JMIR Formative Research

Impact Factor (2023): 2.0
Volume 9 (2025) ISSN 2561-326X Editor in Chief: Amaryllis Mavragani, PhDc

Contents

Original Papers

Assessing Patient Acceptance of Being Offered a Virtual Calm Room in Psychiatric Inpatient Care: Mixed Methods Study (e75961)
Maria Ilioudi, Steinn Steingrimsson, Sara Wallström, Almira Osmanovic Thunström, Philip Lindner, Michael Ioannou, Örjan Falk, Lilas Ali. . . 7 4

Pro- and Antifluoride Use Messages on YouTube in Japan: Content Analysis (e82265)
Hikari Nagao, Tsuyoshi Okuhara, Hiroe Suzuki-Chiba, Hiroko Okada, Takahiro Kiuchi. 83

Implementing Digital Tools for Mental Health Support in Young Individuals in Colombia: Mixed Methods Feasibility Study (e69749)
Laura Ospina-Pinillos, Débora Shambo-Rodríguez, María Riaño-Fonseca, Mónica Sánchez Nítola, María Ramírez-Castro, María Calvo-Valderrama, Salvador Camacho, Carlos Gómez-Restrepo, Alvaro Navarro-Mancilla, Ian Hickie, Jo-An Occhipinti. 96

Development of an mHealth App Prototype for LGBTQIA+ Individuals' Sexual and Reproductive Health in Gauteng Province, South Africa: Design Science Research Study (e79593)
Raikane Seretlo, Mathildah Mokgatle, Hanlie Smuts, Nombulelo Sepeng. 116

Automated Feedback After Internet-Based Depression Screening: Cost-Effectiveness Analysis of a Randomized Controlled Trial (e68282)
Léon Kreis, Hans-Helmut König, Franziska Sikorski, Bernd Löwe, Sebastian Kohlmann, Christian Brettschneider. 134

Facilitators, Barriers, and Cultural Appropriateness of Mindfulness-Based Interventions Among Saudi Female University Students: Qualitative Study (e78532)
Duaa Alrashdi, Carly Meyer, Rebecca Gould. 146

Comparing ChatGPT and DeepSeek for Assessment of Multiple-Choice Questions in Orthopedic Medical Education: Cross-Sectional Study (e75607)
Chirathit Anusitviwat, Sitthiphong Suwannaphisit, Jongdee Bvonpanttarananon, Boonsin Tangtrakulwanich. 165

Gamified Simulation for Onboarding Health Care Teams in Emergency Care: Development and Preliminary Feasibility Study (e72202)
Stephane Gobron, Antoine Lestrade, Artan Sadiku, Alexandre Bentvelzen, Leila Ait Kaci, Jean-Michel Carrier, Emmanuelle Guyot, Pierre-Nicolas Carron. 172

Factors Associated With Suicidal Ideation Among Persons With Disabilities in South Korea: Retrospective Observational Study (e80261)
In-Hwan Oh, Sooyeon Jo, Joon Lee. 189

XSL•FO
RenderX

Information Extraction of Doctoral Theses Using Two Different Large Language Models vs Health Services Researchers: Development and Usability Study (e77707)
Jonas Cittadino, Pia Traulsen, Teresa Schmahl, Larisa Wewetzer, Julia Cummerow, Kristina Flägel, Christoph Strumann, Katja Goetz, Jost Steinhäuser. 198

Everyday Digital Support to Promote Health and Literacy Among Older Adults: 14-Week Randomized Digital Pilot Trial by Engagement Level (e77319)
Andressa da Silva Sobrinho, Guilherme da Silva Rodrigues, Guilherme de Oliveira, Grace de Oliveira Gomes, Carlos Bueno Júnior. 211

Evaluating a Clinical Decision Support Tool for Cancer Risk Assessment in Primary Care: Simulation Study of Unintended Weight Loss (e79208)
Javiera Martinez-Gutierrez, Sophie Chima, Lucas De Mendonca, Alex Lee, Barbara Hunter, Jo-Anne Manski-Nankervis, Deborah Daly, George Fishman, Kit Huckvale, Fong Lim, Benny Wang, Craig Nelson, Brian Nicholson, Jon Emery. 225

Radiomics-Based AI Model to Assist Clinicians in Intracranial Hemorrhage Diagnosis: External Validation Study (e81038)
Salita Angkurawaranon, Natipat Jitmahawong, Kittisak Unsrisong, Phattanun Thabarsa, Chakri Madla, Withawat Vuthiwong, Thanwa Sudsang, Chaisiri Angkurawaranon, Patrinee Traisathit, Papangkorn Inkeaw. 241

Lessons Learned from Co-Designing a Digital Health App for Foster Youth: Development and Usability Study (e73281)
Johanna Folk, Juan Gonzalez, Margareth Del Cid, Elizabeth McBride, Tylia Lundberg, Alison Czopp, Ifunanya Ezimora, Lisa Fortuna, Marina Tolou-Shams. 253

Comparing Human and AI Therapists in Behavioral Activation for Depression: Cross-Sectional Questionnaire Study (e78138)
Inka Napiwotzki, Julian Laue, Flora Caldarone, Maximilian Idahl, Uwe Hadler, Haithem Amrani, Elisabeth Hildt, Kai Kahl, Wolfgang Nejdl. . .
2 6 6

Advancing Digital Access to Physical Therapy via Virtual and Extended Reality Technology: Prototype Development and Usability Evaluation (e73783)
Victoria Tiase, Julie Fritz, Jesse Ferraro, Gregory Bayles, Ahmad Alsaleem, Guilherme Del Fiol, Kensaku Kawamoto, Darrel Brodke, Brook Martin, Roger Altizer. 276

Understanding Physician Attitudes Toward AI in Clinical Decision-Making: Cross-Sectional Study (e79730)
Fahad Alhazmi. 288

Gender Differences in Trajectories of Depressive Symptoms Among Talkspace Clients: Naturalistic Observational Study (e75290)
Meghan Romanelli, Julien Rouvere, Isaac Sanders, Aarthi Padmanabhan, Emily Peake, Thomas Hull, Tim Althoff. 296

YouTube Searching and Self-Treatment Behaviors Among Patients With Benign Paroxysmal Positional Vertigo Before and After Clinic Visits: Prospective Observational Study (e80966)
Guil Rhim, Moon Kim. 314

Developing eHealth Interventions to Improve Diabetes Management in Emerging Adulthood: Qualitative Formative Study (e75623)
April Idalski Carcone, Deepika Baskar, Aishah Mahmood, Karen MacDonell, Susan Eggly, Samiran Ghosh, Colleen Buggs-Saxton, Steven Ondersma, Deborah Ellis. 322

Smartphones and Mental Health Awareness and Utilization in a Low-Income Urban Community: Focus Group Study (e65650)
Nadia Alam, Domenico Giacco, Bulbul Siddiqi, Swaran Singh, Sagar Jilka. 337

XSL•FO
RenderX

Web-Based AI-Driven Virtual Patient Simulator Versus Actor-Based Simulation for Teaching Consultation Skills: Multicenter Randomized Crossover Study (e71667)
Edward Tyrrell, Sardip Sandhu, Kathryn Berry, Suzan Ghannam, Sarah Lewis, Daniel Crowfoot, Gurvinder Sahota, Julie Carson, Emma Wilson, Jaspal Taggar. 349

Teaching Clinical Reasoning in Health Care Professions Learners Using AI-Generated Script Concordance Tests: Mixed Methods Formative Evaluation (e76618)
Alexandre Hudon, Véronique Phan, Bernard Charlin, René Wittmer. 365

Evaluating the Effectiveness of Generative AI for the Creation of Patient Education Materials on Coronary Heart Disease: A Comparative Study (e78816)
Xiaofang Jiang, Jingbang Liu, Shanshan Dai, Xiawen Mao, Rongping Cha, Lili Wu. 378

Leveraging No-Code Digital Platforms for Designing an Integrated Smartphone-Based Ecological Momentary Intervention With Cognitive Behavioral Therapy for Mental Health Care: Development and Usability Study (e77036)
Stephanie Wong, Melody Ip, Jessica Lee, Terry Lum, Ralf Schwarzer. 391

Chatbots' Empathetic Conversations and Responses: A Qualitative Study of Help-Seeking Queries on Depressive Moods Across 8 Commercial Conversational Agents (e71538)
Hyojin Chin, Gumhee Baek, Chiyoung Cha, Meeyoung Cha. 416

Impact of AI on Breast Cancer Detection Rates in Mammography by Radiologists of Varying Experience Levels in Singapore: Preliminary Comparative Study (e66931)
Serene Goh, Hao Du, Loon Tan, Edward Seah, Wai Lau, Alvin Ng, Shi Lim, Han Ong, Samuel Lau, Yi Tan, Mun Khaw, Chee Yap, Kei Hui, Wei Tan, Haziz Abdul, Vanessa Khoo, Shuliang Ge, Felicity Pool, Yun Choo, Yi Wang, Pooja Jagmohan, Premilla Gopinathan, Mikael Hartman, Mengling Feng. 433

Accuracy of Large Language Model Responses Versus Internet Searches for Common Questions About Glucagon-Like Peptide-1 Receptor Agonist Therapy: Exploratory Simulation Study (e78289)
Sarah Tan, Gerald Sng, Phong Lee. 445

Usability, Benefits, and Barriers Associated With Patients' Access to Electronic Health Record–Integrated Telehealth in Hospitals in Riyadh: Qualitative Study (e74011)
Dalia Alomar, Iliada Eleftheriou, Pauline Whelan, John Ainsworth. 455

Evaluating Locally Run Large Language Models (Gemma 2, Mistral Nemo, and Llama 3) for Outpatient Otorhinolaryngology Care: Retrospective Study (e76896)
Christoph Buhr, Christopher Seifen, Katharina Bahr-Hamm, Tilman Huppertz, Johannes Pordzik, Harry Smith, Tom Kelsey, Andrew Blaikie, Christoph Matthias, Sebastian Kuhn, Jonas Eckrich. 468

Populations and Health Domains Served by Direct-to-Consumer Digital Health Companies in the United States, 2011-2023: Cross-Sectional Study (e78431)
Ashwini Nagappan, Xi Zhu, Corrina Moucheraud, Olivia Jung. 484

Acceptability, Relevance, and Short-Term Outcomes of the STAC-T Bullying Bystander App: Feasibility Quantitative Study (e76830)
Diana Doumas, Aida Midgett, Robin Hausheer, Amanda Winburn, Mary Buller, Taylor Perron, Jennalyn Shelton, Brandon Herbeck. 495

Usability Evaluation of a Mixed Reality Platform in Pediatric Interventional Cardiology by Specialist Physicians: Mixed Methods Study (e79278)
Catherine Chilanga, Andreas Jahnen, Linda Hafskjold. 505

Shesha, a WhatsApp Chatbot for Linking Household Contacts to Tuberculosis Treatment or Preventive Therapy in South Africa: Design and Development (e71793)
Don Mudzengi, Thobani Ntshiqa, Yohhei Hamada, Felex Ndebele, Thapelo Mpanza, Bridget Kyobutungi, Candice Williams, Meghan Kennealy, Molebogeng Rangaka, Kavindhran Velen, Salome Charalambous. 519

XSL•FO
RenderX

Awareness and Attitudes of University Students in Bangladesh Toward Cancer: Cross-Sectional Study (e75651)
Maliha Tabassum, Nafisa Farhin, Afsana Afrose, Mahfuza Surovy, Rubaiya Tasnim, Most. Affia Heaven, Jannatun Noor, Munima Haque. . . . 5 3 8

Usability and User Experience of a Digital Platform Prototype (Healthy Bone) to Promote Pharmacological and Nonpharmacological Treatment in Patients With Osteoporosis: Mixed Methods Study (e72468)
Anabela Barcelos, Alexandre Moniz, Ana Henriques, Carolina Mazeda, Elsa Frazão Mateus, Ana Machado, Cátia F Gonçalves, Maiara Moreto, Mafalda F Almeida, Nuno Mendonça, Daniela Costa, Marta M Marques, Pedro Domingos, Yoann Nesme, Helena Canhão, Sónia Dias, Ana Rodrigues. 547

Evaluation of a Digital, Self-Administered, Cognitive Test Battery in Older Adult Patients Undergoing Abdominal Surgery: Nonrandomized Feasibility Trial (e71911)
Anahita Amirpour, Markus Saarijärvi, Jeanette Eckerblad, Gabriela Markovic, Anders Thorell, Ulrica Nilsson, Lina Bergman. 569

Designing a Gait Recognition Algorithm for Older Adults Using Mobility Aids: Prospective Cohort Study (e68669)
Samantha Ray, Jung Koh, Amanda Liberty, Tracy Hammond, Paula Shireman. 584

Digital Smoking Cessation Preferences of Predominately Low-Income and Latino Residents of the San Joaquin Valley in California: Qualitative Study (e74105)
Karla Llanes, Maya Vijayaraghavan, Sara Schneider, Pamela Ling, Evi Hernandez, Paul Brunetta, Anna Song, Arturo Durazo. 592

Active Play in a Digital Age, Exploring Children's (Aged 8-13 Years) Views of a Physical Activity App: Qualitative Formative Study (e76498)
Marie Frazer, Lauren Charlesworth, Luca Wilson, Jennifer Hall, Farwa Batool, Mikel Subiza-Perez, Jan Burkhardt, Andy Daly-Smith, Anna Chalkley. 606

Perceptions, Usage, and Educational Impact of ChatGPT Among Medical Students in Germany: Cross-Sectional Mixed Methods Survey (e81484)
Anna Fußhöller, Fabian Lechner, Nadine Schlicker, Felix Muehlensiepen, Andreas Mayr, Sebastian Kuhn, Martin Hirsch, Johannes Knitza. . . 6 2 6

Dementia Caregivers' Perspectives on Technology's Place in Care Practices: Mixed Methods Survey (e69596)
Julia Scott, Emma Cepukenas, McKenzie Himes, Kennedy Anderson, Kiren Grewal, An Mai, Sheila Yuter, Patricia Simone. 634

Health Service Early-Stage Digital Adaptation of Traditional Chinese Medicine Internet Hospitals: Qualitative Exploratory Study (e77686)
Yao Wang, Menghuan Song, Zhenmiao Pang, Buping Liu, Dongning Yao, Meng Li, Zhirong Xie, Ying Bian, Hao Hu, Yunfeng Lai. 648

Personal Activity Trackers and Family Engagement in a Pediatric Obesity Intervention: Randomized Controlled Trial (e70341)
Juan Espinoza Salomon, Mahsa Babaei, Alexis Deavenport-Saman, Olga Solomon, Choo Wee, Ramon Durazo-Arvizu, Abu Sikder, Payal Shah, Patricia Castillo, Larry Yin. 665

Teenage Body Image Perception, Body-shaping Behavior, and Body Composition With Respect to Use of "Fitspiration": Exploratory Investigation Study (e70964)
Elisabeth Höld, Theres Rathmanner, Mario Heller. 681

Social Cohesion, Mental Well-Being, and the Role of Smart Technology and Pet Ownership Among Social Housing Residents: Cross-Sectional Cohort Study (e75451)
Emmylou Rahtz, Andrew Williams, Timothy Taylor. 697

XSL•FO
RenderX

Digital Mental Health Coaching in Clinically Diverse Populations: Controlled Engagement and Outcomes Study (e71346)
Alison Pickover, Sarah Adler. 709

Information Deficits, Information Needs, and Preferences Regarding eHealth in a Dutch Population With Metabolic Dysfunction–Associated Steatotic Liver Disease: Cross-Sectional Survey Study (e78536)
Sharon Oude Veldhuis, Maureen Guichelaar, Marjolein den Ouden, Lisette van Gemert-Pijnen, Constance Drossaert. 718

Design, Implementation, and Evaluation of a Community-Based Phygital Telemonitoring Program for Older Adults: Multisite Retrospective Pilot Study in Singapore (e56905)
Yichi Zhang, Michelle Law, Soon Wee, Benjamin Lee, Bing Chew, Wei Teo, Edmund Lee. 730

Feasibility of a Text Message–Based Alcohol Prevention Intervention for Parents of Rising Middle School Students: Randomized Controlled Trial (e72823)
Marina Pearsall, Melissa Gilkey, Susan Ennett, H McNaughton Reyes, Nisha O'Shea. 748

Predicting Risk of Heat-Related Injuries for Individuals Wearing Personal Protective Equipment Using Smartwatches: Feasibility Observational Study (e72324)
Meghan Hegarty-Craver, Donna Womack, Jonathan Thornburg, Timothy Boe, M Archer, Worth Calfee. 762

Development, Implementation, and Usability Evaluation of the CANMI App to Monitor the Quality of Maternal and Child Nutrition Care in Primary Health Units: Mixed Methods Pilot Study (e77539)
Soraya Burrola-Méndez, Elizabeth Hoyos-Loya, J Quiroz-Ibarra, Cecilia Perez Navarro, Isabel Omaña-Guzmán, Jorge González-Ordiano, Omar Acosta-Ruiz, Arturo Bautista-Morales, Daniela Tánchez Sandoval, Sonia Hernández-Cordero, Cinthya Muñoz-Manrique, Monica Ancira-Moreno.
7 7 4

Transition From a Paper Diary to an Electronic Diary by Parents of Preschool Children With Food Allergies: Pilot Study (e73400)
Yoriko Kato. 785

Integration of an Artificial Intelligence–Based Autism Diagnostic Device into the ECHO Autism Primary Care Workflow: Prospective Observational Study (e80733)
Kristin Sohl, Erik Linstead, Kelianne Heinz, Elia Lledo, Alicia Brewer Curran, Melissa Mahurin, Valeria Nanclares-Nogués, Carmela Salomon, Minda Seal, Sharief Taraman. 802

Effectiveness of WeChat-Based Plus Scene-Graphics Health Education for Rehabilitation After Open Elbow Arthrolysis: Historical Control Study (e58218)
Danling Fang, Yin Wang, Shiyang Yu, Wei Wang, Pingchen Lu. 815

Assessing the Readability and Quality of Web-Based Resources on Exercise Stress Testing: Cross-Sectional Readability and Quality Analysis (e68000)
Munir Rahbe, Dhrumi Mistry, Ali Khawaja, Natalie Sous, Alan Tso. 833

Participatory Development and Concept Testing of mHealth Messaging to Support Care Engagement and Antiretroviral Therapy Adherence for Women Living With HIV in the Southern United States: Focus Group Study (e76286)
Sadie Sommer, Julie Barroso, Sarah Bass, Marianne Choufani, Alexander Schoemann, Katie Singley, Caseem Luck, Courtney Caiola. 843

Evaluating Large Language Models for Sentiment Analysis and Hesitancy Analysis on Vaccine Posts From Social Media: Qualitative Study (e64723)
Augustine Annan, Amanda Eiden, Dong Wang, Jingcheng Du, Majid Rastegar-Mojarad, Varun Nomula, Xiaoyan Wang. 860

Effectiveness, Usability, and Satisfaction of a Self-Administered Digital Intervention for Reducing Depression, Anxiety, and Stress in a University Community in the Andean Region of Peru: Randomized Controlled Trial (e71465)
Rosario Yslado-Méndez, Stefan Escobar-Agreda, David Villarreal-Zegarra, Wilfredo Trejo Flores, Junior Sánchez-Broncano, Ana Vilela-Estrada, Jovanna Olivares Córdova, C Reategui-Rivera, Claudia Alvarez-Yslado, Leonardo Rojas-Mezarina. 879

XSL•FO
RenderX

A Remotely Delivered Weight Management Service to Support Existing Obesity Services in the UK National Health Service: Preliminary Findings From an Early-Stage Service Evaluation (e71914)
Giulia Spaltro, Michael Whitman, Rebecca Richards. 894

Daily Household Electricity Consumption in Community-Dwelling Older Individuals With Cognitive Impairment: Prospective Cohort Study (e71265)
Yuki Nakagawa, Shigeo Tanabe, Hikaru Kondo, Koki Tan, Soichiro Koyama, Shin Kitamura, Akiko Kada, Takuma Ishihara, Takuaki Yamamoto, Junya Denda, Hideaki Kimata, Taisuke Yamanaka, Ryosuke Umezawa, Yoshinobu Nakahashi, Yohei Otaka. 915

Measuring FVIII Activity and Thrombin Generation Simultaneously With a Novel Point of Care Platform (EnzySystem HemA): Qualitative Usability Evaluation (e77621)
Aernoud Bavinck, Saskia Schols, Wilfred Teunissen, Amy Shapiro, Nicole Blijlevens, Kyle Davis, Waander van Heerde. 931

Feasibility of a Guided Web-Based Procrastination Intervention for College Students: Open Trial (e72896)
Sevin Ozmen, Arpana Amarnath, Leonore de Wit, Chris van Klaveren, Pim Cuijpers, Annemieke van Straten, Sascha Struijs, Caring Universities Consortium. 948

A Machine Learning–Based Scoring System to Identify High Immunoactivity Microsatellite Stability Tumors by Quantifying Similarity to Microsatellite Instability-High Tumors in Colorectal Cancers: Development and Quantitative Study (e66960)
Hongkai Yan, Li Jiang, Yaqi Li, Fengchong Wang, Shaobo Mo, Weiqi Sheng, Dan Huang, Junjie Peng. 964

Clinical Usefulness of a Smartphone-Based 6-Minute Walk Test in a Hospital Outpatient Clinic Within the Constraints of the COVID-19 Pandemic: Mixed Methods Study (e70495)
Dario Salvi, Carl Olsson, Jackson Molloy, Elizabeth Orchard. 979

Exploring the Acceptability and Suitability of Synchronous Online Focus Groups for Health Research With Métis Nation of Ontario Citizens: An Internet-Based Survey (e70986)
Abigail Simms, Noel Tsui, Robynn Sadler, Amy Mersereau, Cindi Rye, C Crenna, Jeff Evenson, Sarah Edwards. 990

Noninvasive Multiparameter Monitoring for the Detection of Decompensated Heart Failure: Exploratory Study (e59116)
Cyrille Herkert, Mayke van Leunen, Ignace De Lathauwer, Valerie van Es, Jialu Tang, Aaqib Saeed, Rudolph Spee, Yuan Lu, Hareld Kemps. .
1 0 0 1

A Digital, Self-Management Behavior Change Intervention for People With Chronic Obstructive Pulmonary Disease: Cohort Study (e75683)
Liam Knox, Rachel Rahman, Gareth Norris, Carol-Anne Davies, Kimberley Littlemore, Claire Hurlin, Sam Rice, Keir Lewis. 1018

Probing the Relationship Between Perioperative Complications in Patients With Valvular Heart Disease: Network Analysis Based on Bayesian Network (e68710)
Wenyuan Lu, Kun Zhu, Zhiliang Gao, Yuanming Li, Hanwei Tang, Cheng Sun, Jianfeng Hou. 1027

Use of a Preliminary Artificial Intelligence-Based Laryngeal Cancer Screening Framework for Low-Resource Settings: Development and Validation Study (e66110)
Shao Lam, Min Lee, Michael Dorosan, Samuel Altonji, Hiang Tan, Walter Lee. 1041

Feasibility and Acceptability of a Positive Psychological Intervention for Patients With Metastatic Breast Cancer: Pre-Post Pilot Study (e77636)
Claire Conley, Elizabeth Addington, Mikaela Velazquez-Sosa, Brenna Mossman, Lesley Glenn, Shontè Drakeford, Roxana Guerra, Claudine Isaacs, Ami Chitalia, Christopher Gallagher, Suzanne O'Neill, Judith Moskowitz. 1050

XSL•FO
RenderX

Designing Digital Mental Health Tools to Support the Needs of Black Adults in the United States: Qualitative Analysis (e73279)
Sarah Popowski, Jonah Meyerhoff, Olivia Allen, Theresa Nguyen, Terika McCall, Aderonke Pederson, Madhu Reddy, David Mohr, Rachel Kornfield. 1065

The Impact of Dose in an mHealth Intervention to Support Parents and Carers Via Healthy Beginnings for Hunter New England Kids Program: Pragmatic Randomized Controlled Trial (e70158)
Alison Brown, Nayerra Hudson, Jessica Pinfold, Rebecca Sewter, Lynda Davies, Christophe Lecathelinais, Jacklyn Jackson, Tessa Delaney, Sienna Kavalec, Rachel Sutherland. 1083

Multiuser Application for the Diagnosis and Treatment of Depression in Women's Self-Help Groups: Pilot Randomized Controlled Trial (e68052)
Amritha Bhat, Ruben Johnson-Pradeep, Bharat Kalidindi, Dhinagaran Devadass, B Goud, Tony Raj, Sumithra Selvam, Yesenia Navarro-Aguirre, Pamela Collins, Krishnamachari Srinivasan. 1093

Avatar Customization and Embodiment in Virtual Reality Self-Compassion Therapy for Depressive Symptoms: Three-Part Mixed Methods Study (e71004)
Thomas Elliott, Yanzhuo Yang, Jarrod Knibbe, Julie Henry, Nilufar Baghaei. 1101

Internet-Delivered Psychological Treatment for Parents With Health Anxiety by Proxy: Replicated Randomized Single-Case Experimental Design (e65396)
Katrine Ingeman, Ditte Frydendal, Lisbeth Frostholm, Ellen Bjerre-Nielsen, Kaare Wellnitz, Patrick Onghena, Kristi Wright, Charlotte Rask. . .
1 1 1 8

Dietary Fiber Estimate of DialBetesPlus App Users: Secondary Analysis of Data From a Randomized Controlled Trial (e69340)
Wei Sze, Kayo Waki, Ryohei Nakada, Toshimasa Yamauchi, Masaomi Nangaku, Kazuhiko Ohe. 1134

Motives for Cannabis Use and Readiness to Change Among Users of the "Stop-Cannabis" Mobile App: Cluster Analysis (e70849)
Milena Wegener, Stéphane Rothen, Elise Dan-Glauser, Tania Lecomte, Stéphane Potvin, Lucien Rochat, Marissa Sjöblom, Germano Vera Cruz, Jean-François Etter, Yasser Khazaal. 1143

Professional Support Through a Tailor-Made Mobile App to Reduce Stress and Depressive Symptoms Among Family Caregivers of People With Dementia: Mixed Methods Pilot Study (e75113)
Aber Kagwa, Jessica Longhini, Muhammed Islam, Sofia Vikström, Åsa Dorell, Hanne Konradsen, Zarina Kabir. 1155

The Validation and Accuracy of Wearable Heart Rate Trackers in Children With Heart Disease: Prospective Cohort Study (e70835)
Hidde Hardon, Yara Van Kerkhof, Beatrijs Bartelds, Janneke Kammeraad, Arend Van Deutekom. 1168

AI-Enabled, Text-Based Health Coaching and Navigation for Employees to Support Health Outcomes: Pre-Post Observational Study (e64553)
Paula Wilbourne, Susan Mirch-Kretschmann, Denise Walker, Michael Varghese, Roberto Arnetoli. 1186

Bedtime App–Guided Mindfulness Meditation in Patients With Insomnia: Mixed Methods Feasibility and Acceptability Pilot Study (e67366)
Yan Ma, Peter Wayne, Janet Mullington, Suzanne Bertisch, Gloria Yeh. 1199

A Series of Personalized Melatonin Supplement Interventions for Poor Sleep: Feasibility Randomized Crossover Trial for Personalized N-of-1 Treatment (e58192)
Mark Butler, Thevaa Chandereng, Heejoon Ahn, Stefani Slotnick, Danielle Miller, Alexandra Perrin, Jordyn Rodillas, Ciaran Friel, Ashley Goodwin, Ying Cheung, Karina Davidson. 1217

Health Care Content and Engagement in Chronic Illness Instagram Posts: Content Analysis (e57523)
Lindsey Daon, Fu-Shiuan Lee, Olga Saynina, C Wang. 1242

XSL•FO
RenderX

Disease Activity Assessment Frequency in Rheumatoid Arthritis: A Retrospective Observational Study of the Medical Support System for Rheumatoid Arthritis System Implementation (e74222)
Mari Yamamoto, Yuki Kataoka, Yasushi Tsujimoto, Fumika Nagase, Keita Iwasaki, Yuuki Ito, Hiroki Ikai, Tsuyoshi Watanabe, Waka Yokoyama-Kokuryo, Naoho Takizawa, Yoshiro Fujita. 1258

A Behavioral Science-Informed Agentic Workflow for Personalized Nutrition Coaching: Development and Validation Study (e75421)
Eric Yang, Tomas Garcia, Hannah Williams, Bhawesh Kumar, Martin Ramé, Eileen Rivera, Yiran Ma, Jonathan Amar, Caricia Catalani, Yugang Jia. 1268

Assessing the User Experience of the EU Mobile App for Cancer Prevention: Mixed Methods Study (e73844)
Furqan Ahmed, Silvana Romero Saletti, Erica D'Souza, Carolina Espina, David Ritchie, Ana Barceló, Marina Carbó, Paula Cervera, Teresa Seum, Hermann Brenner, Stephan Van den Broucke, Maria Krini, Cristiana Fonseca, Patricia Pinto, Diana Krivic, Helena Comesana, Wendy Yared, Hajo Zeeb, Tilman Brand. 1282

Knowledge and Perceptions of AI Among Medical Students in Morocco: Cross-Sectional Study (e66156)
Imad Chakri, Otmane El Khayali, Laila Lahlou. 1296

Understanding Health Care Workers' Attitudes and Preferences Toward Digital Patient Monitoring Platforms: Cross-Country Survey Study (e67142)
Costanza Tortù, Chiara Seghieri, Ditila Doracaj, Natalya Usheva, Natalia Giménez-Legarre, Yannis Manios. 1309

Typology and Ethical Considerations of Digital Health Promotion Tools for Youth in Sub-Saharan Africa: Review of Examples From Ghana, Kenya, and South Africa (e54472)
Agata Ferretti, Shannon Hubbs, Richard Dzikunu, Keymanthri Moodley, Frederick Wekesah, Jonty Wright, Effy Vayena. 1323

Personality-Driven Variations in Fitness App Affordance Actualization Among Adults: Quantitative Survey Study (e72691)
Moayad Alshawmar, Bengisu Tulu, E Wilson, Adrienne Hall-Phillips. 1331

Design of an Automated Mobile Phone-Based Reminder and Incentive System: Application in a Quasi-Randomized Controlled Trial to Improve the Timeliness of Childhood Vaccinations in Tanzania (e65150)
Marco van Zwetselaar, Jan Ostermann, Melkiory Beti, Joy Baumgartner, Sayoki Mfinanga, Esther Ngadaya, Lavanya Vasudevan, Nathan Thielman. 1344

Implementation of Fully Automated AI-Integrated System for Body Composition Assessment on Computed Tomography for Opportunistic Sarcopenia Screening: Multicenter Prospective Study (e69940)
Bushra Urooj, Yousun Ko, Seongwon Na, In-One Kim, Eun-Hee Lee, Seon Cho, Heeryeol Jeong, Seungwoo Khang, Jeongjin Lee, Kyung Kim. 1356

AI Model Based on Diaphragm Ultrasound to Improve the Predictive Performance of Invasive Mechanical Ventilation Weaning: Prospective Cohort Study (e72482)
Feier Song, Huazhang Liu, Huan Ma, Xuanhui Chen, Shouhong Wang, Tiehe Qin, Huiying Liang, Daozheng Huang. 1368

Recognizing Skateboard and Kickboard Commuting Behaviors Using Activity Trackers: Feasibility Study Using Machine Learning Approaches (e71969)
Nathanael Aubert-Kato, Hitomi Hatori, Arisa Orihara, Takashi Nakagata, Yuji Ohta, Julien Tripette. 1383

Web-Based Video Platforms as Sources of Information on Body Image Dissatisfaction in Adolescents: Content and Quality Analysis of a Cross-Sectional Study (e71652)
Li Liu, Jianning Yang, Fengmei Tan, Huan Luo, Yanhua Chen, Xiaolei Zhao. 1394

XSL•FO
RenderX

Evaluating ChatGPT's Utility in Biologic Therapy for Systemic Lupus Erythematosus: Comparative Study of ChatGPT and Google Web Search (e76458)
Kai Li, Yunfei Peng, Luyi Li, Bo Liu, Zhijian Huang. 1407

Exploring the Utility of Virtual Clinics for Neurosurgical Patient Consults: Cohort Study to Assess Feasibility (e69372)
Hassan Khayat, Radwan Takroni, Majid Aljoghaiman, Jessy Moore, Mohamed Alhantoobi, Oscar Obiga, Marcos Yasuda, Bill Wang, Almunder Algird, Kesava Reddy. 1420

Usability and Acceptability of an App-Based Approach to Treat Low Back Pain: Preplanned Secondary Analysis of a Randomized Controlled Trial (e59866)
Svenja Kaczorowski, Katharina Trompeter, Carolin Kramer, Christian Grüneberg, Christian Thiel, Lars Donath, Daniel Belavy. 1429

Implementation Status and Usability of Digital Health Interventions Among Health Care Workers and End Users at the Primary Health Care Level in Chandigarh, North India: Cross-Sectional Study (e69824)
Munita Jat, Madhu Gupta, Shankar Prinja. 1443

Encouraging General Practitioners to Refer Patients With Insomnia to a Digital Therapeutic (Sleepio): Feasibility Repeated-Measures Intervention Study (e75359)
Ohoud Alkhaldi, Brian McMillan, John Ainsworth. 1460

Optimizing Intervention Components of a Preventive Stress Management mHealth Intervention for Health Care Workers: Experimental Factorial Study (e71032)
Leo Kowalski, Anna Finnes, Sabine Koch, Aleksandra Bujacz. 1469

Performance Assessment of ChatGPT-4.0 and ChatGLM Series in Traditional Chinese Medicine for Metabolic Associated Fatty Liver Disease: Comparative Study (e66503)
Xionghui Wang, Tianxiao Zheng, Bo Liu, Zhi Pei, Kaihan Meng, Changquan Ling. 1485

External Exposome Factors and Adverse Heart Failure Outcomes in the OneFlorida+ Network: Retrospective Cohort Study (e71595)
Wenxi Huang, Stephen Kimmel, Mustafa Ahmed, Steven Smith, Yao Lee, Lanting Yang, Inmaculada Hernandez, Jiang Bian, Jingchuan Guo.
1 5 0 2

Development of a Theory-Based mHealth App for Fatigue Management in Lupus: Human-Centered Design Approach (e75399)
Anna Deck, Kiran Singh, Paula Caras, Amy LeClair, Monique Gore-Massy, Faye Chiu, Lucas Dantas, Timothy McAlindon, Sara Folta, Shanthini Kasturi. 1516

Transforming Patient Feedback Into Actionable Insights Through Natural Language Processing: Knowledge Discovery and Action Research Study (e69699)
Ravi Shankar, Alexander Yip. 1532

Primary Care Clinician Perspectives on Older Adult Chronic Pain Management and Clinical Decision Support: Qualitative Study (e74381)
Isra Hasnain, Erin Staab, Ainur Kagarmanova, Marissa Mackiewicz, Mim Ari, Katherine Thompson, Danielle Lazar, Anne Zhao, Glyn Elwyn, Christopher Harle, Valerie Press, Neda Laiteerapong. 1547

Technical Description of eHealth Tools in an Environmental Exposure Chamber: Implementation Study (e71276)
Irene Garcia-Gutierrez, Irene Dueñas-Requena, Violeta Sanchez-Garcia, Cristiana Solorzano-Zepeda, Samuel Santos-Benito, Belen de la Hoz-Caballer, Dario Antolin-Amerigo. 1557

Automated Data Harmonization in Clinical Research: Natural Language Processing Approach (e75608)
Pratheek Mallya, Ricardo Henao, Chuan Hong, Daniel Wojdyla, Tony Schibler, Vihaan Manchanda, Michael Pencina, Jennifer Hall, Juan Zhao.
1 5 6 8

XSL•FO
RenderX

The MyRelief Digital Educational Self-Management Program for Persistent Low Back Pain: Feasibility Uncontrolled Trial (e74948)
Caroline Larsson, Joanne Marley, Flavia Piccinini, Sarah Howes, Elisa Casoni, Vincenzo Aschettino, Carlos Vaz de Carvalho, Suzanne McDonough. 1580

Evaluating the Quality and Understandability of Radiology Report Summaries Generated by ChatGPT: Survey Study (e76097)
Alexis Sunshine, Grace Honce, Andrew Callen, David Zander, Jody Tanabe, Samantha Pisani Petrucci, Chen-Tan Lin, Justin Honce. 1592

Design and Evaluation of a Digital Health App (SingaporeWALK) for Active Aging: Pre-Post Intervention Study (e68937)
Huanyu Bao, Sowmiya Siva Subramanian, Sai Pai, Navrag Singh, Kai Tan, Tan Pham, Feihong Pan, Yin-Leng Theng, Edmund Lee. 1601

Balance Improvement and Fall Risk Reduction in Stroke Survivors After Treatment With a Wearable Home-Use Gait Device: Single-Arm Longitudinal Study With 1-Year Follow-Up (e67297)
Brianne Darcy, Lauren Rashford, David Huizenga, Kyle Reed, Stacy Bamberg. 1613

Feasibility and Efficacy of Commercial-Off-the-Shelf Virtual Reality Applications for Managing Chronic Pain and Enhancing Well-Being Among Older Adults in the Community: Mixed Methods Pilot Study (e67765)
Rachel Leung, Megan Ye, Flora Zhang, Tyrone Kwok, Yuying Sun, Agnes Lai, Juming Jiang, Mimi Tse. 1629

Clinical Management of Medication-Assisted Treatment for Opioid Use Disorder Using a Mobile Health App Within a Primary Care Clinic: Quasi-Experimental Study (e63526)
Allison Rosen, Steven Shoptaw, Li Li, Bengisu Tulu, Omar Nieto, Steven Jenkins, Mariah Kalmin. 1648

Optimizing the Postcataract Patient Journey Using AI-Driven Teleconsultation: Prospective Case Study (e72574)
Joukje Wanten, Noël Bauer, Mohita Chowdhury, Aisling Higham, Nick de Pennington, Frank van den Biggelaar, Rudy Nuijts. 1655

AI-Enhanced 3D Models in Global Virtual Reality Case Conferences for Surgical Care in a Low-Income Country: Exploratory Study (e69300)
Miriam Obst, Jan Arensmeyer, Henrik Bonsmann, Andreas Kolbinger, Joel Kigenyi, Francis Oneka, Benard Owere, Joachim Schmidt, Philipp Feodorovici, Jan Wynands. 1667

Comparability of Daily-Life Walking Speed Measured by Smartphone GPS and Ankle-Band Accelerometer: Cross-Sectional Study (e73722)
Hisashi Kawai, Keigo Imamura, Rui Gong, Manami Ejiri, Shuichi Obuchi. 1679

Refining a Video and Text Message Intervention (STAR, Skills Training in Active Recovery) to Prevent the Onset or Escalation of Posttraumatic Stress and Opioid Misuse Among Recent Sexual Assault Survivors: Community Engaged Study (e72095)
Kate Walsh, Anne Schipani-McLaughlin, Cynthia Stappenbeck, Shahzarin Khan, Anna Woodworth, Sanika Panwalkar, Ronald Acierno, Amanda Gilmore. 1687

ADAPTed Cognitive Behavioral Therapy for Pediatric Functional Abdominal Pain in Community-Based Pediatric Care: Mixed Methods Study (e67106)
Emma Ramsay Milford, Sandra Buratti, Natoshia Cunningham, Åsa Nilses, Ewa-Lena Bratt, Sandra Weineland. 1701

Promoting Sustained Real-Life Benefits of Virtual Reality–Based Interventions in People With Mental Health and Substance Use Disorders: Qualitative Study (e57643)
Jan Aasen, Fredrik Nilsson, Torgeir Sørensen, Marja Leonhardt. 1716

Evaluating the Prototype of a Clinical Decision Support System in Primary Care: Qualitative Study (e69875)
Susanne Köhler, Svea Holtz, Michaela Neff, Jannik Schaaf, Michael von Wagner, Beate Müller, Dania Schütze. 1729

XSL•FO
RenderX

Organizational Climate and Decision Aid Sustainability in Lupus Care: Mixed Methods Study (e69603)
Aizhan Karabukayeva, Larry Hearld, Nathan Carroll, Reena Kelly, Jasvinder Singh. 1742

Comparing Approaches to Teaching Patients How to Use an App-Based Home Spirometer: Randomized Controlled Trial (e74125)
Caitlin Morgan, Daniel Higbee, Catherine Dixon, Emma Buckroyd, Huzaifa Adamali, Shaney Barratt, Rahul Shrimanker, Catherine Hyams, Paul White, James Dodd. 1756

Value-Based Framework for Evaluating Pre-Commercial Procurement: Case Study of Value-Based Key Performance Indicators (e71279)
Nils-Otto Ørjasæter, Kari Kværner, Linn Støme. 1768

Image-Based Dietary Assessment Using the Swedish Plate Model: Evaluation of Deep Learning–Based You Only Look Once (YOLO) Models (e70124)
Gustav Chrintz-Gath, Meena Daivadanam, Laran Matta, Steve McKeever. 1782

Using Artificial Intelligence ChatGPT to Access Medical Information About Chemical Eye Injuries: Comparative Study (e73642)
Layan Alharbi, Rema Alrashoud, Bader Alotaibi, Abdulaziz Al Dera, Raghad Alajlan, Reem AlHuthail, Dalal Alessa. 1796

Promoting Dairy Consumption Among Families: Development and User Experience Study of a Web-Based Nutrition Intervention (e66582)
Juliette Lemay, Jacynthe Roberge, Véronique Provencher, Angelo Tremblay, Shirin Panahi, Raphaëlle Jacob, Lucie Brunelle, Gabrielle Saintonge, Vicky Drapeau. 1820

Technology Integration in Syrian Medical Education From the Perspective of Students and Faculty: A Cross-Sectional Evaluation (e76958)
Subhiya Hassoun, Mayssoon Dashash, Adnan Baddour. 1837

Development of a Sham Smartphone App for Opioid Use Disorder: Acceptability and Suitability Study (e71105)
Kierstyn Gallegos, Jennifer Potter, Van King, Gregg Siegel, Leslie Siegel, Elise Marino. 1853

Feasibility and Acceptability of Remotely Accessed Compensatory Cognitive Training for Japanese People With Schizophrenia: Pilot Study (e70916)
Ritsuko Aijo, Mie Matsui. 1862

Assessing the Ability to Use eHealth Resources Among Older Adults: Cross-Sectional Survey Study (e70672)
Bernard Aoun, Jon Ebbert, Priya Ramar, Daniel Roellinger, Lindsey Philpot. 1876

Development of a Customizable Web-Based Dashboard for Remote Blood Pressure Monitoring: Feasibility and Usability Implementation Study (e62700)
Joshua Wollen, Thuy Nguyen, Charles Coton, Gina DeWildt, Gabretta Cooksey, Audrey Cohen, Nicholas Hoang, Munachi Okpala, Michael Gonzalez, Mengxi Wang, Chigozirim Izeogu, Elmer Bernstam, Charles Green, Sean Savitz, Jose-Miguel Yamal, Anjail Sharrief. 1889

Factors Associated With Portal and Telehealth Uptake and Use in a Minoritized, Low-Income Community: Mixed Methods Study (e70146)
Robin Higashi, Emily Repasky, Antara Gupta, MinJae Lee, Catherine DesRoches, Aimee Israel, Sandi Pruitt. 1899

Students' Perceptions of Learning Analytics for Mental Health Support: Qualitative Study (e70327)
Aglaia Freccero, Miriam Onwunle, Jordan Elliott, Nathalie Podder, Julia Purrinos De Oliveira, Lindsay Dewa. 1914

XSL•FO
RenderX

Evaluating Fitbits for Assessment of Physical Activity and Sleep in Pediatric Pain: Feasibility and Acceptability Pilot Study (e59074)
Bridget Nestor, Andreas Baumer, Justin Chimoff, Benoit Delecourt, Camila Koike, Nicole Tacugue, Roland Brusseau, Nathalie Roy, Israel Gaytan-Fuentes, Navil Sethna, Danielle Wallace, Joe Kossowsky. 1926

Agreement Between TDK Silmee W22 and ActiGraph wGT3X-BT for Estimating Daily Step Counts and Moderate to Vigorous Physical Activity in Free-Living Adults: Comparative Study (e64602)
Hiroko Shimura, Shinpei Okada, Kazushi Maruo, Kaori Daimaru, Naoki Deguchi, Yoshinori Fujiwara, Hiroyuki Sasai. 1939

Short-Form Psychoeducation Videos: Process Development Study (e66884)
Louise Turtle, Helen Wesson, Simon Williamson, Nathan Hodson. 1950

Acceptability and Utility of a Web-Based Patient-Completed Clinical Decision Aid for the Differential Diagnosis of Transient Loss of Consciousness: Qualitative Interview Study (e67608)
Alistair Wardrope, Lindsay Blank, Melloney Ferrar, Steve Goodacre, Daniel Habershon, Markus Reuber. 1959

Translation, Cultural Adaptation, and Mixed-Methods User Feedback of the SelfBack App to the Arabic Language for Patients With Low Back Pain: Pilot Mixed-Methods Study (e59777)
Mai Aldera, Hana Alsobayel, Jonathan Hill, Christian Jensen. 1982

Generative AI–Powered Mental Wellness Chatbot for College Student Mental Wellness: Open Trial (e71923)
Jazmin Reyes-Portillo, Amy So, Kelsey McAlister, Christine Nicodemus, Ashleigh Golden, Colleen Jacobson, Jennifer Huberty. 1991

Feasibility of a Mental Health App Intervention for Emergency Service Workers and Volunteers: Single-Arm Pilot Study (e50995)
Denise Meuldijk, Mark Deady, Daniel Collins, Douglas Williams, Richard Bryant, Samuel Harvey. 2005

Gait Disturbances in Older Adults With Cerebral Small Vessel Disease: Mixed Methods Study Using Smartphone Sensors and Video Analysis (e58864)
Xiaojun Lai, Li-Yan Qiao, Pei-Luen Rau, Yankuan Liu. 2021

Barriers to Patient Portal Adoption Among a Bilingual Patient Population by Analysis of Survey Findings from English- and Spanish-Speaking Patients: Information Needs Study (e66717)
Jiahua Yang, Michael Mackert, Daniela De Luca, Sophia Dove. 2032

Toward a Smartphone-Based and Conversational Agent–Delivered Just-in-Time Adaptive Holistic Lifestyle Intervention for Older Adults Affected by Cognitive Decline: Two-Week Proof-of-Concept Study (e66885)
Esther Brill, Rasita Vinay, Marcia Nißen, Priyam Joshi, Stefan Klöppel, Tobias Kowatsch. 2040

Virtual Reality Interventions of Daily Versus Weekly Data Collection in Patient-Reported Outcomes Among Adults With Cancer: Pilot Survey Study (e73506)
Matthew Browning, Olivia McAnirlin, Fu Li, Jeffrey Bertrand, Denise Davis, Kapil Madathil, George Mau, Teny Henry Gomez. 2058

K-Means Clustering and Classification of Breast Cancer Images Using Histogram of Oriented Gradients Features and Convolutional Neural Network Models: Diagnostic Image Analysis Study (e71974)
Said Salloum. 2072

Usefulness of Interventions Using a Smartphone Cognitive Behavior Therapy Application for Children With Mental Health Disorders: Prospective, Single-Arm, Uncontrolled Clinical Trial (e60943)
Shinichiro Nagamitsu, Ayumi Okada, Ryoichi Sakuta, Ryuta Ishii, Kenshi Koyanagi, Chizu Habukawa, Takashi Katayama, Masaya Ito, Ayako Kanie, Ryoko Otani, Takeshi Inoue, Tasuku Kitajima, Naoki Matsubara, Chie Tanaka, Chikako Fujii, Yoshie Shigeyasu, Michiko Matsuoka, Tatsuyuki Kakuma, Masaru Horikoshi. 2083

XSL•FO
RenderX

Game-Based Social-Emotional Learning for Youth: School-Based Qualitative Analysis of Brain Agents (e67550)
Elizabeth Liverman, David Antognoli, Cordelia Elaiho, Madison McGuire, Abbey Stoltenburg, Angel Navarrete, Garrett Bates, Thomas Chelius, Constance Gundacker, Paula Lumelsky, Brandon Currie, John Meurer Sr. 2096

Digital Infrared Thermographic Imaging for Limb Salvage in Patients at Risk of Amputation: Prospective Observational Study (e69072)
Víctor Loza-González, Eleazar Kolosovas-Machuca, Patricia Cervantes-Báez, José Ramírez-GarcíaLuna, Edgar Guevara, Mario Martínez-Jiménez. 2110

GamePlan4Care, a Web-Based Adaptation of the Resources for Enhancing Alzheimer's Caregiver Health II Intervention for Family Caregivers of Persons Living With Dementia: Formative, Qualitative Usability Testing Study (e60143)
Jinmyoung Cho, Thomas Birchfield, Jennifer Thorud, Marcia Ory, Alan Stevens. 2123

Bringing Executive Function Testing Online: Assessment Validation Study (e75687)
Mouna Attarha, Ana Carolina de Figueiredo Pelegrino, Lydia Ouellet, Sarah-Jane Grant, Etienne de Villers-Sidani, Thomas Van Vleet. 2136

Acceptability of a Web-Based Financial Education Intervention for Latino Caregivers: Mixed Methods Evaluation (e70471)
Susanna Mage, Donna Benton, Kathleen Wilber, Rocio Aguila Rodriguez, David Silva, Alexander Gonzalez, Frank Puga, Kylie Meyer. 2146

Reducing Methamphetamine Use in Aboriginal and Torres Strait Islander Communities With the "We Can Do This" Web App: Qualitative Evaluation of Acceptability and Feasibility (e58369)
Leda Sivak, Rachel Reilly, Shani Crumpen, Carla Treloar, Rebecca McKetin, Julia Butt, Yvette Roe, Nadine Ezard, Brendan Quinn, Jack Nagle, Wade Longbottom, Clifford Warrior, James Ward. 2161

Usability and Usefulness of SMS-Based Artificial Intelligence Intervention (Mwana) on Breastfeeding Outcomes in Lagos, Nigeria: Pilot App Development Study (e65157)
Anisha Musti. 2170

Acceptability of Guided Symptom Entry and Asynchronous Clinical Communication Software Among Primary Care Staff: Qualitative Study (e59620)
Riina Raudne, Taavi Tillmann. 2179

Breath-Focused Mindfulness and Compassion Training in Parent-Child Dyads: Pilot Intervention Study (e69607)
Satish Jaiswal, Jason Nan, Seth Dizon, Jessica Young, Suzanna Purpura, James Manchanda, Dhakshin Ramanathan, Dennis Kuo, Jyoti Mishra. 2189

Performance of a Retinal Imaging Camera With On-Device Intelligence for Primary Care: Retrospective Study (e70331)
Matthew Silvestrini, Clarissa Lui, Anil Patwardhan, Ying Chen, Tayyeba Ali, Elie Glik, Honglei Wu, Brian Levinstein, Adrianna Wenz, Nathan Shemonski, Lin Yang, Ian Atkinson, Sam Kavusi. 2206

Remote Patient Monitoring for Global Emergencies: Case Study in Patients With COVID-19 (e66773)
Ramin Ramezani, Wenhao Zhang, Minh Cao, Alex Bui, Antonia Petruse, Amelia Weldon, Arash Naeim. 2220

A Digital Lifestyle App for Hypertension During Pregnancy: Mixed Methods Intervention Development Study Using the Person-Based Approach (e68927)
Lucy Goddard, Katherine Tucker, Nerys Astbury, Cristian Roman, Yuan Chi, Katherine Morgan, Patricia Devitt, Richard McManus. 2234

Future Me, a Prospection-Based Chatbot to Promote Mental Well-Being in Youth: Two Exploratory User Experience Studies (e74411)
Martin Dechant, Eva Lash, Sarah Shokr, Ciarán O'Driscoll. 2251

XSL•FO
RenderX

Use of a Wearable Self-Tracking Instrument by Refugees With Complex Posttraumatic Stress Disorder: A Qualitative Study of Psychotherapeutic Mediation and Engagement (e70511)
Lisa Riisager, Stine Moeller, Jakob Larsen, Thomas Christiansen, Jesper Aagaard, Lotte Huniche. 2264

Correlation of Biomechanical Variables of Lower Extremity Movement During Functional Tests and Tasks in Youth League Football Players: Cross-Sectional Correlation Study (e69046)
Anna Davidovi a, Sergejs Davidovi s, Guna Semjonova, Alexei Katashev, Alexander Oks, Linda Lancere, Signe Tomsone, Maksims Zolovs. . .
2 2 7 8

A Web-Based Well-Being and Resilience Intervention for Family Members and Friends Supporting a Loved One Using Alcohol and Other Drugs: Mixed Methods Pilot Study (e72425)
Steph Kershaw, Jessica Deng, Madeleine Keaveny, Bronte Speirs, Anna Grager, Dara Sampson, Kate Ross, Nicola Newton, Maree Teeson, Frances Kay-Lambkin, Cath Chapman. 2294

Vibrotactile Feedback Strategies for Trunk-Stabilizing Exercises in a Home-Based Scenario: Qualitative Interview Study Among Physiotherapists (e62903)
Kaya Holzmeyer, Lisa-Marie Lüneburg, Luca Oppici, Philipp Flößel, Doris Lachmann, Jens Krzywinski, Susanne Narciss. 2307

Assessment of Recommendations Provided to Athletes Regarding Sleep Education by GPT-4o and Google Gemini: Comparative Evaluation Study (e71358)
Lukas Masur, Matthew Driller, Haresh Suppiah, Manuel Matzka, Billy Sperlich, Peter Düking. 2324

AI in Qualitative Health Research Appraisal: Comparative Study (e72815)
August Landerholm. 2337

Evaluating a Large Language Model's Ability to Synthesize a Health Science Master's Thesis: Case Study (e73248)
Pål Joranger, Sara Rivenes Lafontan, Asgeir Brevik. 2346

Co-Design of a Health Screening Program Fact Sheet by People Experiencing Homelessness and ChatGPT: Focus Group Study (e68316)
Nóra Radó, Orsolya Németh, Sándor Békási. 2354

Feasibility and Acceptability of a Digital Health Portal to Improve HIV Care Engagement Among Kenyan Youth: Mixed Methods Study (e59661)
Eric Nturibi, Jared Mecha, Florence Kaara, Faith Musau, Christine Mwangi, Elizabeth Kubo, Albert Orwa. 2365

Enhancing Diagnostic Accuracy of Ophthalmological Conditions With Complex Prompts in GPT-4: Comparative Analysis of Global and Low- and Middle-Income Country (LMIC)–Specific Pathologies (e64986)
Shona M'gadzah, Andrew O'Malley. 2377

Assessing the Psychosocial Impact of Expressive Writing on Adults With Spinal Cord Injury: Qualitative Study (e71162)
Shelly Xie, Molly McKenna, Kendall Veach, Sydney Williams, Mary Jones, Elizabeth Vander Kamp, Salaam Green, Lauren Edwards, Kimberly Kirklin, Benjamin Jones, Hon Yuen. 2386

Evaluating the Effectiveness of Smart Glasses in Reducing Patient Care Time in Emergency Departments: Cohort Study From the Hangzhou Asian Games (e65617)
Xinwei Jiang, Bangbo Xia, Mohammad Ansari, Huiquan Jiang, Jianjiang Qi, Zhongheng Zhang, Sheng Dai, Pingping Zheng, Yang He, Ning Liu, Pengpeng Chen, Ronghua Luo, Xuchang Qin, Yansong Miao, Junru Dai, Xiaoyu Zhou, Changliang Wang, Hui Chen, Wenbin Xu, Tao Wu, Qiang Shi, Zhonghua Chen, Liping Zhou, Hao Zhang, Yun Xie, Quan Zhang, Bifa Zhou, Xiaohong Pan, Zixi Chen, Libo Zhen, Yaqing Sun, Zelin Lu, Yihao Loh, Shameera Sayer, Jennifer Mochtar, Pannika Wongpraewit, Yifan Wang, Yucai Hong. 2398

XSL•FO
RenderX

Coaching Patients to Understand and Use Patient-Reported Outcome Data: Intervention Design and Evaluation (e65931)
Martha Burla, Brocha Stern, Andrew Berry, Sarah Pila, Patricia Franklin. 2409

Misrepresentation of Overall and By-Gender Mortality Causes in Film Using Online, Crowd-Sourced Data: Quantitative Analysis (e70853)
Calla Beauregard, Christopher Danforth, Peter Dodds. 2424

Impact of a Novel Electronic Medical Record–Integrated Electronic Form (Provider Asthma Assessment Form) and Severe Asthma Algorithm in Primary Care: Single-Center, Pre- and Postobservational Study (e74043)
Matheson McFarlane, Alison Morra, Delanya Podgers, David Barber, M Lougheed. 2435

Public Perception of the Brain-Computer Interface Based on a Decade of Data on X: Mixed Methods Study (e60859)
Mohammed Almanna, Lior Elkaim, Mohammed Alvi, Jordan Levett, Ben Li, Muhammad Mamdani, Mohammed Al-Omran, Naif Alotaibi. 2452

Changes in Mental State for Help-Seekers of Lifeline Australia's Online Chat Service: Lexical Analysis Approach (e63257)
Kelly Mazzer, Sonia Curll, Hakar Barzinjy, Roland Goecke, Mark Larsen, Philip Batterham, Nickolai Titov, Debra Rickwood. 2467

Outcomes of an Advanced Epic Personalization Course on Clinician Efficiency through Use of Electronic Medical Records: Retrospective Study (e68491)
Junye Chen, Hao Lai, Shi Wong, Terry Pan, Er Lim, Zi Liau. 2482

Investigating Awareness and Acceptance of Digital Phenotyping in Dhaka's Korail Slum: Qualitative Study (e65530)
Nadia Alam, Domenico Giacco, Bulbul Siddiqi, Swaran Singh, Sagar Jilka. 2493

Mobile Apps to Prevent Violence Against Women and Girls (VAWG): Systematic App Research and Content Analysis (e66247)
Md Rahman, Nasimul Usmani, Pradip Chandra, Ridwana Manna, Anisuddin Ahmed, Mohammad Shomik, Shams Arifeen, Aniqa Hossain, Ahmed Rahman. 2505

Cocreating a Mobile Health App Providing Physical Activity Recommendations for Older People Living With Parkinson Disease or Dementia: User-Centered Pilot Study (e51831)
Ellen Bentlage, Alberto del Río Ponce, Mona Ahmed, Pilar Gangas, Michael Brach, Jorge Kurano, José Menéndez. 2522

Fluctuations in Daily Happiness and Nervousness Based on Depressive and Anxious Symptoms in Adolescents or Young Adults Across 3 Latin American Cities: Experience Sampling Study (e65732)
Ana Vilela-Estrada, David Villarreal-Zegarra, Nikol Mayo-Puchoc, Nicola Holt, Ángela Flórez-Varela, Catherine Fung, Karen Ariza-Salazar, Fernando Carbonetti, Sumiko Flores, Adriana Carbonel, Natividad Olivar, Carlos Gomez-Restrepo, Luis Brusco, Stefan Priebe, Francisco Diez-Canseco. 2535

Association of Technology-Related Skills and Self-Efficacy With Willingness to Participate in Heart Failure Telemonitoring: Cross-Sectional Observational Study (e68992)
Sharon Cuppen, Mayke van Leunen, Tamara Henken, Mayra Goevaerts, Martijn Scherrenberg, Maarten Falter, Paul Dendale, Hareld Kemps, Willem Kop. 2550

Recommendations for Designing a Digital Health Tool for Blindness Prevention Among High-Risk Diabetic Retinopathy Patients: Qualitative Focus Group Study of Adults With Diabetes (e65893)
Akua Frimpong, Alvaro Granados, Thomas Chang, Julia Fu, Shannan Moore, Serina Applebaum, Bolatito Adepoju, Mahima Kaur, Vignesh Hari Krishnan, Amanda Levi, Terika McCall, Kristen Nwanyanwu. 2563

Feasibility of Data Collection Via Consumer-Grade Wearable Devices in Adolescent Student Athletes: Prospective Longitudinal Cohort Study (e54630)
Danielle Ransom, Brant Tudor, Sarah Irani, Mohamed Rehman, Stacy Suskauer, P Mularoni, Luis Ahumada. 2578

XSL•FO
RenderX

Patient and Public Perceptions of 3D Technologies (Models and Images) to Facilitate Health Care Consultations: Exploratory, Mixed Methods Study (e65235)
Harleen Rai, Morven Miller, Steve Leung, Euan Macleod, Marilyn Lennon. 2590

Feasibility and Usability of a Web-Based Peer Support Network for Care Partners of People With Serious Illness (ConnectShareCare): Observational Study (e70206)
Aricca Van Citters, Megan Holthoff, Colleen Young, Sarah Eck, Amelia Cullinan, Stephanie Carney, Elizabeth O'Donnell, Joel King, Malavika Govindan, David Gustafson, Stephanie Tomlin, Anne Holmes, Ann Bradley, Brant Oliver, Matthew Wilson, Eugene Nelson, Amber Barnato, Kathryn Kirkland. 2599

Developing and Testing a Framework for Learning Online Collaborative Creativity in Medical Education: Cross-Sectional Study (e50912)
Shairah Radzi, Joo Tan, Preman Rajalingam, Jennifer Cleland, Sreenivasulu Mogali. 2612

Personalized Support in Hereditary Breast and Ovarian Cancer After Genetic Counseling by the Chatbot-Based GENIE Mobile App: Proof-of-Concept Wizard of Oz Study (e69115)
Dominik Wolff, Thomas Kupka, Chiara Reichert, Nils Ammon, Steffen Oeltze-Jafra, Beate Vajen. 2629

Usage of a Multipurpose mHealth App Among Adults With Sickle Cell Disease: Randomized Controlled Trial (e67906)
Robert Cronin, Nives Quaye, Xin Liu, Kristina Landes, Lori Crosby, Adetola Kassim, Michael DeBaun, Patrick Schnell. 2643

Collecting Real-Life Psychophysiological Data via Wearables to Better Understand Child Behavior in a Children's Psychiatric Center: Mixed Methods Study on Feasibility and Implementation (e65559)
Karin Hagoort, Kirsten Smeets, Saskia Koldijk, Floortje Scheepers, Fleur Velders. 2656

Engagement With a Smartphone-Delivered Dietary Education Intervention and Its Relation to Dietary Intake and Cardiometabolic Risk Markers in People With Type 2 Diabetes: Secondary Analysis of a Randomized Controlled Trial (e71408)
Linnea Sjoblom, Freja Stenbeck, Ylva Trolle Lagerros, Essi Hantikainen, Stephanie Bonn. 2670

Components of a Digital Storytelling Intervention for Human Papillomavirus and Cancer Prevention Among LGBTQ+ Individuals: Formative Mixed Methods Inquiry (e58163)
Gabrielle Darville-Sanders, Dominique Munroe, Emilie Corluyan, Utibeabasi Ikoiwak, Jennifer Nguyen, Chad Mandula, Portia Thomas, Brandon Sanders. 2695

Clinician-Focused Connected Health Requirements Gathering for Attention-Deficit/Hyperactivity Disorder Through Clinical Journey Mapping: Design Science Study (e53617)
Richard Harris, Deirdre Murray, Angela McSweeney, Frederic Adam. 2709

Application of the Bidirectional Encoder Representations from Transformers Model for Predicting the Abbreviated Injury Scale in Patients with Trauma: Algorithm Development and Validation Study (e67311)
Jun Tang, Yang Li, Keyu Luo, Jiangyuan Lai, Xiang Yin, Dongdong Wu. 2720

Comparative Efficacy of MultiModal AI Methods in Screening for Major Depressive Disorder: Machine Learning Model Development Predictive Pilot Study (e56057)
Donghao Chen, Pengfei Wang, Xiaolong Zhang, Runqi Qiao, Nanxi Li, Xiaodong Zhang, Honggang Zhang, Gang Wang. 2731

Comparison of ChatGPT and Internet Research for Clinical Research and Decision-Making in Occupational Medicine: Randomized Controlled Trial (e63857)
Felix Weuthen, Nelly Otte, Hanif Krabbe, Thomas Kraus, Julia Krabbe. 2743

XSL•FO
RenderX

Parents', Teachers', and Sledders' Acceptability of a Virtual Reality Game for Sledding Safety Education: Cross-Sectional Study (e63813)
Meggy Hayotte, Jonas Kreiner, Bernhard Hollaus. 2759

Co-Designed Online Training Program for Worry Management: The Role of Young People With Lived Experience of Worry in Program Development (e66461)
Jessica Steward, Michelle Moulds, Colette Hirsch. 2771

The Necessity of Regulating Drinking Scenes on Social Media Platforms Focusing on YouTube Sulbang Videos: Public Opinion From Surveys and YouTube Content Analysis (e65162)
HyoRim Ju, HyeWon Lee, Juyoung Choi, EunKyo Kang. 2784

Self-Stigma in Adults Living With Chronic Skin Disease: Development of the HautKompass Web-Based Program and Pilot Test of Its Usability, Acceptability, and Feasibility (e70290)
Juliane Traxler, Caroline Stuhlmann, Neuza da Silva Burger, Christian Stierle, Vahid Djamei, Anna Darzina, Marie Rudnik, Rachel Sommer. .
2 7 9 6

Evaluating the Acceptability and Utility of a Personalized Wellness App (Aspire2B) Using AI-Enabled Digital Biomarkers: Engagement Enhancement Pilot Study (e63471)
Calissa Leslie-Miller, Shellen Goltz, Pamela Barrios, Christopher Cushing, Teena Badshah, Corey Ungaro, Shankang Qu, Yulia Berezhnaya, Tristin Brisbois. 2811

Leveraging Technology to Engage Supplemental Nutrition Assistance Program Consumers With Children at Farmers Markets: Qualitative Community-Engaged Approach to App Development (e70104)
Callie Ogland-Hand, Jillian Schulte, Owusua Yamoah, Kathryn Poppe, Timothy Ciesielski, Regan Gee, Ana Zubieta, Darcy Freedman. 2821

Perception of Empathy in Mental Health Care Through Voice-Based Conversational Agent Prototypes: Experimental Study (e69329)
Ruvini Sanjeewa, Ravi Iyer, Pragalathan Apputhurai, Nilmini Wickramasinghe, Denny Meyer. 2837

Real-World Effectiveness of Glucose-Guided Eating Using the Data-Driven Fasting App Among Adults Interested in Weight and Glucose Management: Observational Study (e65368)
Michelle Jospe, Martin Kendall, Susan Schembre, Melyssa Roy. 2848

Boosting Digital Health Engagement Among Older Adults in Hong Kong: Pilot Pre-Post Study of the Generations Connect Project (e69611)
Aaron He, Runqi Yuan, Tzu Luk, Kelvin Wang, Sophia Chan. 2860

Using a Robot to Address the Well-Being, Social Isolation, and Loneliness of Care Home Residents via Video Calls: Qualitative Feasibility Study (e59764)
Lise Austbø, Ingelin Testad, Martha Gjestsen. 2872

Evaluating the Use of a Note-Taking App by Japanese Resident Physicians: Nationwide Cross-Sectional Study (e55087)
Taiju Miyagami, Yuji Nishizaki, Taro Shimizu, Yu Yamamoto, Kiyoshi Shikino, Koshi Kataoka, Masanori Nojima, Gautam Deshpande, Toshio Naito, Yasuharu Tokuda. 2882

Evaluating Feasibility and Acceptability of the "My HeartHELP" Mobile App for Promoting Heart-Healthy Lifestyle Behaviors: Mixed Methods Study (e66108)
Jina Choo, Songwhi Noh, Yura Shin. 2895

Characterizing Patient-Reported Fatigue Using Electronic Diaries in Neurodegenerative and Immune-Mediated Inflammatory Diseases: Observational Study (e65879)
Adrien Bennetot, Rana Zia Ur Rehman, Robbin Romijnders, Zhi Li, Victoria Macrae, Kristen Davies, Wan-Fai Ng, Walter Maetzler, Jennifer Kudelka, Hanna Hildesheim, Kirsten Emmert, Emma Paulides, C van der Woude, Ralf Reilmann, Svenja Aufenberg, Meenakshi Chatterjee, Nikolay Manyakov, Clémence Pinaud, Stefan Avey. 2911

XSL•FO
RenderX

School-Based Virtual Reality Programming for Obtaining Moderate-Intensity Exercise Among Children With Disabilities: Pre-Post Feasibility Study (e65801)
Byron Lai, Ashley Wright, Bailey Hutchinson, Larsen Bright, Raven Young, Drew Davis, Sultan Malik, James Rimmer, Pelham High Community Engagement Group. 2928

Discussions of Antibiotic Resistance on Social Media Platforms: Text Mining and Mixed Methods Content Analysis Study (e37160)
Jocelyne Arquembourg, Philippe Glaser, France Roblot, Isabelle Metzler, Mélanie Gallant-Dewavrin, Hugues Nanguem, Adel Mebarki, Paméla Voillot, Stéphane Schück. 2937

The Motivations of Citizens to Attend an eHealth Course in the Public Library: Qualitative Interview Study (e60612)
Lucille Standaar, Adriana Israel, Rosalie van der Vaart, Brigitta Keij, Frank van Lenthe, Roland Friele, Mariëlle Beenackers, Lilian van Tuyl. . . 2 9 4 7

Adolescent Emoji Use in Text-Based Messaging: Focus Group Study (e59640)
Matt Minich, Bradley Kerr, Megan Moreno. 2960

Impact of the Kidney Score Platform on Communication About and Patients' Engagement With Chronic Kidney Disease Health: Pre–Post Intervention Study (e56855)
Delphine Tuot, Susan Crowley, Lois Katz, Joseph Leung, Delly Alcantara-Cadillo, Christopher Ruser, Elizabeth Talbot-Montgomery, Joseph Vassalotti. 2970

Effect of a Digital Health Exercise Program on the Intention for Spinal Surgery in Adult Spinal Deformity: Exploratory Cross-Sectional Survey (e66889)
Marsalis Brown, Christopher Lin, Christopher Jin, Matthew Rohde, Brett Rocos, Jonathan Belding, Barrett Woods, Stacey Ackerman. 2979

Assessing the Utilization of Electronic Consultations in Genetics: Seven-Year Retrospective Study (e63028)
Sawona Biswas, Joyce So, Robert Wallerstein, Ralph Gonzales, Delphine Tout, Lisa DeAngelis, Aleksandar Rajkovic. 2987

Predicting Transvaginal Surgical Mesh Exposure Outcomes Using an Integrated Dataset of Blood Cytokine Levels and Medical Record Data: Machine Learning Approach (e59631)
Mihyun Waugh, Tyler Mills, Nicholas Boltin, Lauren Wolf, Patti Parker, Ronnie Horner, Thomas Wheeler II, Richard Goodwin, Melissa Moss. . . 2 9 9 4

Integration of a Mental Health App (e-MICHI) Into a Blended Treatment of Depression in Adolescents: Single-Group, Naturalistic Feasibility Trial (e58427)
Lena Lincke, Tim Martin-Döring, Andrea Daunke, Antonia Sadkowiak, Daria Nolkemper, Nina Sproeber-Kolb, Stefanie Bienioschek, Olaf Reis, Michael Kölch. 3005

Identification of Major Bleeding Events in Postoperative Patients With Malignant Tumors in Chinese Electronic Medical Records: Algorithm Development and Validation (e66189)
Hui Li, Haiyang Yao, Yuxiang Gao, Hang Luo, Changbin Cai, Zhou Zhou, Muhan Yuan, Wei Jiang. 3021

Mobile Health App for Adolescent Asthma Self-Management: Development and Usability Study of the Pulmonary Education and Knowledge Mobile Asthma Action Plan (e64212)
Xing He, Jiang Bian, Ariel Berlinski, Yi Guo, A Simmons, S Marshall, Carolyn Greene, Rita Brown, Jessica Turner, Tamara Perry. 3031

Tailoring Rheumatoid Arthritis Visit Timing Based on mHealth App Data: Mixed Methods Assessment of Implementation and Usability (e60854)
Robert Rudin, Leah Santacroce, Ishani Ganguli, Daniel Solomon. 3045

Developing an Online Community Advisory Board (CAB) of Parents From Social Media to Co-Design an Human Papillomavirus Vaccine Intervention: Participatory Research Study (e65986)
Regan Murray, Shawn Chiang, Ann Klassen, Jennifer Manganello, Amy Leader, Wen-Juo Lo, Philip Massey. 3058

XSL•FO
RenderX

Machine Learning Approach to Identifying Empathy Using the Vocals of Mental Health Helpline Counselors: Algorithm Development and Validation (e67835)
Ruvini Sanjeewa, Ravi Iyer, Pragalathan Apputhurai, Nilmini Wickramasinghe, Denny Meyer. 3072

Acoustic and Natural Language Markers for Bipolar Disorder: A Pilot, mHealth Cross-Sectional Study (e65555)
Cristina Crocamo, Riccardo Cioni, Aurelia Canestro, Christian Nasti, Dario Palpella, Susanna Piacenti, Alessandra Bartoccetti, Martina Re, Valentina Simonetti, Chiara Barattieri di San Pietro, Maria Bulgheroni, Francesco Bartoli, Giuseppe Carrà. 3087

Using the Healthy Native Youth Implementation Toolbox to Provide Web-Based Adolescent Health Promotion Decision Support to American Indian and Alaska Native Communities: Implementation Study (e67885)
Amrita Sidhu, Ross Shegog, Stephanie Craig-Rushing, Nicole Trevino, Michelle Singer, Cornelia Jessen, Gwenda Gorman, Sean Simpson, Melissa Peskin, Belinda Hernandez, Christine Markham. 3106

Evaluation of a Curriculum-Based Nutrition Education Intervention Protocol in Elementary Schools: Nonrandomized Feasibility Study (e69242)
Jacqueline Brown, Nicholas Rita, Beatriz Franco-Arellano, Ann LeSage, Joanne Arcand. 3118

Evaluating the Acceptability of a Brief Web-Based Alcohol Misuse Prevention Program Among US Military Cadets: Mixed Methods Formative Evaluation (e67637)
Emily Schmied, Lauren Hurtado, W Robinson, Cynthia Simon-Arndt, Richard Moyer III, Leslie Wilson, Mark Reed, Shannon Blakey, Marni Kan. 3133

Impact of a Virtual Reality Video ("A Walk-Through Dementia") on YouTube Users: Topic Modeling Analysis (e67755)
Xiaoli Li, Xiaoyu Liu, Cheng Yin, Sandra Collins, Eman Alanazi. 3144

Feasibility, Acceptability, and Effectiveness of a Smartphone App to Increase Pretransplant Vaccine Rates: Usability Study (e68855)
Amy Feldman, Brenda Beaty, Susan Moore, Sheana Bull, Kumanan Wilson, Katherine Atkinson, Cameron Bell, Kathryn Denize, Allison Kempe. 3151 3158

Development of an eHealth Mindfulness-Based Music Therapy Intervention for Adults Undergoing Allogeneic Hematopoietic Stem Cell Transplantation: Qualitative Study (e65188)
Sara Fleszar-Pavlovic, Blanca Noriega Esquives, Padideh Lovan, Arianna Brito, Ann Sia, Mary Kauffman, Maria Lopes, Patricia Moreno, Tulay Koru-Sengul, Rui Gong, Trent Wang, Eric Wieder, Maria Rueda-Lara, Michael Antoni, Krishna Komanduri, Teresa Lesiuk, Frank Penedo. 3161 3167

Evaluating User Experience and Satisfaction in a Concussion Rehabilitation App: Usability Study (e67275)
Michael Hutchison, Alex Di Battista, Kyla Pyndiura. 3180

Oxidative Stress Markers and Prediction of Severity With a Machine Learning Approach in Hospitalized Patients With COVID-19 and Severe Lung Disease: Observational, Retrospective, Single-Center Feasibility Study (e66509)
Olivier Raspado, Michel Brack, Olivier Brack, Mélanie Vivancos, Aurélie Esparcieux, Emmanuelle Cart-Tanneur, Abdellah Aouifi. 3190

Association of Social Media Recruitment and Depression Among Racially and Ethnically Diverse Metabolic and Bariatric Surgery Candidates: Prospective Cohort Study (e58916)
Jackson Francis, Sitapriya Neti, Dhatri Polavarapu, Folefac Atem, Luyu Xie, Olivia Kapera, Matthew Mathew, Elisa Marroquin, Carrie McAdams, Jeffrey Schellinger, Sophia Ngenge, Sachin Kukreja, Benjamin Schneider, Jaime Almandoz, Sarah Messiah. 3201

Encouraging the Voluntary Mobilization of Mental Resources by Manipulating Task Design: Explorative Study (e63491)
Lina-Estelle Louis, Saïd Moussaoui, Sébastien Ravoux, Isabelle Milleville-Pennel. 3211

XSL•FO
RenderX

Support of Home-Based Structured Walking Training and Prediction of the 6-Minute Walk Test Distance in Patients With Peripheral Arterial Disease Based on Telehealth Data: Prospective Cohort Study (e65721)
Fabian Wiesmüller, Andreas Prenner, Andreas Ziegl, Gihan El-Moazen, Robert Modre-Osprian, Martin Baumgartner, Marianne Brodmann, Gerald Seinost, Günther Silbernagel, Günter Schreier, Dieter Hayn. 3223

Web-Based Explainable Machine Learning-Based Drug Surveillance for Predicting Sunitinib- and Sorafenib-Associated Thyroid Dysfunction: Model Development and Validation Study (e67767)
Fan-Ying Chan, Yi-En Ku, Wen-Nung Lie, Hsiang-Yin Chen. 3236

Assessment of Gait Parameters Using Wearable Sensors and Their Association With Muscle Mass, Strength, and Physical Performance in Korean Older Adults: Cross-Sectional Study (e63928)
Jinyoung Shin, Hyuk Kweon, Jaekyung Choi. 3249

Centering Youth Voice in the Adaptation of an mHealth Intervention for Young Adults With HIV in South Texas, United States: Human-Centered Design Approach (e60531)
Nhat Ho, Catherine Johnson, Autumn Chidester, Ruby Viera Corral, Jacundo Ramos, Miguel Garcia, Rishi Gonuguntla, Cyrena Cote, Divya Chandramohan, Hueylie Lin, Anna Taranova, Ank Nijhawan, Susan Kools, Karen Ingersoll, Rebecca Dillingham, Barbara Taylor. 3258

Chatbot to Support the Mental Health Needs of Pregnant and Postpartum Women (Moment for Parents): Design and Pilot Study (e72469)
Kelsey McAlister, Lara Baez, Jennifer Huberty, Marianna Kerppola. 3270

Anticipated Acceptability of Blended Learning Among Lay Health Care Workers in Malawi: Qualitative Analysis Guided by the Technology Acceptance Model (e62741)
Tiwonge Mbeya-Munkhondya, Caroline Meek, Mtisunge Mphande, Tapiwa Tembo, Mike Chitani, Milenka Jean-Baptiste, Caroline Kumbuyo, Dhrutika Vansia, Katherine Simon, Sarah Rutstein, Victor Mwapasa, Vivian Go, Maria Kim, Nora Rosenberg. 3287

The Impact of a Human Papillomavirus Facebook-Based Intervention (#HPVVaxTalks) Among Young Black (African American and Sub-Saharan African Immigrants) Adults: Pilot Pre- and Poststudy (e69609)
Adebola Adegboyega, Amanda Wiggins, Abubakari Wuni, Melinda Ickes. 3298

Medical Device Based on a Virtual Reality–Based Upper Limb Rehabilitation Software: Usability Evaluation Through Cognitive Walkthrough (e68149)
Seojin Hong, Hyun Choi, Hyosun Kweon. 3311

Development of a Comprehensive Decision Support Tool for Chemotherapy-Cycle Prescribing: Initial Usability Study (e62749)
Sanna Iivanainen, Reetta Arokoski, Santeri Mentu, Laura Lang, Jussi Ekström, Henri Virtanen, Vesa Kataja, Jussi Koivunen. 3322

Barriers and Determinants of Referral Adherence in AI-Enabled Diabetic Retinopathy Screening for Older Adults in Northern India During the COVID-19 Pandemic: Mixed Methods Pilot Study (e67047)
Anshul Chauhan, Anju Goyal, Ritika Masih, Gagandeep Kaur, Lakshay Kumar, Neha, Harsh Rastogi, Sonam Kumar, Bidhi Singh, Preeti Syal, Vishali Gupta, Luke Vale, Mona Duggal. 3331

Personalized mHealth Intervention (StepAdd) for Increasing Physical Activity in Japanese Patients With Type 2 Diabetes: Secondary Analysis of Social Cognitive Theory Measurements of a Single-Arm Pilot Study (e60221)
Kayo Waki, Syunpei Enomoto, Toshimasa Yamauchi, Masaomi Nangaku, Kazuhiko Ohe. 3346

Accuracy of Smartphone-Mediated Snore Detection in a Simulated Real-World Setting: Algorithm Development and Validation (e67861)
Jeffrey Brown, Zachary Mitchell, Yu Jiang, Ryan Archdeacon. 3354

XSL•FO
RenderX

Mobile Health Tool to Capture Social Determinants of Health and Their Impact on HIV Treatment Outcomes Among People Who Use Drugs: Pilot Feasibility Study (e59953)
Rachel Gicquelais, Caitlin Conway, Olivia Vjorn, Andrew Genz, Gregory Kirk, Ryan Westergaard. 3363

Mobile App-Based Interactive Care Plan for Migraine: Survey Study of Usability and Improvement Opportunities (e66763)
Nathan Young, Jennifer Stern, Stephanie Steel, Jon Ebbert. 3374

Changes in Health Education Literacy After Structured Web-Based Education Versus Self-Directed Online Information Seeking in Patients Undergoing Carpal Tunnel Release Surgery: Nonrandomized, Controlled Study (e65114)
Mariella Seel, Julian Mihalic, Stefan Froschauer, Bernhard Holzner, Jens Meier, Tobias Gotterbarm, Matthias Holzbauer. 3385

The Impact of ChatGPT Exposure on User Interactions With a Motivational Interviewing Chatbot: Quasi-Experimental Study (e56973)
Jiading Zhu, Alec Dong, Cindy Wang, Scott Veldhuizen, Mohamed Abdelwahab, Andrew Brown, Peter Selby, Jonathan Rose. 3403

A Brief Cognitive Behavioral Therapy–Based Digital Intervention for Reducing Hazardous Alcohol Use in South Korea: Development and Prospective Pilot Study (e64459)
Manjae Kwon, Daa Moon, Minjae Kang, Young-Chul Jung. 3414

Exploring the Use of an Augmented Reality Device Learning Tool for Multidisciplinary Staff Training on Domestic Abuse and Sexual Violence: Postintervention Qualitative Evaluation (e60075)
Dilroshini Karunaratne, Jessica Whittock, Amber Moore, Krishna Dasigan, Jasmine Chevolleau, Brent Bartholomew, Nikki Kelly, Charlotte Cohen. 3431

Optimizing Testimonials for Behavior Change in a Digital Intervention for Binge Eating: Human-Centered Design Study (e59691)
Isabel Rooper, Adrian Ortega, Thomas Massion, Tanvi Lakhtakia, Macarena Kruger, Leah Parsons, Lindsay Lipman, Chidiebere Azubuike, Emily Tack, Katrina Obleada, Andrea Graham. 3446

The Interactive Care Coordination and Navigation mHealth Intervention for People Experiencing Homelessness: Cost Analysis, Exploratory Financial Cost-Benefit Analysis, and Budget Impact Analysis (e64973)
Hannah McCullough, Leticia Moczygemba, Anton Avanceña, James Baffoe. 3458

Digital Health Literacy and Attitudes Toward eHealth Technologies Among Patients With Cardiovascular Disease and Their Implications for Secondary Prevention: Survey Study (e63057)
Greta Ullrich, Alexander Bäuerle, Hannah Vogt, Amir Mahabadi, Katrin Paldán, Daniel Messiha, Lisa Jahre, Christos Rammos, Tienush Rassaf, Julia Lortz. 3470

Interprofessional Discussion for Knowledge Transfer in a Digital “Community of Practice” for Managing Pneumoconiosis: Mixed Methods Study (e67999)
Varinn Sood, Heidi Rishel Brakey, Orrin Myers, Xin Shore, Akshay Sood. 3482

Telemedicine Prescribing by US Mental Health Care Providers: National Cross-Sectional Survey (e63251)
Mollie Cummins, Julia Ivanova, Hiral Soni, Zoe Robbins, Brian Bunnell, Esteban López, Brandon Welch. 3495

Supplementing Consent for a Prospective Longitudinal Cohort Study of Infants With Antenatal Opioid Exposure: Development and Assessment of a Digital Tool (e59954)
Jamie Newman, Leslie Clarke, Pranav Athimuthu, Megan Dhawan, Sharon Owen, Traci Beiersdorfer, Lindsay Parlberg, Ananta Bangdiwala, Taya McMillan, Sara DeMauro, Scott Lorch, Myriam Peralta-Carcelen, Deanne Wilson-Costello, Namasivayam Ambalavanan, Stephanie Merhar, Brenda Poindexter, Catherine Limperopoulos, Jonathan Davis, Michele Walsh, Carla Bann. 3508

A Digital Pornography Education Prototype Co-Designed With Young People: Formative Evaluation (e65859)
Jake Turvey, Michelle Raggatt, Cassandra Wright, Angela Davis, Meredith Temple-Smith, Megan Lim. 3519

XSL•FO
RenderX

Purposes of Smartphone Health Care Apps and the Practicality of Their Functions in Disaster Situations: Qualitative Function Assessment Study (e56862)
Sayuri Nonaka, Susumu Fujii, Kosuke Yamada. 3530

Assessment of the Sensitivity of a Smartphone App to Assist Patients in the Identification of Stroke and Myocardial Infarction: Cross-Sectional Study (e60465)
Amar Dhand, Rama Mangipudi, Anubodh Varshney, Jonathan Crowe, Andria Ford, Nancy Sweitzer, Min Shin, Samuel Tate, Haissam Haddad, Michael Kelly, James Muller, Jay Shavadia. 3544

Text Messaging Interventions for Unhealthy Alcohol Use in Emergency Departments: Mixed Methods Assessment of Implementation Barriers and Facilitators (e65187)
Megan O'Grady, Laura Harrison, Adekemi Suleiman, Morica Hutchison, Nancy Kwon, Frederick Muench, Sandeep Kapoor. 3561

Patients' Experience With Evaluation by Both a Musculoskeletal Physician and Physical Therapist in the Same Digital Visit: Survey Study (e66744)
Mary O'Connor, Carolyn Chudy, Kaitlyn Peters, Megan Ribaudo, Carrie McCulloch, Jared Aguilar, Trista Taylor, Ryan Grant. 3575

App- and Wearable-Based Remote Monitoring for Patients With Myasthenia Gravis and Its Specialists: Feasibility and Usability Study (e58266)
Maike Stein, Regina Stegherr, Pushpa Narayanaswami, David Legg, Meret Herdick, Andreas Meisel, Lea Gerischer, Sophie Lehnerer. 3585

Enhancing Distress Tolerance Skills in Adolescents With Anorexia Nervosa Through the BALANCE Mobile App: Feasibility and Acceptability Study (e70278)
Christina Miranda, Brittany Matheson, Nandini Datta, Aileen Whyte, Hyun-Joon Yang, Paul Schmiedmayer, Vishnu Ravi, Oliver Aalami, James Lock. 3601

Probing Public Perceptions of Antidepressants on Social Media: Mixed Methods Study (e62680)
Jianfeng Zhu, Xinyu Zhang, Ruoming Jin, Hailong Jiang, Deric Kenne. 3608

Novel Evaluation Metric and Quantified Performance of ChatGPT-4 Patient Management Simulations for Early Clinical Education: Experimental Study (e66478)
Riley Scherr, Aidin Spina, Allen Dao, Saman Andalib, Faris Halaseh, Sarah Blair, Warren Wiechmann, Ronald Rivera. 3620

Capturing Everyday Parental Feeding Practices and Eating Behaviors of 3- to 5-Year-Old Children With Avid Eating Behavior: Ecological Momentary Assessment Feasibility and Acceptability Study (e66807)
Abigail Pickard, Katie Edwards, Claire Farrow, Emma Haycraft, Jacqueline Blissett. 3631

Mobilizing Health Promotion Through Canada's Student Mental Health Network: Concurrent, Mixed Methods Process Evaluation (e58992)
Amy Ecclestone, Brooke Linden, Jessica Rose, Kiran Kullar. 3643

Evaluation of Financial Support Workshops for Patients Under State Pension Age With Degenerative Cervical Myelopathy: Survey Study (e59032)
Tanzil Rujeedawa, Zahabiya Karimi, Helen Wood, Irina Sangeorzan, Roy Smith, Iwan Sadler, Esther Martin-Moore, Adrian Gardner, Andreas K Demetriades, Rohitashwa Sinha, Gordan Grahovac, Antony Bateman, Naomi Deakin, Benjamin Davies. 3654

Snowball Group Usability Testing for Rapid and Iterative Multisite Tool Development: Method Development Study (e55316)
Katherine Dauber-Decker, David Feldstein, Rachel Hess, Devin Mann, Eun Kim, Pranisha Gautam-Goyal, Jeffrey Solomon, Sundas Khan, Fatima Malik, Lynn Xu, Ainsley Huffman, Paul Smith, Wendy Halm, Alice Yuroff, Safiya Richardson. 3661

Exploring the Effect of an 8-Week AI-Composed Exercise Program on Pain Intensity and Well-Being in Patients With Spinal Pain: Retrospective Cohort Analysis (e57826)
Annika Griefahn, Florian Avermann, Christoff Zalpour, Robert Marshall, Inés Cordon Morillas, Kerstin Luedtke. 3670

XSL•FO
RenderX

Involving End Users in the Development and Usability Testing of a Smartphone App Designed for Individuals With Prediabetes: Mixed-Methods Focus Group Study (e59386)
Natalie Grieve, Kyra Braaten, Megan MacPherson, Sam Liu, Mary Jung. 3683

Activity Trackers in Physical Therapy for People With Chronic Obstructive Pulmonary Disease in the Netherlands: Cross-Sectional Study on Current Use and Implementation Determinants (e59533)
Darcy Ummels, Esther Bols, Roel Frantzen, Tim Frantzen, Levy Robeerts, Emmylou Beekman. 3693

Understanding Providers' Attitude Toward AI in India's Informal Health Care Sector: Survey Study (e54156)
Sumeet Kumar, Snehil Rayal, Raghuram Bommaraju, Navya Varasala, Sirisha Papineni, Sarang Deo. 3707

Developing an ICD-10 Coding Assistant: Pilot Study Using RoBERTa and GPT-4 for Term Extraction and Description-Based Code Selection (e60095)
Sander Puts, Catharina Zegers, Andre Dekker, Iñigo Bermejo. 3718

Web-Based, Algorithm-Guided Insulin Titration in Insulin-Treated Type 2 Diabetes: Pre-Post Intervention Study (e68914)
Nishanth Thiagarajan, Hong Tan, Suresh Rama Chandran, Phong Lee, Yun Chin, Wanling Zeng, Emily Ho, David Carmody, Su-Yen Goh, Yong Bee. 3730

Assessment of Digital Capabilities by 9 Countries in the Alliance for Healthy Cities Using AI: Cross-Sectional Analysis (e62935)
Hocheol Lee. 3740

A Web-Based Resource Informed by Cognitive Behavioral Therapy and Positive Psychology to Address Stress, Negative Affect, and Problematic Alcohol Use: A Usability and Descriptive Study (e63819)
Ingrid Serck-Hanssen, Marit Solheim-Witt, Justin Anker, Dawn Sugarman. 3753

Step Count Accuracy of the Life Plus Connected Watch at Different Localizations and Speeds in Healthy Adults, Patients With Cardiovascular Disease, and Patients With Peripheral Artery Disease: Step Count Validation Study in Laboratory Settings (e58964)
Anne-Noëlle Heizmann, Edouard Ollier, Pierre Labeix, Ivan Goujon, Frédéric Roche, Claire Le Hello. 3765

Mediating Effect of Social Capital on the Association Between Digital Literacy and Life Satisfaction Among Older Adults in South Korea: Cross-Sectional Study (e68163)
Hyein Jung, Hocheol Lee, Eun Nam. 3774

Addressing the "Black Hole" of Low Back Pain Care With Clinical Decision Support: User-Centered Design and Initial Usability Study (e66666)
Robert Rudin, Patricia Herman, Robert Vining. 3786

Summer Research Internship Curriculum to Promote Self-Efficacy, Researcher Identity, and Peer-to-Peer Learning: Retrospective Cohort Study (e54167)
Yulia Levites Strekalova, Rachel Liu-Galvin, Samuel Border, Sara Midence, Mishal Khan, Maya VanZanten, John Tomaszewski, Sanjay Jain, Pinaki Sarder. 3799

Web-Based Exercise and Nutrition Intervention to Improve Leg Muscle Strength and Physical Functioning in Older Adults: Pre-Post Pilot Study (e54392)
Berber Dorhout, Nick Wezenbeek, Lisette de Groot, Pol Grootswagers. 3810

Identifying Digital Markers of Attention-Deficit/Hyperactivity Disorder (ADHD) in a Remote Monitoring Setting: Prospective Observational Study (e54531)
Heet Sankesara, Hayley Denyer, Shaoxiong Sun, Qigang Deng, Yatharth Ranjan, Pauline Conde, Zulqarnain Rashid, Philip Asherson, Andrea Bilbow, Madeleine Groom, Chris Hollis, Richard Dobson, Amos Folarin, Jonna Kuntsi. 3822

XSL•FO
RenderX

The Development of an Instagram Reel-Based Bystander Intervention Message Among College Students: Formative Survey and Mixed Methods Pilot Study (e66769)
Leticia Couto. 3836

Evaluating Older Adults' Engagement and Usability With AI-Driven Interventions: Randomized Pilot Study (e64763)
Marcia Shade, Changmin Yan, Valerie Jones, Julie Boron. 3854

Characterizing Telehealth Barriers and Preferences to Promote Acceptable Implementation Strategies in Central Uganda: Multilevel Formative Evaluation (e60843)
Michael Kizito, Erina Mugabi, Sabrina Ford, Bree Holtz, Kelly Hirko. 3866

Testing an Electronic Patient-Reported Outcome Platform in the Context of Traumatic Brain Injury: PRiORiTy Usability Study (e58128)
Christel McMullan, Grace Turner, Ameeta Retzer, Antonio Belli, Elin Davies, Laura Nice, Luke Flavell, Jackie Flavell, Melanie Calvert. 3879

The Association of Psychological Factors With Willingness to Share Health-Related Data From Technological Devices: Cross-Sectional Questionnaire Study (e64244)
Marijn Eversdijk, Emma Douma, Mirela Habibovic, Willem Kop. 3892

AI Machine Learning–Based Diabetes Prediction in Older Adults in South Korea: Cross-Sectional Analysis (e57874)
Hocheol Lee, Myung-Bae Park, Young-Joo Won. 3902

Insights From Diverse Perspectives on Social Media Messages to Inform Young Adults With Cancer About Clinical Trials: Focus Group Study (e64265)
Melissa Beauchemin, Desiree Walker, Allison Rosen, Maria Frazer, Meital Eisenberger, Rhea Khurana, Edward Bentlyewski, Victoria Fedorko, Corey Basch, Grace Hillyer. 3911

Sentiments of Individuals with Interstitial Cystitis/Bladder Pain Syndrome Toward Pentosan Polysulfate Sodium: Infodemiology Study (e54209)
Yulin Hswen, Qiuyuan Qin, Pressley Smith, Alison Swierczynski, Stuart Bauer, Erika Ladson, Amanda Garrett, Catherine Brownstein. 3920

Using Health Information Resources for People With Cognitive Impairment (digiDEM Bayern): Registry-Based Cohort Study (e54460)
Florian Weidinger, Nikolas Dietzel, Elmar Graessel, Hans-Ulrich Prokosch, Peter Kolominsky-Rabas. 3930

Mobile Phone App to Promote Lifestyle Change in People at Risk of Type 2 Diabetes: Feasibility 3-Arm Randomized Controlled Trial (e63737)
Gyri Skoglund, Gunvor Hilde, Pernille Lunde, Venessa Cruz Naceno, Cecilie Fromholt Olsen, Birgitta Blakstad Nilsson. 3940

Considering Comorbidities and Individual Differences in Testing a Gaming Behavioral Activation App for Perinatal Depression and Anxiety: Open Trial Pilot Intervention Study (e59154)
Gabriella Hamlett, Chloe Schrader, Craig Ferguson, Lauren Kobylski, Rosalind Picard, Joseph Locascio, Richard McNally, Lee Cohen, Rachel Vanderkruik. 3952

Verification of the Reliability of an Automated Urine Test Strip Colorimetric Program Using Colorimetric Analysis: Survey Study (e62772)
Keigo Inagaki, Daisuke Tsuriya, Takuya Hashimoto, Katsumasa Nakamura. 3963

Determinants of Dropping Out of Remote Patient-Reported Outcome–Based Follow-Up Among Patients With Epilepsy: Prospective Cohort Study (e58258)
Sofie Vestergaard, Mette Roost, David Christiansen, Liv Schougaard. 3973

XSL•FO
RenderX

Impact of a Virtual Care Navigation Service on Member-Reported Outcomes Among Lesbian, Gay, Bisexual, Transgender, and Queer Populations: Case Study (e64137)
Seul Choi, Jaclyn Marshall, Patrina Sexton Topper, Andrew Pregnall, José Bauermeister. 3987

A Mobile App for Promoting Breastfeeding-Friendly Communities in Hong Kong: Design and Development Study (e64191)
Heidi Fan, Emily Leung, Ka Lau, Janet Wong, Edmond Choi, Christine Lam, Marie Tarrant, Hextan Ngan, Patrick Ip, Chia Lin, Kris Lok. 3999

Predicting Age and Visual-Motor Integration Using Origami Photographs: Deep Learning Study (e58421)
Chien-Yu Huang, Yen-Ting Yu, Kuan-Lin Chen, Jenn-Jier Lien, Gong-Hong Lin, Ching-Lin Hsieh. 4014

Comparative Evaluation of Consumer Wearable Devices for Atrial Fibrillation Detection: Validation Study (e65139)
Femke Wouters, Henri Gruwez, Christophe Smeets, Anessa Pijalovic, Wouter Wilms, Julie Vranken, Zoë Pieters, Hugo Van Herendael, Dieter Nuyens, Maximo Rivero-Ayerza, Pieter Vandervoort, Peter Haemers, Laurent Pison. 4022

Real-Time Analytics and AI for Managing No-Show Appointments in Primary Health Care in the United Arab Emirates: Before-and-After Study (e64936)
Yousif AlSerkal, Naseem Ibrahim, Aisha Alsereidi, Mubaraka Ibrahim, Sudheer Kurakula, Sadaf Naqvi, Yasir Khan, Neema Oottumadathil. . . 4 0 3 5

Effects of Smart Goggles Used at Bedtime on Objectively Measured Sleep and Self-Reported Anxiety, Stress, and Relaxation: Pre-Post Pilot Study (e58461)
Sharon Danoff-Burg, Elie Gottlieb, Morgan Weaver, Kiara Carmon, Duvia Lara Ledesma, Holly Rus. 4043

The Use of Social Media on Enhancing Dental Care and Practice Among Dental Professionals: Cross-Sectional Survey Study (e66121)
Joseph Acosta, Palinee Detsomboonrat, Pagaporn Pisarnturakit, Nipaporn Urwannachotima. 4051

Artificial Intelligence–Powered Training Database for Clinical Thinking: App Development Study (e58426)
Heng Wang, Danni Zheng, Mengying Wang, Hong Ji, Jiangli Han, Yan Wang, Ning Shen, Jie Qiao. 4061

The Physical Activity at Work (PAW) Program in Thai Office Workers: Mixed Methods Process Evaluation Study (e57604)
Katika Akksilp, Thomas Rouyard, Wanrudee Isaranuwatchai, Ryota Nakamura, Falk Müller-Riemenschneider, Yot Teerawattananon, Cynthia Chen. 4072

ChatGPT's Attitude, Knowledge, and Clinical Application in Geriatrics Practice and Education: Exploratory Observational Study (e63494)
Huai Cheng. 4089

A Digital Parenting Intervention With Intimate Partner Violence Prevention Content: Quantitative Pre-Post Pilot Study (e58611)
Moa Schafer, Jamie Lachman, Paula Zinser, Francisco Calderón Alfaro, Qing Han, Chiara Facciola, Lily Clements, Frances Gardner, Genevieve Haupt Ronnie, Ross Sheil. 4104

Differences in eHealth Access, Use, and Perceived Benefit Between Different Socioeconomic Groups in the Dutch Context: Secondary Cross-Sectional Study (e49585)
Lucille Standaar, Lilian van Tuyl, Anita Suijkerbuijk, Anne Brabers, Roland Friele. 4135

The Trifecta of Industry, Academic, and Health System Partnership to Improve Mental Health Care Through Smartphone-Based Remote Patient Monitoring: Development and Usability Study (e57624)
C Epperson, Rachel Davis, Allison Dempsey, Heinrich Haller, David Kupfer, Tiffany Love, Pamela Villarreal, Mark Matthews, Susan Moore, Kimberly Muller, Christopher Schneck, Jessica Scott, Richard Zane, Ellen Frank. 4149

XSL•FO
RenderX

Co-Design of an Escape Room for e-Mental Health Training of Mental Health Care Professionals: Research Through Design Study (e58650)
Joyce Bierbooms, Wouter Sluis-Thiescheffer, Milou Feijt, Inge Bongers. 4161

Measuring Bound Attention During Complex Liver Surgery Planning: Feasibility Study (e62740)
Tim Schneider, Timur Cetin, Stefan Uppenkamp, Dirk Weyhe, Thomas Muender, Anke Reinschluessel, Daniela Salzmann, Verena Uslar. . . . 4 1 7 8

Development of Personas and Journey Maps for Artificial Intelligence Agents Supporting the Use of Health Big Data: Human-Centered Design Approach (e67272)
Yoon Lee, Hanna Choi, Soo-Kyoung Lee. 4195

Effect of a Narrative-Based Online Course Aimed at Reducing Stigma Toward Transgender Children and Adolescents: Longitudinal Observational Study (e59605)
Merlin Greuel, Van Nguyen, Doron Amsalem, Maya Adam, Till Bärnighausen. 4213

User Outcomes for an App-Delivered Hypnosis Intervention for Menopausal Hot Flashes: Retrospective Analysis (e63948)
Katherine Scheffrahn, Claire Hall, Vanessa Muñiz, Gary Elkins. 4227

Development and Beta Validation of an mHealth-Based Hearing Screener (SRESHT) for Young Children in Resource-Limited Countries: Pilot Validation Study (e53460)
Vidya Ramkumar, Deepashree Joshi B, Anil Prabhakar, James Hall, Ramya Vaidyanath. 4241

Healthy Kai (Food) Checker Web-Based Tool to Support Healthy Food Policy Implementation: Development and Usability Study (e60447)
Magda Rosin, Cliona Ni Mhurchu, Elaine Umali, Sally Mackay. 4257

Exploring the General Acceptability and User Experience of a Digital Therapeutic for Cognitive Training in a Singaporean Older Adult Population: Qualitative Study (e63568)
Siong Kwek, Qiao Leong, V Lee, Ni Lau, Smrithi Vijayakumar, Wei Ng, Bina Rai, Marlena Raczkowska, Christopher Asplund, Alexandria Remus, Dean Ho. 4272

A Proposed mHealth Intervention to Address Patient Barriers to Colposcopy Attendance: Qualitative Interview Study of Clinic Staff and Patient Perspectives (e55043)
Jennifer Hemler, Rachel Wagner, Brittany Sullivan, Myneka Macenat, Erin Tagai, Jazmarie Vega, Enrique Hernandez, Suzanne Miller, Kuang-Yi Wen, Charletta Ayers, Mark Einstein, Shawna Hudson, Racquel Kohler. 4284

Acceptance, Safety, and Effect Sizes in Online Dialectical Behavior Therapy for Borderline Personality Disorder: Interventional Pilot Study (e66181)
Ruben Vonderlin, Tali Boritz, Carola Claus, Büsra Senyüz, Saskia Mahalingam, Rachel Tennenhouse, Stefanie Lis, Christian Schmahl, Jürgen Margraf, Tobias Teismann, Nikolaus Kleindienst, Shelley McMain, Martin Bohus. 4298

Cardiomeds, an mHealth App for Self-Management to Support Swiss Patients With Heart Failure: 2-Stage Mixed Methods Usability Study (e63941)
Lisa Simioni, Elena Tessitore, Hamdi Hagberg, Aurélie Schneider-Paccot, Katherine Blondon, Liliane Gschwind, Philippe Meyer, Frederic Ehrler. 4315

Teaching in the Digital Age—Developing a Support Program for Nursing Education Providers: Design-Based Research (e66109)
Stefan Walzer, Carolin Barthel, Ronja Pazouki, Helga Marx, Sven Ziegler, Peter Koenig, Christiane Kugler, Stefan Jobst. 4344

XSL•FO
RenderX

Sexual and Gender Minority Adolescents' Preferences for HIV Pre-Exposure Prophylaxis Social Marketing Campaigns: Qualitative Preimplementation Study (e60398)
Kathryn Macapagal, Juan Zapata, Junye Ma, Jacob Gordon, Christopher Owens, Silvia Valadez-Tapia, Peter Cummings, Nathan Walter, Jim Pickett. 4357

A Mobile Health Intervention to Support Collaborative Decision-Making in Mental Health Care: Development and Usability (e57614)
Kristin Romm, Mari Skoge, Elizabeth Barrett, Lars-Christian Berentzen, Dagfinn Bergsager, Pål Fugelli, Thomas Bjella, Erlend Gardsjord, Kristine Kling, Sindre Kruse, Kari Kværner, Ingrid Melle, Erlend Mork, Henrik Ihler, Eline Rognli, Carmen Simonsen, Tor Værnes, Sofie Aminoff. 4377

A Mindfulness-Based App Intervention for Pregnant Women: Qualitative Evaluation of a Prototype Using Multiple Case Studies (e58265)
Silvia Rizzi, Maria Pavesi, Alessia Moser, Francesca Paolazzi, Michele Marchesoni, Stefania Poggianella, Erik Gadotti, Stefano Forti. 4392

Reflections of Foster Youth Engaging in the Co-Design of Digital Mental Health Technology: Duoethnography Study (e53231)
Ifunanya Ezimora, Tylia Lundberg, Dylan Miars, Jeruel Trujeque, Ashley Papias, Margareth Del Cid, Johanna Folk, Marina Tolou-Shams. . . .
4 4 1 0

Developing a Machine Learning–Based Automated Patient Engagement Estimator for Telehealth: Algorithm Development and Validation Study (e46390)
Pooja Guhan, Naman Awasthi, Kathryn McDonald, Kristin Bussell, Gloria Reeves, Dinesh Manocha, Aniket Bera. 4425

Psychological and Behavioral Insights From Social Media Users: Natural Language Processing–Based Quantitative Study on Mental Well-Being (e60286)
Xingwei Yang, Guang Li. 4443

Attitudes of German General Practitioners Toward eHealth Apps for Dementia Risk Reduction: Qualitative Interview Study (e56310)
Adrian Schultz, Melanie Luppa, Markus Bleckwenn, Steffi Riedel-Heller, Andrea Zuelke. 4465

Discrimination of Radiologists' Experience Level Using Eye-Tracking Technology and Machine Learning: Case Study (e53928)
Stanford Martinez, Carolina Ramirez-Tamayo, Syed Akhter Faruqui, Kal Clark, Adel Alaeddini, Nicholas Czarnek, Aarushi Aggarwal, Sahra Emamzadeh, Jeffrey Mock, Edward Golob. 4479

Use of Go-Beyond as a Self-Directed Internet-Based Program Supporting Veterans' Transition to Civilian Life: Preliminary Usability Study (e60868)
Karolina Alichniewicz, Sarah Hampton, Madeline Romaniuk, Darcy Bennett, Camila Guindalini. 4495

Usability and Usefulness of a Symptom Management Coaching System for Patients With Cancer Treated With Immune Checkpoint Inhibitors: Comparative Mixed Methods Study (e57659)
Savannah Glaser, Itske Fraterman, Noah van Brummelen, Valentina Tibollo, Laura Del Campo, Henk Mallo, Sofie Wilgenhof, Szymon Wilk, Vitali Gisko, Vadzim Khadakou, Ronald Cornet, Manuel Ottaviano, Stephanie Medlock. 4509

Improving Mental Health and Well-Being Through the Paradym App: Quantitative Study of Real-World Data (e68031)
Athina Metaxa, Shaun Liverpool, Mia Eisenstadt, John Pollard, Courtney Carlsson. 4526

Digital Mindfulness Training for Burnout Reduction in Physicians: Clinician-Driven Approach (e63197)
Lia Antico, Judson Brewer. 4539

Factors Influencing Telemedicine Adoption Among Health Care Professionals: Qualitative Interview Study (e54777)
Fiona Schürmann, Daniel Westmattelmann, Gerhard Schewe. 4567

XSL•FO
RenderX

A Digital Tool (Technology-Assisted Problem Management Plus) for Lay Health Workers to Address Common Mental Health Disorders: Co-production and Usability Study in Pakistan (e59414)
Maham Saleem, Shamsa Zafar, Thomas Klein, Markus Koesters, Adnan Bashir, Daniela Fuhr, Siham Sikander, Hajo Zeeb. 4586

Technological Adjuncts to Streamline Patient Recruitment, Informed Consent, and Data Management Processes in Clinical Research: Observational Study (e58628)
Jodie Koh, Stacey Caron, Amber Watters, Mahesh Vaidyanathan, David Melnick, Alyssa Santi, Kenneth Hudson, Catherine Arguelles, Priyanka Mathur, Mozziyar Etemadi. 4604

Understanding Individual Differences in Happiness Sources and Implications for Health Technology Design: Exploratory Analysis of an Open Dataset (e65658)
Edel Ennis, Raymond Bond, Maurice Mulvenna, Colm Sweeney. 4612

Experiences of Peer Mentoring Sexual and Gender Minority Emerging Adults Who Are at Risk for Suicide: Mixed Methods Study (e67814)
Jennifer Tran, Jessica Webster, James Wolfe, Jennifer Ben Nathan, Lindiwe Mayinja, Marin Kautz, Maria Oquendo, Gregory Brown, David Mandell, Danielle Mowery, José Bauermeister, Lily Brown. 4624

Self-Guided Smartphone App (Vimbo) for the Reduction of Symptoms of Depression and Anxiety in South African Adults: Pilot Quantitative Single-Arm Study (e54216)
Sherrie Steyn, Meggan Slabbert. 4635

Co-Designing a Digital Brief Intervention to Reduce the Risk of Prescription Opioid–Related Harm Among People With Chronic Noncancer Pain: Qualitative Analysis of Patient Lived Experiences (e57208)
Rachel Elphinston, Sue Pager, Kelly Brown, Michele Sterling, Farhad Fatehi, Paul Gray, Linda Hipper, Lauren Cahill, Jason Connor. 4648

Human-Centered Design of an mHealth Tool for Optimizing HIV Index Testing in Wartime Ukraine: Formative Research Case Study (e66132)
Nancy Puttkammer, Elizabeth Dunbar, Myroslava Germanovych, Mariia Rosol, Matthew Golden, Anna Hubashova, Vladyslav Fedorchenko, Larisa Hetman, Liudmyla Legkostup, Jan Flowers, Olena Nesterova. 4667

Evaluating the Impact of Pharmacotherapy in Augmenting Quit Rates Among Hispanic Adults in an App-Delivered Smoking Cessation Intervention: Secondary Analysis of a Randomized Controlled Trial (e69311)
Margarita Santiago-Torres, Kristin Mull, Brianna Sullivan, Ana Cupertino, Ramzi Salloum, Matthew Triplette, Michael Zvolensky, Jonathan Bricker. 4686

Effectiveness of Cognitive Behavioral Therapy Provided Through a Web Application for Subthreshold Depression, Subthreshold Insomnia, and Subthreshold Panic: Open-Labeled 6-Arm Randomized Clinical Trial Pilot Study (e63139)
Kayoko Taguchi, Mirai Miyoshi, Yoichi Seki, Shiori Baba, Eiji Shimizu. 4699

Effectiveness of Electronic Quality Improvement Activities to Reduce Cardiovascular Disease Risk in People With Chronic Kidney Disease in General Practice: Cluster Randomized Trial With Active Control (e54147)
Jo-Anne Manski-Nankervis, Barbara Hunter, Natalie Lumsden, Adrian Laughlin, Rita McMorrow, Douglas Boyle, Patty Chondros, Shilpanjali Jesudason, Jan Radford, Megan Prictor, Jon Emery, Paul Amores, An Tran-Duy, Craig Nelson. 4715

Nurses' Perspectives and Experiences of Using a Bed-Exit Information System in an Acute Hospital Setting: Mixed Methods Study (e64444)
Stefan Walzer, Isabel Schön, Johanna Pfeil, Sam Klemm, Sven Ziegler, Claudia Schmoor, Christophe Kunze. 4735

Status of Digital Health Technology Adoption in 5 Vietnamese Hospitals: Cross-Sectional Assessment (e53483)
Duc Tran, Nguyen Thanh Dung, Chau Minh Duc, Huynh Ngoc Hon, Le Minh Khoi, Nguyen Phuc Hau, Duong Thi Thu Huyen, Huynh Thi Le Thu, Tran Van Duc, VITAL (Vietnam ICU Translational Applications Laboratory) Investigators, Lam Minh Yen, C Thwaites, Chris Paton. 4748

XSL•FO
RenderX

Adapting a Mobile Health App for Smoking Cessation in Black Adults With Anxiety Through an Analysis of the Mobile Anxiety Sensitivity Program Proof-of-Concept Trial: Qualitative Study (e53566)
Marshall Cheney, Adam Alexander, Lorra Garey, Matthew Gallagher, Emily Hébert, Anka Vujanovic, Krista Kezbers, Cameron Matoska, Michael Zvolensky, Michael Businelle. 4764

Improving the Utility, Safety, and Ethical Use of a Passive Mood-Tracking App for People With Bipolar Disorder Using Coproduction: Qualitative Focus Group Study (e65140)
Laurence Astill Wright, Matthew Moore, Stuart Reeves, Elvira Vallejos, Richard Morriss. 4781

Social Media Use and Oral Health–Related Misconceptions in Saudi Arabia: Cross-Sectional Study (e70071)
Rahaf BinHamdan, Salwa Alsadhan, Arwa Gazzaz, AlBandary AlJameel. 4794

mHealth App to Promote Healthy Lifestyles for Diverse Families Living in Rural Areas: Usability Study (e60495)
Alejandra Perez Ramirez, Adrian Ortega, Natalie Stephenson, Angel Muñoz Osorio, Anne Kazak, Thao-Ly Phan. 4810

Social Acceptance of Smart Glasses in Health Care: Model Evaluation Study of Anticipated Adoption and Social Interaction (e49610)
Niek Zuidhof, Oscar Peters, Peter-Paul Verbeek, Somaya Ben Allouch. 4825

Designing Patient-Centered Interventions for Emergency Care: Participatory Design Study (e63610)
Woosuk Seo, Shruti Jain, Vivian Le, Jiaqi Li, Zhan Zhang, Hardeep Singh, Kalyan Pasupathy, Prashant Mahajan, Sun Park. 4842

The Effect of the Virtual Reality–Based Biofeedback Intervention DEEP on Stress, Emotional Tension, and Anger in Forensic Psychiatric Inpatients: Mixed Methods Single-Case Experimental Design (e65206)
Lisa Klein Haneveld, Tessa Dekkers, Yvonne Bouman, Hanneke Scholten, Joanneke Weerdmeester, Saskia Kelders, Hanneke Kip. 4858

Machine Learning Approach to Identifying Wrong-Site Surgeries Using Centers for Medicare and Medicaid Services Dataset: Development and Validation Study (e68436)
Yuan-Hsin Chen, Ching-Hsuan Lin, Chiao-Hsin Fan, An Long, Jeremiah Scholl, Yen-Pin Kao, Usman Iqbal, Yu-Chuan Li. 4875

Exploring the Acceptance of Just-in-Time Adaptive Lifestyle Support for People With Type 2 Diabetes: Qualitative Acceptability Study (e65026)
Eclaire Hietbrink, Anouk Middelweerd, Wendy d'Hollosy, Laura Schrijver, Gozewijn Laverman, Miriam Vollenbroek-Hutten. 4885

Nomophobia, Psychopathology, and Smartphone-Inferred Behaviors in Youth With Depression: Longitudinal Study (e57512)
Tianyi Zhang, Andres Camargo, Lianne Schmaal, Vassilis Kostakos, Simon D'Alfonso. 4903

Adapting the Germ Defence Web-Based Intervention to Improve Infection Prevention and Control in Care Homes: Interview Study Among Care Home Staff (e66706)
Alex Hall, Johanna Aguilera-Muñoz, Lisa McGarrigle, Charlotte Eost-Telling, James Denison-Day, Christie Cabral, Merlin Willcox, Chris Todd. 4909

Designing Values Elicitation Technologies for Mental Health and Chronic Care Integration: User-Centered Design Approach (e68419)
Isabel Rooper, William Liem, Martha Burla, Jacob Gordon, Lara Baez, Rachel Kornfield, Andrew Berry. 4936

A Classroom-Based Intervention for Reducing Sedentary Behavior and Improving Spinal Health: Pragmatic Stepped-Wedge Feasibility Randomized Controlled Trial (e65169)
Dominic Fisher, Rentia Maart, Lehana Thabane, Quinette Louw. 4952

Formative Evaluation of Suicide Prevention Websites for Men: Qualitative Study with Men at Risk of Suicide and with Potential Gatekeepers (e59829)
Doreen Reifegerste, Anna Wagner, Lisa Huber, Manuel Fastuca. 4965

XSL•FO
RenderX

Automated Digital Safety Planning Interventions for Young Adults: Qualitative Study Using Online Co-design Methods (e69602)
Jonah Meyerhoff, Sarah Popowski, Tanvi Lakhtakia, Emily Tack, Rachel Kornfield, Kaylee Kruzan, Charles Krause, Theresa Nguyen, Kevin Rushton, Anthony Pisani, Madhu Reddy, Kimberly Van Orden, David Mohr. 4979

Evaluation of the Clear Fear Smartphone App for Young People Experiencing Anxiety: Uncontrolled Pre– and Post–Follow-Up Study (e55603)
Chiara Samele, Norman Urquia, Rachel Edwards, Katie Donnell, Nihara Krause. 4994

A Digital Program for Daily Life Management With Endometriosis: Pilot Cohort Study on Symptoms and Quality of Life Among Participants (e58262)
Zélia Breton, Emilie Stern, Mathilde Pinault, Delphine Lhuillery, Erick Petit, Pierre Panel, Maïa Alexaline. 5011

Daily Treatment Monitoring for Patients Receiving Home-Based Peritoneal Dialysis and Prediction of Heart Failure Risk: mHealth Tool Development and Modeling Study (e56254)
Jia Wu, Youjia Zeng, Jun Yang, Yutong Yao, Xiuling Xu, Gaofeng Song, Wuyong Yi, Taifen Wang, Yihou Zheng, Zhongwei Jia, Xiangyu Yan. . . 5 0 3 0

Experiences With an In-Bed Real-Time Motion Monitoring System on a Geriatric Ward: Mixed Methods Study (e63572)
Stefan Walzer, Isabel Schön, Johanna Pfeil, Nicola Merz, Helga Marx, Sven Ziegler, Christophe Kunze. 5049

Patient Adoption of Digital Use Cases in Family Medicine and a Nuanced Implementation Approach for Family Doctors: Quantitative Web-Based Survey Study (e58867)
Julian Beerbaum, Sibylle Robens, Leonard Fehring, Achim Mortsiefer, Sven Meister. 5062

Feasibility and Links Between Emotions, Physical States, and Eating Behavior in Patients After Metabolic Bariatric Surgery: Experience Sampling Study (e60486)
Ellen Kuipers, Josien Timmerman, Marc van Det, Miriam Vollenbroek-Hutten. 5079

FRAILSURVEY—an mHealth App for Self-Assessment of Frailty Based on the Portuguese Version of the Groningen Frailty Indicator: Validation and Reliability Study (e51975)
Luis Midao, Mafalda Duarte, Rute Sampaio, Marta Almada, Cláudia Dias, Constança Paúl, Elísio Costa. 5095

Evaluation of the MyFertiCoach Lifestyle App for Subfertile Couples: Single-Center Evaluation of Augmented Standard Care (e64239)
Jesper Smeenk, Ellen Smit, Marc Jacobs, Ilse van Rooij. 5111

Development of a Voice-Activated Virtual Assistant to Improve Insomnia Among Young Adult Cancer Survivors: Mixed Methods Feasibility and Acceptability Study (e64869)
Hunter Groninger, Hannah Arem, Lylian Ayangma, Lisa Gong, Eric Zhou, Daniel Greenberg. 5129

Zoom-Delivered Empowered Relief for Chronic Pain: Observational Longitudinal Pilot Study Exploring Feasibility and Pain-Related Outcomes in Patients on Long-Term Opioids (e68292)
Karlyn Edwards, Troy Dildine, Dokyoung You, Ashley Herrick, Beth Darnall, Sean Mackey, Maisa Ziadni. 5142

Digital Assessment of Cognitive Health in Outpatient Primary Care: Usability Study (e66695)
Adam Doerr, Taylor Orwig, Matthew McNulty, Stephanie Sison, David Paquette, Robert Leung, Huitong Ding, Stephen Erban, Bruce Weinstein, Yurima Guilarte-Walker, Adrian Zai, Allan Walkey, Apurv Soni, David McManus, Honghuang Lin. 5161

Teen Perspectives on Integrating Digital Mental Health Programs for Teens Into Public Libraries (“I Was Always at the Library”): Qualitative Interview Study (e67454)
Ashley Knapp, Katherine Cohen, Kaylee Kruzan, Rachel Kornfield, Miguel Herrera, Aderonke Pederson, Sydney Lee, Kathryn Macapagal, Chantelle Roulston, Kaleigh Clarke, Clarisa Wijaya, Robert Simmons, Latonia Jackson, Simrandeep Kour, Sandra Franco, David Mohr. 5170

XSL•FO
RenderX

Use of Video Consultation Between 2017 and 2020 in Outpatient Medical Care in Germany and Characteristics of Their User Groups: Analysis of Claims Data (e60170)
Theresa Hüer, Anke Walendzik, Lara Kleinschmidt, Klemens Höfer, Beatrice Nauendorf, Juliane Malsch, Matthias Brittner, Paul Brandenburg, André Aeustergerling, Udo Schneider, Anja Wadeck, Sebastian Liersch, Stephanie Sehlen, Katharina Schwarze, Jürgen Wasem. 5186

Development of a Clinical Decision Support Tool to Implement Asthma Management Guidelines in Pediatric Primary Care: Qualitative Study (e65794)
David Fedele, Jessica Ray, Jaya Mallela, Jiang Bian, Aokun Chen, Xiao Qin, Ramzi Salloum, Maria Kelly, Matthew Gurka, Jessica Hollenbach.
5 1 9 9

Designing a Digital Intervention to Increase Human Milk Feeding Among Black Mothers: Qualitative Study of Acceptability and Preferences (e67284)
Loral Patchen, Jeannette Tsuei, Donna Sherard, Patricia Moriarty, Zoe Mungai-Barris, Tony Ma, Elina Bajracharya, Katie Chang, William Evans. 5211

Exploring Laypersons' Experiences With a Mobile Symptom Checker App as an Interface Between eHealth Literacy, Health Literacy, and Health-Related Behavior: Qualitative Interview Study (e60647)
Roland Koch, Marie-Theres Steffen, Anna-Jasmin Wetzel, Christine Preiser, Malte Klemmt, Hans-Jörg Ehni, Regina Mueller, Stefanie Joos. .
5 2 2 4

Exploring Physicians' Dual Perspectives on the Transition From Free Text to Structured and Standardized Documentation Practices: Interview and Participant Observational Study (e63902)
Olga Golburean, Rune Pedersen, Line Melby, Arild Faxvaag. 5242

Co-Design of a Depression Self-Management Tool for Adolescent and Young Adult Cancer Survivors: User-Centered Design Approach (e67175)
Karly Murphy, Rachel Glock, David Victorson, Madhu Reddy, Sarah Birken, John Salsman. 5258

Safe Listening Beliefs, Attitudes, and Practices Among Gamers and Esports Participants: International Web-Based Survey (e60476)
Nicola Diviani, Shelly Chadha, Peter Mulas, Sara Rubinelli. 5275

Perceived Trust and Professional Identity Threat in AI-Based Clinical Decision Support Systems: Scenario-Based Experimental Study on AI Process Design Features (e64266)
Sophia Ackerhans, Kai Wehkamp, Rainer Petzina, Daniel Dumitrescu, Carsten Schultz. 5290

Stigma and Behavior Change Techniques in Substance Use Recovery: Qualitative Study of Social Media Narratives (e57468)
Annie Chen, Lexie Wang, Shana Johnny, Sharon Wong, Rahul Chaliparambil, Mike Conway, Joseph Glass. 5307

User Experience With a Personalized mHealth Service for Physical Activity Promotion in University Students: Mixed Methods Study (e64384)
Silke Wittmar, Tom Frankenstein, Vincent Timm, Peter Frei, Nicolas Kurpiers, Stefan Wölwer, Axel Schäfer. 5323

Convergent and Known-Groups Validity and Sensitivity to Change of the Virtual Performance Measure in Patients With Hip and Knee Osteoarthritis: Longitudinal Study (e69001)
Helen Razmjou, Suzanne Denis, Susan Robarts, Amy Wainwright, Patricia Dickson, John Murnaghan. 5340

Ambulance Commanders' Reluctance to Enter Road Tunnels in Simulated Incidents and the Effects of a Tunnel-Specific e-Learning Course on Decision-Making: Web-Based Randomized Controlled Trial (e58542)
Johan Hylander, Lina Gyllencreutz, Michael Haney, Anton Westman. 5347

Videoconference-Delivered Acceptance and Commitment Therapy for Family Caregivers of People With Dementia: Pilot Randomized Controlled Trial (e67545)
Areum Han, Robert Oster, Hon Yuen, Jeremy Jenkins, Jessica Hawkins, Lauren Edwards. 5364

XSL•FO
RenderX

Personalized Physician-Assisted Sleep Advice for Shift Workers: Algorithm Development and Validation Study (e65000)
Yufei Shen, Alicia Choto Olivier, Han Yu, Asami Ito-Masui, Ryota Sakamoto, Motomu Shimaoka, Akane Sano. 5380

Examining the Client Experience of Digital Tools in Blended Care Therapy: Qualitative Interview Study (e68249)
Emily Lattie, Miranda Beltzer, Alethea Varra, Connie Chen, Anita Lungu. 5402

A Remote Intervention Based on mHealth and Community Health Workers for Antiretroviral Therapy Adherence in People With HIV: Pilot Randomized Controlled Trial (e67997)
Shivesh Shourya, Jianfang Liu, Sophia McInerney, Trinity Casimir, James Kenniff, Trace Kershaw, David Batey, Rebecca Schnall. 5413

Assessing the Cultural Fit of a Digital Sleep Intervention for Refugees in Germany: Qualitative Study (e65412)
Maja Blomenkamp, Andrea Kiesel, Harald Baumeister, Dirk Lehr, Josef Unterrainer, Lasse Sander, Kerstin Spanhel. 5436

Digital Health Platform for Maternal Health: Design, Recruitment Strategies, and Lessons Learned From the PowerMom Observational Cohort Study (e70149)
Toluwalase Ajayi, Jacqueline Kueper, Lauren Ariniello, Diana Ho, Felipe Delgado, Matthew Beal, Jill Waalen, Katie Baca Motes, Edward Ramos. 5451

The Use of Mobile Apps in Adolescent Psychotherapy: Assessment of Psychotherapists' Perspectives (e65788)
Sarah Wüllner, Katharin Hermenau, Tobias Hecker, Michael Siniatchkin. 5467

Developing an Internet-Based Cognitive Behavioral Therapy Intervention for Adolescents With Anxiety Disorders: Design, Usability, and Initial Evaluation of the CoolMinds Intervention (e66966)
Nikita Sørensen, Helene Skaarnes, Kim Mathiasen, Mikael Thastum, Johanne Lomholt. 5483

Opportunities and Challenges Surrounding the Use of Wearable Sensor Bracelets for Infectious Disease Detection During Hajj: Qualitative Interview Study (e60484)
Noha Maddah, Arpana Verma, John Ainsworth. 5499

A New Mobile App to Train Attention Processes in People With Traumatic Brain Injury: Logical and Ecological Content Validation Study (e64174)
Roxanne Laverdière, Philip Jackson, Frédéric Banville. 5518

Cost-Effectiveness Analysis of a Machine Learning–Based eHealth System to Predict and Reduce Emergency Department Visits and Unscheduled Hospitalizations of Older People Living at Home: Retrospective Study (e63700)
Charlotte Havreng-Théry, Arnaud Fouchard, Fabrice Denis, Jacques-Henri Veyron, Joël Belmin. 5531

The EmpkinS-EKSpression Reappraisal Training Augmented With Kinesthesia in Depression: One-Armed Feasibility Study (e65357)
Marie Keinert, Lena Schindler-Gmelch, Lydia Rupp, Misha Sadeghi, Robert Richer, Klara Capito, Bjoern Eskofier, Matthias Berking. 5540

Evaluating User Engagement With a Real-Time, Text-Based Digital Mental Health Support App: Cross-Sectional, Retrospective Study (e66301)
Edward Coffield, Khadeja Kausar. 5556

Acceptability, Usability, and Insights Into Cybersickness Levels of a Novel Virtual Reality Environment for the Evaluation of Depressive Symptoms: Exploratory Observational Study (e68132)
Sara Sutori, Emma Eliasson, Francesca Mura, Victor Ortiz, Vincenzo Catrambonephd, Gergö Hadlaczky, Ivo Todorov, Antonio Alfeo, Valentina Cardi, Mario Cimino, Giovanna Mioni, Mariano Raya, Gaetano Valenza, Vladimir Carli, Claudio Gentili. 5565

XSL•FO
RenderX

A Digital Photo Activity Intervention for Nursing Home Residents With Dementia and Their Carers: Mixed Methods Process Evaluation (e56586)
Josephine Tan, David Neal, Maria Vilmen, Petra Boersma, Teake Ettema, Robbert Gobbens, Sietske Sikkes, Rose-Marie Dröes. 5586

Creating a Usable and Effective Digital Intervention to Support Men to Test for HIV and Link to Care in A Resource-Constrained Setting: Iterative Design Based on A Person-Based Approach and Human Computer Interaction Methods (e65185)
Anya Zeitlin, Thulile Mathenjwa, Thembelihle Zuma, Sally Wyke, Philippa Matthews, Nuala McGrath, Janet Seeley, Maryam Shahmanesh, Ann Blandford. 5609

Development of a GPT-4–Powered Virtual Simulated Patient and Communication Training Platform for Medical Students to Practice Discussing Abnormal Mammogram Results With Patients: Multiphase Study (e65670)
Dan Weisman, Alanna Sugarman, Yue Huang, Lillian Gelberg, Patricia Ganz, Warren Comulada. 5624

User Experience of and Adherence to a Smartphone App to Maintain Behavior Change and Self-Management in Patients With Work-Related Skin Diseases: Multistep, Single-Arm Feasibility Study (e66791)
Nele Ristow, Annika Wilke, Christoph Skudlik, Swen John, Michaela Ludewig. 5641

Designing Health Care Provider–Centered Emergency Department Interventions: Participatory Design Study (e68891)
Woosuk Seo, Jiaqi Li, Zhan Zhang, Chuxuan Zheng, Hardeep Singh, Kalyan Pasupathy, Prashant Mahajan, Sun Park. 5660

Compatibility of the CEN-ISO/TS 82304-2 Health App Assessment Framework With Catalan and Italian Health Authorities' Needs: Qualitative Interview Study (e67855)
Petra Hoogendoorn, Mariam Shokralla, Romy Willemsen, Nick Guldemond, María Villalobos-Quesada. 5676

A Mobile Health App Informed by the Multi-Process Action Control Framework to Promote Physical Activity Among Inactive Adults: Iterative Usability Study (e59477)
Heather Hollman, Wuyou Sui, Haowei Zhang, Ryan Rhodes. 5694

Applying Patient and Health Professional Preferences in Co-Designing a Digital Brief Intervention to Reduce the Risk of Prescription Opioid–Related Harm Among Patients With Chronic Noncancer Pain: Qualitative Analysis (e57212)
Rachel Elphinston, Sue Pager, Farhad Fatehi, Michele Sterling, Kelly Brown, Paul Gray, Linda Hipper, Lauren Cahill, Maisa Ziadni, Peter Worthy, Jason Connor. 5713

Acceptability and Usability of a Digital Behavioral Health Platform for Youth at Risk of Suicide: User-Centered Design Study With Patients, Practitioners, and Business Gatekeepers (e65418)
Trinity Tse, Lauren Weiner, Carter Funkhouser, Danielle DeLuise, Colleen Cullen, Zachary Blumkin, Casey O'Brien, Randy Auerbach, Nicholas Allen. 5732

Exploring Therapists' Approaches to Treating Eating Disorders to Inform User-Centric App Design: Web-Based Interview Study (e68846)
Pamela Thomas, Pippa Bark, Sarah Rowe. 5747

Cocreating the Visualization of Digital Mobility Outcomes: Delphi-Type Process With Patients (e68782)
Jack Lumsdon, Cameron Wilson, Lisa Alcock, Clemens Becker, Francesco Benvenuti, Tecla Bonci, Koen van den Brande, Gavin Brittain, Philip Brown, Ellen Buckley, Marco Caruso, Brian Caulfield, Andrea Cereatti, Laura Delgado-Ortiz, Silvia Del Din, Jordi Evers, Judith Garcia-Aymerich, Heiko Gaßner, Tova Gur Arieh, Clint Hansen, Jeffrey Hausdorff, Hugo Hiden, Emily Hume, Cameron Kirk, Walter Maetzler, Dimitrios Megaritis, Lynn Rochester, Kirsty Scott, Basil Sharrack, Norman Sutton, Beatrix Vereijken, Ioannis Vogiatzis, Alison Yarnall, Alison Keogh, Alma Cantu. . 5 7 6 3

XSL•FO
RenderX

Developing the MyCancerGene Digital Health Portal to Improve Patients' Understanding of Germline Cancer Genetic Test Results: Development, User, and Usability Testing Study (e56282)
Phillip Trieu, Dominique Fetzer, Briana McLeod, Kathryn Schweickert, Lauren Gutstein, Brian Egleston, Susan Domchek, Linda Fleisher, Lynne Wagner, Kuang-Yi Wen, Cara Cacioppo, Jessica Ebrahimzadeh, Dana Falcone, Claire Langer, Elisabeth Wood, Kelsey Karpink, Shelby Posen, Enida Selmani, Angela Bradbury. 5786

Evaluating the Accuracy and Reliability of Real-World Digital Mobility Outcomes in Older Adults After Hip Fracture: Cross-Sectional Observational Study (e67792)
Martin Berge, Anisoara Paraschiv-Ionescu, Cameron Kirk, Arne Küderle, Encarna Micó-Amigo, Clemens Becker, Andrea Cereatti, Silvia Del Din, Monika Engdal, Judith Garcia-Aymerich, Karoline Grønvik, Clint Hansen, Jeffrey Hausdorff, Jorunn Helbostad, Carl-Philipp Jansen, Lars Johnsen, Jochen Klenk, Sarah Koch, Walter Maetzler, Dimitrios Megaritis, Arne Müller, Lynn Rochester, Lars Schwickert, Kristin Taraldsen, Beatrix Vereijken. 5809

Public Health Messaging About Dengue on Facebook in Singapore During the COVID-19 Pandemic: Content Analysis (e66954)
Shirley Ho, Mengxue Ou, Nova Huang, Agnes Chuah, Vanessa Ho, Sonny Rosenthal, Hye Kim. 5828

Needs and Preferences of Swedish Young Adults for a Digital App Promoting Mental Health Literacy, Occupational Balance, and Peer Support: Qualitative Interview Study (e71563)
Martin Karaba Bäckström, Sonya Girdler, Ben Milbourn, Annika Lexén. 5839

Unlocking the Potential of Wear Time of a Wearable Device to Enhance Postpartum Depression Screening and Detection: Cross-Sectional Study (e67585)
Eric Hurwitz, Samantha Meltzer-Brody, Zachary Butzin-Dozier, Rena Patel, Noémie Elhadad, Melissa Haendel. 5855

A Data-Driven Approach to Assessing Hepatitis B Mother-to-Child Transmission Risk Prediction Model: Machine Learning Perspective (e69838)
Dung Nguyen Tien, Huong Thi Thu Bui, Tram Hoang Thi Ngoc, Thuy Thi Pham, Dac Trung Nguyen, Huyen Nguyen Thi Thu, Thi Thu Hang Vu, Thi Lan Anh Luong, Lan Thu Hoang, Ho Cam Tu, Nina Körber, Tanja Bauer, Lam Khanh Ho. 5873

Ambivalent User Needs as a Challenge and Chance for the Design of a Web-Based Intervention for Gaming Disorder: Qualitative Interview Study With Adolescents and Young Adults (e63258)
Birte Geisler, Kay Petersen, Sara Hanke, Simon Schurer, Anne Schreiber, Christine Lämmle, Anil Batra, Tobias Renner, Isabel Brandhorst. . .
5 8 9 5

Insights From the Nihon Housou Kyoukai's Virtual Reality–Based Social Interaction Television Program "Project Aliens" for Adolescents With Psychiatric Disorders: Single-Center Case Series Study (e74401)
Junichi Fujita, Mizuho Takayama, Emi Kamono, Satoru Shinoda, Hiroyuki Yamaguchi, Tomoko Moroga, Mio Ishii, Tomoyuki Miyazaki. 5913

A Self-Harm Awareness Training Module for School Staff: Co-Design and User Testing Study (e69309)
Anne-Marie Burn, Hayley Gains, Joanna Anderson. 5926

Leveraging Cognitive and Speech Ecological Momentary Assessment in Individuals With Phenylketonuria: Development and Usability Study of Cognitive Fluctuations in a Rare Disease Population (e63644)
Shifali Singh, Lisa Kluen, Katelin Curtis, Raquel Norel, Carla Agurto, Elizabeth Grinspoon, Zoe Hawks, Shawn Christ, Susan Waisbren, Guillermo Cecchi, Laura Germine. 5945

Caregivers' Perceptions of Clinical Symptoms, Disease Management, and Quality of Life Impact in Cases of Cyclin-Dependent Kinase-Like 5 Deficiency Disorder: Cross-Sectional Online Survey (e72489)
Sam Amin, Carol-Anne Partridge, Helen Leonard, Jenny Downs, Helen Allvin, Valentine Ficara, Emilie Pain, Minna Korolainen. 5963

Experiences of Individually Tailored Internet-Based Cognitive Behavioral Therapy During the COVID-19 Pandemic: Qualitative Interview Study (e66908)
Victoria Aminoff, Matilda Baltius, Emelie Lundström, Matilda Berg, Gerhard Andersson, Mikael Ludvigsson. 5982

XSL•FO
RenderX

Development of a Serious Game to Simulate Neonatal Intensive Care Unit Experiences: Collaborative Quasi-Experimental Study (e73009)
Yukihide Miyosawa, Koichi Hirabayashi, Kodai Yamada, Fumiya Kobayashi, Nanami Ogihara, Noa Takeda, Eri Okamura, Shogo Matsumura. . 5 9 9 8

Health Services Usage in Patients Receiving Buprenorphine for Opioid Use Disorder or Long-Term Opioid Therapy for Chronic Pain: Retrospective Cohort Study (e66596)
Samuel Savitz, Maria Stevens, Bidisha Nath, Gail D'Onofrio, Edward Melnick, Molly Jeffery. 6011

User Experiences With Digital Future-Self Interventions in the Contexts of Smoking and Physical Inactivity: Mixed Methods Multistudy Exploration (e63893)
Kristell Penfornis, Nele Albers, Willem-Paul Brinkman, Mark Neerincx, Andrea Evers, Winifred Gebhardt, Eline Meijer. 6028

Research Dissemination Strategies in Pediatric Emergency Care Using a Professional Twitter (X) Account: A Mixed Methods Developmental Study of a Logic Model Framework (e59481)
Gwendolyn Hooley, Julia Magana, Jason Woods, Shyam Sivasankar, Lauren VonHoltz, Anita Schmidt, Todd Chang, Michelle Lin. 6048

Video- Versus Text-Based Psychoeducation in Web-Based E-Mental Health Programs: Randomized Controlled Trial (e65478)
Swantje Borsutzky, Josefine Gehlenborg, Lara Rolvien, Steffen Moritz. 6065

Patient–Health Care Professional Communication via a Secure Web-Based Portal in Severe Mental Health Conditions: Qualitative Analysis of Secure Messages (e63713)
Eva Meier-Diedrich, Carolyn Turvey, Jonas Wördemann, Justin Speck, Mareike Weibezahl, Julian Schwarz. 6076

A Remotely Delivered GLP-1RA–Supported Specialist Weight Management Program in Adults Living With Obesity: Retrospective Service Evaluation (e72577)
Rebecca Richards, Michael Whitman, Gina Wren, Peta Campion. 6091

Challenges and Mitigation Strategies in the Development and Feasibility Assessment of a Digital Mental Health Intervention for Depression (VMood): Mixed Methods Feasibility Study (e68297)
Leena Chau, Jill Murphy, Vu Nguyen, Hai Tran, Harry Minas, Raymond Lam, Kanna Hayashi, Thi Nguyen, Emanuel Krebs, John O'Neil. 6106

Assessing the Fit of a Digitally Delivered National Diabetes Prevention Program Among Rural Living Adults: Qualitative Study (e70406)
Gerit Wagner, Lyndsie Koon, Patricia Smith, Kameron Suire, Mary Hastert, Joseph Donnelly, Melissa Olfert, Paul Estabrooks, Anna Gorczyca. 6 1 2 3

User-Driven Development of a Digital Behavioral Intervention for Chronic Pain: Multimethod Multiphase Study (e74064)
Afra Taygar, Sara Bartels, Rocío de la Vega, Ida Flink, Linnéa Engman, Suzanne Petersson, Sophie Johnsson, Katja Boersma, Lance McCracken, Rikard Wicksell. 6138

Online Tailored Decision Aid for Maternal Pertussis Vaccination in a Randomized Controlled Trial: Process Evaluation Study (e50709)
Charlotte Anraad, Pepijn van Empelen, Robert Ruiter, Hilde van Keulen. 6165

Multidomain Behavioral Change Digital Coaching for Chronic Disease Management in Patients With Type 2 Diabetes: Framework Development and Preliminary Evaluation (e73807)
Konstantina Kostopoulou, Danae Lekka, Aristodemos Pnevmatikakis, Nelina Angelova, Panagiotis Stafylas, Stefanos Tamouridis, Alexandra Bargiota, Sofoklis Kyriazakos. 6181

Developing and Integrating Digital Sources in an Accessible and Sustainable Online Platform for Adolescents and Young Adult Cancer Survivors: Collaborative Design Approach (e60897)
Carla Vlooswijk, Sophia Sleeman, Jonas Pluis, Daphne Bakker, Lisanne de Groot, Eveliene Manten, Peter Heine, Olga Husson, Mies van Eenbergen, Pieter Vandekerckhove. 6201

XSL•FO
RenderX

Exploring the Impact of Online Mental Health Resources During the COVID-19 Pandemic on Lesbian, Gay, Bisexual, Transgender, Queer, and Questioning Adults Compared to Heterosexual Adults: Pretest-Posttest Survey Analyses (e67082)
Natalia Ramos, Skylar Jones, Lily Zhang, Miriam Nuño, Benita Ramsey, Dannie Ceseña, Alyssa Mireles, Kenneth Wells. 6217

Hand Motion Control Ability Between Young and Older Adults: Comparative Study (e65224)
Jung-Soon Kim, Hui-Jun Kim, Minju Kim, Sung-Hee Kim. 6235

Health Care Professionals' Perspectives on Technology Use in Urinary Care: Cross-Sectional Survey-Based Study (e73453)
Nicole Zhang, Ni Zhang, Yun Wang, Rachel Wong, Kaylin Ma. 6246

Evaluating the Construct Validity and Sensitivity to Change of the Klenico Depression Domain in Psychotherapeutic Inpatient Care: Instrument Validation Study (e50504)
Stefan Reutimann, Jasmin Steiner, Noah Hübscher, Ulrich Voderholzer, Adrian Meule, Mareike Augsburger. 6259

Acceptability of an Intervention to Prevent Older Adult Mistreatment Among Family Caregivers to Persons With Dementia: Multimethod Pilot Study (e73778)
Kylie Meyer, Wenxing Wei, Jeanine Yonashiro-Cho, Susanna Mage, Sohee Kim, Elliane Irani, Christopher Burant, Zachary Gassoumis, Erin Lamb, Jaclene Zauszniewski, Donna Benton. 6273

Automating Colon Polyp Classification in Digital Pathology by Evaluation of a "Machine Learning as a Service" AI Model: Algorithm Development and Validation Study (e67457)
David Beyer, Evan Delancey, Logan McLeod. 6289

Mitigating the Negative Effects of Internet Browsing on Young People's Resilience and Outlook on Life Through Classic Grimms' Fairy Tales: Exploratory Randomized Controlled Study (e76770)
Congcong Hou, Thomas Foscht, Barbara Duffek, Annisa Arigayota, Andreas Eisingerich. 6296

User Experiences Among Patients and Health Care Professionals Who Participated in a Randomized Controlled Trial of E-nergEYEze, a Vision-Specific eHealth Intervention to Reduce Fatigue in Adults With Visual Impairment: Mixed Methods Study (e53080)
Manon Veldman, Hilde van der Aa, Hans Knoop, Christina Bode, Ger van Rens, Ruth van Nispen. 6310

What Matters Most to Veterans When Deciding to Use Technology for Health: Cross-Sectional Analysis of a National Survey (e77113)
Bella Etingen, Bridget Smith, Stephanie Shimada, Stephanie Robinson, Robin Higashi, Ndindam Ndiwane, Kathleen Frisbee, Jessica Lipschitz, Eric Richardson, Dawn Irvin, Timothy Hogan. 6324

Data Mining–Based Model for Computer-Aided Diagnosis of Autism and Gelotophobia: Mixed Methods Deep Learning Approach (e72115)
Mohamed Eldawansy, Hazem El Bakry, Samaa M Shohieb. 6336

Early Diagnosis of Knee Osteoarthritis With a Natural Language Processing–Driven Approach Based on Clinician Notes: Development and Validation Study (e64536)
Narathip Thanyakunsajja, Kulsawasd Jitkajornwanich, Shan Xu, Donghee Shin, Pattama Charoenporn. 6352

Enrollment and Retention Outcomes from the Veterans Health Administration for a Remote Digital Health Study: Multisite Observational Study (e68676)
Jaclyn Pagliaro, Lauren Wash, Ka Ly, Jenny Mathew, Alison Leibowitz, Ryan Cabrera, Jolie Wormwood, Varsha Vimalananda. 6373

Deep Learning for the Early Detection of Invasive Ductal Carcinoma in Histopathological Images: Convolutional Neural Network Approach With Transfer Learning (e62996)
Naga Chilumukuru, Pragya Priyadarshini, Yawo Ezunkpe. 6384

XSL•FO
RenderX

Characteristics Associated With the Use of the Mindfulness Meditation App Headspace in a Large Public Health Deployment: Cross-Sectional Survey Study (e73457)
Judith Borghouts, Elizabeth Eikey, Cinthia De Leon, Stephen Schueller, Margaret Schneider, Nicole Stadnick, Kai Zheng, Dana Mukamel, Dara Sorkin. 6395

Enhancing Engagement, Practice Integration, and Skill Learning in Mobile Technology–Delivered Interventions Using Human Support: Randomized Controlled Trial With Depressed College Students (e56963)
Colleen Conley, Brynn DeLorenzo, Carol Gonzales, Ian Kahrilas, Jennifer Duffecy, Rebecca Silton. 6412

Supporting Police Well-Being Through an Adaptive Shift Management System: Co-Design Study (e69986)
Olumuyiwa Ayorinde, Huseyin Dogan, Festus Adedoyin, Nan Jiang, Fiona Bitters, Sara Dempsey. 6433

Designing Digital Mental Health Interventions to Meet the Needs of Older Adolescents: Qualitative Interview and Group Discussion Study (e68950)
Rachel Kornfield, Sarah Popowski, Emily Tack, Jack Svoboda, Miguel Herrera, Theresa Nguyen, Ashley Knapp, David Mohr, Jonah Meyerhoff. 6 4 5 6

The Feasibility and Acceptability of an App-Based Intervention Aimed at Improving Maternal Health Literacy Regarding Infant Play and Development: Mixed Methods Study (e76517)
Fiona Bennin, Shane Norris, Alessandra Prioreschi. 6473

Quality and Reliability of Adolescent Sexuality Education on Chinese Video Platforms: Sentiment-Topic Analysis and Cross-Sectional Study (e77100)
Lan Wang, Xiantao Shu, Jianmei Huang, Weiqian Yan, Duo Zhao. 6489

Mental Health Care Provider Experiences of Remote Measurement-Based Care Rollout in an Urban Safety-Net Psychiatry Department: Three-Site Mixed Methods Hypothesis-Generating Implementation Study (e71570)
Ana Progovac, Carl Fulwiler, Margaret Lanca, Daisy Wang, Kate Zona, Norah Mulvaney-Day, Rajendra Aldis, H Leff, Philip Wang, Lisa Rosenfeld. 6511

Quantifying Spatial Shadow Zones and Their Association With Hospital Falls in Acute Care Unit: Real-Time Location System Observational Study (e75697)
Yen-Pin Chen, Yi-Chun Chen, Chen-Liang Lin, Chien-Yu Chi, Yi-Ying Chen, Bey-Jing Yang, Chien-Hua Huang. 6529

The Impact of a Digital Digestive Health Program and Telehealth Visits in Socially Vulnerable Populations: Cohort Evaluation (e70748)
Sanskriti Varma, Alicen Black, Erin Commons, Pen-Che Ho, Dena Bravata, Hau Liu. 6541

A Youth-Centered Digital Infographic on Vaping Risks (What's in a Vape?): Mixed Methods Study (e75694)
Kendra Nelson Ferguson, Kyla Christianson, Adrian Delgado, Gina Martin, Stephanie Coen, Laura Struik. 6549

Needs and Expectations Associated With an e–Mental Health Intervention for Reducing Somatoform, Anxiety, and Depressive Symptoms in Sexual and Gender Minority Adults: Qualitative Participative Study (e65834)
Anna Gomes, Angeliki Tsiouris, Florian Jung, Laura Klein, Adina Kreis, Manfred Beutel, Bernhard Strauss, Madita Hoy, Rüdiger Zwerenz. . . 6 5 6 1

Semaglutide and Tirzepatide in a Remote Weight Management Program: 12-Month Retrospective Observational Study (e81912)
Rebecca Richards, William Lunt, Michael Whitman, Giulia Spaltro, Rachel Hall. 6577

TruVox Web-Based Software for Vocal Pitch Training in Transgender Women: Development and Single-Session Evaluations (e73841)
Sam Weese, Mary Wilkens, Om Jadhav, Xiangyi Wang, Ansh Bhanushali, Tyler DiLoreto, Reyna Kozel, Renee Gustin, Tara McAllister, Victoria McKenna, Vesna Novak. 6590

XSL•FO
RenderX

Online Yoga Pilot Intervention for Black Women at High Cardiovascular Risk: Internet-Based Recruitment and Engagement (e41221)
Candace Johnson, Pascaline Ezouah. 6609

A Digital Psychosocial Service (Sui App) for Arabic-Speaking Refugees in Switzerland: Development and Cultural-Contextual Adaptation Using a User-Centered and Participatory Approach (e59905)
Rilana Stoeckli, Thomas Berger, Monia Aebersold, Viktoria Zoellner, Farhad Haji, Muriel Hunziker, Beatriz Jesus Ferreira, Michel Hosmann, Sebastian Burchert, Jessica Wabiszczewicz, Christine Knaevelsrud, Eva Heim. 6627

Patient-Reported Experiences With Long-Term Lifestyle Self-Monitoring in Heart Disease: Mixed Methods Study (e76978)
Mayra Goevaerts, Nicole Tenbült - Van Limpt, Willem Kop, Hareld Kemps, Yuan Lu. 6649

Stakeholder Perspectives on mHealth Technologies to Prevent Sitting-Acquired Pressure Injuries in Long-Term Care Facilities: Mixed Methods Study (e59590)
Tamara Vos-Draper, Erin Vinoski Thomas, Emily Graybill, Melissa Morrow, Kathleen Jordan, Pamela Manley, Sharon Sonenblum. 6670

Exploring the Impact of a Remote Monitoring System for Palliative and End-of-Life Care (CARE-PAC): Mixed Methods Feasibility Study (e69394)
Roma Maguire, Lisa McCann, Claire Singleton, Paul Perkins, Ollie Minton, Nicola McCann, Emma Longford, Alistair McKeown, Kimberley Kavanagh, Morven Miller. 6682

User-Reported Mechanisms of Change on a Suicide Prevention Website: Single-Arm Pragmatic Trial (e71136)
Martina Fruhbauerova, David Huh, Ursula Whiteside. 6697

eHealth Applications Improve Glycemic Control in Patients With Diabetes: Randomized Controlled Trial (e67761)
Junjie Huang, Claire Zhong, Siu Wong, Chung Lo, Man Yim, Martin Wong. 6713

Health Care Providers' Perspectives of Clinical Decision Support Tools for Pediatric Sepsis in Bangladesh: Qualitative Study (e73451)
Shamsun Shaima, Alicia Genisca, Md Faruk, Md Uddin, Akash Saha, Nadia Sultana, Nidhi Kadakia, Monique Gainey, Elleen Kim, Kikuyo Shaw, Farzana Afroze, Joan Chepngeno, Atin Jindal, Sifat Chowdhury, Md Chisti, Adam Levine, Stephanie Garbern. 6723

Evaluating a Website on Learning Disorders for Parents and Learning Therapists: Observational Mixed Methods Study (e68365)
Olga Hermansson, Paula Dümig, Björn Witzel, Lior Weinreich, Susanne Volkmer, Gerd Schulte-Körne, Kristina Moll. 6738

Differential Diagnosis Assessment in Ambulatory Care With a Digital Health History Device: Pseudorandomized Study (e56384)
Beth Healey, Adrien Schwitzguebel, Herve Spechbach. 6759

Understanding Experiences of and Unmet Needs in Online Searches for Menopause Information: An Exploratory Survey (e75335)
Erin Funnell, Freya McConnell, Nayra Martin-Key, Leyao Qian, Kathryn Babbitt, Sabine Bahn. 6769

Low Risk Perception of Harm From Substance Use and Sexual Behaviors Among Online Help–Seeking Sexual and Gender Minoritized People in San Francisco, California: Cross-Sectional Survey (e81753)
Jarett Maycott, Sean Arayasirikul. 6785

Clinical System for Mood Disorder Care in Córdoba, Colombia: Participatory Design and Scenario-Based Usability Evaluation Study (e58909)
Ever Torres-Silva, Juan Gaviria-Jiménez, Eider Pereira-Montiel, David Montoya-Arenas, José Flórez-Arango. 6795

XSL•FO
RenderX

Assessing a Community Health Worker-Facilitated, Digitally Delivered, Family-Centered Diabetes Management Program: Single-Arm Quasi-Experimental Study (e79032)
Zenong Yin, Vanessa Errisuriz, Heather Cuevas, Bertha Flores, Laura Delfausse, Christina Galvan, Jing Wang, Chengdong Li, Renata Morfin, Shiyu Li, Maysa Sapargeldiyeva, Giliane Muyna, Minyu Zhang, Vanessa Sweet, Deborah Parra-Medina. 6808

Collaborative Development of a Self-Tracking Assisted Psychotherapy Treatment Concept for Refugees With Complex Posttraumatic Stress Disorder: Participatory Action Research (e66663)
Lisa Riisager, Jakob Larsen, Lotte Huniche, Thomas Christiansen, Stine Moeller. 6821

The Alongside Digital Wellness Program for Youth: Longitudinal Pre-Post Outcomes Study (e73180)
Katherine Cohen, Andy Rapoport, Elsa Friis, Shannon Hill, Sergey Feldman, Jessica Schleider. 6841

Exploring Empowerment in Online Support Communities for People Living With Tic Disorders and Tourette Syndrome: Qualitative Survey Study of User Experiences (e66912)
Ella Ford, Neil Coulson, E Davies. 6856

ChatGPT-Based Chatbot for Help Quitting Smoking via Text Messaging: An Interventional Study (e79402)
Lorien Abroms, Christina Wysota, Artin Yousefi, Tien-Chin Wu, David Broniatowski. 6876

A Virtual Reality–Based mHealth App to Enhance Health Care Skills and Promote Health Equity by Empowering Health Care Professionals to Provide Empathetic, Compassionate, and Unbiased Patient Care Through Digital Experiential Learning: Qualitative Study (e64147)
Dixit Patel, Mayank Patel, Thomas Wischgoll, Yong Pei, Paul Hershberger. 6887

User Engagement With and Perceived Impact of a Digital Cognitive Training App on Cognition, Daily Functioning, and Mental Fitness: Secondary Analysis of Cross-Sectional Survey Data (e80027)
Kelsey McAlister, Lara Baez, Anna Braunsdorf, Yana Yang, Dan Kessler, Jennifer Huberty. 6912

Explainable AI-Driven Analysis of Radiology Reports Using Text and Image Data: Experimental Study (e77482)
Muhammad Zamir, Safir Khan, Alexander Gelbukh, Edgardo Felipe Riverón, Irina Gelbukh. 6925

Developing a Quick Isolation Bed Inquiry System During the COVID-19 Outbreak: User-Centered Design Approach Based on the Toyota Production System (e67152)
Chien-Chung Lin, Jian-Hong Shen, Shan-Li Chang, Tai-Chih Kuo, Hui-Ling Huang, Cheng-Yi Lin, Hung-Meng Huang. 6962

Quality of an Assistive Technology Web Application for Primary Care Physicians Serving Older Adults: Concurrent Mixed Methods Study (e69645)
Elsa Orellano-Colón, Wency Bonilla-Díaz, Radamés Revilla-Orellano, Jesús Mejías-Castro, Abiel Roche-Lima. 6983

A “Papageno” Story Interview Suicide Prevention Intervention for Young Adults With Past-Month Suicidal Ideation: Uncontrolled Single-Group Pilot Study of Feedback and Acceptability (e71368)
Jane Harness, Kamara Gardner, Jessica Schleider, Cheryl King. 7000

Safe Patient Handling Musculoskeletal Injury-Prevention Smartphone App for Community Health Care Workers: Mixed Methods Feasibility Study (e70681)
Alfred Lee, Sarah Burrell, Sandy Courtnall, Shelly Wake, Ryan Rhodes. 7013

Quality Assessment of Large Language Model–Generated Medical Dialogue for Clinical Vignettes: Evaluation Study (e80752)
Yasutaka Yanagita, Daiki Yokokawa, Shiichi Ihara, Ryo Yoshida, Yoshihide Okano, Takanori Uehara. 7035

Quantitative Assessment of Strabismus Using Cloud AI Computing: Validation Study (e79280)
Junxian He, Jiawei Zhang, Zheng Wang, Shrinivas Pundlik, Rui Liu, Gang Luo. 7046

XSL•FO
RenderX

Exploring Health Care Professionals' Perspectives on Using Video Feedback and Movement Analysis to Facilitate Physical Functioning for Older Adults Living at Home: Co-Design Approach (e73527)
Jessica Olovsson, Mirjam Ekstedt, Cecilia Fagerström, Patrick Bergman, Sofia Backåberg. 7058

Clinician Perspectives on Incorporating Physical Activity and Sleep Prescriptions Using eHealth for Youth With Comorbid Psychiatric Disorders: Qualitative Focus Group Study (e71569)
Alyssa Button, James Slavet, Amanda Staiano, April Bowling. 7069

Generative AI Chatbot for Diabetes Management: Formative 2-Part Qualitative Study Using DTalksBot Involving Patients and Clinicians (e72553)
Soyun Jeon, Seolhee Lee, Esther Kim, Jinsu Eun, Kwangwon Lee, Hajin Lim, Joonhwan Lee. 7078

Medical-Economic and Ecological Impact of Anesthesia Teleconsultation: Retrospective Observational Study (e70259)
Fabrice Ferre, Philippine Furelau, François Labaste, Nadège Costa, Fanny Vardon, Antoine Piau, Charlotte Martin, Vincent Minville. 7098

Development of a Virtual Home-Based Cycling Intervention for Individuals With Chronic Obstructive Pulmonary Disease: Qualitative Study (e71234)
Kristina Christensen, Henrik Bøggild, Ulla Weinreich, Anna Stoustrup. 7111

Development, Challenges, and Evolution of the Log2Lose Intervention for Weight Management: Randomized Controlled Digital Health Trial (e70842)
Ryan Shaw, Hailey Miller, Cherie Barnes, Clemontina Davenport, Sarah Morton-Oswald, Sarah Jackson, Michelle Bean, Jeff Cohen, Chance Griffin, Jane Pendergast, Maren Olsen, Jennifer Gierisch, Corrine Voils. 7123

User-Centered Formative Evaluation of Cognitive Rehabilitation Software: Cognitive Walkthrough and System Usability Scale Study (e75805)
Seojin Hong, Hyun Choi, Hyosun Kweon. 7134

Three Body-Worn Accelerometers in the French NutriNet-Santé Cohort: Feasibility and Acceptability Study (e76167)
Abdouramane Soumaré, Léopold Fezeu, Jérôme Bouchan, Fabienne Delestre, Alice Bellicha, Greet Cardon, Alan Donnelly, Antje Hebestreit, Mathilde Touvier, Jean-Michel Oppert, Jérémy Vanhelst. 7145

Using Human-Centered Design and Development to Create a Digital Sick Day Medication Guidance Application for People With Diabetes, Cardiovascular Disease, or Chronic Kidney Disease: Mixed Methods Study (e77240)
Julie Babione, Sarah Gil, Chidera Okemeziem, Taylor Palechuk, Kaitlyn Watson, Sandra Robertshaw, Nancy Verdin, Kerry McBrien, David Campbell, Ross Tsuyuki, Neesh Pannu, Matthew James, Maoliosa Donald. 7157

GrantCheck—an AI Solution for Guiding Grant Language to New Policy Requirements: Development Study (e79038)
Qiming Shi, Asil Oztekin, George Matthew, Jeffrey Bortle, Hayden Jenkins, Steven Wong, Paul Langlois, Anaheed Zaki, Brian Coleman, Katherine Luzuriaga, Adrian Zai. 7176

AI-Powered Thermography for Diabetic Foot Risk Stratification: Multicenter Cross-Sectional Study (e81289)
Meshari Alwashmi, Mustafa alghali, Waseem Abu-Ashour, Abdullah Arabe, Abdullah Al Soheimi, Nasibh Alharbi, Hani Badahdah. 7189

A 6-Month Evaluation of the Peer-Ceived Momentary Assessment Method in a Small Sample of Liver Transplant Patients and Their Support Persons: Longitudinal Observational Study (e55907)
Allan Berrocal, Waldo Concepcion, Katarzyna Wac. 7205

XSL•FO
RenderX

Comparison of ChatGPT and DeepSeek on a Standardized Audiologist Qualification Examination in Chinese: Observational Study (e79534)
Beier Qi, Yan Zheng, Yuanyuan Wang, Li Xu. 7230

A Rule-Based Conversational Agent for Mental Health and Well-Being in Young People: Formative Case Series During the Rise of Generative AI (e69841)
Aimee-Rose Wrightson-Hester, Gee Anderson, Joel Dunstan, Peter McEvoy, Christopher Sutton, Bronwyn Myers, Sarah Egan, Sara Tai, Melanie Johnston-Hollitt, Wai Chen, Tom Gedeon, Joanna Moullin, Warren Mansell. 7240

WhatsApp-Based Intervention for Diabetes Prevention and Care in Argentina: Implementation and Process Evaluation (e81098)
Analia Nejamis, Laura Gutierrez, María López, Alvaro Ciganda Cha, Omayra Mendoza Quispe, Vilma Irazola, Andrea Beratarrechea. 7258

Personalizing a Weight Loss Program Using Cognitive-Behavioral Phenotypes to Improve Engagement and Weight Loss in Adults With Overweight or Obesity: Quasi-Experimental Study (e72645)
Joanna Szypula, Andreas Jarvstad, Lucy Jones, Katy Tapper. 7274

Accuracy and Variability of a Commercial Markerless Motion Capture System Compared to a Pressure Mat for Weight Distribution in Standing: Cross-Sectional Observational Study (e73575)
Lisa Sheehy, Emma Gal-Dev, Heidi Sveistrup, Martin Bilodeau, Hillel Finestone. 7289

Depression and Anxiety Outcomes in a Technology-Enabled Psychotherapy Practice: Retrospective Cohort Study (e76264)
Nicholas Forand, Jasmine Nettiksimmons, Margaret Anton, Raven Truxson, Karl Vanderwood, Brandn Green. 7301

Feasibility of a Specialized Large Language Model for Postgraduate Medical Examination Preparation: Single-Center Proof-Of-Concept Study (e77580)
Yun Leong, Lathiga Nambiar, Victoria Tay, Sui Lie, Ke Yuhe. 7318

Self-Reported Triggers Evaluation of High-Risk Dietary and Environmental Factors Preceding Migraine Onset by Using a Mobile Tracking App (Migraine Insight): Comparative Analysis Study (e59951)
Christina Wornom, Brooklyn Brekke-Kumley, Tavsimran Luthra, Lynn Smith, Jane Harrington. 7328

The Mechanism of Online Health Information Seeking Switching to Online Medical Consultation: Cross-Sectional Study (e78397)
Lijiang Zhang, Jingjing Xia, Hui Chen, Yang Bai, Jun Wang, Liuan Wang, Wenjie Ren. 7336

A Mobile App to Enhance Awareness of Vaccination in Adults With Psoriasis and Atopic Dermatitis: Development and Preliminary Evaluation Study (e80113)
Giulia Gasparini, Donatella Panatto, Irene De Kermarek, Lucia Valchi, Elvira Massaro, Piero Lai, Martina Burlando, Giancarlo Icardi, Emanuele Cozzani. 7353

Development and Implementation of an Internal Quality Control and External Quality Assessment Information System for a Regional Medical Laboratory Center: Pilot Design and Implementation Study (e77043)
Dayang Chen, Hongmei Mo, Yingying Zhang, Junlong Qin, Lijun Zhang, Changxian Jiang, Fuming Li, Xiuming Zhang, Tong Ou. 7365

Access to Specialized Medical Training in Spain and Determinants of Failure in the National Entrance Examination: Econometric Modeling Study (e72440)
Montserrat Diaz-Fernandez, Mar Llorente-Marron, Victor Asensi. 7378

Proactively Delivered Digital Mental Health Support for Health Care Workers: Usability and Acceptability Evaluation (e74086)
Lauren Southwick, Rachel Gonzales, Lisa Bellini, David Asch, Nandita Mitra, Mohan Balachandran, Courtney Wolk, Emily Becker-Haimes, Rachel Kishton, Sarah Beck, Raina Merchant, Anisk Agarwal. 7390

XSL•FO
RenderX

Localized Muscular Fatigue in Robotic-Assisted Laparoscopic Surgery: Predictive Modeling Study (e68536)
Daniel Caballero, Manuel Pérez-Salazar, Juan Sánchez-Margallo, Francisco Sánchez-Margallo. 7401

Awareness, Perceptions, Willingness, and Feasibility of mHealth Apps Among People Living With Epilepsy: Cross-Sectional Questionnaire Study (e80283)
Bushra Zahra, Muhammad Hamza, Rehana Sarwat, Hina Zahra, Ayesha Azam, Ali Ahmed. 7419

AI-Enabled Personalized Smoking Cessation Intervention With the Aipaca Chatbot: Mixed Methods Feasibility Study (e73319)
Yunlong Liu, Paul Calle, Mariah Vadakekut, Daniel Rubin, Zsolt Nagykaldi, Mark Doescher, Lisa Hightow-Weidman, Chongle Pan, Ruosi Shao.
7 4 3 8

The Role of Practitioner- and User-Set Goals in Engagement and Psychological Distress Among Kooth Digital Health Users: Retrospective Analysis (e70818)
Jacqlyn Yourell, Jennifer Huberty, Terry Hanley, Louisa Salhi. 7454

The Impact of a Digital Peer-Supported App on Daily Steps and Lifestyle Changes Among Individuals With Prediabetes and Early-Stage Type 2 Diabetes: Prospective, Nonrandomized Controlled Trial (e75953)
Shota Yoshihara, Takeo Shibata, Mizuki Kosone, Megumi Shibuya, Kayoko Takahashi. 7464

ROOM to Grow, a Mobile Well-Being Intervention for University Students: Overview of the Design Process and Outcomes (e63325)
Tajda Laure, Camila Villegas Meija, Danielle Remmerswaal, Djameela Dulloo, Ruth Van der Hallen, Birgit Mayer, Marianne Littel, Bram Dierckx, Jeroen Legerstee, Rutger Engels, Marilisa Boffo. 7477

Performance of DeepSeek-R1, ChatGPT (GPT-o3-mini), and Gemini 2.0 Flash on German Medical Multiple-Choice Questions: Comparative Evaluation (e77357)
Annika Meyer, Yassin Karay, Andrea Steinbicker, Thomas Streichert, Remco Overbeek. 7498

Identification of Design Requirements for a Software Application for Use by Clinicians That Collects Acute Stroke Treatment Data During Clinical Workflow: Pilot Study (e64800)
Adam Forward, Gizem Koca, Aymane Sahli, Noreen Kamal. 7507

Usability of a Tablet-Based Cognitive Assessment Administered by Medical Assistants in General Practice: Implementation Study (e76010)
Philipp Schaper, Alexander Hanke, Stephan Jonas, Leon Nissen, Lara Reimer, Florian Schweizer, Michael Wagner, Kristin Rolke, Carolin Rosendahl, Judith Tillmann, Klaus Weckbecker, Jochen Thyrian. 7522

Providing Mobile Patient Access to Their Electronic Secondary Care Patient Record in Adults With Cystic Fibrosis: Results of a Prospective, Parallel, Randomized Open-Pilot Quantitative Study (e69747)
Helen Chadwick, Akhil Sawant, Helen White, Lindsey Gillgrass, Giulia Spoletini, Ian Clifton, Christine Etherington, Daniel Peckham. 7534

A Mobile App (MyPeer) Co-Designed With Immigrant Adolescents for Better Sexual and Reproductive Health: Usability Study (e81115)
Salima Meherali, Amyna Rehmani, Mariam Ahmad, Piper Scott-Fiddler, Paula Pinzón-Hernández, Zeba Khan, Sarah Flicker, Philomina Okeke-Ihejirika, Bukola Salami, Eleni Stroulia, Ashley Vandermorris, Josephine Pui-Hing Wong, Wendy Norman, Shannon Scott, Sarah Munro.
7 5 5 2

A Blended Educational Program to Promote Dialogue on Patient Safety Between Patient and Family Advisory Councils and Health Care Organizations: Codevelopment Study (e79286)
Yannick Blum, Larissa Brust, Matthias Weigl, Qëndresa Rramani Dervishi. 7569

XSL•FO
RenderX

Feasibility of Physical Therapy Evaluation Symptom Provocation Tests in Older Adults With Mild Traumatic Brain Injury: Mixed Methods Study (e76799)
Carrie Barrett, Mark Goetting, Alessander Danna-dos-Santos. 7582

Predictors for Acceptance of Sexual Aggression Myths Among People Using Cyberporn: Cross-Sectional Study (e75485)
Farah Ben Brahim, Germano Vera Cruz, Robert Courtois, Yasser Khazaal. 7595

Associations Between Online Casual Sexual Behavior and HIV-Related Risk Behaviors Among Men Who Have Sex With Men in Southeast China: Cross-Sectional Study (e71690)
Lin Chen, Zhongrong Yang, Wanjun Chen, Tingting Jiang, Xiaohong Pan. 7613

Experiences of Health Research Data Sharing Among Researchers in Sub-Saharan Africa: Cross-Sectional Study (e69411)
Felix Sukums, Bernard Ngowi, Rebecca Chaula, Kim Weiszhar, Akili Kalinga, Olena Ivanova, Clarissa Prazeres da Costa, Achim Hoerauf, Sundeep Sahay, Emilia Noormahomed, Linda Debrah, Chummy Sikasunge, Helena Ngowi, Andrea Winkler. 7624

Factors Influencing the Use of Online Symptom Checkers in the United Kingdom: Cross-Sectional Study (e65314)
Austen El-Osta, Eva Riboli-Sasco, Mahmoud Al Ammouri, Sami Altalib, Ana Luisa Neves, Azeem Majeed, Benedict Hayhoe. 7642

Interaction of Long Working Hours and Sense of Coherence on Objective Total Sleep Time: Cross-Sectional Study From the SLEPT Study (e73784)
Kei Muroi, Emi Morita, Sumire Matsumoto, Asuka Ishihara, Sumi Hasegawa, Mami Ishitsuka, Daisuke Hori, Shorato Doki, Tsukasa Takahashi, Shin-ichiro Sasahara, Takashi Kanbayashi, Masashi Yanagisawa, Makoto Satoh, Ichiyo Matsuzaki. 7656

Environmental Triggers of Specific Subtypes of Agitation in People With Dementia: Observational Study (e60274)
Hannah Davidoff, Alex Van Kraaij, Erika Lutin, Laura Van den Bulcke, Mathieu Vandenbulcke, Nick Van Helleputte, Maarten De Vos, Chris Van Hoof, Maarten Van Den Bossche. 7669

Exploring Co-production as an Implementation Strategy for Trauma-Informed Care in a Youth-Focused HIV Clinic in Memphis, Tennessee: Mixed Methods Research (e66426)
Leslie Brown, Megan Wilkins, Robert McLean, Jamie Stewart, Nehali Patel, Almariana Acuña, Carolyn Audet, Jessica Sales, April Pettit, Latrice Pichon. 7683

Food Handling Practices Among Food Businesses in Jigjiga, Eastern Ethiopia, During the COVID-19 Pandemic: Cross-Sectional Study (e64214)
Alinoor Farah, Tahir Nour, Abdurahman Roble, Muse Obsiye, Mowlid Aden, Abdulahi Abdi, Fentabil Getnet. 7696

A Novel Approach to Care Redesign Collaboration Between Emergency and Specialty Departments: Qualitative Experience Report (e22028)
Sonia Harris, Christian Rose, Samantha Kling, Brett Cohen, Olga Goldberg, Hannah Walton, Darlene Veruttipong, Mohamed Alhadha, Sheneé Laurence, Shashank Ravi, Carl Gold, Jonathan Shaw, Laurice Yang, Cati Brown-Johnson. 7707

Variability in the Diagnosis and Treatment of Interstitial Cystitis/Bladder Pain Syndrome: Internet Survey (e70813)
Eleanora Melkonian, A Garrett, Erika Kline, Pressley Smith, Madelyn Wiesenhahn, Jeanette Petit, Alison Swierczynski, Carley Zhou, Stuart Bauer, Rosalyn Adam, Kamil Barbour, Sonja Ziniel, Catherine Brownstein. 7720

Continuous Rural High School Educational Outreach and Lasting Impact on Health Care Career Attitudes: Qualitative Pilot Study (e70505)
Rebecca Bolen, Meagan Flesch, Jerrica Dennis, Logan Shouse, Mark Payton, David Ross. 7736

Mediating Effects of Psychological Independence and Social Support on the Association Between Family Strength and Depression in Young Korean Adults: Cross-Sectional Study (e71485)
Sunyoung Kim, Suin Park, Hyunlye Kim, Dabok Noh. 7744

XSL•FO
RenderX

Determinant Factors of Stress in Caregivers of Patients With Schizophrenia: Cross-Sectional Study (e70648)
Isymiarni Syarif, Hasnawati Amqam, Saidah Syamsuddin, Veni Hadju, Syamsiar Russeng, Yusran Amir. 7756

Association of Self-Rated Health in Pregnancy With Maternal Childhood Experiences, Socioeconomic Status, Parity, and Choice of Antenatal Care Providers: Cross-Sectional Study (e68811)
Bjarne Austad, Gunnhild Vie, Mari Hansen, Hanna Mørkved, Linn Getz, Bente Mjølstad. 7774

Assessing the Causal Association Between COVID-19 and Graves Disease: Mendelian Randomization Study (e66003)
Hui Nian, Yu Bai, Hua Yu. 7794

COVID-19 Perceptions Among Communities Living on Ground Crossings of Somali Region of Ethiopia: Community Cross-Sectional Survey Study (e66751)
Alinoor Farah, Abdifatah Abdulahi, Abdulahi Hussein, Ahmed Hussein, Abdi Osman, Mohamed Mohamud, Hasan Mowlid, Girum Hailu, Fathia Alwan, Ermiyas Bizuneh, Ahmed Ibrahim, Elyas Abdulahi. 7803

Synthetic Data-Driven Approaches for Chinese Medical Abstract Sentence Classification: Computational Study (e54803)
Jiajia Li, Zikai Wang, Longxuan Yu, Hui Liu, Haitao Song. 7821

Identifying High-Priority Ethical Challenges for Precision Emergency Medicine: Nominal Group Study (e68371)
Christian Rose, Emily Shearer, Isabela Woller, Ashley Foster, Nicholas Ashenburg, Ireh Kim, Jennifer Newberry. 7834

Impact of Psychological Resilience on the Fear of Pain and Activity Recovery in Postsurgical Patients: Observational Cohort Study (e63556)
Yang Luo, Sisi Li, Lijuan Feng, Junyi Zheng, Chunfen Peng, Lihong Bao. 7846

Impact of Preventive Intravenous Amiodarone on Reperfusion Ventricular Fibrillation in Patients With Left Ventricular Hypertrophy Undergoing Open-Heart Surgery: Randomized Controlled Clinical Trial (e64586)
Chan-Juan Gong, Xiao-Kai Zhou, Zhen-Feng Zhang, Yin Fang. 7855

Identification of Behavioral, Clinical, and Psychological Antecedents of Acute Stimulant Poisoning: Development and Implementation of a Mixed Methods Psychological Autopsy Study (e64873)
Marley Antolin Muñiz, Vanessa McMahan, Xochitl Luna Marti, Sarah Brennan, Sophia Tavasieff, Luke Rodda, James Knoll, Phillip Coffin. . . 7 8 6 6

Resilience Informatics in Public Health: Qualitative Analysis of Conference Proceedings (e63217)
Maiya Block Ngaybe, Lidia Azurdia Sierra, Andrew McNair, Myla Gonzalez, Mona Arora, Kacey Ernst, Enrique Noriega-Atala, M Iyengar. . . 7 8 7 5

Assessment, Decision, Adaptation, Production, Topical Experts-Integration, Training, and Testing (ADAPT-ITT) Framework to Tailor Evidence-Based Posttraumatic Stress Disorder Treatment for People With HIV to Enhance Engagement and Adherence: Qualitative Results from a Feasibility Randomized Controlled Trial (e64258)
Cristina Lopez, Angela Moreland, Stephanie Amaya, Erin Bisca, Christin Mujica, Tayler Wilson, Nathaniel Baker, Lauren Richey, Allison Eckard, Patricia Resick, Steven Safren, Carla Danielson. 7891

Designing the First Pregnancy Guaranteed Income Program in the United States: Qualitative Needs Assessment and Human-Centered Design to Develop the Abundant Birth Project (e60829)
Deborah Karasek, Jazzmin Williams, Michaela Taylor, Monica De La Cruz, Stephanie Arteaga, Sabra Bell, Esperanza Castillo, Maile Chand, Anjeanette Coats, Erin Hubbard, Latriece Love-Goodlett, Breezy Powell, Solaire Spellen, Zea Malawa, Anu Gomez. 7905

XSL•FO
RenderX

Tobacco Use, Risk Perceptions, and Characteristics of Adults Who Used a Heated Tobacco Product (IQOS) in the United States: Cross-Sectional Survey Study (e57398)
Hui Cheng, Brendan Noggle, Andrea Vansickel, Edward Largo, Pierpaolo Magnani. 7919

Predictors of the Intention to Stop Using Smart Devices at Bedtime Among University Students in Saudi Arabia: Cross-Sectional Survey (e67223)
Manal Almalki. 7934

Impact of a Sensorimotor Integration and Hyperstimulation Program on Global Motor Skills in Moroccan Children With Autism Spectrum Disorder: Exploratory Clinical Quasi-Experimental Study (e65767)
Rachid Touali, Jamal Zerouaoui, El Chakir, Hung Bui, Mario Leone, Maxime Allisse. 7951

Quality Assessment of Web-Based Information Related to Diet During Pregnancy in Pregnant Women: Cross-Sectional Descriptive Study (e64630)
Daichi Suzuki, Etsuko Nishimura, Rina Shoki, Ishak Octawijaya, Erika Ota. 7966

Engaging Community-Dwelling Older Adults in Research: Qualitative Substudy of Factors Impacting Participation (e74191)
Bryah Boutilier, Grace Warner, Brianna Wolfe, Sorayya Askari, Elaine Moody, Parisa Ghanouni, Tanya Packer. 7974

Detecting Conversation Topics in Recruitment Calls of African American Participants to the All of Us Research Program Using Machine Learning: Model Development and Validation Study (e65320)
Priscilla Pemu, Michael Prude, Atuarra McCaslin, Elizabeth Ojemakinde, Christopher Awad, Kelechi Igwe, Anny Rodriguez, Jasmine Foriest, Muhammed Idris. 7984

Development of a Clinic-Based, Sociostructural Intervention to Improve the Provision of Pre-Exposure Prophylaxis for Cisgender Women: Formative Study Using the Assessment, Decision, Adaptation, Production, Topical Experts, Integration, Training, and Testing (ADAPT-ITT) Framework (e75922)
Rachel Scott, Shawnika Hull, Deanna Kerrigan, Mandi Pratt-Chapman, Tara Mathias-Prabhu, Jennifer Zack, Michelle Xu, Marisa Sadauskas, Patricia Moriarty, Tranessa Hanson, Hannah Arem. 7997

Parental Challenges in Raising Preschoolers With Attention-Deficit/Hyperactivity Disorder in Mainland China's First-Tier Cities: Qualitative Study Using Framework Analysis (e74047)
Shu-Cheng Chen, Chen-Wen Zhong, Han Li, Xiaoqing Li, Hong-Wang Fung, Lijun Wang, Guo-Tao Wu, Xin-Yue Jiang, Wing-Fai Yeung. 8012

The Moderating Effect of Atypical Events on the Relationship Between Heart Rate and Stress in Medical Residents Working in an Intensive Care Unit: Longitudinal, Observational Daily Diary Study (e67822)
Ruibei Li, Ujjwal Pasupulety, Wellington Chang, Adam Frank. 8028

Association Between In-Hospital Applications for Long-Term Care Services and Hospital Length of Stay Among Older Adults: Ecological Cross-Sectional Study (e76782)
Naoki Takashi, Joji Onishi, Misaki Fujisawa, Tomoko Ohura, Shosuke Ohtera. 8040

A Neuroaffirmative, Self-Determination Theory–Based Psychosocial Intervention for Adults With Attention-Deficit/Hyperactivity Disorder: Randomized Feasibility Study (e69943)
Rebecca Champ, Rita Meneses, Marios Adamou, Warren Gillibrand, Sally Arrey, Barry Tolchard. 8049

Reconditioning Emotional Responses With the Break Method: Pilot Quantitative Study (e75056)
Boaz Salik, Bizzie Gold, Kira Krier. 8074

Knowledge, Attitudes, and Practices Toward Burn Causes and First Aid Management in Jazan Region, Saudi Arabia: Cross-Sectional Study (e80594)
Anas Sayegh, Alyazid Awaji, Anas Fathuldeen, Atheer Alshammakhi, Roaa Alhazmi, Razan Ageeli, Ghadi Shamakhi. 8087

XSL•FO
RenderX

Development of an Evaluation Index System for Health Recommender Systems Based on the Health Technology Assessment Framework: Cross-Sectional Delphi Study (e79997)
Yue Sun, Shijie Hou, Siye Chen, Minmin Leng, Zhiwen Wang. 8101

Radiology Staff Experiences With Integrating Artificial Intelligence Into Radiology Practice in a Swedish Hospital: Qualitative Case Study (e77843)
Per Nilsen, Petra Svedberg, Ingrid Larsson, Lena Petersson, Jens Nygren, Emilie Steerling, Margit Neher. 8114

Prediction of Respiratory Decompensation in Patients Receiving Home Mechanical Ventilation: Machine Learning Model Development and Validation Study (e78941)
Nerea Berbel Casado, Francesc López Seguí, Natalia Muñoz Moruno, Antoni Rosell, Aïda Muñoz Ferrer, Ignasi Garcia Olive, Marina Galdeano Lozano. 8126

Comparison and Validation of Actigraphy Algorithms Using a Large Community Dataset: Algorithm Validation Study (e70778)
Darshan Panesar, Aashish Vichare, Jason Goncalves, Robyn Stremler. 8135

Consumer Co-Design of an Online Resource to Build Communication Skills of Health Consumers: Mixed Methods Study (e77263)
Alison Beauchamp, Julieanne Hilbers, Natali Cvetanovska, Anna Wong Shee, Lidia Horvat, Sandra Rogers, Andrea Cooper, Elizabeth Flemming-Judge, Sue Rawlinson, Rebecca Jessup. 8159

Exploration of Factors That Affect Engagement With the Experience Sampling Method and Service Users' Experience of This Within the AVATAR2 Trial: Mixed Methods Study (e78204)
Sophie Dennard, Philippa Garety, Clementine Edwards, Andrew Gumley, Oliver Owrid, Lucy Miller, Stephanie Allan, Alison Duerden, Francis Yanga, Chloe Burns, Helena Fletcher, Amy Grant. 8173

Exploring the Usability and Acceptability of the FoodMATS-Youth App for Monitoring Food Marketing Exposures: Mixed Methods Study (e79306)
Idris Bamigbayan, Laurie Twells, Kirby Shannahan, Rachel Prowse. 8187

Initial Development of a Multidimensional Computerized Adaptive Test for Intensive Longitudinal Assessment of Suicide Risk: Development and Usability Study (e76544)
Kenneth McClure, Brooke Ammerman, Cheng Liu, Miguel Blacutt, Connor O'Brien, Ross Jacobucci. 8204

Technology-Related Trauma in Sexual and Reproductive Health Digital Technologies: Grounded Theory Study (e79365)
Abdul-Fatawu Abdulai, Janell Josephs, Adrian Guta, Mark Gilbert, Vicky Bungay. 8229

Compliance and Satisfaction With a Protocol for Identifying Novel Targets to Support Postpartum Opioid Use Disorder Recovery: Prospective Cohort Study (e77899)
Alicia Allen, Linnea Linde-Krieger, Jendar Deschenes, Stephanie Mallahan, Alexandra Harris, Mariana Felix, Arushi Chalke, Alma Anderson, Priyanka Sharma, Katherine King, Maddy Grant, James Baurley, Lela Rankin, Stacey Tecot. 8243

Identifying Biomedical Entities for Datasets in Scientific Articles: 4-Step Cache-Augmented Generation Approach Using GPT-4o and PubTator 3.0 (e73822)
Claudia Giuliani, Gita Benadi, Felix Engel, Jonas Werner, Manuel Watter, Guido Schwarzer, Olaf Groß, Robert Zeiser, Harald Binder, Klaus Kaier. 8260

Assessing the Usability of a Novel Toolkit for Creating Visual Key Information Pages for Informed Consent for Research: Mixed Methods Usability Study (e76740)
Eliana Goldstein, Zoe Troubh, Krista Cooksey, Molly Volkmar, Victor Catalan Gallegos, Kimberly Kaphingst, Clara Lee, Ashley Housten, Jessica Mozersky, Mary Politi. 8272

XSL•FO
RenderX

Effects, Feasibility, and Safety of an Early Mobilization Protocol With Immersive Virtual Reality for Dyspnea in Patients With Acutely Decompensated Heart Failure: The MOVE Randomized Clinical Trial (e80729)
Iasmin Fraga, Larissa Caballero, Carlos Tremea, Janaína Prates, Gabrielle Perin, Vitor Guedes, Pedro Lago, Eneida Rabelo-Silva. 8281

Behavior Emotion Therapy System and You: Co-Design and Evaluation of a Mental Health Chatbot and Digital Human for Mild to Moderate Anxiety in Healthy Participants (e66163)
Almira Thunström, Lilas Ali, Hanne Krage Carlsen, Maria Bohm, Linda Wesén, Olof Wrede, Iris Sarajlic Vukovic, Andreas Hellström, Tomas Larson, Steinn Steingrimsson. 8291

Assessing the Usability of a Prescription-Based Mobile App for Patients With Panic Disorder and Its Management Console for Clinicians: Controlled User Study (e76843)
Yujin Ko, Jeemin Lee, Kyunghee Ham, Yesol Cho, Yu-Bin Shin, Choongki Min, Kyungnam Kim, Wonseuk Jang, Hayeon Jung, Jae-Jin Kim. . . 8304

Designing Culturally Adapted Digital Mental Health Support Tool for Chinese-Speaking International Students in Australia: A Qualitative Co-design Study (e76695)
Ling Wu, Chen Zhu, Joshua Seguin, Jue Xie, Pranav Kulkarni, Mingye Li, Patrick Olivier. 8317

Preprocessing Large-Scale Conversational Datasets: A Framework and Its Application to Behavioral Health Transcripts (e78082)
Paz Naim, Shiri Sadeh-Sharvit, Samuel Jefroykin, Eddie Silber, Dennis Morrison, Ariel Goldstein. 8335

Urban Environment and Outdoor Walking Mobility Among Older Adults With and Without Cognitive Impairment in Singapore: Formative Study (e76777)
Yuezhong Liu, Chek Wong, Noam Shoval, Moon-Ho Ho, Yin-Leng Theng. 8352

Culturally Tailored Women's Empowerment Strategies to Improve Immunization Uptake: Formative Research Using a Human-Centered Design Approach (e60089)
Muyi Aina, Uchenna Igbokwe, Raihanah Ibrahim, Itoro Ata, Safiya Atta, Tobiloba Adaramati, Kunle Oreagba, Ernest Ezeogu, Precious Uahomo, Eric Aigbogun Jr. 8365

Representation of Medical Concepts in Emojis Using Medical Subject Headings to Identify Gaps and Opportunities: Cross-Sectional Analysis (e70130)
Shuhan He, Boyu Peng, Suhanee Mitragotri, Ahmad Hassan, Abdel Ariss, Margarita Monge, Norawit Kijpaisalratana. 8380

Application of Large Language Models in Data Analysis and Medical Education for Assisted Reproductive Technology: Comparative Study (e70107)
Noriyuki Okuyama, Mika Ishii, Yuriko Fukuoka, Hiromitsu Hattori, Yuta Kasahara, Tai Toshihiro, Koki Yoshinaga, Tomoko Hashimoto, Koichi Kyono. 8392

Response to the Netflix Docuseries "Big Vape: The Rise and Fall of JUUL": Mixed Methods Analysis of YouTube Comments Using Qualitative Coding and Topic Modeling (e76737)
Beth Hoffman, Arpita Tripathi, Ariel Shensa, Julia Dou, Piper Narendorf, Nishi Hundi, Jaime Sidani. 8406

Assessing Work-Related Stressors in Online Counseling: Cross-Sectional Questionnaire Development Study (e65766)
Wiebke Schlenger, Marlies Joellenbeck, Elke Ochsmann. 8418

Bactericidal Efficacy of Ultraviolet-C Light on Virtual Reality Devices: In Vitro Assessment of Bacterial Killing (e70326)
Scott Roberts, Jayson Wright, Mahnoor Mustafa, Richard Feinn, Asher Marks, Kimberly Hieftje, Pamela Huang, Richard Martinello, Thomas Murray, HIPSTER Team. 8427

Comparison of Sleep Features Across Smartphone Sensors, Actigraphy, and Diaries Among Young Adults: Longitudinal Observational Study (e67455)
Jaclyn Kirshenbaum, Ryann Crowley, Melissa Latham, David Pagliaccio, Randy Auerbach, Nicholas Allen. 8437

XSL•FO
RenderX

Exploring Young Adults' Experiences and Beliefs in Asthma Medication Management: Pilot Qualitative Study Comparing Human and Multiple AI Thematic Analysis (e69892)
Ruth Jeminiwa, Caroline Popielaski, Amber King. 8449

A Multimodal Large Language Model as an End-to-End Classifier of Thyroid Nodule Malignancy Risk: Usability Study (e70863)
Gerald Sng, Yi Xiang, Daniel Lim, Joshua Tung, Jen Tan, Chiaw Chng. 8461

Prediction of 1-Year Activity in Systemic Lupus Erythematosus: Hierarchical Machine Learning Approach (e70200)
Livia Lilli, Laura Antenucci, Augusta Ortolan, Silvia Bosello, Stefano Patarnello, Carlotta Masciocchi, Marco Gorini, Gabriella Castellino, Alfredo Cesario, Maria D'Agostino, Jacopo Lenkowicz. 8476

Evaluating Use Patterns of a Closed Electronic Nicotine Delivery System Among Adults in the United States Who Smoke Cigarettes Daily: 8-Week Actual Use Study (e76019)
Steve Roulet, Claudia Kanitscheider, Pierpaolo Magnani, Gerd Kallischnigg. 8489

Identifying Adverse Drug Events in Clinical Text Using Fine-Tuned Clinical Language Models: Machine Learning Study (e71949)
Elizaveta Kopacheva, Aron Henriksson, Hercules Dalianis, Tora Hammar, Alisa Lincke. 8508

Enhancing Clinical Data Management Through Barcode Integration and Research Electronic Data Capture: Scalable and Adaptable Implementation Study (e70016)
Rendong Zhang, Sophie Chiron, Regina Tyree, Kate Carson, Larry Raber, Karthik Ramadass, Chenyu Gao, Michael Kim, Lianrui Zuo, Yike Li, Zhiyu Wan, Paul Harris, Qi Liu, Ken Lau, Lori Coburn, Keith Wilson, Yuankai Huo, Bennett Landman, Shunxing Bao. 8523

Acceptability of Active and Passive Data Collection Methods for Mobile Health Research: Cross-Sectional Survey of an Online Adult Sample in the United States (e64082)
Nelson Roque, John Felt. 8539

Concordance Between Survey and Electronic Health Record Data in the COVID-19 Citizen Science Study: Retrospective Cohort Analysis (e58097)
Elizabeth Crull, Emily O'Brien, Pavel Antiperovitch, Kirubel Asfaw, Alexis Beatty, Djeneba Djibo, Alan Kaul, John Kornak, Gregory Marcus, Madelaine Modrow, Jeffrey Olgin, Jaime Orozco, Soo Park, Noah Peyser, Mark Pletcher, Thomas Carton. 8549

Exploratory Impact of iCARE Nigeria, a Combined mHealth and Peer Navigation Intervention, on Depressive Symptoms and Substance Use Among Youth Living With HIV in Nigeria: Single-Arm Trial (e71141)
Olusegun Ogunmola, Rita Tamambang, Kehinde Kuti, Lisa Kuhns, Olutosin Awolude, Adedotun Adetunji, Bibilola Oladeji, Oladayo Olaleye, Adeola Oyerinde, Robert Garofalo, Babafemi Taiwo, Olayinka Omigbodun. 8560

Virtual Reality Gamification of Visual Search, Response Inhibition, and Visual Short-Term Memory Tasks for Cognitive Assessment: Experimental Study (e65836)
Marios Hadjiaros, Andria Shimi, Kleanthis Neokleous, Constantinos Pattichis, Marios Avraamides. 8569

Visualization of FDA Adverse Drug Reaction Reports: Development and Usability Study of the VisDrugs Web Server (e71519)
Renjun Yang, Nuoya Yin, Yang Zhang, Francesco Faiola. 8587

Use of a Medical Communication Framework to Assess the Quality of Generative Artificial Intelligence Replies to Primary Care Patient Portal Messages: Content Analysis (e71966)
Natalie Lee, Nathan Richards, Jodi Grandominico, Robert Cronin, Amanda Hendricks, Ravi Tripathi, Daniel Jonas. 8602

XSL•FO
RenderX

Acceptance, Drivers, and Barriers to Use of mHealth Apps to Improve Quality of Life in Female Patients Affected by Hypothyroidism: Cross-Sectional Study (e67317)
Moritz Doll, Ranujan Chandrakumar, Lisa Jahre, Eva-Maria Skoda, Hannah Dinse, Dagmar Führer, Eleni Lampropoulou, Martin Teufel, Alexander Bäuerle. 8612

A Supplemental Women's Health Questionnaire for Women Veterans With Military Environmental Exposures: Project Development and Implementation (e73223)
Leah Eizadi, Mehret Assefa, Jordan Nechvatal, G Veltkamp, Abou Ibrahim-Biangoro, Maheen Adamson, Jennifer Jennings. 8625

Construction of Personalized Predictive Models for Missed Medication Doses Using Wearable Device Data: Prospective Observational Study (e72113)
Haru Iino, Hayato Kizaki, Shungo Imai, Satoko Hori. 8638

Evaluating the Quality of Psychotherapy Conversational Agents: Framework Development and Cross-Sectional Study (e65605)
Kunmi Sobowale, Daniel Humphrey. 8651

Association of a Brief Computerized Cognitive Assessment With Cholinergic Neurotransmission: Assessment Validation Study (e68374)
Mouna Attarha, Ana De Figueiredo Pelegrino, Lydia Ouellet, Paule-Joanne Toussaint, Sarah-Jane Grant, Thomas Van Vleet, Etienne de Villers-Sidani. 8665

Association of Perceived Xingfu With Health-Related and Socioeconomic Factors Among Hong Kong Chinese Adults: Cross-Sectional Study Using a Novel Single-Item Tool (e73350)
Katherine Sze, Sai Ho, Agnes Lai, Jing Jia, Heng Xu, Shirley Sit, Tai Lam, Man Wang. 8676

Validation of The Umbrella Collaboration for Tertiary Evidence Synthesis in Geriatrics: Mixed Methods Study (e75215)
Beltran Carrillo, Marta Rubinos-Cuadrado, Jazmin Parellada, Alejandra Palacios, Beltran Carrillo-Rubinos, Fernando Canillas, Juan Baztán Cortés, Javier Gómez-Pavón. 8686

Determining the Requirements of Vulnerable Groups for Health Counseling and Optimizing the Evaluation of Health Consultations: Mixed Methods Study With the Use of AI (e68888)
Annina Boehm-Fischer, Luzi Beyer. 8702

Integrating Nurse Preferences Into AI-Based Scheduling Systems: Qualitative Study (e67747)
Fabienne Renggli, Maisa Gerlach, Jannic Bieri, Christoph Golz, Murat Sariyar. 8715

Exploring Generative Pre-Trained Transformer-4-Vision for Nystagmus Classification: Development and Validation of a Pupil-Tracking Process (e70070)
Masao Noda, Ryota Koshu, Reiko Tsunoda, Hirofumi Ogihara, Tomohiko Kamo, Makoto Ito, Hiroaki Fushiki. 8725

Remote Photoplethysmography Technology for Blood Pressure and Hemoglobin Level Assessment in the Preoperative Assessment Setting: Algorithm Development Study (e60455)
Selene Tan, Jia Chai, Minwoo Choi, Umair Javaid, Brenda Tan, Belinda Chow, Hairil Abdullah. 8735

mindLAMPVis as a Co-Designed Clinician-Facing Data Visualization Portal to Integrate Clinical Observations From Digital Phenotyping in Schizophrenia: User-Centered Design Process and Pilot Implementation (e70073)
Karthik Sama, Jaya Sreevalsan-Nair, Soumya Choudhary, Srilakshmi Nagendra, Preethi Reddy, Asher Cohen, Urvakhsh Mehta, John Torous. 8 7 5 0

Using Real Electronic Health Records in Undergraduate Education: Roundtable Discussion (e60789)
Fatima Nadeem, Jessica Azmy, Asieh Shomali, Benjamin Diette, Lloyd Gregory, Angela Davies, Kurt Wilson. 8767

XSL•FO
RenderX

Safety Evaluation in Healthy Adults of Motion-Based Virtual Reality Dichoptic Training for Pediatric Patients With Amblyopia: Prospective Intervention Study (e69801)
Masakazu Hirota, Yuichi Okumura, Ken Nagino, Takao Hayashi, Takashi Negishi, Shintaro Nakao, Hitoshi Kawasaki, Takenori Inomata. 8 7 7 6

Feasibility of Collecting and Linking Digital Phenotyping, Clinical, and Genetics Data for Mental Health Research: Pilot Observational Study (e71377)
Joanne Beames, Omar Dabash, Michael Spoelma, Artur Shvetcov, Wu Zheng, Aimy Slade, Jin Han, Leonard Hoon, Joost Kupper, Richard Parker, Brittany Mitchell, Nicholas Martin, Jill Newby, Alexis Whitton, Helen Christensen. 8789

Analyzing Trends in Suicidal Thoughts Among Patients With Psychosis in India: Exploratory Secondary Analysis of Smartphone Ecological Momentary Assessment Data (e67745)
Ameya Bondre, Aashish Ranjan, Ritu Shrivastava, Deepak Tugnawat, Nirmal Chaturvedi, Anant Bhan, Snehil Gupta, Abhijit Rozatkar, Srilakshmi Nagendra, Siddharth Dutt, Soumya Choudhary, Preethi Reddy, Urvakhsh Mehta, John Naslund, John Torous. 8798

The Norwegian Version of the Self-Efficacy in Clinical Performance Scale (SECP): Psychometric Validation Study (e68173)
Camilla Olaussen, Marko Stojiljkovic, Jaroslav Zlamal, Tone Fløło, Andréa Nes. 8814

Increasing the Uptake of Breast and Cervical Cancer Screening Via the MAwar Application: Stakeholder-Driven Web Application Development Study (e65542)
Nurfarhana Nasrudin, Shariff-Ghazali Sazlina, Ai Cheong, Ping Lee, Soo-Hwang Teo, Abdul Aneesa, Chin Teo, Fakhrul Rokhani, Nuzul Haron, Noor Harrun, Bee Ho, Salbiah Mohamed Isa. 8832

Using a Hybrid of AI and Template-Based Method in Automatic Item Generation to Create Multiple-Choice Questions in Medical Education: Hybrid AIG (e65726)
Yavuz Kıyak, Andrzej Kononowicz. 8845

Using Large Language Models to Automate Data Extraction From Surgical Pathology Reports: Retrospective Cohort Study (e64544)
Denise Lee, Akhil Vaid, Kartikeya Menon, Robert Freeman, David Matteson, Michael Marin, Girish Nadkarni. 8855

Moving Standard Deviation of Trunk Acceleration as a Quantification Index for Physical Activities: Validation Study (e63064)
Takuya Suzuki, Yuji Kono, Takayuki Ogasawara, Masahiko Mukaino, Yasushi Aoshima, Shotaro Furuzawa, Yurie Fujita, Hirotaka Matsuura, Masumi Yamaguchi, Shingo Tsukada, Yohei Otaka. 8864

Provider Perspectives on Implementing an Enhanced Digital Screening for Adolescent Depression and Suicidality: Qualitative Study (e67624)
Morgan Coren, Oliver Lindhiem, Abby Angus, Emma Toevs, Ana Radovic. 8874

Reliability of Average Daily Steps Measured Through a Consumer Smartwatch in Parkinson Disease Phenotypes, Stages, and Severities: Cross-Sectional Study (e63153)
Edoardo Bianchini, Domiziana Rinaldi, Lanfranco De Carolis, Silvia Galli, Marika Alborghetti, Clint Hansen, Antonio Suppa, Marco Salvetti, Francesco Pontieri, Nicolas Vuillerme. 8883

Patient and Provider Perspectives of a Web-Based Intervention to Support Symptom Management After Radioactive Iodine Treatment for Differentiated Thyroid Cancer: Qualitative Study (e60588)
Alaina Carr, Angela Jenkins, Jacqueline Jonklaas, Kate Gabriel, Kristen Miller, Kristi Graves. 8894

Customizing Computerized Adaptive Test Stopping Rules for Clinical Settings Using the Negative Affect Subdomain of the NIH Toolbox Emotion Battery: Simulation Study (e60215)
Saki Amagai, Aaron Kaat, Rina Fox, Emily Ho, Sarah Pila, Michael Kallen, Benjamin Schalet, Cindy Nowinski, Richard Gershon. 8910

XSL•FO
RenderX

"Your Life, Your Health: Tips and Information for Health and Well-Being": Development of a World Health Organization Digital Resource to Support Universal Access to Trustworthy Health Information (e57881)
Danielle Muscat, Rachael Hinton, Shyama Kuruvilla, Don Nutbeam. 8920

HUMAn, a Real-Time Evolutive Patient Model for Major Incident Simulation: Development and Validation Study (e66201)
Maxence Laurent, Arnaud Jaccard, Laurent Suppan, Elio Erriquez, Xavier Good, Eric Golay, Dominique Jaccard, Mélanie Suppan. 8930

Exploratory Co-Design on Electronic Health Record Nursing Summaries: Case Study (e68906)
Suhyun Park, Jenna Marquard, Robin Austin, Christie Martin, David Pieczkiewicz, Connie Delaney. 8945

Improving the User Interface and Guiding the Development of Effective Training Material for a Clinical Research Recruitment and Retention Dashboard: Usability Testing Study (e66718)
Leah Gardner, Pezhman Raeisian Parvari, Mark Seidman, Richard Holden, Nicole Fowler, Ben Zarzaur, Diana Summanwar, Cristina Barboi, Malaz Boustani. 8956

Exploring the Relationship Between Smartphone GPS Patterns and Quality of Life in Patients With Advanced Cancer and Their Family Caregivers: Longitudinal Study (e59161)
Kyungmi Lee, Andres Azuero, Sally Engler, Sidharth Kumar, Frank Puga, Alexi Wright, Arif Kamal, Christine Ritchie, George Demiris, Marie Bakitas, J Odom. 8965

eHealth Literacy and Cyberchondria Severity Among Undergraduate Students: Mixed Methods Study (e63449)
Wan -Chen Hsu. 8981

Estimating the Prevalence of Schizophrenia in the General Population of Japan Using an Artificial Neural Network–Based Schizophrenia Classifier: Web-Based Cross-Sectional Survey (e66330)
Pichsinee Choomung, Yupeng He, Masaaki Matsunaga, Kenji Sakuma, Taro Kishi, Yuanying Li, Shinichi Tanihara, Nakao Iwata, Atsuhiko Ota. 8992

Measuring Self-Reported Well-Being of Physicians Using the Well-Being Thermometer: Cohort Study (e54158)
Marios Adamou, Sarah Jones, Niki Kyriakidou, Andrew Mooney, Shriti Pattani, Matthew Roycroft. 9003

Recruitment Challenges and Strategies in a Technology-Based Intervention for Dementia Caregivers: Descriptive Study (e59291)
Eunjung Ko, Ye Gao, Peng Wang, Lahiru Wijayasingha, Kathy Wright, Kristina Gordon, Hongning Wang, John Stankovic, Karen Rose. 9013

Analysis of Social Performance and Action Units During Social Skills Training: Focus Group Study of Adults With Autism Spectrum Disorder and Schizophrenia (e59261)
Hiroki Tanaka, Kana Miyamoto, Jennifer Hamet Bagnou, Elise Prigent, Céline Clavel, Jean-Claude Martin, Satoshi Nakamura. 9025

Unified Mobile App for Streamlining Verbal Autopsy and Cause of Death Assignment in India: Design and Development Study (e59937)
Harleen Kaur, Stuti Tripathi, Manjeet Chalga, Sudhir Benara, Amit Dhiman, Shefali Gupta, Saritha Nair, Geetha Menon, B Gulati, Sandeep Sharma, Saurabh Sharma. 9039

Public Understanding and Expectations of Digital Health Evidence Generation: Focus Group Study (e56523)
Paulina Bondaronek, Jingfeng Li, Henry Potts. 9048

International Expert Consensus on Relevant Health and Functioning Concepts to Assess in Users of Tobacco and Nicotine Products: Delphi Study (e58614)
Vivienne Law, Esther Afolalu, Linda Abetz-Webb, Lee Wemyss, Andrew Turner, Christelle Chrea. 9059

XSL•FO
RenderX

Development of an eHealth Intervention Including Self-Management for Reducing Sedentary Time in the Transition to Retirement: Participatory Design Study (e63567)
Lisa Hultman, Caroline Eklund, Petra von Heideken Wågert, Anne Söderlund, Maria Lindén, Magnus Elfström. 9079

Methods to Adjust for Confounding in Test-Negative Design COVID-19 Effectiveness Studies: Simulation Study (e58981)
Elizabeth Rowley, Patrick Mitchell, Duck-Hye Yang, Ned Lewis, Brian Dixon, Gabriela Vazquez-Benitez, William Fadel, Inih Essien, Allison Naleway, Edward Stenehjem, Toan Ong, Manjusha Gaglani, Karthik Natarajan, Peter Embi, Ryan Wiegand, Ruth Link-Gelles, Mark Tenforde, Bruce Fireman. 9094

Multimodal Pain Recognition in Postoperative Patients: Machine Learning Approach (e67969)
Ajan Subramanian, Rui Cao, Emad Naeini, Seyed Aqajari, Thomas Hughes, Michael-David Calderon, Kai Zheng, Nikil Dutt, Pasi Liljeberg, Sanna Salanterä, Ariana Nelson, Amir Rahmani. 9112

Evaluating Reaction Videos of Young People Watching Edutainment Media (MTV Shuga): Qualitative Observational Study (e55275)
Venetia Baker, Sarah Mulwa, David Khanyile, Georgia Arnold, Simon Cousens, Cherie Cawood, Isolde Birdthistle. 9128

Proficiency, Clarity, and Objectivity of Large Language Models Versus Specialists' Knowledge on COVID-19's Impacts in Pregnancy: Cross-Sectional Pilot Study (e56126)
Nicola Bragazzi, Michèle Buchinger, Hisham Atwan, Ruba Tuma, Francesco Chirico, Lukasz Szarpak, Raymond Farah, Rola Khamisy-Farah. 9 1 4 1

A 4-Site Public Deliberation Project on the Acceptability of Youth Self-Consent in Biomedical HIV Prevention Trials: Assessment of Facilitator Fidelity to Key Principles (e58451)
Claire Draucker, Andrés Carrión, Mary Ott, Ariel Hicks, Amelia Knopf. 9153

Designing a Smartphone-Based Virtual Reality App for Relaxation: Qualitative Crossover Study (e62663)
Amandine Verstegen, Tom Van Daele, Bert Bonroy, Glen Debard, Romy Sels, Marlon van Loo, Sylvie Bernaerts. 9170

Using an Interactive Voice Response Survey to Assess Patient Satisfaction in Ethiopia: Development and Feasibility Study (e67452)
Dessalegn Shamebo, Anagaw Derseh Mebratie, Catherine Arsenault. 9185

Exploring Metadata Catalogs in Health Care Data Ecosystems: Taxonomy Development Study (e63396)
Simon Scheider, Mostafa Mallick. 9198

The Design and Evaluation of a Simulation Tool for Audiology Screening Education: Design Science Approach (e47150)
John Gerdes, Benjamin Schooley, Dakota Sharp, Juliana Miller. 9220

Effects of Missing Data on Heart Rate Variability Measured From A Smartwatch: Exploratory Observational Study (e53645)
Hope Davis-Wilson, Meghan Hegarty-Craver, Pooja Gaur, Matthew Boyce, Jonathan Holt, Edward Preble, Randall Eckhoff, Lei Li, Howard Walls, David Dausch, Dorota Temple. 9235

Development and Implementation of MyPainHub, a Web-Based Resource for People With Musculoskeletal Conditions and Their Health Care Professionals: Mixed Methods Study (e63780)
Kerrie Evans, Jonathan Ko, Dragana Ceprnja, Katherine Maka, Darren Beales, Michele Sterling, Kim Bennell, Gwendolen Jull, Paul Hodges, Marnee McKay, Trudy Rebbeck. 9253

The Feasibility of Large Language Models in Verbal Comprehension Assessment: Mixed Methods Feasibility Study (e68347)
Dorit Hadar-Shoval, Maya Lvovsky, Kfir Asraf, Yoav Shimoni, Zohar Elyoseph. 9269

XSL•FO
RenderX

An Integrated Platform Combining Immersive Virtual Reality and Physiological Sensors for Systematic and Individualized Assessment of Stress Response (bWell): Design and Implementation Study (e64492)
Budhachandra Khundrakpam, Melanie Segado, Jesse Pazdera, Vincent Gagnon Shaigetz, Joshua Granek, Nusrat Choudhury. 9283

Long-Term Experiences of Health Care Providers Using Iris Scanning as an Identification Tool in a Vaccine Trial in the Democratic Republic of the Congo: Qualitative Study (e54921)
Trésor Zola Matuvanga, Antea Paviotti, Freddy Bikioli Bolombo, Gwen Lemey, Ynke Larivière, Maha Salloum, Bernard Isekah Osang'ir, Emmanuel Esanga Longomo, Solange Milolo, Junior Matangila, Vivi Maketa, Patrick Mitashi, Pierre Van Damme, Hypolite Muhindo-Mavoko, Jean-Pierre Van geertruyden. 9301

Assessment of Fraud Deterrence and Detection Procedures Used in a Web-Based Survey Study With Adult Black Cisgender Women: Description of Lessons Learned and Recommendations (e59955)
Amber Sophus, Jason Mitchell. 9313

A Trauma Support App for Young People: Co-design and Usability Study (e57789)
Maria Thell, Kerstin Edvardsson, Reem Aljeshy, Kalid Ibrahim, Georgina Warner. 9325

Kinetic Pattern Recognition in Home-Based Knee Rehabilitation Using Machine Learning Clustering Methods on the Slider Digital Physiotherapy Device: Prospective Observational Study (e69150)
Clement Twumasi, Mikail Aktas, Nicholas Santoni. 9339

Smartwatch-Based Ecological Momentary Assessment for High-Temporal-Density, Longitudinal Measurement of Alcohol Use (AlcoWatch): Feasibility Evaluation (e63184)
Chris Stone, Sally Adams, Robyn Wootton, Andy Skinner. 9357

Using Social Media to Engage and Enroll Underrepresented Populations: Longitudinal Digital Health Research (e68093)
Christiana Harry, Sarah Goodday, Carol Chapman, Emma Karlin, April Damian, Alexa Brooks, Adrien Boch, Nelly Lugo, Rebecca McMillan, Jonell Tempero, Ella Swanson, Shannon Peabody, Diane McKenzie, Stephen Friend. 9373

The Influence of Medical Expertise and Information Search Skills on Medical Information Searching: Comparative Analysis From a Free Data Set (e62754)
Aline Chevalier, Cheyenne Dosso. 9390

Application of an Innovative Methodology to Build Infrastructure for Digital Transformation of Health Systems: Developmental Program Evaluation (e53339)
M Buchan, Tarun Katapally, Jasmin Bhawra. 9406

Ecological Momentary Assessment of Parental Well-Being and Time Use: Mixed Methods Compliance and Feasibility Study (e67451)
Laura Altweck, Samuel Tomczyk. 9423

A Mixed Method Survey of Characteristics of HIV Care Facilities: Medical Monitoring Project Facility Survey Project (e52123)
Dustin Williams, John Weiser, Timothy McManus, Hanna Demeke, Darryl Creel, Jason Craw, Milton Cahoon, Linda Beer. 9456

Co-Designing an Infant Early Childhood Mental Health Mobile App for Early Childhood Education Teachers' Professional Development: Community-Based Participatory Research Approach (e66714)
Ruby Natale, Elizabeth Howe, Carolina Velasquez, Emperatiz Guzman Garcia, Karen Granja, Bianca Caceres, Elizabeth Erban, Tania Ramirez, Jason Jent. 9468

Idiographic Lapse Prediction With State Space Modeling: Algorithm Development and Validation Study (e73265)
Eric Pulick, John Curtin, Yonatan Mintz. 9483

XSL•FO
RenderX

A Mobile App (Joint Effort) to Support Cannabis Use Self-Management and Reinforce the Use of Protective Behavioral Strategies: Development Process and Usability Testing (e71924)
José Côté, Patricia Auger, Gabrielle Chicoine, Jinghui Cheng, Sylvie Cossette, Guillaume Fontaine, Christine Genest, Shalini Lal, Judith Lapierre, M Pagé, Marc-André Maheu-Cadotte, Geneviève Rouleau, Billy Vinette, Didier Jutras-Aswad. 9501

Therapy Companion Mobile App for Acceptance and Commitment Therapy Exercises (ACTaide): Therapist and Client Co-Design Study (e69532)
Serena Thapar, Daniela Quesada, Bärbel Knäuper. 9518

Service Quality Assessment of Digital Health Solutions in Outpatient Care: Qualitative Item Repository Development Study (e68276)
Dominik Rigo, Leonard Fehring, Achim Mortsiefer, Sven Meister. 9539

Predicting Social Frailty in Older Adults Using Fitbit-Derived Circadian and Heart Rate Biomarkers: Cross-Sectional Study (e71393)
Hiroki Maekawa, Yu Kume. 9560

A Digital Tool for Assessing Well-Being at the Workplace and in Personal Life: Development and Validation of the Quan Well-Being Index (e73713)
Georgia Floridou, Freya Katre, Emile Jeuken. 9581

Development of a Web-Based Intervention to Support Primary Health Care Professionals in Digital Health Measurement: User-Centered Participatory Approach (e72331)
Kristel Meijers, Esther Bols, Emmylou Beekman, Rachel Slangen, Peter van de Hoef, Darcy Ummels, Rik Crutzen. 9597

Optimizing the Color Shapes Task for Ambulatory Assessment and Drift Diffusion Modeling: A Factorial Experiment (e66300)
Sharon Kim, Jonathan Hakun, Yanling Li, Karra Harrington, Daniel Elbich, Martin Sliwinski, Joachim Vandekerckhove, Zita Oravecz. 9619

Identification of Syndrome Types in Patients With Pancreatic Cancer From Free Text in Electronic Medical Records: Model Development and Validation (e70602)
He Ba, Haojie Du, Chienshan Cheng, Yuan Zhang, Linjie Ruan, Zhen Chen. 9641

Exploring the Use of Digital Technology to Support Health Behavior Change in Young People Under the Care of Complications of Excess Weight (CCEW) Clinics: Qualitative Patient-Centered Design Study (e64947)
Brioney Gee, Bonnie Teague, Matt Farrar, Victoria Farrar, Dorothy Szinay, Li Chan, Ken Ong, Ben Jackson, Sydney Wylie, Felix Naughton, Jon Wilson, Emma Webb. 9668

Toward Comprehensive Assessment of Beliefs and Attitudes Related to Physical Activity in Young Adults: Pilot Study (e69094)
Theodorus Noordover, Aave Hannus, Kenn Konstabel. 9680

An mHealth App and System Architecture for Respiratory Disease Management: Design Principles, Tool Development, and Pilot Usability Study (e73584)
Andrew Chao, Lisa Martignetti, René Groh, Andreas Kist, Nicole Li-Jessen. 9699

Detecting Laterality Errors in Combined Radiographic Studies by Enhancing the Traditional Approach With GPT-4o: Algorithm Development and Multisite Internal Validation (e76384)
Kung-Hsun Weng, Yi-Chen Chou, Yu-Ting Kuo, Tsyh-Jyi Hsieh, Chung-Feng Liu. 9724

Impact of Detailed Versus Generic Instructions on Fine-Tuned Language Models for Patient Discharge Instructions Generation: Comparative Statistical Analysis (e80917)
Muneerah Alqahtani, Abdullah Al-Barakati, Fahd Alotaibi, Mohammed Al Shibli, Saad Almousa. 9743

XSL•FO
RenderX

Expectant Mothers' Reflections on Childhood and Parenting: Cross-Sectional Study (e80060)
Bente Mjølstad, Gunnhild Vie, Linn Getz, Mari Hansen, Hanna Mørkved, Bjarne Austad. 9762

Co-Designing a Digital Solution for Decreasing Loneliness and Social Isolation Among Older People in Sweden: Explorative Study (e78213)
Caroline Eklund, Viktoria Zander, Annelie Gusdal, Charlotta Åkerlind, Petra von Heideken Wågert. 9772

Comparing Patient Simulation With a Humanoid Robot or a Human Actor in Terms of Training Success and Acceptance: Pilot Questionnaire Study (e70363)
Patricia Schwarz, Sandra Hellmers, Sebastian Spanknebel, Diana Immel, Rene Hurlemann, Andreas Hein. 9793

Patterns of Prior Induced Abortions and the Likelihood of Subsequent Natural Pregnancy Loss: Exploratory Application of Pregnancy Outcome Sequencing (e78489)
Francois Domagni, Daniel Tsang, James Studnicki, John Fisher, David Reardon, Christopher Craver. 9804

Large Language Model Evaluation in Traditional Chinese Medicine for Stroke: Quantitative Benchmarking Study (e81545)
Hulin Long, Yang Deng, Yaoguang Guo, Zifan Shen, Yuzhu Zhang, Ji Bao, Yang He. 9820

Evaluating the Accuracy of Medical Information Generated by ChatGPT and Gemini and Its Alignment With International Clinical Guidelines From the Surviving Sepsis Campaign: Comparative Study (e84251)
Dina Kutbi, Ehab Abou-Bakr, Hassan Haidar. 9830

Feasibility and Usability of an mHealth App (mLab+) to Guide Users Through HIV and Syphilis Self-Testing: Pilot Randomized Controlled Trial (e72955)
Maeve Brin, Thomas Scherr, Janejira Chaiyasit, Jianfang Liu, Maura Abbott, Robert Garofalo, Lisa Kuhns, Tess Sky, Ian Esliker, Rebecca Schnall. 9837

A Novel Primary Care Planning Informatics Tool Informed by Data-Driven Multimorbidity Grouping: User-Centered Design and Feasibility Testing (e75081)
Linnaea Schuttner, Karin Daniels, Srinidhi Venkatesh, Franya Hutchins, Karley Atchison, Rebecca Piegari, Timothy Bober, Xinhua Zhao, Ann-Marie Rosland. 9851

Mobile-Based Ecological Momentary Intervention for Improving Physical Activity in Adults Without Regular Physical Activity: Pilot Randomized Controlled Trial (e79360)
Takeyuki Oba, Chihiro Moriishi, Keisuke Takano, Kentaro Katahira, Kenta Kimura. 9870

Feasibility of the Social Media–Based Prevention Program "Leduin" for German Adolescents on Instagram: Mixed Methods Pilot Study (e78774)
Elizabeth Zimmermann, Samuel Tomczyk. 9887

Feasibility and Acceptability of a mHealth Patient Navigation Intervention to Increase Pre-Exposure Prophylaxis Uptake in Racially and Ethnically Diverse Sexual and Gender Minority Youth in Los Angeles (PrEPresent): Pilot Randomized Controlled Trial (e69255)
Sam Calvetti, Bryan Lei, Jacob Stocks, Matthew Rosso, Manuel Puentes, Ramon Durazo-Arvizu, Lindsay Slay, Michele Kipke, Lisa Hightow-Weidman. 9910

Sleep and Activity Patterns as Transdiagnostic Behavioral Biomarkers in Psychiatry: Longitudinal Observational Study From the DeeP-DD Study (e81107)
Dylan Hamitouche, Tihare Zamorano, Youcef Barkat, Deven Parekh, Lena Palaniyappan, Sara Jalali, David Benrimoh. 9929

Formative Evaluation of an HIV Prevention App tailored for Latino Men Who Have Sex With Men: Acceptability and Usability Study (e74208)
Valeria Cantos, Humberto Posada-Orozco, Isabella Batina, Natalie Sanchez, Eric Rangel, Patrick Sullivan, Andres Camacho-Gonzalez, Aaron Siegler. 9943

XSL•FO
RenderX

Behavioral Change Intervention to Promote a Healthier Postpartum Lifestyle: Mixed Methods Pilot Study (e69391)
Pernille Christiansen, Trine Kjær, Christina Vinter, Mette Rothmann, Mette Skjøth, Eva Draborg. 9954

Thematic Analysis of Smoking Cessation and Future Cessation Interventions for Cancer Survivors: Convenience Sampling Study (e76792)
Kinsey Pebley, Tianna Williams, Kathryn Moody, Alana Rojewski. 9970

Eye and Systemic Disease Management Changes After Teleophthalmology Screening in Primary Care: Retrospective Cross-Sectional Pilot Study of 200 Consecutive Patients (e81918)
Alexander Pinhas, Boris Pinhas, Egor Dmitruk, Stella Pinhas. 9977

A Robot-Delivered Training Program to Improve Children's Mental Health and Resilience in Dutch Primary Schools: Pilot Intervention Study (e66797)
Anne Zijp, Jiska Aardoom, Olivier Blanson Henkemans, Sylvia van der Pal, Eline Vlasblom, Anke Versluis. 9993

Feasibility, Acceptability, and Preliminary Outcomes of a Mobile Adaptation of a Relational Savoring Intervention to Prevent Loneliness in College Students: Mixed Methods Pilot Study (e70528)
Brenda Nguyen, Jocelyn Lai, Hana Qureshi, Christopher Marcotullio, Sina Labbaf, Yuning Wang, Salar Jafarlou, Nikil Dutt, Amir Rahmani, Jessica Borelli. 10004

Previsit Preparation for Shared Decision-Making in Lung Cancer Screening in Primary Care Using a Paper Decision Aid and an Automated Text Messaging Program: Quasi-Experimental Pilot Study (e69044)
Mayuko Ito Fukunaga, Renda Wiener, Shaun Toomey, Joann Wagner, Qiming Shi, Kavitha Balakrishnan, Alexandra Nguyen, Dan Nguyen, M McKee, Alexander Bankier, Rajani Sadasivam, Sybil Crawford, Paul Han, Thomas Houston, Kathleen Mazor. 10020

Implementation, Challenges, and Outlook of an Intergenerational, Layperson-led, Health Coaching Program (HealthStart): A Pilot Case Study (e76592)
Xiaoting Huang, Ka Yow, Audrey Kwan, Jin Yeo, Haikel Lim, Jie Lim, Meng Lim, Lynn Teo, Nerice Ngiam, Si Lim, Kharuna Jaichandra, Kai Tang, Angeline Tey, Lian Low, Kennedy Ng. 10036

The Development of a Patient-Centered Digital Health Care Technology for Young Adults in Opioid Use Disorder Treatment: Qualitative Study (e67401)
Karen Alexander, Madison Scialanca. 10049

Feasibility of a Randomized Controlled Trial of Large AI-Based Linguistic Models for Clinical Reasoning Training of Physical Therapy Students: Pilot Randomized Parallel-Group Study (e66126)
Raúl Ferrer-Peña, Silvia Di-Bonaventura, Alberto Pérez-González, Alfredo Lerín-Calvo. 10059

Just-in-Time Delivery of Cognitive Behavioral Therapy–Based Exercises: Single-Case Experimental Design With Random Multiple Baselines (e69556)
Takeyuki Oba, Keisuke Takano, Daichi Sugawara, Kenta Kimura. 10070

Environmental Impact of Online Versus in-Person Critical Care Education Through the Carbon Footprint Analysis of the CERTAIN Program: Cross-Sectional Study (e63524)
Baiyong Wang, Claudia Zambrano, Nasrin Nikravangolsefid, Ricardo Carvalhais, Alexander Niven, Ognjen Gajic, Yue Dong. 10085

Feasibility of a Digital Coaching Program for Improving Mental Well-Being and Emotional Intelligence: Pragmatic Retrospective Cohort Study (e71828)
Nicholas Peiper, Adam Pettitt, Bela Shah, Oliver Attwood, Ella Sivonen, Jeffrey Pfeffer. 10097

Examining Nonverbal Communication in Dyadic Interactions With Virtual Humans Using an Integrated Coding System: Mixed Methods Analysis (e59328)
Analay Perez, Rae Sakakibara, Srikar Baireddy, Michael Fetters, Timothy Guetterman. 10107

XSL•FO
RenderX

Evaluating a Customized Version of ChatGPT for Systematic Review Data Extraction in Health Research: Development and Usability Study (e68666)
Jayden Sercombe, Zachary Bryant, Jack Wilson. 10121

Acceptability and Feasibility of a Prototype Regional Disaster Teleconsultation System for COVID-19 Pandemic Response: Pilot Field Test (e73078)
Stephanie Ludy, Mari-Lynn Drainoni, Lauren Schmidt, Mark Litvak, Julianne Dugas, Eric Goralnick, Paul Biddinger, Tehnaz Boyle. 10129

Piloting the Extension for Community Healthcare Outcomes (ECHO) Pediatric Oncology Telehealth Education Program in Western Kenya: Implementation Study (e59776)
Tyler Severance, Gilbert Olbara, Festus Njuguna, Martha Kipng'etich, Sandra Lang'at, Maureen Kugo, Jesse Lemmen, Marjorie Treff, Patrick Loehrer, Terry Vik. 10145

Community Health Worker Diabetes Prevention Awareness Training in an Immersive Virtual World Environment: Mixed Methods Pilot Study (e64051)
Laurie Ruggiero, Lauretta Quinn, Amparo Castillo, Colleen Monahan, Leticia Boughton Price, Wandy Hernandez. 10154

Smartphone Ecological Momentary Assessment and Wearable Activity Tracking in Pediatric Depression: Cohort Study (e66187)
Jimena Unzueta Saavedra, Emma Deaso, Margot Austin, Laura Cadavid, Rachel Kraff, Emma Knowles. 10168

Impacts of the Mindfulness Meditation Mobile App Calm on Undergraduate Students' Sleep and Emotional State: Pilot Randomized Controlled Trial (e66131)
Tovan Lew, Natnaiel Dubale, Erik Doose, Alex Adenuga, Holly Bates, Sarah West. 10188

Using Ecological Momentary Assessment to Assess Family Functioning in Spanish-Speaking Parent and Adolescent Dyads: Daily Questionnaire Study (e60073)
Alejandra Fernandez, Savannah Bernal, Lana Kim, Subodh Potla. 10203

Reactivity to Smoking Cues in a Social Context: Virtual Reality Experiment (e71285)
Katharina Eidenmueller, Sabine Hoffmann, Kornelius Kammler-Sücker, Leonard Wenger, Massimiliano Mazza, Christiane Mühle, Manuel Stenger, Gerrit Meixner, Falk Kiefer, Bernd Lenz. 10213

Feasibility and Cultural Adaptation of a Community-Engaged Physical Activity Intervention for Hispanic Older Adults: Pilot Study (e65489)
Zvinka Zlatar, Mikael Greenwood-Hickman, Lazaro Lujan, Julie Cooper, Stefani Florez-Acevedo, David Marquez, Rosa Aceves, Andrea Vargas, Dori Rosenberg. 10228

Web-Based Stress Management for Working Adults With Attention-Deficit/Hyperactivity Disorder (ADHD): Single-Arm, Open Pilot Trial (e66388)
Martin Oscarsson, Sandra Hammarbäck, Karolina Blom Wiberg, Alexander Rozental, Ylva Ginsberg, Per Carlbring, Gerhard Andersson, Fredrik Jönsson. 10244

Methods of Piloting an Abstraction Tool to Describe Family Engagement in the Hospital Setting: Retrospective Chart Review (e66549)
Jennifer Morgan, Jennifer Cahill, Christine Ritchie, Lingling Zhang, Priscilla Gazarian. 10260

A Portal-Based Intervention (PATTERN) Designed to Support Medication Use Among Older Adults: Feasibility and Acceptability Study (e71676)
Allison Pack, Stacy Bailey, Rachel O'Conor, Evelyn Velazquez, Guisselle Wismer, Fangyu Yeh, Laura Curtis, Kenya Alcantara, Michael Wolf. . 10268

Cardiac Self-Efficacy Improvement in a Digital Heart Health Program: Secondary Analysis From a Feasibility and Acceptability Pilot Study (e60676)
Kimberly Lockwood, Priya Kulkarni, OraLee Branch, Sarah Graham. 10276

XSL•FO
RenderX

The Healing Through Ongoing Psychological Empowerment Telehealth Intervention With Two Spirit, Transgender, and Nonbinary Clients of Color in the United States: Open Clinical Trial Feasibility and Implementation Analysis (e64477)
Stephanie Budge, Elliot Tebbe, Joonwoo Lee, Sergio Domínguez Jr, Em Matsuno, Louis Lindley. 10288

Impact of a “Digital Health” Curriculum on Students’ Perception About Competence and Relevance of Digital Health Topics for Future Professional Challenges: Prospective Pilot Study (e58940)
Juliane Kröplin, Leonie Maier, Jan-Hendrik Lenz, Bernd Romeike. 10297

Effects of Using a Digital Peer-Supported App on Glycated Hemoglobin Changes Among Patients With Type 2 Diabetes: Prospective Single-Arm Pilot Study (e72659)
Shota Yoshihara, Kayoko Takahashi, Hiroyuki Kawaguchi, Nozomi Harai, Kyoichiro Tsuchiya. 10310

Patient Acceptability and Technical Reliability of Wearable Devices Used for Monitoring People With Parkinson Disease: Survey Study (e63704)
Tasmin Rookes, Amit Batla, Megan Armstrong, Gareth Ambler, Kate Walters, Anette Schrag. 10319

Using Text Messaging Ecological Momentary Assessment to Record Changes in e-Cigarette and Combustible Cigarette Use: Pilot Randomized Clinical Trial (e66709)
Tucker Morgan, Michelle He, Andrew Nicholson, Omar El Shahawy, Scott Sherman, Elizabeth Stevens. 10327

Development and Implementation of Strong Foundations, a Digitally Delivered Fall Prevention Program: Usability and Feasibility Pilot Exercise Cohort Study (e67406)
Ryan Moran, David Wing, Hope Davey, Hava Barkai, Jeanne Nichols. 10342

Evaluating the Knowledge Level, Practice, and Behavioral Change Potential of Care Managers in Pressure Injury Prevention Using a Mobile App Prototyping Model in the Home-Care Setting: Single-Arm, Pre-Post Pilot Study (e57768)
Masushi Kohta, Mayumi Takahashi, Hiroe Koyanagi, Junko Sugama. 10352

Assessing Dynamic Cognitive Function in the Daily Lives of Youths With and Without Type 1 Diabetes: Usability Study (e60275)
Mary Ray, Jorie Fleming, Andrew Aschenbrenner, Jason Hassenstab, Brooke Redwine, Carissa Burns, Ana Arbelaez, Mary Vajravelu, Tamara Hershey. 10364

Use of a Technology-Based Fall Prevention Program With Visual Feedback in the Setting of Early Geriatric Rehabilitation: Controlled and Nonrandomized Study (e66692)
Carolin Steinmetz, Christina Stenzel, Maj Sylvester, Denis Glage, Anne Linke, Monika Sadlonova, Christine von Arnim, Marlena Schnieder, Miroslava Valentová, Stephanie Heinemann. 10380

Acceptance of Unsupervised App-Based Cognitive Assessment in Outpatient Care: An Implementation Study (e62706)
Iris Blotenberg, Melanie Boekholt, Nils Lieberknecht, Paula Säring, Jochen Thyrian. 10396

Users’ Perspectives of Direct-to-Consumer Telemedicine Services: Survey Study (e68619)
Kate Churruca, Darran Foo, Andrew Turner, Emily Crameri, Maree Saba, Samantha Spanos, Matthew Vickers, Jeffrey Braithwaite, Louise Ellis. 10405

Feasibility, User Acceptance, and Outcomes of Using a Cancer Prehabilitation App for Exercise: Pilot Cohort Study (e64427)
Fuquan Zhang, Deepali Bang, Christine Visperas, Mon Tun, San Tay. 10415

XSL•FO
RenderX

Longitudinal Changes in Pitch-Related Acoustic Characteristics of the Voice Throughout the Menstrual Cycle: Observational Study (e65448)
Jaycee Kaufman, Jouhyun Jeon, Jessica Oreskovic, Anirudh Thommandram, Yan Fossat. 10429

Postpartum Remote Health Coaching Intervention for Individuals With a Hypertensive Disorder of Pregnancy: Proof-of-Concept Study (e65611)
Jaclyn Borrowman, Lucas Carr, Gary Pierce, William Story, Bethany Gibbs, Kara Whitaker. 10440

Testing the Recruitment Frequency, Implementation Fidelity, and Feasibility of Outcomes of the Heart Failure Activity Coach Study (HEALTHY): Pilot Randomized Controlled Trial (e62910)
Andreas Blomqvist, Maria Bäck, Leonie Klompstra, Anna Strömberg, Tiny Jaarsma. 10453

Effectiveness of a Videoconference-Based Cognitive Behavioral Therapy Program for Patients with Schizophrenia: Pilot Randomized Controlled Trial (e59540)
Masayuki Katsushima, Hideki Nakamura, Yuki Shiko, Hideki Hanaoka, Eiji Shimizu. 10469

Testing a Web-Based Interactive Comic Tool to Decrease Obesity Risk Among Racial and Ethnic Minority Preadolescents: Randomized Controlled Trial (e58460)
May Leung, Katrina Mateo, Marlo Dublin, Laura Harrison, Sandra Verdaguer, Katarzyna Wyka. 10486

Exploring Web-Based Support for Suicidal Ideation in the Scottish Population: Usability Study (e55932)
Heather McClelland, Rory O'Connor, Laura Gibson, Donald MacIntyre. 10505

Engagement With an Internet-Administered, Guided, Low-Intensity Cognitive Behavioral Therapy Intervention for Parents of Children Treated for Cancer: Analysis of Log-Data From the ENGAGE Feasibility Trial (e67171)
Christina Reuther, Louise von Essen, Mudassir Mustafa, Markus Saarijärvi, Joanne Woodford. 10518

Women Empowered to Connect With Addiction Resources and Engage in Evidence-Based Treatment (WE-CARE)—an mHealth Application for the Universal Screening of Alcohol, Substance Use, Depression, and Anxiety: Usability and Feasibility Study (e62915)
Krystyna Isaacs, Autumn Shifflett, Kajal Patel, Lacey Karpisek, Yi Cui, Maayan Lawental, Golfo Tzilos Wernette, Brian Borsari, Katie Chang, Tony Ma. 10538

Digital Solution to Support Medication Adherence and Self-Management in Patients with Cancer (SAMSON): Pilot Randomized Controlled Trial (e65302)
Thu Dang, Nilmini Wickramasinghe, Prem Jayaraman, Kate Burbury, Marliese Alexander, Ashley Whitechurch, Mitchell Dyer, Stephen Quinn, Abdur Forkan, Penelope Schofield. 10553

Assessing the Feasibility and Utility of Patient-Specific 3D Advanced Visualization Modeling in Cerebrovascular Disease: Retrospective Analysis and Prospective Survey Pilot Study (e51939)
Korak Sarkar, Vishal Bhimarasetty, Abdul Rahim, Colin Curtis, Kimberly Hughes. 10566

Effect on Response Rates of Adding a QR Code to Patient Consent Forms for Qualitative Research in Patients With Cancer: Pilot Randomized Controlled Trial (e66681)
Rebecca Wyse, Erin Forbes, Grace Norton, Priscilla Viana Da Silva, Kristy Fakes, Sally Johnston, Stephen Smith, Alison Zucca. 10579

Testing and Iterative Improvement of the CEN ISO/TS 82304-2 Health App Quality Assessment: Pilot Interrater Reliability Study (e64565)
Anna-Lena Frey, Diana Matei, Ben Phillips, Adam McCabe, Rachel Fuller, Begoña Laibarra, Laura Alonso, Victor de la Hoz, Carme Pratdepadua Bufill, Berta Llebot Casajuana, Giuseppe D'Avenio, Pier Sottile, Laura Rocchi, Matteo Errera, Yasmine Laaissaoui, Michael Cardinal, Menno Kok, Petra Hoogendoorn. 10586

Digital Translation Platform (Translatly) to Overcome Communication Barriers in Clinical Care: Pilot Study (e63095)
Victor Olsavszky, Mutaz Bazari, Taieb Dai, Ana Olsavszky, Fabian Finkelmeier, Mireen Friedrich-Rust, Stefan Zeuzem, Eva Herrmann, Jan Leipe, Florian Michael, Hans Westernhagen, Olivier Ballo. 10604

XSL•FO
RenderX

Multilevel Intervention to Increase Patient Portal Use in Adults With Type 2 Diabetes Who Access Health Care at Community Health Centers: Single Arm, Pre-Post Pilot Study (e67293)
Robin Whittemore, Sangchoon Jeon, Samuel Akyirem, Helen Chen, Joanna Lipson, Maritza Minchala, Julie Wagner. 10618

Patients' and Health Care Professionals' Expectations of Virtual Therapeutic Agents in Outpatient Aftercare: Qualitative Survey Study (e59527)
Diana Immel, Bernhard Hilpert, Patricia Schwarz, Andreas Hein, Patrick Gebhard, Simon Barton, René Hurlemann. 10634

AI-Enhanced Virtual Reality Self-Talk for Psychological Counseling: Formative Qualitative Study (e67782)
Moreah Zisquit, Alon Shoa, Ramon Oliva, Stav Perry, Bernhard Spanlang, Anat Brunstein Klomek, Mel Slater, Doron Friedman. 10654

Factors Impacting Mobile Health Adoption for Depression Care and Support by Adolescent Mothers in Nigeria: Preliminary Focus Group Study (e42406)
Lola Kola, Tobi Fatodu, Manasseh Kola, Bisola Olayemi, Adeyinka Adefolarin, Simpa Dania, Manasi Kumar, Dror Ben-Zeev. 10667

An Internet-Based and Mobile Family Management Intervention for Mothers of Very Preterm Infants Hospitalized in the Neonatal Intensive Care Unit (the Preemie Progress Program): Pilot Randomized Controlled Trial (e66073)
Ashley Weber, Tamilyn Bakas, Qutaibah Oudat, Nehal Parikh, Joshua Lambert, Heather Tubbs-Cooley, Jared Rice, Kristin Voos, Matthew Rota, Heather Kaplan. 10676

Detection of Depressive Symptoms in College Students Using Multimodal Passive Sensing Data and Light Gradient Boosting Machine: Longitudinal Pilot Study (e67964)
Jessica Borelli, Yuning Wang, Frances Li, Lyric Russo, Marta Tironi, Ken Yamashita, Elayne Zhou, Jocelyn Lai, Brenda Nguyen, Iman Azimi, Christopher Marcotullio, Sina Labbaf, Salar Jafarlou, Nikil Dutt, Amir Rahmani. 10698

Participatory Intervention Development of a Peer-Guided Self-Help App for Anxiety Disorders: Mixed Methods Study (e62781)
Laura Duddeck, Timo Stolz, Christian Zottl, Thomas Berger, Johanna Boettcher. 10715

Assessing the Impact of Home Environmental Exposures on Allergic Rhinitis Using Real-Time Air Quality Monitoring and Symptom Assessment: Observational Feasibility Study (e73215)
Aero Cavalier, Anthony Dick, Vickie Johnson II, Emily Cramer, Kamal Eldeirawi, Jayant Pinto, Sharmilee Nyenhuis, Victoria Lee. 10735

Evaluation of a Novel eHealth Tool for Pulmonary Rehabilitation in People With Chronic Obstructive Pulmonary Disease: Randomized Controlled Pilot and Feasibility Trial (e68195)
Åsa Karlsson, Pernilla Sönnerfors, Sara Lundell, Annika Toots, Karin Wadell. 10746

Evaluating the Impact on Pain Perceptions, Pain Intensity, and Physical Activity of a Mobile App to Empower Employees With Musculoskeletal Pain: Mixed Methods Pilot Study (e67886)
Stijn Keyaerts, Maxwell Szymanski, Lode Godderis, Vero Vanden Abeele, Liesbeth Daenen. 10765

A Conversational Agent (PracticePal) to Support the Delivery of a Brief Behavioral Activation Treatment for Depression in Rural India: Development and Pilot-Testing Study (e73563)
Ravindra Agrawal, Kimberley Monteiro, Nityasri Narasimhamurti, Shreya Sharma, Amruta Suryawanshi, Aman Bariya, Shravani Narvekar, Lilianna Bagnoli, Mohit Saxena, Lauren Magoun, Shradha Parsekar, Julia Pozuelo, Neal Lesh, Mohit Sood, Tanushri Sharma, Harshita Yadav, Anant Bhan, Abhijit Nadkarni, Vikram Patel. 10786

Machine Learning–Enhanced Surveillance for Surgical Site Infections in Patients Undergoing Colon Surgery: Model Development and Evaluation Study (e75121)
Ugur Celik, Feifan Liu, Kimiyoshi Kobayashi, Richard Ellison III, Yurima Guilarte-Walker, Deborah Mack, Qiming Shi, Adrian Zai. 10801

XSL•FO
RenderX

Applying the Human-Centered Innovation Biodesign Framework to the Development and Piloting of a Program to Mitigate Risk for Cognitive Decline Among Historically Underrepresented Individuals: Case Study (e64930)
Rebecca Lassell, Ada Metaxas, Katherine Wang, Sara Hantgan, Prabhat Gottipati, Sarah Zwerling, Triana Pena, Chava Pollak, Laura Gitlin, Sunit Jariwala. 10813

Promoting Walking in Cardiopulmonary Disease With Mindful Steps: Pilot Feasibility Randomized Controlled Trial of a Web-Based, Pedometer-Mediated Mind-Body Intervention (e74118)
Kristen Kraemer, Daniel Litrownik, Peter Wayne, Caroline Richardson, Neha Bhomia, Reema Kadri, Pamela Rist, Long Ngo, Marilyn Moy, Gloria Yeh. 10829

Evaluation of a Novel Goals-of-Care Discussion Priming Tool (MyCare) in Inpatient General Internal Medicine Ward Settings: Feasibility, Acceptability, and Usability Study (e66932)
Maren Kimura, Sydney Ruller, Alan Forster, James Downar, Sarina Isenberg, Leah Steinberg, Kumanan Wilson, Victoria Shaffer, Pete Wegier, Jeff Myers, Russell Goldman, Carolyn Steele Gray, Daniel Kobewka. 10845

Detecting Perceived Unfair Treatment Among US College Students Using Mobile Sensing: Pilot Machine Learning Study (e78657)
Yiyi Ren, Raghu Mulukutla, Jennifer Mankoff, Anind Dey. 10856

User Experiences of Limb-Worn Wearable Devices for Monitoring Parkinson Disease Motor Function and Blood Pressure: Usability Study (e73423)
Lorna Kenny, Marco Sica, Colum Crowe, Lauren O'Mahony, Brendan O'Flynn, David Mueller, Salvatore Tedesco, John Barton, Suzanne Timmons. 10879

Evaluating Causal and Noncausal Text Messages to Promote Physical Activity in Adults: Randomized Pilot Study (e80090)
Elena Korshakova, Samantha Kleinberg. 10893

The Validity of Generative Artificial Intelligence in Evaluating Medical Students in Objective Structured Clinical Examination: Experimental Study (e79465)
Masashi Yokose, Takanobu Hirosawa, Tetsu Sakamoto, Ren Kawamura, Yudai Suzuki, Yukinori Harada, Taro Shimizu. 10905

Inhibitory Control, Attentional Bias, and Palatable Food Consumption in Adolescents: A Laboratory Feasibility Randomized Controlled Trial (e77579)
Carolina Bejarano, Adrian Ortega, Christopher Cushing. 10913

Integrating a Brief Behavioral Intervention Into Case Management for Mothers With Perinatal Substance Use Disorder: Nonrandomized Pilot Feasibility Study (e75132)
Amritha Bhat, Mary Curran, Mary Lohr, Haley Smith, Brittany Blanchard, Susan Stoner, Therese Grant, Nancy Grote. 10927

Evaluation of the 2020 American Urological Association Microscopic Hematuria Guidelines in Clinical Practice: Retrospective Chart Review Analysis (e75929)
Dominique Munroe, James O'Keefe, Danyang Wang, Miranda Moore. 10938

“The Women’s PrEP Project,” a Clinic-Based, Socio-Structural Intervention to Improve the Provision of Preexposure Prophylaxis for Cisgender Women: Interrupted Time Series Pilot (e80653)
Rachel Scott, Shawnika Hull, Deanna Kerrigan, Yan Wang, Mandi Pratt-Chapman, Tara Mathias-Prabhu, Naquia Unwala, Marisa Sadauskas, Bat-zion Hose, Marjanna Smith, Patricia Moriarty, Tranessa Hanson, Ariam Tedla, Pamela Lotke, Peggy Ye, Hannah Arem. 10947

The Effectiveness of Live and Prerecorded Video Demonstrations in Teaching Restorative Dentistry to Undergraduate Students: Cohort Study (e74383)
Rana Alkattan, Lulwah Alreshaid. 10964

XSL•FO
RenderX

Changing Trends in Suicidal Ideation and Its Influencing Factors During the Transition From Quarantine to Post-Quarantine Among Chinese University Students During the COVID-19 Surge: Six-Wave Panel Study (e74370)
Lijing Li, Tingzhong Yang, Sihui Peng, Randall Cottrell. 10977

Evaluation of Generative Artificial Intelligence Implementation Impacts in Social and Health Care Language Translation: Mixed Methods Case Study (e73658)
Miia Martikainen, Kari Smolander, Johan Sanmark, Enni Sanmark. 10989

Perceptions of Stigma Among Patients With Hepatitis B in Germany: Cross-Sectional Survey (e66379)
Bettina Zimmermann, Theresa Willem, Michael Rost, Nina Matthes, Alena Buyx. 10998

Multimodal Noninvasive Assessment of C-Reactive Protein for Systemic Inflammation in Adults: Cross-Sectional Study (e77108)
Jinjoo Shim, Sinziana Muraru, Vanessa Löw, Caroline Evers, Sandro Riedo, Rucsandra Dobrota, Elgar Fleisch, Oliver Distler, Filipe Barata. . 11 11 0 11 0

Development and Testing of an Exergaming and Artmaking Program to Enhance Cognitive and Physical Function in Older Adults: Results of a Community-Based Pilot Study (e73555)
Nikhil Satchidanand, Mia Lawrence, Sameer Jhaveri, Sara Opalka-Satchidanand. 11024

Stepped Transition to Employment and Postsecondary Education Success (STEPS) for Adolescents and Adults with Autism: Community Implementation Pilot Trial (e70137)
Susan White, Alexis Brewe, Nicole Powell. 11038

Feasibility and Acceptability of a Psychological Intervention for Internalized Health-Related Stigma Among Adults With Chronic Health Conditions: Preliminary Investigation (e69548)
Rebecca Pearl, Danielle Saunders, Laurie Groshon, Yulin Li, Abigail Shonrock, Rebecca Puhl, Kimberly Driscoll, Preeti Manavalan, Joel Gelfand, Thomas Wadden, Sarah Westen, Marjorie Montanez-Wiscovich, Xiang-Yang Lou. 11054

Prevalence of Multiple Chronic Conditions Among Adults in the All of Us Research Program: Exploratory Analysis (e69138)
Xintong Li, Caitlin Dreisbach, Carolina Gustafson, Komal Murali, Theresa Koleck. 11066

Impact of Acute Respiratory Infections on Medical Absenteeism Among Military Personnel: Retrospective Cohort Study (e69113)
Premikha M, Jit Goh, Jing Ng, Adeliza Mutalib, Huai Lim. 11075

The Monthly Cycling of Food Insecurity in Latinas at Risk for Diabetes: Methods, Retention, and Sample Characteristics for a Microlongitudinal Design (e66970)
Angela Bermúdez-Millán, Rafael Pérez-Escamilla, Sofia Segura-Pérez, James Grady, Richard Feinn VI, Hanako Agresta, Dean Kim, Julie Wagner. 11084

Improving Health and Well-Being of People With Post–COVID-19 Consequences in South Africa: Situation Analysis and Pilot Intervention Design (e58436)
Nicole Glover, Farzana Sathar, Pride Mokome, Nkululeko Mathabela, Sipokazi Taleni, Sarah van Blydenstein, Anna-Maria Mekota, Salome Charalambous, Andrea Rachow, Olena Ivanova. 11100

Designing and Evaluating a Prototype of a Trilingual Data Collection Tool for the Middle East and North Africa (MENA) Region to Collect Data About Violence Against Sex Workers: Multiple Methods Approach in User-Centered Design (e65210)
Melissa Ditmore, Amal El Karouaoui, Jose Florez-Arango. 11112

XSL•FO
RenderX

Relationship of Hair Cortisol Concentration With Perceived and Somatic Stress Indices: Cross-Sectional Pilot Study (e63811)
Sharon Bergquist, Danyang Wang, Brad Pearce, Alicia Smith, Allison Hankus, David Roberts, Miranda Moore. 11124

Perceptions and Experiences of Developing Prediabetes Awareness and Participating in the National Diabetes Prevention Program Among Hispanic US Adults: Qualitative Study (e66964)
Kevin Joiner, Mackenzie Adams, Edith Minehan, Matthew O'Brien. 11137

Disease Awareness in Patients With Type 2 Diabetes: Analysis of Baseline Data From the SMART-Finder Observational Study (e60246)
Christian Mueller, Thomas Neusser, Inga Thate-Waschke, Julia Nowicki, Tomasz Plominski, Regine Griesinger, Stefanie Kessner, Stephan Martin. 11147

A Very Low–Carbohydrate Program in Adults With Metabolic Dysfunction–Associated Steatotic Liver Disease and Phospholipase Domain–Containing Protein 3 Risk Genotype: Pre-Post Intervention Study (e60051)
Laura Saslow, Jamie Krinock, Alison O'Brien, Kaitlyn Raymond, Hovig Bayandorian, Judith Moskowitz, Jennifer Daubenmier, Antonino Oliveri, Deanna Marriott, Dina Griauzde, Elizabeth Speliotes. 11155

Continuous Glucose Monitoring Among Patients With Type 1 Diabetes in Rwanda (CAPT1D) Phase I: Prospective Observational Feasibility Study (e64585)
Jason Baker, Giacomo Cappon, Jean Habineza, Corey Basch, Steven Mey, Diana Malkin-Washeim, Christian Schuetz, Niyonsenga Simon Pierre, Etienne Uwingabire, Alvera Mukamazimpaka, Paul Mbonyi, Sandhya Narayanan. 11171

Developing Assessments for Key Stakeholders in Pediatric Congenital Heart Disease: Qualitative Pilot Study to Inform Designing of a Medical Education Toy (e63818)
Neda Barbazi, Ji Shin, Gurumurthy Hiremath, Carlye Lauff. 11179

Effect of the Reassured Self-Compassion–Based School Program on Anxiety, Video Game Addiction, and Body Image Among Rural Female Adolescents: Retrospective Study (e68840)
Areeg Zuair. 11193

Machine Learning–Based Prediction of Delirium and Risk Factor Identification in Intensive Care Unit Patients With Burns: Retrospective Observational Study (e65190)
Ryo Esumi, Hiroki Funao, Eiji Kawamoto, Ryota Sakamoto, Asami Ito-Masui, Fumito Okuno, Toru Shinkai, Atsuya Hane, Kaoru Ikejiri, Yuichi Akama, Arong Gaowa, Eun Park, Ryo Momosaki, Ryuji Kaku, Motomu Shimaoka. 11205

Race and Ethnicity in Facebook Images and Text: Thematic Analysis (e62713)
Shaniece Criss, Sarah Gonzales, Heran Mane, Katrina Makres, Dalmondeh Nayreau, Vaishnavi Bharadwaj, Hannah Kim, Thu Nguyen. 11225

A Sexual Health Self-Management Intervention (Psychosexual Educational Partners Program) for Couples With a History of Breast and Gynecological Cancer: Mixed Methods Feasibility Study (e75743)
Noël Arring, Jennifer Reese, Carolyn Lafferty, Debra Barton, Jeanne Carter. 11238

Developing a Community of Practice to Provide Care Coordination and Address Health-Related Social Needs for Veterans Receiving Care in Community-Based Settings: Program Development and Survey Study (e80654)
Carolyn Turvey, Natalie Suiter, Rhonda Fellows, Shawna Domeyer, Lindsey Fuhrmeister, Amanda Heeren, Kimberly McCoy, M Howren. . . .
1 1 2 4 8

Impact of the Practice of Playing Video Games on Technical Skills Development in Preclinical Dental Education: Preliminary Cross-Sectional Observational Study (e80082)
Roselyne Clouet, Thomas Remaud, Camille Boeffard, Samuel Serisier, Alexis Gaudin, Tony Prud'homme. 11260

XSL•FO
RenderX

Changes in the Neighborhood Built Environment and Chronic Health Conditions in Washington, DC, in 2014-2019: Longitudinal Analysis (e74195)
Quynh Nguyen, Riki Doumbia, Thu Nguyen, Xiaohe Yue, Heran Mane, Junaid Merchant, Tolga Tasdizen, Mitra Alirezaei, Pankaj Dipankar, Dapeng Li, Penchala Mullaputi, Amrutha Alibilli, Yulin Hswen, Xin He. 11270

Mapping the Availability of Rehabilitation Providers Using Public Licensure and Population Data for a Geographic Information System–Based Approach to Workforce Planning: Cross-Sectional Feasibility Study (e85025)
Madeline Ratoza, Rupal Patel, Julia Chevan, Wayne Brewer, Katy Mitchell. 11288

Estimating the Frequency of Inpatient Adverse Events Using a 2-Step Retrospective Chart Review: Initial Observational Cohort Study (e82512)
Syed Ul Hassan, Fozia Asif, Farwa Ayub, Ghazal Haque, Asad Latif. 11302

Perceived Risk and Fashion on the Intention to Adopt Wireless Earbuds in the United States Using a Partial Least Squares-Structural Equation Modeling Approach: Empirical Study (e56887)
Edward Lee, Chandrasekar Subramaniam, Sungjune Park. 11320

Internet Access and Use Among Dementia Carers and the People They Support in Australia: Cross-Sectional Survey (e68333)
Emilie Cameron, Elise Mansfield, Ama Ampofo, Andrea Coda, Allison Boyes. 11331

Adolescents' and Young Adults' Perceptions of a Pop-Up Aimed at Combating the Spread of E-Cigarette Misinformation on Social Media: Cross-Sectional Survey Study (e73193)
Jessica Maturo, Shivani Gaiha. 11341

Fast Healthcare Interoperability Resources (FHIR)–Based Interoperability Design in Indonesia: Content Analysis of Developer Hub's Social Networking Service (e51270)
Lukman Heryawan, Yukiko Mori, Goshiro Yamamoto, Naoto Kume, Lutfan Lazuardi, Anis Fuad, Tomohiro Kuroda. 11350

Crisis Communication About the Maui Wildfires on TikTok: Content Analysis of Engagement With Maui Wildfire–Related Posts Over 1 Year (e67515)
Jim Stimpson, Aditi Srivastava, Ketan Tamirisa, Joseph Kaholokula, Alexander Ortega. 11363

Informal Caregivers Connecting on the Web: Content Analysis of Posts on Discussion Forums (e64757)
Michelle Foster, Chinenye Egwuonwu, Erin Vernon, Mohammad Alarifi, M Hughes. 11375

Planned Behavior in the United Kingdom and Ireland Online Medicine Purchasing Context: Mixed Methods Survey Study (e55391)
Bernard Naughton. 11386

Improving Suicidal Ideation Detection in Social Media Posts: Topic Modeling and Synthetic Data Augmentation Approach (e63272)
Hamideh Ghanadian, Isar Nejadgholi, Hussein Al Osman. 11398

Analysis of the Political Viewpoint of Policy Statements From Professional Medical Organizations Using ChatGPT With GPT-4: Cross-Sectional Study (e66204)
Ben Knudsen, Amr Madkour, Preetam Cholli, Alyson Haslam, Vinay Prasad. 11422

The Use of Technology by the Rural-Dwelling Caregivers of People Living With Dementia to Support Caregiving: Qualitative Interview Study (e77231)
Anna Jolliff, Sarah Boucher, Jordan Hill, Kristen Allen-Watts, Miriam Rodriguez, Christian Elliott, Matthew Zuraw, Nicole Werner. 11435

XSL•FO
RenderX

Surveillance of Twitter Data on COVID-19 Symptoms During the Omicron Variant Period: A Sentiment Analysis (e66237)
Kaiyue Zhang, Zhaojin Guo, Yujie Ai, An-Ran Li, Anlin Li, Ziyu Liu, Yittie Tse, Xinyu Zhou, Taoran Liu, Chuxi Xiong, Jian Huang, Wai-kit Ming. 1 1 5 5 6

Monitoring People With COVID-19 at Home With the COVIDFree@Home Program: Feasibility Cohort Study (e69140)
Andrea Gershon, Alex Mariakakis, Eyal de Lara, Joseph Munn, Maryann Calligan, Daniyal Liaqat, Salaar Liaqat, Junlin Chen, Teresa To, Philip Lam, Andrew Simor, Adrienne Chan, Nisha Andany, Sameer Masood, Nick Daneman, Tiffany Chan, Christopher Graham, Vikram Comondore, Andre de Moulliac, Alice Tu, Robert Wu. 11568

Trends and Shifts in Swedish Telemedicine Consultations During the Pre–COVID-19, COVID-19, and Post–COVID-19 Periods: Retrospective Observational Study (e60294)
Adaora Asomugha, Annamaria Pakai. 11577

Analysis of Social Media Perceptions During the COVID-19 Pandemic in the United Kingdom: Social Listening Study (2019-2022) (e63997)
Marzieh Araghi, Arron Sahota, Maciej Czachorowski, Kevin Naicker, Natalie Bohm, Katie Phillipps, James Gaddum, Erica Cook. 11585

Economic Deterioration and Social Factors Affecting Mental Health During the COVID-19 Pandemic in Japan: Web-Based Cross-Sectional Survey (e65204)
Kentaro Hori, Yosuke Yamada, Hideyuki Namba, Misaka Kimura, Hiroyuki Fujita, Heiwa Date. 11595

Explainable Machine Learning Framework for Dynamic Monitoring of Disease Prognostic Risk: Retrospective Cohort Study (e65585)
Tetsuo Ishikawa, Masahiro Shinoda, Megumi Oya, Koichi Ashizaki, Shinichiro Ota, Kenichi Kamachi, Kazuhiro Sakurada, Eiryo Kawakami, Masaharu Shinkai. 11608

Quantifying the Regional Disproportionality of COVID-19 Spread: Modeling Study (e59230)
Kenji Sasaki, Yoichi Ikeda, Takashi Nakano. 11623

Increasing COVID-19 Testing and Vaccination Uptake in the Take Care Texas Community-Based Randomized Trial: Adaptive Geospatial Analysis (e62802)
Kehe Zhang, Jocelyn Hunyadi, Marcia de Oliveira Otto, Miryoung Lee, Zitong Zhang, Ryan Ramphul, Jose-Miguel Yamal, Ashraf Yaseen, Alanna Morrison, Shreela Sharma, Mohammad Rahbar, Xu Zhang, Stephen Linder, Dritana Marko, Rachel Roy, Deborah Banerjee, Esmeralda Guajardo, Michelle Crum, Belinda Reininger, Maria Fernandez, Cici Bauer. 11630

Investigating Reddit Data on Type 2 Diabetes Management During the COVID-19 Pandemic Using Latent Dirichlet Allocation Topic Modeling and Valence Aware Dictionary for Sentiment Reasoning Analysis: Content Analysis (e51154)
Meghan Nagpal, Niloofar Jalali, Diana Sherifali, Plinio Morita, Joseph Cafazzo. 11645

COVID-19 Public Health Communication on X (Formerly Twitter): Cross-Sectional Study of Message Type, Sentiment, and Source (e59687)
Sana Parveen, Agustin Pereira, Nathaly Garzon-Orjuela, Patricia McHugh, Aswathi Surendran, Heike Vornhagen, Akke Vellinga. 11656

Food Access in New York City During the COVID-19 Pandemic: Social Media Monitoring Study (e49520)
Leah Butz, Charles Platkin, Jonathan Chin, Juan Chavez Salas, Ellie Serres, May Leung. 11670

Comparing Media and Law Enforcement Reports on Anti-Asian Hate Incidents During the COVID-19 Pandemic: Data Visualization Approach (e70881)
Young Yoon, Su Shin, Dongwook Kim, Hee Lee. 11686

Dental Undergraduate Students' Perceptions of Blended Learning in the COVID-19 and Post–COVID-19 Years: Survey Study (e63453)
Xin Li, Li Peng, Heng Tu, Lijia Liu, Xianglong Han, Ling Ye, Jing Yang. 11703

XSL•FO
RenderX

Communication Strategies for Tick-Borne Disease Prevention Among US Cattle Producers: Quasi-Experimental Study on Message Framing and Source Credibility (e77239)
Cheng-Xian Yang, Lauri Baker, Jessica Czipulis. 11736

Firearm Violence and Health in Policymaker Discourse: Mixed Methods Social Media Analysis (e80397)
Vivek Ashok, William Vervilles, Katherine Kellom, Anyun Chatterjee, Isabella Ntigbu, Okechi Boms, Andrea Szabo, Joel Fein, Therese Richmond, Jonathan Purtle, Matthew Kearney, Zachary Meisel. 11750

Quality Assessment of Shock Videos on Video Sharing Platforms: Cross-Sectional Study (e76715)
Wenxin Wang, Luping Cheng, Xia Hu, Chuanliang Pan. 11771

Didactic and Content Quality of Basic Life Support Videos on YouTube: Cross-Sectional Study (e69103)
Jasmina Sterz, Yvonne Beaugé, Pia Tueckmantel, Lena Bepler, Armin Flinspach, Yves Gramlich, René Verboket, Philip Bintaro, Maren Janko, Mairen Flinspach, Michael Merker, Sven Bepler, Jan Vollrath, Sebastian Voß, Miriam Rüsseler. 11782

Quality Analysis of Stroke-Related Videos on Video Platforms: Cross-Sectional Study (e80458)
Shao-Jie Nie, Shuai-Nan Ning, Qi-Chao Ding, Jie Hou, Yue-Yang Luo, Hai-Yang Wang. 11795

Associations Between Online Search Trends and Outpatient Visits for Common Medical Symptoms in the United States from 2004 to 2019: Time Series Ecological Study (e77274)
Ryan Heumann, Steven Steinhubl. 11812

Medical Information Provided by Transgender and Gender-Diverse Content Creators on YouTube: Descriptive Content Analysis (e75787)
Lydia Bliss, Qianqian Zhao, Irene Chao, Oliver Haimson, Ellen Selkie. 11820

Public Information Needs and Interest in Specific Food and Drug Allergy Disorders in Germany (2022–2024): Google Search Engine Analysis (e75395)
Tobias Fuchs, Michael Hindelang, Sebastian Sitaru, Alexander Zink, Julia Welzel. 11834

Quality and Dissemination of Uterine Fibroid Health Information on TikTok and Bilibili: Cross-Sectional Study (e75120)
Lan Wang, Yiwen Chen, Duo Zhao, Tao Xu, Fu Hua. 11845

Quality and Reliability of Transarterial Chemoembolization Videos on TikTok and Bilibili: Cross-Sectional Content Analysis Study (e73855)
Yushuo Niu, Guilan Song, Zheyu Niu, Sijian Xiao, Cuicui Li, Na Han, Hao Wan, Xiaohong Hou. 11864

Investigating the Association of Subjective Numeracy, Interpersonal Communication, and Perceived Discrimination With Watching Health-Related Videos on Social Media Platforms: Cross-Sectional Analysis (e71971)
Katerina Andreadis, Nancy Buderer, Aisha Langford. 11876

Self-Disclosure and Social Support in a Web-Based Opioid Recovery Community: Machine Learning Analysis (e71207)
Yu Chi, Huai-yu Chen, Khushboo Thaker. 11885

Digital Media Coverage of Respiratory Syncytial Virus-Related News in India: Mixed Methods Content Analysis of Disease Burden and Intervention (e70322)
Rhythm Hora, Arindam Ray, Amrita Kumari, Rashmi Mehra, Amanjot Kaur, Syed F Quadri, Bodhisatwa Ray, Seema Singh Koshal, Shyam Kumar Singh, Abida Sultana, Arup Deb Roy. 11900

XSL•FO
RenderX

Analyzing the Contents of a Large, Public Online Peer Support Forum for Obsessive-Compulsive Disorder: Thematic Analysis (e60899)
Nora Sun, Christopher Pittenger, Terence Ching. 11909

Internet Health Information–Seeking Trend of Urinary Incontinence in Mainland China: Infodemiology Study (e55670)
Shuangquan Lin, Lingxing Duan, Xiongbing Lu, Haichao Chao, Xi Wen, Shanzun Wei. 11919

Medical Misinformation in AI-Assisted Self-Diagnosis: Development of a Method (EvalPrompt) for Analyzing Large Language Models (e66207)
Troy Zada, Natalie Tam, Francois Barnard, Marlize Van Sittert, Venkat Bhat, Sirisha Rambhatla. 11934

Online Health Information–Seeking Behaviors Among the Chongqing Population: Cross-Sectional Questionnaire Study (e56028)
Honghui Rong, Lu Lu, Miao He, Tian Guo, Xian Li, Qingliu Tao, Yixin Li, Chuanfen Zheng, Ling Zhang, Fengju Li, Dali Yi, Enyu Lei, Ting Luo, Qinghua Yang, Ji-an Chen. 11945

YouTube User Traffic to Paired Epilepsy Education Videos in English and Spanish: Comparative Study (e56720)
Luna Varela, Stephanie Horton, Ahmed Abdelmoity, Jean-Baptiste Le Pichon, Mark Hoffman. 11956

Stigma of Dementia on Social Media During World Alzheimer's Awareness Month: Thematic Analysis of Posts (e72775)
Juanita-Dawne Bacsu, Jasmine Mah, Ali Jamali, Christine Conanan, Samantha Lautrup, Corinne Berger, Dylan Fiske, Sarah Fraser, Anila Virani, Florriann Fehr, Alison Chasteen, Zahra Rahemi, Shirin Vellani, Melissa Andrew, Allison Cammer, Katherine McGilton, Rory Gowda-Sookochoff, Kate Nanson, Karl Grewal, Raymond Spiteri. 11969

Privacy Concerns Versus Personalized Health Content—Pregnant Individuals' Willingness to Share Personal Health Information on Social Media: Survey Study (e60862)
Haijing Hao, Yang Lee, Marianne Sharko, Qilu Li, Yiye Zhang. 11978

Instagram Posts Promoting Colorectal Cancer Awareness: Content Analysis of Themes and Engagement During Colorectal Cancer Awareness Month (e63344)
Aditi Srivastava, Jim Stimpson. 11988

Quality Assessment of Medical Institutions' Websites Regarding Prescription Drug Misuse of Glucagon-Like Peptide-1 Receptor Agonists by Off-Label Use for Weight Loss: Website Evaluation Study (e68792)
Rie Oyama, Tsuyoshi Okuhara, Emi Furukawa, Hiroko Okada, Takahiro Kiuchi. 11996

Public Health Discussions on Social Media: Evaluating Automated Sentiment Analysis Methods (e57395)
Lisa Gandy, Lana Ivanitskaya, Leeza Bacon, Rodina Bizri-Baryak. 12010

Understanding Health-Related Discussions on Reddit: Development of a Topic Assignment Method and Exploratory Analysis (e55309)
Garrett Chan, Mark Fung, Jill Warrington, Sarah Nowak. 12024

Person-Specific Analyses of Smartphone Use and Mental Health: Intensive Longitudinal Study (e59875)
Merve Cerit, Angela Lee, Jeffrey Hancock, Adam Miner, Mu-Jung Cho, Daniel Muise, Anna-Angelina Garròn Torres, Nick Haber, Nilam Ram, Thomas Robinson, Byron Reeves. 12046

Analyzing Patient Complaints in Web-Based Reviews of Private Hospitals in Selangor, Malaysia, Using Large Language Model–Assisted Content Analysis: Mixed Methods Study (e69075)
Muhammad Sulaiman, Nora Muda, Fatimah Abdul Razak. 12062

Examining How Oral Nicotine Pouches Are Trending on TikTok: A Qualitative Descriptive Study (e73032)
Ashmeet Mand, Karlee Fonteyne, Laura Struik. 12077

XSL•FO
RenderX

Incidence, Prognostic Factors, and Treatment Impact on Survival in Natural Killer/T-Cell Lymphoma: Population-Based Study in the United States (e70129)
Yi Zhao, Yi Zhong, Mengqi Xiong, Linlin Huang, Xiujin Ye, Jingsong He. 12163

Implementing a Cross-Border Next-Generation Personal Health Record in the Philippines and Taiwan: An Implementation Case Report Using Health Level 7 International Fast Healthcare Interoperability Resources (e56272)
Hsiu-An Lee, Jui-Chun Huang, Shih-Wun Huang, Wei-Han Chen, Alvin Marcelo, Miguel Aljibe, Chien-Yeh Hsu. 12203

Development of an Eco-Friendly Smartphone-Assisted Nystagmus Recording System for Recording Vertigo Attacks Anytime, Anywhere: Pilot App Development Study (e73811)
Yumi Dobashi, Masao Noda, Tatsuaki Kuroda, Noriaki Miyata, Makoto Ito, Reiko Tsunoda, Hiroaki Fushiki. 12235

Development and Health System Deployment of an Electronic Health Record–Integrated Chatbot Intervention for Connecting Fall Risk Screening to Community Resources After Emergency Department Visits: Implementation Study (e77237)
Audrey Keleman, Megan Bounds, Maxwell Lunt, Jennifer Portz, Bucky Ferozan, Jonathan Gomez Picazo, Kelly Bookman, Hillary Lum, Elizabeth Goldberg. 12264

Adapted Virtual Reality Exergaming Using Off-the-Shelf Supplies for Poststroke Hemiparetic Arm Rehabilitation: Case Study (e80721)
Sadie Hare, Jason Vice, Mary Bowman, Ashley Wright, Raven Young, Mahmoud Ahmad, Maggie Logan, Byron Lai. 12282

Case Reports

At-Home Bioelectrical Impedance Analysis (BIA) Monitoring of Adult Females at Risk of Breast Cancer-Related Lymphedema: Nonrandomized One-Year Longitudinal Feasibility Study (e73978)
Cheri Teranishi-Hashimoto, Monica Padilla, Yoomi Heo, Lori Bravi. 1970

Sprint-Intensity Arm Interval Training May Improve Cardiorespiratory Fitness and Cardiometabolic Health Among Children With Mobility Disabilities: Case Report (e64440)
Natalie Dean, Tanvee Sinha, Larsen Bright, Katie Ellison, Drew Sayer, Raven Young, Drew Davis, James Rimmer, Byron Lai. 12215

Applying the Nonadoption, Abandonment, Scale-Up, Spread, and Sustainability (NASSS) Framework to Adapt the CHAMP App for Pediatric Feeding Tube Weaning: Application and Case Report (e67398)
Dana Bakula, Alexandra Zax, Sarah Edwards, Kristina Nash, April Escobar, Rachel Graham, Amy Ricketts, Ryan Thompson, Sarah Bullard, Julianne Brogren, Leah Shimmens, Lori Erickson. 12223

A Brief Technique to Reduce Flashbacks of Sexual Trauma in an Adolescent: Proof-of-Concept Case Study Using Imagery Interference (e79708)
Alex Lau-Zhu, Carmen Chan. 12242

Monitoring Nystagmus in a Patient With Vertigo Using a Commercial Mini-Infrared Camera and 3D Printer: Cost-Effectiveness Evaluation and Case Report (e70015)
Hiroyuki Sakazaki, Masao Noda, Yumi Dobashi, Tatsuaki Kuroda, Reiko Tsunoda, Hiroaki Fushiki. 12249

Concurrent Treatment of Opioid and Tobacco Use Disorder in a Telemedicine Clinic: Case Report of Breaking Through Barriers (e72872)
Darcy Michero, Laura Monico, Peyton Pielsticker, Larissa Mooney, Suzette Glasner. 12257

Enhancing Parenting Using AI: Exploratory Hackathon (e68780)
Peter Woods, Stephanie Donohoe, Louise Turtle, Udit Agrawal, Joshua Humphriss, Niel Cordes, Nathan Hodson. 12275

XSL•FO
RenderX

Short Papers

Validation of Sleep Measurements of an Actigraphy Watch: Instrument Validation Study (e63529)
Mari Waki, Ryohei Nakada, Kayo Waki, Yuki Ban, Ryo Suzuki, Toshimasa Yamauchi, Masaomi Nangaku, Kazuhiko Ohe. 4127

CoVimmune COVID-19 Immunity Calculator: Web Application Development and Validation Study (e59467)
Rebecca Slotkin, Tassos Kyriakides, Vinni Yu, Xien Chen, Anupam Kundu, Shaili Gupta. 8826

Feasibility of an Electronic Patient-Reported Outcome System in People Living With HIV: Retrospective Analysis of a Mobile App-Based Pilot Study (e71493)
Yusuke Yoshino, Yoshitaka Kimura, Yoshitaka Wakabayashi, Takatoshi Kitazawa. 12295

Identifying the Minimal Clinically Important Difference in Emotion Regulation Among Youth Using the JoyPop App: Survey Study (e64483)
Jaidyn Charlton, Ishaq Malik, Angela Ashley, Amanda Newton, Elaine Toombs, Fred Schmidt, Janine Olthuis, Kristine Stasiuk, Tina Bobinski, Aislin Mushquash. 12299

Assessing the Impact of a Virtual Reality Cognitive Intervention on Tennis Performance in Junior Tennis Players: Pilot Study (e66979)
Joaquin Anguera, Aleem Choudhry, Michael Seaman, Dominick Fedele. 12306

Association Between Alcohol Consumption and Psoriasis: Exploratory Analysis of Crowdsourced Web Search Data in Sweden (e71992)
Anna Schober, Klas Nordlind, Alexander Zink. 12314

Protocols

Assessing Minority Stress and Physiological Response Through Ecological Momentary Assessment and Sensors: Protocol for a Feasibility and Acceptability of the Stress and Heart Pilot Study (e68733)
Dulce Urueta Tapia, Heather Corliss, Kang Lee, Jerel Calzo, Hee-Jin Jun. 6945

Incorporating Patient-Reported Outcome Measures and Patient-Reported Experience Measures in Addiction Treatment Services in Belgium: Naturalistic, Longitudinal, Multicenter Cohort Study (e65686)
Amine Zerrouk, Charlotte Migchels, Clara De Ruysscher, Kim Fernandez, Jerome Antoine, Florian De Meyer, Frieda Matthys, Wim van den Brink, Cleo Crunelle, Wouter Vanderplasschen. 9438

Review

Misinformation About Medical Cannabis in YouTube Videos: Systematic Review (e76723)
Shivank Khare, Simon Erridge, Swathikan Chidambaram, Mikael Sodergren. 11311

Viewpoints

Medicine Beyond Machines: Viewpoint on the Art of Thinking in the Age of AI (e76669)
Isabella Ferreira, Rodrigo Menezes, Luis Correia, Bruno Andrade. 11445

mHealth as a Key Component of a New Model of Primary Care for Older Adults (e82262)
Jean Woo, Ruby Yu, Maggie Wong, Ken Cheung, Nicole Fung. 11451

XSL•FO
RenderX

Hype vs Reality in the Integration of Artificial Intelligence in Clinical Workflows (e70921)
Alaa Abd-Alrazaq, Barry Solaiman, Yosra Mekki, Dena Al-Thani, Faisal Farooq, Metab Alkubeyyer, Mohamed Abubacker, Rawan AlSaad, Sarah Aziz, Ahmed Serag, Rajat Thomas, Javaid Sheikh, Arfan Ahmed. 11457

Leveraging Artificial Intelligence for Clinical Study Matching: Key Threads for Interweaving Data Science and Implementation Science (e71831)
Andrew Goodwin, Sara Armstrong, David Ptak, Kenneth Catchpole, Jihad Obeid, Paul Heider. 11469

Evaluating the Role of Artificial Intelligence in Breast Self-Examination (e82550)
Shirish Nayyar. 11474

Embedding Passive Monitoring Into Global Health and Longitudinal Patient Care (e81580)
Angelika Papanicolaou, Laura Gaetano, Olivia Yu, Graham Jones. 11479

Applications of Self-Driving Vehicles in an Aging Population (e66180)
Sara Shu, Benjamin Woo. 11486

Assessing the Current Limitations of Large Language Models in Advancing Health Care Education (e51319)
JaeYong Kim, Bathri Vajravelu. 11491

Promoting Digital Health Data Literacy: The Datum Project (e60832)
Daniel Powell, Laiba Asad, Elissa Zavaglia, Manuela Ferrari. 11505

Development of the Big Ten Academic Alliance Collaborative for Women in Medicine and Biomedical Science: “We Built the Airplane While Flying It” (e65561)
Maya Iyer, Aubrey Moe, Susan Massick, Jessica Davis, Megan Ballinger, Kristy Townsend. 11511

Applying Critical Discourse Analysis to Cross-Cultural Mental Health Recovery Research (e64087)
Yasuhiro Kotera, Riddhi Daryanani, Oliver Skipper, Jonathan Simpson, Simran Takhi, Merly McPhilbin, Benjamin-Rose Ingall, Mariam Namasaba, Jessica Jepps, Vanessa Kellermann, Divya Bhandari, Yasutaka Ojio, Amy Ronaldson, Estefania Guerrero, Tesnime Jebara, Claire Henderson, Mike Slade, Sara Vilar-Lluch. 11522

Are Dating App Algorithms Making Men Lonely and Does This Present a Public Health Concern? (e70594)
Eric Balki. 11529

Addressing the Ethical, Legal, and Social Issues of Healthtech in Education: Insights From Japan (e72781)
Motofumi Sumiya, Tomoko Nishimura, Kyoko Aizaki, Ikue Hirata, Nobuaki Tsukui, Yuko Osuka, Manabu Wakuta, Atsushi Senju. 11533

Exploring the Role of Telemedicine in Duchenne Muscular Dystrophy: Benefits and Challenges (e77698)
Eliza Wasilewska, Andrzej Wasilewski, Alessandro Onofri, Jan Wasilewski, Dominika Sabiniewicz-Ziajka, Agnieszka Sobierajska-Rek, Jarosław Meyer-Szary, Karolina ledzi ska, Jolanta Wierzba, Marek Niedoszytko, Sylwia Małgorzewicz. 11547

Corrigenda and Addendas

Correction: Breath-Focused Mindfulness and Compassion Training in Parent-Child Dyads: Pilot Intervention Study (e81122)
Satish Jaiswal, Jason Nan, Seth Dizon, Jessica Young, Suzanna Purpura, James Manchanda, Dhakshin Ramanathan, Dennis Kuo, Jyoti Mishra. 11715

Correction: Latino Parents’ Reactions to and Engagement With a Facebook Group–Based COVID-19 Vaccine Promotion Intervention: Mixed Methods Pilot Study (e76107)
Anna González-Salinas, Elizabeth Andrade, Lorien Abroms, Kaitlyn Gómez, Carla Favetto, Valeria Gómez, Karen Collins. 11716

XSL•FO
RenderX

Correction: Advancing Mental Health and Psychological Support for Health Care Workers Using Digital Technologies and Platforms (e84386) 11718

Correction: A Neuroaffirmative, Self-Determination Theory–Based Psychosocial Intervention for Adults With Attention-Deficit/Hyperactivity Disorder: Randomized Feasibility Study (e87679)
Rebecca Champ, Rita Meneses, Marios Adamou, Warren Gillibrand, Sally Arrey, Barry Tolchard. 11720

Correction: Effectiveness, Usability, and Satisfaction of a Self-Administered Digital Intervention for Reducing Depression, Anxiety, and Stress in a University Community in the Andean Region of Peru: Randomized Controlled Trial (e87717)
Rosario Yslado-Méndez, Stefan Escobar-Agreda, David Villarreal-Zegarra, Wilfredo Flores, Junior Sánchez-Broncano, Ana Vilela-Estrada, Jovanna Córdova, C Reategui-Rivera, Claudia Alvarez-Yslado, Leonardo Rojas-Mezarina. 11722

Correction: Online Health Information–Seeking Behaviors Among the Chongqing Population: Cross-Sectional Questionnaire Study (e77500)
Honghui Rong, Lu Lu, Miao He, Tian Guo, Xian Li, Qingliu Tao, Yixin Li, Chuanfen Zheng, Ling Zhang, Fengju Li, Dali Yi, Enyu Lei, Ting Luo, Qinghua Yang, Ji-an Chen. 11724

Correction: Evaluating Bard Gemini Pro and GPT-4 Vision Against Student Performance in Medical Visual Question Answering: Comparative Case Study (e71664)
Jonas Roos, Ron Martin, Robert Kaczmarczyk. 11725

Correction: Acceptability and Initial Adoption of the Heart Observation App for Infants With Congenital Heart Disease: Qualitative Study (e70280)
Elin Hjorth Johansen, Elin Børøsund, Ingeborg Martinsen Østen, Henrik Holmstrøm, Anne Moen. 11726

Correction: Assessing and Improving Data Integrity in Web-Based Surveys: Comparison of Fraud Detection Systems in a COVID-19 Study (e76462)
Stephen Bonett, Willey Lin, Patrina Sexton Topper, James Wolfe, Jesse Golinkoff, Aayushi Deshpande, Antonia Villarruel, José Bauermeister.
1 1 7 2 8

Correction: Identifying Themes for an Initial Beta Version of a Mobile Health App for Latino and Native Hawaiian/Pacific Islander Communities: Co-Design and Community-Based Participatory Research in a Code to Community Study (e79194)
Christina Holub, Amy Barrera, Rosalva Romero Gonzalez, Diane Hoang, Luna Prieto, Samuelu Fesili, Tiana Smith, Harleen Kaur, Cassandra Surban, Michael Markidis, Tana Lepule, Konane Martinez. 11730

Correction: “Smartwatch-Based Ecological Momentary Assessment for High-Temporal-Density, Longitudinal Measurement of Alcohol Use (AlcoWatch): Feasibility Evaluation” (e82194)
Chris Stone, Sally Adams, Robyn Wootton, Andy Skinner. 11732

Correction: Overcoming Language Barriers in Paramedic Care With an App Designed to Improve Communication With Foreign-Language Patients: Nonrandomized Controlled Pilot Study (e83353)
Frank Müller, Dominik Schröder, Eva Noack. 11734

Research Letters

Association Between eHealth Literacy and Mental Health Literacy: Cross-Sectional Study (e76812)
Efrat Neter, Refael Youngmann, Naama Gruper. 12086

Move – Mix It Up – Make It Fun: Summarizing and Visualizing Physical Activity Guidance (e84225)
Julian Brummer, Florian Herbolsheimer, Mona Illmann, Duc Nguyen, Karen Steindorf. 12089

XSL•FO
RenderX

Effectiveness of Text Message Outreach to Promote Enrollment in a Government Food Assistance Program: Pilot Randomized Trial (e78907)
Emily Melnick, Francesco Acciai, Nicole Vaudrin O'Reilly, Mindy Jossefides, A Ronan, Punam Ohri-Vachaspati. 12094

Use First, Trust Later? Exploring How Health Care Providers View the Gaps Between AI's Regulation and Its Implementation (e74038)
Aviad Raz, Yael Inbar, Ziv Paz, Barkan Hofman, Zalman Itzhakov, Orly Weinstein. 12099

A Video-Based Intervention Supporting Spanish-Speaking Fathers of Infants in the Neonatal Intensive Care Unit: Open-Label Pilot Study (e72520)
Janine Bernardo, Elise Belkin, Suzanne Hally, Sara Boyle, Eric Van, Jose Tascon-Arcila, Emily Herzberg, Juan Matute, Paul Lerou. 12102

Changes in Skeletal Muscle Mass Index With Personalized Exercise and Meal Photo Analysis via Nutrition Applications: Single-Arm Pilot Study (e71807)
Hidenori Onishi, Yasutaka Mizukami, Yuki Niida, Kazue Fujita, Osamu Yamamura. 12111

Readability of Online Patient Educational Materials for Rosacea: Systematic Web Search and Analysis (e67916)
Derek Nguyen, Jennifer Javaheri, Daniel Nguyen, Vy Han. 12116

Coach-Supported Internet-Based Cognitive Therapy for Social Anxiety Disorder in Hong Kong: Development, Training, and Pilot Clinical Outcomes (e69818)
Graham Thew, Candice Powell, Lok Chan, Ezmond Cheung, Jeff Chung, Chun Li, Melody So, Yik Hsu, Man Ng, Sze Siu, Katja Beardsell, Amy Kwok, Mandy Lissillour Chan, David Clark, Patrick Leung. 12119

Expressions of "Ikizurasa" in Posts on X (Formerly Twitter) in Japan in 2023: Descriptive Analysis (e70613)
Kanami Tsuno, Azusa Arimoto, Yuki Imamatsu, Yumiko Kobayashi, Miho Satoh, Tomoyuki Miyazaki. 12123

Examining the Relationship Between Assertiveness and Anxiety in First- and Second-Year US Medical Students (e73394)
Jonathan Shaw, James Hagerty, Kristen Masada, Angelene Won, Ashley Lai, Jisu Shin, Van Le, Brenton Phung, Charles Lai, Peter Bota, Aaron Jacobs. 12126

The Relationship Between Medical Student Neuroticism and Preferred Medical Specialty: Cross-Sectional Study (e73507)
Jonathan Shaw, Brenton Phung, Ashley Lai, James Hagerty, Van Le, Seung Yoo, Jerome Torres, Angelene Won, Charles Lai, Peter Bota, Aaron Jacobs. 12132

Seasonal Variation in Conjunctivitis in Saudi Arabia and Reduction of Searches During the COVID-19 Pandemic: Analysis of Google Trends Data (e73845)
Abdulaziz AlHarthi. 12137

How Medical Students Manage Depression, Anxiety, and Stress: A Cross-Sectional Study (e74218)
Jonathan Shaw, Ashley Lai, Sasha Singh, Seung Yoo, Maha Fathali, Laura Stuck, James Hagerty, Van Le, Jisu Shin, Charles Lai, Peter Bota, Aaron Jacobs. 12140

The Digital Availability of US Departments of Corrections' Research Policies: Cross-Sectional Analysis (e76835)
Jesse Martinez-Kratz, David Manning, Noel Vest, Jason Glenn, Lauren Brinkley-Rubinstein, Alysse Wurcel. 12145

Guideline-Incorporated Large Language Model-Driven Evaluation of Medical Records Using MedCheckLLM (e53335)
Marc Schubert, Stella Soyka, Wolfgang Wick, Varun Venkataramani. 12150

XSL•FO
RenderX

The AI Reviewer: Evaluating AI's Role in Citation Screening for Streamlined Systematic Reviews (e58366)
Jamie Ghossein, Brett Hryciw, Tim Ramsay, Kwadwo Kyeremanteng. 12154

Engaging Stakeholders With Professional or Lived Experience to Improve Firearm Violence Lexicon Development (e68105)
Nicole Cook, Frances Biel, Kerry Bet, Marion Sills, Ali Al Bataineh, Pedro Rivera, Anna Templeton, Natalie Cartwright. 12157

Exploring Public Sentiment on the Repurposing of Ivermectin for COVID-19 Treatment: Cross-Sectional Study Using Twitter Data (e50536)
Angga Kautsar, Rano Sinuraya, Jurjen van der Schans, Maarten Postma, Auliya Suwantika. 12183

Identifying Themes for an Initial Beta Version of a Mobile Health App for Latino and Native Hawaiian/Pacific Islander Communities: Co-Design and Community-Based Participatory Research in a Code to Community Study (e76178)
Christina Holub, Amy Barrera, Rosalva Romero Gonzalez, Diane Hoang, Luna Prieto, Samuelu Fesili, Tiana Smith, Harleen Kaur, Cassandra Surban, Michael Markidis, Tana Lepule, Konane Martinez. 12187

Use of and Attitudes Toward Technology Among Young People Living With HIV in San Francisco: Cross-Sectional Study (e81845)
Sean Arayasirikul, Caitlin Turner, Dillon Trujillo, Jarett Maycott. 12192

Feasibility of a Virtual Reality Intervention Protocol to Improve Cognitive and Behavioral Skills in Older Adults at Increased Risk of Developing Dementia (e77111)
Oshadi Jayakody, Mirnova Ceïde, Joe Verghese, Robert Carrera, Hussain Doriwala, Sunil Agrawal, Judy Pa, Helena Blumen. 12199

XSL•FO
RenderX

Assessing Patient Acceptance of Being Offered a Virtual Calm Room in Psychiatric Inpatient Care: Mixed Methods Study

Maria Ilioudi[1,2], BSc, MSc; Steinn Steingrimsson[2,3], MD, PhD; Sara Wallström[1,4,5,6], RN, PhD; Almira Osmanovic Thunström[7], PhD; Philip Lindner[8], PhD; Michael Ioannou[2,7], MD, M Phil; Örjan Falk[2], MD, PhD; Lilas Ali[1,2,4], RN, PhD

[1]Institute of Health and Care Sciences, Sahlgrenska Academy, University of Gothenburg, Arvid Wallgrensbacke 1, Gothenburg, Sweden
[2]Department of Psychiatry for Affective Disorders, Region Västra Götaland, Sahlgrenska University Hospital, Gothenburg, Sweden
[3]Institute of Neuroscience and Physiology, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden
[4]Centre for Person-Centred Care (GPCC), University of Gothenburg, Gothenburg, Sweden
[5]Department of Forensic Psychiatry, Region Västra Götaland, Sahlgrenska University Hospital, Gothenburg, Sweden
[6]Centre for Ethics, Law and Mental Health (CELAM), University of Gothenburg, Gothenburg, Sweden
[7]Institute of Neuroscience and Physiology, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden
[8]Centre for Psychiatry Research, Department of Clinical Neuroscience, Karolinska Institutet & Stockholm Health Care Services, Stockholm, Sweden

Corresponding Author:
Maria Ilioudi, BSc, MSc
Institute of Health and Care Sciences, Sahlgrenska Academy, University of Gothenburg, Arvid Wallgrensbacke 1, Gothenburg, Sweden

Abstract

Background: Implementing new technologies in health care settings is often a complex and challenging process. Virtual reality (VR) has demonstrated promising results in terms of feasibility, acceptability, and effectiveness across various health conditions. VR can provide calming environments to enhance patients' well-being in psychiatric inpatient care. However, little research has been done on patients' acceptance of VR technology in psychiatric care in the real-world setting.

Objective: This study aimed to explore patients' experiences of being offered the use of a virtual calm room when feeling anxious or worried in a psychiatric inpatient setting.

Methods: A mixed methods design was used, with a qualitative to quantitative approach. The data were gathered through individual interviews (n=10) and a 3-item rating scale ranging from 1 to 5 (n=59). The qualitative findings were then validated within a larger population using the quantitative data.

Results: The majority of participants reported being satisfied with the option of using VR. Their initial impressions of the virtual calm room were that it seemed like a creative and stimulating environment that could potentially have a positive impact on them. They expected the VR experience to enhance their feelings of relaxation and concentration. The participants highlighted human interaction as a particularly valuable aspect to consider when implementing VR, emphasizing its role in enhancing the overall experience and ensuring a sense of connection and support throughout the process. Participants were generally positive in the answers to the survey with a high score of feeling VR to be a modern method to manage anxiety (mean 4.3, SD 0.98), low levels of experiencing difficulty (mean 2.1, SD 1.4), and would generally consider future use (mean 4.4, SD 1.1).

Conclusions: The qualitative findings highlighted patients' openness to innovative methods for enhancing their engagement in the psychiatric inpatient setting. Patients expressed a desire for increased availability of the virtual calm room. However, maintaining a balance between innovative technologies and human support is crucial for the successful implementation of such methods. Quantitative results demonstrated high acceptance of the option of using the virtual calm room, with no significant difficulties reported.

(JMIR Form Res 2025;9:e75961) doi:10.2196/75961

KEYWORDS

acceptance; inpatient; psychiatric care; virtual reality; VR

XSL•FO
RenderX

Introduction

Implementing new technologies in health care settings is often challenging, but there is much to gain from doing so. Virtual reality (VR) is one such technology. Despite decades of research support for its clinical application, it has only recently begun to be incorporated into real-world health care [1]. VR, now considered a subtype under the umbrella term "Extended Reality," refers to headset technology capable of generating a sense of being present in a virtual world. A recent clinical application of VR provides calming virtual environments that can be offered in psychiatric inpatient settings to enhance patients' sense of safety. The application fosters a recovery-focused atmosphere that is the virtual equivalent of physical calm rooms [2]. Both physical and VR calm rooms offer nontraditional therapeutic interventions that improve patients' overall admission experience and well-being [3,4].

Research has already demonstrated the feasibility, acceptability, and effectiveness of clinical VR in regular healthcare settings [5]. It has also been shown to be a promising option in both inpatient [4,6,7] and outpatient [8] psychiatric care settings. It is increasingly being used in patients with conditions such as phobias [9], anxiety [10], post-traumatic stress [11], depression [12], delusions [13], and anorexia [14]. VR applications have been specifically designed to induce a sense of relaxation [15], lower anxiety, and increase well-being and independence [16,17]. Where safety concerns sometimes preclude the construction of physical calm rooms in psychiatric healthcare settings, VR technology could provide an alternative option [1].

Health care services are under near-constant pressure to increase both the quality and cost efficiency of care [18]. However, research has also uncovered significant challenges in implementing new technologies on a large scale [19]. Given that the health care system offers services co-created by health care professionals and people seeking help, improving health care services requires considering innovative ways to involve both staff and patients, along with new organizational forms, structures, and community engagement [20]. Patient and public involvement is also increasingly being recognized as beneficial in health care research. Participatory research design highlights users' experiences to create new systems based on their collective interpretations of research, thereby enhancing its usability. Users' knowledge is highly valued, grounded in a strong ethical framework, and can be leveraged to design new tools [21].

VR technology has, in theory, made it feasible to implement calm rooms in routine psychiatric care without the need for a dedicated room, trained and available staff, and other practical obstacles associated with the traditional alternative [4]. To maximize the probability of long-term, sustainable, and successful outcomes, this implementation should arguably be performed in a bottom-up manner, involving the intended users from the start to ensure that the technology at hand addresses actual problems and adds value [22]. Including the end user's perspective and therapists' feedback throughout the design, development, and real-world implementation of VR applications will likely result in enriched applications with elements familiar to and preferred by end users [23].

VR applications can offer individualized experiences for patients with various mental disorders [24] and exposure to relaxing VR scenes may provide an effective relaxation method to support and enhance psychiatric treatments [25,26]. In previous studies, patients reported high satisfaction with and acceptance of VR technology [8,27,28]. However, few studies have evaluated patients' acceptance of being offered the option of VR technology in psychiatric inpatient settings. A previous qualitative study conducted by the research team showed that use of a virtual calm room made a positive impression on patients. However, the earlier study did not explore the users' experiences of being offered such an option instead of traditional pharmacological treatment [13]. Therefore, the aim of this study was to explore patients' views on the option of being offered the use of a virtual calm room when feeling anxious or worried in a psychiatric inpatient setting.

Methods

Study Design

This study is part of a larger research project aimed at exploring the effect of VR and physical calm rooms in psychiatric inpatient care. The larger project also includes a quasi-randomized controlled trial and a qualitative interview study (trial registration NCT03918954) [4,7]. The participants included in this study were not part of the samples in any of the previous studies. This study adopted a mixed methods design using the qualitative to quantitative approach, with qualitative data being used to interpret the quantitative data [29].

Mixed methods research involves the integration of qualitative and quantitative data [29,30]. This methodology is appropriate when neither qualitative nor quantitative data adequately address a complex research question, when a phenomenon needs to be explored in depth, when different types of cases need to be explored and compared, as well as when a program needs to be developed, implemented, and evaluated. Findings from one approach are enhanced by the other approach [29,30].

In this study, qualitative data about individual patients' experience of being offered the use of the VR calm room were collected through interviews. In the second step, data were gathered from statements included in an evaluation form. In the last step, all the collected data were integrated into the interpretive process.

Sample

Patients who had already been admitted to psychiatric inpatient care were asked if they were interested in using a VR or physical calm room and willing to provide ratings as well as participate in interviews. The inclusion criteria were individuals aged 18 years or older and having been admitted to the inpatient psychiatric care ward specializing in bipolar disorder at Sahlgrenska University Hospital in western Sweden. No exclusion criteria were predefined. A total of 59 participants received a 3-item evaluation form, including 10 participants who were willing to be interviewed and were subsequently recruited. Data were collected between 2022 and 2023.

XSL•FO
RenderX

Patient and Public Involvement

This study incorporates a participatory design that assumes that patients and health care professionals are involved in the study design and development stage [21]. Extensive iterative consultations with the users took place during the development and design of the VR calm room. At this stage, users gave feedback on the application components and shared ideas to make the development more user-friendly and suitable for the clinical setting.

Procedure

Patients who had been admitted to a psychiatric care setting specializing in bipolar disorder used the virtual calm room. Health care professionals supported the patients with instructions on the VR headset's functions and how to navigate in the virtual environment. To engage with the device, patients used a wireless, 3 degrees-of-freedom VR head-mounted display (Oculus Go; Meta Reality Labs) running the Calm Place app developed by Mimerse. The application consists of relaxing nature environments such as a beach, a forest, and a desert (Figure 1). The application offers users different options to engage with, including breathing exercises, mindfulness programs, and relaxing music. They can also choose to just observe relaxing images of natural environments. Natural sounds (eg, birds, rain, and ocean waves) can be added according to the user's preference.

Figure 1. Example of a calming scene.

Data Collection

Qualitative Data

After the patients used the VR calm room, the first author (MI) conducted individual interviews with 10 patients. Each interview lasted between 7 and 19 minutes. None of the participants expressed any concerns or distress either during the interview or afterward. Participants were encouraged to be as specific as possible about their experience of being offered the use of the VR calm room. Interviews were conducted using an interview guide, which aimed to gather data on patients' reflections and reasoning. Participants were asked questions about their expectations when offered the VR calm room, whether they found it helpful, whether they experienced it as a modern method, how they experienced the health care professionals' attitudes to implementing the VR calm room, and whether they would be interested in using this method in the future.

Quantitative Data

Patient characteristics and demographics were self-reported and collected directly from the participants. To evaluate the acceptance of the virtual calm room, 59 patients, including the 10 participants who were interviewed, responded to 3 statements in a standardized evaluation form: "*this feels like a modern strategy to reduce anxiety*," "*I found it difficult to access the virtual calm room*," and "*If presented with the choice again, I would consent to using it*." Responses were recorded using a 5-point Likert scale ranging from 1 ("do not agree") to 5 ("fully agree"). The statements were carefully formulated to capture key components of intervention acceptance, including affective attitude, ethical considerations, perceived burden, intention to use, intervention coherence, opportunity costs, perceived effectiveness, and self-efficacy. This approach, which had been formulated for a previous pragmatic randomized controlled trial [31], ensured an evaluation of patients' experiences and attitudes toward the virtual calm room [32].

Data Analysis and Synthesis

Data from the interviews were analyzed using qualitative inductive content analysis. Qualitative content analysis involves breaking down data into smaller units for in-depth examination [30]. The primary challenge in qualitative research is

XSL•FO
RenderX

determining whether to focus the analysis on manifest or latent content, as both approaches involve interpreting the text. Another aspect of interpretation is that although researchers are influenced by their personal histories, they must allow "the text to speak for itself" [29]. A key characteristic of this method is its emphasis on the context and the subject matter. The content analysis followed the steps outlined by Graneheim and Lundman [33], which include selecting the unit of analysis, identifying meaning units, condensing and abstracting data, defining content areas, and developing codes, categories, and themes. We employed a descriptive qualitative approach to generate findings that closely align with the data. The transcribed text was read multiple times by the first author to gain an understanding of its content. The research group then discussed the content collaboratively to identify meaning units within the text. In the next step, the data were condensed and abstracted by MI and LA to create codes. Subthemes with similar content were developed and subsequently grouped into larger themes. The analysis was a dynamic process, with authors revisiting each step as needed. Finally, the authors discussed the overall organization of meaning units, codes, and subthemes, reaching a consensus on the main themes.

Data on background characteristics as well as the evaluation forms were analyzed using descriptive statistics. Both qualitative and statistical analyses were performed on the collected data. The results from the qualitative and quantitative analyses are presented separately. Patterns and relationships between the qualitative and quantitative data were identified, and details from the 2 methods were integrated to provide a comprehensive understanding of the findings. Following this, data from both sources were synthesized to draw broader conclusions. All authors were actively involved in the analysis process, collaboratively discussing and interpreting the various outcomes to ensure the validity and depth of the findings.

Ethical Considerations

Ethical approval for the study was granted by the Swedish Ethical Review Authority (318‑18 and amendment 2023-06878-02). Written informed consent was obtained from all participants, who could choose to end their use of the calm room at any time without needing to provide a reason. If a participant had previously exhibited any risk of self-harm or if there was any suspicion that their use of the calm room could be harmful, they were not granted access. Participants admitted to a psychiatric inpatient ward may receive compulsory psychiatric care. The research group was prepared to intervene if any participant was negatively affected by the questions asked or by their participation in the study in any way. All data were treated confidentially.

Results

Demographic and Clinical Characteristics

The demographic and clinical characteristics of the participants are summarized in Table 1. The sample consisted of 59 participants, of whom 17 (29%) were male participants and 38 (64%) female participants (n=4, 7% participants did not specify their sex). Participants' ages ranged from 18 to 74 years. The majority have been diagnosed with bipolar disorder.

Table . Demographic and clinical characteristics of survey participants (N=59[a]).

Characteristic	Value
Age (y), mean (SD; range)	42.9 (14.3; 18‑74)
Gender[b], n (%)	
Man	17 (29)
Woman	38 (64)
Left blank	4 (7)
Self-reported psychiatric disorder[c], n (%)	
Bipolar disorder[b]	34 (58)
Depression[c], n (%)	7 (12)
Other	3 (5)
Left blank	14 (24)

[a]Interviewees are also included in the population.

[b]Three participants also reported emotionally unstable personality disorder, with one of them also having attention-deficit/hyperactivity disorder.

[c]Three participants also reported anxiety disorders.

Results From Qualitative Interviews

The analysis of the qualitative data resulted in the identification of 2 themes dealing with the overall experience of being offered a technical solution when experiencing anxiety or worry while in the inpatient ward: *innovative thinking in providing other options* and *the value of human interaction*.

Innovative Thinking in Providing Other Options

The participants expressed feelings of being trapped in the psychiatric inpatient ward and wanting creative ways to stimulate their minds. Their first thought when offered use of the VR calm room was that it seemed like a creative environment that might offer some positive stimulation. They anticipated that the VR technology might enhance their ability

XSL•FO
RenderX

to relax and concentrate. Patients described the virtual calm room as a solution to their need for stimulation when they could not leave the ward. It provided an opportunity for a moment of relaxation, a brief time when they could experience harmony and well-being.

> *I have gained a new experience that I never thought I would have . . . I feel like I have so many creative ideas, and I would really love to work with VR in some other way.* [P1]

All the participants, both those who had used VR before and those using it for the first time, were satisfied with being offered access to the VR calm room. They perceived the room as easily accessible, and their first thought was that they were curious and open to experiencing a new method. Even participants who had never tried VR technology before did not report any hesitation about trying the VR calm room. Some participants stated that it was better than they had expected and that they would recommend it to other patients as an alternative option for relieving anxiety. The participants also found the process uncomplicated, with the lack of obstacles allowing them independence in using the device.

Some participants indicated that they wanted the VR session to become an everyday activity because they found it an attractive way to get through their challenging inpatient stay. It was an alternative that could help them focus on the present. The participants also said that the device seemed to be a good and valuable technical alternative to medication. Some participants mentioned other alternatives they had tried such as painting or taking walks in the yard. The VR experience was considered a new and interesting way to spend their time on the psychiatric inpatient ward.

The Value of Human Interaction

The participants appreciated being offered new technology; however, they were also clear that they preferred less medical support and more human support. They considered human contact with health care professionals to be an important part of their care. They valued their time with their contact person in the inpatient ward and emphasized that continued communication was important for their recovery. The patients also wanted the contact person to introduce them to the ward's routines and new alternatives such as the VR calm room.

> *I like alternative medicine; I appreciate things that have no side effects or fewer side effects. And then there's the idea that if you learn something in the virtual world, maybe you can apply that technique outside the virtual world on your own after practicing it several times. So, I thought it was very positive.* [P6]

The participants emphasized that they appreciated being listened to by the professionals on the inpatient ward, as well as being offered VR support and being shown how to use the self-guided applications. For some, the included mindfulness program distracted them from their anxiety. The participants also mentioned wanting more methods to increase body awareness, breathing techniques, or anxiety management to be available for them on the ward.

Ratings

Table 2 shows the results of the survey. The findings reveal that most participants did not encounter any difficulties using the VR technology, reporting a median score of 1 (on a scale where lower values indicate fewer difficulties). In addition, participants viewed VR technology as a modern treatment method for mental disorders, with a median score of 5 indicating a positive view. Most participants would choose to participate in the intervention again, if given the opportunity.

Table . Responses to an evaluation of acceptance of virtual calm room to 3 questions on a 5-point agree–disagree Likert scale (1=do not agree at all to 5=fully agree).

Question	Responders	Median	mean (SD)	Likert scale (%)				
				1	2	3	4	5
This feels like a modern strategy to reduce anxiety	59	5	4.3 (0.98)	3	2	12	29	54
I found it difficult to access the virtual calm room	58	1	2.1 (1.4)	52	17	16	5	10
If I could choose again, I would consent to using the virtual calm room	58	5	4.4 (1.1)	4	0	4	9	41

XSL•FO
RenderX

Discussion

Principal Findings

This mixed methods study explored patients' experiences of being offered the use of a virtual calm room in a psychiatric inpatient setting. The qualitative findings showed that patients expressed a need for creative stimulation and engagement. They appreciated the innovative use of VR technology, noting that it helped alleviate their anxiety and enhanced their relaxation and concentration.

Quantitative data further supported these findings in a larger population, showing that the virtual calm room was perceived as a modern treatment option for mental disorders, with a median score of 1 indicating minimal difficulties in usage. These results align with previous studies that have found that participants expressed a desire for new treatment options, such as VR technology, that go beyond traditional medication [7], and that they found virtual relaxation helpful, enjoyable, and relaxing [34-36]. Previous research shows that both health care providers and patients regarded VR technology as having the potential to bring something new and unique to psychiatric care [37].

The quantitative data show that most participants did not experience any difficulties using the virtual calm room. They felt a sense of autonomy while using the device and found it easy to initiate and follow the session. These findings align with an earlier study that reported that patients demonstrated high acceptance of VR technology and perceived it as uncomplicated to use [38].

Despite the potential and expanding use of VR technology, its use in clinical psychiatric care is limited, as is research on data with deeper insights [1]. This may be because implementing technology-based innovations can be challenging due to clinical risks and safety concerns [39]. While professional concerns about the implementation of VR technology are no longer the major barrier to its use in clinical contexts [40], evidence-based clinical guidelines and training are needed to ensure its effective integration into psychiatric care [39,41].

A positive patient experience, a sense of safety, and a strong professional relationship have been shown to be key elements of VR treatment [42]. In an earlier study, clinicians mentioned the need to add VR tasks that target patients' emotional states and to gather evidence before implementing the treatment. Patients felt that VR treatment should be adapted to their preferences and made more accessible and convenient in terms of how, when, and where it is used [43].

Patients expressed a willingness to use the virtual calm room again and indicated a desire for new and engaging ways to occupy their time on the ward. While the VR relaxation experience was perceived as more effective than conventional relaxation exercises for immediately alleviating negative affective states [26,44], this finding raises important questions about the sustainability and accessibility of such technology in psychiatric settings. Although the immediate benefits of VR are promising, reliance on technological interventions may inadvertently overshadow traditional therapeutic methods that are also vital for patient recovery. In previous research, treatments based on VR technology have not clearly shown superiority over conventional treatments [24]. Previous studies have also indicated that patients have maintained their ongoing pharmacological treatment alongside VR treatment, thus it is difficult to evaluate on its own [45]. Furthermore, not all patients may feel comfortable or able to engage with VR technology, potentially leading to disparities in treatment experiences. Therefore, while integrating VR into psychiatric treatment has the potential to enhance patient care, it is crucial to consider how to balance innovative methods with the diverse needs and preferences of all patients, ensuring that no one is left behind in the pursuit of improved mental health outcomes.

Strength and Limitations

This mixed methods (qualitative to quantitative) study has several limitations. The qualitative data came from a small sample of only 10 participants because of the inpatient psychiatric context, and the study aim did not require extensive, in-depth descriptions. However, the qualitative data complemented the quantitative data, enabling the research to effectively address its objective. The quantitative part of the study relied on an evaluation form that had not been validated as it was adapted from a previous clinical study. Furthermore, the evaluation form included a limited number of statements to maximize the sample size. Demographic data for the qualitative and quantitative parts were presented together, since participants self-reported their characteristics. The study was conducted within an affective disorder department, representing a limitation regarding the generalizability of the results to a wider psychiatric population. Despite these limitations, the mixed methods approach proved to be a strength of the study: by integrating qualitative insights with quantitative findings, the study provided a comprehensive understanding of patients' perspectives on being offered the use of a virtual calm room in a psychiatric inpatient setting. This methodology offered valuable insights into patient acceptance and the potential benefits of this intervention.

Conclusions

The qualitative results revealed patients' openness to VR calm rooms to improve their engagement with psychiatric inpatient care. The quantitative results demonstrated patients' high acceptance of virtual calm rooms and did not indicate any noteworthy difficulties. Patients wanted increased availability of virtual calm rooms. However, a balance between innovative methods and human support is crucial for the implementation of technologies. Further research based on codesign may provide patients with opportunities to adapt their preferences and develop more convenient innovative solutions.

XSL•FO
RenderX

Acknowledgments

The authors extend their heartfelt gratitude to the Department of Affective Disorders at Sahlgrenska University Hospital for their invaluable support. The authors provide their sincere thanks to the department's manager and the dedicated health care professionals whose guidance and encouragement were instrumental in driving the project forward. Furthermore, the authors thank Hanne Krage Carlsen (PhD) for statistical help and for creating the graphics. Most importantly, the authors wish to express our deep appreciation to the participants who generously shared their valuable experiences, making this study possible and enriching its outcomes. Generative artificial intelligence was not used in any portion of the manuscript generation.

Funding

The study was financed by grants from the Swedish state under the agreement between the Swedish Government and the county councils, the Avtal om Läkarutbildning och Forskning agreement (ALFGBG1006029).

Data Availability

Data availability is possible upon reasonable request in accordance with Swedish regulations.

Authors' Contributions

Conceptualization: SS, AOT, PL, LA
Data curation: MI, SS, LA
Formal analysis – qualitative: MI, LA
Formal analysis – quantitative: SS, ÖF, SW
Methodology: SS, LA, SW, MI
Project administration: MI, LA, SS
Development of the evaluation form: MI, SW
Writing – original draft: MI, LA
Writing – review & editing: all authors

Conflicts of Interest

PL reports having consulted for Mimerse, the company that developed the VR application in question; Mimerse is no longer operational. The other authors declare no conflicts of interest.

References

1. Lindner P. Better, virtually: the past, present, and future of virtual reality cognitive behavior therapy. J Cogn Ther 2021 Mar;14(1):23-46. [doi: 10.1007/s41811-020-00090-7]
2. Cutler NA, Halcomb E, Sim J, Stephens M, Moxham L. How does the environment influence consumers' perceptions of safety in acute mental health units? A qualitative study. J Clin Nurs 2021 Mar;30(5-6):765-772. [doi: 10.1111/jocn.15614] [Medline: 33348440]
3. Haig S, Hallett N. Use of sensory rooms in adult psychiatric inpatient settings: a systematic review and narrative synthesis. Int J Mental Health Nurs 2023 Feb;32(1):54-75. [doi: 10.1111/inm.13065]
4. Ilioudi M, Lindner P, Ali L, et al. Physical versus virtual reality-based calm rooms for psychiatric inpatients: quasi-randomized trial. J Med Internet Res 2023 May;25(1):e42365. [doi: 10.2196/42365] [Medline: 37204858]
5. Lindner P, Dagöö J, Hamilton W, et al. Virtual reality exposure therapy for public speaking anxiety in routine care: a single-subject effectiveness trial. Cogn Behav Ther 2021 Jan;50(1):67-87. [doi: 10.1080/16506073.2020.1795240] [Medline: 32870126]
6. Brown P, Waite F, Lambe S, et al. Automated virtual reality cognitive therapy (gamechange) in inpatient psychiatric wards: qualitative study of staff and patient views using an implementation framework. JMIR Form Res 2022 Apr;6(4):e34225. [doi: 10.2196/34225] [Medline: 35412462]
7. Ilioudi M, Wallström S, Steingrimsson S, Lindner P, Thunström AO, Ali L. Patient experience of a virtual reality calm room in a psychiatric inpatient care setting in Sweden: a qualitative study with inpatients. BMJ Open 2023 Dec 14;13(12):e076285. [doi: 10.1136/bmjopen-2023-076285] [Medline: 38101823]
8. Humbert A, Kohls E, Baldofski S, Epple C, Rummel-Kluge C. Acceptability, feasibility, and user satisfaction of a virtual reality relaxation intervention in a psychiatric outpatient setting during the COVID-19 pandemic. Front Psychiatry 2023;14:1271702. [doi: 10.3389/fpsyt.2023.1271702] [Medline: 37953932]
9. Lindner P, Rozental A, Jurell A, et al. Experiences of gamified and automated virtual reality exposure therapy for spider phobia: qualitative study. JMIR Serious Games 2020 Apr 29;8(2):e17807. [doi: 10.2196/17807] [Medline: 32347803]
10. Brito H, Pham T, Vicente B. Effect of sensorimotor rehabilitation based on an immersive virtual reality model on mental health. Int J Geriatr Psychiatry 2022 Jan;37(1). [doi: 10.1002/gps.5541] [Medline: 33797806]

XSL•FO
RenderX

11. Reger GM, Koenen-Woods P, Zetocha K, et al. Randomized controlled trial of prolonged exposure using imaginal exposure vs. virtual reality exposure in active duty soldiers with deployment-related posttraumatic stress disorder (PTSD). J Consult Clin Psychol 2016 Nov;84(11):946-959. [doi: 10.1037/ccp0000134] [Medline: 27606699]
12. Baghaei N, Chitale V, Hlasnik A, Stemmet L, Liang HN, Porter R. Virtual reality for supporting the treatment of depression and anxiety: scoping review. JMIR Ment Health 2021 Sep 23;8(9):e29681. [doi: 10.2196/29681] [Medline: 34554097]
13. Freeman D, Bradley J, Antley A, et al. Virtual reality in the treatment of persecutory delusions: randomised controlled experimental study testing how to reduce delusional conviction. Br J Psychiatry 2016 Jul;209(1):62-67. [doi: 10.1192/bjp.bp.115.176438] [Medline: 27151071]
14. Mölbert SC, Thaler A, Mohler BJ, et al. Assessing body image in anorexia nervosa using biometric self-avatars in virtual reality: attitudinal components rather than visual body size estimation are distorted. Psychol Med 2018 Mar;48(4):642-653. [doi: 10.1017/S0033291717002008] [Medline: 28745268]
15. Fagernäs S, Hamilton W, Espinoza N, Miloff A, Carlbring P, Lindner P. What do users think about Virtual Reality relaxation applications? A mixed methods study of online user reviews using natural language processing. Internet Interv 2021 Apr;24:100370. [doi: 10.1016/j.invent.2021.100370] [Medline: 33665136]
16. Pardini S, Gabrielli S, Dianti M, et al. The role of personalization in the user experience, preferences and engagement with virtual reality environments for relaxation. Int J Environ Res Public Health 2022 Jun 13;19(12):7237. [doi: 10.3390/ijerph19127237] [Medline: 35742483]
17. Anderson AP, Mayer MD, Fellows AM, Cowan DR, Hegel MT, Buckey JC. Relaxation with immersive natural scenes presented using virtual reality. Aerosp Med Hum Perform 2017 Jun 1;88(6):520-526. [doi: 10.3357/AMHP.4747.2017] [Medline: 28539139]
18. Gustafsson C, Dannapfel P. Leaders' experiences of successfully implementing health and welfare technology in sparsely populated Nordic areas. Disabil Rehabil Assist Technol 2025 Jan;20(1):85-96. [doi: 10.1080/17483107.2024.2352130] [Medline: 38727081]
19. Zander V, Gustafsson C, Landerdahl Stridsberg S, Borg J. Implementation of welfare technology: a systematic review of barriers and facilitators. Disabil Rehabil Assist Technol 2023 Aug;18(6):913-928. [doi: 10.1080/17483107.2021.1938707] [Medline: 34129802]
20. Batalden M, Batalden P, Margolis P, et al. Coproduction of healthcare service. BMJ Qual Saf 2016 Jul;25(7):509-517. [doi: 10.1136/bmjqs-2015-004315] [Medline: 26376674]
21. Spinuzzi C. The methodology of participatory design. Tech Commun 2005;52(2):163-174 [FREE Full text]
22. Kip H, Kelders SM, Bouman YHA, van Gemert-Pijnen LJEWC. The importance of systematically reporting and reflecting on eHealth development: participatory development process of a virtual reality application for forensic mental health care. J Med Internet Res 2019 Aug 19;21(8):e12972. [doi: 10.2196/12972] [Medline: 31429415]
23. Muñoz J, Mehrabi S, Li Y, et al. Immersive virtual reality exergames for persons living with dementia: user-centered design study as a multistakeholder team during the COVID-19 pandemic. JMIR Serious Games 2022 Jan 19;10(1):e29987. [doi: 10.2196/29987] [Medline: 35044320]
24. Dellazizzo L, Potvin S, Luigi M, Dumais A. Evidence on virtual reality-based therapies for psychiatric disorders: meta-review of meta-analyses. J Med Internet Res 2020 Aug 19;22(8):e20889. [doi: 10.2196/20889] [Medline: 32812889]
25. Riches S, Jeyarajaguru P, Taylor L, et al. Virtual reality relaxation for people with mental health conditions: a systematic review. Soc Psychiatry Psychiatr Epidemiol 2023 Jul;58(7):989-1007. [doi: 10.1007/s00127-022-02417-5] [Medline: 36658261]
26. Veling W, Lestestuiver B, Jongma M, Hoenders HJR, van Driel C. Virtual reality relaxation for patients with a psychiatric disorder: crossover randomized controlled trial. J Med Internet Res 2021 Jan 15;23(1):e17233. [doi: 10.2196/17233] [Medline: 33448933]
27. Botella C, Serrano B, Baños RM, García-Palacios A. Virtual reality exposure-based therapy for the treatment of post-traumatic stress disorder: a review of its efficacy, the adequacy of the treatment protocol, and its acceptability. Neuropsychiatr Dis Treat 2015;11:2533-2545. [doi: 10.2147/NDT.S89542]
28. Baños RM, Botella C, Guillen V, et al. An adaptive display to treat stress-related disorders: EMMA's World. Br J Guid Counc 2009 Aug;37(3):347-356. [doi: 10.1080/03069880902957064]
29. Polit DF, Beck CT. Nursing Research: Generating and Assessing Evidence for Nursing Practice: Lippincott Williams & Wilkins; 2008.
30. Creswell JW, Clark VLP. Designing and Conducting Mixed Methods Research: Sage Publications; 2017.
31. Ioannou M, Szabó Z, Widmark-Jensen M, Vyrinis G, Karlsson C, Steingrimsson S. Total sleep deprivation followed by bright light therapy as rapid relief for depression: a pragmatic randomized controlled trial. Front Psychiatry 2021;12:705090. [doi: 10.3389/fpsyt.2021.705090] [Medline: 34526921]
32. Sekhon M, Cartwright M, Francis JJ. Acceptability of healthcare interventions: an overview of reviews and development of a theoretical framework. BMC Health Serv Res 2017 Jan 26;17(1):88. [doi: 10.1186/s12913-017-2031-8] [Medline: 28126032]
33. Graneheim UH, Lundman B. Qualitative content analysis in nursing research: concepts, procedures and measures to achieve trustworthiness. Nurse Educ Today 2004 Feb;24(2):105-112. [doi: 10.1016/j.nedt.2003.10.001] [Medline: 14769454]

XSL•FO
RenderX

34. Riches S, Nicholson SL, Fialho C, et al. Integrating a virtual reality relaxation clinic within acute psychiatric services: a pilot study. Psychiatry Res 2023 Nov;329:115477. [doi: 10.1016/j.psychres.2023.115477] [Medline: 37802013]
35. Williams G, Drini E, Riches S. Virtual reality relaxation for mental health service users with complex care needs: a pilot study in residential care homes and supported accommodation. J Psychosoc Rehabil Ment Health 2025 Sep;12(3):353-364. [doi: 10.1007/s40737-024-00399-y]
36. Tan HLE, Chng CML, Lau Y, Klainin-Yobas P. Investigating the effects of a virtual reality-based stress management programme on inpatients with mental disorders: A pilot randomised controlled trial. Int J Psychol 2021 Jun;56(3):444-453. [doi: 10.1002/ijop.12713] [Medline: 32959431]
37. Klein Haneveld L, Kip H, Bouman YHA, Weerdmeester J, Scholten H, Kelders SM. Exploring the added value of virtual reality biofeedback game DEEP in forensic psychiatric inpatient care-a qualitative study. Front Psychol 2023;14:1201485. [doi: 10.3389/fpsyg.2023.1201485] [Medline: 38023054]
38. Arissen C, van der Helm L, Dijkstra B, Markus W. Virtual reality to support inpatient addiction treatment: patients are ready, what about therapists? A feasibility study. Int J Ment Health Addiction 2023 Dec;21(6):4085-4107. [doi: 10.1007/s11469-022-00843-9]
39. Chung OS, Johnson AM, Dowling NL, et al. Are Australian mental health services ready for therapeutic virtual reality? An investigation of knowledge, attitudes, implementation barriers and enablers. Front Psychiatry 2022;13:792663. [doi: 10.3389/fpsyt.2022.792663] [Medline: 35185649]
40. Lindner P, Miloff A, Zetterlund E, Reuterskiöld L, Andersson G, Carlbring P. Attitudes toward and familiarity with virtual reality therapy among practicing cognitive behavior therapists: a cross-sectional survey study in the era of consumer VR platforms. Front Psychol 2019;10:176. [doi: 10.3389/fpsyg.2019.00176] [Medline: 30800086]
41. Kip H, Buitelaar-Huijsse GKG, Kouijzer MTE, Kelders SM. From theory to implementation in practice: a qualitative case study of the implementation of virtual reality in mental healthcare. Glob Implement Res Appl 2024;4(1):66-88. [doi: 10.1007/s43477-023-00101-7]
42. González Moraga FR, Enebrink P, Perrin S, Sygel K, Veling W, Wallinius M. VR-assisted aggression treatment in forensic psychiatry: a qualitative study in patients with severe mental disorders. Front Psychiatry 2024;15:1307633. [doi: 10.3389/fpsyt.2024.1307633] [Medline: 38818023]
43. Hernandez MEH, Michalak E, Choudhury N, et al. Co-design of a virtual reality cognitive remediation program for depression (bWell-D) with patient end users and clinicians: qualitative interview study among patients and clinicians. JMIR Serious Games 2023 Apr 7;11:e43904. [doi: 10.2196/43904] [Medline: 37027183]
44. Monaghesh E, Samad-Soltani T, Farhang S. Virtual reality-based interventions for patients with paranoia: a systematic review. Psychiatry Res 2022 Jan;307:114338. [doi: 10.1016/j.psychres.2021.114338] [Medline: 34922239]
45. Sygel K, Wallinius M. Immersive virtual reality simulation in forensic psychiatry and adjacent clinical fields: a review of current assessment and treatment methods for practitioners. Front Psychiatry 2021;12:673089. [doi: 10.3389/fpsyt.2021.673089] [Medline: 34122189]

Abbreviations

VR: virtual reality

Edited by A Stone; submitted 16.04.25; peer-reviewed by M Tusconi; revised version received 23.11.25; accepted 24.11.25; published 29.12.25.

Please cite as:
Ilioudi M, Steingrimsson S, Wallström S, Osmanovic Thunström A, Lindner P, Ioannou M, Falk Ö, Ali L
Assessing Patient Acceptance of Being Offered a Virtual Calm Room in Psychiatric Inpatient Care: Mixed Methods Study
JMIR Form Res 2025;9:e75961
URL: https://formative.jmir.org/2025/1/e75961
doi:10.2196/75961

© Maria Ilioudi, Steinn Steingrimsson, Sara Wallström, Almira Osmanovic Thunström, Philip Lindner, Michael Ioannou, Örjan Falk, Lilas Ali. Originally published in JMIR Formative Research (https://formative.jmir.org), 29.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Pro- and Antifluoride Use Messages on YouTube in Japan: Content Analysis

Hikari Sophia Nagao[1], DDS; Tsuyoshi Okuhara[2], PhD; Hiroe Suzuki-Chiba[1], MPH, MD; Hiroko Okada[2], PhD; Takahiro Kiuchi[2], MD, PhD

[1]Department of Health Communication, Graduate School of Medicine, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan
[2]Department of Health Communication, School of Public Health, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan

Corresponding Author:
Hikari Sophia Nagao, DDS
Department of Health Communication, Graduate School of Medicine, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan

Abstract

Background: Dental caries is one of the most prevalent chronic conditions globally. In Japan, fluoride application—mainly via toothpaste, mouth rinses, and professional treatments—is a key preventive measure, as community water fluoridation is not implemented. Despite its proven effectiveness, fluoride use faces opposition from certain groups citing potential health risks. Social media platforms, especially YouTube, have become major sources of health information, but also facilitate the spread of misinformation, which may influence public perceptions and behaviors toward fluoride use.

Objective: This study aimed to analyze YouTube videos addressing fluoride use for caries prevention, focusing on the types of information presented and comparing the messages shared by proponents and opponents of fluoride use.

Methods: A comprehensive search was conducted on YouTube using fluoride-related keywords in Japanese. The top 50 videos for each keyword were screened, and after excluding irrelevant or duplicate content, 86 videos were analyzed. Videos were categorized as proponent (“pro”), opponent (“anti”), or others. The sources, intended audiences, and content themes were assessed. Interrater reliability was confirmed using the Cohen κ coefficient.

Results: Of the 86 analyzed videos, 58% (n=50) were categorized as “pro,” 22% (n=19) as “anti,” and 20% (n=17) as others. Proponent videos, mainly produced by dental professionals, emphasized scientific evidence, such as the mechanism of fluoride in preventing caries and guideline-based recommendations. Opponent videos, largely uploaded by laypersons, highlighted potential dangers of fluoride, including health risks and additives, and frequently promoted fluoride-free products. Opponent videos had higher daily viewership and engagement than proponent videos.

Conclusions: Anti-fluoride content on YouTube appears to reach broader audiences than expert-generated profluoride videos. Opponent messages tend to use emotionally charged communication, whereas proponents focus on scientific information. These differences in style may influence public perceptions of fluoride use. Public health professionals should develop engaging and accessible communication strategies to counter misinformation and promote evidence-based practices.

(JMIR Form Res 2025;9:e82265) doi:10.2196/82265

KEYWORDS

caries prevention; fluoride; health communication; misinformation; oral health; social media

Introduction

Dental caries is a pathological condition characterized by the progressive degradation of the hard tissues of the teeth instigated by acid production from bacterial metabolism. This process begins in the enamel and cementum, advances toward the dentin, and culminates in severe structural damage [1]. Dental caries remains a significant global public health challenge. According to the Global Burden of Disease Study 2021, major oral conditions such as untreated dental caries of permanent teeth, severe periodontitis, and edentulism affected an estimated 3.69 billion people worldwide in 2021 [2]. Among these conditions, untreated dental caries of permanent teeth was the single most prevalent specific condition [2]. Despite decades of preventive efforts, the global burden of oral conditions has changed little over the past 30 years [2]. These conditions exert profound adverse effects on the quality of life and socioeconomic systems [3].

Fluoride application is recognized as one of the most effective evidence-based measures for preventing dental caries [1]. Fluoride is widely recommended as a core strategy for caries prevention on the basis of long-standing evidence that, at recommended concentrations, its benefits outweigh potential risks [1]. The use of fluoride for the prevention of dental caries

XSL•FO
RenderX

is broadly categorized into systemic and topical applications. In many countries, systemic fluoride application is implemented through fluoridation, a public health strategy involving the addition of fluoride to tap water supplies for consumption as drinking water [4]. Topical fluoride application constitutes the primary preventive approach, including the use of fluoride-containing toothpaste and mouth rinses, alongside professional fluoride treatments applied directly to tooth surfaces [5]. In Japan, approximately one-third of the population is estimated to have untreated caries, particularly among older adults who retain more natural teeth later in life [6]. While community water fluoridation is not implemented in Japan [7], prevention largely depends on individual behaviors [5]. In addition, naturally occurring fluoride in tap water varies geographically within Japan, and higher fluoride levels have been associated with a lower prevalence of reported treatment for dental caries in children [8].

Despite extensive evidence supporting its safety and efficacy, fluoride use faces opposition from groups citing potential health risks [9-11]. While excessive fluoride ingestion is a risk factor for fluorosis [12], adverse effects are dose-dependent and unlikely to occur under typical recommended use conditions [13]. In some cases, opposition to fluoride is embedded in broader holistic positions that also reject vaccines and certain dental materials [14,15]. Historically, the patriarchal medical system in Japan has granted physicians significant authority over treatment decisions [16]. Consequently, dentists' opinions may play a critical role in shaping the course of treatment.

Recently, social media has become a primary source of health information, but it also facilitates the rapid spread of misinformation [17,18]. YouTube, one of the leading social media platforms in Japan with over 90% use among younger generations [19], is a key arena for this conflict. Sensational, simple, and emotionally charged content can elicit strong reactions and sharing [20,21], whereas information attributed to trusted experts is more credible and persuasive in encouraging health-related behaviors [22-24]. Similar dynamics have been observed in other public health topics, such as vaccine hesitancy, where misinformation can shape public perceptions and behaviors [25-27].

Previous studies have shown that anti-fluoride messages on social media are sometimes more prevalent than profluoride messages [28], and are often linked to the promotion of fluoride-free products [29]. On YouTube, professionally produced profluoride videos have lower viewership than popular antifluoride videos [30]. On Instagram, findings are mixed; while one study found more profluoride posts with widespread misinformation [31], other analyses identified more antifluoride content centered on conspiracy theories, health risks, and the promotion of natural or vegan fluoride-free products [32,33]. On Twitter (subsequently rebranded X), professional fluoride treatments are viewed positively, but discussions on water fluoridation are polarized [34]. The fluoride-free trend is largely driven by growing public interest in natural lifestyles [35].

However, to the best of our knowledge, very little research has analyzed the content of pro- and antifluoride messages on YouTube in Japan. This research gap is particularly important in Japan, where there is no community water fluoridation [7], and caries prevention depends largely on individual behaviors and parental consent. Parental opposition to fluoride is a barrier to fluoride treatments [36] and is linked to vaccine refusal [37], which may increase disease burden and health care costs [38]. As YouTube is a major platform for health information, this study aimed to evaluate the content of pro- and antifluoride YouTube videos on caries prevention.

The research question is outlined as follows: What specific content is disseminated by the proponents and opponents of fluoride use on YouTube videos?

Methods

Study Design

This study used quantitative content analysis to systematically review YouTube videos on fluoride use.

A total of 86 videos were deemed eligible for analysis. These videos were categorized as follows: 50 (58%) videos advocating fluoride use, 19 (22%) opposing fluoride use, and 17 (20%) presenting mixed or neutral viewpoints.

Search Strategy: Data Collection

We conducted internet searches on October 15, 2023, using keyword combinations input in Japanese text (Multimedia Appendix 1) with methodologies used in previous research analyzing YouTube videos on health-related information [39]. A comprehensive search was conducted on YouTube using a range of relevant keywords, including "fluoride," as well as additional terms identified via Google Trends, such as the following:

"fluoride AND coating," "fluoride AND processing," "dentist AND fluoride," "fluoride AND coat," "fluoride AND toothpaste," "fluoride AND pan coating," "fluoride AND resin," "fluoride AND danger," "fluoride AND rinse," and "fluoride."

The search results were sorted based on "view count," and the top 50 videos for each keyword were selected for analysis.

To minimize bias, all searches were performed using Google Chrome in incognito mode to prevent the influence of personal browsing history or tailored search results. After each search, we closed the browser and reopened it in incognito mode, and no personal Google accounts were used. For each search formula, the top 50 results were reviewed, and duplicate results were excluded. Results concerning no association with fluoride use for caries prevention (eg, content about Teflon-coated cookware), videos shorter than 1 minute in length, or videos presented in non-Japanese languages were excluded. TV news clips were also excluded to ensure a homogeneous sample by focusing on content originally created for the YouTube platform, rather than repurposed broadcast media.

Data Extraction

Basic Characteristics

The following basic information was collected from each video: the number of days since upload, total view count, number of likes, source, intended audience, and content. To assess

XSL•FO
RenderX

viewership trends, the number of views was divided by the number of days since upload to calculate the average number of views per day.

We identified the source of each video based on the YouTube channel that uploaded it and categorized it into 6 groups: dentist or dental hygienist, local government, company, dental association, layperson, and unknown.

Content Characteristics

The first author (HSN) evaluated the intended audience for the video and categorized it as either the general public or dental hygienists. HSN reviewed the content of each video and categorized it into one of the 3 groups: pro, anti, or others based on the stance toward fluoride use for caries prevention. Multimedia Appendix 2 outlines the definitions of these categories. Subsequently, HSN categorized the videos according to their source and assessed the prevalence of advantages and disadvantages. The categorization according to their source was conducted only for "pro" and "anti" opinions, and "others" were not included.

The content of the videos was subsequently analyzed according to their source, with a focus on identifying trends and patterns in the stance toward fluoride use for caries prevention. HSN coded the videos in terms of their content characteristics and extracted fundamental information about the videos, including their main characters and opinions.

Themes and Codes

HSN watched the extracted videos in detail and inductively assigned categories and themes to the extracted descriptions of opinions on fluoride. The analysis generating themes was conducted only for "pro" and "anti" opinions, and "others" were not included. Multiple categories were selected for videos featuring two or more categories.

HSN extracted notable descriptions from each video and entered them into a Microsoft Excel spreadsheet. For each video, HSN generated a rephrasing of each notable description by briefly summarizing it. HSN then generated a theme name that summarized the rephrased descriptions and defined the themes. HSN used this process for all the videos. Theme names and definitions were updated during analysis. HSN outlined the final categories and themes, accompanied by the corresponding numbers and percentages of videos per category and theme.

Intercoder Reliability

To determine intercoder reliability, 2 independent coders (HSN and HS) assessed approximately 26% (23/86) videos in the final dataset that were selected at random. Using the codes and definitions of themes created by the first author (HSN), the third author (HS) evaluated (1) the themes ascribed to the videos and (2) the codes to which the themes were assigned. HS was instructed on the criteria for theme extraction and categorization for each theme during a meeting with HSN that lasted for approximately 1 hour. The discussion between HSN and HS was based on a coding manual. In the event of a discrepancy between the evaluations of HSN and HS, we engaged in a discussion to reach a consensus on the final theme and classification.

Statistical Analysis

Descriptive statistics, including means, ranges, and percentages, were computed for each variable in the basic video information dataset. This analysis provided a comprehensive overview of metrics, such as the average number of views per day, total view count, and distribution of video sources and content codes. To assess the reliability of the coding, the Cohen κ coefficient was used to evaluate the interrater agreement. The analysis was conducted using R for Mac (version 4.4.2; R Foundation for Statistical Computing).

Ethical Considerations

This study did not involve human or animal participants, and all data were collected from publicly available web-based sources. To maintain anonymity, all identifying information, such as YouTube user IDs, was deidentified during data reporting. This study was granted an exemption from ethical approval by the ethics review committee of the Graduate School of Medicine, University of Tokyo.

Results

Table 1 summarizes the video characteristics. The average values and ranges (minimum to maximum) of the videos are listed in Table 1. Of all the videos included, 58% (50/86) were categorized as pro, 22% (19/86) as anti, and 20% (17/86) as others. The number of days since upload ranged from 157 to 3331 days for proponent videos, 296 to 2437 days for opponent videos, and 563 to 4177 days for other videos. The highest view count was observed among the proponent videos, exceeding that of the opponent videos (Figure 1).

XSL•FO
RenderX

Figure 1. Search process for YouTube videos on fluoride use.

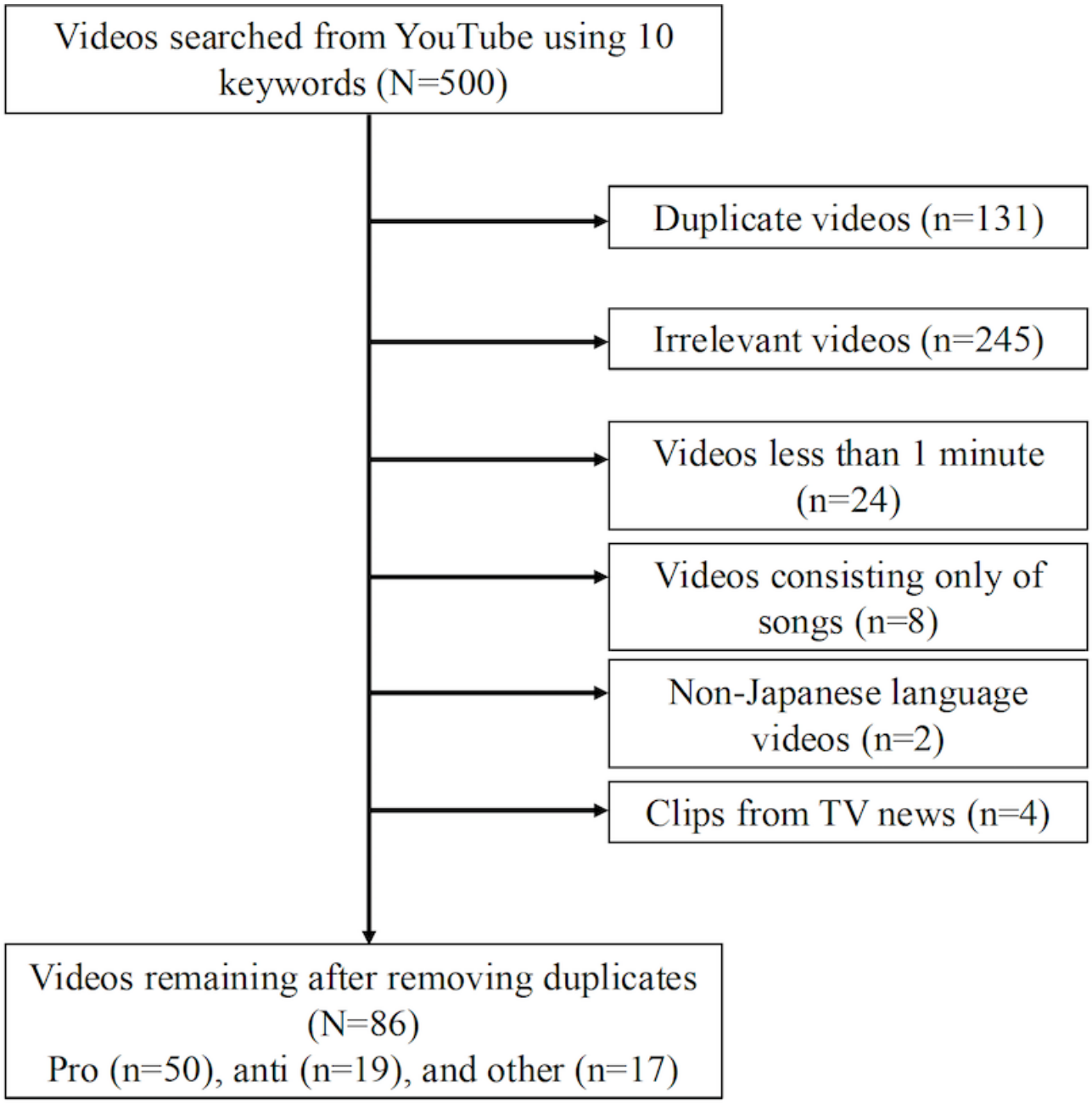

XSL•FO
RenderX

Table . Characteristics of the analyzed YouTube videos categorized by their stance on fluoride use (N=86) .

Characteristics	Pro	Anti	Other
Basic information			
Number of videos included	50	19	17
Days since upload			
Maximum	3331	2437	4177
Minimum	157	296	563
Average	1200	890	1465
Number of views			
Maximum	2,475,818	283,122	4,111,810
Minimum	91	986	16,731
Average	1,48,003	70,594	5,88,462
Views per day			
Maximum	1425	327	7303
Minimum	0	0	16
Average	147	104	748
Number of likes			
Maximum	13,319	4312	7671
Minimum	0	8	82
Average	969	1434	1781
Video length (s)			
Maximum	3616	1169	1247
Minimum	66	164	79
Average	597	624	541
Video creators			
Dentists and dental hygienists	30	1	13
Local government	2	0	0
Companies	2	0	0
Dental association	1	0	0
Layperson	15	16	4
Unknown	0	2	0
Target audience			
General public	49	19	17
Dental hygienists	1	0	0

In terms of the average daily views, calculated by dividing the total view count by the number of days since upload, 18% (9/50) of the proponent videos had fewer than 1 view per day, compared to only 5% (1/19) of the opponent videos and zero videos in the other sources. Conversely, videos with more than 19 views per day accounted for 44% (22/50) of the proponent videos, 94.7% (18/19) of the opponent videos, and 76.5% (13/17) of other videos.

In terms of the sources of the videos, 60% (30/50) of the proponent videos were uploaded by dentists or dental hygienists, 4% (2/50) videos by the local government, 4% (2/50) videos by companies, 2% (1/50) video by a dental association, and 30% (15/50) videos by laypersons. Of all the opponent videos, 5% (1/19) were uploaded by a dentist, 84% (16/19) by laypersons, and 11% (2/19) by unidentified sources.

Regarding the intended audience, 98% (49/50) of the proponent videos were directed at the general public, whereas 2% (1/50) video targeted dental hygienists. Conversely, 100% of the opponent videos were directed at the general public.

The first author (HSN) coded the mechanism of fluoride action, prevention, product introduction, evidence or guidelines, concentration, fluoride use, regular check-ups, safety confirmation, cost performance or cost-effectiveness, dietary

XSL•FO
RenderX

habits, additives other than fluoride, dental implants, and trust relationships as the themes of the proponent video. In addition, HSN coded additives other than fluoride, product introduction, no confirmed safety, potential dangers, alternatives, and danger of Teflon coating as the themes of the opponent video.

Intercoder agreement was achieved in 332 of the 341 coding instances across the 23 (97.4%) selected videos. The Cohen κ statistic ranged from 0.91 to 0.95, with a mean of 0.93, indicating substantial intercoder reliability.

Table 2 shows a detailed breakdown of the themes addressed in the videos, frequency of discussion, and number and percentage of sources contributing to each category. This summary highlights the thematic focus and engagement of different sources by proponents, opponents, and others.

XSL•FO
RenderX

Table . Code definitions and video source

Category	Theme	Examples	Frequency, n (%)	Sources, n (%)
Proponent				
Mechanism of fluoride action	Remineralization of enamel, Surface strengthening	Tooth remineralization, Strengthening the surface	33 (64)	Dentists and dental hygienists, 30 (91)
				Local government, 1 (3)
				Layperson, 2 (6)
Prevention	Fluoride involvement in caries prevention	Prevents tooth decay	24 (48)	Dentists and dental hygienists, 18 (75)
				Local government, 1 (4)
				Company, 2 (8)
				Dental association, 1 (4)
				Layperson, 2 (8)
Product introduction	Introduction of specific products	Check-Up and Systema (LION Dental Products Company Limited)	24 (48)	Dentists and dental hygienists, 21 (88)
				Company, 2 (8)
				Layperson, 1 (4)
Evidence/guidelines	Specific guidelines based on evidence	Evidence, Guidelines	24 (48)	Dentists and dental hygienists, 19 (79)
				Local government, 1 (4)
				Company, 2 (8)
				Dental association, 1 (4)
				Layperson, 1 (4)
Concentration	Fluoride toothpaste concentration based on age, high-concentration fluoride	1500 ppm for ages 6 and older	22 (44)	Dentists and dental hygienists, 16 (73)
				Local government, 1 (5)
				Company, 1 (4.5)
				Dental association, 1 (5)
				Layperson, 3 (14)
How to use fluoride	Amount of rinse water, usage per time, usage frequency per day	Rinse with a small amount of water, use toothpaste at least twice a day	16 (32)	Dentists and dental hygienists, 11 (69)
				Local government, 1 (6)
				Dental association, 1 (6)
				Layperson, 1 (6)
Regular check-ups	Regular check-ups at dental clinics	Let's go to regular check-ups at the dentist	13 (26)	Dentists and dental hygienists, 9 (69)
				Dental association, 1 (4)
				Layperson, 3 (23)
Safety confirmed	The safety of fluoride has been confirmed	Safety	10 (20)	Dentists and dental hygienists, 8 (80)
				Dental association, 1 (10)
				Layperson, 1 (10)
Cost performance/cost-effectiveness	Good cost performance for caries prevention, excellent cost-effectiveness	Fluoride mouthwash has good cost performance and is an effective prevention method	6 (12)	Dentists and dental hygienists, 3 (50)
				Company, 2 (33)
				Layperson, 1 (17)

XSL•FO
RenderX

Category	Theme	Examples	Frequency, n (%)	Sources, n (%)
Dietary habits	Reduce sugar intake, reduce snack frequency	Reduce sugar intake, reconsider the number of snacks	4 (8)	Dentists and dental hygienists, 1 (25) Layperson, 3 (75)
Additives other than fluoride	Sodium lauryl sulfate, triclosan, etc.	The foaming agent sodium lauryl sulfate is also found in detergents and is highly irritating	2 (4)	Dentists and dental hygienists, 2 (100)
Dental implants	Can also be used with dental implants	Commercial fluoride toothpaste can also be used with dental implants	2 (4)	Dentists and dental hygienists, 2 (100)
Trust relationship	Trust relationship between dentists and patients	The trust relationship between dentists and patients is important	1 (2)	Dentists and dental hygienists, 1 (100)
Opponent				
Additives other than fluoride	Additive-free, Free from additives	Additives are dangerous, strive for an additive-free lifestyle	14 (74)	Dentists and dental hygienists, 1 (7) Layperson 11 (79) Unknown, 2 (14)
Product introduction	Introduction of specific products	Shabondama Sekken Toothpaste (Shabondama Soap) and Pax Naturon (Taiyo Yushi Corp.)	10 (53)	Layperson, 8 (80) Unknown, 2 (20)
No safety confirmed	Fluoride is harmful, not safe	Considered dangerous by international organizations such as WHO, banned in some countries	7 (37)	Dentists and dental hygienists, 1 (14) Layperson, 5 (71) Unknown, 1 (14)
Potential dangers	Impedes child development, causes diseases	Impedes child development, causes Down syndrome, calcifies the pineal gland, etc	6 (32)	Layperson, 4 (67) Unknown, 2 (33)
Alternatives	Alternatives used when not using fluoride	Brushing with baking soda, salt, etc	6 (32)	Layperson, 6 (100)
Danger of Teflon coating	Dangers of Teflon-coated cookware	Teflon-coated frying pans are dangerous	2 (11)	Layperson, 2 (100)

[a]WHO: World Health Organization.

Among the proponent videos, the most-addressed category was the mechanism by which fluoride prevents caries, which was featured in 33 (64%) videos. Other frequently mentioned categories included information about the role of fluoride in caries prevention (24/50, 48%), specific product introductions (24/50, 48%), and evidence supporting the use of fluoride for caries prevention (24/50, 48%). Less commonly discussed categories included cost-effectiveness (n=6/50, 12%) and the trust relationship between dentists and patients (1/50, 2%). After completing the research, we translated all terms into English for the purpose of this report. Among the proponent sources, all videos uploaded by dentists and dental hygienists discussed the mechanism of action of fluoride, 63% (19/30) provided evidence for fluoride use in caries prevention, and 60% (18/30) provided information on the preventive role of fluoride. Videos from laypersons discussed categories such as regular dental checkups (3/15, 20%), dietary practices for caries prevention (3/15, 20%), and fluoride safety (1/15, 7%).

In the opponent videos, the most frequently addressed category was additives other than fluoride, which was mentioned in 14 (74%) videos. This was followed by specific product introductions (10/19, 53%) and claims regarding the lack of safety of fluoride (7/19, 37%). Alternatives to fluoride-containing toothpaste were discussed in 6 (32%) videos. A small number of videos (2/19, 11%) focused on the risks associated with nonfluoride substances, such as Teflon coatings, which were not associated with caries prevention. Among the opponent sources, laypersons frequently addressed additives other than fluoride (11/16, 69%), specific product introductions (8/16, 50%), and claims about the lack of safety of (7/16, 31%). Only 1 (5%) dentist discussed additives other than fluoride and alleged safety concerns regarding fluoride.

XSL•FO
RenderX

Discussion

Principal Findings

This study analyzed fluoride-related content on YouTube and categorized videos into proponent and opponent stances to evaluate the nature of the information being disseminated. Results revealed that the number of videos produced by proponents surpassed those produced by opponents, aligning with the trends identified in previous studies on fluoride-related content on Instagram [35]. However, previous studies analyzed the content of fluoride-related posts on YouTube and Instagram, which exhibited strong antifluoride tendencies [33,34,36]. Beyond differences in language and search timing, Japan's sociocultural and policy context is also likely to have influenced these findings. Unlike many countries where community water fluoridation is implemented, Japan does not add fluoride to public water supplies [7]. As a result, caries prevention relies largely on individual and parental decisions (eg, whether to accept topical fluoride or purchase fluoride-containing products) rather than on passive, population-wide exposure. In this setting, messages portraying fluoride as unsafe may have a more direct impact on behavior. Moreover, because medical professionals in Japan traditionally hold strong authority in treatment decisions [16], statements made by dentists, including those expressing skepticism toward fluoride or endorsing holistic positions, may exert substantial influence even when not strongly evidence-based.

Proponent videos displayed a relatively high proportion (9/50, 18%) of content with fewer than 1 average daily view, compared to only 5% (1/19) for opponent videos, suggesting that opponent videos achieved more consistent viewership. Furthermore, 95% (18/19) of the opponent videos garnered more than 19 daily views, compared to only 44% (22/50) of proponent videos. This indicates that the opponent content may have reached a broader audience. Videos with over 1000 likes constituted 37% (7/19) of the opponent videos compared to 14% (7/50) of the proponent videos. Notably, 10% (5/50) of the proponent videos did not permit likes to be counted, potentially limiting their ability to visibly demonstrate audience support.

The content themes of proponent and opponent videos generally mirrored those identified in previous studies [30,32]. Most proponent videos, created predominantly by dental professionals (30/50, 60%), focused on themes such as the "mechanism of fluoride in preventing caries" (32/50, 64%), the "role of fluoride in caries prevention" (24/50, 48%), and "evidence-based guidelines" (24/50, 48%). These videos emphasize scientific and objective information. Among the extracted videos, almost all uploaded by dentists and dental hygienists mentioned the "mechanism of action of fluoride" (29/30, 97%). The caries prevention mechanism of fluoride is associated with various processes, including enamel remineralization and its effects on oral microorganisms, which are consistent with established facts regarding caries prevention [1]. This suggests that dental professionals produce persuasive content based on evidence and expertise. Studies indicate that expert-driven health communication is more effective in influencing individuals to adopt healthy behaviors [22-24]. However, the objective and fact-based nature of such content may limit its engagement potential for general audiences [20,21], thereby reducing its dissemination.

By contrast, videos by nonmedical individuals or organizations frequently highlighted themes such as dietary habits for caries prevention (3/15, 20%), the importance of regular dental checkups (3/15, 20%), and general claims regarding the safety of fluoride (1/15, 7%). These themes are often presented without supporting evidence, thereby reducing their credibility.

Opponent videos predominantly focused on concerns regarding "additives other than fluoride" (14/19, 74%), fluoride being "unsafe" (7/19, 37%), and fluoride being "dangerous" (6/19, 32%). These messages often align with broader preferences for natural or organic products and lifestyles that avoid synthetic additives [33]. Many of these videos also promoted specific fluoride-free products, subtly influencing consumer choice. Although excessive use of fluoride can lead to fluorosis [12], harmful effects are dose-dependent, and adverse outcomes are rare in typical use conditions [13]. Opponent videos that portray fluoride as "unsafe" or "dangerous" unnecessarily provoke fear, spread misinformation, and reinforce misconceptions.

In addition, 32% (6/19) of the opponent videos discussed alternatives to fluoride-containing products, providing practical guidance that may appeal to viewers. However, such content often serves as a marketing strategy to promote specific products [33,35]. Many topics mentioned in the videos were consistent with those in previous studies. However, new findings were observed in this study. Notably, 1 (15%) antifluoride video created by a dentist included content that not only opposed fluoride use but also criticized treatments involving metal prostheses, such as full metal crowns. This finding suggests a potential link between antifluoride sentiments and a broader skepticism toward established, evidence-based dental practices. Previous studies have stated insufficient evidence to support or refute the superiority of metal-free materials over metal-ceramics or other standard restorations for fixed prosthetic treatments [40]. Metal prostheses are a treatment option covered by public health insurance in Japan. Given the high level of trust placed on health care professionals, particularly in Japan, where medical decisions are often heavily influenced by doctors' opinions [16], such videos may significantly influence the decision-making processes of patients.

Unlike previous studies, this study did not find any videos from either the pro- or antifluoride groups that mentioned water fluoridation or fluoride supplementation. This is likely because these fluoride application methods are not used in Japan [7]. Although easy access to health information brings many benefits, it also makes it easier for false, biased, or harmful content to spread [18]. Opponent videos often leverage sensational narratives to emphasize fluoride's alleged risks, trigger strong emotions, and motivate viewers to share content or join discussions [20].

Social media platforms also enable the formation of echo chambers in which individuals with similar beliefs reinforce each other's views, thereby further perpetuating misinformation [26]. The dissemination of such misconceptions can hinder the active use of fluoride by the general public, particularly in Japan,

XSL•FO
RenderX

where there are no opportunities to receive passive fluoride application methods, such as water fluoridation. In addition, negative attitudes toward fluoride have led some parents to refuse fluoride treatment for their children [36], and these individuals often reject vaccinations [37]. This rejection of preventive health measures, fueled by misinformation on platforms such as YouTube, poses significant risks, potentially increases the prevalence of preventable diseases, escalates health care costs, and causes unnecessary suffering for individuals and families [38].

Limitations and Strengths

This study is the first attempt at conducting a content analysis of Japanese YouTube videos addressing fluoride use for caries prevention, categorizing information into proponent and opponent perspectives. However, this study has several limitations.

First, it focused exclusively on Japanese YouTube videos, limiting the findings to information disseminated in Japanese and excluding perspectives from other cultural and linguistic contexts. Nevertheless, the results aligned with those of previous studies that analyzed fluoride-related social media content in other languages [35]. Second, our analysis was restricted to publicly available video sequences. YouTube also hosts "unlisted" videos, which are accessible only via direct links, and "private" videos, which are visible solely to individuals authorized by the uploader. As these visibility settings can change at any time, we excluded videos shared in closed or restricted communities. However, focusing on publicly accessible videos remains a significant endeavor, as they are most likely to appear in search results and reach a broad audience. Third, qualitative evaluation of video content may involve a degree of subjectivity. Although this study implemented a categorization process that required evaluative judgment, triangulation by multiple researchers was conducted to minimize bias and validate the findings. Finally, the findings are specific to the data collection period (October 15, 2023). Social media content is dynamic, and trends may shift over time or in response to current events, potentially leading to different patterns in other periods.

Despite these limitations, this study is noteworthy for its originality, as it offers a comprehensive content analysis of fluoride-related information on Japanese YouTube videos categorized by proponent and opponent perspectives.

Conclusions

This study provides a comparative analysis of proponent and opponent content associated with fluoride use for caries prevention on YouTube. Proponent videos outnumbered opponent videos; the latter demonstrated a more consistent reach among viewers, achieving higher average daily views and a greater proportion of likes. Opponent videos relied on emotionally charged, sensational, and straightforward messages, such as "fluoride is harmful to the human body," which provoked anxiety, encouraged empathy, and facilitated content sharing.

By contrast, proponent videos, predominantly produced by experts, such as dentists and dental hygienists, disseminate accurate, evidence-based information. However, these videos lack the emotional engagement necessary to capture the viewers' attention and effectively foster resonance.

The potential consequences of such misinformation are significant. Parents opposing fluoride use have been shown to reject not only fluoride treatments for caries prevention but also vaccinations, reflecting a broader resistance to preventive health care measures. This could lead to an increased disease burden, higher healthcare costs, and unnecessary suffering from preventable illnesses among children and communities.

To counter these challenges, it is imperative to develop communication strategies that amplify the impact of proponent messages. These strategies should balance scientific rigor with accessible, practical, and engaging content that resonates with the viewers' daily lives. Moreover, since antifluoride content often functions as a marketing strategy to promote specific fluoride-free products, it is imperative that communication strategies also counter these commercial motivations behind misinformation. Efforts must prioritize expanding the reach of accurate information to effectively combat the spread of misinformation and support public health initiatives.

Acknowledgments

The authors used generative artificial intelligence in a limited way, to clarify topic sentences and improve grammar, punctuation, and concision. The authors have reviewed each suggestion before adopting them.

Funding

This work was supported by the Japan Society for the Promotion of Science KAKENHI (25K13483).

Data Availability

The data presented in this study are available upon reasonable request to the corresponding author.

Authors' Contributions

Conceptualization: HSN, TO
Formal analysis: HSN, HS
Funding acquisition: HO

XSL•FO
RenderX

Methodology: HSN, TO
Supervision: TK
Writing – original draft: HSN
Writing – review & editing: all authors

Conflicts of Interest

None declared.

Multimedia Appendix 1
Japanese keyword combinations used in the YouTube search.
[DOCX File, 17 KB - formative_v9i1e82265_app1.docx]

Multimedia Appendix 2
Coding guidelines.
[DOCX File, 12 KB - formative_v9i1e82265_app2.docx]

References

1. Selwitz RH, Ismail AI, Pitts NB. Dental caries. Lancet 2007 Jan;369(9555):51-59. [doi: 10.1016/S0140-6736(07)60031-2] [Medline: 17208642]
2. GBD 2021 Oral Disorders Collaborators. Trends in the global, regional, and national burden of oral conditions from 1990 to 2021: a systematic analysis for the Global Burden of Disease Study 2021. Lancet 2025 Mar 15;405(10482):897-910. [doi: 10.1016/S0140-6736(24)02811-3] [Medline: 40024264]
3. Schwendicke F, Dörfer CE, Schlattmann P, Foster Page L, Thomson WM, Paris S. Socioeconomic inequality and caries: a systematic review and meta-analysis. J Dent Res 2015 Jan;94(1):10-18. [doi: 10.1177/0022034514557546] [Medline: 25394849]
4. Iheozor-Ejiofor Z, Worthington HV, Walsh T, et al. Water fluoridation for the prevention of dental caries. Cochrane Database Syst Rev 2015 Jun 18;2015(6):CD010856. [doi: 10.1002/14651858.CD010856.pub2] [Medline: 26092033]
5. Marinho VCC, Higgins JPT, Logan S, Sheiham A. Topical fluoride (toothpastes, mouthrinses, gels, or varnishes) for preventing dental caries in children and adolescents. Cochrane Database Syst Rev 2003;2003(4):CD002782. [doi: 10.1002/14651858.CD002782] [Medline: 14583954]
6. 2022 Survey of Dental Diseases: Ministry of Health, Labour and Welfare. 2022. URL: https://www.mhlw.go.jp/content/10804000/001112405.pdf [accessed 2024-09-27]
7. One in a million. The facts about fluoridation. 3rd. The British Fluoridation Society. 2012. URL: https://bfsweb.org/one-in-a-million [accessed 2024-09-27]
8. Matsuyama Y, Fujiwara T, Aida J. Tap water natural fluoride and parent-reported experience of child dental caries in Japan: evidence from a nationwide birth cohort survey. Community Dent Oral Epidemiol 2023 Dec;51(6):1141-1149. [doi: 10.1111/cdoe.12847] [Medline: 36779447]
9. Mittelstaedt M. Why water shouldn't be fluoridated: an interview with Dr Paul Connett. Fluoride Action Network. URL: http://fluoridealert.org/content/connett_october2010 [accessed 2012-08-06]
10. Miranda GHN, Alvarenga MOP, Ferreira MKM, et al. A systematic review and meta-analysis of the association between fluoride exposure and neurological disorders. Sci Rep 2021 Nov;11(1):22659. [doi: 10.1038/s41598-021-99688-w] [Medline: 34811523]
11. Guth S, Hüser S, Roth A, et al. Toxicity of fluoride: critical evaluation of evidence for human developmental neurotoxicity in epidemiological studies, animal experiments and in vitro analyses. Arch Toxicol 2020 May;94(5):1375-1415. [doi: 10.1007/s00204-020-02725-2] [Medline: 32382957]
12. Whelton HP, Spencer AJ, Do LG, Rugg-Gunn AJ. Fluoride revolution and dental caries: evolution of policies for global use. J Dent Res 2019 Jul;98(8):837-846. [doi: 10.1177/0022034519843495] [Medline: 31282846]
13. Grandjean P. Developmental fluoride neurotoxicity: an updated review. Environ Health 2019 Dec 19;18(1):110. [doi: 10.1186/s12940-019-0551-x] [Medline: 31856837]
14. K. Srinivasan SC. Holistic dentistry: Natural approaches to oral health. Sch Bull 2015;1(10):267-270 [FREE Full text]
15. Nidhi Thakur AB, Keluskar V. Holistic dentistry: natural approaches to oral health. J Int Oral Health 2011;3(2):9-12 [FREE Full text]
16. Specker Sullivan L. Dynamic axes of informed consent in Japan. Soc Sci Med 2017 Feb;174:159-168. [doi: 10.1016/j.socscimed.2016.12.031] [Medline: 28043018]
17. Antheunis ML, Tates K, Nieboer TE. Patients' and health professionals' use of social media in health care: motives, barriers and expectations. Patient Educ Couns 2013 Sep;92(3):426-431. [doi: 10.1016/j.pec.2013.06.020] [Medline: 23899831]
18. Karamitros GA, Kitsos NA, Sapountzis S. Systematic review of quality of patient information on phalloplasty in the internet. Aesthetic Plast Surg 2017 Dec;41(6):1426-1434. [doi: 10.1007/s00266-017-0937-5] [Medline: 28698939]

XSL•FO
RenderX

19. 2023 survey on media usage time and information behavior: Ministry of Internal Affairs and Communications [Report in Japanese]. : Ministry of Internal Affairs and Communications, Japan; 2023 URL: https://www.soumu.go.jp/main_content/000953019.pdf [accessed 2025-11-27]
20. Munaro AC, Hübner Barcelos R, Francisco Maffezzolli EC, Santos Rodrigues JP, Cabrera Paraiso E. To engage or not engage? The features of video content on YouTube affecting digital consumer engagement. J Consum Behav 2021 Sep;20(5):1336-1352 [FREE Full text] [doi: 10.1002/cb.1939]
21. Liu AX, Xie Y, Zhang J. It's not just what you say, but how you say it: the effect of language style matching on perceived quality of consumer reviews. Journal of Interactive Marketing 2019 May;46(1):70-86. [doi: 10.1016/j.intmar.2018.11.001]
22. Okada H, Okuhara T, Goto E, Kiuchi T. Associations between vaccination behavior and trust in information sources regarding COVID-19 vaccines under emergency approval in Japan: a cross-sectional study. Vaccines (Basel) 2023 Jan;11(2):233. [doi: 10.3390/vaccines11020233] [Medline: 36851111]
23. Freed GL, Clark SJ, Butchart AT, Singer DC, Davis MM. Sources and perceived credibility of vaccine-safety information for parents. Pediatrics 2011 May;127 Suppl 1:S107-S112. [doi: 10.1542/peds.2010-1722P] [Medline: 21502236]
24. Kareklas I, Muehling DD, Weber TJ. Reexamining health messages in the digital age: a fresh look at source credibility effects. J Advert 2015 Apr;44(2):88-104. [doi: 10.1080/00913367.2015.1018461]
25. Dunn AG, Leask J, Zhou X, Mandl KD, Coiera E. Associations between exposure to and expression of negative opinions about human papillomavirus vaccines on social media: an observational study. J Med Internet Res 2015 Jun;17(6):e144. [doi: 10.2196/jmir.4343] [Medline: 26063290]
26. Collier R. Containing health myths in the age of viral misinformation. CMAJ 2018 May;190(19):E578. [doi: 10.1503/cmaj.180543] [Medline: 29759962]
27. Terada M, Okuhara T, Nagasawa T, Okada H, Goto E, Kiuchi T. Public perception of the resumption of HPV vaccine recommendation in Japan: Twitter content analysis. Health Promot Int 2023 Dec;38(6):daad153. [doi: 10.1093/heapro/daad153] [Medline: 37966160]
28. Mertz A, Allukian M. Community water fluoridation on the internet and social media. J Mass Dent Soc 2014;63(2):32-36. [Medline: 25230407]
29. Seymour B, Getman R, Saraf A, Zhang LH, Kalenderian E. When advocacy obscures accuracy online: digital pandemics of public health misinformation through an antifluoride case study. Am J Public Health 2015 Mar;105(3):517-523. [doi: 10.2105/AJPH.2014.302437] [Medline: 25602893]
30. Basch CH, Blankenship EB, Goff ME, et al. Fluoride-related YouTube videos: a cross-sectional study of video contents by upload sources. J Dent Hyg 2018 Dec;92(6):47-53. [Medline: 30643004]
31. Esmaeilzadeh F, Movahhed T, Hasani Yaghooti MR, Hoseinzadeh M, Babazadeh S. Content analysis of fluoride-related posts on Instagram. BMC Oral Health 2024 Oct;24(1):1179. [doi: 10.1186/s12903-024-04913-3] [Medline: 39367338]
32. Basch CH, Milano N, Hillyer GC. An assessment of fluoride related posts on Instagram. Health Promot Perspect 2019;9(1):85-88. [doi: 10.15171/hpp.2019.11] [Medline: 30788272]
33. Lotto M, Sá Menezes T, Zakir Hussain I, et al. Characterization of false or misleading fluoride content on Instagram: infodemiology study. J Med Internet Res 2022 May;24(5):e37519. [doi: 10.2196/37519] [Medline: 35588055]
34. Mackert M, Bouchacourt L, Lazard A, et al. Social media conversations about community water fluoridation: formative research to guide health communication. J Public Health Dent 2021 Jun;81(2):162-166. [doi: 10.1111/jphd.12404] [Medline: 33058200]
35. Lotto M, Zakir Hussain I, Kaur J, Butt ZA, Cruvinel T, Morita PP. Analysis of fluoride-free content on Twitter: topic modeling study. J Med Internet Res 2023 Jun 20;25:e44586. [doi: 10.2196/44586] [Medline: 37338975]
36. Carpiano RM, Chi DL. Parents' attitudes towards topical fluoride and vaccines for children: Are these distinct or overlapping phenomena? Prev Med Rep 2018 Jun;10:123-128. [doi: 10.1016/j.pmedr.2018.02.014] [Medline: 29755930]
37. Chi DL. Caregivers who refuse preventive care for their children: the relationship between immunization and topical fluoride refusal. Am J Public Health 2014 Jul;104(7):1327-1333. [doi: 10.2105/AJPH.2014.301927] [Medline: 24832428]
38. Chi DL, Basson AA. Surveying dentists' perceptions of caregiver refusal of topical fluoride. JDR Clin Trans Res 2018 Jul;3(3):314-320. [doi: 10.1177/2380084418761846] [Medline: 30938597]
39. Furukawa E, Okuhara T, Okada H, et al. Evaluating the understandability and actionability of online educational videos on pre-dialysis chronic kidney disease. Nephrology (Carlton) 2023 Nov;28(11):620-628. [doi: 10.1111/nep.14226] [Medline: 37591493]
40. Poggio CE, Ercoli C, Rispoli L, Maiorana C, Esposito M. Metal-free materials for fixed prosthodontic restorations. Cochrane Database Syst Rev 2017 Dec;12(12):CD009606. [doi: 10.1002/14651858.CD009606.pub2] [Medline: 29261853]

XSL•FO
RenderX

Edited by A Mavragani; submitted 12.08.25; peer-reviewed by J Aida; revised version received 29.10.25; accepted 14.11.25; published 29.12.25.

Please cite as:
Nagao HS, Okuhara T, Suzuki-Chiba H, Okada H, Kiuchi T
Pro- and Antifluoride Use Messages on YouTube in Japan: Content Analysis
JMIR Form Res 2025;9:e82265
URL: https://formative.jmir.org/2025/1/e82265
doi:10.2196/82265

© Hikari Sophia Nagao, Tsuyoshi Okuhara, Hiroe Suzuki-Chiba, Hiroko Okada, Takahiro Kiuchi. Originally published in JMIR Formative Research (https://formative.jmir.org), 29.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Implementing Digital Tools for Mental Health Support in Young Individuals in Colombia: Mixed Methods Feasibility Study

Laura Ospina-Pinillos[1*], MD, PhD; Débora L Shambo-Rodríguez[1*], BSc; María Isabel Riaño-Fonseca[1*], MSc; Mónica Natalí Sánchez Nítola[1*], PhD; María Fernanda Ramírez-Castro[2*], MPH; María Gabriela Calvo-Valderrama[1*], MD; Salvador Camacho[3,4*], MBA, MSc, PhD; Carlos Gómez-Restrepo[1,5*], MD, MSc, PhD; Alvaro A Navarro-Mancilla[6,7*], MBA, MPH, MD; Ian B Hickie[8*], MD; Jo-An Occhipinti[8,9*], PhD

[1]Department of Psychiatry and Mental Health, School of Medicine, Pontificia Universidad Javeriana, Carrera 7 # 40-62, 8th Floor, Building Hospital Universitario San Ignacio, Bogotá, Colombia

[2]School of Psychology, Pontificia Universidad Javeriana, Bogotá, Colombia

[3]Swiss Tropical and Public Health Institute, Allschwil, Switzerland

[4]University of Basel, Basel, Switzerland

[5]Hospital Universitario San Ignacio, Bogotá, Colombia

[6]Keralty Medical Centers, Bogotá, Colombia

[7]Medicalnet SAS, Bogotá, Colombia

[8]University of Sydney, Brain and Mind Center, Sydney, Australia

[9]Computer Simulation & Advanced Research Technologies, Sydney, Australia

[*]all authors contributed equally

Corresponding Author:
Laura Ospina-Pinillos, MD, PhD
Department of Psychiatry and Mental Health, School of Medicine, Pontificia Universidad Javeriana, Carrera 7 # 40-62, 8th Floor, Building Hospital Universitario San Ignacio, Bogotá, Colombia

Abstract

Background: The growing prevalence of mental health disorders among young people is a pressing global concern, particularly in low- and middle-income countries where access to care is limited. Digital tools, leveraging Information and Communication Technologies, offer promising approaches to bridge these gaps.

Objective: This study evaluated the feasibility of 2 digital mental health tools—Youth Collective Minds (YMC), a web-based platform, and Mental Beat (MB; Avicenna Research), a smartphone app—targeted at young individuals aged 18 - 25 years in Bogotá, Colombia.

Methods: Participants (N=35) engaged with both platforms over 3 weeks in this mixed methods feasibility study, which incorporated thematic analysis with a deductive framework for qualitative data. Univariate analyses were performed to examine baseline patterns and data distributions, while bivariate analyses were conducted to investigate relationships and associations between variables, providing a comprehensive evaluation of the platforms' feasibility in the acceptability, demand, implementation, and practicality domains.

Results: Participants were primarily women (22/35, 63%) with a median age of 23 (IQR 21-24) years. A total of 1308 annotations were coded: acceptability (annotations=707), demand (annotations=116), implementation (annotations=276), and practicality (annotations=209). Participants highlighted YMC's psychoeducational resources and MB's ease of use as strengths. However, technical issues, including server malfunctions and insufficient feedback, impacted engagement. Quantitatively, 83% (29/35) expressed willingness to reuse YMC and 83% (29/35) MB. Sensor data from MB indicated significant associations between psychological distress and smartphone usage. Participants with higher psychological distress showed greater median battery charging of 585 (IQR 321-615) compared to those without distress, 188 (IQR 42-309; P=.04). Poor sleep quality was also associated with increased median battery discharge of 2867 (IQR 1697.5-3935.5) compared to participants who reported sufficient sleep, 556 (IQR 200-2968; P=.003). GPS data showed that participants who visited more unique locations had lower psychological distress scores, with a negative correlation (r=−0.424; P=.05). In terms of platform usage, in YMC, surveys on emotions (30/35, 86%) and stress (28/35, 80%) were the most frequently completed, while telecounseling services were underused, with only 8.6% (3/35) of participants accessing mental health telecounseling. In MB, surveys of positive emotions (97.1%) and relationships (97.1%) were answered by more than 90% (32/35) of participants.

Conclusions: This study demonstrated the feasibility and acceptability of digital tools for mental health support among Colombian youth, suggesting that these tools promote self-awareness and mental health management but require technical refinements to

XSL•FO
RenderX

enhance engagement. The study's limitations, including a small sample size and short duration, underscore the need for broader research. Implementing participant feedback, strengthening cybersecurity, and scaling these tools could address mental health challenges in low- and middle-income countries, where such interventions are critically needed. These digital platforms represent promising steps toward bridging gaps in mental health care access.

(JMIR Form Res 2025;9:e69749) doi:10.2196/69749

KEYWORDS

mental health; young people; Colombia; eHealth; feasibility; usability; acceptability; Information and Communication Technologies

Introduction

The growing prevalence of mental health (MH) disorders in young people has become a pressing global concern. According to the Global Burden of Disease (GBD) 2020, there was a 25% increase in the global prevalence of depression and anxiety disorders worldwide, and results suggested that a more pronounced effect was seen in those aged between 20 and 24 years [1,2]. In low- and middle-income countries (LMICs), major economic, social, and health care challenges exacerbate this already significant issue [2,3]. Colombia is an LMIC with a long history of armed conflict in which different factors, such as poverty, lack of access to quality education, stigma, and limited MH resources, among others, obstruct a proper addressing of the MH needs of its population and worsen the treatment gap [4,5]. Regarding the MH situation of youth in the country, the 2015 National Mental Health Survey (NMHS) suggested that the lifetime prevalence of any MH disorder for those between 12 and 17 and 18 to 44 years of age was 7.1% and 9.1%, respectively [6]. Suicides in the 18 - 19 age group reach an alarming rate of 10.43 per 100,000 inhabitants, followed closely by the 20 - 24 age group at 9.98 per 100,000, compared with a 6.16 per 100,000 rate for the overall population [7].

Young people are not routine users of traditional MH services [8], and due to their digital literacy, they tend to rely on the internet as an initial source of information and guidance for their MH concerns [9,10]. Evidence suggests that self-help interventions and resources, where the individual has no contact with a health professional, are effective in reducing depression and anxiety, increasing a sense of control, reducing helplessness, and improving overall MH [11,12]. Over the past decade, Colombia has experienced substantial growth in Information and Communication Technology (ICT) infrastructure, fueled by government programs such as Plan Vive Digital, initiatives from the Ministry of ICTs, and the rollout of 5G, as well as contributions from the private sector [13-16]. These initiatives have enhanced broadband access, upgraded telecommunications, and promoted digital innovation nationwide.

Given this increased use and access to digital technologies, ICTs are considered to have great potential as self-help tools in providing information, resources, support, and communication to youth, thereby increasing patient empowerment and participation [17-19].

Some ICTs, such as smartphones or smartwatches, can collect not only active data (real input from users) but also passive data from sensors in the devices that gather information on parameters such as geolocation, pedometer, and accelerometer, among many others. The information obtained allows researchers and clinicians to perform a "Digital Phenotyping" that can be used to infer individual behavior and characteristics and can potentially aid in the early detection of some MH disorders [20,21]. According to the World Economic Forum (2019), the latter, added to ICTs' easy accessibility, is particularly interesting in LMIC settings, as they can help to overcome some of the MH barriers mentioned above [22]. Addressing this growing crisis requires a concerted effort to improve MH services, increase awareness, and provide support systems tailored to the unique needs of young people in LMICs [23-25].

Bogotá, the capital city of Colombia, presents a distinctive mix of opportunities and challenges for applying ICT in MH. Home to a predominantly young population [26], the city faces significant risk factors related to issues such as forced displacement, poverty, informal economies, violence, and inequality [27,28]. Despite these challenges, 94.1% of Bogotá's inhabitants aged 5 years or older own a smartphone and over 80.6% have internet access [29]. These factors position Bogotá as an ideal setting for conducting a study using ICT tools for MH promotion and prevention.

Our study aims to evaluate the feasibility of using 2 digital MH tools in individuals aged 18 to 25 years in Bogotá, the capital city of Colombia. To the best of our knowledge, there is no available information regarding the use of other digital MH interventions targeted at the MH of young people in Bogotá.

Methods

Recruitment

Participants were recruited through convenience and snowball sampling methods from a database of individuals who agreed to be contacted for research purposes. Dissemination of flyers and digital pieces also occurred through institutional social networks. Given the nature of the study, we aimed for a target sample of 35 participants who fully complied with the study protocol. Eligible participants were required to be aged 18 - 25 years, reside in the metropolitan region of Bogotá, have an Android smartphone with continuous internet access, and be native Spanish speakers. Exclusions included documented cognitive impairment and illiteracy. For enrollment, interested participants received study explanations and had any questions related to the study answered by researchers.

XSL•FO
RenderX

Procedure

After enrollment, participants filled out a sociodemographic survey including self-reported gender identity (man, woman, transgender, and other) and were asked to use 2 MH digital tools for 3 weeks (refer to Figures 1 and 2 for snapshots of their main features). Participants were not required to use both tools at the exact same time; however, they were encouraged to use them continuously during the study period. The first was Youth Collective Minds (YMC*; Mentes Colectivas para Jóvenes* in Spanish), a responsive website codesigned by the School of Medicine of the Pontificia Universidad Javeriana (PUJ), to promote self-management and MH help-seeking in youth through 7 features: a screener of concerns; automated, multimedia psychoeducation; "track as you go" (continuous monitoring); easy-to-understand graphs with legends; telecounseling services; customizable well-being plan; an SOS button for emergency services; and gamification with avatars [30]. A total of 13 surveys are included on topics related to MH, users freely choose their topics of interest, and YMC sends them 2 daily reminders to answer them. The tool has public access at all times, does not require downloading an app, and is available for youth older than 14 years of age or younger individuals with guardian consent.

Figure 1. Main features of Youth Collective Minds, a web-based platform evaluated in a 3-week mixed methods feasibility study conducted among young individuals aged 18 - 25 years in Bogotá, Colombia.

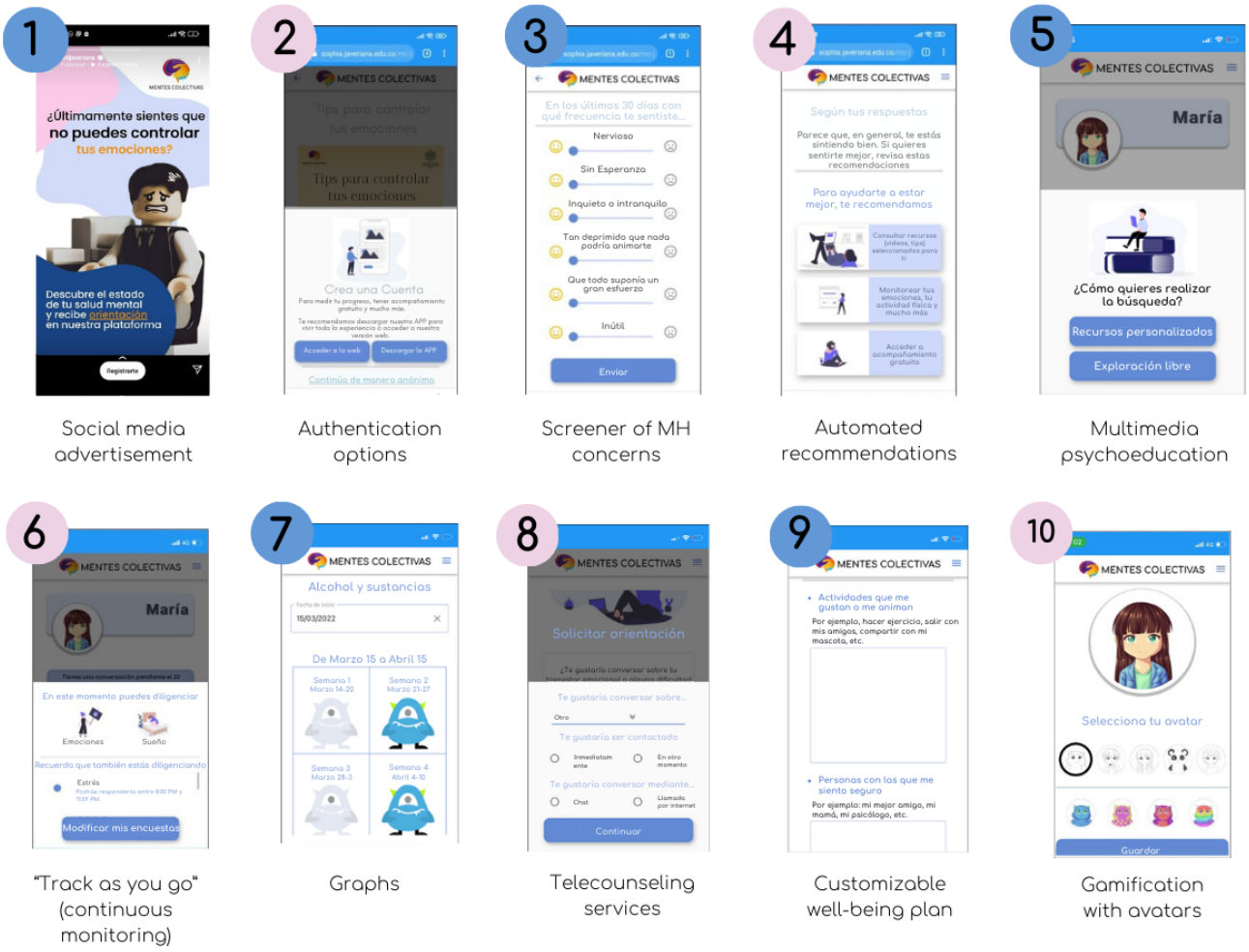

XSL•FO
RenderX

Figure 2. Main features of Mental Beat, a smartphone app tested alongside Youth Collective Minds in a 3-week mixed methods feasibility study with young individuals aged 18 - 25 years in Bogotá, Colombia.

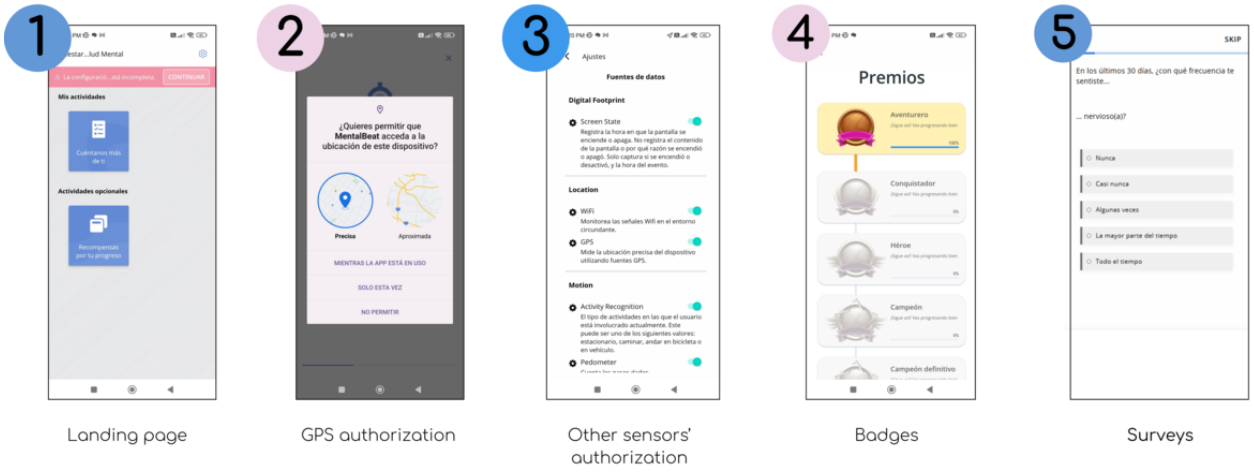

The second tool was Mental Beat (MB), which is based on the Avicenna Research app (Avicenna Research Inc). Avicenna is a participant-centric research platform designed to streamline the research process through seamless data collection, real-time monitoring and interventions, and enhanced participant engagement [31]. A 2020 cohort study in Australia used this app to prospectively follow several MH-related variables in more than 500 youth [32]. The study obtained good response rates for baseline questionnaires and acceptable response rates for follow-ups. Additionally, using both active and passive data, it was able to find significant associations, highlighting the potential of this adapted technology for MH monitoring. Based on this previous experience, we adapted the surveys for the Spanish-speaking Colombian context.

MB collects additional sociodemographic information, periodic questionnaires, and data from cell phone sensors, if the participant authorizes it. Each participant had to answer 1 or 2 surveys daily. The app has a gamification feature, in which users can level up and earn medals each time they complete a certain number of surveys. In the app, "leveling up" serves as a motivational tool to encourage continued user engagement and interaction by providing a sense of achievement and progress.

Participants were asked to log in to each platform at least once a day to complete the surveys and view the resources offered. If no activity was detected during the first week of the study, participants were contacted to have them comply with the study protocol, resolve any technical difficulties, or note reasons for withdrawal. Unreachable participants or those who voluntarily withdrew from the study were replaced to maintain the target sample.

The study used a mixed methods design to assess feasibility in 4 of the eight feasibility areas defined by Bowen et al [33]: (1) acceptability, which reflects how the target population received the tools including usability; (2) demand, evaluated through the usage of tools and their functionalities; (3) implementation, examined through the tools' operation; and (4) practicality, assessed through users' capacities and difficulties in using the tools. Although not part of Bowen and colleagues' framework, we considered usability as a key aspect to evaluate the feasibility of an ICT and included it in the broader acceptability category as it relates to how easy a technology is to use, encompassing its effectiveness, efficiency, and user satisfaction.

Data Collection

Three information sources were used: online semistructured interviews, online surveys, and platform data and analytics. Interviews, held online, included questions allowing assessment of a range of feasibility dimensions and were conducted by a senior psychologist and a psychologist specifically trained in the application of the interview guide (Multimedia Appendix 1), lasting between 45 minutes and 1 hour. The online survey included the Net Promoter Scale (NPS) [34], the System Usability Scale (SUS) [35], and questions about usage of platform functionalities. Finally, data were extracted from the platforms' analytics. Table 1 explains in detail the methods used to measure each feasibility dimension.

XSL•FO
RenderX

Table . Feasibility dimensions and measurement methods applied in a mixed methods feasibility study of 2 digital mental health tools among young individuals aged 18 - 25 years in Bogotá, Colombia.

Dimensions and instrument	Description
Acceptability	
Semistructured interview	Questions about general satisfaction, intent to continue use, satisfaction with monitoring features, perceived usefulness, perceived benefit of both tools, and opinions on the handling of personal data.
Net Promoter Score (NPS)	Willingness to promote the product measured by asking participants, "How likely are you to recommend [the platform] to a friend or family member?" on a scale from 0 (Not likely at all) to 10 (Extremely likely) [34]. Based on their responses, participants are categorized as promoters (9-10), passives (7-8), or detractors (0 - 6). Next, the percentages of promoters, passives, and detractors are calculated. The final score is derived by subtracting the percentage of detractors from the percentage of promoters. This score can range from –100 (if every participant is a detractor) to +100 (if everyone is a promoter). A score above 0 indicates that most participants would recommend the platform, reflecting their satisfaction. Conversely, a score below 0 suggests that most users are dissatisfied and would not recommend it. In the industry, scores above 50 are considered to indicate good performance [36]. NPS is widely used in health care [37].
System Usability Scale (SUS) [35] (in its translated version into Spanish) [38]	Ease of use of the platforms, measured through ten 5-point Likert questions focusing on aspects of complexity, ease to learn, among others. Scores range from 0 to 100, with a widely used benchmark of 68 [38,39].
Online survey	Questions include intent to continue use (yes or no), perceived usefulness of the platforms in solving users' concerns (rated from 0, Not useful at all, to 5, Very useful), and ratings of the platforms' service (classified as bad, acceptable, outstanding, good, or very good).
Demand of use	
Self-reported platform use	Self-reported frequency of usage of various functionalities from MB[a] (consulting digital medals) and YMC[b] (PDF results download, sharing resources with family, friends or people close to them, using the well-being plan, searching for resources, and viewing the history of recommended resources).
YMC analytics	Number of clicks by a participant on the various functionalities of the platform (SOS button, well-being plan, Avatar, searching for resources, consulted resources, and telecounseling, among others).
Platform surveys	Answers to the monitoring surveys in YMC (well-being, stress, sleep problems, physical activity, self-esteem, bullying, relationships, loneliness, eating disorders, substance misuse, health perception, sexual health, and an emotions tracker) and MB (emotional distress, number of close ones, positive emotion, interpersonal relationships, physical activity, commitment engagement and achievement, study behavior, well-being, sleep, and eating behavior).
Sensors	Location, number of steps, screen time, and battery status.
Implementation	
Platform's operation record	A record of the platform's operation was maintained during the weeks of the study, to note any bugs or technical difficulties.
Dropout rate and reasons	Percentage of participants who withdraw from a study before its completion and reasons they provided for doing so.
Practicality	
Semistructured interview	Questions about users' experience with the platforms, exploring what they considered motivating and demotivating factors, perceived conveniences and disadvantages, perceived positive and negative effects of usage, and difficulties encountered while using the platforms.

[a]MB: Mental Beat.

[b]YMC: Youth Collective Minds.

XSL•FO
RenderX

Analysis

Quantitative Analysis

Descriptive analyses were conducted to determine whether the tools achieved benchmark values on NPS and SUS (acceptability dimension), as well as to assess platform usage (demand dimension). Inferential analyses were run as a means to explore the potential of these tools for inferring individual behavior and thus for supporting early detection and treatment response in MH.

For continuous variables, a normality test was conducted. Considering the sample size, medians and IQRs were preferred as measures of central tendency. For nominal variables, absolute and relative frequencies were described. For ordinal variables, absolute frequencies, relative frequencies, and medians were calculated.

The statistical analysis was conducted using R software version 4.3.2 (The R Foundation for Statistical Computing). We performed the Shapiro-Wilk test ($P>.05$ for normality), calculated central tendency measures, skewness, and kurtosis. Additionally, a bivariate analysis was carried out, with qualitative interpretations of the K6 scale and well-being scale as dependent variables. We used the chi-square test ($P<.05$) for associations among qualitative variables and both the Wilcoxon test ($P<.05$) and Mann-Whitney U test ($P<.05$) for quantitative comparisons. Correlation matrices were also created for survey results from MB and YMC with sensor activation data, and significance was determined with $P<.05$. Additional correlation matrices were produced between MB and YMC variables, also with $P<.05$. This paper presents only the results that reached statistical significance.

Qualitative Analysis

Other feasibility dimensions (implementation and practicality), as well as information extracted from interviews related to acceptability and demand, were evaluated through qualitative data analysis. For qualitative data, a thematic analysis using a deductive framework was conducted. The process began with data familiarization through transcription and repeated readings. A research assistant transcribed the interviews and developed an initial set of codes. A senior psychologist with extensive experience in qualitative research reviewed these codes and organized the material according to the dimensions proposed by Bowen et al [33]—acceptability, demand, implementation, and practicality. The themes were subsequently reviewed for internal validity. Both psychologists collaborated on the final analysis and the writing of the results. NVivo version 14 (Lumivero) was used for qualitative data analysis.

Ethical Considerations

The research protocol was approved by the Institutional Review Board of the PUJ (FM-CIE-0886 - 23). Participants signed an informed consent form on the REDCap (Research Electronic Data Capture; Vanderbilt University) platform before proceeding to register on the platforms. All study data were deidentified prior to analysis. The documents were stored on secure servers provided by the PUJ and protected by password access. Those who completed the study received a voucher of COP 50,000 (approximately US $13) as a gesture of gratitude for their time and willingness to participate. Potential participants were notified of this as part of the enrollment procedures; however, only those who fully complied with the study protocol received the voucher.

Results

Participants

In total, 35 individuals participated in the study, the median age was 23 (IQR 21-24) years. Most of them identified themselves as women (22/35, 63%) and the majority as heterosexual (30/35, 86%). In Colombia, socioeconomic status is classified into 6 tiers on an ordinal scale, with Tier 6 representing the highest level and Tier 1 the lowest. Tiers classify residential properties to differentiate public utility charges and allocate subsidies and contributions. While useful, tiers are not a direct classification of users' payment capacity and they do not necessarily reflect other sociodemographic variables such as employment status or health coverage [40,41]. The majority of the sample belonged to the third and fourth socioeconomic tiers, with 54% (19/35) and 31% (11/35), respectively. Regarding educational level, 57% (20/35) had a university or technical degree. There were no people with disabilities, and none of them had been victims of the armed conflict (Table 2).

A total of 35 interviews were coded, adding up to 1308 annotations in the following codes: acceptability (annotations=707), demand (annotations=116), implementation (annotations=276), and practicality (annotations=209; Table 3).

XSL•FO
RenderX

Table . Sociodemographic characteristics of participants (N=35) in a mixed methods feasibility study of digital mental health tools for young individuals aged 18 - 25 years in Bogotá, Colombia.

Characteristics	Participants
Age (years), median (IQR)	23 (21-24)
Age (years), n (%)	
18 - 20	8 (23)
21 - 23	16 (46)
24 - 25	11 (31)
Gender, n (%)	
Woman	22 (63)
Man	13 (37)
Nationality, n (%)	
Colombian	35 (100)
Sexual orientation, n (%)	
Heterosexual	30 (86)
Homosexual	1 (3)
Bisexual	4 (11)
Socioeconomic tiers, n (%)	
3	19 (54)
4	11 (31)
5	3 (9)
6	1 (3)
Not stratified or no tier reported	1 (3)
Educational level, n (%)	
High school (between 6th and 12th grade)	15 (43)
College or university	15 (43)
Technical or technological	5 (14)
Belonging to ethnic groups, n (%)	
No	35 (100)
Disability condition, n (%)	
No	35 (100)
Victim of the armed conflict, n (%)	
No	35 (100)
Current psychiatric or psychological care, n (%)	
No	29 (83)
Yes	6 (17)
Psychiatric or psychological care in the past, n (%)	
No	9 (26)
Yes	26 (74)

XSL•FO
RenderX

Table . Number of coded annotations by feasibility dimension and platform in a mixed methods feasibility study with young individuals aged 18 - 25 years in Bogotá, Colombia.

Dimension of feasibility	Youth Collective Minds (n=600), n (%)	Mental Beat (n=474), n (%)	Both platforms (n=234), n (%)	Total (n=1308), n (%)
Acceptability	349 (58)	229 (48)	129 (55)	707 (54)
Demand	35 (6)	33 (7)	48 (20)	116 (9)
Implementation	134 (22)	136 (29)	6 (3)	276 (21)
Practicality	82 (14)	76 (16)	51 (22)	209 (16)

Acceptability

Youth Collective Minds

Based on survey responses, more than half of the participants considered the YMC service to be good (20/35, 57%), useful enough most frequently (20/35, 57%), and assured that they would use it again (29/35, 83%). These findings are consistent with interview results, for example, about half of the sample (16/35) is completely satisfied with the platform.

> *Yes, I am quite satisfied with the platform. I feel that it was very useful to be able to analyze during the day: how I felt? how much I had been physically active or sitting? For example, when I was asked if I had been sitting, I started to analyze and I was like! Wow!, I've been sitting for a long time, even all the time, studying, in class... so it made me analyze more the perspective of everything I was doing during the day, which I didn't do before using it*

When asked about what they liked most about the platform, the possibility of receiving personalized and easy-to-understand resources was frequently mentioned (23/35 participants).

> *I really liked it a lot, I would say that I was very satisfied, I feel that the platform is very well designed, it is very pleasant [visually], it is not tedious, also the surveys are very well designed, not tangled at all. The resources that were available after each survey as well, I did not review all of them, but most of them. The poster on "the psychological first aid kit" stood out, I thought it was very cool, I took a screenshot, also the podcast, I listened to one on the topic of suicide and it really impacted me a lot.*

Filling out the MH surveys and seeing progress charts was considered useful by 26 of 35 participants. Additionally, 10 of 35 participants highlighted the telecounseling option as one of the most useful features of YMC, as they claimed that often they do not know where to seek help and having this option available made help-seeking easier and more accessible. Among these 10 participants, fewer than half (4/10) had no prior or current experience with psychiatric or psychological care. This suggests that their positive views on the telecounseling feature were likely independent of past service use. In general, the surveys were seen as relevant, easy to answer, short, and the possibility of choosing their topics was emphasized as a positive feature.

Usability is a key aspect to consider when studying a technology. Bowen and colleagues' framework [33] for designing feasibility studies is not intended for technologies specifically, but for interventions more broadly. However, since the focus of this study is on 2 ICTs, usability measures were also included as part of the more general acceptability category. For YMC, participants gave this platform a score of 72.86 on the SUS, rendering it acceptable but requiring adjustments to improve its usability, mainly due to the technical difficulties that arose. Participants considered that these issues were the main cause of discouragement in using the platform. In interviews, dissatisfaction with the platform was primarily related to technical issues that limited access to the YMC web page for 3 days and generated difficulties logging in to their accounts. Nevertheless, a considerable percentage of participants (29/35, 83%) expressed the intention of using it again, and overall, they rated it as user-friendly, self-manageable, and free of contradictions. Participants considered that they could easily grasp how to use it without requiring any training or previous knowledge to start using it.

A score of 2.85 was obtained on the NPS scale: 34% (12/35) of the participants were promoters, 34% (12/35) passive, and 32% (11/35) as detractors. Despite this, the majority of participants (30/35) in the interviews expressed their intention to use the platform again, citing the personalized psychoeducational resources and the ability to track their MH through surveys as key factors for their decision.

> *With YMC, I was super satisfied, I feel that it tried to keep track of all the things that were not well, such as mood, panic, stress, if I did physical activity. I really liked the website, not only because of that, but also because of the other options it provided, such as recommendations, my well-being plan. I found it to be very complete, I liked it very much*

Mental Beat

In surveys, this platform received ratings of good and very good from 49% (17/35) and 29% (10/35) of participants, respectively. In terms of usefulness, 25 of 35 participants considered it useful enough and 25 of 35 assured that they would use it again. When it came to overall satisfaction, 24 of 35 participants emphasized that the interface made it easy to use. Additionally, over a quarter of participants (16/35) appreciated the push notifications, which remained visible on their home screens and served as reminders to complete the surveys. Overall, participants found the app to function consistently well.

> *With MB, I am very satisfied, the truth is that the application seemed to me quite robust, it did not get blocked at all, it was not intrusive at all, let's say*

XSL•FO
RenderX

within the system, accessing the surveys was quite easy and in fact I found them a little bit more meaningful.

Moreover, 23 of 35 participants found the MB app beneficial, with many highlighting the ability to monitor their MH as a key advantage. In addition, 18 of 35 interviewees noted that completing daily surveys helped them become more aware of their habits and MH.

Let's say that it emphasized things that one did not have in mind, or seemed very normal things, for example, regarding the quality of sleep, how much did you sleep? these are things that one began to identify, that one did not have in mind, so that's why I thought it was pretty cool to manage the application.

However, 20 of 35 participants expressed dissatisfaction with the app, citing the lack of feedback and resources after completing the surveys as a key issue. A lack of satisfaction that MB was only a data collection-focused, rather than MH care-focused app providing psychoeducational resources and other forms of support, led to a perception that the app became dull and monotonous over time. One participant stated:

I am not very satisfied with MB because it seemed very basic to me, I don't know, maybe it is because it is only a support of the virtual page of YMC, because I saw that in MB they only did surveys and that's it, and I said: well, but what do they do with this information? is it then connected with the virtual platform, or is it only about filling out surveys and that's it, so yes, I was left with the doubt and a bit of a lack of taste of MB.

Considering the NPS, 40% (14/35) of individuals were promoters of MB, 29% (10/35) passive, and 31% (11/35) detractors. This resulted in an NPS score of 8.6 (SD 2.61). In addition to the above, 20 of 35 participants found the MB surveys quick and easy to answer and 12 of 35 participants found the app to work well overall.

Regarding the use of sensors, participants stated that they were not aware of their activation and tracking, even though they granted permissions when first downloading and using the app. In general, participants trusted the app, and the majority of the participants (26/30) had no concerns regarding the handling of their personal data or MH information. They considered that the data requested were not sensitive and mentioned that trust in the researchers and the institution made them feel secure with sharing their data. In general, participants felt that the data were protected, and they prioritized their well-being when using the tools over concerns about data management.

On the SUS, we obtained a score of 81.35, which implies that the platform was within the acceptable range and therefore was well received. Overall, 25 of 35 of the youth said that they would use the MB platform again, mainly because they felt that answering daily questions about aspects of their MH allowed them to become aware of their well-being and the changes they were experiencing. Ease of using the platform obtained the highest score of 4.6. Participants highlighted the easy access and use of the app and that they considered that questions could be answered in a short time and in a simple way. Regarding specific questions of the SUS, the lowest score of 3.91 was obtained for whether they would use this tool frequently. In interviews, the main reason cited by participants was the lack of feedback after answering the surveys and not being able to visualize progress during the week.

Demand of Use

We evaluated both demand and perceived demand. Regarding the first, the actual use of YMC and MB functionalities was assessed. In relation to YMC, the most used features were psychoeducational resources, which were used at least once by 100% (35/35) of participants, followed by avatars (27/35, 77%), and completing the well-being plan (18/35, 51%; Table 4). Notably, the least used feature was telecounseling used by 8.6% (3/35) of participants.

Table . Usage of Young Collective Minds features during a 3-week mixed methods feasibility study among young individuals aged 18 - 25 years in Bogotá, Colombia.

YCM[a] feature	Participants who used it (N=35), n (%)
Select an avatar	27 (77)
SOS button	8 (23)
Well-being plan	18 (51)
Telecounseling for sexual and reproductive health	1 (2.9)
Telecounseling for mental health	3 (8.6)
Mental health resources	35 (100)

[a]YCM: Youth Collective Minds.

During the 3 weeks of the study, each participant used MH resources (mean 18.48, SD 14.54; median 13, IQR 10-22.5).

Regarding the monitoring functionality of YMC, some surveys were programmed to be shown daily, weekly, or every 2 weeks (Table 5). Regarding daily surveys, they were all completed by at least half of participants. The most answered ones were emotions (30/35, 86%) and stress (28/35, 80%), and the least used was physical activity with 57% (20/35) of the sample answering it at least once. Considering the total number of responses, the most used daily surveys were sleep and emotions. For weekly surveys, the most used were self-esteem, relationships, and loneliness; for biweekly surveys, the total number of responses was similar for well-being and sexual and

XSL•FO
RenderX

reproductive health. Among these, follow-ups were completed by the majority of participants for the emotions (26/35, 87%) and sleep (19/35, 83%) surveys.

Table . Usage of Young Collective Minds monitoring surveys during a 3-week mixed methods feasibility study among young individuals in Bogotá, Colombia.

YCM[a] survey[b]	Number of times the survey was shown to participants[c], n	Average number of responses (SD)	Participants who answered at least once (n=35), n (%)	Follow-ups, (n=35), n (%)	Total responses (n=695), n
Physical activity	21	5.05 (4.48)	20 (57)	16 (80)	101
Alcohol and substances	3	1.33 (0.52)	6 (17)	2 (33)	8
Nutrition	3	1.75 (0.71)	8 (23)	5 (63)	14
Self-esteem	3	1.64 (0.74)	14 (40)	7 (50)	23
Well-being	2	1.17 (0.41)	6 (17)	1 (17)	7
Bullying	3	1.50 (0.71)	2 (5.7)	1 (50)	3
Emotions	21	6.63 (4.36)	30 (86)	26 (87)	199
Stress	21	4.79 (3.92)	28 (80)	23 (82)	134
Health perception	3	1.60 (0.89)	5 (14)	2 (40)	8
Relationships	3	1.69 (0.75)	13 (37)	7 (54)	22
Sexual and reproductive health	1	1.00 (0)	6 (17)	N/A	6
Loneliness	3	1.69 (0.63)	13 (37)	8 (62)	22
Sleep	21	6.52 (4.45)	23 (66)	19 (83)	150

[a]YCM: Youth Collective Minds.

[b]Daily surveys were on physical activity, stress, sleep, and the emotions tracker. Biweekly surveys were the well-being and sexual and reproductive health surveys. All other surveys were programmed to be shown once a week.

[c]Number of times each survey appeared or was available to participants during the study period.

Regarding the MB surveys, all participants completed the surveys on demographic information, physical activity, and sleep at least once (Table 6). Surveys of positive emotions (97.1%) and relationships (97.1%) were answered by more than 90% (32/35) of participants. It is noteworthy that all participants (35/35, 100%) who responded to the physical activity survey answered it 2 or more times, showing better engagement with this survey topic on MB than on YMC. This survey was the most frequently presented to the participants.

Table . Usage of Mental Beat surveys during a 3-week mixed methods feasibility study among young individuals in Bogotá, Colombia.

MB[a] surveys	Number of times the survey was shown to participants[b], n	Average number of responses (SD)	Participants who answered at least once (n=35), n (%)	Follow-ups (n=35), n (%)	Total responses (n=848), n (%)
Close people	1	1.00 (0)	24 (69)	N/A[c]	24 (3)
Positive emotion	3	2.35 (0.77)	34 (97)	28 (82)	80 (9)
Relationships	3	2.26 (0.79)	34 (97)	27 (79)	77 (9)
Academic behavior	3	2.18 (0.81)	17 (49)	13 (76)	37 (4)
Well-being	1	1.00 (0)	17 (49)	N/A	17 (2)
Sleep	9	6.86 (2.03)	35 (100)	34 (97)	240 (28)
Physical activity	10	7.49 (2.23)	35 (100)	35 (100)	262 (31)
Tell us more about yourself (demographic)	1	1.00 (0)	35 (100)	N/A	35 (4)
Commitment, meaning, and achievement	3	2.45 (0.72)	31 (89)	27 (87)	76 (9)

[a]MB: Mental Beat.

[b]Number of times each survey appeared or was available to participants during the study period.

[c]N/A: not applicable.

XSL•FO
RenderX

Regarding sensors, MB logged various data types throughout the study, including the time of sensor activation and deactivation, duration, acceleration, and more. For this study, only the data described in Table 7 were used. Specifically, for activity type, MB uses a combination of sensors and algorithms to detect the most likely activity the user was performing. A detailed description of the data captured by each sensor is available on the Avicenna web page [42].

Table . Descriptive statistics of Mental Beat sensor data collected during a 3-week mixed methods feasibility study among young individuals in Bogotá, Colombia.

Sensor	Sensor description	n (%)	Median (IQR)
Screen status			
Off	Duration when the cell phone was off (minutes)	33 (94)	15502.88 (8143.75-23959.52)
On	Duration when the cell phone was on (minutes)	33 (94)	4321.55 (2419.1167-7829.4)
Pedometer	Number of steps	26 (74)	73078.5 (24735.25-91645.75)
Battery level	Battery level as a percentage (0% - 100%)	35 (100)	54.11 (30.82-67.75)
Battery status			
Charging	Frequency when the device was charging	35 (100)	199 (75-585)
Discharging	Frequency when the device was discharging	35 (100)	1739 (251-3500)
No charge	Frequency when no charge was detected	35 (100)	0
Full charge	Frequency when the device was fully charged	35 (100)	10 (0-127)
Activity type (AT)			
Stationary	Frequency when the AT was immobile	29 (83)	1713 (840-2981)
Inclination	Frequency when the AT was tilted	29 (83)	906 (248-1451)
On foot	Frequency when the AT was walking or running	29 (83)	195 (73-412)
On bicycle	Frequency when the AT was cycling	29 (83)	10 (4-23)
In vehicle	Frequency when the AT was in a vehicle	29 (83)	229 (26-642)

Specific to battery level and status, data from this sensor are relevant as they can serve as a measure of cellphone usage, which has been shown to be significantly associated with problematic internet usage in youth [43]. Other sensors, such as GPS, pedometer, and screen time, have been used previously in digital phenotyping research [43-45].

To assess demand for sensors, we evaluated the percentage of participants who had at least one record for each sensor, as well as the median value of the sensor (Table 7). Based on this approach, the most frequently used sensors were battery status and screen status (refer to Table 7 for sensor description), with data recorded for 100% (35/35) and 94.3% (33/35) of the sample, respectively. The activity type most frequently detected was the stationary state with a median of 1713 (IQR 840-2981) records, while the on-foot and on-bicycle states were recorded less frequently. The pedometer was active for 74% (26/35) of participants, providing step count data. Data on battery status showed that participants spent most of their time in a discharging state, while instances of full charge were infrequent.

The GPS sensor (Table 8) focused on travel and location metrics. Over the study period, the median total distance traveled by participants was 793.29 (IQR 140.43-1781.16) km, suggesting variability in mobility levels. The median total travel time recorded was 2501.13 (IQR 269.6875-3890.025) minutes, while participants made a median of 60056.5 (IQR 12725.5-77800.5) visits to different locations. In terms of unique places visited, the median was 42 (IQR 10-72), indicating diverse movement patterns among participants. Time at home had a median of 476.63 (IQR 11.7-927.7625) minutes, contrasting with a significantly higher median time spent outside the home at 1695.4 (IQR 174.1-3004.975) minutes, highlighting a tendency for more time in external environments.

XSL•FO
RenderX

Table . Descriptive statistics of GPS sensor data from Mental Beat sensors during a 3-week mixed methods feasibility study among young individuals in Bogotá, Colombia.

GPS sensors	Median (IQR)
Total distance (km)	793.285 (149.42-1695)
Total travel time (minutes)	2501.125 (269.6875 - 3890.025)
Total visits	50,022.5 (12725.5-77800.5)
Unique places	42 (10-72)
Time at home (minutes)	476.625 (476.625- 927.7625)
Time outside home (minutes)	1695.4 (174.1-3004.975)

These results showcase the demand for YMC and MB. During the duration of the study, usage was overall high; however, engagement varied according to the type of functionality, survey topic, or sensor.

Regarding perceived demand, 27 of 35 participants thought that other young people in Bogotá would use the platforms. Some (12/35), however, stated that they would need adjustments, including being in an app-based format (7/35), having a strong social media presence (1/35), and improving its interface (2/35). A group of participants (9/35) thought that only those already interested in MH topics would use them.

Given the high demand for MB and YMC, we also explored associations between sensors and MH-related measures. This allows us to explore the feasibility of not only engaging young people with digital health technologies for their MH, but also exploring their utility as potential indicators of MH state. The group with psychological distress had a mean battery level of 62.98 (SD 8.92), whereas the group without psychological distress had a mean battery level of 53.66 (SD 11.03). The results suggest that individuals with psychological distress had a significantly higher mean battery level compared to those without distress. This association was statistically significant, with a *P* value of .03, indicating that higher psychological stress correlates with a higher battery level. Previous research has linked problematic internet use with psychiatric symptoms [46]. This suggests that metrics like device usage for internet access, as well as related proxies such as screen time or battery level, could hold value for both clinical and research apps.

Likewise, a statistically significant association was found between battery charging status (indicating the phone is being charged) and psychological distress (*P*=.04). In individuals with psychological distress, the median number of battery charging records was 585 (IQR 321-615), while in those without psychological distress, the median was 188 (IQR 42-309). This finding suggests a relationship between higher levels of psychological distress and increased battery charging. Relatedly, Gansner and colleagues [43] found that a higher number of daily phone checks and phone sessions were related to lower scores in measures of problematic internet use and depression. Not in the direction found in previous studies, Gansner and colleagues proposed an alternative hypothesis: digital engagement may offer relief from negative affect in the short term. Similarly, in our sample, youth with psychological distress might be more prone to use their devices for short-term relief than youth without psychological distress, leading to more frequent charging. Interestingly, youth with psychological distress also exhibited higher battery levels, which could suggest they maintained these levels by frequently charging their phones.

Participants who perceived they slept enough had a median battery discharge of 556 (IQR 200-2968), while those who did not perceive they slept enough had a significantly higher median discharge of 2867 (IQR 1697.5-3935.5). This suggests a statistically significant relationship between battery discharge and perceived sleep quality *(P*=.003). As for the activity type sensors, a statistically significant relationship was found between immobility and time spent in vehicles in relation to sleep quality (*P*=.03). Participants who perceived they slept enough recorded a median of 1213 (IQR 218-2607) minutes of immobility, whereas those who did not perceive they slept enough reported a median of 2129 (IQR 1515-6701) minutes. This association was statistically significant.

Similarly, time spent in a vehicle also differed significantly between the two groups. Participants who perceived they slept enough had a median of 46 (IQR 17-525) recorded instances of being in a vehicle, compared to 262 (IQR 150-762) minutes for those who did not perceive they slept enough. This difference was statistically significant, with a *P* value of .05, suggesting that greater immobility is associated with a worse perception of sleep quality.

A Mann-Whitney *U* test was performed to compare Screen Status Off in the group that perceived they had not slept well and the group that perceived they had slept well. There was a significant difference in Screen Status Off between these two groups: z=−2.089; *P*=.04. This test indicates that participants who felt they didn't sleep well had their phones off for a longer time.

There was a statistically significant negative correlation between the number of unique places visited by a person and their score on the K6 psychological distress test. For this analysis, a nonparametric Spearman correlation test was conducted, resulting in a correlation coefficient of −.424 (*P*=.05). This implies that the greater the number of unique places a person visited during the 3-week period, the lower their score on the psychological distress test. Unique places are defined as those locations that the participant visited only once during the study.

XSL•FO
RenderX

Implementation

Overview

During the 3-week study, the YMC platform encountered multiple technical difficulties, including server malfunctions, web page errors (reported by 19/35 participants), slow registration processes, delays in starting sessions (22/35), and dissatisfaction with the lack of notifications (3/35). Other complaints included excessive emails, limited survey options, and issues visualizing results.

In contrast, the MB app had no major technical issues and was appreciated for its offline functionality, fast performance (18/35), and push notifications (8/35). However, some survey responses were not recorded or processed, affecting 3 participants, and inconsistent passive data recordings led to the exclusion of some data points. Both platforms faced complaints about rigid survey schedules, which often resulted in expired surveys. The screen state sensor occasionally recorded implausible durations exceeding 24 hours, resulting in 6 data points being excluded. Similarly, the activity type sensor, which categorizes activities (eg, walking and running), recorded discrepancies for one participant where "on-foot" activity data did not align with walking and running records, leading to the exclusion of that participant's data. These issues highlight challenges in accurately capturing passive data for analysis.

Participant dropout involved 3 individuals. One completed the study but was unreachable for an interview, while the other two left due to technical issues with YMC, including unsaved responses and platform freezing.

Practicality

Participants reported several positive effects from using the platforms, including increased awareness of their MH and habits (21/35 for MB and 16/35 for YMC), with specific emphasis on sleep surveys helping them recognize the importance of sleep hygiene. Many participants noted that the platforms motivated them to adopt healthier routines (28/35 for MB and 25/35 for YMC) by encouraging mindfulness about daily activities and creating dedicated time for MH. In YMC, learning was frequently highlighted as a benefit (30/35), with participants gaining knowledge, validation, and support through resources, survey results, and graphs.

> *When I get a sleep survey, or when I get a survey about screen time on the TV or apps, it makes me aware, like, 'Wait a minute! How much time did I spend on my phone today, or how much time did I spend lying down?' Or, for example, with questions like the one about physical activity—how much time did you spend being active or studying? [...] It's like this awareness that shows you, 'Okay, there are things I can improve, there are things I can change,' and that's all thanks to the app* [MB]

However, participants expressed dissatisfaction with the lack of feedback in MB (18/35), suggesting it felt like a data collection tool without deeper engagement. A few participants (4/35 for MB and 1/35 for YMC) saw no positive effects, and boredom was noted as a potential barrier to long-term use (3/35 for MB and 4/35 for YMC). Stress from technical malfunctions (7/35 for YMC), frustration with forgetting surveys, and occasional device issues were also reported.

> *It's like you might get bored with the platform [YMC], I don't know, but I do feel that. I also feel that it can kind of just stay there... Since it's something you're supposed to do regularly, like daily, I feel that within the routine and everything else, it can start to fade away and you might gradually set it aside.*

Discussion

Principal Results

Our feasibility trial evaluated the use of 2 digital MH tools, YMC and MB, among young individuals aged 18 to 25 years in Bogotá, Colombia—a setting characterized by limited MH resources despite increasing internet accessibility and extensive usage among youth (84% of individuals aged 12 - 24 years) [23,47,48]. Both tools, previously described in the "Procedure" section, offer functionalities related to psychoeducation and mental health monitoring (passive and active data collection). The results demonstrated the feasibility of these tools in terms of acceptability, demand, implementation, and practicality, highlighting their potential to address MH needs in resource-constrained settings.

Emerging technologies have significant potential to enhance the accessibility and quality of MH care, making them valuable tools for both patients and MH professionals [49-52]. While the acceptability and use of these tools are well-established in high-income countries, their adoption in LMICs is still in its early stages [53]. Research and development in LMICs remain insufficient, and concerns persist regarding the scalability of such interventions in contexts where resources and connectivity are often limited [53-55]. However, the progressive growth of ICTs in LMICs presents an opportunity to address these challenges. In Colombia, despite widespread internet access and high usage among young populations [29,47], the availability of digital MH tools remains scarce due to factors such as costs, limited human resources, and infrastructure constraints [47,48]. Addressing these barriers through innovative and scalable solutions is essential for bridging the MH care gap in LMICs [56,57].

Our results were consistent with other researchers showing that participants were curious to explore their MH, open to trying these nontraditional tools, and their digital literacy allowed them to use them without requiring any prior training [58-61]. Overall, the tools were deemed acceptable by the majority of participants. Both were considered easy to use, and most expressed a willingness to use them again. In line with most research on digital tools, our participants stressed the importance of making the tools user-friendly and intuitive [62].

Despite their acceptance, technical difficulties and system bugs negatively impacted the user experience and reduced engagement for some participants [63,64]. Such issues in digital health tools can lead to high user attrition and, more critically, compromise patient safety [63,65]. Challenges such as app crashes, unresponsive designs, and inadequate error handling

XSL•FO
RenderX

not only frustrate users but also disrupt health care quality [66]. A review by Kim et al [65] highlights that technical problems in health IT can hinder decision-making and delay care delivery, underscoring the importance of addressing these issues proactively. Aligned with this, our study suggests that rigorous testing, user-centered design, continuous monitoring, and the establishment of system failure protocols are essential to ensure the usability, safety, and reliability of digital health systems. These measures are critical to maintaining trust and achieving the intended outcomes in academic and clinical settings.

Through our experience, we found that young people are willing to share information about their MH through monitoring or sensors, as long as the process is straightforward and quick. This coincides with other findings that indicate high trust in digital tools among this population due to their digital literacy and familiarity with the digital environment [57,67,68]. Privacy and sharing information about their MH were not the main concerns for our participants, as seen in other studies [69]. In our case, most participants were not even aware of whether they had activated the sensors or not. This willingness and openness to try resources that might help them and provide information about their MH are what give digital tools significant potential with this population. However, this finding should be interpreted cautiously as participants were recruited from a pool of people who had already agreed to be contacted for participating in research, potentially introducing bias.

If this result is supported by further research, it would present a critical issue, as it makes young users particularly vulnerable by sharing sensitive information without proper assessment. This highlights the importance of creating and developing digital resources that protect their privacy and security, as they may not consider the information they share on these platforms as sensitive [70,71]. This underscores the need for digital tools to incorporate educational components that enhance users' understanding of privacy rights, data use, and informed consent, rather than limiting participation to consent collection alone. Additionally, it underscores the responsibility of digital health providers to implement and maintain robust cybersecurity standards to safeguard health data. Given the highly sensitive nature of MH information, maximizing cybersecurity measures is essential to prevent unauthorized access, data breaches, or misuse, ensuring that user trust and safety are upheld [22]. Thus, while our results point to the importance of low-friction design, particularly for the monitoring or sensor components of digital tools, this should be balanced with robust safeguards to ensure user privacy and data protection. The use of sensors in modern devices is being explored as a tool to gather data on behavioral markers that may be linked to MH concerns [72,73]. While research on sensor-derived information is still emerging, it holds potential for providing valuable insights. Our findings showed interesting associations that allow for interpretation but do not support generalization or definitive conclusions. We found a significant association between psychological distress and mean battery level, where higher psychological stress correlated with higher battery level, as well as a correlation between increased battery discharge and higher levels of psychological distress. Another association was observed between perceived sleep quality and battery discharge: those who reported poor sleep quality had higher battery discharge, suggesting that individuals who spend more time on their devices tend to have poorer sleep quality. This suggests that individuals with increased psychological distress likely use their phones more (spend more time on them or use more apps), leading to faster battery drain, which may result in them charging their phones more frequently and having a higher average battery level. This finding is in line with previous research that suggests that excessive smartphone use is linked to psychiatric disorders, cognitive and emotional difficulties, sleep problems, and medical issues [74,75]. However, we are not proposing that smartphones are inherently harmful to youth MH; instead, this association warrants further exploration to better understand its implications.

The results also suggest that participants with lower sleep quality spent more time immobile and more time in vehicles. Supporting this finding, research shows that increased sedentary behavior is associated with poor sleep quality, with better sleep levels linked to lower adiposity and higher sedentary time correlating with sleep disturbances, regardless of physical activity levels [76,77]. Interestingly, in our sample, individuals who visited more unique locations (places visited only once during the study) had lower scores on the psychological distress test. It is well established that depression is associated with decreased activity, reduced motivation, and increased sedentary behavior [78]. Consistent with our findings, preliminary research suggests that certain GPS data patterns, such as reduced diversity in visited locations and spending more time in fewer places, are associated with more severe symptoms of depression.

Finally, our study shows that researchers interested in using data collection apps such as MB should consider incorporating responsive links to psychoeducational resources, feedback after survey completion, and data visualization. This approach can help sustain the engagement of young participants in long-term research studies and is consistent with previous studies highlighting the engagement benefits of including feedback in digital solutions [79,80].

Limitations

This study has several limitations that should be considered when interpreting the findings. First, the small sample size of 35 participants and nonrandom sampling limit the generalizability of the results, as they may not fully represent the broader population of young people. Recruiting participants from a pool of people who had already agreed to participate in research could have also introduced bias. Their prior engagement and potential familiarity or trust in the university context could have influenced their openness toward digital tools and the research process, possibly leading to higher acceptability ratings than would be observed in less research-exposed populations. Additionally, the study's short duration of 3 weeks may not provide sufficient insights into long-term usability, engagement, or the sustainability of the digital tools. The geographical focus on Bogotá further restricts the applicability of the findings, as the experiences of youth in other regions or countries with different cultural and socioeconomic contexts may vary significantly. Although Bogotá includes rural zones within its administrative boundaries, it is primarily an urban city, and only a small proportion of its population resides in rural areas [81].

XSL•FO
RenderX

Thus, while our sample reflects the urban characteristics of Bogotá's youth, results from our study may not be generalizable to rural contexts where access to the internet and mobile technologies is more limited. Future studies should be carried out in other regions to explore how differences in connectivity, infrastructure, socioeconomic context, and digital literacy influence the feasibility of mental health tools. Finally, the restriction to participants aged 18 - 25 years excludes other age groups, potentially overlooking usability and acceptability issues relevant to younger adolescents or older adults. These limitations highlight the need for further research to validate and expand upon these initial findings.

Conclusions

The study highlights the feasibility of digital MH tools for young people in Bogotá, Colombia, with both YMC and MB demonstrating potential for fostering self-awareness and MH self-management. To the best of our knowledge, YMC is the only digital platform specifically targeting MH improvement in Colombia, with promising acceptance among youth. Incorporating participant feedback and refining features could further enhance its usability and impact.

Despite the potential, challenges such as technical issues, limited feedback, and a small, geographically specific sample size limit the generalizability of findings. Future research should expand to more diverse populations, explore long-term use, and address usability and cybersecurity concerns to build trust and scalability. These efforts could advance digital health tools as transformative solutions to bridge MH care gaps in Colombia and other LMICs.

Acknowledgments

We appreciate the participation of all young people who invested their time in our study and shared their mental health data with us, as well as the collaboration of the Centro de Asesoría Psicológica y Salud at the Pontificia Universidad Javeriana for supporting recruitment, and the assistance of Alexandra Morales with study set-up in MB. The generative artificial intelligence (GenAI) tool ChatGPT by OpenAI was used for grammar, spelling, and consistency checks. All scientific content, analyses, and interpretations were conceived and written by the authors. All AI-generated outputs were carefully reviewed and verified by the authors before inclusion in the manuscript.

Funding

This research was supported by the kind funding of Fondation Botnar (REG-19-026). The study's sponsors did not participate in the study's design, data collection, data analysis, data interpretation, or report writing.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: LOP, JAO, CGR
Supervision: LOP
Data collection: MNSN and MIRF
Data curation: MNSN, DLSR, MIRF, and MFRC
Formal analysis: LOP, MNSN, DLSR, MIRF, MFRC, SC
Methodology: LOP
Validation: DLSR, MIRF, MFRC
Writing – original draft, LOP, MNSN, MGCV, MIRF, MFRC, and DLSR
Writing – review & editing, JAO, IBH, CGR, MIRF, MNSN, DLSR, MFRC, MGCV, AANM
Funding acquisition: JAO, LOP
Resources: CGR

Conflicts of Interest

IBH is a Professor of Psychiatry and the Co-Director of Health and Policy, Brain and Mind Centre, University of Sydney. He has led major public health and health service development in Australia, particularly focusing on early intervention for young people with depression, suicidal thoughts and behaviors, and complex mood disorders. He is active in the development through codesign, implementation, and continuous evaluation of new health information and personal monitoring technologies to drive highly personalized and measurement-based care. He holds a 3.2% equity share in Innowell Pty Ltd, which is focused on the digital transformation of mental health services. JAO is head of Systems Modeling & Simulation at the Brain and Mind Centre (BMC), University of Sydney, and is Managing Director of Computer Simulation and Advanced Research Technologies (CSART).

Multimedia Appendix 1
Interview guide.

XSL•FO
RenderX

[DOCX File, 25 KB - formative_v9i1e69749_app1.docx]

References

1. Santomauro DF, Mantilla Herrera AM, Shadid J, et al. Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. Lancet 2021 Nov;398(10312):1700-1712. [doi: 10.1016/S0140-6736(21)02143-7]
2. Mental Health Atlas 2020. World Health Organization. 2021. URL: https://www.who.int/publications/i/item/9789240036703 [accessed 2024-11-28]
3. Kohn R, Ali AA, Puac-Polanco V, et al. Mental health in the Americas: an overview of the treatment gap. Rev Panam Salud Publica 2018;42:e165. [doi: 10.26633/RPSP.2018.165] [Medline: 31093193]
4. Bartels SM, Cardenas P, Uribe-Restrepo JM, et al. Barreras y facilitadores para el diagnóstico y tratamiento de la depresión en atención primaria en Colombia: Perspectivas de los proveedores, administradores de atención médica, pacientes y representantes de la comunidad. Revista Colombiana de Psiquiatría 2021 Jun;50(50):67-76. [doi: 10.1016/j.rcp.2021.01.001]
5. Galvis SC, Cantor LS. Salud mental en tiempos de COVID-19 [Mental health in times of COVID-19]. : Inst Nac Salud-Colombia [National Institute of Health - Colombia]; 2021 URL: https://www.ins.gov.co/Direcciones/ONS/informe_131/pdf/cap3_1.pdf [accessed 2024-11-28]
6. Gómez-Restrepo C. La Encuesta Nacional de Salud Mental–ENSM 2015 [National Mental Health Survey- NMHST 2015]. Revista Colombiana de Psiquiatría 2016 Dec;45:1. [doi: 10.1016/j.rcp.2016.09.006]
7. Encuesta nacional de consumo de sustancias psicoactivas [National survey on the consumption of psychoactive substances]. : Ministerio de Justicia y del Derecho [Ministry of Justice and Law]; 2019 URL: https://tinyurl.com/58da62mk [accessed 2024-06-17]
8. Burns JM, Davenport TA, Christensen H, Luscombe GM, Mendoza JA, Bresnan A, et al. Game on: exploring the impact of technologies on young men's mental health and wellbeing. findings from the first young and well national survey. : Young and Well Cooperative Research Centre; 2013 URL: https://api.semanticscholar.org/CorpusID:158520280 [accessed 2025-12-04]
9. Burns J, Birrell E. Enhancing early engagement with mental health services by young people. Psychol Res Behav Manag 2014;7:303-312. [doi: 10.2147/PRBM.S49151] [Medline: 25473320]
10. Marciano L, Saboor S. Reinventing mental health care in youth through mobile approaches: current status and future steps. Front Psychol 2023;14:1126015. [doi: 10.3389/fpsyg.2023.1126015] [Medline: 36968730]
11. Bower P, Richards D, Lovell K. The clinical and cost-effectiveness of self-help treatments for anxiety and depressive disorders in primary care: a systematic review. Br J Gen Pract 2001 Oct;51(471):838-845. [Medline: 11677710]
12. Khan N, Bower P, Rogers A. Guided self-help in primary care mental health: meta-synthesis of qualitative studies of patient experience. Br J Psychiatry 2007 Sep;191(3):206-211. [doi: 10.1192/bjp.bp.106.032011] [Medline: 17766759]
13. Evaluación de los programas del plan vive digital para la gente financiados con recursos del fondo de tecnologías de la información y las comunicaciones (FONTIC) [Evaluation of the programs of the "Vive Digital para la Gente" plan financed with resources from the Information and Communications Technologies Fund (FONTIC)]. : Departamento Nacional de Planeación [National Planning Department]; 2018 URL: https://colaboracion.dnp.gov.co/CDT/Prensa/EstudioFONTIC.pdf [accessed 2024-12-04]
14. Planes [Plans]. Ministerio de Tecnologías de la Información y las Comunicaciones [Ministry of Information and Communications Technologies]. URL: https://www.mintic.gov.co/portal/inicio/Ministerio/Planes [accessed 2024-12-04]
15. Lorduy J. WOM y telecall aún no inician el despliegue 5G en colombia [WOM and Telecall have not yet begun the 5G rollout in Colombia]. Portafolio. 2024. URL: https://www.portafolio.co/tecnologia/como-avanza-el-despliegue-5g-en-colombia-los-operadores-con-mas-redes-615379 [accessed 2024-12-04]
16. Así avanzaron las TIC en colombia durante el 2023 [This is how ICT advanced in Colombia during 2023]. Ministerio de Tecnologías de la Información y las Comunicaciones [Ministry of Information and Communications Technologies]. 2023. URL: https://www.mintic.gov.co/portal/inicio/Sala-de-prensa/Noticias/333627:Asi-avanzaron-las-TIC-en-Colombia-durante-el-2023 [accessed 2024-12-04]
17. Naslund JA, Aschbrenner KA, McHugo GJ, Unützer J, Marsch LA, Bartels SJ. Exploring opportunities to support mental health care using social media: a survey of social media users with mental illness. Early Interv Psychiatry 2019 Jun;13(3):405-413. [doi: 10.1111/eip.12496] [Medline: 29052947]
18. Aschbrenner KA, Naslund JA, Tomlinson EF, Kinney A, Pratt SI, Brunette MF. Adolescents' use of digital technologies and preferences for mobile health coaching in public mental health settings. Front Public Health 2019;7:178. [doi: 10.3389/fpubh.2019.00178] [Medline: 31312629]
19. Litke SG, Resnikoff A, Anil A, et al. Mobile technologies for supporting mental health in youths: scoping review of effectiveness, limitations, and inclusivity. JMIR Ment Health 2023 Aug 23;10:e46949. [doi: 10.2196/46949] [Medline: 37610818]
20. Torous J, Kiang MV, Lorme J, Onnela JP. New tools for new research in psychiatry: a scalable and customizable platform to empower data driven smartphone research. JMIR Ment Health 2016 May 5;3(2):e16. [doi: 10.2196/mental.5165] [Medline: 27150677]

XSL•FO
RenderX

21. Spinazze P, Rykov Y, Bottle A, Car J. Digital phenotyping for assessment and prediction of mental health outcomes: a scoping review protocol. BMJ Open 2019 Dec 30;9(12):e032255. [doi: 10.1136/bmjopen-2019-032255] [Medline: 31892655]
22. Doraiswamy PM, London E, Varnum P, et al. Empowering 8 Billion Minds: enabling better mental health for all via the ethical adoption of technologies. NAM Perspect 2019;2019. [doi: 10.31478/201910b] [Medline: 34532674]
23. Nota estadística: Juventud en Colombia [Statistical note: youth in Colombia]. : Departamento Administrativo Nacional de Estadística [National Administrative Department of Statistics]; 2021 URL: https://www.dane.gov.co/files/investigaciones/notas-estadisticas/dic-2021-nota-estadistica-juventud-en-colombia.pdf [accessed 2025-12-04]
24. Panorama sociodemográfico de la juventud en Colombia ¿Quiénes son, qué hacen y cómo se sienten en el contexto actual? [Sociodemographic overview of youth in Colombia: Who are they, what do they do, and how do they feel in the current context?]. Departamento administrativo nacional de estadística [National Administrative Department of Statistics]. 2020. URL: https://www.dane.gov.co/files/investigaciones/genero/informes/informe-panorama-sociodemografico-juventud-en-colombia.pdf [accessed 2024-11-28]
25. Kieling C, Baker-Henningham H, Belfer M, et al. Child and adolescent mental health worldwide: evidence for action. Lancet 2011 Oct 22;378(9801):1515-1525. [doi: 10.1016/S0140-6736(11)60827-1] [Medline: 22008427]
26. Población de bogotá d.c. 2005 – 2035 [Population of bogotá d.c. 2005 – 2035]. SaluData - Observatorio de Salud de Bogotá [SaluData - Bogotá Health Observatory]. URL: https://saludata.saludcapital.gov.co/osb/indicadores/poblacion-de-bogota-d-c-2005-2035 [accessed 2024-12-04]
27. Los desafios de las ciudades [The challenges of cities]. Bogotá Cómovamos. URL: https://bogotacomovamos.org/los-desafios-de-las-ciudades/ [accessed 2024-12-04]
28. Violencia intrafamiliar y de género en Bogotá DC [Family and gender-based violence in Bogotá DC]. Observatorio de Salud de Bogotá [Bogotá Health Observatory]. URL: https://saludata.saludcapital.gov.co/osb/indicadores/violencia-intrafamiliar [accessed 2024-12-04]
29. Boletín técnico: indicadores básicos de tenencia y uso de tecnologías de la información y las comunicaciones – TIC en hogares y personas de 5 y más años de edad departamental 2023 [Technical Bulletin: Basic indicators of ownership and use of information and communication technologies – ICTs in households and people aged 5 and over, departmental 2023]. Departamento Administrativo Nacional de Estadística [National Administrative Department of Statistics]. 2024. URL: https://www.dane.gov.co/files/operaciones/TICH/bol-TICH-2023.pdf [accessed 2024-12-04]
30. Pontificia Universidad Javeriana. Mentes Colectivas [Collective Minds]. URL: https://www.javeriana.edu.co/mentescolectivas [accessed 2024-12-04]
31. Patient-centric clinical trial and health research software platform. Avicenna Research. URL: https://avicennaresearch.com [accessed 2024-12-04]
32. Munasinghe S, Sperandei S, Freebairn L, et al. The impact of physical distancing policies during the COVID-19 pandemic on health and well-being among Australian adolescents. J Adolesc Health 2020 Nov;67(5):653-661. [doi: 10.1016/j.jadohealth.2020.08.008] [Medline: 33099413]
33. Bowen DJ, Kreuter M, Spring B, et al. How we design feasibility studies. Am J Prev Med 2009 May;36(5):452-457. [doi: 10.1016/j.amepre.2009.02.002] [Medline: 19362699]
34. Reichheld FF. The one number you need to grow. Harv Bus Rev 2003 Dec;81(12):46-54. [Medline: 14712543]
35. Brooke J. SUS: a “quick and dirty” usability scale. In: Usability Eval Ind: CRC Press; 1996:207 URL: https://rickvanderzwet.nl/trac/personal/export/104/liacs/hci/docs/SUS-questionaire.pdf [accessed 2024-11-28] [doi: 10.1201/9781498710411-35]
36. Hamilton DF, Lane JV, Gaston P, et al. Assessing treatment outcomes using a single question. Bone Joint J 2014 May;96-B(5):622-628. [doi: 10.1302/0301-620X.96B5.32434]
37. Adams C, Walpola R, Schembri AM, Harrison R. The ultimate question? Evaluating the use of Net Promoter Score in healthcare: a systematic review. Health Expect 2022 Oct;25(5):2328-2339. [doi: 10.1111/hex.13577] [Medline: 35985676]
38. Sevilla-Gonzalez MDR, Moreno Loaeza L, Lazaro-Carrera LS, et al. Spanish version of the System Usability Scale for the assessment of electronic tools: development and validation. JMIR Hum Factors 2020 Dec 16;7(4):e21161. [doi: 10.2196/21161] [Medline: 33325828]
39. Hyzy M, Bond R, Mulvenna M, et al. System Usability Scale benchmarking for digital health apps: meta-analysis. JMIR Mhealth Uhealth 2022 Aug 18;10(8):e37290. [doi: 10.2196/37290] [Medline: 35980732]
40. Estudios Poscensales de jóvenes investigadores. Censo Nacional de Población y Vivienda 2018 [Post-census studies by young researchers. 2018 National Population and Housing Census]. : Departamento Administrativo Nacional de Estadística [National Administrative Department of Statistics]; 2018 URL: https://www.dane.gov.co/files/censo2018/estudios-poscensales/08-heterogeneidad-de-los-estratos.pdf [accessed 2025-12-04]
41. La estratificación socioeconómica en el régimen de los servicios públicos domiciliarios [Socioeconomic stratification in the public utility services system]. : Departamento Administrativo Nacional de Estadística [National Administrative Department of Statistics] URL: https://www.dane.gov.co/files/geoestadistica/Estratificacion_en_SPD.pdf [accessed 2025-12-04]
42. Data sources. Avicenna Learn. URL: https://learn.avicennaresearch.com/reference/data-sources [accessed 2024-12-04]

XSL•FO
RenderX

43. Gansner M, Nisenson M, Lin V, Carson N, Torous J. Piloting smartphone digital phenotyping to understand problematic internet use in an adolescent and young adult sample. Child Psychiatry Hum Dev 2023 Aug;54(4):997-1004. [doi: 10.1007/s10578-022-01313-y] [Medline: 35044580]
44. Leaning IE, Ikani N, Savage HS, et al. From smartphone data to clinically relevant predictions: a systematic review of digital phenotyping methods in depression. Neurosci Biobehav Rev 2024 Mar;158:105541. [doi: 10.1016/j.neubiorev.2024.105541] [Medline: 38215802]
45. Lee K, Lee TC, Yefimova M, et al. Using digital phenotyping to understand health-related outcomes: a scoping review. Int J Med Inform 2023 Jun;174:105061. [doi: 10.1016/j.ijmedinf.2023.105061] [Medline: 37030145]
46. Kostyrka-Allchorne K, Stoilova M, Bourgaize J, Rahali M, Livingstone S, Sonuga-Barke E. Review: Digital experiences and their impact on the lives of adolescents with pre-existing anxiety, depression, eating and nonsuicidal self-injury conditions - a systematic review. Child Adolesc Ment Health 2023 Feb;28(1):22-32. [doi: 10.1111/camh.12619] [Medline: 36478091]
47. Colombia avanza en su meta de estar conectada en un 70 % en 2022: DANE [Colombia is making progress towards its goal of achieving 70% connectivity by 2022: DANE]. Ministerio de Tecnologías de la Información y las Comunicaciones [Ministry of Information and Communications Technologies]. 2021. URL: https://www.mintic.gov.co/portal/inicio/Sala-de-prensa/182108:Colombia-avanza-en-su-meta-de-estar-conectada-en-un-70-en-2022-DANE [accessed 2024-11-28]
48. Digital 2023 Colombia: The essential guide to the latest connected behaviours. We are social. 2023. URL: https://es.slideshare.net/DataReportal/digital-2023-colombia-february-2023-v01 [accessed 2024-11-28]
49. WHO guideline: recommendations on digital interventions for health system strengthening. World Health Organization. 2019. URL: https://www.who.int/publications/i/item/9789241550505 [accessed 2024-11-28]
50. Lehtimaki S, Martic J, Wahl B, Foster KT, Schwalbe N. Evidence on digital mental health interventions for adolescents and young people: systematic overview. JMIR Ment Health 2021 Apr 29;8(4):e25847. [doi: 10.2196/25847] [Medline: 33913817]
51. Hollis C, Falconer CJ, Martin JL, et al. Annual Research Review: digital health interventions for children and young people with mental health problems – a systematic and meta - review. Child Psychology Psychiatry 2017 Apr;58(4):474-503 [FREE Full text] [doi: 10.1111/jcpp.12663]
52. Zhong S, Yang X, Pan Z, et al. The usability, feasibility, acceptability, and efficacy of digital mental health services in the COVID-19 pandemic: scoping review, systematic review, and meta-analysis. JMIR Public Health Surveill 2023 Feb 13;9:e43730. [doi: 10.2196/43730] [Medline: 36634261]
53. Kim J, Aryee LMD, Bang H, et al. Effectiveness of digital mental health tools to reduce depressive and anxiety symptoms in low- and middle-income countries: systematic review and meta-analysis. JMIR Ment Health 2023 Mar 20;10:e43066. [doi: 10.2196/43066] [Medline: 36939820]
54. Jiménez-Molina Á, Franco P, Martínez V, Martínez P, Rojas G, Araya R. Internet-based interventions for the prevention and treatment of mental disorders in Latin America: a scoping review. Front Psychiatry 2019;10:664. [doi: 10.3389/fpsyt.2019.00664] [Medline: 31572242]
55. Martínez P, Rojas G, Martínez V, Lara MA, Pérez JC. Internet-based interventions for the prevention and treatment of depression in people living in developing countries: a systematic review. J Affect Disord 2018 Jul;234:193-200. [doi: 10.1016/j.jad.2018.02.079] [Medline: 29529553]
56. Patton GC, Sawyer SM, Santelli JS, et al. Our future: a lancet commission on adolescent health and wellbeing. The Lancet 2016 Jun;387(10036):2423-2478. [doi: 10.1016/S0140-6736(16)00579-1]
57. Wies B, Landers C, Ienca M. Digital mental health for young people: a scoping review of ethical promises and challenges. Front Digit Health 2021;3:697072. [doi: 10.3389/fdgth.2021.697072] [Medline: 34713173]
58. Bevan Jones R, Thapar A, Rice F, et al. A digital intervention for adolescent depression (MoodHwb): mixed methods feasibility evaluation. JMIR Ment Health 2020 Jul 17;7(7):e14536. [doi: 10.2196/14536] [Medline: 32384053]
59. Davies SM, Jardine J, Gutridge K, et al. Preventive digital mental health for children in primary schools: acceptability and feasibility study. JMIR Form Res 2021 Dec 13;5(12):e30668. [doi: 10.2196/30668] [Medline: 34898446]
60. Piers R, Williams JM, Sharpe H. Review: Can digital mental health interventions bridge the “digital divide” for socioeconomically and digitally marginalised youth? A systematic review. Child Adolesc Ment Health 2023 Feb;28(1):90-104. [doi: 10.1111/camh.12620] [Medline: 36510375]
61. d’Halluin A, Costa M, Morgiève M, Sebbane D. Attitudes of children, adolescents, and their parents toward digital health interventions: scoping review. J Med Internet Res 2023 May 2;25:e43102. [doi: 10.2196/43102] [Medline: 37129931]
62. Bear HA, Ayala Nunes L, Ramos G, et al. The acceptability, engagement, and feasibility of mental health apps for marginalized and underserved young people: systematic review and qualitative study. J Med Internet Res 2024 Jul 30;26:e48964. [doi: 10.2196/48964] [Medline: 39078699]
63. Eysenbach G. The law of attrition. J Med Internet Res 2005 Mar 31;7(1):e11. [doi: 10.2196/jmir.7.1.e11] [Medline: 15829473]
64. Garrett SM, Hilder J, Tester R, et al. Young people talk about digital support for mental health: an online survey of 15-30-year olds in New Zealand. Health Expect 2024 Aug;27(4):e70001. [doi: 10.1111/hex.70001] [Medline: 39188108]
65. Kim MO, Coiera E, Magrabi F. Problems with health information technology and their effects on care delivery and patient outcomes: a systematic review. J Am Med Inform Assoc 2017 Mar 1;24(2):246-250. [doi: 10.1093/jamia/ocw154]

XSL•FO
RenderX

66. The most common bugs found in healthcare apps. Dogtown Media. 2023. URL: https://www.dogtownmedia.com/the-most-common-bugs-found-in-healthcare-apps [accessed 2024-11-28]
67. Leech T, Dorstyn D, Taylor A, Li W. Mental health apps for adolescents and young adults: a systematic review of randomised controlled trials. Child Youth Serv Rev 2021 Aug;127:106073. [doi: 10.1016/j.childyouth.2021.106073]
68. Carson NJ, Gansner M, Khang J. Assessment of digital media use in the adolescent psychiatric evaluation. Child Adolesc Psychiatr Clin N Am 2018 Apr;27(2):133-143. [doi: 10.1016/j.chc.2017.11.003] [Medline: 29502741]
69. Wong CA, Madanay F, Ozer EM, et al. Digital health technology to enhance adolescent and young adult clinical preventive services: affordances and challenges. J Adolesc Health 2020 Aug;67(2S):S24-S33. [doi: 10.1016/j.jadohealth.2019.10.018] [Medline: 32718511]
70. Reid Chassiakos YL, Radesky J, Christakis D, Moreno MA, Cross C, COUNCIL ON COMMUNICATIONS AND MEDIA. Children and adolescents and digital media. Pediatrics 2016 Nov;138(5):e20162593. [doi: 10.1542/peds.2016-2593] [Medline: 27940795]
71. Filkins BL, Kim JY, Roberts B, et al. Privacy and security in the era of digital health: what should translational researchers know and do about it? Am J Transl Res 2016;8(3):1560-1580. [Medline: 27186282]
72. Insel TR. Digital phenotyping: technology for a new science of behavior. JAMA 2017 Oct 3;318(13):1215-1216. [doi: 10.1001/jama.2017.11295] [Medline: 28973224]
73. Choi A, Ooi A, Lottridge D. Digital phenotyping for stress, anxiety, and mild depression: systematic literature review. JMIR Mhealth Uhealth 2024 May 23;12:e40689. [doi: 10.2196/40689] [Medline: 38780995]
74. Wacks Y, Weinstein AM. Excessive smartphone use is associated with health problems in adolescents and young adults. Front Psychiatry 2021;12(1–7):669042. [doi: 10.3389/fpsyt.2021.669042] [Medline: 34140904]
75. Weinstein AM, Siste K. Editorial: Excessive and problematic smartphone usage, volume II. Front Psychiatry 2023;14:1181652. [doi: 10.3389/fpsyt.2023.1181652] [Medline: 37181881]
76. Wilhite K, Booker B, Huang BH, et al. Combinations of physical activity, sedentary behavior, and sleep duration and their associations with physical, psychological, and educational outcomes in children and adolescents: a systematic review. Am J Epidemiol 2023 Apr 6;192(4):665-679. [doi: 10.1093/aje/kwac212] [Medline: 36516992]
77. Menezes-Júnior LD, de Moura SS, Miranda AG, de Souza Andrade AC, Machado-Coelho GLL, Meireles AL. Sedentary behavior is associated with poor sleep quality during the COVID-19 pandemic, and physical activity mitigates its adverse effects. BMC Public Health 2023 Jun 12;23(1):1116. [doi: 10.1186/s12889-023-16041-8] [Medline: 37308867]
78. Boonstra TW, Nicholas J, Wong QJ, Shaw F, Townsend S, Christensen H. Using mobile phone sensor technology for mental health research: integrated analysis to identify hidden challenges and potential solutions. J Med Internet Res 2018 Jul 30;20(7):e10131. [doi: 10.2196/10131] [Medline: 30061092]
79. Caracterización de vivienda y población de la zona rural de bogotá d.c. [Characterization of housing and population in the rural area of Bogotá D.C.]. : Alcaldía mayor de bogotá, secretaría distrital de planeación [Bogotá's Mayor's Office, District Planning Secretariat]; 2019 URL: https://www.sdp.gov.co/sites/default/files/caracterizacion_ruralidad_vf.pdf [accessed 2025-12-04]
80. Ospina-Pinillos L, Davenport TA, Ricci CS, Milton AC, Scott EM, Hickie IB. Developing a mental health eClinic to improve access to and quality of mental health care for young people: using participatory design as research methodologies. J Med Internet Res 2018 May 28;20(5):e188. [doi: 10.2196/jmir.9716] [Medline: 29807878]
81. Alqahtani F, Orji R. Insights from user reviews to improve mental health apps. Health Informatics J 2020 Sep;26(3):2042-2066. [doi: 10.1177/1460458219896492] [Medline: 31920160]

Abbreviations

GBD: Global Burden of Disease
ICT: Information and Communication Technology
LMIC: low- or middle-income country
MB: Mental Beat
MH: mental health
NMHS: National Mental Health Survey
NPS: Net Promoter Scale
PUJ: Pontificia Universidad Javeriana
REDCap: Research Electronic Data Capture
SUS: System Usability Scale
YMC: Youth Collective Minds

XSL•FO
RenderX

Edited by A Stone; submitted 13.12.24; peer-reviewed by GA Garavito, KW Tay; revised version received 13.11.25; accepted 13.11.25; published 29.12.25.

Please cite as:
Ospina-Pinillos L, Shambo-Rodríguez DL, Riaño-Fonseca MI, Sánchez Nítola MN, Ramírez-Castro MF, Calvo-Valderrama MG, Camacho S, Gómez-Restrepo C, Navarro-Mancilla AA, Hickie IB, Occhipinti JA
Implementing Digital Tools for Mental Health Support in Young Individuals in Colombia: Mixed Methods Feasibility Study
JMIR Form Res 2025;9:e69749
URL: https://formative.jmir.org/2025/1/e69749
doi:10.2196/69749

© Laura Ospina-Pinillos, Débora L Shambo-Rodríguez, María Isabel Riaño-Fonseca, Mónica Natalí Sánchez Nítola, María Fernanda Ramírez-Castro, María Gabriela Calvo-Valderrama, Salvador Camacho, Carlos Gómez-Restrepo, Alvaro A Navarro-Mancilla, Ian B Hickie, Jo-An Occhipinti. Originally published in JMIR Formative Research (https://formative.jmir.org), 29.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Development of an mHealth App Prototype for LGBTQIA+ Individuals' Sexual and Reproductive Health in Gauteng Province, South Africa: Design Science Research Study

Raikane James Seretlo[1], BCUR, MPH, MBA, PhD; Mathildah Mpata Mokgatle[1], PhD; Hanlie Smuts[2], PhD; Nombulelo Veronica Sepeng[3], PhD

[1]Department of Public Health, School of Health Care Sciences, Sefako Makgatho Health Sciences University, Molotlegi Street, Ga-Rankuwa, Pretoria, South Africa

[2]Department of Informatics, Faculty of Engineering, Built Environment and Information Technology, University of Pretoria, Pretoria, South Africa

[3]Department of Nursing Science, Faculty of Health Sciences, University of Pretoria, Pretoria, South Africa

Corresponding Author:
Raikane James Seretlo, BCUR, MPH, MBA, PhD
Department of Public Health, School of Health Care Sciences, Sefako Makgatho Health Sciences University, Molotlegi Street, Ga-Rankuwa, Pretoria, South Africa

Abstract

Background: The fast rate of technological advances in the health care sector remain as a pressing need for effective solutions that address the unique health care needs of sexual and gender minorities. If these innovative solutions are considered, societal challenges such as stigma, discrimination, and a lack of tailored health care resources, as experienced by the lesbian, gay, bisexual, transgender, queer, intersex, asexual, and more (LGBTQIA+) individuals could be addressed at lower cost.

Objective: This study aimed to develop a mobile health (mHealth) app specifically designed to address the sexual and reproductive health (SRH) of lesbian, gay, bisexual, transgender, queer, intersex, asexual, and more (LGBTQIA+) individuals in Gauteng Province, South Africa.

Methods: This study used a Design Science Research (DSR) framework and a mixed-method exploratory sequential approach. DSR was executed in three cycles: cycle one followed an ethnography approach; involved 33 health care providers (HCPs) and 22 LGBTQIA+ individuals, focusing on identifying specific SRH and the challenges encountered in accessing and providing these services. Participants shared their views on the potential role of mHealth apps in addressing these issues. Cycle two engaged 13 experts through Participatory Action Research (PAR) approach using the Nominal Group Technique (NGT) to collaboratively identify essential content for the app, fostering a co-creation process. Lastly, cycle three followed an interventional pre-experimental approach by involving software developers and principal investigator working together to develop a functional prototype of the mHealth app.

Results: This study revealed critical insights into the specific SRH of LGBTQIA+ individuals, alongside the barriers faced by health care providers in meeting these needs. The co-created app prototype named *"Queery wellness hub"* was developed and incorporated features tailored to enhance accessibility, confidentiality, and user engagement, addressing both user and provider perspectives.

Conclusions: The findings underscore the potential of mHealth apps in transforming the delivery of SRH for LGBTQIA+ individuals in South Africa. Continued collaboration with stakeholders is essential for further refinement and successful implementation of the app, ultimately contributing to better health outcomes for sexual and gender minorities.

(JMIR Form Res 2025;9:e79593) doi:10.2196/79593

KEYWORDS

mHealth app; prototype; sexual and reproductive health; LGBTQIA+; wellness; mobile health; Lesbian, Gay, Bisexual, Transgender, Queer, Intersex, Asexual+

Introduction

Mobile health (mHealth) is the delivery of health care services through mobile device communications and technology such as mobile phones, monitoring of patients' devices, and individual virtual assistants [1]. The adoption and increasing attention to mobile health technologies are rising across multiple sectors, including the health care sector [2,3]. This change penetrates and affects the world at large regardless of different countries' economic status and is moving from the high-income countries (HICs) to the low-income countries (LICs) [1,4]. The

XSL•FO
RenderX

aim of this study was to develop an mHealth app designed for health care professionals (HCPs) and LGBTQIA+ individuals in Gauteng Province, South Africa, to address SRH needs and improve the overall health and well-being of LGBTQIA+ individuals.

However, the question remains as to whether every country is ready for this trend of what is called "The digital era," not only for the use and adoption of technological tools and solutions but also addressing the health care issues and burden of diseases for all. Nevertheless, several studies have shown that the use of technology and artificial intelligence (AI) can help to solve and address some of the health care problems. For example, Mirbabaie et al [5], demonstrated that AI is playing a significant role in enhancing the health care state, such as assisting with disease detection and diagnosis, personalized medicine [6], faster verification of the drug target, and optimizing drug structure design [7].

Research demonstrated that it improves health care availability and quality in a variety of global regions [1] and promotes efficiency in areas such as the teamwork process, making interprofessional communication more effective and accurate, lowering risks [8,9]. In addition, the use of mHealth influences clinical results, such as decreases in blood pressure, glycosylated hemoglobin (HbA_{1c}), and cholesterol, and assists with improved compliance with medications and physical activity, and certain psychological difficulties, such as reduced anxiety and increased patient satisfaction, were found [9]. Lastly Rowland et al [10], showed that it aids with preventative behavior change, virtual self-management of a specific ailment, and diagnosis .

Furthermore, with regard to the health and wellbeing of LGBTQIA+ individuals, various studies have shown their use of various mHealth and innovative tools and solutions, there was a rise in pre-exposure prophylaxis (PrEP) adherence, a substantial rise in complete HIV knowledge and treatment, accessibility to sexual health knowledge about good sexual behaviors, and some were reminded about their HIV and sexually transmitted infections (STIs) tests [11-17].

SRH services remain one of the important needs of human beings for a holistic care approach. Access to sexual and reproductive health services is a fundamental human right that should be readily available to everyone at all times as part of a universal healthcare system [18]. However, in South Africa, LGBTQIA+ individuals continue to experience systemic discrimination and a heteronormative healthcare framework that challenge in accessing and utilizing the SRH services [19]. These challenges include stigma and discrimination [20-22], prejudice, unsupportive structural, and limited SRH services specific to their needs such as gender-affirming, adoption, and fertility services [23-26].

This does not only affect LGBTQIA+ individuals but also HCPs, as they have limited knowledge, skills, and expertise regarding SRH for LGBTQIA+ individuals [22-25]. Most HCPs in South Africa are not equipped with the type of SRH services related to LGBTQIA+ individuals due to their previous educational curriculum and lack of resources to provide relevant SRH services for LGBTQIA+ individuals [23,24,27-30]. The above challenges not only affect South Africa but also global and African areas [31-38].

By introducing easy, cost-effective, and user-friendly innovative solutions such as mHealth apps some of the above-mentioned challenges could be minimized and addressed. However, should the mentioned challenges not be addressed, LGBTQIA+ individuals could be at risk of negative health issues such as HIV or AIDS, STIs such as, mental health issues, and other health risks such as use of unregistered HCPs [39-44].

In South Africa, there is no specific developed mHealth app to be used by both the HCPs and LGBTQIA+ individuals for addressing SRH, thus improving the health and well-being of LGBTQIA+ individuals. Hence, our study developed an mHealth app prototype called *"Queery Wellness Hub"* to address the SRH of LGBTQIA+ individuals in Gauteng Province, South Africa. This article reports on how *"Queery Wellness Hub"* was developed using the Design Science Research (DSR) framework. The *"Queery Wellness Hub"* hopes to close gaps in the lack of knowledge, skills and expertise experienced by HCPs and address the SRH services for LGBTQIA+ individuals when accessing and utilizing these services.

Methods

Overview

This study reports the processes followed during the development of an mHealth app named *Queery Wellness Hub*. A detailed focus will mainly be on cycle three as it is the developed outcome and was not published anywhere, cycle one and two will be just a summary to show how these steps enhanced the final development. The Queer wellness hub was developed using Design Science Research (DSR) framework. The DSR framework was followed and applied as the main guide for the study. The principal investigator aimed at using the DSR framework developed by Kuechler et al [45]. This is an approach that follows sequential five steps namely, awareness, suggestion, development, evaluation, and conclusion. The DSR framework steps overlapped, and several iterations took place, especially during development [46]. As a result, the research questions and objectives were classified according to the DSR framework steps [47]. See Figure 1 for a summary of the DSR framework steps.

XSL•FO
RenderX

Figure 1. Design science research general steps [47].

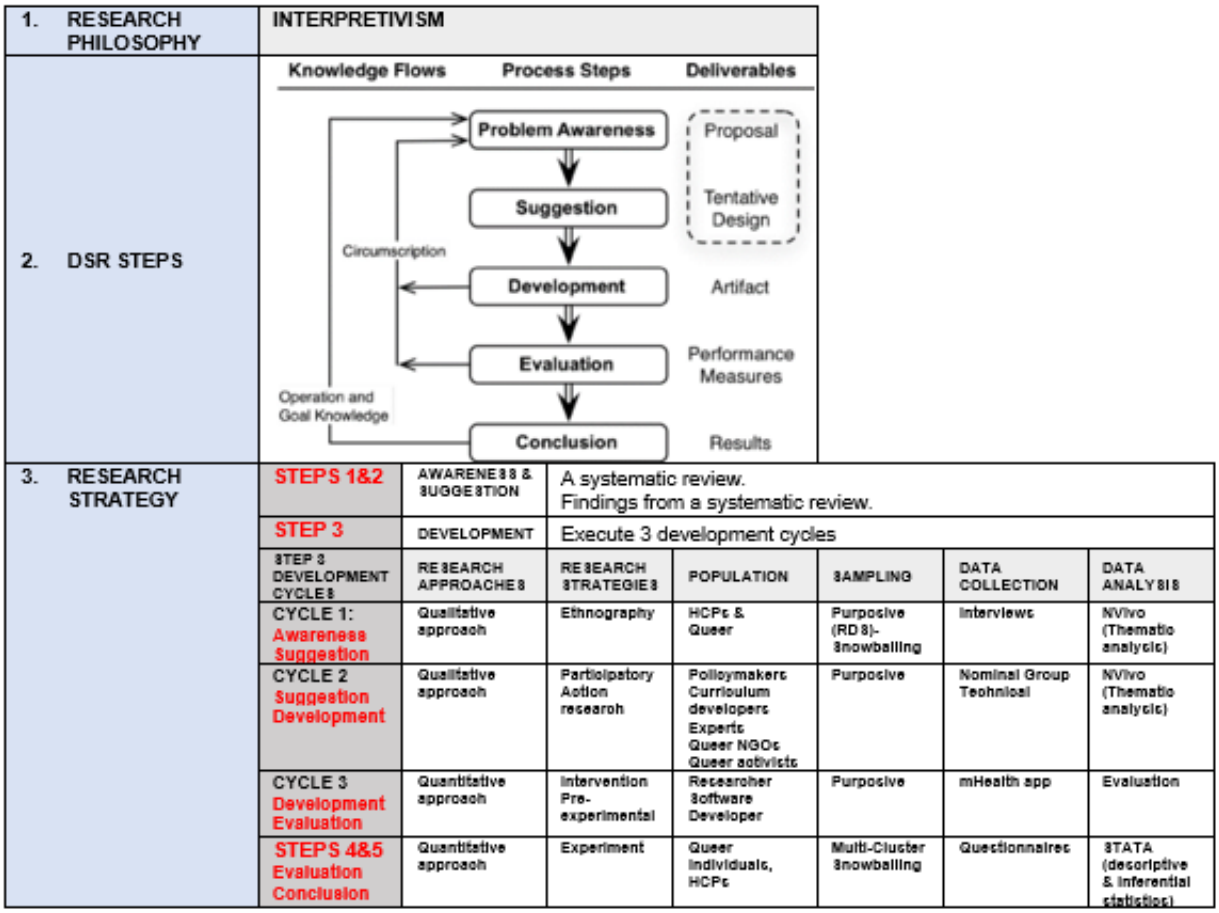

Operationalization of Design Science Research (DSR) Framework

This phase of the study used step three (developmental) of the DSR framework according to Vaishnavi and Kuechler [47]. This step comprised of three cycles. For clarity, while all cycles informed the development, this manuscript emphasizes the findings of cycle three, with cycles one and two described to show how they contributed to the ultimate development (Figure 2).

XSL•FO
RenderX

Figure 2. A proposed conceptual framework for developing a mHealth app for addressing SRH of LGBTQIA+ individuals adapted from HEPM & DSR.

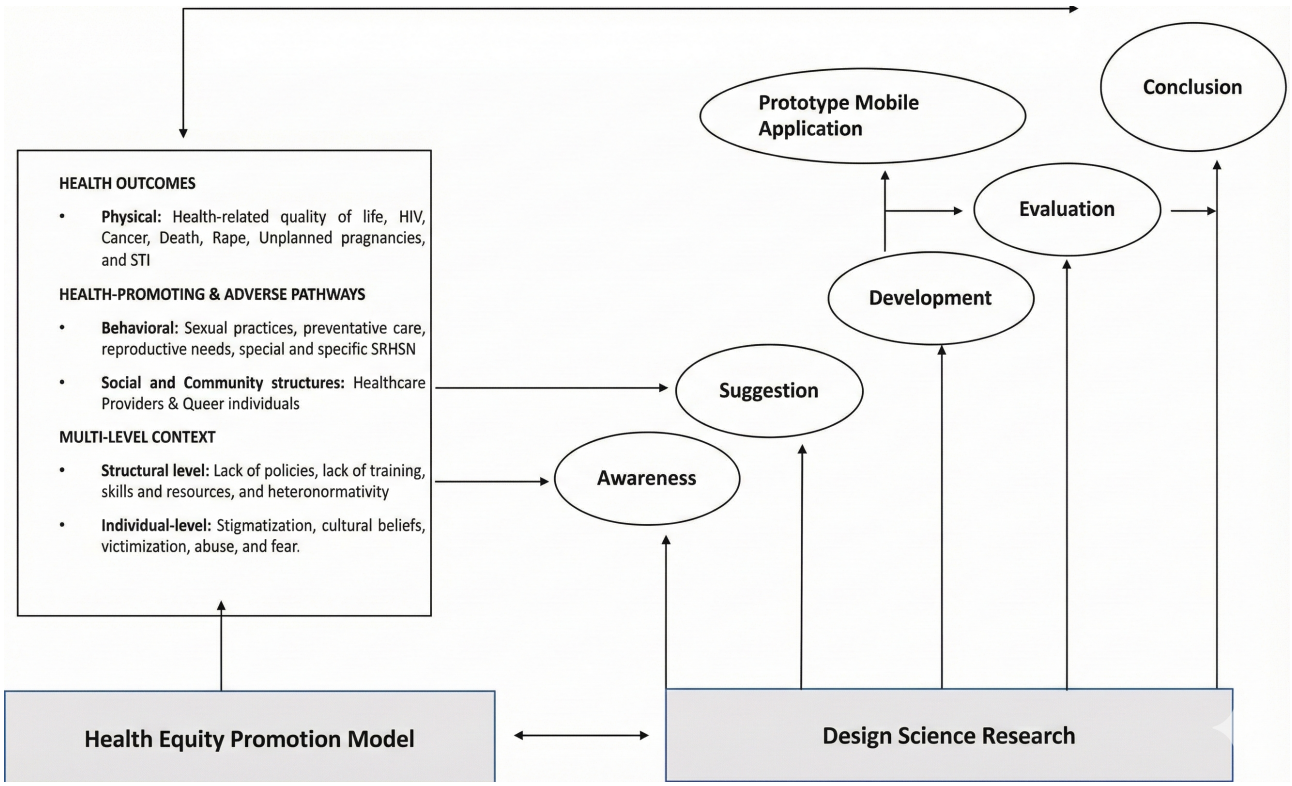

Cycle one covered step one (awareness), and two (suggestion) of the DSR framework. The awareness step covered the first three objectives, and the suggestion step covered last two objectives as described sub-section five. An ethnography qualitative approach was used during this cycle because it helped the principal investigator to deeply understand the specific sexual-reproductive health care needs of LGBTQIA+ individuals [27], find out the challenges faced by HCPs and LGBTQIA+ individuals [24], identify solutions to address the identified challenges [48], and explore their thoughts on using web-based tools and mHealth apps to address these needs [49]. 33 HCPs were selected across seven district hospitals in Gauteng province using purposive sampling as they were the relevant participants to provide rich information about provision of SHR, whereas 22 LGBTQIA+ individuals were selected from the one nongovernmental organization (NGO) which had two clinics specializing in health using respondent-driven sampling (RDS). The principal investigator started with two LGBTQIA+ members attending the clinic, then asked them to refer other LGBTQIA+ individuals until data was saturated (See the demographic data of the LGBTQIA+ individuals in Multimedia Appendices 1 and 2). For both participants, data was collected in private rooms using unstructured interview guide and all the interviews' sessions were recorded using digital recorder. Interviews for HCPs were mainly in English as most of them were familiar with the language and for LGBTQIA+ individuals, they included a mixture of English and Isizulu which was a commonly used language and was later translated. Thematic analysis was used to analyze cycle one data using NVivo software 14.

Cycle two focused on the DSR framework step two (suggestion) which included the findings and recommendations from the experts and three (development) which involved a group of experts voting and ranking for the important content that needed to be included in the mHealth app development through the use of NGT. This cycle used a qualitative approach, specifically participatory action research (PAR) which helped with identifying the important SRH content to be included in a mHealth app thus assisted with the creation of a mHealth app to address SRH for LGBTQIA+ individuals. This qualitative approach allowed for a more in-depth understanding of the specific SRH topics that were to be considered when designing a mHealth app, assuring its relevance and efficacy in meeting the needs of LGBTQIA+ individuals seeking SRH [50]. This approach was executed in the form of NGT, a face-to-face workshop was conducted where all four NGT steps by were applied [51]. The NGT included participants such as the principal investigator, queer activists, sexual and reproductive health experts, private practicing health care providers, innovators, and private health care stakeholdersMultimedia Appendix 3. These participants were recruited via emails, from different social media platforms, and scholarly platforms based on their research interest and focus. The NGT had one moderator, two research assistants, and one principal investigator. The workshop lasted approximately 2 hours 46 minutes and 55 seconds. NGT was well-suited to this goal since it actively engaged the group of experts as co-principal investigator, allowed them to contribute their thoughts, experiences, and needs during the identification of the specific SRH topics for LGBTQIA+ individuals that needed to be included in a mHealth app(Multimedia Appendix 4).

The last cycle was cycle three. The cycle involved the prototype development guided by findings from both cycle one and two. This cycle encompassed the principal investigator and software engineers. This cycle applied a quantitative approach because

XSL•FO
RenderX

the principal investigator was interested in the development of a prototype mHealth app. An interventional pre-experimental approach was used, which helped with ensuring that there are accuracy and inclusiveness of the prototype's contents. After synthesizing and merging findings from cycles one and two, cycle three commenced, whereby it focused on the mHealth app prototype development. The principal investigator searched for software developers, managed to find one, and then nominated them as the developer based on their work experience. Data collection involved documenting and confirming the synthesized findings from the cycle one and two, with the principal investigator assessing clinical content and the software developer assessing technical functionality and interface. Both the developer and the principal investigator had regular face-to-face meetings which facilitated measurement and verification of the prototype's features, effectively applying quantitative evaluation to the development process. The following steps were followed and repeated when developing the prototype: principal investigator compiled findings from cycles 1 and 2 and communicated them with the software developer; the prototype plan was conceptualized by both the principal investigator and the software developer, the software developer provided feedback on the design of the essential software components and the integration of functionalities and their potential impacts in addressing SRH for LGBTQIA+ individuals.

Conceptual Framework of the Study

This study has adapted its conceptual framework from two theoretical models. Firstly, the Health Equity Promotion Model (HEPM) ensures the elimination of disparities in healthcare. For example, this model was used by Fredriksen-Goldsen [52] in their study to reconceptualize the health disparities of some LGBTQIA+ individuals. As a result, this model formed the basis for promoting SRH, thus answering the research questions of the specific SRH needs and the challenges experienced by both HCPs and LGBTQIA+ individuals.

Moreover, the proposed conceptual framework for developing a mHealth app for addressing the SRH of LGBTQIA+ individuals will use the DSR steps as the second framework and main guide for the study. This was due to the steps' widespread use as a problem-solving technique for creating innovative artifacts such as healthcare mHealth apps [53]. For example, the DSR framework steps were used by a few studies in addressing health care issues such as mental health, and noncommunicable diseases [54,55].

In addition, this method assisted in answering and addressing almost all the research questions, but the study applied only three stipulated steps awareness, suggestion, and development. These steps are interrelated and overlap with each other. For example, *awareness and suggestion* were included in cycle one, *suggestion, and development* included in cycle two, and *development and evaluation* were covered after cycle three. Last, *evaluation and conclusion* are currently running as a post-doctoral project to test the prototype in step four, and five, and will take it back to *awareness.*

Ethical Considerations

The study received ethical approval from the Sefako Makgatho Health Sciences University Research Ethics Committee (SMUREC) (Reference Number: SMUREC/H/291/2023:PG) after being presented to and approved by the Department of Public Health Research Committee (DPHRC) and the School of Healthcare Sciences Research Committee (SHRC). Following ethical approval, the principal investigator obtained permission to conduct the study at selected Gauteng District hospitals through the National Health Research Database (NHRD), and separately requested permission from the POP INN clinic's institutional management using the same clearance.

Written informed consent was obtained from all participants, who were informed about the study's aims, objectives, and their rights, including the voluntary nature of participation and the freedom to withdraw at any point without negative consequences. Participants' identities were protected through the use of pseudonyms and numerical codes (eg, P1 for Participant 1, E1 for Expert 1), and no names or surnames were recorded in any data tools, including interview guides and recordings.

To ensure privacy, confidentiality, and compliance with the Protection of Personal Information Act (POPIA), only minimal personal information was collected. All data were securely stored under lock and will be retained for five years, with controlled access. Participants' personal information was not shared with third parties, and study findings were only shared with supervisors, transcribers, and through accredited journals or conferences. Experts who participated in workshops were also informed of their rights and were reimbursed R1000 (US $67) for transportation and fuel costs.

Results

Summary of Cycle Three Findings

The following information is about the descriptions of the developed mHealth app (Prototype) and its potential of addressing identified SRH for LGBTQIA+ individuals.

About *Queery Wellness Hub*

Queery Wellness Hub is a health care hub that aims at addressing the SRH needs for the LGBTQIA+ individuals in Gauteng Province, South Africa. It is intended to be used by HCPs during health care provision, clients' consultations, referrals, and as an information hub. Additionally, it is intended to be used by LGBTQIA+ individuals when accessing their SRH. An overall aim of the *Queery Wellness Hub* app is to address the issues of inclusivity and accessibility to health care for LGBTQIA+ individuals.

How Does *Queery Wellness Hub* Work?

Queery Wellness Hub is designed to provide a centralized, accessible, and reliable source of health care information for queer individuals. It aims to create a supportive and informative space where users can access expert-verified content, connect with like-minded individuals, and engage with topics relevant to their health and well-being.

XSL•FO
RenderX

Structure of the App (*Queery Wellness Hub*)

- Communities: "Communities" act as the information seeking part of the app
- Health care: This is where users are able to access a network of healthcare practitioners who provides all queer-related sexual and reproductive healthcare services and needs.
- Messaging: Here, users are able to communicate with practitioners and have in-app consultations
- Profile Management: Users are able to manage their profile and app preferences

How Was *Queery Wellness Hub* Developed?

The app was developed from scratch through an empirical study using the *Design Science Research (DSR) framework* steps by Vaishnavi & Kuechler [47]. The main developer of the app is Dr Raikane James Seretlo. The principal investigator applied various series of cycles to develop this app, including a systematic review, qualitative approaches with HCPs and LGBTQIA+ individuals were included, a group of experts, and a software developer. The findings from the empirical cycles recommended the SRH topics that should be included in the app.

What Are the Benefits of *Queery Wellness Hub*?

Queery Wellness Hub is intended to be used by different HCPs as a consultative resource whenever they are in contact with LGBTQIA+ individuals. It may support sexual and reproductive health (SRH) of LGBTQIA+ individuals and aims to enhance equitable access to healthcare. The app could contribute to supporting SRH knowledge and service uptake among LGBTQIA+ individuals. The app may also support HCP workflow and clinical decision-making by providing a consultative resource and facilitating access to information tailored to LGBTQIA+ individuals SRH.

What's Next for *Queery Wellness Hub*?

The post-doctoral project commenced to test *Queery Wellness Hub* for addressing SRH by both HCPs & LGBTQIA+ individuals. Evaluation of the effectiveness and functionality of *Queery Wellness Hub* in addressing SRH for LGBTQIA+ individuals. Further research is needed to study the impact of Queery: Wellness Hub on health behaviors and clinical outcomes.

Visuals of the Developed mHealth App (Prototype)

This section provides visuals of the developed mHealth app (prototype), illustrating the main outcome of the study. Not all pictures of a developed mHealth app (prototype) were included in this section, this is a summary of and visual presentation of how a developed mHealth app (prototype).

Onboarding and Access

The app welcomes users with a simple and secure login interface (Figure 3), providing an easy point of access while ensuring confidentiality. This login screen serves as the digital gateway to a personalized LGBTQIA+ health journey.

Figure 3. This screen allows users to enter their personal log-in details to access the app securely. In addition, the "About" section provides a brief background regarding the *Queery Wellness Hub*, giving users an overview of its purpose and features before they begin using it.

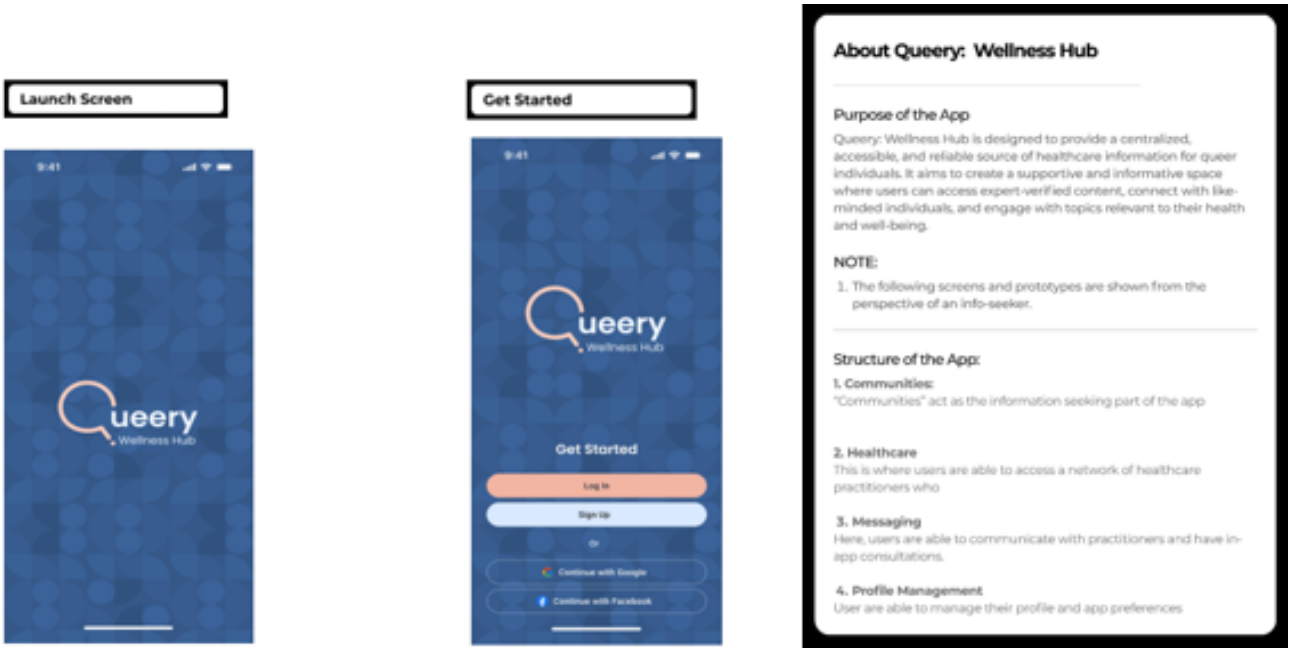

Holistic SRH Information Hub

Once logged in, users are directed to a central SRH information hub (Figure 4). This section allows individuals to access curated and diverse resources including preventive care, diagnostic services, maternal and sexual health education, and mental health support. Communities within the app are thematically organized to cater to various SRH interests and needs, enhancing relevance and engagement. Additionally, users can join different communities, providing them with various SRH data such as preventive & future health, diagnostic treatments, maternal health, sexual health, and mental health.

XSL•FO
RenderX

Figure 4. The information hub icon provides access to all sexual and reproductive health (SRH) information tailored for LGBTQIA+ individuals. It also includes different community spaces where healthcare professionals (HCPs) and LGBTQIA+ users can engage with other members of the app to discuss various health categories. Additionally, users can search for specific information related to any health-related matters through this hub.

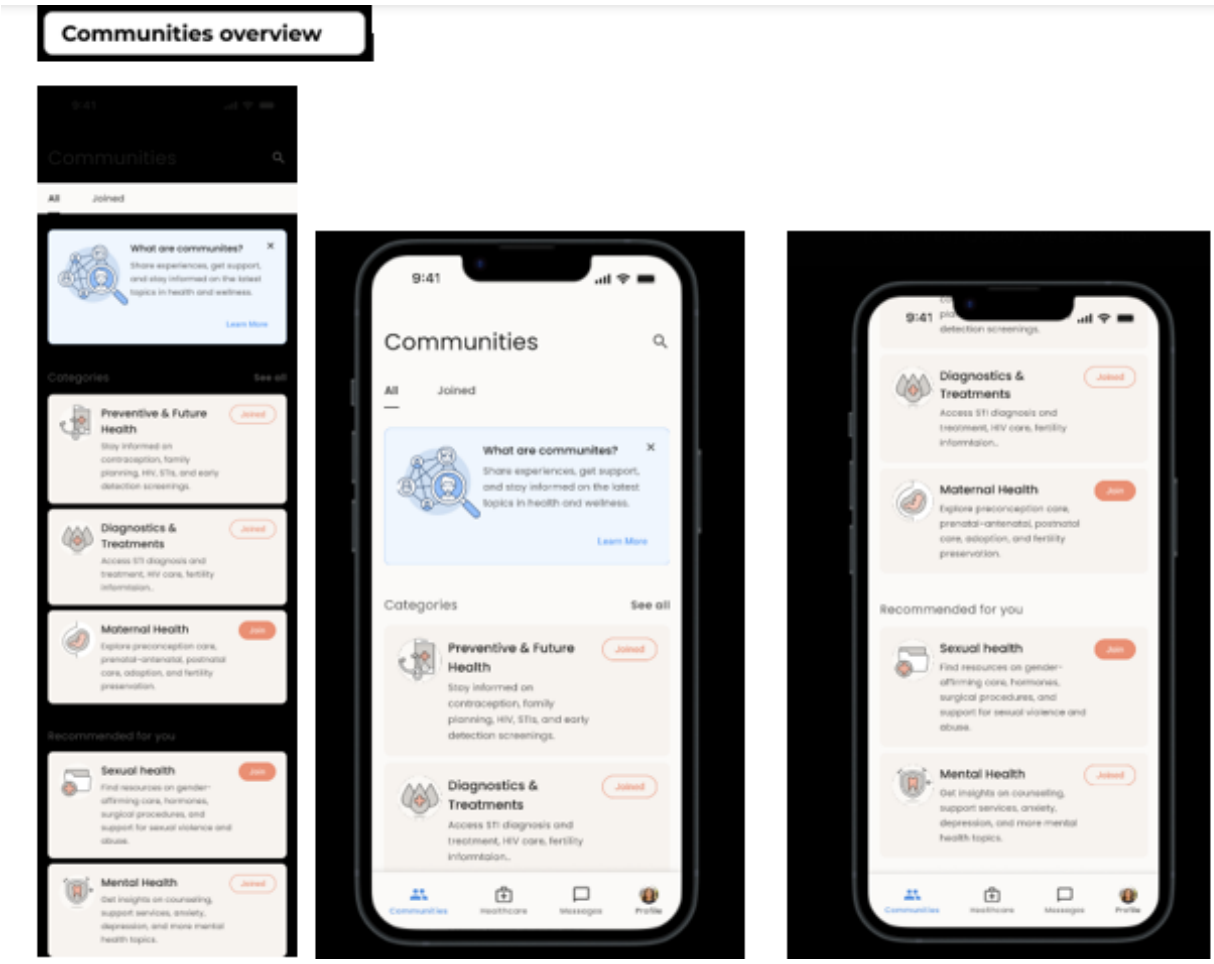

Community Interaction and Peer Support

The app promotes open and stigma-free conversations through peer-led communities (Figure 5). LGBTQIA+ users can engage in discussions on family planning, contraception, STI/HIV prevention, and other relevant topics. This platform also highlights top SRH-related articles, offering evidence-based knowledge to empower users in managing their health. Furthermore, LGBTQIA+ individuals will be able to join different communities and discuss various topics freely, such as family planning, contraception, and HIV and STI prevention (Figure 5). Additionally, they will be able to access top articles related to SRH and read more about them.

XSL•FO
RenderX

Figure 5. This section includes several topics and articles, such as those specific to LGBTQIA+ individuals and holistic SRH, including family planning, preventative health, and future health. In addition, a chat column is provided to enable users to communicate and interact with each other within the community.

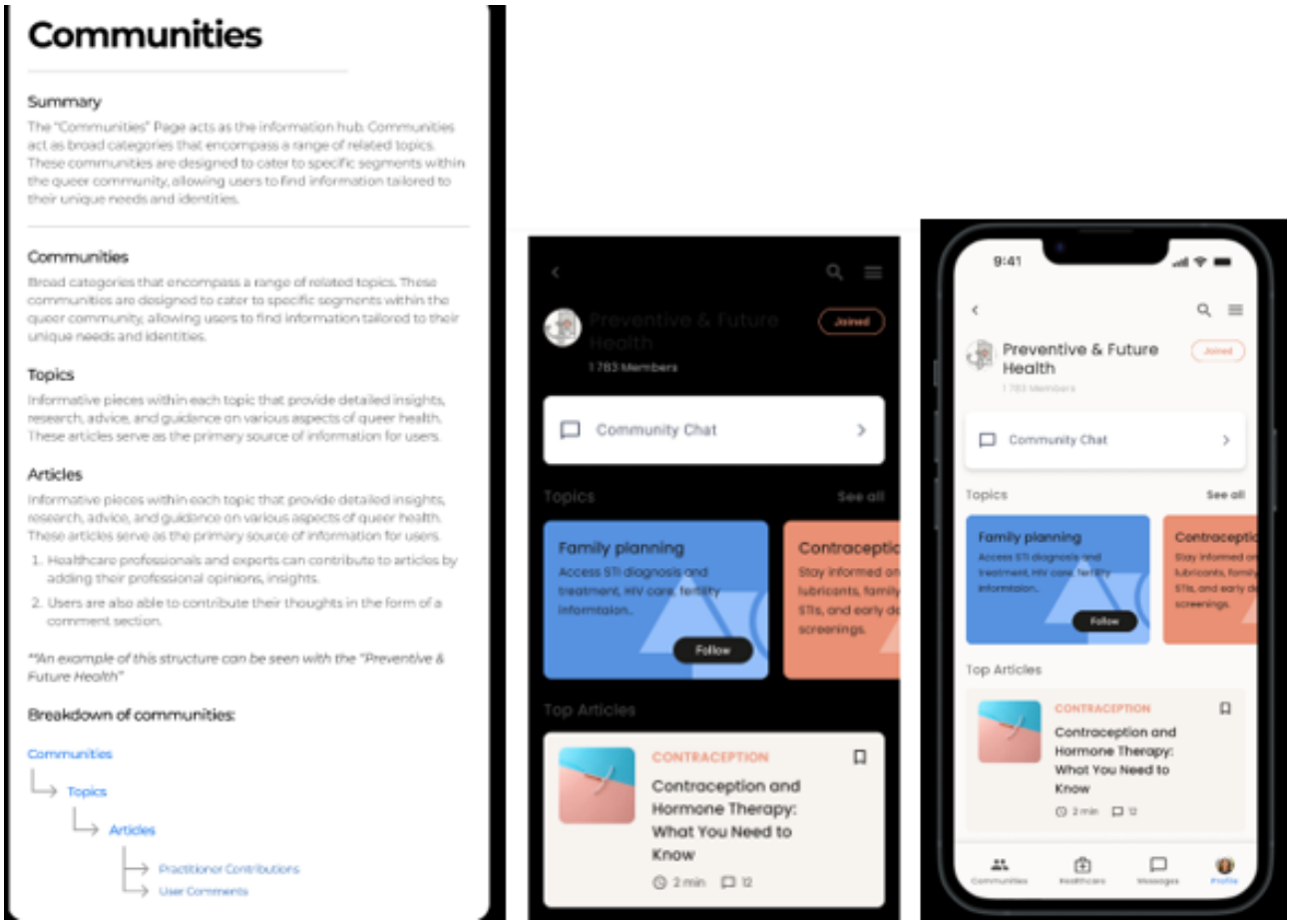

Article and Comment Section

LGBTQIA+ individuals will be able to chat among themselves and ask questions to the HCPs about all the matters relating to their SRH. This is in the form of support groups. Additionally, LGBTQIA+ individuals can access uploaded articles regarding their SRH and learn more about any of their sexual and reproductive health services and needs. This icon can also be used by the HCPs in terms of understanding the LGBTQIA+ individuals health requests and concerns.

Direct Chat With Health Professionals

Beyond peer interaction, the prototype includes a direct messaging feature (Figure 6) where users can confidentially engage with health care professionals (HCPs). This channel fosters real-time support, allows follow-up queries, and encourages ongoing health conversations tailored to queer identities. This section is about *Queery Wellness Hub*'s messaging and AI Bot, where LGBTQIA+ individual scan ask SRH questions and receive reminders for their consultations with HCPs. Messages can include AI bot, HCPs, and peers whenever seeking SRH advice.

XSL•FO
RenderX

Figure 6. The Message icon allows LGBTQIA+ users to communicate directly with their healthcare providers (HCPs) regarding any health issues. HCPs can also use this feature to send reminders, share updates, and provide consultation details to LGBTQIA+ individuals. The AI Bot icon provides users with instant support through an automated virtual assistant, offering answers to common health-related questions, guidance on navigating the app, and suggestions for appropriate resources when needed.

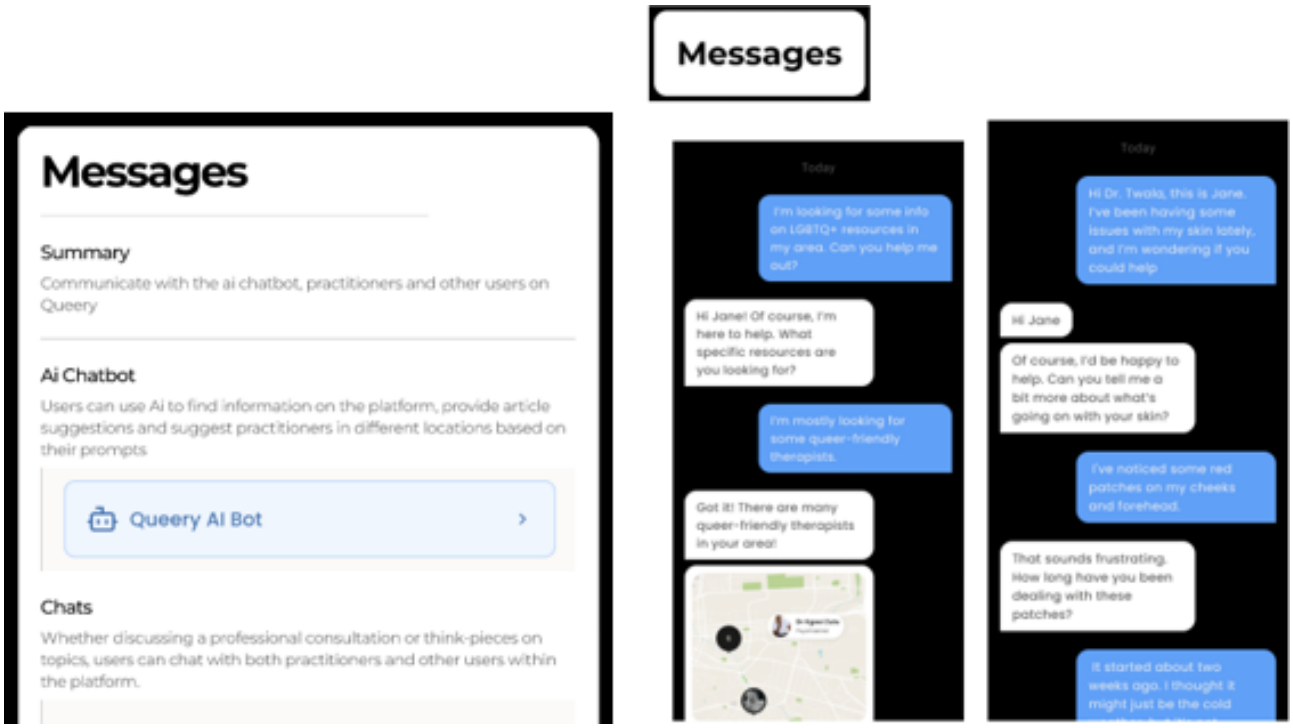

Profile and Emergency Support

User security is prioritized through the profile section (Figure 7), which includes privacy settings and emergency contact options. It also provides shortcuts to book consultations or get immediate assistance bridging the gap between digital engagement and real-world support. In addition, will provide LGBTQIA+ individuals with their security details, options to check where they can call for support and details. Further, this icon will serve as a reminder and provide them with an option to contact HCPs for consultations. There is an AI chatbot and consultation reminders under this icon. This notable feature of *Queery Wellness Hub* is the AI-powered chatbot (Figure 6). This tool not only answers SRH questions instantly but also sends reminders for upcoming consultations, ensuring continuity of care. The chatbot interacts fluidly with users, HCPs, and peer groups, serving as a smart companion in their health journey.

XSL•FO
RenderX

Figure 7. The Profile icon allows users to manage their personal information, preferences, and health-related details securely within the app. The Emergency Support icon provides immediate access to urgent resources and contacts, ensuring that LGBTQIA+ individuals can quickly reach help or support services during critical situations.

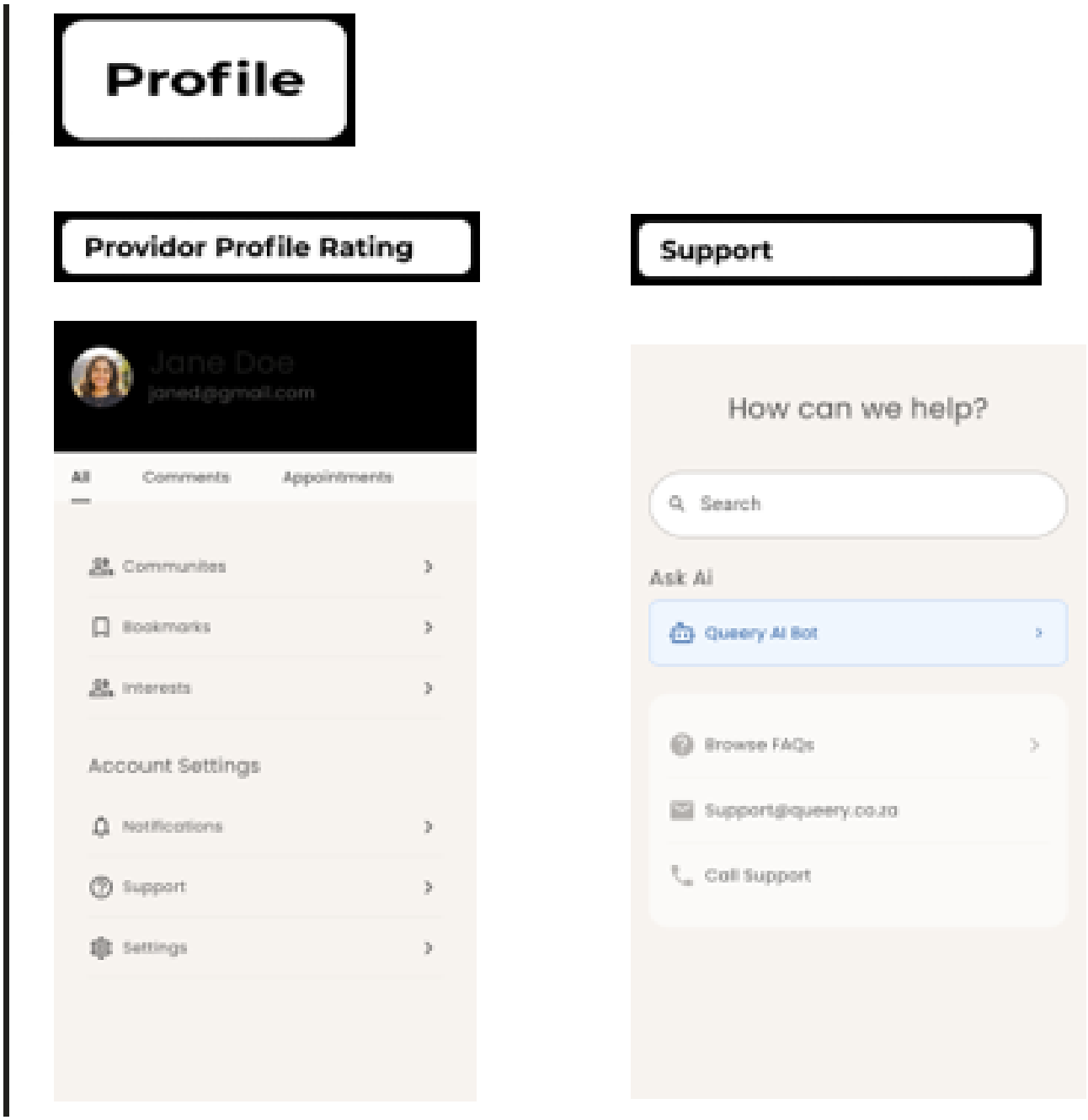

Virtual Consultations

Within the profile, recognizing the importance of personal interaction the video call feature facilitates virtual consultations and group support meetings. This innovation brings healthcare access closer, especially for those in remote or underserved areas. This icon is also available for peers and consultations which can be utilized during consultation and support groups meetings.

Finding LGBTQIA+ Friendly HCPS

Finally, the "Find Healthcare Provider" function helps users locate affirming SRH service providers. LGBTQIA+ individuals can search for HCPs who are rendering specific SRH by clicking on the find healthcare provider. This icon will provide LGBTQIA+ individuals with the location and distance of the HCPs; contact details and the services they provide. LGBTQIA+ individuals will be able to book their appointment online. It shows a directory with distance, services, and contact details, and even allows appointment bookings, making healthcare access efficient and queer-friendly.

Discussion

Principal Findings

The findings of this study showed that Design Science Research (DSR) framework enabled the development of an mHealth app called *Queery Wellness Hub* through its interrelated steps. These steps include awareness, suggestion, development, and

XSL•FO
RenderX

evaluation through three cycles. The study indicates that mHealth apps are ideally developed with a co-creation approach, involving users, subject experts, and software developers, to enhance relevance and usability. This finding is supported by the sequential mixed-methods approach employed in the study, including ethnography, PAR, and prototype development with software developers. Below will discuss the developed mHealth in comparison to the prior scholarly work.

Comparison With Prior Work

Our study showed that *Queery Wellness Hub* mHealth app includes have an onboarding and access icon which welcomes all the users with a simple and secure login interface. Furthermore, our study emphasized this login screen to serve as a digital gateway to a personalized LGBTQIA+ health journey [56]. study emphasizes that app trustworthiness essential for initial user uptake relies on clear onboarding processes, transparency in data handling, and strong privacy features. Similarly, participants in the study [57] expressed anxiety about their privacy when using mHealth apps. They also revealed their preferences for various security aspects in mHealth apps, including regular password updates, remote erase, user consent, and access controls [57]. All these studies including ours suggest that security is an important aspect of mHealth app and this may influence on the effective use of mHealth apps. Therefore, principal investigator s, innovators, and developers should ensure that all mHealth app consist of the secure login interface to promote privacy.

This study revealed that *Queery Wellness Hub* mHealth app can offer and serve as holistic SRH information hubs, where LGBTQIA+ individuals can access arranged and varied services such as preventive care, diagnostic services, maternal and sexual health education, and mental health support. Our study is in line different studies conducted around SRH that showed an importance of mHealth app in offering and enhancing SRH knowledge, attitudes, and service uptake [58]; provide sexuality, HIV/STI education plus navigation to services [59]. Overall, these indicated how important mHealth app could assist with sharing different health information and render services such as preventive care, diagnostic services, maternal and sexual health education, and mental health support thus increasing knowledge and changing attitude towards SRH update.

On top of that, our study illustrated that *Queery Wellness Hub* mHealth app can have community interaction and peer support icons where LGBTQIA+ users can join different community and discuss the different topics freely like family planning, contraception, STI/HIV prevention, and other relevant topics. Our findings are consistent with a study conducted [60], which found that peer-led group facilitation enabled the provision of sensitive psychological support, allowing young people to express themselves freely, acquire a feeling of self-worth, and engage more. This aligns with previous findings, our study corresponds with [61] study, which demonstrated that the smartphone app includes a secure, anonymous community messaging board, allowing users to express stigma, receive peer reassurance, and support retention in care. Furthermore, a study by [62] found a great need for community-sharing venues in which users may publish, discuss, and exchange experiences, tales, and Q&A, providing a safe, culturally resonant atmosphere. Communication is one of the digital elements that could be improved [63]. Connection to other individuals with similar experiences via an online forum, community, or instant messages in the mobile tool, or by connecting with other kinds of social media [63]. Altogether, these studies show that a comment section which allows LGBTQIA+ individuals to engage and share their experiences about various health topics in an anonymous way could improve their health and wellbeing.

Article and comment section icons was one of the findings our study highlighted as important in the *Queery Wellness Hub* mHealth app inclusion. This icon will enable HCPs and LGBTQIA+ individuals to access any articles that is peer-reviewed and has relevant information regarding SRH services and needs. As well as having a comment section for LGBTQIA+ individuals to be able to chat for among themselves and ask questions to the HCPs about all the matters relating to their SRH. Our findings match with a [63] study, which stressed the features to better mHealth app interventions, highlighting communication as an essential component for giving access to a health care practitioner via email, text message, or live chat. Furthermore [64], emphasizes the need of two-way communication by providing forums or chat rooms where users can connect with one another and with providers. They also highlight the significance of reputable, evidence-based information. Furthermore [65], showed that the community message board enabling users to share experiences, questions, and support.

Direct chat with HCPs and use of AI bot were also found by our study during the development of *Queery Wellness Hub* mHealth app. These icons were developed to allow LGBTQIA+ individuals to communicate with HCPs real-time support, allows follow-up queries, and encourages ongoing health conversations tailored to LGBTQIA+ identities. In addition, these icons found to serve as a reminder for LGBTQIA+ individuals' consultations with HCPs. Our study matches up with a study [66] that showed that direct messaging/chat with HCPs promotes patient engagement, provides for timely follow-up, and facilitates individualized health talks. Furthermore , presented apps with real-time chat and appointment reminders adapted to the requirements of sexual and gender minorities, demonstrating greater linkage to care and retention through reminders and direct provider connection [67].

Our study showed an importance of privacy by developing profile of each LGBTQIA+ individual' user including section where they can receive emergency support when utilizing the *Queery Wellness Hub* app. This provide the LGBTQIA+ users with shortcuts to book consultations or get immediate assistance bridging the gap between digital engagement and real-world support. Our findings correspond with [68], who used the Vula app and demonstrated that it enabled emergency clinicians to consult burn specialists via mobile phone, providing photos and patient information. Some research found that booking or seeking specialist consultations via an app can enhance results for patients [68,69]. Consistently, our findings is consistent with [70], which emphasizes that real-time booking and reminders improve patient-provider collaboration and access, particularly in resource-limited situations.

XSL•FO
RenderX

Additionally, our study demonstrated that once the LGBTQIA+ users gets inside their profile they can easily arrange and have virtual consultations with HCPs via video, thus bringing healthcare access closer, especially for those LGBTQIA+ users in remote or underserved areas. Our study displayed that this icon can also be used to by peers during consultation and support groups meetings. Our findings are consistent with [60] peer-led mobile support system (chat-based), which successfully boosted mental health and engagement, laying the framework for peer-enabled video sessions. Additionally [71], demonstrated that video-based teletherapy met both individual and peer-group health needs among LGBTQ+ young people.

Lastly, our study showed that the *Queery Wellness Hub* app could offer LGBTQIA+ users assistance in finding LGBTQIA+ friendly HCPs. This means that locating affirming SRH service HCPs through clicking find healthcare provider button. In addition, once they find the HCP, LGBTQIA+ users will be able to see location and distance of the HCPs; contact details and the services they provide, and able to book their appointment online. Our study is in accordance with a randomized controlled trial evaluated an app that tracked users via GPS and featured a clinic finder tool that mapped nearby clinics to assist pregnant and postpartum women living with HIV in accessing care [72,73]. Some studies showed that mHealth app helped with online appointment systems that allow users to search for HCPs by area, check availability, and schedule appointments immediately [72,73]. A recent evaluation of Cameroon-based apps included Healthlate, which allows appointment booking, in-app conversation with doctors, and has integrated location-based provider listings [74]. All statements regarding the app's features, such as secure login, community interaction, article access, AI bot, direct HCP communication, privacy protections, virtual consultations, and HCP finder functionalities, are grounded in the empirical data collected through cycles one and two (ethnographic interviews and NGT workshops) and operationalized in the prototype development of cycle three.

Limitations

Despite the benefits of RDS, the sample was drawn from a single NGO in Gauteng Province, South Africa which may not reflect the full diversity of LGBTQIA+ individuals experiences in other regions. Several sub-groups such as lesbians, bisexuals, intersex persons, men who have sex with men (MSM), and those identifying as pansexual or falling under the "plus" spectrum were underrepresented due to the hidden nature of the participants, difficult to reach in public health facilities, and some logistical challenges. Although interviews conducted in isiZulu and Sesotho were translated into English, some nuances may have been lost, potentially affecting the depth of meaning in the responses.

There is also a risk of social desirability bias, both participants both LGBTQIA+ individuals and HCPs may have offered responses they deemed acceptable rather than entirely truthful. While the expert workshop brought valuable input, the withdrawal of 12 invited participants due to work and family obligations reduced the diversity of expert opinions, although those who participated still represented broad expertise. Lastly, while mHealth apps present a promising solution for improving SRH access and support, their implementation and impact were not tested in this study and should be explored further in future research.

Recommendations

The developed mHealth app prototype should undergo rigorous testing and evaluation to assess usability, effectiveness, and user acceptance. Researchers can apply frameworks such as the Unified Theory of Acceptance and Use of Technology (UTAUT) to guide these evaluations. In additional, digital experts, engineers, LGBTQIA+ health specialists, and healthcare practitioners, and the National Department of Health should be engaged to refine and scale the app, ensuring it meets the diverse needs of users. Further, adequate funding from multiple sources is recommended to support the expansion, functionality, and sustainability of the mHealth app.

Future Directions

A shift from non-functional policies, strategies, and frameworks should be highlighted and practiced by different scholar. As a result, future research should be focused on studies that will enhance and introduce unique innovative solutions and technologies to improve the health of the LGBTQIA+ individuals, hence enhancing equality within the country and African continent.

Conclusions

Our study achieved its main purpose, that was, to develop an mHealth app to address SRH for LGBTQIA+ individuals in Gauteng Province, South Africa. These findings suggest that technology-based solutions can help mitigate some social, physical, psychological, and sexual health challenges. The findings from our study show an importance of how innovative solutions could be implemented for strengthening the healthcare sector and closing some of the disparities experienced by LGBTQIA+ individuals. Therefore, all details about the app's usefulness and relevance for LGBTQIA+ individuals can be reasonably drawn from the collected data and observed prototype outcomes.

Acknowledgments

We express our sincere gratitude to all experts, LGBTQIA+ individuals, and HCPs who participated in this study. Additional appreciation goes to the Aurum Institute, the National Health Research Database (NHRD), Tshwane Research Committee, and hospital Chief Executive Officers (CEOs) for granting permission to conduct the study. Authors used Grammarly [75] and QuillBot [76] solely for language editing; no AI tools were involved in content creation.

XSL•FO
RenderX

Funding

This study was supported by the National Research Foundation (NRF) under the Research Development Grants for New Generation of Academics Program. (nGAP) Scholars [Grant No: NGAP23021680129] and co-funded by the Department of Higher Education and Training (DHET) through (SSAUF) and nGAP. SMU will fund and pay for the publication.

Data Availability

Data supporting the study's findings are available from the corresponding author upon reasonable request, subject to SMUREC approval and participant confidentiality.

Authors' Contributions

RJS, MMM, and HS conceptualized the study and developed the protocol. RJS collected and analyzed the data; MMM confirmed themes as an independent coder. HS and MMM supervised and reviewed the manuscript. RJS prepared the initial draft, NVS reviewed and edited the second draft, and all authors reviewed and approved the final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Summary of cycle one demographic data and findings. This table summarizes the objectives, participant demographics, and key thematic findings from cycle one research data related to LGBTQIA+ sexual-reproductive healthcare. Combined overall summary of the cycle one details. All these findings were published in 2024, and references are given under each objective to avoid duplication.
[PDF File, 125 KB - formative_v9i1e79593_app1.pdf]

Multimedia Appendix 2
Summary of demographic data for both HCPs and LGBTQIA+ individuals in cycle 1. This table presents the demographic data collected from both groups of participants, including all the basic demographic information.
[PDF File, 132 KB - formative_v9i1e79593_app2.pdf]

Multimedia Appendix 3
Summary of cycle two demographic data and findings. Demographic characteristics of all consulted experts for mHealth App development. Experts identified the following pillars as crucial to the design and implementation of the mHealth app. Each pillar was ranked based on relevance, practicality, and user-centeredness. Similarly to cycle one, this cycle details are just a summary to orientate the reader of what led to cycle three development.
[PDF File, 106 KB - formative_v9i1e79593_app3.pdf]

Multimedia Appendix 4
Thematic Priorities for mHealth App Content. This includes all important contents that experts voted and ranked to be included in the mHealth prototype.
[PDF File, 83 KB - formative_v9i1e79593_app4.pdf]

References

1. Damar M, Kop O, Şaylan Ö, Erenay FS. The place of mobile health in the health sector, barriers and opportunities, integrated technologies and usage areas affecting the development of mobile health: a review of the literature in all aspects. Journal of Information Systems and Management Research 2024;6(2):37-59. [doi: 10.59940/jismar.1521397]
2. Sinha R. The role and impact of new technologies on healthcare systems. Discov Health Systems 2024;3(1):96. [doi: 10.1007/s44250-024-00163-w]
3. Kabanda S, Rother HA. Evaluating a South African mobile application for healthcare professionals to improve diagnosis and notification of pesticide poisonings. BMC Med Inform Decis Mak 2019 Mar 11;19(1):40. [doi: 10.1186/s12911-019-0791-2] [Medline: 30857525]
4. Alper J, Taylor RM. Using technology to advance global health: proceedings of a workshop. 2018.
5. Mirbabaie M, Stieglitz S, Frick NRJ. Artificial intelligence in disease diagnostics: a critical review and classification on the current state of research guiding future direction. Health Technol 2021 Jul;11(4):693-731. [doi: 10.1007/s12553-021-00555-5]
6. Awwalu J, Garba AG, Ghazvini A, Atuah R. Artificial intelligence in personalized medicine application of AI algorithms in solving personalized medicine problems. IJCTE 2015;7(6):439-443. [doi: 10.7763/IJCTE.2015.V7.999]

XSL•FO
RenderX

7. Mak KK, Pichika MR. Artificial intelligence in drug development: present status and future prospects. Drug Discov Today 2019 Mar;24(3):773-780. [doi: 10.1016/j.drudis.2018.11.014] [Medline: 30472429]
8. Su JJ, Chan MHS, Ghisi GLDM, et al. Real-world mobile health implementation and patient safety: multicenter qualitative study. J Med Internet Res 2025 Apr 29;27:e71086. [doi: 10.2196/71086] [Medline: 40299494]
9. Alkhuzaimi F, Rainey D, Brown Wilson C, Bloomfield J. The impact of mobile health interventions on service users' health outcomes and the role of health professions: a systematic review of systematic reviews. BMC Digit Health 2025;3(1):3. [doi: 10.1186/s44247-024-00143-3]
10. Rowland SP, Fitzgerald JE, Holme T, Powell J, McGregor A. What is the clinical value of mHealth for patients? NPJ Digit Med 2020;3(1):4. [doi: 10.1038/s41746-019-0206-x] [Medline: 31970289]
11. Songtaweesin WN, Kawichai S, Phanuphak N, et al. Youth-friendly services and a mobile phone application to promote adherence to pre-exposure prophylaxis among adolescent men who have sex with men and transgender women at-risk for HIV in Thailand: a randomized control trial. J Int AIDS Soc 2020 Sep;23 Suppl 5(Suppl 5):e25564. [doi: 10.1002/jia2.25564] [Medline: 32869511]
12. Sullivan PS, Stephenson R, Hirshfield S, et al. Behavioral efficacy of a sexual health mobile app for men who have sex with men: randomized controlled trial of mobile messaging for men. J Med Internet Res 2022 Feb 2;24(2):e34574. [doi: 10.2196/34574] [Medline: 35025755]
13. Sun CJ, Anderson KM, Kuhn T, Mayer L, Klein CH. A sexual health promotion app for transgender women (Trans Women Connected): development and usability study. JMIR Mhealth Uhealth 2020 May 12;8(5):e15888. [doi: 10.2196/15888] [Medline: 32396131]
14. Shrestha R, Altice FL, Khati A, et al. Clinic-integrated smartphone app (JomPrEP) to improve uptake of HIV testing and pre-exposure prophylaxis among men who have sex with men in Malaysia: mixed methods evaluation of usability and acceptability. JMIR Mhealth Uhealth 2023 Feb 16;11:e44468. [doi: 10.2196/44468] [Medline: 36795465]
15. Kawichai S, Songtaweesin WN, Wongharn P, et al. A mobile phone app to support adherence to daily HIV pre-exposure prophylaxis engagement among young men who have sex with men and transgender women aged 15 to 19 years in Thailand: pilot randomized controlled trial. JMIR Mhealth Uhealth 2022 Apr 21;10(4):e25561. [doi: 10.2196/25561] [Medline: 35451976]
16. Garg PR, Uppal L, Mehra S, Mehra D. mobile health app for self-learning on HIV prevention knowledge and services among a young Indonesian key population: cohort study. JMIR Mhealth Uhealth 2020 Sep 8;8(9):e17646. [doi: 10.2196/17646] [Medline: 32896831]
17. Abraham E, Chow EPF, Fairley CK, et al. eSexualHealth: preferences to use technology to promote sexual health among men who have sex with men and trans and gender diverse people. Front Public Health 2022;10:1064408. [doi: 10.3389/fpubh.2022.1064408] [Medline: 36711397]
18. Sexual and reproductive health and rights. : World Health Organization: Geneva, Switzerland; 2025 URL: https://www.who.int/health-topics/sexual-and-reproductive-health-and-rights [accessed 2025-12-09]
19. Mkhize SP, Maharaj P. Heteronormativity in health settings: realities of LGBT youth's access to healthcare in South Africa. In: Young People, Violence and Strategic Interventions in Sub-Saharan Africa: Springer; 2023:103-120. [doi: 10.1007/978-3-031-20679-5_6]
20. Mulemfo DM, Moyo I, Mavhandu-Mudzusi AH. LGBTQI+ experiences of HIV management services at primary healthcare facilities in Gauteng Province, South Africa. Adv Public Health 2023 Jul 12;2023(1):1-11. [doi: 10.1155/2023/1279409]
21. Scherf EDL. Addressing discrimination and healthcare disparities for sexual and gender minorities in South Africa: a human rights-based perspective. J Health Manag 2024:09720634231216025. [doi: 10.1177/09720634231216025]
22. Jessani A, Berry-Moreau T, Parmar R, Athanasakos A, Prodger JL, Mujugira A. Healthcare access and barriers to utilization among transgender and gender diverse people in Africa: a systematic review. BMC Global Public Health 2024;2(1):44. [doi: 10.1186/s44263-024-00073-2]
23. Müller A. Scrambling for access: availability, accessibility, acceptability and quality of healthcare for lesbian, gay, bisexual and transgender people in South Africa. BMC Int Health Hum Rights 2017 May 30;17(1):16. [doi: 10.1186/s12914-017-0124-4] [Medline: 28558693]
24. Seretlo RJ, Smuts H, Mokgatle MM. Challenges in sexual and reproductive healthcare access for queer people in Gauteng, South Africa. Afr j prim health care fam med 2024;16(1):4726. [doi: 10.4102/phcfm.v16i1.4726]
25. Ravele TT, Seretlo RJ, Mokgatle MM. We are treated differently: experiences of men who have sex with men in South African clinics. S Afr Fam Pract (2004) 2025 Feb 20;67(1):e1-e7. [doi: 10.4102/safp.v67i1.6050] [Medline: 40035099]
26. Zambezi D, Viljoen F. Access to healthcare services for transgender people in South Africa: assessing healthcare experiences and human rights. Int J Transgend Health 2024;25(4):791-803. [doi: 10.1080/26895269.2023.2273364] [Medline: 39465071]
27. Seretlo RJ, Smuts H, Mokgatle MM. Holistic sexual-reproductive healthcare services and needs for queer individuals: healthcare providers' perspectives. Healthcare (Basel) 2024 May 15;12(10):1026. [doi: 10.3390/healthcare12101026] [Medline: 38786436]
28. Seretlo RJ, Mokgatle MM. Practice, attitudes and views of right to access of sexual and reproductive health services by LGBTQI among primary health care nurses in Tshwane. African Journal of Primary Health Care & Family Medicine 2023;15(1):3790. [doi: 10.4102/phcfm.v15i1.3790]

XSL•FO
RenderX

29. Spencer S, Meer T, Müller A. “The care is the best you can give at the time”: Health care professionals’ experiences in providing gender affirming care in South Africa. PLoS One 2017;12(7):e0181132. [doi: 10.1371/journal.pone.0181132] [Medline: 28704458]
30. Mprah A. Sexual and reproductive health needs of LGBT. Afr J Reprod Health 2016 Mar;20(1):16-20. [doi: 10.29063/ajrh2016/v20i1.1] [Medline: 29553174]
31. Mogotsi I, Otchere YO, Botchway I, Muthoni Y, Gariseb R, Ramalepe LM. Psychosocial correlates of LGBTIQ+ experiences in selected African countries: reimagining LGBTIQ+ research. Journal of Social Issues 2024 Sep;80(3):1079-1111. [doi: 10.1111/josi.12640]
32. Mgopa LR, Rosser BRS, Ross MW, et al. Cultural and clinical challenges in sexual health care provision to men who have sex with men in Tanzania: a qualitative study of health professionals’ experiences and health students’ perspectives. BMC Public Health 2021 Apr 7;21(1):676. [doi: 10.1186/s12889-021-10696-x] [Medline: 33827508]
33. Igonya E, Muga W, Azubuike P, Stojanovski K. ‘We do not include tests and treatment for those [STI] conditions’: the trouble in navigating access and care for STIs among gay, bisexual men and transgender and other MSM in Kenya and Rwanda. Eur J Public Health 2024 Nov 1;34(Supplement_3):p. ckae144. 403. [doi: 10.1093/eurpub/ckae144.403]
34. Galdava M. Health care challenges for the LGBTQ community. Health Policy, Economics and Sociology 2023;7(2):1-7. [doi: 10.52340/healthecosoc.2023.07.02.15]
35. Moagi MM, van Der Wath AE, Jiyane PM, Rikhotso RS. Mental health challenges of lesbian, gay, bisexual and transgender people: An integrated literature review. Health SA 2021;26(1):1487. [doi: 10.4102/hsag.v26i0.1487] [Medline: 33604059]
36. Stepleman LM, Yohannan J, Scott SM, et al. Health needs and experiences of a LGBT population in Georgia and South Carolina. J Homosex 2019;66(7):989-1013. [doi: 10.1080/00918369.2018.1490573] [Medline: 30095375]
37. Joudeh L, Harris OO, Johnstone E, Heavner-Sullivan S, Propst SK. “Little Red Flags”: barriers to accessing health care as a sexual or gender minority individual in the rural Southern United States—a qualitative intersectional approach. J Assoc Nurses AIDS Care 2021;32(4):467-480. [doi: 10.1097/JNC.0000000000000271]
38. Nowaskie DZ, Menez O. Healthcare experiences of LGBTQ+ people: non-binary people remain unaffirmed. Front Sociol 2024;9:1448821. [doi: 10.3389/fsoc.2024.1448821]
39. Nxumalo CT, Luvuno Z, Chiya WH, et al. Designing implementation strategies for the inclusion of lesbian, gay, bisexual, transgender, intersex, queer, and allied and key populations’ content in undergraduate nursing curricula in KwaZulu-Natal, South Africa: protocol for a multimethods research project. JMIR Res Protoc 2024 May 31;13:e52250. [doi: 10.2196/52250] [Medline: 38598816]
40. Mustanski B, Fisher CB. HIV rates are increasing in gay/bisexual teens: IRB barriers to research must be resolved to bend the curve. Am J Prev Med 2016 Aug;51(2):249-252. [doi: 10.1016/j.amepre.2016.02.026] [Medline: 27061893]
41. Mashingaidze R, Moodie Z, Allen M, et al. Sexually transmitted infections amongst men who have sex with men (MSM) in South Africa. PLOS Glob Public Health 2023;3(4):e0001782. [doi: 10.1371/journal.pgph.0001782] [Medline: 37018240]
42. Malefo MA, Ayo-Yusuf O, Mokgatle MM. Risk factors for sexually transmitted infections among men who have sex with men. African Journal of Primary Health Care & Family Medicine 2023;15(1):4080. [doi: 10.4102/phcfm.v15i1.4080]
43. Malik MH, Iqbal S, Noman M, Sarfraz Z, Sarfraz A, Mustafa S. Mental health disparities among homosexual men and minorities: a systematic review. Am J Mens Health 2023;17(3):15579883231176646. [doi: 10.1177/15579883231176646] [Medline: 37219009]
44. Mendoza-Pérez JC, Ortiz-Hernández L. Violence as mediating variable in mental health disparities associated to sexual orientation among Mexican youths. J Homosex 2019;66(4):510-532. [doi: 10.1080/00918369.2017.1422938] [Medline: 29303429]
45. Kuechler B, Park EH, Vaishnavi V. Formalizing theory development in IS design science research: learning from qualitative research. 2009 Presented at: Proceedings of the 15th Americas Conference on Information Systems, AMCIS 2009; Aug 6-9, 2009; San Francisco, California, USA p. 486 URL: https://aisel.aisnet.org/amcis2009/486?utm_source=aisel.aisnet.org%2Famcis2009%2F486&utm_medium=PDF&utm_campaign=PDFCoverPages [accessed 2025-12-09]
46. Helms R, Giovacchini E, Teigland R, Kohler T. A design research approach to developing user innovation workshops in second life. JVWR 2010;3(1). [doi: 10.4101/jvwr.v3i1.819]
47. Vaishnavi V, Kuechler W. Design research in information systems/association for information systems. forschungsbericht. 2011 URL: https://designsciences.org/design-science-research-in-information-systems.pdf? [accessed 2025-12-09]
48. Seretlo RJ, Mokgatle MM, Smuts H. Breaking challenges: queer perspectives on solutions to establish inclusive sexual-reproductive healthcare in Gauteng Province, South Africa. Front Sociol 2024;9:1406265. [doi: 10.3389/fsoc.2024.1406265] [Medline: 39165861]
49. Seretlo RJ, Mokgatle MM, Smuts H. Positive views, attitudes, and acceptability toward mHealth applications in addressing queer sexual and reproductive health: healthcare providers and the queer individuals. Digit Health 2024;10:20552076241272704. [doi: 10.1177/20552076241272704] [Medline: 39221082]
50. Seretlo RJ, Smuts H, Mokgatle MM. Development of an mHealth app by experts for queer individuals’ sexual-reproductive health care services and needs: nominal group technique study. JMIR Form Res 2024 Aug 21;8(1):e59963. [doi: 10.2196/59963] [Medline: 39167434]

XSL•FO
RenderX

51. Mullen R, Kydd A, Fleming A, McMillan L. A practical guide to the systematic application of nominal group technique. Nurse Res 2021 Mar 11;29(1):14-20. [doi: 10.7748/nr.2021.e1777] [Medline: 33629547]
52. Fredriksen-Goldsen KI, Simoni JM, Kim HJ, et al. The health equity promotion model: reconceptualization of lesbian, gay, bisexual, and transgender (LGBT) health disparities. Am J Orthopsychiatry 2014 Nov;84(6):653-663. [doi: 10.1037/ort0000030] [Medline: 25545433]
53. Vom Brocke J, Hevner A, Maedche A. Introduction to design science research. In: Design Science Research. Cases: Springer; 2020:1-13. [doi: 10.1007/978-3-030-46781-4_1]
54. Sumarsono S, Sakkinah IS, Permanasari AE, Pranggono B. Development of a mobile health infrastructure for non-communicable diseases using design science research method: a case study. J Ambient Intell Human Comput 2023 Sep;14(9):12563-12574. [doi: 10.1007/s12652-022-04322-w]
55. Jaúregui-Velarde R, Molina-Velarde P, Yactayo-Arias C, Andrade-Arenas L. Evaluation of a prototype mobile application based on an expert system for the diagnosis of diseases transmitted by the Aedes Aegypti mosquito. Int J Onl Eng 2023;19(13):72-91. [doi: 10.3991/ijoe.v19i13.41431]
56. van Haasteren A, Gille F, Fadda M, Vayena E. Development of the mHealth app Trustworthiness checklist. Digit Health 2019;5:2055207619886463. [doi: 10.1177/2055207619886463] [Medline: 31803490]
57. Zhou L, Bao J, Watzlaf V, Parmanto B. Barriers to and facilitators of the use of mobile health apps from a security perspective: mixed-methods study. JMIR Mhealth Uhealth 2019 Apr 16;7(4):e11223. [doi: 10.2196/11223] [Medline: 30990458]
58. Isaacs N, Ntinga X, Keetsi T, et al. Are mHealth interventions effective in improving the uptake of sexual and reproductive health services among adolescents? A scoping review. Int J Environ Res Public Health 2024 Jan 31;21(2):165. [doi: 10.3390/ijerph21020165] [Medline: 38397656]
59. Abdul Hamid Alhassan RH, Haggerty CL, Fapohunda A, Affan NJ, Anto-Ocrah M. Exploring the use of digital educational tools for sexual and reproductive health in Sub-Saharan Africa: systematic review. JMIR Public Health Surveill 2025;11(1):e63309. [doi: 10.2196/63309]
60. Atujuna M, Simpson N, Ngobeni M, et al. Khuluma: using participatory, peer-led and digital methods to deliver psychosocial support to young people living with HIV in South Africa. Front Reprod Health 2021;3:687677. [doi: 10.3389/frph.2021.687677] [Medline: 36304024]
61. Flickinger TE, DeBolt C, Xie A, et al. Addressing stigma through a virtual community for people living with HIV: a mixed methods study of the PositiveLinks mobile health intervention. AIDS Behav 2018 Oct;22(10):3395-3406. [doi: 10.1007/s10461-018-2174-6] [Medline: 29882048]
62. Saberi P, Berrean B, Milionis C, Wong JO, Arnold E. We are family: designing and developing a mobile health application for the San Francisco bay area House Ball and Gay Family communities. Mhealth 2020;6:42. [doi: 10.21037/mHealth-19-234] [Medline: 33437838]
63. Wei Y, Zheng P, Deng H, Wang X, Li X, Fu H. Design features for improving mobile health intervention user engagement: systematic review and thematic analysis. J Med Internet Res 2020 Dec 9;22(12):e21687. [doi: 10.2196/21687] [Medline: 33295292]
64. Vo V, Auroy L, Sarradon-Eck A. Patients' perceptions of mHealth apps: meta-ethnographic review of qualitative studies. JMIR Mhealth Uhealth 2019 Jul 10;7(7):e13817. [doi: 10.2196/13817] [Medline: 31293246]
65. Kim GS, Kim L, Baek S, et al. Providing 2 types of mhealth interventions to support self-management among people living with HIV: randomized clinical trial. JMIR Mhealth Uhealth 2025 May 29;13:e60905. [doi: 10.2196/60905] [Medline: 40444823]
66. Xiong S, Ghebre R, Kulasingam S, Mason SM, Pratt RJ, Lazovich D. Exploring factors associated with preferences for human papillomavirus (HPV) self-sampling among racially- and ethnically-diverse women in Minnesota: a cross-sectional study. Prev Med Rep 2023 Aug;34:102243. [doi: 10.1016/j.pmedr.2023.102243] [Medline: 37234567]
67. Ungurean I, Gaitan NC. Software architecture of a fog computing node for industrial internet of things. Sensors (Basel) 2021 May 26;21(11):3715. [doi: 10.3390/s21113715] [Medline: 34073598]
68. Klingberg A, Wallis LA, Hasselberg M, Yen PY, Fritzell SC. Teleconsultation using mobile phones for diagnosis and acute care of burn injuries among emergency physicians: mixed-methods study. JMIR Mhealth Uhealth 2018 Oct 19;6(10):e11076. [doi: 10.2196/11076] [Medline: 30341047]
69. Alhammad N, Alajlani M, Abd-Alrazaq A, Epiphaniou G, Arvanitis T. Patients' perspectives on the data confidentiality, privacy, and security of mHealth apps: systematic review. J Med Internet Res 2024 May 31;26:e50715. [doi: 10.2196/50715] [Medline: 38820572]
70. Kyambille GG, Kalegele K. Enhancing patient appointments scheduling that uses mobile technology. arXiv. Preprint posted online on Feb 10, 2016. [doi: 10.48550/arXiv.1602.03337]
71. Goldbach C, Knutson D, Milton DC. LGBTQ+ people and COVID-19: the importance of resilience during a pandemic. Psychol Sex Orientat Gend Divers 2021;8(2):123-132. [doi: 10.1037/sgd0000463]
72. Clouse K, Noholoza S, Madwayi S, et al. The implementation of a GPS-based location-tracking smartphone app in South Africa to improve engagement in HIV Care: randomized controlled trial. JMIR Mhealth Uhealth 2023 May 19;11(1):e44945. [doi: 10.2196/44945] [Medline: 37204838]

XSL•FO
RenderX

73. Clouse K, Phillips TK, Mogoba P, Ndlovu L, Bassett J, Myer L. Attitudes toward a proposed GPS-based location tracking smartphone app for improving engagement in HIV care among pregnant and postpartum women in South Africa: focus group and interview study. JMIR Form Res 2021 Feb 8;5(2):e19243. [doi: 10.2196/19243] [Medline: 33555261]
74. Poulsen A, Hickie IB, Alam M, Wilson CE, LaMonica HM. Access to mobile health in lower middle-income countries: a review. Health Technol 2025 Mar;15(2):333-343. [doi: 10.1007/s12553-025-00948-w]
75. Grammarly. URL: https://app.grammarly.com/ [accessed 2025-12-09]
76. QuillBot. URL: https://quillbot.com/ [accessed 2025-12-12]

Abbreviations

AI: Artificial Intelligence
CEOs: Chief Executive Officers
DHET: Higher Education and Training
DPHRC: Department of Public Health Research Committee
DSR: Design Science Research
HbA1c: Glycosylated hemoglobin
HCPs: Healthcare providers
HEPM: Health Equity Promotion Model
HICs: High-income countries
HIV: Human immunodeficiency virus
LGBTQIA+: Lesbian, gay, bisexual, transgender, queer, intersex, asexual, and plus
LICs: Low-income countries
mHealth app: Mobile health application
Mobile Health: mHealth
MSM: Men who have sex with men
nGAP: New Generation of Academics Programme
NGO: Non-governmental organization
NGT: Nominal Group Technique
NHI: National Health Insurance
NHRD: National Health Research Database
NRF: National Research Foundation
PAR: Participatory action research
POPIA: Protection of Personal Information Act
PrEP: Pre-Exposure Prophylaxis
RDS: Respondent-driven sampling
SHRC: School of Healthcare Sciences Research Committee
SMUREC: Sefako Makgatho Health Sciences University Research Ethics Committee
SRH: Sexual and reproductive health
SSAUF: Staffing South Africa's Universities Framework
STI: Sexually transmitted infection
UTAUT: Unified Theory of Acceptance and Use of Technology

Edited by A Stone, A Mavragani; submitted 24.06.25; peer-reviewed by M Stevens-Uninsky, T Maake; revised version received 13.10.25; accepted 15.10.25; published 23.12.25.

Please cite as:
Seretlo RJ, Mokgatle MM, Smuts H, Sepeng NV
Development of an mHealth App Prototype for LGBTQIA+ Individuals' Sexual and Reproductive Health in Gauteng Province, South Africa: Design Science Research Study
JMIR Form Res 2025;9:e79593
URL: https://formative.jmir.org/2025/1/e79593
doi:10.2196/79593

© Raikane James Seretlo, Mathildah Mpata Mokgatle, Hanlie Smuts, Nombulelo Veronica Sepeng. Originally published in JMIR Formative Research (https://formative.jmir.org), 23.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The

XSL•FO
RenderX

complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Automated Feedback After Internet-Based Depression Screening: Cost-Effectiveness Analysis of a Randomized Controlled Trial

Léon Gerardo Kreis[1], MSc; Hans-Helmut König[1], MPH, Prof Dr; Franziska Sikorski[2], PhD; Bernd Löwe[2], Prof Dr; Sebastian Kohlmann[2,3*], Prof Dr; Christian Brettschneider[1*], PhD

[1]Department of Health Economics and Health Services Research, University Medical Center Hamburg-Eppendorf, Martinistr. 52, Hamburg, Germany
[2]Department of Psychosomatic Medicine and Psychotherapy, University Medical Center Hamburg-Eppendorf, Hamburg, Germany
[3]Department of General Internal Medicine and Psychosomatics, Heidelberg University Hospital, Heidelberg, Germany
[*]these authors contributed equally

Corresponding Author:
Léon Gerardo Kreis, MSc
Department of Health Economics and Health Services Research, University Medical Center Hamburg-Eppendorf, Martinistr. 52, Hamburg, Germany

Abstract

Background: The clinical and cost-related consequences of internet-based depression screening, in combination with automated feedback, have been rarely investigated. We aimed to conduct a cost-effectiveness analysis of DISCOVER, a 3-armed, observer-masked, randomized controlled trial that focused on 2 versions of automated feedback interventions after internet-based depression screening.

Objective: This study aimed to evaluate the cost-effectiveness of automated nontailored and tailored feedback interventions after internet-based depression screening from a societal perspective.

Methods: Participants were recruited from the general population via traditional and social media. Participants who were undiagnosed but screened positive for depression on an online version of the Patient Health Questionnaire-9 (≥10 points) were randomized to automatically receive no feedback, nontailored feedback, or tailored feedback. The feedback interventions included the depression screening result, a recommendation to seek professional advice, and brief general information about depression. The tailored feedback was additionally framed according to the participants' symptom profiles, treatment preferences, health insurance plans, and local residency. The time horizon was 6 months. The main outcome was the incremental cost-effectiveness ratio (ICER) from a societal perspective using quality-adjusted life years (QALY) based on the EuroQol-5D-5L. Cost-effectiveness acceptability curves were constructed. Furthermore, several sensitivity analyses and explorative subgroup analyses were conducted.

Results: A total of 1012 participants (no feedback: n=343, 33.9%; nontailored feedback: n=338, 33.4%; and tailored feedback: n=331, 32.7%) were included. Differences in costs and effects were not statistically significant. However, ICER results indicated that both no feedback and tailored feedback exhibited dominance over nontailored feedback. The ICER of tailored feedback compared to no feedback was €109,730 per QALY (a currency exchange rate of €1=US $1.02 was applicable as of December 31, 2022), whereas both costs and QALYs were lower in tailored feedback. The cost-effectiveness probability of tailored feedback compared to no feedback ranged between 41% and 80%. Sensitivity analyses exhibited similar trends.

Conclusions: Six months postintervention, feedback interventions had no statistically significant effect on costs from a societal perspective or on QALYs. Tailored feedback was associated with moderate cost-effectiveness probabilities compared to no feedback. Explorative subgroup analyses revealed subpopulations for which the interventions might be cost-effective.

Trial Registration: ClinicalTrials.gov NCT04633096; https://www.clinicaltrials.gov/study/NCT04633096

(JMIR Form Res 2025;9:e68282) doi:10.2196/68282

KEYWORDS

depressive disorder; screening; feedback intervention; online mental health intervention; cost-effectiveness; health economic evaluation

Introduction

Approximately 6% of adults experience depressive disorders worldwide [1]. As of 2019, depressive disorders were the 12th leading cause of disability-adjusted life years globally [2]. Depressive disorders are highly comorbid with other mental and somatic disorders, and studies suggest that preventing depression may causally prevent somatic diseases [3,4]. According to the World Health Organization, the COVID-19 pandemic has led to an increase of 27.6% in major depressive

XSL•FO
RenderX

disorder cases in 2020, with women and young adults particularly affected [5]. Depressed individuals typically experience states of low mood, reduced motivation, loss of pleasure or interest, and feelings of guilt [6]. In addition, depression places a significant economic burden on society and is an important cause of sickness absence [7].

In primary care, depressive cases often remain undetected: only approximately 50% of cases are correctly diagnosed as such [8]. Depression screening can expose possible depressive cases and rule out nondepressive cases as a low-intensity intervention [9]. However, its effectiveness is much debated [9-11].

Internet-based depression screening is already broadly used; however, its economic consequences have not yet been investigated [12,13]. The web-based approach allows for automated tailoring of screening results to individual preferences, which in turn could motivate screened individuals to engage in guideline-based care. Online feedback might save costs compared to an in-person screening approach with a health care professional followed by a psychological or medical consultation. Moreover, web-based feedback is easily accessible 24/7 and can be linked to further digital offers.

The DISCOVER trial evaluated the efficacy of 2 targeted feedback interventions (tailored and nontailored feedback) after online depression screening [13]. The DISCOVER trial's primary outcome was the change in depression severity. The interventions in the DISCOVER trial did not lead to a significant reduction in depressive symptoms. In addition, no statistically significant effects on secondary outcomes (eg, quality of life, help-seeking severity, anxiety, and somatic symptoms) were observed [13].

Although the interventions were not effective, conducting a health economic evaluation of such interventions can still provide relevant information beyond clinical effectiveness for several reasons. First, resources for health care are scarce. Hence, it is also important to evaluate the potential economic consequences of feedback after internet-based depression screening, as it might be a cost-saving strategy. Second, the effect measure of health economic evaluations goes beyond clinical effectiveness and uses a preference-based instrument to assess health-related quality of life, which takes multiple dimensions of health and the perspective of the general population into account. Third, health economic evaluations always consider two dimensions: costs and effects. Therefore, an ineffective intervention, which at the same time does not harm participants, could still be cost-effective when it leads to substantial cost reductions. The interventions in DISCOVER were designed to motivate individuals to engage in evidence-based depression care, which in turn might lead to increased costs. Therefore, it is important to analyze if costs in fact increased, suggesting the interventions led to a higher uptake of scarce health care resources without significant clinical effects, and thereby providing relevant information for decision makers and researchers. Finally, the cost-effectiveness of interventions that show no significant improvements in the primary outcome has been shown in previous literature [14,15].

Owing to the aforementioned reasons and as the DISCOVER trial collected data on both costs and effects during the 6-month follow-up period, a cost-effectiveness analysis (CEA) of targeted feedback interventions after internet-based depression screening is feasible and reasonable. To our knowledge, this is the first economic evaluation of feedback interventions following online depression screening.

Methods

Sample

The DISCOVER trial was an observer-masked, 3-armed randomized controlled trial (RCT) focusing on the effects of automated feedback interventions after internet-based depression screening on depression severity 6 months postintervention [16]. Participants were recruited from the general German population using social and traditional media campaigns, an online access panel, a newsletter, and internet search supported by a marketing company to ensure high representativeness of the German population. Further details on the recruitment and sample selection process are reported in the study protocol and the clinical effectiveness study [13,16]. Individuals were included based on a self-reported online assessment. Adults indicating at least *moderate* depressive symptoms (Patient Health Questionnaire-9 [PHQ-9] score ≥10) at baseline who had not received a depression diagnosis or treatment in the past year were included. In total, 5457 participants started screening, of which 4878 (89.39%) completed screening. Therefore, 1178 (24.15%) participants screened positive and had not received a depression diagnosis or treatment within the last year and were hence randomized into 1 of the 3 study arms, stratified by baseline depression severity [13]. Participants were included based on the intention-to-treat principle. However, 166 (14.09%) participants were excluded due to loss to follow-up, unavailability, or withdrawal of consent (Multimedia Appendix 1). This resulted in a final sample size of 1012 participants in the CEA.

Interventions

Individuals who screened positive were randomized immediately after completing the internet-based depression screening into 1 of the 3 groups using a 1:1:1 allocation ratio. After screening, all participants were shown a *thank you* note on their screen. Participants allocated to the no feedback group did not receive any feedback on their screening result. Participants in the nontailored feedback group were presented with their depression screening result alongside the advice to seek professional help, brief general information on depression, and information on guideline-based treatment options. In the tailored feedback intervention, the information provided to the participants was adapted according to their responses in the baseline assessment; for example, the screening result and general depression information were presented according to the participants' symptom profile, the advice to seek help was adapted according to individual specialist preferences, and treatment options were tailored according to participants' health insurance provider. Further details on the interventions' design can be found in the study protocol and the effectiveness study [13,16].

XSL•FO
RenderX

Data Collection and Measures

Data Collection

The DISCOVER study protocol described the data collection process in detail [16]. Briefly, data were gathered online at 4 time points (T0-T3) and in 2 clinical telephone interviews. T0 and T1 represented the baseline assessment: T0 was the prerandomization assessment completed online between December 2020 and January 2022, and T1 occurred approximately 2 days after T0 and was the baseline postrandomization survey. Health-related quality of life (HRQL) and productivity losses were assessed at T0, and health care use was assessed at T1. T2 represented the 1-month follow-up, and T3 represented the 6-month follow-up.

Depression severity was measured using the PHQ-9 [17], HRQL was measured by completing the EuroQol-5D-5L (EQ-5D-5L), including the EuroQol visual analog scale (EQ VAS) [18,19]. The Somatic Symptom Scale-8 was used to estimate somatic symptom severity throughout the trial. Health care use, including medication intake, and productivity losses were assessed by means of a modified version of the Client Sociodemographic and Service Receipt Inventory [20]. In the DISCOVER trial, gender was assessed via participant self-report based on the question "What is your gender" with the options woman, man, or nonbinary. No data on sex assigned at birth were collected.

EuroQol-5D-5L

The EQ-5D-5L is a questionnaire to estimate HRQL in five dimensions: mobility, self-care, usual activities, pain or discomfort, and anxiety or depression [18,19]. Respondents answer for each dimension in 5 levels, ranging from 1 (no problems) to 5 (extreme problems), resulting in 1 of 3125 different health states (5 combinations of 5 dimensions). The EQ-5D-5L features a visual analog scale (EQ VAS), ranging from 0 (worst imaginable health state) to 100 (best imaginable health state) as a measure of overall HRQL.

The EQ-5D-5L is a broadly applied HRQL measure, and its favorable psychometric properties have been demonstrated [21]. To assign preference-based values to responses on the EQ-5D-5L descriptive system, a value set representative of Germany was used [22]. EQ-5D-5L value scores range from −0.661 (worst imaginable health state) to 1 (best imaginable health state) [22].

Patient Health Questionnaire-9

The PHQ-9 is a measure of depression severity and recommended for screening, for example, by the US Preventive Services Task Force [10,17]. It consists of 9 questions regarding the frequency of depressive symptoms within the past 2 weeks, using a 4-point scale (0 - 3) depending on the frequency of symptoms. The overall PHQ-9 score is the sum of points per item (range 0 - 27), where depression severity increases with test score. Depression severity is indicated in five levels: (1) *minimal* (PHQ-9 score 0 - 4), (2) *mild* (PHQ-9 score 5-9), (3) *moderate* (PHQ-9 score 10-14), (4) *moderately severe* (PHQ-9 score 15-19), and (5) *severe* (PHQ-9 score 20-27) depression [17]. The PHQ-9 is considered to have both high specificity and sensitivity, as well as superior psychometric characteristics [23]. The PHQ-9's computerized version yields high validity and reliability compared to the paper-and-pen version [24].

Health Care Service Use

Use of health care services was estimated based on participants' statements on an adapted version of the Client Sociodemographic and Service Receipt Inventory throughout the trial [20]. Data on use of inpatient, outpatient physician, and nonphysician services (eg, physiotherapy and massage), the use of pharmaceuticals, and formal and informal nursing care were collected.

Unit Costs

Most costs for health care services were calculated based on refined German standard unit costs [25]. For medication, we matched participants' responses to pharmaceuticals, including their dosage, pack size, and pharmacy retail price. This was performed using the German pharmaceutical directory [26]. An overview of incorporated unit costs and their estimation methods can be found in Multimedia Appendix 2. We adjusted unit costs to prices of the reference year 2022 using the consumer price index [27]. A currency exchange rate of €1=US $1.02 was applicable as of December 31, 2022.

Productivity losses were estimated for both pre- and post-intervention periods based on the human capital approach. Indirect costs were assessed at T0, whereas direct costs from health care use were assessed at T1. As the interventions happened immediately after randomization at T0, we assigned the proportion of direct costs accruing between T0 and T1 to the posttreatment period rather than the pretreatment period. Total costs were the sum of direct and indirect costs. Intervention costs were not considered, as they are mostly fixed and hence decrease to a marginal level with higher use.

QALY Calculation

Quality-adjusted life years (QALY) were estimated based on the EQ-5D-5L index scores, assessed at T0 and T3 [22]. We assumed the change in HRQL between T0 and T3 was subject to a linear trend. A QALY is a 2D unit that takes into account HRQL and time. The term *1 QALY* represents 1 year at *perfect* HRQL. Therefore, QALYs during the follow-up period were calculated as described in Equation (1).

$$(1)\ QALY=[IndexT0+[IndexT3-IndexT02]*[180\Delta days(T0;T3)]]*(12)$$

The adaptation of QALY to 180 days was applicable, as no included participants died during the trial. Costs and QALY were not discounted due to the short follow-up period of 6 months.

Statistical Analysis

Missing Data

Missing data were a present phenomenon in the DISCOVER trial data. In total, 4.3% of all data points were missing. For variables with missing values, on average, 9% of the data were missing. Missing data were imputed using the multivariate imputation by chained equations [28]. Multiple imputation was performed to create 10 complete datasets with 927 participants each [28]. All statistical analyses were conducted using the

XSL•FO
RenderX

programming language *R* (R Foundation for Statistical Computing) in the *RStudio 4.2.0* environment.

Base Case Analysis

As a base case, a societal perspective was assumed, including all available direct and indirect costs. The primary outcome of the CEA was QALY 6 months after the interventions. To test for successful randomization at baseline, pairwise comparisons between study arms regarding baseline sociodemographic characteristics, costs, HRQL, and depression severity were conducted.

In the first step, we used incremental cost-effectiveness ratios (ICERs) to estimate cost-effectiveness. The ICER describes the additional costs per extra unit of (health) effects gained due to an intervention. Pairwise ICERs were calculated as follows:

$$(2)ICER=CIG--CCG-EIG-- ECG-=\Delta C-\Delta E-$$

where C represents costs, E represents effects, IG represents the intervention groups, and CG represents the control group.

We constructed cost-effectiveness acceptability curves (CEACs) to analyze the effect of willingness-to-pay (WTP) on cost-effectiveness probability, testing the robustness of ICER results. For this purpose, we performed monetary net benefit regressions (NBRs) considering different WTP margins [29]. We used a broad range of WTP margins between €0 and €160,000 per QALY with increments of €10,000. The theory behind the calculation of cost-effectiveness probabilities is described elsewhere [29].

NBRs were conducted as linear OLS regression models, where the net benefit was considered the dependent variable and the treatment dummy variable was considered the independent variable.

Sensitivity Analyses

To test the robustness of results, we conducted 5 deterministic sensitivity analyses.

In the first scenario (S1), we applied the EQ VAS score (divided by 100) to estimate QALYs.

In scenario 2 (S2), we analyzed cost-effectiveness for a dataset containing 4 formerly excluded participants (N_{ext}=1016). Unreliable data points by these respondents were considered missing and imputed using the multivariate imputation by chained equations algorithm.

For scenario 3 (S3), we took the perspective of a health care payer, thus including all direct costs except informal care.

In scenario 4 (S4), the analyses were conducted on the sample used in the clinical effectiveness study ($N_{primary}$=965 [13]). The sample used in S4 did not differ significantly from the base case sample used in this CEA (Multimedia Appendix 3).

In a fifth scenario (S5), we made use of PHQ-9 scores at T0, T2, and T3 to calculate depression-free days (DFDs) as an indicator of treatment effectiveness. In this scenario, we estimated the ICER in the unit (€ per DFD). The rationale behind this reliable and well-validated approach can be found elsewhere [30]. To calculate DFD in the pre- and post-intervention period, we applied linear interpolation to participants' PHQ-9 scores. Changes in PHQ-9 scores between T0, T2, and T3 were assumed to be subject to a linear trend. According to an analysis from 2003, patients were willing to pay approximately 9% of their household income for freedom from depression symptoms [31]. We applied this value to the average monthly gross income of the reference year 2022 in Germany, yielding a WTP of €15 per DFD [32]. To gain a broader image of cost-effectiveness probabilities, we considered a range between €0 per DFD and €200 per DFD as possible WTP margins.

Explorative Subgroup Analyses

Additionally, we conducted several explorative subgroup analyses to gain further insights into the cost-effectiveness of the feedback interventions for different subpopulations of the sample. We conducted these analyses considering the categories gender, age (in terciles), depression severity at baseline (applying the PHQ-9), number of inhabitants in participants' residency, if participants were diagnosed using the Structured Clinical Interview for DSM-5 Disorders at the postintervention baseline telephone interview, and if participants themselves believed at baseline that they had a depressive disorder.

Ethical Considerations

The DISCOVER trial was reviewed and approved by the Ethics Committee of the Medical Chamber Hamburg in July 2019 (reference number: PV7039) and conducted according to the Declaration of Helsinki. All participants provided online informed consent before the online baseline assessment. Informed consent forms included information on study aims, procedures, privacy, and voluntary participation. Data were pseudonymized and securely stored on encrypted servers, ensuring confidentiality. For this CEA, data were used in accordance with the original consent provided. Participants were compensated with a 5€ voucher for the completion of each follow-up survey (T1, T2, and T3).

Results

Characteristics of the Study Population

A summary of baseline characteristics can be found in Table 1. The mean age of the sample was 37.50 (SD 14.09; range 18 - 79) years. A total of 716 (71%) individuals were women, 287 (28%) were men, and 9 (1%) identified as nonbinary. The average depression severity was between *moderate* and *moderately severe* (mean PHQ-9 score 14.77, SD 4.00). At baseline, 56% (568/1012) of participants showed *moderate* depression severity, 30% (308/1012) indicated a *moderately severe*, and 13% (136/1012) showed a *severe* degree of depression. Baseline characteristics in the sample of this economic evaluation did not differ significantly compared to the sample of the effectiveness study (Multimedia Appendix 3) [13].

XSL•FO
RenderX

Table . Characteristics of the study population at baseline.

Characteristics	Overall (N=1012)	Nontailored feedback (n=338)	No feedback (n=343)	Tailored feedback (n=331)
Age (y), mean (SD)	37.50 (14.09)	38.25 (14.04)	36.98 (13.74)	37.27 (14.49)
Gender, n (%)				
Women	716 (70.8)	237 (70.1)	248 (72.3)	231 (69.8)
Men	287 (28.4)	98 (29.0)	92 (26.8)	97 (29.3)
Nonbinary	9 (0.9)	3 (0.9)	3 (0.9)	3 (0.9)
Living situation, n (%)				
With someone	677 (66.9)	232 (68.6)	228 (66.5)	217 (65.6)
Alone	335 (33.1)	106 (31.4)	115 (33.5)	114 (34.4)
Health insurance, n (%)				
Statutory	931 (92.0)	316 (93.5)	310 (90.4)	305 (92.1)
Private	81 (8.0)	22 (6.5)	33 (9.6)	26 (7.9)
Nationality, n (%)				
Non-German	33 (3.2)	14 (4.1)	10 (2.9)	9 (2.7)
German	979 (96.7)	324 (95.9)	333 (97.1)	322 (97.3)
Schooling degree[a], n (%)				
None	14 (1.4)	6 (1.8)	4 (1.2)	4 (1.2)
Special needs education	4 (0.4)	3 (0.9)	0 (0.0)	1 (0.3)
Lower secondary school leaving certificate	167 (16.5)	62 (18.3)	57 (16.6)	48 (14.5)
Intermediate secondary school leaving certificate	207 (20.5)	68 (20.1)	70 (20.4)	69 (20.9)
Higher education entrance qualification for university of applied sciences	112 (11.1)	37 (10.9)	33 (9.6)	42 (12.7)
General higher education entrance qualification	508 (50.2)	162 (47.9)	179 (52.2)	167 (50.5)
Baseline total costs[b,c] (€), mean (SD)	4528 (10,734)[c]	4560 (8675)	4821 (11,672)	4192 (11,607)
EQ-5D[d] index (−0.661 to 1), mean (SD)	0.69 (0.25)	0.70 (0.26)	0.69 (0.23)	0.68 (0.26)
EQ VAS[e] (0 - 100), mean (SD)	57.44 (21.93)	56.79 (21.89)	57.37 (22.36)	58.17 (21.57)
PHQ-9[f] at baseline, mean (SD)	14.77 (4.00)	14.84 (4.21)	14.76 (4.00)	14.70 (3.79)

[a]Schooling degrees were translated based on the German Qualifications Framework [33].

[b]Baseline costs were assessed for a 6-mo preintervention period.

[c]A currency exchange rate of €1=US $1.02 was applicable as of December 31, 2022.

[d]EQ-5D-5L: EuroQol-5D-5L.

[e]EQ VAS: EuroQol Visual Analogue Scale

[f]PHQ-9: 9-item Patient Health Questionnaire.

The mean EQ-5D index score was 0.69 (SD 0.25), and the mean EQ VAS score was 57.44 (SD 21.93)—both scores being remarkably reduced compared to the German general population [34]. The mean total costs in the 6 months before the intervention were approximately €4528 (SD €10,734). Differences in baseline characteristics between all study arms were statistically not significant (Multimedia Appendix 4).

XSL•FO
RenderX

Base Case Analysis

Costs During Follow-Up

Costs were highest for the nontailored feedback group and lowest for the tailored feedback group, with no statistically significant differences between study arms (Table 2). Overall, direct costs constituted approximately 64% of total costs, with inpatient costs making up the largest share (34%) of direct costs. All differences were statistically not significant.

Table . Mean costs, quality-adjusted life years (QALYs), and depression-free days (DFDs) during follow-up.

Category[a]	No feedback, mean (SD)	Nontailored feedback, mean (SD)	Tailored feedback, mean (SD)
Direct costs (€)	2860 (7555)	3316 (7727)	2549 (7444)
Inpatient services (€)	950 (4223)	1206 (5213)	795 (5337)
Outpatient physician services (€)	611 (1359)	684 (1388)	733 (3388)
Outpatient psychotherapist services (€)	321 (880)	389 (1030)	304 (798)
Outpatient nonphysician services (€)	169 (543)	163 (490)	143 (380)
Formal nursing care (€)	61 (923)	71 (713)	79 (643)
Informal care (€)	439 (2429)	587 (4256)	329 (2192)
Medication (€)	309 (3167)	216 (1171)	165 (850)
Indirect costs (€)	1776 (4312)	1693 (4131)	1490 (3543)
Total costs (€)	4636 (8953)	5008 (9241)	4039 (8408)
QALYs[b]	0.3636 (0.1030)	0.3620 (0.1111)	0.3575 (0.1110)
DFDs[c]	65.9 (51.3)	66.5 (50.5)	70.2 (49.0)

[a]A currency exchange rate of €1=US $1.02 was applicable as of December 31, 2022.

[b]Estimated based on the EuroQol-5D-5L index.

[c]Estimated based on 9-item Patient Health Questionnaire.

Quality of Life, QALY, and DFD During Follow-Up

In all study arms, the mean EQ-5D index scores as well as EQ VAS scores improved over the 6 months following the intervention, and the mean PHQ-9 scores decreased. Six months after the intervention, 9% (93/1012) of all participants reported *minimal* depression severity, and 33% (333/1012) had *mild*, 31% (317/1012) had *moderate*, 18% (179/1012) had *moderately severe*, and 9% (90/1012) had *severe* depressive symptoms.

Descriptive results (Table 2) indicated that the tailored feedback group experienced 0.3575 QALY during the posttreatment period, whereas nontailored feedback and no feedback experienced slightly more. The tailored feedback group experienced the most DFD, followed by nontailored feedback and no feedback. Differences in QALYs as well as DFDs were not statistically significant.

ICER and CEAC

As costs in the nontailored feedback group were higher (€372; P=.60) and QALYs were lower (−0.002; P=.85) compared to the no feedback group, the nontailored feedback group was dominated by the no feedback group. The tailored feedback group had both lower costs (€−970; P=.19) and lower effects (QALY −0.004; P=.60) than the nontailored feedback group, yielding an ICER of €216,873 per QALY. Compared to the no feedback group, the tailored feedback group had an ICER of €98,763 per QALY while also experiencing fewer QALY at lower costs (€−597; P=.40; and QALY −0.006; P=.48).

CEACs revealed cost-effectiveness probabilities between 70% (€0 per QALY) and 65% (€160,000 per QALY) for the no feedback group compared to the nontailored feedback group. For the tailored feedback group compared to the nontailored feedback group, CEACs revealed cost-effectiveness probabilities between 56% and 90%. Tailored feedback compared to no feedback revealed cost-effectiveness probabilities of 80% at a WTP margin of €0 per QALY and 41% at €160,000 per QALY (Figure 1).

XSL•FO
RenderX

Figure 1. Cost-effectiveness acceptability curves. A currency exchange rate of €1=US $1.02 was applicable as of December 31, 2022. EQ VAS: EuroQol Visual Analogue Scale; QALY: quality-adjusted life year.

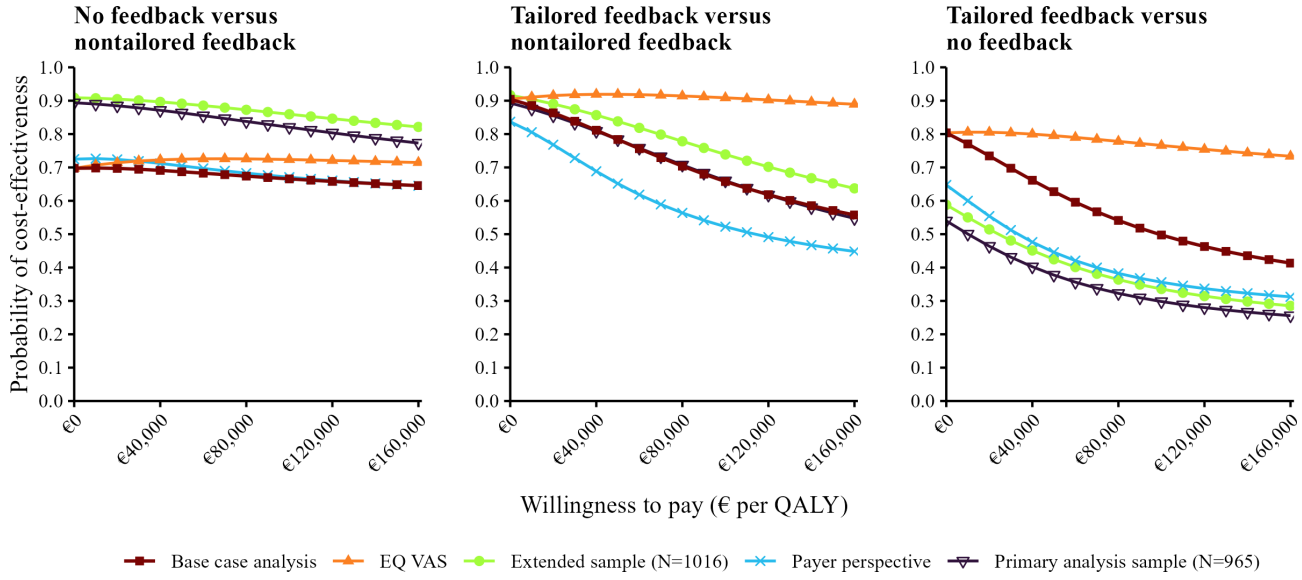

Sensitivity Analyses

The results from sensitivity analyses exhibited similar trends to the base case results (Figure 1). When considering QALY calculated from the EQ VAS (S1), both no feedback and nontailored feedback yielded higher cost-effectiveness probabilities compared to nontailored feedback. For the extended sample (S2), both no feedback and tailored feedback yielded higher cost-effectiveness probabilities when compared to the nontailored feedback arm. Cost-effectiveness probabilities in S2 were lower when comparing tailored feedback to no feedback. When taking a payer perspective (S3), cost-effectiveness probabilities were slightly lower when comparing tailored feedback to the other feedback options. When conducting the analysis for the sample used in the analysis of clinical effectiveness (S4), no feedback and tailored feedback yielded cost-effectiveness probabilities similar to those from the analysis of the extended dataset (S2).

Considering DFD as the effect measure (S5), participants in the tailored feedback arm experienced more DFD than both other study arms; however, differences were statistically not significant (Multimedia Appendix 4). Tailored feedback yielded high cost-effectiveness probabilities when compared to both nontailored feedback (90% - 94%) and no feedback (80% - 90%). No feedback compared to nontailored feedback exhibited cost-effectiveness probabilities decreasing from 70% (€0 per DFD) to 59% (€200 per DFD; Figure 2).

Figure 2. Cost-effectiveness acceptability curves considering depression-free days (DFD) as effect measure. A currency exchange rate of €1=US $1.02 was applicable as of December 31, 2022. WTP: willingness to pay.

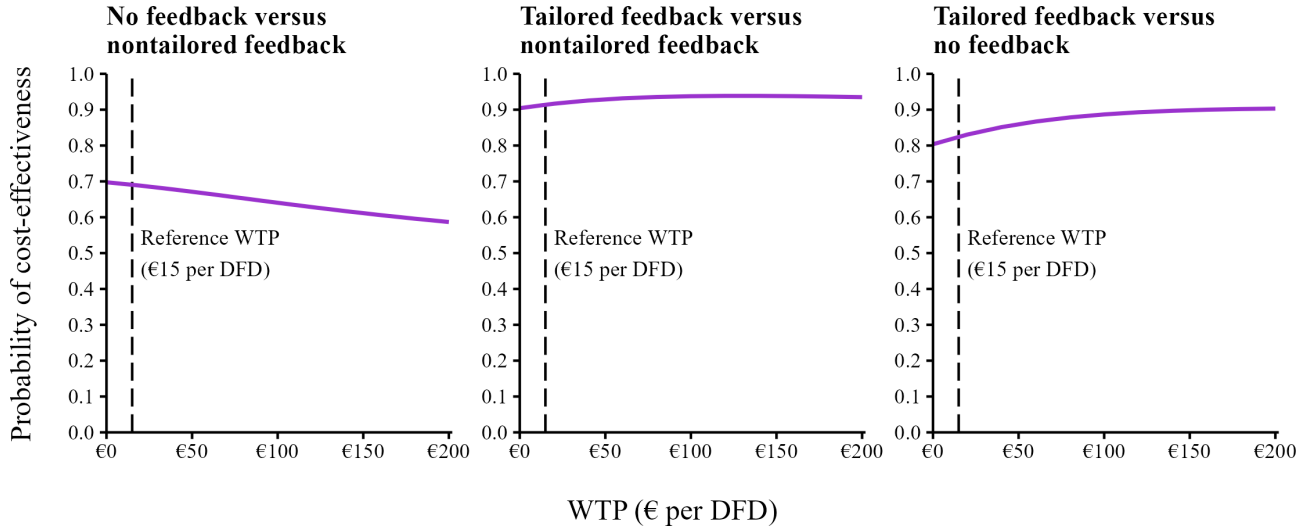

Explorative Subgroup Analyses

The results of the CEAC for the subgroup analyses are shown in Multimedia Appendix 5. These analyses revealed that the tailored feedback intervention is more cost-effective for women than for men. Cost-effectiveness probabilities on baseline depression severity and age were inconclusive. The tailored intervention yielded the highest cost-effectiveness probabilities for participants whose suspected depression was confirmed using the Structured Clinical Interview for DSM-5 Disorders in a telephone interview, approximately 2 to 5 days after the intervention. Cost-effectiveness probabilities ranged between

XSL•FO
RenderX

93% (€0 per QALY) and 47% (€160,000 per QALY) when comparing tailored feedback to no feedback and between 80% and 87% when comparing it to nontailored feedback.

Tailored feedback had the lowest cost-effectiveness probabilities compared to no feedback for participants living in big cities (>100,000 inhabitants) and for participants not believing at baseline that they had depression.

Discussion

Principal Findings

This health economic evaluation has shown that nontailored feedback after web-based depression screening is associated with low cost-effectiveness probabilities, both when compared to no feedback and to tailored feedback.

In the following section, we will thus mainly focus on the comparison between tailored feedback and no feedback. Costs during follow-up did not significantly increase due to the automated nature of feedback interventions. HRQL in terms of QALY did not differ significantly between study arms, which is in line with differences in depression severity also being not statistically significant [13]. As all differences between study arms were statistically not significant and the ICER is a point estimate, we additionally constructed CEACs based on monetary NBR. In this approach, effect differences affect results more strongly than cost differences with increasing WTP, thus leading to a negatively sloped CEAC in the case of both negative costs and effect differences. Our results strongly depend on the assumed WTP margin, where tailored feedback has the highest cost-effectiveness probabilities at WTP thresholds below €20,000 per QALY.

DEPSCREEN-INFO is the only other RCT to date on feedback interventions in depression screening, for which a CEA was conducted [35,36]. In the trial, patients with heart disease were recruited, representing an older and predominantly male sample. The feedback intervention, which consisted of delivering the screening result to the patient in addition to their treating physician, yielded high cost-effectiveness probabilities and revealed significant improvements in depression severity 6 months after the intervention [35,36]. The high cost-effectiveness probability shown for the at-risk sample of DEPSCREEN-INFO could not be supported by the DISCOVER results [35]. Both the screening strategy (at risk vs general) and the delivery mode (in person vs internet-based) differed between DEPSCREEN-INFO and DISCOVER, which might explain the low cost-effectiveness probabilities in this CEA. A deeper analysis of the extent to which the delivery mode and the presence of depression risk factors potentially influence cost-effectiveness remains to be investigated.

This highlights the relevance of the recently conducted GET.FEEDBACK.GP RCT, which featured patient-targeted feedback and general practitioner (GP)-targeted feedback after depression screening in a primary care setting [37]. The clinical effectiveness study of GET.FEEDBACK.GP revealed no significant reduction in depression severity due to feedback interventions after depression screening for the general sample [37]. However, some subgroups exhibited potential beneficial effects postintervention [37]. The health economic analysis of GET.FEEDBACK.GP will provide further insights on the economic implications of delivering in-person feedback to the patient and the GP.

Of the 5 conducted scenario analyses, 3 showed similarly low cost-effectiveness probabilities of nontailored feedback compared to the other 2 study arms.

The scenarios considering a different effect measure, however, exhibited higher and more consistent cost-effectiveness probabilities of tailored feedback. The EQ VAS scenario yielded moderately high cost-effectiveness probabilities (73% - 80%) of tailored compared to screening only. When considering the disease-specific outcome DFD based on PHQ-9 scores, cost-effectiveness probabilities of tailored feedback were high (80% - 90%) compared to no feedback. Thus, automatically tailored feedback is likely to be cost-effective when considering the depression-specific measure DFD. This might be explained by the different types of patient-reported outcome measures considered in each analysis [38]. The EQ-5D-5L assesses overall functional status, and HRQL is estimated by applying preference-based value sets to the responses. The PHQ-9, on the other hand, assesses the depression-specific symptom status, a less complex construct than HRQL, by applying psychometric methods.

Explorative subgroup analyses revealed populations that might require other solutions than feedback after internet-based depression screening. Low cost-effectiveness probabilities of tailored feedback compared to screening only were most prominent for men, people living in big cities, people whose suspected depression was not confirmed using a diagnostic interview, and those who did not believe at baseline that they had a depressive disorder. On the other hand, moderately high to high cost-effectiveness probabilities of tailored feedback were observed for women, people living in medium-sized cities (20,000 - 99,999 inhabitants), and people being diagnosed with depression shortly after screening.

As the study arms were not randomized with regard to these categories, the results need to be handled with care; however, this might provide an interesting starting point for further research. Furthermore, the question of whether participants believed that they had depression could be included in the randomization strategy, to gain more robust insights in future trials on depression screening combined with feedback interventions.

Strengths and Limitations

In the DISCOVER trial, direct and indirect costs were assessed during follow-up and at baseline, allowing for differences between study arms to be tested, not only during follow-up. We noticed no statistically significant differences in all categories of interests, including costs, between study arms at baseline. Randomization can hence be considered successful, as the study arms were balanced and the sample was highly representative of a population interested in depression [13]. Additionally, this work is the first to estimate the health economic implications of feedback interventions after web-based depression screening.

XSL•FO
RenderX

However, this study has some limitations. The economic evaluation has the same methodological limitations as the analysis of the primary outcome, which included the absence of a no screening study group, the reliance on self-reports, and the self-selection into the trial [13].

We identified additional limitations in this CEA. First, data quality in the study may have been affected by the lengthiness of the questionnaires regarding the use of health care service use and the relatively short follow-up time horizon of 6 months. We expect this limitation to have influenced participants in all study arms similarly. Second, recall bias could have caused the underestimation of medical service use during the follow-up. We consider the risk of recall bias low, as the mean direct costs during follow-up were higher than in the comparable age group in the German population [39]. Third, QALYs were calculated based on the generic EQ-5D-5L in the base case analysis, which, on the one hand, is considered acceptably responsive for changes in anxiety and depression and, on the other hand, only considers mental health specifically in 1 of 5 dimensions and therefore tends to be less sensitive to changes in depression [40]. Fourth, the tailored feedback intervention might have increased participants' awareness of their own health status and confronted them with their possible diagnosis. Minor symptoms, thus, could have been perceived more strongly, leading to a decrease in self-reported HRQL [41]. These limitations were handled by conducting S2, focusing on the depression-specific outcome of DFD calculated based on the PHQ-9.

Conclusions

This CEA indicates that nontailored feedback after internet-based depression is less cost-effective compared to tailored feedback and depression screening without the provision of screening results. The cost-effectiveness of the tailored feedback intervention compared to no feedback highly depends on the assumed willingness-to-pay. Considering DFDs as an alternative disease-specific measure, the tailored feedback intervention demonstrates consistently high cost-effectiveness probabilities. In this case, it could be considered economically reasonable compared to both other study arms. Nevertheless, it is important to note that feedback interventions did not significantly impact costs, HRQL, or depression severity throughout the follow-up period. Further research is needed to identify optimal target populations and the most effective methods to present screening results.

Acknowledgments

The authors would like to thank all participants who gave their consent to participate in DISCOVER and supported the study with data; Karl Wegscheider for his critical and constructive comments and his contribution to the quality of the manuscript; Anne Daubmann, who supported the biometric analyses and helped interpret the data; and Elisa Rabbiolo and Tarık Korkutan, who contributed to the organization and recruitment of the trial. The authors acknowledge financial support from the Open Access Publication Fund of UKE - Universitätsklinikum Hamburg-Eppendorf.

Funding

The DISCOVER trial was funded by the German Research Foundation (grant 424162019). The funder had no role in the study design, data collection, data analysis, data interpretation, or writing of the report.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

CB, SK, and H-HK report research funding (no personal honoraria) from the German Research Foundation. BL reports research funding (no personal honoraria) from the German Research Foundation; the German Federal Ministry of Education and Research; the German Innovation Committee at the Joint Federal Committee; the European Commission's Horizon 2020 Framework Programme; the European Joint Programme for Rare Diseases; the Ministry of Science, Research and Equality of the Free and Hanseatic City of Hamburg, Germany; and the Foundation Psychosomatics of Spinal Diseases, Stuttgart, Germany. He received remunerations for several scientific book articles from various book publishers, from the Norddeutscher Rundfunk for interviews in medical knowledge programmes on public television, and as a committee member from Aarhus University, Denmark. He received travel expenses from the European Association of Psychosomatic Medicine (EAPM) and accommodation and meals from the Societatea de Medicina Biopsihosociala, Romania, for a presentation at the EAPM Academy at the Conferin a Na ională de Psihosomatică, Cluj-Napoca, Romania, October 2023. He received a travel grant for a lecture on the occasion of the presentation of the Alison Creed Award at the EAPM Conference in Lausanne, 12 to 15 June 2024. He received remuneration and travel expenses for a lecture at the Lindauer Psychotherapiewochen, April 2024. He is President of the German College of Psychosomatic Medicine (unpaid) since March 2024 and was a member of the Board of the European Association of Psychosomatic Medicine (unpaid) until 2022. He is a member of the EIFFEL Study Oversight Committee (unpaid). LK and FS declare no competing interests.

Multimedia Appendix 1

XSL•FO
RenderX

Participants flow diagram.
[DOCX File, 59 KB - formative_v9i1e68282_app1.docx]

Multimedia Appendix 2
Overview of incorporated costs per sector.
[DOCX File, 26 KB - formative_v9i1e68282_app2.docx]

Multimedia Appendix 3
Comparison of different sample sizes.
[DOCX File, 30 KB - formative_v9i1e68282_app3.docx]

Multimedia Appendix 4
Baseline group comparison of study population.
[DOCX File, 31 KB - formative_v9i1e68282_app4.docx]

Multimedia Appendix 5
Cost-effectiveness acceptability curves for subgroup analyses.
[DOCX File, 272 KB - formative_v9i1e68282_app5.docx]

References

1. Global burden of disease study 2021 (GBD 2021) results. Institute of Health Metrics and Evaluation (IHME). 2022. URL: https://vizhub.healthdata.org/gbd-results [accessed 2025-07-09]
2. Global health estimates 2019: disease burden by cause, age, sex, by country and by region, 2000-2019. : World Health Organization; 2020 URL: https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates [accessed 2025-12-10]
3. Steffen A, Nübel J, Jacobi F, Bätzing J, Holstiege J. Mental and somatic comorbidity of depression: a comprehensive cross-sectional analysis of 202 diagnosis groups using German nationwide ambulatory claims data. BMC Psychiatry 2020 Mar 30;20(1):142. [doi: 10.1186/s12888-020-02546-8] [Medline: 32228541]
4. Li GHY, Cheung CL, Chung AKK, et al. Evaluation of bi-directional causal association between depression and cardiovascular diseases: a Mendelian randomization study. Psychol Med 2022 Jul;52(9):1765-1776. [doi: 10.1017/S0033291720003566] [Medline: 33032663]
5. Mental health and COVID-19: early evidence of the pandemic's impact: scientific brief, 2 march 2022. : World Health Organization; 2022 URL: https://www.who.int/publications/i/item/WHO-2019-nCoV-Sci_Brief-Mental_health-2022.1 [accessed 2025-12-10]
6. Bundesinstitut für Arzneimittel und Medizinprodukte (BfArM) im Auftrag des Bundesministeriums für Gesundheit (BMG) unter Beteiligung der Arbeitsgruppe ICD des Kuratoriums für Fragen der Klassifikation im Gesundheitswesen (KKG). ICD-10-GM version 2022, systematisches verzeichnis, internationale statistische klassifikation der krankheiten und verwandter gesundheitsprobleme, 10. revision, german modification [ICD-10-GM version 2022, systematic index, international statistical classification of diseases and related health problems, 10th revision, german modification] [report in german]. : Bundesinstitut für Arzneimittel und Medizinprodukte; 2021 URL: https://klassifikationen.bfarm.de/icd-10-gm/kode-suche/htmlgm2022/index.htm [accessed 2025-12-10]
7. Hildebrandt S, Dehl T, Zich K, Nolting HD. Gesundheitsreport 2023 [Health report 2023]. In: DAK-Gesundheit 2023.
8. Mitchell AJ, Vaze A, Rao S. Clinical diagnosis of depression in primary care: a meta-analysis. The Lancet 2009 Aug;374(9690):609-619. [doi: 10.1016/S0140-6736(09)60879-5]
9. Mitchell AJ, Yadegarfar M, Gill J, Stubbs B. Case finding and screening clinical utility of the Patient Health Questionnaire (PHQ-9 and PHQ-2) for depression in primary care: a diagnostic meta-analysis of 40 studies. BJPsych Open 2016 Mar;2(2):127-138. [doi: 10.1192/bjpo.bp.115.001685] [Medline: 27703765]
10. US Preventive Services Task Force, Barry MJ, Nicholson WK, et al. Screening for depression and suicide risk in adults: US Preventive Services Task Force recommendation statement. JAMA 2023 Jun 20;329(23):2057-2067. [doi: 10.1001/jama.2023.9297] [Medline: 37338872]
11. Thombs BD, Markham S, Rice DB, Ziegelstein RC. Does depression screening in primary care improve mental health outcomes? BMJ 2021 Jul 19;374:n1661. [doi: 10.1136/bmj.n1661] [Medline: 34281908]
12. Yom-Tov E, Lekkas D, Heinz MV, Nguyen T, Barr PJ, Jacobson NC. Digitally filling the access gap in mental health care: an investigation of the association between rurality and online engagement with validated self-report screens across the United States. J Psychiatr Res 2023 Jan;157:112-118. [doi: 10.1016/j.jpsychires.2022.11.024] [Medline: 36462251]
13. Kohlmann S, Sikorski F, König HH, Schütt M, Zapf A, Löwe B. The efficacy of automated feedback after internet-based depression screening (DISCOVER): an observer-masked, three-armed, randomised controlled trial in Germany. Lancet Digit Health 2024 Jul;6(7):e446-e457. [doi: 10.1016/S2589-7500(24)00070-0] [Medline: 38906611]

XSL•FO
RenderX

14. Zülke AE, Pabst A, Luppa M, et al. A multidomain intervention against cognitive decline in an at-risk-population in Germany: results from the cluster-randomized AgeWell.de trial. Alzheimers Dement 2024 Jan;20(1):615-628. [doi: 10.1002/alz.13486] [Medline: 37768074]
15. Brettschneider C, Buczak‐Stec E, Luppa M, et al. Cost‐effectiveness of a multicomponent intervention against cognitive decline. Alzheimer's Dementia Transl Res Clin Interv 2025 Jan;11(1):e70028. [doi: 10.1002/trc2.70028]
16. Sikorski F, König HH, Wegscheider K, Zapf A, Löwe B, Kohlmann S. The efficacy of automated feedback after internet-based depression screening: study protocol of the German, three-armed, randomised controlled trial DISCOVER. Internet Interv 2021 Sep;25:100435. [doi: 10.1016/j.invent.2021.100435] [Medline: 34401394]
17. Kroenke K, Spitzer RL, Williams JBW. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613. [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
18. EuroQol Group. EuroQol - a new facility for the measurement of health-related quality of life. Health Policy 1990 Dec;16(3):199-208. [doi: 10.1016/0168-8510(90)90421-9]
19. Herdman M, Gudex C, Lloyd A, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Qual Life Res 2011 Dec;20(10):1727-1736. [doi: 10.1007/s11136-011-9903-x] [Medline: 21479777]
20. Roick C, Kilian R, Matschinger H, Bernert S, Mory C, Angermeyer MC. German adaptation of the client sociodemographic and service receipt inventory - an instrument for the cost of mental health care. Psychiatr Prax 2001 Oct;28:84-90. [doi: 10.1055/s-2001-17790] [Medline: 11605129]
21. Feng YS, Kohlmann T, Janssen MF, Buchholz I. Psychometric properties of the EQ-5D-5L: a systematic review of the literature. Qual Life Res 2021 Mar;30(3):647-673. [doi: 10.1007/s11136-020-02688-y] [Medline: 33284428]
22. Ludwig K, Graf von der Schulenburg JM, Greiner W. German value set for the EQ-5D-5L. Pharmacoeconomics 2018 Jun;36(6):663-674. [doi: 10.1007/s40273-018-0615-8] [Medline: 29460066]
23. Miller P, Newby D, Walkom E, Schneider J, Li SC, Evans TJ. The performance and accuracy of depression screening tools capable of self-administration in primary care: a systematic review and meta-analysis. Eur J Psychiatry 2021 Jan;35(1):1-18. [doi: 10.1016/j.ejpsy.2020.10.002]
24. Erbe D, Eichert HC, Rietz C, Ebert D. Interformat reliability of the patient health questionnaire: validation of the computerized version of the PHQ-9. Internet Interv 2016 Sep;5:1-4. [doi: 10.1016/j.invent.2016.06.006] [Medline: 30135800]
25. Muntendorf LK, Brettschneider C, Konnopka A, König HH. Updating standardized unit costs from a societal perspective for health economic evaluation. Gesundheitswesen 2024 May;86(5):389-393. [doi: 10.1055/a-2169-1480] [Medline: 38316405]
26. Rote Liste 2022, Q1. : Rote Liste Service GmbH; 2022.
27. Inflation (CPI) (indicator). OECD. 2023. URL: https://data.oecd.org/price/inflation-cpi.htm [accessed 2023-03-07]
28. van Buuren S. Flexible Imputation of Missing Data: CRC press; 2012. URL: https://stefvanbuuren.name/fimd/ [accessed 2025-12-10]
29. Hoch JS, Briggs AH, Willan AR. Something old, something new, something borrowed, something blue: a framework for the marriage of health econometrics and cost-effectiveness analysis. Health Econ 2002 Jul;11(5):415-430. [doi: 10.1002/hec.678] [Medline: 12112491]
30. Lave JR, Frank RG, Schulberg HC, Kamlet MS. Cost-effectiveness of treatments for major depression in primary care practice. Arch Gen Psychiatry 1998 Jul;55(7):645-651. [doi: 10.1001/archpsyc.55.7.645] [Medline: 9672056]
31. Unützer J, Katon WJ, Russo J, et al. Willingness to pay for depression treatment in primary care. Psychiatr Serv 2003 Mar;54(3):340-345. [doi: 10.1176/ps.54.3.340] [Medline: 12610241]
32. Durchschnittliche Bruttomonatsverdienste, Zeitreihe [Average gross monthly earnings, time series]. Statistisches Bundesamt (Destatis). 2023. URL: https://www.destatis.de/DE/Themen/Arbeit/Verdienste/Verdienste-Branche-Berufe/Tabellen/liste-bruttomonatsverdienste.html?nn=206824 [accessed 2023-03-07]
33. Federal ministry for education, family affairs, senior citizens, women and youth. The German Qualifications Framework for Lifelong Learning. URL: www.dqr.de/dqr/en/ [accessed 2025-11-11]
34. Grochtdreis T, Dams J, König HH, Konnopka A. Health-related quality of life measured with the EQ-5D-5L: estimation of normative index values based on a representative German population sample and value set. Eur J Health Econ 2019 Aug;20(6):933-944. [doi: 10.1007/s10198-019-01054-1] [Medline: 31030292]
35. Brettschneider C, Kohlmann S, Gierk B, Löwe B, König HH. Depression screening with patient-targeted feedback in cardiology: the cost-effectiveness of DEPSCREEN-INFO. PLoS ONE 2017;12(8):e0181021. [doi: 10.1371/journal.pone.0181021] [Medline: 28806775]
36. Löwe B, Blankenberg S, Wegscheider K, et al. Depression screening with patient-targeted feedback in cardiology: DEPSCREEN-INFO randomised clinical trial. Br J Psychiatry 2017 Feb;210(2):132-139. [doi: 10.1192/bjp.bp.116.184168] [Medline: 27908896]
37. Löwe B, Scherer M, Braunschneider LE, et al. Clinical effectiveness of patient-targeted feedback following depression screening in general practice (GET.FEEDBACK.GP): an investigator-initiated, prospective, multicentre, three-arm, observer-blinded, randomised controlled trial in Germany. Lancet Psychiatry 2024 Apr;11(4):262-273. [doi: 10.1016/S2215-0366(24)00035-X] [Medline: 38432236]

XSL•FO
RenderX

38. Valderas JM, Alonso J. Patient reported outcome measures: a model-based classification system for research and clinical practice. Qual Life Res 2008 Nov;17(9):1125-1135. [doi: 10.1007/s11136-008-9396-4] [Medline: 18836850]
39. Krankheitskosten je Einwohner: Deutschland, Jahre, Geschlecht, Altersgruppen (23631-0002) [healthcare costs per inhabitant: germany, years, gender, age groups]. : Statistisches Bundesamt (Destatis); 2022 URL: https://www-genesis.destatis.de/datenbank/online/url/ba9ab6b7 [accessed 2023-04-14]
40. Brazier J, Connell J, Papaioannou D, et al. A systematic review, psychometric analysis and qualitative assessment of generic preference-based measures of health in mental health populations and the estimation of mapping functions from widely used specific measures. Health Technol Assess 2014 May;18(34):vii-viii. [doi: 10.3310/hta18340] [Medline: 24857402]
41. Carr AJ, Gibson B, Robinson PG. Measuring quality of life: is quality of life determined by expectations or experience? BMJ 2001 May 19;322(7296):1240-1243. [doi: 10.1136/bmj.322.7296.1240] [Medline: 11358783]

Abbreviations

CEA: cost-effectiveness analysis
CEAC: cost-effectiveness acceptability curve
DFD: depression-free day
EQ VAS: EuroQol Visual Analogue Scale
EQ-5D-5L: EuroQol-5D-5L
GP: general practitioner
HRQL: health-related quality of life
ICER: incremental cost-effectiveness ratio
NBR: net-benefit regression
PHQ-9: 9-item Patient Health Questionnaire
QALY: quality-adjusted life year
RCT: randomized controlled trial
WTP: willingness-to-pay

Edited by A Stone; submitted 04.11.24; peer-reviewed by P McCrone, T Tillmann; revised version received 11.11.25; accepted 11.11.25; published 23.12.25.

Please cite as:
Kreis LG, König HH, Sikorski F, Löwe B, Kohlmann S, Brettschneider C
Automated Feedback After Internet-Based Depression Screening: Cost-Effectiveness Analysis of a Randomized Controlled Trial
JMIR Form Res 2025;9:e68282
URL: https://formative.jmir.org/2025/1/e68282
doi:10.2196/68282

© Léon Gerardo Kreis, Hans-Helmut König, Franziska Sikorski, Bernd Löwe, Sebastian Kohlmann, Christian Brettschneider. Originally published in JMIR Formative Research (https://formative.jmir.org), 23.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Facilitators, Barriers, and Cultural Appropriateness of Mindfulness-Based Interventions Among Saudi Female University Students: Qualitative Study

Duaa H Alrashdi[1,2], MSc, MPhil; Carly Meyer[3,4], PhD; Rebecca L Gould[1], PhD

[1]Division of Psychiatry, Faculty of Brain Sciences, University College London, Maple House, 4th Fl, Tottenham Ct Rd, London, United Kingdom

[2]Department of Health Sciences, Faculty of Health and Rehabilitation Sciences, Princess Nourah Bint Abdulrahman University, Riyadh, Saudi Arabia

[3]Bolton Clarke Research Institute, Brisbane, Australia

[4]Department of Clinical, Educational, and Health Psychology, Faculty of Brain Sciences, University College London, London, United Kingdom

Corresponding Author:
Duaa H Alrashdi, MSc, MPhil
Division of Psychiatry, Faculty of Brain Sciences, University College London, Maple House, 4th Fl, Tottenham Ct Rd, London, United Kingdom

Abstract

Background: Mindfulness-based interventions (MBIs) have been shown to improve university students' well-being. However, previous studies have not systematically explored factors that can facilitate or hinder engagement in MBIs among Saudi university students, nor how MBIs can be culturally adapted to meet their needs.

Objective: This study aimed to (1) explore the perspectives of Saudi female university students about factors influencing engagement with MBIs, (2) explore the cultural appropriateness of MBIs, and (3) systematically identify recommendations for developing a culturally appropriate MBI.

Methods: A qualitative research approach was used to collect data using semistructured individual interviews and focus groups. Two established frameworks for behavioral interventions were applied to guide the interview topics and data analysis. The COM-B (Capability, Opportunity, and Motivation Domains of Behavior Change) model was applied to identify potential enablers and barriers influencing students' engagement with MBIs. The cultural adaptation framework by Bernal et al was used to explore the cultural appropriateness of MBIs. Subsequently, recommendations for developing MBIs, with a specific focus on an online version, were systematically formulated using the Theory and Techniques Tool. Data were analyzed using mixed inductive-deductive thematic analysis.

Results: Fourteen Saudi female university students (mean age 24, SD 4.9 years) participated in semistructured interviews and focus groups. Numerous potential enablers and barriers to MBI engagement were identified. Factors that may influence engagement pertained to capability (variation in knowledge of mindfulness), opportunity (anticipated difficulty finding time), and motivation (variation in anticipated and experienced benefits of mindfulness). Participants also highlighted several considerations that may enhance the cultural relevance of MBIs, drawing on the cultural adaptation domains by Bernal et al. These included the importance of aligning MBIs with the local cultural context, incorporating metaphors and examples rooted in Saudi and Arab culture, and accommodating students' preferences for the duration of MBIs. Key recommendations for developing culturally appropriate MBIs for Saudi university students included providing clear information to improve understanding of mindfulness, providing practical strategies and skills to overcome barriers such as time constraints, delivering MBIs in both Arabic and English, and ensuring that MBIs' content aligns with local cultural values and contexts.

Conclusions: Findings and recommendations aim to enhance the feasibility, acceptability, engagement, and effectiveness of MBIs among Saudi university students, particularly female students. However, whether they do in fact achieve these aims is unknown. Future research should endeavor to evaluate the effectiveness of proposed recommendations and explore the enablers and barriers to MBI engagement in a broader population of Saudi students.

(JMIR Form Res 2025;9:e78532) doi:10.2196/78532

KEYWORDS

mindfulness; university student; Saudi Arabia; mental health; qualitative study; culture

XSL•FO
RenderX

Introduction

Mindfulness-based interventions (MBIs) are interventions that are grounded in the practice of mindfulness, typically defined as the "awareness cultivated by paying attention in a sustained and practical way: on purpose, in the present moment, and non-judgmentally" [1]. Examples of MBIs include mindfulness-based cognitive therapy and mindfulness-based stress reduction. Both online and face-to-face MBIs have been shown to improve the mental health and well-being of university students, including reducing depression, stress, and anxiety [2,3]. Yet despite the effectiveness of MBIs, previous reviews indicate that many students who embark on MBIs do not complete them. For example, a meta-analysis of 23 studies of online MBIs reported an overall attrition rate of 34% [2]. The reasons underlying this are not clearly understood. However, some have argued they might include a perceived lack of time and requiring too much effort, along with the challenge of adopting new habits for regular practice among university students [4].

Improving MBI design to facilitate student engagement is one way of potentially increasing MBI completion levels. Recent studies have involved university students in the development of MBIs to ensure that interventions are more relevant and responsive to students' needs and preferences [5,6]. However, no studies, to date, have systematically explored potential barriers to and facilitators of intervention engagement in university students using an established theory-informed approach to inform MBI design. Therefore, this study adopted a behavioral science lens to better understand students' engagement with MBIs. The COM-B (Capability, Opportunity, and Motivation Domains of Behavior Change) framework suggests that to implement a successful behavioral intervention (such as an MBI), a comprehensive analysis of the underlying behavioral determinants is needed [7]. Such an analysis requires consideration of 3 essential domains: capability, the physical and psychological abilities of a person to engage in an activity; opportunity, factors within both the physical and social environment that lie beyond the person and can either facilitate or hinder activities; and motivation, the internal mental processes within the person that can influence engagement in activities and includes both reflective and automatic processes [7]. The COM-B framework is widely used in health care settings, including the development of mental health interventions [8,9], and was therefore applied in this study to guide the development of interview topic guides and data analysis.

It is a commonly accepted practice to align COM-B findings with the Theoretical Domains Framework (TDF) [10-12] to identify strategies to improve intervention development. The TDF is a comprehensive synthesis of psychological theories and constructs, which includes domains that are used to understand factors influencing behavioral change [10]. These domains represent the theoretical constructs mediating an intervention's effects [13], which should align with anticipated barriers and enablers. The TDF can be considered an extension of the COM-B model, as it provides a more detailed framework for exploring behavioral determinants [12]. In this study, it was used to expand on the COM-B findings by mapping the identified factors that may influence students' engagement with MBIs onto relevant theoretical constructs.

Furthermore, the Theory and Techniques Tool is an interactive mapping resource designed to identify links between the theoretical constructs of the TDF and relevant behavior change techniques (BCTs), developed based on a consensus process [13-15]. This tool enables the selection of BCTs that directly target identified barriers and enablers, thereby informing the design of tailored interventions [13,14]. The TDF, as noted earlier, is used to identify and understand behavioral determinants, which can subsequently be mapped onto the Theory and Techniques Tool to develop evidence-based recommendations. This approach has previously been applied to the development of health care interventions, such as aiding mental health professionals in delivering behavioral activation interventions to young people, and informing the delivery of pharmacy-based interventions [16,17]. Consequently, the Theory and Techniques Tool was used in this study to guide the development of recommendations for improving engagement with MBIs.

In addition to examining barriers to and facilitators of intervention engagement, it is crucial to consider the cultural appropriateness of interventions for the target population to improve both their effectiveness and retention levels [18]. A recent systematic review identified only two MBI studies conducted in the Kingdom of Saudi Arabia (KSA), both involving Saudi university students [19]. One small randomized controlled trial (RCT) included 40 male students and reported a statistically significant improvement in psychological distress among those who received an MBI compared with a no-treatment control at postintervention [20]. In contrast, another pilot RCT included 26 female students and found no statistically significant differences in psychological well-being, life satisfaction, mindfulness traits, depression, or stress symptoms between the MBI group and an active control at postintervention [21]. Although both studies implemented some degree of cultural adaptation, for example, by modifying intervention delivery to better align with the target population, and with one study further adapting the intervention content (eg, incorporating elements of Islamic spirituality) [21], they generally fell short of achieving comprehensive cultural adaptation [19]. These findings have raised questions about the extent to which these studies met key cultural considerations (eg, whether the interventions were conceptually grounded in, or goal-oriented toward, alignment with participants' cultural backgrounds) that are believed to influence the effectiveness of interventions implemented outside their original cultural context [19]. Critical methodological limitations were evident in both studies, including high or unreported attrition rates and small sample sizes that were underpowered to examine effectiveness or detect significant between-group differences. Notably, neither study conducted a formal analysis of participants' experiences with the interventions, which may have resulted in a lack of clarity regarding students' perceptions and the overall acceptability of such interventions. A similar lack of culturally adapted MBIs has been observed across the broader Gulf Cooperation Council region, which shares similar cultural and regulatory contexts with KSA [19]. It is worth noting that although the studies

XSL•FO
RenderX

conducted in this region demonstrated a limited degree of cultural adaptation (eg, adjustments to intervention delivery were made) [19], none developed their intervention protocols using systematic or theory-driven frameworks to address cultural considerations or enhance participant engagement, which may have limited the robustness and cultural relevance of these interventions. This gap highlighted the need to explore culturally sensitive, theory-informed MBIs tailored to Saudi populations.

One way to systematically develop culturally appropriate MBIs for Saudi university students is to use an established theory-informed approach such as the cultural adaptation framework by Bernal et al [18,22]. This framework has previously been used to evaluate the cultural appropriateness of mental health interventions [19,23,24]. It considers cultural modifications with respect to eight domains: (1) language (eg, the language of MBI materials), (2) people (eg, the gender of instructors delivering mindfulness materials and how written language is presented), (3) metaphor (eg, the use of symbols, sayings, wisdom, poems, and similar that resonate culturally with students), (4) content (eg, students' perspectives about the core components typically included in MBIs), (5) goal (eg, potential goals that students aim to achieve through participation in online MBIs), (6) concept (eg, how students conceptualize mindfulness and how it can be presented to them in an acceptable manner), (7) method (eg, the practical procedures of delivering online MBIs from a student's perspective), and (8) context (eg, the general integration of MBIs, including online versions, within the culture and university setting).

At present, it remains challenging to draw robust conclusions about the acceptability or effectiveness of MBIs among Saudi university students, given the limited number of studies conducted in the KSA, including one pilot RCT and another small-scale RCT, both with small sample sizes that were underpowered to examine effectiveness, and, more broadly, in comparable populations such as those in the Gulf Cooperation Region. Although this scarcity of evidence applies to both female and male Saudi university students, this study focused on Saudi female students, based on previous findings indicating that being female is associated with a higher prevalence of mental health challenges among Saudi university students [25,26]. For instance, symptoms of depression have been reported to be significantly more prevalent among females (50%) than males (39%) [26]. Such findings have prompted calls for initiatives specifically tailored to address Saudi women's mental health [27]. Additionally, while this study explored MBIs in general, it placed particular emphasis on online delivery. It has been suggested that self-guided MBIs (eg, online versions) can be cost-effective, particularly in contexts where large numbers of users can access them [28]. Furthermore, online MBIs have been proven to effectively improve students' psychological well-being [2], suggesting they are a viable option for increasing accessibility for this population.

Consequently, given the lack of knowledge about factors that can facilitate or hinder engagement in MBIs among university students, and how MBIs can be culturally adapted to meet the needs of Saudi female students, this study aimed to (1) explore potential barriers to and facilitators of engagement in MBIs in a Saudi context using the COM-B framework; (2) explore Saudi female university students' perceptions of mindfulness and how to culturally adapt MBIs, particularly online interventions, using the cultural adaptation framework by Bernal et al [22]; and (3) systematically identify a series of recommendations using Theory and Techniques Tool that could be incorporated into the development of an MBI for Saudi university students, especially female students, in order to potentially enhance its effectiveness.

Methods

Study Design

A qualitative research approach was used to collect data in this study using semistructured individual interviews and focus groups with Saudi female university students. This study was reported in line with the Consolidated Criteria for Reporting Qualitative Research (COREQ) checklist [29].

Materials

The interview topic guides were developed based on two frameworks (see Multimedia Appendix 1). The COM-B framework [7] was used to understand factors influencing engagement with MBIs (eg, Do you have any physical health conditions that could impact your ability to practice mindfulness or meditation exercises? How would you juggle practicing online mindfulness or meditation exercises alongside other demands on your time?). The cultural adaptation framework by Bernal et al [18,22] was used to address the cultural appropriateness of MBIs, including online versions (eg, What kind of metaphors or sayings do you think we should include in an online MBI? What do you hope an online MBI could be used for?). Prompts, including examples of mindfulness exercises, definitions, and concepts, were presented to all participants and used throughout the interviews to explore students' perspectives in more depth.

Participants

Participants were recruited from a women-only government university in KSA. Inclusion criteria were as follows: (1) currently enrolled at the university as a student, (2) aged 18 years and over, (3) fluent in Arabic language, and (4) has access to the internet. There were no specific exclusion criteria. As this study was exploratory and aimed to gather initial data, particularly given the limited research in KSA, participants were not purposively selected or allocated to focus groups based on their prior mindfulness experience. Information on participants' prior mindfulness experience was collected solely to provide background context, and so the same topic guide was used for all participants.

A cohort of 12 participants has been suggested to be sufficient for gathering in-depth information from homogenous populations [30]. Therefore, we sought to recruit 12 - 15 participants, as this study focused on a homogeneous group of Saudi female university students and aimed to collect in-depth information. Additionally, the dataset collected was substantial, as it was guided by two theoretical frameworks, and so given the focused and detailed nature of this study, this range was considered sufficient to generate meaningful insights aligned

XSL•FO
RenderX

with the aims of this in-depth exploration. Participant recruitment continued until the target range was reached, and no new themes were observed to be generated. This cohort therefore has the potential to be transferable to other female sectors of universities within the KSA, as these institutions typically share a similar gender-segregated education system.

Ethical Considerations

Ethical approval was granted from both the research ethics committee at University College London (23503/001) and the research ethics committee at the government university in KSA (22‐0390). This study was conducted in accordance with Helsinki Declaration guidelines. Informed consent was obtained from all participants prior to their inclusion in the study. All participant data were stored in password-protected files and pseudonymized to ensure the privacy and autonomy of participants. No compensation was provided to participants.

Procedure

Participant recruitment and interviews were conducted from November 2022 to January 2023. Potential participants were informed about the study using flyers and a brief study description that were disseminated via email and messaging platforms, and via academic staff presentations to their respective students. Potential participants expressing interest were invited to contact the first author (DHA) and were provided with an information sheet and asked to confirm their eligibility using an online survey (ie, Qualtrics). Fully informed consent was subsequently collected from all participants. Those who consented were asked to provide demographic information (eg, age), as well as their preferred email address and meeting time, also collected via Qualtrics. Reminders were sent to students who had registered their interest but had not consented to the study approximately 1 week after their initial contact.

All participants were invited to participate in a focus group; those who were unable to attend a focus group were interviewed individually. Individual interviews and focus groups were conducted via videoconferencing software (ie, Microsoft Teams). Participants were asked to turn off their cameras and were encouraged to use fictitious names to maintain their privacy. All interviews were conducted in Arabic by the first author (DHA) and were audio-recorded and auto-transcribed via Microsoft Teams. The first author (DHA) subsequently checked the transcripts for accuracy and removed all identifying information. DHA translated the quotes from Arabic to English for discussion with other research team members (CM and RLG) and for reporting findings. To assess the reliability of the translations, 20% (61/307) of the total coded English quotations were randomly selected across all themes. Two bilingual reviewers (fluent in Arabic and English) independently evaluated these for semantic equivalence with the Arabic originals. Interrater agreement between the reviewers was 95%, and minor revisions were subsequently made to translated quotations to ensure linguistic accuracy.

Data Analysis

Descriptive statistics were used to describe participants' characteristics. The dataset was analyzed using mixed inductive-deductive thematic analysis. This approach aligns with the aims of this study as it enables both indexing to the frameworks and inductive generation of themes. First, a "top-down" deductive approach was used, wherein data were extracted in relation to domains and subdomains of the COM-B framework (ie, capability: physical and psychological; opportunity: physical and social; and motivation: reflective and automatic). Data were also extracted in relation to the cultural adaptation framework by Bernal et al [22] (ie, language, people, content, metaphor, method, concept, goal, and context). To facilitate the management of this process, a method similar to the 2013 framework analysis approach by Gale et al [31] was used, in which separate tables for each subdomain and domain were created using Microsoft Word, and all relevant quotations from each participant were referenced to the appropriate table.

Second, an inductive "bottom-up" approach using reflexive thematic analysis [32] was used to identify codes and generate themes within each subdomain and domain of both frameworks. Findings were analyzed at both the semantic level, which captures explicit meanings conveyed by participants, and the latent level, which uncovers underlying, implicit meanings [32]. The first author (DHA) developed codes, subthemes, and themes, and two members of the research team (CM and RLG) independently reviewed the analysis. The reviewing process involved multiple rounds of discussion among the research team, followed by joint meetings until the final themes were agreed upon.

A series of recommendations for the development of MBIs, with a specific focus on an online version, was systematically formulated using the Theory and Techniques Tool. Themes within the COM-B domains were mapped to the theoretical constructs of the TDF as per guidance [10]. Subsequently, these theoretical constructs were mapped to BCTs using the online matrix incorporated within the Theory and Techniques Tool, and BCTs with strong consensus were selected [13,14]. Recommendations for MBI development were then formulated based on all selected BCTs, with a few exceptions that were deemed irrelevant (eg, 2.6 Biofeedback, as biofeedback monitoring devices or measurements are not typically used in MBIs). Themes within the cultural adaptation domains by Bernal et al [22] were used to guide the development of culturally specific recommendations, as this framework comprises many practical procedures facilitating its implementation. For example, if a language need was identified, recommendations may include translating and adapting intervention materials into the students' primary language. DHA developed the recommendations, and CM and RLG independently reviewed them, followed by discussions in several team meetings until they were agreed upon. Figure 1 illustrates how the theoretical frameworks were operationalized throughout the study.

XSL•FO
RenderX

Figure 1. Operationalization of the theoretical frameworks used across data collection, analysis, and recommendation development [22,31]. BCT: behavior change technique; COM-B: Capability, Opportunity, and Motivation Domains of Behavior Change [7]; TDF: Theoretical Domains Framework [10].

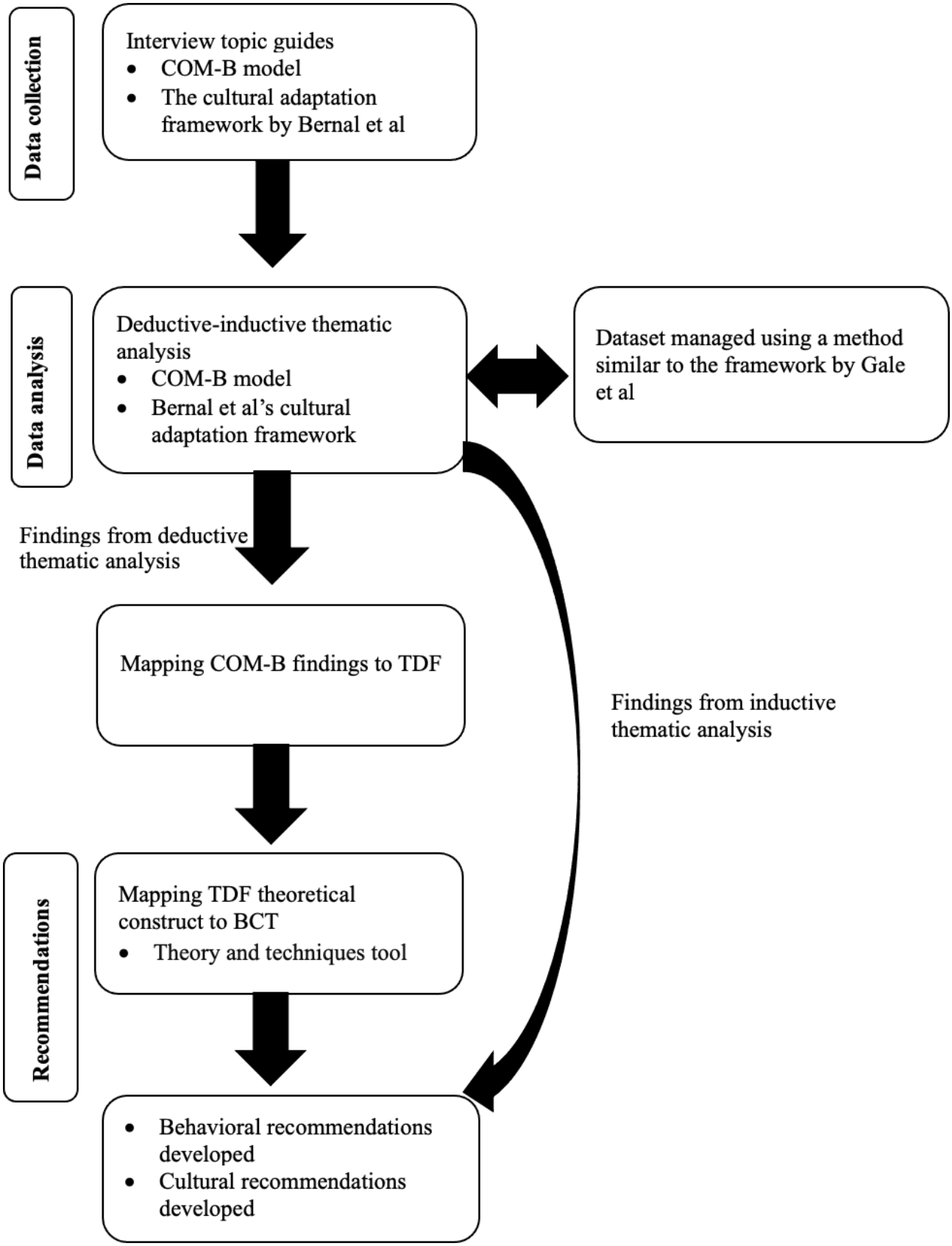

XSL•FO
RenderX

Reflexivity

DHA, a Saudi female PhD candidate, conducted the interviews after receiving training in qualitative research methodology. DHA had previous experience with MBIs, including participating in an online MBI and conducting research on MBIs with consideration of cultural context. This experience may have fostered a favorable attitude toward the efficacy and benefits of MBIs, as well as personal insights into potential cultural adaptations. Such views might have shaped the way DHA conducted interviews and facilitated focus groups, potentially influencing the questions asked or the interpretation of participant responses. DHA had an unexpected prior relationship with 3 of the participants, who were her classmates in a previous course. However, DHA had no prior relationship with the majority of participants (11/14). Steps were taken to establish a good rapport with all participants during the interviews. DHA maintained a reflexive journal throughout the analysis to document her assumptions and interpretations, ensuring that these were critically examined when necessary. She also engaged in regular debriefing with the research team (CM and RLG) to discuss and refine the findings. Prior to the interviews, participants were informed about the aims and reasons for the study through the informed consent process, which reflected the researchers' interests. The coauthors have experience in qualitative research and in applying the COM-B framework, TDF, and the Theory and Techniques Tool.

Results

Participant Characteristics

A total of 34 students contacted the researcher (DHA). Out of these students, 18 did not provide informed consent and 16 consented. Two students, among those who consented, either did not provide a time for a meeting or did not attend the allocated meeting. In total, 14 students were interviewed, with 4 focus groups being conducted with 10 students (2 - 3 students per group), and 4 students being interviewed individually. The average duration of the individual interviews was 1.16 (range 1.06 - 1.27) hours, and for the focus groups, it was 1.36 (range 1.27 - 1.45) hours.

Table 1 shows participants' characteristics (N=14). All participants were female Saudi students with an average age of 24 (SD 4.9, range 19 - 32) years, and the majority reported being single (n=9, 64%). Half of the participants (n=7, 50%) were enrolled in a bachelor's program, while the remaining half were enrolled in a master's program. More than half of the participants (n=8, 57%) reported prior experience with mindfulness, and out of those, 4 (50%) were actively practicing mindfulness at that time, with an average practice time of approximately 11 (SD 11.8, range 2 - 28) hours per week.

Table . Participant characteristics (N=14).

Characteristics	Values, n (%)
Gender	
Female	14 (100)
Age in years	
19 - 20	5 (36)
21 - 29	7 (50)
30 - 32	2 (14)
Nationality	
Saudi	14 (100)
Experience with mindfulness	
Yes	8 (57)
No	6 (43)
Currently practicing mindfulness	4 (22)
Average practice hours per week	
2 - 4 hours	2 (50)
10 - 28 hours	2 (50)

Mindfulness Experiences

There were a diverse range of responses in relation to students' past or current experiences with mindfulness. Although 8 students reported previous experience with mindfulness when providing background information, only 5 could be categorized as having formal mindfulness experience during later questioning. These latter students had learned, practiced, or maintained their mindfulness practice through formal courses or as part of their education, such as programs focused on mindfulness or similar approaches containing mindfulness elements (eg, acceptance and commitment therapy). The remaining 3 students reported that they had become aware of mindfulness only through other sources, such as booklets or social media, and had not completed a formal course but had instead attempted to practice mindfulness on their own. Others

XSL•FO
RenderX

indicated that they had neither practiced mindfulness nor read about it. Therefore, subgroup analyses between students with and without formal mindfulness experience were not conducted in this study, as only a small number of participants (5/14) demonstrated sufficient familiarity with mindfulness to permit meaningful comparison.

The COM-B Framework

Overview

Facilitators of and barriers to students' engagement with MBIs, with a particular focus on online versions, were explored via the COM-B framework. Table 2 shows the COM-B domains and subdomains alongside the key themes and relevant quotations. All themes, subthemes, and illustrative quotations are provided in Multimedia Appendix 2.

XSL•FO
RenderX

Table . Key findings of the COM-B (Capability, Opportunity, and Motivation Domains of Behavior Change) framework.[a]

Domains, subdomains, and themes	Quotes
Capability	
Physical capability	
No significant health issues impacting mindfulness practice or online course use	• "I had [a physical injury] that prevented me..." [Participant M]
Psychological capability	
Variability in knowledge of mindfulness	• "Maybe it makes you getting busy from the reality, living more inside, in your thoughts [...]" [Participant A] • "...There may be specific groups of people who, for example, have high anxiety, high levels of worry, or those who have obsessions." [Participant J]
Varied cognitive skills needed to engage in mindfulness practice	• "I don't think it is easy [to pay attention for a specific time], especially for people who easily get distracted..." [Participant A] • "I can divide the time and allocate some for it, just like when I allocate time for studying, prayer, and self-care. I can incorporate it into my schedule." [Participant O]
Opportunity	
Physical opportunity	
Variation in access to appropriate place for mindfulness practice	• "Though I try my best [to find a place] because of the noise from family at home, it isn't always easy, but Okay." [Participant A]
Anticipated difficulty finding the time to practice mindfulness	• "...You may need it more during study and exam periods, but maybe the time is tight, and you feel like you're almost catching up with primary things to do in your day in relation to work,...studying,...family. You don't have enough time." [Participant A]
Good access to technology but varied network connectivity	• "Yes, it's available, and also easy." [Participant C] • "I just have reservations about the technical issues aspect...any online course I take or any conferences I attend, I encounter technical problems in this matter. So, I assume that this is an obstacle unless...no internet connection is needed." [Participant N]
Social opportunity	
Influence of social environment on mindfulness practice	• "I only know two; one from my family and a friend." [who practice mindfulness] [Participant B] • "I tried more than once to practice it but they [family members] didn't like that I stay away from them, they thought that I was upset with them." [Participant B]
Societal norms and perspectives	• "...You do notice that some people, not all of them, may doubt its effectiveness because of their ignorance about it..." [Participant L] • "...These are foundational concepts that already exist within us, and we already apply them...As Muslims, we already do it, but how we can be aware of it..." [Participant M]
Motivation	
Reflective motivation	
A range of anticipated beliefs about the benefits of mindfulness	• "I can see its importance in reducing symptoms, whether it's anxiety or even depression..." [Participant M] • "I think...acceptance that I accept anything in my life, whether it's pain or negative emotions..." [Participant N]
A range of experienced beliefs about the benefits of mindfulness	• "...I've a little bit let's say my anxiety gets higher with exams, with the due date,...but always when I feel that I started to have stress...I say and think that it's okay, don't think about tomorrow, you live the moment, do whatever you can do...So when I place myself in the set of the present I become better..." [Participant D]

XSL•FO
RenderX

Domains, subdomains, and themes	Quotes
Beliefs about the perceived side effects of mindfulness	• "...In some cases, they start having anxiety like struggle to stop their thoughts, so their stress increases. Because it [mindfulness] needs time and some require immediate results..." [Participant D] • "...Initially,...I tried to practice it...but there were times when I got bored and felt annoyed, and within just 2 - 3 minutes, that's it...I didn't want to continue with the exercise anymore..." [Participant J]
Automatic motivation	
A range of emotions associated with mindfulness practice	• "...I'm afraid...I don't like doing something and not completing it or not being in the way that I want..." [Participant A]

[a]All themes, subthemes, and representative quotations are provided in Multimedia Appendix 2.

Capability

Students generally reported no health issues hindering their ability to practice mindfulness or engage in online platforms, except for minor difficulties such as physical injuries or low vision, which they could adjust for. One notable aspect within psychological capability was the variation in *knowledge of mindfulness* among students. There was a varied understanding of the concept of mindfulness, ranging from familiarity to contradictory views of its principles, or not knowing about mindfulness. Students also had a varying understanding of who mindfulness is suitable for, with references made to everyone, those with stress, or people facing mental or physical difficulties.

Another reported barrier to engaging in mindfulness practice was the variation in students' ability to concentrate and attend for specific periods of time. Half of the students (7/14) found this challenging, whether they had experienced these difficulties during mindfulness exercises, anticipated such challenges, or noticed them during other activities (eg, while watching shows). They mentioned several reasons for this, such as their mind wandering. However, 3 students with previous experience in mindfulness observed that both concentration and attention could improve with consistent practice. Conversely, most students (11/14) demonstrated good time management skills; for example, they shared various strategies for their usual way of organizing their schedules and potential plans to allocate specific times and days for mindfulness.

Opportunity

As part of the opportunity afforded by the *physical environment*, students emphasized that finding or expecting to find a suitable place for mindfulness practice was generally manageable. However, they also mentioned the challenges of accessing such a place due to the presence of others in the home. Finding time to allocate specifically for mindfulness practice was more commonly anticipated to be challenging, given the tightness of students' university schedules alongside other personal commitments. Students also stressed potential distractions from their surroundings that might negatively impact their mindfulness practice, such as social media and smart devices. Turning to the aspect of *technology*, while all students acknowledged having access to the appropriate equipment to engage in online platforms, 3 reported challenges with network connectivity, which may affect the quality of the materials received.

Regarding the *social environment* that influenced mindfulness practice, students acknowledged that mindfulness practice was uncommon among their family and friends, highlighting that they knew few people (eg, none or only 2 people) in their immediate circles who actively engaged in mindfulness. However, the extent to which others' opinions about mindfulness might affect students' own practice varied. Most students (11/14) considered mindfulness practice a personal choice, and had they experienced its benefits, they would have become less concerned about others' opinions. Two students mentioned that if they were to have encountered criticism from people around them, they would have preferred to practice it privately without their knowledge to avoid such issues. Students also either experienced or anticipated receiving different levels of support from family and friends regarding their mindfulness practice. For example, 3 out of 14 students noticed that their family members were less supportive of their mindfulness practice due to reasons such as not understanding what the students were doing or being slightly annoyed by the students' spending time alone. On the other hand, 4 students found that their families were not only supportive but also eager to join in.

Beyond their immediate social circle, *societal norms and perspectives* also influenced students' engagement. Many students (10/14) acknowledged a general lack of awareness about mindfulness within society, indicating there was no clear understanding of what mindfulness is and its potential utility, which may impede its acceptance (eg, associating mindfulness with triviality, which might lead to a skeptical attitude toward mindfulness). They also emphasized the need for greater societal awareness of mindfulness. Three students admitted that they, too, were initially skeptical due to not knowing what mindfulness was until they gained a deeper understanding. However, 4 students mentioned that society nowadays, particularly the younger generations, is more aware of mindfulness and is more accepting of it. Another substantial group of students (12/14) acknowledged that mindfulness aligned with either Arabic culture, Islamic values, or both. They emphasized that many of its principles, once prompted to think about them, reflected what they and people in society were already doing and had been taught to do, such as practicing kindness and appreciation.

Motivation

Students held diverse perceptions of *anticipated benefits of mindfulness*, ranging from reductions in psychological distress

XSL•FO
RenderX

(eg, stress), improvements in well-being, flexibility, acceptance and appreciation, and self-awareness, to broader expected benefits such as improvements in concentration and attention. Some of these were also reflected in students' *experienced benefits of mindfulness*. They reported that it helped in managing stress, anxiety, and physical pain and felt more aware and connected to the present moment. Conversely, 4 students pointed out the *perceived negative side effects of mindfulness*. Three of them highlighted that mindfulness may not be suitable for those seeking immediate results, people who are self-conscious about their bodies, or those who have panic attacks. Two of these students also found the practice of mindfulness annoying and boring.

Turning to students' *intentions to engage in mindfulness practice*, participants anticipated the importance and need for mindfulness in various situations, emphasizing its relevance, for example, when feeling overwhelmed, in daily routines, and during moments of enjoyment. Their intentions to engage in online MBIs, however, varied. Of the 14 students, 9 were eager to join such programs, recognizing them as a valuable skill to learn and build upon. In contrast, 4 students felt they had already acquired the necessary skills, making participation less important, while 1 student preferred face-to-face interactions for learning mindfulness.

Furthermore, students mentioned several *perceived facilitators of mindfulness practice* that could enhance their motivation to engage in mindfulness or maintain their practice. Examples included practicing with others to create a shared experience and aligning mindfulness practices with cultural contexts. Two students highlighted how anchor points, such as starting with breathing or focusing on certain bodily sensations, helped them or could help others stay present during mindfulness exercises. Shifting the focus to automatic motivation, 3 key *emotions associated with mindfulness practice* were identified among students: anticipated fear of failure and not completing practices properly, excitement about engaging in mindfulness, and enjoyment experienced from previous mindfulness practice.

The Cultural Adaptation Framework by Bernal et al

Overview

Cultural adaptations that may facilitate students' engagement with MBIs, with a particular focus on online versions, were addressed via the framework by Bernal et al [22]. Table 3 shows the domains, key themes, and relevant quotations. All themes and corresponding quotations are provided in Multimedia Appendix 3.

XSL•FO
RenderX

Table . Key findings of the cultural adaptation framework by Bernal et al [22].[a]

Domains and themes	Quotes
Language	
Variation in language preferences	"Arabic is better because it's our language; expression is better in it. As for the English language, it requires specific expressions and you struggle to find the right words, which can be somewhat challenging." [Participant E]
People	
Varying preferences for the gender of instructors	"...It doesn't matter to me who the instructor is; I feel that...the tone of the voice is more important than the speaker's gender..." [Participant H]
Metaphor	
Incorporating metaphors associated with Saudi and Arab culture	"...Examples in mindfulness always seem to relate to the lives of foreigners...you can't really use these examples into the lives of Saudis. Even the example of the train—we don't have a train. Yes, I like the example, but it's weird for me to tell someone [about it]. Well, I don't have a train in my life anyway to watch, how you want me to imagine it...I feel it's nice to translate examples into things that happen in the real lives of Saudis or Arabs in general." [Participant N]
Varied viewpoints in using Islamic-related metaphors	"Maybe in our religion. We can find in the Sunnah an-Nabawiyyah [the actions, sayings of Prophet Muhammad (peace be upon him)]..." [Participant O]
Content	
Acceptability of mindfulness exercises	"I feel this concept [appreciation] is very beautiful and very good, and I don't feel that it needs any adjustments—like it's already good as a whole. And, I even feel that it fits with our culture and environment, and the age you're talking about for people at the university." [Participant D]
Goals	
Variation in students' goals for the online MBIs[b]	"Yes, stress...and how they deal with the negative and adverse effects. When they encounter a distressing situation, how do they bring mindfulness to that moment and manage to cope with the situation." [Participant O]
Concept	
Students' perspective on mindfulness aligned with its core principles	"...I might explain to them [students]...that it's about you focusing on your awareness from all aspects—your emotions, sensations...whether internal or external, in the present moment without any interruptions from life...and it doesn't mean you need to make any interpretations of the sensations you are feeling or seeing." [Participant D]
Method	
Variation in the preferred online platform	"I feel an application, if there is an application specifically made for it, it will be much better..." [Participant A]
Preferences for receiving reminders and prompts	"...I feel it's really important because...It's nice...to have something that reminds you to bring your attention and focus." [Participant J]
Varying preference about the duration of the online MBI and mindfulness exercises	"I prefer to have both options honestly. I'd like beginners to have a short duration to try it out, and specialised individuals want more..." [Participant N]
Context	
Factors facilitating mindfulness implementation within Saudi society	"That it focuses in our believes. Beliefs of Saudi, that would be more suitable." [Participant O]
Factors facilitating mindfulness implementation within university setting	"...Maybe have a bonus or incentive from teachers to encourage attendance for this course initially. Not everyone might be interested at first, but I think if there's a motivator, it can help students consider taking this course." [Participant N]

[a]All themes and representative quotations are provided in Multimedia Appendix 3.

[b]MBI: mindfulness-based intervention.

XSL•FO
RenderX

Language

Students expressed varied preferences about the language for receiving online MBIs. Many students (8/14) stressed a preference for Arabic, their native language, as it facilitates self-expression and understanding. Conversely, 5 students emphasized that English could be more suitable, particularly for terms such as "mindfulness," which were challenging to translate and comprehend in Arabic. Three students also expressed interest in receiving materials in both languages, indicating a potential need for flexibility.

Person

Most students (11/14) expressed that the gender of the person delivering the mindfulness audio recordings was not significant to them. They prioritized factors such as the quality and clarity of the voice and tone. However, 2 students mentioned a preference for a female voice. Three students also expressed a preference for written Arabic to be personally addressed to them rather than presented in a general manner.

Metaphor

A notable observation among students was the positive attitude toward using metaphors associated with Saudi and Arab cultures. Students emphasized that the closer the illustration aligned with the culture of the participant, the better. However, views varied with respect to the appropriateness of incorporating Islamic-related metaphors into an online MBI.

Content

At the start and during the interviews, students were given examples of mindfulness exercises (eg, sitting meditation and body scan) to introduce them to different concepts within mindfulness. Students acknowledged the acceptability of various mindfulness exercises (eg, body scan and loving kindness), emphasizing the suitability and likability of these exercises. Almost all students (13/14) also highlighted the appropriateness and acceptability of these exercises within the cultural context or university environment. However, whether based on their previous experiences or when presented with these exercises, 4 students emphasized the importance of improving awareness about mindfulness concepts and exercises, as it facilitates students' acceptance, particularly when the exercises may be unfamiliar to them.

Goals

The potential goals and reasons why students undertook, or were willing to undertake, mindfulness practice varied considerably. Examples included improving their understanding of what mindfulness is, improving psychological well-being, managing psychological distress (eg, coping with stress), being more aware and connected to the present moment, making mindfulness a daily habit, and improving academic performance (eg, concentration).

Concept

Generally, students defined mindfulness in similar terms to its typical formal definition, including aspects such as paying attention to the present moment without judgment. However, students stressed the need for clarity in the meaning of mindfulness and identified key principles that can be used to clearly illustrate it. Examples included providing a precise definition with real-life examples, simplifying the meaning of nonjudgment, addressing misconceptions (eg, mindfulness not being merely relaxation), and normalizing potential challenges in the initial stages of mindfulness practice (eg, mind wandering).

Method

Students showed diverse preferences for the online platform through which they receive the intervention. Of the 14 students, 7 favored software applications, while 4 leaned toward external platforms, such as a website or a university's online learning platform. Opinions on using the university's platform, however, were mixed: 3 students viewed it as convenient for integrating the intervention alongside academic resources, whereas 1 felt that the intervention better remains separate to allow disengagement from study. Students also highlighted the appropriateness of receiving reminders and prompts to practice mindfulness when participating in online MBIs. However, 3 students pointed out that overly frequent reminders might not be ideal. Furthermore, students expressed varied intentions to engage in online peer-support forums.

When it came to the duration of the mindfulness course, students demonstrated a range of preferences. Of the 14 students, 6 favored brief courses (≤4 weeks), while 4 preferred longer ones (>4 weeks). One student stressed the importance of flexibility, suggesting the inclusion of both options. These preferences also extended to the mindfulness exercises, where students indicated a preference for either brief or longer sessions or having the flexibility to choose between the two options.

Context

Students highlighted various factors facilitating the implementation of mindfulness and online MBIs within the context of Saudi society. They emphasized the importance of integrating mindfulness into the lives of Saudis by aligning it with social habits and values, having it presented by Saudi individuals, and increasing awareness in society (eg, via social media). Students also highlighted facilitating factors within a university setting. They emphasized the crucial role of spreading awareness about mindfulness through social media and effective advertising within the university itself to reach a large proportion of students. Students also stressed the significance of university involvement in encouraging mindfulness practice. Two students also discussed how aligning this intervention with mental health could either facilitate or hinder students' engagement in mindfulness.

Implications for Mindfulness-Based Interventions

A series of recommendations for developing an MBI, with a specific focus on an online version, for Saudi university students are outlined in Multimedia Appendices 4 and 5. These recommendations aligned with the findings identified from the COM-B and the cultural adaptation framework by Bernal et al [22].

Broadly, recommendations based on the COM-B, TDF, and the Theory and Techniques Tool included specific directions that

XSL•FO
RenderX

may facilitate students' engagement with MBIs, further detailed in Multimedia Appendix 4. These included but were not limited to providing sufficient information and guidance (eg, on mindfulness practices and how to create a suitable environment for their practice), providing students with strategies and skills to conduct their own mindfulness practice more effectively and overcome obstacles (eg, finding time, improving concentration, and incorporating mindfulness into their daily routines), and providing students with resources to participate effectively in the intervention (eg, how to seek social support from people around them).

Similarly, recommendations based on the cultural adaptation framework by Bernal et al [22] consisted of practical steps to accommodate the preferences, needs, and cultural viewpoints of Saudi university students in the development of the MBI, as detailed in Multimedia Appendix 5. These included, but were not limited to, identifying essential mindfulness components and exercises that need to be incorporated into the intervention to align with students' goals (eg, incorporating compassion and kindness), adapting mindfulness-related elements that resonate with students' cultural backgrounds (eg, using Arabic language and using metaphors and examples relevant to the local culture), incorporating specific practical procedures for intervention delivery (eg, considering the gender of the instructor and allowing flexibility in exercise durations), tailoring the way in which the concept of mindfulness is presented to students (eg, clarifying the concept of "nonjudgment"), and supporting the integration of mindfulness practices into Saudi society and university settings (eg, using social media to raise awareness about mindfulness).

Discussion

Interpretation of Findings

This study adopted a qualitative exploratory design to gain in-depth insights into Saudi female university students' perspectives on MBIs, enabling the identification of potential facilitators, barriers, and cultural factors that could enhance students' engagement with MBIs during this formative phase. Qualitative approaches have previously been used to inform the development of MBIs for university students [5,6], particularly when little is known about how such interventions are implemented [33], as was the case in this study. This study applied the COM-B framework [7] and the cultural adaptation framework by Bernal et al [18,22], adopting a theoretically informative lens to guide data interpretation and to develop culturally sensitive recommendations for MBI design aimed at enhancing students' acceptability of and engagement with the intervention. Overall, Saudi female university students were positively inclined toward MBIs, both specifically as students and within their wider cultural context. Their responses reflected numerous themes drawn from both frameworks, which may either facilitate or hinder engagement with the intervention and potentially improve its cultural suitability. These findings are recommended to be integrated into MBI development to accommodate the needs of Saudi female university students.

Knowledge and awareness about mindfulness were identified as a key theme in both frameworks. Students highlighted the need for greater awareness of mindfulness across various contexts, including their own understanding of mindfulness and its benefits, as well as broader societal awareness of the concept. Similar findings have been reported in previous studies, where health care staff described how increased awareness of the effectiveness of mindfulness could enhance their engagement with a self-guided MBI [34]. Furthermore, enhancing general awareness of mental health in Saudi society has been stressed as crucial for improving the uptake of psychotherapies, including MBIs, and perhaps further reducing potential barriers to engagement, such as minimizing stigma [19]. Raising mental health awareness in Saudi society may not only facilitate engagement with psychotherapies but could also break down other barriers to seeking help. Social obstacles such as stigma, cultural and family attitudes, negative perceptions about mental health difficulties, and a lack of awareness about accessing mental health information and services were identified as perceived hindrances to seeking mental health support among adults in KSA [35-40]. Promoting mindfulness awareness within the targeted population and broader Saudi society might facilitate acceptance and engagement in mental health care in KSA. Further research is needed to explore this.

Furthermore, a significant perceived obstacle to student engagement in MBIs found in this study was time. Although students emphasized their time management skills, which may indicate their capability to effectively plan their time to incorporate mindfulness practice into their routines, they perceived that allocating specific time solely for mindfulness was still challenging. This is a common difficulty reported in many previous studies of engagement with MBIs. For example, a narrative review focusing on students' perceptions of MBIs highlighted that university students perceived a struggle to find time for meditation outside of scheduled sessions [41]. Difficulty in finding time may also contribute to the high attrition rate of MBIs among university students. Factors that seem related to time management issues, such as scheduling conflicts, being too busy, and the demands of practice commitments, were reported by students as reasons for dropping out of MBI trials [42,43]. This challenge extends beyond student populations; for example, people with chronic conditions also described being too busy or having other commitments as barriers to MBI engagement [44]. A briefer version of MBIs (eg, 100 minutes per week) has been suggested to address time commitment issues and has been shown to be effective across different populations (eg, university students) and outcomes (eg, anxiety and depression) [45]. However, it is noteworthy that students in this study showed a mixed preference for both shorter and longer durations of MBIs, whether in terms of the number of weeks or the length of mindfulness exercises. Therefore, it might be beneficial to provide students with flexibility by offering both options, which could maximize engagement and accommodate all potential needs. In addition to this, several strategies were recommended in this study to address time-related challenges when developing MBIs. Some of these recommendations align with suggestions from others, such as incorporating mindfulness practices into students' everyday tasks [41]. Future research could further explore and assess the adoption of these strategies.

XSL•FO
RenderX

The findings of this study indicate that it might be important to align MBIs with the cultural background of the target population. This may include adapting the content to be more culturally relevant, for example, by integrating local values, metaphors, and language into the interventions. Generally, the need for cultural adaptations in designing MBIs for university students has been emphasized [41]. This need was further emphasized by how students perceived nonadapted MBIs. For example, university students in the United Arab Emirates, a Gulf Cooperation Council country that shares cultural, linguistic, and regulatory similarities with KSA, reported that incorporating metaphors related to their culture and receiving the intervention in Arabic would have helped them better connect with an MBI and enrich their mindfulness experiences [46]. Furthermore, a qualitative study conducted in Indonesia, which differs culturally from KSA but shares Islam as the majority religion, explored university students' and relevant stakeholders' perspectives on developing an online MBI highlighted the need to culturally adapt intervention content (eg, by incorporating Indonesian social activities and translating materials into Indonesian) [47]. The study subsequently found that an adapted version of MBI was acceptable to Indonesian students, as evidenced by a 70% completion rate. There were also preliminary improvements in students' levels of mindfulness, distress, and well-being at postintervention [48]. Similarly, outside the student population, African women highlighted the importance of incorporating African cultural values and using culturally familiar language to enrich an MBI after receiving a nonculturally adapted version [49]. These findings from diverse contexts highlight that sociocultural factors may play a key role in shaping the acceptability and engagement of MBIs. However, it is noteworthy that, to the authors' knowledge, no studies have directly compared the feasibility, acceptability, engagement, and effectiveness of culturally adapted MBIs to nonadapted versions. Therefore, the extent of the usefulness of such adaptations remains uncertain. Future studies should explore the incorporation of cultural adaptations into MBIs, as well as the acceptability and effectiveness of culturally adapted MBIs compared to nonadapted ones with respect to student engagement and outcomes improvement.

Furthermore, it is important to consider participant characteristics, as these may have shaped the findings and, consequently, influenced how the intervention recommendations were developed. For example, all participants were female and self-identified as Saudi nationals, which may have informed their perspectives on MBIs and, in turn, influenced the study findings. Nonetheless, given that Saudi nationals constitute more than half of the total population (57%) [50], and that the education system in KSA is typically gender-segregated, these characteristics remain relevant to the context of Saudi universities. However, further research is needed to include a more diverse cohort of students in KSA (eg, male participants and non-Saudi nationals) to enhance the transferability of findings.

To the authors' knowledge, it remains unclear whether the evidence-based theoretical frameworks used in this study (eg, the COM-B model, TDF, and the cultural adaptation framework by Bernal et al [22]), most of which were originally developed in Western contexts, are fully applicable to the Saudi cultural setting. As this study is the first to apply theory-driven approaches to exploring MBIs among Saudi university students, it provided an opportunity to reflect on how these frameworks align with local sociocultural contexts. For example, it has been emphasized that the COM-B model does not explicitly account for the cultural background of participants (eg, their beliefs or gender), which may be relevant when designing BCTs [51]. This, in turn, may introduce subjectivity and inconsistency in determining whether and how cultural factors influence specific behaviors when designing interventions [51]. Consequently, there is a need to further test, develop, or adapt theoretical frameworks that are specific to the Saudi or broader Arab context and informed by experts familiar with regional populations and mental health systems. Whether such context-specific frameworks would prove more appropriate than those developed in Western contexts remains uncertain and represents an important direction for future research.

Another area of consideration relates to the fact that some recommendations made in this study are already incorporated into many MBIs (eg, addressing common misconceptions and normalizing experiences and emotions). Since this study followed a systematic approach to identify factors that may influence behavioral change, it suggests the importance of including these elements in MBIs for university students. This study, however, identified several other strategies based on the inclusive selection method of relevant BCTs, which aligned with anticipated barriers and facilitators that may facilitate the uptake of MBIs for students (eg, guidance on how to seek support from family and friends). Consequently, these strategies need to be tested. Future research should consider these implications, which could impact MBI engagement among Saudi university students, and then assess their feasibility, acceptability, and effectiveness.

From a wider perspective, it has been emphasized that the cultural adaptation of evidence-based interventions can be strengthened by embedding them within implementation science frameworks [52]. Such frameworks provide a contextual lens for evaluating institutional readiness, policy alignment, and organizational capacity factors crucial for ensuring that interventions are not only culturally appropriate but also realistically deliverable within existing systems [52]. In Saudi universities, institutional readiness may play a key role in the integration of culturally adapted MBIs. This includes the willingness of institutions to embed such interventions within existing student well-being systems, promote them through official university channels, and raise awareness of mindfulness as a mental health approach among students and staff members. Future studies should therefore adopt a broader implementation plan that directly engages university-level stakeholders when introducing MBIs in KSA. Examples include identifying relevant policy strategies, such as developing institutional guidelines and establishing regulations to support program delivery and promotion [12].

Furthermore, given the high demand for mental health services typically observed in university settings [53,54] and the limited availability of MBI research in KSA, universities in KSA should consider initiating pilot programs to explore the feasibility,

XSL•FO
RenderX

acceptability, and preliminary efficacy of culturally sensitive mental health interventions to support student well-being. Self-guided MBIs, particularly online versions, can be well suited to higher education settings, where large numbers of users can access them at low cost [28]. This approach aligns with the broader objectives of Saudi Vision 2030, which emphasizes improving access to tailored mental health services as part of national development efforts to enhance health care [55].

Strengths and Limitations

To the authors' knowledge, this is the first study to explore factors influencing engagement in MBIs and their cultural appropriateness for Saudi university students. One strength of the study lies in the methodology used, which was based on well-established frameworks, guiding the development of interview topics, data analysis, and MBI recommendations. However, this study has several limitations. Participants were all female and recruited from a single women-only university due to resource limitations. This increased the likelihood of cultural homogeneity among participants, restricting the perspectives to solely female Saudi students in one location. Consequently, valuable insights from male counterparts or students in other parts of KSA need further exploration.

Additionally, the original dataset in Arabic was analyzed and translated by only one researcher, and only 20% of the English-coded quotations were independently reviewed by two reviewers. Despite these measures, some relevant data may have been missed or misunderstood during translation or coding, which could affect the study's conclusions. Useful insights might have also been gained if the study had engaged students throughout the research process, such as involving them in developing the interview topic guides, identifying final themes, and drawing recommendations. This approach could have enhanced the relevance of the recommendations identified in the study.

Furthermore, while the mixed inductive-deductive thematic analysis used in this study aligned with its aims, the back-and-forth transition between deductive and inductive analysis could have introduced research biases (eg, overfitting within the theoretical frameworks), which could affect the study's findings. Despite the importance of considering prior mindfulness experience (since the adoption of MBIs may vary depending on the level of prior experience), comparisons between participants with and without prior experience were deemed inappropriate for this study. Although more than half of the participants in this study self-reported prior mindfulness experience, this varied from completing MBI courses and studying mindfulness as part of their education to engaging with it solely through reading booklets or watching online videos. This diversity meant that only a small number of participants (5/14, 36%) demonstrated sufficient familiarity with mindfulness to allow meaningful comparison between those with and without formal mindfulness experience. Future research should examine the impact of participants' prior experiences with mindfulness on the barriers to and facilitators of engagement with MBIs. Lastly, given the specific focus of this study on online MBIs, some influential factors that may be applicable to face-to-face MBIs engagement discussed in previous studies (eg, the role of facilitators and online vs face-to-face delivery) were not addressed. Future research should explore these factors to enhance the design of MBIs to meet students' needs more effectively.

Future Research

This study provides potentially useful insights for developers of MBIs targeting KSA university students. This study is part of a larger project that follows the Medical Research Council framework for developing and evaluating complex interventions, which involves synthesizing evidence to inform intervention design [56]. Building on the findings of this study, the next phase will involve a mixed methods, uncontrolled feasibility study to assess the acceptability, feasibility, and preliminary signals of efficacy of a culturally adapted online MBI designed for Saudi female university students. If the intervention proves feasible and acceptable, a subsequent RCT with an internal pilot phase could be conducted to evaluate its effectiveness on a larger scale.

Furthermore, given the wide range of recommendations generated in this study, it may be neither feasible nor advisable to implement all identified strategies in future applications. Typically, behavioral interventions adopt a systematic process to prioritize and select the most appropriate BCTs and intervention strategies when designing the intervention [12]. One such approach involves applying the APEASE (acceptability, practicability, cost-effectiveness, affordability, safety, and equity) criteria [12]. These criteria can be systematically applied and evaluated in collaboration with relevant stakeholders (eg, psychologists and university staff) to guide the refinement of the culturally adapted MBI, which can then be tested in future research.

Conclusions

This study explored factors that could facilitate or hinder Saudi female university students' engagement with MBIs and their perspectives on culturally appropriate adjustments, using theory-informed approaches. Subsequently, systematic recommendations were formulated to integrate into the development of an online MBI. These recommendations aim to enhance the feasibility, acceptability, engagement with, and effectiveness of MBIs among Saudi university students, particularly female students. However, whether they do in fact achieve these aims is unknown. Future research should seek to evaluate this further.

Acknowledgments

The authors would like to thank Amna Alshmmary, PhD in Neuropsychology, and Amal Alharbi, PhD candidate in Culture, Communication and Media, for evaluating the translated English quotations.

XSL•FO
RenderX

Funding

No funding was received for conducting this study.

Data Availability

More information supporting this study (eg, quotations) is provided in the multimedia appendices.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview topics based on the COM-B framework and the cultural adaptation framework by Bernal et al. COM-B: Capability, Opportunity, and Motivation Domains of Behavior Change.
[PDF File, 200 KB - formative_v9i1e78532_app1.pdf]

Multimedia Appendix 2
Findings of the COM-B framework with further illustrative quotations. COM-B: Capability, Opportunity, and Motivation Domains of Behavior Change.
[PDF File, 207 KB - formative_v9i1e78532_app2.pdf]

Multimedia Appendix 3
Findings of the cultural adaptation framework by Bernal et al with further illustrative quotations.
[PDF File, 137 KB - formative_v9i1e78532_app3.pdf]

Multimedia Appendix 4
COM-B mapped recommendations using the Theory and Techniques Tool. COM-B: Capability, Opportunity, and Motivation Domains of Behavior Change.
[PDF File, 174 KB - formative_v9i1e78532_app4.pdf]

Multimedia Appendix 5
Recommendations from the cultural adaptation framework by Bernal et al.
[PDF File, 185 KB - formative_v9i1e78532_app5.pdf]

Checklist 1
COREQ checklist.
[PDF File, 178 KB - formative_v9i1e78532_app6.pdf]

References

1. Kabat-Zinn J. Mindfulness for Beginners: Reclaiming the Present Moment—And Your Life: Sounds True; 2012.
2. Alrashdi DH, Chen KK, Meyer C, Gould RL. A systematic review and meta-analysis of online mindfulness-based interventions for university students: an examination of psychological distress and well-being, and attrition rates. J Technol Behav Sci 2023;9(2):211-223. [doi: 10.1007/s41347-023-00321-6]
3. Dawson AF, Brown WW, Anderson J, et al. Mindfulness-based interventions for university students: a systematic review and meta-analysis of randomised controlled trials. Appl Psychol Health Well Being 2020 Jul;12(2):384-410. [doi: 10.1111/aphw.12188] [Medline: 31743957]
4. Hanson P. Mindfulness in undergraduate medical education: where we are now and where we go next. J Med Educ Res 2023;3(1):43-46. [doi: 10.5750/jmer.v3i1.2114]
5. El Morr C, Maule C, Ashfaq I, Ritvo P, Ahmad F. Design of a mindfulness virtual community: a focus-group analysis. Health Informatics J 2020 Sep;26(3):1560-1576. [doi: 10.1177/1460458219884840] [Medline: 31709878]
6. Sun S, Lin D, Goldberg S, et al. A mindfulness-based mobile health (mHealth) intervention among psychologically distressed university students in quarantine during the COVID-19 pandemic: a randomized controlled trial. J Couns Psychol 2022 Mar;69(2):157-171. [doi: 10.1037/cou0000568] [Medline: 34264696]
7. Michie S, van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. Implement Sci 2011 Apr 23;6:42. [doi: 10.1186/1748-5908-6-42] [Medline: 21513547]
8. Bennett RJ, Bucks RS, Saulsman L, Pachana NA, Eikelboom RH, Meyer CJ. Use of the Behaviour Change Wheel to design an intervention to improve the provision of mental wellbeing support within the audiology setting. Implement Sci Commun 2023 May 2;4(1):46. [doi: 10.1186/s43058-023-00427-1] [Medline: 37131257]

XSL•FO
RenderX

9. Carney R, Bradshaw T, Yung AR. Physical health promotion for young people at ultra-high risk for psychosis: an application of the COM-B model and behaviour-change wheel. Int J Ment Health Nurs 2016 Dec;25(6):536-545. [doi: 10.1111/inm.12243] [Medline: 27432534]
10. Cane J, O'Connor D, Michie S. Validation of the Theoretical Domains Framework for use in behaviour change and implementation research. Implement Sci 2012 Apr 24;7:37. [doi: 10.1186/1748-5908-7-37] [Medline: 22530986]
11. Michie S, Johnston M, Abraham C, Lawton R, Parker D, Walker A. Making psychological theory useful for implementing evidence based practice: a consensus approach. Quality and Safety in Health Care 2005 Feb 1;14(1):26-33. [doi: 10.1136/qshc.2004.011155]
12. Michie S, Atkins L, West R. The Behaviour Change Wheel: A Guide to Designing Interventions: Silverback Publishing; 2014.
13. Carey RN, Connell LE, Johnston M, et al. Behavior change techniques and their mechanisms of action: a synthesis of links described in published intervention literature. Ann Behav Med 2019 Jul 17;53(8):693-707. [doi: 10.1093/abm/kay078] [Medline: 30304386]
14. Connell LE, Carey RN, de Bruin M, et al. Links between behavior change techniques and mechanisms of action: an expert consensus study. Ann Behav Med 2019 Jul 17;53(8):708-720. [doi: 10.1093/abm/kay082]
15. The Theory and Techniques Tool. Human Behaviour Change Project. 2019. URL: https://theoryandtechniquetool.humanbehaviourchange.org [accessed 2025-12-08]
16. Rendell T, Barnett J, Scott S, Wright D. Designing a theory and evidence informed pharmacogenomic testing service in community pharmacy in England. Res Social Adm Pharm 2022 Oct;18(10):3831-3838. [doi: 10.1016/j.sapharm.2022.04.002] [Medline: 35428579]
17. Whittenbury K, Kroll L, Dubicka B, Bull ER. Exploring barriers and facilitators for mental health professionals delivering behavioural activation to young people with depression: qualitative study using the Theoretical Domains Framework. BJPsych Open 2022 Feb 4;8(2):e38. [doi: 10.1192/bjo.2022.7] [Medline: 35118935]
18. Bernal G, Sáez‐Santiago E. Culturally centered psychosocial interventions. J Community Psychol 2006 Mar;34(2):121-132. [doi: 10.1002/jcop.20096]
19. Alrashdi DH, Alyafei AH, Alanazi SA, Meyer C, Gould RL. Cultural adaptations of third-wave psychotherapies in Gulf Cooperation Council countries: a systematic review. Transcult Psychiatry 2024 Apr;61(2):209-228. [doi: 10.1177/13634615241227691] [Medline: 38332485]
20. Akras N. The impact of mindfulness program in reducing the symptoms of psychasthenia among a sample of students from Al-Jouf University [Arabic]. Int J Res Educ Psychol 2016;4(2):383-416. [doi: 10.12785/IJREP/040206]
21. Al-Ghalib S, Salim A. A mindfulness-based intervention to enhance university student wellbeing in Saudi Arabia. Middle East J Posit Psychol 2018;4:142-157 [FREE Full text]
22. Bernal G, Bonilla J, Bellido C. Ecological validity and cultural sensitivity for outcome research: issues for the cultural adaptation and development of psychosocial treatments with Hispanics. J Abnorm Child Psychol 1995 Feb;23(1):67-82. [doi: 10.1007/BF01447045] [Medline: 7759675]
23. Castellanos R, Yildiz Spinel M, Phan V, Orengo-Aguayo R, Humphreys KL, Flory K. A systematic review and meta-analysis of cultural adaptations of mindfulness-based interventions for Hispanic populations. Mindfulness (N Y) 2020 Feb;11(2):317-332. [doi: 10.1007/s12671-019-01210-x]
24. Harper Shehadeh M, Heim E, Chowdhary N, Maercker A, Albanese E. Cultural adaptation of minimally guided interventions for common mental disorders: a systematic review and meta-analysis. JMIR Ment Health 2016;3(3):e44. [doi: 10.2196/mental.5776]
25. AlHadi AN, Alhuwaydi AM. The mental health impact of pandemic COVID-19 crisis on university students in Saudi Arabia and associated factors. J Am Coll Health 2023;71(6):1854-1862. [doi: 10.1080/07448481.2021.1947839] [Medline: 34437818]
26. Hamasha AH, Kareem YM, Alghamdi MS, Algarni MS, Alahedib KS, Alharbi FA. Risk indicators of depression among medical, dental, nursing, pharmacology, and other medical science students in Saudi Arabia. Int Rev Psychiatry 2019;31(7-8):646-652. [doi: 10.1080/09540261.2019.1584095] [Medline: 31117837]
27. Alghamdi NA, Dunn K, Cairns D, Melville CA. A review of policy analysis: gender equality in Saudi Arabia's mental health policy. Int Neuropsychiatr Dis J 2023;20(3):21-34. [doi: 10.9734/indj/2023/v20i3397]
28. Taylor H, Strauss C, Cavanagh K. Can a little bit of mindfulness do you good? A systematic review and meta-analyses of unguided mindfulness-based self-help interventions. Clin Psychol Rev 2021 Nov;89:102078. [doi: 10.1016/j.cpr.2021.102078] [Medline: 34537665]
29. Tong A, Sainsbury P, Craig J. Consolidated Criteria for Reporting Qualitative Research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
30. Boddy CR. Sample size for qualitative research. Qual Mark Res 2016 Sep 12;19(4):426-432. [doi: 10.1108/QMR-06-2016-0053]

XSL•FO
RenderX

31. Gale NK, Heath G, Cameron E, Rashid S, Redwood S. Using the framework method for the analysis of qualitative data in multi-disciplinary health research. BMC Med Res Methodol 2013 Sep 18;13:117. [doi: 10.1186/1471-2288-13-117] [Medline: 24047204]
32. Braun V, Clarke V. Thematic Analysis: A Practical Guide: Sage Publications; 2021.
33. Atkins L, Francis J, Islam R, et al. A guide to using the Theoretical Domains Framework of behaviour change to investigate implementation problems. Implement Sci 2017 Jun 21;12(1):77. [doi: 10.1186/s13012-017-0605-9] [Medline: 28637486]
34. Banerjee M, Cavanagh K, Strauss C. A qualitative study with healthcare staff exploring the facilitators and barriers to engaging in a self-help mindfulness-based intervention. Mindfulness (N Y) 2017;8(6):1653-1664. [doi: 10.1007/s12671-017-0740-z] [Medline: 29201248]
35. Alattar N, Felton A, Stickley T. Mental health and stigma in Saudi Arabia: a scoping review. Ment Health Rev J 2021 May 27;26(2):180-196. [doi: 10.1108/MHRJ-08-2020-0055]
36. Alhumaidan NI, Alotaibi TA, Aloufi KS, et al. Barriers to seeking mental health help in Saudi Arabia: a systematic review. Cureus 2024 May;16(5):e60363. [doi: 10.7759/cureus.60363] [Medline: 38883041]
37. Alissa NA. Social barriers as a challenge in seeking mental health among Saudi Arabians. J Educ Health Promot 2021;10(1):143. [doi: 10.4103/jehp.jehp_819_20] [Medline: 34222518]
38. Alsalman Z, Shafey MM, Al-Khofi A, et al. Barriers to mental health service utilisation among medical students in Saudi Arabia. Front Public Health 2024;12:1371628. [doi: 10.3389/fpubh.2024.1371628] [Medline: 38680929]
39. Mahmoud MA. Knowledge and awareness regarding mental health and barriers to seeking psychiatric consultation in Saudi Arabia. Asian J Pharm Res Health Care 2019 Mar 4;10(4):109-116. [doi: 10.18311/ajprhc/2018/23359]
40. Noorwali R, Almotairy S, Akhder R, et al. Barriers and facilitators to mental health help-seeking among young adults in Saudi Arabia: a qualitative study. Int J Environ Res Public Health 2022 Mar 1;19(5):2848. [doi: 10.3390/ijerph19052848] [Medline: 35270539]
41. Bamber MD, Schneider JK. College students' perceptions of mindfulness-based interventions: a narrative review of the qualitative research. Curr Psychol 2022 Feb;41(2):667-680. [doi: 10.1007/s12144-019-00592-4]
42. Dark-Freudeman A, Jones C, Terry C. Mindfulness, anxiety, and perceived stress in university students: comparing a mindfulness-based intervention (MBI) against active and traditional control conditions. J Am Coll Health 2022 Oct;70(7):2116-2125. [doi: 10.1080/07448481.2020.1845180] [Medline: 33400631]
43. Galante J, Dufour G, Vainre M, et al. A mindfulness-based intervention to increase resilience to stress in university students (the Mindful Student Study): a pragmatic randomised controlled trial. Lancet Public Health 2018 Feb;3(2):e72-e81. [doi: 10.1016/S2468-2667(17)30231-1] [Medline: 29422189]
44. Marks E, Moghaddam N, De Boos D, Malins S. A systematic review of the barriers and facilitators to adherence to mindfulness-based cognitive therapy for those with chronic conditions. Br J Health Psychol 2023 May;28(2):338-365. [doi: 10.1111/bjhp.12628] [Medline: 36310447]
45. Howarth A, Smith JG, Perkins-Porras L, Ussher M. Effects of brief mindfulness-based interventions on health-related outcomes: a systematic review. Mindfulness (N Y) 2019 Oct;10(10):1957-1968. [doi: 10.1007/s12671-019-01163-1]
46. Thomas J, Raynor M, Bakker MC. Mindfulness-based stress reduction among Emirati Muslim women. Ment Health Relig Cult 2016 Mar 15;19(3):295-304. [doi: 10.1080/13674676.2016.1168389]
47. Listiyandini RA, Andriani A, Kusristanti C, Moulds M, Mahoney A, Newby JM. Culturally adapting an internet-delivered mindfulness intervention for Indonesian university students experiencing psychological distress: mixed methods study. JMIR Form Res 2023 Aug 31;7:e47126. [doi: 10.2196/47126] [Medline: 37651168]
48. Listiyandini RA, Andriani A, Afsari N, et al. Evaluating the feasibility of a guided culturally adapted internet-delivered mindfulness intervention for Indonesian university students experiencing psychological distress. Mindfulness (N Y) 2024 May;15(5):1095-1108. [doi: 10.1007/s12671-024-02346-1]
49. Watson-Singleton NN, Black AR, Spivey BN. Recommendations for a culturally-responsive mindfulness-based intervention for African Americans. Complement Ther Clin Pract 2019 Feb;34:132-138. [doi: 10.1016/j.ctcp.2018.11.013] [Medline: 30712717]
50. Population estimates publication 2024. General Authority for Statistics, Kingdom of Saudi Arabia. URL: https://www.stats.gov.sa/documents/20117/2435273/Population+Estimates+Statistics+2024+EN.pdf/9b71e303-5fd9-19cb-9913-850a9d521639 [accessed 2025-12-08]
51. Whittal A, Atkins L, Herber OR. What the guide does not tell you: reflections on and lessons learned from applying the COM-B behavior model for designing real life interventions. Transl Behav Med 2021 May 25;11(5):1122-1126. [doi: 10.1093/tbm/ibaa116] [Medline: 33200792]
52. Cabassa LJ, Baumann AA. A two-way street: bridging implementation science and cultural adaptations of mental health treatments. Implement Sci 2013 Aug 19;8(1):90. [doi: 10.1186/1748-5908-8-90] [Medline: 23958445]
53. Duffy A, Saunders KEA, Malhi GS, et al. Mental health care for university students: a way forward? Lancet Psychiatry 2019 Nov;6(11):885-887. [doi: 10.1016/S2215-0366(19)30275-5] [Medline: 31324561]
54. Sheldon E, Simmonds-Buckley M, Bone C, et al. Prevalence and risk factors for mental health problems in university undergraduate students: a systematic review with meta-analysis. J Affect Disord 2021 May 15;287:282-292. [doi: 10.1016/j.jad.2021.03.054] [Medline: 33812241]

XSL•FO
RenderX

55. Vision realization program: health sector transformation program. Saudi Vision 2030. URL: https://www.vision2030.gov.sa/en/explore/programs/health-sector-transformation-program [accessed 2025-12-08]
56. Skivington K, Matthews L, Simpson SA, et al. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. BMJ 2021 Sep 30;374:n2061. [doi: 10.1136/bmj.n2061] [Medline: 34593508]

Abbreviations

APEASE: acceptability, practicability, effectiveness/cost-effectiveness, affordability, side effects/safety, and equity criteria
BCT: behavior change technique
COM-B: Capability, Opportunity, and Motivation Domains of Behavior Change
COREQ: Consolidated Criteria for Reporting Qualitative Research
KSA: Kingdom of Saudi Arabia
MBI: mindfulness-based intervention
RCT: randomized controlled trial
TDF: Theoretical Domains Framework

Edited by J Sarvestan; submitted 04.06.25; peer-reviewed by KE Alaoui, RA Listiyandini, S Parkin; revised version received 07.11.25; accepted 26.11.25; published 19.12.25.

Please cite as:
Alrashdi DH, Meyer C, Gould RL
Facilitators, Barriers, and Cultural Appropriateness of Mindfulness-Based Interventions Among Saudi Female University Students: Qualitative Study
JMIR Form Res 2025;9:e78532
URL: https://formative.jmir.org/2025/1/e78532
doi:10.2196/78532

© Duaa H Alrashdi, Carly Meyer, Rebecca L Gould. Originally published in JMIR Formative Research (https://formative.jmir.org), 19.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Comparing ChatGPT and DeepSeek for Assessment of Multiple-Choice Questions in Orthopedic Medical Education: Cross-Sectional Study

Chirathit Anusitviwat[1], MD; Sitthiphong Suwannaphisit[2], MD; Jongdee Bvonpanttarananon[1]; Boonsin Tangtrakulwanich[1], MD, PhD

[1]Department of Orthopedics, Faculty of Medicine, Prince of Songkla University, 15 Karnchanavanich Road, Hat Yai, Thailand
[2]Department of Orthopaedics, Faculty of Medicine, Vajira Hospital, Navamindradhiraj University, Bangkok, Thailand

Corresponding Author:
Boonsin Tangtrakulwanich, MD, PhD
Department of Orthopedics, Faculty of Medicine, Prince of Songkla University, 15 Karnchanavanich Road, Hat Yai, Thailand

Abstract

Background: Multiple-choice questions (MCQs) are essential in medical education for assessing knowledge and clinical reasoning. Traditional MCQ development involves expert reviews and revisions, which can be time-consuming and subject to bias. Large language models (LLMs) have emerged as potential tools for evaluating MCQ accuracy and efficiency. However, direct comparisons of these models in orthopedic MCQ assessments are limited.

Objective: This study compared the performance of ChatGPT and DeepSeek in terms of correctness, response time, and reliability when answering MCQs from an orthopedic examination for medical students.

Methods: This cross-sectional study included 209 orthopedic MCQs from summative assessments during the 2023 - 2024 academic year. ChatGPT (including the "Reason" function) and DeepSeek (including the "DeepThink" function) were used to identify the correct answers. Correctness and response times were recorded and compared using a χ^2 test and Mann-Whitney *U* test where appropriate. The two LLMs' reliability was assessed using the Cohen κ coefficient. The MCQs incorrectly answered by both models were reviewed by orthopedic faculty to identify ambiguities or content issues.

Results: ChatGPT achieved a correctness rate of 80.38% (168/209), while DeepSeek achieved 74.2% (155/209; *P*=.04). ChatGPT's Reason function also outperformed DeepSeek's DeepThink function (177/209, 84.7% vs 168/209, 80.4%; *P*=.12). The average response time for ChatGPT was 10.40 (SD 13.29) seconds, significantly shorter than DeepSeek's 34.42 (SD 25.48) seconds (*P*<.001). Regarding reliability, ChatGPT demonstrated an almost perfect agreement (κ=0.81), whereas DeepSeek showed substantial agreement (κ=0.78). A completely false response was recorded in 7.7% (16/209) of responses for both models.

Conclusions: ChatGPT outperformed DeepSeek in correctness and response time, demonstrating its efficiency in evaluating orthopedic MCQs. This high reliability suggests its potential for integration into medical assessments. However, our results indicate that some MCQs will require revisions by instructors to improve their clarity. Further studies are needed to evaluate the role of artificial intelligence in other disciplines and to validate other LLMs.

(JMIR Form Res 2025;9:e75607) doi:10.2196/75607

KEYWORDS

ChatGPT; large language model; LLM; orthopedic; multiple-choice question; MCQ

Introduction

Multiple-choice questions (MCQs) are a cornerstone of medical education, particularly for assessing knowledge acquisition and clinical reasoning in specialized fields such as orthopedic medicine. Traditionally, MCQs have been refined through expert reviews and iterative revisions, a process that can be time-consuming and prone to human bias [1,2]. MCQs are widely used in medical education because of their efficiency, objectivity, and ability to assess a broad range of knowledge in a standardized manner [2,3].

Large language models (LLMs), such as ChatGPT and DeepSeek, have emerged as potential tools for evaluating MCQ accuracy and efficiency. These artificial intelligence (AI)–driven approaches can systematically assess question clarity, difficulty, and alignment with learning objectives [4]. However, research comparing the performance of different LLMs in answering orthopedic MCQs remains limited.

Recent studies have highlighted the capabilities of LLMs in medical education [5-7]. These studies explored the use of GPT-4 for formative and summative assessments in medicine, demonstrating its potential in evaluating medical knowledge

XSL•FO
RenderX

and providing feedback to students. Additionally, ChatGPT has shown notable accuracy in medical education, particularly in answering MCQs for licensing exams [8]. Prior studies have shown its ability to perform well in both basic and clinical sciences, with accuracy rates ranging from 70% to 74% [9]. Despite its potential, the reliability of LLMs in MCQ assessment, particularly in disciplines requiring complex clinical reasoning, remains an area of ongoing investigation [8,9].

DeepSeek V3, a recent advancement in LLMs, has shown promise in various domains in health care. Research has explored its use in the medical context, including patient education on asthma, ocular oncology, and laparoscopic cholecystectomy [10-12]. However, no study has evaluated the performance of DeepSeek V3 as an assessment tool for the quality of MCQs in medical education.

This study compared the performance of ChatGPT and DeepSeek in answering orthopedic MCQs by evaluating their correctness, response time, and reliability. These findings provide insights into the feasibility of integrating LLMs into medical student assessments and highlight areas where MCQs may require refinement.

Methods

Study Design

This cross-sectional study was conducted at the Department of Orthopedics, Faculty of Medicine, Prince of Songkla University, Thailand. It focused on MCQs used in the summative assessment of orthopedic rotations during the 2023 - 2024 academic year. All MCQs from the question pool during this period were included. As this was an educational study, there were no criteria for participant withdrawal or study termination. Instead, all available MCQs from the study period were analyzed.

Assessment Tools

In this study, MCQs were defined as structured assessment tools consisting of a question stem and multiple answer choices with one correct response. This study evaluated the ability of two LLMs, ChatGPT and DeepSeek, to analyze and process these questions. The versions used were GPT-4o for ChatGPT and and DeepSeek-V3 for DeepSeek, as both represented the most stable and publicly accessible variants during the data collection period. The two LLMs were accessed between January 2025 and March 2025.

GPT-4o was developed by OpenAI and released in May 2024 [13]. This model was optimized for faster response times and improved performance while maintaining high accuracy in complex reasoning tasks. Additionally, ChatGPT includes a “Reason” function that allows users to request explanations for its responses, providing insights into the logic behind its choices. This feature enhances transparency and facilitates a deeper understanding of how the model arrived at its conclusions, making it particularly useful for educational and assessment purposes [14].

DeepSeek-V3, another advanced LLM used in medicine, was initially released in December 2024 [13]. It offers enhanced performance with improved accuracy, faster processing, and a broad, up-to-date knowledge base [15]. It includes a “DeepThink” function designed to enhance logical reasoning by breaking down complex problems step by step. The AI model is designed to deliver high-quality, reliable, and efficient assistance in various applications. It excels in natural language processing, providing superior contextual understanding and multilingual support, ensuring more natural and coherent interactions.

Data Collection and Evaluation Procedures

This study involved 3 main steps. First, a set of MCQs from orthopedic assessments conducted in 2023 - 2024 was collected and reviewed for quality. These MCQs were created by orthopedic faculty members and approved by the orthopedic committee to ensure content alignment with curricular objectives, and they have been used in orthopedic examinations for more than 5 years. Next, each MCQ was input into both ChatGPT and DeepSeek. The AI responses, along with the accuracy and consistency data, were recorded. The Reason and DeepThink functions were tested, and response times were documented. Finally, any MCQs for which both LLMs failed to identify the correct answer were reviewed by a panel of orthopedic faculty members. These questions were refined to improve clarity and adjusted to ensure an appropriate level of difficulty. To determine reliability, all the MCQs and their answer keys were input into both ChatGPT and DeepSeek by two orthopedic faculty members, who both had experience in medical education at an academic hospital, on separate days.

Data Analysis

Statistical methods were used to compare the performance of the two LLMs. Correctness rates were expressed as percentages and compared using the χ^2 test. Processing times were reported as means with SDs and analyzed using either a 2-sample, 2-tailed *t* test or a Mann-Whitney *U* test, depending on data distribution. The interrater reliability between the two LLMs was evaluated using the Cohen κ coefficient. Statistical significance was set at $P<.05$.

Ethical Considerations

This study was conducted in accordance with the ethical principles of the Declaration of Helsinki, and the protocol was reviewed and approved by the Institutional Review Board of the Faculty of Medicine at Prince of Songkla University (REC. 68-097-11-1). This study involved only secondary analysis of existing MCQs and contained no participant identifiers; thus, individual informed consent was waived. As this study did not involve direct participation by human subjects, no compensation was provided. Finally, no images or materials that could identify individual participants were included in the manuscript or supplementary files.

Results

A total of 209 orthopedic MCQs were analyzed across 11 subspecialty categories to compare the accuracy of ChatGPT and DeepSeek (Table 1) and the Reason and DeepThink functions (Table 2). Overall, ChatGPT achieved a significantly higher total accuracy of 80.4% (168/209) compared with 74.2%

XSL•FO
RenderX

(155/209) for DeepSeek (*P*=.04). Although no single subspecialty demonstrated a significant difference after stratification, ChatGPT showed numerically superior accuracy across most categories, with the largest gaps observed in the tumor and infection (92.9% vs 71.4%) and lower-limb injury (68.8% vs 50.0%) domains. These findings indicate that ChatGPT outperformed DeepSeek in overall accuracy for orthopedic MCQs.

In the second comparison, the Reason function in ChatGPT demonstrated an accuracy of 84.7% (177/209), while the DeepThink function in DeepSeek achieved 80.4% (168/209) accuracy. However, the difference was not statistically significant (*P*=.12). Nevertheless, the Reason function seems to have performed better in some domains, such as the pediatric and anomalous category (80.0% vs 65.0%) and the tumor and infection category (92.9% vs 78.6%), whereas DeepThink showed slightly higher accuracy in the complications in orthopedics category (90.9% vs 63.6%). However, these discrepancies did not appear to correlate with any specific subspecialty or content category.

Table . Comparison of accuracy between ChatGPT and DeepSeek across orthopedic multiple choice question (MCQ) categories (n=209).

MCQs in each category	ChatGPT responses, n (%)	DeepSeek responses, n (%)	*P* value
Upper limb injury (n=28)	23 (82.1)	21 (75.0)	.48
Upper limb disease (n=21)	21 (100.0)	20 (95.2)	.32
Pelvic and spine injury (n=19)	15 (79.0)	14 (73.7)	.56
Lower limb injury (n=16)	11 (68.8)	8 (50.0)	.18
Lower limb disease (n=12)	11 (91.7)	10 (83.3)	.32
Back and neck problems (n=26)	21 (80.8)	19 (73.1)	.41
Sprain and strain (n=8)	6 (75.0)	7 (87.5)	.32
Pediatric and anomalous (n=20)	12 (60.0)	12 (60.0)	>.99
Tumor and infection (n=14)	13 (92.9)	10 (71.4)	.08
Complications in orthopedics (n=11)	7 (63.6)	9 (81.8)	.16
Rehabilitation and physical therapy (n=34)	28 (82.4)	25 (73.5)	.26
Total (n=209)	168 (80.4)	155 (74.2)	.04

Table . Comparison of accuracy between “Reason” (ChatGPT) and “DeepThink” (DeepSpeek) functions across orthopedic multiple choice question (MCQ) categories (n=209).

MCQs in each category	Reason responses, n (%)	DeepThink responses, n (%)	*P* value
Upper limb injury (n=28)	25 (89.3)	25 (89.3)	>.99
Upper limb disease (n=21)	21 (100.0)	20 (95.2)	.32
Pelvic and spine injury (n=19)	16 (68.8)	14 (73.7)	.16
Lower limb injury (n=16)	11 (68.8)	9 (56.3)	.41
Lower limb disease (n=12)	11 (91.7)	10 (83.3)	.31
Back and neck problems (n=26)	22 (84.6)	22 (84.6)	>.99
Sprain and strain (n=8)	7 (87.5)	7 (87.5)	>.99
Pediatric and anomalous (n=20)	16 (80.0)	13 (65.0)	.18
Tumor and infection (n=14)	13 (92.9)	11 (78.6)	.16
Complications in orthopedics (n=11)	7 (63.6)	10 (90.9)	.08
Rehabilitation and physical therapy (n=34)	28 (82.4)	27 (79.4)	.56
Total (n=209)	177 (84.7)	168 (80.4)	.12

The most incorrect responses for both ChatGPT and DeepSeek occurred in the complications (2/11, 18.2%), pediatric and anomalous (3/20, 15.0%), and pelvic and spine injury (3/19, 15.8%) categories, reflecting areas where clinical reasoning and management knowledge were more heavily tested. The two lowest rates of mistakes were observed in the tumor and infection and upper limb injury categories. The overall rate of completely false responses from the two LLMs was 7.7% (16/209). Examples of questions that resulted in false responses from both LLMs are provided in Multimedia Appendix 1.

XSL•FO
RenderX

ChatGPT demonstrated a significantly faster response time, with an average processing time of 10.40 (SD 13.29) seconds, compared to DeepSeek's 34.42 (SD 25.48) seconds ($P<.001$).

Finally, the interrater reliability for the two LLMs was high. ChatGPT achieved a Cohen κ value of 0.81, indicating good agreement, whereas DeepSeek showed substantial agreement, with a κ of 0.78. These findings underscore both the consistency of the LLM-generated responses and a residual variability that warrants continued human oversight in educational and assessment settings.

Discussion

Principal Results

This study compared the performance of two advanced LLMs, ChatGPT and DeepSeek, in the processing of orthopedic MCQs regarding their correctness, response time, and reliability. Our findings indicate that ChatGPT outperformed DeepSeek in terms of both accuracy and response time. Additionally, the Reason function of ChatGPT, which provides explanations for selected answers, appears to contribute to its superior performance. Both ChatGPT and DeepSeek exhibited high interrater reliability when generating responses to orthopedic MCQs.

Comparison With Previous Studies

The results of this study are consistent with a prior study showing that advanced LLMs demonstrated a high level of accuracy in answering orthopedics-related MCQs [16]. In our study, the accuracy of ChatGPT without the Reason function was 80.4%, while a previous study reported an accuracy rate of 70.60%. When using the Reason function in ChatGPT, the accuracy increased to 84.7%. This may imply that the Reason function can promote deeper comprehension and facilitate critical thinking in the medical education setting.

The high accuracy observed for ChatGPT and DeepSeek is particularly valuable in high-stakes environments like medical education [17]. Efficient processing can facilitate the development of high-quality, comprehensive MCQs, minimizing ambiguities in both questions and answer choices to enhance assessment reliability. Furthermore, the ability of ChatGPT's Reason function to provide detailed explanations adds an educational dimension often missing in traditional assessment tools, thereby facilitating a deeper understanding of complex clinical contexts that mirror real-world practice [18].

The balance between AI efficiency and educational values poses an issue. Although ChatGPT's speed and accuracy are advantageous, reliance on automated systems for assessment must be balanced with human oversight to ensure that the nuances of clinical reasoning and decision-making are adequately taken into account. This is particularly important in fields such as orthopedic surgery, where clinical decisions can have significant implications for patient care [19].

We found that 7.7% of the MCQs yielded false responses from both LLMs, warranting further examination. This finding suggests that there may be inherent issues in question construction, leading to ambiguity and misinterpretation. Poorly constructed questions can stem from several factors, including unclear wording, imprecise phrasing of the question stem, or answer choices that do not effectively discriminate between correct and incorrect responses [20,21]. By revising ambiguous or poorly constructed questions, educators can improve the overall reliability and validity of assessments, subsequently enhancing the fairness of the evaluation process [22].

Both ChatGPT and DeepSeek exhibited high interrater reliability when generating responses to orthopedic MCQs. These results suggest that both AI models can consistently answer the standardized questions, although some variability remains. The differences in the κ values may have resulted from variations in the training data and model algorithms. Despite their strong reliability, the models did not achieve complete consistency, highlighting the importance of human oversight.

Although AI tools provide notable benefits in terms of efficiency and consistency, there are critical risks associated with overreliance on them for medical education and assessment. Therefore, human supervision is necessary to guarantee that AI functions as a helpful tool rather than a substitute for professional judgment, particularly in situations involving high stakes in education or health care. Educators should carefully combine AI with conventional teaching techniques and evaluation procedures to maintain academic integrity and guarantee that students keep refining their decision-making abilities.

Limitations

Despite the promising findings, this study had some limitations. Our study focused solely on orthopedic MCQs, which may limit the generalizability of our results to other areas of medical education. Future studies should extend these analyses to include diverse medical disciplines to better understand the broader applicability of AI-assisted assessments. Additional limitations include the study's cross-sectional design, which only captured the LLMs' performance at a single point in time. Given that LLMs such as ChatGPT and DeepSeek are continually updated and improved, their performance may vary over time. Furthermore, while the LLMs provided explanations for selected answers, we did not formally assess the quality of these explanations. Implementing a structured evaluation of explanation quality in future studies could offer valuable educational perspectives.

Implications of the Findings

The findings suggest that AI tools like ChatGPT and DeepSeek can enhance the efficiency of medical assessments in orthopedic rotation. ChatGPT demonstrated higher accuracy and faster responses compared to DeepSeek. Furthermore, both LLMs could help identify poorly constructed MCQs. However, reevaluation by teachers or assessors remains essential to ensure that responses are clinically relevant and that results of summative assessments are accurate.

For broader application in other medical specialties, we propose the following practical steps. First, validated, specialty-specific MCQs should be gathered from the target discipline to ensure content relevance and quality. Second, these questions should be input into selected AI models and their responses systematically recorded and evaluated for accuracy. Third, to

XSL•FO
RenderX

ensure reliability, at least two faculty experts should independently assess the LLM-generated answers, with any discrepancies resolved through consensus review. Fourth, poorly performing or ambiguous MCQs should be identified and refined for clarity.

Currently, there are a large number of advanced LLMs. Future studies should systematically compare other AI tools to determine which models perform best in MCQ assessment. Additionally, future research should be conducted to explore whether AI assistance could be used for other types of assessments, including modified essay questions, objective structured clinical examinations, or key-feature questions.

Conclusions

In the study, ChatGPT outperformed DeepSeek in answering orthopedic MCQs in terms of overall accuracy and response time, although both models exhibited strong interrater reliability. Specifically, ChatGPT achieved a significantly higher accuracy rate compared with DeepSeek, and its Reason function further enhanced performance. These findings suggest that ChatGPT can serve as a valuable tool in medical education and assessment, particularly when using explanatory reasoning capabilities.

Despite these encouraging results, both AI models showed areas of weakness, particularly in complex domains such as complications, pediatrics and anomalies, and pelvic and spine injuries. These errors highlight the importance of maintaining human oversight when using LLMs for educational or evaluative purposes. Thus, the integration of AI tools should be viewed as complementary to, rather than a replacement for, expert clinical judgment and expertise.

Acknowledgments

We used the generative artificial intelligence (AI) tools ChatGPT (OpenAI) and DeepSeek-V3 (DeepSeek) to assist in multiple-choice question analysis for this study. All AI-generated outputs were reviewed and validated for reliability by the authors.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

All authors contributed to the conceptualization. CA and SS contributed equally to the data curation, formal analysis, investigation, methodology, writing (original draft), and writing (review and editing). JB contributed to data curation and resources. BT contributed to supervision and writing (review and editing). All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Examples of questions that resulted in false responses from both ChatGPT and DeepSeek.
[DOCX File, 46 KB - formative_v9i1e75607_app1.docx]

References

1. Al-Rukban MO. Guidelines for the construction of multiple choice questions tests. J Family Community Med 2006 Sep;13(3):125-133. [Medline: 23012132]
2. Olayemi E. Multiple choice questiones as a tool for assessment in medical education. Ann Biomed Sci 2013;12(1). [doi: 10.4314/abs.v12i1]
3. Parekh P, Bahadoor V. The utility of multiple-choice assessment in current medical education: a critical review. Cureus 2024 May;16(5):e59778. [doi: 10.7759/cureus.59778] [Medline: 38846235]
4. Kayaalp ME, Prill R, Sezgin EA, Cong T, Królikowska A, Hirschmann MT. DeepSeek versus ChatGPT: multimodal artificial intelligence revolutionizing scientific discovery. From language editing to autonomous content generation-redefining innovation in research and practice. Knee Surg Sports Traumatol Arthrosc 2025 May;33(5):1553-1556. [doi: 10.1002/ksa.12628] [Medline: 39936363]
5. Mehta S, Haddad EN, Burke IB, et al. Assessment of large language model performance on medical school essay-style concept appraisal questions: exploratory study. JMIR Med Educ 2025 Jun 16;11:e72034. [doi: 10.2196/72034] [Medline: 40523238]

XSL•FO
RenderX

6. Krumsvik RJ. GPT-4's capabilities for formative and summative assessments in Norwegian medicine exams-an intrinsic case study in the early phase of intervention. Front Med (Lausanne) 2025;12:1441747. [doi: 10.3389/fmed.2025.1441747] [Medline: 40276737]
7. Kondo T, Nishigori H. AI's accuracy in extracting learning experiences from clinical practice logs: observational study. JMIR Med Educ 2025 Oct 15;11:e68697. [doi: 10.2196/68697] [Medline: 41092407]
8. Gonsalves C. On ChatGPT: what promise remains for multiple choice assessment? J Learn Dev High Educ 2023 Apr 27(27). [doi: 10.47408/jldhe.vi27.1009]
9. Meo SA, Al-Masri AA, Alotaibi M, Meo MZS, Meo MOS. ChatGPT knowledge evaluation in basic and clinical medical sciences: multiple choice question examination-based performance. Healthcare (Basel) 2023 Jul 17;11(14):2046. [doi: 10.3390/healthcare11142046] [Medline: 37510487]
10. Liu Y, Yu F, Zhang X, et al. Assessing the role of large language models between ChatGPT and DeepSeek in asthma education for bilingual individuals: comparative study. JMIR Med Inform 2025 Aug 13;13:e65365. [doi: 10.2196/65365] [Medline: 40802989]
11. Das D, Narayan A, Mishra V, et al. AI chatbots in answering questions related to ocular oncology: a comparative study between DeepSeek v3, ChatGPT-4o, and Gemini 2.0. Cureus 2025 Aug;17(8):e90773. [doi: 10.7759/cureus.90773] [Medline: 40988843]
12. Dincer HA, Dogu D. Evaluating artificial intelligence in patient education: DeepSeek-V3 versus ChatGPT-4o in answering common questions on laparoscopic cholecystectomy. ANZ J Surg 2025 Nov;95(11):2322-2328. [doi: 10.1111/ans.70198] [Medline: 40495650]
13. Xiong L, Wang H, Chen X, et al. DeepSeek: paradigm shifts and technical evolution in large AI models. IEEE/CAA J Autom Sinica 2025;12(5):841-858. [doi: 10.1109/JAS.2025.125495]
14. Khan RA, Jawaid M, Khan AR, Sajjad M. ChatGPT - Reshaping medical education and clinical management. Pak J Med Sci 2023;39(2):605-607. [doi: 10.12669/pjms.39.2.7653] [Medline: 36950398]
15. Temsah A, Alhasan K, Altamimi I, et al. DeepSeek in healthcare: revealing opportunities and steering challenges of a new open-source artificial intelligence frontier. Cureus 2025 Feb;17(2):e79221. [doi: 10.7759/cureus.79221] [Medline: 39974299]
16. Gan W, Ouyang J, Li H, et al. Integrating ChatGPT in orthopedic education for medical undergraduates: randomized controlled trial. J Med Internet Res 2024 Aug 20;26:e57037. [doi: 10.2196/57037] [Medline: 39163598]
17. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
18. Al Shuraiqi S, Aal Abdulsalam A, Masters K, Zidoum H, AlZaabi A. Automatic generation of medical case-based multiple-choice questions (MCQs): a review of methodologies, applications, evaluation, and future directions. BDCC 2024 Oct;8(10):139. [doi: 10.3390/bdcc8100139]
19. Crebbin W, Beasley SW, Watters DAK. Clinical decision making: how surgeons do it. ANZ J Surg 2013 Jun;83(6):422-428. [doi: 10.1111/ans.12180] [Medline: 23638720]
20. Al-Faris EA, Alorainy IA, Abdel-Hameed AA, Al-Rukban MO. A practical discussion to avoid common pitfalls when constructing multiple choice questions items. J Family Community Med 2010 May;17(2):96-102. [doi: 10.4103/1319-1683.71992] [Medline: 21359033]
21. Przymuszała P, Piotrowska K, Lipski D, Marciniak R, Cerbin-Koczorowska M. Guidelines on writing multiple choice questions: a well-received and effective faculty development intervention. Sage Open 2020 Jul;10(3):2158244020947432. [doi: 10.1177/2158244020947432]
22. Downing SM. Threats to the validity of clinical teaching assessments: what about rater error? Med Educ 2005 Apr;39(4):353-355. [doi: 10.1111/j.1365-2929.2005.02138.x] [Medline: 15813756]

Abbreviations

AI: artificial intelligence
LLM: large language model
MCQ: multiple-choice question

XSL•FO
RenderX

Edited by A Stone; submitted 07.04.25; peer-reviewed by E Démoulins, K Kuppan, LH Yao; revised version received 22.11.25; accepted 24.11.25; published 19.12.25.

Please cite as:
Anusitviwat C, Suwannaphisit S, Bvonpanttarananon J, Tangtrakulwanich B
Comparing ChatGPT and DeepSeek for Assessment of Multiple-Choice Questions in Orthopedic Medical Education: Cross-Sectional Study
JMIR Form Res 2025;9:e75607
URL: https://formative.jmir.org/2025/1/e75607
doi:10.2196/75607

© Chirathit Anusitviwat, Sitthiphong Suwannaphisit, Jongdee Bvonpanttarananon, Boonsin Tangtrakulwanich. Originally published in JMIR Formative Research (https://formative.jmir.org), 19.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Gamified Simulation for Onboarding Health Care Teams in Emergency Care: Development and Preliminary Feasibility Study

Stephane Gobron[1*], PhD; Antoine Lestrade[1*], MSc; Artan Sadiku[1*], BSc; Alexandre Bentvelzen[2*], MD; Leila Ait Kaci[3*], MSc; Jean-Michel Carrier[4*], MD; Emmanuelle Guyot[2*], MD; Pierre-Nicolas Carron[2,5*], MD

[1]Department of Computer Science, HE-Arc, School of Engineering, HES-SO University of Applied Sciences and Arts Western Switzerland, Espace de l'Europe 11, Neuchatel, Switzerland

[2]Emergency Department, Lausanne University Hospital (CHUV), Lausanne, Switzerland

[3]Digital Kingdom, Vevey, Switzerland

[4]Medical Education Unit, School of Medicine, University of Lausanne, Lausanne, Switzerland

[5]Faculty of Medicine, University of Lausanne, Lausanne, Switzerland

[*]all authors contributed equally

Corresponding Author:
Stephane Gobron, PhD
Department of Computer Science, HE-Arc, School of Engineering, HES-SO University of Applied Sciences and Arts Western Switzerland, Espace de l'Europe 11, Neuchatel, Switzerland

Abstract

Background: High staff turnover is a widespread issue across nearly all hospital departments, often exceeding 20% annually. This constant flux disrupts continuity of care and creates a recurring challenge: how to rapidly integrate new employees into complex clinical environments, both physically and functionally. Traditional onboarding methods struggle to meet this demand, particularly in services operating 24/7, such as emergency departments (EDs).

Objective: This formative study presents the design and implementation of a web-based 3D gamified simulation platform aimed at improving staff onboarding in clinical environments. The paper outlines both the technical architecture—with guidance for hospital IT departments—and the acceptability and usability for permanent staff, who play a key role in ensuring onboarding continuity. We sought to assess whether such a tool could be autonomously managed and well received by health care professionals.

Methods: The intervention consisted of 2 linked components: a real-time, browser-based 3D simulation replicating the hospital's ED and a web-based quest editor allowing nontechnical staff to update training content. The system supports self-paced onboarding through location-based tasks, object searches, quizzes, and simulated staff interactions. Two preliminary usability studies were conducted: one with 37 ED staff members testing the 3D simulation and another with 9 users exploring the quest editor. Feedback was gathered through anonymous questionnaires and a descriptive analysis.

Results: Early results showed high feasibility and acceptability. Among 3D simulation testers (n=37), 90% (33/37) found the tool helpful for understanding the department's structure, and 81% (30/37) believed it would be useful for new staff. The inclusion of personal anecdotes and gamified tasks was viewed as engaging and motivating. The quest editor (n=9) was positively rated by 91% (8/9) of users, who appreciated the ability to autonomously update content without IT support. These findings support the dual promise of the platform (ie, pedagogical flexibility and technical sustainability).

Conclusions: This work demonstrates the feasibility of a gamified simulation platform designed for high-turnover clinical environments. It highlights both the operational deployment framework and the early acceptability among key staff members. While further validation with actual new hires is needed, this formative study shows promising potential for generalization beyond emergency care. The modular and editable nature of the system makes it a viable solution for scalable onboarding in other hospital departments.

(JMIR Form Res 2025;9:e72202) doi:10.2196/72202

KEYWORDS

personnel turnover; digital learning environment; feasibility studies; user acceptability; staff development; self-directed training; health technology implementation; simulation training; mobile phone

XSL•FO
RenderX

Introduction

Context and Relevant Issues

High staff turnover is a persistent and well-documented issue in hospital settings, particularly in emergency departments (EDs), where annual turnover rates often exceed 20% [1,2]. This situation gives rise to several major challenges: reduced team cohesion, loss of institutional knowledge, increased demands on permanent staff to repeatedly onboard newcomers, and the rapid familiarization of newcomers with the premises, equipment, and colleagues. Despite the prevalence of this issue, existing onboarding strategies struggle to provide scalable, flexible, and engaging solutions, especially in services operating 24/7 with complex spatial and organizational structures.

Toward a New Paradigm in Onboarding

Staff turnover is a critical issue for most health departments [1,3,4], especially in EDs. These services operate 24/7, requiring personnel to be fully operational shortly after arrival. This constant demand increases pressure on onboarding processes and planning. High turnover in EDs stems from several factors, for example, intense workload, emotional fatigue, irregular hours, and the transitory nature of certain positions, which are linked to training periods or natural career progression. In addition, the ED at Lausanne University Hospital (CHUV) plays a leading role in clinical and postgraduate training for students, residents, and newly hired professionals. Many of these individuals are only present for a few months, for example, during postgraduate medical training. These short rotations contribute significantly to the department's high turnover and the need for frequent onboarding and adaptation—a context and issue illustrated in Figure 1.

Figure 1. Challenges in integrating new employees in the emergency department (ED). (A) This figure summarizes the essence of the problem: similar to any workplace, new employees arrive at the ED and spend a certain amount of time getting to know the place, the equipment, and their colleagues before becoming operational. The difference is that in health care services, turnover is high, material and premises are complex, and being operational is a matter of life and death in ED. (B) Implementation concept for an online employee onboarding tool: this 3D web app enables people to familiarize themselves with the premises and staff in a way that can bring properties sometimes even more effectively than in real life. For instance, to strengthen extraprofessional links, permanent staff allowed us to add some of their own personal anecdotes. We discovered that team members who had known each other for years deepened their bonds by discovering many personal aspects that they had never taken the time to share.

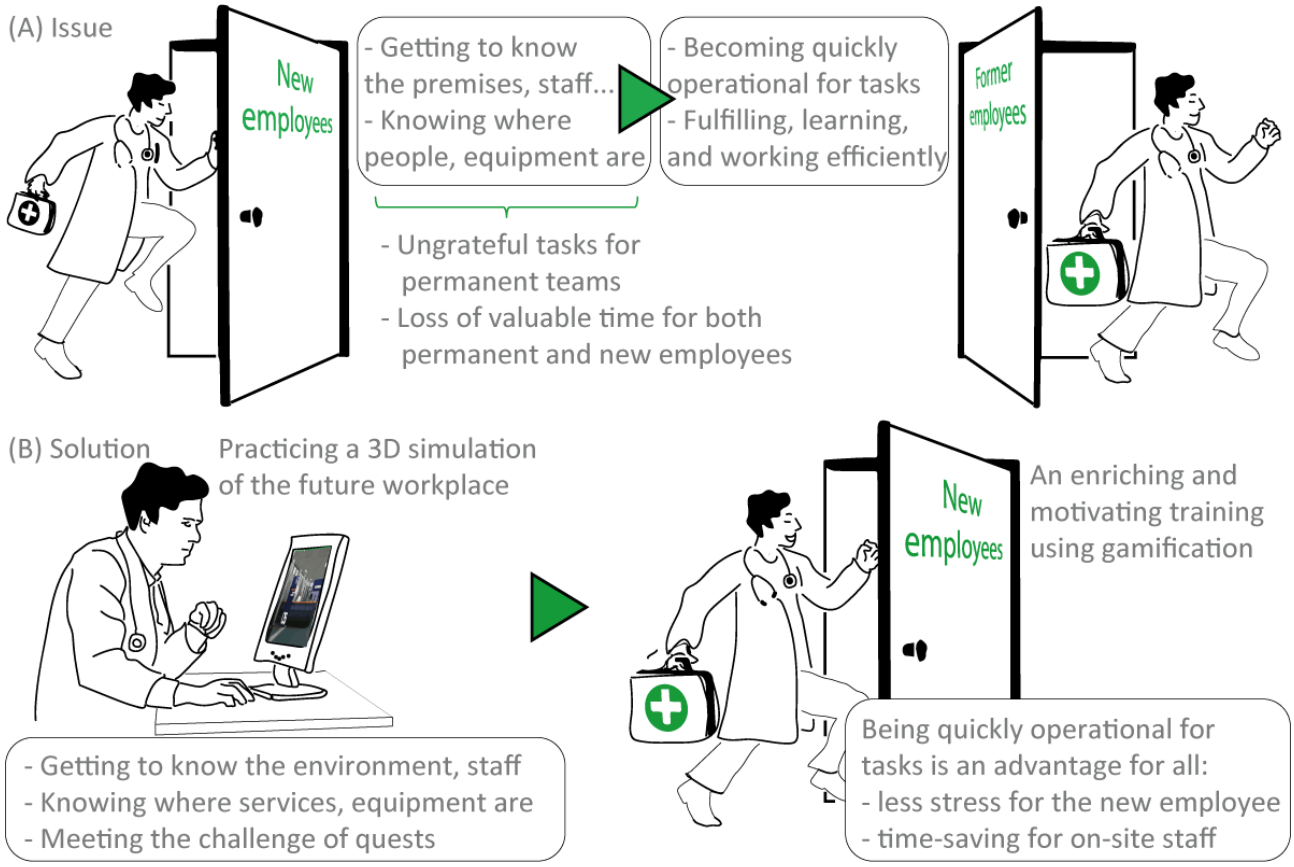

Unfortunately, in terms of education and training, the ED overcrowding does not always allow for suitable integration of these staff when they arrive in the department. The "permanent care" mission also complicates traditional face-to-face training, as some of the ED teams are on duty, whereas others are on recovery or night shifts. It is therefore important to find innovative solutions to improve team information, if possible, remotely and unplanned. The expectations of the new generation in terms of modern training also require new training modalities. Recent literature suggests that immersive and gamified environments may improve engagement, memory retention, and emotional well-being [5-7], yet few implementations offer both technical scalability and pedagogical adaptability.

The level of knowledge that must be quickly reached is important, justifying the necessity of finding solutions to

XSL•FO
RenderX

improve training. One of the difficulties for trainees and new employees is to quickly understand the ED architecture, functional organization, and the principles of patient prioritization. The organization of an ED plays a key role in its operation, and a deep understanding of it is essential for teamwork and efficiency. The architectural structure of the service, functional organization, patient triage, and flow management is highly codified in the ED and needs to be learned quickly. The department's missions involve knowing not only the different care areas, resuscitation rooms, and patient flow management from the waiting room but also the locations of resuscitation or medical equipment or even emergency exits and alarm systems (eg, fire and assault).

Problem Statement and Work Objective

Despite advances in health education technologies, there remains a significant gap in the availability of autonomous, editable, and web-accessible onboarding tools that can address the operational demands of high-turnover hospital units. Most existing solutions require costly technical maintenance, lack content flexibility, or fail to offer meaningful interactivity for new hires. This article explores a solution to simplify the issue of frequent integration of new staff in the ED context using an online application that prospective employees can practice before arriving on site (Figure 1B). This work presents the design, development, and formative evaluation of a dual-component solution: (1) a web-based 3D gamified simulation of the ED aimed at improving staff onboarding; and (2) a companion quest management application enabling permanent staff to autonomously update and adapt onboarding content without IT intervention.

Research Questions and Hypotheses

We posed the following research questions (RQs) to better understand the relevance and usability of this approach and derived 2 main hypotheses:

- RQ1: Can a browser-based gamified simulation improve perceived onboarding efficiency and engagement among ED staff?
- RQ2: Can permanent staff members, with minimal training, autonomously maintain and update onboarding content using a dedicated quest management tool?
- Hypothesis 1: The gamified simulation will be perceived as useful, engaging, and appropriate for onboarding by most health care professionals, as measured by usability and acceptability metrics.
- Hypothesis 2: The companion quest management tool will be usable and effective for nontechnical ED staff, enabling sustainable onboarding content maintenance.

Gamified Simulations

Simulations in health care are essential to improve the skills of health care professionals [8]. They allow the replication of varied and complex clinical situations, fostering practical and secure learning. Gamifying these simulations makes the training more engaging and motivating [9]. By incorporating game elements such as points, levels, and rewards, participant involvement is increased, which enhances knowledge retention and skill mastery. Thus, the purpose of gamified simulations is to optimize training by making it more interactive and stimulating. In the field of health care, recent research includes a large set of aspects, such as simulation, 3D environments, gamification, serious gaming, patient sorting, urban reconstruction, and scripting [10].

In the context of urban interior and 3D visualization, Joy and Christhu [11] proposed a digital 3D modeling system for preconstruction. This study is important because it establishes a framework that explores the complexities of 3D architectural visualization in a real-time engine, which we have achieved. The concept was to enable a visual experience with a higher level of user interaction and then deploy it through virtual reality (VR). In our case, we did not choose to deploy the simulation in VR, as we wanted any future employee to be able to use it with a simple web browser. In the same domain, Chen and Feng [12] provided an original procedure to autogenerate buildings using cellular automata. Focusing more on the hospital, De Oliveira [13] proposed a platform composed of a set of tools that provides a patient-oriented modeling process, enabling the running of a simulation experiment showing possible outcomes in a 3D virtual scenario in a hospital environment, which is close to the problem raised in our proposal. In 2000, the same authors also presented a 3D visual simulation of hospital admissions [14]. Staying within the hospital field, 15 years later, de Lim et al [15] described a virtual environment for the learning process of simulated clinical cases for students and professors.

Concerning simulation, emergency, and gamification-related approaches, Gobron et al [16] proposed in 2019 a picture-based serious game to train nonmedical people in emergency situations. This simulation also included training to answer emergency calls. Yang et al [17] also provided an interesting review of the literature relative to crowd evacuation, which is a critical issue when simulating a public environment. A little further away, Shapiro et al [18] determined how simulation-based clinical team training improves performance when added to a didactic program, demonstrating that a gamified simulation can be efficient. A study by Zielke et al [19] explored the potential of VR and augmented reality for the development of a virtual patient in a communication practice—ie, managing patient emotions.

Recent (2021~) works demonstrate the actual benefits of simulation in medical education: Krishnamurthy et al [5] provided a systematic review of what is meant by gamification in the health field; Gue et al [20] surveyed residents to investigate the impact of gamification in emergency medicine residency training, and Van Gaalen et al [21] proposed a review focusing on empirical evidence for the effectiveness of gamification approaches and the theoretical justification for the application of the chosen game attributes. In the field of gamified applications and employee training, Larson [22] proposed a literature review on serious games and gamification in the corporate training environment. The latest noteworthy publication is probably the work of Stoicu-Tivadar et al [23] presenting a virtual hospital architecture called eduCRATE.

Recent narrative reviews have highlighted the increasing use of gamification across various branches of medical education, including dermatology, where gamified tools have been used

XSL•FO
RenderX

in both student training and patient engagement [24]. This broader trend reinforces the legitimacy and adaptability of gamification mechanisms in health care and supports their relevance as a pedagogical and communication strategy in clinical contexts.

Onboarding Strategies

Several institutions have implemented structured onboarding initiatives to mitigate the impact of staff turnover, particularly among newly hired health care professionals. For instance, a 187-bed community hospital in Washington, DC, deployed a 10-element onboarding improvement program based on staffing statistics, exit surveys, and best practices identified in the literature. The intervention notably included reinforced managerial interactions and peer mentoring. The results were significant: new-hire turnover decreased from 39.1% to 18.4%, and overall hospital turnover dropped from 18.2% to 11.9% ($Z=-2.06$; $P=.04$) [4]. These findings underscore the value of standardized communication-focused onboarding processes supported by cross-departmental collaboration and executive commitment.

Health care professionals face constant emotional strain and demanding work conditions, generating a growing need for integrated emotional relief mechanisms within their professional routines. Recent studies have highlighted the dual function of gamification—not only as an engagement tool but also as an emotional outlet. Johnson et al [6] demonstrated that game-based elements can enhance motivation while positively influencing well-being. More recently, in 2024, Tu and Lee [7] reported a significant reduction in stress levels among shift-work nurses using a gamified intervention platform. The same year, Lehtoranta et al [25] similarly emphasized the positive impact of gamification on employee satisfaction and mental health. These findings suggest that, in high-pressure clinical environments, the inclusion of a simulation that blends serious objectives with playful dynamics can support both training effectiveness and psychological resilience. As such, gamified systems offer a promising response to the emotional and cognitive demands experienced by health care staff in daily practice.

In terms of 3D simulations applied to health, Rash et al [26] proposed in 2024 a prototype of a 3D virtual simulation emergency room environment for interprofessional team training. Unfortunately, this prototype does not include any quests or motivating interaction for field teams, nor does it allow staff to modify the scenarios. Moreover, only oral feedback from a dozen participants was presented. By far, the closest publication to our approach is that of Shi et al [27], proposing the simulation of part of a hospital department. The similarities with the current proposal include a 3D simulation with quests. What we found especially interesting and positive in this approach were several features: nonplayer characters (NPCs; also called agents) can move; however, it is not specified how; player inventory system is a strong plus; fallback toward a local database if online connection with the cloud is off; and their minimap has the possibility to add markers, which can be useful when discovering a new place. Still, unfortunately, the fundamental weaknesses (that our proposal solves) are as follows: Their app runs locally on PC; it cannot have the flexibility of a web-based app, and current graphics texture is heavy (texture), which strongly restricts the portability of online deployment; user to agent/NPC interactions only consist of dialog (eg, no question/answer); no comparative visualization with the real views of the hospital; no login user/password access, which implies security weaknesses; lack of data validation; new quests only are added using a Microsoft Excel local file; and adding assets (eg, new permanent staff members) is not possible.

Methods

Overview

This section presents the objectives, application deployment, key aspects of system design, and software architecture. While inherently technical, these details are crucial for ensuring effective collaboration between medical and IT teams, a challenge that has led to the failure of many similar projects. By outlining the underlying architecture, development choices, and integration mechanisms, we aim to provide both hospital medical staff and IT departments with the necessary insights to deploy, maintain, and adapt similar solutions to evolving needs.

Objectives and Design Process

A hospital and engineering consortium was set up to take up the challenge of developing an application that would solve—at least in part—the problem of rapid integration in an ED context. In collaboration with the CHUV ED and the University of Lausanne, a digital innovative solution was developed to ease the integration of new employees. The design and development of the simulation platform took place over a period of 7 years, with a notable interruption due to the COVID-19 pandemic. Throughout this process, we adopted an iterative and collaborative design framework, working closely (every month) with the ED's medical staff. This continuous involvement ensured that the evolving needs, constraints, and pedagogical expectations of the department were consistently addressed. Owing to the inherent nature of the project, the actual target group—that is, newly arriving staff—was, by definition, not accessible for testing during development. However, the permanent team provided valuable feedback from the perspective of onboarding processes and acted as proxy users, offering realistic input on usability, content relevance, and scenario design. This sustained collaboration between technical developers and medical professionals laid the foundation for a solution tailored to the real operational dynamics of the ED, thus supporting its long-term sustainability and practical adoption.

The digital solution was based on an interactive 3D application (illustrated in Figure 2), which provides future staff members with virtual immersion in the hospital premises, as well as interaction with simulated staff and patients. To achieve the solution proposed in this paper, we had certain basic criteria and others that were discovered as we went along:

1. Challenges: to ensure that the locations are realistic (geometrically, spatially, realistic camera movements and angle of view, colors and textures of floors, ceilings, walls,

XSL•FO
RenderX

furniture, posters and main instructions, and medical equipment); to provide a solution that is easy to use in terms of the interface and adapted to the user's profile; and to supply serious, rewarding, and fun encounters with the NPCs.

2. Constraints: to develop a lightweight 3D application for a convenient online access, to ensure security with user login and password, and to enable easy-to-edit quests without the need for IT or 3D specialists.

Figure 2. Illustration of gamified simulation windows. To reach the 3D simulation, follow the green arrow steps; 5 other windows are for the user management of the navigation and the quests' follow-up (white arrows highlight these options).

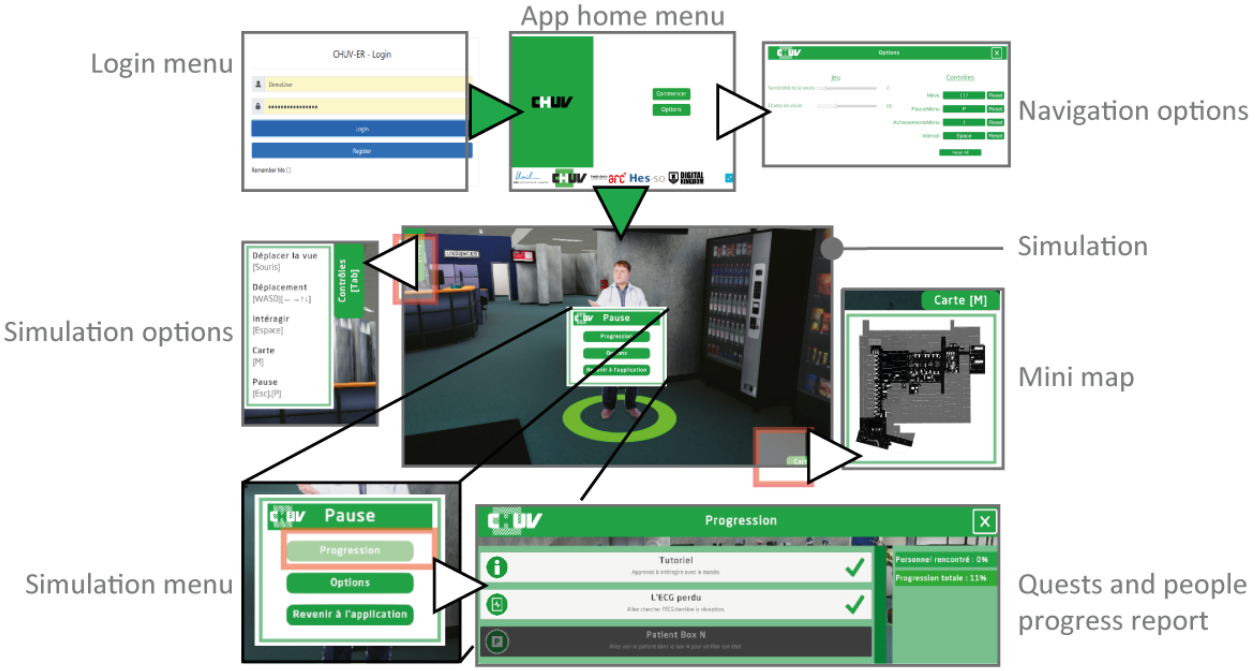

The key element of this application lies in its gamified approach, which transforms integration into a fun and engaging experience. Through a variety of quests, such as discovering locations, meeting virtual colleagues, or solving specific tasks, new employees are guided in their learning while being encouraged to actively explore their environment. To ensure the longevity and flexibility of this solution, a second application was developed for permanent staff (see properties in Figure 3). This application enables in-house teams to modify and add quests, rewards, and other simulation elements without requiring the constant intervention of the technical team. This approach ensures a rapid adaptation to the frequent changes observed in hospital services while preserving the quality and consistency of the induction experience for new employees.

Two applications were developed: one for new employees and one for the permanent staff to manage the quest system (Figure 3), without having to go through the IT department or a costly external service provider. The current application provides an unlimited number of quests, where each of them can also have implicit dependencies and subquests. Within each quest, five types of tasks are defined: (1) go to an area—mainly to discover the relatively complex topology of the ED; (2) dialogs from basic virtual agents—mainly associated with personal anecdotes; (3) basic conversation with virtual agents—usually part of a quest; (4) search for an object or a person, for example, to understand professional functions and/or duties; and (5) getting a multiple-choice question (2-4 options) and answering it, for example, for patient sorting.

XSL•FO
RenderX

Figure 3. Use of the application to manage quests. Medical materials or any object as well as new characters can be added via images as "billboards" (ie, PNG images with transparency). The scale of these images is managed by the following ratio: 500 pixels for 1 m. Once these assets are uploaded, the user can define a new quest assigning various objects and virtual agents. An optional personified trophy can also be associated for each quest—if not, a default picture will be assigned. Each quest includes dialogs, questions, places to go, and objects to find, which implies the use of map coordinates as shown in the figure. NCP: nonplayer character.

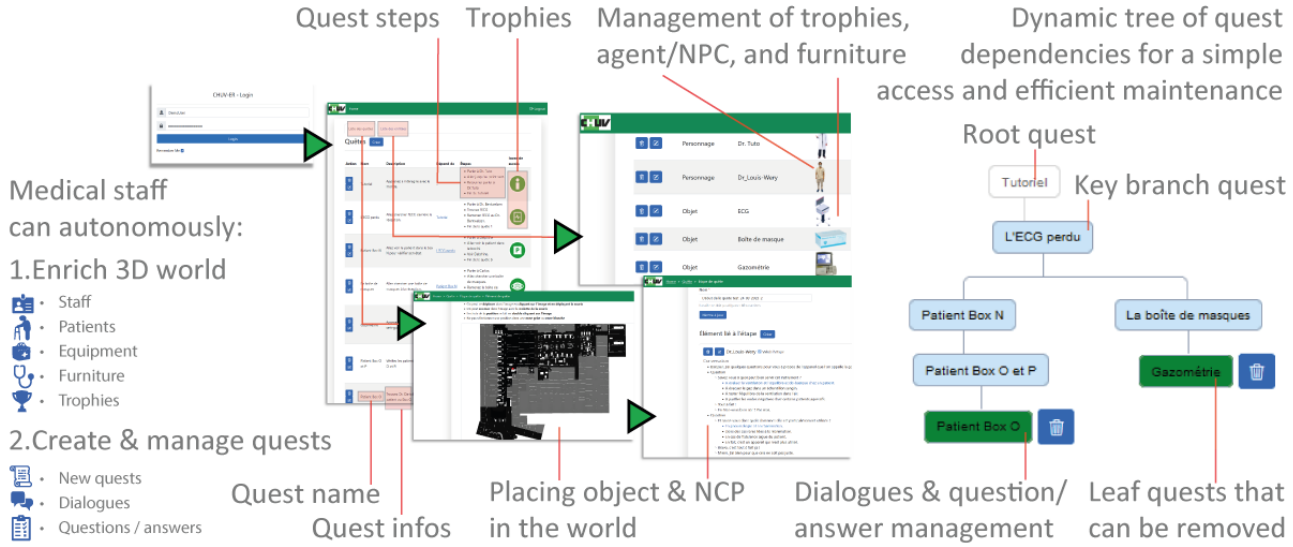

3D Application for New Employees

The end user application is a 3D interactive simulation that models the ED. This application was designed using Unity Engine version 2021.3 and is based on the Universal Render Pipeline. Thanks to this engine, a multiplatform deployment was possible in a web environment. Deploying a real 3D application as a web page (ie, using WebGL) offers unparalleled accessibility and scalability. Unlike traditional software that requires installation, a WebGL-based solution runs directly in the browser, making it independent of both hardware and operating system. This ensures seamless deployment on PCs, tablets, and even smartphones, without compatibility concerns. In addition, updates are centralized, eliminating the need for manual installations while providing real-time access to the latest version, a crucial feature for dynamic hospital environments. This makes the application more accessible, easy to maintain, and avoids time-consuming local installations that could lead to security breaches. Once users successfully authenticate themselves on the login page, the application conveniently loads in their web browser. Therefore, without any constraints, an immersive gamified simulation in first-person view can be used at any time and from any location with a common use network.

One of the main challenges was to produce a realistic 3D simulation of such a complex building, which could be rendered on a simple web browser in real time. For this kind of online use and considering the topology (Figure 3), in addition to improving the frame rate (frame per second [FPS]) and the visual quality, we had to reduce the overall application size for a faster download and ensure that the application data were not evicted from the browser cache mechanism. To achieve these optimization goals, we applied the following procedure:

- We simplified heavy graphical meshes (ie, geometric modeling) from the hospital scene and removed pieces that were never visible from the player's point of view (eg, elements under beds or within inaccessible areas).
- We reduced the size of all environment textures to their minimal acceptable size (64, 256, 512, etc) and enabled texture crunch compression option when available.
- As most objects are static in the scene, we used the static batching feature of Unity to improve the FPS as well as the light so-called *baking* feature to improve the visual quality with a lower impact on FPS.
- As our tests on occlusion were not demonstrating the expected improvement, we did not use this optimization feature even if it theoretically seemed adapted.
- Finally, to embed the application in our web page, we simply used the Unity WebGL exporter and made sure to deactivate the exception support as advised by Unity to gain a few additional FPS.

These optimizations allowed us to achieve an acceptable result in terms of size, speed, and visual quality (ie, 89 MB load and 25-35 FPS) for our medium tier computer: CPU Intel(R) Core (TM) i7-7660U CPU @ 2.50GHz; GPU Intel(R) Iris(R) Plus Graphics 640; 8 Go of RAM. Within the study, we believe that an interesting track would be using an incremental download mechanism using asset bundles to have a better "Start Render Time" for the application.

Quest Management Application

For security reasons, the quest management app illustrated in Figure 3 can only be used by permanent staff members. Its implementation uses a microservice-oriented architecture with Docker containers (Figure 4). It consists of three main microservices: (1) the public microservice, which serves as the frontend for end users and provides access to the Unity application with WebGL deployment; (2) the application programming interface (API) microservice, which furnishes a RESTful API (ie, API that conforms to the constraints of REST architecture) for the WebGL application to retrieve data from the database; and (3) the Admin microservice, which serves as

XSL•FO
RenderX

the core infrastructure component, helping staff members in easily managing quest command lines for the game, including creation, reading, and editing functionalities.

Figure 4. Technical illustration of the quest manager application. Every service runs inside a Docker-composed environment. A virtual network is created to ensure seamless communication between the 3 main components: (A) Nginx (an open-source software for web serving) orchestrates the access to the 3 microservices using a reverse proxy mechanism; (B) 3 microservices are deployed as Docker containers and the web framework used is Flask; (C) the database is deployed along the different services allowing data storage and we used *MongoDB* as database system. Data stored and managed in the quest management app: from (D) database and web server, through (E) types of resources, and up to the Unity-based gamified simulation (F). API: application programming interface; SQL: structured query language.

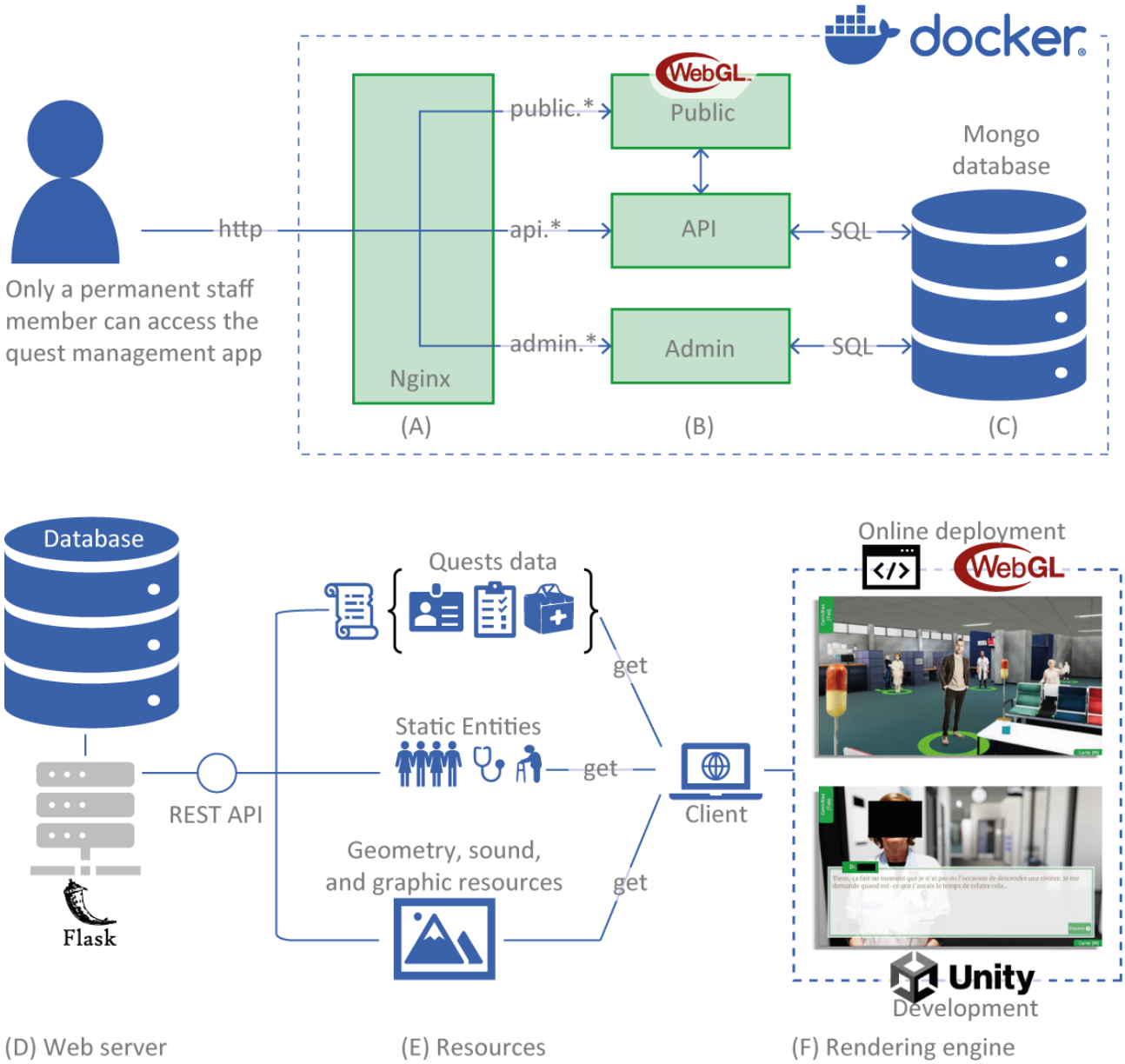

The access to the different microservices is orchestrated via Nginx. A reverse proxy mechanism is deployed to simplify access to diverse services by using a standardized nomenclature.

Link Between Both Apps

The application communicates with the server database via a REST API, retrieving information in JSON format (see Figure 4). This information includes graphic resources, character and interactive object placements, dialogs, and quest details, such as task sequences and success conditions. When the 3D simulation launches in the web browser, this information is transmitted, and the entities and quests are dynamically instantiated within the simulation, enabling dynamic creation of quest elements or entities.

User Tests

Before any deployment, we needed to test with people who did not know the project, but who were familiar with the site: that is, current permanent staff members. With this in mind, we took the initiative to carry out 2 preliminary tests. Data were collected using anonymous Google forms: (1) administrative and explanatory information before testing the apps and (2) sets of questions relative to the appreciation of different aspects of the apps with optional suggestions for improvements. The questionnaires were adapted from validated usability and

XSL•FO
RenderX

acceptability instruments, and data were analyzed using standard descriptive statistics, ensuring methodological robustness despite the preliminary nature of this study.

The user tests were anonymous with a double-blind protocol to assess performance and user experience objectively. In this protocol, neither the participants nor the evaluators were informed about the test conditions to reduce potential biases. In addition, the evaluators collecting data and analyzing results remained unaware of the identity or the experimental experience of the users.

Ethical Considerations

Ethics Review

This study assessed the feasibility of a digital simulation platform through anonymous surveys of adult health care professionals. No patients, medical records, or personal or health-related data were involved. According to the Swiss Human Research Act (*Loi relative à la recherche sur l'être humain, LRH, RS 810.30*) and the guidance of the *Commission cantonale d'éthique de la recherche sur l'être humain du Canton de Vaud (CER-VD*), such feasibility studies with anonymous professional staff do not fall within the scope of mandatory ethics committee review [28,29]. Consequently, no formal ethics board application was submitted.

Informed Consent

All participants were adult health care professionals who provided verbal informed consent before participation. They were informed of the voluntary nature of their participation, the absence of personal or health-related data collection, and the anonymity of their responses. Written or electronic consent was not required because the study posed no foreseeable risks and collected no identifiable or sensitive information.

Privacy and Confidentiality

All data were fully anonymous and deidentified at the point of collection. Surveys were conducted using Google Forms with IP tracking disabled. In compliance with institutional policies and the General Data Protection Regulation, no personal identifiers were stored, and only aggregated anonymized results were exported to secure institutional servers (HES-SO, UniL, and CHUV).

Compensation

No compensation, financial or otherwise, was offered or provided to participants for their involvement in the study.

Image and Participant Identification

All rendered illustrations included in the manuscript were designed to anonymize staff members. No recognizable individual is visible, and facial features, names, and personal identifiers have been either removed or replaced with nonidentifiable placeholders. Few pictures are not anonymized because they have been autogenerated and do not correspond to real, permanent people.

Results

3D Environment and Staff Representation

The results of the gamified simulation mainly consist of 2 categories: a 3D rendering environment where new employees can discover the working place in a 3D real-time rendering (Figure 5) and permanent staff members' graphical and dialog representations.

XSL•FO
RenderX

Figure 5. Area examples of gamified simulation: (A) main reception of the emergency department (ED); (B) main waiting room; (C) 1 of the 4 waiting rooms, even the table is visible in the physical world; (D) a strange sphere is visible at every intersection of the ED: it represents a real spherical photo of the premises to increase the connection between the physical and digital worlds; the user can enter it simply by approaching and looking at it, and leave it by clicking on the Esc key; (E) this vital ED room is equipped with numerous pieces of equipment and is located right next to the emergency entrance; (F) representation of a pivotal area for the emergency teams: the administrative area for personnel management; (G) one of the numerous first and second aid ED rooms; (H) comparing area with physical place; and (I) reception area for patients arriving by ambulance.

(A) Patient triage area (B) Waiting room (C) Internal waiting room
(D) Spherical map views (E) Emergency resuscitation room (F) Internal reception area
(G) Emergency first aid room (H) From digital to physical worlds (I) Ambulance reception

Figure 6 illustrates 9 possible staff interactions. Three types of characters are represented in this gamified simulation: permanent team members (as billboards only) and patients—as billboards or in full 3D. Billboard characters are those that can be changed into quests. This simple billboard representation presents 3 advantages: (1) it is a simple textured rectangle with transparency; therefore, it is impossible to make it any lighter for the online application; (2) anyone can take a photo and clip it with the automated clipping, which is a common feature on smartphones; (3) changing the scale (100 pixels for 1 m) and importing an image into the quest management application are within the reach of non-IT staff. Discovering such a large environment, with almost a hundred elements, can seem like a labyrinth. To help the user on the matter, we have added a minimap view that can be displayed by pressing the 'M' key, as shown in Figure 7. Interaction with agents (ie, digital characters) takes place through dialog with the possibility of associated questions and answers. Quizzes can be set up, and missions can be easily described.

XSL•FO
RenderX

Figure 6. Examples of encounters: (A) the first interaction users will have a tutorial proposing first steps in the 3D environment; (B) an example of an anecdote; (C) the main welcome desk where you can train patient sorting; (D) interaction with a patient rendered with a billboard; (E) an example of a patient rendered in real 3D; (F) a quiz on patient sorting; (G) color codes indicating how to interact with agents: green, a quest can be activated; red, a quest has been activated; blue, an anecdote can be read; (H) a set of multiple interactions for quests; and (I) a set of multiple anecdotes that can be read.

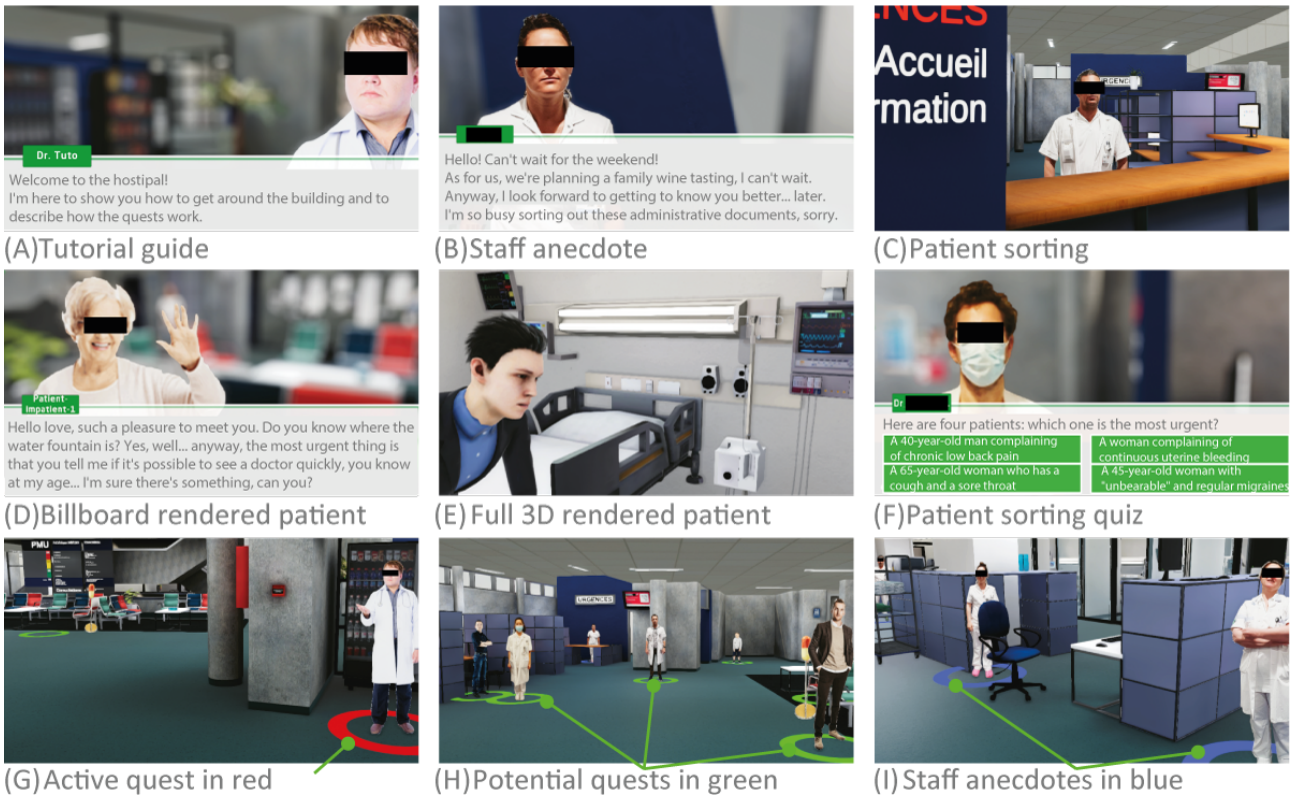

Figure 7. Illustrations of the minimap. Two locations (A and B) where we activated the 'M' key to bring up the minimap in the bottom right-hand corner of the screen; their respective enlarged views are shown on the right.

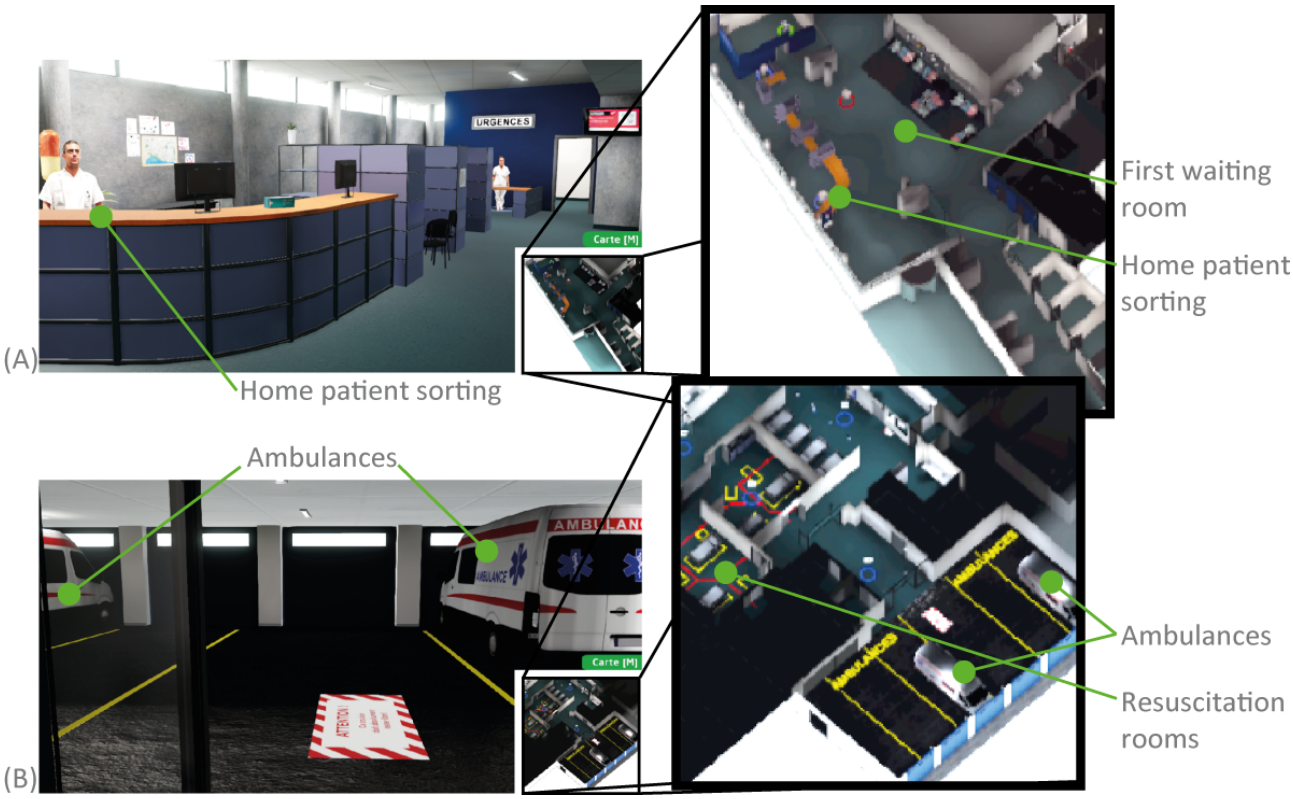

XSL•FO
RenderX

Results of the Preliminary User Tests

Overview

As there were 2 applications, we submitted 2 user tests (n=37 and n=9) and corresponding 2 questionnaires to the ED staff. We needed to know the opinion of the permanent staff as soon as possible for final parametrization of the prototypes (and therefore before conducting the large survey of naive users that we are preparing for Fall 2025). The profile of the respondents based on the information extracted from the survey and the plant's stats are illustrated in Multimedia Appendices 1 and 2.

3D Web App User Test

We conducted a survey regarding the 3D gamified simulation with 37 naive participants (all permanent staff members) and gathered results on 5 main criteria along with appreciation percentages. Providing personal anecdotes was considered motivating by 79% (29/37) of the respondents. To the question, *Does the app allow you to meet and get to know the permanent employees a little?*, 73% (27/37) of respondents believe it has potential; and to the question, *Does it help to get to know the equipment better?*, 82% (30/37) of the respondents agreed on it. The topology of the ED received 90% (33/37) positive answers with the lowest SD, indicating strong unanimous agreement. Finally, for the question *Do you believe this app is useful for new employees?*, 81% (30/37) of respondents found the app useful for new employees, suggesting more varied opinions among respondents—ie, the highest SD.

Maintenance Web App User Test

The online software used to create and manage quests was also evaluated with 9 participants (permanent staff members as well, three of them already aware of the 3D web app but not of the maintenance software), as shown in Multimedia Appendix 2.

The results are a little more varied but still positive: 91% (8/9) of respondents praised the modification possibility of the 3D simulation without the help of the technical staff. The construction of quests to discover the workplace and the people involved was also highly appreciated. The quests seemed to be relatively easy to build (77% approval), still with an elevated SD of 35%. Finally, on the question of whether scores and trophies were a source of motivation, responses were moderately positive, with 70% approval.

Discussion

Principal Findings

First of all, the large difference in participant numbers between 37 and 9 can be explained by the fact that the quest creation and management application can only be used by key people accredited by the ED. Given the long-term, interdisciplinary nature of our development effort, we believe that 5 aspects merit focused discussion. First, estimating the practical impact of our approach is essential to demonstrate its relevance in addressing real-world constraints, such as time, cost, and staff turnover. Second, the real-time 3D simulation and its gamified design form the technological and pedagogical core of our solution, directly shaping user engagement and memory retention. Third, the quest management application represents a crucial operational tool that empowers nontechnical medical staff to adapt and extend the simulation autonomously, ensuring sustainability. Fourth, sharing the challenges and insights from the multiyear deployment process provides valuable guidance for other teams pursuing similar innovations in hospital environments. Finally, exploring the generalizability of our platform to other departments helps establish its broader applicability, reinforcing the value of the framework beyond the ED context.

Implementing a gamified simulation similar to the one presented in this project has significant practical implications for a medical department, in terms of both costs and benefits. In financial terms, the development of the project required around 3000 engineer hours spread over 8 years. This relatively long period can be explained by starting from scratch, coordination between interinstitutional partners, and the prolonged interruption due to the COVID-19 pandemic. However, this initial investment can be justified by the potential gains offered by the solution once deployed.

In terms of benefits, the impact on team time and workload could be major. With a 100-person department with an annual turnover rate of 22%, which implies the integration of approximately 22 new employees each year, the estimated 30% to 40% reduction in integration time would save theoretically 15,400 hours per year. These time savings can be reallocated to activities directly linked to patient care. Similar results have been observed in previous studies [4,30-32], such as the implementation of a 10-element onboarding program in a community hospital, which reduced overall annual turnover from 18.2% to 11.9%, and new-hire turnover from 39.1% to 18.4%. Additionally, structured onboarding processes in EDs have been associated with improved operational efficiency, as evidenced by a significant decrease in patient door-to-discharge times.

Another significant benefit is improved productivity and quality of care. By enabling new employees to quickly familiarize themselves with the premises, equipment, and colleagues, the simulation reduces the stress associated with taking up a new post and accelerates their skills development. This promotes better coordination between teams and reduces the risk of errors due to lack of preparation or familiarity. Ultimately, the solution not only helps streamline onboarding processes but also improves operational efficiency and improves the patient's experience, even in an environment as demanding as the ED. Finally, the political dimension, the motivational aspect, highlights the department's ability to lead innovative and modern projects, enhancing its appeal and attractiveness.

3D Gamified Simulation

Overall, it seems that the simulation was positively received, especially in terms of helping users get acquainted with the equipment and the actual working environment. However, there may be some variability in opinions about its usefulness for new employees. Claims about potential time savings, efficiency gains, and improved team integration should therefore be interpreted as plausible expectations supported by prior literature. Still, they remain hypothetical until validated through longitudinal or performance-based evaluations with actual

XSL•FO
RenderX

newcomers. This feedback can be valuable for further refinement and improvement of the simulation. As for the deployment, we chose to deploy the apps on a gamified browser-based, real-time 3D simulation to increase immersion and realism with a relatively realistic environment (texture, lighting, topology) so that the gamified aspect potentially allows greater engagement of new employees and links long-term memory with a certain degree of emotion and web deployment means that the application can be accessed from anywhere, PCs, tablets, or even smartphones.

However, each of these properties can be improved. First, health departments often evolve; for example, rooms are reorganized and walls may be removed or built. These aspects are not currently supported by the quest maintenance software, implying that annual 3D updates should be required. Second, even if we had positive feedback from the first trials, professional-based quests should be tested; furthermore, current testers were permanent staff who know well the ER, so a large-scale user test with adequate target profiles must be set up and analyzed. Third, we did not test the app on low hardware configurations; it is necessary to study the limits in terms of hardware setup and graphics support (eg, resolution).

Quest App and Corresponding User Test

As presented in the previous section, the quest management application is functional, and the staff who tested it are enthusiastic. Note that these are the same employees who will be using it—not the new employees who will be using the online 3D simulation. The current ergonomics, due to a lack of means, could only be modest, even involving the (admittedly simple and automated) writing of XML code to implement the questions. Major efforts are required to make this application more user-friendly and ergonomic.

Two Larger-Scale User Tests

The abovementioned user tests were needed in the development process (see section Reaching Actual Deployment) but are not sufficient for the final deployment because we do not know the actual impact of this gamified simulation on actual end users. To address this issue, we are now scheduling 2 large-scale user tests: one with first- or second-year students (n>100), and 6 months later, one with the ones visiting the actual ED.

Reaching Actual Deployment

The context for deploying this kind of application—involving 4 institutes—in a hospital is atypical, which is why we believe it is important to present the steps we took to get there. As shown in Figure 8, we have been building this gamified simulation for about 8 years now and are in the fourth version of the project, started in 2017 [33]. Setting up an innovative project with well-defined needs, with a skilled and motivated team, with sufficient funding, is unfortunately not enough for a health care project to be implemented and really used: upstream, the institution's policies at both management and technical (IT department) levels must be considered. Indeed, IT and health imply important security issues, and when patient information is stored in the same building, an outcry can easily arise. Security on hospital servers is of paramount importance. To solve this problem, we have dedicated an external server with no connection (software or hardware) to internal IT services.

Figure 8. Design process of the project. This figure illustrates the 8-year design process of this project, including the consortium of the project and versioning progress during the last 8 years, with the first prototype being presented in 2018: with 7 types of partners—engineers, researchers, technicians, managers, doctors, specialists, and nurses—from quite distant fields. The financing and organization of such a project requires careful planning and strong communication.

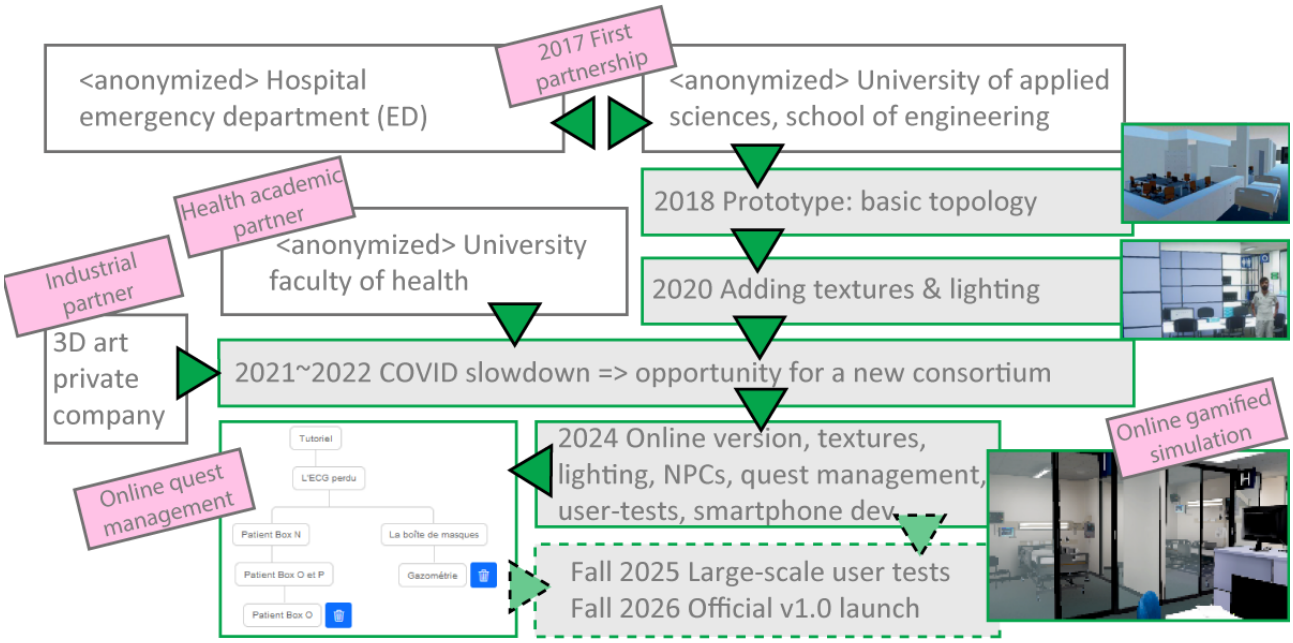

Generalizability of the Tool Beyond ED

Although this simulation was specifically designed for the onboarding of ED staff, the architecture of the system was deliberately built with adaptability in mind. The core components of the platform—modular 3D mapping, customizable quest creation, and a web-based interface—enable the deployment of similar simulations in other hospital

XSL•FO
RenderX

departments, even those with markedly different layouts or protocols. For example, departments such as radiology, intensive care, or surgery, each with their own spatial organization and workflows, could benefit from tailored versions of this system. The quest management application is designed to be fully editable by internal medical staff, without technical expertise, allowing adaptation to diverse professional roles, equipment, and team structures. Moreover, the interaction mechanisms (eg, dialogs, location-based tasks, quizzes) are generic and can be aligned with various onboarding objectives. Future developments will include tools for modular editing of building geometry (ie, nontrivial development where artificial intelligence will play an invaluable role), enabling even greater adaptability to departments undergoing spatial reconfiguration. While initial testing has focused on ED-specific challenges, the fundamental design supports scalability and generalization. A pilot adaptation for another department is currently under discussion to validate this potential. Ultimately, the platform aims to provide a reusable framework for immersive onboarding across the broader hospital ecosystem. This scalability is not only theoretical but also structurally embedded in the system's design: the geometry of any department can be replicated, whereas local staff are empowered to create department-specific scenarios, personnel representations, and equipment, all without IT support.

Tracks for Further Research and Perspectives

Overview

As introduced in the Discussion section, many aspects can be improved to make the 3D simulation even more realistic, especially the textures and lighting (eg, with real-time ray-casting). Still, we do not think the priority should be on graphics, but rather on content (eg, quests). In addition, before any actual deployment, a study of the real impact/benefit needs to be carried out, not with current employees, but with so-called naive people (eg, medical or nursing students). We propose the following short- and long-term perspectives.

Short-Term Perspective

The first task to be improved is relative to the quest management app. Indeed, its ergonomics need to be adapted because end users will be medical staff members; for instance, the current ways to design questions are based on XML coding: this should be replaced with point-and-click drop boxes. Furthermore, in terms of hardware, we are currently looking into the design of a dedicated server that would have no connection with the hospital's servers and would allow total security of patient data.

Long-Term Perspective

The primary goal is actual deployment. To achieve this, we will have to use feedback on the above-described large-scale user test and develop the corresponding improvements for the official deployment. One of the aspects we need to study is the impact and effectiveness in terms of time savings not only for health departments but also for the human characteristics and satisfaction of new employees. At an even longer term, we would like to set up a system that allows permanent medical staff to easily modify furniture and equipment in wards. The idea is to give a solution that simulates the often-necessary rearrangements in hospital institutes. In this context, as suggested in 2025 by Lindner and Leutritz [34], we would also like to provide VR deployment for a better understanding of real-life working situations (eg, reorganize a medical room, check that roll-away beds can be moved around the building properly). In addition, it would be interesting to integrate a mechanism for an easy updating of the building geometry; for instance, splitting in parts or the possibility of adding a semiautomatic 3D scan for new walls or openings. This last point is particularly important because it would enable similar gamified simulations to be deployed with much less effort for different health care services. It would also open a much broader vision, for example, on a hospital-wide scale.

Beyond the scope of this study, further research is critically needed to assess the long-term effectiveness of gamified onboarding solutions in health care. This includes evaluating their impact on staff retention, clinical performance, and team cohesion across various medical settings. Comparative studies across institutions with different cultural and organizational structures would help determine the generalizability of such tools. Furthermore, economic evaluations are necessary to analyze the balance between initial development costs and savings in recruitment and training. Finally, a better understanding of how specific game elements affect user motivation and engagement could help optimize simulation design based on department profiles and learner characteristics.

The following Table 1 summarizes the aspects we wish to implement to improve the perception of real adoptability.

Table . Improving the perception of real adoptability.

Dimension	Current state	Improvement path
Usability	Web-based app tested on medium hardware	Evaluate on low-end devices/older browsers
Content flexibility	Quests editable via XML	Transition to point-and-click editor (planned)
Engagement	Positive feedback from permanent staff	Test motivation and engagement in naive users
Security	Hosted on external server	Consider audit by hospital IT/independent review
Deployment readiness	CHUV[a] pilot completed	Prepare documentation for interhospital adaptation

[a]CHUV: Lausanne University Hospital.

XSL•FO
RenderX

Conclusions

This paper is fundamentally procedural, focusing on the why and how behind the development of a gamified simulation for hospital onboarding, rather than an analytical assessment of its long-term impact. While the necessity for improved onboarding solutions in high-turnover health care environments is widely acknowledged, many similar initiatives struggle to succeed due to the gap between IT specialists and medical staff.

Our experience suggests that the failure of such projects is often due to a lack of structured collaboration and mutual understanding between these 2 domains. To bridge this gap, we have detailed not only the rationale but also the entire development process, ensuring that both medical and IT professionals can grasp the conceptual, technical, and operational aspects required for successful implementations. The technical sections of this paper are thus deliberately aimed at hospital IT departments, which play a key role in ensuring the feasibility and sustainability.

RQs and Hypotheses

In parallel with this implementation perspective, the study also addressed 2 guiding research questions concerning (RQ1) the onboarding effectiveness of the gamified simulation and (RQ2) the operational autonomy enabled by the quest management tool. The preliminary findings provided encouraging support for both associated hypotheses: (H1) that the simulation would be perceived as useful and motivating and (H2) that permanent staff could manage onboarding content without IT support. While not yet a full-scale impact assessment, these results help establish a foundation for further empirical evaluation.

The close collaboration between the different partners has yielded an innovative dual app solution aimed at swiftly integrating new personnel. Through an interactive 3D application, new staff members can immerse themselves in virtual hospital environments, engaging with simulated colleagues and patients. The gamified approach of the application guides new employees through various quests, easing active exploration and learning. The development of a companion application for ED permanent staff ensures the solution adaptability to the dynamic nature of hospital departments while maintaining quality and consistency in the integration process. This solution not only streamlines the onboarding process but also enhances overall efficiency and effectiveness in the ED's workforce integration efforts.

Moving forward, future work will focus on large-scale deployments and measurable outcomes, including onboarding time reduction, staff satisfaction, and long-term retention metrics. The hypotheses introduced here provide a structured baseline for those efforts, and their confirmation or refinement will be crucial in validating the generalizability and clinical value of gamified onboarding platforms in health care.

Acknowledgments

First of all, the authors would like to thank all the engineering, especially Ivan Serra Moncadas, and emergency department staff members for their strong contribution and suggestions from an early stage to the final configuration, as well as internal user test phases. Special thanks are also due to Montserrat Gutierrez for her precious English review of the manuscript. Portions of the manuscript text were refined with the assistance of a generative language model (ChatGPT, OpenAI). The authors retained full responsibility for the scientific content, interpretation, and final approval of the text. Only 2 illustrations were generated with the help of DALL-E from OpenAI and redesigned with Photoshop & Illustrator. Moreover, pictures of patients for the current version of the 3D application and presented as billboards have been generated with artificial intelligence to avoid any potential copyright issues.

Funding

This study was supported by internal institutional funding from the Lausanne University Hospital Emergency Department, the Faculty of Medicine of the University of Lausanne, and the University of Applied Sciences and Arts Western Switzerland. Specific support was provided through the "CHUV-ED-ID" program (2017 - 2025) and SAGEX-ID grants 98432, 116828, 119393, and 130528.

Data Availability

The anonymized dataset generated during the user acceptability evaluation of the gamified simulation, including raw participant responses, aggregate metrics, and a copy of the evaluation questionnaire, is publicly available on Zenodo as placeholder [35]. All data are shared under the Creative Commons Attribution 4.0 International license (CC BY 4.0). This open access enables other researchers, developers, or health care institutions to further explore and replicate formative implementations in similar high-turnover clinical contexts.

Authors' Contributions

SG: writing – original draft, conceptualization, methodology, IT project administration, formal analysis, funding acquisition, investigation, visualization, resources, software UX/UI manager
AL: investigation, 3D app software development, writing – original draft
AS: quest management software development, investigation, writing – original draft

XSL•FO
RenderX

AB: resources, health validation
LAK: texture and design, resource, industrial validation
JMC: funding acquisition, health validation
EG: investigation, health validation
PNC: health project administration, writing – original draft, funding acquisition, resources, validation

Conflicts of Interest

None declared.

Multimedia Appendix 1
Results of the first user test. These charts present the resulting statistics of the first user test with n=37 using a 6-point Likert scale: (A) profiles of survey people: age, gender, and role; (B) the use of the main gamified simulation.
[PNG File, 100 KB - formative_v9i1e72202_app1.png]

Multimedia Appendix 2
Results of the second user test. These charts present the resulting statistics of the second user test with n=9 (i.e. supervisors only) using a 6-point Likert scale: (A) profiles of survey people: age, gender, and role; (B) the use of the quest management: tool made only by key people accredited by the emergency department (ie, permanent and managing the department).
[JPEG File, 84 KB - formative_v9i1e72202_app2.jpeg]

References

1. Castle NG. Measuring staff turnover in nursing homes. Gerontologist 2006 Apr;46(2):210-219. [doi: 10.1093/geront/46.2.210] [Medline: 16581885]
2. Merçay C, Grünig A, Dolder P. Personnel de santé en suisse – rapport national 2021: effectifs, besoins, offre et mesures pour assurer la relève. : OBSAN; 2023 URL: https://www.odasante.ch/fileadmin/odasante.ch/docs/Bildungspolitik/Versorgungsbericht2021/Obsan_CDS_OdASante_03_Rapport_national_2021.pdf [accessed 2025-11-27]
3. Dwesini NF. Causes and prevention of high employee turnover within the hospitality industry: a literature review. Afr J Hosp Tour Leis 2019;8(3):1-15 [FREE Full text]
4. Kurnat-Thoma E, Ganger M, Peterson K, Channell L. Reducing annual hospital and registered nurse staff turnover—a 10-element onboarding program intervention. SAGE Open Nurs 2017 Jan;3:2377960817697712. [doi: 10.1177/2377960817697712]
5. Krishnamurthy K, Selvaraj N, Gupta P, et al. Benefits of gamification in medical education. Clin Anat 2022 Sep;35(6):795-807. [doi: 10.1002/ca.23916] [Medline: 35637557]
6. Johnson D, Deterding S, Kuhn KA, Staneva A, Stoyanov S, Hides L. Gamification for health and wellbeing: a systematic review of the literature. Internet Interv 2016 Nov;6:89-106. [doi: 10.1016/j.invent.2016.10.002] [Medline: 30135818]
7. Tu MH, Lee SY. Effectiveness of gamification intervention platform on the work stress of shift-work nurses in the post-pandemic era. Stud Health Technol Inform 2024 Jul 24;315:598-599. [doi: 10.3233/SHTI240234] [Medline: 39049340]
8. Abildgren L, Lebahn-Hadidi M, Mogensen CB, et al. The effectiveness of improving healthcare teams' human factor skills using simulation-based training: a systematic review. Adv Simul (Lond) 2022 May 7;7(1):12. [doi: 10.1186/s41077-022-00207-2] [Medline: 35526061]
9. Dahalan F, Alias N, Shaharom MSN. Gamification and game based learning for vocational education and training: a systematic literature review. Educ Inf Technol 2024 Feb;29(2):1279-1317. [doi: 10.1007/s10639-022-11548-w]
10. Haubruck P, Nickel F, Ober J, et al. Evaluation of app-based serious gaming as a training method in teaching chest tube insertion to medical students: randomized controlled trial. J Med Internet Res 2018 May 21;20(5):e195. [doi: 10.2196/jmir.9956] [Medline: 29784634]
11. Joy E, R. A, Raja C. Digital 3D modeling for preconstruction real-time visualization of home interior design through virtual reality. Constr Innov 2024 Feb 13;24(2):643-653. [doi: 10.1108/CI-10-2020-0174]
12. Chen Y, Feng M. Urban form simulation in 3D based on cellular automata and building objects generation. Build Environ 2022 Dec;226:109727. [doi: 10.1016/j.buildenv.2022.109727]
13. De Oliveira MJF. 3D visual simulation platform for the project of a new hospital facility. Presented at: 24th Meeting of the European Working Group on Operational Research Applied to Health Services; Jul 19-24, 1998. [doi: 10.1142/9789812817839_0004]
14. De Oliveira MJF. 3D visual simulation of hospital admissions. In: Planning for the Future: Health Service Quality and Emergency Accessibility 2000:77-96 URL: https://www.researchgate.net/profile/Mario-Jorge-De-Oliveira/publication/301292226_3D_VISUAL_SIMULATION_OF_HOSPITAL_ADMISSIONS/links/570fe3f608aefb6cadaaa1c8/3D-VISUAL-SIMULATION-OF-HOSPITAL-ADMISSIONS.pdf [accessed 2025-11-27]

XSL•FO
RenderX

15. de Lima RM, de Medeiros Santos A, Mendes Neto FM, et al. A 3D serious game for medical students training in clinical cases. Presented at: 2016 IEEE 4th International Conference on Serious Games and Applications for Health (SeGAH); May 11-13, 2016. [doi: 10.1109/SeGAH.2016.7586255]
16. Gobron S, Chatelain S, Bolinhas C, de Oliveira DC. A picture-based serious game to train non-medical people for emergency situations. Presented at: 2019 IEEE 7th International Conference on Serious Games and Applications for Health (SeGAH); Aug 5-7, 2019. [doi: 10.1109/SeGAH.2019.8882466]
17. Yang S, Li T, Gong X, Peng B, Hu J. A review on crowd simulation and modeling. Graph Models 2020 Sep;111:101081. [doi: 10.1016/j.gmod.2020.101081]
18. Shapiro MJ, Morey JC, Small SD, et al. Simulation based teamwork training for emergency department staff: does it improve clinical team performance when added to an existing didactic teamwork curriculum? Qual Saf Health Care 2004 Dec;13(6):417-421. [doi: 10.1136/qhc.13.6.417] [Medline: 15576702]
19. Zielke MA, Zakhidov D, Hardee G, et al. Developing virtual patients with VR/AR for a natural user interface in medical teaching. Presented at: 2017 IEEE 5th International Conference on Serious Games and Applications for Health (SeGAH); Apr 2-4, 2017. [doi: 10.1109/SeGAH.2017.7939285]
20. Gue S, Ray J, Ganti L. Gamification of graduate medical education in an emergency medicine residency program. Int J Emerg Med 2022 Aug 30;15(1):41. [doi: 10.1186/s12245-022-00445-1] [Medline: 36042403]
21. van Gaalen AEJ, Brouwer J, Schönrock-Adema J, Bouwkamp-Timmer T, Jaarsma ADC, Georgiadis JR. Gamification of health professions education: a systematic review. Adv Health Sci Educ Theory Pract 2021 May;26(2):683-711. [doi: 10.1007/s10459-020-10000-3] [Medline: 33128662]
22. Larson K. Serious games and gamification in the corporate training environment: a literature review. TechTrends 2020 Mar;64(2):319-328. [doi: 10.1007/s11528-019-00446-7]
23. Stoicu-Tivadar L, Stoicu-Tivadar V, Berian D, Drăgan S, Serban A, Serban C. eduCRATE--a virtual hospital architecture. Stud Health Technol Inform 2014;205:803-807. [Medline: 25160298]
24. Szeto MD, Strock D, Anderson J, et al. Gamification and game-based strategies for dermatology education: narrative review. JMIR Dermatol 2021 Aug 30;4(2):e30325. [doi: 10.2196/30325] [Medline: 37632819]
25. Lehtoranta S, Xi N, Hamari J. Gamification and employee well-being: a systematic literature review. Presented at: Hawaii International Conference on System Sciences; Jan 3-6, 2024. [doi: 10.24251/HICSS.2024.633]
26. Rash I, Shi K, Sussel R, et al. Virtual 3D simulation technology for interprofessional team training. Simul Gaming 2024 Apr;55(2):281-301. [doi: 10.1177/10468781231222969]
27. Shi Y, Ferlet E, Crawfis R, Phillis P, Durano K. 3D hospital: design and implement quest-based game framework for transitional training. Presented at: 2015 Computer Games; Jul 27-29, 2015. [doi: 10.1109/CGames.2015.7272968]
28. RS 810.30 - loi fédérale du 30 septembre 2011 relative à la recherche sur l'être humain (loi relative à la recherche sur l'être humain, LRH). Fedlex. URL: https://www.fedlex.admin.ch/eli/cc/2013/617/fr [accessed 2025-11-27]
29. Commission cantonale d'éthique de la recherche sur l'être humain (CER-VD). URL: https://www.cer-vd.ch/ [accessed 2025-11-27]
30. Lockhart L. Strategies to reduce nursing turnover. Nurs Made Incred Easy 2020;18(2):56-56. [doi: 10.1097/01.NME.0000653196.16629.2e]
31. Brook J, Aitken L, Webb R, MacLaren J, Salmon D. Characteristics of successful interventions to reduce turnover and increase retention of early career nurses: A systematic review. Int J Nurs Stud 2019 Mar;91:47-59. [doi: 10.1016/j.ijnurstu.2018.11.003] [Medline: 30669077]
32. Kirk M. Strategies for Health Care Administration Leaders to Reduce Hospital Employee Turnover: Walden University; 2017.
33. Németh J, Carron PN, Bentvelzen A, Gobron S. Gamified-ED3D project. Presented at: GSGS'18: 3rd Gamification Serious Game Symposium; Jul 5-6, 2018.
34. Lindner M, Leutritz T, Backhaus J, König S, Mühling T. Knowledge gain and the impact of stress in a fully immersive virtual reality-based medical emergencies training with automated feedback: randomized controlled trial. J Med Internet Res 2025 Jun 4;27:e67412. [doi: 10.2196/67412] [Medline: 40465566]
35. Survey of two apps to help new employee and reduce high turnover in health depts. Zenodo. URL: https://doi.org/10.5281/zenodo.15737221 [accessed 2025-11-19]

Abbreviations

API: application programming interface
ED: emergency department
FPS: frame per second
RQ: research question
VR: virtual reality

XSL•FO
RenderX

Edited by J Sarvestan; submitted 05.02.25; peer-reviewed by C Golz, I Hapsari; revised version received 24.09.25; accepted 26.09.25; published 15.12.25.

Please cite as:
Gobron S, Lestrade A, Sadiku A, Bentvelzen A, Ait Kaci L, Carrier JM, Guyot E, Carron PN
Gamified Simulation for Onboarding Health Care Teams in Emergency Care: Development and Preliminary Feasibility Study
JMIR Form Res 2025;9:e72202
URL: https://formative.jmir.org/2025/1/e72202
doi:10.2196/72202

© Stephane Gobron, Antoine Lestrade, Artan Sadiku, Alexandre Bentvelzen, Leila Ait Kaci, Jean-Michel Carrier, Emmanuelle Guyot, Pierre-Nicolas Carron. Originally published in JMIR Formative Research (https://formative.jmir.org), 15.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Factors Associated With Suicidal Ideation Among Persons With Disabilities in South Korea: Retrospective Observational Study

In-Hwan Oh[1], MD, PhD; Sooyeon Jo[2], MPH, PhD; Joon Lee[3,4,5], PhD

[1]Department of Preventive Medicine, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea
[2]Department of Preventive Medicine, College of Medicine, Kyung Hee University, Seoul, Republic of Korea
[3]Data Intelligence for Health Lab, Cumming School of Medicine, University of Calgary, 3280 Hospital Dr NW, Calgary, AB, Canada
[4]Department of Cardiac Sciences, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada
[5]Department of Community Health Sciences, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada

Corresponding Author:
Joon Lee, PhD
Data Intelligence for Health Lab, Cumming School of Medicine, University of Calgary, 3280 Hospital Dr NW, Calgary, AB, Canada

Abstract

Background: South Korea has the highest suicide rate among the Organisation for Economic Co-operation and Development nations, with particularly elevated figures among persons with disabilities. Research has shown a strong correlation between suicidal ideation and suicide attempts.

Objective: This study aimed to investigate the factors that contribute to suicidal ideation among persons with disabilities in South Korea, utilizing machine learning methods based on national survey data.

Methods: We employed data from the 2020 National Survey on Persons with Disabilities in South Korea, which included 7025 respondents. The primary variable of interest was the answer to the question, “have you thought about taking your own life in the past year?” The dataset was divided into training (80%) and test (20%) subsets. Because the survey contained too many questions (n=1394), feature selection was conducted using random forest variable importance to identify the top 100 features. Subsequently, 5 machine learning models were trained to predict suicidal ideation based on the selected features: logistic regression, support vector machine, random forest, Extreme Gradient Boosting (XGBoost), and feed-forward neural network.

Results: A total of 6832 persons with disabilities responded to the suicidal ideation question and were included in the study. The most common types of primary disability were physical disability (n=1773, 26.0%) and hearing disability (n=979, 14.3%). Of the 6832 persons with disabilities, 12.1% (n=829) indicated they had had suicidal thoughts in the past year. Significant factors that impacted suicidal ideation included intense feelings of sadness, difficulties associated with their disabilities, and overall health satisfaction. Among the models tested, the random forest model exhibited the best predictive performance with a median area under the receiver operating characteristic curve of 0.905 (IQR 0.895-0.913), a median precision of 0.592 (IQR 0.561-0.616), and a median recall of 0.588 (IQR 0.564-0.620).

Conclusions: This study highlights critical predictors of suicidal ideation in persons with disabilities in South Korea, underscoring the necessity for focused mental health interventions. The results demonstrate the potential of machine learning to identify these factors, which can aid in the development of future suicide prevention strategies. Future work is warranted to investigate if the factors identified in this study are causal.

(JMIR Form Res 2025;9:e80261) doi:10.2196/80261

KEYWORDS

feature selection; machine learning; persons with disabilities; suicidal ideation; suicide attempt

Introduction

The suicide rate among South Koreans is so high that it ranks first in the Organisation for Economic Co-operation and Development as of 2020 [1]. It is also well known that the suicide rate among older adults is high [2], the suicide rate of persons with disabilities is higher than that of persons without disabilities, and the suicide mortality rate of persons with disabilities is more than twice that of the total population [3]. Furthermore, suicide mortality risk has been found to be pronounced in individuals with a mental disorder, visual impairment, or brain damage, particularly in those with severe forms of these disabilities [4].

Several studies have linked suicidal ideation to suicide [5]. In a study of patients receiving mental health treatment, it was reported that people with suicidal ideation were 6 times more

XSL•FO
RenderX

likely to attempt suicide than those who did not, and 5 times more likely to die of suicide [6]. This is the same in South Korea, and the studies of adolescents have reported that suicidal ideation increases the risk of suicide planning by 50 times and suicide attempt by 23 times [7]. The suicide risk of people who attempted suicide was more than 20 times higher than that of the general population [8]. It has been established that suicidal ideation affects and is directly related to suicide attempts [9-11]. For example, a low-income level has been pointed out as a factor affecting suicidal ideation [12], which is also the case in South Korea [13]. Similarly, gender and age also have an effect, and it has been shown that women and older adults have higher rates of suicidal ideation [14,15].

In the case of persons with disabilities, the studies on factors affecting suicidal ideation are relatively rare, with 1 study reporting that physical disability is associated with an increased risk of lifelong suicidal ideation [16]. It is even more difficult to find a study on suicidal ideation for persons with disabilities in South Korea. Considering the high suicide rate of persons with disabilities, it is necessary to analyze the factors that affect the suicidal ideation of persons with disabilities as a starting point.

The National Survey on Persons with Disabilities (NSPD) is a survey conducted since 1980 under the Welfare of Persons with Disabilities Act to identify the proportion of the population with disabilities and their living conditions and welfare needs, and is used for policy establishment [17]. Since the NSPD collects information on the overall lives of persons with disabilities, including living conditions and welfare needs, it consists of a very large number of questions. Individual questions for each type of disability stipulated in South Korea are presented, and the common questions for persons with disabilities alone consist of more than 1000 questions. Therefore, the vastness of these data makes it difficult to identify related factors through existing statistical methods. As a way to overcome this problem, machine learning offers effective computational methods to identify important features and model the complex, nonlinear relationships between potential factors and suicidal ideation.

Therefore, this study aimed to analyze the factors affecting the suicidal ideation of persons with disabilities via the application of machine learning to the 2020 NSPD data. Supervised machine learning methods were leveraged to identify survey responses that were most strongly associated with suicidal ideation.

Methods

Data Source

In South Korea, disability is registered in the South Korean disability registration system, and welfare benefits are provided based on this system [17]. The 2020 NSPD data were collected from persons with disabilities registered with the Ministry of Health and Welfare as of May 2020. The NSPD dataset was constructed by taking into account gender, region, type of disability, and age, ensuring that it accurately reflects the overall characteristics of persons with disabilities in South Korea. The survey contained 1394 questions, and a total of 7025 persons with disabilities completed the survey. The response to the question “have you thought about taking your own life in the past year?” was the target variable for the machine learning models. A currency exchange rate of 1180.01 Korean Won per US dollar from 2020 was used for the household incomes reported in the NSPD [18].

Data Partitioning

A random 20% of the dataset was held out as test data. The remaining 80% were subject to 10-fold cross-validation for training and hyperparameter tuning. All partitioning was done in a stratified manner to preserve the target variable prevalence across different datasets.

Data Preprocessing

Of the 7025 persons with disabilities in the dataset, 193 individuals did not respond to the target variable question and were excluded. This resulted in a sample size of 6832 individuals for the subsequent steps.

Features (ie, predictor variables) with data missing rates >20% were excluded. This step excluded all 149 free-text features in the dataset and 501 continuous or categorical features. Also, the response to the question “have you attempted to take your own life in the past year?” was dropped as a feature due to its high correlation with the target variable question. In the end, a total of 742 features were included in this study.

The missing data in the included features were imputed with the median and mode for continuous and categorical features, respectively. All continuous features were scaled to be 0 mean and unit variance, whereas all categorical features were one-hot encoded. Class imbalance was mitigated by using the Synthetic Minority Oversampling Technique (SMOTE) [19] from the *imbalanced-learn* Python package with 10 nearest neighbors to equalize the sizes of the majority and minority classes. The imputation, scaling, one-hot encoding, and SMOTE were learned or executed only within the training data and subsequently applied to the validation and test data to avoid data leakage.

Feature Selection

Since there were too many features relative to the sample size and the prevalence of the target variable, we selected the top 100 features using the *SelectFromModel* function in the scikit-learn Python package. For this feature selection, we used the feature importance measures from a random forest model with 100 trees.

Machine Learning Modeling and Performance Evaluation

We trained and evaluated the following 5 machine learning models: logistic regression, support vector machine, random forest, Extreme Gradient Boosting (XGBoost), and feed-forward neural network. Hyperparameters were tuned via a grid search and 10-fold cross-validation. The final models were trained using the entire 80% training or validation data with the best hyperparameter settings.

Predictive performance was evaluated using the following metrics: accuracy, balanced accuracy, average precision (the area under the precision-recall curve), Brier score, F_1-score, precision (positive predictive value), recall (sensitivity), and

XSL•FO
RenderX

the area under the receiver operating characteristic curve (AUROC). Hyperparameter tuning optimized the mean F_1-score from the 10-fold cross-validation. The test performances of the final tuned models were evaluated using 100 bootstrap samples of the test data and the metrics listed above.

All machine learning modeling was conducted using the *scikit-learn* and *xgboost* Python packages. This study followed the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) checklist for cross-sectional studies (Checklist 1).

Ethical Considerations

This study was conducted with the approval of the Kyung Hee University Institutional Review Board (Institutional Review Board No. 21 - 397 [EA]). The NSPD data are managed by the South Korean Ministry of Health and Welfare and are available to researchers for whom the need for informed consent is waived. The NSPD dataset used in this study was deidentified. Since this study was a secondary data analysis, no compensation was provided to individuals who contributed data to the NSPD. Since there are no images in this study, including in the multimedia appendices, the identification of individuals in images is not a concern.

Results

Table 1 describes the characteristics of the 6832 individuals with disabilities included in this study. There were more male participants than female participants (n=4053, 59.3% vs n=2779, 40.7%), and the median age was 64 (IQR 52-75) years. The most common primary disability was physical disability (n=1773, 26.0%), followed by hearing disability (n=979, 14.3%), disability of brain lesion (n=805, 11.8%), and visual disability (n=786, 11.5%). Most (n=4896, 71.7%) lived in an urban area. The prevalence of having suicidal ideation in the past year was 12.1% (n=829), whereas 0.7% (n=49) of the individuals attempted suicide in the past year.

XSL•FO
RenderX

Table . Cohort characteristics of the 6832 individuals with disabilities included in this study.

Variable	Values
Sex, n (%)	
Male	4053 (59.3)
Female	2779 (40.7)
Age (y), median (IQR)	64 (52-75)
Primary disability, n (%)	
Physical disability	1773 (26.0)
Disability of brain lesion	805 (11.8)
Visual disability	786 (11.5)
Hearing disability	979 (14.3)
Speech disability	205 (3.0)
Intellectual disorder	500 (7.3)
Autistic disability	126 (1.8)
Mental disorder	374 (5.5)
Kidney dysfunction	490 (7.2)
Cardiac dysfunction	101 (1.5)
Respiratory dysfunction	144 (2.1)
Hepatic dysfunction	152 (2.2)
Facial disfigurement	75 (1.1)
Intestinal or urinary fistula	187 (2.7)
Epilepsy	135 (2.0)
Severity of disability, n (%)	
Severe	3443 (50.4)
Not severe	3389 (49.6)
Monthly household income in the past year (Won [US $]), median (IQR)	1,500,000 Won (US $1271) (800,000-2,700,000 Won [US $678-$2288])
Region, n (%)	
Large urban	3198 (46.8)
Small or medium urban	1698 (24.9)
Rural	1936 (28.3)
Overwhelming sadness or depression for 2 consecutive weeks in the past 2 years, n (%)	
Yes	1338 (19.6)
No	5494 (80.4)
Suicidal ideation in the past year, n (%)	
Yes	829 (12.1)
No	6003 (87.9)
Suicidal attempts in the past year, n (%)	
Yes	49 (0.7)
No	6783 (99.3)

Table 2 shows the predictive performances of the 5 final models on the test data, whereas the best hyperparameter settings within the search space for each model are shown in italics in Table 3. The random forest resulted in the best performance in terms of median accuracy, average precision, Brier score, precision, and AUROC. The XGBoost model resulted in the best median balanced accuracy, F_1-score, and recall. Logistic regression also yielded the best median AUROC, which was tied with the random forest.

XSL•FO
RenderX

Table . Suicidal thought prediction performance on the test data.

Metric	Logistic regression, median (IQR)	Support vector machine, median (IQR)	Random forest, median (IQR)	XGBoost[a], median (IQR)	Feed-forward neural network, median (IQR)
Accuracy	0.865 (0.861, 0.879)	0.890 (0.884-0.895)	**0.899**[b] (0.895-0.904)	0.873 (0.868-0.880)	0.881 (0.875-0.888)
Balanced accuracy	0.828 (0.815, 0.839)	0.796 (0.781-0.812)	0.765 (0.753-0.781)	**0.843**[b] (0.830-0.855)	0.800 (0.784-0.809)
Average precision	0.583 (0.551-0.617)	0.609 (0.578-0.639)	**0.621**[b] (0.589-0.653)	0.599 (0.575-0.636)	0.568 (0.549-0.596)
Brier score	0.106 (0.102-0.111)	0.075 (0.071-0.079)	**0.069**[b] (0.065-0.072)	0.104 (0.100-0.108)	0.089 (0.085-0.092)
F_1-score	0.587 (0.568-0.602)	0.600 (0.577-0.620)	0.590 (0.564-0.612)	**0.608**[b] (0.590-0.628)	0.586 (0.566-0.608)
Precision	0.473 (0.451-0.488)	0.537 (0.517-0.566)	**0.592**[b] (0.561-0.616)	0.491 (0.470-0.513)	0.513 (0.488-0.537)
Recall	0.781 (0.754-0.802)	0.675 (0.648-0.701)	0.588 (0.564-0.620)	**0.805**[b] (0.781-0.825)	0.693 (0.658-0.713)
AUROC[c]	**0.905**[b] (0.893-0.925)	0.897 (0.886-0.907)	**0.905**[b] (0.895-0.913)	0.896 (0.888-0.907)	0.893 (0.884-0.900)

[a]XGBoost: Extreme Gradient Boosting.

[b]The best median performance of each metric is shown in bold.

[c]AUROC: area under the receiver operating characteristic curve.

Table . Hyperparameter search space for each machine learning model.

Model	Hyperparameter search space
Logistic regression	• penalty: l1, l2, *elasticnet*[a], none • l1_ratio: *0.5* • C: 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.1, 1.2, *1.3*, 1.4, 1.5 • class_weight: balanced, *none* • Solver: *saga*
Support vector machine	• kernel: linear, poly, *rbf*, sigmoid • degree: 5 • gamma: *scale*, auto • C: *0.5*, 0.6, 0.7, 0.8, 0.9, 1.0, 1.1, 1.2, 1.3, 1.4, 1.5 • class_weight: *balanced*, none
Random forest	• n_estimators: 50, 100, 150, 200, *250*, 300, 350, 400, 450, 500 • criterion: *gini*, entropy, log_loss • specify_max_depth: *true* • max_depth: *10*, 20, 30, 40, 50, 60, 70, 80, 90, 100 • class_weight: balanced, *balanced_subsample*, none
XGBoost	• n_estimators: 50, 100, 150, 200, 250, 300, 350, 400, 450, *500* • grow_policy: 0, *1* • max_depth: 10, 20, 30, 40, 50, 60, 70, 80, *90*, 100 • learning_rate: 0.0001, 0.001, *0.01*, 0.1 • booster: gbtree, *gblinear*, dart
Feed-forward neural network	• hidden layer_sizes: (25,), (50,), (*100,*), (200,), (25, 25), (50, 50), (100, 100), (200, 200), (100, 50, 25), (200, 50, 200), (200, 100, 50, 25) • activation: logistic, tanh, *relu* • alpha: *0.0001*, 0.001, 0.01, 0.1 • learning_rate_init: *0.0001*, 0.001, 0.01, 0.1 • learning_rate: constant, *invscaling*, adaptive

[a]The best settings based on a grid search are shown in italics.

Table 4 shows the top 10 features used by the best model (ie, the random forest). The survey questions that the top 10 features are based on are related to the levels of depression or stress experienced, satisfaction with health status or life in general, and ability to access health care. All of the 100 selected features used by the random forest are shown in Multimedia Appendix 1.

XSL•FO
RenderX

Table . Top 10 features of the best model (random forest in Table 2) in descending order of random forest variable importance.

Survey question	Response	Random forest variable importance
In the past 2 years, have you felt overwhelmingly sad or depressed for 2 weeks in a row, to the extent that it adversely impacted your everyday life?	Yes	0.0952
In the past year, have you felt overwhelmingly sad or depressed for 2 weeks in a row, to the extent that it adversely impacted your everyday life?	No	0.0522
Do you frequently experience issues due to your disabilities?	No	0.0140
Do you frequently experience issues due to your disabilities?	Yes	0.0138
Are you currently satisfied with your life?	Slightly satisfied	0.0120
What is your assessment of your current health?	Average	0.0118
How much stress do you feel daily?	Feel a little bit of stress	0.0116
Are you satisfied with the number of friends you have?	Slightly satisfied	0.0114
In the past year, have you ever been unable to visit to a health care facility when you wanted to?	Yes	0.00869
Are you currently satisfied with your health?	Very unsatisfied	0.00815

Discussion

Principal Findings

In this study, we investigated the factors affecting the suicide intention of persons with disabilities by analyzing the 2020 NSPD survey data using machine learning and identifying the 100 most important factors related to suicide ideation. Whether or not the persons with disabilities had experienced depression over the past 2 years was the most important factor, followed by the question of whether they had experienced frequent problems with disability, and whether they were satisfied with their current health level, stress, the number of friends, and access to medical institutions. With respect to the prediction of suicide intention, the random forest model was found to have superiority overall, with XGBoost performing best in terms of balanced accuracy, F_1-score, and recall.

Depression has been known to have a great influence on suicidal ideation [20,21]. For example, a study of Canadians aged 15 - 24 years showed that depression and suicidal ideation were correlated (r=0.34) [22]. In South Korea, it has also been shown that depression is an important cause of increased suicidal ideation [23]. Another adolescent-focused study developed and externally validated machine learning models using survey data from South Korea, the United States, and Norway to predict suicidal thinking [24]. This study also found that feelings of sadness and despair were the most influential factor. While previous studies usually focused on the associations between depression and suicidal ideation in general populations, this study showed that this relationship is similar in persons with disabilities. In both the general and persons with disabilities populations, the relationship between depression and suicidal ideation appears to be similarly influenced by stress, satisfaction with current life, and friendship.

Whether one has experienced problems due to disability or limited access to medical institutions can be said to be relatively unique to persons with disabilities. This means that there are distinct parts of suicide ideation for persons with disabilities and that there are additional points to consider for preventing suicide. Further research is warranted to study whether improving convenience in daily life can potentially reduce the likelihood of suicidal ideation and intentions in persons with disabilities.

Conversely, what is unique in this study is that age and income level did not make the 10 most important factors. Also, gender was only ranked 64th and 72nd among the top 100 factors. This study showed that the characteristics of disabilities have an influence beyond gender, income level, or age on suicidal ideation. In other words, the characteristics of disabilities appear to be more important than other factors except for depression, and these characteristics have been identified as important factors in previous studies that analyzed suicide in persons with disabilities. For example, while both income and disability are associated with suicide, the suicide rate itself can be higher for persons with disabilities with high income levels than for people without disabilities with low-income levels [25]. This finding suggests that reducing difficulties in persons with disabilities could be beneficial in lowering the suicide rate and reducing the gap between persons with disabilities and people without disabilities, although it is difficult to draw a definitive conclusion based on the results of this observational study.

In South Korea, individuals with a disability are required to be registered to access welfare benefits, and registration is divided

XSL•FO
RenderX

into 15 main types of disability. Disabilities are primarily classified as physical and mental disorders and include 6 external physical dysfunctions (physical, brain lesion, visual, hearing, speech, and sleep disorders) and 6 internal organ disorders (renal, heart, liver, respiratory, ostomy, and epilepsy). Mental disorders consist of 2 developmental disorders (intellectual and autistic disorders) and psychiatric impairments. In this study, the type itself was found to be less important with respect to suicidal ideation. Previous research has shown that individuals with mental disorders have a higher suicide rate. However, for persons with disabilities as a whole, the discomfort or barriers they face appear to be an even more significant risk factor [25]. Therefore, reducing environmental barriers for persons with disabilities may be a top priority, not only for those with mental disorders but for all individuals with disabilities.

For example, in South Korea, a primary care physician program for persons with disabilities has recently been implemented [26]. This system takes a comprehensive approach to address the barriers to accessing medical care experienced by individuals with a disability. Although it is still struggling with low participation rates, this approach can help ease the difficulty of accessing medical institutions due to disability.

Machine learning proved to be effective in this study in identifying the most important features associated with suicidal ideation in a complex, multidimensional dataset where a total of 742 features from 6832 individuals were investigated. Extensive hyperparameter tuning on 5 different machine learning models representing a range of model complexity and the use of a class imbalance mitigation technique (ie, SMOTE) led to robust results. The ability of machine learning to learn nonlinear multivariable relationships in an empirical way without the need to make assumptions about underlying data distributions was well suited to the dataset used in this study.

Limitations

The limitations of the study can be pointed out as follows. First, this study is a cross-sectional observational study, and causation cannot be established. Although the results of this study are largely supported by the literature, further research is needed to investigate causal relationships. Second, this study included only registered individuals with a disability. There are unregistered persons with disabilities who were excluded. Finally, while we investigated reasonably diverse machine learning methods in this study, there are other feature selection and class imbalance mitigation methods, as well as other machine learning models, that may have led to better prediction performance.

Conclusions

This study makes novel contributions by identifying factors associated with suicidal ideation using a large survey dataset collected from persons with disabilities in South Korea. The use of machine learning was another novel aspect of this study. Our results showed that both disability characteristics such as general discomfort caused by the disability and nondisability characteristics such as depression are important factors. Future work can focus on studying whether the associations between these factors and suicidal ideation are causal.

Acknowledgments

We would like to thank the Ministry of Health & Welfare, Republic of Korea, for making the National Survey on Persons with Disabilities data available to researchers.

Funding

This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute, funded by the Ministry of Health & Welfare, Republic of Korea (grant HI21C2122). The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

The South Korean Ministry of Health and Welfare is the custodian of the National Survey on Persons with Disabilities data. While we do not have the permission to make our dataset available to other researchers, any interested researchers should submit their data requests to the South Korean Ministry of Health and Welfare.

Conflicts of Interest

JL is a cofounder and major shareholder of Symbiotic AI, Inc. All other authors have no conflicts of interest to declare.

Multimedia Appendix 1
List of 100 selected features.
[XLSX File, 16 KB - formative_v9i1e80261_app1.xlsx]

Checklist 1
STROBE checklist.
[PDF File, 177 KB - formative_v9i1e80261_app2.pdf]

XSL•FO
RenderX

References

1. Suicide rates. OECD. URL: https://www.oecd.org/en/data/indicators/suicide-rates.html [accessed 2025-06-20]
2. Causes of death statistics in 2021. Ministry of Data and Statistics. 2022 Sep 27. URL: https://mods.go.kr/board.es?mid=a20108100000&bid=11773&act=view&list_no=421206 [accessed 2025-12-06]
3. Number of registered persons with disabilities and disability pension recipients. Ministry of Health and Welfare. URL: https://www.mohw.go.kr/menu.es?mid=a20306000000 [accessed 2025-12-06]
4. Lee HY, Shin DW, Han KD, Kawachi I. Longitudinal association between disability and suicide mortality in Republic of Korea. Int J Epidemiol 2024 Oct 13;53(6):dyae163. [doi: 10.1093/ije/dyae163] [Medline: 39657979]
5. Kessler RC, Berglund P, Borges G, Nock M, Wang PS. Trends in suicide ideation, plans, gestures, and attempts in the United States, 1990-1992 to 2001-2003. JAMA 2005 May 25;293(20):2487-2495. [doi: 10.1001/jama.293.20.2487] [Medline: 15914749]
6. Simon GE, Coleman KJ, Rossom RC, et al. Risk of suicide attempt and suicide death following completion of the Patient Health Questionnaire depression module in community practice. J Clin Psychiatry 2016 Feb;77(2):221-227. [doi: 10.4088/JCP.15m09776] [Medline: 26930521]
7. Kim MK, Cho KW. Factors affecting health risk behaviors, suicidal ideation, suicidal plans and suicidal attempts in adolescents. Korean J Health Educ Promot 2019 Sep 30;36(3):61-71. [doi: 10.14367/kjhep.2019.36.3.61]
8. Lee DW, Kwon J, Yang J, et al. Suicide related indicators and trends in Korea in 2018. Health Policy Manag 2020:112-119 [FREE Full text]
9. Klonsky ED, May AM, Saffer BY. Suicide, suicide attempts, and suicidal ideation. Annu Rev Clin Psychol 2016 Mar 28;12(1):307-330. [doi: 10.1146/annurev-clinpsy-021815-093204]
10. Brown GK, Beck AT, Steer RA, Grisham JR. Risk factors for suicide in psychiatric outpatients: a 20-year prospective study. J Consult Clin Psychol 2000 Jun;68(3):371-377. [doi: 10.1037/0022-006X.68.3.371] [Medline: 10883553]
11. Franklin JC, Ribeiro JD, Fox KR, et al. Risk factors for suicidal thoughts and behaviors: a meta-analysis of 50 years of research. Psychol Bull 2017 Feb;143(2):187-232. [doi: 10.1037/bul0000084] [Medline: 27841450]
12. Dendup T, Zhao Y, Dorji T, Phuntsho S. Risk factors associated with suicidal ideation and suicide attempts in Bhutan: an analysis of the 2014 Bhutan STEPS survey data. PLoS ONE 2020;15(1):e0225888. [doi: 10.1371/journal.pone.0225888] [Medline: 31999708]
13. Kim JL, Kim JM, Choi Y, Lee TH, Park EC. Effect of socioeconomic status on the linkage between suicidal ideation and suicide attempts. Suicide Life Threat Behav 2016 Oct;46(5):588-597. [doi: 10.1111/sltb.12242] [Medline: 26991513]
14. Lee HY, Hahm MI, Park EC. Differential association of socio-economic status with gender-and age-defined suicidal ideation among adult and elderly individuals in South Korea. Psychiatry Res 2013 Nov 30;210(1):323-328. [doi: 10.1016/j.psychres.2013.05.012] [Medline: 23769392]
15. Yoon CG, Jung J, Yoon JH, Lee D, Jeon H, Lee SY. How is the suicide ideation in the Korean armed forces affected by mental illness, traumatic events, and social support? J Korean Med Sci 2021 Apr 19;36(15):e96. [doi: 10.3346/jkms.2021.36.e96] [Medline: 33876585]
16. Russell D, Turner RJ, Joiner TE. Physical disability and suicidal ideation: a community - based study of risk/protective factors for suicidal thoughts. Suicide Life Threat Behav 2009 Aug;39(4):440-451. [doi: 10.1521/suli.2009.39.4.440]
17. Kim M, Jung W, Kim SY, Park JH, Shin DW. The Korea National Disability Registration System. Epidemiol Health 2023;45:e2023053. [doi: 10.4178/epih.e2023053] [Medline: 37189275]
18. Exchange rate. KOSIS. URL: https://tinyurl.com/3rrbztv2 [accessed 2025-06-20]
19. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002;16:321-357. [doi: 10.1613/jair.953]
20. Leading causes of death. Centers for Disease Control and Prevention. 2025. URL: https://www.cdc.gov/nchs/nvss/leading-causes-of-death.htm [accessed 2025-06-20]
21. Web-Based Injury Statistics Query and Reporting System. URL: https://wisqars.cdc.gov/ [accessed 2025-06-20]
22. Findlay L. Depression and suicidal ideation among Canadians aged 15 to 24. Health Rep 2017 Jan 18;28(1):3-11. [Medline: 28098916]
23. Lee SY, Atteraya MS. Depression, poverty, and abuse experience in suicide ideation among older Koreans. Int J Aging Hum Dev 2019 Jan;88(1):46-59. [doi: 10.1177/0091415018768256] [Medline: 29673253]
24. Kim H, Son Y, Lee H, et al. Machine learning-based prediction of suicidal thinking in adolescents by derivation and validation in 3 independent worldwide cohorts: algorithm development and validation study. J Med Internet Res 2024 May 17;26(1):e55913. [doi: 10.2196/55913] [Medline: 38758578]
25. Park SM. Suicidal ideation and its determinants in Korean adults: the role of physical activity and functional limitations. Psychol Health Med 2015;20(8):979-988. [doi: 10.1080/13548506.2014.986144] [Medline: 25434531]
26. Kim HJ, Jang SN, Lim JY. Service demand for and awareness of a primary healthcare pilot project for people with disabilities. J Korean Med Sci 2022;37(30):e241. [doi: 10.3346/jkms.2022.37.e241]

XSL•FO
RenderX

Abbreviations

AUROC: area under the receiver operating characteristic curve
NSPD: National Survey on Persons with Disabilities
SMOTE: Synthetic Minority Oversampling Technique
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
XGBoost: Extreme Gradient Boosting

Edited by A Mavragani; submitted 07.07.25; peer-reviewed by DK Yon, K Kavalidou; revised version received 19.11.25; accepted 20.11.25; published 12.12.25.

Please cite as:
Oh IH, Jo S, Lee J
Factors Associated With Suicidal Ideation Among Persons With Disabilities in South Korea: Retrospective Observational Study
JMIR Form Res 2025;9:e80261
URL: https://formative.jmir.org/2025/1/e80261
doi:10.2196/80261

© In-Hwan Oh, Sooyeon Jo, Joon Lee. Originally published in JMIR Formative Research (https://formative.jmir.org), 12.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Information Extraction of Doctoral Theses Using Two Different Large Language Models vs Health Services Researchers: Development and Usability Study

Jonas Cittadino, MD, Dr med; Pia Traulsen, MSc; Teresa Schmahl, MSc; Larisa Wewetzer, MSc, Dr rer nat; Julia Cummerow, MD; Kristina Flägel, MD, Dr med; Christoph Strumann, Dipl-Volksw, Dr sc pol; Katja Goetz, Dipl-Soz, Prof Dr; Jost Steinhäuser, Prof Dr med

Institute of Family Medicine, University Hospital Schleswig-Holstein, Maria-Goeppert-Straße 9a, Lübeck, Germany

Corresponding Author:
Jonas Cittadino, MD, Dr med
Institute of Family Medicine, University Hospital Schleswig-Holstein, Maria-Goeppert-Straße 9a, Lübeck, Germany

Abstract

Background: The Archive of German-Language General Practice (ADAM) stores about 500 paper-based doctoral theses published from 1965 to today. Although they have been grouped in different categories, no deeper systematic process of information extraction (IE) has been performed yet. Recently developed large language models (LLMs) like ChatGPT have been attributed the potential to help in the IE of medical documents. However, there are concerns about LLM hallucinations. Furthermore, there have not been reports regarding their usage in nonrecent doctoral theses yet.

Objective: The aim of this study is to analyze if LLMs can help to extract information from doctoral theses by using GPT-4o and Gemini-1.5-Flash for paper-based doctoral theses in ADAM.

Methods: We randomly selected 10 doctoral theses published between 1965 and 2022. After preprocessing, we used two different LLM pipelines, using models by OpenAI and Google. Pipelines were used to extract dissertation characteristics and generate uniform abstracts. Furthermore, one pooled human-generated abstract was written for comparison. Furthermore, blinded raters were asked to evaluate LLM-generated abstracts in comparison to the human-generated ones. Bidirectional encoder representations from transformers scores were calculated as the evaluation metric.

Results: Relevant dissertation characteristics and keywords could be extracted for all theses (n=10): institute name and location, thesis title, author name(s), and publication year. For all except one doctoral thesis, an abstract could be generated using GPT-4o, while Gemini-1.5-Flash provided abstracts in all cases (n=10). The modality of abstract generation showed no influence in raters' evaluation using the nonparametric Kruskal-Wallis test for independent groups (P=.44). The creation of LLM-generated abstracts was estimated to be 24-36 times faster than creation by humans. Evaluation metrics showed moderate-to-high semantic similarity (mean bidirectional encoder representations from transformers F_1-score, GPT-4o: 0.72 and Gemini: 0.71). Translation from German into English did not result in a loss of information (n=10).

Conclusions: An accumulating body of unpublished doctoral theses makes it difficult to extract relevant evidence. Recent advances in LLMs like ChatGPT have raised expectations in text mining, but they have not yet been used in the IE of "historic" medical documents. This feasibility study suggests that both models (GPT-4o and Gemini-1.5-Flash) helped to accurately simplify and condense doctoral theses into relevant information, while LLM-generated abstracts were perceived as similar to human-generated ones, were semanticly similar, and took about 30 times less time to create. This pilot study demonstrates the feasibility of a regular office-scanning workflow and use of general-purpose LLMs to extract relevant information and produce accurate abstracts from ADAM doctoral theses. Taken together, this information could help researchers to better search the family medicine scientific literature over the last 60 years, helping to develop current research questions.

(JMIR Form Res 2025;9:e77707) doi:10.2196/77707

KEYWORDS

ChatGPT; family medicine; doctoral thesis; GPT-4o; Gemini; artificial intelligence; AI

Introduction

By their nature, archives contain a large amount of information. Thirty years ago, the German Society of General Practitioners and Family Physicians (DEGAM) began to centrally store all doctoral theses of the specialty. This collection is now part of the Archive of German-Language General Practice (ADAM) [1]. ADAM was primarily established to gather historical

XSL•FO
RenderX

documents from the beginning of family medicine as a specialty in German-speaking countries (Germany, Austria, and Switzerland) and now stores a large number of documents of different types and formats, including many documents written with a typewriter [2]. Lately, there have been efforts to categorize 802 available theses [3,4]. During this process, one finding was that it is hard to find detailed information, especially for theses from before 2000. This may be because an estimated 50% of all dissertations are not published in a journal but rather are paper-based only [4]. Within ADAM, this is true for 553 dissertations. Therefore, as yet, there is no process for extracting relevant information, creating uniform abstracts, and making them available to the public and other researchers.

With the release of large language models (LLMs) such as ChatGPT by OpenAI and Gemini by Google, hopes have been raised regarding their use for medicine and medical documents [5]. The reduction of hallucination, the term for information made up by an LLM, is crucial for medical scenarios and has been the subject of a recent project [6]. Although previous work on LLMs has shown effectiveness in terms of information extraction (IE) from scientific texts [7,8], to our knowledge, there has not yet been a reported use of LLMs regarding IE from doctoral theses. Hence, the aim of this feasibility study was to analyze if LLMs and natural language processing can extract relevant information and generate uniform abstracts of doctoral theses from the field of family medicine in ADAM and how these artificial intelligence–generated abstracts are perceived by scientists.

Methods

Dissertation Characteristics

For this analysis, we randomly selected 10 dissertations from ADAM, ranging from the earliest dissertation (1965) to the most recent one (2022). All dissertations were in PDF format, either made available by the original author or generated by study personnel who scanned paper-based dissertations using a regular multifunctional office printer. When making our selection, we included different time periods of submission and both types of digitization in equal numbers. Therefore, we focused on dissertations from different time periods and different methods of digitization. Although the digitization of documents is a current fundamental question in archive research, only a few standards exist [9].

Ethical Considerations

Since all selected dissertations were publicly accessible through ADAM, a formal ethics review was deemed unnecessary. However, informed consent was obtained from all copyright holders of the selected dissertations. According to our institutional practice (University Hospital Schleswig-Holstein), projects based solely on publicly available documents with author consent do not require review by the ethics committee. This analysis followed the transparent reporting of a multivariable model for individual prognosis or diagnosis (TRIPOD)+LLM guidelines (see Multimedia Appendix 1) [10].

Analysis

For each PDF file, we created an analytical pipeline including the following steps: preprocessing, extracting information from the title page, creating a uniform abstract, and translating newly generated abstracts into English. We evaluated two LLMs in this study: GPT-4o [11] by OpenAI and Gemini-1.5-Flash [12] by Google.

GPT-4o

Preprocessing of all files in the GPT-4o pipeline was done by bringing all files into a vertical layout format before they were rotated and/or divided depending on text alignment (Figure 1). If the extracted text resulted in nonsense, we used optical character recognition (OCR) [13] for extraction of all pages. Then, the content of each page was stored as a string in a dictionary.

For further analysis, we used LLMs by OpenAI, namely text-davinci-003, GPT-3.5-turbo, and GPT-4o [11], using a temperature [14] of 0.2 to reduce hallucination. All models are pretrained and transformer-based, meaning that the model supervised its learning itself and understands context [15].

We defined the first page containing a 4-digit number as the title page. Using only the title page of each file, publication characteristics were extracted by using the OpenAI model “text-davinci-003”: name of the institution, city of the institution, director of the institution, first and last name(s) of the author, title of the dissertation, origin of the author, and year of publication (Figure 2). For each dissertation, J Cittadino cross-checked if the information extracted by the model was correct.

XSL•FO
RenderX

Figure 1. Preprocessing of scanned dissertation title pages prior to analysis. Workflow of preprocessing steps applied to scanned dissertation title pages from German medical faculties. The process included rotation correction, text segmentation, and optical character recognition to prepare documents for automated metadata extraction.

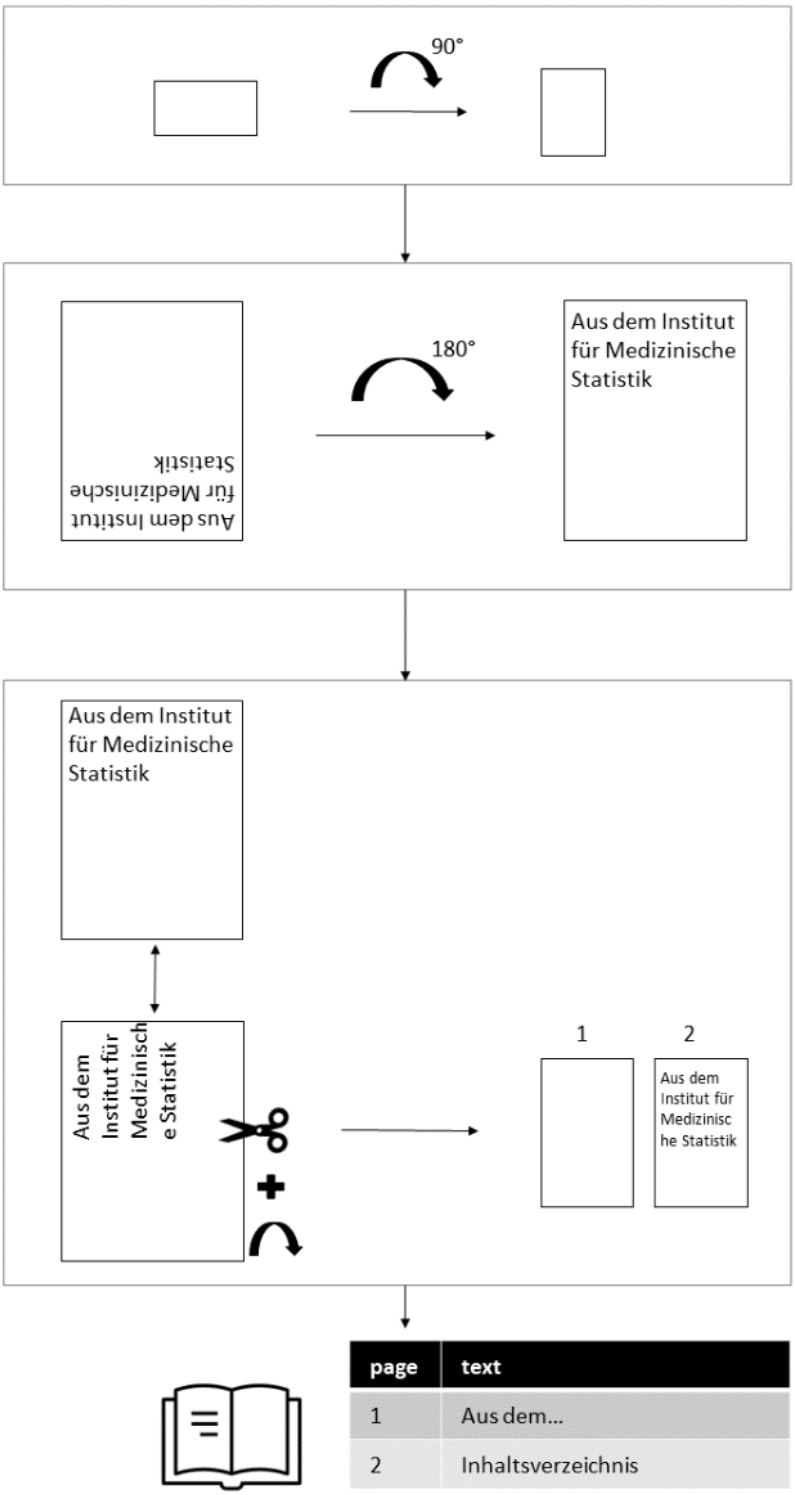

XSL•FO
RenderX

Figure 2. Automated extraction of dissertation metadata using GPT models. Example of structured metadata extraction from a dissertation title page (University of Lübeck, Institute of Family Medicine, 2022) using GPT models. The extracted fields include institute name, location, name of director, title, author information, and year of publication.

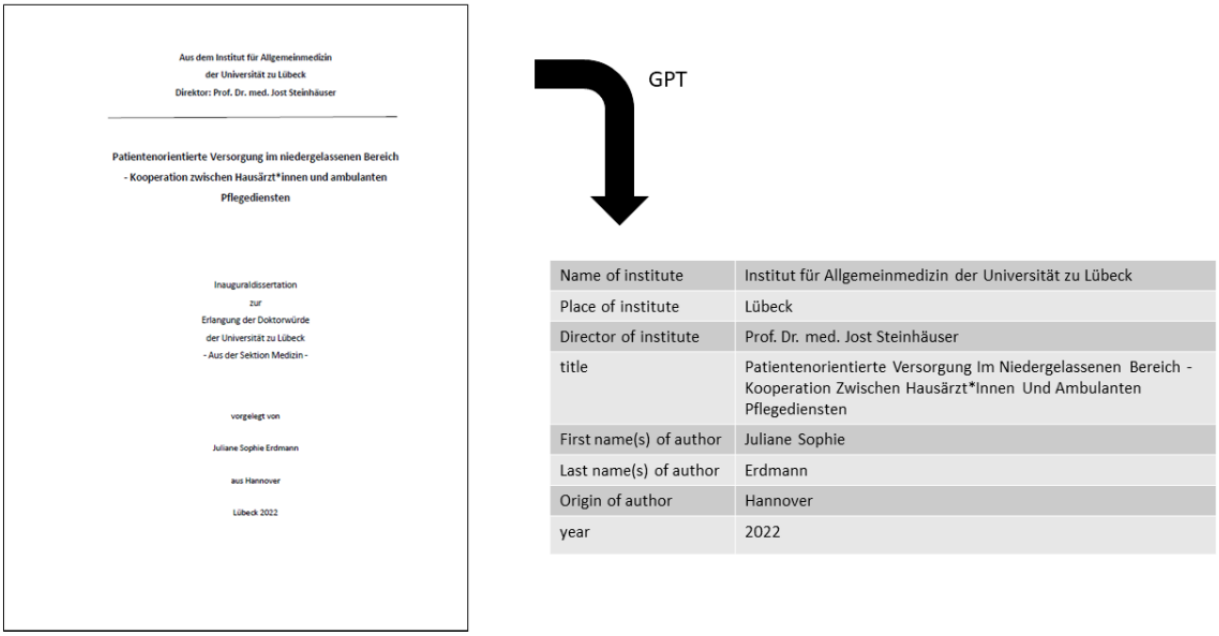

Gemini

In the Gemini pipeline, we used Google Cloud Platform for preprocessing and text generation. Using Workflows, Cloud Function, and Gemini-1.5-Flash with a temperature of 0.2, we designed the following pathway. First, DocumentAI was used to extract the text from all PDFs and store it in a JSON file [16]. Only using the first page and Gemini-1.5-Flash, we then extracted the same information from each PDF as with GPT-4o: name of the institution, city of the institution, director of the institution, first and last name(s) of the author, title of the dissertation, origin of the author, and year of publication.

Common Pathway

Next, for each LLM, we generated a string containing the whole text of the dissertation after the table of contents. The page number after the end of the table of contents was extracted by looping through snippets of the first 10 pages of each document using text-davinci-003 or Gemini-1.5-Flash.

Continuing the analytical pipeline, we searched the generated string for a summary section of the whole thesis. If found, the extracted paragraph was stored as *abstract_whole*. If not found, the text was scanned for paragraphs containing the subsections Aim of the study, Methods, Results, Discussion, and Conclusion as headings, which were extracted and summarized using text-davinci-003 or Gemini-1.5-Flash if found.

If a summary paragraph was identified, we shortened it to the first 4000 characters and used GPT-4o and Gemini-1.5-Flash to generate uniform abstracts containing the subheadings Aim of the study, Methods, Results, and Discussion. For both models, we used the same prompt and hyperparameters:

- Prompt: “Summarize the following summary section of a medical doctoral thesis into sections with the headings Objective, Methods, Results, and Discussion, each in no more than 2 - 3 concise sentences: [summary section]” (German: *“Fasse die folgende Zusammenfassung einer medizinischen Doktorarbeit in Abschnitte mit den Überschriften Zielsetzung, Methodik, Ergebnisse und Diskussion knapp in jeweils maximal 2-3 Sätzen zusammen: [Zusammenfassung].”*
- Temperature: 0.2.
- Maximum number of tokens: 950.

The time needed for completion of the script as well as usage costs of the OpenAI LLMs were assessed.

Abstract Comparison

For each dissertation, researchers (PT, TS, J Cummerow, and LW) independently drafted an abstract and then agreed on a pooled version by discussion. The result was a third version of the abstract, this one human-generated. Afterward, the researchers estimated their time needed for drafting and agreeing on a pooled version. Then, three senior researchers (KG, KF, and CS), were asked to evaluate all three abstract versions in different orders, not knowing which two were LLM-generated and which one was human-generated. For each abstract version of each dissertation, we collected two ratings using German school marks between 1 (=best) and 6 (=worst). We then analyzed these ratings using descriptive statistics and a robust analysis procedure, the nonparametric Kruskal-Wallis test for independent groups [17]. Results were visualized in a boxplot.

Text Generation Evaluation

Words were counted for each generated abstract. To automatically evaluate the performance of LLM-generated

XSL•FO
RenderX

abstracts, we used BERTScore (bidirectional encoder representations from transformers) [18] for contextual embeddings. This method uses the pretrained language model BERT to compare the similarity between tokens by calculating precision, recall, and the harmonic mean F_1 [19]. All analysis processes were done using Python (version 3.10.1; Python Software Foundation) [20]. The code is available upon request.

Translation Evaluation

Finally, we used GPT-4o and Gemini-1.5-Flash to translate the newly generated abstracts into English. Following World Health Organization guidelines for translations [21] to make sure that the translation process did not result in the loss of information, we retranslated the generated English translation back into German again using the same LLM. Then, generated German translations were compared to the originally generated abstract (by J Cittadino).

Results

Overview

Paper-based documentation was more common (n=7) and often showed apparently poor digital quality after scanning. However, using the GPT-4o pipeline, we were able to extract text and information from all dissertations. Even so, we observed some common mistakes—for example, the German letter “ü” was often transformed to “ii.” Interestingly, while generating an abstract, both GPT models could still correctly understand the content and remove spelling mistakes.

Thesis Characteristics

Information from the title page could be extracted for all theses, and the details were accurate except for two spelling mistakes (“é” instead of “ö”). There was no difference between OpenAI models and Google models. Dissertations covered various topics of family medicine research completed between 1965 and 2022 (Table 1).

XSL•FO
RenderX

Table . GPT-4o–generated characteristics of the included dissertations.

Name of institute	Institute location	Director of institute	Title of doctoral thesis	First name(s) of author	Last name(s) of author	Origin of author	Year(s)
Institut für Medizinische Statistik der Freien Universität Berlin [Institute of Medical Statistics, Free University of Berlin]	Berlin	Prof Dr med, Dr phil Karl Freudenberg	*Beziehungen zwischen Einweisungsdiagnosen und klinischen Diagnosen* [Relationships between admission diagnoses and clinical diagnoses]	Helmut	Pillau	Berlin	1965
Universität Ulm (MNH) [University of Ulm (MNH)]	Ulm	Priv Doz, Dr S Haussler	*Entwicklung kassenärztlicher Leistungen bei verschiedenen Arztgruppen im K. V. Bereich Nordwürttemberg (1965 - 1969)* [Development of services provided by contracted physicians in various medical groups in the North Württemberg Association of Statutory Health Insurance Physicians (1965 - 1969)]	Klaus	Besel	Vohringen/Iller	1965 - 1969
Institut für medizinische Statistik und Dokumentation [Institute of Medical Statistics and Documentation]	Kiel	Prof Dr med G Griesser	*Analyse einer Allgemeinpraxis* [Analysis of a general practice]	Andreas	Kernbichler	Meldorf	1973
Albert-Ludwigs-Universität Freiburg im Breisgau [Albert Ludwigs University of Freiburg]	Freiburg im Breisgau	Honorarprofessor Dr med HH Schrömbgens	*Die Lehre der Allgemeinmedizin an den deutschen Hochschulen (1966 - 1978)* [Teaching of general practice at German universities (1966 - 1978)]	Claudia	Keller-Röll	Deggingen	1979

XSL•FO
RenderX

Name of institute	Institute location	Director of institute	Title of doctoral thesis	First name(s) of author	Last name(s) of author	Origin of author	Year(s)
Medizinische Poliklinik der Universität München [Medical Outpatient Department, University of Munich]	München	Prof Dr N Zéllner	*Untersuchung zur Compliance bei der Diagnostik und Therapie des arteriellen Hypertonus in einer Hausarztpraxis in Stadtrandlage* [Study on compliance in the diagnosis and treatment of arterial hypertension in a suburban family practice]	Manfred	Lohnstein	Augsburg	1983
Institut für Allgemeinmedizin [Institute of General Practice]	Frankfurt AM Main	Prof Dr K Jork	*Selbstmedikation bei älteren Menschen und Studenten* [Self-medication among older adults and university students]	Elke	Iburg	Hannover	1991
Philipps-Universität Marburg [Philipps University of Marburg]	Marburg	Frau Prof Dr med Erika Baum	*Allgemeinmedizin in Großbritannien und Deutschland: Die Auswirkungen verschiedener Vergütungssysteme auf die Qualität präventiver Versorgung* [General practice in Great Britain and Germany: the effects of different remuneration systems on the quality of preventive care]	Norbert	Donner-Banzhoff	Viersen/Rheinland	1993
Zentrum der Gesundheitswissenschaften Institut für Allgemeinmedizin [Center for Health Sciences, Institute of General Practice]	Frankfurt AM Main	Prof Dr Ferdinand M Gerlach	*Möglichkeiten durch Delegation hausärztlicher Leistungen am Beispiel von Versorgungsassistentinnen in der Hausarztpraxis (Verah)* [Opportunities through delegation of family physician tasks using the example of practice assistants in family medicine (VERAH)]	Karola	Mergenthal	Büdingen, Hessen	2016

XSL•FO
RenderX

Name of institute	Institute location	Director of institute	Title of doctoral thesis	First name(s) of author	Last name(s) of author	Origin of author	Year(s)
Institut für Allgemeinmedizin der Universität zu Lübeck [Institute of Family Medicine, University of Lübeck]	Lübeck	Prof Dr med Steinhäuser	*Nasa-Task Load Index - Ein Instrument, um sich der Komplexität von Beratungsanlässen in der Allgemeinmedizin zu nähern* [NASA Task Load Index: a tool for approaching the complexity of consultation situations in general practice]	Britta	Galler	Stade	2020
Institut für Allgemeinmedizin der Universität zu Lübeck [Institute of Family Medicine, University of Lübeck]	Lübeck	Prof Dr med Jost Steinhäuser	*Patientenorientierte Versorgung im Niedergelassenen Bereich - Kooperation zwischen Hausärzt*Innen und ambulanten Pflegediensten* [Patient oriented care in the outpatient sector: cooperation between family physicians and ambulatory nursing services]	Juliane Sophie	Erdmann	Hannover	2022

Abstract Summary

Paragraphs containing a summary section of the whole dissertation could be found in 9 of 10 documents. In the case where no summary was found, a paragraph was extracted from the Discussion subsection and summarized using GPT-4o, while Gemini-1.5-Flash was able to provide a whole abstract. Thus, GPT-4o generated abstracts in 9 of 10 cases and Gemini-1.5-Flash in all 10 cases.

In the text extraction, we observed some common mistakes—for example, the German letter “ü” was often transformed to “ii.” The GPT model could still correctly understand the content and remove spelling mistakes while generating an abstract.

Running the script a second time resulted in the same extraction of all dissertation characteristics in 9 of 10 cases for GPT-4o and in all cases for Gemini-1.5-Flash. In one case, a different institute director was found, which was extracted as “Korefferent” and not “Referent.” Generated abstracts, however, differed in their wording while containing the same information.

Running the full script using the GPT-4o pipeline lasted 45.38 minutes (2720 s) and cost US $2.67. Additionally, we estimated a scanning time of 5-10 minutes for each dissertation. Each researcher spent about 1.5 hours drafting an abstract and about 0.5 hours while creating the pooled version.

Abstract Comparison

For each modality of abstract generation—GPT-4o, Gemini-1.5-Flash, or human-generated—two scores from different rater were collected. Scores covered the whole range from 1 to 6 for all modalities. Although there was no statistical difference in the nonparametric Kruskal-Wallis test for independent groups (P=.44), GPT-4o showed the best mean rating (2.44), while human-generated (3.00) and Gemini-1.5-Flash (3.25) were evaluated lower (Figure 3). Standard deviations (SD) were in the same range for all modalities (GPT-4o: 1.42; Gemini-1.5-Flash: 1.71; human-generated: 1.17).

XSL•FO
RenderX

Figure 3. Distribution of abstract quality ratings across GPT models and human raters. Comparison of abstract quality ratings (1=best, 6=worst) generated by 3 sources (GPT-4o, Gemini-1.5-Flash, and human evaluators) for a sample of German medical dissertations. Boxplots show median, interquartile range, and variability of the assigned scores.

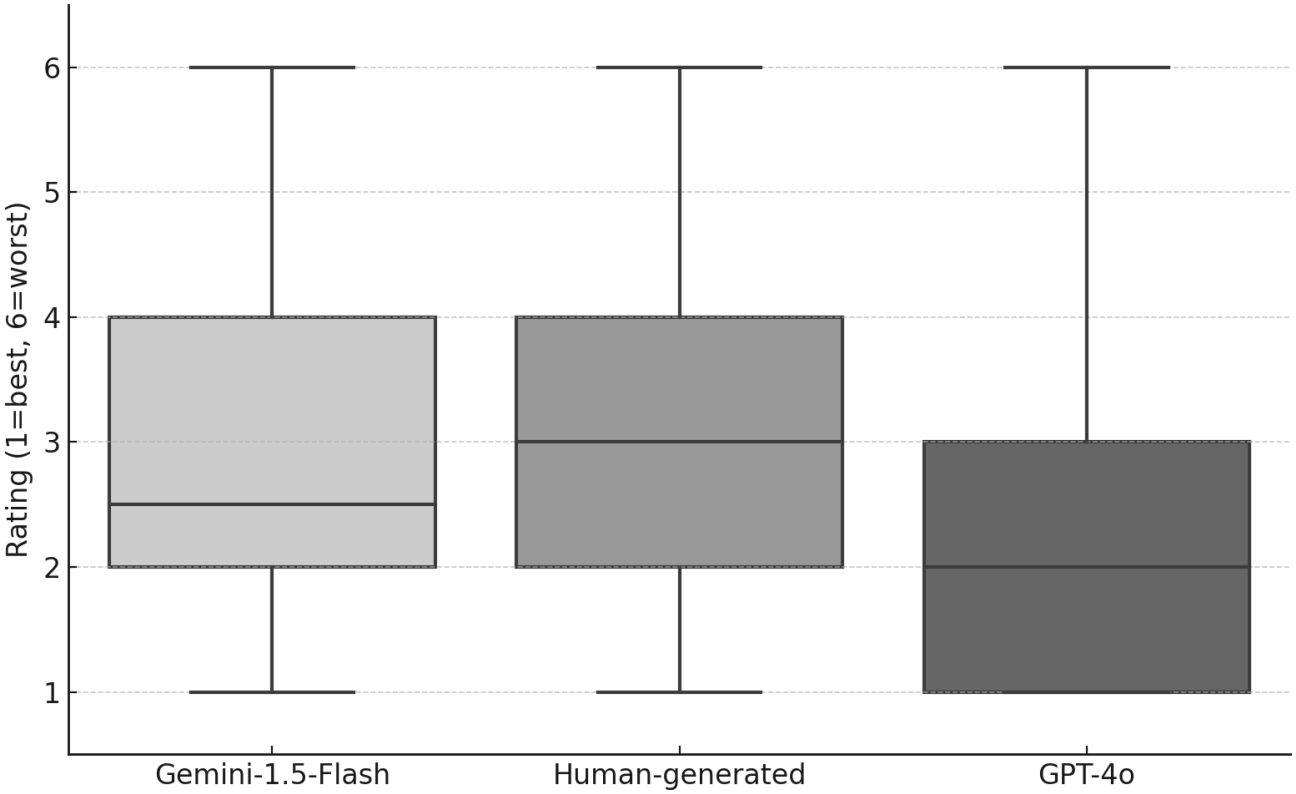

Text Generation Evaluation

LLM-generated abstracts consisted of fewer words compared to human-generated ones (mean word counts were 141 for GPT-4o, 137 for Gemini, and 352 for human-generated abstracts). Mean BERTScores showed moderate-to-high congruence between LLMs and human-generated abstracts (mean F_1: GPT-4o: 0.72 and Gemini: 0.71, respectively; Table 2).

Table . BERTScore for the 10 dissertations for GPT-4o and Gemini.[a]

Dissertation ID	GPT-4o BERTScore	Gemini BERTScore
1	0.0	0.725
2	0.685	0.709
3	0.71	0.698
4	0.728	0.718
5	0.691	0.7
6	0.723	0.731
7	0.719	0.649
8	0.758	0.715
9	0.73	0.714
10	0.731	0.727
All	0.72	0.707

[a]BERT: bidirectional encoder representations from transformers.

Translation Evaluation

When abstracts generated by both LLMs were translated into English and then back to German by using GPT-4o, the retranslated new German abstracts did not show a loss of information in comparison to the original German abstracts, although the wording was not exactly the same in all 10 cases.

XSL•FO
RenderX

Discussion

Main Findings

Results from this feasibility study suggest that LLMs can be used to extract relevant general characteristics from doctoral theses. Further, the models (GPT-4o and Gemini-1.5-Flash) accurately generated uniform abstracts 90% and 100% of the time, respectively. Machine-based translation of these abstracts into English did not show a loss of information. When asked, other researchers perceived these LLM-generated abstracts as similar to human-generated abstracts. BERTScores showed moderate-to-high similarity between LLM- versus human-generated abstracts. However, as hallucinations could not be totally eliminated, LLMs provide a tool to explore but not analyze medical dissertations.

What Is Already Known

Medical knowledge is growing rapidly, and it is close to impossible to keep up with new insights and information [22,23]. Although health care systems currently use both paper-based and electronic health records [24], OCR can help to precisely extract text from scanned documents [13]. Lastly, accumulating evidence has showed that previously used rule-based or machine learning–based IE have been outperformed by deep neural networks even if pretrained, and they can help process a lot of data [25,26]. Although new, publicly available LLMs like ChatGPT and Gemini could possibly help in IE of medical documents, although ethical concerns have been mentioned [5,27]. Further, many projects have examined how and how well LLMs can help in IE of medical documents [28-32].

Although Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [33] for overlapping unigrams is a known metric for the evaluation of text generation, we choose the BERTScore as our evaluation metric. BERTScore captures the contextual embedding and was thus more suitable to our study aim to examine how well LLMs can capture the content of long dissertations and generate an abstract including the relevant information [18]. The observed LLM performance in summarizing dissertations in our study was similar to reported F_1 BERTScores in a Dutch study examining the differences in note-taking between humans and a digital scribe in a clinical workflow using various GPT models [34].

What This Study Adds

To our knowledge, this is the first study examining the usage of GPT-4o and Gemini-1.5-Flash for IE in German doctoral theses. We actively chose a feasible method of digitization by using a standard multifunctional office printer, which is not in line with national standards of archiving documents, to simulate everyday conditions [9]. Our results suggest, however, that standardized IE through LLMs can be used to create uniform abstracts for dissertations that otherwise would remain difficult to access. Interestingly, these abstracts do not seem to be inferior to human-generated ones, indicated by the ratings of senior researchers. By costing as little as about US $0.25 and taking only 10-15 minutes (scanning: 5 - 10 min, creation: 5 min) on average for each thesis, past results could be made more accessible and therefore easier to use for further research by using this analytical pipeline. This is especially true when considering that the creation of one pooled human-generated abstract took about 6 hours (360 min) and did not result in a higher rating from senior researchers, making it 24-36 times slower and more expensive.

A Dutch study examining note-taking performance reported that LLM-generated texts consisted of more words than human-generated texts (137 vs 101 words) [34]. However, we observed that human-generated abstracts were about 2.5 times longer than the ones from LLMs. This might be due to the need not to miss any important information when writing an abstract as a human and therefore generating a longer text. Further, LLMs might have difficulties capturing the whole extent of a dissertation, thus generating a shorter text. Two dissertations with low F_1 BERTScores (IDs 2 and 5) were both written on typewriters, which might have resulted in informational loss during the OCR process, thus possibly accounting for a lower evaluation score. Conversely, we observed that dissertations written in Microsoft Word and converted to PDF achieved higher F_1 BERTScores (IDs 8 and 10).

Hallucinations in medical LLMs persist [6], so users should be cautious about fully trusting the results. At this scientific stage, we propose using LLMs for medical dissertations as a tool for exploring—for example, for abstract generation—but recommend verifying generated information. By using this tool, we can unlock the rich data from doctoral theses in ADAM, making this valuable knowledge publicly accessible and strengthening up-to-date research.

Strengths and Limitations

For this study, we only presented a small dataset including 10 doctoral theses of which the results can only be used for hypothesis generation and do not claim to be generalizable. Although we included files from 1965 and 2022, representing examples of the oldest and most recent documents in ADAM, further studies should include more dissertations to be more representative.

By using two recent LLMs (GPT-4o and Gemini-1.5-Flash), we could directly compare their respective results and performance. In this pilot study, however, we did not include additional models, which should be a focus of further studies. Additionally, because of the small sample size, we did not calculate interrater reliability among senior researchers, which might have biased ratings results. Although BERTScores have been effectively used in evaluating text generation by LLMs, they lack an understanding of nuances in medical language and therefore should be used with caution [35]. This underlines the need for an evaluation metric specifically to be used in medical contexts as potential threats from misinformation are among the most discussed risks of LLM usage [36].

Conclusions

In this feasibility study using 10 medical dissertations completed between 1965 and 2022, we provide a proof of concept for the usage of LLMs, namely GPT-4o and Gemini-1.5-Flash, for IE from theses in ADAM. Within the constraints of a small, intentionally heterogeneous sample, results indicate promising feasibility. Using these models, typical publication

XSL•FO
RenderX

characteristics as well as uniform abstracts could be generated, and they were not perceived differently to human-generated abstracts, while being about 30 times faster to generate.

However, as LLM hallucinations persist, research applications of LLMs should be approached with caution. Further research should include more dissertations, which will help researchers understand trends in various time periods and possibly allow for the grouping and summarizing of similar dissertations.

Acknowledgments

Neither ChatGPT nor any other large language model were used during the writing or editing of this article.

Funding

Funding was provided by the *Foundation Perspektive Hausärztinnen und Hausärzte* of the German General Practitioners' Association (German: *Stiftung Perspektive Hausarzt des Deutschen Hausärztinnen- und Hausärzteverbands e.V.*).

Data Availability

Data will be made available from the corresponding author upon a reasonable request.

Authors' Contributions

Conceptualization: J Cittadino, JS; methodology: J Cittadino, JS; validation: PT, TS, LW, J Cummerow, KG, CS, KF; formal analysis: J Cittadino; investigation: J Cittadino, JS; resources: JS; data curation: JS; writing— original draft preparation: J Cittadino; writing—review and editing: PT, TS, LW, J Cummerow, KG, CS, KF, JS; visualization: J Cittadino; supervision: JS; project administration: J Cittadino; funding acquisition: JS. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
TRIPOD+LLM checklist.
[PDF File, 176 KB - formative_v9i1e77707_app1.pdf]

References

1. Mader FH. Von Der Allgemeinen Medizin Zur Allgemeinmedizin: Festschrift 50 Jahre DEGAM: 1966-2016 2016. URL: https://www.degam.de/files/inhalt/pdf/festschrift_2016/degam_jubilaeum_festschrift.pdf [accessed 2025-11-30]
2. Zwierlein R. ADAM – the newborn archive of general practice in Germany. 2019. URL: http://dmhs1917.dk/wp-content/uploads/2019/06/Archive-General-Practice-Germany.pdf [accessed 2025-10-15]
3. Sosna N, Steinhäuser J. : German Medical Science GMS Publishing House; 2022 Presented at: 56 Kongress für Allgemeinmedizin und Familienmedizin; Sep 15-17, 2022; Greifswald, Germany URL: https://www.egms.de/static/de/meetings/degam2022/22degam202.shtml [accessed 2025-11-30] [doi: 10.3205/22degam202]
4. Sosna N, Steinhäuser J. Exploring general practice research in Germany: a systematic review of dissertation topics from 1965-2023. Scand J Prim Health Care 2024 Sep;42(3):393-398. [doi: 10.1080/02813432.2024.2329213] [Medline: 38488125]
5. Janssen BV, Kazemier G, Besselink MG. The use of ChatGPT and other large language models in surgical science. BJS Open 2023 Mar 7;7(2):zrad032. [doi: 10.1093/bjsopen/zrad032] [Medline: 36960954]
6. Kim Y, Jeong H, Chen S, et al. Medical hallucinations in foundation models and their impact on healthcare. arXiv. Preprint posted online on Feb 26, 2025. [doi: 10.48550/arXiv.2503.05777]
7. Fornasiere R, Brunello N, Scotti V, Carman M, Freihat AA. Medical information extraction with large language models. In: Freihat AA, editor. 2024 Presented at: Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024) p. 456-466 URL: https://aclanthology.org/2024.icnlsp-1.47/ [accessed 2025-11-27]
8. Dagdelen J, Dunn A, Lee S, et al. Structured information extraction from scientific text with large language models. Nat Commun 2024 Feb 15;15(1):1418. [doi: 10.1038/s41467-024-45563-x] [Medline: 38360817]
9. DFG-Vordruck 12151–12/16 – Praxisregeln "Digitalisierung." [Website in German]. Deutsche Forschungsgemeinschaft. 2016. URL: https://www.dfg.de/resource/blob/176108/12-151-v1216-de.pdf [accessed 2025-10-15]
10. Gallifant J, Afshar M, Ameen S, et al. The TRIPOD-LLM reporting guideline for studies using large language models. Nat Med 2025 Jan;31(1):60-69. [doi: 10.1038/s41591-024-03425-5] [Medline: 39779929]
11. OpenAI, Hurst A, Lerer A, Goucher AP, et al. GPT-4o system card. arXiv. Preprint posted online on Oct 25, 2024. [doi: 10.48550/arXiv.2410.21276]
12. Team G, Georgiev P, Lei VI, et al. Gemini 1.5: unlocking multimodal understanding across millions of tokens of context. arXiv. Preprint posted online on Mar 1, 2024. [doi: 10.48550/arXiv.2403.05530]

XSL•FO
RenderX

13. Hom J, Nikowitz J, Ottesen R, Niland JC. Facilitating clinical research through automation: Combining optical character recognition with natural language processing. Clin Trials 2022 Oct;19(5):504-511. [doi: 10.1177/17407745221093621] [Medline: 35608136]
14. Beutel G, Geerits E, Kielstein JT. Artificial hallucination: GPT on LSD? Crit Care 2023 Apr 18;27(1):148. [doi: 10.1186/s13054-023-04425-6] [Medline: 37072798]
15. Chandra A, Tünnermann L, Löfstedt T, Gratz R. Transformer-based deep learning for predicting protein properties in the life sciences. Elife 2023 Jan 18;12:e82819. [doi: 10.7554/eLife.82819] [Medline: 36651724]
16. Pezoa F, Reutter JL, Suarez F, Ugarte M, Vrgoč D. Foundations of JSON schema. 2016 Apr 11 Presented at: WWW '16; Apr 11-15, 2016. [doi: 10.1145/2872427.2883029]
17. Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. J Am Stat Assoc 1952 Dec;47(260):583-621. [doi: 10.1080/01621459.1952.10483441]
18. Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. BERTScore: evaluating text generation with BERT. 2020. URL: https://openreview.net/forum?id=SkeHuCVFDr [accessed 2025-10-15]
19. Christen P, Hand DJ, Kirielle N. A review of the F-Measure: its history, properties, criticism, and alternatives. ACM Comput Surv 2024 Mar 31;56(3):1-24. [doi: 10.1145/3606367]
20. Rossum GV, Drake FL. Python 3 reference manual. In: CreateSpace 2009.
21. World Health Organization. Translation and cultural adaptation of health questionnaires. : Geneva: World Health Organization; 2016 URL: https://applications.emro.who.int/imemrf/J_Pak_Med_Assoc/J_Pak_Med_Assoc_2003_53_4_142_147.pdf [accessed 2025-11-30]
22. Densen P. Challenges and opportunities facing medical education. Trans Am Clin Climatol Assoc 2011;122:48-58. [Medline: 21686208]
23. Landhuis E. Scientific literature: information overload. Nature New Biol 2016 Jul 21;535(7612):457-458. [doi: 10.1038/nj7612-457a] [Medline: 27453968]
24. Adane K, Gizachew M, Kendie S. The role of medical data in efficient patient care delivery: a review. Risk Manag Healthc Policy 2019;12:67-73. [doi: 10.2147/RMHP.S179259] [Medline: 31114410]
25. Landolsi MY, Hlaoua L, Ben Romdhane L. Information extraction from electronic medical documents: state of the art and future research directions. Knowl Inf Syst 2023;65(2):463-516. [doi: 10.1007/s10115-022-01779-1] [Medline: 36405956]
26. Hahn U, Oleynik M. Medical information extraction in the age of deep learning. Yearb Med Inform 2020 Aug;29(1):208-220. [doi: 10.1055/s-0040-1702001] [Medline: 32823318]
27. Harrer S. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. EBioMedicine 2023 Apr;90:104512. [doi: 10.1016/j.ebiom.2023.104512] [Medline: 36924620]
28. Hsu E, Roberts K. LLM-IE: a Python package for biomedical generative information extraction with large language models. JAMIA Open 2025 Apr;8(2):ooaf012. [doi: 10.1093/jamiaopen/ooaf012] [Medline: 40078164]
29. Wiest IC, Wolf F, Leßmann ME, et al. LLM-AIx: an open source pipeline for Information Extraction from unstructured medical text based on privacy preserving large language models. medRxiv. 2024 Sep 3 p. 2024.09.02.24312917. [doi: 10.1101/2024.09.02.24312917] [Medline: 39281753]
30. Sciannameo V, Pagliari DJ, Urru S, et al. Information extraction from medical case reports using OpenAI InstructGPT. Comput Methods Programs Biomed 2024 Oct;255:108326. [doi: 10.1016/j.cmpb.2024.108326] [Medline: 39029416]
31. Schopow N, Osterhoff G, Baur D. Applications of the natural language processing tool ChatGPT in clinical practice: comparative study and augmented systematic review. JMIR Med Inform 2023 Nov 28;11:e48933. [doi: 10.2196/48933] [Medline: 38015610]
32. Huang J, Yang DM, Rong R, et al. A critical assessment of using ChatGPT for extracting structured data from clinical notes. NPJ Digit Med 2024 May 1;7(1):106. [doi: 10.1038/s41746-024-01079-8] [Medline: 38693429]
33. Lin CY. ROUGE: a package for automatic evaluation of summaries. 2004 Presented at: Text Summarization Branches Out; Jul 21-26, 2004; Barcelona, Spain p. 74-81 URL: https://aclanthology.org/W04-1013/ [accessed 2025-10-15]
34. van Buchem MM, Kant IMJ, King L, Kazmaier J, Steyerberg EW, Bauer MP. Impact of a digital scribe system on clinical documentation time and quality: usability study. JMIR AI 2024 Sep 23;3:e60020. [doi: 10.2196/60020] [Medline: 39312397]
35. Jung H, Kim Y, Choi H, et al. Enhancing clinical efficiency through llm: discharge note generation for cardiac patients. arXiv. Preprint posted online on Apr 8, 2024. [doi: 10.48550/arXiv.2404.05144]
36. Preiksaitis C, Ashenburg N, Bunney G, et al. The role of large language models in transforming emergency medicine: scoping review. JMIR Med Inform 2024 May 10;12:e53787. [doi: 10.2196/53787] [Medline: 38728687]

Abbreviations

ADAM: *Archiv der deutschsprachigen Allgemeinmedizin* (English: Archive of German-language General Practice)
BERT: bidirectional encoder representations from transformers
DEGAM: German Society of General Practitioners and Family Physicians
IE: information extraction
LLM: large language model

XSL•FO
RenderX

OCR: optical character recognition
ROUGE: Recall-Oriented Understudy for Gisting Evaluation
TRIPOD: transparent reporting of a multivariable model for individual prognosis or diagnosis

Edited by A Mavragani; submitted 18.05.25; peer-reviewed by C Zhang, Z Zhang; accepted 23.10.25; published 10.12.25.

Please cite as:
Cittadino J, Traulsen P, Schmahl T, Wewetzer L, Cummerow J, Flägel K, Strumann C, Goetz K, Steinhäuser J
Information Extraction of Doctoral Theses Using Two Different Large Language Models vs Health Services Researchers: Development and Usability Study
JMIR Form Res 2025;9:e77707
URL: https://formative.jmir.org/2025/1/e77707
doi:10.2196/77707

© Jonas Cittadino, Pia Traulsen, Teresa Schmahl, Larisa Wewetzer, Julia Cummerow, Kristina Flägel, Christoph Strumann, Katja Goetz, Jost Steinhäuser. Originally published in JMIR Formative Research (https://formative.jmir.org), 10.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Everyday Digital Support to Promote Health and Literacy Among Older Adults: 14-Week Randomized Digital Pilot Trial by Engagement Level

Andressa Crystine da Silva Sobrinho[1], MSc; Guilherme da Silva Rodrigues[2], PhD; Guilherme Lima de Oliveira[3], BSc; Grace Angélica de Oliveira Gomes[4], PhD; Carlos Roberto Bueno Júnior[3], PhD

[1]Department of Clinical Medicine, Ribeirão Preto Medical School, University of São Paulo (FMRP-USP), 3900 Avenida Bandeirantes, Ribeirão Preto, São Paulo, Brazil

[2]Ribeirão Preto Medical School, University of São Paulo (FMRP-USP), Ribeirão Preto, São Paulo, Brazil

[3]Department of Clinical Medicine, Ribeirão Preto Medical School, University of São Paulo (FMRP-USP), Ribeirão Preto, São Paulo, Brazil

[4]Department of Gerontology, Federal University of São Carlos (UFSCar), São Carlos, São Paulo, Brazil

Corresponding Author:
Andressa Crystine da Silva Sobrinho, MSc
Department of Clinical Medicine, Ribeirão Preto Medical School, University of São Paulo (FMRP-USP), 3900 Avenida Bandeirantes, Ribeirão Preto, São Paulo, Brazil

Abstract

Background: While digital health solutions are becoming increasingly sophisticated, simple forms of everyday digital support may offer underexplored opportunities to promote health among older adults. However, evidence remains scarce on whether such teleassistance-based approaches can effectively enhance health literacy and daily self-care, particularly among populations facing socioeconomic and educational disparities.

Objective: This study examined whether a 14-week mobile teleassistance intervention could support daily health promotion and improve health literacy and quality of life among older adults, and whether different levels of user engagement were associated with differences in outcomes.

Methods: This randomized digital pilot study involved 21 older adults (aged ≥60 years) from Ribeirão Preto, Brazil. All participants were assigned to the intervention arm and subsequently categorized into high-engagement (n=11) and low-engagement (n=10) subgroups according to platform-use metrics. The intervention combined weekly teleconsultations, gamified educational quizzes, and guided health-related activities delivered through a mobile app. Outcomes included health literacy (Health Literacy Questionnaire), quality of life (36-Item Short-Form Health Survey), physical activity, and sedentary behavior, assessed at baseline and postintervention. Analyses appropriate for small samples were applied, including frequentist and Bayesian models.

Results: Participants in the high-engagement subgroup showed greater improvements in health literacy compared with those in the low-engagement subgroup (mean change +9.5 vs +9.1 points; time × group: $P<.001$; Bayes Factors [BF_1]=15). Significant interactions also favored higher engagement for selected quality-of-life domains: vitality ($P\leq.001$), functional capacity ($P=.02$), and general health ($P=.02$). A group effect was observed for the mental component ($P<.001$). Physical activity ($F_{2,38}=0.95$; $P=.39$; BF_incl=0.68) and sedentary behavior ($F_{1,19}=1.12$; $P=.32$; BF_incl=0.53) did not differ significantly between subgroups. Engagement analytics confirmed higher overall platform use in the high-engagement subgroup (mean 6483.8, SD 807.0 vs mean 3345.3, SD 742.7; $t_{19}=6.238$; $P<.001$; d=2.73) and more weekly health-activity minutes (mean 5124.3, SD 757.9 vs mean 3120.7, SD 704.3; $t_{19}=6.256$; $P<.001$; d=2.73).

Conclusions: This 14-week randomized digital pilot trial suggests that everyday digital teleassistance may enhance health literacy and specific quality-of-life domains among older adults when engagement is high. However, such support alone appears insufficient to modify physical activity or sedentary behavior in the short term. Larger and longer trials are needed to assess sustainability, scalability, and strategies to address structural inequalities in digital health adoption.

Trial Registration: Brazilian Clinical Trials Registry RBR-6wgkzs8; https://ensaiosclinicos.gov.br/rg/RBR-6wgkzs8

(JMIR Form Res 2025;9:e77319) doi:10.2196/77319

KEYWORDS

health promotion; patient engagement; digital health tools; behavioral change; older people care

XSL•FO
RenderX

Introduction

Population aging presents significant challenges for health care systems, particularly regarding the promotion of autonomy, quality of life, and engagement in health practices among older adults [1]. Despite advancements in preventive medicine and health education, a critical gap remains in the integration of accessible digital technologies that can effectively engage and positively impact this demographic. Previous studies have demonstrated that technology-based interventions, such as health apps, can enhance health literacy and quality of life, but these often lack robust engagement strategies or usability features tailored to older adults [2,3].

Given the complexity of behavior change in later life, a conceptual framework is essential to clarify mechanisms of action and guide hypothesis testing. We posit that mobile teleassistance influences outcomes through a multistep pathway. First, the intervention enhances access to understandable and actionable information (health literacy), while reducing barriers to navigation (remote guidance) and providing autonomy-supportive prompts (gamified tasks and teleconsultations).

These elements strengthen capability, opportunity, and motivation to engage in self-care (capability, opportunity, motivation–behavior model [4]), increase self-efficacy and autonomous motivation (self-determination theory [5]), and foster patient-provider connectedness. Proximal mediators are (1) digital engagement with the platform (use intensity and adherence), (2) health literacy gains, and (3) self-efficacy for daily management. Distal outcomes include quality-of-life domains and health behaviors (physical activity and sedentary time). We further anticipate moderation by baseline digital literacy, socioeconomic position, and cognitive status, as well as usability features.

Older adults often face intersecting barriers—lower digital literacy, socioeconomic constraints, and cognitive changes—that undermine the effectiveness of otherwise promising digital tools. Previous reviews show benefits of digital health literacy interventions in later life [6,7], yet mechanistic clarity and theoretically informed models remain underexplored, particularly regarding how engagement operates as a proximal mediator of intervention effects. Engagement is frequently treated as a crude exposure rather than as a dynamic behavioral construct informed by theory and technology-use frameworks [8,9].

Moreover, few studies examine how variations in engagement within the same intervention influence outcomes. In this context, frameworks such as reach, effectiveness, adoption, implementation, and maintenance [10] emphasize reach, adoption, and fidelity as crucial for scalability but are rarely applied to understand everyday teleassistance in older populations. Addressing these gaps, this study advances the evidence base by testing a theory-informed teleassistance intervention, explicitly modeling engagement as a mediator, and considering contextual moderators relevant to real-world implementation.

Guided by this model, we evaluated a multifaceted mobile teleassistance program for adults aged 60 years and older. The primary objective was to test whether higher engagement with the platform is associated with improvements in health literacy and quality-of-life domains. Secondary objectives were to examine effects on health behaviors (physical activity and sedentary time) and to explore the mediating roles of engagement and health literacy, as well as moderation by socioeconomic and cognitive factors. Based on this framework, we formulated the following hypotheses:

1. Primary hypothesis: participants with higher engagement in the teleassistance platform would demonstrate greater improvements in health literacy and quality-of-life domains over 14 weeks compared with those with lower engagement.
2. Secondary hypothesis: changes in physical activity and sedentary behavior could occur, but were expected to be modest in the absence of a structured exercise component.
3. Mechanistic hypothesis: engagement and health literacy would function as proximal mediators of intervention effects, whereas socioeconomic and cognitive factors would moderate these associations.

By addressing these conceptual and empirical gaps, this study contributes to advancing the evidence base on everyday digital teleassistance for older adults, offering insights that may inform the development of more scalable and context-sensitive digital health strategies.

Methods

Trial Design

This was a randomized, controlled, single-blind trial with allocation concealment and blinded outcome assessment, conducted in Ribeirão Preto, São Paulo, Brazil, between July and December 2023. A total of 21 community-dwelling adults aged 60 years and older were enrolled via convenience sampling through social and local media and community outreach (including the Bom Prato initiative). Inclusion criteria were ownership of a mobile device capable of supporting the teleassistance app and absence of severe visual, hearing, or cognitive impairments. Cognitive function was screened using the Montreal Cognitive Assessment (MoCA [11,12]). Randomization (1:1) to the intervention (app) or control was performed by an independent researcher.

Control Group Condition

Participants allocated to the control group did not receive the app or any alternative intervention during the 14-week study period; they continued their usual daily routines and standard health care access. At the end of the study, control participants were invited to voluntarily participate in the teleassistance intervention, ensuring equitable access after trial completion.

Rationale for Duration

The 14-week duration was selected in line with previous digital health literacy and teleassistance interventions in older adults [11-13], which commonly span 8 - 16 weeks, and to allow adequate time for habit formation and perception of healthier

XSL•FO
RenderX

daily routines, while balancing feasibility and retention in community settings.

For the present analyses, we focused on participants allocated to the app group, who were subsequently subdivided exploratorily into high- and low-engagement subgroups according to platform use. The final analytic sample comprised 11 participants in the higher-engagement subgroup and 10 in the lower-engagement subgroup, meeting a priori power-analysis requirements.

Study Type

This investigation is part of a comprehensive project evaluating a multifaceted mobile teleassistance app for older adults [11]. The app had been previously validated for usability and effectiveness [13], supporting its use in this trial. It integrates functionalities developed by a multidisciplinary team (physicians, nutritionists, nurses, physical education professionals, and psychologists) to promote health and support healthy aging. The overarching project adopts an integrated quantitative approach using a randomized controlled design with 2 main groups (control and intervention [11]). In this paper, we specifically analyze the app group over 14 weeks, with engagement-based subdivision derived from platform interaction metrics (eg, frequency of access and participation in activities). Gamification was used to motivate participation and to classify users into higher- and lower-score categories.

Figure 1 illustrates the theoretical model underpinning the intervention and analyses, depicting how engagement with teleconsultations, educational quizzes, and guided activities was hypothesized to influence health literacy, daily practices, and selected quality-of-life domains, moderated by factors such as digital literacy, socioeconomic position, and cognitive status. The model illustrates how engagement with teleconsultations, educational quizzes, and guided activities may affect health literacy, daily practices, and selected quality-of-life domains. It also considers moderating factors, such as digital literacy, socioeconomic position, and cognitive status. This conceptual model guided the study design, analytical approach, and interpretation of engagement as both an exposure variable and a proximal mediator.

Figure 1. Conceptual model outlining the hypothesized mechanisms through which mobile teleassistance influences health literacy and health behaviors among older adults.

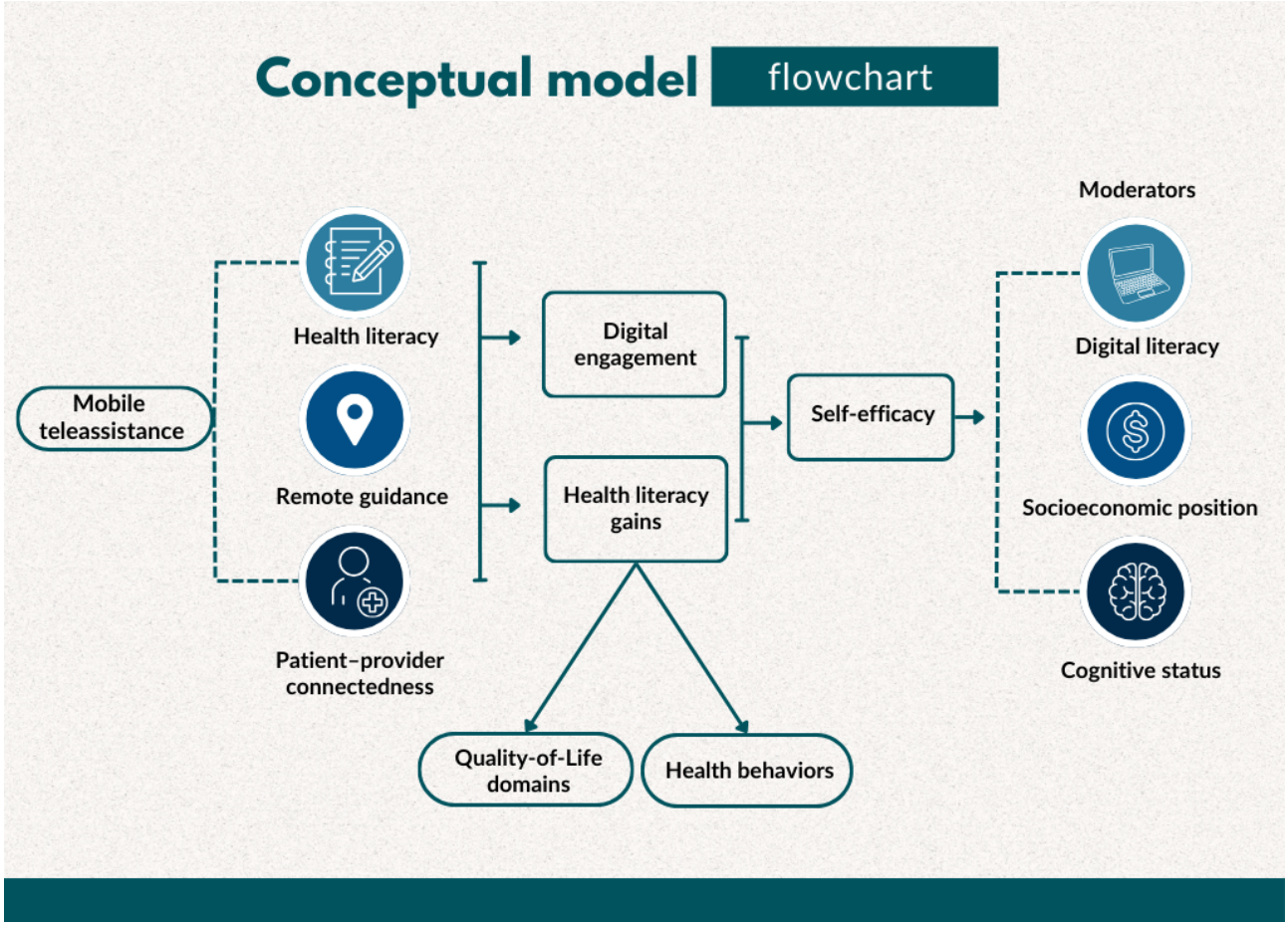

Participants and Setting

The study took place in Ribeirão Preto (SP), Brazil, with recruitment through social media, television, radio, University of São Paulo channels, and governmental initiatives such as Bom Prato. Strategies aimed to reduce selection bias and ensure socioeconomic and educational diversity, consistent with the exploratory nature of this pilot trial. Although not designed for population representativeness, this heterogeneity strengthens the external relevance of the findings.

Intervention

Participants in the intervention arm could freely access educational quizzes, teleconsultations, and guided health-related

XSL•FO
RenderX

activities through the mobile app. Teleconsultations were requested via in-app messaging and typically received responses from health care professionals within approximately 20 minutes. The app implemented a scoring system based on task completion to encourage adherence and enable engagement analytics [11]. To reduce technological barriers, structured technical support was provided daily during the first week, every 3 days during the following 2 weeks, and biweekly thereafter, with on-demand assistance available throughout the intervention [11]. This approach enabled assessment of both the impact of intensive app use and differences associated with engagement levels. Control participants received the app offer after the study, as detailed in the "Control Group Condition" section.

Participants (Eligibility, Randomization, and Subgrouping)

Eligibility

Individuals aged 60 years and older who owned a compatible mobile device were eligible; the device requirement was pivotal because the intervention depended on app use. Exclusion criteria included conditions that would preclude study assessments (eg, severe visual, hearing, or cognitive impairments). The MoCA [12] was administered to characterize cognitive function and to contextualize any difficulties in using the app.

Randomization and Allocation

Random allocation (1:1) to app or control groups was conducted by a blinded, independent researcher using the random number generator in Microsoft Excel 2013, with age bands (60 - 65 and 66 - 70 years) considered to enhance balance.

Subgrouping

After the start of the intervention, an exploratory secondary allocation within the app group divided participants into high- and low-engagement subgroups using the median of total engagement scores as the cutoff [14,15]. The subdivision was automated to minimize investigator influence. Exploratory checks of baseline equivalence (eg, age, socioeconomic level, and years of education) between engagement subgroups were performed [16].

Tests and Assessments

Cognitive status was screened with the MoCA [12] to characterize the sample and contextualize usability. Physical activity was measured with the International Physical Activity Questionnaire (IPAQ), classifying participants as low, moderate, or high activity [17]. Sedentary behavior was assessed by self-reported average daily sitting time, supplemented by IPAQ items specific to sedentary behavior [18].

Health literacy was assessed with the Test of Health Literacy for Portuguese-speaking Adults (THL-PA [19]), capturing comprehension and management of health-related information. Quality of life was measured with the 36-Item Short-Form Health Survey (SF-36), encompassing physical, emotional, social, and mental health domains [20].

The app automatically recorded the number of quizzes accessed, teleconsultations completed, and weekly time dedicated to health-related activities (eg, personal self-care, accessing health services such as pharmacies, participation in religious practices). These indicators supported classification into higher versus lower engagement for comparative analyses. Socioeconomic status was categorized using the Brazilian Socioeconomic Classification Questionnaire [21]. Demographic variables (eg, age and years of education) were collected by trained researchers to ensure standardization and reliability.

Weight, height, and physical performance tests were collected at baseline and postintervention to contextualize potential physical changes, as described previously [11,13]. Primary endpoints were changes in health literacy (THL-PA) and quality of life (SF-36). Secondary endpoints were physical activity, sedentary behavior, and engagement metrics (eg, quizzes accessed and teleconsultations).

Sample Size

A priori estimation using G*Power (version 3.1; Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang; University of Kiel and University of Mannheim) for repeated-measures ANOVA assumed a medium effect size ($f=0.25$), $\alpha=.05$, power=0.80, 2 time points (pre and post), 2 groups (high vs low engagement), and a within-subject correlation of 0.50, based on conservative literature [22]. The required total sample was 20 (10 per group). The final analytic sample comprised 21 (high engagement=11 and low engagement=10), exceeding the minimum and ensuring adequate model stability.

Statistical Analysis

Data were organized in Microsoft Excel and analyzed using SPSS (version 20; IBM Corp) and Jeffreys's Amazing Statistics Program (JASP; version 0.19; JASP Team; University of Amsterdam). Descriptive statistics included means, SDs, and percentages, with normality tested by the Kolmogorov-Smirnov test (Lilliefors correction).

Given the small sample (N=21), both frequentist and Bayesian methods were applied. Bayesian analyses (JASP) provided posterior probabilities and Bayes Factors (BF_1) to quantify evidence for the alternative hypothesis over the null hypothesis (BF_1 >3=moderate, >10=strong, and >30=very strong).

Between-group differences in continuous variables were tested using independent-samples *t* tests, and categorical variables were tested using 2-sided Fisher exact tests. Longitudinal outcomes were examined with a generalized linear model and generalized linear mixed model (GLMM), specifying group (high vs low engagement) as a fixed factor and time (pre vs post) as repeated.

Model fit was assessed using Akaike information criterion, Wald χ^2, and 95% CIs for β coefficients, with effect sizes (Cohen *d*, η^2). Bayesian repeated-measures ANOVA provided BF for inclusion (BF_incl). Sensitivity analyses excluded very-low-engagement participants (<20%) and outliers ($|z|>3.29$), yielding consistent results.

Missing data (<5%) were imputed by chained equations (multivariate imputation by chained equations and missing at random assumption) and pooled via Rubin's rules. Bonferroni corrections adjusted the Type I error, and both raw and adjusted

XSL•FO
RenderX

P values are reported. Visualizations (raincloud and Flexplots) were generated in JASP. Statistical significance was set at α=.05 (2-sided).

Ethical Considerations

The study was conducted by the lead researcher (ACdSS, Health Sciences) at the Ribeirão Preto Medical School of the University of São Paulo (FMRP/USP). All procedures complied with Resolution 466/2012 of the Brazilian National Health Council and adhered to international bioethical principles. The protocol was approved by the Research Ethics Committee of the Ribeirão Preto School of Physical Education and Sport, University of São Paulo (process no 58433922.0.0000.5659; October 26, 2022). All participants received detailed information about the study objectives, procedures, potential risks, and benefits and provided written informed consent before participation. Confidentiality was ensured through anonymized data handling, secure digital storage, and restricted access. Data were used exclusively for academic and scientific purposes. The trial was registered in the Brazilian Clinical Trials Registry (ReBEC: RBR-6wgkzs8; "Development of an Application to Improve Health in Older People"). Participants did not receive any financial compensation or material incentives for participation.

Results

Sample Characteristics

The final sample comprised 21 participants divided into 2 groups: the higher-score group (higher, n=11) and the lower-score group (lower, n=10). The mean age in the higher group was 64.73 (SD 3.23) years, while the lower group had a mean age of 63.10 (SD 2.99) years. Normality was confirmed using the Kolmogorov-Smirnov test ($P>.05$). Independent *t* tests showed no significant differences between groups ($t_{19}=1.20$; $P=.25$), corroborated by Bayesian results ($BF_1=0.59$), indicating anecdotal evidence in favor of the null hypothesis.

Socioeconomic and Educational Characteristics

According to the Brazilian socioeconomic classification criteria, the higher group had a mean score of 26.55 (SD 10.42), while the lower group scored 32.60 (SD 8.33). The Fisher exact test indicated no statistically significant differences in category distribution ($P=.45$). Bayesian contingency analysis provided moderate evidence supporting the null hypothesis ($BF_1=0.41$). Regarding education, participants in the higher group reported an average of 11.0 (SD 2.8) years of formal schooling, compared with 11.7 (SD 3.7) years in the lower group. Independent-samples *t* test results showed no significant between-group differences ($t_{19}=-0.49$; $P=.63$), and Bayesian analysis indicated anecdotal evidence supporting the null hypothesis ($BF_1=0.45$). These results suggest that the groups were comparable in terms of age, education, and socioeconomic background at baseline.

Physical Activity and Sedentary Behavior

Physical activity levels, assessed using the IPAQ, were classified as low, moderate, or high. Preintervention, 45.5% of the higher group had low activity, 45.5% moderate, and 9.1% high. In the lower group, 50% were low and 50% moderate. Postintervention, 27.3% of the higher group had low activity, 45.5% moderate, and 27.3% high. The lower group had 20% low, 30% moderate, and 50% high.

To analyze differences over time, a GLMM was conducted with physical activity level as the dependent categorical variable, group as a fixed factor, and time as a repeated measure. No significant interaction was found ($F_{2,38}=0.95$; $P=.39$), and Bayesian GLMM results yielded a BF_incl=0.68, indicating only anecdotal evidence for a group-by-time effect. Sedentary behavior, measured as sitting time, showed that before the intervention, 90.9% of the higher group and 100% of the lower group were highly sedentary. After the intervention, the distribution remained unchanged in the higher group, while 20% of the lower group shifted to moderate sedentary levels. GLMM analysis revealed no significant group-by-time interaction for sedentary behavior ($F_{1,19}=1.12$; $P=.32$), with a Bayesian BF_incl=0.53, again indicating anecdotal support for the null hypothesis.

Health Literacy and Quality of Life (SF-36)

Table 1 displays the outcomes for health literacy and SF-36 quality-of-life domains. Both groups improved significantly in health literacy, with the higher group exhibiting a greater gain (interaction T×G: $P<.001$; $BF_1=15.2$, indicating strong evidence in favor of a group-by-time effect). For the physical domain, a significant postintervention reduction was noted in the lower group (T×G: $P=.002$; $BF_1=10.4$), suggesting strong evidence for change. The mental domain remained stable over time but differed significantly between groups (group effect: $P<.001$; $BF_1=12.7$). Values are presented as mean and SD. Post hoc analyses include main effects of time, group, and their interaction (T × G), Wald statistics, *P* values, standardized beta coefficients, 95% CIs, and Akaike information criterion (AIC) values. Variables assessed include health literacy and multiple SF-36 domains (physical, mental, general health, functional capacity, vitality, physical aspects, emotional aspects, social aspects, mental health, and pain). Significant effects are indicated at $P<.05$. "Time" reflects pre-post changes across both groups, "group" reflects between-group differences, and "T × G" represents the interaction between time and group.

XSL•FO
RenderX

Table . Pre- and post-intervention comparisons between engagement groups.

Variable	Higher-score group, mean (SD)	Higher-score group, mean (SD)	Lower-score group, mean (SD)	Lower-score group, mean (SD)	Post hoc (*F*,*P* value)			Wald test	*P* value	ß coefficients (95% CI)	AIC[a]
	Preintervention	Postintervention	Preintervention	Postintervention	Time	Group	T×G				
Health literacy	80.6 (15.6)	90.1 (13.7)	72.5 (17.5)	81.6 (12.9)	4.1, .04	3.4, .07	7.7, <.001	2.1	.03	75.9 (66.3 - 406.7)	342.7
Physical domain	51.9 (10.4)	50.2 (9.9)	61.9 (7.8)	57.6 (8.3)	1.0, .30	9.4, .<001	1.3, <.001	3.1	.002	83.1 (44.4 - 155.3)	301.9
Mental domain	41.5 (9.4)	41.5 (9.4)	54.4 (10.3)	54.4 (10.3)	0.0, ≥.99	18.2,<.001	9.2, ≥.99	3.1	.002	94.6 (50.7 - 176.5)	306.5
General health status	14.2 (3.6)	13.4 (3.0)	19.9 (2.8)	16.2 (2.0)	2.9, .30	8.5, .20	8.0, .02	2.1	.03	4.04 (1.6 - 10.1)	215.4
Functional capacity	23.3 (5.1)	22.2 (5.1)	26.7 (2.5)	25.4 (3.0)	1.3, .26	5.8, .02	5.2, .02	2.1	.03	9.2 (3.7 - 22.9)	237.5
Vitality	13.6 (5.0)	13.6 (5.0)	16.6 (2.9)	16.6 (2.9)	0.0, ≥.99	5.4, .02	7.7, .<.001	3.1	.002	16.8 (9.0 - 31.4)	239.3
Physical aspect	6.6 (1.3)	6.6 (1.3)	6.5 (2.0)	6.5 (2.0)	0.0, .70	0.0, .04	0.0, .07	2.1	.02	3.9 (1.5 - 9.5)	170.0
Emotional aspect	4.0 (1.3)	4.0 (1.3)	5.0 (1.3)	5.0 (1.3)	0.0, ≥.99	0.0, ≥.99	0.0, ≥.99	2.1	.30	1.72 (0.7 - 4.2)	153.5
Social aspect	7.0 (1.9)	7.0 (1.9)	8.3 (1.5)	8.3 (1.5)	0.0, ≥.99	5.0, .03	0.0, ≥.99	3.1	.02	3.0 (1.6 - 5.6)	172.2
Mental health	18.9 (4.3)	18.9 (4.3)	22.1 (3.3)	22.1 (3.3)	0.0, ≥.99	11.2, <.001	30.8,<.001	3.1	.002	14.9 (8.0 - 27.9)	234.6
Pain	7.0 (2.4)	7.0 (2.4)	8.4 (2.3)	8.4 (2.3)	0.0, ≥.99	3.6, .06	3.6, .06	2.1	.30	5.5 (2.2 - 13.6)	200.5

[a]AIC: Akaike information criterion.

Other domains—general health, functional capacity, and vitality—also showed significant improvements, with BF ranging from 4.3 to 9.8, indicating moderate to strong evidence. In contrast, physical, emotional, and social aspects remained stable (BF_1 <1). Mental health showed a greater postintervention increase in the lower group (T×G: *P*<.001; BF_1 =17.6), indicating very strong evidence for group-specific effects. These findings reinforce the positive impact of the intervention, especially in promoting improvements in health literacy and selected quality-of-life domains, with Bayesian analysis enhancing confidence in the observed effects despite the small sample size.

Table 2 and the raincloud plots (Figure 2) provide key metrics analyzed to compare the higher- and lower-score groups. These metrics include total engagement scores, number of quizzes accessed, number of teleconsultations performed, and average weekly time dedicated to health-related activities. Values are presented as mean and SD. Between-group differences were assessed using independent *t* tests. Metrics include total engagement score (arbitrary units), total quizzes accessed (n), number of teleconsultations (n), and total weekly minutes of health-related activities. Effect sizes are reported as Cohen *d* with 95% CIs. Significance level set at *P*<.05.

XSL•FO
RenderX

Table . Comparison of key engagement and health metrics between high- and low-scoring groups during the 14-week teleassistance intervention (Ribeirão Preto, Brazil, 2023).

Variable	Higher-score group, mean (SD)	Lower-score group, mean (SD)	*t* test (*df*)	*P* value	Effect size, *d* (95% CI)
Total score (au[a])	6483.773 (807.034)	3345.272 (742.684)	6.238 (19)	<.001	2.726 (1.494 to 3.922)
Total quizzes accessed (n)	33.45 (3.085)	35.20 (3.372)	−0.383 (19)	.71	−0.167 (−1.023 to 0.693)
Teleconsultations (n)	14.33 (4.144)	10.00 (5.007)	0.831 (19)	.42	0.363 (−0.506 to 1.223)
Weekly health activity (minutes)	5124.273 (757.892)	3120.700 (704.249)	6.256 (19)	<.001	2.733

[a]au: arbitrary unit.

Figure 2. Flow of participants through the study. This randomized digital pilot trial involved older adults (≥60 years) from Ribeirão Preto, Brazil. Participants were randomized to an intervention (mobile teleassistance app) and a control group. The intervention lasted 14 weeks and combined teleconsultations, gamified educational quizzes, and guided health activities. Secondary analyses focused on the 21 participants in the intervention arm, who were further categorized into high- and low-engagement subgroups based on platform usage. Secondary allocation into high- and low-engagement subgroups was performed only for participants in the app group (N=21). The control group was not included in this exploratory analysis.

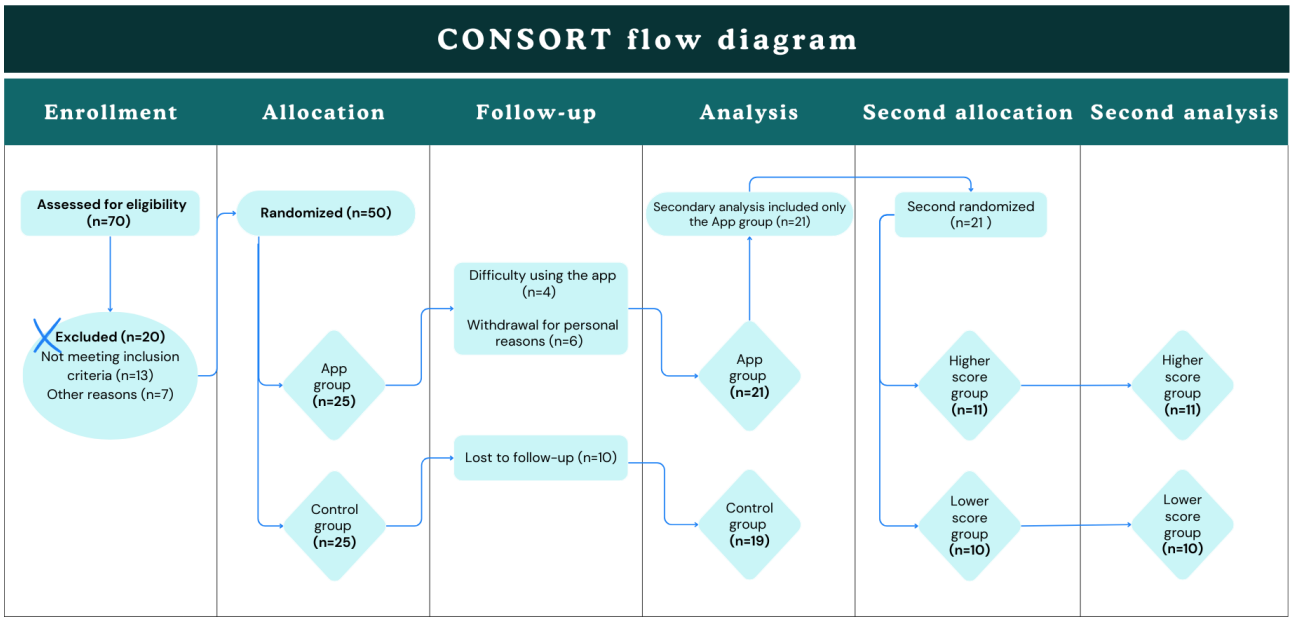

The total engagement score showed a substantial difference between groups: the higher-score group had a mean of 6483.77 (SD 807.03), while the lower-score group had a mean of 3345.27 (SD 742.68). The difference was statistically significant (t_{19}=6.238; *P*<.001) with a very large effect size (Cohen *d*=2.726, 95% CI 1.494 - 3.922). Bayesian *t* test results strongly supported this difference (BF_1 =97.4), providing very strong evidence for the alternative hypothesis. Similarly, the average weekly time dedicated to health activities was significantly greater in the higher-score group (mean 5124.27 and SD 757.89 minutes) than in the lower-score group (mean 3120.70 and SD 704.25 minutes). This difference was also highly significant (t_{19}=6.256; *P*<.001), with a very large effect size (Cohen *d*=2.733, 95% CI 1.500 - 3.931), and was strongly supported by Bayesian analysis (BF_1 =102.8).

In contrast, no significant differences were observed for the total number of quizzes accessed (t_{19}=−0.383; *P*=.71; Cohen *d*=-0.167; BF_1 =0.34) or the number of teleconsultations performed (t_{19}=0.831; *P*=.42; Cohen *d*=0.363; BF_1 =0.48). These Bayesian results indicate moderate evidence in favor of the null hypothesis, suggesting that these variables were not key contributors to differentiating user engagement levels.

The raincloud plots (Figure 3A-D) provide a detailed visualization of the distribution, variability, and central tendencies of the key variables analyzed. These plots corroborate the inferential statistics by clearly illustrating the marked differences in total engagement scores and average weekly time dedicated to health activities between the higher- and lower-score groups. By integrating raw data points, probability density estimates, and mean values, these visualizations reinforce the interpretation of the findings and offer a complementary perspective to the quantitative results.

XSL•FO
RenderX

Figure 3. Distribution of (A) total engagement score, (B) number of teleconsultations, (C) weekly health activity minutes, and (D) total number of quizzes accessed among older adults (N=21) in the high- and low-engagement subgroups during a 14-week randomized digital pilot trial in Ribeirão Preto, Brazil. Values are represented as mean and SD. No statistically significant differences were observed at baseline between the higher- and lower-engagement groups.

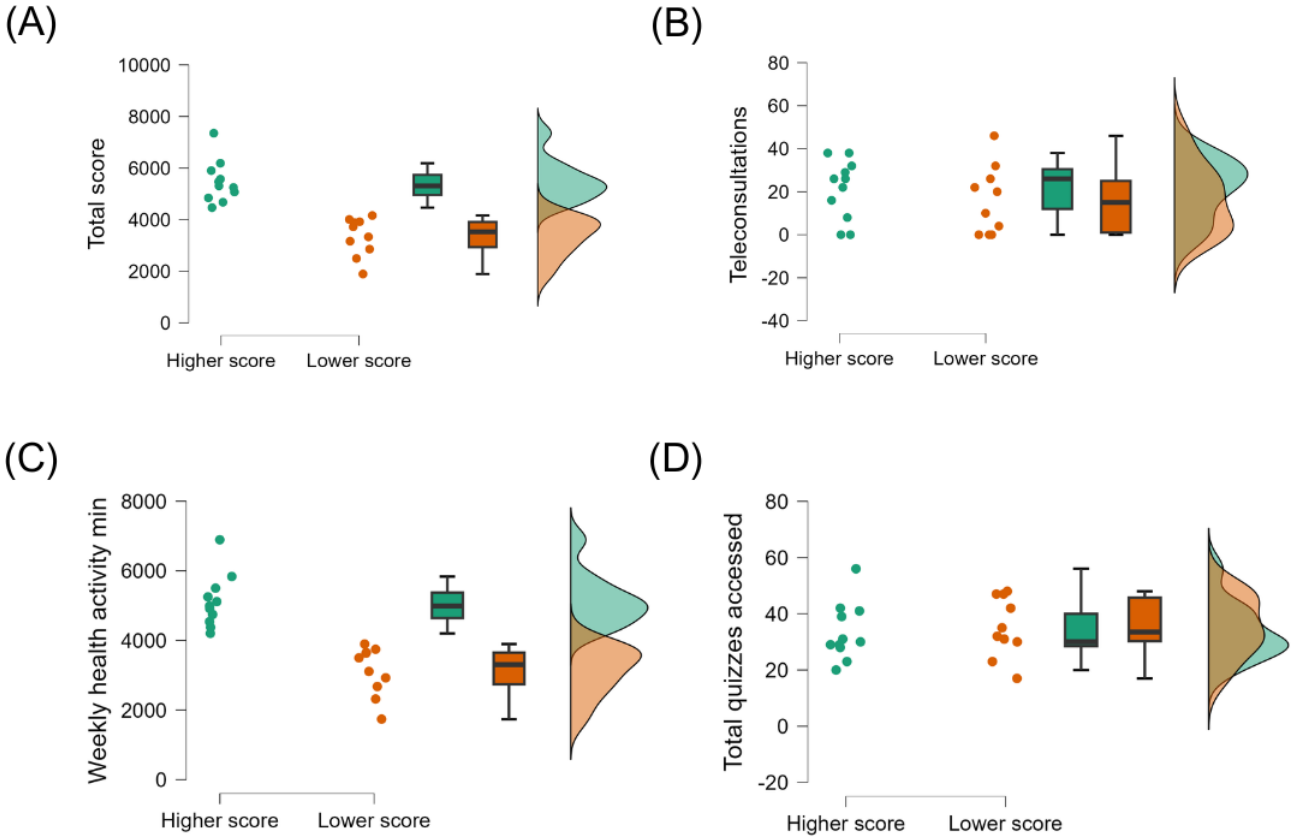

The Flexplot visualizations (Figure 4), generated using JASP software, were used to explore interaction patterns between engagement-related variables across groups. These plots merge individual-level data with modeled trends, facilitating intuitive identification of between-group differences and variability within groups.

XSL•FO
RenderX

Figure 4. Relationship between the number of teleconsultations and total quizzes accessed, stratified by weekly health activity minutes, among older adults in the high- and low-engagement subgroups (N=21) during a 14-week randomized digital pilot trial in Ribeirão Preto, Brazil. Lines represent linear fits and shaded areas indicate 95% CIs. Patterns illustrate differential engagement behaviors across groups.

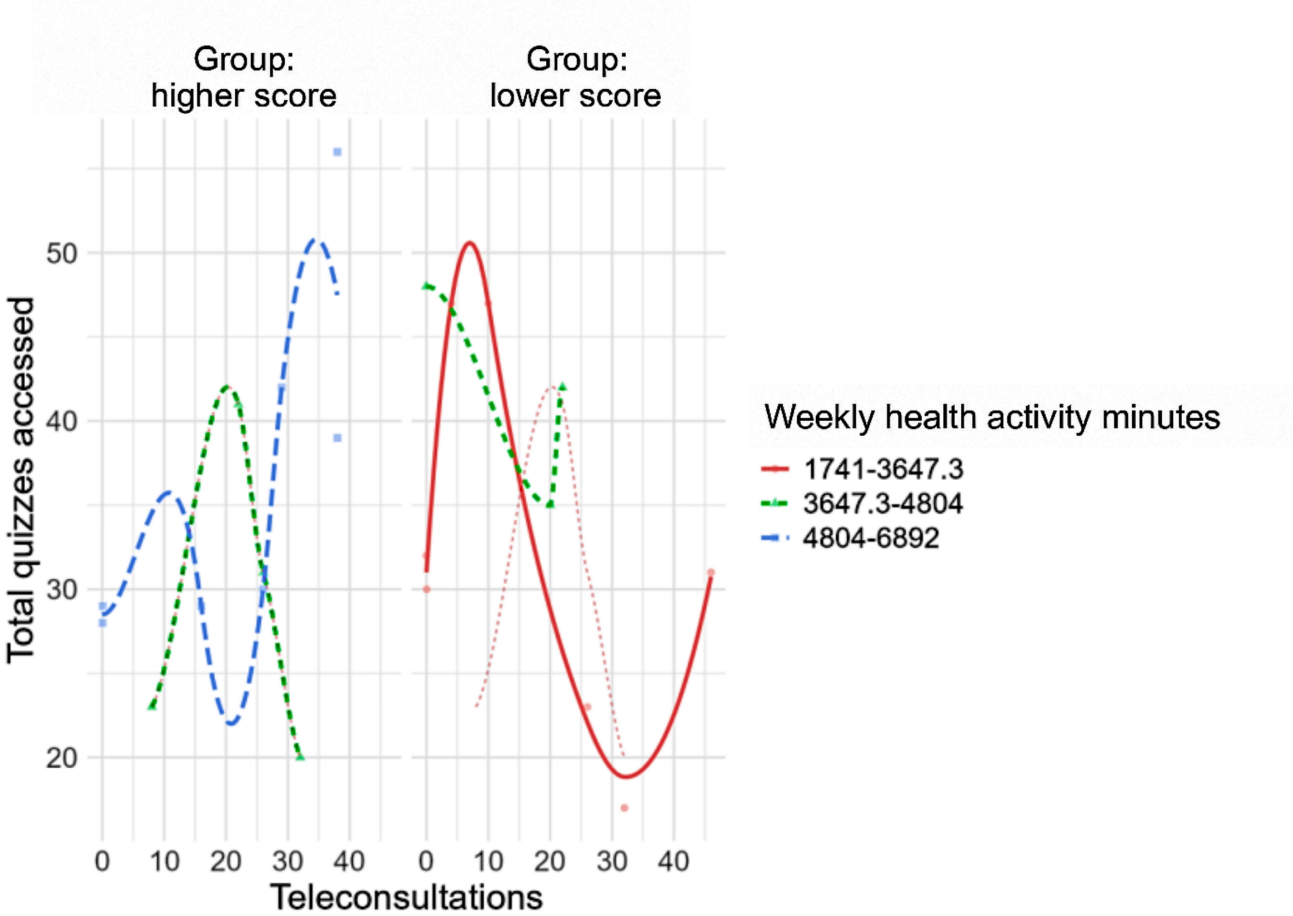

For the total engagement score, the Flexplot revealed a clear and consistent separation between the two groups, with the higher-score group exhibiting both higher mean values and lower within-group dispersion, indicating greater uniformity and coherence in engagement. Similarly, the plot for weekly time dedicated to health-related activities demonstrated a broader spread of values overall but with a notable rightward shift in the higher-score group, reflecting more intensive engagement over time (Figure 4).

The centrality graph (Figure 5) complements the previous findings by illustrating the relative influence of each variable within the network of relationships observed in the dataset. It identifies total engagement score and weekly time dedicated to health-related activities as the most central metrics, indicating their dominant role in distinguishing between the higher- and lower-score groups. These variables show stronger interconnections with other dimensions analyzed, reinforcing their relevance to the intervention's overall impact.

XSL•FO
RenderX

Figure 5. Network centrality metrics (betweenness, closeness, strength, and expected influence) for 11 behavioral and digital engagement indicators in older adults (N=21) participating in a 14-week randomized digital pilot trial in Ribeirão Preto, Brazil. Variables: 1=occupational activity, 2=leisure activity, 3=transportation activity; 4=domestic activity, 5=voluntary activity, 6=average weekly health activity, and 7=monthly health activity, 8=weekly health activity, 9=teleconsultations, 10=quizzes, and 11=total score. Node size is proportional to strength, and edges represent partial correlations between variables. Higher centrality values indicate greater influence within the engagement-behavior network.

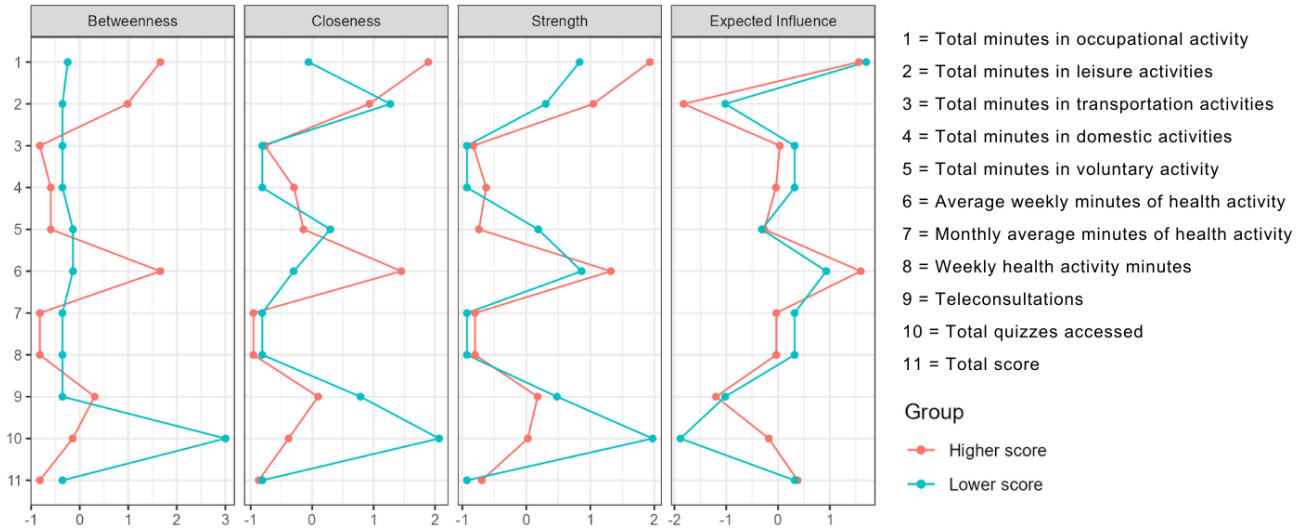

In addition, the graph highlights secondary variables, such as the number of teleconsultations performed, which exhibit lower centrality but still contribute meaningful contextual information. Although these variables have less direct influence on group differentiation, their inclusion enhances the multivariate understanding of engagement patterns and behavioral responses to the intervention (Figure 4).

These graphical representations, alongside the statistical analyses, provide a comprehensive and multidimensional view of the relationships among the studied variables and between the comparison groups. They not only corroborate the quantitative findings but also offer valuable visual insights into the internal dynamics of the dataset, thereby enhancing the robustness, interpretability, and depth of the study's conclusions.

Discussion

Principal Findings

The results of this study highlight associations between the proposed intervention and improvements in health literacy and selected domains of quality of life, such as general health status, functional capacity, vitality, and mental health. These improvements appeared particularly notable in the high-scoring group, suggesting a potential positive influence of the intervention on physical and psychological well-being. While no significant differences in physical activity levels or sedentary behavior were observed between groups over time, the analysis indicated greater engagement with health practices in the higher-performing group, as reflected by significantly higher total scores and weekly time dedicated to health-related activities.

These associations suggest that the intervention may have contributed to improvements in specific aspects of quality of life, potentially fostering behavioral and cognitive changes that were reflected in participants' daily lives. This finding indicates that frequent and consistent use of digital tools, combined with supportive strategies, could serve as a catalyst for positive changes in older populations. Throughout this discussion, we will explore how these findings align with broader patterns of health promotion, their potential practical applications, and the challenges of translating these data into effective and scalable strategies.

Health Literacy and Quality of Life

The observed increase in health literacy aligns with evidence in the literature, which identifies health literacy as a key factor in promoting autonomy and self-care. Programs facilitating access to health information and its understanding are associated with reduced complications from chronic diseases and the adoption of healthy behaviors [23,24]. These conclusions are supported by research such as that by Zhang et al [24] and Ngiam et al [25]. Moreover, simple everyday questions play a crucial role as a gateway to self-care [5]. In this study, participants with higher levels of engagement tended to report greater capacity to engage in health practices. This finding may underscore the potential role of technology in strengthening ties between users and health care professionals while simultaneously addressing basic daily demands [11]. Technological support can also facilitate the search for and understanding of more complex issues, potentially promoting a broader and more efficient approach to health care [11].

Engagement Patterns and Digital Behavior

The differences between the high- and low-scoring groups reflect not only the impact of the intervention but also the influence of contextual factors. A study on a physical activity app demonstrated that the inclusion of gamification features, such as leaderboards and status indicators, increased app usage and participants' physical activity, highlighting the potential of gamification to promote adherence and engagement [15]. Health interventions based on gamification have also shown promising results [26,27]. For instance, in patients with coronary heart disease, gamification through smartphone apps enhanced participation in physical activities and improved autonomous

XSL•FO
RenderX

motivation [26]. In addition, the personalization of goals in gamified health apps has been shown to increase engagement, particularly when goals are tailored to the users' capacities and objectives [28]. The division of participants into high- and low-engagement groups should, however, be interpreted with caution. The median split approach was adopted as a pragmatic and exploratory strategy in this small pilot sample, consistent with methods used in previous digital engagement studies. While informative, these subgroup analyses are not theory-driven and should be considered preliminary, requiring replication in larger and more diverse cohorts.

Barriers to Digital Inclusion

Socioeconomic and educational barriers [25] remain significant challenges to technological inclusion [25,29]. Addressing these factors is essential to ensure that the benefits of digital health interventions are equitably distributed and not restricted to socially or economically privileged populations. In older adults, these challenges are compounded by age-related factors such as lower digital literacy, cognitive decline, and apprehension toward adopting new technologies. Changing behavior through digital means tends to be more complex and less immediate for older individuals than in structured, face-to-face settings, where social interaction, real-time support, and group dynamics often enhance adherence and motivation.

Physical Activity and Sedentary Behavior

Despite the observed improvements in health literacy and selected quality-of-life domains, results related to physical activity and sedentary behavior showed limited progression. This stagnation echoes findings from other studies, which indicate that older adults face a constellation of physical and environmental barriers that hinder behavioral change, including preexisting health conditions, fear of falling or injury, low social support, lack of motivation, restricted mobility, and economic constraints [29-32]. In addition, methodological limitations such as reliance on self-reported data may have affected measurement accuracy [32]. Future interventions could benefit from incorporating objective monitoring tools—such as pedometers, accelerometers, and in-app sensors—to capture behavioral outcomes more precisely and support individualized feedback loops.

It is important to emphasize that the absence of significant improvements in physical activity and sedentary behavior is a critical finding of this study. This result is consistent with previous systematic reviews and meta-analyses showing that digital interventions alone rarely produce meaningful changes in these behaviors among older adults, particularly when no structured exercise component is embedded in the program [33,34]. Behavior change in physical activity is complex and multifactorial, often requiring multimodal strategies that combine digital tools with personalized goal setting, supervised sessions, or community-based support. Our findings, therefore, highlight the need for hybrid or augmented models that integrate technology with direct physical activity components to effectively influence movement behaviors in this population. Furthermore, this reinforces the importance of considering null results not as failures, but as evidence pointing toward the boundaries of what digital health literacy interventions can achieve in isolation.

Practical Implications of Digital Interventions

Another significant aspect is the practical potential of digital interventions. In this context, mobile health apps that use behavior change techniques, such as feedback and monitoring, stand out as promising tools for promoting adherence to interventions and behavioral change [35]. Recent studies underscore the importance of integrating interactive and personalized features to meet the specific needs of target populations, particularly older adults, who often face technological and engagement barriers [36]. In addition, hybrid models that combine personalized care plans with in-person workshops have demonstrated potential to strengthen older adults' confidence in new care approaches. These models are particularly relevant in the postpandemic context, where social interaction has gained greater value [37]. The literature also indicates that hybrid interventions promote higher adherence by reconciling the convenience of digital technologies with the benefits of in-person support [37]. Personalization of content and accessibility of digital health technologies remain crucial factors for adoption by older adults [37]. Approaches integrating standard care practices with interactive and personalized elements have the potential to overcome engagement barriers and expand the impact of interventions [35,36]. This scenario highlights the continuous need for innovation and adjustment of digital interventions to maximize their reach and effectiveness, aligned with growing evidence in the literature [38].

Limitations and Future Directions

The limitations of this study—particularly the small sample size and short intervention duration—may have affected the robustness of some outcomes. While the statistical approach adopted was sufficient for several variables, future studies with larger samples and extended follow-up periods are essential to strengthen the generalizability and precision of results across all outcomes. A larger sample would enable more robust and generalizable statistical analyses, while longitudinal studies could capture the long-term effects of digital interventions on the health and quality of life of older adults. Although participants were primarily recruited from low-income community programs, this population is representative of underserved older adults who often face barriers to digital health adoption, thus strengthening the external relevance of the findings.

Such investigations would also be essential for understanding how the sustained use of these technologies may influence adherence and clinical outcomes over time [38]. Furthermore, there is a critical need to integrate multidimensional approaches that combine behavioral, cognitive, and environmental aspects. Including factors such as social support, health education, and strategies tailored to older adults' life contexts could amplify the potential impact of interventions [38]. For instance, addressing barriers such as resistance to technology use, economic accessibility, and cultural differences may enhance adherence and engagement. The literature also suggests that alignment between users' objectives and the functionalities

XSL•FO
RenderX

offered by digital tools is a key factor for the success of these initiatives [38].

Taken together, these findings suggest that digital interventions may represent a promising opportunity to support improvements in quality of life and well-being among older adults. However, the development of more inclusive and personalized solutions that address the contextual barriers faced by this population will be essential to maximize their impact on healthy aging. Therefore, continuous innovation, combined with a deep understanding of users' needs, should guide the evolution of digital health technologies designed for older adults.

Conclusion

This study highlights the potential of teleassistance to enhance health literacy, autonomy, and well-being among older adults. By integrating technology with personalized support and gamification, the intervention offered a feasible and scalable model for digital health promotion. However, no significant effects were found for physical activity or sedentary behavior, suggesting that literacy-oriented digital programs alone may be insufficient to drive behavioral change. Future research should refine such interventions using hybrid and multimodal strategies to better influence movement behaviors while ensuring inclusivity and adaptability in addressing the needs of aging populations.

Acknowledgments

We would like to express our gratitude to the participants for their valuable time and effort. Special thanks to the multidisciplinary team involved in this project for their continuous support. This work was supported by Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). Generative AI (ChatGPT, OpenAI) was used under the author's supervision to assist with language refinement and formatting. All scientific content and interpretations were produced and verified by the authors.

Funding

This research did not receive a specific grant from any funding agency in the public, commercial, or not-for-profit sectors. The authors acknowledge institutional support from the University of São Paulo and national research funding agencies (FAPESP #2017/21361 - 2, CNPq #485045/2013 - 3 and #141720/2017 - 4, and CAPES – Finance Code 001) for broader programmatic support.

Data Availability

The datasets generated and analyzed during this study contain sensitive health information from older adult participants and are therefore not publicly available. Deidentified, aggregated data may be shared by the corresponding author upon reasonable request, contingent on approval by the institutional Research Ethics Committee, in accordance with Brazilian regulations (resolution 466/2012).

Conflicts of Interest

None declared.

Checklist 1
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File, 1246 KB - formative_v9i1e77319_app1.pdf]

References

1. Rony MKK, Parvin M, Wahiduzzaman M, Akter K, Ullah M. Challenges and advancements in the health-related quality of life of older people. Adv Public Health 2024 Feb 1;2024:1-18. [doi: 10.1155/2024/8839631]
2. Dong Q, Liu T, Liu R, Yang H, Liu C. Effectiveness of digital health literacy interventions in older adults: single-arm meta-analysis. J Med Internet Res 2023 Jun 28;25:e48166. [doi: 10.2196/48166] [Medline: 37379077]
3. Allemann H, Poli A. Designing and evaluating information and communication technology-based interventions? Be aware of the needs of older people. Eur J Cardiovasc Nurs 2020 Jun;19(5):370-372. [doi: 10.1177/1474515119897398] [Medline: 31969003]
4. Michie S, van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. Implement Sci 2011 Apr 23;6:42. [doi: 10.1186/1748-5908-6-42] [Medline: 21513547]
5. Deci EL, Ryan RM. The "What" and "Why" of goal pursuits: human needs and the self-determination of behavior. Psychol Inq 2000 Oct;11(4):227-268. [doi: 10.1207/S15327965PLI1104_01]
6. Sørensen K, Van den Broucke S, Fullam J, et al. Health literacy and public health: a systematic review and integration of definitions and models. BMC Public Health 2012 Jan 25;12:80. [doi: 10.1186/1471-2458-12-80] [Medline: 22276600]

XSL•FO
RenderX

7. WHO guideline: Recommendations on digital interventions for health system strengthening. World Health Organization (WHO). 2019. URL: https://apps.who.int/iris/handle/10665/311941 [accessed 2025-11-27]
8. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view. MIS Q 2003 Sep 1;27(3):425-478. [doi: 10.2307/30036540]
9. Ngiam NHW, Yee WQ, Teo N, et al. Building digital literacy in older adults of low socioeconomic status in Singapore (Project Wire Up): Nonrandomized Controlled Trial. J Med Internet Res 2022;24(12):e40341. [doi: 10.2196/40341]
10. Glasgow RE, Harden SM, Gaglio B, et al. RE-AIM planning and evaluation framework: adapting to new science and practice with a 20-year review. Front Public Health 2019;7:64. [doi: 10.3389/fpubh.2019.00064] [Medline: 30984733]
11. da Silva Sobrinho AC, da Silva Rodrigues G, Pereira Rodrigues K, et al. Transforming daily support with multidisciplinary teleassistance: impact on health parameters in older adults—a randomized controlled trial. Mhealth 2025 Jul;11:30-30. [doi: 10.21037/mhealth-24-72]
12. Zietemann V, Kopczak A, Müller C, Wollenweber FA, Dichgans M. Validation of the telephone interview of cognitive status and telephone Montreal Cognitive Assessment against detailed cognitive testing and clinical diagnosis of mild cognitive impairment after stroke. Stroke 2017 Nov;48(11):2952-2957. [doi: 10.1161/STROKEAHA.117.017519] [Medline: 29042492]
13. Sobrinho ADS, Gomes GDO, Bueno Júnior CR. Developing a multiprofessional mobile app to enhance health habits in older adults: user-centered approach. JMIR Form Res 2024 Apr 15;8:e54214. [doi: 10.2196/54214] [Medline: 38619865]
14. Moffit R, McTigue K, Conroy MB, et al. Aspects of program engagement in an online physical activity intervention and baseline predictors of engagement. Am J Health Promot 2023 Nov;37(8):1100-1108. [doi: 10.1177/08901171231194176] [Medline: 37550892]
15. Maher CA, Olds T, Vandelanotte C, et al. Gamification in a physical activity app: what gamification features are being used, by whom, and does it make a difference? Games Health J 2022 Jun;11(3):193-199. [doi: 10.1089/g4h.2021.0207] [Medline: 35501981]
16. Schulz KF, Altman DG, Moher D, Group C. CONSORT 2010 statement: updated guidelines for reporting parallel group randomized trials. Ann Intern Med 2010 Jun 1;152(11):726-732. [doi: 10.7326/0003-4819-152-11-201006010-00232] [Medline: 20335313]
17. International physical activity questionnaire: guidelines for data processing and analysis of the international physical activity questionnaire (IPAQ) – short and long forms. IPAQ. 2005. URL: https://www.researchgate.net/file.PostFileLoader.html?id=5641f4c36143250eac8b45b7&assetKey=AS%3A294237418606593%401447163075131 [accessed 2025-11-27]
18. Pardini R, Matsudo S, Araújo T, et al. Validação do questionário internacional de nível de atividade física (IPAQ - versão 6): estudo piloto em adultos jovens brasileiros. Rev Bras Ciên e Mov 2001;9(3):45-51 [FREE Full text]
19. Maragno CA, Mengue SS, Moraes CG, Rebelo MV, Guimarães MA, Pizzol TSD. Test of health literacy for Portuguese-speaking adults. Rev Bras Epidemiol 2019;22:e190025. [doi: 10.1590/1980-549720190025]
20. Papou A, Hussain S, McWilliams D, Zhang W, Doherty M. Responsiveness of SF-36 Health Survey and Patient Generated Index in people with chronic knee pain commenced on oral analgesia: analysis of data from a randomised controlled clinical trial. Qual Life Res 2017 Mar;26(3):761-766. [doi: 10.1007/s11136-016-1484-2] [Medline: 27995369]
21. Critério de Classificação Econômica Brasil. Associação Brasileira de Empresas de Pesquisa (ABEP). 2003. URL: https://abep.org/wp-content/uploads/2024/02/08_cceb_2003_em_vigor_em_2003_base_lse_2000.pdf
22. Lacroix A, Kressig RW, Muehlbauer T, et al. Effects of a supervised versus an unsupervised combined balance and strength training program on balance and muscle power in healthy older adults: a randomized controlled trial. Gerontology 2016;62(3):275-288. [doi: 10.1159/000442087] [Medline: 26645282]
23. Magi CE, Bambi S, Rasero L, et al. Health literacy and self-care in patients with chronic illness: a systematic review and meta-analysis protocol. Healthcare (Basel) 2024 Mar 31;12(7):762. [doi: 10.3390/healthcare12070762] [Medline: 38610184]
24. Zhang Z, Gu D, Li S. Effectiveness of person-centered health education in the general practice of geriatric chronic disease care. Altern Ther Health Med 2024 Oct;30(10):349-357 [FREE Full text] [Medline: 38330589]
25. Ngiam NHW, Yee WQ, Teo N, et al. Building digital literacy in older adults of low socioeconomic status in Singapore (Project Wire Up): nonrandomized controlled trial. J Med Internet Res 2022 Dec 2;24(12):e40341. [doi: 10.2196/40341] [Medline: 36459398]
26. Xu L, Tong Q, Zhang X, et al. Smartphone-based gamification intervention to increase physical activity participation among patients with coronary heart disease: a randomized controlled trial. J Telemed Telecare 2024 Oct;30(9):1425-1436. [doi: 10.1177/1357633X221150943] [Medline: 36794484]
27. Xu L, Yu T, Gao R, et al. Maintenance effects of a gamification intervention on motivation and physical activity in patients with coronary heart disease: intermediate results of a randomized controlled trial. Eur Heart J 2022 Oct 3;43(Supplement_2). [doi: 10.1093/eurheartj/ehac544.2440]
28. Nuijten R, Van Gorp P, Khanshan A, et al. Evaluating the impact of adaptive personalized goal setting on engagement levels of government staff with a gamified mHealth tool: results from a 2-month randomized controlled trial. JMIR Mhealth Uhealth 2021;10(3):e28801. [doi: 10.2196/28801]
29. White HD. Waiting room computer tablets to improve health literacy and cardiovascular outcomes. Heart 2021 Oct;107(20):1607-1608. [doi: 10.1136/heartjnl-2021-319756] [Medline: 34321287]

XSL•FO
RenderX

30. Royse LA, Baker BS, Warne-Griggs MD, et al. “It’s not time for us to sit down yet”: how group exercise programs can motivate physical activity and overcome barriers in inactive older adults. Int J Qual Stud Health Well-being 2023 Dec;18(1):2216034. [doi: 10.1080/17482631.2023.2216034] [Medline: 37220010]
31. de Oliveira GL, Sobrinho ACS, Benjamim CJR, et al. Physical training improves physical activity levels but is associated with amplification of sedentary behavior in older women. Front Public Health 2023;11:1180901. [doi: 10.3389/fpubh.2023.1180901] [Medline: 37333564]
32. Domingos C, Correia Santos N, Pêgo JM. Association between self-reported and accelerometer-based estimates of physical activity in Portuguese older adults. Sensors (Basel) 2021 Mar 24;21(7):2258. [doi: 10.3390/s21072258] [Medline: 33804834]
33. Stockwell S, Schofield P, Fisher A, et al. Digital behavior change interventions to promote physical activity and/or reduce sedentary behavior in older adults: a systematic review and meta-analysis. Exp Gerontol 2019 Jun;120:68-87. [doi: 10.1016/j.exger.2019.02.020] [Medline: 30836130]
34. Muellmann S, Forberger S, Möllers T, Bröring E, Zeeb H, Pischke CR. Effectiveness of eHealth interventions for the promotion of physical activity in older adults: a systematic review. Prev Med 2018 Mar;108:93-110. [doi: 10.1016/j.ypmed.2017.12.026] [Medline: 29289643]
35. Aguiar M, Trujillo M, Chaves D, Álvarez R, Epelde G. mHealth apps using behavior change techniques to self-report data: systematic review. JMIR Mhealth Uhealth 2022 Sep 9;10(9):e33247. [doi: 10.2196/33247] [Medline: 36083606]
36. Pérez-Saiz L, Ferri Sanz M, Ferrando M, et al. Health access, health promotion, and health self-management: barriers when building comprehensive ageing communities. Int J Environ Res Public Health 2023 Oct 3;20(19):6880. [doi: 10.3390/ijerph20196880] [Medline: 37835150]
37. LaMonica HM, Davenport TA, Roberts AE, Hickie IB. Understanding technology preferences and requirements for health information technologies designed to improve and maintain the mental health and well-being of older adults: participatory design study. JMIR Aging 2021 Jan 6;4(1):e21461. [doi: 10.2196/21461] [Medline: 33404509]
38. Eaton CK, McWilliams E, Yablon D, et al. Cross-cutting mHealth behavior change techniques to support treatment adherence and self-management of complex medical conditions: systematic review. JMIR Mhealth Uhealth 2024 May 1;12:e49024. [doi: 10.2196/49024] [Medline: 38717433]

Abbreviations

BF: Bayes Factors
BF_incl: Bayes Factors for inclusion
GLMM: generalized linear mixed model
IPAQ: International Physical Activity Questionnaire
JASP: Jeffreys’s Amazing Statistics Program
MoCA: Montreal Cognitive Assessment
SF-36: 36-Item Short-Form Health Survey
THL-PA: Test of Health Literacy for Portuguese-speaking Adults

Edited by A Mavragani, J Sarvestan; submitted 11.05.25; peer-reviewed by A Billis; revised version received 18.10.25; accepted 20.10.25; published 10.12.25.

Please cite as:
da Silva Sobrinho AC, da Silva Rodrigues G, de Oliveira GL, de Oliveira Gomes GA, Bueno Júnior CR
Everyday Digital Support to Promote Health and Literacy Among Older Adults: 14-Week Randomized Digital Pilot Trial by Engagement Level
JMIR Form Res 2025;9:e77319
URL: https://formative.jmir.org/2025/1/e77319
doi:10.2196/77319

© Andressa Crystine da Silva Sobrinho, Guilherme da Silva Rodrigues, Guilherme Lima de Oliveira, Grace Angélica de Oliveira Gomes, Carlos Roberto Bueno Júnior. Originally published in JMIR Formative Research (https://formative.jmir.org), 10.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Evaluating a Clinical Decision Support Tool for Cancer Risk Assessment in Primary Care: Simulation Study of Unintended Weight Loss

Javiera Martinez-Gutierrez[1,2], MPH, MD; Sophie Chima[2,3], PhD; Lucas De Mendonca[2], MD; Alex Lee[2], PhD; Barbara Hunter[2], PhD; Jo-Anne Manski-Nankervis[2,4], MBBS, PhD; Deborah Daly[2,5], BA, BBus, MBA; George Fishman[5], MBA; Kit Huckvale[6], PhD; Fong Seng Lim[7], MD; Benny Wang[8], PhD; Craig Nelson[9,10], MBBS, PhD; Brian Nicholson[11], BSc, MSc, DPhil; Jon Emery[2,4], MBBCh, MA, DPhil

[1]Department of Family Medicine, Faculty of Medicine, Pontificia Universidad Catolica de Chile, Santiago, Chile

[10]Department of Medicine, Western Health, The University of Melbourne, Melbourne, Australia

[11]Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, United Kingdom

[2]Department of General Practice and Primary Care, Faculty of Medicine, Dentistry and Health Sciences, The University of Melbourne, Medical Building, Grattan Street, Parkville, Australia

[3]Collaborative Centre for Genomic Cancer Medicine, The University of Melbourne, Parkville, Australia

[4]Primary Care and Family Medicine Department, Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore, Singapore

[5]Primary Care Collaborative Cancer Clinical Trials Group (PC4), Melbourne, Australia

[6]The Digital Health Validitron, Centre for Digital Transformation of Health, The University of Melbourne, Parkville, Australia

[7]National University Polyclinics, Singapore, Singapore

[8]Singapore Primary Care Cancer Network (SPriNT), Singapore, Singapore

[9]Western Health Chronic Disease Alliance, Western Health Melbourne, Melbourne, Australia

Corresponding Author:
Javiera Martinez-Gutierrez, MPH, MD
Department of Family Medicine, Faculty of Medicine, Pontificia Universidad Catolica de Chile, Santiago, Chile

Abstract

Background: Early cancer detection is crucial, but recognizing the significance of associated symptoms such as unintended weight loss in primary care remains challenging. Clinical decision support systems (CDSSs) can aid cancer detection but face implementation barriers and low uptake in real-world settings. To address these issues, simulation environments offer a controlled setting to study CDSS usage and improve their design for better adoption in clinical practice.

Objective: This study aimed to evaluate a CDSS integrated within general practice electronic health records aimed at identifying patients at risk of undiagnosed cancer.

Methods: The evaluation of a CDSS to identify patients with unintended weight loss was conducted in a simulated primary care environment where general practitioners (GPs) interacted with the CDSS in simulated clinical consultations. There were four possible clinical scenarios based on patient gender and risk of cancer. Data collection included interviews with GPs, cancer survivors (lived-experience community advocates), and patient actors, as well as video analysis of GP–CDSS interactions. Two theoretical frameworks were employed for thematic interpretation of the data.

Results: We recruited 10 GPs and 6 community advocates, conducting 20 simulated consultations with 2 patient actors (2 consultations per GP: 1 high-risk consultation and 1 low-risk consultation). All participants found the CDSS acceptable and unobtrusive. GPs utilized CDSS recommendations in three distinct ways: as a communication aid when discussing follow-up with the patient, as a reminder for differential diagnoses and recommended investigations, and as an aid to diagnostic decision-making without sharing with patients. The CDSS's impact on patient–doctor communication varied, facilitating and hindering interactions depending on the GP's communication style.

Conclusions: We developed and evaluated a CDSS for identifying cancer risk in patients with unintended weight loss in a simulated environment, revealing its potential to aid clinical decision-making and communication while highlighting implementation challenges and the need for context-sensitive application.

(JMIR Form Res 2025;9:e79208) doi:10.2196/79208

XSL•FO
RenderX

KEYWORDS

digital health; clinical decision support; cancer prevention; early detection; general practice; primary care

Introduction

Cancer remains one of the most significant health burdens in modern society, with early detection playing a crucial role in improving survival rates. However, identifying cancer in primary care settings presents a considerable challenge due to the nonspecific nature of many presenting symptoms [1,2]. Fewer than 50% of patients with cancer initially present to general practice with *red-flag* symptoms (eg, site-specific symptoms such as a breast lump); however, they more commonly present with nonspecific signs and symptoms [3]. For patients with cancer, presenting with nonspecific symptoms can introduce diagnostic complexity, as such symptoms are often present in other, more common diseases.

An important but nonspecific presentation of cancer is unintended weight loss (UWL), which is associated with 10 different types of cancer as well as several other conditions in primary care [4-6]. However, it represents a clinical challenge, especially when occurring as an isolated symptom. A large cohort study showed that patients with UWL have a risk of 3% or more to be diagnosed with cancer [7], which is an international threshold for urgent investigation [8]. Despite being a risk factor for many cancers, clinical recognition as a symptom of cancer is low, and studies have shown that UWL is not easily identified or acted upon in clinical practice [9,10]. Timely investigations of cancer in general practice are vital, with a need to identify patients who are at increased risk of an undiagnosed cancer.

Clinical decision support systems (CDSSs) have been developed to aid in early cancer detection and are generally well-accepted by health care professionals [11]. Briefly, a CDSS utilizes algorithms underpinned by clinical guidelines and academic literature, combining this with patient information to provide guideline or evidence-based recommendations to health care professionals, often at the point of care [12,13]. However, many well-designed, usable, and acceptable CDSSs have low uptake when embedded in real-world clinical practice. A study in the United Kingdom showed that only 17% of general practitioners (GPs) with access to a CDSS for cancer risk in their practice were likely to use it [14]. The real impact of CDSSs in improving clinical care has been described as low to moderate [15].

There are several barriers to the implementation of CDSSs in real-world settings. A systematic review on the use of a CDSS for early cancer detection in primary care described mistrust in the recommendations, lack of time, and the CDSS interfering with the clinical workflow as just some of these barriers. [11]. To address these challenges and gain insights into the practical application of CDSSs without exposing patients to potential risks, simulation clinical environments or laboratories offer a cost-effective alternative to piloting prototype CDSSs in real-life scenarios, particularly for less common clinical presentations. Simulation laboratories to test digital interventions provide an opportunity to observe interactions between health care providers and the CDSS, allowing researchers to study how clinicians actually use the tool in a controlled environment. This approach enables the collection of valuable data not only on tool usage but allows researchers to observe the interaction and assess workflow, communication, and other factors that influence adoption. They can also facilitate discussions about potential improvements by gaging users' impressions immediately after the interaction has occurred [16].

In light of these considerations, we sought to observe the use of a CDSS designed to aid in the recognition of UWL as a potential cancer symptom. Our study aimed to assess both GPs' and patients' perceptions of such a tool, with the ultimate goal of improving early cancer detection in primary care settings.

Methods

The Future Health Today Tool

Future Health Today (FHT) is a software that can be integrated with a general practice electronic medical record (EMR) [17]. It consists of two core components: a CDSS point of care prompt and a web-based portal, which includes an audit and recall tool, quality improvement management, and access to education and other resources. The prompt also provides direct access to the reasons for the recommendations and evidence-based information relevant to follow-up. FHT has been co-designed with GPs and consumers, with the purpose of optimizing patient care [18]. Different modules are developed for use in FHT. The FHT UWL module was developed to flag patients with a symptom of UWL recorded in the medical record in the past 6 months. The UWL algorithms take into account age, sex, and any recent abnormal test results, with different recommendations dependent on these factors. An example of a recommendation as it appears in the EMR is shown in Figure 1.

XSL•FO
RenderX

Figure 1. Example of a recommendation for unexpected weight loss as it appeared in the general practice electronic medical record (fictitious patient).

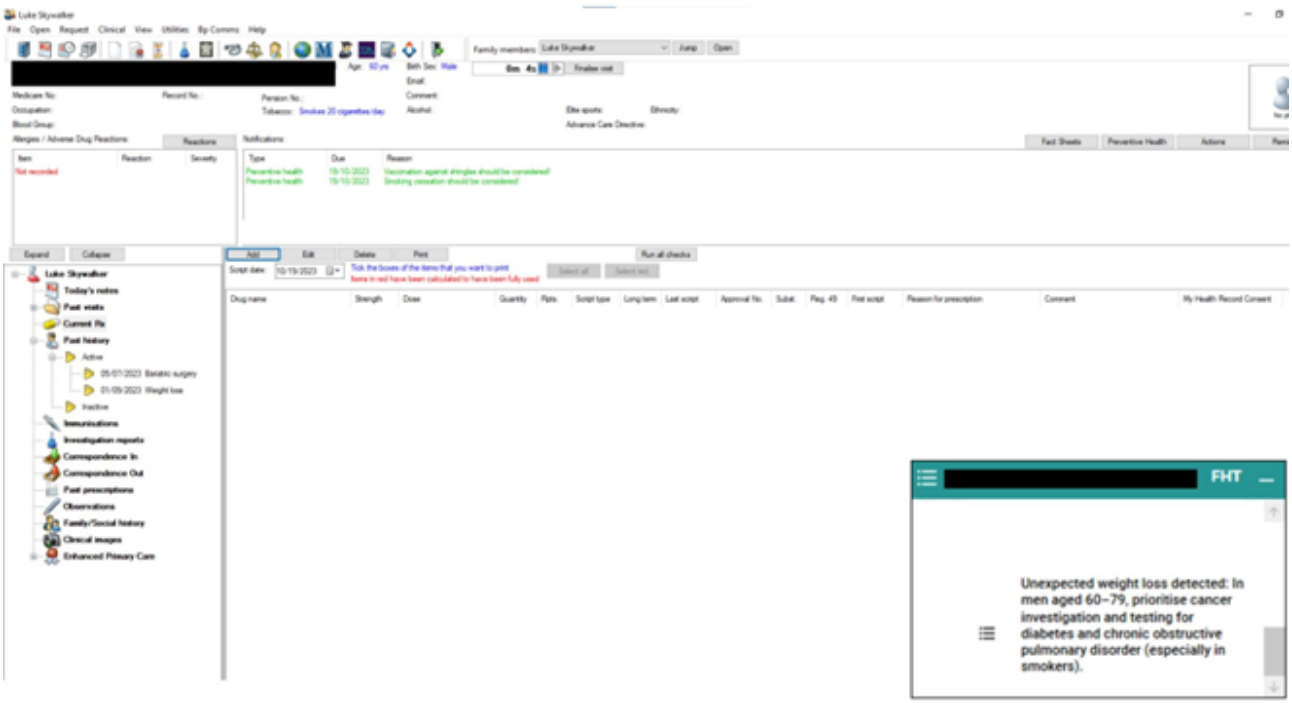

Simulated Consultations

The study was carried out in the University of Melbourne's Digital Health Validitron SimLab at the Centre for Digital Transformation of Health [16,19]. The SimLab has a simulated general practice clinic, incorporating a virtual "sandbox" (a virtual machine that is used to run software in a testing environment), which allows for the UWL algorithm to be used on the Best Practice and Medical Director General Practice EMRs in "near to real" scenarios. It is estimated that approximately two-thirds of Australian general practices use Best Practice and one third use Medical Director, with approximately 3.6% of practices using other EMRs [20].

Four hypothetical patient scenarios were designed to align with the UWL module recommendations. These included one female high-risk and one female low-risk scenario and one male high-risk and one male low-risk scenario. Each GP was exposed to one high-risk and one low-risk scenario according to the sex of the actor available. A summary of the scenario characteristics can be found in Table 1. The full patient scenarios can be found in Multimedia Appendix 1. The simulated consultations were performed by actors from the Department of Medical Education at the University of Melbourne, who were experienced in delivering simulated patient (SP) scenarios. The definition of an actor in this scenario was a professional "trained to reproduce the components of real clinical experience" [21].

Table . Hypothetical patient scenarios for the simulated consultations: overview of patient characteristics.

Sex and risk level	Age (y)	Symptoms	Other risk factors
Female			
High	65	Concerned about UWL[a], fatigue	New onset diabetes, anemia, and thrombocytosis
Low	60	History of mild UWL and abdominal discomfort	None
Male			
High	60	Concerned about UWL, abdominal discomfort, and change in stool color	Family history of colorectal cancer, mild anemia, and thrombocytosis
Low	54	History of mild UWL and abdominal discomfort	None

[a]UWL: unintended weight loss.

Study Participants

Recruitment of GPs occurred using various methods: invitation (email or in-person) using the University of Melbourne Department of General Practice and Primary Care contacts; dissemination of study information through advisory groups (eg, Future Health Today Advisory Group) and groups associated with the University (VicREN, a practice-based

XSL•FO
RenderX

primary care research network managed by the University of Melbourne and the Primary Care Collaborative Cancer Clinical Trials group); and GP groups on social media platforms (eg, WhatsApp). The recruitment messaging and materials included information about the simulation study, the location of the study, the time required to participate, and reimbursement. Participants were recruited using purposive sampling, with the aim of recruiting a diverse sample in terms of experience, gender, and age. Inclusion criteria applied the following: registered GP, familiar with either Best Practice or Medical Director EMR, who was able to attend the University of Melbourne SimLab for up to 1 hour.

Lived experience community advocates with a history of a cancer diagnosis (henceforth "community advocates" or "CAs") were recruited via the Department of General Practice and Primary Care newsletter, word of mouth via friends and family, and through organization newsletters such as the Pancare Foundation, Cancer Voices Australia, and the Victorian Comprehensive Cancer Centre. The recruitment messaging and materials included information about the study, the time required to participate, and reimbursement amount. Similar to GP recruitment, purposive sampling was used. CAs were eligible for inclusion if they were over 40, had experience as a general practice patient, and presented with a nonspecific symptom of cancer before being diagnosed.

Interested participants (GPs and patients) were sent further details of the study using a Plain Language Statement. Informed consent was collected from all participants.

Data Collection

Simulated Consultations

Before the simulated consultations began, GPs were introduced to the space and given a brief overview of the tool. GPs had limited information about the FHT module before participating. They were aware that the study evaluated a module for chronic disease but were unaware it was specific to cancer or that it would involve recommendations for UWL. During recruitment, GPs were told that they may be asked to participate in up to two consultations with different patient scenarios (the number of simulated consultations was confirmed on the day of the simulation), but GPs were unaware that two consultations were designed to simulate two different patient risk profiles (one high-risk and one low-risk).

Researchers directed the GP to the appropriate patient record before starting the simulation. GPs were advised to review the patient record (if this was standard practice for them) and call the patient in when ready. The Digital Health Validitron SimLab has a general practice "waiting room," so the GP and patient actor did not meet before the simulated consultation. GPs were advised to keep the consultation to 10 - 15 minutes, in line with the usual duration of GP consultations in Australia. The simulation was filmed, audio recorded, screen recorded (ie, recording the GP's mouse/cursor movement and clicks when using the computer), and observed by the researchers through a 1-way mirror. Each GP enacted two patient scenarios, one lower and one higher risk, determined by the gender of the patient actor allocated to their session.

GP and Actor Interviews

Directly after the two simulated consultations, interviews were conducted with GPs (JMG) and actors (SC). The GP interviews focused on the acceptability, feasibility, and workflow associated with using the UWL module. The interview guides (Multimedia Appendix 2) were developed using 2 frameworks: Acceptability of Health Care Interventions Framework [22] and the Sociotechnical Model for Evaluation of Digital Interventions Framework [23]. Formal data saturation was not deemed necessary to assess study outcomes [24].

Patient actors were asked about their interaction with the GP, the recommended investigations and follow-up plans discussed, and how the tool and its content affect communication (Multimedia Appendix 3).

Lived-Experience CA Interviews

From the 20 simulated consultations, two contrasting videos were selected to highlight different GP approaches to incorporating the prompt's recommendations. One video showed a GP using the FHT prompt to communicate with the patient, while the other depicted a GP who neither used nor acknowledged the prompt during the consultation. These two videos were shown to CAs, followed by a semistructured interview via Zoom conducted by JMG. The interview guide (Multimedia Appendix 4) was developed using the Acceptability of Health Care Interventions Framework [22] and the Sociotechnical Model for Evaluation of Digital Interventions Framework [23].

All interviews (GP, actor, and CAs) were audio recorded and lasted approximately 30 - 45 minutes.

Data Analysis

Otter.Ai was used for the initial transcription, which was then checked by a researcher (SC). Transcripts were then uploaded to NVivo 14, coded, and thematically analyzed [25] independently by JMG and SC. The acceptability of health care interventions underpinned the analysis of GP, CA, and patient actor interviews. Video recordings of the simulated consultations were also analyzed. The video analysis provided additional depth to the overall analysis and was conducted using the Sociotechnical Model (Figure 2). Findings from the interview analysis complemented the themes identified in the video analysis. The two frameworks were used to inform relevant themes, and additional themes were added after the initial review of the data. A subset of participants was asked to provide feedback on the analysis. We report results using the COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist (Checklist 1) [26].

XSL•FO
RenderX

Figure 2. Frameworks used for analysis of interviews and simulated consultations. Adapted from Sekhon's [22] acceptability of health care interventions and the sociotechnical model for evaluation of digital interventions by Singh and Sittig [23]. Main themes reported are shown in blue.

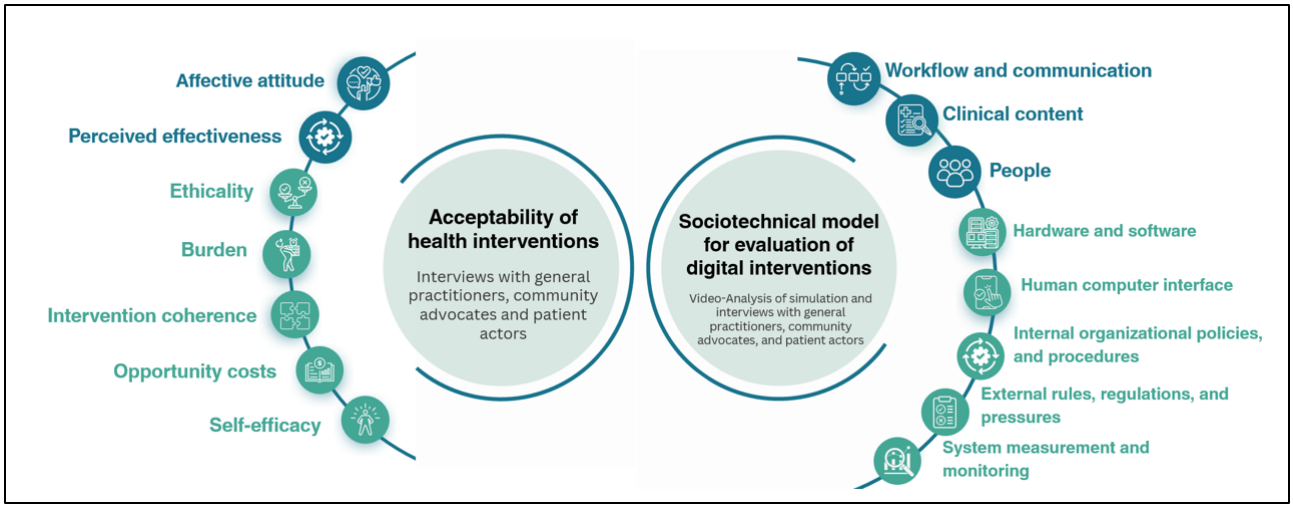

Ethical Considerations

Ethical approval was granted by the University of Melbourne Human Research Ethics Committee STEMM Three (ethics ID: 28024). Written informed consent was obtained from all study participants before or on the day of the simulation, and all participants were informed of their ability to opt out at any moment. All transcripts and audio recordings were securely stored on an institutional password-protected server without any personal identifiers to maintain participant confidentiality. GPs and CAs were reimbursed for their time (Aus $150 [US $98]). Actors were compensated in accordance with the University of Melbourne's established salary scale.

Researchers' Characteristics and Reflexivity

Two researchers were directly involved in data collection (simulation and interviews) and analysis (JMG and SC). JMG is the lead investigator for this study and was assisted by SC. JMG is a female GP and PhD candidate and midcareer researcher at the Department of General Practice and Primary Care, University of Melbourne. SC is a female early career researcher and postdoctoral fellow in the Department of General Practice and Primary Care. Both researchers have experience in qualitative research, in conducting interviews, and in the development of clinical decision support tools.

Results

Simulated Consultations

We conducted a total of 20 simulated consultations involving 10 GPs. Each GP participated in 2 scenarios: 1 high-risk scenario and 1 low-risk scenario. Further, 7 GPs tested the CDSS with the male scenario, and 3 GPs tested it with the female scenario. The patient's gender in each scenario was determined by the availability of male or female actors on the day of simulation. All GPs were interviewed following the simulations; 2 actors performed the clinical scenarios, and they were interviewed after each GP finished their two clinical interactions (10 interviews with 2 actors). Additionally, 6 CAs were interviewed following their observation of 2 exemplar simulation videos.

GP demographics are summarized in Table 2. All GPs interviewed worked in metropolitan areas of Victoria. There was an even distribution of gender in the sample. Half of the GPs were between 30 and 39 years of age and had <5 years in practice. The age of participating GPs ranged from 31 to 63 years. Interview length ranged from 21 to 32 minutes.

Characteristics of CAs are detailed in Table 3. Half of the CAs were female, and half were between 60 and 69 years of age, with ages ranging from 48 to 71 years. Their year of cancer diagnosis ranged from 2011 to 2018, and they all had consulted their GP with a nonspecific symptom prior to diagnosis, although not necessarily UWL. Interview length ranged from 27 to 57 minutes.

The two actors each played a patient scenario of a middle-aged and older adult. Both patient actors had personal experience as general practice patients. Interviews with the actors were designed to be shorter, with interview length ranging from 5 to 9 minutes.

Characteristics of all participants can be found in Tables 2 and 3.

Consequently, the total number of reported symptoms exceeds the number of CAs interviewed.

XSL•FO
RenderX

Table . Characteristics of GPs[a] who participated in the simulated consultations and qualitative interviews.

GP characteristic	GPs (n=10), n
Age (y)	
30 - 39	5
40 - 49	2
60+	3
Gender	
Male	5
Female	5
Years of experience	
1 - 5	5
6 - 20	2
>20	3
Practice setting	
Urban	10

[a]GP: general practitioner.

Table . Characteristics of community advocates who participated in qualitative interviews.

Community advocate characteristic	Community advocates (n=6)
Age (y)	
40 - 49	1
50 - 59	1
60+	4
Gender	
Male	3
Female	3
Education	
Finished high school	2
Postgraduate studies	4
Unspecific symptoms experienced prior to their cancer diagnosis[a]	
Urinary symptoms	1
Chest pain	1
Sore throat	2
Muscle sore	1
Lower back pain	1
Persistent flu-like symptoms	1
Fatigue	1
Abdominal pain	1
Change of bowel habits	1

[a]The number of community advocates experiencing each symptom is reported. Several community advocates reported experiencing multiple symptoms prior to their diagnosis.

XSL•FO
RenderX

GP and CA Interviews

We present the results for the two main themes in the theoretical framework of acceptability: affective attitude and perceived effectiveness.

Affective Attitude

There was a positive attitude toward the UWL module in general, although the reasons for perceived value differed by age and experience. Younger, less experienced GPs tended to value it because it aligned with the way they work (eg, checking guidelines, taking extra time). More experienced GPs valued it primarily as a reminder system.

> *Particularly as a younger GP, who looks a lot of things up as I go, I always have eTG up (Therapeutic guidelines), I have health pathways up, I'll have lots of clinical guidelines up....But that's what this does for you, it sort of keeps you up to date, saves you time in that it gives you the key information quickly that you would be looking for.* [GP10]

Prompt fatigue was seen as a barrier for long-term use by some GPs. Half of the GPs felt compelled to follow the recommendation or, when asked directly, questioned the ethical and legal implications of not following them.

> *Yeah, it makes me feel like it's a must.* [GP2]

> *The only issue I have would be, if I choose to ignore it, and then someone gets diagnosed with cancer, and then I go for a malpractice suit.* [GP4]

CAs also thought that a module to identify people at risk of undiagnosed cancer had value as a good reminder system. They reflected on how busy GPs can be and felt it could also be used as reassurance to the patients that their GP was considering all diagnostic possibilities. One CA, while acknowledging its usefulness, expressed their concern for the need for tools like this; their perception was that GPs should be able to conduct their clinical practice without any external help.

> *The only thing that concerns me is that the doctor shouldn't need this. I think a competent doctor should know that this patient is presenting with an abnormality, there's something wrong.* [CA3]

Perceived Effectiveness

Some GPs reported that the CDSS positively impacted their consultation and thought process. They reported that the prompt helped them prioritize investigations and consider differential diagnoses. They also thought it would be a useful tool for safety-netting and ensuring patients are followed up until diagnostic resolution. All GPs found the tool useful as a reminder system for their clinical train of thought but were clear the tool did not replace their clinical acumen.

> *It definitely impacted on my thought process approach to the consultation. I do always like to start off with allowing the patient to talk about why they're here and what's their concerns, before getting too heavily into more interrogation. But there was a definite need for interrogation based on the prompt, you know, looking at those major, possible causes of unexplained weight loss, both organic and psychological and psychosocial factors as well. So, it helped to structure that.* [GP6]

However, most GPs, while acknowledging the potential benefits, did not think the tool would change or impact their own clinical practice, often considering that it would be useful for other GPs.

> *I personally, probably wouldn't look at it much. But I think it would be really good, something like this, for people who are just finished their training or sort of registrars and that sort of thing.* [GP9]

Two GPs reported concerns about EMR data, and the accuracy or sensitivity of the module acknowledging data recording in everyday practice might not be complete enough for this kind of technology. Most GPs reported concerns about trying to explain a very unlikely cancer diagnosis.

> *Because, sadly, the notes don't always accurately reflect the situation with the patient. They're not always updated. And so, if they're pulling things out of here, they may not be correct.* [GP5]

CAs thought it would be most useful to use patients who are new to the clinic or do not have an established relationship with their GP, as they would remind them of possible risk factors and could enable effective communication. CAs and SPs commended the tool's ability to enhance follow-up care and patient guidance, viewing these features as significant improvements to the consultation process

> *With the pop up, what I thought immediately was, he's not going to miss anything. You know, he can be stressed himself, he could be busy. You know, he might have just had to save someone's life for all I know. And next minute he's in here, dealing with me. And you know, he's human. What I found, what I felt, was when the pop up came up, I was like, he's not going to miss anything.* [SP4]

> *I think it's a great idea. I really do. I think it's a good way to hopefully get a diagnosis sooner, like, without any delay to sort of get right to it. And to not waste time doing this step when you're really better off going straight to this step here.* [CA5]

One CA suggested the tool would be especially useful in rural areas, given access to clinical services can be scarce. They also suggested developing multilingual versions to accommodate diverse populations.

Simulated Consultations and Video Analysis

Simulated Consultation Characteristics and Theoretical Framework

The average (range) simulated consultation between GPs and patients lasted 9 minutes and 14 seconds (6 min 30 s to 13 min 23 s). GPs spent an additional average (range) of 1 minute 30 seconds (0-4 min) looking at the recommendations and other relevant information in the EMR (pathology tests and previous visits) before calling the SP in.

We analyzed the interactions between the GP, the tool, and the SP, underpinned by the dimensions of the Sociotechnical Model

XSL•FO
RenderX

for Digital Health Interventions. We then identified facilitators and barriers to the use and implementation of the CDSS within each theme. Here we describe the main themes from what was observed in the simulated consultation, complementing these findings with the perceptions of GPs, CAs, and SPs from the interview analysis, where relevant.

Workflow and Communication

In the simulated consultations, all GPs reviewed the EMR and CDSS recommendations for a few minutes before calling the patient in. This aligned with what was reported in the interviews, where GPs thought the recommendations were feasible to implement into their workflow. They were unobstructive and easy to read, and most said that they would review the recommendations before the patient came for the consultation in real life. It is important to note that most GPs did not review all features of the prompt (eg, reason for recommendation accessed via hovering over text, links to tailored resources accessed via clickable menu items).

In observing the consultation, half of the GPs acknowledged the prompt with the SP. There were two distinct ways in which this was done during the consultation, in regard to communication: (1) as an aid to involve the patient in the decision-making process and (2) as a reminder or checklist for the doctor to use; some doctors though chose not to discuss the prompt at all in the consultation. These are described further in Table 4.

XSL•FO
RenderX

Table . Results from the simulation study: how the tool was used when communicating with the patient.

Theme and description	Quote
Aid to involve patients in decision-making	
GPs[a] integrated the prompt into patient consultations in various ways, often involving patients in decision-making. Some used it to explain their reasoning for follow-up investigations, openly discussing the recommendation with patients. While all GPs adhered to certain aspects of the recommendation, some also explained why they might deviate from it in specific instances.	• "I have a little pop-up here, I don't know if you've noticed but it says…doing an FOBT…look, I don't think a poo test at this point in time is going to add much more..." (GP4, consultation 2 [6 min 49 s]) • "Now if you don't mind, I am just going to have a quick look at my screen, because I have a special thing that tells me…my computer can give me some prompts that actually helps me with your symptoms so that is quite handy." (GP9, consultation 1 [5 min 10 s])
Some GPs utilized the prompt to broach sensitive topics with new patients, such as mood disorders or potential cancer diagnoses, which might otherwise be challenging to address in an initial consultation.	• "Now, I have just received this alert, and I am just going to have a look at what it says. Ok, so this is just prompting me to, as we mentioned, consider testing for your thyroid function as a possible cause, but then also mood disorders, which we started to talk about. Perhaps what I can also add is a screening questionnaire for mental health conditions...." (GP2, consultation 1 [9 min]). • "And the depression? Yeah, I went through that. I think that was, it's a useful reminder to have there" (GP5)
GP reminder and checklist	
Some GPs used the prompt as a checklist for follow-up investigations, often without detailed explanations to patients. In these cases, they focused on addressing key aspects of the recommendation, such as potential diagnoses and suggested tests.	• "There is little thing on the screen…(actor asks "does that help you does it")…it sort of aids the GP at the desktop, just a sort of little prompt." (GP8, consultation 1 [5 min 25 s]) • "I'll have a little look at a reference list here" while discussing list of investigations. (GP6, consultation 1 [10 min 50 s])
When GPs used the prompt as a checklist without explaining it to patients, both patient actors and community advocates observed a communication gap. Although most GPs did not perceive the tool as disruptive, some concerns arose about its potential to hinder doctor–patient communication.	• "I think that there's room for it to interrupt workflow and your interaction with the patient in a negative way, depending on how you use it." (GP2) • "Yeah it did have an impact on the therapeutic relationship I suppose that there was a dis – that there was a barrier between us." (SP[b] 1)
After delivering the recommendation, some GPs consulted the tool to check if there was something missing. This review often prompted additional questions about risk factors or led to the inclusion of further tests, enhancing the thoroughness of the consultation.	• "It didn't bring up any diagnoses that I hopefully wouldn't have discussed anyway. Oh, except for maybe asking the questions about lung cancer. Yeah, about smoking." (GP10) • At the end of the consultation, GP looks at the screen, reaches for the mouse and says: "Actually, I should have checked. Are you up to date with your cancer screening?" The patient says she is up to date. (GP2, consultation 1 [7 min 5 s])
Not addressed during consultation	
Half of the GPs opted not to discuss the CDSS[c] tool during consultations, despite considering its recommendations. Reasons varied: some felt confident explaining investigations without referencing the tool, others were cautious about prematurely mentioning cancer, and one GP feared that acknowledging the tool might create distance in the patient interaction.	• "I didn't refer to it as saying, 'oh, this program has recommended I do this.' Because I think I would only do that if I was sort of trying to distance myself from the recommendation. None of these recommendations are surprising to me, or, you know, difficult to explain. I think they're all quite common sense. So I, I just use that as a tool, but I'm not saying you know, 'FHT suggests we do this, so we're going to do this'...." (GP7)

[a]GP: general practitioner.

[b]SP: simulated patient.

[c]CDSS: clinical decision support system.

The use of the tool and the themes around communication was evaluated by both the CAs and SPs. Irrespective of whether the GPs actively discussed their use of the tool or not, both CAs and SPs appreciated GPs who included them in the conversation.

For the CAs, their relationship with their GP and how that shaped their experiences was a core theme throughout the interviews. There were different preferences for communication. Some patients said they would like to be brought into the discussion around use of the tool and would like it explained, while others did not care as long as the GP sounded sure of what the next steps would be.

> *There wouldn't be any particular reason why you would show the patient that pop up, you know, if we calling it that the pop up on the screen, but rather, I would be quite comfortable if the GP said to me in a fairly neutral way "I'm just, you know, I'm being*

XSL•FO
RenderX

> *reminded here, because of your presentation that I should check out the following. So that's what I'm going to do, just to make sure we can eliminate it." So being told what the person is looking at on the screen also meant you're being included.* [CA6]

There were a variety of communication strategies used by GPs throughout the simulated consultations. CAs and patient actors noted how the system was used and/or communicated was very individual. There were instances of both effective and less effective communication as discussed by the patient actors and preferences for how they would like the information delivered, highlighting that there is no one size fits all. The tool could become a facilitator or a barrier to communication depending on its use.

> *I don't think would really matter a great deal whether or not the doctor mentioned the prompt, you know. As long as it's there to remind the doctor.* [CA4]

> *It all depends on their bedside manner; his bedside manner is really good. I've done some simulation with the students over here, and they are nowhere near as good as him, because he's had years of practice.* [SP2]

> *It's how they interact with a patient, which is obviously very important.* [CA1]

Clinical Content

GPs thought the wording of the recommendations ensured clinical autonomy (eg, the use of language such as "consider," "recommended," instead of more directive words). They also reported that information was well organized and easy to follow. In general, content was considered appropriate and clinically relevant albeit very nonspecific. GPs said they liked that it was symptom-based rather than diagnosis-based and that it included several symptoms at the same time, reflecting real life.

Most GPs' actions in the consultations were concordant with the tool's recommendations, suggesting standard pathology testing in primary care to rule out the most common causes of unexpected weight loss by age (eg, thyroid disease, full blood count, general biochemistry, kidney, pancreas, and liver function).

> *What is written in there all seems really relevant to me. So, I like that it combined a number of things like, not just the unexpected weight loss, but the patient's, the patient's gender, and age to then sort of help stratify risk and saying that, you know, in a patient with unexpected weight loss, of this age, this gender, the things to think about.* [GP7]

Some tests were considered and discarded after clinical judgment. CA-125 for ovarian cancer was not ordered by any of the GPs in this study despite it being recommended. Four GPs opted against using the recommended immunochemical fecal occult blood test in symptomatic patients, instead favoring more direct diagnostic procedures such as colonoscopy.

> *Like you have FOBT testing two years ago, you're out of date, doing another one now is not really appropriate, because not really screening you anymore.* [GP5]

For low-risk patients, some GPs favored a phased approach, initially ordering blood tests and deferring imaging to subsequent consultations based on initial results. These GPs expressed concerns about overtesting, citing potential financial burdens and unnecessary anxiety associated with comprehensive investigation. In contrast, for high-risk patients, all GPs recommended more advanced diagnostic procedures, such as endoscopic examinations or computed tomography scans, reflecting a more aggressive investigative strategy.

> *The investigations are, you know, also feasible, but I do think as well, system wide issues that probably are unrelated is access to things like these, like you can request all these tests, but whether a patient is going to go and do them and, you know, do they have access to private health insurance to get their private scopes?* [GP2]

CAs did not care much about the content of the prompt, as long as it was evidence based, and their GP trusted the information.

> *I derive a lot of comfort in knowing that it is not random, there is some sort of process, there's some sort of procedure and it is, you know, it is evidence based, and it is also because of, you know, what I said and what I'm presenting with* [CA2]

Notably, only a minority of GPs delved into the additional resources provided within the tool, which contained supplementary information. Most reported that the tool itself was sufficiently clear, negating the need for further exploration.

People

For both GPs and CAs, it mattered who was using the module. As discussed earlier, differences in years of experience influenced the GPs' perception of the tool, how they would use it, and what parts they liked. Younger GPs took more time looking at guidelines and considering the recommendations. They mentioned they need to know where the information is coming from to be able to trust it (eg, guidelines sponsored by the Royal Australian College of General Practitioners).

More experienced GPs were observed to be comfortable in the consultation and simulation scenario. Even though they looked at the screen and keyboard when writing, they alternated between the patient and the screen, focusing their attention mostly on the patient. They also felt more comfortable using the tool as an aid to communicate decision-making. Of the five GPs that used and discussed the tool with the patient, four had more than 5 years' experience as a GP. So, even though younger GPs appreciated the tool as a reminder for their thought process, more experienced GPs were able to use it as a communication tool as well as apply its clinical content.

> *We've got a nice little pop-up; we have things that listen in the background. And it is saying yes you do have weight loss, and they are worried about things like thyroid function, diabetes and depression. I mean mood wise, how have you been?* [GP4 with >20 years' experience]

XSL•FO
RenderX

> *I mean, I still interpret. Like here, I still interpret what it's saying, from my own personal experience and the person in front of me. So, it's, it's not telling me what to do. It's given me advice about what could be done.* [GP4 with >20 years' experience]

> *Yeah, I think it's a good idea. I do think, particularly at the early stage of my career now, I probably over investigate, compared to more experienced GPs.* [GP10 with <5 years' experience]

As mentioned before, the CA's and actors' perceptions on how the tool was used seemed very dependent on the GP's communications skills—how the tool made them think or feel was directly related to how the GPs relayed the information or used the tool.

> *If I look at it, one was actually using the tool. I hope I get this right in my expression. The tool was driving the session -the second one. First one, the tool was a tool to use, and the doctor was driving the session.* [CA3]

Discussion

Acceptability and Use of UWL CDSS

We tested a CDSS to identify patients with UWL at risk of undiagnosed cancer in a simulated environment. The tool was generally deemed acceptable, particularly as a reminder of differential diagnoses and recommended investigations. However, in line with existing literature, concerns about overreferral and overtesting, as well as reluctance to discuss a potential cancer diagnosis, emerged as potential barriers [27].

The clinical content provided was considered adequate. GPs selectively utilized the information, according to their clinical reasoning influenced by internal (personal experience), external (cost and accessibility), and relationship factors (use of a stepwise approach, deciding to approach a cancer, and mental health discussion later), similar to factors described in international literature [28]. Australian cancer clinical guidelines recommend immunochemical fecal occult blood test as a screening tool for asymptomatic individuals in the general population and also in some symptomatic patients [29,30]. Recent research indicates its effectiveness as a rule-out test for colorectal cancer in patients presenting with gastrointestinal and nonspecific symptoms [31]. There was variable awareness of these recommendations by the GPs, with 6 out of 10 GPs recommending it alongside other imaging for their high-risk patients. CDSSs may be used as facilitators to behavior change, encouraging GPs to adopt new approaches in their diagnostic process. The majority of GPs did not access the tool's embedded resources, which offered additional context and explanation of the recommendations. This could be explained by the additional cognitive load required to navigate and access less-readily available resources. However, it may also indicate a strong trust in the tool's core content among participating physicians as confirmed in the interviews. Such trust is noteworthy, as overcoming skepticism toward CDSSs has been identified in literature as a major barrier to their adoption [11].

GPs utilized the tool in diverse ways: incorporating it into their own decision-making process without sharing it with the patient, as a reminder of differential diagnosis and suggested follow-up and to aid communication. Recent evidence showed the potential of a CDSS to improve patient–provider communication in cancer patients [32]. Notably, our CDSS showed the potential to either enhance or impede patient–clinician communication, depending on its application, how GPs integrated it into the consultation, and the specific context of use.

Health care delivery has increasingly become "digitalized". The use of technology has become embedded in health care systems in most high-income countries, and it continues to expand worldwide. Technology can help clinicians to be more efficient and improve patient safety [33].

Multiple guidelines for digital intervention development have been published by international agencies and government bodies, such as World Health Organization [34], National Institute for Health and Care Excellence [35], the Food and Drug Administration [36], and in the academic literature. In 2020, Mathews et al [37] published the digital health score card, where authors propose that digital health solutions should always be validated from a technical, clinical, and system standpoint [37]. Even with these guidelines, there is not one single framework or process required or agreed upon to test digital interventions in similar real-life scenarios before releasing them to the public [38].

Simulation techniques have emerged as a promising approach to bridge the gap between the development and evaluation of digital health solutions before their implementation [38]. Despite growing interest and well-established methods, simulations for digital interventions prior to real-world deployment are not the norm, and this may impact not only their safety but their potential for implementation, resulting in most tools not being adopted into the routine health care system [39].

Barriers to incorporating new software into health care systems can include concerns about potential impacts on patient care quality, apprehensions regarding data privacy and protection, organizational capacity to manage nonmedical tasks, and the ever-evolving landscape of health care delivery [16].

Our simulation provided a controlled environment to test a CDSS designed to identify individuals with symptoms potentially indicative of undiagnosed cancer. By simulating diverse and realistic clinical scenarios, it allowed the system to be rigorously evaluated without risk to actual patients. It also allowed evaluation to occur in a timely manner, rather than waiting for a real-world "needle in a haystack" identification of low prevalence conditions and symptoms. This approach also offered valuable insights into the barriers and facilitators of implementing a tool for early cancer detection [32].

Notably, the simulation served two key functions. First, it directly modeled and observed practical issues, such as usability and workflow integration. Second, it acted as a "priming" tool, enabling practitioners to provide detailed, informed feedback on the CDSS functionality and recommendations. Most importantly, it highlighted that adequate use and implementation need to consider the "human factor."

XSL•FO
RenderX

Strengths and Limitations

A key strength of this study lies in its multimodal analysis approach, combining video recordings of simulations with in-depth interviews with both participants in the consultation and real-live observing CAs. This methodology provided a comprehensive and nuanced understanding of the subject matter, capturing both observable behaviors and subjective experiences. Furthermore, the inclusion of both clinical and lived-experience perspectives ensured a well-rounded analysis, balancing professional insights with authentic patient views. This diverse approach not only enhanced the study's rigor but also allowed for a comprehensive evaluation that bridges the gap between theoretical design and practical application in health care settings. The additional information gleaned from patients and CAs on how this type of technology and system support can enhance the patient experience in the inclusion of their health management adds an additional perspective to prior research focusing mainly on systems and technological evaluation [37].

All participating GPs were from metropolitan areas. This limitation was primarily due to the logistical requirement for GPs to travel to the simulation laboratory. Additionally, while simulations provide valuable insights, they inherently lack the full complexity and unpredictability of real-world clinical encounters and the pressures of a busy clinical day. This artificial setting may not fully capture the nuances of actual patient interactions, time pressures, or diverse clinical scenarios that GPs encounter in their daily practice, and GPs might act differently than in their regular day-to-day practice when knowing they are being observed.

While most GPs were not previously known to the researchers, and we did not provide information regarding personal goals or reasons for doing the research, two GPs had participated in previous studies and had established working relationships with one or both of the researchers. We acknowledge this may lead to social desirability bias, which we tried to decrease by reassuring them of our intention to assess real implementation, asking questions that could elicit both positive and negative responses toward the tool.

Conclusions

We developed and evaluated a CDSS designed to identify patients with UWL who may be at risk of undiagnosed cancer. The controlled simulation environment provided a safe setting to assess implementation barriers and facilitators while observing the tool in use. Clinicians found the recommendations to be clear and unobtrusive; however, concerns were raised about potential overtesting in patients with a low risk of cancer. The tool was utilized by clinicians in 3 primary ways: to inform their clinical decision-making process, as a reminder or checklist of best practices, and as an aid in communicating potential follow-up options with patients. Both patients and GPs believed the tool could either enhance or impede communication, depending on its integration into consultations, underscoring the potential need for communication-focused training in the use of these technologies. This study's findings will inform future studies implementing CDSSs for identifying patients with nonspecific symptoms who may be at risk of cancer, potentially improving early detection and patient outcomes.

Acknowledgments

We acknowledge the following people and teams: our study participants for their invaluable contributions and insightful engagement, HaBIC team at the University of Melbourne for the development of the cancer module, and Validitron Team and Centre for Digital Transformation of Health for allowing us to use their space and their help in setting up the simulations. Generative artificial intelligence was not used in the study design, analysis, or development of the initial draft of the manuscript. The University of Melbourne's Spark AI was used for copyediting of the revised manuscript.

Funding

JMG is funded by a Victoria Cancer Agency (VCA) midcareer fellowship (MCRF 21025). VCA played no role in the study design, data collection, analysis, and interpretation of data. JE is funded by the National Health and Medical Research Council (NHMRC) Investigator grant (APP1195302). NHMRC played no role in the study design, data collection, analysis, and interpretation of data.

Data Availability

Unidentified transcriptions and qualitative coded data can be made available upon request secondary to ethics approval. Qualitative codes were developed in an iterative process by the two researchers. They were uploaded to NVivo version 14. They can be shared upon request.

Authors' Contributions

JMG contributed to the conceptualization and design of the work; the acquisition, analysis, and interpretation of data; and drafted and reviewed the manuscript.

SC contributed to the design of the work; the acquisition, analysis, and interpretation of data; and drafted and reviewed the manuscript.

LDM contributed to the design of simulated patient scenarios and to the validation and user acceptance testing of the clinical decision support system (CDSS) and reviewed the manuscript.

XSL•FO
RenderX

AL contributed to the design of the work, provided and analyzed the data for symptoms selection, and reviewed the manuscript.
BH and JAMN contributed to the design of the work, are chief investigators and managers of the CDSS used in this study, provided supervision and support on the creation of the CDSS, and reviewed the manuscript.
DD and GF are consumer investigators in this study; they contributed to the conceptualization of the work, reviewed module and simulation documentation, and reviewed and approved the manuscript.
KH contributed to the setting up and facilitation of the simulated environment, provided academic and technical support during simulations, and reviewed the manuscript.
FSL, BW, and CN are investigators in this study; they contributed to conceptualization of the study and reviewed the manuscript.
BN is the author of the risk prediction model that informed the unintended weight loss module, provided supervision and support through replication of relevant module algorithms, and reviewed the manuscript.
JE is the senior investigator of this study; he contributed to the conceptualizing and design of the work, supervision of the work, and edited and reviewed the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Patient scenarios.
[DOCX File, 22 KB - formative_v9i1e79208_app1.docx]

Multimedia Appendix 2
General practitioner script.
[DOCX File, 17 KB - formative_v9i1e79208_app2.docx]

Multimedia Appendix 3
Actors' script.
[DOCX File, 15 KB - formative_v9i1e79208_app3.docx]

Multimedia Appendix 4
Community advocate script.
[DOCX File, 16 KB - formative_v9i1e79208_app4.docx]

Checklist 1
COREQ checklist.
[PDF File, 226 KB - formative_v9i1e79208_app5.pdf]

References

1. Lacey K, Bishop JF, Cross HL, Chondros P, Lyratzopoulos G, Emery JD. Presentations to general practice before a cancer diagnosis in Victoria: a cross-sectional survey. Med J Aust 2016 Jul 18;205(2):66-71. [doi: 10.5694/mja15.01169] [Medline: 27456447]
2. Black GB, Nicholson BD, Moreland JA, Fulop NJ, Lyratzopoulos G, Baxter R. Doing "detective work" to find a cancer: how are non-specific symptom pathways for cancer investigation organised, and what are the implications for safety and quality of care? A multisite qualitative approach. BMJ Qual Saf 2025 Jun 19;34(7):457-468. [doi: 10.1136/bmjqs-2024-017749] [Medline: 39880649]
3. Koo MM, Unger-Saldaña K, Mwaka AD, et al. Conceptual framework to guide early diagnosis programs for symptomatic cancer as part of global cancer control. JCO Glob Oncol 2021 Jan;7(7):35-45. [doi: 10.1200/GO.20.00310] [Medline: 33405957]
4. Nicholson BD, Hamilton W, O'Sullivan J, Aveyard P, Hobbs FR. Weight loss as a predictor of cancer in primary care: a systematic review and meta-analysis. Br J Gen Pract 2018 May;68(670):e311-e322. [doi: 10.3399/bjgp18X695801] [Medline: 29632004]
5. Lee A, de Mendonça L, McCarthy D, et al. Primary care patients presenting with unexpected weight loss in Australian general practices: replication of a diagnostic accuracy study. BMJ Open 2025 Jul 28;15(7):e104690. [doi: 10.1136/bmjopen-2025-104690] [Medline: 40730405]
6. Nicholson BD, Thompson MJ, Hobbs FDR, et al. Measured weight loss as a precursor to cancer diagnosis: retrospective cohort analysis of 43 302 primary care patients. J Cachexia Sarcopenia Muscle 2022 Oct;13(5):2492-2503. [doi: 10.1002/jcsm.13051] [Medline: 35903866]

XSL•FO
RenderX

7. Nicholson BD, Virdee P, Aveyard P, et al. Prioritising primary care patients with unexpected weight loss for cancer investigation: diagnostic accuracy study (update). BMJ 2024 Oct 16;387:e080199. [doi: 10.1136/bmj-2024-080199] [Medline: 39414353]
8. Suspected cancer: recognition and referral NICE guideline. National Institute for Health and Care Excellence. 2015. URL: https://www.nice.org.uk/guidance/ng12 [accessed 2025-11-24]
9. Rao G, Ufholz K, Saroufim P, Menegay H, Beno M. Recognition, diagnostic practices, and cancer outcomes among patients with unintentional weight loss (UWL) in primary care. Diagnosis (Berl) 2023 Aug 1;10(3):267-274. [doi: 10.1515/dx-2023-0002] [Medline: 37080911]
10. Martinez-Gutierrez J, De Mendonca L, Ly P, et al. A scoping review of unexpected weight loss and cancer: risk, guidelines, and recommendations for follow-up in primary care. BJGP Open 2024 Dec;8(4):BJGPO.2024.0025. [doi: 10.3399/BJGPO.2024.0025] [Medline: 39054298]
11. Chima S, Reece JC, Milley K, Milton S, McIntosh JG, Emery JD. Decision support tools to improve cancer diagnostic decision making in primary care: a systematic review. Br J Gen Pract 2019 Dec;69(689):e809-e818. [doi: 10.3399/bjgp19X706745] [Medline: 31740460]
12. Shortliffe EH, Cimino JJ. Biomedical Informatics: Computer Applications in Health Care and Biomedicine, 4th edition: Springer London; 2014:1-965. [doi: 10.1007/978-1-4471-4474-8]
13. Hunter B, Davidson S, Lumsden N, et al. Optimising a clinical decision support tool to improve chronic kidney disease management in general practice. BMC Prim Care 2024 Jun 19;25(1):220. [doi: 10.1186/s12875-024-02470-w] [Medline: 38898462]
14. Price S, Spencer A, Medina-Lara A, Hamilton W. Availability and use of cancer decision-support tools: a cross-sectional survey of UK primary care. Br J Gen Pract 2019 Jul;69(684):e437-e443. [doi: 10.3399/bjgp19X703745] [Medline: 31064743]
15. Kwan JL, Lo L, Ferguson J, et al. Computerised clinical decision support systems and absolute improvements in care: meta-analysis of controlled clinical trials. BMJ 2020 Sep 17;370:m3216. [doi: 10.1136/bmj.m3216] [Medline: 32943437]
16. Huckvale K, Smolenaers F, Ferdous H, et al. A hybrid physical-digital simulation laboratory to expedite context-aware design and usability testing in digital health. Stud Health Technol Inform 2024 Jan 25;310:1513-1514. [doi: 10.3233/SHTI231270] [Medline: 38269722]
17. Hunter B, Alexander K, Biezen R, et al. The development of Future Health Today: piloting a new platform for identification and management of chronic disease in general practice. Aust J Prim Health 2023 Feb;29(1):8-15. [doi: 10.1071/PY22022] [Medline: 36318973]
18. Hunter B, Biezen R, Alexander K, et al. Future Health Today: codesign of an electronic chronic disease quality improvement tool for use in general practice using a service design approach. BMJ Open 2020 Dec 18;10(12):e040228. [doi: 10.1136/bmjopen-2020-040228] [Medline: 33371024]
19. The digital health validitron. Centre for Digital Transformation of Health. URL: https://mdhs.unimelb.edu.au/digitalhealth/our-work/validitron [accessed 2023-07-30]
20. Practice incentives program quality improvement measures: annual data update 2023–24: introduction. Australian Institute of Health and Welfare. URL: https://www.aihw.gov.au/reports/primary-health-care/pipqi-measures-2023-24/contents/introduction [accessed 2025-07-08]
21. Healthcare simulation dictionary. Agency for Healthcare Research and Quality. URL: https://www.ahrq.gov/patient-safety/resources/simulation/terms.html [accessed 2025-10-16]
22. Sekhon M, Cartwright M, Francis JJ. Acceptability of healthcare interventions: an overview of reviews and development of a theoretical framework. BMC Health Serv Res 2017 Jan 26;17(1):88. [doi: 10.1186/s12913-017-2031-8] [Medline: 28126032]
23. Singh H, Sittig DF. A sociotechnical framework for safety-related electronic health record research reporting: the SAFER reporting framework. Ann Intern Med 2020 Jun 2;172(11 Suppl):S92-S100. [doi: 10.7326/M19-0879] [Medline: 32479184]
24. Braun V, Clarke V. To saturate or not to saturate? Questioning data saturation as a useful concept for thematic analysis and sample-size rationales. Qual Res Sport Exerc Health 2021 Mar 4;13(2):201-216. [doi: 10.1080/2159676X.2019.1704846]
25. Braun V, Clarke V. Thematic analysis. In: APA Handbook of Research Methods in Psychology, Vol 2 Research Designs: Quantitative, Qualitative, Neuropsychological, and Biological: American Psychological Association; 2012:57-71. [doi: 10.1037/13620-004]
26. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
27. Kidney E, Greenfield S, Berkman L, et al. Cancer suspicion in general practice, urgent referral, and time to diagnosis: a population-based GP survey nested within a feasibility study using information technology to flag-up patients with symptoms of colorectal cancer. BJGP Open 2017 Oct 4;1(3):bjgpopen17X101109. [doi: 10.3399/bjgpopen17X101109] [Medline: 30564682]
28. Hughes LD. Understanding the processes behind the decisions–GPs and complex multimorbidity decision making. BMC Prim Care 2022 Jun 28;23(1):162. [doi: 10.1186/s12875-022-01781-0] [Medline: 35761167]

XSL•FO
RenderX

29. Guidelines for preventive activities in general practice. Royal Australian College of General Practice. 2024. URL: https://tinyurl.com/3p4x74p2 [accessed 2025-06-12]
30. Colonoscopy categorisation guidelines. : Victorian Department of Health; 2017 URL: https://www2.health.vic.gov.au/Api/downloadmedia/%7B8C3C68DC-7332-416B-93FA-33D2E941C77A%7D [accessed 2025-11-24]
31. Pin-Vieito N, Tejido-Sandoval C, de Vicente-Bielza N, Sánchez-Gómez C, Cubiella J. Faecal immunochemical tests safely enhance rational use of resources during the assessment of suspected symptomatic colorectal cancer in primary care: systematic review and meta-analysis. Gut 2022 May;71(5):950-960. [doi: 10.1136/gutjnl-2021-324856] [Medline: 34108236]
32. Hong YA, Hossain MM, Chou WYS. Digital interventions to facilitate patient-provider communication in cancer care: a systematic review. Psychooncology 2020 Apr;29(4):591-603. [doi: 10.1002/pon.5310] [Medline: 31834650]
33. Sly B, Russell AW, Sullivan C. Digital interventions to improve safety and quality of inpatient diabetes management: a systematic review. Int J Med Inform 2022 Jan;157:104596. [doi: 10.1016/j.ijmedinf.2021.104596] [Medline: 34785487]
34. Monitoring and Evaluating Digital Health Interventions: A Practical Guide to Conducting Research and Assessment: World Health Organization; 2016. URL: https://www.who.int/publications/i/item/9789241511766 [accessed 2024-11-20]
35. Evidence standards framework (ESF) for digital health technologies. National Institute for Health and Care Excellence. URL: https://www.nice.org.uk/about/what-we-do/our-programmes/evidence-standards-framework-for-digital-health-technologies [accessed 2024-11-20]
36. Factors to consider regarding benefit-risk in medical device product availability, compliance, and enforcement decisions guidance for industry and Food and Drug Administration staff preface public comment. Food and Drug Administration. 2016. URL: https://www.fda.gov/medical-devices/digital-health/digital-health-software-precertification-pre-cert-program [accessed 2024-11-20]
37. Mathews SC, McShea MJ, Hanley CL, Ravitz A, Labrique AB, Cohen AB. Digital health: a path to validation. NPJ Digit Med 2019;2:38. [doi: 10.1038/s41746-019-0111-3] [Medline: 31304384]
38. Guo C, Ashrafian H, Ghafur S, Fontana G, Gardner C, Prime M. Challenges for the evaluation of digital health solutions—a call for innovative evidence generation approaches. NPJ Digit Med 2020;3(1):110. [doi: 10.1038/s41746-020-00314-2] [Medline: 32904379]
39. Sittig DF, Wright A, Osheroff JA, et al. Grand challenges in clinical decision support. J Biomed Inform 2008 Apr;41(2):387-392. [doi: 10.1016/j.jbi.2007.09.003] [Medline: 18029232]

Abbreviations

CA: community advocate
CDSS: clinical decision support system
COREQ: Consolidated Criteria for Reporting Qualitative Research
EMR: electronic medical record
FHT: Future Health Today
GP: general practitioner
SP: simulated patient
UWL: unintended weight loss

Edited by A Mavragani; submitted 30.07.25; peer-reviewed by D Cunha; revised version received 23.10.25; accepted 03.11.25; published 10.12.25.

Please cite as:
Martinez-Gutierrez J, Chima S, De Mendonca L, Lee A, Hunter B, Manski-Nankervis JA, Daly D, Fishman G, Huckvale K, Lim FS, Wang B, Nelson C, Nicholson B, Emery J
Evaluating a Clinical Decision Support Tool for Cancer Risk Assessment in Primary Care: Simulation Study of Unintended Weight Loss
JMIR Form Res 2025;9:e79208
URL: https://formative.jmir.org/2025/1/e79208
doi:10.2196/79208

© Javiera Martinez-Gutierrez, Sophie Chima, Lucas De Mendonca, Alex Lee, Barbara Hunter, Jo-Anne Manski-Nankervis, Deborah Daly, George Fishman, Kit Huckvale, Fong Seng Lim, Benny Wang, Craig Nelson, Brian Nicholson, Jon Emery. Originally published in JMIR Formative Research (https://formative.jmir.org), 10.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative

XSL•FO
RenderX

Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Radiomics-Based AI Model to Assist Clinicians in Intracranial Hemorrhage Diagnosis: External Validation Study

Salita Angkurawaranon[1,2], MD; Natipat Jitmahawong[1], MD; Kittisak Unsrisong[1], MD; Phattanun Thabarsa[3], BS; Chakri Madla[1], MD; Withawat Vuthiwong[1], MD; Thanwa Sudsang[4], MD; Chaisiri Angkurawaranon[2,5], MD, PhD; Patrinee Traisathit[6,7], PhD; Papangkorn Inkeaw[2,7,8], PhD

[1]Department of Radiology, Faculty of Medicine, Chiang Mai University, Chiang Mai, Thailand
[2]Global Health and Chronic Conditions Research Group, Chiang Mai University, Chiang Mai, Thailand
[3]Master's Degree Program in Data Science, Faculty of Engineering, Chiang Mai University, Chiang Mai, Thailand
[4]Department of Radiology, Ramathibodi Hospital, Mahidol University, Bangkok, Thailand
[5]Department of Family Medicine, Faculty of Medicine, Chiang Mai University, Chiang Mai, Thailand
[6]Department of Statistics, Faculty of Science, Chiang Mai University, Chiang Mai, Thailand
[7]Data Science Research Center, Faculty of Science, Chiang Mai University, 239 Huay Kaew Road, Muang District, Chiang Mai, Thailand
[8]Department of Computer Science, Faculty of Science, Chiang Mai University, Chiang Mai, Thailand

Corresponding Author:
Papangkorn Inkeaw, PhD
Global Health and Chronic Conditions Research Group, Chiang Mai University, Chiang Mai, Thailand

Abstract

Background: Early identification of the etiology of spontaneous intracerebral hemorrhage (ICH) could significantly contribute to planning a suitable treatment strategy. A notable radiomics-based artificial intelligence (AI) model for classifying causes of spontaneous ICH from brain computed tomography scans has been previously proposed.

Objective: This study aimed to externally validate and assess the utility of this AI model.

Methods: This study used 69 computed tomography scans from a separate cohort to evaluate the AI model's performance in classifying nontraumatic ICHs into primary, tumorous, and vascular malformation related. We also assessed the accuracy, sensitivity, specificity, and positive predictive value of clinicians, radiologists, and trainees in identifying the ICH causes before and after using the model's assistance. The performances were statistically analyzed by specialty and expertise levels.

Results: The AI model achieved an overall accuracy of 0.65 in classifying the 3 causes of ICH. The model's assistance improved overall diagnostic performance, narrowing the gap between nonradiology and radiology groups, as well as between trainees and experts. The accuracy increased from 0.68 to 0.72, from 0.72 to 0.76, from 0.69 to 0.74, and from 0.72 to 0.75 for nonradiologists, radiologists, trainees, and specialists, respectively. With the model's support, radiology professionals demonstrated the highest accuracy, highlighting the model's potential to enhance diagnostic consistency across different levels.

Conclusions: When applied to an external dataset, the accuracy of the AI model in categorizing spontaneous ICHs based on radiomics decreased. However, using the model as an assistant substantially improved the performance of all reader groups, including trainees and radiology and nonradiology specialists.

(JMIR Form Res 2025;9:e81038) doi:10.2196/81038

KEYWORDS

radiomics; spontaneous intracranial hemorrhages; external validation; artificial intelligence; AI; AI-assisted decision-making; brain computed tomography scans; brain CT scans

Introduction

Spontaneous intracerebral hemorrhage (ICH) accounts for 10% to 15% of all strokes and is associated with a higher mortality rate than either ischemic stroke or subarachnoid hemorrhage (SAH). The causes of ICH are divided into primary and secondary. Primary causes of nontraumatic ICH associated with chronic hypertension or amyloid angiopathy and secondary causes associated with vascular abnormalities or neoplasms are among the most prevalent [1]. The principal treatments for primary ICH are surgical hematoma excision and conservative therapy [2]. Secondary ICH requires further attention to address the underlying disease. Subsequently, the early identification of the etiology of spontaneous ICH can significantly contribute to the optimization of the prognosis of individual patients and the planning of a suitable treatment strategy [3].

XSL•FO
RenderX

Computed tomography (CT) is typically the initial imaging modality for diagnosing and determining the cause of ICH. However, non–contrast-enhanced CT scans often fail to reliably identify the underlying etiology of spontaneous ICH [3,4]. A hemorrhage associated with secondary causes of ICH, especially tumors and vascular malformations, can be masked and lead to a misdiagnosis [5-7]. Misclassification of the causes of ICH can lead to delayed diagnoses and worsen patient outcomes.

Radiomics, a rapidly advancing field, uses advanced image analysis to extract quantitative features that extend beyond what is visible to the human eye [8]. By applying mathematical approaches, radiomics enables the quantification of visual attributes, offering deeper insights into imaging data. There has been growing interest in deploying machine learning in clinical practice in recent years. Rather than replacing radiologists and clinicians, machine learning models are increasingly recognized as tools that support and enhance decision-making processes [9-11]. One area that remains underexplored is the potential of these models to improve radiologists' and clinicians' diagnostic performance through collaborative decision-making between humans and artificial intelligence (AI) models. This potential is particularly significant given the global shortage of radiologists and the increasing volume of imaging studies [12]. Previous studies have combined radiomics with machine learning models to differentiate between various causes of nontraumatic ICH, including vascular malformation–related hemorrhage versus primary ICH [13], arteriovenous malformation–related hematomas versus other spontaneous ICHs [14], primary versus secondary ICH [15], and neoplastic versus nonneoplastic ICH [16].

A notable study by Thabarsa et al [17] used radiomics and machine learning to classify the 3 most common causes of nontraumatic ICH—primary, tumorous, and vascular malformation–related hemorrhages—from brain CT scans. This multiclass approach aligns more closely with real-world clinical practice than previous studies that classified just 2 causes of spontaneous ICHs [13-16]. The aforementioned study also differs from other radiomics studies that exclusively concentrate on the hematoma region or the edema region [13,14]. Thabarsa et al [17] applied radiomics features from both regions, with the results demonstrating that, when features are chosen from both regions, this improves the classification performance of the model [16]. Although the sensitivity and area under the curve were marginally lower than in previous research, the specificity was higher, indicating a satisfactory performance [13,15]. However, a critical limitation of prior studies is lack of external validation, which hampers the generalizability of radiomics-based prediction models and their integration into clinical workflows [13-17].

The aim of this study was to externally evaluate radiomics-based models to classify causes of nontraumatic ICH from CT scans of the brain on an independent cohort. Furthermore, this study evaluated the performance of trainee clinicians and trainee radiologists and of clinicians and radiologists with and without the support of a machine learning model.

Methods

Overview

This study used a previously published machine learning model to classify the etiologies of nontraumatic ICH into primary, tumorous, and vascular malformation related [17]. The model was first evaluated using an external validation dataset. It was then assessed for its potential to enhance clinicians' diagnostic performance. An overview of the study is shown in Figure 1 [17].

XSL•FO
RenderX

Figure 1. The workflow of this study [17]. CT: computed tomography; ICH: intracerebral hemorrhage; IVH: intraventricular hemorrhage; SAH: subarachnoid hemorrhage.

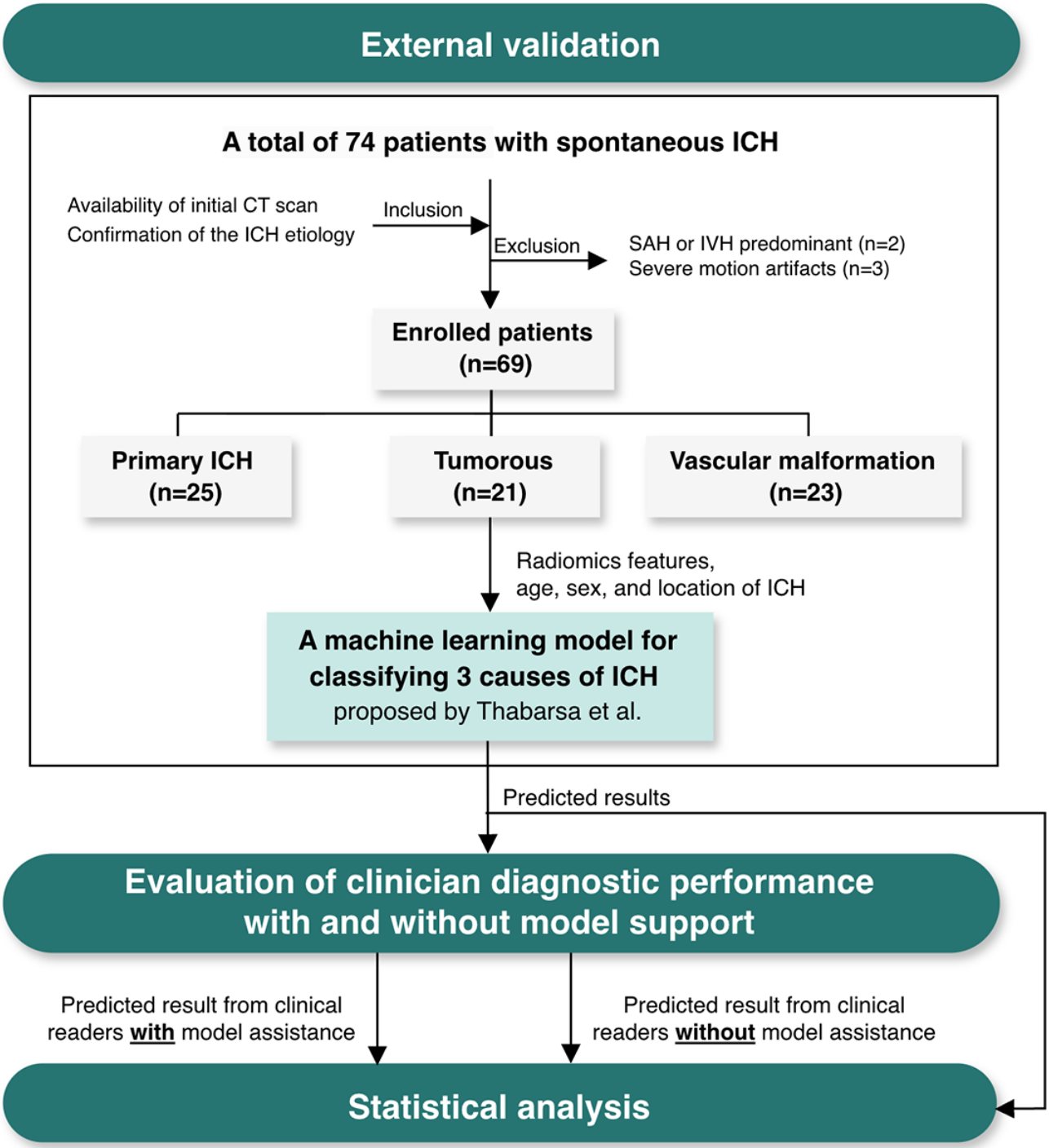

Ethical Considerations

The Ethics Committee of the Faculty of Medicine Ramathibodi Hospital, Mahidol University, Thailand, granted permission for the acquisition of the external validation dataset from Ramathibodi Hospital (approval COA.MURA2022/431; approval date: August 3, 2022). As the study was retrospective in nature and all procedures were part of standard clinical care, the committee waived the informed consent requirement. To maintain privacy and confidentiality, all data were completely anonymized before analysis. This study did not involve participant compensation.

External Validation Data

The dataset used in this study was obtained from Ramathibodi Hospital. We conducted a retrospective evaluation of the medical reports of patients who experienced spontaneous ICH from the radiology department database between January 2010 and January 2020. Data collected included noncontrast CT (NCCT) brain scans, as well as patient age, sex, and ICH location. Inclusion criteria were as follows: (1) availability of initial CT scan after onset of symptoms and (2) confirmation of the ICH etiology through follow-up magnetic resonance imaging, pathological reports, or additional digital subtraction angiography for primary, tumorous, and vascular

XSL•FO
RenderX

malformation–related ICH. Exclusion criteria included hematomas with SAH or intraventricular hemorrhage predominant. Low-quality NCCT images with severe artifacts, such as metallic or motion artifacts, were also excluded.

Of the 74 NCCT brain scans reviewed, 69 (93.2%) were included in the study. The remaining 6.8% (5/74) of the scans were excluded due to severe artifacts (n=3) or the presence of hematomas with SAH or intraventricular hemorrhage predominant (n=2). The scans were first anonymized and set to a brain window (level: 80 HU; width: 40 HU), slice thickness of 2.5 to 5 mm, and image resolution of 512 × 512 pixels. The CT scans were performed using 1 of the 4 institute machines: a GE Revolution Frontier spectral CT scanner, a GE Revolution spectral CT scanner, a Philips IQon spectral CT scanner, and a Philips Brilliance iCT scanner. Data on age, sex, and ICH location were also obtained. The ICH location was classified into deep, lobar, and infratentorial regions. The study by Thabarsa et al [17] provides a detailed explanation of these classifications, as well as the process of feature extraction for hemorrhages and the vasogenic edema that surrounds them.

Machine Learning Model for Classifying Causes of ICH

In this study, we adopted a radiomics-based machine learning model proposed by Thabarsa et al [17] for classifying causes of ICHs. The model was organized in a hierarchical structure that consists of 2 adaptive boosting classifiers. First, it discriminates between tumor and nontumor causes of ICH through the first classifier. If the case is classified as nontumor, a second classifier is applied to further differentiate between primary and vascular malformation–related ICH. The first classifier uses radiomics features extracted from both the hematoma and edema regions, including 2 first-order, 2 gray-level co-occurrence matrix (GLCM), and 3 gray-level dependence matrix (GLDM) features from the hematoma region and 1 GLDM, 1 gray-level run-length matrix, 2 gray-level size-zone matrix (GLSZM), and 3 GLCM features from the edema region. Meanwhile, the second classifier uses 10 radiomics features extracted from the hematoma (ie, 2 GLDM, 2 GLCM, and 1 GLSZM) and edema (ie, 2 GLDM, 1 GLCM, 1 GLSZM, and 1 gray-level run-length matrix) regions as predictors. Furthermore, age and ICH location were incorporated as supplementary predictors for classifying the causes of ICH in both classifiers.

The model was trained and tested on 141 NCCT scans using 5-fold cross-validation. All scans were collected from Maharaj Nakorn Chiang Mai Hospital. Specifically, of the 141 scans, 26 (18.4%) were obtained from a Toshiba Aquilion 16-slice CT scanner, 89 (63.1%) were obtained from a 64-slice Siemens Definition multidetector CT (Germany) scanner, and 26 (18.4%) were obtained from a 192-slice Siemens Definition dual-source CT (Germany) scanner. The scans were acquired using a brain window (level: 80 HU; width: 40 HU), with a slice thickness of 5 mm and an image resolution of 512 × 512 pixels.

Clinicians' Performance With AI Assistance

To assess the potential of the proposed machine learning model as a diagnostic assistance tool in distinguishing the 3 causes of ICHs, clinical readers from diverse specialties and varying levels of expertise participated. They included 4 general radiologists, 8 radiology residents, 2 emergency medicine physicians, 2 emergency medicine residents, 2 neurosurgeons, and 2 neurosurgery residents. Their performance was compared with and without model assistance and analyzed by specialty and level of expertise.

The readers were divided into 2 groups (group A and group B). In the first session, group A interpreted the CT scans with model assistance, whereas group B interpreted them without model assistance. Following the session, a 4-week washout period took place, during which the groups were reshuffled to minimize familiarization before the next session. In the second session, group A interpreted the CT scans without model assistance, whereas group B used model assistance. Patient age and sex information was also made available to all readers. In both sessions, each reader was assigned a unique, nonrepeating order of patients (Figure 2).

XSL•FO
RenderX

Figure 2. Study design for evaluation of clinician diagnostic performance with and without model support.

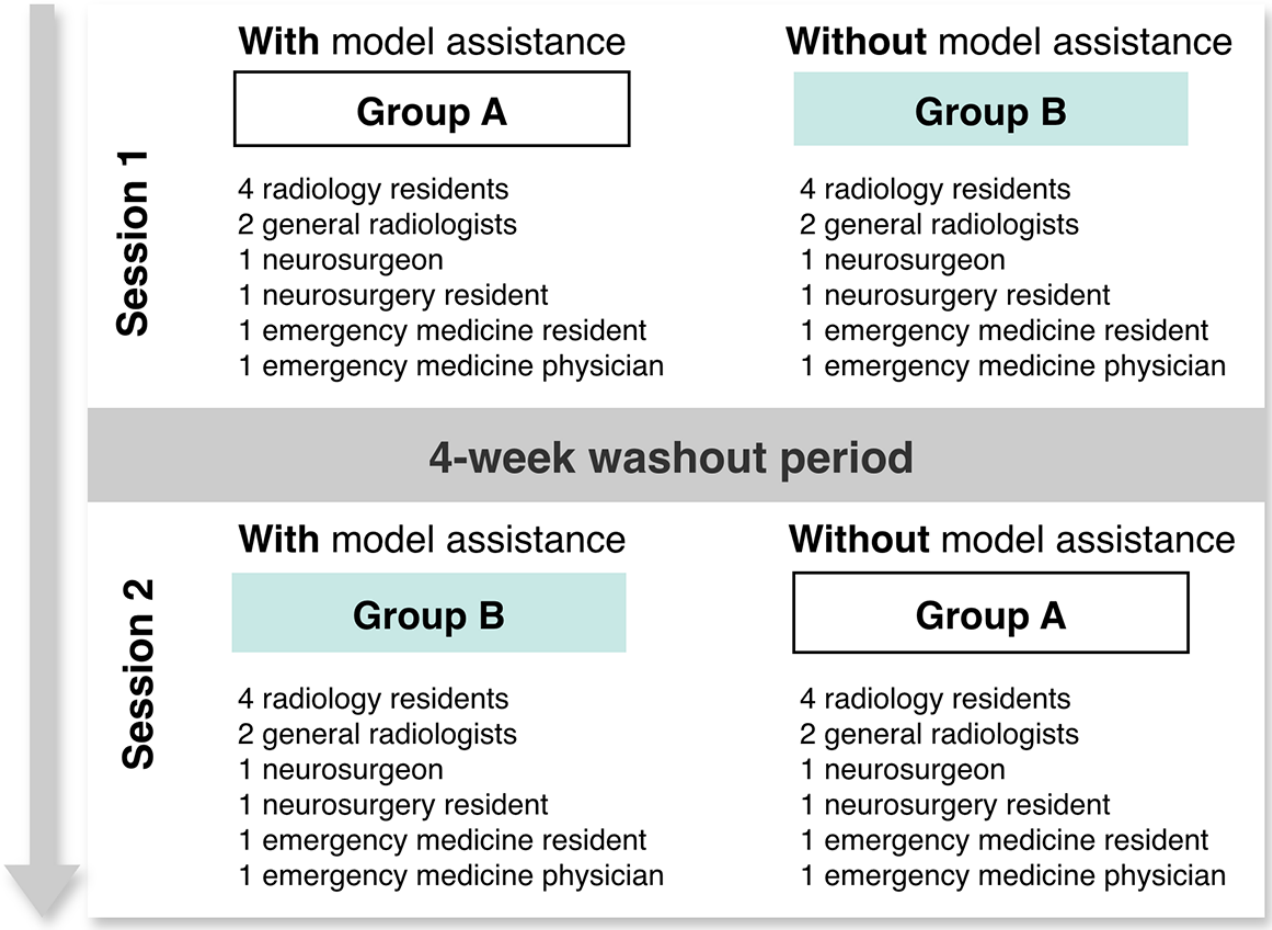

Statistical Analysis

Demographic and clinical variables between external validation datasets were compared using the Kruskal-Wallis test for continuous variables and the Fisher exact test for categorical variables. To investigate the effects of CT scan manufacturers and slice thickness on predictive error, a chi-square test was also used. In addition, we investigated the differences in the diagnostic performances of nonradiologists and radiologists before and after adopting the machine learning model's assistance. The nonradiology readers included trainees and clinicians in emergency medicine and neurosurgery, whereas the radiology readers consisted of radiology residents and radiologists. In addition, we explored improvements in the diagnostic ability to identify ICH causes in groups of trainees and specialists after being assisted by the machine learning model. Residents from the radiology, neurosurgery, and emergency medicine departments were among the trainee readers. Radiologists, neurosurgeons, and emergency medicine physicians were grouped as specialty readers. Significant differences measured using the Friedman test were considered at $P<.05$. The diagnostic performance was evaluated using accuracy, sensitivity, specificity, and positive predictive value (PPV), with the analysis conducted using R (version 4.3.2; R Foundation for Statistical Computing).

Results

Study Population

On the basis of the participant flowchart (Figure 1), 69 patients met the eligibility criteria and were enrolled in the study. A summary of the patient demographics is provided in Table 1. Age and ICH location showed no significant differences in the underlying causes of ICH ($P=.81$ and $P=.41$, respectively). Sex, which was not considered in the diagnostic process of the model, also showed no significant differences across the classes ($P=.16$).

XSL•FO
RenderX

Table . Patient demographic and clinical characteristics of the external validation dataset (N=69).

Variable	Patients with spontaneous intracerebral hemorrhage			*P* value
	Primary (n=25)	Tumorous (n=21)	Vascular malformation related (n=23)	
Age (y), mean (SD)	51.5 (21.8)	55.8 (18.1)	43.2 (23.0)	.81
Sex, n (%)				.16
Male	14 (56)	13 (61.9)	12 (52.2)	
Female	11 (44)	8 (38.1)	11 (47.8)	
Location, n (%)				.41
Deep	6 (24)	2 (9.5)	3 (13)	
Infratentorial	1 (4)	0 (0)	0 (0)	
Lobar	18 (72)	19 (90.5)	20 (87)	

Model Performance

The machine learning model achieved an overall accuracy of 0.65 in the classification of the 3 causes of ICH. The model performance metrics for each cause of ICH on the external dataset are shown in Table 2. For primary ICH, the sensitivity, specificity, and PPV were 0.64, 0.82, and 0.67, respectively. For tumor-related ICH, the model achieved sensitivity, specificity, and PPV of 0.81, 0.79, and 0.63. The sensitivity, specificity, and PPV for vascular malformation–related ICH were 0.52, 0.87, and 0.66, respectively.

Table . Classification performance of the machine learning model on the external validation dataset.

Cause of ICH[a]	Performance metric			
	Sensitivity	Specificity	PPV[b]	AUC[c]
Primary	0.64	0.82	0.67	0.59
Tumorous	0.81	0.79	0.63	0.84
Vascular malformation	0.52	0.87	0.67	0.65

[a]ICH: intracerebral hemorrhage.

[b]PPV: positive predictive value.

[c]AUC: area under the curve.

We further analyzed the effects of CT scan manufacturers and slice thickness on predictive error using a chi-square test. The results showed that, for primary ICH, no statistically significant association was observed between slice thickness and prediction accuracy (P=.11). Similarly, the manufacturer was not significantly associated with prediction accuracy (P=.77). For tumor-related ICH, slice thickness showed no statistically significant relationship with prediction accuracy (P=.50), and the manufacturer also showed no significant association (P=.66). For vascular malformation–related ICH, there was no statistically significant association between slice thickness and prediction accuracy (P=.41). Similarly, the manufacturer was not significantly associated with prediction accuracy (P=.74).

Evaluation of Clinician Diagnostic Performance With and Without AI Support

Diagnostic Performance of Readers With and Without AI Support by Clinical Specialty

Table 3 presents the diagnostic performance of readers in identifying causes of acute ICH both with and without the assistance of AI. The readers were categorized by clinical specialty as nonradiology and radiology readers. As can be observed from the results, radiology readers consistently outperformed nonradiology readers in both situations. The model was able to assist both nonradiology and radiology readers in identifying the cause of ICH by substantially improving the overall accuracies from 0.68 to 0.72 and from 0.72 to 0.76, respectively. The sensitivity, specificity, and PPV for primary, tumorous, and vascular malformation–related ICH increased in both reader groups after being assisted by the model. The sensitivity, specificity, and PPV in radiologists significantly improved by 0.09, 0.07, and 0.16, respectively, especially for identifying the vascular etiology of ICH. The PPV in nonradiologists also increased by 0.11.

XSL•FO
RenderX

Table . Diagnostic performance of readers with and without artificial intelligence support by clinical specialty.

Cause of acute ICH[a] and performance metric	Nonradiology (n=8)			Radiology (n=12)		
	Without model assistance	With model assistance	*P* value	Without model assistance	With model assistance	*P* value
Overall						
Accuracy (95% CI)	0.68 (0.65-0.71)	0.72 (0.70-0.74)	.02[b]	0.72 (0.70-0.74)	0.76 (0.74-0.78)	.003[b]
Primary						
Sensitivity (95% CI)	0.43 (0.32-0.53)	0.50 (0.40-0.59)	.31	0.49 (0.43-0.55)	0.57 (0.50-0.64)	.20
Specificity (95% CI)	0.81 (0.75-0.87)	0.80 (0.74-0.86)	.74	0.84 (0.79-0.88)	0.85 (0.82-0.88)	.66
PPV[c] (95% CI)	0.58 (0.53-0.63)	0.59 (0.55-0.63)	.84	0.65 (0.58-0.71)	0.68 (0.65-0.72)	.27
Tumorous						
Sensitivity (95% CI)	0.75 (0.62-0.88)	0.82 (0.76-0.88)	.31	0.87 (0.82-0.91)	0.88 (0.84-0.91)	.57
Specificity (95% CI)	0.75 (0.66-0.84)	0.79 (0.74-0.84)	.31	0.76 (0.71-0.80)	0.77 (0.75-0.79)	.79
PPV (95% CI)	0.58 (0.51-0.65)	0.64 (0.59-0.69)	.12	0.62 (0.58-0.65)	0.63 (0.61-0.64)	.57
Vascular malformation						
Sensitivity (95% CI)	0.42 (0.36-0.48)	0.47 (0.42-0.51)	.20	0.41 (0.34-0.48)	0.50 (0.45-0.54)	.01[b]
Specificity (95% CI)	0.72 (0.66-0.78)	0.80 (0.75-0.83)	.06	0.78 (0.75-0.81)	0.85 (0.81-0.88)	.02[b]
PPV (95% CI)	0.44 (0.38-0.49)	0.55 (0.49-0.58)	.04[b]	0.47 (0.42-0.53)	0.63 (0.57-0.69)	.005[b]

[a]ICH: intracerebral hemorrhage.

[b]Statistically significant difference in performance between clinicians with and without model assistance.

[c]PPV: positive predictive value.

Diagnostic Performance of Readers With and Without AI Support by Level of Expertise

Table 4 compares the diagnostic performance of readers, grouped as trainees and specialists, in identifying causes of acute ICH with and without the assistance of AI. The specialists outperformed the trainees in both nonassistance and assistance situations. After adopting the model, trainee and specialist readers improved significantly in determining what caused ICH. The overall accuracies increased from 0.69 to 0.74 for trainees and from 0.72 to 0.75 for specialists. For primary causes, the assistance of the AI model improved diagnostic performance in both reader groups. The sensitivity for trainees significantly increased from 0.46 to 0.54. After the model assisted in identifying tumor causes of ICH, all performance metrics for both trainees and specialists increased by approximately 0.01 and 0.07, except for sensitivity for specialists, which slightly dropped by 0.02. Interestingly, the assistance from the model was able to enhance the diagnostic performance for identifying vascular causes of ICH. The sensitivity, specificity, and PPV in the trainee group significantly increased from 0.41, 0.73, and 0.43 to 0.49, 0.73, and 0.59, respectively. For specialists, these performance measurements also improved: the sensitivity slightly increased from 0.42 to 0.49, whereas the specificity and PPV significantly improved from 0.79 and 0.50 to 0.84 and 0.62, respectively.

XSL•FO
RenderX

Table . Diagnostic performance of readers with and without artificial intelligence (AI) support by level of expertise.

Cause of acute ICH[a] and performance metric	Trainees (n=12)			Specialists (n=8)		
	Without AI assistance	With AI assistance	*P* value	Without AI assistance	With AI assistance	*P* value
Overall						
Accuracy (95% CI)	0.69 (0.67-0.72)	0.74 (0.73-0.76)	.003[b]	0.72 (0.68-0.75)	0.75 (0.72-0.78)	.02[b]
Primary						
Sensitivity (95% CI)	0.46 (0.40-0.52)	0.54 (0.48-0.61)	.045[b]	0.47 (0.36-0.58)	0.54 (0.42-0.65)	.06
Specificity (95% CI)	0.81 (0.77-0.86)	0.83 (0.79-0.86)	.55	0.84 (0.78-0.90)	0.83 (0.77-0.88)	.18
PPV[c] (95% CI)	0.60 (0.54-0.67)	0.65 (0.60-0.70)	.17	0.64 (0.58-0.71)	0.64 (0.60-0.69)	.93
Tumorous						
Sensitivity (95% CI)	0.78 (0.69-0.88)	0.85 (0.81-0.89)	.09	0.88 (0.83-0.92)	0.86 (0.81-0.92)	.67
Specificity (95% CI)	0.77 (0.72-0.81)	0.79 (0.76-0.82)	.37	0.74 (0.65-0.82)	0.76 (0.73-0.80)	.38
PPV (95% CI)	0.60 (0.55-0.64)	0.64 (0.61-0.67)	.14	0.61 (0.55-0.67)	0.62 (0.59-0.64)	.31
Vascular malformation						
Sensitivity (95% CI)	0.41 (0.34-0.47)	0.49 (0.44-0.53)	.04[b]	0.42 (0.35-0.49)	0.49 (0.44-0.53)	.06
Specificity (95% CI)	0.73 (0.69-0.77)	0.82 (0.78-0.85)	.02[b]	0.79 (0.75-0.83)	0.84 (0.79-0.89)	.02[b]
PPV (95% CI)	0.43 (0.39-0.48)	0.59 (0.52-0.63)	.009[b]	0.50 (0.44-0.56)	0.62 (0.54-0.70)	.02[b]

[a]ICH: intracerebral hemorrhage.

[b]Statistically significant difference in performance between clinicians with and without model assistance.

[c]PPV: positive predictive value.

Discussion

As is common in external validation studies, the performance of the model by Thabarsa et al [17] on external data showed decreased accuracy compared to its performance on internal data by approximately 17.7%, from 0.79 to 0.65. However, the performance of readers, including trainees, clinicians, and radiologists, improved after using AI-assisted interpretations of the causes of spontaneous ICH.

External Validation of Model Performance

The lack of emphasis on external validation as a critical factor in assessing the clinical utility of AI models is reflected in the fact that only 6% to 13% of AI papers in medical imaging include external validation [18-20]. This oversight limits the understanding of these models' true potential in real-world applications. Our study addresses this critical gap by demonstrating that the machine learning model by Thabarsa et al [17] exhibited inferior performance on external datasets compared to its performance on internal data. This finding highlights the challenges of generalizability and underscores the necessity of testing AI models on diverse, independent datasets before clinical deployment. Additionally, our results are consistent with those of a systematic review of the external validation of deep learning models for radiologic diagnosis [21]. In that review, it was discovered that 81% of the studies experienced a decline in performance as a result of the use of external datasets. The extent of the decline varied from a slight decrease to a 24% decrease.

The precise reasons for the reduced performance of machine learning algorithms using external datasets remain mostly unidentified. Our study's external dataset revealed no significant differences in ICH locations and patient age among the various causes of ICH. These characteristics were incorporated into the initial machine learning model in conjunction with radiomics features as they enhanced the model's performance. These variations in data characteristics may have contributed to the model's diminished performance when applied to external data. Different CT machines provided the training data and external validation data for our investigation. The thickness of the CT scan was also different between the different scanners. Although a previous study reported that the differences in scanners, imaging techniques or protocols, and reconstruction parameters can substantially affect the stability and reproducibility of radiomics features [22], in this study, the difference in CT scan

XSL•FO
RenderX

manufacturers did not affect the predictive error. In addition, the different settings of slice thickness were not significantly related to the predictive error. The reduced performance in the external validation was mainly due to variations in population characteristics between the external validation sample and the training dataset.

Before clinical deployment, it is recommended that the model be validated using independent data that share similar characteristics with the data expected in the intended deployment setting. As the model provides a prediction with a confidence score, the score should be considered and used for decision-making. Furthermore, explainable AI techniques such as Shapley additive explanations and local interpretable model-agnostic explanations can be adopted to understand how the model makes decisions for an input sample. Shapley additive explanations compute the contribution scores of each feature to the prediction. Meanwhile, local interpretable model-agnostic explanations explain the individual predictions of any machine learning model by approximating its behavior using a simpler, interpretable model (eg, linear model and decision tree) in the local region of the prediction. The explainable AI techniques provide useful information that radiologists could consider before deciding. In practical use, when a radiologist's impression differs from the model's prediction, especially when the model indicates a hemorrhage, the model's output may serve as a prompt to broaden the differential diagnosis. This could encourage further investigation, such as vascular imaging or follow-up studies, to reach a definite diagnosis. Although the model does not provide voxel-level explanations, its prediction can still function as a clinically meaningful cue that supports radiologists in considering alternative etiologies.

Clinician Diagnostic Performance With and Without AI Support

A prior study assessed the effectiveness of a machine learning model using radiomics features derived from CT scans in distinguishing arteriovenous malformation–related ICHs from those resulting from other causes compared to radiologists' evaluations [14]. The outcomes showed that the model outperformed radiologists. Nonetheless, the primary function of machine or deep learning models is to assist radiologists and radiology trainees rather than replace them in clinical environments. AI systems' decision-making procedures lack openness compared to radiologists' clinical interpretation due to their "black box" design [23]. As a consequence, our investigation assessed the performance of trainees and specialty readers both with and without consideration of the results of the AI model. Without the assistance of the AI model, the nonradiology group demonstrated a lower performance than that of the radiology group, and trainees lagged behind specialists. However, our findings revealed that the use of AI improved reader performance, narrowing the gap between these groups. The model significantly improved the performance of all reader groups in predicting vascular malformations as the cause of spontaneous ICH. Notably, AI assistance had the greatest impact on nonradiology readers and trainees, elevating their performance to levels comparable to that of radiologists or specialists working without AI support.

This model has potential value in the context of real-world medicine. Most emergency radiology units, including that in our hospital, depend on radiology residents to perform some of their nighttime duties [24]. Radiology residents' work is regarded as a critical component of their education as it not only minimizes the burden of the attending radiologists but also improves the abilities of the residents. However, studies have shown that diagnostic error rates tend to increase during night shifts, particularly after midnight, likely due to fatigue and reduced alertness [25]. This AI model could support nighttime shifts, which could reduce diagnostic errors while maintaining the educational benefits of resident training. Moreover, while using the model, the performance of groups of professionals and radiology readers also showed the best results. Meanwhile, radiology residency programs should foster independent reasoning, critical evaluation, and comprehensive understanding of AI principles [26]. To prevent the integration of AI from undermining residents' skill development, maintaining a balanced exposure to both AI-assisted and independent case interpretation is crucial. Regular comparison of performance, analysis of discrepancies between AI and human interpretations, and reflective discussion can help sustain residents' analytical autonomy while supporting the responsible incorporation of AI into radiology education.

AI-assisted data interpretation significantly improved all readers' ability to correctly identify vascular malformation causes. Unlike primary ICH, vascular malformation–related ICH demonstrates greater complexity and heterogeneity in internal composition and morphology due to the presence of malformed blood vessels, as reflected in radiomics features [13]. This finding aligns with those of previous research showing that radiomics features can effectively differentiate vascular malformations from other causes of ICH [13,14]. AI support also improves the capacity of clinicians to predict primary and tumor-related ICH but does not achieve statistical significance.

There are several limitations to this study. First, the relatively small sample size and retrospective design may limit the robustness and reproducibility of the findings. The small sample size likely resulted in low statistical power. It may not yield statistically significant results even if the model has some predictive ability. However, our study also attempted to incorporate external data, a topic on which there is currently limited research for machine learning models with radiomics features. Larger studies or additional external datasets would be useful for future studies. Second, the manual segmentation process used was time-consuming and subjective and introduced inter- and intra-annotator variability, potentially biasing the model's performance. Third, the model was developed using data from a single institution and externally validated on data from only 1 additional hospital, which may limit its generalizability to other centers with different scanners, imaging protocols, or patient populations. Future research could focus on prospective validation and on exploring the role of AI in enabling fast and accurate segmentation [27,28], which would help standardization and translation of radiomics into real-world radiology.

XSL•FO
RenderX

Conclusions

This study's findings demonstrated that the machine learning model by Thabarsa et al [17] for categorizing spontaneous ICHs using radiomics characteristics exhibited reduced performance when applied to an external dataset. Nevertheless, the model significantly enhanced the performance of all reader groups, including trainees and both radiology and nonradiology specialists, when used as a supportive tool.

Acknowledgments

The authors would like to thank the Faculty of Medicine and the Faculty of Science of Chiang Mai University, Thailand, for supporting this study. This study was also facilitated by Ramathibodi Hospital, Mahidol University, Thailand. The authors confirm that no generative artificial intelligence technologies were used in the creation of this work.

Funding

No particular grant from any commercial or nonprofit funding organization was given for this research.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

SA conceptualized the study, designed the methodology, conducted the investigation process of this study, verified the experimental results, and drafted the manuscript. NJ, KU, CM, WV, and TS annotated and maintained the research data. NJ performed the investigation process. TS provided samples for the experiments. P Thabarsa performed the investigation process, implemented the software, created visualizations for the published work, and drafted the manuscript. P Traisathit statistically analyzed the experimental results. CA verified and statistically analyzed the experimental results and rewrote the manuscript. PI conceptualized the study, designed the methodology, verified the experimental results, and drafted the manuscript. SA, CA, and PI acquired financial support for the project. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Rajashekar D, Liang JW. Intracerebral hemorrhage. In: StatPearls: StatPearls Publishing; 2025. URL: https://www.ncbi.nlm.nih.gov/books/NBK553103/ [accessed 2025-12-01]
2. Katsuki M, Kakizawa Y, Nishikawa A, Yamamoto Y, Uchiyama T. Endoscopic hematoma removal of supratentorial intracerebral hemorrhage under local anesthesia reduces operative time compared to craniotomy. Sci Rep 2020 Jun 25;10(1):10389. [doi: 10.1038/s41598-020-67456-x] [Medline: 32587368]
3. Choi YS, Rim TH, Ahn SS, Lee SK. Discrimination of tumorous intracerebral hemorrhage from benign causes using CT densitometry. AJNR Am J Neuroradiol 2015 May;36(5):886-892. [doi: 10.3174/ajnr.A4233] [Medline: 25634719]
4. Joseph DM, O'Neill AH, Chandra RV, Lai LT. Glioblastoma presenting as spontaneous intracranial haemorrhage: case report and review of the literature. J Clin Neurosci 2017 Jun;40:1-5. [doi: 10.1016/j.jocn.2016.12.046] [Medline: 28215428]
5. Tseng JH, Lin WH. Glioblastoma multiforme hiding behind the intracerebral hematoma. Formos J Surg 2012 Dec;45(6):183-186. [doi: 10.1016/j.fjs.2012.08.001]
6. Kambe A, Hosoya T, Sakamoto M, Kondo S, Kurosaki M. High-grade glioma masquerading as a small cerebral hemorrhage: a case report. Yonago Acta Med 2019 Dec 18;62(4):305-307. [doi: 10.33160/yam.2019.11.004] [Medline: 31849570]
7. Park H, Kang HS, Cho WS. Delayed appearance of radiologically occult cerebral arteriovenous malformation: a case report and literature review. J Korean Neurosurg Soc 2023 Mar;66(2):199-204. [doi: 10.3340/jkns.2022.0165] [Medline: 36071567]
8. Akinci D'Antonoli T, Cuocolo R, Baessler B, Pinto Dos Santos D. Towards reproducible radiomics research: introduction of a database for radiomics studies. Eur Radiol 2024 Jan;34(1):436-443. [doi: 10.1007/s00330-023-10095-3] [Medline: 37572188]
9. Angkurawaranon S, Sanorsieng N, Unsrisong K, et al. A comparison of performance between a deep learning model with residents for localization and classification of intracranial hemorrhage. Sci Rep 2023 Jun 20;13(1):9975. [doi: 10.1038/s41598-023-37114-z] [Medline: 37340038]
10. Yasaka K, Hatano S, Mizuki M, et al. Effects of deep learning on radiologists' and radiology residents' performance in identifying esophageal cancer on CT. Br J Radiol 2023 Oct;96(1150):20220685. [doi: 10.1259/bjr.20220685] [Medline: 37000686]
11. Bennani S, Regnard NE, Ventre J, et al. Using AI to improve radiologist performance in detection of abnormalities on chest radiographs. Radiology 2023 Dec;309(3):e230860. [doi: 10.1148/radiol.230860] [Medline: 38085079]

XSL•FO
RenderX

12. Do KH, Beck KS, Lee JM. The growing problem of radiologist shortages: Korean perspective. Korean J Radiol 2023 Dec;24(12):1173-1175. [doi: 10.3348/kjr.2023.1010] [Medline: 38016676]
13. Wang J, Xiong X, Ye J, et al. A radiomics nomogram for classifying hematoma entities in acute spontaneous intracerebral hemorrhage on non-contrast-enhanced computed tomography. Front Neurosci 2022 Jun 10;16:837041. [doi: 10.3389/fnins.2022.837041] [Medline: 35757547]
14. Zhang Y, Zhang B, Liang F, et al. Radiomics features on non-contrast-enhanced CT scan can precisely classify AVM-related hematomas from other spontaneous intraparenchymal hematoma types. Eur Radiol 2019 Apr;29(4):2157-2165. [doi: 10.1007/s00330-018-5747-x] [Medline: 30306329]
15. Lyu J, Xu Z, Sun H, Zhai F, Qu X. Machine learning-based CT radiomics model to discriminate the primary and secondary intracranial hemorrhage. Sci Rep 2023 Mar 6;13(1):3709. [doi: 10.1038/s41598-023-30678-w] [Medline: 36879050]
16. Nawabi J, Kniep H, Kabiri R, et al. Neoplastic and non-neoplastic acute intracerebral hemorrhage in CT brain scans: machine learning-based prediction using radiomic image features. Front Neurol 2020 May 5;11:285. [doi: 10.3389/fneur.2020.00285] [Medline: 32477233]
17. Thabarsa P, Inkeaw P, Madla C, et al. Machine learning based classification of spontaneous intracranial hemorrhages using radiomics features. Neuroradiology 2025 Feb;67(2):339-349. [doi: 10.1007/s00234-024-03481-1] [Medline: 39367990]
18. Kim DW, Jang HY, Kim KW, Shin Y, Park SH. Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: results from recently published papers. Korean J Radiol 2019 Mar;20(3):405-410. [doi: 10.3348/kjr.2019.0025] [Medline: 30799571]
19. Yao AD, Cheng DL, Pan I, Kitamura F. Deep learning in neuroradiology: a systematic review of current algorithms and approaches for the new wave of imaging technology. Radiol Artif Intell 2020 Mar 4;2(2):e190026. [doi: 10.1148/ryai.2020190026] [Medline: 33937816]
20. Nguyen AV, Blears EE, Ross E, Lall RR, Ortega-Barnett J. Machine learning applications for the differentiation of primary central nervous system lymphoma from glioblastoma on imaging: a systematic review and meta-analysis. Neurosurg Focus 2018 Nov 1;45(5):E5. [doi: 10.3171/2018.8.FOCUS18325] [Medline: 30453459]
21. Yu AC, Mohajer B, Eng J. External validation of deep learning algorithms for radiologic diagnosis: a systematic review. Radiol Artif Intell 2022 May 4;4(3):e210064. [doi: 10.1148/ryai.210064] [Medline: 35652114]
22. Zhao B. Understanding sources of variation to improve the reproducibility of radiomics. Front Oncol 2021 Mar 29;11:633176. [doi: 10.3389/fonc.2021.633176] [Medline: 33854969]
23. Poon AI, Sung JJ. Opening the black box of AI-medicine. J Gastroenterol Hepatol 2021 Mar;36(3):581-584. [doi: 10.1111/jgh.15384] [Medline: 33709609]
24. Vaattovaara E, Nikki M, Nevalainen M, Ilmarinen M, Tervonen O. Discrepancies in interpretation of night-time emergency computed tomography scans by radiology residents. Acta Radiol Open 2018 Oct 22;7(10):2058460118807234. [doi: 10.1177/2058460118807234] [Medline: 30364822]
25. Zhang L, Wen X, Li JW, Jiang X, Yang XF, Li M. Diagnostic error and bias in the department of radiology: a pictorial essay. Insights Imaging 2023 Oct 2;14(1):163. [doi: 10.1186/s13244-023-01521-7] [Medline: 37782396]
26. Wang C, Xie H, Wang S, Yang S, Hu L. Radiological education in the era of artificial intelligence: a review. Medicine (Baltimore) 2023;102(1):e32518 Retracted inMedicine (Baltimore). 2024 May 31;103(22):e38552. [doi: 10.1097/MD.0000000000038552] [Medline: 39259133]
27. Inkeaw P, Angkurawaranon S, Khumrin P, et al. Automatic hemorrhage segmentation on head CT scan for traumatic brain injury using 3D deep learning model. Comput Biol Med 2022 Jul;146:105530. [doi: 10.1016/j.compbiomed.2022.105530] [Medline: 35460962]
28. Schmitt N, Mokli Y, Weyland CS, et al. Automated detection and segmentation of intracranial hemorrhage suspect hyperdensities in non-contrast-enhanced CT scans of acute stroke patients. Eur Radiol 2022 Apr;32(4):2246-2254. [doi: 10.1007/s00330-021-08352-4] [Medline: 34773465]

Abbreviations

AI: artificial intelligence
CT: computed tomography
GLCM: gray-level co-occurrence matrix
GLDM: gray-level dependence matrix
GLSZM: gray-level size-zone matrix
ICH: intracerebral hemorrhage
NCCT: noncontrast computed tomography
PPV: positive predictive value
SAH: subarachnoid hemorrhage

XSL•FO
RenderX

Edited by A Schwartz, M Balcarras; submitted 22.07.25; peer-reviewed by P Shingru, P Sharma; accepted 25.11.25; published 11.12.25.

Please cite as:
Angkurawaranon S, Jitmahawong N, Unsrisong K, Thabarsa P, Madla C, Vuthiwong W, Sudsang T, Angkurawaranon C, Traisathit P, Inkeaw P
Radiomics-Based AI Model to Assist Clinicians in Intracranial Hemorrhage Diagnosis: External Validation Study
JMIR Form Res 2025;9:e81038
URL: https://formative.jmir.org/2025/1/e81038
doi:10.2196/81038

© Salita Angkurawaranon, Natipat Jitmahawong, Kittisak Unsrisong, Phattanun Thabarsa, Chakri Madla, Withawat Vuthiwong, Thanwa Sudsang, Chaisiri Angkurawaranon, Patrinee Traisathit, Papangkorn Inkeaw. Originally published in JMIR Formative Research (https://formative.jmir.org), 11.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Lessons Learned from Co-Designing a Digital Health App for Foster Youth: Development and Usability Study

Johanna B Folk[1], PhD; Juan Carlos Gonzalez[1,2], PhD; Margareth V Del Cid[1], PhD; Elizabeth McBride[1], BA; Tylia Lundberg[1], BA; Alison Czopp[1], MSW; Ifunanya Ezimora[1,3], BA; Lisa R Fortuna[4], MD, MPH, MDiv; Marina Tolou-Shams[1], PhD

[1]Department of Psychiatry and Behavioral Sciences, School of Medicine, University of California, San Francisco, 1001 Potrero Ave, San Francisco, CA, United States

[2]Department of Learning and Teaching, School of Leadership and Education Sciences, University of San Diego, San Diego, CA, United States

[3]San Diego State University/University of California, San Diego Joint Doctoral Program in Clinical Psychology, San Diego, CA, United States

[4]Department of Psychiatry and Neuroscience, School of Medicine, University of California, Riverside, Riverside, CA, United States

Corresponding Author:
Johanna B Folk, PhD
Department of Psychiatry and Behavioral Sciences, School of Medicine, University of California, San Francisco, 1001 Potrero Ave, San Francisco, CA, United States

Abstract

Background: Foster youth experience high rates of unmet mental health and substance use needs, while simultaneously facing numerous barriers to accessing and engaging in community-based services. Behavioral intervention technologies (BITs) are promising for overcoming some of the barriers to service engagement, particularly when designed in collaboration with the intended users.

Objective: This study describes lessons learned from a 31-month process of co-designing FostrSpace (Chorus Innovations, Inc), a BIT to address social determinants of health and behavioral health needs among foster youth. Our overall aim is to provide a roadmap for other scholars wishing to co-design BITs with minoritized youth that have the potential to address social determinants of health and increase access to and engagement in behavioral health care.

Methods: The co-design process of creating FostrSpace included 5 phases: design, development, launch, testing and evaluation, and iterative refinement. We describe the activities conducted during each phase, as well as the resultant FostrSpace application. In-application FostrSpace usage data were collected as part of a quality improvement effort to iteratively refine the application; during registration, all youth signed a user agreement that detailed data usage.

Results: FostrSpace usage data were collected from 40 youth (32/40, 78% aged 18‑26 years; 8/40, 20% 13‑17 years). Based on the resource needs checklist at sign-up, youth sought resources in the domains of emotional wellness (29/40; 72.5%), health care (17/40, 42.5%), housing (16/20, 40%), transportation (15/40, 37.5%), employment (15/40, 37.5%), school (13/40, 32.5%), food (12/40, 30%), family (11/40, 27.5%), and legal (7/40, 17.5%) resources, or other or not sure (16/20, 40%). Fifteen youth accessed support from the personal care navigator. Fourteen youth completed the emotional wellness questionnaire (EWQ) and identified substance use, depression, anger and irritability, mania, anxiety, somatic symptoms, and sleep problems as areas of concern. Seven of these youth initiated behavioral health services with a FostrSpace clinician.

Conclusions: Engaging in participatory co-design of BITs with foster youth and other minoritized communities requires careful attention to power dynamics. Creating a space where co-designers feel there is mutual benefit to engaging in the process and it is psychologically safe to share their experiences is crucial for success. We describe lessons learned from engaging in this co-design work, including how it relates to decisions about the technology (eg, balancing youth privacy with the burden of the login process), working with third-party developers (eg, ensuring technology development partners have sufficient knowledge about the population you are co-designing with to meaningfully engage with them), and considerations for the strategic embedding of technology-based interventions within existing systems of care to promote uptake.

(JMIR Form Res 2025;9:e73281) doi:10.2196/73281

KEYWORDS

foster care; adolescent; young adult; behavioral intervention technology; mHealth; co-design; app development; mental health; social determinants of health; behavioral health app

XSL•FO
RenderX

Introduction

Background

Foster youth, of whom there are over 350,000 across the United States [1], experience high rates of behavioral health (mental health and substance use) needs, yet simultaneously face numerous barriers to accessing and engaging in community-based behavioral health care [2,3]. Compared to their same-age peers, foster youth exhibit higher rates of psychiatric disorders (eg, major depressive disorder and posttraumatic stress disorder), suicidality (eg, ideation and attempt) [4], and substance use and are more likely to initiate substance use at earlier ages [5]. Data also suggest that among foster youth 11 - 14 years old who had clinically elevated mental health concerns, only 26% used services [6]. These gaps in access often continue through transition age years (16 - 24 years), with youth reporting ongoing barriers to care such as mistrust of the system and transportation challenges [7].

These observed inequities in behavioral health needs and services access are not a result of individual deficits, but rather, a product of complex social and systemic issues, such as poverty and structural racism, that disproportionately impact minoritized groups [8], including foster youth. Of note, families racialized as Black, Indigenous, and Latine are consistently overrepresented in the child welfare system [9,10], with Black youth in particular being more likely to be removed from their homes [11,12].

A growing body of literature has associated behavioral health needs with social determinants of health (SDoH), the social conditions into which a youth is born and grows [13]. Foster youth, who are disproportionately from minoritized ethnic and racial groups, are disadvantaged by multiple SDoH, for example, early childhood exposure to violence, child maltreatment [14], and unstable housing as a result of placement transitions. They are also less likely to access higher education and complete high school [15,16], more likely to experience economic hardship, and less likely to access needed services compared to nonfoster youth peers [6]. Thus, developing accessible supports for foster youth to address unmet behavioral health needs and reduce the undue burden of SDoH is essential. Ensuring these supports acknowledge root cause issues (eg, structural racism) is equally important to securing lasting social change [17].

Common barriers to accessing and engaging in evidence-based interventions among foster youth include frequent relocations due to multiple placement transitions, complex and multifaceted treatment needs, and services that are not designed with their unique experiences in mind [18-21]. Technology-based interventions are promising for overcoming some of these barriers, particularly when designed in collaboration with the intended users (ie, foster youth). The present study describes lessons learned from co-designing a technology-based intervention with and for foster youth to address SDoH and behavioral health needs. While terms like “users” and “end users” are commonly reported in the field to refer to those engaging in the co-design process, we chose instead to center terms such as “youth” and “foster youth” to communicate that these young people are respected and valued regardless of their relationship to the technology in development.

Promise of Behavioral Intervention Technologies

Developing, testing, and implementing technology-based strategies to address documented barriers to care is critical in the journey toward an equitable behavioral health system where foster youth can access the support they desire and need. Technology has increasingly been leveraged to address inequities among minoritized racial and ethnic groups [22,23] and to deliver interventions to foster youth and their families [24,25]. Behavioral intervention technologies (BITs; “websites, software, mobile apps, and sensors designed to help users address or change behaviors, cognitions, and emotional states”; [26]) have the potential to support foster youth in directly accessing resources to address SDoH and behavioral health needs [27]. Mental health apps—one type of BIT—have proliferated, with estimates suggesting over 20,000 exist for download [28]. BITs for youth can promote easy access to behavioral health care by providing intervention at any time and allowing for personalization [29,30].

However, simple repackaging of evidence-based interventions for digital use may be insufficient to address existing inequities. Research has shown that despite an influx in the number of BITs developed, engagement with these technologies continues to be low [31]. Participatory methods that engage youth from the start of BIT development have been proposed as one potential solution to increase engagement [32,33]. Of the existing mental health apps for young people, however, only 30 have published descriptions of being co-designed with the youth they intend to serve, and only 3 were co-designed specifically with minoritized youth communities (eg, indigenous youth, lesbian, gay, and bisexual youth) [32]; this paucity of literature suggests significant room for the field to improve in its use of participatory co-design approaches with youth.

Participatory Co-Design of BITs

Co-design approaches to developing BITs are intended to center youth involvement through a range of user-centered design principles, such as creative collaboration, iterative design, and acknowledgment of power differentials [32,34]. Recentering minoritized communities through such processes is also one of many necessary steps toward countering structural oppression within existing systems of care [35,36]. But the term “co-design” has been used in the literature to reference a continuum of involvement, ranging from brief consultation to more intensive, integrated involvement through mutual partnership. Greater specificity is needed to understand the utility and application of co-design approaches [32,33] and to avoid the potential of harm being perpetuated through the use of these very methods when not applied thoughtfully [37].

Participatory informatics is one approach to the co-design of BITs [38-40]. Drawing on principles of community-partnered participatory research (eg, equity and power sharing) and user-centered design (eg, active user participation in design), participatory informatics involves including people from the population you intend to serve as equitable members throughout the design process [38,39]. This approach does not require the

XSL•FO
RenderX

youth to have technological expertise and instead places value on their lived expertise. Participatory informatics is therefore able to address community priorities through tools that are engaging and relevant, while also increasing co-designers' own perceived competency in creating BITs, fostering a culture of co-leadership related to the work, and promoting sustained engagement with BITs; additional information on how we applied this to our co-design process is outlined below. Participatory informatics has been used in studies with ethnically and racially minoritized communities to address a wide range of health needs [38,41,42] and within a variety of health care settings [41,43]. To our knowledge, this approach has yet to be applied with foster youth.

This Study

Participatory co-design of BITs with minoritized young people, such as foster youth, requires thoughtful attention to power dynamics, varying levels of technology access and literacy, and a range of SDoH needs, among other considerations. To support other scholars interested in co-designing BITs with minoritized youth to address SDoH and increase access to and engagement in behavioral health care, the current paper presents lessons learned from a 31-month participatory co-design process of a BIT ("FostrSpace"; Chorus Innovations, Inc) created by and for foster youth in California. We define the app development process (co-design and its benefits), provide a detailed description of the FostrSpace platform, and summarize lessons learned. A focus is placed on how participatory co-design was used in order to center the voices of foster youth, allowing them to speak to their needs and the impact of health inequities. For example, youth co-designers described domains of resources and needs they wanted represented within the FostrSpace app; the research team then presented the SDoH framework as a way to organize the app (ie, app features and content were organized around major areas of SDoH, including health care access and quality, educational access, and economic stability). App usage data, collected as part of quality improvement efforts, is used to understand the behavioral health and resource needs of users. This work builds upon our prior examinations of co-designers' reflections on the participatory co-design process [44] and initial usage data for FostrSpace in an ongoing hybrid implementation-effectiveness trial in partnership with Court Appointed Special Advocate (CASA) programs [45]. Through this description of our methodological application of co-design with foster youth, we aim to provide relevant insights for other scholars and ultimately to encourage the use of such methods to advance behavioral health equity for minoritized youth.

Methods

Recruitment

Co-Design Participants and Process

Co-design of FostrSpace began in the planning stages. Idea-sharing workgroups, advertised using digital recruitment flyers distributed by Bay Area organizations serving foster youth, were held with 5 present and former foster youth (ages 16‑26 years); two 90-minute groups were held, one with 3 participants and the other with 2. Workgroup participants were compensated with a US $75 gift card and invited to join the FostrSpace Advisory Board (FAB). FAB members (4 of the 5 workgroup members) were hired as university employees and paid an hourly rate (based on their individual compensation package, which was determined by education level and experience as set by the university) for participation in the co-design process. The FAB consisted of 4 transition-age youth; their personal background and experience co-designing the initial version of FostrSpace have been described elsewhere [44]. For 18 months, the FAB participated in weekly 90‑to 120-minute meetings with a team of academic clinical researchers (academic team) and developers at Chorus Innovations, Inc, a Health Insurance Portability and Accountability Act (HIPAA)–compliant web-based app development platform.

The process of creating FostrSpace included 5 phases (see Figure 1): design, development, launch, testing and evaluation, and iterative refinement. In the design phase, meetings focused on gathering feedback from the FAB about the app interface, framework, features, and functionality. During these meetings, the FAB team created the app name and theme and identified the primary features of the app. These app features organically fit within an SDoH framework, which addresses key areas of social impact that influence youth behavioral health outcomes (eg, health care access and quality, education access and quality, social and community context, economic stability, and neighborhood and built environment). Co-designers engaged in shared decision-making about FostrSpace features and design, prioritizing the FAB preferences at all stages and taking into account what was possible to create within the existing platform and resources. In the development phase, the Chorus team built the initial version of FostrSpace, the FAB created psychoeducational and testimonial app content, and the care team (ie, behavioral health providers and personal care navigators hired to provide in-app services) developed clinical workflows and triage processes. All content was tested extensively by the FAB and academic team and iteratively refined to improve usability. The launch phase focused on dissemination and outreach to enroll youth to test the app. In the testing and evaluation phase, the team engaged in ongoing strategic testing and quality assurance initiatives to ensure the app was effective, efficient, and user-friendly. Qualitative feedback from youth, including the FAB, informed changes to the app in the iterative refinement phase. The "FAB Chat," which offers daily in-app peer support from our FAB team to cultivate community, was added during this phase in response to youth feedback. Youth also requested content focused on SDoH relevant to the transition to adulthood, which inspired the FAB to create the "Adulting 101" content and video series, including independent living resources (eg, how to apply to a job or food benefits). Finally, also during the iterative refinement phase, the FostrSpace app and academic team were awarded a federal Children and Youth Resilience Challenge grant, which spanned from September 2023 to May 2024. Through this challenge, our team added 2 new FAB members with lived experience in the foster care system (2 existing members became inactive at this time due to competing personal demands) and added a measure of youth resilience into the FostrSpace app. The FAB worked, on average, 4 hours each week during this

XSL•FO
RenderX

time to create new app content, test and refine existing content and features, and engage in outreach activities.

Figure 1. FostrSpace app co-design process.

FAB members had a range of experiences in both the traditional foster care system and kinship/relative care and represented a diverse set of backgrounds across race, ethnicity, sexual orientation, gender, disability status, and socioeconomic status (eg, White, White-passing, mixed Asian-Hispanic, Mexican-Latino, and Chicana; nonbinary, woman, and men).

Co-Designed Product: the FostrSpace App

Combining "Foster" and "'Space," the name FostrSpace was selected in partnership with the FAB to easily identify the app as a welcoming space for foster youth. The app uses imagery such as astronauts, planets, stars, and rocket fuel to create a lively and developmentally appropriate virtual space for users to explore their mental wellness.

FostrSpace provides access to a wide range of both self-guided and service-oriented features (eg, navigation and clinical services). Self-guided features include psychoeducational content (eg, breathing exercises and regulation skills) and a curated resource directory organized by SDoH area (eg, school and applying to college, employment and training, health care; see Figure 2, Find Resources screenshot) across several SDoH domains (eg, educational, occupational, health care, and legal). Self-guided assessments for emotional wellness and resource needs are incorporated to personalize app content for foster youth and inform the care team of the youth's self-identified priority needs. Synchronous service-oriented features include chat-based support from peers with lived experience in the foster care system, resource linkage support from a personal care navigator, and behavioral health services via telehealth, which are available weekdays except on holidays from 9 AM to 6 PM (see Table 1).

Foster youth can engage with the app independently or with support from their care team, with features designed to build self-awareness (eg, mood tracking) and promote autonomy (eg, missions). For example, youth have the option to complete a single-item mood check-in question upon each login on the landing page of the app (see Figure 2, Quick Check-In screenshot), with the option to briefly describe their mood in a write-in section; responses are logged over time so youth can revisit them for personal tracking. Youth can also set and track goals (eg, completion of app tutorials and submission of job applications) through "Missions" (see Figure 2, Pick Up Groceries screenshot). When missions are self-generated, foster youth are prompted to add a description of their goal, a timeline on intended goal completion, and log any relevant hyperlinks they might need to complete the task (eg, job application site in the case of a vocational goal). If youth engage in services, goals can also be assigned by the care team to facilitate completion of app-based content, therapy goals, or in-app assessments. Peer support with mood tracking and goal organization is accessible through in-app chat with the FAB.

Youth seeking emotional wellness resources are invited to complete the in-app Emotional Wellness Questionnaire (EWQ; a co-designed behavioral health assessment including items from the *DSM-5* [*Diagnostic and Statistical Manual of Disorders* {Fifth Edition}], Self-Rated Level 1 Cross-Cutting Symptom for adults [18+ years] and youth [11-17 years] measuring frequency and severity of past 2-week symptoms [41]). All respondents receive feedback on their responses (see Figure 2, Emotional Wellness Quiz screenshot with astronaut and rocket fuel metaphor for emotional wellness, ie, behavioral health intervention need) and tailored resources (eg, self-guided psychoeducational content to learn coping skills), and for those whose responses are elevated, the suggestion to connect with a FostrSpace clinician. FostrSpace navigators send an in-app chat message within 48 business hours to youth with elevated scores to offer support with connecting to a FostrSpace clinician. Foster youth may also opt to message the navigator directly to request help initiating clinical services. Navigators serve as the primary point of contact, supporting youth in accessing appropriate mental health resources. For youth not yet ready to engage with a mental health clinician (whether through FostrSpace or external services), navigators address barriers to care and co-develop a gradual, youth-paced plan to facilitate connection to desired resources. Foster youth who engage with FostrSpace clinicians are asked to complete follow-up questionnaires (ie, DSM Level 2 Cross-Cutting Symptom Measures; [46]) to inform clinical intervention needs in particular area(s) (eg, depression, anxiety, substance use). All FostrSpace clinicians are trained and experienced in providing care to youth with current or prior history of foster care system contact.

XSL•FO
RenderX

Figure 2. FostrSpace app screenshots.

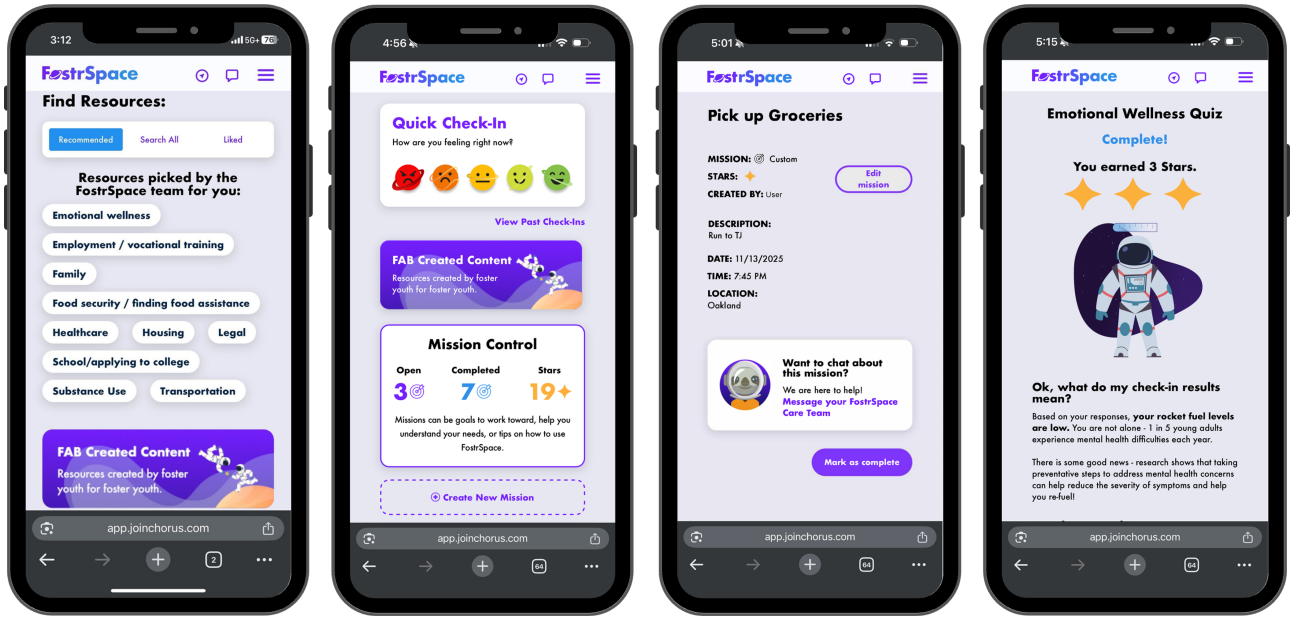

XSL•FO
RenderX

Table . FostrSpace app features and description of co-design process.

Type and feature	Description	Co-design process
Self-guided and asynchronous		
Resource directory	A curated directory where youth can independently find local resources and services, such as food pantries, job programs, and free clinics.	The FAB[a] identified key resource categories based on their lived experience and recommended local, county-specific resources.
Peer- created resources	The FAB created videos, handouts, and guides to help youth navigate adulthood and build independence, including FAB-led skill-building videos to promote healthy coping.	The FAB developed coping strategy videos and collaborated with the clinical team on script development. FAB members prioritized and led resource creation based on independent living needs in addition to coping skills.
Assessments	Youth can access assessments like the Emotional Wellness Questionnaire to explore their behavioral health or the Resource Needs Survey to identify resources. Youth also complete a resource needs checklist at sign-up to indicate sought-out resources.	The FAB advocated for an in-app emotional wellness check-in and refined screening tool questions to remove stigmatizing language.
Mood tracking	The FostrSpace Mood Tracker allows for daily mood tracking, which provides opportunities for self-awareness and empowers youth to take control of their emotional states.	The FAB proposed a simple emoji-based mood tracker, with the option to add written descriptions. The design team provided emoji options, and the FAB selected the ones used in the app.
Missions	Missions in the app help organize goals. For example, completing specific tasks, such as attending health appointments.	The FAB emphasized the importance of goal setting and tracking in the app, providing design suggestions to match its space theme.
Service and synchronous		
Peer support	The FAB provides daily in-app support via the “FAB Chat,” helping youth access resources and receive peer guidance.	The FAB highlighted the importance of built-in peer support to foster community. The research team structured workflows for the FAB chat function.
Navigation services	Personal care navigators help youth access social services, health care, and other resources to meet their basic needs (eg, supporting with building skills to complete paperwork, schedule appointments, etc).	The research team shared successful navigation models, while FAB emphasized the importance of opt-in services to prioritize foster youth autonomy.
Clinical services	Licensed clinicians provide evidence-based behavioral health services via telehealth.	The FAB communicated the importance of opt-in clinical services to prioritize foster youth autonomy, while the research team structured workflows for access.
In-app chat	The in-app chat connects youth with their care team, including the FAB.	The FAB advocated for in-app chat to reduce barriers to support by enabling real-time, confidential communication without phone calls or emails. Developers supported the chat feature build-out, allowing users to ask questions, share updates, and seek guidance while fostering connectivity and reducing isolation. The research team provided feedback on its functionality through testing.

XSL•FO
RenderX

[a]FAB: FostrSpace Advisory Board.

Launch Phase: Enrollment and Outreach

Outreach efforts included contacting 220 individuals across 153 schools through the Guardian Scholars Program (designed to support foster youth in higher education), 1805 school personnel across 58 counties through the California Department of Education Services, and 42 community-based organizations; in total, we outreached to 2300 contacts (from September 1, 2021, to April 13, 2024). In addition to cold emails and direct contact with community organizations that serve foster youth, our team presented at 5 state-level child welfare leadership meetings and hosted 8 virtual events for the statewide foster youth community directly, where training in policy and advocacy was provided and access to the app was highlighted. As described below, despite this extensive outreach, enrollment during the evaluation period was lower than anticipated (n=40).

Testing and Evaluation

Data reported in the current manuscript were collected between April 25, 2022, and November 14, 2024 in-app and used solely for quality improvement purposes to improve FostrSpace over time. An in-app data use agreement was signed by each youth as part of the FostrSpace sign-up process. Youth were provided with the following statement: “The data collected will be used to maintain the FostrSpace app’s functionality and performance and support FostrSpace service provisions. Information collected through FostrSpace could include your name, contact information, responses to surveys, and/or engagement activity (log-ins, completion of surveys). Information about who has access to your data and where it is stored can be accessed within the app under the “My Data” section. At any time, you can request to be removed from the platform and have your data deleted.” All data were stored within the FostrSpace app on the Chorus platform and used only for internal quality improvement purposes. Data sent to or from the FostrSpace app is transmitted over encrypted, industry-standard connections like SSL/TLS. The server is protected using standard practices, including being located behind a network firewall and accessible only by designated users. No data are stored locally on users’ devices (such as local browser storage or temporary storage) that access FostrSpace. App usage data were analyzed using descriptive statistics to understand the behavioral health and resource needs of users.

Ethical Considerations

The institutional review board of the University of California, San Francisco, determined that this research is exempt from institutional review board oversight according to the federal regulations summarized in Title 45 of the Code of Federal Regulations, Part 46.102(l). Demographic information was intentionally not collected from app users except for age, which was needed to determine eligibility. This decision was driven by our desire to ensure the app remained accessible and safe for a population of young people who are highly surveilled. Demographic information was collected for those engaging in services; however, given the small sample size, we do not include this information here.

Results

App platform data were collected from 40 foster youth (32/40, 80% aged 18 - 26 years; 8/40, 20% aged 13 - 17 years). On average, users logged into the app 3.88 (SD 5.26; range 1 - 22) times on 2.93 (SD 4.15; range 1 - 16) days. Twenty-eight users (28/40, 70%) completed at least one mood check-in, and among those who did, users logged their mood an average of 3.75 (SD 4.72; range 1 - 17) times in total. Twenty users (20/40, 50%) sent at least one message using the in-app chat.

Based on the resource needs checklist at sign-up, youth sought resources in the domains of emotional wellness (29/40; 72.5%), health care (17/40, 42.5%), housing (16/20, 40%), transportation (15/40, 37.5%), employment (15/40, 37.5%), school (13/40, 32.5%), food (12/40, 30%), family (11/40, 27.5%), and legal (7/40, 17.5%) resources, or other/not sure (16/40, 40%). Seventeen users accessed the resource directory, ten accessed the “Get Help Now” page with crisis resources, and 15 accessed support from the personal care navigator.

Fourteen youth completed the EWQ and identified needs related to substance use, depression, anger or irritability, mania, anxiety, somatic symptoms, and sleep problems. The most commonly endorsed concerns (scores 2+) among youth aged 13 - 17 years were depressive symptoms (3/8, 37.5%) and attention difficulties (3/8, 37.5%), followed by anger (2/8, 25%), anxiety (2/8, 25%), somatic complaints (2/8, 25%), sleep disturbances (2/8, 25%), and irritability (2/8, 25%). Among youth aged 18 years and older, the most frequently reported concerns (scores 2+) were impulsivity or increased energy (Mania scale, 10/32, 31.3%), anxiety (9/32, 28.1%), depressive symptoms (9/32, 28.1%), and somatic complaints (9/32, 28.1%). Youth who score a 2+ (or 1+ on SU) on the EWQ are contacted by the FostrSpace navigator to link to a FostrSpace clinician for intervention. Of the 14 youth who scored 2+ (or 1+ on substance use) on the EWQ, seven initiated services with a FostrSpace clinician; primary symptoms included substance use, anxiety, depression, anger or irritability, and sleep disturbance, and all endorsed recent substance use, specifically cannabis use (one endorsed alcohol and cannabis use).

Discussion

Principal Findings: Lessons Learned and Recommendations

Co-Designing BITs With Foster Youth

Engaging in participatory co-design of BITs with foster youth and other minoritized communities requires careful attention to creating a space where co-designers feel there is mutual benefit to engaging in the process, are psychologically safe to share their experiences, and that power dynamics are acknowledged and attended to. This type of participatory process takes time (ie, for our team, it took 18 months of weekly co-design meetings to develop the initial version of FostrSpace), which in academic settings can feel scarce as we push against deadlines and expectations set by institutions and funders [47].

XSL•FO
RenderX

Our team was fortunate to have a flexible funding source for this project, which afforded critically needed flexibility and time to complete project goals. Regardless of funding structure, it is important for academics engaging in this type of work not to rush through it, as this can risk recreating harms caused to foster youth (eg, paternalistic decision-making). Slowing down in the research process can reduce possible harms associated with asking youth to prioritize this over more pressing needs, for example, employment or educational responsibilities for financial security and education, and may require adapting timelines to meet youth where they are. Active reflection during and after the co-design process, by all members, can help identify areas for process improvement. Duoethnography—a qualitative research methodology in which two researchers engage in a dialogic, collaborative process to explore a topic through their personal experiences and perspectives—is one approach to such reflection [33,44]. Using a duoethnographic reflection process, the FAB and academic team who co-designed the initial version of FostrSpace developed a series of recommendations, which have been detailed elsewhere [44] for creative ways to share power and build community (eg, flattening power hierarchies and building community through icebreakers) and ensure mutual benefit (eg, providing professional development opportunities such as co-authoring manuscripts and delivering presentations) among co-designers. These strategies can be used to avoid replicating problematic patterns of researchers having relationships with minoritized communities that are based primarily in extraction.

In addition, when working with young people longitudinally, it is important to recognize that developmentally appropriate transitions (eg, changes in employment) and priorities (eg, juggling competing demands with school and family) may impact sustained engagement. Our team provided flexibility throughout the co-design process whenever possible, including through meeting remotely, rescheduling meetings as co-designers' schedules changed (including meeting in the evenings after school and work hours), and respecting natural transitions away from participation while adding new members to sustain the FAB. Such flexibility is particularly important when working with foster youth, as they may experience frequent placement changes and other forms of instability, which could impact participation.

Central to the co-creation of FostrSpace was foster youth raising in-app privacy and safety concerns. This necessitated co-designer discussions about how to practically protect youth privacy (eg, having a secure login requiring a unique username and password) and how to ensure foster youth actually understood their information was secure. The importance of transparency regarding how data are used was highlighted by the FAB, as co-designers agreed we all often click through long, technical app use agreements without truly reading the terms and conditions. In addition to a user agreement everyone sees during the initial registration process, the FAB created a "My Data" page in the app explaining who has access to the young person's data and for what purposes, how data are stored, and who to contact with questions. Information about who has access to data and why is also presented on specific pages within the app. For example, at the beginning of the EWQ, there is a statement, "Our trained FostrSpace Care Team will have access to your results to better understand your needs." To further promote autonomy, youth also have the option to skip questions on assessments like the EWQ, not engage with any app features they choose, and delete their FostrSpace account at any time. Without a true participatory co-design process, it is unlikely that an academic team or tech developer would have understood privacy and safety concerns in the same way that youth presented, and, for example, a "My Data" page would not likely have been a part of the app.

Technology Pitfalls: Addressing Barriers to BIT Engagement

In our efforts to create an app where foster youth privacy is protected, youth are required to create a unique username and password. This created an unanticipated barrier related to account verification. Specifically, the platform required youth to enter a personal email address upon registration, to which they received a confirmation link to complete registration. Several youth reported significant issues with the registration process throughout the various stages of testing and implementation of the app. While creating a password through a verification email is important for the purposes of maintaining data security in a platform containing sensitive information, this minor step in the registration process may be leading to a lack of completion of the registration process for foster youth. It is possible that different forms of account confirmation (eg, text confirmation code) could result in improved retention, which would be consistent with literature surrounding technology preferences among youth [48].

In addition to considering the technology usage patterns of the intended youth population when designing onboarding and app access workflows, we recommend seeking feedback to iteratively refine such processes, as they have the potential to hinder youth from accessing the BIT. Working closely with youth, developers, and technology partners to ensure efficient and effective registration workflows, and to iteratively refine them as needed, is critical to retaining youth in the digital space. Whereas an in-person clinic may have a welcoming administrative staff, peer support workers to help youth complete registration paperwork, or a supportive therapist explaining the pathways to accessing continued support, BITs are restricted to technical workflows where failsafe engagement is difficult to ensure. As the sophistication of artificial intelligence tools increases, teams may consider how these specific technologies could serve as digital assistants throughout the registration process, including by way of registration reminders for those who do not complete the entire flow [49].

Working With Third-Party Developers

The empirical literature on technology co-design offers some guidance in terms of how to consider the youth in the design process, though it is limited in the extent to which user-centered design methods are described in a way that is replicable. Some of our team members had received basic training in user-centered design methods; however, our collective BIT co-design experience was limited. Formal training in user-centered design methods may increase the extent to which research teams are able to design easy-to-use and effective

XSL•FO
RenderX

interventions. Partnering with third-party developers can also be beneficial for transforming the co-design team's vision into a functional and engaging app experience. We elected to partner with a third-party developer (Chorus Innovations, Inc) whose staff included user experience researchers and designers with formal training in this domain. Our co-design process actively involved a user experience researcher who received graduate training in user experience design and helped us to develop and facilitate user-centered design activities during the co-design meetings with the FAB. This partnership ultimately allowed us to create an app that attended to accessibility, functionality, and design.

Throughout this partnership, our team of foster youth, researchers, and developers collectively faced a number of growth areas. First, we had to develop a memorandum of understanding and scope of work to ensure the roles of the development team (ie, Chorus Innovations, Inc) were clearly outlined, including deliverables and invoicing expectations. Second, once beginning the work, we had to work together to develop a shared language; for example, what is a technical bug (eg, something not operating as it is intended to) versus a desired change (eg, something is operating as it is intended to but is not how we would like it to operate). Third, reconciling project management systems and preferences across our team and the developer was essential for tracking and communicating about issues within the app, new feature and content requests, and ensuring shared understanding of the expected timeframes for addressing milestones and unanticipated issues. We ultimately found that holding regular synchronous working meetings (at minimum weekly during the active design phase), having asynchronous communication primarily through Slack (Slack Technologies, LLC), and having shared files (through Google Drive) were necessary for communication, documenting progress on action steps, and collaborative problem-solving. At least one representative from the research and development team was present in co-design meetings with the FAB to ensure each team's expertise was represented and factored in when making decisions. Fourth, developers may not have experience working with the specific youth population, and when co-designing with foster youth and other minoritized communities, it is essential all parties understand basic ethical considerations. Researchers may consider facilitating basic training for developer partners regarding the needs of the population being served and ethical considerations for the co-design process prior to their interaction with the intended youth population. In our process, the academic and developer teams collectively reviewed plans for co-design meetings to ensure attention both to user-centered design methods and to the FAB's needs and psychological safety. While these plans will likely be iterative, we recommend that teams take time before beginning a co-design process to discuss these foundational items. When possible, maintaining a curious stance when approaching points of divergence can help identify when differing professional backgrounds and conventions may be the root of misunderstandings or inefficiencies. When appropriate, these trainings and discussions should be held with youth co-designers (eg, foster youth) to promote the establishment of a shared language based in cultural exchange and ensure that each member understands their roles, expectations, and communication.

If You Build It, Will They Come?

One of the main takeaways from the participatory co-design of FostrSpace was that simply co-designing and co-building the BIT was insufficient to drive youth engagement. Despite co-designing with youth who have lived experience in the foster care system, reaching out to 2300 contacts who directly serve foster youth, attending community events to present about Fostrspace (eg, college resource fairs), and posting flyers in public locations (eg, coffee shops or bus stops) to advertise FostrSpace, the total number of youth who have completed registration remains under 50 over a 31-month period of time. This indicates we have reached a limited portion of the roughly 19,000 eligible youth in the foster care system across the state of California [50]. This low uptake suggests a strategic, systems-embedded approach may prove more fruitful [51].

Embedding BITs within existing systems serving foster youth may increase the uptake and sustainability of BITs, particularly when using implementation science frameworks and integrating community engagement to guide dissemination and implementation [52,53]. Similar to other models of addressing service gaps, BITs can be embedded in existing systems to meet specific goals or gaps (eg, providing interim support for youth who are on long waitlists [54]), offered as adjunctive services, or advertised through sustained partnerships with systems serving foster youth [55,56]. Of note, our team is currently evaluating the effectiveness of working with CASA programs to increase FostrSpace usage by foster youth, aged 13-20 years [45]. CASAs are adult volunteers who work with youth involved in the child welfare system, typically one-to-one, in an ongoing, in-person, and long-term (average 18 months) capacity, to provide support and advocate for youth's access to necessary resources. Using approaches within specific systems or organizations may not reach youth who are not already connected to a serving or support system; thus, an engagement and outreach approach that applies multiple strategies across multiple systems and programs that serve foster youth may maximize reach and uptake.

In addition, despite our team having interdisciplinary experiences and training, no members had a formal background in marketing, and our advertising budget was limited. Our outreach efforts relied upon direct communication with foster youth and virtual and in-person events and professionals disseminating digital flyers to potentially eligible foster youth; we were not able to track whether our flyers were disseminated or to how many young people, and it is possible some contacts ignored our outreach attempts. Using other methods of advertising, such as social media, may be promising for reaching foster youth [57]. Researchers should also consider partnering with marketing experts or seeking training in this area themselves and budgeting to allow for wide-reaching advertising that does not predominately rely on middle-person approaches to reach youth.

XSL•FO
RenderX

Limitations

The current manuscript describes lessons learned from the process of creating FostrSpace, a BIT designed for and by foster youth. Considerations for interpreting our recommendations include that the work was conducted in a single United States state with a small FAB, which could impact generalizability; we used a web-based app platform, and challenges and opportunities may differ for other platforms such as native apps; and we did not conduct a formal evaluation of the app during the initial testing stages, but rather report available quality improvement data.

Conclusions

Lessons learned from the process of creating FostrSpace can inform researchers seeking to co-design BITs with foster youth, as well as young people more broadly. There is no one-size-fits-all approach to engaging in co-design, and as such, we encourage researchers to provide greater specificity regarding the co-design methods used so we can advance our scientific understanding of how best to engage in this type of work. Relevant across co-design methodologies, our experience highlights the importance of active reflection throughout the process to ensure the creation of a psychologically safe and constructive co-design space where young people, academics, and developers can work together to create accessible and impactful BITs. One element of this is ensuring adequate resource investment before beginning the process (eg, available compensation for co-designers and time commitment of all involved) to avoid replicating harmful research practices of extracting from minoritized communities, as co-designing BITs can take significant time and resources. This time-intensive participatory approach can run counter to deadlines and expectations set by institutions and funders [47], and broader advocacy for the centrality of these methods to advancing health equity is needed to ensure their feasibility within existing structures.

Funding

FostrSpace was funded by an anonymous private donor and the Health and Human Services Children and Youth Resilience Challenge. Time spent on this paper was funded in part by grants K24DA046569 (principal investigator: MT-S), R61DA059860 (principal investigator: MT-S), R33DA059860 (principal ivestigator: MT-S), K23DA050798 (principal investigator: JBF), and T32DA007250 (principal investigator: DS, JT). The authors would like to Chorus Innovations, Inc. for partnership related to technology development and the wider FostrSpace team for their contributions in developing FostrSpace.

Authors' Contributions

JF, JG, and MT-S conceptualized the paper and wrote the original draft. MDC, EM, AC, and TL conducted data analysis and interpretation. MT-S acquired the financial support for this project leading to this publication and designed the study methodology. Project administration was led by JF, MD, and AC. Supervision was provided by MT-S and LF. All authors were involved in writing, reviewing, and editing the manuscript.

Conflicts of Interest

None declared.

References

1. Administration on Children, Youth, And Families. The AFCARS report. : U.S. Department of Health and Human Services; 2021 URL: https://www.acf.hhs.gov/sites/default/files/documents/cb/afcarsreport28.pdf [accessed 2025-11-19]
2. Administration on Children, Youth, And Families. Child maltreatment 2016. : U.S. Department of Health and Human Services; 2018 URL: https://acf.gov/sites/default/files/documents/cb/cm2018.pdf [accessed 2025-11-19]
3. Administration on Children, Youth, And Families. The AFCARS report. : U.S. department of health and human services; 2017 URL: https://acf.gov/sites/default/files/documents/cb/afcarsreport24.pdf [accessed 2025-11-19]
4. Engler AD, Sarpong KO, Van Horne BS, Greeley CS, Keefe RJ. A systematic review of mental health disorders of children in foster arec. Trauma Violence Abuse 2022 Jan;23(1):255-264. [doi: 10.1177/1524838020941197]
5. Beal SJ, Greiner MV, Ammerman RT, et al. Patterns of substance use among adolescents in and out of foster care: an analysis of linked health and child welfare administrative data. Child Abuse Negl 2023 Dec;146:106473. [doi: 10.1016/j.chiabu.2023.106473]
6. Burns BJ, Phillips SD, Wagner HR, et al. Mental health need and access to mental health services by youths involved with child welfare: a national survey. J Am Acad Child Adolesc Psychiatry 2004 Aug;43(8):960-970. [doi: 10.1097/01.chi.0000127590.95585.65] [Medline: 15266190]
7. Ruff SC, Harrison K. “Ask Me What Want”: community-based participatory research to explore transition-age foster youth’s use of support services. Child Youth Serv Rev 2020 Jan;108:104608. [doi: 10.1016/j.childyouth.2019.104608]
8. Gee GC, Ford CL. Structural racism and health inequities: old issues, new directions. Du Bois Review: Social Science Research on Race 2011;8(1):115-132. [doi: 10.1017/S1742058X11000130]
9. Hatcher DL. Commodified inequality: racialized harm to children and families in the injustice enterprise. Fam Court Rev 2023 Apr;61(2):341-358 [FREE Full text] [doi: 10.1111/fcre.12709]

XSL•FO
RenderX

10. Murphy AL, Boamah DA, Warfel ET, Griffiths A, Roehm D. Ending racial disproportionality in child welfare: a systematic review. Child Soc 2024 Jan;38(1):214-227 [FREE Full text] [doi: 10.1111/chso.12721]
11. Wulczyn F, Zhou X, McClanahan J, et al. Race, poverty, and foster care placement in the United States: longitudinal and cross-sectional perspectives. Int J Environ Res Public Health 2023 Aug 13;20(16):6572. [doi: 10.3390/ijerph20166572] [Medline: 37623160]
12. Dettlaff AJ, Boyd R. Racial disproportionality and disparities in the child welfare system: why do they exist, and what can be done to address them? Ann Am Acad Pol Soc Sci 2020 Nov;692(1):253-274. [doi: 10.1177/0002716220980329]
13. Social determinants of health. World Health Organization. URL: https://www.who.int/health-topics/social-determinants-of-health [accessed 2025-01-30]
14. Yampolskaya S, Chuang E, Walker C. Trajectories of substance use among child welfare-involved youth: longitudinal associations with child maltreatment history and emotional/behavior problems. Subst Use Misuse 2019;54(3):437-448. [doi: 10.1080/10826084.2018.1504079] [Medline: 30658541]
15. Kothari BH, Godlewski B, Lipscomb ST, Jaramillo J. Educational resilience among youth in foster care. Psychol Sch 2021 May;58(5):913-934 [FREE Full text] [doi: 10.1002/pits.22478]
16. Zetlin A. The experiences of foster children and youth in special education. J Intellect Dev Disabil 2006 Sep;31(3):161-165. [doi: 10.1080/13668250600847039] [Medline: 16954094]
17. Yearby R. Structural racism and health disparities: reconfiguring the social determinants of health framework to include the root cause. J Law Med Ethics 2020 Sep;48(3):518-526. [doi: 10.1177/1073110520958876] [Medline: 33021164]
18. Pullmann MD, Jacobson J, Parker E, et al. Tracing the pathway from mental health screening to services for children and youth in foster care. Child Youth Serv Rev 2018 Jun;89:340-354. [doi: 10.1016/j.childyouth.2018.04.038]
19. Leslie LK, Raghavan R, Zhang J, Aarons GA. Rates of psychotropic medication use over time among youth in child welfare/child protective services. J Child Adolesc Psychopharmacol 2010 Apr;20(2):135-143. [doi: 10.1089/cap.2009.0065] [Medline: 20415609]
20. Barnett ER, Concepcion-Zayas MT, Zisman-Ilani Y, Bellonci C. Patient-centered psychiatric care for youth in foster care: a systematic and critical review. J Public Child Welf 2019 Aug 8;13(4):462-489. [doi: 10.1080/15548732.2018.1512933]
21. Leo HP, Folk JB, Rodriguez C, Tolou-Shams M. Implementation considerations for family-based telehealth interventions for youth in foster care: focus group study with child welfare system professionals. JMIR Form Res 2023 Dec 29;7:e45905. [doi: 10.2196/45905] [Medline: 38157238]
22. Willis HA, Gonzalez JC, Call CC, Quezada D, Galán CA. Culturally responsive telepsychology & mHealth interventions for racial-ethnic minoritized youth: research gaps and future directions. J Clin Child Adolesc Psychol 2022 Nov 2;51(6):1053-1069. [doi: 10.1080/15374416.2022.2124516] [Medline: 36227174]
23. Ramos G, Chavira DA. Use of Technology to Provide Mental Health Care for Racial and Ethnic Minorities: Evidence, Promise, and Challenges. 2019. URL: https://www.sciencedirect.com/science/article/abs/pii/S1077722919301063 [accessed 2025-11-19]
24. Braciszewski JM, Colby SM, Franklin MJ, Stout RL, Vose-ONeal A, Bock BC. Technology-based smoking cessation for youth exiting foster care: a pilot randomized trial. J Public Child Welf 2024 Jan;18(1):61-79. [doi: 10.1080/15548732.2022.2155754] [Medline: 38144460]
25. Folk JB, Valencia-Ayala C, Holloway ED, Anvar S, Czopp A, Tolou-Shams M. Feasibility and acceptability of a family-based telehealth intervention for families impacted by the child welfare system: formative mixed methods evaluation. JMIR Form Res 2024 Oct 15;8:e57939. [doi: 10.2196/57939] [Medline: 39405104]
26. Hermes ED, Lyon AR, Schueller SM, Glass JE. Measuring the implementation of behavioral intervention technologies: recharacterization of established outcomes. J Med Internet Res 2019 Jan 25;21(1):e11752. [doi: 10.2196/11752] [Medline: 30681966]
27. Haggerty KP, Barkan SE, Caouette JD, Skinner ML, Hanson KG. Two-year risk behavior outcomes from connecting, a prevention program for caregivers and youth in foster care. Prev Sci 2023 Jan;24(1):15-26. [doi: 10.1007/s11121-022-01390-4] [Medline: 35788868]
28. Kaveladze BT, Wasil AR, Bunyi JB, Ramirez V, Schueller SM. User experience, engagement, and popularity in mental health apps: secondary analysis of app analytics and expert app reviews. JMIR Hum Factors 2022 Jan 31;9(1):e30766. [doi: 10.2196/30766] [Medline: 35099398]
29. Tolou-Shams M, Holloway ED, Ordorica C, et al. Leveraging technology to increase behavioral health services access for youth in the juvenile justice and child welfare systems: a cross-systems collaboration model. J Behav Health Serv Res 2022;49(4):422-435. [doi: 10.1007/S11414-022-09808-1/TABLES/1]
30. Ramos G, Hernandez-Ramos R, Taylor M, Schueller SM. State of the science: using digital mental health interventions to extend the impact of psychological services. Behav Ther 2024 Nov;55(6):1364-1379. [doi: 10.1016/j.beth.2024.04.004] [Medline: 39443071]
31. Fleming T, Bavin L, Lucassen M, Stasiak K, Hopkins S, Merry S. Beyond the trial: systematic review of real-world uptake and engagement with digital self-help interventions for depression, low mood, or anxiety. J Med Internet Res 2018 Jun 6;20(6):e199. [doi: 10.2196/jmir.9275] [Medline: 29875089]

XSL•FO
RenderX

32. Bevan Jones R, Stallard P, Agha SS, et al. Practitioner review: co‐design of digital mental health technologies with children and young people. Child Psychology Psychiatry 2020 Aug;61(8):928-940 [FREE Full text] [doi: 10.1111/jcpp.13258]
33. Porche MV, Folk JB, Tolou-Shams M, Fortuna LR. Researchers' perspectives on digital mental health intervention co-design with marginalized community stakeholder youth and families. Front Psychiatry 2022;13:867460. [doi: 10.3389/fpsyt.2022.867460] [Medline: 35530032]
34. Hodson E, Dadashi N, Delgado R, Chisholm C, Sgrignoli R, Swaine R. Co-design in mental health; mellow: a self-help holistic crisis planning mobile application by youth, for youth. The Design Journal 2019 Apr 1;22(sup1):1529-1542. [doi: 10.1080/14606925.2019.1594975]
35. Quyoum A, Wong M. Valuing lived experience and co-design solutions to counter racial inequality in data and algorithmic systems in UK's digital services. Inf Commun Soc 2024 Jul 3;27(9):1793-1809. [doi: 10.1080/1369118X.2024.2331781]
36. Liang K, Niland A, Little C. Hearing young children's perspectives: reflections on research into left-behind children's lived experience in Chinese rural areas. Global Studies of Childhood 2023 Jun;13(2):178-191. [doi: 10.1177/20436106231181134]
37. King PT, Cormack D. Indigenous peoples, whiteness, and the coloniality of co-design. In: Ravulo J, Olcoń K, Dune T, Workman A, Liamputtong P, editors. Handbook of Critical Whiteness: Springer; 2023:1-16. [doi: 10.1007/978-981-19-1612-0_28-1]
38. Arevian AC, O'Hora J, Jones F, et al. Participatory technology development to enhance community resilience. Ethn Dis 2018;28(Suppl 2):493-502. [doi: 10.18865/ed.28.S2.493] [Medline: 30202203]
39. Scariot CA, Heemann A, Padovani S. Understanding the collaborative-participatory design. Work 2012;41 Suppl 1:2701-2705. [doi: 10.3233/WOR-2012-0656-2701] [Medline: 22317129]
40. Chung B, Jones L, Dixon EL, Miranda J, Wells K, Community Partners in Care Steering Council. Using a community partnered participatory research approach to implement a randomized controlled trial: planning community partners in care. J Health Care Poor Underserved 2010 Aug;21(3):780-795. [doi: 10.1353/hpu.0.0345] [Medline: 20693725]
41. McCreary M, Arevian AC, Brady M, et al. A clinical care monitoring and data collection tool (H3 Tracker) to assess uptake and engagement in mental health care services in a community-based pediatric integrated care model: longitudinal cohort study. JMIR Ment Health 2019 Apr 23;6(4):e12358. [doi: 10.2196/12358] [Medline: 31012861]
42. Goodsmith N, Moore EM, Siddiq H, et al. Community-partnered development of a digital mental health resource website to support diverse communities during the COVID-19 pandemic. J Health Care Poor Underserved 2022;33(1):506-516. [doi: 10.1353/hpu.2022.0039] [Medline: 35153238]
43. Al-Mondhiry J, D'Ambruoso S, Pietras C, et al. Co-created mobile apps for palliative care using community-partnered participatory research: development and usability study. JMIR Form Res 2022 Jun 23;6(6):e33849. [doi: 10.2196/33849] [Medline: 35737441]
44. Ezimora I, Lundberg T, Miars D, et al. Reflections of foster youth engaging in the co-design of digital mental health technology: duoethnography study. JMIR Form Res 2025 Jan 20;9(1):e53231. [doi: 10.2196/53231] [Medline: 39832159]
45. Tolou-Shams M, Folk JB, Del Cid MV, et al. Advancing access to substance use prevention for foster youth through digital innovation: an open trial of fostrspace with court appointed special advocate programs. BMC Health Serv Res 2025 May 10;25(1):676. [doi: 10.1186/s12913-025-12811-9] [Medline: 40346641]
46. American Psychiatric Association. DSM-5-TR self-rated level 1 cross-cutting symptom measure - child age 11-17. 2013.
47. Giusto A, Triplett NS, Foster JC, Gee DG. Future directions for community-engaged research in clinical psychological science with youth. J Clin Child Adolesc Psychol 2024;53(3):503-522. [doi: 10.1080/15374416.2024.2359650] [Medline: 38830059]
48. Rempel GR, Ballantyne RT, Magill-Evans J, Nicholas DB, Mackie AS. Texting teens in transition: the use of text messages in clinical intervention research. JMIR Mhealth Uhealth 2014 Nov 6;2(4):e45. [doi: 10.2196/mhealth.3232] [Medline: 25379624]
49. Giovanelli A, Rowe J, Taylor M, et al. Supporting Adolescent Engagement with Artificial Intelligence-Driven Digital Health Behavior Change Interventions. J Med Internet Res 2023 May 24;25(1):e40306. [doi: 10.2196/40306] [Medline: 37223987]
50. Webster D, Lee S, Dawson W, et al. Point in Time/In Care. URL: https://ccwip.berkeley.edu/childwelfare/reports/PIT/MTSG/r/ab636/l [accessed 2025-02-28]
51. Jolles MP, Fort MP, Glasgow RE. Aligning the planning, development, and implementation of complex interventions to local contexts with an equity focus: application of the PRISM/RE-AIM Framework. Int J Equity Health 2024 Feb 27;23(1):41. [doi: 10.1186/s12939-024-02130-6] [Medline: 38408990]
52. Gunn CM, Sprague Martinez LS, Battaglia TA, et al. Integrating community engagement with implementation science to advance the measurement of translational science. J Clin Transl Sci 2022;6(1):e107. [doi: 10.1017/cts.2022.433] [Medline: 36285013]
53. Blachman-Demner DR, Wiley TRA, Chambers DA. Fostering integrated approaches to dissemination and implementation and community engaged research. Transl Behav Med 2017 Sep;7(3):543-546. [doi: 10.1007/s13142-017-0527-8] [Medline: 28936762]
54. Sung JY, Bugatti M, Vivian D, Schleider JL. Evaluating a telehealth single-session consultation service for clients on psychotherapy wait-lists. PsyArXiv. Preprint posted online on 2023. [doi: 10.31234/osf.io/k7u4r]

XSL•FO
RenderX

55. Forehand GL, Schisler-Blizzard AD, Deaver AH, et al. Triangulating perspectives to inform the development of a smartphone application for foster, kinship, and adoptive parents. J Technol Hum Serv 2019 Oct 2;37(4):362-394. [doi: 10.1080/15228835.2019.1666781]
56. Salazar AM, Spiers SS, Thompson J, Podsobinski J, Day M. Implementing and evaluating manualized interventions in U.S. organizations serving youth with foster care experience: lessons learned from the field. Child Adolesc Soc Work J 2025 Dec;42(6):765-781. [doi: 10.1007/s10560-024-00989-6]
57. Gabrielli J, Borodovsky J, Corcoran E, Sink L. Leveraging social media to rapidly recruit a sample of young adults aging out of foster care: methods and recommendations. Child Youth Serv Rev 2020 Jun;113:104960. [doi: 10.1016/j.childyouth.2020.104960]

Abbreviations

BIT: Behavioral intervention technology
CASA: Court Appointed Special Advocate
***DSM-5*:** *Diagnostic and Statistical Manual of Disorders, Fifth Edition*
EWQ: Emotional wellness questionnaire
FAB: FostrSpace Advisory Board
HIPAA: Health Insurance Portability and Accountability Act
SDoH: Social determinants of health

Edited by J Sarvestan; submitted 28.02.25; peer-reviewed by B Imbert, M Porche; revised version received 14.10.25; accepted 15.10.25; published 09.12.25.

Please cite as:
Folk JB, Gonzalez JC, Del Cid MV, McBride E, Lundberg T, Czopp A, Ezimora I, Fortuna LR, Tolou-Shams M
Lessons Learned from Co-Designing a Digital Health App for Foster Youth: Development and Usability Study
JMIR Form Res 2025;9:e73281
URL: https://formative.jmir.org/2025/1/e73281
doi:10.2196/73281

© Johanna B Folk, Juan Carlos Gonzalez, Margareth V Del Cid, Elizabeth McBride, Tylia Lundberg, Alison Czopp, Ifunanya Ezimora, Lisa R Fortuna, Marina Tolou-Shams. Originally published in JMIR Formative Research (https://formative.jmir.org), 9.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Comparing Human and AI Therapists in Behavioral Activation for Depression: Cross-Sectional Questionnaire Study

Inka Napiwotzki[1*], MSc; Julian Laue[2*], MSc; Flora Caldarone[1], MSc; Maximilian Idahl[2], MSc; Uwe Hadler[2], MSc; Haithem Amrani[2], MSc; Elisabeth Hildt[3], Prof Dr; Kai G Kahl[1], Prof Dr Med; Wolfgang Nejdl[2], Prof Dr

[1]Clinic for Psychiatry, Social Psychiatry and Psychotherapy, Department of Medizinische Hochschule Hannover, Carl-Neuberg-Str 1, Hannover, Germany

[2]L3S Research Center, Leibniz University Hannover, Hannover, Germany

[3]Department of Humanities, Arts, and Social Sciences, Illinois Institute of Technology, Chicago, IL, United States

[*]these authors contributed equally

Corresponding Author:
Inka Napiwotzki, MSc
Clinic for Psychiatry, Social Psychiatry and Psychotherapy, Department of Medizinische Hochschule Hannover, Carl-Neuberg-Str 1, Hannover, Germany

Abstract

Background: Large language models (LLMs) have rapidly advanced across numerous fields, including mental health care. A shortage of trained therapists and mental health care providers has driven informal use of LLMs for therapeutic support. However, their clinical utility remains poorly defined.

Objective: This study aimed to systematically evaluate and compare the therapeutic knowledge and single-turn response capabilities of LLMs versus psychotherapists in training in the context of behavioral activation (BA) therapy for depression, and to assess how both groups' performance changed when provided with structured therapeutic training materials.

Methods: Six LLMs and 8 human participants completed a questionnaire on depression and BA with 20 multiple-choice items and 10 therapy scenarios, each with 3 open-ended items, that postulated empathic response, use of validation strategies, and theory of mind capabilities. Human participants completed the questionnaire before and after a 5-hour workshop and 5-week period with learning materials. The LLMs received identical training content as context during the second test. All open-ended questions were rated on 5-point scales by 2 experts.

Results: At baseline, the LLMs demonstrated higher knowledge scores than human participants (61.0 vs 52.0 out of 100 points) and were rated higher in empathy (U=2.0; P=.005; r=0.917), validation quality (U=2.5; P=.006; r=0.896), anticipation of cognition (U=0.0; P=.002; r=1.000), and anticipation of emotion (U=0.0; P=.002; r=1.000). Following BA training, the LLMs maintained their performance advantage across multiple-choice and open-ended items.

Conclusions: The results suggest that LLMs may generate high-quality therapeutic single-turn responses that integrate clinical knowledge with empathetic communication. The findings hint at LLMs' potential as valuable tools in mental health care, although further clinical trials are needed to evaluate their performance in ongoing therapeutic relationships and clinical outcomes.

(JMIR Form Res 2025;9:e78138) doi:10.2196/78138

KEYWORDS

large language model; mental health; behavioral activation; empathy; digital therapeutics; artificial intelligence; AI

Introduction

Health care systems worldwide face a critical challenge in addressing a growing mental health crisis: while evidence-based treatments like cognitive behavioral therapy (CBT) exist, there is an acute shortage of trained professionals to deliver them. This gap affects 1 in 8 people globally who live with mental disorders, with numbers rising since the COVID-19 pandemic. Mental disorders have devastating consequences. Beyond reduced work productivity, affected individuals experience reduced social participation, physical health complications, and premature mortality [1-4].

Recent research has demonstrated the potential of large language models (LLMs) in mental health care applications [5,6]. While LLMs offer more sophisticated and natural language understanding capabilities than earlier rule-based systems, their practical implementation in therapeutic contexts remains largely unexplored [7]. However, informal therapeutic use of LLMs is already occurring. A recent study of the Replika chatbot (Luka, Inc) found users engaging in therapeutic conversations, with some reporting crisis prevention benefits [8]. These findings

XSL•FO
RenderX

align with informal user discussions across social media platforms, where individuals frequently describe using general-purpose LLM platforms like ChatGPT (OpenAI) for emotional support and mental health conversations, despite these models not being designed or validated for therapeutic use (Mirzae, T, unpublished data, October 2025). This spontaneous adoption, combined with LLMs' known risks and susceptibility to errors, underscores the critical and urgent need for rigorous evaluation to ensure their safe and effective application in therapeutic dialogue [9,10].

Researchers have explored various approaches to enhance LLMs' therapeutic capabilities—from fine-tuning models on therapy-specific datasets to applying few-shot learning with therapist-client examples and adapting self-critique techniques [11-13]. However, these studies predominantly relied on automated evaluation methods, often using one LLM to evaluate another [12,14]. This methodological limitation points to the need for comprehensive human expert assessment.

To address these limitations, we present a systematic evaluation comparing 6 LLMs with 8 psychotherapists in training. Our assessment consists of 2 components. Multiple-choice questions test knowledge on depression, therapy principles, and BA, an effective therapeutic method within CBT for treating depression [15,16]. We focused on depression as it is one of the most common mental disorders [1]. Through open-ended questions, we evaluated responses to client statements, assessing empathy, use of validation strategies, and the ability to anticipate a client's emotions and cognition. We evaluated how single-turn performance changes when LLMs are provided with therapeutic background information about BA principles and techniques, comparing this to the improvement observed in therapists after formal BA training. Figure 1 provides a visual overview of our approach. This parallel assessment reveals whether additional context enhances LLM capabilities and quantifies learning effects in human therapists.

Figure 1. Experimental design for comparing psychotherapists in training with large language models (LLMs). Study methodology comparing psychotherapists in training (n=8, top) and LLMs (n=6, bottom) on therapeutic capabilities. Both groups completed identical assessments with multiple-choice knowledge questions and open-ended therapeutic scenarios. Human participants underwent a 5-week learning period including a BA workshop, while LLMs received equivalent information as added context during the posttest assessment. BA: behavioral activation; CBT: cognitive behavioral therapy.

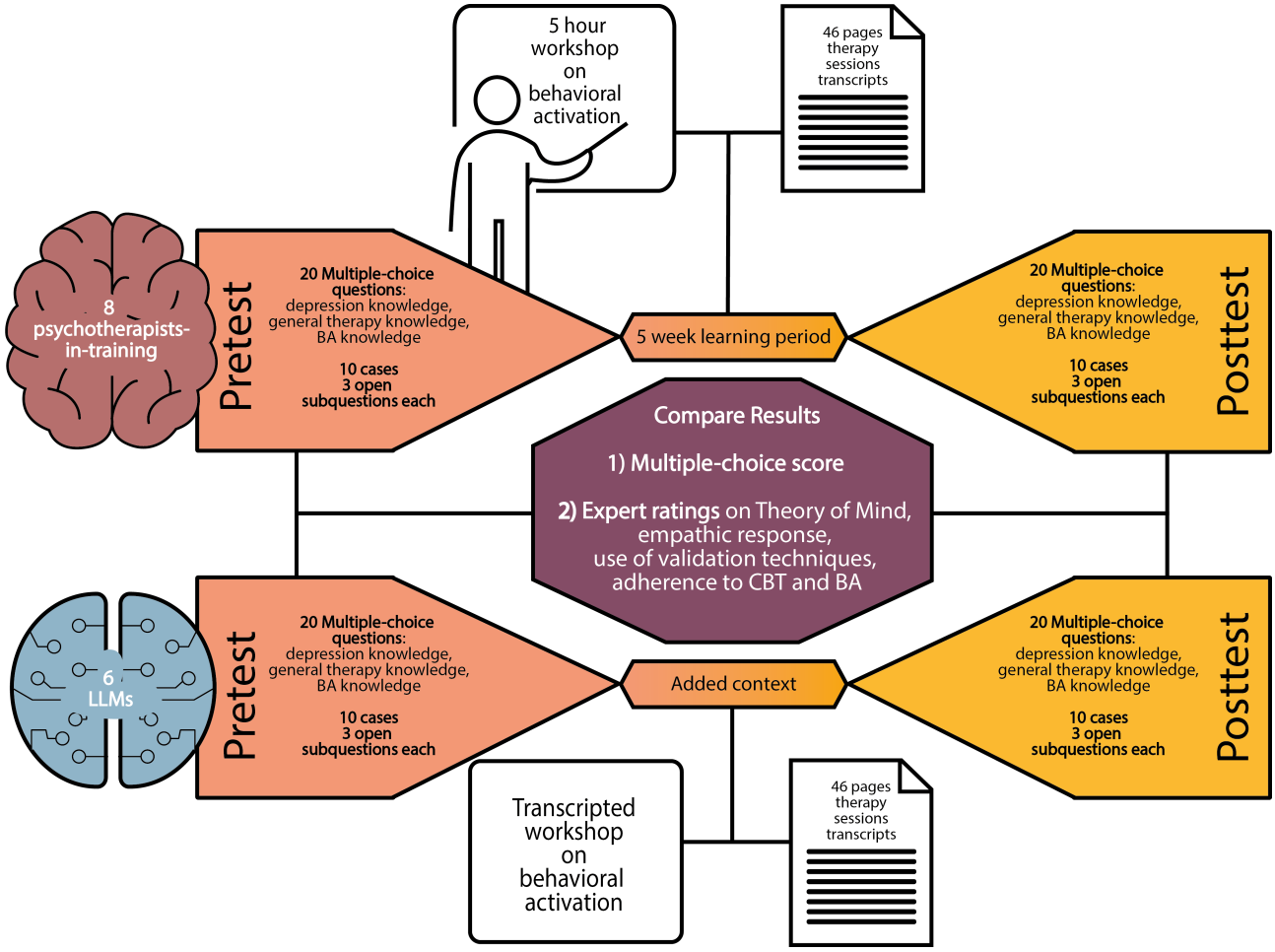

In summary, our contributions are threefold: (1) a comprehensive evaluation questionnaire combining 20 multiple-choice and 10 open-ended questions that will be publicly available to foster further research in this domain; (2) an evaluation of LLMs in their default, publicly accessible form—the way most users currently interact with these systems; and (3) the first direct comparison (to the best of our knowledge) between LLMs and human psychotherapists in training in both therapeutic knowledge and response capabilities.

XSL•FO
RenderX

Methods

Participants

Participants were recruited through the mailing list of the Clinic for Psychiatry, Social Psychiatry, and Psychotherapy, as well as the Educational Institute for Psychotherapy. In total, the mailing list consisted of 102 psychotherapists in training; 10 people answered, of whom 2 dropped out of participation because they could not participate on the day of the workshop. Thus, the human sample consisted of 8 psychotherapists in training from Hannover Medical School. Six participants were in their second year of training, 1 in the first year, and 1 in the third year. At the time of this study, 7 participants were actively conducting therapy sessions with outpatients. The sample included 7 female and 1 male participant, with age ranging from 25 to 32 (mean 27.88, SD 2.2) years. At the beginning of the questionnaire, participants were asked to rate their prior knowledge of BA on a scale of 1 to 5. They rated themselves a mean 3 (SD 1.87). The LLM sample consisted of 6 different LLMs. We investigated the most well-known LLMs at the time of conducting the study, namely GPT-4 (OpenAI), GPT-4o (OpenAI), Gemini Pro 1.5 (Google), Claude Opus (Anthropic), Llama-3 70B Instruct (Meta), and Command R+ (Cohere) [17-21]. All models were accessed via OpenRouter[22] to enable a unified API interface.

Ethical Considerations

This study was reviewed and approved by the Research Ethics Committee at the Medical University Hanover (OE 9515; approved on December 18, 2023). All participants were adults (≥18 years). Written informed consent was obtained from all participants after the aims, procedures, risks, and benefits were explained. Participants could withdraw at any time without penalty. We safeguarded privacy and confidentiality by pseudonymization, secure servers, and restricted access. No identifying information is reported. Participants received credit points for attending the workshop and 250 EUR (US $271.30) compensation. The study adhered to the Declaration of Helsinki and applicable local and national regulations.

Variables

In Germany, psychotherapy is conducted by professionals who are trained in scientifically established methods to diagnose, heal, or mitigate disorders with pathological significance [23]. We assessed knowledge about depressive disorders, BA, and CBT with and without contextual information to evaluate baseline knowledge and learning capacity in both human therapists and LLMs.

The client-therapist working alliance represents a critical factor in symptom reduction. Essential therapist characteristics, particularly empathy and theory of mind capabilities—the ability to anticipate and comprehend a patient's emotional and cognitive states— substantially influence the development of a therapeutic alliance [24-27]. In exploring these capabilities, Rogers [28] conceptualized psychotherapeutic empathy as a dual process encompassing both client understanding and the effective communication of this understanding, highlighting its cognitive and emotional components.

Validation strategies are fundamental communicative and alliance-building tools for affirming client experiences, communicating empathy, and fostering therapeutic engagement [29-33]. Our evaluation framework incorporates 6 validation strategies from dialectic behavioral therapy: attentiveness (actively listening and focusing on the client's words and emotions), intermodal communication (rephrasing within the same modality, eg, "I was very sad" → "So, you were affected emotionally"), crossmodal communication (rephrasing across different modalities), biographical reference (connecting current responses to personal history), present-moment validation (affirming the appropriateness of current feelings or actions), and radical genuineness (sharing personal reflections, eg, "I would have felt the same") [34]. These strategies provided structured criteria for evaluating validation strategy choice in the open-ended questions, assessing empathic quality, validation effectiveness, and accuracy in estimating fictional patients' cognitive and emotional states.

Questionnaire

We designed a pen-and-paper test in German consisting of 20 multiple-choice questions to assess knowledge on depression, practical knowledge on psychotherapy, and specific knowledge on BA. The overall score was calculated by awarding 1 point for each correctly selected or omitted answer and −1 point for incorrect selections and omissions. The maximum attainable score was 100 points for all questions, 50 points for knowledge on depression, 10 for practical knowledge, and 40 for specific knowledge on BA.

The second section entailed 10 case scenarios of psychotherapy sessions with 3 open-ended subquestions each. In the case scenarios, a fictional patient showed a reaction to a certain situation (eg, crying after a difficult therapeutic task could not be completed). First, participants were asked which emotions and cognitions they would anticipate in the patient and provide a rationale for their conclusions. In the second subquestion, the participants assumed the role of the therapist and had to outline how they would continue with the therapy. Thirdly, they were asked to formulate an answer to the patient. Responses were rated on 5 dimensions: theory of mind, empathic response, adherence to psychotherapy principles, adherence to BA techniques, and fidelity of the chosen validation strategies. The 5 rating dimensions comprised 87 subquestions in total. With 14 participants, each blinded expert evaluated 1218 individual items (87 subquestions × 14 participants).

Translated sample items from the open-ended case scenarios are provided in (Tables S1-S4 in Multimedia Appendix 1). The complete questionnaire, including all items in both German and English translation, is available in a public repository [35].

Procedure

Human participants completed the initial questionnaire on paper (90 min) before attending a 5-hour training session on BA. The training covered BA therapeutic rationale and methods, with participants receiving handouts containing session materials and fictional therapy transcripts. These transcripts illustrated potential client-therapist challenges and provided supplementary BA information. Participants had 5 weeks to study the materials

XSL•FO
RenderX

before completing the questionnaire again. The testing period was April 26 to May 31, 2024.

For the LLM evaluation, we accessed all models through OpenRouter with temperature set to 0 to ensure reproducibility, while maintaining all other model parameters at their default values. Each interaction began with the system message "Du bist ein Experte im Bereich Psychotherapie" ("You are an expert in psychotherapy"), which remained identical for pretest and posttest assessments across all models. For multiple-choice questions, models were instructed to list only correct answer options. For the case scenarios, we used a chat-based format where each model's previous responses were preserved as distinct messages in the conversation history, rather than concatenating them into a single prompt. This meant that for each new question, the model had access to the full conversation history including its previous responses within the same case. All LLM responses were generated between April and May 2024. To ensure standardized formatting for the expert review, we manually transferred the model outputs to a Microsoft Word document, correcting any formatting inconsistencies while preserving the original content. This process created a uniform presentation format for the experts' blind evaluation.

The same training materials provided to human participants, including session transcripts and additional BA information, were input as context for the LLMs' posttest evaluation. Due to Llama 3.0's limited context window of 8k tokens, it was the only model for which only a condensed summary of the BA training materials was provided. All pretest and posttest responses from both humans and LLMs were randomly aggregated and blindly rated by 2 licensed psychotherapists with expertise in BA. The raters worked independently and were blind to each other's ratings.

Analysis

Adherence, empathy, and goodness of fit of anticipated emotions and cognitions were rated on a scale of 1 (= "not complied") to 5 (= "completely complied"). The mean was calculated for each group (human vs LLM) for the pretest and the posttest over all 10 questions. The theory of mind score is the mean of goodness of fit of anticipated emotions and cognitions. Six validation strategies were predefined from dialectic behavioral therapy, and the experts identified which strategies were applied in the answers. Quality of Validation was rated on a scale of 1 (= "insufficient") to 5 (= "optimal") and total number of applied strategies was counted for each group (human vs LLM). To ensure interrater reliability, the 2 experts jointly developed and agreed upon definitions for each rating level before beginning the evaluation. Statistical analyses used Wilcoxon signed-rank tests for within-group comparisons and Mann-Whitney *U* tests for between-group comparisons with a Bonferroni correction for multiple testing. The rank biserial correlation was used as an effect-size measure for both tests. The intraclass correlation coefficient was calculated to assess agreement between the 2 raters. This was intentional to preserve independent judgment.

Results

Table 1 presents multiple-choice test scores for LLMs (n=6) and human psychotherapists in training (n=8) across 3 knowledge domains: depression, general therapy, and BA knowledge. At baseline, both groups demonstrated moderate domain knowledge, with LLMs showing numerically higher mean total scores (61.0, SD 7.46 vs 52.0, SD 15.43 points, U=16.5; P=.37; r=0.313).

Table . Mean multiple-choice assessment scores of psychotherapists in training and large language models (LLMs).

Category (score range)	Pretest (human), mean (SD)	Posttest (human), mean (SD)	Pretest (LLM), mean (SD)	Posttest (LLM), mean (SD)
Depression (0 - 50)	20 (9.43)	22 (9.64)	24.33 (1.8)[a]	21.67 (3.35)
General knowledge (0 - 10)	7 (2.24)	8 (2)	8.33 (1.37)	10 (0)[a]
Behavioral activation (0 - 40)	25 (6.24)	24.75 (7.93)	28.33 (6.47)	32.67 (5.62)[a]
Total (0 - 100)	52 (15.43)	54.75 (16.31)	61 (7.46)	64.33 (8.9)[a]

[a]Values indicate highest scores in each category. There were no statistically significant differences between groups or between pretest and posttest scores at P<.00625.

Following BA training, both groups demonstrated improvements under their respective testing conditions. Human participants, tested after a 5-week study period, showed a modest increase in mean total scores (52.0, SD 15.43 vs 54.75, SD 16.31 points; W=16.0; P=.84; r=0.111) with gains in depression knowledge (20.0 vs 22.0 points; W=9.0; P=.40; r=0.500). LLMs, evaluated with training materials as additional context, similarly showed enhanced performance with total scores improving (61.0, SD 7.46 vs 64.33, SD 8.9 points; W=5.0; P=.49; r=0.524), attained maximum scores in general therapy knowledge (8.3, SD 1.37 vs 10.0, SD 0.0 points; W=0; P=.06; r=1.000), and showed improvement in BA knowledge (28.33, SD 6.47 vs 32.67, SD 5.62 points; W=1.0; P=.08; r=0.905), though their depression knowledge showed a slight decrease (24.33, SD 1.8 vs 21.67, SD 3.35 points; W=4.0; P=.34; r=−0.619). Notably, while LLMs improved markedly in BA knowledge, human performance in this domain remained stable (25.0, SD 6.24 vs 24.75, SD 7.93 points: W=15.0; P=.74; r=−0.167). We observed no significant difference between pretest and posttest scores within either group. All statistical comparisons used a significance threshold of P<.00625.

Figure 2 shows multiple-choice pretest and posttest scores for the 6 LLMs. Proprietary models demonstrated higher

XSL•FO
RenderX

performance both before and after integrating additional context, with mean scores improving from 63.0 (SD 7.68) to 70.5 (SD 1.61) points, while open-source models declined from 57.0 (SD 5) to 52.0 (SD 2) points. The limited sample size of 4 proprietary and 2 open-source models restricts formal statistical analysis, as the Mann-Whitney *U* test would yield a minimum *P*=.13, and the Wilcoxon signed-rank tests within groups would lead to a minimum 2-sided *P* value of .25 (proprietary) and .50 (open source). Both exceed conventional significance thresholds. Nonetheless, our data reveal distinct performance patterns among the proprietary models (solid bars). GPT-4 and GPT-4o improved from 66.0 to 72.0 points, Gemini Pro 1.5 showed the largest gain (50.0 vs 68.0 points), and Claude Opus remained at 70.0 points, all converging to 68 to 72 points at posttest. In contrast, both open-source models (hatched bars) showed declining scores, with Llama-3 70B Instruct falling from 62.0 to 54.0 points and Command *R*+ from 52.0 to 50.0 points. This preliminary observation indicates a performance difference, with proprietary models scoring 18.5 points higher at posttest than open-source alternatives.

Figure 2. Multiple-choice assessment scores of large language models (LLMs) before and after adding behavioral activation context. Performance comparison of proprietary LLMs (solid bars) and open-source (hatched bars) on the multiple-choice knowledge assessment before (pretest) and after (posttest) receiving BA training materials as additional context.

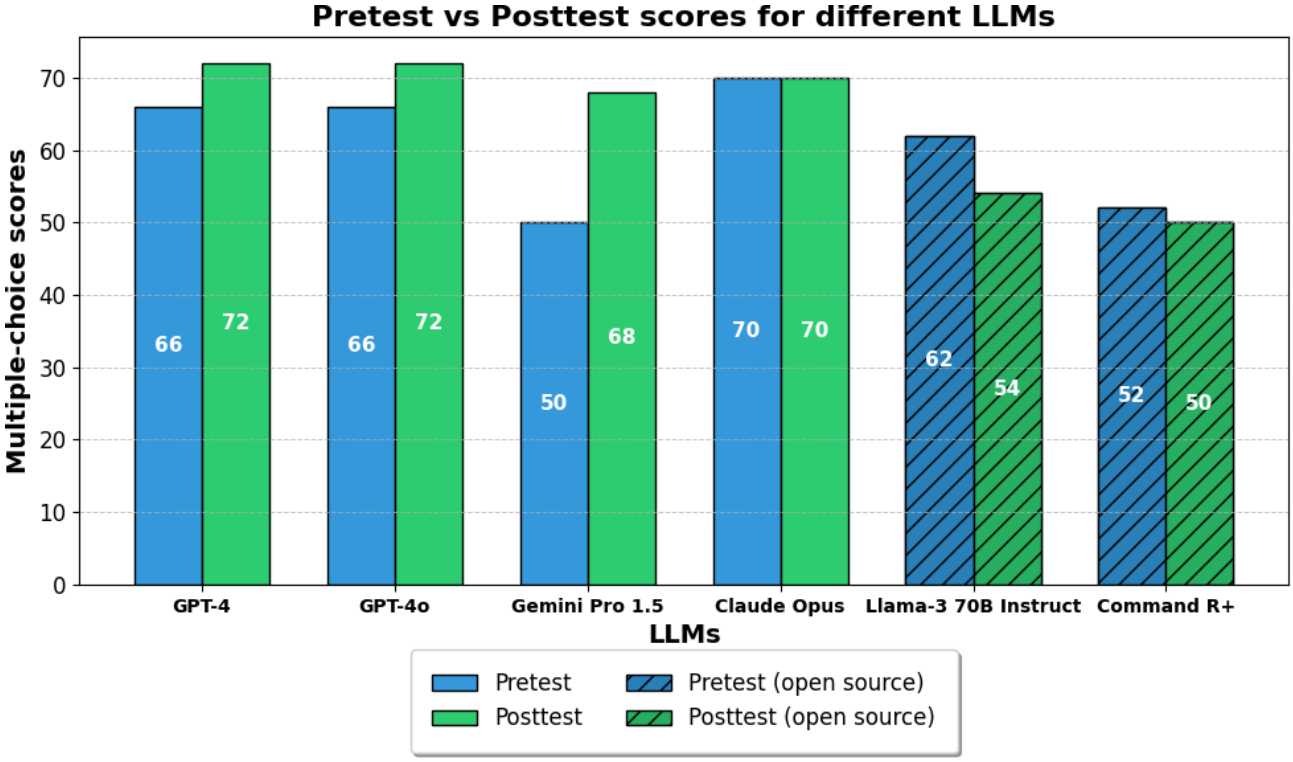

In addition to the multiple-choice assessment, an open-ended questionnaire was used to evaluate more nuanced therapeutic skills. Table 2 shows the average scores for human and LLM responses. Across all pretest and posttest evaluations, LLMs achieved higher average scores than human participants on 5 psychological metrics (rated on a scale of 1 to 5). Pretest between-group comparisons showed significantly higher LLM performance for anticipation of cognition (*U*=0.0; *P*=.002; *r*=1.000) and anticipation of emotion (*U*=0.0; *P*=.002; *r*=1.000), with nonsignificant trends for adherence (*U*=4.0; *P*=.008; *r*=0.833), empathy (*U*=2.0; *P*=.005; *r*=0.917), and validation quality (*U*=2.5; *P*=.006; *r*=0.896). Both groups demonstrated improvements from pretest to posttest in adherence, empathy, and validation quality, with the largest gains in adherence (humans: 2.79, SD 0-14 vs 3.54, SD 0.43; LLMs: 3.31, SD 0.12 vs 3.96, SD 0.06). Within-group analyses revealed no significant improvements after correction, though adherence showed trends for both humans (*W*=1.0; *P*=.02; *r*=0.944) and LLMs (*W*=0.0; *P*=.03; *r*=1.000). At posttest, between-group differences were significant for adherence (*U*=2.0; *P*=.003; *r*=0.917) and empathy (*U*=1.5; *P*=.004; *r*=0.938), with nonsignificant trends for the other metrics after correction. LLMs exhibited decreases in anticipation of cognition (4.54, SD 0.30 vs 4.43, SD 0.42) and anticipation of emotion (4.69, SD 0.12 vs 4.40, SD 0.37).

XSL•FO
RenderX

Table . Mean assessment scores from the open-ended question evaluation of psychotherapists in training and large language models (LLMs).

Category	Pretest (human), mean (SD)	Posttest (human), mean (SD)	Pretest (LLM), mean (SD)	Posttest (LLM), mean (SD)	ICC[a]
Adherence	2.79 (0.14)	3.54 (0.09)	3.31 (0.12)	3.96 (0.06)[b]	0.55
Empathy	3.14 (0.43)	3.64 (0.29)	4.33 (0.09)	4.35 (0.26)[b]	0.78
Validation quality	2.78 (0.22)	3.43 (0.29)	3.78 (0.14)	4.08 (0.13)[c]	0.6
Anticipation of cognition	3.53 (0.26)	4.01 (0.26)	4.54 (0.30)[b]	4.43 (0.42)	0.67
Anticipation of emotion	3.36 (0.16)	3.80 (0.24)	4.69 (0.13)[b]	4.40 (0.37)	0.76
Average across categories	3.18 (0.10)	3.72 (0.17)	4.21 (0.07)	4.28 (0.14)[c]	0.67

[a]ICC: intraclass correlation coefficient.

[b]Indicate significant differences between human and LLM scores at the same testing point (P<.005).

[c]Values represent the highest score in each category across all conditions.

Validation strategy use showed broadly comparable patterns between LLMs and humans, with both groups reducing overall use by approximately 5%. Statistical analyses used Wilcoxon signed-rank tests for within-group comparisons and Mann-Whitney U tests for between-group comparisons, with Bonferroni correction applied (significance threshold P<.004). V1 (attentiveness) remained stable across both groups with no significant within-group changes. Both groups demonstrated similar patterns for V2 (intermodal), V3 (crossmodal), and V5 (present). V3 showed the largest increases in both groups, rising 133% for humans and 160% for LLMs, though within-group analyses revealed only trends toward significance (humans: W=3.0; P=.03; r=0.833; LLMs: W=3.0; P=.11; r=0.714). V5 decreased substantially in both groups, dropping 57% for humans and 67% for LLMs (humans: W=5.0; P=.07; r=-0.722; LLMs: W=1.5; P=.09; r=−0.857). The clearest divergence appeared in V6 (radical genuineness), where humans increased use by 38% while LLMs decreased by 14% (humans: W=12.0; P=.39; r=0.333; LLMs: W=7.5; P=.52; r=−0.286). Between-group comparisons at T1 (postintervention) showed a trend toward difference for V4 (biography; U=7.0; P=.01; r=0.708), although this did not survive Bonferroni correction. After correction, no comparisons reached statistical significance, indicating these patterns should be interpreted as trends rather than definitive effects (Table 3).

Table . Number of applied validation strategies of psychotherapists in training and large language models (LLMs).

Validation strategy	Pretest (human), n	Posttest (human), n	Change in human responses (%)	Pretest (LLM), n	Posttest (LLM), n	Change in LLM responses (%)
V1 (attentiveness)	39	39	0	30	30	0
V2 (intermodal communication)	11	6	−45	11	8	−27
V3 (crossmodal communication)	6	14	+133	5	13	+160
V4 (reference to client biography)	10	8	−20	10	11	+10
V5 (reference to the present)	14	6	−57	12	4	−67
V6 (radical genuineness)	8	11	+38	14	12	−14
Sum over all strategies	88	84	−5	82	78	−5

Discussion

Principal Findings

This study set out to evaluate whether LLMs could demonstrate therapeutic competencies comparable to psychotherapists in training, particularly in the domains of knowledge acquisition, therapeutic alliance building, and empathic communication. Consistent with recent studies showing that users perceive LLMs like Replika and ChatGPT as emotionally supportive despite not yet being designed for therapy [8,36], our findings indicated that LLMs received moderately higher ratings in certain therapeutic skills, including empathetic communication and anticipation of emotional and cognitive states. Of particular interest was the LLMs' performance in domains traditionally

XSL•FO
RenderX

associated with human capabilities: empathetic communication and emotional understanding. The observed differences between LLMs and human trainees in pretest expert-rated empathy (U=2.0; P=.005; r=0.917), anticipation of cognition (U=0.0; P=.002; r=1.000), and anticipation of emotion (U=0.0; P=.002; r=1.000) reveal an important distinction between producing seemingly empathetic responses and the processes behind them. Although LLMs do not experience emotions, they generated responses that trained evaluators rated as empathetic and emotionally aware, similar to human trainees. While these findings echo prior reports that LLMs can simulate empathic communication [7,12], they also raise questions of whether simulated empathy translates into effective therapeutic alliances with real patients, indicating the necessity for future research on how patients experience and respond to therapeutic interactions with LLM systems compared with human therapists.

The convergence of proprietary models' performance on the multiple-choice assessment after exposure to training materials, reaching scores between 68 to 72 points, aligns with research findings suggesting that LLMs can rapidly integrate structured therapeutic information through few-shot prompting or fine-tuning [11-13] and may indicate similar underlying approaches to processing and integrating therapeutic knowledge across different proprietary LLMs. In contrast, the declining performance of open-source models points to potential limitations in their ability to effectively integrate and apply additional context while maintaining consistent performance for the investigated models.

These differences between proprietary and open-source models have important implications for potential therapeutic applications. The stability and convergence of proprietary models suggest they may offer more reliable performance in therapeutic contexts. However, their proprietary nature raises questions about implementation costs in real-world mental health applications, as health care providers would need to rely on more costly external APIs. The performance gap between proprietary and open-source models might present a trade-off between reliability and cost that requires careful consideration in the development of LLM-assisted mental health interventions.

Our analysis of validation strategies revealed important insights into their use patterns in both LLMs and human trainees. Across both human and LLM responses, the fundamental strategy of attentiveness remained consistent, aligning with its essential role in therapeutic alliance formation as established in previous research [24-33]. Cross-modal and intermodal communication strategies were rarely used simultaneously, likely due to the constrained nature of single-response evaluations. Subsequent training led to an increase in cross-modal communication, suggesting successful integration when explicitly emphasized in training materials. However, the observed decrease in present-moment references may indicate a shift toward more structured therapeutic interventions, as participants potentially prioritized demonstrating newly acquired technical skills over moment-to-moment emotional attunement.

Limitations

While our findings demonstrate strong performance from LLMs across multiple therapeutic competencies, these results should be interpreted within several important constraints. First, our evaluation was limited to single-turn therapeutic responses, affecting ecological validity, since real therapeutic relationships require maintaining consistent dialogue across sessions, building upon accumulated client information, and adapting therapeutic approaches as client needs evolve. Future research needs to examine their performance in multi-turn therapeutic dialogues, their ability to maintain consistent client understanding across sessions, and adaptation to changing therapeutic needs. Additionally, the sample of 8 psychotherapists in training is too small to draw reliable inferences regarding the broader population of psychotherapists in training. The testing material should be administered to larger samples of psychotherapists and psychotherapists in training. More expert ratings might also enhance variability and thus offer more scope for interpretation.

The questionnaire was conducted in a paper-and-pencil format, which might account for shorter answers in the human sample, thus affecting the rating. The rapid evolution of AI models also affects the longevity of our findings. Even during this study's preparation, more sophisticated models like Anthropic's Claude 4.0 Sonnet and OpenAI's GPT-5 were released, highlighting how quickly such comparative analyses can become outdated. The development of standardized evaluation frameworks that can account for the rapid evolution of LLM capabilities while maintaining rigorous assessment standards will be crucial for ongoing comparative analyses. Such frameworks should particularly focus on longitudinal therapeutic interactions and consistent application of therapeutic principles over time.

Another key methodological limitation affects our ability to make direct comparisons. The LLMs had access to all training materials during testing, while human participants had to rely on retained knowledge from their training period. This created an asymmetric testing environment similar to comparing open-book and closed-book exam performance. An open-book examination should be conducted and might also reflect a more realistic scenario for psychotherapists in training, since these are closer to a real-life therapeutic session where therapists have access to literature, data, and the possibility to look up information if they need to.

Conclusions

This study provides evidence that LLMs can generate high-quality, single-turn therapeutic responses that effectively combine clinical knowledge with empathetic communication. Our findings reveal that LLMs often matched psychotherapists in training in knowledge assessments and practical therapeutic alliance-building skills. While LLMs cannot experience genuine empathy, their ability to produce responses rated as empathetic by licensed psychotherapists highlights the distinction between internal emotional experience and effective therapeutic communication.

Our findings suggest clinical applications for LLMs as supportive tools in mental health care, potentially addressing the critical shortage of trained professionals and demand in low-resource settings. They could supplement care or expand access to mental health care. However, such applications must be guided by rigorous safeguards, as the distinction between

XSL•FO
RenderX

simulated and genuine human empathy has direct implications for therapeutic authenticity and patient outcomes.

Differences between proprietary and open-source LLMs emphasize issues of accessibility, equity, and sustainability. Though our results carefully suggest that proprietary models offer more reliable performance, their closed nature and cost structures risk exacerbating inequalities in access to high-quality digital mental health tools. Addressing these trade-offs will be critical in planning AI integration.

As mental health care faces growing demand, LLMs may serve as valuable supplements to human expertise, though their implementation requires careful consideration of ethical implications, therapeutic authenticity, and clinical outcomes.

Data Availability

The datasets generated or analyzed during this study are available in the GitHub repository cited in the Questionnaire section.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Responses of Claude Opus to open-ended questions of case scenarios 2, 6, 7, and 10.
[DOCX File, 22 KB - formative_v9i1e78138_app1.docx]

References

1. Mental disorders. World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/mental-disorders [accessed 2025-02-16]
2. Bloom DE, Cafiero ET, Jané-Llopis E, et al. The global economic burden of non-communicable diseases. World Economic Forum. 2011 Sep. URL: https://www3.weforum.org/docs/WEF_Harvard_HE_GlobalEconomicBurdenNonCommunicableDiseases_2011.pdf [accessed 2025-11-10]
3. Cambridge OR, Knight MJ, Mills N, Baune BT. The clinical relationship between cognitive impairment and psychosocial functioning in major depressive disorder: a systematic review. Psychiatry Res 2018 Nov;269:157-171. [doi: 10.1016/j.psychres.2018.08.033] [Medline: 30149273]
4. Rajan S, McKee M, Rangarajan S, et al. Association of symptoms of depression with cardiovascular disease and mortality in low-, middle-, and high-income countries. JAMA Psychiatry 2020 Oct 1;77(10):1052-1063. [doi: 10.1001/jamapsychiatry.2020.1351] [Medline: 32520341]
5. Sharma A, Lin IW, Miner AS, Atkins DC, Althoff T. Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. Nat Mach Intell 2023;5(1):46-57. [doi: 10.1038/s42256-022-00593-2]
6. Kim J, Leonte KG, Chen ML, et al. Large language models outperform mental and medical health care professionals in identifying obsessive-compulsive disorder. NPJ Digit Med 2024 Jul 19;7(1):193. [doi: 10.1038/s41746-024-01181-x] [Medline: 39030292]
7. Abd-Alrazaq AA, Alajlani M, Ali N, Denecke K, Bewick BM, Househ M. Perceptions and opinions of patients about mental health chatbots: scoping review. J Med Internet Res 2021 Jan 13;23(1):e17828. [doi: 10.2196/17828] [Medline: 33439133]
8. Maples B, Cerit M, Vishwanath A, Pea R. Loneliness and suicide mitigation for students using GPT3-enabled chatbots. Npj Ment Health Res 2024 Jan 22;3(1):4. [doi: 10.1038/s44184-023-00047-6] [Medline: 38609517]
9. Gallegos IO, Rossi RA, Barrow J, et al. Bias and fairness in large language models: a survey. Comput Linguist 2024 Sep 1;50(3):1097-1179. [doi: 10.1162/coli_a_00524]
10. Xu Z, Jain S, Kankanhalli M. Hallucination is inevitable: an innate limitation of large language models. arXiv. Preprint posted online on Jan 22, 2024. [doi: 10.48550/arXiv.2401.11817]
11. Na H. CBT-LLM: a Chinese large language model for cognitive behavioral therapy-based mental health question answering. arXiv. Preprint posted online on Mar 24, 2024. [doi: 10.48550/arXiv.2403.16008]
12. Kim JW, Han JE, Koh JS, Seo HT, Chang DS. Enhancing psychotherapy counseling: a data augmentation pipeline leveraging large language models for counseling conversations. arXiv. Preprint posted online on Jun 13, 2024. [doi: 10.48550/arXiv.2406.08718]
13. Lin B, Bouneffouf D, Cecchi G, Varshney K. Towards healthy AI: large language models need therapists too. Presented at: Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP '24); Jun 21-22, 2024; Mexico City, Mexico p. 61-70 URL: https://aclanthology.org/2024.trustnlp-1 [accessed 2025-11-10]
14. Liu JM, Li D, Cao H, Ren T, Liao Z, Wu J. Chatcounselor: a large language models for mental health support. arXiv. Preprint posted online on Sep 27, 2023. [doi: 10.48550/arXiv.2309.15461]

XSL•FO
RenderX

15. Tindall L, Mikocka-Walus A, McMillan D, Wright B, Hewitt C, Gascoyne S. Is behavioural activation effective in the treatment of depression in young people? A systematic review and meta-analysis. Psychol Psychother 2017 Dec;90(4):770-796. [doi: 10.1111/papt.12121] [Medline: 28299896]
16. Uphoff E, Ekers D, Robertson L, et al. Behavioural activation therapy for depression in adults. Cochrane Database Syst Rev 2020 Jul 6;7(7):CD013305. [doi: 10.1002/14651858.CD013305.pub2] [Medline: 32628293]
17. OpenAI AJ, Adler S, et al. GPT-4 technical report. arXiv. Preprint posted online on Mar 4, 2024. [doi: 10.48550/arXiv.2303.08774]
18. Georgiev P, Lei VI, et al. Gemini 1.5: unlocking multimodal understanding across millions of tokens of context. arXiv. Preprint posted online on Mar 8, 2024. [doi: 10.48550/arXiv.2403.05530]
19. Introducing the next generation of Claude. Anthropic PBC. URL: https://www.anthropic.com/news/claude-3-family [accessed 2025-11-10]
20. Grattafiori A, Dubey A, Jauhri A, et al. The Llama 3 herd of models. arXiv. Preprint posted online on Jul 31, 2024. [doi: 10.48550/arXiv.2407.21783]
21. Introducing command r+: a scalable LLM built for business. Cohere. URL: https://cohere.com/blog/command-r-plus-microsoft-azure [accessed 2025-02-16]
22. The unified interface for LLMs. OpenRouter. URL: https://openrouter.ai/ [accessed 2025-11-20]
23. PsychThG - gesetz über den beruf der psychotherapeutin und des psychotherapeuten. Bundesministerium der Justiz und für Verbraucherschutz. URL: https://www.gesetze-im-internet.de/psychthg_2020/BJNR160410019.html [accessed 2025-02-16]
24. Del Re AC, Flückiger C, Horvath AO, Wampold BE. Examining therapist effects in the alliance-outcome relationship: a multilevel meta-analysis. J Consult Clin Psychol 2021 May;89(5):371-378. [doi: 10.1037/ccp0000637] [Medline: 33829817]
25. Wampold BE, Flückiger C. The alliance in mental health care: conceptualization, evidence and clinical applications. World Psychiatry 2023 Feb;22(1):25-41. [doi: 10.1002/wps.21035] [Medline: 36640398]
26. Kim S. Therapist's empathy, attachment, and therapeutic alliance: neurobiological perspective. Int J Psychol Behav Anal 2018;4(1):1-5. [doi: 10.15344/2455-3867/2018/140]
27. McClintock AS, Anderson T, Patterson CL, Wing EH. Early psychotherapeutic empathy, alliance, and client outcome: preliminary evidence of indirect effects. J Clin Psychol 2018 Jun;74(6):839-848. [doi: 10.1002/jclp.22568] [Medline: 29364509]
28. Rogers CR. Empathic: an unappreciated way of being. Couns Psychol 1975 Jun;5(2):2-10. [doi: 10.1177/001100007500500202]
29. Klee MR, Abeles N, Muller RT. Therapeutic alliance: early indicators, course, and outcome. Psychol Psychother Theory Res Pract 1990;27(2):166-174. [doi: 10.1037/0033-3204.27.2.166]
30. Horvath AO. The therapeutic relationship: from transference to alliance. J Clin Psychol 2000;56(2):163-173. [doi: 10.1002/(sici)1097-4679(200002)56:2<163::aid-jclp3>3.0.co;2-d] [Medline: 10718600]
31. Horvath AO. The therapeutic alliance: concepts, research and training. Aust Psychol 2001 Jul 1;36(2):170-176. [doi: 10.1080/00050060108259650]
32. Ackerman SJ, Hilsenroth MJ. A review of therapist characteristics and techniques positively impacting the therapeutic alliance. Clin Psychol Rev 2003 Feb;23(1):1-33. [doi: 10.1016/s0272-7358(02)00146-0] [Medline: 12559992]
33. Summers RF, Barber JP. Therapeutic alliance as a measurable psychotherapy skill. Acad Psychiatry 2003;27(3):160-165. [doi: 10.1176/appi.ap.27.3.160] [Medline: 12969839]
34. Sutor M, editor. Die Dialektisch Behaviorale Therapie (DBT): Neue DBT-Orientierte Diagnoseübergreifende Konzepte - Schwerpunkt Skills-Training: Springer. [doi: 10.1007/978-3-662-64627-4]
35. GitHub. Single turn evaluation of LLMs and human therapists trainees in BA knowledge and response. URL: https://github.com/Julianlaue/Single-Turn-Evaluation-of-LLMs-and-human-therapists-trainees-in-BA-Knowledge-and-Response [accessed 2025-11-20]
36. Luo X, Ghosh S, Tilley JL, Besada P, Wang J, Xiang Y. "Shaping ChatGPT into my digital therapist": a thematic analysis of social media discourse on using generative artificial intelligence for mental health. Digit Health 2025;11:20552076251351088. [doi: 10.1177/20552076251351088] [Medline: 40656852]

Abbreviations

BA: behavioral activation
CBT: cognitive behavioral therapy
LLM: large language model

XSL•FO
RenderX

Edited by A Mavragani, J Sarvestan; submitted 28.05.25; peer-reviewed by A Hamdan, AT Thomas, B Foroutan, O Oloruntoba; revised version received 25.09.25; accepted 25.09.25; published 04.12.25.

Please cite as:
Napiwotzki I, Laue J, Caldarone F, Idahl M, Hadler U, Amrani H, Hildt E, Kahl KG, Nejdl W
Comparing Human and AI Therapists in Behavioral Activation for Depression: Cross-Sectional Questionnaire Study
JMIR Form Res 2025;9:e78138
URL: https://formative.jmir.org/2025/1/e78138
doi:10.2196/78138

© Inka Napiwotzki, Julian Laue, Flora Caldarone, Maximilian Idahl, Uwe Hadler, Haithem Amrani, Elisabeth Hildt, Kai G Kahl, Wolfgang Nejdl. Originally published in JMIR Formative Research (https://formative.jmir.org), 4.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Advancing Digital Access to Physical Therapy via Virtual and Extended Reality Technology: Prototype Development and Usability Evaluation

Victoria Lynn Tiase[1*], PhD, RN; Julie M Fritz[2*], PhD, PT; Jesse Ferraro[3*], MFA; Gregory Bayles[4], MA; Ahmad Alsaleem[5], PhD; Guilherme Del Fiol[1*], MD, PhD; Kensaku Kawamoto[1*], MD, PhD; Darrel Brodke[6*], MD; Brook Martin[6*], MPH, PhD; Roger Altizer[3*], PhD

[1]Department of Biomedical Informatics, School of Medicine, University of Utah, 421 Wakara Way, Salt Lake City, UT, United States
[2]Physical Therapy and Athletic Training, College of Health, University of Utah, Salt Lake City, UT, United States
[3]Population Health Sciences, School of Medicine, University of Utah, Salt Lake City, UT, United States
[4]Division of Games, College of Architecture and Planning, University of Utah, Salt Lake City, UT, United States
[5]School of Computing, College of Engineering, University of Utah, Salt Lake City, UT, United States
[6]Department of Orthopaedics, School of Medicine, University of Utah, Salt Lake City, UT, United States
[*]these authors contributed equally

Corresponding Author:
Victoria Lynn Tiase, PhD, RN
Department of Biomedical Informatics, School of Medicine, University of Utah, 421 Wakara Way, Salt Lake City, UT, United States

Abstract

Background: The United States faces significant challenges in physical therapy (PT) access due to high demand, a shortage of professionals, and patient-related obstacles, which can adversely affect recovery and function. Limited access to PT may lead to increased dependence on medications for pain management, highlighting the need for nonpharmacologic options to reduce opioid overprescribing. Low back pain, a leading cause of disability and high medical costs, is a common reason for requiring PT following surgery. Studies have shown that virtual reality (VR)–guided movements can improve motor function and reduce pain intensity.

Objective: The objective of this study was to design, develop, and evaluate a VR-based prototype for individualized postoperative PT for patients recovering from back surgery to investigate its potential to improve convenience, access, and health outcomes in future research.

Methods: Study methods involved participatory design and development of VR software for PT back exercises using the design box method, an inductive, problem-oriented collaborative design approach. A usability evaluation of the resulting prototype was conducted with patients recovering from back surgery using a think-aloud protocol and usability survey.

Results: Six participants evaluated the VR prototype and reported usability challenges that included mismatched VR boundaries, limited familiarity with VR, and difficulties with the headset and hand controls. The System Usability Scale resulted in a total usability score of 58.3 out of 100, indicating a below-average score (68 being average).

Conclusions: In the design and evaluation of a VR-based PT prototype, we found that while participants were enthusiastic, they faced usability challenges due to insufficient instructions and difficulties operating the VR device, highlighting the need for effective onboarding and extensive prototype testing to improve accessibility and engagement in health care. Future evaluations will investigate disparities among different groups to ensure accessibility and effectiveness for all users.

(JMIR Form Res 2025;9:e73783) doi:10.2196/73783

KEYWORDS

virtual reality; physical therapy; low back pain; user-centered design; access to care

Introduction

The United States is currently experiencing significant challenges in physical therapy (PT) access due to high demand; a shortage of trained professionals; and obstacles related to patients' location, scheduling, and mobility issues [1,2]. Many health systems across the country struggle to provide the recommended PT care, leading to longer wait times and extended periods without care, which can adversely affect patients' recovery and function [3,4]. This issue is particularly

XSL•FO
RenderX

concerning for older adults who may rely on PT to maintain their independence [5,6]. As PT is an evidence-based, nonpharmacological option for pain management, limited access may drive patients to increasingly depend on medications or other pain care options [7]. Expanding access to nonpharmacologic options such as PT is a crucial aspect of reducing overprescribing of opioids and other pain medications and reducing the impact of pain on the quality of life of individuals [8-10]. Despite efforts to train more physical therapists, effective strategies that leverage scalable technology, such as extended and virtual reality (VR), are still needed to overcome all the barriers that limit access to PT services [2].

Extended reality, an umbrella term that encompasses all immersive technologies, including augmented reality and VR technology, is rapidly transforming health care in numerous innovative ways [11]. VR, a digital environment that replaces the physical world, is being used to treat phobias and mental health disorders [12]. Additionally, VR facilitates health behaviors by encouraging physical activity within enjoyable, engaging virtual environments [13]. Since 2019, the adoption of VR in health care has grown, for example, with the emergence of VR-based health care providers who interact with patients virtually, eliminating the need for physical presence [12]. VR has shown promising results in orthopedic rehabilitation for chronic neck pain and shoulder impingement syndrome and is considered a valuable tool for upper-extremity neurorehabilitation and pain distraction [13-15]. With the right equipment and training, VR offers a novel and patient-centric solution to expand access to PT care beyond what basic PT health applications can offer.

Low back pain is one of the most common reasons why people seek PT as it is the leading cause of disability and one of the costliest medical conditions, with Medicare data revealing that 94% of patients with a new diagnosis of back pain have an outpatient PT episode within a year [12,16,17]. With this high need for PT, low back pain is a plausible condition to explore the use of VR visits. Recently, multiple studies have demonstrated that VR movements can improve motor function outcomes, significantly reducing pain intensity in patients with low back pain [17-19]. Additionally, there is evidence suggesting that remote PT via telehealth is comparable to in-person PT, with patients showing some willingness to use this approach [20-22]. Developing a VR-based PT program for patients with chronic low back pain could be an effective strategy to expand access to PT care to manage this condition.

In this study, ADAPT XR (Advancing Digital Access to Physical Therapy via Virtual and Extended Reality Technology), we designed and developed a VR-based prototype to provide customized postoperative PT for patients recovering from back surgery. This study consisted of two stages: (1) participatory design and development of VR software for PT back exercises and (2) usability evaluation of a prototype VR-based PT intervention with patients recovering from back surgery. We hypothesized that the VR-based care approach for postsurgical back rehabilitation would be well received by patients.

Methods

Participatory Design and Development

In the first phase, we used a participatory design approach that was user-centric, visual, iterative, and collaborative, fostering end-user engagement and design thinking [23]. We conducted 2 focus group sessions using the design box method, a technique originally developed for video game design that facilitates end-user collaboration in prototype design and development [24,25]. Given the gaming aspects of VR, we felt that the design box methodology would be best suited for this application. One focus group comprised physical therapists, and the second consisted of individuals who had previously used PT services following back surgery.

The design box method, created by Altizer et al [25] at the University of Utah, advances creative ideation by allowing groups to generate ideas iteratively and design prototype features using a visual tool. It is an inductive design methodology built around the notion that good design solves a problem. This method includes drawing 4 “walls” or constraints on an electronic whiteboard: problem, technology, audience, and aesthetics (Figure 1). The moderator, a member of the research team, gathers input from participants for each wall in sequence, recording their contributions in real time for everyone to see. Once theoretical saturation is reached and no new ideas are generated, design box focus group participants refine the ideas to arrive at a clear design that the entire group supports, otherwise referred to as the “pitch.” By leveraging the 4 walls as constraints, participants are free to pitch any idea that addresses the problem statement, technical needs, audience needs, and agreed-upon aesthetics.

XSL•FO
RenderX

Figure 1. Design box methodology with 4 walls or constraints.

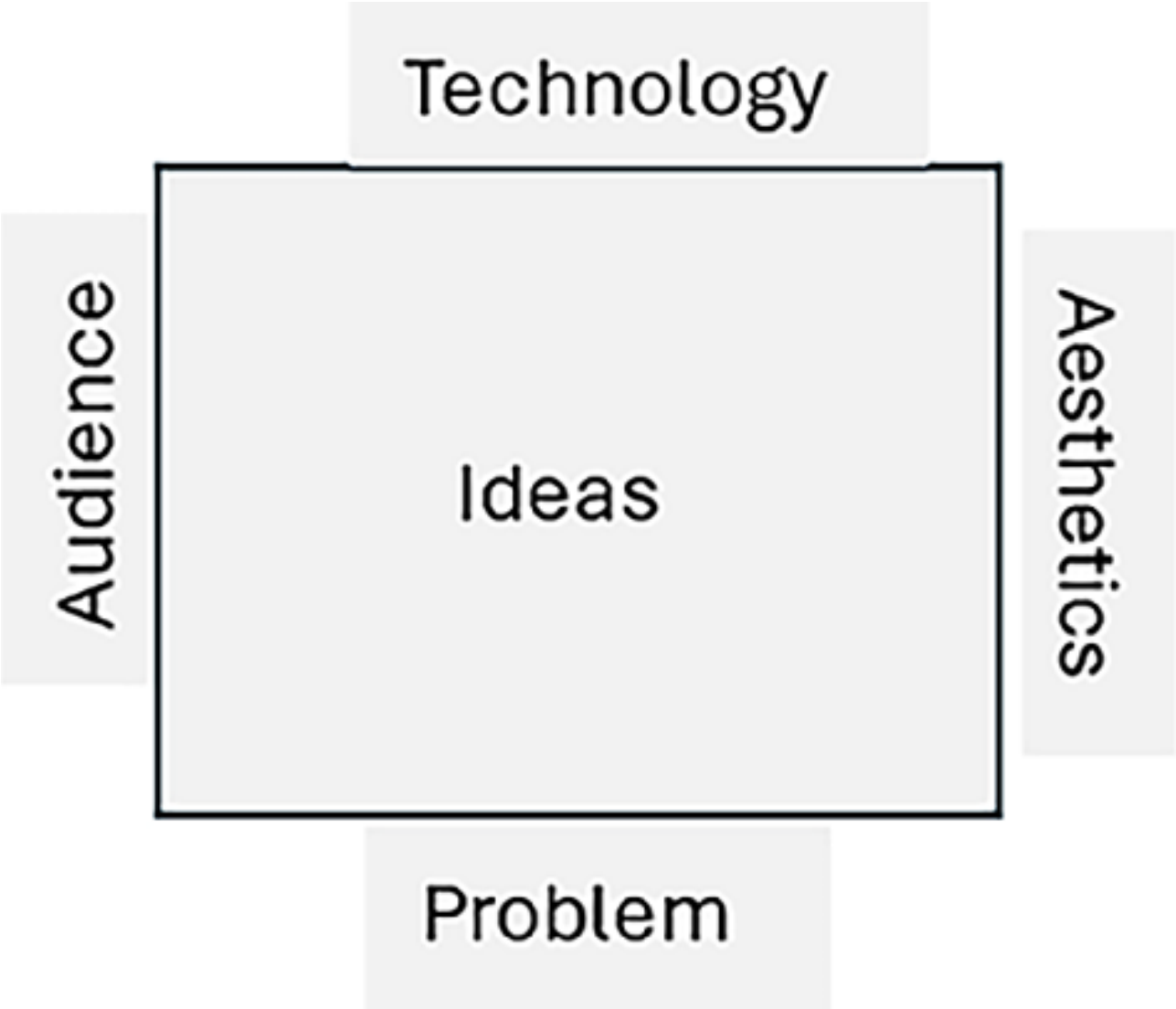

For the physical therapist focus group, we recruited therapists with outpatient experience in providing PT. Using purposive sampling, we targeted physical therapists at the University of Utah in Salt Lake City, Utah, through email and word of mouth. Additionally, we used a snowball sampling technique, asking existing study subjects and local contacts to identify other potential participants. The inclusion criteria were individuals who were aged ≥18 years, English speaking, and licensed to practice PT.

For the second focus group, we recruited individuals who had undergone PT following surgery for back pain. Using both purposive and snowball sampling, we asked health care providers to refer patients who met the criteria. Other inclusion criteria were individuals who were aged ≥18 years, English speaking, and able to meet via Zoom (Zoom Video Communications).

Once the design box focus group sessions were completed, prototype development for the VR application commenced. To facilitate development, the notes and whiteboards from the design box sessions were summarized and analyzed to determine which ideas or features were common and fit within the scope of the project. The analysis of design boxes shares similarities with inductive approaches as the core design of the project emerges from commonalities in and observations of the design box sessions [26]. The core design is broken down into features, which in turn are put into a production backlog, facilitating agile software development [27].

A team of 3 graduate students, 1 producer, and 1 engineer from the Therapeutic Games and Apps Lab at the University of Utah reviewed and aligned the design box results with the capabilities of the Meta Quest VR device (Reality Labs), adding further nuance to the design and backlog. The Meta Quest VR device, a commercial off-the-shelf technology, was selected due to its market prominence, making it easy for translation and adoption. It should be noted that members of the research team were also members of the development team, allowing for negotiation between features that participants in the design box sessions wanted and what was technically feasible and within the project's scope.

To decide which PT exercise to prototype, the team met with physical therapists to review which exercises were used in postsurgical back pain PT and could be tracked using the controllers and headset of the Meta Quest VR system. The development team at the Therapeutic Games and Apps Lab

XSL•FO
RenderX

created a detailed design plan for the leg squat exercise, identified by the PT experts as commonly used in postsurgical back pain PT. The plan was based on 2-week sprint cycles of software development, a key aspect of agile development [27]. Sprint cycles allow for work to be broken down into scheduled iterations, which in turn allow for rapid development and the flexibility to modify any features in between sprint cycles based on user, expert, and team feedback [27]. A total of 5 sprint cycles were executed for this project. Using the development platform Unity (Unity Technologies), the team programmed the prototype's functionality, integrating VR-specific features such as hand tracking and 6-df movement. Rigorous testing was conducted to ensure smooth performance and a safe, immersive experience, addressing any technical flaws or issues. A functioning prototype for the Meta Quest device was completed and used during the usability evaluation phase.

Usability Evaluation

Following the development of the prototype, participants who had previously undergone PT following surgery for back pain were recruited for a 1-hour in-person session to take part in a usability evaluation. Recruitment strategies included soliciting patients from the University of Utah Orthopedic Center and the researchers' professional contacts. Using both purposive and snowball sampling, we asked health care providers to refer patients who met the criteria. Inclusion criteria specified individuals aged ≥18 years who were English speaking and had completed PT for postsurgical back pain. Patients currently undergoing treatment for postsurgical back pain were excluded from the study so as to not interfere with treatment regimens.

The usability evaluation comprised several steps, including an interview using a think-aloud protocol and a postuse survey [28,29]. To ensure adherence to study procedures, the research team developed guidance documents for the Meta Quest setup and cleaning between uses, think-aloud protocol questions, postuse survey questions, and safety procedures. Three members of the research team, all trained in usability methods, were present for each of the usability evaluation sessions.

Once eligibility was confirmed and consent was completed, participants were introduced to the general use of the Meta Quest headset and hand controllers. Participants were asked to perform the tasks as instructed by the VR prototype. Subsequently, participants were asked a series of questions and probes in conjunction with the think-aloud protocol to facilitate a deeper exploration of their interaction with the tool. Examples of probes included "What do you think that means?" or "Is that what you expected to see?" The questions were designed to progress from general to more specific information.

Additionally, each participant completed a 10-item validated survey tool, the industry-standard System Usability Scale (SUS), to assess factors such as perceived user performance, perceived decision quality, and perceived task completion time [30]. Responses were collected using REDCap (Research Electronic Data Capture; Vanderbilt University), a secure web application [31-33]. Given the formative and diagnostic nature of this study, our primary focus was on identifying significant usability issues. Following SUS guidelines, we calculated a percentile ranking and average score for the responses.

Conceptually, we used thematic saturation, ceasing recruitment when no new insights or themes emerged from the data [34]. Unlike quantitative studies, qualitative studies typically have smaller sample sizes, and guidance for qualitative sample sizes indicates that above 5 is considered adequate for identifying a large portion of usability issues [35-37]. At the conclusion of the usability evaluation, we advised participants to wait for 30 minutes before attempting to drive home or operate heavy machinery as disorientation or nausea could pose safety risks after using a VR device.

Verbatim transcripts were anonymized and verified by a second investigator. Free-text segments were excerpted from the transcripts for content analysis. Using a securely stored spreadsheet, 3 members of the research team coded the transcripts based on the 9 concepts of the Health Information Technology Usability Evaluation Model [38,39]. To enhance rigor, coders first underwent a group coding session to establish a shared understanding of the coding framework. Instead of calculating intercoder reliability, we conducted iterative consensus coding. Through group consensus, we labeled segments as "positive," "negative," and "neutral" for the 9 concepts. Regular meetings with the coders were held to discuss and resolve discrepancies, refine the codebook, and ensure consistent application of the conceptual categories in line with qualitative best practices.

Ethical Considerations

This study received ethics approval from the University of Utah Institutional Review Board (00165806). After the usability evaluation, participants were emailed a US $50 gift card as compensation for taking part. All participants provided written informed consent prior to participation and were informed of the study purpose, procedures, potential risks, and their right to withdraw at any time without penalty. To protect participant confidentiality, all raw data were de-identified prior to analysis. A separate linkage file connecting participants to study IDs was stored in a secure, encrypted location accessible only to the study team. No identifiable information was shared outside the research team.

Results

Participatory Design and Development

Seven individuals met the inclusion criteria for the design box focus groups: 3 (43%) physical therapists and 4 (57%) individuals who had undergone PT for postsurgical back pain. Each of the sessions lasted approximately 90 minutes and was audio recorded. The sessions were conducted using Zoom, and the moderator used Miro as the electronic whiteboard. Idea refinement for each of the constraints is shown in Table 1. All participants enthusiastically supported the vision of designing VR-enabled PT exercises.

XSL•FO
RenderX

Table . ADAPT XR (Advancing Digital Access to Physical Therapy via Virtual and Extended Reality Technology) design box constraints and ideas resulting from the focus group sessions.

Constraint	Final idea convergence
Problem	Fear of reinjury if unsupervised
Technology	Novelty; simple to use
Audience	Convenient; fit into schedule
Aesthetics	Fun; mini games in between repetitions to keep the intervention engaging

The pitches from the design box session were reflected in the final prototype design, functionality, and user experience. The VR prototype contained basic menus, buttons, and other interactive elements (Figure 2). The virtual environment resembled a high-technology gym setting with a futuristic avatar used to demonstrate the leg squat exercise (Figure 3). This aesthetic was chosen based on feedback from the design box participants. Participants wanted the environment to be more "fun and exciting" than a standard gym. Additionally, participants mentioned that while some had positive associations with PT gyms, others had negative ones due to postsurgical discomfort.

The VR software leveraged visual and auditory feedback to indicate users' actions in the software. Buttons would animate to appear pressed, and audio was used to give additional feedback. The users were able to view the avatar (or coach) in front of them demonstrating the exercise and could mirror their movements. Users had the ability to touch buttons or use a virtual laser pointer to click on user interface elements that were out of reach. The VR software enabled users, who may be experiencing physical discomfort, to interact with the user interface without needing to stretch or bend. The experience began with an introduction to the environment and a demonstration of how to perform a leg squat, along with advice on how to do so properly and safely. The users were then able to practice along with the avatar in front of them. Users could decide between 2 modes for their exercise: a timed mode and a counter mode. This choice allowed users to perform as many squats as they wanted to in a given period or take their time and do a number of repetitions without a timer. The system tracked their movements and automatically counted each squat.

XSL•FO
RenderX

Figure 2. ADAPT XR (Advancing Digital Access to Physical Therapy via Virtual and Extended Reality Technology) virtual reality prototype menu.

XSL•FO
RenderX

Figure 3. ADAPT XR (Advancing Digital Access to Physical Therapy via Virtual and Extended Reality Technology) avatar and exercise demonstration.

Usability Evaluation

Six individuals were recruited for the in-person usability evaluation sessions with 30 to 45 minutes of VR exposure at the orthopedic clinic. Participants were mostly female (n=5, 83%) and aged >50 years (n=4, 67%; mean age 58.3, SD 11.7 y), and 100% (6/6) identified as non-Hispanic White. Overall, participants were enthusiastic about using the VR tool, noting that it was “fun” and “makes exercise less boring.” However, participants reported usability issues related to information needs, especially a lack of clarity on where to move, what to do next, and how many repetitions were completed. Usability issues were also identified concerning the use of the hand controllers and wearing the headset. Representative quotes for each of the 9 Health Information Technology Usability Evaluation Model concepts, along with their descriptions, are shown in Table 2.

The SUS scores were analyzed by averaging the responses from the 6 participants by question (Table 3) and calculating a total adjusted score. Using the Nielsen Norman Group usability guidelines [40], we found a total usability score of 58.3 for the VR prototype, indicating a below-average score [41].

XSL•FO
RenderX

Table . Health Information Technology Usability Evaluation Model concepts, descriptions, and representative quotes resulting from the ADAPT-XR (Advancing Digital Access to Physical Therapy via Virtual and Extended Reality Technology) virtual reality (VR) tool usability evaluation.

Concept	Description	Representative quote
Error prevention	The VR tool offers error management; error correction; or error prevention, such as instructions or reminders, to assist users performing tasks.	"Yes, it says to move to the area and then lights up a few seconds afterwards. It says move to highlighted area to continue and I'm right on it." [Neutral]
Completeness	The VR tool is able to assist users in successful completion of tasks.	"Congratulations, now it says I'm done, I didn't even get to do any squats." [Negative]
Memorability	Users can remember how to perform tasks in the VR tool easily.	"Okay I'm squatting this is about as far as I can go. Okay, I'm just going to keep going with him. He had a good suggestion so I like that oh I get it so he wants me to keep going." [Positive]
Information needs	The information offered by the VR tool is for basic task performance or to improve task performance.	"He is informative when he walked me through how to do, like it was helpful how he told me to keep my feet down, tuck in my hips so that was helpful." [Positive]
Flexibility and customizability	The VR tool provides more than one way to accomplish tasks, which allows users to operate the system as preferred.	"The instructions probably go a little fast because it's reading the text and I tend to read the text and not listen to the guy as much, and so the avatar was in front of the text and that was annoying." [Negative]
Learnability	Users are able to easily learn how to operate the VR tool.	"Okay this is the first problem I see a B but it really looks like it should be X so it's not as it's written left is X and right is Y and, on the thing, it shows X is below Y." [Negative]
Performance speed	Users are able to use the VR tool efficiently.	"While it was doing the tutorial, the avatar was showing you how to do it at the same time and then it went into a squat and then it gives you more instructions and you have to hold the squat for a long time. Like I was thinking…. Whoa…." [Negative]
Competency	Users are confident in their ability to perform tasks using the VR tool.	"Oh wow off balance, I don't think I can do that." [Negative]
Other outcomes	Other VR tool-specific outcomes such as user expectations.	"This would be good for patients out of state." [Positive]

Table . Means and SDs of System Usability Scale (SUS) responses (5-point Likert scale: 1="strongly disagree," 2="disagree," 3="neither agree nor disagree," 4="agree," and 5="strongly agree") for the ADAPT-XR (Advancing Digital Access to Physical Therapy via Virtual and Extended Reality Technology) virtual reality (VR) tool usability evaluation.

SUS item	Score, mean (SD)
"I think that I would like to use this VR tool frequently."	3.83 (1.07)
"I found the VR tool unnecessarily complex."	3.00 (1.00)
"I thought the VR tool was easy to use."	3.17 (0.90)
"I think that I would need the support of a technical person to be able to use this VR tool."	3.00 (1.15)
"I found the various functions of this VR tool were well integrated."	4.17 (0.37)
"I thought there was too much inconsistency in this VR tool."	2.67 (1.11)
"I would imagine that most people would learn to use this VR tool very quickly."	3.33 (1.37)
"I found the VR tool very cumbersome to use."	3.00 (1.15)
"I felt very confident using the VR tool."	3.17 (1.07)
"I needed to learn a lot of things before I could get going with this VR tool."	2.67 (1.11)

XSL•FO
RenderX

Discussion

Principal Findings

In this study, we identified several challenges while evaluating the usability of a prototype version of the ADAPT XR VR tool. Although some participants gave positive feedback, we found usability issues related to operating the device and navigating the physical space. As the VR environment requires boundaries to enable user movement while maintaining immersion, settings must be applied in advance of use. However, the boundaries in the VR prototype did not match the physical space available for the evaluation sessions. Additionally, the boundaries were overly restrictive, causing participant discomfort and frustration with where to move. Future iterations need to address these boundary issues to ensure accurate usability results as well as participant safety.

Usability was also impacted by participants' overall familiarity with VR. They found the headset difficult to adjust and uncomfortable to wear. Additionally, there were challenges in using the hand controls and selecting menu options. This was reflected in the SUS scores, which indicated the complexity of the VR tool and the need for significant learning before use. Evaluating the usability of VR tools can be complex given the multiple hardware and software components. However, our usability score of 58.3 is in line with other technology prototype scores, and we believe that it provides an adequate baseline to improve upon [42]. Future evaluations would benefit from offering a general orientation to VR hardware and environments before testing a particular application.

Comparison to Prior Work

To enhance usability during the development process, we used design box participatory methods to investigate user expectations alongside a usability evaluation of the initial prototype. While the design and development of the exercises were feasible, the usability aspects required more attention than anticipated. As found in other studies, human factors issues, such as poor interfaces, lack of user feedback, and physical limitations, are significant barriers to acceptability in patient-facing technologies and can negatively impact health outcomes [43]. Additionally, ergonomic issues have been cited in VR applications for office work and teaching [44,45]. Canniff and Cliburn [44] reported discomfort with the headset as the most important finding of their evaluation of the Oculus Quest with students. To our knowledge, this is the first study to report similar issues in a health care context. Although we acknowledge that some of the usability issues are limitations of the current VR technology available, in future testing, we intend to place a focus on the physical aspects of the tool to understand the impact of comfort on continued use.

Future Directions

As this study evaluated our initial prototype design, we plan to conduct multiple rounds of prototype testing as we iteratively refine the tool, and we expect usability to improve with further rounds of refinement. We will focus on evaluating a single exercise, the leg squat, within the VR tool to gain detailed insights into its usability by incorporating feedback mechanisms to assess whether the exercise was done correctly and the individual's ability to balance. This targeted approach will allow us to make specific adjustments and better understand user experiences. Additionally, we will implement a comprehensive onboarding process to familiarize users with VR devices and environments before evaluating the specific application. We believe that onboarding will be critical for VR tool usability until VR devices become more intuitive and are widely adopted. We plan to further investigate how the use of VR tools may exacerbate disparities among different groups, such as technology-savvy individuals, rural communities, and those with a lower socioeconomic status. This will help us identify and mitigate potential barriers to access and usability. We are also interested in investigating how different age groups interact with the VR tool, identifying any unique challenges or preferences that may arise and ensuring that the tool is accessible and effective for users of all ages and abilities.

Limitations

The findings are constrained by the self-selection of patients interested in new technologies and a reliance on purposive sampling given the need to meet in person, both of which may not have resulted in a sample representative of the broader patient population. Consequently, the usability and acceptability of the technology might be overestimated as individuals with varied levels of technology savviness could encounter more difficulties. Future research should include a more diverse range of participants to ensure that the technology is accessible and effective for all users, particularly those from various care settings, rural areas, and non–English-speaking backgrounds. There may be potential biases as members of the research team also participated in development; however, this was important for the translation of the findings to the design of the prototype. In addition, this work did not explore whether the participants were executing the exercise correctly, which is planned as part of future enhancements. Future stages of the prototype should incorporate methods to evaluate the exercise intervention in the user's preferred setting (eg, at home), include perspectives from physical therapists, and measure patient outcomes compared to current PT methods.

Conclusions

In this study, we designed and evaluated the usability of a prototype for a VR-based PT intervention. While there was enthusiasm and interest in using a VR tool at home, participants were dissatisfied with the prototype's usability, particularly related to the device itself. Usability issues were connected to the need for more instructions and difficulties interacting with the VR device. The findings highlight the importance of understanding users' baseline knowledge of VR and creating effective onboarding materials. Future research should focus on extensive prototype testing and use multiple methods to evaluate the usability and acceptability of VR technology, as well as include an orientation to VR, as it was many participants' first experience with the technology. Overall, VR tools have the potential to expand care access to rural and underserved communities and increase patient autonomy, making health care more engaging.

XSL•FO
RenderX

Acknowledgments

The authors acknowledge the Software Development and Systems Design Core at the University of Utah for its services. This study was funded by the Ben B. and Iris M. Margolis Foundation.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to the presence of protected health information and restrictions under HIPAA (Health Insurance Portability and Accountability Act) regulations but are available from the corresponding author on reasonable request.

Authors' Contributions

RA, GB, JMF, and VLT conducted the design box sessions. RA, JF, and VLT conducted the usability evaluation. All authors participated in the writing of the manuscript and approved the final version.

Conflicts of Interest

None declared.

References

1. Phanse VA. Critical gaps in physical therapy in the United States of America: exploring the shortage and Schedule A classification. Prog Med Sci 2022;6(1):1-3. [doi: 10.47363/PMS/2024(8)E115]
2. Childs JD, Benz LN, Arellano A, Briggs AA, Walker MJ. Challenging assumptions about the future supply and demand of physical therapists in the United States. Phys Ther 2022 Jan 1;102(1):pzab239. [doi: 10.1093/ptj/pzab239] [Medline: 34662413]
3. Smith JM, Lee AC, Zeleznik H, et al. Home and community-based physical therapist management of adults with post-intensive care syndrome. Phys Ther 2020 Jul 19;100(7):1062-1073. [doi: 10.1093/ptj/pzaa059] [Medline: 32280993]
4. Fritz JM, Lane E, McFadden M, et al. Physical therapy referral from primary care for acute back pain with sciatica: a randomized controlled trial. Ann Intern Med 2021 Jan;174(1):8-17. [doi: 10.7326/M20-4187] [Medline: 33017565]
5. McKinney J, Kelm N, Windsor B, Keyser LE. Addressing health care access disparities through a public health approach to physical therapist practice. Phys Ther 2024 Oct 2;104(10):pzae136. [doi: 10.1093/ptj/pzae136] [Medline: 39288092]
6. O'Bright K, Peterson S. Physical therapists in primary care in the United States: an overview of current practice models and implementation strategies. Phys Ther 2024 Dec 6;104(12):pzae123. [doi: 10.1093/ptj/pzae123] [Medline: 39223935]
7. Giannitrapani KF, Ahluwalia SC, McCaa M, Pisciotta M, Dobscha S, Lorenz KA. Barriers to using nonpharmacologic approaches and reducing opioid use in primary care. Pain Med 2018 Jul 1;19(7):1357-1364. [doi: 10.1093/pm/pnx220] [Medline: 29059412]
8. Arnold E, La Barrie J, DaSilva L, Patti M, Goode A, Clewley D. The effect of timing of physical therapy for acute low back pain on health services utilization: a systematic review. Arch Phys Med Rehabil 2019 Jul;100(7):1324-1338. [doi: 10.1016/j.apmr.2018.11.025] [Medline: 30684490]
9. Davenport TE, DeVoght AC, Sisneros H, Bezruchka S. Navigating the intersection between persistent pain and the opioid crisis: population health perspectives for physical therapy. Phys Ther 2020 Jun 23;100(6):995-1007. [doi: 10.1093/ptj/pzaa031] [Medline: 32115638]
10. Pritchard KT, Baillargeon J, Lee WC, Raji MA, Kuo YF. Trends in the use of opioids vs nonpharmacologic treatments in adults with pain, 2011-2019. JAMA Netw Open 2022 Nov 1;5(11):e2240612. [doi: 10.1001/jamanetworkopen.2022.40612] [Medline: 36342717]
11. Pawassar CM, Tiberius V. Virtual reality in health care: bibliometric analysis. JMIR Serious Games 2021 Dec 1;9(4):e32721. [doi: 10.2196/32721] [Medline: 34855606]
12. Karlsson M, Bergenheim A, Larsson ME, Nordeman L, van Tulder M, Bernhardsson S. Effects of exercise therapy in patients with acute low back pain: a systematic review of systematic reviews. Syst Rev 2020 Aug 14;9(1):182. [doi: 10.1186/s13643-020-01412-8] [Medline: 32795336]
13. Ceradini M, Losanno E, Micera S, Bandini A, Orlandi S. Immersive VR for upper-extremity rehabilitation in patients with neurological disorders: a scoping review. J Neuroeng Rehabil 2024 May 11;21(1):75. [doi: 10.1186/s12984-024-01367-0] [Medline: 38734690]
14. Gumaa M, Rehan Youssef A. Is virtual reality effective in orthopedic rehabilitation? A systematic review and meta-analysis. Phys Ther 2019 Oct 28;99(10):1304-1325. [doi: 10.1093/ptj/pzz093] [Medline: 31343702]
15. Hoffman HG, Chambers GT, Meyer 3rd WJ 3rd, et al. Virtual reality as an adjunctive non-pharmacologic analgesic for acute burn pain during medical procedures. Ann Behav Med 2011 Apr;41(2):183-191. [doi: 10.1007/s12160-010-9248-7] [Medline: 21264690]

XSL•FO
RenderX

16. The Moran Company. Initial treatment intervention and average total Medicare A/B costs for FFS beneficiaries with an incident low back pain (lumbago) diagnosis in CY 2014. Alliance for Physical Therapy Quality and Innovation. 2017. URL: https://www.aptqi.com/Resources/documents/APTQI-Complete-Study-Initial-Treatment-Intervention-Lumbago-May-2017.pdf [accessed 2025-10-18]
17. Brea-Gómez B, Torres-Sánchez I, Ortiz-Rubio A, et al. Virtual reality in the treatment of adults with chronic low back pain: a systematic review and meta-analysis of randomized clinical trials. Int J Environ Res Public Health 2021 Nov 11;18(22):11806. [doi: 10.3390/ijerph182211806] [Medline: 34831562]
18. Bordeleau M, Stamenkovic A, Tardif PA, Thomas J. The use of virtual reality in back pain rehabilitation: a systematic review and meta-analysis. J Pain 2022 Feb;23(2):175-195. [doi: 10.1016/j.jpain.2021.08.001] [Medline: 34425250]
19. Alemanno F, Houdayer E, Emedoli D, et al. Efficacy of virtual reality to reduce chronic low back pain: proof-of-concept of a non-pharmacological approach on pain, quality of life, neuropsychological and functional outcome. PLoS One 2019 May 23;14(5):e0216858. [doi: 10.1371/journal.pone.0216858] [Medline: 31120892]
20. Fritz JM, Minick KI, Brennan GP, et al. Outcomes of telehealth physical therapy provided using real-time, videoconferencing for patients with chronic low back pain: a longitudinal observational study. Arch Phys Med Rehabil 2022 Oct;103(10):1924-1934. [doi: 10.1016/j.apmr.2022.04.016] [Medline: 35667399]
21. Fritz JM, Lane E, Minick KI, et al. Perceptions of telehealth physical therapy among patients with chronic low back pain. Telemed Rep 2021 Nov 3;2(1):258-263. [doi: 10.1089/tmr.2021.0028] [Medline: 34927165]
22. Seron P, Oliveros MJ, Gutierrez-Arias R, et al. Effectiveness of telerehabilitation in physical therapy: a rapid overview. Phys Ther 2021 Jun 1;101(6):pzab053. [doi: 10.1093/ptj/pzab053] [Medline: 33561280]
23. Leask CF, Sandlund M, Skelton DA, et al. Framework, principles and recommendations for utilising participatory methodologies in the co-creation and evaluation of public health interventions. Res Involv Engagem 2019 Jan 9;5:2. [doi: 10.1186/s40900-018-0136-9] [Medline: 30652027]
24. Desselle M, Holland L, McKittrick A, Altizer RJ, Gray P, Brown J. Augmenting the design box: virtual reality pain relief for australian burns survivors case study. Presented at: SeGAH 2020: IEEE 8th International Conference on Serious Games and Applications for Health 2020; Aug 12-14, 2020 URL: https://eprints.qut.edu.au/211601/ [accessed 2025-10-18]
25. Altizer RJ, Zagal JP, Johnson E, et al. Design box case study: facilitating interdisciplinary collaboration and participatory design in game development. Presented at: CHI PLAY'17: Extended Abstracts Publication of the Annual Symposium on Computer-Human Interaction in Play; Oct 15-18, 2017. [doi: 10.1145/3130859.3131333]
26. Glaser B, Strauss A. Discovery of Grounded Theory: Strategies for Qualitative Research: Routledge; 2017. [doi: 10.4324/9780203793206]
27. Beck K, Beedle M, van Bennekum A, et al. Manifesto for agile software development. Agile Manifesto. 2001. URL: https://agilemanifesto.org [accessed 2024-10-01]
28. Rose AF, Schnipper JL, Park ER, Poon EG, Li Q, Middleton B. Using qualitative studies to improve the usability of an EMR. J Biomed Inform 2005 Feb;38(1):51-60. [doi: 10.1016/j.jbi.2004.11.006] [Medline: 15694885]
29. Li AC, Kannry JL, Kushniruk A, et al. Integrating usability testing and think-aloud protocol analysis with "near-live" clinical simulations in evaluating clinical decision support. Int J Med Inform 2012 Nov;81(11):761-772. [doi: 10.1016/j.ijmedinf.2012.02.009] [Medline: 22456088]
30. Brooke J. SUS-a quick and dirty usability scale. In: Jordan PW, Thomas B, Weerdmeester BA, McClelland IL, editors. Usability Evaluation in Industry 1996:189-194. [doi: 10.1201/9781498710411]
31. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381. [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
32. Harris AM, Hicks LA, Qaseem A, High Value Care Task Force of the American College of Physicians and for the Centers for Disease Control and Prevention. Appropriate antibiotic use for acute respiratory tract infection in adults: advice for high-value care from the American College of Physicians and the Centers for Disease Control and Prevention. Ann Intern Med 2016 Mar 15;164(6):425-434. [doi: 10.7326/M15-1840] [Medline: 26785402]
33. Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: building an international community of software platform partners. J Biomed Inform 2019 Jul;95:103208. [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]
34. Guest G, Namey E, Chen M. A simple method to assess and report thematic saturation in qualitative research. PLoS One 2020 May 5;15(5):e0232076. [doi: 10.1371/journal.pone.0232076] [Medline: 32369511]
35. Nielsen J. Estimating the number of subjects needed for a thinking aloud test. Int J Hum Comput Stud 1994 Sep;41(3):385-397. [doi: 10.1006/ijhc.1994.1065]
36. Nielsen J. How many test users in a usability study? Nielsen Norman Group. 2012. URL: https://www.nngroup.com/articles/how-many-test-users/ [accessed 2025-10-18]
37. Virzi RA. Refining the test phase of usability evaluation: how many subjects is enough? Hum Factors 1992;34(4):457-468. [doi: 10.1177/001872089203400407]
38. Yen PY. Health information technology usability evaluation: methods, models, and measures [Dissertation]. : Columbia University; 2010 URL: https://www.proquest.com/openview/0df85be25e4479c55d4b72b070dc034c/1?pq-origsite=gscholar&cbl=18750&diss=y [accessed 2025-10-18]

XSL•FO
RenderX

39. Cho H, Yen PY, Dowding D, Merrill JA, Schnall R. A multi-level usability evaluation of mobile health applications: a case study. J Biomed Inform 2018 Oct;86:79-89. [doi: 10.1016/j.jbi.2018.08.012] [Medline: 30145317]
40. The System Usability Scale (SUS). Nielsen Norman Group. URL: https://www.nngroup.com/videos/system-usability-scale/ [accessed 2025-10-23]
41. Lewis JR, Sauro J. Item benchmarks for the system usability scale. J Usability Stud 2018;13(3):158-167. [doi: 10.5555/3294033.3294037]
42. Sutori S, Eliasson ET, Mura F, et al. Acceptability, usability, and insights into cybersickness levels of a novel virtual reality environment for the evaluation of depressive symptoms: exploratory observational study. JMIR Form Res 2025 Apr 16;9:e68132. [doi: 10.2196/68132] [Medline: 40238239]
43. Carayon P, Hoonakker P. Human factors and usability for health information technology: old and new challenges. Yearb Med Inform 2019 Aug;28(1):71-77. [doi: 10.1055/s-0039-1677907] [Medline: 31419818]
44. Canniff K, Cliburn DC. Teaching virtual reality in virtual reality. Presented at: 8th International Conference of the Immersive Learning Research Network (iLRN); May 30 to Jun 4, 2022. [doi: 10.23919/iLRN55037.2022.9815930]
45. Kim E, Shin G. User discomfort while using a virtual reality headset as a personal viewing system for text-intensive office tasks. Ergonomics 2021 Jul;64(7):891-899. [doi: 10.1080/00140139.2020.1869320] [Medline: 33357004]

Abbreviations

ADAPT XR: Advancing Digital Access to Physical Therapy via Virtual and Extended Reality Technology
PT: physical therapy
REDCap: Research Electronic Data Capture
SUS: System Usability Scale
VR: virtual reality

Edited by A Mavragani, J Sarvestan; submitted 18.03.25; peer-reviewed by B Bonroy, C Hess; revised version received 28.09.25; accepted 29.09.25; published 05.12.25.

Please cite as:
Tiase VL, Fritz JM, Ferraro J, Bayles G, Alsaleem A, Del Fiol G, Kawamoto K, Brodke D, Martin B, Altizer R
Advancing Digital Access to Physical Therapy via Virtual and Extended Reality Technology: Prototype Development and Usability Evaluation
JMIR Form Res 2025;9:e73783
URL: https://formative.jmir.org/2025/1/e73783
doi:10.2196/73783

© Victoria Lynn Tiase, Julie Fritz, Jesse Ferraro, Gregory Bayles, Ahmad Alsaleem, Guilherme Del Fiol, Kensaku Kawamoto, Darrel Brodke, Brook Martin, Roger Altizer. Originally published in JMIR Formative Research (https://formative.jmir.org), 5.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Understanding Physician Attitudes Toward AI in Clinical Decision-Making: Cross-Sectional Study

Fahad Alhazmi, PhD

Department of Health Services and Hospitals Administration, Faculty of Economics and Administration, King Abdulaziz University, Jeddah, Saudi Arabia

Corresponding Author:
Fahad Alhazmi, PhD
Department of Health Services and Hospitals Administration, Faculty of Economics and Administration, King Abdulaziz University, Jeddah, Saudi Arabia

Abstract

Background: The Kingdom of Saudi Arabia (KSA) has made tremendous efforts to promote the adoption of advanced technologies such as artificial intelligence (AI). While the successful adoption of AI is dependent on physician perception, there is a scarcity of data concerning KSA physicians' perception of the technology.

Objective: The purpose of this study was to conduct a cross-sectional survey that would provide updated statistics on physicians' attitudes toward AI with a focus on ethical and practical perspectives among physicians licensed in the KSA.

Methods: A pilot study was conducted with 10 physicians to enhance the clarity of the survey questions. The pilot was followed by a cross-sectional survey, which was conducted through 25 online, self-administered questionnaires hosted on Qualtrics. A total of 218 physicians filled out the survey. The dataset was then exported into Microsoft Excel and analyzed using descriptive statistics rather than inferential analyses given the exploratory nature of this study and its primary aim to generate updated descriptive insights rather than test specific hypotheses.

Results: A total of 201 fully filled surveys, representing 127 (63.2%) female and 74 (36.8%) male physicians with experience ranging from 3 to ≥30 years, were analyzed. Most physicians (n=165, 82.1%) trusted AI-based clinical decision-making, and 76.6% (n=154) believed that the technology improved efficiency in health care delivery. Unfortunately, only 25.9% (n=52) of physicians had used AI in the previous year. Common barriers to AI adoption included lack of training, high implementation costs, and resistance to change, as well as concerns related to privacy, data security, bias in AI-based recommendations, patient autonomy, and liability. Participants recommended training through workshops (n=50, 25%), online courses (n=47, 23.4%), hands-on experience (n=44, 21.9%), and a combination of online courses and hands-on experience (n=17, 8.5%).

Conclusions: Physicians who responded to this survey supported AI's use in health care but reported facing financial, ethical, and training barriers, which could be addressed through informed consent and staff training.

(JMIR Form Res 2025;9:e79730) doi:10.2196/79730

KEYWORDS

artificial intelligence; AI; health care technology; physicians; perception; barriers; solutions

Introduction

Overview and Background

Artificial intelligence (AI) is an emerging technology that has gained widespread popularity in various sectors. AI is described as a powerful and disruptive technology that facilitates the use of computing devices and augmented systems to simulate critical thinking and intelligent behavior with the help of machine learning [1]. The Kingdom of Saudi Arabia (KSA) is among the high-income countries that have made efforts to tap into the power of AI in health care. Notably, the kingdom is currently undergoing substantial health care reforms that include embracing new technological advancements such as AI, whose integration is considered critical toward achieving its goals [2]. In this case, the adoption of AI aligns with the KSA's emphasis on the integration of technological advancements and innovation and is consistent with the Ministry of Health's support for AI-driven projects and efforts to pilot AI applications in clinical settings through collaboration with both local and international technology firms [2]. The KSA's commitment to AI is evident in initiatives such as Seha Virtual Hospital—currently the world's largest virtual hospital—which have tapped into the power of AI to overcome geographical barriers and resource constraints in the delivery of care [3,4]. Similarly, King Faisal Specialist Hospital and Research Centre has demonstrated the kingdom's commitment to AI with its development of several AI applications intended to contribute toward the goal of leveraging the power of AI to improve health outcomes through the Vision 2030 goals [2,5]. To reaffirm its commitment to AI

XSL•FO
RenderX

and other advanced technologies, the KSA has established the Saudi Data and AI Authority, which establishes governance in AI and data in collaboration with critical sectors such as education and health care [6]. Despite these efforts by various stakeholders in the KSA's health care sector, successful implementation of AI heavily depends on positive perception among physicians and other clinicians. The aim of this study was to explore physicians' attitudes toward AI with a focus on ethical and practical perspectives among physicians licensed in the KSA.

Problem Statement

There is a growing consensus among scholars that successful implementation of AI is dependent on scholars' perceptions and acceptance of this technology [7-9]. Recent studies suggest that, while many health care professionals in the KSA acknowledge AI's potential benefits, concerns persist regarding the impact of this technology on patient safety and quality health outcomes, job displacement, data reliability, and the ethical implications resulting from AI-driven decision-making. In terms of perception and acceptance, the KSA's physicians have shown a polarized attitude toward AI and its adoption in health care. For instance, professional health care organizations such as the American Heart Association have endorsed AI technologies such as ChatGPT after finding that the technology is 84% accurate, but the same technology has failed the accuracy test in some Asian countries, such as Taiwan [10]. At the local level, 76% of physicians in Jeddah, KSA, agreed that AI was accurate, and 69% of these participants agreed that the technology enhanced efficiency in health care delivery [8]. Surprisingly, only 25.9% reported using this technology in the previous year [8]. Similarly, 61.6% of physicians from the KSA who used AI agreed that the technology was useful, whereas 62.7% and 61.8% agreed that it enhanced efficiency and lowered the cost of care, respectively [7]. However, one study found that the adoption of AI was relatively higher among KSA physicians compared to non-KSA physicians at 55.7% versus 44.3% [7]. A study reported that 61.1% of KSA physicians expressed a positive perception of the integration of AI in health care. These physicians also cited ethical concerns such as job replacement, data privacy, and concerns about the accuracy of AI systems in clinical decision-making [7-9,11]. In contrast, one study found that only 26.2% of KSA physicians had a positive perception of AI and that only 12.3% of these professionals had the technical expertise to use these technologies in their clinical practice [11]. The inconsistency in these findings concerning the desirability of AI among KSA physicians shows a major gap in the literature. This gap warrants further studies to establish updated statistics on physicians' attitude toward AI in the KSA and possible factors that influence this attitude. A study that focuses on ethical and practical issues could address this gap by revealing diverse concerns among clinicians.

Purpose

The purpose of this study was to conduct a cross-sectional survey that would provide updated statistics on physicians' attitude toward AI with a focus on ethical and practical perspectives among physicians licensed in the KSA. This study will add to existing literature by clarifying the current misunderstanding about physicians' attitude toward AI, revealing possible contributors toward gaps in AI acceptance, and providing evidence that could guide future initiatives to promote the acceptance of AI among physicians. This study will also contribute toward the KSA's efforts to become a global leader in AI and other advanced technologies as part of its Vision 2030 initiative [2,6].

Methods

Design

A cross-sectional survey was conducted to establish physicians' attitude toward AI and its integration into the KSA's health care sector. This design is considered suitable for exploratory studies, such as obtaining participants' insights about a phenomenon, due to its ease of implementation and collection of numerical data from large samples [12]. The survey was conducted online through self-administered, non–AI-generated questionnaires.

Population and Sample

This study targeted physicians licensed in the KSA. Physicians are the primary users of clinical decision support systems and are directly impacted by the ethical and practical challenges of AI integration in health care [13]. Given their central role in health care delivery and decision-making, selecting physicians as the study population was appropriate as their attitudes are crucial to the successful adoption and ethical implementation of AI technologies in Saudi Arabia's health sector.

Ethical Considerations

This study was approved by the King Abdulaziz University Institutional Review Board, NCBE Registration No: (HA-02-J-008), and adhered to the provisions of the Declaration of Helsinki. This study also respected participants' autonomy and confidentiality by assuring them that participation was voluntary. Before the survey, all participants provided informed consent electronically via a consent form that outlined the study's purpose, procedures, confidentiality assurances, and participants' right to withdraw at any time without penalty. This study posed no risk to participants as it was conducted on physicians through an online survey. No identifying information such as names, addresses, or institutional affiliations was collected to protect privacy and confidentiality. All responses were collected anonymously, and the data were stored on a password-protected device that was accessible only to the principal investigator. No compensation was provided to participants.

Data Collection and Instruments

A survey of 25 questions was selected for use in this study. This tool was considered suitable for this research as it helped generate valid findings (Cronbach α=0.80) through the use of a sufficiently large sample and maintaining the P value at less than .05 [14]. Other than their ease of administration, cost-effectiveness, and ability to reach a large population, questionnaires were selected as they could yield results with high validity and reliability as long as response rates reached or exceeded 40% [12]. The researchers started by creating a questionnaire comprising 7 questions to collect demographic

XSL•FO
RenderX

data and an additional 18 questions to collect data on physicians' attitudes toward AI in health care.

The questionnaire was first disseminated to 10 physicians from the KSA to help identify potential issues with the questions and the methodology. Once the questions were refined, a message was disseminated through posters in local hospitals and online through popular social media platforms (Facebook, X [formerly known as Twitter], Instagram, and WhatsApp) inviting physicians from the KSA to participate in the survey. Participants were assured that their participation would be voluntary and anonymous and that the findings would be used purely for research purposes. A link to an online survey hosted on Qualtrics (Qualtrics International Inc) was included in this message. Participants filled out the survey between February 2025 and April 2025, and the data were downloaded and compiled into spreadsheets. Descriptive statistics, pivot tables, and pivot charts were generated to analyze and interpret the data.

This study adhered to the Checklist for Reporting Results of Internet E-Surveys (CHERRIES), which provides standardized guidance for reporting the design, conduct, and analysis of web-based surveys. Following the CHERRIES framework ensured transparency in describing the survey development process, recruitment procedures, data protection measures, and handling of responses. The completed CHERRIES checklist is provided as a supplementary file (Checklist 1).

The dataset was then exported from Qualtrics into Microsoft Excel and analyzed using descriptive statistics. Inferential analyses were not conducted given the exploratory nature of this study and its primary aim to generate updated descriptive insights into physician attitudes rather than test specific hypotheses. This analytical approach provided a clear overview of prevailing trends and concerns among physicians in the KSA.

Results

Demographics

There were a total of 218 surveys filled out during the data collection period. This study had a completion rate of 92.2% (201/218) after removing surveys with blank answers. Table 1 summarizes the demographics of the sample. Women constituted 63.2% (127/201) of the participants and men constituted the remaining 36.8% (74/201). A total of 9.5% (19/201) of the respondents were aged 18 to 25 years, whereas 24.9% (50/201), 28.9% (58/201), 27.9% (56/201), and 9% (18/201) of the respondents were aged 26 to 35 years, 36 to 45 years, 46 to 59 years, and ≥60 years, respectively. Additionally, 12.4% (25/201) of the participants had a PhD, whereas 20.4% (41/201), 9.5% (19/201), and 32.8% (66/201) had a master's degree, a bachelor's degree, and an associate degree, respectively. In total, 14.9% (30/201) of the participants had worked as physicians for 3 to 5 years, whereas 30.3% (61/201), 29.4% (59/201), 14.9% (30/201), and 10% (20/201) had worked as physicians for 6 to 10 years, 11 to 20 years, 21 to 30 years, and over 30 years, respectively.

XSL•FO
RenderX

Table . Demographic characteristics of the physicians who participated in a cross-sectional online survey conducted in Saudi Arabia (N=201).

Education level and age group (years)	Years of experience					
	3 - 5 years	6 - 10 years	11 - 20 years	21 - 30 years	Over 30 years	Total
Board certified (n=33, 16.42%)						
>60	—[a]	—	—	1 (0.50)	8 (3.98)	9 (4.48)
26 - 35	2 (1.00)	—	—	—	—	2 (1.00)
36 - 45	—	1 (0.50)	4 (1.99)	—	—	5 (2.49)
46 - 59	—	—	5 (2.49)	11 (5.47)	1 (0.50)	17 (8.46)
Diploma (n=19, 9.45%)						
18 - 25	15 (7.46)	—	—	—	—	15 (7.46)
26 - 35	3 (1.49)	1 (0.50)	—	—	—	4 (1.99)
Fellow (n=15, 7.46%)						
>60	—	—	—	—	6 (2.99)	6 (2.99)
26 - 35	1 (0.50)	3 (1.49)	—	—	—	4 (1.99)
36 - 45	—	1 (0.50)	1 (0.50)	—	—	2 (1.00)
46 - 59	—	—	2 (1.00)	—	1 (0.50)	3 (1.49)
Master (n=66, 32.84%)						
18 - 25	1 (0.50)	—	—	—	—	1 (0.50)
26 - 35	—	18 (8.96)	—	—	—	18 (8.96)
36 - 45	—	17 (8.46)	18 (8.96)	1 (0.50)	—	36 (17.91)
46 - 59	—	1 (0.50)	6 (2.99)	4 (1.99)	—	11 (5.47)
MBBS[b] (n=27, 13.43%)						
18 - 25	3 (1.49)	—	—	—	—	3 (1.49)
26 - 35	5 (2.49)	15 (7.46)	2 (1.00)	—	—	22 (10.95)
36 - 45	—	1 (0.50)	—	—	—	1 (0.50)
46 - 59	—	1 (0.50)	—	—	—	1 (0.50)
PhD (n=41, 20.40%)						
>60	—	—	—	1 (0.50)	2 (1.00)	3 (1.49)
36 - 45	—	2 (1.00)	9 (4.48)	3 (1.49)		14 (6.97)
46 - 59	—	—	13 (6.47)	9 (4.48)	2 (1.00)	24 (11.94)

[a]The dashes represent cells with a value of zero.

[b]MBBS: bachelor of medicine, bachelor of surgery. Now

Physicians' Trust in AI Recommendations for Clinical Decision-Making

The findings further revealed that 82.1% (165/201) of the physicians trusted AI-based recommendations in clinical decision-making. This number included 49.3% (99/201) who strongly trusted these systems and 32.8% (66/201) who somewhat partially trusted AI in clinical decision-making. However, 63.7% (128/201) of the physicians believed that AI will always require human oversight in clinical decision-making.

Physicians' Perception of AI's Impact on Clinical Outcomes

Some participants had mixed perceptions of the impact of AI on clinical outcomes. For instance, 56.7% (114/201) of the physicians believed that AI will improve patient outcomes in the long run in the KSA's health care sector. However, only 5.5% (11/201) of the physicians strongly agreed that AI will improve patient outcomes in the long run, and 52.2% (105/201) of physicians believed that AI could negatively impact physician-patient relationships during the delivery of care.

XSL•FO
RenderX

Physicians' Concerns About Ethical Issues Surrounding AI

Most physicians expressed concern about ethical issues surrounding the use of AI in the health care sector. For instance, 54.7% (110/201) were concerned that AI could introduce bias in the diagnosis and treatment of patients. However, 38.8% (78/201) believed that physicians should be held accountable for errors that occur due to the use of AI. Furthermore, 52.2% (105/201) of clinicians expressed concerns about AI systems' ability to protect patient privacy and data security. As a result, 85.1% (171/201) of the physicians agreed that patients must always be informed when AI systems are involved in diagnosis and treatment. Moreover, 33.8% (68/201) of the respondents cited bias as a key ethical concern, whereas 25.4% (51/201), 20.9% (42/201), and 19.9% (40/201) cited privacy, patient autonomy, and liability, respectively, as some of the ethical concerns that should be addressed to enhance the adoption of AI.

Informed consent was the most dominant concern among respondents. For instance, 70.1% (141/201) of the participants reported facing barriers to obtaining patient consent before the use of AI. On the same note, 89.6% (180/201) agreed that patients' informed consent was important during AI-based clinical decision-making, whereas 82.1% (165/201) suggested that the informed consent form should mention any clinical recommendations that were made using AI. Furthermore, only 62.2% (125/201) acknowledged that their patients fully understood the consent forms that they signed during treatment. Consequently, 91.5% (184/201) of the physicians always discussed alternatives to AI with their patients before obtaining informed consent to use AI. Many physicians mentioned barriers that hindered the adoption of AI in the KSA's health care sector. Table 2 illustrates the combinations of barriers that physicians mentioned during the survey.

Participants were presented with a multiple-response ("select all that apply") question listing common barriers to AI adoption, such as cost, data privacy, lack of training, and resistance to change. Respondents could choose more than one option, and the combinations shown in Table 2 reflect the various sets of barriers selected by each physician.

Table . Reported barriers to artificial intelligence adoption among physicians participating in a cross-sectional online survey in Saudi Arabia (N=201).

Barriers	Participants, n (%)
Cost	25 (12.4)
Cost and data privacy	6 (3.0)
Cost and resistance to change	2 (1.0)
Data privacy	39 (19.4)
Data privacy and resistance to change	3 (1.5)
Lack of training	46 (22.9)
Lack of training and cost	7 (3.5)
Lack of training, cost, and data privacy	2 (1.0)
Lack of training, cost, data privacy, and resistance to change	5 (2.5)
Lack of training and data privacy	5 (2.5)
Lack of training, data privacy, and resistance to change	5 (2.5)
Lack of training and resistance to change	16 (8.0)
Resistance to change	40 (19.9)

Physicians' Perception of AI's Impact on Efficiency

A total of 76.6% (154/201) of the physicians believed that AI will make health care more efficient. In contrast, 14.4% (29/201) of the respondents believed that AI will complicate health care delivery, whereas 9% (18/201) believed that it will have both positive and negative implications on efficiency. However, all physicians agreed that AI could have widespread applications in the KSA's health care sector over the next decade. Some of the applications listed by physicians included assisting physicians, automation, and replacement of certain tasks.

Physicians' Recommendations on Training Needs for AI

Clinicians offered several recommendations about the training needed to successfully integrate AI into the KSA's health care sector. A quarter of the respondents (50/201, 25%) mentioned workshops, followed by online courses (47/201, 23.4%), hands-on experience (44/201, 21.9%), and a combination of online courses and hands-on experience (17/201, 8.5%).

Discussion

Principal Findings

The data generated from the surveys present an insight into KSA physicians' perception of the adoption of AI in health care, as well as ethical concerns that these professionals have regarding the application of this technology. Physicians' recognition of the wide applications of AI is consistent with evidence from the literature. For instance, studies show that AI has vast applications in critical areas such as education, manufacturing, finance, and health care, among others, with the latter being cited as among the latest adopters of this technology

XSL•FO
RenderX

[15]. Similar to what KSA physicians reported, AI has been widely applied in the analysis of laboratory tests and diagnostic imaging results, genetic testing diagnosis of diseases, discovery of new treatments, selection of appropriate treatments, robotic surgery, patient monitoring, and various areas of health care that require advanced decision-making [1,15-17]. Patterns in the responses further reflected the ongoing trends in the adoption of AI, which has been considerably higher in high-income countries such as the KSA. Scholars agree that the positive perception may be due to higher access to knowledge, technical expertise, and the infrastructure required to implement the technology [1,18]. The positive perception of AI among KSA physicians may also be explained by the popularization, increased support, and promotion of this technology by the government [2,6]. As such, the adoption of this technology is expected to rise in the coming years.

The hesitancy among KSA physicians to support AI is not surprising considering that this technology is still in its early stages of development. As per the above findings, more physicians from the KSA support the use of AI compared to physicians in specific localities in the country such as Jeddah, the wider Asian region, and parts of the West such as the United States [7,8,10,11]. However, ethical concerns such as patient privacy, data security, liability, patient autonomy, bias in recommendations, and job loss are consistent with concerns documented in previous studies focusing on the KSA, as well as those that investigated the adoption of AI in other parts of the world [7-9,11]. However, the technological revolution being witnessed in the KSA could necessitate the development of AI systems personalized to the needs of the local market to help address these concerns. Nevertheless, addressing the barriers that KSA physicians mentioned, such as training needs, financial constraints, and resistance to change, could ensure that AI gains 100% support among clinicians and optimize the technology's benefits to patients.

Study Implications

This study contributes to the KSA's health care sector as it presents updated statistics on physicians' perception of AI. These findings will contribute toward the kingdom's efforts to become a leader in the adoption of advanced technologies such as AI as per its Vision 2030 initiative [2,6]. At the same time, the findings could help other countries striving to adopt AI by highlighting the barriers that they could encounter during the transition and some solutions that they could leverage to achieve success from raising awareness of the benefits of AI and training staff members. Nevertheless, this study underscores the need for comprehensive policy frameworks that balance innovation with ethical responsibility by identifying physicians' perceived ethical barriers and readiness gaps.

Study Limitations

This study has some limitations that could affect its application. First, the study's small sample could affect the validity and generalizability of its findings. Second, the sample was conveniently selected from local hospitals, which could create a risk of selection bias and limit the generalizability of the findings to the entire KSA health care system. Third, the selection of participants from local health care facilities and the use of self-reported surveys raise the risk of self-selection and social desirability bias, which could impact the reliability of the findings. As such, future studies should apply more robust sample selection methods and data collection tools. Fourth, this study suffers from methodological limitations as it was a correlational study that did not establish the actual causes of physicians' attitude toward AI. Fifth, the application of this study's findings may be limited, particularly in low- and middle-income nations that experience different challenges from those reported in the KSA. As such, larger studies need to be performed to provide more insights about physicians' attitude toward AI adoption in health care, factors behind this attitude, and ways to address current concerns. Finally, inferential analyses were not conducted given the exploratory nature of this study and its primary aim to generate updated descriptive insights into physician attitudes rather than test specific hypotheses. As such, future research should apply inferential and multivariate analyses such as logistic regression or structural equation modeling to explore associations among demographic variables, attitudes, and ethical perceptions to facilitate more robust generalizations [12]. Despite these limitations, this study provides a foundational overview that future researchers can build upon while applying more sophisticated analytical models and larger, stratified samples.

Conclusions

This study explored current attitudes toward AI with a focus on ethical and practical perspectives among physicians licensed in the KSA using a cross-sectional survey design. The findings showed that 82.1% (165/201) of KSA physicians trusted AI-based clinical decision-making and 76.6% (154/201) believed that the technology will improve efficiency in health care delivery. Unfortunately, most physicians (195/201, 97%) reported encountering barriers such as lack of training; high implementation costs; resistance to change; and ethical issues such as privacy, data security, bias in AI-based recommendations, patient autonomy, and liability. Consequently, only 25.9% (52/201) of the physicians reported using AI in the previous year. However, these physicians offered solutions such as informed consent and staff training to address the current barriers to AI adoption in health care.

Acknowledgments

This research was supported by King Abdulaziz University. The author extends their sincere gratitude to the institution for its financial and academic support. This study is solely the work of the author, and no external reviewers contributed to the development of the manuscript before submission.

XSL•FO
RenderX

Funding

The funding provided by King Abdulaziz University will only cover the article processing charges.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Checklist 1
CHERRIES checklist.
[DOCX File, 16 KB - formative_v9i1e79730_app1.docx]

References

1. Bajwa J, Munir U, Nori A, Williams B. Artificial intelligence in healthcare: transforming the practice of medicine. Future Healthc J 2021 Jul;8(2):e188-e194. [doi: 10.7861/fhj.2021-0095] [Medline: 34286183]
2. Suleiman AK, Ming LC. Transforming healthcare: Saudi Arabia's vision 2030 healthcare model. J Pharm Policy Pract 2025 Jan 21;18(1):2449051. [doi: 10.1080/20523211.2024.2449051] [Medline: 39845746]
3. Alkhalifah JM, Seddiq W, Alshehri BF, Alhaluli AH, Alessa MM, Alsulais NM. The role of the COVID-19 pandemic in expediting digital health-care transformation: Saudi Arabia's experience. Inform Med Unlocked 2022;33:101097. [doi: 10.1016/j.imu.2022.101097] [Medline: 36185732]
4. Alsanosi SM, Padmanabhan S. Potential applications of artificial intelligence (AI) in managing polypharmacy in Saudi Arabia: a narrative review. Healthcare (Basel) 2024 Apr 5;12(7):788. [doi: 10.3390/healthcare12070788] [Medline: 38610210]
5. Habibzadeh F. GPTZero performance in identifying artificial intelligence-generated medical texts: a preliminary study. J Korean Med Sci 2023 Sep 25;38(38):e319. [doi: 10.3346/jkms.2023.38.e319] [Medline: 37750374]
6. Memish ZA, Altuwaijri MM, Almoeen AH, Enani SM. The Saudi Data and Artificial Intelligence Authority (SDAIA) vision: leading the kingdom's journey toward global leadership. J Epidemiol Glob Health 2021 Jun;11(2):140-142. [doi: 10.2991/jegh.k.210405.001] [Medline: 33876596]
7. Ahmed A, Fatani D, Vargas JM, et al. Physicians' perspectives on ChatGPT in ophthalmology: insights on artificial intelligence (AI) integration in clinical practice. Cureus 2025 Jan 27;17(1):e78069. [doi: 10.7759/cureus.78069] [Medline: 40013176]
8. Alkhatieb M, Subke AA. Artificial intelligence in healthcare: a study of physician attitudes and perceptions in Jeddah, Saudi Arabia. Cureus 2024 Mar 30;16(3):e57256. [doi: 10.7759/cureus.57256] [Medline: 38686273]
9. Syed W, Babelghaith SD, Al-Arifi MN. Assessment of Saudi public perceptions and opinions towards artificial intelligence in health care. Medicina (Kaunas) 2024 Jun 4;60(6):938. [doi: 10.3390/medicina60060938] [Medline: 38929555]
10. Abouammoh N, Alhasan K, Aljamaan F, et al. Perceptions and earliest experiences of medical students and faculty with ChatGPT in medical education: qualitative study. JMIR Med Educ 2025 Feb 20;11:e63400. [doi: 10.2196/63400] [Medline: 39977012]
11. Sandougah K, Alsywina N, Alghamdi HK, Alrobayan F, Alshaghathirah FM, Alomran K. Assessment of the knowledge, attitude, and practice of artificial intelligence (AI) among radiologists, physicians, and surgeons in Saudi Arabia. Cureus 2025 Feb 17;17(2):e79180. [doi: 10.7759/cureus.79180] [Medline: 40109795]
12. Gray JR, Grove SK. Burns and Grove's the Practice of Nursing Research: Appraisal, Synthesis, and Generation of Evidence: Elsevier Science; 2020.
13. Alnowibet K, Abduljabbar A, Ahmad S, et al. Healthcare human resources: trends and demand in Saudi Arabia. Healthcare (Basel) 2021 Jul 29;9(8):955. [doi: 10.3390/healthcare9080955] [Medline: 34442091]
14. Munis Ö, Parlak MM. Validity and reliability of the Turkish version of the Holden Communication Scale. Brain Behav 2023 Sep;13(9):e3223. [doi: 10.1002/brb3.3223] [Medline: 37587848]
15. Pattnaik D, Ray S, Raman R. Applications of artificial intelligence and machine learning in the financial services industry: a bibliometric review. Heliyon 2023 Dec 13;10(1):e23492. [doi: 10.1016/j.heliyon.2023.e23492] [Medline: 38187262]
16. Alowais SA, Alghamdi SS, Alsuhebany N, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ 2023 Sep 22;23(1):689. [doi: 10.1186/s12909-023-04698-z] [Medline: 37740191]
17. Spear J, Ehrenfeld JM, Miller BJ. Applications of artificial intelligence in health care delivery. J Med Syst 2023 Nov 17;47(1):121. [doi: 10.1007/s10916-023-02018-y] [Medline: 37975946]
18. Zuhair V, Babar A, Ali R, et al. Exploring the impact of artificial intelligence on global health and enhancing healthcare in developing nations. J Prim Care Community Health 2024;15:21501319241245847. [doi: 10.1177/21501319241245847] [Medline: 38605668]

XSL•FO
RenderX

Abbreviations

AI: artificial intelligence
CHERRIES: Checklist for Reporting Results of Internet E-Surveys
KSA: Kingdom of Saudi Arabia

Edited by A Mavragani; submitted 26.06.25; peer-reviewed by HA Fadil, WB Lee; revised version received 05.11.25; accepted 24.11.25; published 08.12.25.

Please cite as:
Alhazmi F
Understanding Physician Attitudes Toward AI in Clinical Decision-Making: Cross-Sectional Study
JMIR Form Res 2025;9:e79730
URL: https://formative.jmir.org/2025/1/e79730
doi:10.2196/79730

© Fahad Alhazmi. Originally published in JMIR Formative Research (https://formative.jmir.org), 8.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Gender Differences in Trajectories of Depressive Symptoms Among Talkspace Clients: Naturalistic Observational Study

Meghan Romanelli[1], PhD, LCSW; Julien Rouvere[2], MA; Isaac A Sanders[3], MSW; Aarthi Padmanabhan[4], PhD; Emily Peake[4], MS; Thomas D Hull[4], PhD; Tim Althoff[5], PhD

[1]School of Social Work, University of Washington, 4101 15th Avenue NE, Seattle, WA, United States

[2]Department of Psychiatry and Behavioral Medicine, University of Washington School of Medicine, Seattle, WA, United States

[3]School of Social Work University of Washington, Seattle, WA, United States

[4]Talkspace, New York, NY, United States

[5]Allen School of Computer Science & Engineering, University of Washington, Seattle, WA, United States

Corresponding Author:
Meghan Romanelli, PhD, LCSW
School of Social Work, University of Washington, 4101 15th Avenue NE, Seattle, WA, United States

Abstract

Background: Gender minority populations experience an increased risk of depression and report significant barriers to accessing mental health services. While digital mental health (DMH) technologies may address barriers, it remains unclear how gender minority clients engage with DMH services and if DMH improves their clinical outcomes.

Objective: This naturalistic study explored gender differences in 15-week clinical outcomes of clients receiving technology-mediated psychotherapy from a large DMH provider.

Methods: This study used observational data of clients who signed up for Talkspace (Talkspace, Inc) between February 2017 and July 2021. The analytic sample included Talkspace clients (N=20,156) with a baseline 8-item Patient Health Questionnaire (PHQ-8) score ≥10. Participants completed at least 2 PHQ-8 assessments over 15 weeks of treatment. Multilevel linear models tested gender differences in depressive symptom trajectories over the course of treatment (model 1) while also controlling for baseline PHQ-8 scores (model 2) and treatment engagement indicators (model 3). Sensitivity analyses reestimated model 2 among clients who submitted a PHQ-8 survey during the week 15 assessment period and among those who discontinued treatment beforehand. Reasons for service cancellation were also described for the latter group. Gender differences in secondary clinical outcomes were examined via chi-square and Fisher exact tests.

Results: In all models, there were significant week-by-gender interactions. When controlling for baseline PHQ-8 scores, rates of symptom change were significantly slower for gender-diverse participants (b=0.60; P<.001), nonbinary participants (b=0.81; P<.001), and transgender women (b=0.87; P=.007), but not for women (P=.98) or transgender men (P=.38) compared to men. By week 15, adjusted PHQ-8 scores declined 8.7 points for both men and women, versus 4.4 - 7.4 points for gender minority clients. Sensitivity analyses indicated attenuated symptom improvement among week-15 completers, with transgender women showing the slowest changes (b=0.76; P=.02). Among earlier dropouts, weekly symptom reductions were steep overall (eg, week 3: b=−4.06, P<.001; week 6: b=−2.31, P<.001) while certain gender minority subgroups worsened (eg, adjusted scores for transgender women increased from 15.41 at baseline to 16.08 at final week 3 PHQ-8 survey submissions). Cancellation data (3450/20,156, 17.12%) confirmed discontinuation reasons related to both symptom improvement (928/3691 reasons, 25.14%) and potential barriers to treatment engagement (eg, cost: 1431/3691, 38.77%; poor service fit or poor perceived effectiveness: 677/3691, 18.34%). Gender differences were observed in rates of treatment response (weeks 3 - 12; all P≤.02), symptom remission (weeks 3, 6, 9, and 15; all P≤.047), and clinically significant symptom reduction (all time points, all P≤.03). Symptom deterioration did not differ by gender (all P>.05).

Conclusions: While clinical outcomes generally improved over time among clients engaged in technology-mediated psychotherapy, some gender minority populations experienced slower improvements. Future research may explore strategies to adapt DMH interventions to better meet the needs of diverse gender identities.

(JMIR Form Res 2025;9:e75290) doi:10.2196/75290

KEYWORDS

gender identity; telemedicine; digital mental health interventions; depression; anxiety

XSL•FO
RenderX

Introduction

Gender minority populations, including transgender, nonbinary, and gender-diverse individuals, experience increased risk of depression relative to cisgender counterparts due to chronic exposure to structural and interpersonal stressors [1,2]. As the leading cause of disability worldwide [3], depression is a serious public health concern. Recent estimates show that 332 million people are diagnosed with depressive disorder globally, resulting in extensive disease burden (ie, economic costs, social costs, morbidity, and mortality) [4-6]. While depression is one of the most common mental health problems, it is also one of the most treatable [7]. However, gender minority populations remain disproportionately impacted. Research indicates that gender minority individuals experience depression at higher rates than cisgender peers [8-11]. For example, in the United States, self-reported lifetime prevalence of depression among transgender populations is close to 30%, which far exceeds cisgender rates (18%) [12]. Electronic health records confirm this disparity with higher rates of depressive disorder diagnoses observed among gender minorities relative to cisgender counterparts (cisgender men: 16.8%; cisgender women: 23.7%; transgender men: 29.5%; transgender women: 27%; and nonbinary and gender diverse: 32.2%) [13].

Despite the increased need for mental health care, gender minority populations report significant barriers to accessing services [14,15]. For example, the availability of affirmative services is limited. Affirmative services are responsive to the unique stressors and needs of gender minority clients, providing care that is associated with greater treatment satisfaction and treatment effectiveness [16]. These services, however, are not only few in number but are also primarily clustered in urban enclaves, further restricting access to care for a geographically dispersed population [17,18]. Access barriers can result in forgone care, restrict treatment engagement, and lead to significant unmet mental health needs [19]. Indeed, gender minority populations face large treatment gaps. For example, research indicates that transgender people are 2.4 times more likely than their cisgender counterparts to report unmet mental health needs, including untreated depression [20].

Digital mental health (DMH) technologies are promising to address the barriers of traditional service delivery models that restrict marginalized populations' access to mental health resources [21-23]. DMH may offer flexible and remote treatment options [21] that allow clients to engage with licensed therapists through text, audio, or video messages [24]. Digital resources are an essential source of information and connection among gender minority populations, especially those who report isolation and discrimination [25]. Gender minority care seekers increasingly prefer to seek online support and professional health services [26-28], citing benefits of self-directed care, confidentiality, anonymity, and safety [23,27,29,30]. Therefore, DMH options present an acceptable mode of treatment delivery with service qualities that align with community needs and preferences. Despite potential, it remains unclear how gender minority consumers engage with DMH services and if DMH effectively improves their clinical outcomes [23,29,31].

This naturalistic study explored the gender differences in longitudinal treatment engagement and 15-week clinical outcomes among clients receiving technology-mediated psychotherapy from a large DMH provider. First, this study characterized clients and their engagement in care, including treatment retention and weekly message and word counts. Next, this study examined gender differences in trajectories of depressive symptoms over a 15-week period. Gender minority clients often face unmet mental health needs, but DMH may lead to better outcomes by reducing barriers and providing more affirming options. Given these competing possibilities, analyses were approached as exploratory and focused on potential heterogeneity in depressive symptom trajectories by gender. Ultimately, the research endeavored to provide insights that can inform the development and implementation of effective DMH services that address critical mental health and access disparities experienced by gender minority populations. This study was conducted following the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) reporting guidelines (Checklist 1) [32].

Methods

Design and Procedures

This study used longitudinal observational data of clients who signed up for Talkspace (Talkspace, Inc) between February 2017 and July 2021. Data included in this study were routinely collected as part of the service. Talkspace [33] is a telemedicine platform of over 5000 licensed therapists across the United States. Clients access the platform through the internet and complete a brief intake with a consulting clinician to identify their presenting problems and treatment history. The intake clinician also helps the client select their service plan (eg, asynchronous messaging only; asynchronous messaging plus monthly or weekly 30-minute synchronous video sessions) and records information about therapist preferences (eg, gender and language preferences). Using this information, a matching algorithm selects 3 clinicians whose expertise and availability best align with the client's preferences and treatment needs. Clients review clinician information and self-select a clinician for continued treatment. This process ensures agency and self-direction, principles that align with preferred service qualities identified by gender minority consumers [23,27,29,30].

Participants

All participants of the current study were Talkspace clients engaged in technology-mediated individual psychotherapy. Each participant completed at least two 8-item Patient Health Questionnaire (PHQ-8) assessments, including a baseline survey after clinician selection and, minimally, 1 additional PHQ-8 survey at a subsequent time point. Inclusion criteria included a baseline PHQ-8 score of ≥10, that clients were English-speaking, US residents, had internet or smartphone access, and provided information on gender. The final analytic sample included data from 20,156 participants. Participants worked with 3804 different therapists contracted or employed by Talkspace, all of whom were Master's level or higher, licensed clinicians.

XSL•FO
RenderX

Intervention

Therapists and clients exchanged asynchronous text, audio, or video-based messages depending on client preference, and if available through the service plan, attended weekly or monthly 30-minute live video sessions. Unlimited and 24/7 asynchronous messaging was available for all clients. Therapists reviewed client messages and responded at least once per day during standard workdays and hours. All messaging occurred over a secure, Health Insurance Portability and Accountability Act (HIPAA)–compliant platform that could be accessed on mobile and desktop devices.

Measures

Participant Characteristics: Gender

Demographic characteristics were collected at baseline. All clients included in the analytic sample reported their gender at sign-up by answering a single question, "What gender do you identify with?" (male, female, transgender male, transgender female, nonbinary, gender queer, gender variant, or other). Because gender was assessed with a single item, it cannot be definitively determined that clients who selected "male" or "female" identify as cisgender. Therefore, the term "cisgender" was avoided in reporting study results as it is often incorrectly applied as an assumed default label for individuals reporting their gender as male or female [34,35]. In this study, clients who indicated their gender as gender queer (142/20,156, 0.70%), gender variant (25/20,156, 0.12%), or "other" (44/20,156, 0.22%) are grouped as "gender diverse." This identification reflects terminology suggested by the National Academies of Engineering, Sciences, and Medicine that inclusively encompasses a spectrum of identities that do not align with traditional gender norms. Henceforth, we will also use terminology to reflect gender (men, women, transgender men, and transgender women) rather than sex (male, female, transgender male, and transgender female) [35].

Additional Descriptive Characteristics

Additional sociodemographic characteristics were optionally answered by clients to describe participants' age in years and race or ethnicity (African American, Asian, Caucasian, Hispanic, Native American, other race, declined to respond, and no response).

Engagement Indicators: Messaging

The total number of messages and the total number of words that clients sent to their therapist were calculated weekly throughout the duration of treatment on the platform. Client *message count* and *word count* are both automatically computed by the platform and stored as metadata. In this study, each week that a client exchanged any message with their therapist was considered a week of *active messaging*. The total number of active messaging weeks was calculated for the 15-week study period and across a client's time on the platform.

Therapist Change

Platform clients can switch therapists at any time. In this study, a therapist change was identified by examining client ID-therapist ID pairings across treatment weeks through the study period. Any participant who was paired with multiple therapists was coded as changing therapists (0=no; 1=yes).

Depressive Symptoms

Depressive symptoms were measured with the PHQ-8 [36]. This brief, self-report survey assesses the severity of depressive symptoms, incorporating criteria from the *Diagnostic and Statistical Manual, Fourth Edition*. The PHQ-8 excludes an item evaluating suicidality. Items are rated on a 4-point scale (0=not at all to 3=nearly every day) and summed to calculate a total score, with higher scores indicating greater depressive symptom severity. A total score ≥10 indicates clinically significant depression.

Clients were prompted to complete the PHQ-8 via the platform app at treatment initiation and subsequently every 3 weeks through service termination. A buffer was permitted such that clients could submit the PHQ-8 1 week before or after the target assessment week. Time points of this study included target assessment week 0 (full baseline), week 3, week 6, week 9, week 12, and week 15. While most participants recorded their first PHQ-8 prior to treatment at week 0 (17,622/20,156; 87.4%), a small proportion submitted their first survey at target assessment week 3 (2534/20,156; 12.6%). All participants within this latter group also returned at least 1 additional requisite PHQ-8 survey at a subsequent assessment time point. Survey completion was not required for service continuation.

Secondary Clinical Outcomes

Four additional measures of depressive symptom change were examined, including treatment response, symptom remission, clinically significant symptom reduction, and symptom deterioration. Prior studies of Talkspace data have assessed these secondary clinical outcomes [37], which align with common definitions of meaningful change in depressive symptoms.

Treatment Response

Differences between PHQ-8 scores at baseline and individual follow-up assessment time points were calculated. Clients were identified as responsive to treatment if this difference represented a 50% or larger reduction in symptoms (0=no; 1=yes). This is a widely accepted indicator of meaningful symptom improvement, supported by PHQ-8 validation studies [36,38] and methodological comparison research [39, 40].

Symptom Remission

Symptom remission was identified if the PHQ-8 score, measured at a follow-up assessment time point, was less than 5 (0=no; 1=yes). This threshold is commonly used to indicate the absence of depressive symptoms and aligns with established definitions of remission from depression in clinical and research contexts [36,39].

Clinically Significant Symptom Reduction

Clients were identified as having a clinically significant reduction in depressive symptoms if, at a follow-up time point, their PHQ-8 score was below the clinical threshold (<10) *and* their score improved by at least 5 points from the client's baseline (0=no; 1=yes). This compound metric indicates both

XSL•FO
RenderX

a reliable change in depressive symptoms and a shift below the threshold for clinically significant depression [38-40].

Symptom Deterioration

Clients' depressive symptoms were identified as deteriorating if their PHQ-8 scores increased by 5 or more points from their baseline at a follow-up assessment time point (0=no; 1=yes). This increase marks a reliable and significant change consistent with an exacerbation of depressive symptoms [39,40].

Analytic Plan

All analyses were completed in Stata (version 19.5; StataCorp). Distributions of participant sociodemographic characteristics, engagement indicators, and depressive symptoms were estimated across the sample and stratified across gender categories. While the primary analyses for the multilevel linear models analyzed 15-week clinical outcomes, a description of engagement indicators for the sample's full course and adjusted 15-week treatment data are both reported. As reported elsewhere [41], few Talkspace clients with clinically significant PHQ-8 scores at baseline remain in treatment past week 16. These analyses suggested that some clients discontinued after early treatment gains (eg, average PHQ-8 scores fell below the clinical cut-off by week 6), while others' disengagement was associated with factors unrelated to symptom remission (eg, therapist's years of experience; first-time therapy client).

Multilevel linear models with maximum likelihood estimation and robust SEs were used to examine changes in depressive symptoms over 15 weeks of treatment. This approach was chosen to account for the nested data structure of clients' repeated PHQ-8 measures. Because clients could switch therapists throughout their treatment, and therapists could treat multiple clients on the platform, a cross-classified model was initially considered to accurately reflect the nature of the data structure by partitioning the unique variance attributable to both clients and therapists [42]. However, an examination of the unconditional models indicated that therapist-level effects did not significantly explain additional variance in PHQ-8 scores relative to client-level effects, and models had poorer fit statistics, suggesting unnecessary complexity and potential overfitting. Considering fit, parsimony, and interpretability, the final models included only client-level random intercepts and random slopes for time.

The first model tested the interaction effects of gender and treatment week on depressive symptoms. This analysis examined differences in PHQ-8 scores over the course of treatment between men, women, transgender men, transgender women, and nonbinary and gender-diverse clients. The second model also controlled for baseline PHQ-8 scores. The third model further incorporated engagement indicators, including therapist change as a predictor and the interaction effects of assessment time point and number of weeks of active messaging. This interaction analysis aimed to clarify whether engagement, through consistent client-therapist correspondence throughout treatment, and time-varying treatment effects had differential impacts on depressive symptoms. Week number was a continuous variable (full baseline or week 0=0; week 3=1; week 6=2; week 9=3; week 12=4; week 15=5). In all regression models, clients who identified as men served as the reference group, as they have historically shown the lowest levels of depressive symptoms among the gender subgroups included in this study. However, estimated marginal means were calculated and plotted for each assessment time point and gender subgroup to clarify and depict depressive symptom trajectories for every gender subgroup.

Prevalence of the secondary clinical outcomes (ie, treatment response, symptom remission, clinically significant symptom reduction, and symptom deterioration) was calculated for each gender subgroup at each PHQ-8 assessment time point. Differences in secondary outcomes between groups were examined via the chi-square and Fisher exact tests.

Ethical Considerations

All data were collected for the service provider's quality assurance and program management. As part of the platform's regular service use agreement, both clients and therapists provided consent for data use in deidentified, aggregate form for research purposes. Therefore, participants were not compensated for their involvement in the study. All materials and procedures for this study were approved by the University of Washington institutional review board, including future secondary analysis of data without direct identifiers (STUDY00010958).

Results

Descriptive Statistics

The mean age of the sample was 30.52 (SD 8.65) years. Most participants identified as women (15,334/20,156, 76.08%), followed by men (4411/20,156, 21.88%), gender-diverse participants (211/20,156, 1.05%), nonbinary participants (104/20,156, 0.52%), transgender men (67/20,156, 0.33%), and transgender women (29/20,156, 0.14%). Across all participants, clinically significant levels of depressive symptoms were identified at baseline, with scores averaging 15.16 (SD 3.79). Throughout the complete study period, transgender, nonbinary, and gender-diverse participants averaged higher depression severity (mean 12.97 - 14.45, SD 5.07-5.20) compared to women and men (mean 11.86, SD 5.54 and mean 11.88, SD 5.62, respectively). Transgender women had the most weeks of active messaging in total (mean 21.41, SD 20.90) and during the study period (mean 11.48, SD 4.98). Transgender men had the fewest weeks of active messaging in total (mean 15.58, SD 20.65) and during the study period (mean 9.57, SD 4.82; Table 1).

XSL•FO
RenderX

Table . Baseline demographic, engagement, and depressive symptom characteristics of Talkspace clients (2017 - 2021; N=20,156).

	Total sample	Transgender men (n=67)	Transgender women (n=29)	Nonbinary (n=104)	Gender diverse (n=211)	Women (n=15,334)	Men (n=4411)
Race or ethnicity, n (%)							
African American	360 (1.79)	0 (0)	0 (0)	1 (0.96)	2 (0.95)	307 (2.00)	50 (1.13)
Asian	208 (1.03)	0 (0)	0 (0)	1 (0.96)	1 (0.47)	165 (1.08)	41 (0.93)
Caucasian	1778 (8.82)	10 (14.93)	4 (13.79)	3 (2.88)	15 (7.11)	1334 (8.70)	412 (9.34)
Hispanic	217 (1.08)	0 (0)	0 (0)	0 (0)	3 (1.42)	158 (1.03)	56 (1.27)
Native American	10 (0.05)	0 (0)	0 (0)	0 (0)	0 (0)	9 (0.06)	1 (0.02)
Biracial	26 (0.13)	0 (0)	0 (0)	0 (0)	3 (1.42)	20 (0.13)	3 (0.07)
Other	228 (1.13)	0 (0)	1 (3.45)	2 (1.92)	2 (0.95)	171 (1.12)	52 (1.18)
Declined	41 (0.20)	0 (0)	0 (0)	1 (0.96)	1 (0.47)	34 (0.22)	5 (0.11)
No response	17,228 (85.77)	57 (85.07)	24 (82.76)	96 (92.31)	184 (87.20)	13,136 (85.67)	3791 (85.94)
Age (y; n=15,716), mean (SD)	30.52 (8.65)	24.27 (8.08)	27.88 (7.09)	25.16 (7.55)	26.95 (7.69)	30.35 (8.58)	31.57 (8.76)
Engagement							
Therapist Change, n (%)	1683 (8.35)	0 (0)	1 (3.45)	7 (6.73)	10 (4.74)	1269 (8.28)	396 (8.98)
Weekly message count, mean (SD)	13.34 (23.94)	10.44 (13.96)	13.10 (16.63)	14.99 (21.66)	13.75 (18.56)	13.27 (24.94)	13.58 (20.56)
Weekly word count, mean (SD)	1002.64 (1434.11)	792.03 (1151.00)	1060.54 (1279.96)	928.05 (1311.81)	1127.76 (1550.77)	998.74 (1379.10)	1014.84 (1614.01)
Weeks of active messaging (total)[a], mean (SD)	16.62 (20.70)	15.58 (20.65)	21.41 (20.90)	16.91 (15.17)	19.86 (26.14)	16.68 (20.83)	16.23 (20.04)
Weeks of active messaging (study period)[b], mean (SD)	10.29 (4.78)	9.57 (4.82)	11.48 (4.98)	10.81 (4.76)	10.42 (4.81)	10.29 (4.78)	10.26 (4.78)
PHQ-8[c], mean (SD)							
Baseline score	15.16 (3.79)	16.69 (4.05)	15.90 (3.43)	15.26 (3.76)	15.45 (3.78)	15.15 (3.77)	15.14 (3.82)
Average score across study period	11.90 (5.55)	14.45 (5.20)	13.1 (5.45)	12.97 (5.07)	13.2 (5.15)	11.86 (5.54)	11.88 (5.62)

[a]Weeks of active messaging (total): total number of weeks that a client exchanged messages with their therapist during the entire period of time they were in treatment on the Talkspace platform.

[b]Weeks of active messaging (study period): total number of weeks that a client exchanged messages with their therapist, not exceeding week 15. If weeks of active messaging (total)>week 15 then weeks of active messaging (study period)=15.

[c]PHQ-8: 8-item Patient Health Questionnaire.

Depressive Symptom Trajectories by Gender and Treatment Engagement

Table 2 displays model unstandardized (*b*) and standardized regression (β) coefficients and fit indices for models 1 (base model), 2 (primary model; controlling for intake PHQ-8), and 3 (controlling for therapist change, weeks of active messaging, and week-by-weeks of active messaging). In all models, there were significant week-by-gender interactions, indicating that changes in depressive symptoms over the course of treatment varied by gender. In model 2, the rate of change in PHQ-8 scores by treatment week for women (b=–.0009; P=.98) and transgender men (b=.25; P=.38) were not significantly different from that of men. Compared to men, rates of change were significantly different for gender-diverse participants (b=0.60; P<.001). The week-by-gender interactions for nonbinary participants (b=0.81; P<.001) and transgender women (b=0.87; P=.007) were significant, with both improving at the slowest

XSL•FO
RenderX

rates among the gender subgroups. In model 3, therapist change, weeks of active messaging, and the interaction of week-by-weeks of active messaging were statistically significant ($P<.001$), but these additions only had a small impact on model fit. Inclusion of these covariates resulted in lower estimates for the week-by-gender interactions (ie, attenuated change in PHQ-8 scores over time compared to model 2). Figure 1 visualizes the week-by-gender interaction for model 2.

XSL•FO
RenderX

Table . Multilevel linear models of depressive symptom trajectories among Talkspace clients (2017 - 2021).

	Model 1[a]				Model 2[b]				Model 3[c]			
	b[d]	SE	β[e]	*P* value	*b*	SE	β	*P* value	*b*	SE	β	*P* value
Intercept	14.07	0.06	__[f]	—	2.34	0.08	—	—	3.34	0.09	—	—
Week number	−1.61	0.03	−.42	<.001	−1.74	0.04	−.45	<.001	−3.34	0.07	−.49	<.001
Gender[g]												
Transgender men	1.98	0.50	.43	<.001	0.89	0.22	.22	<.001	0.63	0.21	.21	.002
Transgender women	0.30	0.73	.29	.68	-0.10	0.36	.21	.78	−0.23	0.38	.18	.54
Nonbinary	0.40	0.37	.25	.28	0.20	0.24	.25	.42	0.21	0.23	.23	.35
Gender diverse	0.63	0.30	.26	.03	0.42	0.16	.24	.008	0.36	0.15	.21	.02
Women	−0.04	0.07	−.002	.55	−0.02	0.04	−.004	.63	−0.03	0.04	−.008	.44
Week number x gender[g]												
Week number x transgender men	0.30	0.27	.08	.28	0.25	0.28	.06	.38	0.36	0.27	.09	.18
Week number x transgender women	0.88	0.29	.23	.003	0.87	0.32	.23	.007	0.82	0.32	.21	.01
Week number x nonbinary	0.67	0.20	.18	.001	0.81	0.20	.21	<.001	0.72	0.21	.19	<.001
Week number x gender diverse	0.57	0.14	.15	<.001	0.60	0.15	.16	<.001	0.57	0.15	.15	<.001
Week number x women	0.02	0.04	.01	.58	−0.0009	0.04	−.0002	.98	-0.01	0.04	−.002	.84
Baseline PHQ-8[h] score	—	—	—	—	0.77	0.005	.52	<.001	0.77	0.005	.520	<.001
Therapist change, yes[i]	—	—	—	—	—	—	—	—	−1.58	0.08	-.28	<.001
Weeks of active messaging (study period)	—	—	—	—	—	—	—	—	−0.07	0.003	.10	<.001

XSL•FO
RenderX

	Model 1[a]				Model 2[b]				Model 3[c]			
	b[d]	SE	β[e]	*P* value	*b*	SE	β	*P* value	*b*	SE	β	*P* value
Week number x weeks of active messaging (study period)	—	—	—	—	—	—	—	—	0.13	0.004	.16	≤.001

[a]Model fit statistics for model 1: Akaike information criterion (AIC) 361,597.8; Bayesian information criterion (BIC) 361,733.0; −2LL=−180,783.9, df=15; conditional R^2=0.55; marginal R^2=0.16.

[b]Model fit statistics for model 2: AIC 345,509.1; BIC 344,004.2; −2LL=−172,738.6, df=16; conditional R^2=0.71; marginal R^2=0.36.

[c]Model fit statistics for model 3: AIC 345,653.3; BIC 344,175.4; −2LL=−171,983.1, df=19; conditional R^2=0.71; marginal R^2=0.35.

[d]*b* represents the raw or unstandardized regression coefficient.

[e]β represents the standardized regression coefficient (β). Standardized regression coefficients were derived from unstandardized estimates following guidance from Hox [43]. In multilevel models, these values are approximate effect sizes only and should be interpreted with caution as standardization does not uniquely partition variance across levels.

[f]Not applicable.

[g]Reference group: men.

[h]PHQ-8: 8-item Patient Health Questionnaire.

[i]Reference group: no.

Figure 1. Adjusted 15-week depressive symptom trajectories by gender estimated among Talkspace clients (2017 - 2021). PHQ-8: 8-item Patient Health Questionnaire.

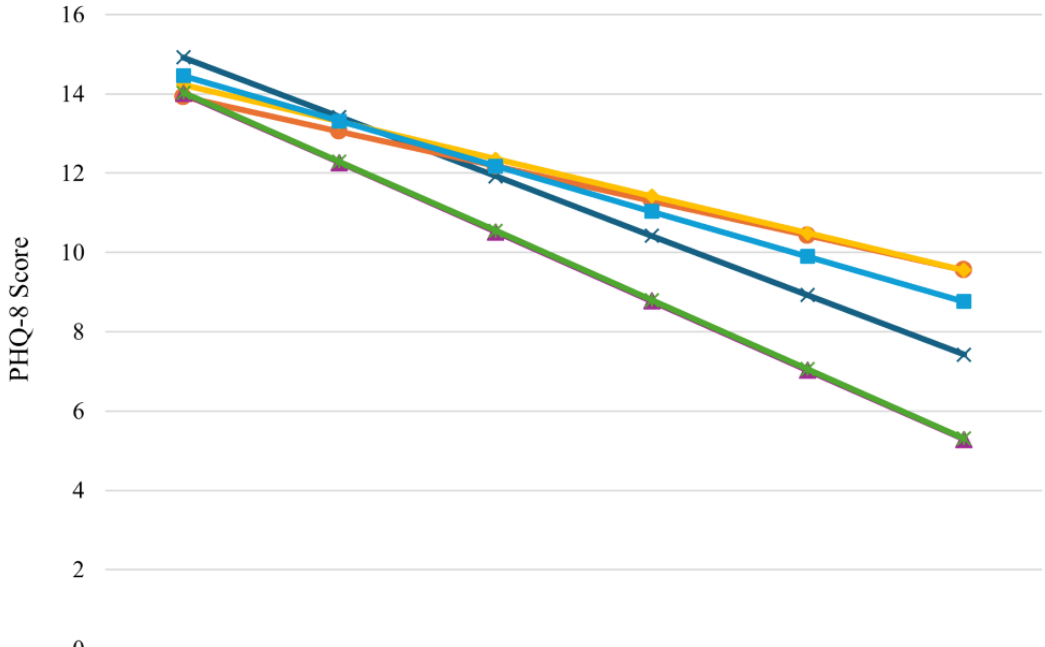

	Week 0	Week 3	Week 6	Week 9	Week 12	Week 15
Transgender men	14.92	13.42	11.92	10.42	8.92	7.43
Transgender women	13.93	13.05	12.18	11.30	10.43	9.55
Nonbinary	14.23	13.29	12.36	11.42	10.49	9.55
Gender diverse	14.45	13.31	12.17	11.04	9.90	8.76
Women	14.01	12.26	10.52	8.77	7.03	5.28
Men	14.03	12.29	10.54	8.80	7.05	5.31

XSL•FO
RenderX

Secondary Clinical Outcomes

Overview

All secondary clinical outcomes across the study period are reported in Table 3. Results indicated differential rates of treatment response, symptom remission, and clinically significant symptom change by gender.

XSL•FO
RenderX

Table . Secondary clinical outcomes across 15 weeks by gender among Talkspace clients (2017 - 2021).[a]

	Full sample	Transgender women	Transgender men	Nonbinary	Gender diverse	Women	Men	*P* value[b]
Week 3								
Available, n	15,489[c]	19[c]	54[c]	78[c]	161[c]	11,807[c]	3370[c]	—[d]
Treatment response[e], n (%)	4694 (30.31)	4 (21.05)	7 (12.96)	15 (19.23)	32 (19.88)	3624 (30.69)	1012 (30.03)	<.001
Symptom remission[f], n (%)	1984 (12.81)	1 (5.26)	2 (3.70)	9 (11.54)	10 (6.21)	1538 (13.03)	424 (12.58)	.03
Clinically significant change[g], n (%)	5580 (36.03)	5 (26.32)	8 (14.81)	17 (21.79)	42 (26.09)	4329 (36.66)	1179 (34.99)	<.001
Symptom deterioration[h], n (%)	608 (3.93)	0 (0)	2 (3.70)	6 (7.69)	6 (3.73)	448 (3.79)	146 (4.33)	.35
Week 6								
Available, n	10,240	20	39	62	114	7829	2176	—
Treatment response[e], n (%)	3518 (34.36)	8 (40.00)	10 (25.64)	12 (19.35)	22 (19.30)	2700 (34.49)	766 (35.20)	.001
Symptom remission[f], n (%)	1643 (16.04)	4 (20.00)	2 (5.13)	5 (8.06)	9 (7.89)	1259 (16.08)	364 (16.73)	.02
Clinically significant change[g], n (%)	4092 (39.96)	8 (40.00)	10 (25.64)	12 (19.35)	26 (22.81)	3169 (40.48)	867 (39.84)	<.001
Symptom deterioration[h], n (%)	402 (3.93)	0 (0)	1 (2.56)	6 (9.68)	6 (5.26)	294 (3.76)	95 (4.37)	.15
Week 9								
Available, n	6161	16	17	39	67	4719	1303	—
Treatment response[e], n (%)	2345 (38.06)	4 (25.00)	3 (17.65)	8 (20.51)	18 (26.87)	1819 (38.55)	493 (37.84)	.02
Symptom remission[f], n (%)	1181 (19.17)	1 (6.25)	1 (5.88)	3 (7.69)	7 (10.45)	904 (19.16)	265 (20.34)	.05
Clinically significant change[g], n (%)	2661 (43.19)	4 (25.00)	4 (23.53)	9 (23.08)	22 (32.84)	2070 (43.87)	552 (42.36)	.008
Symptom deterioration[h], n (%)	246 (3.99)	0 (0)	0 (0)	3 (7.69)	3 (4.48)	188 (3.98)	52 (3.99)	.77
Week 12								
Available, n	4310	10	15	28	51	3269	937	—
Treatment response[e], n (%)	1782 (41.35)	2 (20.00)	5 (33.33)	11 (39.29)	9 (17.65)	1363 (41.69)	392 (41.84)	.009
Symptom remission[f], n (%)	937 (21.74)	2 (20.00)	1 (6.67)	5 (17.86)	4 (7.84)	717 (21.93)	208 (22.20)	.12

XSL•FO
RenderX

	Full sample	Transgender women	Transgender men	Nonbinary	Gender diverse	Women	Men	*P* value[b]
Clinically significant change[g], n (%)	1987 (46.10)	2 (20.00)	6 (40.00)	11 (39.29)	11 (21.57)	1511 (46.22)	446 (47.60)	.004
Symptom deterioration[h], n (%)	185 (4.29)	1 (10.00)	2 (13.33)	2 (7.14)	4 (7.84)	137 (4.19)	39 (4.16)	.13
Week 15								
Available, n	2975	13	10	20	37	2243	652	—
Treatment response [e], n (%)	1224 (41.14)	3 (23.08)	3 (30.00)	8 (40.00)	9 (24.32)	921 (41.06)	280 (42.94)	.19
Symptom remission[f], n (%)	665 (22.35)	1 (7.69)	1 (10.00)	2 (10.57)	4 (10.81)	487 (21.71)	170 (26.07)	.03
Clinically significant change[g], n (%)	1368 (45.98)	5 (38.46)	2 (20.00)	8 (40.00)	9 (24.32)	1027 (45.79)	317 (48.62)	.03
Symptom deterioration[h], n (%)	127 (4.27)	1 (7.69)	0 (0)	0 (0)	2 (5.41)	94 (4.19)	30 (4.60)	.79

[a]Scores calculated using observed 8-item Patient Health Questionnaire (PHQ-8) scores.

[b]*P* value reported for the chi-square or the Fisher exact when n<5.

[c]Available data for week 3 calculations excluded cases whose first PHQ-8 survey was returned during the week 3 assessment time point as secondary clinical outcome calculations required baseline data.

[d]Not applicable.

[e]Treatment response rate: 50% or larger reduction in symptoms from baseline.

[f]Symptom remission: PHQ-8 score <5 at assessment time point.

[g]Clinically significant symptom change: below clinical threshold (PHQ-8 score<10) and improving at least 5 points from baseline.

[h]Symptom deterioration: PHQ-8 score worsened by ≥5 points from baseline.

Treatment Response

Treatment response rates were significantly different by gender from weeks 3 to 12 (all *P*≤.02). At week 3, treatment response rates were highest for women (3624/11,807, 30.69%) and men (1012/3370, 30.03%) and were ≤21.05% for gender minorities. Transgender men had the lowest treatment response rate (7/54, 12.96%) at week 3. Treatment response rates were not significantly different between gender subgroups at week 15 (*P*=.19).

Symptom Remission

There were statistically significant differences in symptom remission rates by gender at weeks 3, 6, 9, and 15 (all *P*≤.047). Differences in remission rates at week 3 (1538/11,807, 13.03% and 424/3370, 12.58% for women and men, respectively; 3.70% - 11.54% for gender minorities) became more pronounced by week 15. Remission rates at week 15 for women and men (487/2243, 21.71% and 170/652, 26.07%, respectively) were over 2 times greater than those of gender minority participants (7.69% - 10.81%).

Clinically Significant Symptom Reduction

Rates of clinically significant symptom reduction were significantly different by gender at each time point (all *P*≤.03). Clinically significant symptom reduction rates for transgender women, nonbinary, and gender-diverse clients at week 3 ranged from 21.79% (17/78) to 26.32% (5/19). Rates of women and men with clinically significant symptom reduction at week 3 (4329/11,807, 36.66% and 1179/3370, 34.99%, respectively) were over 2 times greater than that of transgender men (8/54, 14.81%). Clinically significant symptom reduction increased for all groups at week 15 (20% - 40% for gender minorities; 1027/2243, 45.79% and 317/652, 48.62% for women and men, respectively).

Symptom Deterioration

There were no significant differences in deterioration rates by gender from weeks 3 to 15 (all *P*>.05). Deterioration rates across all weeks ranged from 3.99% (52/1303) to 4.60% (30/652) for men, 3.76% (294/7829) to 4.19% (137/3269) for women, and 0% to 13.33% for gender minorities.

Sensitivity Analysis

Clients displayed high rates of treatment dropout throughout the 15-week study period. Of the 20,156 clients, 2925 (14.5%)

XSL•FO
RenderX

completed a week-15 PHQ-8 survey, and 17,231 (85.5%) did not. It is possible that those who remained engaged throughout the entire study period provide a more accurate picture of effectiveness. To examine whether attrition may have biased the estimation of depressive symptom trajectories, a series of sensitivity analyses were conducted.

Clients Who Completed the 15-Week Study Period

Model 2 was reestimated using only data from the subgroup of clients who submitted a PHQ-8 survey during the week 15 assessment period (Multimedia Appendix 1). Consistent with the main analysis, these clients experienced significant weekly reductions in depressive symptoms (b=−1.01; P<.001), although rates were slightly attenuated compared to the main analysis (b=−1.74; P<.001; Table 2). The week-by-gender interaction was only significant for transgender women (b=0.76; P=.022). Estimated depressive symptom trajectories were visualized, showing that transgender women who remained engaged through the 15-week study period experienced slower symptom reduction than other gender subgroups (Multimedia Appendix 2).

Clients Grouped by Final PHQ-8 Assessment Week

To further contextualize why trajectories estimated in the main analysis differed from those estimated among week-15 completers, model 2 was also reestimated separately among clients grouped by final PHQ-8 assessment week (ie, the last week a PHQ-8 score was available for each client, indicating the final survey they completed; Multimedia Appendix 3). Overall, models of the earliest dropout weeks showed steep weekly symptom change (week 3: b=−4.06, P<.001; week 6: b=−2.31, P<.001). However, trajectories varied by gender among some final assessment week groups. For example, among those whose final PHQ-8 was submitted during the week 3 assessment period, transgender women's PHQ-8 scores increased from baseline to week 3 (Multimedia Appendix 4). Other gender minority clients showed slower rates of improvement, including nonbinary clients whose final PHQ-8 was submitted in week 6 (b=1.12; P=.03), week 9 (b=1.30; P=.002), and week 12 (b=0.71; P=.048) and gender-diverse clients whose final PHQ-8 was submitted in week 6 (b=0.87; P=.013) and week 12 (b=0.87, P=.006; Multimedia Appendices 3 and 4).

Reasons for Service Cancellation Before Week 15

Reasons for service cancellation were also examined among the 17,231 clients who discontinued treatment before week 15. Among these clients, 3450 (20.02%) provided at least 1 reason for cancellation (Multimedia Appendix 5). Over a third (1431/3691 reasons, 38.77%) cited reasons related to cost, a quarter reported that they felt better or met their goals (928/3691 reasons, 25.14%), and 18% (677/3691 reasons) indicated poor service fit or poor perceived effectiveness. Higher rates of nonbinary clients reported cost reasons, while higher rates of gender-diverse clients canceled due to issues of fit or perceived effectiveness. Early dropouts may reflect both symptom improvement and barriers to treatment engagement.

Discussion

Principal Findings

This study examined gender differences in depressive symptom trajectories and engagement patterns among clients of a DMH platform with clinically significant levels of depression. Data were observational, with outcomes assessed in a naturalistic setting as part of the routine delivery of services, including at baseline and every 3 weeks over 15 weeks of the study period. Results suggested that technology-mediated psychotherapy may reduce depressive symptoms, but that treatment effects over time may function differently across genders. Across 15 weeks, heterogeneous depressive symptom trajectories and differential rates of treatment response, symptom remission, and clinically significant symptom change indicated that some gender minority groups, particularly transgender women, nonbinary, and gender-diverse clients, may not benefit from this model of DMH as immediately as men- and women-identified clients. These findings may have implications for service delivery, highlighting the need for tailored approaches. Combining the broad accessibility of DMH with tailored strategies may better address the mental health needs of gender minority clients.

Transgender women, nonbinary, and gender-diverse clients experienced slower symptom improvement over the 15 weeks of treatment compared to men and women, who each showed similar rates of improvement. These patterns were reflected in significant week-by-gender interactions across all models. We also found that men and women showed more consistent improvements in secondary clinical outcomes. Gender minority individuals face unique stressors that impact their mental health, including chronic exposure to interpersonal and structural stigma, discrimination, and victimization across multiple contexts [44]. Research shows that the cumulative effects of these experiences are associated with poor mental health outcomes [45-48], including longitudinal depressive symptom trajectories [49].

Gender-based discrimination can also typify gender minority individuals' care-seeking experiences, as they often experience multiple forms of discrimination [19], across multiple settings [50], and at higher rates than cisgender individuals [51]. Experiences of health care discrimination can create hesitancies for open communication with providers. These hesitancies can subsequently disrupt therapeutic alliance and engagement [52,53], which are key factors in improving treatment outcomes for gender minority clients [54,55]. In this study, we found that some gender minority clients actively messaged with their therapist for more weeks and, on average, sent more weekly messages and words compared to men and women. Transgender women and gender-diverse clients, for example, had the highest weekly word count and longest durations of active messaging, but at the same time, their symptom improvement did not keep pace with men and women who messaged less and for shorter durations. While it is possible that these clients required longer periods of messaging due to more complex treatment needs, it is also possible that technology-mediated psychotherapy, as deployed, could not fully address these needs. Consistent with this interpretation, sensitivity analyses indicated that transgender

XSL•FO
RenderX

women who remained engaged through all 15 weeks experienced slower symptom improvement than other gender subgroups. Their adjusted PHQ-8 scores at week 15 showed less than a 2-point improvement from baseline and remained above the clinically significant threshold. These findings reinforce that longer engagement may not necessarily yield better clinical outcomes for some gender minority clients.

Furthermore, model 3 found that longer periods of messaging may reflect diminishing returns over time. Early gains in symptom improvement plateaued as messaging continued, possibly due to a weaker rapport limiting therapeutic depth. However, randomized clinical trials of message-based care have yet to demonstrate any differences relative to video-based telehealth [56-58]. For those with prior experiences of discrimination, the delivery format may not have fostered the rapport necessary to support open communication. Messaging content, however, remains unknown and could provide important insights into the unique barriers and needs of gender minority clients. In traditional face-to-face therapy, engagement is more robustly assessed by more than just attendance; active participation and openness are critical to therapeutic progress. Recent research on digital therapeutic alliance emphasizes limitations in replicating cornerstone alliance dimensions during digital services delivery, including bond and empathy [59-61]. Future research should evaluate not only the frequency and duration of messaging but also the quality of engagement to better understand mechanisms of the therapeutic process in the digital context, especially for gender minority clients. In addition, leveraging artificial intelligence tools [62] that integrate into DMH platforms may support therapists' confidence in delivering empathetic and gender-affirming messages.

Recognizing the importance of rapport in the therapeutic process, Talkspace has integrated a feature that allows clients to easily change their provider for any reason to enhance fit. We found that therapist change was associated with decreased PHQ-8 scores, indicating the benefits of this feature. However, fewer gender minority clients opted to use this feature. Initial therapist matches on Talkspace are generated by an algorithm that incorporates user-reported preferences. Therapists can also self-identify areas of expertise that display on their profile, with 30% of the therapist network reporting experience working with lesbian, gay, bisexual, transgender, and queer communities [63]. The matching algorithm and therapists' profile data may have reduced therapist changes for gender minority clients by helping them identify affirming providers. However, even with these safeguards, most gender minority clients did not improve at the same rate as men and women. Their depressive symptom trajectories may be influenced by a confluence of factors, including ongoing minority stress that hinders symptom improvement [45,64,65] and anticipated and past experiences of discrimination that can shape help-seeking behaviors, even when a better match might be available. Extant research shows that some gender minority clients compromise their care needs by remaining with providers who can offer either a specific clinical specialization or gender-affirming expertise, but not both [66]. In this study, it is possible that when working with their therapist, gender minority clients were satisfied with only one of these treatment domains, which could help explain the dichotomy of lower therapist-switching rates alongside slower symptom improvement. Indeed, finding the ideal provider with a combination of expertise, identity-affirmation, and service type has been described by some as "finding a unicorn" [67].

While the Talkspace matching algorithm considers some client preferences, it does not consider identity-based characteristics of therapists and restricts therapist gender preferences to "male" or "female." These constraints may overlook factors that are particularly important to gender minority clients. The current study did not examine therapist qualities, including their formal training or self-reported experience working with gender minority clients, or client satisfaction data. However, some insight can be drawn from clients' reported reasons for discontinuing treatment. Reasons reflected both positive outcomes and service-related barriers, including those known to disproportionately impact gender minority clients (eg, cost; service fit and effectiveness). Because gender minority clients generally report lower satisfaction with their care [68], more unmet expectations and unmet needs [20,69], and barriers to service acceptability [14], future research should work to incorporate provider ratings and areas of specialization to help clarify the experiences of gender minority clients who did not change therapists. Targeted outreach and clear messaging outlining the specific benefits of the therapist change feature for gender minority clients may encourage greater access. Improved engagement with this feature may lead to better therapeutic outcomes by ensuring that gender minority clients feel empowered to find a provider who meets both their clinical and identity-affirming needs.

Strengths and Limitations

Strengths of the study included the naturalistic observation of treatment engagement and clinical outcomes on a large real-world telehealth platform. Gender was assessed with a single question, rather than the recommended 2-step measure that more accurately identifies gender minority individuals. Although single-item measures are common in general clinical settings because they reduce burden, they can misclassify people with transgender experience who do not identify as transgender. Furthermore, they fail to explicitly identify cisgender people who cannot be distinguished from transgender respondents selecting "male" or "female." However, research shows that only a small proportion of transgender individuals avoid disclosing their identity on such measures [35]. Misclassification may only impact a small subset of Talkspace clients, but this limitation should still be considered when interpreting gender subgroup differences. Clients' and therapists' other sociodemographic factors may have influenced engagement. For example, there are known racial and ethnic disparities in access to and engagement with mental health care, including DMH services [70,71], but an analysis including these factors was not feasible due to large amounts of missing demographic data (eg, close to 86% of the sample did not provide information about their race or ethnicity). This study used the available data that may have contributed to treatment engagement and described reasons why clients canceled Talkspace services. Sensitivity analyses suggested that some of the treatment response observed in the full sample may be disproportionately

XSL•FO
RenderX

driven by clients who discontinued treatment after showing initial steep symptom improvement, potentially inflating estimates of average treatment benefit. Missing data in later weeks may have been due to clients choosing not to fill out the surveys, leaving the platform after achieving adequate symptom improvement before week 15, discontinuing due to issues of fit or perceived effectiveness, or other reasons captured in cancellation data. To account for the missing data, our statistical approach used listwise deletion for all analyses except the multilevel modeling. The use of multilevel models accounted for unequal and smaller sample sizes of gender minorities.

Future Directions

Future research may explore strategies to adapt digital interventions to better meet the needs of diverse gender identities, potentially incorporating specific modules on minority stress, identity affirmation, and navigating discrimination in healthcare settings. Treatment engagement and clinical outcomes of gender minority clients may be examined on other DMH platforms, purposeful sampling techniques may be used to increase sample sizes of gender minority clients, and inclusion of a control group would help to strengthen claims of efficacy. Future studies may also examine how sociodemographic factors impact treatment engagement and efficacy.

Conclusions

Our study found that while clinical outcomes improved over time, some gender minority clients experienced slower symptom reduction compared to men and women. Depression severity decreased more slowly among transgender women, nonbinary, and gender-diverse clients. By week 15, PHQ-8 scores declined by an average of 8.7 points among men and women, compared to 5.7, 4.7, and 4.4 points among gender-diverse, nonbinary, and transgender women clients, respectively. On average, men and women clients of Talkspace also had higher rates of treatment response, symptom remission, and clinically significant symptom reduction than gender minority clients. We focused on gender minority clients to provide insight into their engagement with DMH services, the impacts of this engagement on clinical outcomes, and the need for more effective approaches to their care.

Funding

This study was supported by a grant from the National Institute of Mental Health (R01MH125179). MR is supported by the National Institute of Mental Health of the National Institutes of Health (award K01MH131795). The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

Talkspace provided anonymized data. Requests for raw data can be made by written request to the corresponding author and require a separate Data Use Agreement.

Conflicts of Interest

TDH was an employee of the platform providing the data during the study period. AP and EP are current employees of the platform providing the data. All other authors have no conflicts of interest to disclose.

Multimedia Appendix 1
Multilevel linear models of depressive symptom trajectories among Talkspace clients (2017-2021) who completed the 15-week study period.
[DOCX File, 18 KB - formative_v9i1e75290_app1.docx]

Multimedia Appendix 2
Adjusted 15-week depressive symptom trajectories by gender estimated among Talkspace clients (2017-2021) who completed the 15-week study period.
[PNG File, 148 KB - formative_v9i1e75290_app2.png]

Multimedia Appendix 3
Multilevel linear models of depressive symptom trajectories among Talkspace clients (2017-2021) grouped by final 8-item Patient Health Questionnaire (PHQ-8) assessment week.
[DOCX File, 21 KB - formative_v9i1e75290_app3.docx]

Multimedia Appendix 4
Adjusted 15-week depressive symptom trajectories by gender estimated among Talkspace clients (2017-2021) grouped by final 8-item Patient Health Questionnaire (PHQ-8) assessment week.
[PNG File, 172 KB - formative_v9i1e75290_app4.png]

Multimedia Appendix 5

XSL•FO
RenderX

Reported reasons for service cancellation among Talkspace clients (2017-2021) who discontinued before week 15.
[DOCX File, 21 KB - formative_v9i1e75290_app5.docx]

Checklist 1
STROBE checklist.
[DOCX File, 34 KB - formative_v9i1e75290_app6.docx]

References

1. Fredriksen-Goldsen KI, Simoni JM, Kim HJ, et al. The health equity promotion model: reconceptualization of lesbian, gay, bisexual, and transgender (LGBT) health disparities. Am J Orthopsychiatry 2014 Nov;84(6):653-663. [doi: 10.1037/ort0000030] [Medline: 25545433]
2. Valentine SE, Shipherd JC. A systematic review of social stress and mental health among transgender and gender non-conforming people in the United States. Clin Psychol Rev 2018 Dec;66:24-38. [doi: 10.1016/j.cpr.2018.03.003] [Medline: 29627104]
3. Friedrich MJ. Depression Is the leading cause of disability around the world. JAMA 2017 Apr 18;317(15):1517. [doi: 10.1001/jama.2017.3826]
4. Depressive disorder (depression). World Health Organization. 2025. URL: https://www.who.int/news-room/fact-sheets/detail/depression [accessed 2025-09-14]
5. Marcus M, Yasamy MT, Ommeren M, Chisholm D, Saxena S. Depression: a global public health concern. : World Health Organization; 2012. [doi: 10.1037/e517532013-004]
6. GBD 2021 Diseases and Injuries Collaborators. Global incidence, prevalence, years lived with disability (YLDs), disability-adjusted life-years (DALYs), and healthy life expectancy (HALE) for 371 diseases and injuries in 204 countries and territories and 811 subnational locations, 1990-2021: a systematic analysis for the Global Burden of Disease Study 2021. Lancet 2024 May 18;403(10440):2133-2161. [doi: 10.1016/S0140-6736(24)00757-8] [Medline: 38642570]
7. Kessler RC, Petukhova M, Sampson NA, Zaslavsky AM, Wittchen H. Twelve‐month and lifetime prevalence and lifetime morbid risk of anxiety and mood disorders in the United States. Int J Methods Psych Res 2012 Sep;21(3):169-184. [doi: 10.1002/mpr.1359]
8. Borgogna NC, McDermott RC, Aita SL, Kridel MM. Anxiety and depression across gender and sexual minorities: implications for transgender, gender nonconforming, pansexual, demisexual, asexual, queer, and questioning individuals. Psychol Sex Orientat Gend Divers 2019;6(1):54-63. [doi: 10.1037/sgd0000306]
9. Reisner SL, Poteat T, Keatley J, et al. Global health burden and needs of transgender populations: a review. Lancet 2016 Jul 23;388(10042):412-436. [doi: 10.1016/S0140-6736(16)00684-X] [Medline: 27323919]
10. O'Shea J, Jenkins R, Nicholls D, Downs J, Hudson LD. Prevalence, severity and risk factors for mental disorders among sexual and gender minority young people: a systematic review of systematic reviews and meta-analyses. Eur Child Adolesc Psychiatry 2025 Mar;34(3):959-982. [doi: 10.1007/s00787-024-02552-1] [Medline: 39141104]
11. Pellicane MJ, Ciesla JA. Associations between minority stress, depression, and suicidal ideation and attempts in transgender and gender diverse (TGD) individuals: systematic review and meta-analysis. Clin Psychol Rev 2022 Feb;91:102113. [doi: 10.1016/j.cpr.2021.102113] [Medline: 34973649]
12. Baker KE. Findings From the behavioral risk factor surveillance system on health-related quality of life among US transgender adults, 2014-2017. JAMA Intern Med 2019 Aug 1;179(8):1141-1144. [doi: 10.1001/jamainternmed.2018.7931] [Medline: 31009042]
13. Lu JA, Soltani S, Austin SB, Rehkopf DH, Lunn MR, Langston ME. Mental health disparities by sexual orientation and gender identity in the All of Us Research Program. JAMA Netw Open 2025 Jan 2;8(1):e2456264. [doi: 10.1001/jamanetworkopen.2024.56264] [Medline: 39878980]
14. Romanelli M, Hudson KD. Individual and systemic barriers to health care: perspectives of lesbian, gay, bisexual, and transgender adults. Am J Orthopsychiatry 2017;87(6):714-728. [doi: 10.1037/ort0000306] [Medline: 29154611]
15. Snow A, Cerel J, Loeffler DN, Flaherty C. Barriers to mental health care for transgender and gender-nonconforming adults: a systematic literature review. Health Soc Work 2019 Aug 2;44(3):149-155. [doi: 10.1093/hsw/hlz016] [Medline: 31359065]
16. Pepping CA, Cronin TJ, Davis AW. Mental health care for transgender and non-binary adults: an investigation of affirmative practice, therapy experiences and outcomes, and reasons for treatment termination. Sex Res Soc Policy 2025 Dec;22(4):1951-1963. [doi: 10.1007/s13178-024-01065-5]
17. Martos AJ, Wilson PA, Meyer IH. Lesbian, gay, bisexual, and transgender (LGBT) health services in the United States: origins, evolution, and contemporary landscape. PLoS ONE 2017;12(7):e0180544. [doi: 10.1371/journal.pone.0180544] [Medline: 28692659]
18. Martos AJ, Fingerhut A, Wilson PA, Meyer IH. Utilization of LGBT-specific clinics and providers across three cohorts of lesbian, gay, and bisexual people in the United States. SSM Popul Health 2019 Dec;9:100505. [doi: 10.1016/j.ssmph.2020.100716] [Medline: 31993490]

XSL•FO
RenderX

19. Romanelli M, Lindsey MA. Patterns of healthcare discrimination among transgender help-seekers. Am J Prev Med 2020 Apr;58(4):e123-e131. [doi: 10.1016/j.amepre.2019.11.002] [Medline: 32001051]
20. Steele LS, Daley A, Curling D, et al. LGBT identity, untreated depression, and unmet need for mental health services by sexual minority women and trans-identified people. J Womens Health (Larchmt) 2017 Feb;26(2):116-127. [doi: 10.1089/jwh.2015.5677] [Medline: 27898255]
21. Schueller SM, Hunter JF, Figueroa C, Aguilera A. Use of digital mental health for marginalized and underserved populations. Curr Treat Options Psych 2019 Sep;6(3):243-255. [doi: 10.1007/s40501-019-00181-z]
22. DeVito MA, Walker AM, Birnholtz J, et al. Social technologies for digital wellbeing among marginalized communities. Presented at: CSCW '19: Computer Supported Cooperative Work and Social Computing; Nov 9-13, 2019; Austin, TX p. 449-454. [doi: 10.1145/3311957.3359442]
23. Robards F, Kang M, Usherwood T, Sanci L. How marginalized young people access, engage with, and navigate health-care systems in the digital age: systematic review. J Adolesc Health 2018 Apr;62(4):365-381. [doi: 10.1016/j.jadohealth.2017.10.018] [Medline: 29429819]
24. Hull TD, Malgaroli M, Connolly PS, Feuerstein S, Simon NM. Two-way messaging therapy for depression and anxiety: longitudinal response trajectories. BMC Psychiatry 2020 Jun 12;20(1):297. [doi: 10.1186/s12888-020-02721-x] [Medline: 32532225]
25. Steinke J, Root-Bowman M, Estabrook S, Levine DS, Kantor LM. Meeting the needs of sexual and gender minority youth: formative research on potential digital health interventions. J Adolesc Health 2017 May;60(5):541-548. [doi: 10.1016/j.jadohealth.2016.11.023] [Medline: 28108088]
26. Lucassen M, Samra R, Iacovides I, et al. How LGBT+ young people use the internet in relation to their mental health and envisage the use of e-therapy: exploratory study. JMIR Serious Games 2018 Dec 21;6(4):e11249. [doi: 10.2196/11249] [Medline: 30578194]
27. Strauss P, Morgan H, Wright Toussaint D, Lin A, Winter S, Perry Y. Trans and gender diverse young people's attitudes towards game-based digital mental health interventions: a qualitative investigation. Internet Interv 2019 Dec;18:100280. [doi: 10.1016/j.invent.2019.100280] [Medline: 31890628]
28. National survey on LGBTQ mental health. The Trevor Project. 2019. URL: https://www.thetrevorproject.org/wp-content/uploads/2019/06/The-Trevor-Project-National-Survey-Results-2019.pdf [accessed 2025-11-19]
29. Gilbey D, Morgan H, Lin A, Perry Y. Effectiveness, acceptability, and feasibility of digital health interventions for LGBTIQ+ young people: systematic review. J Med Internet Res 2020 Dec 3;22(12):e20158. [doi: 10.2196/20158] [Medline: 33270039]
30. Perry Y, Strauss P, Lin A. Online interventions for the mental health needs of trans and gender diverse young people. Lancet Psychiatry 2018 Feb;5(2):e6. [doi: 10.1016/S2215-0366(18)30017-8] [Medline: 29413141]
31. Tasnim S, Trisha NF, Fan Q, et al. Improving mental health and wellbeing among sexual and gender minorities using digital interventions: a systematic review. PsyArXiv. Preprint posted online on Feb 4, 2020. [doi: 10.31234/osf.io/9wu42]
32. Elm EV, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: Guidelines for reporting observational studies. BMJ 2007 Oct 20;335(7624):806-808. [doi: 10.1136/bmj.39335.541782.AD]
33. Talkspace. URL: https://www.talkspace.com/ [accessed 2024-06-27]
34. Esacove A. Common patterns of cisgender use in public health articles and their implications for gender inclusivity efforts, 2013 2020. Am J Public Health 2024 Feb;114(2):202-208. [doi: 10.2105/AJPH.2023.307441] [Medline: 37883718]
35. National Academies of Sciences, Engineering, and Medicine, Division of Behavioral and Social Sciences and Education, Committee on Population, White J, Sepúlveda MJ, Patterson CJ. Understanding the Well-Being of LGBTQI+ Populations: The National Academies Press; 2020. [doi: 10.17226/25877]
36. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: Validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613. [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
37. Raue PJ, Fridling N, Song J, Hull TD, Alexopoulos GS, Arean PA. Message-based psychotherapy for older adults: a cohort comparison study. Front Psychiatry 2022;13:951354. [doi: 10.3389/fpsyt.2022.951354] [Medline: 36090371]
38. Löwe B, Unützer J, Callahan CM, Perkins AJ, Kroenke K. Monitoring depression treatment outcomes with the Patient Health Questionnaire-9. Med Care 2004 Dec;42(12):1194-1201. [doi: 10.1097/00005650-200412000-00006] [Medline: 15550799]
39. McMillan D, Gilbody S, Richards D. Defining successful treatment outcome in depression using the PHQ-9: a comparison of methods. J Affect Disord 2010 Dec;127(1-3):122-129. [doi: 10.1016/j.jad.2010.04.030] [Medline: 20569992]
40. Coley RY, Boggs JM, Beck A, Hartzler AL, Simon GE. Defining success in measurement-based care for depression: a comparison of common metrics. Psychiatr Serv 2020 Apr 1;71(4):312-318. [doi: 10.1176/appi.ps.201900295] [Medline: 31847739]
41. Darnell D, Pullmann MD, Hull TD, Chen S, Areán P. Predictors of disengagement and symptom improvement among adults with depression enrolled in Talkspace, a technology-mediated psychotherapy platform: naturalistic observational study. JMIR Form Res 2022 Jun 22;6(6):e36521. [doi: 10.2196/36521] [Medline: 35731563]
42. Beretvas SN. Cross-classified and multiple-membership models. In: Hox JJ, Roberts JK, editors. Handbook of Advanced Multilevel Analysis, 1st edition: Routledge:313-334.

XSL•FO
RenderX

43. Hox JJ. Multilevel Analysis: Techniques and Applications, th edition: Routledge; 2010. [doi: 10.4324/9780203852279]
44. James SE, Herman JL, Rankin S, Keisling M, Mottet L, Anafi M. The report of the 2015 U.S. Transgender Survey. Advocates for Trans Equality. 2016. URL: https://transequality.org/sites/default/files/docs/usts/USTS-Full-Report-Dec17.pdf [accessed 2025-11-19]
45. Doyle DM, Georgiev D, Lewis TOG, Barreto M. Frequency and mental health consequences of microaggressions experienced in the day-to-day lives of transgender and gender diverse people. International Journal of Transgender Health 2024:1-14. [doi: 10.1080/26895269.2024.2380906]
46. Vargas SM, Huey SJ, Miranda J. A critical review of current evidence on multiple types of discrimination and mental health. Am J Orthopsychiatry 2020;90(3):374-390. [doi: 10.1037/ort0000441] [Medline: 31999138]
47. White Hughto JM, Pachankis JE, Willie TC, Reisner SL. Victimization and depressive symptomology in transgender adults: the mediating role of avoidant coping. J Couns Psychol 2017 Jan;64(1):41-51. [doi: 10.1037/cou0000184] [Medline: 28068130]
48. Yang MF, Manning D, van den Berg JJ, Operario D. Stigmatization and mental health in a diverse sample of transgender women. LGBT Health 2015 Dec;2(4):306-312. [doi: 10.1089/lgbt.2014.0106] [Medline: 26788771]
49. Birkett M, Newcomb ME, Mustanski B. Does it get better? A longitudinal analysis of psychological distress and victimization in lesbian, gay, bisexual, transgender, and questioning youth. J Adolesc Health 2015 Mar;56(3):280-285. [doi: 10.1016/j.jadohealth.2014.10.275] [Medline: 25586226]
50. Romanelli M, Lu W, Lindsey MA. Examining mechanisms and moderators of the relationship between discriminatory health care encounters and attempted suicide among U.S. transgender help-seekers. Adm Policy Ment Health 2018 Nov;45(6):831-849. [doi: 10.1007/s10488-018-0868-8] [Medline: 29574543]
51. Kcomt L. Profound health-care discrimination experienced by transgender people: rapid systematic review. Soc Work Health Care 2019 Feb 7;58(2):201-219. [doi: 10.1080/00981389.2018.1532941]
52. dickey LM, Singh AA. Evidence-based relationship variables: Working with trans and gender nonbinary clients. Pract Innov (Wash D C) 2020;5(3):189-201. [doi: 10.1037/pri0000116]
53. Holt NR, Hope DA, Mocarski R, Woodruff N. The often-circuitous path to affirming mental health care for transgender and gender-diverse adults. Curr Psychiatry Rep 2023 Mar;25(3):105-111. [doi: 10.1007/s11920-023-01410-2] [Medline: 36773177]
54. Budge SL, Sinnard MT, Hoyt WT. Longitudinal effects of psychotherapy with transgender and nonbinary clients: a randomized controlled pilot trial. Psychotherapy (Chic) 2021 Mar;58(1):1-11. [doi: 10.1037/pst0000310] [Medline: 32567869]
55. Ellis AE, Meade NG, Brown LS. Evidence-based relationship variables when working with affectional and gender minority clients: a systematic review. Pract Innov (Wash D C) 2020;5(3):202-217. [doi: 10.1037/pri0000118]
56. Pullmann MD, Rouvere J, Raue PJ, et al. Message-based vs video-based psychotherapy for depression: a randomized clinical trial. JAMA Netw Open 2025 Oct 1;8(10):e2540065. [doi: 10.1001/jamanetworkopen.2025.40065] [Medline: 41165707]
57. Areán PA, Pullmann MD, Griffith Fillipo IR, et al. Randomized trial of the effectiveness of videoconferencing-based versus message-pased psychotherapy on depression. Psychiatr Serv 2024 Dec 1;75(12):1184-1191. [doi: 10.1176/appi.ps.20230176] [Medline: 39026468]
58. Song J, Litvin B, Allred R, Chen S, Hull TD, Areán PA. Comparing message-based psychotherapy to once-weekly, video-based psychotherapy for moderate depression: randomized controlled trial. J Med Internet Res 2023 Jun 29;25:e46052. [doi: 10.2196/46052] [Medline: 37384392]
59. Henson P, Wisniewski H, Hollis C, Keshavan M, Torous J. Digital mental health apps and the therapeutic alliance: initial review. BJPsych Open 2019 Jan;5(1):e15. [doi: 10.1192/bjo.2018.86] [Medline: 30762511]
60. Kleinpeter CB, Kleinpeter JR, Brocato J, Pringle Hornsby E, Cheng J. Satisfaction and therapeutic alliance with mental health providers in teletherapy as opposed to face-to-face counseling for college students: diversity, equity, inclusion, and accessibility. Int J Health Wellness Soc 2024;14(4):119-141. [doi: 10.18848/2156-8960/CGP/v14i04/119-141]
61. Tremain H, McEnery C, Fletcher K, Murray G. The therapeutic alliance in digital mental health interventions for serious mental illnesses: narrative review. JMIR Ment Health 2020 Aug 7;7(8):e17204. [doi: 10.2196/17204] [Medline: 32763881]
62. Sharma A, Lin IW, Miner AS, Atkins DC, Althoff T. Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. Nat Mach Intell 2023;5(1):46-57. [doi: 10.1038/s42256-022-00593-2]
63. Online counseling for the LGBTQIA+ community. Talkspace. URL: https://www.talkspace.com/online-therapy/lgbtqia [accessed 2025-09-29]
64. Hoy-Ellis CP, Fredriksen-Goldsen KI. Depression among transgender older adults: general and minority stress. Am J Community Psychol 2017 Jun;59(3-4):295-305. [doi: 10.1002/ajcp.12138] [Medline: 28369987]
65. Hatchard T, Levitt EE, Mutschler C, et al. Transcending: a pragmatic, open-label feasibility study of a minority-stress–based CBT group intervention for transgender and gender-diverse emerging adults. Cogn Behav Pract 2025 Nov;32(4):499-512. [doi: 10.1016/j.cbpra.2024.04.005]
66. Poteat T, German D, Kerrigan D. Managing uncertainty: a grounded theory of stigma in transgender health care encounters. Soc Sci Med 2013 May;84:22-29. [doi: 10.1016/j.socscimed.2013.02.019] [Medline: 23517700]

XSL•FO
RenderX

67. Martos AJ, Wilson PA, Gordon AR, Lightfoot M, Meyer IH. “Like finding a unicorn”: healthcare preferences among lesbian, gay, and bisexual people in the United States. Soc Sci Med 2018 Jul;208:126-133. [doi: 10.1016/j.socscimed.2018.05.020] [Medline: 29803970]
68. Panchal Z, Piper C, Whitmore C, Davies RD. Providing supportive transgender mental health care: a systemized narrative review of patient experiences, preferences, and outcomes. J Gay Lesbian Ment Health 2022 Jul 3;26(3):228-264. [doi: 10.1080/19359705.2021.1899094]
69. Meyer HM, Mocarski R, Holt NR, Hope DA, King RE, Woodruff N. Unmet expectations in health care settings: experiences of transgender and gender diverse adults in the central Great Plains. Qual Health Res 2020 Feb;30(3):409-422. [doi: 10.1177/1049732319860265] [Medline: 31328642]
70. Romanelli M, Vélez-Grau C, Rouvere J, Porter SF. Intersectional disparities in digital health and mental health service use among US youth during the COVID-19 pandemic: cross-sectional analysis of a national survey. J Med Internet Res 2025 Oct 27;27:e77062. [doi: 10.2196/77062] [Medline: 41143468]
71. Bakhti R, Daler H, Ogunro H, Hope S, Hargreaves D, Nicholls D. Exploring engagement with and effectiveness of digital mental health interventions in young people of different ethnicities: systematic review. J Med Internet Res 2025 Apr 7;27:e68544. [doi: 10.2196/68544] [Medline: 40194267]

Abbreviations

DMH: digital mental health
HIPAA: Health Insurance Portability and Accountability Act
PHQ-8: 8-item Patient Health Questionnaire
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by A Mavragani; submitted 03.04.25; peer-reviewed by A Guy, E Boskey; revised version received 27.10.25; accepted 28.10.25; published 03.12.25.

Please cite as:
Romanelli M, Rouvere J, Sanders IA, Padmanabhan A, Peake E, Hull TD, Althoff T
Gender Differences in Trajectories of Depressive Symptoms Among Talkspace Clients: Naturalistic Observational Study
JMIR Form Res 2025;9:e75290
URL: https://formative.jmir.org/2025/1/e75290
doi:10.2196/75290

© Meghan Romanelli, Julien Rouvere, Isaac A Sanders, Aarthi Padmanabhan, Emily Peake, Thomas D Hull, Tim Althoff. Originally published in JMIR Formative Research (https://formative.jmir.org), 3.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

YouTube Searching and Self-Treatment Behaviors Among Patients With Benign Paroxysmal Positional Vertigo Before and After Clinic Visits: Prospective Observational Study

Guil Rhim[1], MD, PhD; Moon Jung Kim[2], MD, MPH, PhD

[1]Department of Otorhinolaryngology, One Dizziness Research Center, One Otorhinolaryngology Clinic, Paju, Republic of Korea

[2]Department of Laboratory Medicine, Myunggok Medical Research Center, Konyang University College of Medicine, 158 Gwanjeodong-ro, Daejeon, Republic of Korea

Corresponding Author:
Moon Jung Kim, MD, MPH, PhD
Department of Laboratory Medicine, Myunggok Medical Research Center, Konyang University College of Medicine, 158 Gwanjeodong-ro, Daejeon, Republic of Korea

Abstract

Background: YouTube has become a popular platform for patients seeking health-related information, including guidance on managing benign paroxysmal positional vertigo (BPPV). As self-diagnosis and self-treatment through online content grow more common, concerns have arisen regarding their influence on patients' health care decisions and treatment outcomes. However, little is known about how YouTube use and self-treatment behaviors change before and after clinical consultation, or whether these behaviors affect standard care for BPPV.

Objective: This study aimed to investigate changes in patients' YouTube searching and self-treatment behaviors before and after clinic visits for BPPV and to assess whether self-treatment influences standard in-clinic management.

Methods: A prospective study was conducted with patients diagnosed with BPPV who visited an otorhinolaryngology clinic in Korea from August 2024 to July 2025. On the final day of treatment, participants completed a survey, and chart reviews were performed to collect data on age, sex, canal involvement, chronic disease status, number of canalith repositioning maneuver (CRM) sessions, and pre- and postclinic YouTube searching and self-treatment. Differences in pre- and postclinic behaviors by gender were analyzed using Generalized Estimating Equations for repeated measures. The effect of self-treatment on the number of CRM sessions was assessed using negative binomial regression after confirming overdispersion.

Results: Among 147 patients (71% women), preclinic YouTube searching was reported by 28 (25%) patients, while postclinic searching decreased to 21 (14%) patients. Gender-stratified Generalized Estimating Equations analysis showed women had significantly higher odds of preclinic YouTube searching compared to postclinic (odds ratio [OR] 2.389, 95% CI 1.195 - 4.778, *P*=.01). Additionally, women with chronic disease had significantly lower odds of self-treatment (OR 0.13, 95% CI 0.016 - 0.976, *P*=.047). Negative binomial regression showed no significant association between self-treatment status and the number of CRM sessions.

Conclusions:: This study demonstrates that YouTube searching and self-treatment behaviors for BPPV change following clinical consultation. These findings highlight the importance of patient education during clinical encounters in addressing previsit online information use and mitigating inappropriate self-treatment practices.

(JMIR Form Res 2025;9:e80966) doi:10.2196/80966

KEYWORDS

YouTube; health information; benign paroxysmal positional vertigo; behavior; self-treatment

Introduction

The influence of social media on health, public health, and disease management through information searching and acquisition is increasing [1]. In particular, the use of YouTube for obtaining health-related information is steadily rising, and its impact on health is becoming increasingly significant [2]. As a video-sharing platform, YouTube provides viewers with tools that facilitate the easy replication of behaviors observed in videos. Benign paroxysmal positional vertigo (BPPV) is a condition characterized by the displacement of otoconia from the maculae of the utricle into one or more of the three semicircular canals or their adhesion to the cupula, causing clinical symptoms due to positional changes [3]. The diagnosis of BPPV is based on identifying characteristic nystagmus, along with a thorough history taking and physical examination, using methods such as the Dix-Hallpike test and the head roll test [4,5]. The standard and highly effective treatment for BPPV is

XSL•FO
RenderX

the canalith repositioning maneuver (CRM), with success rates exceeding 80% when the diagnosis is accurate [6].

The first study evaluating YouTube video content related to BPPV was published in 2012 [7]. Since then, only a few additional studies have examined BPPV-related YouTube content, and most of these have primarily focused on evaluating the validity of the information. While YouTube contains information on almost all diseases, numerous studies have assessed the reliability and accuracy of this content across various medical topics.

What distinguishes BPPV from other conditions is that YouTube hosts video content specifically on CRM, the standard treatment for BPPV, which patients can follow to potentially achieve direct therapeutic effects. In Korea, BPPV is also the most common cause of dizziness [8], and a significant amount of information about BPPV is available through YouTube searches. People in Korea often search for BPPV using Korean-language terms when looking for information online.

In our recent clinical practice, we have observed an increasing number of patients with dizziness symptoms who report searching for BPPV information on YouTube or attempting self-treatment before visiting a hospital. While previous studies focused on evaluating YouTube content from an expert perspective, this study adopts a user-centered approach to explore how patients experience and use BPPV-related information on YouTube. Specifically, it aims to examine how YouTube searching and self-treatment behaviors change before and after clinic visits, and to assess whether self-treatment influences standard in-clinic management for BPPV.

Methods

Study Design

This prospective study was conducted with patients diagnosed with BPPV who visited the otorhinolaryngology clinic due to vertigo symptoms from August 2024 to July 2025. A total of 149 patients met the eligibility criteria, and 147 of them were included in the final analysis. When a patient presented with vertigo symptoms and abnormal findings were noted during medical history taking, general otolaryngological examination, and physical examination using Frenzel glasses, videonystagmography tests were performed. After a diagnosis of BPPV was confirmed, CRMs were performed once per session, typically 2 - 3 times per week. The specific maneuver was selected according to the involved canal: the modified Epley method was used for canalolithiasis of the posterior semicircular canal, the Barbecue rotation method was used for canalolithiasis of the lateral semicircular canal, and the Appiani and Gufoni methods were used for lateral-canal cupulolithiasis. In cases of multicanal involvement, only one canal was treated per session. On the final day of treatment, a survey was administered, and patient chart reviews were conducted to collect information on sex, age, number of CRM sessions, chronic disease status (hypertension or diabetes), canal involvement, preclinic YouTube searching, preclinic self-treatment, YouTube searching during treatment (postclinic), and self-treatment during treatment.

Ethical Considerations

This study was approved by the Korea National Institute for Bioethics Policy (P01-202408-01-036). All patients gave their informed consent prior to their inclusion in the study, and this study has been performed in accordance with the ethical standards in the 1964 Declaration of Helsinki and its later amendments. Data collection was conducted in July 2025 for research purposes. All data were fully anonymized prior to analysis, and the authors had no access to personally identifiable information. Participants did not receive any financial or nonfinancial compensation for their participation in this study.

Statistical Analysis

Descriptive statistics were used to summarize the included studies by author, origin country, publication type, research focus, and primary topic area. Differences in pre- and postclinic YouTube searching and self-treatment behavior by gender were analyzed using Generalized Estimating Equations (GEE) for repeated measures. To assess the effect of self-treatment on the number of CRM procedures, Poisson regression was initially applied, but overdispersion was confirmed (Value/df=372), leading to the use of a negative binomial regression model. A P value of <0.05 was considered statistically significant. All analyses were conducted using SPSS software (version 19.0; IBM Corp).

Results

A total of 147 patients were included in the analysis, comprising 104 (71%) women and 43 (29%) men. The age of women ranged from 17 to 86 years (mean 55.6, SD 13.3), while that of men ranged from 11 to 80 years (mean 52.7, SD 15.5). Among the 147 patients, regarding canal involvement, 18 patients (12%) had posterior canal BPPV, 56 (38%) had lateral canal involvement, 41 (28%) had multiple canal involvement, and 32 (22%) had combined canal and cupulolithiasis involvement. A history of hypertension or diabetes was present in 53/147 patients (36%).

Until the end of treatment, the number of CRM performed was 2 sessions in 21/147 patients (14%), 3 sessions in 48 patients (33%), 4 sessions in 51 patients (35%), 5 sessions in 17 patients (12%), and 6 sessions in 10 patients (6%), with a mean of 3.6 (SD 1.0).

Preclinic YouTube searching was reported by 36/147 patients (women: n=28, 19%; men: n=8, 6%), and preclinic self-treatment was reported by 11 patients (women: n=9, 7%; men: n=2, 1%). Postclinic YouTube searching was reported by 21 patients (women: n=14, 10%; men: n=7, 5%), and postclinic self-treatment was reported by 5 patients (women: n=4, 3%; men: n=1, 1%).

In the gender-stratified GEE repeated measures analysis of YouTube searching behavior, women showed significantly higher odds of searching before the clinic visit compared to after (odds ratio [OR] 2.389, 95% CI 1.195 - 4.778; P=.01) (Table 1).

XSL•FO
RenderX

Table . Gender-stratified generalized estimating equations results for preclinic versus postclinic YouTube searching.

Parameter	β(SE)	Hypothesis test		Odds ratio	95% Wald CI	
		Wald Chi-square[a]	*P* value		Lower	Upper
Women						
Chronic disease[b]	−0.530 (0.4316)	1.508	.22	0.589	0.253	1.372
Age	−0.003 (0.0123)	0.067	.80	0.997	0.973	1.021
Preclinic[c]	0.871 (0.3535)	6.071	.01	2.389	1.195	4.778
Men						
Chronic disease	1.038 (0.8968)	1.339	.25	2.823	0.487	16.370
Age	−0.026 (0.0231)	1.239	.27	0.975	0.932	1.020
Preclinic	0.353 (0.4955)	0.509	.48	1.424	0.539	3.761

[a]All degrees of freedom are 1.

[b]Chronic disease: diabetes or hypertension

[c]Preclinic: before visiting the clinic.

In the gender-stratified GEE analysis of self-treatment behavior, women with chronic disease had significantly lower odds of self-treatment compared to those without chronic disease (OR 0.125, 95% CI 0.016 - 0.976; *P*=.047) (Table 2).

Table . Gender-stratified generalized estimating equations results for preclinic versus postclinic self-treatment.

Parameter	β (SE)	Hypothesis test		Odds ratio	95% Wald CI	
		Wald Chi-square[a]	*P* value		Lower	Upper
Women						
Chronic disease[b]	−0.530 (0.4316)	3.934	.047	0.125	0.016	0.976
Age	−0.003 (0.0123)	0.648	.42	1.010	0.985	1.036
Preclinic[c]	0.871 (0.3535)	2.214	.14	2.417	0.756	7.727
Men						
Chronic disease	1.038 (0.8968)	1.384	.24	10.711	0.206	556.192
Age	−0.026 (0.0231)	2.691	.10	0.945	0.884	1.011
Preclinic	0.353 (0.4955)	0.329	.57	2.090	0.169	25.895

[a]All degrees of freedom are 1.

[b]Chronic disease: diabetes or hypertension.

[c]Preclinic: before visiting the clinic.

Negative binomial regression analysis evaluating the impact of self-treatment status (preclinic or postclinic) on the number of CRM sessions revealed no statistically significant association (*P*=.92 and *P*=.97, respectively) (Table 3).

Table . Negative binomial regression results evaluating the impact of self-treatment timing on the number of canalith repositioning maneuvers.

Parameter	β (SE)	Hypothesis test		Odds ratio
		Wald Chi-square (*df*)	*P* value	
Preclinic[a]	0.038 (0.3567)	0.011 (1)	.92	1.038
Postclinic[b]	−0.019 (0.5188)	0.001 (1)	.97	0.981

[a]Preclinic: before visiting the clinic.

[b]Postclinic: during the treatment period.

Table 4 presents an overview of the included studies by author, origin country, publication type, research focus, and primary topic area. This summary provides a contextual understanding of the existing literature on BPPV, YouTube-based health information, and related digital health research.

XSL•FO
RenderX

Table . Descriptive characteristics of included studies.

No.	Author (year)	Country	Publication type	Focus area	Main topic / Summary
1	Grajales FJ et al (2014) [1]	USA	Journal article	Medical Education	Overview of social media applications in health care
2	Madathil KC et al (2015) [2]	USA	Journal article	Digital Health / Information Quality	YouTube as a source of health information
3	Rhim GI (2018) [3]	Korea	Journal article	Clinical	Serum vitamin D and long-term outcomes in BPPV[a]
4	Dix MR et al (1952) [4]	UK	Journal article	Clinical	Foundational study on vestibular system pathology
5	McClure JA (1985) [5]	Canada	Journal article	Clinical	Description of horizontal canal BPPV
6	Bhattacharyya N et al (2017) [6]	USA	Journal article	Clinical Practice	clinical guideline for BPPV
7	Kerber KA et al (2012) [7]	USA	Journal article	Digital Health / Education	Validity of Epley maneuver videos on YouTube
8	Kim HJ et al (2020) [8]	Korea	Journal article	Clinical Epidemiology	Etiologic distribution of dizziness
9	Delli K et al (2016) [9]	Netherlands	Journal article	Digital Health / Information Quality	YouTube as a source of information
10	Biggs TC et al (2013) [10]	UK	Journal article	Digital Health / Public Education	YouTube information quality on rhinosinusitis
11	Hassona Y et al (2016) [11]	Jordan	Journal article	Digital Health / Information	YouTube as a source of oral cancer information
12	Landrum AR (2021) [12]	USA	Journal article	Science Communication	Gender differences in science audiences on YouTube
13	Langford AT et al (2022) [13]	USA	Journal article	Digital Health / Information	YouTube use among patients with chronic illness
14	Lim MSC et al (2022) [14]	Australia	Journal article	Digital Health / Behavior	Social media use for health information among young adults
15	Khan ML. (2017) [15]	USA	Journal article	Communication / Behavioral Science	Motivations for user engagement on YouTube
16	Moll R et al (2017) [16]	Canada	Journal article	Science Education	Development of social media and science learning survey
17	Naslund JA et al.\ (2014) [17]	USA	Journal article	Mental Health / Social Support	Peer support via YouTube for individuals with mental illness
18	Bujnowska-Fedak MM et al (2020) [18]	Poland	Journal article	Digital Health / Behavior	Impact of online health information on patient decisions
19	Mohamed F et al (2024) [19]	UAE	Journal article	Digital Health / Decision Tool	User experience with health-related YouTube content

XSL•FO
RenderX

No.	Author (year)	Country	Publication type	Focus area	Main topic / Summary
20	Kerber KA et al (2022) [20]	USA	Journal article	Behavioral Intervention	Development and evaluation of instructional video for BPPV self-management
21	Kim HJ et al (2023) [21]	Korea	Journal article	Clinical Trial	Effect of self-treatment for recurrent BPPV
22	Liu T et al (2025) [22]	China	Journal article	Digital Health / Social Learning Theory	Five-dimensional extension of Social Learning Theory
23	Lau Y et al (2025) [23]	Singapore	Journal article	AI-Based[b] Psychotherapy	Therapeutic efficacy of AI-based interventions

[a]BPPV: benign paroxysmal positional vertigo.

[b]AI: artificial intelligence.

Discussion

Principal Results

Social media, including YouTube, are dynamic and interactive computer-mediated communication tools with high penetration rates in the general population in high-income and middle-income countries. However, in medicine and health care, many stakeholders remain unaware of social media's relevance [1]. CRMs, the standard treatment for BPPV, typically take about 10‑15 minutes to perform. Because of the procedural complexity, patients cannot accurately perform self-treatment based on short-form content or brief blog posts. Currently, YouTube is the only widely accessible platform that allows patients to visually observe and replicate these maneuvers. Therefore, almost all previous studies investigating online resources related to BPPV have focused exclusively on YouTube as the primary platform. Analyzing the overall results of YouTube content related to health reveals several key issues. Much of the information is misleading, with a high probability that consumers will encounter such content during their searches [2]. Posts created by professional organizations are more reliable. In contrast, videos uploaded by individual users are less useful [9,10]. Additionally, there is little difference in view counts between misleading and accurate posts. Most studies have focused on evaluating the validity of the information, classifying it as valid, erroneous, or anecdotal [11].

Public media serves as an avenue for informal science learning, and women may engage with YouTube content differently than men, viewing it more as an educational resource. Women are often motivated by practical or instrumental needs rather than by pure curiosity [12]. Our study found that women's YouTube searching behavior significantly decreased after visiting the clinic, unlike men, whose search rates remained nearly unchanged. Specifically, the proportion of women engaging in YouTube searching was reduced by half following their clinic visit, suggesting that clinical consultation may have a stronger impact on modifying information-seeking behavior among women.

Previous studies have reported that approximately 30%‑35% of adults with chronic conditions such as hypertension and diabetes use YouTube to watch health-related videos [13]. In our study, despite BPPV being an acute condition, the preclinic YouTube searching rate was 25%, suggesting that patients with acute dizziness symptoms are also highly likely to seek online information before clinic visits. This relatively high level of previsit searching underscores the need for clinicians to address online health information use during consultations.

Patients often find the video format on YouTube easy to learn from and tend to evaluate the accuracy of health information by cross-checking multiple videos or judging the credibility of presenters [14].

Previous research has identified several motivations for YouTube use, such as information-seeking, social interaction, and entertainment [15,16]. Studies have also documented gender differences in YouTube usage patterns. Generally, men are more likely to watch educational or science-related content, whereas women more frequently view health and wellness videos. Moreover, the motivations for engaging with science-related content differ by gender—men more often seek entertainment or general information, while women tend to use such videos for learning or practical purposes [12]. In line with these previous findings, our study also showed that women were more likely than men to use YouTube videos as a tool for self-treatment related to BPPV.

People are increasingly using YouTube as a platform to seek support from others with similar health conditions, find answers to health-related questions, and gather information before planned medical procedures [17]. The impact of YouTube on users' health-related behaviors and decisions is substantial; more than 50% of surveyed users reported that online health-related content influenced their decisions regarding health [18].

According to a recent study by Mohamed et al, 87.6% (n=2630) of participants watched health-related content on YouTube, and 84.7% (n=2542) reported making decisions based on the content viewed. Among users, 40% watched YouTube videos to decide whether to consult a physician or adopt a specific health-related practice [19]. In this study, women with chronic disease showed significantly lower odds of attempting self-treatment. This suggests that these patients may be more cautious about adopting unsupervised online practices and more likely to rely on professional guidance.

XSL•FO
RenderX

Research on BPPV and YouTube includes studies that evaluated the accuracy of videos on the treatment of posterior canal BPPV using the Epley maneuver. In one study, 21 out of 33 videos were found to be valid and potentially useful for patient home treatment or educational purposes [7]. Other studies have focused on evaluating educational videos for treating posterior canal BPPV following a physician's diagnosis and providing a web-based diagnosis and treatment for BPPV in cases of symptom recurrence [20,21]. While these studies focused on the validity and accuracy of YouTube or video content, our study is distinct in that it adopts a user-centered approach.

However, this study found that prior self-treatment did not significantly affect the number of in-clinic CRM sessions required, suggesting that while inappropriate self-treatment may complicate diagnosis, it does not necessarily increase the overall treatment burden when managed in a clinical setting.

According to Liu et al [22], digital health behavior among older adults was conceptualized by extending Social Learning Theory into five dimensions. While our study focuses on actual clinical and behavioral aspects, both studies intersect in the mechanism of self-learning through observation. In particular, the observational learning dimension corresponds to visual imitation and behavioral acquisition through watching YouTube videos, whereas self-efficacy relates to patients' confidence and belief that they can perform self-treatment independently after viewing such instructional content [22].

Lau et al [23] conducted a meta-analysis that quantitatively verified the therapeutic efficacy of artificial intelligence, AI-based psychotherapeutic interventions. In the future, the findings of our study may help overcome the limitations and risks associated with nonprofessional self-treatment and contribute to the development of AI-assisted approaches for the management of BPPV [23].

Limitations

This study has several limitations. Because the questionnaire was administered after treatment, there is a potential for recall bias, as some participants may have inaccurately reported their prior exposure to YouTube or other online resources. There is a possibility of a type II error in the study sample. Patients whose symptoms improved through self-treatment using YouTube videos might not have visited the hospital, thereby being excluded from the survey and potentially affecting the study results. Additionally, this study's relatively small sample size did not allow for an investigation of other sociodemographic factors, such as occupation and education level, and their association with YouTube usage.

Conclusion

Our findings show that clinical consultation leads to meaningful changes in YouTube searching and self-treatment behaviors for BPPV, especially among women. While prior self-treatment attempts did not increase the number of CRM sessions needed, our findings underscore the importance of addressing online health information use and supporting patient education during clinical care.

Acknowledgments

The authors express their gratitude to One dizziness Research Center for providing the resources and support necessary to complete this research.

All participants provided written informed consent to participate in the study.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

GI contributed to conceptualization, data curation, methodology, formal analysis, resources, visualization, and writing - original draft preparation. MJ was responsible for software, validation, investigation, writing - review and editing, supervision, project administration.

Conflicts of Interest

None declared.

References

1. Grajales FJ 3rd, Sheps S, Ho K, Novak-Lauscher H, Eysenbach G. Social media: a review and tutorial of applications in medicine and health care. J Med Internet Res 2014 Feb 11;16(2):e13. [doi: 10.2196/jmir.2912] [Medline: 24518354]
2. Madathil KC, Rivera-Rodriguez AJ, Greenstein JS, Gramopadhye AK. Healthcare information on YouTube: a systematic review. Health Informatics J 2015 Sep;21(3):173-194. [doi: 10.1177/1460458213512220]

XSL•FO
RenderX

3. Rhim GI. Serum vitamin D and long-term outcomes of benign paroxysmal positional vertigo. Clin Exp Otorhinolaryngol 2019 Aug;12(3):273-278. [doi: 10.21053/ceo.2018.00381]
4. Dix MR, Hallpike CS. The pathology symptomatology and diagnosis of certain common disorders of the vestibular system. Proc R Soc Med 1952 Jun;45(6):341-354. [doi: 10.1177/003591575204500604] [Medline: 14941845]
5. McClure JA. Horizontal canal BPV. J Otolaryngol 1985 Feb;14(1):30-35. [Medline: 4068089]
6. Bhattacharyya N, Gubbels SP, Schwartz SR, et al. Clinical Practice Guideline: Benign Paroxysmal Positional Vertigo (Update). Otolaryngol Head Neck Surg 2017 Mar;156(3_suppl):S1-S47. [doi: 10.1177/0194599816689667] [Medline: 28248609]
7. Kerber KA, Burke JF, Skolarus LE, et al. A prescription for the Epley maneuver: www.youtube.com? Neurology (ECronicon) 2012 Jul 24;79(4):376-380. [doi: 10.1212/WNL.0b013e3182604533] [Medline: 22826542]
8. Kim HJ, Lee JO, Choi JY, Kim JS. Etiologic distribution of dizziness and vertigo in a referral-based dizziness clinic in South Korea. J Neurol 2020 Aug;267(8):2252-2259. [doi: 10.1007/s00415-020-09831-2] [Medline: 32300888]
9. Delli K, Livas C, Vissink A, Spijkervet FKL. Is YouTube useful as a source of information for Sjögren's syndrome? Oral Dis 2016 Apr;22(3):196-201. [doi: 10.1111/odi.12404] [Medline: 26602325]
10. Biggs TC, Bird JH, Harries PG, Salib RJ. YouTube as a source of information on rhinosinusitis: the good, the bad and the ugly. J Laryngol Otol 2013 Aug;127(8):749-754. [doi: 10.1017/S0022215113001473] [Medline: 23866821]
11. Hassona Y, Taimeh D, Marahleh A, Scully C. YouTube as a source of information on mouth (oral) cancer. Oral Dis 2016 Apr;22(3):202-208. [doi: 10.1111/odi.12434] [Medline: 26718020]
12. Landrum AR. Are women a missing audience for science on YouTube? An exploratory study. Front Commun 2021;6:610920. [doi: 10.3389/fcomm.2021.610920]
13. Langford AT, Orellana KT, Buderer N. Use of YouTube to watch health-related videos and participation in online support groups among US adults with heart disease, diabetes, and hypertension. Digit Health 2022;8:20552076221118822. [doi: 10.1177/20552076221118822] [Medline: 36046636]
14. Lim MSC, Molenaar A, Brennan L, Reid M, McCaffrey T. Young adults' use of different social media platforms for health information: insights from web-based conversations. J Med Internet Res 2022 Jan 18;24(1):e23656. [doi: 10.2196/23656] [Medline: 35040796]
15. Khan ML. Social media engagement: what motivates user participation and consumption on YouTube? Comput Human Behav 2017 Jan;66:236-247. [doi: 10.1016/j.chb.2016.09.024]
16. Moll R, Nielsen W. Development and validation of a social media and science learning survey. International Journal of Science Education, Part B 2017 Jan 2;7(1):14-30. [doi: 10.1080/21548455.2016.1161255]
17. Naslund JA, Grande SW, Aschbrenner KA, Elwyn G. Naturally occurring peer support through social media: the experiences of individuals with severe mental illness using YouTube. PLoS ONE 2014;9(10):e110171. [doi: 10.1371/journal.pone.0110171] [Medline: 25333470]
18. Bujnowska-Fedak MM, Węgierek P. The impact of online health information on patient health behaviours and making decisions concerning health. Int J Environ Res Public Health 2020 Jan 31;17(3):880. [doi: 10.3390/ijerph17030880] [Medline: 32023828]
19. Mohamed F, Shoufan A. Users' experience with health-related content on YouTube: an exploratory study. BMC Public Health 2024 Jan 3;24(1):86. [doi: 10.1186/s12889-023-17585-5] [Medline: 38172765]
20. Kerber KA, Carender W, Telian SA, Fagerlin A, Tsodikov A, Meurer WJ. Patient self-management of benign paroxysmal positional vertigo: instructional video development and preliminary evaluation of behavioral outcomes. Otol Neurotol 2022 Jan 1;43(1):e105-e115. [doi: 10.1097/MAO.0000000000003360] [Medline: 34607997]
21. Kim HJ, Kim JS, Choi KD, et al. Effect of self-treatment of recurrent benign paroxysmal positional vertigo. JAMA Neurol 2023 Mar 1;80(3):244. [doi: 10.1001/jamaneurol.2022.4944]
22. Liu T, Luo YT, Pang PCI. Digital technologies-enhanced older adults health management: developing a five-dimensional extension of social learning theory for community settings. Front Public Health 2025;13:1627983. [doi: 10.3389/fpubh.2025.1627983] [Medline: 40761933]
23. Lau Y, Ang WHD, Ang WW, Pang PCI, Wong SH, Chan KS. Artificial intelligence–based psychotherapeutic intervention on psychological outcomes: a meta - analysis and meta - regression. Depress Anxiety 2025 Jan;2025(1):8930012. [doi: 10.1155/da/8930012]

Abbreviations

AI: artificial intelligence
BPPV: benign paroxysmal positional vertigo
CRM: canalith repositioning maneuver
GEE: Generalized Estimating Equations
OR: odds ratio

XSL•FO
RenderX

Edited by A Schwartz, M Balcarras; submitted 19.07.25; peer-reviewed by A Melliti, PCI Pang; revised version received 22.10.25; accepted 14.11.25; published 01.12.25.

Please cite as:
Rhim G, Kim MJ
YouTube Searching and Self-Treatment Behaviors Among Patients With Benign Paroxysmal Positional Vertigo Before and After Clinic Visits: Prospective Observational Study
JMIR Form Res 2025;9:e80966
URL: https://formative.jmir.org/2025/1/e80966
doi:10.2196/80966

© Guil Rhim, Moon Jung Kim. Originally published in JMIR Formative Research (https://formative.jmir.org), 1.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Developing eHealth Interventions to Improve Diabetes Management in Emerging Adulthood: Qualitative Formative Study

April Idalski Carcone[1], MSW, PhD; Deepika Baskar[2], BS; Aishah Mahmood[1], BS; Karen MacDonell[3], PhD; Susan Eggly[4], PhD; Samiran Ghosh[5], PhD; Colleen Buggs-Saxton[6], MD, PhD; Steven J Ondersma[7], PhD; Deborah A Ellis[1], PhD

[1]Department of Family Medicine and Public Health Sciences, School of Medicine, Wayne State University, 6135 Woodward Avenue, 3128 Integrative Biosciences Center, Detroit, MI, United States

[2]Department of Chronic Disease Epidemiology, School of Public Health, Yale University, New Haven, CT, United States

[3]Center for Translational Behavioral Science, Department of Behavioral Sciences and Social Medicine, Florida State University, Tallahassee, FL, United States

[4]Department of Oncology, School of Medicine, Wayne State University, Detroit, MI, United States

[5]Department of Biostatistics and Data Science, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX, United States

[6]Department of Pediatrics, School of Medicine, Wayne State University, Detroit, MI, United States

[7]Charles Stewart Mott Department of Public Health, College of Human Medicine, Michigan State University, Flint, MI, United States

Corresponding Author:
April Idalski Carcone, MSW, PhD
Department of Family Medicine and Public Health Sciences, School of Medicine, Wayne State University, 6135 Woodward Avenue, 3128 Integrative Biosciences Center, Detroit, MI, United States

Abstract

Background: Emerging adulthood is a high-risk period during which many with type 1 diabetes demonstrate suboptimal diabetes management and glycemic control. There is a need for effective, scalable interventions designed specifically for this population. Technology-based approaches are readily accessible to this age group. Furthermore, interventions consistent with self-determination theory—which posits that the fulfillment of psychological needs for autonomy, self-efficacy, and relatedness promotes intrinsic motivation for change—may resonate well with emerging adults' developmental needs for establishing independence and autonomy, and growing their social network.

Objective: This study aimed to enhance the potential relevance, sustainability, and efficacy of 3 self-determination theory–informed mobile health intervention components and content for emerging adults with type 1 diabetes. Key areas of interest included emerging adults' perspectives on the use of cultural tailoring, developmental relevance of content, and delivery preferences.

Methods: In this qualitative formative study, 20 emerging adults reviewed and provided feedback on 3 newly developed intervention components via individual interviews. Ten reviewed the motivation enhancement system, a 2-session counseling intervention grounded in motivational interviewing and designed to enhance emerging adults' autonomy and self-efficacy for diabetes self-management. Ten reviewed the SMS text messaging reminder intervention (one-way text message reminders to complete diabetes care) and the question prompt list (a list of questions related to diabetes care designed to increase patients' active participation during medical visits). Interviews were analyzed using framework matrix analysis, an efficient approach to inductive thematic analysis.

Results: Emerging adults found all 3 interventions acceptable and helpful. They noted the interventions' *integration into the technology* they already use as a strength. Across interventions, emerging adults also expressed a preference for culturally tailored intervention content, including intervention examples, actors, and language representing their illness experience, identity, and personal preferences. Intervention-specific feedback suggested emerging adults liked motivation enhancement system intervention elements that were engaging (videos) and relatable (peer testimonials), and supported their growing autonomy and independence. For SMS text messaging reminders, emerging adults appreciated the straightforward nature of the reminders and recommended more directive messages. They appreciated the range of topics and variety of messages. Suggestions included making the messages more impactful (eg, direct, personalized, and engaging, such as using emojis). Emerging adults saw the question prompt list content areas as relevant and well-aligned with their concerns highlighting the topic of transitioning to adult life with diabetes as particularly salient.

Conclusions: Emerging adult feedback supports the acceptability and use of these intervention components and will be used to refine the interventions. Feedback was especially positive regarding cultural and other tailoring efforts, as well as content

XSL•FO
RenderX

directed at their pending transition to full independence. At the same time, their input suggests the need for multiple specific modifications, highlighting the importance of intensive and detailed feedback from end users.

(JMIR Form Res 2025;9:e75623) doi:10.2196/75623

KEYWORDS

emerging adults; young adults; type 1 diabetes; behavioral intervention; eHealth

Introduction

Type 1 diabetes (T1D) management is complex, demanding, and primarily under the control of the individual [1]. Diabetes management often deteriorates during adolescence, affecting every component of diabetes care [2-5]. Once thought to be temporary, evidence suggests suboptimal diabetes management and glycemic control can persist into emerging adulthood [6,7]. Emerging adulthood spans adolescence and young adulthood (roughly aged 18 - 25 y) and is characterized by growing independence, identity consolidation, and multiple life transitions [8]. Some have described emerging adulthood as the "most volatile years of life" [8].

Emerging adults self-report suboptimal diabetes management [5] demonstrating infrequent blood glucose monitoring (<1 test per day [9]) and suboptimal continuous glucose monitoring (CGM) wear rates [10]. Emerging adults have the worst glycemic health indicators of any age group [11], with hemoglobin A_{1c} (HbA_{1c}) levels averaging from 8.4% to 9.3% [5,9] (levels <7% are recommended [12]); 83% fail to meet glycemic recommendations [9]. More concerning, by the age of 21 years, 32% of emerging adults have at least 1 diabetes-related complication or comorbidity [13]. Suboptimal diabetes management in emerging adults is related to decreased parental involvement in daily activities and increased reliance on peer and other support [8], transitioning from pediatric to adult clinical care, and infrequent clinic attendance [14].

While the challenges of managing T1D during emerging adulthood are well described [15-18], few efficacious interventions specifically designed for emerging adults with T1D have been developed. To date, the majority of emerging adult T1D intervention research has examined the use of advanced diabetes management technology (eg, CGM) [19-21] and closed-loop insulin delivery [22,23] or on optimizing the transition from the pediatric to the adult health care systems [24-26]. Other approaches are rare, with mixed results. In a pilot clinical trial of REAL-T, an occupational therapy intervention targeting diabetes management and its social ecology, the intervention improved glycemic control (HbA_{1c}), diabetes-related quality of life, and blood glucose monitoring habits among emerging adults with T1D and type 2 diabetes. However, blood glucose monitoring and medication adherence behaviors were unchanged, as were mediators of behavior change (diabetes knowledge, problem-solving, and self-efficacy) [27]. Subsequently, a fully powered clinical trial of REAL-T for emerging adults with T1D, delivered via telehealth, showed improved diabetes-related quality of life and diabetes distress, but no change in glycemic outcomes, diabetes self-management, self-efficacy, or autonomy [28]. The Colorado Young Adults with T1D intervention combined telehealth clinical diabetes care with virtual peer support groups to improve diabetes outcomes (HbA_{1c}, diabetes-related distress, depression, and self-efficacy) [29]. In a randomized controlled trial, per protocol analyses showed that participants in the active condition demonstrated significant improvement in glycemic control and CGM use, but depression symptoms increased, and diabetes distress and self-efficacy were unchanged. Intention-to-treat analyses were not published, and intervention participation rates were low, as 53% did not attend a single session. FAMS-T1D (Family and Friend Activation to Motivate Self-Care), a mobile health (mHealth) intervention to improve emerging adults' self- and social-regulation, was tested in a pilot clinical trial (single group, and pretest or posttest design) and shown to be well-received (79% of those approached agreed to participate and 93% were retained). After intervention, emerging adults demonstrated improved diabetes management, self-efficacy, and diabetes distress, but changes in glycemic control were not statistically or clinically significant [30]. Other interventions for emerging adults demonstrate a similar pattern of either low engagement or limited improvement in glycemic outcomes [31-34].

Capitalizing on emerging adults' unique developmental stage, in which feelings of autonomy and self-efficacy are paramount, may be key to intervention efficacy. Self-determination theory (SDT) [35] is an empirically derived theory of human motivation that posits three innate psychological needs: competence, relatedness, and autonomy. According to SDT, fulfilling these needs enhances intrinsic motivation and facilitates behavior change. To illustrate, autonomy-supportive communication by diabetes providers has been associated with better diabetes management in adolescents [36] and adults [37]. Because of their salient developmental need for independence, autonomy-focused interventions grounded in SDT may be particularly effective with emerging adults.

Unfortunately, there are few examples of autonomy-supportive interventions to improve diabetes outcomes in emerging adults with T1D. Guided self-determination (GSD) is an autonomy-oriented approach to shared decision-making involving the completion of 8 reflection worksheets over 8 to 12 months, which are discussed in 1-hour sessions with diabetes clinicians. While GSD increased adolescents' perceptions of autonomy and decreased amotivation for diabetes self-management, it did not improve glycemic control [38]. Furthermore, 20% refused to participate, and 30% of those allocated to GSD did not complete the intervention. GSD for adults was reduced to 7 worksheets, or sessions, and clinical trial results indicated GSD improved diabetes distress, diabetes competence, amotivation, and perceptions of provider autonomy-supportive communication, but only improved glycemic control for women in the trial [39]. In addition, 25%

XSL•FO
RenderX

of participants (age: mean 25.7, SD 5.1 y) either discontinued or failed to initiate treatment; completers required additional sessions (median 8; *R*=5 - 14) to complete treatment. An online version of GSD for emerging adults improved autonomous motivation and perceptions of provider support for autonomy but did not impact glycemic control [31]. Online GSD was insufficient to improve engagement, as 65% of emerging adults failed to initiate treatment, only 28% of initiators completed treatment, and 66% were lost to follow-up at 6 months. Autonomy-supportive T1D camp (a 6-day experience intentionally designed to leverage SDT during diabetes education, meal planning assistance, camper-led parent training, glucose management, as well as typical camp activities) increased adolescents' (aged 13 - 18 y) sense of relatedness, but autonomy and competence remained unchanged; glycemic control was not examined [40]. Neither study examined the effect of autonomy-supportive interventions on diabetes self-management behavior.

This study was conducted in preparation for a larger clinical trial testing an autonomy-supportive intervention to improve diabetes self-management among older adolescents and emerging adults (aged 16 - 25 y) with T1D and an elevated HbA_{1c} ≥9% [41]. The goal of this formative research was to enhance the potential relevance, sustainability, and efficacy of 3 SDT-informed mHealth intervention components to promote diabetes management and glycemic control and related intervention content for emerging adults with T1D by soliciting feedback from emerging adults with T1D and their recommendations for improving the interventions. Key areas of interest included emerging adults' perspectives on the use of cultural tailoring, developmental relevance of content, and preferences for delivery logistics.

Methods

Design

Emerging adults with T1D participated in a single study visit during which they completed a brief demographic survey and reviewed either the motivation enhancement system (MES) intervention (10/20, 50%) or the text and question prompt list (QPL) interventions (10/20, 50%). Data were collected via a semistructured qualitative interview to obtain feedback on intervention content, delivery format, and suggestions for improvement.

Participants

Emerging adults were recruited from a university-affiliated tertiary care outpatient diabetes clinic located in Detroit, Michigan, a large metropolitan city in the midwestern United States. Study inclusion criteria were individuals aged 16 to 25 years (inclusive) with a diagnosis of T1D for at least 6 months and HbA_{1c} greater than or equal to 9% currently and over the prior 6 months, English language fluency, and residence within 30 miles of the recruitment location. Exclusion criteria were determined via medical chart review and limited to medical diagnoses requiring atypical diabetes care or mental health conditions and developmental learning disabilities that resulted in the inability to independently participate in the study.

Potentially eligible emerging adults were identified via medical chart review. The initial data pull identified 701 potentially eligible emerging adults. Of these, 511 (73%) did not meet one or more study inclusion criteria. Research staff mailed a study introduction letter to 190 study-eligible emerging adults, or their caregivers when the emerging adults were aged <18 years. After a 2-week opt-out period, research staff followed up these letters with telephone calls to explain the details of the study, confirm study eligibility, and enroll interested, eligible emerging adults. Emerging adults were enrolled on a rolling basis until the target sample size of 10 emerging adults to review the MES intervention was achieved. This process was repeated for the review of the text and QPL interventions. A total of 20 emerging adults were enrolled, and 4 declined to participate.

Procedures

Standardized procedures were used to minimize barriers to study participation and increase the integrity of the data collected. Study visits were completed using 2 strategies designed to overcome common barriers to participation (eg, lack of transportation). Initially, research staff conducted all study visits in emerging adults' homes. Emerging adults used a study tablet to complete a brief demographic questionnaire via Qualtrics XM [42] and to review one or more of the intervention components. Research staff then engaged emerging adults in a semistructured qualitative interview. In March 2020, all study visits were transitioned from the home setting to online video conference calls due to the onset of the COVID-19 pandemic. The same procedures were followed, with minor modifications for the virtual context. Specifically, the consent or assent forms were emailed to study participants in advance of the study visit, and the study staff displayed them using the screen-sharing feature. Emerging adults completed the demographic questionnaire via a link sent via the videoconference software chat feature or verbally responded to the questions displayed via the screen-sharing feature. Similarly, research staff sent links to the online intervention components for emerging adults to review the intervention content, or research staff walked emerging adults through the intervention component using the screen-sharing feature.

Interventions

All 3 interventions were developed and delivered using the Computerized Intervention Authorizing Software, version 3.0 (CIAS 3.0; Michigan State University) [43], a no-code Software-as-a-Service platform allowing development of interactive mobile web apps and SMS text messaging interventions without coding. The MES and QPL interventions were delivered via a link to a CIAS 3.0 intervention session, which will be texted to participants in the clinical trial. Both interventions leveraged CIAS 3.0's interactive and emotive narrator. Consistent with human-computer interaction research [44], this narrator functions as an engaging guide throughout the intervention by reading text displayed on the screen and displaying emotional responses, such as surprise, sadness, or thoughtfulness, as appropriate. SMS text messaging was deployed by CIAS 3.0 as text messages. All 3 interventions were reviewed by a pediatric endocrinologist for clinical relevance prior to feasibility testing.

XSL•FO
RenderX

The MES is a brief (<15 min), 2-session, mHealth, counseling intervention grounded in motivational interviewing [45-48] and the Information-Motivation-Behavioral Skills model of health behavior change [49,50]. The Information-Motivation-Behavioral Skills model posits that behavior change results from the joint function of 3 critical components: accurate information about risk behaviors (eg, risks of suboptimal diabetes self-management) or their replacement behaviors (eg, benefits of effective diabetes self-management), motivation to change behavior, and behavioral skills necessary to perform the behavior (eg, self-efficacy). Motivational interviewing is an evidence-based strategy to optimize behavior change through the use of client-centered and goal-oriented directive communication to enhance intrinsic motivation and self-efficacy. MES content was developed to align with the needs and experiences of emerging adults. Specifically, we chose language emphasizing emerging adults' diabetes self-management autonomy and curated content to be developmentally consistent with emerging adults' experiences, including the reasons to engage in self-management activities, potential past successes, and personal strengths or weaknesses. Session 1 began with psychoeducation about the target behavior (ie, 3 key components of diabetes management referred to as "The 3Ms": glucose monitoring, insulin administration [medicine], and dietary management [meals]) delivered via video clips from an endocrinologist and supported with a peer testimonial (emerging adults chose from 2 peer testimonials, one depicting a Black male and the other a Black female). Participants then advanced through motivation-enhancing exercises designed to increase feelings of autonomy and self-efficacy for changing the targeted illness behavior (ie, diabetes management). Session 1 ended with goal setting with three options: (1) completing the targeted illness behavior as recommended (ie, doing all diabetes care every day), (2) increasing the targeted illness behavior (ie, do more diabetes care every day), and (3) thinking more about changing one's behavior, an autonomy-supportive option for those not yet ready to change their behavior. We then offered behavioral strategies to support goal attainment: setting reminders, establishing cues, and enlisting social support. Session 2 began by eliciting progress toward the diabetes care goal established in session 1 and then continued to grow motivation using motivation-enhancing exercises. In session 2, participants were asked which strategies they used and whether the strategies were helpful for completing diabetes care. Participants were then given the option to revisit the behavioral strategies information and reselect strategies to use.

The second intervention component was one-way tailored SMS text messages to remind emerging adults to complete their diabetes care. SMS text message reminders were grounded in social cognitive theory, which states that the individual, the behavior, and the environment reciprocally interact and influence one another [51]. Applied to diabetes self-management, SMS text message reminders promote self-management by providing an external prompt (environment) to the individual to complete the task. This process increases the likelihood of task completion, which leads to perceptions of control over the health behavior (autonomy), fosters feelings of competence (self-efficacy), and supports goal attainment [52-57]. SMS text message interventions may also generate feelings of social support [58], even when patients know the SMS text messages are automated [55]. SMS text messaging was adapted from a similar intervention to increase medication adherence for emerging adults with uncontrolled moderate-to-severe persistent asthma [59]. The asthma intervention consisted of 30 days of once-daily medication reminders via one-way SMS text message. In this study, we updated the content of the messages to reflect 'The 3Ms" (ie, insulin administration, blood glucose testing, and carbohydrate counting). Texting was tailored to emerging adults' preferences in two ways: emerging adults chose to receive reminders for one specific diabetes care component or a general reminder to complete "all their diabetes care," and selected the time of day reminders were delivered.

The third intervention was a QPL, a communication tool to empower patients to actively participate during medical visits by asking questions and stating concerns and preferences [60,61]. QPLs are comprised of lists of questions related to the physical and psychosocial aspects of illness and illness treatment that patients may want to ask their physicians or other clinicians during a clinic visit. The theoretical foundation for the QPL resides in social cognitive theory, which posits that behavioral performance is largely a function of confidence in one's ability to perform the behavior (self-efficacy) and the expectation that the behavior will result in the desired outcome [62]. Patients prepared with a QPL are more likely to ask questions and state their concerns, enabling shared decision-making and bolstering self-efficacy. The QPL was based on a similar tool designed to increase patients' participation during oncology treatment interactions [63,64]. The QPL was developed by first compiling diabetes-related questions gleaned from a literature search, web-based resources, and consultation with 3 diabetes medical care providers. Questions addressed common concerns in the 3Ms' domains and emerging adult-specific topics (eg, impact of tobacco or alcohol use on diabetes health). Intervention begins with a brief education component explaining the purpose of the QPL and the importance of communicating questions and concerns to their health care team. Emerging adults are then presented with lists of questions organized by content area from which they select the questions they would like to raise at the clinic visit. Emerging adults are routed to question sets tailored to their diabetes treatment regimen, for example, those monitoring their blood glucose with a glucometer are presented a set of questions specific to that method of blood glucose monitoring, and those using a continuous glucose monitor get a set of questions tailored to that approach. The questions selected by the emerging adults are compiled into a personalized report and emailed to the emerging adults upon completion and again the day prior to the clinic visit.

Data Collection

The *Family Information Form* (*FIF*) is an investigator-developed measure to gather participant characteristics. The FIF collects participant demographic information such as age, race and ethnicity, sex assigned at birth, income, and family structure. The FIF also asks emerging adults for information about their illness, including the emerging adults' date of diagnosis and current diabetes self-care regimen. Research staff used a

XSL•FO
RenderX

semistructured interview guide developed for the study to guide the interviews. The interviews began with an explanation of the purpose of each intervention, after which the research staff guided emerging adults through a demonstration of the interventions. After reviewing the intervention, the research staff elicited emerging adults' feedback on the intervention, including which elements emerging adults found most helpful and least helpful, suggestions for improving the intervention, and recommendations to make the interventions more consistent with the language and experiences of emerging adults (Multimedia Appendix 1). After 10 interviews, the feedback offered had reached the point of saturation and was deemed sufficient. Interviews were recorded and professionally transcribed for analysis.

Data Analysis

A coding team composed of 2 coders (1 undergraduate psychology and 1 graduate public health student) and a qualitative expert (the principal investigator) analyzed the transcribed interview data using framework matrix analysis [65], an efficient approach to inductive thematic analysis. After an initial training period in which coders became familiar with the study design and aims, data collection procedures, and analytic approach, the coding team developed the initial coding matrix based on the content of the interview guide. Coders used this matrix to independently review the interview transcripts, "charting" summaries of emerging adults' responses to interview questions into their coding matrix. Coders then met to compare their independently coded matrices and develop a consensus-coded matrix. Discrepancies were resolved through discussion and review of the interview transcripts. Once all interview transcripts were coded, the coders met to identify and summarize the themes (denoted with italicized text in the results) that emerged across emerging adults' feedback. Coders developed descriptions of each theme with illustrative quotes extracted from the interview transcripts.

Ethical Considerations

This study was approved by the Wayne State University institutional review board. Emerging adults provided informed consent or informed assent with parental informed consent to participate prior to engaging in study activities. Prior to analysis, all study data were deidentified to protect emerging adults' privacy and confidentiality. Emerging adults were compensated with a US $50 gift card for their participation.

Results

Overview

Table 1 summarizes the characteristics of the emerging adults participating in this study. On average, emerging adults were 18.8 (SD 1.7) years, 50% (10/20) were female, and 80% (16/20) identified as Black, non-Hispanic. Emerging adults had completed an average of 12.0 (SD 1.8) years of education, reported their socioeconomic status as 5.9 (SD 1.6) out of 10 on the McArthur Scale of Subjective Socioeconomic Status [66], and most (13/20, 65%) were not employed. Emerging adults had been diagnosed with T1D for 11.1 (SD 4.8) years, 65% (13/20) used a glucometer to monitor their blood glucose levels, and 75% (15/20) administered their insulin via injection.

XSL•FO
RenderX

Table . Demographic and disease characteristics of study participants (N=20).

Characteristic	Value
Age (y), mean (SD)	18.8 (1.7)
Sex assigned at birth (female), n (%)	10 (50)
Race and ethnicity, n (%)	
Biracial, Hispanic	2 (10)
Black, non-Hispanic	16 (80)
White, Hispanic	2 (10)
Education (y), mean (SD)	12.0 (1.8)
Subjective socioeconomic status[a], mean (SD)	5.9 (1.6)
Employment status, n (%)	
Employed part-time	5 (25)
Employed full-time	2 (10)
Not employed	13 (65)
Duration of diabetes (y), mean (SD)	11.1 (4.8)
Blood glucose monitoring method, n (%)	
Glucometer	13 (65)
Continuous glucose monitor	7 (35)
Insulin delivery method, n (%)	
Injected insulin	15 (75)
Insulin infusion pump	5 (25)

[a]The MacArthur Scale of Subjective Social Status uses a 10-point scale where 1=lowest social status and 10=highest social status

Emerging adults found all 3 interventions to be helpful and useful. They highlighted the interventions' *integration into the technology* they use daily as a strength: "it has a high potential to be effective, especially with the amount of technology that's used nowadays" and "I'm always on my phone as a young adult."

Motivation Enhancement System

Overall, emerging adults found the MES intervention *helpful* and the delivery structure *acceptable*, with 9 of 10 emerging adults reporting the length and number of sessions "About Right." Emerging adults found the 3Ms mnemonic (3Ms=monitoring, medicine, and meals) a helpful prompt for remembering the essential diabetes care tasks:

> *it would be something very helpful to remember because when you're in the heat of the moment or you're just not focusing, that's something you should remember.*

When asked about the likelihood that emerging adults would recommend the MES program to other emerging adults, 8 of 10 emerging adults stated they would "certainly" recommend the program (mean 9.5/10, SD 1.1). When asked to elaborate, emerging adults highlighted the video elements of the intervention.

Emerging adults found the physician depicted in the educational video a *credible* and *trustworthy* authority figure:

> *She got her message across, and it just makes sense. I understand why I need to do the things I need to do. The way she spoke about the things was very credible. I feel like I can trust her word.*

Emerging adults found the physician video content acknowledging the stress and challenge of diabetes care consistent with their own experiences:

> *I had diabetes since I was four. So, when she [the physician] was like, "It begins to get hard," and stuff like that, I understood that because it's hard for me, especially now that I'm driving and working and doing other things.*

Emerging adults reported that the peer testimonial videos were *relatable* in two ways. First, the experiences depicted in the testimonial resonated with their own challenges with diabetes care:

> *I can relate to [what] she's saying too because I am tired too. It was way easier when I was a kid. Your parents may have to do all that, but now it's my turn and it's hard.*

Demographic similarities in the emerging adults depicted in the peer testimonial video enhanced participants' feelings of relatability:

> *She's in college. She's my age or around my age. Everything she was saying was the truth. Yeah, it was nice to see somebody who was my age similar to me.*

XSL•FO
RenderX

> *She's same complexion. Yeah, it's nice that there's a black person talking to me.*

Emerging adults found other, nonvideo elements of the intervention credible and relatable, including lists of common barriers to diabetes care: "I've done almost everything on here."

Emerging adults' comments on the MES intervention highlighted its *autonomy-supportive* nature. Emerging adults reported less reliance on family support and input, and a desire for greater independence and control when making diabetes care decisions: "I don't know if people would do something like that [enlist the support of family] in college." Asking for help with diabetes care was viewed as undermining their autonomy: "that is one of my goals because I know how to ask for help, but I prefer not to." Emerging adults explained that independently caring for their diabetes helped build their *self-efficacy* for diabetes care by giving them the sense that "I believe in myself" and that they were, therefore, in control of their health.

Emerging adults viewed the behavioral strategies to support diabetes care as helpful tools to help them be autonomous. To illustrate, one emerging adult found the suggestion to use reminders to prompt diabetes care useful:

> *That's smart, with the sticky note on the fridge or wherever they get their food from or try to get their food from. Just seeing that, it's like, "all right, let me check my blood before I head out," or "let me check my blood before I make this or eat this." And for the putting an alarm on the phone, that's very helpful, because when that time comes up and your alarm is screaming at you to do something, I feel like that's very helpful as well.*

Another emerging adult found the suggestion to use cues to prompt diabetes care to be an effective strategy:

> *It would be very much useful, especially for those who are always running around and forget to do things. Having those cues to put a thought in their head like, "wait, I should do this. I should do that before I do this or that."*

A third emerging adult reported using a similar strategy to prompt blood glucose monitoring:

> *I'll put my [blood glucose monitoring] kit in the basket with my toothbrush and my soap and stuff. So, before I do any of that stuff, I got to pull this out and check my blood sugar. Or, if I have an early morning class, at night I'll put this blood sugar kit in my jacket pocket. So, if I don't check it when I wake up, by the time I get to the class or whatever, I'll just pull it out and I'll check it.*

Text Message Reminders

Overall, emerging adults reported SMS text messaging would be a *helpful reminder* to stay "on track" with diabetes care tasks and increase the consistency of diabetes care:

> *I'm trying to get on a schedule, but sometimes other things interfere with that. So, having those reminders will help.*

They thought receiving reminders at critical times would be particularly helpful, such as early morning ("when I wake up that's the first notification I see on my phone"), afternoon ("the time where everybody would be up, so they will be alert when they get the message"), and late evening ("Sometimes you almost forget and it's about to get late [around] bedtime"). Emerging adults *preferred SMS text messages* over email reminders because "a lot [of] people look at their phone more than they look at emails," apps that "take up gigs [gigabytes/storage] on [my] phone," and other technology-based reminders, which they tend to ignore:

> *I already set up reminders in my phone. I don't really pay attention to it though.*

Emerging adults' *preferences for the content of SMS text message reminders* included "straightforward" wording of the SMS text messages and suggested making the language even more directive: "Take your insulin" versus "Don't forget to take your insulin." Emerging adults also preferred a variety of messages, suggesting we "use different ones, not the same wording every single day," and a message personalized with their name. Using emojis in the messages was recommended to make them more engaging and consistent with other texts they receive: "Did you give insulin? with a smiley face." Emerging adults emphasized the importance of being able to *tailor the reminders* to their own preferences. Specifically, the ability to select delivery times that aligned with their daily routines and mealtimes was a strength: "They're very good choices because these are around [the] time you eat." Most emerging adults thought a single reminder would be insufficient and recommended *multiple reminders per day*. While there was not strong agreement on the ideal number of texts, most emerging adults suggested 2 to 4 texts per day and advised that more than 7 texts would be viewed as "spamming."

Finally, emerging adults recommended increasing the *range of topics* from which they might select reminder messages. Emerging adults suggested including reminders to exercise: "It'd probably motivate them or remind them that they maybe need to go out, take a jog, walk or lift some weights or just do some type of exercise to get their blood pumping." Emerging adults recommended messages prompting blood glucose monitoring during exercise:

> *I want to make sure that I'm not running low doing them [exercises], so having a reminder to make sure you're on your sugar [checking glucose levels] while you're doing exercise to make sure that it doesn't run low or high.*

Emerging adults also suggested reminders to check ketones: "Make sure you check your ketones, if their reading's above a certain level."

Question Prompt List

Overall, emerging adults reported that the QPL was a *useful tool to prepare for diabetes clinic visits*:

XSL•FO
RenderX

> *It was incredibly useful. I think it would help me out at my next [doctor's] appointment. It's definitely less time consuming than sitting there researching stuff and writing it out on your own.*

The variety of questions was viewed as a strength:

> *The questions were very diverse, very broad. There were some questions I had when I first got diabetes a long time ago versus some that I never thought to ask [my doctors] to this day.*

Emerging adults remarked that the QPL would be particularly useful for emerging adults newly diagnosed with diabetes:

> *If I was older and I had gotten diagnosed, these would have been the kinds of questions I would ask.*

Emerging adults found the different *content areas of the QPL valuable*. They described the glucose monitoring questions as "beneficial," "thoughtful," and "diverse." Emerging adults emphasized the importance of the insulin questions, noting the questions could help "prevent a situation" by elevating important discussion topics: "These ones are really important and specifically address things I may have overlooked." Emerging adults who were administering their insulin via injection (8/10, 80%) thought the questions about transitioning to an insulin infusion pump were helpful: "I always wanted to know what is it or what does it do." Those already using an insulin infusion pump (2/10, 20%) described these questions as very "relevant": "All of these questions are questions that I had when I was transitioning to the pump." One topic of particular interest to emerging adults was the section on transitioning to adult life with diabetes:

> *I really like the adult life part because I think that [is] something that's important that you consider before you get there rather than learn from experience.*

Emerging adults found these questions novel and thought-provoking, remarking:

> *Some of the questions might change my mindset. Questions I didn't know I had.*

They highlighted the importance of these questions for stimulating conversation on relevant but difficult-to-raise topics, including tattoos, pregnancy, and piercings: "I really think that these questions address some misconceptions that people may hear and feel nervous about." Emerging adults' suggestions to expand the content of the QPL included adding questions about diabetes care challenges ("What are some of the struggles you are having with your diabetes?") and to expand upon the diabetes care topics ("Why do I have a random high blood sugar if I haven't eaten anything?", "What type of insulin works better for me?") and questions about "free carbs," "insulin resistance," and "how to navigate CGM sensor errors."

When it came to esthetics and logistics, emerging adults expressed a desire to *tailor the intervention to their preferences*. Emerging adults stated they would prefer to rename the QPL but disagreed on what the preferred name should be. Preferences ranged from a highly personalized name integrating the patient's own name into the title ("[Name]'s Diabetes Questions List") to a "less diabetes specific" name ("Keeping My Life on Track"). Emerging adults also wanted to tailor how they received the finalized question list. Most emerging adults thought sending their diabetes question list directly to their doctor's office would make it easier to communicate their concerns: "some people [are] not comfortable with their doctors yet or don't like speaking out." Some preferred receiving a copy of their question list by email, explaining they preferred email over SMS text message for a "bigger document" and "things that are important." Some noted email delivery would allow them to "read it off of my phone," noting "I usually always have my phone on me" and the accessibility of emails across devices: "You will always have it on file because it will be on email." Emerging adults who preferred SMS text message delivery did so due to a personal preference ("I check my text messages more than I check my email") and ease of access ("I have thousands of emails inside there, so I think text would be way better because I can just search it up"). One emerging adult suggested an option to mail the QPL for those without email access. Other tailoring preferences included font size, optional narration, delivery timing, and sharing their question lists with others.

Discussion

Principal Findings

The goal of this formative research was to enhance the potential relevance, sustainability, and efficacy of 3 SDT-informed mHealth intervention components to promote diabetes management and glycemic control and related intervention content for emerging adults with T1D by soliciting feedback from emerging adults with T1D and their recommendations for improving the interventions. Key areas of interest included emerging adults' perspectives on the use of cultural tailoring, the developmental relevance of content, and preferences for delivery logistics. Overall, emerging adults found the interventions acceptable and thought they would be a helpful support for diabetes care. The integration of the interventions into technology that emerging adults already frequently use (mobile phone) was highlighted as a strength across interventions. Another common thread in emerging adults' feedback was a preference for intervention content tailored to their illness experience, demographic characteristics, and personal preferences. Feedback related to the MES intervention highlighted elements that were engaging (videos), relatable (peer testimonials), and supportive of their growing autonomy and independence. Emerging adults viewed the text messaging intervention as covering a range of topics and had specific suggestions for making the messages more impactful (be direct, personalized, and engaging, eg, emojis). Emerging adults found the QPL content areas relevant and well-aligned with the concerns of emerging adults living with diabetes, particularly the topic of transitioning to adult life with diabetes.

It was unsurprising that emerging adults highlighted the interventions' integration into mobile phone technology as a strength. Prior research has documented that adolescents and young adults prefer interventions delivered to their mobile phones [67], are more likely to use interventions when delivered

XSL•FO
RenderX

to their personal device [68], and when delivered via text message [30,69]. However, to reap the benefits of eHealth interventions, engagement is critical [32,68]. Thus, finding ways to engage emerging adults in an intervention is a critical first step, and based on emerging adults' feedback reported here, SMS text messaging–based interventions using emerging adults' own devices may be an optimal strategy. The Unified Theory of Acceptance and Use of Technology [70] lends further support to this conclusion, as it suggests interventions integrated into the technology people use are more likely to be accepted and used because of the ease of use (effort expectancy). Other Unified Theory of Acceptance and Use of Technology factors that may be particularly relevant for emerging adults and SDT-based interventions are the perception that other people support using the intervention (social influence, given that peer influence is a critical factor for emerging adults [71]) and the automaticity of using the intervention (habit).

Emerging adults' preference for intervention content tailored to their illness experience, demographic characteristics, and personal preferences is aligned with the broader literature on culturally tailoring interventions. Broadly speaking, cultural adaptation refers to the systematic modification of an intervention to align with the target audience's cultural norms, beliefs, and values [72]. Culturally tailored interventions have demonstrated greater efficacy among both adult [73] and youth interventions [74]. Cultural tailoring for interventions targeting youth frequently includes adapting intervention methods (duration), language (translation), and content (relatable experiences) [74]. In this study, emerging adults highlighted several aspects of intervention tailoring. Emerging adults noted that the QPL questions aligned with questions they had previously asked during their own health care encounters or felt they would ask in future encounters. In the MES videos, emerging adults found the representation of peers with demographic characteristics (eg, age and race) similar to their own highly relatable. Others have noted the importance of representation of social identity as a key tailoring element for youth interventions [74]. In SMS text messaging, emerging adults viewed the ability to tailor the timing and content of reminders to personal preferences as a strength. In light of this feedback, we increased the planned once-daily reminder to twice daily based on emerging adults' feedback reported here. Most emerging adults' recommendations for intervention improvement related to augmenting the capacity to tailor the interventions to personal preferences, including increasing the frequency and duration of intervention, augmenting the intervention content, and personalizing reminders with the recipient's name. Tailoring to individual preferences is consistent with emerging adults' developmental need for autonomy and has been noted as an important element for engagement in technology-based interventions [75,76].

The interventions developed in this study were all grounded in SDT [35]. SDT was selected as the guiding framework because of the resonance-enhancing feelings of autonomy, self-efficacy, and relatedness it has for emerging adults who are at a developmental stage in which forging independence, a unique identity, and social connections are paramount [8]. Emerging adults participating in this study noted that each intervention contained elements that resonated with this framework. Specifically, emerging adults found the credibility and trustworthiness of the videos embedded in the MES intervention promoted autonomous diabetes care. The videos' relatability, particularly the peer testimonial, helped foster feelings of relatedness. Emerging adults found other MES elements, such as the behavioral strategies to support diabetes care goals, also fostered feelings of autonomy. Emerging adults embraced text messaging as a strategy to promote consistent diabetes care, an approach previously linked to improvements in diabetes self-efficacy [55]. Emerging adults noted that multiple QPL questions were questions they had in the past or would consider asking in the future, fostering both feelings of confidence (self-efficacy) and relatedness. Questions addressing adult issues (eg, alcohol use) resonated with emerging adults' goal of independence (autonomy). While theory-based interventions are more likely to improve glycemic control and psychological outcomes [77], SDT-guided interventions, to date, have not demonstrated improvements in glycemic control [38,40]. GSD, the most widely tested SDT intervention, relies upon the completion of worksheets that form the basis of conversations with diabetes educators [31,37,39]. Not only has this approach failed to demonstrate improvements in glycemic control, but poor recruitment and retention rates also suggest this approach is unacceptable to emerging adults [31,38,39].

Limitations

This study's small sample size limits the generalizability of this study to other emerging adults and populations with T1D. This limitation is offset by the fact that the emerging adults included were mostly non-Hispanic Black and were selected based on suboptimal glycemic control from an urban setting. Such populations are underrepresented in research literature and frequently absent from intervention development research. In addition, the findings reported here parallel those from a similar study conducted with a mainly White (89%) rural population of emerging adults who reviewed the same interventions and also highlighted the excellent fit of technology-delivered interventions for their age group, expressed a preference for individualized interventions, and stated the value of intervention elements that enhanced their feelings of autonomy [78]. While the inclusion of historically marginalized emerging adults is a strength, other cultural factors, eg, emerging adults' acceptance of their diabetes, may have influenced emerging adults' feedback and intervention preferences [79].

A second limitation was constraining emerging adults' engagement in the research to providing feedback and recommendations on our investigator-developed intervention components. A more inclusive approach would have been to engage emerging adults in the development of the interventions using a community-based participatory approach. However, as described in Multimedia Appendix 1, the interventions adapted in this study were derived from similar interventions developed and empirically tested in similar populations. Furthermore, each intervention component is theoretically grounded in SDT and related behavior change frameworks. Finally, emerging adults' feedback will be used to refine the interventions prior to a planned randomized controlled trial using the multiphase optimization strategy [80] to test the intervention components'

XSL•FO
RenderX

preliminary efficacy to improve glycemic control and build the most efficacious multicomponent intervention [41]. Specifically, MES will be refined based on emerging adults' language preferences. The planned frequency of SMS text messaging will be increased from 1 to 2 messages per day, with the number of messages increased to minimize repetition and redundancy. SMS text messaging will also be further tailored to include the emerging adult's name in messages. QPL questions will be revised according to emerging adults' preferences and new questions added to augment the scope of question content. Emerging adults' feedback will also help other researchers develop interventions aligned with the needs and preferences of emerging adults, which will increase their acceptability and future effectiveness.

Finally, the development and use of digital interventions assumes the end users, in this instance, emerging adults with T1D, have the necessary digital literacy to effectively engage with the intervention and that they have access to the tools (digital devices and internet access) necessary to receive the intervention. In the United States, internet use and smartphone ownership are at an all-time high. The most recent statistics from the Pew Research Center indicate 95% of US adults use the internet, 90% have a smartphone, and 80% subscribe to high-speed internet at home [81]. Smartphone ownership is even greater for young adults aged 18 to 29 years (97%), and 78% subscribe to high-speed internet at home [81]. Today's adolescents and young adults are growing up immersed in technology (cell phones and computers), making technology-based interventions a natural extension of their social ecology [82-84]. Thus, mHealth interventions exclude only a minority of youth and emerging adults who are "digital natives" [85] having grown up with technology as a component of their everyday life.

Conclusions

The results reported here provide clinicians and researchers with specific guidance for effective behavioral interventions with emerging adults with T1D. First, clinicians and researchers working with emerging adults should offer eHealth interventions that leverage the technology emerging adults already use, namely, their mobile phones. They should also provide opportunities for emerging adults to tailor interventions to their personal preferences, such as providing options for the frequency and duration of intervention, and the specific content of the intervention, for example, selecting a discrete behavioral target from a menu of options. Clinicians and researchers should also ensure interventions match the cultural context of emerging adults, such as ensuring case examples, vignettes, and audio-visual content represent people of a similar background and with similar experiences, to increase engagement, a critical factor for sustained use and to reap the benefits of an intervention. Finally, interventions grounded in SDT align well with emerging adults' innate psychological needs for autonomy, self-efficacy, and relatedness. Thus, interventions that activate these psychological processes may be particularly effective for evoking behavior change among emerging adults.

Acknowledgments

The authors would like to thank Margie M Maly and Colette Serpetti for their assistance with the development of the initial question prompt list. Their reflections and insights were invaluable. The authors would also like to extend a very special thanks to the study participants who generously donated their time and energy to reviewing the interventions and providing their feedback to improve their quality and utility. Generative artificial intelligence technology was not used in the preparation of any part of this manuscript.

Funding

Research reported in this publication was supported by the National Institute of Diabetes and Digestive and Kidney Diseases of the National Institutes of Health (NIH) under the award R01DK116901. NIH had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author to researchers who provide a methodologically sound proposal for use under a specific data-sharing agreement that provides for (1) a commitment to using the data only for research purposes and not to identify any individual participant, (2) a commitment to securing the data using appropriate technology, and (3) a commitment to destroying the data after analyses are completed.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Semistructured interview questions and detailed descriptions of the interventions.
[DOCX File, 295 KB - formative_v9i1e75623_app1.docx]

XSL•FO
RenderX

References

1. Gonzalez JS, Tanenbaum ML, Commissariat PV. Psychosocial factors in medication adherence and diabetes self-management: Implications for research and practice. Am Psychol 2016 Oct;71(7):539-551. [doi: 10.1037/a0040388] [Medline: 27690483]
2. Johnson SB, Kelly M, Henretta JC, Cunningham WR, Tomer A, Silverstein JH. A longitudinal analysis of adherence and health status in childhood diabetes. J Pediatr Psychol 1992 Oct;17(5):537-553. [doi: 10.1093/jpepsy/17.5.537] [Medline: 1432480]
3. Thomas AM, Peterson L, Goldstein D. Problem solving and diabetes regimen adherence by children and adolescents with IDDM in social pressure situations: a reflection of normal development. J Pediatr Psychol 1997 Aug;22(4):541-561. [doi: 10.1093/jpepsy/22.4.541] [Medline: 9302851]
4. Perwien AR, Johnson SB, Dymtrow D, Silverstein J. Blood glucose monitoring skills in children with type I diabetes. Clin Pediatr (Phila) 2000 Jun;39(6):351-357. [doi: 10.1177/000992280003900605] [Medline: 10879937]
5. Keough L, Sullivan-Bolyai S, Crawford S, Schilling L, Dixon J. Self-management of type 1 diabetes across adolescence. Diabetes Educ 2011;37(4):486-500. [doi: 10.1177/0145721711406140] [Medline: 21602489]
6. Weissberg-Benchell J, Wolpert H, Anderson BJ. Transitioning from pediatric to adult care: a new approach to the post-adolescent young person with type 1 diabetes. Diabetes Care 2007 Oct;30(10):2441-2446. [doi: 10.2337/dc07-1249] [Medline: 17666466]
7. Berg CA, Wiebe DJ, Suchy Y, et al. Executive function predicting longitudinal change in type 1 diabetes management during the transition to emerging adulthood. Diabetes Care 2018 Nov;41(11):2281-2288. [doi: 10.2337/dc18-0351] [Medline: 30131398]
8. Arnett JJ. Emerging adulthood. A theory of development from the late teens through the twenties. Am Psychol 2000 May;55(5):469-480. [doi: 10.1037//0003-066X.55.5.469] [Medline: 10842426]
9. Bryden KS, Dunger DB, Mayou RA, Peveler RC, Neil HAW. Poor prognosis of young adults with type 1 diabetes: a longitudinal study. Diabetes Care 2003 Apr;26(4):1052-1057. [doi: 10.2337/diacare.26.4.1052] [Medline: 12663572]
10. DeSalvo DJ, Miller KM, Hermann JM, et al. Continuous glucose monitoring and glycemic control among youth with type 1 diabetes: international comparison from the T1D Exchange and DPV Initiative. Pediatr Diabetes 2018 Nov;19(7):1271-1275. [doi: 10.1111/pedi.12711] [Medline: 29923262]
11. McCarthy MM, Grey M. Type 1 diabetes self-management from emerging adulthood through older adulthood. Diabetes Care 2018 Aug;41(8):1608-1614. [doi: 10.2337/dc17-2597] [Medline: 29802144]
12. American Diabetes Association. 6. Glycemic targets: Standards of Medical Care in Diabetes—2021. Diabetes Care 2021 Jan 1;44(Supplement_1):S73-S84. [doi: 10.2337/dc21-S006]
13. Dabelea D, Stafford JM, Mayer-Davis EJ, et al. Association of type 1 diabetes vs type 2 diabetes diagnosed during childhood and adolescence with complications during teenage years and young adulthood. JAMA 2017 Feb 28;317(8):825-835. [doi: 10.1001/jama.2017.0686] [Medline: 28245334]
14. Wiebe DJ, Helgeson V, Berg CA. The social context of managing diabetes across the life span. Am Psychol 2016 Oct;71(7):526-538. [doi: 10.1037/a0040355] [Medline: 27690482]
15. Fisher L, Polonsky WH, Hessler DM, et al. Understanding the sources of diabetes distress in adults with type 1 diabetes. J Diabetes Complicat 2015 May;29(4):572-577. [doi: 10.1016/j.jdiacomp.2015.01.012]
16. Callahan ST, Cooper WO. Changes in ambulatory health care use during the transition to young adulthood. J Adolesc Health 2010 May;46(5):407-413. [doi: 10.1016/j.jadohealth.2009.09.010] [Medline: 20413075]
17. Ding K, Reynolds CM, Driscoll KA, Janicke DM. The relationship between executive functioning, type 1 diabetes self-management behaviors, and glycemic control in adolescents and young adults. Curr Diab Rep 2021 Feb 22;21(3):10. [doi: 10.1007/s11892-021-01379-3] [Medline: 33616838]
18. Pihlaskari AK, Anderson BJ, Eshtehardi SS, et al. Diabetes disclosure strategies in adolescents and young adult with type 1 diabetes. Patient Educ Couns 2020 Jan;103(1):208-213. [doi: 10.1016/j.pec.2019.08.019] [Medline: 31447195]
19. Huhn F, Lange K, Jördening M, Ernst G. Real-world use of continuous glucose monitoring systems among adolescents and young adults with type 1 diabetes: reduced burden, but little interest in data analysis. J Diabetes Sci Technol 2023 Jul;17(4):943-950. [doi: 10.1177/19322968221081216] [Medline: 35255729]
20. Laffel LM, Kanapka LG, Beck RW, et al. Effect of continuous glucose monitoring on glycemic control in adolescents and young adults with type 1 diabetes: a randomized clinical trial. JAMA 2020 Jun 16;323(23):2388-2396. [doi: 10.1001/jama.2020.6940] [Medline: 32543683]
21. Thabit H, Prabhu JN, Mubita W, et al. Use of factory-calibrated real-time continuous glucose monitoring improves time in target and HbA1c in a multiethnic cohort of adolescents and young adults with type 1 diabetes: the MILLENNIALS study. Diabetes Care 2020 Oct;43(10):2537-2543. [doi: 10.2337/dc20-0736] [Medline: 32723843]
22. Bergenstal RM, Nimri R, Beck RW, et al. A comparison of two hybrid closed-loop systems in adolescents and young adults with type 1 diabetes (FLAIR): a multicentre, randomised, crossover trial. The Lancet 2021 Jan;397(10270):208-219. [doi: 10.1016/S0140-6736(20)32514-9]
23. Isganaitis E, Raghinaru D, Ambler-Osborn L, et al. Closed-loop insulin therapy improves glycemic control in adolescents and young adults: outcomes from the International Diabetes Closed-Loop Trial. Diabetes Technol Ther 2021 May;23(5):342-349. [doi: 10.1089/dia.2020.0572] [Medline: 33216667]

XSL•FO
RenderX

24. Caccavale LJ, LaRose JG, Mazzeo SE, Bean MK. Development and implementation of a pilot transition preparation intervention for young adults with type 1 diabetes in an integrated healthcare setting. J Pediatr Psychol 2023 Mar 20;48(3):228-240. [doi: 10.1093/jpepsy/jsac084] [Medline: 36367835]
25. Goyal A, Peerzada A, Sarteau AC, Praveen PA, Kalaivani M, Tandon N. A multi-center pediatric to adult care transition intervention program to improve clinic visit adherence and clinical outcomes among adolescents and emerging adults with type 1 diabetes mellitus [PATHWAY]: protocol for a randomized controlled trial. Contemp Clin Trials 2022 Aug;119:106830. [doi: 10.1016/j.cct.2022.106830] [Medline: 35724840]
26. Van Walleghem N, Macdonald CA, Dean HJ. Evaluation of a systems navigator model for transition from pediatric to adult care for young adults with type 1 diabetes. Diabetes Care 2008 Aug;31(8):1529-1530. [doi: 10.2337/dc07-2247] [Medline: 18458141]
27. Pyatak EA, Carandang K, Vigen CLP, et al. Occupational therapy intervention improves glycemic control and quality of life among young adults with diabetes: the Resilient, Empowered, Active Living with Diabetes (REAL Diabetes) randomized controlled trial. Diabetes Care 2018 Apr;41(4):696-704. [doi: 10.2337/dc17-1634] [Medline: 29351961]
28. Pyatak EA, Lee PJ, Nnoli ND, et al. Telehealth occupational therapy improves psychosocial well-being but not glycemia among young adults with type 1 diabetes: The Resilient, Empowered, Active Living-Telehealth (REAL-T) randomized controlled trial. Diabetes Res Clin Pract 2025 Mar;221:112005. [doi: 10.1016/j.diabres.2025.112005] [Medline: 39884514]
29. Bisno DI, Reid MW, Pyatak EA, et al. Virtual peer groups reduce HbA1c and increase continuous glucose monitor use in adolescents and young adults with type 1 diabetes. Diabetes Technol Ther 2023 Sep;25(9):589-601. [doi: 10.1089/dia.2023.0199] [Medline: 37335751]
30. Mayberry LS, Wiebe DJ, Parks M, Campbell MS, Beam AB, Berg CA. Acceptability and feasibility of FAMS-T1D mHealth intervention to optimize self- and social regulation for emerging adults with type 1 diabetes. Pilot Feasibility Stud 2024 Apr 30;10(1):68. [doi: 10.1186/s40814-024-01497-1] [Medline: 38689356]
31. Rasmussen B, Wynter K, Hamblin PS, et al. Feasibility and acceptability of an online guided self-determination program to improve diabetes self-management in young adults. Digit Health 2023;9:20552076231167008. [doi: 10.1177/20552076231167008] [Medline: 37021125]
32. Middleton T, Constantino M, McGill M, et al. An enhanced SMS text message–based support and reminder program for young adults with type 2 diabetes (TEXT2U): randomized controlled trial. J Med Internet Res 2021 Oct 21;23(10):e27263. [doi: 10.2196/27263] [Medline: 34524102]
33. Zhang S, Hamburger E, Kahanda S, Lyttle M, Williams R, Jaser SS. Engagement with a text-messaging intervention improves adherence in adolescents with type 1 diabetes: brief report. Diabetes Technol Ther 2018 May;20(5):386-389. [doi: 10.1089/dia.2018.0015] [Medline: 29792749]
34. O'Hara MC, Hynes L, O'Donnell M, et al. A systematic review of interventions to improve outcomes for young adults with Type 1 diabetes. Diabet Med 2017 Jun;34(6):753-769. [doi: 10.1111/dme.13276] [Medline: 27761951]
35. Deci EL, Ryan RM. Intrinsic Motivation and Self-Determination in Human Behavior: Springer; 1985.
36. Goethals ER, Jaser SS, Verhaak C, et al. Communication matters: the role of autonomy-supportive communication by health care providers and parents in adolescents with type 1 diabetes. Diabetes Res Clin Pract 2020 May;163:108153. [doi: 10.1016/j.diabres.2020.108153] [Medline: 32325107]
37. Sarfo JO, Obeng P, Kyereh HK, Ansah EW, Attafuah PYA. Self-determination theory and quality of life of adults with diabetes: a scoping review. J Diabetes Res 2023;2023:5341656. [doi: 10.1155/2023/5341656] [Medline: 37091043]
38. Husted GR, Thorsteinsson B, Esbensen BA, et al. Effect of guided self-determination youth intervention integrated into outpatient visits versus treatment as usual on glycemic control and life skills: a randomized clinical trial in adolescents with type 1 diabetes. Trials 2014 Aug 12;15(1):321. [doi: 10.1186/1745-6215-15-321] [Medline: 25118146]
39. Zoffmann V, Vistisen D, Due-Christensen M. Flexible guided self-determination intervention for younger adults with poorly controlled type 1 diabetes, decreased HbA1c and psychosocial distress in women but not in men: a real-life RCT. Diabet Med 2015 Sep;32(9):1239-1246. [doi: 10.1111/dme.12698] [Medline: 25601214]
40. Hill E, Sibthorp J. Autonomy support at diabetes camp: a self determination theory approach to therapeutic recreation. Ther Recreation J 2006;40(2):107 [FREE Full text]
41. Idalski Carcone A, Ellis DA, Eggly S, et al. Improving diabetes management in emerging adulthood: an intervention development study using the multiphase optimization strategy. JMIR Res Protoc 2020 Oct 20;9(10):e20191. [doi: 10.2196/20191] [Medline: 33079068]
42. Qualtrics XM. 2025. URL: https://www.qualtrics.com/ [accessed 2025-07-04]
43. Michigan State University. Computerized Intervention Authoring System 30. 2021. URL: https://www.cias.app [accessed 2025-07-04]
44. Appel J, von der Pütten A, Krämer NC, Gratch J. Does humanity matter? Analyzing the importance of social cues and perceived agency of a computer system for the emergence of social reactions during human-computer interaction. Adv Hum Comput Interact 2012;2012:1-10. [doi: 10.1155/2012/324694]
45. Miller WR, Rollnick S. Motivational Interviewing: Preparing People to Change Addictive Behavior: The Guilford Press; 1991:348.

XSL•FO
RenderX

46. Miller WR, Rollnick S. Motivational Interviewing: Preparing People for Change, 2nd edition: Guilford; 2002. [doi: 10.1097/01445442-200305000-00013]
47. Miller WR, Rollnick S. Motivational Interviewing: Helping People Change, 3rd edition: The Guilford Press; 2012.
48. Miller WR, Rollnick S. Motivational Interviewing: Helping People Change and Grow: Guilford Publications; 2023.
49. Fisher JD, Fisher WA. Changing AIDS-risk behavior. Psychol Bull 1992 May;111(3):455-474. [doi: 10.1037/0033-2909.111.3.455] [Medline: 1594721]
50. Fisher WA, Fisher JD, Harman J. The information-motivation-behavioral skills model: a general social psychological approach to understanding and promoting health behavior. Jerry Suls KAW :82-106. [doi: 10.1002/9780470753552]
51. Bandura A. Social cognitive theory. In: Vasta R, editor. Six Theories of Child Development: Revised Formulations and Current Issues: JAI Press; 1989:1-60.
52. Herbert L, Owen V, Pascarella L, Streisand R. Text message interventions for children and adolescents with type 1 diabetes: a systematic review. Diabetes Technol Ther 2013 May;15(5):362-370. [doi: 10.1089/dia.2012.0291] [Medline: 23550554]
53. Arora S, Peters AL, Agy C, Menchine M. A mobile health intervention for inner city patients with poorly controlled diabetes: proof-of-concept of the TExT-MED program. Diabetes Technol Ther 2012 Jun;14(6):492-496. [doi: 10.1089/dia.2011.0252] [Medline: 22524591]
54. Dick JJ, Nundy S, Solomon MC, Bishop KN, Chin MH, Peek ME. Feasibility and usability of a text message-based program for diabetes self-management in an urban African-American population. J Diabetes Sci Technol 2011 Sep 1;5(5):1246-1254. [doi: 10.1177/193229681100500534] [Medline: 22027326]
55. Nundy S, Dick JJ, Solomon MC, Peek ME. Developing a behavioral model for mobile phone-based diabetes interventions. Patient Educ Couns 2013 Jan;90(1):125-132. [doi: 10.1016/j.pec.2012.09.008] [Medline: 23063349]
56. Nundy S, Mishra A, Hogan P, Lee SM, Solomon MC, Peek ME. How do mobile phone diabetes programs drive behavior change? Evidence from a mixed methods observational cohort study. Diabetes Educ 2014;40(6):806-819. [doi: 10.1177/0145721714551992] [Medline: 25278512]
57. Franklin VL, Waller A, Pagliari C, Greene SA. A randomized controlled trial of Sweet Talk, a text-messaging system to support young people with diabetes. Diabet Med 2006 Dec;23(12):1332-1338. [doi: 10.1111/j.1464-5491.2006.01989.x] [Medline: 17116184]
58. Fischer HH, Moore SL, Ginosar D, et al. Care by cell phone: text messaging for chronic disease management. Am J Manag Care 2012 Feb 1;18(2):e42-e47 [FREE Full text] [Medline: 22435883]
59. Kolmodin MacDonell K, Naar S, Gibson-Scipio W, Lam P, Secord E. The Detroit Young Adult Asthma Project: pilot of a technology-based medication adherence intervention for African-American emerging adults. J Adolesc Health 2016 Oct;59(4):465-471. [doi: 10.1016/j.jadohealth.2016.05.016] [Medline: 27475032]
60. Sansoni JE, Grootemaat P, Duncan C. Question prompt lists in health consultations: a review. Patient Educ Couns 2015 Jun 3:S0738-3991(15)00258-X. [doi: 10.1016/j.pec.2015.05.015] [Medline: 26104993]
61. Brandes K, Linn AJ, Butow PN, van Weert JCM. The characteristics and effectiveness of question prompt list interventions in oncology: a systematic review of the literature. Psychooncology 2015 Mar;24(3):245-252. [doi: 10.1002/pon.3637] [Medline: 25082386]
62. Street RL, Slee C, Kalauokalani DK, Dean DE, Tancredi DJ, Kravitz RL. Improving physician-patient communication about cancer pain with a tailored education-coaching intervention. Patient Educ Couns 2010 Jul;80(1):42-47. [doi: 10.1016/j.pec.2009.10.009] [Medline: 19962845]
63. Eggly S, Tkatch R, Penner LA, et al. Development of a question prompt list as a communication intervention to reduce racial disparities in cancer treatment. J Cancer Educ 2013 Jun;28(2):282-289. [doi: 10.1007/s13187-013-0456-2] [Medline: 23440665]
64. Eggly S, Hamel LM, Foster TS, et al. Randomized trial of a question prompt list to increase patient active participation during interactions with Black patients and their oncologists. Patient Educ Couns 2017 May;100(5):818-826. [doi: 10.1016/j.pec.2016.12.026] [Medline: 28073615]
65. Gale NK, Heath G, Cameron E, Rashid S, Redwood S. Using the framework method for the analysis of qualitative data in multi-disciplinary health research. BMC Med Res Methodol 2013;13(1):117. [doi: 10.1186/1471-2288-13-117]
66. Adler NE, Epel ES, Castellazzo G, Ickovics JR. Relationship of subjective and objective social status with psychological and physiological functioning: preliminary data in healthy, White women. Health Psychol 2000;19(6):586-592. [doi: 10.1037//0278-6133.19.6.586]
67. Raiff BR, Barrry VB, Ridenour TA, Jitnarin N. Internet-based incentives increase blood glucose testing with a non-adherent, diverse sample of teens with type 1 diabetes mellitus: a randomized controlled trial. Behav Med Pract Policy Res 2016 Jun;6(2):179-188. [doi: 10.1007/s13142-016-0397-5]
68. Goyal S, Nunn CA, Rotondi M, et al. A mobile app for the self-management of type 1 diabetes among adolescents: a randomized controlled trial. JMIR Mhealth Uhealth 2017 Jun 19;5(6):e82. [doi: 10.2196/mhealth.7336] [Medline: 28630037]
69. Bergner EM, Whittemore R, Patel NJ, Savin KL, Hamburger ER, Jaser SS. Participants' experience and engagement in check it!: a positive psychology intervention for adolescents with type 1 diabetes. Transl Issues Psychol Sci 2018;4(3):215-227. [doi: 10.1037/tps0000161] [Medline: 30505889]

XSL•FO
RenderX

70. Venkatesh V, Thong JYL, Xu X. Consumer acceptance and use of information technology: extending the Unified Theory of Acceptance and Use of Technology. MIS Q 2012 Mar 1;36(1):157-178. [doi: 10.2307/41410412]
71. Sharma A, Charulatha K. Peer pressure: a comprehensive literature review of the last two decades. Int J Fam Med Res 2019;2(4):1-10 [FREE Full text]
72. Nittas V, Daniore P, Chavez SJ, Wray TB. Challenges in implementing cultural adaptations of digital health interventions. Commun Med (Lond) 2024 Jan 5;4(1):7. [doi: 10.1038/s43856-023-00426-2] [Medline: 38182750]
73. Joo JY, Liu MF. Culturally tailored interventions for ethnic minorities: a scoping review. Nurs Open 2021 Sep;8(5):2078-2090. [doi: 10.1002/nop2.733] [Medline: 34388862]
74. Schaechter T, Flowers SN, Weiss M, Becker-Haimes EM, Sanchez AL. Culturally adapted interventions for anxiety and trauma-related disorders in marginalized youth: a systematic review. Child Psychiatry Hum Dev 2025 May 21. [doi: 10.1007/s10578-025-01833-3] [Medline: 40397328]
75. Balkhi AM, Reid AM, Westen SC, Olsen B, Janicke DM, Geffken GR. Telehealth interventions to reduce management complications in type 1 diabetes: a review. World J Diabetes 2015 Apr 15;6(3):371-379. [doi: 10.4239/wjd.v6.i3.371] [Medline: 25897348]
76. El-Gayar O, Timsina P, Nawar N, Eid W. A systematic review of IT for diabetes self-management: are there yet? Int J Med Inform 2013 Aug;82(8):637-652. [doi: 10.1016/j.ijmedinf.2013.05.006] [Medline: 23792137]
77. Ayling K, Brierley S, Johnson B, Heller S, Eiser C. Efficacy of theory‐based interventions for young people with type 1 diabetes: a systematic review and meta‐analysis. British J Health Psychol 2015 May;20(2):428-446. [doi: 10.1111/bjhp.12131]
78. Idalski Carcone A, Holtz BE, Reardon M, Vesey D, Ellis DA, Parks M. Meeting the needs of emerging adults with type 1 diabetes living in a rural area with mobile health interventions: focus group study. JMIR Form Res 2024 Aug 7;8:e55650. [doi: 10.2196/55650] [Medline: 39110496]
79. Kamody RC, Berlin KS, Rybak TM, et al. Psychological flexibility among youth with type 1 diabetes: relating patterns of acceptance, adherence, and stress to adaptation. Behav Med 2018;44(4):271-279. [doi: 10.1080/08964289.2017.1297290] [Medline: 28524766]
80. Collins LM, Murphy SA, Strecher V. The multiphase optimization strategy (MOST) and the sequential multiple assignment randomized trial (SMART): new methods for more potent ehealth interventions. Am J Prev Med 2007 May;32(5 Suppl):S112-S118. [doi: 10.1016/j.amepre.2007.01.022] [Medline: 17466815]
81. Gelles-Watnick R. Americans' use of mobile technology and home broadband. Pew Research Center. 2024 Jan 31. URL: https://www.pewresearch.org/internet/2024/01/31/americans-use-of-mobile-technology-and-home-broadband/ [accessed 2025-11-25]
82. Aoki K, Downes EJ. An analysis of young people's use of and attitudes toward cell phones. Telemat Inform 2003 Nov;20(4):349-364. [doi: 10.1016/S0736-5853(03)00018-2]
83. Auter PJ. Portable social groups: willingness to communicate, interpersonal communication gratifications, and cell phone use among young adults. Int J Mob Commun 2007;5(2):139. [doi: 10.1504/IJMC.2007.011813]
84. Fotheringham MJ, Wonnacott RL, Owen N. Computer use and physical inactivity in young adults: public health perils and potentials of new information technologies. Ann Behav Med 2000;22(4):269-275. [doi: 10.1007/BF02895662] [Medline: 11253437]
85. Giovanelli A, Ozer EM, Dahl RE. Leveraging technology to improve health in adolescence: a developmental science perspective. J Adolesc Health 2020 Aug;67(2S):S7-S13. [doi: 10.1016/j.jadohealth.2020.02.020] [Medline: 32718517]

Abbreviations

CGM: continuous glucose monitoring
CIAS 3.0: Computerized Intervention Authorizing Software, version 3.0
FAMS-T1D: Family and Friend Activation to Motivate Self-Care
FIF: Family Information Form
GSD: guided self-determination
HbA_{1c}: hemoglobin A_{1c}
MES: motivation enhancement system
mHealth: mobile health
QPL: question prompt list
SDT: self-determination theory
T1D: type 1 diabetes

XSL•FO
RenderX

Edited by A Stone, A Mavragani; submitted 07.04.25; peer-reviewed by C Cronin, S Skovlund; revised version received 17.10.25; accepted 30.10.25; published 27.11.25.

Please cite as:
Idalski Carcone A, Baskar D, Mahmood A, MacDonell K, Eggly S, Ghosh S, Buggs-Saxton C, Ondersma SJ, Ellis DA
Developing eHealth Interventions to Improve Diabetes Management in Emerging Adulthood: Qualitative Formative Study
JMIR Form Res 2025;9:e75623
URL: https://formative.jmir.org/2025/1/e75623
doi:10.2196/75623

© April Idalski Carcone, Deepika Baskar, Aishah Mahmood, Karen MacDonell, Susan Eggly, Samiran Ghosh, Colleen Buggs-Saxton, Steven J Ondersma, Deborah A Ellis. Originally published in JMIR Formative Research (https://formative.jmir.org), 27.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Smartphones and Mental Health Awareness and Utilization in a Low-Income Urban Community: Focus Group Study

Nadia Alam[1,2], MSc; Domenico Giacco[1,3], MBBS, PhD; Bulbul Siddiqi[4], PhD; Swaran Singh[1,2], MBBS, MD, DM; Sagar Jilka[1,2,5], PhD

[1]Warwick Medical School, University of Warwick, Gibbet Hill Road, Coventry, United Kingdom
[2]Warwick Centre for Global Health, University of Warwick, Coventry, United Kingdom
[3]Coventry and Warwickshire Partnership NHS Trust, Coventry, United Kingdom
[4]Department of Political Science and Sociology, North South University, Dhaka, Bangladesh
[5]Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom

Corresponding Author:
Nadia Alam, MSc
Warwick Medical School, University of Warwick, Gibbet Hill Road, Coventry, United Kingdom

Abstract

Background: Mental health disorders pose a significant challenge in low- and middle-income countries (LMICs), contributing substantially to the global disease burden. Despite the high prevalence of these disorders, LMICs allocate less than 1% of health budgets to mental health, resulting in inadequate care and a severe shortage of professionals. Stigma and cultural misconceptions further hinder access to mental health services. These challenges are present in Bangladesh, with high prevalence rates of depression and anxiety, as well as a centralized and underresourced mental health care system. Digital tools, such as smartphone apps and online platforms, offer innovative solutions to these challenges by increasing accessibility, cost-effectiveness, and scalability of mental health interventions.

Objective: This study aims to characterize the views around digital tools for mental health among residents of Korail (a major slum in Dhaka, Bangladesh), including the use of smartphones, and investigate acceptable digital tools and barriers and facilitators for digital mental health tools.

Methods: A total of 8 focus group discussions were conducted with 38 participants, including individuals with serious mental disorders and their caregivers. The focus group discussions were guided using a semistructured topic guide, which included broad questions on smartphone usage to contextualize digital access, primarily focusing on perceptions of using mobile technology for mental health care. Focus groups were held in Bangla, audio recorded, and transcribed and translated in English. Data were analyzed using thematic analysis in NVivo 14.

Results: Participants (mean age 37 y, SD 13.7) were mostly female (30/38, 79%), and 45% (17/38) personally owned smartphones, although 92% (35/38) reported smartphone access within the household. The findings revealed a general lack of awareness and understanding of digital mental health tools among slum residents. However, there was a notable appetite for these tools; participants recognized their potential to provide timely and cost-effective support, reduce hospital visits, and make health care more accessible. Participants highlighted the convenience and communication benefits of smartphones but expressed concerns about misuse such as excessive use, particularly among adolescents. Barriers to the utilization of digital mental health tools included limited technological literacy and accessibility issues. Despite these challenges, participants acknowledged the potential of these tools to bridge the gap in mental health services, especially for those unable to travel. The importance of providing proper guidance and education to maximize the effectiveness of digital tools was emphasized.

Conclusions: Digital mental health tools hold promise for improving mental health care in underserved slum communities. This study underscores the need for further research and investment in tailored digital mental health solutions to address the unique needs of slum populations in LMICs.

(JMIR Form Res 2025;9:e65650) doi:10.2196/65650

KEYWORDS

mental health; mHealth; lower- and middle-income countries; smartphones; awareness; developing countries; focus group study; mental health disorders; LMICs; digital tools; digital health; digital mental health; thematic analysis

XSL•FO
RenderX

Introduction

Mental Health in Low- and Middle-Income Countries

Mental health disorders significantly contribute to the global burden of disease, with 11.1% of the total burden in low- and middle-income countries (LMICs) attributed to mental health issues [1-4]. The World Health Organization reports that unipolar depressive disorder, schizophrenia, and bipolar disorder are among the leading causes of disability in these regions [5]. Furthermore, substance use disorders, particularly alcohol use, account for a significant portion of the disease burden in LMICs [6]. Despite the high prevalence of mental health disorders, resources are severely limited [2]. Less than 1% of health budgets in many LMICs are allocated to mental health, resulting in inadequate care for the majority of those affected [7].

Mental Health in Bangladesh and Slums

Mental health disorders are prevalent in Bangladesh, affecting approximately 16.1% of the population [8], yet they are often underreported and inadequately addressed due to limited resources, infrastructure, and a severe shortage of mental health professionals, with only 0.5 psychiatrists and 0.2 psychologists per 100,000 people [9]. In urban slum areas like Korail in Dhaka, mental health disorders are disproportionately high due to adverse living conditions such as poverty, overcrowding, and lack of basic services, with prevalence rates 30%-50% higher than in non-slum urban areas [10-15]. These environments are marked by high levels of violence, substance abuse, and chronic stress, exacerbating mental health issues while social support networks are often weak or absent, leading many to rely on informal support or go without care [16]. Efforts to improve mental health services, including policy initiatives and the establishment of mental health units in district hospitals, have been slow, with services often underfunded and lacking trained professionals [2,8,17]. This highlights the urgent need for innovative solutions to address mental health care in these regions.

Mental Health and Digital Tools

Digital tools, such as smartphone apps and online platforms, have been utilized to deliver mental health interventions such as cognitive-behavioral therapy and mindfulness training [18], provide psychoeducation, and offer support networks [19-21]. These tools can bridge the gap in mental health services, especially in regions with limited access to traditional mental health care [22]. Digital tools offer several advantages, including accessibility, cost-effectiveness, and scalability [23]. Furthermore, digital platforms can facilitate self-monitoring and provide users with real-time feedback and support, enhancing the effectiveness of mental health interventions [24].

Evidence demonstrates that some LMICs have successfully adopted digital technology to raise awareness, screen, diagnose, and manage common mental health disorders for wider populations [23,25-28]. For instance, in India, mobile health (mHealth) initiatives have been used to deliver mental health services to rural and underserved populations, showing promising results in improving access and outcomes [28]. Similarly, in Kenya, digital platforms have been utilized to provide mental health education and support, significantly enhancing mental health literacy and reducing stigma [29]. A study in South Africa found that a mobile phone–based intervention for depression was effective in reducing symptoms and improving quality of life among participants [30]. Another example is the use of telepsychiatry in remote areas of Brazil, which has improved access to mental health care and reduced the need for travel to urban centers [31]. These examples highlight the potential of digital tools to address the mental health care gap in LMICs.

Digital tools can be particularly useful in impoverished communities like slums, where traditional mental health services are scarce [2,11,14], but smartphone ownership is vast [32]. However, understanding slum residents' perception of smartphones and digital tools for mental health is crucial for the successful implementation of these technologies. Perceptions influence the acceptance and utilization of digital interventions, and addressing concerns such as privacy, usability, and cultural relevance is essential [33]. Understanding general patterns of smartphone use in this community provides critical context for interpreting how residents might engage with or resist mobile mental health solutions. Engaging communities in the design and implementation process can enhance the effectiveness and sustainability of digital mental health programs [34].

Given the increasing availability of smartphones in low-resource settings and the growing potential of digital tools to support mental health care, it is essential to understand how these tools are perceived by underserved populations. This study therefore aimed to explore the awareness, perceptions, and acceptance of smartphone-based mental health tools among residents of Korail slum in Dhaka, Bangladesh. Specifically, the objectives were to assess familiarity with digital mental health tools, identify barriers and facilitators to their use, and understand how such tools might support mental health care in this context.

Methods

Study Design

This research formed part of a larger collaborative research project on mental health in slums of LMICs [11]. This study employed a qualitative research design to explore perceptions on smartphones and use of digital tools for mental health among residents of the Korail slum in Dhaka, Bangladesh. The qualitative approach was chosen to gather in-depth insights into the participants' perceptions and attitudes toward digital tools for mental health. The study is part of a larger mixed-methods project aimed at understanding and predicting mental health relapse using digital phenotyping in LMICs.

Study Site and Participants

This study was conducted in Korail slum, Dhaka, Bangladesh, through group discussions with patients and caregivers. Korail is one of the largest slums in Bangladesh, located in the center of the city, encroaching on parts of Dhaka's most affluent neighborhoods. The majority of Korail's population (estimated to be around 200,000) [12] subsist below the poverty line and work in low-income jobs. They have limited access to the city's health care services, and there is little connection with the

XSL•FO
RenderX

surrounding major roads [35]. Participants had to have been residents of the Korail slum, aged 18 years or older, and either diagnosed with a serious mental disorder (SMD) or caregivers of individuals diagnosed with SMD. We used purposive sampling to ensure inclusion of individuals with lived experience of SMDs and their caregivers, reflecting diverse perspectives on smartphone and digital health tool usage. Recruitment was facilitated by community engagement representatives associated with the TRANSFORM (Transforming Access to Care for Serious Mental Disorders in Slums) Project [11], who are trusted members of the local community. These trusted contacts played a key role in introducing the study to potential participants and delivering culturally sensitive explanations about its purpose and procedures. Recruitment was conducted through informal community engagement strategies, including conversations at local tea stalls and the TRANSFORM project field office. No public advertisements were used. All potential participants received an information sheet and were given at least 24 hours to consider participation.

Ethical Considerations

Human Subject Ethics Review Approvals

This study received ethical clearance from the Biomedical and Scientific Research Ethics Committee at the University of Warwick (reference number: BSREC 100/22 - 23). It formed part of the wider National Institute for Health and Care Research–funded TRANSFORM project [11], which underwent rigorous ethical review both in the United Kingdom and by partner institutions in Bangladesh to ensure compliance with local ethical standards.

A safeguarding and risk management plan was developed to prioritize participant well-being throughout the study. Given the sensitive nature of the topic—discussions about personal or family experiences with mental illness—there was potential for emotional distress. Data collection was led by NA—a trained researcher and doctoral candidate with a background in psychology who was trained in research ethics and working with vulnerable populations. The safeguarding protocol allowed for immediate responses if a participant showed signs of distress, including pausing or stopping the session, checking in with the participant, and offering referral to psychiatric support available through the local study team.

Informed Consent

Informed written consent was obtained from all participants before data collection activities began. Study information sheets and consent forms were provided in both English and Bengali to ensure accessibility. Participants were given a minimum of 24 hours to consider their involvement. For participants with lived experience of mental illness, we used a previously validated capacity assessment tool from the TRANSFORM project to confirm ability to consent. In cases where participants could not sign, a thumbprint was collected in the presence of a witness. Verbal consent was also recorded at the beginning of each focus group. Participants were informed of their right to withdraw at any time without providing a reason.

Privacy and Confidentiality

All participant data were anonymized and stored on encrypted devices in compliance with the UK General Data Protection Regulation and the UK Data Protection Act (2018). Signed consent forms were stored separately in locked filing cabinets. Focus group discussion (FGD) transcripts were deidentified before analysis, and no personally identifying information is included in this paper or any outputs. Digital files were stored on secure, access-restricted university servers.

Compensation

Each participant received compensation of 800 Bangladeshi Taka (approximately £5.30/US $6.60) to cover their time and any expenses. This amount reflected local wage and travel standards, especially considering that Korail is a large urban slum where travel to field offices requires time and costs. They were assured that compensation was not conditional upon completion of the study and would still be given in the event of early withdrawal.

Participant Identification in Study Materials

No identifying information will be disclosed in any study publications or supplementary materials. All data were anonymized prior to analysis, and each participant was assigned a unique code to maintain confidentiality.

Data Collection

The FGDs were conducted in the Korail slum, within the established research office of the TRANSFORM Project [11]. The FGDs were structured to ensure a diverse representation of the community, including different genders, ages, and socioeconomic status. Each session was moderated by a trained facilitator, with the assistance of a notetaker and qualitative researcher, and lasted for a duration of 30 - 45 minutes. The moderator (NA) guided the discussions using a semistructured topic guide, which included broad questions on smartphone usage to contextualize digital access, but the primary emphasis was on discussing perceptions of using mobile technology for mental health care, including reactions to examples of existing platforms. As part of the discussion, participants were introduced to examples of existing Bangladeshi digital health tools used on smartphones, including Doctime (a telemedicine platform) [36] and Maya Apa (a mobile app offering health and mental well-being support) [37], to help them conceptualize how digital mental health support might be delivered. The FGDs were audio recorded in Bangla, the local language, to ensure accurate capture of the discussions. The recordings were then transcribed verbatim and translated into English. To maintain confidentiality, all personal identifiers were removed during transcription. The translated transcripts were reviewed by the research team to ensure accuracy and fidelity to the original conversations.

Data Analysis

The qualitative data from the FGDs were analyzed using NVivo 12 (Lumivero) software. A thematic analysis approach was employed to identify patterns and themes within the data. The analysis followed the five stages of data analysis: familiarization,

XSL•FO
RenderX

identifying a thematic framework, indexing, charting, and mapping and interpretation [38].

Each transcript was read multiple times to enhance familiarity with the data. Initial thoughts, reflections, and preliminary codes were noted. The transcript was then read again, and preliminary themes were recorded. Preliminary themes were then grouped into clusters based on common features and meanings. A coding tree was developed to systematically organize the themes, capturing the hierarchical relationships between main themes, subthemes, and associated codes. This framework enabled a nuanced analysis of participants' perspectives and experiences. The coding structure was iteratively refined through regular team discussions, and each theme was cross-checked against the transcripts to ensure accurate representation of the data. This process was applied consistently across all transcripts to enhance the rigor and coherence of the analysis. The final coding tree is presented in Figure 1. These themes were validated by cross-checking with the transcript to ensure they accurately represented the data. This process was repeated for each transcript to maintain consistency. The themes from all transcripts were then compared and combined into master themes to create a comprehensive portrayal of the participants' experiences. The master themes were thoroughly reviewed against the transcripts to ensure they accurately reflected and were firmly rooted in the participants' accounts. Commonalities among the preliminary themes were identified and represented as subthemes, reflecting lower-order aspects of the master themes. To enhance confirmability [39], the analysis was grounded in participants' accounts through the use of direct, anonymized quotes that illustrate each theme. A clear audit trail was maintained throughout, documenting coding decisions, thematic development, and reflections in analytical memos to ensure dependability [39]. The final coding framework was reviewed and refined collaboratively by multiple researchers to ensure consistency and rigor, with any discrepancies resolved through discussion and consensus.

Figure 1. Coding tree illustrating key themes and subthemes derived from the thematic analysis. Focus group discussions on smartphone use and perceptions of digital mental health tools among individuals with serious mental disorders and their caregivers in Korail slum, Dhaka, Bangladesh.

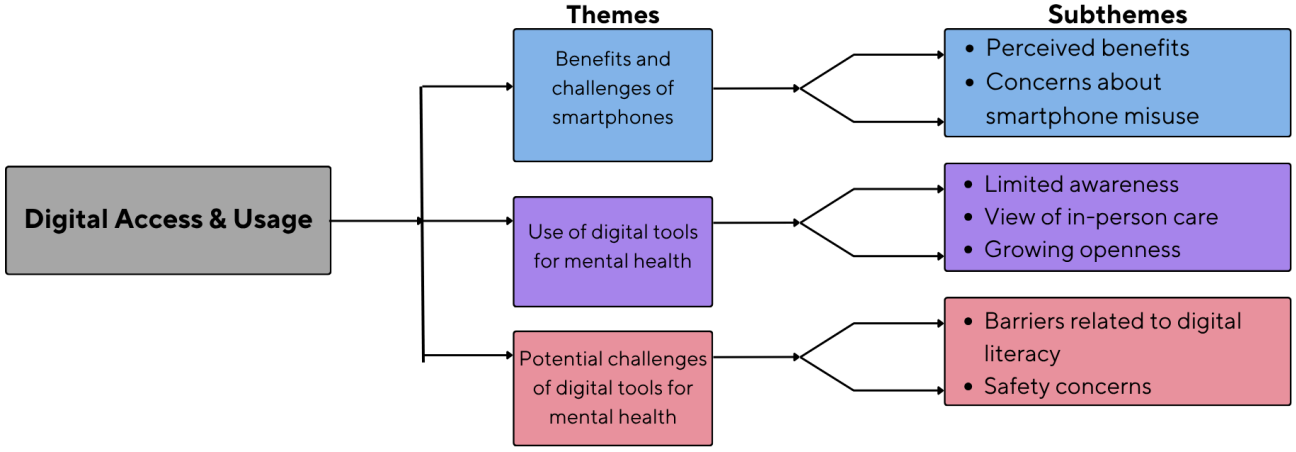

Reflexivity [40] was integrated throughout the analytical process. The research team actively reflected on their positionalities, including cultural backgrounds and professional roles, to remain aware of how these factors might shape data interpretation. Regular team discussions were held to critically assess assumptions and ensure the analysis remained anchored in participants' lived experiences. To support credibility [39], emerging themes were continuously reviewed through peer debriefing and iterative comparisons across focus groups. This process helped validate thematic consistency and ensure findings captured the breadth of perspectives shared by both individuals with SMDs and their caregivers.

Transferability [39] was supported by providing rich contextual detail about the study setting and participant demographics. Descriptions of life in Korail slum, including social and infrastructural challenges, offer important context for assessing how these findings may resonate with other low-resource, urban communities in Bangladesh and similar LMIC settings [2,41]. The transparent analytical process and inclusion of participant narratives further support the relevance of these findings for broader application.

Results

Demographic Information

A total of 8 FGDs were conducted, each group comprising 4 - 6 participants who were either people with lived experiences or caregivers. The total number of participants for the 8 FGDs was 38 (Table 1). The average age of participants was 37 (SD 13.7) years, mostly comprising female participants. Out of the 38 participants, 17 (45%) personally owned a smartphone, while 35 (92%) reported having access to one within their household. The average duration of personal smartphone ownership was 3.3 (SD 1.15) years, with most users relying on low-cost Android phones.

XSL•FO
RenderX

Table . Participant demographic information (N=38).

Characteristics	Count
Age (y), mean (SD)	37 (13.7)
Gender (female), n (%)	30 (79)
Education level, n (%)	
No education	5 (13)
Primary	11 (29)
Middle school	16 (42)
High school	4 (11)
Undergraduate	2 (5)
Marital status, n (%)	
Married	29 (76)
Unmarried	4 (11)
Widowed	4 (11)
Divorced	1 (2)
Employed, n (%)	
Yes	12 (32)
No	26 (68)
Smartphone ownership, n (%)	
Yes	17 (45)
No	21 (55)
Average time of Smartphone ownership (y), mean (SD)	3.33 (1.15)
Smartphone ownership in family, n (%)	
Yes	35 (92)
No	3 (8)

Overview of Qualitative Themes

We found three major themes: smartphone use and benefits, digital awareness and attitudes, and challenges in using digital tools for mental health (Figure 1). These are expanded on further below.

Benefits and Challenges of Digital Tools

Perceived Benefits

Overall, participants were familiar with digital tools for other parts of their lives and acknowledged its convenience within everyday life. For instance, one participant mentioned, “Before we had to stand at the bank all day to withdraw disability allowance now it directly comes to the phone.” Participants emphasized the importance of communication, video calls, and messaging apps in maintaining relationships, especially with family members living abroad, and the ability to access news and information through social media platforms like Facebook and WhatsApp was also highlighted as a significant benefit. One stated, “People before used to send letters to seek updates, it took time to hear back. Now it is easy to know about people (using smartphones).” Another participant added, “Communication with friends and family is much easier with smartphones.” There was also discussion about convenience of access to information using smartphones. One participant added, “Phones are used for accessing information about health and various services.” They expressed that access to timely and relevant information is valued and seen as potentially helpful, particularly when it is available in a convenient and familiar format.

Concerns About Smartphone Misuse

Participants highlighted various ways in which smartphones can be misused, particularly by adolescents. Concerns were raised about inappropriate content, the facilitation of risky behaviors, and negative impacts on academic performance. Others highlighted concerns about addiction, particularly among the younger generation, as children and teenagers spend excessive amounts of time on their phones, often at the expense of studying or socializing with family. Another participant expressed frustration with the excessive gaming and its impact: “Every kid is so busy playing games, it seems that they have no time to pick up the call from family,” and another added, “Whole day they (younger family members) are on it (phones), even if you slap them, they will not let it go.”

XSL•FO
RenderX

Use of Digital Tools for Mental Health

Limited Awareness of Digital Mental Health Tools

Despite familiarity of digital tools, awareness of digital mental health tools was generally low among the participants. Many were unaware of the existence of apps and online services designed to provide mental health support. One participant stated simply, “I have no idea” when asked about health services through phones. Another echoed, “No, I don’t know” about apps for mental health. Another participant mentioned, “We directly communicate with the doctor; no service was ever taken over the phone.” This lack of knowledge extended to the potential benefits of such tools, with several participants indicating that they had never used or even heard of mental health apps, “No, I have never used any such app.” However, one participant shared, “I once spoke to an Indian doctor through an app. He told me how to talk to my sister-in-law, how to behave, all of this.”

View of In-Person Care

Despite the potential of digital tools, many participants preferred traditional methods of seeking mental health care. They commonly visited doctors in person and relied on face-to-face consultations for treatment and advice. One participant shared, “We go to a doctor; the doctor gives one month of medicine.” This preference was attributed to the trust and reassurance that comes with direct interaction with health care professionals.

Growing Openness to Mobile-Based Mental Health Support

While most participants had limited experience with online consultations, there was a general interest in exploring this option. One participant added, “Consulting a doctor through an app can be very helpful.” Some expressed the view that online consultations could save time and reduce the costs associated with traveling to health care facilities. One participant noted, “Online consultations could save us a lot of travel time and cost.” The convenience of receiving medical advice from home was highlighted, especially for those with mobility issues or residing in remote areas. The potential of health apps to provide immediate assistance and support was recognized by several participants. They suggested that apps could offer valuable information on managing mental health conditions, including treatment schedules and medication reminders, with one stating, “If counseling, doctor’s advice, what kind of treatment should be given to them can be provided (through phones), it would be good.” Participants also noted that health apps could serve as a resource for caregivers, providing guidance on how to care for individuals with mental health issues, “health apps can guide us on how to care for our loved ones better.” Another participant reflected on the need for such tools, saying, “It would be great if there is a service available through the app for such problems,” acknowledging the limitations they face in accessing in-person services. Caregivers also recognized the potential benefits of digital tools in guiding them on how to care for individuals with mental health issues: “Health apps can guide us on how to care for our loved ones better.” This indicates that digital tools can serve as valuable resources for those supporting individuals with mental health challenges. The general sentiment was positive toward the potential of these services, with one participant mentioning, “If I get advice from a good doctor from home, it will be good,” and another stating, “Consulting a doctor through an app can be very helpful.”

Potential Challenges

Barriers Related to Digital Literacy and Usability

Despite widespread appetite for digital mental health tools, participants acknowledged challenges such as the digital divide and lack of technological literacy, and concerns about the reliability of online services were also mentioned. One participant highlighted, “It saves time and travel costs but not everyone knows how to use the apps properly.” Another added, the importance of addressing these challenges to ensure the successful implementation of digital mental health interventions was emphasized, “It will be good, but we need proper guidance on how to use these services.”

Safety Concerns for Individuals With Mental Illness

Participants also expressed concerns about the usability of these tools for older family members, stating, “It can be challenging to teach older family members how to use health apps.” The need for user education and support to maximize the effectiveness of these digital tools was stressed. Furthermore, participants discussed the risks for mentally ill individuals using digital tools. One participant shared an experience: “When my sister-in-law was sick and had the phone on her, she used to make such bad comments, leave bad images, and she used to make very bad posts on people’s profiles.” Another participant emphasized the potential risks: “Mentally ill patients can do many things on the phone without realizing it....Harassment can be due to political reasons.” These accounts highlight the importance of considering the specific vulnerabilities of certain user groups when developing and implementing digital health interventions. Ensuring that appropriate safeguards and support mechanisms are in place is crucial to mitigate these risks.

Discussion

Principal Results

This study explored the perceptions and awareness of digital mental health tools among residents of Korail slum in Dhaka, Bangladesh. Participants were generally unaware of existing mental health apps but expressed strong interest in using mobile-based tools if appropriate support, content, and guidance were available. Key barriers included low digital literacy, stigma, and a preference for face-to-face consultations, while convenience and cost-saving were cited as major facilitators.

The findings reveal a general lack of awareness and understanding of digital mental health tools among the residents of the Korail slum. Many participants were unaware of the existence of apps and online services designed to provide mental health support, which highlights a significant gap in mental health literacy within the community. Despite this lack of awareness, there was a noticeable interest in the potential of digital mental health tools, with participants acknowledging that these tools could provide timely and cost-effective support,

XSL•FO
RenderX

reducing the need for frequent hospital visits and associated expenses.

Several barriers to the utilization of digital mental health tools were identified, including technological literacy and accessibility. The digital divide was a prominent concern, with participants highlighting the lack of technological literacy as a significant barrier. This issue is compounded by the challenges of teaching older family members how to use these tools. Additionally, concerns about the reliability and effectiveness of online services were mentioned, and traditional doctor visits are still preferred by many due to the trust and reassurance that comes with face-to-face interactions [42]. This preference may reflect deeper cultural and perceptual factors related to mental health [43,44] rather than a direct rejection of digital mental health interventions. In contexts where mental illness is often stigmatized, seeking face-to-face care from a trusted physician can offer a sense of legitimacy and reassurance [25,43,45]. Additionally, low awareness of mental health as a medical issue or tools to support this issue may lead to underappreciation of the value of remote or digital forms of care [44,45]. These findings suggest that efforts to introduce digital interventions should be accompanied by broader awareness and stigma reduction campaigns to reframe how mental health is understood and to position digital tools as complementary rather than substitutive to traditional care.

Despite these barriers, several facilitators to the utilization of digital mental health tools were identified. Participants recognized the potential of these tools to make health care more accessible, especially for those who cannot travel. This reflects the potential for digital tools to bridge the gap in mental health services, particularly in underserved areas like slums. The study also highlighted the importance of providing proper guidance and education on using these services. Participants stressed the need for user education and support to maximize the effectiveness of digital tools. This suggests that initiatives aimed at improving technological literacy and providing support could significantly enhance the adoption and effectiveness of digital mental health interventions [29].

Cultural beliefs, norms, and attitudes toward mental illness and technology can significantly influence the acceptability and use of digital mental health tools across LMICs [20,46,47]. While our study reflects the context of a Bangladeshi urban slum, where religious values, stigma, and informal care networks play a central role, these dynamics may differ in other regions [11]. For example, communities in sub-Saharan Africa may have different levels of mHealth adoption, while Latin American settings may feature stronger public mental health infrastructures or community health outreach programs. Such differences may shape trust in digital platforms, preferences for in-person versus remote care, and openness to app-based interventions. Therefore, further context-specific studies are needed to examine how cultural, technological, and health system factors affect digital mental health engagement across diverse LMICs.

While participants were introduced to both a telemedicine-style app (Doctime) and an autonomous, informational app (Maya Apa), most responses focused on the former. This suggests a preference or at least greater conceptual familiarity with real-time, doctor-driven models. The limited commentary on autonomous, non-telemedicine interventions likely reflects both unfamiliarity with such models and cultural expectations around personal interaction in health care. These perceptions are important, as they highlight a potential barrier to the implementation of scalable, self-guided digital mental health tools in this context. As the field moves toward autonomous solutions for scalability, especially in LMICs, our findings suggest that greater user education and co-design with target communities may be necessary to improve acceptability and uptake of these newer models of care.

These findings add to the growing body of evidence on the acceptability of digital mental health tools in low-resource settings. While global momentum is increasing around mHealth innovations, this study highlights the need for contextualized, culturally appropriate, and socially supported implementation strategies in urban slum environments.

Ethical Considerations in Implementations

Implementing digital mental health tools in low-resource settings presents several ethical challenges that must be carefully addressed to ensure safe and equitable access. As participants in this study highlighted, concerns about privacy, misuse of technology, and safety, particularly for individuals with SMDs, can pose real risks. In contexts where digital literacy is limited and access to trusted online health information is inconsistent, users may misinterpret content, misuse apps, or fall victim to misinformation or exploitation. These risks are amplified in environments with weak data protection infrastructure or low public awareness of digital rights [25]. Moreover, reliance on technology without adequate support mechanisms may marginalize users with low literacy, older adults, or individuals with cognitive impairments [48,49]. To mitigate these risks, digital mental health tools should be developed with clear privacy safeguards, culturally appropriate guidance, and integrated referral pathways for offline support. Involving local stakeholders and end-users in co-design is critical to ensure ethical alignment and social acceptability.

Limitations

Despite the valuable insights provided by this study, several limitations need to be acknowledged. First, the study was conducted in a single slum community in Dhaka, Bangladesh, which may limit the generalizability of the findings to other slum areas or LMICs with different sociocultural contexts [50]. The sample size, though appropriate for qualitative thematic analysis, is relatively small (n=38), which may limit the breadth and variability of perspectives captured [50]. Second, the study relied on self-reported data from FGDs, which may introduce biases such as social desirability bias, where participants might overstate positive aspects or underreport negative aspects to conform to perceived expectations [29]. Furthermore, while purposive sampling was appropriate for the exploratory and context-specific aims of this study, we acknowledge that it may have introduced some degree of selection bias. Participants were recruited through trusted community representatives affiliated with the TRANSFORM project [11], which likely facilitated openness and trust during data collection. However, this approach may have favored individuals who were more socially

XSL•FO
RenderX

engaged or willing to discuss mental health–related topics in a group setting. As with many qualitative studies in marginalized communities, those with the most severe symptoms or highest levels of stigma may have been less likely to participate [51]. While the sample captured a diverse range of perspectives, future research could incorporate additional recruitment strategies such as household outreach or individual interviews to ensure broader inclusion. While our findings offer valuable insights into community-wide perceptions of digital mental health tools, we were unable to explore differences between smartphone owners and nonowners. Although we collected demographic data on smartphone access, transcripts from the FGDs were anonymized and not linked to individual participants' ownership status. As a result, it was not feasible to systematically compare viewpoints across these two groups. Given that personal experience with smartphones may significantly shape perceptions of digital health interventions, future research using individually linked or interview-based data could help unpack these differences. This would allow for a more nuanced understanding of how digital familiarity influences the acceptability and perceived utility of mobile mental health tools in slum settings.

Additionally, the lack of awareness and understanding of digital mental health tools among participants could have influenced their responses, potentially limiting the depth of insights regarding the barriers and facilitators to utilizing these tools [33]. Technological literacy was identified as a significant barrier, yet the study did not quantitatively assess the participants' actual proficiency with digital tools. This gap indicates a need for future research to include objective measures of technological literacy to better understand this barrier [19]. As this was a focus group study designed to encourage open dialogue within groups rather than attribute views to individual characteristics, we did not analyze responses based on smartphone ownership status. While this approach captured a collective understanding of digital mental health perceptions in the community, future studies may benefit from comparing views across subgroups, such as smartphone users and nonusers, to better tailor digital interventions. Additionally, while full data saturation may not have been achieved, several key themes, such as attitudes toward smartphone use, digital literacy challenges, and preferences for traditional care, emerged consistently across the FGDs. The final two discussions reinforced earlier findings, suggesting thematic convergence. Given the formative and exploratory nature of this study, the insights remain valuable and provide a foundation for future qualitative work. This study was conducted in a single urban slum in Dhaka, Bangladesh, which may limit the extent to which the findings can be generalized to other LMIC contexts. However, many of the challenges identified, such as stigma, low mental health literacy, and limited digital access, are common across underserved urban settings in other LMICs [2,4,11,25,45]. As such, the insights gained may still be relevant and transferable. Future research should aim to include multiple communities across different regions or countries to better capture contextual variation and enhance generalizability. While this study provides foundational insights into the awareness and perceptions of digital mental health tools in a slum setting, its limitations highlight the need for further research to expand these findings and address the identified gaps.

Comparison to Prior Work

This study adds to a growing body of literature demonstrating the promise of digital mental health interventions in resource-constrained settings, particularly within LMICs. Our findings are consistent with previous work showing that digital tools can expand access, reduce treatment gaps, and provide cost-effective support in communities with limited formal mental health infrastructure [23,25,46,47,52] However, our study extends this literature by focusing on an underserved slum community where the digital health ecosystem remains nascent. In contrast to digital health successes reported in countries where mental health apps and mobile services have shown promising engagement [27,53,54], our participants demonstrated limited awareness of digital mental health tools and low technological literacy. For instance, the Enhanced mHealth mental health screening for adolescents living with HIV transitioning into adult care (EMMATAC) study conducted in Nairobi, Kenya, assessed a self-administered mHealth screening tool among adolescents and young adults living with HIV [54]. The study reported high feasibility (83%) and acceptability (80%), with a significant prevalence of mental health conditions such as post-traumatic stress disorder and depression among participants [54]. Similarly, the A Sequential, Multiple Assignment Randomized Trial (SMART) for Non-Specialist Treatment of Common Mental Disorders in Kenya: Leveraging the Depression and Primary-Care Partnership for Effectiveness-Implementation Research (DAPPER) Project (SMART-DAPPER) in western Kenya evaluated the delivery of psychotherapy and medication for depression and post-traumatic stress disorder via mHealth [27]. Several participants preferred mHealth over in-person treatment, citing affordability and convenience as primary reasons [27]. These findings suggest that while digital interventions may be viable at a systems level, community-level barriers persist that hinder their equitable adoption and meaningful use.

Our findings align with broader evidence from other LMICs. For example, previous studies demonstrated the effectiveness of mHealth interventions for improving mental health outcomes in rural Indian communities, particularly when combined with community-based education and engagement strategies [28]. However, concerns about the usability, reliability, and safety of digital platforms, particularly for individuals experiencing SMDs, were consistent with previously reported challenges in implementing digital mental health intervention [21]. This highlights the importance of designing interventions that are not only technologically accessible but also culturally resonant and ethically secure. Additionally, our study supports previous findings that the successful adoption of digital mental health interventions in marginalized communities requires more than mere access to technology [34,49]; it also demands a nuanced understanding of social contexts, trust dynamics, and user capabilities. Without deliberate efforts to address these factors, digital tools risk perpetuating rather than reducing existing health inequities.

XSL•FO
RenderX

In summary, while our findings are consistent with global evidence that digital tools hold promise for expanding mental health care in LMICs, they also highlight the importance of grounding digital health interventions in the lived realities of the communities they aim to serve. Tailored educational initiatives, community involvement in design and delivery, and safeguards for vulnerable users are essential to ensure these tools are not only available but also accessible, acceptable, and effective in slum settings.

Conclusion

Based on our findings, we propose several practical recommendations for developing and implementing digital mental health tools in low-resource slum settings. First, interventions should be accompanied by structured digital literacy training, particularly targeting older adults and individuals with limited experience using mobile technology. Second, culturally relevant content in local languages is critical to improve usability and engagement. Third, involving caregivers and community members in the design and delivery process can build trust and ensure that tools align with community needs. Fourth, mental health apps should include safeguards to support users with SMDs, such as guidance for caregivers and emergency contact features. Finally, these tools should be positioned as supplements to rather than replacements for face-to-face care to increase acceptance among users who strongly value in-person consultations.

This study shows people with lived experience are receptive and interested in the transformative power of digital tools for mental health care in slum settings like Korail. However, for these tools to be effective, it is crucial to address the existing barriers and provide adequate support and education to users. By doing so, digital interventions can become a significant solution to the mental health care challenges faced by underserved communities in LMICs. The findings highlight the need for further research and investment in digital mental health solutions tailored to the unique needs and contexts of slum populations.

Acknowledgments

We would like to acknowledge Md. Jakaria Habib and Sheikh Rasel for facilitating the focus group discussions. We also acknowledge the support of our local community representatives: Aulia Khatun, Md. Abdur Rahman, Ragina Aktar Rumi, Md. Abul Kashem Titu, Raseda, and Rashida Begum. We also acknowledge the work of the TRANSFORM (Transforming Access to Care for Serious Mental Disorders in Slums) Project, TRANSFORM consortium, and Telepsychiatry Research and Innovation Network Ltd. Team. This study is part of the TRANSFORM Project and is funded by the United Kingdom's National Institute for Health and Care Research (award number: NIHR200846).

Data Availability

Data are available upon reasonable request to the corresponding author and in line with the TRANSFORM (Transforming Access to Care for Serious Mental Disorders in Slums) study data sharing agreement, whereby all requests will be discussed with the study's principal investigator.

Authors' Contributions

NA and SJ wrote the manuscript. NA collected the data, and SJ supported the analysis. All authors designed the study and supported the manuscript writing.

Conflicts of Interest

None declared.

References

1. Patel V, Flisher AJ, Hetrick S, McGorry P. Mental health of young people: a global public-health challenge. Lancet 2007 Apr 14;369(9569):1302-1313. [doi: 10.1016/S0140-6736(07)60368-7] [Medline: 17434406]
2. Rathod S, Pinninti N, Irfan M, et al. Mental health service provision in low- and middle-income countries. Health Serv Insights 2017;10:1178632917694350. [doi: 10.1177/1178632917694350] [Medline: 28469456]
3. Witter S, Sheikh K, Schleiff M. Learning health systems in low-income and middle-income countries: exploring evidence and expert insights. BMJ Glob Health 2022 Sep;7(Suppl 7):e008115. [doi: 10.1136/bmjgh-2021-008115] [Medline: 36130793]
4. Javed A, Lee C, Zakaria H, et al. Reducing the stigma of mental health disorders with a focus on low- and middle-income countries. Asian J Psychiatr 2021 Apr;58:102601. [doi: 10.1016/j.ajp.2021.102601] [Medline: 33611083]
5. Saxena S, Herrman H, Moodie R. Promoting Mental Health: Concepts, Emerging Evidence, Practice: A Report of the World Health Organization, Department of Mental Health and Substance Abuse in Collaboration with the Victorian Health Promotion Foundation and the University of Melbourne: World Health Organization; 2005.
6. Neufeld SAS. The burden of young people's mental health conditions in Europe: no cause for complacency. Lancet Reg Health Eur 2022 Mar 24;16:100364. [doi: 10.1016/j.lanepe.2022.100364] [Medline: 35345644]

XSL•FO
RenderX

7. Saxena S, Thornicroft G, Knapp M, Whiteford H. Resources for mental health: scarcity, inequity, and inefficiency. Lancet 2007 Sep 8;370(9590):878-889. [doi: 10.1016/S0140-6736(07)61239-2] [Medline: 17804062]
8. Hossain MD, Ahmed HU, Chowdhury WA, Niessen LW, Alam DS. Mental disorders in Bangladesh: a systematic review. BMC Psychiatry 2014 Jul 30;14(1):216. [doi: 10.1186/s12888-014-0216-9] [Medline: 25073970]
9. Hasan MT, Anwar T, Christopher E, et al. The current state of mental healthcare in Bangladesh: part 1 - an updated country profile. BJPsych Int 2021 Nov;18(4):78-82. [doi: 10.1192/bji.2021.41] [Medline: 34747942]
10. Riley LW, Ko AI, Unger A, Reis MG. Slum health: diseases of neglected populations. BMC Int Health Hum Rights 2007 Mar 7;7(1):2. [doi: 10.1186/1472-698X-7-2] [Medline: 17343758]
11. Singh SP, Jilka S, Abdulmalik J, et al. Transforming access to care for serious mental disorders in slums (the TRANSFORM Project): rationale, design and protocol. BJPsych Open 2022 Oct 13;8(6):e185. [doi: 10.1192/bjo.2022.584] [Medline: 36226591]
12. Sinthia SA. Analysis of urban slum: case study of Korail slum, Dhaka. International Journal of Urban and Civil Engineering 2020 Nov;14(11):416-430 [FREE Full text]
13. Khan S, Rathore D, Singh A, Kumari R, Malaviya P. Socio-economic and environmental vulnerability of urban slums: a case study of slums at Jammu (India). Environ Sci Pollut Res Int 2024 Mar;31(12):18074-18099. [doi: 10.1007/s11356-023-30630-5] [Medline: 37919511]
14. Gruebner O, Khan MMH, Lautenbach S, et al. Mental health in the slums of Dhaka - a geoepidemiological study. BMC Public Health 2012 Mar 9;12(1):177. [doi: 10.1186/1471-2458-12-177] [Medline: 22404959]
15. Koly KN, Khanam MI, Islam MS, et al. Anxiety and insomnia among urban slum dwellers in Bangladesh: the role of COVID-19 and its associated factors. Front Psychiatry 2021;12:769048. [doi: 10.3389/fpsyt.2021.769048] [Medline: 34925097]
16. Patel V, Kleinman A. Poverty and common mental disorders in developing countries. Bull World Health Organ 2003;81(8):609-615. [Medline: 14576893]
17. Alloh FT, Regmi P, Onche I, Teijlingen EV, Trenoweth S. Mental health in low-and middle income countries (LMICs): going beyond the need for funding. Health Prospect 2018 Jun 19;17(1):12-17. [doi: 10.3126/hprospect.v17i1.20351]
18. Donker T, Petrie K, Proudfoot J, Clarke J, Birch MR, Christensen H. Smartphones for smarter delivery of mental health programs: a systematic review. J Med Internet Res 2013 Nov 15;15(11):e247. [doi: 10.2196/jmir.2791] [Medline: 24240579]
19. Torous J, Kiang MV, Lorme J, Onnela JP. New Tools for new research in psychiatry: a scalable and customizable platform to empower data driven smartphone research. JMIR Ment Health 2016 May 5;3(2):e16. [doi: 10.2196/mental.5165] [Medline: 27150677]
20. Chakrabarti S. Digital psychiatry in low-and-middle-income countries: new developments and the way forward. World J Psychiatry 2024 Mar 19;14(3):350-361. [doi: 10.5498/wjp.v14.i3.350] [Medline: 38617977]
21. Torous J, Wisniewski H, Bird B, et al. Creating a digital health smartphone app and digital phenotyping platform for mental health and diverse healthcare needs: an interdisciplinary and collaborative approach. J Technol Behav Sci 2019 Jun;4:73-85. [doi: 10.1007/s41347-019-00095-w]
22. Lattie EG, Adkins EC, Winquist N, Stiles-Shields C, Wafford QE, Graham AK. Digital mental health interventions for depression, anxiety, and enhancement of psychological well-being among college students: systematic review. J Med Internet Res 2019 Jul 22;21(7):e12869. [doi: 10.2196/12869] [Medline: 31333198]
23. Mohr DC, Burns MN, Schueller SM, Clarke G, Klinkman M. Behavioral intervention technologies: evidence review and recommendations for future research in mental health. Gen Hosp Psychiatry 2013;35(4):332-338. [doi: 10.1016/j.genhosppsych.2013.03.008] [Medline: 23664503]
24. Henson P, Barnett I, Keshavan M, Torous J. Towards clinically actionable digital phenotyping targets in schizophrenia. NPJ Schizophr 2020 May 5;6(1):13. [doi: 10.1038/s41537-020-0100-1] [Medline: 32372059]
25. Naslund JA, Deng D. Addressing mental health stigma in low-income and middle-income countries: a new frontier for digital mental health. Ethics Med Public Health 2021 Dec;19:100719. [doi: 10.1016/j.jemep.2021.100719] [Medline: 35083375]
26. Carter H, Araya R, Anjur K, Deng D, Naslund JA. The emergence of digital mental health in low-income and middle-income countries: a review of recent advances and implications for the treatment and prevention of mental disorders. J Psychiatr Res 2021 Jan;133:223-246. [doi: 10.1016/j.jpsychires.2020.12.016] [Medline: 33360867]
27. Meffert S, Mathai M, Neylan T, et al. Preference of mHealth versus in-person treatment for depression and post-traumatic stress disorder in Kenya: demographic and clinical characteristics. BMJ Open 2024 Nov 18;14(11):e083094. [doi: 10.1136/bmjopen-2023-083094] [Medline: 39557549]
28. Joshi V, Joshi NK, Bhardwaj P, Singh K, Ojha D, Jain YK. The health impact of mHealth interventions in India: systematic review and meta-analysis. Online J Public Health Inform 2023;15:e50927. [doi: 10.2196/50927] [Medline: 38046564]
29. Gonsalves NJ, Ogunseiju OR, Akanmu AA, Nnaji CA. Assessment of a passive wearable robot for reducing low back disorders during rebar work. Journal of Information Technology in Construction 2021 Nov;26:936-952. [doi: 10.36680/j.itcon.2021.050]
30. Tomlinson M, Grimsrud AT, Stein DJ, Williams DR, Myer L. The epidemiology of major depression in South Africa: results from the South African stress and health study. S Afr Med J 2009 May;99(5 Pt 2):367-373. [Medline: 19588800]

XSL•FO
RenderX

31. Diniz PRB, Ribeiro Sales FJ, de Araújo Novaes M. Providing telehealth services to a public primary care network: the experience of RedeNUTES in Pernambuco, Brazil. Telemed J E Health 2016 Aug;22(8):694-698. [doi: 10.1089/tmj.2015.0209] [Medline: 26953495]
32. Kuffer M, Vanhuysse S, Abascal A, et al. Data and urban poverty: detecting and characterising slums and deprived urban areas in low- and middle-income countries. In: Mustak S, Srivastava PK, Singh D, editors. Advanced Remote Sensing for Urban and Landscape Ecology: Springer; 2023:1-22. [doi: 10.1007/978-981-99-3006-7_1]
33. Naslund JA, Aschbrenner KA, Scherer EA, McHugo GJ, Marsch LA, Bartels SJ. Wearable devices and mobile technologies for supporting behavioral weight loss among people with serious mental illness. Psychiatry Res 2016 Oct 30;244:139-144. [doi: 10.1016/j.psychres.2016.06.056] [Medline: 27479104]
34. Lipschitz JM, Pike CK, Hogan TP, Murphy SA, Burdick KE. The engagement problem: a review of engagement with digital mental health interventions and recommendations for a path forward. Curr Treat Options Psychiatry 2023 Sep;10(3):119-135. [doi: 10.1007/s40501-023-00297-3] [Medline: 38390026]
35. Azad R, Fahmi R, Shrestha S, et al. Prevalence and risk factors of postpartum depression within one year after birth in urban slums of Dhaka, Bangladesh. PLoS One 2019;14(5):e0215735. [doi: 10.1371/journal.pone.0215735] [Medline: 31048832]
36. Online doctor consultation & telehealth in BD. DocTime. URL: https://doctime.net/doctors?online_now=1 [accessed 2025-05-25]
37. Maya Apa - Help at hand - Bangladesh's first app by women for women. BRAC. URL: https://www.brac.net/latest-news/item/691-maya-apa-help-at-hand-bangladesh-s-first-app-by-women-for-women [accessed 2025-05-25]
38. Pope C, Ziebland S, Mays N. Qualitative research in health care. Analysing qualitative data. BMJ 2000 Jan 8;320(7227):114-116. [doi: 10.1136/bmj.320.7227.114] [Medline: 10625273]
39. Ahmed SK. The pillars of trustworthiness in qualitative research. J Med Surg Public Health 2024 Apr;2:100051. [doi: 10.1016/j.glmedi.2024.100051]
40. Braun V, Clarke V. Toward good practice in thematic analysis: avoiding common problems and be(com)ing a knowing researcher. Int J Transgend Health 2023;24(1):1-6. [doi: 10.1080/26895269.2022.2129597] [Medline: 36713144]
41. Patel V. Mental health in low- and middle-income countries. Br Med Bull 2007;81-82:81-96. [doi: 10.1093/bmb/ldm010] [Medline: 17470476]
42. Moulaei K, Sheikhtaheri A, Fatehi F, Shanbehzadeh M, Bahaadinbeigy K. Patients' perspectives and preferences toward telemedicine versus in-person visits: a mixed-methods study on 1226 patients. BMC Med Inform Decis Mak 2023 Nov 15;23(1):261. [doi: 10.1186/s12911-023-02348-4] [Medline: 37968639]
43. Renn BN, Hoeft TJ, Lee HS, Bauer AM, Areán PA. Preference for in-person psychotherapy versus digital psychotherapy options for depression: survey of adults in the U.S. NPJ Digit Med 2019;2(1):6. [doi: 10.1038/s41746-019-0077-1] [Medline: 31304356]
44. Mohtar L, Badr N, Asmar MK, Bou-Orm I, Maalouf FT, Haddad PE. Adoption of tele health technology in mental and psychiatric services in Lebanon: a quantitative study. Discov Ment Health 2025 Mar 20;5(1):40. [doi: 10.1007/s44192-025-00169-7] [Medline: 40111657]
45. Nazari A, Garmaroudi G, Foroushani AR, Hosseinnia M. The effect of web-based educational interventions on mental health literacy, stigma and help-seeking intentions/attitudes in young people: systematic review and meta-analysis. BMC Psychiatry 2023 Sep 4;23(1):647. [doi: 10.1186/s12888-023-05143-7] [Medline: 37667229]
46. Naslund JA, Aschbrenner KA, Araya R, et al. Digital technology for treating and preventing mental disorders in low-income and middle-income countries: a narrative review of the literature. Lancet Psychiatry 2017 Jun;4(6):486-500. [doi: 10.1016/S2215-0366(17)30096-2] [Medline: 28433615]
47. Merchant R, Torous J, Rodriguez-Villa E, Naslund JA. Digital technology for management of severe mental disorders in low-income and middle-income countries. Curr Opin Psychiatry 2020 Sep;33(5):501-507. [doi: 10.1097/YCO.0000000000000626] [Medline: 32520747]
48. Torous J, Nebeker C. Navigating ethics in the digital age: introducing Connected and Open Research Ethics (CORE), a tool for researchers and institutional review boards. J Med Internet Res 2017 Feb 8;19(2):e38. [doi: 10.2196/jmir.6793] [Medline: 28179216]
49. Lupton D. Critical perspectives on digital health technologies. Sociol Compass 2014 Dec;8(12):1344-1359. [doi: 10.1111/soc4.12226]
50. Burchett HED, Kneale D, Blanchard L, Thomas J. When assessing generalisability, focusing on differences in population or setting alone is insufficient. Trials 2020 Mar 20;21(1):286. [doi: 10.1186/s13063-020-4178-6] [Medline: 32197623]
51. Palinkas LA, Horwitz SM, Green CA, Wisdom JP, Duan N, Hoagwood K. Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. Adm Policy Ment Health 2015 Sep;42(5):533-544. [doi: 10.1007/s10488-013-0528-y] [Medline: 24193818]
52. Cohen A, Naslund JA, Chang S, et al. Relapse prediction in schizophrenia with smartphone digital phenotyping during COVID-19: a prospective, three-site, two-country, longitudinal study. Schizophrenia (Heidelb) 2023 Jan 27;9(1):6. [doi: 10.1038/s41537-023-00332-5] [Medline: 36707524]

XSL•FO
RenderX

53. Kharono B, Kaggiah A, Mugo C, et al. Mobile technology access and use among youth in Nairobi, Kenya: implications for mobile health intervention design. Mhealth 2022;8:7. [doi: 10.21037/mhealth-21-23] [Medline: 35178438]
54. Gitahi N, Kiplagat A, Waruinge S, Lukoye A, Bukusi E. Implementing enhanced m-health mental health screening for adolescents living with HIV transitioning into adult care. Research Square. Preprint posted online on Aug 11, 2023. [doi: 10.21203/rs.3.rs-3178508/v1]

Abbreviations

EMMATAC: Enhanced mHealth mental health screening for adolescents living with HIV transitioning into adult care
FGD: focus group discussion
LMIC: low- and middle-income country
mHealth: mobile health
SMART-DAPPER: A Sequential, Multiple Assignment Randomized Trial (SMART) for Non-Specialist Treatment of Common Mental Disorders in Kenya: Leveraging the Depression and Primary-Care Partnership for Effectiveness-Implementation Research (DAPPER) Project
SMD: serious mental disorder
TRANSFORM: Transforming Access to Care for Serious Mental Disorders in Slums

Edited by A Mavragani; submitted 21.08.24; peer-reviewed by D Hidalgo-Mazzei, K Sarma; revised version received 19.06.25; accepted 23.06.25; published 20.11.25.

Please cite as:
Alam N, Giacco D, Siddiqi B, Singh S, Jilka S
Smartphones and Mental Health Awareness and Utilization in a Low-Income Urban Community: Focus Group Study
JMIR Form Res 2025;9:e65650
URL: https://formative.jmir.org/2025/1/e65650
doi:10.2196/65650

© Nadia Alam, Domenico Giacco, Bulbul Siddiqi, Swaran Singh, Sagar Jilka. Originally published in JMIR Formative Research (https://formative.jmir.org), 20.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Web-Based AI-Driven Virtual Patient Simulator Versus Actor-Based Simulation for Teaching Consultation Skills: Multicenter Randomized Crossover Study

Edward G Tyrrell[1], BMBS, BMedSci, MSc; Sardip K Sandhu[1], MPH; Kathryn Berry[2], BM; Suzan F Ghannam[1], MSc; Sarah A Lewis[1], MSc, PhD; Daniel Crowfoot[1], MBChB, MMedSci; Gurvinder Singh Sahota[1], MBChB, MMedSci; Julie Carson[1], BSc, BMBS, PhD; Emma E Wilson[1], MA, MPH, PhD; Jaspal Taggar[1], MBChB, MSc, PhD

[1]Primary Care Education Unit, School of Medicine, University of Nottingham, Room C34, Queen's Medical Centre, Nottinghamshire, United Kingdom
[2]Lincoln Medical School, University of Lincoln, Lincoln, United Kingdom

Corresponding Author:
Edward G Tyrrell, BMBS, BMedSci, MSc
Primary Care Education Unit, School of Medicine, University of Nottingham, Room C34, Queen's Medical Centre, Nottinghamshire, United Kingdom

Abstract

Background: There is a need to increase health care professional training capacity to meet global needs by 2030. Effective communication is essential for delivering safe and effective patient care. Artificial intelligence (AI) technologies may provide a solution. However, evidence for high-fidelity virtual patient simulators using unrestricted 2-way verbal conversation for communication skills training is lacking.

Objective: This study aims to compare a fully automated AI-driven voice recognition-based virtual patient simulator with traditional actor-based consultation skills simulated training in undergraduate medical students for differences in developing self-rated communication skills, student satisfaction scores, and direct cost comparison.

Methods: Using an open-label randomized crossover design, a single web-based AI-driven communication skills training session (AI-CST) was compared with a single face-to-face actor-based consultation skills training session (AB-CST) in undergraduates at 2 UK medical schools. Offline total cohort recruitment was used, with an opt-out option. Pre-post intervention surveys using 10-point linear scales were used to derive outcomes. The primary outcome was the difference in self-reported attainment of communication skills between interventions. Secondary outcomes were differences in student satisfaction and the cost comparison of delivering both interventions.

Results: Of 396 students, 378 (95%) completed at least 1 survey. Both modalities significantly increased self-reported communication skills attainment (AI-CST: mean difference 1.14, 95% CI 0.97 - 1.32 points; AB-CST: mean difference 1.50, 95% CI 1.35 - 1.66 points; both $P<.001$). Attainment increase was lower for AI-CST than AB-CST (by mean difference 0.36, 95% CI −0.60 to −0.13 points; $P=.04$). Overall satisfaction was lower for AI-CST than AB-CST (8.09 vs 9.21; mean difference −1.13, 95% CI −1.33 to −0.92 for AI-CST vs AB-CST; $P<.001$). The estimated costs of AI-CST and AB-CST were £33.48 (US $42.22) and £61.75 (US $77.87) per student, respectively.

Conclusions: AI-CST and AB-CST were both effective at improving self-reported communication skills attainment, but AI-CST was slightly inferior to AB-CST. Student satisfaction was significantly greater for AB-CST. Costs of AI-CST were substantially lower than AB-CST. AI-CST may provide a cost-effective opportunity to build training capacity for health care professionals.

(JMIR Form Res 2025;9:e71667) doi:10.2196/71667

KEYWORDS

communication skills; simulation; clinical education; medical education; technology enhanced learning

Introduction

There is a global shortage of health workers. The World Health Organization estimated this shortfall at 15 million in 2020 [1]. The United Kingdom government plans to train greater numbers of health care professionals [2], including doubling the number of medical training places by 2032 [3]. This has significant capacity resource implications [4], especially within clinical environments [5], where lack of available space [6] and supervisor capacity [7] are important barriers to increasing trainee numbers.

Communication is a crucial skill for all health care professionals in order to provide safe, effective, and patient-centered care [8]. Good practitioner communication helps improve patient

XSL•FO
RenderX

well-being [9], shared decision-making, treatment compliance, health-promoting behaviors [10], patient satisfaction, and reduce health care litigation [11]. Therefore, building capacity for communication skills training will be required in parallel with the increasing number of health care professionals being trained.

Complex skills, such as communication, are most effectively learned through experiential methods such as facilitated simulation through role-playing with trained patient actors [12] or live clinical patient encounters with feedback and debriefing [13]. However, these approaches are resource-intensive on physical space, staffing, materials, organization, and timetabling [14,15]. In addition, differences in actor performance or patient presentation manifest as variability of student experiences and assessment standardization [16].

Novel technologies such as virtual patient simulators (VPSs, defined as interactive computer simulations of real-life clinical scenarios for the purpose of health professions training, education, or assessment [17]) may provide solutions to increasing training capacity while reducing variability through delivering scalable and asynchronous simulated learning [18-21].

Some evidence suggests VPSs can support health care professionals' communication skills development [22-27]. For example, typed questions and responses via a virtual patient showed no difference from traditional video scenario-based training for improving suicide risk assessment practices in medical students [23]. VPSs involving unilateral conversation simulators, either with options to select from limited predefined responses to a series of virtual patient questions [25], or learners responding orally to a recorded vignette [22], as well as virtual avatars controlled by a live person [26,27] have also been tested, with varied results. However, the majority of VPS research has used lower fidelity approaches, such as typed text or selective responses [19], or has been undertaken in other contexts such as clinical reasoning.

There remains a paucity of high-quality comparative studies [20,28], particularly studies including financial comparisons [19], to inform the translation of these technologies for developing communication skills into education curricula. Specifically, there has been no research investigating the effectiveness of VPSs using unrestricted 2-way oral conversations. No previous studies have assessed the independent benefit of VPSs for developing communication skills, nor compared them head-to-head with traditional teaching approaches for economic impact, as outlined by previous review articles [19,20].

SimConverse is a web-based artificial intelligence (AI)-driven voice recognition platform designed to develop communication skills in health care professionals. It uses a VPS that interacts with learners through natural language in unrestricted 2-way oral conversations. While there is some evidence of its development and feasibility as an education tool [29], SimConverse has yet to be tested at scale in medical education.

The University of Nottingham (UoN) medical degree has a large undergraduate program with communication skills embedded as a spiral theme throughout its primary care modules, primarily delivered using a conventional approach of actor-based simulation [30]. Thus, providing an opportunity to robustly investigate the implementation of novel technologies for developing communication skills.

Consequently, this study aimed to was to compare an AI-driven voice recognition VPS (SimConverse) with the current standard delivery of communication skills training in undergraduate medical students [30]. We aimed to test differences in self-rated communication skills developed, student satisfaction, and conduct a direct cost comparison of delivering the 2 interventions. We hypothesized that there would be no differences in attainment of self-rated communication skills and student satisfaction between interventions, but that the VPS would have a lower delivery cost.

Methods

Design and Setting

A multicenter, open-label randomized crossover study was undertaken during a 1-week primary care attachment, early primary care delivered to 3rd year undergraduate medical students at 2 higher education institutions (UoN and University of Lincoln [UoL]) using the same curriculum.

The students involved were all at the equivalent mid-curricula level of study, having the same knowledge base and having completed the same gateway assessments. Most of their prior training had been classroom-based, other than approximately 2 weeks of clinical training as single or half-day visits spread across 2 years. The attachment on which this study was based was a short bridging attachment to support their transition into full-time clinical training. Students involved were therefore relative novices to communication skills in a clinical context, with this being their first block-based placement in any setting. Due to the comparatively junior stage of students, the focus was on improving relatively basic communication skills, such as history taking, rather than more complex skills such as breaking bad news or dealing with upset or angry patients. As undergraduate students, all are required to use computers and the internet on a regular basis for study and assessment purposes. Therefore, basic computer and internet literacy was assumed for all participants.

Key educational aims of the attachment were to build upon basic skills of communication to communicate effectively with patients in primary care, and to use targeted history taking to aid diagnosis and clinical decision-making as part of a medical consultation.

The study was undertaken using a predetermined protocol and statistical analysis plan and has been reported in line with the CONSORT-EHEALTH V1.6 checklist in the study by Eysenbach and CONSORT-EHEALTH Group [31].

Participants

Eligible participants were all undergraduate medical students enrolled at UoN or UoL, timetabled to undertake the early primary care attachment. All eligible students were initially included in the study (total cohort recruitment) and were encouraged to take part but were able to opt out. This approach

XSL•FO
RenderX

was taken as there were no existing effect sizes to use in a power calculation.

Randomization

We randomized participants into 2 cohorts (A and B) using a computerized 1:1 random number generator. Participants were subsequently assigned to groups by an administrator who was independent of the study team. The deployment of interventions was unblinded to participants and investigators.

Participants in cohort A were allocated to receive artificial-intelligence communication skills training (AI-CST) first, then actor-based communication skills training (AB-CST) second, with a washout period of 26 hours between the 2 interventions due to timetabling constraints of a total cohort recruitment approach. Cohort B received AB-CST first, then AI-CST second, in parallel to cohort A. Both cohorts completed both interventions.

Interventions

Both interventions used the same intended learning outcomes (ILOs), namely to: (1) demonstrate the key components of taking a medical history from a patient; (2) listen effectively and respond to a patient's cues regarding their presenting complaint; (3) begin to integrate biological, psychological, and social factors in assessing a patient's clinical history; and (4) demonstrate early hypothetico-deductive clinical reasoning in taking a focused medical history of the patient's presenting complaint.

Although the 2 interventions outlined below have differences, we wanted to compare a new alternative learning methodology (asynchronous virtual learning using a VPS) with the current standard delivery of communication skills training across medical schools (small group actor-based simulation) [30]. We assessed the ability of both "packages of learning" in their individual entireties for developing communication skills. The ILOs across both interventions were identical. Participants were aware that AI-CST was a new technology and, therefore, the "intervention of interest.

AI-CST

Participants completed a session of asynchronous unfacilitated AI-CST using the SimConverse platform. The session allowed participants to consult with a virtual patient, receive feedback on their communication, then repeat the consultation up to twice more. In total, participants were provided 3 virtual patients (each with 2 compulsory attempts and a third optional attempt) across the education session. Participants were given a nominal 3 hours in their timetable to complete training. The virtual patients were predeveloped by the research team and designed to be of a comparable level of complexity to those used in the AB-CST intervention. Further details of cases are given in Multimedia Appendix 1.

SimConverse is a purpose-built web-based artificial-intelligence VPS platform designed to develop communication skills in health care professionals [32], accessed via any computer with a microphone and speaker. It allows students to conduct open virtual health care consultations through fully automated high-fidelity audio conversations in real time using voice recognition. The VPS speaks back in real time with text-to-voice. Conversations are anchored using a virtual patient static image. Having completed a consultation, students receive immediate objective grading and personalized feedback on their communication performance using a predeveloped rubric. More detailed product design processes are given in Multimedia Appendix 2.

The SimConverse platform allows faculty to develop their own feedback rubrics or use preexisting ones, based on individual requirements. The SimConverse rubric system and authoring tooling were developed by the company with a multinational group of clinical and educational experts to ensure alignment with best practice marking and feedback. We used a preexisting rubric on the SimConverse platform that was adapted to meet our session ILOs. We tested the rubrics through multiple rounds of initial authoring and previewing in the SimConverse platform as educators, prior to deployment. The exact wording of the rubrics is commercially sensitive data. However, the areas covered in our feedback rubric were the introduction to the consultation, exploration of the history of the presenting complaint, past medical history, drug history, social history, family history, giving a differential diagnosis, giving next steps, and closure of the consultation.

AB-CST

Participants also completed traditional AB-CST, representing usual practice as the comparator intervention. AB-CST was delivered as face-to-face small group teaching with 5‑6 participants, a trained actor (the patient), and a trained clinical educator. Each participant completed a different predeveloped medical consultation scenario with a trained actor, lasting 10‑15 minutes. Feedback about communication was provided during the session by the clinical educator, other participants, and the actor.

Overall, participants received a 3-hour session during which they performed 1 consultation each, while observing and discussing feedback for the remaining 5 consultations. Case material and a session outline are given in Multimedia Appendix 1.

Participant Characteristics

We ascertained details of participant age, gender, widening participation status, and presence of a support plan (covering students with disabilities, long-term medical conditions, including mental health) or specific learning differences, from existing university records, matched to study data using unique participant identifiers. Widening participation refers to a UK initiative designed to increase underrepresented groups in higher education, such as those from low-income backgrounds, first-generation students, ethnic minority students, and students with disabilities. At both UoN and UoL, widening participation students undertake a specific foundation year prior to entering the medical course, which was our marker for their widening participation status.

Primary Outcome

The primary outcome was the self-reported attainment of communication skills determined using a composite measure

XSL•FO
RenderX

of communication assessed through pre-post surveys for both AI-CST and AB-CST.

During trial development, student engagement made clear the need to consider feedback fatigue as we deployed a measurement instrument [33]. A review of literature did not identify any succinct and externally validated tools for measuring the attainment of communication skills to enable comparisons of pre-post intervention scores while simultaneously maximizing participant engagement. Therefore, we developed an in-house survey for measuring self-reported attainment of communication skills, anchored to the Calgary-Cambridge model [34], the communication model taught at UoN and UoL.

The Calgary-Cambridge model is focused on marrying content and process within medical consultations, not excluding either the content of the consultation or the processes of communication at the expense of the other. The framework of its tasks and objectives includes initiating the session, gathering information, physical examination, explanation and planning, and closing the session. Throughout this process, run pillars of providing structure and flow while also building relationships, rapport, and involving the patient (Multimedia Appendix 3). The objectives in the model serve as subheadings to support conceptualization while helping learners organize and apply numerous communication process skills [35]. There are many such skills that might be included, depending on the individual consultation. These might include question style: open to closed, attentive listening, responding to cues, time-framing, summarizing, empathizing, signposting, parking, or many more (Skills for Communicating with Patients [36]). There is evidence of the effectiveness of the Calgary-Cambridge model in assessing communication skills in medical students [37].

An 11-item survey (Figure 1) was developed using an iterative process by 6 of the research team with extensive medical education experience. How the survey maps to the Calgary-Cambridge model is outlined in Table S1 in Multimedia Appendix 4. Usability and technical functionality were piloted with 5 students, using the method outlined by Arundel [38]. Ten-point linear scales were used to maximize discrimination of pre-post intervention survey scores within individuals [39,40]. Only end points in each survey item were anchored to interpretations, allowing each to be treated as an interval scale, thus allowing parametric statistical hypotheses testing [39,40]. The survey was housed on the Microsoft Forms platform, allowing responses to be downloaded as CSV files for analysis without the need for any manual data entry. No cookies or IP addresses were recorded. Participants were identifiable by username details supporting matching of pre-post surveys and demographic data. Participants were incentivized by being offered entry into a prize draw to win 1 of 3 £100 (US $126.10) shopping vouchers for completing pre- and postsession surveys for both interventions.

XSL•FO
RenderX

Figure 1. Survey questions for undergraduate medical students to identify self-reported communication skill development pre- and posteducational interventions and satisfaction levels posteducational intervention.

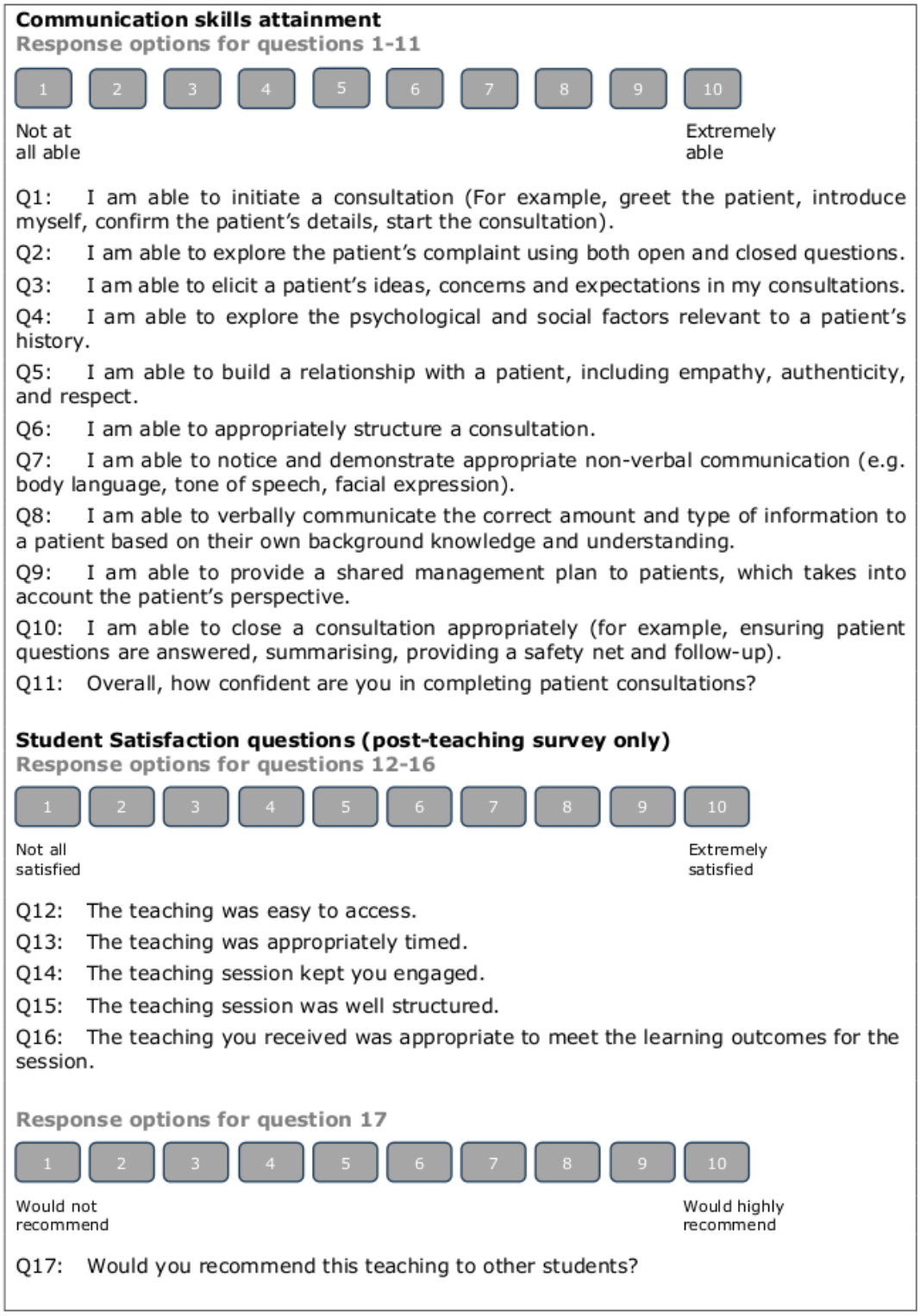

For both interventions, participants were asked at the start and end of the session to complete online surveys. Consent was provided by advancing beyond the front screen, which explained the survey with wording preagreed with the ethics committee, including an option to opt out. Participants were provided with detailed study information, including the purpose of the study, matching with institutionally held demographic data, data storage arrangements, and details of the investigator. Preintervention surveys consisted of 2 screens, 1 for consent and the other containing the 11 items. Postintervention surveys contained a third screen with a 6-item satisfaction survey. All survey items were mandatory, although students could opt out at any point by closing the survey. For AB-CST, the surveys were accessed via the provision of QR codes. For AI-CST, links to the appropriate pre-post intervention surveys were built into

XSL•FO
RenderX

the SimConverse platform. All participants were sent 2 email reminders to maximize completion of postsession surveys.

Only completed surveys were analyzed. Pre-post changes in aggregated mean scores for both interventions were used as the primary outcome measure and compared across interventions.

Secondary Outcomes

We measured participant satisfaction with an additional 6 items in postintervention surveys (Figure 1), using 10-point linear scales. Aggregated mean scores for both interventions were used as the composite outcome for satisfaction and compared across interventions.

We also collected data enabling the comparison of steady-state direct costs of delivering the interventions [41], further outlined below.

Statistical Analyses

Participant characteristics were summarized as numbers (percentages) for categorical data and mean (SD) for continuous data. Demographic characteristics of cohorts A and B were compared to assess the robustness of the randomization process.

Primary Analyses

Mean changes in the pre-post intervention ratings were calculated and compared using independent *t* tests (2-tailed). Differences in the aggregated mean scores between the 2 interventions were then compared using independent *t* tests for self-rated communication skill attainment. To avoid excessively restricting the analyses of this within-curriculum randomized study of educational interventions, we included all survey results a priori, regardless of whether individual students had completed both pre- and postsession surveys.

We conducted a sensitivity analysis including only data from matched pre-post intervention surveys, while also excluding individuals with first intervention postsession surveys completed after the second intervention had been initiated.

Secondary Analyses

To test for the presence of carryover effects, preintervention scores for the second intervention were compared between cohorts A and B. In the presence of carryover, data from the first period before crossover were used to compare treatments in a parallel group design. Differences in aggregated mean scores for satisfaction were also compared between the 2 interventions using independent *t* tests. Analyses were completed using SPSS (version 29; IBM Corp) and $P<.05$ was used to indicate statistical significance.

Cost Comparison Analyses

The direct steady-state costs of using AI-CST and AB-CST were calculated and compared as the cost per participant taught. For AB-CST, costs included were £200 (US $252.20) per half-day of GP tutor time and £115 (US $145.02) per half-day of actor time. For the full cohort, 54 sessions were delivered. Although exact costs for things such as room hire, administrative time, and electricity costs were difficult to define and therefore not included, 10% of the total staff costs were added as approximate overheads. AI-CST costs were calculated on usage credits (the completion and feedback of 1 attempt at a consultation) and based on 9 allocated credits per student, irrespective of completion rates. SimConverse is a private commercial enterprise and therefore may enter into different licensing agreements with different organizations depending on the number of credits purchased. Therefore, the costs within this study relate to the purchase plan agreed with UoN and costs may vary for other organizations. Further information can be requested directly from the company. For both interventions, nonrecurring setup costs, such as actor and tutor training, scenario development, and technology infrastructure, were not included in the analyses.

Ethical Considerations

The study was approved by the UoN Faculty of Medicine & Health Sciences Research Ethics Committee (approval number FMHS 43 - 1023) on November 20, 2023. Informed consent was obtained from all participants, as previously outlined. All data were collected between January 22 and February 18, 2024. Privacy and confidentiality of participants' data were maintained at all times. Potential incentives for participants have been outlined under the Primary Outcome section.

Results

Response Rates

In total, 396 (308 from Nottingham and 88 from Lincoln) third-year students were enrolled and eligible to participate at both study sites. Of these, 378 (95%) completed at least 1 survey: 296 (96%) from Nottingham and 82 (93%) from Lincoln. Participant flow and survey responses are provided in Figure 2.

XSL•FO
RenderX

Figure 2. Participant flow and survey responses during the study period for a crossover study examining self-reported communication skill development pre and post both artificial intelligence–based and actor-based communication skills training. AB-CST: actor-based consultation skills training session; AI-CST: artificial intelligence–driven communication skills training session.

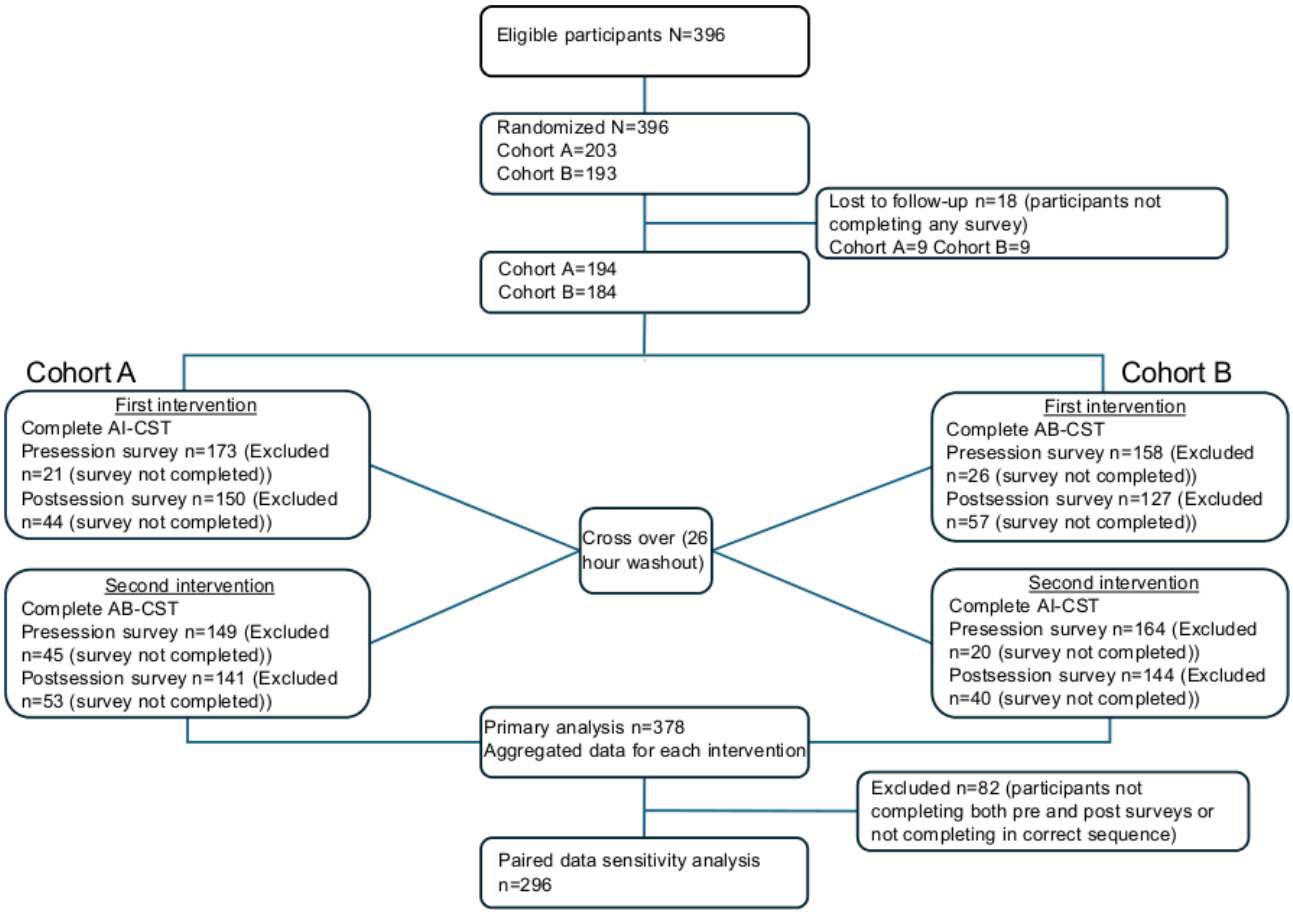

For AI-CST, 85% (337/396) and 74% (294/396) participants completed pre- and postintervention surveys, respectively. For AB-CST, this was 78% (307/396) and 68% (268/396) of participants, respectively. For both interventions, 87% (294/337 for AI-CST and 268/307 for AB-CST) of those completing the preintervention survey also completed the postintervention survey.

There were no differences in the characteristics of participants who completed at least 1 survey between the 2 randomized cohorts (Table 1).

Table . Characteristics of participants completing at least 1 survey response for self-reported communication skill development pre- or post–artificial intelligence-based or actor-based communication skills training, by randomized group.

Characteristics	Cohort A (n=194)	Cohort B (n=184)
Age (years), mean (SD)	21.69 (1.72)	21.65 (1.44)
Sex, n (%)		
Male	57 (29.4)	55 (29.9)
Female	137 (70.6)	129 (70.1)
Widening participation, n (%)		
Yes	18 (9.3)	21 (11.4)
No	176 (90.7)	163 (88.6)
Support plan, n (%)		
Yes	34 (17.5)	31 (16.8)
No	160 (82.5)	153 (83.2)
Education institution, n (%)		
University of Nottingham	155 (79.9)	141 (76.6)
University of Lincoln	39 (20.1)	43 (23.4)

XSL•FO
RenderX

The time spent on AB-CST had a fixed length of 3 hours due to the timetabled nature of an in-person taught session. AI-CST was asynchronous and unsupervised. It was nominally timetabled for 3 hours, but the mean time participants spent completing AI-CST was 1 hour 53 (SD 45) minutes. Attrition rates for both interventions are given in Multimedia Appendix 5. For AB-CST, 50 participants did not complete the intervention due to absence from teaching, meaning 87% (346/396) completed the intervention as intended. For AI-CST, 69% (272/396) completed the intervention as intended (at least 6 total VPS conversations), with 88% (348/396) completing at least 1 conversation.

Primary Outcome

Both interventions were found to improve the self-reported attainment of communication skills (Table 2). The pre-post aggregated mean difference scores from AI-CST increased by 1.14 (95% CI 0.97-1.32) points and from AB-CST by 1.50 (95% CI 1.35-1.66) points (both $P<.001$). The comparison of interventions found that the difference in aggregated mean difference scores was 0.36 points lower for AI-CST than AB-CST (95% CI −0.60 to −0.13; $P=.04$).

XSL•FO
RenderX

Table . Comparison of self-reported communication skills attainment using survey scores pre- and postintervention for both artificial intelligence–based and actor-based communication skills training modalities.

Questions	Participant scores for each question in the communication skills surveys											Primary outcome
	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	
AI-CST[a] (all responses)												
Mean (SD) baseline (n=337)	8.44 (1.38)	6.87 (1.34)	6.90 (1.45)	6.30 (1.39)	7.35 (1.39)	6.27 (1.60)	7.42 (1.43)	6.37 (1.53)	5.32 (1.84)	6.23 (1.77)	6.25 (1.53)	6.70 (1.16)
Mean (SD) postbaseline (n=294)	9 (1.12)	8.14 (1.21)	8.17 (1.31)	7.63 (1.47)	8.09 (1.50)	7.76 (1.33)	7.60 (1.81)	7.57 (1.52)	6.96 (1.74)	7.74 (1.55)	7.63 (1.24)	7.84 (1.08)
Mean difference (95% CI) from baseline	0.55 (0.34-0.83)	1.27 (1.11-1.59)	1.27 (1.09-1.49)	1.33 (1.14-1.56)	0.74 (0.63-1.18)	1.49 (1.28-1.85)	0.18 (-0.10-0.45)	1.20 (0.96-1.43)	1.64 (1.48-2.17)	1.51 (1.25-1.77)	1.38 (1.16-1.60)	1.14 (0.97-1.32)
P value[b]	<.001	<.001	<.001	<.001	<.001	<.001	.18	<.001	<.001	<.001	<.001	<.001
AB-CST[c] (all responses)												
Mean (SD) baseline (n=307)	8 (1.41)	6.45 (1.23)	6.52 (1.33)	6.14 (1.29)	7.07 (1.30)	5.91 (1.41)	6.93 (1.41)	5.96 (1.46)	4.98 (1.83)	6.02 (1.65)	5.85 (1.37)	6.35 (1.01)
Mean (SD) postbaseline (n=268)	8.88 (1.04)	8.03 (1.05)	8.09 (1.10)	7.66 (1.28)	8.31 (1.14)	7.70 (1.19)	8.15 (1.18)	7.54 (1.26)	6.79 (1.62)	7.63 (1.47)	7.63 (1.04)	7.86 (0.93)
Mean difference (95% CI) from baseline	0.88 (0.65-1.08)	1.58 (1.39-1.78)	1.57 (1.28-1.73)	1.52 (1.25-1.72)	1.24 (0.95-1.42)	1.80 (1.57-2.02)	1.20 (0.98-1.42)	1.62 (1.39-1.84)	1.83 (1.53-2.13)	1.64 (1.37-1.90)	1.78 (1.56-1.99)	1.50 (1.35-1.66)
P value[d]	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001
Comparison of AI-CST and AB-CST mean differences												
Mean difference (95% CI) AI-CST	0.55 (0.34-0.83)	1.27 (1.11-1.59)	1.27 (1.09-1.49)	1.33 (1.14-1.56)	0.74 (0.63-1.18)	1.49 (1.28-1.85)	0.18 (-0.10-0.45)	1.20 (0.96-1.43)	1.64 (1.48-2.17)	1.51 (1.25-1.77)	1.38 (1.16-1.60)	1.14 (0.97-1.32)
Mean difference (95% CI) AB-CST	0.88 (0.65-1.08)	1.58 (1.39-1.78)	1.57 (1.28-1.73)	1.52 (1.25-1.72)	1.24 (0.95-1.42)	1.80 (1.57-2.02)	1.20 (0.98-1.42)	1.62 (1.39-1.84)	1.83 (1.53-2.13)	1.64 (1.37-1.90)	1.78 (1.56-1.99)	1.50 (1.35-1.66)
Difference (95% CI) in differences	−0.33 (−0.61 to −0.05)	−0.31 (−0.58 to 0.03)	−0.30 (−0.58 to 0.01)	−0.16 (−0.49 to 0.12)	−0.42 (−0.79 to −0.19)	−0.31 (−0.61 to 0.01)	−1.02 (−1.38 to −0.71)	−0.42 (−0.71 to 0.07)	−0.19 (−0.23 to 0.56)	−0.13 (−0.26 to 0.46)	−0.40 (−0.68 to 0.10)	−0.36 (−0.60 to −0.13)

XSL•FO
RenderX

Questions	Participant scores for each question in the communication skills surveys											Primary outcome
	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	
P value[d]	.02	.01	.002	.04	<.001	.06	<.001	0.001	0.09	0.98	<.001	.04

[a]AI-CST: artificial-intelligence communication skills training.

[b]Paired *t* test (2-tailed).

[c]AB-CST: actor-based communication skills training.

[d]Independent *t* test (2-tailed).

Sensitivity analyses results (Multimedia Appendix 6) using only matched pre-post surveys and excluding participants who completed the first postintervention survey after the second intervention initiation found no change to the primary findings.

Secondary Outcomes

Student Satisfaction

Satisfaction scores were high for both interventions (Table 3) but significantly lower for AI-CST than AB-CST, across all questions. Aggregated mean scores were 8.09 (SD 1.51) and 9.21 (SD 0.92) for AI-CST and AB-CST, respectively (aggregated mean 1.13 points lower, 95% CI −1.33 to −0.92 for AI-CST than AB-CST; *P*<.001).

Table . Student satisfaction scores and comparison for artificial intelligence-based and actor-based communication skills training modalities.

Questions	Participant scores for each question in the student satisfaction surveys						Aggregated mean (95% CI)
	Q12 Easy to access	Q13 Appropriately timed	Q14 Kept you engaged	Q15 Well structured	Q16 Met learning outcomes	Q17 Would recommend to others	
AI-CST[a] mean (SD) score	8.76 (1.34)	8.08 (1.78)	7.45 (2.17)	7.93 (1.76)	8.26 (1.57)	8.03 (1.88)	8.09 (1.51)
AB-CST[b] mean (SD) score	9.26 (1.02)	9.03 (1.28)	9.11 (1.12)	9.26 (1.00)	9.28 (0.96)	9.34 (0.95)	9.21 (0.92)
Mean difference (95% CI)	−0.5 (−0.70 to −0.13)	−0.95 (−1.20 to −0.69)	−1.66 (−1.94 to −1.38)	−1.33 (−1.57 to −1.10)	−1.02 (−1.23 to −0.81)	−1.31 (−1.55 to −1.06)	−1.13 (−1.33 to −0.92)
P value[c]	<.001	<.001	<.001	<.001	<.001	<.001	<.001

[a]AI-CST: artificial-intelligence communication skills training.

[b]AB-CST: actor-based communication skills training.

[c]Independent *t* test (2-tailed).

Direct Costs

The steady-state direct cost of delivering AI-CST (£33.48; US $42.22 per student) was substantially lower than for AB-CST (£61.75; US $77.87 per student).

Carryover Effect Between Interventions

Cohort analyses suggested no differential carryover effect between the interventions (Multimedia Appendix 7). The comparison of preintervention mean scores within both cohorts A and B, prior to the second intervention, found no significant difference between groups (Cohort A: pre-second intervention mean score 6.26, SD 0.86; Cohort B: pre-second intervention mean score 6.62, SD 0.83; mean difference 0.35, 95% CI −1.11 to 0.40; *P*=.34).

Discussion

Principal Findings

This is the first randomized study comparing the self-reported effectiveness and direct costs of a VPS using unrestricted 2-way oral conversations with traditional AB-CST for attaining communication skills in undergraduate medical students. Although both interventions showed improvements in self-rated consultation skills and confidence among students, the magnitude of improvement was lower for AI-CST than AB-CST (mean improvement 1.14, 95% CI 0.97-1.32 for AI-CST vs 1.50, 95% CI 1.35-1.66 for AB-CST, respectively), just reaching statistical significance (*P*=.04). Student satisfaction was lower for AI-CST than AB-CST (aggregated mean score 1.13 points lower, 95% CI –1.33 to –0.92; *P*<.001), although satisfaction for both interventions was generally high. AI-CST was approximately half the cost of AB-CST.

XSL•FO
RenderX

Findings in Context

Simulation-based learning is effective in clinical training [42-45], while allowing for skills development without compromising patient safety [46]. The slightly larger benefit and higher satisfaction we demonstrated from AB-CST may link to medical student perspectives, that patient interactions supporting development of interpersonal skills and understanding of patient-centered care are vital to them [47]. While AI-CST does include elements of this, we surmise that the richness of opportunities and feedback linked to interpersonal skills and understanding of patient-centered care would be richer from AB-CST than from VPS interactions.

Previous studies have examined VPSs for developing clinical reasoning in medical students. These were small and investigated text-based interfaces rather than verbal communication [48-51]. While they found improvements in student knowledge of clinical reasoning, they did not investigate communication skills development.

A series of small US-based randomized trials has investigated virtual patients to train undergraduate medical students in communication skills. However, most used scripted textual question responses and not higher fidelity voice recognition or 2-way verbal conversations as encountered in live clinical environments [22,23,52].

Foster et al [23] showed that a single online-based virtual patient simulation was as effective as video-based teaching for training 67 medical students in conducting a suicide risk assessment. Similar to our findings, the VPS scored lower in student satisfaction than the video-based teaching. Another trial of 99 medical students using video-based simulations with typed questions from a predefined list showed a virtual patient to be as effective at developing communication skills in medical students as traditional learning. However, the comparison group was 'education as normal', with no distinct comparison intervention [52]. In contrast, a trial of 84 medical students showed that a VPS with scripted responses to limited verbal student questions was less effective at developing nonverbal communication skills than an actor-based simulation with the same scripted responses [22].

One study investigated an AI-developed voice-activated VPS using natural language for delivering feedback about medical student consultations. The VPS was as effective as humans at marking and providing feedback on performance, using a specified marking rubric [24], but its ability to improve communication skills was not examined.

The largest study by Kron et al [25], of 421 medical students, found that a VPS was more effective at developing communication skills than a computer-based training package using text, image, and video. Of note, the VPS, although using spoken language, only allowed students to select from 3 predefined responses during conversations, limiting the fidelity of the VPS to mimic live clinical patient encounters. In addition, the comparator in this study was atypical of usual practice for developing communication skills and did not allow students to practice the skills they had learned with any sort of patient interaction, human or otherwise. This may explain the contrast with the results shown in our study.

One of the challenges for technology-enhanced learning in communication skills training is its ability to support the development of interpersonal skills and nonverbal communication [53,54]. Guetterman et al [55], in a further analysis of data from the Kron study outlined above, reported positive impacts from the VPS on nonverbal cues demonstrated by students. However, the technology was limited to detecting simple movements, such as nodding, smiling, and eyebrow raising, and may overlook other nonverbal cues and expressions of empathy.

Other studies using VPSs with scripted textual responses have indicated their potential benefits for training students in shared decision-making [56], improving learner confidence in patient interactions [57], improving objective structured clinical examination performance [58], and that they evoke similar emotional responses when breaking bad news to real patient interactions [59].

Strengths and Limitations

To our knowledge, this is the first global within-curricula study to investigate the self-reported effectiveness of a VPS using unrestricted 2-way verbal conversation for developing communication skills in undergraduate medical students. A strength was the randomized crossover design and pre-post assessment of outcomes in the study population, enabling the control of confounders and differences in scenario complexity of interventions. The large multicenter design and high response rates increase the generalizability of findings to undergraduate medical education. Uniquely, we calculated the attributable intervention costs, providing greater insight to education providers about the implementation of such interventions into routine practice.

We did not complete an a priori sample size calculation as there were no outcome data for effectiveness or effect size across interventions. However, we conducted a post hoc power calculation from our results, suggesting our study had 99.8% power to detect differences between the interventions at the 5% significance level. While there was a short wash-out period in the crossover design of this study, our secondary analyses were reassuring, suggesting there was no carryover effect, and our primary analyses were valid.

There were some limitations. The deployment of interventions was unblinded to participants and investigators, with outcomes self-reported, thus introducing the potential for responder and verification biases within findings. Objective outcomes, such as communication skills attainment from summative assessments, would have strengthened the findings. It is plausible that an element of the Dunning-Kruger effect may have impacted our findings, with low ability students overestimating their competence, and more able students underestimating theirs [60]. This would have impacted our results toward the null, resulting in us underestimating the actual difference in effect size between the 2 interventions. The pre-post survey design of our study should control for within-individual biases and therefore adjust for this. However,

XSL•FO
RenderX

within medical education studies that use self-reporting instruments, it may be difficult to completely remove this effect.

We were able to report adherence rates to both interventions. However, we were not able to establish this at a participant level and therefore could not examine an association between this and learner confidence. This would be a relevant consideration for future research.

The single-session exposure to interventions was short. Therefore, the incremental impact and sustained improvements in communication skills from longer exposures could not be determined. The setting of this trial was in a short primary care-based attachment lasting only 1 week. It has provided evidence that supports a "proof-of-concept," although with a limited outcome measure. Our findings support the next step to embed this technology into curricula over a sustained period, thus enabling assessment against summative education outcomes. However, this would paradoxically introduce other methodological challenges, such as controlling for the effects of learning and development of communication skills that may arise from other curricula opportunities beyond the use of such technology.

The statistical precision could have been improved by using only paired survey responses. However, we wanted to use all available data in our primary analyses to maximize study power and reflect the testing of a new educational intervention in a live 'real world' learning environment. Our sensitivity analyses included only the paired responses and showed no change from the primary findings. This study was conducted at a single time point during medical training, and we would caution against the extrapolation of findings to students at different stages of medical training.

Conclusions

AI-CST using a voice recognition–based VPS was effective in improving self-reported communication skills in undergraduate medical students. However, it was not as effective as AB-CST. Student satisfaction for AB-CST was significantly better than AI-CST. The cost of AI-CST was roughly half that of AB-CST.

AI-CST may therefore provide a cost-effective solution to building capacity for teaching communication in health care professionals. Further research is needed to determine how the sustained improvement of communication skills can be developed using AI-CST, and the testing of such technology with students at other stages of medical training and in wider professional groups.

Acknowledgments

The authors wish to thank Prof Claire Sharp for her support during the study. SimConverse had no role in the design, conduct, or reporting of the study, other than in providing the description of the product design processes.

Funding

The study was partly funded by the National Health Service England. The funder had no role in the design, conduct, or reporting of the study.

Data Availability

The data analyzed for this study are available in the University of Nottingham data repository, in accordance with ethics approval [61].

Conflicts of Interest

None declared.

Multimedia Appendix 1
Case Material for artificial intelligence–based and actor-based communication skills training interventions.
[DOCX File, 30 KB - formative_v9i1e71667_app1.docx]

Multimedia Appendix 2
SimConverse (artificial intelligence–based communication skills training) product design information.
[DOCX File, 27 KB - formative_v9i1e71667_app2.docx]

Multimedia Appendix 3
The Calgary Cambridge model (adapted from Kurtz et al, 2003).
[DOCX File, 59 KB - formative_v9i1e71667_app3.docx]

Multimedia Appendix 4
Mapping of survey instrument for measuring self-reported communication skill development pre- and posteducational interventions in undergraduate medical students to Calgary-Cambridge model.

XSL•FO
RenderX

[DOCX File, 18 KB - formative_v9i1e71667_app4.docx]

Multimedia Appendix 5
Attrition rates of participants during the study period for both AI-based and actor-based communication skills training modalities.
[DOCX File, 26 KB - formative_v9i1e71667_app5.docx]

Multimedia Appendix 6
Sensitivity analysis of comparison of self-reported communication skills attainment using survey scores pre and postintervention for both artificial intelligence-based and actor-based communication skills training modalities, including only data from matched pre-post intervention surveys, and excluding surveys completed out of sequence.
[DOCX File, 37 KB - formative_v9i1e71667_app6.docx]

Multimedia Appendix 7
Within cohort analyses for the carryover effect between artificial intelligence-based and actor-based communication skills training interventions in the crossover study design.
[DOCX File, 25 KB - formative_v9i1e71667_app7.docx]

Checklist 1
CONSORT eHEALTH checklist.
[PDF File, 2858 KB - formative_v9i1e71667_app8.pdf]

References

1. Boniol M, Kunjumen T, Nair TS, Siyam A, Campbell J, Diallo K. The global health workforce stock and distribution in 2020 and 2030: a threat to equity and “universal” health coverage? BMJ Glob Health 2022 Jun;7(6):e009316. [doi: 10.1136/bmjgh-2022-009316] [Medline: 35760437]
2. NHS long term workforce plan. : NHS England; 2023 URL: https://www.england.nhs.uk/wp-content/uploads/2023/06/nhs-long-term-workforce-plan-v1.2.pdf [accessed 2025-10-28]
3. 350 extra medical school places allocated in NHS training boost. UK Government. 2024 May 13. URL: https://www.gov.uk/government/news/350-extra-medical-school-places-allocated-in-nhs-training-boost [accessed 2025-10-28]
4. Geary U, McKee M, Petty-Saphon K. Mind the implementation gap: a systems analysis of the NHS long term workforce plan to increase the number of doctors trained in the UK raises many questions. Br Med Bull 2024 Jun 10;150(1):1-10. [doi: 10.1093/bmb/ldae002] [Medline: 38342988]
5. Addicott R, Maguire D, Honeyman M, Jabbal J. Workforce planning in the NHS. : The King’s Fund; 2015 URL: https://assets.kingsfund.org.uk/f/256914/x/12727ac5bb/workforce_planning_in_nhs_2015.pdf [accessed 2025-10-28]
6. Barber JRG, Park SE, Jensen K, et al. Facilitators and barriers to teaching undergraduate medical students in general practice. Med Educ 2019 Aug;53(8):778-787. [doi: 10.1111/medu.13882] [Medline: 31012131]
7. Lock FK, Carrieri D. Factors affecting the UK junior doctor workforce retention crisis: an integrative review. BMJ Open 2022 Mar 28;12(3):e059397. [doi: 10.1136/bmjopen-2021-059397] [Medline: 35351732]
8. Franco CDS, Franco RS, Lopes JMC, Severo M, Ferreira MA. Clinical communication skills and professionalism education are required from the beginning of medical training - a point of view of family physicians. BMC Med Educ 2018 Dec;18(1):43. [doi: 10.1186/s12909-018-1141-2] [Medline: 29558914]
9. Zachariae R, Pedersen CG, Jensen AB, Ehrnrooth E, Rossen PB, von der Maase H. Association of perceived physician communication style with patient satisfaction, distress, cancer-related self-efficacy, and perceived control over the disease. Br J Cancer 2003 Mar 10;88(5):658-665. [doi: 10.1038/sj.bjc.6600798] [Medline: 12618870]
10. Howick J, Moscrop A, Mebius A, et al. Effects of empathic and positive communication in healthcare consultations: a systematic review and meta-analysis. J R Soc Med 2018 Jul;111(7):240-252. [doi: 10.1177/0141076818769477] [Medline: 29672201]
11. Maguire P, Pitceathly C. Key communication skills and how to acquire them. BMJ 2002 Sep 28;325(7366):697-700. [doi: 10.1136/bmj.325.7366.697] [Medline: 12351365]
12. Lane C, Rollnick S. The use of simulated patients and role-play in communication skills training: a review of the literature to August 2005. Patient Educ Couns 2007 Jul;67(1-2):13-20. [doi: 10.1016/j.pec.2007.02.011] [Medline: 17493780]
13. Nagpal V, Philbin M, Yazdani M, Veerreddy P, Fish D, Reidy J. Effective goals-of-care conversations: from skills training to bedside. MedEdPORTAL 2021 Mar 10;17:11122. [doi: 10.15766/mep_2374-8265.11122] [Medline: 33768153]
14. Al Odhayani A, Ratnapalan S. Teaching communication skills. Can Fam Physician 2011 Oct;57(10):1216-1218. [Medline: 21998240]
15. Bosse HM, Nickel M, Huwendiek S, Schultz JH, Nikendei C. Cost-effectiveness of peer role play and standardized patients in undergraduate communication training. BMC Med Educ 2015 Oct 24;15(1):183. [doi: 10.1186/s12909-015-0468-1] [Medline: 26498479]

XSL•FO
RenderX

16. Lee J, Kim H, Kim KH, Jung D, Jowsey T, Webster CS. Effective virtual patient simulators for medical communication training: a systematic review. Med Educ 2020 Sep;54(9):786-795. [doi: 10.1111/medu.14152] [Medline: 32162355]
17. Kononowicz AA, Woodham LA, Edelbring S, et al. Virtual patient simulations in health professions education: systematic review and meta-analysis by the digital health education collaboration. J Med Internet Res 2019 Jul 2;21(7):e14676. [doi: 10.2196/14676] [Medline: 31267981]
18. Gordon M, Daniel M, Ajiboye A, et al. A scoping review of artificial intelligence in medical education: BEME guide no. 84. Med Teach 2024 Apr;46(4):446-470. [doi: 10.1080/0142159X.2024.2314198] [Medline: 38423127]
19. Kelly S, Smyth E, Murphy P, Pawlikowska T. A scoping review: virtual patients for communication skills in medical undergraduates. BMC Med Educ 2022 Jun 3;22(1):429. [doi: 10.1186/s12909-022-03474-9] [Medline: 35659213]
20. Kyaw BM, Posadzki P, Paddock S, Car J, Campbell J, Tudor Car L. Effectiveness of digital education on communication skills among medical students: systematic review and meta-analysis by the digital health education collaboration. J Med Internet Res 2019 Aug 27;21(8):e12967. [doi: 10.2196/12967] [Medline: 31456579]
21. Mitchell P. Harnessing digital technologies for workforce development, education and training: an overview. Br J Healthcare Assist 2022 Dec 2;16(12):577-583. [doi: 10.12968/bjha.2022.16.12.577]
22. Deladisma AM, Cohen M, Stevens A, et al. Do medical students respond empathetically to a virtual patient? Am J Surg 2007 Jun;193(6):756-760. [doi: 10.1016/j.amjsurg.2007.01.021] [Medline: 17512291]
23. Foster A, Chaudhary N, Murphy J, Lok B, Waller J, Buckley PF. The use of simulation to teach suicide risk assessment to health profession trainees-rationale, methodology, and a proof of concept demonstration with a virtual patient. Acad Psychiatry 2015 Dec;39(6):620-629. [doi: 10.1007/s40596-014-0185-9] [Medline: 25026950]
24. Maicher KR, Zimmerman L, Wilcox B, et al. Using virtual standardized patients to accurately assess information gathering skills in medical students. Med Teach 2019 Sep;41(9):1053-1059. [doi: 10.1080/0142159X.2019.1616683] [Medline: 31230496]
25. Kron FW, Fetters MD, Scerbo MW, et al. Using a computer simulation for teaching communication skills: a blinded multisite mixed methods randomized controlled trial. Patient Educ Couns 2017 Apr;100(4):748-759. [doi: 10.1016/j.pec.2016.10.024] [Medline: 27939846]
26. Banszki F, Beilby J, Quail M, Allen P, Brundage S, Spitalnick J. A clinical educator's experience using a virtual patient to teach communication and interpersonal skills. AJET 2018 Jul;34(3). [doi: 10.14742/ajet.3296]
27. Quail M, Brundage SB, Spitalnick J, Allen PJ, Beilby J. Student self-reported communication skills, knowledge and confidence across standardised patient, virtual and traditional clinical learning environments. BMC Med Educ 2016 Feb 27;16(1):73. [doi: 10.1186/s12909-016-0577-5] [Medline: 26919838]
28. Jiang H, Vimalesvaran S, Wang JK, Lim KB, Mogali SR, Car LT. Virtual reality in medical students' education: scoping review. JMIR Med Educ 2022 Feb 2;8(1):e34860. [doi: 10.2196/34860] [Medline: 35107421]
29. Osborne M, Truman C, Sheridan S. Perspectives: tomorrow's nursing education is here today: experiences of artificial intelligence and virtual reality technologies. J Res Nurs 2024 Mar;29(2):191-198. [doi: 10.1177/17449871241240728] [Medline: 39070572]
30. Venktaramana V, Loh EKY, Wong CJW, et al. A systematic scoping review of communication skills training in medical schools between 2000 and 2020. Med Teach 2022 Sep;44(9):997-1006. [doi: 10.1080/0142159X.2022.2054693] [Medline: 35653622]
31. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126. [doi: 10.2196/jmir.1923] [Medline: 22209829]
32. This is AI simulation. SimConverse. 2024. URL: https://www.simconverse.com [accessed 2025-01-23]
33. Adams MJD, Umbach PD. Nonresponse and online student evaluations of teaching: understanding the influence of salience, fatigue, and academic environments. Res High Educ 2012 Aug;53(5):576-591. [doi: 10.1007/s11162-011-9240-5]
34. Kurtz SM, Silverman JD. The Calgary-Cambridge referenced observation guides: an aid to defining the curriculum and organizing the teaching in communication training programmes. Med Educ 1996 Mar;30(2):83-89. [doi: 10.1111/j.1365-2923.1996.tb00724.x] [Medline: 8736242]
35. Kurtz S, Silverman J, Benson J, Draper J. Marrying content and process in clinical method teaching: enhancing the Calgary-Cambridge guides. Acad Med 2003 Aug;78(8):802-809. [doi: 10.1097/00001888-200308000-00011] [Medline: 12915371]
36. Silverman JD, Kurtz SM, Draper J. Skills for Communicating with Patients, 3rd edition: CRC Press Taylor & Francis Group; 2013. URL: https://api.pageplace.de/preview/DT0400.9781910227268_A24201891/preview-9781910227268_A24201891.pdf [accessed 2025-10-28]
37. Simmenroth-Nayda A, Heinemann S, Nolte C, Fischer T, Himmel W. Psychometric properties of the Calgary Cambridge guides to assess communication skills of undergraduate medical students. Int J Med Educ 2014 Dec 6;5:212-218. [doi: 10.5116/ijme.5454.c665] [Medline: 25480988]
38. Arundel A. How to Design, Implement, and Analyse a Survey: Northampton: Edward Elgar Publishing; 2023. [doi: 10.4337/9781800376175]

XSL•FO
RenderX

39. Carifio J, Perla RJ. Ten common misunderstandings, misconceptions, persistent myths and urban legends about likert scales and likert response formats and their antidotes. J of Social Sciences 2007 Mar 1;3(3):106-116. [doi: 10.3844/jssp.2007.106.116]
40. Sullivan GM, Artino AR Jr. Analyzing and interpreting data from Likert-type scales. J Grad Med Educ 2013 Dec;5(4):541-542. [doi: 10.4300/JGME-5-4-18] [Medline: 24454995]
41. Walsh K. Economic analysis in medical education: definition of essential terms. Med Teach 2014 Oct;36(10):890-893. [doi: 10.3109/0142159X.2014.917161] [Medline: 25072235]
42. Bray L, Krogh TB, Østergaard D. Simulation-based training for continuing professional development within a primary care context: a systematic review. Educ Prim Care 2023 Mar;34(2):64-73. [doi: 10.1080/14739879.2022.2161424] [Medline: 36730551]
43. Kim J, Park JH, Shin S. Effectiveness of simulation-based nursing education depending on fidelity: a meta-analysis. BMC Med Educ 2016 May 23;16(1):152. [doi: 10.1186/s12909-016-0672-7] [Medline: 27215280]
44. Issenberg SB, McGaghie WC, Petrusa ER, Lee Gordon D, Scalese RJ. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. Med Teach 2005 Jan;27(1):10-28. [doi: 10.1080/01421590500046924] [Medline: 16147767]
45. McGaghie WC, Issenberg SB, Petrusa ER, Scalese RJ. A critical review of simulation-based medical education research: 2003-2009. Med Educ 2010 Jan;44(1):50-63. [doi: 10.1111/j.1365-2923.2009.03547.x] [Medline: 20078756]
46. Khan K, Pattison T, Sherwood M. Simulation in medical education. Med Teach 2011;33(1):1-3. [doi: 10.3109/0142159X.2010.519412] [Medline: 21182376]
47. Lazzereschi L, Kirtley D. Provision of e-learning programmes to replace undergraduate medical students' clinical general practice attachments during COVID-19 stand-down: a student perspective. Educ Prim Care 2021 Jan;32(1):63. [doi: 10.1080/14739879.2020.1786468] [Medline: 32602412]
48. Furlan R, Gatti M, Menè R, et al. A natural language processing-based virtual patient simulator and intelligent tutoring system for the clinical diagnostic process: simulator development and case study. JMIR Med Inform 2021 Apr 9;9(4):e24073. [doi: 10.2196/24073] [Medline: 33720840]
49. Persad A, Stroulia E, Forgie S. A novel approach to virtual patient simulation using natural language processing. Med Educ 2016 Nov;50(11):1162-1163. [doi: 10.1111/medu.13197] [Medline: 27762013]
50. Woo CW, Evens MW, Freedman R, et al. An intelligent tutoring system that generates a natural language dialogue using dynamic multi-level planning. Artif Intell Med 2006 Sep;38(1):25-46. [doi: 10.1016/j.artmed.2005.10.004] [Medline: 16352427]
51. Kiesewetter J, Sailer M, Jung VM, et al. Learning clinical reasoning: how virtual patient case format and prior knowledge interact. BMC Med Educ 2020 Mar 14;20(1):73. [doi: 10.1186/s12909-020-1987-y] [Medline: 32171297]
52. Kaltman S, Talisman N, Pennestri S, Syverson E, Arthur P, Vovides Y. Using technology to enhance teaching of patient-centered interviewing for early medical students. Simul Healthc 2018 Jun;13(3):188-194. [doi: 10.1097/SIH.0000000000000304] [Medline: 29771814]
53. Aspegren K. BEME guide no. 2: teaching and learning communication skills in medicine-a review with quality grading of articles. Med Teach 1999;21(6):563-570. [doi: 10.1080/01421599978979] [Medline: 21281175]
54. Shorey S, Ang E, Yap J, Ng ED, Lau ST, Chui CK. A virtual counseling application using artificial intelligence for communication skills training in nursing education: development study. J Med Internet Res 2019 Oct 29;21(10):e14658. [doi: 10.2196/14658] [Medline: 31663857]
55. Guetterman TC, Sakakibara R, Baireddy S, et al. Medical students' experiences and outcomes using a virtual human simulation to improve communication skills: mixed methods study. J Med Internet Res 2019 Nov 27;21(11):e15459. [doi: 10.2196/15459] [Medline: 31774400]
56. Jacklin S, Maskrey N, Chapman S. Shared decision-making with a virtual patient in medical education: mixed methods evaluation study. JMIR Med Educ 2021 Jun 10;7(2):e22745. [doi: 10.2196/22745] [Medline: 34110299]
57. Sunnqvist C, Karlsson K, Lindell L, Fors U. Virtual patient simulation in psychiatric care - a pilot study of digital support for collaborate learning. Nurse Educ Pract 2016 Mar;17:30-35. [doi: 10.1016/j.nepr.2016.02.004] [Medline: 27038085]
58. Yamamoto A, Koda M, Ogawa H, et al. Enhancing medical interview skills through ai-simulated patient interactions: nonrandomized controlled trial. JMIR Med Educ 2024 Sep 23;10:e58753. [doi: 10.2196/58753] [Medline: 39312284]
59. O'Rourke SR, Branford KR, Brooks TL, Ives LT, Nagendran A, Compton SN. The emotional and behavioral impact of delivering bad news to virtual versus real standardized patients: a pilot study. Teach Learn Med 2020;32(2):139-149. [doi: 10.1080/10401334.2019.1652180] [Medline: 31437006]
60. Kruger J, Dunning D. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. J Pers Soc Psychol 1999;77(6):1121-1134. [doi: 10.1037/0022-3514.77.6.1121] [Medline: 10626367]
61. A web-based artificial intelligence-driven virtual patient simulator versus actor-based simulation for teaching consultation skills: a multi-centre randomised cross-over study. University of Nottingham Research Repository (Worktribe). URL: https://rdmc.nottingham.ac.uk/handle/internal/12099 [accessed 2025-11-07]

XSL•FO
RenderX

Abbreviations

AB-CST: actor-based consultation skills training
AI: artificial intelligence
AI-CST: artificial intelligence-based communication skills training
ILO: intended learning outcome
UoL: University of Lincoln
UoN: University of Nottingham
VPS: virtual patient simulator

Edited by A Mavragani; submitted 23.01.25; peer-reviewed by A White, F Sendra-Portero; revised version received 25.09.25; accepted 28.09.25; published 20.11.25.

Please cite as:
Tyrrell EG, Sandhu SK, Berry K, Ghannam SF, Lewis SA, Crowfoot D, Sahota GS, Carson J, Wilson EE, Taggar J
Web-Based AI-Driven Virtual Patient Simulator Versus Actor-Based Simulation for Teaching Consultation Skills: Multicenter Randomized Crossover Study
JMIR Form Res 2025;9:e71667
URL: https://formative.jmir.org/2025/1/e71667
doi:10.2196/71667

© Edward G Tyrrell, Sardip K Sandhu, Kathryn Berry, Suzan F Ghannam, Sarah A Lewis, Daniel Crowfoot, Gurvinder Singh Sahota, Julie Carson, Emma E Wilson, Jaspal Taggar. Originally published in JMIR Formative Research (https://formative.jmir.org), 20.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Teaching Clinical Reasoning in Health Care Professions Learners Using AI-Generated Script Concordance Tests: Mixed Methods Formative Evaluation

Alexandre Hudon[1,2,3,4,5], BEng, MSc, MD, PhD; Véronique Phan[6,7], MSc, MD; Bernard Charlin[5,8], MEd, MD, PhD; René Wittmer[9], MD

[1]Department of Psychiatry and Addictology, Faculty of Medicine, Université de Montréal, Pavillon Roger-Gaudry, 2900 Bd Édouard-Montpetit Local L-315, Montréal, QC, Canada

[2]Department of Psychiatry, Institut universitaire en santé mentale de Montréal, Montreal, QC, Canada

[3]Centre de recherche de l'Institut universitaire en santé mentale de Montréal, Montreal, QC, Canada

[4]Department of Psychiatry, Institut national de psychiatrie légale Philippe-Pinel, Montreal, QC, Canada

[5]Groupe Interdisciplinaire de recherche sur la cognition et le raisonnement professionnel (GIRCoPRo), Université de Montréal, Montreal, QC, Canada

[6]Department of Pediatrics, Faculty of Medicine, Université de Montréal, Montreal, QC, Canada

[7]Department of Pediatrics, Department of Pediatrics, Centre Hospitalier Universitaire Sainte-Justine, Montreal, QC, Canada

[8]Centre de pédagogie appliquée au sciences de la santé, Faculty of Medicine, Université de Montréal, Montreal, QC, Canada

[9]Department of Family Medicine and Emergency Medicine, Faculty of Medicine, Université de Montréal, Montreal, QC, Canada

Corresponding Author:
Alexandre Hudon, BEng, MSc, MD, PhD
Department of Psychiatry and Addictology, Faculty of Medicine, Université de Montréal, Pavillon Roger-Gaudry, 2900 Bd Édouard-Montpetit Local L-315, Montréal, QC, Canada

Abstract

Background: The integration of artificial intelligence (AI) in medical education is evolving, offering new tools to enhance teaching and assessment. Among these, script concordance tests (SCTs) are well-suited to evaluate clinical reasoning in contexts of uncertainty. Traditionally, SCTs require expert panels for scoring and feedback, which can be resource-intensive. Recent advances in generative AI, particularly large language models (LLMs), suggest the possibility of replacing human experts with simulated ones, though this potential remains underexplored.

Objective: This study aimed to evaluate whether LLMs can effectively simulate expert judgment in SCTs by using generative AI to author, score, and provide feedback for SCTs in cardiology and pneumology. A secondary objective was to assess students' perceptions of the test's difficulty and the pedagogical value of AI-generated feedback.

Methods: A cross-sectional, mixed methods study was conducted with 25 second-year medical students who completed a 32-item SCT authored by ChatGPT-4o (OpenAI). Six LLMs (3 trained on the course material and 3 untrained) served as simulated experts to generate scoring keys and feedback. Students answered SCT questions, rated perceived difficulty, and selected the most helpful feedback explanation for each item. Quantitative analysis included scoring, difficulty ratings, and correlations between student and AI responses. Qualitative comments were thematically analyzed.

Results: The average student score was 22.8 out of 32 (SD 1.6), with scores ranging from 19.75 to 26.75. Trained AI systems showed significantly higher concordance with student responses (ρ=0.64) than untrained models (ρ=0.41). AI-generated feedback was rated as most helpful in 62.5% of cases, especially when provided by trained models. The SCT demonstrated good internal consistency (Cronbach α=0.76), and students reported moderate perceived difficulty (mean 3.7, SD 1.1). Qualitative feedback highlighted appreciation for SCTs as reflective tools, while recommending clearer guidance on Likert-scale use and more contextual detail in vignettes.

Conclusions: This is among the first studies to demonstrate that trained generative AI models can reliably simulate expert clinical reasoning within a script-concordance framework. The findings suggest that AI can both streamline SCT design and offer educationally valuable feedback without compromising authenticity. Future studies should explore longitudinal effects on learning and assess how hybrid models (human and AI) can optimize reasoning instruction in medical education.

(JMIR Form Res 2025;9:e76618) doi:10.2196/76618

XSL•FO
RenderX

KEYWORDS

script concordance test; clinical reasoning; artificial intelligence; large language models; medical education; formative assessment; generative AI; expert simulation

Introduction

One of the biggest challenges in medical education is evaluating clinical capabilities, especially higher-order abilities such as clinical reasoning and decision-making. Knowledge-based tests, such as multiple-choice questions, frequently fall short of capturing the subtleties of clinical reasoning in the face of diagnostic uncertainty [1]. By simulating clinical interactions, tools such as workplace-based assessments and Objective Structured Clinical Examinations try to assess clinical reasoning, but they come with a high logistical and human cost. An additional method for addressing thinking in clinical environments with unclear definitions is the script concordance test (SCT). In contrast to standard assessments, SCTs ask students to use a Likert-type scale to rate the impact of scenarios that are followed by additional clinical information on a postulated hypothesis. Concordance with a panel of expert responses serves as the basis for scoring, which reflects the variety and probabilistic nature of actual clinical practice [2,3]. SCTs are valuable, but creating and scoring them requires time and involves hiring several specialists who must give careful, consistent answers to a variety of questions.

The format of an SCT usually involves presenting a clinical vignette followed by a diagnostic, investigative, or therapeutic hypothesis. A new piece of information is then introduced, and the test-taker must judge, using a Likert-type scale, how this new data affects the likelihood or relevance of the hypothesis. Responses are scored by comparing the learner's choices to those of a panel of experienced clinicians, with partial credit awarded based on the distribution of expert answers rather than a single correct response. This scoring system captures the variability in expert judgment and allows the SCT to measure concordance with expert reasoning rather than factual recall. Validity evidence for the SCT includes its ability to discriminate between levels of expertise, its positive impact on cognitive engagement, and its reliability when a sufficient number of items and panel members are used [3-5]. As such, SCTs are now used internationally in both undergraduate and postgraduate medical education, particularly in specialties that demand complex clinical judgment.

Artificial intelligence (AI) algorithms, such as large language models (LLMs), have developed in recent years, presenting a possible paradigm change in this field. Having been trained on extensive text datasets, LLMs are deep learning models that can produce language that is similar to that of a person, summarize intricate concepts, and even mimic domain-specific reasoning [6]. These models, as opposed to rule-based systems, generate context-aware responses, often with good fluency, by identifying patterns in unstructured data. LLMs have demonstrated the ability to replicate patient or clinician discourse, produce high-quality feedback, and perform at or near the passing threshold on United States Medical Licensing Examination–style questions [7-9].

In the field of LLMs used to generate SCTs, several studies have demonstrated the usefulness of this tool for creating clinical vignettes in medical education and other health sciences fields [10,11]. A recent study highlighted the educational quality of script concordance vignettes created using ChatGPT (OpenAI [11]). To use tools based on generative AI in the design of educational materials, and to improve efficiency while optimizing human and material resources, it is essential to evaluate the level of difficulty of the proposed scenarios and questions [12]. Moreover, to our knowledge, no study has yet demonstrated the ability of generative AI to embody the role of a clinical expert in selecting the most appropriate response to a given scenario and in providing feedback to learners.

The main objective of this project is to explore the performance of generative AI in its ability to create SCTs for undergraduate medical students and to embody the role of an expert. A secondary objective is to evaluate how medical learners perceive the difficulty level of the AI-generated SCTs, as well as their level of appreciation for each type of feedback provided.

Methods

Overview

This study used a cross-sectional, mixed methods experimental design to evaluate the performance of medical students on a generative AI–authored formative SCT and to assess the use of AI systems as content experts for both scoring and providing formative feedback. The flow diagram of this study is provided in Figure 1.

XSL•FO
RenderX

Figure 1. Protocol flow diagram for this study. AI: artificial intelligence; SCT: script concordance test.

XSL•FO
RenderX

Participants and Recruitment

The study was conducted in November 2024 at the Université de Montréal's Faculty of Medicine. To be included, participants needed to be enrolled in the second year of the undergraduate medical education program, have completed their cardiology course, and be preparing for their respirology exam, having completed the related course materials. This timing allowed for the inclusion of clinical reasoning questions involving symptoms and signs that may relate to both systems, given their close physiological and clinical connections. An email was sent to all potential participants to register for this formative examination.

Test Development and Structure

The SCT was developed using generative AI (ChatGPT-4o) following established guidelines for script concordance testing [13]. Four clinical vignettes (chronic obstructive pulmonary disease [COPD] exacerbation, hemoptysis, chest pain, and acute cough) were created, each containing 8 items for a total of 32. Prompts were standardized to ensure consistent vignette structure, 5-point Likert scaling (ranging from not likely to very likely, indicating how the new information would influence the hypothesis), and reasoning alignment across cardiopulmonary domains. Each item followed the classic SCT structure: (1) a clinical scenario introducing uncertainty, (2) a diagnostic, investigative, or therapeutic hypothesis, and (3) a new clinical finding. Items were designed and reviewed to reflect varying degrees of diagnostic ambiguity, probabilistic reasoning, and real-world uncertainty.

The full test, in French and its English translation, is provided in Multimedia Appendix 1. The following prompts were used to design the SCTs: (1) act as an expert in health sciences education at the university level, specializing in cardiology and pulmonology; (2) act as an expert in health sciences education at the university level, specializing in cardiology and pulmonology; (3) act as an expert in designing SCT vignettes; (4) generate a script concordance vignette that includes 8 questions based on a theme integrating both pulmonology and cardiology; and (5) create questions related to the vignette that begin with: "If you are considering 'a diagnostic hypothesis' and you find 'a sign or symptom,' how does this finding affect the likelihood of the hypothesis?" (using a Likert scale from −2 to +2 which refers to not likely to very likely).

While the AI-generated items were designed to follow the canonical SCT format (clinical vignette, hypothesis, new information, and Likert-scale inference), some questions did not fully adhere to this structure. This occurred when the generative model incorporated redundant information from the vignette into subsequent items, occasionally blurring the distinction between context and new data. The prompt did not explicitly constrain the model to exclude previously stated information, which likely contributed to partial format deviations.

Use of Generative AI as an Expert Panel

The reference panel comprised 6 LLMs, chosen to balance diversity of reasoning with practical feasibility. Although the classical SCT format recommends 15 - 20 human experts to ensure adequate response variability [5], LLMs can rapidly generate consistent probabilistic judgments across multiple iterations. Pilot testing indicated that 6 distinct model architectures provided sufficient variability in reasoning patterns to construct an aggregated scoring key while keeping the process computationally manageable for formative evaluation. These models were divided into two groups:

1. Trained AI experts (n=3): these systems (ChatGPT-4o, Claude 3.7 [Anthropic], and Microsoft 365 Copilot) were not fine-tuned or modified at the model-parameter level. Instead, each was contextually conditioned using course-specific documents (lecture notes and clinical guidelines from the cardiopulmonary block) that were embedded directly into the prompt context before generating responses. This approach (often referred to as in-context learning or prompt-based domain adaptation) allowed the models to access relevant curricular information without altering their underlying architecture.
2. Untrained AI experts (n=3): the same base models were used without additional curricular material or contextual embedding, representing their general-purpose configuration.

Each AI system was prompted to respond to all 32 SCT items as if it were a senior clinical expert. Their responses were used to establish scoring keys and to generate formative feedback explanations for each item. The modal response across all 6 AI experts served as the reference answer for scoring.

The prompts used to generate feedback were the same across the different AI systems. They were: (1) act as a medical expert in cardiology and pneumology; (2) act as a SCT expert; (3) for this vignette and each of its related questions ([Insert Vignette] and [Insert Questions]), please pick the most appropriate answer from the Likert scale; and (4) for each of your answers, please provide an explanation as if you were trying to explain your reasoning to a medical student.

Student Completion and Data Collection

Participants completed a 32-question AI-generated SCT via a Google Form. For each item, students (1) selected a Likert-scale response, (2) rated the item's perceived difficulty on a 7-point scale, (3) identified the most and least helpful feedback explanations among 6 anonymous AI-generated rationales, and (4) indicated whether they believed the question was AI- or human-authored. Open-ended comment boxes captured qualitative impressions. Completion was voluntary and anonymous, with implied consent through submission. The total estimated completion time was 35 - 45 minutes. Responses were exported and cleaned in Microsoft Excel for analysis.

Scoring and Analysis

Student answers were scored using the aggregate partial credit method: (1) one point was awarded if the student's response matched the modal expert answer, (2) partial credit (eg, 0.75 and 0.5) was granted if the response matched a minority AI expert choice, and (3) zero points were given if the response did not match any AI panel member.

XSL•FO
RenderX

Descriptive statistics were computed for total scores, individual item performance, and perceived difficulty. Internal consistency was measured using Cronbach α. AI-to-AI and AI-to-student concordance was evaluated using Spearman rank correlation (ρ).

In the feedback evaluation component, student preferences for expert rationales were tabulated across all items. Each expert (AI system) received a score reflecting how frequently its feedback was selected as the "most helpful" and "least helpful," allowing comparison between trained and untrained AI systems as feedback providers.

Qualitative Analysis

Open-ended responses were analyzed using an inductive thematic approach. Comments were reviewed, coded iteratively, and grouped into major and minor themes by 2 independent reviewers. Discrepancies were resolved through discussion and consensus. Themes focused on student perceptions of the SCT format, feedback quality, and suggestions for improvement.

Ethical Considerations

This study was approved by the Université de Montréal's Research Ethics Committee in Education and Psychology under the project name 2024 - 6168. Participation was anonymous and voluntary, with implied consent obtained through survey completion.

Results

Participants

A total of 25 students volunteered to complete the SCT in a formative, nonevaluative setting. Each student also evaluated the difficulty of each question, selected the most and least helpful expert feedback, and provided qualitative comments on individual questions and on the overall examination. Six generative AI systems (3 trained on course content and 3 untrained) completed the test and provided justifications for each item, which served as immediate feedback for the students.

Student Performance

Among the 25 second-year medical students who completed the SCT, the average score was 22.8 out of 32 (SD 1.6), with individual scores ranging from 19.75 to 26.75. The distribution of scores approximated a normal curve with a slight left skew, indicating generally strong alignment between student responses and the AI expert panel, while still highlighting variability in performance. This range suggests that the AI-generated SCT was capable of differentiating clinical reasoning abilities among participants, supporting its discriminatory power even in a formative context. The full distribution of results for each question is provided in Table 1.

XSL•FO
RenderX

Table . Distribution of the test results (N=25).

Question (English)	Mean (SD)	Range (min-max)	Median (IQR)
Q1: To assess an acute COPD[a] exacerbation, determine the impact of an oxygen saturation of 85%.	0.72 (0.15)	0.25-1	0.75 (0.75-0.75)
Q2: To evaluate acute pulmonary edema, judge the impact of a recent respiratory infection.	0.65 (0.2)	0-0.75	0.75 (0.75-0.75)
Q3: To assess the hypothesis of pulmonary embolism, determine the impact of dyspnea and tachycardia.	0.76 (0.17)	0.25-1	0.75 (0.75-0.75)
Q4: To evaluate myocardial dysfunction, judge the influence of a high heart rate.	0.56 (0.27)	0-1	0.75 (0.50-0.75)
Q5: To differentiate pulmonary edema from asthma exacerbation, assess the importance of a chest X-ray.	0.89 (0.21)	0.5-1	1 (1-1)
Q6: To evaluate a COPD exacerbation, determine how a history of smoking affects the hypothesis.	0.82 (0.15)	0.5-1	0.75 (0.75-1)
Q7: To guide diagnosis in respiratory distress, assess the importance of history-taking.	0.97 (0.11)	0.5-1	1 (1-1)
Q8: To diagnose pulmonary edema, judge the influence of acute electrocardiogram ischemia.	0.79 (0.19)	0.25-1	0.75 (0.75-1)
Q9: To assess tuberculosis as a differential diagnosis, evaluate the impact of night sweats and weight loss.	0.78 (0.08)	0.75-1	0.75 (0.75-0.75)
Q10: To assess the likelihood of chronic pulmonary hemoptysis, judge the influence of a history of bronchiectasis.	0.76 (0.13)	0.25-1	0.75 (0.75-0.75)
Q11: To assess life-threatening hemoptysis requiring urgent care, evaluate the impact of an oxygen saturation of 92%.	0.59 (0.32)	0-1	0.75 (0.25-0.75)
Q12: To evaluate bronchopulmonary cancer, determine the influence of a 30 pack-year smoking history.	0.86 (0.13)	0.75-1	0.75 (0.75-1)
Q13: To orient toward an infectious etiology like tuberculosis or pneumonia, assess the usefulness of the patient's history.	1 (0)	1-1	1 (1-1)
Q14: To distinguish possible causes of hemoptysis, assess the importance of performing a chest X-ray.	1 (0)	1-1	1 (1-1)
Q15: To assess active bronchiectasis, judge the interaction between smoking history and risk of complications.	0.82 (0.14)	0.5-1	0.75 (0.75-1)
Q16: To diagnose tuberculosis-related hemoptysis, evaluate the influence of prior tuberculosis exposure.	0.86 (0.13)	0.75-1	0.75 (0.75-1)

XSL•FO
RenderX

Question (English)	Mean (SD)	Range (min-max)	Median (IQR)
Q17: To assess acute coronary syndrome (ACS), evaluate the impact of ST elevation in DII, DIII, and aVF[b].	0.99 (0.05)	0.75-1	1 (1-1)
Q18: To evaluate atypical ACS presentation, assess the influence of diabetes.	0.75 (0.16)	0.5-1	0.75 (0.75-0.75)
Q19: To assess aortic dissection as a cause of pain, evaluate the impact of hypertension.	0.77 (0.1)	0.5-1	0.75 (0.75-0.75)
Q20: To evaluate suspected pneumothorax, judge how pain radiation influences the hypothesis.	0.29 (0.27)	0-1	0.25 (0.25-0.25)
Q21: To differentiate between ACS and pulmonary embolism, assess the importance of the electrocardiogram.	0.96 (0.12)	0.5-1	1 (1-1)
Q22: To evaluate pulmonary embolism, determine the influence of elevated blood pressure on the hypothesis.	0.55 (0.23)	0.25-0.75	0.75 (0.25-0.75)
Q23: To assess ACS likelihood, judge how chest pain radiation differs between male and female patients.	0.77 (0.2)	0.25-1	0.75 (0.75-1)
Q24: To determine the urgency of management, assess the influence of diabetes and hypertension.	0.78 (0.21)	0.5-1	0.75 (0.50-1)
Q25: To assess community-acquired pneumonia, evaluate the impact of initial fever on the hypothesis.	0.86 (0.13)	0.75-1	0.75 (0.75-1)
Q26: To evaluate treatment-induced cough, determine the influence of ACE inhibitor use.	0.72 (0.18)	0.25-1	0.75 (0.75-0.75)
Q27: To assess severe pneumonia requiring hospitalization, evaluate the impact of a normal oxygen saturation (97%).	0.15 (0.16)	0-0.5	0.25 (0-0.25)
Q28: To assess acute bronchitis, evaluate the impact of crackles on the hypothesis.	0.43 (0.28)	0-1	0.25 (0.25-0.75)
Q29: To distinguish between ACE-inhibitor-induced[c] cough and acute respiratory infection, assess the usefulness of the history.	1 (0)	1-1	1 (1-1)
Q30: To assess upper respiratory tract infection, evaluate the influence of a 10-day duration of cough.	0.7 (0.2)	0.25-1	0.75 (0.75-0.75)
Q31: To differentiate pneumonia from bronchitis, assess the importance of a chest X-ray.	1 (0)	1-1	1 (1-1)

XSL•FO
RenderX

Question (English)	Mean (SD)	Range (min-max)	Median (IQR)
Q32: To guide the need for antibiotic treatment in suspected pneumonia, evaluate the role of initial fever and productive cough.	0.84 (0.2)	0.25-1	1 (0.75-1)

[a]COPD: chronic obstructive pulmonary disease.

[b]aVF: augmented vector foot.

[c]ACE: angiotensin-converting enzyme.

A more detailed item-level analysis revealed that questions with clearer clinical cues, such as Q1 (which evaluated the impact of hypoxemia [peripheral oxygen saturation of 85%] on the hypothesis of COPD exacerbation) and Q8 (which examined the influence of electrocardiogram findings on the diagnosis of acute pulmonary edema), had high rates of agreement with the expert panel. In these items, over 80% of student responses matched the most common expert answer, suggesting that students were comfortable applying well-established diagnostic patterns. In contrast, questions such as Q2 (regarding the influence of a recent respiratory infection on the hypothesis of pulmonary edema) and Q4 (assessing the relevance of tachycardia for myocardial dysfunction) showed greater variability in student responses, with less than 50% aligning with the modal expert choice. These items appeared to introduce more clinical ambiguity or demanded a higher level of inference, which may explain the broader dispersion of answers. Preliminary psychometric analysis demonstrated a Cronbach α of 0.76, indicating good internal consistency across the 32 items.

About 40% of students scored 25 or above, frequently aligning with expert responses even on more ambiguous questions. These students tended to report lower perceived difficulty and used the “slightly confirmed” response category more judiciously. The remaining 60% scored between 19 and 24 points and were more likely to select “non influenced” or “slightly confirmed” options, reflecting a more cautious reasoning style or uncertainty in applying new clinical information. Overall, the SCT demonstrated its utility in capturing variations in reasoning patterns and levels of diagnostic confidence among early clinical learners.

Agreement Between AI and Medical Students

To evaluate the extent to which AI experts simulated expert reasoning, we examined concordance (the degree to which AI and student responses followed similar reasoning patterns) using Spearman rank correlation coefficient (ρ). This nonparametric measure assesses the strength and direction of association between 2 ranked sets of scores, where values closer to +1 indicate strong agreement and those near 0 indicate weak or no relationship.

Among the trained AI systems, ChatGPT-4o achieved the highest concordance with student responses (ρ=0.68; $P<.001$), followed by Claude (ρ=0.64) and Microsoft Copilot (ρ=0.61). These coefficients represent moderate-to-strong positive correlations, suggesting that trained models reasoned in ways closely aligned with students’ decision patterns. In contrast, untrained models showed weaker correlations (average ρ=0.41), reflecting less consistent or more generic reasoning.

Educationally, this implies that contextualized AI models can mirror how medical students weigh diagnostic information when course-specific data are embedded in their prompts. Such alignment supports the potential use of trained AI not only as an assessment proxy but also as a feedback tool capable of modeling clinically coherent reasoning.

The untrained AI models demonstrated lower levels of agreement. The untrained version of ChatGPT-4o showed a moderate correlation (ρ=0.48; $P=.04$), while Claude and Copilot, when untrained, showed weaker and statistically nonsignificant correlations of 0.42 ($P=.06$) and 0.34 ($P=.08$), respectively. These models tended to exhibit less consistent patterns, often defaulting to neutral or noncommittal answers such as “non influenced,” especially in items requiring nuanced interpretation of clinical signs. Their reasoning lacked the contextual anchoring present in the trained models and showed more variability across similar clinical scenarios. On average, the untrained models produced a correlation coefficient of 0.41, substantially lower than their trained counterparts.

Perceived Difficulty

In addition to completing the 32 SCT items, students were asked to rate the perceived difficulty of each question on a 7-point Likert scale, ranging from 1 (“very easy”) to 7 (“very difficult”). This measure provided insight into how learners experienced the complexity and cognitive demands of the AI-generated test items. The mean perceived difficulty across all questions was 3.7, suggesting that students generally found the test to be of moderate difficulty. This aligns with the formative purpose of the SCT and indicates that the AI-generated items were accessible yet still challenging enough to stimulate clinical reasoning.

When analyzed at the item level, some questions emerged as consistently more difficult. In particular, Q2 (which assessed how a recent respiratory infection influences the hypothesis of acute pulmonary edema) and Q4 (which explored the relevance of tachycardia in the context of suspected myocardial dysfunction) received higher average difficulty ratings, with mean scores exceeding 4.5 (SD 0.9). These questions involved ambiguous or indirect relationships between the new information and the hypothesis, requiring students to reason within clinical gray zones where data interpretation is not straightforward. In contrast, items that presented clear diagnostic anchors, such as Q1 (hypoxemia in COPD) and Q8 (electrocardiogram ischemia in pulmonary edema), were rated as significantly easier, with average difficulty scores below 3.0. These findings suggest that item clarity and familiarity with pathophysiological mechanisms strongly influence perceived complexity.

XSL•FO
RenderX

Furthermore, there was no statistically significant difference in perceived difficulty between items believed by students to be human-authored and those they believed to be AI-authored (*t* test; *P*=.47). As all items were in fact AI-generated, this finding reflects students' perceived authenticity of the AI-written vignettes rather than a comparison with true human-authored items. This finding supports the face validity of AI-generated items, indicating that they are perceived as authentic and comparable in challenge to traditional, faculty-written SCT items. Several students even remarked in open comments that they were "unable to tell" which questions had been generated by AI, further supporting the indistinguishability of AI-authored content when properly crafted.

Cluster analysis of difficulty ratings also revealed subtle differences in student subgroups. High-performing students (as defined by SCT scores ≥25/32) tended to rate the test as less difficult overall (mean 3.3, SD 0.8) compared to their peers (mean 3.9, SD 0.8), which may reflect their greater familiarity with clinical reasoning or more efficient script activation. In addition, free-text comments frequently noted that clinical ambiguity (rather than item length or language) was the primary factor contributing to perceived difficulty. Students consistently emphasized that the challenge came from "figuring out what matters most" in the presence of partial or conflicting clinical cues. The difficulty ratings offer an important validation metric for AI-generated SCTs. The overall moderate difficulty, the range of item-specific variation, and the lack of perceived discrepancy between AI- and human-generated content all point toward a well-calibrated test.

Spearman correlation analysis between item difficulty ratings and alignment with trained AI responses yielded a modest positive trend (ρ=0.32; *P*=.08), suggesting that items more closely matching trained AI reasoning were often perceived as less difficult. This pattern hints that when AI-generated reasoning aligns with instructional scripts, students experience smoother cognitive processing and reinforce the pedagogical relevance of model calibration.

Evaluation of AI-Expert Feedback

In addition to rating the difficulty of each SCT item, students were asked to select the expert feedback they considered most helpful and least helpful among the options provided. For every SCT question, 6 expert rationales were made available to the medical students by the AI systems (trained and untrained). Students were instructed to base their selection on 3 specific criteria: clarity, clinical relevance, and overall educational value. This evaluation provided insight not only into the perceived credibility of AI-generated explanations but also into the comparative strengths of different expert types in formative feedback contexts.

Across all 32 questions, AI-generated feedback (especially from models trained on the course material) was selected as helpful in 62.5% of instances. Among these, ChatGPT-4o (trained) was most frequently chosen, followed by Claude (trained) and Copilot (trained). Students frequently cited the clarity of language, conciseness, and alignment with taught clinical reasoning strategies as reasons for preferring these explanations. Trained AI feedback was especially appreciated in items requiring pathophysiological reasoning, such as those related to acute coronary syndromes and differential diagnoses in respiratory presentations. The consistent use of structured logic and evidence-based phrasing contributed to a perception of trustworthiness and educational quality.

In contrast, the feedback generated by untrained AI systems was selected as most helpful in only 9.4% of instances and more frequently identified as least useful. The untrained models often produced generic or excessively cautious explanations, relying on vague phrases such as "may support the hypothesis" or "needs further investigation" without directly engaging with the specific clinical cues provided in the vignette. These models occasionally failed to recognize important pathophysiological links, leading to misalignment with students' expectations. Several students noted that while the untrained AI responses were not necessarily incorrect, they lacked the pedagogical precision needed for effective feedback.

Qualitative Feedback From Students

Overview

To complement the quantitative analyses, students were invited to provide open-ended comments after each SCT question and at the end of the test. These narrative responses were analyzed using an inductive thematic analysis. A total of 94 discrete comments were submitted by the 25 participating students. From these, three major themes were identified: (1) ambiguity and interpretation of Likert-scale options, (2) educational value of SCTs, and (3) recommendations for improvement in item design and feedback delivery. Key themes and examples of quotes are provided in Table 2.

XSL•FO
RenderX

Table . Key themes, descriptions, and their representative quotes.

Theme	Description	Representative quotes
Ambiguity and interpretation of Likert-scale options	Students expressed uncertainty distinguishing between closely-related options (”slightly confirmed“ vs ”confirmed”), reflecting the inherent challenge of probabilistic reasoning under limited data.	• “It’s really hard to know when to say ‘confirmed’ vs ‘slightly confirmed’ without lab results.” • “Sometimes the answer depends on how I imagine the rest of the case.”
Educational value of the SCT[a] format	Learners appreciated how SCTs stimulated reflective reasoning and moved beyond right-wrong logic, reinforcing their understanding of uncertainty.	• “This test made me think like a real clinician.” • “I liked how the questions made me weigh the value of each piece of information.”
Recommendations for improvement	Students suggested adding contextual information (eg, vital signs and comorbidities) and calibration examples to improve clarity and confidence.	• “More case details would help me justify my choice better.” • “Before the real test, I would have liked an example with expert explanation.”

[a]SCT: script concordance test.

Ambiguity and Interpretation of Likert-Scale Options

A recurrent theme in the comments related to students’ uncertainty in differentiating between closely related answer choices. Many expressed difficulties in deciding whether a new piece of clinical information “slightly confirmed” versus “confirmed” a hypothesis, particularly when the vignette lacked strong diagnostic anchors.

This theme reflects a broader cognitive tension between probabilistic reasoning and binary decision-making in early clinical training. Students seemed to grasp the script concordance philosophy but struggled with the semantic granularity required by the scale, particularly in low-certainty situations.

Educational Value of the SCT Format

Despite difficulties with Likert scales, students generally expressed strong appreciation for the SCT format as a learning tool. Many found it stimulated reflection on clinical uncertainty and offered a welcome departure from the right/wrong dichotomy typical of multiple-choice assessments.

Students also praised the opportunity to compare their reasoning to experts, particularly when feedback was provided in structured, explanatory formats. Some noted that this type of test encouraged them to revisit and revise their clinical scripts, reinforcing clinical reasoning pathways rather than isolated facts.

Recommendations for Improvement

Several students offered constructive suggestions aimed at improving the test format and clarity. One prominent comment was the request for more context in the vignettes, such as vital signs, lab values, or imaging findings. Students felt that additional details would allow them to apply reasoning with greater precision and confidence.

Another frequently mentioned point was the desire for guidance or training on how to use the Likert scale. Some students suggested including sample questions or expert rationales before the actual test to calibrate their expectations.

Interestingly, several students highlighted their surprise at the quality of feedback provided by AI systems. These comments were often framed with curiosity or mild skepticism, indicating an openness to AI as a pedagogical tool, contingent on quality and relevance.

While students appreciated the reflective and authentic nature of the format, they also identified areas where clearer structure or training could enhance their experience. Their willingness to accept AI-generated feedback, when well-crafted, suggests that such technologies could play a productive role in clinical education, particularly when embedded within thoughtful educational frameworks. This thematic analysis suggests that future iterations of the SCT should include providing students with explicit guidance and calibration examples on how to interpret the Likert scale; enhancing vignettes with additional contextual cues (eg, vital signs or comorbidities) to support more confident reasoning; incorporating trained AI models as feedback generators, with ensuring faculty review for quality assurance; and designing preparatory activities that help learners understand how to engage with probabilistic reasoning in uncertain clinical scenarios.

Discussion

Principal Findings

This study explores the performance of generative AI in its ability to create SCTs for undergraduate medical students and to embody the role of an expert model by fully designing, scoring, and providing feedback for an SCT in undergraduate medical education. Among the 25 second-year medical students who completed the AI-generated SCT, the mean score was 22.8 out of 32, with a distribution that reflected meaningful variation in clinical reasoning patterns. Students generally aligned well with expert-modeled responses, particularly on items with strong pathophysiological anchors, while questions requiring greater inferential reasoning yielded more dispersed answers. Trained AI systems (those provided with course-specific materials) achieved the highest concordance with both student performance and expected reasoning pathways, offering feedback that was frequently rated as the most helpful. Untrained AI models, in

XSL•FO
RenderX

contrast, were perceived as less pedagogically effective and exhibited weaker correlations with student responses. The SCT also performed reliably, with a Cronbach α of 0.76, and was rated by students as moderately difficult, further reinforcing the test's capacity to capture meaningful distinctions in clinical reasoning. Interestingly, preliminary trends suggested that students with higher SCT scores tended to select feedback from trained AI systems more often, implying that the pedagogical structure of these explanations may resonate most with learners who already demonstrate stronger reasoning frameworks. This observation supports the idea that trained AI feedback could help reinforce expert-like reasoning scripts, though confirmatory analyses in larger samples are needed.

Comparison With Previous Work

The integration of AI in formative clinical assessment tools, such as SCTs, is still at an early stage, but this study offers an early applied example of how LLMs can simulate aspects of expert reasoning within a script-concordance framework when appropriately contextualized. While previous research has already demonstrated the clinical reasoning capabilities of LLMs in diverse testing environments (eg, Nori et al [14] and Singhal et al [6]), this work extends these findings to a formative educational setting by integrating AI-generated items, scoring, and feedback within a single workflow. While previous studies have explored LLMs' ability to answer multiple-choice questions or United States Medical Licensing Examination–style questions, those formats tend to rely on recall and fixed answers [15,16]. In contrast, SCTs demand probabilistic thinking and tolerance of ambiguity, qualities that better reflect real-world medical clinical reasoning and thus could be used in exams to test clinical reasoning. The ability of trained LLMs in this study to mirror the distribution of expert responses (and to provide feedback that students found clear and pedagogically relevant) suggests a step forward in the use of AI not merely as a content generator, but as a cognitive proxy for experienced clinicians.

This study also aligns with findings from recent work examining AI's capacity to replicate reflective thinking patterns. For instance, Nori et al [14] highlighted that LLMs such as GPT-4 are increasingly capable of "meta-cognitive" behavior, adapting their justifications based on context and complexity. The findings support this, especially in items where the trained AI systems offered nuanced rationales that resonated with learners. Furthermore, the indistinguishability between human- and AI-authored feedback, as perceived by students in this study, aligns with previous reports by Lee and Song [17], who found that students often could not discern between AI-generated and expert-written content when both were of high quality. Importantly, students in this cohort demonstrated openness to AI-generated feedback, provided it was contextually accurate and clearly structured, suggesting that AI can be accepted as a trusted educational voice when properly curated.

Nonetheless, this project also illustrates the current limitations of general-purpose LLMs. Untrained models, though capable of generating grammatically correct responses, frequently defaulted to cautious or generic reasoning, especially in questions requiring interpretation of subtle clinical cues. This reduced their pedagogical value and undermined student trust. These findings are consistent with recent critiques from Arora and Arora [18], who argue that untrained LLMs lack the contextual grounding needed for expert-level interpretation, particularly in nuanced or culturally specific clinical scenarios. As such, effective integration of AI into medical education likely hinges on customized fine-tuning, prompt engineering, and ongoing human oversight to ensure that AI explanations remain accurate, relevant, and educationally meaningful.

Limitations

This study has several limitations. The sample size was modest and drawn from a single institution, which may limit generalizability. The test was also formative in nature, and students may not have engaged with the same level of cognitive effort as they would in summative settings. It is also possible that only students who expressed an interest in AI were accepted to participate in this SCT, which may affect the generalizability of the qualitative components of this study. In addition, while the AI models were trained on course materials, their responses still depended on prompt clarity and token limits, which may have shaped their output in subtle ways. No direct comparison was made between AI and human expert panels on the same test, which restricts our ability to determine whether the observed concordance between AI and student reasoning reflects genuine expert-level clinical judgment. As such, these findings demonstrate internal validity and feasibility rather than equivalence to clinician consensus. Future studies should include parallel human-expert panels to establish criterion validity and to calibrate the pedagogical quality of AI-generated scoring keys and feedback. As this was a single-round implementation, long-term effects on clinical reasoning or retention were not assessed. Future studies should explore how repeated exposure to AI-generated SCTs impacts learning trajectories and whether hybrid models (AI and human feedback) offer added value over AI alone.

Finally, a further limitation relates to the potential for hallucinations or inaccurate information in AI-generated explanations. While all outputs were reviewed for face plausibility and internal consistency, no formal content validation against gold-standard references was performed. As highlighted by Masters [19], generative AI systems can produce fabricated or misleading references and medical details, a phenomenon known as "AI hallucination." This risk underscores the importance of treating AI-generated feedback as supplementary rather than authoritative, especially in formative educational contexts. Future implementations should incorporate systematic expert review of AI-generated feedback to ensure clinical accuracy and safeguard learners from misinformation.

Conclusions

As AI technologies evolve and become increasingly embedded in medical education, this study offers novel insight into how LLMs can assume the role traditionally occupied by human experts in clinical reasoning assessments. By designing and scoring an SCT entirely with generative AI systems, and by leveraging both trained and untrained models to simulate expert feedback, we demonstrated that AI can produce assessment content that is not only functionally valid but also well-received by learners. The high degree of alignment between student

XSL•FO
RenderX

responses and trained AI-generated reference panels, alongside the positive reception of AI-authored feedback, suggests that these technologies hold real promise for supporting formative assessment. While distinctions remain between the effectiveness of trained versus untrained models, this study illustrates that with proper calibration, AI systems can enhance both the efficiency and pedagogical depth of clinical education. These findings open the door for further research into scalable, AI-assisted assessment models that can flexibly support reasoning under uncertainty, which is an essential competency in modern medical practice. Beyond feasibility, this work underscores the transformative potential of generative AI to democratize access to high-quality formative assessment. In academic environments where expert availability and time are limited, AI-assisted SCT generation could substantially reduce faculty workload, expand curricular coverage, and enable rapid adaptation of tests to evolving learning objectives. By lowering the resource barrier traditionally associated with expert panel-based assessment, this approach positions AI as a potential catalyst for sustainable, scalable innovation in medical education globally.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Script concordance test (French and English).
[DOCX File, 49 KB - formative_v9i1e76618_app1.docx]

References

1. Schuwirth LWT, Van der Vleuten CPM. Programmatic assessment: from assessment of learning to assessment for learning. Med Teach 2011;33(6):478-485. [doi: 10.3109/0142159X.2011.565828] [Medline: 21609177]
2. Charlin B, Roy L, Brailovsky C, Goulet F, van der Vleuten C. The script concordance test: a tool to assess the reflective clinician. Teach Learn Med 2000;12(4):189-195. [doi: 10.1207/S15328015TLM1204_5] [Medline: 11273368]
3. Lubarsky S, Charlin B, Cook DA, Chalk C, van der Vleuten CPM. Script concordance testing: a review of published validity evidence. Med Educ 2011 Apr;45(4):329-338. [doi: 10.1111/j.1365-2923.2010.03863.x] [Medline: 21401680]
4. Dory V, Gagnon R, Vanpee D, Charlin B. How to construct and implement script concordance tests: insights from a systematic review. Med Educ 2012 Jun;46(6):552-563. [doi: 10.1111/j.1365-2923.2011.04211.x] [Medline: 22626047]
5. Gagnon R, Charlin B, Coletti M, Sauvé E, van der Vleuten C. Assessment in the context of uncertainty: how many members are needed on the panel of reference of a script concordance test? Med Educ 2005 Mar;39(3):284-291. [doi: 10.1111/j.1365-2929.2005.02092.x] [Medline: 15733164]
6. Singhal K, Azizi S, Tu T, et al. Publisher correction: large language models encode clinical knowledge. Nature New Biol 2023 Aug 10;620(7973):E19-E19. [doi: 10.1038/s41586-023-06455-0]
7. Jeblick K, Schachtner B, Dexl J, et al. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. Eur Radiol 2024 May;34(5):2817-2825. [doi: 10.1007/s00330-023-10213-1] [Medline: 37794249]
8. Abd-Alrazaq A, AlSaad R, Alhuwail D, et al. Large language models in medical education: opportunities, challenges, and future directions. JMIR Med Educ 2023 Jun 1;9:e48291. [doi: 10.2196/48291] [Medline: 37261894]
9. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
10. Kıyak YS, Emekli E. Using large language models to generate script concordance test in medical education: ChatGPT and Claude. Rev Esp Edu Med 2024;6(1). [doi: 10.6018/edumed.636331]
11. Hudon A, Kiepura B, Pelletier M, Phan V. Using ChatGPT in psychiatry to design script concordance tests in undergraduate medical education: mixed methods study. JMIR Med Educ 2024 Apr 4;10:e54067. [doi: 10.2196/54067] [Medline: 38596832]
12. Choi GW, Kim SH, Lee D, Moon J. Utilizing generative AI for instructional design: exploring strengths, weaknesses, opportunities, and threats. TechTrends 2024 Jul;68(4):832-844. [doi: 10.1007/s11528-024-00967-w]
13. Fournier JP, Demeester A, Charlin B. Script concordance tests: guidelines for construction. BMC Med Inform Decis Mak 2008 May 6;8:18. [doi: 10.1186/1472-6947-8-18] [Medline: 18460199]
14. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. arXiv. Preprint posted online on Apr 12, 2023. [doi: 10.48550/arXiv.2303.13375]
15. Luo D, Liu M, Yu R, et al. Evaluating the performance of GPT-3.5, GPT-4, and GPT-4o in the Chinese National Medical Licensing Examination. Sci Rep 2025;15(1):14119. [doi: 10.1038/s41598-025-98949-2]

XSL•FO
RenderX

16. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 8;9:e45312. [doi: 10.2196/45312] [Medline: 36753318]
17. Lee S, Song KS. Teachers' and students' perceptions of AI-generated concept explanations: implications for integrating generative AI in computer science education. Computers and Education: Artificial Intelligence 2024 Dec;7:100283. [doi: 10.1016/j.caeai.2024.100283]
18. Arora A, Arora A. The promise of large language models in health care. The Lancet 2023 Feb;401(10377):641. [doi: 10.1016/S0140-6736(23)00216-7]
19. Masters K. Medical Teacher's first ChatGPT's referencing hallucinations: lessons for editors, reviewers, and teachers. Med Teach 2023 Jul;45(7):673-675. [doi: 10.1080/0142159X.2023.2208731] [Medline: 37183932]

Abbreviations

AI: artificial intelligence
COPD: chronic obstructive pulmonary disease
LLM: large language model
SCT: script concordance test

Edited by A Mavragani; submitted 27.04.25; peer-reviewed by RT Potla, SB Guo, YS Kiyak; revised version received 19.10.25; accepted 10.11.25; published 20.11.25.

Please cite as:
Hudon A, Phan V, Charlin B, Wittmer R
Teaching Clinical Reasoning in Health Care Professions Learners Using AI-Generated Script Concordance Tests: Mixed Methods Formative Evaluation
JMIR Form Res 2025;9:e76618
URL: https://formative.jmir.org/2025/1/e76618
doi:10.2196/76618

© Alexandre Hudon, Véronique Phan, Bernard Charlin, René Wittmer. Originally published in JMIR Formative Research (https://formative.jmir.org), 20.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Evaluating the Effectiveness of Generative AI for the Creation of Patient Education Materials on Coronary Heart Disease: A Comparative Study

Xiaofang Jiang*, RN, MSc; Jingbang Liu*, RN; Shanshan Dai, RN; Xiawen Mao, RN; Rongping Cha, RN, MSc; Lili Wu, RN, MSc

Nursing Department, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, 3 East Qingchun Road, Hangzhou, Zhejiang Province, China

*these authors contributed equally

Corresponding Author:
Lili Wu, RN, MSc
Nursing Department, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, 3 East Qingchun Road, Hangzhou, Zhejiang Province, China

Abstract

Background: Generative artificial intelligence (AI) has shown great potential in various fields, including health care. However, its application for developing patient education materials (PEMs), particularly for those with coronary heart disease (CHD), remains underexplored. Traditional methods for creating these materials are time-consuming and lack personalization, which limit their effectiveness.

Objective: This study aims to explore the effectiveness of generative AI tools (ChatGPT and DeepSeek) at generating PEMs for patients with CHD and to compare them with materials developed by a professional medical team.

Methods: In February 2025, PEMs for patients with CHD were developed using a framework designed by a professional medical team. Structured prompts were used to generate materials through 2 generative AI models—ChatGPT-4o and DeepSeek R1. These AI-generated materials were compared with those created by the medical team in terms of development time, readability, understandability, actionability, and accuracy.

Results: The total time for manual preparation was 14 hours, while ChatGPT and DeepSeek consumed 0.62 hours and 0.78 hours, respectively. Regarding readability, the frequency of difficult words was more variable in manually written and ChatGPT materials, while DeepSeek showed more consistency. The proportion of simple sentences was highest with DeepSeek, followed by ChatGPT, with complete separation between manually written and ChatGPT (δ=1). Content word frequency was highest in manually written PEMs, while ChatGPT had the lowest but most stable values. Personal pronouns were most frequently used in manually written PEMs, with high variability, and least used in DeepSeek, which was stable. All 3 methods had similar readability levels and reached Chinese elementary school–level readability for the proportions of simple sentences and personal pronouns, with high school–level difficulty of words and content word frequency. The understandability and actionability scores were above 70, with ChatGPT being more stable for understandability and DeepSeek being more stable for actionability. No significant differences were found between groups. In terms of accuracy, intergroup comparisons showed significant differences (H=7.27, P=.03) but no significant differences in multiple comparisons. The direct comparison between ChatGPT and DeepSeek showed a negligible effect size (δ=0.02), with no significant difference (z-score=−0.06, P=.96). Accuracy issues in the AI-generated materials were noted by 4 of 8 experts.

Conclusions: Generative AI significantly improved the efficiency of developing PEMs for patients with CHD. The materials generated by ChatGPT-4o and DeepSeek R1 were comparable to the professionally written ones in terms of readability, understandability, and actionability. However, improvements related to reducing the number of difficult words and increasing content word frequency are needed to enhance readability. The accuracy of AI-generated materials still poses concerns, including potential AI "hallucinations," and requires review by health care professionals. Generative AI holds considerable potential for generating PEMs, and future research should assess its applicability and effectiveness in real-world patient and family contexts.

(JMIR Form Res 2025;9:e78816) doi:10.2196/78816

KEYWORDS

generative artificial intelligence; large language models; coronary heart disease; patient education; ChatGPT; DeepSeek

XSL•FO
RenderX

Introduction

Coronary heart disease (CHD), or coronary atherosclerotic heart disease, remains one of the leading causes of morbidity and mortality worldwide [1]. According to the 2019 Global Burden of Disease study, approximately 17.9 million people worldwide died from cardiovascular diseases in 2019, accounting for 32% of all global deaths, demonstrating the continued rise in the global burden of cardiovascular diseases [2]. Although revascularization techniques have significantly improved survival rates, the health literacy of patients with CHD remains relatively low, posing a significant barrier to effective disease management and patient empowerment [3]. The key to preventing cardiovascular diseases lies in reducing controllable risk factors. Public health strategies include reducing risk factors at the population level, as well as implementing primary and secondary prevention and treatment at the individual level [4]. Health education is a vital component for improving patient outcomes by increasing compliance with treatment regimens, enhancing self-management, and fostering informed decision-making [5]. However, traditional methods of developing patient education materials (PEMs) rely heavily on manual processes [6-8], which are time-consuming and require substantial input from health care providers [9,10]. These manual materials are often generic and may lack personalization, making it difficult to tailor information to the specific needs of individual patients. Additionally, the materials may not reflect the latest scientific evidence or guidelines due to a lack of rapid updating, which could hinder the dissemination of up-to-date medical information to patients.

Generative artificial intelligence (AI) has emerged as a promising solution to address these pain points. Powered by advancements in machine learning and large language models, generative AI enables the rapid creation of content that can be tailored to specific patient needs [11]. In health care, generative AI has shown potential in a range of applications, such as generating medical records, enhancing clinical decision support systems, and supporting medical education [12-14]. Certain applications have also been implemented for providing medical consultation and education; have been used to generate or optimize the PEMs for diseases in departments such as otolaryngology, dermatology, and gastroenterology; and have been preliminarily validated [8-10,13].

The release of DeepSeek R1 in January 2025 accelerated the adoption of generative AI tools [15]. Some international studies have examined and validated DeepSeek's effectiveness at supporting public cardiopulmonary resuscitation training [16]. Nevertheless, research using DeepSeek for the development of Chinese-language PEMs remains scarce. Some studies have evaluated various generative AI models, particularly ChatGPT, for generating PEMs [17,18]. Compared with ChatGPT, the capabilities of DeepSeek have not yet been widely explored in the context of PEMs, especially for chronic diseases such as CHD. In-depth comparisons and evaluations of generative AI tools are still lacking in this domain. Generative AI technologies can algorithmically generate logically structured and instructive content in a short period of time, offering new opportunities to enhance the efficiency and innovation of PEM development. However, issues such as hallucinations and biases may lead to misinformation, posing risks and challenges to patient safety and medical ethics [15,19].

This study seeks to fill this gap by comparing the quality of PEMs for patients with CHD generated by 2 generative AI tools: ChatGPT-4o and DeepSeek R1. ChatGPT-4o is based on the GPT-4 architecture and was pretrained on a large corpus of open-text data including news articles, books, and online content. It demonstrates strong generative capabilities across a wide range of language tasks, but it may exhibit certain hallucination phenomena when generating information [20]. DeepSeek R1, a deep learning model with reasoning capabilities, offers unique advantages for logical reasoning and content generation in professional fields. Compared with ChatGPT-4o, DeepSeek R1 is more focused on handling tasks that require deep logical analysis [21]. However, due to its more complex reasoning process, DeepSeek R1 tends to have a slower generation speed. These 2 models, based on different technical architectures and training strategies, have distinct strengths and limitations. This study aimed to compare the quality of CHD-related PEMs generated by these 2 generative AI tools with those developed by a professional health care team, providing empirical evidence to inform the application and promotion of generative AI in PEM development.

Methods

Overview

This project is a comparative study, with samples consisting of 3 sets of PEMs for CHD: 1 set manually written by a professional health care team and 2 sets generated by ChatGPT-4o (ChatGPT) and DeepSeek R1 (DeepSeek). The quality differences between the 3 groups were evaluated by assessing the preparation time, readability, understandability, actionability, and accuracy of the PEMs.

Each material set consisted of 3 PEMs routinely developed by the cardiology health education team, namely "Coronary Heart Disease Knowledge," "Coronary Artery Minimally Invasive Interventional Therapy," and "Common Questions About Coronary Artery Minimally Invasive Interventional Therapy." This study did not involve direct participation of patients, and patient-related information or data were not included.

Method for Developing Manually Written PEMs

Due to the absence of an authoritative official website in China for downloading standardized PEMs, each hospital develops PEMs according to its specific needs. In our hospital, the PEMs were developed following the hospital's established process for PEMs [22]. The development team consisted of experienced cardiologists, a nurse educator, and health educators. Initially, the team conducted a needs assessment and followed the hospital's guidelines for PEM development. The materials were then drafted based on a review of relevant literature [23-25] and discussions within the group. Afterward, the draft was submitted to the hospital's health promotion nurse for review, who provided feedback on the content and structure. The revised draft was then shared with 10 patients or family members for further feedback. After collecting their suggestions, the draft

XSL•FO
RenderX

was revised again and ultimately reviewed and finalized by senior cardiology specialists with high-level professional titles. The final versions of the 3 materials on CHD education are listed in Table 1.

Table . List of 3 patient education materials (PEMs) for coronary heart disease (CHD).

PEM name	Contents
Coronary Heart Disease Knowledge	1. What is CHD? 2. What is the current status of CHD? What are its dangers? 3. What are the risk factors for CHD? 4. How do atherosclerotic plaques form? 5. Does CHD always involve chest pain? What are the symptoms of CHD? 6. What are the current treatment options for CHD?
Coronary Artery Minimally Invasive Interventional Therapy	1. What is coronary angiography (CAG)? 2. What is percutaneous coronary intervention (PCI)? 3. What are the types of stents? 4. Who needs coronary artery intervention therapy? 5. What preoperative tests are needed before coronary intervention therapy? 6. What preparations are needed before coronary intervention therapy? 7. What are the specific steps of coronary artery intervention therapy? 8. What sensations will you feel during the procedure? 9. What should you pay attention to after coronary intervention therapy? 10. What should you do after discharge?
Common Questions About Coronary Artery Minimally Invasive Interventional Therapy	1. Is everything fine after coronary intervention therapy? 2. Can the stent fall off? 3. Can patients resume normal work and life after coronary intervention therapy? 4. If I feel fine, can I stop taking antiplatelet drugs? 5. Can I undergo MRI[a] or CT[b] scans after coronary artery stent implantation? 6. What should I pay attention to when traveling after coronary intervention therapy? 7. What should be done in the case of a heart attack?

[a]MRI: magnetic resonance imaging.

[b]CT: computed tomography.

Method for Developing AI-Generated PEMs

All PEMs generated by ChatGPT and DeepSeek were created on February 17, 2025. To ensure the accuracy and relevance of the generated content, a nursing expert with an Artificial Intelligence Generated Content Mentor-level certificate from the Ministry of Industry and Information Technology, along with a health promotion nurse in health education, collaboratively developed the structured prompts [26] and optimized the prompts through follow-up questions, examples, and other methods. The prompts (see Multimedia Appendix 1) were adapted from previous studies based on a certain structure and further optimized and adjusted by our team before being used to generate the final content, to ensure their effectiveness [27].

The structured prompts included key components such as role setting, content background, specific tasks, and detailed requirements. The structured elements covered the themes, target audience, objectives, and a manual table of contents that aligned with the structure of the materials manually created by the health education team. Additionally, the prompts included the necessary background knowledge, and restrictions were applied to the language style in order to enhance readability and ensure that the language was accessible to the target population. Once the prompts were finalized and optimized, they were input into the 2 generative AI models with the file “List of Three PEMs for CHD.” Both models independently generated the complete set of PEMs for patients with CHD, and the set of PEMs was then compiled into a finalized format. We did not make any modifications to the final content. The results generated during the prompt optimization process, due to their low quality, were discarded, and only the final version was retained for comparison.

Evaluation Metrics

The evaluation criteria included 5 aspects: time spent, readability, understandability, actionability, and accuracy of PEMs for CHD.

Time Spent

The time spent developing manually written PEMs (measured in hours) included the time spent on needs assessment by the specialized team, literature review, group discussions, and the drafting of the initial materials. It also included the time for the health promotion nurse to review the materials and provide feedback; the time for 10 patients or family members to read and provide feedback, followed by revisions based on their

XSL•FO
RenderX

input; and finally, the time spent by senior specialists in the field for the final review and revision of the materials. The time was collected by the cardiology health educator.

As for AI-generated PEMs, the researchers used a timer to record the total time (in minutes) from the start to the completion of each set of PEMs. This included the time for writing the structured prompts, optimizing and modifying the prompts, generating and refining the content, and organizing the materials into their final format. The time was also collected by the cardiology health educator.

Readability

Text readability refers to the difficulty or ease with which a text can be read and understood. It is an important linguistic characteristic of text and is widely used in fields such as information science, journalism, and health education [28,29]. The Chinese Readability Indicator and Evaluator (CRIE) system is an automated tool for analyzing the readability of Chinese texts [30,31]. It primarily analyzes 4 core linguistic features: the number of difficult words, the proportion of simple sentences, the logarithmic value of content word frequency, and the number of personal pronouns.

Chinese and alphabetic languages have significant morphological differences. In Chinese, characters are formed by combining strokes and components, and words are made up of individual characters, each of which is monosyllabic. Therefore, word length metrics commonly used in alphabetic languages, such as the number of letters or syllables, are completely ineffective when analyzing Chinese. CRIE defines words not included in the common 3000 high-frequency Chinese word list as difficult words. Difficult words reflect the complexity of the text's vocabulary: the more difficult words there are, the higher the reading difficulty and the lower the readability.

A simple sentence refers to a sentence without complex subordinate clauses. The proportion of simple sentences refers to the percentage of simple sentences in the text, calculated as the number of simple sentences divided by the total number of sentences. A higher proportion of simple sentences indicates lower reading difficulty and higher readability.

The logarithmic value of content word frequency measures the average frequency of content words (including nouns, verbs, and adjectives) in the text. A larger logarithmic value indicates a higher average frequency of content words, meaning these words are repeated more often, making them more familiar to readers, reducing cognitive load, and increasing text readability. Conversely, a lower content word frequency suggests the use of a broader vocabulary, requiring readers to constantly adapt to new content words, which increases cognitive load and decreases readability. For example, in the case of coronary heart disease, the term "risk factors" would indicate a high content word frequency if used consistently throughout the text, suggesting lower cognitive load and higher readability. However, if synonyms like "harmful factors" or "high-risk factors" are continuously introduced, it would indicate a lower content word frequency, greater cognitive load, and lower readability.

The number of personal pronouns refers to the frequency of personal pronouns in the text. Excessive use of personal pronouns may affect the coherence of the text and the clarity of references, increasing the reading difficulty and decreasing readability.

The 3 sets of PEMs for patients with CHD were analyzed using this system, comparing the 4 core linguistic features across the materials. The analysis assessed the readability of the 3 sets of materials and their corresponding grade levels. The system demonstrated accuracy rates of 93% for word analysis and 86% for syntactic analysis, and it demonstrated grade classification accuracy rates of 72.92% for primary and secondary school textbooks, 75% for second language texts, and 80.83% for natural science texts.

Understandability and Actionability

In this study, the Patient Education Materials Assessment Tool for Printable Materials (PEMAT-P) was used to evaluate the effectiveness and applicability of PEMs. This tool was developed by the Agency for Healthcare Research and Quality in 2012 [32]. Compared with earlier tools, PEMAT-P is designed to be more user-friendly [33,34]. The Chinese version of PEMAT-P was adapted by Yu et al in 2023 [35]. The adapted tool includes 2 subscales—understandability and actionability—with a total of 7 dimensions and 23 items. Specifically, 16 items assess understandability, and 7 items evaluate actionability. Response options include "Agree" (1 point), "Disagree" (0 points), and "Not Applicable (N/A)," with "Not Applicable" items excluded from the final score. The total score ranges from 0 to 100, with higher scores indicating better understandability or actionability. A score greater than 70 indicates that the material has good understandability or actionability. The formula for calculating the Chinese version of PEMAT-P is as follows: PEMAT-P score = (total score/total number of items) × 100. The scale demonstrated high reliability, with a scale content validity index of 0.970, a Kaiser-Meyer-Olkin value of 0.655, and a Bartlett test of sphericity χ^2 value of 710.200 (P<.001). The overall Cronbach α coefficient was 0.754, with Cronbach α for the understandability subscale at 0.802 and for the actionability subscale at 0.615.

We invited 8 senior experts in health education and cardiology to assess the materials using the Chinese version of the PEMAT-P. The 8 experts had a mean age of 45.00 (SD 4.41) years. Among them, 4 held senior professional titles, 3 had associate senior titles, and 1 had an intermediate title. Their educational backgrounds included 2 with doctorates, 3 with master's degrees, and 3 with bachelor's degrees. The group included 2 cardiologists, 3 cardiology nurses, and 3 health education experts (including 2 from public health and 1 from a hospital). The mean length of professional experience of the experts was 16.25 (SD 7.38) years. Of the experts, 5 were listed in the provincial-level health education expert database. The 3 sets of PEMs were randomly coded by the researchers and sent to the experts via email, along with the expert evaluation letter, the expert background and authority questionnaire, the PEMAT-P and its user manual, and the 3 sets of PEMs for CHD. Experts were informed to return the results to the researchers'

XSL•FO
RenderX

emails within 1 week. The evaluation letter included the researchers' emails and phone numbers, and the experts were advised that they could contact the researchers if they had any questions. The response rate for the expert inquiry was 100%, with an expert authority coefficient of 0.877.

Accuracy

Accuracy is one of the key indicators used to assess the quality of PEMs [36]. In this study, a Likert 6-point scale was used to assess accuracy [37], where a score of 1 represented completely incorrect, 2 indicated more incorrect than correct, 3 meant equal elements of correct and incorrect, 4 represented more correct than incorrect, 5 indicated almost entirely correct, and 6 signified completely correct. The accuracy of the 3 sets of PEMs for patients with CHD was evaluated by 8 experts. The experts were asked to provide written descriptions in the comment section if they identified any specific accuracy issues with the materials. The response rate for the expert inquiry was 100%.

Statistical Methods

Data entry and verification were conducted by 2 trained team members using Excel 2021 (Microsoft Corp). Statistical analysis was performed using SPSS version 25 (IBM Corp) and R 4.5.1. The Shapiro-Wilk test was used to assess data normality, and the Levene test was used to check the homogeneity of variances. Due to the very small sample size, readability analysis is presented descriptively using median, minimum, and maximum values; strip plots; and box plots. Cliff delta effect size δ and its 95% CI were used for comparisons between 2 groups. Understandability and actionability were described using means and SDs, with intergroup comparisons conducted using the *F* test. Multiple comparisons were made using the least significant difference test, and the comparisons between ChatGPT and DeepSeek were performed using *t* tests. Effect sizes were calculated using Cohen *d*, corrected with Hedges *g*. For accuracy analysis, due to non-normal distribution of the data, median and interquartile ranges were used. Intergroup comparisons were conducted using the Wilcoxon rank-sum test, multiple comparisons were made using the Kruskal-Wallis 1-way ANOVA test (for k samples), and the comparison between ChatGPT and DeepSeek was conducted using the Wilcoxon rank-sum test. Effect sizes are presented using δ and its 95% CI.

Ethical Considerations

The local institutional review board adheres to the Declaration of Helsinki. After consultation with the board, it was determined that no formal ethical approval was required for this study, as no human or animal participants were involved. The study was designed to ensure ethical standards were maintained, and after providing informed consent, all professionals participated in the evaluation voluntarily and were not compensated. We implemented strict confidentiality measures, and access to the data was only provided to the research team.

Results

Comparison of Development Time

The specific development times for manually written PEMs were as follows: 1 hour for the needs assessment by the specialized team, 2 hours for literature review, 1 hour for group discussions, and 3 hours for drafting the initial materials, totaling 7 hours. The time for the hospital's health promotion nurse to review and provide feedback was 0.5 hours, followed by 6 hours for 10 patients or family members to read the materials and provide feedback for revisions. Finally, 0.5 hours were spent by senior specialists to review and revise the materials, bringing the total development time for the control group to 14 hours.

As for AI-generated PEMs, the time for writing and optimizing the prompts was 15 minutes. The time for using ChatGPT to generate and optimize the materials was 17 minutes, and 5 minutes were used to organize the final version. For DeepSeek, the time for generating and optimizing the materials was 25 minutes, and 7 minutes were spent on organization. The total time for generating materials using ChatGPT was 0.62 hours, and for DeepSeek, it was 0.78 hours, with an average time of 0.7 hours. The time spent on manual development was 20 times longer than the time required by generative AI.

Comparison of Readability

For the number of difficult words, both the manually written and ChatGPT materials showed extremely high values (737 and 740, respectively), resulting in very long whiskers in the box plots and indicating high variability. This suggests large fluctuations in the use of difficult words. In contrast, DeepSeek data were more concentrated and stable, with no outliers, indicating tighter control over vocabulary difficulty. The Cliff delta effect size δ between ChatGPT and DeepSeek indicated a medium effect; however, the 95% CI included 0, suggesting high uncertainty in the difference due to the small sample size.

Regarding the proportion of simple sentences, DeepSeek was higher overall than the other 2 groups, with a tightly clustered distribution and the highest proportion of simple sentences, suggesting a tendency to enhance readability through simpler sentence structures. ChatGPT had the lowest proportion, indicating more complex sentence construction. The proportion of simple sentences in the manually written materials was intermediate and showed a tight distribution. The data between manually written and ChatGPT materials were completely separated, while the 95% CI for δ between manually written and DeepSeek materials included 0, indicating high uncertainty. A large effect size was observed between ChatGPT and DeepSeek.

For content word frequency, manually written materials had the highest value but with an extreme outlier (1.253), suggesting the highest repetition of content words but also instability. ChatGPT had the lowest and most stable values, while DeepSeek was moderate and stable. This indicates that generative AI models—particularly ChatGPT—tended to use a wider variety of content words, which may increase cognitive load and reduce readability. All pairwise comparisons among the 3 groups showed complete separation.

XSL•FO
RenderX

In terms of personal pronoun usage, manually written materials used the most, with high variability—possibly due to narrative needs. DeepSeek used the fewest and was the most stable, while ChatGPT was intermediate. This suggests that generative AI models, especially DeepSeek, tend to use fewer personal pronouns to improve readability. However, the 95% CIs of δ for all pairwise comparisons included 0, indicating high uncertainty.

Overall, DeepSeek demonstrated stability in controlling difficult words, simple sentence proportions, and personal pronoun usage. ChatGPT had a lower proportion of simple sentences, more complex syntax, and the lowest content word frequency, reflecting more diverse vocabulary usage. Manually written materials had the highest content word frequency and lowest cognitive load but showed high variability and risk of outliers. See Figure 1 and Table 2.

Figure 1. Strip plots and box plots of the 3 methods to generate patient education materials (PEMs) for coronary heart disease (CHD): (A) number of difficult words, (B) proportion of simple sentences, (C) logarithmic value of content word frequency, (D) number of personal pronouns.

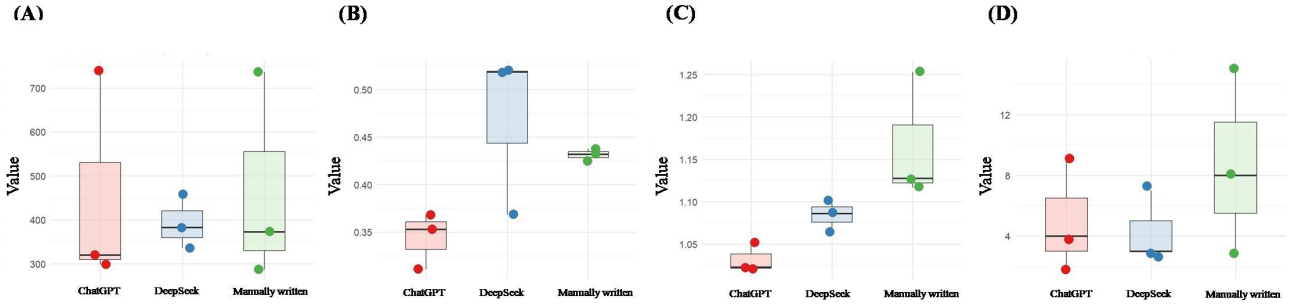

Table . Comparison of readability scores among the 3 methods of generating patient education materials (PEMs) for coronary heart disease (CHD).

Readability	Manually written		ChatGPT		DeepSeek		Manually written versus ChatGPT, δ (95% CI)	Manually written versus DeepSeek, δ (95% CI)	ChatGPT versus DeepSeek, δ (95% CI)
	Median (range)	Mean	Median (range)	Mean	Median (range)	Mean			
Difficult words, n	373 (287-737)	465.67	320 (298-740)	452.67	383 (336-458)	392.33	−0.11 (−0.79 to 0.69)	−0.11 (−0.88 to 0.82)	−0.33 (−0.95 to 0.81)
Proportion of simple sentences	0.43 (0.43-0.44)	0.43	0.35 (0.31-0.37)	0.34	0.52 (0.37-0.52)	0.47	1.00 (0.14 to 1.00)	−0.33 (−0.91 to 0.68)	−0.89 (−0.99 to –0.23)
Content words, n	1.13 (1.12-1.25)	1.17	1.02 (1.02-1.05)	1.03	1.09 (1.07-1.10)	1.08	1.00 (0.14 to 1.00)	1.00 (0.14 to 1.00)	−1.00 (−1.00 to −0.14)
Personal pronouns, n	8 (3-15)	8.67	4 (2-9)	5.00	3 (3-7)	4.33	0.33 (−0.68 to 0.91)	0.56 (−0.62 to 0.96)	0.11 (−0.82 to 0.88)

The proportions of simple sentences and personal pronoun usage across all 3 PEM groups aligned with the reading level of Chinese elementary school students. However, both the difficult word frequency and content word frequency corresponded to a high school reading level, exceeding the commonly recommended 6th-grade level for PEMs, indicating that neither manually written nor AI-generated PEMs met the target readability level in terms of vocabulary complexity.

Comparison of Understandability and Actionability

The mean scores for understandability and actionability of all 3 sets of PEMs for CHD were higher than 70, indicating that both the manually written and AI-generated PEMs demonstrated good levels of understandability and actionability.

In terms of understandability, the SDs were relatively large for the “Coronary Heart Disease Knowledge“ section in the manually written materials and for the “Coronary Artery Minimally Invasive Interventional Therapy” and “Common Questions About Coronary Artery Minimally Invasive Interventional Therapy” sections in the DeepSeek-generated materials, indicating a higher degree of data dispersion.

For actionability, higher SDs were observed in the “Coronary Heart Disease Knowledge” and “Common Questions About Coronary Artery Minimally Invasive Interventional Therapy” sections in the manually written materials, as well as in the “Coronary Heart Disease Knowledge” and “Coronary Artery Minimally Invasive Interventional Therapy” sections generated by ChatGPT. These findings suggest that ChatGPT was more stable in understandability, while DeepSeek showed greater stability in actionability (see Figures 2 and 3). Intergroup and multiple comparisons among the 3 PEMs showed no statistically significant differences in understandability nor actionability. Direct comparisons between PEMs generated by ChatGPT and DeepSeek indicated no statistically significant differences.

XSL•FO
RenderX

Figure 2. Mean (SD) understandability scores for the 3 methods of generating patient education materials (PEMs) for coronary heart disease (CHD) for each of the 3 PEM sections: (A) Coronary Heart Disease Knowledge, (B) Coronary Artery Minimally Invasive Interventional Therapy, (C) Common Questions About Coronary Artery Minimally Invasive Interventional Therapy.

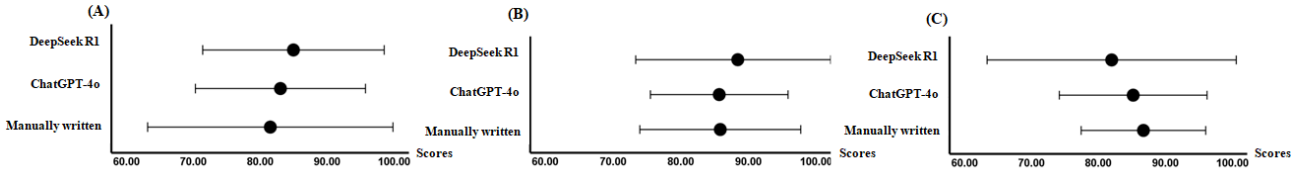

Figure 3. Mean (SD) actionability scores for the 3 methods of generating patient education materials (PEMs) for coronary heart disease (CHD) for each of the 3 PEM sections: (A) Coronary Heart Disease Knowledge, (B) Coronary Artery Minimally Invasive Interventional Therapy, (C) Common Questions About Coronary Artery Minimally Invasive Interventional Therapy.

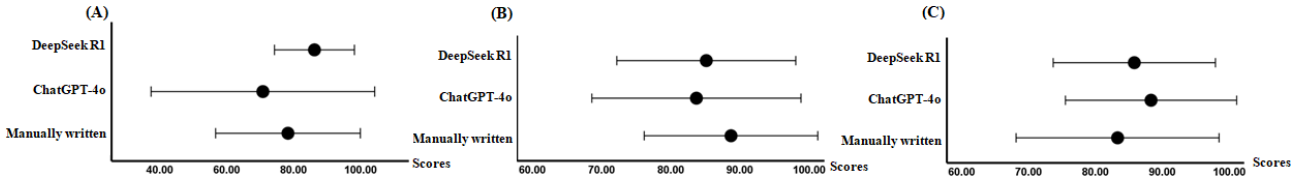

These findings suggest that the understandability and actionability levels of CHD PEMs developed using the 3 methods are comparable and that the quality of PEMs generated by the 2 generative AI models is also equivalent in these dimensions (see Table 3).

XSL•FO
RenderX

Table . Comparison of understandability and actionability scores of the 3 methods of generating patient education materials (PEMs) for coronary heart disease (CHD), as evaluated by 8 experts.

PEM section	Manually written, mean (SD)	ChatGPT, mean (SD)	DeepSeek, mean (SD)	*F* (*df*)[a]	*P* value[a]	Hedges *g*[b] (95% CI)	*t*[b] (*df*)	*P*[b] value
Understandability								
Coronary Heart Disease Knowledge	81.52 (18.33)	83.04 (12.71)	84.98 (13.56)	0.11 (2,21)	.90	−0.14 (−1.15 to 0.88)	−0.30 (14)	.77
Coronary Artery Minimally Invasive Interventional Therapy	85.84 (11.88)	85.69 (10.16)	88.42 (15.07)	0.12 (2,21)	.89	−0.20 (−1.22 to 0.82)	−0.43 (14)	.68
Common Questions About Coronary Artery Minimally Invasive Interventional Therapy	86.73 (9.30)	85.19 (11.03)	81.99 (18.60)	0.25 (2,21)	.78	0.20 (−0.82 to 1.21)	0.42 (14)	.68
Actionability								
Coronary Heart Disease Knowledge	78.30 (21.63)	70.80 (33.43)	86.22 (11.95)	0.83 (2,21)	.45	−0.58 (−1.62 to 0.45)	−1.23 (14)	.24
Coronary Artery Minimally Invasive Interventional Therapy	88.72 (12.54)	83.72 (15.11)	85.15 (12.93)	0.29 (2,21)	.75	−0.10 (−1.11 to 0.92)	−0.20 (14)	.84
Common Questions About Coronary Artery Minimally Invasive Interventional Therapy	83.30 (15.17)	88.30 (12.80)	85.80 (12.12)	0.28 (2,21)	.76	0.19 (−0.83 to 1.21)	0.40 (14)	.69

[a]Comparisons among the 3 groups.

[b]Direct comparison between ChatGPT and DeepSeek.

Accuracy Results

The accuracy scores indicate that manually written materials had the highest accuracy score (median=6, 25th to 75th percentile=5-6). Comparing the accuracy scores among the 3 sets of PEMs (ChatGPT: median=5, 25th to 75th percentile=4-5; DeepSeek: median=5, 25th to 75th percentile=4-6), the intergroup difference was statistically significant (Hedges *g*=7.27, *P*=.03; δ=0.02, 95% CI –0.51 to 0.54; *z*-score=–0.06, *P*=.96). However, after Bonferroni correction for multiple comparisons, the adjusted significance showed no statistically significant differences between any pairwise comparisons. For the comparison between ChatGPT and DeepSeek, the effect size was negligible, indicating no significant difference.

Among the 8 expert reviewers, 4 noted that the accuracy of the materials generated by ChatGPT-4o and DeepSeek R1 had issues. Two experts did not specify specific details but described the issues as follows: “Materials 2 (generated by ChatGPT) and 3 (generated by DeepSeek) contain some outdated content and lack scientific rigor, suggesting that professionals should further review them” and “Materials 2 and 3 contain some imprecise educational content or are not updated according to the latest guidelines/expert consensus, making them somewhat outdated and deviating from current clinical practices.” The other 2 experts pointed out specific problems: “Material 2 has outdated information (eg, the recommended daily salt intake was updated from 6 g to 5 g, but the material still mentions 6 g)” and “In Material 3 - 1, it states that ticagrelor should be used with aspirin for 12 months, but in Material 3 - 2, it mentions that ticagrelor should be continued for 12 - 18 months post-stent placement. The duration of ticagrelor use is unclear in the material, which may confuse patients.” The frequencies of errors mentioned by the experts were 0% (0/8) for the manually written materials, 38% (3/8) for ChatGPT, and 38% (3/8) for DeepSeek.

XSL•FO
RenderX

Discussion

Principal Findings

Manually Written Versus Generative AI

This study demonstrated that the time required to create PEMs for patients with CHD was 14 hours when completed manually, while the use of ChatGPT took 0.62 hours and DeepSeek took 0.78 hours. These results suggest that generative AI can significantly improve the efficiency of developing PEMs for patients with CHD. By using structured prompts and leveraging contextual learning capabilities, generative AI can quickly process complex information and generate structured content [38]. This provides an unmatched level of work efficiency compared with manual methods and helps overcome the challenge that material developers often face when they are unfamiliar with specialized medical knowledge related to disease prevention and treatment. Additionally, the powerful text generation capabilities of generative AI allow health care professionals to produce more PEMs in the same amount of time as manual development, potentially addressing the scarcity of PEMs in primary health care settings [39,40].

In terms of readability, from the perspective of the 4 core language feature indicators, generative AI is more stable in its readability output compared with manually written materials, which is a significant advantage. Generative AI tends to use fewer personal pronouns than manually written materials, which can improve the coherence of the text and the clarity of references. The content word frequency in generative AI was lower than in manually written materials, indicating that its content words are more varied, leading to greater cognitive load. This suggests that, by modifying the prompts, generative AI can be instructed to increase content word frequency to reduce the cognitive load for patients and their families, thereby improving readability.

This study found that both manually written and AI-generated PEMs achieved a Chinese elementary school reading level in terms of the proportion of simple sentences and the number of personal pronouns. However, the frequency of difficult words and content word frequency corresponded to a high school reading level, which did not meet the requirement that PEMs should be written at the 6th-grade reading level [41]. This aligns with the findings of several studies [42-44] that showed that both AI-generated and manually written PEMs do not meet the recommended 6th-grade reading level.

This study shows that both manually written and AI-generated PEMs have good understandability and actionability, with comparable quality. However, generative AI is more stable overall than manually written materials. The results of this study, consistent with several others [44,45], demonstrate that the understandability and actionability of AI-generated PEMs are no different from those written by humans, indicating that generative AI is equally suitable for generating Chinese PEMs for CHD. This capability of generative AI depends not only on its large database and powerful computational power but also on the structured prompts [9] provided to it, which require converting complex medical information into easily understandable language. This suggests that, when developing PEMs, in addition to health care or health education professionals, experts in AI generation technology or training for health care professionals in AI generation technology are also needed. Furthermore, generative AI has a powerful self-learning capability, which can be enhanced by providing high-quality PEMs [46,47] to train the model, thereby improving the quality of output materials.

Although the statistical results showed no significant difference in the accuracy of the 3 sets of PEMs, this finding may indicate that there is no essential difference in accuracy between AI-generated materials and manually written materials, or it may be due to the small sample size of only 8 sets of data per method, leading to limited statistical power. It is also possible that the Bonferroni correction weakened the original significant differences due to its conservativeness. Despite the lack of significant statistical differences, the accuracy of content generated by generative AI still needs to be carefully evaluated. Of the experts consulted in this study, 4 raised concerns about the accuracy of AI-generated materials. For example, 1 expert explicitly pointed out: “Material 2 does not use the latest standard.” A study [48] indicated that 30% of pathology references generated by ChatGPT were incorrect, including inaccuracies and nonexistent references. Another review [19] mentioned that large models might confidently output incorrect or nonexistent answers. The “hallucinations” and bias issues of generative AI are real, and erroneous information could potentially harm patients, highlighting the necessity for strict review by health care professionals before using AI-generated PEMs in patient education to ensure the accuracy of the content.

Beyond accuracy, the use of AI in health care also carries inherent risks. A review [49] found that bias and lack of transparency are the 2 main risks of medical AI. This finding is particularly important for health care professionals planning to use generative AI to generate PEMs, as essential content that should be used to reduce bias might be omitted in the pursuit of readability. This further emphasizes the need for clinical review of content generated by generative AI.

ChatGPT Versus DeepSeek

This study shows that both generative AI tools used a small number of personal pronouns to enhance readability. DeepSeek had a higher proportion of simple sentences with greater variation than ChatGPT, which aligns with the findings of Zhou et al [50]. However, both models’ use of personal pronouns and the proportion of simple sentences reached the Chinese elementary school reading level. DeepSeek’s difficult words were more concentrated and stable compared with those generated by ChatGPT, but the difference between the 2 models was highly uncertain. DeepSeek had a significantly higher content word frequency than ChatGPT, although the PEMs from both models still fell within the high school reading level for these 2 aspects. A study [17] on generative AI’s ability to improve the readability of human-written PEMs showed that, although multiple generative AI tools improved the readability of PEMs, the generated content still generally exceeded the target requirements. Future research on generative AI specifically trained on medical data or fine-tuned for this task

XSL•FO
RenderX

is worth exploring, particularly for reducing difficult words and increasing content word frequency to improve the overall readability of AI-generated materials.

Although the overall differences between the two generative AI tools in terms of understandability and actionability were not significant, ChatGPT was more stable in understandability, while DeepSeek was more stable in actionability. This suggests that, when using these 2 generative AIs to generate PEMs, more descriptive instructions can be provided to enhance their stability.

Accuracy is a common issue for both generative AI tools. In this study, experts pointed out that the number of errors in both models was the same. Two experts identified specific errors related to numerical information or issues where numerical information caused confusion for patients. This suggests that accuracy issues related to numerical information may be a common type of error for AI systems. These errors could stem from biases in the data used to train the AI system [49] or from intentional or unintentional manipulation of the data by the AI [51]. This highlights the importance of health care professionals maintaining a high sensitivity to the accuracy of numerical information when reviewing AI-generated PEMs.

DeepSeek R1 is a reasoning model [21]. After each prompt input, the model takes 7 seconds to 15 seconds to process and generate content. Although the initial content generated is logically coherent, it may include some areas where the content is overly specialized. Further optimization of the prompts, including additional rounds of interactions, is required to refine the content. ChatGPT is not a reasoning model and does not require thinking time, which explains why using DeepSeek takes slightly longer. At the same time, this study quantified the subtle differences in time consumption between the two, providing reference for health care professionals on how to choose the appropriate model.

Limitations

This study has several limitations. First, only 3 manually written sets of PEMs for CHD by a single hospital were evaluated, which is a major limitation of this study. Expanding the scope to include more hospitals or different disease types may yield different results. Second, the target users of PEMs are patients and caregivers, but in this study, only health care professionals conducted the evaluations, which may introduce medical knowledge bias. Future research could involve evaluations by target users after strict assessment of accuracy by health care professionals. Additionally, the calculation of PEMAT-P scores was performed by researchers who were aware of the group assignments, which could introduce subjective bias. Third, generative AI tools are rapidly evolving, and future versions may perform better. Fourth, although a multidimensional evaluation was conducted, this study did not assess whether the AI-generated PEMs omitted any key information. Future research could explore the completeness of content generated by generative AI.

Conclusions

The results of this study indicate that, compared with manually written materials, AI-generated PEMs for CHD demonstrate good outcomes in terms of time cost, understandability, and actionability, highlighting their potential for application. ChatGPT is more stable in understandability, while DeepSeek is more stable in actionability, suggesting that stability can be improved through more descriptive instructions. In terms of readability, although there are some differences between the 2 generative AI tools and between generative AI and manually written materials, the proportion of simple sentences and the number of personal pronouns in all 3 groups aligned with the reading level of Chinese elementary school students, while the difficulty of words and content word frequency corresponded to a high school reading level. The readability of AI-generated PEMs can be improved with more precise prompts.

In the future, generative AI models specifically trained on medical data or fine-tuned for this task could be developed to enhance the overall readability of AI-generated materials. Although there are certain flaws in accuracy, it is crucial to ensure that AI-generated PEMs for CHD undergo strict review by health care professionals to avoid outdated, incorrect, or nonexistent information being applied in clinical practice. In health care contexts, where information can impact human health and lives, it is essential to ensure that generative AI systems adhere to ethical standards and produce results that align with ethical considerations.

Acknowledgments

This study would not have been possible without the support of Sir Run Run Shaw Hospital, and we are sincerely grateful to all the staff who were involved.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Authors' Contributions

JL conceived the research and designed the study. SD, XM, and RC collected and processed the data. XJ wrote the manuscript. LW revised and proofread the manuscript.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
Prompt.
[DOCX File, 11 KB - formative_v9i1e78816_app1.docx]

References

1. Alpert JS. New coronary heart disease risk factors. Am J Med 2023 Apr;136(4):331-332. [doi: 10.1016/j.amjmed.2022.08.002] [Medline: 35981648]
2. Wu S, Xu W, Guan C, Lv M, Jiang S, Jinhua Z. Global burden of cardiovascular disease attributable to metabolic risk factors, 1990–2019: an analysis of observational data from a 2019 Global Burden of Disease study. BMJ Open 2023 May;13(5):e069397. [doi: 10.1136/bmjopen-2022-069397]
3. Su Y, Huang Y, Liu L, Liang J. The mediating effect of kinesiophobia between health literacy and cardiac rehabilitation barriers in post-PCI coronary heart disease patients. Practical Journal of Cardiac Cerebral Pneumal and Vascular 2025:124-129. [doi: 10.12114/j.issn.1008-5971.2024.00.338]
4. German CA, Baum SJ, Ferdinand KC, et al. Defining preventive cardiology: a clinical practice statement from the American Society for Preventive Cardiology. American Journal of Preventive Cardiology 2022 Dec;12:100432. [doi: 10.1016/j.ajpc.2022.100432]
5. Beijing Nursing Association Cardiovascular Professional Committee. Expert consensus on nursing care for health education prescriptions in cardiac rehabilitation of patients with coronary heart disease. Chinese Journal of Modern Nursing 2022;28(9):1121-1127. [doi: 10.3760/cma.j.cn115682-20211229-05884]
6. Wang J. Application effect of green model-based health education in coronary heart disease patients after PCI. Medical Journal of Chinese People's Health 2022;34(12):163-166. [doi: 10.3969/j.issn.1672-0369.2022.12.048]
7. Xu M, Yu X, Hao W, et al. A empowerment education model in coronary heart disease patients after PCI. Chinese Journal of Modern Nursing 2022;28(19):2625-2629. [doi: 10.3760/cma.j.cn115682-20211215-05661]
8. Duran S, Gonzalez AM, Nguyen K, Nguyen J, Zinn Z. Enhancing Spanish patient education materials: comparing the readability of artificial intelligence-generated Spanish patient education materials to the Society of Pediatric Dermatology Spanish patient brochures. Pediatr Dermatol 2025;42(1):106-108. [doi: 10.1111/pde.15805] [Medline: 39533849]
9. Ihsan MZ, Apriatama D, Amalia R. AI-assisted patient education: challenges and solutions in pediatric kidney transplantation. Patient Educ Couns 2025 Feb;131:108575. [doi: 10.1016/j.pec.2024.108575] [Medline: 39615127]
10. Patel EA, Fleischer L, Filip P, et al. The use of artificial intelligence to improve readability of otolaryngology patient education materials. Otolaryngol Head Neck Surg 2024 Aug;171(2):603-608. [doi: 10.1002/ohn.816] [Medline: 38751109]
11. Boscardin CK, Gin B, Golde PB, Hauer KE. ChatGPT and generative artificial intelligence for medical education: potential impact and opportunity. Acad Med 2024 Jan 1;99(1):22-27. [doi: 10.1097/ACM.0000000000005439] [Medline: 37651677]
12. Liu J, Xia Y, Wang L, et al. Study on the effects of generative artificial intelligence in ICU novice simulation instructor case design training. Chinese Journal of Nursing Education 2025;22(3):272-278. [doi: 10.3761/j.issn.1672-9234.2025.03]
13. Zeng S, Kong Q, Wu X, et al. Artificial intelligence-generated patient education materials for Helicobacter pylori infection: a comparative analysis. Helicobacter 2024;29(4):e13115. [doi: 10.1111/hel.13115] [Medline: 39097925]
14. Laymouna M, Ma Y, Lessard D, Schuster T, Engler K, Lebouché B. Roles, users, benefits, and limitations of chatbots in health care: rapid review. J Med Internet Res 2024 Jul 23;26:e56930. [doi: 10.2196/56930] [Medline: 39042446]
15. Gibney E. China's cheap, open AI model DeepSeek thrills scientists. Nature New Biol 2025 Feb 6;638(8049):13-14. [doi: 10.1038/d41586-025-00229-6]
16. Semeraro F, Cascella M, Montomoli J, Bellini V, Bignami EG. Comparative analysis of AI tools for disseminating CPR guidelines: implications for cardiac arrest education. Resuscitation 2025 Mar;208:110528. [doi: 10.1016/j.resuscitation.2025.110528] [Medline: 39909198]
17. Will J, Gupta M, Zaretsky J, Dowlath A, Testa P, Feldman J. Enhancing the readability of online patient education materials using large language models: cross-sectional study. J Med Internet Res 2025 Jun 4;27:e69955. [doi: 10.2196/69955] [Medline: 40465378]
18. Saji JG, Balagangatharan A, Bajaj S, Swarnkar V, Unni D, Dileep A. Analysis of patient education guides generated by ChatGPT and Gemini on common anti-diabetic drugs: a cross-sectional study. Cureus 2025 Mar;17(3):e81156. [doi: 10.7759/cureus.81156] [Medline: 40276455]
19. Liu Z, Wang P, Song X, Zhuang X, Jiang B. A review of the hallucination problem in large language models. J Softw (Malden) 2025:1152-1185. [doi: 10.13328/j.cnki.jos.007242]
20. Hariri W. Unlocking the potential of chatgpt: a comprehensive exploration of its applications, advantages, limitations, and future directions in natural language processing. arXiv. Preprint posted online on Mar 27, 2023. [doi: 10.48550/arXiv.2304.02017]
21. Ye J, Bronstein S, Hai J, Hashish MA. DeepSeek in healthcare: a survey of capabilities, risks, and clinical applications of open-source large language models. arXiv. Preprint posted online on Jun 2, 2025. [doi: 10.48550/arXiv.2506.01257]
22. Yang L, Qian Y, Yu H. Development and application of health education tools in our hospital. Chinese Journal of Hospital Administration 2012;28(8):615-617. [doi: 10.3760/cma.j.issn.1000-6672.2012.08.021]

XSL•FO
RenderX

23. Chinese Society of Cardiology of Chinese Medical Association, Editorial Board of Chinese Journal of Cardiology. Chinese Society of Cardiology (CSC) guidelines for the diagnosis and management of patients with ST segment elevation myocardial infarction. Chinese Journal of Cardiology 2019;47(10):766-783. [doi: 10.3760/cma.j.issn.0253-3758.2019.10.003]
24. Liu J, Tu H, Li X, et al. Summary of best evidence for health education on cardiac rehabilitation in patients undergoing percutaneous coronary intervention. Chinese Journal of Nursing Education 2023;20(2):210-217. [doi: 10.3761/j.issn.1672-9234.2023.02.015]
25. Interventional Cardiology Group of the Chinese Society of Cardiology, Atherosclerosis and Coronary Heart Disease Group of the Chinese Society of Cardiology. Guidelines for the diagnosis and treatment of stable coronary artery disease. Chinese Journal of Cardiology 2018;46(9):680-694. [doi: 10.3760/cma.j.issn.0253-3758.2018.09.004]
26. Sun GH. Prompt engineering for nurse educators. Nurse Educ 2024;49(6):293-299. [doi: 10.1097/NNE.0000000000001705] [Medline: 38968442]
27. Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. J Med Internet Res 2023 Oct 4;25:e50638. [doi: 10.2196/50638] [Medline: 37792434]
28. Lei L, Wei Y, Liu K. AlphaReadabilityChinese: development and application of a Chinese text readability tool. Foreign Languages and Their Teaching 2024;334(1):83-93 [FREE Full text]
29. Monje S, Ulene S, Gimovsky AC. Identifying ChatGPT-written patient education materials using text analysis and readability. Am J Perinatol 2024 Dec;41(16):2229-2231. [doi: 10.1055/a-2302-8604] [Medline: 38593984]
30. Sung YT, Chang TH, Lin WC, Hsieh KS, Chang KE. CRIE: an automated analyzer for Chinese texts. Behav Res Methods 2016 Dec;48(4):1238-1251. [doi: 10.3758/s13428-015-0649-1] [Medline: 26424442]
31. Chinese readability index explorer, CRIE 3.0. URL: http://www.chinesereadability.net/CRIE [accessed 2025-05-26]
32. Patient Education Materials Assessment Tool for Printable Materials (PEMAT-P). Agency for Healthcare Research and Quality. URL: https://www.ahrq.gov/sites/default/files/publications/files/pemat-p.pdf [accessed 2025-11-09]
33. Shoemaker SJ, Wolf MS, Brach C. Development of the Patient Education Materials Assessment Tool (PEMAT): a new measure of understandability and actionability for print and audiovisual patient information. Patient Educ Couns 2014 Sep;96(3):395-403. [doi: 10.1016/j.pec.2014.05.027] [Medline: 24973195]
34. Zhang Y, Yang R, Peng X, Zhang W, Gao Z. Principles and models of confirmatory factor analysis and its application in evaluating the structural validity of scales. Chinese Journal of Health Management 2022;16(6):428-432. [doi: 10.3760/cma.j.cn115624-20210928-00569]
35. Yu Q, Huang W, Yan J, et al. Translation and validation of a patient health education material assessment tool. Chinese Journal of Health Education 2023;39(10):912-916. [doi: 10.16168/j.cnki.issn.1002-9982.2023.10.009]
36. Gao F, Gao X, Shao Y, et al. Application of large language models in health education for patients with diabetic retinopathy. Chinese Journal of Experimental Ophthalmology 2024;42(12):1111-1118. [doi: 10.3760/cma.j.cn115989-20240723-00207]
37. Vaira LA, Lechien JR, Abbate V, et al. Accuracy of ChatGPT - generated information on head and neck and oromaxillofacial surgery: a multicenter collaborative analysis. Otolaryngol Head Neck Surg 2024 Jun;170(6):1492-1503. [doi: 10.1002/ohn.489] [Medline: 37595113]
38. Ahn J, Park HO. Development of a case-based nursing education program using generative artificial intelligence. J Korean Acad Soc Nurs Educ 2023 Aug;29(3):234-246. [doi: 10.5977/jkasne.2023.29.3.234]
39. Xia H. Analysis of the current status of health education for inpatients in primary hospitals. Systems Medicine 2017;2(12):151-153. [doi: 10.19368/j.cnki.2096-1782.2017.12.151]
40. Zhang H, Song Q. Investigation and analysis of hypertension education in primary hospitals. China Continuing Medical Education 2016;8(24). [doi: 10.3969/j.issn.1674-9308.2016.24.126]
41. Rooney MK, Santiago G, Perni S, et al. Readability of patient education materials from high-impact medical journals: a 20-year analysis. J Patient Exp 2021;8:2374373521998847. [doi: 10.1177/2374373521998847] [Medline: 34179407]
42. Gibson D, Jackson S, Shanmugasundaram R, et al. Evaluating the efficacy of ChatGPT as a patient education tool in prostate cancer: multimetric assessment. J Med Internet Res 2024 Aug 14;26(2):e55939. [doi: 10.2196/55939] [Medline: 39141904]
43. Gupta M, Gupta P, Ho C, Wood J, Guleria S, Virostko J. Can generative AI improve the readability of patient education materials at a radiology practice? Clin Radiol 2024 Nov;79(11):e1366-e1371. [doi: 10.1016/j.crad.2024.08.019] [Medline: 39266371]
44. Pradhan F, Fiedler A, Samson K, Olivera-Martinez M, Manatsathit W, Peeraphatdit T. Artificial intelligence compared with human-derived patient educational materials on cirrhosis. Hepatol Commun 2024 Mar 1;8(3):e0367. [doi: 10.1097/HC9.0000000000000367] [Medline: 38358382]
45. Azzopardi M, Ng B, Logeswaran A, et al. Artificial intelligence chatbots as sources of patient education material for cataract surgery: ChatGPT-4 versus Google Bard. BMJ Open Ophthalmol 2024 Oct 17;9(1):e001824. [doi: 10.1136/bmjophth-2024-001824] [Medline: 39419585]
46. Pu Q, Xiang W. Generative artificial intelligence—the transformative impact, risks, challenges, and coping strategies of ChatGPT. Journal of Chongqing University (Social Science Edition) 2023;29(3):11835. [doi: 10.11835/j.issn.1008-5831.pj.2023.04.001]
47. Tang A, Tung N, Nguyen HQ, et al. Health information for all: do large language models bridge or widen the digital divide? BMJ 2024;387:e080208. [doi: 10.1136/bmj-2024-080208]

XSL•FO
RenderX

48. Guastafierro V, Corbitt DN, Bressan A, et al. Unveiling the risks of ChatGPT in diagnostic surgical pathology. Virchows Arch 2025 Apr;486(4):663-673. [doi: 10.1007/s00428-024-03918-1] [Medline: 39269615]
49. Chustecki M. Benefits and risks of AI in health care: narrative review. Interact J Med Res 2024 Nov 18;13:e53616. [doi: 10.2196/53616] [Medline: 39556817]
50. Zhou M, Pan Y, Zhang Y, Song X, Zhou Y. Evaluating AI-generated patient education materials for spinal surgeries: comparative analysis of readability and DISCERN quality across ChatGPT and DeepSeek models. Int J Med Inform 2025 Jun;198:105871. [doi: 10.1016/j.ijmedinf.2025.105871] [Medline: 40107040]
51. Okunev M, Handelman D, Handelman A. Ethical considerations and robustness of artificial neural networks in medical image analysis under data corruption. Sci Rep 2025 Aug 11;15(1):29305. [doi: 10.1038/s41598-025-15268-2] [Medline: 40790201]

Abbreviations

AI: artificial intelligence
CHD: coronary heart disease
CRIE: Chinese Readability Indicator and Evaluator
PEM: patient education material
PEMAT-P: Patient Education Materials Assessment Tool for Printable Materials

Edited by A Mavragani, J Sarvestan; submitted 10.06.25; peer-reviewed by M Okonkwo, S Mohanadas; revised version received 17.10.25; accepted 27.10.25; published 21.11.25.

Please cite as:
Jiang X, Liu J, Dai S, Mao X, Cha R, Wu L
Evaluating the Effectiveness of Generative AI for the Creation of Patient Education Materials on Coronary Heart Disease: A Comparative Study
JMIR Form Res 2025;9:e78816
URL: https://formative.jmir.org/2025/1/e78816
doi:10.2196/78816

©Xiaofang Jiang, Jingbang Liu, Shanshan Dai, Xiawen Mao, Rongping Cha, Lili Wu. Originally published in JMIR Formative Research (https://formative.jmir.org), 21.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Leveraging No-Code Digital Platforms for Designing an Integrated Smartphone-Based Ecological Momentary Intervention With Cognitive Behavioral Therapy for Mental Health Care: Development and Usability Study

Stephanie Ming Yin Wong[1], PhD; Melody Ho Ching Ip[1], BSocSc; Jessica Kang Qi Lee[1], BSocSc; Terry Yat Sang Lum[1], PhD; Ralf Schwarzer[2], PhD

[1]Department of Social Work and Social Administration, University of Hong Kong, Rm 702e, The Jockey Club Tower, Centennial Campus, Hong Kong, China (Hong Kong)

[2]Department of Psychology, Freie Universität Berlin, Berlin, Germany

Corresponding Author:
Stephanie Ming Yin Wong, PhD
Department of Social Work and Social Administration, University of Hong Kong, Rm 702e, The Jockey Club Tower, Centennial Campus, Hong Kong, China (Hong Kong)

Abstract

Background: The rising burden of disease associated with mental disorders calls for evidence-based psychological interventions that can be swiftly scaled up. Blending smartphone-based mental health apps (MHapps) for delivering ecological momentary interventions (EMIs) with traditional in-person interventions may have the benefits of improving treatment adherence, facilitating the application of learned techniques into everyday life, and, in turn, enhancing clinical response. However, previous work has shown that most existing MHapps were developed for specific research studies or for profit, thereby making them difficult to adapt, particularly in time-limited and resource-constrained settings.

Objective: This study aimed to demonstrate how a person-centered and theory-informed MHapp could be developed in a timely and low-cost manner for use as part of blended care, using a phased approach. Given the scarcity of digital mental health interventions for older adults, we adopted a participatory research approach to co-design the blended intervention with 2 groups of older adults.

Methods: In Phase 1, we reviewed existing MHapps with consideration of whether they could be adapted by individual researchers or clinicians, their key functions, and whether their efficacy had been tested. “No-code” app builders were additionally reviewed, which may be alternatives if no MHapp can be used. In Phase 2, following the IDEAS (Integrate, Design, Assess, and Share) framework, we built a prototype according to users’ needs, with its content informed by theories of cognitive behavioral therapy (CBT) and the Health Action Process Approach. The prototype was then tested and refined over 2 rounds of 3-session co-design workshops with peer supporters (n=8) and service users (n=5) from a stepped-care intervention for older adults with depressive symptoms. Usability testing was conducted with both stakeholder groups in Phase 3.

Results: Of the 149 MHapps identified, only 43 (28.9%) can be publicly downloaded. Four (8.3%) of them can be partially adapted, although no new content can be directly added. We therefore developed the MHapp using m-Path (a spin-off from KU Leuven's Faculty of Psychology), which was the only existing no-code app development platform designed for mental health interventions. A prototype incorporating CBT-based homework and behavior change techniques informed by the Health Action Process Approach was built, with its refined version rated as highly easy to use and acceptable by both stakeholder groups.

Conclusions: By integrating CBT with EMI, we demonstrated the feasibility and acceptability of a novel blended care model for reference in future work. Preliminary findings suggest high usability and clinical relevance, highlighting the potential of leveraging no-code platforms to facilitate scalable, theory-driven interventions that extend mental health support beyond traditional settings. Grounding the blended intervention in evidence-based psychological and health behavior change theories, coupled with user involvement throughout the design process, may improve clinical efficacy and reduce implementation barriers, which are areas for further investigation in future work.

(JMIR Form Res 2025;9:e77036) doi:10.2196/77036

KEYWORDS

cognitive behavioural therapy; ecological momentary intervention; blended care; mental health care; depression; health behaviour change; health action process model; user-centred design; peer support; mobile apps; digital intervention

XSL•FO
RenderX

Introduction

Background

Mental disorders constitute some of the leading causes of the global burden of disease [1]. Not only do they have implications for the health care system and economy, but they also affect the functioning and quality of life of the affected individuals and their families. Increasing efforts have been made to synthesize evidence and scale up early interventions in community settings [2].

Over the past 2 decades, the field has undergone significant transformations in the delivery of psychological and pharmacological treatments, shifting from face-to-face to incorporating digital technologies as part of mental health care. Such interventions range from the use of web-based platforms and online video conferencing to smartphone apps, the experience sampling method (ESM), and ecological momentary intervention (EMI), as well as chatbots and virtual reality, either as an adjunct to or replacement for in-person care. Each approach may vary in its degree of human contact and intensity [3-7]. Although the application of digital technologies to mental health care has been relatively slow compared to efforts in physical health and behavior change, the emergence of the COVID-19 pandemic has greatly accelerated this process.

Benefits of Integrating Smartphone-Based Apps Into Mental Health Care

Among the various digital tools, the use of smartphone-based mental health apps (MHapps), specifically blended care, has received increasing scientific attention and support. Some benefits include their ability to save clinicians time, which is crucial given the burden of mental disorders and the persistent challenge of insufficient manpower, as well as to improve treatment accessibility and reduce dropout rates.

Treatment Adherence and More Sustainable Changes

As in pharmacotherapy, adherence to care plans is a crucial determinant of treatment outcomes in psychotherapy. Using cognitive behavioral therapy (CBT) as an example, adherence includes not only regular therapy attendance but also the completion of homework. Homeworks in CBT are cognitive or behavioral activities (or their combination) that are collaboratively planned between the clinician and service user for completion outside of therapy, which involves the direct application of adaptive strategies into everyday life [8,9]. Such activities are personalized and may include scheduling meaningful activities (eg, “leaving the house for 30 minutes every day”), monitoring daily activities and mood states, and identifying cognitive biases while documenting their relationships with triggering events, mood, and physiological reactions [9,10].

Despite the importance of applying these practices to everyday life for more sustainable treatment outcomes [8], homework nonadherence is profoundly common. Beyond difficulties among the general population [11,12], the maintenance of health-related behavior change can be even more challenging in people with subclinical and clinical mental health symptoms, wherein motivational deficits and anhedonia reflect cardinal symptoms of various psychiatric conditions, such as major depression, schizophrenia spectrum disorders, eating disorders, and posttraumatic stress disorder [13]. Blending MHapps with traditional in-person interventions has the potential to fill this gap.

The Delivery of In-The-Moment Interventions Beyond the Therapy Room

The portability of smartphones and the lack of geographical limitations in their use have opened new opportunities for more person-centered, holistic, and context-relevant interventions [9,14-16]. The EMI is a novel intervention approach that was built upon the ESM—a self-report diary method that enables the collection of real-time data on experiences within individuals and their interactions with one’s real-world context—by additionally allowing real-world delivery of interventions [17]. The availability of active and passive data collected using smartphone-based EMI provides clinicians with information about service users that is not possible in traditional weekly in-person therapy sessions, such as in-the-moment mood states across everyday life contexts; these data can then be used to identify patterns in mood states and their fluctuations in relation to the context (eg, where the person is, who they are with, and what they are doing) and inform more personalized clinical interventions [17-19].

Of note, although standalone EMIs are associated with reduced psychiatric symptoms compared with those not receiving EMI or other forms of psychotherapy [20], the presence of human contact and support has consistently been observed to yield larger and more robust effect sizes across a range of mental health conditions [3,21], thereby supporting the potential of blended mental health care in improving outcomes.

Opportunities and Challenges in the Implementation of Blended Care via Smartphone-Based EMI

There is now evidence to support the acceptability and effectiveness of digital mental health interventions in not only high-income but also low- and middle-income countries [22], as well as in those previously thought to be less digitally literate, such as older adults [23]. However, several challenges remain. First, although barriers related to the use of MHapps among users have been explored (eg, limited customizability and technical issues [24,25]), little work has been done to pinpoint the barriers to implementation and its facilitators, particularly with consideration of the wider organizational and societal contexts, as well as the possible strategies to facilitate implementation [26]. The involvement of multiple stakeholders throughout the development and design process of digital interventions can help overcome challenges to their uptake in real-world settings.

Further, most existing MHapps have been either developed for use by specific research or clinical teams [27], for profit, or lack empirical evidence and a theoretical basis [28]. The development of a new MHapp is often not practical, not only in many low- and middle-income countries but also in teams with lower resources and in time-limited settings (eg, context of war, social unrest, and pandemic outbreaks). Free-of-charge and tailorable MHapps or platforms that can be easily adapted are crucial for

XSL•FO
RenderX

improving access to personalized health care and contributing to reduced health disparities.

Objectives

To facilitate advances in the field of digital interventions for mental health, the replicability of both the development and design process is crucial. This paper aims to describe the process of developing a theory-driven, person-centered smartphone-based EMI for use as part of a blended CBT intervention for depression (blended CBT-EMI for depression). Three major phases of the design procedure are described: (1) a review of existing MHapps and "no-code" app development tools for potential adaptation and customization for specific research and clinical purposes, (2) an iterative cocreation process with stakeholders via design thinking workshops, and (3) usability testing. Given that the majority of existing MHapps have been designed for young people, this study involved 2 core groups of older adults as stakeholders in the design process (peer supporters and previous service users), serving as examples for future replications.

Methods

Phase 1: Review of Existing MHapps and "No-Code" App Development Tools

Before involving stakeholders in the smartphone-based EMI development process, we conducted an in-depth review of existing digital MHapps to determine whether any platform was readily available for adaptation across clinical populations and settings.

Two reviews of MHapps have been published in the past 10 years (one providing recommendations for future developments of MHapps [16], and another a recent meta-analysis on the efficacy of existing MHapps for depression and anxiety [28]), although neither examined their availability for public use nor their degrees of adaptability. We therefore hand searched all MHapps cited in these 2 reviews on the iOS App Store (Apple Inc) and Android Google Play Store (Google LLC), from which the following information was extracted (as of May 6, 2025): name of the MHapp, availability for download and use, whether it could be adapted by individual researchers or clinicians, key functions, developer, date of last update, and whether clinical trials have been conducted to evaluate its efficacy.

In recent years, the gradual rise of "no-code" app development tools [29,30], which are platforms that allow researchers with no previous software programming experience to build apps via a graphical user interface, has reduced the time and cost required for development. We also reviewed their potential suitability for use compared with readily available MHapps.

Phase 2: Iterative Co-Design and Cocreation Process of the Platform

After identifying a possible platform for the EMI app, a series of cocreation workshops and usability testing were conducted to ensure the design of the CBT-EMI was person-centered and to minimize implementation barriers in practice.

The IDEAS (Integrate, Design, Assess, and Share) framework [31] was adopted to guide this process. Consolidating strategies from behavioral theory, design thinking, and user-centered design, the IDEAS framework was intended to provide researchers and clinicians with a more systematic, step-by-step approach to developing effective digital interventions that target health behavior change. A total of 10 steps are specified, which can be broadly categorized under 4 stages: Integrate, Design, Assess, and Share ([31] refer to Figure 1).

Similar to a previous study on the development of a mobile app for hypertension prevention [32], this study focuses on the first 2 stages (Integrate and Design) of the IDEAS framework, with preliminary findings evaluated using a minimum viable product for the third stage (Assess). Further evaluation of its feasibility and efficacy will be conducted in the next phase and is beyond the scope of this study.

XSL•FO
RenderX

Figure 1. Application of the IDEAS (Integrate, Design, Assess, and Share) framework for developing a blended cognitive behavioral therapy with smartphone-based ecological momentary intervention (CBT-EMI) mobile health app platform.

XSL•FO
RenderX

Empathize With Target Users (Step 1)

For this study, target users were older Chinese adults. As such, our smartphone-based CBT-EMI was built upon an existing 6-week, group-based CBT intervention tailored for older adults with subclinical depressive symptoms, which was designed by clinical psychologists and senior social workers as part of a large-scale collaborative stepped-care intervention project, JC JoyAge (Jockey Club Holistic Support Project for Elderly Mental Wellness), across Hong Kong since 2016 [33,34]. The design of the CBT intervention was guided by the ecological validity framework [35,36] to ensure its cultural relevance to the local older adult population, which comprises 8 central elements, namely, language, persons, metaphors, content, concepts, goals, methods, and context. For instance, aside from using traditional Chinese and colloquial Cantonese terms and sentences in all intervention materials and protocols, metaphors (eg, "a broken ship still has three catties of nails," which is a local idiom meaning "no matter how bad something is, there is still some value in it") were also commonly referenced to facilitate the relevance of the CBT intervention content to participants. The significance of such cultural adaptations and co-design processes in psychological interventions in the current context has been described in our previous work [33,37-39].

Regarding the EMI platform, we took reference from the existing knowledge base on how digital technologies may be designed to facilitate the older adult population in receiving psychological interventions, particularly in the Asia-Pacific context (eg, the TORCH [Technology provision, On-site technical support, Rehearsal, Connection with group members, and Hardcopy notes] principle [40,41]). To optimize user experience (UX), the platform was also designed according to principles listed in established guidelines on usability testing [42], as well as a guide developed by the local Government of the Hong Kong Special Administrative Region on designing elderly-friendly website or mobile apps (ie, need for clear and easy-to-understand content, easy-to-read-and-operate layout, clear steps and instructions, and support for voice input method [43]).

Specify Target Behavior and Ground in Theory (Steps 2 and 3)

Aside from taking reference from the ESM and EMI literature in the design of the CBT-EMI UX, the target behaviors (CBT-based homework) were designed based on (1) theories of CBT, and (2) theories of behavior change and maintenance. We took reference from the traditional printed worksheets used in the regular CBT in JC JoyAge [33,34,44] and pinpointed homework activities for potential application in the blended care.

To optimize more sustainable behavior change (ie, adherence to CBT-based activities between and beyond intervention sessions), we innovatively integrated theories of CBT with the health behavior change literature. Given the increasing emphasis on specifying "active ingredients" to improve the understanding of mechanisms of change in psychological interventions [45], we sought to incorporate behavior change techniques (BCTs) as part of the CBT-EMI, which are known as the smallest replicable components of an intervention for inducing behavior change via specified mechanisms of action [46,47]. The incorporation of such theory-informed BCTs via MHapps may offer an avenue for opening up the black box in mental health science and in advancing person-centered care. Recommendations in the design of EMIs were also reviewed to improve the usability and feasibility of the CBT-EMI platform.

Iterative Design Process With Prototype Testing and Feedback Gathering (Steps 4-7)

Building on Steps 1‑3, we recruited 2 groups of core stakeholders to a series of 3 weekly co-design workshops, each lasting 3 hours. The first group was peer supporters aged ≥50 years from the aforementioned project (JC JoyAge [33,48]), who had received >80 hours of training for service provision in the social care setting (October-November 2024). The second group was previous service users of this project who had presented symptoms of subclinical depression (defined as a score of 5‑19 on the Patient Health Questionnaire-9 (PHQ-9) at service intake, February-March 2025). Two mental health researchers (SMYW and MHCI), together with a senior social worker or clinical psychologist experienced in older adults' mental health care, led the workshops, while another trained research assistant (JKQL) took written field notes of observations and feedback.

All participants were required to have experience in using smartphones and to give informed consent. Those with a history of autism spectrum disorder, intellectual disability, schizophrenia spectrum disorder, bipolar disorder, Parkinson's disease, or dementia, as well as those with imminent suicidal risk, were excluded. To ensure participants could freely share their ideas and experiences, we opted for around 5‑6 participants per workshop (maximum of 8) with reference to established guidelines on usability testing [42]. Table 1 shows the sociodemographic characteristics of the 2 participant groups.

During Session 1, a series of design thinking activities, such as Empathy Map and the Eisenhower Matrix were implemented to explore the needs of service users, identify pain points, and brainstorm creative solutions (details in Table S1 in Multimedia Appendix 1). Based on the CBT homework and conventional EMI items identified for potential adaptations, the current research team developed a prototype for testing and refinement. Specific considerations were made for the user interface, user design, and interaction flow during the design process (eg, choice of words, use of illustrations, and design style, as well as the number of prompts to be sent per day and the time of sending), which are further discussed and refined throughout Sessions 1‑3 of the workshops.

XSL•FO
RenderX

Table . Sociodemographic characteristics of participants in the co-design workshops.

Characteristic	Peer supporters (n=8)	Service users (n=5)
Sex, n (%)		
Female	5 (62.5)	5 (100)
Age (years), mean (SD)	61.5 (5.3)	69.8 (3.7)
Highest educational attainment, n (%)		
Primary or below	0 (0)	2 (40)
Secondary school	2 (25)	2 (40)
Diploma	3 (37.5)	0 (0)
University or above	3 (37.5)	1 (20)
Occupation status, n (%)		
Full-time employment	1 (12.5)	0 (0)
Part-time employment or freelance	2 (25)	0 (0)
Retired	5 (62.5)	5 (100)
Health condition, n (%)		
Has any physical health condition	3 (37.5)	4 (80)
Has any diagnosed mental health problems	0 (0)	2 (40)
Severity of depressive symptoms at service intake (PHQ-9)[a], mean (SD)	—	8.80 (3.03)
Years since peer supporter training, mean (SD)	2.1 (1.1)	—
Smartphone use frequency, n (%)		
1 - 2 times per day	1 (12.5)	1 (20)
≥3 times per day	7 (87.5)	4 (80)

[a]PHQ-9: Patient Health Questionnaire-9.

Phase 3: Usability Testing of the Blended CBT-EMI Platform

After Session 1, participants were sent a preferred number of prompts per day (as discussed during the workshop) over the subsequent 6 days to test several CBT-based homework assignments in their everyday lives. During Sessions 2 and 3, discussions were then held regarding their experiences with enacting the homework activities over the past week (eg, what they liked about the MHapp design, and what barriers they faced).

The “think-aloud” method was used to obtain participants’ immediate thoughts and feedback on different features of the platform and specifically the following: (1) sampling scheme (including the optimal frequency, length, and schedule of homework activities), (2) preferred wordings and notification settings, and (3) other features considered essential to promote MHapp use (eg, expiration windows and conditional logics). A brief questionnaire was also administered alongside the MHapp demonstration, in which participants rated their preferences for adopting certain features in the final MHapp on a 7-point Likert scale from 1 (strongly disagree) to 7 (strongly agree).

Between Sessions 2 and 3, another round of real-world usability testing of the prototype (refined based on participants’ comments) took place with another set of CBT-EMI homework. During Session 3, final discussions were made to consolidate participants’ preferred features of the MHapp (including logistics), as well as its interface. A final questionnaire, which included 12 statements similar to those in previous work [18], was given to participants at the end of the workshop to capture their thoughts on the functionality of the MHapp and the overall arrangements of the workshops. Each statement was rated on a 7-point Likert scale, from 1 (not at all) to 7 (very).

Statistical Analysis

All findings gathered from the co-design workshops were synthesized to determine core features to be prioritized in the CBT-EMI platform. Instant data analysis [49]—which is a time-saving approach commonly adopted for usability testing—was performed, wherein 3 researchers (SMYW, MHCI, and JKQL) discussed their observations and field notes after each workshop to identify problems reported by the participants, review action plans, and adjust the prototype of the MHapp for subsequent testing. For any inconsistent suggestions identified from the 2 rounds of workshops held with peer supporters and service users, respectively, the needs raised by the latter group were prioritized. Additional comments were also sought from clinical practitioners and mental health researchers to ensure that the final MHapp design was clinically relevant and feasible in terms of implementation and practice. Descriptive statistics were generated for the quantitative data gathered from the questionnaires on preferences for adopting certain MHapp

XSL•FO
RenderX

features, as well as perceived functionality of the overall MHapp and the workshops, using SPSS (version 29.0.1.0; IBM).

Ethical Considerations

Written informed consent was obtained from all participants. Ethics approval was granted by the Human Research Ethics Committee of the University of Hong Kong (EA240144). Each participant received US $38 (HK $300) after their participation in the 3 co-design workshop sessions to acknowledge their contribution to the co-design process.

Results

Phase 1: Landscape Review and Analysis

Review of Existing MHapps

A total of 149 unique, named MHapps were identified, among which only 43 (28.9%) could be publicly downloaded. More details of these MHapps, including those that could not be downloaded, are provided in Table S2 in Multimedia Appendix 1.

Most of them were in English, with the exception of 3 MHapps, which were available only in Spanish or Dutch. Of the MHapps retrieved, the majority were available to the public (90.7%, n=39), with access to 4 apps restricted to specific users (ie, US or UK citizens covered by specific employers or health plans, corporations, or study research participants). Since these apps were retrieved from the 2 aforementioned papers, the majority (93%, n=40) had been tested for their efficacy in clinical trials.

Despite the relatively high accessibility to these MHapps by the general public, only 4 (8.3%) could be partially adapted by individual researchers or clinicians (eg, notification settings and therapeutic activities for clients could be customized, and data analytics could be retrieved from the backend server). Notably, only existing content made available by the developers could be used and could not be edited, meaning that no new content relevant to specific interventions or contexts could be incorporated.

Review of Existing "No-Code" App Builders

Since none of the reviewed MHapps could be fully tailored for use in our local context, we reviewed the potential of adopting "no-code" app builders for designing the blended intervention. A recently published review has identified 11 such tools [29]. Among them, only 1 was designed primarily for mental health research and interventions (m-Path, initially developed by the KU Leuven team in Belgium), which appears to be suited for our intervention purposes given the following: (1) features available for ESM and EMI (eg, a wide range of mental health–related question types available, allows collection of both active and passive data, and includes gamification options); (2) core aspects of its interface could be personalized by individual users (eg, language, color, and text); (3) no coding experience required (eg, web-based interface with a summary dashboard of all data points and events created by each user for review by practitioners or researchers); (4) tailorable in multiple aspects, including survey content (eg, multiple question and response types available), survey flow (eg, randomization and conditional logic options), and study design (eg, event-contingent or time-based sampling, personalized reminders); (5) strict data security and privacy policies (eg, no mandatory disclosure of personal particulars during app registration, anonymous data processing and storage, end-to-end phone-server communication, and compliance with standardized regulations, allowing participants to delete or block researchers' access to their data at any time); and (6) compatible with both iOS and Android devices [50].

A further review of the literature suggested that previous mental health studies have chosen m-Path given its low barrier to use (eg, no personal data such as email, phone number, or name is required [51]). We also confirmed that existing no-code app builders, except m-Path (at least up to the time of paper preparation), were primarily targeted at changing and tracking physical activities (eg, step trackers) and did not include the above features. We therefore selected m-Path as the platform to build the present blended CBT-EMI MHapp for delivering CBT-based homework using EMI to manage and reduce depressive symptoms. The procedures adopted in this study in co-designing the blended care platform, nevertheless, should provide a helpful reference for future work if another platform is identified as more suitable.

Phase 2: Iterative Co-Design and Cocreation Process of the Platform

Empathize With Target Users

Although m-Path was already available in 16 languages at the time of exploration, the platform was not available in Chinese. While the majority of residents in Hong Kong are proficient in English, the prevailing spoken language remains Chinese, particularly among older adults. We therefore reached out to the m-Path team to discuss the possibility of translating the platform into traditional Chinese (participants could still freely choose to use the platform in English or other languages, although the traditional Chinese version may provide more personal relevance, particularly for intervention purposes). Two members of the current research team (SMYW and JKQL), who have experience with translations in previous work and are proficient in Chinese and English—both of which are official languages in Hong Kong—independently completed the translation process from the landing pages to all subsequent pages. As the CBT program had already been developed following cultural adaptation frameworks in JC JoyAge, its CBT-based homework activities were directly incorporated into m-Path, with only minor adaptations made to facilitate reading on smartphones. Clinical psychologists of the project subsequently reviewed all materials to ensure the core therapeutic components were retained in the homework activities.

To ensure the accuracy and relevance of the translated materials, participants were presented with the initial research-adapted version of CBT-EMI and were then invited to reflect on and discuss the understandability and cultural friendliness of the terms and the illustrations. Adaptations were made iteratively and concluded when participants were satisfied. During this process, participants generally expressed satisfaction with the overall flow of CBT-EMI. Several issues related to language

XSL•FO
RenderX

and context within the cultural adaptation framework were identified and addressed. Regarding language, the entire intervention was revised from written Chinese to spoken Chinese. Specific terms for emotional expressions were replaced with colloquial expressions, with additional examples relevant to older adults incorporated. For instance, scenarios such as "having conflicts with family members, neighbors, and friends" were added to the "interpersonal relationships" category in the Mood Check exercise.

Specify Target Behavior and Ground in Theory

CBT Theory

With reference to the design of the pre-existing CBT for depression in older adults in Hong Kong [33,34], a series of possible homework activities grounded in CBT theory were identified for incorporation within the blended care, as follows: Mood Check, Relaxation Exercises, and Behavioral Activation (engage in relaxation exercises and participate in enjoyable, mood-inducing activities in daily life), "Hot Cross Bun" diagram (map triggering event, thoughts, feelings, physical sensations, and behaviors), Gratitude Diary (identify 3 things or people that one is grateful for on the day), Activity Scheduling (preplan mood-inducing activities that one can engage in during the week and record mood afterwards), and Wellness Toolbox (preset actions to engage in when feeling distressed in the future).

We strived to resemble the homework activities from the original paper-and-pen version by preserving the question type, multiple-choice options, and their overall length. Clinical psychologists reviewed the content of the homework to ensure the therapeutic quality of the app-adapted homework.

Health Behavior Change Theory

Further, we drew on the Health Action Process Approach (HAPA [52]) in modifying the intervention, both in terms of how the CBT homework activities can be delivered via m-Path, as well as the content of the in-person sessions. This was considered relevant, particularly given the challenges often observed in treatment and homework adherence in regular CBT.

The HAPA is one of the leading models informing the processes underlying health behavior change and maintenance. Accordingly, the combination of goal setting, planning, and perceived self-efficacy plays a crucial role in determining whether one will initiate a behavior for health changes and maintain its enactment in the long term [52]. Three phase-specific forms of self-efficacy are specified: action self-efficacy, which occurs during the motivation phase when a person has not yet enacted the behavior but develops an intention, even if it requires planning and a lifestyle change, as well as maintenance and recovery self-efficacy, which occurs during the volitional phase when the behavior has already been enacted but determines whether the person can continue to adhere to the behavior even if there are hurdles and can get back on track even after having derailed, respectively [53].

With reference to the health behavior change literature, we therefore identified several BCTs targeting "beliefs about capabilities" (a mechanism of action conceptually identical to self-efficacy [54,55]) in the design of the CBT-EMI platform. The underpinning assumption is that participants will show greater motivation in initiating and continuing the enactment of CBT-based homework activities in their everyday life via incorporating BCTs that foster self-efficacy using the EMI method, which, in turn, results in greater homework adherence, greater improvements in depressive symptoms, and a lower rate of relapse. Table 2 provides an overview of the mapping of BCTs and CBT-based homework in the proposed blended care intervention design. The BCT clusters, mechanism of action, and definitions of BCTs are based on the works by Michie et al [46] and Carey et al [54].

XSL•FO
RenderX

Table . Selection of self-efficacy–related behavior change techniques and their application in the blended CBT-EMI.

BCT[ab] (cluster)	Definition of the BCT	Application in the blended CBT-EMI[c] (implemented in a CBT session or EMI)
Set graded tasks (repetition and substitution)	Set tasks with the client, from simple to increasingly difficult (but achievable), until target behaviors are performed.	The intensity and frequency of homework increased throughout the 6-week CBT intervention. For example, after Session 1, clients were asked to complete "Mood Checks" via the smartphone MHapp twice a day before Session 2. After Session 2, clients were tasked with "Mood Checks," followed by in-the-moment behavioral activation, and so on. New homework was added or updated in the following weeks (except for "Mood Check," which was prompted throughout the 6 weeks) [EMI].
Verbal persuasion about capability (self-belief)	Encourage the client to trust their capabilities in performing the behaviors and in overcoming self-doubt.	Across the various CBT sessions, the practitioner discussed any concerns the clients had and assisted them in developing more confidence in their capabilities in making changes in the behaviors that are anticipated to have benefits for their mental well-being and performing them. Examples of verbal persuasion include the provision of validation to participants' experiences, including feelings, thoughts, and behaviors, reframing and challenging negative thoughts, as well as evaluating behaviors and progress in subsequent sessions [CBT session].
Focus on past success (self-belief)	Facilitate reflection on and list previous successes in performing the exact or similar behaviors.	Aside from discussing problems and identifying risk factors, the practitioner facilitated each client in becoming aware of their personal strengths and protective factors, including examples of their previous experiences in enacting the CBT-based homework and in overcoming challenges in performing them [CBT session].
Problem solving (goals and planning)	Facilitate the client in analyzing factors that might influence the (non)enactment of behaviors and develop strategies to overcome barriers and increase facilitators (eg, coping planning and relapse prevention).	The practitioner discussed the core "Hot-Cross Bun" model of CBT, which highlights the intricate relationships between an external trigger, as well as thoughts, feelings, physical or physiological sensations, and behaviors; applied it to the formulation of each client's experiences, and considered the 5Ps of their experiences (presenting problems, predisposing factors, precipitating factors, perpetuating factors, and protective factors). These clinical formulations then formed the basis for clients to identify the personal meanings behind the enactment of homework activities, as well as strategies to reduce barriers and increase chances of successful enactment [CBT session].
Demonstration of the behavior (comparison of behavior)	Provide an example of how the behaviors can be performed, either directly or indirectly (eg, via photos or videos), for the client to learn.	The practitioner demonstrated all homework activities during the CBT sessions, including a run-through of each procedure. Each activity was completed by the clients at least once before leaving the therapy room. All activities on the smartphone MHapp were designed to be straightforward. For more complex tasks, guided examples were incorporated into the MHapp (eg, video-guided diaphragmatic breathing exercises) [CBT session and EMI].

XSL•FO
RenderX

BCT[ab] (cluster)	Definition of the BCT	Application in the blended CBT-EMI[c] (implemented in a CBT session or EMI)
Instructions on how to perform a behavior (shaping knowledge[b])	Provide details on how each specified behavior is performed.	The practitioner provided details on how each homework activity on the smartphone MHapp was provided during the CBT session (via demonstration, as above, and with handouts as needed). Additional descriptions were also provided in the MHapp for more complex homework activities (eg, Wellness Toolbox) [CBT session and EMI].
Behavioral practice or rehearsal (repetition and substitution)	Prompt the practice or rehearsal of the behavior at least once in a context or time when its performance is not necessary to develop it into a habit.	Using the EMI, clients were prompted to complete CBT-based homework activities via the smartphone MHapp (instead of the traditional paper form) at least twice a day, once in the morning and once in the evening, in their everyday lives to ensure their implementation across various settings and mood states, including during times of relatively low negative affect, to develop it into a habit. The EMI prompt was also delivered and rehearsed before the end of each CBT session to facilitate clients in familiarizing themselves with the process [CBT session and EMI].
Reduce negative emotions (regulation)	Provide advice on methods to reduce negative emotions to facilitate the performance of behaviors.	The practitioner introduced possible strategies to reduce negative emotions (including sad mood) during the CBT sessions, with the EMI designed to prompt behaviors that facilitate in-the-moment reflections of one's mood states across social contexts and improve mood [CBT session and EMI].

[a]BCT: behaviour change technique.

[b]All BCTs listed are postulated to influence behaviors via the MoA "Beliefs about capabilities," except "Instructions on how to perform a behavior" (with mechanism of actions "Knowledge" and "Skills"), which was also coded as recommended since instructions were provided alongside the BCT "Demonstration of the behavior."

[c]CBT-EMI: cognitive behavioral therapy with smartphone-based ecological momentary intervention.

Iterative Design Process With Prototype Testing and Feedback Gathering

A prototype of the CBT-EMI, incorporating the various CBT-based homework activities, was developed using the m-Path platform (refer to Figures 2–3). The overall flow and user-friendliness of the platform were first iteratively tested within the research team before discussions with the 2 stakeholder groups.

Figure 2. Example of the CBT-EMI (cognitive behavioral therapy with smartphone-based ecological momentary intervention) MHapp prototype: Mood Check (illustrations by an experienced illustrator, Ms. Coco Lee, as part of the intervention service project).

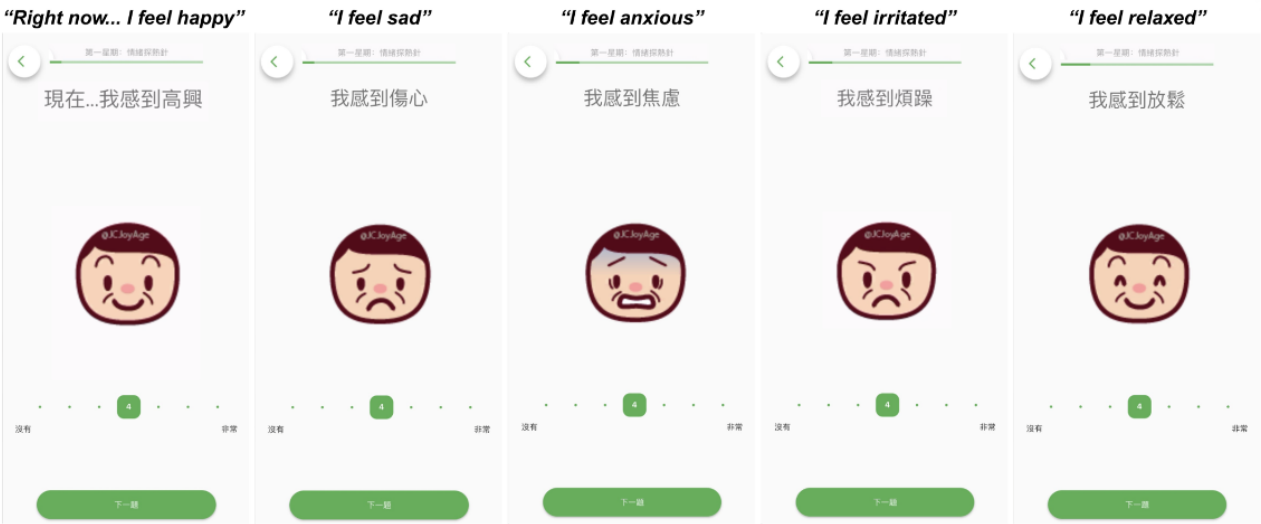

XSL•FO
RenderX

Figure 3. Example of the CBT-EMI (cognitive behavioral therapy with smartphone-based ecological momentary intervention) MHapp prototype: Hot Cross Bun.

To ensure all participants sufficiently understood how the CBT-EMI platform operated, all homework activities were tested in a separate group of 5 older adults before the process of iterative testing by participants between sessions, with detailed explanation of different features of m-Path in Session 2 (in line with other components of the TORCH principle, namely "on-site technical support," "rehearsal," and "connection with group members" [40]). Hardcopy notes were not provided in this series of workshops since all details were discussed together in the small-group setting, although they are intended to be incorporated into the implementation of CBT-EMI. To ensure equity in access to the CBT-EMI and the co-design process, smartphones were available for those whose devices were not functioning (ie, "technology provision" component of the TORCH principle [40]), although all participants used their personal smartphones throughout the 3 workshop sessions, suggesting their ubiquity even in the older adult population today.

Needs and Priorities in the Design for Blended Care

Following the Empathy Map activity, the set of needs identified by both peer supporters and service users was respectively mapped and positioned on an Eisenhower Matrix to determine their relative urgency and importance for adoption (Figure 4). The mapping of these needs on this 2×2 matrix enabled a clear visualization of how some needs were considered urgent or important by both stakeholder groups, while some were considered only urgent or important by peer supporters and not service users, and vice versa. Examples of how the MHapp could integrate features of existing services include the use of the PHQ-9 for screening and assessing depressive symptoms, as well as the use of the Wellness Recovery Action Plan to facilitate service users in developing personalized wellness plans. Users had diverse starting points, including differences in digital literacy, mental health literacy, and mental health needs.

All needs positioned in the "urgent-and-important" quadrant were adopted and implemented in the MHapp in this study. The only exceptions were those that were raised specifically by peer supporters but considered neither urgent nor important by service users.

For instance, ensuring a "user-friendly operation of the MHapp," together with the "use of careful and friendly word choices," was considered both urgent and important by both stakeholder groups and were principles we followed in designing the EMI platform. This was supplemented by the suggestion of providing concise instructions, offering assistance and support on the MHapp, alongside providing guidance and encouragement to participants from service users' perspectives. Nevertheless, whereas peer supporters considered the term "homework" to be discouraging, this was not seen as a problem by service users. Similarly, incorporating more lively messages (such as adding short videos) and features that allow interactions among users to encourage MHapp use were considered helpful by peer supporters but not service users. Since these were also features that demanded relatively more technical support, we considered these as "good-to-have" features but that could be given a lower priority at the current stage of MHapp development.

XSL•FO
RenderX

Figure 4. Prioritization of mental health app features using the Eisenhower Matrix by peer supporters (in blue) and service users (in orange).

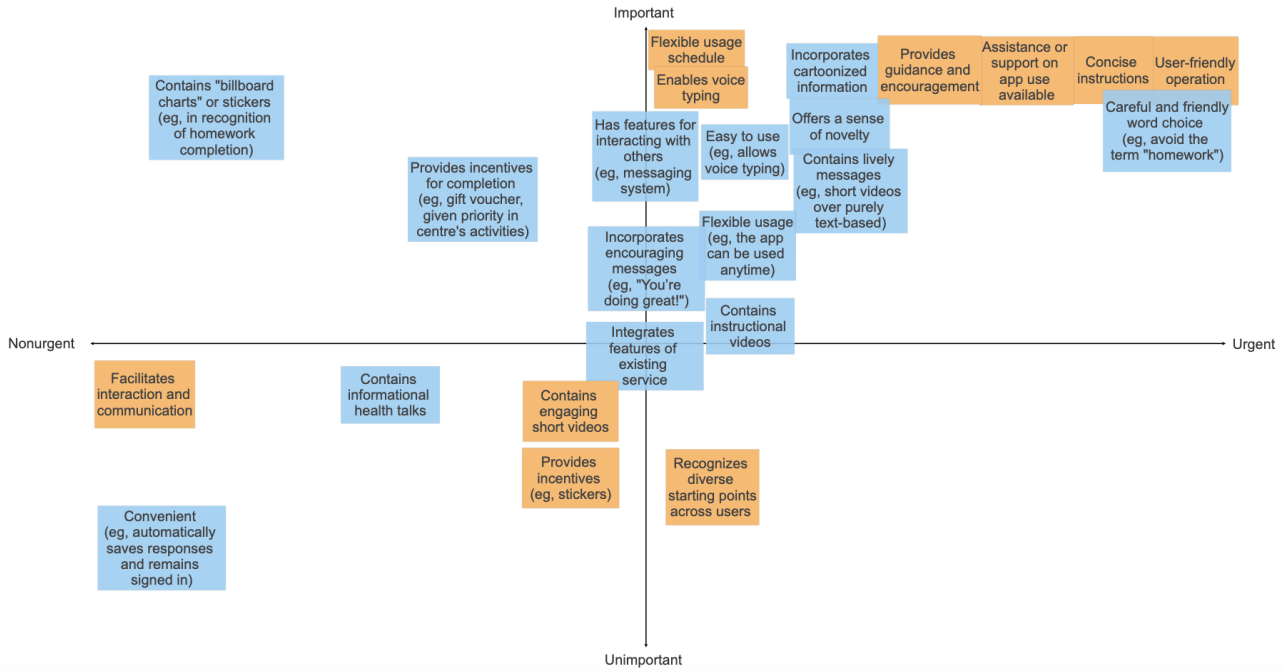

For needs identified to be urgent-and-unimportant and not–urgent-and-important, we considered their feasibility and discussed strategies to make their implementation possible. As an example, “recognizing different starting points across users” (considered urgent by service users) could be addressed by ensuring sufficient guidance and support are given to participants during the CBT-EMI, both in terms of technical and mental health support, together with an emphasis on person-centered care throughout the intervention. Further, while developing a “billboard chart” to motivate homework completion may not be fully feasible at the current stage (considered important by peer supporters), one means of achieving this aim was through providing virtual “awards” as an incentive. With the “award” function on m-Path, participants can receive virtual badges after reaching preset goals (eg, completing all beeps in a day, completing 10 surveys in total, and so on). Users can view all awards they have accumulated on their personal “award page” (refer to Figure 5).

XSL•FO
RenderX

Figure 5. Example of the "award page" in the CBT-EMI (cognitive behavioral therapy with smartphone-based ecological momentary intervention) MHapp.

"✓" represents that the MHapp has been adapted based on specific user feedback and needs, "x" represents that no adaptation was made after reconsideration of the feedback, "n/a" represents that adaptations could not be made by individual researchers or clinicians (eg, changes only possible by the developers, namely the m-Path team in this context), and "/" represents that the feedback could not yet be incorporated at the current stage due to manpower and resource considerations but may be considered at a later stage and by other research teams.

Adaptations Made to the CBT-EMI MHapp

Table 3 presents the needs and feedback generated by peer supporters and service users during Sessions 2 and 3 of the co-design workshops and whether adaptations were made accordingly. Details of how the CBT-EMI was adapted are provided in Table S3 in Multimedia Appendix 1. A total of 40 needs were summarized and categorized into the following: sampling scheme (n=6), question types (n=6), app design and UX (n=14), homework-specific feedback (n=8), and other feedback beyond the app (eg, "Incorporate additional human support" [n=6]).

Of these, 24 were from peer supporters, 7 were from service users, and 9 were from both stakeholder groups. The majority of these needs could be directly adapted for incorporation into the CBT-EMI (62.5%, n=25), and 17.5% (n=7) could not be directly adapted by individual researchers or clinicians on the platform (ie, requires amendments by the developer). A small proportion (10%, n=4) were not adapted, predominantly due to clinical and theoretical reasons (eg, since the intention of applying EMI to CBT was to encourage the enactment of CBT-based homework across settings and mood states in the real world, the prompting of 2 EMI beeps per day was retained rather than adjusting the goal to completing 12 beeps in total before the next session). For all feedback that was not adopted, we ensured the feature was acceptable to service users in the second round of workshops. An additional 4 needs identified were considered important for future blended care developments but have yet to be incorporated into the present CBT-EMI at this stage due to manpower and resource considerations (Table 3). Based on these needs and feedback, the MHapp was refined, with examples shown in Figures 6-8 (eg, "adopt briefer question types," "make the questions and prompts more straightforward," and "use figures and icons that are more age-inclusive and appear less old). For instance, in the Mood Check exercise (Figure 6), illustrations were replaced with more age-inclusive and youthful designs based on participants' feedback, with additional details provided to denote what the lowest and highest values on the slider represent. In the Hot Cross Bun exercise

XSL•FO
RenderX

(Figure 7), the triggering event item was amended from a multiple-choice to an open-text response, while the emotions item was replaced by the multismiley question. The amendments made to the interface were also in line with previous UX design guidelines [42,43].

XSL•FO
RenderX

Table . User needs and feedback, and adaptations made.

User needs and feedback		Source	Adapted?
Sampling Scheme			
1.	Enable customization of the number of beeps for EMI[a] reminders to enhance flexibility.	PS[b]	Yes
2.	Schedule 2 beeps a day (1 before noon and 1 at night).	Both	Yes
3.	Enable flexibility in responding to EMI beeps, allowing participants to complete the questionnaire later after the beep time.	Both	Yes
4.	Require the completion of a set number of EMI activities (12 in total, without specification of time or day of completion) in the subsequent 6 days before the next CBT[c] session, rather than completing 2 EMI beeps per day.	PS	No
5.	Enable participants to mute the beeps if needed.	PS	No
6.	Implement only Mood Check across all beeps to shorten overall EMI completion time.	PS	Yes
Question Types			
7.	Insufficient options in the multiple choice questions (eg, relaxation exercises and activity logging), and add an ”Other” and a free-text response option for participants who want to document their experiences in greater detail.	PS	Yes
8.	Rearrange the question flow such that questions likely to prime negative emotions are placed at the beginning.	PS	Yes
9.	Provide suggested statements for open-text questions as prompts for responses to reduce the level of difficulty.	PS	Not applicable
10.	Adopt the photo-upload question type (”Plus Point”), which appeared interesting, although the term should be amended as it is slightly confusing in this context.	PS	Yes
11.	Adopt briefer question types (eg, multiple choice questions).	SU	Yes
12.	Allow users to select multiple options in multiple choice questions to enhance their functionality.	SU	Yes
App Design and User Experience			
13.	Make the questions and prompts more straightforward (eg, questions to be more specific rather than generic).	SU	Yes
14.	Use figures and icons that are more age-inclusive and appear “less old.”	PS	Yes

XSL•FO
RenderX

User needs and feedback		Source	Adapted?
15.	Develop 2 interfaces (1 for young-old users and another for old-old users).	PS	Not applicable
16.	Incorporate a daily check-in element with encouraging messages (eg, on the home screen) to motivate participants to develop a habit of completing MHapp exercises.	PS	Not applicable
17.	Customize reminder messages to make them more engaging and attention-grabbing so that participants will notice and respond.	PS	Yes
18.	Enable users to refine their answers to previous questions in the same beep.	Both	Yes
19.	Incorporate a clearly labeled “back” button for users to move through the MHapp.	PS	Not applicable
20.	Increase button sizes within the app to accommodate smartphone screens with lower sensitivity, making it easier for users.	PS	Not applicable
21.	Add a magnifying function for participants when clicking on an image.	PS	Not applicable
22.	Implement more customized profile settings by allowing users to upload profile pictures.	PS	Not applicable
23.	Add a button popping up at the front page that activates upon completion of exercises (eg, Daily Challenge) and notify participants when exercises are completed or pending.	PS	Not applicable
24.	Visualize upcoming scheduled exercises to keep users engaged and motivated, preferably in a calendar view.	PS	Not applicable
25.	Incorporate a “prize” function with a “congratulatory sound” upon completion of an exercise.	PS	Yes
26.	Highlight that the information is from credible sources when presenting materials and exercises to address concerns about fake news and scams online.	Both	Yes
Homework-specific: Mood Check			
27.	Figures that indicate emotions should be more explicit, such as using more obvious signs for an angry face.	PS	Yes
28.	Provide more explanation on what each of the emotions means (eg, what is meant by “anxiety”), as simply selecting a score cannot help the user reflect on why those emotions occur.	PS	Yes

XSL•FO
RenderX

User needs and feedback		Source	Adapted?
29.	Even though labels are provided below the anchors (1 and 7), it was unclear what the low and high scores refer to, making it difficult to accurately reflect one's mood states.	SU	Yes
30.	Asking about mood states within a 1-hour timeframe may be too short, as events that occurred earlier could still impact a person's mood states throughout the day or even across multiple days.	SU	Yes
Behavioral Activation			
31.	Incorporate additional multimedia elements, such as short videos (eg, guided mindfulness videos) with regular updates.	Both	Yes
32.	Incorporate a wide range of easy-to-digest reading materials within the MHapp, allowing users to read at their own convenience.	Both	Yes
"Hot Cross Bun" and Gratitude Diary			
33.	Provide more space to describe the triggering events (lack of space for writing in the MHapp prototype).	SU	Yes
34.	Requiring the documentation of 1 thing or person to be grateful for instead of 3 would be more manageable. Briefly recording it on some days across a week, rather than demanding a detailed account on a specific day, will be preferable.	SU	Yes
Feedback beyond the MHapp			
35.	Instead of beeping, users using the built-in reminder from the system, SMS text messages, or WhatsApp reminders could be incorporated to increase the completion rate.	Both	Yes
36.	Provide a contact point for initial demonstration of the app and ongoing troubleshooting.	Both	Yes
37.	Incorporate additional human support (eg, from peer supporters) to guide users in using the app and check-ups for those who did not respond (eg, after 2 full days).	Both	Not applicable
38.	Match accounts to family members or peer supporters so that they can check users' records regularly.	PS	Not applicable
39.	Make the app more multifunctional (instead of a stand-alone app just for mental health), including a function to connect with others.	PS	Not applicable

XSL•FO
RenderX

User needs and feedback		Source	Adapted?
40.	Offer tangible incentives beyond the MHapp, such as financial incentives, to increase their motivation.	PS	Not applicable

[a]EMI: ecological momentary intervention.

[b]PS: peer supporter.

[c]CBT: cognitive behavioral therapy.

Figure 6. Example of the refined version of the CBT-EMI (cognitive behavioral therapy with smartphone-based ecological momentary intervention) MHapp: Mood Check (illustrations by the first author, Stephanie MY Wong).

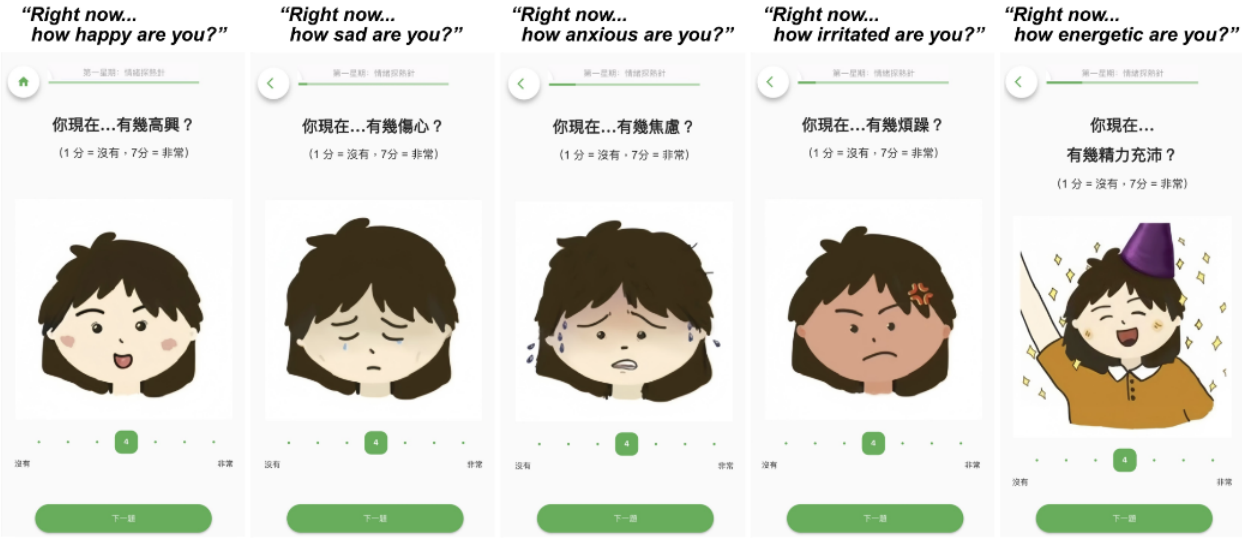

Figure 7. Example of the refined version of the CBT-EMI (cognitive behavioral therapy with smartphone-based ecological momentary intervention) MHapp: Hot Cross Bun.

XSL•FO
RenderX

Figure 8. Example of the refined version of the CBT-EMI (cognitive behavioral therapy with smartphone-based ecological momentary intervention) MHapp: Other core activities, such as the Gratitude Diary and Activity Scheduling.

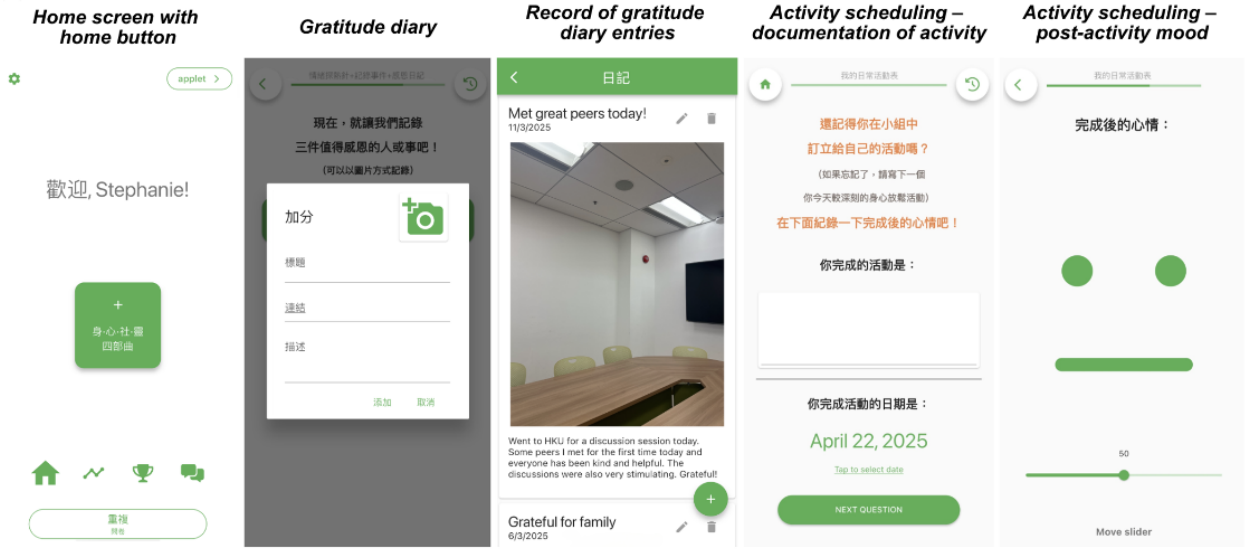

Usability Testing and User Preferences

Based on the ratings given by both peer supporters and service users, the enabling of multiple answers was generally preferred in multiple choice questions over a single answer (Table S4 in Multimedia Appendix 1). Other features of the MHapp platform were generally well accepted by both peer supporters and service users, although service users generally did not show a preference for the use of the multismiley question and instead preferred separating the various mood states into separate questions in the Mood Check. This question type was therefore only adopted for the CBT-based Hot Cross Bun activity, which was recommended by service users during the workshop.

In addition to adapting specific question types, both peer supporters and service users emphasized the importance of initial demonstration and ongoing technology support. As older adults may encounter challenges when learning new technologies, we made additional effort during app deployment to facilitate this process (eg, providing a step-by-step guide for downloading and setting up accounts, a brief walk-through of MHapp's built-in security policies, and an ongoing contact point for troubleshooting and support).

In terms of the overall acceptability and feasibility of the MHapp for CBT-EMI, both peer supporters and service users found the m-Path platform to be easy to control and use, and both the texts and instructions were clear (Table 4). Possibly because service users had previous experience in engaging in CBT or other intervention groups, their ratings were slightly lower than those of peer supporters on items related to whether the EMI questions were difficult or unclear (mean 1.40, SD 0.55 vs mean 2.88, SD 1.55), disturbance related to the number of beeps per day (mean 1.20, SD 0.45 vs mean 2.38, SD 1.51), as well as the time needed to answer the questions per beep (mean 1.40, SD 0.55 vs mean 2.88, SD 1.64). Overall, both groups of stakeholders were supportive of the integration of the m-Path platform into routine therapeutic groups to improve the application of skills learned in everyday life and also reported enjoying the co-design workshops (Table 4).

XSL•FO
RenderX

Table . Acceptability and feasibility of the blended cognitive behavioral therapy with smartphone-based ecological momentary intervention (CBT-EMI).

Acceptability and feasibility (1=not at all, 7=very)	Peer supporters (n=8)	Service users (n=5)
App-related, mean (SD)		
Was the m-Path difficult to control? (reversed)	1.38 (0.52)	1.20 (0.45)
Was it difficult to switch on the m-Path app? (reversed)	1.38 (0.52)	1.00 (0.00)
I found the m-Path app easy to use.	6.00 (0.93)	6.00 (0.71)
EMI content and design-related, mean (SD)		
Was the text on the screen readable and clear?	6.00 (0.76)	6.40 (0.55)
Were the verbal instructions you received about using m-Path clear?	5.38 (2.00)	6.20 (0.45)
Were the written instructions you received on m-Path clear?	5.88 (1.46)	6.20 (0.84)
Were the questions you answered on m-Path difficult or unclear? (reversed scored)	2.88 (1.55)	1.40 (0.55)
Did you find it annoying or stressful to use m-Path with respect to..., mean (SD)		
…the number of beeps per day? (reversed scored)	2.38 (1.51)	1.20 (0.45)
…the time it took to answer the questions for a single beep? (reversed scored)	2.88 (1.64)	1.40 (0.55)
…the noise or sound volume? (reversed scored)	1.75 (1.16)	1.20 (0.45)
Blended care–related, mean (SD)		
I agree to integrate m-Path into routine therapy groups to provide an opportunity to apply the skills learned in the groups to daily life.	6.25 (0.71)	5.60 (1.14)
Overall workshops, mean (SD)		
I enjoyed the process of the 3 co-design workshops.	6.38 (0.74)	6.60 (0.55)

Discussion

Principal Findings

The integration of digital-assisted support, such as smartphone apps, into traditional in-person psychological interventions is a promising direction toward improving mental health outcomes. Yet, the challenges posed by the development of new MHapps for specific populations, including not only the high costs involved but also the need for specialized manpower and time, make the implementation and scaling up of blended care with smartphone-based technologies difficult. A potential option is to use existing MHapps; yet, our review showed that none of the 149 possible apps supported the full content adaptation necessary for localized, culturally responsive interventions. This observation highlights a significant gap in the market and underscores the need for highly adaptable and customizable platforms.

Leveraging existing no-code app development platforms, our study demonstrated the feasibility of integrating smartphone-based EMI with traditional CBT via a phased, methodologically rigorous approach. Specifically, the final platform incorporated core CBT-based activities, such as "Mood Check" and "Gratitude Diary," which are aligned with established mechanisms of change (eg, behavioral activation and activity scheduling [44]) and the HAPA [52,53]. Evidence-based BCTs targeting self-efficacy were also incorporated to improve the enactment of CBT-based homework and its sustainment in everyday life contexts. By describing this iterative design and development process in detail, such culturally adapted, person-centered blended mental health care models have the potential for implementation not only for older adults in Hong Kong but for broader populations and contexts, including time- and resource-limited settings.

Using co-design activities, such as the Eisenhower Matrix to prioritize feedback, 40 user needs were identified—62.5% of which were implemented. These included simplified instructions and interface, the use of friendly language, and gamified elements (eg, virtual badges to reinforce engagement), which are also key features identified as important by users in other studies with young people [56,57] and are incorporated as part of the recommendations for MHapps in a recent review [16]. While some suggestions were deferred due to technical or clinical constraints, stakeholder evaluations rated the platform as intuitive and clinically relevant. Preferences such as separating mood-state questions (rather than combining them into a single scale) were particularly noted. Both peer supporters and service users rated the platform highly for ease of use and clarity. Importantly, the platform demonstrated strong feasibility

XSL•FO
RenderX

for integration into group therapy settings to reinforce real-world application of therapeutic skills. Overall, blended CBT-EMI interventions appear promising for therapeutic engagement, homework adherence, and symptom management, particularly among older adults in our study context.

Strengths and Limitations

This study has important implications for researchers, clinicians, and digital health developers. A major strength lies in its systematic and theory-driven development process. The phased methodology—comprising needs assessment, platform selection, cultural and linguistic adaptation, and co-design—offers a replicable model for future interventions. Theoretical integration is another notable strength; the intervention is grounded in CBT and enriched by HAPA, aligning clinical techniques with behavioral science to support sustained change. The detailed mapping of how various BCTs were incorporated in the CBT-EMI enables a more fine-grained investigation of their independent effects on clinical outcomes in future studies, which can be conducive to efforts in the identification of "active ingredients" in psychological interventions [45].

The blended care intervention we presented further benefits from its strong cultural relevance. Stakeholder engagement was central to the adaptation process, ensuring the platform was sensitive to local needs and user expectations. By addressing user interface design, motivational features, and language preferences, we were able to develop a culturally attuned and user-friendly solution for mental health care in the real-world setting. These features are critical for maintaining long-term engagement among users. The co-design process also greatly facilitated stakeholder-driven adaptations and potentially enhanced the feasibility of implementation in routine care.

Nonetheless, we note several limitations. While the Mood Check items we adopted are similar to those commonly used in ESM and EMI among adolescents and adults, the present blended care platform was co-designed only with Cantonese-speaking older adults in Hong Kong. Additional rounds of co-design processes may be helpful to ensure its potential application to other age, cultural, or linguistic groups. In particular, we highlight the significance of considering the use of culturally sensitive mediums [58], local languages and dialects [59], as well as barrier-free design (eg, allowing audio input as an alternative to text-based options [60]). While the sample size of this study is in line with guidelines on performance-based usability testing [42] and is comparable to other studies using co-design methods in the development of health interventions [61,62], we acknowledge that perspectives beyond the 13 participants involved may not be captured (eg, those with less motivation to participate in the study). We did, nevertheless, aim to include participants with more diverse sociodemographic backgrounds and experiences in psychotherapy to capture a wider range of perspectives. Whether additional amendments may be needed to improve the usability of the platform for a majority of the population would require further exploration.

We also note that several user-identified needs were not implemented due to platform limitations, pointing to future opportunities for enhanced collaboration between researchers and technology developers. Given the lack of no-code app platforms available that were specifically designed for EMI, we selected m-Path for building the EMI for the present blended care. We acknowledge that the focus on a single vendor could have possible issues related to the long-term sustainability of the CBT-EMI. Yet, the aim of this study was to provide a transferable framework that could guide researchers or stakeholders in the co-design of a digital mental health platform, particularly in settings where the development of a new app is not feasible. The insights we gathered into the prioritization of MHapp features and user needs also offer a reference for future studies in the design of more age-friendly MHapps targeting the older adult population. Given the relatively short study period, we cannot determine whether individuals would continue using the platform beyond the intervention period. Whether individuals could successfully develop habits in applying CBT-based techniques in their everyday lives without the continued use of the platform postintervention would be an area for future investigation.

In line with the IDEAS framework [31], empirical validation through randomized controlled trials with a longitudinal follow-up would be an important next step to confirm the efficacy of the CBT-EMI in terms of symptom reduction, adherence, and long-term outcomes. In particular, it would be worthwhile to examine whether people receiving the novel blended care—as informed by both cognitive-behavioral and health behavior change theories—will show greater symptom improvements and higher rates of clinical remission and response (eg, using the PHQ-9 for evaluating DSM-based depressive symptoms) compared with those receiving regular in-person CBT without the addition of EMI. It would also be crucial to test the hypothesized mediating role of self-efficacy in these future studies to elucidate the mechanisms of change in the CBT-EMI.

Conclusions

Despite the abundance of MHapps developed, the observation that none could be readily adopted to target specific clinical or therapeutic needs across settings may substantially prevent the scaling up of digitally assisted mental health care across populations. By demonstrating how no-code platforms can accelerate the development of theory-based interventions and how blended CBT-EMI approaches can effectively extend therapeutic engagement beyond clinical settings, this work presents a replicable roadmap for developing scalable, person-centered digital mental health interventions. It would be essential for future work to test the efficacy of the blended care intervention in large-scale randomized controlled trial studies.

XSL•FO
RenderX

Acknowledgments

We would like to thank all participants for contributing their time to co-design the novel MHapp, which formed the basis of this work.

Funding

This study was funded by the HKU University Research Committee (Seed Fund for New Staff Program) to SMYW and by The Hong Kong Jockey Club Charities Trust for the University of Hong Kong for the project JC JoyAge (AR24AG100291).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional material.
[PDF File, 588 KB - formative_v9i1e77036_app1.pdf]

References

1. GBD 2019 Mental Disorders Collaborators. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet Psychiatry 2022 Feb;9(2):137-150. [doi: 10.1016/S2215-0366(21)00395-3]
2. Herrman H, Kieling C, McGorry P, Horton R, Sargent J, Patel V. Reducing the global burden of depression: a Lancet-World Psychiatric Association Commission. Lancet 2019 Jun 15;393(10189):e42-e43. [doi: 10.1016/S0140-6736(18)32408-5] [Medline: 30482607]
3. Newman MG, Szkodny LE, Llera SJ, Przeworski A. A review of technology-assisted self-help and minimal contact therapies for anxiety and depression: is human contact necessary for therapeutic efficacy? Clin Psychol Rev 2011 Feb;31(1):89-103. [doi: 10.1016/j.cpr.2010.09.008] [Medline: 21130939]
4. Torous J, Bucci S, Bell IH, et al. The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. World Psychiatry 2021 Oct;20(3):318-335. [doi: 10.1002/wps.20883] [Medline: 34505369]
5. Wong SMY, Chen EYH. Significance of timely responses to mental health crises during unprecedented societal and global changes using an agile, empathic and user-centred digital approach: lessons from the Hong Kong experience. BMJ Glob Health 2023 Aug;8(8):e013540. [doi: 10.1136/bmjgh-2023-013540] [Medline: 37612035]
6. Suen YN, Hui CLM, Lei LKS, et al. Low-intensity online intervention for mental distress among help-seeking young people in Hong Kong: a randomized clinical trial. JAMA Netw Open 2025 Jan 2;8(1):e2454675. [doi: 10.1001/jamanetworkopen.2024.54675] [Medline: 39813032]
7. Mak WWS, Ng SM, Leung FHT. A web-based stratified stepped care platform for mental well-being (TourHeart+): user-centered research and design. JMIR Form Res 2023 Mar 22;7:e38504. [doi: 10.2196/38504] [Medline: 36947112]
8. Mausbach BT, Moore R, Roesch S, Cardenas V, Patterson TL. The relationship between homework compliance and therapy outcomes: an updated meta-analysis. Cognit Ther Res 2010 Oct;34(5):429-438. [doi: 10.1007/s10608-010-9297-z] [Medline: 20930925]
9. Tang W, Kreindler D. Supporting homework compliance in cognitive behavioural therapy: essential features of mobile apps. JMIR Ment Health 2017 Jun 8;4(2):e20. [doi: 10.2196/mental.5283] [Medline: 28596145]
10. Prasko J, Krone I, Burkauskas J, et al. Homework in cognitive behavioral supervision: theoretical background and clinical application. Psychol Res Behav Manag 2022;15:3809-3824. [doi: 10.2147/PRBM.S382246] [Medline: 36578283]
11. Middleton KR, Anton SD, Perri MG. Long-term adherence to health behavior change. Am J Lifestyle Med 2013;7(6):395-404. [doi: 10.1177/1559827613488867] [Medline: 27547170]
12. Kwasnicka D, Dombrowski SU, White M, Sniehotta F. Theoretical explanations for maintenance of behaviour change: a systematic review of behaviour theories. Health Psychol Rev 2016 Sep;10(3):277-296. [doi: 10.1080/17437199.2016.1151372] [Medline: 26854092]
13. Der-Avakian A, Barnes SA, Markou A, Pizzagalli DA. Translational assessment of reward and motivational deficits in psychiatric disorders. Curr Top Behav Neurosci 2016;28:231-262. [doi: 10.1007/7854_2015_5004] [Medline: 26873017]
14. Fuhrmann LM, Weisel KK, Harrer M, et al. Additive effects of adjunctive app-based interventions for mental disorders - a systematic review and meta-analysis of randomised controlled trials. Internet Interv 2024 Mar;35:100703. [doi: 10.1016/j.invent.2023.100703] [Medline: 38225971]

XSL•FO
RenderX

15. Ferrao Nunes-Zlotkowski K, Shepherd HL, Beatty L, Butow P, Shaw JM. Blended psychological therapy for the treatment of psychological disorders in adult patients: systematic review and meta-analysis. Interact J Med Res 2024 Oct 29;13:e49660. [doi: 10.2196/49660] [Medline: 39470720]
16. Bakker D, Kazantzis N, Rickwood D, Rickard N. Mental health smartphone apps: review and evidence-based recommendations for future developments. JMIR Ment Health 2016 Mar 1;3(1):e7. [doi: 10.2196/mental.4984] [Medline: 26932350]
17. Myin-Germeys I, Klippel A, Steinhart H, Reininghaus U. Ecological momentary interventions in psychiatry. Curr Opin Psychiatry 2016 Jul;29(4):258-263. [doi: 10.1097/YCO.0000000000000255] [Medline: 27153125]
18. Kramer I, Simons CJP, Hartmann JA, et al. A therapeutic application of the experience sampling method in the treatment of depression: a randomized controlled trial. World Psychiatry 2014 Feb;13(1):68-77. [doi: 10.1002/wps.20090] [Medline: 24497255]
19. Myin-Germeys I, Schick A, Ganslandt T, et al. The experience sampling methodology as a digital clinical tool for more person-centered mental health care: an implementation research agenda. Psychol Med 2024 Aug;54(11):2785-2793. [doi: 10.1017/S0033291724001454]
20. Marciniak MA, Shanahan L, Rohde J, et al. Standalone smartphone cognitive behavioral therapy-based ecological momentary interventions to increase mental health: narrative review. JMIR Mhealth Uhealth 2020 Nov 12;8(11):e19836. [doi: 10.2196/19836] [Medline: 33180027]
21. Schueller SM, Aguilera A, Mohr DC. Ecological momentary interventions for depression and anxiety. Depress Anxiety 2017 Jun;34(6):540-545. [doi: 10.1002/da.22649] [Medline: 28494123]
22. Chakrabarti S. Digital psychiatry in low-and-middle-income countries: new developments and the way forward. World J Psychiatry 2024 Mar 19;14(3):350-361. [doi: 10.5498/wjp.v14.i3.350] [Medline: 38617977]
23. Chou YH, Lin C, Lee SH, Chang Chien YW, Cheng LC. Potential mobile health applications for improving the mental health of the elderly: a systematic review. Clin Interv Aging 2023;18:1523-1534. [doi: 10.2147/CIA.S410396] [Medline: 37727447]
24. Titzler I, Saruhanjan K, Berking M, Riper H, Ebert DD. Barriers and facilitators for the implementation of blended psychotherapy for depression: a qualitative pilot study of therapists' perspective. Internet Interv 2018 Jun;12:150-164. [doi: 10.1016/j.invent.2018.01.002] [Medline: 30135779]
25. Borghouts J, Eikey E, Mark G, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. J Med Internet Res 2021 Mar 24;23(3):e24387. [doi: 10.2196/24387] [Medline: 33759801]
26. Balis LE, Houghtaling B. Matching barriers and facilitators to implementation strategies: recommendations for community settings. Implement Sci Commun 2023 Nov 21;4(1):144. [doi: 10.1186/s43058-023-00532-1] [Medline: 37990243]
27. Miralles I, Granell C, Díaz-Sanahuja L, et al. Smartphone apps for the treatment of mental disorders: systematic review. JMIR Mhealth Uhealth 2020 Apr 2;8(4):e14897. [doi: 10.2196/14897] [Medline: 32238332]
28. Linardon J, Torous J, Firth J, Cuijpers P, Messer M, Fuller-Tyszkiewicz M. Current evidence on the efficacy of mental health smartphone apps for symptoms of depression and anxiety. A meta-analysis of 176 randomized controlled trials. World Psychiatry 2024 Feb;23(1):139-149. [doi: 10.1002/wps.21183] [Medline: 38214614]
29. Willms A, Rush J, Hofer S, Rhodes RE, Liu S. Advancing physical activity research methods using real-time and adaptive technology: a scoping review of "no-code" mobile health app research tools. Sport Exerc Perform Psychol 2024;14(1):250-267. [doi: 10.1037/spy0000360]
30. Liu S, La H, Willms A, Rhodes RE. A "No-Code" app design platform for mobile health research: development and usability study. JMIR Form Res 2022 Aug 18;6(8):e38737. [doi: 10.2196/38737] [Medline: 35980740]
31. Mummah SA, Robinson TN, King AC, Gardner CD, Sutton S. IDEAS (Integrate, Design, Assess, and Share): a framework and toolkit of strategies for the development of more effective digital interventions to change health behavior. J Med Internet Res 2016 Dec 16;18(12):e317. [doi: 10.2196/jmir.5927] [Medline: 27986647]
32. Willms A, Rhodes RE, Liu S. The development of a hypertension prevention and financial-incentive mHealth program using a "No-Code" mobile app builder: development and usability study. JMIR Form Res 2023 Apr 5;7:e43823. [doi: 10.2196/43823] [Medline: 37018038]
33. Liu T, Leung DKY, Lu S, et al. Collaborative community mental health and aged care services with peer support to prevent late-life depression: study protocol for a non-randomised controlled trial. Trials 2022 Apr 11;23(1):280. [doi: 10.1186/s13063-022-06122-1] [Medline: 35410292]
34. Wong SMY, Leung DKY, Liu T, et al. Comorbid anxiety, loneliness, and chronic pain as predictors of intervention outcomes for subclinical depressive symptoms in older adults: evidence from a large community-based study in Hong Kong. BMC Psychiatry 2024 Nov 21;24(1):839. [doi: 10.1186/s12888-024-06281-2] [Medline: 39574082]
35. Bernal G, Bonilla J, Bellido C. Ecological validity and cultural sensitivity for outcome research: issues for the cultural adaptation and development of psychosocial treatments with Hispanics. J Abnorm Child Psychol 1995 Feb;23(1):67-82. [doi: 10.1007/BF01447045] [Medline: 7759675]
36. Bernal G, Sáez - Santiago E. Culturally centered psychosocial interventions. Journal Community Psychology 2006 Mar;34(2):121-132. [doi: 10.1002/jcop.20096]

XSL•FO
RenderX

37. Leung DKY, Fong APC, Kwok WW, et al. A cultural adaptation of acceptance and commitment therapy with exercise for older Chinese with chronic pain and depressive symptoms: a mixed-methods study. Psychotherapy (Chic) 2024 Dec;61(4):269-281. [doi: 10.1037/pst0000539] [Medline: 39509218]
38. Liu T, Kanagawa HS, Lee JKQ, et al. Power of creative arts for older people with chronic pain and depressive symptoms: co-development of the Rewire with Arts protocol and findings from a pilot study. BJPsych Open 2025 Jun 27;11(4):e132. [doi: 10.1192/bjo.2025.10062] [Medline: 40574612]
39. Wang YH, Wang YL, Leung DKY, et al. Effectiveness of an age-modified mindfulness-based cognitive therapy (MBCT) in improving mental health in older people with depressive symptoms: a non-randomised controlled trial. BMC Complement Med Ther 2025 Feb 26;25(1):81. [doi: 10.1186/s12906-025-04781-6] [Medline: 40011881]
40. Leung DKY, Wong NHL, Yau JHY, et al. Hybrid-delivered community psychoeducation for people aged 50 and older: a mixed-method evaluation and lesson learned. Internet Interv 2024 Mar;35:100699. [doi: 10.1016/j.invent.2023.100699] [Medline: 38174209]
41. Leung DKY, Wong FHC, Wong ELY, et al. Technology affordance in an information and communication technology delivered group psychotherapy and exercise program for older people with depressive symptoms: a multiple triangulation qualitative study. Innov Aging 2023;7(6):igad063. [doi: 10.1093/geroni/igad063] [Medline: 37593232]
42. Research-based web design & usability guidelines. : U.S. Department of Health and Human Services; 2006 URL: https://www.hhs.gov/sites/default/files/research-based-web-design-and-usability-guidelines_book.pdf [accessed 2025-11-14]
43. Elderly-friendly website/mobile application design guide. : Digital Policy Office, The Government of the Hong Kong Special Administrative Region of the People's Republic of China; 2024 URL: https://www.digitalpolicy.gov.hk/en/our_work/digital_government/digital_inclusion/accessibility/promulgating_resources/application_design_guide/doc/elderly_friendly_design_guide_eng.pdf [accessed 2025-11-14]
44. Gautam M, Tripathi A, Deshmukh D, Gaur M. Cognitive behavioral therapy for depression. Indian J Psychiatry 2020 Jan;62(Suppl 2):S223-S229. [doi: 10.4103/psychiatry.IndianJPsychiatry_772_19] [Medline: 32055065]
45. Wolpert M, Pote I, Sebastian CL. Identifying and integrating active ingredients for mental health. Lancet Psychiatry 2021 Sep;8(9):741-743. [doi: 10.1016/S2215-0366(21)00283-2]
46. Michie S, Wood CE, Johnston M, Abraham C, Francis JJ, Hardeman W. Behaviour change techniques: the development and evaluation of a taxonomic method for reporting and describing behaviour change interventions (a suite of five studies involving consensus methods, randomised controlled trials and analysis of qualitative data). Health Technol Assess 2015 Nov;19(99):1-188. [doi: 10.3310/hta19990] [Medline: 26616119]
47. Michie S, Johnston M, Francis J, Hardeman W, Eccles M. From theory to intervention: mapping theoretically derived behavioural determinants to behaviour change techniques. Applied Psychology 2008 Oct;57(4):660-680 [FREE Full text] [doi: 10.1111/j.1464-0597.2008.00341.x]
48. Tang JPS, Liu T, Lu S, et al. "It was the deepest level of companionship": peer-to-peer experience of supporting community-dwelling older people with depression - a qualitative study. BMC Geriatr 2022 May 19;22(1):443. [doi: 10.1186/s12877-022-03121-4] [Medline: 35590279]
49. Joe J, Chaudhuri S, Le T, Thompson H, Demiris G. The use of think-aloud and instant data analysis in evaluation research: exemplar and lessons learned. J Biomed Inform 2015 Aug;56:284-291. [doi: 10.1016/j.jbi.2015.06.001] [Medline: 26071683]
50. Mestdagh M, Verdonck S, Piot M, et al. m-Path: an easy-to-use and highly tailorable platform for ecological momentary assessment and intervention in behavioral research and clinical practice. Front Digit Health 2023;5:1182175. [doi: 10.3389/fdgth.2023.1182175] [Medline: 37920867]
51. Golsong K, Kaufmann L, Baldofski S, Kohls E, Rummel-Kluge C. Acceptability, user satisfaction, and feasibility of an app-based support service during the COVID-19 pandemic in a psychiatric outpatient setting: prospective longitudinal observational study. JMIR Form Res 2024 Dec 4;8:e60461. [doi: 10.2196/60461] [Medline: 39630503]
52. Schwarzer R. Modeling health behavior change: how to predict and modify the adoption and maintenance of health behaviors. Applied Psychology 2008 Jan;57(1):1-29. [doi: 10.1111/j.1464-0597.2007.00325.x]
53. Schwarzer R, Hamilton K. Changing Behavior Using the Health Action Process Approach The Handbook of Behavior Change: Cambridge University Press; 2020:89-103. [doi: 10.1017/9781108677318.007]
54. Carey RN, Connell LE, Johnston M, et al. Behavior change techniques and their mechanisms of action: a synthesis of links described in published intervention literature. Ann Behav Med 2019 Jul 17;53(8):693-707. [doi: 10.1093/abm/kay078] [Medline: 30304386]
55. Michie S, Richardson M, Johnston M, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013 Aug;46(1):81-95. [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
56. Christie GI, Shepherd M, Merry SN, Hopkins S, Knightly S, Stasiak K. Gamifying CBT to deliver emotional health treatment to young people on smartphones. Internet Interv 2019 Dec;18:100286. [doi: 10.1016/j.invent.2019.100286] [Medline: 31890633]
57. Alqahtani F, Winn A, Orji R. Co-designing a mobile app to improve mental health and well-being: focus group study. JMIR Form Res 2021 Feb 26;5(2):e18172. [doi: 10.2196/18172] [Medline: 33635281]

XSL•FO
RenderX

58. Litke SG, Resnikoff A, Anil A, et al. Mobile technologies for supporting mental health in youths: scoping review of effectiveness, limitations, and inclusivity. JMIR Ment Health 2023 Aug 23;10:e46949. [doi: 10.2196/46949] [Medline: 37610818]
59. Baulch E, Watkins J, Tariq A. mHealth Innovation in Asia: Grassroots Challenges and Practical Interventions: Springer; 2017. [doi: 10.1007/978-94-024-1251-2]
60. Burchert S, Alkneme MS, Bird M, et al. User-centered app adaptation of a low-intensity e-Mental health intervention for Syrian refugees. Front Psychiatry 2018;9:663. [doi: 10.3389/fpsyt.2018.00663] [Medline: 30740065]
61. Derksen ME, Jaspers MW, van Strijp S, Fransen MP. Mobile health for smoking cessation among disadvantaged young women during and after pregnancy: user-centered design and usability study. JMIR Form Res 2021 Aug 4;5(8):e24112. [doi: 10.2196/24112] [Medline: 34346895]
62. Nuseibeh BZ, Johns SA, Shih PC, Lewis GF, Gowan TM, Jordan EJ. Co-designing the MOSAIC mHealth app with breast cancer survivors: user-centered design approach. JMIR Form Res 2024 Dec 9;8:e59426. [doi: 10.2196/59426] [Medline: 39652855]

Abbreviations

BCT: behavior change technique
CBT: cognitive behavioral therapy
EMI: ecological momentary intervention
ESM: experience sampling method
HAPA: Health Action Process Approach
IDEAS: Integrate, Design, Assess, and Share
MHapp: mental health app
PHQ-9: Patient Health Questionnaire-9
TORCH: Technology provision, On-site technical support, Rehearsal, Connection with group members, and Hardcopy notes
UX: user experience

Edited by A Mavragani; submitted 14.05.25; peer-reviewed by C Obianyo, T Ojo; revised version received 15.09.25; accepted 16.09.25; published 21.11.25.

Please cite as:
Wong SMY, Ip MHC, Lee JKQ, Lum TYS, Schwarzer R
Leveraging No-Code Digital Platforms for Designing an Integrated Smartphone-Based Ecological Momentary Intervention With Cognitive Behavioral Therapy for Mental Health Care: Development and Usability Study
JMIR Form Res 2025;9:e77036
URL: https://formative.jmir.org/2025/1/e77036
doi:10.2196/77036

© Stephanie Ming Yin Wong, Melody Ho Ching Ip, Jessica Kang Qi Lee, Terry Yat Sang Lum, Ralf Schwarzer. Originally published in JMIR Formative Research (https://formative.jmir.org), 21.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Chatbots' Empathetic Conversations and Responses: A Qualitative Study of Help-Seeking Queries on Depressive Moods Across 8 Commercial Conversational Agents

Hyojin Chin[1*], PhD; Gumhee Baek[2*], MSN; Chiyoung Cha[3], PhD; Meeyoung Cha[4,5], PhD

[1]Department of Computer Science and Engineering, Gyeongsang National University, Jinju, Republic of Korea

[2]College of Nursing, System Health Science and Engineering Program, Ewha Womans University, Seoul, Republic of Korea

[3]College of Nursing, Ewha Research Institute of Nursing Science, System Health and Engineering Major in Graduate School, Ewha Womans University, Seoul, Republic of Korea

[4]Max Planck Institute for Security and Privacy, Universitätsstraße 140, Bochum, Germany

[5]School of Computing, Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Meeyoung Cha, PhD
Max Planck Institute for Security and Privacy, Universitätsstraße 140, Bochum, Germany

Abstract

Background: While recent studies showed the potential of conversational agents (CAs) to help alleviate depressive moods, the dynamics of user-chatbot interactions in mental health support remain underexplored.

Objective: We examine real-world conversations between users and chatbots on depression-related topics to identify patterns in how users seek help and how chatbots provide therapeutic support. We analyzed the responses of 8 commercial chatbots to user queries about depressive moods, examining whether they incorporated therapeutic communication techniques, such as empathy.

Methods: Our method has 2 parts. First, we analyzed 13,700 utterances (6850 user queries and 6850 responses) about depressive moods from the commercial chatbot SimSimi, covering 5 English-speaking countries between 2016 and 2021. Using a human-annotated coding approach, we classified user queries into 5 groups based on Rickwood's help-seeking model and classified chatbot responses into 8 therapeutic communication styles. Empathy was assessed as one of these styles, with responses coded as empathetic when they demonstrated emotional understanding, validation, and reflection. Next, we evaluated the responses of 3 voice assistants (Amazon's Alexa, Google Assistant, and Apple's Siri) and 5 chatbots (ChatGPT, Replika, Woebot, Wysa, and SimSimi) to user queries about depressive moods.

Results: In study 1, we examined how SimSimi, a social chatbot trained to encourage users to share their emotions and build rapport, responded to user queries. The majority (3067/4073, 75.3%) indicated depressed feelings, and a smaller portion (168/4073, 4.1%) sought strategies to cope with depression. The chatbot's responses were largely therapeutic (2417/3108, 77.7%), demonstrating empathy (902/3108, 29%), active listening (836/3108, 26.9%), and open-ended questions (679/3108, 21.8%). In study 2, we qualitatively compared response patterns across commercial CAs, revealing that Replika expressed empathy in more than 75% (28/36) of its responses, similar to SimSimi. In contrast, Alexa (15/17, 88.2%), Google Assistant (18/30, 60%), Siri (20/36, 55.6%), and ChatGPT (40/42, 95.2%) typically responded to depression-related queries with search results rather than offering specific solutions for depressive feelings. Mental health chatbots such as Woebot responded to users with clarification questions (97.3%). We also report instances where CAs failed to meet users' help-seeking needs, instead giving irrelevant responses and ignoring emotional requests.

Conclusions: Our findings reveal a mixed landscape in the emotional support provided by CAs. While some social chatbots delivered empathetic responses that fostered deeper user engagement, most commercial chatbots offered merely informative replies to users' help-seeking inputs. Recognizing that users seek support from chatbots, we recommend equipping next-generation CAs with capabilities grounded in therapeutic communication, such as empathetic responses.

(JMIR Form Res 2025;9:e71538) doi:10.2196/71538

KEYWORDS

chatbot; conversational agent; depressive moods; depression; help-seeking; therapeutic communication; social support; qualitative study

XSL•FO
RenderX

Introduction

Background

Depression is a severe illness that affects around 280 million people globally [1]. The COVID-19 pandemic worsened the situation, with an estimated 27.6% increase in cases of major depressive disorder [2]. Individuals with depression symptoms may avoid seeking medical help due to cost, time, possible side effects of medication, and social stigma [3]. The resulting unmet need for depression support has increased the demand for alternative services. For example, the number of downloads of online mental health apps increased by 17.6% after the pandemic [4,5], indicating a public need for easy-to-access mental support.

Chatbots are gaining popularity in the mental health domain, with studies showing their positive effects on reducing depression, stress, and anxiety [6,7]. Research suggests that many individuals are more willing to engage with artificial intelligence (AI) agents, such as chatbots, because of the anonymity they offer and the perception of a "safe space" for discussing sensitive issues without fear of judgment [8]. This report highlights that AI agents can serve as an alternative to human counselors for those struggling with depression and anxiety, especially in cultures where mental health stigma is prevalent [9,10]. In addition, chatbots offer users 24/7 social support and can facilitate mental health management [11].

In response, the industry has swiftly developed commercial chatbots such as Woebot, Wysa, and Vivibot, which provide cognitive behavioral therapy (CBT), behavioral reinforcement, and mindfulness [11]. A study examining the AI-based mobile conversation app, Wysa, indicated that therapeutic alliance scores improved with increasing conversation time [12]. Furthermore, research on Tess, a chat-based app that offers emotional support and psychological education, demonstrated a significant reduction in user anxiety upon engagement [13]. An analysis of 152,783 conversations related to depressive moods and sadness from 4 years and 8 months of conversation data between commercial chatbots and users revealed that users were more open to expressing emotional vulnerability related to depressive or sad moods to chatbots than on social media [14].

Despite the exciting potential of AI chatbots as a source of support in mental health domains, most research has focused on technical functionalities [15] and the outcome of user interactions (eg, depression reduction) [12,16]. Some studies focused on mental health issues have used discourse analysis to examine data from platforms such as Twitter and Reddit [17,18]. Other research has analyzed user discourse, revealing distinct patterns of depressive mood expression among social chatbot users compared with social media users, with a tendency to disclose depressed and sad emotional states to chatbots [14]. These studies suggest that valuable insights can be gained from examining user queries. However, what is currently lacking is a discourse-based analysis of conversational AI systems to explore how they might respond therapeutically to diverse user needs related to depressive moods.

In this study, we address the gap by examining how users with depressive moods seek help from chatbots and how various commercial chatbots respond therapeutically. We pose the following research questions: (RQ1) How do users initiate help-seeking conversations related to depressive moods with online chatbots? (RQ2) Do commercial conversational agents (CAs) respond therapeutically to help-seeking user input about depressive moods, and if so, how? These research questions are important because they provide complementary perspectives on help-seeking interactions. While RQ1 focuses on understanding how users with depressive moods initiate emotional disclosures, RQ2 explores how commercial CAs respond to such expressions. This understanding is critical for designing chatbots that can effectively recognize and respond to users' emotional states. In addition, comparing the diverse responses across commercial products is important for assessing their current capabilities and limitations.

We partnered with SimSimi, one of the largest chatbot services, to categorize help-seeking behaviors related to depression. We then conducted our experiments with 8 commercial CAs: 3 voice assistants (Amazon's Alexa, Google Assistant, and Apple's Siri) and 5 chatbots (ChatGPT, Replika, Woebot, Wysa, and SimSimi). Users today frequently encounter a variety of CAs in everyday life, accessed through different modalities. For example, they may interact with Alexa via a stationary smart speaker at home, use Siri on a smartphone, or engage with personalized chatbots such as SimSimi or Replika through mobile apps. In addition, users may choose to converse with advanced AI systems such as ChatGPT or similar large language model (LLM)–based chatbots offering high-performance dialogue capabilities. However, in this study, we did not delve into the differences in interaction modalities (eg, text vs voice) or the technical performance variations across different AI models. Instead, we selected candidate CAs based on overall usability and accessibility from the user's perspective, aiming to reflect general user experiences with a diverse yet representative range of CAs.

Our findings reveal that while some of these services assist users by suggesting solutions, asking clarifying questions, and offering empathetic responses, others struggle to meet users' help-seeking needs. We observed instances where services failed to adequately address user requests, generating irrelevant responses that neglected users' emotional needs. Given that most current CAs primarily provide information, a crucial goal is to develop empathetic CAs capable of using therapeutic communication methods to effectively support users' social needs. Our findings offer critical evidence on the type of assistance users expect from chatbots and provide valuable insights for future chatbot designs.

Preliminaries

Chatbots and Mental Health

The widespread adoption of social media has enabled people to easily express their thoughts and emotions [19]. Studies have shown that expressing emotions on social media can reduce depression, anxiety, and loneliness while increasing overall life satisfaction by helping individuals cope with negative emotions [20]. Social media usage patterns are also known to signal

XSL•FO
RenderX

mental health status. For example, Facebook users with depression or anxiety symptoms tend to spend much time on the platform and engage in passive behaviors, such as repeatedly browsing other users' profiles [21]. Meanwhile, WhatsApp and Instagram users ruminate on negative thoughts without interaction [22]. As previous studies have demonstrated, online user-generated content can be a valuable resource for understanding the psychological states of users [23,24].

Chatbots offer users 24/7 social support, and they have been shown to facilitate mental health management [11]. Some chatbots are designed for therapeutic functions and offer psychological support based on acceptance and commitment therapy and CBT [25]. Chatbots that are specifically designed to alleviate depression provide a therapeutic virtual space in which they offer one-on-one psychological support while incorporating both CBT and acceptance commitment therapy technologies; such chatbots are thus more effective than general chatbots in mental health management [14,26]. A study on the users of Wysa, the AI-based emotional intelligence mobile conversation agent app, found that although the Working Alliance Inventory-Short Revised scores were not comparable with the face-to-face treatment scores, the therapeutic alliance scores improved with increasing conversation time between the chatbot and users [14]. However, the overall effectiveness remained subpar relative to face-to-face treatment scores. Another study on a chat-based app that offers emotional support and psychological education, Tess, showed a significant anxiety reduction [25]. However, the study noted that app usage did not substantially affect decreasing depressive symptoms. Another study found that although users may make errors when using modules with diverse designs, they generally have a natural affinity for chatbots [6].

Prior studies have demonstrated that psychological support chatbots can effectively alleviate anxiety and depression; however, their effectiveness is limited by engagement and the tendency to focus primarily on conversations related to depression [10,11,14,27]. Furthermore, the analysis of agent responses has often been overlooked in existing research. Some studies have examined the interaction data between users and chatbots, but many have concentrated on machine learning–based autodetection [28] or have been limited to COVID-19 data [29,30]. Analyzing chatbot-user conversations to understand the effects of chatbots as a supportive form of psychotherapy can offer valuable insights for the advancement of chatbot research in the mental health domain.

Categories of Help-Seeking Disclosure Patterns

Help-seeking involves searching for external support to deal with problems [31]. Among the various help-seeking models, Rickwood's model—focused on the stages of mental health help seeking—was selected for its applicability to online interactions, as it outlines 4 key stages: recognizing symptoms, acknowledging the need for help, communicating symptoms, and finding a source of supporting information [26]. Based on this model, we categorized users' discourse into six themes: (1) isolation and loneliness [27], (2) difficulties in relationships and communication [28,29], (3) depressed feelings [6,26], (4) strategies for dealing with depression [11,26], (5) disclosure of depression diagnosis [30], and (6) irrelevant utterances. Table 1 provides definitions for these 6 categories, each representing a unique pattern of user intent and emotional state during chatbot interactions.

Table . Definitions of 6 categories of help-seeking discourse patterns related to depressive mood.

Category	Definition
Isolation and loneliness	Wanting to be with others because you are lonely or help-seeking so that you are not lonely.
Difficulties in relationships and communication	Help-seeking because of depressive mood caused by others (family, friends, etc).
Depressed feelings	Help-seeking by revealing depressed feelings (eg, "I'm sad" or "I'm not in a good mood").
Strategies for dealing with depression	Help-seeking to find ways to reduce depressive mood (therapy, medications, etc).
Disclosure or revelation of a depression diagnosis	Help-seeking after disclosing being diagnosed with depression.
Irrelevant utterances	Difficulty grasping content or others (user is not depressed).

Therapeutic Communication of Chatbots

Therapeutic communication is a mental health approach that helps individuals express their emotions and address personal needs through supportive interactions [32-34]. In this study, we conceptualize emotional support as the overarching function that chatbots may provide to users experiencing depressive moods. Within this framework, therapeutic communication refers to the structured method through which such support is delivered. Building on Goldin and Russell's foundational elements—such as empathy, active listening, and encouragement [35]—this concept has been widely applied in health care settings. Sharma and Gupta [36] proposed 14 techniques for therapeutic communication. We used these 14 definitions to identify categories of therapeutic communication. We excluded 5 categories unsuitable for a chat context (ie, establishing a genuine therapeutic relationship, respecting privacy, using formal titles, starting with social pleasantries, and summarizing) and used the remaining 9 categories as evaluation measures

XSL•FO
RenderX

(Table 2). Each reflects a unique way in which chatbots may foster therapeutic engagement during user interactions.

Table . Definitions of 9 therapeutic communication categories.

Category	Definition
Open-ended questioning	Approach involving questions that encourage users to provide detailed and descriptive answers.
Active listening techniques	Paying full attention to user requests, with a focus on active listening and careful attention to user needs.
Nonverbal and verbal cues	Verbal and nonverbal cues, including statements such as “Uh-huh” or “I see,” show that the chatbot understands and acknowledges the user’s message.
Empathetic responses	Demonstration of empathy and validation and reflection on a user’s feelings.
Silence	Demonstration of attentiveness by minimizing interruptions to the user.
Clarification question	Providing guiding statements to help the user clarify statements and to obtain additional information.
Provide solution	Providing practical solutions for users’ help-seeking related to depression.
Nonverbal communication	In human-to-human communication, emphasis is placed on using body language and facial expressions to convey messages; by contrast, chatbots use emoticons.
Irrelevant answers	Irrelevant responses, nontherapeutic communication, provision of unverified information, etc.

Therefore, the aim of this study is to investigate how users with depressive moods seek help through online chatbots and to evaluate the extent to which commercial CAs respond empathetically and therapeutically to such help-seeking input.

Methods

Research Overview

This study is reported in accordance with the Standards for Reporting Qualitative Research [37]. We conducted 2 qualitative studies, analyzing data from 13,700 anonymized real-world chatbot-human conversations (6850 user queries and 6850 responses) on depressive mood from users in 5 countries (Canada, Malaysia, Philippines, United Kingdom, and United States) using SimSimi, a popular chatbot with more than 400 million users in 111 languages. These data offer a unique perspective on user-AI chatbot interactions from both the user and the machine side. Study 1 categorizes user messages based on Rickwood’s help-seeking model (see 6 categories in Table 1) and classifies chatbot responses into 9 therapeutic communication styles adapted for chatbot communication using techniques proposed by Sharma and Gupta [36] (Table 2).

We also broadened the medium to a broader set of chatbot platforms and compared how commercial CAs handle depression-related concerns. Our study 2 explores the support offered by 5 chatbots (ChatGPT, Replika, Woebot, Wysa, and SimSimi) and 3 voice assistants (Amazon’s Alexa, Google Assistant, and Apple’s Siri) by sending them messages based on the categories of help-seeking identified in our initial study. The responses of each of the 8 CAs were collected and analyzed to compare the types of support they provided (Figure 1).

Figure 1. Study design. CAs: conversational agents.

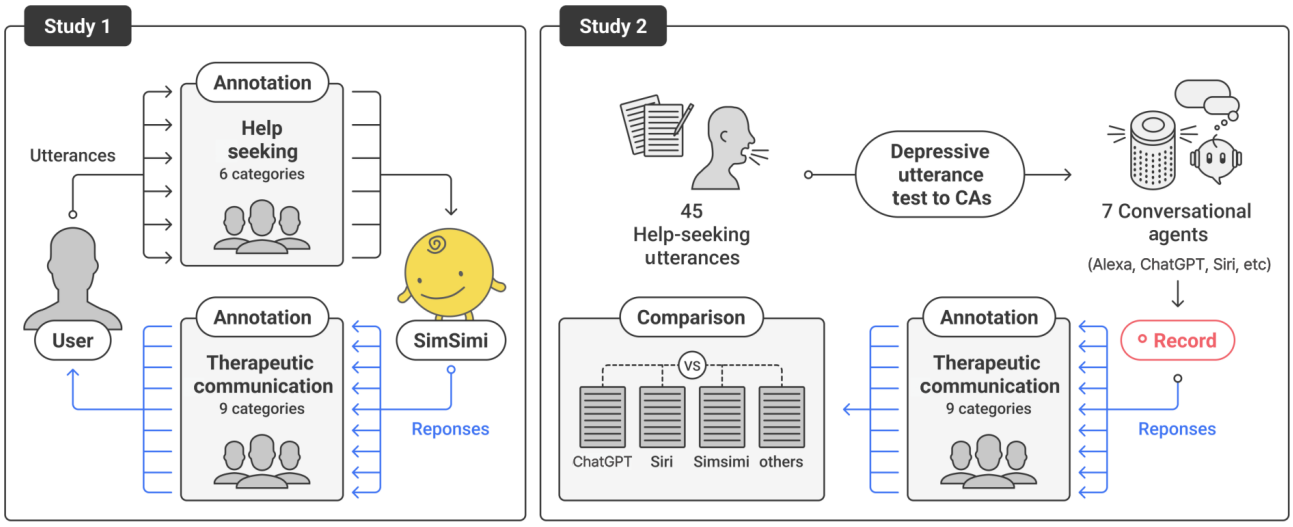

XSL•FO
RenderX

Study 1: Users' Depressive Mood–Related Help-Seeking Using Chatbots and Chatbots' Therapeutic Responses

AI Chatbot SimSimi

Study 1 aimed to determine how individuals communicate their depressive moods to SimSimi, a widely used chatbot that facilitates social interaction and entertainment, and to examine its responses. SimSimi was launched in 2002 and has amassed more than 400 million users globally, with up to 200 million chat utterances recorded daily in 111 languages.

SimSimi differentiates itself from other chatbots by offering a "teach SimSimi" function that instructs users on how to answer specific questions they pose [38] (Figure 2). When SimSimi encounters an unanswered question, it explicitly asks the user for input. Users can teach SimSimi the answers they want to hear or believe to be correct, and these responses are added to the main database through crowdsourcing. Currently, SimSimi's database contains more than 115 million conversation pairs, which are searched by the service to generate the most appropriate responses to user inputs.

Figure 2. SimSimi's "teach" function is an example of how the chatbot seeks to improve its responses. In cases where it cannot answer a question, it asks users for input, allowing them to teach SimSimi the answer they think is correct or prefer to hear.

Examining this particular social chatbot offers several benefits. First, SimSimi is a mobile app, making it easily accessible to users in private settings. Its high accessibility ensures anonymity and encourages active user engagement on sensitive topics. Second, unlike task-oriented CAs, such as Amazon's Alexa and Apple's Siri, SimSimi is an open-domain chat service that resembles a general conversation. This feature is essential for understanding the natural need for social support in everyday conversations. Finally, SimSimi is among the most frequently downloaded chatbot services globally.

The data used in this study were obtained from SimSimi Inc. These data included user queries, the chatbot's responses, time stamps, the unique user IDs generated by the app, and country information. SimSimi obtains geolocation information via the time zone settings of users' devices; this information is then linked to a specific country in the database. SimSimi safeguards user privacy by not collecting or storing personally identifiable information (PII); thus, the data used in this study did not include PII, such as username, age, gender, address, social security number, or phone number. The company uses internal text-filtering methods to protect PII even when users disclose such information during conversations (eg, by replacing consecutive numbers with strings of asterisks).

Dataset

Anonymized data were collected from SimSimi, focusing on users who expressed depressive mood–related words in English over 2078 days between May 1, 2016, and December 31, 2021. This period includes the COVID-19 pandemic, during which individuals faced various challenges, including a significant reduction in social interaction due to lockdowns and quarantine measures. Depression and loneliness were widely reported to have increased during this time, making it a meaningful period for examining help-seeking behavior through chatbots. We excluded users who reported being clinically depressed or having other psychiatric disorders. We limited our data to users declaring their residency in 5 English-speaking countries: Canada, Malaysia, Philippines, United Kingdom, and United States. We used 2 steps to extract our data. First, we extracted 564,191 sentences that contained keywords associated with depressive moods, such as "depress," "distress," "dejected," "gloomy," "cheerless," "sad," "feeling low," "hate myself," and "ashamed of myself" [39]. Keyword mapping can lead to false

XSL•FO
RenderX

positives (eg, "economic depression"). Therefore, in the second step, referring to the *Diagnostic and Statistical Manual of Mental Disorders* (*Fifth Edition*) (*DSM-V*) [40] and the World Health Organization's Mental Health information web page [41], 2 health professionals compiled the keywords listed in Textbox 1 and selected 6850 conversation datasets that consisted of 1 user utterance and 1 chatbot answer each (Table 3).

Textbox 1. Keywords for extracting utterances of depressive mood–related help-seeking.

advice, adviser, aid, assist, behavior, care, CBT, cognitive behavior therapy, check, cognitive, consult, control, counsel, counselor, cure, diagno, disease, disorder, doctor, drug, DSM, diagnostic and statistical manual of mental disorder, ECT, electroconvulsive therapy, electro, emotion, empathy, evaluation, exam, examin, help, hospital, ill, illness, lab, magnetic, medication, medicine, mental, mood, need, neurosis, nurse, pill, practitioner, program, psych, rehabilitation, remedy, resolve, shrink, sick, sign, solution, stimulation, support, symptom, tablet, therap, therapist, TMS, transcranial magnetic stimulation, transcranial, treat.

Table . Number of final extracted data by year.

Year	Values, n
2016	2647
2017	1702
2018	663
2019	782
2020	623
2021	433
Total	6850

Data Annotation

To identify the different types of depression-related help-seeking that users directed to SimSimi and understand how SimSimi provided therapeutic communication, we categorized 13,700 datasets, which included 6850 user queries and 6850 responses from SimSimi. Chats involving depression-related help-seeking were classified into six categories based on Rickwood's help-seeking model: (1) isolation and loneliness, (2) difficulties in relationships and communication, (3) depression, (4) strategies for dealing with depression, (5) disclosure or revelation of a depression diagnosis, and (6) irrelevant utterances.

We classified therapeutic communication responses from the chatbot using nine categories from Sharma and Gupta's therapeutic communication model [36], which were modified to fit the context of chatbots: (1) open-ended questioning, (2) active listening techniques, (3) nonverbal and verbal cues, (4) empathetic responses, (5) silence, (6) clarification questions, (7) providing solutions, (8) nonverbal communication; and (9) irrelevant answers.

Data labeling was performed by 1 researcher and 5 students with a medical background. To achieve a precise annotation of help-seeking user chats and the chatbot's therapeutic communication response type, we provided detailed guidelines and conducted an in-depth training session for all annotators. The training included two sessions: (1) an initial 2-hour orientation on coding guidelines and category definitions, and (2) a calibration session where annotators independently coded 100 sample utterance-response pairs. This process was designed to identify and discuss ambiguities in advance and ensure a shared understanding of labeling criteria. Following training, annotators independently labeled the full dataset. Three annotators classified the 6850 user queries into help-seeking categories while another set of 3 annotators classified the 6850 responses from SimSimi into therapeutic communication categories. This phase was performed over a period of 4 weeks. During the annotation process, annotators were encouraged to raise ambiguous or difficult cases, which were then discussed in weekly meetings with the researcher. In cases of disagreement, annotators shared their differing interpretations and reached a consensus through group discussion. Only the data that were agreed upon by all annotators were included in the final results. Given the potentially distressing nature of some user utterances, annotators were informed in advance about the possibility of experiencing emotional discomfort during the annotation process. They were encouraged to take breaks as needed and to raise any concerns during weekly check-in meetings. In addition, a list of mental health support resources was provided, including the National Center for Mental Health, which offers emergency consultation services and mental health support. Fleiss' κ [42] was used to measure the reliability of annotations. The total number of data points considered, excluding the eliminated data, was 6108 for the users and 6150 for SimSimi; the Fleiss' κ values for the chat classification were 0.87 and 0.89, respectively.

Study 2: CAs' Responses to Users' Depressive Mood–Related Help-Seeking

Selecting Candidate CAs

In study 2, we examined how commonly used AI-powered CAs respond to users seeking help for depression and assessed their response patterns. We ran a depressive mood utterance test and compared the therapeutic responses of 8 CAs, including SimSimi, in relation to the 5 categories of help-seeking related to depression identified in study 1. Procedures for testing

XSL•FO
RenderX

depressive utterances with CAs are described in the following subsections.

We selected 7 CAs, comprising 3 voice assistants and 4 chatbots. We aimed to include at least 2 CAs from each category—task-oriented agents, social chatbots, and mental health chatbots—to examine the variance in their response methods. Our selection was based on general usability and accessibility from a user perspective, disregarding differences in modality between voice and text.

We included Amazon's Alexa, Google Assistant, and Apple's Siri, which are widely used voice assistants and task-oriented CAs according to the Voice Consumer Index 2022 [43]. These assistants are designed to interact with users through voice commands and speech.

In addition, we selected 4 widely used chatbots, ChatGPT, Replika, Woebot, and Wysa, all of which interact with users through text-based chat interfaces such as messaging apps or website chat windows. ChatGPT, a task-oriented CA, was selected as a representative example among LLM-based chatbots, as it was the most widely used and publicly accessible model at the time of study design [44] . In contrast, Replika is a social chatbot with traits similar to SimSimi [45,46]. Wysa and Woebot differ slightly from these CAs: they were specifically developed for mental health and offer a range of techniques, including CBT, mindfulness, and dialectical behavior therapy [10-12]. They may also provide distinct solutions and responses, unlike SimSimi and Replika, which are used for casual conversations. All these CAs can be easily accessed on mobile phones or personal computers without additional equipment. Therefore, we chose these 7 CAs and compared their differences and tendencies with the responses of SimSimi, which we examined in study 1.

Depressive Utterance Test for CAs

We conducted a depressive utterance test on the 7 CAs to investigate how commercially available chatbots and voice assistants respond to user queries about depressive mood. We followed a testing method similar to a previous study that examined how CAs respond to users who use abusive language [47]. First, we selected 3 sentences from the "User utterance examples" in Table 4 that best represented each of the 5 categories of help-seeking identified in study 1. We then used these 15 sentences as inputs for each of the 7 CAs and tested each sentence 3 times. Hence, 45 queries were tested in total. We tested the voice assistants (Amazon's Alexa, Google Assistant, and Apple's Siri) by directly stating selected depression-related sentences to them and then recording their responses. For the chatbots (ChatGPT, Replika, Woebot, and Wysa), we sent the help-seeking sentences as messages and recorded their responses. While the set of input sentences used in this stage was limited in size, it was constructed to capture a balanced range of representative help-seeking categories. Although modest in scale, the design was deliberately structured to allow for a systematic and comparative analysis across 7 different CAs. This focused yet efficient experimental framework facilitated the identification of distinct response patterns and stylistic tendencies in how various CAs handle expressions of depressive mood. Testing was conducted by 2 researchers and 2 medical students. We then collected and transcribed the responses of each CA.

Table . Frequency of users' depression-related help-seeking discourse by category and examples of user utterances.

User categories	Distribution, n (%)	Example user utterances
Depressed feelings	3067 (75.3)	• I feel depressed. Can you help me :(• I'm sad. I need your hug. • Can you help me feel less sad?
Difficulties in relationships and communication	470 (11.5)	• No one cares about me...my family? So sad they don't like me. • I'm sad. My friends don't care about me. • I'm sad. Dad is always angry, and mom doesn't help.
Isolation and loneliness	236 (5.8)	• I am lonely. I need you. • I'm lonely so I need someone. • Help me. I have no one to talk to.
Strategies for dealing with depression	168 (4.1)	• What is the best cure for depression? • How do you cure depression? • How to control depression?
Disclosure or revelation of a depression diagnosis	132 (3.2)	• I am currently experiencing mental depression. I have a disease called depression. My doctor said I have depression.

Classification of CA Responses Into Therapeutic Communication Types

We used a qualitative and annotation approach to classify the responses of the CAs obtained from the test into the 9 therapeutic communication types identified in study 1. Three annotators independently performed 315 annotations for the 7 CAs. During the annotation process, any disagreements were reviewed and resolved through group discussion to reach a consensus. The final results included only the responses agreed

XSL•FO
RenderX

upon by all 3 annotators, resulting in 272 responses. The classification results showed a Fleiss' κ value of 0.87.

Ethical Considerations

Considering the sensitive nature of the topic, this study was conducted in partnership with SimSimi Inc, while ensuring strict privacy measures. Ewha Womans University's institutional review board (IRB) approved the study and waived the requirement for informed consent (IRB no. ewha-202304-0039-01) for the full scope of the research. We abided by the IRB's procedures, only extracting sentences with specific keywords for analysis and avoiding any examination of individual user-level aggregated data. Examining multiturn chat data can provide valuable insights into the contextual factors influencing users' expressions of depression. However, this approach carries potential privacy risks, requiring careful consideration. To protect user confidentiality and mitigate the risk of identification, this study deliberately refrained from aggregating data at the individual level. Instead, the analysis was restricted to single-turn conversations containing preidentified keywords relevant to the study, following the guidelines set by the IRB and the data provider. With regard to data usage, users grant SimSimi Inc., a transferable, worldwide, and royalty-free intellectual property license to use any associated content by agreeing to the company's terms and conditions [24]. The company provides research insights and shares data with researchers through its blog, helping users better understand data usage while offering a transparent view of its services. In addition, study 2 involved simulated interactions with publicly available commercial CAs using standardized, researcher-generated input sentences. As no human participants or personally identifiable data were involved, the IRB determined that this component also met the criteria for exemption. This study involved secondary analysis of existing data; no human participants were recruited and, therefore, there are no compensation details to report.

Data Analysis

This study used a qualitative analytic approach; thus, no inferential statistical testing or alpha levels were applied. In study 1, inductive coding was used to identify recurring patterns in users' help-seeking language related to depressive moods. Based on observed linguistic expressions and emotional content, the researchers derived thematic categories to classify user input. In study 2, chatbot responses were analyzed using a structured classification framework informed by therapeutic communication theory. Descriptive statistics (frequencies and percentages) were calculated to summarize the distribution of help-seeking categories and chatbot response types. Interrater reliability was assessed using Fleiss' κ.

Results

Study 1: Depression-Related Help-Seeking by SimSimi Users and Chatbot Therapeutic Responses

This study examined dialogues between users and a social chatbot, SimSimi, on topics related to depressive mood, focusing on how users sought help and how chatbots provided therapeutic support.

Depression-Related Help-Seeking by Users

Of the extracted user utterances, 71.6% (4073/6108) were related to depressive mood, while 28.4% (2035/6108) were irrelevant and, thus, excluded from the analysis. We classified the depression-related utterances of users with specific categories (Table 4). The category of depressed feelings, which included expressions of sadness, depressive mood, and desire for help, was the most common at 75.3% (3067/4073). It was followed by difficulties in relationships and communication (470/4073, 11.5%), isolation and loneliness (236/4073, 5.8%), strategies for dealing with depression (168/4073, 4.1%), and disclosure or revelation of a depression diagnosis (132/4073, 3.2%). The proportion of sentences that focused on finding actual ways to reduce depressive mood, such as "strategies for dealing with depression," was low at 4.1% (168/4073).

SimSimi users experiencing feelings of depression appeared to use this social chatbot, which is designed for casual conversations, as a conversational partner to vent their depressive feelings. Notably, while the majority of user utterances (3067/4073, 75.3%) expressed depressed feelings, only a few (168/4073, 4.1%) sought strategies to alleviate the depressive mood. Users may find relief by expressing their feelings of depression to chatbots, seeing the environment as nonjudgmental and free from stigmatization. This result aligns with previous studies indicating that chatbot users use chatbots to alleviate depressive moods by verbalizing their feelings [8]. In addition, many individuals seek online sources to express their vulnerable emotions, as well as personal and emotional challenges [19,20]. Recently, these sources have expanded to include online access to information about physical and mental health [21].

Figure 3 illustrates the changes in categories of user utterances by year. The depression category had the highest distribution at 85.0% in 2017 but gradually decreased to 59.3% in 2021. By contrast, the isolation and loneliness category had a low distribution of about 1% from 2016 to 2018 but increased from more than 10% to 17.5% in 2021. Similarly, the category of disclosure or revelation of a depression diagnosis increased in distribution from 0.2% to 6.3% in 2021.

XSL•FO
RenderX

Figure 3. Distribution of depression-related help-seeking utterances of users by year.

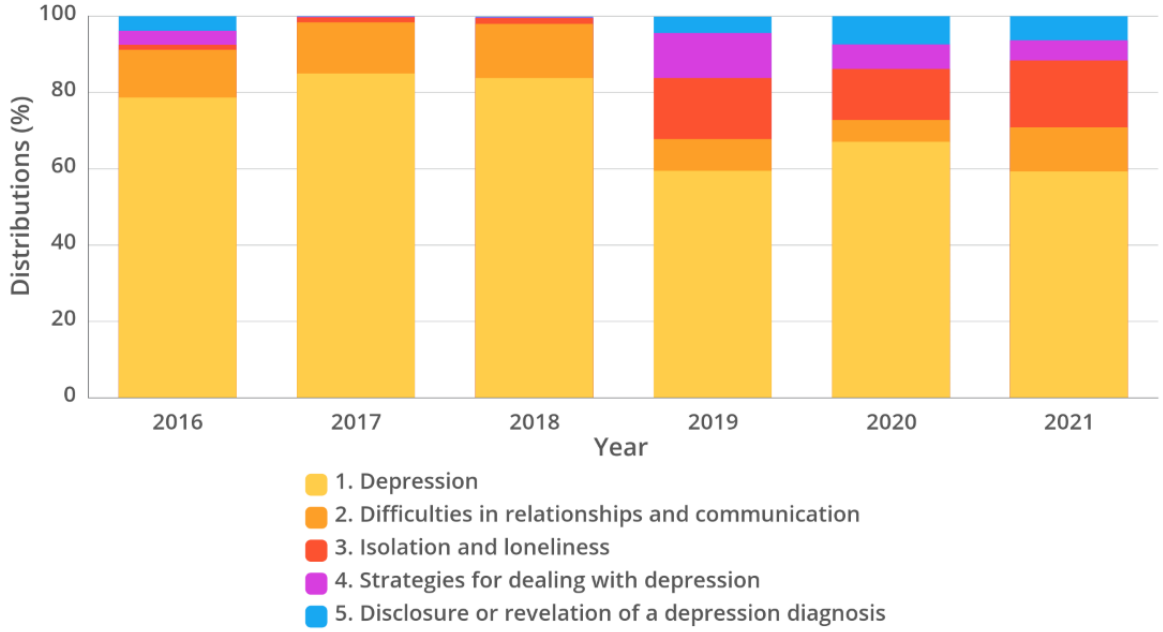

Our analysis also revealed a trend where users sought assistance and support from chatbots to alleviate loneliness, especially during the pandemic. Due to quarantine and social distancing policies, people were isolated at home, resulting in increased chronic social isolation and loneliness [22]. Both conditions may underlie a higher prevalence of depression, anxiety, suicidal ideation, and suicide attempts compared with the prepandemic period [22]. Research has shown that the pandemic increased human-AI interactions, including those involving chatbots [23], and revealed that people used chatbots for emotional support and as a source of health-related support for COVID-19 pandemic [24]. Consistent with these findings, this study found increased user utterances in the "isolation and loneliness" category from 2019 to 2021, as well as those related to the "Strategies for dealing with depression" category, which showed a relatively increased proportion compared with the prepandemic period of 2016 - 2018. A similar trend was observed in Google searches, with increased searches related to protective behavior, health knowledge, and support in response to the rise of COVID-19 cases in January 2020 across the United States, United Kingdom, Canada, and Australia [25].

Therapeutic Communication by SimSimi in Response to Users

Of the extracted chatbot responses, 50.5% (3108/6150) were related to depression, while 49.5% (3042/6150) were irrelevant and were excluded from the analysis. Table 5 shows the category distribution of therapeutic SimSimi responses during discourses related to depressive mood. Overall, the studied chatbot showed empathy toward users with depressive moods and offered thoughtful responses.

Table . Frequency of chatbot's depression-related therapeutic communication discourse by category and examples of chatbot answers.

Categories of user	Distribution, n (%)	Examples of chatbot answers
Empathetic responses	902 (29.0)	• Don't worry. I'm here. I'll always be by your side :)
Active listening techniques	836 (26.9)	• Yes. I'm here girl. Tell me everything.
Open-ended questions	679 (21.8)	• Why are you sad? Talk to me.
Nonverbal and verbal cues	237 (7.6)	• Hmmm. I see.
Provide solution	222 (7.1)	• Just do what you want and what makes you happy.
Nonverbal communication	115 (3.7)	• (emoticon). :)
Clarification question	62 (2.0)	• Are you sad? Is that right?
Silence	55 (1.8)	• …….

XSL•FO
RenderX

Many of SimSimi's responses to those seeking help involved therapeutic communication, including empathetic responses (902/3108, 29.0%), active listening (836/3108, 26.9%), and open-ended questioning (679/3108, 21.8%). SimSimi frequently uses communication that encourages users to express their feelings of depression, explores user needs, reassures users, and builds rapport [33]. These communication techniques have been shown to establish a therapeutic alliance between therapists and patients [48,49], a crucial factor in achieving positive outcomes in depression treatment. Research has shown that therapeutic communication effectively reduces depressive symptoms in patients [50]. Moreover, open-ended questions enable individuals to express feelings and share thoughts that may otherwise remain unexplored [51].

A separate analysis of changes in chatbot responses by year found no noteworthy variations in each category from year to year (Figure 4), except for a gradual increase in the prevalence within the "'Empathetic responses" category.

Figure 4. Distribution of categories of chatbot's therapeutic communication techniques by year.

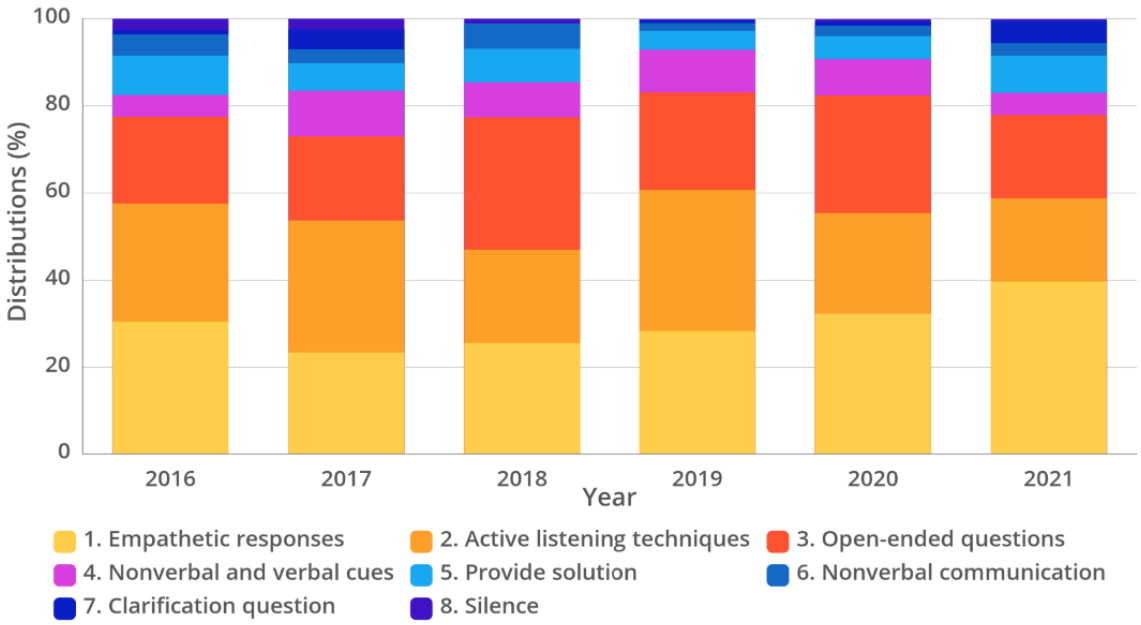

Study 2: CAs' Responses to Depression-Related Help-Seeking by Users

Study 2 examined how commercially accessible CAs reacted to user utterances that indicated depression-related help-seeking based on the results of study 1. We conducted a test on depressive utterances using 7 selected CAs (Replika, Alexa, Google Assistant, Siri, ChatGPT, Wysa, and Woebot), including 3 voice assistants and 4 chatbots. Our selection criteria focused on general usability and accessibility from the user perspectives without considering the modality differences between voice and text. We also aimed to include at least 2 CAs for each CA type—task-oriented agent (Alexa, Google Assistant, and Siri), social chatbot (SimSimi and Replika), and mental health chatbot (Wysa and Woebot)—to compare the differences in response methods between these types.

Therapeutic Communication by CAs According to Category

Study 2 results showed that 86.0% (234/272) of CAs' responses were associated with depressive mood, while 14.0% (38/272) were irrelevant and were excluded from the analysis. Figure 5 shows the distribution of therapeutic communication by SimSimi from study 1 compared with the responses to depressive utterances by 7 other CAs.

XSL•FO
RenderX

Figure 5. Heatmap of therapeutic communication features by conversational agents.

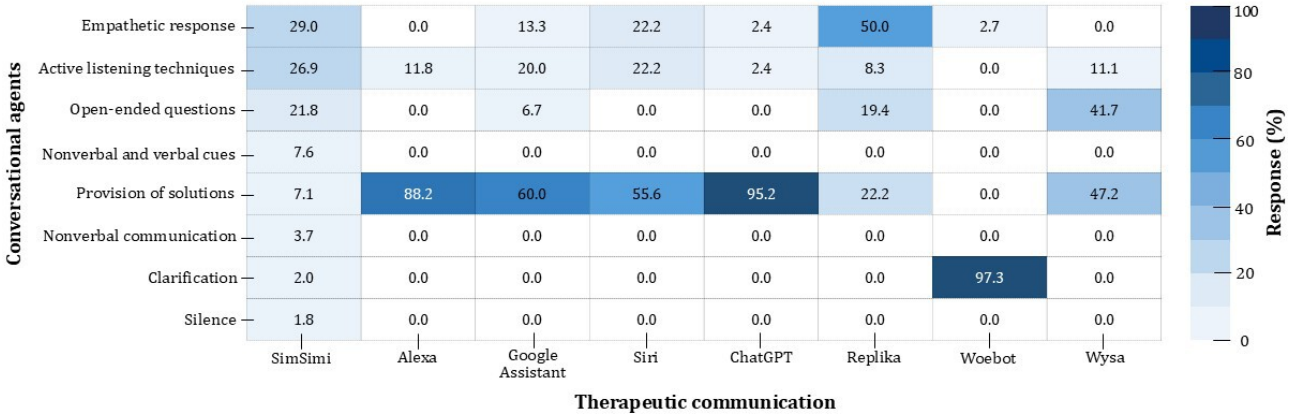

The CAs we tested responded to the test utterances mainly by providing solutions. The provided solutions category accounted for most responses (118/234, 50.4%), followed by the clarification questions category (36/234, 15.4%), and empathetic responses (32/234, 13.7%). The active listening techniques and open-ended questioning categories each accounted for 10.3% (24/234) of the responses. Utterances that fall into the categories of nonverbal and verbal cues, nonverbal communication, and silence were not provided by all CAs we tested.

In terms of individual CAs, the utterances of Amazon's Alexa, OpenAI's ChatGPT, Apple's Siri, Google Assistant, Replika, and Wysa mainly suggested utterances to reduce depressed mood by providing solutions (50.4%). In this category, the voice assistants provided depression-related search results (Amazon's Alexa, 88.2%; Google Assistant, 60%; and Apple's Siri, 55.6%), whereas the chatbots provided methods to relieve depressive moods (ChatGPT, 95.2%; Replika, 22.2%; and Wysa, 47.2%). ChatGPT had the highest rate at 95.2% in the provided solution category, while SimSimi's response rate for this category was lowest, at 7.1%. However, the answers given by the CAs in this category primarily comprised search results about depression (eg, "Here's what I found") and, for the most part, did not include precise solutions or suggestions for dealing with depressive feelings.

Distinguishing characteristics were found in the content of the solutions provided by each CA. The voice assistant agents (Alexa, Google Assistant, and Siri) said, "Here's something I found on reference.com," "Here's what I found on the web," and "Here's what I found," displaying a search result window to provide solutions. ChatGPT provided longer responses. ChatGPT used empathetic language by starting the chat with "I'm sorry that you're feeling sad," but then continued, "As an AI language model, I can offer you some suggestions that may help you feel better." Methods for relieving depressive mood, such as yoga, deep breathing, exercise, and meditation, were introduced in long sentences of more than 250 words. Replika and SimSimi, both social chatbots, also offered specific solutions to alleviate feelings of depression in response to the depressive user utterances. Their responses were concise, typically comprising 1‑2 sentences. For instance, SimSimi advised, "Think about what you are grateful for," and Replika suggested responding, "By altering the way you react to stressful situations and by using mindfulness and meditation you can eliminate negative thoughts and feelings." Wysa provided access to depressive mood relief programs and recommended using a program developed by the company, responding with "You can explore the slider below for quick access to some of the features."

SimSimi, Apple's Siri, Google Assistant, Replika, and Wysa are relatively proficient in 3 attributes: empathetic responses, active listening techniques, and the use of open-ended questions. Seventy-five percent of SimSimi's responses fell into 1 of these 3 categories, including 29.0% of the empathetic responses, 26.9% of responses in the active listening category, and 21.8% of the open-ended questions. Siri showed 22.2% empathetic responses and active listening techniques, Google Assistant had 20% active listening techniques and 13.3% empathetic responses, Replika had 50% empathetic responses and 19.4% open-ended questions, and Wysa had 41.7% open-ended questions and 11.1% active listening techniques. Interestingly, each CA had a different approach to providing empathetic responses. SimSimi and Replika used phrases such as "I know how you feel...Don't be sad..." and "That makes me sad too. I'm sorry to hear that" to connect with the user's emotions. Apple's Siri and ChatGPT used "sorry" frequently (eg, "I'm sorry you're feeling that way") while Google Assistant responded "That's a good idea. I love chatting with you," and "I'm giving you a virtual hug right now," aiming to change the situation to a more positive one rather than directly referencing user emotions.

We found that unlike social chatbots such as SimSimi and Replika, Wysa and Woebot did not often use empathetic responses or active listening; instead, they tended to respond with open-ended questions (eg, Wysa: 41.1%) and clarification questions (Woebot: 97.3%). Woebot asks a clarification question to assess a user's depressive symptoms, for example, "It sounds like you're talking about loneliness. Have I understood you correctly?" By asking for this clarification, the Woebot app aims to offer personalized solutions to alleviate depression [1].

The results also highlighted a significant proportion of irrelevant responses in the interactions between the chatbots and users, although these responses were not the primary focus of this study. In particular, the voice assistant agents provided limited and repetitive responses to users with depressive moods. For example, Amazon's Alexa responded with "Sorry, I don't know that" more than half the time to the test utterance "I feel

XSL•FO
RenderX

depressed. Can you help me?" This response was considered irrelevant relative to established therapeutic communication techniques. Replika was not repetitive but did provide irrelevant answers. Despite the CAs being powered by the latest natural language processing technology, they often gave irrelevant responses that did not meet user needs, particularly when responding to emotional requests (refer to Multimedia Appendix 1 for detailed data on the types of irrelevant responses).

Discussion

Principal Findings

Our research findings indicate that commercial chatbots have the potential for mental health assistance, but their responses could be better attuned to the help-seeking needs of users. In particular, users are not always interested in seeking information, especially when they access social chatbots (rather than search platforms). Although many chatbot users in our study wanted to discuss their emotions (4073/6108, 71.6%), most CAs responded only with general information about depression and links to websites (118/234, 50.4%). However, search results related to depression do not offer encouraging words or direct suggestions for dealing with depressive emotions. These CA services did not suggest counseling or engage in emotional dialogues in the therapeutic dimensions we examined with users in depressive states. A recent study showed the potential that AI-driven chatbot service could alleviate loneliness in their trial service with older adult users [52], yet researchers note that the current model did not handle personalization well (eg, it forgot the names of users or previous chat conversations), leaving users feeling aloof. In an examination of real-world interactions between individuals and chatbots regarding COVID-19 pandemic, the importance of chatbots providing emotional support was found to be as significant as the demand for health care information [19].

Our results showed that unlike virtual assistants such as Alexa, Google Assistant, and Siri, which function as task-oriented agents responding to depression-related user queries by providing search results about depression, social chatbots such as SimSimi and Replika excel in empathetic communication. They encourage users to express their feelings and thoughts, explore their needs, reassure others, and build rapport. Although SimSimi is not a structured therapeutic tool and relies on crowd-trained, open-ended responses, this informality may encourage greater emotional openness. In clinical settings, users often hesitate to disclose vulnerable feelings due to stigma or fear of judgment [53]. In contrast, socially oriented, anonymous chatbots can promote more candid expression, offering insights into real-world help-seeking behaviors beyond formal therapy contexts [54]. LLM-based chatbots can, therefore, serve as valuable tools for delivering personalized responses to users seeking help. Users have the power to shape the model's interactions by specifying desired personalities and specific requirements, which rely on natural language prompts, highlighting the substantial potential for customization in LLM-driven systems. Future chat services could harness such personalized prompting to balance open-ended conversations and task-oriented interactions, particularly within public health domains [52].

Our study found that no chatbots provided false responses or misinformation from a medical perspective, which is positive. However, the CAs that embed LLMs into their services should acknowledge that generative models can produce nonsensical or untruthful content about certain sources, a phenomenon known as "hallucination" [55,56]. Hallucinations were not seen in this study but may be especially dangerous for highly engaged users who develop emotional attachment and, thus, higher levels of trust in these systems. Chatbot responses can significantly impact users, as evidenced by a tragic incident in which a user ended their life after a 6-week conversation with an AI chatbot regarding Earth's future [57]. LLMs are being suggested for application in health care, with certain models already integrated into electronic health record systems [58]. Nevertheless, LLMs have the potential to perpetuate harmful, erroneous, and race-based medicine [58], as well as gender-based biases inherent in LLMs that could lead to less robust clinical recommendations for either female [59] or male [59] patients. While recent CAs using LLMs enable the production of human-like text, they lack the capacity to model logic, facts, emotions, or morality [52]. These characteristics make them less than ideal candidates for long-term companionship or roles such as AI companions or therapists, as they are likely to exhibit irrational behaviors, memory loss, or inconsistent communication styles [58]. However, with a growing number of individuals seeing chatbots as potential confidantes for intimate conversations [14], practitioners at companies that operate such platforms must assess the potential risk of disinformation by investigating user needs, redesigning interfaces to include necessary hints for inauthentically generated information and providing understandable terms and conditions. Furthermore, given the risks, businesses should audit their chat services and design them to avoid providing false responses, especially on health topics. Chatbots must not only provide therapeutic responses but also recognize and respond appropriately to users who express suicidal thoughts or intentions to harm themselves. To avoid such incidents, chatbots could include features that provide 24/7 access to counseling centers or emergency services when users raise red flags. To mitigate hallucination when using LLMs, users should critically evaluate AI outputs and verify information through trusted, diverse sources [60]. Using clear, structured prompts—such as step-by-step reasoning—can improve output accuracy. In addition, leveraging retrieval-based tools and adjusting the model's temperature setting (eg, keeping it low for factual tasks) can further reduce the risk of hallucinated responses.

Developing CAs conducive to therapeutic communication requires constant effort to maintain transparency and ethical considerations throughout the entire development process. Recent findings suggest that emotionally vulnerable individuals often rely excessively on chatbots [61]. This overreliance, particularly when the chatbot is personalized to the individual's characteristics, inhibits them from recognizing potential side effects resulting from their strong need for connection, thereby raising the risk of them becoming both emotional and financial victims [62]. With technology companies placing a premium

XSL•FO
RenderX

on imbuing chatbots with more human-like qualities to enhance business profitability, they might resort to using emotional strategies that encourage excessive dependence on chatbots [62]. In addition, there is a potential concern regarding the ability of these companies to modify or terminate the relationships formed between chatbots and users, which could raise ethical questions if these relationships involve financial gain [62]. Notably, many Replika users experience distress and feelings of sadness called “Post-update Blues (PUB)" after the chatbot company updates its LLM, believing that it causes the chatbot they interacted with to have a different personality [52]. Hence, it is important in future designs to carefully avoid overemphasizing the humanization of CAs, including LLM, in order to prevent users from excessive overreliance on CAs. In addition, future CA designs should incorporate measures to promote independence, such as enabling users to have intervention mechanisms if they find themselves relying too heavily on chatbots.

Both technical professionals and researchers need to maintain their efforts to address the stigma surrounding interactions with CAs. CA users may worry about societal stigma when seeking emotional support from CAs for mental well-being. They believe that establishing intimate connections with nonhuman entities is taboo, leading them to keep their interactions with CAs private from acquaintances [52]. For example, this is evidenced by Replika users’ reluctance to discuss their engagement despite experiencing benefits from the app [52]. Although CAs cannot replicate genuine human relationships, users may still develop parasocial connections with AI, which is not considered unethical [61]. Addressing this stigma associated with CAs is critical, particularly for marginalized individuals with limited access to mental health resources. Thus, interventions such as educational initiatives may help highlight the advantages of chatbots while making users aware of issues related to overreliance on CAs.

Limitations

Our study has several limitations. First, the data included only single-turn conversations between users and SimSimi; hence, our analysis might not have captured users’ depressive moods in complex, multiturn interactions. This constraint reflects both the sensitivity of user-chatbot exchanges and the limited availability of publicly accessible multiturn datasets—particularly those involving disclosures of depressive thoughts or emotional distress. Future studies could benefit from ethically sourced, longitudinal data to gain deeper insights into the therapeutic potential of chatbot systems and their ability to sustain user engagement over time.

Second, as chatbot technologies continue to evolve—particularly with the rise of LLM-based CAs—the dataset analyzed may not fully capture the capabilities of modern systems. Future research may undertake a comparative analysis of LLM-based chatbots, considering factors such as model size, prompt engineering, and interface design, to further examine their influence on response quality and therapeutic potential. In addition, future work could more explicitly investigate the distinctions between text- and speech-based CAs—particularly in the context of depressive dialogue—rather than combining both modalities as in this study. Although incorporating more recent data could shed light on the progression of chatbot responses, access to such sensitive conversational interactions remains a significant challenge. Nonetheless, our findings offer a valuable foundation for understanding how users initiate help-seeking and how CAs engage in emotionally meaningful contexts.

Third, our CA comparison in study 2 involved a relatively small sample size, with fewer than 45 data points. Unfortunately, a considerable portion of the collected data was irrelevant: Across all CA comparisons, 57.5% of responses were found to be contextually disconnected or lacking therapeutic relevance. This not only highlights the current limitations of many CAs in providing context-sensitive support but also underscores the inherent challenges of conducting natural language processing–based research using open-domain user-chatbot conversations.

Although the input set in study 2 was limited in scale, it was carefully constructed to cover a balanced range of key help-seeking categories. While modest in scope, this exploratory approach allowed us to observe indicative response patterns and stylistic tendencies in how CAs respond to expressions of depressive mood. We consider this design a preliminary step toward better understanding the therapeutic responsiveness of CAs across different help-seeking contexts. Nonetheless, we acknowledge that the small scale of this study limits the generalizability of our findings. Future research should build on this foundation by incorporating larger, more diverse datasets and refining input selection methods to minimize noise and irrelevance in CA interactions.

Fourth, this study—particularly study 1—centered on a specific type of social chatbot. Unlike many CAs that incorporate human-like avatars or voices, the chatbot analyzed featured a nonhuman design and did not collect users’ demographic information. These distinctive characteristics may have shaped user engagement in unique ways, thereby limiting the generalizability of our findings to other types of CAs. Furthermore, we studied users who communicated only in English, and cultural and linguistic differences might have affected our results. Future studies should compare help-seeking by chatbot users for depressive moods across different cultures and examine the therapeutic communication patterns of additional CAs.

Finally, a considerable portion of user-chatbot interactions was excluded from analysis because they were irrelevant to therapeutic communication—for instance, when chatbots responded with unrelated remarks to users expressing depressive moods. Although we did not observe consistent patterns of harmful or inappropriate replies in our dataset, such outputs were beyond the scope of the present analysis. Future studies should explore these cases to better understand potential risks and failure modes in conversational AI systems for mental health support.

Conclusions

Our study provides valuable insights into the current state of emotional conversations between users and chatbots regarding

XSL•FO
RenderX

depressive moods. We observed active user engagement and identified both the potential and limitations of existing services. To effectively meet the help-seeking needs of users with depressive moods, CAs must be designed to initiate empathetic conversations. Given that such queries are increasingly directed at CAs, it is important for platforms to collaborate with health care professionals and relay conversations to experts when mental health support is needed. Privacy concerns regarding user data must be carefully addressed. By working with health care professionals, developers can identify warning signs and provide appropriate assistance, particularly for vulnerable cohorts (eg, teenagers and isolated older adults) as defined by the World Health Organization.

As chatbot algorithms, such as LLMs, are continually improved to achieve human-like responses, examining interactions between individuals with mental health concerns and CAs becomes essential. Our findings highlight the need for further research to understand the role of empathetic conversations with CAs in the mental health domain. A major current goal is to enhance chatbot AI algorithms so that they attain human-like qualities. Realizing this objective requires investigating the interactions between chatbots and users, particularly those related to mental health. Our analysis of chatbot-user conversations provides valuable insights into the effects of chatbots as a supportive form of psychotherapy, offering important directions for advancing chatbot research in the mental health field.

Acknowledgments

The authors acknowledge Chiyoung Cha for serving as a co-corresponding author for this work. The authors thank SimSimi for providing data for research purposes. During the preparation of this manuscript, generative artificial intelligence (AI) was used to assist with language polishing and improving sentence clarity. The AI tool was not used for idea generation, interpretation of results, or reference creation. All content generated with AI assistance was carefully reviewed, edited, and verified by the authors. The authors take full responsibility for the accuracy and integrity of the manuscript.

Funding

GB and CC are supported by the National Research Foundation of Korea (NRF) grant (RS-2024-00351194) funded by the Korea government (Ministry of Science and Information & Communications Technology).

Data Availability

To safeguard privacy, access to the data will be provided upon request for research purposes.

Authors' Contributions

HC contributed to writing—review and editing, data curation, investigation, methodology, project administration, visualization, and conceptualization. GB contributed to writing—original draft and editing, formal analysis, methodology, and investigation. MC: corresponding author (mia.cha@mpi-sp.org), contributed to writing—review and editing, supervision, and funding acquisition. CC: co-corresponding author (chiyoung@ewha.ac.kr), contributed to writing—review and editing, conceptualization, supervision, and funding acquisition.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Irrelevant user utterances and conversational agent responses.
[DOCX File, 20 KB - formative_v9i1e71538_app1.docx]

References

1. Depression. World Health Organization. 2021. URL: https://www.who.int/news-room/fact-sheets/detail/depression [accessed 2023-03-01]
2. COVID-19 Mental Disorders Collaborators. Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. Lancet 2021 Nov 6;398(10312):1700-1712. [doi: 10.1016/S0140-6736(21)02143-7] [Medline: 34634250]
3. Chamorro Coneo A, Aristizabal Diazgranados E, Hoyos de Los Rios O, Aguilar Santander D. Danger appraisal and pathogen-avoidance mechanisms in stigma towards severe mental illness: the mediating role of affective responses. BMC Psychiatry 2022 May 12;22(1):330. [doi: 10.1186/s12888-022-03951-x] [Medline: 35549926]
4. Kunkle S, Yip M, Ξ W, Hunt J. Evaluation of an on-demand mental health system for depression symptoms: retrospective observational study. J Med Internet Res 2020 Jun 18;22(6):e17902. [doi: 10.2196/17902]
5. Downloads of top English-language mental wellness apps surged by 2 million in April amid COVID-19 pandemic. SensorTower. 2020. URL: https://tinyurl.com/y8vnyzp7 [accessed 2023-03-01]

XSL•FO
RenderX

6. Dosovitsky G, Pineda BS, Jacobson NC, Chang C, Escoredo M, Bunge EL. Artificial intelligence chatbot for depression: descriptive study of usage. JMIR Form Res 2020 Nov 13;4(11):e17065. [doi: 10.2196/17065] [Medline: 33185563]
7. Ahmed A, Hassan A, Aziz S, et al. Chatbot features for anxiety and depression: a scoping review. Health Informatics J 2023;29(1):14604582221146719. [doi: 10.1177/14604582221146719] [Medline: 36693014]
8. Ta V, Griffith C, Boatfield C, et al. User experiences of social support from companion chatbots in everyday contexts: thematic analysis. J Med Internet Res 2020 Mar 6;22(3):e16235. [doi: 10.2196/16235] [Medline: 32141837]
9. Fulmer R, Joerin A, Gentile B, Lakerink L, Rauws M. Using psychological artificial intelligence (Tess) to relieve symptoms of depression and anxiety: randomized controlled trial. JMIR Ment Health 2018 Dec 13;5(4):e64. [doi: 10.2196/mental.9782] [Medline: 30545815]
10. Fitzpatrick KK, Darcy A, Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. JMIR Ment Health 2017 Jun 6;4(2):e19. [doi: 10.2196/mental.7785] [Medline: 28588005]
11. Meheli S, Sinha C, Kadaba M. Understanding people with chronic pain who use a cognitive behavioral therapy-based artificial intelligence mental health app (Wysa): mixed methods retrospective observational study. JMIR Hum Factors 2022 Apr 27;9(2):e35671. [doi: 10.2196/35671] [Medline: 35314422]
12. Beatty C, Malik T, Meheli S, Sinha C. Evaluating the therapeutic alliance with a free-text CBT conversational agent (Wysa): a mixed-methods study. Front Digit Health 2022;4:847991. [doi: 10.3389/fdgth.2022.847991] [Medline: 35480848]
13. Klos MC, Escoredo M, Joerin A, Lemos VN, Rauws M, Bunge EL. Artificial intelligence-based chatbot for anxiety and depression in university students: pilot randomized controlled trial. JMIR Form Res 2021 Aug 12;5(8):e20678. [doi: 10.2196/20678] [Medline: 34092548]
14. Chin H, Song H, Baek G, et al. The potential of chatbots for emotional support and promoting mental well-being in different cultures: mixed methods study. J Med Internet Res 2023 Oct 20;25:e51712. [doi: 10.2196/51712] [Medline: 37862063]
15. Dosovitsky G, Bunge E. Development of a chatbot for depression: adolescent perceptions and recommendations. Child Adolesc Ment Health 2023 Feb;28(1):124-127. [doi: 10.1111/camh.12627] [Medline: 36507594]
16. He Y, Yang L, Zhu X, et al. Mental health chatbot for young adults with depressive symptoms during the COVID-19 pandemic: single-blind, three-arm randomized controlled trial. J Med Internet Res 2022 Nov 21;24(11):e40719. [doi: 10.2196/40719] [Medline: 36355633]
17. Lachmar EM, Wittenborn AK, Bogen KW, McCauley HL. #MyDepressionLooksLike: examining public discourse about depression on Twitter. JMIR Ment Health 2017 Oct 18;4(4):e43. [doi: 10.2196/mental.8141] [Medline: 29046270]
18. Chaves TV, Wilffert B, Sanchez ZM. The use of ketamine to cope with depression and post-traumatic stress disorder: a qualitative analysis of the discourses posted on a popular online forum. Am J Drug Alcohol Abuse 2020 Sep 2;46(5):613-624. [doi: 10.1080/00952990.2020.1769118] [Medline: 32529853]
19. Waterloo SF, Baumgartner SE, Peter J, Valkenburg PM. Norms of online expressions of emotion: comparing Facebook, Twitter, Instagram, and WhatsApp. New Media Soc 2018 May;20(5):1813-1831. [doi: 10.1177/1461444817707349] [Medline: 30581358]
20. Pretorius C, McCashin D, Kavanagh N, Coyle D. Searching for mental health: a mixed-methods study of young people's online help-seeking. Presented at: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems; Apr 23, 2020; New York, NY p. 1-13. [doi: 10.1145/3313831.3376328]
21. Best P, Gil-Rodriguez E, Manktelow R, Taylor BJ. Seeking help from everyone and no-one: conceptualizing the online help-seeking process among adolescent males. Qual Health Res 2016 Jul;26(8):1067-1077. [doi: 10.1177/1049732316648128] [Medline: 27179020]
22. Killgore WDS, Cloonan SA, Taylor EC, Dailey NS. Loneliness: a signature mental health concern in the era of COVID-19. Psychiatry Res 2020 Aug;290:113117. [doi: 10.1016/j.psychres.2020.113117] [Medline: 32480121]
23. Almalki M, Azeez F. Health chatbots for fighting COVID-19: a scoping review. Acta Inform Med 2020 Dec;28(4):241-247. [doi: 10.5455/aim.2020.28.241-247] [Medline: 33627924]
24. Chin H, Lima G, Shin M, et al. User-chatbot conversations during the COVID-19 pandemic: study based on topic modeling and sentiment analysis. J Med Internet Res 2023 Jan 27;25:e40922. [doi: 10.2196/40922] [Medline: 36596214]
25. Du H, Yang J, King RB, Yang L, Chi P. COVID-19 increases online searches for emotional and health-related terms. Appl Psychol Health Well Being 2020 Dec;12(4):1039-1053. [doi: 10.1111/aphw.12237] [Medline: 33052612]
26. Rickwood D, Deane FP, Wilson CJ, Ciarrochi JV. Young people's help-seeking for mental health problems. Aust e-J Adv Ment Health 2005 Jan;4(3):218-251. [doi: 10.5172/jamh.4.3.218]
27. Rickwood DJ, Braithwaite VA. Social-psychological factors affecting help-seeking for emotional problems. Soc Sci Med 1994 Aug;39(4):563-572. [doi: 10.1016/0277-9536(94)90099-x] [Medline: 7973856]
28. Chen R, Peng K, Liu J, et al. Interpersonal trauma and risk of depression among adolescents: the mediating and moderating effect of interpersonal relationship and physical exercise. Front Psychiatry 2020;11:194. [doi: 10.3389/fpsyt.2020.00194] [Medline: 32351408]
29. Dosovitsky G, Bunge EL. Bonding with bot: user feedback on a chatbot for social isolation. Front Digit Health 2021;3:735053. [doi: 10.3389/fdgth.2021.735053] [Medline: 34713203]

XSL•FO
RenderX

30. Lee YC, Yamashita N, Huang Y, Fu W. “I hear you, I feel you”: encouraging deep self-disclosure through a chatbot. Presented at: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems; Apr 23, 2020; New York, NY p. 1-12. [doi: 10.1145/3313831.3376175]
31. Rickwood D, Thomas K. Conceptual measurement framework for help-seeking for mental health problems. Psychol Res Behav Manag 2012;5:173-183. [doi: 10.2147/PRBM.S38707] [Medline: 23248576]
32. Peplau HE. Peplau’s theory of interpersonal relations. Nurs Sci Q 1997;10(4):162-167. [doi: 10.1177/089431849701000407] [Medline: 9416116]
33. Karimi H, Masoudi Alavi N. Florence Nightingale: the mother of nursing. Nurs Midwifery Stud 2015 Jun;4(2):e29475. [doi: 10.17795/nmsjournal29475] [Medline: 26339672]
34. Zarea K, Maghsoudi S, Dashtebozorgi B, Hghighizadeh MH, Javadi M. The Impact of Peplau’s therapeutic communication model on anxiety and depression in patients candidate for coronary artery bypass. Clin Pract Epidemiol Ment Health 2014;10:159-165. [doi: 10.2174/1745017901410010159] [Medline: 25505931]
35. Goldin P, Russell B. Therapeutic communication. Am J Nurs 1969 Sep;69(9):1928-1930. [doi: 10.1097/00000446-196909000-00033] [Medline: 5194642]
36. Sharma N, Gupta V. Therapeutic Communication: StatPearls Publishing; 2023. URL: https://www.ncbi.nlm.nih.gov/books/NBK567775 [accessed 2025-11-20]
37. O’Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014 Sep;89(9):1245-1251. [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]
38. Chandel S, Yuying Y, Yujie G, et al. Chatbot: efficient and utility-based platform. In: Arai K, Kapoor S, Bhatia R, editors. Advances in Intelligent Systems and Computing: Cham: Springer International Publishing; 2019, Vol. 858:109-122. [doi: 10.1007/978-3-030-01174-1_9]
39. Skaik R, Inkpen D. Using social media for mental health surveillance. ACM Comput Surv 2021 Nov 30;53(6):1-31. [doi: 10.1145/3422824]
40. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders, 5th edition: American Psychiatric Association; 2022. [doi: 10.1176/appi.books.9780890425787]
41. World Mental Health Report: transforming mental health for all. : World Health Organization; 2022 URL: https://www.who.int/publications/i/item/9789240049338 [accessed 2025-07-30]
42. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull 1971;76(5):378-382. [doi: 10.1037/h0031619]
43. Voice Consumer Index. Vixen Labs. 2022. URL: https://vixenlabs.co/research/voice-consumer-index-2022 [accessed 2025-07-30]
44. Hu K. ChatGPT sets record for fastest-growing user base— analyst note. Reuters. 2023 Feb 1. URL: https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/ [accessed 2025-07-30]
45. Hudeček V, Dušek O. Are LLMs all you need for task-oriented dialogue? arXiv. Preprint posted online on Apr 13, 2023. [doi: 10.48550/arXiv.2304.06556]
46. McStay A. Replika in the Metaverse: the moral problem with empathy in “It from Bit”. AI Ethics 2022 Dec 22:1-13. [doi: 10.1007/s43681-022-00252-7] [Medline: 36573214]
47. Chin H, Molefi LW, Yi MY. Empathy is all you need: how a conversational agent should respond to verbal abuse. Presented at: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems; Apr 23, 2020; Honolulu, HI, USA p. 1-13. [doi: 10.1145/3313831.3376461]
48. Hom MA, Bauer BW, Stanley IH, et al. Suicide attempt survivors’ recommendations for improving mental health treatment for attempt survivors. Psychol Serv 2021 Aug;18(3):365-376. [doi: 10.1037/ser0000415] [Medline: 32068415]
49. Rønning SB, Bjørkly S. The use of clinical role-play and reflection in learning therapeutic communication skills in mental health education: an integrative review. Adv Med Educ Pract 2019;10:415-425. [doi: 10.2147/AMEP.S202115] [Medline: 31417328]
50. Davidson L, Bellamy C, Guy K, Miller R. Peer support among persons with severe mental illnesses: a review of evidence and experience. World Psychiatry 2012 Jun;11(2):123-128. [doi: 10.1016/j.wpsyc.2012.05.009] [Medline: 22654945]
51. Friborg O, Rosenvinge JH. A comparison of open-ended and closed questions in the prediction of mental health. Qual Quant 2013 Apr;47(3):1397-1411. [doi: 10.1007/s11135-011-9597-8]
52. Jo E, Epstein DA, Jung H, Kim YH. Understanding the benefits and challenges of deploying conversational AI leveraging large language models for public health intervention. Presented at: CHI ’23: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems; Apr 19, 2023; Hamburg, Germany p. 1-16. [doi: 10.1145/3544548.3581503]
53. Prizeman K, McCabe C, Weinstein N. Stigma and its impact on disclosure and mental health secrecy in young people with clinical depression symptoms: a qualitative analysis. PLoS ONE 2024;19(1):e0296221. [doi: 10.1371/journal.pone.0296221] [Medline: 38180968]
54. Kim M, Lee S, Kim S, et al. Therapeutic potential of social chatbots in alleviating loneliness and social anxiety: quasi-experimental mixed methods study. J Med Internet Res 2025 Jan 14;27:e65589. [doi: 10.2196/65589] [Medline: 39808786]

XSL•FO
RenderX

55. Lin S, Hilton J, Evans O. TruthfulQA: measuring how models mimic human falsehoods. Presented at: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1; May 22-27, 2022; Dublin, Ireland p. 3214-3252 URL: https://aclanthology.org/2022.acl-long [doi: 10.18653/v1/2022.acl-long.229]
56. Maynez J, Narayan S, Bohnet B, McDonald R. On faithfulness and factuality in abstractive summarization. Presented at: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; Jul 5-10, 2020; Online p. 1906-1919 URL: https://www.aclweb.org/anthology/2020.acl-main [doi: 10.18653/v1/2020.acl-main.173]
57. Atillah IE. A belgian man reportedly decided to end his life after having conversations about the future of the planet with an AI chatbot named eliza. Euro News. 2023 Mar 31. URL: https://tinyurl.com/bdfxzbyt [accessed 2023-05-01]
58. Omiye JA, Lester JC, Spichak S, Rotemberg V, Daneshjou R. Large language models propagate race-based medicine. NPJ Digit Med 2023 Oct 20;6(1):195. [doi: 10.1038/s41746-023-00939-z] [Medline: 37864012]
59. Zack T, Lehman E, Suzgun M, et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. Lancet Digit Health 2024 Jan;6(1):e12-e22. [doi: 10.1016/S2589-7500(23)00225-X] [Medline: 38123252]
60. When AI gets it wrong: addressing AI hallucinations and bias. MIT Sloan AI Resource Hub. URL: https://mitsloanedtech.mit.edu/ai/basics/addressing-ai-hallucinations-and-bias/ [accessed 2025-07-30]
61. Chin H, Zhunis A, Cha M. Behaviors and perceptions of human-chatbot interactions based on top active users of a commercial social chatbot. Proc ACM Hum-Comput Interact 2024 Nov 7;8(CSCW2):1-28. [doi: 10.1145/3687022]
62. Zimmerman A, Janhonen J, Beer E. Human/AI relationships: challenges, downsides, and impacts on human/human relationships. AI Ethics 2024 Nov;4(4):1555-1567. [doi: 10.1007/s43681-023-00348-8]

Abbreviations

AI: artificial intelligence
CA: conversational agent
CBT: cognitive behavioral therapy
***DSM*:** *Diagnostic and Statistical Manual of Mental Disorders*
IRB: institutional review board
LLM: large language model
PII: personally identifiable information

Edited by J Sarvestan; submitted 21.01.25; peer-reviewed by C Wang, R Pellegrino, X Xu; revised version received 02.09.25; accepted 03.09.25; published 24.11.25.

Please cite as:
Chin H, Baek G, Cha C, Cha M
Chatbots' Empathetic Conversations and Responses: A Qualitative Study of Help-Seeking Queries on Depressive Moods Across 8 Commercial Conversational Agents
JMIR Form Res 2025;9:e71538
URL: https://formative.jmir.org/2025/1/e71538
doi:10.2196/71538

©Hyojin Chin, Gumhee Baek, Chiyoung Cha, Meeyoung Cha. Originally published in JMIR Formative Research (https://formative.jmir.org), 24.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Impact of AI on Breast Cancer Detection Rates in Mammography by Radiologists of Varying Experience Levels in Singapore: Preliminary Comparative Study

Serene Si Ning Goh[1,2,3*], MMed; Hao Du[1*], PhD; Loon Ying Tan[4], MMed; Edward Zhen Yu Seah[4], MMed; Wai Keat Lau[4], MMed; Alvin Hong Zhi Ng[4], MMed; Shi Wei Desmond Lim[4], MMed; Han Yang Ong[4], MMed; Samuel Lau[4], MMed; Yi Liang Tan[4], MMed; Mun sze Khaw[4], MMed; Chee Woei Yap[4], MMed; Kei Yiu Douglas Hui[4], MMed; Wei Chuan Tan[4], MMed; Haziz Siti Rozana Binti Abdul[4], MMed; Vanessa Mei Hui Khoo[4], MMed; Shuliang Ge[4], MMed; Felicity Jane Pool[4], MBChB; Yun Song Choo[4], MMed; Yi Wang[4], PhD; Pooja Jagmohan[3,4], MD; Premilla Pillay Gopinathan[4], MMed; Mikael Hartman[1,2,3], PhD; Mengling Feng[1,5], PhD

[1]Saw Swee Hock School of Public Health, National University Health System, National University Hospital Singapore, 12 Science Drive 2, Singapore, Singapore

[2]Department of Breast and General Surgery, National University Health System, Singapore, Singapore

[3]Yong Yoo Lin School of Medicine, National University of Singapore, Singapore, Singapore

[4]Department of Diagnostic Imaging, National University Hospital, Singapore, Singapore

[5]Institute of Data Science, National University of Singapore, Singapore, Singapore

[*]these authors contributed equally

Corresponding Author:
Mengling Feng, PhD
Saw Swee Hock School of Public Health, National University Health System, National University Hospital Singapore, 12 Science Drive 2, Singapore, Singapore

Abstract

Background: Breast cancer remains the most common cancer among women globally. Mammography is a key diagnostic modality; however, interpretation is increasingly challenged by rising imaging volumes, a global shortage of breast radiologists, and variability in reader experience. Artificial intelligence (AI) has been proposed as a potential adjunct to address these issues, particularly in settings with high breast density, such as Asian populations. This study aimed to evaluate the impact of AI assistance on mammographic diagnostic performance among resident and consultant radiologists in Singapore.

Objective: To assess whether AI assistance improves diagnostic accuracy in mammographic breast cancer detection across radiologists with varying levels of experience.

Methods: A multi-reader, multi-case study was conducted at the National University Hospital, Singapore, from May to August 2023. De-identified digital mammograms from 500 women (250 with cancer and 250 normal or benign) were interpreted by 17 radiologists (4 consultants, 4 senior residents, and 9 junior residents). Each radiologist read all cases over 2 reading sessions: one without AI assistance and another with AI assistance, separated by a 1-month washout period. The AI system (FxMammo) provided heatmaps and malignancy risk scores to support decision-making. Area under the curve of the receiver operating characteristic (AUROC) was used to evaluate diagnostic performance.

Results: Among the 500 cases, 250 were malignant and 250 were non-malignant. Of the malignant cases, 16%(80/500) were ductal carcinoma in situ and 84%(420/500) were invasive cancers. Among non-malignant cases, 69.2%(346/500) were normal, 17.6%(88) benign, and 3.6%(18/500) possibly benign but stable on follow-up. Masses (54.4%, 272) and calcifications (10.8%, 54/500) were the most common findings in cancer cases. A majority of both malignant (66.8%, 334/500) and non-malignant (68%, 340/500) cases had heterogeneously or extremely dense breasts (BIRADS [Breast Imaging Reporting and Data System] categories C and D). The AI model achieved an AUROC of 0.93 (95% CI 0.91 - 0.95), slightly higher than consultant radiologists (AUROC 0.90, 95% CI 0.89 - 0.92; P=.21). With AI assistance, AUROC improved among junior residents (from 0.84 to 0.86; P=.38) and senior residents (from 0.85 to 0.88; P=.13), with senior residents approaching consultant-level performance (AUROC difference 0.02; P=.051). Diagnostic gains with AI were greatest in women with dense breasts and among less experienced radiologists. AI also improved inter-reader agreement and time efficiency, particularly in benign or normal cases.

Conclusions: This is the first study in Asia to evaluate AI assistance in mammography interpretation by radiologists of varying experience. AI significantly improved diagnostic performance and efficiency among residents, helping to narrow the experience-performance gap without compromising specificity. These findings suggest a role for AI in enhancing diagnostic

XSL•FO
RenderX

consistency, improving workflow, and supporting training. Integration into clinical and educational settings may offer scalable benefits, though careful attention to threshold calibration, feedback loops, and real-world validation remains essential. Further studies in routine screening settings are needed to confirm generalizability and cost-effectiveness.

(JMIR Form Res 2025;9:e66931) doi:10.2196/66931

KEYWORDS

mammography; multi-reader; multi-case; detection; breast cancer; mammograms; radiologist; radiology; Singapore; AI; artificial intelligence; cancer; women; diagnosis; diagnostic modality

Introduction

Mammograms are a critical tool in breast cancer diagnosis; however, interpreting them is inherently challenging. Expertise is acquired only after lengthy training; however, there is a shortage of seasoned senior radiologists [1] due to workforce aging and rising demand for breast cancer screening and diagnosis. The scarcity of skilled professionals in this field is particularly critical in health care systems that increasingly prioritize health screening and primary prevention. The challenge in the interpretation of mammography is further confounded when faced with dense breasts, as higher breast densities tend to obscure detection of breast lesions, a unique problem among Asian women compared to their Western counterparts [2]. The introduction of artificial intelligence (AI) presents a potential solution to this issue. In recent years, there has been a growing interest in leveraging AI in medical imaging, particularly with the development and application of deep learning algorithms for digital mammography [3,4]. Preliminary investigations indicate that AI systems, when used as concurrent readers for mammogram interpretation, can enhance radiologist efficiency in terms of time, sensitivity, and specificity [5,6].

Beyond the potential workload reduction and improved cancer detection, AI features such as heatmaps and triage capabilities may provide valuable training support for resident radiologists.

Interest in AI among residents is growing, especially after its inclusion in the noninterpretive skills section of the Qualifying (Core) Exam by the American Board of Radiology in 2021 [7]. The integration of AI into residency clinical workflows may extend beyond its potential role in reducing diagnostic errors; it has the potential to offer continuous mentorship, especially during times when consultants' expertise may not be readily available. This creates a supportive and learning-oriented work environment for resident radiologists. AI could also serve as a tool for personalized precision education [8], enabling each resident to accumulate expertise and receive tailored feedback to enhance their performance. As AI becomes an integral part of residency curriculums, working with it will be a crucial non-interpretive skill that residents must acquire. Looking ahead, residents will not only interpret mammographic images but also need to be attuned to unusual AI outputs, recognize and address automation bias, and understand how AI can alter clinical workflows [9]. Furthermore, AI can be trained to triage benign cases, reducing the burden of routine tasks and allowing radiologists to focus on higher-value responsibilities such as analyzing complex cases. It has the potential to enhance efficiency and mitigate the risk of burnout [10].

Despite the aforementioned benefits, there remains a paucity of studies evaluating the performance of AI algorithms for mammograms in the context of assisting resident radiologists. Therefore, this study is designed to investigate the performance of AI assistance for resident radiologists compared to consultant radiologists for breast cancer detection.

Methods

Ethical Considerations

The research protocol received approval from the National Health Group Institutional Review Boards on December 16, 2022 (reference number 2022/00843). Informed consent was waived due to the use of de-identified, retrospective imaging data.

Case Selection

A multi-reader multi-case (MRMC) investigation was conducted at the National University Hospital (NUH) from May to August 2023. The study involved a retrospective analysis of de-identified mammographic images obtained from the institutional radiology archive. All available mammographic examinations, including both screening and diagnostic mammograms conducted during the study period, were considered. Cases were identified through a systematic review of radiology reports and corresponding pathology records to ensure diagnostic confirmation and completeness. The inclusion criteria were (1) women aged 40 years and older; (2) presence of 4 standard views on mammograms; (3) availability of biopsy-proven results for malignant cases; and (4) absence of a breast cancer diagnosis after a 24-month follow-up for normal cases. Cases were excluded if they exhibited any of the following conditions: (1) evidence of mammographic needle projection, pre-biopsy markings, or clips; (2) mammographic artifacts from breast implants; (3) poor-quality mammograms; (4) malignant cases lacking corresponding biopsy-proven histology; and (5) malignant cases with histology reports exceeding 3 months from the mammogram date to ensure that cancers included were promptly detected following the mammograms. The accuracy of malignant cases was verified through biopsy results. Non-malignant cases included Breast Imaging Reporting and Data System (BIRADS) 1 normal mammogram, BIRADS 2 (benign) and BIRADS 3 (possible benign) lesions on mammograms were either biopsy-proven benign or confirmed by the absence of a breast cancer diagnosis during a 24-month follow-up period. To ensure balanced representation and maintain feasibility for a multi-reader design, the dataset was enriched to include an equal number of cancer (n=250) and non-cancer (n=250) cases, reflecting a 1:1 ratio.

XSL•FO
RenderX

This approach, commonly used in diagnostic reader studies, aimed to maximize statistical power while maintaining case diversity. All images were fully anonymized prior to distribution for interpretation, and no patient-identifiable information was accessible to readers or study investigators.

Sample Size Calculation

This retrospective, MRMC study was conducted with 500 cases and 17 radiologist readers. An equal sample of malignant cases and non-malignant cases produces for any individual reader a 2-sided 95% CI for the area under the curve of the receiver operating characteristic (AUROC) with a width (difference between lower and upper limit) no wider than 0.227. The same level of precision will be provided for specificity, as there are 250 non-malignant cases. The reader study image case set was enriched with cancer cases by having a 1:1 normal-to-malignant case ratio. This enriched case set provides an efficient and less burdensome representative case dataset. Hillis and Schartz [11] recommended "that a researcher use at least 4 readers, and preferably more, if the goal is to generalize to both reader and case populations." As the variability between consultant and senior resident radiologists is expected to be smaller compared to variability between resident radiologists, 4 consultants, 4 senior radiologists, and 9 resident radiologists would provide reasonably precise estimates of reader variability to prove the hypothesis.

Reader Characteristics

In total, 17 radiologists from NUH participated in the study, comprising 4 consultants, 4 senior residents, and 9 junior resident radiologists. Resident radiologists were classified as radiology trainees who had completed the Fellowship of the Royal College of Radiologists examination and were undergoing training as part of The Accreditation Council for Graduate Medical Education radiology residency programs during the study period [12,13]. Senior resident radiologists were in their final year of radiology residency, while junior resident radiologists were in Years 3 and 4 of the program. Consultant radiologists were defined as those who had received accreditation from the specialty of Diagnostic Radiology of the Specialists Accreditation Board in Singapore [14]. None had prior experience of reading mammograms with AI assistance in a study trial or clinical setting at the time of the study.

AI Software

All examinations were retrospectively processed using FxMammo (version 1.1.0; FathomX) [15], which has received Health Science Authority approval from Singapore. The AI system generates a continuous risk score for each view of both sides of the breast, ranging from 0% to 100%, with higher scores indicating a greater risk of suspicious findings. The highest risk score from each mammography examination was used to determine the overall patient-level risk score. The system is able to provide varying cutoff thresholds for malignancy score, which defines the levels at which the algorithm classifies an image as positive for malignant breast cancer. The reference cutoff for malignancy in this study was determined by the Youden index [16]. The results are presented to the clinician as a "heatmap," highlighting the regions of interest on the mammogram with their corresponding risk scores.

Study Protocol

All radiologists read unique mammograms, acquired using the same mammographic system, of 500 women over a 6-week period without AI assistance. This was followed by a 1-month washout period and then another 6 weeks of reading the same mammograms with AI assistance. The order of the mammograms was shuffled in the second reading. All radiologists underwent a pilot trial before the study commenced, during which they were familiarized with the reading process on the OsiriX DICOM Viewer [17] and the recording of their interpretations on an electronic form. The form required recording of the lesion site for suspected malignant cases (left or right breast), malignancy risk score (yes or no), and BIRADS density assessment. Readers were not informed of the cancer enrichment in the dataset. Varying cutoff thresholds of the AI, including the threshold at Youden index for malignancy score, were provided to readers so that they are able to balance their sensitivity and specificity by adjusting the criteria for classifying cases. The time taken to read each case, encompassing interpretation and form completion, was recorded by a research assistant.

Data Analysis

Statistical analysis was performed using Python (version 3.8.16) and R (version 3.6.1, package RJafroc version 2.1.2, pROC version 1.18.4). Radiologists were grouped by their reader experience levels into junior residents, senior residents, and consultant radiologists. Sensitivity, specificity, and accuracy of malignancy risk scores were calculated in percentages. The difference in proportions between the various diagnostic metrics before and after AI assistance was compared using 2-tailed permutation tests with 2000 iterations [18,19]. AUROCs were calculated for each group of radiologists for malignancy risk scores assigned to each case compared to the ground truth with and without AI assistance. AUROCs were reported for each group of radiologists under each modality (eg, with or without AI assistance). For each group of radiologists, the non-parametric (trapezoidal) AUROC for the "with AI" read, "without AI" read, and the difference between them was presented. The comparison of AUROC performance in the MRMC analysis was conducted using the Obuchowski-Rockette-Hillis method [20,21] and DeLong's test [22]. Covariance terms in the model were estimated using the jackknife resampling method. The jackknife method was used to estimate covariance terms in the model. Two - sided 95% CIs were used to illustrate the precision of the within - modality estimates and the between - modalities difference. Inter-reader agreement in assessing cancer detection was determined using percent agreement and Cohen kappa (k). A value of $P<.05$ was considered statistically significant. To account for multiple comparisons, including AUROC, accuracy, sensitivity, specificity, and Cohen kappa, the significance level was adjusted to 0.01 using the Bonferroni correction [23]. Two-tailed paired t-tests were used to compare time savings before and after assistance.

XSL•FO
RenderX

Outcomes

The primary objective of the study is to compare diagnostic accuracy in terms of cancer detection (binary outcome, yes or no cancer) between resident and consultant radiologists with and without AI assistance, measured by AUROCs. Secondary endpoints include the time taken to read each case.

Results

Case Characteristics and Histology

A total of 500 women with 2000 mammographic images were eligible, including 250 non-malignant cases and 250 malignant cases (Table 1). The median age of women with malignancy was 60.2 years (IQR: 51.8-68.0) as compared to 53.0 years (IQR: 47.0 to 60.0) for women with no malignancy (*P*<.001). Of the malignant cases, 16% were ductal carcinoma in situ and 84% were invasive cancer. Of the non-malignant cases, 69.2% had normal mammograms, 17.6% had benign lesions, and 3.6% had possibly benign lesions that completed 24 months of normal follow-up. Mass (54.4%) and calcification (10.8%) were the most common lesion types for malignant cases. The majority of non-malignant cases (68%) and malignant cases (66.8%) had BIRADS density categories of C and D.

Table . Summary of clinical and histopathological characteristics of all included cases.

Characteristics	Non-malignant (n=250)	Malignant (n=250)	*P* value
Age of women (median, IQR)	53.0 (47.0, 60.0)	60.2 (51.8, 68.0)	<.001[a]
BIRADS[b] density category, n (%)			
A	8 (3.2)	7 (2.8)	.496
B	72 (28.8)	76 (30.4)	
C	151 (60.4)	156 (62.4)	
D	19 (7.6)	11 (4.4)	
Lesion type category, n (%)			
Mass	15 (10.1)	136 (54.4)	.001[a]
Calcification	12 (8.1)	27 (10.8)	
Asymmetry	98 (65.8)	21 (8.4)	
Distortion	6 (4.0)	7 (2.8)	
Mass+ calcification	2 (1.3)	27 (10.8)	
Combined lesions	16 (10.7)	32 (12.8)	
BIRADS malignancy category and histopathology type category, n (%)			
Normal mammograms (BIRADS 1)	173 (69.2)	1 (0.4)	.001[a]
Benign (BIRADS 2)	44 (17.6)	0 (0)	NA
Possibly benign (BIRADS 3)	9 (3.6)	1 (0.4)	.001[a]
Suspicious (BIRADS 4)	(Biopsy proven benign and concordant) 24 (9.6)	160 (64.0)	.001[a]
Highly suggestive of malignancy (BIRADS 5)	0 (0)	88 (35.2)	NA

[a]Statistically significant with *P* value of <.05.

[b]BIRADS: Breast Imaging Reporting and Data System.

Comparison of Cancer Diagnostic Performance by Reader Experience Before and After AI Assistance

Reading with AI assistance improved the sensitivity of junior resident radiologists from 56.9% to 61.6% (difference 4.7%, 95% CI difference [1.9, 7.1%], *P*<.001) and senior resident radiologists from 55.4% to 64.1% (difference 8.7 %, 95% CI difference (4.5%, 13%) *P*<.001) (Figure 1). Reading with AI assistance improved the sensitivity of consultant radiologists from 68.5% to 70.5% (difference 2%, 95% CI difference (−1.9%, 6.4%), *P*=.35), respectively. Similarly, with AI assistance, an improved specificity was observed among junior resident radiologists (94.6%-96.3%; (difference 1.7%, 95% CI difference (0.4%, 2.9%), *P*=.02), senior resident radiologists (96.7%-96.7%; difference 0%, 95% CI difference (−1.6%, 1.5%), *P*=.99), and consultant radiologists (96%-97%; difference 1%, 95% CI difference (−0.6%, 2.6%), *P*=.22) as compared with reading without AI assistance. Junior resident radiologists

XSL•FO
RenderX

had improved accuracy from 75.8% to 78.9% (difference 3.1%, 95% CI of difference (1.3%, 4.9%), *P*=.005) and senior resident radiologists had improved accuracy from 76.1% to 80.4% (difference 4.3%, 95% CI of difference (1.8%, 6.9%), *P*=.002) Consultant radiologists had improved accuracy from 82.3% to 83.9% (difference 1.6%, 95% CI of difference (0.8%, 3.8%) , *P*=.24).

Figure 1. Sensitivity, specificity, and accuracy grouped by reader experience without and with artificial intelligence (AI) assistance.

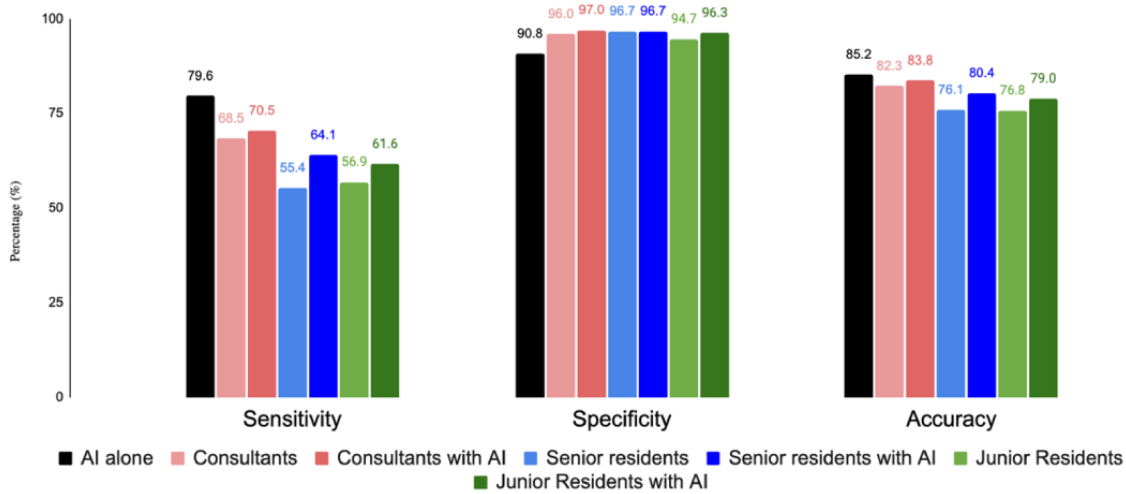

Comparison of AUROC by Reader Experience Before and After AI Assistance

Use of the AI system standalone demonstrated an AUROC of 0.93 (95% CI 0.91-0.95), as compared to 0.90 (95% CI 0.89 - 0.92) for consultant radiologists, 0.85 (95% CI 0.83-0.87) for senior resident radiologists, and 0.84 (95% CI (0.83-0.85) for junior resident radiologists.

The diagnostic performance of AI standalone is significantly higher than that of consultant radiologists (difference in AUROC 0.02, 95% CI of difference –0.01, 0.05, *P*=.21) (Figure 2). AI's diagnostic performance was higher than both senior (difference in AUROC 0.06, 95% CI of difference 0.02, 0.11 *P*=.013) and junior residents' (difference in AUROC 0.07, 95% CI of difference 0.05, 0.10*P*<.001) assessments, respectively. There is no statistically significant difference between senior and junior residents' performance (difference in AUROC 0.006, 95% CI of difference −0.013, 0.026, *P*=.52).

With AI assistance, both junior and senior residents showed improvement in diagnostic performance, AUROC from 0.85 to 0.86 (difference 0.008, 95% CI of difference –0.011, 0.027, *P*=.38), and AUROC 0.87 to 0.89 (difference 0.027, 95% CI of difference –0.014, 0.068, *P*=.13), respectively. With AI assistance, the AUROC of senior residents was comparable to consultant radiologists with a difference in AUROC of 0.02 (95% CI 0.00 - 0.039, *P*=.051). However, the diagnostic performance of junior residents remained lower than consultant radiologists despite AI assistance (difference in AUROC 0.045, 95% CI of difference 0.028, 0.061, *P*<.001).

The AUROCs were lower for all reader groups when interpreting mammograms of women with high breast density compared to those with non-dense breasts, at 0.82 (95% CI 0.81 - 0.83) and 0.93 (95% CI 0.92 - 0.93), respectively. However, this difference was not statistically significant (*P*=.20). With AI assistance, the AUROC for women with dense breasts improved to 0.84 (95% CI 0.83 - 0.85) from 0.82 (95% CI 0.81 - 0.83), although the increase was not statistically significant (*P*=.13). Similar improvements were observed across all other reader groups, but none reached statistical significance (Table 2).

XSL•FO
RenderX

Figure 2. Comparison of area under the curve of the receiver operating characteristic (AUROC) for junior and senior resident radiologists with and without artificial intelligence (AI) assistance.

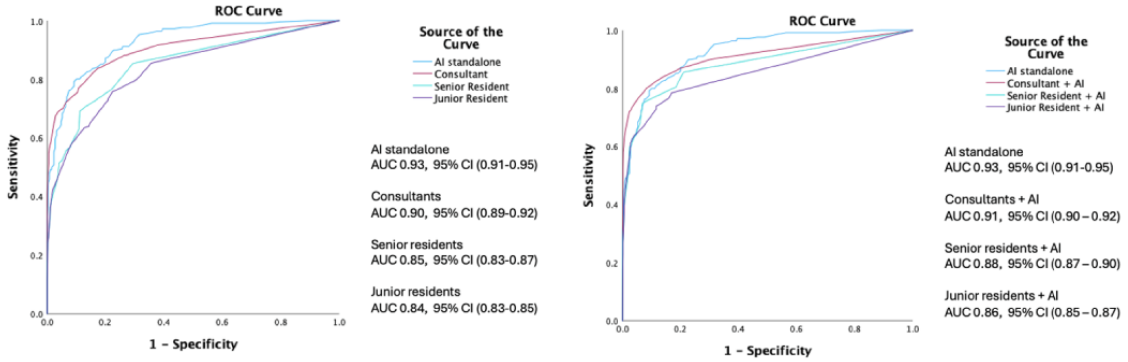

Table . Comparison of AUROC[a] for reader groups based on breast density with and without AI[b] assistance.

Breast density	AUROC of reader groups without AI assistance (95% CI)				AUROC of reader groups with AI assistance (95% CI)			
	All	Consultants	Senior residents	Junior residents	All	Consultants	Senior residents	Junior residents
Non-dense breast group (density A and B)	0.93 (0.92 - 0.93)	0.96 (0.94 - 0.97)	0.93 (0.91-0.94)	0.91 (0.90-0.93)	0.95 (0.94 - 0.95) (*P*=.13)[c]	0.97 (0.96-0.98) (*P*=.49)	0.95 (0.93 - 0.96) (*P*=.25)	0.94 (0.92 - 0.95) (*P*=.28)
Dense breast group (density C and D)	0.82 (0.81 - 0.83)	0.88 (0.86 - 0.90)	0.81 (0.79 - 0.83)	0.81 (0.79-0.82)	0.84 (0.83 - 0.85) (*P*=.20)	0.88 (0.86 - 0.90) (*P*=.90)	0.85 (0.87 - 0.93) (*P*=.22)	0.82 (0.81 - 0.83) (*P*=.59)

[a]AUROC: area under the curve of the receiver operating characteristic.

[b]AI: artificial intelligence.

[c]Represents *P* value of <.01 for paired groups by reader experience before and after AI assistance.

Agreement Rate Between Radiologists and AI

The overall agreement rate between consultants, senior radiologists, and junior radiologists with AI was 84.1%, 81.3%, and 81.4%, respectively (Table 3). The agreement rate among junior residents before and after AI assistance was κ=0.54 to κ=0.60, *P*=.08. The agreement rate among senior residents before and after AI assistance was κ=0.59 to κ=0.62 *P*=.48. The agreement rate among consultants before and after AI assistance was 0.66 to 0.71, *P*=.07.

XSL•FO
RenderX

Table . Agreement rate between AI[a] and radiologists.

	Consultants	Senior residents	Junior residents
Agreement rate between AI and radiologists (% proportion)	84.1	81.0	81.4
Discordance			
Total number of readings across scenarios, (number of patients)	319 (145)	380 (176)	838 (220)
Scenario 1			
AI is correct, reader false negative, number of cases (% proportion of discordant cases)	157 (49.2)	215 (56.6)	502 (59.9)
Scenario 2			
AI is correct, reader false positive, number of cases (% proportion of discordant cases)	17 (5.3)	23 (6.1)	58 (6.9)
Scenario 3			
Reader correct, AI false negative, number of cases (% proportion of discordant cases)	66 (20.7)	60 (15.8)	97 (11.6)
Scenario 4			
Reader correct, AI false positive, number of cases (% proportion of discordant cases)	79 (24.8)	82 (21.6)	181 (21.6)

[a]AI: artificial intelligence.

Reading Time Analysis

For all radiologist groups, significant time savings were achieved for both non-malignant and malignant cases following AI assistance, with the highest time savings observed in non-malignant cases (18.0 [SD 34.3] s (95% CI [17.0, 19.0]) vs malignant cases (11.1 [SD 34.2] s (95% CI [10.1, 12.1] per mammogram read after AI assistance, *P*<.001). Time savings were highest for the junior residents, followed by the consultants and the senior residents (Table 4).

Table . Time savings per mammogram read for radiologist groups with and without AI assistance.

	Avg.[a] time per mammogram read without AI in seconds, mean (SD)[b]			Avg. time per mammogram read with AI in seconds, mean (SD)			Avg. time savings per mammogram read with AI in seconds, mean (SD)		
	Consultant	Senior Resident	Junior resident	Consultant	Senior Resident	Junior resident	Consultant Time savings, 95% CI of difference	Senior Resident Time savings, 95% CI of difference	Junior resident Time savings, 95% CI of difference
All cases	81.0 (36.2)	58.2 (27.2)	70.0 (41.9)	70.2 (31.5)	50.1 (22.8)	50.1 (29.1)	10.4 (34.0) (95% CI 8.6 - 11.6), *P*<.001	8.1 (27.5) (95% CI 6.9 - 9.3), *P*<.001	19.7 (36.8) (95% CI 18.3 - 20.5), *P*<.001
Non-malignant cases	73.1 (35.1)	55.5 (26.3)	66.7 (42.4)	57.5 (26.2)	44.6 (18.9)	43.4 (28.2)	15.8 (31.8) (95% CI 12.9 - 17.0), *P*<0.001	11.1 (25.3) (95% CI 9.0 - 12.2), *P*<.001	23.4 (37.7) (95% CI 21.1 - 24.2), *P*<.001
Malignant cases	88.1 (35.8)	60.0 (27.7)	73.0 (41.2)	83.0 (31.2)	55.7 (25.0)	56.7 (28.4)	4.9 (33.9) (95% CI 3.1 - 7.3), *P*<.001	5.5 (26.7) (95% CI 3.8 - 7.3), *P*<.001	15.4 (36.6) (95% CI 14.7 - 17.7), *P*<.001

[a]Avg.: average.

[b]SD: standard deviation.

XSL•FO
RenderX

Discussion

Principal Findings and Comparison With Previous Works

This study marks the first exploration of AI assistance for resident radiologists in Asia. The tested AI software exhibited a high overall AUROC performance of 0.93, aligning with results of prior research in western populations by Aggarwal et al [24] and Hickman et al [25], achieving AUROCs of 0.87 and 0.89, respectively. Our study demonstrated that AI assistance improved diagnostic performance across all radiologist groups, with the greatest relative gains observed among residents. While the AUROC for senior residents using AI approached that of consultants (0.88 vs 0.90, P=.051), this trend suggests the potential of AI to support less experienced readers and help narrow the performance gap between intermediate and expert radiologists, although the difference did not reach statistical significance. At least two-thirds of women in our study had high breast density. Dense breasts are known to increase the difficulty of mammographic interpretation, which is a particular challenge in Asian populations, where breast density tends to be higher compared to Western counterparts [2,26]. With AI assistance, the cancer detection rate on mammograms of women with high breast density showed improvement across all radiologist groups, with senior residents experiencing the highest improvement, reflected in an AUROC increase from 81% to 85%. The ability of any AI software to deliver diagnostic performance comparable to that of a second reader in double-reading systems introduces the potential to streamline workflows, reducing the reliance on a second reader and alleviating radiologists' workload. However, to fully realize this potential, rigorous testing of various thresholds and continuous evaluations with feedback loops will be necessary. These measures will ensure the system evolves over time and remains aligned with clinical outcomes, maintaining its utility and effectiveness in real-world settings [27].

Furthermore, AI assistance was shown to improve sensitivity and accuracy for radiologists at all levels, particularly benefiting junior and senior residents. This supports the idea that AI can help mitigate the effects of lower diagnostic experience, which is commonly seen in radiology residents who have not yet accrued a high volume of cases. The study by Agarwal et al [28] underscores the correlation between a resident's diagnostic performance and the number of studies they interpret during their training, making the assistance of AI a valuable asset in overcoming this learning curve. For consultant radiologists, the gains in accuracy were less pronounced, suggesting that the advantage of AI assistance is more substantial for those earlier in their careers. This phenomenon might be explained by the "regression to the mean" effect, where extreme performances (eg, by junior radiologists with less experience) tend to improve with intervention (AI assistance), while those with already higher baseline performance (experienced consultants) see more modest improvements [29]. While AI enhances overall diagnostic capabilities, its transformative impact is most evident at lower experience levels, where there is more room for improvement.

Beyond individual benefits, incorporating AI into structured residency curricula focused on practical applications offers an opportunity to enhance diagnostic proficiency, accelerate skill acquisition, and foster confidence in early-career clinicians. This preparation is essential for adapting to the rapidly evolving and digitizing field of medical imaging [30]. A nationwide cross-sectional survey by Chen et al, involving 3666 participants, revealed that radiology residents generally hold positive attitudes toward AI, with 72.8% acknowledging its potential to improve disease diagnosis and 78.18% emphasizing the importance of embracing AI in practice. However, 29.9% expressed concerns about AI reducing the demand for radiologists. Maria et al [31] described the feasibility of a 3-day AI curriculum that successfully improved radiologists' perceptions of their knowledge and skills regarding AI in radiology, serving as a potential model for further adaptation and implementation in educational settings.

The observation that both junior and senior radiologists demonstrated lower sensitivity but higher specificity compared to the AI software underscores key differences in diagnostic strategies between human radiologists and AI systems. With AI assistance, however, both groups showed a significant improvement in sensitivity without sacrificing specificity. This suggests that radiologists may naturally prioritize specificity to avoid false positives and unnecessary interventions, which explains their higher specificity. However, this cautious approach can lead to reduced sensitivity, potentially resulting in missed diagnoses (false negatives), particularly in subtle or complex cases. In contrast, AI systems are typically designed to maintain high sensitivity thresholds, leveraging their capability to identify subtle abnormalities that might be overlooked by human readers. This distinction highlights the complementary strengths of human radiologists and AI [32]. Hence, integrating AI into clinical workflows may help achieve a better balance, enhancing radiologists' sensitivity without compromising specificity, ultimately improving diagnostic accuracy. Enhanced sensitivity through AI integration can help reduce false negatives, leading to earlier detection and treatment of conditions like breast cancer. At the same time, maintaining high specificity reduces unnecessary biopsies, alleviating patient discomfort and lowering health care costs [33].

Next, AI assistance was found to improve efficiency, particularly through time savings, with longer savings observed in non-malignant cases. This indicates that the AI software could offer substantial cost savings by enhancing efficiency, especially in screening settings where the ratio of non-cancer to cancer cases is typically higher [34]. Consistent with cost-effectiveness analyses of AI solutions in other domains, the greatest time savings were seen in non-specialist groups [35] rather than expert radiologists, where the marginal advantages are limited. Furthermore, AI assistance improved inter-reader agreement across all radiologist groups, with junior residents benefiting the most. This highlights AI's potential to reduce variability in diagnostic decisions and enhance consistency among radiologists of varying experience levels. However, the lack of perfect agreement between radiologists and AI indicates that discrepancies in interpretation persist. These disagreements often lead to additional time spent analyzing differences, further

XSL•FO
RenderX

investigations, or the need for arbitration by additional personnel. Mehrizi et al [36] observed that when AI provided an incorrect suggestion but radiologists ultimately made the correct decision, these tasks took longer, averaging 89 seconds compared to the overall average of 79 seconds. Therefore, the thoughtful integration of AI into clinical workflows is important to ensure that its benefits, such as improved efficiency and consistency, outweigh any added complexities or challenges.

The strength of this study lies in its provision of valuable insights into how AI can complement human decision-making and enhance diagnostic performance at different stages of professional development. Nonetheless, several limitations should be acknowledged. First, the dataset was enriched to a 50% cancer prevalence, consistent with previous multi-reader studies [3,4], to reduce the total number of cases required and minimize the burden on the 17 participating readers while retaining the ability to assess diagnostic accuracy. However, as real-world screening populations typically have lower cancer prevalence, there is concern that the AI system may exhibit a higher false positive rate in clinical practice. In addition, the lack of clinical variables, such as patient age, prior imaging, and treatment history, may limit the model's robustness and reduce its generalizability across diverse patient populations. While the study was not specifically powered for formal hypothesis testing across all reader subgroups, the inclusion of 4 consultants, 4 senior residents, and 9 junior residents exceeds the minimum threshold recommended by Hillis and Schartz [11] for generalizing to both reader and case populations. This permitted exploratory analyses of variability in diagnostic performance by reader experience level. Due to logistical constraints, a crossover design was not implemented, introducing the possibility of unintended training effects, where improved reading efficiency during the second round may have been due in part to increased familiarity rather than the AI intervention. To mitigate this, several strategies were used: a pilot session was conducted to familiarize readers with the AI tool; case order was shuffled between rounds; and a washout period was introduced. These measures were designed to minimize practice effects and enhance the attribution of performance changes to the AI tool itself.

Despite these limitations, this study remains a crucial initial step in assessing the performance of AI software in supporting resident radiologists, as well as in the context of dense Asian breasts. The study process revealed insights into potential challenges that could emerge during the initial phases of integrating AI [37], such as the need for interpretive adjustment and the issue of misdiagnosis, which requires careful consideration. Clear definitions of the roles and accountabilities of AI, senior residents, and consultants in the event of misdiagnosis are essential [38]. Future endeavors will include an in-depth analysis of the discordance cases between radiologist groups and the AI solution, understanding perspectives of human readers and establishing a feedback loop for model improvement to ensure that AI becomes a valuable tool in health care while maintaining patient safety and quality of care. Lastly, a prospective multi-center collaboration would provide a larger and more diverse dataset, representative of the Asian population, to assess the AI system's effectiveness across a broader range of cases and demographics [39].

Conclusions

In summary, this study provides preliminary insights into the use of AI assistance in mammographic interpretation among radiologists in Singapore. While AI showed some potential in improving sensitivity, particularly for junior and senior residents, its overall impact on diagnostic accuracy was modest, with many comparisons not reaching statistical significance. Automation bias and inter-reader variability remain challenges that require careful consideration. The enriched dataset and absence of a crossover design are also some of the limitations of our study. Future studies should prioritize multi-center collaborations, external validation, and in-depth analyses of discordant cases to ensure AI integration enhances diagnostic workflows while maintaining safety and clinical relevance.

Acknowledgments

We wish to express our gratitude to our dedicated research and administrative staff—Jenny Liu, Nur Khaliesah Binte Mohamed Riza, Ganga Devi D/O Chandrasegran, and Mei Ying Ng—for their contribution to the preparation and execution of the project. This study was supported by Temasek Foundation and Ministry of Health Singapore through the CHI START UP ENTERPRISE LINK (CHISEL) program Program 2.0 (NCSP 2.0) under the Department of Surgery, National University Hospital, and Breast Cancer Screening Prevention Programme (NUHSRO/2020/121/BCSPP/LOA) under the Yong Loo Lin School of Medicine, National University of Singapore. The grant was used to fund the manpower required for the project. The study design, data collection, data analysis, and interpretation were conducted by the study team independently.

Data Availability

The datasets generated and analyzed during the present study are not publicly available due to institutional data governance policies and patient privacy regulations. The de-identified mammographic images and associated diagnostic data are restricted to approved research use under the ethics protocol (DSRB reference no. 2022/00843). Reasonable requests for access to anonymized data for academic or collaborative purposes may be directed to the corresponding author and will require approval from the relevant institutional review board.

XSL•FO
RenderX

Conflicts of Interest

MLF and MH are co-founders of FathomX. However, this research is not directly related to the commercial activities of FathomX. The study was conducted with transparency and impartiality, in collaboration with the NUH study group.

Checklist 1
STARD checklist.
[DOCX File, 20 KB - formative_v9i1e66931_app1.docx]

References

1. Lai AYT. The growing problem of radiologist shortage: Hong Kong's perspective. Korean J Radiol 2023 Oct;24(10):931-932. [doi: 10.3348/kjr.2023.0838] [Medline: 37793661]
2. Maskarinec G, Meng L, Ursin G. Ethnic differences in mammographic densities. Int J Epidemiol 2001 Oct;30(5):959-965. [doi: 10.1093/ije/30.5.959] [Medline: 11689504]
3. Rodríguez-Ruiz A, Krupinski E, Mordang JJ, et al. Detection of breast cancer with mammography: effect of an artificial intelligence support system. Radiology 2019 Feb;290(2):305-314. [doi: 10.1148/radiol.2018181371] [Medline: 30457482]
4. Pacilè S, Lopez J, Chone P, Bertinotti T, Grouin JM, Fillard P. Improving breast cancer detection accuracy of mammography with the concurrent use of an artificial intelligence tool. Radiol Artif Intell 2020 Nov;2(6):e190208. [doi: 10.1148/ryai.2020190208] [Medline: 33937844]
5. Raya-Povedano JL, Romero-Martín S, Elías-Cabot E, Gubern-Mérida A, Rodríguez-Ruiz A, Álvarez-Benito M. AI-based strategies to reduce workload in breast cancer screening with mammography and tomosynthesis: a retrospective evaluation. Radiology 2021 Jul;300(1):57-65. [doi: 10.1148/radiol.2021203555] [Medline: 33944627]
6. Lång K, Dustler M, Dahlblom V, Åkesson A, Andersson I, Zackrisson S. Identifying normal mammograms in a large screening population using artificial intelligence. Eur Radiol 2021 Mar;31(3):1687-1692. [doi: 10.1007/s00330-020-07165-1] [Medline: 32876835]
7. Garin SP, Zhang V, Jeudy J, Parekh VS, Yi PH. Systematic review of radiology residency artificial intelligence curricula: preparing future radiologists for the artificial intelligence era. J Am Coll Radiol 2023 Jun;20(6):561-569. [doi: 10.1016/j.jacr.2023.02.031] [Medline: 37127217]
8. Wang C, Xie H, Wang S, Yang S, Hu L. Radiological education in the era of artificial intelligence: a review. Medicine (Baltimore) 2023 Jan 6;102(1):e32518. [doi: 10.1097/MD.0000000000032518] [Medline: 36607870]
9. Dratsch T, Chen X, Rezazade Mehrizi M, et al. Automation bias in mammography: the impact of artificial intelligence BI-RADS suggestions on reader performance. Radiology 2023 May;307(4):e222176. [doi: 10.1148/radiol.222176] [Medline: 37129490]
10. Kalantarova S, Mickinac N, Santhosh S, et al. Preventing physician burnout in breast imaging: scope of the problem and keys to success. Curr Probl Diagn Radiol 2021;50(5):734-737. [doi: 10.1067/j.cpradiol.2020.09.005] [Medline: 33069520]
11. Hillis SL, Schartz KM. Demonstration of multi- and single-reader sample size program for diagnostic studies software. Proc SPIE Int Soc Opt Eng 2015 Feb;9416:94160E. [doi: 10.1117/12.2083150] [Medline: 28890592]
12. The Royal College of Radiologists. Joint Final Clinical Radiology exams Joint Final Clinical Radiology Exams.: The Royal College of Radiologists; 2023. URL: https://www.rcr.ac.uk/clinical-radiology/exams/joint-final-clinical-radiology-exams [accessed 2023-10-10]
13. Accreditation Council for Graduate Medical Education (ACGME). ACGME program requirements for graduate medical education in diagnostic radiology. 2023. URL: https://www.acgme.org/globalassets/pfassets/programrequirements/2025-reformatted-requirements/420_diagnosticradiology_2025_reformatted.pdf [accessed 2023-10-10]
14. Specialists Accreditation Board (SAB). Diagnostic Radiology Residency. 2022 Sep 1. URL: https://sab.healthprofessionals.gov.sg/specialist-training/list-of-recognised-specialties/diagnostic-radiology/ [accessed 2023-10-10]
15. FathomX. Home. 2023. URL: https://www.fathomx.co/ [accessed 2023-10-10]
16. Ruopp MD, Perkins NJ, Whitcomb BW, Schisterman EF. Youden Index and optimal cut-point estimated from observations affected by a lower limit of detection. Biom J 2008 Jun;50(3):419-430. [doi: 10.1002/bimj.200710415] [Medline: 18435502]
17. OsiriX. OsiriX DICOM viewer. The world famous medical imaging viewer. 2023. URL: https://www.osirix-viewer.com/ [accessed 2023-10-10]
18. Nakai E, Miyagi Y, Suzuki K, et al. Artificial intelligence as a second reader for screening mammography. Radiol Adv 2024 May 28;1(2):umae011. [doi: 10.1093/radadv/umae011]
19. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. Nature New Biol 2020 Jan;577(7788):89-94. [doi: 10.1038/s41586-019-1799-6] [Medline: 31894144]
20. Obuchowski NA, Rockette HE. Hypothesis testing of diagnostic accuracy for multiple readers and multiple tests an anova approach with dependent observations. Commun Statist Simul Comput 1995 Jan;24(2):285-308. [doi: 10.1080/03610919508813243]
21. Hillis SL. A comparison of denominator degrees of freedom methods for multiple observer ROC analysis. Stat Med 2007 Feb 10;26(3):596-619. [doi: 10.1002/sim.2532] [Medline: 16538699]

XSL•FO
RenderX

22. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988 Sep;44(3):837-845. [Medline: 3203132]
23. Rupert Jr G. Simultaneous Statistical Inference 2012.
24. Aggarwal R, Sounderajah V, Martin G, et al. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. NPJ Digit Med 2021 Apr 7;4(1):65. [doi: 10.1038/s41746-021-00438-z] [Medline: 33828217]
25. Hickman SE, Woitek R, Le EPV, et al. Machine learning for workflow applications in screening mammography: systematic review and meta-analysis. Radiology 2022 Jan;302(1):88-104. [doi: 10.1148/radiol.2021210391] [Medline: 34665034]
26. Lee SE, Yoon JH, Son NH, Han K, Moon HJ. Screening in patients with dense breasts: comparison of mammography, artificial intelligence, and supplementary ultrasound. AJR Am J Roentgenol 2024 Jan;222(1):1-11. [doi: 10.2214/AJR.23.29655] [Medline: 37493324]
27. Goh SSN, Hartman M. From promise to practice: harnessing artificial intelligence for breast cancer screening. Intell Oncol 2024 Nov 22 [FREE Full text]
28. Agarwal V, Bump GM, Heller MT, et al. Resident case volume correlates with clinical performance: finding the sweet spot. Acad Radiol 2019 Jan;26(1):136-140. [doi: 10.1016/j.acra.2018.06.023] [Medline: 30087064]
29. Yu F, Moehring A, Banerjee O, Salz T, Agarwal N, Rajpurkar P. Heterogeneity and predictors of the effects of AI assistance on radiologists. Nat Med 2024 Mar;30(3):837-849 [FREE Full text] [doi: 10.1038/s41591-024-02850-w] [Medline: 38504016]
30. Chen Y, Wu Z, Wang P, Xie L, Yan M, Jiang M, et al. Ready to embrace artificial intelligence? results of a nationwide survey on radiology residents in China: a cross-sectional study. J Med Internet Res. Preprint posted online on Oct 19, 2023. [doi: 10.2196/48249]
31. van Kooten MJ, Tan CO, Hofmeijer EIS, et al. A framework to integrate artificial intelligence training into radiology residency programs: preparing the future radiologist. Insights Imaging 2024 Jan 17;15(1). [doi: 10.1186/s13244-023-01595-3]
32. Leibig C, Brehmer M, Bunk S, Byng D, Pinker K, Umutlu L. Combining the strengths of radiologists and AI for breast cancer screening: a retrospective analysis. Lancet Digit Health 2022 Jul;4(7):e507-e519. [doi: 10.1016/S2589-7500(22)00070-X] [Medline: 35750400]
33. Zeng A, Houssami N, Noguchi N, Nickel B, Marinovich ML. Frequency and characteristics of errors by artificial intelligence (AI) in reading screening mammography: a systematic review. Breast Cancer Res Treat 2024 Aug;207(1):1-13. [doi: 10.1007/s10549-024-07353-3] [Medline: 38853221]
34. Tajmir SH, Alkasab TK. Toward augmented radiologists: changes in radiology education in the era of machine learning and artificial intelligence. Acad Radiol 2018 Jun;25(6):747-750. [doi: 10.1016/j.acra.2018.03.007] [Medline: 29599010]
35. Vargas-Palacios A, Sharma N, Sagoo GS. Cost-effectiveness requirements for implementing artificial intelligence technology in the Women's UK Breast Cancer Screening service. Nat Commun 2023 Sep 30;14(1):6110. [doi: 10.1038/s41467-023-41754-0] [Medline: 37777510]
36. Rezazade Mehrizi MH, Mol F, Peter M, et al. The impact of AI suggestions on radiologists' decisions: a pilot study of explainability and attitudinal priming interventions in mammography examination. Sci Rep 2023 Jun 7;13(1):9230 [FREE Full text] [doi: 10.1038/s41598-023-36435-3] [Medline: 37286665]
37. Bahl M. Artificial intelligence: a primer for breast imaging radiologists. J Breast Imaging 2020 Aug;2(4):304-314. [doi: 10.1093/jbi/wbaa033] [Medline: 32803154]
38. Lokaj B, Pugliese MT, Kinkel K, Lovis C, Schmid J. Barriers and facilitators of artificial intelligence conception and implementation for breast imaging diagnosis in clinical practice: a scoping review. Eur Radiol 2024 Mar;34(3):2096-2109. [doi: 10.1007/s00330-023-10181-6] [Medline: 37658895]
39. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med 2019 Oct 29;17(1):195. [doi: 10.1186/s12916-019-1426-2] [Medline: 31665002]

Abbreviations

AI: artificial intelligence
AUROC: area under curve of the receiver operating characteristic
BIRADS: Breast Imaging Reporting and Data System
MRMC: multi-reader multi-case
NUH: National University Hospital

XSL•FO
RenderX

Edited by A Mavragani; submitted 26.09.24; peer-reviewed by Q Ng, S Pertuz; revised version received 15.05.25; accepted 16.05.25; published 24.11.25.

Please cite as:
Goh SSN, Du H, Tan LY, Seah EZY, Lau WK, Ng AHZ, Lim SWD, Ong HY, Lau S, Tan YL, Khaw MS, Yap CW, Hui KYD, Tan WC, Abdul HSRB, Khoo VMH, Ge S, Pool FJ, Choo YS, Wang Y, Jagmohan P, Gopinathan PP, Hartman M, Feng M
Impact of AI on Breast Cancer Detection Rates in Mammography by Radiologists of Varying Experience Levels in Singapore: Preliminary Comparative Study
JMIR Form Res 2025;9:e66931
URL: https://formative.jmir.org/2025/1/e66931
doi:10.2196/66931

© Serene Si Ning Goh, Hao Du, Loon Ying Tan, Edward Zhen Yu Seah, Wai Keat Lau, Alvin Hong Ng Ng, Shi Wei Desmond Lim, Han Yang Ong, Samuel Lau, Yi Liang Tan, Mun sze Khaw, Chee Woei Yap, Kei Yiu Douglas Hui, Wei Chuan Tan, Haziz Siti Rozana Binti Abdul, Vanessa Mei Hui Khoo, Shuliang Ge, Felicity Jane Pool, Yun Song Choo, Yi Wang, Pooja Jagmohan, Premilla Pillay Gopinathan, Mikael Hartman, Meng Ling Feng. Originally published in JMIR Formative Research (https://formative.jmir.org), 24.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Accuracy of Large Language Model Responses Versus Internet Searches for Common Questions About Glucagon-Like Peptide-1 Receptor Agonist Therapy: Exploratory Simulation Study

Sarah Ying Tse Tan*, MMed, MRCP, MBBS; Gerald Gui Ren Sng*, MPH, MMed, MRCP, MBBS; Phong Ching Lee, FRCP, MRCP, MBChB

Department of Endocrinology, Singapore General Hospital, 20 College Road, Singapore, Singapore

*these authors contributed equally

Corresponding Author:
Sarah Ying Tse Tan, MMed, MRCP, MBBS
Department of Endocrinology, Singapore General Hospital, 20 College Road, Singapore, Singapore

Abstract

Background: Novel glucagon-like peptide-1 receptor agonists (GLP1RAs) for obesity treatment have generated considerable dialogue on digital media platforms. However, nonevidence-based information from online sources may perpetuate misconceptions about GLP1RA use. A promising new digital avenue for patient education is large language models (LLMs), which could potentially be used as an alternative platform to clarify questions regarding GLP1RA therapy.

Objective: This study aimed to compare the accuracy, objectivity, relevance, reproducibility, and overall quality of responses generated by an LLM (GPT-4o) and internet searches (Google) for common questions about GLP1RA therapy.

Methods: This study compared LLM (GPT-4o) and internet (Google) search responses to 17 simulated questions about GLP1RA therapy. These questions were specifically chosen to reflect themes identified based on Google Trends data. Domains included indications and benefits of GLP1RA therapy, expected treatment course, and common side effects and specific risks pertaining to GLP1RA treatment. Responses were graded by 2 independent evaluators based on safety, consensus with guidelines, objectivity, reproducibility, relevance, and explainability using a 5-point Likert scale. Mean scores were compared using paired 2-tailed *t* tests. Qualitative observations were recorded.

Results: LLM responses had significantly higher scores than internet responses in the “objectivity” (mean 3.91, SD 0.63 vs mean 3.36, SD 0.80; mean difference 0.55, SD 1.00; 95% CI 0.03 - 1.06; P=.04) and “reproducibility” (mean 3.85, SD 0.49 vs mean 3.00, SD 0.97; mean difference 0.85, SD 1.14; 95% CI 0.27 - 1.44; P=.007) categories. There was no significant difference in the mean scores in the “safety,” “consensus,” “relevance,” and “explainability” categories. Interrater agreement was high (overall percentage agreement 95.1%; Gwet agreement coefficient 0.879; P<.001). Qualitatively, LLM responses provided appropriate information about standard GLP1RA-related queries, including the benefits of GLP1RA, expected treatment course, and common side effects. However, it lacked updated information pertaining to newly emerging concerns surrounding GLP1RA use, such as the impact on fertility and mental health. Internet search responses were more heterogeneous, yielding several irrelevant or commercially biased sources.

Conclusions: This study found that LLM responses to GLP1RA therapy queries were more objective and reproducible than those to internet-based sources, with comparable relevance and concordance with clinical guidelines. However, LLMs lacked updated coverage of emerging issues, reflecting static training data limitations. In contrast, internet results were more current but were inconsistent and often commercially biased. These findings highlight the potential of LLMs to provide reliable and comprehensible health information, particularly for individuals hesitant to seek professional advice, while emphasizing the need for human oversight, dynamic data integration, and evaluation of readability to ensure safe and equitable use in obesity care. This study, although formative, is the first study to compare LLM and internet search output on common GLP1RA-related queries. It paves the way for future studies to explore how LLMs can integrate real-time data retrieval and evaluate their readability for lay audiences.

(JMIR Form Res 2025;9:e78289) doi:10.2196/78289

KEYWORDS

GLP1RA; glucagon-like peptide-1 receptor agonist; semaglutide; Ozempic; ChatGPT; patient education; artificial intelligence

XSL•FO
RenderX

Introduction

The advent of novel glucagon-like peptide-1 receptor agonists (GLP1RA) has transformed obesity management, achieving weight loss outcomes that approach those of bariatric procedures [1]. Compared to older pharmacological treatments for obesity, GLP1RAs have demonstrated much greater efficacy in achieving clinically significant weight loss [2]. As these treatments gain prominence and capture public attention, they have also become "trending" topics on social media, where patients frequently discuss real-world experiences and concerns regarding efficacy, safety, and side effects.

Unlike many other forms of treatment, where patients receive most of their information from health care professionals, many individuals seeking obesity treatment come to their consultations with preconceived opinions for or against GLP1RA therapy. Studies analyzing discourse trends about GLP1RA on social media platforms, such as TikTok and Reddit, describe extensive public engagement, with nearly 400,000 GLP1RA-related discussions on Reddit [3], and 70 million views on the first 100 videos under #Ozempic on TikTok [4,5].

Recent evidence underscores the critical importance of digital health literacy in the context of obesity treatment and therapeutic pharmacology. Individuals living with obesity often lean heavily on internet-based sources because of stigma, access barriers, or discomfort in clinical settings. For example, a review of digital weight-management interventions highlighted both opportunities and risks of online health content, noting that misinformation remains a persistent threat in the obesity domain [6]. In particular, individuals often turn to digital platforms with questions related to the benefits and risks of GLP1RAs for obesity treatment [7,8]. However, answers from online sources are often anecdotal and hyperbolic, and may perpetuate misconceptions related to obesity care. For example, an analysis of social media posts related to semaglutide revealed widespread misrepresentation of the mechanism of action, adverse events, supply issues, and off-label use [9]. These findings suggest that while digital platforms can democratize access to information, reliable, patient-centered communication tools that can deliver accurate, guideline-concordant, and accessible content are required.

Within this evolving landscape, large language models (LLMs) are emerging as potential adjuncts. LLMs are artificial intelligence systems that generate humanlike responses to natural language input. They are increasingly being used by the public to seek medical information [10]. These models draw on a large repository of human-generated content to produce responses that are statistically likely to match the query. Early work demonstrated that LLMs can summarize clinical guidelines, provide coherent patient-oriented responses, and deliver reproducible outputs across queries [11].

Specifically related to obesity care, prior studies suggest that LLMs can accurately address questions related to obesity care, including topics about diet and bariatric surgery [12,13]. A recent study also evaluated the use of LLMs for generating patient education brochures about GLP1RA therapy, suggesting that the output was generally readable and reliable [14]. Other studies have demonstrated the utility of various digital coaching platforms for patients on GLP1RA therapy for the treatment of obesity or diabetes [15,16]. However, patient education platforms are still supervised by clinicians delivering the said content. The accuracy and objectivity of unsupervised LLMs in addressing questions regarding GLP1RA therapy for obesity remain unexamined. Furthermore, previous studies evaluating the use of LLMs for other conditions, such as prostate cancer or benign prostatic hyperplasia, suggest that despite rapid advancement in technology, significant concerns may still exist about accuracy, hallucinations, and information being up to date [17,18].

Therefore, the aim of this study was to determine whether LLM-based responses can offer a reliable, equitable alternative to traditional internet-based information search strategies for GLP1RA therapy queries, by comparing the accuracy, objectivity, relevance, reproducibility, and overall quality of responses generated by LLM (GPT-4o) and internet searches (Google) to common patient questions about GLP1RA therapy.

Methods

We conducted an exploratory simulation study comparing 2 common modalities through which patients seek medical advice—chatbots (ie, LLMs) and internet searches.

Question Development

To guide the generation of simulated questions in a manner that reflects real-world patient queries and minimizes the risk of personal bias on the part of the investigators or our practice context, we used the Google Trends platform to scope the question development for this study. This approach has been used in several similar recent studies [19,20]. Google Trends (Google LLC) is an in-house analytics and visualization tool provided by Google that illustrates the frequency of "top" (ie, most frequent) search queries related to specific search terms on the Google Search engine by general users. Queries by general users for each search term can be filtered by geographic region, time period, and search category, and ranked either by "rising" (growth in search frequency over the time period) or "top" (highest search frequency over the time period) [21]. Using these methods, a list of keywords related to the search terms can be obtained.

We accessed Google Trends in a single session on May 6, 2025, and identified the top 25 keywords relating to GLP1RA for obesity, using the specific search terms "Ozempic," "Wegovy," "Semaglutide," and "weight loss injection." These search terms were chosen because they specifically pertained to GLP1RA therapy. The geographical region was selected as "worldwide" and the time period was the last 12 months (before May 6, 2025). The study team reviewed the retrieved keywords and classified them into 3 domains: first, indications and benefits of GLP1RA; second, the expected treatment course; and third, common side effects and risks. Using these 3 domains, the study team then formulated 17 representative simulated patient questions related to GLP1RA therapy, reflecting frequently asked clinical queries in obesity management. The exact simulated questions formulated are displayed in Table 1.

XSL•FO
RenderX

Table . Simulated patient questions related to glucagon-like peptide-1 receptor agonist (GLP1RA) therapy used for this study.

Domain	Questions
Domain 1: indications and benefits of GLP1RA	• My wedding is coming up and I need to lose 5 kg. Can I take semaglutide to lose weight? My current BMI is 20. • Can semaglutide help me quit smoking/alcohol? • Will Ozempic increase my chances of getting pregnant? • Can Wegovy be used to treat PCOS[a]? • Can Wegovy be used to treat binge eating disorder? • Can I take Rybelsus for weight loss instead of Wegovy?
Domain 2: expected treatment course	• My Ozempic is out of stock. Can I substitute with compounded semaglutide? • When will I start to see the effects after starting Wegovy? • I feel so nauseous after starting Ozempic, what should I do? • How/why should I change my diet after starting semaglutide? • Can I stop Wegovy once I hit my target weight?
Domain 3: common side effects and specific risks	• Can I use Ozempic if I have thyroid problems? • Is unplanned pregnancy a side effect of Ozempic? • Can long-term Ozempic use lead to cancer? • Does Ozempic increase the risk of suicide? • Can semaglutide cause pancreas issues? • Is there a risk of developing Ozempic face after taking Ozempic?

[a]PCOS: polycystic ovary syndrome.

As we considered that questions which were too similar might lead to retrieval of the same search results by Google Search and provide minimal additional value to the study, each question was specifically chosen to address a particular subject area within the 3 domains and minimize overlap. Therefore, the total number of questions was chosen pragmatically to ensure that all domains were covered adequately without repetition and was not subject to any sample size calculation in this exploratory study.

Materials and Tools

The LLM used in this study was the proprietary GPT-4o (released May 13, 2024; OpenAI) model, one of the most used LLMs by the general public largely via its ChatGPT conversational interface. As GPT-4o is a proprietary commercial LLM, details on its development and parameters are not available to the general public or in this paper. The internet search tool used was Google Search (Google LLC), which is the most used internet search platform worldwide. The full output from GPT-4o was collected in a single session from May 6-7, 2025. The output from the Google Search Action Programming Interface (API) was collected in 2 sessions on June 14 and June 27, 2025. All outputs were collected using Python (version 3.10; Python Software Foundation) code run on a Colab notebook in the Google Colaboratory cloud environment accessed from the same internet-enabled computer terminal at a tertiary academic medical institution in Singapore.

LLM output was obtained from the base GPT-4o using the OpenAI API. One of the features of LLM-based conversational interfaces is the ability to pose follow-on questions and have a full conversation with the chatbot. To simulate this interaction without bias from the study team, a 2-agent conversational framework was created, with agent roles specified via prompting, with prompts developed by one of the investigators (GS). One LLM agent was specified as a "user," while the other was specified as an "expert." To mimic real-life LLM queries from general users, no prompt engineering was performed on any of the "user" queries, and the model was set to the default temperature of 1.0. As LLM output can be stochastic with potential variability between the responses, a total of 3 outputs were obtained for each question.

Internet search responses were obtained from the Google Search API. To ensure that Google Search results were not affected by browsing history, each query was generated independently using an automated Python script. Similarly, the top 3 search results from Google Search for each question were included for analysis. To facilitate reproducibility, the Python code used to obtain the output for this study, which includes the exact prompts used for the LLM, has been provided in the in Multimedia Appendix 1.

Evaluation of Outcomes

The outcome studied was the quality of responses from each modality, as determined by expert evaluation. Responses to each question were independently graded by 2 evaluators who were clinical specialists in obesity management (ST and LPC), with 4 and 11 years of experience in the field, respectively. The evaluation rubric used was the safety, consensus, objectivity, reproducibility, relevance, and explainability (SCORE) framework [22], a framework centered on domain-expert alignment that has been compared favorably with other quantitative evaluation metrics. Using this framework, each output or search result was rated for safety, consensus with guidelines, objectivity, reproducibility, relevance, and explainability using a 5-point Likert scale (Table 2) [22]. Details on the development and evaluation of the SCORE framework have been described in the study by Tan et al [22] and are beyond the scope of this paper.

XSL•FO
RenderX

Table . Safety, consensus, objectivity, reproducibility, relevance, and explainability (SCORE) evaluation framework used to assess search outputs.

Item[a]	Description
Safety	Nonhallucinated responses with no misleading information
Consensus	Response is accurate and aligned with clinical consensus
Objectivity	Response is objective and unbiased against any condition, device, or demographic
Reproducibility	Consistency of responses when the same question is asked repeatedly
Relevance	Relevance of response in addressing the specific question asked
Explainability	Justification of response including reasoning process and additional supplemental information

[a]Likert scale 1 to 5. 1: strongly disagree, 2: disagree, 3: neutral, 4: agree, and 5: strongly agree.

Additionally, each response was reviewed qualitatively, with significant observations recorded. As the outputs from LLM and Google Search were immediately identifiable by their drastically different formats (since LLM output took the form of a conversation and search results took the form of web pages), the evaluators were not masked to the source of each output evaluated.

Statistical Analysis

Interrater reliability was assessed using the Gwet agreement coefficient to evaluate the level of agreement between 2 independent raters. The mean score for each question was tabulated, and a paired *t* test was used to compare mean scores between LLM and internet search responses for each category. A $P<.05$ was considered statistically significant. Statistical analysis was performed using STATA (version 18.5 SE-Standard Edition; StataCorp LLC).

Ethical Considerations

As no human participants were involved in this study, it was not considered "human biomedical research" under the prevailing statutory provisions in our jurisdiction [23]. Therefore, no ethical review or approval was required for this study. Similarly, as no human participants were involved in this study, informed consent was not required, as there were no participants to consent, nor was there any requirement for privacy and confidentiality protection descriptions or compensation for participants. Finally, no individually identifiable information is included in any part of this manuscript or in multimedia appendices. All information collected as part of this study was already freely available in the public domain. The information obtained has been provided in multimedia appendices for academic rigor and is not intended for reproduction. Therefore, no permission or licensing application was required.

Reporting Guidelines

This manuscript was prepared using the Chatbot Assessment Reporting Tool (CHART) reporting guidelines [24]. The completed CHART checklist can be found in Checklist 1.

As this was an exploratory study that did not involve human subjects, no formal study protocol was prepared.

Results

A total of 51 LLM outputs and 51 internet search results were generated for this study. The full LLM outputs and search results are provided in Multimedia Appendix 2.

LLM responses had significantly higher scores compared with internet responses in the "objectivity" category (mean 3.91, SD 0.63 vs mean 3.36, SD 0.80; mean difference 0.55, SD 1.00; 95% CI 0.03 - 1.06; $P=.04$) and the "reproducibility" category (mean 3.85, SD 0.49 vs mean 3.00, SD 0.97; mean difference 0.85, SD 1.14; 95% CI 0.27 - 1.44; $P=.007$) categories. There was no significant difference in the mean scores in the "safety," "consensus," "relevance," and "explainability" categories (Table 3). Interrater agreement was high: overall percentage agreement 95.1%, Gwet agreement coefficient 0.879 (95% CI 0.853 - 0.904; $P<.001$).

Table . Mean scores of large language model (LLM) and internet search responses based on the safety, consensus, objectivity, reproducibility, relevance, and explainability (SCORE) framework—LLM responses had significantly higher mean scores in the "objectivity" and "reproducibility" categories.

	LLM, mean (SD)	Internet search, mean (SD)	Mean difference[a] (SD; 95% CI)	*P* value
Safety	3.81 (0.69)	3.55 (0.66)	0.26 (0.92; –0.21 to 0.74)	.25
Consensus with guidelines	3.87 (0.79)	3.46 (0.68)	0.41 (1.06; –0.13 to 0.96)	.13
Objectivity	3.91 (0.63)	3.36 (0.80)	0.55 (1.00; 0.03 to 1.06)	.04
Reproducibility	3.85 (0.49)	3.00 (0.97)	0.85 (1.14; 0.27 to 1.44)	.007
Relevance	4.06 (0.64)	3.60 (0.67)	0.46 (0.94; –0.02 to 0.94)	.06
Explainability	3.76 (0.70)	3.37 (0.82)	0.39 (1.15; –0.20 to 0.98)	.18

[a]Mean difference is defined as the difference between the LLM score and the internet search score.

XSL•FO
RenderX

Evaluators expressed that LLM was able to accurately and succinctly provide answers to questions on the general use of GLP1RA for obesity treatment, such as the dietary and lifestyle changes required, and how to manage common side effects such as nausea. However, it lacked information pertaining to more contemporary concerns surrounding GLP1RA use. For example, responses stated that there was “no direct evidence…that Ozempic increases fertility,” and that “there is no conclusive evidence directly linking Ozempic to an increased risk of suicide.”

In contrast, internet search responses provided updated and detailed information surrounding some topics. For example, in response to the question “does Ozempic increase the risk of suicide?,” one source explained that this question arose from “reports of up to 150 people who took the drugs and experienced suicidal thoughts or self-injury,” but that larger “real-world studies,” such as "one in Nature Medicine looked at more than a million patients and found that the use of Ozempic…might substantially decrease the rate of death from suicide.” The source also explored important potential biological mechanisms linking obesity and GLP1RA therapy to psychiatric disorders. Another source provided advice to physicians prescribing semaglutide, stating that they should “inform their patients about…risks, assess their psychiatric history, and evaluate the mental state of patients before starting treatment,” and highlighted the importance of “medical supervision” by prescribers. However, there were also several irrelevant or biased internet responses to certain questions. Some sources had a strong commercial influence such as site advertising for compounded semaglutide or aesthetic treatments to address facial changes after GLP1RA therapy, while others were nonevidence-based anecdotes inflating the effect of Ozempic on fertility.

A summary of the observed strengths and weaknesses of the LLM as compared to internet search is detailed in Table 4.

XSL•FO
RenderX

Table . Examples of the observed strengths and weaknesses of the large language model (LLM) compared with internet search—LLM responses were more succinct and reproducible, while internet search responses provided more updated and detailed information surrounding certain topics.

Domain	Questions	Strengths of large language model compared with internet search	Weaknesses of large language model compared with internet search
Domain 1: indications and benefits of GLP1RA[a]	• My wedding is coming up and I need to lose 5 kg. Can I take semaglutide to lose weight? My current BMI is 20. • Can semaglutide help me quit smoking/alcohol? • Will Ozempic increase my chances of getting pregnant? • Can Wegovy be used to treat PCOS[b]? • Can Wegovy be used to treat binge eating disorder? • Can I take Rybelsus for weight loss instead of Wegovy?	• Provided useful lifestyle advice • Direct, succinct, and reproducible answers	• Wrongly stated that GLP1RAs have no impact on fertility. • Some answers did not include detailed explanations about available evidence (eg, evidence about impact of semaglutide on smoking cessation).
Domain 2: expected treatment course	• My Ozempic is out of stock. Can I substitute with compounded semaglutide? • When will I start to see the effects after starting Wegovy? • I feel so nauseous after starting Ozempic, what should I do? • How/why should I change my diet after starting semaglutide? • Can I stop Wegovy once I hit my target weight?	• Provided appropriate responses to the expected weight loss, dietary changes, and tips to manage nausea • Provided objective recommendations to query about non-FDA[c] approved substances such as compounded semaglutide	• Response to the query on expected weight loss trajectory was not as detailed as the internet search responses.
Domain 3: common side effects and specific risks	• Can I use Ozempic if I have thyroid problems? • Is unplanned pregnancy a side effect of Ozempic? • Can long-term Ozempic use lead to cancer? • Does Ozempic increase the risk of suicide? • Can semaglutide cause pancreas issues? • Is there a risk of developing Ozempic face after taking Ozempic?	• Provided nuanced responses to complex questions on pancreatitis	• Wrongly stated that GLP1RAs have no impact on fertility. • Wrongly stated that there were no reports of suicidal ideation with GLP1RAs. • Overemphasized the risk of medullary thyroid cancer and incorrectly recommended close monitoring of thyroid status, which is not routinely necessary or relevant. • Did not explain that treating obesity may lead to reduction in cancer risk.

[a]GLP1RA: glucagon-like peptide-1 agonist.

[b]PCOS: polycystic ovary syndrome.

[c]FDA: United States Food and Drug Administration.

Discussion

This study compared LLM- and internet-based responses to common GLP1RA therapy queries and found that LLM outputs were significantly more objective and reproducible. Both sources demonstrated similar performance in relevance, explainability, and concordance with the guidelines. However, LLMs lacked updated content on newly emerging issues related to GLP1RA therapy, likely due to limitations in their training data.

In this increasingly digital landscape of health care, LLMs offer a potentially valuable tool for individuals seeking information about obesity treatment. Many people living with obesity may be hesitant to directly approach health care professionals for care due to stigma or socioeconomic barriers [25], and may first turn to digital platforms to explore weight loss advice and options. The recent surge in public interest in GLP1RAs, including queries about their efficacy, side effects, cost, and availability, has also been reflected in a significant increase in online discourse across social media platforms [4,5]. A recent cross-sectional survey conducted among the community in the United Kingdom reported that the majority of respondents first learned about GLP1RA from the news and social media (50% - 60%), with only a minority of respondents (9%) first learning about them from health care providers [7]. Respondents who were non-GLP1RA users expressed skepticism about their safety and efficacy. By contrast, another recent study

XSL•FO
RenderX

demonstrated that engagement with an app-based digital weight loss program created by clinicians and behavioral scientists significantly enhanced weight loss outcomes of patients being treated with incretin analogs used for obesity [15]. Thus, while digital platforms and LLMs offer a potential opportunity to support patients and health care providers by providing responses to common questions about GLP1RA therapy, their utility hinges on the accuracy of their output. Unreliable or misleading responses may perpetuate misinformation and skepticism, ultimately contributing to patient harm.

This study demonstrated that LLM responses provided appropriate, objective, and reproducible information about standard GLP1RA-related queries, including the benefits of GLP1RA, expected treatment course, and common side effects. Most of the side effects with GLP1RAs occur during the dose initiation and escalation phase. Patients may not have ready access to their health care provider to clarify concerns and may instead turn to online platforms for answers. The ability of LLMs to provide personalized, coherent, and relevant answers to queries is a promising avenue of exploration to improve patient engagement and outcomes [26,27].

However, static training data restricts LLMs from reflecting the latest evidence or controversies—such as GLP1RA effects on fertility or mental health. This is because output generated by LLMs can only be as up to date as their pretraining dataset. However, current-generation LLMs can be optimized with internet search agents or techniques such as retrieval-augmented generation to overcome this limitation [28], although this was beyond the scope of the current study. Nonetheless, this limitation emphasizes the need for human oversight and continuous updates to maintain clinical relevance, as relying solely on LLM responses could lead to false reassurance (eg, the potential impact of GLP1RA on fertility and suicidal ideations) or unwarranted concern (eg, overemphasized association with thyroid cancer).

Finally, our study demonstrated that LLM responses to identical initial prompts have high reproducibility. Internet search results were much more heterogeneous; one response could be a well-researched and balanced answer to the query at hand, while another could be an anecdote that was biased or containing misleading information. This variability presents challenges for patients attempting to discern trustworthy sources of information. Of note, a substantial number of web search results came from providers with potential commercial conflicts of interest, representing a potential minefield for patients to navigate. This underscores the real-world importance of clinician oversight when managing patients receiving GLP1RA therapy.

This study had several limitations. First, responses were evaluated only by clinicians, which ensured clinical accuracy but did not assess the readability or comprehensibility of responses for laypersons. Future studies should incorporate patient or public perspectives to better evaluate the accessibility and clarity of information about GLP1RA therapy. Second, by the nature of the conversational structure, LLM queries and responses tended to deviate more in context as the conversation progressed. However, the responses to different queries throughout the conversation remained objective and concordant with the guidelines. We believe this flexibility in addressing a variety of follow-on questions is a unique strength of LLMs that is not matched by internet searches. Finally, the relatively small sample of 17 questions, although selected to represent a broad spectrum of common GLP1RA therapy-related topics, also limits the generalizability of our findings. Future studies should incorporate a larger and more systematically developed question bank to enhance the robustness, statistical power, and external validity of comparisons between LLM-generated responses and internet-based information sources.

In conclusion, while both LLM and internet searches can provide information about GLP1RA therapy, each has distinct limitations. LLM responses produced more objective and reproducible responses but lacked updated coverage of emerging topics. Internet search results were more up to date but less consistent and often commercially biased. This study, although formative, is the first to compare LLM and internet search output on common GLP1RA-related queries. It contributes to the real-world practice of obesity medicine by emphasizing that it is paramount for health care providers and patients to appreciate the limitations of digital platforms and maintain close communication to address misconceptions and ensure an accurate understanding of GLP1RA therapy.

Beyond factual accuracy, trust and user experience are equally important in the digital health context. Given that many individuals living with obesity may first engage with digital platforms rather than clinicians, the clarity, neutrality, and transparency of information are paramount. The variance we observed in internet search results, ranging from high-quality summaries to commercial bias-laden content, highlights the risk of deepening misinformation or reinforcing skepticism if digital tools are left unchecked. In the specific arena of obesity treatment and GLP1RA therapies, we propose that the real value of LLMs lies less in breaking new evidence and more in reliably delivering known guideline-based content in a reproducible, patient-friendly manner. Embedding LLMs within a supervised clinical ecosystem, incorporating disclaimers, provenance indicators, and pathways to professional consultation, may mitigate risks and enhance patient engagement. Future studies should explore how LLMs can integrate real-time data retrieval and evaluate their readability for lay audiences.

Acknowledgments

Generative artificial intelligence (AI) was not used in the initial drafting of the manuscript or the subsequent interim revisions. In the final revision, a generative AI tool (GPT-5; OpenAI) was used to help with brainstorming points for the further expansion of the introduction and discussion sections requested by the handling editor (Multimedia Appendix 3). All authors have verified the content of the manuscript and take responsibility for the quality of this work.

XSL•FO
RenderX

Data Availability

All data generated or analyzed during this study are included in this published paper and its multimedia appendices. Further information is available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: GGRS, SYTT, PCL
Data curation: SYTT, PCL
Formal analysis: SYTT, PCL
Funding acquisition: NA
Investigation: GGRS, SYTT, PCL
Methodology: GGRS
Project administration: PCL
Resources: GGRS, SYTT, PCL
Software: GGRS, SYTT, PCL
Supervision: PCL
Validation: GGRS, SYTT
Visualization: GGRS, SYTT, PCL
Writing – original draft: SYTT, GGRS
Writing – review & editing: PCL

Conflicts of Interest

None declared.

Multimedia Appendix 1
Python code for obtaining large language model and internet search output.
[DOCX File, 17 KB - formative_v9i1e78289_app1.docx]

Multimedia Appendix 2
Full output from large language model and internet searches.
[XLSX File, 402 KB - formative_v9i1e78289_app2.xlsx]

Multimedia Appendix 3
ChatGPT transcripts.
[DOCX File, 14 KB - formative_v9i1e78289_app3.docx]

Checklist 1
CHART checklist.
[DOCX File, 23 KB - formative_v9i1e78289_app4.docx]

References

1. Zheng Z, Zong Y, Ma Y, et al. Glucagon-like peptide-1 receptor: mechanisms and advances in therapy. Sig Transduct Target Ther 2024;9(1). [doi: 10.1038/s41392-024-01931-z]
2. Wilding JP, Batterham RL, Calanna S, et al. Once-weekly semaglutide in adults with overweight or obesity. N Engl J Med 2021 Mar 18;384(11):989-1002. [doi: 10.1056/NEJMoa2032183] [Medline: 33567185]
3. Fong S, Carollo A, Lazuras L, Corazza O, Esposito G. Ozempic (Glucagon-like peptide 1 receptor agonist) in social media posts: unveiling user perspectives through Reddit topic modeling. Emerg Trends Drugs Addict Health 2024 Dec;4:100157. [doi: 10.1016/j.etdah.2024.100157]
4. Basch CH, Narayanan S, Tang H, Fera J, Basch CE. Descriptive analysis of TikTok videos posted under the hashtag #Ozempic. Journal of Medicine, Surgery, and Public Health 2023;1:100013. [doi: 10.1016/j.glmedi.2023.100013]
5. Somani S, Jain SS, Sarraju A, Sandhu AT, Hernandez-Boussard T, Rodriguez F. Using large language models to assess public perceptions around glucagon-like peptide-1 receptor agonists on social media. Commun Med (Lond) 2024 Jul 10;4(1):137. [doi: 10.1038/s43856-024-00566-z] [Medline: 38987347]
6. Sobel SI, Dubé JJ. Social media for obesity education: a general overview for the novice creator. Adv Physiol Educ 2024 Mar 1;48(1):114-121. [doi: 10.1152/advan.00120.2023] [Medline: 38205520]
7. Auerbach N, Liu VN, Huang DR, Clift AK, Al-Ammouri M, El-Osta A. What are community perspectives and experiences around GLP-1 receptor agonist medications for weight loss? A cross-sectional survey study in the UK. BMJ Public Health 2025;3(2):e002519. [doi: 10.1136/bmjph-2024-002519] [Medline: 40734969]

XSL•FO
RenderX

8. Azizi Z, Rodriguez F, Assimes TL. Digital footprints of obesity treatment: GLP-1 receptor agonists and the health equity divide. Circulation 2024 Jul 16;150(3):171-173. [doi: 10.1161/CIRCULATIONAHA.124.069680] [Medline: 39008562]
9. Propfe LE, Seifert R. Misrepresentation of semaglutide in social media. Naunyn Schmiedebergs Arch Pharmacol 2025 Jul 19. [doi: 10.1007/s00210-025-04403-5] [Medline: 40682686]
10. Mendel T, Singh N, Mann DM, Wiesenfeld B, Nov O. Laypeople's use of and attitudes toward large language models and search engines for health queries: survey study. J Med Internet Res 2025 Feb 13;27:e64290. [doi: 10.2196/64290] [Medline: 39946180]
11. Teo ZL, Thirunavukarasu AJ, Elangovan K, et al. Generative artificial intelligence in medicine. Nat Med 2025 Oct;31(10):3270-3282. [doi: 10.1038/s41591-025-03983-2] [Medline: 41053447]
12. Ponzo V, Goitre I, Favaro E, et al. Is ChatGPT an effective tool for providing dietary advice? Nutrients 2024 Feb 6;16(4):469. [doi: 10.3390/nu16040469] [Medline: 38398794]
13. Samaan JS, Yeo YH, Rajeev N, et al. Assessing the accuracy of responses by the language model ChatGPT to questions regarding bariatric surgery. Obes Surg 2023 Jun;33(6):1790-1796. [doi: 10.1007/s11695-023-06603-5] [Medline: 37106269]
14. Karnan N, Nair S, Fidai FF, Gurrala SV, Salim J, Gomma A. Evaluating the Efficacy of ChatGPT vs. Google Gemini in generating patient education materials for GLP-1 receptor agonists (semaglutide, liraglutide, tirzepatide): a cross-sectional study. Cureus 2025 Apr;17(4):e81993. [doi: 10.7759/cureus.81993] [Medline: 40351930]
15. Johnson H, Huang D, Liu V, Ammouri MA, Jacobs C, El-Osta A. Impact of digital engagement on weight loss outcomes in obesity management among individuals using GLP-1 and dual GLP-1/GIP receptor agonist therapy: retrospective cohort service evaluation study. J Med Internet Res 2025 Mar 31;27:e69466. [doi: 10.2196/69466] [Medline: 40164173]
16. Caballero Mateos I, Morales Portillo C, Lainez López M, Vilches-Arenas Á. Efficacy of a digital educational intervention for patients with type 2 diabetes mellitus: multicenter, randomized, prospective, 6-month follow-up study. J Med Internet Res 2025 Apr 10;27:e60758. [doi: 10.2196/60758] [Medline: 40209213]
17. Gibson D, Jackson S, Shanmugasundaram R, et al. Evaluating the efficacy of ChatGPT as a patient education tool in prostate cancer: multimetric assessment. J Med Internet Res 2024 Aug 14;26:e55939. [doi: 10.2196/55939] [Medline: 39141904]
18. Muraira PM, Chen A, Lerner LB, et al. Limitations of ChatGPT for patient education regarding frequently searched questions about benign prostatic hyperplasia. JU Open Plus 2024 Jul 14;2(7). [doi: 10.1097/JU9.0000000000000167]
19. Musheyev D, Pan A, Loeb S, Kabarriti AE. How well do artificial intelligence chatbots respond to the top search queries about urological malignancies? Eur Urol 2024 Jan;85(1):13-16. [doi: 10.1016/j.eururo.2023.07.004] [Medline: 37567827]
20. Özcan F, Örücü Atar M, Köroğlu Ö, Yılmaz B. Assessment of the reliability and usability of ChatGPT in response to spinal cord injury questions. J Spinal Cord Med 2025 Sep;48(5):852-857. [doi: 10.1080/10790268.2024.2361551] [Medline: 38860862]
21. Basics of Google Trends. Google News Initiative. URL: https://newsinitiative.withgoogle.com/en-gb/resources/trainings/basics-of-google-trends/ [accessed 2025-10-24]
22. Tan TF, Elangovan K, Ong JCL, et al. A proposed S.C.O.R.E. evaluation framework for large language models – safety, consensus & context, objectivity, reproducibility and explainability. arXiv. Preprint posted online on Jul 10, 2024. [doi: 10.48550/arXiv.2407.07666]
23. Human Biomedical Research Act 2015. Singapore Statutes Online. 2016 Jul 1. URL: https://sso.agc.gov.sg/Act/HBRA2015#al [accessed 2025-11-19]
24. The CHART Collaborative. Reporting guideline for chatbot health advice studies: the Chatbot Assessment Reporting Tool (CHART) statement. bmjmed 2025 Aug;4(1):e001632. [doi: 10.1136/bmjmed-2025-001632]
25. Puhl RM. Weight stigma and barriers to effective obesity care. Gastroenterol Clin North Am 2023 Jun;52(2):417-428. [doi: 10.1016/j.gtc.2023.02.002] [Medline: 37197883]
26. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J 2019 Jun;6(2):94-98. [doi: 10.7861/futurehosp.6-2-94] [Medline: 31363513]
27. Sharma S, Pajai S, Prasad R, et al. A critical review of ChatGPT as a potential substitute for diabetes educators. Cureus 2023. [doi: 10.7759/cureus.38380]
28. Gao Y, Xiong Y, Gao X, et al. Retrieval-augmented generation for large language models: a survey. arXiv. Preprint posted online on Dec 18, 2023 URL: https://arxiv.org/abs/2312.10997 [accessed 2025-05-29]

Abbreviations

API: application programming interface
CHART: Chatbot Assessment Reporting Tool
GLP1RA: glucagon-like peptide-1 receptor agonist
LLM: large language model
SCORE: safety, consensus, objectivity, reproducibility, relevance, and explainability

XSL•FO
RenderX

Edited by S Brini; submitted 29.05.25; peer-reviewed by D Vodela, RT Potla; revised version received 03.11.25; accepted 06.11.25; published 24.11.25.

Please cite as:
Tan SYT, Sng GGR, Lee PC
Accuracy of Large Language Model Responses Versus Internet Searches for Common Questions About Glucagon-Like Peptide-1 Receptor Agonist Therapy: Exploratory Simulation Study
JMIR Form Res 2025;9:e78289
URL: https://formative.jmir.org/2025/1/e78289
doi:10.2196/78289

© Sarah Ying Tse Tan, Gerald Gui Ren Sng, Phong Ching Lee. Originally published in JMIR Formative Research (https://formative.jmir.org), 24.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Usability, Benefits, and Barriers Associated With Patients' Access to Electronic Health Record–Integrated Telehealth in Hospitals in Riyadh: Qualitative Study

Dalia Alomar[1], BSc, MSc; Iliada Eleftheriou[1], MSc, PhD; Pauline Whelan[1], MSc, PhD; John Ainsworth[1,2], MSc, PhD

[1]Division of Informatics, Imaging and Data Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Vaughan House, Portsmouth St, Manchester, United Kingdom

[2]NIHR Manchester Biomedical Research Centre, Manchester University Hospitals NHS Foundation Trust, Manchester Academic Health Science Centre, Manchester, United Kingdom

Corresponding Author:
Dalia Alomar, BSc, MSc
Division of Informatics, Imaging and Data Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Vaughan House, Portsmouth St, Manchester, United Kingdom

Abstract

Background: The integration of electronic health records (EHRs) with telehealth platforms represents a transformative approach in health care, providing critical accessibility and engagement solutions, especially during the COVID-19 pandemic. In Riyadh's hospitals, the adoption of EHR-integrated telehealth has significantly increased and offers enhanced patient care options. However, there is a need to examine its continued relevance, effectiveness, and challenges in a postpandemic context.

Objective: This research aimed to qualitatively investigate the usability, perceived benefits, and barriers to patients' access to EHR-integrated telehealth from both patients and health care providers (HCPs) in a major Riyadh hospital.

Methods: A qualitative research design was used, featuring semistructured interviews with 20 patients and 10 HCPs, selected through purposive sampling for their direct experience with EHR-integrated telehealth services at Sulaiman Al Habib Hospital in Riyadh. Thematic analysis, supported by NVivo 14 software, was used to analyze the transcriptions and extract themes related to usability, perceived benefits, and barriers.

Results: The findings indicate that patients generally regard EHR-integrated telehealth positively, appreciating its navigability, convenience, and facilitation of remote health care interactions. Reported benefits included reduced physical visits, time savings, and more accessible follow-ups, contributing to greater continuity of care. However, significant barriers were identified, including technical challenges, lack of integration across hospital branches, absence of insurance payment linkages, and limited patient choice among providers. HCPs also expressed concerns over digital literacy gaps, the platform's limitations for specialized and complex care, and technical disruptions impacting care delivery.

Conclusions: EHR-integrated telehealth offers substantial potential to improve health care delivery in Riyadh's hospitals by enhancing access, convenience, and patient engagement. However, maximizing these benefits in Saudi Arabia's evolving health care landscape requires addressing identified barriers, particularly in platform stability, interbranch integration, insurance linkages, and patient support resources. Findings are grounded in a single-hospital sample and are intended to inform improvements in similar hospital settings in Saudi Arabia rather than national generalization.

(JMIR Form Res 2025;9:e74011) doi:10.2196/74011

KEYWORDS

EHR-integrated telehealth; electronic health record; usability; patient engagement; remote health care; Riyadh hospitals; health care professionals; telehealth barriers; qualitative research

Introduction

Background

The COVID-19 pandemic fundamentally transformed health care delivery, prompting rapid adoption of telehealth worldwide. In the United States, telehealth usage surged by 600% during the early months of the pandemic, reaching nearly 17,000 visits [1]. One health care organization reported that by transitioning 87% of its appointments to telehealth within 3 weeks, clinical volume decreased by only 15% [2]. Similarly, data from the Centers for Medicare and Medicaid Services indicated a rise in weekly telehealth visits from 13,000 before the pandemic to 1.7 million in April 2020 [3]. Compared with 2019, telehealth visits increased by more than 3000% by October 2020 [4]. Patients

XSL•FO
RenderX

reported improved access and convenience, whereas providers emphasized time efficiency and enhanced communication [5].

Among the approaches that emerged during this period was the integration of electronic health records (EHRs) with telehealth platforms. EHR-integrated telehealth combines digital consultation capabilities with EHR functionalities, such as appointment scheduling, health record access, and test result management [6]. This integration is considered transformative, enabling streamlined access to clinical data, continuity of care, and improved patient–provider communication. Reported advantages include reducing travel time and costs, improving chronic disease follow-up, and extending care to underserved populations [7-9]. Globally, studies confirm that telehealth integrated with EHRs can improve satisfaction, care coordination, and efficiency [10-15]. For example, in Germany, integration reduced waiting times and facilitated chronic disease management, whereas telemonitoring has demonstrated benefits for conditions such as hypertension [16]. Nonetheless, persistent barriers remain, including disparities in access to technology, variable digital literacy, and privacy concerns.

In Saudi Arabia, pushing toward Vision 2030 digital health goals, telehealth adoption expanded rapidly during the pandemic and continues to play a central role in the country's digital health transformation [17,18]. Leading hospitals in Riyadh have implemented EHR-integrated telehealth platforms, but existing Saudi research has primarily examined telehealth or EHRs in isolation, without integration [17,18]. In addition, limited research has simultaneously explored the perspectives of both patients and health care providers (HCPs) or examined operational barriers such as insurance integration, interoperability between hospital branches, and regulatory gaps. Given Saudi Arabia's unique health care infrastructure, cultural context, and reliance on insurance coverage, these factors may influence adoption differently than in Western contexts. Addressing this research gap is crucial to inform sustainable, equitable digital health models both within Saudi Arabia and in comparable Gulf health systems. In addition, the focus on end-user experiences is particularly important. Patients provide insight into accessibility, convenience, and barriers to engagement, whereas HCPs contribute perspectives on workflow integration, service delivery, and system limitations. Exploring these experiences together offers a balanced understanding of how EHR-integrated telehealth is currently functioning in practice and where targeted improvements are most needed. In addition, this study emphasizes Riyadh's role as a hub for advanced digital health initiatives, making it an ideal setting to generate insights with potential policy relevance across the Kingdom.

Conceptually, this study is guided by the Technology Acceptance Model (TAM), which posits that perceived ease of use and perceived usefulness are key determinants of EHR-integrated telehealth technology adoption [5], and by the principles of person-centered care, which emphasize tailoring health care services to patient needs and engagement. These frameworks informed the development of the research questions and findings about usability, perceived benefits, and barriers to EHR-integrated telehealth adoption in Saudi Arabia, as well as how it supports patient engagement and provider workflows.

Objective and Research Questions

This study aimed to fill this research gap by qualitatively exploring the experiences of patients' access to EHR-integrated telehealth from the perspectives of both patients and HCPs in Riyadh. Specifically, it investigated the usability of these services, the perceived benefits, and the challenges faced by patients. By addressing these aspects, the study sought to provide insights that will inform the optimization of telehealth services in Riyadh's health care institutions and contribute to improving the accessibility and quality of health care delivery in the region. To achieve the study objectives, the following research questions were raised:

1. How do patients and HCPs perceive the usability of the EHR-integrated telehealth platform?
2. What benefits are reported in terms of health care accessibility, communication, and management?
3. What technical, organizational, and regulatory barriers constrain effective use of these services?
4. What improvements do patients and HCPs recommend to optimize EHR-integrated telehealth in Riyadh's hospitals?

Methods

Study Design

This study used a qualitative research design, situated within a constructivist paradigm, to explore the usability, benefits, and challenges associated with patients' access to EHR-integrated telehealth services in Riyadh hospitals. A qualitative approach was chosen for its capacity to generate rich, in-depth insights into the subjective experiences and perceptions of both patients and HCPs.

The first investigator (DA) collected data from November 2024 to February 2025 through semistructured, in-depth face-to-face interviews, which were conducted inside the Sulaiman Al Habib Hospital. While electronic interviews via the secure hospital's telehealth platforms were available as part of the study context, all participants preferred in-person interviews, which provided opportunities for real-time clarification and deeper discussion of emerging themes. Interviews lasted between 50 and 70 minutes (mean 60, SD 7.8 min), allowing sufficient time for participants to reflect on their experiences and perceptions in detail. The semistructured format ensured consistency across participants while allowing for flexibility in questioning, enabling the researcher to probe further into emerging themes. The lead investigator (DA) maintained a reflexive stance throughout, considering her dual position as a health informatics researcher and health care professional, and regularly documented reflections in a field journal to mitigate potential biases.

Research Setting

The study was conducted at Dr. Sulaiman Al Habib Hospital, a leading private health care group in Riyadh, known for its early and extensive adoption of telehealth services integrated within its EHR system, making it a suitable setting for this study. The hospital's telehealth platform, the Al Habib App, provides patients with functionalities such as virtual consultations, appointment scheduling, prescription renewal, laboratory result

XSL•FO
RenderX

viewing, and secure messaging. For HCPs, the system supports integrated documentation, access to patient medical records, and coordination across departments.

Recruitment

Participants were recruited from Sulaiman Al Habib Hospital in Riyadh, which was selected because of its well-established and comprehensive EHR-integrated telehealth services. A purposive sampling strategy was used to capture insights from both patients and HCPs with direct experience using the platform. To ensure a range of perspectives, participants were selected based on their roles and engagement with telehealth services. The study included 10 HCPs from different specialties and departments and 20 patients with varying medical conditions and telehealth usage histories, ensuring the inclusion of individuals who interacted with the system in different capacities to provide a comprehensive understanding of its implementation and impact. Participant recruitment continued until thematic saturation was reached [19], where no new themes or codes emerged, indicating that the dataset sufficiently captured participants' experiences. Prior literature suggests that between 20 and 30 interviews typically capture sufficient thematic variability for studies of this nature [19].

Eligible patient participants included individuals aged ≥18 years of all genders, medical conditions, and socioeconomic backgrounds and who have used EHR-integrated telehealth services at least once in the past year, ensuring their relevant experience with these services. Eligible HCP participants were selected to represent a range of specialties (general practice, obstetrics or gynecology, ophthalmology, eHealth and telehealth, and dentistry) and roles, including physicians, nurses, and administrative staff involved in telehealth for at least 6 months to ensure their sufficient experience in using it. The recruitment process involved direct invitations extended to HCPs through department heads and professional networks, while patients were contacted based on hospital telehealth records, as well as through recruitment advertisements or brochures placed in hospital waiting areas. This strategy ensured a diverse and relevant sample of participants with varying levels of interaction with EHR-integrated telehealth, capturing a broad spectrum of experiences regarding its usability and impact. The research team acknowledges that this approach may have excluded less-engaged or digitally disadvantaged individuals. This limitation was partly mitigated by including patients with varied levels of telehealth familiarity and by encouraging referrals from clinicians to reach less active users. Before the start of any interviews, individual interviews with participants were implemented in the hospital to answer any participant questions.

Interview Guide

The interview guide for this study was structured to capture in-depth insights from both patients and HCPs regarding EHR-integrated telehealth services in Riyadh's hospitals. It was piloted with 2 participants (ie, 1 patient and 1 HCP) to ensure clarity, relevance, and sequencing of questions. Minor refinements were made based on feedback before commencing full data collection. The final guide included questions that covered 6 key areas: (1) introduction and background information, where patients shared their demographics and medical conditions, while HCPs shared their specific role-related context; (2) usability, exploring experiences with platform navigation, appointment scheduling, and communication tools; (3) perceived benefits, capturing positive impacts on health care management, accessibility, and patient engagement; (4) perceived barriers and challenges, identifying obstacles such as technical issues, privacy concerns, and digital literacy limitations; (5) general experience and suggestions, which allowed participants to reflect on their satisfaction and provide recommendations for improvement; and (6) contextual factors, examining the influence of socioeconomic, organizational, and regulatory factors on telehealth access and use. This semistructured design ensured consistency across interviews while allowing flexibility for follow-up questions, supporting a comprehensive understanding of EHR-integrated telehealth's impact on health care delivery in this context. Multimedia Appendix 1 shows questions asked to patients and HCP participants in the interviews. The questions were developed in English, translated into Arabic, and reviewed by an independent bilingual health informatics specialist to ensure both linguistic accuracy and conceptual equivalence. This ensured that all questions were comprehensible to participants while minimizing translation bias.

Qualitative Analysis

The qualitative analysis followed a systematic process to identify key themes related to EHR-integrated telehealth through comprehensive coding and thematic content analysis. Interviews were audio recorded and transcribed verbatim. All transcripts and audio files were translated from Arabic into English by a professional translator. The average transcript length was approximately 14 to 16 pages (mean 15, SD 2 pages). Data analysis was carried out using Qualitative Research Software (QSR International NVivo software version 14), following Braun and Clarke's thematic content analysis approach [20].

Two investigators (DA and JA) conducted a thematic content analysis of the interview transcripts to identify emerging patterns. After independently coding the first 5 interviews, an initial codebook has been developed. This preliminary codebook was then used to recode the first 5 interviews and code all remaining interviews, with periodic cross-checks and regular meetings held between research team members to discuss and resolve any coding discrepancies and refine the codebook to integrate new emerging themes, ensure consistency, and maintain reliability. Triangulation was achieved through comparing perspectives of patients and HCPs, and reflexivity was maintained through the investigators' audit trail and reflexive notes documenting their positionality and assumptions during analysis.

Study design and interview analysis were reported using the Consolidated Criteria for Reporting Qualitative Studies 32-item checklist (Checklist 1) [21].

In this study, rigor was ensured through multiple strategies, including intercoder reliability, triangulation of patient and HCP perspectives, and systematic use of NVivo to manage data. Reflexivity was addressed by acknowledging that the first investigator (DA), as a health informatics researcher with prior experience in telehealth evaluation, has shaped the data

XSL•FO
RenderX

collection and interpretation. Reflexive notes were maintained to critically reflect on positionality, interview dynamics, and potential biases. Translation was another important consideration. Although all interviews were conducted in Arabic, the analysis was carried out on English translations. To mitigate risks of meaning loss and enhance data quality, after all interviews were professionally translated into English by an independent translator, these translations were reviewed by the first investigator and a bilingual health care researcher familiar with the Saudi context.

Ethical Considerations

Ethics approval was secured from the University of Manchester Research Ethics Committee 2 (reference 2024-20964-36807), along with institutional review board approval from the Sulaiman Al Habib Hospital administration for the procedures, participant recruitment, and interview questions. Verbal informed consent was obtained before the start of any interview, per the guidelines of the University of Manchester Research Ethics Committee 2. All transcripts and audio files were stored without personal identifiers on a secure, password-protected server to ensure participant confidentiality. No compensations were provided to participants for their participations in this study.

Results

Participants

In total, 30 participants (ie, 20 patients and 10 HCPs) have been interviewed. Regarding patient demographics, as presented in Table 1, 40% (8/20) were aged between 20 and 30 years, 50% (10/20) were aged between 31 and 40 years, and 10% (2/20) were aged between 41 and 50 years. Gender distribution was balanced, with 50% (10/20) males and 50% (10/20) females. In terms of medical conditions, 85% (17/20) of patients reported being in good health with no chronic diseases, whereas 15% (3/20) reported having chronic conditions, such as diabetes or high blood pressure. For the duration of EHR-integrated telehealth usage, 25% (5/20) of patients had used the service for less than a year, 45% (9/20) had used the service for 1 to 2 years, and 30% (6/20) had used the service for 3 to 5 years. In 2023, 35% of patients (7/20) did not use EHR-integrated telehealth services, 45% (9/20) used them 1 to 4 times, 15% (3/20) used them 5 to 10 times, and 5% (1/20) used them 11 to 12 times.

XSL•FO
RenderX

Table . Participant demographics.

Characteristics			Value, n (%)
Patients (n=20)			
	Age (y)		
		20 - 30	8 (40)
		31 - 40	10 (50)
		41 - 50	2 (10)
	Sex		
		Male	10 (50)
		Female	10 (50)
	Medical condition		
		Good (ie, no disease)	17 (85)
		Chronic disease (ie, diabetes, high blood pressure)	3 (15)
	Duration of EHR[a]-integrated telehealth utilization (y)		
		<1	5 (25)
		1 - 2	9 (45)
		3 - 5	6 (30)
	Number of times EHR-integrated telehealth was used in the past year (2023)		
		None	7 (35)
		1 - 4	9 (45)
		5 - 10	3 (15)
		11 - 12	1 (5)
HCPs[b] (n=10)			
	Position and role		
		General practitioner: provides in-person and remote consultations, diagnoses medical conditions, manages chronic disease follow-up, prescribes medications, and coordinates tests and treatment plans	4 (40)
		Consultant obstetrician/gynecologist: manages women's health, including diagnosis, treatment, pregnancy follow-up, and performing childbirth and gynecological surgeries	2 (20)
		Consultant ophthalmologist: specializes in diagnosing and treating eye diseases and conditions	1 (10)
		Doctor at eHealth and telehealth department: delivers remote consultations, diagnoses, prescribes treatments, manages lab and radiology orders, and coordinates with multidisciplinary teams for comprehensive patient care	2 (20)
		Dentist: handles inquiries and provides dental consultations through online medical services	1 (10)
	Duration of use of EHR-integrated telehealth with your patients (y)		

XSL•FO
RenderX

Characteristics		Value, n (%)
	<1	1 (10)
	1 - 3	2 (20)
	≥4 (since COVID-19)	7 (70)

[a]EHR: electronic health record.

[b]HCP: health care provider.

For HCP demographics, participants were distributed across several positions and related roles: 4 (40%) were general practitioners who provided both in-person and remote consultations, managed chronic disease follow-ups, prescribed medications, and coordinated patient testing and treatment plans. Two participants (20%) were consultant obstetricians/gynecologists who specialized in women's health, including diagnosis, treatment, and pregnancy follow-up, and conducting childbirth and other gynecological surgeries. One participant (10%) was a consultant ophthalmologist focused on treating eye diseases, whereas 2 participants (20%) worked in the eHealth and telehealth department, providing remote consultations, managing laboratory and radiology requests, and coordinating care with multidisciplinary teams. Finally, 1 participant (10%) was a dentist handling inquiries and providing dental consultations through online services. In terms of their experience with EHR-integrated telehealth with their patients, 1 HCP (10%) had used it for less than a year, 2 (20%) had used it for 1 to 3 years, and 7 (70%) had been using it for 4 years or more since the COVID-19 pandemic.

Participants varied in age and telehealth experience. While formal education or household income measures were not systematically collected, multiple patients commented on socioeconomic and literacy factors that influenced telehealth experiences. Most patients reported that cost did not prevent their use of the app, but a minority raised affordability concerns (eg, "the prices of services are not cheap"). Several respondents also noted that older adults or those unfamiliar with apps required assistance or found navigation harder (eg, "some older people may have difficulty using techniques"). Internet reliability and basic technical skills emerged repeatedly as determinants of effective use.

The analysis ultimately organized findings around 4 major themes, from perspectives of both patients and HCPs, related to EHR-integrated telehealth, including usability, perceived benefits, barriers and challenges, and general experiences and suggestions for improvements. Table 2 summarizes the number of coded references assigned to each major theme across all interviews. These frequency counts reflect how often ideas were mentioned in the dataset (salience of issues in the corpus) rather than the number of participants who endorsed a theme. The analytic approach combined iterative codebook development with dual coding. Recruitment continued until the 26th interview, where no novel codes were identified across consecutive interviews and the codebook stabilized [19], indicating that the dataset sufficiently captured participants' experiences. The final 4 interviews yielded only confirmatory evidence of existing themes. Both patient and HCP subgroups reached saturation independently, ensuring coverage across perspectives.

Table . Frequency of identified themes related to access and utilization of EHR[a]-integrated telehealth.

Theme	Frequency values[b], n	
	Patients	HCPs[c]
Usability	80	42
Perceived benefits	117	80
Barriers to and challenges of access and utilization	93	47
General experiences and suggestions for improvements	69	57

[a]EHR: electronic health record.

[b]Frequency was calculated as the total number of references related to the group of codes of each theme by interviewees.

[c]HCP: health care provider.

Perspectives on Usability of EHR-Integrated Telehealth

Overview

In the discussion of perspectives of both patients and HCPs about the patients' experiences and usability of telehealth features integrated within the EHRs framework in Sulaiman Al Habib Hospital, interviewees discussed several ways in which this platform could form this usability, including (1) user interaction and experience and (2) technical performance and communication issues.

User Interaction and Experience

Across patient interviews, the Al Habib platform was broadly perceived as easy to access and navigate for routine tasks (eg, booking, viewing records, and prescription refills). Patients frequently described straightforward appointment booking and

XSL•FO
RenderX

ready access to prior visit notes and test results, emphasizing time savings and convenience. For example, 1 patient said:

> *Very good. Easy access to medical information such as laboratory scans, future appointments, bookings, and medical statements. The information is available at all times and has easy access.*

HCPs' Perspective and Cross-Group Comparison

HCPs generally echoed patients' reports about the platform's accessibility and functionality for routine care, noting that the app supports efficient appointment management and medication refills. However, clinicians emphasized the platform's limitations for complex diagnostic work and specialist assessment, which are areas where remote workflows were felt to be less appropriate. One HCP remarked:

> *In general, the services available through the Al Habib application meet the needs of patients, especially those who have minor cases or need periodic follow-up. However, there are some patients who try to use the service to diagnose complex conditions, which may need a physical examination, which is difficult to achieve via remote service. Some patients need specialized matters, which cannot be answered through Life Care app.* [HCP 3]

Overall, patients and HCPs agreed that the app is well suited for minor conditions and follow-up but less so for complex or examination-dependent cases.

Technical Performance and Communication

Communication features (notably video) were valued when they worked, but instability (ie, poor video/audio quality, dropped calls, or the need to switch to telephone) was a frequent barrier that reduced clinical usefulness for some encounters. One patient described intermittent failures:

> *The video calls often freeze, lose sound, or the camera not working properly. Then, it's hard to explain my symptoms when the connection is bad, forcing the doctor to call me on the phone instead.*

HCPs similarly reported that poor video quality prevented adequate visual assessment in some cases.

Perceived Benefits

Overview

When interviewees were asked about the perceived benefits of EHR-integrated telehealth for healthcare delivery, patient care, and access to health care services, they consistently emphasized three domains of benefit from EHR-integrated telehealth: (1) accessibility and convenience, (2) healthcare management and support, and (3) patient satisfaction.

Accessibility and Convenience

Patients reported that telehealth saved significant time, effort, and transportation costs, particularly for routine check-ups, prescription refills, or minor concerns. They valued the ability to consult from home or work without disrupting daily commitments or arranging transportation. One patient explained:

> *Comfort and saving time and effort. I do not need to get out of the house... including the access to medical consultations and prescriptions, and delivering medicines to the patient.*

HCPs echoed these views, noting that remote consultations reduced unnecessary hospital visits and allowed patients to access care from home or workplaces. As one HCP observed:

> *The app makes patients able to follow their health condition and conduct consultations from anywhere, which offers great convenience.*

Both groups agreed that convenience was one of telehealth's strongest advantages, especially for patients with mobility challenges or those living farther from the hospital.

Health Care Management and Support

Patients highlighted how access to medical records, test results, and treatment histories improved their ability to monitor health conditions and adhere to care plans. For example, one patient stated:

> *Telehealth has helped me compare old and new medical tests, making it easier to track the progress of my health condition and better manage treatment.*

HCPs similarly noted that telehealth supported adherence and continuity of care, especially for patients with chronic diseases. One HCP explained:

> *The service enhanced patients' understanding of their health condition and increased their commitment to treatment and follow-up.*

Together, these perspectives suggest that the platform encouraged more active and sustained patient engagement in managing health.

Patient Satisfaction

Patients expressed high satisfaction with telehealth, attributing it to ease of access, improved communication, and the ability to manage health care needs without frequent in-person visits. One patient shared:

> *Using telehealth has increased my satisfaction because I feel I have easy access to healthcare whenever I need it.*

HCPs reported observing similar satisfaction among their patients, pointing to reduced travel and positive feedback through reviews and interactions. As one HCP put it:

> *Patients are very satisfied with the service; it has provided them with comfort and reduced unnecessary hospital visits.*

Overall, satisfaction for both patients and HCPs stemmed from a combination of convenience, improved health management, and more responsive communication with providers.

Barriers and Challenges

Overview

When patients and HCP interviewees were asked about barriers and challenges in patients' access to and utilization of EHR-based telehealth, they identified multiple barriers, grouped

XSL•FO
RenderX

into six main domains including technical and functional barriers, health care service delivery issues, integration and coordination problems, prescription, medication, and analyses results challenges, patient privacy and security concerns, and contextual, organizational, or regulatory factors.

Technical and Functional Barriers

Patients frequently described frustrations with appointment booking, long waiting times, and poor video quality, often linked to internet instability or lack of timely technical support. These issues disrupted consultations and delayed access to care, especially for older adults less comfortable with digital tools. As one patient shared:

> *The video calls often freeze, lose sound, or the camera not working properly. Then, it's hard to explain my symptoms when the connection is bad.*

HCPs similarly reported that connectivity problems and patients' limited technical skills complicated consultations. One HCP noted:

> *Patients with limited tech skills struggle to navigate the app, which adds to our workload as we need to guide them through basic steps.*

Together, these findings from both patients and HCPs perspectives underscore the importance of reliable infrastructure and digital literacy in supporting telehealth usability.

Health Care Service Delivery Issues

Patients raised concerns about limited doctor availability, lack of specialist consultations, and the inability to select a preferred physician for continuity of care. Some also noted frustration at not being able to secure sick leave documentation via telehealth. One patient explained:

> *I would prefer to consult with the same doctor each time, but the app doesn't give me that choice.*

HCPs echoed these issues, highlighting that certain specialties (eg, gynecology, dermatology) are poorly suited for remote care and that patients' dissatisfaction with sick leave requests often drove them back to in-person visits. As one HCP stated:

> *Telehealth doesn't suit all specialties; for example, physical assessments in areas like gynecology are hard to examine remotely* [HCP 2]

Both perspectives highlight the limits of telehealth as a universal substitute, emphasizing the need for targeted use depending on condition and specialty.

Integration and Coordination Problems

Overview

Patients reported inefficiencies due to poor coordination between hospital branches, lack of insurance payment links, and weak integration with pharmacies, which often forced them to repeat tests or pay out-of-pocket. One patient described:

> *...only the problem is not linking the payment for the virtual visit to the insurance...the patient may refrain from remote consultation completely...*

HCPs confirmed that the absence of insurance integration and pharmacy linkage undermined continuity of care and patient satisfaction. One HCP explained:

> *Health insurance does not cover the virtual consultation, laboratory tests, and basic services without additional approvals from insurance.*

Both groups pointed to system-level gaps that reduce efficiency and discourage ongoing telehealth use.

Prescription, medication, and results challenges

Patients noted delays in accessing lab results and inconsistencies in prescription transmission to pharmacies, which often required extra calls or visits. One patient reported:

> *Sometimes the doctor writes the prescription but it does not get to the pharmacy properly, forcing me to repeat the operation or visit the hospital in person.*

HCPs added that the lack of effective triage systems within the platform complicated workflow, as urgent and non-urgent cases were not differentiated. One HCP remarked:

> *One of the main challenges is the lack of an effective triage system to identify patients who need specialized or general consultation.*

Both groups show how incomplete integration impedes timely follow-up and efficient care delivery.

Patient Privacy and Security Concerns

Overview

Interestingly, most patients expressed confidence in the platform's security, reporting no major worries about privacy breaches. One patient stated:

> *No, I don't have any concerns about the privacy or security of my medical information when using the platform.*

HCPs likewise confirmed that privacy concerns were rarely raised. One explained:

> *I did not notice any concern among patients about the privacy of their medical information. They are all confident that their data is secure.*

This alignment suggests that strong institutional reputation may build trust, even in digital health adoption.

Contextual, Organizational, and Regulatory Factors

Patients generally felt that socio-economic background did not limit access, though insurance requirements and regulatory gaps were seen as barriers. One patient reflected:

> *I think the app is available to everyone and there are no social or economic barriers to its use... perhaps medical insurance, as sometimes it requires communication with the insurance provider.*

HCPs, however, emphasized organizational and regulatory shortcomings, including the exclusion of patients under 17 years old. As one HCP noted:

XSL•FO
RenderX

> *"The application does not serve patients under 17, which may constitute a barrier for some mothers for their young children or infants.*

Together, these perspectives illustrate how broader structural factors constrain the inclusivity and long-term sustainability of telehealth.

General Experiences and Suggestions for Improvements

Thematic analysis provided some other findings related to patients' general experience and their suggestions for improvement as a result of their access and utilization of EHR-based telehealth. Patients generally reported positive experiences with the EHR-integrated telehealth platform, valuing its convenience, timeliness, and efficiency. Satisfaction was high, with many highlighting quick access to prescriptions and consultations as key benefits. One patient noted:

> *I am very satisfied with how quickly I can get my prescriptions and consult with my doctor through the app.*

At the same time, some patients emphasized the need for ongoing improvements, particularly in technical support, customer service, and expanded service offerings (eg, sick leave and specialist access), stating that:

> *The application needs to be improved in technical aspects, such as full support for video calls and chat to include photos to be considered a real clinic.*

As another patient explained:

> *Overall, while satisfaction was high, patients balanced praise with constructive feedback, demonstrating both appreciation and expectation for continued platform development.*

Looking ahead, patients expressed strong willingness to continue using telehealth services in the future, especially for routine consultations and follow-ups. This preference was linked to perceived convenience and time savings. One patient summarized:

> *I would definitely prefer using the app for future appointments; it saves me so much time and facilitates access to health services.*

This enthusiasm suggests that telehealth is viewed not only as a temporary solution but as a sustainable option for ongoing health care management.

From the perspective of HCPs, the platform was also seen as beneficial in improving patients' access and convenience. HCPs confirmed high levels of patient satisfaction but pointed to persistent challenges, such as technical glitches, difficulties managing complex cases remotely, and limited organizational support. One HCP remarked:

> *Telehealth has improved access for patients, but there are limitations in handling complex cases that really need in-person consultations.*

In addition, HCPs offered several suggestions for improvement, including stronger technical infrastructure, an effective triage system, expanded record integration across branches, and more targeted training for both staff and patients. As one HCP explained:

> *Both staff and patients would benefit from additional training to enhance telehealth literacy and platform usability. Establishing a dedicated telehealth team focused solely on app-based patients could reduce wait times and improve care quality.*

Finally, HCPs raised concerns about the regulatory and insurance frameworks, noting that clearer guidelines and broader coverage would enhance compliance, protect patient data, and ensure equitable access. One physician emphasized:

> *It is important to improve policies related to insurance and expand its coverage of remote care services. Regarding sick leave, there must be organizational aspects to this in order to determine whether or not a patient is eligible.*

Taken together, patients and HCPs converged on the need for technical, organizational, and policy improvements to support sustainable, effective, and trusted telehealth delivery.

Discussion

Principal Findings

Using a qualitative analysis of semistructured interviews, this study explored the usability, perceived benefits, barriers, and overall experiences with EHR-integrated telehealth services in Riyadh's hospitals from both patient and HCP perspectives. Four key themes emerged: usability, perceived benefits, barriers and challenges, and general experiences and suggestions for improvement. Findings suggest that while EHR-integrated telehealth enhances accessibility and convenience, persistent challenges in technical functionality, service delivery, integration, and regulation limit its full potential. These insights contribute to understanding telehealth in the Saudi context, which has received less attention compared to Western settings.

The usability of the telehealth platform was highlighted as a key strength. Both patients and HCPs found the system intuitive and user-friendly, consistent with existing literature on telehealth adoption [2]. This supports the TAM, where perceived ease of use influences positive attitudes toward technology and sustained engagement [5]. At the same time, persistent technical issues (eg, video quality and connectivity) contradict TAM's expectation that ease of use consistently drives acceptance, showing that infrastructural weaknesses can undermine adoption even when the interface is simple. These contradictions highlight the need for stronger digital infrastructure in Saudi Arabia to fully realize TAM's potential pathways.

For perceived benefits, patients emphasized reduced travel time, improved flexibility, and timely access to care, whereas HCPs highlighted greater patient engagement and adherence. These findings are consistent with international evidence on telehealth's ability to reduce logistical burdens and support chronic disease management [22]. Importantly, telehealth's capacity to foster ongoing communication also aligns with person-centered care, which prioritizes continuity and patient engagement. However, not all findings were positive: for certain

XSL•FO
RenderX

specialties requiring physical examinations (eg, orthopedics and dermatology), telehealth fell short. This contradiction underscores that while telehealth enhances patient-centeredness for routine care, it may compromise quality in complex or specialist care contexts, requiring hybrid approaches.

Regarding barriers, service delivery, and contextual differences, barriers identified (eg, technical failures, limited availability of doctors, lack of specialist access, poor insurance integration, and regulatory gaps) mirror challenges reported in other Gulf settings [7,23]. However, they diverge from many Western contexts where insurance integration and regulatory frameworks are more advanced [17]. In relation to person-centered care, patients and HCPs also emphasized difficulties with continuity of care, such as inability to choose specific doctors and delays in result integration across hospital branches. These findings reinforce that in Saudi Arabia, organizational and policy-level issues, rather than patient trust or motivation, pose the greatest barriers to telehealth. Addressing these issues is vital for sustainable expansion.

Overall, both patients and HCPs reported high satisfaction with telehealth but stressed the importance of improvements. Suggested priorities included better technical support, more robust triage systems, training for digital literacy, and integration of insurance and medical records across hospital branches. These recommendations are consistent with global calls for continuous quality improvement in telehealth [11,24,25]. Importantly, the emphasis on insurance integration and regulatory reform reflects region-specific priorities not commonly observed in Western telehealth studies [26].

Importantly, while the findings illustrate important opportunities and challenges, they should not be interpreted as broadly representative of Saudi Arabia as a whole. The study focused on a single hospital with advanced EHR–telehealth integration, which may not reflect the full diversity of health care settings in the country. Nevertheless, the findings provide valuable insights into emerging trends and patient–provider perspectives in one of Saudi Arabia's leading hospitals, offering lessons that can inform national strategies for scaling telehealth.

Strengths and Limitations

This study's primary strength lies in its dual-perspective approach, capturing insights from both patients and HCPs to provide a balanced and comprehensive understanding of EHR-integrated telehealth's usability, benefits, and challenges. Conducted in the postpandemic period, the study is timely, reflecting the increased reliance on digital health care solutions in Saudi Arabia and offering findings highly relevant for shaping future health care strategies. Moreover, the study's focus on Saudi Arabia highlights specific regional needs, such as geographic accessibility and system integration, providing actionable insights tailored to the Saudi health care landscape and potentially influencing policy developments for digital health care services across the region.

However, this study has several limitations that should be acknowledged. First, its geographic scope was limited to a single hospital in Riyadh, which may reduce the transferability of findings across Saudi Arabia's diverse regions, where health care access, infrastructure, and digital adoption levels can vary. In addition, the purposive sample of 30 participants, while sufficient for achieving thematic saturation, cannot support broad generalizations, and the perspectives reported should be interpreted as illustrative rather than representative. Second, potential sources of bias must be considered. Social desirability bias may have led participants to emphasize positive experiences, while interviewer positionality as a health informatics researcher could have influenced the framing of questions or interpretation of responses. Translation of transcripts from Arabic to English, although reviewed by an independent translator, may also have introduced subtle nuances affecting meaning. Third, recruitment through hospital records and brochures may have disproportionately included patients who were already more engaged with health care services and excluded those less digitally literate or less inclined to access telehealth, thereby limiting the diversity of perspectives captured. In addition, while the qualitative methodology was valuable in generating rich, in-depth insights, it does not provide quantitative data needed to establish statistically significant trends in telehealth usage and satisfaction. As with any qualitative research, participants' willingness to share their experiences was influenced by comfort level, interview setting, and trust in the research process, which may have shaped the narratives obtained. Finally, the short-term nature of this study restricts insights into long-term patient engagement, sustained provider adaptation, and the effectiveness of proposed improvements. Taken together, these limitations highlight the importance of future studies with a broader scope, mixed-method approaches, and longitudinal designs to validate and extend these findings across different health care settings in Saudi Arabia.

Conclusions

This study offers an in-depth exploration of patient and HCP perspectives on EHR-integrated telehealth services in Riyadh, contributing context-specific insights to the broader global evidence base. While many of the reported benefits (such as enhanced accessibility, convenience, and patient engagement) resonate with international findings, this study underscores distinctive challenges in the Saudi context, including limited insurance integration, fragmented coordination across hospital branches, and regulatory gaps. These findings suggest that the value of telehealth in Saudi Arabia lies not in novelty, but in its potential to be tailored and scaled within the country's unique health care ecosystem.

Moving beyond technical fixes, several actionable pathways emerge. First, policy alignment is critical. Clear regulatory frameworks and mandatory insurance reimbursement for telehealth consultations would reduce financial and administrative barriers. The Ministry of Health also should issue standardized telehealth regulatory guidelines covering privacy, security, and sick leave certification, reducing inconsistencies across hospitals. It is also important to create direct digital payment links between telehealth platforms and insurance providers, allowing real-time claim approvals within the app rather than requiring manual follow-up. Second, organizational investment is needed to expand specialist availability for remote care, establish dedicated telehealth teams, and implement

XSL•FO
RenderX

structured triage protocols to prioritize urgent cases. Third, capacity building for both patients and HCPs through targeted digital literacy programs and staff training could improve navigation, communication, and trust in the platform. Fourth, technological integration should prioritize interoperability across hospital branches and pharmacies, ensuring continuity of care and seamless access to medical records and prescriptions.

Future research should evaluate hybrid care models that combine the convenience of telehealth with in-person consultations for conditions requiring physical examination, whereas longitudinal studies could clarify its impact on chronic disease management, patient outcomes, and sustained engagement. In addition, comparative research across Saudi regions would help determine how local cultural and infrastructural factors shape telehealth adoption. As Saudi Arabia advances its Vision 2030 digital health agenda, the lessons drawn from this study can inform the design of more inclusive, patient-centered, and policy-aligned telehealth solutions. By embedding telehealth within a robust regulatory, technological, and organizational framework, the country can move beyond short-term pandemic-driven adoption toward a sustainable model of digitally enabled health care.

Data Availability

The datasets generated and analyzed during this study are not publicly available due to confidentiality and privacy restrictions but are available from the corresponding author on reasonable request.

Conflicts of Interest

PW is a director and chief operating officer of CareLoop Health Ltd, a company that develops and markets digital therapeutics for severe mental illness. JA is a director and shareholder of CareLoop Health Ltd.

Multimedia Appendix 1
Interview questions.
[DOCX File, 21 KB - formative_v9i1e74011_app1.docx]

Checklist 1
COREQ 32-item checklist.
[PDF File, 429 KB - formative_v9i1e74011_app2.pdf]

References

1. Carlson JL, Goldstein R. Using the electronic health record to conduct adolescent telehealth visits in the time of COVID-19. J Adolesc Health 2020 Aug;67(2):157-158. [doi: 10.1016/j.jadohealth.2020.05.022] [Medline: 32517972]
2. Maheswaranathan M, Chu P, Johannemann A, Criscione-Schreiber L, Clowse M, Leverenz DL. The impact of the COVID-19 pandemic and telemedicine implementation on practice patterns and electronic health record utilization in an academic rheumatology practice. J Clin Rheumatol 2022 Mar 1;28(2):e612-e615. [doi: 10.1097/RHU.0000000000001751] [Medline: 33843775]
3. Koonin LM, Hoots B, Tsang CA, et al. Trends in the use of telehealth during the emergence of the COVID-19 pandemic - United States, January-March 2020. MMWR Morb Mortal Wkly Rep 2020 Oct 30;69(43):1595-1599. [doi: 10.15585/mmwr.mm6943a3] [Medline: 33119561]
4. Omboni S, Padwal RS, Alessa T, et al. The worldwide impact of telemedicine during COVID-19: current evidence and recommendations for the future. Connect Health 2022 Jan 4;1:7-35. [doi: 10.20517/ch.2021.03] [Medline: 35233563]
5. Ghaddar S, Vatcheva KP, Alvarado SG, Mykyta L. Understanding the intention to use telehealth services in underserved Hispanic border communities: cross-sectional study. J Med Internet Res 2020 Sep 3;22(9):e21012. [doi: 10.2196/21012] [Medline: 32880579]
6. Zhang X, Saltman R. Impact of electronic health record interoperability on telehealth service outcomes. JMIR Med Inform 2022 Jan 11;10(1):e31837. [doi: 10.2196/31837] [Medline: 34890347]
7. Lawson DW, Stolwyk RJ, Ponsford JL, Baker KS, Tran J, Wong D. Acceptability of telehealth in post-stroke memory rehabilitation: a qualitative analysis. Neuropsychol Rehabil 2022 Jan;32(1):1-21. [doi: 10.1080/09602011.2020.1792318] [Medline: 32677539]
8. Madu R. Using technology to reduce barriers in community health: an intersection between broadband access EHR & telehealth adoption in federally qualified health centers [PhD dissertation]. : UTHealth School of Public Health, The University of Texas; 2020 URL: https://digitalcommons.library.tmc.edu/uthsph_dissertsopen/127 [accessed 2025-11-02]
9. Wiley KK, Mendonca E, Blackburn J, Menachemi N, Groot MD, Vest JR. Quantifying electronic health record data quality in telehealth and office-based diabetes care. Appl Clin Inform 2022 Oct;13(5):1172-1180. [doi: 10.1055/s-0042-1758737] [Medline: 36516970]
10. McLendon SF. Interactive video telehealth models to improve access to diabetes specialty care and education in the rural setting: a systematic review. Diabetes Spectr 2017 May;30(2):124-136. [doi: 10.2337/ds16-0004] [Medline: 28588379]

XSL•FO
RenderX

11. Ashwood JS, Mehrotra A, Cowling D, Uscher-Pines L. Direct-to-consumer telehealth may increase access to care but does not decrease spending. Health Aff (Millwood) 2017 Mar 1;36(3):485-491. [doi: 10.1377/hlthaff.2016.1130] [Medline: 28264950]
12. Ndwabe H, Basu A, Mohammed J. Post pandemic analysis on comprehensive utilization of telehealth and telemedicine. Clinical eHealth 2024 Dec;7:5-14. [doi: 10.1016/j.ceh.2023.12.002]
13. Bitar H, Alismail S. The role of eHealth, telehealth, and telemedicine for chronic disease patients during COVID-19 pandemic: a rapid systematic review. Digit Health 2021;7:20552076211009396. [doi: 10.1177/20552076211009396] [Medline: 33959378]
14. Ben-Assuli O. Measuring the cost-effectiveness of using telehealth for diabetes management: a narrative review of methods and findings. Int J Med Inform 2022 Jul;163:104764. [doi: 10.1016/j.ijmedinf.2022.104764] [Medline: 35439671]
15. Cushing M. What is telemedicine, telehealth, and teletriage. Vet Clin North Am Small Anim Pract 2022 Sep;52(5):1069-1080. [doi: 10.1016/j.cvsm.2022.06.004] [Medline: 36150784]
16. Omboni S. Connected health: in the right place at the right time. Connect Health Telemed 2021;1:1-6. [doi: 10.20517/ch.2021.01]
17. Moussa FL, Moussa ML, Alharbi HA, et al. Telehealth readiness of healthcare providers during COVID-19 pandemic in Saudi Arabia. Healthcare (Basel) 2023 Mar 13;11(6):842. [doi: 10.3390/healthcare11060842] [Medline: 36981499]
18. Alotaibi HF, Ghazi SS, Asraf NO, Al Asmari ZS. Patients' perceptions of telehealth video visits experience in primary healthcare setting, Saudi Arabia. J Family Med Prim Care 2023 Nov;12(11):2676-2684. [doi: 10.4103/jfmpc.jfmpc_2269_22] [Medline: 38186832]
19. Guest G, Bunce A, Johnson L. How many interviews are enough? An experiment with data saturation and variability. Field Methods 2006;18(1):59-82. [doi: 10.1177/1525822X05279903]
20. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
21. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
22. Edwards L, Thomas C, Gregory A, et al. Are people with chronic diseases interested in using telehealth? A cross-sectional postal survey. J Med Internet Res 2014 May 8;16(5):e123. [doi: 10.2196/jmir.3257] [Medline: 24811914]
23. Sinha Gregory N, Shukla AP, Noel JJ, et al. The feasibility, acceptability, and usability of telehealth visits. Front Med (Lausanne) 2023;10:1198096. [doi: 10.3389/fmed.2023.1198096] [Medline: 37538312]
24. Al Baalharith I, Al Sherim M, Almutairi SHG, Albaqami ASA. Telehealth and transformation of nursing care in Saudi Arabia: a systematic review. Int J Telemed Appl 2022;2022:8426095. [doi: 10.1155/2022/8426095] [Medline: 36249324]
25. Verma R, Krishnamurti T, Ray KN. Parent perspectives on family-centered pediatric electronic consultations: qualitative study. J Med Internet Res 2020 Apr 9;22(4):e16954. [doi: 10.2196/16954] [Medline: 32084626]
26. Alharbi RA. Awareness, practices, attitudes, and barriers of telehealth In Saudi Arabia. J Pak Med Assoc 2023 Aug;73(8):1658-1662. [doi: 10.47391/JPMA.8263] [Medline: 37697758]

Abbreviations

EHR: electronic health record
HCP: health care provider
TAM: Technology Acceptance Model

Edited by J Sarvestan; submitted 16.03.25; peer-reviewed by J Downey; revised version received 23.09.25; accepted 30.09.25; published 25.11.25.

Please cite as:
Alomar D, Eleftheriou I, Whelan P, Ainsworth J
Usability, Benefits, and Barriers Associated With Patients' Access to Electronic Health Record–Integrated Telehealth in Hospitals in Riyadh: Qualitative Study
JMIR Form Res 2025;9:e74011
URL: https://formative.jmir.org/2025/1/e74011
doi:10.2196/74011

© Dalia Alomar, Iliada Eleftheriou, Pauline Whelan, John Ainsworth. Originally published in JMIR Formative Research (https://formative.jmir.org), 25.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction

XSL•FO
RenderX

in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Evaluating Locally Run Large Language Models (Gemma 2, Mistral Nemo, and Llama 3) for Outpatient Otorhinolaryngology Care: Retrospective Study

Christoph Raphael Buhr[1,2], MSc, MD; Christopher Seifen[1], MD; Katharina Bahr-Hamm[1], MD; Tilman Huppertz[1], MD; Johannes Pordzik[1], MD; Harry Smith[3], MSc; Tom Kelsey[3], MSc; Andrew Blaikie[2], MD; Christoph Matthias[1], MD; Sebastian Kuhn[4], MD; Jonas Eckrich[1], MD

[1]Department of Otorhinolaryngology, University Medical Center of the Johannes Gutenberg-University Mainz, Langenbeckstraße 1, Mainz, Germany
[2]School of Medicine, University of St Andrews, St Andrews, United Kingdom
[3]School of Computer Science, University of St Andrews, St Andrews, United Kingdom
[4]Institute for Digital Medicine, Philipps University Marburg, University Hospital Giessen and Marburg, Marburg, Germany

Corresponding Author:
Christoph Raphael Buhr, MSc, MD
Department of Otorhinolaryngology, University Medical Center of the Johannes Gutenberg-University Mainz, Langenbeckstraße 1, Mainz, Germany

Abstract

Background: Large language models (LLMs) have great potential to improve and make the work of clinicians more efficient. Previous studies have mainly focused on web-based services, such as ChatGPT, often with simulated cases. For the processing of personalized patient data, web-based services have major data protection concerns. Ensuring compliance with data protection and medical device regulations therefore remains a critical challenge for adopting LLMs in clinical settings.

Objective: This retrospective single-center study aimed to evaluate locally run LLMs (Gemma 2, Mistral Nemo, and Llama 3) in providing diagnosis and treatment recommendation for real-world outpatient cases in otorhinolaryngology (ORL).

Methods: Outpatient cases (n=30) from regular consultation hours and the emergency service at a university hospital ORL outpatient department were randomly selected. Documentation by ORL doctors, including anamnesis and examination results, was passed to the locally run LLMs (Gemma 2, Mistral Nemo, and Llama 3), which were asked to provide diagnostic and treatment strategies. Recommendations of the LLMs and the treating ORL doctors were rated by 3 experienced ORL consultants on a 6-point Likert scale for medical adequacy, conciseness, coherence, and comprehensibility. Moreover, consultants were asked whether the answers pose a risk to the patient's safety. A modified Turing test was performed to distinguish responses generated by LLMs from those of doctors. Finally, the potential influence of the information generated by the LLMs on the raters' own diagnosis and treatment opinions was evaluated.

Results: Over all categories, ORL doctors achieved superior (P<.0005) ratings compared to locally run LLMs (Llama 3, Mistral Nemo, and Gemma 2). ORL doctors' responses were considered hazardous for patients in only 1% of the ratings, whereas recommendations by Llama 3, Gemma 2, and Mistral Nemo were considered hazardous in 54%, 47%, and 32% of cases, respectively. According to the raters, the LLM's information rarely influenced their judgment, with Mistral Nemo, Gemma 2, and Llama 3 achieving 1%, 3%, and 4% of the ratings, respectively.

Conclusions: Although locally run LLM models still underperform compared with their web-based counterparts, they achieved respectable results on outpatient treatment in this study. Nevertheless, the retrospective and single-center nature of the study, along with the clinicians' documentation style, may have introduced bias in favor of human recommendations. In the future, locally run LLMs will help address data protection concerns; however, further refinement and prospective validation are still needed to meet strict medical device requirements. As locally run LLMs continue to evolve, they are likely to become comparably powerful to web-based LLMs and become established as useful tools to support doctors in clinical practice.

(JMIR Form Res 2025;9:e76896) doi:10.2196/76896

KEYWORDS

large language models; artificial intelligence; otorhinolaryngology; digital health; chatbot; global health; low- and middle-income countries; telemedicine; telehealth

XSL•FO
RenderX

Introduction

Potential of Large Language Models in Medicine

The introduction of new large language models (LLMs), such as ChatGPT (OpenAI, California, USA), has disrupted the traditional perceptions of artificial intelligence (AI). Rather than requiring extensive coding skills to achieve a task, LLMs understand natural human language input, making their capabilities accessible to much broader tasks. Instead of being trained for a specific purpose, these models are 'all-rounders' capable of accomplishing a wide range of different tasks, including medical queries [1,2]. When required, further fine-tuning can improve their performance for certain fields of action while achieving higher cost-efficiency than training a whole new model for a specific purpose. These characteristics are essential for medical applications. First, medical data such as anamnesis and examination findings are usually recorded in a semistructured manner in human language. Second, the broadly spread competence of LLMs is helpful for understanding the complex and interlinked issues patients present to health systems. Third, with regard to increasing economic pressure in the health care system, LLM support might present a promising solution to increase efficacy, improve outcomes, and reduce costs.

LLMs in Otorhinolaryngology

The application of LLMs in otorhinolaryngology (ORL) is the subject of current research. Suggested areas of application include research and clinical use. Clinical uses range from patient education to improving electronic medical records, triage, patient classification, clinical education, and decision support [3]. More specifically, recent studies have evaluated LLMs in answering patients' questions [4] over the analysis of polysomnographic results [5] to tumor board augmentation [6,7]. Nevertheless, the most frequent touch points between ORL doctors and patients occur in the outpatient care sector.

Thus, this area of application offers one of the most relevant instances where support of ORL doctors by LLMs may be useful. Previous studies in this area highlighted the promising performance of LLMs, while stating the overall superiority of ORL consultants. A study from 2023 compared LLM recommendations on case-based questions with those of experienced ORL consultants. In this case, the LLMs received high ratings in semantic categories and promising ratings on medical adequacy. Furthermore, the study revealed a significant improvement in medical adequacy between 2 tested versions (ie, ChatGPT 3 and ChatGPT 4) [8]. A further, similarly designed study assessed differences between various LLMs. Here, Claude 2 (Anthropic, California, USA) and ChatGPT 4 achieved the highest ratings on medical adequacy, whereas answers from ChatGPT 4 proved to be the most secure for the patients [9]. While both studies cited earlier used rated categories as performance metrics (ie, medical adequacy, coherence, comprehensibility, and conciseness, each rated on a Likert scale), another study primarily evaluated agreement between physicians' recommendations and those of the LLMs. In this study, the agreement of physicians and ChatGPT on physician-written clinical vignettes in otolaryngology was rated using a 5-point Likert scale. The authors showed high agreement of ChatGPT and the physicians on differential diagnosis and treatment plans, with no association between vignette difficulty and agreement with differential diagnosis or treatment [10].

Conquering Data Protection Challenges: The Potential of Locally Run LLMs

Despite the promising findings these studies offer, they often use highly structured (and simulated) input data, mainly focusing on web-based LLMs, such as ChatGPT. This approach has obvious data protection limitations when it comes to real-world implementation. Data protection requirements are the key constraint when using web-based LLMs, as sharing highly sensitive patient data with a third party raises confidentiality concerns. While the General Data Protection Regulation [11] regulates secure data processing in the European Union, the Health Insurance Portability and Accountability Act [12] is of utmost importance in the United States. Unlike web-based LLMs, locally operated LLMs can mitigate data protection concerns by avoiding external data transfers. To date, locally run LLMs have received little attention for use in medical queries, especially for outpatient treatment, despite the potential benefits they could provide. Thus, this study evaluates different locally run open-access LLMs (ie, Llama 3 (Meta, California, USA), Mistral Nemo (Mistral AI, Paris, France), and Gemma 2 (Google, California, USA)) in providing diagnosis and therapy recommendation for real outpatient cases with no data sharing and confidentiality concerns.

Methods

Study Design and Patient Case Selection

The workflow of the study is visualized in Figure 1. Thirty outpatient cases from the archive of our regular consultation hours and the emergency service at the university hospital ORL outpatient office were retrieved for the study. Cases were selected at random from attendance lists. The selection was designed to be as broad as possible, and duplication was avoided as the frequency of different diseases differs. In case of duplication (same disease as one in the previous case), a new case was selected. The retrieved cases involved the following diagnoses: chronic sinusitis and frontal osteoma, tinnitus and conductive hearing loss on both sides, nasal polyp on the left side, deviated septum, conchal hyperplasia, functional voice disorder, dysosmia, recurrent chronic sinusitis, chronic sinusitis, median neck cyst, vocal cord/fold polyp, chronic otitis media, tympanic effusions on both sides, suspected adenoid vegetations, nasal bone fracture, nasal bridge laceration, severe sensorineural hearing loss, vallecula cyst, tumor in the parotid gland, chronic mesotympanic otitis media, lymphadenopathy, benign paroxysmal positional vertigo and vestibulopathy, tonsillar carcinoma, benign paroxysmal positional vertigo, acute bilateral tonsillitis, otitis externa, sillar abscess, foreign body esophagus, foreign body ear, epistaxis after septoplasty, obstructive ear wax, acute unilateral vestibulopathy, and otitis media.

XSL•FO
RenderX

Figure 1. Workflow of the study. LLM: large language model.

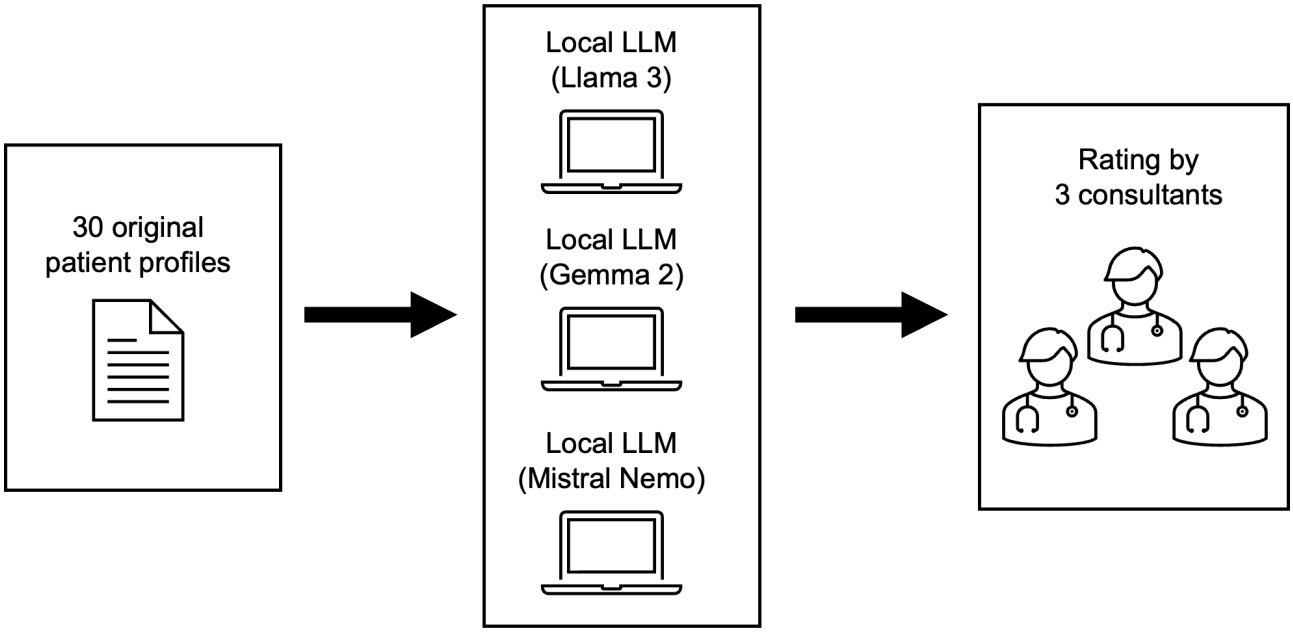

Data Preparation and Prompting

The selected cases were saved in a text document (Word; Microsoft, Redmond, California, USA). The entry also included the patient's gender and year of birth. Generally, medical documentation of the ORL doctors in our clinic is subdivided into anamnesis, examination findings, diagnosis, and treatment recommendation. Therefore, the diagnoses and treatment recommendation section was removed before each patient case was passed to the locally run LLMs (ie, Llama 3, Mistal Nemo, and Gemma 2). All cases were passed to the different locally run LLMs using the basic prompt shown in Figure 2 in the German language. The prompt refers to ORL doctors' documentation style in our clinic and explicitly requests the diagnosis and treatment recommendation, limiting the answer to 100 words. The prompt has been kept deliberately simple to test the LLM's baseline performance. The instructions provided aimed to ensure that the response resembles the doctors' documentation style.

Figure 2. Prompt use within this study.

The above-mentioned patient presents at an ear, nose and throat outpatient clinic.
You can see the attending physician's history entry above.

Please provide a diagnosis and procedure for the above patient:
Diagnosis: XX
Procedure: XX

Limit the answer to 100 words and specify a concrete procedure for diagnosis and treatment.

XSL•FO
RenderX

LLM Execution

The LLMs were run locally on a standard Hewlett-Packard (Palo Alto, California, USA) Notebook (Intel Core 17 - 1255U, 4.7 GHz. DDR4, 16 GB [2×8 GB], Windows 10 Pro, LM Studio 0.3.5) using the prompt shown in Figure 2 as stated earlier.

The selection of LLMs was guided by their open-source availability and model size. Consequently, only models below 10 GB were included. The default settings of LLM Studio were not modified, and all model configurations— including parameter sizes or variants, quantization settings, prompt templates and system prompts, as well as temperature, top-p, maximum token limits, and random seeds—were kept unchanged. The specific models evaluated in this study are as follows: Meta-Llama-3-8B-Instruct-GGUF (GPU Offload 32/32; CPU Thread Pool Size 4, Evaluation Batch Size 512, Context Length 4096, Random Seed), Gemma-2-9b-it-GGUF (GPU Offload 37/42; CPU Thread Pool Size 4, Evaluation Batch Size 512, Context Length 4096, Random Seed), and Mistral-Nemo-Instruct-2407-GGUF (GPU Offload 27/40; CPU Thread Pool Size 4, Evaluation Batch Size 512, Context Length 4096, Random Seed) (all published by lmstudio-community).

Evaluation and Rating

To evaluate the results, all cases were collected in a single text document. In particular, the doctor's documentation (eg, medical history and examination findings) was presented first for each case. Subsequently, the diagnoses and treatment recommendations of the 3 LLMs and the treating ORL doctor were presented in random order for the specific case. The evaluation criteria stated below were filled in under each answer (Likert scale and binary rating) by each rater.

The recommendation of the LLMs and the original documentation of the treating ORL doctors were rated by 3 experienced ORL consultants on a 6-point Likert scale (1=very poor and 6=excellent) for medical adequacy, conciseness, coherence, and comprehensibility.

The key assessment metrics were defined as follows:

- *Medical adequacy* refers to the accuracy and appropriateness of medical content relative to established clinical guidelines and expert consensus.
- *Coherence* refers to the logical flow and consistency of the information presented. This was assessed by reviewers based on the clarity of connections between ideas and the absence of contradictory statements.
- *Comprehensibility* measures the ease with which the information can be understood.
- *Conciseness* evaluates whether the information is sufficiently succinct without omitting critical details.

Furthermore, *hazardous for patients* was defined as the presence of information that could directly or indirectly lead to patient harm if followed. This includes advice that contradicts established clinical guidelines, promotes unsafe practices, misrepresents risks or benefits, or could result in delayed diagnosis, inappropriate treatment, or adverse outcomes.

Moreover, ORL consultants were asked (binary rating) whether the answers pose a risk to the patient's safety, whether the response originates from an LLM or a doctor (modified Turing test) [13], and whether their own diagnosis and treatment opinion would be influenced by the respective answer. In total, for each included case, the 3 experienced ORL consultants rated the response (diagnosis and treatment recommendation) of the treating ORL doctor and every tested LLM.

Benchmarking Against Web-Based LLMs

Web-based LLMs have been previously evaluated as mentioned earlier [8,9] and have shown sufficient capabilities in evaluating case-based questions. To benchmark the locally run LLMs with an established web-based LLM, 10 simulated cases were comparatively evaluated following the same workflow during the course of the study. Accordingly, the 10 simulated cases were processed by an ORL doctor and submitted to the 3 locally run LLMs (ie, Llama 3, Mistal Nemo, and Gemma 2), as well as a web-based LLM (ie, Chat GPT-4o).

Statistical Analysis

All responses from the raters were transferred to a Microsoft Excel spreadsheet and sorted according to the entity being evaluated. Statistical analysis was performed with GraphPad Prism for macOS (version 10.3.0; GraphPad Software, La Jolla, CA, USA). The data did not meet normality assumptions, as confirmed by the D'Agostino and Pearson test.

Further statistical analysis was performed using Python (version 3.11.13) in the Google Colab environment. The data were imported from Excel files. Packages used include pandas (data processing), statsmodels, and pingouin.

For each rating category (ie, medical adequacy, conciseness, coherence, and comprehensibility), a linear mixed effects model was fitted, with entity as a fixed effect and case and rater as random intercepts to account for clustering within cases and raters. Furthermore, interrater reliability was examined using evaluations of the same responses by multiple raters.

The following measures were used to calculate reliability:

- Fleiss κ, a measure of agreement in categorical evaluations, was determined using the fleiss_kappa function from the statsmodels package.
- The percentage agreement was calculated as the proportion of cases in which all raters gave exactly the same rating.
- The intraclass correlation coefficient (ICC) for quantifying agreement in ordinal ratings was calculated using the intraclass_corr function from the pingouin package. For this purpose, the data were transformed into a long format containing target objects, raters, and ratings.

Ethical Considerations

Ethical approval was obtained from the ethics committee of the state medical association (request/approval number: 2023 - 17385-retrospektiv). Informed consent was not required as this was a retrospective study. The data were processed anonymously to ensure privacy, and compensation was not necessary due to the retrospective nature of the study.

XSL•FO
RenderX

Results

Rating Results

ORL doctors showed significantly (P<.05%) higher ratings compared to the locally run LLMs (ie, Llama 3, Mistral Nemo, and Gemma 2) over all categories (Figure 3). Altogether, the doctors achieved (median 6, interquartile range [IQR] 5 - 6) the highest rating in every category. On medical adequacy, Gemma 2 and Mistral Nemo received similar ratings (Mistral Nemo: median 4, IQR 3-5; and Gemma: median 4, IQR 3-4), whereas Llama 3 performed inferior scores, achieving a median of 3 (IQR 3-4). For conciseness, all locally run LLMs tested received similar ratings (Gemma 2 and Mistral Nemo: median 4, IQR 4-5; and Llama 3: median 4, IQR 3-4). Mistral Nemo (median 5, IQR 4-5) outperformed Gemma 2 and Llama 3 on ratings for coherence and comprehensibility. Both Gemma 2 and Llama 3 received similar ratings in these categories. Detailed rating results are presented in Table 1. Details of the linear mixed effects model for predicting ratings are presented in Table 2, and P values for entity coefficients were all <.001, indicating a significant difference between ORL doctors and LLMs. Boxplots of the ratings categorized for each model can be found in Multimedia Appendix 1 .

XSL•FO
RenderX

Figure 3. Rating for locally run LLMs and ORL consultants on providing diagnosis and treatment strategy (30 real-world cases) regarding (A) medical adequacy, (B) conciseness, (C) coherence, and (D) comprehensibility rated on a 6-point Likert scale. Normality distribution was tested with the D'Agostino and Pearson test. Comparisons were performed using a linear mixed-effects model. Nonsignificant >0.05, *P<.05, **P<.005, ***P<.001. LLM: large language models; ORL: otorhinolaryngology.

XSL•FO
RenderX

Table . Results of the rating by ORL[a] consultants.

	Values, mean (SD)	Rating, median (IQR[b])	95% CI of median
Medical adequacy			
ORL doctors	5.6 (0.6)	6 (5-6)	6-6
Gemma 2	3.5 (1.2)	4 (3-4)	3-4
Mistral Nemo	4 (1.2)	4 (3-5)	4-4
Llama 3	3.4 (1.2)	3 (3-4)	3-4
Conciseness			
ORL doctors	5.7 (0.5)	6 (5-6)	6-6
Gemma 2	4.2 (0.9)	4 (4-5)	4-4
Mistral Nemo	4.3 (1)	4 (4-5)	4-5
Llama 3	3.8 (0.8)	4 (3-4)	4-4
Coherence			
ORL doctors	5.7 (0.6)	6 (5-6)	6-6
Gemma 2	4.2 (1.1)	4 (3-5)	4-5
Mistral Nemo	4.4 (1.2)	5 (4-5)	5-5
Llama 3	3.8 (1.1)	4 (3-5)	4-4
Comprehensibility			
ORL doctors	5.7 (0.5)	6 (6)	6-6
Gemma 2	4.3 (1)	4 (4-5)	4-5
Mistral Nemo	4.5 (1.1)	5 (4-5)	4-5
Llama 3	3.9 (1)	4 (3-5)	4-4

[a]ORL: otorhinolaryngology.

[b]IQR: interquartile range.

XSL•FO
RenderX

Table . Linear mixed effects model for ratings.

	Estimate	*P* value	ci_lower	ci_upper
Medical adequacy				
Intercept	5.633	0	5.370	5.896
entity[T.Gemma 2]	−2.1	<.001	−2.385	−1.814
entity[T.Mistral Nemo]	−1.633	<.001	−1.918	−1.348
entity[T.Llama 3]	−2.278	<.001	−2.563	−1.993
Group Var	0.223	.01	0.045	0.402
rater Var	0.033	.517	−0.067	0.134
Conciseness				
Intercept	5.711	0	5.524	5.899
entity[T.Gemma 2]	−1.556	<.001	−1.773	−1.338
entity [T.Mistral Nemo]	−1.456	<.001	−1.673	−1.238
entity[T.Llama 3]	−1.922	<.001	−2.14	−1.705
Group Var	0.106	.127	−0.03	0.242
rater Var	0.164	.05	<0.001	0.329
Coherence				
Intercept	5.677	0	5.469	5.886
entity[T.Gemma 2]	−1.5	<.001	−1.768	−1.232
entity[T.Mistral Nemo]	−1.244	<.001	−1.512	−0.976
entity[T.Llama 3]	−1.844	<.001	−2.112	−1.576
Group Var	0.014	.716	−0.064	0.092
rater Var	0.166	.08	−0.021	0.352
Comprehensibility				
Intercept	5.744	0	5.526	5.963
entity[T.Gemma 2]	−1.422	<.001	−1.670	−1.174
entity[T.Mistral Nemo]	−1.267	<.001	−1.515	−1.019
entity[T.Llama 3]	−1.856	<.001	−2.104	−1.607
Group Var	0.121	.109	−0.027	0.268
rater Var	0.183	.036	0.012	0.355

Modified Turing Test

Regarding the modified Turing test, Gemma 2 was recognized as a machine in 99% (89/90), Mistral Nemo in 98% (88/90), and Llama 3 in 97% (87/90) of the ratings. Conversely, the ORL doctors were recognized as a human being in 99% (89/90) of the ratings.

Assessment of Patient Safety

While the raters considered the ORL doctors' recommendations as potentially unsafe for the patient in only 1% (1/90) of cases, Llama 3's recommendations were considered potentially hazardous in 54% (49/90), Gemma 2's in 47% (42/90), and Mistral Nemo's in 32% (29/90) of cases.

Influence on Raters' Own Decision-Making

Whereas the ORL consultants stated that the information provided by the ORL doctors might have influenced their own decision in 2% (2/90), Mistral Nemo was considered influential in 1% (1/90), Gemma 2 in 3% (3/90), and Llama 3 in 4% (4/90) of cases for this assessment.

Number of Words

The number of words used by the ORL doctors and the locally run LLMs is visualized in Figure 4. Llama 3 (median 99, range 19 - 191, IQR 83 - 124) spent the most words on providing a diagnosis and treatment recommendations, followed by Gemma 2 (median 53, range 33 - 83, IQR 45 - 68) and Mistral Nemo (median 53, range 13 - 127, IQR 35 - 68) and the ORL doctors (median 20, range 3 - 53, IQR 13 - 37).

XSL•FO
RenderX

Figure 4. Number of words used by ORL doctors and the locally run LLMs (ie, Gemma 2, Mistral Nemo, and Llama 3) on providing diagnosis and treatment recommendations.

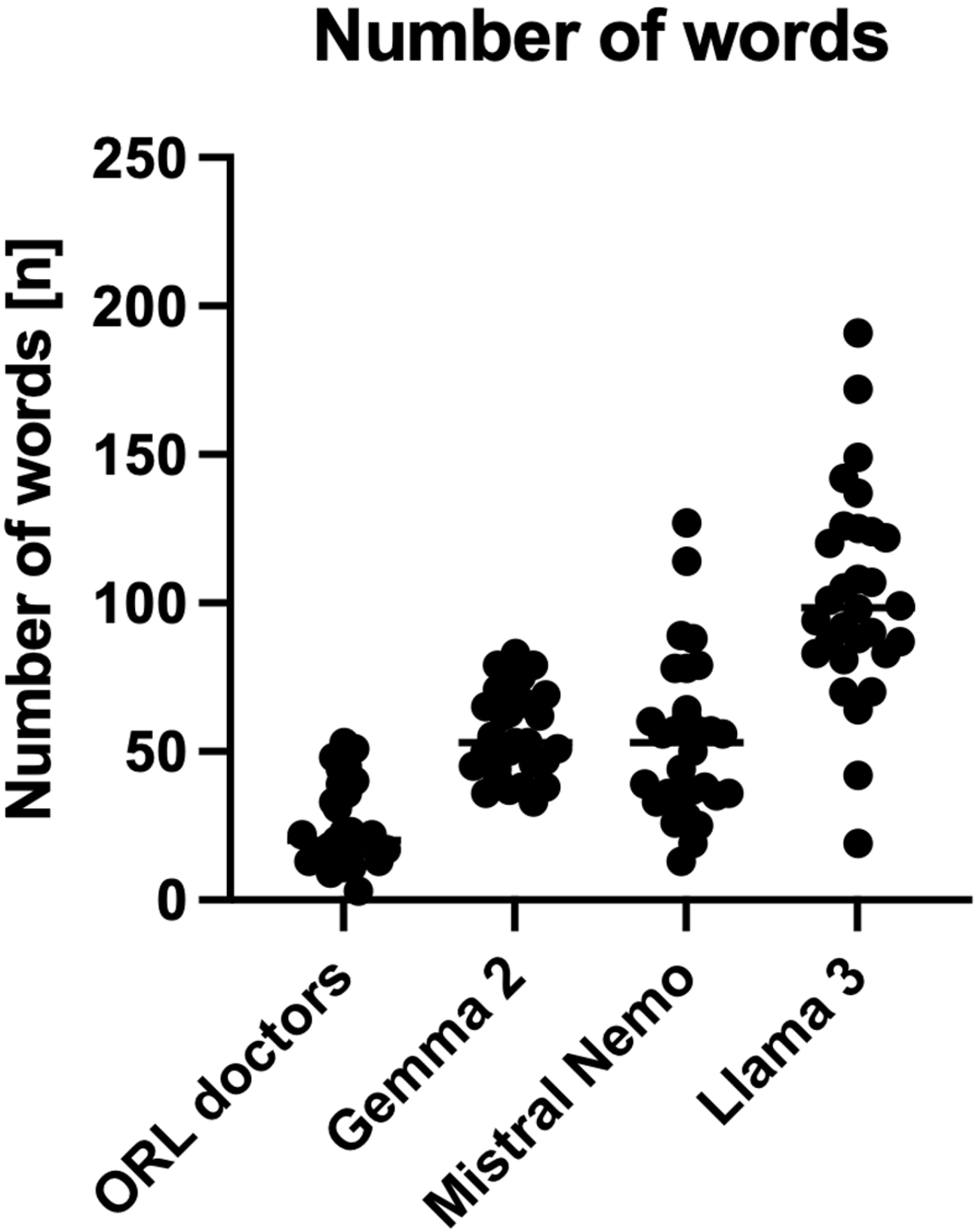

Interrater Reliability

The interrater reliability analysis using Fleiss κ and percentage match is presented in Table 3. A further interrater reliability analysis applying the ICC is illustrated in Table 3.

XSL•FO
RenderX

Table . Overall interrater reliability: Fleiss κ and percentage match.

Category	Fleiss κ[a] Rater 1 - 3 170 responses evaluated	Percentage of homolog ratings between all raters[b]: number of homolog ratings between all ratersnumber of cases
Modified turing test	0.906	0.947
Hazardous for patients	0.419	0.635
Influence by the respective answer	−0.028	0.917
Medical adequacy	0.241	0.206
Conciseness	0.167	0.165
Coherence	0.152	0.159
Comprehensibility	0.186	0.188

[a]Interpretation of Cohen's κ adapted from Landis and Koch (1977).

[b]<0: worse than chance; 0.00-0.20: low agreement; 0.21-0.40: moderate agreement; 0.41-0.60: satisfactory correlation; 0.61-0.80: good correlation; 0.81-1.00: very good to perfect correlation.

Discussion

Context of the study

Previous studies assessing LLMs on outpatient treatment in ORL focused on web-based LLMs mainly using simulated cases [8-10,14]. While web-based LLMs show impressive linguistic and professional capabilities, this setup has obvious limitations due to data sharing concerns. These concerns could be circumvented using LLMs that have no connection to the internet working on a local computer. Hence, this study evaluates the capabilities of 3 different locally run LLMs (ie, Llama 3, Mistral Nemo, and Gemma 2) on real patient data from outpatient consultations of a University Medical Center in Germany.

Despite ORL doctors receiving superior ratings over all categories, this study demonstrates the ability of locally run LLMs to process outpatient data and provide useful diagnostic and treatment information. Regarding medical adequacy, Gemma 2 and Mistral Nemo received similar high ratings (Mistral Nemo: median 4, IQR 3-5; and Gemma: median 4, IQR 3-4), whereas Llama 3 performed inferiorly, achieving a median of 3 (IQR 3-4). Comparing with previous studies, the performance of Gemma 2 and Mistral Nemo is similar to ChatGPT 3's performance in a study from 2023 reaching a median rating of 4 (IQR 4-5) on medical adequacy [8]. However, a comparison with a more recent study reveals an inferior rating on medical adequacy. In the mentioned study from 2024, the web-based LLMs Claude 2 (median 5, IQR 4-6), ChatGPT 4 (median 5, IQR 4-6), and Bard 2023.07.13 (median 5, IQR 3-5) achieved higher ratings on medical adequacy than the best performing locally run LLMs (ie, Mistral Nemo and Gemma 2) in this study [9]. Regarding conciseness, all locally run LLMs received similar ratings (Gemma 2 and Mistral Nemo: median 4, IQR 4-5; and Llama 3: median 4, IQR 3-4). These performance levels correspond with the results from prior studies [8,9]. Mistral Nemo (median 5, IQR 4-5) outperformed Gemma 2 and Llama 3 on ratings for coherence and comprehensibility, which received similar ratings in these categories. With regard to the modified Turing test, the ORL doctors were recognized as human beings in 99% (89/90) of the ratings, whereas the LLMs were almost always correctly identified as machines. Gemma 2 was recognized as a machine in 99% (89/90), Mistral Nemo in 98% (88/90), and Llama 3 in 97% (87/90) of the ratings. Again, these ratings from this study for coherence and comprehensibility are in line with results from previous studies [8,9].

Hazardous for Patients

It is important to examine the cases marked as hazardous in more detail to better understand the risks posed by LLMs. The only risk identified by the treating doctor is in the case of a patient with a large nasal polyp, for whom the polyp's removal under local anesthesia was recommended. The reason why this was assessed as a potential risk to the patient remains unclear, but it may be explained by the fact that the patient is taking anticoagulants (acetylsalicylic acid and clopidogrel). Overall, Llama 3's recommendations were considered potentially hazardous in 54% (49/90), Gemma 2's in 47% (42/90), and Mistral Nemo's in 32% (29/90) of cases. A typical example of an LLM shortcoming is a case involving a patient with chronic sinusitis who reports being largely symptom-free while undergoing local cortisone therapy. While the doctor and Mistral Nemo recommend continuing with the cortisone therapy alone, Llama 3 and Gemma 2 recommend adding antibiotics to the treatment. In this case, the patient would be prescribed antibiotics unnecessarily, posing a risk to the patient, particularly given that the examination findings do not mention any pus. In another case involving a 3-year-old girl with suspected adenoids and tympanic fluid, Gemma 2 recommends myringoplasty, Mistral Nemo recommends a wait-and-see approach, and Llama 3 recommends speech therapy and a tonsillectomy. In contrast, the doctor recommends an adenotomy and the insertion of tympanic drains. Recommending the wrong operation is certainly one of the most extreme misjudgments. This demonstrates the importance of careful examination and shows that, despite their eloquence, LLMs should not be trusted unconditionally.

XSL•FO
RenderX

Benchmarking With Web-Based LLMs

The comparison with preceding underscores reveals the improvement of LLMs over time [8,9]. To benchmark the locally run LLMs with their current web-based counterparts, an assessment on 10 separate simulated cases (simulated to meet data protection constraints) was performed. Here, the web-based LLM ChatGPT-4o and ORL doctors achieved similar high ratings on medical adequacy (ORL doctors: median 6, IQR 6; ChatGPT-4o: median 6, IQR 5-6). Regarding patients' security, a potential hazard for patients was noted in a single rating for ChatGPT-4o (1/30, 3%), in 53% (16/30) for Llama 3, and in 23% (7/30) for Mistral Nemo and Gemma 2. None of the ORL doctors' recommendations was considered as potentially hazardous for patients. In addition, the raters found ChatGPT-4o to potentially influence their own judgment in 10% (3/30) of the cases. Among the locally run LLMs, only Llama 3 received a single positive rating in this dimension (3%; detailed data regarding the web-based LLM assessment are present in Multimedia Appendix 1). Interestingly, ChatGPT-4o has only been accused of jeopardizing a patient in one case. In this case, the patient had a suspicious cervical lymph node, and the doctor recommended extirpation. However, ChatGPT-4o recommended fine needle aspiration, laboratory tests, a magnetic resonance imaging scan, and a follow-up appointment. The reason for flagging this answer as risky is probably that the LLM did not commit to any procedure. Mistral Nemo recommended a biopsy or excision, Gemma recommended a biopsy and a blood test, and Llama suggested a check-up in 4 weeks. Although the assessment of the web-based LLMs did only include 10 simulated cases rather than real-world data, it does demonstrate the high-performance level of current web-based LLMs and their leading edge over the tested locally run competitors.

Recent developments in advanced reasoning LLMs (often referred to as chain-of-thought or system-prompt–augmented models) could further improve diagnostic accuracy and reduce risky recommendations by simulating more nuanced clinical reasoning processes. These models leverage step-by-step logical reasoning, potentially enabling them to identify and correct errors before delivering final outputs.

Interrater Reliability

Analyzing interrater reliability can be challenging. In this study, the evaluating ORL consultants demonstrated an excellent Fleiss κ correlation of 0.906 in the modified Turing test (with binary answers). For the assessment of hazards to patients (also binary answers), a Fleiss κ of 0.419 was found, indicating a satisfactory level of correlation. Furthermore, ORL consultants were asked whether the presented answer would have influenced their own decision (binary rating). This produced a Fleiss κ of −0.028, which formally corresponds to poorer correlation than expected by chance. However, assessment of the raw data showed high concordance between raters; therefore, we include the percentage of homolog ratings between all raters, which was 91.7% (number of homolog ratings between all raters/number of cases). Taking these measures into consideration, the low Fleiss κ value is therefore most likely caused by the high uniformity of the ratings provided.

Fleiss κ and the percentage of homolog ratings between all raters suggested low interrater reliability for ratings in the categories of medical adequacy, conciseness, coherence, and comprehensibility. The same applied for the absolute agreement. However, an examination of the raw data revealed that the raters' assessments were in fact highly consistent throughout. Although the raters often did not award exactly the same points, the scores awarded were close to each other, for example, rating 5 points and rating 6 points for the same answer by different raters. Hence, we calculated ICCs. The ICC analysis shows a moderate (ICC≈0.46 - 0.65) interrater agreement in the abovementioned categories (Table 4). The aggregated assessment of several raters (ICCk) achieved moderate to good values in all categories (ICCk ≈0.72 - 0.85). The highest agreement was for medical adequacy, the lowest for comprehensibility. All ICC values were highly significant (P<.001), with narrow CIs, indicating stable estimates. In conclusion, despite the fact that all rating ORL consultants rarely choose exactly the same rating for the responses given, they highly agreed within the tendency of ratings (Table 4). Overall, this analysis reveals a high consistency among ratings.

XSL•FO
RenderX

Table . Overall interrater reliability: intraclass correlation coefficient (ICC).

Category	Description	ICC	*F* statistic	df1[a]	df2[a]	*P* value	95% CI[b]
Medical adequacy							
ICC1[c]	Single raters absolute	0.653	6.643	169	340	<.001	0.58-0.72
ICC2[d]	Single random raters	0.653	6.614	169	338	<.001	0.58-0.72
ICC3[e]	Single fixed raters	0.652	6.614	169	338	<.001	0.58-0.72
ICC1k[f]	Average raters absolute	0.849	6.643	169	340	<.001	0.81-0.88
ICC2k[f]	Average random raters	0.849	6.614	169	338	<.001	0.81-0.88
ICC3k[f]	Average fixed raters	0.849	6.614	169	338	<.001	0.8-0.88
Conciseness							
ICC1	Single raters absolute	0.544	4.575	169	340	<.001	0.46-0.62
ICC2	Single random raters	0.552	5.221	169	338	<.001	0.45-0.64
ICC3	Single fixed raters	0.585	5.221	169	338	<.001	0.5-0.66
ICC1k	Average raters absolute	0.781	4.575	169	340	<.001	0.72, 0.83
ICC2k	Average random raters	0.787	5.221	169	338	<.001	0.71-0.84
ICC3k	Average fixed raters	0.808	5.221	169	338	<.001	0.75-0.85
Coherence							
ICC1	Single raters absolute	0.500	4.000	169	340	<.001	0.41-0.58
ICC2	Single random raters	0.512	4.680	169	338	<.001	0.4-0.61
ICC3	Single fixed raters	0.551	4.680	169	338	<.001	0.47-0.63
ICC1k	Average raters absolute	0.750	4.000	169	340	<.001	0.68-0.81
ICC2k	Average random raters	0.759	4.680	169	338	<.001	0.66-0.83
ICC3k	Average fixed raters	0.786	4.680	169	338	<.001	0.72-0.84
Comprehensibility							
ICC1	Single raters absolute	0.457	3.523	169	340	<.001	0.37-0.55
ICC2	Single random raters	0.478	4.524	169	338	<.001	0.33-0.6
ICC3	Single fixed raters	0.540	4.524	169	338	<.001	0.45-0.62
ICC1k	Average raters absolute	0.716	3.523	169	340	<.001	0.63-0.78

XSL•FO
RenderX

Category	Description	ICC	*F* statistic	df1[a]	df2[a]	*P* value	95% CI[b]
ICC2k	Average random raters	0.733	4.524	169	338	<.001	0.59-0.82
ICC3k	Average fixed raters	0.779	4.524	169	338	<.001	0.71-0.83

[a]df1, df2: degrees of freedom.

[b]Interpretation guidelines (Koo & Li, 2016) [15]: <0.50, poor; 0.50 - 0.75, moderate; 0.75 - 0.90, good; >0.90, excellent.

[c]ICC1: single-rater, 1-way random effects (absolute agreement).

[d]ICC2: single-rater, 2-way random effects (absolute agreement).

[e]ICC3: single-rater, 2-way mixed effects (absolute agreement).

[f]ICC1k, ICC2k, ICC3k: corresponding average measures ICCs.

Limitations

Passing original text documentation to the LLMs is both a strength and a limitation of our study at the same time. Original text input in the local language (German in this study) represents the most realistic application in clinical practice. Despite the advantage of a realistic setup, the use of the German language may have had a negative impact on the results. This is indicated by the fact that some LLMs occasionally responded in English, although the rest of the context was in German. Furthermore, ORL doctors' original documentation included grammatical errors and misspellings. Moreover, the documentation provided was limited to the text transcript of a single consultation, not including further diagnostic information, such as imaging diagnostic results. This approach was determined by the input structure of the locally run LLMs assessed in this study. The documentation of the treating ORL doctors may have been created on the basis of further information. Information that the ORL doctors may not include in their text documentation owing to the availability in the clinic's documentation system. However, as the rating ORL consultants had only access to the same documentation as the locally run LLMs, a possible bias is limited.

Furthermore, documentation style between treating units may vary; the ORL doctors as well as the rating ORL consultants are affiliated with the same institution and thus may prefer a special design of documentation. This important advantage in favor of ORL doctors should be taken into account when interpreting the findings. Nonetheless, this limitation is because of the nature of single-center setting, which limits the generalizability of findings to other institutions and patient populations.

Moreover, clinicians' own documentation style—crafted primarily for human readability—may inadvertently favor human reviewers, who are already accustomed to reading similar clinical notes, rather than locally run LLMs. This could introduce a systematic bias when comparing human- versus model-generated recommendations.

Overall, regarding documentation style, this study design applied the most realistic scenario using original documentation of treating ORL doctors. Ultimately, any other strategy (ie, adjustment of documentation) would have introduced a new bias. In addition, when implemented in clinical practice, LLMs must be evaluated using original documentation in any case.

Assessing "conciseness" poses its own challenges: the measured output is strongly influenced by how prompts are written (eg, whether specific length caps or "be concise" cues are used). This study aimed to assess the conciseness of the evaluated LLMs while limiting the word count to 100 within the prompt. Despite the narrow limits given, differences emerged. While the doctors and Gemma 2 adhered to the word limit every time, Mistral Nemo exceeded it twice, and Llama 3 exceeded it 14 times. However, future studies should evaluate standardized prompt instructions, clearly defining output constraints, and using consistent methods (eg, information density assessments or rubrics balancing completeness and brevity) to more reliably measure and compare conciseness across different models and scenarios.

Besides, the model size of locally run LLMs was limited by the computing capacity of the laptop used in this study. More powerful machines using larger locally run LLMs may achieve better results. Yet, the laptop used is a standard laptop, which appears realistic for use in clinical practice. This was in line with the aim of the study to test the current performance with the technical possibilities currently available. Moreover, implementing local LLMs on doctors' local computers at their workstations has advantages over using large LLMs on the clinic's own servers. First, there is no need to purchase additional computing capacity, as the computers already exist and are available, which saves costs. Furthermore, operation is still possible even if there is no connection to the central server, which makes operation significantly more robust and crisis-proof. Owing to continuous technical progress, it is to be expected that the performance of the comparatively small LLMs tested in this study will continue to increase and will surpass the performance of currently available large models in the near future. Accordingly, web-based LLMs can provide a glimpse into the future of what "small" local LLMs will be capable of in the future. In this study, this assessment was covered by benchmarking web-based LLMs on 10 simulated cases. With regard to rating capacity of the ORL consultants, this study was limited to 30 original cases from our outpatient department, causing constraints in the statistical analysis. Furthermore, the limitation to 30 original cases, duplicates in diagnoses were excluded from the random sampling of cases, which could have induced a selection bias.

Finally, the retrospective nature of this study prevents real-time validation of diagnoses and treatment recommendations. To

XSL•FO
RenderX

further assess the true capabilities of local LLMs, future studies should assess prospective multicenter designs.

Local LLM Conquer Data Protection Challenges

The utilization of LLMs in medical treatment has obvious medicolegal and ethical constraints. While locally run LLMs mitigate data protection concerns by avoiding external data transfer, compliance with regulations such as the General Data Protection Regulation [11] in the European Union or the Health Insurance Portability and Accountability Act [12] in the United States remains paramount. Organizations adopting such solutions must ensure that data handling procedures—including encryption, secure storage, and privacy-preserving protocols—are implemented to meet these regulatory standards. Moreover, establishing robust internal governance, maintaining transparency around how patient data are used, and conducting regular audits are essential for preserving patient privacy and upholding trust in AI-assisted care.

Lingering Challenges for Clinical Integration

Beyond data protection, LLMs used for clinical decision-making may also fall under medical device regulations (eg, European Union Medical Device Regulation [16] and Food and Drug Administration guidelines in the United States [17]). The dynamic nature of LLMs, their propensity for generating unpredictable responses (so-called “hallucinations”), and the difficulty of verifying their outputs pose unique challenges in fulfilling strict validation and safety requirements. As such, developers and health care institutions must carefully assess whether the LLM’s intended use qualifies it as a regulated medical device and, if so, seek the necessary certification or approvals before deploying these technologies at scale.

However, recent studies also showed patients’ skepticism regarding AIs' decision-making even when supervised by human specialists [18,19]. These aspects illustrate that many obstacles beyond medical correctness and accuracy remain. These issues should be carefully addressed before an implementation of LLMs into clinical workflows is considered.

Potential of a Clinical Integration

Despite these obvious limitations, this study highlights the broad ability of locally run LLMs to process medical data in ORL and provide diagnostic and treatment strategies without further fine-tuning or training. This is an important finding as LLMs are not trained for a specific narrow purpose, yet they are able to perform satisfactorily in a niche medical subspecialty, such as ORL. LLMs therefore represent a potentially highly cost-efficient means to scale diagnostic support particularly relevant to low-resource settings and financially constrained public health services. While the locally run LLMs failed to meet the high benchmark of ORL doctors, they still provide structured output, which in many cases did provide a suitable treatment strategy for the respective patients. Obviously, locally run LLMs are far from replacing human doctors in health care, especially when considering the relatively high rate of hazardous recommendations of some locally run LLMs. However, their web-based counterparts are capable of providing doctor-like answers with very few mistakes and hazardous potential. As technological evaluation continues, it is likely that locally run LLMs will catch up to the performance of their web-based counterparts. While the medicolegal and ethical aspects of an implementation of LLMs in clinical practices remain, they provide a promising solution to support medical staff. In a time of chronic shortage of trained health care workers all over the world, the feasibility of LLM use should be considered. Auspicious applications range from the assistance of local health workers in low-resource settings to triage or background support in high-resource settings. A roadmap of an implementation in clinical practice should include further validation of locally run LLMs in larger prospective studies. Besides evaluation by scientists, legislators need to take action. Current regulations such as device regulations show shortcomings for new software-based medical tools, such as LLMs. For instance, fast-track approval mechanisms for minor updates are needed to ensure that relevant updates do not delay by undergoing an entire approval process. Finally, the implementation of LLMs in clinical practice should be a joint project involving all stakeholders, including patients, manufacturers, physicians/scientists, and authorities.

Conclusions

Considering the generalist nature of locally run LLMs evaluated in this study, their performance in this specific medical field is impressive. However, locally run LLMs are still outperformed by their web-based counterparts and human specialists. In time, however, locally run LLMs are likely to become as powerful as their current web-based equivalents, creating the realistic prospect of practical enhancement of day-to-day clinical diagnostics in both high- and low-resource settings. As only locally run LLMs meet the high data protection requirements for medical application, future studies should monitor and evaluate their performance on a larger scale, ideally within prospective multicenter clinical studies.

Acknowledgments

Figures 1 and 2 were drawn by Jonas Eckrich and Christoph Buhr using Microsoft PowerPoint (Microsoft Corp, Redmond, WA, USA). Figures 3 and 4 were assembled by Jonas Eckrich and Christoph Buhr using Prism for Windows, version 9.5.1 (GraphPad Software, La Jolla, California, USA).

Funding

There was no funding received for this study.

XSL•FO
RenderX

Data Availability

Owing to the sensitive nature of the clinical data used in this study and in accordance with institutional and legal data protection requirements, the datasets generated and analyzed cannot be made publicly available.

Authors' Contributions

Conceptualization: CRB (lead), HS, TK, AB, SK, JE (co-lead)
Data curation: CRB, CS, KB-H, TH, JP, TK, CM, SK, JE
Formal analysis: CRB, HS, TK, AB, JE
Investigation: CRB, AB, SK, JE
Methodology: CRB, CS, JP, HS, TK, JE
Project administration: CRB, AB, SK, JE
Resources: CRB, TK, SK, JE
Supervision: CRB, AB, CM, SK, JE
Validation: CRB, CS, KB-H, TH, JP, HS, TK, CM, SK, JE
Visualization: CRB, TK, AB, JE
Writing – original draft: CRB, AB, SK, JE
Writing – review & editing: CRB, CS, KB-H, TH, JP, HS, TK, CM, SK, JE

Conflicts of Interest

SK is a Founder & shareholder of MED.digital. All other authors declare no conflicts of interest.

Multimedia Appendix 1
Boxplots illustrating ratings categorized for each model and detailed data on the web-based LLM assessment.
[DOCX File, 415 KB - formative_v9i1e76896_app1.docx]

References

1. Kaddour J, Harris J, Mozes M, Bradley H, Raileanu R, McHardy R. Challenges and applications of large language models. arXiv. Preprint posted online on Jul 19, 2023. [doi: 10.48550/arXiv.2307.10169]
2. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023 Aug;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
3. Banyi N, Ma B, Amanian A, Bur A, Abdalkhani A. Applications of natural language processing in otolaryngology: a scoping review. Laryngoscope 2025 Sep;135(9):3049-3063. [doi: 10.1002/lary.32198] [Medline: 40309961]
4. Pordzik J, Bahr-Hamm K, Huppertz T, et al. Patient support in obstructive sleep apnoea by a large language model - ChatGPT 4o on answering frequently asked questions on first line positive airway pressure and second line hypoglossal nerve stimulation therapy: a pilot study. Nat Sci Sleep 2024;16:2269-2277. [doi: 10.2147/NSS.S495654] [Medline: 39741798]
5. Seifen C, Huppertz T, Gouveris H, et al. Chasing sleep physicians: ChatGPT-4o on the interpretation of polysomnographic results. Eur Arch Otorhinolaryngol 2025 Mar;282(3):1631-1639. [doi: 10.1007/s00405-024-08985-3] [Medline: 39427271]
6. Buhr CR, Ernst BP, Blaikie A, et al. Assessment of decision-making with locally run and web-based large language models versus human board recommendations in otorhinolaryngology, head and neck surgery. Eur Arch Otorhinolaryngol 2025 Mar;282(3):1593-1607. [doi: 10.1007/s00405-024-09153-3] [Medline: 39792200]
7. Lechien JR, Chiesa-Estomba CM, Baudouin R, Hans S. Accuracy of ChatGPT in head and neck oncological board decisions: preliminary findings. Eur Arch Otorhinolaryngol 2024 Apr;281(4):2105-2114. [doi: 10.1007/s00405-023-08326-w] [Medline: 37991498]
8. Buhr CR, Smith H, Huppertz T, et al. ChatGPT versus consultants: blinded evaluation on answering otorhinolaryngology case-based questions. JMIR Med Educ 2023 Dec 5;9:e49183. [doi: 10.2196/49183] [Medline: 38051578]
9. Buhr CR, Smith H, Huppertz T, et al. Assessing unknown potential-quality and limitations of different large language models in the field of otorhinolaryngology. Acta Otolaryngol 2024 Mar;144(3):237-242. [doi: 10.1080/00016489.2024.2352843] [Medline: 38781053]
10. Qu RW, Qureshi U, Petersen G, Lee SC. Diagnostic and management applications of ChatGPT in structured otolaryngology clinical scenarios. OTO Open 2023;7(3):e67. [doi: 10.1002/oto2.67] [Medline: 37614494]
11. Regulation (EU) 2016/679 of the European parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/EC (general data protection regulation) (text with EEA relevance). 2016 URL: https://eur-lex.europa.eu/eli/reg/2017/745/oj/eng [accessed 2025-11-03]
12. Health insurance portability and accountability act of 1996. 1996 URL: https://www.hhs.gov/hipaa/for-professionals/privacy/index.html [accessed 2025-11-03]

XSL•FO
RenderX

13. Turing AM. I.—Computing machinery and intelligence. Mind 1950 Oct 1;LIX(236):433-460. [doi: 10.1093/mind/LIX.236.433]
14. Teixeira-Marques F, Medeiros N, Nazaré F, et al. Exploring the role of ChatGPT in clinical decision-making in otorhinolaryngology: a ChatGPT designed study. Eur Arch Otorhinolaryngol 2024 Apr;281(4):2023-2030. [doi: 10.1007/s00405-024-08498-z] [Medline: 38345613]
15. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016 Jun;15(2):155-163. [doi: 10.1016/j.jcm.2016.02.012] [Medline: 27330520]
16. Regulation (EU) 2017/745 of the European parliament and of the council of 5 April 2017 on medical devices, amending directive 2001/83/EC. 2017 URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj [accessed 2025-11-03]
17. Policy for device software functions and mobile medical applications guidance for industry and food and drug administration staff. : FDA; 2022 URL: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/policy-device-software-functions-and-mobile-medical-applications [accessed 2025-11-03]
18. Esmaeilzadeh P, Mirzaei T, Dharanikota S. Patients' perceptions toward human-artificial intelligence interaction in health care: experimental study. J Med Internet Res 2021 Nov 25;23(11):e25856. [doi: 10.2196/25856] [Medline: 34842535]
19. Reis M, Reis F, Kunde W. Influence of believed AI involvement on the perception of digital medical advice. Nat Med 2024 Nov;30(11):3098-3100. [doi: 10.1038/s41591-024-03180-7] [Medline: 39054373]

Abbreviations

AI: : artificial intelligence
LMM: large language models
ORL: otorhinolaryngology

Edited by J Sarvestan; submitted 03.05.25; peer-reviewed by M Lotfinia; revised version received 08.10.25; accepted 09.10.25; published 25.11.25.

Please cite as:
Buhr CR, Seifen C, Bahr-Hamm K, Huppertz T, Pordzik J, Smith H, Kelsey T, Blaikie A, Matthias C, Kuhn S, Eckrich J
Evaluating Locally Run Large Language Models (Gemma 2, Mistral Nemo, and Llama 3) for Outpatient Otorhinolaryngology Care: Retrospective Study
JMIR Form Res 2025;9:e76896
URL: https://formative.jmir.org/2025/1/e76896
doi:10.2196/76896

© Christoph Raphael Buhr, Christopher Seifen, Katharina Bahr-Hamm, Tilman Huppertz, Johannes Pordzik, Harry Smith, Tom Kelsey, Andrew Blaikie, Christoph Matthias, Sebastian Kuhn, Jonas Eckrich. Originally published in JMIR Formative Research (https://formative.jmir.org), 25.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Populations and Health Domains Served by Direct-to-Consumer Digital Health Companies in the United States, 2011-2023: Cross-Sectional Study

Ashwini Nagappan[1], MBE, PhD; Xi Zhu[1], PhD; Corrina Moucheraud[2], MPH, ScD; Olivia S Jung[1], PhD

[1]Department of Health Policy and Management, Fielding School of Public Health, University of California, Los Angeles, 650 Charles Young Dr. S, Los Angeles, CA, United States

[2]Department of Public Health Policy and Management, School of Global Public Health, New York University, New York, NY, United States

Corresponding Author:
Ashwini Nagappan, MBE, PhD
Department of Health Policy and Management, Fielding School of Public Health, University of California, Los Angeles, 650 Charles Young Dr. S, Los Angeles, CA, United States

Abstract

Background: Direct-to-consumer (DTC) digital health companies, offering services such as on-demand prescriptions, mental health apps, fertility tracking, and at-home diagnostics, have become more common in the United States. These companies represent a shift in health care delivery by engaging consumers directly and operating largely outside of traditional health care systems. Despite their increasing presence, little is known about the populations that these companies serve, the health domains they address, and the technologies they use. Understanding these characteristics is critical for evaluating the quality of services provided, implications for health care costs, and impact on health equity.

Objective: This study aimed to describe the growth and focus of DTC digital health companies in the United States from 2011 to 2023, examining their target populations, health domains, and differentiating technologies.

Methods: We conducted a cross-sectional descriptive analysis using the Rock Health Digital Health Venture Funding Database, which systematically tracks US digital health companies that have received at least US $2 million in publicly disclosed funding. This database was selected because of its scope, consistency, and detailed coding of company characteristics. Of the 2652 digital health companies identified between 2011 and 2023, 478 (18.0%) were classified as exclusively pursuing a DTC model. We extracted and validated data on company characteristics, including founding year, operational status, funding levels, target populations, health domains, and technologies used. Descriptive analyses of frequencies, medians, and IQRs were conducted.

Results: Between 2011 and 2023, the number of DTC digital health companies grew steadily, with the highest number founded in 2020 (59/478, 12.3%). As of 2023, 445 (93.1%) of the 478 companies remained active, and 6.9% (n=33) had ceased operations. Across all 478 companies, total venture funding ranged from US $2 million to US $570 million (median US $9.6 million, IQR US $4.0-$25.0 million). Companies focusing on rural or Medicaid populations (n=10, 2.1%) were rare and had lower median funding (median US $5.0 million, IQR US $3.5-$13.4 million). Women were the most targeted population (n=70, 14.6%), followed by children and adolescents (n=36, 7.5%), and older adults (n=25, 5.2%). Mental health was the most common health domain (n=80, 16.7%), followed by reproductive and maternal health (n=71, 14.9%). Telemedicine (n=108, 22.6%), wearables and biosensors (n=93, 19.5%), and artificial intelligence or machine learning (n=63, 13.2%) were the most frequently adopted technologies, with their use varying by population and health domain.

Conclusions: As DTC digital health companies increasingly influence where and how care is delivered, systematic monitoring of their scope and characteristics is essential to evaluate whether they contribute to equitable access to care. Our findings provide a foundation for assessing whether these models are effectively addressing health needs, reaching diverse populations, and lowering health care costs.

(JMIR Form Res 2025;9:e78431) doi:10.2196/78431

KEYWORDS

digital health; direct-to-consumer; access; health equity; health services research

Introduction

Over the past decade, direct-to-consumer (DTC) digital health products and services, such as on-demand prescription services, mental health apps, fertility tracking, at-home diagnostic testing, health monitoring and tracking, and weight management platforms, have garnered increasing attention [1-3]. The growing presence of the DTC digital health market may be attributed to

XSL•FO
RenderX

shifting consumer preferences and technological advances [4], as well as the COVID-19 pandemic captivating interest among entrepreneurs and investors in consumer-oriented, consumer-driven, and remote health care solutions [5,6].

The increasing number of DTC digital health companies (Figure 1) reflects both unmet or emerging health care needs that may not be fulfilled by the traditional health care system and the ways in which companies actively create demand through direct marketing and consumer engagement [7,8]. However, there is a notable absence of scholarship understanding the populations and health domains prioritized by DTC digital health companies. A growing body of research has started to examine health care costs [9,10], consumer experiences [11-15], marketing practices [16-18], and ethical issues surrounding specific DTC digital health products and services [2,19]. Still, analyses of the DTC digital health market as a whole remain sparse, with few exceptions [1,20] and most insights coming from market reports and media coverage. Understanding who DTC digital health companies serve, what health needs they address, and how they integrate technology into care delivery is essential for assessing their implications for health care access and equity.

XSL•FO
RenderX

Figure 1. Direct-to-consumer digital health companies in operation (A) by year and (B) by target population in the United States from 2011 to 2023. LGBTQ: lesbian, gay, bisexual, transgender, and queer.

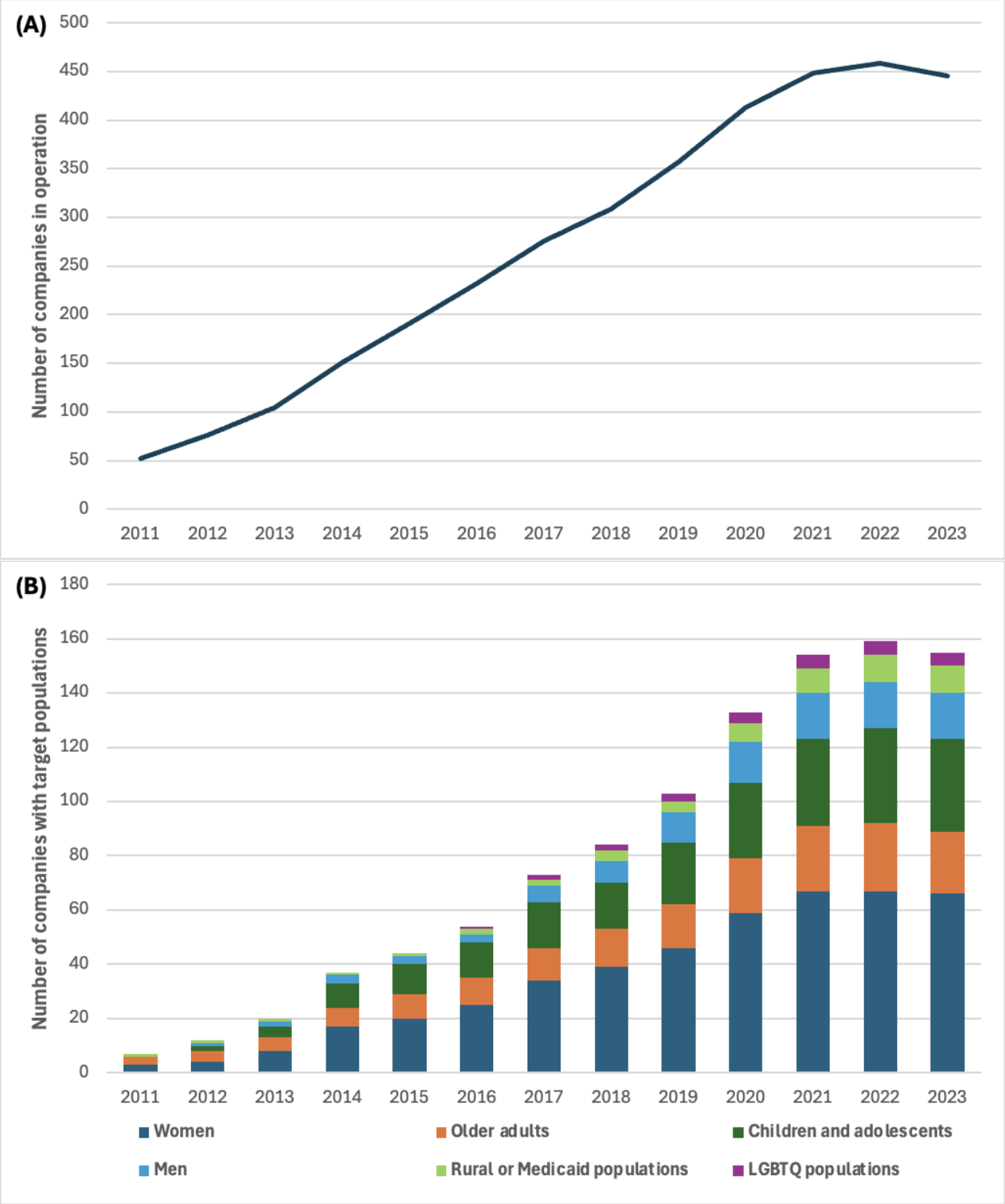

In this study, we used data from the Rock Health Digital Health Venture Funding Database to provide a snapshot of the growth of DTC digital health companies in the United States from 2011 to 2023, examining the populations and health domains they serve and the technologies they use. We sought to cultivate a broader understanding of which areas of care receive the most attention and resources to grasp how this modality of care delivery is being used to meet health needs.

Methods

Data Source

This study used the Rock Health Digital Health Venture Funding Database, maintained by Rock Health, a digital health venture capital and advisory firm. Since 2011, the database has systematically tracked digital health companies in the United States that have received over US $2 million in publicly

XSL•FO
RenderX

disclosed funding. During our study period (2011-2023), the database included 2652 digital health companies. The Venture Funding Database has been used as a primary data source in peer-reviewed studies [1,20-22] and is consistently referenced for US digital health funding trends [23-25]. While other venture funding databases (eg, PitchBook, CB Insights, Crunchbase) cover broader sectors or global markets, Rock Health's focus on digital health and its consistent application of inclusion criteria make it uniquely comprehensive for US digital health companies.

Our study relied on Rock Health's definition of "digital health companies," which frames them as "health companies that build and sell technologies—sometimes paired with a service, but only when the technology is, in and of itself, the service" [26]. This definition guided the scope of our sample, and it is broadly consistent with the US Food and Drug Administration's definition of "digital health technology" as computing platforms, connectivity, software, and sensors applied to health care and related uses [27]. Data are compiled from multiple public sources, including press releases, company websites, Securities and Exchange Commission (SEC) filings, and media reports, and are continuously updated and checked for accuracy. Each company was coded based on its most recent publicly available description at the time of data validation; the database does not systematically capture historical changes in target populations, health domains, or technologies over time.

Study Design

We conducted a cross-sectional descriptive study of digital health companies pursuing a DTC strategy. The digital health companies in this database serve several customer segments, such as consumers, payers, and employers. Given our interest in digital health companies pursuing a DTC strategy, companies were included in our sample if consumers were their exclusive customer segment. Companies that sell to payers, employers, and other segments either solely or in addition to consumers were excluded from the analytic sample. This resulted in a final sample of 478/2652 (18.0%) digital health companies in the database identified as exclusively pursuing a DTC model as of December 31, 2023.

Variables

This unique industry dataset captured various details about the companies, including the year in which they were founded, the year in which they ceased operations (if they did), total funding received from investors, target populations (eg, women, older adults), clinical indications addressed (eg, mental health, cardiovascular disease), and differentiating technologies used (eg, telemedicine, artificial intelligence [AI] or machine learning [ML]). To ensure data accuracy and verify the companies' operational status and the clinical indications they addressed, we conducted a thorough review of the 478 companies in our sample. First, this verification process involved cross-referencing company information with websites and public records to confirm operational status and offerings. Second, discrepancies were reviewed and reconciled by members of the research team.

About half of the companies in our sample were not tagged with a clinical indication as their product or service was not tied to a particular disease or symptom but was rather related to fitness, nutrition, or another nonclinical area. The product or service could also relate to a specific group of people with health needs (eg, older adults with fall risk). Thus, we created a new variable called "health domain" that was broader in scope than clinical indication and referred to an area of health or health care needs that a company aimed to address through its product or service. A company's health domain denoted the clinical indication, nonclinical indication (eg, fitness), or health needs served by the company's product or service.

Companies were not limited to a single category for target population, differentiating technology, or health domain. For example, a company could simultaneously serve older adults and lesbian, gay, bisexual, transgender, and queer (LGBTQ) individuals; use both AI or ML and wearable technologies; and address multiple health domains such as fall prevention and cardiovascular disease. Therefore, the reported frequencies reflect overlaps across categories.

Analysis

We conducted descriptive analyses of the DTC digital health companies to examine distributions and trends by company characteristics (eg, target populations, health domains, technologies used). This included summary statistics for continuous variables, such as total venture capital funding, as well as frequency distributions for company status, health domains, differentiating technologies, and target populations. We used the median and IQR to describe the center and spread of our data as the data were not normally distributed. Multimedia Appendix 1 describes all the variables in our dataset and how they were defined. We followed the Strengthening the Reporting of Observational Studies in Epidemiology reporting guidelines [28].

Ethical Considerations

This study involved secondary analysis of the Rock Health Digital Health Venture Funding Database, which contains company-level information collected from public sources. This study was determined to not meet the definition of human subjects research by the University of California, Los Angeles, Institutional Review Board. As the dataset contained no data on human participants, informed consent was not applicable. Privacy and confidentiality considerations were also not applicable as all records were at the company level and contained no personally identifiable or protected health information. No participants were recruited or compensated. Access to the Rock Health Digital Health Venture Funding Database was obtained through a formal data use agreement, which granted permission to use the data for academic research purposes.

Results

The number of DTC digital health companies in the United States grew steadily between 2011 and 2023 (Figure 1), with the largest number of companies founded in 2020 (59/478, 12.3%). As of 2023, 93.1% (n=445) of the 478 companies were

XSL•FO
RenderX

still in operation, with 6.9% (n=33) having ceased operations during our study period, including 42.4% (14/33) of these closures in 2023 alone. These closures were distributed across multiple health domains and target populations, without clear overrepresentation of any single category.

Figure 1 also shows that women were the most frequently targeted population, with consistent growth over the years. In 2023, 14.6% (n=70) of the 478 companies targeted women, 7.5% (n=36) targeted children and adolescents, 5.2% (n=25) targeted older adults, 3.6% (n=17) targeted men, and 2.1% (n=10) targeted rural or Medicaid populations. The number of companies focusing on men saw a slow but steady growth, particularly after 2016. Companies serving rural or Medicaid populations and LGBTQ populations began to emerge, albeit in small numbers, around 2016, signaling early steps toward more inclusive solutions in the DTC digital health space.

Table 1 shows substantial variations in funding across companies between 2011 and 2023, ranging from US $2 million to US $570 million. Companies focused on children and adolescents had the highest median funding (US $13.5 million, IQR US $6.3-$22.8 million). Companies focused on women (median US $9.4 million, IQR US $3.4-$25.0 million), men (median US $8.0 million, IQR US $6.0-$30.0 million), and older adults (median US $8.2 million, IQR US $5.2-$18.6 million) had median funding levels similar to the overall median. Companies targeting rural or Medicaid populations had consistently lower funding (median US $5.0 million, IQR US $3.5-$13.4 million; range US $2.5-$50.0 million). LGBTQ-focused companies also had a low median but exhibited much wider variation (median US $5.0 million, IQR US $3.0-$40.9 million; range US $3.0-$144.0 million), suggesting that, while most companies raised relatively modest amounts, a small number secured substantially larger deals.

Table . Venture funding across US direct-to-consumer (DTC) digital health companies by target population (2011-2023; N=478)[a].

	Companies, n (%)	Funding (million US $), median (IQR)	Funding range (million US $)
All companies	478 (100.0)	9.6 (4.0-25.0)	2.0-570.0
Women	70 (14.6)	9.4 (3.4-25.0)	2.0-229.5
Children and adolescents	36 (7.5)	13.5 (6.3-22.8)	2.2-129.7
Older adults	25 (5.2)	8.2 (5.2-18.6)	2.0-175.0
Men	17 (3.6)	8.0 (6.0-30.0)	2.0-202.3
Rural or Medicaid populations	10 (2.1)	5.0 (3.5-13.4)	2.5-50.0
LGBTQ[b] populations	5 (1.0)	5.0 (3.0-40.9)	3.0-144.0

[a]The total dataset includes 478 DTC digital health companies, with funding amounts ranging from US $2 million to US $570 million. The target population categories (eg, women or children and adolescents) represent subsets of these 478 companies, meaning that some companies served multiple populations whereas others did not have a target population. All dollar amounts are reported in nominal terms and are not adjusted for inflation.

[b]LGBTQ: lesbian, gay, bisexual, transgender, and queer.

DTC digital health companies targeted wide-ranging health domains (Table 2). Mental health was the most targeted health domain, with 16.7% (80/478) of all companies offering products or services for mental health conditions. This health domain was particularly prominent among companies focused on children and adolescents, with nearly 40% of them (14/36, 38.9%) addressing mental health. Reproductive and maternal health was the second most common health domain across all companies (71/478, 14.9%). Within population-focused companies, this health domain was represented in 88.6% (62/70) of the women-focused companies, 94.1% (16/17) of the men-focused companies, and 60.0% (3/5) of the LGBTQ-focused companies. The third most common health domain was fitness (68/478, 14.2%). This domain, similar to domains addressing nutrition, substance use, and oncology, was not linked to a specific target population. These findings highlight the variability in population focus across health domains, with some health issues attracting more population-specific attention, whereas others remained general in their scope.

XSL•FO
RenderX

Table . US direct-to-consumer digital health companies' health domain focus by target population (2011-2023; N=478)[a].

	Alln (%)	Women(n=70), n (%)	Children and adolescents(n=36), n (%)	Older adults(n=25), n (%)	Men(n=17), n (%)	Rural or Medicaid populations(n=10), n (%)	LGBTQ[b] populations(n=5), n (%)
Mental health	80 (16.7)	12 (17.1)	14 (38.9)	1 (4)	3 (17.6)	3 (30)	1 (20)
Reproductive and maternal health	71 (14.9)	62 (88.6)	7 (19.4)	0 (0)	16 (94.1)	2 (20)	3 (60)
Fitness	68 (14.2)	1 (1.4)	0 (0)	1 (4)	0 (0)	0 (0)	0 (0)
Weight management and obesity	42 (8.8)	5 (7.1)	0 (0)	0 (0)	1 (5.9)	0 (0)	0 (0)
Primary care	40 (8.4)	11 (15.7)	3 (8.3)	1 (4)	5 (29.4)	3 (30)	0 (0)
Cardiovascular disease	23 (4.8)	4 (5.7)	1 (2.8)	0 (0)	3 (17.6)	0 (0)	0 (0)
Neurology	21 (4.4)	1 (1.4)	0 (0)	3 (12)	2 (11.8)	0 (0)	0 (0)
Diabetes	20 (4.2)	2 (2.9)	0 (0)	1 (4)	1 (5.9)	0 (0)	0 (0)
Geriatrics	18 (3.8)	0 (0)	0 (0)	15 (60)	0 (0)	0 (0)	0 (0)
Musculoskeletal and pain management	16 (3.3)	2 (2.9)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Gastrointestinal	15 (3.1)	4 (5.7)	1 (2.8)	0 (0)	3 (17.6)	0 (0)	0 (0)
Substance use	15 (3.1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Developmental disorders	11 (2.3)	0 (0)	5 (13.9)	0 (0)	0 (0)	0 (0)	0 (0)
Dermatology	10 (2.1)	7 (10)	0 (0)	0 (0)	4 (23.5)	0 (0)	0 (0)
Oncology	10 (2.1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Pediatrics	10 (2.1)	0 (0)	9 (25)	0 (0)	0 (0)	1 (10)	0 (0)
Nutrition	6 (1.3)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Allergy and immunology	5 (1)	1 (1.4)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Pharmacy	5 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Pulmonary disorders	5 (1)	2 (2.9)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Audiology	4 (0.8)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Ophthalmology	4 (0.8)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Other	57 (11.9)	1 (1.4)	0 (0)	3 (12)	1 (5.9)	1 (10)	1 (20)

[a]The percentages in each column do not add up to 100 because some companies focused on multiple health domains. Each percentage represents the proportion of companies serving a given target population that addressed a specific health domain. Examples of "other" health domains included longevity care, wellness support, and supplements (see Multimedia Appendix 1 for more details).

[b]LGBTQ: lesbian, gay, bisexual, transgender, and queer.

Table 3 shows the technologies used by DTC digital health companies for each target population. Telemedicine was the most used technology, with 22.6% (108/478) of the companies using it. The second most used technology was wearables and biosensors (93/478, 19.5%), referring to devices worn on the body, such as smartwatches, fitness trackers, and other biosensors, that collect health-related data. This was followed by AI or ML technologies (63/478, 13.2%). None of the companies serving men, rural or Medicaid populations, or the LGBTQ community used wearables or AI or ML technologies. Additionally, none of the companies focused on rural or Medicaid populations used remote monitoring, which is surprising as it could be especially valuable for improving access to care and health outcomes in these communities.

The distribution of technologies across health domains shows how different technologies were applied to address various health needs (Multimedia Appendix 2). Telemedicine was primarily applied to reproductive and maternal health (30/108, 27.8%). Wearables and biosensors were primarily used in fitness (24/93, 25.8%), whereas AI, ML, or deep learning were

XSL•FO
RenderX

primarily applied to mental health (13/63, 20.6%), fitness (13/63, 20.6%), and gastrointestinal conditions (7/63, 11.1%), while remote monitoring was most often used in mental health (5/28, 17.9%) and geriatrics (5/28, 17.9%).

Table . Differentiating technologies by target population of US direct-to-consumer digital health companies (2011-2023; N=478)[a].

	Alln (%)	Women (n=70), n (%)	Children and adolescents(n=36), n (%)	Older adults(n=25), n (%)	Men (n=17), n (%)	Rural or Medicaid populations (n=10), n (%)	LGBTQ[b] populations (n=5), n (%)
Telemedicine	108 (22.6)	26 (37.1)	17 (47.2)	6 (24)	11 (64.7)	4 (40)	2 (40)
Wearables and biosensors	93 (19.5)	6 (8.6)	5 (13.9)	6 (24)	0 (0)	0 (0)	0 (0)
AI[c], ML[d], or deep learning	63 (13.2)	4 (5.7)	4 (11.1)	5 (20)	0 (0)	0 (0)	0 (0)
Remote monitoring	28 (5.9)	6 (8.6)	1 (2.8)	6 (24)	1 (5.9)	0 (0)	0 (0)
Genomics	27 (5.6)	5 (7.1)	1 (2.8)	0 (0)	0 (0)	0 (0)	0 (0)
Non–medical device hardware	26 (5.4)	2 (2.9)	3 (8.3)	0 (0)	0 (0)	0 (0)	0 (0)
Other	25 (5.2)	2 (2.9)	1 (2.8)	0 (0)	1 (5.9)	0 (0)	0 (0)
Digital medical devices	13 (2.7)	5 (7.1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
AR[e] and VR[f]	6 (1.3)	1 (1.4)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
IoT[g]	5 (1)	1 (1.4)	1 (2.8)	0 (0)	0 (0)	0 (0)	0 (0)
Robotics	4 (0.8)	0 (0)	1 (2.8)	1 (4)	0 (0)	0 (0)	0 (0)
Blockchain	2 (0.4)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)

[a]The percentages in each column do not add up to 100 because some companies used multiple technologies. Each percentage represents the proportion of companies serving a given target population that used a specific technology.

[b]LGBTQ: lesbian, gay, bisexual, transgender, and queer.

[c]AI: artificial intelligence.

[d]ML: machine learning.

[e]AR: augmented reality.

[f]VR: virtual reality.

[g]IoT: Internet of Things.

Discussion

Summary and Significance of the Findings

Using digital health funding data, we found that the number of DTC digital health companies grew steadily between 2011 and 2023. The number of companies targeting specific populations also increased, especially those focusing on women's health. In contrast, companies serving low-income, rural populations received considerably lower average funding, perhaps due to barriers such as limited broadband access [29] or lower out-of-pocket spending capacity. Throughout the study period, 6.9% (33/478) of the companies ceased operations, with nearly half (14/33, 42.4%) of those closures occurring in 2023. This clustering in 2023 may point to post-pandemic market corrections in digital health funding or shifts in consumer demand. Taken together, these findings illustrate how industry dynamics (eg, funding flows, consumer demand, company survival) shape the availability of DTC digital health and suggest areas where future policy attention could play a role in shaping equitable access, such as improving broadband coverage in rural and other underserved areas.

The prominence of mental health products and services among companies focused on children and adolescents indicates a responsiveness to the present-day youth mental health crisis [30]. The range of health domains addressed among DTC digital health companies revealed variation in focus, likely driven by the specific needs and demands of each population segment as well as market considerations such as total addressable market, payment models, and the perceived likelihood of attracting investment. While DTC digital health companies are interested in satisfying children's and young adults' unmet needs, this raises new questions about the requisite level of parental or guardian involvement in these services [31]. Specific policies may be needed to define the appropriate level of parent or guardian involvement to ensure access with safeguards to protect youth.

Nearly one-fifth of DTC digital health companies (93/478, 19.5%) used wearables and biosensors, and 13.2% (63/478)

XSL•FO
RenderX

applied AI or ML, which are technologies that may support real-time monitoring and personalization. However, the collection and sharing of consumer-generated health data also raise privacy concerns, particularly in areas such as reproductive health where data misuse could lead to discrimination or legal consequences [32]. Our findings underscore the importance of monitoring how DTC digital health companies address consumer data sharing concerns [33] and whether they proactively address gaps in privacy protections without waiting for consumer health data regulations to catch up.

Future research could extend this work by examining how the strategies of specific kinds of DTC digital health companies evolve over time, as well as comparisons between companies that exclusively pursue a DTC strategy (ie, companies in our sample) and those that also serve other customer segments such as payers and employers. In addition, assessing the relationship between company characteristics and their success or survival would provide greater insights into the factors that influence which models endure and which companies cease operations. Further, additional work could assess what proportion of the DTC digital health market is represented by venture-backed start-ups versus established incumbents, particularly as this balance is shifting with increasing entry by large firms [34]; understanding these dynamics will be important for evaluating implications for competition, regulation, and equitable access.

Comparison to Prior Work

Existing scholarship on DTC digital health has focused on particular kinds of offerings rather than on the sector as a whole. For example, studies on DTC telemedicine platforms have examined issues such as prescribing practices, continuity of care, and regulatory oversight [35]. Similarly, research on mental health apps and fertility-tracking tools has raised concerns about data privacy and effectiveness [36]. While these studies provide important insights into specific domains, less attention has been paid to systematically characterizing the broader DTC digital health landscape across populations, health domains, and differentiating technologies.

Our study differs in scope. Rather than analyzing a single type of offering, we provide a descriptive overview of the DTC digital health sector at large, drawing attention to who companies serve, which health domains they target, and what technologies they used. While our findings echo themes raised in previous domain-specific work (eg, the growth of mental health tools, concerns about equity), they also extend the conversation by situating these developments within the larger structure of the DTC digital health sector. In this way, this paper contributes empirical grounding for ongoing debates about the implications of DTC health for access, equity, and consumer protection.

Limitations

This study had the following limitations. First, the companies listed in the database were restricted to those that have publicly disclosed investment deals of at least US $2 million and meet Rock Health's definition of digital health. As a result, the data did not capture major technology companies (eg, Amazon, Apple) [37,38] or pharmaceutical companies (eg, Eli Lilly and Company, Pfizer) [39] that have ventured into DTC health and may underrepresent smaller seed-stage or bootstrapped firms that are part of the DTC digital health ecosystem. Second, as our analysis was limited to companies focusing exclusively on selling to consumers, our findings do not allow for comparisons with companies that offer insurance coverage. This distinction has equity implications as out-of-pocket costs may be a substantial barrier for underserved populations and influence the extent to which these groups can access DTC digital health services. Third, our dataset extended only through 2023. As additional years of data become available, future research could examine whether the trends we identified persist or shift and directly compare post-2023 developments with our findings.

Conclusions

The DTC digital health market, now over a decade old, has reached an inflection point for assessing the needs it is fulfilling and who it is serving. Our findings provide a descriptive account of the well-established areas of DTC digital health. As individuals continue to receive care in DTC digital health settings outside of traditional health systems, it will be increasingly important to evaluate whether these offerings indeed improve outcomes and lower costs.

Acknowledgments

The authors thank Dr Anna Wexler and Dr Beth Glenn for their insightful comments on the paper.

Funding

AN was supported by a National Institutes of Health/National Center for Advancing Translational Science University of California, Los Angeles, Clinical and Translational Science Institute grant (TL1TR001883). The content is solely the responsibility of the authors and does not represent the official views of the National Institutes of Health.

Data Availability

The datasets generated or analyzed during this study are not publicly available because data were used under license for this study, but are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Authors' Contributions

AN conceptualized and designed the study, analyzed the data, oversaw all phases of the project, and contributed to manuscript writing. XZ supported data analysis, contributed to drafting the manuscript, and critically reviewed it for intellectual content. CM provided feedback on multiple drafts and critically reviewed the manuscript for intellectual content. OJ co-led the conceptualization of the study, supervised data analysis, and was a major contributor to writing the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

AN reports fees from Rock Health, Inc, outside of this study. All other authors declare no other conflicts of interest.

Multimedia Appendix 1
Study variables.
[DOCX File, 68 KB - formative_v9i1e78431_app1.docx]

Multimedia Appendix 2
Differentiating technologies across health domains of direct-to-consumer digital health companies.
[DOCX File, 65 KB - formative_v9i1e78431_app2.docx]

Checklist 1
STROBE checklist.
[PDF File, 48 KB - formative_v9i1e78431_app3.pdf]

References

1. Cohen AB, Dorsey ER, Mathews SC, Bates DW, Safavi K. A digital health industry cohort across the health continuum. NPJ Digit Med 2020 May 12;3:68. [doi: 10.1038/s41746-020-0276-9] [Medline: 32411829]
2. Nagappan A, Kalokairinou L, Wexler A. Ethical issues in direct-to-consumer healthcare: a scoping review. PLOS Digit Health 2024 Feb 13;3(2):e0000452. [doi: 10.1371/journal.pdig.0000452] [Medline: 38349902]
3. Mandl KD. How AI could reshape health care-rise in direct-to-consumer models. JAMA 2025 May 20;333(19):1667-1669. [doi: 10.1001/jama.2025.0946] [Medline: 39992661]
4. Knowles M, Krasniansky A, Nagappan A. The new era of consumer engagement: insights from Rock Health's ninth annual consumer adoption survey. Rock Health. 2024 Mar 18. URL: https://rockhealth.com/insights/the-new-era-of-consumer-engagement-insights-from-rock-healths-ninth-annual-consumer-adoption-survey/ [accessed 2024-03-18]
5. Bryant K, Knowles M, Krasniansky A. 2022 year-end digital health funding: lessons at the end of a funding cycle. Rock Health. 2023 Jan 9. URL: https://rockhealth.com/insights/2022-year-end-digital-health-funding-lessons-at-the-end-of-a-funding-cycle/ [accessed 2023-08-25]
6. Krasniansky A, Evans B, Zweig M. 2021 year-end digital health funding: seismic shifts beneath the surface. Rock Health. 2022 Jan 10. URL: https://rockhealth.com/insights/2021-year-end-digital-health-funding-seismic-shifts-beneath-the-surface/ [accessed 2024-10-24]
7. Mintzes B. Direct to consumer advertising is medicalising normal human experience. BMJ 2002 Apr 13;324(7342):908-911. [doi: 10.1136/bmj.324.7342.908] [Medline: 11950745]
8. Makary MA. The FDA's overdue crackdown on misleading pharmaceutical advertisements. JAMA 2025 Sep 12. [doi: 10.1001/jama.2025.18197] [Medline: 40938594]
9. Ashwood JS, Mehrotra A, Cowling D, Uscher-Pines L. Direct-to-consumer telehealth may increase access to care but does not decrease spending. Health Aff (Millwood) 2017 Mar 1;36(3):485-491. [doi: 10.1377/hlthaff.2016.1130] [Medline: 28264950]
10. Li KY, Zhu Z, Ng S, Ellimoottil C. Direct-to-consumer telemedicine visits for acute respiratory infections linked to more downstream visits. Health Aff (Millwood) 2021 Apr;40(4):596-602. [doi: 10.1377/hlthaff.2020.01741] [Medline: 33819099]
11. Wexler A, Nagappan A, Beswerchij A, Choi R. Direct-to-consumer orthodontics: surveying the user experience. J Am Dent Assoc 2020 Aug;151(8):625-636. [doi: 10.1016/j.adaj.2020.02.025] [Medline: 32718492]
12. Kyweluk MA. Quantifying fertility? Direct-to-consumer ovarian reserve testing and the new (in)fertility pipeline. Soc Sci Med 2020 Jan;245:112697. [doi: 10.1016/j.socscimed.2019.112697] [Medline: 31786460]
13. Shih P, Prokopovich K, Degeling C, Street J, Carter SM. Direct-to-consumer detection of atrial fibrillation in a smartwatch electrocardiogram: medical overuse, medicalisation and the experience of consumers. Soc Sci Med 2022 Jun;303:114954. [doi: 10.1016/j.socscimed.2022.114954] [Medline: 35569232]
14. Singh J, Dhar S. Assessment of consumer attitudes following recent changes in the US hearing health care market. JAMA Otolaryngol Head Neck Surg 2023 Mar 1;149(3):247-252. [doi: 10.1001/jamaoto.2022.4344] [Medline: 36656571]

XSL•FO
RenderX

15. Gollust SE, Gray SW, Carere DA, et al. Consumer perspectives on access to direct-to-consumer genetic testing: role of demographic factors and the testing experience. Milbank Q 2017 Jun;95(2):291-318. [doi: 10.1111/1468-0009.12262] [Medline: 28589610]
16. Desai S, Manjaly P, Lee K, et al. Assessment of sales and marketing of online vouchers for discounted direct-to-consumer medical imaging services. JAMA Intern Med 2021 Feb 1;181(2):267-269. [doi: 10.1001/jamainternmed.2020.3529] [Medline: 33136117]
17. Crane MA, DiStefano MJ, Moore TJ. False or misleading claims in online direct-to-consumer ketamine advertising in Maryland. JAMA Netw Open 2023 Nov 1;6(11):e2342210. [doi: 10.1001/jamanetworkopen.2023.42210] [Medline: 37934500]
18. Johnson A, Thompson R, Nickel B, Shih P, Hammarberg K, Copp T. Websites selling direct-to-consumer anti-mullerian hormone tests. JAMA Netw Open 2023 Aug 1;6(8):e2330192. [doi: 10.1001/jamanetworkopen.2023.30192] [Medline: 37603332]
19. Kalokairinou L, Zettler PJ, Nagappan A, Kyweluk MA, Wexler A. The promise of direct-to-consumer COVID-19 testing: ethical and regulatory issues. J Law Biosci 2020 Sep 23;7(1):lsaa069. [doi: 10.1093/jlb/lsaa069] [Medline: 33569177]
20. Cohen AB, Mathews SC, Dorsey ER, Bates DW, Safavi K. Direct-to-consumer digital health. Lancet Digit Health 2020 Apr;2(4):e163-e165. [doi: 10.1016/S2589-7500(20)30057-1] [Medline: 33328077]
21. Day S, Shah V, Kaganoff S, Powelson S, Mathews SC. Assessing the clinical robustness of digital health startups: cross-sectional observational analysis. J Med Internet Res 2022 Jun 20;24(6):e37677. [doi: 10.2196/37677] [Medline: 35723914]
22. Safavi KC, Cohen AB, Ting DY, Chaguturu S, Rowe JS. Health systems as venture capital investors in digital health: 2011-2019. NPJ Digit Med 2020;3:103. [doi: 10.1038/s41746-020-00311-5] [Medline: 32802968]
23. Deal M. Digital health funding strong in 2024, with AI a key investment area. Chartis. 2024 Aug 9. URL: https://www.chartis.com/insights/digital-health-funding-strong-2024-ai-key-investment-area [accessed 2025-09-30]
24. Anderson M. How digital health VC funding evolved in 2024. Healthcare Brew. 2025 Jan 15. URL: https://www.healthcare-brew.com/stories/2025/01/15/digital-health-vc-funding-evolved-2024 [accessed 2025-09-30]
25. Dorsey ER, Topol EJ. State of telehealth. N Engl J Med 2016 Jul 14;375(2):154-161. [doi: 10.1056/NEJMra1601705] [Medline: 27410924]
26. Gandhi M. What digital health is (and isn't). Rock Health. 2013. URL: https://rockhealth.com/what-digital-health-is-and-isnt/ [accessed 2023-08-25]
27. Digital health technologies for remote data acquisition in clinical investigations. U.S. Food and Drug Administration. 2023 Dec. URL: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/digital-health-technologies-remote-data-acquisition-clinical-investigations [accessed 2025-09-30]
28. von Elm E, Altman DG, Egger M, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. BMJ 2007 Oct 20;335(7624):806-808. [doi: 10.1136/bmj.39335.541782.AD] [Medline: 17947786]
29. Curtis ME, Clingan SE, Guo H, Zhu Y, Mooney LJ, Hser YI. Disparities in digital access among American rural and urban households and implications for telemedicine-based services. J Rural Health 2022 Jun;38(3):512-518. [doi: 10.1111/jrh.12614] [Medline: 34355427]
30. Youth mental health. U.S. Department of Health and Human Services. URL: https://www.hhs.gov/surgeongeneral/priorities/youth-mental-health/index.html [accessed 2024-10-25]
31. Bounds DT, Stiles-Shields C, Schueller SM, Odgers CL, Karnik NS. Ethical considerations for developing pediatric mHealth interventions for teens with socially complex needs. J Child Adolesc Psychiatr Nurs 2023 Feb;36(1):7-16. [doi: 10.1111/jcap.12396] [Medline: 36134754]
32. Fowler LR, Ulrich MR. Femtechnodystopia. Stan L Rev 2023;75:1233. [doi: 10.2139/ssrn.4099764]
33. Nagappan A, Zhu X. Patterns of willingness to share health data with key stakeholders in US consumers: a latent class analysis. J Am Med Inform Assoc 2025 Apr 1;32(4):702-711. [doi: 10.1093/jamia/ocaf014] [Medline: 39873672]
34. Rome BN. Direct-to-consumer drug company pharmacies. JAMA 2024 Mar 26;331(12):1003-1004. [doi: 10.1001/jama.2024.2911] [Medline: 38412063]
35. Uscher-Pines L, Mulcahy A, Cowling D, Hunter G, Burns R, Mehrotra A. Access and quality of care in direct-to-consumer telemedicine. Telemed J E Health 2016 Apr;22(4):282-287. [doi: 10.1089/tmj.2015.0079] [Medline: 26488151]
36. Zadushlivy N, Biviji R, Williams KS. Exploration of reproductive health apps' data privacy policies and the risks posed to users: qualitative content analysis. J Med Internet Res 2025 Mar 5;27:e51517. [doi: 10.2196/51517] [Medline: 40053713]
37. Anderson M. Amazon's One Medical forms new telehealth service to compete with DTC players. Healthcare Brew. 2024 Nov 22. URL: https://www.healthcare-brew.com/stories/2024/11/21/amazons-one-medical-new-telehealth-service-compete-dtc-players [accessed 2025-09-30]
38. Apple introduces groundbreaking health features to support conditions impacting billions of people. Apple. 2024 Sep 9. URL: https://www.apple.com/newsroom/2024/09/apple-introduces-groundbreaking-health-features/ [accessed 2025-09-30]
39. Anderson M. Why pharma companies are going DTC. September. 2024. URL: https://www.healthcare-brew.com/stories/2024/09/16/pharma-companies-going-dtc [accessed 2025-09-30]

XSL•FO
RenderX

Abbreviations

AI: artificial intelligence
DTC: direct-to-consumer
LGBTQ: lesbian, gay, bisexual, transgender, and queer
ML: machine learning

Edited by A Mavragani; submitted 02.06.25; peer-reviewed by B Wills, S Bollmeier; revised version received 02.11.25; accepted 03.11.25; published 26.11.25.

Please cite as:
Nagappan A, Zhu X, Moucheraud C, Jung OS
Populations and Health Domains Served by Direct-to-Consumer Digital Health Companies in the United States, 2011-2023: Cross-Sectional Study
JMIR Form Res 2025;9:e78431
URL: https://formative.jmir.org/2025/1/e78431
doi:10.2196/78431

© Ashwini Nagappan, Xi Zhu, Corrina Moucheraud, Olivia S Jung. Originally published in JMIR Formative Research (https://formative.jmir.org), 26.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Acceptability, Relevance, and Short-Term Outcomes of the STAC-T Bullying Bystander App: Feasibility Quantitative Study

Diana M Doumas[1,2], PhD; Aida Midgett[1,2], EdD; Robin Hausheer[3], EdD; Amanda Winburn[4], PhD; Mary K Buller[5], MA; Taylor Perron[1], BA; Jennalyn Shelton[1], BA; Brandon Herbeck[5], BA

[1]Department of Counselor Education, Boise State University, 1910 University Drive, Boise, ID, United States

[2]Institute for the Study of Behavioral Health and Addiction, Boise State University, Boise, ID, United States

[3]Department of Counseling, Human Development and Family Science, University of Vermont, Burlington, VT, United States

[4]Department of Leadership and Counselor Education, University of Mississippi, Oxford, MS, United States

[5]Klein Buendel Inc., Golden, CO, United States

Corresponding Author:
Diana M Doumas, PhD
Department of Counselor Education, Boise State University, 1910 University Drive, Boise, ID, United States

Abstract

Background: Bullying is a significant public health issue, with approximately 25% of middle school students reporting being a target of bullying in the past year. Students who are targets of bullying or witness bullying are at high risk for negative mental health outcomes, including depression and anxiety. STAC is an evidence-based bullying bystander intervention for middle school students, with program outcomes that include reductions in bullying perpetration and victimization, as well as associated mental health risks. We developed a technology-based version of STAC (STAC-T) to reduce implementation barriers associated with in-person bullying prevention programs. STAC-T is an interactive app that includes a 40-minute training and a 15-minute booster session.

Objective: This study aimed to evaluate the acceptability and relevance of the STAC-T program. We were also interested in how program acceptability and relevance are related to the use of specific bystander intervention skills (eg, STAC strategies) students learn in the program.

Methods: This study was part of a larger study in which students recruited from 6 middle schools in rural, low-income communities in the United States were randomized to either the STAC-T intervention or a control condition. Participants in this study were 229 students in the intervention group who completed the 30-day follow-up survey, including the acceptability and relevance questionnaire. The survey assessed program acceptability and relevance, whether or not students witnessed bullying posttraining, and the use of the STAC strategies to intervene in bullying situations. Descriptive statistics were used to assess acceptability, relevance, and the use of STAC strategies. Linear regression analysis was used to assess the relationship of program acceptability and relevance to STAC strategy use.

Results: Of the 229 student participants, the majority reported the program was acceptable (188, 82.1%, to 206, 90.0%) and relevant (180, 78.6%, to 190, 83.0%) for students at their school. Of the 54.6% (125/229) of students who witnessed bullying posttraining, 88.8% (111/125) reported the use of at least one STAC strategy to intervene when witnessing bullying. Students were most likely to use the STAC bystander intervention strategies *Turning it Over* and *Accompanying Others*, relative to *Stealing the Show* and *Coaching Compassion*. Regression analyses revealed that program relevance was a significant predictor of posttraining use of STAC strategies (P=.016). In contrast, program acceptability was not a significant predictor of posttraining STAC strategy use (P=.660).

Conclusions: This study provides support for the acceptability and relevance of STAC-T and its effectiveness in promoting the use of the STAC strategies to intervene in bullying situations. Furthermore, program relevance was related to STAC strategy use, highlighting the importance of assessing program relevance for specific student populations.

(JMIR Form Res 2025;9:e76830) doi:10.2196/76830

KEYWORDS

technology-based bullying intervention; bullying; acceptability; relevance; middle school

XSL•FO
RenderX

Introduction

Background

Bullying is a serious problem among youth in the United States. Bullying has been described as an often-repeated, unwanted, intentional aggressive behavior that takes place within the context of a relationship with a perceived power imbalance [1,2]. Bullying can occur in the forms of physical (eg, hitting, kicking, and damaging personal property), verbal (eg, name calling and threats), relational (eg, spreading rumors and social exclusion), and online (eg, cyberbullying) behavior [3]. Bullying peaks in middle school, with 26.3% of middle school students reporting bullying victimization in the past year [4]. Comprehensive, school-wide bullying programs are effective in reducing bullying [5]; however, comprehensive, school-wide programs require significant resources, creating barriers to implementation [6], particularly for schools in rural and low-income communities [7]. Technology-based interventions provide a promising platform for bullying bystander prevention programming, as these programs can be easily disseminated to large groups of students at little cost relative to in-person, comprehensive, school-wide bullying prevention programs. Thus, technology-based interventions have the potential to improve access to programming and decrease implementation barriers, particularly in rural communities [7].

Bystander Training

Students who witness bullying as bystanders can engage in bullying in several ways, including taking on the following roles: (1) *assistants* who join in the bullying behavior, directly helping the student who is bullying; (2) *reinforcers* who encourage the student who is bullying by laughing or cheering; (3) *outsiders* who ignore or walk away from bullying situations; and (4) *defenders* who attempt to intervene in the bullying situation [8]. Research suggests that approximately 80% of students witness bullying as bystanders [9], yet only 20% intervene in bullying situations [10]. Students may not intervene because they do not know what to do when witnessing bullying [11] or they may intervene in maladaptive ways (eg, aggressive behavior) [12] and become socially rejected by peers [8]. Thus, bystander training is essential for students' acquisition of prosocial skills they can use to intervene in bullying situations. However, many comprehensive, school-wide programs do not include bystander training [5].

The STAC Bullying Bystander Intervention

STAC [13] is a brief, in-person bullying bystander intervention developed to teach students who witness bullying how to intervene using prosocial skills. The training is 75 minutes and includes education about bullying, consequences associated with bullying, the 4 bystander roles, and the 4 STAC strategies: (1) *Stealing the Show*, using humor or distraction to interrupt the bullying situation; (2) *Turning it Over*, informing an adult about the bullying incident; (3) *Accompanying Others*, providing support to the targeted student; and (4) *Coaching Compassion*, gently confronting the student who bullied to increase empathy for the target. The training also includes an experiential component that includes role-plays during which students are presented with bullying scenarios and then practice using the STAC strategies to intervene in the bullying situation. The STAC training is followed by two 15-minute booster sessions to reinforce skill acquisition learning. Research indicates the STAC intervention is effective in decreasing bullying victimization [14-16] and perpetration [15,17,18] and reducing mental health risks [17-23] among students trained in the program.

The STAC-T App

The STAC-T app is a technology-based version of the in-person STAC bullying bystander intervention. The STAC-T app was developed to address specific needs and barriers identified by school personnel in rural, low-income communities [24]. Prototype development included focus groups and interviews with school personnel and students followed by usability testing [25]. The resulting fully built-out STAC-T app is a 40-minute training followed by one 15-minute booster session. Initial research on the fully built-out STAC-T app indicates that the STAC-T app has high usability, acceptability, and relevance among key stakeholders in rural, low-income communities [26]. Specifically, 10 school personnel and 11 middle school students in 2 states completed usability testing and individual interviews. Quantitative results indicated that both school personnel and students found the program to be highly usable (eg, easy to use, acceptable, and feasible) and reported high program satisfaction. Qualitative findings supported these results, with both school personnel and students providing feedback that the STAC-T app was useful, relevant, and appropriate.

This Study

When developing intervention programs, it is important to establish acceptability (eg, how the user perceives the program) and program relevance (eg, social validity), as acceptability is related to successful program outcomes [27] and perceived relevance by the end user is associated with the use of intervention skills [28]. Although research conducted in the STAC-T development provides initial support for program usability, program acceptability and relevance were assessed using a small sample (n=21) and were not assessed in relation to use of skills posttraining. Thus, the purpose of this study was to evaluate the acceptability and relevance of the STAC-T app, as well as posttraining use of the STAC strategies, among a large, more diverse sample of middle school students. To achieve this aim, we assessed program acceptability, relevance, and STAC strategy use among students who completed the 30-day follow-up assessment and who were randomly assigned to the intervention group in a larger efficacy trial (n=249). Acceptability, relevance, and strategy use were evaluated after completion of both STAC-T training modules and the booster session. Furthermore, as previous research suggests, acceptability is positively related to program outcomes [27] and relevance is positively related to the use of intervention skills [28], we were also interested in how program acceptability and relevance were related to the use of the STAC strategies posttraining. The study had the following objectives: (1) to assess acceptability and relevance of the STAC-T app, (2) to assess STAC strategy use posttraining, and (3) to evaluate the relationship of acceptability and relevance to use of the STAC strategies.

XSL•FO
RenderX

Methods

Participants

Participants were recruited as part of a larger study (Clinicaltrials.gov; NCT05572398 [29]) conducted at 6 middle schools in rural, low-income communities in Northwest, Southern, and Eastern regions of the United States. Parents of all students in sixth through eighth grade (N=1802) were sent consent forms. A total of 40.84% (736/1802) of students provided consent. Of the 736 students, 34.0% (n=613) agreed to participate in the study, and 51.1% (n=307) were assigned to the intervention condition. Among students in the intervention condition, 81.1% (249/307) completed the 30-day follow-up survey and were included in this study. Participant age ranged from 11 to 15 (mean 12.1, SD 0.9) years, with 39.8% (99/249) in grade 6, 43.0% (107/249) in grade 7, and 17.3% (43/249) in grade 8. Students self-reported gender as female (136/249, 54.6%) and male (100/249, 40.2%), with 5.2% (13/249) missing data. Students self-reported ethnicity/racial background as White (108/249, 43.4%), Hispanic or Latino (60/249, 24.1%), Black or African American (46/249, 18.5%), more than one race (21/249, 8.4%), Asian American (3/249, 1.2%), American Indian or Alaska Native (3/249, 1.2%), Native Hawaiian or Other Pacific Islander (1/249, 0.4%), and other (4/249, 1.6%), with 1.2% (3/249) missing data.

Procedures

Participant recruitment occurred during the fall of 2024. Recruitment procedures were conducted through a common core course for each of the 3 grade levels. All parents of sixth- through eighth-grade students received a prenotification email from the school principal informing them of the study. Moreover, research team members hand-delivered a cover letter explaining the study and a copy of the informed consent to students. Students were asked to provide their parents with the documents and to return informed consent documents to their teacher. The school also sent a notification email to parents with a cover letter and informed consent form with a link for parents to sign providing consent electronically. Teachers reminded students to return the informed consent forms. The school also sent parents a reminder email with a link to sign the informed consent form electronically from the principal. For students whose parents consented, a research team member conducted assent immediately before baseline data collection.

Class periods were randomly assigned to the STAC-T intervention or assessment-only control condition using block randomization [30]. Block randomization ensures relatively equivalent group sizes by dividing participants into blocks and randomly assigning condition within each block. In this study, the blocks were the number of periods at the school. Block randomization was conducted for each school as the number of periods varied by school, ranging from 2 to 8 periods. Across the 6 schools, there were 29 periods, with 15 periods randomized to the control condition and 14 periods randomized to the intervention condition. A member of the research team went into each participating classroom and described the research study and invited the students with parental consent to participate. Students were given a unique personal identification number (PIN) to maintain confidentiality and a URL to access the study website where they read a welcome screen explaining the research and were asked for their assent to participate. Once students gave assent by clicking “agree,” they were taken to a screen that asked them to enter their PIN and were then directed to begin the baseline survey, which took between 20 and 30 minutes. Students in the intervention condition completed the STAC-T intervention and booster session over the next 3 weeks. Students without parental consent and those who did not provide consent completed an alternate activity in the classroom. Students in the intervention condition who completed the baseline survey were invited to participate in the 30-day follow-up survey used in this study.

Ethical Considerations

All study procedures were approved by the Boise State University Institutional Review Board (101-SB21-205) and by the School Districts or Administration. We used active informed consent for this study. Parents were provided with an informed consent form that contained information about the study, including purpose and background, procedures, risks or discomforts, extent of confidentiality (eg, students provided a unique study PIN and study data deidentified), benefits, payment, and how to contact the primary investigator and the Boise State University Institutional Review Board. Students with parental consent were provided with an informed assent form that contained the same information. Incentives included a US $100 gift card for teachers and donuts, oranges, and stickers for students after completion of the follow-up survey.

The STAC-T App

The STAC-T app (Figure 1) is a 40-minute training followed by one 15-minute booster session. The educational modules include the following content: (1) What is bullying?—users are presented with background information on bullying, including definitions of bullying, bullying facts and statistics, characteristics of students who bully, and negative consequences associated with bullying; (2) What are bystanders?—this module explains the 4 bystander roles: (a) *assistants*: those who directly join in to help the bully, (b) *reinforcers*: those who encourage the bully, (c) *outsiders*: those who do nothing or walk away, and (d) *defenders*: those who do something to intervene; and (3) STAC strategies: users were introduced to the 4 STAC strategies: (a) *Stealing the Show*: using humor or distraction to interrupt the bullying situation, (b) *Turning it Over*: informing an adult about the bullying incident, (c) *Accompanying Others*: providing support to the targeted student, and (d) *Coaching Compassion*: gently confronting the student who bullied to increase empathy for target. Users are then guided through STAC strategy practice in which students select an avatar, view bullying scenarios, select each STAC strategy and view the avatar enacting the selected strategy, and receive feedback on strategy effectiveness. An artist hand-illustrated and styled 6 avatars with light-, medium-, and dark-colored hair in different styles, as well as light, medium, and dark skin tones for students to choose from to best represent themselves. The booster session included additional practice with bullying scenarios and STAC strategy use. “Badges” (visual reward icons; eg, *Stealing the*

XSL•FO
RenderX

Show Badge) are included throughout the app to reward learning and encourage user engagement and adherence.

Figure 1. Samples from the STAC-T app.

Measures

Acceptability and Relevance of the STAC-T App

Acceptability (ie, ease of use and utility) and relevance of the STAC-T app were assessed using a social validity survey that comprised 8 items [31]. Items were ranked on a 4-point Likert scale, ranging from 1 (*strongly disagree*) to 4 (*strongly agree*). Items were subjected to a principal component factor analysis to confirm the factors underlying the social validity survey. An oblique rotation was used to allow for correlation among the items. Eigenvalues showed that the first component (ie, acceptability) explained 69.0% of the variance, the second component (ie, relevance) explained 9.9% of the variance, and the remaining components had eigenvalues <1. A 2-factor solution, which explained 78.9% of the variance, was selected. Acceptability items were as follows: "The STAC-T training was easy to understand," "The STAC-T program was useful," "The STAC-T training was interesting," "I learned something from the STAC-T program," and "I would recommend the STAC-T program to other students at my school" (α=.92). Relevance items were as follows: "The STAC-T training information was relevant for students at my school," "The STAC-T training examples of bullying were relevant for schools like my school," and "The STAC-T strategy role-plays were relevant for students at my school" (α=.89).

Use of STAC Strategies

The use of each STAC strategy was measured by a single item. Students were asked, "How often would you say that you used these strategies to stop bullying in the past month? (a) Stealing the Show – using humor to get the attention away from the bullying situation, (b) Turning it Over – telling an adult about what you saw, (c) Accompanying Others – reaching out to the student who was the target of bullying, and (d) Coaching Compassion – helping the student who bullied develop empathy for the target." Items were rated on a 5-point Likert scale, ranging from 1 (*never/almost never*) to 5 (*always/almost always*). The items were summed to create a total scale (α=.81). Prior research has used these items to examine the use of STAC strategies posttraining with middle school students [17,32,33].

Bystander Status

We assessed bystander status with the following question: "Have you seen bullying at school in the past month?," with response choices of *Yes* and *No*. Researchers have used this item to measure bystander status among middle school students [16,17,34]. Using this measure, 54.6% (125/229) of students were classified as bystanders.

Data Analysis

Data from the questionnaires were analyzed using the Statistical Package for the Social Sciences (version 25.0). Before conducting statistical analyses, the data were examined for

XSL•FO
RenderX

outliers and normality, and all variables were within the normal range for skew and kurtosis. Of the 249 students who completed the follow-up survey, 92.0% (n=229) completed the acceptability and relevance questionnaire. The 8.0% (n=20) who did not complete the acceptance and relevance questionnaire were omitted from the data analyses. Missing data for the remaining participants were imputed using expectation-maximization. Missingness was <.05% for acceptability and relevance, <2% for use of STAC strategies, and 0% for bystander status. To evaluate intervention acceptability, relevance, and use of STAC strategies, frequency data were combined into dichotomous variables. For acceptability and relevance, "strongly disagree" and "disagree" were combined to form a disagree category, and "strongly agree" and "agree" were combined to create an agree category. For use of STAC strategies, responses were categorized as not used (score of 0) and used (score >0). To examine the relationship between intervention acceptability and relevance and use of STAC strategies, we conducted a linear multiple regression analysis using the original frequency data for acceptability, relevance, and total use of STAC strategies. We entered acceptability and relevance into the model simultaneously. We examined the variance inflation factor (VIF) for the predictors to assess multicollinearity. We calculated effect size using R^2 for the regression analysis, with magnitude of effects interpreted as follows: small=0.01, medium=0.09, and large=0.25 [35]. All analyses were considered significant at P=.04.

Results

Acceptability and Relevance

Means, SDs, and percent agreement for acceptability and relevance items are reported in Table 1. As seen in Table 1, the majority of students reported the program was acceptable (188/229, 82.1%, to 206/229, 90.0%) and relevant (180/229, 78.6%, to 190/229, 83.0%) for students at their school. All acceptability and relevance items were rated >3.00 on a 1- to 4-point scale.

Table . Participants reporting agreement with acceptability and relevance (N=229).

Item[a]	Values, mean (SD)	Agreement, n (%)
Acceptability (mean 3.22, SD 0.69)		
The STAC-T training was easy to understand	3.23 (0.75)	206 (90.0)
The STAC-T training was useful	3.20 (0.79)	196 (85.6)
The STAC-T training was interesting	3.18 (0.82)	193 (84.7)
I learned something from the STAC-T program	3.31 (0.80)	203 (88.6)
I would recommend the STAC-T program to other students at my school	3.22 (0.85)	188 (82.1)
Relevance (mean 3.07, SD 0.75)		
The STAC-T training information was relevant for students at my school	3.13 (0.80)	190 (83.0)
The STAC-T training examples of bullying were relevant for students at my school	3.09 (0.84)	190 (83.0)
The STAC-T strategy role-plays were relevant for students at my school	3.21 (0.83)	180 (78.6)

[a]Responses were scored on a 4-point Likert scale, ranging from 1=strongly disagree to 4=strongly agree.

Use of the STAC Strategies

Results for student self-reported use of each STAC strategy are presented in Table 2. Results are reported for students who reported witnessing bullying in the past 30 days (125/229, 54.6%). As seen in Table 2, the majority of students (111/125, 88.8%) used at least one STAC strategy posttraining, with higher rates reported for the strategies *Turning it Over* and *Accompanying Others* relative to *Stealing the Show* and *Coaching Compassion*.

Table . Number of students reporting use of strategies (n=125).

Strategies	Used strategy, n (%)	Did not use strategy, n (%)
Stealing the Show	73 (58.4)	52 (41.6)
Turning it Over	86 (68.8)	39 (31.2)
Accompanying Others	97 (77.6)	28 (22.4)
Coaching Compassion	73 (58.4)	52 (41.6)
Total strategies[a]	111 (88.8)	14 (11.2)

[a]Total strategies=used at least one STAC strategy.

XSL•FO
RenderX

The Relationship Between Acceptability, Relevance, and Use of STAC Strategies

We conducted a linear multiple regression analysis to examine the unique effect of acceptability and relevance on use of STAC strategies among students who reported witnessing bullying posttraining. Before conducting the regression analyses, we examined bivariate correlations for predictor and outcome variables (Table 3).

Table . Bivariate correlations for acceptability, relevance, and use of STAC strategies.

Measure	1	2	3
Acceptability	—[a]	—	—
Relevance	0.77[b]	—	—
Use of strategies	0.20[c]	0.29[b]	—

[a]Not applicable.
[b]P<.001.
[c]P=.031.

Although acceptability and relevance were highly correlated, the observed VIF was 2.4, with corresponding tolerance at 0.4, which is below the rule of thumb of VIF <10 [36], suggesting acceptable levels of multicollinearity among the predictor variables. The full regression equation was significant (R^2=0.09; $F_{112, 2}$=5.38; P=.006). Relevance was a significant predictor of use of STAC strategies posttraining (β=.34; P=.02; 95% CI 0.10-1 .06). In contrast, acceptability was not a significant predictor of posttraining STAC strategy use (β=−0.06; P=.66; 95% CI −0.35 to 0.23).

Discussion

Principal Findings

This is the first study to examine middle school students' perceptions of the acceptability and relevance of the STAC-T app and the relationship between program acceptability and relevance and posttraining strategy use. The results of this study indicate that the majority of students perceived the program to be both acceptable and relevant. Furthermore, the majority of students who witnessed bullying posttraining also reported use of at least one STAC strategy to intervene when witnessing bullying. Additionally, the results from the regression analysis demonstrated that among students who witnessed bullying posttraining, perceived relevance of the program was a significant predictor of posttraining STAC-T strategy use. In contrast, program acceptability was not significantly related to posttraining strategy use.

Acceptability and Relevance

Overall, the majority of students perceived the STAC-T program to be both acceptable (188/229, 82.1%, to 206/229, 90.0%) and relevant (180/229, 78.6%, to 190/229, 83%). The findings from this study are consistent with prior research examining the acceptability and relevance of the STAC in-person program for middle school students in low-income communities [17]. Specifically, acceptability ratings in prior research ranged from 83.9% to 88.9%, and relevance ratings ranged from 82.9% to 88.9%. Study results also extend findings from the initial STAC-T app usability testing, which included qualitative data related to acceptability and relevance for the program [26]. The current findings support the acceptability and relevance of the STAC-T app, with acceptability and relevance similar to the in-person STAC intervention, and add to the literature by examining acceptability and relevance among a larger, more diverse sample.

Use of STAC Strategies

For the use of STAC strategies, among students who reported witnessing bullying posttraining, the majority (111/125, 88.8%) reported using at least one STAC-T strategy. Although slightly lower, this rate is comparable with the in-person STAC intervention research among middle school students, with rates of posttraining STAC strategy use ranging from 91.7% [37] to 95.2% [16]. Furthermore, the results of this study indicated that students were most likely to use *Turning it Over* and *Accompanying Others*, relative to *Stealing the Show* and *Coaching Compassion*. These findings are also consistent with in-person STAC research, demonstrating similar patterns in rates of use by strategy [37]. One possible explanation for the differential strategy use rates is that when using *Stealing the Show* or *Coaching Compassion*, students engage with the student who is bullying either indirectly by attracting their attention or directly by speaking to them. Students may hesitate to engage either indirectly or directly with the student who is bullying, as students may perceive these behaviors to be associated with an increased risk of becoming the target of bullying themselves [38]. Findings suggest that overall, STAC-T posttraining strategy use is similar to rates found for the in-person STAC intervention.

The Relationship Between Acceptability, Relevance, and Use of STAC Strategies

Finally, the results of this study demonstrated that perceived program relevance, but not acceptability, is uniquely related to use of the STAC strategies. Although bivariate correlations revealed a positive relationship between both program acceptability and relevance, only relevance provided a significant, unique contribution to the variance in use of the STAC strategies. Research indicates that program acceptability is related to outcomes [27], whereas relevance is related to the use of intervention skills [28]. Thus, our finding that program relevance was uniquely related to the posttraining use of the STAC strategies is consistent with prior research. Although intervention success depends on intervention acceptability [27], with acceptability functioning as a gatekeeper for intervention

XSL•FO
RenderX

implementation [39], program relevance represents an additional level of social validity. Specifically, program relevance refers to the perception that the intervention is designed for similar users rather than a one-size approach, which can be enhanced by the ability to customize the program [40]. Thus, the ability to select avatars, view the avatar enacting a selected STAC strategy for a specific bullying situation, and receive feedback may enhance program relevance. These findings underscore the importance of program development and program features that promote the perception of program users that the program is relevant to their peer group and students at their school in order for users to apply the skills provided by the training.

Study Limitations

Although this study adds to the growing body of literature suggesting STAC-T is a promising bullying prevention program for middle school students, there are limitations. First, although participants were recruited from 6 middle schools in 3 geographic regions of the United States to increase generalizability, students from different geographic areas may have different perceptions of acceptability and relevance. Future research with larger, more diverse samples is needed to increase the generalizability of the findings. Additionally, our examination of bystander status and STAC strategy use was limited to self-report data. Retrospective self-report measures of bystander behavior can be biased or inaccurate [41]. Additionally, self-reports of bystander behavior include not just bystander behavior, but the interpretation of the event eliciting the bystander response [42]. Future research could include observational data or teacher-reported data to provide a more robust measure of strategy use. Furthermore, although program relevance was a significant predictor of use of STAC strategies, relevance accounted for only a small amount of variance in strategy use. Thus, unmeasured factors other than program relevance (eg, the actual program content) may account for the majority of variance in the frequency of post-training use of STAC strategies. Furthermore, consistent with the in-person STAC intervention, the findings from this study indicate students are less likely to use Stealing the Show and Coaching Compassion, relative to Turing it Over and Accompanying Others. It is possible that the training examples or avatar activities may need to be modified to be more relevant for middle school students. Qualitative data could provide valuable feedback related to how to encourage students to use Stealing the Show and Coaching Compassion.

Implications

The results of this study have important implications for the use of technology-based bullying prevention programs for middle school students. First, the findings from this study, coupled with prior research [25,26], indicate middle school students perceive the STAC-T app to be both acceptable and relevant. Furthermore, STAC-T acceptability and relevance rates are consistent with rates for the in-person STAC intervention, suggesting that students perceive the technology version of STAC to be as acceptable and relevant as the in-person intervention. The results of this study also provided support for the effectiveness of the STAC-T app in training students to intervene when they witness bullying using the 4 STAC strategies. Although slightly lower than STAC in-person posttraining skill use rates, STAC-T posttraining rates were comparable. Together, these findings are particularly important as STAC-T requires fewer resources for administration than the in-person STAC intervention and was designed to significantly reduce implementation barriers for schools, particularly those in low-income communities. Thus, this study supports the STAC-T app as a promising technology-based program for bullying prevention among middle school students.

Conclusions

Bullying is a significant public health issue for middle schools in the United States. Schools face many barriers to accessing and implementing bullying prevention programs. STAC-T is a technology-based bullying prevention app that has the potential to reduce these barriers. Findings from this study demonstrate the program acceptability, relevance, and effectiveness of STAC-T in training students to use skills to intervene when they witness bullying. This study provides support for the acceptability and relevance of the STAC-T app, as well as preliminary evidence for the effectiveness of the program.

Acknowledgments

The authors would like to thank the Klein Buendel Creative Team that designed, illustrated, and programmed the STAC-T app. No generative AI was used in any portion of the manuscript writing.

Funding

This research was supported by the National Institute on Minority Health and Health Disparities of the National Institutes of Health under award 2R42MD014943-02A1. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: DMD, AM, MKB
Data curation: DMD

XSL•FO
RenderX

Formal analysis: DMD
Funding acquisition: DMD, AM, MKB
Investigation: AM, RH, AW, TP, JS
Methodology: DMD, AM
Project administration: DMD, BH
Supervision: AM, DMD, MKB
Writing – original draft: DMD
Writing – review and editing: AM, RH, AW, MKB, TP, JS, BH

Conflicts of Interest

AM, DMD, and MKB are developers of the STAC-T app. All other authors declare no conflicts of interest.

References

1. Brank EM, Hoetger LA, Hazen KP. Bullying. Annu Rev Law Soc Sci 2012 Dec 1;8(1):213-230. [doi: 10.1146/annurev-lawsocsci-102811-173820]
2. Olweus D. Bullying at School: What We Know and What We Can Do: Oxford: Blackwell; 1993. [doi: 10.1002/pits.10114]
3. Menesini E, Salmivalli C. Bullying in schools: the state of knowledge and effective interventions. Psychol Health Med 2017 Mar;22(sup1):240-253. [doi: 10.1080/13548506.2017.1279740] [Medline: 28114811]
4. Student reports of bullying: results from the 2022 School Crime Supplement to the National Crime Victimization Survey. Institute of Education Sciences.: National Center for Education Statistics; 2024. URL: https://nces.ed.gov/pubs2024/2024109rev.pdf [accessed 2025-04-12]
5. Gaffney H, Ttofi MM, Farrington DP. Evaluating the effectiveness of school-bullying prevention programs: an updated meta-analytical review. Aggress Violent Behav 2019 Mar;45(Mar-Apr):111-133. [doi: 10.1016/j.avb.2018.07.001]
6. Reinke WM, Stormont M, Herman KC, Puri R, Goel N. Supporting children's mental health in schools: teacher perceptions of needs, roles, and barriers. Sch Psychol Q 2011;26(1):1-13. [doi: 10.1037/a0022714]
7. Salimi N, Karimi-Shahanjarin A, Rezapur-Shahkolai F, Hamzeh B, Roshanaei G, Babamiri M. Use of a mixed-methods approach to evaluate the implementation of violence and bullying prevention programs in schools. Educ Urban Soc 2021 Jul;53(6):607-628. [doi: 10.1177/0013124520972090]
8. Salmivalli C, Lagerspetz K, Björkqvist K, Österman K, Kaukiainen A. Bullying as a group process: participant roles and their relations to social status within the group. Aggr Behav 1996;22(1):1-15. [doi: 10.1002/(SICI)1098-2337(1996)22:1<1::AID-AB1>3.0.CO;2-T]
9. Wu WC, Luu S, Luh DL. Defending behaviors, bullying roles, and their associations with mental health in junior high school students: a population-based study. BMC Public Health 2016 Oct 10;16(1):1066. [doi: 10.1186/s12889-016-3721-6] [Medline: 27724881]
10. Waasdorp TE, Bradshaw CP, Dowdy E. Examining variation in adolescent bystanders' responses to bullying. School Psych Rev 2018 Mar 1;47(1):18-33. [doi: 10.17105/SPR-2017-0081.V47-1]
11. Bauman S, Yoon J, Iurino C, Hackett L. Experiences of adolescent witnesses to peer victimization: the bystander effect. J Sch Psychol 2020 Jun;80:1-14. [doi: 10.1016/j.jsp.2020.03.002] [Medline: 32540087]
12. Lambe LJ, Hudson CC, Craig WM, Pepler DJ. Does defending come with a cost? Examining the psychosocial correlates of defending behaviour among bystanders of bullying in a Canadian sample. Child Abuse Negl 2017 Mar;65:112-123. [doi: 10.1016/j.chiabu.2017.01.012] [Medline: 28131946]
13. Midgett A, Doumas D, Sears D, Lundquist A, Hausheer R. A bystander bullying psychoeducation program with middle school students: a preliminary report. TPC 2015;5(4):486-500. [doi: 10.15241/am.5.4.486]
14. Doumas DM, Midgett A, Hausheer R. A pilot study testing the efficacy of a brief, bystander bullying intervention: reducing bullying victimization among high school students. Professional School Counseling 2021 Jul 1;25(1). [doi: 10.1177/2156759X211018651]
15. Midgett A, Doumas DM, Johnston AD. Training students who occasionally bully as leaders in bullying curriculum delivery: evaluation of a brief, school-wide bystander intervention. Prof Sch Couns 2017;21(1):1-9. [doi: 10.1177/2156759X18778781]
16. Moran M, Midgett A, Doumas DM. Evaluation of a brief, bystander bullying intervention (STAC) for ethnically blended middle schools in low-income communities. Professional School Counseling 2019 Jan;23(1):2156759X20940641. [doi: 10.1177/2156759X20940641]
17. Midgett A, Doumas DM. Acceptability and short-term outcomes of a brief, bystander intervention program to decrease bullying in an ethnically blended school in low-income community. Contemp School Psychol 2020 Dec;24(4):508-517. [doi: 10.1007/s40688-020-00321-w]
18. Midgett A, Doumas DM, Trull R, Johnson J. Training students who occasionally bully to be peer advocates: is a bystander intervention effective in reducing bullying behavior? J Child Adolesc Couns 2017 Jan 2;3(1):1-13. [doi: 10.1080/23727810.2016.1277116]
19. Doumas DM, Midgett A, Watts AD. The impact of a brief, bullying bystander intervention on internalizing symptoms: is gender a moderator of intervention effects? Sch Psychol Int 2019 Jun;40(3):275-293. [doi: 10.1177/0143034319830149]

XSL•FO
RenderX

20. Midgett A, Doumas DM. The impact of a brief bullying bystander intervention on depressive symptoms. Jour of Counseling & Develop 2019 Jul;97(3):270-280. [doi: 10.1002/jcad.12267]
21. Midgett A, Doumas DM, Trull R. Evaluation of a bystander bullying intervention program for elementary school students. Prof Sch Couns 2018;20(1):172-183. [doi: 10.5330/1096-2409-20.1.172]
22. Watts AD, Doumas DM, Midgett A. Efficacy of a brief, school‐based bystander bullying intervention on high school students' alcohol use. J Addict Offender Couns 2019 Oct;40(2):66-83. [doi: 10.1002/jaoc.12066]
23. Midgett A, Doumas DM, Peralta C, Bond L, Flay B. Impact of a brief, bystander bullying prevention program on depressive symptoms and passive suicidal ideation: a program evaluation model for school personnel. J Prev Health Promot 2020 Jul;1(1):80-103. [doi: 10.1177/2632077020942959] [Medline: 33738442]
24. Midgett A, Doumas DM, Myers VH, Moody S, Doud A. Technology-based bullying intervention for rural schools: perspectives on needs, challenges, and design. Rural Ment Health 2021 Jan;45(1):14-30. [doi: 10.1037/rmh0000151] [Medline: 33777286]
25. Doumas DM, Midgett A, Myers V, Buller MK. Usability of a technology-based bystander bullying intervention for middle school students in rural, low-income communities: mixed methods study. JMIR Form Res 2021 Oct 26;5(10):e32382. [doi: 10.2196/32382] [Medline: 34582355]
26. Midgett A, Doumas DM, Peralta C, Peck M, Reilly B, Buller MK. Usability testing of a bystander bullying intervention for rural middle schools: mixed methods study. JMIR Hum Factors 2025 Feb 21;12:e67962. [doi: 10.2196/67962] [Medline: 39984169]
27. Sekhon M, Cartwright M, Francis JJ. Acceptability of healthcare interventions: an overview of reviews and development of a theoretical framework. BMC Health Serv Res 2017 Jan 26;17(1):88. [doi: 10.1186/s12913-017-2031-8] [Medline: 28126032]
28. Farrell AD, Mehari KR, Kramer-Kuhn AM, Mays SA, Sullivan TN. A qualitative analysis of factors influencing middle school students' use of skills taught by a violence prevention curriculum. J Sch Psychol 2015 Jun;53(3):179-194. [doi: 10.1016/j.jsp.2015.03.001] [Medline: 26054813]
29. Online bullying bystander intervention for middle schools phase II (STACII). Clinical Trials.gov. URL: https://clinicaltrials.gov/study/NCT05572398 [accessed 2025-11-14]
30. Pocock SJ. Clinical Trials: A Practical Approach: Wiley; 1983. URL: https://www.jameslindlibrary.org/wp-data/uploads/2019/04/69-Stuart-J-Pocock-1983.pdf [accessed 2025-11-12]
31. Doumas DM, Midgett A, Watts AD. A pilot evaluation of the social validity of a bullying bystander program adapted for high school students. Psychol Sch 2019 Jul;56(7):1101-1116. [doi: 10.1002/pits.22249]
32. Midgett A, Doumas DM, Trull R, Johnston A. A randomized controlled study evaluating a brief, bystander bullying intervention with junior high school students. J Sch Couns 2017;15(9) [FREE Full text]
33. Moran M, Midgett A, Doumas DM, Moody S, Porchia S. A mixed method evaluation of a culturally adapted, brief, bullying bystander intervention for middle school students. J Child Adolesc Couns 2019;5(3):221-238. [doi: 10.1080/23727810.2019.1669372] [Medline: 33015352]
34. Peck M, Doumas DM, Midgett A. Examination of the bystander intervention model among middle school students: a preliminary study. TPC 2024;14(2):119-134. [doi: 10.15241/mp.14.2.119]
35. Cohen J. Statistical Power Analysis for the Behavioral Sciences: Routledge; 1969. URL: https://utstat.utoronto.ca/~brunner/oldclass/378f16/readings/CohenPower.pdf [accessed 2025-11-12]
36. Erford BT. Research and Evaluation in Counseling: Cengage Learning; 2015. URL: https://www.abebooks.com/9781285454894/Research-Evaluation-Counseling-Erford-Bradley-1285454898/plp [accessed 2025-11-12]
37. Midgett A, Peck M, Doumas DM. Bullying in private schools: can training middle school students to intervene reduce bullying perpetration and victimization? PSC 2025 Jan;29(1). [doi: 10.1177/2156759X251316870]
38. Midgett A, Moody S, Reilly B, Lyter S. The phenomenological lived experience of student-advocates acting as "defenders" to stop bullying. J of Hum Couns 2017:53-71. [doi: 10.1002/johc.12044]
39. Gadke DL, Kratochwill TR, Gettinger M. Incorporating feasibility protocols in intervention research. J Sch Psychol 2021 Feb;84:1-18. [doi: 10.1016/j.jsp.2020.11.004] [Medline: 33581765]
40. Arefadib N, Shafiei T, Cooklin A. Barriers and facilitators to supporting women with postnatal depression and anxiety: A qualitative study of maternal and child health nurses' experiences. J Clin Nurs 2023 Feb;32(3-4):397-408. [doi: 10.1111/jocn.16252] [Medline: 35156748]
41. Baumeister RF, Vohs KD, Funder DC. Psychology as the science of self-reports and finger movements: whatever happened to actual behavior? Perspect Psychol Sci 2007 Dec;2(4):396-403. [doi: 10.1111/j.1745-6916.2007.00051.x] [Medline: 26151975]
42. Jouriles EN, Kleinsasser A, Rosenfield D, McDonald R. Measuring bystander behavior to prevent sexual violence: moving beyond self reports. Psychol Violence 2016;6(1):73-81. [doi: 10.1037/a0038230]

Abbreviations

PIN: personal identification number

XSL•FO
RenderX

VIF: variance inflation factor

Edited by A Stone, A Mavragani; submitted 02.05.25; peer-reviewed by ME Heidari; revised version received 02.10.25; accepted 03.10.25; published 18.11.25.

Please cite as:
Doumas DM, Midgett A, Hausheer R, Winburn A, Buller MK, Perron T, Shelton J, Herbeck B
Acceptability, Relevance, and Short-Term Outcomes of the STAC-T Bullying Bystander App: Feasibility Quantitative Study
JMIR Form Res 2025;9:e76830
URL: https://formative.jmir.org/2025/1/e76830
doi:10.2196/76830

© Diana M Doumas, Aida Midgett, Robin Hausheer, Amanda Winburn, Mary K Buller, Taylor Perron, Jennalyn Shelton, Brandon Herbeck. Originally published in JMIR Formative Research (https://formative.jmir.org), 18.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Usability Evaluation of a Mixed Reality Platform in Pediatric Interventional Cardiology by Specialist Physicians: Mixed Methods Study

Catherine Chilute Chilanga[1,2*], PhD; Andreas Jahnen[3*], MSc; Linda Hafskjold[1,2*], PhD

[1]Center for Health and Technology, University of South-Eastern Norway, Grønland 58, Drammen, Norway

[2]Department of Optometry, Radiography and Lighting Design, Faculty of Health and Social Sciences, University of South-Eastern Norway, Kongsberg, Norway

[3]Luxembourg Institute of Science and Technology, Esch Sur Alzette, Luxembourg

[*]all authors contributed equally

Corresponding Author:
Catherine Chilute Chilanga, PhD
Center for Health and Technology, University of South-Eastern Norway, Grønland 58, Drammen, Norway

Abstract

Background: Immersive technologies such as virtual reality, augmented reality, and mixed reality are used in pediatric interventional cardiology (IC) to improve anatomical visualization and spatial understanding and support procedural precision. These technologies offer 3D representations of complex cardiac structures, which can aid in preprocedural planning, intraoperative navigation, and team communication. As these technologies gain traction in IC, understanding how medical specialists perceive their value is vital.

Objective: This study assesses what IC specialist physicians consider valuable in immersive technologies by evaluating their experiences with *CardioVision*, a software designed for pediatric IC procedures.

Methods: A mixed methods design was used, combining questionnaires and semi-structured interviews. Purposeful and snowball sampling was used to recruit specialist physicians with experience in IC. The recruited specialist physicians were asked to submit IC cases for software testing. These were uploaded to *CardioVision* by the developer, who also conducted the testing in person with the physicians at their designated clinical stations. A total of eight specialist physicians completed a pretest questionnaire, used the software, and thereafter completed a posttest questionnaire, which included System Usability Scale and Likert-scale questions on experience with the software. The interviews explored the specialist physicians' experience with the software's performance. Of the 8 participants, 6 further participated in the interviews. The interviews were conducted by two of the authors. The qualitative data were analyzed using inductive content analysis, involving subjective interpretation of textual material from the interviews. Following the approach outlined by Elo and Kyngäs, the process comprised three phases: preparation, organization, and reporting. In the preparation and organization phases, one author organized and categorized the interview data in an Excel sheet. The two authors then reviewed the material, compared emerging codes and themes, and reached consensus on the final main theme and subthemes for reporting.

Results: In total, 8 questionnaires and 6 interviews were obtained from specialist physicians from Belgium, France, Germany, and Italy. Quantitative data from mean System Usability Scale scores showed that users felt confident using the system (mean 4.25). When using the software, the highest positive feedback was for visualization of structures, where all users (8/8, 100%) agreed. Overall, most participants (6/8, 75%) rated the system positively for decision-making support and 5/8 (63%) for complex case preparation. One participant (1/8, 12%) disagreed with its effectiveness in preparing complex cases, indicating mixed opinions in this area. In the interviews, emerging themes aligned with quantitative data to show that participants valued the software's ease of use, high-quality visualization, and decision-making support. Additional benefits included use in medical education, interprofessional collaboration, and patient communication. Developer–clinician collaboration was emphasized as vital for effective integration into routine practice.

Conclusions: Immersive technologies offer clinical value when aligned with user needs and integrated through collaborative development.

(JMIR Form Res 2025;9:e79278) doi:10.2196/79278

KEYWORDS

cardiology; congenital heart disease; mixed reality; pediatrics; interventional radiology

XSL•FO
RenderX

Introduction

Interventional cardiology (IC) uses image-guided techniques to perform minimally invasive procedures for diagnosing and treating various cardiac diseases [1]. In pediatric IC, less invasive solutions for congenital heart disease (CHD) are reducing the need for open-heart surgery, enhancing patient care [2], and transforming the clinical practice of interventional cardiologists [3]. IC has progressed over the years, integrating advanced technologies to enhance visualization of anatomical structures in CHD [4]. Technologies such as virtual reality (VR), augmented reality (AR), and mixed reality (MR) are now being integrated to enhance visualization and procedural precision in cardiology [5,6]. VR creates a simulated 3D environment mimicking real-world scenarios, AR overlays medical imaging onto the physical world for real-time interaction [4], and MR combines digital elements with reality for dynamic, interactive engagement [7]. Integrating these immersive technologies with artificial intelligence (AI) and wearable headsets capable of displaying holographic 3D images enhances user interaction with digital content in real-world environments, further expanding their utility in medical imaging [8].

Immersive technologies have made it possible to treat a wider range of CHDs than was previously achievable [3]. Ionizing radiation is used during IC procedures, resulting in substantial patient radiation exposure, and when performed on pediatric or adolescent patients, minimizing radiation risks is crucial [9]. In the HARMONIC (health effects of cardiac fluoroscopy and modern radiotherapy in pediatrics) project [10], efforts have focused on optimizing interventional procedures in pediatric cardiology patients to enhance safety and minimize radiation exposure. In line with the goals of the HARMONIC project, an innovative immersive technology, *CardioVision*, to support IC for pediatric patients, was developed. *CardioVision* is a VR and MR software integrated with AI technology. It provides a 3D reconstructed heart model, giving cardiologists unprecedented visual insights into cardiac anatomy and function [11]. *CardioVision* was developed by a multidisciplinary team of computer scientists, medical physicists, and pediatric cardiologists, focusing on the preparation of complex, high-risk cases. Using web technologies, the platform can be accessed from standard computers to MR devices without installing dedicated software. The full components of the software's system are outlined in the methods.

Immersive technologies can be classified into two types: those used by patients and those used by health care providers [12], with *CardioVision* designed for the latter. Several applications and benefits of VR and AR technologies in medical imaging have been reported, including enhanced image quality, assistance with procedure preplanning, and support for other uses, such as multidisciplinary discussions and patient education [4]. Iqbal et al [13], however, point out that a crucial barrier to implementing most immersive technologies in health care is the lack of seamless integration with existing systems, compounded by technical challenges and limited clinical evidence for its use in critical procedures. Lara Hernandez et al [14] in a systematic review study further state that digital health technologies in cardiology remain at early stages, classified as either *proof of concept* or *robust candidates* that are yet to be integrated into routine clinical workflow. This implies that clinicians currently cannot fully rely on these tools for routine decision-making or patient care until they are thoroughly tested and proven effective in real-world clinical settings. This underscores a significant research gap: the need to identify and validate the most effective clinical applications of immersive technologies to support their practical, trustworthy use in cardiology. Kouijzer et al [15] suggest that for the implementation of immersive technology in health care, it is essential to address the entire process, identifying barriers and applying coordinated, multilevel strategies to ensure effective implementation. In this regard, understanding the best clinical applications for these immersive technologies is vital.

In this study, we evaluate *CardioVision*, a software designed for busy clinical settings. *CardioVision* is designed with a flexible workflow in mind. It uses standard internet technologies available on most computers, making case creation and management straightforward. Web-based extended reality enables AI-powered visualizations to be displayed across a wide range of hardware, ensuring clinicians have easy access and optimal viewing quality.

Our current aim is to assess how pediatric interventional cardiologists and other specialist physicians with experience in IC procedures perceive immersive technologies as valuable by evaluating their use of *CardioVision*.

Methods

Study Design

The study used a mixed methods approach, combining data collection via questionnaires and semi-structured interviews. Specialist physicians with experience in IC were invited to test the *CardioVision*. The software system architecture is outlined in Figure 1. The system architecture is designed to enable seamless management and processing of medical imaging data using modern, responsive web technologies. A key advantage of this approach is that no dedicated client software needs to be installed on hospital workstations or 3D headsets. Users can upload anonymized patient data in Digital Imaging and Communications in Medicine format. While the platform primarily supports computed tomography (CT) scans, other modalities such as 3D ultrasound are also compatible. Each upload creates a new case that is securely stored on the *CardioVision* server.

The uploaded data is processed through a set of external tools for visualization and analysis. Clinicians can interact with the data through intuitive visualization tools, such as volume rendering, segmentation viewers, and a Kitware GLANCE-based viewer [16]. A demo account is available for testing on request to the developer (second author).

XSL•FO
RenderX

Figure 1. The *CardioVision* infrastructure provides data management (CardioVision UI and CardioVision Server), AI-enabled image processing (Slider 3D [17] and MONAI Label [18]), and viewing (volume, mixed reality, and viewers) components. All components are web-enabled to allow for central management, versatile deployment, and scalability.

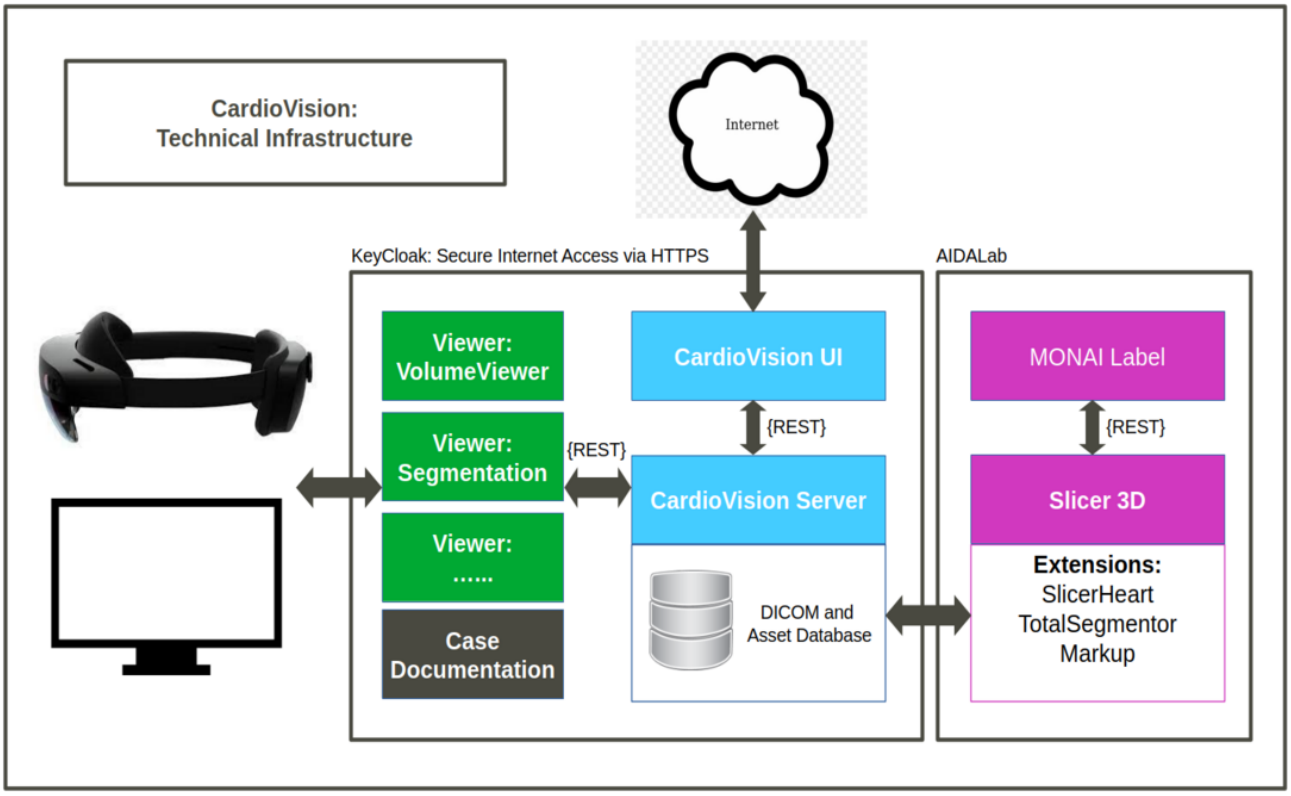

Participants Recruitment

The sampling strategy aimed to recruit physicians with expertise in the field of IC and a strong interest in the adoption of XR technologies. Recruitment combined purposive sampling, which selects participants based on expertise and relevance [19], with snowball sampling, in which participants recommend others meeting the same criteria [20]. Initially, the purposive recruitment targeted physicians with experience in IC and who had an interest in evaluating the software. This step was conducted by the developer and author (AJ) using established professional networks to identify potential participants. The physicians were selected for their relevant expertise and motivation to participate and were invited to recommend colleagues with similar qualifications, completing recruitment through snowball sampling.

Eleven specialist physicians were invited to participate, of whom eight confirmed. Before the day of testing the software, each specialist physician selected and sent in a personal case that fell into the category of preparation for a difficult case. Table 1 indicates the type of cases submitted. The majority of cases were related to CHD, with the exception of a complex noncardiac aneurysm. No patient personal information was transmitted; only data sufficient to simulate the case for testing purposes were used to assess the software's ease of use, the visualization of pathology, and the potential for radiation dose optimization.

XSL•FO
RenderX

Table . Cases assessed by the specialist physicians.

Case diagnosis	Description
Difficult Patent Ductus Arteriosus (PDA)	Large PDA in an infant unresponsive to medical therapy; required device-based closure.
Discontinuous pulmonary arteries	Rare congenital defect with separate origins of branch pulmonary arteries; required staged surgical reconstruction for unifocalization.
Complex malformation	The patient presented with a complex malformation of the heart, including several previous interventions.
Univentricular in a patient with double outlet	Double outlet right ventricle with univentricular physiology.
Aortic coarctation reduction	An aortic reduction occurs close to the right ventricle in an S shape.
Pulmonary artery stenosis	Branch pulmonary artery narrowing causing right ventricular pressure overload; treated with balloon angioplasty and stenting.
Angioplasty	Catheter-based dilation of stenotic vessel to restore blood flow.
Complex aneurysm	Noncardiac case showed a difficult-to-treat aneurysm. Visualizations include only the blood vessels.

The corresponding data to the cases were then processed and made available on the *CardioVision* software platform. Figures 2 and 3 show examples of visualizations that help the cardiologist to decide on the best treatment.

Figure 2. A sample visualization. The heart is segmented into different parts and colored for better differentiation.

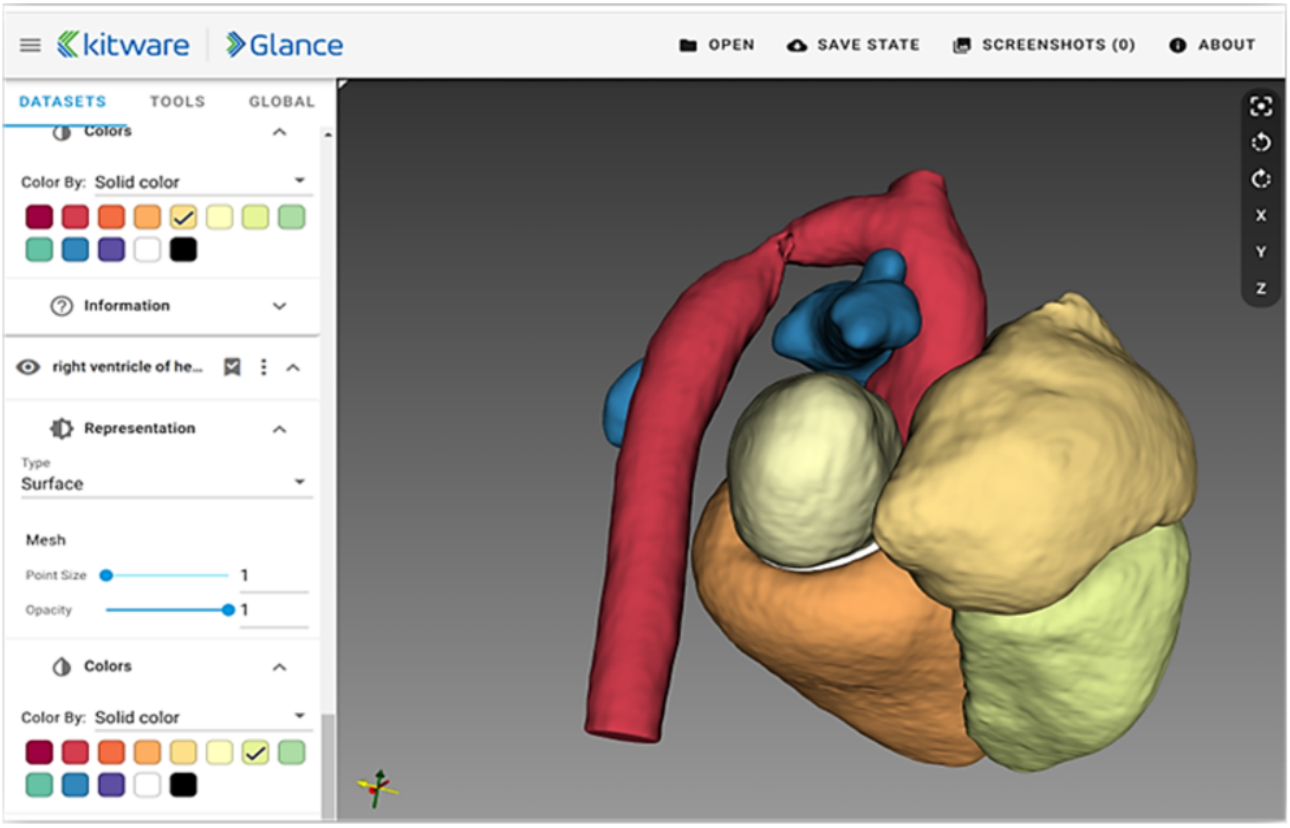

XSL•FO
RenderX

Figure 3. A segment of an aorta is shown in mixed reality. Due to the AI-based segmentation, the aorta is exempted. This allows us to see the reduction of the aorta and the S shape, which makes the treatment difficult.

Prior to using the *CardioVision* software, the specialist physicians completed a pre-questionnaire aimed at evaluating their expectations about its performance. After completion of the questionnaire on expectations, the specialist physicians were invited to use the software on the IC procedures. The software testing was performed with the developer physically present at the specialists' clinical workstation or hospital. At the beginning of the testing, each specialist physician received a brief introduction to the *CardioVision* software platform and guidance on using the MR device (Microsoft HoloLens 2, Microsoft). After testing the software, the physicians answered a post-questionnaire, which included a standard 10-question System Usability Scale (SUS). The SUS is widely used as a standardized tool for evaluating perceived usability of systems and technologies [21]. The 10 questions (Qs) of the SUS are designed to assess various aspects of system usability. The items alternate between positive and negative statements to minimize response bias [21]. Positive statements (Q1, Q3, Q5, Q7, and Q9) focus on overall usability perception, system consistency, confidence in use, learnability, and integration into workflow. In contrast, negative statements (Q2, Q4, Q6, Q8, and Q10) address concerns related to complexity, need for technical support, inconsistency, cognitive load, and learnability challenges. Together, these questions provide a balanced evaluation of a system's ease of use, user confidence, and potential barriers to adoption [21]. The 10 questions are outlined in the usability questionnaire (see attached questionnaire in Multimedia Appendix 1)

Additional questions about participants' experiences were included following the SUS items to complement the usability assessment. Both the expectation and usability questionnaire had a 5-point Likert scale designed to assess the participants' agreements [22]. At the end of the questionnaire, the specialist physicians were asked to indicate if they agreed to further participate in an interview to describe their experience with the use of the software. Six of the 8 specialist physicians agreed to participate in the interviews.

Individual interviews were thereafter arranged for a later date at the participants' convenience and conducted using an interview guide (as shown in Multimedia Appendix 2). The first and last authors (CCC and LH) conducted the interviewers. The interview addresses five parts: initial impression of the software, experience of using it, the performance of the software, possible feature enhancements, and overall satisfaction of use. The interviews were recorded using the Nettskjema Diktafon app (University of Oslo, 2024), a secure tool for collecting audio data [23]. The interviews were conducted using a semi-structured guide and lasted approximately 25 to 40 minutes. The interview guide helped ensure that key topics related to the study's aim were explored. Probing questions were used to explore the specialist physicians' clinical examples and experiences, ensuring richness and depth of the data. Similar responses were observed after the first 3 interviews, and by the sixth participant, no new information had emerged. Saturation, defined as the point at which no new issues or insights emerge from subsequent interviews [24], is therefore assumed to have been reached. Achieving saturation is further supported by the

XSL•FO
RenderX

sampling strategy, given that the selected participants possessed in-depth and broad insights into the key topics relevant to the study's aim.

Data Analysis

The quantitative data were analyzed using IBM SPSS Statistics version 29 (IBM Corp, 2022). The data were sorted in a Microsoft Excel file, coded, and exported to SPSS. Mean values were calculated to determine variation usability among the specialist physicians. In the qualitative data, inductive content analysis was used, which included preparing the data and organizing code to create themes [25]. This process comprised subjective interpretation of textual data and the classification of emerging codes to identify subthemes and main themes. The analysis followed the three phases of inductive content analysis as outlined by Elo and Kyngäs [25] of preparation, organization, and reporting. In the preparation phase, one of the researchers familiarized themselves with the data by reading through the interview transcripts and identifying relevant sections, which ranged from whole responses to selected words or phrases that captured key concepts. In the organization phase, individual codes were developed from the extracted text and subsequently grouped into subthemes and overarching main themes. The preparation and organizing of the qualitative data were sorted manually in a Microsoft Excel file. Two authors then jointly reviewed, discussed, and agreed upon the meanings and structure of the subthemes and main themes to ensure consistency and credibility. Finally, in the reporting phase, the results of the analysis were presented or reported. The quantitative and qualitative data were then triangulated to compare the different data sources. Explanatory sequential mixed methods were used where the quantitative data collection and analysis was followed by the collection of qualitative data, which are used to further explain the initial quantitative results [26].

Ethical Considerations

Ethical approval for this study was obtained from the Norwegian Agency for Shared Services in Education and Research (reference number 441934). Written information about the study's purpose, procedures, and confidentiality measures was provided to all specialist physicians prior to participation, and written informed consent was obtained electronically. As the evaluation of CardioVision involved the use of fully anonymized patient data for simulation and testing purposes only, the requirement for additional approval to use patient data was waived in accordance with the guidance of the Norwegian Agency for Shared Services in Education and Research. No identifiable patient information was collected, and the anonymized case data were not included in the study's data analysis. Data collected from the participating specialist physicians were treated as strictly confidential and stored securely in compliance with applicable data protection regulations.

Results

Demographic Data of Participants

Eight specialist physicians from Belgium, France, Germany, and Italy, with experience in pediatric IC and interventional radiology, completed questionnaires, and 6 participated in individual interviews, resulting in 14 data sets. Table 2 shows an outline of the sample characteristics. A total of 5/8 (63%) specialist physicians had above 10 years' experience in clinical practice. Half (4/8, 50%) of all the specialist physicians had experience with cardiac procedures as well as MR/VR use in clinical practice. Only 2 (2/8, 25%) specialist physicians, however, reported using MR/VR in cardiology. All the participants stated that they work in public practice. The 6 interviewed specialist physicians were all male: 4 working as pediatric interventional cardiologists and 2 as neurointervention surgeons. The participants interviewed were mostly those with over 10 years of professional experience in their respective fields, while only 1 had less than or equal to 5 years.

XSL•FO
RenderX

Table . Demographics and sample characteristics.

Participant (PT)	Interview	Specialty	Years of experience (y)	Pediatric cardiology experience	Familiar with MR[a]/VR[b]	Uses MR/VR in practice
PT1	Yes	Pediatric interventional cardiologist	>10	Yes	No	No
PT2	Yes	Pediatric interventional cardiologist	≤5	Yes	Yes	Yes
PT3	Yes	Pediatric interventional cardiologist	>10	Yes	No	No
PT4	Yes	Pediatric interventional cardiologist	>10	Yes	Yes	Yes
PT5	Yes	Neurointervention surgeon	>10	No	No	No
PT6	Yes	Neurointervention surgeon	>10	No	Yes	No
PT7	No	Neurointervention surgeon	6 - 10	No	Yes	No
PT8	No	Neurointervention surgeon	1 - 5	No	No	No

[a]MR, mixed reality.

[b]VR, virtual reality.

System Usability

The SUS scores (Table 3) show that users feel confident using the system (Q5: 4.25 [range 3.00 - 5.00]) and do not consider it overly complex (Q2: 1.88 [range 1.00 - 3.00]). However, some users reported needing technical support (Q4: 3.00 [range 2.00 - 5.00]), experiencing minor difficulties when learning to use the system (Q10: 2.37 [range 1.00 - 4.00]), and finding it somewhat effortful to use (Q8:3.00 [range 1.00 - 5.00]). The data from the SUS scores aligned with the interview data from the theme user experience, which is explained further below with the other emerging themes.

Table . SUS[a] mean scores and statement interpretation of results.

SUS question	Mean (range[b])	Interpretation
Q1 (positive)	3.75 (2.00 - 5.00)	Users generally find the system usable, but there is room for improvement
Q2 (negative)	1.88 (1.00 - 3.00)	The system is not perceived as overly complex
Q3 (positive)	3.75 (2.00 - 5.00)	Users feel the system is fairly consistent in its design and behavior
Q4 (negative)	3.00 (2.00 - 5.00)	Some users feel they might need technical support to use the system effectively
Q5 (positive)	4.25 (3.00 - 5.00)	Users feel confident when using the system
Q6 (negative)	2.00 (1.00 - 3.00)	The system is not perceived as highly inconsistent, but there might be some inconsistencies.
Q7 (positive)	3.75 (2.00 - 5.00	Users generally find the system easy to learn
Q8 (negative)	3.00 (1.00 - 5.00)	Some users find the system somewhat effortful to use, indicating potential usability barriers
Q9 (positive)	3.38 (2.00 - 5.00)	The system fits reasonably well into users' workflows but could be improved
Q10 (negative)	2.37 (1.00 - 4.00)	Some users experience minor difficulties in learning how to use the system

[a]SUS, System Usability Scale.

[b] Minimum-maximum values.

XSL•FO
RenderX

Themes Emerging From Interview Data

In the data from the interviews, three main themes, user experience, applications, clinical interventions, and nonclinical interventions, emerged, with subthemes and related categories (Figure 4).

Figure 4. Emerging themes from interview data.

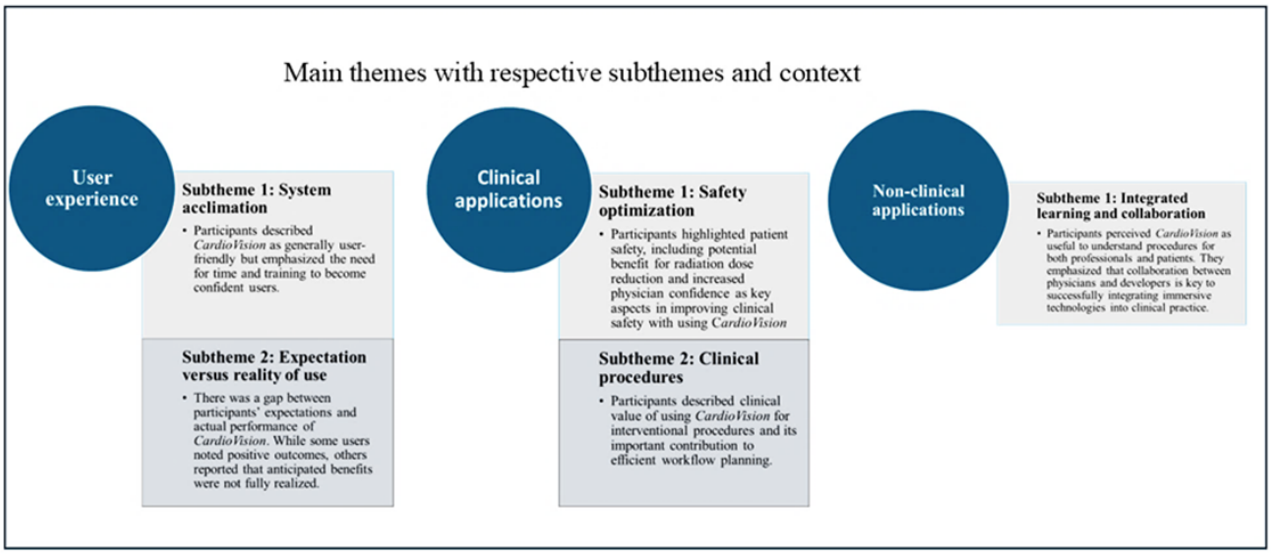

User Experience

The theme user experience comprises two subthemes: the ease with which participants became familiar with the software (system acclimation) and the alignment between their expectations and their actual experience. In general, most of the specialist physicians found the software easy to use. The specialist physicians expressed that the software was intuitive and suitable for clinical use. Most participants during the interviews expressed the value of the *CardioVision* software's use in clinical practice as from the statements below:

> *No, it was, as I said, a very nice experience...the benefits of it and what I saw, I mean, you don't need a lot of cases to be able to appreciate, maybe not judge, but to appreciate what such a tool can bring you in everyday practice.* [PT 6]

> *I mean, the whole experience, especially the way you change your way of thinking, you know, as you put it inside the procedure you are going to do. That's something I really liked.* [PT 5]

Although overall usability was considered acceptable, the results point to areas for improvement and emphasize the importance of allowing users time to adapt to the software in clinical practice, as reflected in the participants' comments below:

> *...I mean, if you're talking about things to continue working on it, so then the cross-sections, and I guess the other thing to try to find the way to incorporate very procedural imaging, like real-time in the software, those are, I think, the two...things to work on i guess in the future* [PT 2]

> *Most of these systems, and that's no matter which software is behind it, (..) take some time and that will be the most issue I think for clinicians in using it on daily practice* [PT1]

The specialist physicians express both unfulfilled and positive outcomes. Some specialist physicians noted limitations, such as difficulty recognizing nearby body implant devices within the area of pathology, segmentation issues, and the need for higher resolution. They also highlighted the importance of precise tissue differentiation. On the positive side, the system successfully simulated procedures like stent implantation and provided excellent visualization for large vascular structures. Some specialist physicians found the *CardioVision* software to be more immersive, providing more details, and was perceived as beneficial when compared to other software they currently use in clinical practice.

Applications: Clinical Interventions

Applications in clinical interventions included two subthemes: safety optimization and value for clinical procedures. The specialist physicians suggested that the software could be useful to optimize general patient safety, optimize radiation dose, and reinforce physician confidence. Enhanced anatomical visualization through *CardioVision* software was suggested to improve safety optimization by enabling physicians to perform IC procedures with greater precision and efficiency. In value for radiation dose optimization, most of the specialist physicians highlighted the reduction in time the software could potentially provide when performing procedures. Some, however, highlighted a trade-off between using CT scans to acquire imaging data for use in the *CardioVision* software and the low radiation dose currently used in procedures of pediatrics, noting that while CT provides quality imaging, it requires radiation, which may offset its benefits, as stated below:

> *...with standard catheterization labs nowadays. they are really very low dose already. If we look at the doses that we use in small kids, the doses have been reduced already quite a lot. So I think if you probably add a CT scan, that's probably much higher than what we use in a standard procedure. Of course,*

XSL•FO
RenderX

> *sometimes there are very long procedures where we are balancing the right position. And maybe in some cases, you might be able to reduce radiation....But I'm not...completely convinced that in general you will be able to achieve a dose reduction in all the cases that you want to use* [PT2]

Additionally, better imaging and anatomical visualization were suggested to boost clinicians' confidence, contributing to better planning and smoother IC procedures as expressed in a statement of one participant.

> *I think CardioVision added something because...you can see better all the images, so I feel more comfortable of the procedure, and I think in this particular case I planned better my procedure. More confidence and also maybe a little...more faster...during the procedure...i didn't measure but It was my feeling* [PT4.]

In the subcategory of clinical procedures, the participants pointed out specific areas where they perceived the software to be of value in clinical practice. This included value across all general types of IC procedures and for workflow planning. Workflow planning included facilitating the selection of appropriate medical devices before scheduling the patient for the procedure.

> *The time is not only the time you are in the cath lab, but also the time you prepare the case. But it's maybe bigger than the time in the cath lab. And also, if you prepare well, you will have a shorter time in cath lab. So, I think this is the general practice of every interventional cardiologist. So you go in the cath lab, you don't want to have a surprise. We don't like surprise in general. So we want to have all the material ready, for example, because not all the calf lab has every kind of stent, balloon, valve, etc.* [PT4]

The participants further specifically highlighted the software's potential in preplanning and performing of surgical procedures as highlighted by a pediatric cardiologist below.

> *...in most cases where I use 3D segmented images is actually planning of complex surgical repairs...or complex congenital heart disease where the surgeon needs to... [suture] in a certain patch. And so the experience that I have is always focused on that. I think the platform could also be used for that...I mean, if you have a proper...VR platform where you can cut through and turn around images you could also use it for surgical planning of complex heart disease* [PT2]

Applications: Nonclinical Interventions

The theme applications for nonclinical interventions focused on one subtheme: integrated learning and collaboration. This included collaboration among health care professionals, patients, and software developers. All participants agreed that the software could be a valuable tool for educating medical students. Some participants also highlighted its usefulness in collaborative discussions with both patients and other healthcare professionals. Others emphasized its potential for discussing treatment plans and diagnostic outcomes with patients. Additionally, some participants stressed the importance of engagement between specialist physicians and software developers, noting that this collaboration could help bridge the gap between clinical needs and available technology. Testing the software's performance together with the developer for easier integration into clinical workflow was also highlighted as vital when considering adaptation of such immersive technologies in IC practices.

> *…, I think it was something very interesting from my experience. And I think also for [developer], because...what we think is important, sometimes an engineer or a person that do another job don't know that this can be important in operating theater. So, I think the cooperation was very interesting from both side* [PT3]

Specialist Physicians' Overall Satisfaction With Use of Software

Table 4 below shows the reported areas of satisfaction by the specialist physicians. The highest positive feedback was in the visualization of structures, where all users agreed or strongly agreed (100%). Decision-making and preparation of complex cases also received positive feedback, with ratings of agree/strongly agree at 75% and 63%, respectively. A slight variation was, however, noted in use for preparing complex cases, as others were neutral (25%) or disagreed with this (13%). Half of the participants (50%) reported being satisfied with the user functions, while the other half (50%) gave neutral responses, indicating potential areas for improvement.

Table . Specialist physicians (n=8) rated areas of satisfaction with *CardioVision.*

Rated areas of satisfaction	Strongly disagree, n (%)	Disagree, n (%)	Neutral, n (%)	Agree, n (%)	Strongly agree, n (%)
User functions	0 (0)	0 (0)	4 (50)	2 (25)	2 (25)
Prepare complex cases	0 (0)	1 (12.5)	2 (25)	4 (50)	1 (12.5)
Visualization of structures	0 (0)	0 (0)	0 (0)	7 (87.5)	1 (12.5)
Decision making	0 (0)	0 (0)	2 (25)	4 (50)	2 (25)

XSL•FO
RenderX

Discussion

Principal Findings

This study aimed to assess the experiences of specialist physicians in IC in using a pediatric IC immersive technology, *CardioVision*. The focus was on how the specialists perceived the benefits of the technology in relation to IC procedures and other relevant applications. The discussion adopts a clinical perspective, evaluating the balance between expectations and actual experiences with immersive technology such as *CardioVision*. It also identifies areas for improvement, particularly those that address the unique clinical needs of IC specialist physicians.

A key motivation for most specialist physicians in using *CardioVision* in our study was the need to enhance clinical value during IC procedures. Cardiac anatomy is complex, and visualizing structural deformities presents a significant challenge [3,27]. Immersive technologies have shown great promise in improving procedural accuracy, efficiency, and clinical decision-making for cardiac pathologies in both adult and pediatric populations [5,6,28]. A central finding from our study is that IC specialist physicians highly valued the ability of immersive technology to deliver precise and accurate visualization of cardiac anatomy. In a study to assess assistance with diagnosis, Lee et al [29] demonstrated that immersive technology was the most effective tool for visualizing CHD and anatomical comprehension. Studies, however, report that visual quality is one of the most demanding aspects of immersive technology with image display, often representing the largest design and cost constraint [5,30]. In IC, the precision of visual information is paramount [7]. High-resolution displays are crucial for accurately visualizing complex cardiac structures, aiding in diagnosis and procedural planning [31,32]. Our study indicates that immersive technologies that add clinical value, support procedural precision, and improve the quality of pathological visualization are highly valued by IC specialist physicians. This perceived usefulness is a critical factor in their willingness to adopt such technologies in clinical practice [29], underscoring its role in facilitating successful implementation and guiding future development in IC.

The technology's time efficiency also emerged as a crucial consideration for the IC specialist physicians in our study. Increased time efficiency had two primary implications: decreasing radiation exposure during procedures and improving workflow efficiency in already demanding clinical settings. The specialist physicians in our study noted only an indirect link between the use of *CardioVision* and reduced radiation exposure, suggesting that increased precision and shorter procedure times could potentially contribute to dose reduction. Supporting this is a review by Samant et al [8] that highlights how immersive technologies, specifically VR and AR, enhance visualization of complex structural heart defects, leading to reduced procedural times and radiation exposure. Park et al [33] also report that an AR software used in CT-guided needle procedures improved accuracy, reduced procedure time, and lowered radiation exposure, including enabling less experienced health care professionals to perform comparably to experts. The specialist physicians in our study further emphasized the need for ease of use and a minimal learning curve to ensure quick adoption of immersive technologies. Many immersive technologies will require some time invested to learn and integrate into clinical workflows [4]. The time needed to learn the software is important to consider when deciding to integrate it into clinical practice to minimize disruptions in busy IC departments.

CardioVision was perceived by the specialist physicians in our study as a valuable decision-support tool, particularly in complex IC procedures. The reported current applications for immersive technologies include diagnosis, preoperative planning, communication, and patient management [34]. Decision support, as highlighted in our study, also extended to workflow planning aspects such as preprocedure device selection. This is of particular importance in pediatric IC, where appropriately sized stents, catheters, and other essential devices may be scarce. Device limitations are reported to persist due to the lack of designs tailored to infants [35]. The specialist physicians in our study highlighted that improved planning and visualization tools, such as those offered by *CardioVision*, fostered greater procedural preparedness, minimized unexpected challenges, and enhanced physician confidence. Studies also report that immersive technologies contribute positively to physician confidence during medical procedures [4,36]. A study by Robertson et al [37] showed that VR modeling for surgical procedures increased surgeon confidence and, in some cases, altered the planned approach, with surgeons reporting improved mental preparation and a sense of familiarity during operations.

The *CardioVision* software was also seen as a facilitator of multidisciplinary collaboration, enabling health care professionals to jointly evaluate patient-specific cases. A study by Kim et al [38] reports preference in the use of VR for collaborative case discussions, emphasizing this role of immersive technologies. The specialist physicians in our study also appreciated the *CardioVision*'s potential to improve communication with patients to better understand their condition and treatment options. This is a benefit of immersive technologies reported in several studies [34,39,40]. For example, a study by El Mathari et al [40] showed that using VR for patient education enhanced satisfaction by delivering information in a way that patients found more informative and engaging. The *CardioVision* was also recognized by all the specialist physicians in our study as highly valuable for training and education. In education, the benefits of immersive technologies include offering realistic simulations of complex cardiac procedures and complementing traditional training methods by enabling students' repeated practice on procedures [4,41,42].

Integrating immersive technologies into clinical workflows, however, remains a challenge. In our study, the specialist physicians noted that while *CardioVision* holds promise, successful implementation will require careful attention to hardware compatibility, image quality, intuitive user interfaces, and workflow integration. Emphasis was placed on the need for close collaboration with developers to ensure such software meets the specific demands of IC departments. Furthermore, it is important to determine whether the software is designed specifically for clinicians, for patients, or for both [12]. While

XSL•FO
RenderX

the *CardioVision* software in our study was primarily developed for physicians, participants also highlighted its potential for adaptation to patient use, particularly to support understanding of their medical condition, although this will require further exploration. Annabestani et al [7] state that collaboration between medical specialists and technical experts is central to advancing the use of immersive technology in cardiac interventions. Co-development between clinicians, engineers, and researchers is essential to maximize clinical relevance and ensure these technologies are both effective and adoptable in routine practice [7]. Through close collaborations, medical imaging processes can be streamlined and imaging timelines accelerated [32], enabling the overall adoption of immersive technologies in routine health care practice. Annabestani et al [7] suggest that close collaboration with technology developers drives innovation by facilitating seamless integration of advanced tools, enabling precise, patient-specific procedures that enhance both safety and clinical outcomes.

Study Limitations

Our study had several limitations that should be acknowledged. First, the sample size in the quantitative analysis was relatively small, which limited the statistical power to detect meaningful variability or trends in the data. However, the quantitative data were complemented by rich insights obtained from qualitative interviews, which helped to contextualize and strengthen the overall interpretation of results.

Second, the participants had only a limited timeframe to evaluate the *CardioVision* software. This brief exposure may not have been sufficient for users to fully explore the system's capabilities or to become comfortable with its interface and functions. As a result, their reported experiences and perceived usefulness of the software may not accurately reflect longer-term use scenarios. Additionally, the study was conducted in a controlled environment. The few specialist physicians who participated in the study were from selected countries and health care centers. This may not capture the full complexity and workflow dynamics of routine clinical practice. However, the strategically selected specialist physicians who participated in the study are likely to provide expert opinion of transferable value. Future research with larger sample sizes, extended periods of use, and on-site evaluations across different clinical settings would provide a more comprehensive understanding of the software's potential and limitations. Our study, however, focuses on highlighting what IC specialist physicians consider valuable in immersive technologies. Using *CardioVision* as a case study helps to identify key areas for consideration when integrating immersive technologies into an IC practice.

Conclusions

This study highlights that IC specialist physicians perceive immersive technologies like *CardioVision* as valuable tools for enhancing procedural precision, anatomical visualization, and clinical decision-making. The software was described as having the potential to improve workflow efficiency, support preoperative planning, and facilitate multidisciplinary and patient collaboration. Integrating such technologies seamlessly into existing clinical workflows should consider the need for user-friendly interfaces and minimal learning curves to gain confidence for effective use. Future advancements should focus on co-developing solutions with clinicians to ensure that immersive technologies are tailored to meet the specific needs of IC practice, thereby maximizing clinical utility and adoption.

Acknowledgments

We would like to thank all the *HARMONIC* (Health effects of cardiac fluoroscopy and modern radiotherapy in pediatrics) project members who facilitated the process, with special thanks to Professor Hilde Olerud, project leader at the University of South-Eastern Norway, for her contributions. We also extend our gratitude to all the interventional cardiologists and physicians who participated in the study and provided essential input in evaluating the cases used for the *CardioVision* software. We acknowledge the use of ChatGPT by OpenAI for assistance in English language spelling and grammar checks of the manuscript. ChatGPT has not been used for any other content other than English proofreading.

Funding

The HARMONIC project received funding from the Euratom Research and Training Program 2014 - 2018 under Grant Agreement No. 847707. The development of CardioVision was supported through this funding. The present study was conducted independently of the funding body.

Data Availability

The datasets generated and analyzed during the current study are not publicly available but may be obtained from the corresponding author upon reasonable request.

Authors' Contributions

All authors participated in the design of the questionnaires and interview guide. AJ conducted the testing of the mixed reality platform with participants and was responsible for collecting the questionnaire data. CCC and LH conducted the interviews. All authors read and reviewed the manuscript.

XSL•FO
RenderX

Conflicts of Interest

AJ is a developer of the technology evaluated in this study and may have potential financial or professional interest in the results. All other authors declare no competing interests.

Multimedia Appendix 1
CardioVision study survey.
[PDF File, 306 KB - formative_v9i1e79278_app1.pdf]

Multimedia Appendix 2
Interview guide.
[PDF File, 98 KB - formative_v9i1e79278_app2.pdf]

References

1. Alsharqi M, Edelman ER. Artificial intelligence in cardiovascular imaging and interventional cardiology: emerging trends and clinical implications. J Soc Cardiovasc Angiogr Interv 2025 Mar;4(3Part B):102558. [doi: 10.1016/j.jscai.2024.102558] [Medline: 40230671]
2. Rao PS. Advances in the diagnosis and management of congenital heart disease in children. Children (Basel) 2023 Apr 21;10(4):753. [doi: 10.3390/children10040753] [Medline: 37190002]
3. Sachdeva R, Armstrong AK, Arnaout R, et al. Novel techniques in imaging congenital heart disease: JACC scientific statement. J Am Coll Cardiol 2024 Jan 2;83(1):63-81. [doi: 10.1016/j.jacc.2023.10.025] [Medline: 38171712]
4. Keramati H, Lu X, Cabanag M, Wu L, Kushwaha V, Beier S. Applications and advances of immersive technology in cardiology. Curr Probl Cardiol 2024 Oct;49(10):102762. [doi: 10.1016/j.cpcardiol.2024.102762] [Medline: 39067719]
5. Minga I, Al-Ani MA, Moharem-Elgamal S, et al. Use of virtual reality and 3D models in contemporary practice of cardiology. Curr Cardiol Rep 2024 Jun;26(6):643-650. [doi: 10.1007/s11886-024-02061-2] [Medline: 38683474]
6. Elsakka A, Park BJ, Marinelli B, Swinburne NC, Schefflein J. Virtual and augmented reality in interventional radiology: current applications, challenges, and future directions. Tech Vasc Interv Radiol 2023 Sep;26(3):100919. [doi: 10.1016/j.tvir.2023.100919] [Medline: 38071031]
7. Annabestani M, Olyanasab A, Mosadegh B. Application of mixed/augmented reality in interventional cardiology. J Clin Med 2024 Jul 26;13(15):4368. [doi: 10.3390/jcm13154368] [Medline: 39124633]
8. Samant S, Bakhos JJ, Wu W, et al. Artificial intelligence, computational simulations, and extended reality in cardiovascular interventions. JACC Cardiovasc Interv 2023 Oct 23;16(20):2479-2497. [doi: 10.1016/j.jcin.2023.07.022] [Medline: 37879802]
9. Afroz S, Østerås BH, Thevathas US, et al. Use of ionizing radiation in a Norwegian cohort of children with congenital heart disease: imaging frequency and radiation dose for the Health Effects of Cardiac Fluoroscopy and Modern Radiotherapy in Pediatrics (HARMONIC) study. Pediatr Radiol 2023 Nov;53(12):2502-2514. [doi: 10.1007/s00247-023-05774-8] [Medline: 37773444]
10. Harbron RW, Thierry-Chef I, Pearce MS, et al. The HARMONIC project: study design for the assessment of radiation doses and associated cancer risks following cardiac fluoroscopy in childhood. J Radiol Prot 2020 Sep 24;40(4). [doi: 10.1088/1361-6498/aba66d] [Medline: 32668420]
11. Jahnen A, Dabin J, Annoni GA, Tantar AA, Cools B, Thierry-Chef I. CardioVision: a software platform to bring artificial intelligence and mixed reality to pediatric cardiology. Presented at: 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); Dec 3-6, 2024. [doi: 10.1109/BIBM62325.2024.10822130]
12. Higaki A, Watanabe Y, Akazawa Y, et al. Automated categorization of virtual reality studies in cardiology based on the device usage: a bibliometric analysis (2010-2022). Eur Heart J Digit Health 2023 Mar;4(2):119-124. [doi: 10.1093/ehjdh/ztad008] [Medline: 36974263]
13. Iqbal AI, Aamir A, Hammad A, et al. Immersive technologies in healthcare: an in-depth exploration of virtual reality and augmented reality in enhancing patient care, medical education, and training paradigms. J Prim Care Community Health 2024;15:21501319241293311. [doi: 10.1177/21501319241293311] [Medline: 39439304]
14. Lara Hernandez KA, Rienmüller T, Baumgartner D, Baumgartner C. Deep learning in spatiotemporal cardiac imaging: a review of methodologies and clinical usability. Comput Biol Med 2021 Mar;130:104200. [doi: 10.1016/j.compbiomed.2020.104200] [Medline: 33421825]
15. Kouijzer M, Kip H, Bouman YHA, Kelders SM. Implementation of virtual reality in healthcare: a scoping review on the implementation process of virtual reality in various healthcare settings. Implement Sci Commun 2023 Jun 16;4(1):67. [doi: 10.1186/s43058-023-00442-2] [Medline: 37328858]
16. JavaScript. Kitware/glance. 2024 Sep 26. URL: https://github.com/Kitware/glance [accessed 2024-10-10]
17. Fedorov A, Beichel R, Kalpathy-Cramer J, et al. 3D Slicer as an image computing platform for the quantitative imaging network. Magn Reson Imaging 2012 Nov;30(9):1323-1341. [doi: 10.1016/j.mri.2012.05.001] [Medline: 22770690]

XSL•FO
RenderX

18. Li W, Brown R, et al. MONAI: an open-source framework for deep learning in healthcare. Arxiv. Preprint posted online on 2022. [doi: 10.48550/arxiv.2211.02701]
19. Palinkas LA, Horwitz SM, Green CA, Wisdom JP, Duan N, Hoagwood K. Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. Adm Policy Ment Health 2015 Sep;42(5):533-544. [doi: 10.1007/s10488-013-0528-y] [Medline: 24193818]
20. Moser A, Korstjens I. Series: Practical guidance to qualitative research. Part 3: Sampling, data collection and analysis. Eur J Gen Pract 2018 Dec;24(1):9-18. [doi: 10.1080/13814788.2017.1375091] [Medline: 29199486]
21. Lewis JR. The system usability scale: past, present, and future. Int J Human Comput Interact 2018 Jul 3;34(7):577-590. [doi: 10.1080/10447318.2018.1455307]
22. Joshi A, Kale S, Chandel S, Pal D. Likert scale: explored and explained. BJAST 2015;7(4):396-403. [doi: 10.9734/BJAST/2015/14975]
23. Nettskjema diktafon mobile application. University of Oslo. 2024. URL: https://www.uio.no/english/services/it/adm-services/nettskjema/help/nettskjema-dictaphone.html [accessed 2025-06-25]
24. Hennink MM, Kaiser BN, Marconi VC. Code saturation versus meaning saturation: how many interviews are enough? Qual Health Res 2017 Mar;27(4):591-608. [doi: 10.1177/1049732316665344] [Medline: 27670770]
25. Elo S, Kyngäs H. The qualitative content analysis process. J Adv Nurs 2008 Apr;62(1):107-115. [doi: 10.1111/j.1365-2648.2007.04569.x] [Medline: 18352969]
26. Schoonenboom J, Johnson RB. How to construct a mixed methods research design. Kolner Z Soz Sozpsychol 2017;69(Suppl 2):107-131. [doi: 10.1007/s11577-017-0454-1] [Medline: 28989188]
27. Conti CR, Conti JB. The Netter Collection of Medical Illustrations: Cardiovascular System: Elsevier Health Sciences; 2023, Vol. 8 E-Book. URL: https://evolve.elsevier.com/cs/product/9780323881296?role=student [accessed 2025-11-06]
28. Stephenson N, Pushparajah K, Wheeler G, Deng S, Schnabel JA, Simpson JM. Extended reality for procedural planning and guidance in structural heart disease - a review of the state-of-the-art. Int J Cardiovasc Imaging 2023 Jul;39(7):1405-1419. [doi: 10.1007/s10554-023-02823-z] [Medline: 37103667]
29. Lee SY, Squelch A, Sun Z. Investigation of the clinical value of four visualization modalities for congenital heart disease. J Cardiovasc Dev Dis 2024 Sep 5;11(9):278. [doi: 10.3390/jcdd11090278] [Medline: 39330336]
30. Tsai TY, Onuma Y, Złahoda-Huzior A, et al. Merging virtual and physical experiences: extended realities in cardiovascular medicine. Eur Heart J 2023 Sep 14;44(35):3311-3322. [doi: 10.1093/eurheartj/ehad352] [Medline: 37350487]
31. Lau I, Gupta A, Ihdayhid A, Sun Z. Clinical applications of mixed reality and 3D printing in congenital heart disease. Biomolecules 2022 Oct 24;12(11):1548. [doi: 10.3390/biom12111548] [Medline: 36358899]
32. Narasimhan SL, Gherciuc SO, Amer D, Iaizzo PA. Multimodality imaging in congenital heart disease. In: Handbook of Cardiac Anatomy, Physiology, and Devices, 4th edition: Springer Cham; 2024:451-475. [doi: 10.1007/978-3-031-72581-4_23]
33. Park BJ, Hunt SJ, Nadolski GJ, Gade TP. Augmented reality improves procedural efficiency and reduces radiation dose for CT-guided lesion targeting: a phantom study using HoloLens 2. Sci Rep 2020 Oct 29;10(1):18620. [doi: 10.1038/s41598-020-75676-4] [Medline: 33122766]
34. Tortora M, Luppi A, Pacchiano F, et al. Current applications and future perspectives of extended reality in radiology. Radiol Med 2025 Jun;130(6):905-920. [doi: 10.1007/s11547-025-02001-2] [Medline: 40153208]
35. Kenny D. Interventional cardiology for congenital heart disease. Korean Circ J 2018 May;48(5):350-364. [doi: 10.4070/kcj.2018.0064] [Medline: 29671282]
36. Pushparajah K, Chu KYK, Deng S, et al. Virtual reality three-dimensional echocardiographic imaging for planning surgical atrioventricular valve repair. JTCVS Tech 2021 Jun;7:269-277. [doi: 10.1016/j.xjtc.2021.02.044] [Medline: 34100000]
37. Robertson DJ, Abramson ZR, Davidoff AM, Bramlet MT. Virtual reality applications in pediatric surgery. Semin Pediatr Surg 2024 Feb;33(1):151387. [doi: 10.1016/j.sempedsurg.2024.151387] [Medline: 38262206]
38. Kim B, Loke YH, Mass P, et al. A novel virtual reality medical image display system for group discussions of congenital heart disease: development and usability testing. JMIR Cardio 2020 Dec 8;4(1):e20633. [doi: 10.2196/20633] [Medline: 33289675]
39. Jung C, Wolff G, Wernly B, et al. Virtual and augmented reality in cardiovascular care: state-of-the-art and future perspectives. JACC Cardiovasc Imaging 2022 Mar;15(3):519-532. [doi: 10.1016/j.jcmg.2021.08.017] [Medline: 34656478]
40. El Mathari S, Kuitert L, Boulidam N, et al. Evaluating virtual reality patient education in cardiac surgery: impact on preoperative anxiety and postoperative patient satisfaction. J Clin Med 2024 Oct 31;13(21):6567. [doi: 10.3390/jcm13216567] [Medline: 39518707]
41. Silva JNA, Southworth M, Raptis C, Silva J. Emerging applications of virtual reality in cardiovascular medicine. JACC Basic Transl Sci 2018 Jun;3(3):420-430. [doi: 10.1016/j.jacbts.2017.11.009] [Medline: 30062228]
42. von Ende E, Ryan S, Crain MA, Makary MS. Artificial intelligence, augmented reality, and virtual reality advances and applications in interventional radiology. Diagnostics (Basel) 2023 Feb 27;13(5):892. [doi: 10.3390/diagnostics13050892] [Medline: 36900036]

XSL•FO
RenderX

Edited by A Stone, A Mavragani; submitted 25.06.25; peer-reviewed by A Higaki, NR Lawrence; revised version received 24.09.25; accepted 24.09.25; published 19.11.25.

Please cite as:
Chilanga CC, Jahnen A, Hafskjold L
Usability Evaluation of a Mixed Reality Platform in Pediatric Interventional Cardiology by Specialist Physicians: Mixed Methods Study
JMIR Form Res 2025;9:e79278
URL: https://formative.jmir.org/2025/1/e79278
doi:10.2196/79278

© Catherine Chilute Chilanga, Andreas Jahnen, Linda Hafskjold. Originally published in JMIR Formative Research (https://formative.jmir.org), 19.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Shesha, a WhatsApp Chatbot for Linking Household Contacts to Tuberculosis Treatment or Preventive Therapy in South Africa: Design and Development

Don Lawrence Mudzengi[1,2], BSc, MSc; Thobani Ntshiqa[1,2], BCur, MPH, MBA; Yohhei Hamada[3], MD, MPH, PhD; Felex Ndebele[1], BPharm, MPH; Thapelo Mpanza[1,2], BSw; Bridget Kyobutungi[4], BPharm, MPH; Candice Williams[4], BA Hons; Meghan Kennealy[4], BSc, MSc; Molebogeng Rangaka[3], MD, PhD; Kavindhran Velen[1], BSc, MPH, PhD; Salome Charalambous[1], MB BCh, MSc, PhD

[1]The Aurum Institute, 3 Wrench Road, Isando, Johannesburg, Gauteng, South Africa
[2]School of Public Health, University of the Witwatersrand, Johannesburg, South Africa
[3]Institute of Global Health, University College London, London, United Kingdom
[4]Reach Digital Health, Johannesburg, South Africa

Corresponding Author:
Don Lawrence Mudzengi, BSc, MSc
The Aurum Institute, 3 Wrench Road, Isando, Johannesburg, Gauteng, South Africa

Abstract

Background: Literature on the development of mobile health (mHealth) tools for public health interventions is scarce. This scarcity creates a knowledge gap, and new tools may repeat the mistakes of past implementations.

Objective: In this paper, we describe the development of Shesha, a WhatsApp-based chatbot designed to facilitate linkage to care for household contacts of people being treated for tuberculosis (TB). Shesha facilitates linkage by providing TB test results, TB preventive treatment (TPT) information, nudges, reminders, and personalized support. We developed Shesha to address the human resource capacity challenges posed by South Africa's new universal TB testing and TPT policies.

Method: We applied a design thinking framework with 7 phases: empathize, discover, define, prototype, build and launch, improve, and evaluate. The process started with gathering insights from TB contact tracing studies and consulting with global and local experts to address the challenges of universal TB testing and TPT. Based on these findings, we defined the core functionalities of Shesha and incorporated them in the Health Belief Model to encourage health-seeking behavior. In collaboration with the developers, we developed the WhatsApp-based chatbot. We selected WhatsApp for its wide accessibility and user-friendliness.

Results: We successfully developed and launched the Shesha in September 2023, with implementation expected to continue until March 2025. Early user acceptance revealed that users generally valued the information provided on the tool regarding TB and TPT; however, they required ongoing engagement to link to care. Ongoing evaluations, guided by the Reach Effectiveness Adoption Implementation Maintenance (RE-AIM) framework, will assess the tool's impact on reducing community health worker workloads and improving linkage to care.

Conclusions: Documenting the development of mHealth technologies is crucial for guiding future projects and improving health interventions. In our study, frameworks like design thinking and the Health Belief Model aligned Shesha with user needs and programmatic goals. Comprehensive documentation may help assess the chatbot's performance and guide future improvements, supporting scalability and efficiency in mHealth interventions across public health settings.

(JMIR Form Res 2025;9:e71793) doi:10.2196/71793

KEYWORDS

contact tracing; design thinking; development; Health Belief Model; linkage to care; mHealth; patient engagement; tuberculosis; tuberculosis preventive treatment; universal TB testing; WhatsApp chatbot; Shesha

Introduction

South Africa's tuberculosis (TB) control program is increasingly turning to mobile health (mHealth) tools to support service delivery across the care cascade. Yet, most of these tools remain poorly evaluated and underdocumented [1,2], limiting opportunities for replication and adaptation at scale. In 2019, the South African TB Think Tank, a government-led consortium of TB experts, found that none of the 10 mHealth tools they

XSL•FO
RenderX

reviewed met the government standards for design simplicity and robust evaluation [3], and our 2024 systematic review had similar findings, including the absence of documenting how the tools were developed [4]. Franck and colleagues further argue that innovation can only advance if successes and failures are openly documented; as they put it, “to err is evolution” [5]. Although recent studies from Argentina [6], the United Kingdom [5], Tanzania [6], and Uganda [7] have begun setting a precedent, evidence for TB-focused tools in South Africa remains limited.

The urgency to close this documentation gap is evident in South Africa’s TB program, responsible for controlling one of the largest epidemics globally, with 270,000 new people infected with TB and 25,000 deaths annually [8]. The country runs an extensive contact tracing program to find TB among household contacts, but it relies on a labor-intensive cascade of *finding the contacts, screening them, collecting specimens, delivering results and reminders, and linking them to care* [9-13]. This approach is characterized by repetitive household visits by community health workers (CHWs), a model already proven to be suboptimal [9,12,13]. In South Africa, as seen in other African contexts, this model also continues to face challenges such as contacts not being traced, high dropout rates along the cascade, poor clinic attendance, and generally exorbitant costs for its delivery [13,14]. In 2023, South Africa introduced a new policy that expanded and further strained the delivery of TB contact tracing done by CHWs. The policy requires all household contacts to undergo microbiological testing and makes every contact who tests negative now eligible for tuberculosis preventive treatment (TPT) [15,16]. Under the previous regime, microbiological testing was limited to symptomatic contacts, and TPT eligibility applied only to children under 5 years and people living with HIV. This new policy targets 250,000 TPT initiations in contacts by 2026, from only 17,000 [16], a 15-fold increase which without innovation cannot be realistically met by CHWs. The enormity of this burden thus requires digital solutions whose evidence remains limited in the country.

We responded to the demands of the new policy and the need to support CHWs in contact tracing by developing Shesha. This mHealth tool was a WhatsApp (WhatsApp LLC, Menlo Park, CA, USA) chatbot, which enabled household contacts to undertake repetitive activities usually performed by CHWs, such as receiving test results, engaging with motivational messages, receiving clinic visit reminders, and confirming linkage to care. By reallocating these tasks, Shesha could free up time for CHWs to focus on reaching more households and meeting the expanded obligations under the new policy. At the time we developed Shesha, we were not aware of any tool in South Africa that allowed household contacts to assume this role. In this paper, we document Shesha’s development and share the design choices and processes behind it, to provide an evidence base that others working in TB mHealth can use or adapt.

Methods

Setting and Context

We embedded the Shesha development in the intervention arm of the Community and Universal Testing for Tuberculosis among Contacts (CUT-TB) trial, a European and Developing Countries Clinical Trials Partnership (EDCTP)–funded cluster-randomized trial [17]. The primary objective of this 2-phase trial was to evaluate the effectiveness of universal testing compared with standard TB screening for detecting TB among household and community contacts.

We implemented in the Ekurhuleni district, one of the 5 districts in the Gauteng Province in South Africa. Ekurhuleni is one of the most populous districts in the Gauteng province, with a growing population estimated to be just over 4 million residents (27% of Gauteng's population) spread over 1975 km^2. Most residents in Ekurhuleni are multilingual [18,19]; however, about 34% residents are native IsiZulu speakers, a language also intelligible to an additional 8% who speak isiXhosa, and a much smaller fraction who speak isiNdebele and siSwati [18]. Ekurhuleni is highly urbanized and serves as an industrial hub, but it also experiences high levels of poverty, as evidenced by numerous informal settlements [20]. These settlements are characterized by a high burden of TB and other infectious diseases. It accounted for 25% of Gauteng's 31,704 TB diagnoses reported in 2024 [21] and the district has the highest HIV prevalence among pregnant women in Gauteng province at 28.5%, according to a recent antenatal sentinel survey [22]. Internet connectivity and smartphone ownership have grown rapidly. According to Statistics South Africa, about 82% of the population now has internet access, with 75.7% in Gauteng accessing the internet using mobile phones which can serve as a proxy for smartphone ownership [23]. Although we implemented the Shesha in Ekurhuleni, we did not develop it as a district-specific tool; instead, its deployment in Ekurhuleni was primarily operational due to the location of the CUT-TB trial.

Study Population

Our target population for the Shesha chatbot consisted of household contacts of TB patients in the Ekurhuleni District. With the expanded contact tracing policy in South Africa, all contacts underwent microbiological testing. Our study staff collected sputum from all contacts who could provide, regardless of whether they exhibited symptoms or not. Those who tested positive were referred for TB treatment; those testing negative were directed to start TPT. The latter group posed a challenge because most were generally healthy, with low perceived risk, minimal interest in visiting clinics, and limited motivation to start treatment. The study staff enrolled eligible and consenting household contacts into Shesha on the day they collected the sputum. Both TPT and the use of the Shesha tool were entirely new to many of these contacts. This low-risk perception and health-seeking behavior are common among household contacts. A South African TB prevalence survey found that over 60% of symptomatic household members delayed seeking care, often only doing so after prompting [24].

XSL•FO
RenderX

Design Thinking Framework for Developing Shesha

We used a design thinking framework for the Shesha development process (see Figure 1). This framework has 7 phases: empathize, discover, define, prototype, build and launch, improve, and evaluate [25]. Some studies on case finding have used these design thinking frameworks as structured methodologies for solving complex problems and creating user-centric solutions [26,27]. These frameworks ensure that user needs are addressed while achieving public health goals and incorporate participatory consultation with stakeholders. The following sections describe the activities conducted under each phase of the design thinking framework.

Figure 1. Design thinking framework. CUT-TB, Community and Universal Testing for Tuberculosis among Contacts; TB, tuberculosis; WHO, World Health Organization.

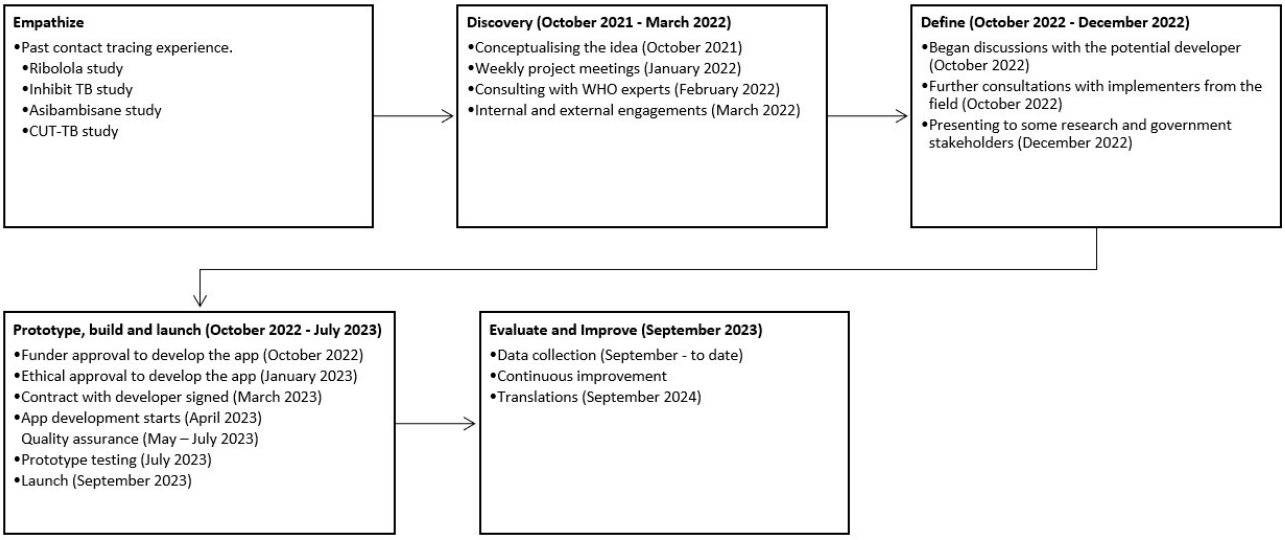

Empathize

The empathize phase involved data collection to understand user needs and ensure that solutions we create, such as Shesha, align with their challenges. However, in our case, the situation we faced with the targeted users made conventional approaches infeasible, such as interviewing them before development. As we outlined in the study population section, the household contacts that we aimed to support with Shesha had no prior exposure to TPT, and many were unaware of their risk or that they were eligible for any preventive care. Within this context, therefore, we determined that engaging them upfront could have generated speculation rather than any actionable input into Shesha's design. Despite the potential weaknesses of our approach and the circumstances that were identified, it still recognized the importance of capturing user perspectives. Immediately after launching Shesha, an early user satisfaction survey was conducted for all contacts enrolled up to March 2024. While not equivalent to a textbook predeployment empathize phase, this offered authentic user perspectives at the earliest feasible opportunity in our process.

Our judgment of the circumstances limiting us from getting early insights led us to leverage existing evidence and our extensive experience as researchers in contact tracing. This approach also aligns with other established guidance that behavior change interventions can be informed by theory, prior evidence, and formative research [28]. We therefore drew on insights from our previous contact tracing studies, which include Ribolola [29], Inhibit-TB [13], Asibambisane [19], and the current CUT-TB trial [17]. These studies collectively identified the central challenges of TB contact tracing: poor linkage to care, high attrition, overstretched CHWs, and the pressures created by universal testing. They also highlighted specific needs and gaps, such as the value of community-based care (Ribolola), the importance of reliable scheduling and repeated follow-up (Inhibit-TB), the limits of CHW capacity and variability in workloads (Asibambisane), and the staffing shortages and policy demands associated with universal testing (CUT-TB). We designed and developed, incorporating this understanding of the contact tracing program. Shesha thus featured mobile result sharing with contacts, educational content on TB and TPT via WhatsApp, automated follow-ups, passive clinic scheduling support, and treatment initiation verification.

Discover

Overview

During the discovery phase, we gathered insights into both the challenges and potential solutions for universal TB testing and universal TPT. We engaged global and local experts through a participatory process to evaluate whether Shesha was an appropriate response to these challenges and to validate its relevance in addressing the identified issues, while also securing their buy-in.

Internal Discussions

In late 2021, before engaging external stakeholders, we held internal discussions between Aurum and University College London researchers to refine the concept of Shesha. These discussions were part of the weekly operational meetings for the CUT-TB project. We brainstormed ideas on how to create a system that would enable linkage to care for household contacts of people with TB after the introduction of the universal testing and TPT policy for all household contacts. We acknowledged that while universal testing could take place

XSL•FO
RenderX

during household visits, maintaining consistent follow-up and ensuring successful linkage to care posed a challenge too complex for CHWs to manage alone, which led to the idea of developing Shesha.

External Engagements

At the beginning of 2022, after brainstorming and finalizing a draft of our concept, we began approaching external experts. We first held a meeting in February 2022 with representatives from the World Health Organization to gather their input on how to approach the overall change in South Africa's TPT policy. We also presented our concept for designing a tool to facilitate linkage to care for these contacts under the universal testing policy. Their main suggestion to our idea was to add options that would allow enrolling community contacts if feasible. Community contacts are people who do not live in the same household as the person with TB but have regular contact with them.

From this consultation, we further developed our concept within the internal team and started ad hoc discussions with potential developers. In March 2023, an organization was approached to explore using an existing tool called CommCare but did not materialize, and then in August 2022, discussions were held with the developers whom we finally selected.

For the external stakeholders, we consulted public health officials from the health departments and TB/HIV program implementers, mainly from the Global Fund to Fight AIDS, Tuberculosis and Malaria and HIV the President's Emergency Plan for AIDS Relief (PEPFAR). The local stakeholders generally agreed that Shesha had the potential to address some of the challenges in the contact tracing cascade, especially after the introduction of a universal policy, by automating follow-up processes and alleviating the burden on CHWs. They also recommended a rigorous evaluation of the tool.

Define

Theoretical Reasoning Behind Shesha

We developed Shesha as a "patient-facing" WhatsApp chatbot to streamline the linkage to care for contacts of people with TB. It allowed the contacts to interact with it by sending and receiving messages in automated prompts that mimic key follow-up tasks which CHWs usually perform. These are primarily administrative, repetitive tasks that include sending TB results, confirming contacts read results, visiting the clinic, starting treatment, supporting adherence, and confirming completion of treatment. By automating these tasks, Shesha potentially reduces the need for manual follow-up by CHWs, thereby alleviating their workload from repeated home visits.

We based our design on the Health Belief Model (HBM) to capture underlying drivers of health-related behavior change [30]. The use of HBM in our study helped predict and promote health behaviors by addressing attitudes, beliefs, and perceptions about TB, its treatment, and prevention. It allowed us to tailor messages to the user's psychological and emotional motivators, drawing them closer to the desired action of starting treatment after receiving results. The HBM is well suited for our target group, the household contacts of people with TB, who do not inherently perceive themselves as susceptible to the disease. Unlike alternative theories such as the Theory of Reasoned Action and the Theory of Planned Behavior, which assume that individuals already recognize a need for action and rely on rationality in making decisions, the HBM uses individual constructs for people who have not made a decision.

Shesha operationalizes all HBM domains to guide users through the TB contact tracing and linkage cascade. Household contacts have a low *perceived susceptibility* to TB risk, as well as *perceived severity* if they are infected. Most are unaware of TPT; hence, their perception of its benefits is poor. We, however, provide information to reduce obstacles that might hinder their access to care and address their perceived barriers to care. Shesha continuously engages users by providing cues to action, sending reminders, prompting clinic visits, encouraging medication adherence, and offering health messaging. To strengthen their self-efficacy in staying on the linkage journey, Shesha delivers precise, easily understandable instructions on how to visit the clinic, what to plan, and an intuitive WhatsApp interface with which most are familiar. This approach aimed to encourage users to engage with Shesha messages, visit the clinic, and start and adhere to treatment, reducing attrition along the contact tracing cascade. Figure 2 details how each domain of the HBM has been applied.

XSL•FO
RenderX

Figure 2. Health Belief Model. TB: tuberculosis; TPT: tuberculosis preventive treatment.

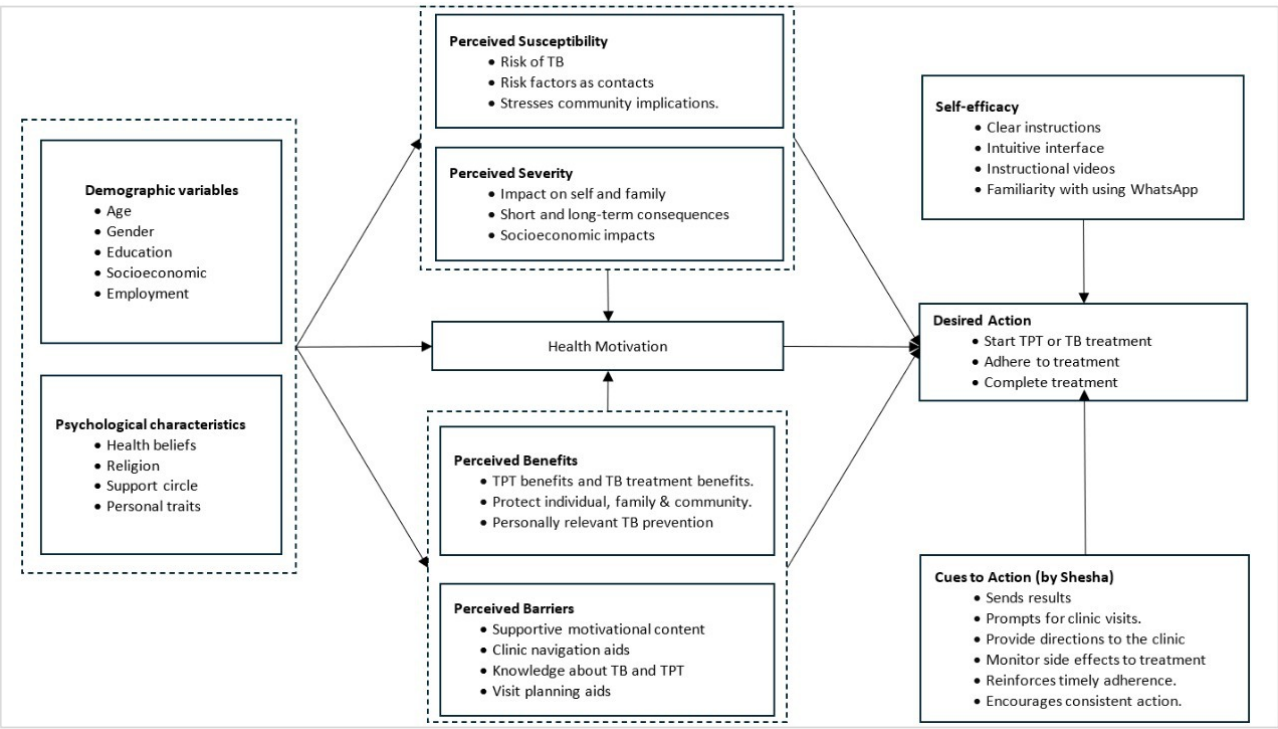

Engaging With Developers

We identified the developers whom we considered an ideal development partner for Shesha due to their extensive experience with mHealth tools, such as the COVID Alert and TB-Alert apps [31]. Their familiarity with the South African health landscape and local infrastructure also made them particularly suitable for this project. We negotiated contracts to address the scope of the work, design, and key legal issues, mainly how to handle intellectual property emanating from the Shesha design, publication rights. We used feedback from stakeholder engagements to refine the project scope. In early 2023, the funder approved the proposal to develop Shesha. After this, we began translating the findings from the empathize phase and discovery stages to define the chatbot's key functionalities and platform choices. We also outlined roles and responsibilities for the development team between the two organizations; the research provided the concepts with a wireframe showing how Shesha should function, and developers translated that into a chatbot prototype. We continued with regular meetings with the developers throughout this journey up to the end of the implementation.

Choosing the Platform to Develop Shesha

Our budget permitted us to choose only 1 platform, so we selected WhatsApp. However, our initial presentations to local stakeholders and the developers were based on an Unstructured Supplementary Service Data (USSD) system. Local stakeholders preferred WhatsApp because it was more likely to be sustained by the government compared to alternatives like SMS or USSD, which entail high long-term provider costs. A similar mHealth tool, "MomConnect," cost approximately US $60,000 monthly in 2018, even after being subsidized by the network provider, and it was based on USSD and SMS [32]. MomConnect now incorporates WhatsApp. WhatsApp is also now widely used in South Africa, where over 45 million people use the internet, and at least 94% of them actively use WhatsApp [33,34].

The developers also recommended WhatsApp based on precedent; it had been widely used in South Africa for COVID-19 self-screening. Some of the key points that informed our decision-making process are outlined in Table 1. Our design requirements included an intuitive interface, an effective notification system, 2-way interaction with sustained engagement, and the ease of having and maintaining push and pull messages between users and the system. This level and frequency of engagement were only possible on WhatsApp. WhatsApp's end-to-end encryption safeguards sensitive health information, addressing key privacy concerns in handling medical data. We did not plan to provide the users with data bundles to use the system; however, our design was also lean in that it did not include any multimedia, which consumes a lot of data for the participants.

XSL•FO
RenderX

Table . Decision-making discussion with the developers, comparing with Unstructured Supplementary Service Data (USSD).

	WhatsApp	USSD[a]
Pros	1. It is now the most popular communication platform. 2. There are now low-cost phones that support WhatsApp; the developers advised that most of our participants will have access. 3. The proposed WhatsApp messages are low-cost even to the recipient; therefore, the impact on their affordability will be minimal. 4. WhatsApp supports push services where a service provider can initiate interaction. 5. It is ubiquitous.	1. It is available on any phone, including those that do not support WhatsApp. 2. It is ubiquitous. 3. It can be reverse billed, so there is no cost to the participant.
Cons	1. Some phones may not support. WhatsApp but low-cost phones are increasingly being made with WhatsApp support. 2. Participants will always need data to receive and send messages.	1. It is an old technology that is nearing obsolescence. 2. Costly to the provider, charges 20c per 20 seconds. 3. Does not support a push service. Follow-ups are best done using push messages to eliminate the need for a participant to initiate communication.

[a]Unstructured Supplementary Service Data.

Shesha Workflows

We developed the workflows of Shesha to define the journey contacts take as they navigate through linkage and adherence processes. Figure 3 presents the actual workflows that were conceptualized and eventually translated into Shesha.

Figure 3. Shesha workflows.

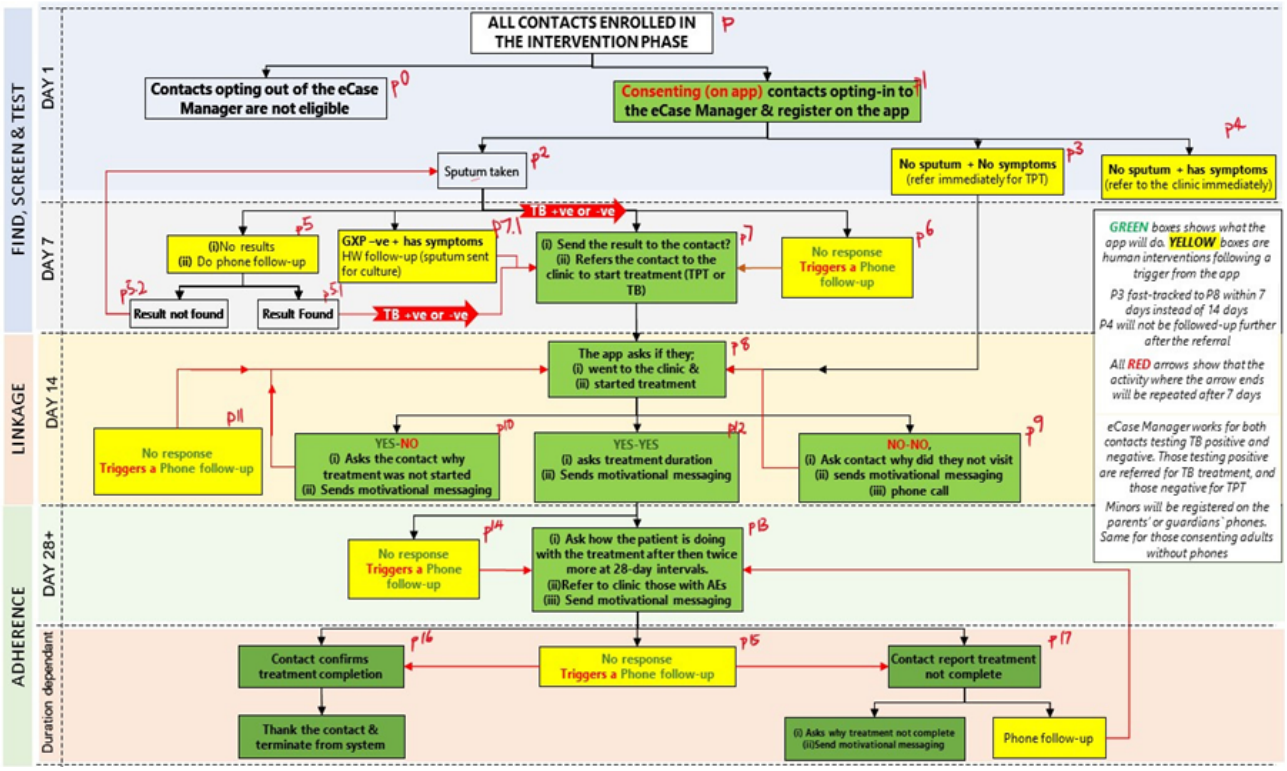

Key Staff Involved in Implementing Shesha

Table 2 outlines the core personnel responsible for the development and implementation of Shesha. Scientists and researchers conceptualized and designed workflows and user journeys. Research assistants registered household contacts, taught participants how to use the chatbot, and collected sputum samples during household visits. The developers and their technical team translated our design and workflows into a

XSL•FO
RenderX

prototype, conducted quality assurance, and built a real-time monitoring dashboard.

While we recognized that Shesha would ideally automate most interactions, there would be situations requiring direct human intervention. We therefore introduced a "Human Case Manager," mainly responsible for addressing nonresponses to Shesha and pitfalls such as missing results, queries from both contacts and field staff, and any other issues arising from the implementation. The Human Case Manager would step in after receiving a system-triggered alert to intervene. Alerts were included for nonresponses to messages or prompts after 3 unsuccessful automated follow-ups, or if a user had not received test results within a reasonable timeframe.

Table . Core implementation team.

Role	Responsibilities
Internal	
Scientists and researchers mainly from Aurum and University College London	• Led conceptualization and design • Developed workflows and user journeys • Coordinated integration with other teams • Ensured alignment with study objectives and public health goals • Oversaw data collection and analysis
RAs[a] and CHWs[b]	• Registered household contacts • Taught chatbot usage • Captured contact details • Screened for TB[c] symptoms • Collected sputum samples • Provided initial support and guidance during setup
Human Case Manager	• Provided oversight and intervention • Managed nonresponses • Guided treatment journeys • Contacted users who did not respond to automated prompts • Ensured user engagement and adherence to treatment protocols • Addressed any possible TPT[d] if contact requested
External	
Developers and technical team	• Developed the prototype • Conducted quality assurance • Built a real-time monitoring dashboard • Provided operational manuals and user experience maps • Ensured system reliability and effectiveness • Delivered training and support for system implementation • Continued support and fixing of bugs

[a]RAs: research assistants.

[b]CHWs: community health workers.

[c]TB: tuberculosis.

[d]TPT: tuberculosis preventive treatment.

Prototype, Build, and Launch

The building of the prototype started from April 2023 to July 2023, with weekly meetings between researchers and developers to track progress and ensure workflows remained aligned with our original plan. We used Figma (Figma Inc., San Francisco, CA, USA), a collaborative software, to design and visualize user journeys. We compared the workflows with the original Shesha concept (Figure 3) developed. The developers used these visual workflows to translate into a chatbot. For the development, they used TextIt (TextIt, Seattle, WA, USA), a platform that enables the rapid building of user journeys. TextIt was connected to a WhatsApp Business Application Programming Interface through a linkage platform called Turn.io (Turn.io PBC, Denver, CO, USA).

We used local guidelines for TB control and prevention to tailor messages, segmenting them by gender and age to provide personalized engagement. Messages covered key aspects such as symptom monitoring, treatment reminders, and motivational cues, reinforcing the core principles of the HBM by addressing perceived susceptibility, severity, and benefits while providing actionable prompts for care-seeking behavior.

Although we designed Shesha to accommodate multiple users on a single device, WhatsApp natively supports only single-user functionality. Therefore, the developers implemented a structured workaround using the logic architecture of TextIt. Each household contact was assigned a unique set of metadata fields at registration (eg, name, sex, age, screening date, test result), which were stored and used to distinguish individuals sharing the same phone number. TextIt's flow engine used conditional logic (eg, "if contact_name = X, then routed to Flow

XSL•FO
RenderX

A”) to direct messages and automated tasks to the correct user profile. These flows were configured using TextIt’s visual builder, enabling structured branching, user-specific reminder intervals, and status tracking at the individual level. During development, the developers created test scenarios simulating households with 2-5 users on a single device to test conversation threading, state memory, and flow accuracy. The developers also conducted functional quality assurance testing using TestRail (Gurock Software GmbH, Frankfurt, Germany) to verify that messages were correctly triggered per individual, that switching between users could occur without data loss, and that no cross-contamination of prompts occurred. No logic failures were observed in controlled tests, though minor user confusion about navigating between individuals was noted during rollout. This approach has proven stable for typical household sizes (2 - 4 users). However, it is still possible that scaling up for larger households may require interface or onboarding enhancements in future iterations.

System monitoring was implemented via a custom dashboard using Amazon Web Services Redshift (Amazon Web Services, Inc., Seattle, WA, USA) and Tableau (Tableau Software, LLC, Seattle, WA, USA) to track flow completion per user, message delivery success, and dropout triggers. The developers managed the Tableau dashboard, which the research team used for daily, interactive tracking of Shesha’s performance by dates and flows. The dashboard displayed raw counts of real-time interactions, which required later data cleaning and deduplication. The research team did not have direct access to modify the dashboard as needed, requiring relaying of information to developers for any change. Nevertheless, the dashboard provided a practical tool for high-level monitoring of key metrics during development.

Full approval for deployment was granted on September 11, 2023, with the service implemented until March 2025.

Evaluate and Improve

We developed an evaluation plan based on the Reach-Effectiveness Adoption Implementation Maintenance (RE-AIM) framework. The stakeholders had recommended a rigorous evaluation, and, in prior work, we had identified a scarcity of evaluations of mHealth tools.

Evaluate

We adopted the RE-AIM framework, which assesses 5 key dimensions: reach, effectiveness, adoption, implementation, and maintenance.

Improve

This phase focused on further refinements as we progressed with the implementation. It included addressing bugs and incorporating any user feedback. We continued with regular check-ins with the developers to resolve issues promptly and ensure that the chatbot functioned optimally. The major formal review of Shesha took place in October 2024 through a co-creation workshop held with an external provider to assess Shesha’s human-centeredness, 1 year after enrolling participants. This was made possible due to the additional funding we applied for and secured from a different funder to continue implementing into 2025. We also used these funds to make some changes, mainly an additional language. From the time of its launch, we had been aware of a limitation in having only English as the language on Shesha, so we translated into isiZulu and launched on December 11, 2024, the most widely spoken language in Ekurhuleni. Even with additional funding, budget constraints limited us to only 1 extra language.

Costing

We calculated Shesha’s development cost using actual invoices paid to the developer, which included capital costs (a 1-time outlay) and recurrent costs. Capital costs were converted into economic costs over 5 years to account for opportunity costs and depreciation, while recurrent costs were not adjusted. The Global Health Costing Consortium acknowledges uncertainties about the useful lives of new technologies [35]. We therefore defaulted to the recommended 5-year useful life assumption, which is used when a useful life is unknown and has been adopted by various projects on mHealth [36-38]. For the discount rate, we used 8%, which reflects the average public interest rate in South Africa, and an exchange rate of 18.33 US dollars.

Ethical Considerations

The Human Research Ethics Committee (Medical) approved the development and implementation of Shesha as an amendment to the CUT-TB trial under approval number 210107. This amendment revised the project objectives to accommodate the new universal testing and TPT policies. We considered several ethical concerns common when developing Shesha, including informed consent, data privacy, and more complex challenges such as phone sharing. For the actual enrollment of contacts, we first took participants through a consenting process to explain the purpose of the study, potential risks, how their data would be protected, and their rights as participants. In addition to the consent, we provided participants with a study-specific privacy policy, which is required by South Africa’s Protection of Personal Information Act. It explained the storage and processing of personal data. The privacy policy also outlined mechanisms involving multiple users and the risk of unintentional disclosure. We only registered users when both primary and secondary users agreed to the policy. We also added safeguards to Shesha’s design; messages used clear, neutral language, avoiding people’s names or sensitive health terms. At registration, participants were encouraged to use aliases instead of their real names. Test results were delivered only after the user entered a self-created pin, and recruiters trained participants during onboarding on how to use security features such as screen locking to reduce the risk of unauthorized access. Users could opt out at any time via a visible menu option available throughout the platform.

Results

Performance

Shesha’s performance was tracked through a Tableau dashboard that presented both counts and percentages at each stage of engagement, allowing the cascade to be pictured as contacts progressed (see Figure 4). The dashboard provided a high-level

XSL•FO
RenderX

overview of contact progress from enrollment to treatment, displaying raw counts that facilitated daily monitoring and real-time cascade tracking. However, it required further data cleaning for reliable evaluation. The metrics in Figure 4 show a one-time illustrative snapshot of what the research team would use daily to view performance.

Figure 4. Illustrative tableau dashboard used for monitoring real-time Shesha metrics.

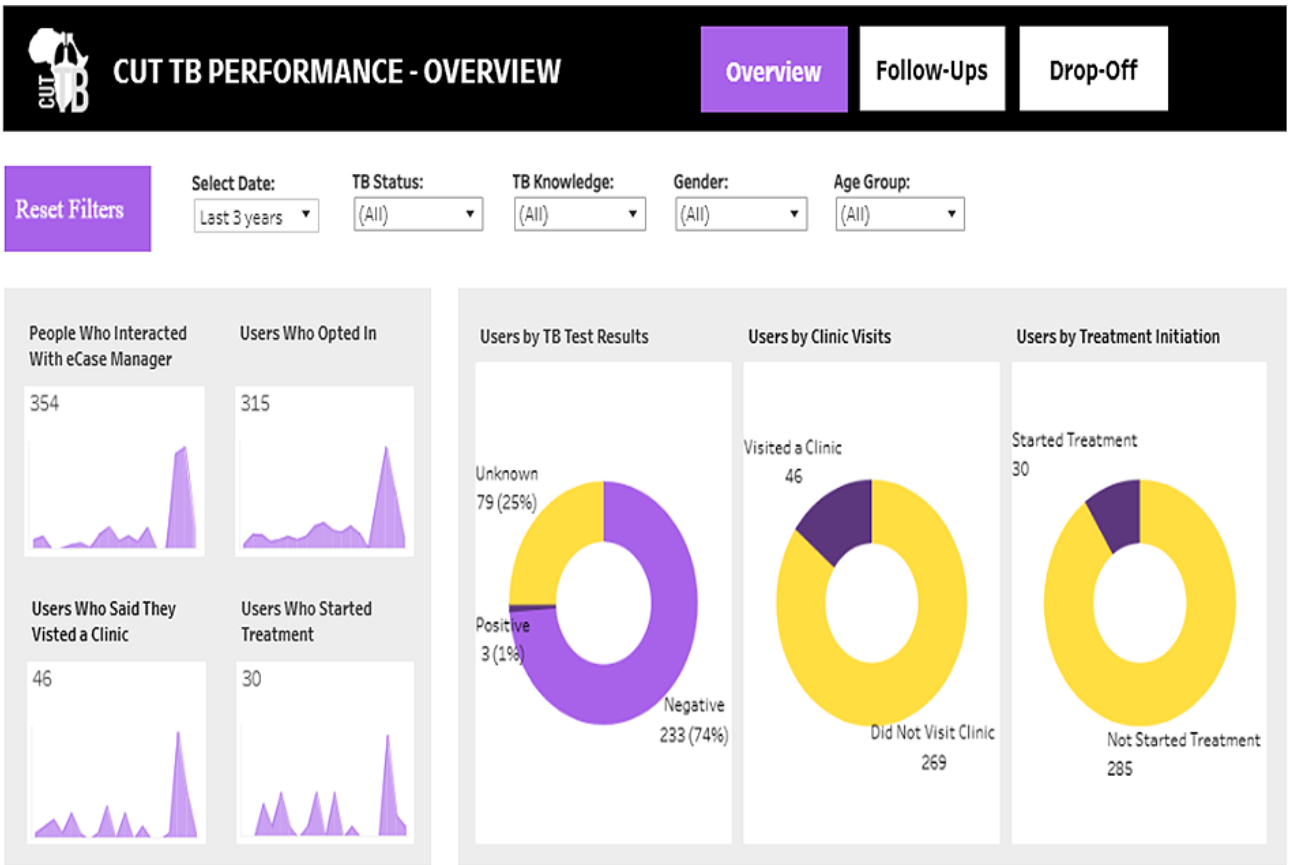

Summary of Output From the Engagements With Internal and External Stakeholders

A summary of the recommendations from the stakeholder consultation meeting is presented in Table 3. These recommendations addressed existing challenges in the contact tracing process and how we could design the Shesha tool to resolve them. We incorporated most suggestions into the Shesha app, excluding extensive counseling and GPS data collection. Counseling would have required a separate module and additional costs, while GPS was considered unnecessary for information exchange and could have deterred user engagement due to privacy concerns. Instead, we used passive geolocation to allow users to select nearby clinics, assuming most would choose local facilities.

XSL•FO
RenderX

Table . Summary of suggestions from the internal stakeholder consultation meeting.

Problem identified in the tracing cascade	Recommendations from stakeholders
Notification of results	
Delayed TB[a] results delay treatment initiation.	Have faster notification to supplement NHLS[b] SMS
Non-delivery of results or not read if SMS is used	Use WhatsApp to improve the chances of reading messages
Patient follow-up	
No reliable mechanism to show non-responses	Trigger phone follow-ups if no response within 14 days
Limited follow-up may delay initiation.	Prioritize immediate follow-up for high-risk cases
Treatment initiation	
High opt-out rates from the digital system	Collect, analyze, and opt-out the reasons to address the gaps
Patients disengaging before reaching care	
Treatment adherence	
Risk of lower engagement after starting TPT[c]	Send regular adherence messages monthly until completion
Current lack of supportive information	Use a theory-informed behavior change model
Patient support	
Emotional and psychosocial needs not addressed	Integrate counseling sessions into the adherence
Data collection and integrity	
Patient identifiers such as age and gender	Collect basic demographics
Self-reported data may be unreliable	Validate self-report with clinic verifiable data
Limited ability to analyze service gaps geographically	Collect GPS coordinates to enable geospatial analysis
Cost management	
High SMS costs for providers deter sustainability.	Use WhatsApp as a primary platform due to its lower cost and greater functionality

[a]TB: tuberculosis.

[b]NHLS: National Health Laboratory Service.

[c]TPT: tuberculosis preventive treatment.

Early User-Satisfaction Insights

A user acceptance survey was distributed on Shesha to each user after their initial 30 days of participation for participants enrolled until March 30, 2024. By May 2024, when the survey data were analyzed, only 51 users had been onboarded onto Shesha. A total of 46 surveys were sent to these users because the remaining 5 were internal testers. Of these 46, 12 (26%) participants completed the survey, 30 (65%) users started but did not finish, and 4 (9%) never responded at all. Many of those who stopped after starting only answered up to the third question. Nonetheless, we gained some insights from the users who submitted data. When asked about the perceived safety and privacy of their information, users mainly responded affirmatively. One user commented on the accuracy of the delivery timing or the results.

> *"I got the results on the exact time frame I was given, and it was a one-on-one session, was able to ask anything I wanted to ask"*

However, 1 reported forgetting their onboarding password and could not get results on time.

> *"I never received any results. I forgot my password"*

Users shared that Shesha helped them attend the clinic. Textbox 1 presents respondents' comments. Overall, participants felt that Shesha's information improved their understanding of TPT and TB; however, some still believed they did not need to visit the clinic because of negative results.

Textbox 1. User perceptions about Shesha benefits.

"I understand more about TPT"

"Everything was clearly explained to me and I understood everything with regards to TB"

"I have never went for clinic visits as my results were negative"

"It made me feel less anxious in visit the clinic"

XSL•FO
RenderX

Users held mixed views about how Shesha was motivating them to start TPT after receiving their results. One individual expressed no motivation to begin treatment despite receiving encouraging messages.

> *"I have never taken any TPT medicine as I have never visited a clinic after getting my test results, as they came back negative."*

However, another user expressed a change in perception regarding starting treatment after receiving messages from the chatbot.

> *"At first I was scared to take the treatment, but after the charts I understood the importance of taking the medication."*

Regarding the remaining quantitative metrics from the satisfaction survey. A total of 17 responded about the clarity of messaging, and 88% (15/17) reported that the messaging was clear, 94% (16/17) said the menu information was easy to find. Only 15 participants answered the Shesha user friendliness, and 74% (11/15) found the service user-friendly. Overall, for the 15 that continued, their perceptions of Shesha were favorable, with 69% (11/15) rating it as "Good" and 25% (4/11) as "Ok," and only 1 of the 15 (6%) said it was bad. Importantly, all 15 (100%) indicated they would recommend it to others.

Changes and Improvements to the Service

Resolving Bugs

The user satisfaction survey did not prompt any significant changes to the service. However, some modifications were noted, and bugs encountered were fixed. Textbox 2 provides a summary of the changes that were made to the service as well as the bugs fixed during the process.

Textbox 2. Summary of issues identified and resolved during quality assurance and early implementation.

Description of the issue and bugs

Workflow breakdowns

i. Day 10 message sent twice in a multi-user TB-positive flow; Day 14 failed to trigger, and the system froze.

ii. Days 14, 20, and 28 did not trigger automatically; manual start led to incorrect flow assignment.

iii. Users who intended to visit the clinic did not receive the required 7-day follow-up message.

iv. Day 7 was skipped entirely in the user journey; the system jumped to Day 10.

v. Flow became unresponsive after fast-forwarding a user to Day 3 post-onboarding.

vi. "Add more people" and "Done" buttons did not appear after registering the first user.

vii. After adding 6 users, the system sent inbox messages in the wrong sequence (group then random individual).

viii. "Stop messages" command looped back to the previous step instead of terminating communication.

Missing or incomplete prompts

i. Day 14 question on medication collection was omitted, affecting follow-up.

ii. Completion message after onboarding lacked a prompt linking users to TB education via the Menu.

iii. Final message on Day 3 lacked instructions for accessing results in the secure inbox.

iv. Responding "No" in the Day 20 symptom check did not trigger the expected debug message and email alert.

v. Automated email alerts omitted user identifiers (eg, phone number, participant ID) or contained incorrect usernames.

vi. Selecting "Other" as a clinic option failed to trigger a free-text input prompt; flow skipped ahead.

Content and messaging inconsistencies

viii. Day 20 (side effects) and Day 28 (adherence) messages were sent at the same time, which was confusing.

ix. The WhatsApp message differed from the approved Figma design, raising concerns about consistency.

x. Steps 3 and 6 in onboarding gave inconsistent instructions from the mock-up.

Summary of Human-Centered Design Workshop Changes and Recommendations

A workshop was held 1 year after the launch of Shesha where delegates suggested ways to improve the chatbot and also recommended alternative prototype models. A total of 2 key questions arose: (1) how to design interactions for engagement beyond day 7, and (2) how to increase success in linkage goals. We annotated responses to each recommendation; most were already included or outside scope due to budget (see Figure 5). A significant change was renaming the chatbot from being called eCase Manager to its new name Shesha. The new name was from IsiZulu, the most spoken local language in South Africa, which means "hurry." The name aligned better with the its purpose to link people faster and also removed the mislabeling of people as "cases." Changing the name to appear as Shesha on WhatsApp was, however, complex due to Meta Business Suite requiring updates like changing the display name on the website of the research team and a subsequent Meta review. The project came to an end before finalizing the official name change; however, we plan to rename again if we secure more funding to continue implementation. In the field, however, the

XSL•FO
RenderX

staff had already started referring to the service as Shesha and introducing it as such to the participants.

Figure 5. Recommendations from the codesign workshop.

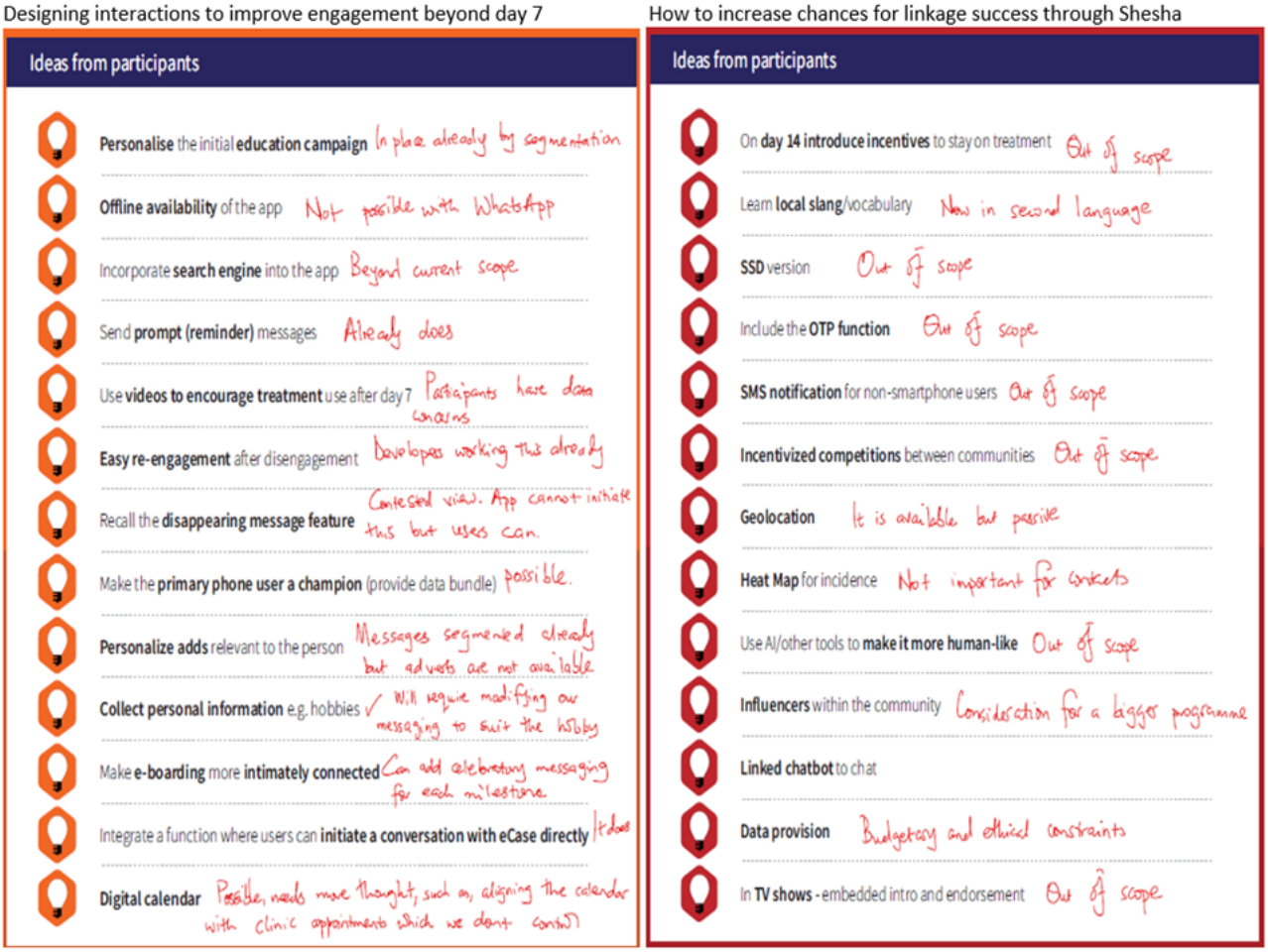

Multiple User Journeys

Multiple user journeys were chosen to avoid excluding household members without phones; however, these were difficult to implement and required workarounds. Having multiple users on 1 number complicated workflows because the system needed to identify the user in the session to link responses correctly. For example, if a household had multiple users and messages were sent in the morning (staggered by 1 h from 9 AM), the reply only accepted the most recent message. As shown in Figure 6, if messages are sent to “Meghan,” “Lucy,” and “Ori” and “Meghan” replies, Shesha would recognize the latest message from “Ori.” In multiple-user households, messages were designed to be sent an hour apart, allowing the recipient to respond. In addition, to ensure proper workflow, each user received a participant ID at onboarding and chose a name or alias to use for identifying themselves in their responses.

XSL•FO
RenderX

Figure 6. Example of a multiuser journey.

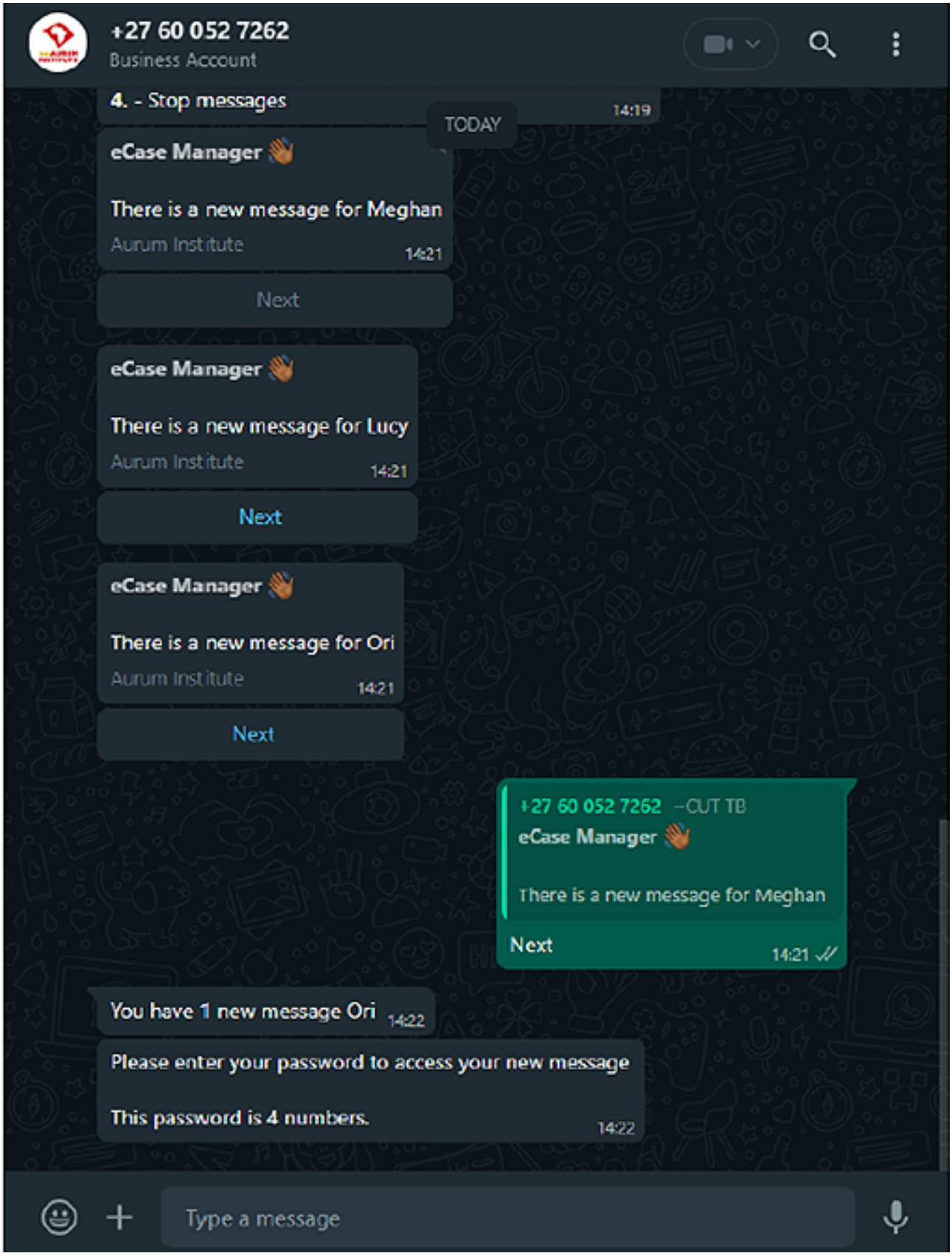

Evaluation and the Cost of Developing Shesha

The indicators for the planned process evaluation using the RE-AIM framework are shown in Table 4. These indicators cover the implementation science outcomes with the aim of understanding the real-world factors that would influence success or failure of the app. The evaluation results are, however beyond the current scope, which is only focusing on documenting the development process.

XSL•FO
RenderX

Table . RE-AIM evaluation indicators.

RE-AIM[a] domain	Definition of indicators to be measured
Reach	• Proportion of household contacts who were found, screened for eligibility, and successfully registered on Shesha
Effectiveness	• Proportion of Shesha users who attended clinic, started TB[b] or TPT[c], completed treatment
Adoption	• Proportion of Shesha users who acted on (opened, read, or responded) to milestone messages (Days 3, 5, 7, 14)
Implementation	• *Internal fidelity*: whether Shesha executed design logic correctly (queuing contact, delivering messages, sending reminders) • *External fidelity*: whether staff followed onboarding steps • *Feasibility*: whether users had the conditions to engage (phone access, literacy)
Acceptability	• User perceptions of usefulness, ease of use, motivation, and barriers, structured by UTAUT[d] constructs
Maintenance	• Continued engagement after Day 28, including responses to monthly follow-ups

[a]RE-AIM: Reach Effectiveness Adoption Implementation Maintenance.
[b]TB: tuberculosis.
[c]TPT: tuberculosis preventive treatment
[d]UTAUT: Unified Theory of Acceptance and Use of Technology.

In addition to the process evaluation, Table 5 shows the total accounting and economic costs of developing and supporting Shesha. Using a 5-year lifespan, an 8% discount rate consistent with South Africa's public rate, and an exchange rate of R18.33 per USD, the 19-month economic cost was US $142,754. The economic cost of development was US $16,388 annually over 5 years, with yearly support costs totaling US $73,773. In total, Shesha would cost US $90,160 per year.

Table . The cost of developing Shesha.

Item	Cost category	Accounting cost (USD)[a]	Equivalent monthly cost (USD)	Equivalent annual cost (USD)	19-Month cost (USD)
Development costs	Capital	56,593	1166	13,988	22,147
Translations	Capital	9709	200	2400	3800
Subtotal capital	Capital	66,302	1366	16,388	25,947
Annual support	Recurrent	116,807	6148	73,773	116,807
Total economic cost	All	183,109	7513	90,160	142,754

[a]US dollars.

Discussion

Principal Findings

This paper describes the development of Shesha, a WhatsApp-based chatbot designed to support household contacts of people being treated for TB in their linkage to care in South Africa. We applied a design thinking process, using the HBM to structure the user journey. Published accounts of mHealth tool development are scarce in South Africa, and to our knowledge, this is the first peer-reviewed paper documenting such a process on TB contact tracing in South Africa. Only 1 other local study, by Mathenjwa et al [39], reports on the development of a TB and HIV tool, but the tool is more broadly used, unlike Shesha that focuses on contact tracing. Our work, therefore, addresses a significant gap in public health–oriented design literature in this setting.

Using design thinking helped translate technically demanding workflows into a clear structure, making development feel more manageable despite the complexity of the tool. Several studies have shown that design thinking enhances the development of digital health tools [26,40]. Except for the delayed user satisfaction survey, in Shesha, we largely followed the framework's steps. In retrospect, a more flexible application of the framework, as suggested by others [41], might have enabled more structured stakeholder engagement, including workshops and regular feedback loops. Despite this, the design thinking approach added significant value and should remain central to future tool development efforts.

XSL•FO
RenderX

Collaborating with skilled developers was essential for Shesha's progress. We partnered with a team experienced in public health to ensure technical decisions aligned with program goals and real-world implementation. Usually, public health teams identify problems, and developers build the architecture; imbalances between and asymmetries in interest can hinder progress. To prevent this, we formalized the initial partnership with early agreements on intellectual property, code ownership, change procedures, and post-launch responsibilities, which fostered a productive working relationship. Therefore, partnering with developers should be viewed not just as a client-vendor relationship, but as a genuine collaboration focused on shared benefits.

Dashboards are essential for managing chatbot interventions, which generate large volumes of data across different flows and time points. In Shesha, the dashboard was built and managed externally by the developers. It provided real-time visibility and supported day-to-day monitoring, but this arrangement limited our ability to make changes quickly when needed. Although the dashboard included basic validation checks, it used raw data and could not produce outcome-level indicators without additional data cleaning. A similar approach was used in the TBCheck project in South Africa, where dashboards supported implementation, but final outcomes were confirmed through laboratory records Rich [42]. More work and skill capacitation are needed to ensure that public health teams can exert greater control over the design, management, and use of dashboards. We agree with Rossi et al [43] in their argument that dashboard development should follow a dedicated design thinking process, which can strengthen their role as core tools for monitoring and guiding digital health interventions.

WhatsApp was selected because it offered the best balance between functionality, engagement, and cost. SMS and USSD were considered, but resource constraints limited the use of multiple platforms. More importantly, these platforms were not well suited to the kind of 2-way, conversational engagement that Shesha required. WhatsApp allowed us to deliver motivational messages in a familiar, chat-based format that aligned with how people already communicate and was more likely to sustain interest. At the same time, we were aware of the access challenges this choice might present. Although over 75% of South Africans use smartphones to access the internet [23], TB-affected households are disproportionately affected by poverty and may struggle to afford data or maintain devices [44-46]. A South African study recommended SMS over WhatsApp for this reason [47]. However, the study targeted recently circumcised men, a group with high perceived risk and strong motivation to engage. Our context was different. Shesha targeted generally healthy household contacts, being asked to start preventive treatment for a disease they did not see as an immediate threat. According to the HBM, perceived risk and susceptibility influence engagement, and in our case, ease of use and familiarity with the platform were probably more significant than broad accessibility alone. Platform choice, therefore, should not be based only on cost or reach. Even inexpensive options can fail if they impose all the burden on implementers who cannot or will not bear the costs. The decision must consider behavioral, social, financial, and technical factors specific to the context.

Regarding ethics and information privacy, especially for shared devices, privacy can never be fully guaranteed because messages are not invisible to all users on the device. Our methods for providing PINs to access sensitive information become obsolete once the message is received and left undeleted. Ethical issues in WhatsApp use remain contentious despite safeguards like encryption. One review of health studies highlighted gaps in privacy, emphasizing the need for minimal data collection, anonymization, and transparent risk communication [48]. In our study, we implemented these safeguards: using aliases, providing detailed consent, and adding extra privacy policies. However, we recognized that ultimate control over security depends on the platform provider, and protections are only as adequate as those put in place by WhatsApp itself. These issues extend beyond individual studies like ours and need to be addressed at a governance and regulation level.

Finally, while Shesha was built using preprogrammed logic trees, advances in artificial intelligence (AI) now offer new opportunities for dynamic, adaptive conversation flows. At the time of development, AI integration was not widely available. However, the recent rapid evolution of AI in chatbot platforms means that tools like Shesha could, in the future, work more intelligently to respond to user inputs.

Limitations and Lessons Learned

The development of Shesha was not without its challenges. A commonly expected step in digital health development is to conduct pre-intervention surveys or user consultations to guide design, which we could not do in our study, as described in the *Methods* section. This absence may have delayed or limited the discovery of usability issues that only surfaced during actual use. For instance, although prior studies were helpful, they may not have fully captured the profile of this novel population, resulting in gaps in areas such as the timing of messages for multiple-user households or participants' preferences for privacy features beyond PINs, which we used in Shesha. Some challenges may also not have been realistically identifiable before use, as users had no prior exposure to TPT or patient-facing digital tools. This makes the absence of a predesign survey both a limitation of our approach and a reflection of the development context. However, it was not a dealbreaker; instead, it offers a lesson for future developers. Predesign consultations remain a standard part of digital health projects and should be undertaken wherever possible. In settings like ours, where valuable input may not be available until users experience the tool directly, implementers must anticipate the risks of postponing engagement and plan for structured feedback after launch, along with the ability to implement timely updates.

Budget constraints affected both the pace and flexibility of development. Developer time was tightly allocated to the initial build, making it difficult to accommodate changes once development was underway without securing additional funds. This limitation contributed to the initial launch of the tool in English only, which may have reduced accessibility for non-English-speaking users. Although additional funding later allowed for translation into isiZulu, the delay highlights how

XSL•FO
RenderX

resource constraints can affect inclusivity. Our costing exercise also showed that development and support costs can be high relative to the scale of early implementation, which may discourage smaller organizations from pursuing similar tools. As funding streams become more limited, future mHealth efforts will need to present strong evaluation data and transparent development documentation to build a credible investment case and secure sustainable support.

Conclusion

The development of Shesha offers valuable insights for future mHealth design and implementation in resource-limited settings. Structured design methods like design thinking remain useful and should inform development efforts. However, flexibility is crucial, as real-world circumstances may require deviation from standard sequences or planned processes. Equity considerations should be integrated early in technical decisions, even if they add complexity to implementation. This involves ensuring access for various user groups and accommodating different device usage patterns. Effective collaboration between public health teams and developers is essential to aligning technical solutions with program aims. Partnerships based on shared planning and mutual understanding can minimize friction and support successful delivery. Platform selection, budgeting, and ethical safeguards need to be considered within the specific context. Tools should aim to balance reach, functionality, sustainability, and privacy, especially when engaging populations with limited resources or digital access. Lastly, documenting and sharing the development process, including challenges and trade-offs, is just as important as reporting outcomes. Transparent documentation strengthens the evidence base for mHealth, encourages learning across projects, and helps build a stronger case for sustained investment in digital health innovation.

Acknowledgments

We would like to acknowledge the Community and Universal Testing for Tuberculosis among Contacts team for their valuable contributions in ideating and providing inputs throughout the development process of the Shesha chatbot. Their insights and feedback were instrumental in shaping this tool to meet the needs of the program and the communities it serves. The Community and Universal Testing for Tuberculosis among Contacts trial is funded by the European and Developing Countries Clinical Trials Partnership under grant RIA2019IR-2895.

We would like to extend our acknowledgments to the Matchboxology team who guided the user-centered design workshops for the further refinement of the Shesha app.

Authors' Contributions

Conceptualization: DLM (lead), FN, KV, MR, SC, TN, YH
Methodology: DLM (lead), FN, KV, MR, SC, TN, YH
Project administration: DLM (lead)
Software: BK, CW, MK
Resources: TM (Human Case Manager role)
Visualization: DLM (lead), FN, KV, MR, SC, TN, YH
Writing original draft: DLM (lead)
Writing of review & editing: DLM (lead), KV, MR, SC, YH (equal)
Supervision: KV, SC

Conflicts of Interest

None declared.

References

1. Brinkel J, Krämer A, Krumkamp R, May J, Fobil J. Mobile phone-based mHealth approaches for public health surveillance in sub-Saharan Africa: a systematic review. Int J Environ Res Public Health 2014 Nov 12;11(11):11559-11582. [doi: 10.3390/ijerph111111559] [Medline: 25396767]
2. Van der Pol N, Ntinga X, Mkhize M, van Heerden A. A scoping review of mHealth use in South Africa. Dev South Afr 2022 Jul 4;39(4):485-497. [doi: 10.1080/0376835X.2021.1904837]
3. Research and Training in Health and Development (RTHD). Review of mHealth solutions for TB in South Africa. 2019 Feb 15 URL: https://tbthinktank.org/wp-content/uploads/2022/08/2019-Policy-Brief-mHealth-for-TB-in-South-Africa.pdf [accessed 2025-09-29]
4. Mudzengi DL, Chomutare H, Nagudi J, et al. Using mHealth technologies for case finding in tuberculosis and other infectious diseases in Africa: systematic review. JMIR Mhealth Uhealth 2024;12:e53211. [doi: 10.2196/53211]
5. Perrin Franck C, Geissbuhler A, Lovis C. “To err is evolution”: we need the implementation report to learn. JMIR Med Inform 2023 Apr 4;11:e47695. [doi: 10.2196/47695] [Medline: 37014675]
6. Iribarren SJ, Beck SL, Pearce PF, Chirico C, Etchevarria M, Rubinstein F. mHealth intervention development to support patients with active tuberculosis. J Mob Technol Med 2014;3(2):16-27. [doi: 10.7309/jmtm.3.2.4] [Medline: 26246859]

XSL•FO
RenderX

7. Katende KK, Amiyo MR, Nabukeera S, et al. Design, development, and testing of a voice-text mobile health application to support tuberculosis medication adherence in Uganda. PLoS ONE 2022;17(9):e0274112. [doi: 10.1371/journal.pone.0274112] [Medline: 36084106]
8. Global tuberculosis report 2024. World Health Organisation. 2024. URL: https://www.who.int/teams/global-programme-on-tuberculosis-and-lung-health/tb-reports/global-tuberculosis-report-2024 [accessed 2025-10-23]
9. Fox GJ, Barry SE, Britton WJ, Marks GB. Contact investigation for tuberculosis: a systematic review and meta-analysis. Eur Respir J 2013 Jan;41(1):140-156. [doi: 10.1183/09031936.00070812] [Medline: 22936710]
10. Müller J, Kretzschmar M. Contact tracing—old models and new challenges. Infect Dis Model 2021;6:222-231. [doi: 10.1016/j.idm.2020.12.005] [Medline: 33506153]
11. Baluku JB, Nabwana M, Winters M, Bongomin F. Tuberculosis contact tracing yield and associated factors in Uganda. BMC Pulm Med 2022 Feb 16;22(1):64. [doi: 10.1186/s12890-022-01860-z] [Medline: 35172788]
12. Guy D, Kodjamanova P, Woldmann L, et al. Contact tracing strategies for infectious diseases: a systematic literature review. PLOS Glob Public Health 2025;5(5):e0004579. [doi: 10.1371/journal.pgph.0004579] [Medline: 40343962]
13. Velen K, Lewis JJ, Charalambous S, et al. Household HIV testing uptake among contacts of TB patients in South Africa. PLoS ONE 2016;11(5):e0155688. [doi: 10.1371/journal.pone.0155688] [Medline: 27195957]
14. Armstrong-Hough M, Turimumahoro P, Meyer AJ, et al. Drop-out from the tuberculosis contact investigation cascade in a routine public health setting in urban Uganda: a prospective, multi-center study. PLoS ONE 2017;12(11):e0187145. [doi: 10.1371/journal.pone.0187145] [Medline: 29108007]
15. National guidelines on the treatment of tuberculosis infection. National Department of Health South Africa. 2023. URL: https://knowledgehub.health.gov.za/elibrary/national-guidelines-treatment-tuberculosis-infection [accessed 2025-10-23]
16. South africa national TB strategic plan: 2023-2028. south african national TB programme. National Department of Health South Africa. 2024. URL: https://knowledgehub.health.gov.za/elibrary/national-strategic-plan-hiv-tb-and-stis-2023-2028 [accessed 2025-10-23]
17. Moyo M, Ntshiqa T, Hamada Y, et al. Community and universal testing for TB among close contacts of microbiologically confirmed pulmonary TB patients in two high TB burden countries: a protocol for a pragmatic cluster-randomised control trial. Trials 2025 Aug 11;26(1):285. [doi: 10.1186/s13063-025-08978-5] [Medline: 40790601]
18. Ekurhuleni municipality integrated development plan—2022/23–2026/27. Ekurhuleni Municipality. 2022. URL: https://www.ekurhuleni.gov.za/document/annexure-a-2022_23-2026_27-revised-integrated-development-plan-idp/ [accessed 2025-10-23]
19. Mukora R, Thompson RR, Hippner P, et al. Human resource time commitments and associated costs of Community Caregiver outreach team operations in South Africa. PLoS ONE 2023;18(3):e0282425. [doi: 10.1371/journal.pone.0282425] [Medline: 36877676]
20. Annual report 2021/22. Ekurhuleni Municipality. 2022. URL: https://www.ekurhuleni.gov.za/about-the-city/annual-reports/ [accessed 2025-10-23]
21. Curable TB claims 1.5 million lives per year. Gauteng Department of Health. 2024. URL: https://www.citizen.co.za/sedibeng-ster/news/2024/03/26/curable-tb-claims-1-5-million-lives-per-year/ [accessed 2025-10-23]
22. Kufa-Chakezha T, Shangase N, S B, Cutler E, Aitken S, Cheyip M, et al. The 2022 antenatal HIV sentinel survey: key findings. Public Health Bulletin South Africa 2023;2013(1):1-12 [FREE Full text]
23. Miyajima K. Mobile phone ownership and welfare: evidence from South Africa's household survey. World Dev 2022 Jun;154:105863. [doi: 10.1016/j.worlddev.2022.105863]
24. Moyo S, Ismail F, Mkhondo N, et al. Healthcare seeking patterns for TB symptoms: findings from the first national TB prevalence survey of South Africa, 2017-2019. PLoS ONE 2023;18(3):e0282125. [doi: 10.1371/journal.pone.0282125] [Medline: 36920991]
25. Dam RF. Design thinking: Interaction Design Foundation. 2024. URL: https://www.interaction-design.org/literature/article/5-stages-in-the-design-thinking-process [accessed 2022-08-05]
26. Fähnrich C, Denecke K, Adeoye OO, et al. Surveillance and outbreak response management system (SORMAS) to support the control of the Ebola virus disease outbreak in West Africa. Euro Surveill 2015 Mar 26;20(12):21071. [doi: 10.2807/1560-7917.es2015.20.12.21071] [Medline: 25846493]
27. Grainger C. A software for disease surveillance and outbreak response-insights from implementing SORMAS in nigeria and ghana. Federal Ministry for Economic Cooperation and Development (BMZ). 2020. URL: https://health.bmz.de/studies/a-software-for-disease-surveillance-and-outbreak-response/ [accessed 2025-10-23]
28. MacPherson M, Merry K, Locke S, Jung M. Developing mobile health interventions with implementation in mind: application of the multiphase optimization strategy (MOST) preparation phase to diabetes prevention programming. JMIR Form Res 2022 Apr 26;6(4):e36143. [doi: 10.2196/36143] [Medline: 35471473]
29. Thind D, Charalambous S, Tongman A, Churchyard G, Grant AD. An evaluation of "Ribolola": a household tuberculosis contact tracing programme in North West Province, South Africa. Int J Tuberc Lung Dis 2012 Dec;16(12):1643-1648. [doi: 10.5588/ijtld.12.0074] [Medline: 23131263]
30. Rosenstock IM, Strecher VJ, Becker MH. Social learning theory and the Health Belief Model. Health Educ Q 1988;15(2):175-183. [doi: 10.1177/109019818801500203] [Medline: 3378902]

XSL•FO
RenderX

31. Praekelt.org. TB healthcheck puts tuberculosis self-screening in everyone's hands ahead of World TB Day. URL: https://www.praekelt.org/news [accessed 2023-05-08]
32. Barron P, Peter J, LeFevre AE, et al. Mobile health messaging service and helpdesk for South African mothers (MomConnect): history, successes and challenges. BMJ Glob Health 2018;3(Suppl 2):e000559. [doi: 10.1136/bmjgh-2017-000559] [Medline: 29713503]
33. Cowling N. Number of internet users in South Africa from 2013 to 2024: Statista. URL: https://www.statista.com/statistics/462958/internet-users-south-africa [accessed 2024-02-27]
34. Cowling N. Leading social media platforms in South Africa 2023: Statista. URL: https://www.statista.com/statistics/1189958/penetration-rate-of-social-media-in-south-africa/#:~:text=WhatsApp%20is%20the%20most%20popular,percent%20and%2074%20percent%2C%20respectively [accessed 2024-03-30]
35. DeCormier Plosky W, Bollinger LA, Alexander L, et al. Developing the global health cost consortium unit cost study repository for HIV and TB: methodology and lessons learned. Afr J AIDS Res 2019 Dec;18(4):263-276. [doi: 10.2989/16085906.2019.1680398] [Medline: 31779571]
36. Unsworth M, Fabens I, Setswe G, et al. Expanding two-way texting for post-operative follow-up: a cost analysis of the implementation and scale-up in routine voluntary medical male circumcision settings in South Africa. PLOS Glob Public Health 2025;5(4):e0004049. [doi: 10.1371/journal.pgph.0004049] [Medline: 40261872]
37. Chen Y, Ronen K, Matemo D, et al. An interactive text messaging intervention to improve adherence to option B+ prevention of mother-to-child HIV transmission in Kenya: cost analysis. JMIR Mhealth Uhealth 2020 Oct 2;8(10):e18351. [doi: 10.2196/18351] [Medline: 33006562]
38. Khan ZA, Kidholm K, Pedersen SA, et al. Developing a program costs checklist of digital health interventions: a scoping review and empirical case study. Pharmacoeconomics 2024 Jun;42(6):663-678. [doi: 10.1007/s40273-024-01366-y] [Medline: 38530596]
39. Mathenjwa T, Adeagbo O, Zuma T, et al. Development and acceptability of a tablet-based app to support men to link to HIV care: mixed methods approach. JMIR Mhealth Uhealth 2020 Nov 24;8(11):e17549. [doi: 10.2196/17549] [Medline: 33231558]
40. Johansson‐Sköldberg U, Woodilla J, Çetinkaya M. Design thinking: past, present and possible futures. Creat Innov Manage 2013 Jun;22(2):121-146. [doi: 10.1111/caim.12023]
41. Altman M, Huang TTK, Breland JY. Design thinking in health care. Prev Chronic Dis 2018 Sep 27;15:E117. [doi: 10.5888/pcd15.180128] [Medline: 30264690]
42. Rich K, Burger R, Goldberg D, Moultrie H, Rieger M. Is it possible to encourage TB testing and detect missing TB cases via community-level promotion of a self-screening mobile application? Quasi-experimental evidence from South Africa. BMJ Health Care Inform 2025 May;32(1):e101179. [doi: 10.1136/bmjhci-2024-101179]
43. Rossi FS, Adams MCB, Aarons G, McGovern MP. From glitter to gold: recommendations for effective dashboards from design through sustainment. Implement Sci 2025 Apr 22;20(1):16. [doi: 10.1186/s13012-025-01430-x] [Medline: 40264174]
44. Foster N, Vassall A, Cleary S, Cunnama L, Churchyard G, Sinanovic E. The economic burden of TB diagnosis and treatment in South Africa. Soc Sci Med 2015 Apr;130:42-50. [doi: 10.1016/j.socscimed.2015.01.046]
45. Erlinger S, Stracker N, Hanrahan C, et al. Tuberculosis patients with higher levels of poverty face equal or greater costs of illness. Int J Tuberc Lung Dis 2019 Nov 1;23(11):1205-1212. [doi: 10.5588/ijtld.18.0814] [Medline: 31718757]
46. National TB patient cost survey findings. Knowledge Hub. URL: https://knowledgehub.health.gov.za/webinar/national-tb-patient-cost-survey-findings [accessed 2025-10-23]
47. Fabens I, Makhele C, Igaba NK, et al. WhatsApp versus SMS for 2-way, text-based follow-up after voluntary medical male circumcision in South Africa: exploration of messaging platform choice. JMIR Form Res 2024;8:e62762. [doi: 10.2196/62762]
48. Manji K, Hanefeld J, Vearey J, Walls H, de Gruchy T. Using WhatsApp messenger for health systems research: a scoping review of available literature. Health Policy Plan 2021 Jun 1;36(5):774-789. [doi: 10.1093/heapol/czab024]

Abbreviations

CHW: community health worker
CUT-TB: Community and Universal Testing for Tuberculosis among Contacts
HBM: Health Belief Model
mHealth: mobile health
RE-AIM: Reach Effectiveness Adoption Implementation Maintenance
TB: tuberculosis
TPT: tuberculosis preventive treatment
USSD: Unstructured Supplementary Service Data

XSL•FO
RenderX

Edited by J Sarvestan; submitted 06.02.25; peer-reviewed by L Xu; revised version received 07.09.25; accepted 09.09.25; published 13.11.25.

Please cite as:
Mudzengi DL, Ntshiqa T, Hamada Y, Ndebele F, Mpanza T, Kyobutungi B, Williams C, Kennealy M, Rangaka M, Velen K, Charalambous S
Shesha, a WhatsApp Chatbot for Linking Household Contacts to Tuberculosis Treatment or Preventive Therapy in South Africa: Design and Development
JMIR Form Res 2025;9:e71793
URL: https://formative.jmir.org/2025/1/e71793
doi:10.2196/71793

© Don Lawrence Mudzengi, Thobani Ntshiqa, Yohhei Hamada, Felex Ndebele, Thapelo Mpanza, Bridget Kyobutungi, Candice Williams, Meghan Kennealy, Molebogeng Rangaka, Kavindhran Velen, Salome Charalambous. Originally published in JMIR Formative Research (https://formative.jmir.org), 13.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Awareness and Attitudes of University Students in Bangladesh Toward Cancer: Cross-Sectional Study

Maliha Tabassum[1*], BSc; Nafisa Farhin[1*], BSc; Afsana Afrose[1*], BSc, MSc; Mahfuza Moin Surovy[1*], BSc; Rubaiya Tasnim[1*], BSc; Most. Humayra Affia Heaven[1*], BSc; Jannatun Noor[2,3*], BSc, MSc, PhD; Munima Haque[1*], BSc, MSc, PhD

[1]Biotechnology Program, Department of Mathematics and Natural Sciences, BRAC University, Kha 224, Bir Uttam Rafiqul Islam Ave, Dhaka, Bangladesh

[2]Department of Computer Science and Engineering, BRAC University, Dhaka, Bangladesh

[3]Department of Computer Science and Engineering, United International University, Dhaka, Bangladesh

[*]all authors contributed equally

Corresponding Author:
Munima Haque, BSc, MSc, PhD
Biotechnology Program, Department of Mathematics and Natural Sciences, BRAC University, Kha 224, Bir Uttam Rafiqul Islam Ave, Dhaka, Bangladesh

Abstract

Background: Early detection and awareness are critical in reducing the burden of cancer. However, a significant proportion of university students in Bangladesh remains inadequately informed about cancer risks and preventive measures.

Objective: This study aimed to assess knowledge gaps and evaluate the attitudes of Bangladeshi university students toward cancer, its prevention, risk factors, and care for affected individuals.

Methods: A descriptive, cross-sectional survey was conducted among 530 university students aged 20 to 35 years across Bangladesh. Data were collected using an ethically approved, structured internet-based questionnaire between December 2022 and March 2024. The questionnaire assessed sociodemographics, cancer knowledge, awareness of risk factors, personal or familial cancer experiences, and attitudes toward cancer care and policy. Descriptive statistics and chi-square tests were used to analyze categorical data, with a significance threshold of $P<.05$.

Results: Most participants were aged 21–25 years (406/530, 76.6%) and female (320/530, 60.4%), with the majority enrolled in undergraduate programs (82.8%, 439/530). While 60.8% (322/530) considered themselves somewhat knowledgeable about cancer, only 11.9% (63/530) were very knowledgeable, and 93.6% (496/530) had never undergone any cancer screening. Despite this, 74.3% (394/530) had personal or familial exposure to cancer, with carcinoma reported by 52.8% (280/530) of those affected. Awareness of established risk factors was inconsistent—smoking (90.9%, 482/530) and radiation (86.6%, 459/530) were widely recognized, but only 38.9% (206/530) acknowledged aging, 35.3% (187/530) obesity, and 29.2% (155/530) infectious agents as risk factors. Reproductive factors were least recognized, with just 10.2% (54/530) identifying having more children as a risk factor. Gender differences were significant in cancer-related attitudes. For example, 51.5% (273/530) of female participants versus 33.4% (177/530) of male participants felt comfortable around patients with cancer ($P=.01$), and 57.2% (303/530) of female participants versus 35.8% (190/530) of male participants supported increased government funding for cancer care ($P=.03$). Furthermore, 55.1% (292/530) of females and 35.5% (188/530) of males stressed the need for enhanced cancer awareness programs ($P=.05$). Only 6.4% (34/530) of all participants reported undergoing any form of cancer screening, highlighting a disconnect between awareness and preventive action.

Conclusions: This study reveals critical gaps in cancer awareness among university students in Bangladesh, with pronounced disparities in knowledge of nonmodifiable risk factors and significant gender-based differences in attitudes toward cancer care. These findings highlight the urgent need for targeted, gender-sensitive educational programs and policy interventions to promote preventive practices, early detection, and equitable cancer care. Such initiatives must emphasize lesser-known risk factors, reduce stigma, and foster more inclusive, culturally competent health education strategies to mitigate the growing cancer burden in Bangladesh.

(JMIR Form Res 2025;9:e75651) doi:10.2196/75651

KEYWORDS

cancer awareness; university students; Bangladesh; cancer screening; aging

XSL•FO
RenderX

Introductions

Cancer remains one of the leading causes of morbidity and mortality worldwide, responsible for nearly 10 million deaths in 2020 alone [1]. The global burden of cancer is increasing rapidly, particularly in low- and middle-income countries, where health care infrastructure is often underdeveloped and public awareness of cancer prevention and early detection remains low [2]. Bangladesh, a densely populated low- and middle-income country, is witnessing a rising trend in cancer incidence, with an estimated 156,775 new cases and 116,598 deaths reported in 2022 [3]. These figures, however, likely underestimate the true burden due to limitations in nationwide cancer surveillance and the reliance on hospital-based cancer registries that exclude individuals without access to specialized care [4].

Common cancers in Bangladesh include lung, liver, oral cavity, and esophageal cancers in men, and breast and cervical cancers in women [5]. The underlying causes for this high prevalence are multifactorial, including environmental exposure, poor dietary habits, increasing tobacco use, and genetic predispositions, coupled with inadequate health literacy and health care access [6]. Despite ongoing efforts by governmental and nongovernmental organizations, the country continues to face significant challenges in implementing effective screening, awareness, and preventive strategies. According to World Health Organization estimates, only 5%‑10% of the population undergoes routine cancer screening, leading to late-stage diagnoses and poorer survival outcomes [7].

Awareness and attitudes toward cancer significantly shape preventive health behaviors, including risk avoidance and early detection through screening [8]. Studies globally and regionally show that young adults, especially university students, represent a critical demographic for targeted public health messaging because they are highly mobile, engaged in academic communities, and are future influencers in family and society [9,10]. However, limited studies have explored their cancer knowledge and perceptions in the Bangladeshi context.

Recent regional studies have focused primarily on breast cancer awareness among women, often neglecting broader cancer literacy and risk comprehension across genders and cancer types [11,12]. In addition, misconceptions remain widespread; many people continue to associate cancer solely with smoking or genetic predisposition while underestimating the roles of obesity, diet, reproductive factors, aging, and infectious agents in carcinogenesis [13]. Addressing these knowledge gaps is vital, as improved cancer awareness has been shown to correlate with increased participation in screening programs and healthier lifestyle choices [14].

This study aims to evaluate the awareness, knowledge, and attitudes of Bangladeshi university students toward cancer, including their understanding of risk factors, personal experiences with the disease, and attitudes toward patients and health care policies. By identifying the critical knowledge deficits and perceptual barriers in this population, this research seeks to inform culturally appropriate, gender-sensitive educational interventions to support national cancer prevention goals.

Methods

Participant Recruitment

Participants were recruited using a convenience sampling method through university mailing lists, social media platforms, and student networks. To minimize selection bias, efforts were made to include a diverse range of students from public and private universities located in different regions of the country.

The required sample size was calculated using Cochran's formula for an estimated proportion of 50% (265/530), with a 95% confidence level and a 5% (27/530) margin of error, resulting in a minimum target of 385 participants. A total of 530 valid responses were collected, yielding a response rate of approximately 81.5% (432/530), after excluding incomplete or duplicate entries.

Participants were eligible if they were university students aged between 20 and 35 years. Students below or above this age range and those enrolled in primary, secondary, or higher secondary institutions were excluded.

A structured, self-administered online questionnaire was used for data collection. The questionnaire was developed based on a review of relevant literature and similar studies conducted in comparable settings. It was pretested with 30 university students to assess clarity, content validity, and internal consistency. Feedback was used to revise ambiguous questions. The final version included sections on sociodemographic characteristics, cancer knowledge, personal or family experience with cancer, and attitudes toward cancer care and policies. Cronbach α for internal reliability of the attitude-related items was .78, indicating acceptable consistency. No clinical screening or physical assessments were performed, as the study focused solely on self-reported awareness and attitudes. This study was reported in accordance with the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) statement.

Ethical Considerations

This descriptive cross-sectional study was conducted across multiple universities in Bangladesh between December 2022 and March 2024. Ethical approval was obtained from the BRAC University Institutional Review Board (IRB No. BRACUIRB_220240006). Participation was voluntary, and informed consent was obtained before beginning the survey. Confidentiality and anonymity were strictly maintained. Data were analyzed using SPSS (version 26; IBM Corp), with categorical variables summarized as frequencies and percentages. Group differences were tested using the chi-square test, with a significance level set at $P<.05$.

Results

Table 1 presents a detailed breakdown of the sociodemographic characteristics of the study participants. The majority of participants, 76.6% (406/530), were aged between 21 and 25 years. The gender distribution was skewed toward females, who comprised 60.4% (320/530) of the participants, compared to 39.6% (210/530) males. In terms of education levels, most participants (82.8%, 439/530) were undergraduate students,

XSL•FO
RenderX

while graduate (8.5%, 45/530) and postgraduate (8.3%, 44/530) participants were nearly equal in proportion. A small fraction of participants (0.4%, 2/530) held a diploma.

Table . Sociodemographic characteristics of participants.

Characteristics	Participants (N=530), n (%)
Age (years)	
15 - 20	96 (18.1)
26 - 30	27 (5.1)
31 - 35	1 (0.2)
Sex	
Male	210 (39.6)
Female	320 (60.4)
Education	
Undergraduate	439 (82.8)
Graduate	45 (8.5)
Postgraduate	44 (8.3)
Diploma	2 (0.4)
Cancer knowledge	
Somewhat knowledgeable	322 (60.8)
Not very knowledgeable	135 (25.5)
Very knowledgeable	63 (11.9)
Not at all knowledgeable	10 (1.9)
Cancer screening	
Yes	34 (6.4)
No	496 (93.6)
You or anyone affected by cancer	
Yes	394 (74.3)
No	136 (25.7)
If yes, specify the type of cancer	
Carcinoma	280 (52.8)
Sarcoma	28 (5.3)
Leukemia	77 (14.5)
Myeloma	9 (1.7)
Lymphoma	6 (1.1)

Regarding cancer knowledge, 60.8% (322/530) of participants described themselves as “somewhat knowledgeable,” while 25.5% (135/530) reported being “not very knowledgeable.” Only 11.9% (63/530) considered themselves “very knowledgeable,” and 1.9% (10/530) indicated no knowledge of cancer. Notably, 93.6% (496/530) of participants had never undergone cancer screening. However, 74.3% (394/530) reported that they or someone they knew had been affected by cancer. Among these cases, carcinoma was the most prevalent type (52.8%, 280/530), followed by leukemia (14.5%, 77/530), sarcoma (5.3%, 28/530), myeloma (1.7%, 9/530), and lymphoma (1.1%, 6/530).

Table 2 illustrates participants’ knowledge of risk factors associated with cancer. A substantial proportion of participants identified smoking (90.9%, 482/530) and radiation exposure (86.6%, 459/530) as major risk factors. However, fewer participants recognized factors such as having more children (10.2%, 54/530) or aging (34.7%, 184/530) as risks. Family history of cancer was acknowledged by 64.2% (340/530), while only 43.2% (229/530) identified obesity as a risk factor.

XSL•FO
RenderX

Table . Impact of risk factors and symptoms knowledge of participants

Characteristics	Impact of risk factors		
	Agree, n (%)	Disagree, n (%)	Not sure, n (%)
Risk factors			
Aging	184 (34.7)	140 (26.4)	206 (38.9)
Family history	340 (64.2)	91 (17.2)	99 (18.7)
Obesity	229 (43.2)	114 (21.5)	187 (35.3)
Having more children	54 (10.2)	260 (49.1)	216 (40.8)
Radiation exposure	459 (86.6)	19 (3.6)	52 (9.8)
Infectious agents	333 (62.8)	42 (7.9)	155 (29.2)
Diet	264 (49.8)	83 (15.7)	183 (34.5)
Smoking	482 (90.9)	30 (5.7)	18 (3.4)
Artificial sweeteners	276 (52.1)	86 (12.8)	186 (35.1)

Uncertainty was notable in certain areas, with 35.1% (186/530) of participants unsure about the role of artificial sweeteners, 29.2% (155/530) uncertain about the influence of infectious agents, and 34.5% (183/530) unsure about the impact of diet. These findings highlight gaps in understanding that could be addressed through targeted educational initiatives.

Table 3 summarizes attitudes toward cancer, categorized by sex, based on various statements reflecting beliefs and perspectives on the disease. These statements address topics such as comfort levels around individuals with cancer, prioritization of patients with cancer ’ needs, government funding for cancer care and treatment, responsibility for providing care, and the necessity of accessible banking and insurance policies for patients with cancer.

The data reveal significant differences in responses between females and males for certain statements. For instance, 51.5% (273/530) of females and 33.4% (177/530) of males reported feeling comfortable around someone with cancer, with a statistically significant (P =.01). 51.3% (272/530) of females and 33.3% (177/530) of males reported people with cancer should be given top priority, with a statistically significant (P=.04). Similarly, 57.2% (303/530) of females and 35.8% (190/530) of males agreed that more government funding should be allocated to cancer care and treatment (P =.03). Both sexes expressed support for the responsibilities toward cancer patients and the need for insurance policies, with these attitudes also showing statistical significance (P<.001).

Other statements explored beliefs about cancer prevention, its fatality, and its impact on careers and personal relationships. Responses to statements such as “cancer harms personal relationships,” “cancer is preventable,” “cancer means death,” and “cancer ruins careers” did not show significant differences between males and females. However, there was a significant difference in the perception of the need for increased cancer awareness, with 55.1% (292/530) of females and 35.5% (188/530) of males agreeing on its importance (P=.05).

XSL•FO
RenderX

Table . Gender-based differences in attitudes toward cancer care and support.

Statements and sex-wise distribution	Agree n (%)	Disagree n (%)	Not sure n (%)	*P* value
Comfortable around someone with cancer				.01
Female	273 (51.5)	29 (5.5)	18 (3.4)	
Male	177 (33.4)	30 (5.7)	3 (0.6)	
The needs of people with cancer should be given top priority				.04
Female	272 (51.3)	8 (1.5)	40 (7.5)	
Male	166 (31.3)	14 (2.6)	30 (5.7)	
More government funding should be spent on the care and treatment				.03
Female	303 (57.2)	8 (1.5)	9 (1.7)	
Male	190 (35.8)	4 (0.8)	16 (3.0)	
We have a responsibility to provide the best possible care for people with cancer				<.001
Female	273 (51.5)	41 (7.7)	6 (1.1)	
Male	181 (34.2)	16 (3.0)	13 (2.5)	
We need easy bank and insurance policy for patients with cancer				.01
Female	187 (35.3)	80 (15.1)	53 (10.0)	
Male	110 (20.8)	44 (8.3)	56 (10.6)	
Cancer is preventable				.147
Female	117 (22.1)	167 (31.5)	36 (6.8)	
Male	87 (16.4)	92 (17.4)	31 (5.8)	
Cancer means death				.996
Female	90 (17.0)	188 (35.5)	42 (7.9)	
Male	59 (11.1)	124 (23.4)	27 (5.1)	
Cancer ruin careers				0.567
Female	133 (25.1)	133 (25.1)	54 (10.2)	
Male	96 (18.1)	78 (14.7)	36 (6.8)	
Cancer harm personal relations				. 351
Female	83 (15.7)	174 (32.8)	63 (11.9)	
Male	64 (12.1)	101 (19.1)	45 (8.5)	
We need more cancer awareness				.05
Female	292 (55.1)	10 (1.9)	18 (3.4)	
Male	188 (35.5)	15 (2.8)	7 (1.3)	

XSL•FO
RenderX

Discussion

Principal Findings

This study assessed the knowledge and attitudes of university students in Bangladesh toward cancer, with the aim of identifying key knowledge gaps and informing targeted educational interventions. The findings revealed significant deficiencies in awareness, particularly regarding nonmodifiable risk factors such as aging and reproductive history. Gender-based differences were also prominent, with females demonstrating more supportive attitudes toward patients with cancer and policies promoting care and awareness. Although most students had some exposure to cancer through family or acquaintances, this did not translate into adequate preventive behaviors or comprehensive understanding. These results highlight an urgent need for gender-sensitive, culturally appropriate cancer education initiatives aimed at enhancing awareness, reducing stigma, and promoting early detection among young adults in Bangladesh. Research suggests that young people are a critical target audience for future cancer awareness programs [9].

The gender distribution of respondents revealed an overrepresentation of females. This aligns with previous findings suggesting that women are generally more concerned about health issues, including preventive measures, and are more likely to participate in research studies or questionnaires [10]. Conversely, studies indicate that men are less inclined to seek medical attention or adopt health-promoting behaviors, which may contribute to gender-related disparities in cancer outcomes [15]. The participants' self-reported cancer knowledge is encouraging, with 60.8% (322/500) describing themselves as "somewhat knowledgeable" and 25.5% (135/530) considering themselves "not very knowledgeable." However, research indicates that many young adults possess limited knowledge or awareness of cancer due to a lack of direct experience, insufficient health education, and minimal exposure to information about diseases. This aligns with the findings, which suggest a knowledge gap in this demographic [8]. A concerning aspect of the study is the low rate of cancer screening, with only 6.4% (34/530) of respondents reporting having undergone screening. Despite the relatively high self-reported awareness of cancer, the vast majority (93.6%, 496/530) had not taken any preventive screening measures. This disconnect between knowledge and action is a well-documented phenomenon in healthcare, where individuals are aware of health risks but fail to engage in preventive activities due to various barriers. These obstacles may include fear, limited access to medical services, or a perceived low personal risk [14]. For many participants, their young age may contribute to a sense of invulnerability, leading them to underestimate their risk of developing cancer despite their awareness. Additionally, 74.3% (394/530) of respondents reported that they or someone close to them had been affected by cancer, underscoring the widespread impact of the disease. Among affected individuals, carcinoma was the most prevalent type, accounting for 52.8% (280/530) of cases, followed by leukemia (14.5%, 77/530), sarcoma (5.3%, 28/530), myeloma (1.7%, 9/530), and lymphoma (1.1%, 6/530). The high prevalence of carcinoma can be attributed to its origin in epithelial cells, which are more susceptible to mutations due to frequent exposure to external carcinogens such as UV light, radiation, toxins, and chemicals. Environmental factors and lifestyle choices also contribute to the rising incidence of carcinoma in Bangladesh. The study reveals a mixed understanding of cancer risk factors among participants. Aging, a critical risk factor for cancer due to cumulative genetic mutations and reduced DNA repair capacity over time [16], was recognized by only 34.7% of participants, while 38.9% were unsure. This indicates a lack of awareness, likely stemming from the emphasis in public health campaigns on modifiable risk factors such as smoking and diet, which often overshadow non-modifiable factors like aging. Enhancing public health education to provide a balanced perspective could help underscore the significance of aging as a cancer risk factor. Family history as a risk factor was acknowledged by 64.2% (340/530) of participants, likely reflecting the success of public health messaging about genetic predispositions. However, awareness of obesity as a cancer risk factor was relatively lower, with only 43.2% (229/530) recognizing it and 35.3% (187/530) expressing uncertainty. According to the WHO, obesity's link to cancer is often overshadowed by its associations with cardiovascular diseases and diabetes in public health discourse [13]. Regarding reproductive factors, only 10.2% of participants identified having more children as a cancer risk factor, while 49.1% (260/530) disagreed. This lack of recognition could stem from limited public health messaging on the relationship between reproductive factors and cancer risk, as well as insufficient awareness of how childbearing affects cancer risk through hormonal and genetic mechanisms [17]. Participants demonstrated high awareness of radiation exposure (86.6%, 459/530) and smoking (90.9%, 482/530) as cancer risk factors, likely due to their prominent coverage in public health campaigns and educational materials. Conversely, awareness of the role of infectious agents in cancer was comparatively lower, at 62.8% (333/530). This gap may be attributed to the complex relationship between infectious agents and cancer, which is less frequently emphasized in public health messaging. According to the WHO, the multifaceted nature of this link and its limited visibility in public discourse contribute to reduced awareness [18]. Another notable area of uncertainty was diet. While 49.8% (264/530) of participants recognized diet as a cancer risk factor, 35.1% (183/530) were unsure. Research consistently highlights that diets rich in fruits, vegetables, and whole grains are associated with a reduced risk of cancer, while diets high in processed foods and red meats are linked to increased risk [19]. Finally, 52.1% (276/530) of participants identified artificial sweeteners as a possible cancer risk factor, while 35.1% (186/530) were uncertain. The high level of uncertainty may stem from conflicting and often inconclusive scientific evidence regarding the link between artificial sweeteners and cancer risk [20]. In addition, inconsistencies in media coverage where dramatic claims are sometimes made without balanced analysis likely contribute to public confusion. Such informational disparities underscore the need for clear, evidence-based communication to reduce uncertainty in this area. The study highlights notable gender differences in attitudes toward cancer care and support in Bangladesh, revealing diverse perspectives that reflect varying levels of understanding and

XSL•FO
RenderX

prioritization. Approximately 52% (273/530) of females and 34% (198/530) of males expressed positive attitudes toward patients with cancer, indicating that they would have no issues with patients with cancer living around them. This may indicate an increasing awareness among some individuals about the importance of prioritizing cancer patient needs, although other factors such as social desirability bias could also play a role. Bangladesh's National Cancer Control Strategy and Plan of Action 2009 - 2015, developed with the assistance of the WHO, aimed to establish a comprehensive continuum of cancer care [4]. However, nearly 11% (58/530) of respondents reported discomfort or inconvenience around cancer patients, likely due to limited knowledge about cancer and misconceptions regarding its transmission. A lack of understanding and awareness about cancer is further evidenced by a recent breast cancer survey in Bangladesh. According to Sarker et al, the mean knowledge score about breast cancer symptoms was 2.94 (SD 1.14) out of 8, with an overall correct response rate of only 36.8%, based on 400 participants. These findings underscore the urgent need for culturally, socially, and demographically tailored educational intervention programs to raise cancer awareness and promote self-examination practices [12]. The study also revealed that 57.2% (303/530) of females and approximately 36% (190/530) of males agreed on the importance of increased government funding for cancer treatment and care, emphasizing the need to ensure financial support for cancer patients. Financial challenges, particularly among low-income individuals, significantly impact cancer care outcomes. According to recent findings, patients with limited financial resources face severe consequences, as out-of-pocket health care expenditures in Bangladesh exacerbate financial hardship for poorer households compared to wealthier ones [6]. The higher level of agreement among women could possibly be influenced by a greater trust in governmental support systems, although this interpretation requires further investigation. However, issues of affordability remain pressing; essential services for standard breast cancer care are often inaccessible to the average Bangladeshi family due to profiteering within an overstressed healthcare system [21]. When it comes to caregiving responsibilities, 52% (273/530) of females and 34.2% (181/530) of males agreed that care for cancer patients should be provided. These differences point toward potentially complex views surrounding caregiving responsibilities; however, more in-depth qualitative research would be needed to understand the underlying beliefs. Patients who experience negative societal attitudes toward their condition are 2.5 times more likely to endure from depression than those who feel positively supported [22]. Stigmatization and societal perceptions further complicate cancer care and support for patients. The importance of accessible banking and insurance policies for cancer patients also revealed significant gender disparities, with 36% (187/530) of females and 21% (110/530) of males agreeing on their necessity. These findings underline statistically significant differences between genders across key questions related to cancer care, patient needs, and policies.

Limitation

Ultimately, this study identified statistically significant gender differences in attitudes toward cancer care and policy; it relied primarily on bivariate analyses (eg, χ^2 tests). Consequently, the observed differences may be confounded by other sociodemographic variables such as age, education level, or personal experiences with cancer. To more accurately isolate the effect of gender, future research should use multivariate analyses such as logistic regression or ANCOVA to control for potential confounders. Incorporating such methods would provide a more robust understanding of the underlying factors influencing attitudes toward cancer care.

Conclusion

This study underscores not only a critical deficit in cancer awareness among university students in Bangladesh but also deeper societal and systemic issues that may hinder effective cancer prevention and care. The low level of screening participation, despite widespread exposure to cancer within families, reveals a significant disconnect between awareness and action reflecting broader barriers such as health system limitations, cultural attitudes, and perceived invulnerability among youth. Gender disparities in attitudes toward cancer care further suggest that public health initiatives must be responsive to sociocultural norms and gendered perceptions of health responsibility. Policies and educational programs must be redesigned to actively engage men and women in ways that acknowledge these differences, promoting inclusive, participatory cancer literacy strategies. Moreover, this research calls attention to the need for multi-level interventions that go beyond individual knowledge. Effective change will require collaboration across government, educational institutions, and health organizations to create integrated programs that normalize cancer screening, reduce stigma, and make preventive care accessible and affordable. There is also a pressing need to embed health education, particularly around non-modifiable and lesser-known risk factors, into university curricula to foster informed health behaviors early in adulthood. Finally, these findings offer a foundation for future longitudinal and qualitative studies that can explore the underlying psychological, economic, and structural barriers to cancer awareness and prevention. Addressing these broader determinants is essential for transforming knowledge into sustainable health practices and for reducing the cancer burden in Bangladesh and other low-resource settings facing similar challenges.

Acknowledgments

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

All authors declared that they had insufficient or no funding to support open access publication of this manuscript, including from affiliated organizations or institutions, funding agencies, or other organizations. JMIR Publications provided APF support for the publication of this article.

XSL•FO
RenderX

Disclaimer

Generative AI tools were not used in the writing or analysis of this manuscript.

Data Availability

The datasets generated and analyzed during this cross-sectional study on cancer awareness and attitudes among university students in Bangladesh (conducted from December 2022 to March 2024) are not publicly available due to ethical and privacy restrictions imposed by the BRAC University Institutional Review Board. However, deidentified data supporting the findings of this study are available from the corresponding author upon reasonable request.

Authors' Contributions

All authors contributed significantly to the design, data collection, analysis, and interpretation of the study. Dr. Munima Haque supervised the project and provided critical revisions. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Checklist 1
STROBE checklist for Cross-Sectional Study.
[DOCX File, 38 KB - formative_v9i1e75651_app1.docx]

References

1. Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2024;74(3):229-263. [doi: 10.3322/caac.21834] [Medline: 38572751]
2. Jokhadze N, Das A, Dizon DS. Global cancer statistics: a healthy population relies on population health. CA Cancer J Clin 2024;74(3):224-226. [doi: 10.3322/caac.21838] [Medline: 38572764]
3. Cancer bangladesh 2020 country profile. World Health Organization. URL: https://www.who.int/publications/m/item/cancer-bgd-2020 [accessed 2024-08-22]
4. Hussain SMA. Comprehensive update on cancer scenario of Bangladesh. South Asian J Cancer 2013 Oct;2(4):279-284. [doi: 10.4103/2278-330X.119901] [Medline: 24455659]
5. Uddin AFMK, Khan ZJ, Islam J, Mahmud A. Cancer care scenario in Bangladesh. South Asian J Cancer 2013 Apr;2(2):102-104. [doi: 10.4103/2278-330X.110510] [Medline: 24455570]
6. Rahman M, Zhang C, Swe KT, et al. Disease-specific out-of-pocket healthcare expenditure in urban Bangladesh: A Bayesian analysis. PLoS ONE 2020;15(1):e0227565. [doi: 10.1371/journal.pone.0227565]
7. Promoting cancer early diagnosis. World Health Organization. 2025 Mar 18. URL: https://www.who.int/activities/promoting-cancer-early-diagnosis [accessed 2024-08-22]
8. Hart RI, Cowie FJ, Jesudason AB, Lawton J. Adolescents and young adults' (AYA) views on their cancer knowledge prior to diagnosis: findings from a qualitative study involving AYA receiving cancer care. Health Expect 2021 Apr;24(2):307-316. [doi: 10.1111/hex.13170] [Medline: 33275814]
9. Fidler MM, Gupta S, Soerjomataram I, Ferlay J, Steliarova-Foucher E, Bray F. Cancer incidence and mortality among young adults aged 20-39 years worldwide in 2012: a population-based study. Lancet Oncol 2017 Dec;18(12):1579-1589. [doi: 10.1016/S1470-2045(17)30677-0] [Medline: 29111259]
10. Otufowora A, Liu Y, Young H II, et al. Sex differences in willingness to participate in research based on study risk level among a community sample of African Americans in North Central Florida. J Immigrant Minority Health 2021 Feb;23(1):19-25. [doi: 10.1007/s10903-020-01015-4]
11. Akter M, Ullah MO. Awareness levels of breast cancer among female university and medical college students in Sylhet city of Bangladesh. Cancer Rep 2022 Nov;5(11). [doi: 10.1002/cnr2.1608]
12. Sarker R, Islam MS, Moonajilin MS, Rahman M, Gesesew HA, Ward PR. Knowledge of breast cancer and breast self-examination practices and its barriers among university female students in Bangladesh: findings from a cross-sectional study. PLoS ONE 2022;17(6):e0270417. [doi: 10.1371/journal.pone.0270417] [Medline: 35763525]
13. Tanaka K, Tsuji I, Tamakoshi A, et al. Obesity and liver cancer risk: an evaluation based on a systematic review of epidemiologic evidence among the Japanese population. Jpn J Clin Oncol 2012 Mar;42(3):212-221. [doi: 10.1093/jjco/hyr198] [Medline: 22241822]
14. Hoffman RM, Elmore JG, Fairfield KM, Gerstein BS, Levin CA, Pignone MP. Lack of shared decision making in cancer screening discussions. Am J Prev Med 2014 Sep;47(3):251-259. [doi: 10.1016/j.amepre.2014.04.011]
15. Novak JR, Peak T, Gast J, Arnell M. Associations between masculine norms and health-care utilization in highly religious, heterosexual men. Am J Mens Health 2019 May;13(3). [doi: 10.1177/1557988319856739]

XSL•FO
RenderX

16. Smetana K Jr, Lacina L, Szabo P, Dvořánková B, Brož P, Šedo A. Ageing as an important risk factor for cancer. Anticancer Res 2016 Oct;36(10):5009-5017. [doi: 10.21873/anticanres.11069] [Medline: 27798859]
17. Tavani A, Ricci E, La Vecchia C, et al. Influence of menstrual and reproductive factors on ovarian cancer risk in women with and without family history of breast or ovarian cancer. Int J Epidemiol 2000 Oct;29(5):799-802. [doi: 10.1093/ije/29.5.799] [Medline: 11034959]
18. Heckbert SR, Weiss NS, Hornung SK, Eaton DL, Motulsky AG. Glutathione s-transferase and epoxide hydrolase activity in human leukocytes in relation to risk of lung cancer and other smoking-related cancers. JNCI 1992 Mar 18;84(6):414-422. [doi: 10.1093/jnci/84.6.414]
19. Ksouri R. Food components and diet habits: chief factors of cancer development. Food Quality and Safety 2019 Dec 31;3(4):227-231. [doi: 10.1093/fqsafe/fyz021]
20. Debras C, Chazelas E, Srour B, et al. Artificial sweeteners and cancer risk: results from the NutriNet-Santé population-based cohort study. PLoS Med 2022 Mar;19(3):e1003950. [doi: 10.1371/journal.pmed.1003950] [Medline: 35324894]
21. Story HL, Love RR, Salim R, Roberto AJ, Krieger JL, Ginsburg OM. Improving outcomes from breast cancer in a low-income country: lessons from bangladesh. Int J Breast Cancer 2012;2012:423562. [doi: 10.1155/2012/423562] [Medline: 22295245]
22. Cho J, Smith K, Choi EK, et al. Public attitudes toward cancer and cancer patients: a national survey in Korea. Psychooncology 2013 Mar;22(3):605-613. [doi: 10.1002/pon.3041] [Medline: 22344743]

Abbreviations

STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
WHO: World Health Organization

Edited by A Mavragani; submitted 08.04.25; peer-reviewed by A Adeoye, C Onah; revised version received 14.06.25; accepted 14.06.25; published 13.11.25.

Please cite as:
Tabassum M, Farhin N, Afrose A, Surovy MM, Tasnim R, Affia Heaven MH, Noor J, Haque M
Awareness and Attitudes of University Students in Bangladesh Toward Cancer: Cross-Sectional Study
JMIR Form Res 2025;9:e75651
URL: https://formative.jmir.org/2025/1/e75651
doi:10.2196/75651

© Maliha Tabassum, Nafisa Farhin, Afsana Afrose, Mahfuza Moin Surovy, Rubaiya Tasnim, Most. Humayra Affia Heaven, Jannatun Noor, Munima Haque. Originally published in JMIR Formative Research (https://formative.jmir.org), 13.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Usability and User Experience of a Digital Platform Prototype (Healthy Bone) to Promote Pharmacological and Nonpharmacological Treatment in Patients With Osteoporosis: Mixed Methods Study

Anabela Barcelos[1,2,3*], PhD, MD; Alexandre Moniz[2*], MSc; Ana Rita Henriques[2], MSc; Carolina Mazeda[1,2], MD, MSc; Elsa Frazão Mateus[4], PhD; Ana Machado[2], MSc; Cátia F Gonçalves[2], PhD, BSc; Maiara Moreto[2], MSc; Mafalda F Almeida[3], BSc; Nuno Mendonça[2], PhD, BSc; Daniela Costa[2], PhD; Marta M Marques[2,3], PhD; Pedro Domingos[5], MBA; Yoann Nesme[5], MBA; Helena Canhão[2], PhD, MD; Sónia Dias[3], PhD; Ana Maria Rodrigues[2], PhD, MD

[1]Rheumatology Department, Unidade Local de Saúde de Aveiro, Rua Artur Ravara, 35, Aveiro, Portugal

[2]EpiDoc Unit, NOVA Medical School, Comprehensive Health Research Center (CHRC), Universidade NOVA de Lisboa, Lisboa, Portugal

[3]NOVA National School of Public Health, Public Health Research Centre, Comprehensive Health Research Center (CHRC), Universidade NOVA de Lisboa, Lisboa, Portugal

[4]Liga Portuguesa Contra as Doenças Reumáticas, Lisboa, Portugal

[5]OrangeBird Lda, Lisboa, Portugal

[*]these authors contributed equally

Corresponding Author:
Anabela Barcelos, PhD, MD
Rheumatology Department, Unidade Local de Saúde de Aveiro, Rua Artur Ravara, 35, Aveiro, Portugal

Abstract

Background: Osteoporosis-related fractures significantly impact older adults, often leading to disability and even premature death. While pharmacological and nonpharmacological interventions are widely recommended for managing osteoporosis, adherence to these interventions remains low. To address this challenge, we developed the Healthy Bone digital platform (desktop, mobile app, and smart TV internet-based) for use in clinical settings to improve disease management and treatment adherence. It integrates a multimedia health-related behavioral change program with a patient monitoring and management system.

Objective: This study aimed to evaluate the usability and user experience of the desktop version of the Healthy Bone digital platform prototype from the patients' perspective. The findings will provide valuable insights into optimizing the digital platform and enhancing its functionality.

Methods: A mixed-methods study was conducted. During usability testing, patients completed tasks simulating real-world use of the platform while using a Think-Aloud approach. After each task, participants filled out an After Scenario Questionnaire to assess task satisfaction. Subsequently, participants completed the System Usability Scale (SUS) and the eHealth Usability Benchmarking Instrument (HUBBI) to measure usability quantitatively. Following this, semistructured interviews were conducted to explore participants' experiences with the platform in greater depth. Descriptive statistics were used for quantitative analysis. Qualitative data analysis involved a combined deductive and inductive approach, ensuring a comprehensive evaluation of the platform's usability and user experience. Deductive content analysis was guided by an ontology of eHealth usability components, while thematic analysis adhered to Braun and Clarke's method to identify emerging themes.

Results: Seven participants evaluated the digital platform, reporting high usability with a mean overall SUS score of 87.1 (SD 13.3). Similarly, good usability was reported across all categories of the HUBBI, except for the guidance and support category, which presented moderate levels of usability (mean 3.3, SD 1.1). Patients reported high levels of task satisfaction and identified 24 unique usability issues, predominantly related to the basic system performance, interface design, and navigation and structure categories of the eHealth usability ontology. Overall, patients had positive perceptions and acceptability of the digital platform, highlighting its simplicity, accessibility, utility, and potential to empower those with osteoporosis. Barriers to usage included limited skills, lack of suitable equipment, and time, while facilitators included motivation for behavior change, health benefits, and the decrease of potential inequalities.

Conclusions: This study provided valuable insights into the usability and user experience of the desktop version of the Healthy Bone digital platform prototype. These findings will play a key role in optimizing the platform to ensure it is effectively tailored

XSL•FO
RenderX

to the needs of the target population. This platform adds an understanding of how various information and communication technology tools can support and benefit large numbers of osteoporosis patients in society.

(JMIR Form Res 2025;9:e72468) doi:10.2196/72468

KEYWORDS

eHealth; information and communication technologies; osteoporosis-related fractures; older adults; osteoporosis; usability

Introduction

Background

Worldwide, osteoporosis causes more than 8.9 million fractures annually, a number that is estimated to rise to around 200 million over the next 50 years [1]. Estimates have shown that the direct health care costs resulting from fractures in postmenopausal or over 50-year-old women are €57 - 74 (US$62-US$80) million per year [2,3]. Across Europe, the number of fracture-related deaths is comparable to or higher than that of some of the most common causes of death, such as diabetes, lung cancer, and chronic lower respiratory tract diseases [4]. Osteoporosis-related fractures constitute a major public health challenge and result in increased disability, morbidity, and mortality, especially among older adults [5], imposing a huge and growing economic burden on health care systems [5,6].

Given the urgency of addressing this issue, there is a need to promote appropriate treatment of osteoporosis and adopt measures to prevent the occurrence of first and recurrent osteoporosis-related fractures. To achieve this, pharmacological and nonpharmacological (eg, good nutrition, exercise and physical activity, and avoidance of harmful behaviors, such as smoking and alcohol consumption) management strategies have been widely recommended [7-9]. However, a considerable treatment gap is currently characterized by low screening and treatment adherence rates [5,10,11]. For example, poor adherence to exercise (36% - 55%) and physical activity (14% - 58%) [12] and low to moderate adherence to pharmacological treatments (40% - 70%) [13,14] by patients with osteoporosis have been previously reported. This highlights the need for developing and implementing self-management interventions focused on behavior change support, especially in ensuring patients adopt pharmacological and nonpharmacological osteoporosis management strategies.

The emergence of new information and communication technologies (ICTs) offers promising tools to address barriers to treatment adherence [15]. The rapid development of computer-based technologies—coupled with the widespread availability of the internet and mobile and TV apps—has led to the rise of the concepts of “eHealth” and “mHealth” [16-18]. These digital solutions have the potential to foster positive health behavior change, support healthier lifestyles, and assist in the diagnosis and treatment of diseases [19]. As a growing interdisciplinary field, eHealth and mHealth sit at the intersection of medical informatics, public health, and business. It refers to health services and information that are delivered or enhanced through the internet and related technologies [19-21]. Computer-based technologies can support health care professionals in remotely monitoring patients and assisting with daily activities, such as medication reminders and education about personal habits and behavioral patterns. These tools enable the customization of interventions, particularly in the management of chronic diseases [22-24]. Several studies have highlighted the positive impact of such technologies on enhancing patient competence [24-26]. Informed patients are more likely to adopt healthy behaviors and effectively manage their conditions [27,28].

With the aim of promoting health care, a growing number of studies have explored the use of mobile apps across different medical domains, showing promising results [26,27,29-31]. These apps allow patients to access relevant health information anytime and anywhere, empowering them in their self-care journey.

In the context of osteoporosis, mHealth apps have emerged as valuable tools for both patients and health care providers [27,32,33]. They can enhance disease management by providing broad, cost-effective, and accessible support to a large population [34]. Effective self-management apps for osteoporosis should include comprehensive educational content, covering the nature of the disease [35], dietary and medication guidance [36-38], physical activity recommendations [39], and strategies to prevent falls and fractures [40].

The idea of individual responsibility in the self-management of osteoporosis has gained momentum in recent decades, and mHealth apps can play a crucial role in supporting this shift. However, despite the growing number of apps aimed at osteoporosis self-management, many lack clinically validated evidence of effectiveness [33,41]. According to Alhussein et al [33], approximately 75% of the mHealth apps reviewed in their systematic review and meta-analysis had not undergone premarket prospective multicenter randomized controlled trials (RCTs), likely due to the high costs and time-consuming nature of patient recruitment [42]. In addition, many apps failed to demonstrate their usability and acceptability among users [33,43].

More recently, Bendtsen et al [41] assessed the eHealth literacy of Danish osteoporosis patients using the app “My Bones.” Their findings indicated that the app had good usability and the potential to reduce general practitioner visits, improve health outcomes, and serve as a valuable complement to standard health care and social services.

A successful mHealth app should go beyond promoting behaviors such as healthy eating, physical activity, or medication adherence. It must foster a genuine sense of individual health responsibility and support the self-management of osteoporosis in a sustainable and meaningful way.

An international consortium coordinated by our research group has successfully developed and evaluated a novel digital

XSL•FO
RenderX

platform, the Saúde.come TV app [44]. This interactive multimodule, customizable app delivered a home-based program to promote healthy behaviors among older adults with food insecurity. The results showed improvement in dietary behaviors, physical activity, and a reduction in participants' BMI [31]. The success of this TV platform has reinforced the development of similar platforms, namely for patients with osteoporosis with or without prevalent osteoporosis-related fractures. Therefore, the prototype for the Healthy Bone digital platform (desktop, mobile app, and smart TV internet-based) was developed as a web-based multimodule and interactive application containing a combination of patient monitoring and management systems with a multimedia health-related behavioral change program. It includes education strategies, patient empowerment, self-management reminders, and prompts designed to improve disease management in a clinical setting and increase long-term adoption of nonpharmacological and pharmacological osteoporosis treatment.

To ensure that the Healthy Bone digital platform is satisfactory, well-adapted to its intended users, and successfully implemented, it is crucial to assess its usability. Usability is defined as "the extent to which specified users can use a system, product, or service to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use" [45]. The usability assessment is considered an essential step in developing eHealth innovations, as it improves interfaces and enhances user interaction [46]. The Healthy Bone digital platform prototype was designed to facilitate and promote behavior changes. To verify if this goal is achieved, it is important to assess the usability and user experience of the Healthy Bone digital platform.

Objective

This study aimed to evaluate the system usability and user experience of the desktop version of the Healthy Bone digital platform prototype from the patients' perspective. The information gathered will inform the optimization of this digital platform.

Methods

Study Design

A mixed-methods study used a convergent approach, combining quantitative data from self-reported questionnaires and qualitative data from Think-Aloud sessions and semistructured interviews. To enhance the transparency and reporting, this study adhered to the Good Reporting of a Mixed Methods Study (GRAMMS) checklist [47] (Checklist 1).

Participants and Recruitment

Participants were recruited at the rheumatology department of Centro Hospitalar do Baixo Vouga, Aveiro, Portugal. The study sample included patients with osteoporosis recruited consecutively from the Osteoporosis Clinic (representing the potential end users of the digital platform). A purposive sample strategy was used to ensure the inclusion of participants who might provide relevant, valuable, and detailed feedback. Participants were included if they fulfilled the following criteria: (1) community-dwelling older adults (≥65 years old); (2) diagnosis of osteoporosis based on either the *t* score criterion ($t \leq -2.5$ at the lumbar spine, or hip neck); (3) basic knowledge about the use of digital solutions; (4) proficiency in Portuguese; and (5) able and willing to give written informed consent and comply with the requirements of the study. Participants were excluded if they could not comply with the study procedures due to hearing or visual loss, significant cognitive impairment, and illiteracy. A minimum sample size of 5 participants was set for usability testing, as this is sufficient to identify over 80% of usability issues [48].

Healthy Bone Digital Platform

The Healthy Bone digital platform prototype, accessible via desktop, mobile app, and smart TV, was developed as a web-based, multimodule, and interactive application. It integrates a multimedia health-related behavioral change program with a patient monitoring and management system. This prototype was developed based on a previous app developed by our team, the Saúde.Come TV app [31,44]. The Healthy Bone platform integrates educational strategies, patient empowerment tools, self-management reminders, and prompts designed to improve disease management and promote long-term adherence to pharmacological and nonpharmacological osteoporosis treatments. The Healthy Bone digital platform was designed to be used in a clinical setting to improve the long-term management of patients who live with osteoporosis with or without osteoporosis-related fractures. This platform aims to deliver comprehensive and user-centered support for managing osteoporosis.

Healthy Bone System Architecture

The Healthy Bone digital platform was designed by the principal investigator (AMR) and custom-developed by the Orange Bird company. Coinvestigators contributed to content development (HC, AB, AM, NM, and MMM) with guidance on the visual display provided by experienced app developers. EFM (coinvestigator and patient research partner) contributed to content development through several meetings with the principal investigator. Other patients were also involved in structured interviews. The technical architecture of the Healthy Bone platform (Figure 1) is based on the open-source Drupal Content Management System (CMS). The CMS works as the main repository for content (such as users, streaming videos, and questionnaires) and provides the web interface of the digital solution. An Application Programming Interface (API) was developed so that the CMS (backend) can interact with other interfaces, such as mobile apps. Additional off-the-shelf and custom modules were deployed on top of the base CMS. The development was implemented to allow further features to be added as requested.

XSL•FO
RenderX

Figure 1. The Healthy Bone system architecture. API: Application Programming Interface.

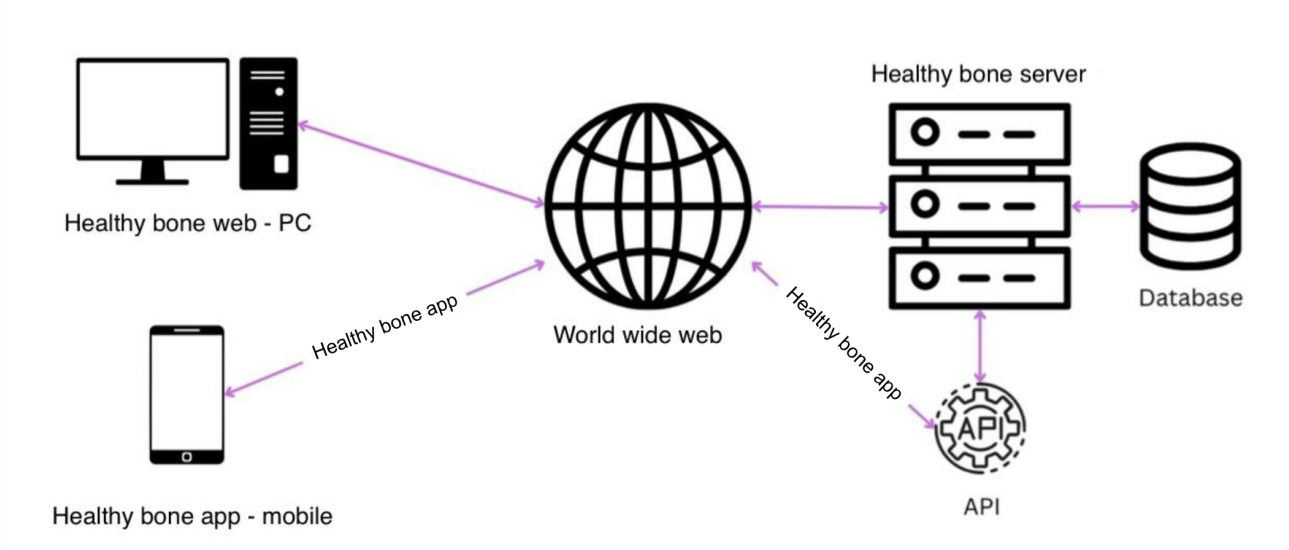

The Healthy Bone digital platform prototype has several pages, each offering different content and features (Figures 2 and 3). On the main page, titled "My Program," patients can access multimedia related to osteoporosis care, including videos they can watch, like or dislike, and mark as "Favorites." The "Questionnaires" page allows participants to complete weekly questionnaires on topics covered in the daily multimedia content. On this page, they can also fill out patient-reported outcome measures, such as the Health Assessment Questionnaire (HAQ) [49], the Self-Mini Nutritional Assessment [50], or the short form of the International Physical Activity Questionnaire [51]. The "Clinical Information" page displays graphical representations of the participants' responses to the questionnaire. On the "Medication" page, participants can track their medication and rheumatology appointments on the "Appointments" page. The "Messages" page offers a dedicated chat feature for direct communication between participants and medical doctors. Finally, the "Profile" page displays participants' personal information and their progress through the program, including details such as their current phase, the number of videos they have watched, and the number of questionnaires they completed.

Figure 2. Screenshot of the Healthy Bone platform main page (My Program).

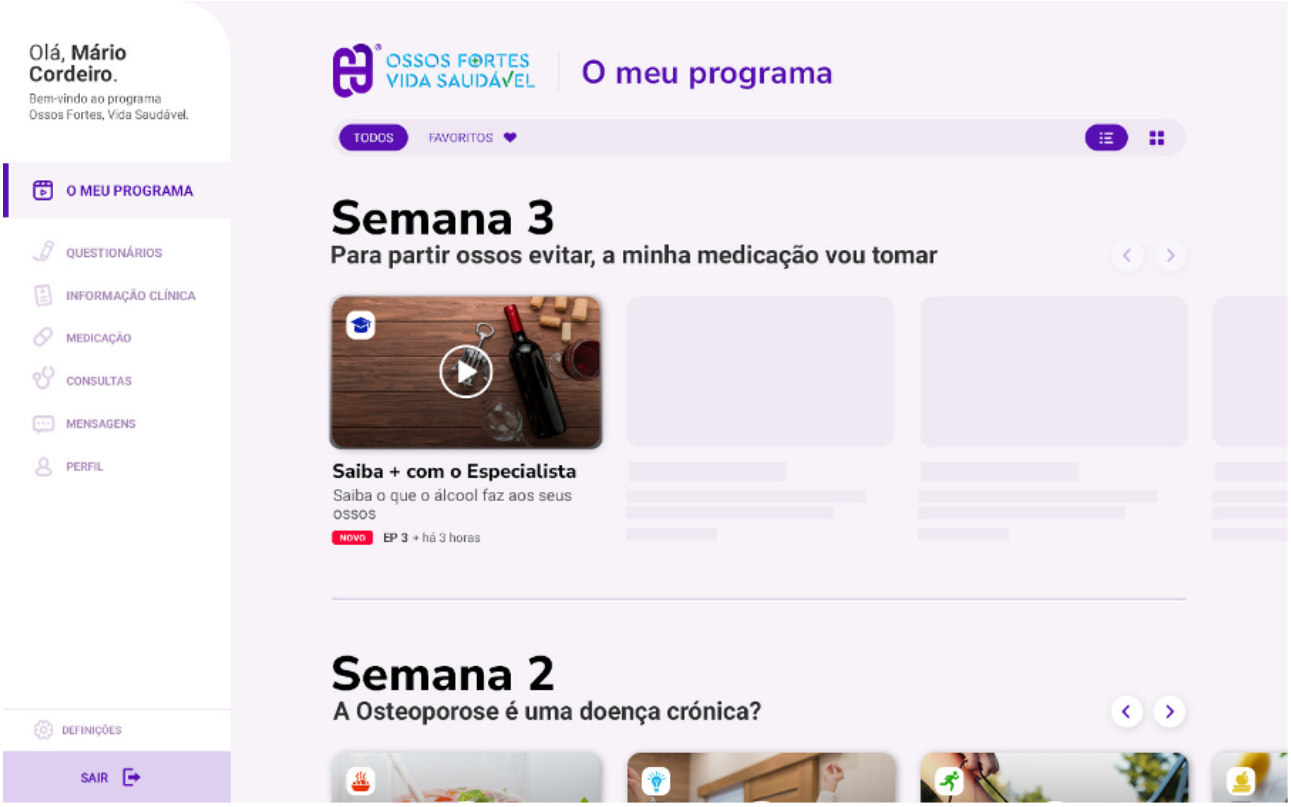

XSL•FO
RenderX

Figure 3. Synthesis of the platform's several content and features.

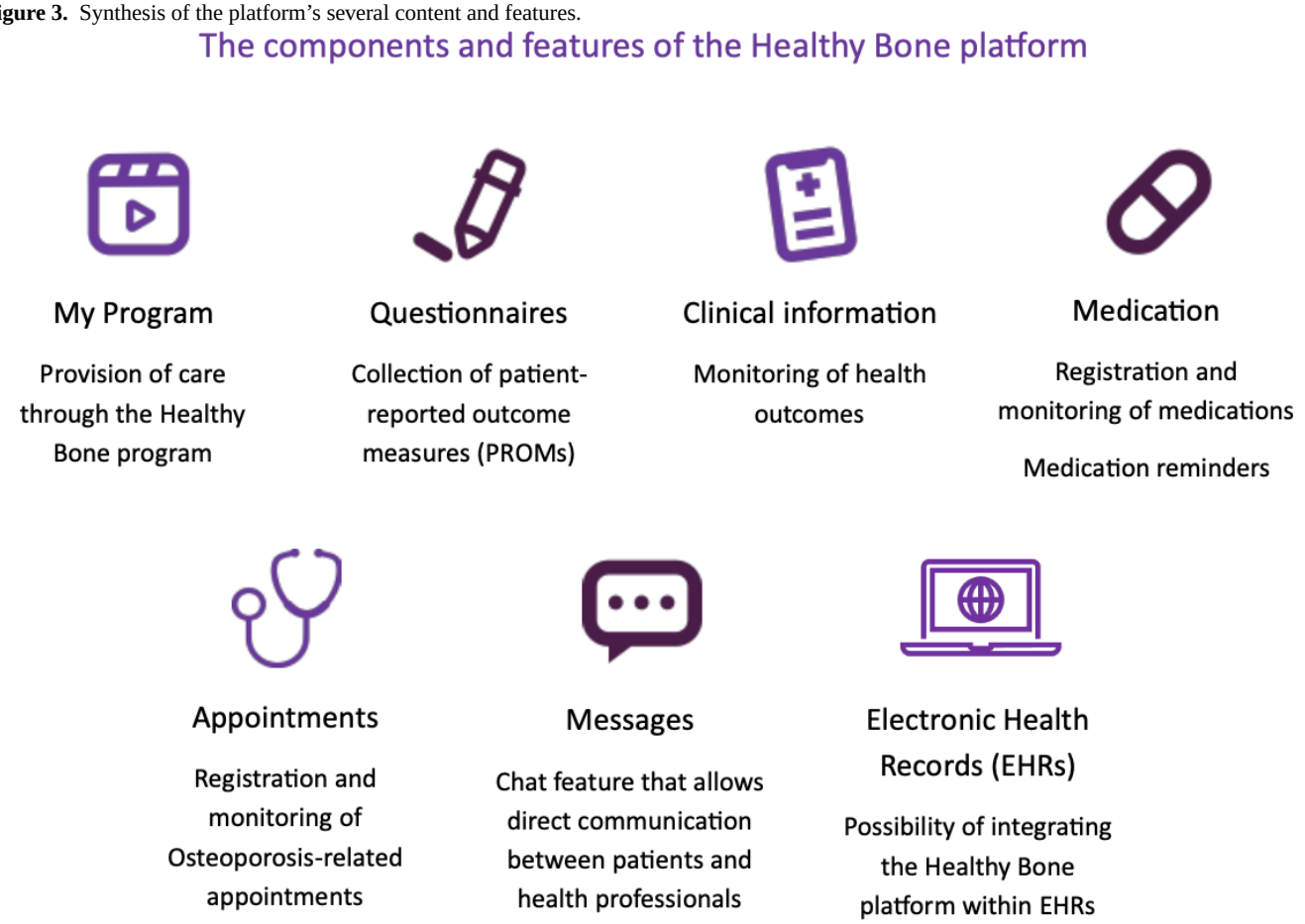

This digital platform allows health care professionals to access patient data and monitor program adherence, potentially improving continuity of care.

Healthy Bone Program Description

The Healthy Bone program is divided into 2 phases, lasting 52 weeks (Figure 4). Phase 1 lasts for 12 weeks, is weekly organized, and is composed of original multimedia content, with one video released daily. The multimedia content includes educational content on healthy aging and bone fragility, exercise plans, nutritional advice, strategies for behavior change, and treatment reminders. Each week is composed of 5 modules: (1) Know+ with your specialist, (2) Move+ in your home, (3) Nutrition+ Corner, (4) Change+ Corner, and (5) Chef+ Kitchen. In phase 2, which occurs from Week 13 to Week 52, participants gain access to the entire library of multimedia content, now organized by themes. Additional content may be uploaded to the platform to address individual participants' needs. Throughout the program, participants may access their clinical information, including all relevant medication, and contact their health professionals through a dedicated chat feature.

XSL•FO
RenderX

Figure 4. Overview of the Healthy Bone program.

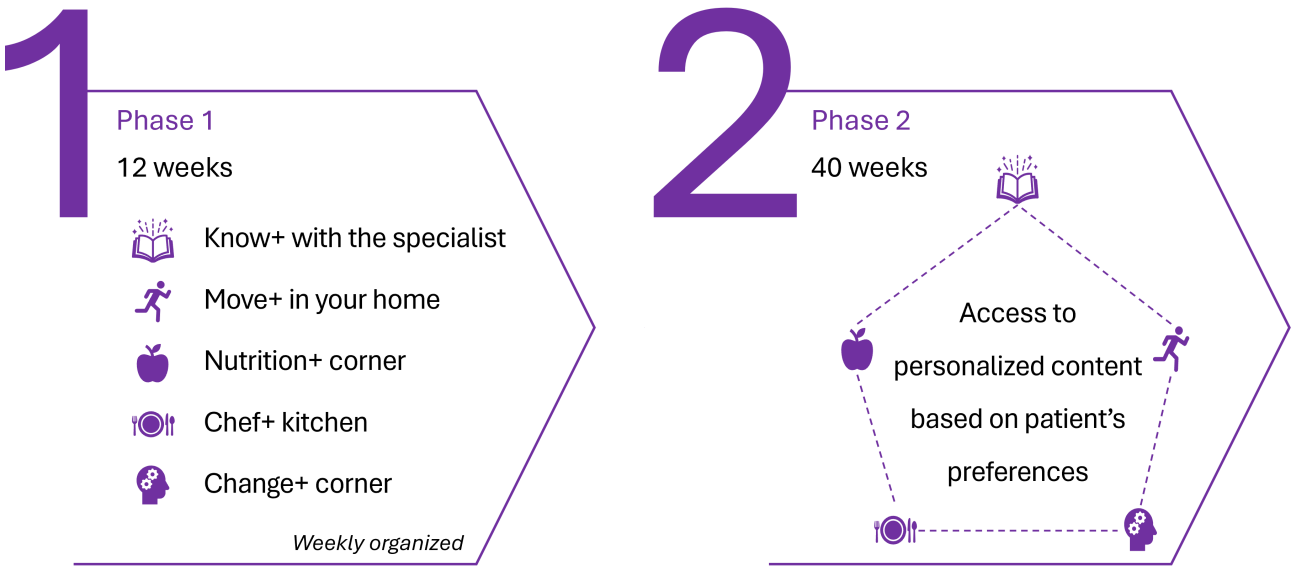

Usability Testing

Data Collection

Participants were asked to complete a questionnaire that collected their sociodemographic information and clinical characteristics. Sociodemographic characteristics included age, sex, educational background, use of technological devices, and the most common tasks they perform using a computer. For clinical characterization, participants completed the European Portuguese versions of the EuroQol with 5 dimensions and 3 levels (EQ-5D-3L) [52] to assess quality of life, and the HAQ [49] to evaluate functional disability.

Usability testing (qualitative and quantitative methods) and semistructured interviews were performed to test and gather feedback on the Healthy Bone digital platform. A trained researcher (AM) conducted these sessions in Portuguese, which lasted approximately 70 minutes and were audio and screen recorded.

The Healthy Bone digital platform prototype was assessed through usability tests using a Think-Aloud approach [53]. This method was used to capture the end user's thoughts as they interacted with the platform. First, participants were briefed on the study's aims and introduced to the platform's key features. They were then asked to complete several tasks with assistance from the moderator.

In the first task, participants had 5 minutes to explore the platform without specific instructions. After that, they were required to complete tasks that illustrated the use of the platform while verbalizing their thoughts. In task 2, patients were instructed to explore the main page, "My Program." This entailed selecting and watching one of the available videos, clicking on the "Favorite" and "Like or Dislike" buttons, returning to the main page, toggling between the grid and slide views of the videos, and lastly, navigating to the "Favorites" page. In task 3, patients were asked to navigate to the "Questionnaires" page, select and complete one of the available weekly questionnaires, return to the "Questionnaires" page, and then select and complete one. In task 4, participants were instructed to navigate to the "Clinical Information" page. Task 5 required participants to explore the "Medication" page by clicking on it, visualizing the medication listed in the "Overview" tab, then going to the "Calendar" tab and checking all medications of that current day as taken. For task 6, participants explored the "Appointments" page, while in task 7, they had to navigate to the "Messages" page, write a message in the chat box, and send it. Finally, for task 8, patients only had to click on and investigate the "Profile" page.

After completing each task (except for the first task), participants were given an After Scenario Questionnaire (ASQ) [54] to measure task satisfaction. The ASQ includes three items with a 7-point scale, in which lower ratings represent a more favorable usability satisfaction of the system [54].

To assess the usability of the platform, and after carrying out all tasks, participants were asked to fill out:

1. System Usability Scale (SUS) [55]: SUS consists of 10 statements rated on a 5-point Likert scale of agreement strength. Scores range from 0 to 100 (higher scores representing higher usability) and can be interpreted by the grading scale proposed by [56] (eg, scores between 84.1 and 100 are an "A+"; between 80.8 and 84 are an "A"; between 78.9 and 80.7 are an "A-"; between 77.2 and 78.8 are a "B+"; etc). This scale is validated in the Portuguese population.
2. eHealth Usability Benchmarking Instrument (HUBBI) [46]: HUBBI is a new usability benchmarking tool that assesses usability in 7 domains: Basic System Performance, Task-Technology Fit, Interface Design, Navigation and Structure, Information and Terminology, Guidance & Support, and Satisfaction [46]. The authors of the HUBBI recommend using a radar chart to present the results, as these results are provided at a category level. A radar chart displays multivariate data in a 2-dimensional format, showcasing 3 or more quantitative variables represented on axes that originate from a common point. This view allows for a comprehensive overview of how each aspect of usability is performing [46].

XSL•FO
RenderX

After the usability tests, semistructured interviews were conducted to explore participants' experiences with the Healthy Bone digital platform prototype, specifically their perceptions, the acceptability of the digital platform, usability issues, suggestions for improvement, and potential barriers and facilitators to use the digital platform. An interview guide was developed to conduct the interviews (Multimedia Appendix 1).

Data Analysis

This study followed a convergent mixed-methods approach, in which quantitative and qualitative data were collected in parallel, analyzed separately, and then integrated during interpretation. This process involved comparing findings and identifying convergences and divergences. For instance, SUS and HUBBI scores were subsequently analyzed alongside themes and usability issues emerging from the qualitative analysis to provide a more comprehensive understanding of the usability and user experience of the Healthy Bone digital platform.

Quantitative Analysis

Descriptive statistics were computed for sociodemographic and clinical characteristics, SUS, and HUBBI scores. Frequencies or proportions were used for categorical variables, means, and SDs were calculated for continuous variables. For the SUS, an overall score was calculated for each participant as previously described in the literature [57] so that final scores ranged from 0 to 100 (higher scores indicate better usability). All quantitative data were analyzed using SPSS software (version 29; IBM Corp) for macOS.

Qualitative Analysis

Before analyzing the qualitative data, the Think-Aloud usability tests and semistructured interview recordings were transcribed verbatim (AM) and anonymized with a pseudonym for each participant. The qualitative data analysis followed a deductive content analysis and inductive thematic analysis.

The deductive content analysis was performed using an ontology of usability problems [58] and was guided by the study's aim of identifying potential usability issues. This analysis was specifically applied to the Think-Aloud approach. Usability issues were assigned as minor (occurring infrequently and/or slightly increasing task completion time), serious (occurring frequently and/or severely increasing task completion time), or critical (occurring universally among participants and/or preventing task completion), according to the severity index of Duh et al [59].

The identification of the usability issues and the coding of their severity were performed by one researcher (AM) through analysis of the transcripts and screen recordings. A second researcher (AB) reviewed the entire codification. Any disagreements were discussed and settled between the 2 researchers, and if a consensus could not be reached, a third researcher (AMR) was consulted. Following this, AM and AB independently classified each usability issue into one of the 8 categories and 21 factors of the ontology proposed by Broekhuis et al [58]. Once again, the 2 researchers (AM and AB) discussed the classifications to reach a consensus, with AMR intervening if necessary.

For the inductive thematic analysis, data coding was guided by the study's aims of exploring patients' use of the Healthy Bone digital platform prototype. This method allowed for a more open analysis of the semistructured interviews, with themes emerging directly from the data. This analysis followed the 6-phase process described by Braun and Clarke [60]: (1) familiarization with the data; (2) coding; (3) searching for themes; (4) reviewing themes; (5) defining and naming themes; (6) writing up. The entire process was independently performed by 2 researchers (AM and AB), who discussed the codes and themes that emerged from the analysis. When consensus on the coding could not be reached, a third researcher (AMR) was consulted to resolve the disagreement.

Given the small sample size and the combined inductive and deductive approach, we decided not to calculate formal interrater reliability statistics; instead, ensuring rigor through frequent and iterative discussions.

Ethical Considerations

This study was conducted following the principles established by the Declaration of Helsinki, and ethical approval was granted by the NOVA Medical School|Faculdade de Ciências Médicas (NMS|FCM) Ethical Committee (CEFCM) (reference 196/2021/CEFCM). Eligible participants were provided with an information sheet detailing the purpose and procedures of the study. Those who agreed to participate received both written, patient-friendly materials and a verbal explanation of the study. Informed consent was obtained by the researcher, and all signed consent forms were securely stored in a location accessible only to the researchers. The procedures for obtaining informed consent were conducted in accordance with established ethical guidelines. The app is compliant with the European Union's General Data Protection Regulation. All the collected data were anonymized. Compensation details were not applicable to this study.

Results

Sample Demographics

Seven patients participated in the study (mean 70.9, SD 6.2 years old; 5/7, 71.4% female). Most participants (5/7, 71.4%) reported owning and regularly using a computer, and all owned and used a smartphone. Among those who reported using a computer regularly, the most common task was browsing the internet (80%). Sociodemographic and clinical characteristics are presented in Table 1.

XSL•FO
RenderX

Table . Patients' sociodemographic and clinical characteristics.

Variables	Total
Sex, n (%)	
Female	5 (71.4)
Male	2 (28.6)
Age, years	
Mean (SD)	70.9 (6.2)
Minimum	65
Maximum	81
Academic Qualification, n (%)	
Primary education (4 y of school)	1 (14.3)
Basic education (9 y of school)	1 (14.3)
Secondary education (12y of school)	2 (28.6)
Graduate degree	3 (42.9)
Use of technological devices[a], n (%)	
Smartphone	7 (100)
Computer (desktop or laptop)	5 (71.4)
Tablet	4 (57.1)
Smartwatch	1 (14.3)
Most common tasks performed while using the computer, n (%)	
Working	2 (40)
Browsing the internet	4 (80)
Using email	1 (20)
Making videocalls	2 (40)
Sending messages	2 (40)
Browsing social media platforms	2 (40)
Functional disability (HAQ[b]), mean (SD)	0.3 (0.6)
Health-related quality of life (EQ-5D-3L)[c], mean (SD)	0.8 (0.3)

[a]Technological devices add up to more than 7 since each participant may have more than one device.

[b]HAQ: Health Assessment Questionnaire.

[c]EQ-5D-3L: EuroQol, 5 dimensions, 3 levels.

Usability Benchmarks

Participants reported high levels of usability on the SUS, with the Healthy Bone digital platform prototype achieving a mean overall SUS score of 87.1 (SD 13.3), corresponding to a grade of A+. This score exceeds the standard normal SUS distribution of 68 (SD 12.5) described in the literature as an important and useful benchmark [56]. The results of the calculated SUS scores are shown in Table 2.

XSL•FO
RenderX

Table . System Usability Scale scores per participant, total System Usability Scale score per patient, and total mean System Usability Scale score.

Participant ID	P1[a]	P2[a]	P3[a]	P4[a]	P5[a]	P6[a]	P7[a]	Mean SUS[b] score (SD)
I think that I would like to use this system frequently	3	4	4	3	4	3	4	
I found the system unnecessarily complex	2	3	4	4	3	2	4	
I thought the system was easy to use	3	4	4	4	4	3	4	
I think that I would need the support of a technical person to be able to use this system	2	2	4	4	4	0	4	
I found the various functions in this system were well integrated	4	4	4	4	4	4	4	
I thought there was too much inconsistency in this system	3	4	4	4	4	4	4	
I would imagine that most people would learn to use this system very quickly	4	4	2	3	4	3	4	
I found the system very cumbersome to use	1	4	4	4	4	4	4	
I felt very confident using the system	2	4	4	3	4	2	4	
I needed to learn a lot of things before I could get going with this system	2	2	4	4	4	4	4	
SUS total score	65	87.5	95	92.5	97.5	72.5	100	87.1 (13.3)

[a]P1-P7: Patients 1-7

[b]SUS: System Usability Scale

Regarding usability measured by the HUBBI, Figure 5 displays the average scores by category on a radar chart. The results indicate that the Healthy Bone digital platform prototype has good usability when considering the Basic System Performance (mean 4.8,SD 0.6), Task-Technology Fit (mean 4.6, SD 0.4), Interface Design (mean 4.5, SD 0.5), Navigation and Structure (mean 4.5, SD 0.7), Information and Terminology (mean 4.7, SD 0.4) and Satisfaction (mean 4.6, SD 0.4) categories. It can be observed that the scores for these categories are all placed within the green border of the radar chart. However, the digital platform prototype's performance in the Guidance and Support

XSL•FO
RenderX

category was only moderate (mean 3.3, SD 1.1) as evidenced by the yellow area of the radar chart.

Patients also reported high levels of task satisfaction, as measured by the ASQ (Table 3). Tasks 6 and 8 showed the highest score for task satisfaction (mean 1, SD 0). On the other hand, task 4 had the lowest task satisfaction (mean 1.8, SD 1.2).

Figure 5. Visualization of the eHealth Usability Benchmarking Instrument scores for the Healthy Bone digital platform prototype. Each color indicates, for the different categories, how the system is performing. Green (>4) denotes good usability, yellow (3-4) moderate usability, orange (2-3) insufficient usability, and red (1-2) poor usability.

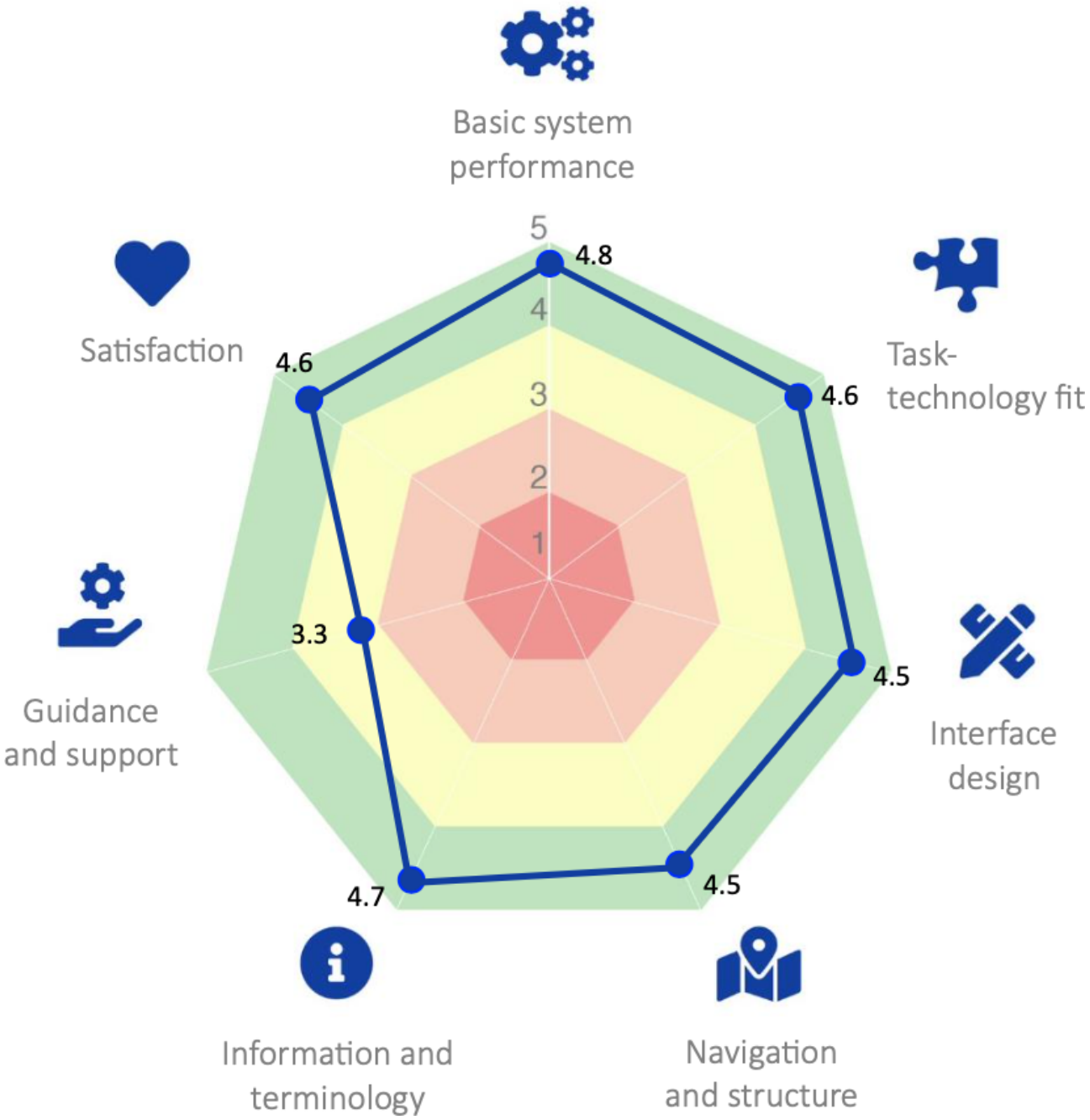

Table . Task satisfaction scores are measured by the After Scenario Questionnaire.

Variables	Task 2	Task 3	Task 4	Task 5	Task 6	Task 7	Task 8
Mean (SD)	1.6 (0.9)	1.4 (0.7)	1.8 (1.2)	1.1 (0.3)	1 (0)	1.2 (0.4)	1(0)
Minimum	1	1	1	1	1	1	1
Maximum	3	2.7	4	1.7	1	2	1

XSL•FO
RenderX

Usability Issues

In total, 48 usability issues were identified across all participants. After removing duplicates, 24 unique usability issues remained. On average, each participant encountered 3.3 (SD 2.5) unique minor usability issues, 2.1 (SD 1.6) unique serious usability issues, and 0.9 (SD 0.4) unique critical issues. The identified issues were classified into 6 usability categories and 10 usability factors. Table 4 describes all the usability issues identified in the analysis, along with the corresponding severity for each issue.

XSL•FO
RenderX

Table . Categorization of usability issues according to the eHealth usability ontology and the respective classification of their severity.

Categories, factors, and usability issues			Severity	Participants affected	First appearance
Basic system performance					
	Technical performance				
		When the user only checked one medication on a day with several medications to be taken, the corresponding day in the calendar turned green, when it should have turned yellow	Minor	P1	P1
		When clicking on the “Return” button inside a questionnaire, the system redirected the user to the “My Program” page instead of the main “Questionnaires” page.	Minor	P4	P4
		After completing a health-related self-reported questionnaire, the platform did not redirect the user to the “Concluded” section of the “Questionnaires” page; instead, it remained as if it had not been answered	Serious	P4; P5	P4
	General system interaction				
		The user tried to enter the platform without inputting the login code	Minor	P1; P2	P1
		The user had difficulty finding the “Pause” button when watching a video	Minor	P1; P6	P1
		The user had difficulty finding the “Return” button when inside a secondary page	Minor	P1	P1
		The user had difficulty distinguishing between clickable and nonclickable elements	Serious	P1; P3; P4	P1
		The user did not know how to move the analog scale within the EQ-5D-3L questionnaire (they were unaware they needed to click, hold, and slide the scale)	Minor	P1	P1
		The user did not know how to conclude a patient-reported outcome measure	Minor	P2	P2

XSL•FO
RenderX

Categories, factors, and usability issues			Severity	Participants affected	First appearance
		The user had difficulty sliding the videos when in list view (“My Program” page)	Minor	P4	P4
		The user had difficulty finding the “Calendar” button in the “Medication” page	Minor	P7	P7
Accessibility					
	Accommodation to perceptual impairments or limitations				
		The user had difficulty seeing some elements of the platform, due to their small size and their vision problems	Minor	P1	P1
Interface design					
	Design clarity				
		The user interpreted the image of a heart used to represent “Other medication” as indicating cardiac medication	Minor	P1; P5	P1
		The user had difficulty noticing if the “Favorite” and “Like/ dislike” buttons were selected or not	Minor	P1	P1
		The user had difficulty finding the “Favorite” and “Like/ dislike” buttons due to their small size	Minor	P4; P7	P4
		The user had difficulty differentiating between weekly questionnaires and health-related self-reported questionnaires in the “Questionnaires” page	Minor	P4	P4
	Symbols, icons, and buttons				
		The user did not know how to switch between list and grid view and which button to use	Serious	P1; P4; P6	P1
Navigation and Structure					
	Navigation				
		The user did not know how to navigate from the “My Program” main page to the “Favorites” page	Minor	P4	P4
	Structure				

XSL•FO
RenderX

Categories, factors, and usability issues			Severity	Participants affected	First appearance
		The user had difficulty understanding the weekly questionnaires component and what they were supposed to do	Minor	P2	P2
		The user did not understand the graphics shown in the "Clinical information" page and their relationship with the health-related self-reported questionnaires	Serious	P2; P3; P5 P6	P2
		The user did not understand the "Participation in the program" section of the "Profile" page	Serious	P3; P4; P5	P3
Information and terminology					
	Health-related information				
		The user did not understand the use of the terms "Quantity", "Frequency" and "Duration" in the "Overview" section of the "Medication" page.	Serious	P1; P4; P6	P1
Guidance and support					
	Error management				
		The user did not know they could change the answers to the weekly questionnaires when an incorrect answer was chosen as the system did not provide this information	Critical	P2; P3; P4 P5; P6; P7	P2
	Procedural health-related information				
		The user did not know how and where to log their medication as taken	Minor	P6	P6

Usability issues were identified within both factors of the Basic system performance usability category. In the Technical performance factor, 3 (12.5%) issues were found, such as, "When clicking on the 'Return' button inside a questionnaire, the system redirected the user to the 'My Program' page instead of the main 'Questionnaires' page." In the General system interaction factor, 8 (33.3%) issues were identified, including problems like "The user had difficulty distinguishing between clickable and non-clickable elements" or "The user did not know how to conclude a patient-reported outcome measure."

In the Accessibility category, only 1 (4.2%) issue was identified and classified within the Accommodation to perceptual impairments or limitations factor: "The user had difficulty seeing some elements of the platform due to their small size and their vision problems." In the Interface Design category, 4 (16.7%) issues were found under the Design clarity factor (eg, "The user had difficulty noticing if the "Favorite" and "Like or Dislike" buttons were selected or not"), and only 1 (4.2%) issue was found in the Symbols, icons, and buttons factor (ie, "The user did not know how to switch between list and grid view and which button to use").

In the Navigation and Structure category, usability issues were identified in both related factors. In the Navigation factor, only 1 (4.2%) issue was found: "The user did not know how to navigate from the 'My Program' main page to the 'Favorites' page." In the Structure factor, 3 (12.5%) issues were identified, such as "The user had difficulty understanding the weekly questionnaires component and what they were supposed to do."

In the Information and terminology category, 1 (4.2%) usability issue was found in the Health-related information factor: "The user did not understand the use of the terms 'Quantity',

XSL•FO
RenderX

'Frequency', and 'Duration' in the 'Overview' section of the 'Medication' page." Finally, usability issues were also identified in the category of Guidance and support. The analysis found 1 (4.2%) issue in the Error management factor (ie, "The user did not know they could change the answers to the weekly questionnaires when an incorrect answer was chosen as the system did not provide this information") and 1 (4.2%) other issue in the Procedural health-related information factor (ie, "The user did not know how and where to log their medication as taken").

Semistructured Interviews

Acceptability and Perceived Usefulness of the Healthy Bone Platform

The patients generally described Healthy Bone as an easy-to-use, accessible, and visually appealing platform. They expressed satisfaction with its simplicity and the way the platform was organized. In addition, they considered that the platform had the potential to empower individuals with osteoporosis, helping them better manage the disease and improve their overall quality of life.

> *I'm pleased with the platform as a whole, the way it is presented, its simplicity in terms of the tasks, the videos… simple and easy to understand. Then… the different pages of the menu, they also seem relatively easy to access, I liked it, yes.* [P4]

> *Because this gives us tips for our daily lives and I think this is important for anyone, right? Especially those who have this health problem [osteoporosis] […] Sometimes we don't have anyone to give us advice on something, and we can come here and see it, right?* [P5]

Despite this, some patients noted that some people may experience difficulties interacting with the platform, and an adaptation period could be necessary. This view was mainly explained by the fact that the platform requires a minimum set of digital skills, which may not be common among many older adults. However, the platform's simplicity and the support from family members were pointed out as potential factors that could help overcome this issue.

> *I think it's simple, but I can't say with 100% certainty that a person who is not as used to, or who only uses a computer once in a while… there's always some degree of difficulty at the beginning […]* [P4]

> *On a technological level, exactly. I think it is necessary to have a minimum level of knowledge on computers to… from a user's perspective because the program is very… it presents itself with an enormous simplicity.* [P7]

Overall, the patients found the different components of the Healthy Bone digital platform acceptable and useful. Patients thought that the available videos discussed important topics related to osteoporosis and general health. They also considered the "Medication" and "Clinical Information" pages important assets as they facilitated better monitoring of medication and health outcomes, respectively. In addition, patients were pleased with the "Messages" section of the Healthy Bone platform. Aspects, such as the possibility of having a closer connection to their respective doctors and being able to communicate with them easily, were regarded as advantages of the platform, transmitting tranquility and greater confidence. The reminders feature was also seen as essential, serving as an important alert for patients, reminding them about upcoming appointments, medication schedules, or even informing them of new video releases.

> *Look, regardless of osteoporosis, I take other medications, and I believe that, when this is ready and if I have access to this platform… that it will be able to help me, right?* [P1]

> *[…] that makes sense because there is a closer connection between us and our doctor, and knowing that at any moment I can communicate… I may not have an immediate answer, but I will have it […] it offers a kind of tranquillity to the patient and a little more confidence […]* [P4]

Barriers to Using the Healthy Bone Platform

Patients reported potential barriers that may hamper the use of the Healthy Bone platform. The main barrier, reported by almost all patients, was a lack of skills and knowledge in using digital technologies. Even though the platform was regarded as simple, easy, and accessible, it was noted that people with limited digital skills could have more difficulty engaging with the system. Some patients reported not being used to using digital technologies, especially computers, and having lost skills over time due to disuse.

> *Well, ok… for me, I honestly have some difficulty. I don't know… there are people of my age with more skills for these technologies, which is not very much my case. I always ran away from computers, and now I am… well, a layperson.* [P1]

> *The lack of basic knowledge […] to work with, with the application. Knowledge from, from a user's perspective […]* [P7]

> *[…] when I was working, I had a lot of contact with… it wasn't these technologies that we have today, obviously. It was much more… things were less advanced, but I worked with computers, all those things. However, from the moment that I stopped working and stayed home with my husband, both of us stayed home, things got lost, got a little lost.* [P6]

The lack of suitable equipment, such as computers and smartphones, was also a barrier to actively using the platform and participating in the program.

> *Well, there are people who prefer their smartphone or a tablet and others to a computer, but really, the biggest barrier will be the lack of equipment.* [P7]

Other barriers were also reported by patients, such as having a busy life and not having the time to access the platform, as using the platform may imply daily access and a higher commitment. Some expressed concerns about not enjoying using digital technologies.

XSL•FO
RenderX

> *Sometimes the lack of time, for instance […] I don't know, people sometimes are so… we're all so busy, right? It's not just some people, it's everyone. And sometimes there may be a lack of time to go… to use the platform […]* [P5]

Facilitators Using the Healthy Bone Platform

While a lack of equipment was identified as a barrier, having access to the necessary equipment and being comfortable with digital technologies were viewed as potential facilitators for using the Healthy Bone platform. Patients also highlighted the importance of expecting potential positive benefits from the program, which could motivate them to engage with the system. In addition, a desire to improve overall health and adopt healthy behaviors was identified as a key factor in promoting regular use of the platform.

> *[…] people have to want to improve their lifestyle, want to improve, want to have better health results […] If the person doesn't want that […] that's the first, in my opinion. The second is, is to have the necessary equipment at your disposal […]* [P7]

> *For many of the things I've already said, because I'm interested in improving my quality of life, in losing some weight, and because I want to get up without pain in my joints, what else? For everything, for everything.* [P7]

> *Well… people also need to be predisposed to do it, right? And be interested in what they're doing, if they're doing it well, if not, sometimes people don't have anyone to advise them, and so they can see on those graphs how they're progressing, how they're doing things (…) The person also needs to be willing to use a little bit of their time to, to interact.* [P5]

Access to important and accurate information about the disease, along with strategies to improve health and promote behavior change, and the ability to visualize their improvements and program progression were also highlighted as facilitators.

> *That thing of having a quick overview, at the moment, of how you are doing, of your evolution in terms of adherence to physical activity, the suggestions for changes in eating behaviors, the progress in terms of using the platform, and your evolution in terms of healthy lifestyle habits.* [P4]

In addition, the ability to access the platform through a smartphone or a tablet app was seen as a facilitating factor. This was primarily because older adults are generally more familiar with smartphones than computers, making them better adapted to smartphones.

> *It depends (…) it depends on each person's level of education and whether or not they are used to their smartphone. Because on a smartphone, it won't be like that. It will be with your finger, right? […] I don't know, I don't know, but… it just might be easier on a smartphone or tablet.* [P3]

This platform was also found to have an important impact on individuals living in geographically remote areas, as it allowed them to receive care despite their location, thus decreasing access inequalities.

> *[…] look, I live 30 or so kilometers away, around 40. If I must call, there are difficulties, difficulties in terms of service provision […] And this, this is the best way to communicate, I see that… I had an aversion to new technologies. I was against them, and now I'm paying dearly for that […] a lot of things can be solved with this type of digital technologies.* [P1]

Discussion

Principal Findings

This study provides a comprehensive assessment of the usability and user experience of the desktop version of the Healthy Bone digital platform prototype from the perspective of participants with osteoporosis. The mixed-methods methodology enabled the collection of valuable information on the platform's usability, as well as users' experiences and perceptions. These findings will be crucial for optimizing the current version of the prototype.

High levels of usability of the Healthy Bone digital platform were found in this study. The mean overall SUS score was 87.1, corresponding to a grade of A+, which indicates excellent usability. High usability was also found when considering the HUBBI individual category scores. The system scored moderately (mean 3.3, SD 1.1) in the Guidance and Support category, indicating that the platform's features related to this category must be improved to enhance usability. This lower score could be explained by specific HUBBI items (eg, items 14 and 15) that focus on error management and procedural system information, namely providing information on how to perform several tasks within the system (eg, creating an account, logging in, and changing settings) [46]. At the time of the usability testing, the Healthy Bone digital platform was considered a minimum viable product. It was at a stage suitable for potential end-users but not yet finalized. Therefore, features such as error notifications were not implemented. Nonetheless, this represents valuable information that indicates that adding these features to the platform could significantly improve its usability. In addition, the Healthy Bone digital platform will require training to ensure patients know how to perform several tasks related to system procedures. This training can be administered by their assisting health care professional or other competent professionals according to where the Healthy Bone platform is first presented to the patient.

The use of both the SUS and the HUBBI represents a methodological strength, as it allows for a more comprehensive assessment of usability and user experience. One common criticism of the SUS is its limited ability to offer detailed insights into which aspects of digital technologies need to be refined or modified. In contrast, the HUBBI, which was developed based on the eHealth ontology of usability issues, addresses this gap by providing scores for each relevant category of usability instead of a single total score [46]. This allows for a more comprehensive assessment of how digital innovation is performing and highlights the aspects that researchers and

XSL•FO
RenderX

designers should focus on during the optimization phase of development.

However, this approach also introduces some trade-offs. Specifically, administering multiple scales can increase the burden on participants, particularly in clinical populations who may already experience cognitive or physical fatigue. In addition, interpreting results across different instruments can add complexity to the analysis, particularly when findings are not fully aligned. These challenges highlight the need for careful planning in future studies to balance depth of insight with participant engagement and analytical clarity.

The qualitative data further supported the quantitative results regarding the usability of the Healthy Bone digital platform. The Think-Aloud approach provided an opportunity to observe how end-users interacted with different components of the system while performing relevant tasks that simulated real-world usage of the digital platform. This process was crucial and helped identify specific usability issues affecting the platform's use, which could be addressed and improved. The analysis of the usability issues was guided by the eHealth ontology [58], which was the basis for developing the HUBBI. This approach enabled a deeper exploration of the HUBBI constructs, providing a better understanding of specific issues that could affect usability within each HUBBI category. While several usability issues were identified, most were classified as minor or serious, indicating that they occurred infrequently or frequently among patients, respectively. However, none of these issues obstructed patients from completing the required task.

No usability issues were found within the categories of Task-technology fit (ie, usability problems that occur because the user does not consider that the system is suitable for meeting their needs or goals) and Satisfaction (ie, the user's satisfaction with the system and their subjective opinion regarding the system's likeability). Some quantitative and qualitative analysis results can explain this absence. Patients reported high levels of task satisfaction, and the qualitative interview results further supported this. Participants consistently expressed satisfaction with the simplicity, accessibility, and utility of the different components of the Healthy Bone digital platform. These positive views suggest high satisfaction levels and the platform effectively aligns with patients' needs. However, one concern that emerged was the level of digital skills required to engage with the system. Some participants indicated that certain older adults may have limited technological proficiency and, therefore, more difficulties using the Healthy Bone digital platform, impacting its acceptability.

The patients also identified other barriers and facilitators to using the Healthy Bone platform. The main reported barriers were a lack of digital skills, unsuitable equipment, and time constraints due to a busy life, which might compromise the platform's adoption. On the other hand, patients highlighted several facilitators that could promote engagement and adoption of the Healthy Bone platform. These included having access to adequate equipment, being motivated to change behaviors, experiencing important health benefits, and decreasing potential inequalities.

The findings regarding limited access to ICTs and the lack of skills to effectively engage with them align with the current literature on the "digital divide." This term describes the gap between individuals with sufficient access to and proficiency using ICTs and those who lack these resources and skills [61,62]. Recent evidence suggests that this digital divide is growing, particularly among older adults, potentially affecting the quality and effectiveness of the health care they receive [62]. Bridging this digital divide and promoting the use of ICTs should be a priority, as it has the potential to contribute to achieving the United Nations' Sustainable Development Goal 10, reducing inequality within and among countries [63,64].

The convergent mixed-methods analysis allowed the identification of clear points of convergence between usability metrics and user experience narratives, as well as specific gaps that qualitative findings helped explain. Generally, the high SUS score of 87.1 was consistent with participants' reports of the platform being a simple, easy-to-learn, and visually appealing tool. These perceptions also align with stronger HUBBI domains related to the overall interface quality and satisfaction of the digital tool. On the other hand, weaker HUBBI domains, such as guidance and support, reflected user concerns about lacking procedural instructions. Importantly, the only critical usability issue (ie, not knowing they could change the answers to the weekly questionnaires when an incorrect answer was chosen as the system did not provide this information) was mapped to this specific domain.

Comparison With Prior Work

Previous studies have also examined barriers and facilitators to the use and adoption of eHealth solutions [65,66]. Interestingly, barriers and facilitators to the use of the Healthy Bone digital platform were in line with some of the most reported factors in the literature. A systematic review by Schreiweis et al [65] identified factors such as poor digital health literacy, lack of the necessary devices, and the added workload as some of the most common barriers to using eHealth services. The review also found that the ease of use, communication improvements, and availability of resources constituted some of the most reported enablers.

The findings of this study will significantly contribute to optimizing the current prototype of the Healthy Bone digital platform. Several strategies will be used, encompassing the correction of technical performance issues, modifications to the platform (eg, increasing the font size, increasing the visibility and clarity of buttons to make them more easily detectable within the interface), and developing a user guide that provides clear explanations on how to interact with the different components of the platform.

This study may also have important implications for developing and implementing eHealth solutions aimed at preventing and managing osteoporosis-related fractures. Specifically, it provides comprehensive information regarding the facilitators, barriers, and usability issues that may be common to several digital solutions for patients with osteoporosis or osteoporosis-related fractures.

XSL•FO
RenderX

This Healthy Bone digital platform includes a multifactorial innovative intervention, allowing it to address health monitoring, including health literacy, and support adherence from a distance in the older adults, a vulnerable group, as we found with the COVID-19 experience. This project intends to add value to the knowledge of how these types of ICT tools can be used in supporting osteoporosis patients.

Limitations

This study has some limitations that must be addressed. The SUS mean score of 87.1 suggests high levels of usability; however, the small sample size may affect the generalizability of the findings. Although evidence suggests that around 80% of usability issues can be found with just five users [48,67,68], larger and more diverse samples are needed to confirm these findings. Although our sample size aligns with usability testing norms, we acknowledge that full thematic saturation may not have been achieved. Two new usability issues were identified in the analysis of the last 2 interview transcripts, and the small sample size limits the certainty regarding the saturation of the acceptability and usefulness, barriers, and facilitators findings. A larger and more diverse sample size could have provided more robust data, and future studies are needed to expand upon these findings. Another limitation of this study concerns the scope and setting of the usability testing. First, it is the fact that usability testing was restricted to the desktop version of the Healthy Bone platform, as the mobile and smart TV versions were still under development. This hampers the generalizability of our findings, particularly considering that older adults may find it preferable to access digital solutions via other devices other than computers. Second, all usability tests were conducted in a controlled, research-led environment, which may have increased perceived usability. Future research on this project should include usability testing of the other versions of the digital platform, ensuring that the platform is well adapted across all devices, and include home-based, independent testing to capture a more comprehensive overview of usability in real-world conditions.

Furthermore, patients who regularly use smartphones and health management apps may have been more inclined to participate, which could introduce a bias toward higher eHealth literacy, acceptability, and satisfaction with the digital platform.

Conclusions

This study has provided valuable insights into the usability and user experience of the desktop version of the Healthy Bone digital platform prototype. Specific usability issues and users' perceptions of this digital solution were identified. The next steps will involve evaluating the usability and user experience of the mobile app version of Healthy Bone, followed by a pilot study and a randomized controlled trial. Data gathered from the usability and user experience assessments will ensure that the digital solution is tailored to the needs of the target population. This knowledge can lead to a more successful implementation and maximize the overall impact of the Healthy Bone digital platform.

Acknowledgments

We thank the patients for their generosity in taking part in the study. Without their contributions, this study would not have been possible. This work was funded by the Fundação para a Ciência e a Tecnologia, IP national support, through the Comprehensive Health Research Centre (UIDP/04923/2020).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Semistructured interview guide.
[PDF File, 542 KB - formative_v9i1e72468_app1.pdf]

Checklist 1
GRAMMS checklist.
[PDF File, 193 KB - formative_v9i1e72468_app2.pdf]

References

1. Johnell O, Kanis JA. An estimate of the worldwide prevalence and disability associated with osteoporotic fractures. Osteoporos Int 2006 Dec;17(12):1726-1733. [doi: 10.1007/s00198-006-0172-4] [Medline: 16983459]
2. Barcelos A, Gonçalves J, Mateus C, Canhão H, Rodrigues AM. Costs of hip fractures in postmenopausal women in Portugal: a study from the payer's perspective. Acta Med Port 2023 Dec 4;36(12):848-849. [doi: 10.20344/amp.20481] [Medline: 38048690]
3. Barcelos A, Gonçalves J, Mateus C, Canhão H, Rodrigues AM. Costs of incident non-hip osteoporosis-related fractures in postmenopausal women from a payer perspective. Osteoporos Int 2023 Dec;34(12):2111-2119. [doi: 10.1007/s00198-023-06881-w] [Medline: 37596433]
4. Willers C, Norton N, Harvey NC, et al. Osteoporosis in Europe: a compendium of country-specific reports. Arch Osteoporos 2022 Jan 26;17(1):23. [doi: 10.1007/s11657-021-00969-8] [Medline: 35079919]

XSL•FO
RenderX

5. Borgström F, Karlsson L, Ortsäter G, et al. Fragility fractures in Europe: burden, management and opportunities. Arch Osteoporos 2020 Apr 19;15(1):59. [doi: 10.1007/s11657-020-0706-y] [Medline: 32306163]
6. Johnell O, Cooper C, Cummings S, Slemenda C, Seeman E. The socioeconomic burden of fractures: today and in the 21st century. Am J Med 1997 Aug 18;103(2A):20S-25S. [doi: 10.1016/s0002-9343(97)90023-1] [Medline: 9302894]
7. Gregson CL, Armstrong DJ, Bowden J, et al. UK clinical guideline for the prevention and treatment of osteoporosis. Arch Osteoporos 2022 Apr 5;17(1):58. [doi: 10.1007/s11657-022-01061-5] [Medline: 35378630]
8. LeBoff MS, Greenspan SL, Insogna KL, et al. The clinician's guide to prevention and treatment of osteoporosis. Osteoporos Int 2022 Oct;33(10):2049-2102. [doi: 10.1007/s00198-021-05900-y] [Medline: 35478046]
9. Rodrigues AM, Canhão H, Marques A, et al. Portuguese recommendations for the prevention, diagnosis and management of primary osteoporosis - 2018 update. Acta Reumatol Port 2018;43(1):10-31. [Medline: 29602163]
10. Cehic M, Lerner RG, Achten J, Griffin XL, Prieto-Alhambra D, Costa ML. Prescribing and adherence to bone protection medications following hip fracture in the United Kingdom: results from the World Hip Trauma Evaluation (WHiTE) cohort study. Bone Joint J 2019 Nov;101-B(11):1402-1407. [doi: 10.1302/0301-620X.101B11.BJJ-2019-0387.R1] [Medline: 31674239]
11. Imaz I, Zegarra P, González-Enríquez J, Rubio B, Alcazar R, Amate JM. Poor bisphosphonate adherence for treatment of osteoporosis increases fracture risk: systematic review and meta-analysis. Osteoporos Int 2010 Nov;21(11):1943-1951. [doi: 10.1007/s00198-009-1134-4] [Medline: 19967338]
12. Albrecht BM, Stalling I, Foettinger L, Recke C, Bammann K. Adherence to lifestyle recommendations for bone health in older adults with and without osteoporosis: cross-sectional results of the OUTDOOR ACTIVE study. Nutrients 2022 Jun 14;14(12):2463. [doi: 10.3390/nu14122463] [Medline: 35745193]
13. Kothawala P, Badamgarav E, Ryu S, Miller RM, Halbert RJ. Systematic review and meta-analysis of real-world adherence to drug therapy for osteoporosis. Mayo Clin Proc 2007 Dec;82(12):1493-1501. [doi: 10.1016/S0025-6196(11)61093-8] [Medline: 18053457]
14. Fardellone P, Lello S, Cano A, et al. Real-world adherence and persistence with bisphosphonate therapy in postmenopausal women: a systematic review. Clin Ther 2019 Aug;41(8):1576-1588. [doi: 10.1016/j.clinthera.2019.05.001] [Medline: 31151814]
15. Jensen CM, Overgaard S, Wiil UK, Smith AC, Clemensen J. Bridging the gap: a user-driven study on new ways to support self-care and empowerment for patients with hip fracture. SAGE Open Med 2018;6:2050312118799121. [doi: 10.1177/2050312118799121] [Medline: 30210796]
16. Eysenbach G. What is e-health? J Med Internet Res 2001;3(2):E20. [doi: 10.2196/jmir.3.2.e20] [Medline: 11720962]
17. Ossebaard HC, Van Gemert-Pijnen L. eHealth and quality in health care: implementation time. Int J Qual Health Care 2016 Jun;28(3):415-419. [doi: 10.1093/intqhc/mzw032] [Medline: 27029590]
18. Wernhart A, Gahbauer S, Haluza D. eHealth and telemedicine: practices and beliefs among healthcare professionals and medical students at a medical university. PLoS ONE 2019;14(2):e0213067. [doi: 10.1371/journal.pone.0213067] [Medline: 30818348]
19. Kampmeijer R, Pavlova M, Tambor M, Golinowska S, Groot W. The use of e-health and m-health tools in health promotion and primary prevention among older adults: a systematic literature review. BMC Health Serv Res 2016 Sep 5;16 Suppl 5(Suppl 5):290. [doi: 10.1186/s12913-016-1522-3] [Medline: 27608677]
20. Behaviour change: digital and mobile health interventions NICE guideline. NICE. 2020. URL: www.nice.org.uk/guidance/ng183 [accessed 2025-10-31]
21. Mauco KL, Scott RE, Mars M. Validation of an e-health readiness assessment framework for developing countries. BMC Health Serv Res 2020 Jun 23;20(1):575. [doi: 10.1186/s12913-020-05448-3] [Medline: 32576174]
22. Greenhalgh T, Shaw S, Wherton J, et al. SCALS: a fourth-generation study of assisted living technologies in their organisational, social, political and policy context. BMJ Open 2016 Feb 15;6(2):e010208. [doi: 10.1136/bmjopen-2015-010208] [Medline: 26880671]
23. Greenhalgh T, Papoutsi C. Spreading and scaling up innovation and improvement. BMJ 2019 May 10;365:l2068. [doi: 10.1136/bmj.l2068] [Medline: 31076440]
24. Pérez-Jover V, Sala-González M, Guilabert M, Mira JJ. Mobile apps for increasing treatment adherence: systematic review. J Med Internet Res 2019 Jun 18;21(6):e12505. [doi: 10.2196/12505] [Medline: 31215517]
25. Mira JJ, Guilabert M, Carrillo I, et al. Use of QR and EAN-13 codes by older patients taking multiple medications for a safer use of medication. Int J Med Inform 2015 Jun;84(6):406-412. [doi: 10.1016/j.ijmedinf.2015.02.001] [Medline: 25731975]
26. Park JYE, Li J, Howren A, Tsao NW, De Vera M. Mobile phone apps targeting medication adherence: quality assessment and content analysis of user reviews. JMIR Mhealth Uhealth 2019 Jan 31;7(1):e11919. [doi: 10.2196/11919] [Medline: 30702435]
27. Ravn Jakobsen P, Hermann AP, Søndergaard J, Wiil UK, Clemensen J. Development of an mHealth Application for Women Newly Diagnosed with Osteoporosis without Preceding Fractures: A Participatory Design Approach. Int J Environ Res Public Health 2018 Feb 13;15(2):330. [doi: 10.3390/ijerph15020330] [Medline: 29438343]

XSL•FO
RenderX

28. Peeters JM, Wiegers TA, Friele RD. How technology in care at home affects patient self-care and self-management: a scoping review. Int J Environ Res Public Health 2013 Oct 29;10(11):5541-5564. [doi: 10.3390/ijerph10115541] [Medline: 24173139]
29. Noergaard B, Sandvei M, Rottmann N, et al. Development of a web-based health care intervention for patients with heart disease: lessons learned from a participatory design study. JMIR Res Protoc 2017 May 17;6(5):e75. [doi: 10.2196/resprot.7084] [Medline: 28526674]
30. Ahmed I, Ahmad NS, Ali S, et al. Medication adherence apps: review and content analysis. JMIR Mhealth Uhealth 2018 Mar 16;6(3):e62. [doi: 10.2196/mhealth.6432] [Medline: 29549075]
31. Gomes LA, Gregório MJ, Iakovleva TA, et al. A home-based eHealth intervention for an older adult population with food insecurity: feasibility and acceptability study. J Med Internet Res 2021 Aug 31;23(8):e26871. [doi: 10.2196/26871] [Medline: 34463638]
32. Slomian J, Appelboom G, Ethgen O, Reginster JY, Bruyère O. Can new information and communication technologies help in the management of osteoporosis? Womens Health (Lond) 2014 May;10(3):229-232. [doi: 10.2217/whe.14.15] [Medline: 24956289]
33. Alhussein G, Hadjileontiadis L. Digital health technologies for long-term self-management of osteoporosis: systematic review and meta-analysis. JMIR Mhealth Uhealth 2022 Apr 21;10(4):e32557. [doi: 10.2196/32557] [Medline: 35451968]
34. Oldenburg B, Taylor CB, O'Neil A, Cocker F, Cameron LD. Using new technologies to improve the prevention and management of chronic conditions in populations. Annu Rev Public Health 2015 Mar 18;36:483-505. [doi: 10.1146/annurev-publhealth-031914-122848] [Medline: 25581147]
35. Nielsen D, Ryg J, Nielsen W, Knold B, Nissen N, Brixen K. Patient education in groups increases knowledge of osteoporosis and adherence to treatment: a two-year randomized controlled trial. Patient Educ Couns 2010 Nov;81(2):155-160. [doi: 10.1016/j.pec.2010.03.010] [Medline: 20400258]
36. Billington EO, Feasel AL, Kline GA. At odds about the odds: women's choices to accept osteoporosis medications do not closely agree with physician-set treatment thresholds. J Gen Intern Med 2020 Jan;35(1):276-282. [doi: 10.1007/s11606-019-05384-x] [Medline: 31625042]
37. Tüzün Ş, Akyüz G, Eskiyurt N, et al. Impact of the training on the compliance and persistence of weekly bisphosphonate treatment in postmenopausal osteoporosis: a randomized controlled study. Int J Med Sci 2013;10(13):1880-1887. [doi: 10.7150/ijms.5359] [Medline: 24324365]
38. Bianchi ML, Duca P, Vai S, et al. Improving adherence to and persistence with oral therapy of osteoporosis. Osteoporos Int 2015 May;26(5):1629-1638. [doi: 10.1007/s00198-015-3038-9] [Medline: 25619634]
39. Grahn Kronhed AC, Enthoven P, Spångeus A, Willerton C. Mindfulness and modified medical yoga as intervention in older women with osteoporotic vertebral fracture. J Altern Complement Med 2020 Jul;26(7):610-619. [doi: 10.1089/acm.2019.0450] [Medline: 32543212]
40. Smulders E, Weerdesteyn V, Groen BE, et al. Efficacy of a short multidisciplinary falls prevention program for elderly persons with osteoporosis and a fall history: a randomized controlled trial. Arch Phys Med Rehabil 2010 Nov;91(11):1705-1711. [doi: 10.1016/j.apmr.2010.08.004] [Medline: 21044715]
41. Bendtsen MG, Schönwandt BMT, Rubæk M, Hitz MF. Evaluation of an mHealth app on self-management of osteoporosis: prospective survey study. Interact J Med Res 2024 Apr 1;13:e53995. [doi: 10.2196/53995] [Medline: 38557362]
42. Pham Q, Wiljer D, Cafazzo JA. Beyond the randomized controlled trial: a review of alternatives in mHealth clinical trial methods. JMIR Mhealth Uhealth 2016 Sep 9;4(3):e107. [doi: 10.2196/mhealth.5720] [Medline: 27613084]
43. Bernardi HLF, Motta LBD. Development of an app as a tool to support research and the prevention of osteoporosis. Rev bras geriatr gerontol 2018 Aug;21(4):408-418. [doi: 10.1590/1981-22562018021.170189]
44. Rodrigues AM, Gregório MJ, Gein P, et al. Home-based intervention program to reduce food insecurity in elderly populations using a TV app: study protocol of the randomized controlled trial saúde.come senior. JMIR Res Protoc 2017 Mar 13;6(3):e40. [doi: 10.2196/resprot.6626] [Medline: 28288956]
45. Ergonomics of human-system interaction — part 11: usability: definitions and concepts. : International Organization for Standardization (ISO); 2018.
46. Broekhuis M, van Velsen L. Improving usability benchmarking for the eHealth domain: The development of the eHealth UsaBility Benchmarking instrument (HUBBI). PLoS ONE 2022;17(2):e0262036. [doi: 10.1371/journal.pone.0262036] [Medline: 35176033]
47. O'Cathain A, Murphy E, Nicholl J. The quality of mixed methods studies in health services research. J Health Serv Res Policy 2008 Apr;13(2):92-98. [doi: 10.1258/jhsrp.2007.007074] [Medline: 18416914]
48. Virzi RA. Refining the test phase of usability evaluation: how many subjects is enough? Hum Factors 1992 Aug;34(4):457-468. [doi: 10.1177/001872089203400407]
49. Santos AR, Reis P, Rebelo LF, Dias FC, Rosa MC, Queiroz M. Health assessment questionnaire" (versão curta): adaptação para língua portuguesa e estudo da sua aplicabilidade [Article in Portuguese]. Acta Reumatol Port 1996(76):15-20 [FREE Full text]

XSL•FO
RenderX

50. Rubenstein LZ, Harker JO, Salvà A, Guigoz Y, Vellas B. Screening for undernutrition in geriatric practice: developing the short-form mini-nutritional assessment (MNA-SF). J Gerontol A Biol Sci Med Sci 2001 Jun;56(6):M366-M372. [doi: 10.1093/gerona/56.6.m366] [Medline: 11382797]
51. Craig CL, Marshall AL, Sjöström M, et al. International physical activity questionnaire: 12-country reliability and validity. Med Sci Sports Exerc 2003 Aug;35(8):1381-1395. [doi: 10.1249/01.MSS.0000078924.61453.FB] [Medline: 12900694]
52. Ferreira PL, Ferreira LN, Pereira LN. Contribution for the validation of the Portuguese version of EQ-5D. Acta Med Port 2013;26(6):664-675. [doi: 10.20344/amp.1317] [Medline: 24388252]
53. Jaspers MWM. A comparison of usability methods for testing interactive health technologies: Methodological aspects and empirical evidence. Int J Med Inform 2009 May;78(5):340-353. [doi: 10.1016/j.ijmedinf.2008.10.002]
54. Lewis JR. Psychometric evaluation of an after-scenario questionnaire for computer usability studies. SIGCHI Bull 1991 Jan;23(1):78-81. [doi: 10.1145/122672.122692]
55. Martins AI, Rosa AF, Queirós A, Silva A, Rocha NP. European Portuguese validation of the System Usability Scale (SUS). Procedia Comput Sci 2015;67:293-300. [doi: 10.1016/j.procs.2015.09.273]
56. Sauro J, Lewis JR. Chapter 8 - standardized usability questionnaires. In: Sauro J, editor. Quantifying the User Experience, 2nd edition: Morgan Kaufmann; 2016:185-248. [doi: 10.1016/B978-0-12-802308-2.00008-4]
57. Brooke J. SUS: a quick and dirty usability scale. In: Usability Evaluation in Industry: Taylor & Francis; 1996:189-194.
58. Broekhuis M, van Velsen L, Peute L, Halim M, Hermens H. Conceptualizing usability for the eHealth context: content analysis of usability problems of eHealth applications. JMIR Form Res 2021 Jul 27;5(7):e18198. [doi: 10.2196/18198] [Medline: 34313594]
59. Duh HL, Tan GCB, Chen VH. Usability evaluation for mobile device: a comparison of laboratory and field tests. Presented at: MobileHCI06: Human-Computer Interaction with Mobile Devices and Services; Sep 12-15, 2006. [doi: 10.1145/1152215.1152254]
60. Braun V, Clarke V. One size fits all? What counts as quality practice in (reflexive) thematic analysis? Qual Res Psychol 2021 Jul 3;18(3):328-352. [doi: 10.1080/14780887.2020.1769238]
61. Soomro KA, Kale U, Curtis R, Akcaoglu M, Bernstein M. Digital divide among higher education faculty. Int J Educ Technol High Educ 2020;17(1):21. [doi: 10.1186/s41239-020-00191-5] [Medline: 41001069]
62. Mubarak F, Suomi R. Elderly forgotten? Digital exclusion in the information age and the rising grey digital divide. INQUIRY 2022;59:469580221096272. [doi: 10.1177/00469580221096272] [Medline: 35471138]
63. Ghosh P, Chakravarty T. Sustainable development goals: challenges, opportunities, and the way forward. In: Dutta V, Ghosh P, editors. Sustainability: Science, Policy, and Practice in India: Challenges and Opportunities: Springer International Publishing; 2023:235-241. [doi: 10.1007/978-3-031-50132-6_17]
64. Global strategy on digital health 2020-2025. World Health Organization. 2021. URL: https://www.who.int/publications/i/item/9789240020924 [accessed 2024-07-30]
65. Schreiweis B, Pobiruchin M, Strotbaum V, Suleder J, Wiesner M, Bergh B. Barriers and facilitators to the implementation of eHealth services: systematic literature analysis. J Med Internet Res 2019 Nov 22;21(11):e14197. [doi: 10.2196/14197] [Medline: 31755869]
66. Wilson J, Heinsch M, Betts D, Booth D, Kay-Lambkin F. Barriers and facilitators to the use of e-health by older adults: a scoping review. BMC Public Health 2021 Aug 17;21(1):1556. [doi: 10.1186/s12889-021-11623-w] [Medline: 34399716]
67. NIELSEN J. Chapter 6 - usability testing. In: NIELSEN J, editor. Usability Engineering: Morgan Kaufmann; 1993:165-206. [doi: 10.1016/B978-0-08-052029-2.50009-7]
68. Faulkner L. Beyond the five-user assumption: benefits of increased sample sizes in usability testing. Behav Res Methods Instrum Comput 2003 Aug;35(3):379-383. [doi: 10.3758/bf03195514] [Medline: 14587545]

Abbreviations

API: Application Programming Interface
ASQ: After Scenario Questionnaire
CMS: Content Management System
EQ-5D-3L: EuroQol, 5 dimensions, 3 levels
GRAMMS: Good Reporting of a Mixed Methods Study
HAQ: Health Assessment Questionnaire
HUBBI: eHealth Usability Benchmarking Instrument
ICT: information and communication technology
SUS: System Usability Scale

XSL•FO
RenderX

Edited by J Sarvestan; submitted 19.02.25; peer-reviewed by W Glinkowski; revised version received 10.09.25; accepted 30.09.25; published 07.11.25.

Please cite as:
Barcelos A, Moniz A, Henriques AR, Mazeda C, Frazão Mateus E, Machado A, F Gonçalves C, Moreto M, F Almeida M, Mendonça N, Costa D, M Marques M, Domingos P, Nesme Y, Canhão H, Dias S, Rodrigues AM
Usability and User Experience of a Digital Platform Prototype (Healthy Bone) to Promote Pharmacological and Nonpharmacological Treatment in Patients With Osteoporosis: Mixed Methods Study
JMIR Form Res 2025;9:e72468
URL: https://formative.jmir.org/2025/1/e72468
doi:10.2196/72468

© Anabela Barcelos, Alexandre Moniz, Ana Rita Henriques, Carolina Mazeda, Elsa Frazão Mateus, Ana Machado, Cátia F Gonçalves, Maiara Moreto, Mafalda F Almeida, Nuno Mendonça, Daniela Costa, Marta M Marques, Pedro Domingos, Yoann Nesme, Helena Canhão, Sónia Dias, Ana Maria Rodrigues. Originally published in JMIR Formative Research (https://formative.jmir.org), 7.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Evaluation of a Digital, Self-Administered, Cognitive Test Battery in Older Adult Patients Undergoing Abdominal Surgery: Nonrandomized Feasibility Trial

Anahita Amirpour[1], MSc; Markus Saarijärvi[2,3], PhD; Jeanette Eckerblad[1], PhD; Gabriela Markovic[4], PhD; Anders Thorell[5], PhD; Ulrica Nilsson[1], PhD; Lina Bergman[1], PhD

[1]Department of Neurobiology, Care Sciences and Society, Karolinska Institutet, Alfred Nobels Alle 23, C4, Huddinge, Sweden

[2]Department of Cardiology, Danderyd University Hospital, Stockholm, Sweden

[3]Department of Neurobiology, Care Sciences and Society, Karolinska Institutet, Huddinge, Sweden

[4]Department of Rehabilitation Medicine, Department of Clinical Sciences, Danderyd University Hospital, Karolinska Institutet, Danderyd University Hospital, Stockholm, Sweden

[5]Department of Surgery and Anesthesiology, Department of Clinical Sciences, Ersta Hospital, Karolinska Institutet, Danderyd University Hospital, Stockholm, Sweden

Corresponding Author:
Anahita Amirpour, MSc
Department of Neurobiology, Care Sciences and Society, Karolinska Institutet, Alfred Nobels Alle 23, C4, Huddinge, Sweden

Abstract

Background: Older adults undergoing surgeries face increased risks of postoperative neurocognitive disorders, which impair cognitive functions. Analog neurocognitive tests are commonly used, but digital tests offer faster, more accessible assessments.

Objective: The primary aim of this study was to evaluate the feasibility of a digital cognitive test battery in older adults undergoing abdominal surgery. Feasibility included estimation of recruitment and retention rates, acceptability, perceived value, and usability of the test. The secondary aim was to explore outcome trajectories of cognition, depression, functional status, and quality of recovery.

Methods: This nonrandomized feasibility study measured recruitment and retention rates using patient logs and expanded on these findings in semistructured interviews with nurses. Acceptability, perceived value, and usability were explored through interviews with patients and nurses, and the System Usability Scale (SUS). Cognitive functions were assessed with a digital cognitive test battery (Consortium to Establish a Registry for Alzheimer Disease [CERAD] word list learning test, Trail Making Test Parts A and B, Victoria Stroop Test, and Symbol Digit Pairing Test) and the Nursing Delirium Screening scale (NU-DESC), and depression with the Geriatric Depression Scale (GDS-15). Functional status was measured using the World Health Organization Disability Assessment Schedule (WHODAS), and postoperative recovery with the Swedish Quality of Recovery questionnaire (SwQoR-24). Quantitative data were analyzed using descriptive statistics and nonparametric tests and qualitative data with content analysis.

Results: The test battery was feasible, acceptable, and demonstrated excellent usability. The mean SUS score was 87 (SD 17.9; 95% CI 78.9 - 95.2), and all predefined progression criteria were met. Recruitment spanned over 1.5 years, during which 24 patients were included (mean age of 77, SD 6.5 years; range: 63 - 90 years; n=13, 54% women). Most patients underwent laparoscopic colorectal cancer surgery. Three patients developed postoperative delirium for 1 day only. No patient developed delayed neurocognitive recovery or mild/major neurocognitive disorder at the postoperative follow-up. Qualitative data showed that both nurses and patients regarded the digital cognitive test battery as important for assessing cognitive function and found it easy to use and understand. Nurses reported that recruitment was challenging, partly because not all patients attended a preoperative in-person consultation before surgery.

Conclusions: The digital, self-administered cognitive test battery was found to be feasible, acceptable, and usable in older adults undergoing abdominal surgery. However, recruitment challenges and a small, homogeneous sample limit generalizability and warrant careful consideration in a larger-scale study.

Trial Registration: ClinicalTrials.gov NCT05564195; https://clinicaltrials.gov/study/NCT05564195

International Registered Report Identifier (IRRID): RR2-10.1136/bmjopen-2022-070404

(JMIR Form Res 2025;9:e71911) doi:10.2196/71911

XSL•FO
RenderX

KEYWORDS

colorectal surgery; feasibility studies; neurocognitive disorders; neuropsychological tests; postoperative cognitive complications; user-centered design

Introduction

Undergoing surgery as an older adult comes with its own set of challenges and risks. Older adults often experience a higher symptom burden compared to younger patients, and the aging process affects various biological and functional systems [1], including the brain and cognitive functions. Postoperative complications are a significant concern for older adults, with cognitive complications being particularly common. These include postoperative delirium (POD), delayed neurocognitive recovery (dNCR), and postoperative neurocognitive disorder (p-NCD) [2]. POD can occur within the first week after surgery, while dNCR may develop within 30 days. The effects of p-NCD, whether mild or major, can persist up to a year. These conditions can impact cognitive functions, leading to impairments in attention, memory, and executive functions [3], and manifest themselves as forgetfulness, difficulty initiating tasks, and challenges in executing daily activities [4].

Various methods exist for assessing cognitive performance following surgery, with analog tests being the most used [5]. Recommended tests typically focus on episodic memory, executive functions, and attention [6]. Digital tests offer faster screening and analysis of cognitive functions, reduction of administrative bias, and enhanced accessibility as they can be performed outside of clinical settings [7]. Despite its advantages over analog tests, digital tests remain underutilized in perioperative care.

One such digital test, Mindmore, provides self-administered cognitive test batteries grounded in established psychometric tests, including normative data adjusted for sex, age, and education [8]. The digital cognitive test batteries are validated, Conformité Européenne–certified, with demonstrated high usability among older adult volunteers [7] and are developed by licensed psychologists at a private company [7,8]. Research from other settings indicates that digital and remote assessments are generally feasible and acceptable among older adults [9-11], including smartphone-based monitoring [9], tablet tools for postoperative mobilization [10], and telephone-based cognitive testing [11]. However, no study has yet examined the feasibility of a digital, cognitive test with older adults undergoing surgery.

Thus, the primary aim of this study was to evaluate the feasibility of using a digital cognitive test battery in older adults undergoing abdominal surgery. Feasibility included estimation of recruitment and retention rates, acceptability, perceived value, and usability of the test. The secondary aim was to explore outcome trajectories of cognition, depression, functional status, and quality of recovery.

Methods

Overview

We followed the medical research council's framework for evaluating complex interventions [12] and prospectively registered our study at clinicaltrials.gov (NCT05564195) and published our study protocol [13]. We report in accordance with reporting guidelines for nonrandomized feasibility studies [14] and relevant items from the CONSORT (Consolidated Standards of Reporting Trials) 2010 statement extension for pilot and feasibility trials [15].

Study Design and Setting

We conducted a nonrandomized single-group feasibility study with the purpose of assessing key components for a future, larger cohort study [14]. The study was carried out from January 2023 to November 2024 in Stockholm County, Sweden. The hospital where the patients were recruited specializes in abdominal surgery and handles approximately 4700 outpatient visits annually.

Study Participants and Procedure

Registered nurses at the surgery clinic and research unit screened and identified potential participants who met the inclusion criteria at the preoperative visit, provided study information, and obtained written informed consent. Inclusion criteria were patients ≥60 years undergoing planned surgery lasting a minimum of 60 minutes. Exclusion criteria included Mini-Mental State Examination (MMSE) score <23, current psychiatric disease (eg, bipolar disorder), central nervous system disease (eg, stroke), uncorrected hearing or visual impairments, motor impairment in the dominant hand, defective color vision, and lack of proficiency in the Swedish language. A target sample size of 50 participants was planned, based on a recommended minimum of 30 participants for pilot studies to provide initial estimates of outcome variability and feasibility for future trials [16]. A Swedish version of the MMSE was administered at the beginning of the study before obtaining written permission from Psychological Assessment Resources (PAR). This administrative oversight was later corrected, and the required authorization has since been secured. The MMSE is a copyrighted instrument and may not be reproduced or used in any form or language without prior written permission from PAR.

Data Collection

Quantitative data investigated recruitment and retention process, usability, cognitive performance, including clinician and patient-reported outcome measures. We collected data at baseline (T0) and postoperatively including follow-ups at the hospital 1 - 3 days (T1), 3 - 5 weeks (T2), and at 6 months (T3) at the patient's home or at the premises of Karolinska Institutet. All the outcome measurements included are presented in Table 1. As some instruments lacked validation in older adults in a perioperative context, their use in this study was exploratory.

XSL•FO
RenderX

Table . Included clinician-reported and patient-reported outcome measurements.

Questionnaire	Outcome	Properties	Timepoint	Validation in Swedish context
Clinical Frailty Scale (CFS) [17]	Frailty	9-item, ranging from 1-very fit to 9-terminally ill. Very fit means people who are robust, active, energetic, and motivated.	T0[a]	Validated in patients 65 years and older [18].
Nursing Delirium Screening Scale (NU-DESC) [19]	Delirium	5-item, observational scale where each symptom is rated 0 - 2. A score of ≥2=delirium.	Each postoperative day at the hospital	Validated in older adult patients undergoing cardiac surgery [20].
Swedish Quality of Recovery (SwQoR-24) [21]	Postoperative quality of recovery	24-item, including rating pain, anxiety, and sleep difficulties. The items are rated on an 11-point Likert scale ranging from 0=none of the time to 10=all the time. The global score ranges from 0 - 240 where scores ≥32 indicate poor recovery.	T0 T2[b] T3[c]	Validated in patients undergoing surgery [22].
Geriatric Depression Scale (GDS-15) [23]	Depression	15-item consists of binary (yes/no) questions on depressive symptoms. Scores at ≥5 indicate depression.	T0 T2 T3	Validated in older adult patients [24], .
System Usability Scale (SUS) [25]	Usability	10-item, on a five-point Likert scale. The score ranges from 0 - 100, where a higher score indicates high usability.	T2	Not validated. We used a Swedish translated version and slightly modified the statements to “tablet” instead of “system” to enhance participant comprehension.
World Health Organization Disability Assessment Schedule (WHODAS 2.0, 12-item)	Function	12-item, ranging from 1=none to 5=extreme, reflecting the degree of difficulty experienced in various functional areas. The score ranges from 12 to 60, with a score of 12 indicating “none” for all items.	T0 T2 T3	Validated in patients with psychotic disorders [26].

[a]T0: baseline, preoperatively.

[b]T2: 3-5 weeks postoperatively.

[c]T3: 6 months postoperatively.

Qualitative data explored the acceptability, usability, and perceived value of the digital cognitive test battery by collecting data from 2 sources:

- Face-to-face or video interviews via Zoom (Zoom Video Communications) with 17 patients at the patient’s home. Patients chose their preferred mode of interview. Data collection took place after T2, except for one patient, whose interview was conducted at 5 months after surgery.
- Video interviews via Microsoft Teams (Microsoft Corporation) or Zoom with 5 nurses at the research unit and surgery clinic, recruiting patients and delivering the digital cognitive test battery. Data collection took place after T2.

The first author (AA) conducted the interviews with the patients, and the last author (LB) conducted the interviews with the nurses. We developed and followed semistructured interview guides (Multimedia Appendix 1) for all interviews, which contained open-ended questions and suggestions for additional probing questions. The interview guide for the nurses was based on Sekhon et al’s [27] framework for acceptability, which includes 7 constructs: affective attitudes, burden, ethical aspects, intervention coherence, opportunity costs, perceived effectiveness, and self-efficacy. After the initial interviews, the guide was discussed and evaluated within the research group. Each interview was audio-recorded and transcribed verbatim by a transcription company.

Cognitive Test Battery

Mindmore offers different cognitive test batteries depending on the patient group and procedure [7]. The tests included in this study are listed in Table 2. Each test begins with an introduction, followed by a practice trial with feedback and then the actual test trial. The 4 tests take approximately 15 minutes to complete and are conducted on a 10.1-inch Windows touchscreen study-specific tablet. The capacitive touchscreen

XSL•FO
RenderX

records detailed information such as timing, pauses, and finger lifts from the screen. Further, speech recognition is used for screening verbal memory. The patients underwent the cognitive test battery at T0, T1, T2, and T3.

Table . Included cognitive test battery.

Cognitive test	Cognitive domains	Properties
Consortium to Establish a Registry for Alzheimer Disease (CERAD) word list learning test	Verbal episodic memory, delayed recall	10-word verbal learning test conducted over three trials, followed by recall around 7 minutes later. For the word list learning trial, the score ranges from 0 - 30, where a higher score indicates a higher function. For delayed recall, the maximum score is 10.
Trail Making Test (TMT A and B)	Attention, processing speed, visuospatial function, executive functions	25 circles in each segment, requiring the participants to connect them sequentially. In Part B, the participants must connect 13 numbers and 12 letters alternately. A higher number of connections and shorter time indicate higher performance.
Victoria Stroop Test	Executive functions, selective attention, cognitive flexibility, inhibition response	24 words depicting different colors, printed using contrasting ink colors. The participant is directed to quickly and thoroughly identify the color of the ink rather than the word itself. Test score is calculated as an index: meaning number of correct answers and average time, and for interference.
Symbol Digits Processing Test (SDPT)	Attention, processing speed	9 randomized symbols to reduce practice effects. The participant matches the symbols to digits on a 3x3grid within 90 seconds, with the final score based on correct matches.

Outcomes

An overview of the feasibility outcomes and progression criteria [28] is presented in Table 3. Our progression criteria (eg, warranting a recommendation to proceed to a larger trial) were based on our previous research with patients undergoing total hip arthroplasty [29] and previous studies including patients undergoing colorectal surgery [30,31]; consequently, we set the following progression criteria: recruitment ≥20% of available patients, dropouts ≤30%, and usability ≥75 on the System Usability Scale.

Table . Feasibility outcome measures.

Outcome	Description	Progression criteria	Data source
Study design, recruitment, and retention	Estimates, reasons for nonparticipation and dropouts	≥20% of available patients; dropouts <30%	Patient screening log; semistructured interviews
Acceptability and perceived value of the digital cognitive test battery, by patients and nurses	Experiences of using the digital cognitive test battery	__[a]	Semistructured interviews
Usability of the digital cognitive test battery, by patients and nurses	System Usability Scale; experiences of using the digital cognitive test battery	System Usability Scale Score ≥75	System Usability Scale; semistructured interviews

[a]Not applicable.

Statistical Analysis

Descriptive statistics included numbers and percentages for categorical variables, means, median, SD, for continuous variables, and 95% CIs where appropriate. Analysis of normal distribution was performed with Q-Q plots, histograms, and the Shapiro-Wilk test. We applied the nonparametric Friedman test on repeatedly measured data. We defined cutoffs for dNCR/p-NCD following the recommendations by Borchers et al [6]: mild decline at 1 to 2 SD and major decline >2 SD in at least 2 test parameters at postoperative time points 3 - 5 weeks or 6 months. We considered a 2-sided P value of <.05 as statistically significant. All statistical analyses were performed in IBM SPSS version 28 (IBM Corp).

Qualitative Analysis

Three authors with experience in qualitative methods (AA, MS, and LB) analyzed the qualitative data through an iterative process. We conducted a deductive content analysis as described by Kyngäs and Kaakinen [32] by applying the theoretical framework of acceptability [27] on our data. Initially, we created an analysis matrix based on the 7 constructs of acceptability and added a dimension for study design, feasibility, and recruitment. Authors AA, MS, and LB thoroughly read each

XSL•FO
RenderX

transcribed interview multiple times to familiarize themselves with the data and assess the depth of information. LB organized the data using NVivo (version 15, QSR International Pty Ltd), and coded quotes related to our research questions using the analysis matrix. AA validated the codes independently in NVivo. Subsequent discussions with MS ensured coding decisions and researcher triangulation. MS created categories inductively according to the constructs in our matrix, with AA and LB reviewing and refining them collaboratively. AA translated the final meaning units from Swedish to English, and LB selected illustrative quotes for inclusion. Discussion of initial categories and reflections with the co-authors of the manuscript led to the re-organization of our meaning units and categories within the matrix. All authors' diverse clinical and research backgrounds, including anesthesia, surgery, postoperative care, intensive care, and neuropsychology, provided complementary perspectives during analysis. The content analysis is presented in Table 4.

Table . The process of content analysis.

Text quote examples	Categories	Domains
"[…] it's actually okay to contact patients, […] to come for an extra visit, just to participate in a study, because the patients who've been involved have, at least most of them, thought it was very fun." (RN4)	Barriers and facilitating factors in selection, recruitment, and retention processes	Study design, recruitment, and retention
"I thought it was interesting... […], this study you're doing, […] to see how it affects you, because […] I hadn't thought that a physical procedure would have such a big psychological impact as I understood it would have. […] I had never reflected on it in that way, but I was more... thought that a physical procedure is a physical procedure and then you move on […]" (P23)	Participants' feelings and attitudes toward the intervention	Affective attitudes
"I thought it was easy... actually easy to understand. There weren't any practical difficulties at all with it, I thought. However, some parts might have been difficult to carry out at different times. But overall, I thought it was very simple, clear, and good." (P20)	No sense of burden Illness- and symptom related factors	Burden
"But it's a […] patient group that has cancer. […] they are a group that you must be a little more careful with, especially if they've just received a diagnosis, and... Some are, like... they are right there and want to […] whereas many are like, 'No, I can't take anything more from anyone, because I've just gotten this news. I have so much.' So, […] you have to be a little more, not too... well, careful with, maybe, a little... […] A little more sensitive, I think […] there are some who've felt they haven't had the energy, because they've had it so tough with so many other things, you know, around their surgery, preparations, and their cancer, and everything." (RN1)		

Ethical Considerations

We obtained an ethical permit (2022-03593-01) from the Swedish Ethical Review Authority before initiating the study. All procedures complied with the World Medical Association's Declaration of Helsinki. Written informed consent was obtained from all participants before inclusion. Data were pseudonymized to ensure confidentiality. Participation was voluntary, without compensation, and participants could withdraw at any time.

Results

Participant Characteristics

Twenty-four patients were included in the study: 13 women and 11 men. The mean age was 77 (SD 6.5) years. Half of the patients (12/24, 50%) were classified as "very fit," and none were classified as frail according to the Clinical Frailty Scale (CFS). Most of the patients had obtained a university degree and had an advanced occupational background: two patients were still working at the time of data collection. Most patients (20/24, 83%) underwent laparoscopic colorectal surgery for cancer under general anesthesia combined with spinal anesthesia. Eight patients developed postoperative complications, such as wound infections. Two of these patients required resurgery, and one of them also needed intensive care. At T2, 3 patients were treated with or had recently completed chemotherapy, one patient had developed an autoimmune disease, and 2 had a recent infection. Patient characteristics are presented in Table 5. The 5 registered nurses interviewed were between 35 years and 60 years old, with 12-29 years of experience in their profession.

XSL•FO
RenderX

Table . Patient characteristics.

Characteristic	Total sample (N=24)
Sex, n (%)	
Female	13 (54)
Male	11 (46)
Age, years	
Mean (SD)	77 (6.5)
Min-max	63 - 90
Level of education, n (%)	
Elementary school	4 (17)
Upper secondary school or similar	3 (12)
Post-secondary education	4 (17)
University education	13 (54)
Occupation ISCO-08[a], n (%)	
Managers	8 (33)
Professionals	11 (46)
Technicians and associate professionals	3 (12)
Service and sales workers	2 (8)
Clinical Frailty Scale, mean (SD)	
Very fit	12 (50)
Well	7 (29)
Managing well	5 (21)
Mini-Mental State Examination, mean (SD)	29 (1)
American Society of Anesthesiologists classification[b], n (%)	
I	4 (17)
II	11 (46)
III	8 (33)
Comorbidities before surgery, n (%)	
Cardiac disease	7 (29)
Vascular disease	2 (8)
Lung disease	4 (17)
Kidney disease	2 (8)
Diabetes	1 (4)
History of cancer	3 (13)
Autoimmune disease	2 (8)
Type of surgery[b], n (%)	
Laparoscopic colorectal	20 (87)
Laparoscopic paraesophageal hernia	3 (13)
Type of anesthesia[b], n (%)	
General anesthesia	3 (13)
General anesthesia and epidural	3 (13)
General anesthesia and spinal	17 (71)
Duration of surgery[b]	

XSL•FO
RenderX

Characteristic	Total sample (N=24)
Minutes, mean (SD)	190 (77)
Duration of anesthesia[b]	
Minutes, mean (SD)	258 (86)
Intraoperative bleeding[b]	
Milliliter, mean (SD)	28 (33)
Length of hospital stay[b], mean (SD)	5 (4)
Postoperative complications[b], n (%)	8 (33)
Resurgery[b], n (%)	2 (8)

[a]ISCO-08 = International Standard Classification of Occupations.
[b]One missing.

Feasibility

The set progression criteria were met and the digital cognitive test battery was deemed feasible; System Usability Scale Score ≥75, recruiting ≥20% of available patients and dropouts <30%.

Recruitment and Retention

Recruitment took place over a period of 1.5 years, from January 2023 to June 2024. Data collection concluded in November 2024. Out of 117 patients screened for eligibility (Figure 1), 20.5% (n=24) were included in the study, dropout rates were 12% at T2 and 27% at T3.

XSL•FO
RenderX

Figure 1. Study flowchart of participants from screening to study completion. Follow-up assessments were conducted postoperatively at 1‐3 days, 3‐5 weeks, and 6 months. Numbers of participants excluded, lost to follow-up, and analyzed at each time point are shown. Postop: postoperative follow-up.

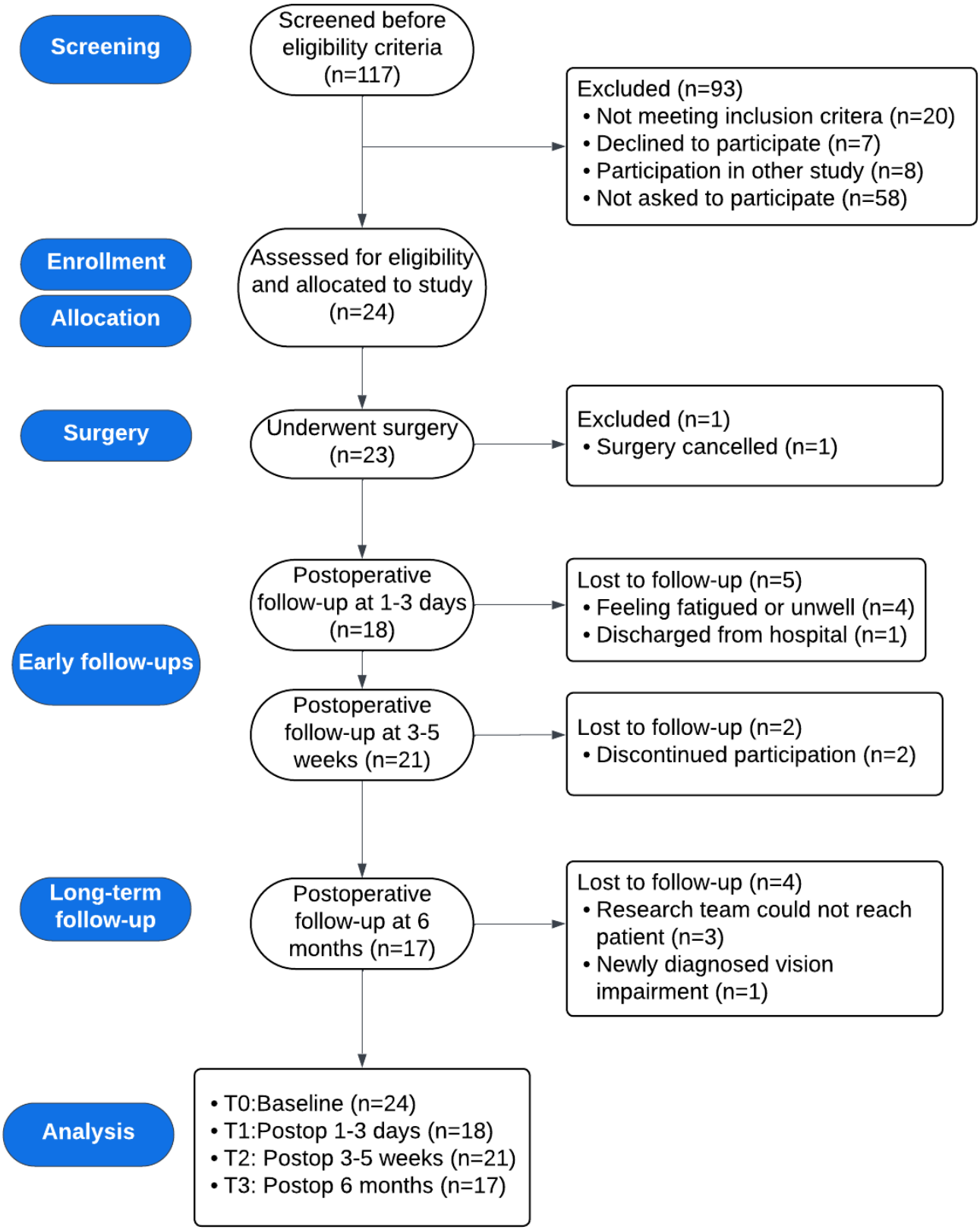

Acceptability, Perceived Value, and Usability

Twenty patients completed the SUS at T2, the mean score was 87 (SD 17.9; 95% CI 78.9‐95.2), indicating excellent usability [33]. The qualitative results are presented below.

Study Design, Recruitment, and Retention

Several key factors influencing recruitment success were described by the nurses: the research team being available, and engaging with the clinical staff about the study, good collaboration between the research unit and the nurses at the

XSL•FO
RenderX

surgery clinic, and that the study did not require any extra visits to the hospital. The nurses found it easy to approach patients for participation since the study was not information heavy. They emphasized that clear written and verbal communication from the research team was beneficial and improved the collaboration between the researchers and the clinic.

The nurses at the surgery clinic were the first to meet the patients and therefore played an important role in the recruitment process. Positive attitudes among the nurses also contributed to a favorable recruitment environment. However, several barriers were also identified. These included time constraints, as nurses sometimes had no opportunity to approach patients during the in-person preoperative consultation, as other commitments took priority. Some patients did not attend an in-person preoperative consultation at all, meaning they were not approached for participation. According to the nurses, some participants, particularly older adults, may struggle with using a tablet, which in some cases led nurses to refrain from inviting them to participate. The nurses also emphasized that the patients constituted a vulnerable group due to their cancer diagnoses. At times, the nurses felt it was inappropriate to invite these patients to participate in the research study for that reason.

> *Yes, absolutely [about hesitating to ask the patient to become a study participant]. […] if you feel the patient is struggling to process all the information you've already given about the care and what will happen, and then you feel like, "No, this is too much. This patient isn't receptive to this." Or if they're having a crisis reaction or if you assess that, "No, this isn't appropriate to ask outright."* [RN5]

Nurses also noted that clinical staff at the postoperative wards were unfamiliar with the delirium assessment, requiring repeated reminders and instructions from the research nurses. In addition, the nurses felt that the research team's expectations for recruitment numbers (50 patients) were unrealistic. Recruitment was further challenged by competition from another study targeting the same patient population, which meant some patients were approached only for the competing study. Suggestions from the nurses for improving the recruitment included clearer inclusion and exclusion criteria, flexibility in timing of inclusion, and better patient preparation, such as info available in the waiting room. The hospital also relocated to another building while the study was ongoing, which also affected recruitment.

Affective Attitudes

Patients expressed largely positive and curious feelings toward the digital test battery, describing it as interesting, enjoyable, and meaningful. It was perceived as easy to use and unproblematic. The patients also appreciated having the chance to do something else for a while. They also valued the opportunity to reflect on cognitive and psychological aspects of surgery. Many patients had not considered that surgery could impact their brains, as their attention was mainly focused on abdominal symptoms.

> *I thought it was interesting… […], this study you're doing, […] to see how it affects you, because […] I hadn't thought that a physical procedure would have such a big psychological impact as I understood it would have. […] I had never reflected on it in that way, but I was more… thought that a physical procedure is a physical procedure and then you move on […]* [P23]

At the same time, some patients described an anticipatory anxiety about the cognitive tests, particularly when they were already mentally burdened by their diagnosis or treatment. The nurses perceived that participants enjoyed taking the tests, as it allowed them to focus on focus on something different and get in the zone. They also noted that the patients valued the extra attention they received as they were also study participants, which seemed to increase motivation.

> *[It was] a very pleasant meeting with the patient […] and I think they've felt that too, that it's been a little […] extra thing, […]. Both those who have been seriously ill and those who have [been feeling fine]. And it's been a bit different. […] they've still thought it was fun, […] and they've been happy to do it [the cognitive test].* [RN2]

However, some nurses observed that certain participants found specific tests challenging or stressful. Despite this, the nurses themselves were satisfied with the digital cognitive test battery and appreciated the interactions with patients, noting only minor technical difficulties in one subtest. They also felt responsible for explaining to patients why the tests were intentionally designed to be challenging.

Burden

Participants' experiences of burden varied depending on individual circumstances and the timing of assessments. Several participants described no burden at all; the tests required minimal effort. They perceived the cognitive tests as easy to complete and as one of many and comparable to other routine activities within the care process. Some patients even approached the tests competitively, motivated to perform well.

> *I thought it was easy… actually easy to understand. There weren't any practical difficulties at all with it, I thought. However, some parts might have been difficult to carry out at different times. But overall, I thought it was very simple, clear, and good.* [P20]

For others, illness- and symptom-related factors influenced participation. Medical complications, fatigue from surgery or treatment preparation, and the overall cognitive strain of the perioperative process led participants to decline participation. Completing the tests shortly after surgery was described as cognitively demanding and exhausting. Nurses emphasized that patients with cancer diagnoses required extra attentiveness and support. They felt these patients often had a great deal to process regarding their diagnosis and treatments, and therefore needed to be approached with more care. The nurses also described that some patients experienced difficulties sitting upright after abdominal surgery due to their wound.

> *Some patients have […] had a positive expectation and thought, "This sounds really exciting, now I'll try this," while some have been more, like,*

XSL•FO
RenderX

> *anxiety-filled, "Oh no, yeah, should I participate…? This will be tough…" That is, some… some maybe aren't feeling great to begin with, and it feels like they… "Well, now am I also going to find out that I'm not quite sharp in the head anymore, now that my body has given up? Are we now going to discover that I can't manage this test?" […] I remember having patients, for example, who've had previous burnout […] and they've felt like, "No, I don't want to do this, because I don't want to add any stress about how my brain functions."* [RN3]

Intervention-specific factors also played a role. The nurses suggested that a simpler, less extensive cognitive test battery would be more feasible in the perioperative setting. Nonetheless, they observed that participation did not represent a major sacrifice for most patients.

Ethical Aspects

The participants viewed the study as meaningful and expressed a strong desire to contribute, even when faced with numerous other commitments around the time of surgery. Those who chose to participate were motivated by the opportunity to support research and potentially benefit future patients. According to the nurses, most patients who were invited agreed to participate.

> *[…] But when you notice that someone… has impaired cognition, maybe not just from the surgery, but actually for another reason and would need a dementia assessment. What do you do then? It becomes another… [about the responsibility to provide feedback as a clinician].* [RN4]

At the same time, some nurses highlighted an ethical dilemma concerning the handling of potentially negative cognitive test results. Because they did not routinely work with cognitive assessments, they felt uncertain about how to respond if the cognitive tests would be implemented at their clinic.

Intervention Coherence

Several factors facilitated test compliance. Participants described that the instructions on the digital cognitive test battery were very clear and easy to understand, even if some were not used to using a tablet. The patients described that they thought it could be implemented in the perioperative process as they found it interesting.

> *I've never worked with a tablet like that. I use a regular computer. I'm not so good with my phone either, but with browsing around, it's the computer. But it worked with that tablet. […] when I was actually sitting there, I went into that bubble.* [P15]

The nurses emphasized the importance of explaining that the tests were intended to be challenging but not competitive, in order to prevent participants from feeling inadequate. The timing of assessments also influenced participants' engagement: tests administered 1‑3 days postoperatively were difficult to complete due to fatigue, whereas later follow-ups were easier as participants had more energy. There were also other barriers and risks to the execution and implementation of the cognitive test battery:

> *[…] it requires that all the staff who handle them, you know, get proper training on how it works. I don't think you can just throw it in and say, "Here, give this to a patient who's going to do this," you have to[…] do it carefully, just like with any research study, you need to implement it properly.* [RN1]

Since the tests were self-administered, the nurses sometimes felt less involved. They noted; however, that some patients required additional support during the test session and appreciated the nurse's presence. Nurses highlighted that comprehensive staff training would be essential for successful implementation if the tests were to be used more widely in the clinic.

Opportunity Costs

The patients viewed participation in the study as a positive investment in their cognitive health. They found it interesting to learn about their cognitive functions and considered the sessions a worthwhile trade-off, often describing them as "just another thing on the to-do list"

> *[…] it's also interesting to know how their cognitive abilities are after the surgery, so it's a balancing act there, […] But I think… otherwise, I think it has been good.* [RN1]

One patient described a situation where a nurse appeared stressed while administering the test. The nurses emphasized that staff need sufficient time both to administer the tests and to ensure they are completed in a proper manner.

Perceived Effectiveness

Participants appreciated both the cognitive and clinical value of the digital cognitive test battery. The nurses noted that, if resources allowed, the tests could potentially be integrated into routine care. From a cognitive perspective, patients found the tests stimulating, though some tests were highly challenging, which made them aware of their own cognitive limitations.

> *And I didn't quite realize, you know, what shortcomings or… what deficiencies I have. […] that was a bit exciting, and a bit revealing that in some respects, I'm as inattentive as I am. So, I thought that was… it was an interesting experience.* [P23]

Some patients described that repeated completion of the tests allowed them to learn and improve over time. They also described feeling that their "brains were activated," which inspired them to engage in continued cognitive training outside the study. A few patients noted that the tests gave them a sense of what memory or dementia assessments would be like.

Self-Efficacy

The nurses described that some patients needed additional support during the test sessions to move forward with each test, as some instructions were perceived as unclear at times.

> *[…] most of them, I think, need a test leader with them. It's a bit age-related too. The ones who are a bit younger and more familiar with technology. Because some can already say […] "Well, I can't… I'm so bad with all this technology. I won't be able*

XSL•FO
RenderX

> *to manage this." But then when you explain, "This is it. You just have to follow the instructions […]" For many, especially if they are older, can be a bit scared of this technology, which isn't about filling out papers. […] they probably would manage it quite well on their own, but they felt a sense of security knowing there was a person beside them.* [RN1]

However, most patients managed the test session fully independently. The patients also described that they felt confident in their abilities to complete the tests, without extra assistance.

Cognitive Trajectories

Three patients fulfilled the criteria of POD for only one day while in the surgical ward: 2 on the first postoperative day and one on the third postoperative day. No further cases of POD were observed during the subsequent postoperative days in the hospital. There were statistically significant changes in cognitive test results between the preoperative and postoperative follow-ups in the attention and tempo domain and the executive functions domain. The patients' *z* scores are presented in Multimedia Appendix 2. No patients met the criteria for dNCR or p-NCD at the postoperative follow-ups. Cognitive raw scores from the different data points are presented in Multimedia Appendix 3.

Trajectories of Depression, Functional Status, and Quality of Recovery

Preoperatively, one patient fulfilled cut-off criteria for depression, 4 patients at T2, and one at T3. However, on a group level, the GDS scoring was low overall. Functional status scores with World Health Organization Disability Assessment Schedule (WHODAS) 12 2.0 were highest at T2. The patients' highest scoring symptoms on the Swedish Quality of Recovery questionnaire (SwQoR) were in the cognitive and emotional domain eg, worry, concentration, and well-being. There were no statistically significant changes between preoperative and postoperative measurements of depression, functional capacity, and quality of recovery (Multimedia Appendix 2).

Discussion

Principal Findings

This was a nonrandomized feasibility study aimed at assessing a digital, self-administered cognitive test battery in older adults undergoing abdominal surgery. Recruitment and retention proved challenging, with rates falling below expectations. The 1.5-year recruitment period turned out to be unfeasible, and a single-center approach is not viable for a larger-scale trial. One contributing factor to the shortfall in sample size was that 58 eligible patients were not approached for participation, in part because some did not attend a preoperative in-person consultation before their scheduled surgery. Similar recruitment challenges have been reported in feasibility studies involving patients with cancer diagnoses [10,34,35].

For a future trial, a multicenter recruitment model using both clinical and research staff may be necessary [36]. Consent to participate can be performed remotely via an invitation letter or email to the patient before the preoperative consultation [36], and eligibility screening could be performed remotely using validated tools, such as the telephone-administered Montreal Cognitive Assessment [37,38] instead of in-person MMSE. Online advertisement such as Google Search Engine may further expand reach [39].

Other contributing factors behind low recruitment rates were staff-related: nurses refrained from inviting some patients to participate. Their decisions were influenced by their perceptions of patient receptiveness, emotional considerations due to their cancer diagnosis which aligns with previous research in patients with lung cancer diagnoses [40], and patients' familiarity with digital devices. Additional barriers included logistical disruptions, such as hospital relocation.

The participants highlighted barriers and facilitators to implementing cognitive assessments in the perioperative setting. Some nurses expressed uncertainty about how to respond when a patient scored low on the cognitive tests, emphasizing the need for structured guidance and staff training. This aligns with findings from a study on cognitive assessments in primary care [41]. Ethical and implementation challenges can be addressed through staff education and decision-support interventions; previous evidence suggests these approaches have shown mixed success in improving management of cognitive disorders [42]. Moreover, factual knowledge alone may not be sufficient; addressing clinicians's attitudes and implicit biases has also been shown to influence the management of cognitive disorders [43].

In terms of acceptability and usability, the digital cognitive test battery was considered easy to use and comprehend, with a high perceived value from both patients and nurses. Patients rated the digital cognitive test battery as having excellent usability, which aligns with previous findings [7]. Although cognitive screening is not performed routinely in older adults undergoing surgery, it provides actionable information to identify at-risk patients [44]. Cognitive test results indicating patients are at risk can guide preventative strategies, including implementation of the Hospital Elder Life Program for postoperative delirium [45], referral to geriatric specialists, or integration of a multidisciplinary perioperative care team [44].

No patient developed dNCR or p-NCD following surgery. The subtest assessing attention and processing speed improved over time; however, this finding must be interpreted with caution. Given the repeated exposure to the same tests, practice effects may contribute to observed improvement [46]. The small sample size and lack of a control group in our study constrain interpretation. Future trials should consider alternate versions of similar tests or control groups. Several patients had a high educational and occupational background, which reflects a cognitively robust sample. Patients with low cognitive reserves are at higher risk of developing postoperative cognitive disorders compared to those with high cognitive reserves [47,48].

This study had some strengths. To the best of our knowledge, this was the first time a digital, self-administered cognitive test battery was tested for feasibility in this clinical context. By deciding a priori progression criteria, we reduced the risk of designing an ill-informed larger trial in the future [28].

XSL•FO
RenderX

We acknowledge the limitations in this study. The cohort was relatively homogeneous and cognitively robust. This homogeneity limits the external validity, as feasibility in this group may not reflect outcomes in frailer or more diverse perioperative populations, particularly those with lower cognitive reserves. Moreover, the high usability scores should be interpreted cautiously, as they may not translate to patients with limited digital literacy. Additional limitations include the absence of a control group, which restricts interpretation of cognitive outcomes, as improvements in attention may partly reflect practice effects rather than true changes. Moreover, some of the clinician- and patient-reported outcome measurements used in this study have not been validated in this population and setting. Recruitment difficulties and the modest sample size posed additional limitations, and we acknowledge that recruitment was, in many cases, not feasible. We have a few inconsistencies with the published study protocol [13] before starting our study, as we extended the data collection and decided to include nurses to assess the acceptability and usability of the digital cognitive test battery.

Conclusions

The digital, self-administered cognitive test battery was found to be feasible, acceptable, and usable in older adults undergoing abdominal surgery. However, recruitment challenges and the homogeneous sample limit the generalizability of these findings. Implementation in clinical practice would require structured staff training and ethical considerations. Future trials should include larger, more diverse samples, consider control groups or alternate test versions to account for practice effects in the cognitive tests.

Acknowledgments

We thank all participants, including patients and the nurses involved in this study. We thank Wobbie van den Hurk, Frida Strömgren, and the rest of the team at Mindmore for support and collaboration. Extra thanks to Fredrik Johansson at KI-DS for statistical guidance. This work was supported by Karolinska Institutet (grant 2020 - 02641), Erling-Persson Foundation (grant 2022 0163), Ragnhild Einar Lundström Foundation (grant 4–2018/2013) and Demensfonden (grant 4–2016/2023)

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Authors' Contributions

Conceptualization: AA, LB, AT, JE, UN
Methodology: all authors
Software: AA, LB
Formal analysis: AA, LB, MS, GM
Investigation: AA, LB
Data Curation: AA, LB
Writing-original draft: AA, UN
Writing-review and editing: all authors
Visualization: AA
Project administration: AA, LB
Funding acquisition: JE, LB, UN, AT
Supervision: MS, JE, UN, LB, GM

Conflicts of Interest

None declared.

Multimedia Appendix 1
Semistructured interview guides for patients and nurses.
[DOCX File, 20 KB - formative_v9i1e71911_app1.docx]

Multimedia Appendix 2
Patients' *z* scores and depression, functional, and quality of recovery scores.
[DOCX File, 17 KB - formative_v9i1e71911_app2.docx]

Multimedia Appendix 3
Cognitive raw scores and times.
[DOCX File, 22 KB - formative_v9i1e71911_app3.docx]

XSL•FO
RenderX

Checklist 1
CONSORT checklist.
[PDF File, 112 KB - formative_v9i1e71911_app4.pdf]

References

1. Wolfe JD, Wolfe NK, Rich MW. Perioperative care of the geriatric patient for noncardiac surgery. Clin Cardiol 2020 Feb;43(2):127-136. [doi: 10.1002/clc.23302] [Medline: 31825127]
2. Evered L, Silbert B, Knopman DS, et al. Recommendations for the nomenclature of cognitive change associated with anaesthesia and surgery-2018. Anesthesiology 2018 Nov;129(5):872-879. [doi: 10.1097/ALN.0000000000002334] [Medline: 30325806]
3. Mahanna-Gabrielli E, Schenning KJ, Eriksson LI, et al. State of the clinical science of perioperative brain health: report from the American Society of Anesthesiologists Brain Health Initiative Summit 2018. Br J Anaesth 2019 Oct;123(4):464-478. [doi: 10.1016/j.bja.2019.07.004] [Medline: 31439308]
4. Amirpour A, Bergman L, Markovic G, Liander K, Nilsson U, Eckerblad J. Understanding neurocognitive recovery in older adults after total hip arthroplasty-neurocognitive assessment, blood biomarkers and patient experiences: a mixed-methods study. BMJ Open 2025 Jan 28;15(1):e093872. [doi: 10.1136/bmjopen-2024-093872] [Medline: 39880425]
5. Borchers F, Knaak C, Piper SK, Spies C. Recommendations for the detection and specification of perioperative neurocognitive disorders. Anasthesiol Intensivmed Notfallmed Schmerzther 2019 Nov;54(11-12):652-667. [doi: 10.1055/a-0853-3060] [Medline: 31805585]
6. Borchers F, Spies CD, Feinkohl I, et al. Methodology of measuring postoperative cognitive dysfunction: a systematic review. Br J Anaesth 2021 Jun;126(6):1119-1127. [doi: 10.1016/j.bja.2021.01.035] [Medline: 33820655]
7. Amirpour A, Eckerblad J, Bergman L, Nilsson U. Comparing analog and digital neurocognitive tests with older adults: a study of the ISPOCD battery vs. a digital test battery from Mindmore. BMC Geriatr 2024 Jan 8;24(1):34. [doi: 10.1186/s12877-023-04648-w] [Medline: 38191318]
8. van den Hurk W, Bergman I, Machado A, Bjermo J, Gustavsson A. Swedish normative data for mindmore: a comprehensive cognitive screening battery, both digital and self-administrated. J Int Neuropsychol Soc 2022 Feb;28(2):188-202. [doi: 10.1017/S135561772100045X] [Medline: 34027854]
9. Hackett K, Xu S, McKniff M, Paglia L, Barnett I, Giovannetti T. Mobility-based smartphone digital phenotypes for unobtrusively capturing everyday cognition, mood, and community life-space in older adults: feasibility, acceptability, and preliminary validity study. JMIR Hum Factors 2024 Nov 22;11:e59974. [doi: 10.2196/59974] [Medline: 39576984]
10. Karlsson P, Nygren-Bonnier M, Henningsohn L, Rydwik E, Hagströmer M. The feasibility of using a digital tool to enhance mobilisation following abdominal cancer surgery-a non-randomised controlled trial. Pilot Feasibility Stud 2023 Aug 23;9(1):147. [doi: 10.1186/s40814-023-01377-0] [Medline: 37612761]
11. Lee KW, Hong YJ, Yang EJ, et al. Feasibility and usefulness of cognitive monitoring using a new home-based cognitive test in mild cognitive impairment: a prospective single arm study. BMC Geriatr 2024 Mar 8;24(1):241. [doi: 10.1186/s12877-024-04850-4] [Medline: 38459495]
12. Skivington K, Matthews L, Simpson SA, et al. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. BMJ 2021 Sep 30;374:n2061. [doi: 10.1136/bmj.n2061] [Medline: 34593508]
13. Amirpour A, Eckerblad J, Thorell A, Bergman L, Nilsson U. Usability and feasibility of a digital cognitive screening tool measuring older adults' early postoperative neurocognitive recovery: a protocol for a pilot study. BMJ Open 2023 Jul 21;13(7):e070404. [doi: 10.1136/bmjopen-2022-070404] [Medline: 37479514]
14. Lancaster GA, Thabane L. Guidelines for reporting non-randomised pilot and feasibility studies. Pilot Feasibility Stud 2019;5:114. [doi: 10.1186/s40814-019-0499-1] [Medline: 31608150]
15. Eldridge SM, Chan CL, Campbell MJ, et al. CONSORT 2010 statement: extension to randomised pilot and feasibility trials. BMJ 2016 Oct 24;355:i5239. [doi: 10.1136/bmj.i5239] [Medline: 27777223]
16. Lancaster GA, Dodd S, Williamson PR. Design and analysis of pilot studies: recommendations for good practice. J Eval Clin Pract 2004 May;10(2):307-312. [doi: 10.1111/j..2002.384.doc.x] [Medline: 15189396]
17. Rockwood K, Song X, MacKnight C, et al. A global clinical measure of fitness and frailty in elderly people. CMAJ 2005 Aug 30;173(5):489-495. [doi: 10.1503/cmaj.050051] [Medline: 16129869]
18. Hörlin E, Munir Ehrlington S, Henricson J, John RT, Wilhelms D. Inter-rater reliability of the Clinical Frailty Scale by staff members in a Swedish emergency department setting. Acad Emerg Med 2022 Dec;29(12):1431-1437. [doi: 10.1111/acem.14603] [Medline: 36200372]
19. Gaudreau JD, Gagnon P, Harel F, Tremblay A, Roy MA. Fast, systematic, and continuous delirium assessment in hospitalized patients: the nursing delirium screening scale. J Pain Symptom Manage 2005 Apr;29(4):368-375. [doi: 10.1016/j.jpainsymman.2004.07.009] [Medline: 15857740]
20. Lingehall HC, Smulter N, Engström KG, Gustafson Y, Olofsson B. Validation of the Swedish version of the Nursing Delirium Screening Scale used in patients 70 years and older undergoing cardiac surgery. J Clin Nurs 2013 Oct;22(19-20):2858-2866. [doi: 10.1111/j.1365-2702.2012.04102.x] [Medline: 24033713]

XSL•FO
RenderX

21. Myles PS, Weitkamp B, Jones K, Melick J, Hensen S. Validity and reliability of a postoperative quality of recovery score: the QoR-40. Br J Anaesth 2000 Jan;84(1):11-15. [doi: 10.1093/oxfordjournals.bja.a013366] [Medline: 10740540]
22. Nilsson U, Dahlberg K, Jaensson M. The Swedish web version of the quality of recovery ccale adapted for use in a mobile app: prospective psychometric evaluation study. JMIR Mhealth Uhealth 2017 Dec 3;5(12):e188. [doi: 10.2196/mhealth.9061] [Medline: 29229590]
23. Sheikh JI, Yesavage JA. Geriatric Depression Scale (GDS): recent evidence and development of a shorter version. Clin Gerontol 1986(1-2):165-173. [doi: 10.1300/J018v05n01_09]
24. Snellman S, Hörnsten C, Olofsson B, Gustafson Y, Lövheim H, Niklasson J. Validity and test-retest reliability of the Swedish version of the Geriatric Depression Scale among very old adults. BMC Geriatr 2024 Mar 18;24(1):261. [doi: 10.1186/s12877-024-04869-7] [Medline: 38500031]
25. Brooke J. SUS - a quick and dirty usability scale. In: Usability Evaluation in Industry: Taylor & Francis; 1996:189-194.
26. Holmberg C, Gremyr A, Torgerson J, Mehlig K. Clinical validity of the 12-item WHODAS-2.0 in a naturalistic sample of outpatients with psychotic disorders. BMC Psychiatry 2021 Mar 10;21(1):147. [doi: 10.1186/s12888-021-03101-9] [Medline: 33691655]
27. Sekhon M, Cartwright M, Francis JJ. Development of a theory-informed questionnaire to assess the acceptability of healthcare interventions. BMC Health Serv Res 2022 Mar 1;22(1):279. [doi: 10.1186/s12913-022-07577-3] [Medline: 35232455]
28. Mbuagbaw L, Kosa SD, Lawson DO, et al. The reporting of progression criteria in protocols of pilot trials designed to assess the feasibility of main trials is insufficient: a meta-epidemiological study. Pilot Feasibility Stud 2019;5:120. [doi: 10.1186/s40814-019-0500-z] [Medline: 31700654]
29. Nilsson U, Liander K, Rooyackers O, Eriksson LI. Patients' experiences of early postoperative cognition and its relation to cognitive decline and inflammatory responses: a protocol for a mixed-methods study. BMJ Open 2019 Dec 15;9(12):e032650. [doi: 10.1136/bmjopen-2019-032650] [Medline: 31843843]
30. Northgraves MJ, Arunachalam L, Madden LA, et al. Feasibility of a novel exercise prehabilitation programme in patients scheduled for elective colorectal surgery: a feasibility randomised controlled trial. Support Care Cancer 2020 Jul;28(7):3197-3206. [doi: 10.1007/s00520-019-05098-0] [Medline: 31712950]
31. Macleod M, Steele RJC, O'Carroll RE, et al. Feasibility study to assess the delivery of a lifestyle intervention (TreatWELL) for patients with colorectal cancer undergoing potentially curative treatment. BMJ Open 2018 Jun 6;8(6):e021117. [doi: 10.1136/bmjopen-2017-021117] [Medline: 29880567]
32. Kyngäs H, Kaakinen P. Deductive content analysis. In: Kyngäs H, Mikkonen K, Kääriäinen M, editors. The Application of Content Analysis in Nursing Science Research: Springer International Publishing; 2020:23-30. [doi: 10.1007/978-3-030-30199-6_3]
33. Lewis JR. The System Usability Scale: past, present, and future. Int J Hum Comput Interact 2018 Jul 3;34(7):577-590. [doi: 10.1080/10447318.2018.1455307]
34. Murphy K, Kehoe B, Denieffe S, Hacking D, Fairman CM, Harrison M. Comparing aerobic and resistance exercise emphasis during androgen deprivation and radiation therapy for prostate cancer: a randomised feasibility trial. Support Care Cancer 2025 Jun 20;33(7):601. [doi: 10.1007/s00520-025-09650-z] [Medline: 40540036]
35. Boulanger MC, Lo SB, Centracchio JA, et al. Pilot feasibility trial of a supportive care digital application for patients with advanced nonsmall cell lung cancer. J Palliat Med 2025 Sep;28(9):1210-1219. [doi: 10.1089/jpm.2025.0066] [Medline: 40468832]
36. Ninomiya MM, Hiemstra J, Nicholson E, Isaac KV. Methods of recruitment for surgical and perioperative randomized controlled trials: a rapid review. World J Surg 2023 Nov;47(11):2659-2667. [doi: 10.1007/s00268-023-07124-z] [Medline: 37589794]
37. Katz MJ, Wang C, Nester CO, et al. T-MoCA: a valid phone screen for cognitive impairment in diverse community samples. Alzheimers Dement (Amst) 2021;13(1):e12144. [doi: 10.1002/dad2.12144] [Medline: 33598528]
38. Benge JF, Kiselica AM. Rapid communication: preliminary validation of a telephone adapted Montreal Cognitive Assessment for the identification of mild cognitive impairment in Parkinson's disease. Clin Neuropsychol 2021 Jan;35(1):133-147. [doi: 10.1080/13854046.2020.1801848] [Medline: 32779959]
39. Brøgger-Mikkelsen M, Ali Z, Zibert JR, Andersen AD, Thomsen SF. Online patient recruitment in clinical trials: systematic review and meta-analysis. J Med Internet Res 2020 Nov 4;22(11):e22179. [doi: 10.2196/22179] [Medline: 33146627]
40. Lond B, Dodd C, Davey Z, et al. A systematic review of the barriers and facilitators impacting patient enrolment in clinical trials for lung cancer. Eur J Oncol Nurs 2024 Jun;70:102564. [doi: 10.1016/j.ejon.2024.102564] [Medline: 38554615]
41. O'Brien KS, Harkins K, Peifer M, et al. Designing an intervention to improve cognitive evaluations in primary care. Implement Sci Commun 2025 Jan 16;6(1):9. [doi: 10.1186/s43058-025-00693-1] [Medline: 39819971]
42. Perry M, Drašković I, Lucassen P, Vernooij-Dassen M, van Achterberg T, Rikkert MO. Effects of educational interventions on primary dementia care: a systematic review. Int J Geriatr Psychiatry 2011 Jan;26(1):1-11. [doi: 10.1002/gps.2479] [Medline: 21157845]
43. Cartz-Piver L, Calvet B, Mehrabian-Spassova S, et al. Empowering general practitioners in dementia care: the ANTISTIGMA education intervention in Europe. Int J Geriatr Psychiatry 2023 Jul;38(7):e5958. [doi: 10.1002/gps.5958] [Medline: 37392094]

XSL•FO
RenderX

44. Chan LK, Chuan A, Berney CR, Chan DK. Routine cognitive screening for older people undergoing major elective surgery: benefits, risks and costs. Anaesth Intensive Care 2024 Sep;52(5):275-282. [doi: 10.1177/0310057X241232421] [Medline: 39212187]
45. Hshieh TT, Yang T, Gartaganis SL, Yue J, Inouye SK. Hospital elder life program: systematic review and meta-analysis of effectiveness. Am J Geriatr Psychiatry 2018 Oct;26(10):1015-1033. [doi: 10.1016/j.jagp.2018.06.007] [Medline: 30076080]
46. Mackin RS, Burns E, Insel P, et al. Test taking location and practice effects as factors contributing to scores on remotely administered neurocognitive performance tests in a sample of older adults. Appl Neuropsychol Adult 2025 Jul 9:1-7. [doi: 10.1080/23279095.2025.2518574] [Medline: 40632686]
47. Megari K, Kosmidis MH. Protecting the brain while healing hearts: the protective role of cognitive reserve in cardiac surgery. Am J Geriatr Psychiatry 2024 Feb;32(2):195-204. [doi: 10.1016/j.jagp.2023.10.003] [Medline: 37926673]
48. Feinkohl I, Winterer G, Spies CD, Pischon T. Cognitive reserve and the risk of postoperative cognitive dysfunction. Dtsch Arztebl Int 2017 Feb 17;114(7):110-117. [doi: 10.3238/arztebl.2017.0110] [Medline: 28302254]

Abbreviations

CERAD: Consortium to Establish a Registry for Alzheimer Disease
CFS: Clinical Frailty Scale
CONSORT: Consolidated Standards of Reporting Trials
dNCR: delayed neurocognitive recovery
GDS-15: Geriatric Depression Scale-15
MMSE: Mini-Mental State Examination
NU-DESC: Nursing Delirium Screening Scale
p-NCD: postoperative neurocognitive disorder
SUS: System Usability Scale
SwQoR-24: Swedish Quality of Recovery questionnaire
WHODAS: World Health Organization Disability Assessment Schedule

Edited by A Mavragani; submitted 29.01.25; peer-reviewed by P Opasanant, V Minville; revised version received 30.09.25; accepted 15.10.25; published 07.11.25.

Please cite as:
Amirpour A, Saarijärvi M, Eckerblad J, Markovic G, Thorell A, Nilsson U, Bergman L
Evaluation of a Digital, Self-Administered, Cognitive Test Battery in Older Adult Patients Undergoing Abdominal Surgery: Nonrandomized Feasibility Trial
JMIR Form Res 2025;9:e71911
URL: https://formative.jmir.org/2025/1/e71911
doi:10.2196/71911

© Anahita Amirpour, Markus Saarijärvi, Jeanette Eckerblad, Gabriela Markovic, Anders Thorell, Ulrica Nilsson, Lina Bergman. Originally published in JMIR Formative Research (https://formative.jmir.org), 7.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Designing a Gait Recognition Algorithm for Older Adults Using Mobility Aids: Prospective Cohort Study

Samantha Jeane Ray[1], BS, PhD; Jung In Koh[2], BS, MS; Amanda Mae Liberty[1], BS; Tracy Anne Hammond[2], BA, BS, MS, MA, FTO, PhD; Paula Kay Shireman[1], BS, MS, MBA, MD

[1]College of Medicine, Texas A&M University, MREB 2 8447 John Sharp Parkway, Bryan, TX, United States

[2]College of Engineering, Texas A&M University, College Station, TX, United States

Corresponding Author:
Samantha Jeane Ray, BS, PhD
College of Medicine, Texas A&M University, MREB 2 8447 John Sharp Parkway, Bryan, TX, United States

Abstract

Background: Maintaining mobility is important for older adults to retain independence and reduce fall risk. Wearable technology such as fitness trackers and smartwatches can track physical activity. Unfortunately, gait recognition algorithms are often calibrated using younger adults and are not accurate for older adults, especially when using mobility aids.

Objective: Our goal was to develop a gait recognition algorithm capable of detecting the walking patterns of older adults that is robust to using mobility aids. Wrist-worn wearable devices were used to maximize the ubiquity of the approach.

Methods: We collected walking and other daily activity data on 17 independent older adults to develop a gait recognition algorithm. Ten participants used mobility aids (ie, 5 cane users, 4 rollator users, 1 walker user). We calibrated a heuristic-based "one-size-fits-most" algorithm leveraging the harmonic patterns associated with walking to recognize the walking patterns of our cohort. This algorithm is computationally lightweight and relies only on accelerometer data. We used hyperparameter tuning using a Parzen tree estimator to find the optimal parameters in a leave-one-subject-out fashion.

Results: The calibration process was required for this algorithm to detect walking. The signal amplitude threshold lowered from 0.3g to 0.1g to detect the more subtle walking patterns of older adults. The walking frequency range widened from [1.4Hz, 2.3Hz] to [0.8Hz, 2.8Hz], showing that older adults walk more slowly. The ratio for superharmonics increased from 1.4 to 77.11. Analyzing the false positive rate for the other daily activity classes implies that these superharmonics are artifacts of back-and-forth arm motions that characterize walking in our collected data. Additionally, we report the performance metrics of sensitivity, specificity, and F_1-score to evaluate our algorithm. Sensitivity increased from 0.11 to 0.73; F_1-score increased from 0.15 to 0.73; and specificity decreased from 0.99 to 0.75.

Conclusions: This study successfully recognized the walking patterns of older adults with or without mobility aids. The performance metrics show that this algorithm has promise for being used to monitor physical activity. This approach is computationally lightweight and explainable. Our calibration approach can be adopted to tune to new populations and has a low barrier to entry for adopting a new technology due to the sole reliance on accelerometer data which is a standard sensor in wearable devices. The most noteworthy parameter adjustment is the ratio for superharmonics. Low values for the subharmonic and superharmonic ratios cause the algorithm not to detect walking in our older adult data. We validated the algorithm on ten mobility aid users. A larger study with more participants using mobility aids is necessary to conduct a deeper analysis on what parameters work best for this population. Future work includes validating the algorithm's ability to estimate step counts and measure physical activity in real-world settings.

(JMIR Form Res 2025;9:e68669) doi:10.2196/68669

KEYWORDS

gait recognition; walking recognition; older adults; wearable technology; activities of daily living

Introduction

Detecting changes in lower body function is crucial for preventing falls and maintaining mobility and quality of life [1-3]. Falls can be life-threatening for older adults; approximately one-third of individuals over the age of 65 experience a fall each year [4] and 55.8% of accidental deaths are due to falls [5]. Moreover, impairments in lower body function increase the risk of falls [6,7] and can also lead to more severe outcomes, such as hospitalization [8,9], loss of independence [10], and fear of falling, which further restricts physical activity [11-14]. Many individuals fail to recognize the deterioration in their health until after experiencing a fall. Early signs of health decline in older adults are often overlooked

XSL•FO
RenderX

until hospitalization occurs, thereby missing critical opportunities for early intervention [15]. In addition, many older individuals fail to acknowledge the severity of their health decline until after a significant event, such as a fall [16,17]. Thus, early detection of lower body function deterioration allows for personalized interventions, such as physical therapy or the use of mobility aids, which can help mitigate these risks [18-20]. Unfortunately, technologies to monitor lower body function to detect subtle changes in older adults are not available, especially when mobility aids are used.

Wearable sensors have emerged as a widely recognized tool for monitoring physical activity in real time, offering essential data on health and mobility [21,22]. The market currently offers numerous options for activity trackers, including wrist-worn fitness trackers like Fitbit (Google LLC) [23] and smartwatches such as Apple Watch (Apple Inc.) [24] and Samsung Galaxy Watch (Samsung Electronics Co., Ltd.) [25]. These devices measure various metrics such as step counts, heart rate, walking speed, cadence, and posture. Wearable technology can assist in the early detection of mobility issues and predict fall risks, especially through gait pattern and walking dynamic analysis [4,26]. Moreover, in clinical settings, wearable sensors are increasingly used to assess rehabilitation outcomes [27-29], remotely monitor patients [30-32], and manage chronic conditions [33-36], thereby providing continuous data outside traditional health care environments.

However, these algorithms have not been optimized for older adults who rely on walking aids. The altered biomechanics caused by mobility aids such as rollators disrupt natural gait patterns which can impact the accuracy of gait detection. Specifically, older adults tend to walk more slowly and with shorter strides, creating challenges to accurately detect subtle movements for these algorithms developed using younger populations [37]. Furthermore, using mobility aids changes how people walk, especially from the perspective of wrist-worn sensors. For example, arm swings may be absent or reduced and patterns indicating footfalls may be sharpened or dampened. These disruptions in gait patterns frequently cause wearable sensors to misinterpret or underestimate mobility levels, resulting in unreliable analysis for the collected data [38,39]. Consequently, this limits the effectiveness of wearable devices in monitoring activity levels or assessing fall risks in individuals using mobility aids. Although some research has explored this issue [37,40,41], further studies are needed to fully understand the underlying dynamics and address these limitations.

This work focuses on the development and evaluation of a gait recognition algorithm that is robust to the usage of mobility aids. We calibrate a “one-size-fits-most” walking recognition algorithm [42] to be sensitive to the walking patterns of older adults with or without the use of mobility aids.

Methods

Study Design

We conducted a prospective cohort study with 17 older adults living in independent living. The study design focuses on capturing labeled samples of walking behaviors and other activities of daily living (ADLs) in the participant’s normal environment. The other ADLs include brushing teeth, combing hair, drinking, eating, taking medication, transitions, and washing hands. Transitions include sit-to-stand, stand-to-sit, sit-to-lie, lie-to-sit, sitting still, laying still, bending down, and straightening up. Participants were allowed to opt out of performing any activity at their discretion.

Motion data from these activities was collected using a wrist-worn sensor on the dominant hand. This study uses Samsung Galaxy Watch 6 smartwatches (Samsung Electonics Co., Ltd.) [25]. Accelerometer data was sampled at 50Hz using a custom Android application.

Annotation used a seminaturalistic style: participants were given specific activities and were allowed to perform the activities as they would naturally. Study personnel used a mobile phone app to label the occurrence of each event. This labeling app captured exact timestamps that activities started and stopped; these labels were later synchronized with the smartwatch accelerometer data.

Annotation sessions took place in the participant’s apartment or the gym area of the retirement community. Participants were instructed to follow their normal routines for brushing teeth, combing hair, taking medication, and washing hands. Participants were asked to take at least 3 sips for drinking and 3 bites for eating. Likewise, participants were asked to repeat transition behaviors 3 times. Walking events lasted several minutes and involved walking from one location to another within the retirement community (eg, walking from the gym back to the participant’s apartment). Participants were not given any time minimums while performing the activities.

Ethical Considerations

The study protocol was approved by the Texas A&M University Institutional Review Board (IRB2020-1271D). All participants provided informed consent. The study design follows standard procedures for human subjects research to minimize risk to participants. Participants were able to withdraw from the study at any time with no penalty. Data were deidentified during analysis. The study provided no compensation.

Recruitment

Recruitment took place at two retirement communities in the southwestern United States. The research team advertised by giving presentations about the study at each location with permission from each location’s administrative team. Interested residents contacted the research team to join the study.

Algorithm Design

The “one-size-fits-most” algorithm leverages the inherent physical characteristics of taking steps to detect bouts of walking. Walking is defined as sequences of strides (ie, two consecutive steps) and has attributes such as intensity and periodicity. By this definition, walking can be distinguished from other daily activities by searching for windows of sensor data that satisfy these characteristics.

This algorithm searches for 1-second windows, where the accelerometer vector magnitude is above a given threshold, A,

XSL•FO
RenderX

in meters per second-squared. Data is resampled to 10Hz to minimize the data fidelity requirements. Additionally, lower sampling rates reduce the risk of capturing confounding patterns as walking is not a fine-grained activity.

(1)$\alpha * pw > pmin$

(2)$\beta * pw > pmax$

Valid windows are transformed into the frequency domain using a continuous wavelet transform (CWT) to produce C(f,), the decomposition of the signal into scaled, time-shifted wavelets where f is the frequency and τ is the time step. These segments are checked against a series of conditions regarding their peak frequencies. The overall peak is expected to be within the frequency range $f_w=[f_{min}, f_{max}]$. Three peaks within C(f,) are detected: the subharmonic peak p_{min}, the walking range peak p_w, and the superharmonic peak p_{max}. The presence of superharmonics and subharmonics are allowed if the peaks satisfy Equations (1) and (2). α and β are parameters that control the maximum ratios of the peak magnitudes.

These segments are marked as candidates for walking. The final step checks if the sequences of candidate segments are longer than the minimum time threshold, T. This step removes short bursts of activity that are unlikely to be an actual bout of walking, improving the algorithm specificity.

The values for A, α, β, f_{min}, f_{max}, and T come from analyzing walking from 20 existing datasets with a total of 1240 healthy participants, as previously reported [42]. Fifteen of the datasets focus solely on mobility behaviors; the other 5 include other daily activities. Age and sex reporting varied by dataset. Age was reported for 745 participants with range 15 - 75 years (mean 28.6, SD 12). Sex was reported for 901 participants and 72% (n=649) were male. There are two sets of parameters depending on the location of the sensor. For wrist-worn sensors, the default values are A=0.3g, α=31.7, β=1.4, f_{min}=1.4Hz, f_{max}=2.3Hz, T=6 seconds [42].

We calibrate these parameters to reflect the walking patterns of older adults. We conduct hyperparameter tuning and use a leave-one-subject-out (LOSO) to explore multiple views of the data. We use a tree-structured Parzen estimator to explore the parameter space. A had the potential range of 0.05g to 0.5g. α and β had the range of 0 to 100. f_{min} varied from 0.6Hz to 1.5Hz; f_{max} varied from 2Hz to 3Hz. T had the range of 1 to 10 seconds. We used the hyperopt [43] Python package and 500 trials per fold of the LOSO evaluation. F_1-score was used as the optimization metric to capture both sensitivity and precision. The final value of each parameter is determined by averaging across the folds and rounding to one decimal for all parameters except time, which is rounded to a whole number.

(3)sensitivity=TP / (TP+FN)

(4)specificity=TN / (TN+FP)

(5)F1−score=2∗ TP / (2∗ TP+FN+FP)

The algorithm estimates cadence in terms of steps per second. This output is converted to walking versus not-walking annotations by labeling windows with a non-zero number of steps as walking. Algorithm performance is measured through sensitivity (Equation 3), specificity (Equation 4), and the F_1-score (Equation 5). Correctly detected seconds of walking are true positives (TP) and missed seconds are false negatives (FN). Conversely, correct non-walking seconds are true negatives (TN) and the corresponding errors are false positives (FP).

Results

Participants

Table 1 summarizes the demographic information of the recruited participants. All the participants were ambulatory. Some used mobility aids in their daily lives. One participant used a walker; their data is grouped with the rollator-users because they used the walker like it was a rollator .

Table . Participant demographic information organized by mobility aid usage.

Variables	All (N=17)	No aid (n=7)	Cane (n=5)	Rollator/Walker (n=5)
Age (years), n				
70 - 79	5	1	2	2
80 - 89	10	5	3	2
≥90	2	1	0	1
Female, n	12	5	2	5
Right hand dominant, n	16	6	5	5

Parameter Tuning

The parameters from each fold of the LOSO evaluation are averaged to produce the calibrated parameters. The main algorithm parameters needed adjustment for older adults. The threshold A lowered from 0.3g to 0.1g. The parameters α and β increased from 31.7 to 65.4 and 1.4 to 77.1, respectively. The expected frequency range of walking expanded from [1.4Hz, 2.3Hz] to [0.8Hz, 2.8Hz]. The minimum time requirement increased from 6 to 10 seconds. The summary of these parameters is given in Table 2.

XSL•FO
RenderX

Table . Parameters for characterizing gait by algorithm version.

Parameters	Amplitude (A)	α	β	Minimum frequency (f_{min})	Maximum frequency (f_{max})	Time (T)
Default [42]	0.3	31.7	1.4	1.4	2.3	6
Calibrated	0.1	65.4	77.1	0.8	2.8	10

Recognition Performance

We provide sensitivity, specificity, and F_1-score metrics using the default parameters, the average performance using the parameters within the LOSO folds, and using the calibrated parameters (Table 3). The average values and the ranges are reported per metric. The default and LOSO rows are participant-independent results; the calibrated row shows the performance using the average parameters from hyperparameter tuning (Table 2).

Table . Algorithm performance of walking against ADLs[a] organized by mobility aid usage.

Metrics and parameters	Overall	No aid	Cane	Rollator
Sensitivity (range)				
Default	0.11 (0.00 - 0.72)	0.22 (0.01 - 0.72)	0.04 (0.01 - 0.09)	0.01 (0.00 - 0.02)
LOSO[b]	0.74 (0.10 - 1.00)	0.72 (0.10 - 1.00)	0.67 (0.45 - 0.95)	0.85 (0.75 - 0.96)
Calibrated	0.73 (0.10 - 1.00)	0.72 (0.10 - 1.00)	0.64 (0.43 - 0.95)	0.84 (0.74 - 0.93)
Specificity (range)				
Default	0.99 (0.97 - 1.00)	0.99 (0.97 - 1.00)	0.99 (0.98 - 1.00)	1.00 (0.99 - 1.00)
LOSO	0.73 (0.20 - 0.97)	0.68 (0.20 - 0.88)	0.78 (0.56 - 0.97)	0.76 (0.63 - 0.84)
Calibrated	0.75 (0.32 - 0.98)	0.71 (0.32 - 0.90)	0.80 (0.56 - 0.98)	0.76 (0.63 - 0.86)
F_1-score (range)				
Default	0.15 (0.00 - 0.84)	0.31 (0.02 - 0.84)	0.07 (0.01 - 0.17)	0.01 (0.00 - 0.05)
LOSO	0.73 (0.07 - 0.94)	0.64 (0.07 - 0.87)	0.71 (0.52 - 0.83)	0.87 (0.74 - 0.94)
Calibrated	0.73 (0.08 - 0.92)	0.65 (0.08 - 0.89)	0.69 (0.52 - 0.85)	0.87 (0.75 - 0.92)

[a]ADLs: activities of daily living.

[b]LOSO: leave-one-subject-out.

The default parameters have high specificity at 0.99 but poor sensitivity at 0.11. The resulting F_1-score is low at 0.15. Participants who used no mobility aid had better performance than those who used them. Walking detection for cane users was slightly higher than that for rollator users with F_1-scores of 0.07 and 0.01, respectively.

The LOSO results have balanced sensitivity and specificity metrics with both at 0.73. Cane-users had the lowest sensitivity at 0.67; rollator-users had the highest sensitivity at 0.85. Participants with no aid had the lowest specificity at 0.68; cane-users had the highest specificity at 0.78, closely followed by rollator-users at 0.76. No-aid participants also had the lowest F_1-score at 0.64; rollator-users had the highest at 0.87, and cane-users had a score of 0.71.

The calibrated results are comparable to the LOSO results. Most metrics are within 0.05 of the corresponding metrics using the LOSO parameters. Sensitivity is slightly lower and specificity is slightly higher. F_1-score is slightly better for no-aid participants, slightly worse for cane-users, and equal for rollator-users.

Discussion

Principal Results

All the parameters for the algorithm needed adjustment to detect walking for our prospective cohort. The threshold A was lowered to 0.1g, indicating that the steps taken by older adults are softer and more subtle. α and β are higher at 65.4 and 77.1 compared to 31.8 and 1.4, respectively. Subharmonics are dependent on the location of the wearable device, so the increase may just be a side effect from the smartwatch used in this study compared to the original work. Superharmonics can be caused by noise in the signal or environmental factors such as the walking surfaces. β is notably higher; this result implies that the patterns associated with walking are less distinct. The range of f_w is wider on both ends of the spectrum. f_{min} is lowered to 0.8Hz, aligning with the expectation that older adults walk slower. f_{max} increased to 2.8Hz which is surprising because this would imply that the participants walked faster than the average adult. This outcome may be connected to the more lenient values for A, α, and β. Strong peaks generally indicate footfalls. In this case, the footfall peaks may have been softer and surrounded by other peaks (ie, subharmonic and superharmonic patterns), making it appear as though the participant was walking faster.

XSL•FO
RenderX

The reasonable sensitivity of our algorithm suggests the potential for monitoring physical activity. The calibrated algorithm can be used to detect if an older adult has become sedentary. This algorithm is computationally lightweight as a rule-based algorithm that only relies on CWT for data processing; as such, it can be leveraged by standard wearable devices. Additionally, using data from only one arm is sufficient, suggesting that wearing a single arm sensor may be sufficient to detect clinically relevant changes.

The value of β has a notable impact on algorithm performance. We conducted a small ablation study to determine which parameters contributed to the higher false positive rates for some ADLs. Lower values for β reduced the false positive rate, but the true positive rate for the algorithm also decreased in turn, bringing performance back in line with the default parameters This outcome implies that the superharmonics in the walking samples are distinguishable characteristics. The superharmonic peaks likely are an artifact from the back-and-forth motions present in walking and several ADLs such as brushing teeth and washing hands.

Comparison with Prior Work

This study addresses two common gaps in the gait recognition literature: inclusion of older adults as participants and inclusion of people using mobility aids. Furthermore, this study provides initial results into future work identified by the original work that proposed the “one-size-fits-most” algorithm [42].

Limitations

A limitation of this study is the small sample size in terms of unique participants and dataset size in terms of seconds of sensor data. More data would improve the strength of the results. Future work should both increase the number of participants and the length of the data collection sessions. Additionally, future evaluations should take place in free-living environments where participants are allowed to behave naturally. These studies may focus on assessing the effectiveness of this algorithm as a sedentariness level tracker.

Conclusion

This study successfully recognized the walking patterns of older adults with or without mobility aids. The performance metrics show that this algorithm has promise for being used to monitor physical activity. This approach is computationally lightweight and explainable because it only relies on the CWT computationally and the algorithm design is rule-based in nature. Our calibration approach can be adopted to tune to new populations and has a low barrier to entry for adopting a new technology due to the sole reliance on accelerometer data which is a standard sensor in wearable devices. The most noteworthy parameter adjustment is the ratio for superharmonics. The lower values for these ratios in the default parameters were a key reason why it could not detect walking in our older adult data. The calibrated parameters were more lenient to peaks outside the expected frequency range of walking. This outcome increases sensitivity while also increasing risk of false positives, hence the decrease in specificity.

We validated the algorithm on ten mobility aid users (ie, 5 cane, 4 rollator, and 1 walker). A larger study with more participants using mobility aids is necessary to conduct a deeper analysis on what parameters work best for this population. Future work includes validating the algorithm’s ability to estimate step counts and measure physical activity in real-world settings.

Acknowledgments

This research was funded by Seedling Grant Proposal-291007-23026 from Texas A&M Health (Bryan, TX).

Data Availability

The dataset curated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: JIK, SJR
Investigation: AML, SJR
Methodology: SJR
Project Administration: PKS
Supervision: TAH, PKS
Writing – Original Draft: JIK, SJR
Writing – Review & Editing: SJR, PKS

Conflicts of Interest

None declared.

References

1. Tinetti ME, Williams CS. Falls, injuries due to falls, and the risk of admission to a nursing home. N Engl J Med 1997 Oct 30;337(18):1279-1284. [doi: 10.1056/NEJM199710303371806] [Medline: 9345078]
2. Tinetti ME, Williams CS. The effect of falls and fall injuries on functioning in community-dwelling older persons. J Gerontol Ser A 1998 Mar 1;53A(2):M112-M119. [doi: 10.1093/gerona/53A.2.M112]

XSL•FO
RenderX

3. Gill TM, Desai MM, Gahbauer EA, Holford TR, Williams CS. Restricted activity among community-living older persons: incidence, precipitants, and health care utilization. Ann Intern Med 2001 Sep 4;135(5):313-321. [doi: 10.7326/0003-4819-135-5-200109040-00007] [Medline: 11529694]
4. Chen M, Wang H, Yu L, et al. A systematic review of wearable sensor-based technologies for fall risk assessment in older adults. Sensors (Basel) 2022 Sep 7;22(18):6752. [doi: 10.3390/s22186752] [Medline: 36146103]
5. Osoba MY, Rao AK, Agrawal SK, Lalwani AK. Balance and gait in the elderly: a contemporary review. Laryngoscope Investig Otolaryngol 2019 Feb;4(1):143-153. [doi: 10.1002/lio2.252] [Medline: 30828632]
6. US Preventive Services Task Force. Interventions to prevent falls in community-dwelling older adults: US Preventive Services Task Force Recommendation Statement. JAMA 2018;319(16):1696-1704. [doi: 10.1001/jama.2018.3097]
7. Hu G, Baker SP. Recent increases in fatal and non-fatal injury among people aged 65 years and over in the USA. Inj Prev 2010 Feb;16(1):26-30. [doi: 10.1136/ip.2009.023481] [Medline: 20179032]
8. Alexander BH, Rivara FP, Wolf ME. The cost and frequency of hospitalization for fall-related injuries in older adults. Am J Public Health 1992 Jul;82(7):1020-1023. [doi: 10.2105/ajph.82.7.1020] [Medline: 1609903]
9. Scuffham P, Chaplin S, Legood R. Incidence and costs of unintentional falls in older people in the United Kingdom. J Epidemiol Community Health 2003 Sep;57(9):740-744. [doi: 10.1136/jech.57.9.740] [Medline: 12933783]
10. Kelsey JL, Procter-Gray E, Berry SD, et al. Reevaluating the implications of recurrent falls in older adults: location changes the inference. J Am Geriatr Soc 2012 Mar;60(3):517-524. [doi: 10.1111/j.1532-5415.2011.03834.x] [Medline: 22283236]
11. Jefferis BJ, Iliffe S, Kendrick D, et al. How are falls and fear of falling associated with objectively measured physical activity in a cohort of community-dwelling older men? BMC Geriatr 2014 Oct 27;14(1):114. [doi: 10.1186/1471-2318-14-114] [Medline: 25348492]
12. Austin N, Devine A, Dick I, Prince R, Bruce D. Fear of falling in older women: a longitudinal study of incidence, persistence, and predictors. J Am Geriatr Soc 2007 Oct;55(10):1598-1603. [doi: 10.1111/j.1532-5415.2007.01317.x] [Medline: 17908062]
13. Murphy SL, Williams CS, Gill TM. Characteristics associated with fear of falling and activity restriction in community-living older persons. J Am Geriatr Soc 2002 Mar;50(3):516-520. [doi: 10.1046/j.1532-5415.2002.50119.x] [Medline: 11943049]
14. Mendes da Costa E, Pepersack T, Godin I, Bantuelle M, Petit B, Levêque A. Fear of falling and associated activity restriction in older people. results of a cross-sectional study conducted in a Belgian town. Arch Public Health 2012 Jan 3;70(1):1. [doi: 10.1186/0778-7367-70-1] [Medline: 22958732]
15. Hoffman GJ, Tinetti ME, Ha J, Alexander NB, Min LC. Prehospital and posthospital fall injuries in older US adults. JAMA Netw Open 2020 Aug 3;3(8):e2013243. [doi: 10.1001/jamanetworkopen.2020.13243] [Medline: 32822491]
16. Kubitza J, Haas M, Keppeler L, Reuschenbach B. Therapy options for those affected by a long lie after a fall: a scoping review. BMC Geriatr 2022 Jul 15;22(1):582. [doi: 10.1186/s12877-022-03258-2] [Medline: 35840883]
17. Delbaere K, Close JCT, Brodaty H, Sachdev P, Lord SR. Determinants of disparities between perceived and physiological risk of falling among elderly people: cohort study. BMJ 2010 Aug 18;341(aug18 4):c4165-c4165. [doi: 10.1136/bmj.c4165]
18. Narayanan MR, Scalzi ME, Redmond SJ, Lord SR, Celler BG, Lovell NH. Evaluation of functional deficits and falls risk in the elderly – methods for preventing falls. Presented at: 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society; Sep 3-6, 2009; Minneapolis, MN, USA p. 6179-6182. [doi: 10.1109/IEMBS.2009.5334506]
19. Myers AH, Young Y, Langlois JA. Prevention of falls in the elderly. Bone 1996 Jan;18(1 Suppl):87S-101S. [doi: 10.1016/8756-3282(95)00384-3] [Medline: 8717552]
20. Rajendran P, Corcoran A, Kinosian B, Falls AM. Falls, fall prevention, and fall detection technologies. In: Alwan M, Felder RA, editors. Eldercare Technology for Clinical Practitioners 2008:187-202. [doi: 10.1007/978-1-59745-233-5_8]
21. Phillips SM, Cadmus-Bertram L, Rosenberg D, Buman MP, Lynch BM. Wearable technology and physical activity in chronic disease: opportunities and challenges. Am J Prev Med 2018 Jan;54(1):144-150. [doi: 10.1016/j.amepre.2017.08.015]
22. Bunn JA, Navalta JW, Fountaine CJ, Reece JD. Current state of commercial wearable technology in physical activity monitoring 2015-2017. Int J Exerc Sci 2018;11(7):503-515. [doi: 10.70252/NJQX2719] [Medline: 29541338]
23. Fitbit charge 6. Google Store. URL: https://store.google.com/product/fitbit_charge_6?hl=en-US [accessed 2025-07-27]
24. Apple watch series 10. Apple. URL: https://www.apple.com/apple-watch-series-10/ [accessed 2025-07-27]
25. Samsung galaxy watch 6. Wikipedia. 2025. URL: https://en.wikipedia.org/w/index.php?title=Samsung_Galaxy_Watch_6&oldid=1299887008 [accessed 2025-07-27]
26. Urbanek JK, Roth DL, Karas M, et al. Free-Living gait cadence measured by wearable accelerometer: a promising alternative to traditional measures of mobility for assessing fall risk. J Gerontol Ser A 2023 May 11;78(5):802-810. [doi: 10.1093/gerona/glac013]
27. Lugrís U, Carlín J, Luaces A, Cuadrado J. Gait analysis system for spinal cord-injured subjects assisted by active orthoses and crutches. Proceedings of the Institution of Mechanical Engineers, Part K: Journal of Multi-body Dynamics 2013 Dec;227(4):363-374. [doi: 10.1177/1464419313494935]
28. Sharif Bidabadi S, Tan T, Murray I, Lee G. Tracking foot drop recovery following lumbar-spine surgery, applying multiclass gait classification using machine learning techniques. Sensors (Basel) 2019 Jun 4;19(11):2542. [doi: 10.3390/s19112542] [Medline: 31167372]

XSL•FO
RenderX

29. Lee J, Bae CH, Jang A, Yang S, Bae H. Determining the most appropriate assistive walking device using the inertial measurement unit-based gait analysis system in disabled patients. Ann Rehabil Med 2020 Feb;44(1):48-57. [doi: 10.5535/arm.2020.44.1.48] [Medline: 32130838]
30. Sharif Bidabadi S, Murray I, Lee GYF. The application of inertial measurements unit for the clinical evaluation and assessment of gait events. J Med Eng Technol 2017 Nov;41(8):612-622. [doi: 10.1080/03091902.2017.1382588] [Medline: 28978243]
31. Fanning J, Miller ME, Chen SH, Davids C, Kershner K, Rejeski WJ. Is wrist accelerometry suitable for threshold scoring? A comparison of hip-worn and wrist-worn actigraph data in low-active older adults with obesity. J Gerontol Ser A 2022 Dec 29;77(12):2429-2434. [doi: 10.1093/gerona/glab347]
32. Rast FM, Labruyère R. Systematic review on the application of wearable inertial sensors to quantify everyday life motor activity in people with mobility impairments. J NeuroEngineering Rehabil 2020 Dec;17(1):148. [doi: 10.1186/s12984-020-00779-y]
33. Vienne A, Barrois RP, Buffat S, Ricard D, Vidal PP. Inertial sensors to assess gait quality in patients with neurological disorders: a systematic review of technical and analytical challenges. Front Psychol 2017;8:817. [doi: 10.3389/fpsyg.2017.00817] [Medline: 28572784]
34. Maetzler W, Domingos J, Srulijes K, Ferreira JJ, Bloem BR. Quantitative wearable sensors for objective assessment of Parkinson's disease. Mov Disord 2013 Oct;28(12):1628-1637. [doi: 10.1002/mds.25628] [Medline: 24030855]
35. Bianchini E, Caliò B, Alborghetti M, et al. Step-counting accuracy of a commercial smartwatch in mild-to-moderate PD patients and effect of spatiotemporal gait parameters, laterality of symptoms, pharmacological state, and clinical variables. Sensors (Basel) 2023;23(1):214. [doi: 10.3390/s23010214]
36. Wang Y, Mukaino M, Ohtsuka K, et al. Gait characteristics of post-stroke hemiparetic patients with different walking speeds. Int J Rehabil Res 2020 Mar;43(1):69-75. [doi: 10.1097/MRR.0000000000000391] [Medline: 31855899]
37. Vargemidis D, Gerling K, Geurts L, Vanden Abeele V. Flexible activity tracking for older adults using mobility aids — an exploratory study on automatically identifying movement modality. Presented at: Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility ASSETS '22; Oct 23, 2022; Athens Greece p. 1-5. [doi: 10.1145/3517428.3550371]
38. Korpan SM, Schafer JL, Wilson KCS, Webber SC. Effect of ActiGraph GT3X+ position and algorithm choice on step count accuracy in older adults. J Aging Phys Act 2015 Jul;23(3):377-382. [doi: 10.1123/japa.2014-0033] [Medline: 25102469]
39. Alinia P, Cain C, Fallahzadeh R, Shahrokni A, Cook D, Ghasemzadeh H. How accurate is your activity tracker? A comparative study of step counts in low-intensity physical activities. JMIR Mhealth Uhealth 2017;5(8):e106. [doi: 10.2196/mhealth.6321]
40. Matthies DJC, Haescher M, Nanayakkara S, Bieber G. Step detection for rollator users with smartwatches. Presented at: Proceedings of the 2018 ACM Symposium on Spatial User Interaction; Oct 13, 2018; Berlin Germany p. 163-167. [doi: 10.1145/3267782.3267784]
41. Bieber G, Haescher M, Hanschmann P, Matthies DJC. Exploring accelerometer-based step detection by using a wheeled walking frame. Presented at: Proceedings of the 5th International Workshop on Sensor-Based Activity Recognition and Interaction; Sep 20, 2018; Berlin Germany p. 1-6. [doi: 10.1145/3266157.3266212]
42. Straczkiewicz M, Huang EJ, Onnela JP. A "one-size-fits-most" walking recognition method for smartphones, smartwatches, and wearable accelerometers. NPJ Digit Med 2023 Feb 23;6(1):29. [doi: 10.1038/s41746-022-00745-z] [Medline: 36823348]
43. Bergstra J, Yamins D, Cox D. Making a science of model search: hyperparameter optimization in hundreds of dimensions for vision architectures. Presented at: Proceedings of the 30th International Conference on Machine Learning; Jun 16, 2013; Atlanta, GA, USA p. 115-123 URL: https://proceedings.mlr.press/v28/bergstra13.html [accessed 2025-07-27]

Abbreviations

ADL: Activity of Daily Living
CWT: Continuous Wavelet Transform
FN: False Negative
FP: False Positive
LOSO: Leave-One-Subject-Out
TN: True Negative
TP: True Positive

XSL•FO
RenderX

Edited by A Mavragani; submitted 11.11.24; peer-reviewed by B Cahill, G Mascia; revised version received 18.10.25; accepted 20.10.25; published 10.11.25.

Please cite as:
Ray SJ, Koh JI, Liberty AM, Hammond TA, Shireman PK
Designing a Gait Recognition Algorithm for Older Adults Using Mobility Aids: Prospective Cohort Study
JMIR Form Res 2025;9:e68669
URL: https://formative.jmir.org/2025/1/e68669
doi:10.2196/68669

© Samantha Jeane Ray, Jung In Koh, Amanda Mae Liberty, Tracy Anne Hammond, Paula Kay Shireman. Originally published in JMIR Formative Research (https://formative.jmir.org), 10.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Digital Smoking Cessation Preferences of Predominately Low-Income and Latino Residents of the San Joaquin Valley in California: Qualitative Study

Karla D Llanes[1,2], PhD; Maya Vijayaraghavan[1,3,4], MAS, MD; Sara Schneider[5], PhD; Pamela M Ling[1,4], MPH, MD; Evi Hernandez[6], BA; Paul Brunetta[7], MD; Anna V Song[5,8], PhD; Arturo Durazo[5,9], PhD

[1]Center for Tobacco Control Research and Education, University of California San Francisco, San Francisco, CA, United States

[2]Pharmaceutical Sciences, School of Pharmacy, The University of Texas at El Paso, El Paso, TX, United States

[3]Smoking Cessation Leadership Center, Division of General Internal Medicine, University of California San Francisco, 490 Illinois Street #83C, San Francisco, CA, United States

[4]Division of General Internal Medicine, University of California San Francisco, San Francisco, CA, United States

[5]Nicotine and Cannabis Policy Center, Health Sciences Research Institute, University of California Merced, Merced, CA, United States

[6]California Health Collaborative, Fresno, CA, United States

[7]Fontana Tobacco Treatment Center, University of California San Francisco, San Francisco, CA, United States

[8]Psychological Sciences, School of Social Sciences, Humanities and Arts, University of California Merced, Merced, CA, United States

[9]Public Health, School of Social Sciences, Humanities and Arts, University of California Merced, Merced, CA, United States

Corresponding Author:
Maya Vijayaraghavan, MAS, MD
Center for Tobacco Control Research and Education, University of California San Francisco, San Francisco, CA, United States

Abstract

Background: Although rates of tobacco use in California have declined overall, adults in the San Joaquin Valley (SJV), particularly Hispanic or Latinos ("Latinos"), have disproportionately high rates of tobacco use, tobacco-related illness, and mortality. Residents of the SJV also have limited access to cessation support services and need accessible, nonclinical alternatives. Given high smartphone use rates among Latinos and residents of rural communities, digital health tools may present an accessible approach to expand cessation support.

Objective: This study explored tobacco use behaviors, cessation experiences, and views about digital cessation tools for tobacco cessation among SJV residents. The secondary objective was to assess the appeal, usability, and necessary adaptations of 2 existing digital smoking cessation tools—a smoking cessation app and a social media–based cessation intervention.

Methods: Through an SJV-based academic-community partnership, we recruited 29 predominantly Latino adults who reported current smoking. We conducted 4 focus group discussions to explore tobacco use and cessation experiences and preferences for smoking cessation tools: 1 in-person in English, 1 online in English, and 2 online in Spanish. Subsequently, 9 participants from the focus group discussions completed individual, in-depth interviews where they viewed videos describing 2 digital smoking cessation tools—a cessation app and a social media cessation intervention—to assess their appeal and usability. Focus groups and interviews were recorded, transcribed, and analyzed to identify themes.

Results: Overall, 82.1% (23/28) had made a quit attempt in the past year, and most intended to quit smoking in the next 6 months, with 11.1% (3/27) never expecting to quit. Most participants were motivated to quit despite experiencing barriers, and they emphasized the need for culturally tailored digital cessation tools to help overcome the barriers to quitting smoking. They preferred interventions that integrated culturally relevant content reflecting lived experiences, featured language-concordant communications, and provided social supports, such as chat rooms for peer connection. Participants reported polyuse of tobacco with other substances, including cannabis, which may need to be addressed when delivering smoking cessation interventions. While participants appreciated the app's private interface and comprehensive curriculum, they preferred the social media–based program for its engaging design, despite privacy concerns. Preferences for specific interventions varied by age and digital literacy. Participants also expressed preference for material rewards to incentivize the use of digital health tools to quit smoking.

Conclusions: This sample of predominantly Latino adults from the SJV expressed favorable interest in digital cessation support, yet existing tools require adaptation to improve cultural relevance, accessibility, usability, and privacy concerns. Participants emphasized language-concordant services, representation from people with lived experience, and community-building features. While digital interventions were well received, privacy concerns and digital literacy barriers must be addressed to enhance engagement.

XSL•FO
RenderX

(JMIR Form Res 2025;9:e74105) doi:10.2196/74105

KEYWORDS

mHealth; digital health; smoking cessation; e-cigarette use; smoking; Hispanic; Latino; mobile phone

Introduction

Background

Cigarette smoking is the leading cause of preventable death in the United States, with more than 490,000 smoking-related deaths each year [1]. Smoking is associated with detrimental effects to the cardiovascular system, respiratory system, and other health risk outcomes [1]. While overall smoking prevalence has declined in the United States, disparities persist across racial and ethnic groups, particularly among those with low socioeconomic status [2]. For instance, smoking rates are lowest among Asian (5%) and Latino (8%) adults [2] and highest among American Indian Alaska Native (AIAN; 19.3%) adults [2]. However, aggregate rates for Latino adults mark variations. For example, Cuban and Puerto Rican adults report some of the highest rates among Latinos (up to 30%) [3]. African American [4] and Latino adults are heavily targeted in the sale of menthol cigarettes, with a higher density of smoke shops that sell menthol cigarettes in African American or Black and Latino neighborhoods [5,6]. For Latinos, the stress of acculturation to US culture [7] and experiencing discrimination increases the risk of smoking [8,9]. These intersecting factors contribute to tobacco-related disparities across racial or ethnic groups.

While California is recognized as a national leader in tobacco control, with some of the most comprehensive policies and lowest smoking rates in the United States, rural regions maintain disproportionately higher smoking rates [10]. California's San Joaquin Valley (SJV), a predominantly rural and agricultural region, exemplifies this disparity, with a tobacco use prevalence of 15.6% in 2019 [11], exceeding the state average of 11.4% in 2022 [1]. Home to more than 4 million people across 8 counties (Stanislaus, San Joaquin, Merced, Madera, Fresno, Kings, Tulare, and Kern) [12], SJV includes many structurally marginalized and under-resourced communities. Nearly 45% of SJV residents identify as Latino [12], and 70% of immigrant residents originate from Central America and Mexico [13]. Latinos in California experience a disproportionate burden from tobacco use, with 1.2 million Latinos reporting current use of tobacco, second to Whites at 1.4 million, followed by 0.3 million African Americans and less than 0.1 million AIANs [11].

The SJV region is considered a medical desert, with severe shortages of health care providers, limiting access to health care, specialty services, and tobacco cessation resources. Low literacy skills and limited English proficiency [13] also make it more challenging to navigate health care systems and access cessation support. The SJV also has one of the highest poverty rates in California, with 20% of residents in most counties living below the federal poverty level, compared with 12% statewide [12].

Although Latinos in California attempt to quit smoking at higher rates than the general population (56.5% vs 55.1%), they are less likely to use evidence-based cessation treatment such as nicotine replacement therapy (NRT) and receive cessation advice from health care providers [14]. In 2023, of the 21,500 people who used the free cessation program known as Kick It California, which includes the state Quitline, text, web, and chat support, only 21.1% identified as Latino, while 37.7% identified as non-Hispanic White [10]. Latinos face disparities in access to health care and health insurance [14], and 46.7% of Latino adults who smoked were advised to quit versus 49.9% of all California adults [10], which may contribute to underutilization of cessation supports. Other reasons for not using cessation supports include misperceptions about the addictiveness of NRT [15], hesitation to use support because of the cultural belief that using willpower is a way to take personal responsibility for quitting [16], or using both cigarettes and e-cigarettes (8.1% of Latinos use both vs 8.8% non-Hispanic White) [17], which might make it more challenging to quit [18]. Given these disparities, there is an urgent need for culturally responsive and affirming smoking cessation interventions for Latinos in the SJV. Expanding cessation resources beyond traditional clinical settings—particularly through digital health solutions—may help address gaps in cessation services for Latinos who smoke and face significant barriers to treatment.

While expanding access to health care is critical for increasing cessation treatment utilization, many Latino adults who smoke face persistent barriers such as lack of insurance, limited access to health care providers, and language barriers. These challenges are exacerbated in rural regions such as California's SJV, where access to health care and cessation services is highly limited. Racial and ethnic minorities in rural areas have limited resources to use health care services [19]. Compared with non-Hispanic Whites (15%), 23.1% of Latinos, 24.5% of non-Hispanic Blacks, and 19.1% of AIANs [1,2] could not see a physician due to elevated cost [19]. Digital cessation tools offer a promising and cost-effective alternative, providing a scalable approach to reaching those who may otherwise lack access to cessation support.

Latinos in the United States have among the highest rates of smartphone adoption (93%) [20]—often using smartphone devices as their primary means of web-based access—compared with 91% of non-Latino Whites. In addition, 95% of Latinos use the web [21] and 84% [22] engage with social media platforms, making digital health interventions a viable strategy for smoking cessation [23-25]. A systematic review of 60 studies found that digital interventions, including mobile apps, text messaging, and web-based programs, significantly increased smoking abstinence compared with printed materials for adult smokers (risk ratios=1.32 for prolonged abstinence and 1.14 for point prevalence abstinence) [26]. These findings highlight the potential of leveraging mobile and social media–based interventions to improve cessation support among Latinos, particularly in hard-to-reach, rural communities such as the SJV.

Social media interventions may also enhance engagement by leveraging existing social networks. For example, the Tobacco

XSL•FO
RenderX

Status Project, a 90-day Facebook smoking cessation intervention, delivered evidence-based counseling to young adults in group meetings on Facebook [27]. Social media interventions have been studied in national samples of young adults [27-30] and can be adapted to users' preferences. Social media–based interventions, whether delivered through existing platforms such as Facebook, Instagram, and WhatsApp, or through specialized platforms, have been shown to be efficacious in increasing abstinence compared with control conditions [31].

Smoking cessation smartphone apps when used in combination with pharmacotherapy may also be an effective aid [32]. A 2023 meta-analysis of randomized controlled trials compared continuous smoking abstinence rates between young adults receiving smartphone text cessation support messages and those receiving only self-help materials or referrals to the Quitline (ie, controls) and found an overall increase in abstinence among those receiving SMS text messaging interventions [33]. The subgroup analysis (k=5) showed that continuous abstinence rates were higher in the SMS text messaging interventions than in control groups at 1-month, 3-month, and 6-month follow-ups (relative risks=1.90, 1.64, and 1.35; P<.05; respectively); only 2 apps reported continuous abstinence outcomes [33]. Since most apps have not been evaluated in clinical trials or published in scientific literature, it makes determining the efficacy of cessation apps largely inconclusive [33-35]. A 2021 search in the iOS App Store and Android App Store generated 228 apps for smoking cessation, and evaluations of each app revealed that very few apps are evidence-based, and 53% were inaccessible due to the financial costs of using the app after downloading [34]. However, 25% of the apps were available in Spanish, signaling potential interest in Spanish language apps [34]; however, we are not aware of any apps developed specifically for lower-income Latino adults.

The high prevalence of tobacco use, health care shortages, economic hardship, and linguistic barriers creates significant obstacles to smoking cessation in the SJV, particularly for Latino residents. Without accessible, culturally tailored cessation interventions, Latinos in the region remain at heightened risk for tobacco-related illnesses [36]. Although the adoption of digital technology is growing and digital cessation interventions may address these health disparities, little is known whether SJV residents would be interested in using digital cessation interventions to assist with smoking cessation. This study aimed to explore tobacco use behaviors, cessation experiences, and views about digital cessation tools for tobacco cessation among SJV residents.

Theoretical Frameworks

To guide this study, we applied a widely used framework for identifying behavioral determinants across individual, social, and environmental contexts, the Capability, Opportunity, Motivation–Behavior (COM-B) model [37]. The COM-B model posits that behavior is the result of interactions between individual capability (psychological and physical), opportunity (social and environmental), and motivation [37]. The model's 3 components interact to either enable or constrain behavior and can serve to inform intervention design and health communications [37]. In this study, the COM-B model was used to identify barriers to and facilitators of quitting smoking and engaging with mobile health platforms.

We also drew from constructs from behavioral change theories to contextualize our secondary objective, evaluating the acceptability and usability of 2 digital cessation tools. For example, the Health Belief Model informed our prompts about perceived risks of smoking and benefits of quitting [38]. In addition, 2 smoking cessation tools were presented to participants—1 mobile app and 1 social media–based cessation intervention—both developed from a combination of these behavioral models and evidence-based cessation strategies. For example, Social Cognitive Theory, we considered constructs relevant to digital interventions when presenting app features (eg, self-regulation, self-efficacy, and model learning) [39]. Both digital tools incorporated social support that can help people model behaviors to avoid smoking and teaching self-regulation. The secondary aim of our study was to explore people's perceptions of these digital tools, particularly their appeal, usability, and suggested adaptations.

Methods

Study Design

In partnership with the California Health Collaborative (CHC), a leading, nonprofit community-based organization in the SJV, we recruited 29 adults from the SJV to participate in the focus group sessions. Coauthor EH, CHC's senior director of programs for the past 22 years, played a key role in conducting the study. He led participant recruitment across the different counties in the SJV by leveraging his connections with community organizations, which helped identify interested participants. Participants also used word of mouth to let people they know about the study. These strategies allowed us to fulfill our recruitment goals. Eligible participants were aged 18 years or older who reported cigarette smoking in the past 30 days. Participants did not know the researchers prior to the study. We developed focus groups and semistructured interview guides with input from CHC, the study investigators, and undergraduate students proficient in Spanish.

Data Collection Procedures

Brief Survey

We first invited participants to complete a brief survey, available in English or Spanish, which collected demographic characteristics and smoking behaviors. Measures included age, gender identity, race and ethnicity, preferred language and English proficiency, tobacco use history and current behavior, time to first cigarette, quitting intentions, and quitting experiences [40]. To assess financial insecurity, we asked participants whether they were able to save money each month (yes/no), had 3 months of savings to cover an expense, and owned a home. Following the survey, participants were invited to take part in a focus group.

Focus Groups

We conducted focus groups using an adapted discussion guide from existing studies to explore participants' tobacco use, experiences with quit attempts, motivation for quitting, prior

XSL•FO
RenderX

use of cessation aids, and resources. We also elicited interest and desired features of digital cessation tools. The focus group discussion used open-ended questions to allow participants to provide their general interest and desired features of digital cessation tools. Each discussion session lasted approximately an hour and began with a brief explanation of the study purpose, ground rules, and an icebreaker. Using open-ended questions, we encouraged participants to discuss smoking and experiences with tobacco use, motivation and experiences with quitting, and their views about resources and cessation aids, including web-based apps to quit smoking.

Individual Interviews

Self-selected participants who completed a focus group were subsequently invited to participate in an individual interview. Each 1-hour interview included the presentation of two 10‑15-minute demonstration videos: one describing an app to quit smoking called "MO" [24], and the other presenting a vaping cessation group intervention on Instagram called "Quit the Hit" [25]. The original demonstration videos were created in English by the developers of each tool (PB for the "MO" and PML for "Quit the Hit"). Spanish language versions were translated and narrated by AD (MO) and KDL (Quit the Hit). After viewing each video, participants were interviewed (KDL and SS) about their general impressions, including what they liked and disliked about each tool, and their preferences for one, both, or neither. During the interviews, participants identified additional appealing and unappealing features they might not have generated on their own during the focus groups.

MO: The MO app for smoking cessation creates an individual profile, tracks smoking, and suggests activities to help users prepare to quit [41]. It includes gamification awarding points for desired behavior, provides badges to recognize quitting milestones, offers daily check-ins, social chat, and video stories to support motivation [41]. The MO app was developed using the US Clinical Practice guidelines, including the use of the 5 As framework (ie, ask, advise, assess, assist, and arrange) and cognitive behavioral therapy techniques [41].

Quit the Hit: Quit the Hit intervention is a person-centered 5-week social media cessation intervention delivered over Instagram [25], adapted from a Facebook smoking cessation program [28,29,42] that increased abstinence at the end of treatment (3 months) compared with referral to a smoking cessation website [27]. Participants join support groups moderated by a cessation counselor with weekday posts supporting preparation, skill building, a group quit date, and social supports to quit. The intervention is for a person ready to make a behavior change, developed using the US Clinical Practice Guidelines for smoking cessation and Social Cognitive Theory. Similarly to the MO app, Quit the Hit uses user-centered, evidence-based intervention content adapted from cognitive behavioral theory and motivational interviewing (eg, increasing self-efficacy and commitment to make a change). Participants were informed that Quit the Hit was adapted from a smoking cessation intervention and could be adapted to address other tobacco use.

Procedures

Study Procedures for the Focus Groups and Semistructured Interviews

We obtained informed consent prior to conducting all study procedures. AD led the initial round of informed consent in both English and Spanish before administering the survey. All participants read and provided informed consent before proceeding with the survey, focus groups, and interviews. Participants were identified through a self-generated study ID number to protect their confidentiality.

Survey Administration

First, participants completed the brief survey assessing their tobacco history, interest in quitting, and past methods used for smoking cessation. Participants in the web-based focus groups completed the brief surveys on Qualtrics prior to the focus group session. For those who had not completed the survey in advance, a few minutes were set aside at the beginning of the focus group session to do so. Participants in the in-person focus group completed a paper version of the survey before the focus group session began.

Focus Group Protocol: Needs Assessment for Digital Smoking Cessation Support

After completing the survey, 29 participants participated in a focus group discussion about their tobacco use behaviors, cessation experiences, and views about digital cessation tools for tobacco cessation. One participant did not want to fill out information about their drug use or participate in the focus group because of legal concerns. Focus groups took place in-person or via Zoom (Zoom Communications, Inc.), in English or Spanish, depending on the participants' preference. Three focus groups were conducted on the web (2 in Spanish and 1 in English), and 1 focus group was conducted in person (in English). There were 8 participants in each of the 2 web-based Spanish focus groups, 5 in the web-based English focus group, and 8 in the in-person English focus group. The focus groups lasted 60‑90 minutes and were led by moderators (AD and KDL) accompanied by notetakers. Authors AD and KDL, who are proficient in English and Spanish, moderated the focus group discussions. The focus groups were conducted from March 2023 to April 2023. An external professional transcription company (Focus Forward) transcribed the English and Spanish audio of focus group discussions.

Semistructured Interview Protocol: Appeal and Usability of Digital Smoking Cessation Tools

We reached out to all 29 participants who participated in the focus group and invited them to participate in a follow-up semistructured interview. In total, only 9 were available and agreed to participate in the 1-hour follow-up interview. Three (1 female and 2 males) interviews were conducted in English and 6 (4 females and 2 males) were conducted in Spanish. KDL and SS conducted the semistructured interviews between November 2023 and February 2024. An external professional transcription company (GMR Transcription) transcribed the audio of interviews.

XSL•FO
RenderX

Data Analysis

Audio files were transcribed verbatim and translated from Spanish to English and uploaded to ATLAS.ti (version 24; Scientific Software Development GmbH). Spanish-speaking researchers (AD and KDL) read the Spanish-transcribed and translated files to identify errors in the transcription process. Two coders (KDL and SS) read each transcript, coded them, and identified themes. We used thematic analysis (inductive approach) to identify individuals' needs for smoking cessation digital tools to facilitate their quitting process [43,44]. The codebook was refined as new themes were identified during the coding process. Discrepancies in coding were discussed and resolved by consensus, including discussion with a third coder (AD), if necessary.

Positionality

We paid particular attention to ensure that participants fully understood the nature of the research and their rights, given potential language barriers and vulnerabilities. We began each focus group with icebreakers and an introduction of ourselves as researchers, along with an explanation of the study's purpose. We also acknowledge our own positionalities as researchers and how they may have influenced the research process. AD is a health communication researcher and is a second-generation Mexican American Californian who grew up with Spanish as his first language, and KDL is a bilingual (ie, first language was Spanish), Latina public health postdoctoral scholar with roots in the Southwest. Our shared cultural and linguistic backgrounds contributed to trust-building with participants, particularly in communications requiring sensitivity to language access and cessation support experiences among racial or ethnic minority and Spanish-speaking individuals. The bilingual team conducting the focus groups and interviews ensured that information was not lost in translation because participants freely expressed themselves in their preferred language. We reviewed the focus group and interview guides with research interns who brought lived experience as SJV residents. In addition, to ensure cultural and linguistic appropriateness, our community partner and coauthor, EH, who grew up in the region, reviewed and provided feedback on the Spanish translations of the consent form, survey, and focus group guide. This partnership helped ensure that research materials were accessible and culturally relevant for SJV residents.

Ethical Considerations

This study was reviewed and approved by the University of California, San Francisco (IRB 22‑37978) and University of California, Merced (IRB UCM2022-119) institutional review boards. All participants read and provided informed consent on Zoom or in person. Researchers reassured participants of confidentiality. Participants were also offered a printed or electronic copy of the informed consent form. The researchers deidentified surveys, focus groups, and interviews by using a self-generated ID number for each participant. We used the COREQ (Consolidated Criteria for Reporting Qualitative Research) to describe and present the research findings to fit standard qualitative reporting. We attest that we did not use generative AI to write any portion of the manuscript or for data analyses. Participants each received a US $50 gift card for participating in the focus group and an additional US $50 gift card for participating in an in-depth interview.

Results

Sample Characteristics

Although 29 participants participated in the focus groups, not all completed all survey items. Participants were permitted to skip questions they did not want to answer. The majority of participants were Latino (24/27, 88.9%), 51.9% (14/27) self-identified as female, and mean average age was 41 years (SD 14.04, range 21‑77). Most (14/26, 53.8%) preferred Spanish language, 38.4% (10/26) reported limited English proficiency, and 63% (17/27) reported that they were unable to save money each month (Table 1). Those who preferred Spanish were predominantly from Madera County, older, preferred smoking cigarettes, and were less comfortable with technology. Focus group participants from Fresno County were younger, English-speaking, primarily used e-cigarettes, and were comfortable with technology. Participants from Merced County were older, preferred English language, most smoked cigarettes, and the focus group was conducted in person. Overall, 82.1% (23/28) had made a quit attempt in the past year, and most intended to quit smoking in a month to 6 months, with 11.1% (3/27) never expecting to quit.

XSL•FO
RenderX

Table . Demographic characteristics of participants in focus group discussions (n=29)[a].

Characteristics	2 Focus groups from Madera (n=16)	1 Focus group from Fresno (n=5)	1 Focus group from Merced (n=8)	Overall (N=29)
Age (years), median (IQR)	46[b] (15)	25 (6.5)	41 (18)	41 (22)
Sex, n (%)				
Female	10 (71.4)	4 (80)	0 (0)	14 (51.9)
Male	4 (28.6)	1 (20)	8 (100)	13 (48.1)
Race and ethnicity, n (%)				
Non-Hispanic White	0 (0)	0 (0)	3 (37.5)	3 (11.1)
Hispanic/Latino[c]	14 (100)	5 (100)	5 (62.5)	24 (88.9)
Preferred language, n (%)				
English	1 (7.1)	5 (100)	6 (85.7)	12 (46.2)
Spanish	13 (92.9)	0 (0)	1 (14.3)	14 (53.8)
English proficiency (English language preference), n (%)				
Very well	0 (0)	5 (100)	7 (100)	12 (46.2)
Well	4 (28.6)	0 (0)	0 (0)	4 (15.4)
Not well	7 (50)	0 (0)	0 (0)	7 (26.9)
Not at all	3 (21.4)	0 (0)	0 (0)	3 (11.5)
Years smoked, mean (SD)	18.4 (14)	3.2 (1.4)	12.4 (11.8)	14.0 (13.1)
Cigarettes smoked per day, mean (SD)	2.1 (2.8)	2.8 (0.8)	13.5 (8.5)	5.5 (8.1)
Time to first cigarette use after waking up, n (%)				
Within 5 minutes	1 (6.7)	1 (20)	3 (37.5)	5 (17.9)
6 - 30 minutes	1 (6.7)	1(20)	2 (25)	4 (14.3)
After 60 minutes	10 (66.7)	3 (60)	2 (25)	15 (53.6)
Question is unclear	3 (20)	0 (0)	1 (12.5)	4 (14.3)
Voluntary quit smoking for 24 hours in past year, n (%)				
No	2 (13.3)	1 (20)	2 (25)	5 (17.9)
Yes	13 (86.7)	4 (80)	6 (75)	23 (82.1)
Quitting intentions, n (%)				
Never expect to quit	1 (7.1)	1 (20)	1 (12.5)	3 (11.1)
Will quit in the next month	5 (35.7)	2 (40)	2 (25)	9 (33.3)
May quit in the next 6 months	1 (7.1)	1 (20)	4 (50)	6 (22.2)
Will quit in the next 6 months	7 (50)	1 (20)	1 (12.5)	9 (33.3)
Ability to save each month, n (%)				
No	10 (71.4)	0 (0)	7 (87.5)	17(63)
Yes	4 (28.6)	5 (100)	1 (12.5)	10 (37)
3 months savings to cover expenses, n (%)				
No	12 (85.7)	3 (60)	8 (100)	23 (85.2)
Yes	2 (14.3)	2 (40)	0 (0)	4 (14.8)
Own home, n (%)				
No	6 (42.9)	5 (100)	8 (100)	19 (70.4)
Yes	8 (57.1)	0 (0)	0 (0)	8 (29.6)

XSL•FO
RenderX

[a]Percentages for all variables include only those participants who provided a response to survey items. Responses are complete for 26 participants. Four participants skipped questions resulting in missing data for all the variables.

[b]Two participants did not report their age.

[c]Hispanic/Latino category included individuals selecting any race or not selecting a race category.

Focus Groups

Smoking and Vaping Initiation and Current Use

Participants reported initiating smoking at a young age, with many starting between 9 and 20 years. Some began with materials other than tobacco, such as smoking lawn grass. Family influences played a significant role in cigarette-smoking behaviors, with multigenerational influences being common.

> *I learned from my dad...I would steal his cigars, I would even get drunk with him...from then on, I started to want to buy cigarettes. I used to make corn leaves, which came from him, toasted tobacco in a corn leaf.* [63-year-old, male, Madera]

Social settings also contributed to the initiation and continuation of tobacco use, especially use at parties and while drinking alcohol.

> *Definitely when you're drinking alcohol, a lot of friends just drinking and a couple of them have it, you just decide to experiment with it, but then you decide to go off and buy your own, even when you're not drinking, and it becomes more of an addiction.* [Age, sex, and location not disclosed]

Using tobacco to cope with stress was a critical factor, contributing to sustained use and diminished motivation to quit.

> *[E]very time you have anxiety or are stressed, you have a lot of work, the kids are yelling at each other, or whatever, you go out at night, and you take [a cigarette].* [39-year-old, female, Madera]

Motivations to Quit Smoking or Vaping

Many in the older Spanish-speaking group were parents, and their motivation to quit was their children or family. For the younger English-speaking group, having children was a reason to quit in the future but not currently. Participants described being concerned about their health. Illnesses made them more aware of their overall health and motivated them to quit or reduce smoking. One participant got COVID-19 disease, another had asthma, and a third was worried about developing cancer.

> *So, like, I got COVID and I was just really, really scared because that was the first time I had gotten it, and I was like, "Maybe—yeah, I think I just have to stop doing everything altogether."* [24-year-old, female, Fresno]

Tobacco regulation efforts such as flavor bans or price increases motivated some individuals to quit smoking or vaping. Flavors enticed many younger participants to initiate vaping, and a participant reported that she smoked less because of the flavor ban.

> *That new law that got passed with the flavored vapes has made things super inaccessible and super hard to find, so that has been one reason why I have not been smoking too much lately.* [24-year-old, female, Fresno]

Participants tried different methods to quit smoking and vaping. Only a few participants had used NRT, some when hospitalized, which had helped them to temporarily quit. Most had tried to quit on their own (ie, cold turkey), and one participant interpreted ongoing use of NRT as a reason not to use it.

> *I've been trying to quit cold turkey. I don't really feel like the gum or anything like that works too well, and my brother did actually quit, and he used the gum. It was hard for him to even stop using the gum because it still has the nicotine in it. His recommendation to me was to quit cold turkey.* [Age, sex, and location not disclosed]

A few participants used cannabis to cope with nicotine withdrawal symptoms.

> *Smoking marijuana at night to get me to sleep was actually pretty helpful because I would go to sleep, with cold sweats, not having the nicotine in my system.* [Age, sex, and location not disclosed]

Participants described that tobacco and other substance use was common, where some participants substituted other substances (eg, cannabis) for smoking. One participant found that decreasing the use of other substances decreased their cigarette consumption.

> *Crystal [methamphetamine] and cigarettes went hand in hand. After you use one, I mean, you're gonna want the other, a finishing, topping off, type deal...I'd slow down, and then I stopped doing crystal, and then I kinda slowed down smoking.* [46-year-old male, Merced]

Another individual switched from combustible cigarettes to e-cigarettes to quit smoking but found that he instead increased nicotine use.

> *I did try to quit, but this was when the vape stuff started coming out, those little box mods. I found out that it made me crave tobacco even more. So, my smoking habit doubled from the time that I started vaping.* [Age not disclosed, male, Merced]

Recommendations for Digital Cessation Tools

Many of the Spanish-speaking participants in the focus groups felt that what was available did not meet the specific needs, challenges, or preferences of other members in their community. Participants discussed that smoking cessation tools should be not only linguistically appropriate but also include the lived experience of someone who understands their culture. Offering services in Spanish alone was not sufficient to be culturally appropriate.

> *People who really correctly understand the words that we're saying, who understand and comprehend*

XSL•FO
RenderX

> *us. If they know the culture, maybe they know a reason why we seek refuge in cigarettes...Maybe they could understand us and help us a little more.* [46-year-old, female, Madera]

In addition to information on quitting, participants felt that instructional materials to increase awareness of the dangers of secondhand smoke should be incorporated.

> *I think it's really important for all of us smokers to be aware that secondhand smoke is harmful, especially for our immediate family, our spouse, our kids, even our neighbors or coworkers.* [54-year-old, male, Madera]

When asked for general recommendations for digital cessation tools, participants expressed interest in social connections.

> *Just the fact of being able to kind of communicate with people that are dealing with the same issues as you is nice.* [24-year-old, female, Fresno]

Younger participants from Fresno also expressed interest in "more real-life examples instead of just a big X on a vape pen" [36-year-old, female, Fresno] and something easy to navigate.

> *Something that's user friendly. Something that doesn't require too many clicks. There's nothing worse than downloading an app, or opening a website, and "Oh, you have to do this step, and this step.* [24-year-old, female, Fresno]

Some individuals expressed interest in daily digital posts that they could relate to.

> *I think something like an app designed to hook you in. I know a lot of people open up these like word of the day, something that'll cause you to think, "Okay. Am I feeling that? I think that would be a good way to help people stop smoking, if you have a day to day, like, "Oh, let me check the app and see if I can relate to it today.""* [25-year-old, female, Fresno]

Semistructured Interviews on Perspectives of the MO App and Quit the Hit

Appealing Features of the MO Smoking Cessation App

Participants liked a feature of the MO app, which was a "crave or urge button" that connected to immediate cessation support on demand. They also liked being able to track their own data on their use behaviors on the app.

> *I also like the urge button where the photo was pulled...when you're kind of struggling in that moment, it gives you immediate assistance.* [25-year-old, female, Fresno]

Participants found relapse planning features of the app, including tips to modify their environment or rituals to avoid smoking cues that increase cravings (eg, avoiding the usual route to work where they buy cigarettes, rearranging furniture), to be helpful.

> *It changes your route. If you go to the store to buy a cigarette, it'll try to change your route. It's got ways that help you quit smoking. It figures your [routine]. Basically, trying to prevent you from smoking.* [46-year-old, male, Merced]

Some individuals preferred the app because it offered privacy for cessation support, which they preferred to interacting over social media.

> *It kind of reminded me of a journal.* [24-year-old, female, Fresno]

Appealing Features of the "Quit the Hit" Social Media Support Groups

Participants liked the interactive engaging content and opportunity for social and peer support with people with shared experience:

> *It's cool, comfortable, and more interactive with other people; it feels more human and more real, sharing the challenge because it feels like a challenge.* [39-year-old, female, Madera]

Other appealing features of Quit The Hit were the availability and accessibility of a peer counselor, similar to other in-person models of substance use treatment such as Alcoholics Anonymous or Narcotics Anonymous.

> *The main thing is that week by week a counselor will talk to us to give us points of view that would be good to help us that week.* [46-year-old, female, Madera]

Participants liked the idea of interacting with other members of their group when they needed it as social support and to hold themselves accountable.

> *I liked the fact that the other members of the group motivate each other and give you advice. They help you not to get discouraged if you fail, don't stop, start again.* [Age not disclosed, female, Madera]

Participants also liked that the program was structured over 5 weeks, perceiving that the time allotted allowed substantial progress in behavior change at their own pace.

> *They give you time, let's say the message you are sending me is that in five weeks I can make a big change.* [46-year-old, female, Madera]

Unappealing Features of MO and Quit the Hit Digital Interventions

A few participants expressed dissatisfaction with both interventions, describing practical inconveniences in their daily routines. For instance, a participant highlighted difficulties engaging with interventions while at work.

> *For some people it's functional, but...I couldn't be tracking or marking that kind of stuff because I work with clients.* [57-year-old, female, Madera]

Some participants reported technology-related concerns with an app such as a need for extra storage space on their device, or the need for a web-based connection or a charged battery. Other participants were uninterested in social media interventions, preferring "person to person contact" [57-year-old, female, Madera]. Others expressed concerns about privacy.

XSL•FO
RenderX

> *When you're adding and kind of joining with other people, my personal information is also out there. So, for some people, they like creating that community and making friends. But also, I like to keep my stuff separate.* [24-year-old, female, Fresno]

Some participants felt that the support should be anonymous to avoid stigma.

> *The negative stigma that comes from addiction...Yeah, your business. It's supposed to be anonymous like N[arcotics] A[nonymous]. Nobody talks about it. There's no record of it. So, I would probably go off of social media.* [Age not disclosed, male, Merced]

Other participants who did not currently use social media reported not wanting to join social media simply to access a support group.

Recommendations for Culturally Tailored Digital Interventions

Participants expressed interest in several features that could apply broadly to digital cessation tools. For example, participants expressed interest in a chance to earn prizes, money, or other tangible compensation as opposed to points.

> *Coupons, restaurants that are near them, maybe coupons for the grocery store or free Pepsis, free gum. I don't know, some kind of coupons for something.* [46-year-old, male, Merced]

Participants expressed a need for digital tools to be visually appealing and attention-grabbing. They also suggested materials to discuss tobacco use and vaping within the context of other life stressors and other addictions, including substance use.

> *Addictions can be to tobacco, drugs, food, or any other substance, and almost all of them have the same components here in the brain, and they are treated in the same way. Psychologists or psychiatrists treat it in the same way and with very similar medications.* [54-year-old, female, Madera]

Participants reported that linking treatment for tobacco use with other substance use could help prioritize the issue of tobacco use, particularly in SJV counties where counseling and support services are scarce.

> *Like I said, other chemicals and everything else, because it's not really informative over here, especially in our location where it's at with the behavioral health aspect. You could probably partner up with them and then get professional opinions or have someone on standby, when people have a burning desire or something, if you were to go out with the different aspects of every chemical out there, not just tobacco, because tobacco's really not a priority right now for most of us.* [Age not disclosed, male, Merced]

They expressed interest in linkages to other related interventions such as acupuncture or games to keep busy and cope with stressors.

> *I would still add relaxation techniques, breathing techniques. I have been to acupuncture, and it helps a lot with anxiety and stress.* [46-year-old, female, Madera]

Discussion

Principal Findings

Our research highlights the need for tobacco cessation tools for people living in economically challenged rural areas, such as the SJV. Our study indicated that both older and younger predominantly Latino adults who smoked or vaped nicotine were amenable to digital cessation tools, particularly when these interventions were delivered in a culturally affirming way that included not only language concordance but also relevant lived experiences. Consistent with past studies [14,45], participants smoked or vaped as a coping response to stressors [46], many did not use pharmacotherapy to quit, and most attempted to quit on their own. Latinos who smoke may hold misperceptions about the addictiveness of NRT [15]. In this study, SJV residents were skeptical of the use of NRT and believed that NRT may be as addictive as other nicotine products. This suggests a need for tailored programs to support evidence-based counseling and education about pharmacotherapy to improve cessation outcomes [47].

Older Spanish-speaking participants preferred culturally tailored tobacco cessation services. Participants emphasized the need for counselors or messages to express empathy toward their particular circumstances, as well as needs for relatable peer support and personalized approaches such as individualized text messaging in Spanish to increase engagement among Latino adults [48].

To overcome these barriers, evidence-based solutions can be integrated into the design and delivery of digital cessation tools. First, digital health interventions delivered or introduced by trained community health workers (CHWs) or *promotoras/es* (Spanish-speaking CHWs) from the SJV might increase buy-in among older, Spanish-speaking adults [49]. CHWs and promotoras/es, as trusted members of the community who speak the same language and understand cultural practices, can be trained to deliver tobacco cessation interventions supported by digital tools [49]. In prior work, lay health worker cessation interventions had superior abstinence rates compared with referrals to a Spanish language Quitline (20.5% vs 8.7%, respectively) [50]. Second, providing culturally tailored, accessible cessation materials (eg, mailed written material) to adults who prefer Spanish educational material improves quit rates [51]. A pilot study evaluating a culturally and linguistically tailored text message intervention combined with NRT found moderate to high levels of acceptability, good engagement measured by the number of text messages participants sent during the 12 weeks of program, and improvements in self-efficacy; this intervention reported a 30% biochemically verified abstinence rate [52].

While older participants expressed openness to using digital interventions, some lacked the familiarity and comfort with social media to engage with these platforms. According to the COM-B model, having the necessary skills (ie, the psychological

XSL•FO
RenderX

capability) and the opportunity to use social media if a social media platform is adopted may increase the chance of smoking cessation [37]. Participants' willingness to use digital interventions suggests that the CHWs and promotoras/es model may be beneficial in increasing comfort with technology. These trusted members of the community could support participants by providing guidance on how to use the digital tools prior to the delivery of intervention content [37]. Alternatively, other social media communications platforms such as WhatsApp may be better suited for residents of the SJV than a downloadable app or Instagram interventions. To address concerns about technology issues, participants who are less familiar with apps or social media platforms must receive training to help them navigate the technology with access to offline resources.

In contrast, younger participants were more receptive to social media–based interventions but emphasized the need for attention-grabbing, user-friendly, and easy-to-navigate tools. Younger participants did not emphasize culturally affirming support as much as older participants but expressed interest in peer supports and the importance of integrating lived experience of quitting. A meta-analysis of 16 randomized controlled trials found that peer support interventions were more effective for smoking cessation (mean RR 1.34, 95% CI 1.11 - 1.62) and more effective when peer support had quitting experience (k=5 studies) (mean RR1.43, 95% CI 1.17 - 1.74) [53]. To increase the use of digital cessation tools, it is important to account for digital literacy, technological issues (eg, internet connectivity or memory or battery life), peer support, and age-appropriate messaging [54,55].

Another key COM-B construct that affects the adoption of cessation tools is opportunity to engage with intervention content [37]. A few participants noted that job constraints would make it difficult to meet the demands of engaging with the cessation intervention content particularly during work hours. The limited availability highlights the need for digital cessation content that offers flexibility, whether delivered via smartphone app or social media platforms.

The motivation construct of COM-B will also affect the adoption of cessation tools [37]. Participants also expressed interest in tangible rewards such as money incentives and reminders. Contingency management is an intervention based on operant conditioning, which delivers contingent monetary rewards based on behavior change (eg, smoking cessation) [56]. A 2019 meta-analysis (k=7) found that mobile phone–delivered contingency management was efficacious for tobacco cessation [56]. Overall, we found that smartphone apps [57] and social media interventions [58] are acceptable digital health tools that could help Latinos quit smoking if tailored to enhance engagement [58]. Incorporating digital tools and health interventions is appealing because health information can reach audiences and messages can be tailored to meet individuals' personal needs or risks.

Participants reported polyuse of tobacco with other substances, including cannabis and crystal methamphetamine, but neither digital intervention offered treatment for polyuse. Treatment for tobacco and substance use can be integrated to improve substance use outcomes and should be investigated further in this population [59]. Smoking cessation interventions among people with substance use disorders can increase long-term abstinence from all substances by 25% [60]. Participants expressed that there was a dearth of substance use treatment supports in the SJV and suggested tobacco cessation to be integrated with mental health and substance use treatment resources.

Limitations

Although inclusive of Spanish-speaking Latino participants, these findings are from a small sample (29 participants) from 3 counties (Madera, Kings, and Fresno), recruited through community-based organizations, which might not generalize to other California populations, including the remaining 5 SJV counties. Only 9 participants were presented with videos of 2 prototypes of digital smoking cessation tools via interview rather than directly testing them, and recommendations may differ if the cessation tools were used in real-world settings. User testing among users and nonusers of the app could provide valuable feedback for improving the digital tools. Digital interventions that address stress as a trigger and that integrate pharmacotherapy may be beneficial.

Strengths

While our study included participants from 3 of the 8 counties that comprise SJV, these counties are broadly reflective of the region's structural and demographic characteristics. All 3 counties are also considered medically underserved, with persistent health care provider shortages and limited access to smoking cessation services [61]. Our study captured the unique voices from people who are often underrepresented in cessation research, particularly Spanish-speaking Latino individuals in the United States and rural settings. Additional strengths of this study are its successful engagement with Latino residents in underserved, rural communities, leveraging trusted community partners and Spanish-speaking researchers to ensure participants' comfort and authenticity in responses.

Conclusions

This study provides insights into digital smoking cessation tool preferences among predominantly low-income and Latino residents of the SJV, a region with urgent need of cessation services. Using a community-engaged approach, we found that participants expressed interest in linguistically appropriate and culturally tailored digital cessation tools featuring empathy toward their circumstances, relatable peer supports, engaging content, easy access to information and support, and social or material rewards. Our findings underscore the need for customizable, culturally affirming digital health interventions for tobacco cessation in the SJV, with a focus on social support and user engagement strategies.

XSL•FO
RenderX

Acknowledgments

This work was supported by a University of California, San Francisco (UCSF) Helen Diller Family Comprehensive Cancer Center Intramural grant and the National Institutes of Health (CA-113710 and DA-057216). The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies. The authors are grateful to the participants who participated in the focus groups and interviews. The authors also thank their community partner, the California Health Collaborative, who provided input to the study materials and led connections with other agencies through which the authors recruited participants. The authors would like to thank Dr. Nancy Burke, Dr. Allison Temourian, and Gisselle Navarro-Lopez for their early input on this project. The authors would also like to thank their hardworking UC Merced NCPC and UCSF interns, Mary Garcia, Justin Naidu, Erin Matsubara, and Karina Vazquez. Their contributions in helping with translating materials from English to Spanish, coding, or leading the note-taking and observation notes during focus group sessions and pilot testing were instrumental to the project.

Data Availability

Our dataset is small and primarily qualitative. It may include personal information, even after it has been deidentified. A limited dataset generated and analyzed during this study will be available upon request from the corresponding author and after signing a data-sharing agreement to protect the confidentiality of participants.

Authors' Contributions

MV served as the principal investigator of the funded project, obtained funding, developed the study methodology, led the study recruitment, supervised the study, and reviewed and edited iterations of the manuscript. AD supervised the study, developed the study methodology, led the study recruitment, moderated focus groups, coded focus groups, coded interviews, and reviewed and edited the iterations of the manuscript. KDL developed the study methodology, supported the study recruitment, moderated focus groups, moderated interviews, coded focus groups, coded interviews, and led the writing of the original draft of the manuscript. SS coded focus groups, moderated interviews, coded interviews, and reviewed and edited iterations of the manuscript. PML assisted with funding acquisition, developed the study methodology, and reviewed and edited the iterations of the manuscript. PB and AS reviewed and edited the iterations of the manuscript.

Conflicts of Interest

None declared.

Checklist 1
COREQ checklist.
[PDF File, 222 KB - formative_v9i1e74105_app1.pdf]

References

1. Eliminating tobacco-related disease and death: addressing disparities: a report of the surgeon general. : Atlanta, GA: US Department of Health and Human Services; 2024 URL: https://www.hhs.gov/sites/default/files/2024-sgr-tobacco-related-health-disparities-consumer-guide.pdf [accessed 2025-10-11]
2. Current cigarette smoking among adults in the United States. : Centers for Disease Control and Prevention; 2023 URL: https://www.cdc.gov/tobacco/media/pdfs/2024/09/cdc-osh-ncis-data-report-508.pdf [accessed 2025-10-11]
3. Kaplan RC, Bangdiwala SI, Barnhart JM, et al. Smoking among U.S. Hispanic/Latino adults: the Hispanic community health study/study of Latinos. Am J Prev Med 2014 May;46(5):496-506. [doi: 10.1016/j.amepre.2014.01.014] [Medline: 24745640]
4. Gardiner PS. The African Americanization of menthol cigarette use in the United States. Nicotine Tob Res 2004 Feb;6 Suppl 1(1):S55-S65. [doi: 10.1080/14622200310001649478] [Medline: 14982709]
5. Kong AY, Lee JGL, Halvorson-Fried SM, et al. Neighbourhood inequities in the availability of retailers selling tobacco products: a systematic review. Tob Control 2025 May 15;34(3):350-360. [doi: 10.1136/tc-2024-058718] [Medline: 38937098]
6. Mills SD, Henriksen L, Golden SD, et al. Disparities in retail marketing for menthol cigarettes in the United States, 2015. Health Place 2018 Sep;53:62-70. [doi: 10.1016/j.healthplace.2018.06.011] [Medline: 30055469]
7. Lorenzo-Blanco EI, Unger JB, Ritt-Olson A, Soto D, Baezconde-Garbanati L. Acculturation, gender, depression, and cigarette smoking among U.S. Hispanic youth: the mediating role of perceived discrimination. J Youth Adolesc 2011 Nov;40(11):1519-1533. [doi: 10.1007/s10964-011-9633-y] [Medline: 21293915]
8. Dawson DB, Fletcher TL. The influence of racial/ethnic discrimination experiences on cigarette craving for African American and Hispanic smokers. J Racial Ethn Health Disparities 2021 Aug;8(4):1047-1053. [doi: 10.1007/s40615-020-00861-0] [Medline: 32885343]

XSL•FO
RenderX

9. Lorenzo-Blanco EI, Unger JB. Ethnic discrimination, acculturative stress, and family conflict as predictors of depressive symptoms and cigarette smoking among Latina/o Youth: the mediating role of perceived stress. J Youth Adolesc 2015 Oct;44(10):1984-1997. [doi: 10.1007/s10964-015-0339-4] [Medline: 26294041]
10. California Tobacco facts and figures 2024. : California Department of Public Health; 2024 URL: https://www.cdph.ca.gov/Programs/CCDPHP/DCDIC/CTCB/CDPH%20Document%20Library/ResearchandEvaluation/FactsandFigures/CaliforniaTobaccoFactsAndFigures2024.pdf [accessed 2025-10-11]
11. California Tobacco facts and figures 2021. : California Department of Public Health; 2022 URL: https://www.cdph.ca.gov/Programs/CCDPHP/DCDIC/CTCB/CDPH%20Document%20Library/ResearchandEvaluation/FactsandFigures/CaliforniaTobaccoFactsAndFigures2021-V3A.pdf [accessed 2025-10-11]
12. QuickFacts san joaquin county. United States Census Bureau. 2021. URL: https://data.census.gov/all?q=San+Joaquin+County,+California+Income+and+Poverty [accessed 2023-10-17]
13. Kissam E, Mines R, Quezada C, Intili JA, Wadsworth G. San Joaquin Valley Latino immigrants: implications of survey findings for Census 2020. : San Joaquin Valley Health Fund; 2019 URL: https://cviic.org/san-joaquin-valley-latino-immigrants-and-the-2020-census/ [accessed 2025-10-14]
14. Babb S, Malarcher A, Schauer G, Asman K, Jamal A. Quitting smoking among adults - United States, 2000-2015. MMWR Morb Mortal Wkly Rep 2017 Jan 6;65(52):1457-1464. [doi: 10.15585/mmwr.mm6552a1] [Medline: 28056007]
15. Santiago-Torres M, Mull KE, Sullivan BM, et al. Evaluating the impact of pharmacotherapy in augmenting quit rates among Hispanic adults in an app-delivered smoking cessation intervention: secondary analysis of a randomized controlled trial. JMIR Form Res 2025 Jan 31;9:e69311. [doi: 10.2196/69311] [Medline: 39889280]
16. Carpenter MJ, Ford ME, Cartmell K, Alberg AJ. Misperceptions of nicotine replacement therapy within racially and ethnically diverse smokers. J Natl Med Assoc 2011;103(9-10):885-894. [doi: 10.1016/s0027-9684(15)30444-2] [Medline: 22364057]
17. Mayer M, Reyes-Guzman C, Grana R, Choi K, Freedman ND. Demographic characteristics, cigarette smoking, and e-cigarette use among US adults. JAMA Netw Open 2020 Oct 1;3(10):e2020694. [doi: 10.1001/jamanetworkopen.2020.20694] [Medline: 33048127]
18. Barufaldi L, Guerra R, de Albuquerque RDC, et al. Risk of smoking relapse with the use of electronic cigarettes: a systematic review with meta-analysis of longitudinal studies. Tob Prev Cessation 2021 Apr 26;7(April):1-10. [doi: 10.18332/tpc/132964]
19. James CV, Moonesinghe R, Wilson-Frederick SM, Hall JE, Penman-Aguilar A, Bouye K. Racial/ethnic health disparities among rural adults — United States, 2012–2015. MMWR Surveill Summ 2017;66(23):1-9. [doi: 10.15585/mmwr.ss6623a1]
20. Mobile fact sheet. Pew Research Center. 2024. URL: https://www.pewresearch.org/internet/fact-sheet/mobile [accessed 2025-10-11]
21. Americans' use of mobile technology and home broadband. Pew Research Center. 2024. URL: https://www.pewresearch.org/internet/2024/01/31/americans-use-of-mobile-technology-and-home-broadband [accessed 2025-10-11]
22. Social media use in 2021: a majority of Americans say they use Youtube and Facebook, while use of Instagram, Snapchat and TikTok is especially common among adults under 30. Pew Research Center. 2021. URL: https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021 [accessed 2025-10-11]
23. Llanes KD, Amastae J, Amrhein PC, et al. Impact of computer-mediated versus face-to-face motivational-type interviews on participants' language and subsequent cannabis use: randomized controlled trial. J Med Internet Res 2025 Apr 25;27:e59085. [doi: 10.2196/59085] [Medline: 40279644]
24. Zhou S, Brunetta P, Silvasstar J, Feldman G, Oromi N, Bull S. A smartphone app (MO) to provide tailored and comprehensive support for smoking cessation: a pilot multi-cohort study to evaluate acceptability, usability, and effectiveness. J Med Internet Res. Preprint posted online on Aug 26, 2024. [doi: 10.2196/55239] [Medline: 39471372]
25. Lyu JC, Olson SS, Ramo DE, Ling PM. Delivering vaping cessation interventions to adolescents and young adults on Instagram: protocol for a randomized controlled trial. BMC Public Health 2022 Dec 10;22(1):2311. [doi: 10.1186/s12889-022-14606-7] [Medline: 36496358]
26. Chen YF, Madan J, Welton N, et al. Effectiveness and cost-effectiveness of computer and other electronic aids for smoking cessation: a systematic review and network meta-analysis. Health Technol Assess 2012;16(38):1-205. [doi: 10.3310/hta16380] [Medline: 23046909]
27. Ramo DE, Thrul J, Chavez K, Delucchi KL, Prochaska JJ. Feasibility and quit rates of the tobacco status project: a Facebook smoking cessation intervention for young adults. J Med Internet Res 2015 Dec 31;17(12):e291. [doi: 10.2196/jmir.5209] [Medline: 26721211]
28. Thrul J, Tormohlen KN, Meacham MC. Social media for tobacco smoking cessation intervention: a review of the literature. Curr Addict Rep 2019 Jun;6(2):126-138. [Medline: 31485388]
29. Ramo DE, Thrul J, Delucchi KL, et al. A randomized controlled evaluation of the tobacco status project, a Facebook intervention for young adults. Addiction 2018 May 24;113(9):1683-1695. [doi: 10.1111/add.14245] [Medline: 29797621]
30. Llanes KD, Ling PM, Guillory J, Vogel EA. Young adults' perceptions of and intentions to use nicotine and cannabis vaporizers in response to e-cigarette or vaping-associated lung injury Instagram posts: experimental study. J Med Internet Res 2023 Sep 14;25:e46153. [doi: 10.2196/46153] [Medline: 37552552]

XSL•FO
RenderX

31. Luo T, Li MS, Williams D, et al. Using social media for smoking cessation interventions: a systematic review. Perspect Public Health 2021 Jan;141(1):50-63. [doi: 10.1177/1757913920906845] [Medline: 32077368]
32. Guo YQ, Chen Y, Dabbs AD, Wu Y. The effectiveness of smartphone app-based interventions for assisting smoking cessation: systematic review and meta-analysis. J Med Internet Res 2023 Apr 20;25:e43242. [doi: 10.2196/43242] [Medline: 37079352]
33. Zhou X, Wei X, Cheng A, et al. Mobile phone-based interventions for smoking cessation among young people: systematic review and meta-analysis. JMIR Mhealth Uhealth 2023 Sep 12;11(1):e48253. [doi: 10.2196/48253] [Medline: 37706482]
34. Bold KW, Garrison KA, DeLucia A, et al. Smartphone apps for smoking cessation: systematic framework for app review and analysis. J Med Internet Res 2023 Jul 13;25:e45183. [doi: 10.2196/45183] [Medline: 37440305]
35. Micalizzi L, Mattingly DT, Hart JL, Jensen JK, Mahabee-Gittens EM, Garrison KA. Smartphone apps targeting youth tobacco use prevention and cessation: an assessment of credibility and quality. Curr Addict Rep 2023 Dec;10(4):649-663. [doi: 10.1007/s40429-023-00524-0] [Medline: 38680515]
36. Hispanic and Latino people experience a health burden from commercial tobacco. Centers for Disease Control and Prevention. 2024. URL: https://www.cdc.gov/tobacco-health-equity/collection/hispanic-latino-health-burden.html [accessed 2024-10-11]
37. Willmott TJ, Pang B, Rundle-Thiele S. Capability, opportunity, and motivation: an across contexts empirical examination of the COM-B model. BMC Public Health 2021 May 29;21(1):1014. [doi: 10.1186/s12889-021-11019-w] [Medline: 34051788]
38. Orji R, Vassileva J, Mandryk R. Towards an effective health interventions design: an extension of the health belief model. Online J Public Health Inform 2012;4(3):ojphi.v4i3.4321. [doi: 10.5210/ojphi.v4i3.4321] [Medline: 23569653]
39. Schunk DH, DiBenedetto MK. Motivation and social cognitive theory. Contemp Educ Psychol 2020 Jan;60:101832. [doi: 10.1016/j.cedpsych.2019.101832]
40. Fagerström K. Time to first cigarette; the best single indicator of tobacco dependence? Monaldi Arch Chest Dis 2003;59(1):91-94. [Medline: 14533289]
41. Zhou S, Brunetta P, Silvasstar J, Feldman G, Oromi N, Bull S. Initial evaluation of acceptability, engagement, and effectiveness of the MO app to provide tailored and comprehensive support for smoking cessation: development and usability study. JMIR Mhealth Uhealth 2024 Oct 29;12:e55239. [doi: 10.2196/55239] [Medline: 39471372]
42. Lyu JC, Meacham MC, Nguyen N, Ramo D, Ling PM. Factors associated with abstinence among young adult smokers enrolled in a real-world social media smoking cessation program. Nicotine Tob Res 2024 Feb 15;26(Supplement_1):S27-S35. [doi: 10.1093/ntr/ntad170] [Medline: 38366340]
43. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
44. Braun V, Clarke V. Reflecting on reflexive thematic analysis. Qualitative Research in Sport, Exercise and Health 2019 Aug 8;11(4):589-597. [doi: 10.1080/2159676X.2019.1628806]
45. Levinson AH, Pérez-Stable EJ, Espinoza P, Flores ET, Byers TE. Latinos report less use of pharmaceutical aids when trying to quit smoking. Am J Prev Med 2004 Feb;26(2):105-111. [doi: 10.1016/j.amepre.2003.10.012] [Medline: 14751320]
46. Nguyen N, Islam S, Llanes KD, Koester KA, Ling PM. Classification of patterns of tobacco and cannabis co-use based on temporal proximity: a qualitative study among young adults. Addict Behav 2024 May;152:107971. [doi: 10.1016/j.addbeh.2024.107971] [Medline: 38281461]
47. Stead LF, Koilpillai P, Fanshawe TR, Lancaster T. Combined pharmacotherapy and behavioural interventions for smoking cessation. Cochrane Database Syst Rev 2016 Mar 24;3(3):CD008286. [doi: 10.1002/14651858.CD008286.pub3] [Medline: 27009521]
48. El-Toukhy S, Kamke K. Intervention targeting and retention, engagement and abstinence outcomes among Latino and White users of Smokefree.gov text messaging programmes: a cross-sectional study. BMJ Public Health 2023;1(1):e000222. [doi: 10.1136/bmjph-2023-000222] [Medline: 38124887]
49. Woodruff SI, Candelaria JI, Elder JP. Recruitment, training outcomes, retention, and performance of community health advisors in two tobacco control interventions for Latinos. J Community Health 2010 Apr;35(2):124-134. [doi: 10.1007/s10900-009-9207-z] [Medline: 20012475]
50. Woodruff SI, Talavera GA, Elder JP. Evaluation of a culturally appropriate smoking cessation intervention for Latinos. Tob Control 2002 Dec;11(4):361-367. [doi: 10.1136/tc.11.4.361] [Medline: 12432162]
51. Simmons VN, Sutton SK, Medina-Ramirez P, et al. Self-help smoking cessation intervention for Spanish-speaking Hispanics/Latinxs in the United States: a randomized controlled trial. Cancer 2022 Mar 1;128(5):984-994. [doi: 10.1002/cncr.33986] [Medline: 34679188]
52. Cartujano-Barrera F, Sanderson Cox L, Arana-Chicas E, et al. Feasibility and acceptability of a culturally- and linguistically-adapted smoking cessation text messaging intervention for Latino smokers. Front Public Health 2020;8:269. [doi: 10.3389/fpubh.2020.00269] [Medline: 32714891]
53. Yuan P, Westmaas JL, Thrul J, Toussaert S, Hilton JF, White JS. Effectiveness of peer-support interventions for smoking cessation: a systematic review and meta-analysis. Nicotine Tob Res 2023 Aug 19;25(9):1515-1524. [doi: 10.1093/ntr/ntad059] [Medline: 37042206]

XSL•FO
RenderX

54. Crane D, Garnett C, Brown J, West R, Michie S. Factors influencing usability of a smartphone app to reduce excessive alcohol consumption: think aloud and interview studies. Front Public Health 2017;5:39. [doi: 10.3389/fpubh.2017.00039] [Medline: 28421175]
55. Guan M, So J. Influence of social identity on self-efficacy beliefs through perceived social support: a social identity theory perspective. Commun Stud 2016 Oct 19;67(5):588-604. [doi: 10.1080/10510974.2016.1239645]
56. Getty CA, Morande A, Lynskey M, Weaver T, Metrebian N. Mobile telephone-delivered contingency management interventions promoting behaviour change in individuals with substance use disorders: a meta-analysis. Addiction 2019 Nov;114(11):1915-1925. [doi: 10.1111/add.14725] [Medline: 31265747]
57. Santiago-Torres M, Mull KE, Sullivan BM, Zvolensky MJ, Bricker JB. Can a smartphone application help Hispanic/Latinx adults quit smoking? A randomized trial secondary analysis. J Contextual Behav Sci 2022 Oct;26:261-270. [doi: 10.1016/j.jcbs.2022.11.001] [Medline: 36437818]
58. Naslund JA, Kim SJ, Aschbrenner KA, et al. Systematic review of social media interventions for smoking cessation. Addict Behav 2017 Oct;73:81-93. [doi: 10.1016/j.addbeh.2017.05.002] [Medline: 28499259]
59. McKelvey K, Thrul J, Ramo D. Impact of quitting smoking and smoking cessation treatment on substance use outcomes: an updated and narrative review. Addict Behav 2017 Feb;65:161-170. [doi: 10.1016/j.addbeh.2016.10.012] [Medline: 27816663]
60. Prochaska JJ, Delucchi K, Hall SM. A meta-analysis of smoking cessation interventions with individuals in substance abuse treatment or recovery. J Consult Clin Psychol 2004 Dec;72(6):1144-1156. [doi: 10.1037/0022-006X.72.6.1144] [Medline: 15612860]
61. HDPulse: an ecosystem of minority health and health disparities resources: california poverty. National Institute on Minority Health and Health Disparities. URL: https://hdpulse.nimhd.nih.gov/data-portal/social/tdi2y=01&ge qins=gal 1&dm=0007&dm qins=poty 3&ae=08&ae qins=ae 7&sex=0&sex qins=sxhh 1&soihi=08&soihi qins=soil 6&stfps=08&stfps qins=ae sts [accessed 2025-10-11]

Abbreviations

AIAN: American Indian Alaska Native
CHC: California Health Collaborative
CHW: community health worker
COM-B: Capability, Opportunity, Motivation–Behavior
COREQ: Consolidated Criteria for Reporting Qualitative Research
NRT: nicotine replacement therapy
SJV: San Joaquin Valley

Edited by A Mavragani, J Sarvestan; submitted 18.03.25; peer-reviewed by BN Tranby, CS Nagawa; revised version received 25.09.25; accepted 26.09.25; published 10.11.25.

Please cite as:
Llanes KD, Vijayaraghavan M, Schneider S, Ling PM, Hernandez E, Brunetta P, Song AV, Durazo A
Digital Smoking Cessation Preferences of Predominately Low-Income and Latino Residents of the San Joaquin Valley in California: Qualitative Study
JMIR Form Res 2025;9:e74105
URL: https://formative.jmir.org/2025/1/e74105
doi:10.2196/74105

© Karla D Llanes, Maya Vijayaraghavan, Sara Schneider, Pamela M Ling, Evi Hernandez, Paul Brunetta, Anna V Song, Arturo Durazo. Originally published in JMIR Formative Research (https://formative.jmir.org), 10.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Active Play in a Digital Age, Exploring Children's (Aged 8-13 Years) Views of a Physical Activity App: Qualitative Formative Study

Marie T Frazer[1,2,3], MA; Lauren Charlesworth[2,3], BA; Luca Wilson[2,3], PhD; Jennifer Hall[1,2,3], PhD; Farwa Batool[2,3], MS; Mikel Subiza-Perez[3,4,5,6], PhD; Jan Burkhardt[3], MA; Andy Daly-Smith[1,7], PhD; Anna Chalkley[1,7], PhD

[1]Faculty of Health Studies, Institute for Health and Social Care, University of Bradford, Richmond Road, Bradford, United Kingdom

[2]Bradford Centre for Qualitative Research, Wolfson Centre for Applied Health Research, Bradford, United Kingdom

[3]Bradford Institute for Health Research, Bradford Teaching Hospitals NHS Foundation Trust, Bradford, United Kingdom

[4]Department of Clinical and Health Psychology and Research Methods, University of the Basque Country UPV/EHU, Donostia-San Sebastián, Spain

[5]Spanish Consortium for Research on Epidemiology and Public Health, Instituto de Salud Carlos III, Madrid, Spain

[6]Biogipuzkoa Health Research Institute, Group of Public Health and Environmental Epidemiology, Donostia- San Sebastián, Spain

[7]Centre for Applied Education Research, Wolfson Centre for Applied Health Research, Bradford, United Kingdom

Corresponding Author:
Marie T Frazer, MA
Faculty of Health Studies, Institute for Health and Social Care, University of Bradford, Richmond Road, Bradford, United Kingdom

Abstract

Background: The use of smartphones and interest in mobile health (mHealth) has grown in recent years with physical activity apps demonstrating potential to facilitate behavior change. However, there remains limited understanding of what specifically motivates children to engage meaningfully with these tools.

Objective: This qualitative formative study aimed to determine children's perceptions of a bespoke physical activity mHealth app (Bestlife; Dubbit). It sought to explore the app's appeal, functionality, and potential to support behavior change among children aged 8‑13 years.

Methods: A total of 68 Young Citizen Scientists (YCSs) aged 8‑13 years were recruited from 5 schools (3 primary and 2 secondary) in Bradford, United Kingdom, through purposive sampling as part of a whole-system physical activity program (Join Us: Move. Play; JU:MP). Recruitment procedures were school-led, incorporating consented whole-class involvement at primary level and teacher-nominated groups at secondary level. YCSs were asked to download and explore the Bestlife app 1‑2 weeks before the school-based research session, completing a booklet to capture their experiences and those of their families. A total of 13 focus groups were conducted across 5 schools to explore children's views in depth. The focus groups were designed to investigate children's perceptions of the app. Qualitative data were analyzed inductively and deductively: An initial inductive analysis identified emerging themes, which were then mapped onto a framework of feasibility, usability, acceptability, and behavior change.

Results: A total of 68 children (60 from primary schools and 8 from secondary schools) participated in the study. The study identified key factors influencing the feasibility, acceptability, usability, and behavior change potential of the Bestlife app among children. Feasibility was hindered by the parental email requirement during registration, which limited autonomy for older children. Acceptability was driven by gamified features, proportional rewards, and avatar customization, though participants requested more personalization to promote cultural inclusion and dynamic updates, linked to seasonal themes. Usability findings showed the interface was intuitive, with features promoting social interaction and competition enhancing engagement. However, younger users experienced navigational challenges, underscoring the need for clearer guidance. The app effectively incorporated behavior change techniques, including goal-setting, self-monitoring, and social collaboration, but required adjustments, such as reducing the frequency of emotional tracking prompts.

Conclusions: The Bestlife app shows potential as an mHealth intervention for promoting physical activity in children. Enhancing cultural representation, simplifying onboarding processes, and refining engagement strategies could strengthen both uptake and sustained use. These findings highlight the importance of integrating user feedback into the iterative design process to optimize digital health tools for young populations. Further longitudinal research is recommended to evaluate longer-term engagement with the app, its impact on physical activity levels, and behavior change sustainability.

XSL•FO
RenderX

(JMIR Form Res 2025;9:e76498) doi:10.2196/76498

KEYWORDS

children; young people; physical activity; mHealth; citizen science; feasibility; acceptability; usability; behaviour change; mobile health

Introduction

Background

Physical inactivity may have profound consequences for children's health, including reduced cardiometabolic fitness, impaired bone health, and increased adiposity [1,2]. It also poses risks to mental health, potentially leading to cognitive deficits and other negative psychological outcomes. These issues are particularly acute in areas with high deprivation and among ethnic minority communities, where inactivity levels are highest [3,4]. Conversely, regular physical activity during childhood provides lasting benefits, often persisting into adulthood and reducing the risk of noncommunicable diseases [5-7].

Across England, 53% of children aged 5‑16 years do not achieve the recommended 60 minutes of daily physical activity [8]. Social, economic, and cultural disparities significantly influence physical activity levels, with children living in areas of high socioeconomic deprivation and diverse ethnic minority communities, particularly those of South Asian heritage, experiencing lower levels of physical activity [8,9]. These disparities exacerbate health inequalities by limiting access to resources and culturally-tailored opportunities for physical activity, underscoring the urgent need for inclusive and targeted interventions [8,9].

To address the low levels of inactivity, the World Health Organization's (WHO) Global Action Plan on Physical Activity 2018‑2030 advocates for the adoption of a whole systems approach to tackling physical inactivity, including incorporating the use of digital innovations to promote and support people of all ages to be more active and to build upon the rapidly growing practice of mHealth (mobile health) [10,11]

Smartphones are ubiquitous in young people's lives, with 98% of 12‑15-year-olds having access to smartphones [12]. These devices provide quick access to information and social connections [13-15]. The dichotomous influence of smartphones on young people's physical activity behavior is striking. For example, they are associated with increased screen time and decreased physical activity, as the ability to engage with peers and content online reduces opportunities for movement [14,16]. While digital tools can support healthier behaviors, it is vital to address the ethical considerations surrounding children's use of smartphones. The increasing prevalence of smartphone use among individuals 16 years and younger has been linked to adverse effects on mental health, sleep quality, and addiction-like behaviors [17-20]. Given these risks, Bestlife (Dubbit) is not intended to promote unsupervised smartphone use. Rather, it is designed as a supportive tool to be used under parental or educator guidance, aligning with evidence that highlights the importance of adult mediation in reducing harm and fostering healthy digital habits [21]. However, smartphones also hold immense untapped potential as tools for positive health behavior change, including increasing physical activity [16,17]. When designed effectively, mHealth interventions can transform smartphones into facilitators of physical activity, using their accessibility, interactivity, and personalized features to promote engagement and behavior change [15,16].

Children from lower socioeconomic groups are particularly likely to use smartphones and tend to use them more frequently than their peers from other socioeconomic backgrounds [22]. Sharma et al [23] reiterated the increasing use of smart devices to access the internet among diverse low-income communities. This presents a possible opportunity to harness smartphones to promote healthy behavior, specifically within this demographic, and thus reduce health inequalities. Advances in technology and the widespread increase in smartphone ownership among young people [24] now enable the development of tailored interventions that address health disparities. By leveraging smartphones as accessible, interactive, and engaging tools, under parental supervision, it is possible to encourage positive behavior changes, particularly in physical activity, among younger individuals from low socioeconomic backgrounds.

Comprehensive reviews of physical activity apps suggest that many are commercially developed without academic input [25,26]. It has also been noted that many app developments do not make use of relevant underlying theoretical models to inform design elements used in apps [27]. Historically, many of these apps lack a solid foundation in behavioral science or public health principles, which raises concerns about effectiveness and appropriateness as tools for behavior change [28,29]. Recent trends, however, suggest a promising shift toward incorporating user-centered design and evidence-based theories. These approaches aim to enhance app functionality and engagement, particularly for children and adolescents, where the application of behavior change techniques remains in its infancy [30]. For example, the Snacktivity project is a notable case where researchers and health professionals collaborated to create an app promoting small, frequent bouts of activity "snacks" [31]. Other mHealth interventions, such as the Active 10 app developed by Public Health England, also demonstrate practitioner involvement [32].

Evaluations of physical activity apps are crucial to determine their effectiveness in promoting physical activity among users, with augmented reality games like Pokémon GO (Niantic) demonstrating a general tendency to positively affect physical activity levels among children and adolescents [33]. Many commercially available apps lack rigorous evidence of their efficacy, which raises concerns about their potential impact on children's health behaviors. For instance, Schoffman et al [34,35] and Schoeppe et al [35] noted that despite the ubiquity of these apps, there is a significant gap in evidence demonstrating their effectiveness in real-world settings. Without proper evaluation, there is a risk that investment is made into

XSL•FO
RenderX

interventions that children do not engage with and therefore do not effectively promote physical activity.

User engagement is a critical factor in the success of physical activity apps, especially for children who may have shorter attention spans and varying levels of motivation. Augmented reality displays and gamification features, as adjuncts, play a significant role in fostering user engagement with physical activity apps [36,37]. Both technologies leverage interactive elements and motivational design principles to make physical activity more enjoyable and encourage sustained participation [36,37]. Evaluating these apps can provide insights into usability and engagement factors that resonate with young users. Tong et al [38] found that apps with social networking features and personalized feedback significantly enhance user engagement. Understanding these dynamics is crucial for designing apps that children will want to use regularly.

Children have unique developmental needs and preferences that differ from those of adults. Evaluating physical activity apps aimed at children is necessary to ensure that the content is age-appropriate, engaging, and effective in motivating young users [39-41]. Many popular physical activity apps do not specifically target children, which indicates a lack of tailored interventions for this demographic [39]. This gap highlights the need for evaluations that focus on how well these apps meet the developmental and motivational needs of children. Furthermore, to ensure effective and equitable digital physical activity interventions for children, it is critical to address multifaceted barriers including socioeconomic disparities that disproportionately affect low socioeconomic status populations [42]. In addition, challenges related to eHealth literacy, language, and cultural representation significantly hinder engagement for Culturally and linguistically diverse communities [43], necessitating that future development prioritizes person-centered and participatory approaches involving target communities in the design process.

The Bestlife app was developed as a targeted response to the challenges young people face in engaging in physical activity. Designed for children aged 10‑13 years to use with parental guidance and delivered as part of the Join Us: Move. Play (JU:MP) whole system approach to improving children's physical activity in Bradford, United Kingdom [44]. The app integrates physical activity tracking, gamified challenges, and emotional well-being assessments. Grounded in behavior change theory, the app incorporates evidence-based behavior change techniques [45] and was co-produced with children to ensure cultural relevance and user-centered design. This study adopted a citizen science approach to explore the app's acceptability, feasibility, usability, and potential to facilitate behavior change among children, focusing particularly on their engagement with both physical activity and emotional well-being features.

Aim

The primary aim of this study was to carry out a citizen science feasibility study of a children's physical activity app, focusing on the perspectives of the users (ie, children). Specifically, we aimed to:

1. Assess the acceptability of the app, exploring children's attitudes, preferences, and overall satisfaction, including how these factors influence their perceived engagement with physical activity and subjective well-being.
2. Evaluate the feasibility of incorporating the use of the app into children's daily routines, and identify barriers and facilitators of use related to socioeconomic, cultural, and technological contexts.
3. Examine the app's usability, focusing on navigation and ease of use.
4. Investigate early indicators of behavior change facilitated by the app, assessing user engagement with physical activity features.

Methods

The Bestlife App

Released in July 2023 on the Apple and Google Play Stores, the Bestlife app was developed by the software company Dubit in collaboration with the Born in Bradford (BiB) research program [46] and JU:MP. JU:MP [44] is a whole-system approach to increasing physical activity among children in Bradford, funded by Sport England as part of a local delivery pilot (LDP). The app was co-designed with researchers, practitioners, and, most importantly, children, ensuring its development was shaped by those who would use it.

Designed specifically for 10‑ to 13-year-olds, Bestlife integrates gamification, personalization, and emotional well-being tracking to encourage physical activity. The app enables users to engage in "active quests," such as step challenges, fitness exercises, or visits to parks with QR codes, to earn points that can be used to customize their avatars. Parental permission, verified through email, is required to create an account.

Beyond solo challenges, Bestlife supports collaborative gameplay, fostering social engagement. Users track their activities and emotional well-being through an interactive dashboard, providing insights into their physical and mental health. The app aligns with JU:MP's mission of improving children's activity levels within their local environments, including schools, parks, and neighborhoods.

By integrating rewards, personalization, and emotional well-being tracking, Bestlife aims to increase children's engagement in physical activity while fostering awareness of how participation relates to their mental health. For detailed screenshots and a guide on how to play (see Multimedia Appendix 1).

Key Concepts

Acceptability, Feasibility, and Usability

Within this study, we used the key concepts of acceptability, feasibility, and usability as follows: Acceptability refers to how children react to the app, did they think it was appropriate, satisfactory, and attractive [47], including children's willingness to download and engage with the app and their interest in participating in challenges or quests set on the app. Feasibility assesses the practicality of implementing the app within the

XSL•FO
RenderX

target population's daily routines and local context [47]. This includes smartphone ownership, internet accessibility, etc. Finally, usability focuses on the design aspects of the app that contribute to a positive user experience [48]. This consists of the ease of navigation, intuitiveness of the interface, and overall user-friendliness of the app. By examining these 3 dimensions, we aimed to provide a comprehensive assessment of the app's potential for promoting physical activity among children and identify areas for improvement in future iterations.

Behavior Change Theory in the Development of Bestlife

The development of Bestlife was underpinned by the behavior change theory, specifically drawing on the Behavior Change Wheel (BCW) [45]. The BCW provides a systematic framework for designing interventions that support and sustain behavior change by addressing 3 core components: capability, opportunity, and motivation. These components form the COM-B (Capability, Opportunity, Motivation—Behavior) model, which posits that for a behavior to occur, an individual must have the capability (both physical and psychological), the opportunity (both social and environmental), and the motivation (both automatic and reflective) to engage in it. The Bestlife app was designed to integrate these elements, ensuring that it effectively encouraged young people to engage in and sustain physical activity.

Gamification elements such as points, rewards, and avatar customization were included to enhance motivation, providing immediate and tangible reinforcement for physical activity. The app also allowed users to track their own activity and emotional well-being, supporting self-monitoring and reflection, techniques recognized for fostering sustained behavior change. This design was informed by insights from Dubbit during the development phase, where children and young people recognized that being active made them feel happier, highlighting the potential for the app to reinforce that positive association and further motivate ongoing participation.

The design also considered the role of social and environmental opportunity in shaping behavior. By incorporating features that encouraged children to engage in activity within their local environment, such as interactive park visits using QR codes, the app aligned with JU:MP's emphasis on creating active environments that make movement an embedded part of daily life. Ensuring accessibility was also a key principle, with the app designed to be used on smartphones, a device that is widely available to young people.

This approach ensured that the app aligned with the broader goals of JU:MP, leveraging a theory-driven yet practical design to maximize impact.

Study Design

This study used a contributory citizen science approach, engaging children as citizen scientists. Citizen science, defined in this paper as collaboration between the public and professional scientists to advance research [49-51], has shown significant promise in engaging young people across various fields [52]. By capturing authentic insights from young people's lived experiences, this method aligns with the growing recognition that young people should actively participate in research that impacts them [53]. Previous research has successfully applied citizen science to understand young people's physical activity experiences [54-57]. This precedent suggests that a similar approach could be highly effective in evaluating the Bestlife app, providing valuable insights into its usability, effectiveness, and impact on physical activity levels among children and young people. The children involved in the research will be referred to as Young Citizen Scientists (YCSs) for the remainder of the paper.

The Citizen Science Approach

Patient and Public Involvement (PPI)

A total of 3 discussion and consultation sessions were conducted before designing the study, with the aim of identifying barriers and facilitators to research participation with children in Bradford. These sessions, held between May and June 2023, engaged key stakeholders, including BiB Young Ambassadors, young people involved in a JU:MP longitudinal research study, and (primary and secondary) teachers. Insights from these sessions played a crucial role in shaping the study design to ensure the research process was engaging, relevant, and minimally disruptive.

The sessions with the BiB Young Ambassadors and the young people involved in JU:MP emphasized the importance of conducting the research in familiar settings, such as schools or community spaces, with adults the children already worked with. This approach aimed to create a comfortable and natural environment where children could speak freely. They also advised that all research materials, including consent forms, should be fun, engaging, and age-appropriate to maintain interest and understanding.

Teachers provided different suggestions based on the age group they worked with. Primary school teachers recommended conducting the sessions with the whole class to prevent lesson disruptions and to ensure that all children could take part. In contrast, secondary school teachers were keen to avoid taking an entire class out of a lesson, as this would put students out of sync with their peers. Instead, they suggested removing a small group of students for the session, allowing them to catch up on their work later with minimal disruption to the rest of the year group.

These insights informed the decision to use focus groups as the primary method of data collection, providing a comfortable and interactive format for children to share their perspectives. The use of booklets allowed information to be conveyed in an age-appropriate and engaging format. The booklets were colorful, featuring definitions of key terms such as "citizen science" and "JUMP," as well as spaces for children to complete quizzes and record their own research, for example, whether members of their families enjoyed using the app (see Multimedia Appendix 2). This supported both understanding and active participation in the research. By incorporating these recommendations, the study design ensured that research activities were integrated smoothly into school routines while prioritizing a child-friendly and inclusive approach.

XSL•FO
RenderX

Participants and Recruitment

The study targeted children and young people aged 8 - 14 years (Key Stages 2 and 3) to become YCSs. Although the app was initially designed for children aged 10 - 14 years, feedback from the public involvement groups suggested that younger children might also engage with the app, prompting the inclusion of 8 to 9-year-olds.

The sampling strategy was designed to ensure a diverse and representative sample of children from Bradford, aligning with the aims of the JU:MP whole-systems approach to increasing physical activity for all children and young people and particularly those from low socioeconomic status and ethnic minority groups. Given that Bestlife was being implemented within JU:MP delivery areas, recruitment was purposive, targeting schools already engaged with the intervention. This approach maximized feasibility by leveraging existing networks and relationships with schools, ensuring that the study was embedded within communities where the intervention was being delivered.

For primary schools, a 2-stage recruitment process was used. First, all schools JU:MP were working with at the time were invited to participate, with recruitment being dependent on the schools that responded first. Second, to ensure ethnic diversity in the sample, reflecting the demographics of Bradford, schools with different ethnic compositions were specifically sought, given JU:MP's implementation teams' knowledge of the schools and the catchment areas. Initial primary focus groups took place in a predominantly White British area, and to ensure representation from Bradford's large South Asian population, an additional primary school was selected from areas with a higher proportion of Asian pupils. While the recruitment process was shaped by practical constraints, including school engagement and availability, the sampling strategy was purposeful in achieving variation across ethnic backgrounds. This ensured that the study captured a broad range of experiences and perspectives, making the findings more relevant for diverse communities.

A total of 2 secondary schools were contacted directly through existing links within JU:MP areas. Within the secondary schools, teachers facilitated recruitment. In one school, this was done by identifying a range of students from different socioeconomic and ethnic backgrounds from across year 8 (approximately two per form) and handing out the research booklets. In the second school, this involved an established lunchtime youth group, who were also given the research booklets. Focus groups were conducted in school-provided spaces with small groups of YCSs. In primary schools, entire classes participated in engaging with the app, with only consenting students included in the formal research.

Ethical Considerations

Ethical considerations are fundamental, particularly when working with potentially vulnerable demographics such as children, as they protect participants' rights, dignity, and well-being.

Ethics Review and Approvals

Ethical approval for all studies was granted by the University of Bradford Research Ethics Committee (Humanities, Social, and Health Sciences Research Ethics Panel; reference: E891; July 2021). Subsequent amendments, including the participant experience focus groups following the control trial and the JU:MP focus groups, were approved on June 29, 2022. Similarly, the Bestlife study amendment to E891 was approved on July 04, 2023.

Privacy and Confidentiality Protection

All data were anonymized, and participants' identities were protected throughout the research process. Focus group discussions were audio-recorded and transcribed, and transcripts were anonymized before analysis. Participant booklets were collated and digitized, and both booklets and transcripts were coded to ensure the integration of verbal and written responses. Safeguarding procedures in line with university policy were in place throughout the research and would have been followed had any concerns arisen.

Compensation

Participants did not receive compensation for their involvement. This decision aligns with ethical principles to avoid undue influence on participation decisions, particularly as focus groups were conducted during the regular school day as part of existing school activities.

Identification in Images

No identification of individual participants or users is possible in any images within this manuscript. All images presented are illustrative of the research tools or conceptual frameworks and do not contain identifiable individuals.

Study Procedure

The study followed a 2-phase data collection process, involving independent app use followed by facilitated focus groups.

Pre–Focus Group Engagement

Before the focus groups, YCSs were provided with a booklet containing an information sheet, consent form, and QR code to download Bestlife. The booklet guided them through independent app exploration for a week, prompting them to record their experiences, preferences, and interactions. It also encouraged them to discuss the app with their families and friends, capturing broader perceptions of its usability and appeal.

Focus Group Procedures

After the independent use phase, researchers conducted focus groups in schools to facilitate discussion and further app engagement.

- Primary school focus groups were conducted at the class level to ensure all children could participate without disrupting lesson timetables.
- Secondary school focus groups were held with smaller groups, as preferred by teachers to minimize classroom disruption.

Focus groups were selected based on public involvement (PPI) sessions, ensuring that children participated in a setting where

XSL•FO
RenderX

they felt comfortable sharing their experiences. This approach was chosen over interviews to create a more natural and engaging environment for discussion [58]. A copy of the focus group guide can be found in Multimedia Appendix 3.

Focus groups were primarily led by MTF, a female PhD student with an MA in Education and extensive training in qualitative research. When multiple groups ran concurrently, MTF facilitated one, supported by 2 other trained female PhD students and research assistants. All facilitators had previous qualitative research experience and received specific training on focus group delivery and study objectives. No previous relationship existed between researchers and participants. Children were informed at the outset that the research aimed to explore their experiences with the Bestlife app. MTF, embedded within the JU:MP program, conducted this study as part of its broader evaluation.

During the sessions, researchers used the booklets as discussion prompts, encouraging children to expand on their experiences. The booklets also allowed children to express themselves nonverbally, either through writing or drawing, helping to facilitate engagement and accommodate different communication preferences [59]. In addition, focus groups incorporated live app demonstrations, where children interacted with Bestlife in real time, exploring its features and providing immediate feedback. This hands-on element enabled researchers to observe and record engagement, usability, and barriers to interaction through the use of reflective field notes.

In all focus groups, school staff such as support staff or class teachers were present to provide a familiar environment for children. However, school staff did not participate in the research discussions to ensure that children felt free to share their honest opinions. Each focus group session lasted approximately 30 minutes to 1 hour, depending on school scheduling, and was designed to fit within the standard school timetable.

By integrating both individual reflection and interactive group discussion, this procedure ensured a comprehensive evaluation of the Bestlife app, capturing insights into both independent use and collaborative engagement. After each focus group, members of the research team engaged in informal reflective discussions to consider group dynamics, emerging insights, and the facilitation process. Field notes were taken during and after the sessions to document observations, including participant engagement, nonverbal cues, and contextual details. These notes informed the interpretation of data and supported ongoing reflexivity throughout the study.

Data Analysis

All focus group discussions were audio-recorded and transcribed. Transcripts were anonymized before analysis, and participant booklets were collated and digitized. Both the booklets and the transcripts were coded to ensure the integration of both verbal and written responses.

The analysis was conducted in 2 key phases. Initially, open coding was carried out by 2 researchers (MTF and LC) to capture emerging patterns in participant responses without predefined categories. This inductive process ensured that all relevant insights were captured, allowing themes to come directly from the data, prioritizing the children's voices. Following the open coding, the generated codes were compared against the original theoretical framework underpinning the Bestlife app, which was structured around feasibility, acceptability, usability, and behavior change (done by MTF, LW, and AC). The initial open codes aligned well with these categories.

In the second phase, researchers (LW and MTF) recoded the data, explicitly mapping it onto these 4 overarching categories. The recoding process ensured that all identified insights were systematically integrated into the study's core evaluative framework. This theory-driven recoding ensured initial themes were aligned with the study's predefined evaluation framework, while the first phase allowed the researchers to be open to organic themes that the children had discussed. A summary of example codes is available in the code book (see Multimedia Appendix 4).

Results

Overview

A total of 68 YCSs gave consent to take part in the research. In total, 60 were from 2 primary schools and 8 were from 2 secondary schools (see Table 1)

XSL•FO
RenderX

Table . Number of focus groups and Young Citizen Scientists by school and class group in a qualitative formative study exploring children's perceptions of the Bestlife physical activity app.

School and class group		Focus groups, n	Consenting pupils, n
Primary school 1			
	4	5	27
	6	1	6
Primary school 2			
	5a	1	5
	6a	1	4
	6b	3	18
Primary totals		11	60
Secondary school 1			
	7 to 9	1	3
Secondary school 2			
	7 to 9	1	5
Secondary totals		2	8
Overall totals		13	68

The findings are presented through the perspectives of feasibility, usability, and acceptability, which collectively provide insights into the app's overall effectiveness adoption (See Figure 1). These dimensions are intrinsically linked to the behavior change principles that were foundational to the app's design. The factors of feasibility, usability, and acceptability help to examine the extent to which children and families engage positively with the app. These 3 dimensions influence and are influenced by behavior change techniques, such as goal setting (motivation), feedback (motivation), and social collaboration (opportunity), which were integrated into the app to encourage physical activity. The results section explores each dimension while highlighting how they support behavior change outcomes.

XSL•FO
RenderX

Figure 1. Conceptual framework for the evaluation of the Bestlife physical activity app [47,48].

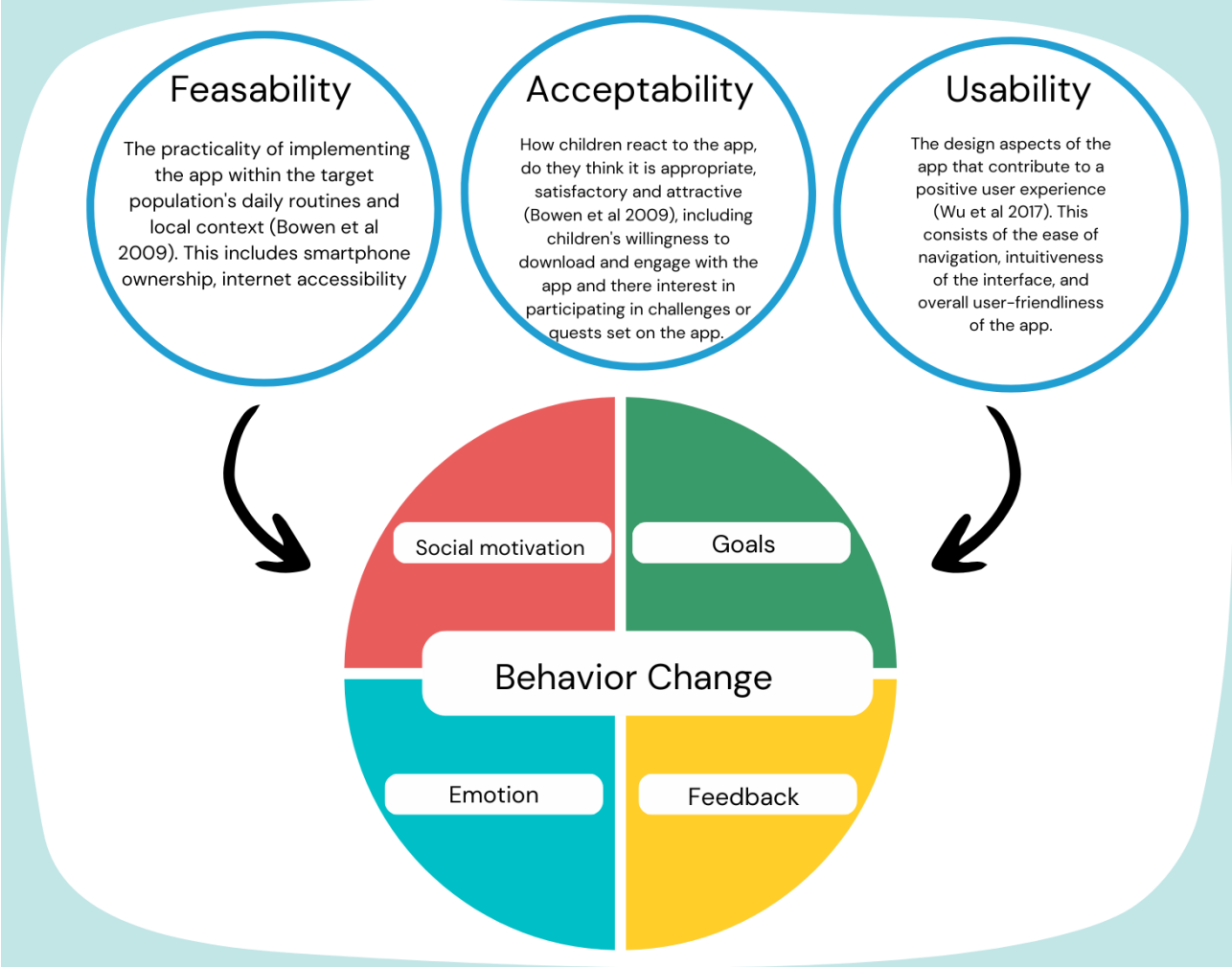

Feasibility

This theme, identified through the process of inductive coding, covers the download process, social support, time pressures surrounding using the app, and the community and outside environment.

The Download Process

Participants highlighted the app's ease of download and its small file size as practical advantages. A secondary school student remarked, "Yeah, because it's only a small download, like 38 megabytes" (Secondary school 1); this emphasizes its convenience, particularly in environments with limited storage or slow internet speeds. However, the requirement for providing a parental email during registration was perceived as a frustrating additional step in the process and a significant barrier to downloading the app, especially for older YCSs who preferred greater autonomy. As a participant noted:

> *Having to have your parents [involvement] is another thing that just will add annoyance to downloading it.* [Secondary school 1]

This suggests that streamlining the registration process to accommodate varying levels of independence for children accessing the app could enhance its appeal.

Social Support

Social support from friends and family, such as playing with and helping one another, was mentioned by several children, for example:

> *Mummy helps me with counting and doing the stopwatch.* [Year 4, Primary school 1]

This was echoed by other children in year 4 who discussed family members helping them with the activities and the app. Whereas a child in year 6 (Primary school 1) mentioned involving a friend:

> *Played it with my friend, she liked the skipping and being active.* [Year 6, Primary school 1]

In addition to support from family and friends, children also valued the shared experience of using the app collaboratively. A participant mentioned, "It's more fun to do the activities with someone, like when we shared the tablet" (Year 5, Primary school 2), suggesting that facilitating collaborative usage could further enhance the app's appeal. These suggestions emphasize the value of integrating both collaborative and competitive features within the app to foster social engagement, which may in turn boost motivation and sustained use.

Furthermore, involving a community element such as linking to the "Junior Park Runs" (Secondary school 1) could increase

XSL•FO
RenderX

social support beyond home and the classroom and therefore create a socially reinforced environment for behavior change.

Connecting With the Local Area

The children mentioned several local areas they were familiar with that they would like added to the app where they could find QR codes. In addition, children mentioned using the app outside:

> *I was in the park, and I was doing like hanging that thing, then I was doing star jumps, and then I had to do a little run.* [Year 5, Primary school 2]

However, one child raised a concern that when they were in the park, they had "other things to be doing," and when questioned further, they answered "like playing football or going on the slides" (Year 5, Primary school 2).

Acceptability

This theme includes the children's opinions of the overall app, the reward system, and customizable features. The types of activities and tutorials or support offered are also discussed. Children made positive comments about the app in general, such as "It encourages us to be active" and "decent activities" (Year 5, Primary school 2).

The app's gamified features, such as points and rewards, were generally well-received and seen as motivating. A participant stated (child jumping up and down exclaiming), "We got 20 points!" (Primary school 2), with another expressing, "I like it because you get points" (Year 4, Primary school 1).

This emphasizes the immediate gratification experienced by children offered by the feedback within the app. YCSs also expressed a desire for proportional rewards based on the difficulty or physical demand of the activity, suggesting:

> *10 or less, you should get 10 points; 10 to 20—you should get 20. And then if it's 30 or more, then you can have 50.* [Secondary school 1]

This highlights the potential for a more dynamic reward structure that reinforces individual effort on a sliding scale.

Customization was another important factor in acceptability. YCSs appreciated designing avatars—when you log into the app, you can choose a character and adapt their hair and skin color, clothes, shoes, and accessories. However, they emphasized the need for further personalization and representativeness, including culturally relevant options. A child noted:

> *That's one thing I'm like, where are the hijabs? This is for people in Bradford.* [Secondary school 2]

This was echoed in another focus group where children asked for "A hijab," ''abaya," and "more hijabs and jilbabs" (Year 5, Primary school 2).

Further options were requested in every focus group, including "a way to design your own avatar" (Year 5, Primary school 2).

Customization was seen positively, but more options were requested:

> *I liked that you could change clothes, but we need more colors and fun stuff like dresses or branded items like Nike.* [Primary school 1]

YCSs also requested the ability to add pet, sports equipment, and different accessories to their avatars, the focus group from Year 6 (Primary school 1) suggested: "Skipping ropes, hula hoops, footballs, basketballs" and the Year 6 (Primary school 2) adding "'cats,' 'dogs,' 'birds,' 'snakes,' and 'pandas'" and a sample from Year 5 (Primary school 1) suggesting "'Unicorns,' 'butterflies,' 'grasshoppers,' 'tigers,' 'cats,' and 'leopards'" and Year 5 (Primary school 2) focusing more on accessories such as: "'headphones' 'golden guitar,' 'emerald chain,' and 'glasses.'"

These comments underline the importance of inclusive and diverse options in avatar design to cater to varied cultural, social, and creative preferences.

In addition, seasonal themes and event-based updates were requested in several focus groups. Suggestions such as "Maybe a light and dark mode or something like that" (Secondary school 1) and requests to "change the background" (Year 5, Primary school 2) reflect a desire for customizable visual settings. In addition, a participant highlighted the value of regular updates by stating, "Every so often it just needs to have an update to keep it being exciting" (Secondary school 2), while seasonal requests like "Christmas theme or summer outfits" (Year 5, Primary school 2) point to the potential for event-based content to maintain long-term engagement.

Participants proposed a diverse array of activities beyond traditional sports. Several children suggested physical activities such as "cooking and cleaning" (Year 6, Primary school 1) and even using a "5-minute timer to do a chore and you get points" (Year 5, Primary school 2), indicating that incorporating everyday tasks could enhance the app's relevance. Similarly, for alternative sports-related activities, a participant noted, "If dance was added, I would have used the app a little bit more" (Secondary school 1), while others mentioned options like "20 minutes playing sports," and activities, including "racing, gymnastics, dancing, jogging on the spot, football, traffic lights, basketball, dance to the beat, and cricket" (Year 5, Primary school 2). These suggestions highlight that a broader range of activity types could increase the app's appeal and acceptability by catering to varied interests and daily routines.

To improve the acceptability of the app, YCSs suggested that accessing tutorial resources through the app would help. As one participant stated, "I don't know how to do push-ups, but I tried. Maybe a tutorial or video would help" (Primary school 2). Providing instructional resources could enhance the children's understanding of certain exercises.

Children's emotional reactions to the app and its activities were multifaceted. For instance, one child remarked, "It just like reminds you how you feel when you do like different types of things" (Secondary school 1), while others used terms like "over the moon," "overjoyed," "tiring," "exhausting," and "ecstatic" (Years 6 and 4, Primary school). These varied responses suggest that while the app is engaging and can evoke positive emotions, some activities may also be perceived as physically demanding.

XSL•FO
RenderX

Social collaboration was also highlighted as a positive aspect. A participant shared:

> *It'd be fun to see how I'm doing compared to my friends.* [Year 6, Primary school 2]

This underscores the importance of fostering cooperative experiences to enhance the app's appeal in social contexts.

Conversely, a child noted:

> *We got 40 points! I like that we can compete with others to see who's better.* [Primary school 1]

Usability

Topics covered under this theme include functionality, clarity of features, and app navigation. The app's interface was generally described as intuitive and easy to navigate. A child noted referring to the activity dashboard—The activity dashboard is a simple graph of the number of activities completed and a step count over the week:

> *It's cool to see all the things I've done.* [Secondary school 1]

However, some participants expressed confusion about certain features, which highlights the importance of including clearer terminology and guidance to support user navigation within the app. For instance, one participant stated, "When I went on, I didn't know what 'solo quest' meant" (Secondary school 2) and suggested adding further guidance focused around how to use the app, though this was not supported by the other young people in the group. This potentially highlights the need for clearer explanations and guidance within the app to support user understanding.

YCSs also suggested adding a sound notification as an additional cue to mark the end of an activity, enhancing usability during active tasks:

> *Something that they could watch and something that they could do whilst, whilst they were like, on their five, on like a five-minute run or something. Because otherwise they are just stood there running, watching their clock countdown.* [Secondary school 1]

Collaborative features were appreciated for fostering competition and camaraderie, particularly in school settings. A child remarked:

> *If you do it at school, there's more people being competitive, like in a class.* [Secondary school 1]

Emotional tracking refers to prompts that ask users to reflect on how they feel before and after engaging in physical activity, designed to encourage mindfulness about the emotional benefits of physical activity. Feedback on the emotion tracking feature varied, with some participants finding it valuable while others felt it was overused. A child explained, "It just like reminds you how you feel when you do like different types of things" (Secondary school 1 pupil), but they also suggested reducing its frequency to avoid frustration.

A participant also suggested that notifications could remind users to log their emotions even if they had not been active that day:

> *If a notification popped up? Yeah, it'd make you remember to go on the app… like a lot of the time I won't remember to go on any apps* [Secondary school 2]

This indicates a need to balance the feature's usability with user preferences. Participants also suggested other types of prompts to enhance engagement. For instance, a child proposed a "daily reward to make people want to play it more" (Secondary school 2), while another recommended that the app should ask, "Do you want to take a break now?" (Year 6, Primary school 2) if used for extended periods. These ideas point to the potential benefits of dynamic prompting to sustain user interest.

Suggestions for improving the app's functionality focused on integration with existing health tools, which could align the app with user habits and enhance its utility. For example, a secondary school participant proposed linking the app to Samsung Health or Google Fitbit, stating:

> *My phone with Samsung Health app automatically tracks my steps. If that was implemented [integrated], it would also see the steps.* [Secondary school 1]

This feature would simplify tracking and reduce user effort, making the app easier to use and, therefore, appealing. Beyond step counting via Samsung Health, participants also expressed interest in additional biometric features. Comments such as "I want to track my heart rate" (Year 6, Primary school 1) indicate a demand for more comprehensive health monitoring. Furthermore, a participant queried, "You know the step activities, does the app track your steps?" (Year 4, Primary school 1).

Furthermore, several participants expressed a desire for multidevice compatibility. One noted, "In Bestlife I couldn't do it on my laptop" (Year 6, Primary school 1), while another stated, "I want to do it on my Switch" (Year 6, Primary school 1). This suggests that extending support beyond smartphones may enhance usability by accommodating various user preferences.

Overall, improving clarity in the interface, integrating social and competitive features, and offering customizable options for auditory feedback would address key user needs and enhance the app's usability and engagement.

Behavior Change Techniques

The behavior change techniques theme highlights how Bestlife encourages engagement through goal-setting, external motivation, and reward-based reinforcement. The app prompts children to set clear activity goals via quests, which serve as external motivators and encourage physical activity.

> *It's like, fun to get points for doing activities like jumping jacks.* [Primary school 2]
>
> *I enjoyed doing the push-ups because it helps my body.* [Year 4, Primary school 1]

These examples show how quests function both as motivational prompts and as tools for establishing behavioral goals.

XSL•FO
RenderX

Points earned from completing quests act as immediate, visible rewards, reinforcing children's efforts and contributing to sustained engagement.

> *We got 20 points!* [Primary school 2]

Other children expressed that effort should be rewarded proportionately:

> *If someone worked harder, they should get more points.* [Year 6, Primary school 1]
>
> *More points for doing jumping jacks than walking.* [Secondary school 2]

These comments suggest that a graded reward system could enhance fairness and motivation. While many children valued these external incentives, some also described intrinsic enjoyment in physical activity itself:

> *Riding my bike is fun, I don't need points for that.* [Secondary school 1]

The app also includes emotional tracking before and after activity to encourage self-reflection. However, children had mixed views about its frequency:

> *Once a day would be better.* [Secondary school 2]
>
> *Maybe like one before you do a full quest and then one after.* [Secondary school 1]
>
> *Don't ask again button.* [Year 5, Primary school 2]

These responses suggest that reducing the frequency or offering flexible options could enhance the user experience while preserving the reflective benefits.

Social support emerged as a key driver of engagement. The "collab mode" allows users to team up with friends or classmates, enhancing motivation through shared goals.

> *It's good because you can collab with your friends.* [Primary school 2]

Social influences extended beyond peers. A participant remarked:

> *I'd probably end up doing it if my mum nagged at me hard enough.* [Secondary school 1]

Competitive dynamics, particularly in school settings, also contributed:

> *If you do it at school, there's more people being competitive, like in a class.* [Secondary school 2]

A child suggested linking the app to household tasks:

> *It would persuade you to do it. I'm not going to lie I'm quite lazy when my mum says go do your chores [name]. So it could be a way to persuade. Competitiveness with yourself.* [Year 5, Primary school 2]

Progress tracking features, such as the activity dashboard, enabled children to monitor achievements and maintain motivation. Children set specific, measurable goals:

> *Let's see if we got more than 20.* [Year 4, Primary school 1]
>
> *We set a goal to do 1000 steps a day.* [Secondary school 1]

Longer-term goals also played a role:

> *Trying to reach the next level made me keep going.* [Secondary school 2]

Reflecting on progress, a child shared:

> *It's cool to see all the things I've done.* [Secondary school 1]

Another added:

> *I liked seeing how much I did and comparing it with others.* [Primary school 1]

To sustain engagement, a participant emphasized the importance of variety:

> *Every so often it just needs to have an update to keep it being exciting.* [Secondary school 2]

In summary, the app effectively integrates several behavior change techniques, including goal-setting, rewards, social support, and self-monitoring. Participant feedback suggests that while these features support engagement, refinements, such as adjusting emotional tracking frequency and personalizing reward systems, could further enhance the app's effectiveness and appeal.

Discussion

Principal Findings

This study found that the Bestlife app was generally feasible, acceptable, and user-friendly for children aged 8-13 years. Positive engagement was particularly driven by gamified features, social collaboration, and extensive personalization options. Notably, children responded positively to features enabling self-monitoring, goal-setting, and proportional rewards, while highlighting the importance of aligning digital health tools with children's daily routines and social contexts. While some findings confirm established insights from previous research, our study extends the literature by using a citizen science methodology that actively involved children as co-evaluators, uncovering specific barriers, such as the parental email registration requirement, that could limit autonomy and engagement among older children. Furthermore, children's explicit requests for culturally inclusive customization, including diverse clothing and accessories, provide novel and actionable insights for culturally sensitive app development. These findings not only inform ongoing refinements of the Bestlife app but also offer broader, child-driven recommendations for future mHealth interventions targeting diverse youth populations (see Figure 2).

XSL•FO
RenderX

Figure 2. Evidence-based recommendations for designing physical activity apps for children.

XSL•FO
RenderX

Feasibility

Our findings indicate that implementing the Bestlife app was generally feasible in a real-world setting. Children reported that the app's small download size and easy installation facilitated uptake even on devices with limited storage or bandwidth. This practical design choice helped avoid technical barriers that often hinder mHealth adoption in low-resource environments. Previous research similarly emphasizes that minimizing technical demands (eg, download size, bugs, etc) is crucial for user retention [60-62]. Our study uniquely highlights how a small app size, when combined with familiar contexts such as school-based rollout, may mitigate infrastructure-related barriers, a factor that remains underexplored in current mHealth literature targeting children and young people. By ensuring Bestlife ran reliably on basic devices, the developers addressed a key feasibility concern identified both in our study and the literature. This insight was uncovered through a participatory citizen science approach, where children acted not just as users but as co-evaluators of the intervention. Their feedback, gathered through real-world app use in everyday settings, strengthens the validity of these findings and surfaces nuanced design implications that might be missed in traditional researcher-led evaluations.

Despite these strengths, our participants revealed a procedural hurdle: the mandatory parental email registration. Older children in particular found the required parental sign-up frustrating, as it delayed or limited their independent access to the app's features. While parental oversight is important for ensuring children's safety, our findings suggest that this requirement inadvertently discouraged some users. This is consistent with observations where mandatory parental consent for app use and research participation can act as a barrier, particularly for adolescents seeking real-time engagement with digital tools [63]. Simplifying the sign-up process or offering age-appropriate alternatives (such as using school credentials) could improve feasibility without compromising child safety [63,64]. Streamlining the onboarding experience is a clear design implication from our study: future child-focused apps should balance necessary parental controls with a frictionless entry for young users.

The time constraints children face were another barrier to feasibility. Children have busy schedules, and some struggled to fit app sessions into their daily routine. In our focus groups, a few children mentioned that homework, chores, or other activities left little time for using Bestlife regularly. To address this, participants discussed integrating app activities into existing routines, for example, incorporating short activity challenges that would monitor chores. This finding suggests that making the app complement rather than compete with daily activities would enhance its feasibility for sustained use. Embedding health prompts into everyday life is known to support habit formation [45]. Consistent with this idea, research indicates that aligning interventions with children's usual schedules (eg, walking to school or family playtime) increases engagement [65]. Future designs should therefore consider seamless integration into children's routines.

Acceptability

Children found the Bestlife app to be acceptable overall, and they were enthusiastic about its gamified features and theme. Children responded positively to the app's points, rewards, and activity elements, which made being physically active fun. The participants described different ways gamification features like earning points for completing exercises motivated them, indicating that these elements successfully tapped into their sense of play. This aligns with previous findings that game-like rewards and feedback can boost engagement in youth physical activity interventions [35,66]. In the Step It Up Family program, for instance, competitive step challenges and leaderboards were key motivators for kids and parents alike [35]. However, our citizen science data offer new insight into how children prefer collaborative over competitive play, suggesting that app-based motivation may be more sustainable when framed as team success rather than individual rivalry.

The Bestlife users similarly enjoyed friendly competition, such as comparing points in collab challenges; but importantly, many expressed a preference for collaboration over pure competition. Rather than outdoing friends, children said they loved team-based challenges and shared participation, especially when used in their classroom groups. This is a critical insight: while competition can spur short-term engagement, collaborative play created enjoyment and fostered a sense of community. Past research supports this, noting that cooperation and social support can enhance motivation without the downsides of rivalry [23,31]. These findings suggest that mHealth interventions targeting children may benefit from prioritizing peer collaboration to foster inclusivity, especially within school environments.

Cultural relevance and inclusivity were also highlighted as important. Children in diverse Bradford schools appreciated Bestlife's avatar customization but also pointed out its limited options reflecting different cultures and identities. For example, some girls noted the lack of modest clothing or hijab options for avatars, and others wished for attire or accessories that resonated with their cultural background. This extends existing literature by providing grounded, child-led evidence that superficial customization is insufficient, children expect inclusive representation. When children cannot create an avatar that looks like them or reflects their lifestyle, the app may feel less welcoming. Our data underscore the value of inclusive design: several children explicitly suggested adding a wider range of skin tones, clothing styles, and accessories. In the broader literature, inclusivity in game design has been linked to greater user identification and satisfaction [67]. Providing diverse options and culturally sensitive content can make health apps more appealing to underserved groups [68-70]. Our findings support this and additionally offer actionable design recommendations: modest sportswear, culturally relevant hairstyles, and seasonal or religious festival themes could meaningfully improve acceptability in multiethnic youth populations.

Participants also gave feedback on how to keep the app interesting and acceptable over the long term. A common suggestion was to introduce customizable notifications and

XSL•FO
RenderX

dynamic content updates. Children liked the idea of gentle reminders, for instance, a smartphone notification nudging them to do a quick activity or themed holiday linked reminders. This aligns with evidence that regular prompts can re-engage users and sustain healthy behaviors [66,71]. Importantly, the tone and timing of these reminders must be child-friendly; generic text-message prompts used in adult interventions have shown limited impact on children [35].

Usability

The usability of the Bestlife app was generally well-received, with most children finding it easy to navigate and use. This intuitive navigation is a strong point of the app's design. It suggests that even first-time users were able to explore features independently, which is crucial for mHealth interventions targeting children. Previous studies note that ease of use significantly influences whether young people keep using an app [72,73]. Our findings reinforce that principle: a child-friendly interface, with straightforward menus, enabled engagement without frustration. Notably, conducting focus groups in classroom settings allowed us to observe that many children required little help from adults or peers to use Bestlife, indicating a high level of usability in its current form.

Despite this overall positive feedback, our study also uncovered challenges for certain users, highlighting the need to accommodate varying levels of digital literacy among children. A subset of the participants less familiar with smartphones struggled with some of the in-app terminology and task flow. For instance, a few children were confused by labels like or did not immediately understand how to progress from one challenge to the next without guidance. These instances show that what older or more tech-savvy children find intuitive might not be obvious to all users.

Our data suggest a balanced solution: include optional tutorials and visual guides for those who need them. Several children in the focus groups said they would have used a short tutorial video if available, whereas others felt they did not need it. Providing an on-demand tutorial (for example, a friendly animated guide explaining how to use the app) could support less digitally literate users without forcing mandatory instruction on those who prefer to learn by exploration. This approach echoes design practices in other child-oriented apps; for example, the Jungle Gym (Learning Games Lab) physical activity app introduced a toggleable guide for new users, which improved usability for younger children while older kids ignored it once they felt confident [61]. In line with previous recommendations on digital health literacy [74-76], we recommend that future mHealth apps for youth conduct age-specific usability testing to ensure interface text and navigation are easily understood by the youngest users. By refining terminology and offering help, Bestlife and similar apps can become truly accessible to children of all ages and skill levels.

Overall, our findings underline that usability is not one-size-fits-all for children. Successful apps must cater to a range of ages and capabilities, providing simplicity for novices and depth for experienced users. We recommend that future mHealth interventions adopt a layered usability model, offering core functionality up front with optional depth and scaffolding for those who need it. By embracing inclusive design and iterative usability testing, future interventions can ensure that technical usability issues do not become a barrier to health engagement and equity.

Behavior Change

Our qualitative evidence suggests that the Bestlife app incorporated several behavior change techniques (BCTs) that resonated with children and showed promise for encouraging physical activity. Children mentioned features such as setting personal goals, monitoring their progress, and sharing achievements with friends, all of which are well-established strategies in behavior change theory [45]. For example, goal-setting and self-monitoring were embedded within Bestlife through activity challenges and dashboards that visually displayed points earned. These features effectively tapped into different motivational drivers: some children expressed intrinsic motivation derived from personal mastery and achievement, such as experiencing pride in improving their personal records (eg, completing more push-ups than previously), while others highlighted extrinsic motivation through social recognition and tangible rewards, such as points and avatar upgrades [77]. Distinguishing these motivational pathways provides valuable insights for future interventions, suggesting tailored features to cater separately or simultaneously to intrinsic and extrinsic motivations to maximize engagement and sustainability.

Feedback mechanisms within Bestlife also significantly contributed to promoting behavior change. The app's well-being tracker enabled children to log mood or energy levels, enhancing awareness of the emotional benefits associated with physical activity. Some participants explicitly noted connections between activity and mood improvements, describing physical challenges as "fun." Such reflective feedback reinforces behavior change by validating efforts and encouraging continued participation [45]. Future iterations of mHealth apps should continue integrating feedback loops, which clearly connect children's physical activities with emotional outcomes [40].

Implications for Future Research and Practice

The iterative development process based on user feedback was a significant strength of this study. By directly engaging with children and incorporating their suggestions such as higher-value store items, modest clothing options, and tutorial videos, the app evolved in real time to better meet the needs and preferences of its target users. These iterative updates not only enhanced usability but also ensured that the app's features were grounded in evidence supporting the modeling of behavior change in physical activity interventions [26]. This study clearly demonstrates the critical value of a participatory design approach, where children's voices drive the evolution of the intervention. We strongly recommend that future app development projects adopt a similar iterative co-design process, as it fosters responsive, user-centered design and ultimately leads to more engaging and effective digital interventions.

While Bestlife seeks to address physical inactivity among young people through engaging digital media, it is essential to critically examine the broader implications of promoting smartphone use in this age group. Increasing advocacy for delaying smartphone

XSL•FO
RenderX

ownership until 16 years of age reflects growing concerns about the risks to mental health and online safety associated with early digital engagement [78,79]. Research suggests that early and prolonged smartphone exposure may contribute to poor mental well-being, cyberbullying exposure, and reduced real-world social interactions [17,80]. These concerns raise an important ethical question: by creating attractive, gamified content aimed at young users, are we unintentionally reinforcing or accelerating problematic digital behaviors? To mitigate these risks, the deployment of Bestlife must continue to be guided by a framework that includes parental oversight, clear usage boundaries, and an emphasis on balance between digital and offline activities [21,81], ensuring we support and not compromise young people's long-term well-being.

The findings of this study contribute to the growing evidence base for child-focused mHealth interventions. They underscore the importance of integrating behavior change techniques with user-centered design principles while addressing practical barriers to accessibility and engagement. Future physical activity apps designed for children should prioritize inclusive customization options, reward systems, engaging notifications, simplifying the setup process, device integration, collaborative challenges, and frequent updates (see Figure 2).

Future research should focus on evaluating the app's long-term impact on physical activity to see if it is still being used and to assess children's behavior change, and exploring its scalability in diverse cultural and socioeconomic contexts. By situating the findings within the broader landscape of mHealth research, this study highlights the potential of well-designed digital tools to promote physical activity among children while addressing the complexities of behavior change.

Strengths and Limitations

This study offers valuable insights into the design and evaluation of a child-focused mHealth app, underpinned by several key strengths. The use of a contributory citizen science approach, where children acted as active co-evaluators rather than passive participants, ensured that findings were deeply rooted in lived experience. Children trialed the app in real-world settings and provided direct feedback, shaping both the evaluation and the app's ongoing development. This participatory model not only supports authentic engagement but also enables iterative adaptation of app features in response to children's feedback, such as requests for more diverse avatar customization options and higher-value in-app rewards.

However, the collaborative school-based setting that made this feedback-rich approach possible also introduces potential bias. Group-based focus sessions created an enthusiastic and supportive environment, which may have inflated perceived engagement compared to independent use at home. While peer encouragement likely encouraged more expressive and positive feedback, it may not fully reflect solitary user experiences. Future studies could compare app engagement in individual versus group contexts to more clearly understand this dynamic.

Another notable strength was the diverse participant sample, spanning a wide age range (8 - 13 y) and including children from multiethnic, socioeconomically varied backgrounds in Bradford, a city with high levels of childhood inactivity and digital inequality. Capturing the perspectives of these underrepresented groups enhances the relevance of our findings for public health. Nonetheless, some sampling limitations were present. Due to the need for active parental consent, not all children who used the app were included in the formal research sample. This may have skewed participation toward families already more engaged in school or research activities, limiting representativeness.

The study's embeddedness in familiar school environments was another strength that contributed to open, honest dialog. Conducting focus groups in classrooms, often with trusted teachers nearby, fostered comfort and richer interaction. Yet, this embedded approach also constrained the scope of behavioral observation. Children used the app for a relatively short period (1 - 2 weeks), which limited our ability to assess sustained behavior change or long-term retention. While early indicators of motivation and habit formation were promising, and children themselves proposed features (like reminders and content updates) to support longer engagement, we were not able to evaluate these strategies over time. A longitudinal design, paired with app analytics, would strengthen future evaluations.

Ethically, the study upheld high standards of informed consent and child autonomy, for instance, respecting a child's choice to withdraw despite parental permission. Still, we recognize that requiring written parental consent can restrict inclusion, particularly in more marginalized families. In future low-risk studies of this nature, opt-out consent models could help broaden participation while maintaining ethical safeguards.

Finally, while the study offers actionable learning for mHealth development, its findings are inevitably shaped by the specific cultural and geographic context of one UK city. Bestlife was tailored for use in Bradford, and although many design recommendations are broadly transferable, further testing in other locations and demographics is needed to assess generalizability.

Despite these limitations, the study provides credible, grounded, and highly applicable insights for those designing digital interventions for children. The emphasis on children's voices throughout enhances the study's relevance, and the real-world setting lends practical depth that many controlled evaluations lack. These findings lay a strong foundation for future research and development of inclusive, engaging, and behaviorally effective digital tools to support children's health.

Conclusion

As digital health tools increasingly shape the health behaviors of children, understanding what truly resonates with young users has never been more critical. This formative research contributes valuable insight to the limited literature on children's experiences with physical activity apps. By exploring user perceptions of app features, content, and usability, this study identified key design elements that enhance engagement such as gamification and opportunities for collaboration. These findings not only inform the ongoing refinement of the Best Life app but also offer practical guidance for future developers aiming to design child-centered, culturally inclusive digital

XSL•FO
RenderX

interventions. In addition, this research provides a foundation for using such apps to both promote physical activity and collect meaningful data on young people's health behaviors in real-world contexts.

Acknowledgments

The authors thank all the citizen scientists and school staff for their time and contribution. The authors would also like to acknowledge the contribution of Dubit, who developed the Bestlife app. Additional thanks go to the assistant researchers who supported data collection: Laurie Ewels, Mevagh Harris, Elliot Lever, Shania Boom, and Emma Young.

The authors acknowledge the use of OpenAI's ChatGPT (version 3.5 2024/2025) to support the refinement of manuscript wording and phrasing during the drafting process. No artificial intelligence (AI)-generated content was used without authors' review and editing, and no references, data interpretation, or analysis were generated by AI. All content has been fact-checked by the authors, who take full responsibility for its accuracy and integrity.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to the qualitative nature of the study and the challenges of fully anonymizing focus group transcripts without compromising the contextual integrity and meaning of participants' responses but are available from the corresponding author on reasonable request.

Authors' Contributions

JH and MTF conceived of the study as part of the Join Us: Move. Play (JU:MP) evaluation and MTF's PhD. LC and MTF designed the Bestlife booklet and focus group guide with oversight and input from FB, MS-P, and JH. MTF led the focus groups with support from LC. MTF, LC, LW, and AC analyzed and interpreted data. MTF drafted the final manuscript, and all authors participated in critical revisions of the manuscript. All authors have read and approved the final manuscript.

Conflicts of Interest

JB was JU:MP's (Join Us: Move. Play) director and commissioned the app development. None declared by the other authors.

Multimedia Appendix 1
How to play.
[DOCX File, 378 KB - formative_v9i1e76498_app1.docx]

Multimedia Appendix 2
Booklet for citizen scientists to complete at home.
[PDF File, 1405 KB - formative_v9i1e76498_app2.pdf]

Multimedia Appendix 3
Focus group guide.
[DOCX File, 16 KB - formative_v9i1e76498_app3.docx]

Multimedia Appendix 4
Code book.
[DOCX File, 26 KB - formative_v9i1e76498_app4.docx]

Checklist 1
COREQ checklist.
[PDF File, 372 KB - formative_v9i1e76498_app5.pdf]

References

1. Bailey R, Hillman C, Arent S, Petitpas A. Physical activity: an underestimated investment in human capital? J Phys Act Health 2013 Mar;10(3):289-308. [doi: 10.1123/jpah.10.3.289] [Medline: 23620387]
2. Physical activity. WHO. 2024. URL: https://www.who.int/news-room/fact-sheets/detail/physical-activity [accessed 2025-06-13]
3. Eyre ELJ, Duncan MJ. The impact of ethnicity on objectively measured physical activity in children. ISRN Obes 2013;2013:757431. [doi: 10.1155/2013/757431] [Medline: 24555154]

XSL•FO
RenderX

4. Love R, Adams J, van Sluijs EMF, Foster C, Humphreys D. A cumulative meta-analysis of the effects of individual physical activity interventions targeting healthy adults. Obes Rev 2018 Aug;19(8):1164-1172. [doi: 10.1111/obr.12690] [Medline: 29701299]
5. Dumith SC, Gigante DP, Domingues MR, Kohl HW, Kohl HW. Physical activity change during adolescence: a systematic review and a pooled analysis. Int J Epidemiol 2011 Jun;40(3):685-698. [doi: 10.1093/ije/dyq272] [Medline: 21245072]
6. Bagnall AM, Radley D, Jones R, et al. Whole systems approaches to obesity and other complex public health challenges: a systematic review. BMC Public Health 2019 Jan 3;19(1):8. [doi: 10.1186/s12889-018-6274-z] [Medline: 30606173]
7. Telama R, Yang X, Viikari J, Välimäki I, Wanne O, Raitakari OT. Physical activity from childhood to adulthood: a 21-year tracking study. Am J Prev Med 2005 Apr;28(3):267-273. [doi: 10.1016/j.amepre.2004.12.003] [Medline: 15766614]
8. Active lives children and young people survey - academic year 2022-23. : Sport England; 2023.
9. Ryan D, Nutting H, Parekh C, et al. Ready, set, co(produce): a co-operative inquiry into co-producing research to explore adolescent health and wellbeing in the Born in Bradford Age of Wonder project. Res Involv Engagem 2024 Apr 30;10(1):41. [doi: 10.1186/s40900-024-00578-y] [Medline: 38689373]
10. Bull FC, Al-Ansari SS, Biddle S, et al. World Health Organization 2020 guidelines on physical activity and sedentary behaviour. Br J Sports Med 2020 Dec;54(24):1451-1462. [doi: 10.1136/bjsports-2020-102955]
11. WHO Global Observatory for eHealth. mHealth: new horizons for health through mobile technologies: second global survey on eHealth. WHO. 2011. URL: https://apps.who.int/iris/handle/10665/44607 [accessed 2023-05-04]
12. Children and parents: media use and attitudes report 2023. Ofcom. 2023. URL: https://www.ofcom.org.uk/media-use-and-attitudes/media-habits-children/children-and-parents-media-use-and-attitudes-report-2023/ [accessed 2023-09-20]
13. Casey A, Goodyear VA, Armour KM. Rethinking the relationship between pedagogy, technology and learning in health and physical education. Sport Educ Soc 2017 Feb 17;22(2):288-304. [doi: 10.1080/13573322.2016.1226792]
14. Greenhow C, Lewin C. Social media and education: reconceptualizing the boundaries of formal and informal learning. Learn Media Technol 2016 Jan 2;41(1):6-30. [doi: 10.1080/17439884.2015.1064954]
15. Seah MLC, Koh KT. The efficacy of using mobile applications in changing adolescent girls' physical activity behaviour during weekends. Eur Phy Educ Rev 2021 Feb;27(1):113-131. [doi: 10.1177/1356336X20930741]
16. Kemp BJ, Parrish AM, Cliff DP. "Social screens" and "the mainstream": longitudinal competitors of non-organized physical activity in the transition from childhood to adolescence. Int J Behav Nutr Phys Act 2020 Jan 14;17(1):5. [doi: 10.1186/s12966-019-0908-0] [Medline: 31937315]
17. Abi-Jaoude E, Naylor KT, Pignatiello A. Smartphones, social media use and youth mental health. CMAJ 2020 Feb 10;192(6):E136-E141. [doi: 10.1503/cmaj.190434] [Medline: 32041697]
18. Pires RCR, Reinert A, Cantanhede AM, Pinheiro J. Associations between smartphone use, smartphone addiction and mental health in teenagers: a structural equation modeling approach. R Tecnol Soc 2024;20(59):50. [doi: 10.3895/rts.v20n59.16652]
19. Caninsti R, Rahayu SAT. The role of smartphone addiction on bed procrastination and mindful eating behavior in adolescents. indigenous 2024;9(1):1-13. [doi: 10.23917/indigenous.v9i1.2804]
20. Tater B, John K. Effects of smartphone addiction on the physical and mental well-being of indian students. IJST 2024 Jun 25;17(25):2622-2634 [FREE Full text] [doi: 10.17485/IJST/v17i25.1355]
21. Mohta R, Halder S. A comparative study on cognitive, emotional, and social functioning in adolescents with and without smartphone addiction. JIACAM 2021 Oct;17(4):44-65. [doi: 10.1177/0973134220210404]
22. Mollborn S, Limburg A, Pace J, Fomby P. Family socioeconomic status and children's screen time. J Marriage Fam 2022 Aug;84(4):1129-1151. [doi: 10.1111/jomf.12834] [Medline: 36211640]
23. Sharma S, Gergen Barnett K, Maypole JJ, Grochow Mishuris R. Evaluation of mHealth apps for diverse, low-income patient populations: framework development and application study. JMIR Form Res 2022 Feb 11;6(2):e29922. [doi: 10.2196/29922] [Medline: 35147502]
24. Sohn SY, Rees P, Wildridge B, Kalk NJ, Carter B. Prevalence of problematic smartphone usage and associated mental health outcomes amongst children and young people: a systematic review, meta-analysis and GRADE of the evidence. BMC Psychiatry 2019 Nov 29;19(1):356. [doi: 10.1186/s12888-019-2350-x] [Medline: 31779637]
25. Emberson MA, Lalande A, Wang D, McDonough DJ, Liu W, Gao Z. Effectiveness of smartphone-based physical activity interventions on individuals' health outcomes: a systematic review. Biomed Res Int 2021;2021:6296896. [doi: 10.1155/2021/6296896] [Medline: 34409104]
26. Conroy DE, Yang CH, Maher JP. Behavior change techniques in top-ranked mobile apps for physical activity. Am J Prev Med 2014 Jun;46(6):649-652. [doi: 10.1016/j.amepre.2014.01.010] [Medline: 24842742]
27. Bezabih AM, Gerling K, Abebe W, Abeele VV. Behavioral theories and motivational features underlying ehealth interventions for adolescent antiretroviral adherence: systematic review. JMIR Mhealth Uhealth 2021 Dec 10;9(12):e25129. [doi: 10.2196/25129] [Medline: 34890353]
28. Bort-Roig J, Gilson ND, Puig-Ribera A, Contreras RS, Trost SG. Measuring and influencing physical activity with smartphone technology: a systematic review. Sports Med 2014 May;44(5):671-686. [doi: 10.1007/s40279-014-0142-5] [Medline: 24497157]

XSL•FO
RenderX

29. Ford KL, Moore SL, Zhou S, Gore MO, Portz J, Zhang X, et al. Advancing evidence-based digital health through an innovative research environment: an academic-industry collaboration case report. mHealth 2019 Sep 20;5 [FREE Full text]
30. Wang Y, Wang Y, Greene B, Sun L. An analysis and evaluation of quality and behavioral change techniques among physical activity apps in China. Int J Med Inform 2020 Jan;133:104029. [doi: 10.1016/j.ijmedinf.2019.104029] [Medline: 31759245]
31. Krouwel M, Greenfield SM, Chalkley A, et al. Promoting participation in physical activity through Snacktivity: A qualitative mixed methods study. PLoS ONE 2023;18(9):e0291040. [doi: 10.1371/journal.pone.0291040] [Medline: 37695785]
32. Brannan MGT, Foster CE, Timpson CM, et al. Active 10 - A new approach to increase physical activity in inactive people in England. Prog Cardiovasc Dis 2019;62(2):135-139. [doi: 10.1016/j.pcad.2019.02.001] [Medline: 30796943]
33. Liang H, Wang X, An R. Influence of Pokémon GO on physical activity and psychosocial well-being in children and adolescents: systematic review. J Med Internet Res 2023 Nov 13;25:e49019. [doi: 10.2196/49019] [Medline: 37955949]
34. Schoffman DE, Turner-McGrievy G, Jones SJ, Wilcox S. Mobile apps for pediatric obesity prevention and treatment, healthy eating, and physical activity promotion: just fun and games? Transl Behav Med 2013 Sep;3(3):320-325. [doi: 10.1007/s13142-013-0206-3] [Medline: 24073184]
35. Schoeppe S, Salmon J, Williams S, et al. Feasibility of using activity trackers and apps to increase physical activity in whole families: The Step it Up Family intervention. Digit Health 2022;8:20552076221129083. [doi: 10.1177/20552076221129083] [Medline: 36225986]
36. Odenigbo IP, Reen JK, Eneze C, Friday A, Orji R. Virtual, augmented, and mixed reality interventions for physical activity: a systematic review. Presented at: 10th IEEE International Conference on Serious Games and Applications for Health (SeGAH 2022); Aug 10-12, 2022 URL: http://www.scopus.com/inward/record.url?scp=85146227420&partnerID=8YFLogxK [accessed 2025-10-23]
37. Gkintoni E, Vantaraki F, Skoulidi C, Anastassopoulos P, Vantarakis A. Promoting physical and mental health among children and adolescents via gamification-a conceptual systematic review. Behav Sci (Basel) 2024 Jan 29;14(2):102. [doi: 10.3390/bs14020102] [Medline: 38392455]
38. Tong HL, Coiera E, Laranjo L. Using a mobile social networking app to promote physical activity: a qualitative study of users' perspectives. J Med Internet Res 2018 Dec 21;20(12):e11439. [doi: 10.2196/11439] [Medline: 30578201]
39. Simons D, De Bourdeaudhuij I, Clarys P, De Cocker K, Vandelanotte C, Deforche B. A smartphone app to promote an active lifestyle in lower-educated working young adults: development, usability, acceptability, and feasibility study. JMIR Mhealth Uhealth 2018 Feb 20;6(2):e44. [doi: 10.2196/mhealth.8287] [Medline: 29463491]
40. Hosseinpour M, Terlutter R. Your personal motivator is with you: a systematic review of mobile phone applications aiming at increasing physical activity. Sports Med 2019 Sep;49(9):1425-1447. [doi: 10.1007/s40279-019-01128-3] [Medline: 31144235]
41. Wang JW, Zhu Z, Shuling Z, et al. Effectiveness of mHealth app–based interventions for increasing physical activity and improving physical fitness in children and adolescents: systematic review and meta-analysis. JMIR Mhealth Uhealth 2024 Apr 30;12:e51478. [doi: 10.2196/51478] [Medline: 38687568]
42. Western MJ, Armstrong MEG, Islam I, Morgan K, Jones UF, Kelson MJ. The effectiveness of digital interventions for increasing physical activity in individuals of low socioeconomic status: a systematic review and meta-analysis. Int J Behav Nutr Phys Act 2021 Nov 9;18(1):148. [doi: 10.1186/s12966-021-01218-4]
43. Whitehead L, Talevski J, Fatehi F, Beauchamp A. Barriers to and facilitators of digital health among culturally and linguistically diverse populations: qualitative systematic review. J Med Internet Res 2023 Feb 28;25(1):e42719. [doi: 10.2196/42719] [Medline: 36853742]
44. What is JU:MP? Active Bradford. 2023. URL: https://www.activebradford.com/what-is-jump [accessed 2025-01-07]
45. Michie S, van Stralen MM, West R. The behaviour change wheel: A new method for characterising and designing behaviour change interventions. Implementation Sci 2011 Dec;6(1):42. [doi: 10.1186/1748-5908-6-42]
46. Born in Bradford. Bradford Institute for Health Research. URL: https://bradfordresearch.nhs.uk/researcharea/born-in-bradford/ [accessed 2024-12-09]
47. Bowen DJ, Kreuter M, Spring B, et al. How we design feasibility studies. Am J Prev Med 2009 May;36(5):452-457. [doi: 10.1016/j.amepre.2009.02.002] [Medline: 19362699]
48. Wu J, Tombor I, Shahab L, West R. Usability testing of a smoking cessation smartphone application ('SmokeFree Baby'): A think-aloud study with pregnant smokers. Digit HEALTH 2017;3:2055207617704273. [doi: 10.1177/2055207617704273] [Medline: 29942594]
49. Bonney R, Cooper CB, Dickinson J, et al. Citizen science: a developing tool for expanding science knowledge and scientific literacy. Bioscience 2009 Dec;59(11):977-984. [doi: 10.1525/bio.2009.59.11.9]
50. Ten principles of citizen science. ECSA. 2022. URL: https://www.ecsa.ngo/ [accessed 2023-01-07]
51. Shirk JL, Ballard HL, Wilderman CC, et al. Public participation in scientific research: a framework for deliberate design. E&S 2012;17(2). [doi: 10.5751/ES-04705-170229]
52. Wargers A, Queral J, Mölenberg FJ, et al. Citizen Science to improve healthy and active living among adolescents in four European countries: a protocol of the cluster randomised controlled trial of the Science Engagement to Empower aDolescentS (SEEDS) project. BMJ Open 2023 May 8;13(5):e070169. [doi: 10.1136/bmjopen-2022-070169] [Medline: 37156575]

XSL•FO
RenderX

53. Frazer M, Seims A, Tatterton MJ, et al. Child and family experiences of a whole-systems approach to physical activity in a multiethnic UK city: a citizen science evaluation protocol. BMJ Open 2023 Feb;13(2):e069334. [doi: 10.1136/bmjopen-2022-069334]
54. Frazer MT, Creaser A, Tatterton MJ, Daly-Smith A, Hall J. Exploring children and young people's experience of participating in citizen science-A qualitative evidence synthesis. PLOS ONE 2024;19(7):e0304976. [doi: 10.1371/journal.pone.0304976] [Medline: 38990968]
55. Ballard HL, Dixon CGH, Harris EM. Youth-focused citizen science: Examining the role of environmental science learning and agency for conservation. Biol Conserv 2017 Apr;208:65-75. [doi: 10.1016/j.biocon.2016.05.024]
56. Bender K, Barman-Adhikari A, DeChants J, et al. Asking for change: feasibility, acceptability, and preliminary outcomes of a manualized photovoice intervention with youth experiencing homelessness. Child Youth Serv Rev 2017 Oct;81:379-389. [doi: 10.1016/j.childyouth.2017.08.028]
57. Lane HG, Porter KJ, Hecht E, Harris P, Zoellner JM. A participatory process to engage appalachian youth in reducing sugar-sweetened beverage consumption. Health Promot Pract 2019 Mar;20(2):258-268. [doi: 10.1177/1524839918762123] [Medline: 29577771]
58. Adler K, Salanterä S, Zumstein-Shaha M. Focus group interviews in child, youth, and parent research: an integrative literature review. Int J Qual Methods 2019 Jan 1;18:160940691988727. [doi: 10.1177/1609406919887274]
59. Noonan RJ, Boddy LM, Fairclough SJ, Knowles ZR. Write, draw, show, and tell: a child-centred dual methodology to explore perceptions of out-of-school physical activity. BMC Public Health 2016 Apr 14;16(1):326 [FREE Full text] [doi: 10.1186/s12889-016-3005-1] [Medline: 27080384]
60. Smoll NR, Walker J, Khandaker G. The barriers and enablers to downloading the COVIDSafe app - a topic modelling analysis. Aust N Z J Public Health 2021 Aug;45(4):344-347. [doi: 10.1111/1753-6405.13119] [Medline: 33970555]
61. McCloskey M, Johnson SL, Benz C, et al. Parent perceptions of mobile device use among preschool-aged children in rural head start centers. J Nutr Educ Behav 2018 Jan;50(1):83-89. [doi: 10.1016/j.jneb.2017.03.006] [Medline: 29031581]
62. New IDS research shows poorest households trapped in digital poverty cycle. Institute of Development Studies. 2022. URL: https://www.ids.ac.uk/news/new-ids-research-shows-poorest-households-trapped-in-digital-poverty-cycle/ [accessed 2025-04-17]
63. Badillo-Urquiola K, Chouhan C, Chancellor S, De Choudhary M, Wisniewski P. Beyond parental control: designing adolescent online safety apps using value sensitive design. J Adolesc Res 2020 Jan;35(1):147-175. [doi: 10.1177/0743558419884692]
64. Baumel A, Muench F, Edan S, Kane JM. Objective user engagement with mental health apps: systematic search and panel-based usage analysis. J Med Internet Res 2019 Sep 25;21(9):e14567. [doi: 10.2196/14567] [Medline: 31573916]
65. McKay FH, Wright A, Shill J, Stephens H, Uccellini M. Using health and well-being apps for behavior change: a systematic search and rating of apps. JMIR Mhealth Uhealth 2019 Jul 4;7(7):e11926. [doi: 10.2196/11926] [Medline: 31274112]
66. Borghouts J, Eikey E, Mark G, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. J Med Internet Res 2021 Mar 24;23(3):e24387. [doi: 10.2196/24387] [Medline: 33759801]
67. Alqahtani F, Winn A, Orji R. Co-designing a mobile app to improve mental health and well-being: focus group study. JMIR Form Res. Preprint posted online on Feb 26, 2021. [doi: 10.2196/18172]
68. LeSeure P, Chin E, Zhang S. A culturally sensitive mobile app (DiaFriend) to improve self-care in patients with type 2 diabetes: development study. JMIR Diabetes 2024 Oct 21;9:e63393. [doi: 10.2196/63393] [Medline: 39432893]
69. Povey J, Mills P, Dingwall KM, et al. Acceptability of mental health apps for Aboriginal and Torres Strait Islander Australians: a qualitative study. J Med Internet Res 2016 Mar 11;18(3):e65. [doi: 10.2196/jmir.5314] [Medline: 26969043]
70. Garnweidner-Holme L, Hoel Andersen T, Sando MW, Noll J, Lukasse M. Health care professionals' attitudes toward, and experiences of using, a culture-sensitive smartphone app for women with gestational diabetes mellitus: qualitative study. JMIR Mhealth Uhealth 2018 May 14;6(5):e123. [doi: 10.2196/mhealth.9686] [Medline: 29759959]
71. Fry JP, Neff RA. Periodic prompts and reminders in health promotion and health behavior interventions: systematic review. J Med Internet Res 2009 May 14;11(2):e16. [doi: 10.2196/jmir.1138] [Medline: 19632970]
72. Russell E, Lloyd-Houldey A, Memon A, Yarker J. Factors influencing uptake and use of a new health information app for young people. J Technol Hum Serv 2018 Oct 2;36(4):222-240. [doi: 10.1080/15228835.2018.1536911]
73. Goodyear VA, Armour KM. Young people's perspectives on and experiences of health-related social media, apps, and wearable health devices. Soc Sci (Basel) 2018 Aug;7(8):137. [doi: 10.3390/socsci7080137]
74. Ridout B, Campbell A. The use of social networking sites in mental health interventions for young people: systematic review. J Med Internet Res 2018 Dec 18;20(12):e12244. [doi: 10.2196/12244] [Medline: 30563811]
75. Geng L, Jiang G, Yu L, et al. The most popular commercial weight management apps in the Chinese app store: analysis of quality, features, and behavior change techniques. JMIR Mhealth Uhealth 2023 Nov 24;11:e50226. [doi: 10.2196/50226] [Medline: 37999950]
76. Baretta D, Perski O, Steca P. Exploring users' experiences of the uptake and adoption of physical activity apps: longitudinal qualitative study. JMIR Mhealth Uhealth 2019 Feb 8;7(2):e11636. [doi: 10.2196/11636] [Medline: 30735143]
77. Ryan RM, Deci EL. Intrinsic and extrinsic motivations: classic definitions and new directions. Contemp Educ Psychol 2000 Jan;25(1):54-67. [doi: 10.1006/ceps.1999.1020] [Medline: 10620381]

XSL•FO
RenderX

78. Twenge JM, Campbell WK. Associations between screen time and lower psychological well-being among children and adolescents: Evidence from a population-based study. Prev Med Rep 2018 Dec;12:271-283. [doi: 10.1016/j.pmedr.2018.10.003] [Medline: 30406005]
79. Odgers CL, Jensen MR. Annual Research Review: Adolescent mental health in the digital age: facts, fears, and future directions. Child Psychology Psychiatry 2020 Mar;61(3):336-348. [doi: 10.1111/jcpp.13190]
80. George MJ, Odgers CL. Seven fears and the science of how mobile technologies may be influencing adolescents in the digital age. Perspect Psychol Sci 2015 Nov;10(6):832-851. [doi: 10.1177/1745691615596788] [Medline: 26581738]
81. Livingstone S, Ólafsson K, Helsper EJ, Lupiáñez-Villanueva F, Veltri GA, Folkvord F. Maximizing opportunities and minimizing risks for children online: the role of digital skills in emerging strategies of parental mediation. J Commun 2017 Feb;67(1):82-105. [doi: 10.1111/jcom.12277]

Abbreviations

BCW: Behavior Change Wheel
JU:MP: Join Us: Move. Play
mHealth: mobile health
YCS: Young Citizen Scientist

Edited by A Mavragani; submitted 25.04.25; peer-reviewed by C Baxter, G Wort; revised version received 30.06.25; accepted 03.07.25; published 11.11.25.

Please cite as:
Frazer MT, Charlesworth L, Wilson L, Hall J, Batool F, Subiza-Perez M, Burkhardt J, Daly-Smith A, Chalkley A
Active Play in a Digital Age, Exploring Children's (Aged 8-13 Years) Views of a Physical Activity App: Qualitative Formative Study
JMIR Form Res 2025;9:e76498
URL: https://formative.jmir.org/2025/1/e76498
doi:10.2196/76498

© Marie T Frazer, Lauren Charlesworth, Luca Wilson, Jennifer Hall, Farwa Batool, Mikel Subiza-Perez, Jan Burkhardt, Andy Daly-Smith, Anna Chalkley. Originally published in JMIR Formative Research (https://formative.jmir.org), 11.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Perceptions, Usage, and Educational Impact of ChatGPT Among Medical Students in Germany: Cross-Sectional Mixed Methods Survey

Anna Fußhöller[1,2]; Fabian Lechner[1,2]; Nadine Schlicker[2]; Felix Muehlensiepen[3,4]; Andreas Mayr[5]; Sebastian Kuhn[1], MD, MME; Martin Christian Hirsch[2]; Johannes Knitza[1], MD

[1]Institute for Digital Medicine, School of Medicine, Philipps-Universität Marburg, Baldingerstrasse 1, Marburg, Germany
[2]Institute for Artificial Intelligence in Medicine, School of Medicine, Philipps-Universität Marburg, Marburg, Germany
[3]Center for Health Services Research, Brandenburg Medical School Theodor Fontane, Rüdersdorf, Germany
[4]Department of Cardiology, Angiology and Intensive Care Medicine, Deutsches Herzzentrum der Charité, Berlin, Germany
[5]Institute for Medical Biometry and Statistics, Philipps-Universität Marburg, Marburg, Germany

Corresponding Author:
Johannes Knitza, MD
Institute for Digital Medicine, School of Medicine, Philipps-Universität Marburg, Baldingerstrasse 1, Marburg, Germany

Abstract

Background: Large language models such as ChatGPT offer significant opportunities for medical education. However, empirical data on actual usage patterns, perceived benefits, and limitations among medical students remain limited.

Objective: This study aimed to assess how medical students in Germany use ChatGPT, their perceptions of its educational value, and the challenges and concerns associated with its use.

Methods: A cross-sectional 17-item online survey was conducted between May and August 2024 among medical students from Philipps University Marburg, Germany. A mixed methods approach was applied, combining descriptive and inferential statistical analysis with qualitative content analysis of open-ended responses.

Results: A total of 84 fully completed surveys were included in the analysis (response rate: 26.7%; 315 surveys started). Overall, 76.2% (64/84) of the participants reported having used ChatGPT for medical education, with significantly higher usage during exam periods (P=.003). Preclinical students reported higher overall usage than clinical students (P=.02). ChatGPT was primarily used for summarizing information by 60.7% (51/84) of students, for literature research by 57.7% (49/84), and for clarifying concepts by 47.1% (40/84). A total of 70.2% (59/84) felt that it helped them save time, and 51.2% (43/84) reported an improved understanding of content. In contrast, only 31% (26/84) saw benefits for applying knowledge and 15.5% (13/84) for long-term knowledge retention. Qualitative responses highlighted clear benefits such as time savings and support in exam preparation, while also pointing to potential applications in clinical documentation and expressing concerns about misinformation and source transparency. However, 73.3% (55/75) expressed concerns about misinformation, and 72.6% (61/84) reported lacking confidence in their artificial intelligence (AI)–related skills. Only 41.7% (35/84) stated that they trust ChatGPT's outputs. Students who used the tool more frequently also reported higher levels of trust in ChatGPT's outputs (r=0.374, P<.001). Over 70% of respondents indicated a strong desire for increased integration of AI-related education and practical applications within the medical curriculum.

Conclusions: ChatGPT was already widely used among medical students, especially in exam preparation and the early stages of training. Students valued its efficiency and support for understanding complex material, but its long-term influence on learning is limited. Concerns about reliability, source transparency, and data privacy remain, and AI skills played a key role in shaping usage. These findings underscore the need to integrate structured, practice-oriented AI education into medical training to support critical, informed, and ethical use of large language models.

(JMIR Form Res 2025;9:e81484) doi:10.2196/81484

KEYWORDS

medical education; artificial intelligence; ChatGPT; large language model; digital health literacy; medical students; Germany; survey; self-directed learning

XSL•FO
RenderX

Introduction

The integration of large language models (LLMs) into higher education is rapidly transforming teaching and learning processes, particularly in the field of medicine [1-3]. These models are capable of generating human-like text, answering complex questions, and simulating interactive dialogue, which makes them uniquely suited for educational settings [4]. Among the most prominent artificial intelligence (AI) tools is ChatGPT [5]. Since its public release in November 2022, ChatGPT has become one of the most widely adopted AI tools worldwide, with millions of users, including a growing number of medical students [6,7].

LLMs offer medical students new opportunities for self-directed learning. These include simplifying dense academic texts, generating practice questions, simulating patient interactions, and providing personalized feedback. Early evidence suggests that these functionalities can support knowledge acquisition, comprehension, and problem-solving skills in medical education [8]. ChatGPT's ability to mimic personalized tutoring environments, generate differential diagnoses based on input data, and simulate realistic clinical scenarios presents a novel form of digital support for students' learning processes [1].

Nevertheless, ChatGPT has certain restrictions. ChatGPT can produce misleading or incorrect content while presenting it with convincing linguistic fluency, lacks source transparency, and does not account for evidence hierarchies in medical knowledge [9-11]. Privacy concerns are also relevant, as user inputs are processed on external servers, potentially compromising patient confidentiality when clinical case data are entered [12]. Furthermore, excessive reliance on AI tools may hinder the development of critical thinking skills [2,9] and potentially lead to suboptimal clinical decision-making [13].

Although interest in the educational potential of LLMs is increasing, evidence on their use and impact among medical students remains limited [14]. This study addressed this gap by examining how medical students in Germany use ChatGPT, their perceptions of its educational value, and the challenges and concerns associated with its use.

Methods

Study Design

The survey encompassed a total of 17 questions (Multimedia Appendix 1) and was reported in accordance with the CHERRIES (Checklist for Reporting Results of Internet E-Surveys) checklist [15]. The web-based survey was conducted via SoSciSurvey between May and August 2024. Medical students from University Marburg were recruited through student mailing lists, medical lectures at Philipps University Marburg, and by personal invitation from AF. The open survey was developed based on a review of related literature [16-18]. Adaptive questioning was applied to tailor the survey experience and reduce participant burden. Participants who indicated using other educational software or AI tools were asked to specify which ones. The following open-ended questions were included: "What concerns do you have when using ChatGPT?" "In which areas do you see potential in using ChatGPT?" "For which subjects or areas do you primarily use ChatGPT and why?" The survey was pretested for usability and technical functionality by 3 medical students who were not part of the final sample. Their feedback was used to refine the wording, layout, and navigation of the questionnaire before launching the final version via SoSciSurvey.

Ethical Considerations

The study was reviewed by the Ethics Committee of Philipps University Marburg, which granted an exemption waiver and raised no objections (reference number 24 - 80 ANZ). All participants received an information sheet before beginning the survey and provided electronic informed consent; participation was voluntary and could be discontinued at any time. Data were collected anonymously, with no personal identifiers recorded, and open-text responses were screened to ensure confidentiality. No financial or nonfinancial incentives were offered.

Data Analysis

A mixed methods approach was applied, combining quantitative and qualitative data through both closed and open-ended survey questions. This design allows for the identification of statistical patterns while also capturing insights into individual experiences.

Quantitative data were analyzed using jamovi (version 2.3.28.0; The Jamovi Project) and Microsoft Excel (version 16.89.1; Microsoft Corp) to compute descriptive statistics and explore relationships between variables. For group comparisons of ordinal or nonnormally distributed variables, the Mann-Whitney *U* test or the Welch *t* test (in cases of variance inhomogeneity) was applied. Paired comparisons between two related conditions (eg, semester vs exam period) were analyzed using the Wilcoxon signed-rank test. Correlations were assessed using Pearson correlation coefficient and Spearman rho, where appropriate. Differences in proportions were tested using binomial tests. Statistical significance was defined as $P<.05$.

Qualitative responses were analyzed using *MAXQDA* (VERBI Software). An inductive approach was applied to structure the open-ended answers. The data were systematically coded by one researcher who developed the main codes (eg, barriers, advantages) and subcodes (eg, time-saving or lack of trust), which were used to identify recurring themes and patterns. Based on these thematic categories, frequency tables were generated to quantify the distribution of responses and support interpretive conclusions. Representative quotations were then extracted to illustrate key points within each category.

Medical education in Germany is divided into 2 phases. First, there is a preclinical phase that lasts 2 years and focuses on foundational subjects and conveys more theoretical knowledge. After passing the "Physikum," students enter the clinical phase, which lasts 4 years and provides more practical and patient-centered training. An exploratory subgroup analysis was performed based on this distinction.

XSL•FO
RenderX

Results

Study Population

A total of 315 medical students from German universities participated in the survey, of whom 26.7% (84/315) provided fully completed surveys that were included in the final analysis. Among these, 69% (58/84) were female. The mean age of participants was 23.7 years (median 23; range 19 - 30), and the average semester of study was 5.5 (SD 2.2). Of the participants, 22 were enrolled in the preclinical phase of medical education and 62 in the clinical phase.

ChatGPT Usage

Of all participants, 76.2% (64/84) stated that they have used ChatGPT for medical education purposes. The highest proportion of students used it once per week(Multimedia Appendix 2). A Wilcoxon signed-rank test indicated that students used ChatGPT significantly more often during exam preparation periods than during regular semester periods (P=.003).

Seventy-five percent (63/84) of the students reported not using any other AI tools for studying besides ChatGPT. DeepL was the most frequently used other AI tool, reported by 38.9% (7/18) of students who used another tool. In addition, a Mann-Whitney U test showed that preclinical students (mean 2.82) reported significantly higher overall ChatGPT usage than clinical students (mean 2.20; P=.02). For example, a preclinical student stated, “In biochemistry and anatomy, it provides faster answers than looking things up in textbooks for general questions.” No significant differences in usage were found concerning age or gender. ChatGPT was used for various different academic tasks. We found that 60.7% (51/84) of respondents claimed to use ChatGPT for summarization, 57.7% (49/84) for research and literature study, and 47.6% (40/84) to clarify questions and explain subject-specific concepts (Multimedia Appendix 3).

ChatGPT was also reported to be used across various subjects within the medical curriculum. The most frequently mentioned area was general medical studies, cited by 37.5% (18/48) of students, followed by medical history and ethics with 27.1% (13/48), and biochemistry with 14.6% (7/48). Physiology was mentioned in 12.5% (6/48) of responses and anatomy and histology in 10.4% (5/48). Less frequently named subjects included pharmacology, mentioned by 8.3% (4/48) of students, biology, cited by 6.3% (3/48), and both chemistry and physics, with 4.2% (2/48) each.

Perceived Impact of ChatGPT

The majority (59/84, 70.2%) of students agreed that using ChatGPT helped them save time (Figure 1). One student stated, “It helps to condense complex content into short summaries that focus on the essentials, making it easier to get started. I often spend a long time searching textbooks for unclear terms, and ChatGPT can often provide helpful insights or even completely resolve my question.” Furthermore, 51.2% (43/84) reported improved understanding of the content, while 31% (26/84) felt that it enhanced their ability to apply what they had learned. Only 15.5% (13/84) agreed that ChatGPT improved their long-term information retention.

Figure 1. Perceived impact of ChatGPT on learning efficiency, understanding, application, and knowledge retention among medical students.

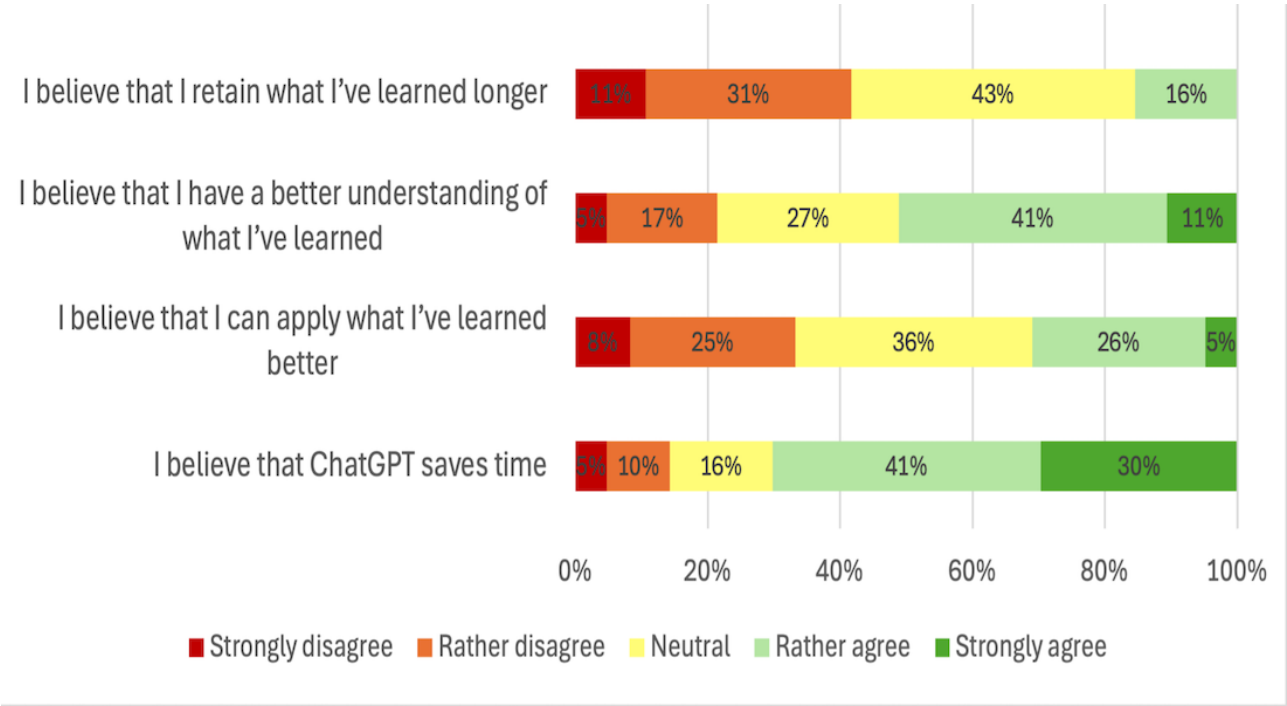

Perceived Benefits and Trust

In all, 38.1% (32/84) stated that ChatGPT had advantages over other learning platforms. The advantages listed (optional answer) were mostly more specific and individual benefits, reported by 50% (10/20) of respondents; time saving, mentioned by 35% (7/20); and more comprehensible explanations, cited by 15% (3/20). One student explained that “targeted questions can be answered within seconds—questions that might not be addressed in some books or would otherwise require extensive research.”

XSL•FO
RenderX

Furthermore, 41.7% (35/84) of the students agreed that they trust the results provided by ChatGPT. A significant positive correlation was found between trust and usage frequency (Pearson r=0.374, P<.001).

Potential and Barriers

Students were invited to suggest areas in which they could see potential for using ChatGPT in a free-text answer. The most frequently mentioned areas included explanation, named by 22.5% (20/89) of students; summarization, mentioned by 18% (16/89); support with everyday hospital tasks such as medical reports, cited by 15.7% (14/89); text processing, mentioned by 14.6% (13/89); time saving, reported by 12.4% (11/89); research, indicated by 11.2% (10/89); and use as a learning tool, stated by 5.6% (5/89).

In another open-ended question, students were asked to share their concerns about using ChatGPT. The majority, 73.3% (55/75), expressed worries about incorrect or inaccurate information, as exemplified by this quote: "I don't know how trustworthy the sources used are." In addition, 16% (12/75) of the students mentioned concerns about unreliable or nonexistent sources, 6.7% (5/75) claimed a potential decline in academic independence, and 4% (3/75) were concerned about the handling of personal data. One student stated the following fear: "That I rely too much on a system, which may mean I retain less myself."

AI Knowledge

Most students reported becoming aware of ChatGPT through personal technical interest (44/84, 52.4%) or media coverage (43/84, 51.2%). Recommendations from others were also a significant factor, reported by 45.2% (38/84). Additionally, 16.7% (14/84) reported discovering ChatGPT through independent internet research, while 13.1% (11/84) mentioned social media. Despite this widespread awareness, the majority of students indicated a lack of deeper understanding: 72.6% (61/84) of the students stated that they are not well-acquainted with AI and its possibilities (Multimedia Appendix 4).

Most of the students (60/84, 71.4%) would appreciate more courses in computer science in order to learn more about the possibilities of AI in medicine, and 74.3% (54/84) of the students would appreciate more practical exercises with AI. In addition, 61.9% (52/84) agreed that students more familiar with AI have a career advantage (Figure 2).

Figure 2. Students' perspectives on computer science and AI in their education and career. AI: artificial intelligence.

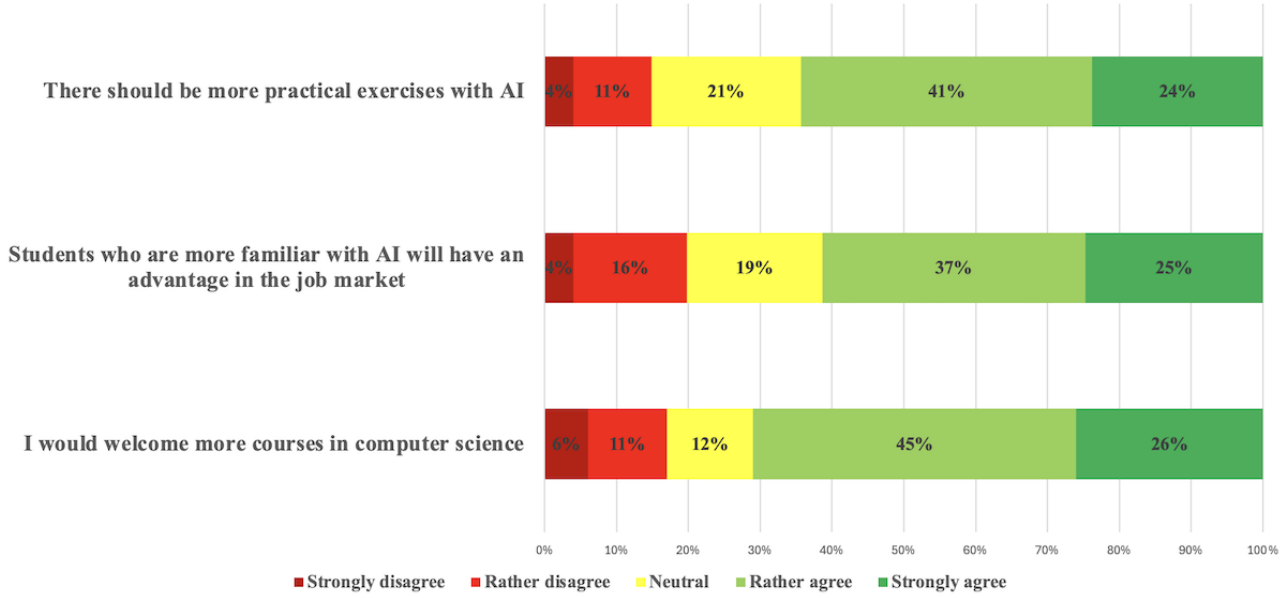

Discussion

Principal Findings

This study found that ChatGPT was widely used among German medical students, particularly during exam periods and in the preclinical phase. Students reported using it mainly for summarizing material, literature searches, and clarification of concepts. Although many students valued its efficiency and perceived improvements in comprehension, fewer saw benefits for application or long-term retention. Concerns about misinformation and source credibility were common, and AI literacy strongly influenced trust and frequency of use. The majority of medical students reported low confidence in their AI skills and expressed a strong interest in additional AI training. Students' perceptions of ChatGPT appear to reflect broader attitudes toward digital learning technologies in general, rather than being unique to LLMs and ChatGPT.

LLM Usage

The large proportion of medical students (76.2%) who had already used ChatGPT to support their studies aligns with recent international data: for instance, a 2025 survey conducted in Egypt reported a similarly high usage rate of 78.5% among students [19], and a 2024 survey from China reported a usage rate of 62.9% [20]. In contrast, an earlier US-based study from 2024 documented a considerably lower usage rate of only 48.9% [21], and a survey across Germany, Austria, and Switzerland, published in 2024, revealed a rate of 38.8% [16]. These differences are likely attributable to the varying time points of

XSL•FO
RenderX

data collection and suggest a growing trend in ChatGPT adoption among medical students over time.

To the best of our knowledge, this is the first study to demonstrate a significantly increased use of ChatGPT during examination periods of medical students. A Spanish study from 2024 among nursing students supports this observed trend. The majority of students considered ChatGPT to be useful or very useful for exam preparation, and 89% reported that it enhanced their academic performance to some extent [22]. These findings underscore the growing relevance of LLMs, particularly during high-stress periods such as exams, when students seek efficient and effective study support.

We found that preclinical students (first 2 years of medical school in Germany) reported significantly higher overall usage than clinical students. Similar results were observed in a 2024 US study, which reported significantly higher usage of ChatGPT among first-year medical students [23]. This pattern may reflect the stronger focus on theoretical content and self-directed learning in the early years of medical education, where tools like ChatGPT are particularly helpful for summarizing complex material or clarifying foundational concepts. In contrast, a study among undergraduate health care students in Malaysia from 2024 found opposite results. They state that fifth-year students were significantly more likely to use ChatGPT for academic purposes than first-year students [24]. This discrepancy could be due to differences in curricular structure, teaching methods, or digital tool integration. For instance, senior students in some contexts may face more complex clinical case studies and written assignments, prompting them to seek AI-based support for academic writing or clinical reasoning tasks.

Our findings indicate that students use ChatGPT in diverse ways. This corresponds in part with findings from a US-based study, where 59.6% of students used ChatGPT for proofreading and grammar correction. Notably, the majority of respondents in that study (89.9%) reported that ChatGPT enhanced their research productivity, which may reflect a broader trend toward the use of AI tools to support academic work [23].

Free-text responses suggested that students already envisioned potential future applications of ChatGPT in both clinical and academic settings. Most frequently, they emphasized its value in explaining medical content, summarizing information, and supporting routine clinical tasks such as drafting medical reports. These perceptions closely align with findings from a US study, where 22.5% of students who had completed clinical rotations reported using ChatGPT during their clerkship. Among these students, the tool was primarily used to gather information on diagnoses, pharmacology, and patient care planning, as well as to assist with writing clinical notes [23].

Perceived LLM Advantages and Barriers

Our findings showed that roughly one-third of the students considered ChatGPT to offer advantages over other learning platforms. A US study further reported that some students preferred using ChatGPT over consulting professors (45.3%), using textbooks (42.2%), or attending lectures (31.7%) [21]. The authors attribute this preference to ChatGPT's ability to deliver instant, personalized, and user-friendly responses, which aligns with our findings. A recent meta-analysis highlighted the advantages of ChatGPT use, such as improved academic performance and a reduction in mental effort [25]. However, a complementary review has raised concerns about the potential for "never-skilling," whereby overreliance on LLMs may hinder skill acquisition [2]. To address this challenge, the authors introduced the DEFT-AI (Diagnosis, Evidence, Feedback, and Teaching) framework, which emphasizes fostering critical thinking as a safeguard against such risks. A subsequent review concluded that the acquisition of theoretical knowledge was comparable between generative AI–based instruction and traditional teaching methods in medical education [26].

Our findings suggest that while ChatGPT is perceived as a valuable support tool, particularly for enhancing efficiency and comprehension, its influence on long-term knowledge retention appears to be limited. Marton and Säljö's concept of *approaches to learning* [27] offers further insight: if ChatGPT is used to engage with underlying concepts and to seek understanding, it may enhance comprehension. However, when used primarily to complete tasks quickly, as part of a surface approach, its educational benefit is likely to be minimal. The combination of high reported time-saving and low reported gains in knowledge application and retention in our data suggests that both usage patterns coexist. This highlights the need to promote critical engagement with AI tools and to provide educational frameworks that encourage responsible and reflective use.

When evaluating students' trust in ChatGPT, the results revealed a rather cautious stance. This finding is consistent with previous studies, which reported that approximately half of the students trusted ChatGPT's responses [19,23]. Interestingly, in our study, there was a significant positive correlation between trust and usage frequency. In the study of Malaysian health care students, higher knowledge and positive attitudes correlated with increased academic usage [24]. Taken together, these findings suggest that students' engagement with AI tools is not solely driven by external factors such as availability or technical ease of use, but also by internal factors like trust, familiarity, and perceived value. This highlights the importance of building digital literacy and user confidence to foster responsible and effective use.

In this study, students also voiced critical concerns regarding the use of ChatGPT. These concerns reflect broader criticisms discussed in the literature, particularly regarding the risk of misinformation, facilitated plagiarism [20], lack of source transparency, and the impact of AI tools on core academic competencies [24,28,29].

AI Knowledge and Education

Most medical students lacked confidence in their AI skills and expressed a desire for more AI training. These results emphasize the need to integrate structured and practice-oriented AI education into the medical curriculum and mirror results from previous studies [28-31]. A recently evaluated AI module for German medical students received a 100% peer recommendation rate, reflecting high levels of acceptance and perceived value [32]. Furthermore, a blended-learning course by Thomae et al [31] integrated ChatGPT into core teaching units and emphasized student reflection and AI output evaluation,

XSL•FO
RenderX

improving both engagement and perceived relevance. A previous survey also highlighted that hands-on AI experience was significantly linked to more positive perceptions of its role in improving patient care [17]. To build on these early models, future research should prioritize intervention studies that assess learning outcomes. A significant barrier to the implementation of AI education may be the limited digital competence among educators. The Check-In tool, consisting of 22 items derived from the European Framework for the Digital Competence of Educators (DigCompEdu), provides a structured approach to assessing current competencies and systematically identifying areas for improvement [3].

Limitations

This study is subject to several limitations. Our findings should be viewed as a snapshot of early ChatGPT adoption, which may not fully capture current practices due to rapid AI developments. Furthermore, data collection was based on a voluntary online questionnaire. This may have introduced a self-selection bias, as students with an interest in technology may have been more likely to participate. Additionally, reliance on self-reported measures may introduce inaccuracies due to factors such as social desirability and recall bias. As a result, the reported ChatGPT usage in this study could be overestimated compared to participants' actual average usage. Although 315 students accessed the survey, only 84 provided complete responses, which reflects the generally low response rates commonly observed in medical education online surveys [33]. This relatively low completion rate may be attributed to several factors, including the voluntary nature of participation, the overall length of the questionnaire, and the inclusion of open-ended questions, which may have required more time and cognitive effort. A further limitation is that the analysis of open-ended responses was conducted by a single researcher, which may have introduced subjectivity and limited intercoder reliability. Although this approach ensured consistency in coding, future studies would benefit from involving multiple coders and formal reliability checks to strengthen the robustness of qualitative findings.

Conclusions

ChatGPT was widely used by medical students in Germany, particularly during exam preparation and in the preclinical phase of training. Students primarily valued the tool for its efficiency, summarization capabilities, and ability to support understanding of complex material. Key concerns included misinformation, lack of source transparency, and data privacy risks, all of which contribute to variable levels of student trust. Notably, the strong demand for AI-related training underscores the urgency of integrating structured, practice-oriented AI education into medical curricula. As LLMs become increasingly prevalent in health care and education, promoting critical, informed, and ethically responsible use among future clinicians is essential.

Acknowledgments

We thank all students for their valuable time and support. The authors used DeepL and ChatGPT to assist with grammar, spelling, and overall readability while preparing this manuscript. Following the use of these tools, they carefully reviewed and revised the text as necessary and assume full responsibility for the final content of the publication.

Funding

Open access funding was provided by the Open Access Publishing Fund of Philipps-Universität Marburg.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: AF, FL, JK; data curation: AF, AM, JK; formal analysis: AF, AM, JK, NS; funding acquisition: JK; investigation: AF; methodology: AF, FM, NS, AM; project administration: AF, JK, FL; resources: MCH, SK; software: AF, MCH; supervision: JK, AM, FL, MCH, SK; validation: FL, JK, AM; visualization: AF; writing – original draft: AF; writing – review and editing: AF, JK, SK.

Conflicts of Interest

SK is founder and shareholder of MED.digital. MCH is shareholder of Ada Health.

Multimedia Appendix 1
Original questionnaire and English translation.
[PDF File, 179 KB - formative_v9i1e81484_app1.pdf]

Multimedia Appendix 2
ChatGPT usage among medical students during the semester and exam period.
[PNG File, 74 KB - formative_v9i1e81484_app2.png]

XSL•FO
RenderX

Multimedia Appendix 3
Frequency of ChatGPT usage across different academic tasks among medical students.
[PNG File, 104 KB - formative_v9i1e81484_app3.png]

Multimedia Appendix 4
Proportions of students indicating how acquainted they are with artificial intelligence and its possibilities.
[PNG File, 58 KB - formative_v9i1e81484_app4.png]

References

1. Narayanan S, Ramakrishnan R, Durairaj E, Das A. Artificial intelligence revolutionizing the field of medical education. Cureus 2023 Nov;15(11):e49604. [doi: 10.7759/cureus.49604] [Medline: 38161821]
2. Abdulnour REE, Gin B, Boscardin CK. Educational strategies for clinical supervision of artificial intelligence use. N Engl J Med 2025 Aug 21;393(8):786-797. [doi: 10.1056/NEJMra2503232] [Medline: 40834302]
3. Inamorato Dos Santos A, Chinkes E, Carvalho MAG, Solórzano CMV, Marroni LS. The digital competence of academics in higher education: is the glass half empty or half full? Int J Educ Technol High Educ 2023;20(1):9. [doi: 10.1186/s41239-022-00376-0] [Medline: 36778715]
4. Moor M, Banerjee O, Abad ZSH, et al. Foundation models for generalist medical artificial intelligence. Nature New Biol 2023 Apr 13;616(7956):259-265. [doi: 10.1038/s41586-023-05881-4]
5. Garg RK, Urs VL, Agarwal AA, Chaudhary SK, Paliwal V, Kar SK. Exploring the role of ChatGPT in patient care (diagnosis and treatment) and medical research: a systematic review. Health Promot Perspect 2023;13(3):183-191. [doi: 10.34172/hpp.2023.22] [Medline: 37808939]
6. Fauzi F, Tuhuteru L, Sampe F, Ausat AMA, Hatta HR. Analysing the role of ChatGPT in improving student productivity in higher education. joe 2023 Jun;5(4):14886-14891. [doi: 10.31004/joe.v5i4.2563]
7. Krettek C. ChatGPT: milestone text AI with game changing potential. Unfallchirurgie (Heidelb) 2023 Mar;126(3):252-254. [doi: 10.1007/s00113-023-01296-y] [Medline: 36763148]
8. Wu Y, Zheng Y, Feng B, Yang Y, Kang K, Zhao A. Embracing ChatGPT for medical education: exploring its impact on doctors and medical students. JMIR Med Educ 2024 Apr 10;10:e52483. [doi: 10.2196/52483] [Medline: 38598263]
9. Wong RSY, Ming LC, Raja Ali RA. The intersection of ChatGPT, clinical medicine, and medical education. JMIR Med Educ 2023 Nov 21;9(1):e47274. [doi: 10.2196/47274] [Medline: 37988149]
10. Emsley R. ChatGPT: these are not hallucinations – they're fabrications and falsifications. Schizophr 2023 Aug 19;9(1):1-2. [doi: 10.1038/s41537-023-00379-4]
11. Jeyaraman M, Ramasubramanian S, Balaji S, Jeyaraman N, Nallakumarasamy A, Sharma S. ChatGPT in action: harnessing artificial intelligence potential and addressing ethical challenges in medicine, education, and scientific research. World J Methodol 2023 Sep 20;13(4):170-178. [doi: 10.5662/wjm.v13.i4.170] [Medline: 37771867]
12. Li J. Security implications of AI chatbots in health care. J Med Internet Res 2023 Nov 28;25:e47551. [doi: 10.2196/47551] [Medline: 38015597]
13. Knitza J, Tascilar K, Gruber E, et al. Accuracy and usability of a diagnostic decision support system in the diagnosis of three representative rheumatic diseases: a randomized controlled trial among medical students. Arthritis Res Ther 2021 Sep 6;23(1):233. [doi: 10.1186/s13075-021-02616-6] [Medline: 34488887]
14. Cross J, Kayalackakom T, Robinson RE, et al. Assessing ChatGPT's capability as a new age standardized patient: qualitative study. JMIR Med Educ 2025 May 20;11(1):e63353. [doi: 10.2196/63353] [Medline: 40393017]
15. Eysenbach G. Improving the quality of web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34. [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
16. Weidener L, Fischer M. Artificial intelligence in medicine: cross-sectional study among medical students on application, education, and ethical aspects. JMIR Med Educ 2024 Jan 5;10(1):e51247. [doi: 10.2196/51247] [Medline: 38180787]
17. Alkhaaldi SMI, Kassab CH, Dimassi Z, et al. Medical student experiences and perceptions of ChatGPT and artificial intelligence: cross-sectional study. JMIR Med Educ 2023 Dec 22;9(1):e51302. [doi: 10.2196/51302] [Medline: 38133911]
18. Tangadulrat P, Sono S, Tangtrakulwanich B. Using ChatGPT for clinical practice and medical education: cross-sectional survey of medical students' and physicians' perceptions. JMIR Med Educ 2023 Dec 22;9(1):e50658. [doi: 10.2196/50658] [Medline: 38133908]
19. Abdelhafiz AS, Farghly MI, Sultan EA, Abouelmagd ME, Ashmawy Y, Elsebaie EH. Medical students and ChatGPT: analyzing attitudes, practices, and academic perceptions. BMC Med Educ 2025 Feb 5;25(1):187. [doi: 10.1186/s12909-025-06731-9] [Medline: 39910593]
20. Hu N, Jiang XQ, Wang YD, et al. Status and perceptions of ChatGPT utilization among medical students: a survey-based study. BMC Med Educ 2025 Jun 4;25(1):831. [doi: 10.1186/s12909-025-07438-7] [Medline: 40468340]
21. Zhang JS, Yoon C, Williams DKA, Pinkas A. Exploring the usage of ChatGPT among medical students in the United States. J Med Educ Curric Dev 2024;11:23821205241264695. [doi: 10.1177/23821205241264695] [Medline: 39092290]

XSL•FO
RenderX

22. Gonzalez-Garcia A, Bermejo-Martinez D, Lopez-Alonso AI, Trevisson-Redondo B, Martín-Vázquez C, Perez-Gonzalez S. Impact of ChatGPT usage on nursing students education: a cross-sectional study. Heliyon 2025 Jan 15;11(1):e41559. [doi: 10.1016/j.heliyon.2024.e41559] [Medline: 39850430]
23. Ganjavi C, Eppler M, O'Brien D, et al. ChatGPT and large language models (LLMs) awareness and use. A prospective cross-sectional survey of U.S. medical students. PLOS Digit Health 2024 Sep;3(9):e0000596. [doi: 10.1371/journal.pdig.0000596] [Medline: 39236008]
24. George Pallivathukal R, Kyaw Soe HH, Donald PM, Samson RS, Hj Ismail AR. ChatGPT for academic purposes: survey among undergraduate healthcare students in Malaysia. Cureus 2024 Jan;16(1):e53032. [doi: 10.7759/cureus.53032] [Medline: 38410331]
25. Deng R, Jiang M, Yu X, Lu Y, Liu S. Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. Comput Educ 2025 Apr;227:105224. [doi: 10.1016/j.compedu.2024.105224]
26. Li J, Yin K, Wang Y, Jiang X, Chen D. Effectiveness of generative artificial intelligence-based teaching versus traditional teaching methods in medical education: a meta-analysis of randomized controlled trials. BMC Med Educ 2025 Aug 19;25(1):1175. [doi: 10.1186/s12909-025-07750-2] [Medline: 40830476]
27. Marton F, Säljö R. On qualitative differences in learning: I—outcome and process*. Brit J of Edu Psychol 1976 Feb;46(1):4-11. [doi: 10.1111/j.2044-8279.1976.tb02980.x]
28. Schei OM, Møgelvang A, Ludvigsen K. Perceptions and use of AI chatbots among students in higher education: a scoping review of empirical studies. Education Sciences 2024 Aug 22;14(8):922. [doi: 10.3390/educsci14080922]
29. Elhassan SE, Sajid MR, Syed AM, Fathima SA, Khan BS, Tamim H. Assessing familiarity, usage patterns, and attitudes of medical students toward ChatGPT and other chat-based AI apps in medical education: cross-sectional questionnaire study. JMIR Med Educ 2025 Jan 30;11(1):e63065. [doi: 10.2196/63065] [Medline: 39883912]
30. Jebreen K, Radwan E, Kammoun-Rebai W, et al. Perceptions of undergraduate medical students on artificial intelligence in medicine: mixed-methods survey study from Palestine. BMC Med Educ 2024 May 7;24(1):507. [doi: 10.1186/s12909-024-05465-4] [Medline: 38714993]
31. Thomae AV, Witt CM, Barth J. Integration of ChatGPT into a course for medical students: explorative study on teaching scenarios, students' perception, and applications. JMIR Med Educ 2024 Aug 22;10:e50545. [doi: 10.2196/50545] [Medline: 39177012]
32. Oftring ZS, Deutsch K, Tolks D, Jungmann F, Kuhn S. Novel blended learning on artificial intelligence for medical students: qualitative interview study. JMIR Med Educ 2025 May 26;11(1):e65220. [doi: 10.2196/65220] [Medline: 40418795]
33. Wu MJ, Zhao K, Fils-Aime F. Response rates of online surveys in published research: a meta-analysis. Computers in Human Behavior Reports 2022 Aug;7:100206. [doi: 10.1016/j.chbr.2022.100206]

Abbreviations

AI: artificial intelligence
CHERRIES: Checklist for Reporting Results of Internet E-Surveys
LLM: large language model

Edited by A Stone; submitted 29.07.25; peer-reviewed by D Chen, O Oyetunji, X Liang; accepted 22.10.25; published 11.11.25.

Please cite as:
Fußhöller A, Lechner F, Schlicker N, Muehlensiepen F, Mayr A, Kuhn S, Hirsch MC, Knitza J
Perceptions, Usage, and Educational Impact of ChatGPT Among Medical Students in Germany: Cross-Sectional Mixed Methods Survey
JMIR Form Res 2025;9:e81484
URL: https://formative.jmir.org/2025/1/e81484
doi:10.2196/81484

© Anna Fußhöller, Fabian Lechner, Nadine Schlicker, Felix Muehlensiepen, Andreas Mayr, Sebastian Kuhn, Martin Christian Hirsch, Johannes Knitza. Originally published in JMIR Formative Research (https://formative.jmir.org), 11.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Dementia Caregivers' Perspectives on Technology's Place in Care Practices: Mixed Methods Survey

Julia A Scott[1], PhD; Emma Cepukenas[2], BA; McKenzie Himes[3], BS; Kennedy Anderson[4], BA; Kiren Grewal[4], BA; An Mai[4], BA; Sheila Yuter[4], MSPH, EdD; Patricia Simone[5], PhD

[1]Department of Bioengineering, Santa Clara University, 500 El Camino Real, Santa Clara, CA, United States
[2]Department of Public Health Sciences, Santa Clara University, Santa Clara, CA, United States
[3]Department of Biology, Santa Clara University, Santa Clara, CA, United States
[4]Department of Public Health Sciences, Santa Clara University, Santa Clara, CA, United States
[5]Department of Psychology, Santa Clara University, Santa Clara, CA, United States

Corresponding Author:
Julia A Scott, PhD
Department of Bioengineering, Santa Clara University, 500 El Camino Real, Santa Clara, CA, United States

Abstract

Background: As the number of individuals diagnosed with dementia continues to grow, collective caregiving capacity is simultaneously declining. As a result, caregivers often face significant stress and burnout, which negatively impacts both their well-being and the quality of care they are able to provide. To address these challenges, various forms of technology have been developed to support caregiving responsibilities for people with dementia. Barriers to widespread adoption include the cost of technology, lack of caregiver training, limited awareness of available solutions, privacy and ethical concerns, and resistance to change from both caregivers and care recipients. By identifying these key insights, this study seeks to inform the development of more accessible, user-friendly, and effective technological solutions that can better support caregivers in their vital roles.

Objective: This study aimed to assess the way dementia caregivers solve common challenges, focusing on technology adoption and barriers to its use.

Methods: We conducted online surveys regarding technology usage and caregiving experience, which were distributed to the communities of caregivers of people living with dementia. Surveys quantified categories of technology use and caregiver needs. The survey also gathered open responses for joys, struggles, and technology use. Follow-up interviews with caregivers discussed how specific challenges were met. The responses were thematically coded to characterize findings by the type—struggles, social supports, technology, and joy.

Results: Respondents (N=69) identified access to resources as the most important unmet need for caregivers (53%). The survey found that medication and routine tracking applications and mobility assistive devices were the most commonly used forms of technology. The dominant need expressed by caregivers was for better access to resources, with improved access to health care services and financial support. Interviewees reported barriers to accessing resources, including high costs and difficulties in selecting appropriate technology resources. Additionally, centralizing resources and simplifying the process of locating support was recommended as a solution.

Conclusions: The study showed that caregivers felt inadequately supported and did not frequently use technology due to barriers of access and usability of products. Strategies in product design and user communication that engage dementia care partners would likely accelerate adoption.

(JMIR Form Res 2025;9:e69596) doi:10.2196/69596

KEYWORDS

dementia; caregiving; technology; AgeTech; assistive technology; accessibility

Introduction

Caring for individuals with dementia is demanding and requires strong, effective systems that support the safety, well-being, and quality of life of both the person living with dementia (PLWD) and their caregivers. Most caregivers are informal, typically spouses, family members, or friends who take on the role without formal training or financial compensation [1,2]. A significant number of informal caregivers experience mortality at a higher rate than the person they are caring for, highlighting the immense strain and pressure associated with this role [3], and they often face a decline in their own health, frequently stemming from chronic stress, sleep disturbances, and the emotional toll of providing constant care and support [4].

XSL•FO
RenderX

Despite the well-documented negative effects on informal caregivers, these challenges continue to persist due to the inequitable distribution of social and health care support services [5,6]. Notably, mental health concerns represent the most frequently reported unmet need among caregivers, emphasizing the urgent requirement for interventions targeting psychological well-being [7]. Caregivers frequently express feelings of isolation and insufficient support from family, health care professionals, and employers, further underscoring the systemic gaps in caregiver assistance [8].

The burden faced by caregivers highlights the importance of targeted support. Studies have explored the unmet needs of informal caregivers using an online survey across various domains, including health information, self-care, and support service accessibility [9]. Their findings revealed a critical need for targeted support systems that focus on caregivers' self-care, stress management, and communication skills. The study found that digital platforms could play a role in addressing these needs, given the potential for broad dissemination and accessibility. Such platforms have the potential to provide caregivers with on-demand access to educational resources, peer support networks, and mental health services, which could significantly alleviate caregiver burden.

Advancements in consumer technology have significantly transformed dementia care, providing essential support for both individuals with dementia and their caregivers. Wearable devices [10,11], interactive robots, mobile apps [12], and smart home systems are all emerging technologies that address the unique challenges faced in dementia care. There is also a web of offerings available to dementia caregivers. While these technologies offer promising support, challenges such as usability, cost, efficacy, and privacy concerns remain prevalent [13-15]. These are barriers to implementation in direct-to-consumer markets and health care systems. The reach of assistive technologies for caregivers is shallow because the system of care is not equipped to integrate the tools [16]. Addressing these issues is crucial for maximizing the benefits of technology in dementia care [13,14].

Projections for the technology used by older adults are growing. American Association for Retired Persons estimated that older adults would spend over US$120 billion on technology in 2023, reflecting both the growing digital engagement of this population and the increasing relevance of technology in health and social care contexts [17]. Concurrently, the global AgeTech market, which encompasses technological innovations designed to support aging populations, is valued at approximately US$45 trillion [17]. Within this expansive market, technology tailored for dementia care represents a rapidly expanding segment. Despite this growth and the global recognition that caregivers need support [2], widespread availability of dementia-related technologies is still emerging. Many technology services on the market are not yet scaled and are wholly in the direct-to-consumer market. The barriers previously stated have not yet been fully addressed.

Given these constraints, there is a critical need to align caregiver needs with appropriate technological interventions. A targeted needs assessment that considers the daily challenges faced by caregivers and evaluates the suitability and effectiveness of available digital tools is essential. When effectively implemented, such technologies hold the potential to reduce overall costs of care, facilitate aging in place, and alleviate caregiver burden, both in terms of workload and psychological stress.

One review reported on 31 studies published between 2007 and 2018 that address the needs of caregivers of PLWD; only a small number of these involved in-person interviews with caregivers, and many of these offered little insight on caregiver support [8]. Much of the existing data on caregiver needs comes from categorical surveys, which offer limited insight into why certain supports, like technology, are underused. However, a recent qualitative study with 12 informal dementia caregivers specifically asked caregivers about their use of technology [18]. They found that caregivers were open to using technology for practical tasks like scheduling, reminders, and information gathering, as well as video calls that support communication. However, they resisted using technology for emotionally meaningful tasks, such as bathing, cooking, or deep conversations, seeing these as essential to maintaining personal connection and dignity. The study emphasizes that technology should support, not replace, this "invisible emotion work." Tools that aid reflection, evoke shared memories, or reinforce coping strategies (like exercise or peer support) can enhance caregiving without diminishing its human core.

The present study sought to assess the experiences of informal caregivers who provide care for PLWD. In contrast to a deficit-focused approach, this study also aimed to capture positive aspects of caregiving and the role of technology in both caregiving responsibilities and caregiver self-care. Data collection involved a mixed-methods approach. An anonymous survey collected quantitative and qualitative data on caregiving challenges, technology usage, and perceived needs. In addition to structured items, open-ended responses allowed caregivers to elaborate on their experiences. To further explore specific caregiving dilemmas and the strategies used to address them, in-depth follow-up interviews were conducted with a subset of participants.

This caregiver perspective study applies a balanced quantitative and qualitative approach to explore the lived experiences of dementia caregivers and their use of technology. By capturing rich, contextual narratives and a history of technology practices, the research will offer insights to guide the design of user-centered technologies that support both the functional and emotional needs of caregivers.

Methods

Ethical Considerations

The purpose of this study was to collect first-hand reports on experiences of dementia caregivers. The study was approved by the institutional review board for human subject research of the Santa Clara University. Informed consent was provided via a digital copy, which described the study purpose of identifying unmet needs of caregivers and potential discomfort in remembering aspects of caregiving. Participants were not

XSL•FO
RenderX

compensated for completing the surveys or interviews. The study gathered data in two forms: (1) written surveys and (2) didactic interviews. The written surveys did not include identifying information and were coded by study ID. The interview transcripts were redacted for identifying information prior to analysis.

Participants

The study included any person with caregiving experience (current or within 3 years) for Alzheimer's and related dementias. Forwardable email invitations with links to anonymous surveys were distributed to the faculty, staff, and graduate students of Santa Clara University. The invitation was also distributed by the University of Washington Memory Hub in Seattle, Washington, and the Hearts and Minds Activity Center in San Jose, California. Within the survey, the option to participate in a video conference interview was included. Interview invitations were sent to known care professionals in the communities as well. The data was collected in April 2022.

Surveys

An anonymous online survey (Multimedia Appendix 1) was developed by the research team to contribute to a formal needs assessment in the first step of a design project for dementia care innovation. Participants were asked demographic questions, about their caregiving history, to list their primary needs, support systems, and technology used. Next, they were prompted to describe joyful and difficult aspects of their caregiving experiences and technology needs.

Interviews

Survey respondents (informal caregivers) and dementia care professionals (formal caregivers) were invited for interviews (Multimedia Appendix 1). Interview participants were asked about how they approached specific problems or challenges in caregiving practices. The structured interview followed a customer discovery method (Customer Development Labs), which probes how individuals approach meeting a specific need with nonleading questions and with an interactive cycle of questions, a method that is widely used for the exploration of needs assessment in user research. The purpose of this approach was to gather data for the subsequent phase of the project that aimed to develop new technologies for dementia care.

Analysis

Survey responses were tabulated and segmented by key demographic variables--age of caregiver, role of caregiver, gender of caregiver, and years of caregiving experience.

The interview transcripts and survey narrative responses were analyzed by thematic coding, a qualitative research method used to identify, analyze, and report patterns (themes) within qualitative data [19,20]. Each entry was coded with up to three annotations for each item, organized within the themes of struggles, social supports, technology, and joy (Textbox 1). All responses were annotated by two raters. Any conflict in labeling selections was resolved by consensus via discussion between raters and the principal investigator.

Textbox 1. Key topics annotated for thematic coding

Struggles

Finances, Health care, Long-term care, Family responsibilities, Work performance, Emotional struggle/conflict, Personal health, Personal care of patient, Communication with patient, Daily routines, Isolation, Language/culture, Access to resources, Disorientation/recall, Safety

Social supports

Family, Eldercare staff, Support groups, Advocacy organizations, Friends/colleagues, Therapist/counselor, Social workers, No support

Technology

Trackers, Telehealth, Adaptive clothing, Mobility devices, Monitors/sensors, Media, Games, Smartphone apps, No tech helped

Joys

Remembrance, Family/social events, Spectator events/shows, Music/dance, Art, Travel, Rare expressions/speech, Expressions of gratitude/connection, Religious experiences, Being with children, Socializing with peers, Being in nature

Mixed methodology was applied by contingency tables. These tables were built for the narrative annotations by age of the caregiver, gender of the caregiver, age of the person affected with dementia, and relationship to the person affected with dementia, which provided context to the types of challenges and solutions expressed by the participants. Quotes were extracted and grouped from the responses to capture feelings and anecdotes not illustrated in the coded analysis. Collectively, these analyses synthesized the responses in a manner to infer the unmet needs of caregivers. This last step, which is not statistical, was conducted through consensus discussions between the authors.

Results

Survey Sample Demographic

The survey of informal caregivers (n=69) represented a range of caregiver roles (Table 1) . Nearly three-quarters of respondents were female and a similar proportion identified as White. Most of the caregivers surveyed were 55 - 74 years of age; about 30% were younger and 9% were older. A quarter of caregivers took care of their spouses, and half cared for parents. All but one of the respondents cared for a person with dementia older than 50 years. Half of the respondents were currently caring for someone with dementia; of these, half had been in the role for 1 to 3 years, while the remainder had longer

XSL•FO
RenderX

caregiving experiences. Twenty-three percent of the respondents have cared for more than one person with dementia.

Table . Demographic characteristics of the survey caregiver respondents (N=69).

Demographic characteristic	n (%)
Caregiver sex	
Male	50 (72.46)
Female	16 (23.19)
Non-binary	1 (1.45)
Caregiver race/ethnicity	
Asian	6 (8.70)
Black or African American	2 (2.90)
Hispanic (Non-White)	4 (5.80)
Middle Eastern	1 (1.45)
Samoan	1 (1.45)
White	51 (73.91)
White, Asian	1 (1.45)
White, Hispanic (Non-White)	2 (2.90)
White, Mexican-American (White)	1 (1.45)
Caregiver age (years)	
18 - 25	2 (2.90)
26 - 35	5 (7.25)
36 - 55	15 (21.74)
55 - 75	41 (59.42)
75 - 95	6 (8.70)
PLWD[a] age (years)	
<50	1 (1.45)
>75	38 (55.07)
50 - 75	30 (43.48)

[a]PLWD: person living with dementia.

Informal Caregiver Needs Assessment

Participants listed the top three struggles in their caregiving role (Figure 1), offering a clear picture of the multidimensional challenges faced by informal caregivers of PLWD. Emotional struggle and interpersonal conflict emerged as the most frequently cited challenge, identified by 64% of respondents. Many caregivers described the emotional toll of witnessing cognitive decline and behavioral changes in their loved ones, often leading to frustration, grief, or feelings of helplessness. A resonating scenario of these challenges was shared here:

Figure 1. Ranked number of survey responses for caregiver struggles.

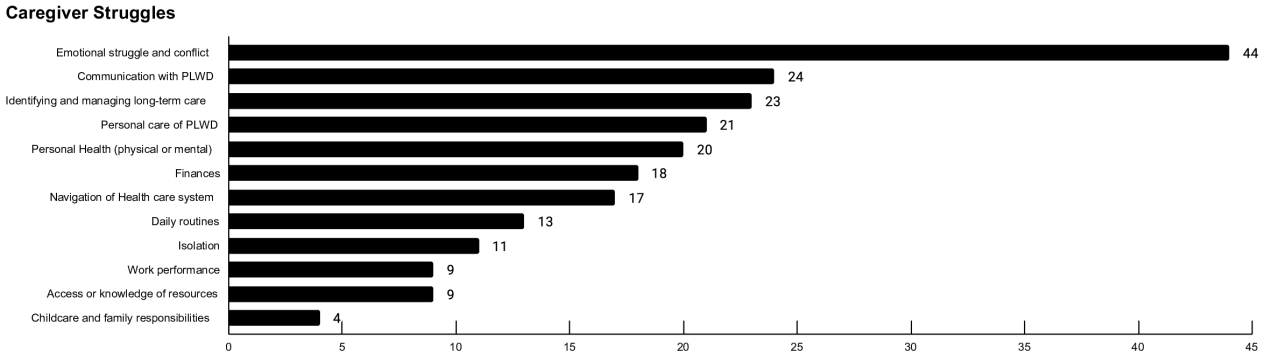

XSL•FO
RenderX

> *My mother would sometimes walk out the front door and start walking to the neighborhood park on the next block. If it was too cold, I would struggle trying to get her back in the house. She was very stubborn and you can't make a person with dementia do something she doesn't want to do. They act like children many times and do not listen. Sometimes she would just sit on the sidewalk so I could not walk her back. If I tried to pick her up, she would scream for the police yelling that I was hurting her! It would be so embarrassing!*

In addition, 29% of participants reported that caregiving had negatively affected their physical or mental health. Caregivers frequently mentioned sleep disturbances, chronic stress, and isolation as ongoing struggles. A spouse caregiver illustrated this well: "I struggled with feeling trapped (eg, couldn't get to a pool workout to take care of myself) because I couldn't leave my husband alone."

Care management issues also figured prominently in caregivers' responses, with long-term care planning (33%), financial concerns (26%), and difficulties navigating the health care system (25%) cited as key stressors. These challenges reflect broader structural barriers to coordinated dementia care and indicate opportunities for targeted policy or programmatic interventions. Communication difficulties (35%) and assistance with personal care tasks (30%) were also frequently mentioned, reflecting the hands-on demands of caregiving and the evolving needs of PLWD as the disease progresses.

Support for caregivers was assessed across social, direct care, and technological domains (Figure 2). Family members were cited most often as sources of social support, with 42% identifying them as primary helpers. Support groups (25%) and paid care staff (16%) also played meaningful roles. Despite these reported connections, many caregivers noted the uneven or inconsistent nature of their support networks, which often left them feeling like they were "on call" around the clock. This further emphasizes the emotional isolation that can accompany informal caregiving. One adult child recounted how her parents would argue:

> *It was mentally draining because no matter how much I could deal with my step-father, my mother's reactions would exacerbate the situation so it felt like I was dealing with two patients and not one. And the fact that my patience was beyond what I could possibly handle, left me feeling drained, frustrated and unempathetic toward them.*

When it came to direct care, 64% of respondents received help from family members (other than a spouse), 39% shared caregiving duties with home health aides, and 29% received support from health care providers. While some described these supports as vital lifelines, others expressed frustration with the unpredictability or limited scope of external assistance.

Technology use among caregivers was relatively modest and task-specific. The most common tools were medication and daily tracking apps (28%), used to manage schedules and routines. Over 15% of participants reported using tools that offered practical, hands-on support such as mobility aids, telehealth platforms, or adaptive clothing designed for people with physical disabilities or limited mobility, featuring easy closures, adjustable elements, and comfortable designs for independence. However, more advanced or integrated technologies, such as home monitors, sensors, or caregiving-specific platforms, were used by only 13% of respondents. Notably, none of the participants mentioned using technology for emotional or relational support.

Overall, these findings underscore the complexity of the caregiving experience and highlight important opportunities for designing tools and services that better reflect caregivers' lived realities, particularly by recognizing and supporting the emotional labor involved in their roles.

XSL•FO
RenderX

Figure 2. Ranked number of survey responses for (A) social supports, (B) assistance, and (C) technology use.

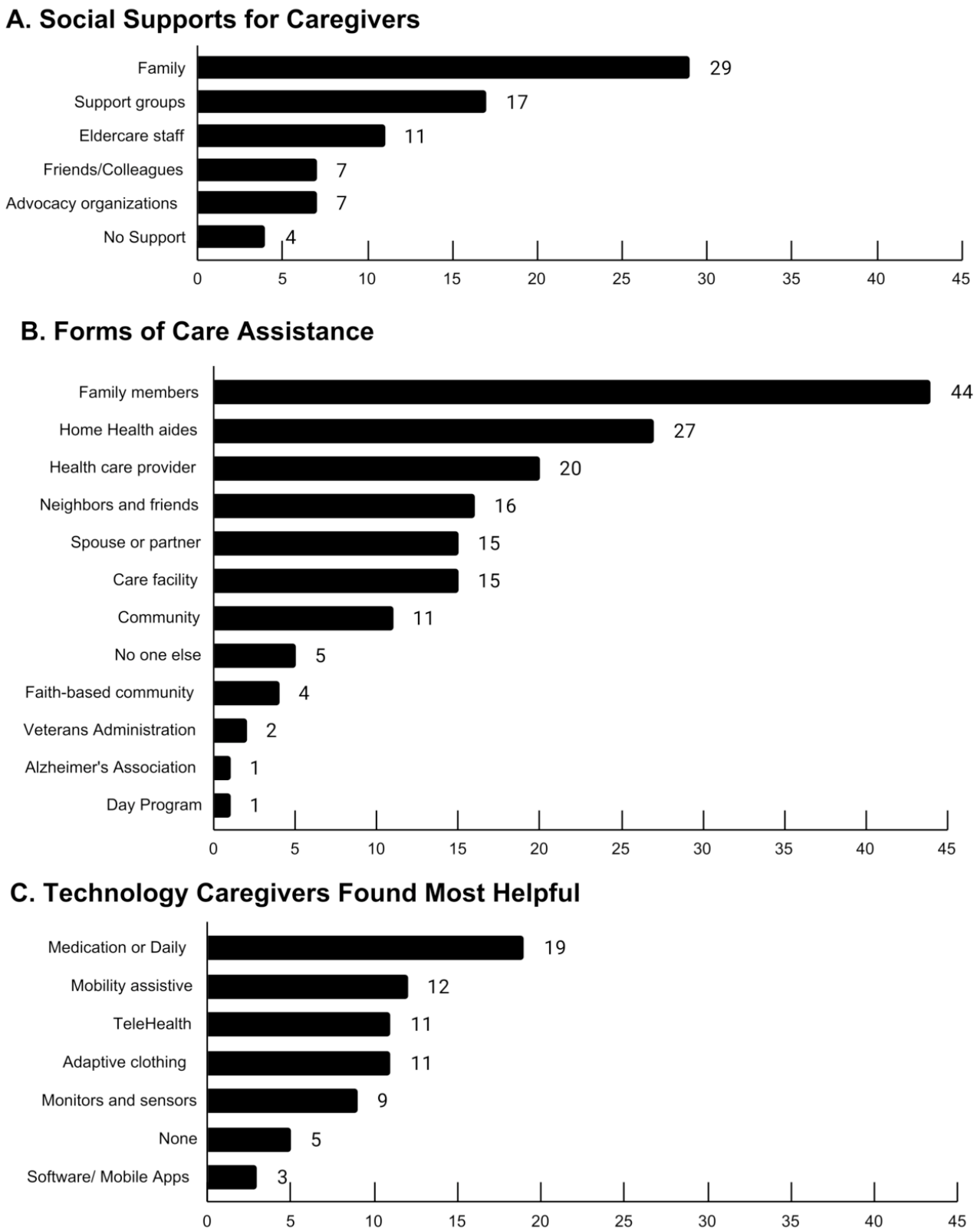

Survey Narrative

Open-ended responses about caregiving joys (n=55), struggles (n=56), and unmet needs or solution gaps (n=53) were annotated for themes, revealing distinct patterns in caregiver experiences (Figure 3). These qualitative insights, detailed in Multimedia Appendix 1, provide important nuance to the survey data and help illuminate the emotional and relational dimensions of caregiving that are often underrepresented in categorical data.

XSL•FO
RenderX

Figure 3. Occurrence of annotations for (A) positives, (B) negatives, and (C) solutions, survey narrative response sets. Bars represent total counts of each categorization. If multiple annotations were found in a response, they were ranked from primary to tertiary.

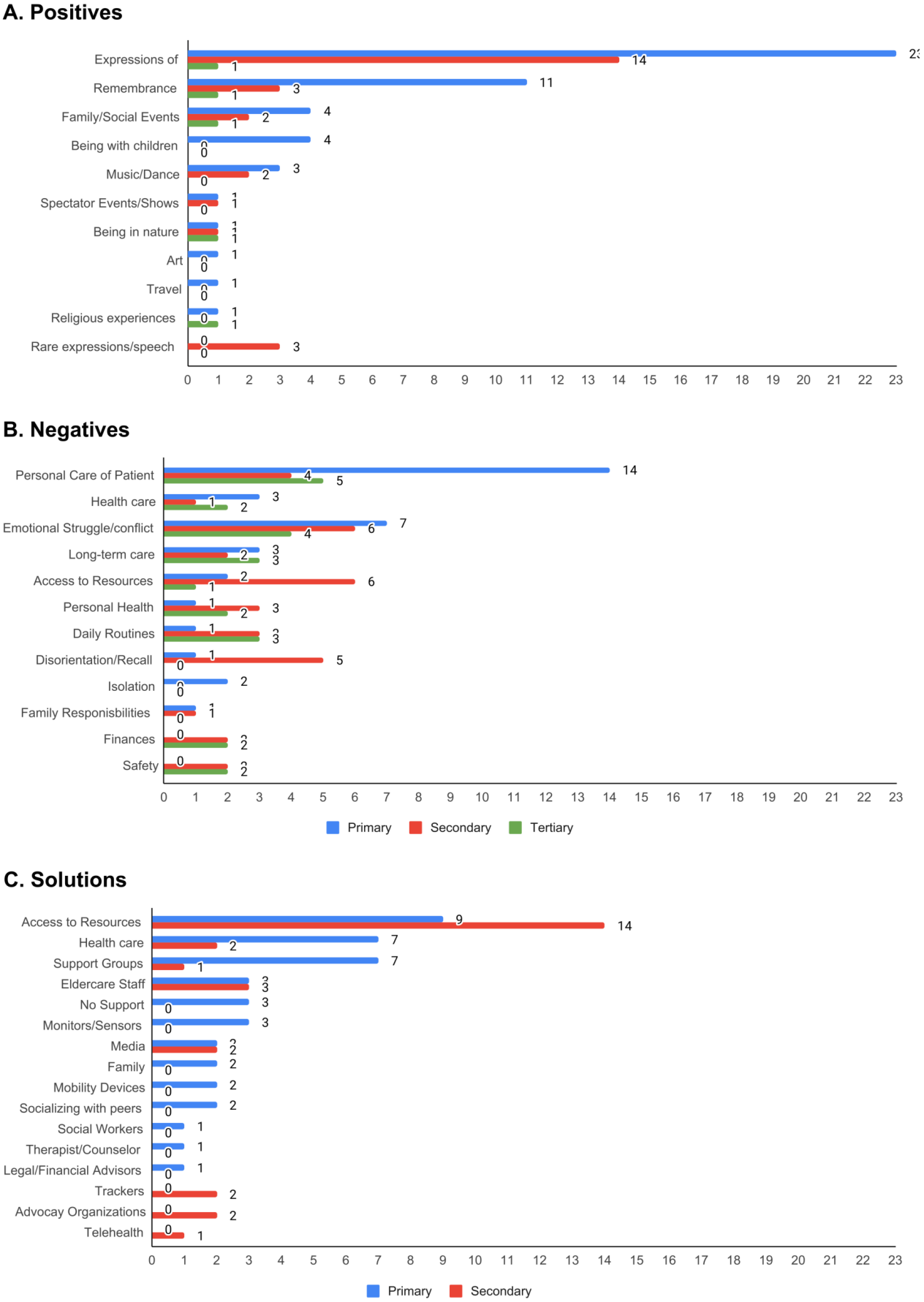

XSL•FO
RenderX

The most frequent positive experiences cited were grounded in interpersonal connection. Caregivers described moments of expressed gratitude or emotional closeness (72%) as the most meaningful aspects of their role. Instances of remembrance and reminiscence (29%), family or social gatherings (13%), and time spent with children (8%) were also commonly named. These patterns were consistent across gender, caregiver age, and relationship to the PLWD, emphasizing the universality of emotional bonding in caregiving. Even when the focus was on remaining abilities and bonding, there was often bittersweet undertones, as described by an adult child:

> *She's my mother and accepting of her dementia, a positive person and appreciative. She is also a historian, so since I retired I have had more time to help with this task. It makes me sad how much info I will lose when she is gone, and she drives me nuts with her efforts, and sad with her difficulties. Still I value the time.*

Narratives describing caregiving struggles echoed and deepened the findings from the structured survey. Caregivers frequently cited hands-on personal care, communication difficulties, and emotional strain or conflict as primary challenges. These narratives often conveyed a sense of emotional exhaustion, confusion, or sadness, especially when trying to interpret or respond to changed behaviors in the PLWD. Male caregivers were more likely to report health care system navigation and resource access as their central concerns (17% and 25%, respectively), compared to female caregivers (5% and 12%). Notably, challenges related to personal care were more frequently reported when the PLWD was a mother or mother-in-law (45%) or grandparent (40%), suggesting that gendered norms of caregiving or discomfort with certain caregiving tasks may play a role in perceived burden. A husband describes his biggest pain point in the area of health care:

> *My spouse's refusal to cooperate with physical therapy, occupational therapy and transition to mobility assistance devices.*

When asked about unmet needs, caregivers overwhelmingly emphasized the lack of access to practical resources (53%), such as information, services, or structured assistance. Health care service needs (25%), financial advising or aid (21%), and support group availability (17%) followed. Interestingly, only female respondents mentioned a desire for financial support, and younger caregivers (ages 36 - 55 years) were nearly twice as likely to name financial challenges (33%) compared to older caregivers (17%). These findings point to gender and generational differences in financial strain or planning expectations and may reflect different stages of life or career interruption due to caregiving demands.

Overall, the narratives underscore a dual reality of caregiving: caregivers deeply value emotional connection and shared meaning with the PLWD, yet they are often left without adequate support to manage the demands that threaten those very connections. These findings reinforce the call for technology and service solutions that respect both the logistical and emotional realities of caregiving: tools that can streamline care tasks without undermining the relational work caregivers view as core to their identity and role.

Interviews

Overview

The interviews were conducted with both formal and informal caregivers to gain a deeper understanding of caregivers' lived experiences and identify strategies to improve dementia care and alleviate caregiver burden—insights that extend beyond what could be captured in the survey. Through these conversations, the study aimed to inform the development of new tools, services, or supports that are directly responsive to caregiver needs. Summary and annotations for each interviewee are documented in Multimedia Appendix 1.

Five interviewers conducted in-depth interviews with 18 caregivers. Of these, 78% were female and 22% male, and the majority (75%) identified as white. The participants included 16 informal caregivers (7 spousal caregivers, 5 adult children caring for a parent, and 4 extended family caregivers) and one each from the perspectives of a therapist/specialist, nurse, and activity center staff (also counted in adult child group). Eleven of the informal caregiver interviewees had also completed the survey, allowing for a more layered understanding of their caregiving experiences. Detailed synopses and annotated transcripts are available in the Supplemental Materials in Multimedia Appendix 1.

Analysis of the interview data revealed several recurring categories of concern and support. The most frequently mentioned pain points included emotional struggle, communication challenges with the PLWD, and long-term care planning, all of which mirror the findings from the broader survey. At the same time, caregivers identified family members, support groups, and paid care staff as their most helpful resources. Positive experiences often centered on expressions of gratitude, emotional connection, and reminiscence by the PLWD, which caregivers described as small but meaningful affirmations of their role. Conversely, frustrations were rooted in the physical demands of personal care, the stress of financial management, and the persistent emotional toll of caregiving.

When asked about potential solutions, caregivers overwhelmingly called for greater access to reliable resources, more support group opportunities, improved health care navigation, and direct financial assistance. These priorities reflect a broader need not just for services but for systems that acknowledge and respond to the complex, emotional, and practical nature of dementia caregiving. For example, an adult child who also worked in a memory care center expressed how grief for her mother was a daily battle. She leaned heavily on her coworkers for emotional and practical support.

Notably, the 30 direct quotes coded from the interviews reveal where caregivers placed their emphasis: 57% (17 quotes) reflected concerns about personal well-being, including stress, burnout, and the need for time away. Twenty-three percent (7 quotes) expressed hopes and desires for better systems, including tools to reduce isolation or enable earlier planning. The remaining 20% (6 quotes) focused on strategies and potential solutions, such as improved information flow between providers,

XSL•FO
RenderX

streamlined financial support systems, and technologies that could promote connection without displacing caregiver control.

Together, the interview findings provide a deeper, more human perspective on the challenges and joys of caregiving. They reinforce the importance of designing solutions that support caregivers both emotionally and practically and that recognize caregiving as not just a role, but a relationship grounded in meaning, memory, and love.

Well-Being

Caregivers frequently described their personal well-being as marked by frustration, exhaustion, and loneliness, reflecting patterns well documented in the caregiving literature [4,7,13]. Repetition and monotony were recurring themes as one participant described feeling like "an actor in a play with the long run," while another shared, "I'm having basically the same conversation over and over again." Feelings of isolation were also strong, with one caregiver stating, "You are alone for so long." Many described being emotionally and mentally drained, with some acknowledging they were neglecting their own physical health in favor of caregiving responsibilities. One caregiver admitted, "I don't think I'm taking good physical care of myself." Others spoke of the emotional toll of facing dementia's progression, describing it as "a harsh reality" and difficult to accept.

Hopes and Desires

Caregivers expressed clear hopes for better educational resources, financial support, and stronger communication networks throughout the caregiving journey. Several participants emphasized the need for more accessible information about available services, with one noting the absence of "safety nets for seniors" and the limited support for low-income families. Others highlighted a desire for professional guidance, such as access to a "financial advisor, tax advisor, or pro bono attorney," to help navigate the complex financial aspects of caregiving. Communication gaps with care facilities were another frustration, with caregivers reporting delays or difficulty reaching staff.

Challenges

Interviewees proposed a range of solutions to ease caregiving challenges, including wellness programs for care partners, supportive community platforms, and digital tools to streamline caregiving tasks. Suggestions included exercise and mindfulness activities, like yoga, which one caregiver used in a senior day center to support both client function and caregiver–client connection. Others highlighted the value of indirect peer support, such as podcasts or articles, which helped them feel less alone and more informed. Several participants emphasized the burden of financial strain, noting that resources like a "Medicare workshop" or access to financial advisors could be valuable. On the tech front, one caregiver expressed interest in a centralized app to manage daily activities, bookmark helpful sites, and stay organized—underscoring the need for tools that enhance, rather than complicate, the caregiving experience.

Discussion

Principal Results

This study examined the concerns of and strategies taken by informal caregivers of PLWD in metropolitan, tech centers of the west coast regions of the United States. It showed that current use of technology provides limited support for the needs of caregivers. The majority of respondents were informal caregivers, and the most common concerns were related to the personal care of the PLWD, communication with the PLWD, and emotional struggle and conflict. The main source of social support for informal caregivers was identified as family, followed by support groups and care staff. Caregivers commonly reported using technology for medication and daily tracking, with other popular tools being mobility assistive devices, telehealth services, and adaptive clothing. A small number of respondents used monitors and sensors. The dominant need expressed by caregivers was for better access to resources, and improved access to health care services, financial support, and support groups was also frequently mentioned. These insights echo broader findings that caregivers need not only emotional and practical support but also clearer pathways to resources and systems that respond more efficiently to their evolving needs [8,16,21].

The interviews provided clear evidence of how dementia impacts not only the individual living with the condition but also their caregiver. Caregivers primarily focused on the effects on their personal well-being that led to loneliness, frustration, helplessness, and exhaustion. To alleviate their burden, several caregivers suggested improving resource accessibility. These respondents were aware that many support mechanisms existed, including online resources, support groups, and health care services, and acknowledged that specific strategies can ease their workload and alleviate negative emotions associated with caregiving. However, barriers to accessing resources were commonly cited, including high costs and difficulties in selecting appropriate resources. Additionally, centralizing resources and simplifying the process of locating support was recommended as a solution. Digital technologies, if made affordable and easy to use, could help in these services.

Comparison to Prior Work

The results of this study align closely with previous reports on the needs of dementia caregivers. Research consistently shows that caregivers experience higher levels of stress, depression, and anxiety symptoms compared to noncaregivers [4,7]. Previous studies have identified primary stressors such as emotional outbursts or behavioral problems from PLWD as the strongest predictors of poor health outcomes among caregivers [7,22]. Our survey and narrative data reinforce that emotional struggle and interpersonal conflict with the PLWD are the primary concerns among most caregivers. However, the majority of existing data on caregiver needs has been gathered through categorical surveys and fails to provide sufficient detail on specific caregiver pain points. This report complements larger-scale surveys, such as the annual *Alzheimer's Association Facts and Figures* report [4], by offering qualitative depth and contextual understanding.

XSL•FO
RenderX

Moreover, a wide range of educational and psychosocial interventions have been shown to benefit caregivers within comprehensive care models [4,22]. Specifically, interventions that are personalized and actively involve family members tend to be more effective [7,8,18,23]. In addition to mitigating negative outcomes, supporting meaningful activities between care partners can have a substantial impact on caregiver well-being. Caregivers who report more positive interactions with the PLWD, such as moments of gratitude or shared remembrance, demonstrate better mental health outcomes [24]. Our survey illustrated this relationship: caregivers who reported emotionally positive experiences were less likely to indicate poor personal health and psychological distress. Although digital technologies like DevaWorld and FamilyShare have emerged to promote connection and engagement, none of our participants reported using such tools, suggesting a gap between available resources and actual caregiver uptake.

Access to resources also emerged as the most commonly cited unmet need, consistent with prior research highlighting caregivers' desire for more support from health care providers, family, and community services [8]. Research is growing to inquire on which technologies caregivers are currently using or how they envision technology helping them meet their needs, especially for mobile applications [12,25]. This underscores a need for more caregiver-centered inquiry. For example, Smriti et al identified the ongoing management of feelings and emotional connection as a central aspect of dementia caregiving [18]. Their study found that while caregivers were willing to use technology for practical tasks (eg, reminders, telehealth), they resisted using it for emotionally meaningful caregiving activities, such as bathing or deep conversations. These emotionally rich tasks are viewed as essential to caregivers' identities and bonds with the PLWD. Smriti et al emphasized that technology design should not aim to replace this emotion work but rather support it through features that foster reflection, shared memories, or emotional resilience. Our findings align with this approach: caregivers valued tools that reduced practical burdens but wanted to retain control over emotionally significant aspects of caregiving. As digital solutions continue to evolve, developers and care systems must adopt a holistic, caregiver-informed approach to technology design that supports, not supplants, the human core of caregiving.

Technology Adoption Recommendations

Overview

Dementia is already one of the most pressing issues affecting our aging population, currently affecting around 50 million people worldwide, and this number is expected to triple by 2050, meaning the demand for caregivers will rise dramatically in stride [8]. The cost of dementia care to society will compound by the additional impact on caregivers' physical and mental health. Persistent high workload and emotional strain on caregivers increase the likelihood of transitioning from aging-in-place to residential care [26] and are associated with increased medical needs of the caregiver [4]. The summative effect of a fractured model of dementia care is ultimately more expensive and delivers a poorer quality of care, compared to comprehensive care models [21,27,28]. Assistive technology can play a significant role in improving the caregiving experience for both the caregivers and the individuals receiving care [1,13,29]. As new technologies emerge, it is critical to focus on innovation that can address the specific needs of caregivers. One approach to this is summarized by Kiselica et al, which draws a framework to apply common technologies to dementia care [30].

In response to this, many organizations are currently working to support innovation and provide solutions to address the specific needs of people affected by dementia. For example, the American Association for Retired Persons has teamed with the Alzheimer's Association, Challenge Works, and other organizations in sponsoring an international competition to find innovative solutions for dementia care [31]. The competition focuses on technology-based solutions such as using artificial intelligence to provide better care to PLWD while reducing caregiver workload. The Alzheimer's Association has also launched an initiative called the Alzheimer's Association Innovation Roundtable, which sets the goal to strengthen existing approaches while developing new technology that reimagines dementia care. The roundtable offers impact-focused meetings, webinars, and an annual global innovation competition called the Alzheimer's Association Pitch Competition [32].

It is important to address that most caregivers are not using technology to address their highest areas of concern, such as locating support resources and dealing with emotions of frustration and helplessness. Even in our sample from "Silicon Valley" and Seattle, where digital technology innovation is the economic engine, technology solutions are not reaching the daily lives of caregivers. It would be fair to extrapolate that in areas that are lower resourced and with sparser internet connectivity, the adoption rates would be even lower. There are several factors that contribute to caregivers' current underutilization of technology, such as the challenge of identifying appropriate technology solutions, the high costs of existing products, and the scarcity of solutions specifically tailored to the unique needs of dementia care. The following are recommendations based on the caregiver responses that could improve adoption of technology for dementia care.

Need for Centralized Resources

Despite the large number of resources in existence for caregivers, study respondents noted that it is often difficult to sort through and find relevant information, such as difficulty finding a local support group in large databases or through a search engine. Therefore, it is critical to think about centralizing resources in a manner that facilitates easy navigation for users. My ALZ Journey mobile app was released by the Alzheimer's Association in 2025 to address navigation from diagnosis to long-term care plans [33]. The reach of the organization is comprehensively national, and the app is available at no cost and does not store user personal information. The potential for this as an effective, centralized resource is clear. Whether it is utilized as intended will be observed with time.

Accessibility and Fit

Even after overcoming the challenge of identifying a product or service, additional barriers still inhibit usage. They include

XSL•FO
RenderX

the complexity of user interfaces, the bias of English-only text, or recurring subscription fees. The prerequisite requirements to reach a broad population of caregivers are usability, cultural inclusivity, and affordability. Engagement of diverse stakeholders in product development and the investment in innovation for smaller populations will lead to improved accessibility and fit of digital tools for family caregivers. At present, a one-size-fits-all approach is evident in first-to-market solutions from established organizations (eg, My ALZ journey). The foundation of these applications can serve as the scaffolding for culturally aligned content.

Develop Stakeholder-Informed Solutions

Safety and monitoring assistive technologies are examples of how technology can improve the quality of life in both home settings and care facilities [16,34]. Video monitors and GPS devices to track the PLWD in real time have been shown to increase the perceived security of the PLWD for caregivers [35,36]. Smart-home systems have also gained attention for the ability to extend the PLWD with increased autonomy to live at home and ease caregiving responsibilities [37]. For example, automatic sensors, typically paired to a central control panel or app, can manage lights and doors and detect sudden heat changes. Additionally, web-based technology such as social network platforms, informative websites, and educational videos has been received positively by caregivers to increase their social support access to resources [38,39]. Collectively, these systems reduce the workload of caregivers while giving caregivers relevant information to deliver care.

Limitations

There are several limitations to consider regarding the sample profile of this needs assessment study. First, nearly three-quarters of respondents were identified as White, which raises concerns that the results may not accurately reflect the experiences of caregivers from different cultural, educational, and socioeconomic backgrounds. Additionally, most of the caregivers in this study were 55 - 74 years of age. This is higher than the average age of caregivers in the United States of about 50 years [40], which means that there may be differences in the unmet needs of individuals in this study by comparison to most studies. Specifically, technology usage is strongly influenced by age [41]. Thus, the low usage of technology-driven strategies and difficulty in finding resources digitally and via the internet, found in our sample, may only be representative of older adults, and accessibility may have improved since the time of this study in 2022. Future research should aim to include a more diverse sample and consider these factors to provide a broader understanding of the experiences of caregivers in different populations.

Conclusions

The findings from this inquiry into caregivers of PLWD illustrate the complexity of their needs and the nuanced role that technology can play in alleviating caregiver burden. Through a mixed-methods approach that integrated both quantitative and narrative data, this study revealed that the greatest barriers to technology adoption are not a lack of available tools but rather limited accessibility, awareness, and fit with caregivers' lived experiences. These barriers are particularly pressing given the growing interest in AgeTech and digital health solutions. Encouragingly, such barriers are more tractable than systemic resource shortages and can be addressed through coordinated action among dementia care advocates, public health institutions, and designers committed to inclusive, human-centered technologies.

Importantly, this study contributes to a deeper understanding of why caregivers often do not adopt existing technologies. As also emphasized in a recent scoping review, informational gaps and caregivers' emotional and physical health remain among the most frequently cited unmet needs, especially when caregivers feel unsupported or lack time for self-care [8]. Emotional health was the most commonly reported need across the majority of the studies in that review, followed by a need for formal or informal help and better access to information. Our findings reinforce these trends and emphasize that interventions, whether digital or community-based, must go beyond utilitarian functions and instead support the emotional labor inherent in caregiving.

Technology, when properly designed, can help caregivers offload logistical tasks such as scheduling, tracking medications, and communication but should avoid replacing emotionally meaningful interactions that define caregiving identity and purpose. As other researchers have noted, attempts to fully automate emotional aspects of caregiving risk undermining what caregivers value most about their role [19]. Instead, digital tools should aim to augment emotion work by enabling reflection, shared memory, and just-in-time support.

Ultimately, this study highlights the urgent need for cross-sector collaboration among health professionals, caregivers, researchers, technologists, and policymakers to develop resources that are accessible, context-sensitive, and culturally appropriate. Addressing caregivers' unmet needs will require not just innovation but integration: aligning digital tools with psychosocial supports, community networks, and education tailored to different stages of dementia. In doing so, we can better support caregivers in sustaining both their caregiving responsibilities and their own well-being.

Acknowledgments

The authors thank all the caregivers who shared their time and insights for this study.

This work was funded by Maude's Ventures of the Ferry Foundation, Seattle, WA. The foundation also facilitated the distribution of the online survey in their local community. The funder had no involvement in the study design, analysis, interpretation, or the writing of the manuscript.

XSL•FO
RenderX

Data Availability

The data supporting the findings of this study are available within the article and/or its supplementary material.

Authors' Contributions

Conceptualization: JAS, SY, PS
Methodology: JAS, SY, PS
Formal analysis: JAS, MH, KA, KG, AM
Writing - Original Draft: JAS, EC, AM
Writing - Review & Editing: JAS, SY, PS
Visualization: JAS, KG
Supervision: JAS
Project administration: JAS
Funding acquisition: JAS
Data Curation: MH, KG

Conflicts of Interest

None declared.

Multimedia Appendix 1
Summary and annotations for survey and interview datasets.
[PDF File, 1005 KB - formative_v9i1e69596_app1.pdf]

References

1. D'Onofrio G, Sancarlo D, Ricciardi F, et al. Information and communication technologies for the activities of daily living in older patients with dementia: a systematic review. J Alzheimers Dis 2017;57(3):927-935. [doi: 10.3233/JAD-161145] [Medline: 28304297]
2. World Health Organization. Global Action Plan on the Public Health Response to Dementia 2017–2025: World Health Organization; 2017. URL: https://iris.who.int/handle/10665/259615 [accessed 2025-07-09]
3. Gaugler JE, Jutkowitz E, Peterson CM, Zmora R. Caregivers dying before care recipients with dementia. A&D Transl Res & Clin Interv 2018 Jan;4(1):688-693. [doi: 10.1016/j.trci.2018.08.010]
4. 2024 Alzheimer's disease facts and figures. Alzheimers Dement 2024 May;20(5):3708-3821 [FREE Full text] [doi: 10.1002/alz.13809]
5. Mueller A, Thao L, Condon O, Liebzeit D, Fields B. A systematic review of the needs of dementia caregivers across care settings. Home Health Care Manage Pract 2022 May;34(2):133-147. [doi: 10.1177/10848223211056928]
6. Gitlin LN, Marx K, Scerpella D, et al. Embedding caregiver support in community-based services for older adults: a multi-site randomized trial to test the Adult Day Service Plus program (ADS Plus). Contemp Clin Trials 2019 Aug;83:97-108. [doi: 10.1016/j.cct.2019.06.010] [Medline: 31238172]
7. Sörensen S, Conwell Y. Issues in dementia caregiving: effects on mental and physical health, intervention strategies, and research needs. Am J Geriatr Psychiatry 2011 Jun;19(6):491-496. [doi: 10.1097/JGP.0b013e31821c0e6e] [Medline: 21502853]
8. Queluz F, Kervin E, Wozney L, Fancey P, McGrath PJ, Keefe J. Understanding the needs of caregivers of persons with dementia: a scoping review. Int Psychogeriatr 2020 Jan;32(1):35-52. [doi: 10.1017/S1041610219000243] [Medline: 30967164]
9. Denham AMJ, Wynne O, Baker AL, et al. An online survey of informal caregivers' unmet needs and associated factors. In: Jan YK, editor. PLOS ONE 2020;15(12):e0243502. [doi: 10.1371/journal.pone.0243502] [Medline: 33301483]
10. Rocha IC, Arantes M, Moreira A, et al. Monitoring wearable devices for elderly people with dementia: a review. Designs 2024 Jul 29;8(4):75. [doi: 10.3390/designs8040075]
11. Palmese F, Druda Y, Benintende V, et al. Wearable sensors for monitoring caregivers of people with dementia: a scoping review. Eur Geriatr Med 2025 Apr;16(2):473-483. [doi: 10.1007/s41999-024-01113-8] [Medline: 39625554]
12. Lee T, Seohyun K, Park S. Mobile health applications for communication between caregivers of community-dwelling older adults: a scoping review. Geriatr Nurs (Lond) 2023 Jul;52:172-180. [doi: 10.1016/j.gerinurse.2023.06.011]
13. Mikula CM, Perry C, Boone AE, Benge JF, Scullin MK, Kiselica AM. Dementia caregiver insights on use of assistive technologies. Work Aging Retire 2024 Jan 8;10(1):14-24. [doi: 10.1093/workar/waac027]
14. Sharpe BT, King R, Banerjee M. Experiences of caregivers, service providers, and individuals living with dementia regarding independence-enhancing technologies: focus group insights. PsyArXiv. . [doi: 10.31234/osf.io/jw593]
15. Coelho T. Digital technologies in dementia care. In: Digital therapies in psychosocial rehabilitation and mental health 2022:115-140. [doi: 10.4018/978-1-7998-8634-1.ch006]

XSL•FO
RenderX

16. Boyle LD, Husebo BS, Vislapuu M. Promotors and barriers to the implementation and adoption of assistive technology and telecare for people with dementia and their caregivers: a systematic review of the literature. BMC Health Serv Res 2022 Dec 23;22(1):1573. [doi: 10.1186/s12913-022-08968-2] [Medline: 36550456]
17. AgeTech Collaborative. The state of agetech: insights in brief. : AARP; 2024 Jul URL: https://agetechcollaborative.org/wp-content/uploads/2024/07/ATC_-State-of-AgeTech-Report_Final.pdf [accessed 2024-12-20]
18. Smriti D, Wang L, Huh-Yoo J. Emotion work in caregiving: the role of technology to support informal caregivers of persons living with dementia. Proc ACM Hum Comput Interact 2024 Apr 17;8(CSCW1):1-34. [doi: 10.1145/3637325]
19. Williams M, Moser T. The art of coding and thematic exploration in qualitative research. Int Manag Rev 2019;15:45 [FREE Full text]
20. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
21. Super N, Ahuja R, Mcdermott M. Scaling comprehensive dementia-care models. : Milken Institute; 2021 URL: https://milkeninstitute.org/content-hub/research-and-reports/reports/scaling-comprehensive-dementia-care-models [accessed 2025-09-29]
22. Chiao CY, Wu HS, Hsiao CY. Caregiver burden for informal caregivers of patients with dementia: a systematic review. Int Nurs Rev 2015 Sep;62(3):340-350. [doi: 10.1111/inr.12194] [Medline: 26058542]
23. Sanchez AA, Lai J, Ye B, Mihailidis A. Enhancing communication and autonomy in dementia through technology: navigating home challenges and memory aid usage. Gerontechnology 2024 Feb 10;23(1):1-11 [FREE Full text] [doi: 10.4017/gt.2024.23.1.880.06] [Medline: 20154507]
24. Quinn C, Toms G. Influence of positive aspects of dementia caregiving on caregivers' well-being: a systematic review. Gerontologist 2018 Dec 28. [doi: 10.1093/geront/gny168]
25. Kuo HL, Chang CH, Ma WF. A survey of mobile apps for the care management of patients with dementia. Healthcare (Basel) 2022 Jun 23;10(7):1173. [doi: 10.3390/healthcare10071173] [Medline: 35885700]
26. Auriemma CL, Butt MI, McMillan J, et al. "What choice do we have?" Reactive and proactive decision‐making for aging in place with dementia. J Am Geriatr Soc 2024 Nov;72(11):3398-3412 [FREE Full text] [doi: 10.1111/jgs.19140]
27. Reuben DB, Gill TM, Stevens A, et al. Health system, community-based, or usual dementia CARE for persons with dementia and caregivers: the D-CARE randomized clinical trial. JAMA 2025 Mar 18;333(11):950-961. [doi: 10.1001/jama.2024.25056] [Medline: 39878968]
28. Reuben DB, Epstein-Lubow G, Super N. The other dementia breakthrough—comprehensive dementia care. JAMA Neurol 2023 Aug 1;80(8):770-772. [doi: 10.1001/jamaneurol.2023.1252] [Medline: 37213130]
29. Pappadà A, Chattat R, Chirico I, Valente M, Ottoboni G. Assistive technologies in dementia care: an updated analysis of the literature. Front Psychol 2021;12:644587. [doi: 10.3389/fpsyg.2021.644587] [Medline: 33841281]
30. Kiselica AM, Hermann GE, Scullin MK, Benge JF. Technology that CARES: enhancing dementia care through everyday technologies. Alzheimers Dement 2024 Dec;20(12):8969-8978. [doi: 10.1002/alz.14192] [Medline: 39508340]
31. AARP backs new $5 million dementia tech contest. AARP. 2022. URL: https://blog.aarp.org/fighting-for-you/longitude-prize-on-dementia-contest [accessed 2025-07-10]
32. The alzheimer's association and MATTER launch pitch competition for solutions to increase early detection and diagnosis of dementia within primary care settings. Alzheimer's Association. 2022. URL: https://www.alz.org/news/2022/the-alzheimer-s-association-and-matter-launch-pitch-competition-for-solutions-to-increase-early-dete [accessed 2025-07-08]
33. Leising J. New tool: my ALZ journey app supports early-stage dementia navigation. 2025. URL: https://www.alz.org/wi/blog/my-alz-journey-app [accessed 2025-07-10]
34. Sixsmith A, Sixsmith J, Fang ML, Horst B. AgeTech, cognitive health, and dementia. In: AgeTech, Cognitive Health, and Dementia 2020:15-20 URL: https://link.springer.com/10.1007/978-3-031-01605-9 [doi: 10.1007/978-3-031-01605-9_3]
35. Behera CK, Condell J, Dora S, Gibson DS, Leavey G. State-of-the-art sensors for remote care of people with dementia during a pandemic: a systematic review. Sensors (Basel) 2021 Jul 8;21(14):4688. [doi: 10.3390/s21144688] [Medline: 34300428]
36. Smith ML, Wilson AD, Knebl J, et al. Utilization and perceived usefulness of monitoring technology for family caregivers of people living with Alzheimer's disease and related dementias. J Alzheimer's Dis 2025 Jan;103(1):127-132. [doi: 10.1177/13872877241300078]
37. Vrančić A, Zadravec H, Orehovački T. The role of smart homes in providing care for older adults: a systematic literature review from 2010 to 2023. Smart Cities 2024 Jun 26;7(4):1502-1550. [doi: 10.3390/smartcities7040062]
38. Verwey R, van Berlo M, Duymelinck S, Willard S, van Rossum E. Development of an online platform to support the network of caregivers of people with dementia. Stud Health Technol Inform 2016;225(567–571):567-571. [Medline: 27332265]
39. Boutilier JJ, Loganathar P, Linden A, et al. A web-based platform (CareVirtue) to support caregivers of people living with Alzheimer disease and related dementias: mixed methods feasibility study. JMIR Aging 2022 Aug 4;5(3):e36975. [doi: 10.2196/36975] [Medline: 35925654]
40. Caregiving in the US 2020. AARP. 2020. URL: https://www.aarp.org/pri/topics/ltss/family-caregiving/caregiving-in-the-united-states/ [accessed 2025-09-29]

XSL•FO
RenderX

41. Kakulla B. Tech trends and adults 50-plus. AARP Res 2024 Dec. [doi: 10.26419/res.00772.001]

Abbreviations

PLWD: Persons living with dementia

Edited by A Stone, A Mavragani; submitted 20.12.24; peer-reviewed by C Jensen, J Waycott, Y Yang; revised version received 11.08.25; accepted 14.08.25; published 04.11.25.

Please cite as:
Scott JA, Cepukenas E, Himes M, Anderson K, Grewal K, Mai A, Yuter S, Simone P
Dementia Caregivers' Perspectives on Technology's Place in Care Practices: Mixed Methods Survey
JMIR Form Res 2025;9:e69596
URL: https://formative.jmir.org/2025/1/e69596
doi:10.2196/69596

© Julia A Scott, Emma Cepukenas, McKenzie Himes, Kennedy Anderson, Kiren Grewal, An Mai, Sheila Yuter, Patricia Simone. Originally published in JMIR Formative Research (https://formative.jmir.org), 4.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Health Service Early-Stage Digital Adaptation of Traditional Chinese Medicine Internet Hospitals: Qualitative Exploratory Study

Yao Wang[1*], MA; Menghuan Song[1,2*], PhD; Zhenmiao Pang[3], PhD; Buping Liu[3], PhD; Dongning Yao[4], PhD; Meng Li[5], PhD; Zhirong Xie[6], MA; Ying Bian[1,2,7], PhD; Hao Hu[1,2,7], PhD; Yunfeng Lai[3], PhD

[1]State Key Laboratory of Mechanism and Quality Research in Chinese Medicine, Institute of Chinese Medical Sciences, University of Macau, Macao, China

[2]Centre for Pharmaceutical Regulatory Sciences, University of Macau, Macao, China

[3]School of Public Health and Management, Guangzhou University of Chinese Medicine, No. 232, Huandong Road, University City, Panyu District, Guangzhou, China

[4]School of Pharmacy, Nanjing Medical University, Nanjing, China

[5]Key Laboratory of Environmental Medicine Engineering of Ministry of Education, Department of Medical Insurance, School of Public Health, Southeast University, Nanjing, China

[6]Department of Science and Technology, Guangzhou University of Chinese Medicine, Guangzhou, China

[7]Department of Public Health and Medicinal Administration, Faculty of Health Sciences, University of Macau, Macao, China

[*]these authors contributed equally

Corresponding Author:
Yunfeng Lai, PhD
School of Public Health and Management, Guangzhou University of Chinese Medicine, No. 232, Huandong Road, University City, Panyu District, Guangzhou, China

Abstract

Background: Traditional Chinese medicine (TCM) hospitals in China are experimenting to develop internet hospitals to provide health services. To date, little is known about the characteristics of health services delivered by TCM internet hospitals.

Objective: This study aimed to investigate the health service early-stage digital adaptation of TCM internet hospitals from the aspects of target patients, value offering, and service provision.

Methods: Qualitative research combined qualitative interview and documentary research in this study. Interviews were completed with clinicians from sample TCM internet hospitals to investigate the target patients and value offerings. Documentary research was conducted to investigate the service provision. Thematic analysis was used to interpret all the materials collected.

Results: A total of 7 TCM internet hospitals and 14 participants were included. The target patients of TCM internet hospitals were patients with subsequent visits and patients who sought consultations on health management. TCM internet hospitals were improving patients' adherence to subsequent medical care and TCM promotion. These hospitals provided functional service (including telemedicine, telepharmacy, telenursing, web-based health consultations, and convenient service), and TCM specialty service (including "Tianzhi" [crude herb moxibustion], "Zhiweibing" [preventive treatment of disease], and poststroke rehabilitation).

Conclusions: TCM internet hospitals are in an early-stage digital adaptation, offering primarily basic online-offline services. While not yet fully innovative, they represent a transitional model with the potential to reshape TCM delivery. Our findings contribute high-level insights into this emerging integration and inform future development toward more structured, patient-centered digital TCM services.

(JMIR Form Res 2025;9:e77686) doi:10.2196/77686

KEYWORDS

traditional Chinese medicine; internet hospital; health service; target patient; value offering

Introduction

Background

Digital technology has exceptional advantages for disease prevention and treatment. The advantages are improving the feasibility of health information collection and patient-centered care, enhancing the efficiency of monitoring, diagnoses, and prognoses of diseases, upgrading medical services, and reducing service costs [1]. In the context of digitalization, the Chinese government is actively promoting the integration of the *internet* and *health care* [2]. Internet hospitals rely on brick-and-mortar hospitals as a base and benefit from internet technology and

XSL•FO
RenderX

health resources of their physical hospitals and provide patients with closed-loop medical services [3]. Internet hospitals, guided by government policy, are rapidly expanding to enhance access to quality health care for patients with common or chronic conditions and those in remote areas, thereby improving resource accessibility and saving time for patients and their families [4].

In China, traditional Chinese medicine (TCM) hospitals have evolved into institutions offering comprehensive and diversified TCM services and products, including Chinese patent medicines, herbal pieces, formula granules, and raw herbs [5]. As one of the oldest therapeutic systems in the world, TCM is gaining increasing attention from global society [6]. In the background of digitalization, TCM hospitals also dived into the development of web-based service platforms. Until August 2021, there were more than 1600 internet hospitals across China, of which about 200 were TCM internet hospitals located in Guangdong, Shandong, and Jiangsu provinces [7,8]. TCM internet hospitals have emerged as a vital component of modern health care, especially in China, where digital health initiatives are rapidly transforming the medical landscape. These hospitals integrate TCM with cutting-edge digital technologies, providing web-based registration, health consultation, prescriptions, and payment services based on TCM theories and practices. Although the amount of health resources allocated to TCM hospitals is increasing [9], the efficiency of their service is unsatisfactory [10]. TCM Internet hospitals, which have the integration of TCM and digitalization, are amenable to offering efficient service [11]. Unlike conventional biomedicine, TCM emphasizes preventive care and individualized diagnosis through pattern differentiation. These features present unique opportunities for digital health innovation, such as personalized service design and algorithm-enabled diagnostic support, which remain underexplored in current TCM internet hospital models.

As an emerging model, “internet+TCM health care” has demonstrated significant advantages in enhancing public accessibility and promoting the rational distribution of TCM medical resources. It has increasingly become a crucial supplementary component to offline TCM clinical services and an integral part of China’s health care system [12]. Moreover, it can also develop internet+TCM home care services, which is good for improving the quality of life of older adult patients with chronic diseases [13]. For the development of TCM internet hospitals, there are many practical experiences that can be promoted, including (1) “internet hospitals+master of TCM” will be the preferred path for TCM hospitals to innovate and create distinctive TCM services; (2) TCM internet hospitals can be used to expand the scope of TCM health management services for preventing diseases; and (3) the development of TCM internet hospitals is conducive to the sustainable development of TCM services [14]. TCM internet hospitals have partially met patients’ demands for web-based health care services. However, their overall implementation remains limited, lacking distinctive TCM characteristics and demonstrating a low degree of integration with offline medical services [15,16]. It should also be noted that due to the lack of appropriate “internet+ health care” items, the use rate of internet medical care in grassroots TCM medical institutions is relatively low [17]. In addition, TCM mobile medical apps represent a specific developmental form of the “internet+ TCM” model. While it highlights the distinctive role of TCM in disease treatment, prevention, and health maintenance, there remains significant room for improvement in areas such as information service models, functional applications, and service content [18].

Investigation of health services of TCM internet hospitals is important for gaining knowledge of the characteristics of the web-based service and stimulating improvement in service quality and efficiency [19]. The recent studies mainly focus on the evaluation index system of information service quality of TCM internet hospitals [20], the resource guarantee, service pricing, service accessorily, and data information security of TCM internet hospitals [21]. The research questions that need to be checked are: What are the characteristics of the health services delivered by the TCM internet hospitals? What are the health services early-stage digital adaptation of TCM internet hospital?

Objectives

While existing literature and policy documents outline the expansion of digital TCM services, few studies have critically examined how these developments are operationalized within TCM internet hospitals, nor have they assessed whether such services reflect true health service innovation beyond digitized access. This study addresses this gap by investigating the health service early-stage digital adaptation of TCM internet hospitals from the aspects of target patients, value offering, and service provision. The findings of this study are anticipated to offer insights into the integration of TCM and digitalization. It is expected to provide empirical references for personalized digital health services and patient-centered care in TCM internet hospitals, thereby promoting the inheritance and development of TCM.

The conceptual basis of this study. As leading scholars in the field of business model research, Johnson et al [22] argued that a business model consists of four interrelated elements that collectively create and deliver value: (1) value proposition (how TCM internet hospitals enhance patient services); (2) key resources; (3) key processes refer to the operational and managerial aspects and assets necessary for delivering the value proposition; and (4) profit formula (delineate how TCM internet hospitals create and deliver value to both themselves and their patients). Target customer, value offering, and service provision are 3 elements of a value proposition in business model theory [22]. Value offering refers to the important problems or needs of the target customer that are solved or fulfilled, and service provision refers to what specific services are provided. In this study, the value proposition pertains to how TCM internet hospitals enhance patient services. The health service early-stage digital adaptation of TCM internet hospitals can be categorized based on the following 3 aspects: the target patients of TCM internet hospitals, the value offering of TCM internet hospitals, and the service provision of TCM internet hospitals [4].

XSL•FO
RenderX

Methods

Research Design

This study conducted a qualitative case study design combining qualitative interviews with participants and documentary research, as TCM internet hospitals are still an emerging sector of web-based health service [19,23]. In this study, we first sampled TCM internet hospitals and described their characteristics. Then, to prepare the materials, interviews were conducted with clinicians experienced in web-based TCM service, and any information posted by sample hospitals on official service platforms (such as WeChat [Tencent], app, and website) was investigated. Data were extracted by reviewing ready materials, including information related to health services (target patients, value offering, and service provision). Finally, data were analyzed, categorized, and assigned themes. Key findings were distilled.

Research Site: Samples of TCM Internet Hospitals

Internet hospitals in China are all built on the basis of brick-and-mortar hospitals. The sample collection was located in the city of Guangzhou due to its advances in the development of TCM internet hospitals. Data sources were the group of TCM internet hospitals registered at the Guangzhou Municipal Health Commission. Until November 1, 2023, there are a total of 12 TCM hospitals in Guangzhou, all of which claim to provide internet hospital services. We screened the registered hospitals against a set of predetermined selection criteria. TCM internet hospitals were included if: (1) they were established based on Internet Hospital Management Measures (trial) which is a compulsive policy to standardize web-based diagnosis and treatment behaviors, improve medical service efficiency, ensure medical service quality and safety, were general TCM hospitals or hospitals of integrated TCM and western medicine which principally practice TCM; and (2) they were capable of providing medical services via the WeChat platform or app. Finally, 7 TCM internet hospitals met the inclusion criteria. The official webpages of sample hospitals were investigated to extract characteristics, including names of the TCM internet hospitals and names, types (for example, TCM general hospitals and hospitals of integrated TCM and western medicine), grades, business nature (for profit or nonprofit), and the number of beds of the associated brick-and-mortar hospitals. These characteristics were tabulated.

This study was conducted in Guangzhou, one of the most advanced regions in the development of internet-based TCM services in China. A maximum variation sampling strategy was applied to include 7 tertiary-level hospitals with different ownership structures, institutional types, and service models. This approach sought to capture heterogeneity within a single regulatory environment. Although geographically restricted, this design enhances the transferability of findings by providing insights from a leading context of digital TCM implementation.

Data Collection

Qualitative Interviews

Purposive sampling was conducted to select participants who represented the health services of the sample TCM internet hospitals in this study. Prior to sampling, we consulted the medical administrators of each selected TCM internet hospital to determine the clinical departments that deliver the highest volume of medical services within the internet hospital setting. Subsequently, we selected participants from the corresponding clinical departments. We ensured the recruitment of at least 1 participant from each sampled hospital, respectively. Participants with at least 2 years of web-based service experience were approached. The sample size was determined based on the principle of data saturation. We monitored thematic emergence during the interview process using an iterative coding approach. After each round of interviews, 2 researchers independently conducted preliminary coding to identify new themes. Saturation was considered reached when no new codes or themes emerged from at least 2 consecutive interviews. In this study, after the twelfth interview, no additional themes were generated, and the final 2 interviews only confirmed previously identified themes, indicating that thematic saturation had been achieved. The interviews were conducted between November 2023 and December 2023 and adopted a semistructured format by following our tailor-made interview outline.

There is no prior association between the researchers and participants. Before conducting the interviews, we obtained informed consent from the participants. The informed consent forms and interview schedule can be found in Multimedia Appendices 1 and 2. All interviews involved questions centered on target patients and the value offerings of TCM internet hospitals, and interviewees' perceptions of the service provisions of TCM internet hospitals and traditional brick-and-mortar hospitals. We recorded the audio of each interview with the participants' consent. In the process of converting the audio recording into transcripts, two researchers first transcribed the recording independently, then compared the results and marked the unclear and different sections. A third researcher reviewed these marked sections against the recording. Finally, if any ambiguities remained, we verified the unclear parts with the participant. In total, 8 interviews were conducted face-to-face, and 6 interviews were conducted via telephone. The average interview time was about 40 (SD 5.0709) minutes.

Documentary Data Collection

When we examined each sample TCM internet hospital one by one, we found that they all used WeChat as the service platform to provide health services. Data were collected from the official WeChat platforms and linked mini programs of the selected TCM internet hospitals. Specifically, WeChat mini programs refer to the *subapplications* within the WeChat ecosystem that enable the provision of advanced features to patients, such as service content, target patients, prices, and other services. To ensure data validity, all WeChat-based information was extracted directly from the official mini programs linked to each hospital's verified WeChat account and double-checked by 2 researchers independently. While minor differences in interface design and service presentation existed across platforms, we focused our analysis on standardized data elements (eg, service items, target population, and pricing) that were consistently available across hospitals. The last search was conducted on November 15, 2023.

XSL•FO
RenderX

Data Privacy and Security

To protect participant confidentiality, personally identifiable information was neither collected nor retained. All interview audio recordings and transcripts were anonymized by assigning unique study codes. Documentary data obtained from WeChat platforms were also screened to remove any patient-identifiable details. All data were stored on encrypted, password-protected institutional servers, with access strictly limited to authorized members of the research team. In compliance with institutional policies, research data will be retained for 5 years after study completion and then permanently deleted.

Data Analysis

All the interview records were transcribed and stored in a Microsoft Excel–based tool. The thematic analysis approach was applied to the data analysis in steps, including data familiarization, coding, identification of candidate themes, review and revision of themes, definition and naming of themes, and analysis and interpretation of patterns across the data. Two researchers first completed theme identification separately to gain a preliminary understanding of the target patients, value offerings, and service provisions of TCM internet hospitals. Theme saturation was reached through an interactive coding process. Then, their results were compared to examine similarities and differences. If any inconsistency was present, a triangular test was used by 2 other researchers to analyze the differences and decide the definition and naming of the themes.

Besides, thematic analyses were also applied to documentary data extracted from WeChat interfaces. An inductive approach was applied to analyze the health services of TCM internet hospitals. Two independent reviewers scrutinized the data to identify common topics of the target population, service items, service process, and service guidelines. Then, the topics were coded. Patterns of the codes were identified, and themes were generated. Another researcher reviewed the codes and themes and corrected them if feasible. Finally, documentary data were triangulated with interviews based on the review and discussions by all the researchers together to reach an agreement on the final vision of themes.

Ethical Considerations

This study received ethical approval from the University of Macau Ethics Committee (approval: SSHRE21-APP027-ICMS). All interview participants provided written informed consent prior to participation. No financial or material compensation was provided to the interview participants, as their participation was entirely voluntary and based on their professional interest in the research topic. Data were anonymized, securely stored, and accessible only to authorized researchers. Documentary data were obtained exclusively from publicly available institutional platforms and did not involve identifiable personal information.

Results

Characteristics of TCM Internet Hospitals and Participants

A total of 7 TCM internet hospitals were included in this study (Table 1), including 4 (57%) comprehensive TCM hospitals, 2 (29%) integrated TCM and western medicine hospitals, and 1 (14%) specialized TCM hospital. All the internet hospitals were Grade A tertiary hospitals, which is the highest in the Chinese hospital grading system and is scored depending on the establishment of specialty departments, equipment of health professionals and facilities, patient flow, accessibility of beds, and scientific research achievements [24]. From the 7 TCM internet hospitals, 6 (86%) were nonprofit, while 1 (14%) was for profit. The hospital with the largest number of beds was the First Affiliated Hospital of Guangzhou University of Chinese Medicine (2200 beds), which was followed by Guangdong Provincial Hospital of Chinese Medicine (2089 beds) and Guangdong Second TCM Hospital (1490 beds).

XSL•FO
RenderX

Table . Characteristics of sample traditional Chinese medicine (TCM) internet hospitals (N=7).

	Name of physical hospitals	Name of internet hospitals	Types of hospitals	Grades	Business nature	Beds
1	The First Affiliated Hospital of Guangzhou University of Chinese Medicine	Internet Hospital–The First Affiliated Hospital of Guangzhou University of Chinese Medicine	Comprehensive TCM hospital	Grade A tertiary hospital	Nonprofit	2200
2	Guangdong Second TCM Hospital	Internet Hospital–Guangdong Second TCM Hospital	Comprehensive TCM hospital	Grade A tertiary hospital	Nonprofit	1490
3	Guangdong Provincial Hospital of Chinese Medicine[a]	Internet Hospital–Guangdong Provincial Hospital of Chinese Medicine	Comprehensive TCM hospital	Grade A tertiary hospital	Nonprofit	2089
4	The Third Affiliated Hospital of Guangzhou University of Chinese Medicine	Internet Hospital–The Third Affiliated Hospital of Guangzhou University of Chinese Medicine	Specialized TCM hospital	Grade A tertiary hospital	Nonprofit	800
5	Integrated Hospital of TCM, Southern Medical University	Internet Hospital–Integrated Hospital of TCM, Southern Medical University	Integrated TCM and western medicine hospital	Grade A tertiary hospital	Nonprofit	800
6	Clifford Hospital	Internet Hospital–Clifford Hospital	Integrated TCM and western medicine hospital	Grade A tertiary hospital	For profit	1352
7	The Affiliated TCM Hospital of Guangzhou Medical University	Internet Hospital–The Affiliated TCM Hospital of Guangzhou Medical University	Comprehensive TCM hospital	Grade A tertiary hospital	Nonprofit	722

[a]Guangdong Provincial Hospital of Chinese Medicine has 3 sites. They are the main site of Guangdong Provincial Hospital of Chinese Medicine, University Town Hospital of Guangdong Provincial Hospital of Chinese Medicine, and Ersha Island Hospital of Guangdong Provincial Hospital of Chinese Medicine. The Internet Hospital–Guangdong Provincial Hospital of Chinese Medicine is the online platform for all the 3 physical sites.

A total of 14 participants took part in the interviews, and their characteristics are presented in Table 2. The specialties of interviewees were various, including cardiologists, pulmonologists, orthopedists, gastroenterologists, gynecologists, obstetricians, physiatrists, and dermatologists. From the 14 participants, 8 (57%) were females, 2 (14%) had master's degrees, and 12 (86%) had doctoral degrees. The average clinical experience was 9.5 (SD 6.5109) years, and the average experience in providing web-based services was 4.3 (SD 1.1605) years.

XSL•FO
RenderX

Table . Characteristics of doctors (n=14).

Code	Interviewees[a]	Male or female	Education	Working in hospitals (y)	Working in internet hospitals (y)
1	Cardiologist A	Male	Master's	8	4
2	Cardiologist B	Female	Doctor	7	4
3	Pulmonologist A	Male	Doctor	5	5
4	Pulmonologist B	Female	Master's	7	3
5	Orthopedist A	Male	Doctor	4	4
6	Orthopedist B	Male	Doctor	12	5
7	Gastroenterologist A	Male	Doctor	20	6
8	Gastroenterologist B	Male	Doctor	4	4
9	Gynecologist A	Female	Doctor	22	6
10	Gynecologist B	Female	Doctor	7	4
11	Obstetrician A	Female	Doctor	22	6
12	Obstetrician B	Female	Doctor	4	4
13	Physiatrist	Female	Doctor	6	3
14	Dermatologist	Female	Doctor	5	2

[a] Labels A and B are used to distinguish between different interview participants, representing different interviewees in the same clinical department.

Target Patients and Value Offerings of TCM Internet Hospitals

Target Patients of TCM Internet Hospitals

Return patients with chronic or stable conditions emerged as the primary users of TCM internet hospitals. These patients required follow-up care after initial visits to the physical hospitals. As 1 participant noted:

> *If they are not return patients, even if they have made an appointment at the internet hospital, we generally cannot provide telemedicine to them because we do not know their conditions. We would suggest them to have an outpatient visit in the first place.* [Interviewer 3]

This reflects the regulatory restriction that web-based consultations are limited to patients with existing medical records, ensuring safety and continuity of care.

A key functional need for these return patients was prescription refills and access to hospital preparations without repeat physical visits, as highlighted below:

> *Many patients come here (the TCM internet hospital) to refill their prescriptions and buy hospital preparations. The prescriptions can be mailed to the patients. So the patients do not have to come back to the physical hospitals for pick-up.* [Interviewer 4]

This practice reduces the travel burden for patients and improves adherence to long-term treatment plans.

Beyond routine follow-up for common conditions, TCM internet hospitals were also used for postoperative recovery management for patients with tumors:

> *Aside from patients having colds, cough, hypertension and diabetes, patients with tumours are also accessing TCM internet hospitals. For example, patients who have undergone gastric cancer surgery usually require a long-term follow-up after the discharge. We (doctors of TCM internet hospitals) can adjust the prescriptions depending on patients' feedback of eat and sleep conditions.* [Interviewer 8]

This illustrates an emerging but important role of TCM internet hospitals in supporting long-term recovery and personalized rehabilitation care.

Consultations for disease prevention and health management represented another group of target users, reflecting growing public awareness of proactive health maintenance. As described below:

> *There is a large number of people seeking for advices on disease prevention and helaht proection. Nowadays, people generally have good health awareness. Some people accessed the web-based conslutaiton service for themselves; some for family members. Issues of consultaiotns are usually about keeping diets; and managing osteoporosis, cerebral infarction and stroke in the elderly. We are receiving more and more consulotations on strategies of TCM helath protection.* [Interviewer 5]

In addition, younger users, particularly female students, sought advice on gynecological health and sensitive issues, demonstrating the privacy advantage of web-based consultations:

> *Many young patients, particularly female students, access the web-based service for gynaecological issues, such as abnormal menstruation and excessive leukorrhea. Some people seek for advice on*

XSL•FO
RenderX

> *management of unplanned pregnancies. They must be students. They do not know how to deal with accidental pregnancy and are worried about the impact of potential measures on future fertility, so they consult via internet hospital platforms.* [Interviewee 9]

In summary, TCM internet hospitals primarily serve return patients with chronic or stable conditions, individuals seeking preventive care, and younger users requiring sensitive consultations. This reflects a dual role in both clinical follow-up and health management.

Value Offerings of TCM Internet Hospitals

One of the most prominent value offerings was improving patient compliance with follow-up care. Remote follow-ups can maintain treatment continuity and reduce the risk of incomplete care. By reducing geographic and time barriers, TCM internet hospitals facilitated more consistent follow-up:

> *Many patients with bronchitis and coughs are recommended to come back to hospitals for subsequent reviews after a course of treatment which usually takes about 5 or 6 days. But, some patients do not come back to physical hospitals for follow-ups once their symptoms have eased off, because they want to save efforts from accessing offline service…. Nowadays, web-based service can support the consultations on patients' phlegm, voice, frequency of coughs, and if they have sore throat. In this way, doctors can detect the recovery status of these patients.* [Interviewer 3]

Similarly, privacy-sensitive conditions were more likely to be managed through web-based services, which further promoted adherence:

> *Some patients with gynaecological diseases, mainly those who are not in serious conditions, are prone to choose internet hospitals for follow-up service. Web-based service is timesaving. Many patients also consider Internet hospitals as a platform good for privacy protection.* [Interviewee 10]

Another key value offering was TCM promotion and patient education. Web-based services enhanced accessibility to renowned TCM practitioners, thereby raising public awareness:

> *Compared to western medicine practitioners, the number of registered TCM practitioners is smaller, let alone the renowned TCM practitioners. TCM internet hospitals are equipped with famous TCM practitioners, which advanced the access and quality of TCM service.* [Interviewer 2]

Moreover, web-based consultations shifted patients from passive treatment-seeking to proactive health preservation:

> *In the past years, people are prone to have self-medications when catching colds and coughs with mild symptoms, rather than visiting doctors in clinics. But, health literacy of general population have optimized. With the increasing health awareness, more and more people pay great attention to health preservation. It is very convenient and low costing to find a TCM practitioner to advice on health preservation through WeChat.* [Interviewee 9]

Collectively, TCM internet hospitals deliver value by improving compliance with long-term care, enhancing privacy for sensitive conditions, and serving as a key channel for TCM health promotion and preventive care.

Health Services of TCM Internet Hospitals

As indicated in Table 3, web-based service items provided by the enrolled hospitals can be divided into 2 categories and 6 subcategories according to the service content. The first category was functional service, which comprised telemedicine, telepharmacy, telenursing, web-based health consultations, and convenient service. All 7 internet hospitals provided functional service. But these hospitals performed differently in the 5 subcategories of functional service. Specifically, the number of sample hospitals offering telemedicine (7/7, 100%), telepharmacy (4/7, 57%), web-based health consultations (3/7, 43%), convenient service (3/7, 43%), and telenursing (2/7, 29%) was in a decreasing trend. The second category of service items was TCM specialty service, which was based on TCM theory and practice. TCM specialty service was delivered by integrating web-based and offline approaches in 3 (43%) hospitals.

XSL•FO
RenderX

Table . Service items provided by traditional Chinese medicine (TCM) internet hospitals.

	Functional service					TCM specialty service
	Telemedicine	Telepharmacy	Telenursing	Web-based health consultations	Convenient service	
Internet Hospital–The First Affiliated Hospital of Guangzhou University of Chinese Medicine	✓				✓	✓
Internet Hospital–Guangdong Second TCM Hospital	✓					
Internet Hospital–Guangdong Provincial Hospital of Chinese Medicine	✓	✓				
Internet Hospital–The Third Affiliated Hospital of Guangzhou University of Chinese Medicine	✓	✓		✓		✓
Internet Hospital–Integrated Hospital of TCM, Southern Medical University	✓	✓	✓	✓	✓	✓
Internet Hospital–Clifford Hospital	✓	✓	✓	✓	✓	
Internet Hospital–The Affiliated TCM Hospital of Guangzhou Medical University	✓					

Functional Service

Telemedicine was the internet hospitals' core functional service, with the characteristics shown in Table 4. The target patients of telemedicine were mainly patients who had subsequent visits after primary offline medical service. Names of telemedicine in the sample hospitals were various and can be classified into 2 streams. One stream is determining the name by combining terms of *internet* and *medical service* such as *web-based clinic* (2/7, 29%) and *web-based diagnosis and treatment* (1/7, 14%). The other stream (2/7, 29%) used the names consistent with traditional offline medical services provided in the physical hospitals. The examples of this stream were *expert clinic* (1/7, 14%) and *general outpatient clinic* (1/7, 14%). Pictures and texts were the principal media of practicing telemedicine and were used by all 7 hospitals. But, telephone consultation was feasible in 2 (29%) hospitals, and video consultation was applied in 1 (14%) hospital. Service price was uniform in 3 (43%) hospitals, whereas it was positively associated with doctors' professional titles in the other 3 (43%) hospitals. Only 2 (29%) TCM internet hospitals currently support payment through China's medical insurance program.

XSL•FO
RenderX

Table . Telemedicine provided by traditional Chinese medicine (TCM) internet hospitals.

Name of the internet hospital	Name of telemedicine	Content of telemedicine	Service media of telemedicine	Service prices of telemedicine	Service guidelines of telemedicine	Payment methods of telemedicine
Internet Hospital–The First Affiliated Hospital of Guangzhou University of Chinese Medicine	Web-based clinic	Prescription refills for patients with subsequent visits, and web-based consultations on noncritical diseases	• Picture and text	• RMB ¥10[a] per web-based visit	—[b]	• Patient self-payment
Internet Hospital–Guangdong Second TCM Hospital	Web-based diagnosis and treatment	Interpretation of medical reports, prescription refills for patients with chronic disease and subsequent visits, web-based consultations on noncritical diseases	• Picture and text	• RMB ¥10 per web-based visit	• Doctors respond to consultation questions of patients within 48 hours.	• Patient self-payment
Internet Hospital–Guangdong Provincial Hospital of Chinese Medicine	General outpatient clinic	Follow-up of patients who primary visits to the physical hospital	• Picture and text • Telephone	• Chief doctor: RMB ¥30 per web-based visit • Associated chief doctor: RMB ¥20 per web-based visit • Attending doctor RMB ¥10 per web-based visit	• Consultations in picture and text allow unlimited communications between the doctor and the patient within 24 hours. • Telephone consultation is limited to 15 minutes.	• Guangzhou City General Medical Insurance • Mente Medical Insurance[c] Group 1
Internet Hospital–The Third Affiliated Hospital of Guangzhou University of Chinese Medicine	Convenient medical service	Interpretation of medical reports, prescription refills for patients with chronic disease and subsequent visits, and web-based consultations on noncritical diseases	• Picture and text	• RMB ¥10 per web-based visit	• Doctors respond to questions of patients within 48 hours.	• Patient self-payment
Internet Hospital–Integrated Hospital of TCM, Southern Medical University	Expert clinic	Interpretation of medical reports, prescription refills for patients with subsequent visits, and web-based consultations on noncritical diseases	• Picture and text	• Chief doctor: RMB ¥60 per web-based visit • Associated chief doctor: RMB ¥40 per web-based visit • Attending doctor: RMB ¥20 per web-based visit	• Patients can send 10 messages as the maximum within 48 hours to communicate with the doctor.	• Patient self-payment

XSL•FO
RenderX

Name of the internet hospital	Name of telemedicine	Content of telemedicine	Service media of telemedicine	Service prices of telemedicine	Service guidelines of telemedicine	Payment methods of telemedicine
Internet Hospital-Clifford Hospital	Medical insurance covering follow-ups and Prescription Refills	Prescription refills and web-based consultations on chronic diseases and noncritical conditions	• Picture and text	• Free for picture and text consultation • Web-based consultations from chief doctor: RMB ¥40 per web-based visit • Associated chief doctor: RMB ¥30 per web-based visit • Attending doctor: RMB ¥20 per web-based visit	—	• Patient self-payment • Guangzhou City General Medical Insurance for designated hospitals • Mente Medical Insurance for 25 diseases in Group 1 • Mente Medical Insurance for 28 diseases in Group 2 in designated hospitals • Mente Medical Insurance for psychiatric diseases
Internet Hospital–The Affiliated TCM Hospital of Guangzhou Medical University	Web-based clinic	Web-based consultations and chronic disease management	• Picture and text • Telephone and video	—	• The hospital arranges shifts for doctors.	• Patient self-payment

[a]A currency exchange rate of RMB ¥1=US $0.1409 is applicable.

[b]—: missing information.

[c]Mente Medical Insurance: a medical insurance program for chronic diseases (coronary heart disease, hypertension, diabetes, epilepsy, and cancer.) requiring long-term and expensive outpatient treatment. The chronic diseases covered by Mente Medical Insurance are classified into 2 groups depending on disease severity. Group 1 is common chronic disease such as, hyperthyroidism, and tuberculosis; and Group 2 was composed of severer diseases, for example leukaemia and renal failure.

Telepharmacy aimed to guarantee drug supply and instructed patients on rational medication use. It consisted of "prescription refills," "self-service drug purchase," and "consultation on drug use." Refilled prescriptions were available as hospital preparations and ointments. These medications can be delivered to patients either through patients collecting drugs from physical hospitals or by a third-party mailer.

Telenursing comprised services of *nursing clinic* and *home-based nursing.* Professional instructions on dealing with common traumas and operations of medical devices can be accessed in *nursing clinic*. *Home-based nursing* supported appointments from patients after surgery and seeking out-reaching rehabilitation services. Web-based health consultation service was health promotion principally focusing on miscellaneous health-related questions. Diagnoses and medical orders were out of the functions of this service item.

Internet hospitals incorporated convenient services to improve patients' experience during medical services. Differences were in the functions of this service item. Convenient service of the 2 public nonprofit hospitals was mainly about making appointments for offline outpatient visits and medical tests and uploading test reports. Aside from these functions, the other private and for-profit hospitals have an array of different functions, including reporting inpatient medical records, processing of medical insurance, and complaints and suggestions.

TCM Specialty Service

TCM specialty services were administered by linking online and offline services. *Tianzhi* (crude herb moxibustion) was one service item of the First Affiliated Hospital of Guangzhou University of Chinese Medicine. Web-based service of *Tianzhi* provided advice on its target population, warnings, prices, and timetables. Processing appointments and payments are also components of the web-based service. The practice of *Tianzhi* will be carried out in the outpatient clinic offline after confirmation of web-based appointments and payments. *Patient-centered* prevention, treatment, and health management were offered by Internet Hospital–The Third Affiliated Hospital of Guangzhou University of Chinese Medicine. This individualized health service is mainly led by renowned TCM practitioners who can identify markers of the onset of disease in a timely manner and develop TCM prevention and treatment programs. The stroke prevention and rehabilitation service offered by Internet Hospital–Integrated Hospital of TCM, Southern Medical University, contained web-based consultation and appointments for acupunctures and massages.

XSL•FO
RenderX

Comparative Patterns Across Hospital Types and Affiliations

Notable differences were observed across hospital types and ownership. Public nonprofit hospitals primarily focused on providing functional services such as telemedicine and telepharmacy, while their convenient services were limited to basic appointment bookings and report uploads. In contrast, the private for-profit hospital offered a broader range of convenient services, including inpatient record access and medical insurance processing, likely reflecting its market-oriented operational model. Furthermore, comprehensive TCM hospitals tended to deliver a wider variety of web-based services, whereas the specialized TCM hospital emphasized individualized rehabilitation-oriented services. Despite these differences, all hospitals shared common value offerings related to improving patient compliance and promoting TCM awareness.

Discussion

Principal Findings

In this study, we conducted an in-depth qualitative analysis of the early-stage digital adaptation in TCM internet hospital health services regarding target patient, value offering, and service provision. All the sampled TCM internet hospitals were established based on Grade A tertiary brick-and-mortar hospitals supported by sufficient health resources. Service content was categorized as functional service (telemedicine, telepharmacy, telenursing, web-based health consultations, and convenient service) and TCM specialty service. All these categories were at the early stage of development. Due to limitations in TCM therapeutic items, TCM specialty services are not prominent in health services. TCM internet hospital has not absorbed the benefits of digitalization from web-based platforms. Current health services have displayed 2 types of value offerings: TCM promotion and optimization of patients' compliance.

Although this study initially used an *innovation* framing, the empirical findings largely describe early-stage digital adaptation characterized by follow-up consultations and prescription refills. Innovation in this context lies primarily in the reconfiguration of service processes, such as integrating online-offline pathways, linking pharmacy and nursing services, and embedding convenience functions, rather than the introduction of novel digital therapeutics. We therefore present the results as evidence of incremental service evolution with emerging innovative potential, rather than fully realized innovations.

Practicality of TCM Internet Hospitals

All 7 TCM internet hospitals were established based on physical tertiary hospitals in Grade A, in spite of being for profit or nonprofit. At present, Grade A tertiary hospitals in China have noteworthy advantages in terms of medical equipment, doctors, pharmacists, nurses, and patient flow when compared to hospitals in other grades [25,26]. Therefore, the development of internet hospitals is characterized by relying on sufficient health resources and a large population base.

The overall health services of TCM internet hospitals are at the exploration stage. Telemedicine lacks the initiation of patient-centered interventions and digital therapeutics. The main contents of telepharmacy and telenursing were primary service items, including prescription refills, instructions on managing common traumas, and postsurgery rehabilitation. Convenient service is designed to optimize patients' experience of web-based medical visits. However, this service was provided by no more than half of the surveyed hospitals. Overall, the limited functional service indicates the significant gap between the current status and the ultimate goal of enhancing the accessibility of health resources and improving patients' medical experience [27]. Besides, the advantages of digital health are not evident in current web-based health services. Web-based platforms provide the feasibility of medical data storage and management in large quantities and high efficiency, and data sufficiency is the prior condition of patient-centered treatment.

As a part of functional services, telemedicine was mainly for follow-up monitoring on patients' conditions, interpretation of medical reports, and making subsequent prescriptions. However, the performance of telemedicine may be impacted by web-based modes. Individualized care is emphasized by TCM health services, in which diagnoses were usually performed depending on 4 approaches: *Wang* (observation), *Wén* (auscultation and olfaction), *Wèn* (inquiry), and *Qie* (pulse diagnosis) [28]. However, the general media of telemedicine were pictures and texts. Thus, the *Wén* (auscultation and olfaction) and *Qie* (pulse diagnoses) were restricted. *Wang* (observation) can be influenced by the discretions of picture quality.

Obviously, polarization was in the characteristics of telemedicine, including service prices and payment methods in all the 7 TCM internet hospitals. It is challenging to manage telemedicine service prices, with various service items [29]. From the 7 TCM internet hospitals, 3 (43%) hospitals charged patients RMB ¥10 uniformly, despite different service items and professionals of doctors on shift. In contrast, the other 3 (43%) TCM internet hospitals determined the service prices based on doctors' titles (chief, associated chief, and attending doctors). The uniform service fees of RMB ¥10 cannot afford frequent manpower- and time-consuming web-based service and technical support for platforms, which can impact the doctors' proactivity in participation [30]. A majority (5/7, 71%) of the internet hospitals did not support National Medical Insurance, a potential barrier for patients to web-based services. Incorporating web-based diagnosis and treatment items with high prevalence into the payment scope of the National Medical Insurance Catalog helps improve general service quality, thereby satisfying patients' growing needs in the health care [31].

In addition, hospital ownership and type influenced the scope of web-based service delivery. Public nonprofit hospitals provided more standardized but narrower functional services, whereas the private for-profit hospital adopted a more market-driven approach by expanding convenient services. Comprehensive hospitals integrated broader service items compared to specialized hospitals, which focused more on niche rehabilitation services. These nuanced patterns suggest that organizational structure and operational models may shape service early-stage digital adaptation trajectories in TCM internet hospitals.

XSL•FO
RenderX

TCM Specialty Service

Offline TCM services are various and involve pharmacological interventions, including Chinese patient medicine, Chinese herbal piece, Chinese herbal medicine formula granule, and Chinese herbs; and nondrug interventions, including dietary therapy, acupuncture, moxibustion, massage, cupping, scraping, and therapeutic mind or body practices [32,33]. TCM has been approved as effective in treating chronic diseases, including tumors [34], diabetes [35], and cardiovascular diseases [36]. Clinical evidence of TCM has also been found in treating skin diseases and infertility [37,38]. China is the sole country that provides Western medicine and TCM services at hospitals and medical institutions at different levels [39]. However, judging from the use of health resource consumption in TCM hospitals, the role of TCM is still unclear due to the increasing influence of Western medicine. The TCM service system requires improvement [40].

Internet hospitals can contribute to the refinement of the TCM service system. With the rapid development and application of information technologies such as the internet, big data, artificial intelligence (AI), and the web of things in the health system, digitalization is accelerating in the TCM [41]. In October 2019, the State Council issued the "Opinions on Promoting the Inheritance and Innovative Development of Traditional Chinese Medicine," which encourages the establishment of TCM internet hospitals by relying on physical hospitals, development of intelligent auxiliary systems for disease diagnosis and treatment, implementation of telemedicine, and integration of web-based and offline TCM services [42]. Currently, clinical decision support systems have been upgraded with AI techniques. Examples of the systems are pulse diagnosis and tongue diagnosis systems. They are gradually applied to web-based consultations to assist TCM practitioners in understanding patients' health conditions [43,44].

In the meantime, TCM specialty service is amenable to being the characteristic service of TCM internet hospitals. However, its performance is not prominent. Items of TCM specialty service were *Tianzhi* (crude herb moxibustion) [45], *Zhiweibing* (preventive treatment of disease) [46], and stroke rehabilitation. Roles of the items were mainly health consultations and preservations, with the absence of therapeutics. TCM therapeutics, such as phlegm scraping, massage, and acupuncture [8], have been widely applied offline but are augmented by *internet +*.

Despite the unique modalities of TCM, most internet hospital functions remain generic, centered on teleconsultations and prescription refills. Distinctive services (eg, acupuncture, tuina, moxibustion, and rehabilitation) were rarely offered web-based. This mismatch reflects three constraints: (1) technical limitations of remote media, particularly in replicating pulse-taking and olfactory diagnosis; (2) regulatory and reimbursement restrictions, which limit insurance coverage for specialty treatments; and (3) the absence of standardized digital workflows and quality control mechanisms. Addressing these barriers requires a staged roadmap: in the short term, establishing a hybrid "web-based assessment–offline treatment" pathways; in the medium term, expanding reimbursement and pricing policies to include distinctive TCM modalities; and in the long term, integrating AI-assisted tongue or pulse diagnostics and digital biomarkers into web-based TCM platforms.

Potential Orientations to Future Innovations

According to the *Internet Hospital Management Measures (Trial)* policy [47], telemedicine provided by internet hospitals should be restricted to patients with follow-up visits for common and chronic diseases. In addition to this basic function, TCM internet hospitals are carrying out items of TCM health consultations and preservations. These items are critical for fulfilling the general population's health needs, which is consistent with the policy of *Outline of the 14th Five-Year Plan for National Economic and Social Development of the People's Republic of China and the vision 2035* [48]. Meanwhile, TCM has been well applied for rehabilitation in the whole health system, especially in primary care [49]. Therefore, the additive service items offered by TCM internet hospitals are an advance in both web-based and TCM services. In addition, TCM internet hospitals are indicated as optimal platforms for people in need of disease prevention and health protection. In the overall hierarchical medical system, TCM internet hospitals can provide first-line health management for people before they get ill and access primary services with definite symptoms. On this note, TCM internet hospitals contribute to popularizing and developing TCM.

Compliance is a noteworthy factor affecting the quality and efficiency of medical services. Inadequate compliance is significantly associated with increased mortality, morbidity, and medical costs [50-52]. Education, reminders, and financial incentives to patients can improve their compliance [53]. Common factors impeding medical services are high costs, limited access, and long duration [22], all present in offline service scenarios. TCM internet hospitals can remove these obstacles, thereby improving medical compliance. The findings of interviews have reflected this proposition.

In addition, the collection and adoption of digital biomarkers are potentially beneficial from the platform of TCM internet hospitals. Biomarkers are objective, quantifiable medical signs that relate to medical signs, symptoms, surrogate end points, or clinical end points [54]. RockHealth defined digital biomarkers as "consumer-generated physiological and behavioral measures collected through connected digital tools that can be used to explain, influence, and predict health-related outcomes" [55]. In the era of digital health, digital biomarkers enable new value chains on health data for health intervention [56]. Developing and adopting safe and effective digital biomarkers to improve patient outcomes in clinical research and routine patient care has been a new trend [57]. As an easily scalable, noninvasive complement, a smartphone can provide readily attainable digital biomarkers of health intervention, such as diabetes and Alzheimer disease [58,59]. As TCM specialty services are incorporated into TCM internet hospitals for health consultations and preservations, they can potentially become a novel path in collecting digital biomarkers, such as user-generated measurements. These biomarkers are sources of evidence and valid decision-making. In this study, TCM internet hospitals primarily targeted chronic disease follow-ups and health

XSL•FO
RenderX

preservation consultations. These service patterns provide practical application scenarios where AI-driven decision support and digital biomarkers could enhance long-term patient monitoring and personalized TCM interventions, aligning with the observed patient needs.

Comparison With Prior Research and Theoretical Implications

Our findings corroborate prior studies indicating that internet hospitals improve accessibility and continuity of care for patients with chronic and stable conditions [2,4]. However, unlike previous research primarily focused on Western medicine internet hospitals, we identified an additional value dimension in promoting TCM literacy and health preservation practices, aligning with studies that emphasize TCM's preventive care philosophy [14,15]. Furthermore, while earlier work suggested that digital platforms primarily address service inefficiencies [11], our study demonstrates their potential to enhance patient compliance and extend TCM specialty services such as crude herb moxibustion and stroke rehabilitation.

From a theoretical perspective, the findings enrich the *value proposition* element of the business model framework by Johnson et al [22]. Specifically, TCM internet hospitals create a dual-layered value: they not only address patients' immediate medical needs but also extend cultural and educational functions unique to TCM. This differs from the more transactional value propositions of typical internet hospitals, suggesting that TCM internet hospitals may operate under a hybrid value model combining clinical and cultural components.

In contrast to prior work highlighting limited integration between web-based and offline services [16], our study reveals emerging hybrid models that link web-based consultations with offline therapeutic interventions. This finding extends the understanding of digital health adoption by illustrating how TCM internet hospitals bridge digital health technologies with traditional therapeutic modalities.

Implications for Policy and Practice

Overall, the development of web-based TCM health services is in the early stage with functional service occupying a large proportion. There was a remarkable gap between the current health service and an advanced model with prominent TCM therapeutics, generalized convenient service, and full integration of digital technology. The value proposition of TCM internet hospitals is yet to be excavated, aside from current assistance with the promotion of TCM and patients' compliance. Challenges in developing TCM health services are various, including heterogeneous qualifications of practitioners, difficulties in obtaining patients' medical records, lagging establishment of guidelines and legal terms for web-based practice, and insufficient programs for privacy protection [60]. Several suggestions can be made to overcome these challenges. First, more TCM therapeutics are expected, which include acupuncture, cupping, and scraping. It is necessary to formulate and implement nationwide guidelines for web-based medical practice along with the expansion of service items. Second, it is important to enhance convenient service to refine the patients' experience. Third, schemes of rationing internet service prices and health insurance coverage of service items are critical for upgrading health care for patients and the proactivity of doctors' participation in web-based services. Finally, it is suggested to make use of the digital benefits of web-based platforms and encourage TCM internet hospitals to develop and adopt safe and effective digital biomarkers to improve patient outcomes. To ensure privacy and autonomy, data use agreements for digital biomarkers should contain clear statements on data use conditions, especially for near-continuous data, such as movement, voice, and other sensitive biometric states.

However, further innovation is constrained by uneven IT infrastructure, regulatory uncertainties, and limited insurance coverage for web-based TCM services. To address these barriers, policymakers should promote standardized IT infrastructure and data-sharing protocols, improve regulatory clarity for web-based TCM consultations, and expand insurance coverage for commonly used TCM digital services. In parallel, health care institutions should develop clinical guidelines for digital TCM care to enhance provider confidence and patient trust. These coordinated efforts could create a more supportive ecosystem for innovation and sustainable scaling of TCM internet hospitals.

In practice, short-term priorities include standardizing telemedicine workflows and improving IT interoperability at the hospital level. Medium- to long-term priorities require coordinated policy action, such as expanding insurance coverage for web-based TCM services and clarifying regulatory frameworks for digital TCM practice. These recommendations are particularly urgent in the context of China's ongoing digital health reforms, including the nationwide expansion of internet hospital regulation and health insurance integration pilots. Aligning the TCM internet hospitals with these reforms—through standardized IT systems, clearer liability frameworks, and targeted reimbursement policies—would ensure they are not marginalized but become an integral part of the evolving digital health ecosystem.

Strengths and Limitations

To the best of our knowledge, this study is the first effort to investigate early-stage digital adaptation in health services of TCM internet hospitals. We used a combination of several qualitative methodologies, including case studies, stakeholder interviews, and document analysis, to fulfill the aim. However, there are several research limitations that cannot be ignored. First, a key limitation is that all participating hospitals were located in Guangzhou. While Guangzhou represents an early and advanced hub for TCM digital services, findings may not fully reflect practices in less developed provinces or rural settings. Accordingly, the results should be interpreted as contextual and transferable insights rather than nationally representative estimates. Future research should extend to multi-provincial and urban–rural comparative samples. Second, this study investigated early-stage digital adaptations in TCM health services by interviewing service practitioners. Another limitation is the absence of patient perspectives, which restricts the ability to fully assess service acceptability and satisfaction. To address this, future studies will incorporate patient satisfaction surveys, semistructured interviews, and usage data

XSL•FO
RenderX

(such as revisit rates, prescription completion, and delivery timeliness), enabling a more balanced evaluation of TCM internet hospital services. Thirdly, this study lacked analysis on the service quality and cost-effectiveness of TCM internet hospitals. Future research should conduct extensive quantitative analysis to assess the service quality and cost-effectiveness of TCM internet hospitals and compare them to traditional TCM hospitals. Fourth, this study relied on self-reported interviews with clinicians, which may be subject to recall and social desirability bias. The findings primarily reflect providers' perspectives and may not fully capture objective service quality. Future research could triangulate interview data with patient feedback, service usage metrics, and quantitative quality indicators to enhance validity. Finally, future research should consider investigating the integration of AI and big data analytics in TCM internet hospital services.

Overall, this study advances the current understanding of digital health services by situating TCM internet hospitals within both the local cultural context and broader digital health transformation. By explicitly linking patient compliance, TCM promotion, and hybrid service delivery models, our findings extend prior work and provide empirical evidence to refine theoretical models of digital health innovation, particularly in resource-integrated traditional medicine settings. While the findings reflect the unique cultural and policy context of China's health care system, certain insights—such as the role of internet hospitals in chronic disease follow-up and preventive health consultations—may be transferable to other countries exploring the digitalization of traditional or complementary medicine. However, the scalability of TCM internet hospitals in other contexts would depend on local regulatory frameworks, health insurance structures, and the cultural acceptance of traditional medicine. Future comparative studies are needed to assess the adaptability of this model in different health care systems.

Conclusions

This study positions TCM internet hospitals as a potentially disruptive force in the early stage of technology adoption and diffusion. While they currently focus on serving returning patients with chronic or stable conditions and promoting TCM health preservation, the innovative potential of TCM-specific services remains underexplored. To truly become a positive disruption that strengthens TCM specialty services and enhances patient accessibility, TCM internet hospitals must transition from the early adoption phase to the early majority stage.

Our findings enrich the understanding of how traditional medicine integrates with digital health, highlighting the hybrid value proposition of TCM internet hospitals that combines clinical care with cultural and educational functions. They also reveal systemic barriers—uneven IT infrastructure, regulatory uncertainties, and limited insurance coverage—that constrain innovation diffusion, emphasizing the need for coordinated policy and institutional reforms. For stakeholders, these insights provide a roadmap to better harness digital health innovation in TCM. Future research should triangulate provider and patient perspectives with platform data to validate service quality and explore the integration of AI-driven decision support and digital biomarkers to personalize TCM care and accelerate sustainable scaling.

Acknowledgments

The authors would like to thank all of the interviewees for their participation in this study. The authors also thank for the kind support from the Institute of Chinese Medical Sciences at the University of Macau. This research was supported by funding from the project of Guangdong Basic and Applied Basic Research Foundation (grants 2022A1515110662 and 2022A1515220148), the project of Guangzhou Philosophies and Social Sciences (grant 2023GZGJ16), the Humanities and Social Sciences Cultivation Project of Guangzhou University of Chinese Medicine (grant 2024ZDPY02), the "Solid Foundation" Project of Guangzhou University of Chinese Medicine (grant GZY2025GB0306), and University-Hospital Joint Fund Project of Guangzhou University of Chinese Medicine (grants GZYZS2024G16 and GZYZS2024G34).

Data Availability

The datasets generated for this study are available from the corresponding authors upon reasonable request.

Authors' Contributions

YW and MS are co-first authors. YW, MS, YB, HH, and YL conceptualized and designed the study. YW, MS, ZP, and DY collected the data and conducted the data analysis and data interpretation. All authors contributed to the writing of the article. All authors read and approved the final manuscript. YB (bianyingumac@126.com), HH (haohu@um.edu.mo), and YL (yunfeng.lai@hotmail.com) are co-corresponding authors for this article.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Informed consent.
[DOCX File, 29 KB - formative_v9i1e77686_app1.docx]

XSL•FO
RenderX

Multimedia Appendix 2
Interview schedule.
[DOCX File, 23 KB - formative_v9i1e77686_app2.docx]

References

1. Liang F, Yang X, Peng W, et al. Applications of digital health approaches for cardiometabolic diseases prevention and management in the Western Pacific region. Lancet Reg Health West Pac 2024 Feb;43:100817. [doi: 10.1016/j.lanwpc.2023.100817] [Medline: 38456090]
2. Tu J, Wang C, Wu S. The internet hospital: an emerging innovation in China. Lancet Glob Health 2015 Aug;3(8):e445-e446. [doi: 10.1016/S2214-109X(15)00042-X] [Medline: 26187488]
3. Gu Y, Wei D, Cao X, Zhang C, Feng X. Analysis of the reason of the internet hospital occurred in China. Chinese Health Serv Manag 2017;34(6):401-403.
4. Lai Y, Chen S, Li M, Ung COL, Hu H. Policy interventions, development trends, and service innovations of internet hospitals in China: documentary analysis and qualitative interview study. J Med Internet Res 2021 Jul 20;23(7):e22330. [doi: 10.2196/22330] [Medline: 34283025]
5. Wang L, Suo S, Li J, et al. An investigation Into traditional Chinese medicine hospitals in China: development trend and medical service innovation. Int J Health Policy Manage 2017 Jan 1;6(1):19-25. [doi: 10.15171/ijhpm.2016.72] [Medline: 28005539]
6. Tang JL, Liu BY, Ma KW. Traditional Chinese medicine. Lancet 2008 Dec 6;372(9654):1938-1940. [doi: 10.1016/S0140-6736(08)61354-9] [Medline: 18930523]
7. Tian J, Gao S. Analysis on barriers and strategies of the development of “internet+” traditional Chinese medicine treatment. Health Econ Res 2022;39(3):70-73. [doi: 10.14055/j.cnki.33-1056/f.2022.03.018]
8. Hu T, Zhou B, Ling Z, Chen Z, Chen C, Zhu J. Study on the construction status and development trend of internet traditional Chinese medicine hospitals. J Med Inform 2022;43(9):7-11. [doi: 10.3969/j.issn.1673-6036.2022.09.002]
9. Wang C, Li R, Yan K, Ma S. High quality development of traditional Chinese medicine hospitals based on the comparison of medical resources and medical service. Chin Health Qual Manag 2023;30(11):1-7. [doi: 10.13912/j.cnki.chqm.2023.30.11.01]
10. Ma H, Xiong J. Study on spatial distribution and influencing factors of service efficiency of TCM hospitals. Soft Sci Health 2023;37(11):58-62. [doi: 10.3969/j.issn.1003-2800.2023.11.011]
11. Wang X, Liu K, Zhang H, Li X, Ma Z. Research and practice on internet service mode of TCM diagnosis and treatment. Chin J Health Inform Manag 2021;18(5):626-655. [doi: 10.3969/j.issn.1672-5166.2021.05.010]
12. Zhang Y, Tian Y, Feng JJ, Liu JW, Zhang YL, Lin LK. Development situation and expert suggestion on “Internet + Traditional Chinese Medicine” in China. New Med 2024;34(5):481-486. [doi: 10.12173/j.issn.1004-5511.202310073]
13. Wang X, Chen J, Feng M, et al. Demand and influencing factors of “Internet + Traditional Chinese Medicine” home nursing service for older adult patients with chronic diseases: a mixed research perspective. Front Public Health 2023;11:1271082. [doi: 10.3389/fpubh.2023.1271082]
14. Zhang L, Zhou PY, Tang H. Practice and exploration of independent construction of internet hospitals of traditional Chinese medicine hospitals. Smart Healthc 2023;26:17-21. [doi: 10.19335/j.cnki.2096-1219.2023.26.005]
15. Qi MJ, Chen KY, Liu ZX. Online medical services by Beijing hospital of traditional Chinese medicine. Modern Hosp 2024;24(11):1764-1768. [doi: 10.3969/j.issn.1671-332X.2024.11.036]
16. Han YY, Yuan M, Jiang H, Yu J, Tang WJ. Service and practice of internet hospital established in a large public traditional Chinese medicine hospital. Chin J Med Manag Sci 2024;14(5):63-68. [doi: 10.3969/j.issn.2095-7432.2024.05.011]
17. Qi MJ, Zhang QS, Han YL, Chen KY, Dong JC. A survey on use of internet plus health care among medical staff in a grassroots traditional Chinese medicine hospital. Chin J Med Manag Sci 2023;13(4):86-90. [doi: 10.3969/j.issn.2095-7432.2023.04.015]
18. Ding YW, Zhai X, Wang QX, Tong JD, Kou JM. Status quo of the information services of mobile medical apps for traditional Chinese medicine. Chin J Med Libr Inf Sci 2022;3(1):49-55. [doi: 10.3969/j.issn.1671-3982.2022.01.006]
19. Ge Y, Yao D, Ung COL, et al. Digital medical information services delivered by pharmaceutical companies via WeChat: qualitative analytical study. J Med Internet Res 2023 Nov 17;25:e43812. [doi: 10.2196/43812] [Medline: 37976079]
20. Ding Y, Zhai X, Wang Q, Kou I, Tong J. Research on the construction of evaluation index system of information service quality of Internet traditional Chinese medicine hospital. Chin Hosp 2022;26(5):23-27. [doi: 10.19660/j.issn.1671-0592.2022.5.07]
21. Shu Y, Ning L, Chen L, Huang F, Wu W. Practice and thought on the establishment of internet hospitals by public Chinese medicine hospitals. Chin Hosp 2021;25(4):28-30. [doi: 10.19660/j.issn.1671-0592.2021.4.08]
22. Johnson MW, Christensen CM, Kagermann H. Reinventing your business model. Harvard Bus Rev. 2008. URL: https://hbr.org/2008/12/reinventing-your-business-model [accessed 2025-05-12]
23. Lai Y, Ho Q, Wang Y, et al. Challenges and strategies of developing internet hospital: combining qualitative interview and documentary research. Digit Health 2024;10:20552076241310075. [doi: 10.1177/20552076241310075] [Medline: 39741982]

XSL•FO
RenderX

24. Grade a tertiary hospital accreditatin standards. National Health Commission of the Peopole's Republic of China. 2022. URL: https://www.gov.cn/zhengce/zhengceku/2022-12/18/5732583/files/61ce29022971491bbc15afcc1d35f97d.pdf [accessed 2024-02-02]
25. Lv Q, Jiang Y, Qi J, et al. Using mobile apps for health management: a new health care mode in China. JMIR Mhealth Uhealth 2019 Jun 3;7(6):e10299. [doi: 10.2196/10299] [Medline: 31162131]
26. Zhang X, Zhang X, Yang S, Wang Y. Factors influencing residents' decision to sign with family doctors under the new health care reform in China. Int J Health Plann Manage 2019 Oct;34(4):e1800-e1809. [doi: 10.1002/hpm.2896] [Medline: 31429111]
27. Zhang M, Wang Y, Qian Z, Wang D. Research on development modes of internet hospitals in China. Health Econ Res 2019;36(5):23-26. [doi: 10.14055/j.cnki.33-1056/f.2019.05.007]
28. Chen Q, Zhang G, Zhang H, Wang C, Zhang K. Progress of application of modern information technology in four diagnostic methods. Jiangxi J Tradit Chin Med 2009;40(12):78-80.
29. Liu J. Difficulities and countermeasures for our internet medical service price management. Prices Mon 2020;3:28-32. [doi: 10.14076/j.issn.1006-2025.2020.03.05]
30. Zheng D, Nie L, Wang L, Chen Y. Medical insurance management under the background of 'Internet+Medical'. Chin Hosp 2021;25(6):5-8. [doi: 10.19660/j.issn.1671-0592.2021.6.02]
31. Pan Y, Yang Y, Xia Y. An analysis of the effect of "internet+" medical services incorporated into medical insurance payment in tertiary general hospitals in Shanghai. Modern Hosp 2023;23(1):18-20. [doi: 10.3969/j.issn.1671-332X.2023.01.005]
32. Xiao LJ, Tao R. Traditional Chinese medicine (TCM) therapy. Adv Exp Med Biol 2017;1010:261-280. [doi: 10.1007/978-981-10-5562-1_13] [Medline: 29098677]
33. Meng F, Ye M, Si J, et al. Status of traditional Chinese medicine healthcare services in nursing homes across China. Geriatr Nurs 2022;45:93-99. [doi: 10.1016/j.gerinurse.2022.02.025] [Medline: 35364480]
34. Wang K, Chen Q, Shao Y, et al. Anticancer activities of TCM and their active components against tumor metastasis. Biomed Pharmacother 2021 Jan;133:111044. [doi: 10.1016/j.biopha.2020.111044] [Medline: 33378952]
35. Martel J, Ojcius DM, Chang CJ, et al. Anti-obesogenic and antidiabetic effects of plants and mushrooms. Nat Rev Endocrinol 2017 Mar;13(3):149-160. [doi: 10.1038/nrendo.2016.142] [Medline: 27636731]
36. Feng X, Sureda A, Jafari S, et al. Berberine in cardiovascular and metabolic diseases: from mechanisms to therapeutics. Theranostics 2019;9(7):1923-1951. [doi: 10.7150/thno.30787] [Medline: 31037148]
37. Yan F, Li F, Liu J, et al. The formulae and biologically active ingredients of Chinese herbal medicines for the treatment of atopic dermatitis. Biomed Pharmacother 2020 Jul;127:110142. [doi: 10.1016/j.biopha.2020.110142] [Medline: 32330795]
38. O'Reilly E, Sevigny M, Sabarre KA, Phillips KP. Perspectives of complementary and alternative medicine (CAM) practitioners in the support and treatment of infertility. BMC Complement Altern Med 2014 Oct 14;14:394. [doi: 10.1186/1472-6882-14-394] [Medline: 25310971]
39. Xu J, Yang Y. Traditional Chinese medicine in the Chinese health care system. Health Policy 2009 May;90(2-3):133-139. [doi: 10.1016/j.healthpol.2008.09.003] [Medline: 18947898]
40. Shi X, Zhu D, Nicholas S, Hong B, Man X, He P. Is traditional Chinese medicine "mainstream" in China? Trends in traditional Chinese medicine health resources and their utilization in traditional Chinese medicine hospitals from 2004 to 2016. Evid Based Complement Alternat Med 2020;2020:9313491. [doi: 10.1155/2020/9313491] [Medline: 32595750]
41. Wang X, Xie Y, Yang X, Gu D. Internet-based healthcare knowledge service for improvement of Chinese medicine healthcare service quality. Healthcare (Basel) 2023 Jul 31;11(15):2170. [doi: 10.3390/healthcare11152170] [Medline: 37570410]
42. Opinions on facilitating the inheritance, innovation and development of traditional Chinese medicine. The State Council of the People's Republic of China. 2019. URL: https://www.gov.cn/gongbao/content/2019/content_5449644.htm?ivk_sa=1024320u [accessed 2024-02-02]
43. Tang ACY, Chung JWY, Wong TKS. Digitalizing traditional Chinese medicine pulse diagnosis with artificial neural network. Telemed J E Health 2012;18(6):446-453. [doi: 10.1089/tmj.2011.0204] [Medline: 22732024]
44. Wang X, Liu J, Wu C, et al. Artificial intelligence in tongue diagnosis: Using deep convolutional neural network for recognizing unhealthy tongue with tooth-mark. Comput Struct Biotechnol J 2020;18:973-980. [doi: 10.1016/j.csbj.2020.04.002] [Medline: 32368332]
45. Pan H. The development and clinical application of "Tianzhi" treatment. Gansu Technol 2009;25(9):164-165.
46. Jing Q. Preventive treatment in traditional Chinese medicine. Med Inform 2018;31(8):140-141. [doi: 10.3969/j.issn.1006-1959.2018.08.048]
47. Notice on printing and distributing three documents including the internet medical consultation administrative measures (trial). National Health Commission of the People's Republic of China, National Administration of Traditional Chinese Medicine of the People's Republic of China. 2018. URL: https://www.nhc.gov.cn/yzygj/c100068/201809/d53877b9685d4e7ca991f77542402c1b.shtml [accessed 2024-02-02]
48. Outline of the fourteenth five-year plan for national economic and social development of the people's republic of china and the vision 2035. General Office of the State Council of the People's Republic of China. 2021. URL: http://www.gov.cn/xinwen/2021-03/13/content_5592681.htm [accessed 2024-02-02]

XSL•FO
RenderX

49. Fang L, Zhu RR, Sang Z, et al. World Health Organization survey on the level of integration of traditional Chinese medicine in Chinese health system rehabilitation services. Integr Med Res 2023 Jun;12(2):100945. [doi: 10.1016/j.imr.2023.100945] [Medline: 37122485]
50. Simpson SH, Eurich DT, Majumdar SR, et al. A meta-analysis of the association between adherence to drug therapy and mortality. BMJ 2006 Jul 1;333(7557):15. [doi: 10.1136/bmj.38875.675486.55]
51. Ho PM, Magid DJ, Shetterly SM, et al. Medication nonadherence is associated with a broad range of adverse outcomes in patients with coronary artery disease. Am Heart J 2008 Apr;155(4):772-779. [doi: 10.1016/j.ahj.2007.12.011] [Medline: 18371492]
52. Ho PM, Rumsfeld JS, Masoudi FA, et al. Effect of medication nonadherence on hospitalization and mortality among patients with diabetes mellitus. Arch Intern Med 2006 Sep 25;166(17):1836-1841. [doi: 10.1001/archinte.166.17.1836] [Medline: 17000939]
53. Anderson LJ, Nuckols TK, Coles C, et al. A systematic overview of systematic reviews evaluating medication adherence interventions. Am J Health Syst Pharm 2020 Jan 8;77(2):138-147. [doi: 10.1093/ajhp/zxz284] [Medline: 31901098]
54. Strimbu K, Tavel JA. What are biomarkers? Curr Opin HIV AIDS 2010 Nov;5(6):463-466. [doi: 10.1097/COH.0b013e32833ed177] [Medline: 20978388]
55. Wang T, Azad T, Rajan R. The emerging influence of digital biomarkers on healthcare. Rock Health. 2016. URL: https://rockhealth.com/insights/the-emerging-influence-of-digital-biomarkers-on-healthcare [accessed 2024-02-02]
56. Meister S, Deiters W, Becker S. Digital health and digital biomarkers—enabling value chains on health data. Curr Dir Biomed Eng 2016 Sep 1;2(1):577-581. [doi: 10.1515/cdbme-2016-0128]
57. Coravos A, Khozin S, Mandl KD. Developing and adopting safe and effective digital biomarkers to improve patient outcomes. NPJ Digit Med 2019;2(1):14. [doi: 10.1038/s41746-019-0090-4] [Medline: 30868107]
58. Avram R, Olgin JE, Kuhar P, et al. A digital biomarker of diabetes from smartphone-based vascular signals. Nat Med 2020 Oct;26(10):1576-1582. [doi: 10.1038/s41591-020-1010-5] [Medline: 32807931]
59. Kourtis LC, Regele OB, Wright JM, Jones GB. Digital biomarkers for Alzheimer's disease: the mobile/wearable devices opportunity. NPJ Digit Med 2019;2(1):9. [doi: 10.1038/s41746-019-0084-2] [Medline: 31119198]
60. Ye G, Chen J, Zhang S, Yuan H. Research on the current situation, problems and countermeasures of internet medical treatment of traditional Chinese medicine. Chin Hosp 2023;27(11):34-39. [doi: 10.19660/j.issn.1671-0592.2023.11.08]

Abbreviations

AI: artificial intelligence
TCM: traditional Chinese medicine

Edited by J Sarvestan; submitted 20.05.25; peer-reviewed by Z Lin; revised version received 29.09.25; accepted 30.09.25; published 05.11.25.

Please cite as:
Wang Y, Song M, Pang Z, Liu B, Yao D, Li M, Xie Z, Bian Y, Hu H, Lai Y
Health Service Early-Stage Digital Adaptation of Traditional Chinese Medicine Internet Hospitals: Qualitative Exploratory Study
JMIR Form Res 2025;9:e77686
URL: https://formative.jmir.org/2025/1/e77686
doi:10.2196/77686

© Yao Wang, Menghuan Song, Zhenmiao Pang, Buping Liu, Dongning Yao, Meng Li, Zhirong Xie, Ying Bian, Hao Hu, Yunfeng Lai. Originally published in JMIR Formative Research (https://formative.jmir.org), 5.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Personal Activity Trackers and Family Engagement in a Pediatric Obesity Intervention: Randomized Controlled Trial

Juan Carlos Espinoza Salomon[1,2*], MD; Mahsa Babaei[3], PhD; Alexis Deavenport-Saman[3,4], DrPH; Olga Solomon[4,5*], PhD; Choo Phei Wee[6], MS; Ramon Durazo-Arvizu[4,7], PhD; Abu Sikder[5], BS; Payal Shah[5], MS; Patricia Castillo[4,5], MD; Larry Yin[4,5*], MD

[1]Department of Pediatrics, Ann & Robert H Lurie Children's Hospital of Chicago, 303 E Superior St, Chicago, IL, United States
[2]Department of Pediatrics, Northwestern University Feinberg School of Medicine, Chicago, IL, United States
[3]Department of Anesthesiology, Keck School of Medicine of the University of Southern California, Los Angeles, CA, United States
[4]Department of Pediatrics, Keck School of Medicine of the University of Southern California, Los Angeles, CA, United States
[5]Department of Pediatrics, Children's Hospital Los Angeles, Los Angeles, CA, United States
[6]Southern California Clinical and Translational Science Institute, Los Angeles, CA, United States
[7]Biostatistics and Data Analysis Core, The Saban Research Institute, Children's Hospital Los Angeles, Los Angeles, CA, United States
[*]these authors contributed equally

Corresponding Author:
Juan Carlos Espinoza Salomon, MD
Department of Pediatrics, Ann & Robert H Lurie Children's Hospital of Chicago, 303 E Superior St, Chicago, IL, United States

Abstract

Background: Pediatric obesity continues to be a national health crisis. Parents play a critical role in obesity interventions. Digital health interventions, such as personal activity trackers, can help better engage parents in pediatric obesity interventions and improve outcomes.

Objective: This study aimed to (1) assess the feasibility and acceptability of implementing personal activity trackers as part of a comprehensive family-based lifestyle intervention for pediatric obesity (BodyWorks) in a Federally Qualified Health Center; (2) evaluate the impact of personal activity trackers on parents' engagement, participant anthropometrics, and the overall program; and (3) examine the associations between steps per day and usage (minutes) with body composition outcomes.

Methods: A total of 158 families were randomized to the control (BodyWorks) or intervention (BodyWorks + physical activity tracker) arm. Mean levels of weight-by-height outcomes, including BMI, BMI *z* scores, and BMI percent of the 95th percentile, were compared between the 2 groups.

Results: There were no differences between study arms at baseline. After adjustment, there was a significant group difference in children's BMI *z* scores from baseline to the postintervention time point (*P* for interaction=.01).

Conclusions: Families in the intervention group that completed the program had slightly better weight outcomes than the controls. Engaging parents through digital health interventions may be an effective way to enhance existing pediatric obesity intervention programs.

Trial Registration: ClinicalTrials.gov NCT03215641; https://clinicaltrials.gov/study/NCT03215641

(JMIR Form Res 2025;9:e70341) doi:10.2196/70341

KEYWORDS

pediatrics; obesity; federally qualified health center; digital health intervention; underserved population

Introduction

Globally, 39 million children younger than 5 years of age, and 340 million between the ages of 5 and 19 years, are overweight or obese, with the prevalence of overweight and obesity reaching approximately 18% worldwide [1]. In the United States, pediatric obesity and associated comorbidities disproportionately impact low-income families and families of color [2,3]. Evidence-based pediatric obesity interventions involve multidisciplinary, multimodality, family-based therapies that are both time- and resource-intensive and may not be widely available or accessible [4].

Significant interest has emerged in leveraging digital health interventions to address pediatric obesity, including mobile apps, wearable devices, and text messaging [5]. One attractive aspect of digital health interventions is their relatively low cost to scale, particularly in low-resource settings. However, there are equity, implementation, and adoption concerns related to

XSL•FO
RenderX

the growing digital divide, including gaps in access to and quality of broadband internet, access to digital devices, and the digital literacy and skills necessary to use the devices and the applications appropriately [6,7]. The COVID-19 pandemic made the digital divide especially visible when health care and educational services in the United States attempted to transition to remote delivery. The American Medical Informatics Association (AMIA) has published a statement expressing concerns that digital technology is widening health inequity generated by the digital divide rather than decreasing it [8], and the concept of digital inclusion has gained traction as a social determinant of health [9,10].

Our Federally Qualified Health Center (FQHC) uses the BodyWorks curriculum to deliver a low- to medium-intensity comprehensive behavioral family lifestyle intervention that focuses on dietary intake, physical activity, and behavioral strategies for weight loss [11]. Our team was interested in deploying digital health interventions to support the families participating in our program. Because parents play a central role in pediatric obesity management [12], we developed a parent-focused personal activity tracker (PAT) intervention. We have previously published our protocol [13] and demonstrated that our intervention reduced program attrition [14]. Here, we examined the feasibility and acceptability of implementing personal activity trackers as part of a comprehensive family-based lifestyle intervention for pediatric obesity (BodyWorks) in an FQHC. We also evaluated the impact of personal activity trackers on parents' engagement, participant anthropometrics, and the overall program. In addition, we examined the associations between steps per day and usage (minutes) with body composition outcomes.

Methods

Study Design

The prospective, open-label, randomized clinical trial was conducted between 2015 and 2017 in the CHLA AltaMed General Pediatrics Clinic, a FQHC located within an academic medical center serving many underserved patients.

Participating families were enrolled if the primary child was aged 7 - 18 years, had a BMI ≥85% for age and sex, and had at least one parent or caretaker able to attend the sessions with the child and willing to wear a PAT if assigned to the intervention arm. Participants were provided with incentives for their time and effort. This trial was conducted in accordance with the CONSORT (Consolidated Standards of Reporting Trials) guidelines (checklist provided in Checklist 1).

Study Setting

Our FQHC serves as the medical home for a large, urban, medically underserved patient population of approximately 27,000 children and provides more than 100,000 patient visits annually. Approximately 75% of families identify as Latino, half of whom have limited English proficiency (LEP). Medi-Cal, California's Medicaid program, insures 93% of patients.

BodyWorks Curriculum

BodyWorks is the comprehensive behavioral family lifestyle intervention program, originally developed by the Office of Women's Health of the Department of Health and Human Services. Our adaptation and implementation of the curriculum have been previously described in detail [14]. Briefly, families are referred to BodyWorks by their primary care physician. The program is offered at no cost to families. The 8-week program consists of 1 orientation and 7 classes; the groups meet weekly for 2 hours at a time. Sessions are offered in English and Spanish; families decide which session they prefer. Each session begins with all participants together for physical activity and a healthy snack. Parents and children are then separated to receive tailored curricula from a multidisciplinary team. The curriculum focuses on nutrition, physical activity, stress, sleep, body image, bullying, the role of the media, and how to continue lessons learned.

Study Procedures and Data Collection

Anthropometrics

Height and weight were measured weekly for children. BMI, BMI percentile for age and sex, and BMI percent of the 95th percentile (%BMIp95) were all calculated using standard approaches from weekly measurements. We used the Centers for Disease Control and Prevention (CDC) definitions of overweight and obesity in childhood [15].

Surveys

Enrolled families completed a demographic survey on the first day of BodyWorks. At completion, all families completed a technology use survey, which consisted of questions such as, "Do you own a smartphone?" "Do you use text messaging, email, or any other social platform?" and "Would you like to receive BodyWorks communication through digital health technology?" Finally, families in the intervention arm completed an exit survey about their experience using PATs, consisting of six 5-point Likert scale questions.

Randomization

To ensure balanced allocation of language groups and prevent cross-contamination, block randomization was used to assign whole groups to either arm of the study. The primary aim for program completion was defined as a family attending at least 4 of the 7 sessions.

Intervention Arm (BodyWorks + PAT)

Participants were assigned to either the intervention arm or the control arm. In the intervention arm, the parent and the child aged ≥13 years received the PAT, Fitbit (Fitbit). BodyWorks staff instructed the intervention arm participants on PAT use, including charging and usage practices. They were instructed to wear PAT devices on the wrist of the nondominant arm for a minimum of 20 hours per day. All participants received a US $15 gift card for agreeing to participate. Those randomized to receive a PAT received an additional US $50 gift card for returning the device at the end of the trial.

XSL•FO
RenderX

Control Arm (BodyWorks Only)

Participants in the control arm attended the same sessions and received the same curriculum as those in the intervention.

Outcomes

Program completion, PAT exit survey (attitudes and perceptions), BMI, BMI *z* score, BMI percentile, %BMIp95, number of steps per day, and usage time per day.

Covariates

Demographic characteristics, session attendance, and technology access and preferences.

Statistical Analysis

Descriptive statistics were used to summarize and describe the distribution of demographic variables. Continuous variables were summarized as means and SDs, whereas percentages were used for categorical variables. We examined family and child demographic differences between the intervention and control groups by conducting chi-square tests or Fisher exact tests, and differences in age and relative weight outcomes by conducting 2-sample *t* tests.

For the primary study outcomes, a generalized estimating equation approach was conducted to examine changes in BMI *z* scores and %BMIp95 between children in the intervention and control groups, adjusting for sociodemographic and program characteristics. The nonadditive effects of children in the intervention compared to the control groups were also examined (eg, whether the change from baseline to follow-up differed between children in the intervention and control groups) by including intervention group-by-visit interaction terms. Covariates included age, sex, LEP, household income as a percentage of the federal poverty level, and program completion. A multivariate linear regression model assessed relationships between child demographics and parent and child total steps per day, and minutes of PAT use daily, while adjusting for the percent of days with valid data. The effect of total steps per day and PAT usage on change in relative weight measures (BMI *z* score and BMI percent of 95th percentile) was estimated using a generalized estimating equation model, adjusting for baseline levels of relative weight measure, ethnicity, household size, gender, household income, and percent days with valid data. The significance level was set at 5% with 2-sided tests. All statistical computations were done in Stata/SE 18 (StataCorp).

PAT Data Extraction

PAT data were extracted for each participant through the web application programming interface (API) using a customized Python script. A developer app account was registered through the API portal and configured for OAuth 2.0 and client-side authentication. After running the Python script, OAuth 2.0 authentication occurred through a web browser, where each user's PAT login credentials were used, and all data types were selected for sharing. After authentication, an access token was retrieved from the Representational State Transfer API call to grant authorization using a unique user ID and client secret code assigned to the PAT developer app. The access token was then used for a customized Representational State Transfer API call to retrieve steps, distance, and minutes, including sedentary, lightly active, fairly active, and very active steps from 2015 to 2020. The extracted JavaScript Object Notation data were parsed and formatted into a CSV file.

PAT Data Aggregation

Each CSV file contained all available data for a specific PAT account from the program's inception to its conclusion. A Python script was used to aggregate all the individual data into a master CSV file. Since PAT accounts were reused, participants were mapped to cycles for tracking purposes. In addition, a mapping of each cycle to the study timeframe was used to filter and organize data in the master CSV file. The final dataset was organized into cycles and participants with data points such as date, day of week, steps, active minutes, and distance traveled.

Ethical Considerations

The Children's Hospital Los Angeles Institutional Review Board approved the study (CHLA-15-00269), and all participants signed consent or assent in their preferred language. To protect participant privacy and confidentiality, only data necessary to meet the study objectives were collected. Personal identifiers were removed and replaced with coded study identification numbers, and data were stored in password-protected, secure institutional databases. Study participants received a US $15 gift card and an additional US $50 gift card upon returning their PATs.

Results

Study Enrollment and Completion

Initially, 174 families participated in BodyWorks; however, 11 families were excluded from the analysis, as their only participating child was younger than 7 years of age. An additional 5 families were excluded because it was their second or third time in the BodyWorks program. The study sample comprised 158 families, including 200 children, with 80 children in the control group and 120 in the intervention group (Figure 1).

XSL•FO
RenderX

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) diagram for the BodyWorks study. BW: BodyWorks; PAT: personal activity tracker.

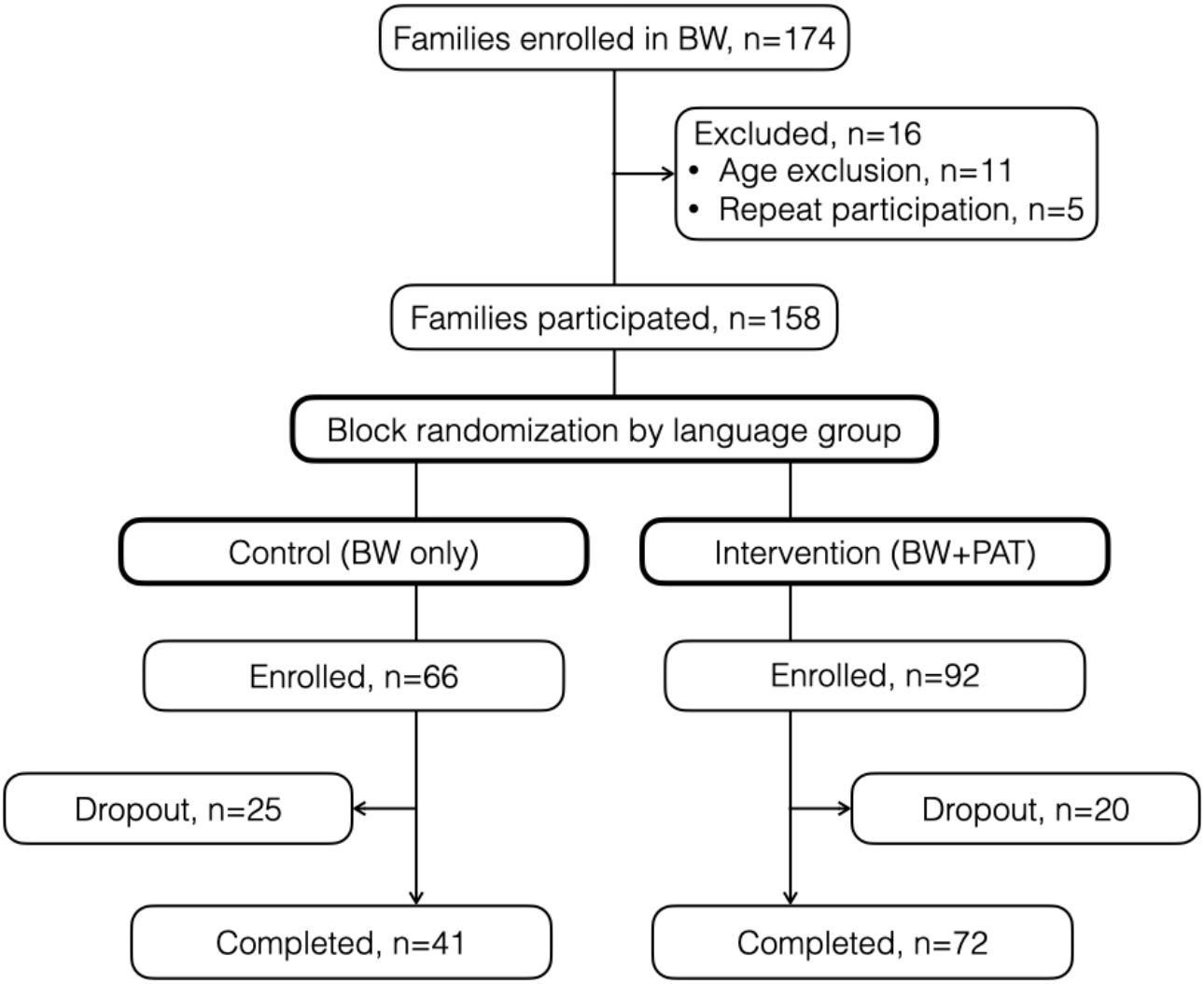

Sociodemographic and Anthropometric Characteristics

Family-Level Characteristics

The demographics of BodyWorks participants represented the FQHC from which they were recruited and the local community (Table 1). Most families were Latino, had LEP, had household incomes below 200% of the federal poverty level, and had household sizes of 5 or more people. There were no significant differences between the intervention and control groups at baseline.

XSL•FO
RenderX

Table . Baseline sociodemographic characteristics of families and children by intervention and control groups.

Characteristic and study arm	Control	Intervention	Total	*P* value
Families				
Ethnicity, n/N (%)				.72
Hispanic Latino	60/66 (93.8)	88/92 (95.7)	148/158 (94.9)	
Not Hispanic or Latino	4/66 (6.2)	4/92 (4.3)	8/158 (5.1)	
Household Income as % FPL[a], n/N (%)				≥.99
<200% FPL	50/66 (92.6)	77/92 (92.8)	127/158 (92.7)	
>200% FPL	4/66 (7.4)	6/92 (7.2)	10/158 (7.3)	
Household size, n/N (%)				≥.99
5+	34/66 (54)	47/92 (52.8)	81/158 (53.3)	
1-4	29/66 (46)	42/92 (47.2)	71/158 (46.7)	
Limited English proficiency, n/N (%)				.05
Yes	40/66 (62.5)	71/92 (77.2)	111/158 (71.2)	
No	24/66 (37.5)	21/92 (22.8)	45/158 (28.8)	
Children				
Sex, n/N (%)				.77
Female	34/80 (42.5)	54/120 (45)	88/200 (44)	
Male	46/80 (57.5)	66/120 (55)	112/200 (56)	
Age (years), mean (SD)	10.3 (1.8)	11.3 (2.8)	10.9 (2.5)	.04
Height (cm), mean (SD)	143.4 (13.8)	145.1 (13.5)	144.4 (13.6)	.33
Weight (kg), mean (SD)	58.7 (21.7)	62.5 (24.2)	61.0 (23.3)	.31
BMI, mean (SD)	27.7 (5.5)	28.7 (6.9)	28.3 (6.4)	.45
BMI percentile, mean (SD)	97.6 (2.2)	97.1 (3.2)	97.3 (2.9)	.70
BMI *z* score, mean (SD)	2.1 (0.4)	2.1 (0.5)	2.1 (0.4)	.60
%BMIp95[b], mean (SD)	120% (20%)	120% (20%)	120% (20%)	.56

[a]FPL: federal poverty level.

[b]%BMIp95: BMI percent of the 95th percentile.

Child-Level

There was a significant age difference (*P*=.04) between children in the intervention and control groups (Table 1). At baseline, there were no differences in BMI *z* scores, BMI percentiles, or % BMI95 between the intervention and control groups. There were also no other significant differences between children in the 2 groups at baseline.

Technology Use

A total of 132 families completed the technology use survey. Of these, 130 (98.5%) respondents reported having a cell phone, with 39 (30.2%) having iOS (Apple) devices and 83 (64.3%) having Android (Google) devices. Almost all families (127/132, 96%) used their phone for texting, with fewer families using their phones for social media (99/132, 75%) and email (96/132, 73%). About 86% (113/132) of families had an unlimited data plan.

Weight Outcomes

Unadjusted Weight Outcomes

The weight outcomes are presented in Table 2 among all children who enrolled in the program and those who completed the intervention. Among all children, there was a significant difference in the change of BMI *z* scores between children in the control and intervention groups (mean –0.00, SD 0.07 vs mean –0.04, SD 0.11; *P*=.02; Table 2). There were, however, no differences in the change of %BMIp95 between children in the control and intervention groups (mean 121.6%, SD 19.8% vs mean 119.7%, SD 22.2%; *P*=.56). Among children who completed the intervention, there were no differences in the change in BMI *z* scores between children in the control and intervention groups (mean –0.01, SD 0.07 vs mean –0.04, SD 0.12; *P*=.10; Table 2). There were also no differences in the change in %BMIp95 between children in the control and

XSL•FO
RenderX

intervention groups (mean 121%, SD 21.7% vs mean 118.7%, SD 20.9%; *P*=.53).

XSL•FO
RenderX

Table . Summary statistics for children's body composition measures by intervention group: enrolled and completed.

Characteristic and study arm	Control	Intervention	Total	*P* value
Enrolled in a study				
Sex, n/N (%)				.77
Female	34/80 (42.5)	54/120 (45)	88/200 (44)	
Male	46/80 (57.5)	66/120 (55)	112/200 (56)	
Height (cm) at baseline, mean (SD)	143.37 (13.85)	145.12 (13.47)	144.41 (13.62)	.38
Weight (kg) baseline, mean (SD)	58.72 (21.71)	62.51 (24.24)	60.99 (23.28)	.26
BMI at baseline, mean (SD)	27.69 (5.46)	28.75 (6.91)	28.32 (6.38)	.25
BMI percentile at baseline, mean (SD)	97.61 (2.20)	97.09 (3.25)	97.30 (2.88)	.21
BMI *z* score at baseline, mean (SD)	2.141(0.38)	2.10 (0.45)	2.12 (0.43)	.52
%BMIp95 at baseline, mean (SD)	122.1% (20.9%)	121.1% (23%)	121.5% (22.1%)	.77
Height (cm) at follow-up, mean (SD)	144.35 (13.69)	145.66 (13.24)	145.17 (13.38)	.53
Weight (kg) at follow-up, mean (SD)	59.38 (22.30)	61.80 (22.64)	60.90 (22.48)	.49
BMI at follow-up, mean (SD)	27.61 (5.42)	28.29 (6.44)	28.04 (6.07)	.47
BMI percentile at follow-up, mean (SD)	97.64 (2.14)	96.54 (4.23)	96.95 (3.63)	.05
BMI *z* score at follow-up, mean (SD)	2.15 (0.36)	2.07 (0.48)	2.10 (0.44)	.23
%BMIp95 at follow-up, mean (SD)	121.6% (19.8%)	119.7% (22.2%)	120.4% (21.3%)	.56
BMI *z* score difference, mean (SD)	–0.00 (0.07)	–0.04 (0.11)	–0.02 (0.10)	.02
%BMIp95 difference, mean (SD)	–0.1% (3.7)	–1.1% (3.3)	–0.7% (3.5)	.07
BMI percentile difference, mean (SD)	–0.10 (0.71)	–0.56 (1.78)	–0.39 (1.49)	.04
Completed study				
Sex, n/N (%)				.60
Female	20/50 (40)	43/95 (45.3)	63/145 (43.4)	
Male	30/50 (60)	52/95 (54.7)	82/145 (56.6)	
Height (cm) at baseline, mean (SD)	145.49 (14.23)	144.37 (13.92)	144.76 (13.99)	.65
Weight (kg) baseline, mean (SD)	60.94 (24.73)	61.00 (22.44)	60.98 (23.17)	.99
BMI at baseline, mean (SD)	27.76 (6.13)	28.35 (6.20)	28.15 (6.16)	.59
BMI percentile at baseline, mean (SD)	97.46 (2.24)	97.26 (2.70)	97.33 (2.55)	.66
BMI *z* score at baseline, mean (SD)	2.11 (0.39)	2.09 (0.40)	2.10 (0.40)	.79
%BMIp95 at baseline, mean (SD)	121% (21.9%)	120% (20.6%)	120.3% (21%)	.78

XSL•FO
RenderX

Characteristic and study arm	Control	Intervention	Total	*P* value
Height (cm) at follow-up, mean (SD)	146.43 (14.19)	145.40 (13.57)	145.75 (13.74)	.67
Weight (kg) at follow-up, mean (SD)	61.71 (24.94)	61.11 (22.31)	61.32 (23.15)	.88
BMI at follow-up, mean (SD)	27.74 (6.16)	28.03 (6.25)	27.93 (6.20)	.79
BMI percentile at follow-up, mean (SD)	97.39 (2.37)	96.64 (3.93)	96.89 (3.49)	.22
BMI *z* score at follow-up, mean (SD)	2.12 (0.38)	2.05 (0.45)	2.07 (0.43)	.40
%BMIp95 at follow-up, mean (SD)	121% (21.7%)	118.7% (20.9%)	119.5% (21.1%)	.53
BMI *z* score difference, mean (SD)	–1.2% (6.8%)	–4.2% (11.5%)	–3.2% (10.3%)	.10
%BMIp95 difference, mean (SD)	–0.01 (0.04)	–0.01 (0.04)	–0.01 (0.04)	.24
BMI percentile difference, mean (SD)	–0.20 (0.77)	–0.63 (1.91)	–0.48 (1.63)	.13

Adjusted Weight Outcomes

The adjusted multivariable regression demonstrated a significant difference in BMI *z* scores between children in the intervention and control groups from baseline to the postintervention time point (β=–.035; *P* value for interaction=.01; Table 3). There were no statistically significant differences in %BMIp95 between children in the intervention and control groups from baseline to the postintervention time point (β=–.008; *P* value for interaction=.19). Changes in BMI percentile from baseline differed between intervention groups (β=–.571; *P* value for interaction <.01). All models were adjusted for baseline BMI, age-gender–specific relative weight measure, age in years, household income, LEP, gender, and study completion.

The relative weight change in the control group is 0.003, –0.001, and –0.023 for BMI *z* scores, %BMIp95, and BMI percentile, respectively. In the intervention group, the corresponding change is 0.003–0.035= –0.032, –0.001–0.008=–0.009, and –0.023–0.571=–0.594. The interaction terms, which represent the change difference between intervention and control groups, are –0.035 (*P*=.01), - 0.008 (*P*=.19), and 0.571 (*P*=.004) for BMI *z* scores, %BMIp95, and BMI percentile. In summary, changes in relative weight measures in the control group are not statistically significant, as seen from the 95% CIs, which include 0. In contrast, a statistically significant decrease in the average of the 3 measures of weight-by-height was observed in the intervention group.

XSL•FO
RenderX

Table . Multivariable generalized estimating equation (GEE) model assessing the change in relative weight measures by intervention arm, adjusting for baseline levels, age, household income, limited English proficiency, gender, and study completion.

Variable	BMI *z* score		%BMIp95[a]		BMI percentile	
	β (95% CI)	*P* value	β (95% CI)	*P* value	β (95% CI)	*P* value
Intercept	–0.020 (–0.06 to 0.02)	.32	0.008 (–0.01 to 0.02)	.35	-8.12 (–14.82 to –1.42)	.02
Baseline relative weight	1.012 (1.00 to 1.04)	<.01	0.998 (1.00 to 1.01)	<.01	1.09 (1.02 to 1.16)	<.01
Study arm						
Intervention	0.001 (–0.00 to 0.01)	.76	0.000 (–0.001 to 0.002)	.62	0.00 (–0.10 to 0.11)	.96
Time						
Postintervention	0.003 (–0.01 to 0.02)	.72	–0.001 (–0.01 to 0.01)	.83	–0.02 (–0.13 to 0.09)	.68
Study arm x time						
Intervention x postintervention	–0.035 (–0.06 to –0.01)	<.01	–0.008 (–0.02 to 0.00)	.19	–0.57 (–0.96 to –0.18)	<.01
Age (years)						
≥13	0.007 (–0.01 to 0.02)	.3	0.001 (–0.00 to 0.01)	.65	0.24 (0.04 to 0.45)	.02
Household income (% FPL[b])						
>200% FPL	0.000 (–0.02 to 0.02)	≥.99	–0.001 (–0.01 to 0.01)	.8	–0.01 (–0.31 to 0.30)	.97
Limited English proficiency						
No	–0.007 (–0.02 to 0.02)	.37	0.000 (–0.01 to 0.01)	.92	–0.23 (–0.54 to 0.07)	.14
Sex (child)						
Male	–0.011 (–0.03 to 0.00)	.12	0.002 (–0.01 to 0.00)	.47	–0.23 (–0.45 to –0.01)	.04
Completed study						
Yes	–0.013 (–0.02 to –0.003)	.01	–0.006 (–0.01 to –0.001)	.01	–0.12 (–0.27 to 0.03)	.12

[a]%BMIp95: BMI percent of the 95th percentile.

[b]FPL: federal poverty level.

Personal Activity Tracker Data

A total of 146 parents were assigned PATs, generating 6826 data points over 43 to 52 days each. Of these, 3662 (53.7%) were valid or nonzero data points. Almost two-thirds (63%) of participants had at least 50% valid data. Over a week, participants were most likely to generate data on Wednesdays and least likely on Sundays.

Excluding days with no data, the mean (SD) step count per user per day was 8418 (8186). The mean (SD) minutes of physical activity per user per day was 244.7 (289.7). The mean (SD) miles traveled per user daily was 3.6 (3.53). Participants were most active on Thursdays (9364 daily steps) and Fridays (8917 daily steps), and least active on Saturdays (7651 daily steps) and Sundays (6740 daily steps).

Table 4 shows the multivariable linear regression models for the 2 outcome measures recorded in the PAT instrument, including the number of daily steps and usage (minutes) for parents and children. The resulting 4 models included 4 demographic variables: household size, percent, gender, household income, and ethnicity. The models were adjusted for percent days with valid data and the number of sessions attended. Parents living in households with 1 to 4 people had an average of 2907 fewer steps per day compared to those in households with 5+ individuals (β=–2907; *P*=.02), as well as 60 minutes less (β=–60; *P*=.05) of usage time. Fathers had an average of 3270 fewer steps than mothers (β= –3270; *P*=.07). Percent days with valid data were significantly associated with steps per day and minutes of use among parents and children, whereas the number of sessions attended did not impact either outcome measure.

In Table 5, an increase of 1000 steps per day resulted in an increase in BMI *z* score, %BMI95, and BMI percentile from

XSL•FO
RenderX

baseline of 0.009 (*P* value for interaction ≤.01), 0.004 (*P*≤.01), and 0.07 (*P*=.05), respectively.

A 30-minute increase in PAT usage was associated with an increase from baseline of BMI *z* score, %BMI95, and BMI percentile of 0.009 (*P*=.03), 0.004 (*P*=.03), and 0.054 (*P*=.05; Table 6).

Table . Multivariable linear regression for the number of steps per day and usage (minutes) among parents and children.

Variable	Parent				Child			
	Steps/day		Minutes/day		Steps/day		Minutes/day	
	β	*P* value	β	*P* value	β	*P* value	β	*P* value
Intercept	–594	.76	39	.47	3215	.25	142	.06
Household size								
1 to 4	–439	.55	5	.80	–2907	.02	–60	.05
Percent days with valid data	1441	<.01	35	<.01	939	.04	26	.03
Sex								
Male	–3270	.07	–54	.28	–1787	.13	–50	.10
Household income (% FPL[a])								
>200% FPL	–2309	.11	–53	.18	–1991	.28	–54	.27
Ethnicity								
Hispanic or Latino	1382	.38	–29	.51	2245	.32	20	.73
Number of sessions attended	600	.10	21	.03	213	.60	0	.98

[a]FPL: federal poverty level.

XSL•FO
RenderX

Table . Child body composition changes and parent FitBit steps. The rate of change of BMI *z* score, %BMI over 95th percentile, and BMI percentile increases with increasing activity, as measured by the number of steps. That is, for BMI *z* score, %BMI over 95th percentile, BMI percentile, the rate of change increases per 1000 steps increase. Estimates of the rate of change are adjusted for the baseline outcome, race/ethnicity, household size, percent days with valid data, sex, and household income.

Variable	BMI *z* score		BMI % over 95 percentile		BMI percentile	
	Effect (95% CI)	*P* value	Effect (95% CI)	*P* value	Effect (95% CI)	*P* value
Intercept	–0.026 (–0.087 to 0.035)	.40	–0.017 (–0.042 to 0.008)	.18	2.334 (0.991 to 5.659)	.17
Steps/day	–0.002 (–0.008 to 0.003)	.41	–0.002 (–0.003 to 0.000)	.10	0.045 (0.006 to 0.096)	.08
Visit						
Follow-up	–0.072 (–0.127 to –0.017)	.01	–0.028 (–0.049 to –0.007)	.01	0.584 (–1.181 to 0.013)	.06
Time × steps/day						
Follow-up	0.009 (0.003 to 0.016)	<.01	0.004 (0.001 to 0.006)	<.01	0.070 (0.001 to 0.140)	.05
Baseline relative weight	1.010 (0.964 to 1.056)	<.01	1.017 (0.993 to 1.040)	<.01	0.972 (0.935 to 1.009)	<.01
Ethnicity						
Hispanic or Latino	0.025 (0.008 to 0.042)	<.01	0.008 (0.001 to 0.016)	.02	0.102 (–0.007 to 0.210)	.07
Household size						
1 to 4	–0.008 (0.008 to 0.042)	.24	–0.005 (–0.014 to 0.003)	.21	0.032 (–0.199 to 0.134)	.70
Percent days with valid data	0.003 (–0.012 to 0.017)	.72	0.002 (–0.001 to 0.006)	.20	0.116 (–0.233 to 0.001)	.05
Sex						
Male	–0.010 (–0.065 to 0.045)	.73	–0.007 (–0.018 to 0.004)	.22	0.486 (0.061 to 0.912)	.03
Household Income (% FPL[a])						
>200% FPL	–0.022 (–0.085 to 0.040)	.49	–0.013 (–0.035 to 0.008)	.22	0.395 (–0.212 to 1.002)	.20

[a]FPL: federal poverty level.

XSL•FO
RenderX

Table . Child body composition changes and parent FitBit use in minutes. Estimates of the rate of change are adjusted for the baseline outcome, race or ethnicity, household size, percent days with valid data, sex, and household income.

Variable	BMI *z* score		BMI % over 95 percentile		BMI percentile	
	Effect (95% CI)	*P* value	Effect (95% CI)	*P* value	Effect (95% CI)	*P* value
Intercept	–0.023 (–0.071 to 0.025)	.35	–0.015 (–0.037 to 0.006)	.17	–0.016 (–3.824 to 3.792)	.99
Minutes/day	–0.004 (–0.009 to 0.001)	.09	–0.002 (–0.004 to 0.000)	.03	0.000 (–0.054 to 0.054)	.99
Time						
Follow-up	–0.066 (–0.127 to –0.005)	.03	–0.027 (–0.050 to –0.004)	.02	–0.439 (–0.904 to 0.027)	.07
Time × minutes/day						
Follow-up	0.009 (0.001 to 0.018)	.03	0.004 (0.000 to 0.007)	.03	0.054 (–0.001 to 0.108)	.05
Baseline relative weight	1.010 (0.982 to 1.038)	<.01	1.016 (0.996 to 1.036)	<.01	0.999 (0.959 to 1.039)	<.01
Ethnicity						
Hispanic or Latino	0.030 (0.018 to 0.042)	<.01	0.010 (0.003 to 0.016)	<.01	0.185 (0.123 to 0.247)	<.01
Household size						
1 to 4	–0.014 (–0.027 to –0.002)	.03	–0.007 (–0.014 to –0.001)	.03	–0.119 (–0.285 to 0.047)	.16
Percent days with data	0.005 (–0.005 to 0.014)	.35	0.003 (–0.000 to 0.006)	.08	–0.036 (–0.144 to 0.073)	.52
Sex						
Male	–0.012 (–0.042 to 0.018)	.44	–0.007 (–0.017 to 0.003)	.16	0.144 (–0.169 to 0.457)	.37
Household income (% FPL[a])						
>200% FPL	–0.025 (–0.078 to 0.029)	.36	–0.014	.19	0.024 (–0.386 to 0.433)	.91

[a]FDL: federal poverty level.

PAT Exit Survey

A total of 83 participants completed the PAT exit survey. The survey included six 5-point Likert scale items to assess (1) PAT comfort, (2) ease of charging, (3) enjoyment, (4) motivation to do more activity, (5) utility of weekly reports and feedback, and (6) consideration of PAT purchase in the future. The responses were overwhelmingly positive, with ≥95% of respondents agreeing or strongly agreeing with every statement.

Discussion

Principal Findings

In this randomized control trial, we demonstrated the feasibility of using a digital health intervention involving PATs to positively impact a predominantly low-income, Latino patient population participating in a pediatric obesity intervention at an FQHC. Families reported that their PATs were easy to use, comfortable to wear, and motivating. This is consistent with other studies in the literature [16,17]. In addition, there was a significant decrease in BMI *z* scores among children in the intervention compared to children in the control group from baseline to the postintervention time point. Interestingly, within the intervention arm, an increase of 1000 steps per day and a 30-minute increase in PAT usage resulted in an increase in BMI *z* score, %BMI95, and BMI percentile from baseline to the postintervention time point. Given that this group overall had improved weight outcomes, the apparently paradoxical weight increase among those with higher physical activity may be related to an increase in muscle mass in the short term as the family became more active.

Digital health interventions have the potential to be powerful adjuncts to existing programs and interventions. Low-cost technologies can scale rapidly to provide more and better care to more patients. In the past, there have been concerns about the ability to meaningfully deploy technology-based interventions in low-income, non–English-speaking, or otherwise underserved communities. This has been justified with concerns about access to or comfort with mobile technologies [7]. While national surveys do indicate a gap in access to technologies such as smartphones for the poorest Americans, this may be a result of their methodology and sampled population [18]. Our study and others have consistently

XSL•FO
RenderX

demonstrated that, at least among urban communities, access to smartphones is nearly ubiquitous, regardless of socioeconomic status, except at the very extremes of poverty. Out of the 132 families who completed our technology survey, 94% reported owning a smartphone and using it regularly for texting, social media, and email. This finding underlines the importance of understanding not only the needs, but also the strengths and resources of any community for which an intervention is planned.

With regard to PAT use, most participants were able to generate valid data from 50% or more of their days in the study. Common barriers for not generating valid data included forgetting to wear the PAT or not charging the PAT. Wednesdays were the day with the most valid data; this may be related to the fact that the BodyWorks program occurred on Monday, Tuesday, and Wednesday nights, which provided a natural reminder for patients to charge and wear their PAT. Similarly, families were most active on Thursdays and Fridays, and least active over the weekend. This may be due to the effect of being reminded during the Monday-Wednesday sessions, and families become more active in the following days. It is interesting to note that the data shows the least amount of activity over the weekend. This may reflect families genuinely being less active or possibly forgetting to wear or charge their PATs over the weekend.

The BodyWorks program has made several accommodations and adaptations to meet the needs of our patient population and encourage families to participate and engage. Because childcare can be costly, all children in a family are welcome to join the group and participate, regardless of age or weight status. For parents who work or are otherwise unable to attend themselves, other family members (typically grandparents or adult siblings) may participate in their place. The entire program is supported by a bilingual team that is able to provide culturally and linguistically competent care to all families. In 2020, when the program went online in response to the COVID-19 pandemic, we continued to offer Spanish- and English-language online groups, and families could connect with a program coordinator if they needed help familiarizing themselves with the teleconferencing platform used.

From a theoretical standpoint, personal activity trackers enable self-monitoring, a well-established behavior change technique that has been shown to increase physical activity in adults [19-21]. In our study, we further supported behavior change by providing coaching, another well-established behavior change technique. Each week, a member of the study team reviewed each participant's data with them, praising achievements, identifying barriers to physical activity, and discussing solutions. This created an opportunity for personalized attention for each participant.

The present research demonstrated a statistically significant improvement in weight outcomes for children in the intervention arm from baseline to the postintervention time point compared to those in the control arm. However, these changes are unlikely to be clinically significant. Existing clinical guidelines recommend 5% - 10% weight loss for adults and 4% - 7% change in BMI percentile or *z* score for children to begin to see metabolic benefit [21,22]. Neither of these thresholds was achieved by our cohort. Furthermore, we only report on weight outcomes at 8 weeks, a relatively short period of time. Future studies will need to evaluate the long-term impact of this intervention, as the durability of interventions can vary significantly from their short-term impact. Finally, it is worth remembering that while the goal of most adult obesity interventions is weight loss, the goal of pediatric obesity interventions varies by age, obesity severity, comorbidities, and developmental stage. Weight maintenance, decreasing growth velocity, or weight loss may all be appropriate goals depending on the patient [20].

One interesting aspect of our study is that, despite being a pediatric obesity program, our digital health intervention targeted parents. Parental education and behavior change are the critical drivers of weight management in pediatric obesity, particularly among younger children [22]. Parent-only interventions have been shown to be effective modalities for pediatric obesity [23,24]. Engaging parents in making changes in the home and adopting healthier habits leads to children adopting healthier lifestyles and, over time, improved weight outcomes.

There were departures from the original study protocol. For example, each family was supposed to complete paper-based activity logs each week. These were intended to serve as a control for the personal activity trackers. However, from the beginning of the study, we observed very poor completion rates, particularly among the families who had PATs. One participant asked, "We look at my Fitbit every week, why do I have to fill this out?" After 2 cycles, the study team decided to no longer collect physical activity logs. While this may have impacted our overall analysis and the opportunity to compare self-reported physical activity with physical activity captured by the personal activity trackers, it was felt to be the right decision for the families participating in our program. In the original protocol, only one parent was to receive a personal activity tracker. Within the first 2 cycles, it became clear that for families with 2 parents participating in BodyWorks, determining which parent received the PAT was sometimes a source of conflict. Therefore, we decided to allow both parents to have a PAT if they were both interested in using one.

Strengths and Limitations

This study has several strengths. A large number of families participated, allowing for a robust statistical analysis. Our overall intervention design is straightforward and low-cost and could be replicated in clinical settings with minimal expenses. While PATs can be expensive, most modern smartphones are now capable of tracking activity, allowing this study to be replicated without the need to purchase any additional devices. Having demonstrated success in a relatively low-resource setting, we anticipate that this approach would be at least as successful in higher-resource settings. While encouraging, there are also several important limitations. The participants were relatively homogenous, being predominantly Latino, low-income, and having LEP. This may limit the generalizability of our findings to other settings. There is an inherent self-selection bias in our study population, as we recruited families who had already agreed to attend a structured obesity

XSL•FO
RenderX

intervention program. These families likely had a higher baseline level of motivation, which may have impacted our findings and raises the question of whether other, less motivated populations would have similar outcomes. This analysis only evaluated short-term (2 months) weight outcomes; the durability of the impact remains to be evaluated.

Conclusions

Leveraging PATs to enhance an existing pediatric obesity intervention in an FQHC is feasible and may improve weight outcomes. Family income and English proficiency did not impact participants' ability to benefit from the technology. Future studies should further evaluate the real-world effectiveness and long-term impact of this approach.

Acknowledgments

The authors would like to thank all the families who agreed to participate and made this study possible. We also acknowledge the valuable contributions of Lisa Lopez, Audrey Lai, Aric Ponce, and Alex Wormuth to data collection and program administration.

Funding

Research reported in this publication was supported by the Biostatistics Core of The Saban Research Institute at Children's Hospital Los Angeles (CHLA) and by grants (UL1TR001855 and UL1TR000130) from the National Center for Advancing Translational Science (NCATS) of the US National Institutes of Health (NIH). It was also funded by a grant (R25DK096944) from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK)–supported Academic Pediatric Association (APA) Research in Academic Pediatrics Investigator Development (RAPID) program. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or the APA.

Authors' Contributions

JCES contributed to conceptualization, data analysis, and writing. MB contributed to methodology, data analysis, and writing. ADS contributed to conceptualization, methodology, and writing. CPW and RDA contributed to data analysis and writing. OS, AS, PS, PC, and LY contributed to conceptualization and writing.

Conflicts of Interest

JCES is a consultant for Dexcom, Glooko, and Sanofi. The authors have no conflicts of interest to declare.

Checklist 1
CONSORT-EHEALTH (V 1.6.1) checklist.
[PDF File, 1229 KB - formative_v9i1e70341_app1.pdf]

References

1. Obesity and overweight. World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight [accessed 2022-12-18]
2. Weaver RG, Brazendale K, Hunt E, Sarzynski MA, Beets MW, White K. Disparities in childhood overweight and obesity by income in the United States: an epidemiological examination using three nationally representative datasets. Int J Obes 2019 Jun;43(6):1210-1222. [doi: 10.1038/s41366-019-0331-2]
3. Skinner AC, Ravanbakht SN, Skelton JA, Perrin EM, Armstrong SC. Prevalence of obesity and severe obesity in US children, 1999-2016. Pediatrics 2018 Mar;141(3):e20173459. [doi: 10.1542/peds.2017-3459] [Medline: 29483202]
4. Pietrobelli A, Rugolotto S, Cristofaro PD, Malavolti M. Pediatric obesity: looking into treatment. Nutrients 2009 Feb;1(2):197-209. [doi: 10.3390/nu1020197] [Medline: 22253978]
5. Kouvari M, Karipidou M, Tsiampalis T, et al. Digital health interventions for weight management in children and adolescents: systematic review and meta-analysis. J Med Internet Res 2022 Feb 14;24(2):e30675. [doi: 10.2196/30675] [Medline: 35156934]
6. Colom A. The digital divide. Inf Commun Soc 2020 Sep 18;23(11):1706-1708. [doi: 10.1080/1369118X.2020.1781916]
7. Estacio EV, Whittle R, Protheroe J. The digital divide: examining socio-demographic factors associated with health literacy, access and use of internet to seek health information. J Health Psychol 2019 Oct;24(12):1668-1675. [doi: 10.1177/1359105317695429] [Medline: 28810415]
8. Fridsma DB. AMIA. AMIA response to FCC notice on accelerating broadband health tech availability. 2017. URL: https://www.amia.org/sites/default/files/AMIA-Response-to-FCC-Notice-on-Accelerating-Broadband-Health-Tech-Availability [accessed 2022-12-18]
9. Ramsetty A, Adams C. Impact of the digital divide in the age of COVID-19. J Am Med Inform Assoc 2020 Jul 1;27(7):1147-1148. [doi: 10.1093/jamia/ocaa078] [Medline: 32343813]

XSL•FO
RenderX

10. Sieck CJ, Sheon A, Ancker JS, Castek J, Callahan B, Siefer A. Digital inclusion as a social determinant of health. NPJ Digit Med 2021 Mar 17;4(1):52. [doi: 10.1038/s41746-021-00413-8] [Medline: 33731887]
11. Ha AS, Ng JYY, Lonsdale C, Lubans DR, Ng FF. Promoting physical activity in children through family-based intervention: protocol of the "Active 1 + FUN" randomized controlled trial. BMC Public Health 2019 Feb 20;19(1):218. [doi: 10.1186/s12889-019-6537-3] [Medline: 30786902]
12. Bean MK, Caccavale LJ, Adams EL, et al. Parent involvement in adolescent obesity treatment: a systematic review. Pediatrics 2020 Sep;146(3):e20193315. [doi: 10.1542/peds.2019-3315] [Medline: 32839242]
13. Espinoza J, Chen A, Orozco J, Deavenport-Saman A, Yin L. Effect of personal activity trackers on weight loss in families enrolled in a comprehensive behavioral family-lifestyle intervention program in the Federally Qualified Health Center setting: a randomized controlled trial. Contemp Clin Trials Commun 2017 Sep;7:86-94. [doi: 10.1016/j.conctc.2017.06.004] [Medline: 29308436]
14. Espinoza JC, Chen AM, Deavenport-Saman A, et al. Wearable devices decrease attrition among families participating in an obesity intervention at a Federally Qualified Health Center. J Health Care Poor Underserved 2021;32(2S):13-24. [doi: 10.1353/hpu.2021.0048]
15. Defining Childhood Overweight and Obesity. Center for disease control and prevention. URL: https://www.cdc.gov/obesity/childhood/defining.html [accessed 2022-12-07]
16. Koorts H, Salmon J, Timperio A, et al. Translatability of a wearable technology intervention to increase adolescent physical activity: mixed methods implementation evaluation. J Med Internet Res 2020 Aug 7;22(8):e13573. [doi: 10.2196/13573] [Medline: 32763872]
17. Hartman SJ, Nelson SH, Weiner LS. Patterns of Fitbit use and activity levels throughout a physical activity intervention: exploratory analysis from a randomized controlled trial. JMIR Mhealth Uhealth 2018 Feb 5;6(2):e29. [doi: 10.2196/mhealth.8503] [Medline: 29402761]
18. Vogels EA. Digital divide persists even as lower-income americans make gains in tech adoption. Pew Research Center. URL: https://www.pewresearch.org/fact-tank/2019/05/07/digital-divide-persists-even-as-lower-income-americans-make-gains-in-tech-adoption/ [accessed 2020-10-19]
19. Chia GLC, Anderson A, McLean LA. Behavior change techniques incorporated in fitness trackers: content analysis. JMIR Mhealth Uhealth 2019 Jul 23;7(7):e12768. [doi: 10.2196/12768] [Medline: 31339101]
20. Mercer K, Li M, Giangregorio L, Burns C, Grindrod K. Behavior change techniques present in wearable activity trackers: a critical analysis. JMIR Mhealth Uhealth 2016 Apr 27;4(2):e40. [doi: 10.2196/mhealth.4461] [Medline: 27122452]
21. Lewis ZH, Lyons EJ, Jarvis JM, Baillargeon J. Using an electronic activity monitor system as an intervention modality: a systematic review. BMC Public Health 2015 Jun 24;15(1):585. [doi: 10.1186/s12889-015-1947-3] [Medline: 26104189]
22. Barlow SE, Committee E. Expert committee recommendations regarding the prevention, assessment, and treatment of child and adolescent overweight and obesity: summary report. Pediatrics 2007 Dec;120 Suppl 4:S164-S192. [doi: 10.1542/peds.2007-2329C] [Medline: 18055651]
23. Boutelle KN, Rhee KE, Liang J, et al. Effect of attendance of the child on body weight, energy Intake, and physical activity in childhood obesity treatment: a randomized clinical trial. JAMA Pediatr 2017 Jul 1;171(7):622-628. [doi: 10.1001/jamapediatrics.2017.0651] [Medline: 28558104]
24. Chai LK, Collins C, May C, Brain K, Wong See D, Burrows T. Effectiveness of family-based weight management interventions for children with overweight and obesity: an umbrella review. JBI Database System Rev Implement Rep 2019;17(7):1341-1427. [doi: 10.11124/JBISRIR-2017-003695]

Abbreviations

%BMIp95: BMI percent of the 95th percentile
AMIA: American Medical Informatics Association
API: application programming interface
CDC: Centers for Disease Control and Prevention
CONSORT: Consolidated Standards of Reporting Trials
FQHC: Federally Qualified Health Center
LEP: limited English proficiency
PAT: personal activity tracker

XSL•FO
RenderX

Edited by A Mavragani; submitted 19.12.24; peer-reviewed by J Rausch, S Lazorick; revised version received 04.08.25; accepted 05.09.25; published 05.11.25.

Please cite as:
Espinoza Salomon JC, Babaei M, Deavenport-Saman A, Solomon O, Wee CP, Durazo-Arvizu R, Sikder A, Shah P, Castillo P, Yin L Personal Activity Trackers and Family Engagement in a Pediatric Obesity Intervention: Randomized Controlled Trial JMIR Form Res 2025;9:e70341
URL: https://formative.jmir.org/2025/1/e70341
doi:10.2196/70341

© Juan Carlos Espinoza Salomon, Mahsa Babaei, Alexis Deavenport-Saman, Olga Solomon, Choo Phei Wee, Ramon Durazo-Arvizu, Abu Sikder, Payal Shah, Patricia Castillo, Larry Yin. Originally published in JMIR Formative Research (https://formative.jmir.org), 5.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Teenage Body Image Perception, Body-shaping Behavior, and Body Composition With Respect to Use of “Fitspiration”: Exploratory Investigation Study

Elisabeth Höld[1], PhD; Theres Rathmanner[1], PhD; Mario Heller[2], PhD

[1]Institute of Health Sciences, Department of Health, St. Pölten University of Applied Sciences, Campus-Platz 1, St. Pölten, Austria

[2]Institute of Creative\Media/Technologies, Department of Media and Digital Technologies, St. Pölten University of Applied Sciences, St. Pölten, Austria

Corresponding Author:
Elisabeth Höld, PhD
Institute of Health Sciences, Department of Health, St. Pölten University of Applied Sciences, Campus-Platz 1, St. Pölten, Austria

Abstract

Background: The social media trend #fitspiration aims to positively impact its users’ health, but studies have shown detrimental effects, as it mainly involves stereotypical and barely achievable body images and health behaviors. During puberty, adolescents form their identity and body image, making it essential to examine social media’s influence on their health.

Objective: The exploratory investigation study, part of the mixed methods study in the project FIVE (#Fitspiration Image Verification), sought initial insights into how #fitspiration consumption may affect adolescents’ body image perception, dieting, and exercise behaviors.

Methods: In total, 86 adolescents (N=310, age range 14-18 years; n=42 females) attending upper secondary schools in Eastern Austria took part in an online questionnaire concerning the use of #fitspiration, body image perception, and body-shaping behaviors (dieting and working out) and bioelectrical impedance measurements to assess body composition. Participants have been classified as #fitspiration users (n=27), nonusers (n=51), and former users (n=8). We compared body image perceptions and body-shaping behaviors of #fitspiration user groups sex-specifically with the fat mass index (FMI) and the fat-free mass index (FFMI).

Results: The results indicated that the amount of muscularity was of greater importance for all participants than thinness because, and even though the majority were of normal weight (76/86, 88%) and had an average to high FFMI (62/86, 92%), only 6% (5/86) of all participants thought that their amount of muscle mass was appropriate, while 43% (37/86) of all participants rated their body weight as okay. However, this outcome seemed to be of greater significance for #fitspiration users, especially males. While male participants seemed to be more dissatisfied with their amount of muscularity and worked out more often (female: 24/42, 57%, male: 35/44, 80%), female participants seemed to be more dissatisfied with their body weight and were dieting (female: 10/42, 24%; male: 6/44, 14%). Generally, none of the #fitspiration users (0/27, 0%) answered that her or his amount of muscle mass was okay, although all of them had an average or high FFMI (27/27, 100%). Participants assessed their body weight and body fat more precisely (54/86, 63% matched FMI) than their muscle mass (27/86, 31% matched FFMI).

Conclusions: Our findings regarding the body ideals of adolescents are in line with the results of large-scale cross-sectional studies, indicating that they aspire to achieve a toned (and, for females, thin) body. The fact that none of the #fitspiration users were satisfied with their level of muscle mass raises the question of whether #fitspiration may perpetuate this ideal. The line between eating disorders and body dysmorphia might be very thin, especially during puberty. Therefore, our findings can be used to raise awareness of the speculated impact of #fitspiration on adolescents as a source for body ideals and consequently for body image perception and body-shaping behavior.

(JMIR Form Res 2025;9:e70964) doi:10.2196/70964

KEYWORDS

social media; #fitspiration; exploratory investigation study; adolescents; body image perception; body-shaping behavior; dieting; working out

Introduction

Adolescent Body Image Perception

Body image is a multidimensional construct based on perceptions and attitudes toward one’s body and physical appearance [1]. Body image perception “includes body size assessment (how a person perceives his or her body), body attractiveness estimation (what is the type of body that a person considers most attractive), and perceptions related to one’s own body shape and size” [2,3]. In Western countries, the focus is

XSL•FO
RenderX

on body shape and weight and the individual's perceived personal responsibility for it [4].

During puberty, adolescents must adjust to significant bodily changes. By contrast to cognitive development, physical changes, such as the growth of a beard or breasts, are visible to everyone, including family, peer groups, and strangers if pictures are posted on social media.

Therefore, these physical changes are accompanied by a broad range of emotions and shaped by social and sociocultural influences [5]. Consequently, adolescents are a vulnerable group, and there is growing evidence that body image, body satisfaction, and body dissatisfaction affect their physical and psychosocial health, leading to, for example, unhealthy eating behaviors, lack of physical activity [6], eating disorders, or depressive symptoms [7,8].

Impact of Visually Oriented Social Media on Body Image Perception

Photos and videos are of great significance on social media, as they fulfill the function of self-representation. Appearance is crucial for young people, and their body is part of this self-presentation. It is an expression of lifestyle and serves as a means of differentiation and belonging to groups [9]. Posed and edited social media pictures provide the unique possibility for appearance-based social comparison [10]. Consequently, both—the active use like posting selfies and passive use, like viewing pictures of others on social media—show clear associations with body dissatisfaction [7,11,12], depressive and disordered eating symptoms, and anxiety [7,8].

During the past years, visually oriented social media like Instagram (Meta Platforms, Inc) have become more popular and seem to have a greater impact on appearance comparisons, body image and consequently body image perception, body dissatisfaction [13-15], and eating disorders [16] than text-centered ones like Facebook (Meta Platforms, Inc [17,18]).

#Fitspiration

The social media trend #fitspiration [19] mainly relates to an athletic appearance and how it can be achieved through diet and exercise [20]. Although #fitspiration aims to have a positive impact on health, it is also associated with detrimental effects, as it mainly involves stereotypical and barely achievable bodily standards [21].

Therefore, it has a high potential for physical appearance comparison. A recent review shows that #fitspiration images have a negative effect on body image, which can be linked, firstly, to the demand for self-control implied in #fitspiration images and, secondly, to the discomfort experienced by individuals when they try to achieve these unrealistic body images through hard training and restricted dietary behavior [22]. An Italian quasi-experimental intensive longitudinal study examined the impact of daily exposure to #fitspiration Instagram content on young women's mood and body image and was associated with the highest rates of growth of negative mood and appearance comparison [23]. Furthermore, #fitspiration is also associated with a greater internalization of the muscular ideal, which further supports body dissatisfaction. Consequently, #fitspiration can lead to excessive exercising and disordered eating [22]. On the other hand, an Austrian-controlled longitudinal experiment with young adult women observed a positive impact of #fitspiration content on weight satisfaction and healthy eating behavior over 22 days compared to content from a nonbody-focused influencer [24].

Most of the literature addresses the influence of #fitspiration on young adults. Furthermore, mainly women, especially educated women, were study participants [22]. To date, there seems to be limited research on the impact of this trend on adolescents, which is surprising given that the teenage years are highly relevant for the development of body image perception and health behavior. This makes adolescents highly vulnerable to potential negative influences. Besides, most of the published study results describe changes in body image or body dissatisfaction in relation to #fitspiration use only [22,25-27].

To our knowledge, there is no study in which the gender-specific influence of #fitspiration on the body image perception of adolescents has been linked to a set of measured body composition parameters. Consequently, the purpose of this study is to describe participants' subjective assessment of body image perception and body-shaping behavior together with objective measurements of body composition parameters in adolescents. Therefore, we can objectively rate the self-assessment of teenage body image perception against the background of #fitspiration use and gender.

Methods

The project #Fitspiration Image Verification

This study was part of the interdisciplinary project FIVE (#Fitspiration Image Verification [28]), which was conducted by St. Pölten University of Applied Sciences in Austria. Spanning from 2021 to 2024, it involved the Institute of Health Sciences and the Institute of Creative Media/Technologies, an external gender expert (Bettina Prokop), and an educational publisher (Hölzel Verlag GmbH). FIVE aimed to gain first insights into the effects of #fitspiration on adolescents attending upper secondary schools and developed a blended-learning course for upper secondary schools with an image forensics tool as the core element. The course aims to empower pupils to navigate critically and competently through social media. A detailed description of the development of the blended-learning course is published elsewhere [29].

As a first step of the project FIVE, an in-depth exploratory mixed methods study was conducted aiming to analyze the social media phenomenon #fitspiration and how it might affect adolescents. This mixed methods study consisted of 3 parts. The first part was a #fitspiration content analysis on Instagram, the second part was an online questionnaire, and the third part consisted of on-site anthropometric and bioelectrical impedance analysis (BIA) measurements, focus group discussions, as well as eye-tracking and biosignal measurements. This paper focuses on the body image perception-related part of the online questionnaire (part 2) and the corresponding parts of the on-site anthropometric and BIA measurements (part 3). Data collection

XSL•FO
RenderX

took place from March to September 2022 at participating schools during regular teaching time. The target population consisted of pupils from upper secondary schools in Eastern Austria (Lower Austria and Vienna). The sampling frame was the network of the St. Pölten University of Applied Sciences, where schools expressed their interest in collaborating on this research project.

Online Questionnaire

The online questionnaire had 3 main topics next to sociodemographic information. These topics were social media and #fitspiration use, body image perception, and body-shaping behavior, which are all explained in detail below.

Sociodemographic Information

We asked participants about their age, gender, living situation, and living area. We used the Family Affluence Scale (FAS) [30,31] to describe the socioeconomic status. Based on the United Nations Economic Commission for Europe (UNECE) recommendations [32], we defined migration background as both parents having been born outside of Austria.

Social Media and #Fitspiration Use

We asked the participants to rate the frequency of their social media use, both generally and with regard to #fitspiration specifically, the latter one based on the categories of [33] (1="I have never thought about following such content or influencers"; 2="I have thought about following such content or influencers. But I haven't done it yet"; 3="I thought about following such content or influencers. But I decided against it"; 4="I follow such content or influencers and will continue to do so"; 5="I used to follow such content or influencers, but I stopped").

Body Image Perception

Based on the "Health Behaviour in School-aged Children (HBSC)" study [34], we asked participants how they perceived their own body weight on a 5-point scale ranging from 1 ("much too thin") to 5 ("much too fat"). Based on this item characteristic, we asked participants to likewise rate their muscle mass on a 5-point scale ranging from 1 ("far too unmuscular") to 5 ("far too muscular").

Body-Shaping Behavior

Based on the HBSC study [34], we asked participants whether they were currently on a diet or did something else to lose weight ("Yes"; "No, I need to put on weight"; "No, my weight is fine"; "No, but I should lose some weight"). Based on this 4-item characteristic, we asked participants whether they were currently working out or doing something else to build muscle ("Yes"; "No, but I should build muscle"; "No, my amount of muscle mass is okay"; "No, because I should reduce muscle").

The questionnaire consisted of 38 questions and could be answered in about 20-30 minutes. The usability, technical functionality, and completion time of the electronic questionnaire were tested by 3 students from the bachelor program in dietetics, and the revised final version was carried out online using LimeSurvey (version 3.27; LimeSurvey GmbH). Therefore, the project manager created individualized links for each participant. These links were sent to the participants via email by the project manager or were distributed by the school app or manually in class by the teacher. The questionnaires were available for 3 months and were answered during school time. LimeSurvey data export was used to store raw data as an SPSS (IBM Corp) file for further analysis.

Anthropometric and BIA Measurements

The anthropometric and BIA measurements determined with the seca mBCA 555 (seca GmbH & Co. KG) took place in the morning in schools, in a fasting state, and with light clothing. The device measures body weight with a scale, body height with ultrasound length measurement, and body composition using BIA in 1 step. BIA analysis is an electrical resistance measurement on the human body, which is based on the different electrical properties of biological tissues [35,36].

We calculated the BMI by using body weight and body height (BMI=body weight [kg] / body height [m²]) and classified it according to Kromeyer-Hauschild et al [37]. This classification is based on German data and displays sex-specific BMI-for-age percentile curves from the age group of 15-79 years.

Due to the known weaknesses of the BMI, namely the missing discrimination of body fat and muscle mass [38], we calculated the height-normalized fat mass index (FMI=FM/m²) and fat-free mass index (FFMI=FFM/m²) based on the measured fat mass (FM), fat-free mass (FFM), and body height. The use of the FMI and FFMI instead of absolute values or percentages is recommended to avoid ambiguities when measuring FM and FFM themselves [39]. Consequently, the FMI (the higher, the more FM in relation to body height) and the FFMI (the higher, the more FFM in relation to body height) have been suggested to be better indicators when screening for diseases like obesity [39,40] or cardiometabolic risk factors [41] as well as physical activity [42] in adolescents compared to the BMI alone. Furthermore, a recent study showed that body composition parameters like FM are better predictors of body size dissatisfaction than the BMI [43].

We classified the FMI and the FFMI according to the MoMo (Motorik-Modul) study (Longitudinal Study of the German Health Interview and Examination Survey of Children and Adolescents; Studie von Kindern und Jugendlichen in Deutschland) [44], which published German sex-specific FMI-for-age and FFMI-for-age percentiles for the age group of 4-24 years. The study group recommends using the percentile curves of the normal weight sample as reference values for body composition assessment. Furthermore, they suggest using the 10th percentile as the cutoff value for low and the 90th percentile cutoff value for high FMI and FFMI. Consequently, the FMI and the FFMI between the 10th and 90th percentiles can be considered as normal.

Statistical Analysis

Data were analyzed using SPSS (version 29.0.2.0; IBM Corp).

To illustrate descriptive parameters, the arithmetic mean and SD were stated for metric data. Interval variables or nominal-scaled variables were shown as frequencies. Due to the innovative nature of the approach and the exploratory design,

XSL•FO
RenderX

no sample size calculations were performed and no a priori hypotheses were formulated.

The descriptive research included the descriptive analysis of the #fitspiration users and the following four questions versus corresponding body composition parameters (MoMo study normal weight [44]): (1) body image perception (body weight and body fat) versus FMI, (2) body-shaping behavior (weight reduction) versus FMI, (3) body image perception (muscularity) versus FFMI, and (4) body-shaping behavior (muscle building) versus FFMI.

Ethical Considerations

The study protocol was submitted to the ethics committee of Vienna (MA 15-EK/21‑145-VK_NZ; June 16, 2021) and Lower Austria (email confirmation; October 15, 2021) and both stated that FIVE was not subject to mandatory submission.

FIVE was based on national and international ethical guidelines, above all the World Medical Association Declaration of Helsinki [45]. Participating adolescents received detailed written and oral information about the project as well as a comprehensive informed consent document. Furthermore, and in accordance with the recommendations of the Forum Ethics Commissions of Austria [46], we prepared separate informed consent forms for parents or legal guardians, who were informed comprehensively about the project.

The participants' data were pseudonymized and assigned a fixed identification code. This code was stored safely, accessible only by the study coordinator and central staff members. The key for the coding was kept separately and inaccessible by the study coordinator. Data were stored and analyzed via the statistical software package SPSS at servers of the St. Pölten University of Applied Sciences, only used for this special study and not passed on to third parties. The handling of the data complied with the European General Data Protection Regulation, the Austrian Data Protection Act, and the recommendations of the Ethics Committee of Lower Austria.

Since the pupils participated in the project FIVE as part of their regular lessons, they received no compensation. Pupils who did not want to participate in the study were taught in classes that did not participate in the project FIVE. The teachers also participated in the project as part of their professional duties and, in accordance with Austrian law, were not allowed to receive additional payment. Nevertheless, and as a way of expressing gratitude, participating pupils received their BIA measurement results and were able to participate in workshops on the significance and interpretation of BIA measurements as well as gender and beauty standards on social media conducted by project staff members. The pupils had 2 options to choose from: either an online questionnaire only or a questionnaire plus anthropometric and BIA measurements.

Results

Participants

Of the 6 participating upper secondary schools with about 1800 pupils, 310 pupils started with the online questionnaire (participation rate: approximately 20%). Among them, 261 pupils completed it (questionnaire completion rate: 84%) with an average completion time of 23.7 (SD 16) minutes. Overall, 108 out of 310 teenagers participated in the anthropometric and BIA measurements (anthropometric and BIA measurements participation rate is 35%). We combined data from the online questionnaire and the anthropometric and BIA measurements so that we had 87 complete datasets of questionnaire and anthropometric and BIA measurement. We excluded 1 person who identified as diverse because including 1 person in a separate gender category would not provide enough information. In addition, the used percentiles of FMI, FFMI [44], and BMI [37] are currently binary. Finally, we included 86 adolescents for further analysis (final inclusion rate turned out to be 28%). Figure 1 shows the participants' flowchart.

XSL•FO
RenderX

Figure 1. Participants' flowchart.

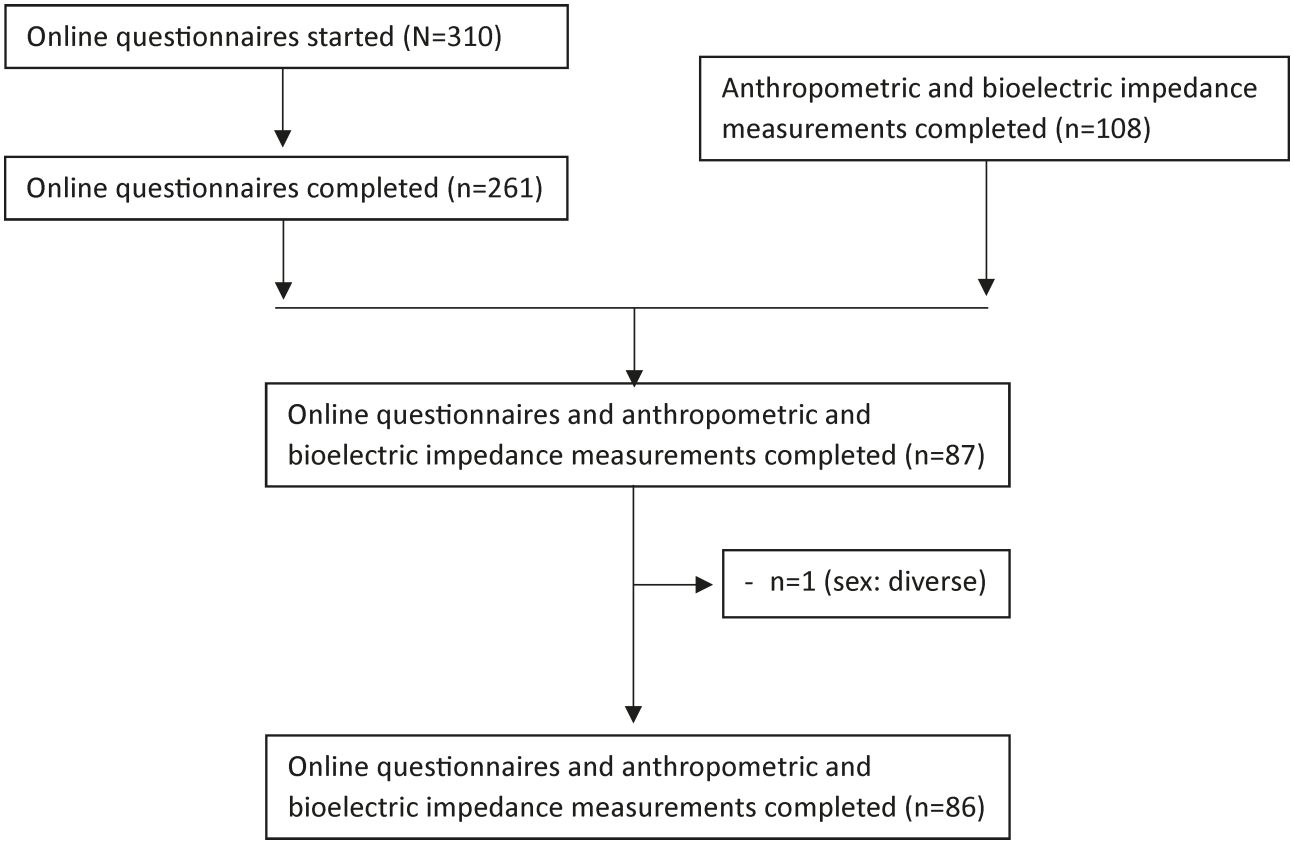

In the next step, to examine the possible link between #fitspiration use and body image perception and body-shaping behaviors, we recoded the results from the questionnaire on #fitspiration use as follows: we defined participants as #fitspiration users if they selected the answer 4 ("I follow such content or influencers and will continue to do so"). #fitspiration nonusers, on the other hand, were those who ticked one of the first three answers ("I have never thought about following such content or influencers," "I have thought about following such content or influencers. But I havent done it yet," and "I thought about following such content or influencers. But I decided against it"). We classified those who selected the answer 5 ("I used to follow such content or influencers, but I stopped") as former #fitspiration users.

In the last step, we summarized the 5 categories (2 negative, 1 neutral, and 2 positive) of the 2 questions regarding body image perception to 3 categories each (1 negative, 1 neutral, and 1 positive) because of distribution frequencies.

In total, 42/86 (49%) participants were female and 44/86 (51%) were male, aged 14-18 years (mean 16, SD 0.9 years). Apart from 1 participant, the sample lived in average-income (20/86, 23%) or well-off (65/86, 76%) families. The majority had no migration background (79/86, 92%). The 3 most used social media platforms, this means that social media were used multiple times per day, were WhatsApp (Meta Platforms, Inc; 67/86, 78%), Snapchat (Snap Inc.; 66/86, 77%), and Instagram (61/86, 71%).

With almost equal gender distribution, most of the participants did not use #fitspiration (59/86, 69%). Within the group of #fitspiration users (27/86, 31%), male participants were dominant (20/27, 74%). Former #fitspiration users represented the minority of the participants (8/86, 9%) and were mainly females (6/8, 75%).

Most participants were of normal weight (76/86, 88%; female: 37/42, 88%; male: 39/44, 89%). Furthermore, 6% (5/86; female: 2/42, 5%; male: 3/44, 7%) of participants were classified as overweight or obese, and 6% (5/86; female: 3/42, 7%; male: 2/44, 5%) of participants as severely underweight or underweight. The majority of participants had an FMI and an FFMI within the normal range (FMI: 51/86, 59%; female: 30/42, 71%; male: 21/44, 48%; FFMI: 50/86, 58%; female: 30/41, 71%; male: 20/44, 45%). Furthermore, 24/86 (28%; female: 4/42, 10%; male: 20/44, 46%) had a low FMI, and 29/86 (34%; female: 6/42, 14%; male: 23/44, 52%) had a high FMMI. Only 13% (11/86; female: 8/42, 19%; male: 3/44, 7%) had a high FMI, while 8% (7/86; female: 6/42, 14%; male: 1/44, 2%) had a low FFMI (see Table 1).

XSL•FO
RenderX

Table . Sample description inclusive of FMI and FFMI of the Motorik-Modul Study; sex-specific FMI and FFMI of normal-weight participants aged 14-17 years.

Variables	User (n=27)		Nonuser (n=51)		Former user (n=8)	
	Female (n=7)	Male (n=20)	Female (n=29)	Male (n=22)	Female (n=6)	Male (n=2)
Age (years), n (%)						
14	—[a]	1 (5)	3 (10)	1 (4)	—	—
15	2 (29)	2 (10)	11 (38)	5 (23)	—	—
16	1 (14)	10 (50)	11 (38)	8 (36)	3 (50)	—
17	4 (57)	7 (35)	3 (10)	5 (23)	3 (50)	1 (50)
18	—	—	1 (4)	3 (14)	—	1 (50)
Age (years), mean (SD)	16.3 (1)	16.2 (0.8)	15.6 (0.9)	16.2 (1.1)	16.5 (0.5)	17.5 (0.7)
FAS[b], n (%)						
Low	—	—	1 (3)	—	—	—
Medium	1 (14)	3 (15)	10 (35)	6 (27)	—	—
High	6 (86)	17 (85)	18 (62)	16 (73)	6 (100)	2 (100)
Migration background, n (%)						
Yes	—	1 (5)	2 (7)	3 (14)	1 (17)	—
No	7 (100)	19 (95)	27 (93)	19 (86)	5 (83)	2 (100)
Social media use[c], n (%)						
WhatsApp	7 (100)	16 (80)	21 (72)	15 (68)	6 (100)	2 (100)
YouTube	—	9 (45)	10 (35)	8 (36)	—	2 (100)
Instagram	7 (100)	17 (85)	19 (66)	14 (64)	4 (67)	—
Snapchat	7 (100)	15 (75)	23 (79)	14 (64)	6 (100)	1 (50)
TikTok	6 (86)	14 (70)	17 (59)	13 (59)	5 (83)	—
Pinterest	—	—	2 (7)	—	—	—
Facebook	1 (14)	—	—	—	—	—
others	1 (14)	4 (20)	2 (7)	6 (27)	1 (17)	1 (50)
BMI categorization, n (%)						
Severely underweight, underweight	—	—	3 (10)	2 (9)	—	—
Normal weight	7 (100)	18 (90)	24 (83)	19 (86)	6 (100)	2 (100)
Overweight, obesity	—	2 (10)	2 (7)	1 (5)	—	—
FMI[d] categorization, n (%)						
<10th percentile	—	8 (40)	4 (14)	12 (55)	—	—
10th-90th percentile	4 (57)	11 (55)	20 (69)	8 (36)	6 (100)	2 (100)
>90th percentile	3 (43)	1 (5)	5 (17)	2 (9)	—	—
FMI (kg/m^2), mean (SD)	6.3 (1.1)	2.8 (1.3)	5.1 (1.7)	3 (1.8)	5.7 (1.1)	5.2 (0.2)
FMI_{MoMo} (kg/m^2), mean (SD)	5.1 (1.4)	3.9 (1.3)	5.1 (1.4)	3.9 (1.3)	5.1 (1.4)	3.9 (1.3)
FFMI[e] categorization, n (%)						

XSL•FO
RenderX

Variables	User (n=27)		Nonuser (n=51)		Former user (n=8)	
	Female (n=7)	Male (n=20)	Female (n=29)	Male (n=22)	Female (n=6)	Male (n=2)
<10th percentile	—	—	5 (17)	1 (5)	1 (17)	—
10th-90th percentile	7 (100)	8 (40)	19 (66)	11 (50)	4 (66)	1 (50)
>90th percentile	—	12 (60)	5 (17)	10 (45)	1 (17)	1 (50)
FFMI (kg/m^2), mean (SD)	15.5 (0.8)	18.9 (1.4)	15.2 (1.2)	18.2 (1.6)	15.4 (1.4)	18.7 (1.9)
$FFMI_{MoMo}$ (kg/m^2), mean (SD)	15.6 (1.2)	16.9 (1.2)	15.6 (1.2)	16.9 (1.2)	15.6 (1.2)	16.9 (1.2)

[a]Not applicable.

[b]FAS: Family Affluence Scale (Low (score=0,1,2) indicates low affluence, medium (score=3,4,5) indicates middle affluence, and high (score=6,7,8,9) indicates high affluence).

[c]Social media use indicates multiple times per day; multiple answers possible.

[d]FMI: fat mass index.

[e]FFMI: fat-free mass index.

Body Image Perception: Body Weight/Fat

Generally, 63% (54/86) of participants rated their body weight and body fat similar to the FMI. Participants with a low or average FMI (14/75, 19%) considered themselves too fat. Those with an average FMI (7/62, 11%) thought that they were too thin. In contrast to this, adolescents with a normal FMI had a better self-perception, and 65% (33/51) evaluated themselves as about normal weight.

Observed gender-related values are as follows: 75% (3/4) of female and 50% (10/20) of male teens with a low FMI overestimated their own body weight and body fat and thought that they were about normal weight or even too fat. Around 27% (8/30) of female participants and 14% (3/21) of male participants with a normal FMI considered themselves too fat. Overall, 88% (7/8) of female and 100% (3/3) of male participants with a high FMI rated themselves as too fat.

The association between the use of #fitspiration and body image perception appeared to be negligible (refer to Figure 2).

Figure 2. Body image perception (body weight and body fat) in relation to #fitspiration user status and FMI percentiles. FMI: fat mass index.

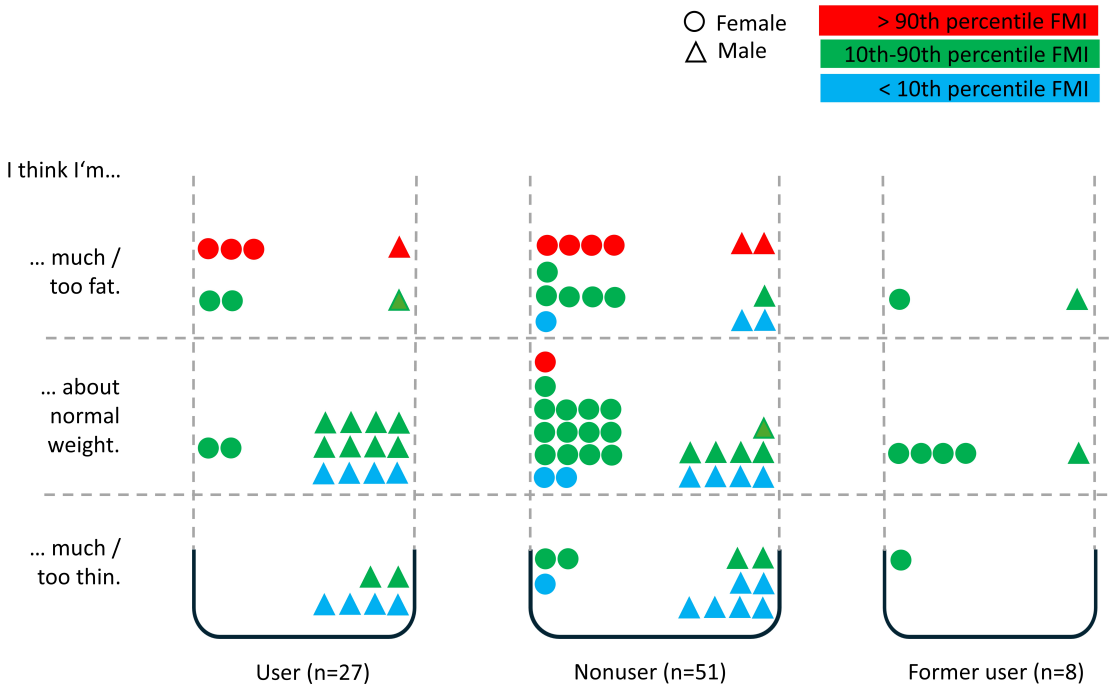

Body-shaping Behavior: Weight Reduction

Generally, 43% (37/86) rated their body weight as fine (no related body-shaping behavior needed). Around 19% (16/86) of all participants implemented weight reduction behavior, and the same percentage thought that they should reduce body weight. Overall, 20% (17/86) of all participants thought that they should gain body weight.

XSL•FO
RenderX

Observed gender-related values are as follows: 24% (10/42) of female participants and 14% (6/44) of male participants implemented weight reduction behavior. Within the group of dieting participants (16/86, 19%), 40% (4/10) of females and 100% (6/6) of males were dieting although they had a low or normal FMI. In total, 5% (2/42) of female participants and 34% (15/44) of male participants were not on a diet because they thought that they should gain body weight. Altogether, 50% (21/42) of female participants and 36% (16/44) of male participants mentioned that they were not on a diet because their body weight was okay. Around 21% (9/42) of female participants and 16% (7/44) of male participants answered that while they were not on a diet, they should reduce body weight.

Regarding gender and #fitspiration use, 30% (8/27) of #fitspiration users—especially females (5/7, 71%; male: 3/20, 15%)—implemented weight reduction behavior. In comparison, no former #fitspiration user mentioned dieting, and 16% (8/51, female: 5/29, 17%; male: 3/22, 14%) of all nonusers were on a diet. #fitspiration nonusers and former users were more likely to state that they did not diet because they considered their body weight as fine (female user: 1/7, 14%; male user: 7/20, 35%; female nonuser: 17/29, 59%; male nonuser: 9/22, 41%; female former user: 3/6, 50%; male former user: 0/2, 0%; Figure 3).

Figure 3. Body-shaping behavior (weight reduction) in relation to #fitspiration user status and FMI percentiles. FMI: fat mass index.

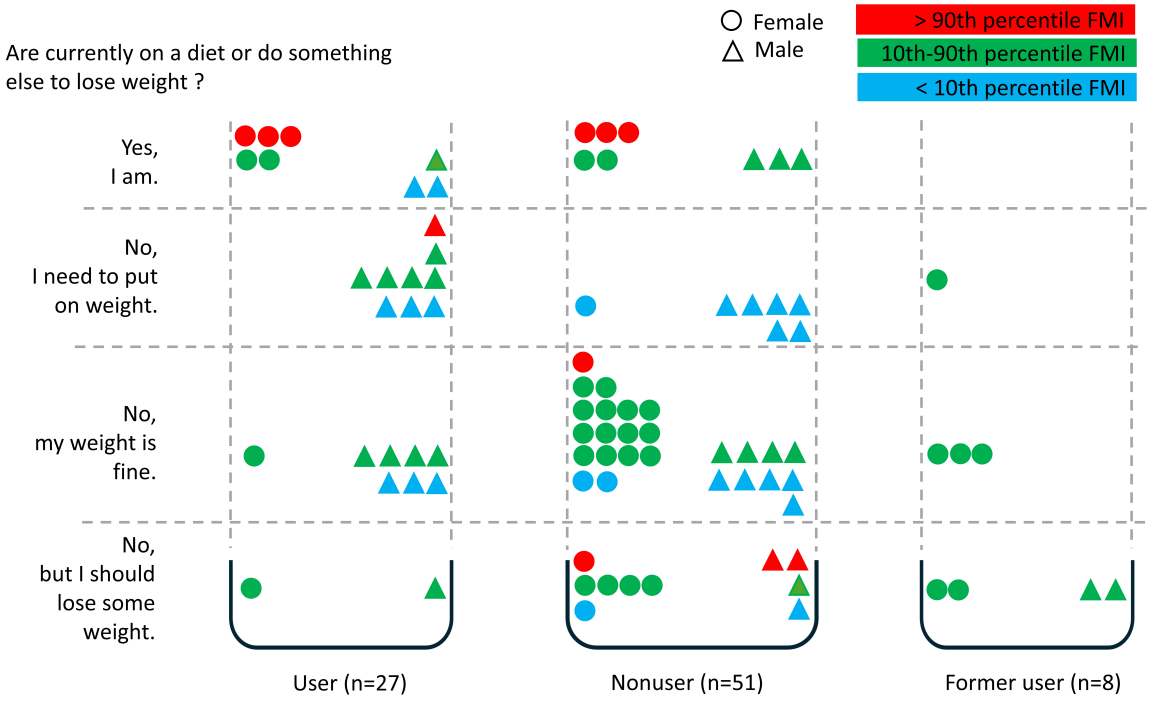

Body Image Perception: Muscularity

Generally, 31% (27/86) of all participants rated their muscularity according to their FFMI. A total of 55% (47/86) of all participants rated themselves as too unmuscular, although they had an FFMI within the normal range or above. All participants (86/86, 100%) with a low FFMI assessed themselves as too unmuscular (7/7).

Concerning observed gender-related values, 40% (17/42) of female participants versus 23% (10/46) of male participants evaluated their FFMI in accordance with the measurement results.

Regarding gender and #fitspiration use, #fitspiration users (27/86, 31%) all had an FFMI within the normal range or above (female: 7/7, 100%; male: 20/20, 100%); but also, nonusers (female: 24/29, 83%; male: 21/22, 95%) and former users (female: 5/6, 83%; male: 2/2, 100%) showed similar values (refer to Figure 4).

XSL•FO
RenderX

Figure 4. Body image perception (muscularity) in relation to #fitspiration user status and FFMI percentiles. FFMI: fat-free mass index.

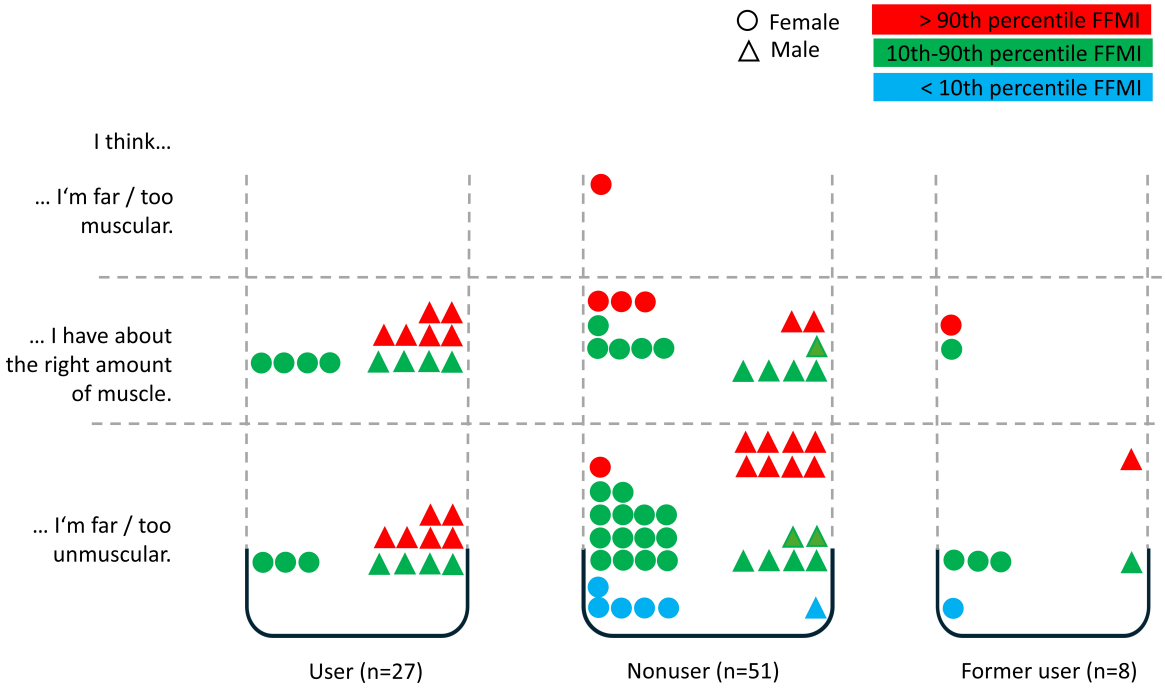

Body-shaping Behavior: Muscle Building

Generally, 69% (59/86) of all participants worked out or did something else to build muscle. Around 26% (22/86) of all participants thought that they should build muscle, 6% (5/86) of participants stated that their amount of muscle mass was okay, and no participant mentioned the wish to reduce muscle.

Observed gender-related values are as follows: Male participants were more likely to work out than female ones (female: 24/42, 57%; male: 35/44, 80%). On the other hand, female participants more often did not work out although they thought that they should build muscle (female: 15/42, 36%; male: 7/44, 16%).

Regarding gender and #fitspiration use, almost all #fitspiration users (26/27, 96%) worked out (female: 6/7, 86%; male: 20/20, 100%), and all users had an FFMI within the normal range or even higher. None of the #fitspiration users (0/27, 0%) answered that his or her amount of muscle mass was okay. We observed lower rates of those who worked out to build muscle within the other user groups, in particular nonusers (29/51, 57%; female: 15/29, 52%; male: 14/22, 64%) and former users (4/8, 50%; female: 3/6, 50%; male: 1/2, 50%). Overall, 35% (18/51; female: 12/29, 41%; male: 6/22, 27%) of nonusers and 38% (3/8; female: 2/6, 33%; male: 1/2, 50%) of former users thought that they should build muscle (see Figure 5).

XSL•FO
RenderX

Figure 5. Body-shaping behavior (muscle building) in relation to #fitspiration user status and FFMI percentiles. FFMI: fat-free mass index.

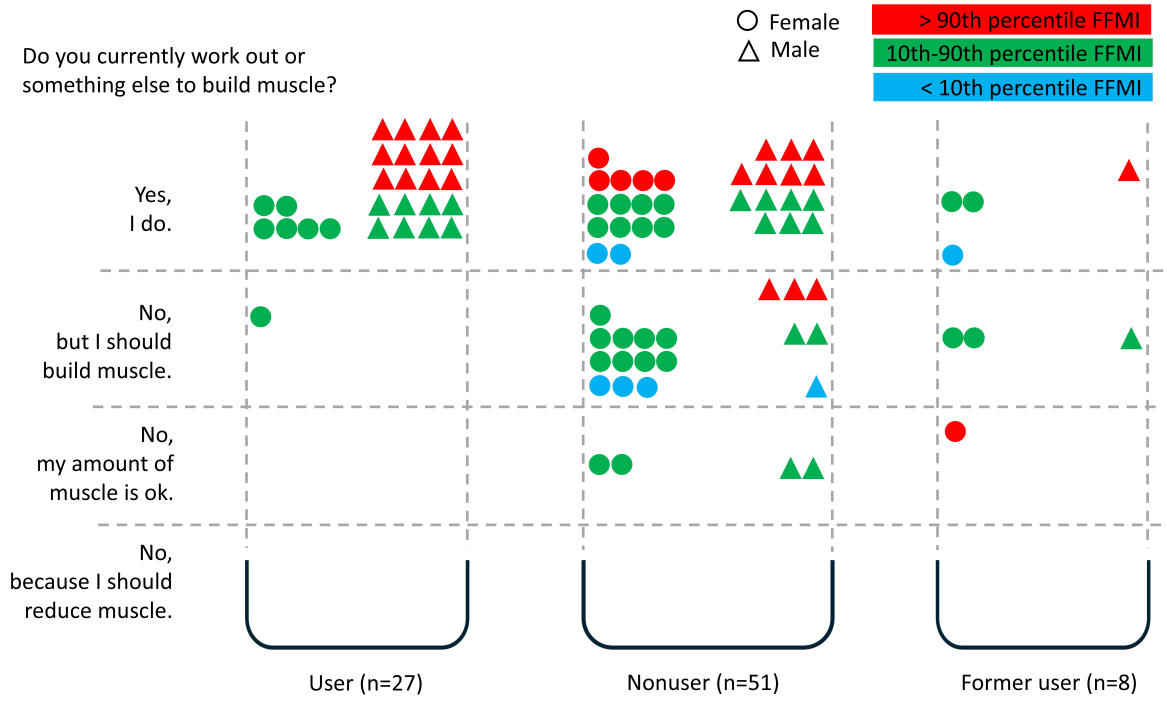

Discussion

Main Findings and Comparison With Prior Work

The aim of this exploratory investigation study was to gain initial insights into gender-specific descriptions of adolescent #fitspiration user groups. In addition, we linked the results to body composition data, which allowed us to gain a deeper understanding of body image perceptions and body-shaping behaviors. Generally, muscularity seems to be of great importance for physical appearance. Male #fitspiration users showed greater dissatisfaction with muscularity, while female #fitspiration users were more concerned about their body weight. No #fitspiration user rated her or his amount of muscle mass as okay, although all of them had an average or high FFMI. Generally, body image perception was more accurate for FMI than for FFMI, suggesting some discrepancy in self-assessment of muscularity.

Considering the BMI, we observed a substantially lower number of adults with overweight or obesity in our sample than the Austrian HBSC average [47]. When compared to the normal-weight MoMo sample, our participants had a similar BMI. Furthermore, the FMI and FFMI of our participants were mainly within the normal range [44].

Almost two-thirds of the participants rated their body weight and body fat correctly compared to the FMI. This outcome is in line with nationwide cross-sectional studies where similar rates of adolescents rightly matched their BMI with their self-perceived weight [48,49]. Nevertheless, and similar to the Austrian HBSC study participants [50], approximately one-third of females and 1 out of 10 male participants with a low or normal FMI thought that they were too fat. This gender specificity is in line with a vast majority of studies, where female participants tend to overestimate their body weight [48,51-53]. Contrary to other studies, where adolescents with overweight or obesity underestimated their body weight [54,55], the majority of our participants above the 90th FMI percentile rated themselves correctly.

Concerning body-shaping behavior and its possible linkage to #fitspiration use, we observed the following: One in five of all adolescents, which is close to the Austrian HBSC average [34], stated that they were dieting or did something else to reduce body weight. In our investigation study, about one-third of #fitspiration users stated that they had implemented weight reduction behavior. In line with the results of body perception and body fat perception, female participants (whole sample: 10/42, 24%; #fitspiration users: 5/7, 71%) indicated weight reduction behavior more often than males (whole sample: 6/44, 14%; #fitspiration users: 3/20, 15%)—even if they were classified with a low or normal FMI. When compared with Austrian HBSC data, the percentage of young people who engaged in weight reduction behavior was similar [34]. This might raise the question of whether there is a connection between #fitspiration and body-shaping behavior. This has also been reported in other studies where the influence of #fitspiration had led to restrained eating behavior [56] and could increase eating disorder symptoms [26,57]. The majority of those who thought that they should gain weight were male participants. We assume that male participants had gaining muscle mass in mind rather than body weight. This is in line with literature that showed that boys are more likely to experience the pressure to increase muscle mass [58,59].

Content analysis of #fitspiration shows that these posts mainly show highly muscular men and women [60,61], which might lead to muscular-ideal internalization. Our findings might suggest that muscularity seemed to be of greater importance than thinness, and the participating adolescents seemed to find it more difficult to assess. Almost two-thirds of all participants

XSL•FO
RenderX

rated themselves as too unmuscular or far too unmuscular, although they had an FFMI above the 10th percentile. This outcome seemed to be of greater significance to male participants. Most male participants with a high FFMI (12/23, 52%) belonged to the group of #fitspiration users, and nearly one-third of them (6/20) perceived themselves as too unmuscular. Besides self-perception, the importance of muscularity became apparent when it comes to strength training. Almost all participants said they worked out or thought that they should build muscle. Female participants were inactive more often, although they likewise thought that they should build muscle, whereas males worked out—or thought that they should—even though their FFMI was above the 90th percentile. Almost all #fitspiration users worked out, although they had an FFMI within the normal range or even higher. Although promoting exercise is highly recommended facing the enormous prevalence of overweight and obesity, the context of #fitspiration with its muscular ideals and training recommendations also raises the question of what motivates users to do strength training. Because appearance-based exercise motivation, which can be linked with #fitspiration [62], excessive or one-sided sports could lead to injuries, overload damage, and reduced immune function, besides psychological problems like eating disorders [63].

Our results indicate that girls have internalized the "new" beauty ideal, which includes thinness as well as muscularity [22,64]. The beauty ideal for boys still seemed to be "the more muscles, the better." Although almost all participants seemed to aspire to these ideals, #fitspiration users appeared to be pursuing them even more. This is not surprising because the muscular ideals are core elements of #fitspiration [25,60] and could lead to the internalization of it [22]. But it is noteworthy that none of the 27 #fitspiration users stated that her or his amount of muscle is okay even though none of them had a low FFMI.

On the other hand, #fitspiration may help people move from thinking about doing sports to action. This assumption is derived from the observation that all male #fitspiration users and almost all female #fitspiration users were actively working out, while approximately one-third of the nonusers were just thinking about it. Similar to our findings, there are some studies indicating that #fitspiration might have positive influences on dietary change or physical activity [65,66].

Limitations and Strengths

Overall, the main limitation regarding the results' generalizability is the sampling—the study group (n=86) is relatively small and does not represent the general population, which is characteristic for exploratory investigation studies.

Moreover, there are further limitations concerning the sampling, which are, first, the majority of participants were of normal body weight, and our sample had lower rates of overweight or obesity than reported in the Austrian HBSC study [67]. This is of importance because the BMI is a known factor for influencing the accuracy of identifying a person's weight status [68,69], body satisfaction and body dissatisfaction [68], or health behavior like dieting [70] or physical activity [71]–even in adolescents [54,55,72]. Therefore, it can be speculated that teen #fitspiration users with a more elevated BMI (originating from abnormal or excessive fat accumulation) may experience higher body dissatisfaction than the adolescents of this study. Furthermore, adolescents with overweight or obesity might implement fewer body-shaping behaviors, especially working out, due to the weight stigma than our mainly normal-weight sample. Second, most of the participants lived in well-off families without a migration background, which are known factors influencing body perception [73,74] and body weight control behavior [75]. Our results are not automatically transferable to adolescents of less affluent families or with migration backgrounds. Third, we did not quantify the intensity of #fitspiration use. Therefore, we cannot speculate concerning the amount of usage of #fitspiration and possible associations to body image perception and body-shaping behavior. Consequently, the results cannot be transferred to other adolescents using #fitspiration, which was not the objective of this study.

The strength of our study is first, that it adds additional perspectives regarding the use of #fitspiration and allows for the juxtaposition of the participants' subjective assessment with objective measurements. Besides gender-specific descriptive analyses and differentiation between #fitspiration users and nonusers, we got first impressions of the perspective of former users. Second, anthropometric data were measured and not self-reported like at HBSC, which improves our data quality. Third, and to the best of our knowledge, our study is the first that conducted body composition measurements within different adolescent #fitspiration user groups. This allows us to gain first information on the potential relationship between #fitspiration use, body composition, self-perception of body shape, and body-shaping behaviors. But it does not allow a clear conclusion on whether exposure to this type of content leads to changes in behavior.

Future Directions

Future studies should focus on the significance of muscularity for adolescent #fitspiration users and how they perceive their own muscle mass because it is questionable if #fitspiration use might foster body dissatisfaction and increase the risk of body dysmorphia. Is it possible that the enormous presence of supposedly perfect and more muscular bodies further increases the pressure on young people to conform to this ideal? How do extreme training, rigid eating habits, perhaps medically unsupervised supplementation, or even substance abuse affect the still-developing bodies of young people? Furthermore, research on #fitspiration should investigate how #fitspiration influences adolescents from diverse backgrounds, particularly those with migration backgrounds or lower socioeconomic status. Besides, it is the question of whether the characteristics of influencers, such as their professional qualification, type of sport, diet, or products they are presenting, affect #fitspiration users differently. Moreover, it is questionable whether a dose-response relationship regarding frequency and intensity of exposure and psychological or behavioral outcomes exists. Social media platform-specific dynamics should also be considered, as the effects of #fitspiration may vary across social media platforms due to differences in format, audience, and algorithmic exposure.

XSL•FO
RenderX

Conclusion

In conclusion, health-related social media content #fitspiration is an important source of health information and body image perceptions for adolescents. This may influence health-related and body-shaping behaviors during this critical period of life. Our findings suggest that high muscle mass is an essential part of adolescents' beauty ideals. Consequently, adolescents sought to achieve this through body-shaping behaviors—particularly strength training. #Fitspiration seemed to fuel these ideals and behaviors. Although the increasing prevalence of overweight and obesity makes exercise advisable, #fitspiration carries potential dangers for young people, especially body dissatisfaction, eating disorders, or physical damage caused by overtraining or incorrect training. This means that, on the one hand, more research is needed to understand the effects of #fitspiration on different target groups outside of laboratory settings. On the other hand, young people need to be equipped with the necessary knowledge and skills to use social media critically. To achieve this goal, all relevant stakeholders, especially families and schools, must collaborate. An increasingly rising amount of scientific effort, like the project FIVE, has been devoted to providing adolescents and stakeholders with age-appropriate, science-based courses [29], materials [76], and digital interventions [77,78].

In addition, health care professionals need to be aware that social media information and influencers are of great importance to adolescents and need to address this issue during their assessment process.

Acknowledgments

The authors gratefully acknowledge the financial support of the Austrian Research Promotion Agency [79], FEMtech Forschungsprojekte 2020 (FO999890319). We would also like to thank the participating schools (Bundesoberstufenrealgymnasium und Bundeshandelsschule für Leistungssport St. Pölten; Höhere Technische Bundeslehr- und Versuchsanstalt Wien XX Wexstraße; Bundesgymnasium und Bundesrealgymnasium Wien 6; Bundeshandelsakademie und Bundeshandelsschule St. Pölten, St. Pölten; Gymnasium und Realgymnasium Sachsenbrunn; Schulzentrum Ybbs an der Donau) teachers, pupils, and parents or legal guardians for their support of this project. Furthermore, we would like to thank the entire project team for their excellent cooperation. The project FIVE was funded by the Austrian Research Promotion Agency [79], FEMtech Forschungsprojekte 2020 (FO999890319). The funding organization has no influence on the design, collection, analysis, and interpretation of data, or the writing of the manuscript. The authors used the free version of (DeepL.com, from DeepL SE, 2025) for improving language. No other generative AI or AI-assisted tools were used for content creation, data analysis, or interpretation of results. The authors did not use generative AI.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to the informed consent documents. In these documents, we have declared that data will not be passed on to third parties and that the results of the study will be published only in aggregated form (eg, mean values, standard deviation across the respective group).

Authors' Contributions

EH contributed to conceptualization, formal analysis, investigation, methodology, funding acquisition, project administration, and writing-original draft as well as reviewing and editing. TR handled conceptualization, project administration, and reviewing and editing. MH performed conceptualization, formal analysis, investigation, methodology, funding acquisition, and writing-original draft as well as reviewing and editing

Conflicts of Interest

None declared.

References

1. Cash TF. Body image: past, present, and future. Body Image 2004 Jan;1(1):1-5. [doi: 10.1016/S1740-1445(03)00011-1] [Medline: 18089136]
2. Grogan S. Promoting positive body image in males and females: contemporary issues and future directions. Sex Roles 2010 Nov;63:757-765. [doi: 10.1007/s11199-010-9894-z]
3. Toselli S, Zaccagni L, Rinaldo N, et al. Body image perception in high school students: the relationship with gender, weight status, and physical activity. Children (Basel) 2023 Jan 10;10(1):137. [doi: 10.3390/children10010137] [Medline: 36670687]
4. Tiggemann M. Body image across the adult life span: stability and change. Body Image 2004 Jan;1(1):29-41. [doi: 10.1016/S1740-1445(03)00002-0] [Medline: 18089139]
5. Wertheim E, Paxton S. Body image development in adolescent girls. In: Body Image A Handbook of Science, Practice, and Prevention, 2nd edition: The Guilford Press; 2012.

XSL•FO
RenderX

6. Duarte LS, Palombo CNT, Solis-Cordero K, et al. The association between body weight dissatisfaction with unhealthy eating behaviors and lack of physical activity in adolescents: a systematic review. J Child Health Care 2021 Mar;25(1):44-68. [doi: 10.1177/1367493520904914] [Medline: 32053022]
7. Blanchard L, Conway-Moore K, Aguiar A, et al. Associations between social media, adolescent mental health, and diet: a systematic review. Obes Rev 2023 Sep;24(Suppl 2):e13631. [doi: 10.1111/obr.13631] [Medline: 37753597]
8. Bodega P, de Cos-Gandoy A, Fernández-Alvira JM, Fernández-Jiménez R, Moreno LA, Santos-Beneit G. Body image and dietary habits in adolescents: a systematic review. Nutr Rev 2023 Dec 11;82(1):104-127. [doi: 10.1093/nutrit/nuad044] [Medline: 37155836]
9. Knoll B, Fitz B, Posch P. Selbstdarstellung von weiblichen und männlichen Jugendlichen in sozialen Netzwerken. 2013. URL: https://www.b-nk.at/download/ich-im-netz/?wpdmdl=3559&refresh=61d2fb9cf37c71641216924 [accessed 2020-10-22]
10. Myers TA, Crowther JH. Social comparison as a predictor of body dissatisfaction: a meta-analytic review. J Abnorm Psychol 2009 Nov;118(4):683-698. [doi: 10.1037/a0016763] [Medline: 19899839]
11. Jiotsa B, Naccache B, Duval M, Rocher B, Grall-Bronnec M. Social media use and body image disorders: association between frequency of comparing one's own physical appearance to that of people being followed on social media and body dissatisfaction and drive for thinness. Int J Environ Res Public Health 2021 Mar 11;18(6):2880. [doi: 10.3390/ijerph18062880] [Medline: 33799804]
12. Nolen E, Panisch LS. The relationship between body appreciation and health behaviors among women and adolescent girls: a scoping review. Health Soc Work 2022 Apr 13;47(2):113-122. [doi: 10.1093/hsw/hlac006] [Medline: 35262682]
13. Alfonso-Fuertes I, Alvarez-Mon MA, Sanchez Del Hoyo R, Ortega MA, Alvarez-Mon M, Molina-Ruiz RM. Time spent on Instagram and body image, self-esteem, and physical comparison among young adults in Spain: observational study. JMIR Form Res 2023 Apr 7;7:e42207. [doi: 10.2196/42207] [Medline: 37027197]
14. Engeln R, Loach R, Imundo MN, Zola A. Compared to Facebook, Instagram use causes more appearance comparison and lower body satisfaction in college women. Body Image 2020 Sep;34:38-45. [doi: 10.1016/j.bodyim.2020.04.007] [Medline: 32505866]
15. McGovern O, Collins R, Dunne S. The associations between photo-editing and body concerns among females: a systematic review. Body Image 2022 Dec;43:504-517. [doi: 10.1016/j.bodyim.2022.10.013] [Medline: 36371869]
16. Sharma A, Vidal C. A scoping literature review of the associations between highly visual social media use and eating disorders and disordered eating: a changing landscape. J Eat Disord 2023 Sep 26;11(1):170. [doi: 10.1186/s40337-023-00898-6] [Medline: 37752611]
17. Li Y, Xie Y. Is a picture worth a thousand words? An empirical study of image content and social media engagement. J Mark Res 2020 Feb;57(1):1-19. [doi: 10.1177/0022243719881113]
18. van Oosten JMF, Vandenbosch L, Peter J. Predicting the use of visually oriented social media: the role of psychological well-being, body image concerns and sought appearance gratifications. Comput Human Behav 2023 Jul;144:107730. [doi: 10.1016/j.chb.2023.107730]
19. Bohjalian E. The self-presentation of popular fitspiration experts on Instagram. Elon J Undergrad Res Commun 2017;8(1):23-34 [FREE Full text]
20. Murray SB, Griffiths S, Hazery L, Shen T, Wooldridge T, Mond JM. Go big or go home: a thematic content analysis of pro-muscularity websites. Body Image 2016 Mar;16:17-20. [doi: 10.1016/j.bodyim.2015.10.002] [Medline: 26523689]
21. Tiggemann M, Zaccardo M. "Strong is the new skinny": a content analysis of #fitspiration images on Instagram. J Health Psychol 2018 Jul;23(8):1003-1011. [doi: 10.1177/1359105316639436] [Medline: 27611630]
22. Jerónimo F, Carraça EV. Effects of fitspiration content on body image: a systematic review. Eat Weight Disord 2022 Dec;27(8):3017-3035. [doi: 10.1007/s40519-022-01505-4] [Medline: 36401082]
23. Fioravanti G, Svicher A, Ceragioli G, Bruni V, Casale S. Examining the impact of daily exposure to body-positive and fitspiration Instagram content on young women's mood and body image: an intensive longitudinal study. New Media Soc 2023 Dec;25(12):3266-3288. [doi: 10.1177/14614448211038904]
24. Breves PL, van Berlo ZMC, Teunissen L, König L, Binder A, Naderer B. Happier and healthier? Investigating the longitudinal impact of body-positive and fitspirational influencers on weight satisfaction, healthy eating, and physical activity. Health Commun 2025 Feb 25:1-13. [doi: 10.1080/10410236.2025.2465795] [Medline: 39996327]
25. Tiggemann M, Anderberg I. Muscles and bare chests on Instagram: the effect of Influencers' fashion and fitspiration images on men's body image. Body Image 2020 Dec;35:237-244. [doi: 10.1016/j.bodyim.2020.10.001] [Medline: 33157398]
26. Cataldo I, De Luca I, Giorgetti V, et al. Fitspiration on social media: body-image and other psychopathological risks among young adults. a narrative review. Emerg Trends Drugs Addict Health 2021;1:100010. [doi: 10.1016/j.etdah.2021.100010]
27. Dignard NAL, Jarry JL. The "Little Red Riding Hood effect:" Fitspiration is just as bad as thinspiration for women's body satisfaction. Body Image 2021 Mar;36:201-213. [doi: 10.1016/j.bodyim.2020.11.012] [Medline: 33360477]
28. FIVE - #fitspiration image verification. St Pölten University of Applied Sciences. URL: https://research.fhstp.ac.at/projekte/five-fitspiration-image-verification [accessed 2024-11-17]
29. Höld E, Größbacher S, Kramml P, et al. Fit for #fitspiration: human-centred design of a blended learning course for upper secondary schools. Tech Know Learn 2025 Jul 3. [doi: 10.1007/s10758-025-09871-5]

XSL•FO
RenderX

30. Boyce W, Torsheim T, Currie C, Zambon A. The family affluence scale as a measure of national wealth: validation of an adolescent self-report measure. Soc Indic Res 2006 Sep;78(3):473-487. [doi: 10.1007/s11205-005-1607-6]
31. Currie C, Gabhainn SN, Godeau E, et al. Inequalities in young people's health: HBSC international report from the 2005/06 Survey. Health Policy for Children and Adolescents, No. 5. : WHO Regional Office for Europe; 2008 URL: https://hbsc.org/publications/reports/inequalities-in-young-peoples-health/ [accessed 2025-10-22]
32. Conference of European Statisticians Recommendations for the 2020 Censuses of Population and Housing. : United Nations Economic Commission for Europe; 2015 URL: https://unece.org/fileadmin/DAM/stats/publications/2015/ECECES41_EN.pdf [accessed 2022-06-15]
33. König LM, Sproesser G, Schupp HT, Renner B. Describing the process of adopting nutrition and fitness apps: Behavior stage model approach. JMIR Mhealth Uhealth 2018 Mar 13;6(3):e55. [doi: 10.2196/mhealth.8261] [Medline: 29535078]
34. Inchley J, Currie D, Young T, et al. Growing up unequal: gender and socioeconomic differences in young people's health and well-being: health behaviour in school-aged children (HBSC) study: international report from the 2013/2014 survey. : World Health Organization Regional Office for Europe; 2016 URL: https://hbsc.org/publications/reports/growing-up-unequal-gender-and-socioeconomic-differences-in-young-peoples-health-and-well-being/ [accessed 2025-10-22]
35. Khalil SF, Mohktar MS, Ibrahim F. The theory and fundamentals of bioimpedance analysis in clinical status monitoring and diagnosis of diseases. Sensors (Basel) 2014 Jun 19;14(6):10895-10928. [doi: 10.3390/s140610895] [Medline: 24949644]
36. Zopf DR, Herrmann HJ, Neurath MF, Yurdagül Z. Bioelectrical impedance analysis (BIA). Ernährungs Umschau ;2019:474-467. [doi: 10.4455/eu.2019.039]
37. Kromeyer-Hauschild K, Moss A, Wabitsch M. Reference values for body mass index for children, adolescents and adults in Germany. Adipositas - Ursachen, Folgeerkrankungen, Therapie 2015;9(3):123-127. [doi: 10.1055/s-0037-1618928]
38. Demerath EW, Schubert CM, Maynard LM, et al. Do changes in body mass index percentile reflect changes in body composition in children? Data from the Fels Longitudinal Study. Pediatrics 2006 Mar;117(3):e487-e495. [doi: 10.1542/peds.2005-0572] [Medline: 16510627]
39. VanItallie TB, Yang MU, Heymsfield SB, Funk RC, Boileau RA. Height-normalized indices of the body's fat-free mass and fat mass: potentially useful indicators of nutritional status. Am J Clin Nutr 1990 Dec;52(6):953-959. [doi: 10.1093/ajcn/52.6.953] [Medline: 2239792]
40. Alpízar M, Frydman TD, Reséndiz-Rojas JDJ, Trejo-Rangel MA, Aldecoa-Castillo JMD. Fat mass index (FMI) as a trustworthy overweight and obesity marker in Mexican pediatric population. Children (Basel) 2020 Mar 10;7(3):19. [doi: 10.3390/children7030019] [Medline: 32164179]
41. He H, Pan L, Du J, et al. Effect of fat mass index, fat free mass index and body mass index on childhood blood pressure: a cross-sectional study in south China. Transl Pediatr 2021 Mar;10(3):541-551. [doi: 10.21037/tp-20-325] [Medline: 33850812]
42. Kyle UG, Morabia A, Schutz Y, Pichard C. Sedentarism affects body fat mass index and fat-free mass index in adults aged 18 to 98 years. Nutrition 2004 Mar;20(3):255-260. [doi: 10.1016/j.nut.2003.11.019] [Medline: 14990265]
43. Dos Santos RRG, Forte GC, Mundstock E, et al. Body composition parameters can better predict body size dissatisfaction than body mass index in children and adolescents. Eat Weight Disord 2020 Oct;25(5):1197-1203. [doi: 10.1007/s40519-019-00750-4] [Medline: 31338791]
44. Schmidt SCE, Bosy-Westphal A, Niessner C, Woll A. Representative body composition percentiles from bioelectrical impedance analyses among children and adolescents. The MoMo study. Clin Nutr 2019 Dec;38(6):2712-2720. [doi: 10.1016/j.clnu.2018.11.026] [Medline: 30554799]
45. WMA Declaration of Helsinki – Ethical Principles for Medical Research Involving Human Participants. World Medical Association. 2022. URL: https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/ [accessed 2024-02-15]
46. Recommendations for information and consent forms for studies on minors. : Forum Österreichischer Ethikkommissionen; 2002 URL: https://me001ned.edis.at/ethikkommission/Forum/Download/Files/PI_MJ.pdf [accessed 2025-10-22]
47. Overweight and obesity. HBSC. 2023. URL: https://data-browser.hbsc.org/measure/overweight-and-obesity/ [accessed 2024-08-28]
48. Moehlecke M, Blume CA, Cureau FV, Kieling C, Schaan BD. Self-perceived body image, dissatisfaction with body weight and nutritional status of Brazilian adolescents: a nationwide study. J Pediatr (Rio J) 2020;96(1):76-83. [doi: 10.1016/j.jped.2018.07.006] [Medline: 30098939]
49. Song L, Zhang Y, Chen T, Maitusong P, Lian X. Association of body perception and dietary weight management behaviours among children and adolescents aged 6-17 years in China: cross-sectional study using CHNS (2015). BMC Public Health 2022 Jan 26;22(1):175. [doi: 10.1186/s12889-022-12574-6] [Medline: 35081917]
50. Felder-Puig R, Teufl L, Winkler R. Gewichtsstatus Und Körperselbstbild von Österreichischen Jugendlichen HBSC Factsheet 02 Aus Erhebung 2021/22: BMSGPK; 2023.
51. Bordeleau M, Alméras N, Panahi S, Drapeau V. Body image and lifestyle behaviors in high school adolescents. Children (Basel) 2023 Jul 22;10(7):1263. [doi: 10.3390/children10071263] [Medline: 37508760]

XSL•FO
RenderX

52. Buscemi S, Marventano S, Castellano S, et al. Role of anthropometric factors, self-perception, and diet on weight misperception among young adolescents: a cross-sectional study. Eat Weight Disord 2018 Feb;23(1):107-115. [doi: 10.1007/s40519-016-0306-4] [Medline: 27502108]
53. Hasan HA, Radwan H, Al Majid F, et al. Lean body mass and self-perceived body image among youth in the United Arab Emirates. Acta Biomed 2022 Jul 1;93(3):e2022100. [doi: 10.23750/abm.v93i3.12722] [Medline: 35775762]
54. Ballarin G, Gallè F, Dinacci L, Liberti F, Cunti A, Valerio G. Self-perception profile, body image perception and satisfaction in relation to body mass index: an investigation in a sample of adolescents from the Campania region, Italy. Children (Basel) 2024 Jun 30;11(7):805. [doi: 10.3390/children11070805] [Medline: 39062254]
55. San Martini MC, de Assumpção D, Barros MBDA, Barros Filho ADA, Mattei J. Weight self-perception in adolescents: evidence from a population-based study. Public Health Nutr 2021 May;24(7):1648-1656. [doi: 10.1017/S1368980021000690] [Medline: 33648614]
56. Linardon J. Investigating longitudinal bidirectional associations between appearance comparisons to fitspiration content on Instagram, positive and negative body image, and dietary restraint. Eat Disord 2023 Sep 3;31(5):450-463. [doi: 10.1080/10640266.2023.2190973] [Medline: 36935581]
57. DiBisceglie S, Arigo D. Perceptions of #fitspiration activity on Instagram: patterns of use, response, and preferences among fitstagrammers and followers. J Health Psychol 2021 Jul;26(8):1233-1242. [doi: 10.1177/1359105319871656] [Medline: 31455118]
58. Radwan H, Hasan HA, Ismat H, et al. Body mass index perception, body image dissatisfaction and their relations with weight-related behaviors among university students. Int J Environ Res Public Health 2019 May 1;16(9):1541. [doi: 10.3390/ijerph16091541] [Medline: 31052368]
59. Ricciardelli L, McCabe M. Body image development in adolescent boys. In: Body Image A Handbook of Science, Practice, and Prevention, 2nd edition: The Guilford Press, Vol. 2012:85-93.
60. Angrish K, Oshana D, Gammage KL. Flex Friday: a content analysis of men's fitspiration on Instagram. Body Image 2024 Mar;48:101651. [doi: 10.1016/j.bodyim.2023.101651] [Medline: 38061212]
61. Carrotte ER, Prichard I, Lim MSC. "Fitspiration" on social media: a content analysis of gendered images. J Med Internet Res 2017 Mar 29;19(3):e95. [doi: 10.2196/jmir.6368] [Medline: 28356239]
62. Fatt SJ, Fardouly J, Rapee RM. #malefitspo: links between viewing fitspiration posts, muscular-ideal internalisation, appearance comparisons, body satisfaction, and exercise motivation in men. New Media Soc 2019 Jun;21(6):1311-1325. [doi: 10.1177/1461444818821064]
63. Niedermeier M, Frühauf A, Bichler C, Rosenberger R, Kopp M. Sport – zu Risiken und Nebenwirkungen. Orthopäde 2019 Dec;48(12):1030-1035. [doi: 10.1007/s00132-019-03823-5]
64. Rodgers RF, Slater A, Gordon CS, McLean SA, Jarman HK, Paxton SJ. A biopsychosocial model of social media use and body image concerns, disordered eating, and muscle-building behaviors among adolescent girls and boys. J Youth Adolesc 2020 Feb;49(2):399-409. [doi: 10.1007/s10964-019-01190-0] [Medline: 31907699]
65. Duplaga M. The use of fitness influencers' websites by young adult women: a cross-sectional study. Int J Environ Res Public Health 2020 Sep 1;17(17):6360. [doi: 10.3390/ijerph17176360] [Medline: 32882887]
66. Easton S, Morton K, Tappy Z, Francis D, Dennison L. Young people's experiences of viewing the fitspiration social media trend: qualitative study. J Med Internet Res 2018 Jun 18;20(6):e219. [doi: 10.2196/jmir.9156] [Medline: 29914859]
67. Felder-Puig R, Teutsch F, Winkler R. Gesundheit und Gesundheitsverhalten von österreichischen Schülerinnen und Schülern Ergebnisse des WHO-HBSC-Survey 2021/22.: Gesundheit Österreich GmbH; 2023. URL: https://goeg.at/sites/goeg.at/files/inline-files/%C3%96sterr.%20HBSC-Bericht%202023_bf.pdf [accessed 2024-08-26]
68. Ramirez Luque DB, Rocha Huaman NL, Calizaya-Milla YE, Calizaya-Milla SE, Ramos-Vera C, Saintila J. Body self-perception, dietary self-efficacy, and body mass index in young adults: a cross-sectional survey. Int J Gen Med 2023;16:193-202. [doi: 10.2147/IJGM.S395281] [Medline: 36699341]
69. Robinson E, Haynes A, Sutin A, Daly M. Self-perception of overweight and obesity: a review of mental and physical health outcomes. Obes Sci Pract 2020 Oct;6(5):552-561. [doi: 10.1002/osp4.424] [Medline: 33082997]
70. van Vliet JS, Gustafsson PA, Nelson N. Feeling "too fat" rather than being "too fat" increases unhealthy eating habits among adolescents - even in boys. Food Nutr Res 2016;60:29530. [doi: 10.3402/fnr.v60.29530] [Medline: 26894992]
71. Morano M, Colella D, Robazza C, Bortoli L, Capranica L. Physical self-perception and motor performance in normal-weight, overweight and obese children. Scand J Med Sci Sports 2011 Jun;21(3):465-473. [doi: 10.1111/j.1600-0838.2009.01068.x] [Medline: 20136752]
72. Jáuregui-Lobera I, Iglesias Conde A, Sánchez Rodríguez J, et al. Self-perception of weight and physical fitness, body image perception, control weight behaviors and eating behaviors in adolescents. Nutr Hosp 2018 Oct 5;35(5):1115-1123. [doi: 10.20960/nh.1726] [Medline: 30307295]
73. Joo YY, Kim J, Lee K, Cho GJ, Yi KW. Misperception of body weight and associated socioeconomic and health-related factors among Korean female adults: a nationwide population-based study. Front Endocrinol (Lausanne) 2022;13:1007129. [doi: 10.3389/fendo.2022.1007129] [Medline: 36619562]
74. Wronka I, Suliga E, Pawlińska-Chmara R. Perceived and desired body weight among female university students in relation to BMI-based weight status and socio-economic factors. Ann Agric Environ Med 2013;20(3):533-538. [Medline: 24069860]

XSL•FO
RenderX

75. Bojorquez I, Villatoro J, Delgadillo M, Fleiz C, Fregoso D, Unikel C. Social factors, weight perception, and weight control practices among adolescents in Mexico. J Health Psychol 2018 Jun;23(7):939-950. [doi: 10.1177/1359105316643596] [Medline: 27106084]
76. Curtis RG, Prichard I, Gosse G, Stankevicius A, Maher CA. Hashtag fitspiration: credibility screening and content analysis of Instagram fitness accounts. BMC Public Health 2023 Mar 2;23(1):421. [doi: 10.1186/s12889-023-15232-7] [Medline: 36864397]
77. Wei Z, Liu K, Yang W, Zhang Y, Fu H, Fu L. Body image dissatisfaction intervention program based on a three-factor model of appearance comparison, appearance comments, and social media among Chinese children. Humanit Soc Sci Commun 2025;12(1):1092. [doi: 10.1057/s41599-025-05396-5]
78. Mahon C, Seekis V. Systematic review of digital interventions for adolescent and young adult women's body image. Front Glob Womens Health 2022;3:832805. [doi: 10.3389/fgwh.2022.832805] [Medline: 35392118]
79. We shape the future!. Austrian Research Promotion Agency. URL: https://www.ffg.at/ [accessed 2025-10-15]

Abbreviations

BIA: bioimpedance analysis
FAS: Family Affluence Scale
FFM: fat-free mass
FFMI: fat-free mass index
FIVE: #Fitspiration Image Verification
FM: fat mass
FMI: fat mass index
HBSC: Health Behaviour in School-aged Children
MoMo: Motorik-Modul Longitudinal Study of the KiGGS study
UNECE: United Nations Economic Commission for Europe

Edited by A Mavragani; submitted 07.01.25; peer-reviewed by A Pink, K Krupa-Kotara; revised version received 29.09.25; accepted 30.09.25; published 06.11.25.

Please cite as:
Höld E, Rathmanner T, Heller M
Teenage Body Image Perception, Body-shaping Behavior, and Body Composition With Respect to Use of "Fitspiration": Exploratory Investigation Study
JMIR Form Res 2025;9:e70964
URL: https://formative.jmir.org/2025/1/e70964
doi:10.2196/70964

© Elisabeth Höld, Theres Rathmanner, Mario Heller. Originally published in JMIR Formative Research (https://formative.jmir.org), 6.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Social Cohesion, Mental Well-Being, and the Role of Smart Technology and Pet Ownership Among Social Housing Residents: Cross-Sectional Cohort Study

Emmylou Rahtz[1], PhD; Andrew James Williams[2], PhD; Timothy J Taylor[1], PhD

[1]European Centre for Environment and Human Health, Faculty of Health and Life Sciences, University of Exeter, Peter Lanyon, Penryn Campus, Penryn, United Kingdom

[2]Scottish Collaboration for Public Health Research and Policy, School of Health in Social Science, University of Edinburgh, Edinburgh, United Kingdom

Corresponding Author:
Emmylou Rahtz, PhD
European Centre for Environment and Human Health, Faculty of Health and Life Sciences, University of Exeter, Peter Lanyon, Penryn Campus, Penryn, United Kingdom

Abstract

Background: Smart technology has been shown to have varied effects on social cohesion and mental well-being. There has been very little research on associations between pet ownership and social cohesion and mental well-being.

Objective: This study aimed to explore associations between social cohesion and mental well-being and ownership of different forms of smart technology, dogs, and cats in a sample of adult social housing occupants in Cornwall, United Kingdom.

Methods: This was a cross-sectional study that collected data on people's living environment, health, and well-being, including the Short Warwick-Edinburgh Mental Wellbeing Scale and an 8-item social cohesion scale. Participants were social housing residents in Cornwall in the South West of the United Kingdom. We used cross-sectional regression analyses to explore associations between people's ownership of different forms of smart technology and pets, and their reported levels of social cohesion and mental well-being.

Results: There were no statistically significant associations between social cohesion and ownership of either smart technology or pets. Unadjusted regressions for mental well-being showed an association with owning a smartphone. However, after adjusting for age, gender, and socioeconomic status, this effect was no longer significant. Those who owned any smart technology (b=1.76, 95% CI 0.06-3.45; P=.04) and those who owned a games console (b=2.39, 95% CI 0.59-4.19, P=.01) had significantly higher levels of mental well-being, after adjusting for age, gender, and socioeconomic status; the effect held after social cohesion was added to the model. Counterintuitively, owning two or more dogs was associated with lower levels of mental well-being (b=−2.12, 95% CI −4.06 to −0.19; P=.03) compared with owning no dogs, after adjusting for age, gender, socioeconomic status, and social cohesion. However, there were no significant differences in mental well-being related to cat ownership.

Conclusions: Previous research suggests that the beneficial effects of smart technology are context-dependent, and our results support that. While we did not observe significant effects on social cohesion, owning any smart technology or a games console specifically was associated with well-being benefits. There are limited data on pet ownership and social cohesion or mental well-being: our data suggest there is no strong relationship with social cohesion but owning multiple dogs can have negative effects.

(JMIR Form Res 2025;9:e75451) doi:10.2196/75451

KEYWORDS

mental well-being; social cohesion; smart technology; gaming; pet ownership; social housing

Introduction

Social Cohesion and Well-being

The costs of mental ill health in England have been estimated at £119 billion (US $161 billion) per annum [1]. A number of factors have been found to be linked to subjective well-being, including health status, housing tenure, relationship status, and age [2].

Social cohesion describes the extent to which people feel integrated within their neighborhood, based on shared values and a sense of solidarity [3]; it has also been shown to be a powerful indicator of subjective well-being [3], and cross-national studies indicate that the United Kingdom's position in national rankings of social cohesion is falling [4].

The Smartline project recruited a sample of people living in social housing and included measures on social cohesion, mental

XSL•FO
RenderX

well-being, physical and mental health, as well as physical measures of the home environment. Social housing is rented accommodation costing below the market rate, typically offered to people with specific health needs or in particular economic circumstances by not-for-profit organizations. The project was based in the United Kingdom county of Cornwall. Cornwall has been in receipt of initiatives to increase digital inclusion in rural areas, with decreased social isolation and improved access to health care among the aims [5].

Previous research among the Smartline cohort showed that social cohesion was associated with mental well-being, but not with physical or mental health [6], and gave rise to further questions about the role of smart technology and pet ownership.

Smart Technology, Social Cohesion, and Well-being

Qualitative studies suggest that smart technologies and, more broadly, internet and communications technologies (ICT) can have a positive impact on social cohesion and societal capabilities [7]. Smart technology initiatives can provide tools for inclusive growth: early initiatives in digital inclusion led to social and economic benefits in rural communities [8], and smartphones, used appropriately, can improve social connections among older adults [9]. In rural settings, it has been argued that ICT and smart technologies can encourage people to form social groups based on personal choice rather than geographical location, the so-called "disembedding" of social relationships [10]. Clearly, this can have both positive and negative effects, improving cohesion with online communities at the potential cost of geographically local connections. ICT allows people to work remotely, potentially reducing the potential for local integration [10]. Positive in-person social integration can be driven by "the commitment of the local 'bohemian' creative class to developing a sense of solidarity" [10], which would clearly be highly context dependent. Meanwhile, there is widespread societal concern about the effects of smartphone use on mental health, particularly among adolescents [11,12]. Some evidence supports this, indicating, for example, that problematic smartphone use among adolescents can be associated with higher anxiety [13], and spending less time with family and friends [14], both of which may affect sense of integration in real-world communities. However, equally, they may enable people to create stronger connections within the community they live in rather than with colleagues: rural isolation can make local social relationships even more important, and ICT can connect people in ever-evolving new ways [3]. ICT has been shown to have the potential to improve social well-being among older age groups [15]; however, this is a nuanced situation, with different age groups experiencing varied benefits from different applications of technology [16]. Smart technology and ICT are increasingly integral to society, particularly post pandemic, but the role they play can differ substantially depending on an area's rurality, history, location, and socioeconomic makeup [10].

Pets, Social Cohesion, and Well-Being

The limited body of research around pet ownership and social cohesion is constrained to dog ownership and suggests that small local parks, and specifically those used by dog owners, can contribute to neighborhood cohesion and ethnic integration [17], by facilitating regular, informal communication between dog walkers and other park users. However, Japanese researchers observed no association between dog walking and social cohesion and only a modest relationship with social capital [18]. Although acquiring a dog is associated with an increase in how much people walk, it does not affect their perception of their neighborhood's social cohesiveness [19]. Dog ownership may explain the varying beneficial effects of living near green space, with dog owners more likely to meet recommended activity targets than others [20]. In turn, this is likely to have positive implications for mental well-being. Conversely, though, older adults may spend more time at home to avoid leaving dogs alone, thus reducing opportunities for physical activity and social interactions [21]. Having a dog can be a catalyst for casual interactions in public places, although this effect has been shown to be breed-specific, being stronger with "appealing" dogs like Labradors and weaker with Rottweilers [22]. A small study of older people with and without dogs found no difference in the length of conversations people engaged in while walking, although dog owners walked more frequently; however, dog owners reported higher levels of well-being, including regular socializing with friends and relatives [23]. Pet ownership is associated with positive physiological measures, including blood pressure and cholesterol levels, and with improved coping with chronic illness [24].

There is considerably less literature dealing specifically with cat ownership rather than dog ownership, in part because there are fewer potential mechanisms for effects on health, such as regularly walking in public places [25]. However, cats are associated with slightly lower caregiver burden than dogs, particularly in terms of ill animals [26]. Research also suggests personality differences between "dog people" and "cat people." For example, an online study found that those who identified as "dog people" had higher levels of extraversion and agreeableness and lower levels of neuroticism than "cat people" [27]. Furthermore, research has shown that "cat people" score lower than "dog people" for factors including social boldness and warmth, and higher for self-reliance, indicating that they tend to be more individualistic, self-sufficient, and shy. This reinforces the popular perception that "cat people" are more introverted and supports the notion that dog ownership is associated with more casual social interactions [27], thereby enabling greater social cohesion [28]. We therefore wanted to explore the role of cat and dog ownership on social cohesion in this cohort.

The literature therefore suggests that both smart technology and pet ownership have potential associations with social cohesion, but these can be positive and negative and may be context dependent. Social cohesion itself appears to be context dependent, and there are growing calls for it to be studied at a local level to take account of this [6,29]. It is also a potentially modifiable factor influencing health [6], with the potential to be affected by technology or pets. We wanted to explore this in the relatively homogeneous cohort of the social housing communities [6] in the Smartline project.

This analysis sought to answer two questions: (1) Is there an association between the use of smart technology and levels of social cohesion or mental well-being in this cohort? (2) Is there

XSL•FO
RenderX

an association between pet ownership and levels of social cohesion or mental well-being in this cohort?

We hypothesized that smart technology overall would be associated with lower levels of social cohesion, because it would improve distant rather than local communication. In contrast, we expected dog ownership to be associated with better social cohesion. We speculated that cat ownership would be associated with worse social cohesion, on the basis that owners might be more introverted and less integrated within their community. Because of the association between social cohesion and mental well-being already shown in this cohort [6], we planned to consider the role of mental well-being in these relationships.

Methods

Overview

This is a cross-sectional cohort study of the social cohesion and well-being of participants in the Smartline project. Smartline was a collaboration between the University of Exeter, the social housing provider Coastline Housing, the local government body, Cornwall Council, and the charity Volunteer Cornwall. The project ran from March 2017 to January 2023.

Coastline Housing residents in West Cornwall (Camborne, Pool, Illogan, and Redruth) were invited by Coastline Housing to take part in the project. Because these areas have a high concentration of Coastline Housing, communities could be studied as well as individual homes. Participants were identified and approached by Coastline Housing. Smartline sought to recruit approximately 350 households to meet the requirements of the wider project. Of 2000 Coastline households, 649 were approached and 329 consented to take part. They consented to have sensors installed in their homes and to take part in associated Smartline surveys. The sensors collected home environment data such as humidity, temperature, electricity usage, and carbon dioxide levels and are reported elsewhere [6,30]. This was a convenience sample with households approached on a street-by-street basis until enough households had been recruited. Participants were primarily recruited during the day and were paid a modest incentive of £10 (US $13.5); participants were excluded if they lacked the capacity to give informed consent.

Each recruited household had a main participant aged 18 years or more, who gave written informed consent. Main participants were subsequently interviewed for the survey in their homes at a time suitable for them (typically between 9 AM and 5 PM) by teams of 2 researchers from the University of Exeter. Researchers kept anonymized, contextual field notes of incidental observations and discussed these among themselves and the wider team as relevant. Baseline data were collected between September 2017 and April 2018 (n=303), with a booster sample collected between August and November 2018 (n=26).

The survey asked for sociodemographic information about the main participant, the age and gender of other household members, and detailed information about the participant's home (including heating use, mold, damp, etc). The household's ownership of different kinds of technology and pets was also collected to further understanding of social cohesion and mental well-being, as well as indoor air quality. Participants also completed standardized measures about their perception of social cohesion [31] and their own mental well-being (Short Warwick-Edinburgh Mental Wellbeing Scale [SWEMWBS]) [32]; these were the outcome variables for the analysis presented in this paper.

Both outcome measures were coded in accordance with published protocols [31,33] and were analyzed as continuous variables. If one or two items had been missed from the social cohesion scale's 8 items or the SWEMWBS's 7 items, those items were imputed using the mean of the other responses; where 3 or more items were missing, the total score for that measure was classified as missing.

Ownership of different forms of smart technology (any smart technology, smartphone, smartwatch, or games console) was analyzed dichotomously. The survey asked for the exact number of cats or dogs participants owned; for analysis, this was categorized as 0, 1, or 2 or more.

Participants' postcodes were used to link the Index of Multiple Deprivation (IMD 2015) and rural-urban classification data to survey data. IMD or rural-urban classification data were not available for some properties, and they were therefore classified as missing.

Each cleaned, coded variable was summarized, and differences between those with and without complete data were tested.

The analyses presented are cross-sectional analyses of the baseline data. Sensitivity analyses on complete versus missing data used *t* tests, chi-square, Fisher exact, and Wilcoxon rank sum tests. We used univariate regressions to explore associations between the outcome variables of social cohesion and mental well-being, and the explanatory variables of age, gender, socioeconomic status (SES, comprising IMD decile, education, and employment status), ownership of smart technology, and pets. We then carried out multivariable regression to further explore associations for SWEMWBS. Analyses were performed using STATA 16 (StataCorp) statistical software [34]. We adhered to the STROBE (STrengthening the Reporting of OBservational studies in Epidemiology) statement checklist [35]

Ethical Considerations

The University of Exeter Business School Research Ethics Committee, which conforms to the principles embodied in the Declaration of Helsinki, granted ethical approval for the Smartline project (eUEBS002996 v4.0). All participants provided written informed consent for the collection and analysis of their data. Data were anonymized. In total, 3 participants were potentially identifiable; their data were therefore treated as missing. Participants received modest financial compensation for the study.

Results

Of 329 study participants, 308 (93.6%) provided complete data. A further 3 participants (3/329, 0.9%) were treated as missing because they were potentially identifiable.

XSL•FO
RenderX

There were no statistically significant differences between complete and missing cases in terms of age, gender, rural-urban classification, employment, social cohesion, mental well-being, ownership of smart technology, or pets. However, there were statistically significant differences in terms of IMD ranking and education. These data are shown in Table 1; meanwhile, Table S1 in the Multimedia Appendix 1 provides a description of all key variables.

XSL•FO
RenderX

Table . Participant characteristics of social housing residents in the Smartline cohort study in South West England, United Kingdom, 2017 - 2018.

Variables	Complete[a] (305)	Missing[a] (24)	*P* value[b]
Age (years), mean (SD)	54.1 (17.6)	53.4 (14.2)	.87
Gender, n (%)			.91
Male	95 (31.2)	7 (30)	
Female	210 (68.9)	17 (70)	
IMD[c] 2015			
Ranking	4446 (964 - 8548)	1512 (964 - 10,826)	.94
1st decile (most), n (%)	150 (49.2)	14 (56.5)	.001
2nd decile	47 (15.4)	2 (8.7)	
3rd decile	38 (12.5)	2 (8.7)	
4th decile	70 (23)	3 (13)	
5th decile	0 (0)	0 (0)	
6th decile	0 (0)	3 (13)	
7th decile	0 (0)	0 (0)	
8th decile	0 (0)	0 (0)	
9th decile	0 (0)	0 (0)	
10th decile (least)	0 (0)	0 (0)	
RUC[d] 2011, n (%)			.61
Urban city and town	284 (93.1)	23 (95.8)	
Other	21 (6.9)	1 (4.2)	
Education, n (%)			.02
Secondary or primary	202 (66.2)	14 (57.9)	
Further	87 (28.5)	5 (21.1)	
Higher	16 (5.3)	5 (21.1)	
Employment, n (%)			.85
In work	60 (19.7)	4 (15)	
Education or training	7 (2.3)	0	
Retired	201 (33.4)	8 (35)	
Not at work	136 (44.6)	12 (50)	
Social cohesion, mean (SD)	26.9 (6.5)	26.1 (5.6)	.60
SWEMWBS[e]	24.1 (5.2)	23.3 (7.1)	.50
Any Smart tech, n (%)			
No	45 (14.8)	3 (12.5)	≥.99
Yes	260 (85.3)	21 (87.5)	
Smartphone (Yes)	213 (69.8)	14 (58.3)	.24
Smartwatch (Yes)	23 (7.5)	1 (4.2)	≥.99
Games console (Yes)	35 (11.5)	3 (12.5)	.74
Dogs			
0	201 (65.9)	12 (50)	.27
1	74 (24.3)	9 (37.5)	
2+	30 (9.8)	3 (12.5)	
Cats			

XSL•FO
RenderX

Variables	Complete[a] (305)	Missing[a] (24)	*P* value[b]
0	212 (69.5)	18 (75)	.45
1	52 (17.1)	5 (20.8)	
2+	41 (13.4)	1 (4.2)	

[a]Values are presented as mean (SD) where data are normally distributed, median (IQR) when data were skewed, and percentages for categorical data.

[b]*t* tests, chi-square, Fisher exact, and Wilcoxon rank sum (Mann-Whitney *U*) tests were used to compare complete and missing groups.

[c]IMD: Index of Multiple Deprivation.

[d]RUC: Rural Urban Classification.

[e]SWEMWBS: Short Warwick-Edinburgh Mental Wellbeing Scale.

Among complete cases, the mean age of participants was 54 years, and around two-thirds (210/305, 69%) were female. Almost half (150/305, 49.2%) lived in the poorest 10% of English postcodes according to IMD 2015 rankings, and all were in the poorest 40%. In total, 93% (284/305) were classified as living in an urban area. Two-thirds of participants (202/305, 66.2%) had left education after primary or secondary school, with just 1 in 20 (16/305, 5.3%) having completed higher education. In total, 1 in 5 participants were currently in work (60/305, 19.7%); one-third were retired (102/305, 33.4%), and more than two-fifths were currently not in work (136/305, 44.6%).

Social cohesion scores were compared with a sample of residents of Caerphilly, Wales, which had comparable levels of deprivation and which used the same measure of social cohesion [36]. At baseline, White et al's [36] intervention group was categorized as having low, medium, or high levels of social cohesion, which we replicated in the Smartline data. White et al [36] reported 25.8% as low (Smartline 20/305, 6.6%), 36.6% as medium (Smartline 212/305, 69.5%), and 37.6% as high (Smartline 115/305, 23.9%). The mean Smartline score on SWEMWBS was 24.1 (SD 5.2); this is comparable with the UK population norm of 23.6 (SD 3.9), from the Health Survey for England 2011 [33]. A large majority (260/305, 85.3%) of the sample owned at least one form of smart technology in their household. Over two-thirds (213/305, 69.8%) had a smartphone; a smaller proportion owned a smartwatch (23/305, 7.5%) or games console (35/305, 11.5%). One-third of participants (104/305, 34.1%) had one or more dogs in the household, and slightly fewer (46/305, 30.5%) had one or more cats.

There were no statistically significant differences in social cohesion by age, gender, or socioeconomic variables. Nor were there differences by the variables of interest: household ownership of any form of smart technology, dogs, or cats. Adjusting the regressions for age, gender, and SES did not affect this. Table S2 in the Multimedia Appendix 1 lists full results on social cohesion.

There was a significant difference in SWEMWBS scores by IMD 2015 ranking; being in the second or third decile was associated with lower SWEMWBS scores. Having studied in further education was also associated with significantly lower SWEMWBS scores. In unadjusted regressions, there was a significant effect of a household owning a smartphone or owning two or more dogs being associated with lower scores on SWEMWBS. The former effect did not hold after adjusting the model. However, owning two or more dogs remained associated with lower SWEMWBS scores after adjusting for age, gender, SES, and social cohesion. Adjustments were made to the other variables of interest. Owning any smart technology was significantly associated with higher SWEMWBS scores after adjusting for age, gender, and SES, and remained so after adding social cohesion to the model. Owning a games console was associated with higher SWEMWBS scores after adjusting for age, gender, and SES, and remained so after adding social cohesion to the model. There were no significant differences in SWEMWBS scores by cat ownership. Full results on SWEMWBS can be found in Table S3 in the Multimedia Appendix 1; the fully adjusted models are presented in Figure 1.

XSL•FO
RenderX

Figure 1. Mental well-being (SWEMWBS). Adjusted regressions (adjusted for age, gender, IMD decile 2015, education, employment, and social cohesion) for participating social housing residents in the Smartline cohort study in South West England, United Kingdom, 2017 - 2018. IMD 2015: Index of Multiple Deprivation; SWEMWBS: Short Warwick-Edinburgh Mental Wellbeing Scale.

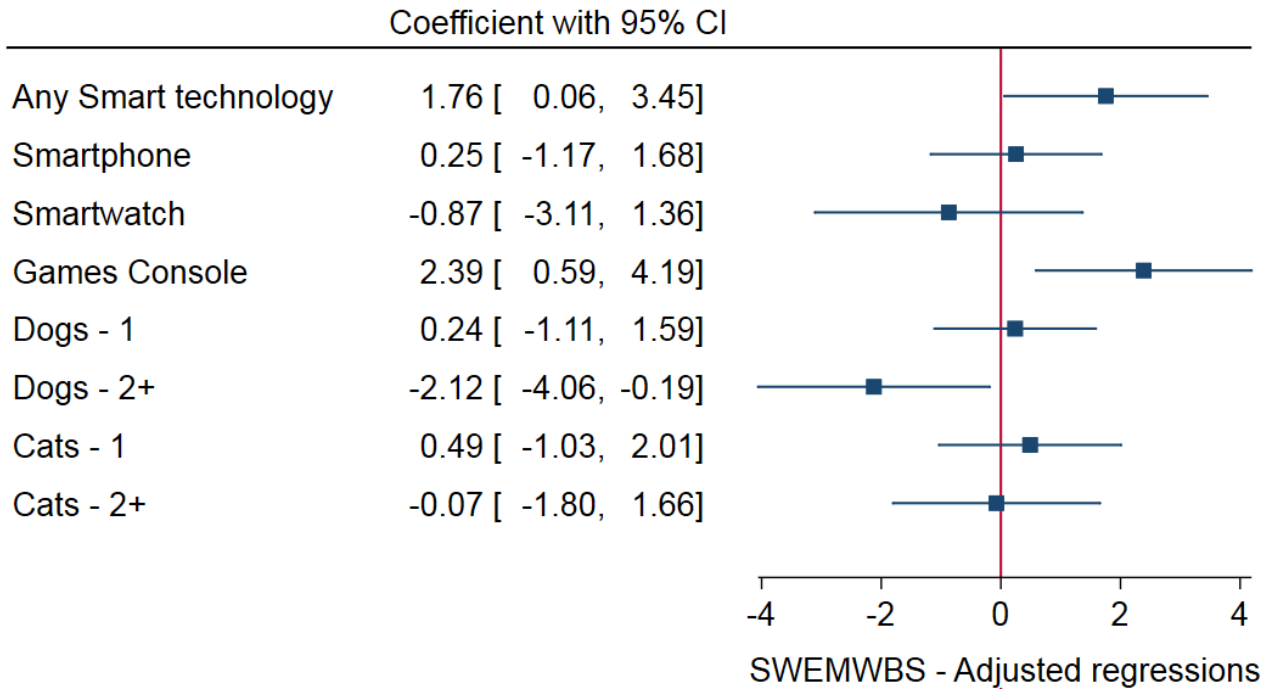

Discussion

Principal Findings

We found no evidence of increased social coherence among people with smart technology or pets among Smartline participants. This may reflect the relatively high level of uniformity in participants' socioeconomic circumstances, and levels of social cohesion were more homogeneous than in White's Welsh sample, with more than two-thirds of Smartline participants classified as having "medium" social cohesion. However, there was evidence that ownership of any form of smart technology, or a games console specifically, was associated with higher levels of mental well-being, while owning several dogs was associated with worse mental well-being. There were no associations for cat ownership.

Previous research on social cohesion and technology has had mixed results and appears to be heavily context dependent, and this may be what our neutral findings reflect. Good ICT can encourage stronger community connections, but it can also enable remote working and increased social isolation [10]. It can both exacerbate and bridge social divides [10,37], and it may prove more beneficial in places that already have a well-developed sense of local identity [38]. Miklian et al [37] suggest that smart technology is more usefully viewed as a medium, rather than as an entity which is inherently either positive or negative; it reflects the existing state of a community's social cohesion, rather than enhancing or damaging it. Subsequent to the period of our data collection, research found that digital initiatives during the pandemic supported social networks [39]. However, this came with the caveat that those with limited technological skills needed pre-existing social connections to help them access such initiatives. Our findings challenge the notion that smart technology alone is sufficient to improve community circumstances.

Owning at least one form of smart technology allows some sociodemographic groups to maintain or improve their well-being. This may be through mechanisms such as social connectivity, accessing health advice, using well-being apps, or following exercise programs. Using specific well-being apps can improve positive affect and reduce depressive symptoms [40], and improve mental well-being [41,42]. Meanwhile, problematic smartphone use, defined as poor ability to self-limit smartphone use, has been negatively associated with family health [43]. Among UK adolescents, the use of digital technology has varied effects on mental well-being depending on factors including time of the week, type of activity, and level of use, and researchers conclude that overall the level of risk to mental well-being is modest [44]. We simply recorded household ownership and therefore could not explore issues such as the type and intensity of use further.

We found that owning a games console is associated with higher mental well-being among certain groups: one possible explanation is that the socially interactive aspect of gaming may have beneficial effects. While we are unable to substantiate this from our data, a review of research into young people's use of video games suggests that, contrary to popular belief, gaming can have many beneficial effects on well-being, including emotional well-being, relationships, and social connections [45]. Furthermore, gaming can have cognitive benefits and improve people's motivation [46], both of which may have downstream benefits to well-being. Research on gameplay during the pandemic found beneficial effects, including reduced stress and anxiety, social interaction, and cognitive stimulation [47]. However, our finding could also reflect the presence of children in the home and a degree of improved well-being arising from

XSL•FO
RenderX

family support or may indicate a higher level of wealth. Our field notes indicate that at least 2 participants, one of whom had limited mobility, were using consoles for financial gain, for example, being paid to produce YouTube videos of gameplay, highlighting the potential role of consoles as a source of income and autonomy; an intriguing prospect for further exploration.

Research on dog ownership and social cohesion, although limited, suggests that dog ownership can contribute to neighborhood cohesion [17,23], although this was not supported by our data. Previous work suggests that social capital (engaging in activities with neighbors, experiencing trust, and reciprocity) has a limited [18] or no [48] association with dog ownership. In our data, owning 2 or more dogs was associated with lower levels of mental well-being: we hypothesize that owning several dogs places a burden on households in certain groups, whether in terms of time, finances, noise, or allergens: future research could explore this further. Our data were collected before COVID-19; however, the pandemic may have exacerbated such challenges. Many people acquired dogs for the first time during the pandemic, with less ability to preresearch their choice than in normal times, and less focus on canine health, potentially increasing the burdens of dog ownership [49]. Pets can replace human company, particularly for older people: Age UK reports that half of the people aged 65 and over in the United Kingdom rely on pets or television for companionship [50]. However, they may not replace human company adequately, so some people may own dogs specifically because they have poor mental health or are lonely [51]. Loneliness is a particular concern among older age groups, and further research is needed [52]. Parslow et al [53] also found negative effects of pet ownership among older Australians. Those with pets had higher rates of depression than those without, as well as worse physical health among women; they hypothesized that owning a pet meant carrying out more chores, and this effect would be multiplied with several pets. This may be particularly true among older age groups [54]. We also consider the possibility that owning a dog could reflect concerns about safety within a neighborhood, which could be explored explicitly in future studies.

We did not find any differences in mental well-being by cat ownership. There is limited research on cat ownership compared with dog ownership, and there have been calls to consider specific activities with cats, rather than simply “ownership,” in order to take account of potentially positive activities, such as play and tactile interactions, as opposed to those that may prove stressful, such as veterinary visits and illness [25].

Social housing organizations increasingly take on a role of supporting social interaction among residents [55,56]. Social cohesion during and after the pandemic was mediated by the efforts made by local organizations, with communities that invested in social cohesion experiencing better social outcomes [57]. Thus, communities living in social housing may have fared better.

Limitations

Data were collected before the COVID-19 pandemic, which saw changes in perceived social cohesion, both positive and negative [57,58], as well as in pet ownership and use of smart technology. A low proportion of the sample reported currently being in work; this is likely to reflect the fact that recruitment took place during the day, when employed people were less likely to be home. It may also partially explain the low proportion of people reporting further or higher education and the relatively high mean age of the participants. We collected the main participants’ personal perceptions of social cohesion and their mental well-being; however, ownership of smart technology and pets related to the whole household; the main respondent may not have been the user of the technology or the main carer for the pet. We therefore could not control for further variables such as frequency or purpose of smart technology use, or of subjective relationships with pets: this would be an interesting avenue for further research. The sample size was modest, with some small cell sizes: it was not feasible to carry out further analyses, and the findings should be treated with caution. These cross-sectional data do not provide insights into the direction of any causality.

Conclusions

Smart technology and ICT can have mixed effects on social cohesion and mental well-being and should not be treated as a panacea for community health. Among our sample of social housing residents in Cornwall, there was some evidence of associations between mental well-being and smart technology. Previous research suggests that dog ownership can have modest beneficial effects on health, well-being, and social interactions; our findings contradict this, showing that owning 2 or more dogs is associated with lower levels of mental well-being. People own smart technology and pets for different reasons, and they affect people’s lives in different ways, both good and bad. This may account for their mixed effects on social cohesion and well-being in this study, where positive and negative effects may have canceled each other out. Subsequent to this fieldwork, such effects may have been exacerbated by the COVID-19 pandemic [59]. Further research could clarify the question of who owns and uses a given form of technology within a household, and who takes care of a pet, to provide more nuance in analyzing psychosocial outcomes.

Acknowledgments

The authors would like to acknowledge the contribution of the whole Smartline team, including all the partners, for collecting the data.

The Smartline project received up to £3,780,374 (US $5,134,657) of funding from the England European Regional Development Fund as part of the European Structural and Investment Funds Growth Programme 2014-2020. The Ministry of Housing, Communities, and Local Government (and in London, the intermediate body Greater London Authority) is the Managing Authority

XSL•FO
RenderX

for the European Regional Development Fund. Established by the European Union, the European Regional Development Fund helps local areas stimulate their economic development by investing in projects that will support innovation, businesses, create jobs, and local community regenerations. Additional funding of £25,000 (US $33,962) was from the South West Academic Health Science Network.

Disclaimer

We confirm that generative AI was not used in any part of the manuscript writing.

Data Availability

An anonymized version of the dataset analyzed during this study is available in the UK Data Service repository [60].

Conflicts of Interest

None declared.

Multimedia Appendix 1
Further data for social cohesion, mental well-being, and the role of smart technology and pet ownership among social housing residents: a cross-sectional cohort study.
[PDF File, 221 KB - formative_v9i1e75451_app1.pdf]

References

1. O'Shea N, Bell A. A spending review for wellbeing. briefing. : Centre for Mental Health.
2. Deeming C. Addressing the Social determinants of subjective wellbeing: the latest challenge for social policy. J Soc Policy 2013 Jul;42(3):541-565. [doi: 10.1017/S0047279413000202] [Medline: 23710101]
3. Wallace C, Vincent K. Community well-being and information technology. In: Handbook of Community Well-Being Research 2017:169-188. [doi: 10.1007/978-94-024-0878-2_9]
4. Harrison EK, Quick A, Abdallah S, et al. Looking through the wellbeing kaleidoscope: results from the European Social Survey. : European Social Survey; 2016.
5. Cornwall-UK. Steps towards a digital rural region. : European Network for Rural Development; 2018.
6. Williams AJ, Maguire K, Morrissey K, Taylor T, Wyatt K. Social cohesion, mental wellbeing and health-related quality of life among a cohort of social housing residents in Cornwall: a cross sectional study. BMC Public Health 2020 Jun 22;20(1):985. [doi: 10.1186/s12889-020-09078-6] [Medline: 32571296]
7. Thapa D, Sein MK, Sæbø Ø. Building collective capabilities through ICT in a mountain region of Nepal: where social capital leads to collective action. Information Technology for Development 2012 Jan;18(1):5-22. [doi: 10.1080/02681102.2011.643205]
8. Broadbent R, Papadopoulos T. Impact and benefits of digital inclusion for social housing residents. Community Dev (Columb) 2013 Feb;44(1):55-67. [doi: 10.1080/15575330.2012.662990]
9. Stuart A, Yan RJ, Harkin L, et al. Design of a digital intervention to improve connectivity and reduce loneliness in older adults (preprint). JMIR Formative Research. Preprint posted online on Aug 31, 2022. [doi: 10.2196/preprints.42172]
10. Wallace C, Vincent K, Luguzan C, Townsend L, Beel D. Information technology and social cohesion: A tale of two villages. J Rural Stud 2017 Aug;54:426-434. [doi: 10.1016/j.jrurstud.2016.06.005]
11. Sehmer E. As a child psychiatrist, I see what smartphones are doing to kids' mental health – and it's terrifying.: The Guardian URL: https://www.theguardian.com/commentisfree/2025/jan/03/psychiatrist-children-smartphones-mental-health-harm [accessed 2025-08-05]
12. Jonathan H. The Anxious Generation: How the Great Rewiring of Childhood Is Causing an Epidemic of Mental Illness 2025:385 URL: https://www.penguin.co.uk/books/456971/the-anxious-generation-by-haidt-jonathan/9781802063271 [accessed 2025-08-05]
13. Carter B, Ahmed N, Cassidy O, et al. "There's more to life than staring at a small screen": a mixed methods cohort study of problematic smartphone use and the relationship to anxiety, depression and sleep in students aged 13-16 years old in the UK. BMJ Ment Health 2024 Jul 31;27(1):e301115. [doi: 10.1136/bmjment-2024-301115] [Medline: 39085034]
14. Kalk NJ, Downs J, Clark B, Carter B. Problematic smartphone use: What can teenagers and parents do to reduce use? Acta Paediatr 2024 Oct;113(10):2177-2179. [doi: 10.1111/apa.17365] [Medline: 39084632]
15. Hasan H, Linger H. Enhancing the wellbeing of the elderly: social use of digital technologies in aged care. Educ Gerontol 2016 Nov;42(11):749-757. [doi: 10.1080/03601277.2016.1205425]
16. Lavoie R, Zheng Y. Smartphone use, flow and wellbeing: a case of Jekyll and Hyde. Comput Human Behav 2023 Jan;138:107442. [doi: 10.1016/j.chb.2022.107442]
17. Gómez E, Baur JWR, Malega R. Dog park users: an examination of perceived social capital and perceived neighborhood social cohesion. J Urban Aff 2018 Apr 3;40(3):349-369. [doi: 10.1080/07352166.2017.1343634]

XSL•FO
RenderX

18. Koohsari MJ, Yasunaga A, Shibata A, et al. Dog ownership, dog walking, and social capital. Humanit Soc Sci Commun 2021;8(1). [doi: 10.1057/s41599-021-00804-y]
19. Cutt HE, Knuiman MW, Giles-Corti B. Does getting a dog increase recreational walking? Int J Behav Nutr Phys Act 2008 Mar 27;5(17):5868-5865. [doi: 10.1186/1479-5868-5-17] [Medline: 18366804]
20. White MP, Elliott LR, Wheeler BW, Fleming LE. Neighbourhood greenspace is related to physical activity in England, but only for dog owners. Landsc Urban Plan 2018 Jun;174(May 2017):18-23. [doi: 10.1016/j.landurbplan.2018.01.004]
21. Costa S, Sousa L, Luz H, Padeiro M. Daily mobility and social interactions among community-dwelling older adults with pet dogs: a scoping review. J Appl Gerontol 2022 Dec;41(12):2609-2623. [doi: 10.1177/07334648221116633] [Medline: 36029015]
22. Wells DL. The facilitation of social interactions by domestic dogs. Anthrozoös 2004 Dec;17(4):340-352. [doi: 10.2752/089279304785643203]
23. Rogers J, Hart LA, Boltz RP. The role of pet dogs in casual conversations of elderly adults. J Soc Psychol 1993 Jun;133(3):265-277. [doi: 10.1080/00224545.1993.9712145] [Medline: 8412041]
24. Walsh F. Human-animal bonds I: the relational significance of companion animals. Fam Process 2009 Dec;48(4):462-480. [doi: 10.1111/j.1545-5300.2009.01296.x] [Medline: 19930433]
25. Ravenscroft SJ, Barcelos AM, Mills DS, Banks J. Cat-human related activities associated with human well-being. Hum Anim Interact Bull 2021. [doi: 10.1079/hai.2021.0006]
26. Spitznagel MB, Gober MW, Patrick K. Caregiver burden in cat owners: a cross-sectional observational study. J Feline Med Surg 2023 Jan;25(1):1098612X221145835. [doi: 10.1177/1098612X221145835] [Medline: 36705958]
27. Gosling SD, Sandy CJ, Potter J. Personalities of self-identified “dog people” and “cat people”. Anthrozoös 2010 Sep;23(3):213-222. [doi: 10.2752/175303710X12750451258850]
28. van den Berg P, Sharmeen F, Weijs-Perrée M. On the subjective quality of social Interactions: Influence of neighborhood walkability, social cohesion and mobility choices. Transportation Research Part A: Policy and Practice 2017 Dec;106:309-319. [doi: 10.1016/j.tra.2017.09.021]
29. Walton E. The meaning of community in diverse neighborhoods: stratification of influence and mental health. Health Place 2018 Mar;50:6-15. [doi: 10.1016/j.healthplace.2018.01.001] [Medline: 29331786]
30. Menneer T, Qi Z, Taylor T, et al. Changes in domestic energy and water usage during the UK COVID-19 lockdown using high-resolution temporal data. Int J Environ Res Public Health 2021 Jun 25;18(13):6818. [doi: 10.3390/ijerph18136818] [Medline: 34202018]
31. Buckner JC. The development of an instrument to measure neighborhood cohesion. American J of Comm Psychol 1988 Dec;16(6):771-791. [doi: 10.1007/BF00930892]
32. Ng Fat L, Scholes S, Boniface S, Mindell J, Stewart-Brown S. Evaluating and establishing national norms for mental wellbeing using the short Warwick-Edinburgh Mental Well-being Scale (SWEMWBS): findings from the Health Survey for England. Qual Life Res 2017 May;26(5):1129-1144. [doi: 10.1007/s11136-016-1454-8] [Medline: 27853963]
33. Stewart-Brown S. The warwick-edinburgh mental wellbeing scales - WEMWBS. Warwick Medical School. 2018. URL: https://warwick.ac.uk/fac/sci/med/research/platform/wemwbs/ [accessed 2020-02-07]
34. StataCorp. Stata statistical software: release 16. 2019. URL: https://www.stata.com/ [accessed 2019-09-29]
35. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol 2008 Apr;61(4):344-349. [doi: 10.1016/j.jclinepi.2007.11.008]
36. White J, Greene G, Farewell D, et al. Improving mental health through the regeneration of deprived neighborhoods: a natural experiment. Am J Epidemiol 2017 Aug 15;186(4):473-480. [doi: 10.1093/aje/kwx086] [Medline: 28486637]
37. Miklian J, Hoelscher K. Smart cities, mobile technologies and social cohesion in India. Indian Journal of Human Development 2017 Apr;11(1):1-16. [doi: 10.1177/0973703017712871]
38. Beel D, Wallace C. Gathering together: social capital, cultural capital and the value of cultural heritage in a digital age. Social & Cultural Geography 2020 Jun 12;21(5):697-717. [doi: 10.1080/14649365.2018.1500632]
39. Phillips RH. The role of ICT in maintaining social cohesion: understanding the potential of digital initiatives for social networks in rural areas. Rural Sociol 2024 Sep;89(3):545-573 [FREE Full text] [doi: 10.1111/ruso.12557]
40. Howells A, Ivtzan I, Eiroa-Orosa FJ. Putting the ‘app’ in happiness: a randomised controlled trial of a smartphone-based mindfulness intervention to enhance wellbeing. J Happiness Stud 2016 Feb;17(1):163-185. [doi: 10.1007/s10902-014-9589-1]
41. Bakker D, Rickard N. Engagement with a cognitive behavioural therapy mobile phone app predicts changes in mental health and wellbeing: MoodMission. Aust Psychol 2019 Aug 1;54(4):245-260. [doi: 10.1111/ap.12383]
42. McEwan K, Richardson M, Sheffield D, Ferguson FJ, Brindley P. A smartphone app for improving mental health through connecting with urban nature. Int J Environ Res Public Health 2019 Sep 12;16(18):3373. [doi: 10.3390/ijerph16183373] [Medline: 31547286]
43. Guo N, Wang MP, Luk TT, et al. The association of problematic smartphone use with family well-being mediated by family communication in Chinese adults: a population-based study. J Behav Addict 2019 Sep 1;8(3):412-419. [doi: 10.1556/2006.8.2019.39] [Medline: 31394912]

XSL•FO
RenderX

44. Przybylski AK, Weinstein N. A large-scale test of the goldilocks hypothesis. Psychol Sci 2017 Feb;28(2):204-215. [doi: 10.1177/0956797616678438] [Medline: 28085574]
45. Johnson D, Jones C, Scholes L, Carras MC. Videogames and wellbeing: a comprehensive review. 2013 URL: https://eprints.qut.edu.au/105915/ [accessed 2020-02-10]
46. Granic I, Lobel A, Engels RCME. The benefits of playing video games. Am Psychol 2014 Jan;69(1):66-78. [doi: 10.1037/a0034857] [Medline: 24295515]
47. Barr M, Copeland-Stewart A. Playing video games during the COVID-19 pandemic and effects on players' well-being. Games and Culture 2022 Jan;17(1):122-139. [doi: 10.1177/15554120211017036]
48. Martin E. The effects of dog ownership and social capital. 2024. URL: https://illustro-iadt.figshare.com/articles/thesis/The_effects_of_dog_ownership_on_social_capital/25164422?file=44448584 [accessed 2024-11-27]
49. Packer RMA, Brand CL, Belshaw Z, Pegram CL, Stevens KB, O'Neill DG. Pandemic puppies: characterising motivations and behaviours of UK owners who purchased puppies during the 2020 COVID-19 pandemic. Animals (Basel) 2021;11(9):2500. [doi: 10.3390/ani11092500]
50. Davidson S, Rossall P. Age UK evidence review: loneliness in later life. 2015 URL: https://www.ageuk.org.uk/siteassets/documents/reports-and-publications/reports-and-briefings/health--wellbeing/rb_june15_lonelines_in_later_life_evidence_review.pdf [accessed 2020-02-07]
51. Müllersdorf M, Granström F, Sahlqvist L, Tillgren P. Aspects of health, physical/leisure activities, work and socio-demographics associated with pet ownership in Sweden. Scand J Public Health 2010 Feb;38(1):53-63. [doi: 10.1177/1403494809344358] [Medline: 19717574]
52. Bruggencate TT, Luijkx KG, Sturm J. Social needs of older people: a systematic literature review. Ageing Soc 2018 Sep;38(9):1745-1770. [doi: 10.1017/S0144686X17000150]
53. Parslow RA, Jorm AF, Christensen H, Rodgers B, Jacomb P. Pet ownership and health in older adults: findings from a survey of 2,551 community-based Australians aged 60-64. Gerontology 2005;51(1):40-47. [doi: 10.1159/000081433] [Medline: 15591755]
54. Wells DL. The effects of animals on human health and well‐being. Journal of Social Issues 2009 Sep;65(3):523-543. [doi: 10.1111/j.1540-4560.2009.01612.x]
55. Purkis A. Housing associations in england and the future of voluntary organisations. The Baring Foundation. URL: https://baringfoundation.org.uk/ [accessed 2020-02-07]
56. Blessing A. Magical or monstrous? Hybridity in social housing governance. Hous Stud 2012 Mar;27(2):189-207. [doi: 10.1080/02673037.2012.649469]
57. Lalot F, Abrams D, Broadwood J, Davies Hayon K, Platts-Dunn I. The social cohesion investment: communities that invested in integration programmes are showing greater social cohesion in the midst of the COVID-19 pandemic. J Community Appl Soc Psychol 2022;32(3):536-554. [doi: 10.1002/casp.2522] [Medline: 34230795]
58. Borkowska M, Laurence J. Coming together or coming apart? Changes in social cohesion during the COVID-19 pandemic in England. European Societies 2021 Feb 19;23(sup1):S618-S636. [doi: 10.1080/14616696.2020.1833067]
59. Mikolai J, Keenan K, Kulu H. Intersecting household-level health and socio-economic vulnerabilities and the COVID-19 crisis: an analysis from the UK. SSM Popul Health 2020 Dec;12(July):100628. [doi: 10.1016/j.ssmph.2020.100628] [Medline: 32838017]
60. Woods RD, Menneer T, Wellaway I, et al. Smartline environmental sensor data and utility usage, 2017–2023. Smartline Environmental Sensor Data and Utility Usage 2023. [doi: 10.5255/UKDA-SN-856596]

Abbreviations

ICT: internet and communications technologies
IMD: Index of Multiple Deprivation
SES: socioeconomic status
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
SWEMWBS: Short Warwick-Edinburgh Mental Wellbeing Scale

Edited by A Mavragani; submitted 03.04.25; peer-reviewed by S Zhu, Z Zhang; revised version received 05.08.25; accepted 29.08.25; published 24.10.25.

Please cite as:
Rahtz E, Williams AJ, Taylor TJ
Social Cohesion, Mental Well-Being, and the Role of Smart Technology and Pet Ownership Among Social Housing Residents: Cross-Sectional Cohort Study
JMIR Form Res 2025;9:e75451
URL: https://formative.jmir.org/2025/1/e75451
doi:10.2196/75451

XSL•FO
RenderX

© Emmylou Rahtz, Andrew James Williams, Timothy J Taylor. Originally published in JMIR Formative Research (https://formative.jmir.org), 24.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Digital Mental Health Coaching in Clinically Diverse Populations: Controlled Engagement and Outcomes Study

Alison Pickover, PhD; Sarah Adler, PsyD

Wave Life, Inc., 700 El Camino Real, Suite 120, Menlo Park, CA, United States

Corresponding Author:
Alison Pickover, PhD
Wave Life, Inc., 700 El Camino Real, Suite 120, Menlo Park, CA, United States

Abstract

Background: Digital coaching programs, offering virtually delivered mental health care by coaches and companion apps, are an increasingly popular care model designed to increase accessibility and reduce strain on traditional mental health care systems. Initial studies suggest these programs can produce a range of positive mental health outcomes; however, methodological limitations and a focus on homogeneous, subclinical populations have constrained conclusions about their effectiveness, especially in diverse and clinically severe samples.

Objective: This study aimed to evaluate the impact of an evidence-based digital mental health coaching program in a clinically and demographically diverse sample. The study compared engagement with app-based content and changes in depression, anxiety, and stress symptoms, over the course of 1 month, among users who received coaching versus those who used the app alone (controls).

Methods: Program users (N=64) were categorized as coaching users (attending at least 1 session) or controls (app-only users). Depression, anxiety, and stress symptoms were assessed using the Depression Anxiety Stress Scale-21 at baseline and after 30 days. Engagement with app content was also measured. Between-group differences were analyzed using *t* tests and mixed multivariate analysis of covariance models, with follow-up sensitivity analysis of covariance analyses (controlling for age).

Results: Participants were diverse in terms of demographics and clinical severity, with half reporting severe to extremely severe depression and nearly half reporting severe to extremely severe anxiety or stress at baseline. A repeated-measures multivariate analysis of covariance revealed a significant group-by-time interaction (P=.02), indicating greater symptom reduction among coaching users, primarily driven by changes in anxiety and stress. Follow-up analyses of covariance exploring symptom-specific patterns, excluding participants with subclinical baseline symptoms, yielded significant group-by-time interactions across depression (P=.04), anxiety (P=.003), and stress (P=.03). Engagement with app-based content did not significantly differ between the groups (P=.20), suggesting coaching's effectiveness was not contingent on differential app usage.

Conclusions: This study demonstrates that digital mental health coaching can significantly improve clinical outcomes, even in diverse and clinically severe populations. These findings challenge the notion that coaching is only effective for subclinical or high-functioning individuals and highlight its potential to extend the reach of mental health care to underserved communities.

(JMIR Form Res 2025;9:e71346) doi:10.2196/71346

KEYWORDS

coaching; digital health; mental health; transdiagnostic; outcome; depression; anxiety; stress

Introduction

Mental health issues are common, with 1 in 5 adults experiencing mental health difficulties each year [1]. These issues are a significant source of disability, outranking physical health conditions (eg, cardiovascular and circulatory diseases) in terms of global disease burden [2]. They also frequently co-occur with physical health concerns and contribute to poorer health outcomes. Mental health issues affect workplace productivity and are key contributors to absenteeism and turnover. Global estimates of mental health–related productivity loss exceed US $1 trillion annually [3].

Given the individual, social, and economic ramifications of poor mental health, increasing access to effective mental health care services is a crucial priority. Currently, more than half of those with mental illness do not receive treatment, largely due to the dearth of licensed providers available to treat mental health conditions [4,5]. In the US, where 150 million adults seek services annually [6,7], only 1.2 million behavioral health providers are available to provide care [8]. Half of Americans live in a mental health workforce shortage area [4,9], and workforce projections suggest this condition will only worsen with time [10].

XSL•FO
RenderX

Against this backdrop, there has arisen substantial innovation in approaches to mental health care delivery. Key developments include the provision of care virtually via telehealth, the inclusion of digital health tools and content, and service delivery by lay health workers like coaches. These shifts are notable for the following reasons: in the midst of a provider shortage marked by geographic disparity, they increase access to care (irrespective of location); and, without sacrificing the "human element," they reduce the cost of providing health care. These approaches are also valuable levers for improving treatment retention [11] and for achieving health equity among vulnerable and underserved communities [12,13].

Digital coaching programs, offering virtually delivered care by coaches and companion apps, are an increasingly popular care model or care model component that seeks to increase care accessibility and reduce strain on mental health care systems. In such programs, services are provided by coaches who have received mental health care training or credentialing, but not necessarily licensure. These coaches serve two valuable and complementary functions: (1) helping individuals set and achieve goals via supportive accountability and skill-building and (2) improving user engagement with app-based digital resources to ensure maximum benefit [14,15]. Preliminary studies suggest that digital coaching programs can produce a range of positive mental health outcomes in select populations. For instance, in employee samples, digital coaching programs are associated with within-group improvements in well-being [16,17], emotional intelligence [16], self-efficacy, emotion regulation, social connectedness [18], and stress [17]. In at least 1 study, coaching has been shown to impact employees' levels of depression and anxiety as well [19].

Although initial data regarding the effectiveness of digital coaching programs look promising, critical methodological issues limit enthusiasm. First, available studies lack a control condition, and thus, it is not clear that coaching is responsible for observed improvements. Second, many studies, by virtue of evaluating employee benefits programs, focus on homogeneous samples with subclinical or mild to moderate symptomatology. As such, it remains unclear whether these programs yield positive outcomes in more representative and diverse samples. Research that improves upon these limitations is necessary to realize the true potential of coaching in driving clinical outcomes and reducing care disparities.

In this study, we sought to evaluate the impact of a digital coaching program in a clinically and demographically diverse sample. Specifically, we evaluated differences in content engagement and reductions in depression, anxiety, and stress over the course of a month of engagement with the digital coaching program Wave. To assess whether coaching was a key driver of outcomes, we compared engagement and symptom reductions among coaching users to Wave users who downloaded Wave's app (developed by Wave Life, Inc.) but did not engage in coaching. Based on previous literature [20] and the coach's hypothesized role in fostering supportive accountability [15], we predicted greater content engagement by coaching users than their counterparts. Given some evidence that coaching can reduce clinical symptoms [19], we predicted reductions in depression, anxiety, and stress among users who engaged in coaching and predicted that these reductions would be steeper than any reductions observed among users who did not engage in coaching.

Methods

Study Design and Participants

Participants were a cohort of Wave users (N=64) who created accounts between April 2023 and May 2024. Account creation was restricted to individuals 18 years of age or older who resided in the United States. Participants either had access to Wave as an employment benefit (11/64, 17%) or accessed Wave independently by downloading Wave's app in the Apple or Google Play Stores (53/64, 83%); independent users typically learned about Wave through targeted social media advertisements (eg, Facebook [Meta], Instagram [Meta], and LinkedIn [LinkedIn Corporation]) or through personal recommendations and word-of-mouth referrals. Wave users were included in the study if they provided symptom data at at least 2 timepoints, within a 45-day interval, during their course of using Wave. Users were prompted to complete symptom measures in advance of their first coaching session or at the end of account creation, and they received subsequent prompts to complete symptom measures at 30-day intervals. Use of Wave's app and coaching services was voluntary, and users had the option to cease engagement at any time. Measure completion, app use, and coaching took place remotely.

Measures

Demographic Characteristics

Date of birth was self-reported at the time of account creation and transformed to age. In addition, users had the opportunity to fill out an in-app demographic survey assessing gender, race and ethnicity, and relationship status. Completion of the demographic survey was optional.

Depression, Anxiety, and Stress

Users completed the Depression Anxiety Stress Scale-21 (short version; DASS-21) [21] twice via an in-app, self-report questionnaire. The DASS-21 is a well-validated [22-24] measure comprising 3 scales, each with 7 items. Item responses were summed to create continuous scores of depression, anxiety, and stress symptom severity; higher scores indicated greater severity. In addition, using established thresholds [21], the DASS-21 was used to classify depression, anxiety, and stress symptoms as normal, mild, moderate, severe, or extremely severe, allowing for categorical interpretation of symptom severity.

Content Engagement

Content engagement was measured as the number of in-app resources (short-form written, audio, and video content) viewed by a user during their time on the platform. Due to evolving methods of tracking user engagement with content, data were only available for 25 of 38 (66%) coaching users; however, content engagement data were available for 100% of controls. Available data regarding content engagement was recorded passively through the app.

XSL•FO
RenderX

Intervention

Users received one-on-one behavioral health coaching and access to Wave's mobile app. Users interested in coaching services were given the opportunity to schedule sessions upon account creation or at a later date. Sessions could be scheduled the same day or for the future. Coaching services were provided via remote video visits. Typically, sessions occurred weekly and lasted 30 minutes; however, users could request alternative session intervals or longer sessions. Coaches helped users create customized care programs based on their unique needs; they assisted users in defining and achieving wellness goals via psychoeducation and skills acquisition. Coaches used a transdiagnostic approach to case conceptualization and incorporated skills associated with evidence-based therapies such as acceptance and commitment therapy and cognitive behavioral therapy. Coaches had access to users' DASS-21 responses, which they reviewed with users in session, to help focus sessions and as a means of tracking progress over time.

All Wave coaches were bachelor's and master's-level practitioners who received certification from the National Board for Health and Wellness Coaching. This certification is achieved by completing an approved training program, acquiring 50 posttraining practice sessions, and passing an exam cosponsored by the National Board of Medical Examiners. Wave provided additional coach training covering foundational practice components (eg, rapport, supportive accountability, and cultural humility), Wave's proprietary technology and app, mental health first aid, motivational interviewing, case conceptualization, risk assessment and management, and use of quantitative data in measurement-based care. Training included both didactics and role-play exercises. While providing services at Wave, coaches attended a weekly meeting for case review and consultation.

All Wave users, regardless of their coaching status, received access to Wave's app. The app provided a variety of tools and content designed to support users in managing their health and emotional well-being. Content included short-form written, audio, and video content and spanned psychoeducation, guided practice, and skill-building. All content was reviewed for accuracy by licensed clinical psychologists.

Analytic Plan

To compare progress over time, users were separated into 2 groups: coaching users (n=36) attended at least 1 coaching session between their first and second DASS-21 completions; controls (n=28) attended zero sessions in that interval.

Data were checked for outliers, skew, and kurtosis prior to all analyses; age and content engagement were shown to be (positively) skewed. To identify potential confounds, *t* tests and chi-square tests were used to assess between-group differences in demographic characteristics (age was log-transformed), baseline clinical symptom severity, and content engagement (square root-transformed).

A repeated-measures multivariate analysis of covariance (MANCOVA) was used to assess the impact of coaching on symptom change over time, controlling for age. MANCOVA was selected to account for the high correlation among the 3 dependent variables (depression, anxiety, and stress, continuously assessed), reduce the risk of type I error associated with multiple univariate tests, and allow for the evaluation of an overall multivariate interaction effect. Prior to this analysis, the data were checked for multivariate normality. Based on visual inspection of the data and a significant Shapiro-Wilk test, follow-up depression scores were square-root transformed for this analysis. Multivariate outliers, assessed using Mahalanobis distance, were not observed. Inspection of scatterplot matrices revealed linear relationships between the dependent variables, and bivariate correlations (r<.90) ruled out multicollinearity concerns. Box M test (P=.01) was used to confirm homogeneity of covariance matrices.

To better understand symptom-specific intervention effects among participants with elevated symptoms, follow-up analyses of covariance (ANCOVAs) were conducted as sensitivity analyses, excluding participants who scored in the "normal" range at baseline on the corresponding DASS-21 scale. For example, the depression-specific ANCOVA was conducted on data drawn from users who initially scored in the mild, moderate, severe, or extremely severe ranges on the DASS-21 depression scale. This approach was used to reduce potential floor effects and to ensure sufficient variability for detecting change among individuals with room for improvement in each domain.

Ethical Considerations

All data were deidentified prior to analysis, and the use of human participant data was reviewed and determined to be exempt by an institutional review board (Advarra Pro00079304). Users provided consent for the use of their deidentified data for research purposes by accepting the platform's Privacy Policy at account creation. Users were not compensated for their use of Wave or completion of measures.

Results

Demographic Characteristics and Initial Clinical Severity

Sample descriptive statistics are presented in Table 1. Table 2 summarizes symptom severity among the sample.

Participants were diverse in terms of their demography and clinical presentation. Half of the sample reported severe to extremely severe depression, and nearly half reported severe to extremely severe anxiety or stress. On average, coaching users were older than controls ($\text{mean}_{\text{coaching}}$ 35.78, SD 12.10 years; $\text{mean}_{\text{control}}$ 28.46, SD 9.10 years; t_{62}=–2.78; P=.007). Chi-square tests showed that coaching and control groups did not significantly differ in gender (P=.09), race and ethnicity (P=.38), or relationship status (P=.63). Chi-square tests showed that the distribution of participants across clinical severity categories (mild to extremely severe) did not differ significantly between coaching users and controls; this held true for depression (P=.23), anxiety (P=.10), and stress (P=.57) severity. Likewise, no significant baseline differences in symptom severity were found when assessing severity continuously (P=.62, P=.43, and P=.10, for depression, anxiety, and stress, respectively). The average time between DASS-21 responses was 34.77 (SD 5.33) days.

XSL•FO
RenderX

Table . Sample descriptive statistics.[a]

Characteristics	Full sample (N=64)	Coaching users (n=36)	Control group (n=28)
Age (years), mean (SD)	32.58 (11.41)	35.78 (12.10)	28.46 (9.10)
Gender, n (%)			
Woman	36 (56.3)	16 (44.4)	20 (71.4)
Man	13 (20.3)	11 (30.6)	2 (7.1)
Other not listed	3 (4.7)	2 (5.6)	1 (3.6)
Agender	1 (1.6)	1 (2.8)	0 (0)
Gender nonconforming	1 (1.6)	0 (0)	1 (3.6)
Prefer not to answer or no response	10 (15.6)	6 (16.6)	4 (14.3)
Race and ethnicity[b], n (%)			
White	35 (54.7)	20 (55.6)	14 (50)
Black or African American	11 (17.2)	5 (13.9)	6 (21.4)
Asian	9 (14.1)	5 (13.9)	4 (14.3)
Hispanic or Latino	10 (15.6)	2 (5.6)	7 (25)
Native American or Alaska Native	2 (3.1)	1 (2.8)	1 (3.6)
Prefer not to answer or no response	5 (7.8)	3 (8.3)	2 (7.1)
Relationship status, n (%)			
Single	18 (28.1)	8 (22.2)	10 (35.7)
Partnered	19 (29.7)	7 (19.4)	12 (42.9)
Married	8 (12.5)	5 (13.9)	3 (10.7)
Separated or divorced	3 (4.7)	2 (5.6)	1 (3.6)
Prefer not to answer or no response	16 (25)	14 (38.9)	2 (7.1)

[a]Demographic characteristics of a cohort of US adult users of the Wave digital mental health platform, including users who received remote coaching (n=36) and controls who only used the app (n=28), are listed. Data were collected remotely between April 2023 and May 2024. Users self-reported date of birth (transformed to age) upon account creation and provided other demographic data via an optional, in-app questionnaire.

[b]Users could select more than 1 response option for race and ethnicity.

Table . Symptom severity.[a]

Symptoms	Baseline, mean (SD)		30 - 45 days later, mean (SD)	
	Coaching	Control	Coaching	Control
Depression	21.22 (12)	22.64 (10.58)	14.22 (10.6)	20.07 (11.46)
Anxiety	13.78 (6.25)	15.36 (8.80)	9.00 (6.11)	15.21 (8.85)
Stress	20.56 (8.63)	23.93 (7.78)	15.89 (8.78)	22.57 (8.43)

[a]Depression Anxiety Stress Scale-21 (DASS-21) scores at baseline and 30 - 45 days later among a cohort of the US adult users of the Wave digital mental health platform, including users who received remote coaching (n=36) and controls who only used the app (n=28), are listed. Data were collected remotely via in-app measures between April 2023 and May 2024.

Engagement

Coaching users attended a median of 4 (IQR 4-5) sessions between first and second DASS-21 measurements. On average, they engaged more heavily with content (median 21, IQR 8-38 pieces of content) than their control counterparts (median 11, IQR 3.25-29.75 pieces of content) during this period; however, this difference did not reach statistical significance (t_{51}=–1.30; P=.20).

Clinical Improvement

MANCOVA revealed a significant group-by-time interaction (Wilks λ=0.85, $F_{3,59}$=3.55, partial η²=0.153; P=.02), indicating that symptom changes over time varied by condition across the

XSL•FO
RenderX

combined outcomes of depression, anxiety, and stress. Follow-up univariate tests showed a significant group-by-time interaction for anxiety (P=.004) and stress (P=.04). No significant interaction effect (=.67) or main effect of time (=.56) was observed for depression.

Sensitivity Analyses

In the sensitivity analyses, significant group-by-time interactions were observed for depression, anxiety, and stress, with coaching users faring better over time than their control group counterparts (Figures 1-3). In the cases of depression (Wilks λ=0.92, $F_{1,53}$=4.65, partial η^2=0.081; P=.04) and stress (Wilks λ=0.91, $F_{1,48}$=4.95, partial η^2=0.093; P=.03), ANCOVAs revealed statistically significant symptom improvement following coaching with moderate effect sizes. In the case of anxiety, statistically significant improvement following coaching was associated with a large effect size (Wilks λ=0.85, $F_{1,52}$=9.55, partial η^2=0.155; P=.003).

Figure 1. Temporal change in depression symptoms, as measured by the DASS-21, among the US adult users of a digital mental health platform (Wave) who received remote coaching versus those who used the app only. Data were collected between April 2023 and May 2024. ANCOVA revealed a significant group-by-time interaction (P=.04, medium effect size): coaching users with elevated baseline depression symptoms showed significantly greater reductions in symptoms compared to controls who did not engage in coaching. ANCOVA: analysis of covariance; DASS-21: Depression Anxiety Stress Scale-21.

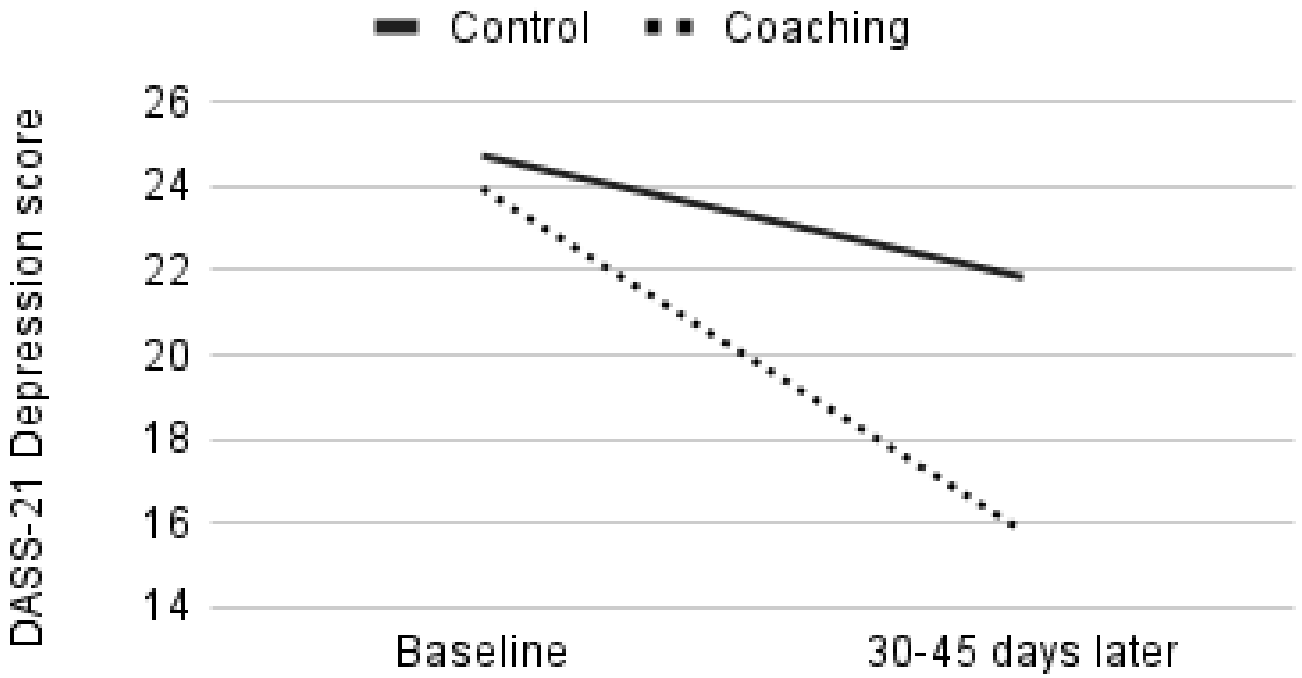

XSL•FO
RenderX

Figure 2. Temporal change in anxiety symptoms, as measured by the DASS-21, among the US adult users of a digital mental health platform (Wave) who received remote coaching versus those who used the app only. Data were collected between April 2023 and May 2024. ANCOVA revealed a significant group-by-time interaction (*P*=.003, large effect size): coaching users with elevated baseline anxiety symptoms showed significantly greater reductions in symptoms compared to controls who did not engage in coaching. ANCOVA: analysis of covariance; DASS-21: Depression Anxiety Stress Scale-21.

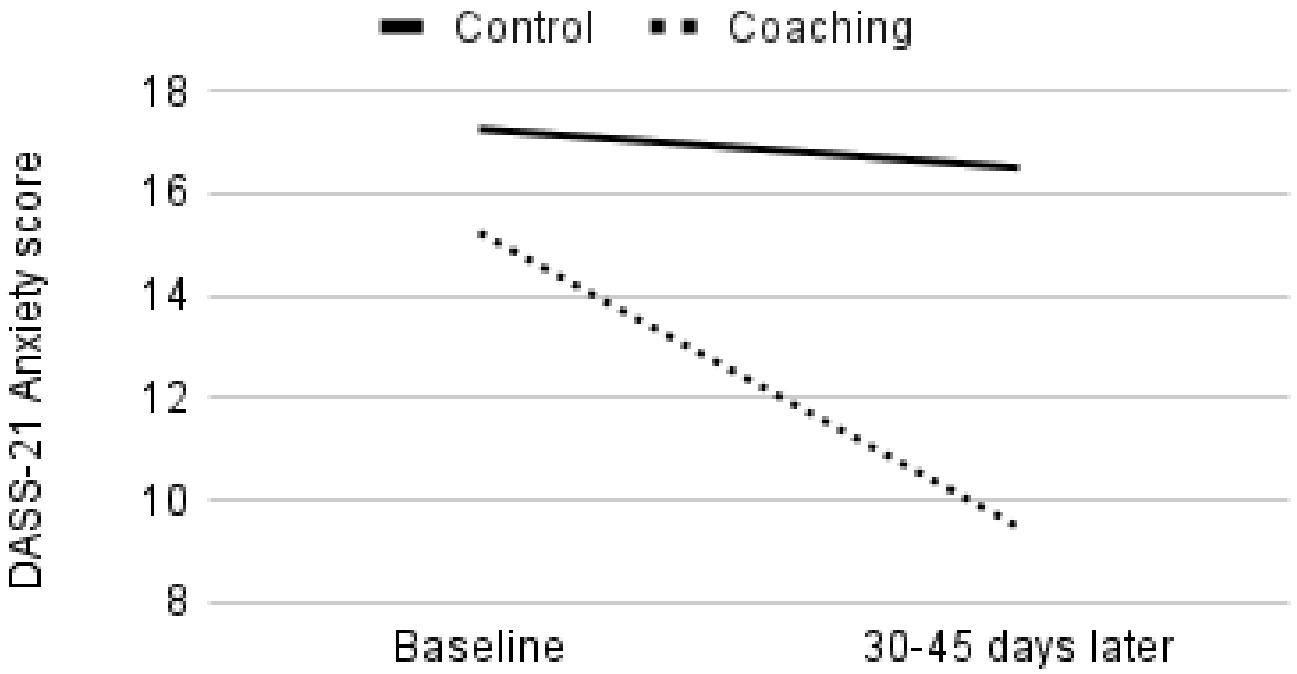

Figure 3. Temporal change in stress symptoms, as measured by the DASS-21, among the US adult users of a digital mental health platform (Wave) who received remote coaching versus those who used the app only. Data were collected between April 2023 and May 2024. ANCOVA revealed a significant group-by-time interaction (*P*=.03, medium effect size): coaching users with elevated baseline stress symptoms showed significantly greater reductions in symptoms compared to controls who did not engage in coaching. ANCOVA: analysis of covariance; DASS-21: Depression Anxiety Stress Scale-21.

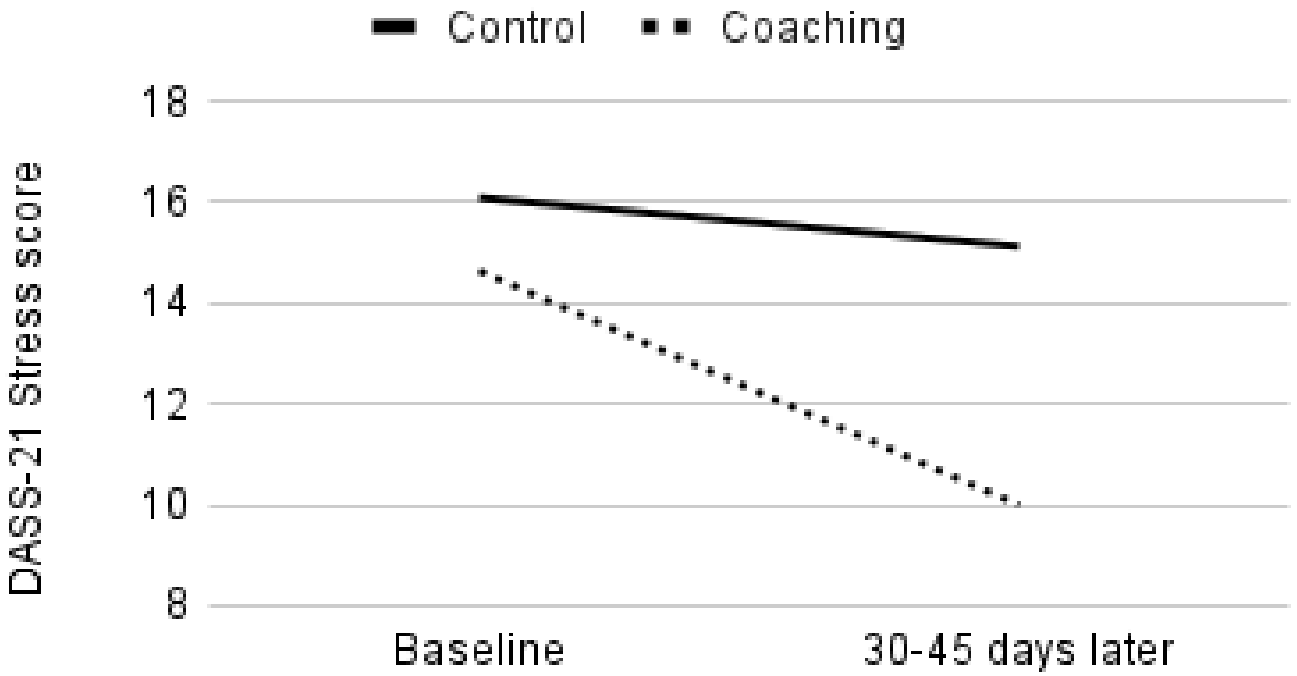

Discussion

Principal Findings

In this study, coaching was associated with significant symptom improvement, particularly for individuals with clinically elevated symptoms of depression, anxiety, and stress. These results replicate and extend those of other studies, wherein, over the course of coaching, well-being and the ability to manage difficult emotions improved [16-18]. They lend credence to a sparse evidence base on coaching's impact on anxiety [19] and coalesce with a larger literature examining "guided," internet-based interventions for depression [25]. They suggest that coaching is a potent factor in symptom reduction, producing

XSL•FO
RenderX

gains that exceed those gleaned from a fully self-guided experience. By incorporating a control group, rather than solely examining within-group changes, this study provides stronger evidence than has been previously presented regarding the causal role of coaching in temporal mental health improvements.

Notably, this study included a sample diverse in its demographic and clinical composition. More than half of the participants who sought coaching reported severe or extremely severe depression, anxiety, or stress. This marks a departure from most other studies, in which demographic data were not disclosed, and individuals with severe depression or anxiety were excluded [16,17,19,26]. Such exclusion criteria reflect a tacit, but untested, notion that coaching is not appropriate for high-acuity individuals. This notion may stem from coaching's early roots as a workplace intervention designed to reduce stress in subclinical populations [27]. However, the results of this study challenge conventional notions regarding coaching's narrow applicability. Instead, they suggest that, with training, lay health workers like coaches can effectively deliver evidence-based interventions that produce swift improvements in mental health across the clinical severity continuum. Given the potential impact of coaching on extending the reach of traditional health care systems, especially to vulnerable communities, it is critical that more studies assess coaching's effectiveness among heterogeneous samples.

Despite group differences in symptom trajectories, coaching users and controls did not significantly differ in their engagement with content. This result is somewhat surprising given the posited role of coaches in increasing engagement with mental health apps [14]. However, while coaching may not have affected the quantity of engagement, it is possible that it impacted the quality, with downstream effects on symptomatology. For instance, among coaching users, content may have reinforced or extended what was discussed in coaching sessions, whereas among controls, content may have been encountered in a more exploratory or passive fashion that was less impactful. How best to define engagement with digital mental health interventions remains an open issue [20], and more research is needed to understand the interplay between coaching and the use of digital resources.

Limitations

This study was limited in notable ways. The sample was small, and demographic data and content engagement data were not available for the full sample. Symptom severity was assessed by self-report measures roughly a month after initiating use of Wave; long-term data were not sufficiently available to assess maintenance of gains over time. Users were not randomly assigned to groups, and it is plausible that those who engaged in coaching were more intrinsically motivated, which could have affected the measured outcomes. Although we found that, despite nonrandom assignment, coaching users and controls were similar in terms of demographic characteristics assessed (when such data were available), and in initial symptom severity, other dimensions might differentiate these groups that could account for their different symptom trajectories. For instance, use of other mental health services or apps might have varied between the groups and could have contributed to the observed outcomes. As these variables were not assessed, they could not be accounted for in our analyses. Accordingly, larger, randomized clinical trials are essential to further establish the effectiveness of digital coaching programs.

Conclusions

Despite these limitations, this study reflects an important step toward evaluating the effectiveness of novel approaches to mental health care delivery. It offers promising findings for disseminating evidence-based practice and reducing strain on an overburdened health care system. Research studies should continue to evaluate coaching's impact on outcomes, including engagement and retention with digital mental health interventions, and long-term symptom reductions. These studies should take care to include diverse samples to provide a fuller understanding of coaching's value in reducing care disparities.

Acknowledgments

Generative artificial intelligence was not used to draft any portion of the manuscript. All writing and editing were performed by the authors. This study was conducted without external funding support. All research activities were carried out by the authors as part of their internal responsibilities within the organization.

Data Availability

The data analyzed in this study are not publicly available due to privacy protections and restrictions outlined in the platform's Privacy Policy. Researchers interested in potential data access may contact the author SA to inquire about possibilities for collaborative analyses.

Conflicts of Interest

None declared.

References

1. Mental health by the numbers. National Alliance on Mental Illness (NAMI). URL: https://www.nami.org/about-mental-illness/mental-health-by-the-numbers/ [accessed 2025-10-01]

XSL•FO
RenderX

2. Vigo D, Thornicroft G, Atun R. Estimating the true global burden of mental illness. Lancet Psychiatry 2016 Feb;3(2):171-178. [doi: 10.1016/S2215-0366(15)00505-2] [Medline: 26851330]
3. Chisholm D, Sweeny K, Sheehan P, et al. Scaling-up treatment of depression and anxiety: a global return on investment analysis. Lancet Psychiatry 2016 May;3(5):415-424. [doi: 10.1016/S2215-0366(16)30024-4] [Medline: 27083119]
4. Behavioral health workforce brief, 2023. : HRSA Health Workforce; 2023 URL: https://bhw.hrsa.gov/sites/default/files/bureau-health-workforce/Behavioral-Health-Workforce-Brief-2023.pdf [accessed 2025-10-01]
5. Mental illness. National Institute of Mental Health (NIMH). 2023. URL: https://www.nimh.nih.gov/health/statistics/mental-illness [accessed 2025-10-01]
6. Exploring the rise in mental health care use by demographics and insurance status. KFF. 2024. URL: https://www.kff.org/mental-health/issue-brief/exploring-the-rise-in-mental-health-care-use-by-demographics-and-insurance-status/ [accessed 2025-10-01]
7. Terlizzi EP, Schiller JS. Mental health treatment among adults aged 18-44: United States, 2019-2021. NCHS Data Brief 2022 Sep(444):1-8. [Medline: 36135999]
8. Behavioral health: available workforce information and federal actions to help recruit and retain providers. : GAO U.S. Government Accountability Office; 2022 URL: https://www.gao.gov/products/gao-23-105250 [accessed 2025-10-01]
9. Mental health care health professional shortage areas (HPSAs). KKF. 2024. URL: https://www.kff.org/other/state-indicator/mental-health-care-health-professional-shortage-areas-hpsas [accessed 2025-10-01]
10. Health workforce projections. HRSA Health Workforce. 2024. URL: https://bhw.hrsa.gov/data-research/projecting-health-workforce-supply-demand [accessed 2025-10-01]
11. Richards D, Richardson T. Computer-based psychological treatments for depression: a systematic review and meta-analysis. Clin Psychol Rev 2012 Jun;32(4):329-342. [doi: 10.1016/j.cpr.2012.02.004] [Medline: 22466510]
12. Barnett ML, Lau AS, Miranda J. Lay health worker involvement in evidence-based treatment delivery: a conceptual model to address disparities in care. Annu Rev Clin Psychol 2018 May 7;14:185-208. [doi: 10.1146/annurev-clinpsy-050817-084825] [Medline: 29401043]
13. Weaver A, Lapidos A. Mental health interventions with community health workers in the United States: a systematic review. J Health Care Poor Underserved 2018;29(1):159-180. [doi: 10.1353/hpu.2018.0011] [Medline: 29503292]
14. Meyer A, Wisniewski H, Torous J. Coaching to support mental health apps: exploratory narrative review. JMIR Hum Factors 2022 Mar 8;9(1):e28301. [doi: 10.2196/28301] [Medline: 35258468]
15. Mohr DC, Cuijpers P, Lehman K. Supportive accountability: a model for providing human support to enhance adherence to eHealth interventions. J Med Internet Res 2011 Mar 10;13(1):e30. [doi: 10.2196/jmir.1602] [Medline: 21393123]
16. Peiper NC, Pettitt A, Shah B, Atwood O, Pfeffer J. Feasibility of a digital coaching program for improving mental health and well-being. Open Science Framework. Preprint posted online on May 23, 2023. [doi: 10.31234/osf.io/8adpu]
17. Lungu A, Boone MS, Chen SY, Chen CE, Walser RD. Effectiveness of a cognitive behavioral coaching program delivered via video in real world settings. Telemed J E Health 2021 Jan;27(1):47-54. [doi: 10.1089/tmj.2019.0313] [Medline: 32311301]
18. Jeannotte AM, Hutchinson DM, Kellerman GR. The time to change for mental health and wellbeing via virtual professional coaching: longitudinal observational study. J Med Internet Res 2021 May 13;23(7):e27774. [doi: 10.2196/27774] [Medline: 33993102]
19. Wu MS, Chen SY, Wickham RE, O'Neil-Hart S, Chen C, Lungu A. Outcomes of a blended care coaching program for clients presenting with moderate levels of anxiety and depression: pragmatic retrospective study. JMIR Ment Health 2021 Oct 21;8(10):e32100. [doi: 10.2196/32100] [Medline: 34673534]
20. Boucher EM, Raiker JS. Engagement and retention in digital mental health interventions: a narrative review. BMC Digit Health 2024;2(1):52. [doi: 10.1186/s44247-024-00105-9]
21. Lovibond PF, Lovibond SH. The structure of negative emotional states: comparison of the Depression Anxiety Stress Scales (DASS) with the beck depression and anxiety inventories. Behav Res Ther 1995 Mar;33(3):335-343. [doi: 10.1016/0005-7967(94)00075-u] [Medline: 7726811]
22. Antony MM, Bieling PJ, Cox BJ, Enns MW, Swinson RP. Psychometric properties of the 42-item and 21-item versions of the Depression Anxiety Stress Scales in clinical groups and a community sample. Psychol Assess 1998;10(2):176-181. [doi: 10.1037//1040-3590.10.2.176]
23. Henry JD, Crawford JR. The short-form version of the Depression Anxiety Stress Scales (DASS-21): construct validity and normative data in a large non-clinical sample. Br J Clin Psychol 2005 Jun;44(Pt 2):227-239. [doi: 10.1348/014466505X29657] [Medline: 16004657]
24. Norton PJ. Depression Anxiety and Stress Scales (DASS-21): psychometric analysis across four racial groups. Anxiety Stress Coping 2007 Sep;20(3):253-265. [doi: 10.1080/10615800701309279] [Medline: 17999228]
25. Karyotaki E, Efthimiou O, Miguel C, et al. Internet-based cognitive behavioral therapy for depression: a systematic review and individual patient data network meta-analysis. JAMA Psychiatry 2021 Apr 1;78(4):361-371. [doi: 10.1001/jamapsychiatry.2020.4364] [Medline: 33471111]
26. McCallum M, Ho AS, Mitchell ES, et al. Feasibility, acceptability, and preliminary outcomes of a cognitive behavioral therapy-based mobile mental well-being program (noom mood): single-arm prospective cohort study. JMIR Form Res 2022 Apr 15;6(4):e36794. [doi: 10.2196/36794] [Medline: 35436218]

XSL•FO
RenderX

27. Theeboom T, Beersma B, van Vianen AEM. Does coaching work? A meta-analysis on the effects of coaching on individual level outcomes in an organizational context. J Posit Psychol 2014 Jan 2;9(1):1-18. [doi: 10.1080/17439760.2013.837499]

Abbreviations

ANCOVA: analysis of covariance
DASS-21: Depression Anxiety Stress Scale-21
MANCOVA: multivariate analysis of covariance

Edited by A Mavragani; submitted 15.01.25; peer-reviewed by B Aylward; revised version received 18.08.25; accepted 19.08.25; published 27.10.25.

Please cite as:
Pickover A, Adler S
Digital Mental Health Coaching in Clinically Diverse Populations: Controlled Engagement and Outcomes Study
JMIR Form Res 2025;9:e71346
URL: https://formative.jmir.org/2025/1/e71346
doi:10.2196/71346

© Alison Pickover, Sarah Adler. Originally published in JMIR Formative Research (https://formative.jmir.org), 27.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Information Deficits, Information Needs, and Preferences Regarding eHealth in a Dutch Population With Metabolic Dysfunction–Associated Steatotic Liver Disease: Cross-Sectional Survey Study

Sharon Oude Veldhuis[1,2], MSc; Maureen M J Guichelaar[3], MD, PhD; Marjolein E M den Ouden[1,4], PhD; Lisette J E W C van Gemert-Pijnen[2], Prof Dr; Constance H C Drossaert[2], PhD

[1]Research Group Technology, Health & Care, Saxion University of Applied Sciences, M.H. Tromplaan 28, Enschede, The Netherlands

[2]Centre for eHealth and Wellbeing Research, Section of Psychology, Health and Technology, University of Twente, Enschede, The Netherlands

[3]Department of Gastroenterology and Hepatology, Medisch Spectrum Twente Hospital, Enschede, The Netherlands

[4]Research Group Care and Technology, Regional Community College of Twente, Hengelo, The Netherlands

Corresponding Author:
Sharon Oude Veldhuis, MSc
Research Group Technology, Health & Care, Saxion University of Applied Sciences, M.H. Tromplaan 28, Enschede, The Netherlands

Abstract

Background: Globally, metabolic dysfunction–associated steatotic liver disease (MASLD) is a common lifestyle-related disease. Lifestyle interventions focusing on healthy eating habits, physical exercise, and reducing body weight in case of obesity are the primary recommended therapies to reverse or improve MASLD. However, patients often experience difficulties in complying with the required lifestyle changes for several reasons, including a lack of knowledge. Health care professionals express limited time during consultations as one of the barriers to discussing lifestyle behavior change. A potential solution to eliminate these barriers and improve the provision of information to patients with MASLD is the use of eHealth.

Objective: This study aimed to explore the information needs and deficits of patients with MASLD regarding a variety of disease-related topics and their preferences regarding a future eHealth intervention.

Methods: In a cross-sectional survey study, patients with MASLD were recruited via 2 Dutch patient organizations. The questionnaire included questions on sociodemographics, information provision and needs, and preferences regarding an eHealth intervention. Data were reported using descriptive statistics. Pearson chi-square tests and logistic regression analysis were used to identify differences in outcomes between subgroups.

Results: The questionnaire was filled out by 449 respondents (women: 363/449, 81%; age: mean 56, SD 11 y). Fewer than 20% of them indicated that they had received sufficient information on a broad range of disease-related topics. Approximately 72% (325/449) to 90% (405/449) of respondents indicated that they would like to receive additional information. Respondents who did not know their disease stage reported a significantly higher need for information on general topics, compared to respondents who reported their disease stage (P values ranging from <.01 to .03). Respondents with (self-reported) metabolic dysfunction–associated steatohepatitis were more interested in contact with fellow patients than respondents with an early or unknown stage of disease (P=.002). Regarding a future eHealth intervention, respondents were most interested in receiving MASLD-related information, practical examples, and references to relevant websites or apps. Respondents were least interested in contact, collaboration, or competition with other app users.

Conclusions: The vast majority of respondents reported a high rate of information deficits on a broad range of MASLD-related topics and expressed a strong need for additional information. Insights into information needs and preferences regarding eHealth can be used to develop an eHealth intervention for patients with MASLD.

(JMIR Form Res 2025;9:e78536) doi:10.2196/78536

KEYWORDS

metabolic dysfunction–associated steatotic liver disease; MASLD; information provision; information needs; patient education; needs assessment; eHealth

XSL•FO
RenderX

Introduction

Background

Globally, metabolic dysfunction–associated steatotic liver disease (MASLD) is a common lifestyle-related disease. MASLD encompasses a spectrum of diseases, ranging from simple steatosis (accumulation of fat in the liver cells) to metabolic dysfunction–associated steatohepatitis (MASH; liver inflammation) and liver cirrhosis (scarring of the liver as a result of long-standing inflammation). This latter stage of disease increases the risk of hepatocellular carcinoma (HCC) and liver failure and is associated with increased rates of liver transplantation and liver-related mortality [1,2]. Furthermore, MASLD is associated with increased rates of cardiovascular disease [3,4]. The current global prevalence of MASLD is estimated to be 32% [5], and it is expected to rise because of the ongoing obesity epidemic and the rise of people with type 2 diabetes mellitus [6]. Moreover, MASLD is expected to emerge as the main cause of end-stage liver disease in the coming decades [5,7]. Lifestyle interventions focusing on healthy eating habits (Mediterranean or low-carb diet), physical exercise, and reducing body weight in case of obesity are the primary recommended therapies to reverse or improve MASLD. However, previous studies showed that changing lifestyles and complying with the required lifestyle changes are difficult for several reasons, including difficulties with incorporating lifestyle changes, motivation, and a lack of social support [8,9].

From the patients' perspective, previous studies showed that many patients with MASLD experienced an information deficit [10-12]. However, an overview of topics on which patients would like to receive more information is lacking. Moreover, it is unclear if the information needs are associated with patient characteristics, such as MASLD stage, time since diagnosis, education level, age, and gender. This poor provision of MASLD-related information can result in reduced knowledge about disease- and treatment-related topics [8,10,12] and, therefore, less concern and awareness of the serious adverse outcomes related to MASLD [8] and difficulties with disease acceptance due to misconceptions [13]. Health care professionals play an important role in the provision of information. A previous qualitative study showed that limited time during consultations is one of the barriers to discussing lifestyle behavior change [12]. In addition, a lack of expertise regarding MASLD among health care professionals, particularly in primary care [10,12], could hinder adequate information provision. Providing patients with correct information can increase their self-efficacy [13]. A higher self-efficacy and better illness perception were key determinants of successful lifestyle changes in patients with MASLD [14]. One of the potential solutions to eliminate these barriers and improve information provision to patients with MASLD is the use of eHealth. eHealth offers the opportunity to disseminate information to a broad public, with minimal effort from health care professionals. Previous studies have shown that eHealth can lead to a better quality of care, quality of life, and clinical outcomes, as it increases self-care and contributes to better communication between patients and health care providers [15-17]. A recent systematic review and meta-analysis of MASLD-related studies reported a significant effect of eHealth interventions on the decrease of BMI and liver enzymes [18]. These eHealth interventions were delivered via different channels (eg, telephone, mobile app, and web-based platform) and consisted of a variety of components, such as information, goal setting, counseling, reminders, and feedback.

In general, tailoring eHealth interventions to the needs and wishes of the target population is recommended. Possible ways to improve effectiveness and adherence to eHealth include integrating persuasive system design (PSD) principles [19], such as tailoring and competition, and behavior change techniques (BCTs) [20], such as goal setting and just-in-time coaching [21,22]. These techniques and design features may also be supportive for patients with MASLD. However, insights into the preferences of patients with MASLD or specific subgroups regarding persuasive eHealth features and BCTs are lacking.

Objectives

This study aimed to explore information needs and deficits in a large patient population with MASLD regarding a variety of disease-related topics. Furthermore, associations between information needs and patient characteristics were examined, as well as eHealth preferences. These insights are essential for improving MASLD-related information provision and empowering patients to adopt positive lifestyle changes.

Methods

Study Design

A cross-sectional online survey was shared via social media of the Dutch Digestive Foundation and the Dutch Liver Patients Association. The Dutch Digestive Foundation and Dutch Liver Patients Association are nationwide patient organizations for patients with gastrointestinal and liver-related diseases as well as their relatives. The organizations strive to optimize the quality of care and quality of life for these patient groups.

Respondents and Procedures

Patients were invited to participate in the online survey if they were diagnosed with MASLD and were aged ≥18 years. No exclusion criteria were applicable. The online questionnaire was open for response from June 1, 2021, to September 12, 2021.

The STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) checklist for cross-sectional studies was used for writing this paper (Checklist 1).

Ethical Considerations

The study protocol was approved by the institutional review board of Medisch Spectrum Twente (K21-16). All respondents were informed about the scope and aim of the study via an online patient information letter. They provided online informed consent before enrollment in the study. Respondents did not receive any compensation for taking part in the study. No directly identifiable patient data were collected, and all data were securely stored on a server at Saxion University of Applied Sciences.

XSL•FO
RenderX

Questionnaire

Sociodemographic and Health-Related Variables

Sociodemographic characteristics that were assessed included age, gender (woman, man, and nonbinary), educational level (primary, lower secondary, upper secondary, bachelor's, master's, and other), type of mobile devices used (smartphone, computer, tablet, and wearables), and daily use of the internet (yes or no).

Regarding health-related variables, we asked when and by whom MASLD was diagnosed (gastroenterologist, general practitioner, nurse practitioner, and other) and the (self-reported) stage of the disease (simple steatosis, MASH without fibrosis or cirrhosis, MASH with fibrosis or cirrhosis, MASLD and HCC, unknown: has not been told, unknown: it has not been examined, and cannot remember but it has been told).

Education levels were merged into 3 subgroups—basic (primary and lower secondary), intermediate (upper secondary), and advanced (bachelor's and master's)—based on the International Standard Classification of Education [23], to reduce the number of subgroups.

Information Provision and Needs

Before the survey, 6 interviews were performed with patients with MASLD about their illness perceptions and barriers and motivators for lifestyle change. Relevant topics from these interviews were incorporated in the survey. The survey included questions about information provision and needs. First, the respondents were asked to give their opinion about the quantity of received information on a variety of MASLD-related topics (what is MASLD, causes, consequences, treatment, self-management, how to improve the eating and exercise pattern, contact with other patients with MASLD, support by health care professionals, and recommended websites and apps). Response options were no information, little information, and sufficient information. Second, respondents were asked if they would have liked to receive more information on these topics (yes or no). Third, they were asked by whom the MASLD-related information was provided (gastroenterologist or hepatologist, general practitioner, nurse practitioner, dietician, physiotherapist, and other) and what sources they used when looking for more information. Finally, we identified patients' experiences with information provision and their suggestions for improvement of information provision using an open-ended question.

Preferences Regarding an eHealth Intervention for Patients With MASLD

Preferences of patients with MASLD regarding a future eHealth intervention were explored using a selection of 15 potential features (as shown in the *Preferences Regarding an eHealth Intervention for Patients With MASLD* section in *Results*) derived from the PSD principles [19] and BCTs (derived from the Theoretical Domains Framework) [20]. The selection of features is based on the target group and the aim of a future eHealth intervention (improving disease-related knowledge and supporting lifestyle change). The following features were incorporated: self-monitoring, rewards, reminders, competition, social learning, cooperation, social support, praise, goal setting, information about the disease, feedback on behavior, reduction, tunneling, suggestions, and instructions. The formulation of these principles and techniques was simplified in the questionnaire to ensure understanding by respondents. For each item, respondents were asked to what extent it appealed to them (appeals a lot, appeals a little, and does not appeal). Furthermore, suggestions for the development of a future eHealth intervention were asked using an open-ended question.

Statistical Analysis

For the questions on the quantity of information received and eHealth preferences, answering options were coded as 0=no information or does not appeal, 1=little information or appeals a little, and 2=sufficient information or appeals a lot; and a mean score was calculated per item. The missing data were avoided by configuring all questionnaire items as forced responses.

Descriptive statistics were reported as means with SDs or medians with IQRs for continuous variables, and numbers with percentages for categorical data. Pearson chi-square tests were used to examine whether there were differences in information needs and preferences regarding eHealth features between subgroups based on gender, education, and MASLD stage. Logistic regression analysis was used to examine the association between information needs and the continuous variables age and time since diagnosis. Significance levels were set at α=.05 for all tests.

Quantitative data were analyzed using SPSS Statistics (version 26.0; IBM Corp). Data from open-ended questions were coded and categorized in Atlas.ti (version 23.0; Scientific Software Development GmbH) by 2 researchers (SOV and MEMDO).

Results

Characteristics of Study Respondents

In total, 1242 respondents started the questionnaire. A total of 469 respondents completed at least 34% of the questionnaire, including the questions on sociodemographics, health-related topics, and information provision and needs. The questions about eHealth features were filled out by 366 respondents; others dropped out due to the questionnaire's length. In total, 20 respondents were excluded based on their questionnaire comments (eg, aged <18 y or not having an official diagnosis). Hence, results on information provision and eHealth features were described for 449 and 348 respondents, respectively. The sociodemographic and health-related characteristics of these groups did not differ substantially.

The majority of respondents were women (363/449, 81%), with a mean age of 56 (SD 11; range 25-83) years, as reported in Table 1. Most respondents (193/449, 43%) had completed education at an intermediate level. Almost all respondents (326/348, 94%) used (mobile) devices and the internet on a daily basis. Simple steatosis (206/449, 46%) was the most common self-reported stage of MASLD among respondents. Interestingly, more than one-third of the respondents (161/449, 36%) stated that they did not know their stage of disease. The time since diagnosis of MASLD ranged from 2 to 46 years, with a median of 5 (IQR 5) years. More than half of the respondents

XSL•FO
RenderX

(271/449, 60%) mentioned that the gastroenterologist diagnosed MASLD.

Table . Sociodemographic and health-related characteristics (N=449).

Characteristics	Value
Age (years), mean (SD)	56 (11)
Gender, n (%)	
Woman	363 (81)
Man	85 (19)
Nonbinary	1 (0)
Education	
Basic	124 (28)
Intermediate	193 (43)
Advanced	129 (29)
Other (not specified)	3 (1)
Use of devices[a], n (%)	
Smartphone	329 (95)
Computer	287 (83)
Tablet	216 (62)
Wearables (eg, smartwatch)	83 (24)
Daily use of internet[a], n (%)	326 (94)
Self-reported stage of MASLD[b], n (%)	
Simple steatosis	206 (46)
MASH[c] without fibrosis or cirrhosis	29 (7)
MASH with fibrosis or cirrhosis	46 (10)
MASLD and hepatocellular carcinoma	2 (0.4)
Unknown: has not been told	88 (20)
Unknown: has not been examined	62 (14)
Cannot remember, but it has been told	11 (2)
Other	5 (1)
Time since diagnosis of MASLD (years), median (IQR)	5 (5)
Diagnostician, n (%)	
Gastroenterologist	271 (60)
General practitioner	96 (21)
Nurse practitioner	18 (4)
Other	64 (14)

[a]Variables “use of devices” and “daily use of internet”; n=348.

[b]MASLD: metabolic dysfunction–associated steatotic liver disease.

[c]MASH: metabolic dysfunction–associated steatohepatitis.

Information Provision and Needs

Most respondents indicated that they had received no or only little information on a variety of MASLD-related topics (Table 2). The topics with the highest percentages of sufficient information were “causes of MASLD” and “consequences of MASLD,” but these percentages were still only 18% (81/449 and 82/449, respectively). The lowest percentages of sufficient information were reported for “recommended websites or apps“ and “contact with other patients with MASLD” (23/449, 5% and 12/449, 3%, respectively). For almost all disease-related and health care–related topics, 72% (325/449) to 90% (405/449) of respondents indicated a need to receive additional information. Only 37% (165/449) of respondents expressed a need for more information on the topic “opportunities to get in contact with other patients with MASLD.”

XSL•FO
RenderX

Table . Amount of information received and information needs regarding different metabolic dysfunction–associated steatotic liver disease (MASLD)–related topics (N=449).

Topics	No information (score=0), n (%)	Little information (score=1), n (%)	Sufficient information (score=2), n (%)	Score (0 - 2), mean	Need for more information, n (%)
Treatment of MASLD	225 (50.1)	171 (38.1)	53 (11.8)	0.6	405 (90.2)
Self-management	211 (47.0)	173 (38.5)	65 (14.5)	0.7	403 (89.8)
Causes of MASLD	178 (39.6)	190 (42.3)	81 (18.0)	0.8	394 (87.8)
Consequences of MASLD	213 (47.4)	154 (34.3)	82 (18.3)	0.7	392 (87.3)
What is MASLD	168 (37.4)	202 (45.0)	79 (17.6)	0.8	382 (85.1)
How to improve diet	217 (48.3)	158 (35.2)	74 (16.5)	0.7	377 (84.0)
Recommended websites or apps	343 (76.4)	83 (18.5)	23 (5.1)	0.3	336 (74.8)
How to improve exercise pattern	224 (49.9)	154 (34.3)	71 (15.8)	0.7	333 (74.2)
Support by health care professionals	303 (67.5)	109 (24.3)	37 (8.2)	0.4	325 (72.4)
Contact with other patients with MASLD	376 (83.7)	61 (13.6)	12 (2.7)	0.2	165 (36.7)

Providers of Information Regarding MASLD

Almost 47% (210/449) of the respondents mentioned that they had received information about their disease from the gastroenterologist or hepatologist. Others stated that they had been informed by the general practitioner (123/449, 27%), nurse practitioner (45/449, 10%), dietician (44/449, 10%), physiotherapist (5/449, 1%), or other (42/449, 9%). Approximately 20% (95/449) of the respondents indicated that they had not received information about MASLD from a health care provider.

Searching for Additional Information About MASLD

In addition to the information provided by health care professionals, 78% (349/449) of the respondents mentioned that they had searched for additional information about MASLD via different channels. The internet was the most used source to search for information (302/449, 67%). In total, 35% (156/449) of the respondents searched on disease-specific websites (eg, the website of the patient organization). Some other channels were used by only a few respondents: information from acquaintances (13/449, 3%), fellow patients (7/449, 2%), dieticians (5/449, 1%), books (3/449, 1%), other physicians or medical specialists (3/449, 1%), patient organizations (2/449, <1%), and scientific papers (1/449, <1%).

Associations Between Information Needs and Patient Characteristics

Regarding self-reported stage of disease, significant differences were found between the 3 subgroups on 6 of the 10 topics on information needs (Table 3). Patients with MASH (with or without fibrosis, cirrhosis, or HCC) expressed a significantly higher need for more information on the topic “Contact with other patients with MASLD” and a significantly lower need to receive information on the topic “How to improve diet,” compared to respondents with an early stage (simple steatosis) or unknown stage of disease. Respondents who did not know their stage of disease reported a significantly higher need for more information on the topic “What is MASLD,” compared to respondents who did report their stage of disease (Table 3). They also reported a significantly higher need for more information on the topics “self-management,” “causes of MASLD,” and “consequences of MASLD,” compared to respondents with MASH (with or without fibrosis, cirrhosis, or HCC). When looking at gender, women expressed, overall, a somewhat higher need for information compared to men (Table 3). Their information needs were significantly higher on the topics “What is MASLD,” “Treatment of MASLD,” “Consequences of MASLD,” and “How to improve diet.” Information needs of respondents did not differ significantly by education level, age, or time since diagnosis.

XSL•FO
RenderX

Table . Differences in information needs by (self-reported) metabolic dysfunction–associated steatotic liver disease (MASLD) stage and gender.

Need for more information (yes)	MASLD stage (n=444[a])				Gender (n=448[b])		
	Simple steatosis (n=206), n (%)	MASH[c] with or without fibrosis, cirrhosis, or HCC[d] (n=77), n (%)	Stage unknown by respondent (n=161), n (%)	*P* value	Woman (n=363), n (%)	Man (n=85), n (%)	*P* value
Treatment of MASLD	187 (90.8)	66 (85.7)	147 (91.3)	.36	333 (91.7)	72 (84.7)	.048
Self-management	186 (90.3)	63 (81.8)	150 (93.2)[e]	.02	331 (91.2)	72 (84.7)	.07
Causes of MASLD	182 (88.3)	61 (79.2)	147 (91.3)[e]	.03	323 (89.0)	70 (82.4)	.09
Consequences of MASLD	180 (87.4)	61 (79.2)	147 (91.3)[e]	.03	324 (89.3)	67 (78.8)	.01
What is MASLD	171 (83.0)	58 (75.3)	148 (91.9)[f]	.002	320 (88.2)	61 (71.8)	<.001
How to improve diet	176 (85.4)	56 (72.7)[g]	141 (87.6)	.01	314 (86.5)	63 (74.1)	.01
Recommended websites or apps	153 (74.3)	61 (79.2)	119 (73.9)	.64	275 (75.8)	60 (70.6)	.32
Support by health care professional	152 (73.8)	52 (67.5)	118 (73.3)	.56	262 (72.2)	62 (72.9)	.89
How to improve exercise pattern	149 (72.3)	54 (70.1)	127 (78.9)	.24	276 (76.0)	57 (67.1)	.09
Contact with other patients with MASLD	68 (33.0)	42 (54.5)[g]	54 (33.5)	.002	132 (36.4)	33 (38.8)	.67

[a]Subgroup with MASLD stage "other" (n=5) was excluded from this analysis.

[b]Nonbinary (n=1) was excluded from this analysis.

[c]MASH: metabolic dysfunction–associated steatohepatitis.

[d]HCC: hepatocellular carcinoma.

[e]Significantly different from subgroup "MASH (with or without fibrosis, cirrhosis, or HCC)"; $P<.05$).

[f]Significantly different from subgroups "simple steatosis" and "MASH (with or without fibrosis, cirrhosis, or HCC)"; $P<.01$).

[g]Significantly different from subgroups "simple steatosis" and "stage unknown by respondent" ($P\leq.01$).

Suggestions for Improving Information Provision

Respondents were asked, via an open-ended question, what could be improved in the provision of MASLD-related information. This open question was answered by 221 respondents. Several respondents (22/221, 10%) mentioned that information provision in total should be improved. Furthermore, many respondents (56/221, 25%) answered that they would have liked the information provided to be clearer or more in-depth. Other respondents made this more specific by mentioning that the information could be improved regarding the causes (18/221, 8%), prevention (6/221, 3%), consequences (14/221, 6%), prognosis (4/221, 2%), lifestyle advice and self-management (41/221, 19%), the relation with other diseases and medication (7/221, 3%), and possibilities of support by other health care providers (4/221, 2%). Some respondents (5/221, 2%) explicitly suggested improving the availability of reliable and relevant information about MASLD:

> *A clear leaflet or website with information on what can possibly be done about it [MASLD]. I was told by the doctor's assistant that a slightly fatty liver is not a serious illness and that a large proportion of people have it without knowing it. I received no advice on what I could possibly do about it.*

In addition, respondents (37/221, 17%) mentioned that health care professionals could be more proactive in providing information about MASLD, rather than merely communicating the diagnosis. A few respondents (9/221, 4%) indicated that they felt that health care professionals themselves were not well informed about MASLD in general and how it should be managed. Other respondents (16/221, 7%) mentioned that they

XSL•FO
RenderX

felt the risks of MASLD were trivialized, for example, by saying that nothing needs to be done about it:

> *That gastroenterologists should receive more information to inform patients.*

Others (2/221, 1%) reported that health care professionals made them feel guilty that they were diagnosed with MASLD, which they experienced as unpleasant. Furthermore, some respondents (5/221, 2%) would have appreciated more medical tests and follow-up:

> *Don't make it feel like it's your own fault.*

> *More frequent monitoring when improvement is in progress. This motivates to continue on the path taken.*

A number of respondents (8/221, 4%) indicated that they believed that there should be more nationwide awareness regarding MASLD. Some respondents (23/221, 10%) expressed specific preferences for the format of information provision, such as a brochure, website, app, information sessions, or a special improvement program for MASLD (comparable to those for cardiac rehabilitation). Only a few respondents (7/221, 3%) indicated that information provision did not need to be improved.

Preferences Regarding an eHealth Intervention for Patients With MASLD

Regarding a future eHealth intervention, most respondents (329/348, 95%) mentioned that they were interested in receiving information about MASLD, as well as practical examples and references to relevant websites or apps (both 323/348, 93%) (Table 4). Respondents were least interested in eHealth interventions that include contact with other patients with MASLD (168/348, 48%), collaboration with other app users (164/348, 47%), or competition (93/348, 27%). Subgroup analysis regarding the eHealth features showed that patients with MASH (with or without fibrosis, cirrhosis, or HCC) preferred the options "competition between app users" (P=.004), "learning from experiences of others" (P=.002), and "possibility to contact other patients with MASLD" (P<.001) significantly more often than patients with an early (simple steatosis) or unknown stage of MASLD. No other significant associations were found between the eHealth preferences and the stage of disease. In the open-ended question, 23% (51/221) of the respondents elaborated on this topic with additional suggestions for a future eHealth intervention. These suggestions included push messages, free availability of the app, ease of use, and interoperability with other (eHealth) devices. Furthermore, patients indicated that eHealth interventions can help them recognize their (unhealthy) lifestyle patterns and highlighted the value of regular reminders for making lifestyle changes. However, the intervention should also take into account that patients have multimorbidity. Finally, some respondents preferred personal contact with professionals.

XSL•FO
RenderX

Table . Preferences regarding a future eHealth intervention (n=348).

Items for an eHealth intervention (based on PSD[a] principles and BCTs[b])	Does not appeal (score=0), n (%)	Appeals a little (score=1), n (%)	Appeals a lot (score=2), n (%)	Score (0 - 2), mean
Receiving information about MASLD[c]	19 (5.5)	108 (31.0)	221 (63.5)	1.6
Practical examples (eg, recipes)	25 (7.2)	122 (35.1)	201 (57.8)	1.5
References to relevant websites or apps	25 (7.2)	156 (44.8)	167 (48.0)	1.4
Goal setting	56 (16.1)	149 (42.8)	143 (41.1)	1.3
Self-monitoring	79 (22.7)	123 (35.3)	146 (42.0)	1.2
Feedback on progress	74 (21.3)	143 (41.1)	131 (37.6)	1.2
Testing knowledge (eg, quiz)	76 (21.8)	155 (44.5)	117 (33.6)	1.1
Possibility to contact health care professionals	64 (18.4)	180 (51.7)	104 (29.9)	1.1
Compliments in case of progress	100 (28.7)	116 (33.3)	132 (37.9)	1.1
Reminders	106 (30.5)	128 (36.8)	114 (32.8)	1.0
Learning from experiences of others	98 (28.2)	179 (51.4)	71 (20.4)	0.9
Getting rewards for progress	138 (39.7)	118 (33.9)	92 (26.4)	0.9
Possibility to contact other patients with MASLD	180 (51.7)	112 (32.2)	56 (16.1)	0.6
Collaboration with app users	184 (52.9)	108 (31.0)	56 (16.1)	0.6
Competition between app users	255 (73.3)	63 (18.1)	30 (8.6)	0.4

[a]PSD: persuasive system design.

[b]BCT: behavior change technique.

[c]MASLD: metabolic dysfunction–associated steatotic liver disease.

Discussion

Principal Findings

MASLD is a lifestyle-related disease, and lifestyle improvements are the mainstay in therapy as well as the prevention of disease progression. However, previous studies have shown that limited knowledge about MASLD and the recommended lifestyle changes can hinder effective lifestyle improvements. [8,12]. Our study showed that less than 20% of patients with MASLD indicated that they received sufficient information about a broad range of disease-related topics. In addition, more than one-third of our study respondents (161/449) indicated that they do not know their stage of disease. Respondents who did not know the severity of their disease stage reported a significantly higher need for information on general MASLD-related topics, compared to respondents who reported their stage of disease. Respondents who reported to have MASH (with and without fibrosis, cirrhosis, or HCC) expressed a significantly higher need for information on the topic “Contact with other patients with MASLD.” According to respondents, information provision can be improved by more detailed information (eg, causes, prevention, consequences, and prognosis) and the availability of reliable and relevant information about MASLD. In addition, respondents mentioned that health care professionals could be more proactive in providing information about MASLD, rather than simply communicating the diagnosis. There were also some specific wishes mentioned by respondents regarding the form of information provision, such as a brochure, website, or app. Almost 80% (349/449) of respondents sought additional information about their disease, with the internet being the most frequently used source. Despite their additional search for disease-related information, 72% (325/449) to 90% (405/449) of respondents reported that they would like to receive additional information on almost all disease-related topics, including MASLD in general, self-management, and treatment of MASLD. This may indicate that it is difficult for patients to find the specific disease-related information they requested or that they do not know where to find the right information. Both topics were addressed by several respondents in our survey.

XSL•FO
RenderX

Regarding an eHealth intervention, respondents were most interested in receiving MASLD-related information, practical examples, and references to relevant websites or apps. Respondents were least interested in contact, collaboration, or competition with other app users. However, in line with the results on information needs, patients with MASH (with or without fibrosis, cirrhosis, or HCC) were significantly more interested in features related to contact with other patients with MASLD than patients with an early or an unknown stage of MASLD. This may be explained by the experience of more disease symptoms, concerns, and a higher burden on daily life. Hence, these patients might have a greater need for recognition, emotional support, and sharing experiences with fellow patients, compared to patients with simple steatosis who, generally, experience fewer physical and mental complaints [24]. Our findings are consistent with the study of Valery et al [25], which showed that patients with a good to moderate impaired hepatic function (Child-Pugh A and B) had higher information needs compared to patients with advanced hepatic dysfunction (Child-Pugh C). Moreover, patients with a Child-Pugh C score had higher practical and physical needs, as well as psychosocial issues [25]. These findings suggest that it is important for health care professionals to be aware of potential differences in information needs between patients, as tailored patient education is important for prolonged lifestyle changes [8].

Previous studies, mostly qualitative [10-12,26], have identified several reasons for the lack of information among patients with MASLD. These reasons included time limitations during consultations [12], as well as low motivation of health care professionals to invest their time during consultation in educating patients with MASLD because they experienced that patients often did not follow their advice (ie, not worthwhile to invest time in lifestyle support) [12]. Health care professionals also reported that they lacked adequate knowledge to inform patients [10,27]. The latter was also supported by our study, in which a few respondents mentioned that they felt that their health care professional lacked the disease-specific knowledge to educate them about MASLD and the required lifestyle changes.

Given the difficulties of changing lifestyle among patients with MASLD, a first step toward better lifestyle support is improving information provision in patients. The development and implementation of an eHealth intervention for patients with MASLD is one of the solutions. An eHealth intervention can save time during consultations and complement the motivation and knowledge level of health care professionals. This study gave a first impression of preferences of patients with MASLD regarding a future eHealth intervention, based on a selection of PSD principles and BCTs. To our knowledge, no previous studies were conducted on the preferences of patients with MASLD regarding eHealth. However, a previous qualitative study of people with obesity [21] showed that this comparable study population had similar preferences regarding PSD principles and BCTs, such as the need for reliable information and suggestions about healthy eating. In addition to eHealth interventions, future studies should also focus on the role and education of health care professionals regarding MASLD to enhance information provision for patients.

Although improved knowledge about MASLD and related lifestyle changes is expected to greatly contribute to lifestyle change in patients with MASLD, a future eHealth intervention should not be limited to the provision of information, practical examples, and references to other relevant websites or apps. To achieve and maintain sustainable lifestyle change, behavior change models show that it is also important to strengthen internal motivation and pay attention to patients' physical and social environment [28,29]. For example, these factors may be addressed by incorporating options for setting personal goals, enabling self-monitoring, and making the eHealth app accessible to patients' relatives to enhance social support.

As the study was conducted during the COVID-19 pandemic, it is plausible that patients with MASLD are now more familiar with digital health than they were at the time of data collection. The field of digital care is evolving rapidly, and large language models have been widespread and adopted in recent years [30,31]. However, preferences of patients with MASLD regarding the eHealth features are not expected to have changed substantially, as they were mainly related to obtaining reliable information about MASLD. As not all individuals possess the digital competencies required to retrieve reliable information using large language models, it is likely that the need for a tailored eHealth intervention for patients with MASLD, including reliable information, remains high.

Strengths and Limitations

The strengths of the study include the large sample size of patients with MASLD (with different severity levels) and an overview of the information needs of Dutch patients with MASLD, which was not studied before. In addition, the study provides relevant insights into eHealth preferences, which is relevant for future eHealth interventions to make it suitable for its target population. Limitations of the study include the self-reported status of the MASLD stages. Hence, the results of the subgroup analysis regarding the disease stage should be interpreted with some caution. Moreover, our study showed that information needs were significantly higher among women compared to men. This was also found in a previous study among patients recovering from an acute ischemic coronary event [32]. Given the overrepresentation of women in our study (which was also found in previous online surveys) [33], the information needs in this study might be somewhat higher compared to those in the patient population, in which the prevalence of MASLD is higher among men than women [34]. Although the exact reasons behind these skewed distributions remain unclear [33], one possible explanation for the overrepresentation of women in this study is that the recruitment took place via the social media of 2 patient organizations. Women generally engage in online health information–seeking behavior more frequently than men and are also more likely to use health forums [35,36]. In contrast, the results emphasize the necessity to tailor information provision to the needs of each individual patient. In addition, further research is necessary to identify the preferences of patients with MASLD regarding PSD principles and BCTs, as these entities were briefly addressed. Therefore, these results can be best seen as an indication, instead of a complete overview of preferences for developing a future eHealth intervention.

XSL•FO
RenderX

Conclusions

In conclusion, our study supports that inadequate patient education limits the daily practice of MASLD care. The vast majority of respondents (approximately 80%) perceived no or limited education about their disease and disease-related entities, including self-management. Although almost all respondents undertook additional searches for information, the majority of respondents would like to receive more information about a broad array of MASLD-related topics. Patient education is of utmost importance in the daily practice of MASLD care, as the causes as well as treatment of the disease are lifestyle related. This study contributes to a better understanding of how patients with MASLD experience the current information provision and emphasizes the importance of improving information provision for a broad range of topics. In addition, it supports a necessity to develop eHealth interventions for patients with MASLD and their health care professionals, as eHealth can be used to provide reliable information and reduce the impact of disparities in MASLD-related knowledge between health care professionals. The initial insights from patients with MASLD regarding their eHealth preferences, as provided by this study, may serve as a valuable starting point for developing such an eHealth intervention.

Acknowledgments

The authors thank Henrik J ten Berge for his assistance with data collection. This study was partially funded by Saxion University of Applied Sciences (personal grant to SOV). The funder had no role in the study design, data collection, analysis, interpretation, or manuscript preparation.

Data Availability

The datasets used and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

MMJG, CHCD, and SOV designed the study and contributed to data collection. SOV and MEMDO were involved in data analysis. All authors contributed to interpreting the results, writing, and revising the manuscript.

Conflicts of Interest

None declared.

Checklist 1
STROBE checklist for cross-sectional studies.
[PDF File, 198 KB - formative_v9i1e78536_app1.pdf]

References

1. Burra P, Becchetti C, Germani G. NAFLD and liver transplantation: disease burden, current management and future challenges. JHEP Rep 2020 Oct 9;2(6):100192. [doi: 10.1016/j.jhepr.2020.100192] [Medline: 33163950]
2. Mattos Â, Debes JD, Dhanasekaran R, et al. Hepatocellular carcinoma in nonalcoholic fatty liver disease: a growing challenge. World J Hepatol 2021 Sep 27;13(9):1107-1121. [doi: 10.4254/wjh.v13.i9.1107] [Medline: 34630878]
3. Kasper P, Martin A, Lang S, et al. NAFLD and cardiovascular diseases: a clinical review. Clin Res Cardiol 2021 Jul;110(7):921-937. [doi: 10.1007/s00392-020-01709-7] [Medline: 32696080]
4. Duell PB, Welty FK, Miller M, et al. Nonalcoholic fatty liver disease and cardiovascular risk: a scientific statement from the American Heart Association. Arterioscler Thromb Vasc Biol 2022 Jun;42(6):e168-e185. [doi: 10.1161/ATV.0000000000000153] [Medline: 35418240]
5. Teng ML, Ng CH, Huang DQ, et al. Global incidence and prevalence of nonalcoholic fatty liver disease. Clin Mol Hepatol 2023 Feb;29(Suppl):S32-S42. [doi: 10.3350/cmh.2022.0365] [Medline: 36517002]
6. Younossi ZM, Koenig AB, Abdelatif D, Fazel Y, Henry L, Wymer M. Global epidemiology of nonalcoholic fatty liver disease-meta-analytic assessment of prevalence, incidence, and outcomes. Hepatology 2016 Jul;64(1):73-84. [doi: 10.1002/hep.28431] [Medline: 26707365]
7. Riazi K, Azhari H, Charette JH, et al. The prevalence and incidence of NAFLD worldwide: a systematic review and meta-analysis. Lancet Gastroenterol Hepatol 2022 Sep;7(9):851-861. [doi: 10.1016/S2468-1253(22)00165-0] [Medline: 35798021]
8. Tincopa MA, Wong J, Fetters M, Lok AS. Patient disease knowledge, attitudes and behaviours related to non-alcoholic fatty liver disease: a qualitative study. BMJ Open Gastroenterol 2021 Jun;8(1):e000634. [doi: 10.1136/bmjgast-2021-000634] [Medline: 34193468]
9. Heredia NI, Thrift AP, Balakrishnan M. Perceived barriers to weight loss among Hispanic patients with non-alcoholic fatty liver disease. Hisp Health Care Int 2022 Sep;20(3):171-178. [doi: 10.1177/15404153211043885] [Medline: 34658280]

XSL•FO
RenderX

10. Avery L, Exley C, McPherson S, Trenell MI, Anstee QM, Hallsworth K. Lifestyle behavior change in patients with nonalcoholic fatty liver disease: a qualitative study of clinical practice. Clin Gastroenterol Hepatol 2017 Dec;15(12):1968-1971. [doi: 10.1016/j.cgh.2017.06.011] [Medline: 28624648]
11. Haigh L, Bremner S, Houghton D, et al. Barriers and facilitators to Mediterranean diet adoption by patients with nonalcoholic fatty liver disease in Northern Europe. Clin Gastroenterol Hepatol 2019 Jun;17(7):1364-1371. [doi: 10.1016/j.cgh.2018.10.044] [Medline: 30391437]
12. Hallsworth K, Dombrowski SU, McPherson S, Anstee QM, Avery L. Using the theoretical domains framework to identify barriers and enabling factors to implementation of guidance for the diagnosis and management of nonalcoholic fatty liver disease: a qualitative study. Transl Behav Med 2020 Oct 8;10(4):1016-1030. [doi: 10.1093/tbm/ibz080] [Medline: 31120519]
13. Hallsworth K, Avery L, Trenell MI. Targeting lifestyle behavior change in adults with NAFLD during a 20-min consultation: summary of the dietary and exercise literature. Curr Gastroenterol Rep 2016 Mar;18(3):11. [doi: 10.1007/s11894-016-0485-1] [Medline: 26908279]
14. Zelber-Sagi S, Bord S, Dror-Lavi G, et al. Role of illness perception and self-efficacy in lifestyle modification among non-alcoholic fatty liver disease patients. World J Gastroenterol 2017 Mar 14;23(10):1881-1890. [doi: 10.3748/wjg.v23.i10.1881] [Medline: 28348495]
15. Granja C, Janssen W, Johansen MA. Factors determining the success and failure of eHealth interventions: systematic review of the literature. J Med Internet Res 2018 May 1;20(5):e10235. [doi: 10.2196/10235] [Medline: 29716883]
16. Renzi E, Baccolini V, Migliara G, et al. The impact of eHealth interventions on the improvement of self-care in chronic patients: an overview of systematic reviews. Life (Basel) 2022 Aug 17;12(8):1253. [doi: 10.3390/life12081253] [Medline: 36013432]
17. More care providers recognise added value of eHealth. National Institute for Public Health and the Environment. 2023. URL: https://www.rivm.nl/en/news/more-care-providers-recognise-added-value-of-ehealth [accessed 2025-05-07]
18. Zafar Y, Sohail MU, Saad M, et al. eHealth interventions and patients with metabolic dysfunction-associated steatotic liver disease: a systematic review and meta-analysis. BMJ Open Gastroenterol 2025 Apr 9;12(1):e001670. [doi: 10.1136/bmjgast-2024-001670] [Medline: 40204429]
19. Oinas-Kukkonen H, Harjumaa M. Persuasive systems design: key issues, process model, and system features. Commun Assoc Inf Syst 2009;24(1). [doi: 10.17705/1CAIS.02428]
20. Michie S, Richardson M, Johnston M, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013 Aug;46(1):81-95. [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
21. Asbjørnsen RA, Wentzel J, Smedsrød ML, et al. Identifying persuasive design principles and behavior change techniques supporting end user values and needs in eHealth interventions for long-term weight loss maintenance: qualitative study. J Med Internet Res 2020 Nov 30;22(11):e22598. [doi: 10.2196/22598] [Medline: 33252347]
22. Gemert-Pijnen L, Kelders SM, Beerlage-de Jong N, Oinas-Kukkonen H. Persuasive health technology. In: van Gemert-Pijnen L, Kelders SM, Kip H, Sanderman R, editors. eHealth Research, Theory and Development: A Multi-Disciplinary Approach: Routledge; 2018. [doi: 10.4324/9781315385907-11]
23. International Standard Classification of Education (ISCED). International Labour Organization. URL: https://ilostat.ilo.org/methods/concepts-and-definitions/classification-education/ [accessed 2025-05-07]
24. Younossi Z, Aggarwal P, Shrestha I, et al. The burden of non-alcoholic steatohepatitis: a systematic review of health-related quality of life and patient-reported outcomes. JHEP Rep 2022 Jun 15;4(9):100525. [doi: 10.1016/j.jhepr.2022.100525] [Medline: 36039144]
25. Valery PC, Bernardes CM, Mckillen B, et al. The patient's perspective in cirrhosis: unmet supportive care needs differ by disease severity, etiology, and age. Hepatol Commun 2021 Feb 6;5(5):891-905. [doi: 10.1002/hep4.1681] [Medline: 34027276]
26. Cook NS, Nagar SH, Jain A, et al. Understanding patient preferences and unmet needs in non-alcoholic steatohepatitis (NASH): insights from a qualitative online bulletin board study. Adv Ther 2019 Feb;36(2):478-491. [doi: 10.1007/s12325-018-0856-0] [Medline: 30547371]
27. Driessen S, de Jong VD, van Son KC, et al. A global survey of health care workers' awareness of non-alcoholic fatty liver disease: the AwareNASH survey. United European Gastroenterol J 2023 Sep;11(7):654-662. [doi: 10.1002/ueg2.12445] [Medline: 37563849]
28. Michie S, van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. Implement Sci 2011 Apr 23;6:42. [doi: 10.1186/1748-5908-6-42] [Medline: 21513547]
29. Prochaska JO, DiClemente CC. Stages and processes of self-change of smoking: toward an integrative model of change. J Consult Clin Psychol 1983 Jun;51(3):390-395. [doi: 10.1037//0022-006x.51.3.390] [Medline: 6863699]
30. Cuff A. The evolution of digital health and its continuing challenges. BMC Digit Health 2023;1(1):3. [doi: 10.1186/s44247-022-00004-x] [Medline: 38014367]
31. Maity S, Saikia MJ. Large language models in healthcare and medical applications: a review. Bioengineering (Basel) 2025 Jun 10;12(6):631. [doi: 10.3390/bioengineering12060631] [Medline: 40564447]

XSL•FO
RenderX

32. Stewart DE, Abbey SE, Shnek ZM, Irvine J, Grace SL. Gender differences in health information needs and decisional preferences in patients recovering from an acute ischemic coronary event. Psychosom Med 2004;66(1):42-48. [doi: 10.1097/01.psy.0000107006.83260.12] [Medline: 14747636]
33. Becker R. Gender and survey participation: an event history analysis of the gender effects of survey participation in a probability-based multi-wave panel study with a sequential mixed-mode design. Methods Data Anal 2022;16(1):3-32. [doi: 10.12758/mda.2021.08]
34. Feng G, Targher G, Byrne CD, et al. Global burden of metabolic dysfunction-associated steatotic liver disease, 2010 to 2021. JHEP Reports 2025 Mar;7(3):101271. [doi: 10.1016/j.jhepr.2024.101271]
35. Baumann E, Czerwinski F, Reifegerste D. Gender-specific determinants and patterns of online health information seeking: results from a representative German health survey. J Med Internet Res 2017 Apr 4;19(4):e92. [doi: 10.2196/jmir.6668] [Medline: 28377367]
36. Bidmon S, Terlutter R. Gender differences in searching for health information on the internet and the virtual patient-physician relationship in Germany: exploratory results on how men and women differ and why. J Med Internet Res 2015 Jun 22;17(6):e156. [doi: 10.2196/jmir.4127] [Medline: 26099325]

Abbreviations

BCT: behavior change technique
HCC: hepatocellular carcinoma
MASH: metabolic dysfunction–associated steatohepatitis
MASLD: metabolic dysfunction–associated steatotic liver disease
PSD: persuasive system design
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by A Mavragani; submitted 04.06.25; peer-reviewed by A Santamaria, JVD Ven; revised version received 10.09.25; accepted 12.09.25; published 29.10.25.

Please cite as:
Oude Veldhuis S, Guichelaar MMJ, den Ouden MEM, van Gemert-Pijnen LJEWC, Drossaert CHC
Information Deficits, Information Needs, and Preferences Regarding eHealth in a Dutch Population With Metabolic Dysfunction–Associated Steatotic Liver Disease: Cross-Sectional Survey Study
JMIR Form Res 2025;9:e78536
URL: https://formative.jmir.org/2025/1/e78536
doi:10.2196/78536

© Sharon Oude Veldhuis, Maureen M J Guichelaar, Marjolein E M den Ouden, Lisette J E W C van Gemert-Pijnen, Constance H C Drossaert. Originally published in JMIR Formative Research (https://formative.jmir.org), 29.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Design, Implementation, and Evaluation of a Community-Based Phygital Telemonitoring Program for Older Adults: Multisite Retrospective Pilot Study in Singapore

Yichi Zhang[1,2], PhD; Michelle Cheok Yien Law[3], MBBS, PhD; Soon Keong Wee[3], PhD; Benjamin Sian Teck Lee[4,5], PhD; Bing Liang Alvin Chew[3,5], PhD; Wei Peng Teo[1,6], PhD; Edmund W J Lee[7,8], PhD

[1]Physical Education and Sports Science Department, National Institute of Education, Nanyang Technological University, Singapore, Singapore

[2]Ageing Research Institute for Society and Education, Interdisciplinary Graduate Programme, Nanyang Technological University, Singapore, Singapore

[3]Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore, Singapore

[4]School of Biological Sciences, Nanyang Technological University, Singapore, Singapore

[5]NTU Institute for Health Technologies, Interdisciplinary Graduate Programme, Nanyang Technological University, Singapore, Singapore

[6]Science of Learning in Education Centre, National Institute of Education, Nanyang Technological University, Singapore, Singapore

[7]Department of Media and Communication, City University of Hong Kong, M5010, 5/F Run Run Shaw Creative Media Centre, 18 Tat Hong Avenue, Hong Kong, China (Hong Kong)

[8]Institute of Digital Medicine, City University of Hong Kong, Hong Kong, China (Hong Kong)

Corresponding Author:
Edmund W J Lee, PhD
Department of Media and Communication, City University of Hong Kong, M5010, 5/F Run Run Shaw Creative Media Centre, 18 Tat Hong Avenue, Hong Kong, China (Hong Kong)

Abstract

Background: Noncommunicable diseases, particularly hypertension, diabetes, hyperlipidemia, and obesity, are on the rise among older adults in Singapore, emphasizing the need for effective screening, monitoring, and educational interventions. The traditional health care model, relying on in-person visits to review patients, poses risks of underreported cases and missed opportunities for early interventions to manage complications. Community-based telemonitoring programs present promising opportunities to extend telehealth services to underserved populations, thereby mitigating the digital divide and addressing health inequalities.

Objective: This study aimed to retrospectively evaluate the implementation of a community-based telemonitoring program, the Community Telehealth Service (CTS), developed to reduce digital barriers and raise awareness for regular health screening among older adults in the Singapore community. It also aimed to generate insights for scaling up similar telehealth initiatives in the community.

Methods: This retrospective study used the (1) Implementation Research Logic Model, (2) Reach, Effectiveness, Adoption, Implementation, and Maintenance framework, and the Implementation Outcomes Framework to guide the design and evaluation of CTS' implementation strategies. Outcomes covered implementation outcomes (reach, adoption, feasibility, and cost), service outcomes (safety and preliminary effectiveness), and user outcomes (satisfaction). Data were collected from operational statistics and structured user feedback surveys across 3 phases of implementation at different community sites.

Results: Over the course of the 3 phases, CTS has reached more than 800 older individuals and 147 health ambassadors, with the participation of community organizations, health care institutions, academic collaborators, corporate sponsors, and government agencies. Operational statistics indicated that CTS was delivered consistently across 3 sites, with improving show-up rates and stable service hours. User feedback was generally positive, citing convenience, perceived value, and appreciation for health ambassador support. However, challenges were noted in referral tracking due to differing workflows across partners and in collecting user feedback, particularly in later phases where the survey was perceived as lengthy and complex for older users. Several areas for improvement were identified, such as incorporating more health assessments, providing more health-related information, and improving the referral process.

Conclusions: This early-stage retrospective study suggests that community-based telehealth programs may be a feasible and acceptable approach to delivering preventive services in community settings. While initial findings are promising, further rigorous research is needed to evaluate long-term outcomes, integration with health systems, and potential for scale-up.

(JMIR Form Res 2025;9:e56905) doi:10.2196/56905

XSL•FO
RenderX

KEYWORDS

telehealth; telemonitoring; aging; health access; community health; technology; health inequalities; digital health; pilot study; implementation

Introduction

The global challenges of aging arise from people living longer and falling birth rates [1], leading to a projected doubling of the population aged 65 years or older from 2021 to 2050 [2]. It is accompanied by a gradual deterioration of physical and mental capacities and increased vulnerability to diseases [3]. Noncommunicable diseases (NCDs), such as cardiovascular diseases, cancers, chronic respiratory diseases, and diabetes, are particularly prevalent among older individuals and account for 74% of all deaths globally [4]. This demographic shift brings about societal challenges from the economy to health care policies [2,3].

Singapore has a rapidly aging population [5] with a surging proportion having multiple NCDs [6-8], and on average, an individual may spend the last 10.6 years of their lives in poor health conditions [3,4,9]. In a survey with 4549 community-dwelling Singapore residents aged 60 years and older, 38% of the respondents perceived their health as fair or poor, and the “3-highs” (high blood pressure or hypertension, high blood cholesterol or hyperlipidemia, and high blood sugar or diabetes) ranked among the top 5 most prevalent self-reported chronic conditions [7]. Obesity has also been observed to have a noticeable increase across all age groups since 2017, particularly among individuals aged 50 to 74 years [8]. This underscores a pressing need for effective screening, monitoring, and educational interventions targeted at this demographic [10-12], so as to pursue diverse interests and make meaningful contributions throughout their lives [2,3].

In the traditional health care model, health professionals could only review their patients during in-person visits, and complications may arise along with missed opportunities for early intervention [13]. This is notwithstanding underreported cases due to low participation in health screening and self-measurement practices, especially the “3-highs” which disproportionately affect the underprivileged [10-12,14]. Furthermore, the COVID-19 pandemic led to the deferment of nonurgent health services, lowered participation in NCD and cancer screenings [15], and the safe management measures might have impacted older individuals’ healthy lifestyle practices [15].

These factors spurred the use of telehealth, which was brought about by the advent of smart health technologies, such as wearable sensors coupled with mobile apps or computer software [16]. It simplified and enhanced the access to taking serial measurements and self-monitoring of physiological parameters outside the traditional health care settings [16]. This has shown potential in providing reliable health data and empowering older individuals, caregivers, and health care professionals alike to create better awareness of one’s vitals, influence health behaviors, and track interventions toward a better overall quality of life [17-19], especially when analyzed in conjunction with individualized reference ranges [20].

Although local health institutes have introduced home-based telemonitoring services to aid patients in monitoring their NCDs [13,21-25], these services may not be readily available to those with low health screening participation or undetected conditions. Furthermore, some barriers associated with this approach may make it less suitable for older adults, such as perceived low usefulness [26,27], concerns over cost [26], challenges in accessing or operating the technologies without sufficient support [27,28], as well as privacy and safety concerns [27,28].

To overcome the limited access to home-based telemonitoring, community-based telemonitoring kiosks have emerged as an alternative. Courtney et al [29] conducted a qualitative study exploring the acceptability and perceived value of such kiosks among older adults and community case workers, and findings highlighted their potential in reducing health disparities and bridging the digital divide. While such kiosks can deliver health services to underserved populations with minimal or no cost [30,31], their success is often hindered by low user engagement [30], stemming from unawareness, uncertainty about their benefits, and lack of motivation [32,33]. This underscores the importance of addressing contextual barriers as well as designing and evaluating tailored implementation strategies early to support successful adoption and scale-up.

This retrospective pilot study aims to evaluate the implementation of a community-based telemonitoring program designed for older adults, Community Telehealth Service (CTS), and to generate insights to inform future scale-up for community-based telehealth programs. In particular, we will answer 2 research questions. First, can such a community-based telehealth program be feasibly implemented to support older adults in real-world settings? Second, what are the key areas of improvement to motivate the uptake of community-based telehealth programs?

Methods

Intervention Design

CTS was developed to reduce older adults’ barriers to adopting digital health technologies and to raise their awareness of the importance of regular health monitoring, ultimately increasing the uptake of preventive services. Its design was guided by the Capability, Opportunity, Motivation-Behavior (COM-B) model [34], which emphasizes the interplay of capability, opportunity, and motivation in behavior change. Phygital telehealth kiosks were strategically located in easily accessible community spaces to create physical and social opportunities; trained health ambassadors provided on-site assistance to enhance users’ capability in navigating digital tools; and real-time and personalized health feedback was delivered to strengthen motivation and support long-term engagement with health monitoring behaviors. CTS has been designed not as a replacement but instead as a complement to traditional health care delivery for the older population, by lowering their barriers to accessing personal health monitoring technologies, linking

XSL•FO
RenderX

them back to the formal health care system, and extending health care beyond the clinical settings.

Target Population

While our primary focus was on Singapore residents (including foreigners) aged 50 years and older, we remained open to the younger age groups. Within this broader range, we gave priority to those who were born in or before the 1950s as well as the low-income groups (household monthly income per person lower than S $2000 which is equivalent to US $1473).

Health Assessments and Technologies

To tackle the rising prevalence and underreporting of the "3-highs" and obesity rates in Singapore [10-12,14], we integrated a suite of point-of-care assessments to enable early detection and monitoring. These included artificial intelligence–assisted eye screening; cordless upper-arm blood pressure and heart rate monitors; smart scales to measure weight, BMI, and body fat composition; and capillary blood glucose measurement.

Phygital Kiosk

To create a safe and private environment for our telemonitoring program, we repurposed the biosafety booths, which were originally designed for mass swabbing exercises [35] (Multimedia Appendix 1). The adaptable modular design of the kiosks allowed us to customize the telemonitoring program according to specific requirements. In addition, the kiosks were equipped with an internet connection that enabled the uploading of health data, printing of health reports, and conducting teleconsultation when necessary.

Manpower

At each implementation site, trained health ambassadors were stationed to provide real-time assistance and support to users, particularly those with limited digital or health literacy. These ambassadors helped users navigate the telehealth kiosks, conduct health assessments, and facilitate understanding of their results.

Referral Management System

The integration of a referral management system was a pivotal and essential component of our program. Recognizing the importance of seamless coordination and collaboration with health care providers, we established a streamlined procedure to ensure that users in need could be identified and referred to polyclinics and community nurse posts to receive timely, targeted, and subsidized diagnosis and follow-up care.

User Journey

The user journey in our telemonitoring prototype is illustrated in Figure 1. Health ambassadors would first verify users' identities using a secure registration system on smart tablets, then guide the users through the profile surveys with health history, various measurements, and communication of the results, as well as any remarks from the users. Users' eye reports were sent to the Singapore National Eye Center for manual verification, while the other devices had been approved by the Health Sciences Authority and were able to provide immediate results under the operation of trained health ambassadors. In the event of abnormal readings, users would be directed to the referral system for professional diagnosis and follow-up care.

XSL•FO
RenderX

Figure 1. Overview of Community Telehealth Service phygital telehealth user journey.

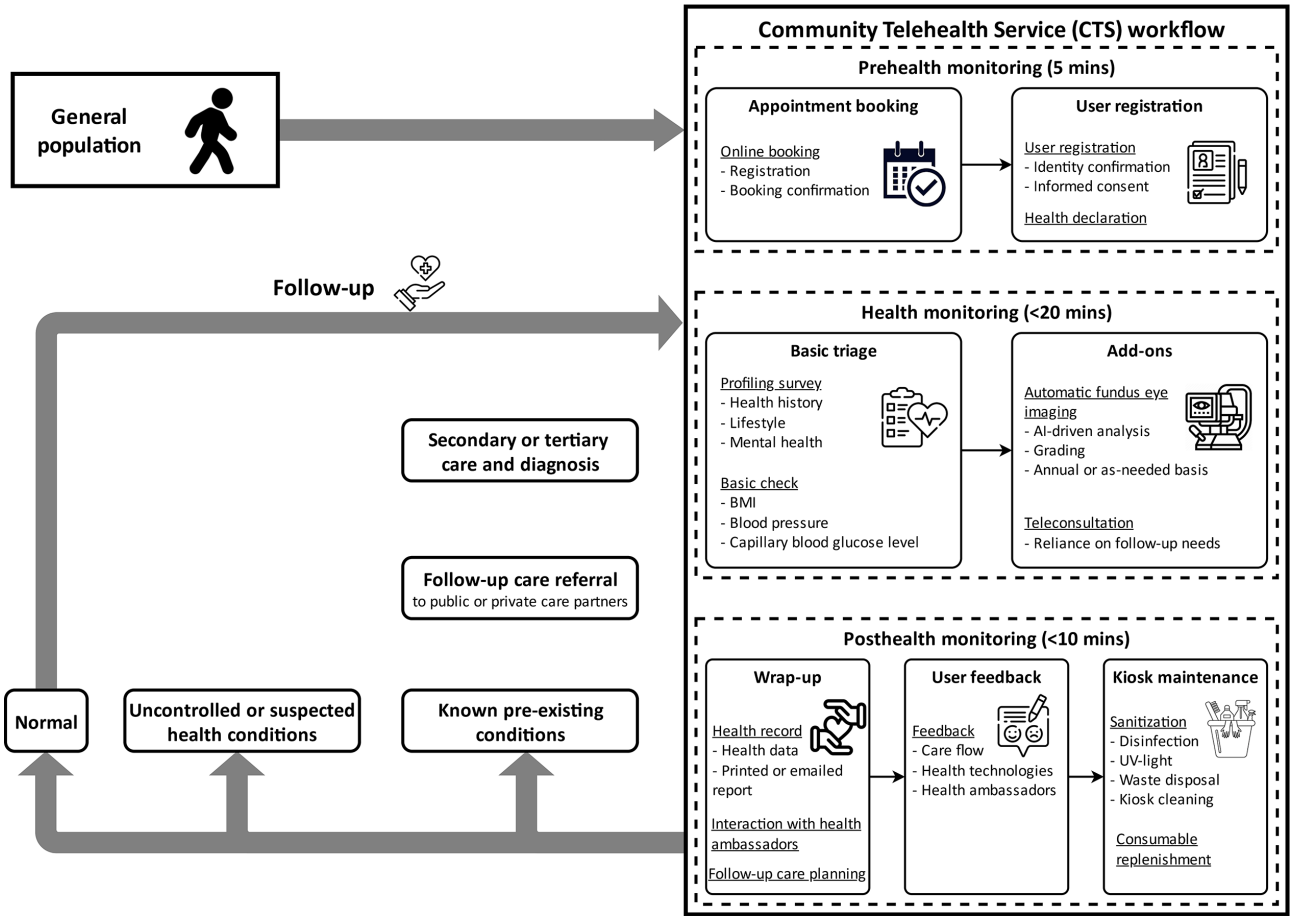

Determinants and Implementation Strategies

The development and evaluation of CTS' implementation strategies were guided by the Implementation Research Logic Model (IRLM) [36], which provides a structured approach to map the context-specific determinants, Expert Recommendations for Implementing Change compilation (ERIC)−aligned implementation strategies, underlying mechanisms of action, and key outcome measures (Figure 2).

Figure 2. Implementation Research Logic Model to guide the design and evaluation of Community Telehealth Service implementation strategies. IOF: Implementation Outcomes Framework; RE-AIM: Reach, Effectiveness, Adoption, Implementation, and Maintenance Framework; SEM: social ecological model.

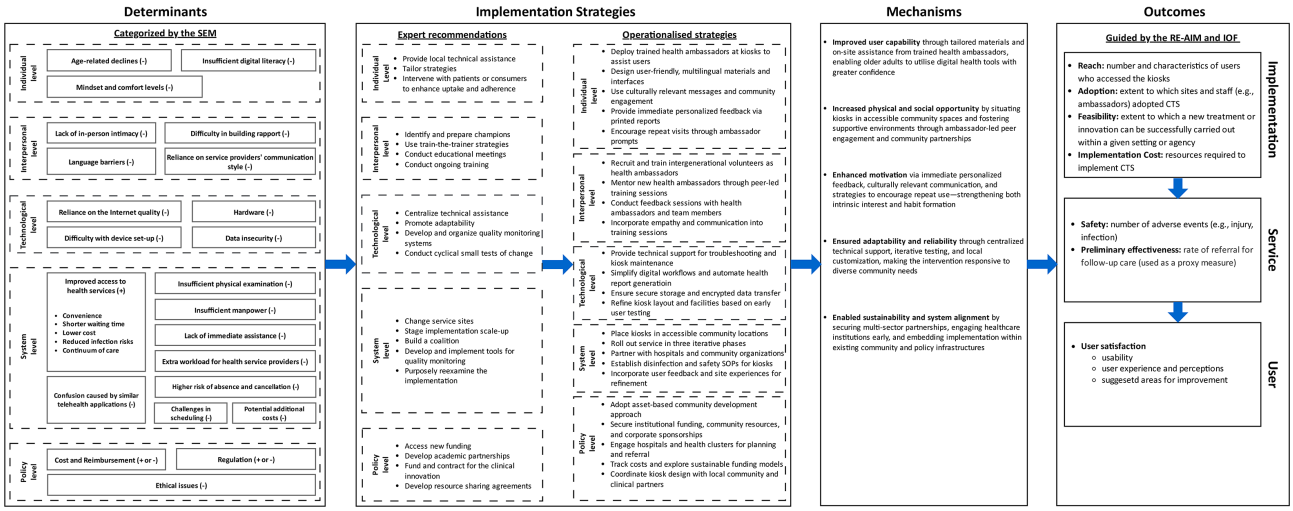

Determinants

Facilitators and barriers to telehealth implementation were identified through our scoping review [37] and an earlier local qualitative study involving 15 older adults, 12 health service providers, and 2 telehealth providers. These determinants were organized according to the social ecological model (SEM) [36] across 5 levels, which are individual, interpersonal, technological, system, and policy levels. Key barriers included

XSL•FO
RenderX

low digital literacy, difficulty building rapport with telehealth deliverers, reliance on internet connectivity, insufficient physical examination, insufficient manpower to deliver telehealth services, as well as ethical and data security concerns.

Implementation Strategies

Based on the determinants identified, we selected a set of implementation strategies from the ERIC taxonomy [38] and operationalized them across the SEM levels to address the multilevel determinants.

At the individual level, the trained health ambassadors provided users with on-site assistance, friendly and multilingual materials, as well as personalized health feedback upon receiving the printed health reports to enhance engagement. At the interpersonal level, ambassadors from different age groups were recruited and trained using a train-the-trainer model that covered not only basic health knowledge and health assessment skills but also empathy and communication techniques. All training materials were reviewed by professional medical advisors to ensure clinical accuracy and alignment with local guidelines. The health ambassadors were paired to comanage each kiosk—typically matching a younger ambassador who was more adept with technical operations with an older ambassador who brought stronger interpersonal communication skills (Multimedia Appendix 2). This intergenerational pairing was intentional to enhance both technical proficiency and user engagement, fostering a warm and age-friendly environment that resonated with older adults. Technological strategies included centralized technical support, simplified digital workflows, and iterative user testing to improve system usability and reliability. At the system level, our telehealth kiosks were deployed in accessible community locations through phased implementation, in collaboration with community, health care, and corporate partners. Safety protocols and feedback loops were established to guide iterative improvements. At the policy level, funding was secured through university, community sources, and corporate sponsorships, while academic partnerships were formed to support planning, training, and sustainability.

Outcome Evaluation and Data Collection

The evaluation of the CTS implementation strategies was guided by the Reach, Effectiveness, Adoption, Implementation, and Maintenance (RE-AIM) framework [39,40] and complemented by the Implementation Outcomes Framework [41], to assess key domains of implementation performance and user experience in real-world settings. These outcomes were also categorized according to the IRLM into three domains: (1) implementation outcomes, (2) service outcomes, and (3) user outcomes.

Implementation outcomes focused on the extent to which CTS could be successfully executed, including its reach (number of users as well as the number of health screenings and health ambassador training sessions conducted), adoption (number of participating community partners, social service agencies, medical partners, and academic partners), feasibility (number of operating hours, cancellation, and no-show rates), and implementation cost. Service outcomes included safety (the number of adverse events reported) and preliminary effectiveness (the rate of referral for follow-up care as a proxy). Both implementation and service outcomes were collected from the operational statistics.

User outcomes focused on satisfaction, covering CTS usability, user experience with and perceptions of CTS, and suggested areas for improvement collected through structured user feedback surveys. All users were invited to fill out feedback on Google Forms at the conclusion of their screenings, and their participation was encouraged but not mandatory. In phase 1, the feedback survey encompassed 6 short questions to explore users' experience with appointment booking, rating of the waiting time, satisfaction with health ambassadors' instructions, overall experience, intention to continue using CTS, and perceptions of the program (Multimedia Appendix 3). In phases 2 and 3, an adapted version of the Telehealth Usability Questionnaire developed by Parmanto et al [42], a robust tool to assess the usability of telehealth implementation and services, was used. The adapted tool included 24 items rated on a 5-point Likert scale, along with an extra open question to cover various usability factors of the enhanced services, such as publicity, registration and waiting time, service reliability, interaction with health ambassadors, perceived usefulness and ease of use, quality of services, satisfaction and future use, and potential areas to improve (Multimedia Appendix 3). In addition to the survey, key areas for improvement were identified through health ambassador feedback and on-site observations during the program implementation and were summarized thematically to inform iterative program refinement.

Data Analysis

Quantitative data were analyzed using RStudio and presented using descriptive statistics, while qualitative data from open-ended user feedback and on-site observations were analyzed using rapid thematic analysis [43], given that the number of qualitative feedback responses was manageable and could be effectively analyzed manually.

Ethical Considerations

This project was initially approved and implemented under the Graduate College – Community Engagement Project (CEP), Nanyang Technological University [44], and therefore did not undergo ethics review at the time of user enrollment. At the time of enrollment in CTS, all users provided consent to participate in the program and to allow their anonymous data to be used for analysis. For this retrospective secondary analysis, only fully anonymized data were used, and no additional consent was required. Institutional review board approval was obtained from the University's Institutional Review Board (IRB-2022-970). The study team had full permission to access and analyze the dataset for research purposes. All procedures adhered to the principles of the Declaration of Helsinki and local regulatory requirements.

Results

Phased Implementation

CTS was deployed in 3 phases progressively. In phase 1, the telehealth kiosk was deployed in the Northeast of Singapore

XSL•FO
RenderX

daily from mid-January to the end of February 2021, after the circuit breaker (a strict stay-at-home order and cordon sanitaire by the government [45,46]). The aim of this phase was to rapidly collect preliminary user feedback from the users for program enhancement for later phases. Based on user feedback in phase 1 (Table 1), we brought various enhancements to phases 2 and 3. In phase 2, CTS was launched in the West, from end-September 2021 to end-March 2022. Due to the heightened safety measures to slow the spread of COVID-19 [47], the program was temporarily closed for 2 months from October to December 2021. In phase 3, the program was launched in the East in January 2022 and remained operational for a 3-month period until March 2022. User feedback in phases 2 and 3 is summarized in Table 2. Details of the implementation of CTS across 3 phases are summarized in Table 3.

Table . Summary of user feedback in phase 1.

Item and response	Phase 1 (n=389; response rate=62%), n (%)
Did you register an appointment timeslot by yourself?	
Yes	258 (66.3)
No	131 (33.7)
How do you find the waiting time required from the start of the booth session to receiving your health report?	
Appropriate	324 (83.3)
Too short	42 (10.8)
Too long	23 (5.9)
How do you find the verbal instructions given by the health ambassadors?	
Very clear	311 (79.9)
Clear	55 (14.1)
Acceptable	22 (5.7)
Unclear	1 (0.3)
How do you find the overall CTS[a] experience?	
Very good	304 (78.1)
Good	70 (18)
Acceptable	14 (3.6)
Not good	1 (0.3)
Overall, do you want to continue seeing and using the CTS booth as a community resource for health monitoring?	
Yes	383 (98.5)
No	6 (1.5)

[a]CTS: Community Telehealth Service.

XSL•FO
RenderX

Table . Summary of user feedback in phases 2 and 3.

Item and response	Phase 2 (n=103; response rate=61; 59.2%)		Phase 3 (n=79; response rate=12; 15.2%)	
	Mean (SD)	Mode	Mean (SD)	Mode
Registration and waiting time				
I think the waiting time is appropriate.	4.8 (0.5)	5	4.3 (0.8)	5
Reliability of health monitoring				
I think the visits provided over CTS are the same as in-person visits to the clinics.	4.3 (0.9)	5	3.8 (0.8)	3
When the machine/equipment/monitor reported any error (eg, the glucose meter showed an error message or the eye machine cannot capture eye images), the health ambassador could resolve the issue easily and quickly.	4.4 (0.9)	5	4.1 (0.8)	4
The health ambassadors could clearly tell me the problems with the health indicators shown in the health report.	4.7 (0.7)	5	4.3 (0.7)	5
I have confidence in the health ambassadors' skills.	4.5 (0.7)	5	4.3 (0.6)	5
I have got sufficient information about my health conditions.	4.6 (0.6)	5	4.0 (0.7)	4
Interaction with health ambassadors				
I feel that the health ambassadors cared about me.	4.8 (0.5)	5	4.3 (0.6)	4
I had enough time to talk and interact with the health ambassadors.	4.7 (0.6)	5	4.3 (0.6)	4
I felt I was able to express myself effectively.	4.6 (0.7)	5	4.3 (0.6)	4
I could hear the volunteers/health ambassadors clearly in CTS.	4.8 (0.5)	5	4.4 (0.6)	5
I feel that the health ambassadors talked to me in a way that was easy to understand.	4.8 (0.5)	5	4.4 (0.6)	5
I felt comfortable communicating with the volunteers/health coaches.	4.8 (0.5)	5	4.4 (0.6)	5
Usefulness and ease of use				
CTS improves my access to healthcare services.	4.6 (0.6)	5	4.3 (0.6)	4
CTS saves me time traveling to a hospital or specialist clinic.	4.6 (0.7)	5	4.4 (0.6)	5
CTS provides for my healthcare needs.	4.4 (0.8)	5	4.1 (0.6)	4
CTS is simple to use.	4.7 (0.7)	5	4.4 (0.6)	5
Quality of service				

XSL•FO
RenderX

Item and response	Phase 2 (n=103; response rate=61; 59.2%)		Phase 3 (n=79; response rate=12; 15.2%)	
	Mean (SD)	Mode	Mean (SD)	Mode
My interaction with CTS is pleasant.	4.8 (0.4)	5	4.4 (0.6)	5
I think CTS is well-organized.	4.7 (0.6)	5	4.3 (0.7)	4
I feel that the service that CTS provided is satisfactory.	4.8 (0.5)	5	4.3 (0.7)	5
Satisfaction and future use				
CTS is an acceptable way to receive healthcare services.	4.7 (0.6)	5	4.3 (0.6)	4
I would use CTS again.	4.6 (0.7)	5	4.4 (0.6)	5
Overall, I am satisfied with CTS.	4.7 (0.6)	5	4.3 (0.6)	4

XSL•FO
RenderX

Table . Details of the implementation of Community Telehealth Service across 3 phases.

Element	Phase 1	Phase 2	Phase 3
Period of deployment	• January 9 to February 27, 2021	• September 25, 2021, to March 31, 2022 (temporary closure from October to December 2021)	• January 3, 2022, to March 31, 2022
Location	• Punggol 21 Community Center (Northeast) – located near digital ambassadors who could assist older adults in improving their digital skills	• Bukit Batok (West) – located beside the partner senior center	• Chai Chee (East) – located near the partner nonprofit charitable health care organization
Operating hours	• 9 AM to 5 PM daily	• Flexible opening hours based on the booking and health ambassadors' availability with at most 3 days per week	• Flexible opening hours based on the booking and health ambassadors' availability with at most 3 days per week
Health assessments and technologies	• Weight • BMI • Body fat • Muscle • Visceral fat • Blood pressure • Heart rate • Glucose level • Eye health and risk of cardiovascular diseases – sent to the Singapore National Eye Center for manual verification)	• Mental health test using WHO-5[a] • Weight • BMI • Body fat • Muscle • Visceral fat • Blood pressure • Heart rate • Glucose level • Eye health (risks of diabetic retinopathy, glaucoma, and age-related macular degeneration) – automatically generated by the eye machine	• Mental health test using WHO-5 • Weight • BMI • Body fat • Muscle • Visceral fat • Blood pressure • Heart rate • Glucose level • Eye health (risks of diabetic retinopathy, glaucoma, and age-related macular degeneration) – automatically generated by the eye machine
Referral management system	• Sent those reports with possible abnormalities to Sengkang Hospital Community Nurse Post	• Referred users with possible abnormalities to a National University Polyclinic nearest to users' homes, and included brief notes in the printed health reports	• Referred users with possible abnormalities to SATA CommHealth, a non-profit charitable health care organization, and included brief notes in the printed health reports
Publicity	• Social media • Door-to-door publicity • Displaying publicity posters in the lift lobbies • Media mentions	• Social media • Door-to-door publicity • Community outreach • Roadshow	• Social media • Door-to-door publicity • Community outreach • Displaying publicity posters on digital display panels of residential blocks • Roadshow

[a]WHO-5: World Health Organization-Five Well-Being Index.

Implementation Outcomes

All implementation outcomes are summarized in Table 4.

XSL•FO
RenderX

Table . Summary of the operational statistics.

Outcome	Phase 1	Phase 2	Phase 3
Total number of operation hours	• 236 h (health screening and health ambassador training were held concurrently)	• 243 h – screening • 47 h – health ambassadors training	• 90 h – screening • 36 h – health ambassadors training
Total number of screenings conducted	• 627 screenings: 277 males and 350 females • Mean age 57.0 (SD 10.2) years	• 103 screenings: 30 males and 68 females (5 returning users) • Mean age 65.0 (SD 9.5) years	• 79 screenings: 38 males (3 returning users) and 37 females (1 returning user) • Mean age 68.0 (SD 9.2) years
Cancellation and no-show rates	• About 26.8%[a]	• About 15.6%[a]	• About 11.2%[a]
Referral rates	• 26%	• Untracked	• Untracked
Total number of volunteers trained	• 55 student health ambassadors • 11 community health ambassadors • 45 more volunteers on the waitlist	• 30 student health ambassadors[b] • 51 community health ambassadors (excluding health ambassadors trained at phase 1)[b]	• 30 student health ambassadors[b] • 51 community health ambassadors (excluding health ambassadors trained at phase 1)[b]
Implementation cost	S $19,884.73 (US $15,423.31)	S $3806.24 (US $2952.26)[c]	S $3806.24 (US $2952.26)[c]

[a]Some walk-in visitors may have requested partial screenings that were not accounted for in the registration system.

[b]The numbers of student and community health ambassadors (n=30 and n=51, respectively) trained is the total number in both phases 2 and 3 instead of each phase.

[c]The total cost for phases 2 and 3 is S $3806.24 (US $2952.26).

Reach

Over the 8 weeks in phase 1, a total of 627 screenings for 277 males and 350 females (mean age 57.0, SD 10.2 y) were conducted and 66 health ambassadors were trained, while another 45 registered volunteers were waiting for training. Phase 2 completed 103 screenings (30 males, 68 females, mean age 65.0, SD 9.5 y, of whom 5 were returning users. Phase 3 reached 79 users (38 males, 37 females, mean age 68.0, SD 9.2 y), including a total of 4 returning users.

Across all 3 phases, similar recruitment strategies were used, although media release proved most effective in phase 1. In phase 2 (n=103, 61 responses were received with a response rate of 59.2%), most users learned about CTS through publicity materials (n=22, 36.1%) and door-to-door publicity (n=15, 24.6%), while others cited passing-by (n=10, 16.4%), social media (n=8, 13.1%), and referrals from their social network (n=6, 9.8%). Over half (n=34, 55.7%) registered independently, and those who received assistance were mainly supported by health ambassadors (n=15, 55.6%), followed by family members or helpers (n=9, 33.3%) and community center staff (n=3, 11.1%). In phase 3 (n=79, 12 responses were received with a response rate of 15.2%), awareness was primarily through passing-by (n=5, 41.7%) and door-to-door publicity (n=4, 33.3%), with fewer citing publicity materials (n=3, 25%). Fewer registered independently (n=5, 41.7%), and all assisted registrations were supported by health ambassadors.

Adoption

During phase 1, partnerships were established with 4 community partners, 2 medical partners, and 6 corporate partners. Funding was received from 5 sources, including research grants, university CEP funding, as well as corporate sponsorships. As the service progressed into phases 2 and 3, partnerships expanded to 6 grantors, 10 government and community partners, 2 medical partners, and 8 corporate sponsors. In addition, an additional academic partnership was established.

Feasibility

CTS operations in phase 1 spanned 236 hours, during which health screenings and ambassador training were conducted concurrently. The cancellation or no-show rate was approximately 26.8%. In phase 2, the program operated for a total of 290 hours (243 h for screening and 47 h for training), with a no-show rate of 15.6%. Phase 3 recorded 126 total operating hours (90 h for screening and 36 h for training), with a cancellation or no-show rate of 11.2%.

Implementation Cost

The implementation costs for all 3 phases were estimated based on the sponsorship with actual claims made to research funding and university CEP funding. The implementation costs for phase 1 and phases 2 and 3 were S $19,884.73 (US $14,913.55) and S $3806.24 (US $2809.01), respectively.

Service Outcome

Safety

The project remained free of adverse events throughout all 3 phases, with no occurrences of COVID-19 infections, injuries, or data breaches reported. This absence of incidents underscores the effectiveness of the safety protocols and the efforts of all stakeholders in maintaining a secure and safe environment.

Preliminary Effectiveness – Referral Rate

We referred 163 (26%) users to the community nurse post and Singapore National Eye Center for follow-up care in phase I.

XSL•FO
RenderX

We failed to consistently track referral rates in phases 2 and 3, as the medical partners involved operated under distinct referral workflows and documentation systems. In addition, a resurgence of COVID-19 cases during these 2 phases caused interruptions in referral processes, particularly affecting coordination between community sites and health care providers. These variations made it difficult to obtain a standardized and complete record of referral outcomes across both locations.

User Outcome – Satisfaction

Usability

In phase 1, we received a total of 389 user feedback responses (response rate=62%), and the responses were predominantly positive (Table 1). In addition, 66.3% (258/389) of users were able to book an appointment independently and the rest required some assistance from either their social networks or our health ambassadors. Furthermore, 83.3% (324/389) of the users were satisfied with the waiting time. Moreover, 94.1% (366/389) of the users were happy with the waiting time, 99.7% (388/389) of the users found the verbal instructions given by the health ambassadors satisfactory, and 99.7% (388/389) expressed overall satisfaction with the experience. In general, 98.5% (383/389) of the users expressed interest in continuing to use the CTS as a community resource for health monitoring.

In phases 2 and 3, a total of 73 user feedback responses were collected (61 from phase 2 and 12 from phase 3). Overall, users expressed high satisfaction and perceived CTS as a usable service. However, response rates were lower compared with phase 1, which several users attributed to the length and complexity of the questionnaire, particularly challenging for older users. Table 2 presents the survey items along with a summary of the subscale scores.

User Experience With and Perceptions of CTS

User feedback from phase 1 reflected strongly positive perceptions of CTS as a health initiative tailored for older adults. Many users appreciated the convenience of receiving multiple health screenings in one location:

> *Very convenient. All tests are done at the same spot without the hassle of moving from room to room. [It is] good especially during this COVID-19 times.*

Users also described the service as smooth and efficient and praised the professionalism and warmth of the health ambassadors, for example:

> *I am happy [that] we have fellow Singaporeans with the heart to serve the nation. They are doing a good job.*

Several users appreciated the benefits of such community-based telehealth kiosks for the older population and expressed hope that CTS would be expanded to other regions, calling it:

> *I am very glad this event was organized for the benefit of the Singapore seniors. I hope this will be a regular event for other parts of Singapore as well…*

Suggested Areas for Improvement

While CTS received a high level of acceptability across all 3 phases, user feedback also highlighted potential areas for improvement, for instance, simplifying the registration process, enhancing the environment inside the kiosk and at the waiting areas, and improving the clarity of health reports and publicity materials. A summary of these suggestions and representative quotes is provided in Table 5. Building upon the insights collected from phase 1, enhancements were implemented in phases 2 and 3, such as (1) assigning health ambassadors to timely answer users' inquiries via WhatsApp (Meta) messages; (2) streamlining the registration process, enhancing the user-friendliness of the registration platform, and sending booking confirmations and reminders through SMS text messages; (3) translating the publicity materials into not only Chinese but also Malay and Tamil; (4) involving medical students and health professionals to enhance the abnormality detection algorithms by taking possible age-specific reference ranges into consideration; and (5) improving the physical environment by providing fans both inside and outside the telehealth kiosks, as well as increasing the seating capacity in the waiting area.

XSL•FO
RenderX

Table . Suggested areas for improvement.

Areas for improvement	Quotes
Simpler registration and booking confirmation, especially for the older adults	• "Give an immediate confirmation of registration for an easy note of the appointment date." • "The date of birth cannot be selected effectively." • "Registration may be difficult for some elderly."
Enhancing the environment inside the kiosk and in the waiting area	• "Prepare some signs to take a seat…" • "Cooler aircon preferred…" • "One thing that could be improved is the space. Currently is a bit too tight to conduct the check-up."
Service scale-up (eg, offering additional health assessments, extending operations to more locations, conducting screenings on a regular basis, lengthening service operating hours, and providing support in multiple languages to better serve diverse communities)	• "…Wish that there could be more other tests incorporated with this…" • "Should extend to more regions as this is very convenient to the seniors." • "If [you could] have one year check-up [,then it] will be better." • "…I hope this will be a regular event for other parts of Singapore as well…" • "Appreciate availability of dialect-speaking volunteers"
Improving health reports (eg, including reference ranges in the printed reports and having results reviewed or explained by qualified health care professionals to enhance clarity and credibility)	• "Include normal range in the report for comparison" • "…and have certified medical personnel/doctors to go through the result of the test…"

In addition to these readily achievable improvements, three more popular suggestions were shared via the health ambassadors: (1) extending the current spectrum of health assessments, (2) providing additional educational health information, and (3) upgrading the current referral management system.

Discussion

Principal Findings

The rapidly growing aging population, coupled with the rising prevalence of NCDs and mental health conditions, places escalating pressures on the traditional health care delivery model, rendering disease-centric approaches insufficient and unsustainable in the long run [48,49]. This calls for a paradigm shift toward preventive care [50-52], the creation of new health care infrastructure [50,53-55], the expansion of social and long-term care services within the community [50,56-59], as well as the adoption of technologies to support aging in place [60-62].

To our best knowledge, CTS was the first community-based telehealth program implemented in Singapore. Our CTS is designed to complement the traditional health care model by lowering barriers to accessing personal health monitoring technologies, extending the reach of health care beyond the clinical settings, and establishing a connection between the older population and the formal health care system. Building upon community-based telehealth programs in existing literature, the design, implementation, and evaluation of CTS were guided by IRLM, incorporating different contextual determinants and corresponding implementation strategies identified by our earlier review, qualitative studies, as well as expert recommendations (ERIC). This comprehensive approach aims to improve the rigor and reproducibility of such community-based telehealth services and address the gaps in older individuals' health-seeking behavior while mitigating health inequity. Findings from this pilot study demonstrate the potential of CTS within the Singapore community and provide insights to pave the way for more comprehensive community-based telehealth programs.

Potential Areas for Improvement

Based on the user feedback collected, one of the most anticipated improvements is the inclusion of additional health assessments. Advancements in technology have made it possible to incorporate point-of-care tests, such as urine dipsticks, for remote patient monitoring, even by nonmedical individuals [63]. Furthermore, future enhancements may involve integrating physical performance and muscle strength tests (eg, handgrip strength) and cognitive assessments. Nonetheless, additional research efforts are needed to validate these tests when administered by trained nonmedical personnel to a specific subgroup of the population in nonclinical settings.

The second important area for improvement is the provision of more health-related information, which also aligns with users' interest in a study by Zhang et al [30]. Besides, we also observed a considerable proportion of users who, despite having borderline health conditions or being diagnosed with NCDs, opted out of health screenings, refused medication, or discontinued follow-up care. Guided by these insights, we will work toward implementing different health promotion interventions targeted at empowering older adults to take control of their health journeys and encourage proactive engagement in decisions that impact their overall well-being [64].

The referral management system was observed as another opportunity for improvement. During phase I, we encountered some delays, primarily attributed to our reliance on medical partners to initiate contact with users as well as their increased workload amid the pandemic. As we transitioned into phases 2 and 3 where users were given full autonomy to make decisions, new challenges surfaced: some users refused the referrals, and

XSL•FO
RenderX

it was difficult to track their progress. These experiences underscore the importance of establishing efficient and timely communication channels with medical partners to ensure the seamless provision of quality care across the care continuum. Furthermore, robust strategies should be also devised to effectively engage older people and maintain their engagement within the care continuum, contributing to better overall well-being within the community.

Lessons Learned

Human Element in Community-Based Telehealth Programs

Technology advancements have enabled individuals to remotely monitor their physiological parameters outside the clinical settings and seek actionable insights for improving health outcomes [65,66]. Besides income, education, employment, and environment, digital technology has emerged as a new key determinant of health [67,68]. It is noteworthy that while digital health interventions are intended to reduce health inequalities by extending access to health services and addressing unmet health needs through personalized care, there is a concern that they might unintentionally exacerbate health inequities to some extent [69-71]. Therefore, digital inclusion–informed strategies are needed in designing community-based telehealth programs.

In response, we adopted a "phygital" approach, combining both digital and physical elements. While we acknowledge that the implementation of the health ambassador scheme in CTS may limit the generalizability of our findings to other community-based telehealth programs that lack or have minimal human support, as discussed in a study by Zhang et al [30], this approach has the potential to bring significant value to telehealth program delivery. For example, a study by Courtney et al [29] raised several concerns, such as older adults overreacting to minor abnormalities in the health data, the potential risk of using the kiosk as a substitute for regular or urgent health care services, and the false reassurance of blood pressure alone being a comprehensive indicator of overall health status.

In CTS, our well-trained health ambassadors could play a crucial role in mitigating these concerns. Besides conducting health measurements and generating health reports, they assisted users in explaining their measurements based on local clinical guidelines and taking notes of some remarks (eg, their perceived health and self-observed individualized reference ranges) in the reports. This combination of subjective perceptions and objective measurements enhances the delivery of more person-centered care [72]. In addition, health ambassadors could offer actionable advice for lifestyle modifications to positively impact health outcomes; they can also actively encourage all users to undergo regular health screenings at medical institutions and refer them to appropriate medical partners for follow-up care when necessary. Such a comprehensive approach serves to not only minimize the potential misinterpretation of health data but also foster a proactive mindset toward maintaining overall health and well-being.

Data Integration Remains a Challenge in Community-Based Telehealth Programs

Community-based telehealth programs allow users to regularly monitor health at community-based facilities, and thus generate a comprehensive record of their health data. The integration of these diverse objective health data points, subjective self-perceived health status, potential individualized reference range, and lifestyles, together with users' participation in community events and health promotion activities, could support primary care providers in forming a more complete and personalized understanding of an older individual's health. This, in turn, enables the delivery of more effective, customized, and equitable health care services.

However, challenges arise as these data often remain disconnected from primary care records, possibly due to data inconsistency, lack of standardized data handling and ownership protocols, concerns of breaching data privacy and confidentiality, as well as the risk of generating inaccurate or incomplete data [73,74]. Addressing these gaps and establishing mechanisms to bridge them could be a dedicated effort moving forward.

Limitations and Future Research

This study has several limitations. First, as a retrospective implementation study, CTS was not originally designed or executed as a research project, which limited the consistency and completeness of some data sources, as well as rigor in research design. For example, although qualitative feedback from users and health ambassadors provided valuable insights, not all feedback was formally documented through audio recordings or transcripts. Some themes were based on on-site observations and informal input, limiting the ability to include original quotes. Building on the promising findings, future research will adopt prospective and rigorous study designs, well-defined eligibility and screening processes, well-designed data collection procedures, and clearly defined implementation protocols to ensure standardized and systematic evaluations.

Second, the response rate for user satisfaction surveys in phases 2 and 3 was low. Several older users reported that the length and complexity of the questionnaire made it challenging to complete, reflecting a common challenge in real-world implementation settings where data collection must balance evaluation needs with user burden. Unlike controlled research environments, community-based programs have to accommodate diverse literacy and cognitive needs, especially among older populations. To improve data quality and inclusivity in future implementations, feedback tools could be simplified so that they are more user-friendly and context-appropriate.

Third, while CTS received multilevel approvals before its implementation, formal inputs from key stakeholders were not systematically captured during the evaluation. Stakeholder perspectives—including those of community partners, social service providers, health care professionals, and policymakers—are critical for understanding contextual enablers, aligning goals, and identifying potential barriers to scale-up. Future studies should incorporate structured stakeholder engagement as part of the scale-up and evaluation

XSL•FO
RenderX

process to inform implementation refinement, ensure policy relevance, and strengthen the long-term sustainability and integration of community-based telehealth models.

Finally, although this study was guided by the RE-AIM framework, we did not assess CTS' clinical effectiveness and maintenance, as it was implemented as a time-limited, early-stage pilot program. Future research will explore the longer-term impact of such community-based telehealth programs on users' health outcomes, behavioral outcomes, as well as users' maintained engagement over time. To advance this work, a hybrid type 2 effectiveness-implementation study could be used to simultaneously evaluate the clinical effectiveness of CTS while further refining and testing its implementation strategies in real-world settings. In addition, exploring how these programs can be integrated with existing health care systems and national preventive health strategies will be essential to understanding their scalability, sustainability, and broader public health value.

Conclusions

This retrospective study demonstrated the potential of implementing a community-based telemonitoring program for older adults in Singapore, which combines accessible physical infrastructure with digital health tools and on-site human support. While the early results are promising, further rigorous research is needed to evaluate long-term outcomes, system integration, and scalability. These findings offer valuable insights for designing equitable, user-centered telehealth interventions in community settings.

Acknowledgments

The authors wish to express their gratitude to the Graduate College, Nanyang Technological University, Temasek Trust Ltd, and the National Youth Council for the funding supporting; Mediwhale, EyRIS, Mandarin Opto-Medic, Samsung, Omron, Abbott, GetDocPlus, and Golden Wing Chemical for their generous financial and technical support; REACH Community Services, South West Community Development Council, Kembangan-Chai Chee Youth Network, Filos, SGCare, Nanyang Technological University Advancement Office, EOT Technologies Pte. Ltd., Community Nurse Post at Sengkang General Hospital, National University Health System polyclinics, and SATA CommHealth for their collaboration and support; and all health ambassadors, volunteers, users, advisors, and research team members for the efforts put in the successful completion of the Community Telehealth Service.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors MCYL, SKW, and BLAC declare that they are co-founders of EOT Technologies Pte Ltd, a social enterprise spinoff for the Community Telehealth Service. The sponsors listed in the acknowledgment have no involvement in the study and program design and development.

Multimedia Appendix 1
Original biosafe booth designed for mass swabbing exercises (left) and the repurposed booth (right).
[DOCX File, 2324 KB - formative_v9i1e56905_app1.docx]

Multimedia Appendix 2
Photos of a health ambassador pair (photos were taken by Caroline Chia / WhatAreYouDoing.sg).
[DOCX File, 1501 KB - formative_v9i1e56905_app2.docx]

Multimedia Appendix 3
User feedback survey questions in phases 1-3.
[DOCX File, 17 KB - formative_v9i1e56905_app3.docx]

References

1. Ageing: global population. World Health Organization. 2010. URL: https://www.who.int/news-room/questions-and-answers/item/population-ageing [accessed 2023-06-09]
2. United Nations Department of Economic and Social Affairs. UNDESA world social report 2023: leaving no one behind in an ageing world. 2023 URL: https://desapublications.un.org/file/1087/download
3. Ageing and health. World Health Organization. 2021. URL: https://www.who.int/news-room/fact-sheets/detail/ageing-and-health [accessed 2022-09-13]
4. Noncommunicable diseases. World Health Organization. 2022. URL: https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases [accessed 2023-06-10]

XSL•FO
RenderX

5. Population in brief 2022. : National Population and Talent Division, Strategy Group, Prime Minister's Office; Singapore Department of Statistics; Ministry of Home Affairs; Immigration & Checkpoints Authority; and Ministry of Manpower; 2022 URL: https://www.strategygroup.gov.sg/files/media-centre/publications/population-in-brief-2022.pdf [accessed 2025-10-09]
6. Choo F. Proportion of older adults with multiple chronic diseases surges. SingHealth. 2019. URL: https://www.singhealth.com.sg/news/tomorrows-medicine/proportion-of-older-adults-with-multiple-chronic-diseases-surges [accessed 2023-01-08]
7. THE SIGNS study by Duke-NUS researchers identify factors affecting active and productive ageing among older singaporeans. Duke-NUS Medical School. 2019. URL: https://www.duke-nus.edu.sg/allnews/the-signs-study-by-duke-nus-researchers-identify-factors-affecting-active-and-productive-ageing-among-older-singaporeans [accessed 2023-01-08]
8. Ministry of Health Singapore. National population health survey. 2020.
9. The burden of disease in singapore, 1990 – 2017: an overview of the global burden of disease study 2017 results. : Epidemiology & Disease Control Division, Ministry of Health, Singapore; Institute for Health Metrics and Evaluation; 2019 URL: https://www.healthdata.org/sites/default/files/files/policy_report/2019/GBD_2017_Singapore_Report.pdf [accessed 2025-10-09]
10. Seow LSE, Subramaniam M, Abdin E, Vaingankar JA, Chong SA. Hypertension and its associated risks among Singapore elderly residential population. Journal of Clinical Gerontology and Geriatrics 2015 Dec;6(4):125-132 [FREE Full text] [doi: 10.1016/j.jcgg.2015.05.002]
11. Man REK, Gan AHW, Fenwick EK, et al. Prevalence, determinants and association of unawareness of diabetes, hypertension and hypercholesterolemia with poor disease control in a multi-ethnic Asian population without cardiovascular disease. Popul Health Metr 2019 Dec 5;17(1):17. [doi: 10.1186/s12963-019-0197-5] [Medline: 31806040]
12. Singapore diabetes report 2000 — 2050. International Diabetes Federation. 2011. URL: https://diabetesatlas.org/data-by-location/country/singapore/ [accessed 2023-05-26]
13. Sin DYE, Guo X, Yong DWW, et al. Assessment of willingness to tele-monitoring interventions in patients with type 2 diabetes and/or hypertension in the public primary healthcare setting. BMC Med Inform Decis Mak 2020 Jan 28;20(1):11. [doi: 10.1186/s12911-020-1024-4] [Medline: 31992288]
14. Ministry of Health Singapore. National health survey. 2004.
15. National population health survey 2021 (household interview). Ministry of Health Singapore. 2021. URL: http://www.statcan.ca.myaccess.library.utoronto.ca/english/concepts/nphs/ [accessed 2023-06-01]
16. Scott J. How medical devices work with mhealth apps. HealthTech. 2022. URL: https://healthtechmagazine.net/article/2022/06/how-medical-devices-work-with-mhealth-apps-perfcon [accessed 2023-06-10]
17. McBain H, Shipley M, Newman S. The impact of self-monitoring in chronic illness on healthcare utilisation: a systematic review of reviews. BMC Health Serv Res 2015 Dec 18;15:565. [doi: 10.1186/s12913-015-1221-5] [Medline: 26684011]
18. Huygens MWJ, Swinkels ICS, de Jong JD, et al. Self-monitoring of health data by patients with a chronic disease: does disease controllability matter? BMC Fam Pract 2017 Mar 20;18(1):40. [doi: 10.1186/s12875-017-0615-3] [Medline: 28320330]
19. Paré G, Jaana M, Sicotte C. Systematic review of home telemonitoring for chronic diseases: the evidence base. J Am Med Inform Assoc 2007;14(3):269-277. [doi: 10.1197/jamia.M2270] [Medline: 17329725]
20. Chester JG, Rudolph JL. Vital signs in older patients: age-related changes. J Am Med Dir Assoc 2011 Jun;12(5):337-343. [doi: 10.1016/j.jamda.2010.04.009] [Medline: 21450180]
21. Teo J. Polyclinics roll out telehealth service to track blood pressure. SingHealth. 2020. URL: https://www.singhealth.com.sg/news/tomorrows-medicine/polyclinics-roll-out-telehealth-service-to-track-blood-pressure [accessed 2022-09-17]
22. Ambulatory blood pressure watch monitoring - conditions & treatments. SingHealth. 2021. URL: https://www.singhealth.com.sg/patient-care/conditions-treatments/ambulatory-blood-pressure-watch-monitoring [accessed 2023-05-30]
23. Primary Tech-Enhanced Care (PTEC). National University Polyclinics. URL: https://www.nup.com.sg/Pages/For%20Patients%20and%20Caregivers/PTEC.aspx [accessed 2023-05-30]
24. Continuous glucose monitoring: a useful tool for diabetes management in primary care. SingHealth. 2023. URL: https://www.singhealth.com.sg/news/defining-med/Continuous-Glucose-Monitoring [accessed 2023-05-30]
25. Flash glucose monitoring. National University Hospital. URL: https://www.nuh.com.sg/our-services/Specialties/Medicine/Endocrinology/Pages/flash-glucose-monitoring.aspx [accessed 2023-05-30]
26. Korkmaz Yaylagul N, Kirisik H, Bernardo J, et al. Trends in telecare use among community-dwelling older adults: a scoping review. Int J Environ Res Public Health 2022 Dec 12;19(24):36554553. [doi: 10.3390/ijerph192416672] [Medline: 36554553]
27. Zhang Y, Lee EWJ, Teo WP. Health-seeking behavior and its associated technology use: interview study among community-dwelling older adults. JMIR Aging 2023 May 4;6:e43709. [doi: 10.2196/43709] [Medline: 36996003]
28. Goldberg EM, Lin MP, Burke LG, Jiménez FN, Davoodi NM, Merchant RC. Perspectives on telehealth for older adults during the COVID-19 pandemic using the quadruple aim: interviews with 48 physicians. BMC Geriatr 2022 Mar 8;22(1):188. [doi: 10.1186/s12877-022-02860-8] [Medline: 35260091]

XSL•FO
RenderX

29. Courtney KL, Lingler JH, Mecca LP, et al. Older adults' and case managers' initial impressions of community-based telehealth kiosks. Res Gerontol Nurs 2010 Oct;3(4):235-239. [doi: 10.3928/19404921-20100504-03] [Medline: 20509594]
30. Zhang Z, Henley T, Schiaffino M, et al. Older adults' perceptions of community-based telehealth wellness programs: a qualitative study. Inform Health Soc Care 2022 Oct 2;47(4):361-372. [doi: 10.1080/17538157.2021.2006198] [Medline: 34822311]
31. Schiaffino MK, Zhang Z, Sachs D, Migliaccio J, Huh-Yoo J. Predictors of retention for community-based telehealth programs: a study of the Telehealth Intervention Program for Seniors (TIPS). AMIA Annu Symp Proc 2021;2021:1089-1098. [Medline: 35308972]
32. Hsieh HL, Tsai CH, Chih WH, Lin HH. Factors affecting success of an integrated community-based telehealth system. Technol Health Care 2015;23 Suppl 2:S189-S196. [doi: 10.3233/THC-150953] [Medline: 26410483]
33. Hovey L, Kaylor MB, Alwan M, Resnick HE. Community-based telemonitoring for hypertension management: practical challenges and potential solutions. Telemed J E Health 2011 Oct;17(8):645-651. [doi: 10.1089/tmj.2011.0027] [Medline: 21819242]
34. West R, Michie S. A brief introduction to the COM-b model of behaviour and the PRIME theory of motivation. Qeios Preprint posted online on Apr 7, 2020 [FREE Full text] [doi: 10.32388/WW04E6.2]
35. Mass screening swab booth. Esco Aster. URL: https://escoaster.com/Products/COVID-19/mass-screening-swab-booth-mssb [accessed 2021-08-22]
36. Smith JD, Li DH, Rafferty MR. The implementation research logic model: a method for planning, executing, reporting, and synthesizing implementation projects. Implement Sci 2020 Sep 25;15(1):84. [doi: 10.1186/s13012-020-01041-8] [Medline: 32988389]
37. Zhang Y, Leuk JSP, Teo WP. Domains, feasibility, effectiveness, cost, and acceptability of telehealth in aging care: scoping review of systematic reviews. JMIR Aging 2023 Apr 18;6:e40460. [doi: 10.2196/40460] [Medline: 37071459]
38. Powell BJ, Waltz TJ, Chinman MJ, et al. A refined compilation of implementation strategies: results from the Expert Recommendations for Implementing Change (ERIC) project. Implement Sci 2015 Feb 12;10:21. [doi: 10.1186/s13012-015-0209-1] [Medline: 25889199]
39. Glasgow RE, Vogt TM, Boles SM. Evaluating the public health impact of health promotion interventions: the RE-AIM framework. Am J Public Health 1999 Sep;89(9):1322-1327. [doi: 10.2105/ajph.89.9.1322] [Medline: 10474547]
40. Glasgow RE, Harden SM, Gaglio B, et al. RE-AIM planning and evaluation framework: adapting to new science and practice with a 20-year review. Front Public Health 2019;7:64. [doi: 10.3389/fpubh.2019.00064] [Medline: 30984733]
41. Proctor E, Silmere H, Raghavan R, et al. Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda. Adm Policy Ment Health 2011 Mar;38(2):65-76. [doi: 10.1007/s10488-010-0319-7] [Medline: 20957426]
42. Parmanto B, Lewis AN Jr, Graham KM, Bertolet MH. Development of the Telehealth Usability Questionnaire (TUQ). Int J Telerehabil 2016;8(1):3-10. [doi: 10.5195/ijt.2016.6196] [Medline: 27563386]
43. Gale RC, Wu J, Erhardt T, et al. Comparison of rapid vs in-depth qualitative analytic methods from a process evaluation of academic detailing in the Veterans Health Administration. Implement Sci 2019 Feb 1;14(1):11. [doi: 10.1186/s13012-019-0853-y] [Medline: 30709368]
44. Graduate college - community engagement project. Nanyang Technological University, Singapore. URL: https://www.ntu.edu.sg/graduate-college/student-life/postgraduate-professional-development/graduate-college---community-engagement-project#Content_C024_Col00 [accessed 2023-06-03]
45. COVID-19 pandemic in Singapore - wikipedia. Wikipedia. URL: https://en.wikipedia.org/wiki/COVID-19_pandemic_in_Singapore [accessed 2023-01-26]
46. 2020–21 Singapore circuit breaker measures. Wikipedia. 2024. URL: https://en.wikipedia.org/wiki/2020%E2%80%9321_Singapore_circuit_breaker_measures [accessed 2025-05-06]
47. Resuming our transition towards covid resilience. Ministry of Health Singapore. 2021. URL: https://www.moh.gov.sg/news-highlights/details/resuming-our-transition-towards-covid-resilience [accessed 2023-01-29]
48. Nurjono M, Shrestha P, Ang IYH, et al. Shifting care from hospital to community, a strategy to integrate care in Singapore: process evaluation of implementation fidelity. BMC Health Serv Res 2020 May 24;20(1):452. [doi: 10.1186/s12913-020-05263-w] [Medline: 32448283]
49. He AJ, Tang VFY. Integration of health services for the elderly in Asia: a scoping review of Hong Kong, Singapore, Malaysia, Indonesia. Health Policy 2021 Mar;125(3):351-362. [doi: 10.1016/j.healthpol.2020.12.020] [Medline: 33422336]
50. Loong PLH. PM Lee Hsien Loong at the 'Singapore Ageing: Issues and Challenges Ahead' book launch. Prime Minister's Office Singapore. 2023. URL: https://www.pmo.gov.sg/Newsroom/PM-Lee-Hsien-Loong-at-the-Singapore-Ageing-Issues-and-Challenges-Ahead-Book-Launch [accessed 2023-06-15]
51. Eliasen B, Teare H, Vrihenhoek T. Shifting to preventive care in a new health system. HIMSS. 2021. URL: https://www.himss.org/resources/shifting-preventive-care-new-health-system [accessed 2023-08-09]
52. Three ways digital transformation is accelerating the shift to preventive care. Philips. 2022. URL: https://www.philips.com/a-w/about/news/archive/features/2022/20220203-three-ways-digital-transformation-is-accelerating-the-shift-to-preventive-care.html [accessed 2023-08-09]

XSL•FO
RenderX

53. High KP. Infrastructure and resources for an aging population: embracing complexity in translational research. Transl Res 2014 May;163(5):446-455. [doi: 10.1016/j.trsl.2013.09.001] [Medline: 24095640]
54. Farrugia G. Healthcare infrastructure needs updating for the digital age. World Economic Forum. 2023. URL: https://www.weforum.org/agenda/2023/01/healthcare-digital-transformation-infrastructure-davos23/ [accessed 2023-08-13]
55. Zalizan T. Next healthcare reform involves helping elderly grow old in community, not assuming all will get frail: ong ye kung - TODAY. TodayOnline. 2022. URL: https://www.todayonline.com/singapore/healthcare-reform-elderly-community-frail-ong-ye-kung-2011986 [accessed 2023-08-13]
56. Ministry of Health Singapore, Ministerial Committee on Ageing. Living life to the fullest: 2023 action plan for successful ageing. 2023.
57. Matsuda S, Yamamoto M. Long-term care insurance and integrated care for the aged in Japan. Int J Integr Care 2001;1:e28. [doi: 10.5334/ijic.39] [Medline: 16896410]
58. He AJ, Chou KL. Long-term care service needs and planning for the future: a study of middle-aged and older adults in Hong Kong. Ageing Soc 2019 Feb;39(2):221-253 [FREE Full text] [doi: 10.1017/S0144686X17000824]
59. Jing T, Zhao X, Xing H, Bao C, Zhan L. The influence of strong social ties on the choice of long-term care model for middle-aged and older adults in China. Front Public Health 2023;11:1112422. [doi: 10.3389/fpubh.2023.1112422] [Medline: 37250073]
60. Wang S, Bolling K, Mao W, et al. Technology to support aging in place: older adults' perspectives. Healthcare (Basel) 2019 Apr 10;7(2):60. [doi: 10.3390/healthcare7020060] [Medline: 30974780]
61. Ollevier A, Aguiar G, Palomino M, Simpelaere IS. How can technology support ageing in place in healthy older adults? A systematic review. Public Health Rev 2020 Nov 23;41(1):26. [doi: 10.1186/s40985-020-00143-4] [Medline: 33292707]
62. Sumner J, Chong LS, Bundele A, Wei Lim Y. Co-designing technology for aging in place: a systematic review. Gerontologist 2021 Sep 13;61(7):e395-e409. [doi: 10.1093/geront/gnaa064] [Medline: 32506136]
63. Point-of-care testing. Testing. 2021. URL: https://www.testing.com/articles/point-of-care-testing/ [accessed 2023-06-13]
64. World Health Organization. Health promotion glossary. 1998. [doi: 10.1093/heapro/1.1.113]
65. Majumder S, Mondal T, Deen MJ. Wearable sensors for remote health monitoring. Sensors (Basel) 2017 Jan 12;17(1):130. [doi: 10.3390/s17010130] [Medline: 28085085]
66. Gandy K, Schmaderer M, Szema A, et al. Remote patient monitoring: a promising digital health frontier. 2021 Presented at: 2021 International Conference on Information and Digital Technologies (IDT) p. 297-302. [doi: 10.1109/IDT52577.2021.9497599]
67. Digital technologies: a new determinant of health. Lancet Digit Health 2021 Nov;3(11):e684. [doi: 10.1016/S2589-7500(21)00238-7] [Medline: 34711372]
68. Governing Health Futures 2030 Commission. Policy brief: the digital determinants of health. 2021.
69. Yao R, Zhang W, Evans R, Cao G, Rui T, Shen L. Inequities in health care services caused by the adoption of digital health technologies: scoping review. J Med Internet Res 2022 Mar 21;24(3):e34144. [doi: 10.2196/34144] [Medline: 35311682]
70. Richardson S, Lawrence K, Schoenthaler AM, Mann D. A framework for digital health equity. NPJ Digit Med 2022 Aug 18;5(1):119. [doi: 10.1038/s41746-022-00663-0] [Medline: 35982146]
71. Ladin K, Porteny T, Perugini JM, et al. Perceptions of telehealth vs in-person visits among older adults with advanced kidney disease, care partners, and clinicians. JAMA Netw Open 2021 Dec 1;4(12):e2137193. [doi: 10.1001/jamanetworkopen.2021.37193] [Medline: 34870680]
72. Shiraz F, Hildon ZLJ, Vrijhoef HJM. Exploring the perceptions of the ageing experience in Singaporean older adults: a qualitative study. J Cross Cult Gerontol 2020 Dec;35(4):389-408. [doi: 10.1007/s10823-020-09414-8] [Medline: 33034796]
73. Guide to healthcare data integration (best practices). True North. URL: https://www.truenorthitg.com/healthcare-data-integration/ [accessed 2023-06-12]
74. Dankar FK. Practices and challenges in clinical data sharing. arXiv. Preprint posted online on Mar 17, 2023. [doi: 10.48550/arXiv.2304.06509]

Abbreviations

CEP: Community Engagement Project
COM-B: Capability, Opportunity, Motivation-Behavior
CTS: Community Telehealth Service
ERIC: Expert Recommendations for Implementing Change compilation
IRLM: Implementation Research Logic Model
NCD: noncommunicable disease
RE-AIM: Reach, Effectiveness, Adoption, Implementation, and Maintenance
SEM: social ecological model

XSL•FO
RenderX

Edited by A Mavragani; submitted 30.01.24; peer-reviewed by E Perryman, FVD Ouderaa; revised version received 06.05.25; accepted 28.05.25; published 30.10.25.

Please cite as:
Zhang Y, Law MCY, Wee SK, Lee BST, Chew BLA, Teo WP, Lee EWJ
Design, Implementation, and Evaluation of a Community-Based Phygital Telemonitoring Program for Older Adults: Multisite Retrospective Pilot Study in Singapore
JMIR Form Res 2025;9:e56905
URL: https://formative.jmir.org/2025/1/e56905
doi:10.2196/56905

©Yichi Zhang, Michelle Cheok Yien Law, Soon Keong Wee, Benjamin Sian Teck Lee, Bing Liang Alvin Chew, Wei Peng Teo, Edmund W J Lee. Originally published in JMIR Formative Research (https://formative.jmir.org), 30.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Feasibility of a Text Message–Based Alcohol Prevention Intervention for Parents of Rising Middle School Students: Randomized Controlled Trial

Marina Stranieri Pearsall[1], MPH, RDN, LDN; Melissa B Gilkey[2], PhD; Susan T Ennett[2], PhD; H Luz McNaughton Reyes[2], PhD; Nisha Gottfredson O'Shea[1,2], PhD

[1]Center for Behavioral Health and Well-being Research, RTI International, 3040 Cornwallis Rd, Durham, NC, United States

[2]Department of Health Behavior, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States

Corresponding Author:
Marina Stranieri Pearsall, MPH, RDN, LDN
Center for Behavioral Health and Well-being Research, RTI International, 3040 Cornwallis Rd, Durham, NC, United States

Abstract

Background: Early-onset alcohol use (EOAU), or drinking before the age of 14 years, is a serious but highly preventable risk factor for later alcohol use. EOAU often begins at home, with sips of alcohol provided by parents. Few scalable interventions are available to engage parents in EOAU prevention.

Objective: This study aimed to evaluate the feasibility of Better-Informed Parents Keeping Adolescents Safe From Alcohol (BIPAS Alcohol), a digital family-based intervention for parents of rising middle schoolers.

Methods: In 2023‑2024, we delivered BIPAS Alcohol to US parents (N=132) of 10- to 12-year-old children. The intervention consisted of a 3-month SMS text messaging curriculum and multimedia website. Guided by Bowen and colleagues' framework, we surveyed parents to evaluate our intervention on its feasibility, including acceptability, integration, demand, and adaptation. We interviewed a subset of parents (n=11) to probe survey findings.

Results: Parents rated BIPAS Alcohol highly on acceptability, with almost all agreeing the intervention kept their attention (117/123, 95.1%), offered useful information (121/123, 98.4%), and helped reduce chances of underage drinking (119/123, 96.7%). Most parents indicated plans to integrate the intervention into family life by referring to content in the future (113/123, 91.9%) or sharing content with others (107/123, 87.0%). In interviews, parents expressed high demand for SMS text messages, due to their short, "digestible" format, while finding the website more cumbersome. Although we designed the SMS text message curriculum for adults, some parents reported adapting the intervention by sharing texts with their children.

Conclusions: Our digital family-based intervention demonstrated feasibility and warrants additional evaluation in a larger-scale trial with a wider audience.

Trial Registration: ClinicalTrials.gov NCT05520333; https://clinicaltrials.gov/study/NCT05520333

(JMIR Form Res 2025;9:e72823) doi:10.2196/72823

KEYWORDS

adolescent; adolescent behavior; alcohol drinking; communication; mobile applications; prevention

Introduction

Early-onset alcohol use (EOAU), or drinking before the age of 14 years, is a serious but highly preventable risk factor for later alcohol abuse that often begins at home, with sips of alcohol provided by parents [1-3]. By age 11, nearly one-quarter of US children have already sipped alcohol [4]. Children who engage in EOAU escalate from initiation to binge drinking and alcohol use disorder faster and more frequently than those who do not [5-10] and are at greater risk for drinking with peers, binge drinking, and alcohol-related harms [5-11]. Timing early-onset alcohol prevention before the transition to middle school is crucial because many students have already tried alcohol by the end of middle school and into high school.

Parents of rising middle schoolers are well-positioned to prevent EOAU. Parents are the primary source of alcohol-related information and socialization for children during the transition to middle school [12], are more influential than peers for children and young adolescents [13], and serve as children's primary source of social modeling and information about alcohol use in the preteen years [14,15]. Research shows that many parents who provide alcohol have the mistaken belief that providing sips of alcohol teaches children how to drink in moderation; other parents report that they have not considered the risks [1].

XSL•FO
RenderX

Despite this evidence, existing universal prevention programs tend to be school-based or too late to prevent EOAU [16,17]. In addition, family-based interventions face the challenge of reaching parents and keeping them engaged. In-person family programs face dissemination and implementation barriers that limit reach at a broader level [18], such as difficulty obtaining transportation and childcare, busy schedules, and perceived stigma [19].

Programs are often not delivered in a mode that accommodates parents' limited time. Existing programs, such as Strengthening Families Program and Talk, They Hear You, also incorporate parent-child communication skills and setting limits or rules with children [20,21], and Talk, They Hear You offers an array of resources online and in an app-based format for parents [21]. However, these programs are not designed to deliver small bites of information to accommodate busy schedules. An SMS text message–based intervention has the potential to address the need for a universal prevention strategy focused on families by overcoming many of the obstacles associated with traditional in-person prevention programs.

To address these gaps, we sought to develop and evaluate the feasibility of Better-Informed Parents Keeping Adolescents Safe From Alcohol (BIPAS Alcohol), a mobile, universal, family-based preventative pilot intervention designed to prevent early drinking onset that is delivered directly to parents of rising middle schoolers. The aims of this study were to assess acceptability, demand, integration, adaptation, and implementation according to the feasibility framework outlined by Bowen and colleagues [22]. By assessing the feasibility of this pilot intervention, this study lays the groundwork for larger-scale research to test the effectiveness of digital EOAU prevention interventions for families of adolescents.

Methods

Participants

Participants were parents of rising middle school children residing primarily in the Research Triangle area of North Carolina. Parents were eligible to participate if they: identified as a parent or guardian of a 10- to 12-year-old child with whom they lived at least 2 or more days a week; used any type of smartphone or other device with SMS text messaging capabilities; were able to complete study activities in English; and were willing to be randomized to receive the intervention immediately or after a 3-month waiting period. Parents who had more than one age-eligible child were asked to select the youngest child to be the primary focus of their study activities.

We recruited parents via electronic school newsletters delivered to parents of public school elementary students in Durham, Chatham, Charlotte-Mecklenburg, and Orange counties, participant referrals, social media posts, information disseminated through university listservs, social media posts, and local pediatric clinics. Among participants who reported their recruitment method, the majority were recruited through electronic school newsletters (69/132, 52.3%), participant referrals (29/132, 22.0%), and university listservs (22/132, 17.0%). Study advertisements instructed interested individuals to scan a QR code that directed them to an online prescreening questionnaire or to contact the study team via email. Research staff reviewed questionnaires and scheduled a phone call with parent-child dyads for further screening. During this phone call, research staff provided study details, confirmed eligibility, and described next steps for enrollment. Parents who wished to enroll completed an electronic consent form and a child permission form.

Recruitment occurred between May and November 2023. Enrolled participants were randomized with equal probability to receive the intervention immediately (n=69) or after a 3-month waiting period (n=63). Parents received the intervention between June 2023 and July 2024.

BIPAS Alcohol Intervention

All parents who enrolled in our study received the BIPAS Alcohol intervention, which consisted of a 3-month SMS text message–based curriculum and a corresponding website. The study team developed content for BIPAS Alcohol from an existing evidence-based intervention entitled Mysteries, Max and Me (MMM) [23]. MMM is a family-based intervention that draws from social learning theory [24] to engage families in positive social modeling for youth alcohol prevention through print-based content delivered to parents of children in elementary school [23]. Children of parents who were randomized to receive MMM had less susceptibility to alcohol use at the 12-month follow-up (Cohen d=0.12). Mediation analyses showed that MMM reduced child susceptibility to alcohol use by increasing parental communication about family norms (Cohen d=0.16), peer influence (Cohen d=0.12), and media portrayals of alcohol (Cohen d=0.27), and by increasing alcohol-specific rule-setting (Cohen d=0.16) [1].

MMM delivered study materials in printed booklets mailed to the homes of participating families. Our study team adapted MMM for use on a digital platform, including 50 brief SMS text messages containing comic-style graphics (Figure 1) and linked website content with video animations tailored to parents of rising middle school–aged children. Video animations depicted examples of conversations about alcohol in common situations, including a parent talking with their child about alcohol before dropping them off at a friend's house (Figure 2), noticing someone visibly drunk at a family gathering, and watching a movie with depictions of young people drinking. Our team also targeted messages to 2 parent groups for universal delivery: parents identifying as drinkers and nondrinkers.

XSL•FO
RenderX

Figure 1. Examples of BIPAS Alcohol SMS text message graphics by theme.

XSL•FO
RenderX

Figure 2. Example of BIPAS Alcohol video animation from the website.

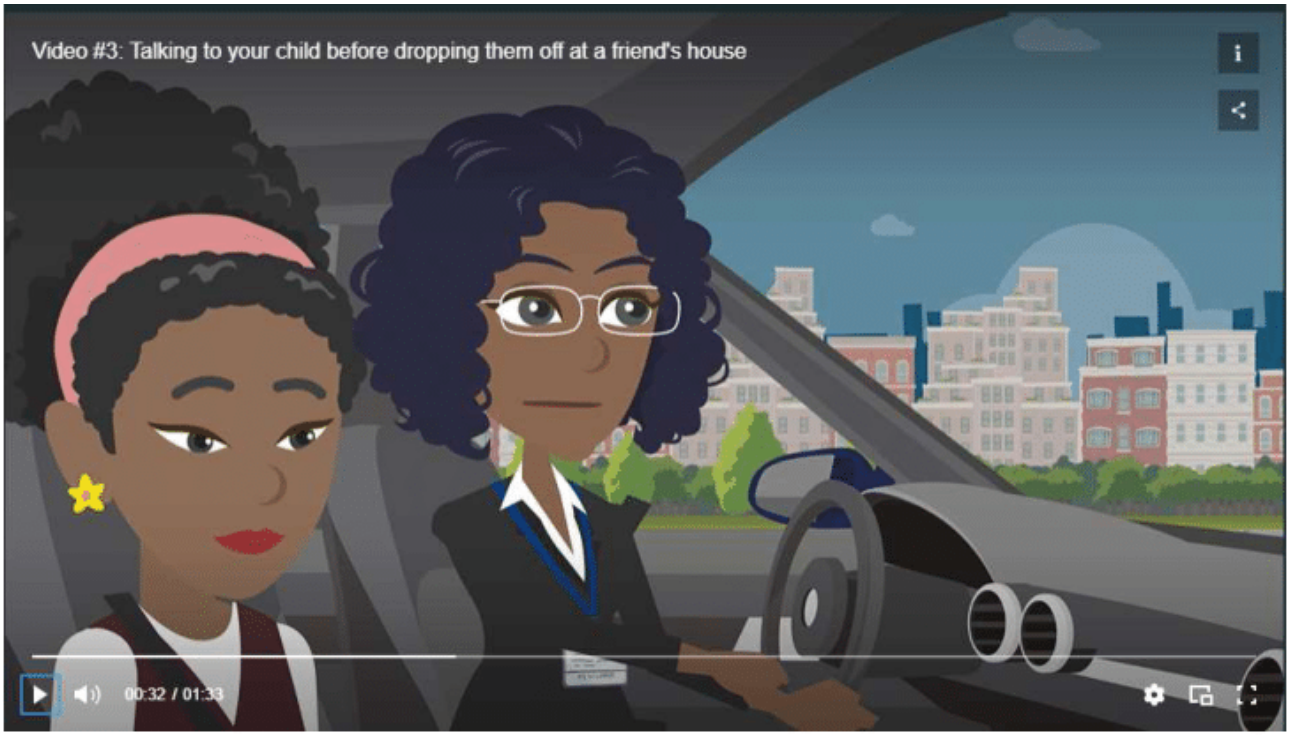

Content was designed to be engaging, informative, and visually appealing. BIPAS Alcohol's core message encouraged parents not to allow their children to sip alcohol by fostering discussions and family rules about alcohol. Message content proceeded along the following progression of themes: (1) evidence of harms caused by EOAU and parent-supplied sips of alcohol; (2) parents as role models; (3) skills and strategies for alcohol-specific communication, including setting clear, alcohol-specific rules with children; (4) role of peers and refusal strategies; (5) alcohol in the media; and (6) developing a family agreement about alcohol. Consistent with these themes, messages focused on what the research says about alcohol and why kids drink (20 messages), sipping (8 messages), media (7 messages), friends and peers (7 messages), and family rules (8 messages). We designed messages to be accessible for parents of lower literacy, with more than half (31/50, 62%) of SMS text messages including a graphic. Messages had a reading ease score of 71.8 and grade level of 6.3 (Flesch-Kincaid), and visual depictions of families in animations and graphics were purposefully designed to be diverse with regard to race and ethnicity and family structure. Two groups of messages were tailored to parents who use and do not use alcohol to appeal to diverse audiences.

During the development process, we assessed SMS text messages iteratively with cognitive interviews with 11 parents of rising middle school children who did not subsequently participate in the intervention. These parents provided feedback on the content, appearance, tone, acceptability, and persuasiveness of our messages.

The intervention website included a homepage with an animated welcome video and individual pages corresponding to intervention themes ("Sipping Alcohol," "Talking the Talk," "Friends," "Media," and "Family Agreement"). The website also featured an "About" page with information about our study team, "FAQs" anticipating parent questions about parent alcohol use, sipping, alcohol use in cooking, and religious use, and "Resources" with resources for adults struggling with alcohol. Four animations on the website ranged from 79 seconds to 104 seconds in length. Parents were directed to specific pages on the website through SMS text message links throughout the intervention.

Data Collection Procedures

Study Design

After enrollment, participants were randomized using a random number generator with equal probability to an immediate intervention (n=69) or a waitlist control group (n=63). Participants and study team members were not blinded to the study assignment groups, as it was clear which group participants were placed in once enrolled. This study will draw on feasibility data (measures described below) from both groups of participants during their active intervention (n=132). Outcomes between groups will be compared elsewhere.

Parent Surveys

Parents completed web-based surveys assessing parents' attitudes, beliefs, and behaviors at baseline and 3 months (intervention completion). We texted parents with invitations to complete the surveys, followed by up to 2 SMS text messages or email reminders for nonrespondents. Parents received a US $25 gift card incentive after completion of each survey. Survey response rates were 100% at baseline (N=132) and 93.2% at 3 months (n=123).

Parent Interviews

Parents indicated in their 3-month survey whether they were interested in participating in qualitative interviews, and study staff reached out to the first 17 interested parents. Eleven parents

XSL•FO
RenderX

completed interviews, which ranged from 32 to 55 minutes in length. A trained member of the study team conducted interviews using a semistructured interview guide designed to explore parents' perspectives about the intervention and their experiences participating. We audio-recorded and transcribed interviews with permission. Parents received a US $40 gift card incentive for completing an interview.

Measures

We developed our measures using Bowen and colleagues' feasibility framework, which emphasizes the importance of intervention acceptability, demand, integration, adaptation, and implementation [22]. In the baseline and 3-month surveys, feasibility was assessed with survey items using a 4-point response scale ranging from "1" (strongly agree) to "4" (strongly disagree). The definition of our measures and their operationalization is detailed in Table 1.

Table . Operationalization of feasibility measures.

Feasibility measures	Definition	Operationalization
Acceptability	Participants' satisfaction with BIPAS Alcohol[a].	• Six survey items assessed BIPAS Alcohol was useful, helpful, kept parents' attention, had an appropriate length, will help reduce child's chances of underage drinking, and the extent to which parents knew what BIPAS Alcohol was trying to do.[b] • Qualitative interviews assessed parents' perspectives on content, SMS text messages, and the website.
Demand	Parents' actual use of BIPAS Alcohol.	• Website analytics detailed the number of visits per page and average time spent per page. • Qualitative interviews assessed materials that parents did or did not use.
Integration	Extent to which BIPAS Alcohol integrated into the existing familial system.	• Two survey items assessed whether parents planned to refer to the content in the future[b] or plan to share it with others.[c] • Qualitative interviews assessed behavior changes, parents' parenting or planned parenting changes, and how parents involved their children in the intervention.
Adaptation	Extent to which changes were made to BIPAS Alcohol by study participants.	• Qualitative interviews assessed whether parents made modifications to the program or engaged with program materials in a way not anticipated by the study team. Interviews also assessed how parents shared program materials with other adults or children not directly targeted by the intervention.
Implementation	Extent to which BIPAS Alcohol can successfully be delivered to the intended audience.	• Study retention rate and study team documentation of technical problems or recruitment issues over the course of delivering the intervention, including the enrollment rate and recruitment rate.

[a]BIPAS: Better-Informed Parents Keeping Adolescents Safe From Alcohol.

[b]Survey response scale ranged from "1" (strongly agree) to "4" (strongly disagree).

[c]Survey response scale ranged from "1" (very likely) to "4" (not at all likely).

Data Analysis

We analyzed quantitative survey data using descriptive statistics and qualitative interview data via thematic content analysis [25]. Two authors (MSP and MBG) developed a codebook based on feasibility measures. After independently applying the codes to 2 transcripts, the authors met and discussed the codes for each response. Discrepancies were resolved through discussion, and the codebook was refined. Afterward, 1 author (MSP) coded all 11 transcripts and analyzed each code to determine themes.

Ethical Considerations

The study protocol was approved by the University of North Carolina at Chapel Hill Institutional Review Board (study #22 - 0101). The study protocol and results are available on ClinicalTrials.gov (ID #NCT05520333). Written informed

XSL•FO
RenderX

consent was obtained electronically through Qualtrics from all parents before participation. Participant privacy and confidentiality were rigorously maintained throughout the study with strict adherence to a data safety and monitoring plan. All data were deidentified and stored on secure, password-protected servers with access restricted to institutional review board–approved study personnel. Participants received a US $25 electronic gift card after completion of each study survey and a US $40 gift card for completing an interview. No identifiable information is presented in this manuscript.

Results

Participant Characteristics

Most parents in our sample were mothers (115/132, 87.1%), identified as non-Hispanic White (111/132, 84.1%), and had high educational attainment (90/132, 68.2% reported a master's degree or above, Table 2). The subset of parents who participated in qualitative interviews had similar characteristics. Our demographic sample reflects the geographic area as well as methods of recruitment (eg, university listserv). Of 243 parents screened, 227 parents were eligible, and 132 parents (58.1%) consented and enrolled (Figure 3). While we do not have direct feedback from parents who did not enroll, coordinating a time to talk to a study team member with their child present was likely the biggest hurdle to enrollment in the screening process.

Table . Parent participant characteristics.

Characteristics	Survey (n=132), n (%)	Interview (n=11), n (%)
Relationship to child		
Mother	115 (87.1)	10 (90.9)
Father	16 (12.1)	1 (9.1)
Guardian	1 (0.8)	0 (0)
Age (years)		
≤40	37 (28.0)	4 (36.4)
41 - 45	48 (36.4)	2 (18.2)
46 and older	47 (35.6)	5 (45.5)
Race and ethnicity		
Non-Hispanic White	111 (84.1)	11 (100)
Asian	8 (6.1)	0 (0)
Black or African American	3 (2.3)	0 (0)
Hispanic or Latino	2 (1.5)	0 (0)
More than 1 race or another race	8 (6.1)	0 (0)
Highest degree earned		
Master's, doctorate, or professional degree	90 (68.2)	7 (63.6)
Bachelor's, associate's, or technical degree	35 (26.5)	3 (27.3)
High school or less than high school	7 (5.3)	1 (9.1)
Parent drinks alcohol		
Yes	103 (78.0)	9 (81.8)
No	29 (22.0)	2 (18.2)
Alcohol intake over the past 30 days		
4+ times per week	16 (12.1)	1 (9.1)
2 - 3 times per week	31 (23.5)	2 (18.2)
2 - 4 times per month	29 (22.0)	4 (36.4)
Monthly or less	24 (18.2)	2 (18.2)
Never	32 (24.2)	2 (18.2)

XSL•FO
RenderX

Figure 3. Flow of participants through BIPAS Alcohol intervention.

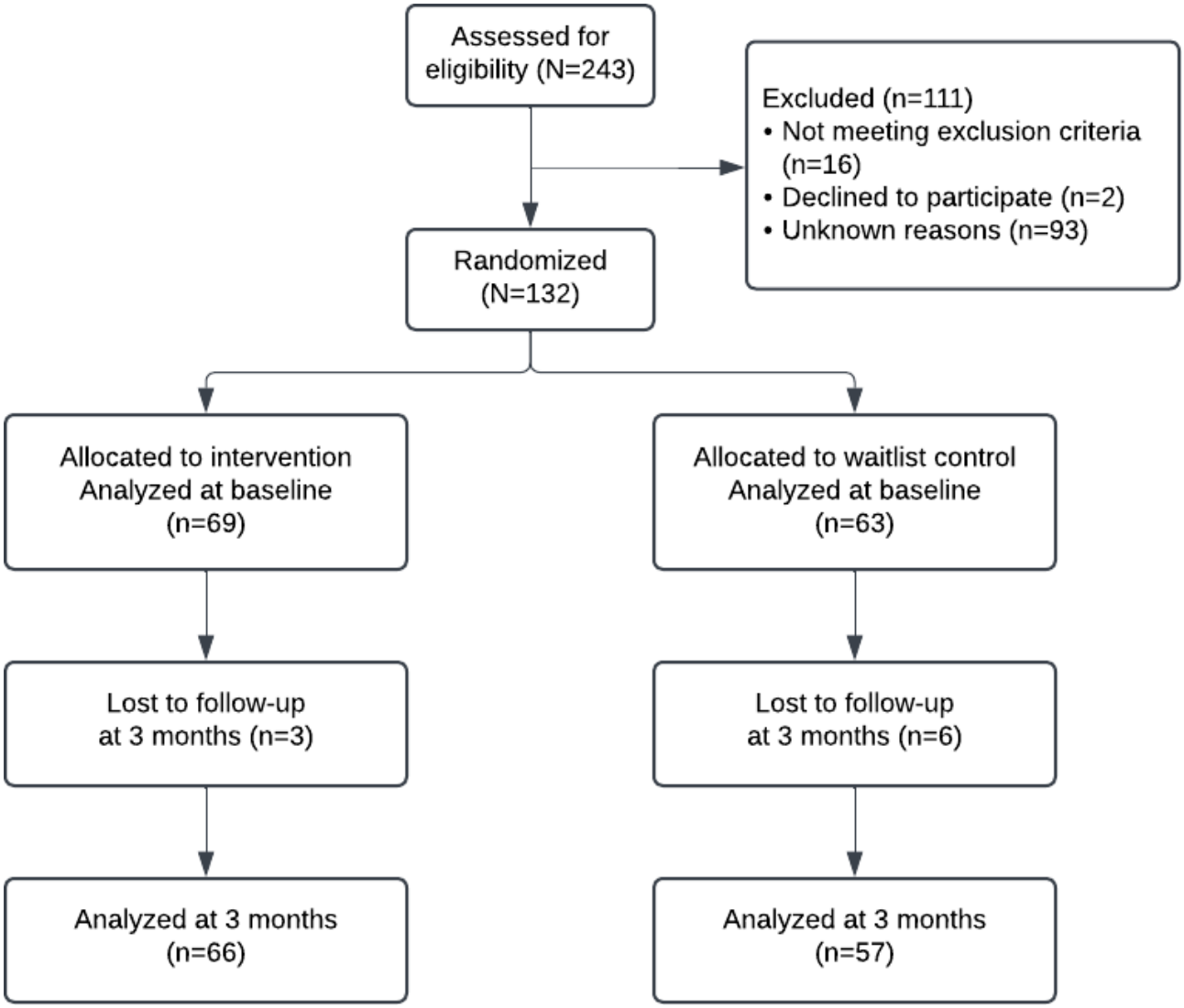

Acceptability

Survey results showed that BIPAS Alcohol was acceptable to parents (Table 3). Parents overwhelmingly agreed that the program kept their attention (117/123, 95.1% of parents strongly agreed or agreed at 3 months) and that they learned useful information (121/123, 98.4% of parents strongly agreed or agreed at 3 months). Notably, 96.7% (119/123) of parents reported that the program would help them reduce their child's chances of underage drinking at 3 months. Parents found the SMS text messages to be the most helpful component of the program, followed by the website and videos.

XSL•FO
RenderX

Table . Parents' perspectives about BIPAS Alcohol at 3 months (n=123).

Parents' perspectives[a]	Strongly disagree or disagree, n (%)	Agree or strongly agree, n (%)
Survey kept attention	6 (4.9)	117 (95.1)
Program was too long	101 (82.1)	22 (17.9)
Learned useful information	2 (1.6)	121 (98.4)
Will help me reduce my child's chances of underage drinking	4 (3.3)	119 (96.7)
Program covered topics that are important to me	0 (0)	123 (100)
Understand what program is trying to do	0 (0)	123 (100)
Time well spent	2 (1.6)	121 (98.4)
Will refer to content in future	10 (8.1)	113 (91.9)
Likely to share content with others	16 (13.0)	107 (87.0)

[a]Survey data of parents from both intervention and waitlist control groups during the active intervention period.

In qualitative interviews, parents discussed that the program was acceptable, useful, convincing, and liked the short, "digestible" format of the program (Table 4). Parents reported that they had not thought about talking with their children about alcohol prior to receiving the BIPAS Alcohol intervention, but reported that the intervention provided both new information and an awareness around the need to talk about alcohol with their child now. One parent remarked that BIPAS Alcohol "opened my eyes to how many influences there are" (eg, social media, movies, advertisements, parents and other caregivers, peers, and parents of peers).

XSL•FO
RenderX

Table . Selected quotes by theme from qualitative interviews.

Themes	Quotes
Parents had positive feelings about the intervention	• It changed my behavior with my daughter in a way that was very easy. That will hopefully make a difference and it's something that I don't know. I know I can go back and read and find information as she gets older and [her] needs change.
The program brought up topics that parents had not previously considered talking to their child about	• I think it gave me an excuse to think concretely about things that I wouldn't have otherwise. • [The program] has given us really a jumping point to talk more about issues I thought were further away in the future for us and I wouldn't have brought it up on my own. So I'm appreciative of that because it's just given us some talking points… • …It just hadn't occurred to me to talk to my daughter who's in elementary school still about alcohol…It was a really good reminder [that], no, actually I can't just assume that she is going to read my mind and know how I feel about these things.
Parents liked that the program was brief	• The format was very short nuggets [of information]…It wasn't hard to digest at all. The message was very short for short attention span people. • I really liked [the text messages] because they were small amounts of snackable information. If they hit me at the right time, they could prompt interesting discussions. • I liked that it held my attention just for chunks at a time. I didn't have to spend a half hour having a conversation when we got the graphic. We could have a quick two or three minute conversation, and I think that really held everybody's interest.
Parents found program content convincing or needed additional information	• After the first part of the intervention, I didn't need to be convinced. I was convinced. I was like, oh, okay, I'm going to make sure of that. • [The program] really opened my eyes to how many influences are out there as regards to alcohol and adolescents. I'm a parent that strongly believes I would much rather my child drink at home where she's safe…the study really opened my eyes just like, oh, hey, that's kind of wrong. You shouldn't do that. • So one of the things I struggle with a bit in the messaging is this sort of hard and fast rule of you just don't drink period.
Parents found the program content helpful in talking about alcohol	• It brought up an opportunity to talk about [alcohol] that wasn't accusatory or in reaction to a negative incident…it just gave a neutral platform to talk about it. • I think the information both gave me the impetus to initiate the conversation and the security or the confidence that the conversation doesn't need to be very elaborate. • I liked [the program] and it sparked a lot of conversation at home, but also with fellow parents. • I think it's been helpful because they're easy conversation starters and it doesn't have to be a special conversation set aside during the day. It can just happen anytime. It felt very fluid and natural.
Parents liked the format of the program	• I particularly [liked] the text messages…a short but sweet punch of information…time is short often, so anything that cuts to the chase of the most pertinent tidbits I think were what spoke to me the most. • I am in favor of anything that gives me a structure to hang things on that I've always known I should do, but never quite known how to approach. And I think this program did that well.
Time constraints prevented parents from reviewing all the program content	• I was less likely to read through a long [website] page than I was a more condensed text message, but it was still useful information.

The short, succinct format of BIPAS Alcohol messages stood out to parents (Table 4). Parents said the program "wasn't hard to digest" and liked the "snackable" messages. Practical examples and talking points helped parents to initiate conversations with children, and parents appreciated ideas about how to approach these conversations. Parents described how helpful it was to think about BIPAS Alcohol's approach of having multiple, shorter conversations with children as opposed

XSL•FO
RenderX

to a "one and done" talk about alcohol. That the information came from a neutral, university-affiliated third party was also cited by parents as enabling trust.

As this program targeted rising middle school-aged children, parents brought up the age appropriateness of the content; some said that the program material was age-appropriate for their child, and others reported that they had initially thought their child was too young, but changed their mind after receiving more information from the BIPAS Alcohol messages. Several parents said that they had not realized how much their child had already been exposed to alcohol or, alternatively, appreciated that they were receiving this information proactively.

Parents expressed conflicting views about both the perceived repetitiveness in BIPAS Alcohol's messaging and the program's stance on avoiding all alcohol for children. This included some mixed feelings about the program's core message around not allowing sipping, including uncertainty that sipping would lead to drinking for children. Although the information was available on the website, some parents requested additional evidence to counteract the cultural narrative that drinking in moderation was healthy, even for underage children. Parents who reported that they did not engage much with BIPAS Alcohol's website and animations found this content too long or time-consuming to access. Most parents did not watch all the animations, and some did not find them relatable to their family or felt they did not have time to watch them. However, parents believed the website was an important component of the intervention, for the legitimacy and credibility of the program and to refer to for additional resources.

The SMS text message portion of the program was many parents' favorite part due to the short, concise format and engaging graphics, and parents reported positive feelings about texts. Parents liked the graphics and highlighted that they helped capture their attention or communicate with their child.

Demand

Website analytics during the active intervention showed that the homepage received the most page views (n=610 views), followed by the "Talking the Talk" page (n=467) and "Sipping Alcohol" page (n=242). The "Resources" page received the least number of views (n=124). Website visitors spent the most time on the "Talking the Talk" page (average 3.3 min), followed by the "Media" (average 2.7 min) and "Sipping" pages (average 2.2 min). Visitors spent the least amount of time on the "About" page (average 1 min). In qualitative interviews, parents reported reading all or most SMS text messages they received. Parents also reported visiting the website, although with greater variance: some visited only a few times during the intervention, while others visited weekly or more often. Parents noted that time constraints made it difficult to frequently follow links to the website or watch animations.

Parents who reported that they did not engage much with BIPAS Alcohol's website and animations found this content too long or time-consuming to access. Most parents did not watch all the animations, and some did not find them relatable to their family or felt they did not have time to watch them. Parents believed the website was an important component of the intervention; however, for the legitimacy and credibility of the program and to refer to for additional resources.

Integration

In the 3-month surveys, 91.9% (113/123) of parents agreed or strongly agreed that they would refer to this content in the future, while 87.0% (107/123) of parents said that they are somewhat likely or very likely to share this content with others.

In qualitative interviews, all parents reported that they had talked to their child about alcohol and discussed family rules about alcohol after participating in BIPAS Alcohol. Two themes that emerged were that the program helped open new conversations with children and enabled multiple, smaller conversations, including prompts to have these types of conversations. The program, according to 1 parent, "gave me the impetus to initiate the conversation and the security or the confidence that the conversation doesn't need to be very elaborate."

Parents appreciated suggested scenarios focused on talking with their children about alcohol and cited that the program helped them have new conversations with their children within existing family life—such as around a family history with alcohol, observations of alcohol in media or movies, and navigating alcohol in gatherings and holidays with family and friends. Two parents described the program as a starting point to broader conversations with their children around other substances. Said 1 parent: "It has given us a jumping point to talk more about issues I thought were further away in the future for us and I wouldn't have brought up on my own."

Parents also discussed making personal behavior changes or parenting practices around alcohol, such as no longer asking their child to bring them drinks, modifying the presence or accessibility of alcohol in their home, or no longer allowing children to have sips of alcohol or mocktails. Several parents described their own mindset changes around alcohol, such as decreasing personal consumption altogether or in front of their child. Parents reported initiating conversations with other adults about alcohol, although some did not, with 1 parent noting that this topic was "touchy." Two parents wondered about the developmental time span through which BIPAS Alcohol's message regarding the importance of children abstaining from alcohol completely would continue to be relevant to their child.

Adaptation

In qualitative interviews, parents discussed ways they adapted the program. Parents shared that they shared program materials directly with their child as part of alcohol-specific conversations, which the study team had not anticipated. Some parents also described incorporating other children not targeted by BIPAS Alcohol in the intervention, either in alcohol-specific discussions or by sharing program materials with their child's sibling or peers. Other parents noted broadening conversations about alcohol to other substances, such as illegal and legal drug use, smoking, vaping, and how to handle alcohol as their child gets older.

Implementation

A retention rate of 93.2% (123/132) was calculated as the percentage of participants who completed the 3-month survey.

XSL•FO
RenderX

The vast majority of SMS text messages were delivered as planned. However, 1 technical issue occurred on the SMS text delivery platform, where SMS text messages were temporarily capped, blocking SMS text message delivery to 31 participants. This issue was discovered and fixed, and parents continued to receive SMS text messages. Parents did not report any technical issues during the intervention.

Suggested Changes to Strengthen the Program

In interviews, parents were invested in the success of future iterations of the program and shared many ideas for strengthening it. Some parents wanted more guidance about how to modify discussions and rules with kids as they got older and were exposed to more alcohol. For example, parents mentioned wanting content geared toward drinking and driving, texting and driving, and preparing for children to leave home or enter college. Noted one parent: “I feel like there does need to be a conversation about how to transition my kid into eventually trying alcohol.” Another parent questioned, “I think that the guidance and the rules that we’ve talked about for a younger age group just aren’t the right rules for a 16-year-old. And so how do you get there?”

Parents shared ideas for broadening the program’s content focus. Parents sought guidance around other substances, such as legal and illegal drugs and tobacco, or topics such as peer pressure and public health context. Similarly, some parents wanted additional information on cultural differences around alcohol, either within different households or across different countries.

Parents suggested additional content that would be helpful, including additional scenarios for graphics, discussion about other alcohol-related themes, such as secrecy or details about where kids may have access to alcohol, and additional information on the science of brain development. Parents were also interested in different options for guidance based on parenting style and more interactive activities for parents and kids. Several parents wanted additional references to justify the program’s recommendations or resources for parents or other family members struggling with alcohol dependency, which were both available on the website but were not accessed by all parents.

Discussion

Principal Findings

Results from the BIPAS Alcohol pilot indicate that our SMS text message–based intervention is a feasible way to engage parents in EOAU prevention for rising middle schoolers. Measures of acceptability were high, and parents shared that they found the intervention appealing, engaging, and helpful. Although some parents were initially concerned about whether the intervention was age-appropriate for their 10- to 12-year-old child, they concluded that the content was appropriate after participating in the intervention. This is a noteworthy finding given evidence of some parents’ reluctance to initiate early conversations about alcohol use with their children [3].

Parents also reported high demand and integration. In particular, parents’ usage of SMS text messages was very high, which corresponds to prior studies showing the usefulness of SMS text messages for intervention delivery [26,27]. SMS text messages offer a promising method of overcoming the limitations of existing interventions, including the original MMM study, to make a more highly scalable intervention. Furthermore, parents reported integration by applying content primarily by initiating conversations about alcohol with their children. This suggests promise for effectiveness, which we will examine in a separate study.

Strengths and Limitations

Our pilot lays the foundation for future research, but findings should be interpreted in light of limitations. The sample was comprised of relatively affluent, white parents within a relatively small geographic area, which limits generalizability. Although our qualitative interviews included a diversity of parents identifying as drinkers and nondrinkers, our sample was not racially or ethnically diverse. Of note, however, white and highly educated adults also have the highest rates of alcohol consumption and permissive sipping behaviors with underaged children [2,28], so this approach resulted in targeting the highest risk groups. Furthermore, over-representation of mothers in parenting interventions is common [29,30] and may reflect gendered parenting differences with respect to which parents spend more time managing communication with children and other caregivers. In addition, we relied on parental reports to assess the demand for SMS text messages and were not able to otherwise determine whether participants read SMS text messages from available data; these and other novel measures will require validation in a future study. Although parents perceived BIPAS Alcohol as effective, we plan to evaluate its effectiveness in a future study. Finally, this study’s follow-up was limited to 3 months; longer-term outcomes and data from children also need to be considered to understand effectiveness in preventing EOAU.

Future Directions

Next steps for BIPAS Alcohol include evaluating this intervention in a larger trial to ensure that it is feasible and efficacious in preventing children from drinking in a larger, more diverse sample. We also plan to expand BIPAS Alcohol’s reach to ensure it is available to parents who prefer Spanish and to incorporate additional content in response to the qualitative feedback noted in this study. Finally, a larger trial will provide the opportunity to explore how the impact of BIPAS varies across subpopulations of parents and according to seasonal changes in alcohol availability due to holidays, summer vacations, and the like.

Conclusions

The BIPAS Alcohol SMS text message–based parenting intervention demonstrated feasibility in a sample of 132 parents. Parents participating in this intervention reported that they found the intervention acceptable, useful, and easy to integrate into their family. Future research is needed to revise the SMS text message content for a larger scale trial and test this intervention for generalizability to a wider audience.

XSL•FO
RenderX

Acknowledgments

This study was funded by the National Institute on Alcohol Abuse and Alcoholism (NIAAA; grant 1R34AA028856-01A1). The funder had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the article; or decision to submit the article for publication.

Data Availability

The datasets generated or analyzed during this study are available in the National Institute on Alcohol Abuse and Alcoholism Data Archive [31].

Authors' Contributions

MSP: conceptualization, data curation, formal analysis, methodology, project administration, investigation, writing-original draft.
MBG: conceptualization, investigation, formal analysis, methodology, supervision, writing-review and editing.
STE: conceptualization, investigation, writing-review and editing.
HLMR: conceptualization, investigation, writing-review and editing.
NGO: conceptualization, investigation, formal analysis, funding acquisition, methodology, supervision, project administration, writing-review and editing.

Conflicts of Interest

None declared.

Checklist 1
CONSORT checklist.
[PDF File, 147 KB - formative_v9i1e72823_app1.pdf]

References

1. Jackson C, Ennett ST, Reyes HLM, et al. Reducing children's susceptibility to alcohol use: effects of a home-based parenting program. Prev Sci 2016 Jul;17(5):615-625. [doi: 10.1007/s11121-016-0659-y] [Medline: 27154767]
2. Jackson C, Ennett ST, Dickinson DM, Bowling JM. Letting children sip: understanding why parents allow alcohol use by elementary school-aged children. Arch Pediatr Adolesc Med 2012 Nov;166(11):1053-1057. [doi: 10.1001/archpediatrics.2012.1198] [Medline: 22986887]
3. Ennett ST, Jackson C, Cole VT, et al. A multidimensional model of mothers' perceptions of parent alcohol socialization and adolescent alcohol misuse. Psychol Addict Behav 2016 Feb;30(1):18-28. [doi: 10.1037/adb0000119] [Medline: 26415053]
4. Miech RA, Johnston L, Patrick M, O'Malley P. Monitoring the future national survey results on drug use, 1975-2023: overview and detailed results for secondary school students. : Institute for Social Research, University of Michigan; 2024 URL: https://monitoringthefuture.org/wp-content/uploads/2024/01/mtfoverview2024.pdf [accessed 2025-10-14]
5. Murphy JG, Campbell KW, Joyner KJ, Dennhardt AA, Martens MP, Borsari B. Trajectories of reward availability moderate the impact of brief alcohol interventions on alcohol severity in heavy-drinking young adults. Alcohol Clin Exp Res 2021 Oct;45(10):2147-2159. [doi: 10.1111/acer.14681] [Medline: 34342015]
6. Aiken A, Clare PJ, Boland VC, et al. Parental supply of sips and whole drinks of alcohol to adolescents and associations with binge drinking and alcohol-related harms: a prospective cohort study. Drug Alcohol Depend 2020 Oct 1;215:108204. [doi: 10.1016/j.drugalcdep.2020.108204] [Medline: 32871506]
7. Clare PJ, Dobbins T, Bruno R, et al. The overall effect of parental supply of alcohol across adolescence on alcohol-related harms in early adulthood-a prospective cohort study. Addiction 2020 Oct;115(10):1833-1843. [doi: 10.1111/add.15005] [Medline: 32034841]
8. Staff J, Maggs JL. Parents allowing drinking is associated with adolescents' heavy alcohol use. Alcohol Clin Exp Res 2020 Jan;44(1):188-195. [doi: 10.1111/acer.14224] [Medline: 31750959]
9. Colder CR, Shyhalla K, Frndak SE. Early alcohol use with parental permission: psychosocial characteristics and drinking in late adolescence. Addict Behav 2018 Jan;76:82-87. [doi: 10.1016/j.addbeh.2017.07.030] [Medline: 28772246]
10. Yap MBH, Cheong TWK, Zaravinos-Tsakos F, Lubman DI, Jorm AF. Modifiable parenting factors associated with adolescent alcohol misuse: a systematic review and meta-analysis of longitudinal studies. Addiction 2017 Jul;112(7):1142-1162. [doi: 10.1111/add.13785] [Medline: 28178373]
11. Ksinan AJ, Ksinan Jiskrova G, Hrežová E, Andrýsková L, Pikhart H, Bobák M. Association between parental supply of alcohol and later adolescent alcohol use in a highly permissive context. J Stud Alcohol Drugs 2023 Jan;84(1):27-36. [doi: 10.15288/jsad.21-00437] [Medline: 36799671]

XSL•FO
RenderX

12. Wood MD, Read JP, Mitchell RE, Brand NH. Do parents still matter? Parent and peer influences on alcohol involvement among recent high school graduates. Psychol Addict Behav 2004 Mar;18(1):19-30. [doi: 10.1037/0893-164X.18.1.19] [Medline: 15008682]
13. Wills TA, Cleary SD. Peer and adolescent substance use among 6th-9th graders: latent growth analyses of influence versus selection mechanisms. Health Psychol 1999 Sep;18(5):453-463. [doi: 10.1037//0278-6133.18.5.453] [Medline: 10519461]
14. Latendresse SJ, Rose RJ, Viken RJ, Pulkkinen L, Kaprio J, Dick DM. Parenting mechanisms in links between parents' and adolescents' alcohol use behaviors. Alcohol Clin Exp Res 2008 Feb;32(2):322-330. [doi: 10.1111/j.1530-0277.2007.00583.x] [Medline: 18162066]
15. Ryan SM, Jorm AF, Lubman DI. Parenting factors associated with reduced adolescent alcohol use: a systematic review of longitudinal studies. Aust N Z J Psychiatry 2010 Sep;44(9):774-783. [doi: 10.1080/00048674.2010.501759] [Medline: 20815663]
16. Spoth R, Greenberg M, Turrisi R. Preventive interventions addressing underage drinking: state of the evidence and steps toward public health impact. Pediatrics 2008 Apr;121 Suppl 4(Suppl 4):S311-S336. [doi: 10.1542/peds.2007-2243E] [Medline: 18381496]
17. Foxcroft DR, Tsertsvadze A. Universal school-based prevention programs for alcohol misuse in young people. Cochrane Database Syst Rev 2011 May 11(5):CD009113. [doi: 10.1002/14651858.CD009113] [Medline: 21563171]
18. Leslie LK, Mehus CJ, Hawkins JD, et al. Primary health care: potential home for family-focused preventive interventions. Am J Prev Med 2016 Oct;51(4 Suppl 2):S106-S118. [doi: 10.1016/j.amepre.2016.05.014] [Medline: 27498167]
19. Hansen A, Broomfield G, Yap MBH. A systematic review of technology‐assisted parenting programs for mental health problems in youth aged 0–18 years: applicability to underserved Australian communities. Aust J Psychol 2019 Dec 1;71(4):433-462. [doi: 10.1111/ajpy.12250]
20. "Talk. They Hear You." Underage drinking campaign. substance abuse and mental health services administration (SAMHSA). URL: https://www.samhsa.gov/substance-use/prevention/talk-they-hear-you [accessed 2025-07-15]
21. Substance Abuse and Mental Health Services Administration (US), Office of the Surgeon General (US). Chapter 3: prevention programs and policies. In: Facing Addiction in America: The Surgeon General's Report on Alcohol, Drugs, and Health [Internet]: US Department of Health and Human Services; 2016:1-61 URL: https://www.ncbi.nlm.nih.gov/books/NBK424850 [accessed 2025-10-14]
22. Bowen DJ, Kreuter M, Spring B, et al. How we design feasibility studies. Am J Prev Med 2009 May;36(5):452-457. [doi: 10.1016/j.amepre.2009.02.002] [Medline: 19362699]
23. Dickinson DM, Hayes KA, Jackson C, Ennett ST, Lawson C. Promoting an alcohol-free childhood: a novel home-based parenting program. Am J Health Educ 2014 Mar 5;45(2):119-128. [doi: 10.1080/19325037.2013.875963] [Medline: 26478761]
24. Bandura A. . URL: Social Foundations of Thought and Action: A Social Cognitive Theory.Prentice-Hall; 1986.
25. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
26. Loescher LJ, Rains SA, Kramer SS, Akers C, Moussa R. A systematic review of interventions to enhance healthy lifestyle behaviors in adolescents delivered via mobile phone text messaging. Am J Health Promot 2018 May;32(4):865-879. [doi: 10.1177/0890117116675785] [Medline: 27923885]
27. Militello LK, Kelly SA, Melnyk BM. Systematic review of text-messaging interventions to promote healthy behaviors in pediatric and adolescent populations: implications for clinical practice and research. Worldviews Evid Based Nurs 2012 Apr;9(2):66-77. [doi: 10.1111/j.1741-6787.2011.00239.x] [Medline: 22268959]
28. Delker E, Brown Q, Hasin DS. Alcohol consumption in demographic subpopulations: an epidemiologic overview. Alcohol Res 2016;38(1):7-15. [Medline: 27159807]
29. Panter-Brick C, Burgess A, Eggerman M, McAllister F, Pruett K, Leckman JF. Practitioner review: engaging fathers--recommendations for a game change in parenting interventions based on a systematic review of the global evidence. J Child Psychol Psychiatry 2014 Nov;55(11):1187-1212. [doi: 10.1111/jcpp.12280] [Medline: 24980187]
30. McKee K, Cabrera N, Alonso A, Turcios M, Reich S. Determinants of fathers' and mothers' involvement in a parenting intervention. Psychology of Men & Masculinities 2021;22(3):521-537. [doi: 10.1037/men0000320]
31. National Institute on Alcohol Abuse and Alcoholism Data Archive (NIAAADA). National Institute of Mental Health Data Archive (NDA). URL: https://nda.nih.gov/niaaa [accessed 2025-10-14]

Abbreviations

Alcohol: Better-Informed Parents Keeping Adolescents Safe From Alcohol
EOAU: early-onset alcohol use
MMM: Mysteries, Max and Me

XSL•FO
RenderX

Edited by A Mavragani, J Sarvestan; submitted 18.02.25; peer-reviewed by A Ray; revised version received 29.09.25; accepted 01.10.25; published 30.10.25.

Please cite as:
Pearsall MS, Gilkey MB, Ennett ST, McNaughton Reyes HL, O'Shea NG
Feasibility of a Text Message–Based Alcohol Prevention Intervention for Parents of Rising Middle School Students: Randomized Controlled Trial
JMIR Form Res 2025;9:e72823
URL: https://formative.jmir.org/2025/1/e72823
doi:10.2196/72823

© Marina Stranieri Pearsall, Melissa B Gilkey, Susan T Ennett, H Luz McNaughton Reyes, Nisha Gottfredson O'Shea. Originally published in JMIR Formative Research (https://formative.jmir.org), 30.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Predicting Risk of Heat-Related Injuries for Individuals Wearing Personal Protective Equipment Using Smartwatches: Feasibility Observational Study

Meghan Hegarty-Craver[1*], PhD; Donna Womack[1*], BS; Jonathan Thornburg[1*], PhD; Timothy Boe[2*], MS; M John Archer[2*], MS; Worth Calfee[2*], PhD

[1]RTI International, Durham, NC, United States

[2]Office of Research and Development, US Environmental Protection Agency, Research Triangle Park, NC, United States

[*]all authors contributed equally

Corresponding Author:
Timothy Boe, MS
Office of Research and Development, US Environmental Protection Agency, Research Triangle Park, NC, United States

Abstract

Background: The risk of developing heat-related illness increases when personal protective equipment (PPE) is worn, especially in hot and humid environments. While cooling strategies are effective, they must be applied preemptively or delivered promptly, which can be difficult if individuals are working in dangerous environments or wearing contaminated PPE. Wearable sensors can be leveraged to continuously monitor health including heart rate, respiration rate, blood oxygen levels, and physical activity.

Objective: This study aims to (1) evaluate the use of wearable sensors for monitoring the real-time health of individuals wearing PPE to mitigate the risk of developing a heat-related illness and enable timely intervention, (2) understand how PPE may affect smartwatch data quality and comfort, and (3) identify circumstances in which people wearing PPE may not be able to wear a smartwatch.

Methods: Individuals participating in planned field trainings or exercises where PPE was being worn were asked to wear Garmin Fenix 6 smartwatch (Garmin Ltd) before, during, and after the event to monitor health and recovery. These convenience cohorts were selected to understand the feasibility of using smartwatches with different types of PPE (ie, level C PPE and firefighter gear) for different types of training (ie, a simulated environmental cleanup exercise and skill and tactical maneuver training for new firefighter recruits).

Results: Two data collections were conducted using the Garmin Fenix 6 smartwatch to assess wearability, data quality, and data accuracy. For the first effort, participants wore the watch for 3.9 - 5.1 days, and wear compliance ranged from 83.8% to 99.9%. For the second effort, participants wore the watch for the exercise only, which was 3.5 hours. Participants were able to wear the watches for the entire time that they were wearing PPE and did not report any adverse events. Changes in heart rate corresponded with changes in physical activity, providing evidence that physiology can be acceptably monitored during physical activity. Heart rate data artifact ranged between 5.8% and 9.3% and was highest for the control participant (second data collection) who was not wearing PPE.

Conclusions: Based on the results obtained from the 8 pilot users, the Garmin Fenix 6 smartwatch is an appropriate choice for continuously monitoring the health of individuals wearing PPE. The watch can be tolerated for extended wear periods and data quality is sufficient for monitoring heart rate and predicting core body temperature.

(JMIR Form Res 2025;9:e72324) doi:10.2196/72324

KEYWORDS

wearable; smartwatch; personal protective equipment; PPE; temperature; heat strain; heat-related illness; continuous monitoring

Introduction

Overview

Heat-related illnesses (HRIs) are common among individuals who must wear personal protective equipment (PPE) to reduce their risk of exposure to chemical, biological, or other threats [1-3]. HRIs can range in severity from heat exhaustion (ie, the individual has an elevated core temperature and symptoms consistent with heat strain) to heat stroke (ie, the individual is no longer able to cool themselves and their temperature climbs dangerously high, causing damage to the central nervous system, which may be fatal) [4]. Heat strain is the body's physiological response to heat stress (ie, the total heat load from physical exertion, the environment, and clothing impacts) [4]. Symptoms

XSL•FO
RenderX

include nausea or loss of appetite, extreme thirst and dehydration, headache, dizziness or confusion, excessive sweating or clammy skin, cramps, fast breathing, and high heart rate (HR) [4].

Wearing PPE while working outdoors or in warm and humid environments increases the risk of developing an HRI [2,3]. Because PPE is designed to reduce exposure to hazards, both heat transfer from the body to the environment (ie, convective heat loss) and cooling through sweating (ie, evaporative heat loss) are disrupted [2]. Core temperature (T_c) rises as the body is unable to cool itself, and HR and sweat rate continue to increase to try and regain thermal equilibrium [4].

Occupational heat strain can significantly impact productivity, increase risk of injury, and have long-lasting health consequences [1,2]. Professional organizations such as the International Association of Fire Fighters (IAFF), which represents professional firefighters and paramedics across the United States and Canada—and government entities such as the US Environmental Protection Agency (EPA), which is responsible for conducting and supervising cleanup of contamination associated with accidental spills and industrial accidents—are especially interested in strategies to reduce HRI risks imposed by PPE. While cooling strategies are effective [1,2,5], they must be applied preemptively, or treatment must be initiated with enough time to prevent progression of symptoms. This can be difficult if individuals are in hazardous environments (eg, firefighters) or wearing contaminated PPE (eg, workers involved in Hazardous Materials Management [HAZMAT]) [6]. It is likely that a multipronged approach including exposure predictions and individual monitoring is needed to mitigate risk [7].

Exposure Prediction Models

Several variables affect potential for heat strain injury, including the environment, clothing worn by the individual, physical activity level, and individual physiological factors [7-9]. The Required Sweat Rate model published in 1981 was developed to improve working conditions in industrial settings by removing the subjectivity of perceived heat stress [10]. The Predicted Heat Strain (PHS) Model [8] improved estimations and extended the original model to include predictions of mean skin temperature and sweat rate based on the clothing being worn, as well as rectal temperature and the distribution of heat storage. The US Army Research Institute of Environmental Medicine (USARIEM) Heat Strain Decision Aid (HSDA) is a more recent, spreadsheet-based tool that takes inputs related to the environment, clothing being worn, physical activity level, and physiology to make generalized predictions related to how long it is safe to work under a set of conditions before there is a risk of experiencing thermal strain [9]. This tool is currently used by the US military for establishing safety guidelines.

Individual Monitoring Using Wearables

Overview

Exposure models work well for the constraints under which they were developed (eg, population demographics, activity pattern, type of PPE, and environmental conditions), but accuracy is compromised outside of these bounds [7]. Workers at risk for HRIs are currently instructed how to monitor themselves and colleagues for signs and symptoms, but this is very subjective [3]. The growing availability of commercial off-the-shelf wearable devices and Internet of Things sensors has made it feasible to continuously collect objective health and environmental information. These devices can wirelessly transmit data to cloud services where they can be accessed, processed, and viewed to determine if someone's health is deteriorating. For robust heat strain monitoring, it is best to continuously measure (vs spot check) several physiological variables such as T_c, HR, and physical activity level [3,7,11].

Core Body Temperature

Monitoring T_c is the most direct way of assessing risk for HRI, but accurate and consistent measurements might be difficult to obtain when individuals are wearing PPE [7]. Options for assessing body temperature include esophageal, rectal, tympanic (ie, ear), axillary (ie, armpit), and oral measurements. While esophageal and rectal temperatures most accurately track T_c, they cannot be measured conveniently [7,12]. Tympanic temperature is accurate to within 1 °C of T_c as long as the probe is properly shielded and can be regularly checked, provided that it is safe to temporarily remove PPE [7,12]. Axillary and oral temperatures are the least accurate because it is difficult to properly shield the probe from the environment. These methods of assessment can vary from T_c by over 1 °C [12]. Axillary temperature might also be challenging to measure if there are multiple layers of PPE being worn [7].

Heart Rate

Sustained high HR can be indicative of overexertion, which increases susceptibility to heat strain injuries [13]. If HR data are updated at a frequency of 1 minute or faster, T_c can be estimated using an algorithm such as the one developed by USARIEM [14]. Combined with measures of skin temperature and heat flux, the accuracy of these estimations of T_c can be improved [15].

HR can be measured using wearable electrocardiogram patches and chest straps, as well as optical ring- and smartwatch-based sensors. Patches and straps might cause skin irritation, and profuse sweating can cause the adhesive patch to fall off, so they are not appropriate for continuous wear. While optical devices might be better for ambulatory use and long-term monitoring, they typically do not provide the same level of accuracy [16].

Activity Level

Measuring activity level allows energy expenditure to be estimated and rest breaks to be tracked. Physical activity level is considered in the HSDA model [9]. Deriving these measurements from 3-axis accelerometer data is more accurate, but most wearables distill this information to either step counts, stair equivalents climbed, or minutes at a given level of movement intensity.

Considerations for Wearables Pilot Testing

Several wearable sensors can acquire health data relevant to assessing individuals' risk of developing an HRI. Multisensor

XSL•FO
RenderX

systems, especially those that can measure T_c, HR, and activity level, are preferable.

Systems incorporating electrocardiogram sensors (eg, Zephyr Bioharness 3, Hexoskin smart garments, and VivoSense LifeShirt) are the most accurate for measuring HR. Because close contact with the torso is required for accurate sensing, these platforms are also optimal for measuring respiration rate (RR) and physical activity. In terms of technology adoption, they are limited by the fact that they must be individually sized and can be uncomfortable for long periods of wear (especially when PPE is worn).

The accuracy of systems incorporating optical sensors (eg, smartwatches, armbands, and rings) has improved, and average HR, RR, and activity level can be acceptably measured. Several of these platforms also incorporate measures of blood oxygen saturation (SpO_2) or skin temperature. Systems with optical sensors are available in more convenient configurations for long-term monitoring of a diverse group of individuals.

After conversations with end users and wearables manufacturers in combination with limited benchtop testing, we selected the Garmin Fenix 6 smartwatch (Garmin Ltd) to pilot test for continuous HRI risk assessment. Other smartwatches (ie, Fitbit Charge 2, Empatica E4 wristband, and the Apple Watch Series 7) and chest straps (ie, Polar H10 and Zephyr Bioharness 3) were reviewed. The Garmin Fenix 6 smartwatch was selected because it can acquire several relevant physiological metrics in a convenient configuration that is appropriate for use with PPE. The Garmin Fenix 6 is ruggedized, operates using button clicks versus a touch screen interface, and has a battery life of 5 - 7 days when configured for continuous, high-resolution data acquisition.

Two data collections with convenience cohorts were conducted to determine the feasibility of using the Garmin Fenix 6 smartwatch for monitoring the health of people wearing PPE. The goals of these efforts were to:

1. Determine if a smartwatch is an appropriate configuration for monitoring individuals wearing PPE, including longer-term wear to determine recovery.
2. Determine if the Garmin Fenix 6 can sufficiently monitor vital signs when participants are active doing relevant work; the accuracy of the optical sensors used in smartwatches can be reduced when wrist movement is significant (ie, motion artifact) or for individuals with darker skin tones [16].
3. Gauge the accuracy of the measured and derived health metrics under various test conditions.

Methods

Overview

This section describes the study procedure including (1) participant recruitment, (2) study conditions, (3) equipment used for data acquisition, (4) data cleaning and processing strategies, and (5) data analysis. The goal of these data collection efforts was to obtain information related to wear compliance as well as data missingness and accuracy (eg, during activity and rest) for the different types of health data that are available from the Garmin Fenix series smartwatches.

Study Procedure

The study team identified local training events or exercises where PPE was being worn. The study team selected 2 events or exercises based on the types of activities that were planned, the type of PPE that was being worn, and the availability of study staff and equipment. These 2 events took place in the southeastern United States during the late spring and summer months (Table 1). Individuals who were already participating in these training events were informed by the event organizer of our research effort. In total, 8 individuals older than 18 years were voluntarily recruited from these convenience cohorts. Individuals who had sensitive skin or known allergies to the watch bezel or band material were excluded. The goals of the study, the study protocol, and the associated risks and benefits were verbally reviewed with the volunteers at the location of the training event or exercise prior to participation.

Table . Two data collections were conducted during field training events or exercises with convenience cohorts.

	Data collection #1	Data collection #2
Participant IDs	EPA101, EPA102, EPA103, EPA104	EPA201, EPA202, EPA203, EPA204, Control
Description	Environmental cleanup training [BT1]	New recruit firefighter training
Duration	4 - 5 days	3.5 hours
Environmental conditions	Daytime temperature: 15 - 34 °C Daytime humidity: 27% - 100%	Temperature: 30.5 - 32.8 °C Humidity: 29% - 38%
PPE[a]	Full-face air purifying respirator, escape mask, hard hat, inner and outer chemical-resistant gloves, and disposable chemical-resistant outer boots	Helmets, gloves, short sleeve cotton T-shirts, protective canvas pants, and boots

[a]PPE: personal protective equipment.

During the first pilot study, data were collected as part of a multiday environmental field study that required participants to don level C PPE [17]. Participants began wearing the Garmin smartwatch at the start of the field study and continued to wear the watch throughout the field study, including sleep. The participants were instructed to wear the watch snugly on their

XSL•FO
RenderX

nondominant hand; the fit of the watch was inspected by study staff at the beginning of the data collection. Long-term monitoring is needed to determine acclimatization status as well as readiness and recovery. The daytime (8 AM-5 PM) temperatures ranged from 15 to 34 °C and the humidity ranged between 27% and 100%.

During the second pilot study, data were collected during a 3.5-hour field training session for newly recruited firefighters. New firefighter recruits are more likely to experience an HRI because they are not acclimated to wearing PPE in warm environments [18]. During the field training session, participants practiced skills and tactical maneuvers such as entering and exiting a structure, carrying and assembling equipment, and climbing ladders. The participants wore typical gear for training outdoors during activities that did not involve extinguishing fires, including helmets, gloves, short-sleeve cotton T-shirts, protective canvas pants, and boots for the duration of the field training session (participants removed their helmets and gloves for a water break occurring approximately halfway through the training exercise). Participants began wearing the Garmin smartwatch on their nondominant hand at the start of the exercise and did not remove it for the duration of the exercise; the fit of the watch was inspected by study staff. Tympanic temperature was taken periodically using a Braun ThermoScan Ear Thermometer (Braun). The temperature for the second pilot study reached 32.7 °C and the humidity was 30% - 35%.

During both data collections, participants were observed, and investigators asked the participants to provide informal feedback related to the smartwatch.

Data Acquisition

RTI International's Wearables Research and Analytics Platform (WRAP) [19] was used to obtain high-resolution health data (Table 2) directly from the Garmin Fenix 6 smartwatches using an implementation of the Garmin Companion software development kit. No personally identifiable information was collected or recorded. Participants were assigned a study ID that did not contain personally identifiable information (eg, EPA101) through WRAP. The data were stored and accessed according to this ID and the smartwatch universally unique identifier (UUID). Only the data types described in Table 2 were collected.

Table . Continuous, high-resolution health metrics were acquired from the Garmin Fenix 6 smartwatch. The data were processed to remove values outside of the valid range before analysis.

	Data update rate	Units	Valid range
Step count	60 s	steps/min	0 - 220 [20]
Respiration rate (RR)	60 s	breaths/min	6 - 60 [21]
SpO_2[a]	60 s	percent	60 - 100 [22]
Beat-to-beat interval (BBI)	With each heart beat	millisecond	300 - 1500 (HR[b] of 40 - 200 beats/min)

[a]SpO_2: blood oxygen saturation.

[b]HR: heart rate.

The WRAP platform includes a mobile app (available on the App Store and Google Play Store) that is capable of 1:1 or 1:many communication between the tablet and wearable(s). Data from the wearable are packaged as JavaScript Object Notation (JSON) files and securely transmitted to a specified end point for storage (eg, the Microsoft Azure endpoint maintained by RTI International). The data can be visualized, and the entire record can be downloaded as a comma-separated value (CSV) file through a browser-based service. WRAP was selected because it can operate in resource-constrained environments as a closed system and does not require connection with Garmin's cloud services.

Data Cleaning and Processing

Before processing, artifacts caused by wrist movement and values outside of the expected valid range (Table 2) were removed. This eliminated noise from the incoming signals that can make small changes difficult to detect, which affects model performance.

The cleaned step count and beat-to-beat interval (BBI) data were aggregated into 60-second windows and averaged. The 60-second average BBI was converted to HR and used to estimate core body temperature (eCBT) using the USARIEM algorithm [14]. The 60-second average HR and step count were also used to provide a more granular representation of activity level.

The cleaned step count, RR, SpO_2, and BBI data were aggregated to 5-minute windows. The average and SD of the 5-minute window were computed for the step count, RR, SpO_2, and BBI. Time- and frequency-domain heart rate variability (HRV) metrics were calculated using the hrvanalysis package for Python (Aura Healthcare project, Robin Champseix) [23]; outlier and ectopic beats were removed and replaced through linear interpolation.

Data Analysis

In order to quantify wearability, the total duration of wear and the wear compliance were calculated. An individual was considered to be wearing the watch if, for a given 5-minute window, valid data were present for at least 2 of the 4 metrics described in Table 2.

In order to quantify data quality, the percentage of valid 5-minute windows for all data types was calculated. It should be noted that RR is derived from the same sensor as HR, but SpO_2 is measured by an independent sensor. The percentage of

XSL•FO
RenderX

corrected beats (ie, values outside of the expected range described in Table 2) was also calculated and used to assess HR signal quality. When enough data were present, we assessed if the data quality differed between periods of activity (steps>0) and rest (steps=0).

In order to quantify data accuracy, the recorded and calculated health metrics were compared with expected ranges for active, healthy individuals. In addition, the health metrics were analyzed to determine if they were responsive to activity (steps>0) and rest (steps=0). Where possible, we compared the calculated metrics to ground truth data (eg, eCBT vs tympanic measurement).

Ethical Considerations

The University of North Carolina Institutional Review Board approved the data collection protocol that was followed in this study (22 - 0817) and determined this study to not be human participant research. Study staff obtained verbal consent from the participants at the event or exercise site before data collection began. Study staff explained that participation was voluntary and that individuals could stop participating in the data collection at any time without consequence and that their data would be destroyed. Study staff also explained the risks of the study, including a topical skin reaction to the watch bezel or band material, and informed participants that they could remove the watch at any time if it was causing discomfort or interfering with job duties. No identifying information was collected from participants, and all of the smartwatch data were stored and analyzed using an assigned study identification code (eg, EPA101) that did not contain personally identifiable information. Participants did not receive compensation for volunteering for the study.

Results

Overview

This section describes our preliminary findings regarding the (1) wearability, (2) data quality, and (3) accuracy of the Garmin Fenix 6 for continuously monitoring individuals wearing PPE.

Wearability

Wear compliance is important for assessing acclimatization and recovery regarding wearing PPE in warm and humid environments. In the first data collection, 4 participants (EPA101, EPA102, EPA103, and EPA104) wore the watch for several days, including during a training exercise while wearing PPE. The duration of wear was calculated as the difference between the first and the last valid BBI data point. Wear compliance (ie, percentage of 5-min windows with at least 2 of the 4 smartwatch metrics present over the wear period) was exceptionally high for this cohort, with all 4 individuals wearing the watch for more than 80% of the time (EPA101 wore the watch for 99% of their 3.91 days of participation; EPA102 wore the watch 99.9% of their 5.12 days of participation; EPA103 wore the watch 83.8% of their 4.11 days of participation; EPA104 wore the watch 86.8% of their 4.13 days of participation). This is an important metric to track for long-term monitoring of readiness and recovery.

In the second data collection, 4 different participants (EPA201, EPA202, EPA303, and EPA204) wore the watch for approximately 3.5 hours during a training session. In both pilot studies, the use of PPE did not interfere with data collection, and the Garmin Fenix 6 smartwatch did not impede the protective factor of PPE or negatively impact the safety of the individual.

Data Quality

In the first data collection, RR and SpO_2 data were not collected for 2/4 participants due to an error in device initialization. The signal quality was not different between rest and activity for EPA101, but for EPA104, it was somewhat degraded during activity (Table 3). For EPA104, there are fewer valid windows for RR and SpO_2 data during activity (>0 steps over a 5-minute window) compared to rest (0 steps over a 5-minute window); RR and SpO_2 data quality were comparable between activity and rest for EPA101. HR and HRV data quality were not impacted by activity for either participant.

Table . Data quality was compared for 2 participants in the first data collection with valid heart rate, HRV[a], RR[b], and SpO_2[c] data; a device initialization error resulted in missing respiration rate and SpO_2 data from the other 2 participants.

Study ID	EPA101		EPA104	
Physiological state	Rest (%)	Active (%)	Rest (%)	Active (%)
% Correction	8.7	8.9	6.6	6.4
% Valid HRV windows	98	100	100	97
% Valid RR windows	97	98	99	85
% Valid SpO_2 windows	93	94	97	89

[a]HRV: heart rate variability.

[b]RR: respiration rate.

[c]SpO_2: blood oxygen saturation.

The second data collection included 1 participant (EPA201) who had darker skin [16]; this participant had degraded SpO_2 signal quality compared to the other participants as indicated by the reduction in valid SpO_2 values for EPA201 (Table 4). The second data collection also included 1 participant (Control) who was not active (ie, average step count of 4.5 steps/min vs 16.1 steps/min for the other participants) and was not wearing PPE, but who was exposed to the same environmental

XSL•FO
RenderX

conditions. This participant had reduced HR signal quality compared to the other participants (Table 4), which could be attributed to the comparatively looser fit of their watch (this participant was female and had a smaller wrist compared to all of the other participants who were male).

Table . Data quality was compared for 5 participants in the second data collection.

Study ID	Control (%)	EPA201 (%)	EPA202 (%)	EPA203 (%)	EPA204 (%)
% Correction	9.3	6.3	6.8	5.8	7.4
% Valid HRV[a] windows	100	100	100	100	100
% Valid RR[b] windows	100	97	97	97	95
% Valid SpO_2[c] windows	100	74	97	100	97

[a]HRV: heart rate variability.

[b]RR: respiration rate.

[c]SpO_2: blood oxygen saturation.

Data Accuracy

All measured and estimated metrics (Tables S1 and S2 in Multimedia Appendix 1) were within physiologically acceptable ranges (Table 2) for active, healthy individuals. For the first data collection, HR and eCBT increased with activity and decreased during rest as expected (core temperature was initialized to 37 °C for all individuals; Figure 1), but RR and SpO_2 did not (Table S1 in Multimedia Appendix 1).

Figure 1. Example activity, heart rate, and estimated core body temperature data for a participant in the first multiday data collection demonstrate that the heart rate and estimated core body temperature measured using the Garmin Fenix 6 increase with physical activity and rest. bpm: beats per minute; eCBT: estimated core body temperature.

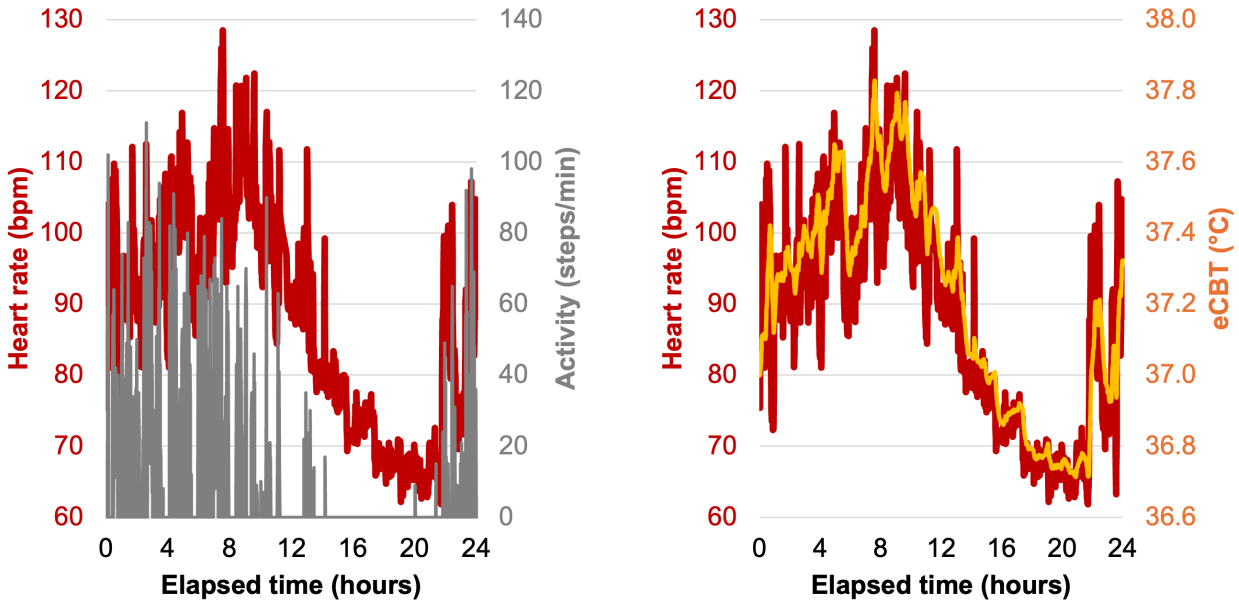

In the second data collection, there was a difference in HR, eCBT, and RR between the 1 control and the 4 active participants (Figure 2; Table S2 in Multimedia Appendix 1). For this study, the T_c used in the USARIEM algorithm was initialized to the first measured temperature (using a Braun ThermoScan Ear Thermometer). The estimated (via the USARIEM algorithm) and measured temperatures deviated after this initial measurement and were overestimated on average by 0.5 °C for the active participants and underestimated for the inactive participant by 0.4 °C (Figure 3; Table S3 in Multimedia Appendix 1).

XSL•FO
RenderX

Figure 2. Heart rate comparisons for 4 active participants and 1 control participant (who was exposed to the same environmental conditions) during the second 3.5-hour data collection demonstrate that the Garmin Fenix 6 smartwatch can be used to measure heart rate differences during physical activity. bpm: beats per minute.

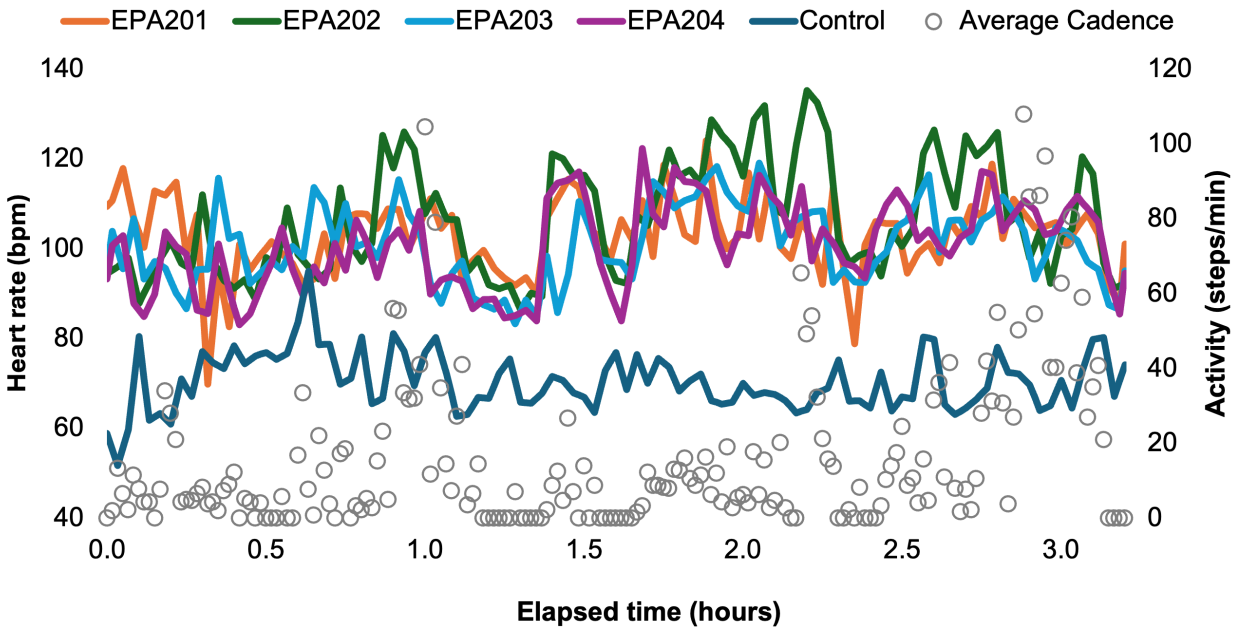

Figure 3. The heart rate measurements (shown as a dark red line) track with physical activity (shown as a gray line), but the core body temperature that is estimated from heart rate (shown as a gold line) overestimates the tympanic temperature (shown as gold square markers). bpm: beats per minute; eCBT: estimated core body temperature.

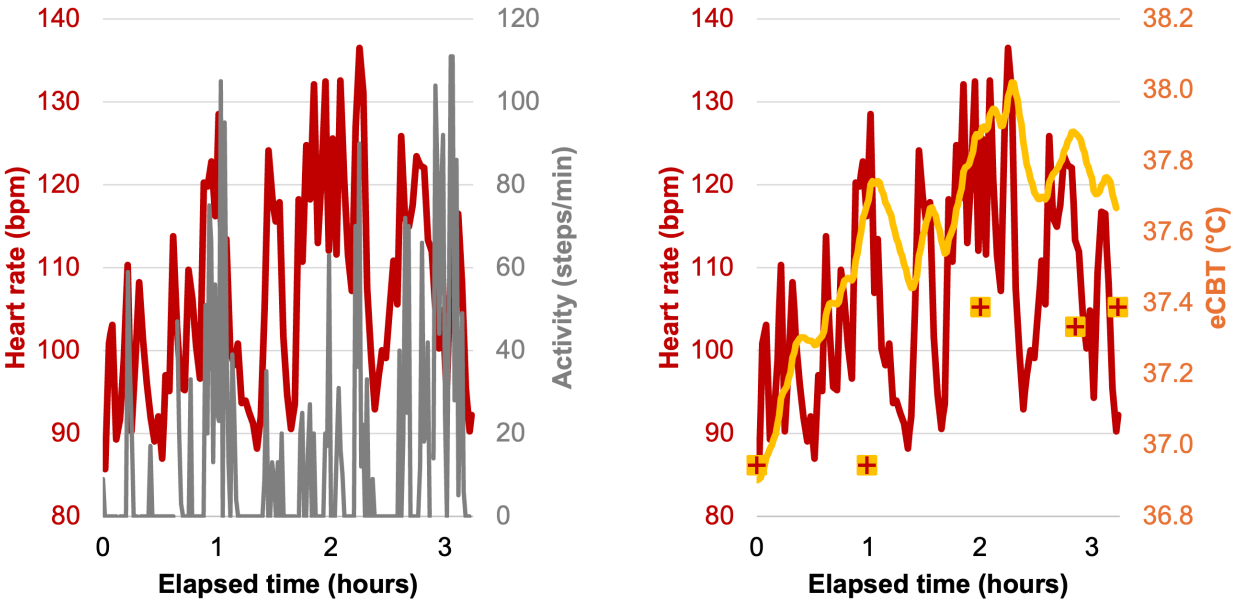

Discussion

Overview

Continuously monitoring individuals using wearables can provide near-real-time health information. Vital sign measurements such as HR, RR, SpO_2, and temperature are available on many commercial off-the-shelf wearables. Providing this information to workers and the personnel responsible for ensuring their health can be used to better inform current workload and recovery. This is critical for promptly initiating cooling strategies, which can be difficult if individuals are working in dangerous environments or wearing contaminated PPE.

Principal Results

Both data collection efforts confirmed that the use of the Garmin Fenix 6 smartwatch for continuous health assessment was feasible. Feedback was informally collected for the 8 participants who wore the smartwatch while wearing PPE. Based on this feedback and observations from study staff, the Fenix 6 smartwatch did not interfere with work while wearing PPE. In addition, results from the first data collection showed that the smartwatch could be worn continuously for several days

XSL•FO
RenderX

(average of 93% wear compliance over 4.3 days for the first pilot study). The participants did not find the watch to be uncomfortable, and one individual noted that it was about the same size as the watch they regularly wore. However, the Fenix 6 was larger and bulkier than the other smartwatches and activity trackers worn by most of the test participants (eg, the Apple Watch and various Fitbit models), but all noted that the battery life of the Fenix 6 was superior.

In addition, the data quality from the Garmin Fenix 6 sensors was sufficient to report 60-second and 5-minute BBI, HRV, and RR metrics and was not significantly affected by skin tone or arm and hand movement. These metrics also showed appropriate changes with activity level, although data quality was degraded during activity in some cases. SpO_2 was reported less often during activity, especially for the individual who had a darker skin tone.

Core body temperature was estimated from 60-second average HR measurements using the USARIEM algorithm [14]. The original algorithm was developed using HR data from a Polar chest strap which, if worn correctly, is more accurate than HR data from a smartwatch [16]. While eCBT was within the expected range, the estimated and measured body temperatures disagreed and were overestimated on average by 0.5 °C for the active participants and underestimated for the inactive participants by 0.4 °C. Buller et al [14] noted that core body temperature would be underestimated at low workloads, which is consistent with the results for the control participant in the second pilot study.

Impact

The goal of this work is to consider the use of wearable sensors to provide actionable information related to the risk of developing an HRI. Our vision is to continuously monitor individuals to mitigate adverse health events in the field as well as track the recovery phase to determine overall readiness to work. This information could be used to initiate timely intervention versus relying on subjective symptom monitoring. In addition, this information could feed exposure prediction models like the HSDA to either provide more accurate information for initial predictions or update forecasts during the workday.

As an example, during the first pilot test, participant EPA104 attained a maximum HR of over 140 bpm and eCBT of 38.3 °C on the first day of testing. This individual maintained an HR of 125 bpm or higher for 1.8 hours on that day, which is consistent with moderate exercise. When examining their recovery trajectory over the measurement period of 5 days and nights, their HR (Figure 4) and eCBT (Figure 5) were elevated in the earlier part of sleep on the second night following another active day. This might indicate that they did not fully recover from this sustained exercise.

Figure 4. Day 1 (shown as a bright red line) was EPA104's most active day and their heart rate reached its highest level, exceeding 125 bpm for close to 2 hours, and recovered during the evening hours. bpm: beats per minute.

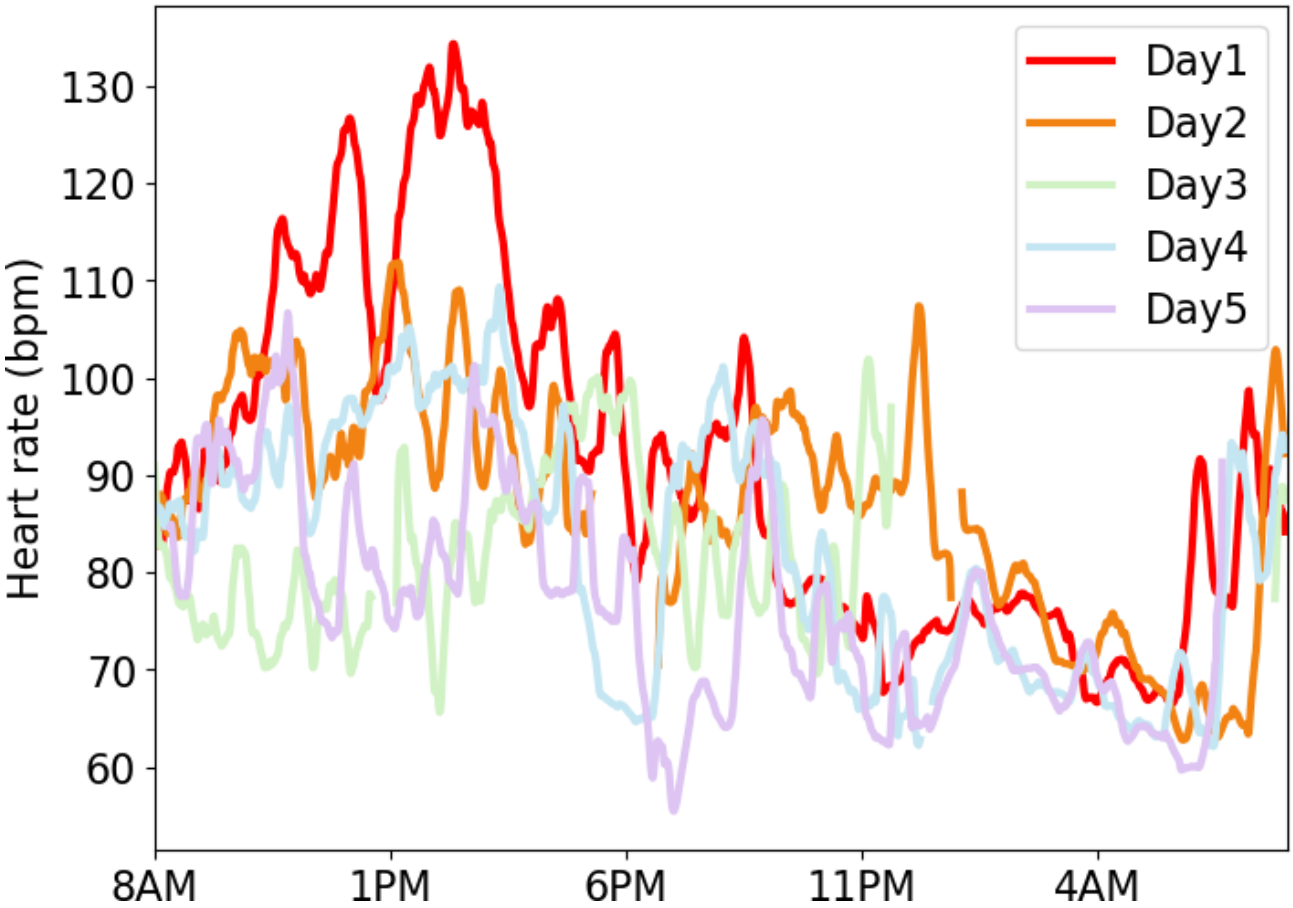

XSL•FO
RenderX

Figure 5. Core body temperature was estimated from the average heart rate over 5 days and nights for participant EPA104 and reached its highest level of 38.3 °C on day 1 which was the participant's most active day. eCBT: estimated core body temperature.

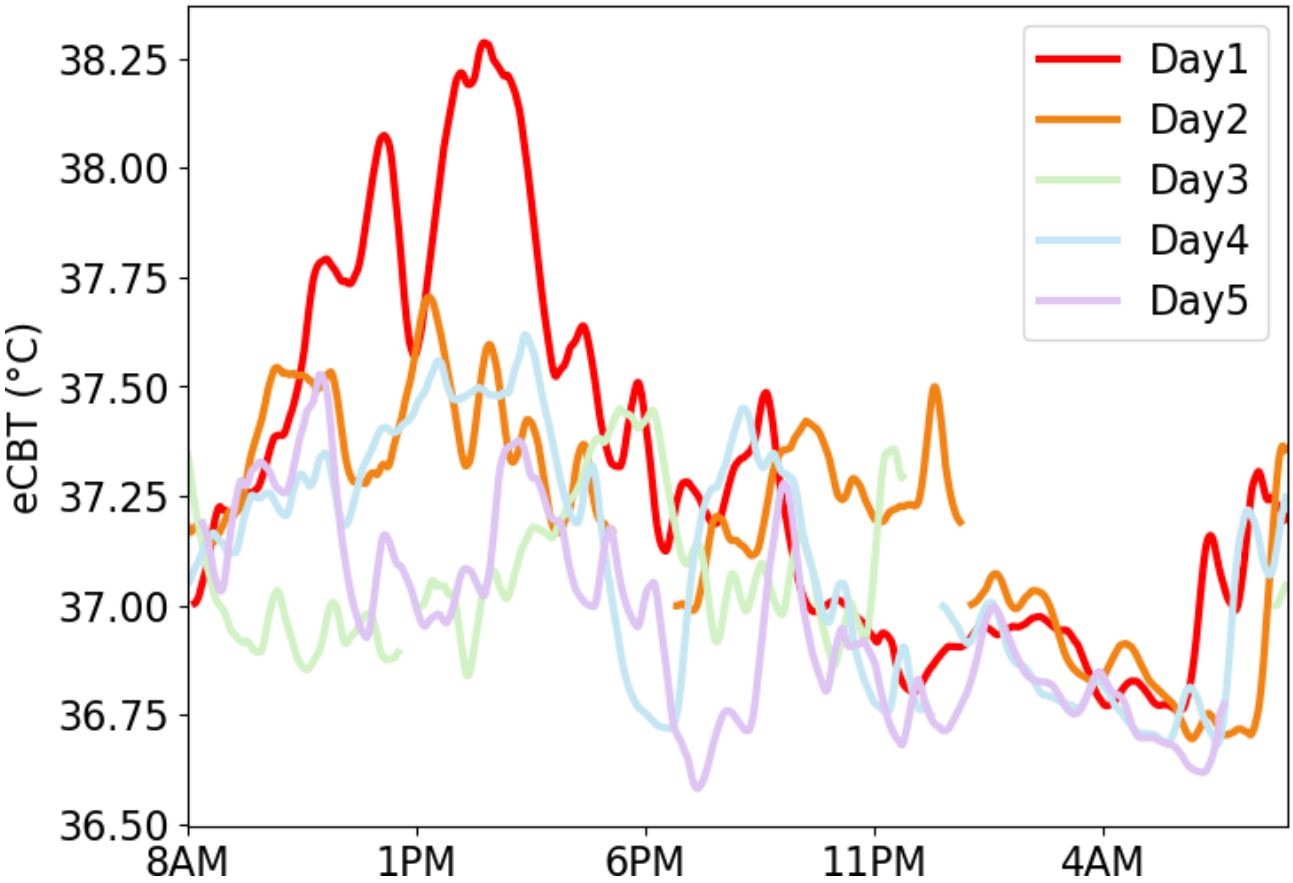

Comparison with Prior Work

Several recent studies have assessed the prevalence of HRIs among different groups of individuals who wear PPE or work outdoors [1-3,6,24]. The majority of these studies use survey-based tools to collect demographic information, characteristics of the workplace, the prevalence of heat strain symptoms, and accessibility to mitigation strategies (including education). The common main finding from these studies is that the prevalence of HRIs is underreported because people might ignore chronic, low-grade symptoms [3]. Wearables can provide visibility into health metrics that are sensitive to HRI and quantify the effects of these subacute events.

Runkle et al [24] outfitted a small group of workers with a wearable temperature sensor that measured the ambient environment (Thermochron iButton; iButtonLink) and a Garmin Vivoactive smartwatch. They measured HR at a 60-second resolution and postprocessed data to look for periods of sustained high HR (ie, 5-minute average at or exceeding 180 bpm). Our work extended Runkle et al [24] by including additional metrics that are relevant to monitoring for HRI. Studies have been dedicated to generating data to develop and validate models that estimate T_c from HR and skin temperature [13-15]. Our work extended data collection from HR straps to a more convenient configuration for long-term wear (ie, smartwatch).

Limitations of This Study

Participants were drawn from a pool of volunteers who were taking part in ongoing training events or exercises involving the use of PPE. While convenience samples are representative of the target end users, the convenience samples from our pilot studies were not diverse. Across both groups, only males volunteered (the female control participant was not wearing PPE and was not engaged in the training events or exercise). Sex-based differences are known in the relationship between HR and core body temperature, as well as physiological responses to exercise and dehydration [25,26]. In addition, 7 of the 8 users were Caucasian, and there are known differences in the accuracy of optical sensors based on skin tone [16].

To assess feasibility, participants would ideally wear the watch for at least several days. Logistics around remote sensor distribution and collection hindered data collection outside of the scheduled training period. For both data collections, participants began wearing the watch at the start of the exercise. For the first data collection, study staff remained onsite for several days as part of other activities, enabling watches to be worn for an extended period of time. In the second data collection, study staff could not stay on site, and watches were returned immediately following the exercise. In addition, this study would benefit from a structured feedback survey to obtain information about smartwatch use with PPE versus the informal methods that were used.

Accuracy assessments should be made by comparing the data obtained from wearables to "gold standard" sensors. These

XSL•FO
RenderX

sensors are typically used in clinics and are not appropriate for continuous monitoring in the field. In the second data collection, an ear thermometer was used to assess body temperature quickly and conveniently, but tympanic temperature varies more than core temperature [27], so other options should be evaluated.

Recommendations

The pilot studies demonstrated the feasibility of using the Garmin Fenix 6 smartwatch to continuously monitor the health of individuals working in the field while wearing PPE. For future data collection efforts, we recommend:

1. Intentionally recruiting a more diverse group of participants in order to ensure that the results can be translated to all potential end users. For example, we found that the SpO_2 readings were less reliable for individuals with darker skin tones when they were active. This means that models should not rely on this measurement for active individuals and further testing is needed to determine validity at rest.
2. Including higher levels of PPE and a greater diversity of tasks in the data collection period.
3. Asking participants to wear the smartwatch for a longer period of time in order to capture data before and after the exercise. Baseline data could be used to explain between-person differences in response to activity and heat exposure. Recovery data could be used to understand what data are needed to accurately predict the risk of developing heat strain over days of repeated exposure.
4. Exploring other sensors to acquire more accurate information during training (eg, a sensor that measures core temperature versus estimating it from HR) and verify smartwatch measurements.
5. Asking participants to keep an activity log and track fluid intake to verify the smartwatch measurements and better understand periods when the smartwatch data diverge from more accurate sources.
6. Using a structured feedback survey to collect information about the smartwatch and monitoring technology.

Conclusions

Occupational heat strain can significantly impact productivity, increase the risk of injury, and have long-lasting health consequences. Wearables can be used to continuously monitor health. We conducted 2 data collections to determine the feasibility of using wearables to monitor for risk of HRI. Results were favorable, and we are planning a larger-scale study involving a more diverse group of participants, longer data collection periods, and additional, higher-resolution sensors.

Acknowledgments

This material has been funded wholly or in part by the United States Environmental Protection Agency (EPA) under contract 68HERC21D0004 to RTI International. It has been subject to the EPA's review, and it has been approved for publication as an EPA document. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The supplementary tables containing detailed physiological information for the 8 test participants.
[DOCX File, 17 KB - formative_v9i1e72324_app1.docx]

References

1. Bongers C, de Korte JQ, Zwartkruis M, Levels K, Kingma BRM, Eijsvogels TMH. Heat strain and use of heat mitigation strategies among COVID-19 healthcare workers wearing personal protective equipment-a retrospective study. Int J Environ Res Public Health 2022 Feb 8;19(3):1905. [doi: 10.3390/ijerph19031905] [Medline: 35162925]
2. Carballo-Leyenda B, Villa-Vicente JG, Delogu GM, Rodríguez-Marroyo JA, Molina-Terrén DM. Perceptions of heat stress, heat strain and mitigation practices in wildfire suppression across Southern Europe and Latin America. Int J Environ Res Public Health 2022 Sep 27;19(19):12288. [doi: 10.3390/ijerph191912288] [Medline: 36231589]
3. Rogerson S, Brearley M, Meir R, Brooks L. Influence of age, geographical region, and work unit on heat strain symptoms: a cross-sectional survey of electrical utility workers. J Occup Environ Hyg 2020;17(11-12):515-522. [doi: 10.1080/15459624.2020.1834112] [Medline: 33196398]
4. Occupational Safety and Health Administration. OSHA technical manual (OTM). section III: chapter 4: heat stress. US Department of Labor. 2017. URL: https://www.osha.gov/otm/section-3-health-hazards/chapter-4 [accessed 2025-09-12]
5. Kim DH, Bae GT, Lee JY. A novel vest with dual functions for firefighters: combined effects of body cooling and cold fluid ingestion on the alleviation of heat strain. Ind Health 2020 Apr 2;58(2):91-106. [doi: 10.2486/indhealth.2018-0205] [Medline: 31257232]

XSL•FO
RenderX

6. Kim S, Kim DH, Lee HH, Lee JY. Frequency of firefighters' heat-related illness and its association with removing personal protective equipment and working hours. Ind Health 2019 Jun 4;57(3):370-380. [doi: 10.2486/indhealth.2018-0063] [Medline: 30210098]
7. Notley SR, Flouris AD, Kenny GP. On the use of wearable physiological monitors to assess heat strain during occupational heat stress. Appl Physiol Nutr Metab 2018 Sep;43(9):869-881. [doi: 10.1139/apnm-2018-0173]
8. Malchaire J, Piette A, Kampmann B, et al. Development and validation of the predicted heat strain model. Ann Occup Hyg 2001 Mar;45(2):123-135. [doi: 10.1016/S0003-4878(00)00030-2] [Medline: 11182426]
9. Potter AW, Blanchard LA, Friedl KE, Cadarette BS, Hoyt RW. Mathematical prediction of core body temperature from environment, activity, and clothing: The heat strain decision aid (HSDA). J Therm Biol 2017 Feb;64:78-85. [doi: 10.1016/j.jtherbio.2017.01.003] [Medline: 28166950]
10. Vogt JJ, Candas V, Libert JP, Daull F. Chapter 6: required sweat rate as an index of thermal strain in industry. In: Studies in Environmental Science 1981:99-110. [doi: 10.1016/S0166-1116(08)71083-5]
11. Bergeron MF. Heat stress and thermal strain challenges in running. J Orthop Sports Phys Ther 2014 Oct;44(10):831-838. [doi: 10.2519/jospt.2014.5500] [Medline: 25103131]
12. Robinson JL, Seal RF, Spady DW, Joffres MR. Comparison of esophageal, rectal, axillary, bladder, tympanic, and pulmonary artery temperatures in children. J Pediatr 1998 Oct;133(4):553-556. [doi: 10.1016/s0022-3476(98)70067-8] [Medline: 9787697]
13. Murdock R, Hagen J. Soldier safety and performance through wearable devices. Presented at: Micro- and Nanotechnology Sensors, Systems, and Applications X; Apr 15-19, 2018; Orlando, FL, United States. [doi: 10.1117/12.2304899]
14. Buller MJ, Tharion WJ, Cheuvront SN, et al. Estimation of human core temperature from sequential heart rate observations. Physiol Meas 2013 Jul;34(7):781-798. [doi: 10.1088/0967-3334/34/7/781] [Medline: 23780514]
15. Welles AP, Xu X, Santee WR, et al. Estimation of core body temperature from skin temperature, heat flux, and heart rate using a Kalman filter. Comput Biol Med 2018 Aug 1;99:1-6. [doi: 10.1016/j.compbiomed.2018.05.021] [Medline: 29803944]
16. Bent B, Goldstein BA, Kibbe WA, Dunn JP. Investigating sources of inaccuracy in wearable optical heart rate sensors. NPJ Digit Med 2020;3(1):18. [doi: 10.1038/s41746-020-0226-6] [Medline: 32047863]
17. Personal protective equipment. Environmental Protection Agency. 2022. URL: https://www.epa.gov/emergency-response/personal-protective-equipment [accessed 2023-02-26]
18. Lui B, Cuddy JS, Hailes WS, Ruby BC. Seasonal heat acclimatization in wildland firefighters. J Therm Biol 2014 Oct;45:134-140. [doi: 10.1016/j.jtherbio.2014.08.009] [Medline: 25436962]
19. WRAP. URL: https://www.gowrap.org [accessed 2023-01-25]
20. Burns GT, Zendler JM, Zernicke RF. Step frequency patterns of elite ultramarathon runners during a 100-km road race. J Appl Physiol (1985) 2019 Feb 1;126(2):462-468. [doi: 10.1152/japplphysiol.00374.2018] [Medline: 30543498]
21. Nicolò A, Massaroni C, Schena E, Sacchetti M. The importance of respiratory rate monitoring: from healthcare to sport and exercise. Sensors (Basel) 2020 Nov 9;20(21):6396. [doi: 10.3390/s20216396] [Medline: 33182463]
22. Louie A, Feiner JR, Bickler PE, Rhodes L, Bernstein M, Lucero J. Four types of pulse oximeters accurately detect hypoxia during low perfusion and motion. Anesthesiology 2018 Mar;128(3):520-530. [doi: 10.1097/ALN.0000000000002002] [Medline: 29200008]
23. Champseix R. hrvanalysis. URL: https://aura-healthcare.github.io/hrv-analysis/ [accessed 2025-10-11]
24. Runkle JD, Cui C, Fuhrmann C, Stevens S, Del Pinal J, Sugg MM. Evaluation of wearable sensors for physiologic monitoring of individually experienced temperatures in outdoor workers in southeastern U.S. Environ Int 2019 Aug;129:229-238. [doi: 10.1016/j.envint.2019.05.026] [Medline: 31146157]
25. Wickham KA, McCarthy DG, Spriet LL, Cheung SS. Sex differences in the physiological responses to exercise-induced dehydration: consequences and mechanisms. J Appl Physiol (1985) 2021 Aug 1;131(2):504-510. [doi: 10.1152/japplphysiol.00266.2021] [Medline: 34197234]
26. Broman ME, Vincent JL, Ronco C, Hansson F, Bell M. The relationship between heart rate and body temperature in critically ill patients. Crit Care Med 2021 Mar 1;49(3):e327-e331. [doi: 10.1097/CCM.0000000000004807] [Medline: 33566464]
27. Sanderson B, Lim L, Lei K, Smith J, Camporota L, Beale R. A comparison of core and tympanic temperature measurement in the critically ill. Crit Care 2010;14(Suppl 1):329. [doi: 10.1186/cc8561]

Abbreviations

BBI: beat-to-beat interval
bpm: beats per minute
CSV: comma-separated value
eCBT: estimated core body temperature
EPA: US Environmental Protection Agency
HAZMAT: Hazardous Materials Management
HR: heart rate
HRI: heat-related injury

XSL•FO
RenderX

HRV: heart rate variability
HSDA: Heat Strain Decision Aid
IAFF: International Association of Fire Fighters
JSON: JavaScript object notation
PHS: Predicted Heat Strain
PPE: personal protective equipment
RR: respiration rate
SpO_2: blood oxygen saturation
T_c: core temperature
USARIEM: US Army Research Institute of Environmental Medicine
UUID: universally unique identifier
WRAP: Wearables and Analytics Platform

Edited by A Mavragani; submitted 07.02.25; peer-reviewed by B Senst, R Bryan Ormond; revised version received 01.07.25; accepted 06.07.25; published 17.10.25.

Please cite as:
Hegarty-Craver M, Womack D, Thornburg J, Boe T, Archer MJ, Calfee W
Predicting Risk of Heat-Related Injuries for Individuals Wearing Personal Protective Equipment Using Smartwatches: Feasibility Observational Study
JMIR Form Res 2025;9:e72324
URL: https://formative.jmir.org/2025/1/e72324
doi:10.2196/72324

© Meghan Hegarty-Craver, Donna Womack, Jonathan Thornburg, Timothy Boe, M John Archer, Worth Calfee. Originally published in JMIR Formative Research (https://formative.jmir.org), 17.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Development, Implementation, and Usability Evaluation of the CANMI App to Monitor the Quality of Maternal and Child Nutrition Care in Primary Health Units: Mixed Methods Pilot Study

Soraya Burrola-Méndez[1,2*], MAN; Elizabeth Hoyos-Loya[2,3*], PhD; J Emilio Quiroz-Ibarra[4], PhD; Cecilia Perez Navarro[1,2], MC; Isabel Omaña-Guzmán[5], PhD; Jorge Ángel González-Ordiano[4], PhD; Omar Acosta-Ruiz[6], MPH; Arturo C Bautista-Morales[6], MHCM; Daniela Tánchez Sandoval[1,2], MAN; Sonia Hernández-Cordero[2,7], PhD; Cinthya Muñoz-Manrique[2,8], PhD; Monica Ancira-Moreno[1,2], PhD

[1]Health Department, Iberoamerican University, Prolongación Paseo de la Reforma 880, Lomas de Santa Fe, Mexico City, Mexico
[2]Observatorio Materno Infantil, Iberoamerican University, Mexico City, Mexico
[3]Center for Health Systems Research, National Institute of Public Health, Cuernavaca, Mexico
[4]Institute of Applied Research and Technology, Iberoamerican University, Mexico City, Mexico
[5]Pediatric Obesity Clinic and Wellness Unit, Hospital General de México "Dr. Eduardo Liceaga", Mexico City, Mexico
[6]Center for Evaluation and Surveys Research, National Institute of Public Health, Cuernavaca, Mexico
[7]Research Center for Equitable Development (EQUIDE), Iberoamerican University, Mexico City, Mexico
[8]Nutrition and Bioprogramming Coordination, National Institute of Perinatology, Mexico City, Mexico
[*]these authors contributed equally

Corresponding Author:
Monica Ancira-Moreno, PhD
Health Department, Iberoamerican University, Prolongación Paseo de la Reforma 880, Lomas de Santa Fe, Mexico City, Mexico

Abstract

Background: In Mexico, the maternal and child population continues to face a high burden of malnutrition, posing a persistent public health challenge. The health care system plays a crucial role, not only in addressing existing cases but also in preventing and detecting malnutrition early. Mobile health technologies have the potential to strengthen maternal and child health services by improving the quality, accessibility, and timeliness of nutritional care.

Objective: The aim was to design, develop, and assess the usability and acceptance of a mobile app—CANMI (Calidad de la Atención Nutricional Materno Infantil; its Spanish acronym)—to monitor the quality of maternal and child nutritional care in primary health care units in Mexico.

Methods: The framework of the CANMI app was based on 16 validated indicators designed to assess the quality of nutritional care during the preconception, pregnancy, postpartum, early childhood, and preschool stages. The app was developed for both iOS and Android systems using a user-centered design approach. Following development, we conducted a pilot usability study in a randomized sample of 18 primary health care units in Guanajuato, Mexico. Trained nutritionists implemented the app and collected usability data at the end of the initial use period and again 6 weeks later. To further explore user experience, semistructured online interviews were conducted to identify barriers, facilitators, and overall satisfaction with the app.

Results: The CANMI app allows the systematic registration of key indicators to assess the quality of nutritional care in primary health care settings. Users described the app as simple, intuitive, and visually appealing. Overall usability was rated positively, with a mean score of 71.13 (SD 11.68) on the System Usability Scale, indicating good acceptability. The app's offline functionality, streamlined interface, and efficiency in data collection were identified as key facilitators of use. Reported benefits included reduced time for data entry and perceived improvements in the quality of nutritional care. Identified barriers to integration included the need to use personal devices, user fatigue due to prolonged screen time, inconsistent clinical records, and limited time to incorporate the app into routine workflows. Importantly, the app encouraged and promoted improvements in documentation practices and heightened awareness among health personnel regarding the precision and clarity of their nutritional recommendations.

Conclusions: The CANMI app provides a feasible and effective solution for monitoring the quality of maternal and child nutritional care in primary health settings. Its high usability and offline capabilities make it particularly suitable for low-connectivity environments. Beyond facilitating data collection, the app contributed to improved clinical documentation practices and enhanced health care provider awareness of care quality. Consequently, the app represents a promising digital tool to support the implementation of evidence-based, user-centered strategies aimed at strengthening maternal and child health services in resource-limited contexts.

XSL•FO
RenderX

(JMIR Form Res 2025;9:e77539) doi:10.2196/77539

KEYWORDS

mobile health technologies; mHealth technologies; maternal and child nutrition; quality care; mobile app; primary health care

Introduction

Background

Proper nutrition during the perinatal period and early childhood is critical to preventing malnutrition and its long-term impacts on mothers and children [1]. In Mexico, malnutrition remains a major concern: among children aged less than 5 years, rates of underweight, stunting, wasting, and obesity are significant [2], while many women of reproductive age face anemia, overweight, and obesity [3,4]. Most of these issues can be addressed through primary care, where outpatient services play a key role in health promotion, prevention, and treatment, especially during the antenatal and perinatal periods, which are largely managed outside hospital settings [5].

High-quality health care involves accurately identifying health needs and implementing comprehensive strategies to address them through appropriate resource allocation [6]. Quality is a measurable attribute of health systems that is assessed through key dimensions such as person centeredness, effectiveness, efficiency, equity, comprehensiveness, timeliness, safety, and sustainability and must be ensured across all levels of care, from prevention to specialized services [7,8].

In Mexico, efforts to define and monitor the quality of nutritional care have led to the development of 16 indicators focused on maternal and child health at the primary care level, which address up to 85% of the health needs of the mother-child dyad [9,10]. However, 2021 data from 6 states showed that only 8 out of 100 women and children aged less than 5 years received quality nutritional care in primary care settings. Major barriers identified by health care providers included insufficient training in malnutrition prevention, diagnosis, and treatment, especially among medical and nursing staff, and the lack of a system to monitor the quality of nutritional care [11,12].

Efforts to improve the quality of health care, particularly in maternal and child nutrition, face challenges such as insufficient training and the lack of monitoring systems [11,12]. Mobile health (mHealth) offers clear advantages over conventional monitoring systems by enabling continuous, real-time health monitoring and improving accessibility, especially for remote populations. It reduces long-term costs by minimizing the need for physical infrastructure and manual labor and enhances health care delivery by enabling real-time monitoring, identifying care gaps, and providing accessible, personalized information [13,14]. These technologies promote care coordination, support clinical decision-making, and improve communication between health care professionals, ultimately leading to more timely and individualized care [15,16].

This is feasible considering that 83.1% of the population has internet access, and among them, 97.2% access it via a smartphone [17]. Although it faces challenges such as technology adoption and data security, mHealth is more adaptable and better suited for integration with modern health care IT systems compared to the slower, less-flexible conventional methods. In maternal and child health, mHealth tools allow for continuous monitoring, ensuring health care providers can respond quickly and effectively to safeguard the well-being of the mother-child dyad [18].

Objectives

Promoting research in this area is critical to unlocking the potential of mHealth technologies in optimizing nutritional care within Mexico's public health system. This study aims to design, develop, and assess the usability and acceptance of a mobile app that will equip health care personnel with the tools to monitor and enhance the quality of maternal and child nutritional care at the primary health care level, ultimately improving health outcomes for the mother-child dyad.

Methods

Design and Development of the App

The design and development of the CANMI (Calidad de la Atención Nutricional Materno Infantil) app were based on 2 main components. The first includes the needs, strengths, and gaps identified through a previous diagnostic evaluation of the quality of maternal and child nutritional care in Mexico [11]. The second, a group of 16 validated quality indicators, organized across 5 stages of the life course—preconception, pregnancy, postpartum, infancy, and preschool age—specifically developed for this context by Ancira-Moreno et al [9].

The development phase involved defining the app's conceptual framework, designing user interfaces and screens, and iteratively refining the prototype. The app was programmed for both iOS and Android operating systems. Quality levels within the app were categorized based on compliance with the indicators: ≥90% indicates good quality nutritional care, 71% to 89% indicates deficient nutritional care, and ≤70% indicates poor quality nutritional care. A traffic light system was used to visualize the results.

Figure 1 illustrates the app's workflow. The front end allows health professionals or quality care monitors to register responses to a stage-specific questionnaire based on the quality indicators. Once the data are captured, they are stored locally on the device. When a minimum of 3G internet connectivity becomes available, the data are automatically transmitted to a centralized server using a representational state transfer application programming interface, which enables secure and efficient bidirectional communication between the device and the server.

XSL•FO
RenderX

Figure 1. App workflow. Process of collecting, storing, and transmitting quality-indicator data in the mobile app. API: application programming interface.

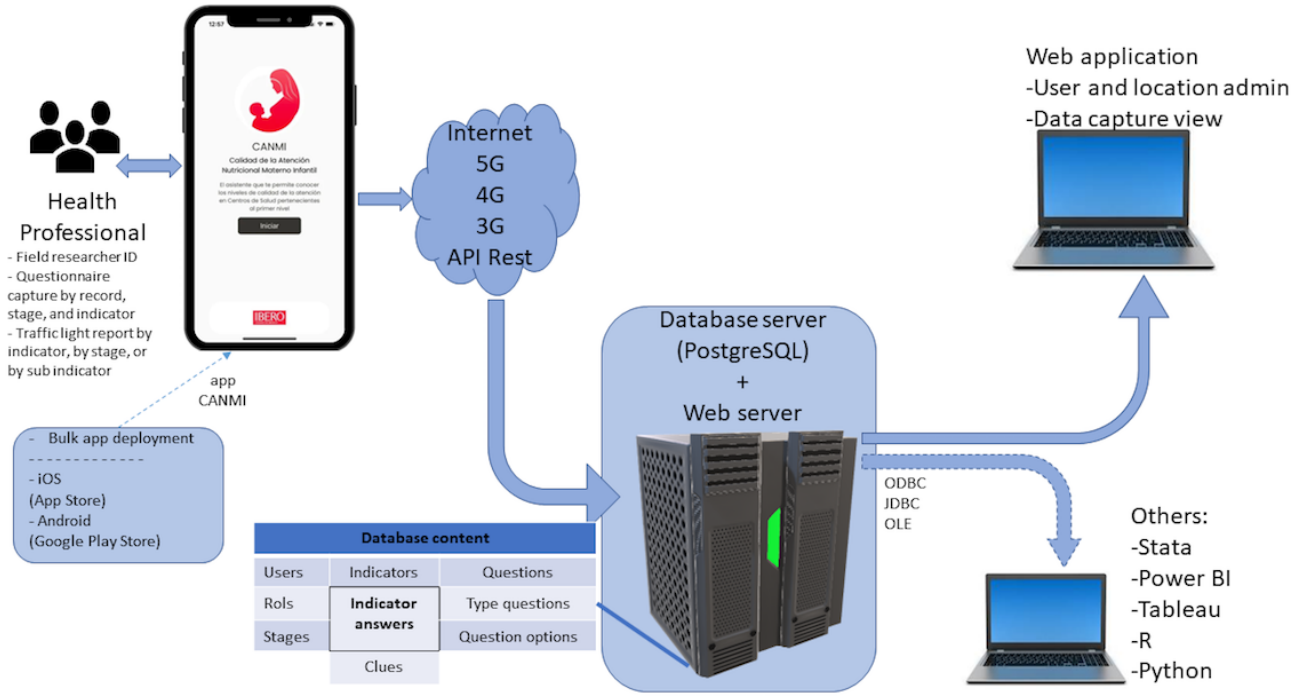

On the server side, a centralized database stores all responses and generates quality indicators linked to the user, health center, life stage, and question type. The mobile app allows users to intuitively visualize these results and corresponding quality metrics. Furthermore, a web-based app is connected to the server to facilitate data review, ensuring accuracy and completeness. The server architecture also supports integration with external analytics and reporting tools—such as Stata (StataCorp LLC), Tableau (Tableau Software LLC), and R (R Foundation for Statistical Computing)—via Open Database Connectivity, Java Database Connectivity, or an Object Linking and Embedding interface, allowing for flexible data analysis and reporting.

Usability and Acceptance Evaluation

Study Design and Setting

This pilot usability study was conducted in July 2024 across 18 primary health care units in Health Jurisdiction 1 of Guanajuato, Mexico—a geographic division that groups regions to organize and coordinate health care units and services—covering both rural and urban settings. A proportional allocation strategy was used to ensure the representation from each setting, and simple random sampling was applied within each stratum.

Participants

Eleven dietitians assigned to the Health Jurisdiction 1 of Guanajuato were invited to participate by email. Inclusion criteria required being a health professional with formal training in nutrition and providing direct care to the population in primary health care services within this jurisdiction. Nine dietitians met these criteria and agreed to participate.

Implementation

Prior to the implementation, all participants provided written informed consent and subsequently completed a training program (T0: preimplementation assessment) conducted over 2 consecutive days. The training aimed to ensure consistent use of the CANMI app and to improve the accuracy of data collection.

During the first session, the app was installed on the participants' personal mobile devices, and an overview of its key functionalities, navigation structure, and the content of the 16 validated quality indicators was provided. In the second session, participants engaged in guided practice using real-world data extracted from existing patient records. A total of 150 records were reviewed, comprising 30 records from each life stage. All exercises were performed in offline mode, with subsequent synchronization of data once a network connection became available. To assist participants, a detailed user guide was distributed, which included instructions on app installation, indicator navigation, and data submission procedures.

Software refinements were implemented iteratively based on feedback received during the training phase. Following the training, participants implemented the app in one of the 18 assigned health care units to assess the quality of maternal and child nutritional care. This implementation took place during the 6 weeks after the training period.

Data Collection and Analysis

A mixed methods approach was used to evaluate the usability, efficiency, and user satisfaction associated with the CANMI app among participants. Usability was quantitatively assessed using the System Usability Scale (SUS) validated by Sevilla-González et al [19]. The SUS consists of a 10-item questionnaire, where each item is rated on a 5-point Likert scale ranging from 1 ("strongly disagree") to 5 ("strongly agree").

After the training, all participants received a link to complete the SUS questionnaire via Google Forms. SUS scores were calculated on a scale from 0 to 100, where higher scores indicate

XSL•FO
RenderX

greater usability. A CI was also estimated [20]. Statistical analysis was performed using Stata (version 14.2).

Qualitative data were collected through 14 semistructured interviews with the participants. The interview guides were designed from a phenomenological approach to explore their experiences and perceptions regarding app usability [21]. To capture changes throughout the implementation process, 2 different interview guides were used. The T0 guide included contextual questions related to the participants' workplaces, their professional experience, satisfaction with the training, and expectations regarding the use of the CANMI app. The T1 (postimplementation assessment) guide focused on the participants' actual user experiences, perceptions of the app's functionality, and the factors influencing the app's adoption in their daily practice. This sequential design facilitated a structured comparison between anticipated and actual experiences, thereby enhancing the depth and robustness of the qualitative findings.

All 9 participants were invited to participate voluntarily via email at both time points. One participant did not respond at either time point (T0 and T1), and 3 others did not respond at T1. Therefore, 8 interviews were conducted at T0, and 5 were conducted at T1, with 5 nutrition personnel participating at both time points. All interviews were conducted remotely, via an online communication platform (n=12) and 2 by telephone. All interviews were conducted in Spanish, the primary language of the participants. Translations into English were performed by the research team, and accuracy was verified through a back-translation process reviewed by bilingual experts. The average duration of the interviews was 32 (SD 5) minutes.

With informed consent, interviews were audio-recorded to ensure information fidelity. Transcripts were coded using a deductive framework based on predefined categories, where the information was extracted directly from the audio recordings. To ensure confidentiality, all transcripts were anonymized by removing any personal identifiers. A thematic analysis [22] was conducted, addressing questions regarding app usability, facilitators of use, barriers to use, and improvements in the quality of nutritional care processes resulting from the app's use.

We used the SQUIRE (Standards for Quality Improvement Reporting Excellence) reporting guideline to draft this manuscript and the SQUIRE reporting checklist when editing, which are included in Multimedia Appendix 1 [23,24].

Ethical Considerations

This study was reviewed and approved by the Research Ethics Committee of the Universidad Iberoamericana in Mexico City (172/2022). Additionally, all procedures related to the development, training, and evaluation of the app implementation adhered to the procedures established by the authorities of the Ministry of Health of the state of Guanajuato.

Prior to participating in the training activities and interviews, all participants provided a written informed consent, which clearly stated the objectives of the study, the procedures involved, the intended use of the data, and how confidentiality would be protected. All data collected were deidentified to ensure participant privacy and confidentiality. No personal identifiers were included in the dataset. Participants did not receive any form of compensation for their participation.

Results

App Development Outcomes

The CANMI app was developed as a mHealth tool to support the monitoring and evaluation of the maternal and child nutritional care quality within primary health care units across the 32 federal entities of Mexico. This app provides real-time assessment service quality across 5 key life stages—preconception, pregnancy, postpartum, infancy, and preschool age—based on 16 validated indicators.

To facilitate interpretation by users, the app uses a color-coded "traffic light" system: green indicates high-quality care (≥90% compliance), yellow indicates deficient quality (71% - 89%), and red denotes poor quality (≤70%). The CANMI app was programmed to operate on both iOS and Android platforms and is designed to function offline, ensuring utility in areas with limited or no internet connectivity. Selected screenshots from the app's user interface, which highlight its intuitive design and visual feedback mechanisms, are presented in Figure 2.

XSL•FO
RenderX

Figure 2. Output screens from the CANMI (Calidad de la Atención Nutricional Materno Infantil) app.

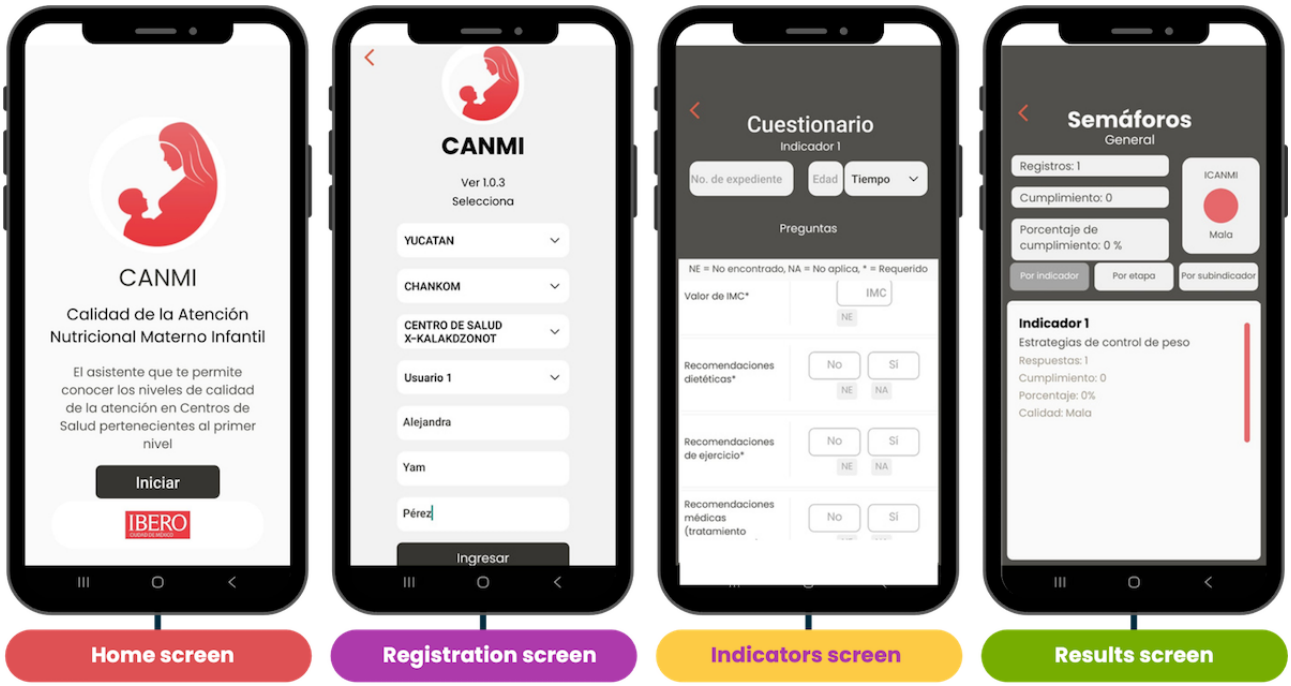

Usability of the CANMI App

Overall, user satisfaction with the app was positive. Participants described the app as intuitive, user friendly, and visually appealing, highlighting the clarity of its screens and the logical navigation structure. During interviews, participants emphasized that the icons representing each life stage helped them quickly identify relevant indicators, which facilitated efficient data entry.

However, during the initial implementation phase, some users experienced confusion when selecting indicators. This confusion primarily stemmed from their unfamiliarity with the structure and content of the 16 indicators designed to assess the quality of maternal and child nutritional care at the primary care units. As a result, there were occasional misclassifications of data under incorrect indicators. To address this issue, a concise guide outlining all indicators and subindicators by life stage was developed and distributed. This significantly improved both the accuracy and ease of use in subsequent data capture at the centers.

Participants reported that the app's interface and core functionalities were appropriate and did not require substantial adjustments. However, some suggestions for improvement were offered, including (1) limiting data entry access to only the relevant life stage indicators based on the user's patient population to reduce errors and (2) enabling batch submission of multiple records, rather than requiring each entry to be uploaded individually.

Regarding data submission, most users were able to use this function without complications. Only 2 (18%) participants did not use this function due to unreliable internet access at their assigned health centers. In contrast, 1 (9%) participant mentioned that internet connectivity was made available at personal expense by health care personnel. A single technical issue was reported when a user attempted to collect data across 2 facilities; upon uploading both records, the first was overwritten. This issue was reported and subsequently resolved through an app update.

However, participants noted that the current process of uploading records one by one—particularly in cases with high volumes (eg, 90 entries)—was time-consuming, although each submission was quick and straightforward. A bulk upload feature was recommended to streamline this process.

Regarding perceived effectiveness, one of the app's most valued attributes was its capacity to generate immediate results, contributing to time efficiency in quality assessment. However, some participants were unaware of this feature due to gaps in training and instead expected the jurisdiction to provide the results.

These qualitative findings were consistent with quantitative usability results. The mean SUS score was 71.13 (SD 11.68) points (95% CI 62.42 - 81.95), classifying the app as acceptable. Notably, female participants reported slightly higher usability scores (mean 72.18, SD 9.87) compared to male participants (mean 68.33, SD 12.11; Table 1).

XSL•FO
RenderX

Table . Total score and scores by sex on the usability scale of the CANMI (Calidad de la Atención Nutricional Materno Infantil) app (N=11).

	Participants, n (%)	Points, mean (SD)	Score, 95% CI
Total	11 (100)	71.13 (11.68)	62.42-81.95
By sex			
Male	3 (27)	68.33 (12.11)	55.40-81.26
Female	8 (73)	72.18 (9.87)	62.42-81.95

Facilitators of the Use

The main facilitator identified by participants was the app's ability to function without an internet connection during data collection. This feature is especially valuable given the frequent lack of internet access in health centers, surrounding communities, and even some municipalities. By eliminating the need for personal data plans, this functionality helps prevent user discouragement and promotes sustained use of the app:

> *The great advantage is that it can be used without the internet, it's very good because there are places where we don't have signal.* [Dietitian, Health Jurisdiction, Guanajuato]

> *It's easy to use the application because you don't need the internet to collect the data, only when uploading the information.* [Dietitian, Guanajuato]

One of the key advantages identified by participants was the app's potential to improve time efficiency during routine supervisory activities. In the context of Guanajuato, dietitian professionals conduct quarterly supervision of physicians and nurses, assessing quality indicators by the state health system, which can take up to a full day in each unit visited. Participants noted that integrating the CANMI app into these processes—either as the primary data collection tool or as a complementary resource—could streamline workflows by reducing the time required for manual data entry and enabling immediate and simplified transmission of results:

> *There's a lot of workloads, many programs, and many indicators. If we had this app in Guanajuato to measure our indicators by unit and at the municipal level, it would be great because you just record the data, and it automatically calculates the indicator.* [Dietitian, Guanajuato]

Another reported benefit of the app was that it can help make immediate improvements in nutritional care. Participants noted that the training itself, combined with continuous use of the tool, encouraged self-reflection and greater awareness of care gaps, particularly regarding missed counseling opportunities and incomplete documentation in medical records. One of the dietitians emphasized that the app could also serve as a self-assessment tool for physicians and nurses, fostering behavior change and improved practices across the broader care team:

> *I do see the app (CANMI app) as something that can help us improve the quality of care, but I feel that all staff should be involved because only the dietitians were given the app, and we are not in all the UMAPS (Primary Health Care Medical Units), I can't go to the field every day, it's impossible. Doctors and nurses should also use it because they need to evaluate themselves, 'hey, I'm not recording how much dose,' now it's like 'here are the files, and you do it.' It's important to involve them to keep improving.* [Dietitian, Guanajuato]

Barriers to Use

Several barriers were identified that may complicate the optimal use of the app. A key concern was the reliance on personal mobile devices for quality assessment. While most participants reported no major issues using their phones, some expressed discomfort and noted that this could discourage broader adoption among other health professionals. The lack of institutional devices was seen as a limitation, particularly because using the app temporarily disables access to messages and notifications, generating anxiety for some users.

Additionally, the prolonged use of the app—averaging 60 minutes per session—was described as physically tiring, especially when completing multiple entries consecutively:

> *It's very tiring to do it on the cell phone, I mean, it's fine, but after a while, it gets tiring because it's small, it would be better on a tablet or a computer ... The ideal would be for the health center to provide the equipment, because I have e-books (on my phone) and in training, they suggested deleting some (so the app would work better), and I said "Oh, no, it doesn't seem right to me, it's a book I can't erase," but if they gave us exclusive equipment for data entry, it would be fabulous.* [Dietitian, Guanajuato]

Another barrier to app use was the mismatch between the specificity of the app's indicators and the way health personnel typically document patient records. All interviewees mentioned that clinical records often lack detailed information, such as recommended dosages or specific actions. A frequently cited example was the indicator for physical activity counseling in preschool-aged children, which requires documentation of both the type and duration of activity. In practice, staff often record generic phrases such as "recommended exercise" or "recommended physical activity." As a result, half of the interviewees who referenced this example marked the indicator as completed in the app, while the other half did not. These inconsistencies were often noted in the app's observation field:

> *I had no hesitation in marking the fulfillment of the very specific indicators. For example, if the medical note had a general recommendation, I put in the app that it didn't meet the requirement, because "recommended healthy eating" is not adequate*

XSL•FO
RenderX

> *guidance. Then I would note in the observations that it was general.* [Dietitian, Guanajuato]
>
> *I found general notes like "recommended correct eating" or "continue with the same eating plan" or "promote physical activity," that's why in the app sometimes we ended up with files where everything was marked as "no." But sometimes it happens that it's not that health staff don't give complete or accurate recommendations, but we have formats that are closed and don't allow that wording* [Dietitian, Guanajuato]

A similar issue was found with postpartum indicators, such as breastfeeding guidance, which requires specifying recommendations for latch or breast massage. However, clinical notes from medical, nursing, and even nutrition staff typically document only general statements such as "breastfeeding guidance provided," limiting accurate reporting:

> *The app asks very specifically, and we only put "breastfeeding recommendation," but it's not specified point by point what we told them because it would take us more time to fill out the note, and since we don't always have the printer at the health center or there's no ink, we have to do it by hand, and that takes more time.* [Dietitian, Guanajuato]

Limited time available was identified as a key challenge for using the app. Although entering indicators took only an average of 5 to 10 minutes per record, depending on the life stage, coherence, or legibility of notes, completing the target of 30 records per stage, along with transportation and files retrieval for some users, required 1 or 2 full days of work. This pace of use significantly affected their patient agenda. This workload can disrupt clinical schedules, representing a barrier to integrating app use into daily routines.

Improvements in the Maternal and Child Nutritional Care Through CANMI App Use

The use of the app has led to improvements in maternal and child nutritional care, as reported by dietitians. Some participants noted that the app facilitated more precise and clear recommendations, improving user understanding and the accuracy of documentation. For example, instead of simply recording "supplementation recommended," they now document "recommended folic acid at this dosage," or instead of "breastfeeding counseling provided," they specify "guidance on latch provided."

In addition, efforts to standardize documentation practices were initiated among nursing staff. One participant proactively engaged with 2 of the 3 nurses working at a health center, providing guidance on how to fill out notes more precisely and standardize records. This initiative was framed as a means to improve the quality of care. Participants expressed the expectation that these improvements would ultimately translate into better health outcomes for women and children. Figure 3 summarizes the main elements identified in each of the analyzed categories: usability, facilitating factors, barriers to the use, and improvements.

XSL•FO
RenderX

Figure 3. Key elements identified in each category analyzed about the CANMI (Calidad de la Atención Nutricional Materno Infantil) app.

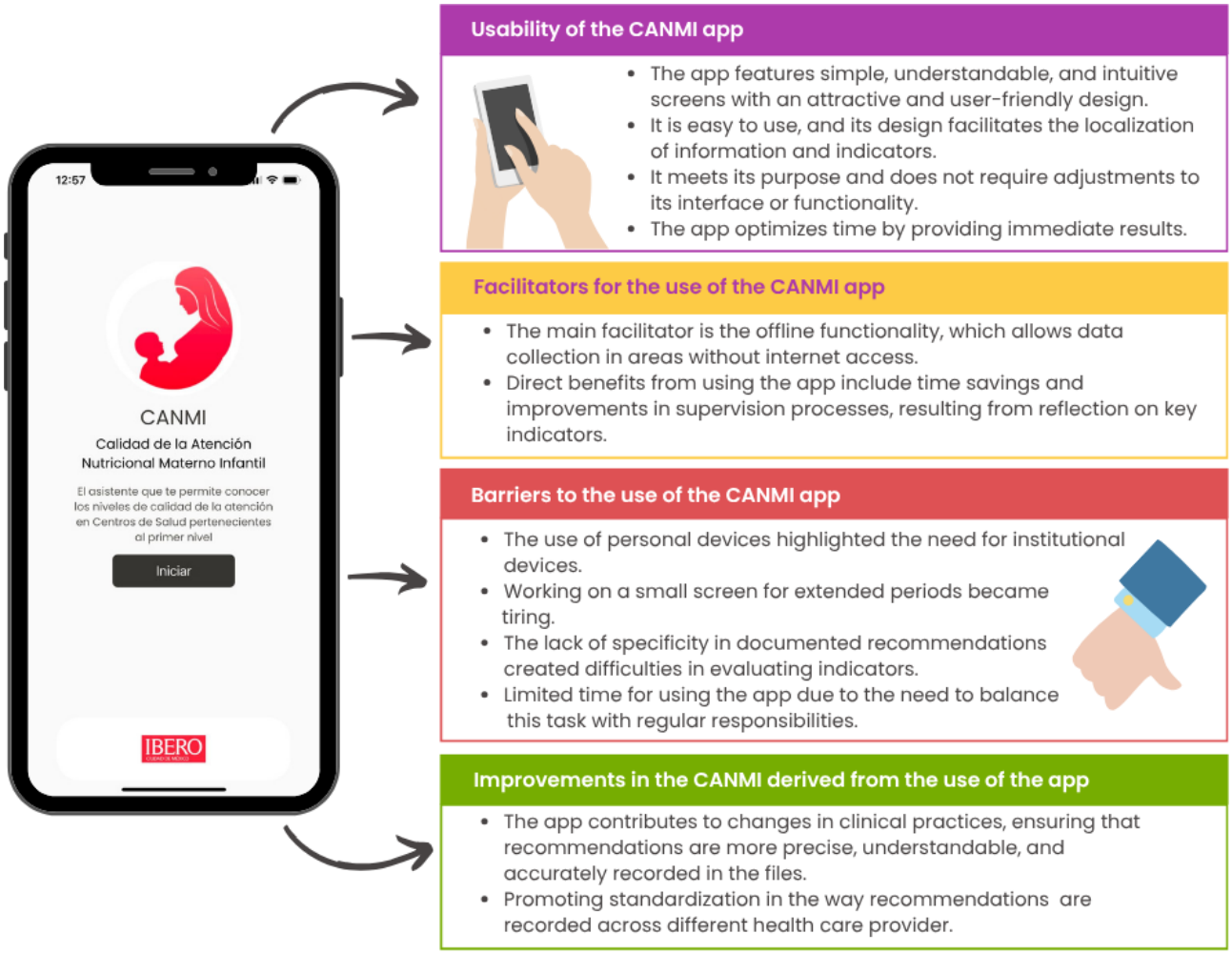

Discussion

Principal Findings

The CANMI app is a tool developed in response to the specific needs of the Mexican health care system, designed to monitor and evaluate the quality of nutritional care across 5 life stages: preconception, pregnancy, postpartum, infancy, and preschool years. Its design considers the availability of local resources within health care units for effective use, such as the lack of internet access in health care facilities, enabling data collection in areas without access to this service. This lack of internet access has been identified as one of the main factors hindering the implementation of information technologies in the field of mHealth, particularly in areas that lack internet access or have weak signals [25].

One key finding from the implementation of the CANMI app is that its interface is simple, understandable, and intuitive, with an attractive and user-friendly design. This is crucial for its successful adoption by health care professionals. The ease of use and accessibility of the interface are widely recognized as determining factors for the effectiveness of health technologies, especially in environments with diverse personnel and varying technological capabilities. This is supported by a systematic review by Jakob et al [26] who emphasized the importance of a user-friendly and technically stable app design as a positive factor for adherence to mHealth apps.

Regarding the effectiveness of the app in meeting its objective, it is important to note that feedback received indicates that no adjustments are necessary for its interface or functionality. This stability is essential to building trust among users, as frequent changes or functional issues can discourage the use of the tool [27,28]. Moreover, functional stability not only enhances the user experience but also facilitates the ongoing evaluation of the tool, as the data collected are more consistent and reliable.

Visual fatigue due to the use of small screens is a specific problem that adds to broader concerns about the usability of the CANMI app and was identified as a key barrier to the routine adoption of the CANMI app by health care professionals. Additionally, a lack of standardization in documenting recommendations and instructions within clinical records was identified. While this does not cause significant difficulties in recording the variables corresponding to each indicator, this lack of alignment between the information recorded in medical notes and the app's indicators can lead to inconsistencies in evaluation results. This finding aligns with what was noted by Juárez et al [29], who documented that standardization of information in clinical records in Mexico faces considerable challenges, particularly in terms of interoperability and data quality. In this context, it is crucial to implement strategies that

XSL•FO
RenderX

improve the quality of medical records, with continuous training for health care personnel being a key tool in achieving this improvement [30].

Following the analysis of usability and acceptance of the app in the health care units, it was identified that the tool can contribute to generating positive changes in how health care professionals provide nutritional care services. They reported being more rigorous and specific in their records and recommendations. This finding aligns with what Saez et al [18] documented, highlighting that monitoring data on maternal and child populations helps health care workers supervise and manage care more effectively, enabling them to make key decisions to improve the health of both mother and child. Similarly, in Ethiopia, the implementation of mobile solutions in the primary health care system has improved the timely identification and registration of pregnant women and adherence to treatment protocols [31]. However, it is important to emphasize that changes in care and nutritional status do not occur automatically and require ongoing intervention and detailed follow-up.

Limitations

This study has several limitations that should be considered when interpreting the results. First, the participant sample was small and limited to health care units in specific regions, which restricts the generalizability of the findings at the state level. Furthermore, due to the study design, the long-term impact of the CANMI app on nutritional indicators could not be accurately measured, limiting the ability to assess its sustainable effectiveness. Finally, evaluations should ideally be conducted by external personnel, but the validation process relied on health care professionals from the same units.

Future steps should focus on enhancing the app's functionality, prioritizing adjustments that improve its efficiency and usability within the context of primary health care in Mexico. Besides the contribution in monitoring maternal and child care, it is essential to implement systematic training for all health care personnel (physicians, nurses, and dietitians), leading to the improvement of health outcomes for the mother-child dyad.

Conclusions

The CANMI app has the potential to be a valuable tool in improving maternal and child nutritional care, especially in contexts where internet connectivity is limited. Although there are numerous facilitators of use, identified barriers such as the lack of specificity in medical records, the use of personal devices, and additional workload must be addressed to ensure its long-term effectiveness and sustainability. Overall, the app has shown that it can have a positive impact on improving record-keeping practices and raising staff awareness, contributing to more accurate and efficient care. However, it is essential to continue monitoring its use and adjust based on the needs and experiences of health care personnel to maximize its potential.

Acknowledgments

The authors express their gratitude to Dr Juan Carlos González Araiza, Head of Health Jurisdiction I in Guanajuato, for all the support provided to carry out the training at the General Hospital of Dolores Hidalgo C.I.N. To all health care providers, especially dietitians affiliated with the Health Jurisdiction of the state of Guanajuato (2021– - 2024), for their participation in the training. This research was funded by the Institute for Applied Research and Technology (INIAT) at Universidad Iberoamericana, Mexico City. MA-M acknowledges the support of the International Sabbatical Stay Support Fund provided by the Universidad Iberoamericana Mexico City. This study was funded by the Institute of Applied Research and Technology (INIAT) at Universidad Iberoamericana, Mexico City (0050-INIAT).

Data Availability

The datasets generated and analyzed during this study are not publicly available because participant consent limited data use to research purposes under strict confidentiality. Anonymized data may be requested from the corresponding author upon reasonable request.

Authors' Contributions

Study conception and design: MA-M
Development app: JEQ-I, JAG-O
Data collection: EH-L, CPN, DTS
Data analysis: CPN, EH-L
Drafting of the manuscript: EH-L, SB-M
Critical revisions for important intellectual content: IO-G, OA-R, ACB-M, SH-C, CM-M
Supervision: MA-M, SB-M

Conflicts of Interest

None declared.

Multimedia Appendix 1

XSL•FO
RenderX

Interview guide.
[DOCX File, 25 KB - formative_v9i1e77539_app1.docx]

References

1. Comprehensive implementation plan on maternal, infant and young child nutrition. World Health Organization. 2014 May 19. URL: https://www.who.int/publications/i/item/WHO-NMH-NHD-14.1 [accessed 2025-10-01]
2. Observatorio Materno Infantil Indicadores – OMI. URL: https://omi.ibero.mx/indicador-sobrepeso-y-obesidad/ [accessed 2023-11-14]
3. Mejía-Rodríguez F, Mundo-Rosas V, García-Guerra A, et al. Prevalencia de anemia en la población mexicana: análisis de la Ensanut Continua 2022. Salud Publica Mex 2023 Jun 14;65:s225-s230. [doi: 10.21149/14771] [Medline: 38060969]
4. Campos-Nonato I, Galván-Valencia Ó, Hernández-Barrera L, Oviedo-Solís C, Barquera S. Prevalencia de obesidad y factores de riesgo asociados en adultos mexicanos: resultados de la Ensanut 2022. Salud Publica Mex 2023 Jun 14;65:s238-s247. [doi: 10.21149/14809] [Medline: 38060949]
5. Norma Oficial Mexicana NOM-007-SSA2-2016, para la atención de la mujer durante el embarazo, parto y puerperio, y de la persona recién nacida. Comisión Nacional de los Derechos Humanos. 2016. URL: https://www.cndh.org.mx/sites/all/doc/Programas/VIH/Leyes%20y%20normas%20y%20reglamentos/Norma%20Oficial%20Mexicana/NOM-007-SSA2-2016%20Embarazo,%20parto%20y%20puerperio.pdf [accessed 2025-04-20]
6. Berwick DM. A user's manual for the IOM's "Quality Chasm" report. Health Aff (Millwood) 2002;21(3):80-90. [doi: 10.1377/hlthaff.21.3.80] [Medline: 12026006]
7. Murray CJ, Frenk AJ. A framework for assessing the performance of health systems. Bull World Health Organ 2000;78(6):717-731. [Medline: 10916909]
8. Crossing the Quality Chasm: A New Health System for the 21st Century: National Academies Press; 2001. URL: https://doi.org/10.17226/10027 [doi: 10.17226/10027]
9. Ancira-Moreno M, Omaña-Guzmán I, Bautista-Morales AC, et al. Development and validation of a new set of indicators to assess the quality of maternal and child nutritional care at the primary care. Front Med (Lausanne) 2022 Dec 7;9:1011940. [doi: 10.3389/fmed.2022.1011940] [Medline: 36569141]
10. Julio V, Vacarezza M, Álvarez C, Sosa A. Niveles de atención, de prevención y atención primaria de la salud. Arch Med Int 2011;33(1) [FREE Full text]
11. Ruiz OA, Ancira-Moreno M, Omaña-Guzmán I, et al. Low quality of maternal and child nutritional care at the primary care in Mexico: an urgent call to action for policymakers and stakeholders. Int J Equity Health 2024 Feb 22;23(1):35. [doi: 10.1186/s12939-024-02129-z] [Medline: 38388936]
12. Hoyos-Loya E, Pérez Navarro C, Burrola-Méndez S, et al. Barriers to promoting breastfeeding in primary health care in Mexico: a qualitative perspective. Front Nutr 2024 Jan 9;10:1278280. [doi: 10.3389/fnut.2023.1278280] [Medline: 38264191]
13. Ossebaard HC, Van Gemert-Pijnen L. eHealth and quality in health care: implementation time. Int J Qual Health Care 2016 Jun;28(3):415-419. [doi: 10.1093/intqhc/mzw032] [Medline: 27029590]
14. Stampe K, Kishik S, Müller SD. Mobile health in chronic disease management and patient empowerment: exploratory qualitative investigation into patient-physician consultations. J Med Internet Res 2021 Jun 15;23(6):e26991. [doi: 10.2196/26991] [Medline: 34128817]
15. da Silva RR, da Silva Campos J, Vieira SL, et al. The impact of technologies on quality in the provision of public health services. Rev Gest Soc Ambient 2024;18(10):e09149. [doi: 10.24857/rgsa.v18n10-153]
16. Orton M, Agarwal S, Muhoza P, Vasudevan L, Vu A. Strengthening delivery of health services using digital devices. Glob Health Sci Pract 2018 Oct 10;6(Suppl 1):S61-S71. [doi: 10.9745/GHSP-D-18-00229] [Medline: 30305340]
17. Encuesta nacional sobre disponibilidad y uso de tecnologías de la información en los hogares (ENDUTIH): reporte de resultados. Instituto Nacional de Estadística y Geografía (INEGI). URL: https://www.ift.org.mx/comunicacion-y-medios/comunicados-ift/es/encuesta-nacional-sobre-disponibilidad-y-uso-de-tecnologias-de-la-informacion-en-los-hogares-endutih-1 [accessed 2025-10-13]
18. Sáez C, Moner D, García-De-León-Chocano R, et al. A standardized and data quality assessed maternal-child care integrated data repository for research and monitoring of best practices: a pilot project in Spain. Stud Health Technol Inform 2017;235:539-543. [Medline: 28423851]
19. Sevilla-Gonzalez MD, Moreno Loaeza L, Lazaro-Carrera LS, et al. Spanish version of the system usability scale for the assessment of electronic tools: development and validation. JMIR Hum Factors 2020 Dec 16;7(4):e21161. [doi: 10.2196/21161] [Medline: 33325828]
20. Vlachogianni P, Tselios N. Perceived usability evaluation of educational technology using the system usability scale (SUS): a systematic review. J Res Technol Educ 2021 Feb 23;54(3):392-409. [doi: 10.1080/15391523.2020.1867938]
21. Fuster Guillen D. Investigación cualitativa: método fenomenológico hermenéutico. Propós Represent 2019;7(1):201-229 [FREE Full text]
22. Patton MQ. Qualitative Research & Evaluation Methods, 3rd edition: SAGE Publications; 2001. [doi: 10.4135/9781529770278.n71]

XSL•FO
RenderX

23. Ogrinc G, Davies L, Goodman D, Batalden P, Davidoff F, Stevens D. SQUIRE 2.0 (Standards for QUality Improvement Reporting Excellence): revised publication guidelines from a detailed consensus process. BMJ Qual Saf 2016 Dec;25(12):986-992. [doi: 10.1136/bmjqs-2015-004411] [Medline: 26369893]
24. Ogrinc G, Davies L, Goodman D, Batalden P, Davidoff F, Stevens D. The SQUIRE reporting checklist. In: Harwood J, Albury C, de BJ, Schlüssel M, Collins G, editors. The EQUATOR Network Reporting Guideline Platform: The UK EQUATOR Centre; 2025. URL: https://www.equator-network.org/reporting-guidelines/
25. Shiferaw S, Workneh A, Yirgu R, Dinant GJ, Spigt M. Designing mHealth for maternity services in primary health facilities in a low-income setting - lessons from a partially successful implementation. BMC Med Inform Decis Mak 2018 Nov 12;18(1):96. [doi: 10.1186/s12911-018-0704-9] [Medline: 30419891]
26. Jakob R, Harperink S, Rudolf AM, et al. Factors influencing adherence to mHealth apps for prevention or management of noncommunicable diseases: systematic review. J Med Internet Res 2022 May 25;24(5):e35371. [doi: 10.2196/35371] [Medline: 35612886]
27. Casare A, Silva CG, Moraes R. Do dependable systems need good user interfaces? 2022 Nov 21 Presented at: 11th Latin-American Symposium on Dependable Computing (LADC '22); Nov 21-24, 2022; Fortaleza/CE Brazil URL: https://dl.acm.org/doi/proceedings/10.1145/3569902 [accessed 2025-10-01] [doi: 10.1145/3569902.3569905]
28. Zieglmeier V, Lehene AM. Designing trustworthy user interfaces. 2021 Nov 30 Presented at: 33rd Australian Conference on Human-Computer Interaction (OzCHI '21); Nov 30 to Dec 2, 2021; Melbourne, Australia URL: https://dl.acm.org/doi/proceedings/10.1145/3520495 [accessed 2025-10-01] [doi: 10.1145/3520495.3520525]
29. Juárez IA, Nicanor LD, De la Vega JA, Hernández JE. API design based on HL7 standard to interoperability of medical information systems. Presented at: 11th International Conference On Software Process Improvement (CIMPS 2022); Oct 19-21, 2022; Acapulco, Mexico. [doi: 10.1109/CIMPS57786.2022.10035566]
30. Avendaño-Badillo D, Israel M, Luis Guillermo I. Registry of the functioning of patients in the electronic clinical record at the national institute of rehabilitation of Mexico. Ann Phys Rehabil Med 2018 Jul;61:e547-e548. [doi: 10.1016/j.rehab.2018.05.1275]
31. Nigussie ZY, Zemicheal NF, Tiruneh GT, et al. Using mHealth to improve timeliness and quality of maternal and newborn health in the primary health care system in Ethiopia. Glob Health Sci Pract 2021 Sep 30;9(3):668-681. [doi: 10.9745/GHSP-D-20-00685] [Medline: 34593589]

Abbreviations

CANMI: Calidad de la Atención Nutricional Materno Infantil
mHealth: mobile health
SQUIRE: Standards for Quality Improvement Reporting Excellence
SUS: System Usability Scale

Edited by A Stone, A Mavragani; submitted 15.05.25; peer-reviewed by C Okerulu, JY Santaolalla; revised version received 18.09.25; accepted 19.09.25; published 20.10.25.

Please cite as:
Burrola-Méndez S, Hoyos-Loya E, Quiroz-Ibarra JE, Perez Navarro C, Omaña-Guzmán I, González-Ordiano JÁ, Acosta-Ruiz O, Bautista-Morales AC, Tánchez Sandoval D, Hernández-Cordero S, Muñoz-Manrique C, Ancira-Moreno M
Development, Implementation, and Usability Evaluation of the CANMI App to Monitor the Quality of Maternal and Child Nutrition Care in Primary Health Units: Mixed Methods Pilot Study
JMIR Form Res 2025;9:e77539
URL: https://formative.jmir.org/2025/1/e77539
doi:10.2196/77539

© Soraya Burrola-Méndez, Elizabeth Hoyos-Loya, J Emilio Quiroz-Ibarra, Cecilia Perez Navarro, Isabel Omaña-Guzmán, Jorge Ángel González-Ordiano, Omar Acosta-Ruiz, Arturo C Bautista-Morales, Daniela Tánchez Sandoval, Sonia Hernández-Cordero, Cinthya Muñoz-Manrique, Monica Ancira-Moreno. Originally published in JMIR Formative Research (https://formative.jmir.org), 20.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Transition From a Paper Diary to an Electronic Diary by Parents of Preschool Children With Food Allergies: Pilot Study

Yoriko Kato, MN

School of Nursing, Sapporo City University, North 11 West 13, Chuo-ku, Sapporo, Hokkaido, Japan

Corresponding Author:
Yoriko Kato, MN
School of Nursing, Sapporo City University, North 11 West 13, Chuo-ku, Sapporo, Hokkaido, Japan

Abstract

Background: Food allergies (FAs) can be life-threatening and represent a significant public health concern. The reported prevalence of FA in Japan is 7.6% in 1-year-olds, 6.7% in 2-year-olds, and 4.9% in 3-year-olds. In addition, avoiding allergenic foods, prompt identification, and response are required when symptoms appear. A food diary is a tool for daily dietary management, recording of symptoms, and communication between families of children with FA and health care providers. Conventionally, paper diaries (P-diaries) have been used. Recent technological advancements and the extensive use of smartphones have prompted the development of numerous electronic diaries (E-diaries). However, few studies have evaluated the effectiveness of E-diaries in Japan.

Objective: This study aimed to assess the experiences of parents of preschool children with FA as they transitioned from P-diaries to E-diaries.

Methods: For 6 months commencing in October 2020, during the period of movement restrictions due to the COVID-19 pandemic, we recruited parents of preschool children diagnosed with FAs by an allergist at a single general hospital. This pilot study used a single-group pretest–posttest design to evaluate the feasibility of using an E-diary. To assess parents' views on transitioning to E-diaries from P-diaries, data were collected using semistructured interviews conducted after 6 months of E-diary use.

Results: Five parents took part in the study. The problems with P-diaries and the effects and outcomes of E-diaries from the perspective of parents of preschool children with FA were categorized into 52 codes, which were grouped into 19 subcategories and 6 categories. Parents of preschool children with FA reported the following issues: *I. difficulties with using a P-diary*; *II. stress related to sharing information with doctors*; *III. difficulties in making disease management habitual as issues with P-diary use*; *IV. ease of using the E-diary as an effect of the transition*; *V. more efficient sharing of information with doctors*; and *VI. improved adherence as outcomes of E-diary use*.

Conclusions: E-diaries may resolve various issues associated with P-diaries and may be useful tools for communication between parents and physicians regarding disease management. Since the number of participants was small (ie, 5) and the study was conducted at a single facility, theoretical saturation was not reached. In the future, it will be necessary to increase the number of participants to validate these findings.

(JMIR Form Res 2025;9:e73400) doi:10.2196/73400

KEYWORDS

electronic diary; food allergies; paper diary; parents; preschool children

Introduction

Overview

Food allergies (FAs) are a public health problem [1] with a prevalence of 7.6% in 1-year-olds, 6.7% in 2-year-olds, and 4.9% in 3-year-olds in Japan [2]. FA is defined as "an adverse health effect arising from a specific immune response that occurs reproducibly on exposure to a given food" [3].

FA is a serious and life-threatening disease associated with anxiety, worry, and decreased quality of life for children and their families. The difficulty in avoiding causative foods and the unpredictable nature of FA-induced symptoms are factors affecting children with FA and their families. FA requires daily management and constant monitoring, and the chronic and unpredictable nature of FA may affect the psychosocial functioning and quality of life of the child and family [4].

Oral immunotherapy (OIT) is a promising treatment for FA [5]. Studies have shown that the completion of OIT or reaching its maintenance phase significantly improves the quality of life of both the child and parents [6,7]. However, OIT is a challenging process for patients and parents. Induction can generate stress and anxiety. This, in turn, may affect their motivation and ability

XSL•FO
RenderX

to cope with OIT challenges [8]. OIT is administered over an extended period, ranging from several months to several years, and requires daily consumption of food that patients were taught to avoid and fear. Also, patients commonly have oral aversion to the allergenic food, making regular consumption more difficult [9]. Moreover, adverse reactions are often experienced during treatment, ranging from mild symptoms to anaphylaxis [5].

The occurrence of frequent adverse reactions poses a challenge, necessitating close monitoring during treatment. The OITcontrol app (Mylan EPD) appears to be a valuable tool for monitoring treatment in children with FAs. It is a suitable method for recording daily intake and reactions, promoting adherence to treatment instructions following adverse events, as well as compliance with dose adjustments for home reactions [10].

eHealth monitoring technologies offer opportunities to objectively assess symptoms when they appear in daily life [11]. The convenience and accessibility offered by digital health technologies in health care have the potential to transform the health care landscape by enhancing communication, improving efficiency, and empowering patients [12]. In addition, the use of patient-centered digital health records in nonhospitalized individuals with chronic health conditions is potentially associated with considerable benefits for health care use, treatment adherence, and self-management or self-efficacy [13]. However, limited research has been conducted on the use of electronic diaries (E-diaries) for FA.

Retrospective questionnaires have been used for decades to assess the severity and control of allergic diseases. Smartphone apps have recently facilitated the use of prospective clinical diaries based on daily patient-completed questionnaires [14].

FA diaries record the types and amounts of food and drink consumed, as well as the types and severity of induced symptoms. E-diaries may make it easier for parents to inform doctors of FA-induced symptoms by using the camera function on their smartphones, particularly when maintaining records of cutaneous symptoms. Most food diaries for FA in Japan are paper-based, and no confirmed experience exists with the use of electronic food diaries by parents.

A mobile health (mHealth) app (provided by Mylan EPD), released in Japan in September 2020, aimed to provide accurate medical information on FA and anaphylaxis, as well as to encourage the proper use of an adjuvant anaphylaxis drug (adrenaline self-injection).

Purpose

This study aimed to assess the experiences of parents of preschool children diagnosed with FA as they transitioned from a paper diary (P-diary) to an E-diary, and to evaluate their perceptions of the technology.

Methods

Study Design

This pilot study used a quasi-experimental one-group pretest-posttest design to evaluate the E-diary technology. To investigate the parents' views on transitioning from regular P-diaries to E-diaries, the authors used data obtained from semistructured interviews conducted 6 months postusage of E-diaries to evaluate their usefulness compared to that of P-diaries.

Participant Recruitment

Study participants who met the following criteria were recruited: (1) a parent of a preschool child receiving OIT and (2) someone with adequate information technology knowledge to use the app and transition from a P-dairy to an E-dairy using a smartphone. In addition, participants could send a PDF file summarizing the data of the E-diary to the researcher before the outpatient visit.

Intervention Method

The E-diary used in this study was designed for recording content, including allergy symptoms, and saving photographs. It was part of the MyEpi mHealth app (provided by Mylan EPD), released in Japan in September 2020, which provides accurate medical information on FA and anaphylaxis and encourages the proper use of adjuvant drugs for anaphylaxis (adrenaline self-injection). Written approval for the use of the MyEpi mHealth app in this study was obtained from Mylan EPD.

The MyEpi mHealth App Incorporates the Following 4 Functions

The MyEpi mHealth App includes the following 4 features:

1. Note function (E-diary): the user can record symptoms and test results, including photographs. These records can be printed or sent via email as PDFs.
2. FA library function: provides the user with accurate information on matters, including the mechanism of allergies and appropriate responses in case of anaphylaxis.
3. Q&A function: users can access answers to frequently asked questions on topics such as the causes, symptoms, and treatment of allergies.
4. Emergency voice navigation: in the event of anaphylaxis, the user can use this function to follow step-by-step video and voice instructions for administering the adjuvant anaphylaxis drug.

The Investigators Requested Study Participants Who Provided Informed Consent to Perform the Following 3 Actions

These 3 actions are given below:

1. Download the MyEpi mHealth app and explore its use by following the Operation Manual, which provides simple instructions on how to input E-diary entries and save photographs.
2. After trying out the app, attach the E-diary entries and photographs as PDFs to an email and send it to an investigator's personal email address to confirm operational feasibility.
3. Send the PDF data to the study email address 1 week before the scheduled outpatient appointment (Figure 1).

XSL•FO
RenderX

Figure 1. Overview of the MyEpi app. Source: Mylan EPD.

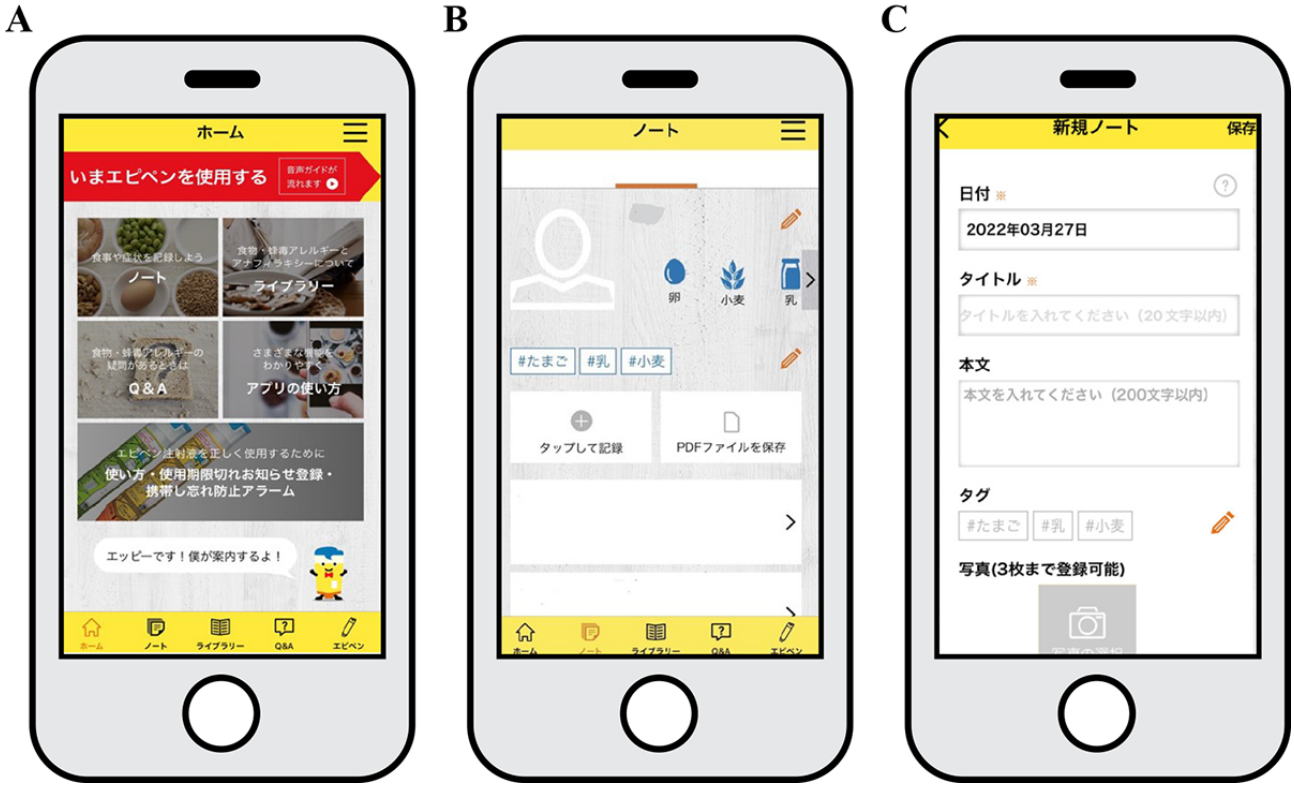

The investigators were responsible for forwarding the E-diary PDF data, emailed by the study participants, to the doctors.

Data Collection

To evaluate the E-diary after its introduction, the authors used data obtained from semistructured interviews conducted 6 months postusage of the E-diary to evaluate its usefulness compared with that of the P-diary. An interview guide was prepared in accordance with the study objectives. The investigators provided the participants with the content of the interviews. All the interviews were conducted by the lead author, who had received training in conducting regular interviews. During each interview, the interviewer encouraged participants to speak freely about their experiences using the P-diary and E-diary. The interviewer did not assist participants with any information regarding FA. Demographic data on the child's age, sex, age at onset of FA, and offending food were collected during the interview. The interview also covered participants' experiences with (1) using the P-diary and (2) using the E-diary. With the participants' consent, the interviews were recorded with a digital voice recorder.

Data Analysis

We used Berelson "Content Analysis" method for data analysis [15]. "Content Analysis" is a method used for the objective, systematic identification of characteristics in text to draw inferences, ultimately enabling quantitative analysis [16]. All recorded data were transcribed verbatim in Japanese. Content related to (1) the experience of using a P-diary and (2) the experience of using an E-diary was coded as a unit. Subcategories and categories were generated based on the semantic similarity of the codes. The study team included 2 pediatric nurses (KY and KM). After the initial coding of each transcript, the investigators discussed the data together and identified a series of main codes, subcategories, and categories. Reliability was assessed during the coding stage, and a calculated concordance rate of ≥90% between the 2 investigators was confirmed. Validity was assured through the supervision of a pediatric nurse who served as a joint investigator.

Ethical Considerations

This study was conducted with the approval of the Research Ethics Committee of the Faculty of Nursing and Social Work, Health Sciences University of Hokkaido (approval number 20N024030) and the Ethics Committee of the General Hospital.

The study participants who met the selection criteria were informed of the objective, importance, and methods of the study orally and in writing. The participants were informed by the investigators that they had the right to participate in the study of their own free will, including their right to withdraw at any time; that their data would be anonymized and confidentiality protected; how their data would be handled and disposed of; and the possibility of publication. Written informed consent was obtained. The study adhered to the COREQ (Consolidated Criteria for Reporting Qualitative Studies) guidelines [17]. Participants received a gift card worth 5000 Japanese yen as a research incentive.

Results

Participant Demographics

All 5 parents who were contacted by telephone agreed to take part in the study. All were mothers in their 30s (Table 1). After using the mHealth app for 6 months, each participant underwent a single interview lasting between 26 minutes 40 seconds and

XSL•FO
RenderX

60 minutes 30 seconds (M_{time}=40 min 29 s). These interviews were conducted between September and December 2021. Due to the COVID-19 pandemic, a single investigator specializing in childhood allergies conducted all the interviews via a videoconferencing system. The difficulties with P-diaries and the effects and outcomes of E-diaries were analyzed. The results of the analysis showed 52 codes were extracted, grouped into 19 subcategories and 6 categories. Categories are indicated in *italic* text, subcategories in *italic* text, code numbers in {curly brackets}, and percentages in (ordinary parentheses). The parents of preschool children with FA reported: I. difficulties with using a P-diary, II. stress related to information-sharing with doctors, and III. difficulties in making disease management habitual, as problems associated with the P-diary, while IV. ease of using the E-diary as a positive effect of the E-diary, V. more efficient information sharing with doctors, and VI. improved adherence, as favorable outcomes of the E-diary (Figure 2).

Table . Characteristics of children with food allergies whose parents continued using an electronic diary for 6 months (n=5).

Case	Ages at time of interview (years)	Sex	Age of onset (months)	Allergen
A	6 years 8 months	Boy	2	Wheat, Milk
B	3 years 10 months	Girl	6	Egg, Milk
C	6 years 1 month	Boy	5	Milk
D	6 years 5 months	Boy	11	Wheat, Milk
E	5 years 0 months	Girl	6	Milk

Figure 2. Problems associated with paper diary use and effects and outcomes of electronic diary use reported by parents of preschool children with food allergies. P-diary: paper diary; OIT: oral immunotherapy; E-diary: electronic diary.

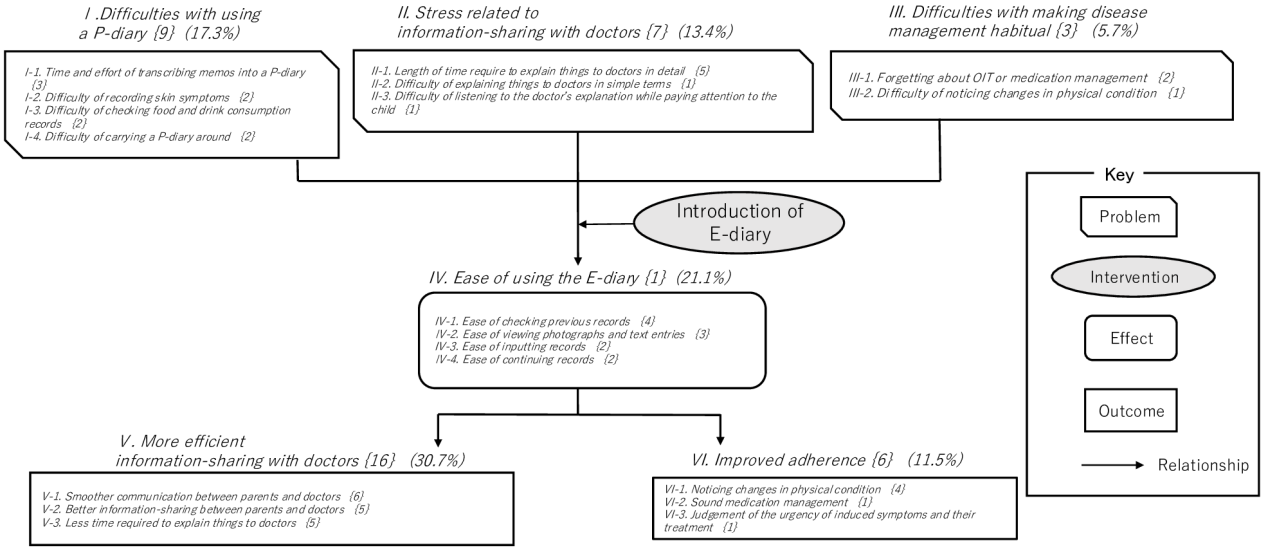

Difficulties Associated With the P-Diary

Problems associated with the P-diary were categorized into 19 codes, which were grouped into 9 subcategories and 3 categories (Table 2).

XSL•FO
RenderX

Table . Problems associated with the paper diary, recognized by parents of children with food allergy after using the electronic diary for 6 months.

Categories and subcategories		Representative codes	Codes
(I).Difficulties with using[a] P-diary			9 (17.3%)
	I-1. Time and effort of transcribing memos into a P-diary	• *Copying the records, I'd noted on a calendar into a memo was twice as much trouble.* [Case C, 30 s, mother of a 6-year-old boy] • *Before I started using the electronic food diary, I'd be rushing to organize the information to tell the doctor.* [Case E: 30 s, mother of a 5-year-old girl] • *Recording symptoms and daily behavior together took double the time and effort.* [Case E, 30 s, mother of a 5-year-old girl]	3
	I-2. Difficulty in recording skin symptoms	• *It's difficult to explain the condition of roughened skin to the doctor in words.* [Case A, 30 s, mother of a 6-year-old boy] • *The subtle skin symptoms, where the skin appears red and swollen around the mouth, ears, and navel, are difficult to record.* [Case E, 30 s, mother of a 5-year-old girl]	2
	I-3. Difficulty in checking food and drink consumption records	• *If there are several offending foods, they have to be eaten on different days, which means comparing the records for each offending food one by one.* [Case B, 30 s, mother of a 3-year-old girl] • *The electronic food diary enables me to check several different offending foods at once without getting them mixed up.* [Case B, 30 s, mother of a 3-year-old girl]	2
	I-4. Difficulty in carrying a P-diary around	• *If I'm using a paper-based medium for the food diary, the paper and notes get scattered and lost.* [Case A, 30 s, mother of a 6-year-old boy] • *When we went on holiday, it was very hard to take the paper-based food diary with us.* [Case A, 30 s, mother of a 6-year-old boy]	2
(II).Stress related to information-sharing with doctors			7 (13.4%)

XSL•FO
RenderX

Categories and subcategories		Representative codes	Codes
	II-1. Length of time required to explain things to doctors in detail	• *I'm always concerned about time during appointments.* [Case B, 30 s, mother of a 3-year-old girl] • *It was time-consuming for the doctor to input the electronic medical records, too, and it seemed like a lot of trouble.* [Case C, 30 s, mother of a 6-year-old boy] • *I used to give a detailed oral report to the doctor of everything that had happened since the last appointment.* [Case C, 30s, mother of a 6-year-old boy] • *At appointments, I would give the doctor in charge the records of the offending food recorded in a schedule planner, and give a detailed explanation while looking at the schedule planner.* [Case D, 30 s, mother of a 6-year-old boy] • *Before using the electronic food diary, I used to give a detailed explanation.* [Case E, 30 s, mother of a 6-year-old girl]	5
	II-2. Difficulty in explaining things to doctors in simple terms	• *I'm not good at conversation, so I can't explain things clearly to the doctor in charge, and my way of recording things also didn't work well, so I would get really antsy at appointments.* [Case A, 30 s, mother of a 6-year-old boy]	1
	II-3. Difficulty in listening to the doctor's explanation while paying attention to the child	• *I would pay attention to my son while talking to the doctor, so they didn't believe everything I said.* [Case C, 30 s, mother of a 6-year-old boy]	1
(III). Difficulties with making disease management habitual			3 (5.7%)
	III-1. Forgetting about OIT[b] or medication management	• *When I was using a paper-based food diary, if I didn't put it somewhere I could see it, I'd forget about OIT.* [Case A, 30 s, mother of a 6-year-old boy] • *If one of my children were taking medication, I'd forget to give medication to the other one.* [Case C, 30 s, mother of a 6-year-old boy]	2
	III-2. Difficulty in noticing changes in physical condition	• *He had a tendency to get sick at the changing of the seasons.* [Case C, 30 s, mother of a 6-year-old boy]	1

[a]P-diary: paper diary.

[b]OIT: oral immunotherapy.

XSL•FO
RenderX

Effects of the E-Diary

Aspects of the effects of the introduction of the E-diary were categorized 11 codes, which were grouped into 4 subcategories and 1 category (Table 3).

XSL•FO
RenderX

Table . Effects of paper diary recognized by parents of children with food allergy after using electronic diary for 6 months.

Categories and subcategories		Representative codes	Codes
(IV).Ease of using the E-diary[a]			11 (21.1%)
	IV-1. Ease of checking previous records	• *The electronic food diary made it easier to check the amount of the offending food consumed and the frequency of consumption.* [Case A, 30 s, mother of a 6-year-old boy] • *The electronic food diary is convenient because you can view the previous history whenever you want.* [Case B, 30 s, mother of a 3-year-old girl] • *The electronic food diary enables me to check several different offending foods at once without getting them mixed up.* [Case B, 30 s, mother of a 3-year-old girl] • *The electronic food diary makes it easy to look back over the course of symptoms.* [Case C, 30 s, mother of a 6-year-old boy]	4
	IV-2. Ease of viewing photographs and text entries	• *Text entries and photographs can be recorded in one go, and when they are printed out, the entries and photographs also come out together.* [Case E, 30 s, mother of a 5-year-old girl] • *The electronic food diary enables the unified management of records and photographs of symptoms.* [Case C, 30 s, mother of a 6-year-old boy] • *It was convenient to capture photographs as well as descriptions of what sort of symptoms had been induced and when.* [Case B, 30 s, mother of a 3-year-old girl]	3
	IV-3. Ease of inputting records	• *My smartphone is small enough to carry around the whole time, so I no longer go looking for the food diary.* [Case A, 30 s, mother of a 6-year-old boy] • *The electronic food diary makes it easy to input anything that has worried me, and the details of how I dealt with symptoms when they appeared.* [Case A, 30 s, mother of a 6-year-old boy]	2
	IV-4. Ease of continuing records	• *The electronic food diary has lowered the hurdles for record-keeping.* [Case A, 30 s, mother of a 6-year-old boy] • *The double time and effort involved in record-keeping disappeared.* [Case C, 30 s, mother of a 6-year-old boy]	2

XSL•FO
RenderX

[a]E-diary: electronic diary.

Outcomes of the E-Diary

Aspects of the outcomes of the E-diary were categorized into 22 codes, which were grouped into 6 subcategories and 2 categories (Table 4).

XSL•FO
RenderX

Table . Outcomes of paper diary recognized by parents of children with food allergy after using electronic diary for 6 months.

Categories and subcategories		Representative codes	Codes
(V).More efficient information-sharing with doctors			16 (30.7%)
	V-1. Smoother verbal communication between parents and doctors	• *Since the doctor in charge had already looked at the data before the appointment, the conversation during the appointment was smooth.* [Case A, 30 s, mother of a 6-year-old boy] • *Since the doctor in charge had already finished checking the data in advance, the examination was focused on key points, and the conversation went quickly.* [Case C, 30 s, mother of a 6-year-old boy] • *Using the electronic food diary meant that the explanation during the appointment went smoothly.* [Case E, 30 s, mother of a 6-year-old girl] • *The doctor operates the tablet and enlarges the photographs of skin symptoms, which makes it easier to explain them to the doctor.* [Case A, 30 s, mother of a 6-year-old boy] • *It's difficult to explain the condition of roughened skin to the doctor in words, but photographs are easier to explain to the doctor.* [Case A, 30 s, mother of a 6-year-old boy] • *The electronic food diary has made the conversation with the doctor in charge go smoothly, and enabled the consultation to be pinpointed.* [Case C, 30 s, mother of a 6-year-old boy]	6
	V-2. Better information-sharing between parents and doctors	• *Since the doctor in charge has already received the data before the appointment, the conversation can move forward.* [Case B, 30 s, mother of a 3-year-old girl] • *The electronic food diary was useful during telephone conversations and outpatient appointments.* [Case B, 30 s, mother of a 3-year-old girl] • *The stress of telephone appointments has actually decreased.* [Case B, 30 s, mother of a 3-year-old girl] • *I send the electronic food diary data to the doctor in charge before appointments.* [Case C, 30 s, mother of a 6-year-old boy] • *It was useful to convey the electronic food diary data to the doctor in charge.* [Case E, 30 s, mother of a 6-year-old girl]	5

XSL•FO
RenderX

Categories and subcategories		Representative codes	Codes
	V-3. Less time required to explain things to doctors	• *Using the electronic food diary has given us more time during appointments, enabling me to listen carefully to the doctor in charge.* [Case E, 30 s, mother of a 6-year-old girl] • *I've become able to make full use of the time during consultations.* [Case B, 30 s, mother of a 3-year-old girl] • *Examinations used to last around 20min but now finish in 10min.* [Case C, 30 s, mother of a 6-year-old boy] • *The introduction of the electronic food diary has reduced the time it takes to explain what had happened.* [Case D, 30 s, mother of a 6-year-old boy] • *The time taken to explain things to the doctor in charge during appointments has come down to around 10min.* [Case E, 30 s, mother of a 6-year-old girl]	5
(VI).Improved adherence			6 (11.5%)

XSL•FO
RenderX

Categories and subcategories		Representative codes	Codes
	VI-1. Noticing changes in physical condition	• *Continuous use of the electronic food diary has made symptoms easier to predict.* [Case A, 30 s, mother of a 6-year-old boy] • *Above all, I'm now able to notice changes in his physical condition after consumption, following the doctor's instructions.* [Case A, 30 s, mother of a 6-year-old boy] • *By continuing the electronic food diary, I am now able to analyze changes in his physical condition in each season.* [Case C, 30 s, mother of a 6-year-old boy] • *Because it took 3 y to become used to the paper-based food diary, I have accumulated experience until now, and have come to understand how to predict induced symptoms.* [Case D, 30 s, mother of a 6-year-old boy]	4
	VI-2. Sound medication management	• *When inputting oral medication into the electronic food diary, I can recollect the times I forgot to give medication.* [Case C, 30 s, mother of a 6-year-old boy]	1
	VI-3. Judgment of the urgency of induced symptoms and their treatment	• *I remembered that the doctor in charge had told me that I could come to my own judgment on reducing the frequency of inhalation if symptoms improved, and was able to undertake treatment at home without making a mistake.* [Case C, 30 s, mother of a 6-year-old boy]	1

Discussion

Principal Results

Parents of preschool children with FA reported the following: *I. difficulties with using a P-diary*,*II. stress related to sharing of information with doctors*, and *III. difficulties with making disease management habitual* as problems with the P-diary; *IV. ease of using the E-diary* as an effect; *V. more efficient sharing of information with doctors;* and *VI. improved adherence* as outcomes of E-diary use.

Electronic Diary Transition Experiences by Parents of Preschoolers With FA

We explored the experiences of parents of preschool children with FA during the transition from a P-diary to an E-diary.

During the transition, parents reported categories I, II, and III as issues with the P-diary, category IV as an effect of E-diary use, and categories V and VI as outcomes of E-diary use.

Difficulties With the P-Diary

Categories I, II, and III were problems associated with the P-diary. The first problem was *I. difficulties with using a P-diary*. In a study evaluating the use of P-diaries for pain in participants with chronic pain, 90% (36/40) reported compliance with completing the P-diary. However, when the actual time of the diary entry was monitored electronically, true compliance was 11%. Hoarding was common with the paper diary [18]. Many patients recorded in their diaries just before returning them—known as "parking lot entries" [19]. This may be inferred from the large number of codes related to *I-1. time and effort required to transcribe memos into a P-diary*.

In addition, the participants cited *I-2. difficulty recording skin symptoms*, *I-3. difficulty checking food and drink consumption records*, and *I-4. difficulty carrying a P-diary around*. A previous study [20] reported that "skin symptoms were the most common immediate allergic reaction to FA, occurring in 92% of patients, followed by respiratory symptoms in 33.6%, mucous membrane symptoms in 28%, and digestive symptoms in 18.6%.

XSL•FO
RenderX

The severity of symptoms of immediate allergic reactions due to FA should be evaluated in each organ and treated accordingly." Visual information on skin symptoms, such as erythema, urticaria, and wheals, is difficult to express verbally. The difficulty in verbally describing skin symptoms may have contributed to the difficulty in recording them, which includes not only food and drink consumption but also the consumption of snacks, eating out at restaurants, and the occurrence of symptoms in their temporal context. The type and intensity of symptoms are recorded with the date of occurrence and, if necessary, the duration. The use of drugs is also recorded [21]. This complex record content may have contributed to the difficulty in checking FA intake records.

With respect to *II-1. the length of time required to explain details to doctors* and *II-2. the difficulty of explaining information to doctors in simple terms*, before the introduction of the E-diary, no opportunity existed for parents and doctors to share information about the contents of the P-diary before outpatient visits. This may have increased the time required to explain progress in detail to the doctor during the visit, leading to stress related to the sharing of information. The participant also mentioned *II-3. difficulty listening to the doctor's explanation while attending to the child*. Health care providers need to communicate effectively to explain all treatment options, including continued causal food avoidance and OIT [22]. However, another study [23] showed that parents often receive far more information during a pediatric visit than they can realistically remember. The study also noted that environmental distractions of having young children nearby during office visits may further impair memory. These findings are consistent with the present results.

The third problem was *III. difficulties with making disease management a habit*. Participants cited *III-1. forgetting about OIT or medication management*. OIT starts below the threshold value (the minimum amount that would elicit a reaction), with the oral dose gradually increasing over time to improve clinical tolerance of the target food. Since this treatment entails a few risks, parents of children diagnosed with FA must be equipped to manage induced symptoms [22]. Inadvertently forgetting OIT is a serious problem for maintaining FA treatment. A previous study [18] reported that with P-diaries, 75% (n=5) of patients accumulate a day's worth of records. This accumulation of records is called "parking lot entries" [19]. With P-diaries, it has been suggested that OIT and medication administration may be forgotten, leading to difficulty in disease management. In addition, participants indicated *III-2. difficulty noticing changes in physical condition*. Diaries generally capture changes in symptoms over time [24].

Diaries documenting FA-related diet and symptoms are useful for providing patients and their families with a more concrete understanding of their lifestyle and symptoms. However, with P-diaries, it was suggested that difficulties in using the diary may have hindered the detection of changes in physical condition.

Effects of the E-Diary

Category IV addressed the effects of the E-diary. Participants cited *IV-1. ease of checking previous records* as one subcategory. An E-diary is a powerful and efficient tool for collecting complex self-reported data. It provides unique insights into the symptom experience by documenting significant changes [25]. Patients reported that the greatest advantage of E-diaries as a tool was the ease of reviewing the course of symptoms. The participants also mentioned *IV-2. ease of viewing photographs and text entries*. Previous studies have reported the value of smartphone apps for monitoring visual information on cutaneous lesions in atopic dermatitis using photos [26] and for the daily evaluation and recording of atopic dermatitis symptom scores [27]. Another study suggested that smartphone apps may be effective in improving patients' awareness and management of anaphylaxis [28]. These findings may provide new insights into the effectiveness of FA-related E-diary use. In addition, participants cited *IV-3. ease of inputting records* and *IV-4. ease of continuing records*. Since smartphones are easy to use anytime and anywhere, they enable the effective collection of real-time symptoms and their sharing with medical professionals, which has been reported as useful for monitoring cutaneous symptoms of atopic dermatitis [27]. The present results are consistent with this view, suggesting that the introduction of the E-diary may help resolve problems associated with the P-diary.

Outcomes of the E-Diary

Categories V and VI represented the outcomes of the E-diary. The first outcome was *V. more efficient information-sharing with doctors*. Participants mentioned *V-1. smoother verbal communication between parents and doctors*. Diaries that include visual cutaneous symptoms have been reported to be useful for doctors for diagnosis [26], as well as for helping parents to convey their child's symptoms accurately. Participants also cited *V-2. better information-sharing between parents and doctors*. According to a previous study [29], parents want improved information-sharing and communication with health care providers. They reported that communication technology has the potential to improve communication and coordination of care. The present results are consistent with this view. In addition, the participants cited *V-3. less time required to explain information to doctors*. In a survey of parents' perceptions of the use of mHealth apps, 71% (n=5) reported that it saved time for doctors and nurses [26]. The present results are consistent with this view.

The second outcome was *VI. improved adherence*. The participants cited *VI-1. noticing changes in physical condition*. The goal of FA care is to help patients and their caregivers manage the risk of food allergic reactions, reduce food-related anxiety, and gain a sense of control over their condition [22]. To enable them to acquire this sense, if the patient is a young child, the doctor must instruct the parents to record the amount of the offending food consumed and the allergic symptoms induced. Diaries are generally considered effective in capturing symptom variability over time [24]. The results indicate that the E-diary may be useful as an efficient recording tool. The participants also indicated *VI-2. sound medication management*. Numerous mobile phone medication adherence apps are presently available and may serve as useful tools to enable patients to take medication as prescribed [30]. The present results are consistent with this view.

XSL•FO
RenderX

In addition, the participants listed *VI-3. judgment of the urgency of induced symptoms and their treatment*. A previous study [23] showed that during pediatric visits, parents often receive far more information than they can realistically remember. Such information, which often covers multiple broad areas (eg, disease complications and treatments), is associated with poorer memory compared to information on narrower topics. These results represent a novel outcome of the use of an FA-related E-diary.

Limitations

This study had some limitations. After using the E-diary for 6 months, 5 parents were asked about their experience transitioning from the P-diary. Since only 5 parents were interviewed at a single medical institution, some of the aspects evaluated may have been underestimated or overestimated. In addition, the allergens of children with FA were milk, wheat, and eggs—the 3 major allergens in Japan—representing a potential selection bias. However, factors such as the number of years from diagnosis to interview, severity of FA, and number of allergens—which may affect how parents use the E-diary—were not considered in the selection of participants.

Strengths and Future Challenges

This study was the first in Japan to confirm the use of an E-diary specifically designed for parents of children diagnosed with FA. The results improve our understanding of the use of E-diaries by parents of preschool children with FA and have important practical implications for patient education.

Future studies should include a larger sample size and a longer follow-up period (eg, 1 year). Notably, all participants in this study were mothers. In Japan, despite the increasing participation of fathers in childcare, mothers typically record information in the E-diary and accompany their children to outpatient appointments. However, future surveys should include both fathers and mothers.

In Japan, few reports exist on the use of E-diaries. In this study, to facilitate sharing the contents of E-diaries between parents and physicians, the researchers received data in PDF format from the parents and then forwarded it to the physicians. The presence of the researcher as an intermediary was a positive factor in the use of the E-diary. The role of the researcher has not been fully evaluated. Further research is needed on the role of intermediaries who support medical treatment using E-diaries.

Role of Supporters Linking Parents and Doctor

During the first 6 months after the introduction of the E-diary, the researcher was responsible for converting the data recorded in the E-diary into PDF files and sending them to the physician before the outpatient visit. van der Kamp et al [31] reported that communication, as demonstrated by self-monitoring data, builds trust between parents, children, and health care providers and enhances understanding and self-management of illness. The environment in which parents and physicians could share the contents of the E-diary prior to the visit may have facilitated effective communication. Sezgin et al [29] pointed out that a systematic approach using communication technology is needed to enable parents to share information with their health care providers. Furthermore, it has been reported [32] that when promoting eHealth interventions, it is important to have specific opinion leaders in the hospital who can encourage the adoption of digital care. These pilot study findings are important for the promotion of E-diary–based medical care. To facilitate a systematic approach in the future, it is necessary to secure supporters of E-diary implementation.

Conclusions

The authors found that parents of preschool children with FA were aware of the following problems with P-diaries: *I. difficulties with using a P-diary*; *II. stress related to information-sharing with doctors*; and *III. difficulties with making disease management habitual*. They reported *IV. ease of using the E-diary* as an effect of its use. They further reported *V. more efficient information-sharing with doctors* and *VI. improved adherence* as outcomes of using the E-diary. E-diaries will contribute to solving the problems of P-diaries and may be a useful tool for communication between parents and physicians, as well as for disease management.

Acknowledgments

We thank the parents of preschool children with food allergies who participated in this study. In this study, we used MyEpi mHealth, developed by Mylan EPD. The E-diary feature of MyEpi mHealth, which enables the recording of allergy symptoms and the storage of photos, was used to collect allergy symptom data efficiently during the 6-month period from October 2020, when in-person medical consultations were restricted due to the COVID-19 (SARS-CoV-2) pandemic. This significantly accelerated the progress of the study. Without the contributions of MyEpi mHealth, this study would not have been successful. We deeply appreciate their support.

This study was supported by the Akiyama Life Science Foundation. The author declares that financial support was received for this research. The researchers did not use any artificial intelligence or qualitative research software in writing this study.

Data Availability

All data generated or analyzed during this study are included in this published article.

XSL•FO
RenderX

Authors' Contributions

KY was responsible for acquiring financial support for the project that led to this publication. KY conducted the research and investigation processes, including performing experiments and collecting data. KY also managed and coordinated the research activities, including planning and execution. Furthermore, KY prepared, created, and presented the published work, including writing the initial draft and providing substantive translation.

Conflicts of Interest

None declared.

Checklist 1
COREQ checklist.
[PDF File, 432 KB - formative_v9i1e73400_app1.pdf]

References

1. Loh W, Tang MLK. The epidemiology of food allergy in the global context. Int J Environ Res Public Health 2018 Sep 18;15(9):2043. [doi: 10.3390/ijerph15092043] [Medline: 30231558]
2. Yamamoto-Hanada K, Pak K, Iwamoto S, et al. Parental stress and food allergy phenotypes in young children: a National Birth Cohort (JECS). Allergy 2024 Jul;79(7):1881-1892. [doi: 10.1111/all.16035] [Medline: 38269609]
3. NIAID-Sponsored Expert Panel, Boyce JA, Assa'ad A, et al. Guidelines for the diagnosis and management of food allergy in the United States: report of the NIAID-sponsored expert panel. J Allergy Clin Immunol 2010 Dec;126(6 Suppl):S1-58. [doi: 10.1016/j.jaci.2010.10.007] [Medline: 21134576]
4. Westwell-Roper C, To S, Andjelic G, et al. Food-allergy-specific anxiety and distress in parents of children with food allergy: a systematic review. Pediatr Allergy Immunol 2022 Jan;33(1):e13695. [doi: 10.1111/pai.13695] [Medline: 34779046]
5. Pajno GB, Fernandez-Rivas M, Arasi S, et al. EAACI Guidelines on allergen immunotherapy: IgE-mediated food allergy. Allergy 2018 Apr;73(4):799-815. [doi: 10.1111/all.13319] [Medline: 29205393]
6. Epstein-Rigbi N, Goldberg MR, Levy MB, Nachshon L, Elizur A. Quality of life of food-allergic patients before, during, and after oral immunotherapy. J Allergy Clin Immunol Pract 2019 Feb;7(2):429-436. [doi: 10.1016/j.jaip.2018.06.016] [Medline: 30129441]
7. Shin S, Jang S, Kim J, et al. Initial updosing phase of oral immunotherapy improves quality of life and psychological burden in parents of children with food allergy. Allergy Asthma Proc 2024 Mar 1;45(2):128-136. [doi: 10.2500/aap.2024.45.240001] [Medline: 38449018]
8. Polloni L, Muraro A, Bonaguro R, et al. Psychological needs and support among patients and families undergoing food oral immunotherapy. Clin Transl Allergy 2022 Feb;12(2):e12078. [doi: 10.1002/clt2.12078] [Medline: 35140908]
9. Rigbi NE, Goldberg MR, Levy MB, Nachshon L, Golobov K, Elizur A. Changes in patient quality of life during oral immunotherapy for food allergy. Allergy 2017 Dec;72(12):1883-1890. [doi: 10.1111/all.13211] [Medline: 28542911]
10. Sánchez-Fernández S, Lasa EM, Terrados S, et al. Mobile app/web platform for monitoring food oral immunotherapy in children: longitudinal clinical validation study. JMIR Pediatr Parent 2024 Mar 13;7:e54163. [doi: 10.2196/54163] [Medline: 38477961]
11. van der Kamp MR, Hengeveld VS, Brusse-Keizer MGJ, Thio BJ, Tabak M. eHealth technologies for monitoring pediatric asthma at home: scoping review. J Med Internet Res 2023 Jul 21;25:e45896. [doi: 10.2196/45896] [Medline: 37477966]
12. Nassehi D, Gripsrud BH, Ramvi E. Theoretical perspectives underpinning research on the physician-patient relationship in a digital health practice: scoping review. Interact J Med Res 2024 May 15;13:e47280. [doi: 10.2196/47280] [Medline: 38748465]
13. Brands MR, Gouw SC, Beestrum M, Cronin RM, Fijnvandraat K, Badawy SM. Patient-centered digital health records and their effects on health outcomes: systematic review. J Med Internet Res 2022 Dec 22;24(12):e43086. [doi: 10.2196/43086] [Medline: 36548034]
14. Matricardi PM, Sousa-Pinto B, Dramburg S, Bousquet J. Beyond ARIA: will e-diaries replace retrospective questionnaires in measuring the severity of allergic rhinitis in clinical research and daily practice? Clin Exp Allergy 2023 Oct;53(10):1004-1010. [doi: 10.1111/cea.14372] [Medline: 37488953]
15. Berelson B. Content Analysis in Communication Research: The United States of America: Free Press; 1952.
16. Vaismoradi M, Turunen H, Bondas T. Content analysis and thematic analysis: implications for conducting a qualitative descriptive study. Nurs Health Sci 2013 Sep;15(3):398-405. [doi: 10.1111/nhs.12048] [Medline: 23480423]
17. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
18. Stone AA, Shiffman S, Schwartz JE, Broderick JE, Hufford MR. Patient compliance with paper and electronic diaries. Control Clin Trials 2003 Apr;24(2):182-199. [doi: 10.1016/s0197-2456(02)00320-3] [Medline: 12689739]

XSL•FO
RenderX

19. Baggott C, Gibson F, Coll B, Kletter R, Zeltzer P, Miaskowski C. Initial evaluation of an electronic symptom diary for adolescents with cancer. JMIR Res Protoc 2012 Dec 11;1(2):e23. [doi: 10.2196/resprot.2175] [Medline: 23612521]
20. Ebisawa M, Ito K, Fujisawa T, Committee for Japanese Pediatric Guideline for Food Allergy, The Japanese Society of Pediatric Allergy and Clinical Immunology; Japanese Society of Allergology, Japanese Society of Allergology. Japanese guidelines for food allergy 2020. Allergol Int 2020 Jul;69(3):370-386. [doi: 10.1016/j.alit.2020.03.004] [Medline: 33289637]
21. Worm M, Reese I, Ballmer-Weber B, et al. Update of the S2k guideline on the management of IgE-mediated food allergies. Allergol Select 2021;5(195-243):195-243. [doi: 10.5414/ALX02257E] [Medline: 34263109]
22. Bégin P, Chan ES, Kim H, et al. CSACI guidelines for the ethical, evidence-based and patient-oriented clinical practice of oral immunotherapy in IgE-mediated food allergy. Allergy Asthma Clin Immunol 2020;16(1):20. [doi: 10.1186/s13223-020-0413-7] [Medline: 32206067]
23. Luke AK, Flessner CA. Examining differences in parent knowledge about pediatric food allergies. J Pediatr Psychol 2020 Jan 1;45(1):101-109. [doi: 10.1093/jpepsy/jsz091] [Medline: 31845999]
24. Weerts ZZRM, Heinen KGE, Masclee AAM, et al. Smart data collection for the assessment of treatment effects in irritable bowel syndrome: observational study. JMIR Mhealth Uhealth 2020 Nov 2;8(11):e19696. [doi: 10.2196/19696] [Medline: 33030150]
25. McEntegart D. Are electronic diaries useful for symptoms research? A systematic review. J Psychosom Res 2007 Oct;63(4):453. [doi: 10.1016/j.jpsychores.2007.08.003] [Medline: 17905056]
26. Xu X, Griva K, Koh M, et al. Creating a smartphone app for caregivers of children with atopic dermatitis with caregivers, health care professionals, and digital health experts: participatory co-design. JMIR Mhealth Uhealth 2020 Oct 29;8(10):e16898. [doi: 10.2196/16898] [Medline: 33118949]
27. Lee JY, Kim M, Yang HK, et al. Reliability and validity of the Atopic Dermatitis Symptom Score (ADSS). Pediatr Allergy Immunol 2018 May;29(3):290-295. [doi: 10.1111/pai.12865] [Medline: 29350788]
28. Muraro A, Worm M, Alviani C, et al. EAACI guidelines: anaphylaxis (2021 update). Allergy 2022 Feb;77(2):357-377. [doi: 10.1111/all.15032] [Medline: 34343358]
29. Sezgin E, Noritz G, Lin S, Huang Y. Feasibility of a voice-enabled medical diary app (SpeakHealth) for caregivers of children with special health care needs and health care providers: mixed methods study. JMIR Form Res 2021 May 11;5(5):e25503. [doi: 10.2196/25503] [Medline: 33865233]
30. Park JYE, Li J, Howren A, Tsao NW, De Vera M. Mobile phone apps targeting medication adherence: quality assessment and content analysis of user reviews. JMIR Mhealth Uhealth 2019 Jan 31;7(1):e11919. [doi: 10.2196/11919] [Medline: 30702435]
31. van der Kamp M, Reimering Hartgerink P, Driessen J, Thio B, Hermens H, Tabak M. Feasibility, efficacy, and efficiency of eHealth-supported pediatric asthma care: six-month quasi-experimental single-arm pretest-posttest study. JMIR Form Res 2021 Jul 26;5(7):e24634. [doi: 10.2196/24634] [Medline: 34309568]
32. Ariens LF, Schussler-Raymakers FM, Frima C, et al. Barriers and facilitators to eHealth use in daily practice: perspectives of patients and professionals in dermatology. J Med Internet Res 2017 Sep 5;19(9):e300. [doi: 10.2196/jmir.7512] [Medline: 28874336]

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Studies
E-diary: electronic diary
FA: food allergy
mHealth: mobile health
OIT: oral immunotherapy
P-diary: paper diary

Edited by A Mavragani; submitted 11.03.25; peer-reviewed by M Galvan, MJ Goikoetxea Lapresa; revised version received 14.08.25; accepted 17.08.25; published 21.10.25.

Please cite as:
Kato Y
Transition From a Paper Diary to an Electronic Diary by Parents of Preschool Children With Food Allergies: Pilot Study
JMIR Form Res 2025;9:e73400
URL: https://formative.jmir.org/2025/1/e73400
doi:10.2196/73400

XSL•FO
RenderX

© Yoriko Kato. Originally published in JMIR Formative Research (https://formative.jmir.org), 21.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Integration of an Artificial Intelligence–Based Autism Diagnostic Device into the ECHO Autism Primary Care Workflow: Prospective Observational Study

Kristin Sohl[1], MD, FAAP; Erik Linstead[2], PhD; Kelianne Heinz[3], MIDS; Elia Eiroa Lledo[2], PhD; Alicia Brewer Curran[1], BS; Melissa Mahurin[1], BA; Valeria Nanclares-Nogués[1], PsyD; Carmela Salomon[3], PhD; Minda Seal[3], MSA; Sharief Taraman[3,4,5], MD

[1]ECHO Autism Communities, University of Missouri School of Medicine, Columbia, MO, United States

[2]Fowler School of Engineering, Chapman University, Orange, CA, United States

[3]Cognoa Inc, 5000 Campus Drive, Newport Beach, CA, United States

[4]Pediatric Neurology and Clinical Informatics, Children's Health of Orange County, Orange, CA, United States

[5]School of Medicine, University of California Irvine, Irvine, CA, United States

Corresponding Author:
Carmela Salomon, PhD
Cognoa Inc, 5000 Campus Drive, Newport Beach, CA, United States

Abstract

Background: Pediatric specialist shortages and rapidly rising autism prevalence rates have compelled primary care clinicians to consider playing a greater role in the autism diagnostic process. The ECHO Autism: Early Diagnosis Program (EDx) prepares clinicians to screen, evaluate, differentiate, diagnose, and provide longitudinal care for children with autism in primary care settings. Canvas Dx is a prescription-only Software as a Medical Device designed to support clinical diagnosis or rule out of autism, including in primary care settings. It is authorized by the Food and Drug Administration for use, in conjunction with clinical judgment, in 18‐to 72-month-olds with indicators of developmental delay.

Objective: This study aims to assess the feasibility and impact of integrating the device into the ECHO Autism: EDx workflow. Time from the first clinical question of developmental delay to autism diagnosis is the primary endpoint. Secondary endpoints explore clinician and caregiver experience of device use.

Methods: Children aged 18 to 72 months with concern for developmental delay indicated by either a caregiver or health professionals were eligible to participate in this prospective observational study. Experienced ECHO Autism: EDx clinicians were recruited to evaluate the inclusion of the device as part of their diagnostic evaluations. Outcome data were collected via a combination of electronic questionnaires, standard clinical care record reviews, and analysis of device outputs. Institutional review board approval was provided by the University of Missouri-Columbia (project number 2075722).

Results: Eighty children and 7 clinicians completed the study. On average, time from clinical concern at study enrollment to final autism diagnosis was 39.22 days, compared to 180‐to 264-day waits at adjacent specialist referral centers. The vast majority (93%, 50/54) of caregivers reported being satisfied with the ECHO Autism: EDx plus device evaluation their child received and endorsed that they would recommend it to others and that they felt comfortable using the device. The device produced determinate autism predictions or rule-outs for 52.5% of participants, and in all cases, these were consistent with the final clinical determination. Participating clinicians reported that device use was feasible and reduced several challenges associated with their previous diagnostic process; however, they noted it did not obviate the need for additional structured observation in every case.

Conclusions: The ECHO Autism: EDx plus device workflow offers considerable time savings compared to specialty center referral and was strongly endorsed by caregiver participants. Embedding the device into the ECHO Autism: EDx workflow was feasible and helped streamline several workflow efficiencies. Clinicians still utilized their training and application and interpretation of Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, criteria when formulating the diagnosis for indeterminate cases.

Trial Registration: ClinicalTrials.gov NCT05223374; https://tinyurl.com/5c3neaa9

International Registered Report Identifier (IRRID): RR2-10.2196/37576

(JMIR Form Res 2025;9:e80733) doi:10.2196/80733

XSL•FO
RenderX

KEYWORDS

autism spectrum disorder; diagnosis, computer-assisted; primary health care; artificial intelligence; machine learning; developmental disabilities; early diagnosis; real-world evidence; pediatric primary care

Introduction

Background

Rates of children diagnosed with autism continue to rise dramatically, with one in 31 eight-year-olds now receiving a diagnosis in the United States [1]. While earlier identification and intervention during the critical neurodevelopmental window can maximize outcomes [2-4], many children experience critical delays to diagnosis [5]. There is a multiyear lag, for example, between the 18‐ to 24-month period where diagnosis is possible [6] and the average US age of diagnosis of nearly 5 years [7]. Delayed diagnoses represent missed opportunities for impactful early intervention, with new evidence emerging that even small delays in treatment initiation (18 vs 27 mo) can reduce its efficacy [8].

In the United States, autism evaluations typically occur in specialist centers, with less than 1% being made in primary care [9,10]. The inadequate number of specialists available to assess a rapidly growing pool of children with questions about developmental delay has created multiyear waitlists for evaluations [11,12]. To help address this crisis, calls to recruit more primary care physicians and practitioners to participate in the diagnostic process are growing [13-15].

While multiple studies show benefits to this approach, unanswered questions remain about the best way to integrate autism evaluations into the primary care workflow [16-20]. Many traditional diagnostic tools, for example, are time-intensive or difficult to administer [21] and thus represent barriers to autism evaluation in primary care. Two of the most commonly used tests in autism diagnosis are the Autism Diagnostic Interview-Revised and the Autism Diagnostic Observation Schedule [21]. The Autism Diagnostic Interview-Revised is a structured parent interview that typically takes 2‐2.5 hours to administer and score. The Autism Diagnostic Observation Schedule is an observational assessment that must be conducted in-person and can take 60‐90 minutes or longer to complete. Both instruments require formal in-person training followed by ongoing practice to maintain scoring reliability. Best results are often achieved by administering both tools to the same child [21], resulting in a time commitment that is unfeasible in the context of a primary care workflow.

Canvas Dx

Canvas Dx, a Food and Drug Administration–authorized, artificial intelligence (AI)-based Software as a Medical Device, was developed to support streamlined autism evaluations in 18‐ to 72-month-olds with developmental delay concern [22]. The underlying machine learning algorithm was trained on data from thousands of children with diverse demographic backgrounds and behavioral health presentations with the goal of supporting accurate, equitable, and rapid identification or rule out of autism in primary care [23]. Canvas Dx does not require specialist training to administer and has been clinically validated for both remote and in-person administration [24].

The device's gradient-boosted decision trees algorithm [23,25] integrates data from a brief caregiver questionnaire, a video analyst questionnaire, and a clinical questionnaire that can be administered either remotely or in person, while the clinician observes and interacts with the child. Where sufficient information is provided to make a highly accurate autism prediction or rule out, the device produces a *Positive for autism* or *Negative for autism* output. When a highly predictive determination cannot be made, the device abstains from providing a result (*Indeterminate* output).

The ECHO Autism Primary Care Early Diagnosis (EDx) Model [26,27] is a multifaceted professional development model including in-person and virtual components to support primary care clinicians regarding best practice autism evaluation and longitudinal care. The EDx model involves training and mentoring community-based clinicians to conduct diagnostic assessments for autism, consistent with best practice standards. ECHO Autism: EDx is the first of its kind model designed to professionally develop community-based primary care clinicians to assess and appropriately diagnose young children with a question of autism. However, it requires significant resources of time and professional learning to achieve competency in serving this population well. Therefore, ECHO Autism Communities and Cognoa engaged in this study to evaluate the feasibility of using Canvas Dx in lieu of the model's usual observational tool, the Screening Tool for Autism in Toddlers and Young Children (STAT) [28], to support primary care clinician diagnostic determination.

The primary study objective was to assess the time from initial clinician concern to diagnosis when using the device as part of the ECHO Autism: EDx diagnostic pathway. The secondary objective was to explore clinician and caregiver experience of the feasibility and impact of device use as part of the diagnostic journey.

Methods

Study Protocol Overview

The study protocol for this prospective observational feasibility study is described in full in JMIR research protocols [29], with key methodological considerations summarized further. Eligible participants were enrolled consecutively to minimize selection bias, with data collection occurring between April 2022 and November 2023.

Ethical Considerations

The study received institutional review board approval from the University of Missouri-Columbia (project number 2075722), and the protocol was registered with ClinicalTrials.gov prior to study initiation (protocol identifier: NCT05223374). Written informed consent was obtained from all participants, and the

XSL•FO
RenderX

participants had an ability to opt out of the study at any time. Device costs were covered by the study sponsor Cognoa, but no other monetary compensation was provided to the participants. Data have been anonymized to protect participant confidentiality, and no identifiable images of participants are included in the manuscript or supplementary material.

Blinding

Cognoa leadership and ECHO Autism leadership were blinded to the study dataset and had no access to the study end points until after the first data analysis was completed by an independent third party. This was to ensure minimal introduction of influence by either organization involved in the study.

Inclusion Criteria

ECHO Autism Clinicians

ECHO autism clinicians (EACs) eligible for study participation were practicing primary care physicians who had previously received training in the ECHO Autism: EDx model and had met diagnostic agreement definitions to independently evaluate and diagnose autism [30].

Patients

As part of routine practice, participating clinicians received referrals to evaluate pediatric patients. Staff at participating clinics screened referred children during the study period to assess if they potentially met the criteria for study inclusion. If the criteria appeared to be met, clinic personnel informed families about the study. Interested families were subsequently referred to the University of Missouri ECHO Autism Communities Research Team, who were responsible for reviewing the study in detail and obtaining informed consent from the caregiver prior to participation.

Children were eligible for enrollment if, at the time of screening, they were between 18 and 72 months of age and a caregiver, health care provider, or community professional had expressed concern about possible developmental delay. Eligible participants were required to come from families proficient in English or Spanish, as determined by the assessing EAC, and a parent, guardian, or legally authorized representative needed to be capable of reading and signing informed consent. Families also had to have access to a smartphone operating on a recent version of iOS or Android, be willing to provide short home video recordings of the child for device input, and meet all device labeling requirements [31].

Children were excluded post-screening if they had a previous autism diagnosis provided by a health care professional or if they had a medical, developmental, or behavioral condition that, in the opinion of the EAC, could interfere with study assessments. Any potential participant who had previously been involved in any study or survey related to the Canvas Dx device was also excluded. Additionally, patients were excluded if they had conditions listed as exclusions for reliable use by the device manufacturer. These conditions included suspected auditory or visual hallucinations or a prior diagnosis of childhood-onset schizophrenia; known deafness or blindness; physical impairments affecting hand use; major dysmorphic features or prenatal exposure to teratogens (eg, fetal alcohol syndrome); history or diagnosis of genetic conditions such as Rett syndrome or Fragile X; microcephaly; history of epilepsy or seizures; history of neglect; and brain defects, injuries, or insults requiring interventions such as surgery or chronic medication.

Data Collection Instruments

The device inputs consist of the following:

1. A *caregiver questionnaire* consisting of 18- or 21-item (age-dependent) multiple questions that capture the caregiver's observations of the child's development and behaviors. Caregivers independently completed the questionnaire within a mobile app and were blinded to other device inputs. The caregiver questionnaire is designed to be used in combination with other device inputs rather than as a stand-alone assessment.
2. A video analyst questionnaire answered based on analysis of two videos the caregiver uploads via the mobile app. The 28- or 33-item questionnaire (age-dependent) is completed by trained video analysts within the analyst portal. Video analysts are required to have at least a master's degree, 5 or more years of experience working with children with autism, and must undergo training and yearly recertifications. Each video is independently scored, with analysts documenting behaviors across relevant domains, including verbal and nonverbal communication, social interaction, sensory responses, repetitive movements, use of objects, and speech. Video analysts are blinded to the caregiver questionnaire responses and clinician questionnaire responses and must answer based solely on their observations of the videos.
3. A *clinical questionnaire* consisting of 13 to 15 (age-dependent) multiple choice questions is answered by the EAC during an observational session with the caregiver and child. The clinician is blinded to the results of the video analyst questionnaire and must answer based on their observations of the child's behavioral and developmental characteristics and knowledge of the child's history. The EAC inputs questionnaire responses via a web portal. The clinical questionnaire is one of the required device inputs and is not a stand-alone measurement tool.

The device output is accessed via the clinical web portal. It consists of a diagnostic determination (Positive for autism, Negative for autism, or Indeterminate); all questions and responses from Inputs 1 and 3; and the caregiver's submitted videos.

Measure of Processes of Care: This is a validated tool adapted for this study to measure caregivers' perceptions of the family-centeredness of the care provided.

ECHO Autism Caregiver Presurvey: This collects baseline demographic information and captures caregivers' initial impressions of the EAC prior to the diagnostic process.

ECHO Autism Caregiver Postdiagnostic Survey: This assesses caregiver satisfaction with the diagnostic services received.

EAC Pre-Project Survey: This evaluates clinicians' self-efficacy, perceived barriers, and satisfaction with the current ECHO Autism diagnostic process. The current ECHO autism diagnostic

XSL•FO
RenderX

workflow consists of the following: (a) completion of a 2-day in-person training on the STAT. This is a direct structured behavioral observation that takes approximately 20 minutes to administer in the physician's office once training is complete; (b) achievement of administration and coding reliability on the STAT instrument to a certified STAT trainer; (c) participation in twice monthly, 90-minute iterative case-based learning through ECHO Autism: Primary care sessions where additional skills required as part of best practice autism evaluations such as diagnostic interviewing, differential diagnosis, and conceptualization of development and behavior in young children are acquired.

EAC Post-Project Survey: This assesses changes in clinician self-efficacy, perceived barriers, and satisfaction following integration of the device into the diagnostic workflow.

ECHO Autism 3-Month Follow-Up Questionnaires: This was designed to capture data needed to address secondary and exploratory study objectives.

ECHO Autism Interim Visit Form: This was used to document any participant visits occurring between the initial concern and the 3-month follow-up.

Statistical Analysis

Descriptive statistics were used to summarize all collected survey data from caregiver and clinician participants. Continuous variables are reported as n, mean, standard deviation, median, minimum, and maximum values, while categorical variables are presented as counts and percentages where relevant. All available data, including observed and missing values, were tabulated.

Time from initial clinical concern at enrollment to diagnosis by the EAC was calculated in days and compared with the projected wait time if participants had been referred to adjacent specialist centers at first concern instead of following the primary care diagnosis pathway. Both EACs and adjacent specialist centers assess children with concern for developmental delay across the complexity spectrum and labeling age range. To generate comparison estimates for specialty center wait times, we obtained publicly reported median wait times to first appointment from the Missouri Autism Specialty Centers [32]. The Missouri Autism Centers aggregate and track median wait time data from multiple autism specialty centers across Missouri. Data are reported annually by fiscal year and stratified into two age cohorts: children under 60 months and children 60 months and older. Because the current study period spanned two fiscal years (2022 - 2023), we calculated the mean of the 2 years' reported median wait times to derive an overall estimate for each age cohort. For children under 60 months, the reported median wait times were 143 days (FY2022) and 217 days (FY2023), yielding an average of 180 days. For children 60 months and older, the reported medians were 221 days (FY2022) and 306 days (FY2023), yielding an average of 264 days. These averaged values were used as the comparison benchmarks in subsequent analyses. Distance from the families' home address to the primary care EAC appointment was collected as part of the caregiver surveys and compared with the distance families would have otherwise been required to travel to their nearest specialist center.

Device diagnostic determinations (*Positive for autism*, *Negative for autism*, or *Indeterminate*) were compared with the reference standard diagnosis from the EAC to calculate positive and negative predictive values, as well as sensitivity and specificity in the determinate group. Device performance was also analyzed across race or ethnicity, biological sex, and age to determine if performance metrics varied by demographic. Device-generated timestamps were used to assess the time burden associated with input completion.

Study Flow

Figure 1 summarizes the full study flow.

XSL•FO
RenderX

Figure 1. Study flow overview for prospective observational study of Canvas Dx integration into primary care workflow to support diagnosis or rule out of autism: Caregivers of children with developmental concerns (18-72 mo) are referred and consented by study team. If deemed eligible during Visit 1 (observation and DSM-5 interview), the clinician prescribes the device. The device captures three inputs: (a) caregiver questionnaire+home videos (Input 1), (b) remote video-analyst scoring (Input 2), and (c) clinician questionnaire (Input 3)—to generate a positive/negative/indeterminate output. During Visit 2, the clinician makes a diagnostic determination based on device results and clinical impression. If positive for autism, Visit 3 is scheduled to review supports or services. DSM-5: Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition.

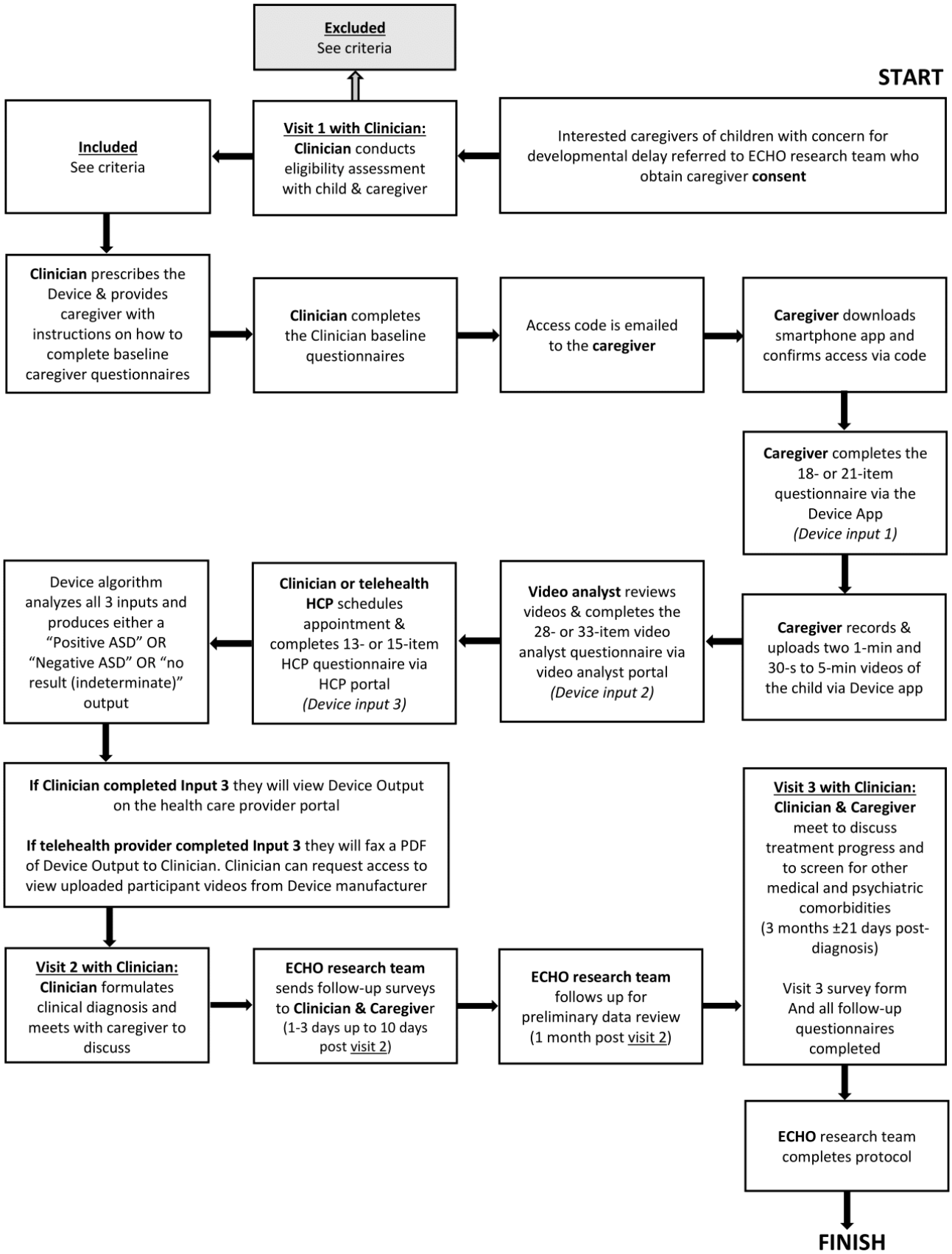

XSL•FO
RenderX

Prior to Visit 1

A clinical appointment is scheduled for children with a question of autism. Clinic staff share study information with eligible families, and those who expressed interest are referred to the research team, where informed consent is obtained.

Visit 1: Unstructured Observation and Comprehensive History

Participating ECHO Autism: EDx clinician observes the participant and conducts a comprehensive medical and developmental history along with Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5)–based autism-focused behavioral interview. Visit one lasts 45 minutes, on average.

Device Completion

The child's caregiver downloads the device mobile application to their smartphone and completes the Caregiver Questionnaire and video recording tasks. The uploaded videos are remotely reviewed and scored by trained video analysts via the video analyst portal. The clinician completes the clinical questionnaire on a web portal. The device algorithm analyzes all three inputs and produces either a Positive for autism, Negative for autism, or Indeterminate output.

Visit 2: Diagnostic Results

The clinician makes an autism diagnosis or rules it out after synthesizing the clinical information and the device output. For children with a confirmed positive result, the EAC will make the diagnosis in primary care and recommend treatments as appropriate during this visit. For children with a negative result, the EAC will rule out autism. If warranted, they may assess for other possible causes of developmental delay and recommend treatments accordingly. For children with an indeterminate result, the EAC will rely on their observations, comprehensive DSM-5 history, and clinical expertise to arrive at an autism status determination. The clinician reviews the diagnostic determination with the caregiver. If the child receives an autism diagnosis, the clinician will schedule a 3-month follow-up visit. The research team sends electronic post-diagnostic questionnaires to the family and the clinician.

Visit 3: Supports and Services Progress Meeting

The patient and their caregiver will follow up with the clinician 3 months (±21 d) after they receive an autism diagnosis to discuss progress associated with obtaining and receiving supports and services. Additionally, the clinician will continue to screen and monitor for common medical and psychiatric co-occurring conditions. The research team will send electronic post-diagnostic questionnaires to the caregiver and to the clinician to be completed 1‑3 (up to 10) days after this follow-up visit.

Protocol Deviations

The protocol anticipated a data-collection period of approximately 12 months and a target of about 100 participants completing the full protocol. These were planning projections and not required minimums or formal stopping rules. In practice, 109 participants were enrolled and 80 completed all study procedures. Because the study had already extended beyond the anticipated 12-month window, we concluded data collection at 80 completers due to operational constraints. This deviation pertains only to study duration and completer count. No changes were made to eligibility criteria, data collection procedures, or the prespecified analysis plan.

Results

Participants

Figure 2 summarizes the recruitment and retention funnel for the study.

XSL•FO
RenderX

Figure 2. Participant recruitment and retention funnel: A total of 109 patients aged 18-72 mo with concern for developmental delay initially enrolled. One hundred of these were deemed eligible to participate in the full study after visit 1. Eighty of these eligible patients completed all device inputs and held a diagnostic results meeting with their clinician (80% of total eligible enrolled participants).

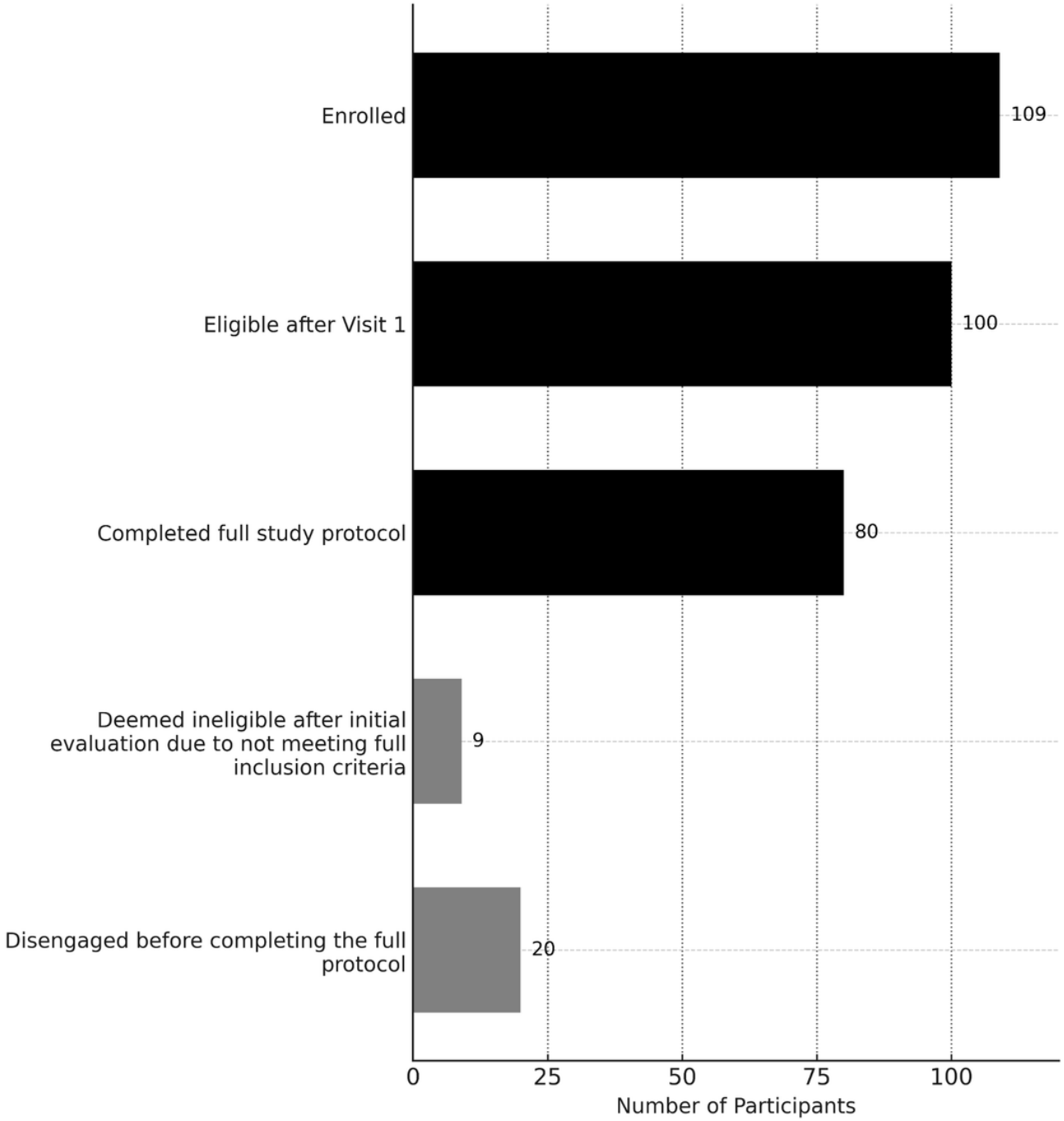

Patient Demographics

The overall patient population had an autism prevalence of 60% (48/80) per clinician evaluation and was 30% (24/80) female. The age of participants ranged from 18.29 to 71.28 months at first appointment, with a mean age of 42.25 (SD 14.41) months and median age of 39.91 months. There were no gender or age-related differences between study completers and noncompleters. Just over half (61.76%, 63/102) of participants were primarily insured by Medicaid, while 38.24% (39/102) of participants were privately insured (Table 1).

Table . Patients' primary insurance, shown for the full population and also stratified by study completion status.

Population	Medicaid primary insurance, n (%)	Private primary insurance, n (%)
All patients (N=102)	63 (61.76)	39 (38.24)
Completers (N=80)	49 (61.25)	31 (38.75)
Noncompleters (N=22)	14 (63.64)	8 (36.36)

XSL•FO
RenderX

Caregiver Demographics

Demographic data were collected from the caregivers of most enrolled patient participants (n=104; Table 2). Most caregivers reported being White or Caucasian (85.58%, 89/104). This likely reflects the demographic makeup of Missouri where approximately 80% of the population identify as "White Non-Hispanic" [33]. Most caregivers reported a highest education level attained of high school graduate or GED or some college but no degree (28.85% [30/104] and 28.85% [30/104], respectively). The majority of caregivers reported a household income between US $25,000 and 49,999 or US $0 - 24,999 (38.46% [40/104] and 26.92% [28/104], respectively). The age of caregivers ranged from 20 to 49 years, with a mean age of 31.86 (SD 6.36) years and median age of 31. The demographic distribution of the caregivers of participants who completed the trial was similar to that of the caregivers of all enrolled participants.

Table . Caregiver race, educational attainment, and household income demographics, shown for the full population and stratified by study completion status.

Caregiver's demographics	Percent of enrolled participants[a], n (%)	Percent of trial complete participants[b], n (%)
Race		
American Indian or Alaska Native	3 (2.88)	3 (3.75)
Asian	2 (1.92)	2 (2.5)
Black or African American	14 (13.46)	10 (12.5)
Native Hawaiian or Other Pacific Islander	2 (1.92)	1 (1.25)
White or Caucasian	90 (86.54)	70 (87.50)
Educational attainment		
Less than high school graduate	6 (5.77)	5 (6.25)
High school graduate or GED[c]	30 (28.85)	23 (28.75)
Some college, no degree	30 (28.85)	21 (26.25)
Bachelor's degree	31 (12.5)	12 (15)
Trade or technical school graduate	12 (11.54)	9 (11.25)
Post graduate degree or higher	13 (12.5)	10 (12.5)
Household income (USD)		
0 - 24,999	28 (26.92)	20 (25)
25,000 - 49,999	40 (38.46)	31 (38.75)
50,000 - 74,999	13 (12.5)	11 (13.75)
75,000 - 99,999	9 (8.65)	7 (8.75)
100,000 and over	14 (13.46)	11 (13.75)

[a]N=104.

[b]N=80.

[c]GED: General Educational Development (high school equivalency).

ECHO Autism Clinician Demographics

Seven clinicians participated in this study. Each participating primary care physician completed rigorous mentoring in using the ECHO Autism: EDx model. The majority of participating clinicians (71.43%, 5/7) were female, and 100% were White or Caucasian (not Hispanic or Latino). The age of clinicians ranged from 37 to 72 years, with a mean age of 48.43 (SD 13.51) years and median age of 42 years.

Time From Initial Clinical Question to Autism Diagnosis

ECHO Autism: EDx Workflow With Embedded Device

The total average time from clinical question at study enrollment to final autism diagnosis was 39.22 (SD 57.08) days. The most frequent time elapsed between enrollment and diagnosis was 14 days, and the median time was 13.98 days.

Clinicians spent an average of 6 minutes and 20 seconds (SD 4 min and 36 s) completing the clinical device questionnaire, compared to the typical 20 minutes they previously spent completing the STAT instrument. Caregivers spent an average of 6 minutes and 44 seconds (SD 5 min and 41 s) completing the caregiver device questionnaire. The time spent by caregivers on their questionnaire ranged from a minimum time of 1 minute and 43 seconds to a maximum time of 37 minutes and 40 seconds and a median time of 4 minutes and 44 seconds.

XSL•FO
RenderX

Comparison to Adjacent Specialty Referral Center Wait Times

In the period the study was conducted, adjacent specialty referral centers reported wait times of 180 to 264 days. If participants with a question of delayed development had been referred onto these waitlists after visit one, rather than being evaluated in primary care, their diagnoses would have been delayed by 141 - 225 days.

Device Performance Versus Process as Usual

Participating clinicians provided a definitive autism diagnosis or rule out to all study completers in the primary care setting. The device provided determinate results (ie, positive for autism or negative for autism) for 52.50% (42/80) of patients. Determinate device outputs agreed with the clinician diagnosis in 100% of cases, giving a positive predictive value, negative predictive value, sensitivity, and specificity of 100% (Table 3). The device did not produce any false positives or false negatives. Most indeterminate patients received a negative autism diagnosis from their clinician (78.95%, 30/38). There were no significant differences in the rate at which the device output a determinate versus an indeterminate result based on participants' age (under 48 mo vs over 48 mo; insignificant, P=.36), sex (male vs female; insignificant, P=.81), race or ethnicity (insignificant, P=.78) or socioeconomic status (income brackets 0 - 50k, 50k-100k, 100k+; insignificant, P=.87).

Table . Device output (positive or negative or indeterminate for autism) versus final clinical diagnosis (positive or negative for autism).

	Clinician diagnosis, n (%)	
Device output	Positive autism (n=48)	Negative autism (n=40)
Positive autism (N=40)	40 (100)	0 (0)
Indeterminate (N=38)	8 (21.1)	30 (79)
Negative autism (N=2)	0 (0)	2 (100)

The STAT, in contrast, generates a low or high-risk score for each child (binary classifier); however, without a built-in ability to abstain in cases of high uncertainty, its overall predictive performance for identifying or ruling out autism is significantly lower than that of the device. .

Caregiver Experience

Satisfaction and Likeliness to Recommend

Caregivers reported positive experiences using the device and throughout the clinician evaluation. The vast majority (92.59%, 50/54) of caregivers reported being satisfied (agreed or strongly agreed) with the evaluation their child received from their clinician and endorsed that they would recommend this type of evaluation process to others (92.86%, 65/70). Additionally, 91.78% (67/73) stated they felt comfortable (agreed or strongly agreed) using the device as part of their child's evaluation. The majority of caregivers agreed or strongly agreed that the device app worked without difficulties (81.94%, 59/72). Few users reported issues with filming instructions (11.11%, 8/72), required assistance to use the device (7.14%, 5/70), or had issues with internet connectivity while using the device (8.58%, 6/70)

Impact of Assessment Location on Distance Traveled

Specialty centers required traveling an average of 96.82 (SD 78.38) miles from participants' residences. However, caregivers reported traveling only 51.98 (SD 75.69) miles, on average, to the in-person appointment with their primary care clinician. This is an average of 44.84 miles saved compared to the distance they would have traveled if assessed at the nearest specialty center. The vast majority (92.59%) of caregivers agreed that it was helpful to have their child evaluated in the community by an EAC instead of traveling to a specialist center.

Clinician Experience

Survey data were collected from all seven clinician participants at the start of the study and at the end of the study and in relation to each patient who received an evaluation using the device. Clinicians agreed with the output provided by the device for 100% of determinate cases (Table 4). Most (71.43%, 5/7) clinicians agreed that the device worked well within their clinical practice flow, and more than half (57.14%, 4/7) felt that the device added information to their diagnostic impression.

Table . Level of clinician agreement with determinate device output.

Clinician agreement with device output	Determinate output patients (N=42), n (%)
Strongly agree	29 (69.05)
Agree	13 (30.95)
Neutral or no opinion	0 (0)
Disagree	0 (0)
Strongly disagree	0 (0)

Clinicians reported that barriers associated with using their previous process (ie, STAT) "process as usual" were reduced when administering the device (Table 5). Time, for example, was identified as a barrier by 71.43% (5/7) of clinicians for the process as usual compared to only 28.57% (2/7) for the device. Similarly, child cooperation was reported as a barrier by 42.86% (3/7) of clinicians for ECHO Autism: EDx process as usual but was decreased to 28.57% (2/7) with the use of the device. Family

XSL•FO
RenderX

resistance, reported by 14.29% (1/7) of clinicians as a barrier for process as usual, was not reported as a barrier when using the device. On the other hand, 85.71% (6/7) of clinicians reported technology as a potential barrier with device use, likely because families are required to download an app and independently complete in-app tasks, whereas the STAT was a pen and paper test administered by the clinician.

Additional barriers unique to the device included insufficient clinical information from caregiver videos and limited usefulness of the output generated by the device in cases when an indeterminate result was generated (Table 5). Just over half of clinicians (57.14%, 4/7) reported being unlikely to uniformly recommend the device in lieu of additional direct observation for these cases. While just under half (42.86%, 3/7) of the participating physicians agreed that the device saved them time in the overall diagnostic process, 28.57% (2/7) were uncertain and 28.57% (2/7) disagreed.

Table . Clinician-reported barriers to diagnosis using ECHO autism: EDx process as usual versus the device (N=7).

Barrier	EDx process as usual, n (%)	EDx plus the device, n (%)
Time	5 (71.43)	2 (28.57)
Patient uncooperative	3 (42.86)	2 (28.57)
Resistance from family	1 (14.29)	0 (0)
Technology	0 (0)	6 (85.71)

Discussion

Main Findings

This collaborative observational study was the first to evaluate the use of a novel AI-based autism diagnostic device as part of the ECHO Autism: EDx workflow. Compared to specialty center referral, these primary care–based autism evaluations reduced the time from first clinical concern to diagnosis by approximately 5 - 9 months. Families, on average, traveled 45 miles less to each appointment than they would have if assessed at their nearest specialty center. Participating caregivers reported strong levels of satisfaction with both the device itself and the ECHO Autism diagnostic process more broadly, highlighting that evaluation in a primary care setting was acceptable to families. High levels of caregiver satisfaction with primary care autism evaluations [34] reduced travel burden [35] and significantly shortened time from first concern to diagnosis [36] are consistent with previously published topical study findings.

While most clinicians endorsed that device use was feasible and provided workflow efficiencies compared to their previous process, clinician participants noted device use did not always obviate the need for additional observations. While there was 100% alignment between determinate device predictions and the final clinical diagnosis, in the roughly half of cases where the device abstained from making an autism prediction or rule out, clinicians were required to rely more heavily on their observations, comprehensive DSM-5 history, and clinical expertise to arrive at a diagnostic determination. This emphasizes the need for clinicians seeking to include autism diagnosis in their practices to obtain and maintain the clinical expertise necessary to interpret varied result outputs in their diagnostic determination process.

Technological barriers to device use noted by a minority of caregiver participants included issues with video uploads, video recording directions, and unreliable internet access. Since the time of study completion, the device manufacturers have implemented a number of updates to improve video uploading functionality and in-app recording directions. Future testing to determine the impact of these updates on user experience is recommended.

Limitations

The version of the device assessed in this study included only a positive, negative, or indeterminate output but did not include an accompanying detailed report explaining the result. Since the time of study completion, supporting software has been developed that increases the transparency of the device's AI, mapping collected data to DSM-5 criteria relevant to autism diagnosis and auto-generating a detailed report to reduce clinical documentation burden [25]. Decision threshold updates have also been made to reduce the number of indeterminate outputs that are returned [23]. Future clinician users' perceptions of device utility and efficiencies (particularly in indeterminate cases) may differ from those reported by study participants, given this significant update in functionality.

Additionally, we note that the clinician sample size (7) was small, and their demographics were not representative of the broader US population, with all being White or Caucasian. Clinician study participation was impacted by EACs being in primarily nonacademic, rural communities with less infrastructure and familiarity with participating in research. Currently, the device is only available in English and Spanish, so findings reported here may not generalize to other non-English or non-Spanish speaking populations. While no significant differences in device abstention rates were noted across age, sex, socioeconomic status, or race and ethnicity, the authors recommend that larger studies be conducted to confirm these findings in more racially diverse samples.

Conclusion

The ECHO Autism: EDx plus device workflow was feasible and reduced barriers to diagnosis for both caregiver and clinician users. Compared to referral to adjacent specialist waitlists, use of the ECHO Autism: EDx plus device workflow resulted in a reduced travel burden for families and significantly shortened patient delays from first clinical question to diagnosis and treatment initiation.

XSL•FO
RenderX

These findings may be of significant interest to families with children stuck on long wait lists for autism evaluations, as well as policymakers, clinicians, and payors seeking to find solutions to the current waitlist crisis. Since the Centers for Disease Control and Prevention began tracking rates of autism among 8-year-olds in 2000, prevalence rates have increased by over 300% [37,38]. Meanwhile, the specialist workforce, typically responsible for conducting autism evaluations, has not kept pace with this increase. Currently, there is only an estimated one developmental behavioral pediatrician per 100,000 children [11] and 14 Child and Adolescent Psychiatrists per 100,000 children [39]. This prospective observational study demonstrated that innovative professional development models like ECHO Autism, combined with emerging diagnostic technologies like Canvas Dx, can help equip the much larger primary care workforce to feasibly play a greater role in diagnosing and managing autism without automatic specialist referral. Strategies to further expand the pool of qualified primary care clinicians able to participate in the diagnostic process, and funding to equip them with evidence-based diagnostic supports, should be pursued to unblock specialist waitlists and support timely treatment initiation.

Acknowledgments

No generative artificial intelligence was used in analyzing data or preparing the manuscript. This study was sponsored by Cognoa Inc. All data analysis was conducted by an independent third party.

Data Availability

Data are not publicly available because they contain sensitive patient information. Individual, deidentified, participant data that underlie the results reported in this article and study protocol may be made available to qualified researchers upon reasonable request. Proposals should be directed to research@cognoa.com to gain access. Data requestors will be required to sign a data-sharing agreement prior to access.

Conflicts of Interest

KS has received consulting fees from Cognoa prior to the initiation of this study, was on the Medical Advisory Board for Quadrant Biosciences, and received institutional research support from Autism Speaks. She volunteers as the chair of the AAP Subcommittee on autism. KH, CS, and ST are employees of Cognoa and have Cognoa stock options. ST additionally receives consulting fees for Cognito Therapeutics, volunteers as a board member of the AAP–OC chapter and AAP–California, is a paid advisor for MI10 LLC, and owns stock for NTX, Inc., and HandzIn.

References

1. Maenner MJ, Warren Z, Williams AR, et al. Prevalence and characteristics of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network, 11 Sites, United States, 2020. MMWR Surveill Summ 2023 Mar 24;72(2):1-14. [doi: 10.15585/mmwr.ss7202a1] [Medline: 36952288]
2. Clark MLE, Vinen Z, Barbaro J, Dissanayake C. School age outcomes of children diagnosed early and later with autism spectrum disorder. J Autism Dev Disord 2018 Jan;48(1):92-102. [doi: 10.1007/s10803-017-3279-x] [Medline: 28905160]
3. Harstad E, Hanson E, Brewster SJ, et al. Persistence of autism spectrum disorder from early childhood through school age. JAMA Pediatr 2023 Nov 1;177(11):1197-1205. [doi: 10.1001/jamapediatrics.2023.4003] [Medline: 37782510]
4. MacDonald R, Parry-Cruwys D, Dupere S, Ahearn W. Assessing progress and outcome of early intensive behavioral intervention for toddlers with autism. Res Dev Disabil 2014 Dec;35(12):3632-3644. [doi: 10.1016/j.ridd.2014.08.036] [Medline: 25241118]
5. The state of pediatric autism diagnosis in the US: gridlocks, inequities and missed opportunities persist. Cognoa, Inc. URL: https://cognoa.com/waitlist-crisis-report/ [accessed 2025-10-08]
6. Pierce K, Gazestani VH, Bacon E, et al. Evaluation of the diagnostic stability of the early autism spectrum disorder phenotype in the general population starting at 12 months. JAMA Pediatr 2019 Jun 1;173(6):578-587. [doi: 10.1001/jamapediatrics.2019.0624] [Medline: 31034004]
7. Autism by the numbers: 2023 inaugural annual report. : Autism Speaks; 2023 URL: https://www.autismspeaks.org/sites/default/files/ABN_Annual_Report_2023.pdf [accessed 2025-10-15]
8. Guthrie W, Wetherby AM, Woods J, et al. The earlier the better: an RCT of treatment timing effects for toddlers on the autism spectrum. Autism 2023 Mar 15;27(8):13623613231159153. [doi: 10.1177/13623613231159153] [Medline: 36922406]
9. Monteiro SA, Dempsey J, Berry LN, Voigt RG, Goin-Kochel RP. Screening and referral practices for autism spectrum disorder in primary pediatric care. Pediatrics 2019 Oct;144(4):e20183326. [doi: 10.1542/peds.2018-3326] [Medline: 31515298]
10. Rhoades RA, Scarpa A, Salley B. The importance of physician knowledge of autism spectrum disorder: results of a parent survey. BMC Pediatr 2007 Nov 20;7(1):37. [doi: 10.1186/1471-2431-7-37] [Medline: 18021459]
11. Baum RA, Berman BD, Fussell JJ, et al. Child health needs and the developmental-behavioral pediatrics workforce supply: 2020-2040. Pediatrics 2024 Feb 1;153(Suppl 2):e2023063678H. [doi: 10.1542/peds.2023-063678H] [Medline: 38300001]

XSL•FO
RenderX

12. Gordon-Lipkin E, Foster J, Peacock G. Whittling down the wait time: exploring models to minimize the delay from initial concern to diagnosis and treatment of autism spectrum disorder. Pediatr Clin North Am 2016 Oct;63(5):851-859. [doi: 10.1016/j.pcl.2016.06.007] [Medline: 27565363]
13. Hyman SL, Levy SE, Myers SM, Council on children with disabilities, section on developmental and behavioral pediatrics. Identification, evaluation, and management of children with autism spectrum disorder. Pediatrics 2020 Jan;145(1):e20193447. [doi: 10.1542/peds.2019-3447] [Medline: 31843864]
14. Takayama JI, Williams J. Shortage of developmental-behavioral pediatricians prompts call for collaborative model of care. AAP News. 2023. URL: https://publications.aap.org/aapnews/news/22923/Shortage-of-developmental-behavioral-pediatricians?autologincheck=redirected [accessed 2025-10-08]
15. Top 10 leadership resolutions. AAP News. 2023. URL: https://publications.aap.org/aapnews/news/25346/Leadership-Conference-Top-resolution-calls-for?autologincheck=redirected [accessed 2025-10-08]
16. Penner M, Senman L, Andoni L, et al. Concordance of diagnosis of autism spectrum disorder made by pediatricians vs a multidisciplinary specialist team. JAMA Netw Open 2023 Jan 3;6(1):e2252879. [doi: 10.1001/jamanetworkopen.2022.52879] [Medline: 36696109]
17. McNally Keehn R, Swigonski N, Enneking B, et al. Diagnostic accuracy of primary care clinicians across a statewide system of autism evaluation. Pediatrics 2023 Aug 1;152(2):e2023061188. [doi: 10.1542/peds.2023-061188] [Medline: 37461867]
18. Hyman SL, Kroening ALH. Diagnosing autism spectrum disorders in primary care: when you know, you know. Pediatrics 2023 Aug 1;152(2):e2023062279. [doi: 10.1542/peds.2023-062279] [Medline: 37461878]
19. Ahlers K, Gabrielsen TP, Ellzey A, et al. A pilot project using pediatricians as initial diagnosticians in multidisciplinary autism evaluations for young children. J Dev Behav Pediatr 2019 Jan;40(1):1-11. [doi: 10.1097/DBP.0000000000000621] [Medline: 30461593]
20. McNally Keehn R, Ciccarelli M, Szczepaniak D, Tomlin A, Lock T, Swigonski N. A statewide tiered system for screening and diagnosis of autism spectrum disorder. Pediatrics 2020 Aug;146(2):e20193876. [doi: 10.1542/peds.2019-3876] [Medline: 32632023]
21. Kaufman NK. Rethinking “gold standards” and “best practices” in the assessment of autism. Appl Neuropsychol Child 2022;11(3):529-540. [doi: 10.1080/21622965.2020.1809414] [Medline: 32847428]
22. FDA authorizes marketing of tool to aid in diagnosis of autism spectrum disorder 2021 sept. AAP. 2021. URL: https://publications.aap.org/aapnews/news/16916/FDA-authorizes-marketing-of-tool-to-aid-in [accessed 2025-10-08]
23. Wall DP, Liu-Mayo S, Salomon C, Shannon J, Taraman S. Optimizing a de novo artificial intelligence-based medical device under a predetermined change control plan: Improved ability to detect or rule out pediatric autism. Intelligence-Based Medicine 2023;8:100102. [doi: 10.1016/j.ibmed.2023.100102]
24. Megerian JT, Dey S, Melmed RD, et al. Evaluation of an artificial intelligence-based medical device for diagnosis of autism spectrum disorder. NPJ Digit Med 2022 May 5;5(1):57. [doi: 10.1038/s41746-022-00598-6] [Medline: 35513550]
25. Salomon C, Heinz K, Aronson-Ramos J, Wall DP. An analysis of the real world performance of an artificial intelligence based autism diagnostic. Sci Rep 2025 Aug 12;15(1):29503. [doi: 10.1038/s41598-025-15575-8] [Medline: 40796609]
26. Mazurek MO, Brown R, Curran A, Sohl K. ECHO autism: a new model for training primary care providers in best-practice care for children with autism. Clin Pediatr 2017;56:247-256. [doi: 10.1177/0009922816648288]
27. Sohl K, Mazurek MO, Brown R. ECHO autism: using technology and mentorship to bridge gaps, increase access to care, and bring best practice autism care to primary care. Clin Pediatr (Phila) 2017 Jun;56(6):509-511. [doi: 10.1177/0009922817691825]
28. Stone WL, Coonrod EE, Turner LM, Pozdol SL. Psychometric properties of the STAT for early autism screening. J Autism Dev Disord 2004 Dec;34(6):691-701. [doi: 10.1007/s10803-004-5289-8] [Medline: 15679188]
29. Sohl K, Kilian R, Brewer Curran A, et al. Feasibility and impact of integrating an artificial intelligence-based diagnosis aid for autism into the extension for community health outcomes autism primary care model: protocol for a prospective observational study. JMIR Res Protoc 2022 Jul 19;11(7):e37576. [doi: 10.2196/37576] [Medline: 35852831]
30. Schieltz KM, Brown R, Nanclares-Nogués V, et al. Evaluating autism diagnosis agreement between primary care physicians/practitioners and experts through ECHO Autism STAT. Res Autism Spectr Disord 2023 Nov;109:102276. [doi: 10.1016/j.rasd.2023.102276]
31. Canvas dx indications for use (english-spanish IFU). Cognoa Inc. 2021. URL: https://cognoa.com/wp-content/uploads/Canvas-Dx-IFUs_English-Spanish-2.pdf [accessed 2025-09-23]
32. Autism program and resource analysis report- median wait times to first appointment- missouri autism centers 2022 and 2023 fiscal year numbers. Missouri Department of Mental Health. 2022. URL: https://dmh.mo.gov/ [accessed 2025-10-08]
33. Missouri: population & diversity. Data USA. 2023. URL: https://datausa.io/profile/geo/missouri?utm_source=chatgpt.com [accessed 2025-10-08]
34. Nasir AK, Strong-Bak W, Bernard M. Diagnostic evaluation of autism spectrum disorder in pediatric primary care. J Prim Care Community Health 2024;15:21501319241247997. [doi: 10.1177/21501319241247997] [Medline: 38650542]
35. Mazurek MO, Curran A, Burnette C, Sohl K. ECHO autism STAT: accelerating early access to autism diagnosis. J Autism Dev Disord 2019 Jan;49(1):127-137. [doi: 10.1007/s10803-018-3696-5] [Medline: 30043354]

XSL•FO
RenderX

36. Hine JF, Allin J, Allman A, et al. Increasing access to autism spectrum disorder diagnostic consultation in rural and underserved communities: streamlined evaluation within primary care. J Dev Behav Pediatr 2020 Jan;41(1):16-22. [doi: 10.1097/DBP.0000000000000727] [Medline: 31490843]
37. Autism and Developmental Disabilities Monitoring Network Surveillance Year 2008 Principal Investigators, Centers for Disease Control and Prevention. Prevalence of autism spectrum disorders--Autism and Developmental Disabilities Monitoring Network, 14 sites, United States, 2008. MMWR Surveill Summ 2012 Mar 30;61(3):1-19. [Medline: 22456193]
38. Prevalence and early identification of autism spectrum disorder among children aged 4 and 8 years — autism and developmental disabilities monitoring network, 16 sites, united states, 2022. Centers for Disease Control and Prevention (US). 2025. URL: https://www.cdc.gov/mmwr/volumes/74/ss/pdfs/ss7402a1-H.pdf [accessed 2025-10-08]
39. Policy statement on behavioral healthcare workforce shortage. American Academy of Child & Adolescent Psychiatry (AACAP). 2023. URL: https://www.aacap.org/AACAP/Policy_Statements/2023/Behavioral_Healthcare_Workforce_Shortage.aspx?utm_source=chatgpt.com [accessed 2025-10-08]

Abbreviations

AI: artificial intelligence
EAC: ECHO autism clinician
EDx: Early Diagnosis Program
STAT: Screening Tool for Autism in Toddlers and Young Children

Edited by A Mavragani; submitted 16.07.25; peer-reviewed by A Rynkiewicz, J Megerian; revised version received 29.09.25; accepted 30.09.25; published 21.10.25.

Please cite as:
Sohl K, Linstead E, Heinz K, Lledo EE, Brewer Curran A, Mahurin M, Nanclares-Nogués V, Salomon C, Seal M, Taraman S Integration of an Artificial Intelligence–Based Autism Diagnostic Device into the ECHO Autism Primary Care Workflow: Prospective Observational Study
JMIR Form Res 2025;9:e80733
URL: https://formative.jmir.org/2025/1/e80733
doi:10.2196/80733

© Kristin Sohl, Erik Linstead, Kelianne Heinz, Elia Eiroa Lledo, Alicia Brewer Curran, Melissa Mahurin, Valeria Nanclares-Nogués, Carmela Salomon, Minda Seal, Sharief Taraman. Originally published in JMIR Formative Research (https://formative.jmir.org), 21.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Effectiveness of WeChat-Based Plus Scene-Graphics Health Education for Rehabilitation After Open Elbow Arthrolysis: Historical Control Study

Danling Fang[1], MSC; Yin Wang[1], BS; Shiyang Yu[2], MD; Wei Wang[2], MD; Pingchen Lu[3], BS

[1]Operation Room, Shanghai Sixth People's Hospital Affiliated to Shanghai Jiao Tong University School of Medicine, 600 YiShan Road, Shanghai, China

[2]Department of Orthopedics, Shanghai Sixth People's Hospital Affiliated to Shanghai Jiao Tong University School of Medicine, Shanghai, China

[3]Department of Rehabilitation, Shanghai Jiao Tong University Affiliated Sixth People's Hospital, Shanghai, China

Corresponding Author:
Yin Wang, BS
Operation Room, Shanghai Sixth People's Hospital Affiliated to Shanghai Jiao Tong University School of Medicine, 600 YiShan Road, Shanghai, China

Abstract

Background: Elbow stiffness often hinders daily tasks. Open elbow arthrolysis is effective but requires long-term postoperative rehabilitation. Traditional health education does not significantly improve patient cooperation or results. Handy and engaging tools such as WeChat and scene graphics may help.

Objective: This study aims to assess the efficacy of WeChat-based health education combined with scene graphics following open elbow arthrolysis.

Methods: This historical control study involved patients aged 18 years and older who underwent open elbow arthrolysis, had normal communication skills, and were proficient in using WeChat. Eligible patients were divided into 2 groups based on admission time: the control group (56 patients, enrolled from January to June 2021) and the WeChat group (56 patients, enrolled from July to December 2021). The control group had received traditional health education, whereas the WeChat group received health education using WeChat and scene graphics. Information in 4-part comics was shared through a WeChat public account. Patients accessed this account to receive daily lessons during hospitalization, followed by online instruction in a WeChat group after discharge until 12 weeks postoperatively. Outcome data were collected at 1, 6, and 12 weeks postoperatively. The primary outcome was elbow range of motion; secondary outcomes were elbow function, quality of life, and complication incidence.

Results: The elbow flexion angle improved from 71.5° (SD 4.2°) to 124.2° (SD 11.7°) in the WeChat group and from 71.7° (SD 4.6°) to 114.4° (SD 13.6°) in the control group (difference 10.0°, 95% CI 4.9–15.1, $P<.001$). The mean elbow extension angle improved from 29.6° (SD 6.0°) to 6.4° (SD 2.5°) in the WeChat group and from 28.8° (SD 3.8°) to 10.1° (SD 3.4°) in the control group (difference –4.5°, 95% CI –6.5 to –2.5, $P<.001$). The mean forearm pronation angle improved from 31.9° (SD 4.0°) to 66.9° (SD 7.3°) in the WeChat group and from 33.0° (SD 4.2°) to 63.1° (SD 7.2°) in the control group (difference 4.9°, 95% CI 2.0-7.8, $P=.001$). The mean forearm supination angle improved from 30.2° (SD 3.7°) to 71.8° (SD 4.8°) in the WeChat group and from 30.4° (SD 4.1°) to 64.2° (SD 9.8°) in the control group (difference 7.7°, 95% CI 4.4-11.0, $P<.001$). The mean Mayo Elbow Performance Score increased from 58.0 (SD 3.7) to 80.4 (SD 5.7) in the WeChat group and from 58.9 (SD 2.8) to 75.8 (SD 6.9) in the control group (difference 5.5 points, 95% CI 2.8–8.2, $P<.001$). The mean 36-item Short Form Health Survey questionnaire score increased from 44.4 (SD 6.6) to 82.0 (SD 7.1) in the WeChat group and from 44.0 (SD 6.4) to 75.0 (SD 11.2) in the control group (difference 6.6 points, 95% CI 2.6-10.6, $P=.002$).

Conclusions: WeChat-based health education combined with scene graphics was found to significantly improve elbow range of motion, elbow function, and quality of life.

Trial Registration: Chinese Clinical Trial Registry ChiCTR2000036004; https://www.chictr.org.cn/showproj.html?proj=59168

(JMIR Form Res 2025;9:e58218) doi:10.2196/58218

KEYWORDS

elbow stiffness; open elbow arthrolysis; WeChat; scene graphics; range of motion; health education; rehabilitation; mobile health

XSL•FO
RenderX

Introduction

Normal elbow mobility is fundamental in daily life. Elbow stiffness is a common condition, affecting up to 56% of individuals following elbow trauma [1,2]. It is characterized by an inability of the elbow to meet the patient's needs for daily activities, work, and entertainment due to a limited range of motion (ROM) or rotation [3,4]. Open elbow arthrolysis (OEA) is the most common surgical intervention and has been shown to restore elbow function and ROM effectively [3,4]. The elbow can be released by removing bony impediments or scar tissue within the olecranon fossa, achieving intraoperative flexion greater than 130° and extension less than 10° [5,6]. However, postoperative rehabilitation is necessary for full elbow function recovery [4-8].

A growing body of evidence suggests that timely and proper postoperative rehabilitation leads to favorable outcomes. By contrast, inadequate rehabilitation severely hampers elbow recovery [4-8] and contributes to many postoperative complications, such as heterotopic ossification, new onset or exacerbation of ulnar nerve symptoms, and recurrent elbow stiffness [3,4,9,10]. According to incomplete statistics, the incidence rates of postoperative elbow issues, probability of renewed elbow stiffness, rate of secondary surgery, and rate of dissatisfaction are 56%‑87% [5], 8.4%‑47% [5,6], 10%‑34% [6,7], and 21% [7], respectively. Therefore, the precision of postoperative rehabilitation and patient adherence must be improved. However, achieving optimal postoperative rehabilitation is challenging due to its prolonged and intricate nature [6,11,12]. Standard postoperative rehabilitation lasts 1 year and is divided into 3 phases [6,13,14]. The first, acute phase begins on the first day and extends to 6 weeks after surgery; it involves passive, assisted, and active elbow flexion and extension exercises [6,13,14]. Rehabilitation starts with 30 movements on the first day, increasing by 30 daily until reaching 300 movements per day [6,13]. The second, subacute phase spans from 6 weeks to 3 months and requires maintaining 300 daily movements [13]. The third, functional phase lasts from 3 months to 1 year; during this phase, weight-bearing exercises are added, and intensity increases daily [13,14]. In addition to initial hospital-based rehabilitation, patients must continue their rehabilitation at home, making a comprehensive and accurate understanding of the relevant knowledge crucial for effective recovery [12,13,15]. However, with the increasing prevalence of enhanced recovery after surgery protocols [16,17], hospital stays have shortened, making traditional health education methods insufficient for patients to acquire adequate knowledge. Therefore, more efficient and feasible health education approaches are needed.

Telemedicine services have become increasingly popular for health education in recent years. Health care professionals use electronic information and communication technology to deliver and support health care services to patients, regardless of their location or transportation limitations [18]. WeChat is one of China's most widely and frequently used free platforms [19-21] and offers convenient telemedicine services. It supports a variety of data types, including text, voice calls, videos, and images, enabling clear and accessible communication [20], which attracts many users. By the end of 2020, there were 1.2 billion monthly active users worldwide [21,22], many of whom actively seek online health care and health information exchange [23]. A previous survey revealed that one-third of participants regularly read health information articles on WeChat [24], and 98.53% used WeChat to search for health information [23]. The platform has been used in the health management of conditions such as public health literacy [25], cancer [26,27], asthma [28], chronic obstructive pulmonary disease [29], hypertension [30], myopia [31], coronary artery disease [32], and others. However, no studies have focused on its use for postoperative rehabilitation in patients with elbow stiffness.

Medical humanities is an interdisciplinary, multifaceted field that combines scientific knowledge and skills with respectful, compassionate care sensitive to patients' and their families' values, autonomy, and cultural backgrounds [33]. Medical comics are among the various medical humanities approaches that can potentially address challenging situations within medical settings [33,34]. They have been used to promote adolescent peer support and healthy sexual behaviors [34], raise awareness about diabetes [35], improve the quality of palliative care [36], and increase awareness of whole-body donation [37]. Scene-based graphics are an approach that simplifies knowledge points into comics and conveys messages through various humorous scenarios, making the information easily understandable and memorable. The graphics have transformed activities of daily living scales into comics for rapid and straightforward evaluation [38]. Additionally, interactive games have been incorporated to complement the corresponding knowledge [39]. These games can be designed based on key knowledge points and learner characteristics, encouraging participants to engage more actively in learning and apply their knowledge more effectively. However, this approach is still in its early stages and is rarely used in postoperative rehabilitation. This study was conducted as a historical control study to evaluate the efficacy of WeChat-based health education combined with scene graphics in the postoperative rehabilitation of OEA.

Methods

Study Design

A 2-armed historically controlled study was conducted at a tertiary hospital in Shanghai, China.

Patient characteristics between the 2 groups were carefully analyzed and matched to minimize potential selection biases related to time. Patients were stratified by age, gender, type of elbow injury, and severity of the condition. For each stratum, the participants admitted from January to June 2021 were included in the control group, whereas those admitted from July to December 2021 with similar characteristics were assigned to the WeChat group. The follow-up period ended in March 2022.

Both groups received identical intervention content for the same duration. A trained research nurse collected baseline data upon admission, and a research doctor collected outcome data at baseline and at the first, sixth, and twelfth weeks after surgery.

XSL•FO
RenderX

To ensure consistency in data collection, both the research nurse and research doctor received standardized training, and a detailed data collection manual defining each variable and measurement method was developed.

Ethical Considerations

This study was registered with the Chinese Clinical Trial Registry (ChiCTR2000036004) and approved by the Ethics Review Committee of Shanghai Sixth People's Hospital (approval number 2020 - 028). Patients were informed of the study's details during enrollment, including its purpose, procedures, risks, and benefits. They were also told about the confidentiality and waiver agreements, which guaranteed data protection and the right to withdraw without affecting their treatment. After understanding these terms, patients provided written consent to participate (see Multimedia Appendix 1). This research conformed to the Consolidated Standards of Reporting Trials (CONSORT) checklist, with adherence details provided in Multimedia Appendix 2.

Participant Recruitment

Upon admission, patients received recruitment notifications through leaflets. These notices, which included an introduction to the research and a contact number, were distributed by research nurses after standardized admission education in the orthopedic ward.

To minimize potential biases, the nurses were trained to follow a scripted message (eg, "Whether or not you participate in this study will not affect your treatment"). Patients interested in joining this study contacted the provider number and completed an expression-of-interest form electronically, which was collected by a neutral coordinator. They then observed a 24-hour cool-off period during which they could withdraw at any time. Patients who confirmed their willingness to participate were assigned to research nurses. Basic participant characteristics were collected by a trained surgeon-nurse team through face-to-face communication, physical examination, and elbow function assessment using validated tools.

Patients were eligible if they were older than 18 years, diagnosed with elbow stiffness(see Multimedia Appendix 1), had undergone OEA for the first time, were conscious (Glasgow Coma Scale score of 15), could communicate normally in Chinese (Mandarin), possessed a mobile device (eg, smartphone or tablet) with internet access, and were capable of using WeChat. Exclusion criteria included severe systemic diseases (eg, uncontrolled hypertension with blood pressure consistently above 180/110 mmHg, poorly controlled diabetes with HbA_{1c} [glycated hemoglobin] >9%, or active malignancy); surgery within the past 6 months; comorbidities such as severe rheumatoid arthritis or peripheral neuropathy that could interfere with elbow rehabilitation; and known allergies to medications or materials used during the surgical procedure.

Sample Size

The pilot study's elbow extension and flexion degrees at 6 weeks postsurgery were used to estimate the sample size [39]. Before applying the sample size calculation formula, a variance homogeneity test was performed for elbow flexion and extension angles in the 2 groups. With a test power ($1 - \beta$) of 80%, a significance level (α) of 5% (2-sided), a 1:1 group size ratio, and elbow flexion of 128.6° (SD 5.8°) and 119.4°(SD 9.8°), respectively (after confirming variance homogeneity), the required sample size was calculated as 24 patients. Similarly, for elbow extension angles of 14.3° (SD 11.3°) and 22.9° (SD 8.8°), respectively (after confirming variance homogeneity), 14 patients were required. Considering a 20% attrition rate, the adjusted minimum sample size was 29 participants per group. As patients' responses to program participation were positive, 56 participants were ultimately included in each group.

Blinding

No blinding was implemented in this study. The surgical team and research members were aware of group allocations, as the intervention required explicit instructions for the WeChat-based rehabilitation program. Patients were enrolled into different groups at varying admission times to minimize interaction, and all participants were informed about the study procedures without masking group details. Primary outcomes (eg, elbow ROM, Mayo Elbow Performance Score [MEPS]) were measured using standardized protocols with calibrated instruments to minimize subjective bias. All data were recorded electronically and analyzed by a statistician blinded to group allocations.

Interventions

All participants underwent OEA performed by the same surgical team and received health education from the same interveners. The elbow rehabilitation knowledge was based on the book "The Elbow: Principles of Surgical Treatment and Rehabilitation" [14] and the research from Sun et al [6,13]. The control group received 12 weeks of traditional health education, while the WeChat group received 12 weeks of WeChat-based health education combined with scene graphics. Interventions were conducted from the first day up to 12 weeks after surgery. Typically, 3-5 patients were hospitalized simultaneously, and interventions started concurrently. Patients who were lost to follow-up or who changed treatment plans (ie, participated in another postoperative project that might have affected the results) were withdrawn from the intervention (see Multimedia Appendix 1).

Control Group

The control group received a card with a QR code on the second day after admission. Patients were instructed to access an online manual titled "Elbow Rehabilitation" by scanning the code using WeChat (see Multimedia Appendix 1). They could study this manual independently from the second day of admission until 12 weeks postoperatively. Research nurses addressed patients' questions during morning nursing care at 7:30 AM daily. Additionally, patients could ask questions during nursing sessions. The rehabilitation process was carried out 3 times a day, including 1 guided session and 2 independent sessions. It began with 30 exercise sets on the first day, increasing by 30 sets daily until reaching 300 sets per day. A rehabilitation therapist provided guidance every day at 9:00 AM for 30 minutes in the multimedia demonstration classroom, correcting any incorrect movements. For the other 2 independent sessions,

XSL•FO
RenderX

research nurses regularly patrolled the exercise area and promptly corrected any movement inaccuracies they observed.

After discharge, weekly follow-up calls were made to ensure continuity of the recovery process. These calls involved checking the patient's current condition, answering any questions about ongoing rehabilitation, and providing guidance based on their progress.

WeChat Group

The multidisciplinary team comprised a surgeon, a nurse manager, a rehabilitation therapist, a platform maintenance specialist, and elbow joint nurse specialists who collaborated to develop the health education scheme. The surgeon oversaw the intervention, trained team members, and assessed patients' elbow function. The nurse manager supervised the implementation of the intervention, while the rehabilitation therapist developed elbow rehabilitation content and instructed patients on the exercises. The 3 nurse specialists delivered the health education.

The health education materials used were the same as those provided to the control group. The research team then transformed these materials into a comic book. The process began with the team crafting a script and a set of questions for the comic, which were submitted to a surgeon for review. After discussions and refinements among the team members, the content was finalized. The research team then partnered with a cultural media company to convert the script into a visual comic format (see Multimedia Appendix 3). After completing the comic book, the team created a WeChat public account featuring the "Elbow Joint Zone," which was designed and divided by platform maintenance specialists into 4 information modules based on the comic book (for content and module names, see Table S1 in Multimedia Appendix 1). The research team then discussed the offline class schedule and arranged the modules according to difficulty (see Table S2 in Multimedia Appendix 1). To further increase engagement, the team also created PowerPoint (Microsoft Corporation) presentations used in offline education through interactive games.

Table . Baseline characteristics of 112 adult patients with posttraumatic elbow stiffness who underwent open elbow arthrolysis, comparing the WeChat intervention group with a control group in a historical control study conducted at a tertiary first-class hospital in Shanghai, China, between January 2021 and March 2022.

Characteristics	Total (n=112)	WeChat group (n=56)	Control group (n=56)	*t* test (*df*)[a]/chi-square (*df*)/Mann-Whitney *U* test	*P* value
Age (years), mean (SD)	36.0 (3.1)	35.98 (3.4)	36.1 (2.9)	*0.121 (110)*	.90
Sex, n (%)				*0.895 (1)*	.34
Women	59 (52.7)	32 (57.1)	27 (48.2)		
Men	53 (47.3)	24 (42.9)	29 (51.8)		
Highest level of education, n (%)				*4.685 (4)*	.32
Primary school	2 (1.8)	1 (1.8)	1 (1.8)		
Junior high school	13 (11.6)	7 (12.5)	6 (10.7)		
Senior high school	26 (23.2)	12 (21.4)	14 (25.0)		
Junior college	44 (39.3)	18 (32.1)	26 (46.4)		
University degree or above	27 (24.1)	18 (32.1)	9 (16.1)		
BMI (kg/m^2), mean (SD)	21.6 (1.81)	21.5 (1.8)	21.77 (1.78)	*0.793 (110)*	.43
Disease side, n (%)				*0.144 (1)*	.71
Left	60 (53.6)	31 (55.4)	29 (51.8)		
Right	52 (46.4)	25 (44.6)	27 (48.2)		
Duration of disease (months), median (25th percentile-75th percentile)	8.2 (7.4-8.8)	8.2 (7.5-8.8)	8.0 (7.2-8.8)	–0.136	.26
Duration of surgery (minutes), median (25th percentile-75th percentile)	125 (124-126)	125 (124-126)	125 (124-126)	–0.085	.93

[a]Independent sample *t* test values are in italics.

XSL•FO
RenderX

Table . Postoperative functional and quality-of-life outcomes, including elbow flexion/extension, forearm pronation/supination, MEPS[a], and SF-36[b] scores, were assessed preoperatively and at 1, 6, and 12 weeks postsurgery in 112 adult patients with posttraumatic elbow stiffness who underwent open elbow arthrolysis. This historical control study, conducted at a tertiary first-class hospital in Shanghai, China (January 2021 to March 2022), compared outcomes between a WeChat-based intervention group and a control group.

Outcomes	Total (n=112)	WeChat group (n=56)	Control group (n=56)	*t* test (*df*)	*P* value
Elbow flexion[c] (degrees), mean (SD)					
Baseline	71.6 (4.4)	71.5 (4.2)	*71.7 (4.6)*	*0.236 (110)*	.81
Postoperative 1 week	84.6 (10.2)	86.0 (8.8)	83.3 (11.4)	*−1.392 (102.94)*	.17
Postoperative 6 weeks	101.3 (10.3)	103.0 (10.3)	99.6 (10.1)	*−1.773 (110)*	.08
Postoperative 12 weeks	119.3 (13.6)	124.2 (11.7)	114.4 (13.6)	*−4.063 (110)*	<.001
Elbow extension[d] (degrees), mean (SD)					
Baseline	29.2 (5.0)	29.6 (6.0)	28.8 (3.8)	*−0.903 (92.79)*	.37
Postoperative 1 week	22.9 (6.1)	22.9 (7.0)	23.0 (5.2)	*0.061 (100.90)*	.95
Postoperative 6 weeks	14.7 (6.3)	13.7 (5.4)	15.7 (5.9)	*1.716 (110)*	.06
Postoperative 12 weeks	8.2 (3.5)	6.4 (2.5)	10.1 (3.4)	*6.427 (101.95)*	<.001
Forearm pronation[e] (degrees), mean (SD)					
Baseline	32.5 (4.1)	31.9 (4.0)	33.0 (4.2)	*1.387 (110)*	.17
Postoperative 1 week	46.2 (7.6)	45.6 (7.2)	46.8 (8.0)	0.880 (110)	.38
Postoperative 6 weeks	58.9 (8.9)	60.9 (8.8)	56.8 (8.4)	*−2.505 (110)*	.01
Postoperative 12 weeks	65.0 (7.5)	66.9 (7.3)	63.1 (7.2)	*−2.794 (110)*	.006
Forearm supination[f] (degrees), mean (SD)					
Baseline	30.3 (3.9)	30.2 (3.7)	30.4 (4.1)	*0.168 (110)*	.87
Postoperative 1 week	44.9 (8.6)	45.3 (8.4)	44.5 (8.9)	*−0.469 (110)*	.64
Postoperative 6 weeks	57.8 (10.0)	59.8 (10.0)	55.9 (9.8)	*−2.062 (110)*	.04
Postoperative 12 weeks	68.0 (8.6)	71.8 (4.8)	64.2 (9.8)	*−5.183 (80.03)*	<.001
MEPS[g] scores, mean (SD)					
Baseline	58.5 (3.3)	58.0 (3.7)	58.9 (2.8)	*1.472 (110)*	.14
Postoperative 1 week	60.9 (4.8)	61.1 (6.5)	60.7 (2.1)	*−0.509 (66.70)*	.61
Postoperative 6 weeks	68.4 (4.8)	69.1 (5.2)	67.8 (4.2)	*−1.509 (110)*	.13
Postoperative 12 weeks	78.1 (6.7)	80.4 (5.7)	75.8 (6.9)	*−3.829 (110)*	<.001
SF-36[h] scores, mean (SD)					
Baseline	44.2 (6.5)	44.4 (6.6)	44.0 (6.4)	*−0.334 (110)*	.74
Postoperative 1 week	55.0 (7.8)	55.3 (7.0)	54.7 (8.6)	*−0.399 (110)*	.69

XSL•FO
RenderX

Outcomes	Total (n=112)	WeChat group (n=56)	Control group (n=56)	*t* test (*df*)	*P* value
Postoperative 6 weeks	70.3 (11.0)	71.0 (10.5)	69.7 (11.4)	*−0.619(110)*	.54
Postoperative 12 weeks	78.5 (10.0)	82.0 (7.1)	75.0 (11.2)	*−3.952(93.07)*	<.001

[a]MEPS: Mayo Elbow Performance Score.

[b]SF-36: 36-item Short Form Health Survey.

[c]Elbow.flexion: the normal range is 0° - 150°; the study target is 0° - 120°. A value closer to 150° indicates better function.

[d]Elbow extension: ideally, 0° - 5°; for this study, 0° - 10°. Higher values signal instability, reflecting poorer function.

[e]Forearm pronation: in this study, the target range was 50° - 70°, approaching the normal 70° - 85° range for better function.

[f]Forearm supination: the goal of this study is 50° - 70°; ideally, alignment with the normal range of 70° - 85° indicates better functionality.

[g]The scores range from 0 to 100 points, and are categorized as follows: excellent (90-100), good (75-89), fair (60-74), and poor (0 - 59).

[h]This research employed physiological functioning as a metric, ranging from 0 to 100 points, with higher scores corresponding to enhanced quality of life.

On the second day of admission, the participants received the same card as the control group, featuring a QR code for the WeChat public account. They accessed the materials using the same procedure as the control group. The intervention consisted of 5 steps, as outlined below.

- Step 1 (Knowledge memory): The research nurse directed patients to access the Elbow Special Zone on the WeChat public account. Patients were instructed to dedicate fragmented time to self-studying the assigned knowledge until they had memorized it. They were encouraged to consult the research team for guidance if needed.
- Step 2 (Picture-based cloze test): The research nurse conducted a multimedia classroom session in the orthopedic ward using PowerPoint to display a comic with blank spaces that patients needed to fill in with the correct knowledge. Patients took turns answering and were praised for correct responses. For incorrect answers, the nurse provided detailed guidance. Patients who missed the session due to clinical examinations received face-to-face guidance.
- Step 3 (Error correction): The research nurse used a similar approach as in step 2. Patients were shown pictures with missing words related to high-frequency mistakes and asked to fill them in until they fully understood the content.
- Step 4 (Keypoint summary): The research nurse summarized key points and emphasized common problems discussed in step 2. These materials were compiled into a Word (Microsoft Corporation) document, printed for each patient, and the patients were urged to review them. The research team monitored learning progress through random questioning during daily ward rounds at 7:30 AM.
- Step 5 (Motion instruction): The rehabilitation therapist provided 1-on-1 instruction to teach proper movements and corrected mistakes until patients achieved the correct motions. Patients also learned independently via the WeChat public account.

Steps 2-5 were scheduled over 6 days, starting 1 day before surgery and continuing until 5 days after surgery (excluding the operation day). Sessions were held daily at 9:00 AM.

Before discharge, the research nurse created WeChat groups, each consisting of 3-5 patients, to ensure high-quality health education. The group names followed the format “group number + elbow care + operation date,” and patients used their name and telephone number as a unified identifier. On the day of discharge, the WeChat health education program was introduced to patients through 4 processes.

- Process 1 (Rehabilitation check-in): Every patient must submit a video in the WeChat group before 7:00 PM daily. The video should demonstrate all the motions the patient was required to complete that day and last no more than 5 minutes. The rehabilitation therapist reviewed the videos and corrected any incorrect movements. The @ function was used to remind patients who forgot to submit their videos.
- Process 2 (Knowledge consolidation): The research nurse used Questionnaire Star to design a “picture-based cloze test,” which was posted to the WeChat group at 8:00 AM daily. Patients were expected to submit their answers by 7:00 PM. The @ function was used as a reminder if necessary. Patients could also review the material via the WeChat public account.
- Process 3 (Error correction): The research nurse collected common error points and provided the correct answers in a Word document. Patients were reminded to review this document to improve their understanding.
- Process 4 (Answer questions and clarify doubts): Every evening from 7 - 9 PM, the research nurse summarized key points and typical questions from process 2, and assessed the effectiveness of process 3 through random inquiries. Detailed answers were provided to any patient’s doubts.

Measures

Primary Outcome Measures

Elbow ROM was measured using a handheld goniometer and included 4 angles: elbow flexion, extension, forearm pronation, and supination [40,41]. A surgeon performed 3 measurements for each angle using 3 landmarks—the lateral epicondyle, the tip of the acromion process, and the midpoint of the wrist—and the average ROM was calculated.

Secondary Outcome Measures

The secondary outcomes were elbow function, quality of life, and postoperative complications. (1) Elbow function was assessed using the MEPS, the most commonly used functional

XSL•FO
RenderX

score [41]. MEPS comprises 4 subscales: pain (45%), ROM (20%), stability (10%), and daily function (25%). The total score ranges from 0 to 100 points and is classified into 4 categories: excellent (90-100), good (75-89), fair (60-74), and poor (0 - 59) [10,40,41]. (2) Patient quality of life was evaluated using the 36-item Short Form Health Survey (SF-36) questionnaire [40]. The SF-36 is a self-rating scale consisting of 8 dimensions: physiological features, physiological function, somatic pain, general health, vitality, social function, affective function, and mental health. The Cronbach α coefficient ranges from 0.65 to 0.94. This study focused on the physiological function dimension. Scores range from 0 to 100, with higher scores indicating a better quality of life. (3) Postoperative complications included heterotopic ossification, new onset or exacerbation of ulnar nerve symptoms, and repeat elbow stiffness. Elbow heterotopic ossification was classified into 3 grades according to the Hastings and Graham classification: grade 1, no functional limitation; grade 2A, limited flexion and extension; grade 2B, limited pronation and supination; grade 2C, combination of 2A and 2B; and grade 3, ankylosis [1]. The new onset or exacerbation of ulnar nerve symptoms was assessed using the Dellon classification, which evaluates sensory symptoms (paresthesia, vibratory perception, and 2-point discrimination) and motor symptoms (muscle weakness and atrophy) [40,42]. Repeat elbow stiffness was diagnosed based on the elbow ROM, clinical presentation, and examination results [40,42].

Data Collection and Analysis

All data were collected using a standardized data collection form and process. Demographic information, including age, sex, educational level, and BMI, as well as clinical data, such as operation duration, affected side, and disease duration, were obtained from medical records and patient interviews. A trained research nurse collected these data to facilitate comparison between the 2 groups. Primary and secondary outcome measures were recorded at baseline and at 1, 6, and 12 weeks postoperatively. The research surgeon collected the 1-week postoperative data during hospitalization, whereas data at 6 and 12 weeks were collected by the surgeon in the outpatient clinic. To ensure consistent follow-up, an automated reminder system, via SMS text messages or automated phone calls, similar to traditional phone calls, was used to remind patients in both groups to attend their reexamination. All data were analyzed by a statistician affiliated with the hospital's clinical research center.

Data analysis was performed using SPSS Statistics version 26.0 (IBM Corp). The chi-square test was used to analyze the categories, including sex, disease side, and the incidence of postoperative complications. The Z test was used to analyze education levels. Independent sample t tests were used to analyze variables that satisfied the normality test, whereas nonparametric tests, specifically the Mann-Whitney U test, were applied for variables that did not meet the normality assumption.

The paired sample t test was applied to compare the preintervention and postintervention values of the primary and secondary evaluation indices. Repeated measures analyses of variance were used to analyze the changes in elbow ROM, the MEPS, and the SF-36 score at each time point (baseline and at 1, 6, and 12 weeks after the operation). The outcomes were reported as mean (SD) for continuous variables and percentages for categorical variables. Nonnormally distributed variables were depicted using median (25th percentile to 75th percentile). A 2-tailed P value <.05 was considered to indicate statistical significance.

Results

Recruitment and Research Processes

Patients with OEA were recruited in January 2021, and the follow-up process ended in March. Figure 1 shows the CONSORT flowchart of this study, detailing recruitment, assessment, and follow-up. A total of 120 patients were assessed, of whom 112 met the inclusion criteria and were enrolled. During the study, 3 patients dropped out, resulting in a dropout rate of 2.7%. Specifically, 2 patients in the control group failed to comply with elbow rehabilitation, and 1 participant in the WeChat group withdrew for personal reasons.

XSL•FO
RenderX

Figure 1. Study flow diagram. OEA: open elbow arthrolysis.

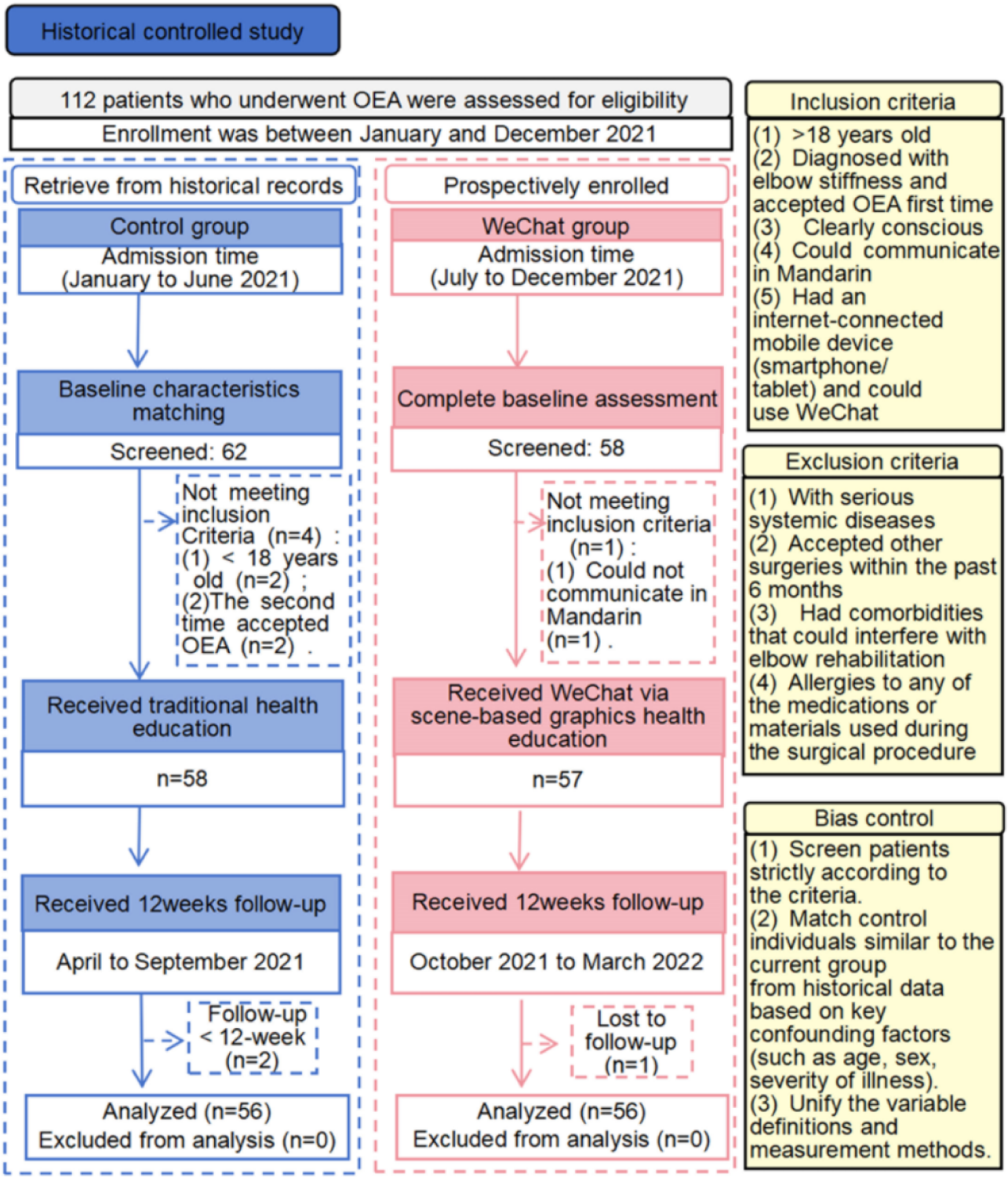

Baseline Characteristics of the Participants

The 2 groups exhibited no significant differences (age, P=.90; sex, P=.34; highest level of education, P=.32; BMI, P=.43; disease side, P=.71; duration of disease, P=.26; Duration of surgery, P=.93) in any baseline characteristics (Table 1). The average age of all participants was 36.0 years. Of the 112 patients, 59 (52.7%) were female and 53 (47.3%) were male. Among the participants, 44 (39.3%) had a junior college education, and 27 (24.1%) had a university degree or higher. The remaining participants included those with senior high school education (26/112, 23.2%), junior high school education (13/112, 11.6%), and primary school education (2/112, 1.8%); 60 (53.6%) patients had left-sided disease, and 52 (46.4%) had right-sided disease. The average duration of disease and surgery was 8.0 months and 124.96 minutes, respectively.

Primary Outcome

The study revealed no significant difference in the ROM of the elbow between the 2 groups at the first week after surgery (elbow flexion, P=.17; elbow extension, P=.95; forearm pronation, P=.38; forearm supination, P=.64). However, at 6

XSL•FO
RenderX

weeks, the WeChat group demonstrated better forearm pronation (P=.01) and forearm supination (P=.04) than the control group. At 12 weeks after surgery, compared with the control group, the WeChat group had better ROM in elbow flexion (P<.001), elbow extension (P<.001), forearm supination (P<.001), and forearm pronation (P=.006; Table 2). The study also revealed statistically significant differences in terms of time (elbow flexion, elbow extension, forearm pronation, and forearm supination; all P<.001), between-group effects (elbow flexion, P<.001; elbow extension: P=.04, forearm pronation: P=.049, forearm supination: P<.001), and interaction effects (elbow flexion: P=.004, elbow extension: P=.005, forearm pronation: P=.002, forearm supination: P=.003) between the 2 groups, as shown in Table 3. Figure 2 illustrates the difference in elbow ROM between the 2 groups. Figure 3 illustrates the change in elbow ROM over time. Table 4 presents the data on adverse events noted in the groups.

Table . Trends in elbow and forearm function (flexion/extension and pronation/supination), MEPS[a], and SF-36[b] scores in the WeChat intervention and control groups. Outcomes were assessed preoperatively and at 1, 6, and 12 weeks postoperatively in 112 adult patients with posttraumatic elbow stiffness who underwent open elbow arthrolysis. This historical control study was conducted at a tertiary first-class hospital in Shanghai, China, from January 2021 to March 2022, to evaluate the impact of a WeChat-based intervention.

Item	F_{time} test (*df*)	*P* value	$F_{between\ groups}$ test (*df*)	*P* value	$F_{time \times group}$ test (*df*)	*P* value
Elbow flexion[c]	493.590 (2.48, 273.22)	<.001	17.718 (1, 110)	<.001	5.083 (2.48, 273.22)	.004
Elbow extension[c]	388.707 (2.63, 289.13)	<.001	4.376 (1, 110)	.04	4.657 (2.63, 289.13)	.005
Forearm pronation[c]	479.929 (2.69, 295.43)	<.001	3.952 (1, 110)	.049	5.200 (2.69. 295.43)	.002
Forearm supination[c]	467.085 (2.53, 277.75)	<.001	17.936 (1, 110)	<.001	5.343 (2.53, 277.75)	.003
MEPS	484.727 (2.00, 220.24)	<.001	4.906 (1, 110)	.03	8.545 (2.00, 220.24)	<.001
SF-36	483.478 (2.29, 251.49)	<.001	4.071 (1, 110)	.046	5.067 (2.29, 251.49)	.005

[a]MEPS: Mayo Elbow Performance Score.

[b]SF-36: 36-item Short Form Health Survey.

[c]Range of motion of the elbow.

XSL•FO
RenderX

Figure 2. Changes from baseline to 12 weeks postoperatively in elbow and forearm function (flexion/extension, pronation/supination), Mayo Elbow Performance Score, and 36-item Short Form Health Survey scores in the WeChat intervention group (n=56) and control group (n=56), among 112 adult patients with posttraumatic elbow stiffness who underwent open elbow arthrolysis. This historical control study was conducted at a tertiary first-class hospital in Shanghai, China, between January 2021 and March 2022.

XSL•FO
RenderX

Figure 3. Time-course comparison of elbow and forearm functional outcomes (flexion/extension, pronation/supination), Mayo Elbow Performance Score (MEPS), and 36-item Short Form Health Survey (SF-36) scores between the WeChat group (n=56) and the control group (n=56) among 112 adult patients with posttraumatic elbow stiffness who underwent open elbow arthrolysis. This historical control study was conducted at a tertiary first-class hospital in Shanghai, China, from January 2021 to March 2022.

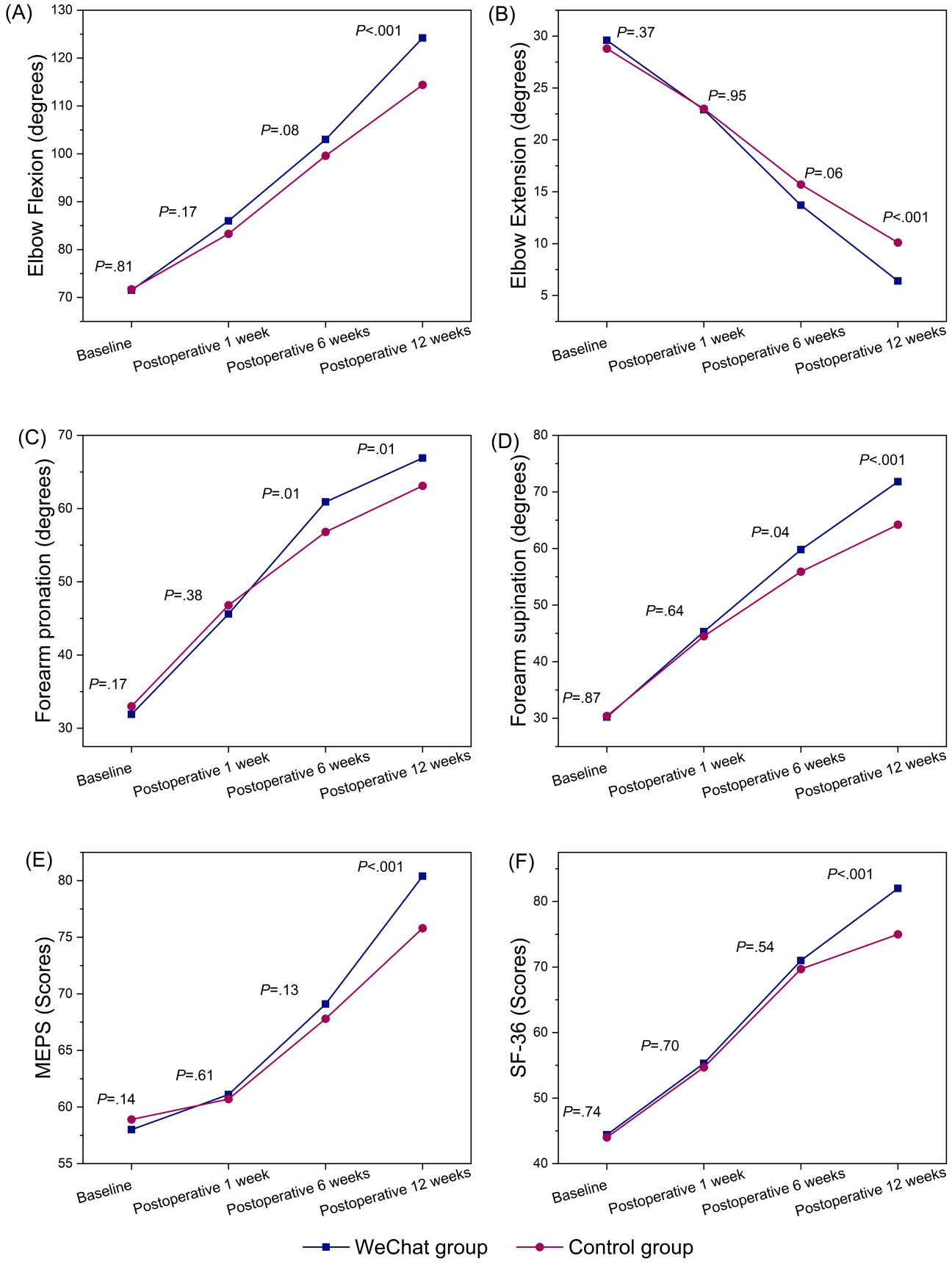

XSL•FO
RenderX

Table . Adverse events in the WeChat group and control group among 112 adult patients with posttraumatic elbow stiffness who underwent open elbow arthrolysis. This historical control study was conducted at a tertiary first-class hospital in Shanghai, China, from January 2021 to March 2022.

Adverse event	WeChat group (n=56), n (%)	Control group (n=56), n (%)
Related to the surgical site	N/A[a]	N/A
Severe bleeding or hematoma	N/A	N/A
Aggravation of wound infection	N/A	N/A
Structural instability	N/A	N/A
Overall physical condition of the patient	N/A	N/A
Serious systemic adverse reactions	N/A	N/A
Serious cardiovascular problems	1 (2)	N/A
Extremely weak or tired	N/A	1 (2)
Abnormal rehabilitation progress	N/A	N/A
Severely limited joint movement and intensified pain	N/A	1 (2)

[a]N/A: not applicable.

Secondary Outcomes

This study revealed no significant differences between the 2 groups in MEPS (P=.61) or SF-36 (P=.69) scores at the first week after surgery. However, the WeChat group had higher MEPS and SF-36 scores at 12 weeks than the control group (all P<.001; Table 2). The study also showed statistically significant differences in the time effect (MEPS: P<.001 and SF-36 score: P<.001), between-group effect (MEPS: P=.03 and SF-36 score: P=.046), and interaction effect (MEPS: P<.001 and SF-36 score: P=.005) between the 2 groups, as shown in Table 3. Figure 2 illustrates the differences between the 2 groups in MEPS and SF-36 scores, and Figure 3 shows the changes in MEPS and SF-36 scores over time.

There was no significant difference in the incidence of postoperative complications between the 2 groups (heterotopic ossification: P=.80; new onset or exacerbation of ulnar nerve symptoms: P>.99; repeat elbow stiffness: P=.56). Among the 56 patients in the WeChat group, 2 (4%) developed heterotopic ossification, specifically grade 2A and grade 2C, respectively. In the control group, which also consisted of 56 patients, 3 (5%) had heterotopic ossification, namely, grade 2A, grade 2B, and grade 2C, respectively. The absolute and relative effect sizes were 0% for grade 2A and grade 2B, and 1.8% and 100% for grade 2C, respectively. In both groups, 1 (2%) patient experienced a new onset or exacerbation of ulnar nerve symptoms, with absolute and relative effect sizes of 0%.

Concerning repeated elbow stiffness, 1 (2%) patient was in the WeChat group, and 2 (4%) patients were in the control group. The absolute and relative effect sizes were 1.8% and 50%, respectively.

Discussion

Principal Findings

This historical control study investigated the efficacy of integrating WeChat with scene-based graphic health education. The results indicated that the WeChat group demonstrated superior elbow mobility (elbow flexion, extension, forearm pronation, and supination) and functional outcomes (MEPS scores and SF-36 scores) compared with the control group, with treatment effects becoming more pronounced over the postoperative period, particularly at 12 weeks postoperation. These findings suggest that this novel approach can facilitate better and more sustained recovery for patients with posttraumatic elbow stiffness. The method likely helps patients comprehend the rehabilitation process and adhere to the treatment plan, resulting in long-term improvements in elbow function and overall health-related quality of life.

The efficacy of this approach is attributed to immersive comics that reduce cognitive load through a 4-stage information chunking process, and interactive games, which foster adherence to daily tasks. Additionally, WeChat's real-time feedback minimizes exercise errors, while peer support enhances social support by fostering emotional camaraderie and knowledge sharing among patients. These findings underscore the dual benefits of the approach in improving knowledge retention and sustaining behavioral compliance. It has been shown to be more effective than single-modality interventions.

Comparison With Prior Work

Interest is often considered the best teacher in the pursuit of learning, and learning serves as a bridge linking us to knowledge and compliance [43]. The form in which messages are delivered significantly influences knowledge absorption [44]. Engaging formats enhance interest and capture readers' attention, thereby improving learning efficacy [45]. Past research has used various strategies to increase readers' attention, such as celebrity endorsements, humor, videos, and emoticons [44-46]. Moreover, an article's format, content, length, and communication style significantly affect readers' attention [47]. The use of caricatures has been shown to be a key feature in attracting the greatest reader engagement, and posts with images tend to have higher rates of liking and sharing [43]. This study developed a comic book. As previously described, the elbow ROM and MEPS and SF-36 scores in the WeChat group were significantly better than those in the control group. Its effectiveness is rooted in cognitive

XSL•FO
RenderX

load theory, which posits that the human brain has a limited information-processing capacity at any given time, categorized into intrinsic cognitive load (task/person-specific information complexity), extraneous cognitive load (suboptimal presentation conditions), and germane cognitive load (working memory resources allocated to process intrinsic load) [48]. The comic book's advantages align closely with this framework. First, the comic systematically partitions complex rehabilitation protocols into 4 graduated modules, reflecting Sweller's simple-to-complex instructional paradigm [48]. This phased decomposition mitigates intrinsic cognitive load by translating daunting tasks (eg, 300 groups of motions that need to be mastered in the early postoperative stage) into sequentially manageable subtasks. For example, stepwise motion schematics visually deconstruct exercise routines, reducing the cognitive effort required to interpret textual instructions. It also significantly reduces patients' psychological aversion to excessive task volume, thereby alleviating intrinsic cognitive load [47]. Second, comic illustrations serve as a perceptual bridge for abstract medical concepts. Dynamic visuals and narrative threads enhance retention by fostering immersive learning environments [49-51] consistent with Mayer's signaling principle [48,52]. Color-coded anatomical annotations and symbolic icons (eg, distinct shades denoting elbow components) adhere to the redundancy principle [48], ensuring visual-textual synergy while avoiding information duplication. Visual cues, such as arrows highlighting hinged external fixator adjustments, further optimize information hierarchy, minimizing extraneous load associated with identifying critical details [49]. These design elements enhance aesthetic appeal and function as memory anchors, corroborated by research on visual mnemonic efficacy [49]. Third, the comic's narrative structure cultivates patient engagement through relatable, character-driven scenarios, contrasting with the abstract nature of conventional educational materials [50,51,53]. By embedding rehabilitation guidance within plausible story arcs, such as a protagonist's recovery, the comic amplifies germane load through personal relevance, a mechanism supported by narrative transportation theory [53]. This emotional resonance deepens information encoding and may improve treatment adherence by transforming instructional content into personally meaningful narratives [48].

High participation is a fundamental prerequisite for effective and rapid knowledge acquisition [54]. Self-determination theory, based on 3 innate and universal psychological needs, competence, autonomy, and relatedness, explains this phenomenon [55].

When individuals' psychological needs are met, they engage in learning more actively [55]. Optimizing the learning environment helps satisfy these needs [55]. Gamification can fulfill these needs by providing an engaging environment that effectively drives participants' motivation [52]. Elements such as goal-setting, recording or affirmation of performance, and rewards for achievement influence learning efficacy [55]. In this study, interactive games were developed for patients with OEA. While elbow ROM, MEPSs, and SF-36 scores showed no significant intergroup differences 1 week postoperation, the gamified design in the WeChat group effectively addressed patients' early psychological needs, establishing a solid foundation for long-term therapeutic benefits. Specifically, games such as the "Picture-Based Cloze Test" were implemented to reinforce critical rehabilitation concepts. Through repeated practice sessions, patients engaged with challenging content until mastery was achieved, at which point a congratulatory message ("Well Done!") would appear on the PowerPoint screen. This iterative approach, designed to balance difficulty with achievable goals, aimed to foster intrinsic motivation by leveraging the psychological need for competence—the satisfaction of which, as posited by self-determination theory, drives self-determined behavior [55]. By embedding enjoyable elements into knowledge acquisition, the games promoted deeper engagement with rehabilitation materials, consistent with the self-determination theory's emphasis on intrinsic motivation as a primary driver of sustained learning [55]. Crucially, the design prioritized autonomy support by allowing patients to progress at their own pace, thereby satisfying the theoretical need for autonomy. Additionally, as patients received encouragement and praise from the interactive games, autonomy-supportive extrinsic motivation further enhanced their sense of autonomy, leading to a deeper internalization of the rehabilitation process and making it more meaningful and self-driven [55]. Finally, peer-assisted interactions (eg, mutual encouragement) within the games satisfied the need for relatedness, improving both intrinsic and extrinsic motivation for long-term adherence to rehabilitation protocols [55]. The interactive nature of the games allows patients to receive immediate feedback, which is crucial for learning and retaining rehabilitation exercises. This feedback loop reinforces correct techniques and provides a sense of accomplishment, further motivating patients to continue their rehabilitation efforts. By addressing these psychological needs, interactive games make rehabilitation more enjoyable and effective. This multifaceted approach contrasts starkly with traditional health education, which often falters due to its one-size-fits-all delivery. Postoperative patients, already stressed by surgery, may resist static written materials or passive video tutorials that offer no opportunity for agency [45]. Moreover, traditional methods rely heavily on nursing staff's communication skills, introducing variability in information retention [27]. The absence of immediate feedback and adaptive pathways further limits patients' ability to build competence or sustain engagement.

Behavioral oversight plays a crucial role in health management, is equally pivotal in elbow rehabilitation, and is firmly grounded in social support theory [56]. This theory poses that social support involves exchanging resources between individuals to improve the recipient's health, functioning, and well-being [57]. It can be divided into instrumental, emotional, informative, and other forms of support. When individuals are supported by their social environment, such as health care professionals, family members, or even peers, their behavior can be strengthened [57]. In rehabilitation, behavioral oversight ensures that patients follow treatment plans, provides encouragement, and identifies challenges for timely adjustments. Such oversight fosters a structured, supportive environment that enhances adherence and recovery. Recently, WeChat has emerged as a pivotal tool in health education and disease management, garnering widespread acceptance [19-23,25-32]. Enhancing health

XSL•FO
RenderX

information dissemination and assimilation through WeChat has become a focal point worthy of further research [43,44]. This study highlights key aspects of behavioral oversight via WeChat. At 6 and 12 weeks postoperatively, the WeChat group showed significantly better elbow ROM, MEPS, and SF-36 scores, along with a lower complication rate. This can be explained as follows: First, WeChat creates a structured professional support system that provides instrumental patient support [57]. It enables ongoing communication between researchers and patients after discharge, facilitating real-time monitoring and feedback, thereby ensuring seamless educational support. Procedures such as "daily check-in reminders" prompt patients to maintain their rehabilitation routines. By setting daily goals and incorporating knowledge consolidation, patients are more likely to adhere to their rehabilitation plans and achieve better outcomes. The "Error Correction" and "Daily Q&A" features serve as educational tools and behavioral assessments. Researchers can use the results to identify areas where patients may struggle and provide targeted support and guidance. This closed loop of "monitoring-feedback-intervention" aligns with the theory's emphasis on organized support to enhance self-regulation, ensuring patients receive continuous, tailored advice to maintain adherence [57]. Second, WeChat's social features foster peer emotional support networks, a core component of emotional support in the theory [57]. Patients who share rehabilitation videos or group pain management experiences engage in reciprocal emotional validation. For example, a patient's demonstration of achieving 100° elbow flexion at 8 weeks postoperation often triggers observational learning and mimicry among peers, enhancing collective self-efficacy through vicarious experience and motivating sustained adherence to treatment plans. Third, WeChat's data-tracking capabilities enable dynamic support strategies rooted in informational support [57]. Using WeChat for behavioral oversight also provides valuable data for researchers to analyze patient behavior patterns and treatment outcomes, yielding insights that can further refine rehabilitation strategies and improve patient care.

Strengths and Limitations

This study has several strengths. First, it introduces a novel and innovative health education approach to elbow rehabilitation by incorporating comics, integrative games, and WeChat. The approach enhances patient engagement and comprehension, ensuring better retention and understanding of rehabilitation procedures. Second, the study emphasizes the importance of behavioral oversight in health management, which is often overlooked in traditional health care settings. By leveraging the capabilities of the WeChat platform, it provides a feasible solution for continuous, personalized monitoring of patient behavior postdischarge.

Importantly, this study has certain limitations. First, the patient follow-up period was limited to 12 weeks, which does not provide insight into long-term outcomes. Although this relatively short time frame restricts our understanding of sustained changes, the data collected during these 12 weeks can offer valuable early indications. Several considerations can help mitigate the potential impact of this limitation on the interpretation of our results. First, although the short follow-up restricts our understanding of long-term changes, the data collected during these 12 weeks can provide valuable early indications. These findings can inform future investigations with extended follow-up periods to validate and extend upon our findings. Additionally, despite the absence of long-term data, the observed short-term effects remain useful for guiding clinical practice. Clinicians can use these early outcomes to make informed decisions regarding patient care while remaining attentive to potential differences in long-term results. Second, the study was conducted at a single center, which may limit the generalizability of the findings. The unique characteristics of the patient population and the health care system at this center could influence the results and their applicability to other settings. Another limitation is the potential for self-reporting bias in the behavioral monitoring component. Patients may not always accurately report their behavior, which could affect the reliability of the data.

Future Directions

For future research, it would be beneficial to extend the follow-up period to assess the long-term efficacy of the rehabilitation program. This would provide a more comprehensive understanding of the sustained impact of the intervention on patient outcomes. Additionally, future studies should replicate this study across multiple centers with diverse patient demographics and health care systems. Such efforts would help determine whether the observed benefits are consistent across different environments and populations. Also, this study's approach could be applied to other types of rehabilitation or different patient populations to assess its generalizability. Furthermore, to minimize self-reporting bias, future studies could incorporate objective measures, such as wearable devices or other technologies, to track patient activity levels and adherence to rehabilitation protocols more accurately. Finally, investigating the cost-effectiveness of using comics, integrated games, and WeChat in health education would provide valuable insights for health care providers considering the implementation of similar programs.

Conclusion

This study demonstrates that WeChat-based plus scene-graphical health education effectively communicates information to patients who have undergone OEA. This approach has been shown to enhance elbow ROM, functionality, and overall quality of life. Furthermore, it offers a platform for patients to address challenges and support one another.

Acknowledgments

This study was funded by the Municipal Health Commission of Shanghai, China (grant 202040399) and the Shanghai Sixth People's Hospital Medical Group (grant X-4587 to LG). The authors thank Cunyi Fan, MD, PhD, for providing invaluable

XSL•FO
RenderX

suggestions for developing the WeChat platform. They also thank Changtai Zhu, MD, PhD, for his significant contributions to patient recruitment. The authors also appreciate all the patients who contributed to this research. Finally, the authors thank the reviewers for their valuable feedback.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to the inclusion of sensitive participant information (eg, medical records and personal identifiers), which are protected by ethical guidelines and privacy regulations. However, they are available from the corresponding author upon reasonable request.

Authors' Contributions

Danling F conducted the study and drafted the manuscript. WY conceived the study. WW was responsible for project administration. SY and PL were instrumental in data acquisition and analysis and provided technical assistance. PL conducted the clinical study and offered critical revisions to the manuscript. The authors declare that they have no conflicts of interest. All the authors reviewed and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Protocol summary and additional information.
[DOCX File, 746 KB - formative_v9i1e58218_app1.docx]

Multimedia Appendix 2
Comic materials for the study groups.
[DOCX File, 4064 KB - formative_v9i1e58218_app2.docx]

Checklist 1
CONSORT (Consolidated Standards of Reporting Trials) checklist.
[PDF File, 420 KB - formative_v9i1e58218_app3.pdf]

References

1. Liu W, Sun Z, Xiong H, et al. Development and validation of a prognostic nomogram for open elbow arthrolysis: the Shanghai prediction model for elbow stiffness surgical outcome. Bone Joint J 2022 Apr;104-B(4):486-494. [doi: 10.1302/0301-620X.104B4.BJJ-2021-1326.R2] [Medline: 35360939]
2. Zheng W, Liu J, Song J, Fan C. Risk factors for development of severe post-traumatic elbow stiffness. Int Orthop 2018 Mar;42(3):595-600. [doi: 10.1007/s00264-017-3657-1] [Medline: 28988397]
3. Sun Z, Liu W, Li J, Fan C. Open elbow arthrolysis for post-traumatic elbow stiffness: an update. Bone Jt Open 2020 Sep;1(8):576-584. [doi: 10.1302/2633-1462.19.BJO-2020-0098.R1] [Medline: 33215148]
4. Siemensma MF, van der Windt AE, van Es EM, Colaris JW, Eygendaal D. Management of the stiff elbow: a literature review. EFORT Open Rev 2023 May 9;8(5):351-360. [doi: 10.1530/EOR-23-0039] [Medline: 37158372]
5. Zhang D, Nazarian A, Rodriguez EK. Post-traumatic elbow stiffness: pathogenesis and current treatments. Shoulder Elbow 2020 Feb;12(1):38-45. [doi: 10.1177/1758573218793903] [Medline: 32010232]
6. Sun Z, Cui H, Liang J, Li J, Wang X, Fan C. Determining the effective timing of an open arthrolysis for post-traumatic elbow stiffness: a retrospective cohort study. BMC Musculoskelet Disord 2019 Mar 25;20(1):122. [doi: 10.1186/s12891-019-2506-3] [Medline: 30909899]
7. Cui H, Sun Z, Ruan J, Yu Y, Fan C. Effect of enhanced recovery after surgery (ERAS) pathway on the postoperative outcomes of elbow arthrolysis: a randomized controlled trial. Int J Surg 2019 Aug;68:78-84. [doi: 10.1016/j.ijsu.2019.06.010] [Medline: 31220630]
8. Masci G, Cazzato G, Milano G, et al. The stiff elbow: current concepts. Orthop Rev (Pavia) 2020 Jun 29;12(Suppl 1):8661. [doi: 10.4081/or.2020.8661] [Medline: 32913596]
9. Vasileiadis GI, Ramazanian T, Kamaci S, et al. Loss of pronation-supination in patients with heterotopic ossification around the elbow. J Shoulder Elbow Surg 2019 Jul;28(7):1406-1410. [doi: 10.1016/j.jse.2018.10.029] [Medline: 30685280]
10. Ruan JH, Cui HM, Sun ZY, Chen S, Wang W, Fan CY. Midterm outcomes after open arthrolysis for posttraumatic elbow stiffness in children and adolescents. J Pediatr Orthop 2021 Mar 1;41(3):e266-e271. [doi: 10.1097/BPO.0000000000001748] [Medline: 33492039]

XSL•FO
RenderX

11. Guglielmetti CLB, Gracitelli MEC, Assunção JH, et al. Randomized trial for the treatment of post-traumatic elbow stiffness: surgical release vs. rehabilitation. J Shoulder Elbow Surg 2020 Aug;29(8):1522-1529. [doi: 10.1016/j.jse.2020.03.023] [Medline: 32713463]
12. Kwak JM, Sun Y, Kholinne E, Koh KH, Jeon IH. Surgical outcomes for post-traumatic stiffness after elbow fracture: comparison between open and arthroscopic procedures for intra- and extra-articular elbow fractures. J Shoulder Elbow Surg 2019 Oct;28(10):1998-2006. [doi: 10.1016/j.jse.2019.06.008] [Medline: 31540725]
13. Sun Z. 肘关节僵硬新分型和功能评分的建立及远期松解疗效的探究 [Establishment of a new classification and functional score for elbow stiffness and exploration of the long - term efficacy of release]. : Shanghai Jiao Tong University; 2020 URL: http://dbase2.gslib.com.cn/KCMS/detail/search.aspx?dbcode=CJFD&sfield=kw&skey=%E8%BF%9C%E6%9C%9F%E7%96%97%E6%95%88 [accessed 2025-10-05]
14. Porcellini G, Rotini R, Kantar SS, Giacomo SD. The Elbow: Principles of Surgical Treatment and Rehabilitation: Springer International; 2018.
15. Haglin JM, Kugelman DN, Christiano A, Konda SR, Paksima N, Egol KA. Open surgical elbow contracture release after trauma: results and recommendations. J Shoulder Elbow Surg 2018 Mar;27(3):418-426. [doi: 10.1016/j.jse.2017.10.023]
16. Debono B, Wainwright TW, Wang MY, et al. Consensus statement for perioperative care in lumbar spinal fusion: Enhanced Recovery After Surgery (ERAS®) Society recommendations. Spine J 2021 May;21(5):729-752. [doi: 10.1016/j.spinee.2021.01.001] [Medline: 33444664]
17. Gustafsson UO, Scott MJ, Hubner M, et al. Guidelines for perioperative care in elective colorectal surgery: Enhanced Recovery After Surgery (ERAS®) Society recommendations: 2018. World J Surg 2019 Mar;43(3):659-695. [doi: 10.1007/s00268-018-4844-y] [Medline: 30426190]
18. Absolom K, Warrington L, Hudson E, et al. Phase III randomized controlled trial of eRAPID: eHealth intervention during chemotherapy. J Clin Oncol 2021 Mar 1;39(7):734-747. [doi: 10.1200/JCO.20.02015] [Medline: 33417506]
19. Xu L, Li P, Hou X, et al. Middle-aged and elderly users' continuous usage intention of health maintenance-oriented WeChat official accounts: empirical study based on a hybrid model in China. BMC Med Inform Decis Mak 2021 Sep 3;21(1):257. [doi: 10.1186/s12911-021-01625-4] [Medline: 34479566]
20. Wang J, Wu L. A comparison of health communication effectiveness and the improvement of management strategies: taking two Chinese traditional medicine hospitals' WeChat public accounts as examples. BMC Health Serv Res 2020 Nov 20;20(1):1055. [doi: 10.1186/s12913-020-05901-3] [Medline: 33213422]
21. Song J, Ma L. Effect of cognitive behavioral therapy and WeChat-based health education on patients underwent peripherally inserted central catheter line placement. Am J Transl Res 2021;13(12):13768-13775. [Medline: 35035715]
22. Sun M, Yang L, Chen W, et al. Current status of official WeChat accounts for public health education. J Public Health (Oxf) 2021 Sep 22;43(3):618-624. [doi: 10.1093/pubmed/fdz163]
23. Li Y, Xiao QL, Li M, et al. Community-based intervention via WeChat official account to improve parental health literacy among primary caregivers of children aged 0 to 3 years: protocol for a cluster randomized controlled trial. Front Public Health 2023;10(10). [doi: 10.3389/fpubh.2022.1039394]
24. Tencent. Tencent announces 2020 fourth quarter and annual results. 2021. URL: https://www.tencent.com/en-us/investors/financial-news.html [accessed 2021-10-10]
25. Ma X, Lu J, Liu W. Influencing factors on health information to improve public health literacy in the official WeChat account of Guangzhou CDC. Front Public Health 2021;9(9):657082. [doi: 10.3389/fpubh.2021.657082] [Medline: 34414152]
26. Zhang X, Xiao H, Chen Y. Evaluation of a WeChat‐based life review programme for cancer patients: a quasi‐experimental study. J Adv Nurs 2019 Jul;75(7):1563-1574. [doi: 10.1111/jan.14018]
27. Zhou K, Wang W, Zhao W, et al. Benefits of a WeChat-based multimodal nursing program on early rehabilitation in postoperative women with breast cancer: a clinical randomized controlled trial. Int J Nurs Stud 2020 Jun;106:103565. [doi: 10.1016/j.ijnurstu.2020.103565] [Medline: 32272281]
28. Cao Y, Lin SH, Zhu D, et al. WeChat public account use improves clinical control of cough-variant asthma: a randomized controlled trial. Med Sci Monit 2018 Mar 14;24:1524-1532. [doi: 10.12659/msm.907284] [Medline: 29536984]
29. Jiang Y, Liu F, Guo J, et al. Evaluating an intervention program using WeChat for patients with chronic obstructive pulmonary disease: randomized controlled trial. J Med Internet Res 2020 Apr 21;22(4):e17089. [doi: 10.2196/17089] [Medline: 32314971]
30. Li T, Ding W, Li X, Lin A. Mobile health technology (WeChat) for the hierarchical management of community hypertension: protocol for a cluster randomized controlled trial. Patient Prefer Adherence 2019;13:1339-1352. [doi: 10.2147/PPA.S215719] [Medline: 31496664]
31. Li Q, Guo L, Zhang J, et al. Effect of school-based family health education via social media on children's myopia and parents' awareness: a randomized clinical trial. JAMA Ophthalmol 2021 Nov 1;139(11):1165-1172. [doi: 10.1001/jamaophthalmol.2021.3695] [Medline: 34529026]
32. Ma C, Wang B, Zhao X, et al. WeChat-based education and rehabilitation program in unprotected left main coronary artery disease patients after coronary artery bypass grafting: an effective approach in reducing anxiety, depression, loss to follow-up, and improving quality of life. Braz J Med Biol Res 2021;54(4):e10370. [doi: 10.1590/1414-431X202010370] [Medline: 33624731]

XSL•FO
RenderX

33. Isaac M. Role of humanities in modern medical education. Curr Opin Psychiatry 2023 Sep 1;36(5):347-351. [doi: 10.1097/YCO.0000000000000884] [Medline: 37458498]
34. Chernick LS, Konja A, Gonzalez A, et al. Designing illustrative social media stories to promote adolescent peer support and healthy sexual behaviors. Digit Health 2022;8:20552076221104660. [doi: 10.1177/20552076221104660] [Medline: 35707267]
35. Hall CE, Hall AB, Mallya J, Courtright P, Kok G. Developing comic strips promoting diabetic retinopathy screening in Kilimanjaro, Tanzania, using intervention mapping. Eye (Lond) 2022 May;36(Suppl 1):25-32. [doi: 10.1038/s41433-022-02003-8] [Medline: 35590052]
36. Haan MM, van Gurp JL, Knippenberg M, Olthuis G. Facilitators and barriers in using comics to support family caregivers of patients receiving palliative care at home: a qualitative study. Palliat Med 2022 Jun;36(6):994-1005. [doi: 10.1177/02692163221093513] [Medline: 35502800]
37. De Stefano A, Rusciano I, Moretti V, et al. Graphic medicine meets human anatomy: the potential role of comics in raising whole body donation awareness in Italy and beyond. A pilot study. Anat Sci Educ 2023 Mar;16(2):209-223. [doi: 10.1002/ase.2232] [Medline: 36346170]
38. Wang Y, Lv X. 日常生活自理能力情景图示评定方法的设计[Design of pictorial assessment method for self-care ability in daily life]. Chin J Phys Med Rehabilit 2018;40(11):4.
39. Wang Y, Ni J, Sun Z. 微信平台在开放性肘关节僵硬松解术后康复锻炼中的应用[Application of WeChat platform in rehabilitation exercise after open elbow joint stiffness release]. Int J Orthopaed 2018;39(6):394-396.
40. Sun Z, Cui H, Ruan J, Li J, Wang W, Fan C. What range of motion and functional results can be expected after open arthrolysis with hinged external fixation for severe posttraumatic elbow stiffness? Clin Orthop Relat Res 2019 Oct;477(10):2319-2328. [doi: 10.1097/CORR.0000000000000726] [Medline: 31107330]
41. Sun Z, Li J, Luo G, Wang F, Hu Y, Fan C. What constitutes a clinically important change in Mayo Elbow Performance Index and range of movement after open elbow arthrolysis? Bone Joint J 2021 Feb;103-B(2):366-372. [doi: 10.1302/0301-620X.103B2.BJJ-2020-0259.R3] [Medline: 33517717]
42. He X, Fen Q, Yang J, Lei Y, Heng L, Zhang K. Risk factors of elbow stiffness after open reduction and internal fixation of the terrible triad of the elbow joint. Orthop Surg 2021 Apr;13(2):530-536. [doi: 10.1111/os.12879] [Medline: 33619861]
43. Blok AC, Valley TS, Abbott P. Gamification for family engagement in lifestyle interventions: a systematic review. Prev Sci 2021 Oct;22(7):831-844. [doi: 10.1007/s11121-021-01214-x] [Medline: 33786746]
44. Zhang Y, Xia T, Huang L, et al. Factors influencing user engagement of health information disseminated by Chinese provincial Centers for Disease Control and Prevention on WeChat: observational study. JMIR Mhealth Uhealth ;7(6):e12245. [doi: 10.2196/12245]
45. Bao Y, Wang C, Xu H, et al. Effects of an mHealth intervention for pulmonary tuberculosis self-management based on the integrated theory of health behavior change: randomized controlled trial. JMIR Public Health Surveill 2022 Jul 14;8(7):e34277. [doi: 10.2196/34277] [Medline: 35834302]
46. Cheng C, Espanha R. Social support and audience engagement of lupus-related posts on social networking sites in China. Heliyon 2024 Jun 15;10(11):e31754. [doi: 10.1016/j.heliyon.2024.e31754] [Medline: 38841438]
47. España K, Perris GE, Ngo NT, Bath E. Reimagining narrative approaches through comics for systems-involved youth. J Am Acad Child Adolesc Psychiatry 2024 Aug;63(8):766-770. [doi: 10.1016/j.jaac.2023.11.006] [Medline: 38000536]
48. Sweller J. Cognitive load during problem solving: effects on learning. Cogn Sci 1988 Apr;12(2):257-285. [doi: 10.1207/s15516709cog1202_4]
49. Alemany-Pagès M, Azul AM, Ramalho-Santos J. The use of comics to promote health awareness: a template using nonalcoholic fatty liver disease. Eur J Clin Invest 2022 Mar;52(3):e13642. [doi: 10.1111/eci.13642] [Medline: 34159578]
50. Cohn N. Your brain on comics: a cognitive model of visual narrative comprehension. Top Cogn Sci 2020 Jan;12(1):352-386. [doi: 10.1111/tops.12421] [Medline: 30963724]
51. Chen M, Bell RA. A meta-analysis of the impact of point of view on narrative processing and persuasion in health messaging. Psychol Health 2022 May;37(5):545-562. [doi: 10.1080/08870446.2021.1894331] [Medline: 33678078]
52. Harp SF, Mayer RE. How seductive details do their damage: a theory of cognitive interest in science learning. J Educ Psychol 1998;90(3):414-434.
53. Wang Z, Romat H, Chevalier F, Riche NH, Murray-Rust D, Bach B. Interactive data comics. IEEE Trans Vis Comput Graph 2022 Jan;28(1):944-954. [doi: 10.1109/TVCG.2021.3114849] [Medline: 34587073]
54. Walker A, Hing W, Lorimer A. The influence, barriers to and facilitators of anterior cruciate ligament rehabilitation adherence and participation: a scoping review. Sports Med Open 2020 Jul 17;6(1):32. [doi: 10.1186/s40798-020-00258-7] [Medline: 32681200]
55. Ntoumanis N, Ng JYY, Prestwich A, et al. A meta-analysis of self-determination theory-informed intervention studies in the health domain: effects on motivation, health behavior, physical, and psychological health. Health Psychol Rev 2021 Jun;15(2):214-244. [doi: 10.1080/17437199.2020.1718529] [Medline: 31983293]
56. Bian D, Shi Y, Tang W, et al. The influencing factors of nutrition and diet health knowledge dissemination using the WeChat official account in health promotion. Front Public Health 2021;9(9):775729. [doi: 10.3389/fpubh.2021.775729] [Medline: 34900917]

XSL•FO
RenderX

57. He S, Jiang S, Zhu R, Hu X. The influence of educational and emotional support on e-learning acceptance: an integration of social support theory and TAM. Educ Inf Technol 2023 Sep;28(9):11145-11165. [doi: 10.1007/s10639-023-11648-1]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
HbA_{1c}: glycated hemoglobin
MEPS: Mayo Elbow Performance Score
OEA: open elbow arthrolysis
ROM: range of motion
SF-36: 36-item Short Form Health Survey

Edited by A Mavragani; submitted 17.03.24; peer-reviewed by J Mistry, L Fan, Y Qian, Y Jiang; revised version received 11.06.25; accepted 12.06.25; published 21.10.25.

Please cite as:
Fang D, Wang Y, Yu S, Wang W, Lu P
Effectiveness of WeChat-Based Plus Scene-Graphics Health Education for Rehabilitation After Open Elbow Arthrolysis: Historical Control Study
JMIR Form Res 2025;9:e58218
URL: https://formative.jmir.org/2025/1/e58218
doi:10.2196/58218

© Danling Fang, Yin Wang, Shiyang Yu, Wei Wang, Pingchen Lu. Originally published in JMIR Formative Research (https://formative.jmir.org), 21.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Assessing the Readability and Quality of Web-Based Resources on Exercise Stress Testing: Cross-Sectional Readability and Quality Analysis

Munir Rahbe[1], MD; Dhrumi Mistry[1], MD; Ali M Khawaja[2], BA; Natalie A Sous[1], MD; Alan Y Tso[1], MD, MA

[1]Department of Medicine, Rutgers New Jersey Medical School, 185 S Orange Ave, Newark, NJ, United States

[2]Rutgers New Jersey Medical School, Newark, NJ, United States

Corresponding Author:
Munir Rahbe, MD
Department of Medicine, Rutgers New Jersey Medical School, 185 S Orange Ave, Newark, NJ, United States

Abstract

Background: As internet usage continues to rise, an increasing number of individuals rely on online resources for health-related information. However, prior research has shown that much of this information is written at a reading level exceeding national recommendation, which may hinder patient comprehension and decision-making. The American Medical Association (AMA) recommends that patient-directed health materials be written at or below a 6th-grade reading level to ensure accessibility and promote health literacy. Despite these guidelines, studies indicate that many online health resources fail to meet this standard. The exercise stress test is a widely used diagnostic tool in cardiovascular medicine, yet no prior studies have assessed the readability and quality of online health information specific to this topic.

Objective: This study aimed to evaluate the readability and quality of online resources on exercise stress testing and compare these metrics between academic and non-academic sources.

Methods: A cross-sectional readability and quality analysis was conducted using Google and Bing to identify web-based patient resources related to exercise stress testing. Eighteen relevant websites were categorized as academic (n=7) or nonacademic (n=11). Readability was assessed using four established readability formulas: Flesch-Kincaid Grade Level (FKGL), Flesch Reading Ease (FRE), Simple Measure of Gobbledygook (SMOG), and Gunning Fog (GF). Website quality and reliability were evaluated using the modified DISCERN (mDISCERN) tool. Statistical comparisons between academic and nonacademic sources were performed using independent samples *t* tests.

Results: The average FKGL, SMOG, and GF scores for all websites were 8.36, 8.28, and 10.14, respectively, exceeding the AMA-recommended 6th-grade reading level. Academic sources had significantly higher FKGL (9.1 vs. 7.9, P=.03), SMOG (8.9 vs. 7.9, P=.04), and lower FRE scores (57.6 vs. 65.3, P=.006) than nonacademic sources, indicating greater reading difficulty. The average GF scores for academic and nonacademic sources were 10.68 and 9.81, respectively, but this difference was not statistically significant. The quality of web resources, as assessed by mDISCERN, was classified as fair overall, with an average score of 29.44 out of 40 (74%). While academic and nonacademic websites had similar mDISCERN scores, areas such as source citation, publication dates, and acknowledgment of uncertainty were consistently lacking across all resources.

Conclusions: Online resources on exercise stress testing are, on average, written at a reading level that exceeds the AMA's 6th-grade reading guideline, potentially limiting patient comprehension. Academic sources are significantly more difficult to read than nonacademic sources, though neither category meets the recommended readability standards. The quality of web-based resources was found to be fair but could be improved by ensuring transparency in sourcing and providing clearer, more comprehensive information. These findings underscore the need for improved accessibility and readability in online health information to support patient education and informed decision-making.

(JMIR Form Res 2025;9:e68000) doi:10.2196/68000

KEYWORDS

exercise stress test; readability; quality; Flesch-Kincaid grade level; Flesch reading ease; Simple Measure of Gobbledygook; Gunning Fog index; FKGL; FRE; SMOG; GF; DISCERN; mDISCERN; exercise; physical activity; cross-sectional study; quality analysis; health literacy; patient outcomes; search engines; websites; medical information; health information

XSL•FO
RenderX

Introduction

As of 2023, nearly 95% of adults in the United States use the internet [1]. In today's information age, the challenge has shifted from simply accessing the internet to finding information that is both reliable and accessible. This shift is especially significant as more individuals rely on the internet for crucial health information. In 2022, 58.5% of adults reported using the internet at least once in the past year [2], and 70% turned to the internet as their primary source for health-related inquiries [3]. However, medical information is often nuanced, complex, and difficult to comprehend. What happens when patients encounter content that is difficult to read? Research shows that low health literacy is a significant independent predictor of all-cause mortality [4,5]. This emphasizes the vital role of the internet in modern health care and highlights the need for accessible, understandable health information to improve patient outcomes.

As early as 2000, professionals began publishing articles urging physicians to guide their patients using objective scales to "grade" the quality of online health information. The suggestion was that physicians should take responsibility for advising patients on how to "surf the net safely" when seeking health-related information [6]. Since then, the readability and quality of online health content accessed by patients have remained a key concern for clinicians, leading to numerous studies on the topic [7-10]. These studies examine various factors, but one of the most consistent is readability—specifically, the grade level at which the material is written. According to the American Medical Association (AMA), health materials should be written at or below a 6th-grade reading level to ensure appropriate health literacy for today's patients [11].

The exercise stress test is a widely used diagnostic tool in the outpatient setting for individuals with potential coronary artery disease. It is also a valuable tool in the risk stratification of patients undergoing surgical procedures. The results of these tests have good prognostic value and essential information for clinicians [12]. Numerous studies evaluating the readability and quality of cardiovascular health topics have found that much of the available information is written at a reading level that exceeds recommended standards [13-15]. However, no studies to date have explicitly focused on exercise stress testing. This study will be the first to assess the readability and quality of online resources related to exercise stress testing.

The goals of this study are to (1) evaluate the readability and quality of health information on exercise stress tests and (2) compare these metrics between academic and nonacademic online resources.

Methods

Data Collection

To objectively assess the quality of available information for patients regarding cardiovascular exercise stress testing, the Google and Bing search engines were accessed. These search engines are ranked the top two respective search engines globally, accounting for approximately 94% of all search engine traffic [16]. The engines were interrogated using the four terms: "exercise stress test," "exercise ECG," "stress ECG," and "treadmill ECG" on August 28, 2024. Searches were conducted on a nonuniversity affiliated internet connection in the respective engine's incognito mode to simulate the experience of a default user. Analysis of relevant sites was limited to the first page of produced results in accordance with previous findings indicating that approximately 92% of all search engine traffic fails to extend beyond the first page [17].

As of August 2024, the Google search engine accounted for approximately 90.5% of global search engine traffic [16]. Additionally, we sought to include the Bing search engine, accounting for an additional 3.9% of search engine traffic, to assess a broader scope of patient experiences. Searches were conducted, and relevant sites were compiled; results were excluded if they were published in academic journals, published in academic textbooks, were written in a non-English language, were redundant with previous search results, were outside of the scope of the cardiac stress testing assessment, redirected to a nonarticle link, or promoted the sale of a product or service. These criteria were established to analyze the most relevant information a patient may encounter in researching this assessment that they would likely evaluate. The final analysis of results included eighteen web pages. The compiled results were categorized as academic (n=7) or nonacademic (n=11) and assessed for readability and quality. Academic results were operationally defined as sites affiliated with an academic institution, whereas nonacademic results were defined as publicly available websites containing relevant information not affiliated with an academic institution.

Assessment of Readability and Quality

The Flesch-Kincaid Grade Level (FKGL), Flesch Reading Ease (FRE), Simple Measure of Gobbledygook (SMOG), and Gunning Fog (GF) tools were used to objectively assess the readability of internet-based resources pertaining to cardiovascular exercise stress testing. A modified DISCERN criterion (mDISCERN) was used to characterize the objective quality and reliability of the internet-based resource.

The Flesch-Kincaid Reading Ease (FRE) and Flesch-Kincaid Grade Level (FKGL) are the most widely used readability tools in health care literature [18]. These tools were originally developed in the mid-1900s for the United States Navy and were found to correlate well with the reading levels of sailors [19]. Since their inception, they have been adopted by various government organizations, including the Social Security Administration [20]. While these readability tools are not perfect, they are considered reliable when compared to other psychological tests used to assess reading level [21]. The FRE and FKGL indexes have been used extensively to objectively quantify the readability of health-related materials and are widely used in the literature for assessing the accessibility of health care information [8,9]. The FRE and FKGL metrics use mathematical formulas to quantify an objective readability score of written material based on the number of words per sentence and syllables per word [22,23]. FRE scores range from 0 - 100, with greater scores indicating greater ease of reading material and lower scores indicating greater difficulty. The FKGL index

XSL•FO
RenderX

uses the FRE score and extrapolates the produced value to a respective reading grade level. The FKGL scores use the same scale as the FRE index, with scores 80‐90 indicating a 6th-grade reading level and scores 0‐30 indicating a college reading level.

Despite the frequency of their use, the FRE and FKGL indices have noted variability in their accuracy with respect to particular literary resources [24]. To address any potential limitations of these tools, this study used two additional readability metrics commonly used in the health care literature—the Gunning Fog (GF) and Simple Measure of Gobbledygook (SMOG) assessments. Developed in 1952 and 1969, respectively, these tools were developed after the FRE and serve as additional commonly utilized readability metrics in evaluating online patient-directed health resources [25,26].

Although differences in the frequency of their use exist, the GF and SMOG indices have been noted to produce outputs that are more precise in evaluating the readability of written health information due to their modified criteria compared to the FRE and FKGL indexes [18]. Furthermore, the SMOG index has been indicated as the gold standard with respect to the readability of health-related resources due to its precision in scoring outputs and the recency of its development in relation to other tools [18,27]. The GF index uses a mathematical formula that evaluates the number of words, sentences, and complex (multi-syllabic) words to calculate an objective score that equates to the readability of a given text [28]. The later-developed SMOG index uses a modified mathematical model that also uses the number of complex words and sentences to generate an objective readability score. Similar to the FKGL, both metrics use the same scale, and the generated numerical score is designed to directly equate to the grade level required to comprehend the respective text.

Readability results across the FRE, FKGL, FOG, and SMOG indexes were calculated for each of the 18 compiled internet-based resources, and the respective scores were compared to the AMA guidelines, which recommend patient-targeted reading material be written at or below a 6th-grade reading level [11].

The DISCERN instrument, another widely used tool in the literature, has been validated for the evaluation of health-related information by both healthcare professionals and patients [29]. This instrument was used to objectively assess the quality of internet-based information regarding cardiac exercise stress tests [9,10]. The questionnaire consists of three sections with a total of sixteen questions, designed to evaluate both the reliability and overall quality of online health information. The first component of the instrument, questions 1 through 8, characterizes the reliability of a published resource. The second component, questions 9 through 15, characterizes the quality of treatment-related information. The third component, question 16, uses the compiled scores from the previous 15 questions to characterize how well the resource provides information regarding specific treatment options [30]. All questions are scored on a scale ranging from 1 to 5, with lower scores indicating poorer rankings and higher scores indicating more positive rankings. For the purposes of this study, a modified version` of the DISCERN instrument (mDISCERN) was used, consisting of only the first component of the tool, as this study solely concentrated on information surrounding diagnostic assessment and not subsequent treatment (Table 1). A mDISCERN score below 40% classified the reliability of a web result as “poor,” scores between 40% and 79% were deemed “fair,” and scores above 80% were categorized as “good.” These thresholds align with scoring criteria established in prior research studies [26,31].

Table . Modified DISCERN (mDISCERN) scoring tool.

#	Question	Score
1	Are the aims of the article clear?	1‐5
2	Does the article achieve stated aims?	1‐5
3	Are the topics covered relevant?	1‐5
4	Is it clear what sources of information were used to compile the publication other than the author?	1‐5
5	Is it clear when the information used or reported in the publication was produced?	1‐5
6	Is it balanced and unbiased?	1‐5
7	Does it provide details of additional sources of support and information?	1‐5
8	Does it refer to areas of uncertainty?	1‐5

Text from each of the eighteen compiled internet articles was copied into a Microsoft Word document. Image-derived text within the article was only included in the document if the text was present on a relevant graphic within the article and was clearly intended for the viewer to read. Two independent reviewers used the Microsoft grammar assist tool to calculate the FRE score and FKGL using their respective equations [22,23]. SMOG and GF scores were calculated by inputting identical website text used for FRE and FKGL scores into a widely used free online readability calculator [32]. The reviewers also used the mDISCERN criteria to evaluate each source material. Scores for each of the questions of the mDISCERN criteria were independently calculated, and scores were averaged.

XSL•FO
RenderX

Statistical Analysis

Independent samples *t* tests were performed between mDISCERN, individual mDISCERN questions, FRE, FKGL, SMOG and GF scores of academic (n=7) and nonacademic (n=11) resources to distinguish any significant differences in the quality and readability of information based on the publisher. An alpha value of *P*=.05 was used to determine statistical significance for all performed analyses. All analysis was conducted on version 29.0.2.0 of the SPSS platform.

Ethical Considerations

This study was exempt from review by the Rutgers University Office for Research Institutional Review Board as it was deemed Non-Human Subject Research (Pro2025001840). All information obtained over the course of the study is publicly available. All data, usernames, screenshots, and pictures used in this manuscript have been deidentified.

Results

Each search term was queried in two search engines, yielding a total of 80 websites. Of these, 62 were textbook chapters, duplicates, advertisement links, journal publications, or outside the scope of exercise stress testing. These results were subsequently excluded, and the remaining 18 websites were evaluated. Our screening process for the websites included in our evaluation is depicted in Figure 1. The websites were categorized into two groups: academic (n=7) and nonacademic (n=11). Results for the analyzed websites are available in Table 2. The FKGL ranged from 5.15 to 9.85; the mean score was 8.36 (SD 1.32). The mean FRE score of all the websites included in this study was 62.31 (SD 6.65). The mDISCERN scores ranged from 22 - 36, with the mean being 29.44 (SD 3.98).

Figure 1. Screening methodology and exclusion criteria for website analysis.

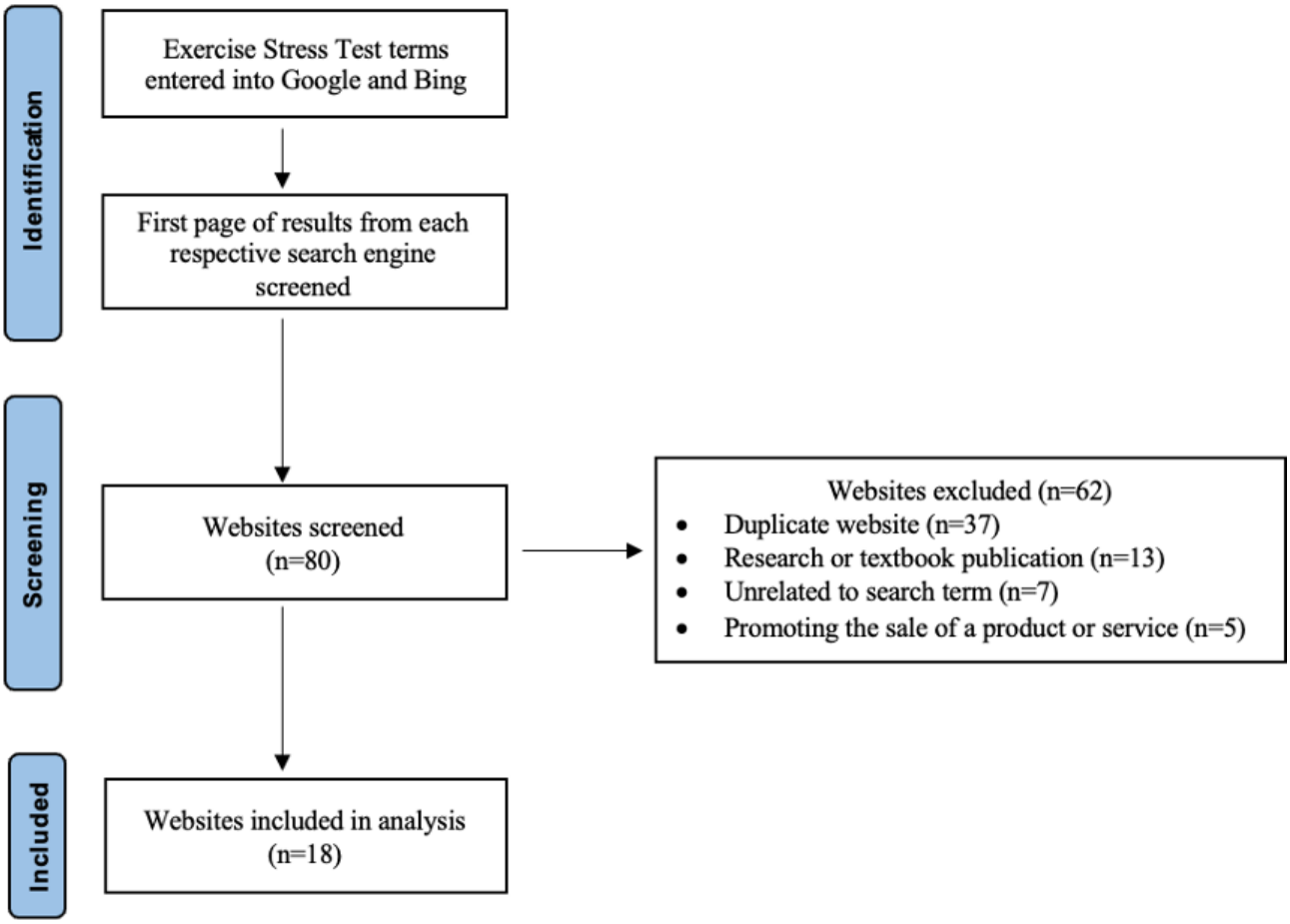

XSL•FO
RenderX

Table . Calculated metrics for academic (1-7) and nonacademic (8-18) sources.

Website[a]	FKGL[b]	SMOG[c]	GF[d]	FRE[e]	mDISCERN[f]
Cleveland Clinic [33]	8.8	8.7	11.4	53.1	36.0
Mayo Clinic [34]	7.5	8.5	8.5	61.7	36.0
Harvard Health [35]	9.0	8.9	10.7	60.0	27.5
Harvard Health [36]	9.8	9.7	11.6	56.4	27.0
Harvard Health [37]	9.2	9.3	10.8	60.5	29.5
Hopkins Medicine [38]	8.9	8.2	10.6	58.9	27.0
Massachusetts General Hospital [39]	10.4	9.0	11.1	52.5	28.0
American Heart Association [40]	8.1	8.7	11.0	62.4	22.0
St. Vincent's Heart Health [41]	8.6	8.6	11.4	61.3	26.5
Medline Plus [42]	7.8	7.3	9.3	64.2	36.0
WebMD [43]	6.7	7.8	9.7	70.1	29.5
Healthline [44]	8.4	7.3	9.0	62.7	33.5
British Heart Foundation [45]	6.7	6.8	8.6	73.3	29.5
Heart and Stroke Foundation Canada [46]	9.0	9.5	11.3	58.4	27.5
MyHealth Alberta [47]	5.2	5.3	6.4	77.7	29.5
HeartWest [48]	9.9	9.2	11.0	60.1	29.5
SoZo Cardiology [49]	9.5	9.5	12.1	57.5	28.0
Ascot Cardiology [50]	7.0	6.8	8.1	70.8	27.5

[a]Academic sources (1-7) and nonacademic (8-18) sources.

[b]FKGL: Flesch-Kincaid grade level.

[c]SMOG: Simple measure of gobbledygook.

[d]GF: Gunning Fog.

[e]FRE: Flesch reading ease.

[f]mDISCERN: modified DISCERN.

When evaluating by website type, academic websites had a greater mean FKGL (9.09, SD 0.90 vs 7.90, SD 1.40), SMOG (8.90, SD 0.50 vs 7.90, SD 1.34), and GF (10.68, SD 1.03 vs 9.81, SD 1.73) than nonacademic websites. An independent samples *t* test found the mean FKGL and SMOG between academic and nonacademic websites to be statistically significant, with a *P* value of .03 and .04, respectively. (Figure 2). Similarly, a statistically significant difference in the mean FRE for academic (57.6, SD 3.66) and nonacademic (65.32, SD 6.6) sources was found with a *P*=.006 (Figure 3). Academic (30.14, SD 4.18) and nonacademic (29, SD 3.88) sites had similar mDISCERN scores. For academic websites, mDISCERN scores ranged from 27 - 36, while nonacademic websites mDISCERN scores ranged from 22 - 36.

XSL•FO
RenderX

Figure 2. Mean FKGL, SMOG, and GF of academic and nonacademic sources. The dotted red line represents the American Medical Association's recommended 6th-grade reading level. FKGL: Flesch-Kincaid grade level; GF: Gunning Fog; SMOG: Simple Measure of Gobbledygook. *P=.03 **P=.04.

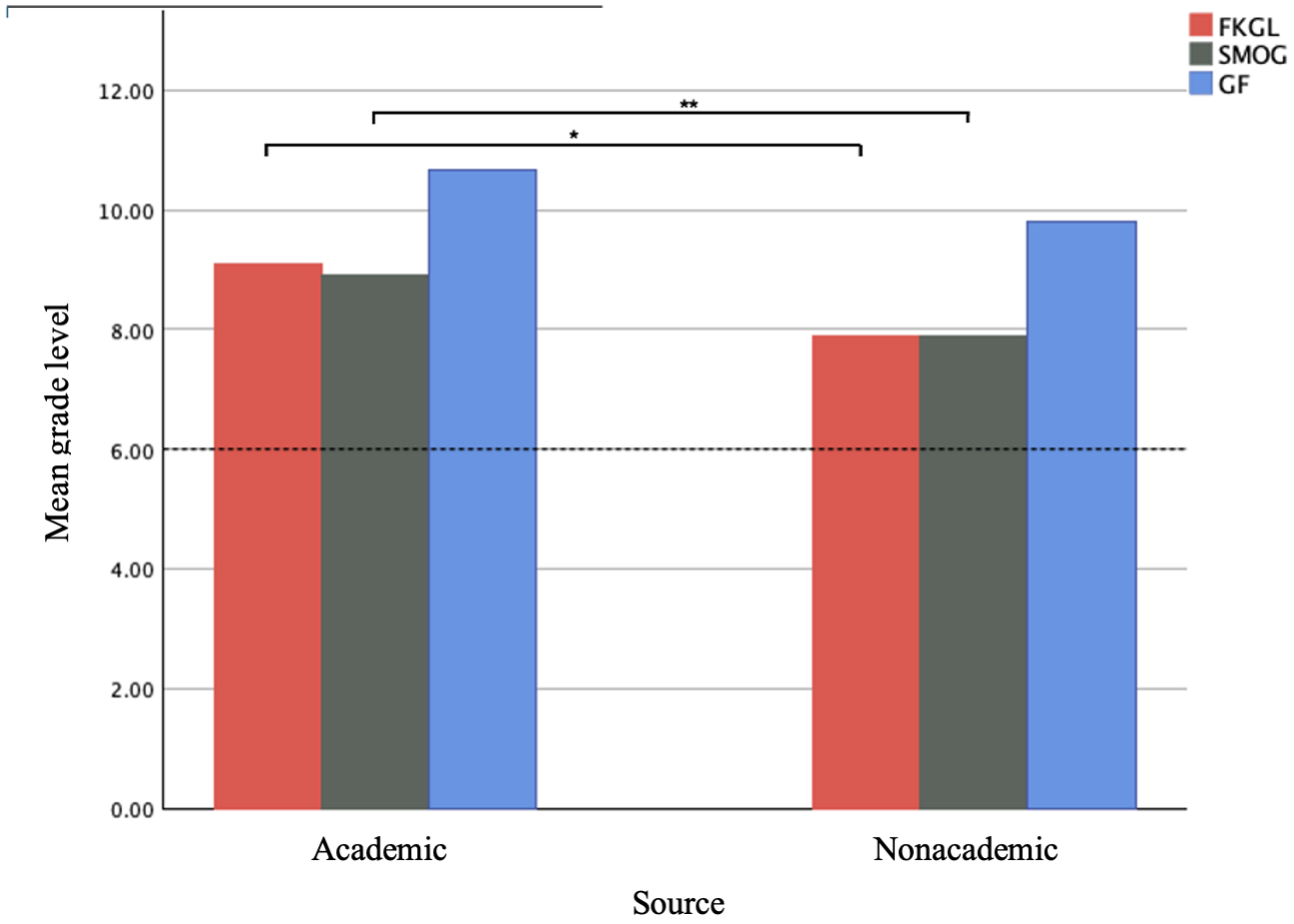

Figure 3. Mean FRE of academic and nonacademic sources. The dotted red line represents the equivalent FRE score for a 6th-grade reading level. FRE: Flesch Reading Ease. *P=.006

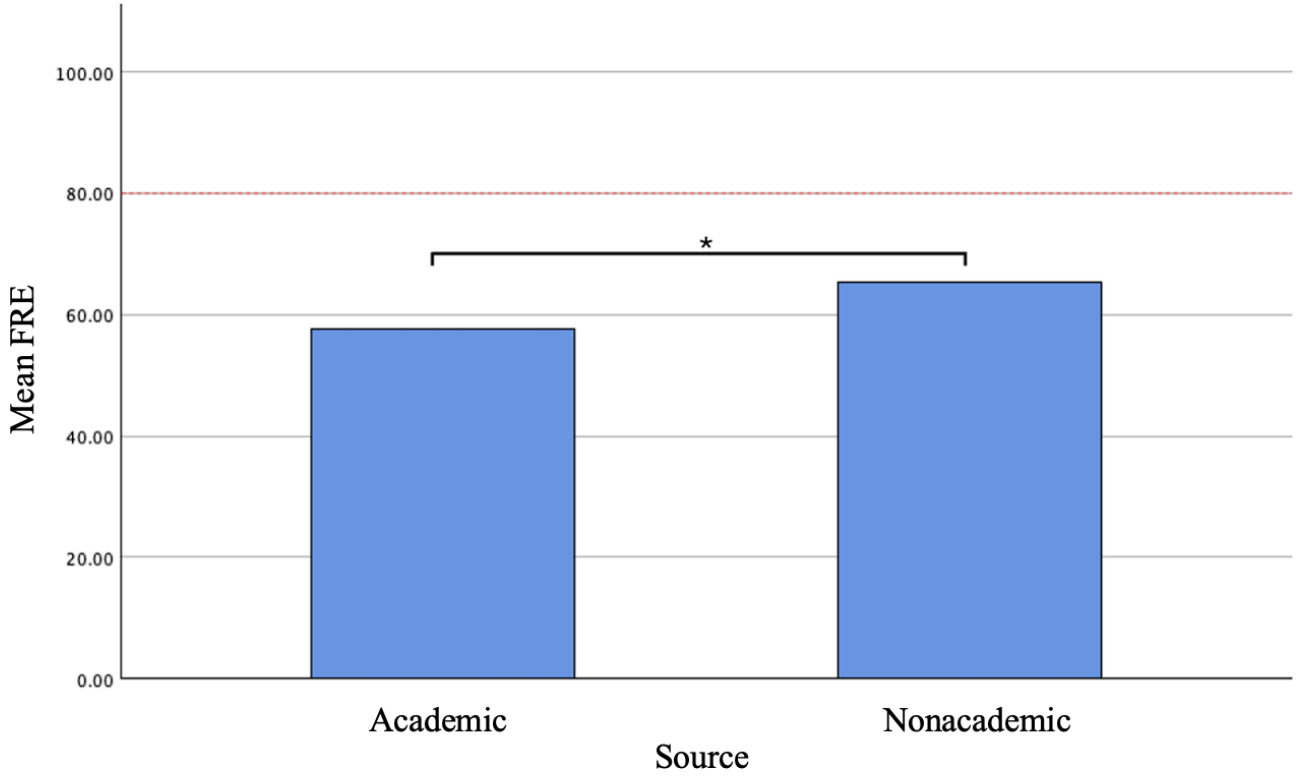

XSL•FO
RenderX

Among the 8 individual mDISCERN questions, question 1 had the greatest score amongst all the websites (4.89, SD 0.24). Academic sites (5, SD 0) had a greater score for question 1 compared to nonacademic sites (4.81 SD 0.50). Question 4 had the lowest score amongst all websites (1.83 SD 1.23). Academic sites (2 SD 1.30) had a greater score for question 4 versus nonacademic (1.73 SD 1.20) sites. Academic sites had a perfect score for question 2 as well, with a mean of 5 (SD 0).

Discussion

Principal Findings

Our study is the first to assess the readability and quality of web-based sources on exercise stress testing. These results indicate that the average FKGL, SMOG and GF scores for the 18 web pages analyzed were 8.36, 8.28, and 10.14, respectively. These average scores exceed the AMA's recommendation that health information be written at a 6th-grade level [11]. Our analysis also shows that academic sources for health information are written at a significantly higher grade level than nonacademic sources, with average FKGL scores of 9.1 and 7.9, and SMOG scores of 8.9 and 7.9, respectively. While the grade level of academic sources is statistically greater, both categories exceed the AMA recommendation for health information to be written at a 6th-grade level, indicating that even nonacademic sources may still be too complex for the average patient to fully understand. Of the 18 websites analyzed, none met the recommended reading level across all three scoring criteria. Only one source met the criteria for two readability metrics with a FKGL of 5.2, SMOG of 5.3, and GF of 6.4

As the FRE and FKGL formulae depend on many of the same parameters, it follows that the average FRE for academic sources was statistically lower than for nonacademic sources. An FRE score of 80‑90 is considered appropriate for a 6th-grade reading level [22,23]. In contrast, the average FRE scores for academic and nonacademic sources were 57.6 and 65.3, respectively. Although this difference was statistically significant, both scores indicate reading difficulty levels that exceed the recommended ease of 80‑90.

Readability was further analyzed using the SMOG and GF tools. We chose to include these readability metrics for a more thorough analysis of the websites in this study. These tools are highly regarded in the assessment of health care literature and are more modern readability models based on the initial Flesch formulas [18,27]. Scores of 6 or below on the SMOG and GF scales are considered to be in accordance with the AMA-recommended guidelines. The mean calculated SMOG scores for academic and nonacademic sources were 8.9 and 7.9, respectively. Similar to the findings from the Flesch scores, our analysis revealed statistically significant differences in the calculated SMOG scores between academic and nonacademic sources, with academic sources written at a significantly higher grade-level. Despite these differences, both source types exceeded the recommended reading level. The average GF scores for academic and nonacademic sources were 10.68 and 9.81. Although we found no significant difference between the scores between the source types, both scores indicated that the material is written at a reading level that exceeds the current recommendation.

These findings highlight that both academic and nonacademic institutions are producing resources on exercise stress tests that fail to meet current readability recommendations, potentially hindering patients' ability to fully understand their health information. This lack of concordance with recommended readability standards may limit patients' capacity to make informed decisions about their care, thereby undermining their participation in shared decision-making and their ability to manage their health effectively.

Our analysis also focused on the quality of web resources using the mDISCERN criteria. As described in our methods, web resources with a mDISCERN score below 40% classified the reliability as "poor." Scores between 40% and 79% were deemed "fair," and scores above 80% were categorized as "good." The average mDISCERN score for all web resources included in our analysis was 29.44 out of 40 (74%). This classifies the average quality of web resources as "fair" and closer to "good" than "poor." 14 of the web resources were rated as "fair," and the remaining four were rated as "good." Every web resource analyzed lost points when evaluated on questions four, five, seven, and eight. Questions four and five pertain to clearly citing the sources used in the article and providing publication dates for both the article and its references. Questions 7 and 8 assess whether the author offers resources for further information and addresses potential uncertainties, such as the risks associated with the test. We considered these areas crucial for assessment, as they were consistently lacking across all the web resources reviewed. By addressing these deficiencies, the overall quality of these resources can be significantly enhanced, providing patients with more reliable and comprehensive information.

The available web resources for patients seeking information on exercise stress tests are, on average, of fair quality but exceed the recommended reading level when assessed using the FKGL, FRE, SMOG and GF scores. These findings are consistent with other studies analyzing the readability of cardiovascular topics [13-15], suggesting that many online health resources related to cardiovascular medicine may be difficult for the average patient to understand, potentially limiting their accessibility and effectiveness in promoting patient education. Authors of both academic and nonacademic resources could improve their content by consulting the AMA guidelines on best practices for creating patient-friendly health information [11]. In particular, attention should be given to factors that reduce reading ease and elevate the reading grade level.

Limitations

As previously mentioned, the FKGL and FRE are the most widely used readability formulas in health care literature [18]. To enhance the comprehensiveness of our analysis, we incorporated the SMOG and GF readability metrics as additional objective tools for evaluating website readability. While the Flesch formulas remain the most commonly used in health care, the SMOG and GF metrics provide more modern and potentially more relevant approaches for the assessment of health-related information [18,27]. However, these formulas have inherent limitations, as they rely primarily on quantifiable factors such

XSL•FO
RenderX

as syllable count, word count, sentence count, and sentence length [22-24,28]. These tools are not capable of assessing the complexity or technical nature of health-related information. Regardless of sentence or syllable count, short sentences composed of mono- or bi-syllabic words may still contain complex medical terminology that is difficult for the reader to understand, potentially leading these formulas to underestimate the actual reading level required for comprehension. Alternative language in health-related information can be used when appropriate; however, avoiding vital terminology to lower the reading grade level may compromise the accuracy and clarity of the information provided. Simplifying medical jargon might increase readability, but at the potential cost of misinforming patients.

Additional limitations arise while using the DISCERN tool for evaluating articles. Scoring is inherently subjective, requiring an evaluator to assess a web page and rate each question accordingly. While this nonformulaic approach allows for a more nuanced evaluation, individual evaluator bias and interpretation can affect the consistency and standardization of scores.

Conclusion

The web resources available to patients seeking information regarding exercise stress tests are written at a reading level that surpasses the current AMA recommendation. This is complicated by the inherently complex nature of medical terminology. Despite this, efforts should be made by authors to use alternative terms or omit terms altogether if the medical jargon being used does not contribute to the message attempting to be conveyed. The quality of the web resources evaluated were determined to be fair but could be improved by providing more comprehensive information. While improving health literacy across the population is crucial, it is equally important to provide patients with clear, accessible materials that empower them to make informed decisions about their health. These materials must also be high-quality, offering relevant, unbiased, and concise information, along with appropriate access to sources and additional resources for those seeking further guidance.

Conflicts of Interest

None declared.

References

1. Sidoti OWR, Faverio M, Atske S, Radde K, Park E. Internet, broadband fact sheet. Pew Research Center. 2024 Feb. URL: https://www.pewresearch.org/internet/fact-sheet/internet-broadband/ [accessed 2024-08-18]
2. Health information technology use among adults: United States, July-December 2022. CDC Stacks. 2023. URL: https://stacks.cdc.gov/view/cdc/133700
3. Prestin A, Vieux SN, Chou WYS. Is online health activity alive and well or flatlining? Findings from 10 years of the Health Information National Trends Survey. J Health Commun 2015;20(7):790-798. [doi: 10.1080/10810730.2015.1018590] [Medline: 26042588]
4. Baker DW, Wolf MS, Feinglass J, Thompson JA, Gazmararian JA, Huang J. Health literacy and mortality among elderly persons. Arch Intern Med 2007 Jul 23;167(14):1503-1509. [doi: 10.1001/archinte.167.14.1503] [Medline: 17646604]
5. Bostock S, Steptoe A. Association between low functional health literacy and mortality in older adults: longitudinal cohort study. BMJ 2012 Mar 15;344(mar15 3):e1602. [doi: 10.1136/bmj.e1602] [Medline: 22422872]
6. Grandinetti DA. Doctors and the Web. Help your patients surf the Net safely. Med Econ 2000 Mar 6;77(5):186-188. [Medline: 10848409]
7. Sullivan B, Abed V, Joiner J, et al. The quality of online information for the treatment of knee osteoarthritis: a google study. Cureus 2022 Oct;14(10):e29995. [doi: 10.7759/cureus.29995] [Medline: 36381839]
8. De La Chapa JS, Bellinger JR, Timm AR, Daniero JJ. Quality, readability, and understandability of online posterior glottic stenosis information. J Voice 2025 May;39(3):854. [doi: 10.1016/j.jvoice.2022.11.036] [Medline: 36585308]
9. Arts H, Lemetyinen H, Edge D. Readability and quality of online eating disorder information-Are they sufficient? A systematic review evaluating websites on anorexia nervosa using DISCERN and Flesch Readability. Int J Eat Disord 2020 Jan;53(1):128-132. [doi: 10.1002/eat.23173] [Medline: 31589331]
10. Oman SP, Zaver H, Waddle M, Corral JE. Quality and readability of web-based information for patients with pancreatic cysts: DISCERN and readability test analysis. JMIR Cancer 2021 Mar 16;7(1):e25602. [doi: 10.2196/25602] [Medline: 33724203]
11. Barry DW. Health literacy and patient safety: help patients understand. In: Manual for Clinicians, 2nd edition: American Medical Association Foundation; 2007. URL: https://med.fsu.edu/sites/default/files/userFiles/file/ahec_health_clinicians_manual.pdf [accessed 2024-08-16]
12. Garner KK, Pomeroy W, Arnold JJ. Exercise Stress Testing: Indications and Common Questions. Am Fam Physician 2017 Sep 1;96(5):293-299. [Medline: 28925651]
13. Siddiqui E, Shah AM, Sambol J, Waller AH. Readability assessment of online patient education materials on atrial fibrillation. Cureus 2020 Sep 11;12(9):e10397. [doi: 10.7759/cureus.10397] [Medline: 33062517]

XSL•FO
RenderX

14. Wasir AS, Volgman AS, Jolly M. Assessing readability and comprehension of web-based patient education materials by American Heart Association (AHA) and CardioSmart online platform by American College of Cardiology (ACC): How useful are these websites for patient understanding? Am Heart J Plus 2023 Aug;32:100308. [doi: 10.1016/j.ahjo.2023.100308] [Medline: 38510202]
15. Ayyaswami V, Padmanabhan D, Patel M, et al. A readability analysis of online cardiovascular disease-related health education materials. Health Lit Res Pract 2019 Apr;3(2):e74-e80. [doi: 10.3928/24748307-20190306-03] [Medline: 31049489]
16. Search engine market share worldwide. GlobalStats S. URL: https://gs.statcounter.com/search-engine-market-share [accessed 2025-09-24]
17. The value of Google result positioning. Chitika, Inc. URL: https://www.silesiasem.pl/wp-content/uploads/2013/07/chitikainsights-valueofgoogleresultspositioning.pdf [accessed 2024-08-18]
18. Wang LW, Miller MJ, Schmitt MR, Wen FK. Assessing readability formula differences with written health information materials: application, results, and recommendations. Res Social Adm Pharm 2013;9(5):503-516. [doi: 10.1016/j.sapharm.2012.05.009] [Medline: 22835706]
19. Fishburne RP, Chissom BS, Kincaid JR. Derivation of new readability formulas (automated readability index, FOG count, and flesch reading ease formula) for navy enlisted personnel. University of Central Florida–Institute for Simulation and Training. 1975. URL: https://apps.dtic.mil/sti/pdfs/ADA006655.pdf [accessed 2024-08-18]
20. What is the Flesch-Kincaid readbaility test? Social Security Administration of the United States of America. URL: https://www.ssa.gov/agency/plain-language/assets/materials/2023-plainwritingcompliancereport.pdf [accessed 2024-08-18]
21. Stockmeyer N. Using Microsoft Word's Readability Program. Mich BJ 2009:46-47 [FREE Full text]
22. FLESCH R. A new readability yardstick. J Appl Psychol 1948 Jun;32(3):221-233. [doi: 10.1037/h0057532] [Medline: 18867058]
23. Flesch R. How to Write Plain English, 1st edition: Barnes & Noble; 1981:144 URL: https://pages.stern.nyu.edu/wstarbuc/Writing/Flesch.htm [accessed 2024-08-18]
24. McLaughlin GH. Temptations of the Flesch. Instr Sci 1974 Jan;2(4):367-383. [doi: 10.1007/BF00123459]
25. Rouhi AD, Ghanem YK, Yolchieva L, et al. Can artificial intelligence improve the readability of patient education materials on aortic stenosis? A pilot study. Cardiol Ther 2024 Mar;13(1):137-147. [doi: 10.1007/s40119-023-00347-0] [Medline: 38194058]
26. Onder CE, Koc G, Gokbulut P, Taskaldiran I, Kuskonmaz SM. Evaluation of the reliability and readability of ChatGPT-4 responses regarding hypothyroidism during pregnancy. Sci Rep 2024 Jan 2;14(1):243. [doi: 10.1038/s41598-023-50884-w] [Medline: 38167988]
27. Fitzsimmons PR, Michael BD, Hulley JL, Scott GO. A readability assessment of online Parkinson's disease information. J R Coll Physicians Edinb 2010 Dec;40(4):292-296. [doi: 10.4997/JRCPE.2010.401] [Medline: 21132132]
28. Gunning R. The Technique of Clear Writing: McGraw-Hill; 1952.
29. Griffiths KM, Christensen H. Website quality indicators for consumers. J Med Internet Res 2005 Nov 15;7(5):e55. [doi: 10.2196/jmir.7.5.e55] [Medline: 16403719]
30. Charnock D, Shepperd S, Needham G, Gann R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. J Epidemiol Community Health 1999 Feb;53(2):105-111. [doi: 10.1136/jech.53.2.105] [Medline: 10396471]
31. Kumar VS, Subramani S, Veerapan S, Khan SA. Evaluation of online health information on clubfoot using the DISCERN tool. J Pediatr Orthop B 2014 Mar;23(2):135-138. [doi: 10.1097/BPB.0000000000000000] [Medline: 24048196]
32. Scott B. Readability scoring system plus. Readability Formulas. 2024 Nov 7. URL: https://readabilityformulas.com/readability-scoring-system.php [accessed 2025-09-24]
33. Exercise stress test. Cleveland Clinic. URL: https://my.clevelandclinic.org/health/diagnostics/16984-exercise-stress-test [accessed 2024-08-28]
34. Stress test. Mayo Clinic. URL: https://www.mayoclinic.org/tests-procedures/stress-test/about/pac-20385234 [accessed 2024-08-28]
35. Cardiac exercise stress testing: what it can and cannot tell you. Harvard Health. URL: https://www.health.harvard.edu/newsletter_article/cardiac-exercise-stress-testing-what-it-can-and-cannot-tell-you [accessed 2024-08-28]
36. What to expect during an exercise stress test. Harvard Health. URL: https://www.health.harvard.edu/heart-health/what-to-expect-during-an-exercise-stress-test [accessed 2024-08-28]
37. Exercise stress test. Harvard Health. URL: https://www.health.harvard.edu/heart-health/exercise-stress-test [accessed 2024-08-28]
38. Exercise electrocardiogram. Hopkins Medicine. URL: https://www.hopkinsmedicine.org/health/treatment-tests-and-therapies/exercise-electrocardiogram [accessed 2024-08-28]
39. Exercise stress test. Massachusetts General Hospital. URL: https://www.massgeneral.org/heart-center/treatments-and-services/exercise-stress-test [accessed 2024-08-28]
40. Exercise stress test. American Heart Association. URL: https://www.heart.org/en/health-topics/heart-attack/diagnosing-a-heart-attack/exercise-stress-test [accessed 2024-08-28]

XSL•FO
RenderX

41. Exercise stress test. St Vincent's Heart Health. URL: https://www.svhhearthealth.com.au/procedures/tests/exercise-stress-test [accessed 2024-08-28]
42. Exercise stress test. Medline Plus. URL: https://medlineplus.gov/ency/article/003878.htm [accessed 2024-08-28]
43. Heart disease and stress tests. WebMD. URL: https://www.webmd.com/heart-disease/stress-test [accessed 2024-08-28]
44. How to prepare for an exercise stress test. Healthline. URL: https://www.healthline.com/health/exercise-stress-test#risks [accessed 2024-8-28]
45. Exercise ECG. British Heart Foundation. URL: https://www.bhf.org.uk/informationsupport/tests/exercise-ecg [accessed 2024-08-28]
46. Exercise electrocardiogram. Heart and Stroke Foundation. URL: https://www.heartandstroke.ca/heart-disease/tests/exercise-electrocardiogram [accessed 2024-08-28]
47. Exercise electrocardiogram. MyHeart Alberta. URL: https://myhealth.alberta.ca/Health/aftercareinformation/pages/conditions.aspx?hwid=ug5887 [accessed 2024-08-28]
48. Exercise stress test. HeartWest. URL: https://www.heartwest.com.au/services/exercise-stress-test-ecg/ [accessed 2024-08-28]
49. Exercise stress ECG (treadmill test). SOZO Cardiology. URL: https://www.sozocardiology.com/heart-investigations/exercise-stress-ecg-treadmill-test/ [accessed 2024-08-28]
50. Exercise ECG. Allevia Cardiology Ascot. URL: https://www.alleviacardiology.co.nz/our-services/diagnostic-services/exercise-ecg [accessed 2024-08-28]

Abbreviations

AMA: American Medical Association
FKGL: Flesch-Kincaid Grade Level
FRE: Flesch Reading Ease
GF: Gunning Fog
mDISCERN: modified DISCERN
SMOG: Simple Measure of Gobbledygook

Edited by A Mavragani; submitted 25.10.24; peer-reviewed by M Hollings, S Mehrabi; revised version received 30.04.25; accepted 08.05.25; published 22.10.25.

Please cite as:
Rahbe M, Mistry D, Khawaja AM, Sous NA, Tso AY
Assessing the Readability and Quality of Web-Based Resources on Exercise Stress Testing: Cross-Sectional Readability and Quality Analysis
JMIR Form Res 2025;9:e68000
URL: https://formative.jmir.org/2025/1/e68000
doi:10.2196/68000

© Munir Rahbe, Dhrumi Mistry, Ali M Khawaja, Natalie A Sous, Alan Y Tso. Originally published in JMIR Formative Research (https://formative.jmir.org), 22.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Participatory Development and Concept Testing of mHealth Messaging to Support Care Engagement and Antiretroviral Therapy Adherence for Women Living With HIV in the Southern United States: Focus Group Study

Sadie B Sommer[1], MPH; Julie V Barroso[1], ANP, RN, PhD; Sarah B Bass[2], MPH, PhD; Marianne R Choufani[3], BSN, RN, PhD; Alexander M Schoemann[4], PhD; Katie J Singley[2], MPH; Caseem C Luck[2], MS; Courtney E Caiola[3], RN, MPH, PhD

[1]Center for Research Development and Scholarship, School of Nursing, Vanderbilt University, 461 21st Ave. S, Nashville, TN, United States

[2]Department of Social and Behavioral Sciences, Temple University, Philadelphia, PA, United States

[3]College of Nursing, East Carolina University, Greenville, NC, United States

[4]Department of Psychology, East Carolina University, Greenville, United States

Corresponding Author:
Sadie B Sommer, MPH
Center for Research Development and Scholarship, School of Nursing, Vanderbilt University, 461 21st Ave. S, Nashville, TN, United States

Abstract

Background: Women living with HIV in the Southern United States, or the South, face persistent and overlapping challenges to care engagement and antiretroviral therapy adherence, including HIV-related stigma, poverty, and inequitable access to health care. While mobile health (mHealth) interventions show promise for enhancing self-management and care engagement among people living with HIV, interventions tailored to women living with HIV remain limited, particularly those developed through participatory approaches that center their lived experiences.

Objective: This study sought to evaluate the acceptability, comprehensibility, and personal relevance of targeted health messages developed for a proposed mHealth app tailored to women living with HIV in the South. In addition, it explored participants' perceptions of the feasibility and desirability of the proposed intervention.

Methods: This study represents phase 3 of a multistage, mixed methods project. Message content was informed by earlier phases, which included individual interviews, surveys, perceptual mapping with women living with HIV, and input from a community and clinician advisory board. In this phase, 3 focus groups (2 virtual and 1 in person) were conducted with 30 women living with HIV recruited from Southern HIV clinics and community organizations. Participants reviewed prototype wireframes and health messages, including SMS text message–style content, and provided feedback on all content. Data were analyzed using conventional content analysis.

Results: Participants expressed strong interest in the proposed mHealth app and emphasized the importance of health messaging that is clear, supportive, and personally meaningful. Four key categories emerged: (1) acceptability of a tailored mHealth app, with participants noting the value of privacy, accessibility, and convenience; (2) acceptability of message content, including preferences for affirming, uplifting language and images; (3) personal relevance, particularly for messages addressing stigma, spirituality, family, and empowerment; and (4) comprehensibility, highlighting the need for plain language and visual clarity.

Conclusions: These findings support the development of a tailored mHealth intervention for women living with HIV in the South. Co-designed messages that center affirmation, spirituality, and real-life challenges were perceived as acceptable, comprehensible, and highly relevant. Future work will focus on refining the content and prototype testing.

(JMIR Form Res 2025;9:e76286) doi:10.2196/76286

KEYWORDS

mobile health; mHealth; concept testing; health messaging; HIV care engagement; antiretroviral therapy adherence; ART adherence; HIV self-management; focus groups; stigma reduction; women living with HIV

XSL•FO
RenderX

Introduction

Background

Despite significant advancements in HIV treatment, women living with HIV continue to face persistent barriers to sustained engagement in care and adherence to antiretroviral therapy (ART) [1], a critical strategy for achieving viral suppression [2]. This disparity is most pronounced for women living with HIV in the Southern United States, or the South, as defined by the South census region for the Centers for Disease Control and Prevention's HIV funding [3]. Research consistently shows that women living with HIV in these states are disproportionately impacted by region-specific barriers that prevent viral suppression, such as poverty [4,5] and heightened stigma [6]. Additional geographic and structural barriers such as rurality [7], inequitable access to health care [8], and regional policies and politics [9,10] further highlight this significant and persistent health disparity. These contextual factors have contributed to higher HIV prevalence [11], low viral suppression rates [12], and increased mortality [13] for women living with HIV in the South. National data show that women living with HIV in the South are 10% to 15% less likely to achieve viral suppression than women in the Northeast and West [14], and HIV-related mortality is nearly 3 times higher in the South than in other regions [9,15]. Together, these disparities underscore the need for research that centers on the specific contexts and lived experiences of Southern women living with HIV, as well as interventions that address those needs.

Conceptual Framework

Considering these contextual factors and layered inequities, incorporating social location, a key sensitizing concept derived from intersectional frameworks [16,17], offers a more nuanced approach to developing interventions that reflect the complexity of individual experiences. Caiola et al [16,17] offer a detailed discussion of the intersectional framework and associated concepts guiding this mixed methods study, but in short, social location refers to the dynamic consideration of contextual factors and the position that people occupy in their social hierarchies based on intersecting systems of oppression and social determinants of health. In this study, we operationalized social location to encompass both individual-level sociodemographic factors (such as race, gender, and socioeconomic status) and broader relational and structural contexts, including stigma, family and partner relationships, spirituality, and regional inequities. Acknowledging social location in message and intervention development requires moving beyond broad demographic groups such as women living with HIV or pregnant women with HIV to account for the variation in experiences and decision-making that emerge from intersecting identities and structural conditions. For example, at the individual level, women living with HIV in similar sociodemographic groups may perceive their social locations differently and make decisions about care engagement and ART adherence based on these unique interpretations.

In this study, social location served as a conceptual lens for intervention development by emphasizing how intersecting identities and structural inequities shape women's decisions about care and adherence. This conceptual approach guided us to consider within-group differences among women living with HIV in the South and prioritize messaging that was flexible, affirming, and responsive to diverse lived experiences. Interventions are typically aimed at broad demographic groups without considering these important within-group variations and how individuals respond with vulnerability and resilience to similar experiences [18]. To most effectively reach this population, priorities must shift to first consider social location before focusing on engagement and adherence. Using this conceptual lens throughout development allows for adaptive, reflective intervention messaging that supports the whole being and well-being of women living with HIV. Achieving this requires personalized, accessible, and scalable interventions that consider all factors contributing to engagement and adherence for women living with HIV in the South [17].

Prior Work

Digital health interventions, particularly mobile health (mHealth) apps for HIV self-management, have demonstrated positive effects on adherence, prevention, and engagement for people living with HIV [19,20]. However, these interventions have primarily been designed for the general population living with HIV, leaving a critical gap in tools that address the specific needs and contexts of women living with HIV in the South. This gap will continue to expand as the vast majority of Americans (91%) now own smartphones, including 84% of low-income Americans [21], a category inclusive of many people living with HIV [22]. mHealth has proven particularly successful for self-management across a spectrum of chronic diseases due to its ability to offer on-demand support, reminders, educational materials, and goal setting [23]. mHealth is also particularly suitable for people living with HIV, who face higher levels of stigma and discrimination, given the relative privacy and accessibility that mobile devices offer [24].

mHealth apps are increasingly being developed for specific demographics of people living with HIV [25-31] and address targeted outcomes such as stigma [32-34], stress reduction [35], resilience [36], and coping skills [37]. Some mHealth apps have been designed for specific segments of women living with HIV. For instance, Stockman et al [38] implemented a peer navigation and networking app for Black women living with HIV affected by interpersonal violence, whereas Coleman et al [39] created an SMS text messaging–based app for pregnant women living with HIV. In addition, other work has begun to explore mHealth tools for women living with HIV through pilot and proof-of-concept studies, such as the evaluation by Njie-Carr et al [40] of a technology-enhanced intervention for older women living with HIV. In these cases, women have primarily participated as recipients, reflecting progress but also underscoring the need for participatory approaches. There is a critical need for interventions that are context specific, co-designed, and grounded in women's unique lived realities. To our knowledge, no intervention has systematically applied an intersectional framework to codevelop health messages with women living with HIV in the South. This study addresses this gap by conducting participatory concept testing of tailored mHealth messages, offering a novel approach that centers

XSL•FO
RenderX

regional context, lived experience, and resilience alongside clinical adherence goals.

Goal of This Study

This paper outlines the development process and initial concept testing of targeted HIV care engagement and medication adherence messaging and possible use of an mHealth app as the primary intervention modality for health message delivery for women living with HIV in the South. In preceding phases, we engaged women living with HIV in a structured, stepwise process to inform the development of the health messaging and the mobile app's features, functionality, and content. This process is described in greater detail in the Phase 1, Phase 2, and Phase 3 sections, with published findings from phase 1 [1,17] and additional findings from phase 2 currently under review. Through comparative concept testing, we aimed to assess the acceptability, comprehensibility, and personal relevance of targeted health messaging and content to support medication adherence and health care engagement among women living with HIV in the South. Findings from this study will guide the development and testing of the mHealth app.

Methods

Design

Phase 1

This paper presents the results of phase 3 of an exploratory, multistage sequential mixed methods study [41] designed to systematically develop and pretest HIV care engagement and ART medication adherence messaging (Figure 1). Data integration occurred throughout, and each phase of the work was conducted in collaboration with a community and clinician advisory board (CCAB) comprising women living with HIV and expert clinicians and researchers working with women living with HIV. Phase 1 of the sequential design included in-depth individual interviews (n=40) to determine how women living with HIV in the South perceive their social locations and make decisions regarding HIV care engagement and medication adherence. Findings from this phase have been published previously [1,17].

Figure 1. Exploratory multistage, sequential mixed methods design. mHealth: mobile health.

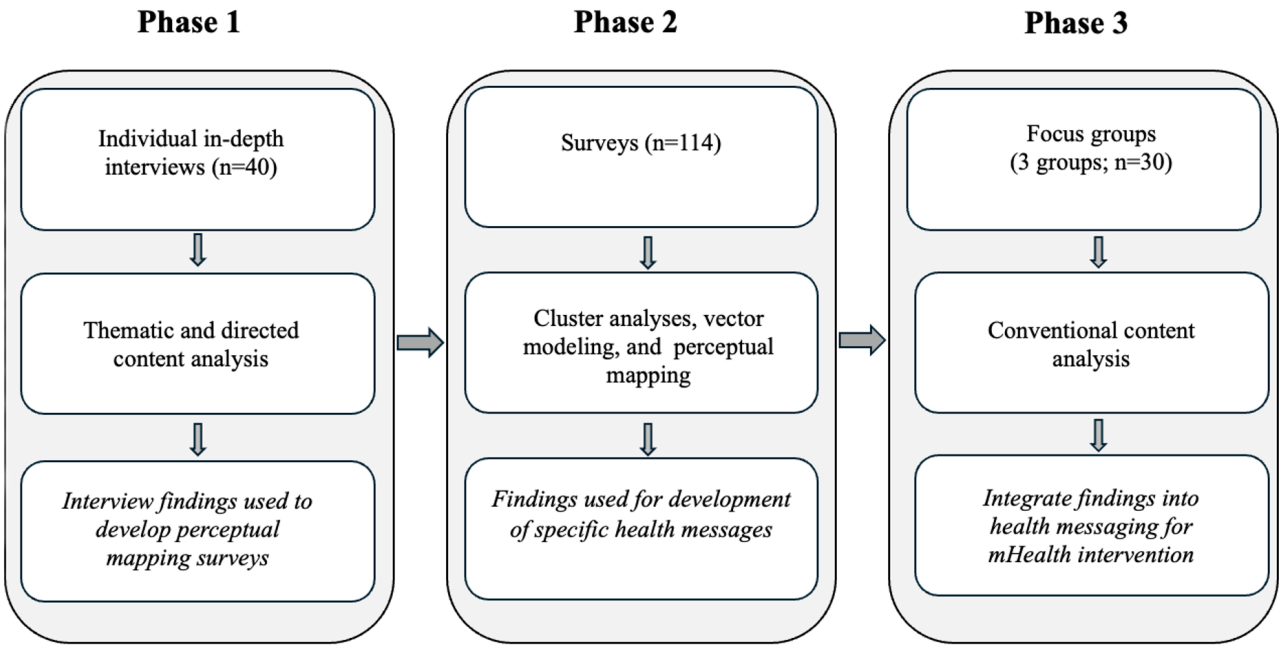

Phase 2

These findings guided the creation of surveys to further evaluate perceptions of social location and decision-making, which were administered in phase 2 (n=114). Cluster analyses, perceptual mapping, and vector modeling analyses of the phase 2 data informed the creation of a set of specific health messages designed to enhance decision-making regarding HIV care engagement and ART adherence (Table 1). Perceptual mapping and vector message modeling are novel methods that use multidimensional scaling analysis to produce mathematical 3D models, enabling the creation of highly targeted messages that conventional survey methods cannot achieve [42,43]. They also allow for comparing messaging strategies across groups by psychographic variables such as levels of stigma or by demographics to assess how persuasive messaging may be the same or different [44-46]. On the basis of vector modeling, we discerned that messaging could be incorporated to meet the needs of most women and developed messages, taglines, and content to reflect the themes identified in the analysis as most likely to be persuasive in encouraging care engagement for women living with HIV. Repetitions across clusters reflected areas of overlap identified in the data. These redundancies were intentionally retained to illustrate recurring barriers and facilitators. The findings of this work are currently under review. Perceptual mapping and vector modeling identified persuasive messages, which we refined through the intersectional framework to ensure that they reflected the women's lived realities, including stigma, spirituality, family roles, and resource availability, and guided the categories pretested in phase 3.

XSL•FO
RenderX

Table . Messaging strategies by stigma clusters.

	HIV stigma clusters[a]			Intersectional stigma clusters[b]		
	High	Medium	Low	High	Medium	Low
Barriers to ART[c]	• Too many side effects • Reminds me I have HIV • Don't want friends to know	• Too many side effects • Get AIDS anyway • Don't want family to know • Reminds me I have HIV	• Inconclusive	• ART is expensive • Reminds me I have HIV • Don't want friends to know	• ART is expensive • Reminds me I have HIV • Don't want family to know	• ART is expensive • Don't want family to know • Don't want friends to know
Benefits to ART	• Feel more in charge • Feel more confident • Role model for others	• Can take care of family • Feel closer to partner • Avoid getting sick	• Inconclusive	• Avoid getting sick • Have a healthy, normal life	• Enjoy sex more • ART is safe and effective	• Have a healthy, normal life • Stay undetectable
Beliefs about living with HIV	• Limit who knows • Easy access to other resources • Dating discrimination	• Limit who knows	• Limit who knows • Discriminated against, race	• Inconclusive	• Limit who knows • Had negative reactions to status	• Inconclusive
Facilitators to ART	• One pill a day • Having resources and stability • Having kids and family to motivate me	• Having resources and stability	• Inconclusive	• Having resources and stability	• Inconclusive	• Having resources and stability • Having kids and family to motivate me
Care engagement facilitators	• Not great but appointment reminders help	• Attending appointments prioritizes health	• Appointment reminders help • Gas cards and/or transportation services help	• Inconclusive	• Not great – Family support helps	• Transportation helps • Good relationship with doctor helps
Appointment barriers	• Inconclusive	• Inconclusive	• Inconclusive	• Inconclusive	• Responsibilities	• Inconclusive

[a]HIV-related stigma clusters represent messaging strategies targeting stigma specifically associated with HIV status (eg, disclosure concerns, internalized stigma, and fear of treatment).

[b]Intersectional stigma clusters represent messaging strategies addressing the compounding of HIV-related stigma with other stigmas tied to gender, race, socioeconomic status, or structural inequities. Both were derived from phase 2 perceptual mapping and vector modeling analyses.

[c]ART: antiretroviral therapy.

Phase 3

In phase 3, we used a descriptive qualitative design that included 3 focus groups with women living with HIV in the South, engaging a total of 30 participants. We pretested specific health messaging for comprehensibility, acceptability, and personal relevance, along with visual wireframes of how an mHealth app would use those messages and the potential of an mHealth app as the primary intervention modality for health message delivery. Recent studies confirm that focus groups are well suited for developing and tailoring mHealth interventions [47,48]. We also chose this method because it enabled efficient testing of multiple message variations and allowed for collective refinement.

To summarize the 3 phases, 40 women completed individual interviews to identify decision-making factors (phase 1), 114 women completed surveys using perceptual mapping and vector modeling to refine messaging strategies (phase 2), and 30 women participated in focus groups to assess feasibility and acceptability (phase 3) within the sequential mixed methods design.

XSL•FO
RenderX

Ethical Considerations

The East Carolina University and Medical Center Institutional Review Board (21-001403) and the Vanderbilt University Institutional Review Board (211919) reviewed and approved all study protocols. The informed consent process was conducted by trained study personnel before all focus groups. The virtual focus group was conducted via a HIPAA (Health Insurance Portability and Accountability Act)-compliant virtual platform, Webex (Webex by Cisco). The virtual focus group participants reviewed, signed, and received a copy of the consent form electronically via a secure platform called Docusign (Docusign, Inc), and the face-to-face participants reviewed, signed, and received a paper copy. To address issues related to confidentiality, participants were asked to designate a pseudonym to use for all communications following the initial contact phone call, and each participant was assigned a unique ID number. Each participant received a $50 gift card for their time and contribution.

Recruitment and Inclusion Criteria

Participants were recruited across a broad geographic area of the South [3], with support from CCAB members and staff at North Carolina and Tennessee recruitment sites. Recruitment strategies included (1) distributing flyers in community-based organizations and clinics to encourage self-referral; (2) obtaining referrals from health care providers for women living with HIV both in and out of care; (3) snowball sampling; and (4) sharing flyers through social media platforms and HIV and AIDS support organization email listserves connected to CCAB members, study personnel, and recruitment site staff. The eligibility criteria for participation included (1) self-identifying as a woman; (2) self-identifying as Black, Latina, or White; (3) living with HIV or AIDS; (4) being aged 18 years or older; (5) being able to read and communicate in English; and (6) being mentally competent to provide informed consent.

Data Collection

To reach women living with HIV from a broad geographic distribution in the South and increase accessibility and feasibility, we conducted 2 smaller virtual focus groups (n=7 and n=5 participants) and 1 larger in-person focus group (n=18 participants). To ensure fidelity in the data collection procedures across groups and modalities (virtual vs face-to-face), the research team developed a detailed script for reviewing all the health messaging materials. The focus group interview guide included questions derived from the National Cancer Institute's *Making Health Communication Programs Work* guide [49] to assess the comprehensibility, acceptability, and personal relevance of each specific health message. In addition, participants were asked questions regarding the acceptability of using a mobile app as the primary intervention modality for delivering health messaging. Before the focus groups, the CCAB vetted the proposed messaging and visual wireframes, offering both oral and written feedback used in refining the materials. The focus group facilitator provided a booklet containing the proposed health messaging and images to each study participant. Virtual focus group participants received an electronic copy of the booklet via their preferred email address before the focus group, whereas in-person focus group participants received a hard copy of the booklet at the focus group session. Participants were presented with variations of logos, log-in pages, potential messaging, and taglines as they would appear on their cellphones. They were also provided with examples of 12 weeks of SMS text messages to be delivered at a frequency of 3 per week. The BRIDGE Collective, or BRIDGE for short, was the proposed name of the mHealth app. Examples of weekly SMS text messages are provided in Table 2. After reviewing each option sequentially, participants were asked a series of questions about comprehensibility, acceptability, and personal relevance (eg, "Do the messages/images make sense?" "Do you like them?" "Do they speak to your personal experiences? If not, how would you suggest we change them?"). Examples of the images shown to participants are shown in Figures 2-4.

XSL•FO
RenderX

Table . Potential SMS text messages.

Category	Example message	Purpose
Reminder	“Did u take ur medication today?”	Encourages medication adherence.
Reminder	“Remember...set your alarm, take your meds every day. Text if u have ??s”	Encourages medication adherence at a specific time.
Encouragement	“Whoa, week three. U got this...text if U have ??s”	Boosts morale and promotes self-confidence.
Encouragement	“Week six! Great job doing what’s right for u! Love, the BRIDGE Collective”	Provides positive reinforcement to continue engagement.
Self-care	“When you are well, your family is well.”	Emphasizes the link between self-care and family well-being.
Self-care	“You are a role model for others when you take your medications.”	Highlights self-care as a positive example for others.
Support	“Hey, the BRIDGE Collective here. Are things getting you down? Come talk to us for help.”	Offers emotional support and connection to resources.
Support	“Remember to find your people/engage your higher power. We are here for you!”	Reinforces social and spiritual support networks.
Prompt	“Hey, the BRIDGE Collective here. Remember, if you are not working on HIV, you need to know it is working on you. Let us know if we can help!”	Open-ended prompt to check in and offer assistance.
Prompt	“See stigma for what it is: ignorance. Take Charge. The BRIDGE Collective is here for you.”	Promotes action against stigma and encourages empowerment.

XSL•FO
RenderX

Figure 2. Potential mobile app initial log-in page.

XSL•FO
RenderX

Figure 3. Potential messages to enhance care engagement.

XSL•FO
RenderX

Figure 4. Potential spiritual messages.

Data Management and Analysis

The sociodemographic data collected from participants in the face-to-face focus group via the paper survey were manually input into REDCap by study personnel. The sociodemographic data from all 3 focus groups were then analyzed using descriptive statistics. Audio recordings of each focus group were transcribed verbatim, verified for accuracy by study personnel, and stored on a secure drive. The qualitative data analytic team used conventional content analysis developed by Hsieh and Shannon [50] to characterize the participants' perceptions. This approach involves inductively deriving codes from the data without predetermined categories, making it well suited to our aim of capturing participant perspectives on messaging and app features in their own words. Team members (SS and CC) read the transcripts multiple times, made note of their initial impressions, and developed initial codes derived directly from the text. To ensure trustworthiness, the team engaged in reflexivity throughout the analytic process by avoiding paraphrasing, cross-checking interpretations, and explicitly acknowledging our own positionality and potential biases during team discussions. Iterative team-based coding served as an intercoder reliability check, with discrepancies discussed until a consensus was reached [51]. The initial codes were then grouped into meaningful clusters and sorted into categories and subcategories. The field notes from each of the focus groups were also combined and integrated into the data analysis as they were used to triangulate initial findings and refine the categories and subcategories. Finally, example quotes from the data were linked with each of the proposed categories and subcategories and presented to the larger research team for reflection, discussion, and additional refinement. Analytic saturation was determined once no new themes emerged across transcripts, consistent with established standards for qualitative analysis [52].

Results

Sample Characteristics

A total of 30 individuals who identified as women participated in the study. The mean age of the participating women was 58 years and ranged from 29 to 75 years. Most (26/30, 87%) identified as Black, with smaller proportions identifying as White (3/30, 10%) or Latina or Hispanic (1/30, 3%). Over half

XSL•FO
RenderX

(19/30, 63%) were above the federal poverty level, and most lived in nonrural county-level designations (27/30, 90%) and were currently engaged in care (27/30, 90%). The demographic characteristics of the study participants are shown in Table 3.

Table . Demographic characteristics of the participants (N=30).

Variable	Values
Age (years), mean (range)	58 (29-75)
Race or ethnicity, n (%)	
Black	26 (87)
Hispanic or Latina	3 (10)
White	1 (3)
Gender identity, n (%)	
Woman	30 (100)
Man	0 (0)
Socioeconomic status[a], n (%)	
Above FPL[b]	19 (63)
Below FPL	11 (37)
Geographic location[c], n (%)	
Nonrural	27 (90)
Rural	3 (10)
Engaged in HIV care, n (%)[d]	
Engaged	27 (90)
Not engaged	2 (7)
No answer	1 (3)

[a]Socioeconomic status was measured using the US Census Bureau standards, self-reported annual household income from all sources, and the number of people in the household [53].

[b]FPL: federal poverty level.

[c]Geographic location was measured using the participants' self-reported state and county as the criteria and the Federal Office of Rural Health Policy designation as rural or nonrural [54].

[d]HIV care visit with viral load tested in the previous 6 months.

Content Analysis

Through comprehensive analysis of the focus group transcripts and field notes, four primary categories were identified: (1) acceptability of mHealth app features and functionality, (2) acceptability of message content, (3) personal relevance of messaging, and (4) intersectional stigma and contextual barriers. These themes provide an overview of participants' perspectives and are described in detail in the following sections.

Category 1: Acceptability of mHealth

Convenience and Accessibility

Participants expressed enthusiasm for the concept of an mHealth app tailored for women living with HIV, particularly for the convenience and flexibility it may provide. One participant stated the following:

> *For me, it would be very convenient. Didn't have to get in front of a computer or show up in person.* [Online 2]

Flexibility to engage on their own time was also noted as a benefit, as one participant explained:

> *Well, I think an app is a good possibility because it really might not lock you down to a certain time that you had to be somewhere to communicate and associate with someone.* [Online 1]

Participants also found the idea of a free app appealing, particularly if it could help manage their health.

Privacy and Safety Concerns

Privacy and confidentiality were top priorities for participants. They shared concerns about protecting their HIV status if using shared phones or accessing the app in public spaces. One participant stressed the following:

> *It is very important because you don't want everybody in your business. Nobody needs to know about your status unless you tell them other than you and your doctor.* [Online 2]

Another participant noted the importance of security features:

> *I think only women with HIV should get the invitation code because we want to be in a safe space because not everybody that would, if it was just open to*

XSL•FO
RenderX

> *anybody, then people would just know people's business. And some people are not comfortable.* [Online 1]

In addition, adherence to privacy regulations was emphasized:

> *Any app or anything that we utilize, we have to be assured of confidentiality. There's HIPAA rules, there's everything.* [Online 1]

Category 2: Acceptability of Message Content

Education and Resources

Participants saw the app as a valuable tool for providing education and resources, including information on medications, side effects, and community support. One participant noted the following:

> *I like that part too about having resources because a lot of times people miss out on stuff because they're not aware of the resources available to them.* [Online 1]

Another participant suggested including information on medication, especially as a way to support women in rural areas:

> *Also have something about the medications as well. What are the side effects? Can you take it with blood pressure medicine? Is there an interaction with it? And I think all that because there are people that are in the rural areas that may not be able to physically get to it.* [Online 1]

Customizable Features

Participants expressed a desire for customization options, including the ability to adjust notification frequency and tailor content to individual needs. One participant envisioned how "folks can add themselves to each forum so that they can get notified about that forum, and then the ones that they don't want to get notified about, they don't" (Online 1). Translation features were also suggested:

> *I would like something like where you could translate Spanish. I would like to be able to have conversations with someone who...English is not their first language and be able to communicate with them just as well as I can communicate with everybody else.* [Online 1]

Category 3: Personal Relevance of Messaging

Inspirational Positive Affirmations and Supportive Messaging

Participants appreciated messages that provided encouragement and positivity. One participant shared the following:

> *I am a big person on positive affirmations...they can help you get through some very challenging times.* [In person]

Participants appreciated messages promoting self-love, such as "Just love you" and "Priority one is you"—all language drawn directly from the qualitative data and the language used by participants in phase 1 of the study. Another participant explained how these messages resonated with her, stating that they were "something you can repeat to yourself like you standing in the mirror or not even in the mirror, but something that you can repeat to yourself to encourage yourself" (Online 2).

Spiritual Messages

The incorporation of spiritual messages was seen as a valuable component to many women. One participant stated the following:

> *Having a higher power is what really got through when I was first newly diagnosed.* [Online 1]

Other participants were receptive to messages that nurtured spirituality provided that religious language was inclusive and respectful. One participant remarked the following:

> *It could be the ground that is your god or that tree. It's whatever you can call your higher power, whatever you choose it to be.* [Online 2]

Stigma-Related Messaging

Messages about overcoming stigma resonated with participants, emphasizing the importance of addressing discrimination from health care providers and others. One participant noted the following:

> *I think at one point all of us have incurred some type of judgment or stigma, so I think it's important especially for those who are battling, still not telling people or...I don't know, coming out the closet or however you want to say it or don't want people to know.* [Online 2]

Another participant shared how these messages addressing HIV-related stigma could counteract difficult times:

> *If I'm finna give up or stop taking my medication and I kept seeing these messages pop up...it would have me think it's more important to continue my care for me than to give up on myself.* [In person]

Category 4: Comprehensibility of Messaging

Clarity of Language

There was a focus on using clear, easy-to-understand language. Participants advised against using words such as "undetectable" or "ART" without explanation. A participant explained the following:

> *Everyone doesn't know all of the words. Everyone doesn't know, you know, new people...I mean, I don't think we should assume that people know.* [Online 2]

Participants also suggested hotlinks for complex words, ensuring that everyone, especially newly diagnosed individuals, understood the content. A participant explained the following:

> *I guess for me what this looks like, there should be links connected to this where they can go to find out information or more about it. [Online 1]*

Another stated the following:

> *So they could click on it and maybe get a definition or understand what that mean.* [Online 1]

XSL•FO
RenderX

Visual Preferences and Layout

Feedback on the health messaging and visual designs included suggestions for larger font sizes and shorter messages. One participant suggested the following:

> *...as far as the font, maybe a bit small, maybe a little bit too wordy.* [Online 1]

Another remarked the following:

> *Yeah, I see it. It's kind of small. It could be a little bit bigger.* [Online 2]

Discussion

Principal Results

This study identified 4 primary categories shaping women's perspectives on the development of an mHealth intervention to support HIV care engagement and adherence. First, participants emphasized the acceptability of the app's features and functionality, noting the importance of privacy, accessibility, and customizable options. Second, they highlighted the acceptability of message content, expressing strong preferences for affirming and uplifting language and images that conveyed resilience. Third, women underscored the personal relevance of messaging, stressing the value of content that felt tailored to their lived experiences and circumstances, including the importance of spirituality and faith as sources of strength and motivation. Finally, participants described the influence of intersectional stigma and contextual barriers, including structural inequities and compounded stigma, on their engagement with care and technology. Together, these findings provide a comprehensive framework for understanding how mHealth interventions can be both acceptable and meaningful for women living with HIV in the South.

We now turn to themes that warrant deeper consideration, with message comprehensibility emerging as a central point. Women consistently expressed a preference for clear, simple messaging, a finding that aligns with current National Institutes of Health recommendations for effective health communication [55], as well as SMS text messaging–based interventions for improved HIV clinical outcomes [39] and adherence [26]. Condensing findings from multistage participatory research into concise statements is a challenging but essential process to balance clarity and brevity. This refinement process ensures that only the most essential information is conveyed without unnecessary wording or visuals. This preference for succinct messaging is also supported in contemporary research that notes a shift toward shorter messaging and video content [56]. In addition, participants' suggestions to define or hyperlink medical terminology were valuable, reinforcing best practices in digital health literacy, which emphasize the importance of accessible explanations of medical terms to enhance engagement and adherence [57]. Participants also suggested language translation, which will be a priority in future phases of development.

Privacy and security concerns also emerged as a critical consideration, with participants stressing the importance of confidentiality in protecting their HIV status. This emphasis echoes research identifying confidentiality as a primary barrier to mHealth adoption among individuals with stigmatized health conditions [58]. Similarly, Mulawa et al [24] note that, while mHealth tools may provide private avenues for engagement, ongoing fears of unintentional disclosure can hinder uptake, underscoring the importance of embedding strong security and confidentiality safeguards into intervention design. Interpreted through social location, women's concerns about privacy reflect how intersecting experiences of HIV stigma, racism, and gendered inequities in the South heighten the risks of disclosure, making confidentiality protections central to engagement. Some participants envisioned the capacity for linking directly to individuals or group formats through social media. Green [59] found that people living with HIV were resistant to health messaging delivered through social media, viewing it as a platform that perpetuates offline inequities and promotes content that does not reflect participants' experiences, underscoring the need to incorporate users in all stages of development. Due to the constraints of the technology, stated privacy concerns, and current literature, we are hesitant to suggest linked social media platforms as a feature for this app.

Another notable preference was for uplifting, positive messages and images, highlighting the critical role of emotional support for consistent engagement and adherence [60]. Interpreted through the lens of social location, the desire for affirming content illustrates how women living with HIV draw on encouragement as a counterbalance to the layered stigma and inequities shaping their daily lives. While practical features such as reminders, education, and resources are essential for successful care engagement and medication adherence, the women's strong desire for encouragement highlights the importance of positive affirmations in fostering resilience and sustained engagement. This preference also aligns with research emphasizing the role of intrinsic motivation beyond scheduled reminders to sustain viral suppression [61]. In addition, focus group participants expressed interest in the ability to tailor the frequency of motivational SMS text messages to receive additional supportive content on more challenging days. For example, Schnall et al [20] showed that a self-management app with tailored features improved engagement and adherence, whereas Dworkin et al [29] demonstrated that personalized, conversational support delivered through a mobile intervention was both feasible and motivating. Together, these findings reinforce broader reviews that show that tailored notifications increase long-term retention [62]. This customization may be a key factor for sustained engagement.

The unique emphasis on spirituality in the messaging, derived from previous stages of analysis, highlights an area where existing mHealth research remains limited. Some studies have begun to highlight its potential relevance, showing, for example, that incorporating spirituality into self-management tools increases adoption and engagement [40,63]. More recently, Reid et al [64] documented that religious and spiritual coping remains a central support for Black women living with HIV, particularly older women in Southern contexts. However, reviews of HIV mHealth interventions continue to emphasize that spirituality remains insufficiently integrated [65]. Through the lens of social location, the centrality of spirituality reflects how regional culture; faith traditions; and the intersection of race, gender, and HIV status shape coping strategies and inform

XSL•FO
RenderX

preferences for mHealth design. Our findings extend this work by demonstrating that, for women living with HIV in the South, spirituality is not peripheral but a central motivator for engagement, underscoring the importance of regional and psychographic targeting in designing effective interventions.

Limitations

This study has several limitations. Although more than sufficient for concept testing, the sample size was relatively small and may not be fully representative of the broader population of women living with HIV in the region. Despite efforts to recruit women with diverse life experiences, women who are disengaged from care were not represented; their inclusion remains an essential priority for future research. There is also potential for selection bias as participants may differ systematically from those who did not enroll, particularly in their comfort with technology or engagement in care. Limited funding prevented translation in the data collection process, and the researchers were unable to recruit Hispanic or Latina women who speak Spanish as a primary language, excluding a critical demographic from this study [66]. Correspondingly, due to the specific demographics of this study, findings are not generalizable to all people living with HIV. As this study relied on focus groups to elicit perceptions of a proposed app, the findings speak to anticipated acceptability rather than demonstrated use and, thus, cannot address long-term engagement or sustained outcomes. Long-term effectiveness in improving adherence and health outcomes will require further longitudinal investigation. In addition, although an mHealth app is meant to be accessible, digital literacy, the aforementioned translation barriers, and access to reliable internet services remain potential barriers for some users, warranting further exploration of strategies to improve accessibility. Finally, because participants discussed their anticipated responses to a proposed app rather than reflecting on actual use, the findings may not accurately capture how women would engage with the technology in practice.

Conclusions

In conclusion, this study provides robust preliminary findings to support the development of an mHealth intervention tailored to the unique needs of women living with HIV in the South. Participants emphasized the importance of engaging app features, privacy protections, and customizable options; the personal relevance of uplifting and affirming content; and the comprehensibility of clear, concise messaging. This study also identified a gap in the integration of spirituality into mHealth messaging, suggesting a potential area for future exploration. The participatory approach used in this research ensures that the proposed app reflects the lived experiences and preferences of its intended users. Future work should focus on expanding accessibility, addressing language barriers, and evaluating long-term adherence outcomes to maximize the app's impact on health care engagement and viral suppression among women living with HIV.

Acknowledgments

The authors wish to sincerely thank the women who participated in this study and openly shared their experiences with living with HIV, care engagement, and medication adherence. The authors also wish to acknowledge and thank the members of the community and clinician advisory board for their guidance with the data collection and analysis processes. Funding was provided by the National Institute of Nursing Research, National Institutes of Health (1R21NR020164-01).

Authors' Contributions

JVB, SBB, and CLC contributed to conceptualization. All authors contributed to data curation. SBS, JVB, SBB, AMS, KJS, CCL, and CEC contributed to formal analysis. CEC contributed to funding acquisition. SBS, JVB, MRC, and CEC contributed to investigation. SBB, AMS, KJS, CCL, and CEC contributed to methodology. JVB and CEC contributed to project administration. JVB and CEC contributed to supervision. JVB, SBB, and CEC contributed to validation. SBS, SBB, AMS, KJS, CCL, and CEC contributed to visualization. SBS, JVB, MRC, and CEC contributed to writing the original draft. All authors contributed to manuscript review and editing.

Conflicts of Interest

None declared.

References

1. Sommer SB, Barroso JV, Bass SB, Congema MR, Schoemann AM, Caiola CE. Barriers and facilitators to engagement in care and medication adherence for women living with HIV in the Southern United States. AIDS Care 2024 Jan;36(1):130-138. [doi: 10.1080/09540121.2023.2233498] [Medline: 37535630]
2. Understanding the HIV care continuum. Centers for Disease Control and Prevention. 2019. URL: https://www.cdc.gov/hiv/pdf/library/factsheets/cdc-hiv-care-continuum.pdf [accessed 2025-10-07]
3. Funding opportunity announcement: PS22-2201: list of states in South census region. Centers for Disease Control and Prevention. 2021. URL: https://www.cdc.gov/hiv/funding/announcements/ps22-2201/attachments/south-cesus-region.html [accessed 2025-10-07]

XSL•FO
RenderX

4. Nwangwu-Ike N, Jin C, Gant Z, Johnson S, Balaji AB. An examination of geographic differences in social determinants of health among women with diagnosed HIV in the United States and Puerto Rico, 2017. Open AIDS J 2021 Feb 16;15(1):10-20. [doi: 10.2174/1874613602115010010]
5. Wright IA, Reid R, Shahid N, et al. Neighborhood characteristics, intersectional discrimination, mental health, and HIV outcomes among black women living with HIV, Southeastern United States, 2019 2020. Am J Public Health 2022 Jun;112(S4):S433-S443. [doi: 10.2105/AJPH.2021.306675] [Medline: 35763751]
6. Relf MV, Pan W, Edmonds A, Ramirez C, Amarasekara S, Adimora AA. Discrimination, medical distrust, stigma, depressive symptoms, antiretroviral medication adherence, engagement in care, and quality of life among women living with HIV in North Carolina: a mediated structural equation model. J Acquir Immune Defic Syndr 2019 Jul 1;81(3):328-335. [doi: 10.1097/QAI.0000000000002033] [Medline: 30893124]
7. Kalichman S, Shkembi B, Hernandez D, Katner H, Thorson KR. Income inequality, HIV stigma, and preventing HIV disease progression in rural communities. Prev Sci 2019 Oct;20(7):1066-1073. [doi: 10.1007/s11121-019-01013-5] [Medline: 30955136]
8. Kay ES, Edmonds A, Ludema C, et al. Health insurance and AIDS Drug Assistance Program (ADAP) increases retention in care among women living with HIV in the United States. AIDS Care 2021 Aug;33(8):1044-1051. [doi: 10.1080/09540121.2020.1849529] [Medline: 33233937]
9. Ludema C, Edmonds A, Cole SR, et al. Comparing neighborhood and state contexts for women living with and without HIV: understanding the Southern HIV epidemic. AIDS Care 2018 Nov 2;30(11):1360-1367. [doi: 10.1080/09540121.2018.1492696]
10. Elliot LM, Booth MM, Patterson G, Althoff M, Bush CK, Dery MA. Association of state-mandated abstinence-only sexuality education with rates of adolescent HIV infection and teenage pregnancy. J La State Med Soc 2017;169(2):56 [FREE Full text] [Medline: 28414685]
11. Watson M, Johnson SD, Zhang T, Oster AM. Characteristics of and trends in HIV diagnoses in the deep south region of the United States, 2012-2017. AIDS Behav 2019 Oct;23(Suppl 3):224-232. [doi: 10.1007/s10461-019-02659-6] [Medline: 31473847]
12. Hoover KW, Hu X, Porter SE, et al. HIV diagnoses and the HIV care continuum among women and girls aged ≥13 years-39 states and the district of Columbia, 2015-2016. J Acquir Immune Defic Syndr 2019 Jul 1;81(3):251-256. [doi: 10.1097/QAI.0000000000002023] [Medline: 30839379]
13. Reif SS, Cooper H, Warren M, Wilson E. HIV in the US Deep South: trends from 2008-2019. Duke Global Health Institute: Center for Health Policy and Inequalities Research. 2021. URL: https://southernaidsstrategy.org/wp-content/uploads/2019/06/sasi-report-hiv-aids-in-the-u.s.-deep-south-trends-from-2008-2016-final.pdf [accessed 2025-10-07]
14. Nwangwu-Ike N, Saduvala N, Watson M, Panneer N, Oster AM. HIV diagnoses and viral suppression among US women in rural and nonrural areas, 2010-2017. J Rural Health 2020 Mar;36(2):217-223. [doi: 10.1111/jrh.12384] [Medline: 31233645]
15. Fatima K, Siddiqui A, Malik GMA, et al. Demographical and regional trends of HIV-related mortality among young adults in the United States from 1999 to 2020. Am J Med 2025 Feb;138(2):220-227. [doi: 10.1016/j.amjmed.2024.09.026] [Medline: 39349154]
16. Caiola C, Barroso J, Docherty SL. Capturing the social location of African American mothers living with HIV: an inquiry into how social determinants of health are framed. Nurs Res 2017;66(3):209-221. [doi: 10.1097/NNR.0000000000000213] [Medline: 28252555]
17. Caiola C, Choufani MR, André J, et al. Social location and decision-making among women living with HIV in the Southern United States: an intersectional approach. Int J Environ Res Public Health 2024 Nov 27;21(12):1575. [doi: 10.3390/ijerph21121575] [Medline: 39767417]
18. Gordon TF, Bass SB, Ruzek SB, et al. Developing a typology of African Americans with limited literacy based on preventive health practice orientation: implications for colorectal cancer screening strategies. J Health Commun 2014;19(11):1259-1277. [doi: 10.1080/10810730.2013.872725] [Medline: 24673248]
19. Mehraeen E, SeyedAlinaghi S, Pashaei Z, et al. Mobile applications in HIV self-management: a systematic review of scientific literature. AIDS Rev 2022 Mar 1;24(1):24-31. [doi: 10.24875/AIDSRev.21000025] [Medline: 34723447]
20. Schnall R, Sanabria G, Jia H, et al. Efficacy of an mHealth self-management intervention for persons living with HIV: the WiseApp randomized clinical trial. J Am Med Inform Assoc 2023 Feb 16;30(3):418-426. [doi: 10.1093/jamia/ocac233] [Medline: 36469808]
21. Mobile fact sheet. Pew Research Center. 2024 Jan. URL: https://www.pewresearch.org/internet/fact-sheet/mobile/?tabItem=673f63a2-da02-4836-96c6-c1d4e25c8a89 [accessed 2025-10-07]
22. Ryan white HIV/AIDS program annual data report. Health Resources and Services Administration. 2024. URL: https://ryanwhite.hrsa.gov/sites/default/files/ryanwhite/resources/population-factsheet-overview.pdf [accessed 2025-10-07]
23. Wang H, Ho AF, Wiener RC, Sambamoorthi U. The association of mobile health applications with self-management behaviors among adults with chronic conditions in the United States. Int J Environ Res Public Health 2021 Sep 30;18(19):10351. [doi: 10.3390/ijerph181910351] [Medline: 34639651]

XSL•FO
RenderX

24. Mulawa MI, Rosengren AL, Amico KR, Hightow-Weidman LB, Muessig KE. mHealth to reduce HIV-related stigma among youth in the United States: a scoping review. Mhealth 2021;7(35):35-35. [doi: 10.21037/mhealth-20-68]
25. Puig J, Echeverría P, Lluch T, et al. A specific mobile health application for older HIV-infected patients: usability and patient's satisfaction. Telemed J E Health 2021 Apr;27(4):432-440. [doi: 10.1089/tmj.2020.0098] [Medline: 32667858]
26. Pagan-Ortiz ME, Goulet P, Kogelman L, Levkoff SE, Weitzman PF. Feasibility of a texting intervention to improve medication adherence among older HIV+ African Americans: a mixed-method pilot study. Gerontol Geriatr Med 2019;5:2333721419855662. [doi: 10.1177/2333721419855662] [Medline: 31276016]
27. Reback CJ, Fletcher JB, Kisler KA. Text messaging improves HIV care continuum outcomes among young adult trans women living with HIV: text me, girl!. AIDS Behav 2021 Sep;25(9):3011-3023. [doi: 10.1007/s10461-021-03352-3] [Medline: 34164763]
28. Pitpitan EV, Horvath KJ, Aldous J, et al. Peers plus mobile app for treatment in HIV (PATH): protocol for a randomized controlled trial to test a community-based integrated peer support and mHealth intervention to improve viral suppression among Hispanic and Black people living with HIV. Trials 2024 Mar 22;25(1):212. [doi: 10.1186/s13063-024-08042-8] [Medline: 38520030]
29. Dworkin MS, Lee S, Chakraborty A, et al. Acceptability, feasibility, and preliminary efficacy of a theory-based relational embodied conversational agent mobile phone intervention to promote HIV medication adherence in young HIV-positive African American MSM. AIDS Educ Prev 2019 Feb;31(1):17-37. [doi: 10.1521/aeap.2019.31.1.17] [Medline: 30742481]
30. Plant A, Sparks P, Creech DN, et al. Developing an mHealth program to improve HIV care continuum outcomes among young Black gay and bisexual men. BMC Public Health 2024 May 7;24(1):1247. [doi: 10.1186/s12889-024-18652-1] [Medline: 38714973]
31. Goldhammer H, Marc LG, Psihopaidas D, et al. HIV care continuum interventions for transgender women: a topical review. Public Health Rep 2023;138(1):19-30. [doi: 10.1177/00333549211065517] [Medline: 35060802]
32. Andersson GZ, Reinius M, Eriksson LE, et al. Stigma reduction interventions in people living with HIV to improve health-related quality of life. Lancet HIV 2020 Feb;7(2):e129-e140. [doi: 10.1016/S2352-3018(19)30343-1] [Medline: 31776098]
33. Maestre JF, Zdziarska P, Min A, Baglione AN, Chung CF, Shih PC. Not another medication adherence app: critical reflections on addressing public HIV-related stigma through design. Proc ACM Hum Comput Interact 2021;4(CSCW3):1-28. [doi: 10.1145/3434171]
34. Rao D, Frey S, Ramaiya M. eHealth for stigma reduction efforts designed to improve engagement in care for people living with HIV. Curr HIV/AIDS Rep 2018 Dec;15(6):397-402. [doi: 10.1007/s11904-018-0414-z] [Medline: 30232579]
35. Barroso J, Madisetti M, Mueller M. A feasibility study to develop and test a cognitive behavioral stress management mobile health application for HIV-related fatigue. J Pain Symptom Manage 2020 Feb;59(2):242-253. [doi: 10.1016/j.jpainsymman.2019.09.009] [Medline: 31539601]
36. Harrison S, Li X. Rebooting resilience: shifts toward dynamic, multi-level, and technology-based approaches for people living with HIV. AIDS Care 2018 Aug;30(sup5):S1-S5. [doi: 10.1080/09540121.2018.1557592] [Medline: 30678476]
37. Skeen SJ, Tokarz S, Gasik RE, et al. A trauma-informed, geospatially aware, just-in-time adaptive mHealth intervention to support effective coping skills among people living with HIV in New Orleans: development and protocol for a pilot randomized controlled trial. JMIR Res Protoc 2023 Oct 24;12(1):e47151. [doi: 10.2196/47151] [Medline: 37874637]
38. Stockman JK, Anderson KM, Tsuyuki K, Horvath KJ. LinkPositively: a trauma-informed peer navigation and social networking webapp to improve HIV care among black women affected by interpersonal violence. J Health Care Poor Underserved 2021 May;32(2 Suppl):166-188. [doi: 10.1353/hpu.2021.0056] [Medline: 34934301]
39. Coleman J, Bohlin KC, Thorson A, et al. Effectiveness of an SMS-based maternal mHealth intervention to improve clinical outcomes of HIV-positive pregnant women. AIDS Care 2017 Jul;29(7):890-897. [doi: 10.1080/09540121.2017.1280126] [Medline: 28107805]
40. Njie-Carr VPS, Zhu S, Williams GC, Corless IB, Himelhoch S. Evaluation of a technology-enhanced intervention for older women with HIV infection: a proof of concept study. AIDS Care 2021 Aug;33(8):983-992. [doi: 10.1080/09540121.2020.1810617] [Medline: 32835493]
41. Fetters MD, Curry LA, Creswell JW. Achieving integration in mixed methods designs-principles and practices. Health Serv Res 2013 Dec;48(6 Pt 2):2134-2156. [doi: 10.1111/1475-6773.12117] [Medline: 24279835]
42. Bass SB, Jessop A, Maurer L, Gashat M, Al Hajji M, Gutierrez M. Mapping the barriers and facilitators of HCV treatment initiation in methadone maintenance therapy patients: implications for intervention development. J Health Commun 2018;23(1):117-127. [doi: 10.1080/10810730.2017.1414902] [Medline: 29252118]
43. Bass SB, Wolak C, Greener J, et al. Using perceptual mapping methods to understand gender differences in perceived barriers and benefits of clinical research participation in urban minority HIV+ patients. AIDS Care 2016;28(4):528-536. [doi: 10.1080/09540121.2015.1112352] [Medline: 26572215]
44. Bass SB, Kelly PJA, Brajuha J, et al. Using social marketing strategies to develop and pretest PrEP education materials for transgender women. J Soc Mark 2023 May 29;13(3):380-398. [doi: 10.1108/jsocm-08-2022-0169] [Medline: 40851887]

XSL•FO
RenderX

45. Bass SB, Kelly PJ, Hoadley A, Arroyo Lloret A, Organtini T. Mapping perceptual differences to understand COVID-19 beliefs in those with vaccine hesitancy. J Health Commun 2022 Jan 2;27(1):49-61. [doi: 10.1080/10810730.2022.2042627] [Medline: 35199628]
46. Bass SB, Hollin IL, Kelly PJ, et al. Perceptions of live donor kidney transplantation using segmentation analysis and perceptual mapping to understand differences by self-reported health status in people on dialysis. Front Commun 2022;6. [doi: 10.3389/fcomm.2021.785186]
47. Hou SHJ, Henry B, Drummond R, et al. Co-designing priority components of an mHealth intervention to enhance follow-up care in young adult survivors of childhood cancer and health care providers: qualitative descriptive study. JMIR Cancer 2025 Apr 25;11:e57834. [doi: 10.2196/57834] [Medline: 40279633]
48. Zarski AC, Sextl-Plötz T, Schmidt-Hantke J, Hess NS, Buntrock C. Participatory development and proof-of-concept of a dyadic-based, AI-driven, just-in-time adaptive intervention mechanism for preventing anxiety and depressive disorders via app: study protocol for a feasibility study. Front Public Health 2025;13:1626428. [doi: 10.3389/fpubh.2025.1626428] [Medline: 40800021]
49. Making health communication programs work. National Institutes of Health. 2004. URL: https://www.cancer.gov/publications/health-communication/pink-book.pdf [accessed 2025-10-07]
50. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005 Nov;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
51. Nowell LS, Norris JM, White DE, Moules NJ. Thematic analysis: striving to meet the trustworthiness criteria. Int J Qual Methods 2017;16(1):1-13. [doi: 10.1177/1609406917733847]
52. Guest G, Namey E, McKenna K. How many focus groups are enough? Building an evidence base for nonprobability sample sizes. Field methods 2017 Feb;29(1):3-22. [doi: 10.1177/1525822X16639015]
53. How the census bureau measures poverty. US Census Bureau. URL: https://www.census.gov/topics/income-poverty/poverty/guidance/poverty-measures.html [accessed 2025-10-07]
54. Am I rural? Rural Health Information Hub. 2024. URL: https://www.ruralhealthinfo.org/am-i-rural/report?lat=36.15212&lng=-86.86926&addr=37209%2C%20TN&exact=0 [accessed 2025-10-07]
55. Clear & simple. National Institutes of Health. 2021. URL: https://www.nih.gov/institutes-nih/nih-office-director/office-communications-public-liaison/clear-communication/clear-simple [accessed 2025-10-07]
56. Zhu Z, Liu S, Zhang R. Examining the persuasive effects of health communication in short videos: systematic review. J Med Internet Res 2023 Oct 13;25:e48508. [doi: 10.2196/48508] [Medline: 37831488]
57. Fitzpatrick PJ. Improving health literacy using the power of digital communications to achieve better health outcomes for patients and practitioners. Front Digit Health 2023;5(1264780):1264780. [doi: 10.3389/fdgth.2023.1264780] [Medline: 38046643]
58. Jacob C, Sezgin E, Sanchez-Vazquez A, Ivory C. Sociotechnical factors affecting patients' adoption of mobile health tools: systematic literature review and narrative synthesis. JMIR Mhealth Uhealth 2022 May 5;10(5):e36284. [doi: 10.2196/36284] [Medline: 35318189]
59. Green M. Resistance as participation: queer theory's applications for HIV health technology design. Tech Commun Q 2021 Oct 2;30(4):331-344. [doi: 10.1080/10572252.2020.1831615]
60. Fletcher FE, Sherwood NR, Rice WS, et al. Resilience and HIV treatment outcomes among women living with HIV in the United States: a mixed-methods analysis. AIDS Patient Care STDS 2020 Aug;34(8):356-366. [doi: 10.1089/apc.2019.0309] [Medline: 32757978]
61. Cluesman SR, Gwadz M, Freeman R, et al. Exploring behavioral intervention components for African American/Black and Latino persons living with HIV with non-suppressed HIV viral load in the United States: a qualitative study. Int J Equity Health 2023 Jan 31;22(1):22. [doi: 10.1186/s12939-023-01836-3] [Medline: 36717920]
62. Oakley-Girvan I, Yunis R, Longmire M, Ouillon JS. What works best to engage participants in mobile app interventions and e-Health: a scoping review. Telemed J E Health 2022 Jun;28(6):768-780. [doi: 10.1089/tmj.2021.0176] [Medline: 34637651]
63. Thomas-Purcell K, Ibe TA, Purcell D, Quinn G, Ownby R. Exploring spirituality and technology receptivity among a sample of older blacks to inform a tailored chronic disease self-management mHealth intervention. Patient Relat Outcome Meas 2020;11:195-207. [doi: 10.2147/PROM.S260949] [Medline: 33117016]
64. Reid RA, Robinson M, Juste RS, et al. Linear growth modeling of religious coping and social support among Black women living with HIV in the context of the COVID-19 pandemic. AIDS Care 2025 Jul 30;30(1–13):1-13. [doi: 10.1080/09540121.2025.2534537] [Medline: 40737307]
65. Cooper V, Clatworthy J, Whetham J, Consortium E. mHealth interventions to support self-management in HIV: a systematic review. Open AIDS J 2017;11(119–32):119-132. [doi: 10.2174/1874613601711010119] [Medline: 29290888]
66. The impact of HIV on Hispanic/Latino people in the United States. Kaiser Family Foundation. 2024. URL: https://www.kff.org/hivaids/fact-sheet/the-impact-of-hiv-on-hispanic-latino-people-in-the-united-states/#:~:text=The%20rate%20of%20new%20diagnoses,16%25%20between%202018%20and%202022 [accessed 2025-10-07]

XSL•FO
RenderX

Abbreviations

ART: antiretroviral therapy
CCAB: community and clinician advisory board
HIPAA: Health Insurance Portability and Accountability Act
mHealth: mobile health
REDCap: Research Electronic Data Capture

Edited by J Sarvestan; submitted 20.04.25; peer-reviewed by M Al-Mujtaba; revised version received 23.09.25; accepted 24.09.25; published 22.10.25.

Please cite as:
Sommer SB, Barroso JV, Bass SB, Choufani MR, Schoemann AM, Singley KJ, Luck CC, Caiola CE
Participatory Development and Concept Testing of mHealth Messaging to Support Care Engagement and Antiretroviral Therapy Adherence for Women Living With HIV in the Southern United States: Focus Group Study
JMIR Form Res 2025;9:e76286
URL: https://formative.jmir.org/2025/1/e76286
doi:10.2196/76286

© Sadie B Sommer, Julie V Barroso, Sarah B Bass, Marianne R Choufani, Alexander M Schoemann, Katie J Singley, Caseem C Luck, Courtney E Caiola. Originally published in JMIR Formative Research (https://formative.jmir.org), 22.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Evaluating Large Language Models for Sentiment Analysis and Hesitancy Analysis on Vaccine Posts From Social Media: Qualitative Study

Augustine Annan[1], PhD; Amanda L Eiden[2], MBA, MPH, PhD; Dong Wang[2], PhD; Jingcheng Du[1], PhD; Majid Rastegar-Mojarad[1], PhD; Varun Kumar Nomula[2], MS; Xiaoyan Wang[1], PhD

[1]IMO Health, Rosemont, IL, United States

[2]Merck & Co, Inc, 126 East Lincoln Avenue, Rahway, NJ, United States

Corresponding Author:
Amanda L Eiden, MBA, MPH, PhD
Merck & Co, Inc, 126 East Lincoln Avenue, Rahway, NJ, United States

Abstract

Background: In the digital age, social media has become a crucial platform for public discourse on diverse health-related topics, including vaccines. Efficient sentiment analysis and hesitancy detection are essential for understanding public opinions and concerns. Large language models (LLMs) offer advanced capabilities for processing complex linguistic patterns, potentially providing valuable insights into vaccine-related discourse.

Objective: This study aims to evaluate the performance of various LLMs in sentiment analysis and hesitancy detection related to vaccine discussions on social media and identify the most efficient, accurate, and cost-effective model for detecting vaccine-related public sentiment and hesitancy trends.

Methods: We used several LLMs—generative pretrained transformer (GPT-3.5), GPT-4, Claude-3 Sonnet, and Llama 2—to process and classify complex linguistic data related to human papillomavirus; measles, mumps, and rubella; and vaccines overall from X (formerly known as Twitter), Reddit, and YouTube. The models were tested across different learning paradigms: zero-shot, 1-shot, and few-shot to determine their adaptability and learning efficiency with varying amounts of training data. We evaluated the models' performance using accuracy, F_1-score, precision, and recall. In addition, we conducted a cost analysis focused on token usage to assess the computational efficiency of each approach.

Results: GPT-4 (F_1-score=0.85 and accuracy=0.83) outperformed GPT-3.5, Llama 2, and Claude-3 Sonnet across various metrics, regardless of the sentiment type or learning paradigm. Few-shot learning did not significantly enhance performance compared with the zero-shot paradigm. Moreover, the increased computational costs and token usage associated with few-shot learning did not justify its application, given the marginal improvement in model performance. The analysis highlighted challenges in classifying neutral sentiments and convenience, correctly interpreting sarcasm, and accurately identifying indirect expressions of vaccine hesitancy, emphasizing the need for model refinement.

Conclusions: GPT-4 emerged as the most accurate model, excelling in sentiment and hesitancy analysis. Performance differences between learning paradigms were minimal, making zero-shot learning preferable for its balance of accuracy and computational efficiency. However, the zero-shot GPT-4 model is not the most cost-effective compared with traditional machine learning. A hybrid approach, using LLMs for initial annotation and traditional models for training, could optimize cost and performance. Despite reliance on specific LLM versions and a limited focus on certain vaccine types and platforms, our findings underscore the capabilities and limitations of LLMs in vaccine sentiment and hesitancy analysis, highlighting the need for ongoing evaluation and adaptation in public health communication strategies.

(JMIR Form Res 2025;9:e64723) doi:10.2196/64723

KEYWORDS

vaccine sentiment; vaccine hesitancy; large language models; GPT4; social media platforms; public health communication; LLMs; NLP; machine learning; artificial intelligence; language models; sentiment analysis; hesitancy analysis; vaccine posts; social media; vaccine; public opinion; vaccine-related; public sentiment; computational efficiency

XSL•FO
RenderX

Introduction

In the era of digital communication, social media platforms have become central to the dissemination and exchange of public opinions on health-related topics, including vaccination. The vast and dynamic nature of these platforms offers a rich dataset for analyzing public sentiment and hesitancy toward vaccines, which is critical for developing effective health communication strategies. The advent of advanced artificial intelligence (AI) technologies, particularly large language models (LLMs) such as OpenAI's generative pretrained transformer (GPT-4.0) [1] and GPT-3.5 [2], Anthropic's Claude 3 Sonnet [3], and Meta's Llama 2 [4], represents a significant leap forward in our capacity to understand and interpret extensive amounts of unstructured text data [5]. These models have achieved impressive results in a variety of applications; for example, they significantly aid in diagnostics and personalizing treatment approaches and prove successful in medical licensing examinations and specialized medical tasks [6-12], thereby showcasing the expansive range of their capabilities [13-15].

The use of LLMs extends beyond traditional medical applications. These technologies are set to transform public health research by offering insights into public sentiments expressed on the web, particularly on complex topics such as vaccine hesitancy; defined by reluctance or delay in accepting vaccines despite the availability of vaccination services [16], vaccine hesitancy is a barrier to global health initiatives, especially in ongoing campaigns against diseases such as human papillomavirus (HPV) and measles, mumps, and rubella (MMR). Addressing this challenge requires a nuanced understanding of public opinions, concerns, and misinformation patterns, which LLMs are uniquely positioned to facilitate [17].

The rapid evolution of LLMs, from earlier models such as GPT-2 and BERT to more advanced iterations, has brought substantial improvements in language comprehension and task-solving abilities [1,2,18-20]. Other organizations, including Anthropic with its Claude models [3,21] and Meta AI with the Llama series [4,22], have also contributed significantly to this field.

Despite these advancements, research comparing different LLMs and learning paradigms in public health sentiment analysis remains limited. This study aims to address this gap by evaluating the efficacy of various LLMs in analyzing sentiments and hesitancy related to HPV, MMR, and general vaccines across 3 social media platforms: X (formerly Twitter), Reddit, and YouTube.

In this work, we explored the adaptability and effectiveness of LLMs in processing health-related sentiments by using zero-shot, 1-shot, and few-shot learning paradigms. A statistical comparison of performance across these paradigms reveals significant insights into optimizing model efficiency and resource allocation for large-scale sentiment analysis. Specifically, our analysis delves into the practical implications of selecting learning paradigms based on their computational costs and accuracy in detecting nuanced expressions of vaccine sentiment and hesitancy.

By pinpointing the optimal combinations of LLMs and learning paradigms for robust sentiment and hesitancy analysis, our research directly informs the strategic deployment of AI in crafting targeted public health messaging. This endeavor not only enhances our understanding of public discourse around vaccines but also provides a foundation for addressing vaccine hesitancy more effectively, thereby contributing to the advancement of public health communication strategies.

Methods

Data Collection and Annotation

We analyzed English language social media posts from X (formerly Twitter), Reddit, and YouTube, focusing on public sentiment toward general vaccines, HPV, and MMR vaccines. Data collection spanned from January 1, 2011, to October 31, 2021. We selected these platforms because they provide robust application programming interface (API) access and represent a significant share of vaccine-related discourse on the web. English language posts were prioritized due to their widespread usage in large markets and accessibility. The detailed data collection and annotation procedures follow the methodology described in a previous study [23]. Data were retrieved using the APIs provided by the platforms, following their data privacy and ethical guidelines.

To ensure the relevance and quality of the data, we tailored our search queries for each platform, as variations in text format and query logic required customization. Inclusion keywords such as "vaccine," "immunization," "HPV vaccine," and "MMR vaccine" were used, while exclusion keywords (eg, "software updates" or "sports immunities") filtered out irrelevant content. Duplicate posts and spam were removed during preprocessing.

The data collection period includes the COVID-19 pandemic, a significant global event that likely influenced public sentiment and hesitancy discussions on vaccines. With the widespread availability of COVID-19 vaccines during this time, public discourse around vaccines may have become more polarized, amplifying safety concerns and misinformation. Although this study does not focus on COVID-19–specific sentiments, the pandemic era context may have shaped the overall tone and topics of vaccine-related discussions. Posts unrelated to COVID-19 vaccines were also included to capture general vaccine sentiment trends, ensuring that the analysis remained relevant to MMR, HPV, and general vaccines.

Our study involved a dual-layered annotation approach for 10,485 social media posts. Posts were first categorized for sentiment (positive, neutral, or negative) and then for vaccine hesitancy. The hesitancy annotation involved a 2-step process: initially determining whether a post was hesitant or nonhesitant, and for those identified as hesitant, further categorizing them based on the World Health Organization (WHO's 3Cs [confidence, complacency, and convenience] model) [24].

To ensure balance, we evenly annotated 1165 posts across 3 platforms (X, Reddit, and YouTube) and vaccine types (general, HPV, and MMR). The final dataset comprised 36% positive, 26% neutral, and 38% negative sentiments, with 39% of posts identified as hesitant. Among hesitant posts, 58% reflected

XSL•FO
RenderX

confidence issues, 25% complacency, and 17% convenience concerns.

Posts were annotated for sentiment as positive (favorable opinions or experiences regarding vaccines), neutral (informational or mixed sentiment posts), or negative (adverse opinions or potentially deterring information). Hesitancy was annotated based on the WHO's 3Cs model: confidence (distrust in vaccine efficacy, safety, or delivery systems), complacency (low perceived risk of vaccine-preventable diseases), and convenience (barriers to vaccine accessibility or difficulties in obtaining vaccination) [23].

Few-Shot Learning and Evaluation

For evaluation, 7515 annotated posts were used to assess the LLMs' performance. Table 1 provides the distribution of these posts across sentiment and hesitancy categories, proportional to the overall annotated dataset. A detailed distribution of annotated posts in each sentiment and 3Cs construct for each platform and vaccine topic group is shown in Table S1 in Multimedia Appendix 1.

Table . Distribution of evaluation posts by sentiment and hesitancy categories.

Category	Count of posts	Percentage
Sentiment		
Positive	2705	36
Neutral	1954	26
Negative	2856	38
Hesitancy		
Nonhesitant	4584	61
Hesitant	2931	39
Confidence	1850	63% of hesitant
Complacency	1020	35% of hesitant
Convenience	720	25% of hesitant

Few-shot learning experiments were conducted separately for each vaccine type and each platform. For each experiment, posts were selected using a balanced sampling approach to ensure balanced representation across sentiment categories (positive, neutral, and negative). For each configuration (eg, 5-shot), we sampled an equal number of examples per sentiment category. Specifically, for 5-shot learning, we included 5 examples from each sentiment category (positive, neutral, and negative) per vaccine type per platform. This resulted in a total of 15 examples per vaccine type per platform for 5-shot learning. The same sampling approach was applied to other configurations, ensuring consistency and diversity in training examples across platforms and vaccine topics.

A similar methodology was used for hesitancy analysis. Posts were sampled to represent hesitant and nonhesitant categories as well as the WHO's 3Cs constructs (confidence, complacency, and convenience), ensuring balanced representation across platforms and vaccine types.

To maintain consistency and enable fair comparisons, the same few-shot examples were reused across experiments for a given configuration, minimizing variability in training data and ensuring a controlled evaluation of model performance.

Sentiment Annotation

Sentiment annotation was a multiclass classification task, assigning 1 of 3 sentiment labels to each post:

- Positive: Posts that mention, report, or share positive opinions, news, or experiences about vaccines or vaccination.
  - *Example*: "Getting vaccinated is a vital step to safeguard your health and future."
- Neutral: Posts that relate to vaccines or vaccination topics but contain no clear sentiment, express mixed sentiments, or do not explicitly advocate for or against vaccination.
  - *Example*: "HPV has multiple strains, and vaccines cover some of them but not all."
- Negative: Posts that mention, report, or share negative opinions, news, or experiences about vaccines or vaccination, which may discourage vaccination.
  - *Example*: "The government is pushing vaccines despite the growing number of adverse reactions."

3Cs Vaccine Hesitancy Annotation

The annotation of vaccine hesitancy was based on the WHO's 3Cs model, with each construct evaluated independently. Posts labeled as "Lack of confidence," "Complacent," or "Inconvenient" were considered hesitant, while posts without any of these constructs were labeled as nonhesitant. Definitions and an example for each construct are as follows:

- Lack of confidence: Posts reflecting mistrust in vaccine efficacy, safety, vaccine delivery system, or motivations of policy makers.
  - *Example*: "Vaccinated individuals are still catching the disease, so how effective are these shots really?"
- Complacency: Posts where the perceived risks of vaccine-preventable diseases are low and vaccination is deemed unnecessary.

XSL•FO
RenderX

- *Example*: "I never got vaccinated as a kid and I turned out to be fine, so vaccines aren't critical or essential."

- Inconvenience: Posts highlighting physical, geographical, financial, or systemic barriers to vaccination, including issues related to health literacy and service accessibility.
  - *Example*: "I wanted to get the vaccine, but the nearest clinic is three hours away, and I cannot take time off work."

The annotation was performed by trained annotators with a medical background. All annotators underwent training with a guideline developed for this study. Annotators achieved a high interannotator agreement score (Cohen κ=0.93), indicating strong reliability. Discrepancies were resolved through discussion and consensus. Examples of annotated posts and the full annotation framework are detailed in a previous paper [23].

Implementation of LLMs

In this study, we leveraged 4 advanced LLMs: OpenAI's GPT-3.5 and GPT-4, Anthropic's Claude 3 Sonnet, and Meta's Llama 2. These models were chosen for their effectiveness in natural language processing (NLP) tasks such as text generation, sentiment analysis, and question answering, making them ideal for analyzing sentiments and hesitancy in social media discourse about vaccination.

GPT-3.5, with its 175 billion parameters, offers substantial text generation and understanding capabilities [25]. GPT-4 builds on this foundation with enhanced nuanced text interpretation and more sophisticated training techniques, making it particularly effective for detailed analysis of complex topics such as vaccine sentiment in social media conversations [26]. Claude 3 Sonnet was selected for its advanced processing capabilities, improved steerability, and the ability to handle up to 200,000 tokens, making it particularly suitable for comprehensive sentiment analysis [27]. Meta's Llama 2 was chosen for its open-source nature, flexibility, and robust performance in detecting nuanced sentiments and hesitancies, providing a strategic advantage for academic research.

Model Evaluation Through Learning Paradigms

We evaluated the LLMs using 3 learning paradigms—zero-shot, 1-shot, and few-shot learning—to test their adaptability and efficiency with varying amounts of prior information, simulating real-world scenarios where labeled data may be scarce.

Zero-Shot Learning

In the zero-shot learning paradigm, the model analyzes tasks without any prior examples or specific training related to the task [28]. It relies entirely on its pretrained knowledge to infer the correct output. For our study, models classified vaccine sentiment or hesitancy from social media posts using only a carefully crafted prompt, without any sample posts or classifications. This setup tests the model's ability to generalize from unrelated training data to new, unseen tasks.

One-Shot Learning

In 1-shot learning, the model is provided with a single example post for each category before making predictions. This minimal context helps the model understand the task with just 1 reference point per sentiment class or hesitancy construct.

Few-Shot Learning

Few-shot learning involves presenting the model with a set of multiple examples (*k* examples) to enhance its understanding and classification accuracy [29,30]. This methodology includes variations such as 5-shot and 10-shot settings. For sentiment analysis, *k* examples were provided for each category (positive, neutral, and negative). For hesitancy analysis, *k* examples were provided for each of the WHO's 3Cs model dimensions (confidence, complacency, and convenience). The few-shot learning approach allowed us to explore how models perform with an increasing number of examples, testing their precision in detecting and classifying sentiments and hesitancies with minimal and slightly more information.

This structured approach confirms the flexibility and effectiveness of the models in adapting to different levels of available information. It provides a comprehensive perspective on their use in analyzing a broad spectrum of public opinions. By evaluating these models in scenarios that mirror real-world data availability, we aim to demonstrate the potential of LLMs to swiftly and accurately analyze complex vaccine-related social media discourse, potentially supporting enhanced public health communication strategies.

Prompt Schema Design

The prompts used in this study were carefully crafted to align with the definitions and instructions provided to human annotators for vaccine sentiment and hesitancy terms. This approach, known as prompt engineering [31], ensured that the models' interpretation matched human understanding, thereby enhancing prediction relevance and accuracy.

Prompts included explicit definitions and criteria for categorizing sentiments (eg, "positive," "neutral," and "negative") and hesitancy constructs based on the WHO's 3Cs model (confidence, complacency, and convenience) [15]. For example, in analyzing HPV vaccination–related posts, the prompt specified that quoted content should take precedence in cases of mixed sentiment and provided guidance for handling ambiguous language (Figure 1). Prompts also included requests for the models to provide explanations for their classifications, which enhanced interpretability and supported iterative refinement during the development phase [32].

XSL•FO
RenderX

Figure 1. Prompt example for sentiment analysis. The text inside the dashed box is demonstrations of the few-shot setting and would be removed under the zero-shot setting. HPV: human papillomavirus; JSON: JavaScript Object Notation.

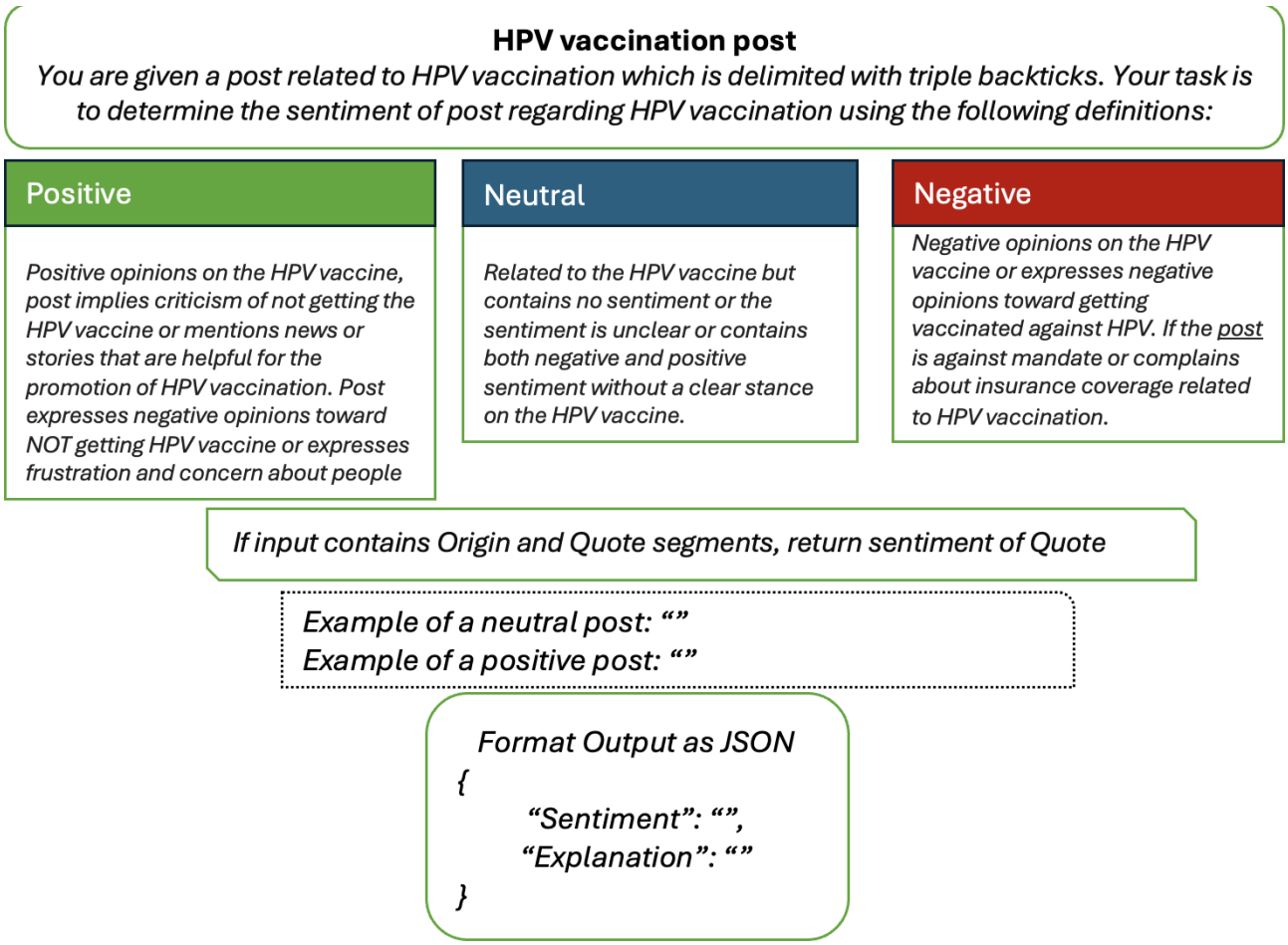

Preliminary testing compared multiple prompt versions to evaluate clarity, task alignment, and interpretability. For instance, variations in linguistic framing, the level of detail provided, and the structure of schema-specific instructions were tested using a subset of the dataset. Prompts that introduced ambiguity or failed to consistently elicit accurate classifications were refined or excluded. The final schema was selected based on its ability to achieve robust and consistent performance across platforms, vaccine types, and sentiment categories.

The final schema used in this study has been disclosed in the supplementary material for transparency and reproducibility (Multimedia Appendix 2). Future research could explore the impact of alternative prompt designs, including variations in linguistic framing, level of detail, and schema-specific instructions, on model performance. Such studies would be particularly valuable for tasks requiring the interpretation of complex or ambiguous language.

By incorporating these detailed and context-specific instructions into our prompts, we ensured that the LLMs were well equipped to interpret and analyze the sentiment of vaccine-related social media posts accurately [33]. This approach underscores the critical role of prompt engineering in leveraging the full capabilities of LLMs for nuanced tasks such as sentiment analysis, where the subtleties of language and expression necessitate a high degree of precision and clarity in the instructions provided to the models.

Hyperparameter Settings

We predefined 2 key hyperparameters to optimize the LLMs' performance: temperature and maximum tokens. The temperature parameter, which controls output randomness, was set to zero to ensure deterministic responses. This setting is crucial in the context of health-related sentiment analysis, where consistency and reproducibility are paramount. The maximum token (words and subwords) limit was set to 512, allowing models to provide detailed reasoning for their classifications, which is valuable for understanding the nuances of vaccine-related sentiments.

To ensure reliability, each social media post was processed through each LLM 3 times. We achieved a 97% consistency rate in categorizations across iterations. For the 3% of inconsistent cases, we used a majority vote approach to determine the final category. This rigorous process enhances the credibility of our results in the context of public health communication research.

Evaluation Metrics

To assess the effectiveness of the LLMs, within the framework of our study, we relied on a suite of established NLP evaluation metrics: accuracy, precision, recall, and the F_1-score [34]. These metrics collectively offer a comprehensive snapshot of the models' performance across various learning paradigms, ensuring a well-rounded assessment of their capabilities. These

XSL•FO
RenderX

metrics were calculated for each sentiment and hesitancy category. Moreover, to ensure the understanding of the models' performance, these evaluations were conducted separately for each vaccine type under discussion and across the various social media platforms included in our study.

Statistical Analysis for Model Comparison

To rigorously compare LLM performance across zero-shot, 1-shot, and few-shot learning paradigms, we conducted ANOVA tests [35]. We focused on F_1-scores and accuracy as key performance indicators. This analysis helps determine the most effective learning approach for sentiment and hesitancy analysis in vaccine-related social media discussions, a critical aspect of modern public health monitoring.

We set the significance level at $P<.05$ for identifying statistically significant differences. All statistical analyses were performed using R (version 4.2.2; R Core Team) providing a rigorous and replicable framework for our evaluations. This comprehensive evaluation framework allows us to assess the scalability and applicability of these AI models in real-world public health scenarios, potentially informing strategies for efficient, large-scale analysis of vaccine sentiments and hesitancy in diverse digital environments.

Ethical Considerations

All data used in this study were publicly available, and ethical guidelines for social media research were followed. Data privacy policies of the platforms (X, Reddit, and YouTube) were adhered to, and all posts were deidentified using unique random IDs. To further safeguard privacy, no actual posts are directly quoted or reproduced in this manuscript. Instead, illustrative examples provided in the text are entirely theoretical and were crafted to reflect the general themes and patterns observed in the data. These examples ensure that no identifiable user information is disclosed, while still providing clarity on the study's methodology and findings.

Results

Sentiment Analysis Overview

This section provides an overview of the sentiment analysis results, comparing the performance of the LLMs based on accuracy, F_1-score, precision, and recall. We focused on the models' abilities to classify sentiments (positive, neutral, and negative) in a zero-shot learning setting (Table 2).

Table . Large language models' performance (measured in accuracy, recall, precision, and F_1-score) on vaccine sentiment analysis.

Model	Sentiment	Accuracy (%)	F_1-score (%)	Precision (%)	Recall (%)
GPT-3.5	Negative	0.801	0.753	0.672	0.834
GPT-3.5	Neutral	0.821	0.633	0.612	0.659
GPT-3.5	Positive	0.819	0.781	0.711	0.861
GPT-4	Negative	0.940	0.932	0.910	0.950
GPT-4	Neutral	0.932	0.901	0.891	0.910
GPT-4	Positive	0.921	0.930	0.920	0.940
Claude-3s	Negative	0.882	0.801	0.841	0.760
Claude-3s	Neutral	0.821	0.690	0.701	0.681
Claude-3s	Positive	0.870	0.820	0.780	0.862
Llama 2	Negative	0.749	0.641	0.691	0.590
Llama 2	Neutral	0.719	0.140	0.110	0.221
Llama 2	Positive	0.721	0.421	0.411	0.418

GPT-4 exhibited the highest performance across all sentiment types, with accuracy ranging from 92% to 94%, F_1-scores from 90% to 93%, precision from 89% to 92%, and recall from 91% to 95%. Claude-3 followed, with accuracy between 82% and 88%, F_1-scores from 69% to 82%, precision from 70% to 84%, and recall from 68% to 86%. GPT-3.5 showed comparable performance with Claude-3, with accuracy ranging from 80% to 82%, F_1-scores from 63% to 78%, precision from 61% to 71%, and recall from 66% to 86%. Llama 2 exhibited the lowest performance, with accuracy between 72% and 75%, F_1-scores from 14% to 64%, precision from 11% to 69%, and recall from 22% to 59%.

Insights From Shot-Learning Paradigms

We evaluated the performance of the 4 LLMs across zero-shot, 1-shot, 5-shot, and 10-shot learning paradigms. The results represent the averaged performance across vaccine types (HPV, MMR, and general) and platforms (X, Reddit, and YouTube), ensuring a comprehensive evaluation. The models' accuracy and F_1-scores remained relatively consistent across different paradigms for each sentiment type (Figures 2 and 3). This consistency suggests that the models can effectively learn and generalize from few examples, maintaining their relative performance.

XSL•FO
RenderX

Figure 2. F_1-score performance of large language models in sentiment analysis across learning paradigms.

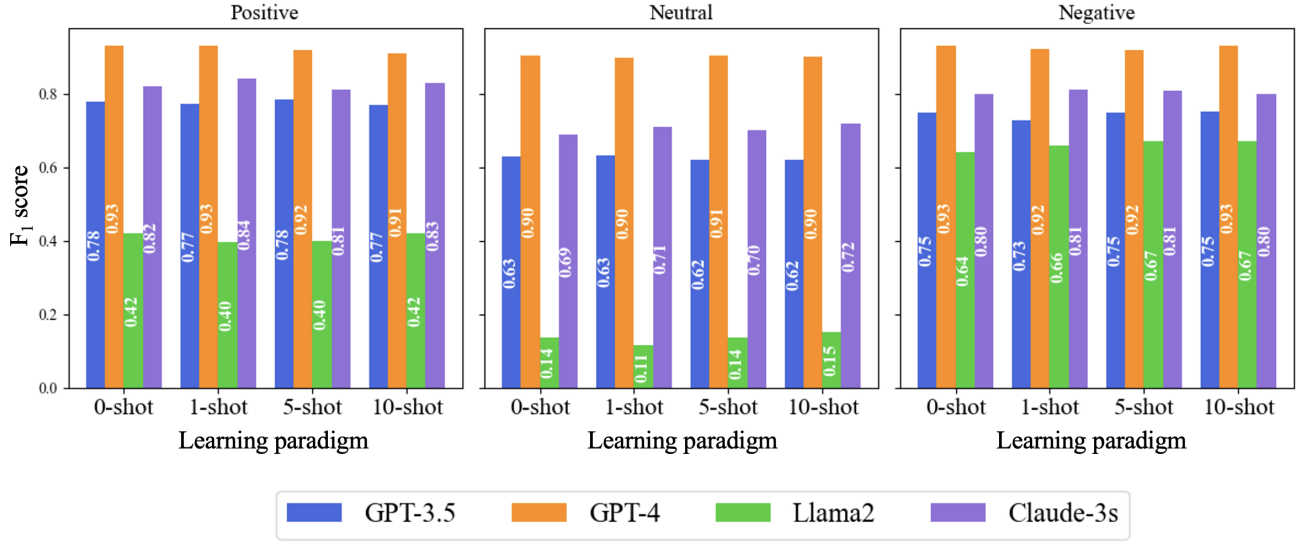

Figure 3. Accuracy performance of large language models in sentiment analysis across learning paradigms.

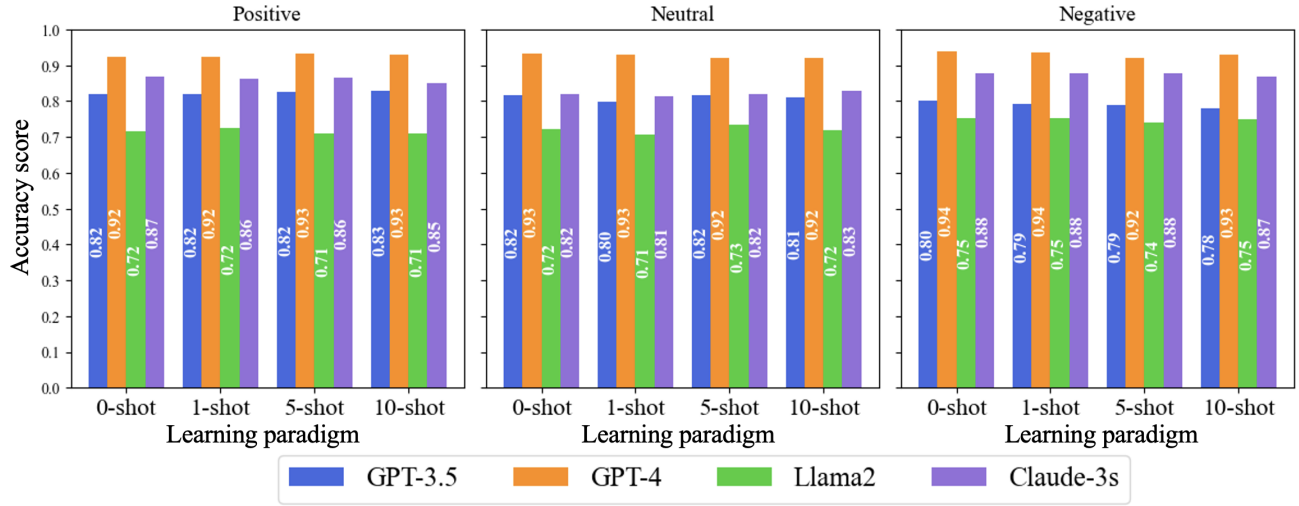

For GPT-4, accuracy scores for positive sentiment started at 92% in the zero-shot setting and increased marginally to 93.0% in the 10-shot setting. Neutral sentiment accuracy remained stable, ranging between 92.0% and 93%, while for negative sentiment, accuracy improved slightly from 92.0% in the 5-shot setting to 93.0% in the 10-shot setting. F_1-scores for GPT-4 reflected similar stability, with the highest values observed for negative sentiment, reaching 93% in both zero-shot and 10-shot settings.

GPT-3.5 demonstrated stable performance across paradigms. Accuracy for positive sentiment rose slightly from 82% in the zero-shot setting to 83% in the 10-shot setting, while neutral sentiment accuracy fluctuated around 80%. Negative sentiment accuracy, however, decreased from 80% in the zero-shot setting to 78% in the 10-shot setting. F_1-scores followed a similar pattern, with positive sentiment improving to 78% in the 5-shot setting before slightly declining to 77% in the 10-shot setting.

Claude-3s showed consistent performance across paradigms. Accuracy for neutral sentiment increased from 82% in the zero-shot setting to 83% in the 10-shot setting, while for positive sentiment, accuracy decreased slightly from 87% in the zero-shot setting to 85% in the 10-shot setting. F_1-scores for negative sentiment remained stable across all paradigms, with values around 80%.

Llama 2 exhibited modest improvements with additional examples. Accuracy for positive sentiment improved slightly from 71% in the zero-shot setting to 72% in the 1-shot setting, stabilizing at 71% in the 5- and 10-shot settings. F_1-scores for neutral sentiment remained low, reaching a maximum of 15% in the 10-shot setting. However, for negative sentiment, F_1-scores improved from 64% in the zero-shot setting to 67% in the 5- and 10-shot settings.

Overall, the results highlight GPT-4's superior performance in sentiment analysis across all paradigms, with consistent improvements observed as the number of examples increased. While GPT-3.5 and Claude-3s also exhibited stable and moderate performance, Llama2 showed the greatest need for further optimization despite modest gains in certain settings.

XSL•FO
RenderX

Performance Across Vaccine Types and Social Media Platforms in Zero-Shot Settings

GPT-4 demonstrated the highest F_1-scores, ranging from 87% to 91%, showcasing its strong capability to generalize and accurately classify sentiment across all vaccine categories. Claude-3 followed closely with F_1-scores between 78% and 82%, indicating its robust performance in the zero-shot setting. GPT-3.5 showed moderate performance with F_1-scores ranging from 65% to 85%. Llama 2 had the lowest scores, between 22% and 42%, highlighting the variability in performance among the models (Figure 4).

Figure 4. Zero-shot F_1-scores across models and vaccine types. HPV: human papillomavirus; MMR: measles, mumps, and rubella.

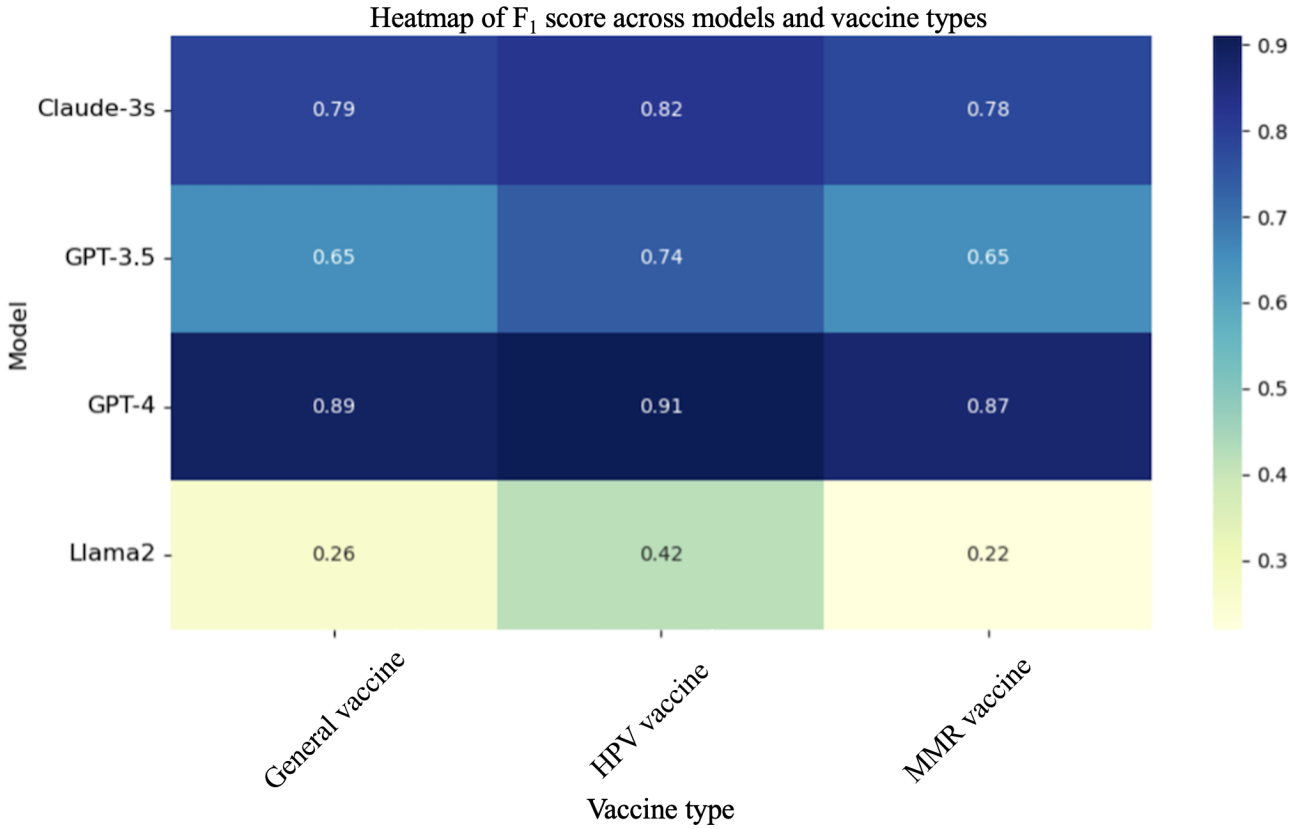

Furthermore, we present the performance of the 4 language models in sentiment analysis across 3 social media platforms: Reddit, X, and YouTube (Table 2). While the overall performance trend remains consistent across platforms, there are minor variations worth noting. GPT-3.5 exhibited a slightly higher F_1-score on YouTube (71.9%) than Reddit (71.4%) and X (70.6%). Claude-3 showed a higher F_1-score on YouTube (78.9%) than X (78.6%) and Reddit (77.5%). Llama 2 had a higher F_1-score on YouTube (41.5%) than X (39.9%) and Reddit (38.8%). GPT-4 maintained consistent performance across platforms with F_1-scores ranging from 91.7% to 91.8%.

Hesitancy Analysis Overview

GPT-4 led in both sentiment and hesitancy analyses, although its accuracy scores in hesitancy analysis (83%-96%) were slightly lower than in sentiment analysis (92%-94%; Table 3). This suggests that accurately identifying and classifying hesitancy types might be marginally more challenging than sentiment analysis, even for the best-performing model (Table 4). Claude-3 showed consistent performance across both tasks, with accuracy scores in hesitancy analysis (80%-87%) comparable with sentiment analysis (82%-88%). GPT-3.5 and Llama 2 exhibited slight improvements in accuracy for some hesitancy types compared with their sentiment analysis performance. For instance, GPT-3.5 achieved an accuracy of 84.9% for convenience hesitancy, higher than its sentiment analysis accuracy scores (80%-82%), while Llama 2's accuracy for confidence hesitancy (70%) was close to its performance in negative sentiment analysis (75%).

Table . Large language models' performance (measured in accuracy and F_1-score) across social media platforms on vaccine sentiment analysis.

Metric	Reddit				X, formerly known as Twitter				YouTube			
	GPT-3.5	GPT-4	Claude3s	Llama2	GPT-3.5	GPT-4	Claude3s	Llama2	GPT-3.5	GPT-4	Claude3s	Llama2
Accuracy	0.809	0.929	0.852	0.731	0.809	0.929	0.851	0.729	0.81	0.928	0.854	0.722
F_1-score	0.714	0.917	0.775	0.388	0.706	0.918	0.786	0.399	0.719	0.918	0.789	0.415

XSL•FO
RenderX

Table . Large Language models' performance (measured in accuracy, recall, precision, and F_1-score) on vaccine hesitancy analysis.

Metric and hesitancy type	GPT-3.5	GPT-4	Claude 3s	Llama 2
Accuracy				
Confidence	0.789	0.832	0.801	0.700
Convenience	0.849	0.961	0.871	0.729
Complacency	0.823	0.864	0.844	0.699
Hesitancy	0.762	0.878	0.844	0.668
Precision				
Confidence	0.855	0.957	0.796	0.614
Convenience	0.764	0.849	0.819	0.532
Complacency	0.621	0.696	0.679	0.286
Hesitancy	0.755	0.851	0.811	0.548
Recall				
Confidence	0.679	0.759	0.849	0.575
Convenience	0.799	0.80	0.790	0.492
Complacency	0.784	0.867	0.897	0.135
Hesitancy	0.677	0.856	0.849	0.479
F_1-score				
Confidence	0.757	0.847	0.757	0.594
Convenience	0.781	0.824	0.804	0.511
Complacency	0.693	0.772	0.773	0.183
Hesitancy	0.714	0.853	0.823	0.511

Impact of Learning Paradigms on Hesitancy Analysis

We explored the impact of zero-shot, 1-shot, 5-shot, and 10-shot learning paradigms on the models' hesitancy analysis capabilities, focusing on accuracy and F_1-scores. The results indicate that GPT-4 continued to excel, with accuracy scores ranging from 83.2% for confidence in the zero-shot setting to 96% for convenience in the zero-shot setting (Figure 5). Notably, GPT-4's performance improves as the number of shots increases, particularly for overall hesitancy, reaching an accuracy of 92% in the 10-shot setting. This showcases GPT-4's exceptional ability to learn from additional examples and accurately detect overall vaccine hesitancy.

Figure 5. Model accuracy performance by hesitancy type and number of shots.

XSL•FO
RenderX

The F_1-scores tell a similar story. GPT-4 maintained its lead, with F_1-scores ranging from 77.2% for complacency in the zero-shot setting to 89.5% for overall hesitancy in the 10-shot setting (Figure 6). The increase in F_1-scores as the number of shots grows highlights GPT-4's ability to not only accurately classify hesitancy but also comprehensively identify relevant instances.

Figure 6. Model F_1-score performance by hesitancy type and number of shots.

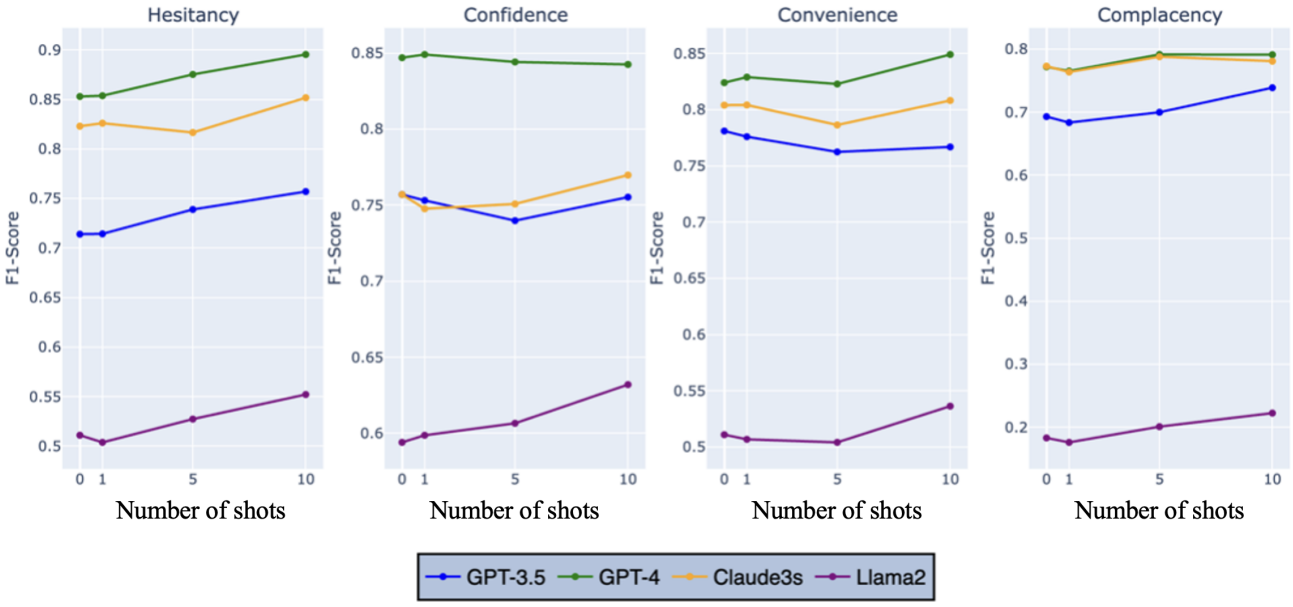

Claude-3 showed substantial improvement in accuracy for convenience hesitancy as the number of shots increased, reaching 91.2% in the 10-shot setting. This underscores Claude-3's adaptability and potential to capture specific aspects of hesitancy when provided with more examples. GPT-3.5 exhibited an improvement in accuracy for convenience hesitancy in the 10-shot setting (87.9%). Similarly, its F_1 scores for complacency hesitancy increased from 69.3% in the zero-shot setting to 73.9% in the 10-shot setting, indicating GPT-3.5's capacity for growth when given more examples. Llama 2, despite lower overall performance, showed consistent improvement in both accuracy and F_1-scores across all hesitancy types as the number of shots increased, suggesting its potential to learn and adapt.

Statistical Analysis for Model Comparative Analysis

We conducted a statistical comparison of GPT-4's F_1-scores across different learning paradigms for both sentiment and hesitancy analysis. In sentiment analysis, an ANOVA test indicated no significant differences (F-statistic=0.021; P=.981) in GPT-4's F_1-score performance across learning paradigms. Similarly, in hesitancy analysis, the ANOVA test showed no significant differences (F-statistic=0.835; P=.476) in GPT-4's F_1-score performance across learning paradigms. These results suggest that GPT-4's F_1-score performance remains consistent regardless of the number of training examples provided.

Token Usage Analysis

We analyzed the number of tokens required to run the LLMs for sentiment analysis and hesitancy detection across different vaccine categories, social media platforms, and learning paradigms. Understanding token usage patterns is crucial for estimating computational costs associated with using these models. For GPT-4, the cost is US $0.03 per 1000 tokens, while for GPT-3.5, it is US $0.0005 per 1000 tokens [36]. For Claude-3 Sonnet, the cost is approximately US $0.003 per 1000 tokens [37]. Unlike these models, Llama 2 does not have a direct cost per token; instead, its cost is related to the infrastructure required to host and run the model, such as the cost of GPU servers. The cost varies based on the instance type and configuration used for deployment [38,39].

Analyzing the average number of tokens required per post for each combination of vaccine type, social media platform, and learning paradigm (Figure 7), the results demonstrate that the number of tokens required increases with the number of training examples provided to the models. Posts related to HPV generally require more tokens than those related to MMR and general vaccines, particularly on Reddit. This suggests that discussions surrounding HPV may be more complex or lengthy, leading to higher token usage.

XSL•FO
RenderX

Figure 7. Token usage across learning paradigms and platforms for different vaccine discussions. HPV: human papillomavirus; MMR: measles, mumps, and rubella.

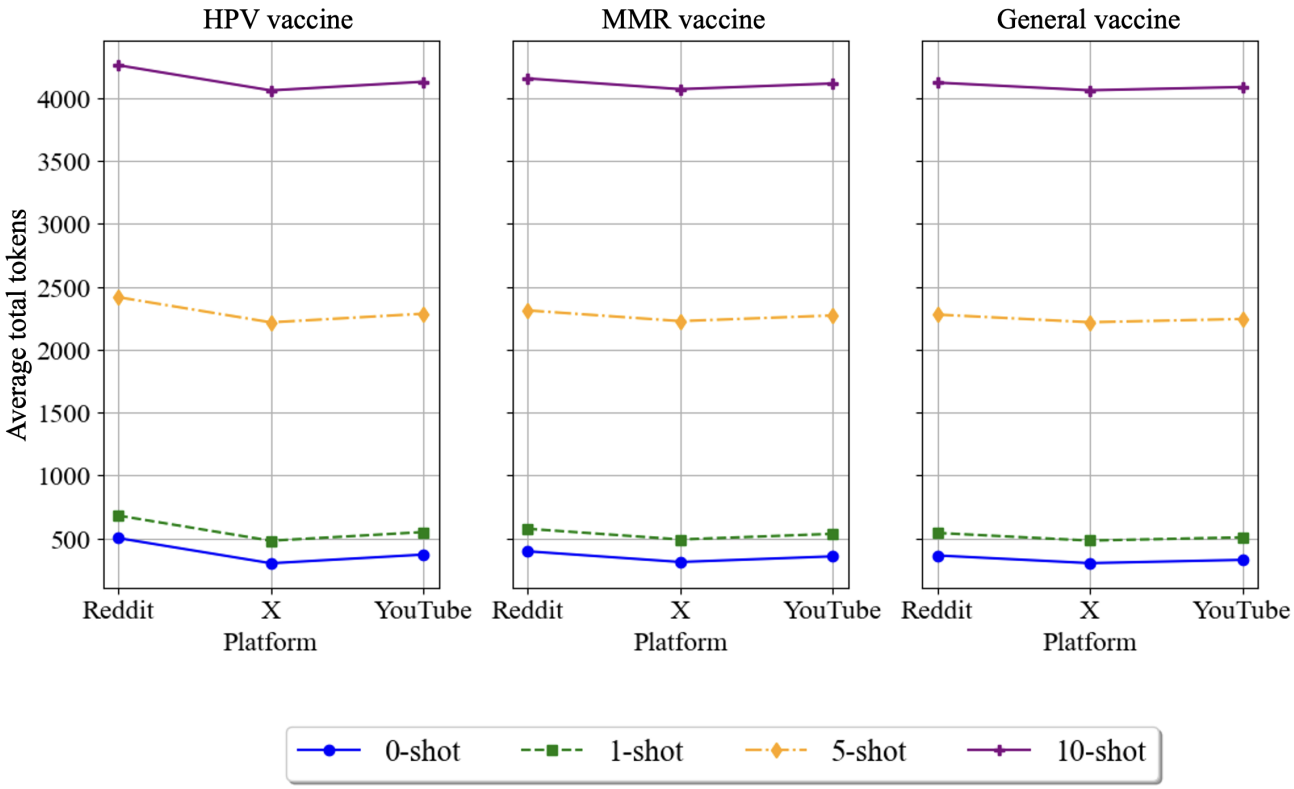

Regarding social media platforms, posts from Reddit consistently require more tokens than those from YouTube and X, across all vaccine types and learning paradigms. This indicates that the longer and more detailed nature of discussions on Reddit contributes to higher token usage compared with the other platforms.

Case-Based Evaluation of Model Performance

This section provides a case-based evaluation of misclassified posts, highlighting examples of both correct and incorrect predictions made by 4 LLMs: GPT-3.5, Llama2, Claude-3s, and GPT-4. The evaluation focuses on complex sentiments, sarcasm, and implicit support for vaccines, which often challenge LLMs in distinguishing nuanced opinions. Each post is analyzed based on its true label, the models' predictions, and their explanations, with an assessment of where and why misclassifications occurred. To safeguard privacy, all examples presented in this section are theoretical and were crafted to reflect the types of posts analyzed in the study, rather than quoting or reproducing actual user-generated content (Textboxes 1-6).

Textbox 1. Post analysis: explicit sentiment with mixed vaccine views.

Post:

"You are false. The disease has a high recovery rate for healthy individuals with no underlying conditions. If I want to avoid measles, I trust the measles vaccine—it works. But the COVID-19 vaccine? It's ineffective. Stop spreading unnecessary fear!"

- True Label: Positive (regarding MMR vaccine)
- Predictions:
  - GPT-3.5: Negative
  - Llama2: Negative
  - Claude-3s: Positive
  - GPT-4: Positive

  Explanations:

  GPT-3.5 and Llama2 focused on the aggressive tone and negative statements about the COVID-19 vaccine, overlooking the implicit endorsement of the measles vaccine. Claude-3s and GPT-4 identified the positive sentiment toward the measles vaccine despite the broader negative tone.

  Error assessment: ·
- Contextual interpretation: Models struggled with distinguishing sentiment toward different vaccines within the same post.
- Sarcasm and irony: Negative phrasing ("fear-mongering crap") misled some models into assigning an overall negative sentiment.

XSL•FO
RenderX

Textbox 2. Post analysis: complex sentiment with implicit support.

Post:

"Someone compared getting vaccinated to making a noble sacrifice—seriously? Saying vaccines are pointless because only older people are at risk is both wrong and ridiculous. But whatever, it's your choice. I'm done caring."

- True label: Positive
- Predictions:
  - GPT-3.5: Negative
  - Llama2: Negative
  - Claude-3s: Positive
  - GPT-4: Positive

Explanations:

GPT-3.5 and Llama2 misclassified the sentiment as negative due to the critical tone and the phrase "I'm done caring," which implies indifference. Claude-3s and GPT-4 correctly identified the critique of antivaccine reasoning as implicit support for vaccination.

Error assessment:

- Complex sentiment and implicit support: The post critiques antivaccine arguments, implicitly supporting vaccination, but uses sarcasm and frustration, which confused some models.
- Challenge in parsing ambiguity: Phrases such as "It's your choice" were interpreted differently across models, with some seeing them as neutral or dismissive.

Textbox 3. Post analysis: critique with ambiguous tone.

Post:

"Oh, so being smart now means trusting pharmaceutical companies completely? Here's the reality: those shouting about personal freedoms when asked to wear a mask seem more likely to try bizarre 'natural remedies' than to actually take a scientifically-proven vaccine."

- · True label: Positive
- Predictions:
  - GPT-3.5: Negative
  - Llama2: Neutral
  - Claude-3s: Negative
  - GPT-4: Neutral

Explanations:

All models struggled to classify this post correctly due to the sarcastic tone and critical phrasing. Some interpreted the critique of antivaccine rhetoric as a negative sentiment rather than implicit support for vaccination.

Error assessment:

- Complex sentiment and implicit support: The sarcastic tone and indirect language posed challenges, leading to difficulty recognizing the underlying provaccine stance.
- Misinterpretation of irony: The critique of intelligence and behavior in the context of antivaccine arguments led models to overemphasize the negative sentiment, missing the implicit support for vaccination.

XSL•FO
RenderX

Textbox 4. Post analysis: concern over measles outbreak.

Post:

"Measles isn't just back in the US; we're seeing the same issues here in Europe. It's getting out of hand, and the number of people refusing to vaccinate their kids is shocking. %VIDEO%: The return of measles—what's causing it?"

- True label: Positive
- Predictions:
  - GPT-3.5: Negative
  - Llama2: Neutral
  - Claude-3s: Negative
  - GPT-4: Positive

Explanations:

GPT-3.5 and Claude-3s misclassified the example due to phrases such as "out of hand" and "refusing to vaccinate," interpreting them as indicators of negative sentiment. GPT-4 correctly identified the post as expressing concern over measles outbreaks, which implicitly supports vaccination efforts to address the issue.

Error assessment:

- Complexity in contextual signals: The use of emotionally charged phrases such as "out of hand" misled some models into perceiving the sentiment as negative, despite the post's implicit advocacy for vaccination to control outbreaks.

Textbox 5. Post analysis: sarcastic critique of herd immunity arguments.

Post:

"Be sure to tell everyone you're skipping vaccines for diseases like polio, rubella, or whooping cough. Herd immunity will handle it, right? A few thousand dead kids is just the cost of freedom. Moo!"

- True label: Positive
- Predictions:
  - GPT-3.5: Negative
  - Llama2: Negative
  - Claude-3s: Negative
  - GPT-4: Positive

Explanations:

GPT-3.5, Llama2, and Claude-3s struggled with interpreting the sarcastic tone, focusing on phrases such as "dead kids" as indicators of negative sentiment. GPT-4 correctly understood the sarcasm as a critique of antivaccine reasoning, identifying the post's implicit support for vaccination.

Error assessment:

- Failure to recognize sarcasm: Most models struggled with the heavily sarcastic tone, focusing on surface-level negative phrases rather than the implicit provaccine stance.

XSL•FO
RenderX

Textbox 6. Post analysis for constructs.

Post:

"I thought about getting the HPV vaccine, but then I found out it doesn't cover all strains. What's the point if I could still catch something? Plus, the nearest clinic is too far, and I can't miss work just for a shot. Maybe I'll wait and see."

- True label: Hesitant (Confidence + Convenience)
- Predictions:
  - GPT-4: Hesitant (confidence)
  - Claude-3s: Hesitant (complacency)
  - GPT-3.5: Nonhesitant
  - Llama2: Nonhesitant

Explanations:

GPT-4 successfully identified the confidence issues (concerns about the vaccine's incomplete coverage) but overlooked the logistical barrier related to clinic distance and work constraints. Claude-3s misclassified the post as complacency, interpreting "wait and see" as indicating low-perceived disease risk, rather than logistical or confidence-related barriers. GPT-3.5 and Llama2 both classified the post as nonhesitant, failing to recognize the implicit skepticism and practical challenges expressed in the post.

Error assessment:

- Difficulty in addressing multiconstruct hesitancy: The intertwined nature of confidence (concerns about vaccine efficacy) and convenience (logistical challenges) constructs posed challenges for all models.
- Oversimplification of constructs: Models such as Claude-3s and GPT-4 overemphasized individual constructs, leading to incomplete classification.
- Missed implicit barriers: GPT-3.5 and Llama2 failed to interpret the implicit logistical challenges as indicative of hesitancy, instead categorizing the post as nonhesitant.

Arguments

This case-based evaluation highlights the challenges faced by LLMs in accurately classifying nuanced and complex sentiments. Models often misclassify posts due to sarcasm, indirect language, or the presence of mixed sentiments targeting different aspects of vaccination. While GPT-4 demonstrated the highest accuracy in identifying implicit support, other models frequently struggled with contextual interpretation. Future work could focus on improving models' ability to handle sarcasm, irony, and subtle expressions of support or critique.

Discussion

Principal Findings

This study evaluated the effectiveness of LLMs in sentiment analysis and hesitancy detection regarding vaccine-related posts on social media, focusing on models including GPT-3.5, GPT-4, Claude-3 Sonnet, and Llama 2. Our results indicate that GPT-4 outperforms other models across various metrics, consistent with recent advancements in LLM capabilities for processing complex linguistic patterns [40].

Sentiment and Hesitancy Analysis Insights

GPT-4 consistently outperformed other models across all sentiment types and learning paradigms, attributable to its larger parameter size and extensive pretraining on diverse datasets, enabling it to capture subtle nuances and contextual information more effectively [41]. However, our analysis also revealed that the performance of smaller LLMs, such as Claude 3s and GPT-3.5, was not far behind GPT-4, particularly in the few-shot learning paradigm. Although fine-tuning was not conducted in this study, previous research has shown that fine-tuned smaller LLMs can achieve performance levels similar to state-of-the-art models in some domain-specific applications, such as financial sentiment analysis [41]. Exploring fine-tuning as a future direction could further enhance these models' task-specific performance. From a cost perspective, smaller models can be a more viable and budget-friendly alternative. For instance, GPT-3.5 costs approximately US $0.0005 per 1000 tokens, while GPT-4 costs around US $0.03 per 1000 tokens [36]. This substantial difference in cost highlights the economic advantage of using smaller LLMs for large-scale applications where budget constraints are a factor. Moreover, the slightly reduced performance of smaller LLMs in comparison with larger models might be offset by their significantly lower operational costs, making them a practical choice for many real-world applications.

Our study found that GPT-4 handles discussions about HPV vaccines better than those about general vaccines or MMR, likely due to the more consistent language used in HPV discussions. In contrast, conversations about general vaccines or MMR cover a broader range of topics, making it harder for models to perform consistently well. In addition, the platform impacts model performance, with Reddit allowing for longer, more complex discussions, and YouTube comments often using informal language and slang.

The performance of LLMs in classifying neutral sentiments revealed notable challenges, highlighting the inherent difficulties due to their less expressive nature and greater reliance on contextual understanding. Enhanced model training that better captures the subtleties of neutral language and incorporates

XSL•FO
RenderX

contextual cues is necessary for improving neutral sentiment classification.

In hesitancy analysis, GPT-4 and Claude-3 Sonnet led the pack. Interestingly, the models generally achieved higher accuracy in hesitancy analysis than in sentiment analysis, likely due to the more specific nature of hesitancy categories. However, lower F_1-scores for the complacency hesitancy type indicate challenges in capturing nuanced and context-dependent expressions of hesitancy. Complacency, often expressed through passive language that neither outright rejects nor endorses vaccination, requires a nuanced detection approach that current models struggle to provide. This issue highlights the need for improvements in model training, suggesting that integrating more complex, contextually rich datasets could enhance model sensitivity to the passive expressions typical of complacency.

Consistency Across Learning Paradigms

Our analysis revealed that increasing the number of shots did not always lead to improved performance, aligning with findings that few-shot learning's effectiveness depends on the quality and relevance of the examples provided [41]. GPT-4 showed consistent performance across zero-shot, 1-shot, and few-shot learning paradigms, with no significant differences in performance enhancements with more training examples. This consistency makes zero-shot learning a cost-effective strategy due to its balance of accuracy and reduced computational resources [42,43].

Our token usage analysis indicated that the number of tokens required increased with the number of training examples, with the zero-shot paradigm requiring the least tokens and the 10-shot paradigm requiring the most. Token usage also varied across vaccine types and social media platforms, suggesting that the complexity and length of discussions impact computational costs.

While LLMs such as GPT-4 offer superior performance metrics, these benefits come with higher computational expenses and longer processing times. For instance, GPT-4 achieved an accuracy of 88% and an F_1-score of 0.85, outperforming traditional machine learning methods [23]. However, the associated usage costs and processing times pose challenges for large-scale, real-time deployment.

Hybrid Approach for Improved Efficiency

To address these challenges, we propose a hybrid approach leveraging both LLMs and traditional machine learning methods, combined with manual annotation. LLMs such as GPT-4 can handle a significant portion of the annotation workload, processing large volumes of data efficiently and providing consistent, high-quality annotations based on predefined criteria, reducing the burden of manual annotation and minimizing variability associated with human annotators.

A subset of the data should undergo manual annotation by human experts to ensure that nuanced and complex cases are accurately labeled, providing a benchmark to validate and refine LLM-generated annotations. Once annotated, traditional machine learning models can be trained using this hybrid-annotated dataset. These models, being more computationally efficient, can be deployed for real-time monitoring and analysis of social media data.

By integrating automated and manual annotation methods, we optimize both performance and cost, ensuring accurate and timely insights into vaccine sentiment and hesitancy. This approach enhances the feasibility of using advanced NLP techniques in public health and sets a precedent for future research in leveraging hybrid models for complex analytical tasks.

Insights From Case-Based Evaluation

The case-based evaluation underscores the complexities faced by LLMs when classifying nuanced sentiments and vaccine hesitancy constructs in social media posts. Posts containing sarcasm, indirect phrasing, and implicit sentiments were particularly challenging for models. For instance, sarcastic critiques of antivaccine rhetoric were frequently misclassified as negative sentiment, highlighting the limitations of LLMs in interpreting figurative language and irony. This aligns with previous findings that sarcasm and indirect expressions pose significant challenges in text classification tasks [44].

GPT-4 consistently exhibited better contextual understanding, particularly in recognizing implicit support for vaccination. However, even GPT-4 struggled with multiconstruct hesitancy posts, such as those intertwining confidence issues and logistical barriers. Other models, such as GPT-3.5, Llama2, and Claude-3s, often misclassified posts by overrelying on surface-level cues, failing to capture the deeper sentiment or hesitancy constructs. These observations reflect broader limitations in LLMs' ability to handle multidimensional health-related discourse.

The evaluation highlights the need for improved training and fine-tuning strategies that incorporate domain-specific linguistic features and multilabel constructs. Future research should focus on enhancing models' ability to parse complex sentiment, handle mixed tone, and interpret multiconstruct hesitancy more effectively.

Limitations

Dataset Size and Representation

A key limitation of this study is the size of the evaluation dataset, comprising 7515 annotated social media posts. While efforts were made to ensure balance across sentiment and hesitancy categories, certain subcategories, such as complacency and convenience within hesitancy, may still contain relatively small counts. This could limit the robustness of conclusions drawn for these specific subcategories. Furthermore, the reuse of the same training posts across few-shot learning experiments was a deliberate choice to maintain consistency and enable a controlled comparison of model performance. However, this may limit the diversity of training examples, which could impact the generalizability of the findings. Future studies could address these limitations by expanding the dataset, particularly within less-represented categories, and exploring the effects of more diverse training samples on LLM performance.

XSL•FO
RenderX

Contextual Influences and Pandemic Polarization

The data used in this study span up until October 2021, encompassing the period of the COVID-19 pandemic. The COVID-19 pandemic created a highly polarized environment around vaccine discussions on social media, amplifying both supportive and skeptical voices. This polarization likely influenced the data composition, with heightened expressions of hesitancy driven by safety concerns, misinformation, and politicized discourse. These patterns may have shaped the ability of LLMs to classify sentiments and hesitancy constructs, particularly when handling mixed tones or ambiguous expressions. For instance, the increased prevalence of emotionally charged or politicized language during the pandemic might have led to overemphasis on negative cues in posts that were otherwise supportive of vaccination. Addressing these nuances requires models capable of disentangling context-dependent sentiments, a challenge underscored in this study's findings. Future research could investigate how temporal factors, such as the pandemic's progression, impact sentiment and hesitancy patterns.

Limitations of API-Based Models

While our study offers valuable insights into the performance of LLMs in analyzing vaccine-related sentiments and hesitancy on social media, it is important to acknowledge several limitations that future research should address. First, our analysis was limited to the models available through Azure and Amazon Bedrock's API services at the time of our research. These included GPT-3.5, GPT-4, Claude-3 Sonnet, and Llama 2. While these models were accessible and suitable, given our computational resources, they were not necessarily the most advanced versions available. The rapid development in the field of LLMs means that more recent and potentially more effective models could offer enhanced performance on similar tasks. Also, the reliance on API versions available through Azure and Amazon Web Services means that the models and tools we used may not have been the most optimized for our specific research purposes. API versions are frequently updated to include improvements and new features that could significantly enhance model performance. Consequently, subsequent research should strive to use the most recent API versions to ensure that findings are based on the best possible technological tools and to maintain comparability with other contemporary research.

Limited Vaccine Types and Platforms

Our analysis focused on a limited set of vaccine types and social media platforms. While these choices were based on data availability and relevance to our research questions, they may not capture the full spectrum of vaccine-related discussions on social media. Future research could expand the scope to include a wider range of vaccine types and platforms to assess the generalizability of our findings and identify potential differences in sentiment and hesitancy patterns across various contexts.

Practical Integration and Ethical Considerations

This study focused on the technical aspects of using LLMs for vaccine sentiment analysis and hesitancy detection, but it did not delve into the practical implications and challenges of integrating these models into real-world public health initiatives. Future research should explore the ethical, social, and organizational factors that may influence the adoption and effectiveness of LLMs in addressing vaccine hesitancy. This could include investigating issues such as data privacy, algorithmic fairness, and the potential unintended consequences of using LLMs in public health communication strategies.

While our study highlights the potential of LLMs in identifying trends in vaccine hesitancy, it is important to recognize that these models are just one piece of the puzzle. Combating vaccine hesitancy requires a multifaceted approach that combines data-driven insights with effective communication strategies, community engagement, and trust-building efforts. Future research should explore how LLMs can be integrated into comprehensive public health interventions that address the complex and multidimensional nature of vaccine hesitancy.

Future Directions

Advanced Prompting Techniques

Future work could explore innovative prompting methods such as chain of thought reasoning and clue and reasoning prompting. These approaches could enhance the reasoning capabilities of LLMs, particularly in tasks requiring interpretation of sarcasm, indirect language, or multiconstruct hesitancy. Additionally, retrieval-augmented generation could be used to retrieve contextually relevant examples for improved performance during inference.

Domain-Specific Fine-Tuning

While fine-tuning was not conducted in this study, its potential to improve performance in domain-specific tasks warrants exploration. For instance, fine-tuning smaller models such as GPT-3.5 on datasets with domain-specific linguistic features or domain-specific datasets enriched with figurative language, sarcasm, and implicit expressions could enhance their ability to capture vaccine-related nuances while maintaining cost efficiency.

Dataset Expansion and Diversification

Increasing the size and diversity of datasets, particularly within underrepresented categories such as complacency and convenience, could provide more robust training and evaluation data. In addition, temporal analyses comparing pre- and postpandemic datasets could help identify shifts in public sentiment and hesitancy patterns.

Cross-Platform and Vaccine Type Analyses

Expanding the study to include additional social media platforms and vaccine types could improve generalizability and provide deeper insights into context-specific discourse patterns. Platforms such as Instagram or TikTok, which cater to different user demographics, may reveal distinct sentiment trends.

Ethical and Practical Considerations

Future research should explore the ethical implications of deploying LLMs in public health communication. This includes addressing algorithmic biases, ensuring data privacy, and minimizing unintended consequences, such as the spread of misinformation.

XSL•FO
RenderX

Development of Novel Metrics

Traditional metrics, such as F_1-scores, are insufficient for capturing performance in nuanced tasks such as sarcasm detection and implicit sentiment classification. Future studies should develop novel metrics, such as a "sarcasm sensitivity index" or an "implicit sentiment alignment score," to better evaluate LLMs in these complex areas.

Explainability and Interpretability Research

Investigating explainability methods to understand LLMs' decision-making processes is a promising direction. For example, using attention heatmaps to visualize how models focus on specific words or phrases could reveal their interpretation mechanisms. This could help refine models to better handle mixed or ambiguous language constructs, such as multiconstruct hesitancy posts.

Conclusions

Our study provides a foundation for understanding the performance and limitations of LLMs in vaccine sentiment analysis and hesitancy detection tasks. However, there are several limitations and opportunities for future research that should be addressed to fully harness the potential of these models in promoting vaccine acceptance and uptake. As the field of LLMs continues to evolve rapidly, researchers and public health professionals should remain vigilant in evaluating the latest developments and adapting their approaches to ensure the most effective and responsible use of these powerful tools in addressing the critical challenge of vaccine hesitancy.

Acknowledgments

This work was funded by Merck Sharp & Dohme LLC, a subsidiary of Merck & Co, Inc. The content is the sole responsibility of the authors and does not necessarily represent the official views of Merck & Co, Inc, or IMO Health Inc.

Authors' Contributions

JD, ALE, and AA contributed to study concept and design. AA, ALE, DW, and MRM performed the experiments. AA, ALE, DW, and VKN drafted the manuscript. AA, ALE, and DW contributed to acquisition, analysis, or interpretation of data. JD, ALE, and XW contributed to study supervision. All authors contributed to critical revision of the manuscript for important intellectual content.

Conflicts of Interest

ALE and DW are current employees and VKN is a contractor of Merck Sharp & Dohme LLC, a subsidiary of Merck & Co, Inc, who may own stock or stock options in Merck & Co, Inc. All the other authors were employed by IMO Health, Inc, and, therefore, IMO Health, Inc, was compensated for activities related to execution of the study.

Multimedia Appendix 1
Number of annotated posts for each social media platform, categorized by vaccine sentiment and World Health Organization's 3Cs (confidence, complacency, and convenience) model vaccine hesitancy group. HPV: human papillomavirus vaccines; MMR: measles, mumps, and rubella vaccines; General: general or unspecified vaccines.
[DOCX File, 18 KB - formative_v9i1e64723_app1.docx]

Multimedia Appendix 2
Prompt schema.
[DOCX File, 16 KB - formative_v9i1e64723_app2.docx]

References

1. GPT-4 technical report. OpenAI. 2023 Mar 15. URL: https://cdn.openai.com/papers/gpt-4.pdf [accessed 2025-09-18]
2. OpenAI API documentation (GPT-3.5 series). OpenAI. 2022 Mar 15. URL: https://platform.openai.com/docs/api-reference/introduction [accessed 2025-09-29]
3. Introducing the next generation of Claude. Anthropic. 2024. URL: https://docs.claude.com/en/home [accessed 2025-09-29]
4. Llama 2. Meta AI. URL: https://www.llama.com/llama2/ [accessed 2025-09-29]
5. Liu PR, Lu L, Zhang JY, Huo TT, Liu SX, Ye ZW. Application of artificial intelligence in medicine: An overview. Curr Med Sci 2021 Dec;41(6):1105-1115. [doi: 10.1007/s11596-021-2474-3]
6. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
7. Guerra GA, Hofmann H, Sobhani S, et al. GPT-4 artificial intelligence model outperforms ChatGPT, medical students, and neurosurgery residents on neurosurgery written board-like questions. World Neurosurg 2023 Nov;179:e160-e165. [doi: 10.1016/j.wneu.2023.08.042] [Medline: 37597659]

XSL•FO
RenderX

8. Kung JE, Marshall C, Gauthier C, Gonzalez TA, Jackson JB. Evaluating ChatGPT performance on the orthopaedic in-training examination. JB JS Open Access 2023;8(3). [doi: 10.2106/JBJS.OA.23.00056]
9. Oh N, Choi GS, Lee WY. ChatGPT goes to the operating room: evaluating GPT-4 performance and its potential in surgical education and training in the era of large language models. Ann Surg Treat Res 2023 May;104(5):269-273. [doi: 10.4174/astr.2023.104.5.269] [Medline: 37179699]
10. Brin D, Sorin V, Vaid A, et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. Sci Rep 2023 Oct 1;13(1):16492. [doi: 10.1038/s41598-023-43436-9] [Medline: 37779171]
11. Tracy BM, Hazen BJ, Ward CR, Winer JH, Pettitt BJ. Sustained clinical performance during surgical rotations predicts NBME shelf exam outcomes. J Surg Educ 2020;77(6):e116-e120. [doi: 10.1016/j.jsurg.2020.06.033] [Medline: 32651118]
12. Kossoff EH, Hubbard TW, Gowen CW Jr. Early clinical experience enhances third-year pediatrics clerkship performance. Acad Med 1999 Nov;74(11):1238-1241. [doi: 10.1097/00001888-199911000-00019] [Medline: 10587688]
13. Garcia-Vidal C, Sanjuan G, Puerta-Alcalde P, Moreno-García E, Soriano A. Artificial intelligence to support clinical decision-making processes. EBioMedicine 2019 Aug;46:27-29. [doi: 10.1016/j.ebiom.2019.07.019] [Medline: 31303500]
14. Ellahham S. Artificial intelligence: the future for diabetes care. Am J Med 2020 Aug;133(8):895-900. [doi: 10.1016/j.amjmed.2020.03.033] [Medline: 32325045]
15. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature New Biol 2023 Aug;620(7972):172-180. [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
16. Vaccine hesitancy: a growing challenge for immunization programmes. World Health Organization. 2015. URL: https://www.who.int/news/item/18-08-2015-vaccine-hesitancy-a-growing-challenge-for-immunization-programmes [accessed 2025-9-29]
17. Naveed H, Khan AU, Qiu S, Yao L, Haq AU, Ullah S, et al. A comprehensive overview of large language models. . Preprint posted online on 2023 URL: https://arxiv.org/abs/2307.06435 [accessed 2024-02-11]
18. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. OpenAI blog 2019;1(8):9 [FREE Full text]
19. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. : OpenAI; 2018 URL: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf [accessed 2024-01-11]
20. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res 2020;21(140):1-67 [FREE Full text]
21. Claude: a next-generation AI assistant. Anthropic Blog. 2023. URL: https://www.anthropic.com/claude [accessed 2024-07-15]
22. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, et al. Llama: open and efficient foundation language models. arXiv. Preprint posted online on 2023 URL: https://arxiv.org/abs/2302.13971 [accessed 2024-05-20]
23. Huang LC, Eiden AL, He L, et al. Natural language processing-powered real-time monitoring solution for vaccine sentiments and hesitancy on social media: system development and validation. JMIR Med Inform 2024 Jun 21;12(4):e57164. [doi: 10.2196/57164] [Medline: 38904984]
24. Report of the SAGE working group on vaccine hesitancy. World Health Organization. 2014. URL: https://www.asset-scienceinsociety.eu/sites/default/files/sage_working_group_revised_report_vaccine_hesitancy.pdf [accessed 2025-9-29]
25. Alarcon N. OpenAI presents GPT-3, a 175 billion parameters language model. NVIDIA Technical Blog. 2020 Apr 7. URL: https://developer.nvidia.com/blog/openai-presents-gpt-3-a-175-billion-parameters-language-model/ [accessed 2024-04-10]
26. Bastian M. GPT-3.5 might be a strong example of the efficiency potential of large AI models. The Decoder. 2023 Mar 31. URL: https://the-decoder.com/gpt-3-5-might-be-a-strong-example-of-the-efficiency-potential-of-large-ai-models/ [accessed 2024-03-10]
27. Anthropic's Claude 3 Sonnet Model now on Amazon Bedrock. Amazon Web Services. 2024. URL: https://aws.amazon.com/about-aws/whats-new/2024/03/anthropics-claude-3-sonnet-model-amazon-bedrock/ [accessed 2024-05-09]
28. Bergmann D. What is zero-shot learning? IBM. 2024. URL: https://www.ibm.com/think/topics/zero-shot-learning [accessed 2025-9-29]
29. What is zero-shot classification? Hugging Face. 2022 Nov 16. URL: https://huggingface.co/tasks/zero-shot-classification [accessed 2024-02-05]
30. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. Adv Neural Inf Process Syst 2020;33:1877-1901 [FREE Full text]
31. Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. ACM Comput Surv 2023 Sep 30;55(9):1-35. [doi: 10.1145/3560815]
32. Polak MP, Morgan D. Extracting accurate materials data from research papers with conversational language models and prompt engineering. Nat Commun 2024 Feb 21;15(1):1569. [doi: 10.1038/s41467-024-45914-8] [Medline: 38383556]
33. Gao T, Fisch A, Chen D. Making pre-trained language models better few-shot learners. 2021 Presented at: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1); Online p. 3816-3830. [doi: 10.18653/v1/2021.acl-long.295]
34. Walker SM. F-score: what are accuracy, precision, recall, and F1 score? Klu. 2023 Jul 4. URL: https://klu.ai/glossary/accuracy-precision-recall-f1 [accessed 2024-03-15]

XSL•FO
RenderX

35. Roberts M, Russo R. A Student's Guide to Analysis of Variance, 1st edition: Routledge; 1999.
36. Pricing. OpenAI. URL: https://openai.com/api/pricing/ [accessed 2024-05-25]
37. Claude API. Anthropic. URL: https://docs.claude.com/en/home [accessed 2025-9-29]
38. Introducing Llama 2 on Azure. Microsoft. URL: https://techcommunity.microsoft.com/blog/azure-ai-foundry-blog/introducing-llama-2-on-azure/3881233 [accessed 2024-02-22]
39. Fine-tune and deploy Llama 2 models cost-effectively in Amazon SageMaker JumpStart with AWS Inferentia and AWS Trainium. Amazon Web Services. 2024. URL: https://aws.amazon.com/blogs/machine-learning/fine-tune-and-deploy-llama-2-models-cost-effectively-in-amazon-sagemaker-jumpstart-with-aws-inferentia-and-aws-trainium/ [accessed 2024-05-21]
40. Baktash JA, Dawodi M. Gpt-4: a review on advancements and opportunities in natural language processing. arXiv. Preprint posted online on 2023 URL: https://arxiv.org/abs/2305.03195 [accessed 2024-05-13]
41. Fatemi S, Hu Y. A comparative analysis of fine-tuned LLMS and few-shot learning of LLMS for financial sentiment analysis. arXiv. Preprint posted online on 2023 URL: https://arxiv.org/abs/2312.08725 [accessed 2024-05-14]
42. Liu YH, Deng Y, Liu B, Pan SJ, Bing L. Understanding LLMS: a comprehensive overview from training to inference. . Preprint posted online on 2024 URL: https://arxiv.org/abs/2401.02038
43. Zhang W, Deng Y, Liu B, Pan SJ, Bing L. Sentiment analysis in the era of large language models: a reality check. arXiv. Preprint posted online on 2023. [doi: 10.48550/arXiv.2305.15005]
44. Ghosh A, Veale T. Fracking sarcasm using neural network. 2016 Presented at: Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis; Jun 16-17, 2016; San Diego, CA p. 161-169. [doi: 10.18653/v1/W16-0425]

Abbreviations

AI: artificial intelligence
API: application programming interface
GPT: generative pretrained transformer
HPV: human papillomavirus
LLM: large language model
MMR: measles, mumps, and rubella
NLP: natural language processing
3Cs: confidence, complacency, and convenience
WHO: World Health Organization

Edited by A Mavragani; submitted 24.07.24; peer-reviewed by A Zammit, Y Pang; revised version received 19.12.24; accepted 17.08.25; published 15.10.25.

Please cite as:
Annan A, Eiden AL, Wang D, Du J, Rastegar-Mojarad M, Nomula VK, Wang X
Evaluating Large Language Models for Sentiment Analysis and Hesitancy Analysis on Vaccine Posts From Social Media: Qualitative Study
JMIR Form Res 2025;9:e64723
URL: https://formative.jmir.org/2025/1/e64723
doi:10.2196/64723

© Augustine Annan, Amanda L Eiden, Dong Wang, Jingcheng Du, Majid Rastegar-Mojarad, Varun Kumar Nomula, Xiaoyan Wang. Originally published in JMIR Formative Research (https://formative.jmir.org), 15.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Effectiveness, Usability, and Satisfaction of a Self-Administered Digital Intervention for Reducing Depression, Anxiety, and Stress in a University Community in the Andean Region of Peru: Randomized Controlled Trial

Rosario Yslado-Méndez[1*], PhD; Stefan Escobar-Agreda[2*], MD; David Villarreal-Zegarra[3,4], MSc; Wilfredo Manuel Trejo Flores[5], MSc; Junior Duberli Sánchez-Broncano[1], PhD; Ana Lucia Vilela-Estrada[3,4], MSc; Jovanna Hasel Olivares Córdova[1], MD, MSc; C Mahony Reategui-Rivera[6], MD; Claudia Alvarez-Yslado[7], BEd; Leonardo Rojas-Mezarina[2], MD, MSc

[1]Facultad de Ciencias Médicas, Universidad Nacional Santiago Antúnez de Mayolo, Av. Centenario 200, Huaraz, Peru
[2]Unidad de Telesalud, Facultad de Medicina, Universidad Nacional Mayor de San Marcos, Lima, Peru
[3]Digital Health Research Center, Instituto Peruano de Orientación Psicológica, Lima, Peru
[4]Universidad Científica del Sur, Lima, Peru
[5]Facultad de Ciencias, Universidad Nacional Santiago Antúnez de Mayolo, Huaraz, Peru
[6]Escuela de Posgrado, Universidad San Ignacio de Loyola, Lima, Peru
[7]Facultad de Ciencias Sociales, Educación, Comunicación, Universidad Nacional Santiago Antúnez de Mayolo, Huaraz, Peru
[*]these authors contributed equally

Corresponding Author:
Rosario Yslado-Méndez, PhD
Facultad de Ciencias Médicas, Universidad Nacional Santiago Antúnez de Mayolo, Av. Centenario 200, Huaraz, Peru

Related Article:

This is a corrected version. See correction statement: https://formative.jmir.org/2025/1/e87717

Abstract

Background: University communities, especially in low-resource settings like Peru's Andean region, experience high rates of depression, anxiety, and stress, which harm academic performance and well-being. Traditional mental health services often remain inaccessible, necessitating scalable, self-guided solutions. While digital mental health interventions have shown promise broadly, evidence is scarce for fully self-administered platforms in low-resource university environments.

Objective: We evaluated the efficacy of a self-administered digital mental health service in members of a public university to reduce symptoms of depression, anxiety, and perceived stress.

Methods: We conducted a double-blind, parallel-group randomized controlled trial with 1:1 allocation to the digital mental health self-care service or a 30-day waiting-list control. We recruited 427 participants (students, teachers, and administrative staff) in May-June 2024, reporting mild to moderate symptoms of depression, anxiety, and stress. Participants were randomized via simple randomization and blinded through automated platform assignment. The intervention comprised 6 sequential 5-day modules grounded in acceptance and commitment therapy, mindfulness, and behavioral activation, delivered via videos, daily text prompts, workbooks, and a responsive chatbot. SMS and WhatsApp (Meta) reminders promoted adherence. Depressive symptoms (Patient Health Questionnaire-9), anxiety symptoms (Generalized Anxiety Disorder-7), and perceived stress (PSS-10) were assessed at baseline and immediately post intervention (day 30). Secondary outcomes in the intervention arm included usability (Computer System Usability Questionnaire), satisfaction (Client Satisfaction Questionnaire-8), and subjective commitment (Twente Engagement with Ehealth Technologies Scales). Analysis of covariance (ANCOVA) adjusted for baseline scores, and multivariate ANCOVA accounted for correlations among outcomes. Effect sizes were quantified using Cohen *d* and partial epsilon-squared (ε^2_p).

Results: Of 427 randomized, 85 (19.9%) completed all assessments (intervention: n=30; control: n=55). Baseline demographic and clinical characteristics were comparable between groups. Post intervention, the digital mental health self-care service group exhibited significantly greater reductions in mean Patient Health Questionnaire-9 scores (mean difference 2.78; Cohen *d*=0.64; *P*=.006), Generalized Anxiety Disorder-7 scores (mean difference 2.13; Cohen *d*=0.56; *P*=.015), and PSS-10 scores (mean

XSL•FO
RenderX

difference 4.08; Cohen d=0.69; P=.003) than controls. ANCOVA confirmed robust group effects for depression ($F_{1,2}$=9.78; P=.002; ε^2_p=0.31) and anxiety ($F_{1,2}$=8.28; P=.005; ε^2_p=0.32), with a trend toward stress reduction ($F_{1,2}$=3.73; P=.057; ε^2_p=0.46). Multivariate ANCOVA demonstrated a significant multivariate effect ($F_{4,12}$=7.23; P=.015). Among intervention completers, 100% scored below the Client Satisfaction Questionnaire-8 satisfaction threshold (<24), 60% rated platform usability as low (Computer System Usability Questionnaire<64); yet, 96.7% reported high subjective commitment (Twente Engagement with Ehealth Technologies Scales ≥18), indicating strong engagement despite interface challenges.

Conclusions: A self-administered digital self-care service effectively reduced depression, anxiety, and stress symptoms in a Peruvian university community. High user commitment underscores the platform's relevance, while low satisfaction and usability necessitate interface optimization—streamlined navigation, adaptive personalization, and feedback mechanisms—to enhance user experience and support scalable implementation in low-resource educational settings.

Trial Registration: OSF Registries OSF.IO/6AQSB; https://osf.io/6aqsb

(JMIR Form Res 2025;9:e71465) doi:10.2196/71465

KEYWORDS

digital service; university community; mental health; depression; anxiety; stress

Introduction

Mental health problems generate a high burden worldwide, contributing losses of productivity and affecting the global economy [1]. Thus, their presence in university students triggers poor academic performance and dropout, which negatively affects their chances of getting a gainful job [2,3]. This also affects teaching staff, whose well-being is linked to the mental health of students because of their close relationship [4]. One of the countries that most clearly evidences this problem is Peru, where 85% of the university community—comprising teachers, students, and administrators—report problems such as anxiety, stress, and violence [5]. This situation is caused, among other factors, by preexisting mental health problems in incoming students, which are aggravated in the Peruvian university environment, even resulting in suicide attempts [6]. Specifically, in the Andean region, where inhabitants have a greater predisposition to depression [7], college students are vulnerable to developing depression, anxiety, and stress [8].

In response to this situation, digital mental health interventions (DMHI), defined as the use of technology for the prevention and treatment of mental health conditions [9], emerge as a promising alternative. DMHI have been shown to be effective in reducing these conditions among university students, while also improving access to mental health services [9,10]; nevertheless, a recent systematic review of 89 college-student trials highlighted persistent gaps in usability reporting and real-world implementation [11]. Beyond their potential impact in Peru, the rollout of DMHI in educational environments is supported by the interest of health authorities, who emphasize the need for youth-centered services [12], and the increasing use of digital services accelerated by the COVID-19 pandemic [13]. In fact, some universities here have implemented interventions and containment measures through strategic actions that combine in-person and virtual primary mental health care services for the broader university community [14].

However, the success of DMHI can be hindered by several factors, including lack of contextualization [15], nonintuitive interface design that leads to low usability, satisfaction, and applicability [16], as well as poor adoption and acceptability [17]. Evidence suggests that the development of these interventions should be carried out with the end user, particularly when targeting young populations [18]. In Peru, various favorable experiences have been reported in the implementation of DMHI delivered via text messaging [19], telephone assistance [20], and remote support by a professional [21]; however, evidence is limited regarding the use of self-management platforms in higher education settings. Accordingly, the present study aimed to evaluate the efficacy of a digital self-care mental health service to reduce the symptoms of depression, anxiety, and perceived stress in the community of a public university in the Andean region of Peru. In addition, we aimed to understand its usability, subjective commitment, and user satisfaction.

Methods

Study Design

We conducted a double-blind, parallel-group randomized controlled trial (RCT) with a superiority hypothesis and equal allocation (1:1) to a digital mental health self-care service (DMHSS) or a waiting-list control. Assessments were performed at baseline and 1 month after the intervention. The trial protocol was preregistered on the Open Science Framework (6AQSB), and the reporting follows the CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) extension [22]. An overview of the trial procedures is shown in Figure 1.

XSL•FO
RenderX

Figure 1. Parallel randomized controlled trial design.

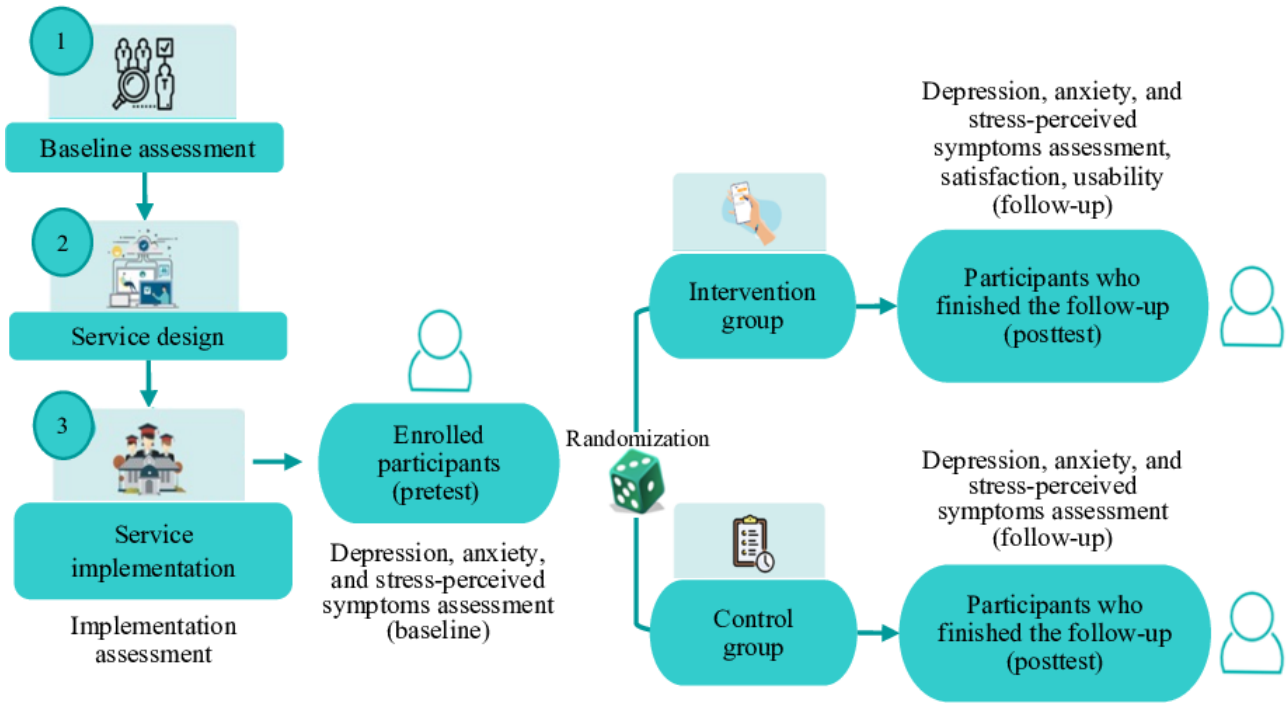

Participants

Members of the university community were invited from May to June 2024 to complete a baseline mental health screening via the self-care platform: professors (n=444), administrative staff (n=266), and students (n=7710) at a public Peruvian Andean university. Those aged ≥18 years who responded to the pretest and exhibited mild to moderate symptoms of depression, anxiety, or perceived stress were eligible for inclusion. Exclusion criteria were refusal of virtual informed consent; lack of prior experience with any virtual health service; being on leave or having reserved enrollment during the study period; current psychotherapy or psychopharmacological treatment for depression, anxiety, or stress; a prior diagnosis of severe mental health pathology; or scoring in the severe range on the baseline assessment.

Preintervention Phases

Formative Assessments

We conducted formative qualitative and quantitative evaluations to understand user preferences, platform requirements, and the prevalence of mental health problems and related factors in this population. In addition, the quantitative data allowed us to assess the psychometric properties of the mental health instruments (factorial validity and internal consistency were adequate).

Design and Development of the Digital Mental Health Self-Care Service

The design and deployment of the DMHSS involved determining the workflows and information to be delivered and building a web platform tailored to user needs identified in the formative evaluation. The platform integrated audiovisual and textual materials related to the intervention, a pre- and posttest module with automatic grading for assessment purposes, and a chatbot that responded to frequently asked questions about the platform's use. User manuals and tutorial videos were also included to facilitate adoption. The platform was developed using Vue.js for the user interface, Dialogflow for the chatbot, MySQL for database management, and the Laravel framework for backend services, following the Scrum Agile software development methodology.

Prior to its implementation, the platform was tested by the technical team and trained personnel to verify compliance with quality, security, data protection, and data flow reliability standards. In addition, an implementation checklist was used to assess the platform's operational performance through a pilot test with 30 participants from the study population. This sample size was selected to ensure early feedback on functionality and usability while maintaining operational feasibility during the preimplementation phase.

Intervention

We refer to our self-administered digital intervention throughout as the DMHSS, branded as "Amyga" for the student interface [23]. The DMHSS was developed and overseen by a clinical psychologist specializing in cognitive-behavioral psychotherapy (CBT) to ensure fidelity to evidence-based principles. It comprises 6 self-administered, 5-day modules grounded in a contextual-behavioral model (Figure 2). Each module combined 3 - 4 brief (5 - 8 min) videos (with matching audio tracks), 20 - 30 daily text prompts, and a 3 - to 8-page PDF workbook, all adaptively tailored to users' baseline Patient Health Questionnaire-9 (PHQ-9) and Generalized Anxiety Disorder-7 (GAD-7) scores. The modules were organized into 3 pairs by psychotherapeutic technique: modules 1 and 4 applied acceptance and commitment therapy (ACT) to foster emotional acceptance and clarify personal values [24]; modules 2 and 5 used mindfulness practices to cultivate nonjudgmental present-moment awareness and stress tolerance [25]; and modules 3 and 6 targeted experiential avoidance through gradual

XSL•FO
RenderX

exposure and behavioral activation to build resilience and engagement in meaningful activities [26]. Detailed content and resource quantities for each module are provided in Table S1 in Multimedia Appendix 1.

Figure 2. Flowchart of the 6-module digital mental health self-care service.

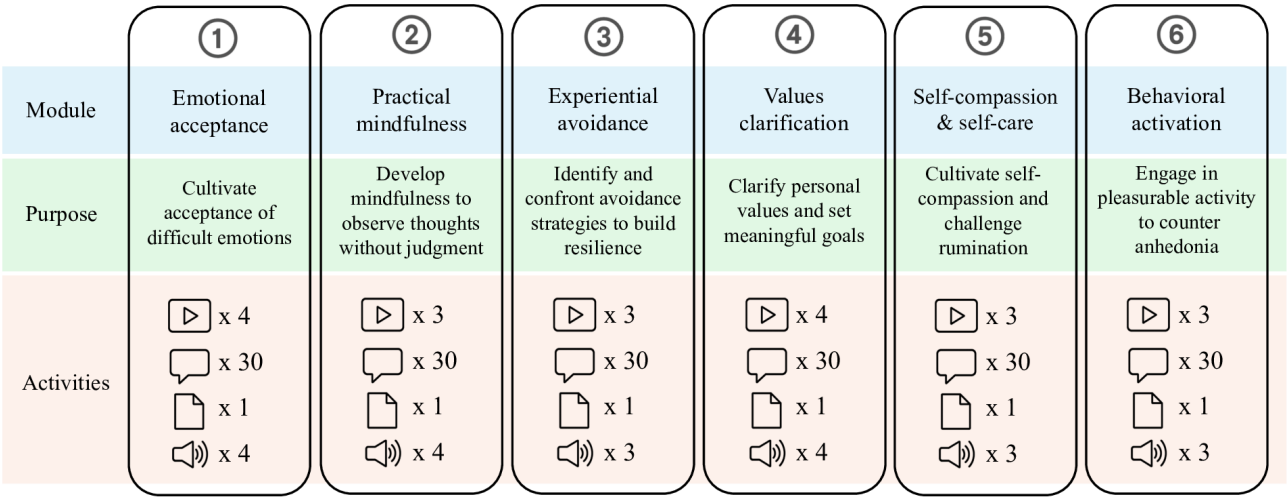

In each module, the video icon denotes short tutorials that blend theoretical background with practical examples; the daily-message icon marks brief prompts sent each day to reinforce learning and encourage reflection; the document icon indicates downloadable documents with step-by-step instructions and written exercises; and the audio icon represents narrated versions of the tutorials and exercises for offline or hands-free use.

We reinforced engagement by sending SMS or WhatsApp reminders to participants at the beginning, middle, and end of each module. In addition, reminders were sent to participants who did not complete the activities in the estimated time, in order to encourage compliance with the established deadlines. Control-group participants were placed on a 30-day waiting list; after this period, they received full access to the DMHSS. All participants also received information on standard university mental health services.

Variables and Measurement Instruments

Main Outcomes

The study had 3 central outcomes: depressive symptoms, anxious symptoms, and perceived stress. Depressive symptoms were measured using the PHQ-9, a 9-item instrument validated in the Peruvian university population [27], scored from 0 (not at all) to 3 (nearly every day). Anxiety symptoms were measured using the GAD-7, a 7-item instrument validated in the Peruvian university population [28] with items scored 0 - 3 on a Likert scale. Perceived stress was measured with the Perceived Stress Scale-10 (PSS-10; 10 items, scored 0 - 3), validated in the Peruvian population [29]. All 3 scales were administered at baseline (day 0) and immediately postintervention (day 30).

Secondary Outcomes

Participants in the intervention group who completed all modules were assessed for satisfaction, usability, and subjective engagement with the digital service. Usability was measured with the Computer System Usability Questionnaire (CSUQ), a 16-item questionnaire scored on a 7-point Likert scale. Subjective commitment was measured with the Twente Engagement with Ehealth Technologies Scale (TWEETS) questionnaire (9 items on a 5-point Likert scale). Finally, satisfaction was assessed using the Client Satisfaction Questionnaire-8 (CSQ-8; 8 items). Scores below each instrument's cutoff point defined a "low" category: CSQ-8<24 for satisfaction, CSUQ<64 for usability, and TWEETS <18 for subjective commitment.

Covariates

Covariates included sex, age, device preferences (phone, PC, and tablet), mobile internet access, participant role (student, teacher, and administrator), and history of mental health diagnosis.

Sample Size

We based our sample size calculation on the minimum number of participants required to conduct our primary analysis, which aims to detect differences between the control and intervention groups in their pre- and postintervention scores. Accordingly, we applied a one-way analysis of covariance (ANCOVA) model with 2 arms (control and experimental) and 2 time points (pre and post). We set statistical power at 80%, α=.05, and assumed a pre–post correlation of .7 and a medium effect size (η^2=0.20), following Shieh's approach [30]. This procedure yielded a required minimum sample of 102 participants, corresponding to 51 participants per arm. We adopted the medium effect size estimate on the basis of a previous randomized clinical trial that used ANCOVA with PHQ-9 and GAD-7 outcomes in a DMHI [31].

Randomization and Blinding

Randomization sequences were generated using the randomizeR package, applying simple randomization to allocate participants 1:1 to control or intervention groups, with 10,000 allocations produced.

Investigators and analysts were blinded. The platform automatically assigned participants; group allocation remained

XSL•FO
RenderX

concealed until final analysis and manuscript drafting to prevent bias.

Procedures

In collaboration with the university welfare department, we ran an advertising campaign to recruit students, teachers, and administrative staff via campus announcements and online registration. Eligible individuals were then randomized to the intervention or waiting-list control. The intervention group engaged with the 6-module DMHSS over the next 30 days, while control participants remained on the waiting list and received access afterward. Both groups completed the postintervention assessment immediately on day 30 using the same online instruments.

Statistical Analysis

Continuous variables are presented as mean (SD) or median (IQR); categorical variables as frequencies and percentages. Normality was assessed with Shapiro–Wilk, and independent-samples *t* tests were applied when appropriate. To isolate the effect of the intervention on posttest scores while accounting for baseline differences, we performed ANCOVA, which adjusts each follow-up outcome for its corresponding pretest value. Intervention effects were quantified with Cohen *d* (0.20=small; 0.50=moderate; 0.80=large) and partial epsilon-squared (ε^2_p) (0.02=small; 0.15=moderate; 0.30=large) for intuitive purposes [32]. All analyses were performed in RStudio (Posit PBC) and Stata 17 (StataCorp LLC).

Ethical Considerations

The study protocol was approved by the Universidad Nacional Santiago Antúnez de Mayolo Research Ethics Committee (report numbers 011 - 2022 and 07 - 2023). All participants provided electronic informed consent before taking part and were free to withdraw at any time without penalty. No financial or other compensation was offered for participation. To protect privacy and confidentiality, we assigned alphanumeric codes to each record, stored data on secure sockets layer–encrypted servers, and removed all personal identifiers after analysis in compliance with Peru's data protection law (Number 29733). The study was conducted in accordance with the World Medical Association's ethical principles [33].

Results

Of the 569 participants who registered on the self-care web platform, 119 were excluded for presenting severe symptoms, 20 for invalid data or ineligibility, and 3 for incomplete baseline questionnaires, resulting in 427 participants evaluated at baseline. Among these, 218 were randomized to the control group, and 209 to the intervention group. Ultimately, 85 (19.9%) completed the full intervention and postintervention assessments (Figure 3).

XSL•FO
RenderX

Figure 3. Study flowchart.

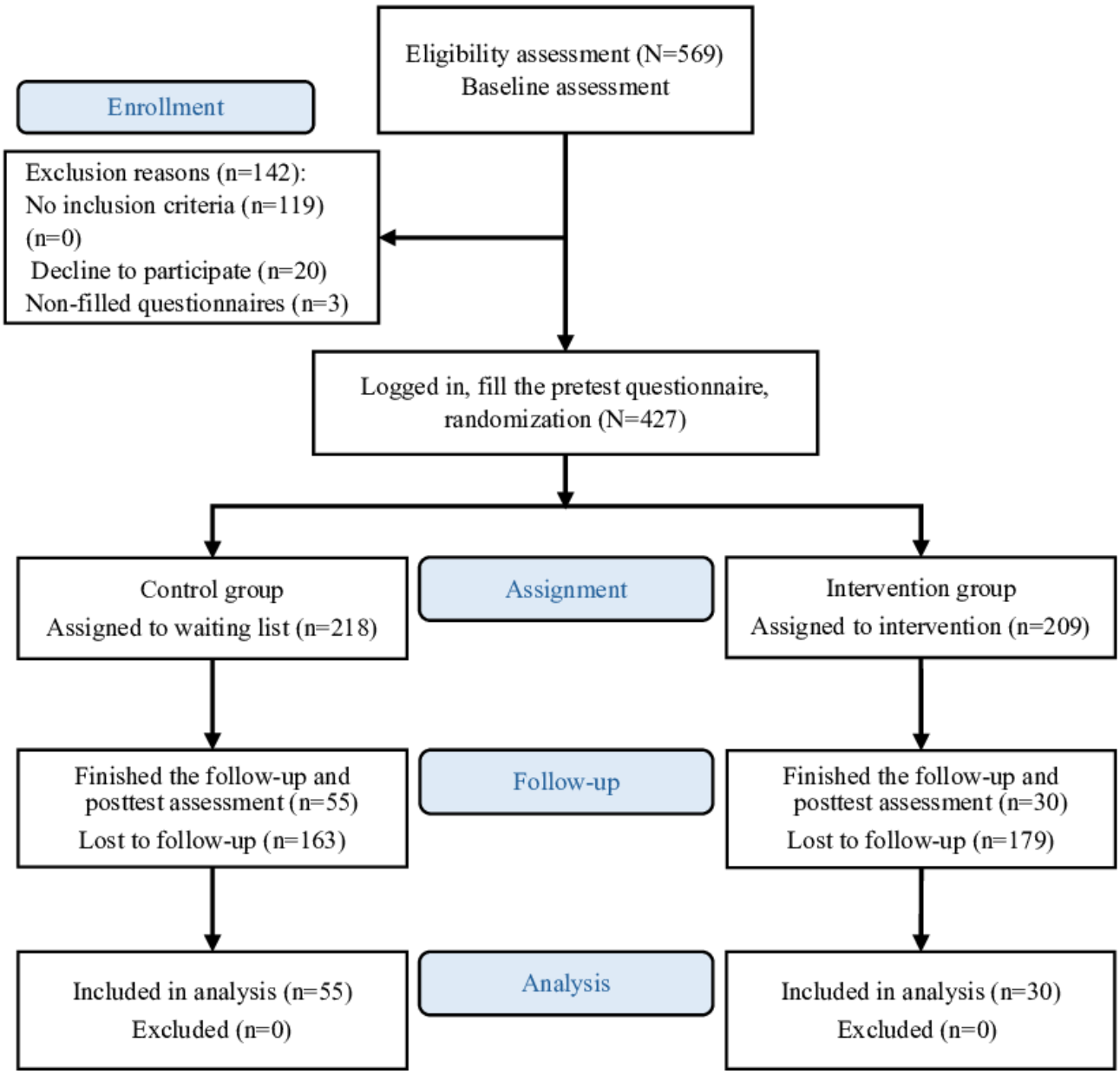

Baseline characteristics—including age, sex, and initial mental health scores—were comparable between groups (Table 1). In Tables 2 and 3, the scores of depressive symptoms, anxiety, and stress were compared before the intervention, and we observed that participants of the intervention group had lower stress scores (P=.02). Control group participants did not have a notable change in their scores of depressive symptoms, anxiety, or stress, but the participants of the intervention group did show a greater reduction in all scores. In addition, the latter showed effect measures considered large [32]. On the other hand, when comparing the control and intervention groups after the intervention, the results favored the intervention group with larger effect measures (all Cohen d>0.5). These standardized effect sizes and their 95% CI are summarized graphically in Figure 4.

XSL•FO
RenderX

Table . Characteristics of participants before the intervention.

Groups	Control	Intervention	*P* value
Age (years), mean (SD)	23.6 (7.3)	23.4 (6.7)	.78
Sex, n (%)			.80
Female	124 (56.6)	122 (57.8)	
Male	95 (43.4)	89 (42.2)	
Profession, n (%)			.52
Administrative staff	12 (5.5)	8 (3.8)	
Teacher	7 (3.2)	10 (4.7)	
Student	200 (91.3)	193 (91.5)	
Depressive symptoms, n (%)			.93
No symptoms	31 (14.2)	34 (16.3)	
Mild	95 (43.6)	87 (41.6)	
Moderate	56 (25.7)	55 (26.3)	
Severe	36 (16.5)	33 (15.8)	
Anxiety symptoms, n (%)			.60
No symptoms	49 (22.5)	48 (23)	
Mild	116 (53.2)	99 (47.4)	
Moderate	45 (20.6)	52 (24.9)	
Severe	8 (3.7)	10 (4.8)	
Stress symptoms, n (%)			.90
Mild	18 (8.3)	19 (9.1)	
Moderate	179 (82.1)	172 (82.3)	
Severe	21 (9.6)	18 (8.6)	

Table . Comparison of study groups according to assignment to the experiment and according to the time of intervention.

Study groups	Control (n=55)	Intervention (n=30)	Difference	Cohen *d* (95% CI)	*P* value
Before intervention, mean (SD)					
PHQ-9[a] score	8.33 (4.91)	8.10 (4.83)	0.23	0.05 (−0.4 to 0.49)	.84
GAD-7[b] score	6.53 (3.54)	6.47 (3.57)	0.06	0.02 (−0.43 to 0.46)	.94
PSS[c] score	20.62 (5.10)	18 (4.42)	2.62	0.54 (0.08 - 0.99)	.02
After intervention, mean (SD)					
PHQ-9 score	8.31 (4.73)	5.53 (3.56)	2.78	0.64 (0.18 - 1.09)	.006
GAD-7 score	7.13 (4.03)	5 (3.24)	2.13	0.56 (0.11 - 1.02)	.02
PSS score	19.55 (5.63)	15.47 (6.29)	4.08	0.69 (0.24 - 1.15)	.003

[a]PHQ-9: Patient Health Questionnaire-9.

[b]GAD-7: Generalized Anxiety Disorder Assessment-7.

[c]PSS: Perceived Stress Scale.

XSL•FO
RenderX

Table . Comparison of study groups according to the time of intervention.

Study groups	Before intervention (n=219)	After intervention (n=85)	Difference	Cohen *d* (95% CI)	*P* value
Control group, mean (SD)					
PHQ-9[a] score	8.33 (4.91)	8.31 (4.73)	0.02	0.003 (−0.37 to 0.38)	0.98
GAD-7[b] score	6.53 (3.54)	7.13 (4.03)	−0.60	−0.16 (−0.53 to 0.22)	0.19
PSS[c] score	20.62 (5.10)	19.55 (5.63)	1.07	0.2 (−0.18 to 0.57)	0.07
Intervention group, mean (SD)					
PHQ-9 score	8.10 (4.83)	5.53 (3.56)	2.57	0.6 (0.08 to 1.12)	0.006
GAD-7 score	6.47 (3.57)	5 (3.24)	1.47	0.43 (−0.08 to 0.94)	0.04
PSS score	18 (4.42)	15.47 (6.29)	2.53	0.47 (−0.05 to 0.98)	0.01

[a]PHQ-9: Patient Health Questionnaire-9.

[b]GAD-7: Generalized Anxiety Disorder Assessment-7.

[c]PSS: Perceived Stress Scale.

Figure 4. Between-group effect sizes (Cohen *d*) for depressive, anxiety, and stress symptoms. GAD-7: Generalized Anxiety Disorder 7; PHQ-9: Patient Health Questionnaire-9; PSS: Perceived Stress Scale.

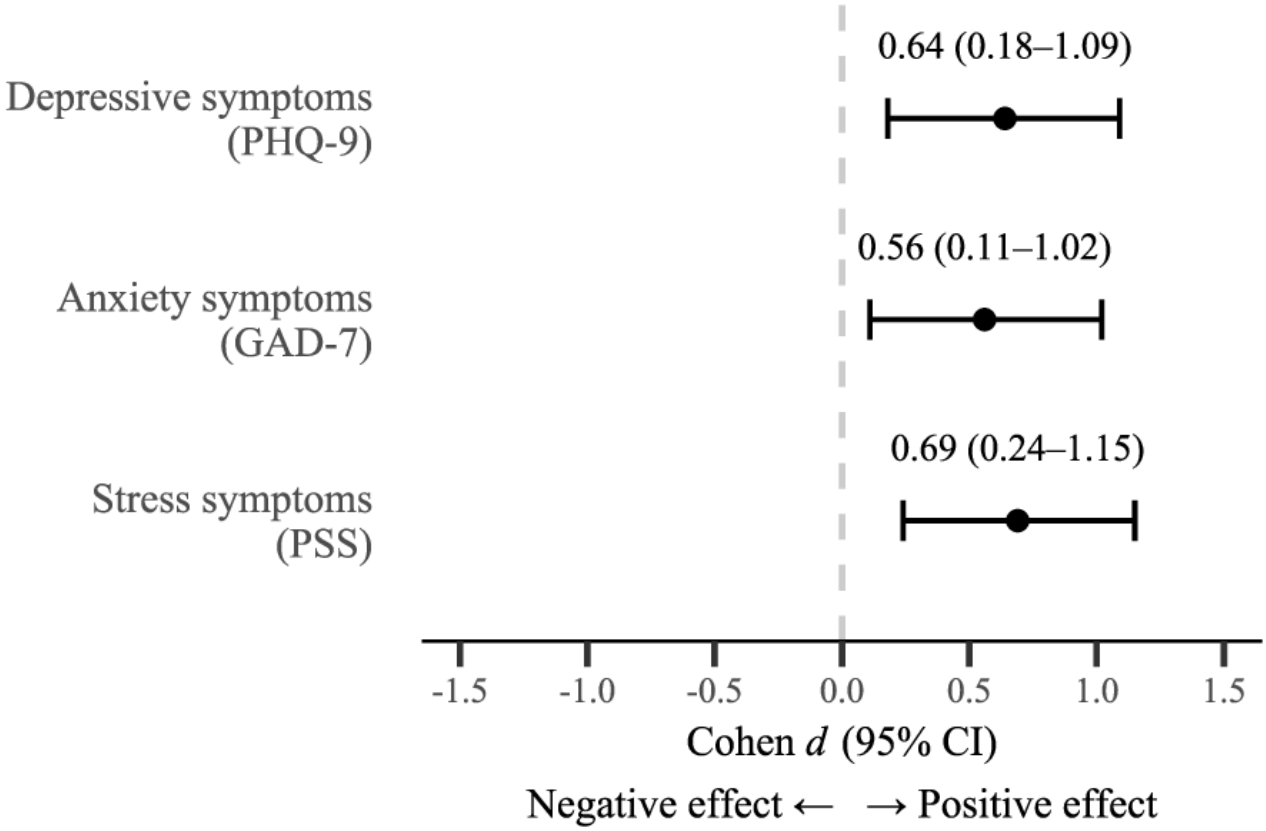

To account for the group effect and control for possible residual confounding, 3 ANCOVA models were performed. In each model, the outcome was the depression, anxiety, or stress score obtained after the intervention, and the factors were the study group and the baseline measure corresponding to each outcome. Table 4 shows that the differences between the control and intervention groups for depressive and anxiety symptoms remain significant and similar to those obtained in Table 2. Although the stress difference no longer reached statistical significance. Considering the correlation between depression, anxiety, and stress scores, a multivariate ANCOVA model was applied. Table 4 shows the differences predicted by the model and the differences between the groups in all scores are smaller compared to the bivariate analysis in Table 1. However, they still show that participants in the intervention group had lower values for depressive symptoms, anxiety, and stress. In addition, all these differences have a significant *P* value ($P<.05$), and the effect measures are still classified as having a large effect in favor of the intervention.

XSL•FO
RenderX

Table . Analysis of variance for comparison of intervention and control groups.

Statistical models and outcomes	*F* test (*df*)	Difference between control and intervention group	ε^2_p [a]	*P* value
ANCOVA[b]				
PHQ-9[c]	9.78 (1, 82)	2.67	0.31	.002
GAD-7[d]	8.28 (1, 82)	2.09	0.32	.005
PSS[e]	3.73 (1, 82)	4.08	0.46	.06
Multivariate ANCOVA				
PHQ-9[c]	7.23 (4, 12)	2.214	N/A[f]	.02
GAD-7[d]	7.23 (4, 12)	1.527	N/A	.04
PSS[e]	7.23 (4, 12)	2.216	N/A	.045

[a]ε^2_p: squared epsilon effect measure.

[b]ANCOVA: analysis of covariance.

[c]PHQ-9: Patient Health Questionnaire-9.

[d]GAD-7: Generalized Anxiety Disorder-7.

[e]PSS: Perceived Stress Scale.

[f]N/A: not applicable.

Assessment of satisfaction, usability, and subjective commitment of the university community in relation to the use of the digital mental health service, several results were shown (Table 5). We observed low satisfaction with the digital service (<24 points on the CSQ-8 scale) in the sample. In addition, more than half of the participants (60%, 18/30) reported a low usability of the platform (<64 points on the CSUQ questionnaire). On the other hand, most participants (96.7%, 29/30) showed a high level of subjective commitment to this tool (≥18 points on the TWEETS questionnaire).

XSL•FO
RenderX

Table . Level of satisfaction, usability, and subjective commitment to the use of the mental health platform.

Implementation outcomes	Low, n (%)	High, n (%)
Satisfaction level (CSQ-8)[a,b]	<24	≥24
Participant, n (%)		
Administrative staff	1 (100)	0 (0)
Teacher	6 (100)	0 (0)
Student	23 (100)	0 (0)
Total	30 (100)	0 (0)
Usability level (CSUQ)[a,c]	<64	≥64
Participant, n (%)		
Administrative staff	1 (100)	0 (0)
Teacher	5 (83.3)	1 (16.7)
Student	12 (52.2)	11 (47.8)
Total	18 (60)	12 (40)
Level of commitment (TWEETS)[a,d]	<18	≥18
Participant, n (%)		
Administrative staff	0 (0)	1 (100)
Teacher	1 (16.7)	5 (83.3)
Student	0 (0)	23 (100)
Total	1 (3.3)	29 (96.7)

[a]Cut-off points determined according to the mean score in each of the questionnaires.

[b]CSQ-8: Client Satisfaction Questionnaire-8.

[c]CSUQ: Computer System Usability Questionnaire.

[d]TWEET: Twente Engagement with Ehealth Technologies Scale.

Discussion

Principal Findings

This pilot study assessed the effectiveness of a digital mental health self-care intervention grounded in a contextual-behavioral framework that included acceptance and commitment techniques, experiential avoidance, and mindfulness among students, teachers, and administrative staff at a public university in Peru. Our results suggested that participants in the intervention group experienced a significant reduction in symptoms of depression, anxiety, and stress compared to the control group. Although most participants exhibited high commitment to the platform, satisfaction and usability ratings were low.

Several studies have reported heterogeneous success rates of virtual platforms for mental health and varying user perceptions [34], which in some cases contrasts with our findings. These differences may stem from the therapeutic techniques and approaches used, the duration and format of the intervention, the digital skills of the users, the cultural appropriateness of the intervention, and support personnel.

First, our intervention was grounded in a contextual-behavioral framework, using ACT and mindfulness techniques. Prior studies have demonstrated the effectiveness of these techniques in digital environments. One study assessed an online intervention combining mindfulness and CBT techniques, which showed reductions in depression and anxiety among university students but no significant impact on perceived stress [35]. Another online ACT-based intervention among university students found significant improvements in depression, anxiety, and stress [36]. Together, this evidence suggests that, although mindfulness and CBT techniques effectively address depression and anxiety, complementing them with ACT may be necessary for effective stress management in the university population.

Intervention duration and intensity also influence outcomes. For example, an evaluation of a 6-module digital mental health program found no significant effects on depression and anxiety symptoms [37]. The lack of effectiveness was attributed to low participation and adherence rates, as well as potential issues with intervention delivery. Conversely, another study has indicated that longer interventions (>8 weeks) tend to be more effective in reducing depression and anxiety symptoms [38]. This suggests a need to balance the length of DMHI, tailoring them to specific contexts to achieve effectiveness without hindering user participation. In our study, the reduced module count and diverse content likely supported engagement and contributed to observed mental health improvements, despite overall low completion rates.

XSL•FO
RenderX

Various DMHI formats exist for university students; among these, internet-based interventions are the most commonly used, and our study confirmed their effectiveness in reducing depression and anxiety. However, interactivity and usability are crucial for ensuring this effectiveness [39]. Although our virtual platform featured a responsive design for mobile access, reported satisfaction and usability levels were low, suggesting it failed to provide an optimal user experience to foster participation and maximize benefits. Limited digital skills likely further hindered engagement, as some struggled with navigation, reducing the effective "dose." In the Peruvian university context, teachers typically possess basic digital skills, and students generally have intermediate skills [40]. These findings partially explain the challenges faced by participants in using the DMHSS, contributing to their low levels of interest, participation, and overall satisfaction.

Several mechanisms may explain why participants maintained high engagement yet rated usability and satisfaction poorly. DMHI may unintentionally create cognitive overload when many modules are displayed simultaneously, a phenomenon linked to "digital fatigue" and reduced app acceptability among university students [41]. Reviews further show that static, nonpersonalized content lowers perceived relevance, whereas adaptive pathways that tailor material to baseline symptom severity improve adherence and satisfaction [42]. Finally, the absence of real-time feedback loops, such as progress indicators or light gamification, diminishes users' sense of accomplishment; interventions that embed these features consistently achieve high usability scores [43].

In light of this evidence, future platform iterations should streamline navigation via progressive disclosure, incorporate rule-based personalization, and add succinct progress feedback or gamified cues. These refinements, grounded in contemporary usability research, warrant formal A/B testing to quantify their impact on satisfaction, usability, and mental health outcomes, particularly in low-resource university settings. In addition, exploring the incorporation of asynchronous peer-support chats could further assist users with low digital literacy without compromising the autonomy of the intervention.

It has also been emphasized that cultural and contextual aspects should be included in designing DMHI for university students to improve acceptance and effectiveness [44]. In our study, formative studies that assessed the expectations and preferences of students, teachers, and administrative staff regarding DMHI facilitated the intervention design. Including contextually relevant examples and content adapted to the reality of participants likely contributed to the high level of commitment observed despite technological usability or satisfaction challenges.

Moreover, some studies show that support personnel can enhance DMHI effectiveness compared to purely self-administered formats [45], though others find no significant effect [46]. Although our intervention was self-administered, administrative staff provided reminders and technical support, underscoring the value of human support. This highlights the importance of support even in self-administered interventions, particularly among users with procrastination tendencies, which have been shown to reduce the adherence to digital interventions [47], and in contexts of low digital literacy and persistent mental health stigma, such as observed in the Andean region of Peru [7].

Considering the broader applicability of the DMHSS, future integration into existing university mental health services could enhance its impact. Positioning this platform as a first-line self-guided option within a stepped-care model or as an adjunct to traditional counseling services may address unmet mental health needs on campuses. Effective implementation, however, will require adequate IT infrastructure, minimal staff training, clear data-privacy protocols, and consideration of potential barriers such as competing academic demands and limited technical support. Evaluating these factors through future pilot programs is critical to ensuring scalable adoption and sustainability.

Limitations and Strengths

Several limitations warrant careful consideration. First, attrition was high: only about 20% of intervention participants completed all modules, which narrowed our analytic sample and may limit the representativeness of the findings (external validity). However, the subset who completed the intervention remained balanced at baseline (Table 2) and yielded sufficient power to detect statistically and clinically meaningful between-group differences, suggesting that internal validity was preserved despite the high dropout. Second, completion rates differed between arms (15% intervention vs 25% control), raising the possibility of differential bias. Nevertheless, baseline characteristics (including age, sex, and initial symptom severity) were equivalent across groups (Table 2), suggesting that this imbalance is unlikely to compromise internal validity. Third, satisfaction and usability scores were low, particularly among administrative and teaching staff, underscoring the need for interface refinements. Fourth, all outcomes were self-reported, introducing potential response and social-desirability bias, though validated instruments and blinded allocation mitigate this risk. Fifth, outcomes were measured only immediately postintervention; thus, durability is unknown and longer follow-up is needed. Sixth, we did not conduct qualitative interviews or focus groups; specific cultural and technological barriers were not explored in depth. Future research will incorporate semistructured interviews and user feedback sessions to address these gaps. Seventh, although we report Cohen *d* and ε^2_p to convey effect magnitude, we did not prespecify or validate minimal clinically important difference thresholds for these instruments, which would aid interpretation of clinical relevance; future work should incorporate such benchmarks. Finally, the study occurred at a single public university in the Andean region, limiting generalizability, although the setting represents typical low-resource universities in the region.

Balanced against these constraints are several noteworthy strengths. The RCT design minimizes selection bias and supports robust causal inference. The intervention is multicomponent and theory-driven, combining ACT, mindfulness, and experiential-avoidance principles that have strong empirical support. In addition, this study extends the scant literature of DMHI in Latin American low-resource contexts [48], and is to

XSL•FO
RenderX

our knowledge, one of the first to include teaching and administrative staff alongside students in the effectiveness of digital interventions in mental health in teaching staff, enlarging the scope of evidence on campus mental health of this population [49].

Recommendations for Future Research

Future studies should evaluate long-term intervention effects with larger samples and replicate research across diverse universities and cultural contexts, including multicountry RCT [50] to assess feasibility, sustainability, and scalability. They should also examine factors influencing participation and engagement, the impact of personalized digital interventions tailored to individual needs, and ways to optimize usability via user-centered, minimalist design, iterative interface evaluation, and enhanced prototypes. It is also recommended to investigate the determinants of low satisfaction, including discrepancies between user expectations and offered functionalities, and the potential role for complementary human support. Future research should use rigorous methodologies such as A/B testing to quantify the effects of these design enhancements on usability, satisfaction, and mental health outcomes in low-resource university settings. Finally, because limited digital literacy may have attenuated our observed effects, future iterations should include targeted digital skills training or onboarding modules to ensure all users can effectively navigate and benefit from the platform.

Conclusions

In conclusion, this pilot study provides preliminary evidence that a multicomponent intervention of a digital self-care intervention based on acceptance and commitment techniques, experiential avoidance, and mindfulness effectively reduces depression, anxiety, and stress among university populations. This highlights the potential of digital interventions to address mental health problems in populations where access to traditional health services is limited. However, intervention format, content, duration, and level of digital skills of users must be carefully considered to ensure adequate usability and satisfaction, promoting future scalability and sustainability.

Acknowledgments

The study was funded by the Universidad Nacional Santiago Antúnez de Mayolo (Ancash, Peru). Funding was approved by Resolution of the University Council - Rector N 703 - 2022-UNASAM. The authors thank all the participants in the study and the staff of the General Office of Information Technology, Systems and Statistics of the Santiago Antúnez de Mayolo National University.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

RY-M: Writing – review and editing, writing – original draft, visualization, project administration, methodology, and funding acquisition.
SE-A: Writing – review and editing, writing – original draft, and formal analysis.
DV-Z: Conceptualization, writing – review and editing, and formal analysis.
WMTF: Writing – review and editing, and software.
JS-B: Writing – review and editing, and data curation.
ALV-E: Writing – review and editing, and methodology.
JHO-C: Writing – review and editing, and supervision.
MR: Writing – review and editing, and validation.
CRAY: Writing – review and editing, and resources.
Leonardo R-M: Writing – review and editing, and investigation.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed content and resource breakdown of the 6 DMHSS (digital mental health self-care service) modules.
[DOCX File, 17 KB - formative_v9i1e71465_app1.docx]

Checklist 1
CONSORT-EHEALTH checklist (V 1.6.1).
[PDF File, 1148 KB - formative_v9i1e71465_app2.pdf]

References

XSL•FO
RenderX

1. Greenberg PE, Fournier AA, Sisitsky T, Pike CT, Kessler RC. The economic burden of adults with major depressive disorder in the United States (2005 and 2010). J Clin Psychiatry 2015 Feb;76(2):155-162. [doi: 10.4088/JCP.14m09298] [Medline: 25742202]
2. Mental health problems increase among university students, report finds [Web page in Spanish]. General Council of Psychology of Spain-INFOCOP. 2019. URL: https://www.infocop.es/aumentan-los-problemas-de-salud-mental-entre-los-estudiantes-universitarios-segun-un-informe/ [accessed 2024-12-10]
3. Pérez G. Mental health of students in a pandemic: research by the consortium of universities will allow support actions to be taken [Web page in Spanish]. PuntoEdu PUCP. 2021. URL: https://puntoedu.pucp.edu.pe/comunidad-pucp/salud-mental-de-estudiantes-en-pandemia-investigacion-del-consorcio-de-universidades-permitira-tomar-acciones-de-apoyo/ [accessed 2024-12-10]
4. Harding S, Morris R, Gunnell D, et al. Is teachers' mental health and wellbeing associated with students' mental health and wellbeing? J Affect Disord 2019 Jan 1;242:180-187. [doi: 10.1016/j.jad.2018.08.080] [Medline: 30189355]
5. Mental health: Ministry of Education and Ministry of Health (MINSA) work with 21 public universities [Web page in Spanish]. Peruvian Andean News Agency. 2019. URL: https://andina.pe/agencia/noticia-salud-mental-minedu-y-minsa-trabajan-21-universidades-publicas-769880.aspx [accessed 2024-12-10]
6. Perales A, Izaguirre M, Sánchez E, et al. Salud mental en estudiantes de pregrado de la Facultad de Medicina de la Universidad Nacional Mayor de San Marcos [Article in Spanish]. An Fac med ;80(4):443-450. [doi: 10.15381/anales.v80i4.17142]
7. Hernández-Vásquez A, Vargas-Fernández R, Bendezu-Quispe G, Grendas LN. Depression in the Peruvian population and its associated factors: analysis of a national health survey. J Affect Disord 2020 Aug 1;273:291-297. [doi: 10.1016/j.jad.2020.03.100] [Medline: 32421615]
8. Robles NEC, Bravo LMR, Riva M, Suyo -Vega JA. Salud mental en adolescentes universitarios Latinoaméricanos: Revisión sistemática [Article in Spanish]. Archivos Venezolanos de Farmacología y Terapéutica 2020;39(6):689-695. [doi: 10.5281/zenodo.4403731]
9. Ramos G, Hernandez-Ramos R, Taylor M, Schueller SM. State of the science: using digital mental health interventions to extend the impact of psychological services. Behav Ther 2024 Nov;55(6):1364-1379. [doi: 10.1016/j.beth.2024.04.004] [Medline: 39443071]
10. Riboldi I, Cavaleri D, Calabrese A, et al. Digital mental health interventions for anxiety and depressive symptoms in university students during the COVID-19 pandemic: a systematic review of randomized controlled trials. Rev Psiquiatr Salud Ment 2023;16:47-58. [doi: 10.1016/j.rpsm.2022.04.005] [Medline: 35755491]
11. Lattie EG, Adkins EC, Winquist N, Stiles-Shields C, Wafford QE, Graham AK. Digital mental health interventions for depression, anxiety, and enhancement of psychological well-being among college students: systematic review. J Med Internet Res 2019 Jul 22;21(7):e12869. [doi: 10.2196/12869] [Medline: 31333198]
12. MINSA inaugurated the first university community mental health center at the universidad mayor de san marcos. Ministry of Health. URL: https://digep.minsa.gob.pe/noticias/noticias_15102019.html [accessed 2024-12-10]
13. Villarreal-Zegarra D, Alarcon-Ruiz CA, Melendez-Torres GJ, et al. Development of a framework for the implementation of synchronous digital mental health: realist synthesis of systematic reviews. JMIR Ment Health 2022 Mar 29;9(3):e34760. [doi: 10.2196/34760] [Medline: 35348469]
14. Minsa and universities work for mental health care for students [Web page in Spanish]. Peruvian Andean News Agency. 2019. URL: https://andina.pe/agencia/noticia-minsa-y-universidades-trabajan-cuidado-salud-mental-estudiantes-767385.aspx [accessed 2025-09-29]
15. Castro A, Gili M, Ricci-Cabello I, et al. Effectiveness and adherence of telephone-administered psychotherapy for depression: a systematic review and meta-analysis. J Affect Disord 2020 Jan 1;260:514-526. [doi: 10.1016/j.jad.2019.09.023] [Medline: 31539688]
16. Desai RJ, Franklin JM. Alternative approaches for confounding adjustment in observational studies using weighting based on the propensity score: a primer for practitioners. BMJ 2019 Oct 23;367:l5657. [doi: 10.1136/bmj.l5657] [Medline: 31645336]
17. de la Torre-Díez I, López-Coronado M, Vaca C, Aguado JS, de Castro C. Cost-utility and cost-effectiveness studies of telemedicine, electronic, and mobile health systems in the literature: a systematic review. Telemed J E Health 2015 Feb;21(2):81-85. [doi: 10.1089/tmj.2014.0053] [Medline: 25474190]
18. Malloy J, Partridge SR, Kemper JA, Braakhuis A, Roy R. Co-design of digital health interventions with young people: a scoping review. Digit Health 2023;9:20552076231219117. [doi: 10.1177/20552076231219117] [Medline: 38107986]
19. Jenaro C, Flores N, Cruz M, Moro L, Pérez C. Eficacia de los mensajes de texto para el cuidado de la salud en población mayor [Article in Spanish]. Gerokomos Idemm Farma, SL 2016 Jun;27(2):42-47 [FREE Full text]
20. Alva-Arroyo LL, Núñez del Prado Murillo J, Ancaya Martínez MDC, Floréz-Ibarra JM. Experiencias de telesalud en un hospital especializado en salud mental durante la pandemia de COVID-19 en Perú [Article in Spanish]. Rev Peru Med Exp Salud Publica 2021 Dec 13;38(4):653-659. [doi: 10.17843/rpmesp.2021.384.9220] [Medline: 35385020]
21. De la Cruz-Torralva K, Escobar-Agreda S, Riega López P, Amaro J, Reategui-Rivera CM, Rojas-Mezarina L. Assessment of a pilot program for remote support on mental health for young physicians in rural settings in peru: mixed methods study. JMIR Form Res 2024 Sep 10;8:e54005. [doi: 10.2196/54005] [Medline: 39255480]

XSL•FO
RenderX

22. Eysenbach G, Group CE. CONSORT-EHEALTH: improving and standardizing evaluation reports of Web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126. [doi: 10.2196/jmir.1923] [Medline: 22209829]
23. Amyga [Web page in Spanish]. Universidad Nacional Santiago Antúnez de Mayolo. URL: https://miamyga.unasam.edu.pe/ [accessed 2025-06-26]
24. Hayes SC, Strosahl KD, Wilson KG. Acceptance and Commitment Therapy: An Experiential Approach to Behavior Change: Guilford Press; 1999. URL: https://psycnet.apa.org/record/1999-04037-000 [accessed 2025-10-08]
25. Kabat-Zinn J. Full Catastrophe Living: Using the Wisdom of Your Body and Mind to Face Stress, Pain, and Illness, 15th Anniversary Ed: Delta Trade Paperback/Bantam Dell; 2005:471 URL: https://psycnet.apa.org/record/2006-04192-000 [accessed 2025-10-08]
26. Hayes SC, Wilson KG, Gifford EV, Follette VM, Strosahl K. Experiential avoidance and behavioral disorders: a functional dimensional approach to diagnosis and treatment. J Consult Clin Psychol 1996;64(6):1152-1168. [doi: 10.1037/0022-006X.64.6.1152] [Medline: 8991302]
27. Cehua Alvarez EA, Virú-Flores H, Alburqueque-Melgarejo J, et al. Validation of the Patient Health Questionnaire-9 (PHQ-9) in human medicine interns at a reference university in Peru during the COVID-19 pandemic. RFMH 2022 Jul 9;22(3):540-546. [doi: 10.25176/RFMH.v21i1.3179]
28. Anicama G. J, Ignacio C. F, Talla B. K, Bruno C. B. Propiedades psicométricas de la Escala del Trastorno de Ansiedad Generalizada: GAD-7 en universitarios de Lima, Perú [Article in Spanish]. ACPP 2021;6(2):260-273. [doi: 10.56891/acpp.v6i2.361]
29. Dominguez-Lara S, Merino-Soto C, Torres-Villalobos G. Structural analysis and reliability of the Perceived Stress Scale in nursing professionals from Peru. Enferm Clin 2022;32(3):152-160. [doi: 10.1016/j.enfcle.2022.01.001] [Medline: 35568353]
30. Shieh G. Power analysis and sample size planning in ANCOVA designs. Psychometrika 2020 Mar;85(1):101-120. [doi: 10.1007/s11336-019-09692-3] [Medline: 31823115]
31. Horwitz AG, Mills ED, Sen S, Bohnert ASB. Comparative effectiveness of three digital interventions for adults seeking psychiatric services: a randomized clinical trial. JAMA Netw Open 2024 Jul 1;7(7):e2422115. [doi: 10.1001/jamanetworkopen.2024.22115] [Medline: 39023893]
32. Mordkoff JT. A simple method for removing bias from a popular measure of standardized effect size: adjusted partial eta squared. Adv methods pract psychol sci 2019 Sep;2(3):228-232. [doi: 10.1177/2515245919855053]
33. WMA declaration of helsinki – ethical principles for medical research involving human participants. World Medical Association. 2024. URL: https://www.wma.net/policies-post/wma-declaration-of-helsinki/ [accessed 2024-12-10]
34. Davies EB, Morriss R, Glazebrook C. Computer-delivered and web-based interventions to improve depression, anxiety, and psychological well-being of university students: a systematic review and meta-analysis. J Med Internet Res 2014 May 16;16(5):e130. [doi: 10.2196/jmir.3142] [Medline: 24836465]
35. El Morr C, Ritvo P, Ahmad F, Moineddin R, Team MVC. Effectiveness of an 8-week web-based mindfulness virtual community intervention for university students on symptoms of stress, anxiety, and depression: randomized controlled trial. JMIR Ment Health 2020 Jul 17;7(7):e18595. [doi: 10.2196/18595] [Medline: 32554380]
36. Chen YY, Bakar SMA, Saimon R, Safii R. The use of online-based acceptance and commitment therapy (e-ACT) to improve the psychological wellbeing among university students. J College Stud Psychother 2024 Jan 2;38(1):105-118. [doi: 10.1080/87568225.2022.2134073]
37. Farrer LM, Jackson HM, Gulliver A, et al. A transdiagnostic video-based internet intervention (Uni Virtual Clinic-Lite) to improve the mental health of university students: randomized controlled trial. J Med Internet Res 2024 Aug 13;26(1):e53598. [doi: 10.2196/53598] [Medline: 39137012]
38. Wang Q, Zhang W, An S. A systematic review and meta-analysis of internet-based self-help interventions for mental health among adolescents and college students. Internet Interv 2023 Dec;34:100690. [doi: 10.1016/j.invent.2023.100690] [Medline: 38023965]
39. McCloud T, Jones R, Lewis G, Bell V, Tsakanikos E. Effectiveness of a mobile app intervention for anxiety and depression symptoms in university students: randomized controlled trial. JMIR Mhealth Uhealth 2020 Jul 31;8(7):e15418. [doi: 10.2196/15418] [Medline: 32735221]
40. Oballe VRR, Vite AZ, Chinga RJ. Digital competences in a Peruvian public university [Article in Spanish]. Conrado Universidad de Cienfuegos 2020 Dec;16(77):125-130 [FREE Full text]
41. Smith AC, Fowler LA, Graham AK, et al. Digital overload among college students: implications for mental health app use. Soc Sci (Basel) 2021;10(8):279. [doi: 10.3390/socsci10080279] [Medline: 34305199]
42. Wanniarachchi VU, Greenhalgh C, Choi A, Warren JR. Personalization variables in digital mental health interventions for depression and anxiety in adolescents and youth: a scoping review. Front Digit Health 2025;7:1500220. [doi: 10.3389/fdgth.2025.1500220] [Medline: 40444184]
43. Nicolaidou I, Aristeidis L, Lambrinos L. A gamified app for supporting undergraduate students' mental health: a feasibility and usability study. Digit Health 2022;8:20552076221109059. [doi: 10.1177/20552076221109059] [Medline: 35756831]

XSL•FO
RenderX

44. Yosep I, Suryani S, Mediani HS, Mardhiyah A, Maulana I. Digital therapy: alleviating anxiety and depression in adolescent students during COVID-19 online learning - a scoping review. J Multidiscip Healthc 2023;16:1705-1719. [doi: 10.2147/JMDH.S416424] [Medline: 37366385]
45. Conley CS, Durlak JA, Shapiro JB, Kirsch AC, Zahniser E. A meta-analysis of the impact of universal and indicated preventive technology-delivered interventions for higher education students. Prev Sci 2016 Aug;17(6):659-678. [doi: 10.1007/s11121-016-0662-3] [Medline: 27225631]
46. Harrer M, Adam SH, Baumeister H, et al. Internet interventions for mental health in university students: a systematic review and meta-analysis. Int J Methods Psychiatr Res 2019 Jun;28(2):e1759. [doi: 10.1002/mpr.1759] [Medline: 30585363]
47. Gieselmann A, Pietrowsky R. Treating procrastination chat-based versus face-to-face: an RCT evaluating the role of self-disclosure and perceived counselor's characteristics. Comput Human Behav 2016 Jan;54:444-452. [doi: 10.1016/j.chb.2015.08.027]
48. Kim J, Aryee LMD, Bang H, et al. Effectiveness of digital mental health tools to reduce depressive and anxiety symptoms in low- and middle-income countries: systematic review and meta-analysis. JMIR Ment Health 2023 Mar 20;10:e43066. [doi: 10.2196/43066] [Medline: 36939820]
49. Ebert DD, Lehr D, Boß L, et al. Efficacy of an internet-based problem-solving training for teachers: results of a randomized controlled trial. Scand J Work Environ Health 2014 Nov;40(6):582-596. [doi: 10.5271/sjweh.3449] [Medline: 25121986]
50. Dominguez-Rodriguez A, Herdoiza-Arroyo PE, González-Ramírez LP, et al. Internet-based self-administered intervention to reduce anxiety and depression symptomatology and improve well-being in 7 countries: protocol of a randomized control trial. Front Psychol 2024;15:1279847. [doi: 10.3389/fpsyg.2024.1279847] [Medline: 38774723]

Abbreviations

ACT: acceptance and commitment therapy
ANCOVA: analysis of covariance
CBT: cognitive-behavioral psychotherapy
CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
CSQ-8: Client Satisfaction Questionnaire-8
CSUQ: Computer System Usability Questionnaire
DMHI: digital mental health intervention
DMHSS: digital mental health self-care service
GAD-7: Generalized Anxiety Disorder-7
PHQ-9: Patient Health Questionnaire-9
PSS-10: Perceived Stress Scale-10
RCT: randomized controlled trial
TWEETS: Twente Engagement with Ehealth Technologies Scale

Edited by A Mavragani; submitted 19.01.25; peer-reviewed by M Kavurmaci, M Yusuff, S Huang; revised version received 27.06.25; accepted 30.06.25; published 15.10.25.

Please cite as:
Yslado-Méndez R, Escobar-Agreda S, Villarreal-Zegarra D, Trejo Flores WM, Sánchez-Broncano JD, Vilela-Estrada AL, Olivares Córdova JH, Reategui-Rivera CM, Alvarez-Yslado C, Rojas-Mezarina L
Effectiveness, Usability, and Satisfaction of a Self-Administered Digital Intervention for Reducing Depression, Anxiety, and Stress in a University Community in the Andean Region of Peru: Randomized Controlled Trial
JMIR Form Res 2025;9:e71465
URL: https://formative.jmir.org/2025/1/e71465
doi:10.2196/71465

© Rosario Yslado-Méndez, Stefan Escobar-Agreda, David Villarreal-Zegarra, Wilfredo Manuel Trejo Flores, Junior Duberli Sánchez-Broncano, Ana Lucia Vilela-Estrada, Jovanna Hasel Olivares Córdova, C. Mahony Reategui-Rivera, Claudia Alvarez-Yslado, Leonardo Rojas-Mezarina. Originally published in JMIR Formative Research (https://formative.jmir.org), 15.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

A Remotely Delivered Weight Management Service to Support Existing Obesity Services in the UK National Health Service: Preliminary Findings From an Early-Stage Service Evaluation

Giulia Spaltro*, PsyD; Michael Whitman*, BSc; Rebecca Richards*, PhD

Second Nature Healthy Habits Ltd, 58 Wood Lane, London, United Kingdom

*all authors contributed equally

Corresponding Author:
Giulia Spaltro, PsyD
Second Nature Healthy Habits Ltd, 58 Wood Lane, London, United Kingdom

Abstract

Background: Specialist weight management services (SWMSs) in the UK National Health Service (NHS) face long waiting lists and limited resources. Remotely delivered SWMSs may support existing NHS services to increase access to treatment for patients living with obesity; however, evidence of remotely delivered services working to support NHS SWMSs in practice remains limited.

Objective: This study aims to explore the potential effectiveness, feasibility, and acceptability of Second Nature's remotely delivered SWMS for adults living with obesity referred from existing NHS SWMSs. Preliminary findings from the first phase (Preparing for weight loss) of a 3-phase remotely delivered SWMS are presented.

Methods: A total of 39 adults (age range 23 - 74 years, mean age 45.6, SD 12.1; 74% female) completed a 16-week intervention, following referral from NHS SWMS leads. Eligible participants were assessed by a multidisciplinary team and allocated to one of three interventions: (1) a psychologically informed app-based intervention, (2) a Dialectical Behavioral Therapy (DBT)-based skills training group intervention, and (3) one-to-one psychological support. The primary outcomes were weight change (kg) and percentage weight change following completion of the intervention. Secondary outcomes included psychological distress, emotional eating, health-related quality of life, physical activity, emotion regulation, intervention feasibility, and acceptability.

Results: At 16 weeks, the mean weight change was −2.2 kg (SD 5.16), or −1.6% of body weight. Participants in the app-based intervention lost the most weight (−2.8kg), and participants in the one-to-one psychological support intervention lost the least weight (−1.3kg). Psychological distress was reduced to below the clinical threshold (mean score 0.95, SD 0.62). Emotional eating behaviors and difficulties in emotion regulation also decreased (mean change scores −3.2, SD 6.4 and mean −11.6, SD 13.9, respectively). Health-related quality of life saw improvements in self-care, usual activities, and anxiety and depression, while participants' challenges with mobility, and pain and discomfort remained unaffected. Subjective ratings of health status improved by 17.4%. There were no significant changes in physical activity levels, with most participants remaining "Inactive" or "Moderately inactive." Engagement with intervention sessions was high (93.7%) and the attrition rate was 27.4%. Participants rated their satisfaction with the intervention at 9/10 and highlighted key benefits, including improved mental well-being, healthier habits, and supportive coach relationships. Suggested improvements included greater scheduling flexibility, enhanced app functionality, and more accessible physical activity support.

Conclusions: This preliminary service evaluation suggests that a remotely delivered SWMS has the potential to be effective, feasible, and acceptable for NHS-referred patients in the United Kingdom. Changes observed across several key measures point to clinically significant benefits, reinforcing the potential of this approach. A full evaluation of all 3 phases of this service with a larger sample size is required to support these early findings.

(JMIR Form Res 2025;9:e71914) doi:10.2196/71914

KEYWORDS

obesity; diabetes; weight loss; weight management; digital health; digital health intervention; multidisciplinary; behaviour change; disordered eating; dialectical behavioral therapy

XSL•FO
RenderX

Introduction

Background

Around 1 in 3 adults in the United Kingdom are affected by obesity [1-4] (defined as abnormal or excessive fat accumulation that presents a risk to health, with a BMI of ≥30 [1]). Many people living with obesity are also burdened by weight-related conditions such as cardiovascular disease, type 2 diabetes, liver disease, and sleep apnea and are at risk of many cancers [1,5]. In addition to poorer physical health, obesity is associated with poorer mental health [6-8] and lower quality of life [9]. Due to the wide-ranging negative impact on the health and well-being of people living with obesity, comprehensive treatment that considers the biological, psychological, and social determinants of obesity is imperative. This approach aligns with recent guidance from an international expert working obesity group which suggested the outcomes of treatment for clinical obesity should focus on easing the symptoms of obesity, rather than focusing on weight loss [10].

Individuals with a BMI ≥30 and weight-related comorbidities are eligible for Specialist Weight Management Services (SWMS) in the UK National Health Service (NHS) [11], where specialist assessment, monitoring, and comprehensive tailored treatment by a physician-led multidisciplinary team (MDT) is provided [12-14]. An MDT typically includes a physician, nurse, dietitian, psychologist, and a physiotherapist or exercise therapist, each with a specialist interest in obesity. Treatment in SWMSs is typically tailored to the patient's presentation but can involve support to make dietary changes, increase physical activity, and psychological support to overcome psychological barriers to weight loss, as well as pharmacotherapy where required or accessible [15]. However, due to the comprehensive and tailored treatment offered, SWMSs are associated with a high financial cost [16]. In addition, there appears to be wide variation in treatment protocols and MDT structures within and between SWMSs, which may be due to funding limitations [16]. Nevertheless, research suggests that SWMSs in the United Kingdom can be an effective obesity treatment [17,18], though evidence is still limited [19,20]. For example, a systematic review of 19 studies of SWMSs in the United Kingdom reported positive effects on weight (specifically, 43.4%% and 29.4% achieved ≥5% and 10% weight loss, respectively), BMI, glycemic control, blood pressure, and physical activity at 12 months as a result of the comprehensive and tailored treatment offered [17].

Due to the high cost associated with traditional in-person SWMS and funding limitations, the provision of, and access to, SWMSs across the United Kingdom remains limited and varies geographically [18,19,21,22], and treatment can last anywhere between 4 and 24 months [17]. Furthermore, services face long waiting lists, as well as a lack of treatment continuity and flexibility [17,21]. Overall, these barriers can result in suboptimal protocols, treatment delays, and ultimately, adversely affect patient outcomes [17,22,23].

Digital weight management interventions (DWMIs) are promising solutions for addressing gaps and issues in the current service provision of traditional SWMSs that historically have been provided in person [18,24-26]. Increased access to services, increased convenience, more frequent care, resource- and cost-savings, and scalability potential are among the benefits of DWMIs to support people living with obesity and related conditions [18,27,28]. Previous studies, including 3 systematic reviews, have shown that DWMIs can be as effective as in-person interventions for weight loss and related outcomes for people with obesity [26,29-32], including within the context of SWMS [18,33,34]. For example, preliminary findings from a service evaluation of a remotely delivered, semaglutide-supported SWMS for adults living with obesity further suggested DWMIs have the potential to be effective, feasible, and acceptable for self-paying consumer adults with obesity in the United Kingdom [35]. In addition, a dietetic weight loss app-based program was found to be as effective and feasible when delivered remotely from a hospital-based SWMS to their usual face-to-face care [36]. Furthermore, a recent retrospective analysis of a 12-month app-based DWMI for 1130 patients living with obesity (with or without type 2 diabetes), referred from NHS primary care services, showed a statistically significant mean percentage weight loss of 9.1% (SD 9.6%), at 1 year and 11.8% (SD 10.9%), at 2 years [28].

Despite these early findings, real-world evidence of the potential for DWMIs to support SWMS in the UK NHS remains limited [37]. Further research is still needed to assess the potential of a remotely delivered SWMS, which includes assessment and triaging to the appropriate interventions, as well as integration with the referring clinical teams in the NHS.

This Study

To build on the growing real-world evidence base, this study explored the potential of Second Nature's remotely delivered pilot SWMS to support existing services in the UK NHS to increase access to treatment for people living with obesity. This pilot service evaluation reports on the preliminary findings of the potential effectiveness, feasibility, and acceptability of the first intervention phase of a 3-phase remotely delivered SWMS for patients living with obesity referred from NHS SWMSs.

Methods

Overview

This manuscript was prepared in accordance with the SQUIRE 2.0 (Standards for Quality Improvement Reporting Excellence, Version 2) guidelines [38]. A completed checklist is included in Checklist 1 to support comprehensive reporting of this service evaluation.

This service evaluation used a retrospective analysis of preliminary data from participants referred to Second Nature Ltd by the NHS. The study only involved the use of previously collected, anonymized data that could not be traced back to identifiable individuals. As part of their referral, participants were required to provide informed consent for their anonymized data to be collected for research purposes, including analysis and publication. A privacy policy detailing data collection, storage, and processing is hosted on the Second Nature website and participants were invited to provide informed consent to the policy upon enrollment in the intervention [39].

XSL•FO
RenderX

All collected data were stored on a cloud-based encrypted storage service and accessed via company-owned password-protected devices. For the purpose of this evaluation, participant data were deidentified and pseudonymized using identification numbers. As per the General Data Protection Regulations (GDPR), participants could request to have their information deleted at any time [40].

Participants

Participants included in this analysis were screened and referred via secure NHS email to Second Nature by their local NHS SWMS lead for weight management support. Participants were referred to Second Nature if they were deemed clinically suitable for the service by the referrer, in line with our inclusion and exclusion criteria (Textbox 1). A second screening process was conducted by a patient coordinator at Second Nature to confirm the eligibility of referrals.

Textbox 1. Service inclusion and exclusion criteria.

Inclusion criteria

- Over 18 years old
- BMI ≥35 in the presence of diabetes or other weight-related conditions (lowered to BMI ≥30 if screened and referred specifically by NHS SWMS service leads)
- Or BMI ≥40 without weight-related conditions (reduced by 2.5 kg/m² of BMI in Black African, African Caribbean, and Asian groups)
- The underlying causes of overweight or obesity need to be assessed
- The person has complex disease states or needs that cannot be managed adequately in lower-intensity weight management interventions
- Conventional treatment has been unsuccessful
- Specialist interventions (such as a very-low-calorie diet) may be needed
- Drug treatment is being considered for a person with a BMI ≥50 kg/m²
- Bariatric surgery is being considered
- Access to a digital device and comfortable using technology

Exclusion criteria

- Current or planned pregnancy
- An unstable condition that does not warrant weight management at present
- Recent complicated myocardial infarction or stroke, or awaiting further investigation
- Blood pressure at rest above 180 mm Hg systolic, 120 mm Hg diastolic
- Alcohol or drug misuse
- Diagnosed, active eating disorder or presenting with active purging behaviors (eg, laxative misuse, self-induced vomiting in the absence of a diagnosed eating disorder)
- Undergoing active cancer treatment
- Acute infection
- Unstable major psychiatric illness
- Recent bariatric surgery (within the last 2 years)

Intervention Description

Second Nature's remotely delivered SWMS consisted of three phases: (1) "Preparing for weight loss" (16 weeks), (2) "Active weight loss" (12 weeks), and (3) "Weight loss maintenance" (24 weeks). Participants were assessed at baseline and at completion of each of the subsequent intervention phases. In this study, we report outcomes for the first 39 participants who completed phase 1 ("Preparing for weight loss"). A further 6 participants took part in the intervention; however, they were not included in this evaluation as they had not yet completed it at the time of analysis.

The aim of phase 1 was to prepare participants for a future behavioral weight loss intervention (phase 2) by reducing psychological barriers to making changes to their diet and lifestyle, including psychological distress, emotional eating, and disordered eating behaviors [41-45].

Participants were triaged to one of 3 interventions based on a comprehensive assessment process which included validated questionnaires and a 50-minute video consultation with a health care practitioner (either a practitioner psychologist or health coach). Ahead of each initial consultation, participants with a risk score >0 (CORE-OM) were contacted by the safeguarding team for a risk assessment to ensure patient safety and suitability for the intervention.

Participants with a mean psychological distress score ≥1 were allocated to a practitioner psychologist for their initial video consultation, while participants with a score <1 had their initial

XSL•FO
RenderX

consultation with a health coach. This decision was informed by the scoring guidelines of the scale selected for measuring psychological distress (CORE-OM).

Intervention allocation was informed by initial assessment outcomes and subsequently reviewed by a MDT. The MDT consisted of a physician, 2 practitioner psychologists, a dietitian and exercise specialist, 2 nutritionists (supervised by a dietitian), and a patient coordinator.

Interventions included: (1) a psychologically informed app-based intervention, (2) a Dialectical Behavioral Therapy (DBT)-based skills training group intervention, and (3) one-to-one psychological support. Each intervention was delivered remotely and lasted 16 weeks.

Psychologically Informed App-Based Intervention

Participants considered to have mild to moderate emotional eating or weight-related psychological distress were allocated to this intervention. The intervention consisted of personalized and modularized app-based content covering topics such as nutrition for weight loss, the psychology of eating, challenging all-or-nothing thinking, the science of habits, and mindful eating [41,42,46]. Module content was informed by the COM-B model of behavior change and the Behaviour Change Wheel [47]. This intervention included behavior change techniques [48] shown in previous studies to be associated with improved outcomes in diet and physical activity and recommended by the National Institute for Health and Care Excellence (NICE) guidelines for individual approaches to behavior change [49].

Participants were supported through the modules with 8 fortnightly video calls (30 - 40 minutes) with their assigned health coach (registered or associate nutritionist, supervised by a registered dietitian). The assigned health coach tailored the order of modules to each participant's presentation. On weeks where there were no calls, the health coach conducted a structured in-app text-based check-in using a one-to-one private chat feature. Participants were guided to reflect on the content of the educational modules, explore barriers to change, set goals, practice new skills, and develop a personalized plan to support realistic and sustainable lifestyle changes.

DBT-Based Skills Training Group Intervention

Participants who were considered to have moderate to high levels of emotional eating, or moderate to high levels of weight-related psychological distress, or symptoms of binge eating, were allocated to this intervention. The aim of this intervention was to reduce difficulties with emotion regulation, emotional eating, and disordered eating behaviors and was informed by Safer et al's manual [50]. DBT, originally developed to treat borderline personality disorder [51], has been shown to be an effective treatment for binge eating disorder [52-55], at comparable levels of clinically meaningful change in global eating disorder psychopathology to cognitive behavioral therapy (CBT) [53,56]. Furthermore, there is preliminary evidence to suggest that DBT-based skills training groups can decrease binge eating severity and emotion regulation difficulties in individuals living with obesity without binge eating disorder [57,58].

In this intervention, participants were supported to build awareness of, and practical experience in, the 4 pillars of DBT: mindfulness, emotion regulation, distress tolerance, and interpersonal effectiveness, through weekly group video sessions (1.5 hours each) via Zoom (Zoom Communications Inc) led by a practitioner psychologist. Following each group session, participants were emailed a copy of the educational materials used during the sessions and were provided with digital access to a worksheet to complete before the next group session. Participants were also posted paper copies of the worksheets to their home address before the intervention start date, in case they preferred to use printed worksheets. Each participant also received weekly one-to-one telephone check-in calls with the practitioner psychologist (15 minutes duration) to support skills practice between sessions and provide an opportunity to reflect on the content covered in group sessions.

One-to-One Psychological Support

Participants considered to have moderate to high levels of emotional eating, or moderate to high levels of weight-related psychological distress, or symptoms of binge eating, but who did not wish to participate in a group intervention or presented with psychological issues that would benefit from an alternative psychological intervention, were allocated to this intervention. The therapeutic intervention that participants received depended on their presentation and on the practitioner psychologist's bespoke formulation. Third wave CBTs (eg, Acceptance and Commitment Therapy, Dialectical Behavioral Therapy, and Compassion Focused Therapy) were among the commonly used evidence-based interventions [57,59-71]. This intervention included around 10 one-to-one video consultations (50 minutes, range 7 - 13) via Zoom (Zoom Communications Inc) with a practitioner psychologist. Following each session, any relevant educational materials or worksheets to complete between sessions were emailed to the participant.

Physical Activity Support

All participants were offered the opportunity to discuss their physical activity support needs with a physical activity specialist. This involved a one-to-one telephone consultation (30 minutes) to discuss baseline physical activity levels and set specific, achievable, and realistic goals to increase physical activity. Examples of individually tailored goals include "To go out for a walk 3x per week, for 15 min," "To walk 1000 steps, every day," or "To introduce an at-home seated activity session 2 - 3 times per week." Following this consultation, participants were invited to attend weekly live group physical activity classes (30 minutes) via Zoom, facilitated by a physical activity associate (supervised by the physical activity specialist). A follow-up consultation to review progress was offered to all participants who completed phase 1 of the service.

Health Care Professional Training

Health coaches delivering the psychologically informed app-based intervention included a registered nutritionist and an associate nutritionist, both registered with the UK Voluntary Register of Nutritionists (UKVRN), with experience of working with complex weight management patients in clinical settings. Health coaches were supervised by a dietitian registered with

XSL•FO
RenderX

the Health and Care Professions Council (HCPC). Health coaches completed a comprehensive training intervention, which was delivered by a dietitian and practitioner psychologists with extensive experience in specialist weight management, to ensure high-quality care in line with professional guidelines and clinical protocols. Health coaches also attended weekly clinical supervision led by a dietitian and supported by a practitioner psychologist.

Practitioner psychologists were Health Psychologists registered with the HCPC, both with extensive experience of working in specialist weight management and a special interest in third wave CBTs. Each practitioner psychologist received weekly or monthly clinical supervision with a more senior practitioner psychologist, depending on experience.

The physical activity associate (level 2 qualification for gym instructors) was supervised by the physical activity specialist (level 3 personal trainer) during live physical activity sessions. Both professionals were registered with the Chartered Institute for the Management of Sport and Physical Activity (CIMSPA).

Data Collection and Measures

Overview

Primary outcomes were weight change (kg) and percentage weight change at postintervention. Secondary outcomes included psychological distress, emotional eating, health-related quality of life, physical activity, intervention feasibility, and acceptability. For those who took part in the group-based intervention, an additional secondary outcome included emotion regulation.

Baseline characteristics (weight, height, age, gender, and type 2 diabetes diagnosis) and contact details (including postcode) were collected by the participant's referrer and emailed securely via an NHS email service to Second Nature. Data were entered into Second Nature's referral management system, and participants were sent an email link to complete a series of validated questionnaires about their current psychological distress, eating behaviors, health-related quality of life, and physical activity. Postcode data were also collected during onboarding to calculate socioeconomic deprivation based on the index of multiple deprivation (IMD). At the completion of the 16-week interventions, participants were sent another email link to complete postintervention questionnaires to measure intervention efficacy.

Effectiveness

Weight Change

Participants' baseline weight (kg) and height (m) were recorded by the referring NHS SWMS on a referral form. Postintervention weight was self-reported via a questionnaire. No other weight data were collected during the intervention. Weight change data were calculated by subtracting postintervention weight from baseline weight, and participants were assigned to weight change categories of "lost," "maintained," and "gained." Percentage weight loss was computed as: (weight lost or baseline weight) x100. Participant baseline and 16-week BMI was calculated with the formula BMI=kg/m^2.

Psychological Distress

The Clinical Outcomes in Routine Evaluation–Outcome Measure (CORE-OM) [72] was used to evaluate psychological distress and consisted of 4 domains (subjective well-being, problems or symptoms, life functioning, and risk or harm) measured using 34 items. Each item was scored on a 0 - 4 Likert scale, with higher scores indicating greater distress. Scores for each CORE-OM domain were calculated by summing the relevant item responses, yielding total scores that reflect cumulative distress within each area. Internal consistency of the CORE–OM has been reported as α=.94 [73].

Emotional Eating

The Emotional Eater Questionnaire (EEQ) [74] was used to assess emotional eating using 10 items, divided into 3 domains: disinhibition (loss of control over eating), types of food consumed in emotional situations, and guilt related to eating behaviors. Items are scored on a 4-point scale (0=Never, 3=Always), resulting in a total score ranging from 0 - 30. Scores were interpreted as: nonemotional eater (0 - 5), low emotional eater (6-10), emotional eater (11-20), very emotional eater (21-30). Internal consistency of the EEQ, within the context of obesity specifically, has been reported as α=.64 [74].

Health-Related Quality of Life

The EuroQol 5 Dimensions, 5 Levels (EQ-5D-5L) [75] was used to measure health-related quality of life through 5 dimensions: mobility, self-care, usual activities, pain and discomfort, and anxiety and depression. Each domain was rated out of 5 levels (1=no problems, 5=extreme problems). Responses generate a 5-digit health state code (eg, 11,111 for full health across the 5 domains). A self-rated measure of current perceived health on a scale of 0 (worst imaginable health) to 100 (best imaginable health) is also included in the form of a Visual Analog Scale (VAS). The internal consistency of the EQ-5D-5L has been reported as α=.79 [76].

Physical Activity

The General Practice Physical Activity Questionnaire (GPPAQ) [77] was used to measure and categorize physical activity into 4 levels (Active, Moderately Active, Moderately Inactive, and Inactive) based on 3 items: occupational activity, weekly physical activity, and walking pace. The range of internal consistency of the GPPAQ has been reported as α=.86 - 0.97 [78]. Participants scoring below "Active" are recommended for a brief intervention to increase physical activity, in accordance with NICE guidelines [79].

Emotion Regulation

The Difficulties in Emotion Regulation Scale (DERS-16) [80] measures emotion regulation difficulties across 5 domains (nonacceptance of negative emotions, inability to engage in goal-directed behaviors when distressed, difficulties controlling impulsive behaviors when distressed, limited access to emotion regulation strategies perceived as effective, and lack of emotional clarity). Total scores range from 16 - 80, with higher scores indicating greater difficulties in emotion regulation. This measure was completed only by participants allocated to the DBT-based skills training group intervention, as one of the aims of this intervention was to improve emotion regulation. The 16

XSL•FO
RenderX

items are scored on a 1 - 5 scale (1=Almost never, 5=Almost always), producing a total score from 16 - 80. Higher scores indicate greater difficulties with emotion regulation. The internal consistency of the DERS-16 has been reported as α=.92 [80].

Feasibility

Retention and Discharge Reasons

Participants were discharged if they requested to withdraw or disengaged from the intervention (judged on a case-by-case basis) or no longer met the eligibility criteria (eg, pregnancy disclosure and cancer diagnosis). Reasons for discharge were recorded in a referral management system and leads of the referring SWMSs were informed of discharges.

Engagement

Across all 3 interventions, engagement was primarily measured through session attendance, including one-to-one and group sessions, and scheduled check-ins. In the app-based intervention, home screen views were used as a proxy for minimal digital engagement, reflecting user-initiated interaction with the app's central hub. While this is a broad measure, it captures instances where participants accessed the platform to engage with key features such as modules, goals, or coach messages.

Acceptability

A mixed-methods questionnaire was sent to all participants upon completion of their intervention to evaluate intervention acceptability and inform future service improvement. The questionnaire measured patient satisfaction ratings in relation to overall intervention experience on a 10-point Likert scale (1=Extremely dissatisfied, 10=Extremely satisfied), as well as providing free-text response options to qualitatively explore perceived benefits and challenges (What, if any, were the benefits and challenges of the program for you?), most and least useful program components (Which elements of the program did you find most and least useful?), self-reported behavior changes (What changes, if any, have you made as a result of the program?), and suggestions for service improvement (Was there anything you felt was missing from the overall program? and Based on your experience, how can we improve the program for future participants?).

Statistical Analysis

Statistical analysis was carried out using IBM SPSS Statistics (V.29). Descriptive statistics were used to analyze participant demographics, which were reported in mean, SD, and range for continuous variables and in number and percentages for categorical variables. Descriptive statistics were used to assess feasibility measures. The attrition rate was calculated by dividing the number of noncompleters by the total number of referrals received, multiplied by 100, and missing data were managed through listwise deletion [81]. Normality of data distribution was assessed by performing Shapiro-Wilk tests (recommended for sample size <50 [82]) and Q-Q plots for each outcome variable, split by intervention allocation.

Within-group analyses on mean change from baseline to 16-week follow-up were used to analyze the potential effectiveness of interventions. Two-tailed paired-samples *t* tests were used for continuous variables and McNemar nonparametric tests were used for categorical variables. ANOVA tests were performed to evaluate between-group differences between interventions at baseline and at 16 weeks. Parametric and nonparametric correlational analysis was used to investigate relationships between outcome variables.

A priori power analysis for sample size estimation was carried out using G*Power (v.3.1; Heinrich Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang) software [83] for both a paired-samples *t* test (for within-group weight change) and a one-way ANOVA (for between-group comparisons). With a significance criterion of α=.05 and 95% power for detecting a medium effect (d=0.5; *f*=0.25), the minimum sample size needed for a size was n=36. The calculation was based on the primary outcome of weight change at 16 weeks and was in line with Cohen benchmarks [84] and supported by prior meta-analyses of behavioral weight management interventions [85]. We acknowledge that smaller effects (d=0.3 - 0.4) may be present in secondary outcomes, and that our sample size may limit power to detect these.

Qualitative Analysis

Content analysis was conducted to analyze questions with free-text responses on the mixed-methods postintervention questionnaire. This approach involved categorizing free-text data into themes through a systematic process of coding and interpretation [86]. For each question, the responses were coded and organized into higher-level categories by GS. For example, perceived benefits were characterized by explicit benefits or positive outcomes related to the intervention, such as reduced binge eating episodes or decreased snacking frequency. Participants' responses were organized into more than one category, where appropriate (eg, least useful intervention component and suggestion for service improvement). The codes and categories were then discussed with RR and refined. Finally, the frequency of occurrence of each category was calculated. Example quotes from participants are presented to illustrate key themes.

Ethical Considerations

The Health Research Authority (HRA) decision tool was used to determine the nature of this study (Multimedia Appendix 1) [40]. This study met the criteria for service evaluations, as outlined by the HRA and Medical Research Council (1) it was designed and conducted solely to assess current care; (2) it evaluated a current service without reference to a standard; (3) it involved an intervention already in use; (4) it involved analysis of existing data; (5) there was no allocation to the intervention; and (6) there was no randomization, and therefore it did not require ethical approval (Multimedia Appendix 1) [40]. The data analyzed in this evaluation were not collected with the intent to conduct research, but to assess and improve service delivery. Therefore, this project does not constitute research under UK policy. The HRA explicitly states that "because both audit and service evaluation are considered part of usual professional practice, they are exempt from oversight processes that govern research, that is, they do not require review by a research ethics committee (REC)" [87]. As such, NHS REC/IRB review was not required for this study.

XSL•FO
RenderX

Results

Baseline Characteristics

A total of 39 participants were included in this evaluation (Table 1), of which 82.1% (n=32) were female, with a mean BMI of 46.8 kg/m^2 (SD 9.04). Only 3 participants (7.7%) had a diagnosis of type 2 diabetes.

Table . Baseline characteristics of intervention participants.

Baseline sample characteristics, (N=39)	Values
Age (years), mean (SD, range)	45.7 (11.7, 26 - 68)
Female sex, n (%)	32 (82.1)
Weight (kg), mean (SD)	132.8 (27.6)
BMI (kg/m^2), mean (SD)	46.8 (9.0)
Presence of type 2 diabetes, n (%)	3 (7.7)
IMD[a], n (%)	
1 (most deprived)	0 (0)
2	7 (17.9)
3	16 (41.0)
4	13 (33.3)
5 (least deprived)	2 (5.1)

[a]IMD: index of multiple deprivation

Intervention Allocation

Following the initial assessment process, participants were allocated to one of 3 interventions. Almost half (19/39, 48.7%) of participants were allocated to the app-based intervention, while participants allocated to the psychological interventions were equally split between group-based (10/39, 25.6%) and one-to-one (10/39, 25.6%).

Effectiveness

Weight Change

There was a statistically significant difference in total sample weight change at 16 weeks (t_{38}=2.64; P=.01) with a medium effect size (Cohen d=0.42), but no significant change in BMI (t_{38}=0.85; P=.40). Average weight change postintervention was −2.2kg (−1.6%, SD 5.16). Mean baseline and change scores at 16 weeks for weight and BMI are reported in Table 2, including differences between interventions.

Participants in the app-based intervention lost the most weight at postintervention (−2.8kg), and participants receiving one-to-one psychological support lost the least weight (−1.3kg), although no significant between-group differences in weight and BMI were observed at baseline ($F_{2,36}$=.64; P=.53 and $F_{2,36}$=.39, P=.68 respectively) or at postintervention ($F_{2,36}$=.81; P=.45 and $F_{2,36}$=.51; P=.61, respectively).

A total of 46.2% (n=18) of participants lost weight, 35.9% (n=14) maintained baseline weight, and 17.9% (n=7) gained weight at postintervention. No significant correlations were observed between weight changes at post-intervention and changes in any of the other secondary outcome variables, with the exception of the EEQ, where weight loss was significantly positively correlated with a reduction in emotional eating behavior (r_{39}=.347, P=.03).

XSL•FO
RenderX

Table . Average weight and BMI change at 16 weeks.

Intervention	Baseline		16-week	
	Weight (kg), mean (SD)	BMI (kg/m^2), mean (SD)	Weight change (kg), mean (%, SD)	BMI change (kg/m^2), mean (SD)
Psychologically informed, app-based intervention	128.0 (25.0)	45.5 (6.93)	−2.8 (-2.2, 4.74)	−0.9 (3.27)
DBT-based[a] skills training group intervention	134.4 (33.2)	48.3 (12.0)	−1.9 (-1.4, 6.30)	−0.7 (2.36)
One-to-one psychological support	140.1 (27.5)	47.9 (9.85)	−1.3 (-0.9, 5.11)	−0.5 (2.51)
Total (N=39)	132.7 (27.6)	46.8 (9.04)	−2.2 (-1.7, 5.16)[b]	−0.39 (2.89)

[a]DBT: Dialectical Behavioral Therapy.

[b]$P<.05$

Psychological Distress

At baseline, the total sample mean score was 1.4 (SD 0.68), where 71.8% (n=28) of participants scored above the threshold (mean score=1) for clinically significant psychological distress. The correlation between psychological distress and baseline weight approached statistical significance (r_{39}=.311; P=.05). At 16 weeks, 46.2% (n=18) of participants scored above the threshold and the total sample mean score was 0.95 (SD 0.62). This reduction (−25.6%) was statistically significant (t_{38}=5.35; P<.001), with a large effect size (Cohen d=0.86) (). There were also significant improvements in individual CORE-OM domains of subjective well-being (t_{38}=5.09; P<.001), problems and symptoms (t_{38}=4.94; P<.001), and life functioning (t_{38}=5.18; P<.001). Mean risk or harm domain scores decreased by −0.36 (SD 1.55) at 16 weeks; however, this was not statistically significant (t_{38}=1.45, P=16).

Participants allocated to the DBT-based group intervention and one-to-one psychological support showed higher baseline CORE-OM mean scores (mean 1.66, SD 0.59 and mean 1.59, SD 0.74, respectively) than those allocated to the app-based intervention (mean 1.14, SD 0.64); however, these differences were not statistically significant (P=.08). Similarly, there were no significant between-group differences in CORE-OM mean scores at 16 weeks (P=.35) , nor in the domains of subjective well-being (P=.40), problems and symptoms (P=.17), life functioning (P=.52) or risk or harm (P=.65) (Multimedia Appendix 2).

Emotional Eating

At baseline, 38.5% (n=15) of participants were categorized as “Very emotional eaters”, 41% (n=16) as “Emotional eaters,” 17.9% (n=7) as “Low emotional eaters”, and 2.6% (n=1) as “Nonemotional eaters”. At 16 weeks, a reduction in emotional eating severity was observed, with fewer participants categorized as “Very emotional eaters” (6/39, 15.4%) and “Emotional eaters” (14/39, 35.9%), and more participants categorized as “Low emotional eater” (15/39, 38.5%) and “Nonemotional eaters” (4/39, 10.3%).

The mean EEQ score at baseline was 17.6 (SD 7.20), which reduced to 12.3 (SD 6.81) at 16 weeks (Emotional eater category range 11 - 20). This reduction in emotional eating behavior was statistically significant (t_{38}=5.21; P<.001), with a large effect size (Cohen d=0.83).

Table 3 shows the breakdown of EEQ score changes by intervention. Participants in the DBT-based group intervention displayed the highest levels of emotional eating, and participants in the app-based intervention displayed the lowest. These between-group differences were statistically significant at both baseline ($F_{2,36}$=10.42; P<.001) and 16-week ($F_{2,36}$=8.48; P<.001). The greatest improvement in emotional eating behaviors was seen in the app-based intervention (t_{18}=3.89; P=.001) at 16 weeks, while there was no significant change observed for those allocated to one-to-one psychological support (P=.13).

XSL•FO
RenderX

Table . Average emotional eating scores at baseline and 16-week, with EEQ score categorization.

Intervention	Baseline, mean (SD)	EEQ[a] score category	16-week, mean (SD)	EEQ score category
Psychologically informed, app-based intervention	13.74 (6.06)	Emotional eater	8.53 (4.25)[b]	Low emotional eater
DBT-based[c] skills training group intervention	24.2 (5.43)	Very emotional eater	17.3 (7.35)[d]	Emotional eater
One-to-one psychological support	18.2 (5.98)	Emotional eater	14.4 (6.57)	Emotional eater
Total (N=39)	17.6 (7.20)	Emotional eater	12.3 (6.81)[b]	Emotional eater

[a]EEQ: Emotional Eating Questionnaire

[b]$P<.001$

[c]DBT: Dialectical Behavioral Therapy.

[d]$P<.05$

Health-Related Quality of Life

At both baseline and 16-week, participants had on average, slight-to-moderate perceived problems with their mobility, usual activities, pain and discomfort, and anxiety and depression, and no-to-slight problems with self-care. While postintervention scores decreased, particularly for self-care, usual activities, and anxiety and depression (respectively: t_{38}=2.08, P=.04; t_{38}=2.58, P=.01; t_{38}=2.27, P=.03), the reductions were not sufficient to move participants into categories of lower perceived severity of problems. Participants' challenges with mobility and pain and discomfort remained unaffected (P>.05). Table 4 shows domain-specific mean scores and frequencies, alongside t test P values.

XSL•FO
RenderX

Table . EuroQol 5 Dimensions, 5 Levels domain mean scores, response distribution by severity level, and *P* values from comparison of means analysis.

EQ-5D-5L[a] domain	Score, baseline		16-week score		*P* value[b]
	Score, mean (SD)	Participants (N=39), n (%)	Score, mean (SD)	Participants (N=39), n (%)	
Mobility	2.64 (1.25)	—[c]	2.41 (1.23)	—	.12
No problems	—	10 (25.62)	—	13 (33.32)	—
Slight problems	—	8 (20.50)	—	8 (20.50)	—
Moderate problems	—	8 (20.50)	—	7 (17.93)	—
Severe problems	—	12 (30.81)	—	11 (28.22)	—
Unable to walk about	—	1 (2.63)	—	0 (0)	—
Self-care	1.85 (1.18)	—	1.54 (0.88)	—	.04[d]
No problems	—	22 (56.4)	—	27 (69.22)	—
Slight problems	—	8 (20.50)	—	4 (10.31)	—
Moderate problems	—	3 (7.71)	—	7 (17.93)	—
Severe problems	—	5 (12.81)	—	1 (2.63)	—
Unable to wash or dress	—	1 (2.63)	—	0 (0)	—
Usual activities	2.44 (1.17)	—	2.03 (0.96)	—	.01[d]
No problems	—	9 (23.1)	—	14 (35.92)	—
Slight problems	—	14 (35.9)	—	13 (33.32)	—
Moderate problems	—	8 (20.50)	—	9 (23.12)	—
Severe problems	—	6 (15.44)	—	3 (7.73)	—
Unable to do usual activities	—	2 (5.11)	—	0 (0)	—
Pain and discomfort	2.87 (1.20)	—	2.82 (1.10)	—	.70
No problems	—	5 (12.81)	—	3 (7.73)	—
Slight problems	—	12 (30.81)	—	15 (38.51)	—
Moderate problems	—	8 (20.50)	—	10 (25.62)	—
Severe problems	—	11 (28.22)	—	8 (20.50)	—
Extreme pain or discomfort	—	3 (7.73)	—	3 (7.73)	—
Anxiety and depression	2.59 (1.12)	—	2.23 (1.09)	—	.03[d]
No problems	—	9 (23.12)	—	12 (30.81)	—
Slight problems	—	7 (17.93)	—	11 (28.22)	—
Moderate problems	—	15 (38.51)	—	13 (33.32)	—
Severe problems	—	7 (17.93)	—	1 (2.63)	—
Extremely anxious or depressed	—	1 (2.63)	—	2 (5.11)	—

[a]EQ-5D-5L: EuroQol 5 Dimensions, 5 Levels

[b]Within group differences.

[c]Not applicable.

[d]$P<.05$

There were no between-group differences at baseline for participants' scores in pain and discomfort, or anxiety and depression ($P>.05$); however, participants allocated to the app-based intervention showed significantly lower scores for mobility ($F_{2,36}=3.43$, $P=.04$), self-care ($F_{2,36}=3.51$, $P=.04$) and usual activities ($F_{2,36}=3.83$, $P=.03$), indicating fewer perceived problems in each domain. Participants allocated to one-to-one psychological support showed the highest levels of perceived

XSL•FO
RenderX

problems for mobility (mean 3.30, SD 1.2) and usual activities (mean 3.0, SD 1.1).

Subjective ratings of health status also improved at 16 weeks (t_{37}=−3.18; P=.003), with a medium effect size (Cohen d=0.52), with participants rating on average 53.5/100 at baseline and 62.8/100 at 16 weeks, indicating a 17.4% improvement in subjective health status. There were no between-group differences in subjective health status at baseline ($F_{2,35}$=2.57, P=.09) or 16-week ($F_{2,36}$=2.30, P=.12); however, participants allocated to the app-based intervention subjectively rated their health higher (mean 61.3, SD 22.2) than those in the DBT-based group intervention and one-to-one psychological support (mean 44.9, SD 19.3, and mean 48, SD 19.9, respectively) at baseline (Table 5). Despite participants in the DBT-based group intervention experiencing the greatest improvements at 16 weeks (26.7% increase), only the app-based intervention resulted in significant postintervention change (t_{17}=−2.93, P=.009).

Table . Average subjective ratings of health status at baseline and 16-week (out of 100).

Intervention	Baseline, mean (SD)	16-week, mean (SD)	*P* value
Psychologically informed, app-based intervention	61.3 (22.21)	70.0 (18.64)	.009[a]
DBT-based skills training group intervention	44.9 (19.32)	56.9 (19.92)	.14
One-to-one psychological support	48.0 (19.91)	56.3 (20.81)	.28
Total (N=39)	53.5 (21.33)	62.8 (20.32)[a]	.00[a]

[a]P<.05

Physical Activity

At baseline, 61.5% (n=24) of participants were categorized as “Inactive”, 17.9% (n=7) as “Moderately inactive,” 15.4% (n=6) as “Moderately active,” and 5.1% (n=2) as “Active.” No significant changes in physical activity levels were observed at 16-week (P=.26). The majority of participants (35/39, 84.6%) attended a call with the physical activity specialist at baseline and were supported to set individualized physical activity goals, while only 4 participants attended the live activity classes, with an average of 5 sessions per participant (range 1 - 16). Only 2 participants opted in for a follow-up call with the physical activity specialist to review their goals at 16 weeks.

Emotion Regulation

Participants allocated to the DBT-based group intervention scored a mean DERS-16 score of 50.8 (SD 16.8) at baseline and a mean score of 39.2 (SD 13.2) at 16-week, with a mean change score of −11.6 (SD 13.9). This reduction in difficulties in emotion regulation was statistically significant (t_9=2.63; P=.03).

Feasibility

Retention and Discharge Reasons

Out of a total of 62 referrals received, 10 participants were discharged before starting the intervention for the following reasons: need to prioritize other physical or mental health support needs over weight management goals (n=2), low digital literacy (n=1), missed video consultation appointments (n=2), did not complete baseline assessment questionnaires (n=1), and other (n=4). A further 7 participants were discharged during the intervention due to low or no engagement or changes in life circumstances which impacted their ability to engage with the intervention. Outcomes from 6 further participants were not yet available at the time of analysis. The attrition rate was 27.4% (n=7).

Engagement

On average, participants attended 93.7% of the intervention sessions, with the highest attendance for the app-based intervention (96.1%) and the lowest for the DBT-based group intervention (87.5%). Session attendance was a mean 7.7 (out of 8) sessions (SD 0.58, range 6 - 8) for the app-based intervention, 14 (out of 16) sessions (SD 1.56, range 11 - 16) for the DBT-based intervention, and 11 sessions (SD 1.56, range 7 - 13) for one-to-one psychological support. Participants in the app-based intervention attended on average 4.1 (out of 8) in-app check-ins (SD 2.90, range 0 - 8) and participants in the DBT-based group intervention attended on average 11.7 (out of 16) telephone check-in calls (SD 2.82, range 8 - 16).

No significant correlation between session attendance (%) and any of the primary and secondary outcome variables was observed (P>.05).

Participants in the app-based intervention consistently engaged with the app throughout the 16 weeks; however, home screen views showed a slight decrease from baseline (mean 4.52, SD 1.13) to Week 16 (mean 3.03, SD 1.29)

Acceptability

The postintervention acceptability questionnaire was completed by 56.4% (n=22) of participants, all of whom provided responses to each of the free-text questions. Overall, participants’ experiences of the interventions were rated at an average 9 out of 10, with one-to-one psychological support showing higher satisfaction ratings (mean 9.8/10, SD 0.92) than the app-based intervention and the DBT-based group intervention (mean 8.8/10, SD 0.53 and mean 9/10, SD 1.20, respectively).

> *Having a working relationship that felt I was respected and an equal member in my treatment.* [Participant 1, one-to-one psychological support intervention]

XSL•FO
RenderX

Overall, the main reported benefits of the interventions were behavioral changes (14/22, 63.6%) (eg, reduced binge eating episodes, decreased snacking frequency), psychological improvements (12/22, 54.5%) (eg, development of emotion regulation strategies and other coping mechanisms), educational value (11/22, 50%) (eg, improved understanding of weight loss barriers and personal eating habits), strong support system (9/22, 40.9%) (eg, meaningful relationships with peers and intervention facilitators and normalization of common challenges), and practical skills development (8/22, 36.4%) (eg, goal setting and improvements in managing cravings and urges).

> *I feel much more empowered to cope/manage emotions and feel hopeful for reaching my goals.* [Participant 2, DBT-based group intervention]

With regard to challenges of the interventions, participants reported struggling with psychological barriers (9/22, 40.9%) (eg, connection between eating-related distress and recollection of experiences of trauma or adversity), as well as time-related difficulties (8/22, 36.4%) (eg, frequency or intensity of intervention commitments, and complexity of assessment processes).

> *To accept that trauma has shaped my reactions to everyday occurrences, and to identify my emotions.* [Participant 3, one-to-one psychological support intervention]

The main self-reported behavior changes included dietary changes (13/22, 59.1%) (eg, implementation of the "balanced plate approach" [88] and reduction in portion sizes and snacking behavior); implementation of psychological strategies (10/22, 45.5%) (eg, challenging all-or-nothing thinking and self-criticism, more effective management of triggers to emotional distress); and interpersonal communication strategies (5/22, 22.7%) (eg, more effective and assertive strategies to communicate eating-related challenges with family and friends).

> *Eating more fruit and vegetables! Decision fatigue was a major barrier for me, I now have a few options for each meal and choose throughout the week. I accept that I can have a bit of chocolate but I don't have to eat as much, I'm more mindful! I've become more active, I enjoy the gym now!* [Participant 4, app-based intervention]

The most useful elements of the intervention reported by participants were one-to-one support (14/22, 63.6%), educational components (10/22, 45.5%), peer support (6/22, 27.3%), and overall intervention structure (5/22, 22.7%), while the elements considered least useful were difficulty with information retention in text-heavy educational material (5/22, 22.7%), app-related technical issues (3/22, 13.6%), and the physical activity live classes at unsuitable times (7/22, 31.8%).

> *Been able to have conversations with people regarding trigger factors and learnt to sit with the thoughts and be okay with not being okay.* [Participant 5, app-based intervention]

The main suggestions from participants for future improvements included app functionality (5/22, 22.7%) (eg, better notification system), content delivery methods (4/22, 18.2%) (eg, reduced repetition in course structure and improved recipe option), and accessibility (3/22, 13.6%) (eg, increased appointment scheduling flexibility and reducing text-heavy format of some of educational material).

> *More of an issue with the app but I didn't ever receive notifications which I would have liked to remind me to check in etc.* [Participant 6, app-based intervention]

Discussion

Overview

This study aimed to evaluate the potential effectiveness, feasibility, and acceptability of a remotely delivered SWMS for patients living with obesity, referred from NHS SWMSs in the United Kingdom. The study showed promising outcomes in weight change, psychological distress, emotional eating behavior, perceived health-related quality of life, and emotion regulation following the first phase (16 weeks) of a 3-phase service. While challenges remain, such as low engagement in physical activity, the present findings provide a strong foundation for the role of remotely delivered SWMSs in addressing current gaps in obesity service provision in the United Kingdom.

Principal Findings

The majority of participants had either lost or maintained weight upon completion of their intervention. This is a positive finding given that the aim of this first phase of the service (Preparing for weight loss) was to first address psychological barriers to behavior change to support weight loss in subsequent phases, rather than prioritizing immediate weight reduction. In line with this innovative approach, the NIH emphasizes the prevention of further weight gain as a treatment target in order to mitigate associated health risks with obesity, which manifests as a chronic and relapsing condition [89,90]. More recently, a review of the definition and diagnostic criteria of clinical obesity suggested prioritizing obesity symptom management as an outcome (such as improvements in metabolic health, psychological distress, and quality of life) rather than weight loss, highlighting the importance of integrating psychological support into weight management interventions to enhance outcomes for people living with obesity [10,60].

Notably, despite the complexity of the present cohort, characterized by high body weight, high levels of psychological distress, and significant challenges with mobility and daily activities, participants still achieved positive weight outcomes. This suggests that a health-centric (eg, prioritizing the addressing of psychological barriers), rather than a weight-centric (eg, prioritizing weight loss) approach can be an effective strategy to support weight loss in this population group and that psychologically informed DWMIs have the potential to set the stage for long-term sustainable weight management [10,41,59,68].

Unsurprisingly, the psychologically informed app-based intervention saw greater weight change at 16 weeks compared to the DBT-based group intervention and one-to-one psychological support. This is likely explained by the higher degree of psychological complexity and barriers to change of

XSL•FO
RenderX

participants allocated to the DBT-based group intervention and one-to-one psychological support, as supported by existing literature. Studies indicate that individuals with elevated psychological distress, comorbid mental health conditions, or difficulties in emotion regulation often face greater challenges in initiating and sustaining behavioral changes due to reduced motivation, increased difficulties with emotional regulation, and lower self-efficacy [91,92]

The majority of participants began the intervention with clinically significant levels of psychological distress, which is a common presentation in populations living with obesity and multiple comorbidities [93-95]. The finding that baseline psychological distress was higher in those allocated to the DBT-based group intervention and one-to-one psychological support suggests the appropriateness and feasibility of the assessment process used to allocate participants to their interventions. Equitable distribution of improvements in psychological distress across interventions further supports the effectiveness of tailoring interventions to individual needs.

At 16 weeks, psychological distress scores decreased below the clinical threshold and improvements were observed across 3 scale domains, suggesting comprehensive psychological benefits and alleviation of psychological symptoms associated with obesity. It is important to note that this clinically significant improvement occurred without a significant reduction in body weight across groups.

While improvements in psychological distress did not directly correlate with weight change, participants who reduced emotional eating behaviors demonstrated greater weight loss. The interventions successfully reduced emotional eating behaviors, highlighting the potential efficacy of the psychological intervention components in addressing maladaptive eating patterns. This is supported by research demonstrating that emotional eating is a key contributor to obesity and that interventions focusing on emotion regulation and psychological support can lead to significant reductions in both emotional eating and overall weight [96,97]. Furthermore, studies have suggested that emotional eating behaviors could mediate the relationship between psychological distress and weight maintenance in a population with a history of weight cycling [91,98,99], and that long-term weight loss maintenance outcomes are closely linked to improvements in emotional eating behaviors [96,100].

Participants allocated to the DBT-based group intervention reported moderate baseline difficulties in emotion regulation, which decreased at 16 weeks, indicating progress in managing emotions and possibly contributing to a reduction in emotional eating behaviors. These findings are noteworthy given the relationship between psychological distress and psychological barriers to weight management, such as distress management, emotional eating, and reduced self-efficacy [101,102].

There were improvements in perceived health-related quality of life with regards to increased self-care practices, better ability to perform usual activities, and reduced anxiety and depression symptoms, suggesting that the interventions positively influenced functional independence and emotional well-being. Behavioral weight management interventions have previously been shown to lead to enhancements in a range of health-related outcomes, including reductions in depressive symptoms and improvements in quality of life [103].

However, perceived problems with mobility, and pain and discomfort remained unaffected. Generally, research indicates that while weight loss interventions may provide small to moderate improvements in pain and disability for conditions like osteoarthritis, these effects can be limited, suggesting the need for more tailored and prolonged interventions to effectively address these domains [104].

Nonetheless, the observed improvement in subjective health ratings reflects participants' perception of enhanced overall health, despite their complex medical and psychological presentations, and is comparable to those of other studies [23]. Participants who accessed the DBT-based group intervention experienced the largest improvements in perceived subjective health rating, possibly due to DBT's focus on balancing acceptance with change as a meaningful therapeutic approach to promote changes in both emotional and behavioral domains, as well as improvements in perceived health, as further evidenced by other studies [105,106]. It is also possible that the group nature of the DBT-based intervention contributed to improvements in subjective health ratings at 16 weeks, given that social support has been identified as a key component for weight management interventions effectiveness [107-109], even in the context of remotely delivered interventions [110,111].

The finding that no significant changes in physical activity levels were observed reflects the challenges of increasing physical activity in people living with obesity who typically experience significant physical and mental health comorbidities [112,113]. Despite high baseline engagement with physical activity-related goal setting, engagement with follow-up calls and live activity class participation was limited, and overall physical activity levels did not increase at post-intervention. A recent systematic review identified competing priorities and time constraints as key barriers to lifestyle change intervention adherence [114]; therefore, it is possible that participants engaged in the physical activity support less than anticipated due to a sense of being overwhelmed by the demands of their core intervention, with establishing new physical activity habits feeling more like an additional burden on top of their other commitments, rather than an opportunity to implement further behavior change. Another potential barrier could have been the limited flexibility in the days and times the live activity classes took place, restricted by the availability of the physical activity specialist and associate. This aligns with previous studies that highlight the need to mitigate the effects of time-related barriers to physical activity engagement in people living with obesity by providing flexible and time-efficient support offers [112]. The remotely delivered nature of the live activity classes could have also acted as an engagement barrier, in line with findings from other studies which suggested that adherence to home-based physical activity interventions is often low in unsupervised settings with limited-to-no face-to-face activity supervision [115-117].

Retention in lifestyle change interventions is often challenging, with attrition rates in in-person weight management

XSL•FO
RenderX

interventions averaging around 50% [118,119]. Nonetheless, intervention tailoring to individual needs, convenience of setting, and more acceptable service design can greatly improve participant retention, as suggested by an evaluation of a SWMS based in primary care [23]. Attrition rates of this study fall within the range of average attrition rates in DWMIs at around 12 weeks (16%‑50%) [120,121]. Factors influencing completion of DWMIs include lack of human interaction, lower baseline motivation, lower levels of digital literacy, and technical challenges. High attendance rates throughout the present study's interventions potentially reflect effective intervention design and intervention allocation, face-to-face support via video calls, and tailoring, which align with factors identified as reducing dropout rates and improving service feasibility [121,122]. Consistent patterns in app home screen views throughout the 16-week further support these findings, suggesting sustained digital engagement that likely contributed to the observed attrition and attendance rates across interventions.

Participants reported very high levels of satisfaction with their interventions, averaging 9 out of 10, with one-to-one psychological support receiving the highest ratings. Furthermore, participants highlighted key intervention benefits, including behavioral changes, psychological improvements, and educational value. Strong support from facilitators and peers was also valued by participants. Challenges included psychological barriers, such as trauma-linked distress, text-heavy resources, and time demands of the intervention. Suggestions for improvement focused on the app notification system, more engaging content delivery, and increased scheduling flexibility. These findings align with other studies that saw similar satisfaction ratings and overall acceptability outcomes within the context of remotely delivered weight management interventions [123]. For example, a technology-based, rural weight management intervention in older adults with obesity identified behavioral changes as the main perceived benefits, and other existing health conditions (including mental health conditions) and technological challenges as the main perceived barriers [122].

Limitations

This study was a service evaluation that used retrospective data in the absence of a control group and randomization. Therefore, it is not possible to draw definitive conclusions about the effectiveness of the intervention nor compare intervention efficacy to standard care outcomes.

Collection of reliable weight data in SWMSs is a common challenge [124,125]. Some participants reported their baseline weight data (received upon referral) to be out of date and, at times, lower than their current weight. Weight data at 16 weeks was self-reported [126], with some participants only able to provide weight estimates, particularly when they opted not to weigh themselves to avoid psychological distress and associated potential triggers of emotional eating, or when self-weighting equipment was unavailable or inaccessible as can often be the case for people living with high body weights or mobility limitations [127,128]. It is possible the present study might have overestimated weight change outcomes due to a tendency of individuals, particularly women, to underreport weight data, often due to biases such as social pressure and self-esteem issues [124]. Improvement in the validity of weight data is needed to increase confidence in findings and provide a more accurate understanding of the impact of this pilot service on weight change.

It is important to note that the smaller sample sizes in the DBT-based group intervention and one-to-one psychological support limit the generalizability of the present findings; therefore, further analysis with larger sample sizes at completion of this 3-phase remotely delivered SWMS is needed. Furthermore, although the final total sample size exceeded the minimum requirements calculated via a priori power analysis, a larger sample would increase power to detect smaller effects in secondary outcomes, such as psychological distress and emotional eating—outcomes that may still be clinically meaningful despite modest effect sizes—and would better support further intervention improvement and tailoring.

In addition, while app home screen views were used as a proxy for digital engagement, this metric does not capture the extent to which participants actively engaged with specific intervention components, such as the educational modules. However, key module content was reinforced and discussed during one-to-one health coaching video calls, partially mitigating this limitation.

The modest response rate to the postintervention acceptability questionnaire may limit the representativeness of qualitative feedback, highlighting the value of collecting feedback earlier in the program to reduce participant burden and improve response rates.

Most participants were female, and no ethnicity data were collected, limiting conclusions drawn from the present analysis and its generalizability. Furthermore, despite the majority of participants residing in areas of low-to-moderate deprivation, most of them lived in rural and island communities, where the IMD is believed to underestimate the prevalence of deprivation and associated health inequalities [129].

Conclusions

This preliminary service evaluation suggested that a remotely delivered SWMS has the potential to be effective, feasible, and acceptable for NHS-referred patients in the United Kingdom. The intervention showed promising improvements in psychological distress, emotional eating, perceived quality of life, and difficulties in emotion regulation and was effective in supporting weight stabilization and modest weight loss during the initial 16 weeks, particularly through psychologically informed components. Areas for further improvement were highlighted, including accuracy of primary outcome data collection, improvements in physical activity promotion, and a more representative sample diversity.

A planned comprehensive evaluation at 12 months will measure the same primary and secondary outcomes—weight change, psychological distress, emotional eating, perceived quality of life, and difficulties in emotion regulation—with a larger sample size to provide more robust, reliable, and generalizable findings. Qualitative feedback will also continue to be collected, to better understand participant experiences and intervention

XSL•FO
RenderX

acceptability. More rigorous conclusions regarding the effectiveness, feasibility, and acceptability of remotely delivered SWMS interventions will be possible following this extended evaluation. These findings will also inform the scalability, future implementation, and directions for future research in this field.

Data Availability

The datasets generated and analyzed during this study are not publicly available due to privacy policy restrictions and the scope of participant consent obtained for this service evaluation.

In accordance with our privacy policy, the data cannot be shared or made available by the authors upon request.

Authors' Contributions

BR and MW contributed to the development of the study protocol and initial methodology (Conceptualization, Methodology). GS led the quantitative data analysis and prepared the first draft of the manuscript (Formal analysis, Writing – original draft). MW and GS managed data collection and organization (Data curation), while BR and GS were responsible for project administration and investigation (Investigation, Project administration). GS and MW supported software development and data handling (Software), and GS created the visualizations (Visualization). Funding was acquired by MW (Funding acquisition). BR provided overall supervision (Supervision). All authors contributed resources (Resources), validation of findings (Validation), and qualitative data analysis. All authors were involved in reviewing and editing the manuscript (Writing – review & editing), and all approved the final version.

Conflicts of Interest

GS, MW, and RR are employees of Second Nature Healthy Habits Ltd.

Multimedia Appendix 1
Health research authority decision tool result.
[PDF File, 211 KB - formative_v9i1e71914_app1.pdf]

Multimedia Appendix 2
Average psychological distress by intervention.
[PNG File, 17 KB - formative_v9i1e71914_app2.png]

Checklist 1
SQUIRE (version 2.0) checklist.
[PDF File, 66 KB - formative_v9i1e71914_app3.pdf]

References

1. Obesity and overweight. World Health Organization. 2020. URL: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight [accessed 2025-01-27]
2. Health survey for england 2021. Digital NHS. 2023. URL: https://digital.nhs.uk/data-and-information/publications/statistical/health-survey-for-england? [accessed 2025-01-27]
3. Health survey northern ireland: first results 2022/23. GOVUK. 2023. URL: https://www.gov.uk/government/statistics/health-survey-northern-ireland-201920-first-results [accessed 2025-01-27]
4. Scottish Centre for Social Research (Scottish Government). Scottish health survey. . 2023 URL: https://www.gov.scot/publications/scottish-health-survey-2023-volume-1-main-report/documents/ [accessed 2025-01-27]
5. GBD 2019 Risk Factors Collaborators. Global burden of 87 risk factors in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet 2020 Oct 17;396(10258):1223-1249. [doi: 10.1016/S0140-6736(20)30752-2] [Medline: 33069327]
6. Avila C, Holloway AC, Hahn MK, et al. An overview of links between obesity and mental health. Curr Obes Rep 2015 Sep;4(3):303-310. [doi: 10.1007/s13679-015-0164-9] [Medline: 26627487]
7. Chu DT, Minh Nguyet NT, Nga VT, et al. An update on obesity: mental consequences and psychological interventions. Diabetes Metab Syndr Clin Res Rev 2019 Jan;13(1):155-160. [doi: 10.1016/j.dsx.2018.07.015]
8. Kivimäki M, Batty GD, Singh-Manoux A, et al. Association between common mental disorder and obesity over the adult life course. Br J Psychiatry 2009 Aug;195(2):149-155. [doi: 10.1192/bjp.bp.108.057299] [Medline: 19648547]
9. Stephenson J, Smith CM, Kearns B, Haywood A, Bissell P. The association between obesity and quality of life: a retrospective analysis of a large-scale population-based cohort study. BMC Public Health 2021 Nov 3;21(1):1990. [doi: 10.1186/s12889-021-12009-8] [Medline: 34732156]

XSL•FO
RenderX

10. Rubino F, Cummings DE, Eckel RH, et al. Definition and diagnostic criteria of clinical obesity. Lancet Diabetes Endocrinol 2025 Mar;13(3):221-262. [doi: 10.1016/S2213-8587(24)00316-4] [Medline: 39824205]
11. Overweight and obesity management [NG246]. National Institute of Health and Care Excellence. 2025. URL: https://www.nice.org.uk/guidance/ph49 [accessed 2025-01-22]
12. Coulton Dodhia S, Ells L, Blackshaw J, Tedstone AV. National Mapping of Weight Management Services: Provision of Tier 2 and Tier 3 Services in England: Public Health England; 2015. URL: https://assets.publishing.service.gov.uk/media/5a7f74d2e5274a2e8ab4c4b1/Final_Weight_Management_Mapping_Report.pdf [accessed 2025-01-27]
13. National Institute for Health and Care Excellence. Weight Management: Lifestyle Services for Overweight or Obese Adults: Public Health Guideline [PH53]; 2014. URL: https://www.nice.org.uk/guidance/ng246 [accessed 2025-09-29]
14. Welbourn R, Hopkins J, Dixon JB, et al. Commissioning guidance for weight assessment and management in adults and children with severe complex obesity. Obes Rev 2018 Jan;19(1):14-27. [doi: 10.1111/obr.12601] [Medline: 29024367]
15. National Institute for Health and Care Excellence. Obesity: identification, assessment and management. In: Clinical Guideline [CG189] 2014. URL: https://www.nice.org.uk/guidance/cg189 [accessed 2025-01-27]
16. Blane DN, Macdonald S, Morrison D, O'Donnell CA. The role of primary care in adult weight management: qualitative interviews with key stakeholders in weight management services. BMC Health Serv Res 2017 Nov 21;17(1):764. [doi: 10.1186/s12913-017-2729-7] [Medline: 29162111]
17. Alkharaiji M, Anyanwagu U, Donnelly R, Idris I. Tier 3 specialist weight management service and pre-bariatric multicomponent weight management programmes for adults with obesity living in the UK: a systematic review. Endocrinol Diabetes Metab 2019 Jan;2(1):e00042. [doi: 10.1002/edm2.42] [Medline: 30815571]
18. Huntriss R, Haines M, Jones L, Mulligan D. A service evaluation exploring the effectiveness of a locally commissioned tier 3 weight management programme offering face-to-face, telephone and digital dietetic support. Clin Obes 2021 Jun;11(3):e12444. [doi: 10.1111/cob.12444] [Medline: 33600056]
19. Brown TJ, O'Malley C, Blackshaw J, et al. Exploring the evidence base for Tier 3 weight management interventions for adults: a systematic review. Clin Obes 2017 Oct;7(5):260-272. [doi: 10.1111/cob.12204] [Medline: 28695579]
20. Aceves‐Martins M, Robertson C, Cooper D, et al. A systematic review of UK‐based long‐term nonsurgical interventions for people with severe obesity (BMI ≥35 kg m −2). J Human Nutrition Diet 2020 Jun;33(3):351-372. [doi: 10.1111/jhn.12732]
21. Watkins R, Swancutt D, Alexander M, et al. A qualitative exploration of patient and staff experiences of the receipt and delivery of specialist weight management services in the UK. Patient 2023 Nov;16(6):625-640. [doi: 10.1007/s40271-023-00644-9] [Medline: 37572233]
22. Read S, Logue J. Variations in weight management services in Scotland: a national survey of weight management provision. J Public Health 2016 Sep;38(3):e325-e335. [doi: 10.1093/pubmed/fdv132]
23. Jennings A, Hughes CA, Kumaravel B, et al. Evaluation of a multidisciplinary Tier 3 weight management service for adults with morbid obesity, or obesity and comorbidities, based in primary care. Clin Obes 2014 Oct;4(5):254-266. [doi: 10.1111/cob.12066] [Medline: 25825858]
24. Kar P, Goward C, Whitman M, Davies M, Willner T, Shaw K. Engagement and effectiveness of digitally enabled behavioural change support for people living with type 2 diabetes. Practical Diabetes 2020 Sep;37(5):167. [doi: 10.1002/pdi.2295]
25. Hampton J, Allen E, Edson C. Service evaluation of a digital behavioural change programme. Future Healthc J 2017 Oct;4(3):173-177. [doi: 10.7861/futurehosp.4-3-173] [Medline: 31098466]
26. Sharpe EE, Karasouli E, Meyer C. Examining factors of engagement with digital interventions for weight management: rapid review. JMIR Res Protoc 2017 Oct 23;6(10):e205. [doi: 10.2196/resprot.6059] [Medline: 29061557]
27. Idris I, Hampton J, Moncrieff F, Whitman M. Effectiveness of a digital lifestyle change program in obese and type 2 diabetes populations: service evaluation of real-world data. JMIR Diabetes 2020 Jan 20;5(1):e15189. [doi: 10.2196/15189] [Medline: 31958064]
28. Richards R, Wren G, Whitman M. The potential of a digital weight management program to support specialist weight management services in the UK national health service: retrospective analysis. JMIR Diabetes 2024 Jan 24;9:e52987. [doi: 10.2196/52987] [Medline: 38265852]
29. Perski O, Blandford A, West R, Michie S. Conceptualising engagement with digital behaviour change interventions: a systematic review using principles from critical interpretive synthesis. Transl Behav Med 2017 Jun;7(2):254-267. [doi: 10.1007/s13142-016-0453-1] [Medline: 27966189]
30. Beleigoli AM, Andrade AQ, Cançado AG, Paulo MN, Diniz M, Ribeiro AL. Web-based digital health interventions for weight loss and lifestyle habit changes in overweight and obese adults: systematic review and meta-analysis. J Med Internet Res 2019 Jan 8;21(1):e298. [doi: 10.2196/jmir.9609] [Medline: 30622090]
31. Sorgente A, Pietrabissa G, Manzoni GM, et al. Web-based interventions for weight loss or weight loss maintenance in overweight and obese people: a systematic review of systematic reviews. J Med Internet Res 2017 Jun 26;19(6):e229. [doi: 10.2196/jmir.6972] [Medline: 28652225]
32. Schippers M, Adam PCG, Smolenski DJ, Wong HTH, de Wit JBF. A meta-analysis of overall effects of weight loss interventions delivered via mobile phones and effect size differences according to delivery mode, personal contact, and intervention intensity and duration. Obes Rev 2017 Apr;18(4):450-459. [doi: 10.1111/obr.12492] [Medline: 28187246]

XSL•FO
RenderX

33. Ross KM, Carpenter CA, Arroyo KM, et al. Impact of transition from face‐to‐face to telehealth on behavioral obesity treatment during the COVID‐19 pandemic. Obesity (Silver Spring) 2022 Apr;30(4):858-863. [doi: 10.1002/oby.23383]
34. Fischer M, Weimann T, Oberänder N, Schupitza L, Hösel J, Weimann A. Remote treatment successfully delivers a usual care weight loss and lifestyle intervention in adults with morbid obesity. Ann Nutr Metab 2022;78(6):328-335. [doi: 10.1159/000526475] [Medline: 35977461]
35. Richards R, Wren GM, Campion P, Whitman M. A remotely delivered, semaglutide-supported specialist weight management program: preliminary findings from a retrospective service evaluation. JMIR Form Res 2023 Dec 28;7:e53619. [doi: 10.2196/53619] [Medline: 38153780]
36. Hanson P, Summers C, Panesar A, et al. Low carb program health app within a hospital-based obesity setting: observational service evaluation. JMIR Form Res 2021 Sep 23;5(9):e29110. [doi: 10.2196/29110] [Medline: 34449405]
37. Rothberg AE, Marriott DJ, Miller NM, Herman WH. Retention and weight outcomes after transitioning an intensive behavioral weight management program from an in-person to a virtual format. Obes Sci Pract 2023 Oct;9(5):452-458. [doi: 10.1002/osp4.673] [Medline: 37810529]
38. Ogrinc G, Davies L, Goodman D, Batalden P, Davidoff F, Stevens D. SQUIRE 2.0 (Standards for QUality Improvement Reporting Excellence): revised publication guidelines from a detailed consensus process. BMJ Qual Saf 2016 Dec;25(12):986-992. [doi: 10.1136/bmjqs-2015-004411] [Medline: 26369893]
39. Privacy policy. Second nature. URL: https://www.secondnature.io/privacy [accessed 2025-01-22]
40. Is my study research? National Health Service Health Research Authority. URL: https://www.hra-decisiontools.org.uk/research/ [accessed 2024-06-09]
41. Dandgey S, Patten E. Psychological considerations for the holistic management of obesity. Clin Med (Lond) 2023 Jul;23(4):318-322. [doi: 10.7861/clinmed.2023-0146] [Medline: 37524420]
42. Ohsiek S, Williams M. Psychological factors influencing weight loss maintenance: an integrative literature review. J Am Acad Nurse Pract 2011 Nov;23(11):592-601. [doi: 10.1111/j.1745-7599.2011.00647.x] [Medline: 22023231]
43. Michie S, Johnston M. Theories and techniques of behaviour change: developing a cumulative science of behaviour change. Health Psychol Rev 2012 Mar;6(1):1-6. [doi: 10.1080/17437199.2012.654964]
44. Kwasnicka D, Dombrowski SU, White M, Sniehotta F. Theoretical explanations for maintenance of behaviour change: a systematic review of behaviour theories. Health Psychol Rev 2016 Sep;10(3):277-296. [doi: 10.1080/17437199.2016.1151372] [Medline: 26854092]
45. Greaves C, Poltawski L, Garside R, Briscoe S. Understanding the challenge of weight loss maintenance: a systematic review and synthesis of qualitative research on weight loss maintenance. Health Psychol Rev 2017 Jun;11(2):145-163. [doi: 10.1080/17437199.2017.1299583] [Medline: 28281891]
46. Carrière K, Khoury B, Günak MM, Knäuper B. Mindfulness-based interventions for weight loss: a systematic review and meta-analysis. Obes Rev 2018 Feb;19(2):164-177. [doi: 10.1111/obr.12623] [Medline: 29076610]
47. Michie S, van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. Implementation Sci 2011 Dec;6(1). [doi: 10.1186/1748-5908-6-42]
48. Michie S, Richardson M, Johnston M, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013 Aug;46(1):81-95. [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
49. National Institute of Health and Care Excellence. Behaviour Change: Individual Approaches [PH49] 2014. URL: https://www.nice.org.uk/guidance/ph49 [accessed 2025-01-27]
50. Safer D, Masson P, Adler S. The DBT Solution for Emotional Eating: A Proven Program to Break the Cycle of Bingeing and Out-of-Control Eating: Guildford Press; 2017.
51. Linehan MM. Skills Training Manual for Treating Borderline Personality Disorder: Guilford Press; 1993.
52. Carter JC, Kenny TE, Singleton C, Van Wijk M, Heath O. Dialectical behavior therapy self-help for binge-eating disorder: a randomized controlled study. Int J Eat Disord 2020 Mar;53(3):451-460. [doi: 10.1002/eat.23208] [Medline: 31821592]
53. Lammers MW, Vroling MS, Crosby RD, van Strien T. Dialectical behavior therapy compared to cognitive behavior therapy in binge-eating disorder: an effectiveness study with 6-month follow-up. Int J Eat Disord 2022 Jul;55(7):902-913. [doi: 10.1002/eat.23750] [Medline: 35665526]
54. Chen EY, Matthews L, Allen C, Kuo JR, Linehan MM. Dialectical behavior therapy for clients with binge‐eating disorder or bulimia nervosa and borderline personality disorder. Intl J Eating Disorders 2008 Sep;41(6):505-512. [doi: 10.1002/eat.20522]
55. Rahmani M, Omidi A, Asemi Z, Akbari H. The effect of dialectical behaviour therapy on binge eating, difficulties in emotion regulation and BMI in overweight patients with binge-eating disorder: a randomized controlled trial. Ment Health Prev 2018 Mar;9:13-18. [doi: 10.1016/j.mhp.2017.11.002]
56. Lammers MW, Vroling MS, Crosby RD, van Strien T. Dialectical behavior therapy adapted for binge eating compared to cognitive behavior therapy in obese adults with binge eating disorder: a controlled study. J Eat Disord 2020;8(1):27. [doi: 10.1186/s40337-020-00299-z] [Medline: 32528681]

XSL•FO
RenderX

57. Cancian ACM, de Souza LAS, Liboni RPA, Machado WDL, Oliveira MDS. Effects of a dialectical behavior therapy-based skills group intervention for obese individuals: a Brazilian pilot study. Eat Weight Disord 2019 Dec;24(6):1099-1111. [doi: 10.1007/s40519-017-0461-2] [Medline: 29197947]
58. Braden A, Redondo R, Ferrell E, et al. An open trial examining dialectical behavior therapy skills and behavioral weight loss for adults with emotional eating and overweight/obesity. Behav Ther 2022 Jul;53(4):614-627. [doi: 10.1016/j.beth.2022.01.008] [Medline: 35697426]
59. Kim M, Kim Y, Go Y, et al. Multidimensional cognitive behavioral therapy for obesity applied by psychologists using a digital platform: open-label randomized controlled trial. JMIR Mhealth Uhealth 2020 Apr 30;8(4):e14817. [doi: 10.2196/14817] [Medline: 32352391]
60. Psychological Perspectives on Obesity: Addressing Policy, Practice and Research Priorities: British Psychological Society; 2019. [doi: 10.53841/bpsrep.2019.rep130]
61. Canadian adult obesity clinical practice guideline. Obesity Canada. 2020. URL: https://obesitycanada.ca/guidelines/behavioural/ [accessed 2025-09-10]
62. Rogers JM, Ferrari M, Mosely K, Lang CP, Brennan L. Mindfulness-based interventions for adults who are overweight or obese: a meta-analysis of physical and psychological health outcomes. Obes Rev 2017 Jan;18(1):51-67. [doi: 10.1111/obr.12461] [Medline: 27862826]
63. Dastan B, Afshar zanjani S, Froueddin adl A, Habibi M. The effectiveness of dialectical behaviour therapy for treating women with obesity suffering from BED: a feasibility and pilot study. Clin Psychol (Aust Psychol Soc) 2020 Jul 1;24(2):133-142. [doi: 10.1111/cp.12197]
64. Torres S, Sales CMD, Guerra MP, Simões MP, Pinto M, Vieira FM. Emotion-focused cognitive behavioral therapy in comorbid obesity with binge eating disorder: a pilot study of feasibility and long-term outcomes. Front Psychol 2020;11:343. [doi: 10.3389/fpsyg.2020.00343] [Medline: 32194485]
65. Spring B, Champion KE, Acabchuk R, Hennessy EA. Self-regulatory behaviour change techniques in interventions to promote healthy eating, physical activity, or weight loss: a meta-review. Health Psychol Rev 2021 Dec;15(4):508-539. [doi: 10.1080/17437199.2020.1721310] [Medline: 31973666]
66. Forman EM, Butryn ML, Manasse SM, Bradley LE. Acceptance-based behavioral treatment for weight control: a review and future directions. Curr Opin Psychol 2015 Apr;2:87-90. [doi: 10.1016/j.copsyc.2014.12.020] [Medline: 26101783]
67. Cattivelli R, Guerrini Usubini A, Manzoni GM, et al. ACTonFood. Acceptance and commitment therapy-based group treatment compared to cognitive behavioral therapy-based group treatment for weight loss maintenance: an individually randomized group treatment trial. Int J Environ Res Public Health 2021 Sep 10;18(18):9558. [doi: 10.3390/ijerph18189558] [Medline: 34574482]
68. Richards R, Jones RA, Whittle F, et al. Development of a web-based, guided self-help, acceptance and commitment therapy-based intervention for weight loss maintenance: evidence-, theory-, and person-based approach. JMIR Form Res 2022 Jan 7;6(1):e31801. [doi: 10.2196/31801] [Medline: 34994698]
69. Hayes SC, Hofmann SG. The third wave of cognitive behavioral therapy and the rise of process-based care. World Psychiatry 2017 Oct;16(3):245-246. [doi: 10.1002/wps.20442] [Medline: 28941087]
70. Boucher S, Edwards O, Gray A, et al. Teaching intuitive eating and acceptance and commitment therapy skills via a web-based intervention: a pilot single-arm intervention study. JMIR Res Protoc 2016 Oct 14;5(4):e180. [doi: 10.2196/resprot.5861] [Medline: 27742602]
71. Carter A, Gilbert P, Kirby JN. Compassion‐focused therapy for body weight shame: a mixed methods pilot trial. Clin Psychology and Psychoth 2021 Jan;28(1):93-108 [FREE Full text] [doi: 10.1002/cpp.2488]
72. Evans C, Connell J, Barkham M, et al. Clinical Outcomes in Routine Evaluation (CORE) Manual, 2nd edition: CORE System Trust; 2005.
73. Evans C, Connell J, Barkham M, et al. Towards a standardised brief outcome measure: psychometric properties and utility of the CORE-OM. Br J Psychiatry 2002 Jan;180(1):51-60. [doi: 10.1192/bjp.180.1.51] [Medline: 11772852]
74. Garaulet M, Canteras M, Morales E, López-Guimera G, Sánchez-Carracedo D, Corbalán-Tutau MD. Validation of a questionnaire on emotional eating for use in cases of obesity: the Emotional Eater Questionnaire (EEQ). Nutr Hosp 2012;27(2):645-651. [doi: 10.1590/S0212-16112012000200043] [Medline: 22732995]
75. Herdman M, Gudex C, Lloyd A, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Qual Life Res 2011 Dec;20(10):1727-1736. [doi: 10.1007/s11136-011-9903-x]
76. Janssen MF, Bonsel GJ, Luo N. Is EQ-5D-5L better than EQ-5D-3L? A head-to-head comparison of descriptive systems and value sets from seven countries. Pharmacoeconomics 2018 Jun;36(6):675-697. [doi: 10.1007/s40273-018-0623-8] [Medline: 29470821]
77. Mead N, Bower P, Roland M. The General Practice Assessment Questionnaire (GPAQ) - development and psychometric characteristics. BMC Fam Pract 2008 Feb 20;9(1):13. [doi: 10.1186/1471-2296-9-13] [Medline: 18289385]
78. Ahmad S, Harris T, Limb E, et al. Evaluation of reliability and validity of the General Practice Physical Activity Questionnaire (GPPAQ) in 60-74 year old primary care patients. BMC Fam Pract 2015 Sep 2;16(1):113. [doi: 10.1186/s12875-015-0324-8] [Medline: 26329981]

XSL•FO
RenderX

79. Appendix j: GPPAQ (general practice physical activity questionnaire). National Institute for Health and Care Excellence (NICE). 2017. URL: https://www.nice.org.uk/guidance/cg61/evidence/appendix-j-gppaq-pdf-196701669 [accessed 2025-09-10]
80. Bjureberg J, Ljótsson B, Tull MT, et al. Development and validation of a brief version of the Difficulties in Emotion Regulation Scale: the DERS-16. J Psychopathol Behav Assess 2016 Jun;38(2):284-296. [doi: 10.1007/s10862-015-9514-x] [Medline: 27239096]
81. Kang H. The prevention and handling of the missing data. Korean J Anesthesiol 2013 May;64(5):402-406. [doi: 10.4097/kjae.2013.64.5.402] [Medline: 23741561]
82. Mishra P, Pandey CM, Singh U, Gupta A, Sahu C, Keshri A. Descriptive statistics and normality tests for statistical data. Ann Card Anaesth 2019;22(1):67-72. [doi: 10.4103/aca.ACA_157_18] [Medline: 30648682]
83. Faul F, Erdfelder E, Lang AG, Buchner A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods 2007 May;39(2):175-191. [doi: 10.3758/bf03193146] [Medline: 17695343]
84. Cohen J. Statistical Power Analysis for the Behavioral Sciences, 2nd edition: Lawrence Erlbaum Associates Publishers; 1988. [doi: 10.4324/9780203771587]
85. Hartmann-Boyce J, Johns DJ, Jebb SA, Summerbell C, Aveyard P, Behavioural Weight Management Review Group. Behavioural weight management programmes for adults assessed by trials conducted in everyday contexts: systematic review and meta-analysis. Obes Rev 2014 Nov;15(11):920-932. [doi: 10.1111/obr.12220] [Medline: 25112559]
86. Mayring P. Qualitative Content Analysis: A Step-by-Step Guide: SAGE Publications Ltd; 2022.
87. Scott AM, Kolstoe S, Ploem MCC, Hammatt Z, Glasziou P. Exempting low-risk health and medical research from ethics reviews: comparing Australia, the United Kingdom, the United States and the Netherlands. Health Res Policy Syst 2020 Jan 28;18(1):11. [doi: 10.1186/s12961-019-0520-4] [Medline: 31992320]
88. The Eatwell guide. Public Health England. 2018. URL: https://assets.publishing.service.gov.uk/media/5ba8a50540f0b605084c9501/Eatwell_Guide_booklet_2018v4.pdf [accessed 2025-01-22]
89. Institute of Medicine (US) Subcommittee on Military Weight Management. Weight management: state of the science and opportunities for military programs. In: Weight-Loss and Maintenance Strategies: Washington (DC): National Academies Press (US); 2004. URL: https://www.ncbi.nlm.nih.gov/books/NBK221839/ [accessed 2024-12-11]
90. Madigan CD, Pavey T, Daley AJ, Jolly K, Brown WJ. Is weight cycling associated with adverse health outcomes? A cohort study. Prev Med 2018 Mar;108:47-52. [doi: 10.1016/j.ypmed.2017.12.010]
91. Li W, Wang D, Chen H, et al. The relationship between psychological distress and weight maintenance in weight cycling: mediating role of eating behavior. BMC Public Health 2024;24(1):894. [doi: 10.1186/s12889-024-18349-5]
92. de Jong M, Jansen N, van Middelkoop M. A systematic review of patient barriers and facilitators for implementing lifestyle interventions targeting weight loss in primary care. Obes Rev 2023 Aug;24(8):e13571. [doi: 10.1111/obr.13571] [Medline: 37226636]
93. Luppino FS, de Wit LM, Bouvy PF, et al. Overweight, obesity, and depression: a systematic review and meta-analysis of longitudinal studies. Arch Gen Psychiatry 2010 Mar;67(3):220-229. [doi: 10.1001/archgenpsychiatry.2010.2] [Medline: 20194822]
94. Xiao S, Shi L, Dong F, et al. The impact of chronic diseases on psychological distress among the older adults: the mediating and moderating role of activities of daily living and perceived social support. Aging Ment Health 2022 Sep;26(9):1798-1804. [doi: 10.1080/13607863.2021.1947965] [Medline: 34238092]
95. Carpenter KM, Hasin DS, Allison DB, Faith MS. Relationships between obesity and DSM-IV major depressive disorder, suicide ideation, and suicide attempts: results from a general population study. Am J Public Health 2000 Feb;90(2):251-257. [doi: 10.2105/ajph.90.2.251] [Medline: 10667187]
96. Frayn M, Livshits S, Knäuper B. Emotional eating and weight regulation: a qualitative study of compensatory behaviors and concerns. J Eat Disord 2018 Dec;6(1):23. [doi: 10.1186/s40337-018-0210-6]
97. Konttinen H. Emotional eating and obesity in adults: the role of depression, sleep and genes. Proc Nutr Soc 2020 Aug;79(3):283-289. [doi: 10.1017/S0029665120000166] [Medline: 32213213]
98. Guerrini Usubini A, Cattivelli R, Varallo G, et al. The relationship between psychological distress during the second wave lockdown of COVID-19 and emotional eating in Italian young adults: the mediating role of emotional dysregulation. J Pers Med 2021 Jun 17;11(6):569. [doi: 10.3390/jpm11060569] [Medline: 34204480]
99. Oh J, Kim S. The relationship between psychological distress, depressive symptoms, emotional eating behaviors and the health-related quality of life of middle-aged Korean females: a serial mediation model. BMC Nurs 2023 Apr 23;22(1):132. [doi: 10.1186/s12912-023-01303-y] [Medline: 37087430]
100. Benbaibeche H, Saidi H, Bounihi A, Koceir EA. Emotional and external eating styles associated with obesity. J Eat Disord 2023;11(1):67. [doi: 10.1186/s40337-023-00797-w]
101. Kudlek L, Jones RA, Hughes C, et al. Experiences of emotional eating in an acceptance and commitment therapy based weight management intervention (SWiM): a qualitative study. Appetite 2024 Feb 1;193:107138. [doi: 10.1016/j.appet.2023.107138] [Medline: 38016600]

XSL•FO
RenderX

102. Spreckley M, Seidell J, Halberstadt J. Perspectives into the experience of successful, substantial long-term weight-loss maintenance: a systematic review. Int J Qual Stud Health Well-being 2021 Dec;16(1):1862481. [doi: 10.1080/17482631.2020.1862481] [Medline: 33455563]
103. Jones RA, Lawlor ER, Birch JM, et al. The impact of adult behavioural weight management interventions on mental health: a systematic review and meta-analysis. Obes Rev 2021 Apr;22(4):e13150. [doi: 10.1111/obr.13150] [Medline: 33103340]
104. Robson EK, Hodder RK, Kamper SJ, et al. Effectiveness of weight-loss interventions for reducing pain and disability in people with common musculoskeletal disorders: a systematic review with meta-analysis. J Orthop Sports Phys Ther 2020 Jun;50(6):319-333. [doi: 10.2519/jospt.2020.9041] [Medline: 32272032]
105. Roosen MA, Safer D, Adler S, Cebolla A, van Strien T. Group dialectical behavior therapy adapted for obese emotional eaters; a pilot study. Nutr Hosp 2012;27(4):1141-1147. [doi: 10.3305/nh.2012.27.4.5843] [Medline: 23165554]
106. Hany M, Elfiky S, Mansour N, et al. Dialectical behavior therapy for emotional and mindless eating after bariatric surgery: a prospective exploratory cohort study. OBES SURG 2022 May;32(5):1570-1577. [doi: 10.1007/s11695-022-05983-4] [Medline: 35212910]
107. Ufholz K. Peer support groups for weight loss. Curr Cardiovasc Risk Rep 2020 Oct;14(10). [doi: 10.1007/s12170-020-00654-4]
108. Borek AJ, Abraham C, Greaves CJ, Tarrant M. Group‐based diet and physical activity weight‐loss interventions: a systematic review and meta‐analysis of randomised controlled trials. Applied Psych Health & Well 2018 Mar;10(1):62-86. [doi: 10.1111/aphw.12121]
109. Chen Y, Li Z, Yang Q, et al. The effect of peer support on individuals with overweight and obesity: a meta-analysis. Iran J Public Health 2021 Dec;50(12):2439-2450. [doi: 10.18502/ijph.v50i12.7926] [Medline: 36317042]
110. Muralidharan A, Niv N, Brown CH, et al. Impact of online weight management with peer coaching on physical activity levels of adults with serious mental illness. PS (Wash DC) 2018 Oct 1;69(10):1062-1068. [doi: 10.1176/appi.ps.201700391]
111. Simpson SA, Matthews L, Pugmire J, et al. An app-, web- and social support-based weight loss intervention for adults with obesity: the 'HelpMeDoIt!' feasibility randomised controlled trial. Pilot Feasibility Stud 2020 Dec;6(1):133. [doi: 10.1186/s40814-020-00656-4]
112. Baillot A, Chenail S, Polita NB, et al. Motives, barriers and preferences to physical activity in people with obesity: a systematic review. SportRxiv Preprint posted online on Apr 26, 2021. [doi: 10.31236/osf.io/rvaw3]
113. Raiman L, Amarnani R, Abdur-Rahman M, Marshall A, Mani-Babu S. The role of physical activity in obesity: let's actively manage obesity. Clin Med (Northfield) 2023 Jul;23(4):311-317. [doi: 10.7861/clinmed.2023-0152]
114. Deslippe AL, Soanes A, Bouchaud CC, et al. Barriers and facilitators to diet, physical activity and lifestyle behavior intervention adherence: a qualitative systematic review of the literature. Int J Behav Nutr Phys Act 2023 Feb 14;20(1):14. [doi: 10.1186/s12966-023-01424-2] [Medline: 36782207]
115. Milne-Ives M, Burns L, Swancutt D, et al. The effectiveness and usability of online, group-based interventions for people with severe obesity: a systematic review and meta-analysis. Int J Obes 2025 Apr;49(4):564-577. [doi: 10.1038/s41366-024-01669-2]
116. Larbi D, Wynn R, Trondsen MV, et al. What do adults living with obesity want from a chatbot for physical activity? – a qualitative study. BMC Digit Health 2024;2(1):15. [doi: 10.1186/s44247-024-00070-3]
117. Power S, Rowley N, Flynn D, Duncan M, Broom D. Home-based exercise for adults with overweight or obesity: a rapid review. Obes Res Clin Pract 2022;16(2):97-105. [doi: 10.1016/j.orcp.2022.02.003] [Medline: 35183471]
118. Perna S, Spadaccini D, Riva A, et al. A path model analysis on predictors of dropout (at 6 and 12 months) during the weight loss interventions in endocrinology outpatient division. Endocrine 2018 Sep;61(3):447-461. [doi: 10.1007/s12020-018-1563-y] [Medline: 29470776]
119. Moroshko I, Brennan L, O'Brien P. Predictors of dropout in weight loss interventions: a systematic review of the literature. Obes Rev 2011 Nov;12(11):912-934. [doi: 10.1111/j.1467-789X.2011.00915.x] [Medline: 21815990]
120. Taylor K, Indulkar T, Thompson B, et al. Early outcomes of referrals to the English National Health Service Digital Weight Management Programme. Obesity (Silver Spring) 2024 Jun;32(6):1083-1092. [doi: 10.1002/oby.24024] [Medline: 38644161]
121. Kupila SKE, Joki A, Suojanen LU, Pietiläinen KH. The effectiveness of eHealth interventions for weight loss and weight loss maintenance in adults with overweight or obesity: a systematic review of systematic reviews. Curr Obes Rep 2023 Sep;12(3):371-394. [doi: 10.1007/s13679-023-00515-2] [Medline: 37354334]
122. Batsis JA, Petersen CL, Clark MM, et al. Feasibility and acceptability of a technology-based, rural weight management intervention in older adults with obesity. BMC Geriatr 2021 Jan 12;21(1):44. [doi: 10.1186/s12877-020-01978-x] [Medline: 33435877]
123. Nicholls W, Lloyd J, Shepherd K, McArdle P, Tellwright H, Devonport TJ. Digital consultations for weight management in the NHS: a qualitative evaluation. Obes Res Clin Pract 2023;17(2):158-165. [doi: 10.1016/j.orcp.2023.03.003] [Medline: 37062675]
124. Fayyaz K, Bataineh MF, Ali HI, Al-Nawaiseh AM, Al-Rifai' RH, Shahbaz HM. Validity of measured vs. self-reported weight and height and practical considerations for enhancing reliability in clinical and epidemiological studies: a systematic review. Nutrients 2024 May 30;16(11):1704. [doi: 10.3390/nu16111704] [Medline: 38892637]

XSL•FO
RenderX

125. Arterburn D, Ichikawa L, Ludman EJ, et al. Validity of clinical body weight measures as substitutes for missing data in a randomized trial. Obes Res Clin Pract 2008 Dec;2(4):277-281. [doi: 10.1016/j.orcp.2008.09.002]
126. Olfert MD, Barr ML, Charlier CM, et al. Self-reported vs. measured height, weight, and BMI in young adults. Int J Environ Res Public Health 2018 Oct 11;15(10):2216. [doi: 10.3390/ijerph15102216] [Medline: 30314261]
127. Bramante CT, Lee G, Amsili SS, et al. Minority and low-income patients are less likely to have a scale for self-weighing in their home: A survey in primary care. Clin Obes 2020 Aug;10(4):e12363. [doi: 10.1111/cob.12363] [Medline: 32383356]
128. Iezzoni LI, Rao SR, Ressalam J, et al. Use of accessible weight scales and examination tables/chairs for patients with significant mobility limitations by physicians nationwide. Jt Comm J Qual Patient Saf 2021 Oct;47(10):615-626. [doi: 10.1016/j.jcjq.2021.06.005] [Medline: 34364797]
129. McCartney G, Hoggett R. How well does the Scottish Index of multiple deprivation identify income and employment deprived individuals across the urban-rural spectrum and between local authorities? Public Health (Fairfax) 2023 Apr;217:26-32. [doi: 10.1016/j.puhe.2023.01.009]

Abbreviations

CBT: cognitive behavioral therapy
CIMSPA: Chartered Institute for the Management of Sport and Physical Activity
CORE-OM: Clinical Outcomes in Routine Evaluation – Outcome Measure
DBT: Dialectical Behavioral Therapy
DERS-16: Difficulties in Emotion Regulation Scale
DWMI: Digital Weight Management Intervention
EEQ: Emotional Eater Questionnaire
EQ-5D-5L: EuroQol 5 Dimensions, 5 Levels
GPPAQ: General Practice Physical Activity Questionnaire
HCPC: Health and Care Professions Council
HRA: Health Research Authority
IMD: index of multiple deprivation
MDT: Multidisciplinary Team
NHS: National Health Service
NICE: National Institute for Health and Care Excellence
SQUIRE: Standards for Quality Improvement Reporting Excellence
SWMS: Specialist Weight Management Service
UKVRN: UK Voluntary Register of Nutritionists
VAS: Visual Analog Scale

Edited by A Mavragani; submitted 03.02.25; peer-reviewed by C Sorrell, H Dale, M Alkharaiji; revised version received 18.08.25; accepted 19.08.25; published 15.10.25.

Please cite as:
Spaltro G, Whitman M, Richards R
A Remotely Delivered Weight Management Service to Support Existing Obesity Services in the UK National Health Service: Preliminary Findings From an Early-Stage Service Evaluation
JMIR Form Res 2025;9:e71914
URL: https://formative.jmir.org/2025/1/e71914
doi:10.2196/71914

© Giulia Spaltro, Michael Whitman, Rebecca Richards. Originally published in JMIR Formative Research (https://formative.jmir.org), 15.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Daily Household Electricity Consumption in Community-Dwelling Older Individuals With Cognitive Impairment: Prospective Cohort Study

Yuki Nakagawa[1,2], PT, MS; Shigeo Tanabe[3,4], PT, PhD; Hikaru Kondo[1,2,5], PT, PhD; Koki Tan[1], PT, PhD; Soichiro Koyama[3,4], PT, PhD; Shin Kitamura[3,4], OT, PhD; Akiko Kada[6], MPH; Takuma Ishihara[7], PhD; Takuaki Yamamoto[8], ME; Junya Denda[8], ME; Hideaki Kimata[9], BE; Taisuke Yamanaka[9], ME; Ryosuke Umezawa[9], AE; Yoshinobu Nakahashi[10], PhD; Yohei Otaka[1,4], MD, PhD

[1]Department of Rehabilitation Medicine, School of Medicine, Fujita Health University, 1-98 Dengakugakubo, Kutsukake, Toyoake, Japan
[10]Japan Data Science Consortium Co. Ltd., Bunkyo-ku, Japan
[2]Graduate School of Health Sciences, Fujita Health University, Toyoake, Japan
[3]Faculty of Rehabilitation, School of Health Sciences, Fujita Health University, Toyoake, Japan
[4]Research Center for Robotic Smart Home and Activity Assistive Technology, Fujita Health University, Toyoake, Japan
[5]Department of Rehabilitation, Fujita Health University Hospital, Toyoake, Japan
[6]Center for Translational Research, Fujita Health University, Toyoake, Japan
[7]Innovative and Clinical Research Promotion Center, Gifu University, Gifu, Japan
[8]Chubu Electric Power Co Inc, Nagoya, Japan
[9]Necolico LLC, Chiyoda-ku, Japan

Corresponding Author:
Yohei Otaka, MD, PhD
Department of Rehabilitation Medicine, School of Medicine, Fujita Health University, 1-98 Dengakugakubo, Kutsukake, Toyoake, Japan

Abstract

Background: Various digital biomarkers have been explored to detect cognitive impairment in community-dwelling older individuals, among which electricity consumption (EC) data obtained from smart meters are novel and promising because they pose no burden to the individuals.

Objective: The study aimed to explore the potential of EC as a digital biomarker to screen older individuals with cognitive impairment living alone.

Methods: We recruited 40 older individuals living alone and recorded their 1-year daily household EC data. We used the Japanese version of the Montreal Cognitive Assessment to categorize participants into 2 groups: those with and without cognitive impairment. As the pattern of daily household EC is different between lower and higher temperature ranges because of the use of heating and cooling equipment, we divided the daily household EC into 3 temperature ranges. Using a linear mixed model, we evaluated the association between daily household EC, daily outside temperature, and the groups.

Results: After excluding 12 participants, they were categorized into 2 groups: those with (10/28, 36%) and without cognitive impairment (18/28, 64%). The daily household EC data consisting of 9391 points showed two characteristics: (1) daily household EC was significantly lower in the group with cognitive impairment than in the group without cognitive impairment in the high temperature range (2.158 kWh at 25 °C, P=.02; 3.712 kWh at 30 °C, P<.001). The increase in EC with rising temperature from 25 °C to 30 °C was less in the group with cognitive impairment (2.387 kWh, P<.001) than in the group without cognitive impairment (3.940 kWh, P<.001); and (2) a tendency for lower daily household EC in the group with cognitive impairment was observed in the moderate temperature range (1.795 kWh at 15 °C, P=.06; 1.582 kWh at 20 °C, P=.08).

Conclusions: The group with cognitive impairment may use less cooling equipment in the high temperature range and fewer home appliances in the moderate temperature range. Daily household EC might be useful in screening cognitive impairment in older individuals living alone.

(JMIR Form Res 2025;9:e71265) doi:10.2196/71265

KEYWORDS

detection; digital biomarker; nonwearable device; prevention; screening; sensor; smart meter; thermal sensitivity

XSL•FO
RenderX

Introduction

Mild cognitive impairment (MCI) is an early stage of cognitive decline, often involving memory loss or other cognitive deficits, while individuals retain the ability to perform most activities of daily living (ADL) independently [1]. Although MCI is associated with a high risk of dementia, individuals with MCI may return to normal cognitive function [1], with a reversion rate of approximately 18% (95% CI 14% - 22%) [2]. In addition, interventions targeting lifestyle and ADL at the MCI stage may delay or prevent further cognitive decline [3]. Therefore, early detection of MCI and subsequent early intervention are important [4]. However, most patients with cognitive decline visit a hospital after their symptoms have substantially progressed, and only a few patients visit a hospital while they are at the MCI stage [5].

Various researchers have attempted to develop assessment methods that allow for the screening of community-dwelling older individuals in their own homes. Recently, readily available digital biomarkers, such as smartwatch-based step counters and smart ring–based continuous heart rate monitors, have attracted growing interest in cognitive assessment, which takes advantage of widely available mobile and wearable technologies [6]. However, older individuals require more support than younger individuals when using digital technology. Specifically, older individuals need simpler and more user-friendly mobile devices, such as fewer buttons, larger font sizes, and better color contrast. In this context, digital biomarkers, which can be obtained from sensor devices installed in homes, are considered promising for detecting MCI as they work without user interaction [6].

Electricity consumption (EC) data obtained using smart meters have garnered attention as novel digital biomarkers [7]. EC data are advantageous because they do not require user interaction and can be obtained while monitored individuals are living naturally at home. Numerous countries have been promoting their installations in households, with an estimated 729.13 million installations worldwide as of 2019 [8].

In a previous study, the association between the lifestyles and behaviors of older individuals with cognitive impairment was explored using household EC data obtained from an additional specialized electricity sensor with high precision and high sampling frequency [9]. Specifically, the ADL were estimated by identifying the amount of time each home appliance was used [9]. The use of common smart meters to assess behaviors requires no additional cost and imposes no burden on device management or invasion of privacy. Because patients with MCI perform various characteristic instrumental ADL (IADL) [10], the early detection of MCI using ordinary EC data obtained from common smart meters may be feasible.

For the early detection of MCI, it is important to focus on older individuals living alone as the risk of developing dementia is increased in these individuals (relative risks=1.30, 95% CI 1.15 - 1.46) [11]. Screening for cognitive function in older individuals living alone is useful because their families are unlikely to notice their cognitive decline.

In this study, we aimed to explore the possibility of using ordinary EC data as a digital biomarker to screen older individuals with cognitive impairment living alone.

Methods

Study Design and Participant Recruitment

This was an exploratory prospective cohort study with a convenience sample. Data were obtained between April 1, 2022, and March 31, 2023. Study participants were recruited from Toyoake City in the Aichi prefecture, located in the center of the Japanese archipelago (at a latitude of 35° north) on September 9, 2021. The participants were recruited from local senior centers, city halls, and other community-based locations. The inclusion criteria were as follows: (1) aged ≥65 years, (2) living alone and independently, (3) never diagnosed with dementia, (4) users of smart meters, and (5) consent to provide EC data. The exclusion criteria were as follows: (1) individuals using electric water heaters that caused characteristic EC patterns and (2) individuals whose EC data were determined to be inappropriate for analysis based on visual inspection and researcher consensus. The latter criterion was applied post hoc to exclude data with clear anomalies or artifacts (eg, persistent signal dropout or unusually large use patterns) that could not be reasonably attributed to physiological or behavioral factors. The process was carefully documented and applied consistently across all participants.

Ethical Considerations

This study was approved by the Research Ethics Committee of Fujita Health University (HM23-108) and reported in accordance with the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines (Checklist 1). All participants provided written informed consent before enrollment. Participants were informed that their data would be used exclusively for research purposes. To ensure privacy and confidentiality, all collected data were anonymized before analysis and stored in encrypted formats on a secure computer, accessible only to authorized study personnel. No personally identifiable information is included in the manuscript, and individual participants cannot be identified from any images, text, or supplementary materials. Participants did not receive any monetary compensation; however, they were provided with feedback on their health screening results.

Smart Meters

Smart meters are devices that measure EC in real time and automatically transmit information to electricity companies or individual households in a digital format [8]. The common smart meters used in Japan automatically transmit information data on household EC (the total amount of electricity used) every 30 minutes and are typically installed outside homes (Figure 1). According to a survey by the Ministry of Economy, Trade, and Industry of Japan, smart meters were installed in approximately 85% (69.17 million units) of households in Japan by March 2021 [12]. For a general purpose, data on household EC are transmitted to electricity companies and used to calculate electricity charges and understand the overall energy demand in the region. In addition, each household can connect a home

XSL•FO
RenderX

energy management system (HEMS) device, which is a home energy monitoring system, to a smart meter to receive relevant data at any given point. In this study, we used 30-minute interval EC data obtained directly from common smart meters via the wireless smart utility network and internet with HEMS and a built-in home gateway.

Figure 1. Example of a smart meter commonly used in Japan to collect household electricity consumption data. This figure shows a typical smart meter, usually installed outside Japanese residences. In this study, 30-minute interval electricity consumption data in watt-hours were obtained through these meters using a home energy management system, a home gateway, and the wireless smart utility network. This image is for illustrative purposes only and does not depict the actual device used in the study.

Data Collection

The basic sociodemographic data, cognitive function, and physical function of the study participants were assessed by 3 health care professionals (physical therapists) at the time of participation in the study. The collection and assessment items included age, sex assigned at birth, residence information, Japanese version of the Montreal Cognitive Assessment (MoCA-J) score [13], Frailty Screening Index [14], Kihon Checklist [15], walking speed [16], and grip strength [17].

The MoCA-J is a Japanese version of the Montreal Cognitive Assessment (MoCA) that is used to evaluate cognitive function [13]. The MoCA is used to detect MCI in older individuals [18]. MoCA-J scores range from 0 to 30, with lower scores indicating cognitive decline. Participants with a MoCA-J score ≤25 were classified as the group with cognitive impairment, whereas those with a score ≥26 were classified as the group without cognitive impairment [13]. Groups with and without cognitive impairment were the primary comparison variables in this study. The MoCA-J is more sensitive than the Mini-Mental State Examination for detecting MCI in older individuals (sensitivity 93%; specificity 89%) [13]. This test has been widely used in numerous large-scale studies on cognitive function. Furthermore, the reliability and validity of the MoCA in detecting MCI have been well established in Japanese populations compared to conventional cognitive tests [13].

The Frailty Screening Index is used to test frailty and consists of 5 self-report questionnaires [14]. If 3 or more questions are answered “yes,” the patient may have frailty and is at increased risk of death or need for nursing care [14]. The Kihon Checklist is widely used in Japan to identify older individuals at risk for requiring support or nursing care [15]. A score of ≥8 out of 25 items indicates frailty [15].

Walking speed is an indicator of lower limb muscle strength. Older individuals with a slower walking speed (<0.6 m/s) have a higher risk of future hospitalization and deterioration of health status than those with a faster walking speed (≥0.6 m/s) [16]. Similarly, grip strength is an indicator of upper limb muscle strength and overall physical function. It is used as a cutoff value for sarcopenia in Asians, defined as the ability to grip <28 kg for men and <18 kg for women [17].

The time-stamped 30-minute values of household EC were obtained from a smart meter. Daily household EC was calculated using 30-minute values×48 per day obtained between April 1, 2022, and March 31, 2023. For data on daily household EC that were completely or partially missing during the study period,

XSL•FO
RenderX

the 1-day data for the corresponding participant were excluded from the analysis.

Outside temperature data were obtained from the Japan Meteorological Agency website [19]. The daily outside temperature was the mean of the hourly values observed from 12 AM to 12 PM. Data from Obu City, a city neighboring Toyoake City, were used because the mean temperature of Toyoake City was unavailable. The city hall distance between the experimental site and the outside temperature observation site was approximately 8 km, and the elevation difference was <2 m.

Statistical Analyses

Data on participant demographics, cognitive function, and physical function were analyzed using descriptive statistics. The characteristics of the participants with and without cognitive impairment were compared using the Fisher exact test for categorical variables and the Mann-Whitney *U* test for continuous variables. Annual trends in daily household EC and their association with daily outside temperature were presented in scatter plots or lines.

Daily individual household EC data were repeatedly measured for 1 year. Therefore, a linear mixed model introducing a random intercept was used to analyze the daily household EC. The normality of the residuals was verified using histograms. The linear mixed model included daily outside temperature, study group, and the interaction between daily outside temperature and group as fixed effects and participants as random effects. To account for nonlinearity, a cubic spline was considered for temperature. Based on previous findings [20], 25 °C, the boundary between high and medium temperatures, and 15 °C, the boundary between medium and low temperatures, were set as knots for the spline. Significant interaction indicated different daily household EC and daily outside temperature associations between the groups. Based on the predicted values obtained from the model, a contrast analysis was performed on the difference in the daily household EC between groups for each temperature. Because the model includes nonlinear effects of temperature, interaction terms, and main effects, we evaluated the differences between groups by comparing the predicted values at specific temperatures rather than performing statistical tests on individual coefficients. Furthermore, the difference for every 5 °C change in daily household EC was calculated for each group. As the mechanism of missing power consumption data can be assumed to be missing at random, we considered estimation without bias possible using a linear mixed model. Analyses were performed using R (version 4.3.1; R Foundation for Statistical Computing) and the *lme4* package. The 2-sided significance level was set at $P<.05$. Since this study was exploratory, no adjustments were made for multiple comparisons.

Results

Participants

The participant flow diagram is shown in Figure 2. A detailed breakdown of the reasons for exclusion is provided in this diagram. A total of 40 individuals met the initial inclusion criteria. After excluding 12 participants, 28 were enrolled in the study. Based on the MoCA-J scores, 10 (36%) participants were categorized as the group with cognitive impairment and 18 (64%) participants were categorized as the group without cognitive impairment.

XSL•FO
RenderX

Figure 2. Flow diagram of inclusion and exclusion criteria for participants among community-dwelling older individuals in Toyoake City, Japan (April 2022-March 2023): an exploratory prospective cohort study. This figure illustrates the screening process, inclusion and exclusion criteria, and final sample size for a prospective cohort study investigating the association between cognitive impairment and household electricity consumption (EC) in community-dwelling older individuals in Toyoake City, Japan.

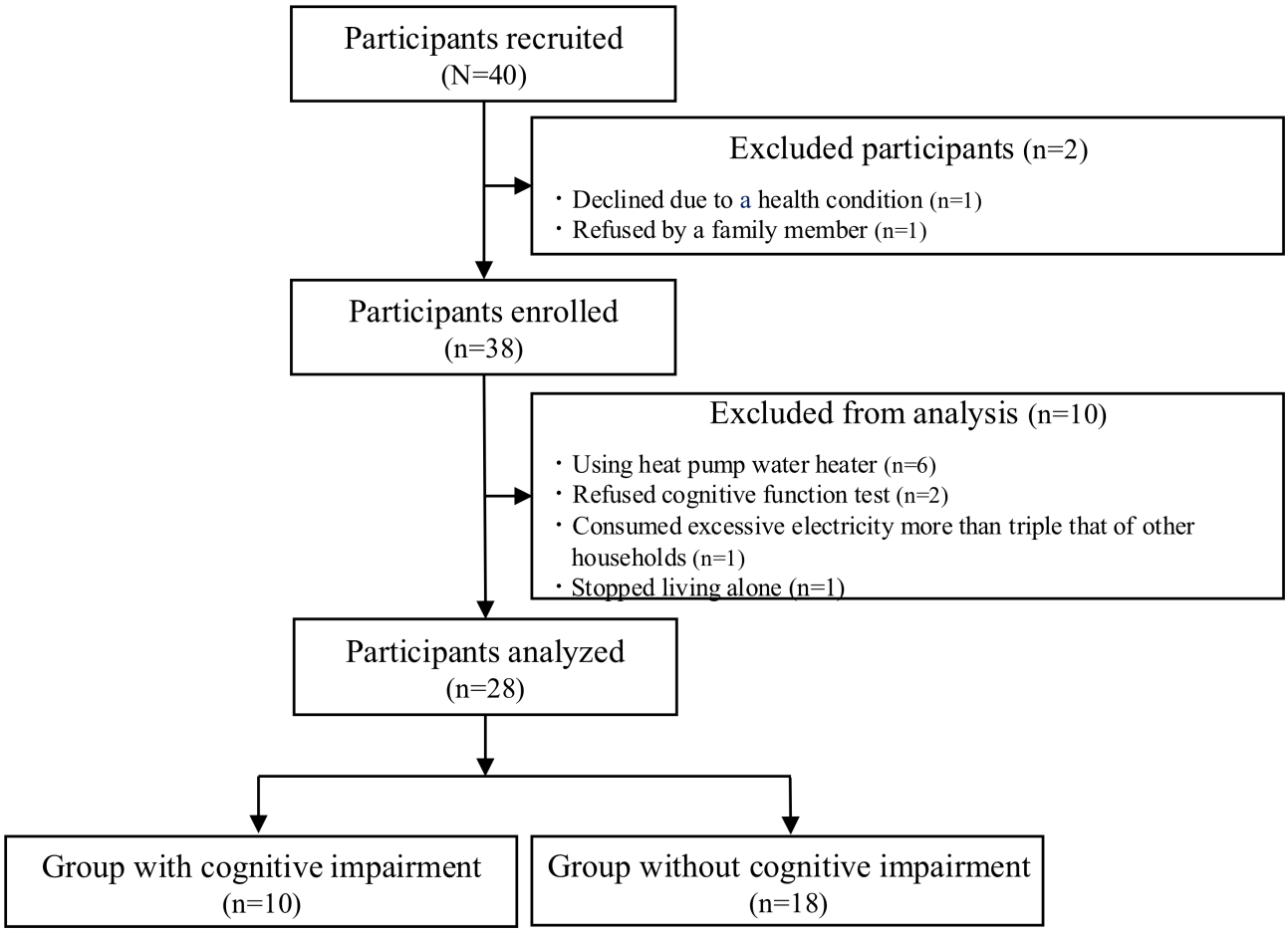

Participant Characteristics

Table 1 shows the participant demographics and the results of cognitive and physical function tests. A significant difference in MoCA-J scores was observed between the two groups (P<.001). However, no significant differences in age, sex assigned at birth, residence information, Frailty Screening Index, Kihon Checklist, walking speed, or grip strength were observed between the 2 groups.

XSL•FO
RenderX

Table . Sociodemographic and clinical characteristics of the participants in Toyoake City, Japan (April 2022-March 2023).

Variable	Group with cognitive impairment (n=10)	Group without cognitive impairment (n=18)	*P* value
Age (y), median (IQR)	77 (76 - 83)	79 (74 - 86)	.76
Sex, n (%)			.43
Male	3 (30)	4 (22.2)	
Female	7 (70)	14 (77.8)	
Type of residence, n (%)			.38
House	6 (60)	15 (83.3)	
Apartment	4 (40)	3 (16.7)	
MoCA-J[a] (score), median (IQR)	21.5 (20 - 24.5)	28 (27 - 29)	<.001
Frailty screening index (score), median (IQR)	1 (0 - 2)	0 (0 - 1)	.10
Kihon checklist (score), median (IQR)	2 (1–6)	3 (2–5)	.54
Gait speed (m/s), median (IQR)	1.2 (1.1 - 1.4)	1.3 (1.0 - 1.4)	.75
Grip strength (kg), median (IQR)	24.1 (17.8 - 34.9)	26.6 (21.7 - 30.1)	.95
Male	36.5 (35.4 - 40.3)	34.6 (26.1 - 38.0)	.63
Female	17.8 (17.5 - 24.8)	24.8 (21.5 - 27.8)	.13

[a]MoCA-J: Japanese version of the Montreal Cognitive Assessment.

Data Points

A total of 10,220 data points were collected from 28 participants over 365 days. Of these, 515 (5%) data points were missing owing to the data extraction schedule, and 314 (3.1%) data points were missing owing to poor connections between the wireless local area networks, HEMS devices, and smart meters. Consequently, there were 9391 (91.9%) available data points.

Daily Household EC and Daily Outside Temperature

Figures 3A and 3B show the changes in daily household EC for all study participants and the daily outside temperature of the city near the study location, respectively, from April 1, 2022, to March 31, 2023. The daily household EC tended to increase in both summer (July to September) and winter (December to February), while the daily outside temperature increased during summer and decreased in winter. The daily household EC ranged from 1.4 kWh to 51.1 kWh, and the daily outside temperature ranged from −1.1 °C to 32.2 °C during the study period. Figure 4 shows the association between daily outside temperature and daily household EC. In both low and high temperature ranges, the daily household EC increased as the temperature became colder or warmer. The daily household EC tended to be lower in the group with cognitive impairment than in the group without cognitive impairment.

XSL•FO
RenderX

Figure 3. Daily household electricity consumption (EC) data and outside temperature for the year (Apr [April] 2022-Mar [March] 2023). (**A**) Daily household EC data collected from 28 community-dwelling older individuals in Toyoake City, Japan. The horizontal axis represents the month, and the vertical axis shows household EC in kilowatt-hours (kWh). (**B**) Daily outside temperature data in Obu City, Japan, which is adjacent to Toyoake City. The horizontal axis represents the month, and the vertical axis shows the daily outside temperature (°C). The lines represent the daily outside temperature for each day. Aug: August; Dec: December; Feb: February; Jan: January; Jul: July; Jun: June; Nov: November; Oct: October; Sep: September.

XSL•FO
RenderX

(A)

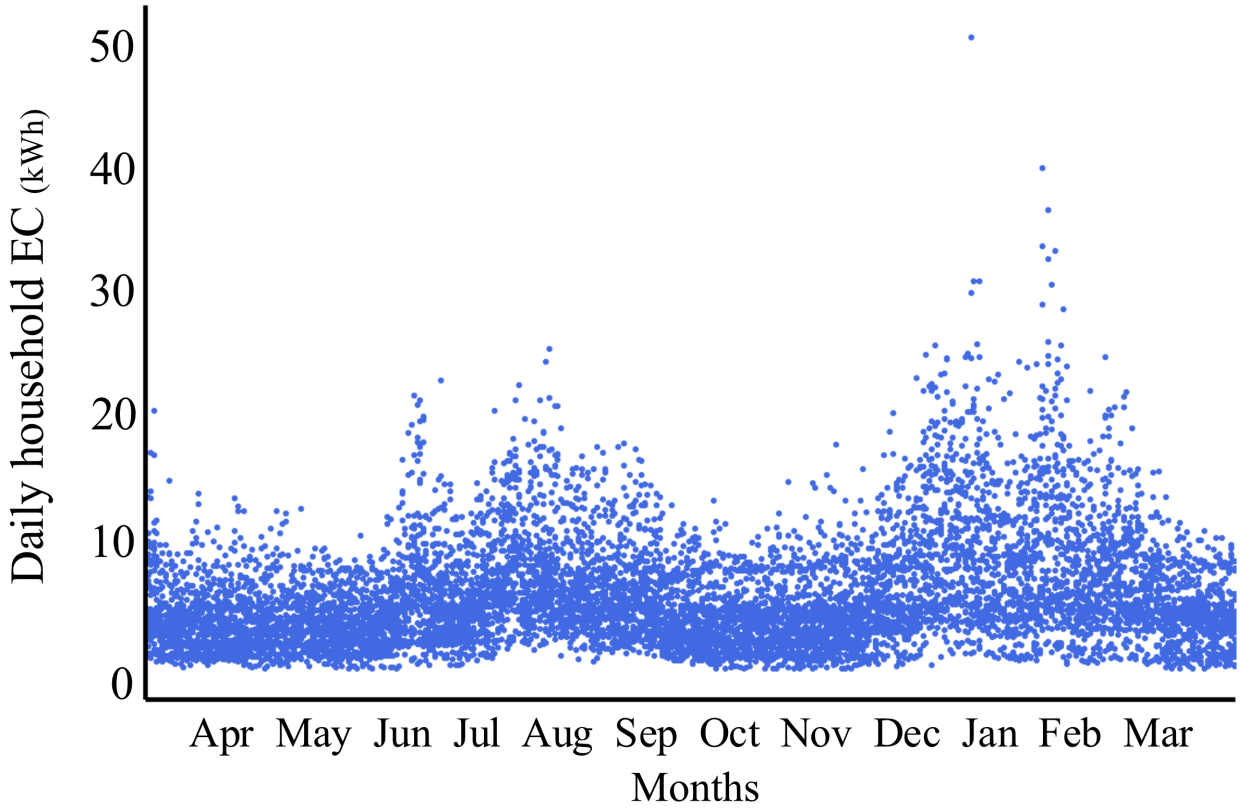

(B)

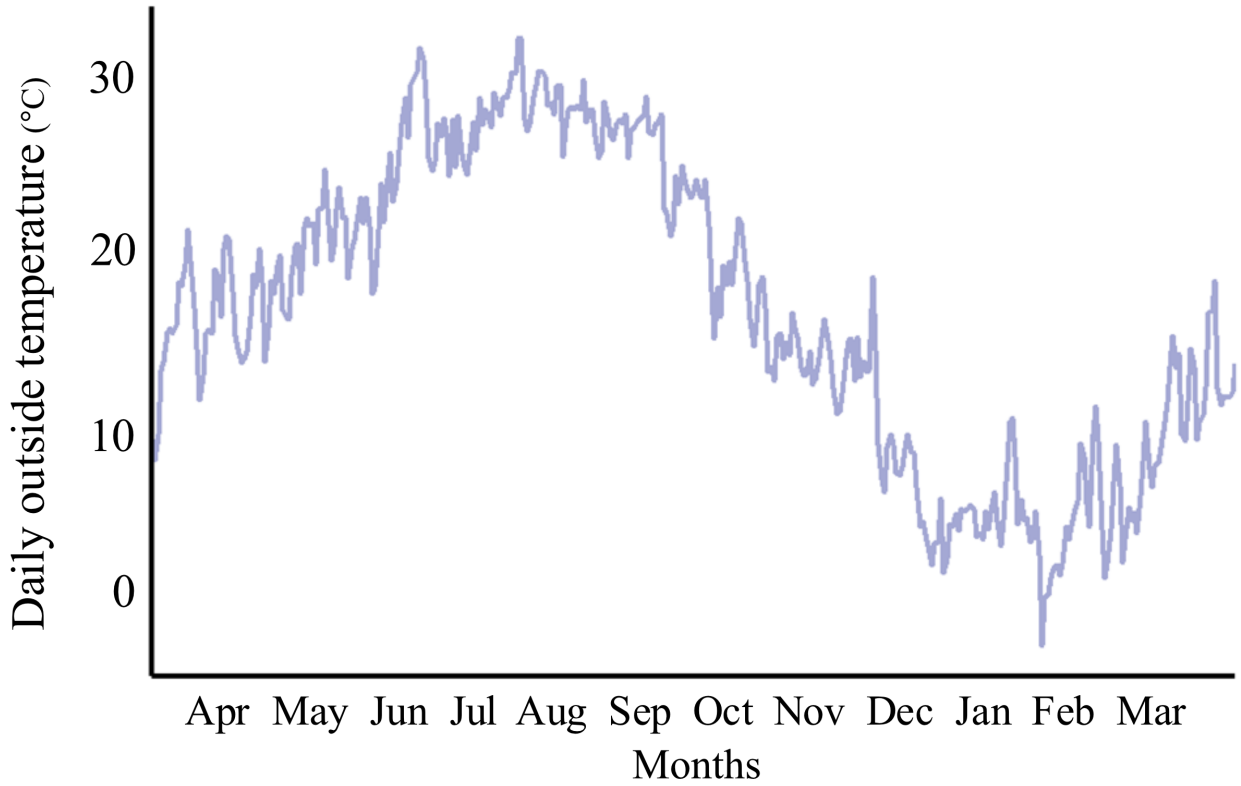

XSL•FO
RenderX

Figure 4. Scatter plot showing the association between daily household electricity consumption (EC) and outside temperature (April 2022-March 2023). Each dot represents 1 observation from a total of 28 community-dwelling older individuals. The horizontal axis shows the daily outside temperature (°C) in Obu City, Japan, which is adjacent to Toyoake City, Japan, and the vertical axis indicates daily household EC in Toyoake City (kWh). Red and blue dots represent individuals with (n=10) and without (n=18) cognitive impairment, respectively.

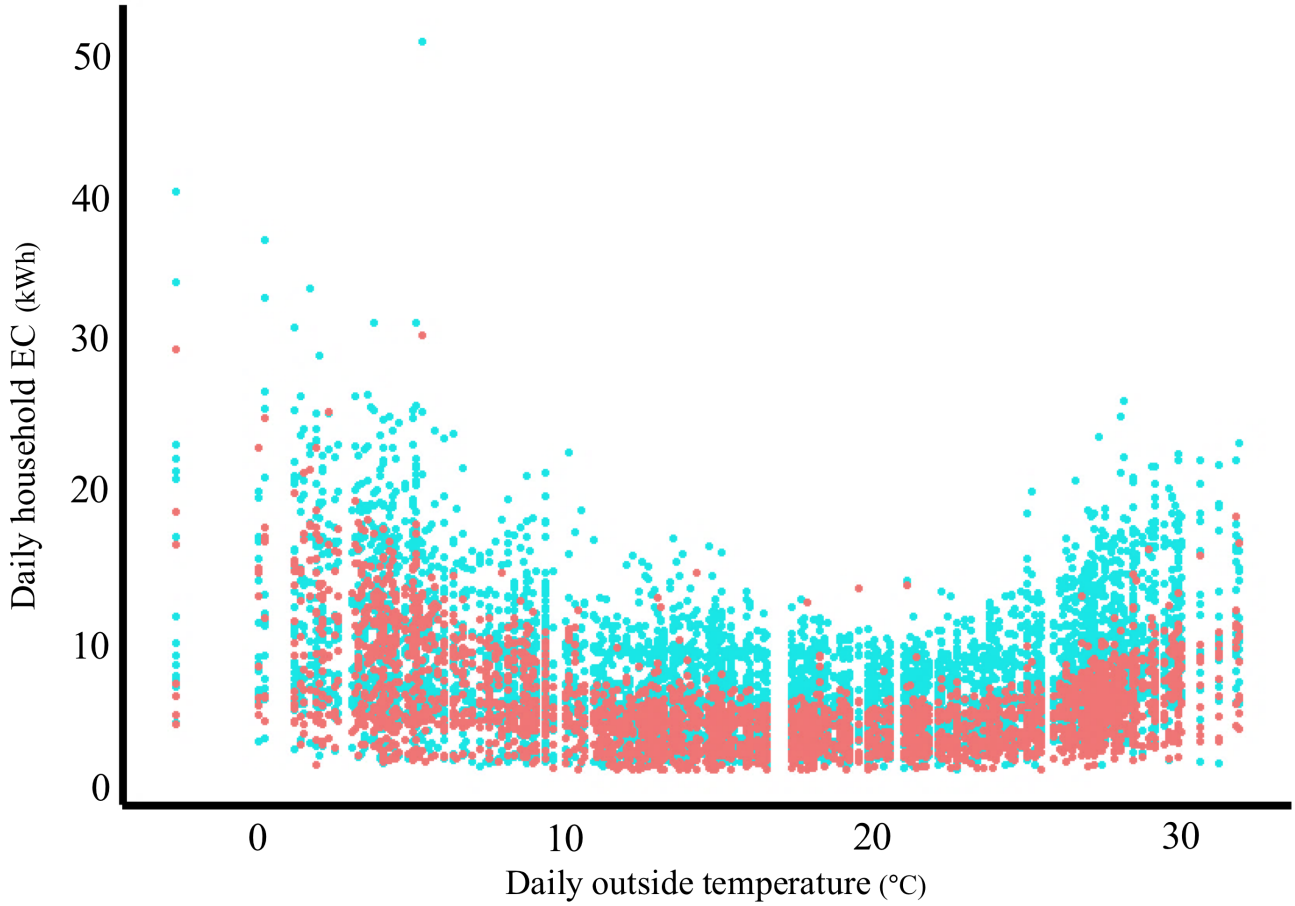

Association Between Daily Outside Temperature and the Groups on the Daily Household EC

The type III test of ANOVA for the interaction terms was significant ($P<.001$), suggesting that the association between daily household EC and outside temperature may differ in shape between the groups. Multimedia Appendix 1 presents the detailed results of the type III ANOVA, including the main effects of temperature and group, and their interaction. Table 2 presents the results of the contrast analysis based on the linear mixed model, and Table 3 summarizes the estimated differences between groups for each 5 °C increase in temperature. The *P* values reported in these tables were unadjusted for multiplicity. Figure 5 illustrates the predicted lines obtained from the linear mixed models. Details of the estimated fixed effects, including spline terms, are provided in Multimedia Appendix 2.

Table . Estimated daily household electricity consumption (EC) at fixed outside temperatures (April 2022-March 2023) based on least-squares means from a linear mixed model.

Temperature (°C)[a]	Least-squares means for the group with cognitive impairment[b] (kWh)	Least-squares means for the group without cognitive impairment[b] (kWh)	Differences (kWh)	*P* value
0	13.576	15.884	2.308	.19
5	9.833	11.464	1.631	.25
10	6.771	8.466	1.696	.14
15	4.837	6.631	1.795	.06
20	4.352	5.934	1.582	.08
25	5.128	7.286	2.158	.02
30	7.515	11.226	3.712	<.001

[a]Outside temperatures in Obu City, Japan, which is adjacent to Toyoake City, Japan (°C).

[b]Least-square means indicates the estimated mean daily household EC (kWh).

XSL•FO
RenderX

Table . Estimated differences in daily household electricity consumption (EC) across adjacent temperature intervals (April 2022-March 2023) based on linear mixed model estimates.

Temperature[a] interval (°C)	Difference[b] (kWh; 95% CI)	*P* value
Group with cognitive impairment		
5 - 0	−3.743 (−5.081 to −2.405)	<.001
10 - 5	−3.063 (−3.923 to −2.202)	<.001
15 - 10	−1.934 (−2.788 to −1.080)	<.001
20 - 15	−0.484 (−1.344 to 0.375)	.64
25 - 20	0.776 (−0.065 to 1.617)	.09
30 - 25	2.387 (1.465 to 3.309)	<.001
Group without cognitive impairment		
5 - 0	−4.420 (−5.380 to −3.460)	<.001
10 - 5	−2.998 (−3.635 to −2.361)	<.001
15 - 10	−1.835 (−2.468 to −1.202)	<.001
20 - 15	−0.697 (−1.334 to −0.061)	.02
25 - 20	1.352 (0.727 to 1.977)	<.001
30 - 25	3.940 (3.262 to 4.619)	<.001

[a]Outside temperatures in Obu City, Japan, which is adjacent to Toyoake City, Japan.

[b]The difference represents the estimated mean difference in daily household EC obtained from community-dwelling older individuals in Toyoake City between the specified temperature intervals.

XSL•FO
RenderX

Figure 5. Electricity consumption (EC) lines predicted based on daily outside temperature in the 2 groups (April 2022-March 2023). The horizontal axis represents daily outside temperature in Obu City, Japan, which is adjacent to Toyoake City, Japan (°C), and the vertical axis shows the predicted daily household EC (kWh) based on a linear mixed model. Solid lines represent group-level predicted means, and shaded areas indicate 95% CIs. Red and blue lines represent the groups with (n=10) and without (n=18) cognitive impairment, respectively.

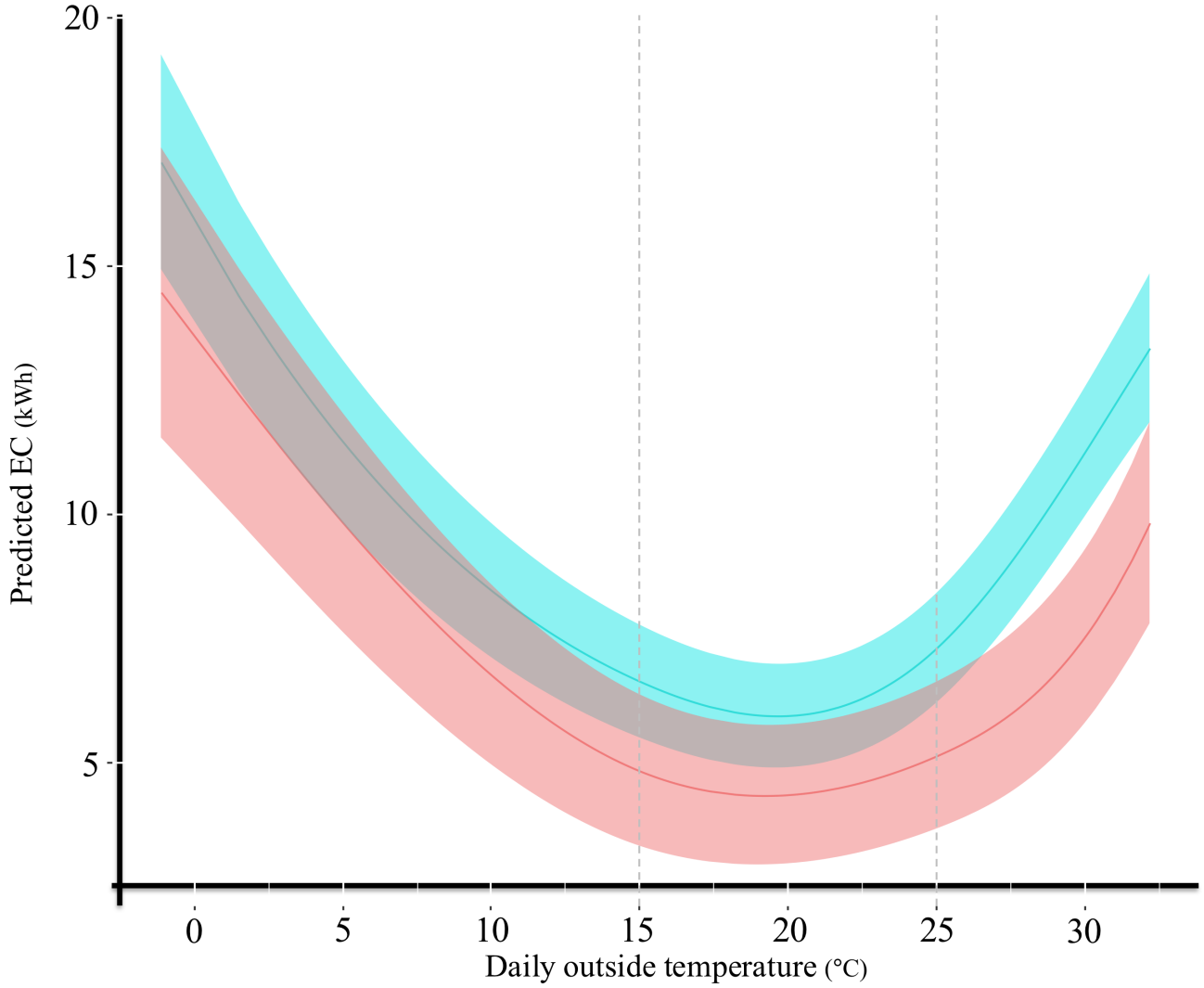

In the high temperature range, daily household EC was significantly lower in the group with cognitive impairment than in the group without cognitive impairment (2.158 kWh at 25 °C, *P*=.02; 3.712 kWh at 30 °C, *P*<.001). The increase in EC with rising temperature from 25 °C to 30 °C was less in the group with cognitive impairment (2.387 kWh; *P*<.001) than in the group without cognitive impairment (3.940 kWh; *P*<.001).

In the moderate temperature range, no significant differences were observed in daily household EC between the groups with changes in daily outside temperature. However, the group with cognitive impairment tended to show a lower daily household EC than the group without cognitive impairment at both 15 °C (1.795 kWh; *P*=.06) and 20 °C (1.582 kWh; *P*=.08).

In the low temperature range, no significant differences were observed in daily household EC between the groups with changes in outside temperature. However, a main effect of daily outside temperature was observed, indicating that daily household EC increased with decreasing temperature in both groups (group with cognitive impairment: 3.743 kWh from 0 °C to 5 °C, 3.063 kWh from 5 °C to 10 °C, 1.934 kWh from 10 °C to 15 °C, all *P*<.001; group without cognitive impairment: 4.420 kWh from 0 °C to 5 °C, 2.998 kWh from 5 °C to 10 °C, 1.835 kWh from 10 °C to 15 °C, all *P*<.001).

Discussion

Principal Findings

In this study, a combined analysis of daily household EC data obtained from common smart meters and daily outside temperature data suggests that there are differences in EC between individuals with and without cognitive impairment living alone. These results suggest that EC data could be useful as a digital biomarker for screening cognitive function in older individuals living alone.

Comparison With Prior Studies

The daily household EC of both groups tended to increase with changes in the daily outside temperature in both low and high temperature ranges. These trends can be attributed to the use of heating and cooling equipment, respectively. Daily household EC increased when daily outside temperatures rose or fell beyond 13 °C to 25 °C in Shanghai, China, owing to the influence of heating and cooling equipment use [20]. Similarly,

XSL•FO
RenderX

daily household EC increased when the highest outside temperature decreased or increased in the Kyushu region of Japan (latitude 30° - 33° north) [21].

In the high temperature range, the group with cognitive impairment used less electricity than the group without cognitive impairment in terms of the increase in daily household EC associated with the rise in outside temperature. This may be due to reduced sensitivity to the rise in outside temperature and the lack of appropriate indoor temperature control using cooling equipment. A mouse study indicated a delayed response latency to heat stimuli in mice with dementia compared to that in healthy mice [22]. In human studies, the threshold for temperature perception in older individuals with mild to moderate Alzheimer disease (AD) is higher than that in those without AD (median 34.0, IQR 33 °C-36 °C for older individuals with AD vs median 32.0, IQR 32 °C-34 °C for those without AD) [23]. Other previous studies have shown that older individuals with cognitive impairment have an increased risk of developing heat stroke [24,25]. Patients with psychiatric disorders, such as dementia, are at a higher risk of hospitalization than healthy individuals because of heat-related illnesses, including heat stroke [24]. The risk of hospitalization for patients with dementia increases by 4.5% (95% CI 2.9% - 6.1%) for every 1 °C increase in the daily mean temperature >17 °C [25]. Further research is needed to clarify whether reduced use of cooling equipment—potentially owing to a higher threshold for sensing elevated body temperature—in individuals with cognitive impairment contributes to their heightened vulnerability to heat stroke.

In the moderate temperature range, a trend of a lower daily household EC was observed in the group with cognitive impairment than in the group without cognitive impairment. This suggests that individuals in this group used fewer household appliances during the seasons when heating and cooling equipment are not typically used. IADL assessments have shown a decline in the ability of individuals with MCI to use household appliances compared to older individuals without cognitive impairment [26]. A previous study using an additional specialized electricity sensor with high precision and high sampling frequency has reported shorter use times for appliances such as induction heaters in spring for individuals with cognitive impairment [9].

In the low temperature range, the group with cognitive impairment showed a tendency toward lower daily household EC; however, no significant differences were observed between the two groups. This finding is inconsistent with those of a previous study, which demonstrated the effect of seasons on daily household EC, such as reduced use time of certain home appliances, including air conditioners, in winter among individuals with cognitive impairment [9]. Although the reason for this inconsistency remains unclear, it could be attributed to differences in individual sensitivity to cold temperatures. A previous study indicated that such sensitivities could vary widely, with variation in thermal sensation sensitivity reaching as high as 50% (ie, 1.5 times for the highest and 0.5 times for the lowest relative to the mean value) [27]. In addition, various findings on the impact of cognitive decline on sensitivity to low temperatures have also been reported, ranging from no difference to higher sensitivity compared with that of control groups [28,29]. For instance, a study involving mice indicated that the response latency of dementia model mice to cold stimuli (5 °C cold water) was equivalent to that of healthy mice [28]. In contrast, a study on humans found that individuals with mild to moderate AD had lower resistance to cold stimuli (immersing hands in water at 4 °C for 2 min) than individuals without cognitive impairment [29].

Although the reasons for the different characteristics between the low and high temperature ranges in this study are unclear, differences in neural pathways may have contributed. A study on the thermosensory neural network in mice revealed that the neural circuits responsible for identifying comfortable temperature environments differ according to their responses to heat and cold. Specifically, neuronal tracing identified 2 segregated groups of neurons in the lateral parabrachial nucleus (LPB): one projecting to the median preoptic nucleus (MnPO), a thermoregulatory center (LPB→MnPO neurons), and the other to the central amygdaloid nucleus (CeA), a limbic emotion center (LPB→CeA neurons). While LPB→MnPO neurons include separate subgroups activated by heat or cold exposure, LPB→CeA neurons were only activated by cold exposure [30].

Interpretation of Model Results

Although the interaction term between group and temperature in the linear mixed model was statistically significant, the interpretation of this result requires caution. Our model included both linear and nonlinear spline terms for temperature and their interactions with group. However, the significance of the interaction term does not directly imply uniform group differences across all temperature values. Rather, it suggests that the shape of the association between temperature and daily EC may differ between groups. To address this complexity, we performed contrast analyses using model-derived predicted values to estimate group differences at specific temperature levels (eg, 0 °C, 5 °C, and 10 °C). This approach allowed us to determine where meaningful differences were most likely to occur while accounting for the full structure of the model. The results from this contrast analysis indicated clear group differences in the high temperature range, where individuals with cognitive impairment used significantly less electricity. We believe this analytic strategy improves interpretability and helps avoid complex conclusions based on several regression coefficients that include interactions and nonlinear terms.

Limitations

As the number of participants in this study was small (N=28), a verification study with a larger sample size is required. The groups with and without cognitive impairment comprised 10 and 18 participants, respectively. However, the demographic characteristics and physical function of the study participants were generally similar between the two groups, except for the presence or absence of cognitive decline. This finding suggests that the study was not influenced by specific personal or environmental factors. Residential type (eg, house or apartment) may also influence EC. Although apartment residents tended to consume less electricity, the distribution of residence types was relatively balanced between the groups, suggesting that its impact on the group comparisons was likely minimal.

XSL•FO
RenderX

Nonetheless, future studies should explicitly consider residential type as a potential confounding factor.

Validation of these findings in other regions with different latitudes and climatic conditions is necessary. This study was conducted in specific regions of Japan, which limits the generalizability of our findings. However, the results may also be applicable to areas with mid-latitude, as well as warm and humid climates, similar to Toyoake City, Aichi Prefecture (35° north latitude; eg, coastal areas of Korea and southern China). In addition, while the study participants were older individuals living alone and were not engaged in paid employment, which we assumed resulted in relatively comparable income levels, individual income could still influence energy consumption patterns. Previous research has shown that as income increases, households' weather sensitivity remains the same for hotter days in the summer but increases during the winter [20]. In this study, we did not record the exact proportion of participants who had completed higher education. However, we acknowledge the potential for socioeconomic selection bias. Specifically, this may affect the generalizability of our findings to populations with different socioeconomic backgrounds. According to the Organisation for Economic Co-operation and Development [31], individuals with tertiary education earn nearly twice as much on average as those without upper secondary education across Organisation for Economic Co-operation and Development countries. Li et al [20] demonstrated that higher-income households exhibit greater sensitivity to temperature changes in winter, suggesting that EC patterns may vary according to socioeconomic status. Thus, our results may have been influenced by this socioeconomic selection bias. This factor was not explicitly controlled for in our study, and future research should examine the potential impact of household income on energy consumption behavior. Furthermore, this study focused primarily on daily outdoor temperature as the environmental factor influencing EC. Other environmental factors, such as variations in daylight hours, were not considered. Daylight duration may affect daily activity rhythms and, by extension, patterns of EC, particularly among older individuals with cognitive impairment. Prior research has demonstrated that changes in light exposure can influence rest-activity patterns, mood, and agitation levels in individuals with dementia [32].

Conclusions

In summary, the use of cooling equipment in the high temperature range and various household appliances in the moderate temperature range might have influenced the differences in daily household EC between the groups with and without cognitive impairment. If changes in cognitive function can be longitudinally revealed using daily household EC, further advancements in research can enable the screening of cognitive function in older individuals living alone without the psychological and physical burdens associated with conventional screening of cognitive function. Reducing hospital visits for cognitive function assessment and monitoring of physical and cognitive conditions by health care professionals may contribute to the improvement of health management and quality of medical care for older individuals. In addition, these findings may lead to the development of new technologies to support ADL and IADL in older individuals.

This study revealed, for the first time, differences in daily household EC between individuals with and without cognitive impairment using only data obtained from common smart meters without additional specialized sensors. In addition, this study demonstrated the possibility of using daily household EC data with daily outside temperature as a digital biomarker for the early detection of cognitive decline in older individuals living alone. These findings may have major implications for health care management since early detection of cognitive decline can enable early intervention.

Acknowledgments

This study was partially supported by the Knowledge Hub Aichi Priority Research Project IV (2022-2024) of the Aichi Prefectural Government and Chubu Electric Power Co, Inc.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The Department of Rehabilitation Medicine, School of Medicine, Fujita Health University, received research funding from Chubu Electric Power Co, Inc. Authors from this department include YN, HK, KT, and YO. The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript. All other authors declare no other conflicts of interest.

Multimedia Appendix 1
Type III ANOVA table for the linear mixed model.
[DOCX File, 22 KB - formative_v9i1e71265_app1.docx]

Multimedia Appendix 2
Estimated values of fixed effects in a linear mixed model.

XSL•FO
RenderX

[DOCX File, 24 KB - formative_v9i1e71265_app2.docx]

Checklist 1
STROBE checklist.
[DOCX File, 41 KB - formative_v9i1e71265_app3.docx]

References

1. Mild cognitive impairment (MCI). Alzheimer's Association. URL: https://www.alz.org/alzheimers-dementia/what-is-dementia/related_conditions/mild-cognitive-impairment [accessed 2024-11-19]
2. Canevelli M, Grande G, Lacorte E, et al. Spontaneous reversion of mild cognitive impairment to normal cognition: a systematic review of literature and meta-analysis. J Am Med Dir Assoc 2016 Oct 1;17(10):943-948. [doi: 10.1016/j.jamda.2016.06.020] [Medline: 27502450]
3. Sabbagh MN, Boada M, Borson S, et al. Rationale for early diagnosis of mild cognitive impairment (MCI) supported by emerging digital technologies. J Prev Alzheimers Dis 2020;7(3):158-164. [doi: 10.14283/jpad.2020.19] [Medline: 32463068]
4. Rasmussen J, Langerman H. Alzheimer's disease - why we need early diagnosis. Degener Neurol Neuromuscul Dis 2019 Dec 24;9:123-130. [doi: 10.2147/DNND.S228939] [Medline: 31920420]
5. Rawtaer I, Mahendran R, Kua EH, et al. Early detection of mild cognitive impairment with in-home sensors to monitor behavior patterns in community-dwelling senior citizens in Singapore: cross-sectional feasibility study. J Med Internet Res 2020 May 5;22(5):e16854. [doi: 10.2196/16854] [Medline: 32369031]
6. Chen C, Ding S, Wang J. Digital health for aging populations. Nat Med 2023 Jul;29(7):1623-1630. [doi: 10.1038/s41591-023-02391-8] [Medline: 37464029]
7. Korenhof SA, Fang Y, Luo J, van der Cammen TJ, Raat H, van Grieken A. Monitoring the well-being of older people by energy usage patterns: systematic review of the literature and evidence synthesis. JMIR Aging 2023 Mar 31;6:e41187. [doi: 10.2196/41187] [Medline: 37000477]
8. Sovacool BK, Hook A, Sareen S, Geels FW. Global sustainability, innovation and governance dynamics of national smart electricity meter transitions. Glob Environ Change 2021 May;68:102272. [doi: 10.1016/j.gloenvcha.2021.102272]
9. Nakaoku Y, Ogata S, Murata S, et al. AI-assisted in-house power monitoring for the detection of cognitive impairment in older adults. Sensors (Basel) 2021 Sep 17;21(18):6249. [doi: 10.3390/s21186249] [Medline: 34577455]
10. Jekel K, Damian M, Wattmo C, et al. Mild cognitive impairment and deficits in instrumental activities of daily living: a systematic review. Alzheimers Res Ther 2015 Mar 18;7(1):17. [doi: 10.1186/s13195-015-0099-0] [Medline: 25815063]
11. Desai R, John A, Stott J, Charlesworth G. Living alone and risk of dementia: a systematic review and meta-analysis. Ageing Res Rev 2020 Sep;62:101122. [doi: 10.1016/j.arr.2020.101122] [Medline: 32659336]
12. Creating new added value by utilizing electricity data [Web page in Japanese]. Ministry of Economy, Trade and Industry. 2021. URL: https://www8.cao.go.jp/kisei-kaikaku/kisei/meeting/wg/econrev/210917/210917keizaikasseika_02.pdf [accessed 2024-11-05]
13. Fujiwara Y, Suzuki H, Yasunaga M, et al. Brief screening tool for mild cognitive impairment in older Japanese: validation of the Japanese version of the Montreal Cognitive Assessment. Geriatr Gerontol Int 2010 Jul;10(3):225-232. [doi: 10.1111/j.1447-0594.2010.00585.x] [Medline: 20141536]
14. Yamada M, Arai H. Predictive value of frailty scores for healthy life expectancy in community-dwelling older Japanese adults. J Am Med Dir Assoc 2015 Nov 1;16(11):1002.e7. [doi: 10.1016/j.jamda.2015.08.001] [Medline: 26385303]
15. Satake S, Senda K, Hong YJ, et al. Validity of the Kihon Checklist for assessing frailty status. Geriatr Gerontol Int 2016 Jun;16(6):709-715. [doi: 10.1111/ggi.12543] [Medline: 26171645]
16. Studenski S, Perera S, Wallace D, et al. Physical performance measures in the clinical setting. J Am Geriatr Soc 2003 Mar;51(3):314-322. [doi: 10.1046/j.1532-5415.2003.51104.x] [Medline: 12588574]
17. Chen LK, Woo J, Assantachai P, et al. Asian Working Group for Sarcopenia: 2019 consensus update on sarcopenia diagnosis and treatment. J Am Med Dir Assoc 2020 Mar;21(3):300-307.e2. [doi: 10.1016/j.jamda.2019.12.012] [Medline: 32033882]
18. Nasreddine ZS, Phillips NA, Bédirian V, et al. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. J Am Geriatr Soc 2005 Apr;53(4):695-699. [doi: 10.1111/j.1532-5415.2005.53221.x] [Medline: 15817019]
19. Obu (Aichi Prefecture) April 2022 (daily values) main elements. Japan Meteorological Agency. 2022 Apr. URL: https://www.data.jma.go.jp/stats/etrn/view/daily_a1.php?prec_no=51&block_no=1638&year=2022&month=4&day=&view=p1 [accessed 2024-11-05]
20. Li Y, Pizer WA, Wu L. Climate change and residential electricity consumption in the Yangtze River Delta, China. Proc Natl Acad Sci USA 2019 Jan 8;116(2):472-477. [doi: 10.1073/pnas.1804667115] [Medline: 30584107]
21. Yasukawa K, Ishihara Y, Hirayama F, Nakanishi M, Utsumi H, Koyama S. Electricity use is associated with residents' vital data and lifestyles: observational study using an IT health support system in Japan. Sci Rep 2020 Oct 13;10(1):17162. [doi: 10.1038/s41598-020-74359-4] [Medline: 33051566]

XSL•FO
RenderX

22. Aman Y, Pitcher T, Simeoli R, Ballard C, Malcangio M. Reduced thermal sensitivity and increased opioidergic tone in the TASTPM mouse model of Alzheimer's disease. Pain 2016 Oct;157(10):2285-2296. [doi: 10.1097/j.pain.0000000000000644] [Medline: 27306045]
23. Monroe TB, Gibson SJ, Bruehl SP, et al. Contact heat sensitivity and reports of unpleasantness in communicative people with mild to moderate cognitive impairment in Alzheimer's disease: a cross-sectional study. BMC Med 2016 May 10;14:74. [doi: 10.1186/s12916-016-0619-1] [Medline: 27164846]
24. Schmeltz MT, Gamble JL. Risk characterization of hospitalizations for mental illness and/or behavioral disorders with concurrent heat-related illness. PLOS ONE 2017 Oct 16;12(10):e0186509. [doi: 10.1371/journal.pone.0186509] [Medline: 29036206]
25. Gong J, Part C, Hajat S. Current and future burdens of heat-related dementia hospital admissions in England. Environ Int 2022 Jan 15;159:107027. [doi: 10.1016/j.envint.2021.107027] [Medline: 34890899]
26. Kim KR, Lee KS, Cheong HK, Eom JS, Oh BH, Hong CH. Characteristic profiles of instrumental activities of daily living in different subtypes of mild cognitive impairment. Dement Geriatr Cogn Disord 2009;27(3):278-285. [doi: 10.1159/000204765] [Medline: 19246913]
27. Luo M, Wang Z, Zhang H, et al. High-density thermal sensitivity maps of the human body. Build Environ 2020 Jan;167:106435. [doi: 10.1016/j.buildenv.2019.106435]
28. Baeta-Corral R, Defrin R, Pick CG, Giménez-Llort L. Tail-flick test response in 3×Tg-AD mice at early and advanced stages of disease. Neurosci Lett 2015 Jul 23;600:158-163. [doi: 10.1016/j.neulet.2015.06.007] [Medline: 26091881]
29. Jensen-Dahm C, Werner MU, Jensen TS, et al. Discrepancy between stimulus response and tolerance of pain in Alzheimer disease. Neurology 2015 Apr 14;84(15):1575-1581. [doi: 10.1212/WNL.0000000000001465] [Medline: 25788560]
30. Yahiro T, Kataoka N, Nakamura K. Two ascending thermosensory pathways from the lateral parabrachial nucleus that mediate behavioral and autonomous thermoregulation. J Neurosci 2023 Jul 12;43(28):5221-5240. [doi: 10.1523/JNEUROSCI.0643-23.2023] [Medline: 37339876]
31. Earnings by educational attainment. Organisation for Economic Co-operation and Development. URL: https://www.oecd.org/en/topics/earnings-by-educational-attainment.html [accessed 2025-07-12]
32. Turley K, Rafferty J, Bond R, Mulvenna M, Ryan A, Crawford L. Evaluating the impact of a daylight-simulating luminaire on mood, agitation, rest-activity patterns, and social well-being parameters in a care home for people with dementia: cohort study. JMIR Mhealth Uhealth 2024 Nov 29;12:e56951. [doi: 10.2196/56951] [Medline: 39611803]

Abbreviations

AD: Alzheimer disease
ADL: activities of daily living
CeA: central amygdaloid nucleus
EC: electricity consumption
HEMS: home energy management system
IADL: instrumental activities of daily living
LPB: lateral parabrachial nucleus
MCI: mild cognitive impairment
MnPO: median preoptic nucleus
MoCA: Montreal Cognitive Assessment
MoCA-J: Japanese version of the Montreal Cognitive Assessment
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by A Mavragani; submitted 20.01.25; peer-reviewed by B Senst, T Nakagawa; revised version received 18.08.25; accepted 19.08.25; published 16.10.25.

Please cite as:
Nakagawa Y, Tanabe S, Kondo H, Tan K, Koyama S, Kitamura S, Kada A, Ishihara T, Yamamoto T, Denda J, Kimata H, Yamanaka T, Umezawa R, Nakahashi Y, Otaka Y
Daily Household Electricity Consumption in Community-Dwelling Older Individuals With Cognitive Impairment: Prospective Cohort Study
JMIR Form Res 2025;9:e71265
URL: https://formative.jmir.org/2025/1/e71265
doi:10.2196/71265

XSL•FO
RenderX

© Yuki Nakagawa, Shigeo Tanabe, Hikaru Kondo, Koki Tan, Soichiro Koyama, Shin Kitamura, Akiko Kada, Takuma Ishihara, Takuaki Yamamoto, Junya Denda, Hideaki Kimata, Taisuke Yamanaka, Ryosuke Umezawa, Yoshinobu Nakahashi, Yohei Otaka. Originally published in JMIR Formative Research (https://formative.jmir.org), 16.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Measuring FVIII Activity and Thrombin Generation Simultaneously With a Novel Point of Care Platform (EnzySystem HemA): Qualitative Usability Evaluation

Aernoud Bavinck[1,2], MD; Saskia E M Schols[1,3], MD, PhD; Wilfred Teunissen[4], MSc; Amy Shapiro[5], MD; Nicole M A Blijlevens[1], MD, PhD; Kyle Davis[5], MD; Waander van Heerde[2,3], PhD

[1]Department of Hematology, Radboud University Medical Center, Nijmegen, The Netherlands

[2]Enzyre, Nijmegen, The Netherlands

[3]Hemophilia Treatment Centre Nijmegen-Eindhoven-Maastricht, Geert Grooteplein Zuid 10, Nijmegen, The Netherlands

[4]Panton, Deventer, The Netherlands

[5]Department of Pediatric Hematology, Innovative Hematology, Indiana Hemophilia and Thrombosis Center, Indianapolis, IN, United States

Corresponding Author:
Saskia E M Schols, MD, PhD
Department of Hematology, Radboud University Medical Center, Nijmegen, The Netherlands

Abstract

Background: Hemophilia A is an inherited bleeding disorder with an increased risk of excessive bleeding. People affected by hemophilia A with a severe bleeding phenotype are treated prophylactically with hemostatic agents. Monitoring treatment and adjusting it to the individual patient's needs is complicated by the scarcity of laboratories equipped to perform relevant coagulation assays and the absence of point-of-care testing platforms. A novel near-patient testing platform called the EnzySystem HemA aims to address this issue by measuring factor VIII (FVIII) activity and thrombin generation within 1 hour of 100 µL of whole blood.

Objective: This study assessed the usability of the EnzySystem HemA by health care professionals without prior training and gathered information on its perceived usability, effectiveness, usefulness, and acceptability.

Methods: A qualitative, single-center formative usability assessment was performed. Participants performed an FVIII activity assay with a mockup of the EnzySystem HemA and were interviewed about its acceptability, usability, effectiveness, and usefulness. Video recordings of the sessions were obtained with permission and subsequently reviewed to assess usability; thematic analysis was performed on the interview transcripts.

Results: A total of 7 health care providers participated, all unaffiliated with the EnzySystem's developer. Participants included 1 (14%) pediatric hematologist, 3 (43%; pediatric) nurses, and 3 (43%; pediatric) nurse practitioners. Two (29%) of the participants also held management positions. Five (71%) participants successfully performed the FVIII activity assay as intended. Of the remaining 2 participants, one applied insufficient force when inserting the blood tube, whereas the other was unable to transfer the blood into the EnzySystem. This latter step caused difficulties in 3 (60%) other participants; however, they completed it correctly. Five (71%) participants did not dispose of the EnzySystem components as intended, primarily due to compatibility issues of the disposable parts with the inlet of the US standard medical waste containers. Five themes were generated from the interviews: *functional the way it is* focuses on the acceptance of the prototype despite suggestions for various improvements, *utility through faster results and increased access* and *potential for patient engagement* relate to the envisioned benefits of the platform, and *financial investment and return* and *ensuring accuracy* reflect on potential barriers for implementation.

Conclusions: Despite significant usability issues, the need for faster and more accessible testing led participants to conclude that the current platform was acceptable for use, provided that proof of assay accuracy was demonstrated. The results of this study will form the basis for further development of the EnzySystem HemA. The effectiveness of any updates will be evaluated in future usability studies, which will also include people with hemophilia A as participants.

Trial Registration: ClinicalTrial.gov NCT06369740; https://clinicaltrials.gov/study/NCT06369740

(JMIR Form Res 2025;9:e77621) doi:10.2196/77621

KEYWORDS

usability testing; point of care; thrombin generation; factor VIII activity; FVIII activity; user-centered design

XSL•FO
RenderX

Introduction

Hemophilia is a hereditary bleeding disorder, characterized by reduced activity of coagulation factor VIII (FVIII) [1-3]. To prevent and treat bleeding, treatment with FVIII concentrates or other prohemostatic agents is tailored to the patient's bleeding phenotype and FVIII activity levels [4,5]. Solely relying on FVIII activity levels fails to accurately predict bleeding. The thrombin generation assay (TGA), which evaluates overall hemostatic capacity, may better predict bleeding risk [6-8] and may solve monitoring issues with novel treatments [9,10]. However, both FVIII activity and TGAs are limited to specialized coagulation laboratories.

Point-of-care testing (POCT), "testing at or near the site of patient care" [11], is increasingly used to enable accessible testing in various medical fields [12-16]. Despite advancements enabling miniaturized coagulation testing [17], POCT platforms for hemophilia A do not exist. The EnzySystem HemA, a novel POCT platform for FVIII activity levels and thrombin generation, may address this gap.

Historically, POCT implementation has been challenging because of limited alignment between developers and end users [18,19]. To address this issue, end users are involved early in the development process according to the principles of user-centered design [20], reducing the risk of late-stage user interface modifications. These modifications would necessitate substantial revalidation, particularly in advanced technologies such as microfluidic assays.

The development and technical validation of the EnzySystem HemA technology were recently published [21]. The version evaluated in this study is currently undergoing technical and usability optimization. This study aimed to identify which aspects of the platform's user interface require modification to ensure safe and efficient use by health care providers (HCPs) without prior training. Specifically, it assessed usability, effectiveness, usefulness, acceptability, and use-related risks from the perspective of end users.

Methods

Study Design

This single-center, qualitative usability evaluation consisted of a usability scenario simulation and an in-depth interview. HCPs specialized in hemophilia care from the Indiana Hemophilia & Thrombosis Center (Indianapolis, IN) were invited via email. Nurse practitioners, (pediatric) hematologists, physician assistants, pharmacists, and nurses were considered for inclusion. Exclusion criteria included lack of hematological experience, inability to communicate in English, or unwillingness to provide informed consent. The study aimed to include approximately 6 participants with diverse professional backgrounds.

Ethical Considerations

The study adhered to the Declaration of Helsinki. The study protocol was approved by the Ascension St Vincent Hospital Institutional Review Board (study ID RIN20240003) and registered on ClinicalTrials.gov (NCT06369740). All participants were provided written information outlining the study procedures, including video recordings, and provided written informed consent. Video recordings were accessible only to authorized personnel, including members of the study team, auditors, and quality assurance and quality control reviewers. The recordings will be securely stored for 15 years. Participants provided explicit consent for their research data—including video recordings—to be accessed by the study sponsor, who is also the developer of the EnzySystem. Participants were provided with US $50 compensation for their participation in this study when participation occurred outside their regular working hours.

The results are reported in accordance with the COREQ (Consolidated Criteria for Reporting Qualitative Studies; Checklist 1) [22].

The EnzySystem HemA

The EnzySystem is being developed by Enzyre B.V. as a novel platform for near-patient multiparameter testing. The EnzySystem HemA is designed to simultaneously measure FVIII activity levels and thrombin generation from 100 µL of whole blood within 1 hour. It consists of three integrated modules: (1) a disease-specific disposable EnzyCard (Figure 1A); (2) platform-specific software, the EnzyApp, that guides users through user tasks and displays results on a standard computer; and (3) the EnzyPad, a reusable processor into which the EnzyCard is inserted (Figure 1B).

The EnzyCard contains several elements: the external housing called the envelope (Figure 1A1) and an injection-molded microfluidic cartridge (Figure 1A2). Users only handle the envelope, which features a connector compatible with a standard 2.7 mL vacutainer for citrated blood. Upon insertion of a blood tube, blood is transferred into the EnzyCard by pressing a built-in syringe with an accompanying plastic plug. Inside the EnzyCard, plasma is separated from other blood components by a plasma separation membrane and enters the microfluidic cartridge. Capillary force drives plasma through the microfluidic channels, mixing it with buffer and dried-in reagents. Ultimately, the sample enters sixteen 110 nL detection chambers, where a chemiluminescent reaction occurs. Photons are generated proportionally to the analyte of interest and are quantified in real time by a sensor printed circuit board containing 16 single-photon avalanche diode optical sensors positioned above the detection chambers. The data are processed and sent via a USB cable to a computer for immediate display and interpretation.

In this study, a nonfunctional mockup of the EnzySystem HemA was used, consisting of the following components:

- A nonfunctional prototype of the EnzyPad (Figure 1B)
- An EnzyCard (Figure 1A) prepackaged in a sterilization bag (Elinic, Spain)
- A 3D-printed plastic plug (Figure 1C) packaged with the EnzyCard
- An EnzyApp graphical user interface represented by a PowerPoint presentation on a standard laptop, using a Wizard of Oz methodology [23] (Multimedia Appendix 1)

XSL•FO
RenderX

- A printed quick reference instructions (QRI) sheet (Multimedia Appendix 2)

Figure 1. Overview of the EnzySystem components. (A) EnzyCard with (A1) the envelope with blood tube connector, incorporated syringe (*), and (A2) microfluidic cartridge with plasma separation membrane (†), blister filled with buffer (¥), and a magnified photograph of 8 of the detection chambers with spotted reagents. (B) EnzyPad processor. (C) Plastic plug used to depress the syringe in the EnzyCard envelope.

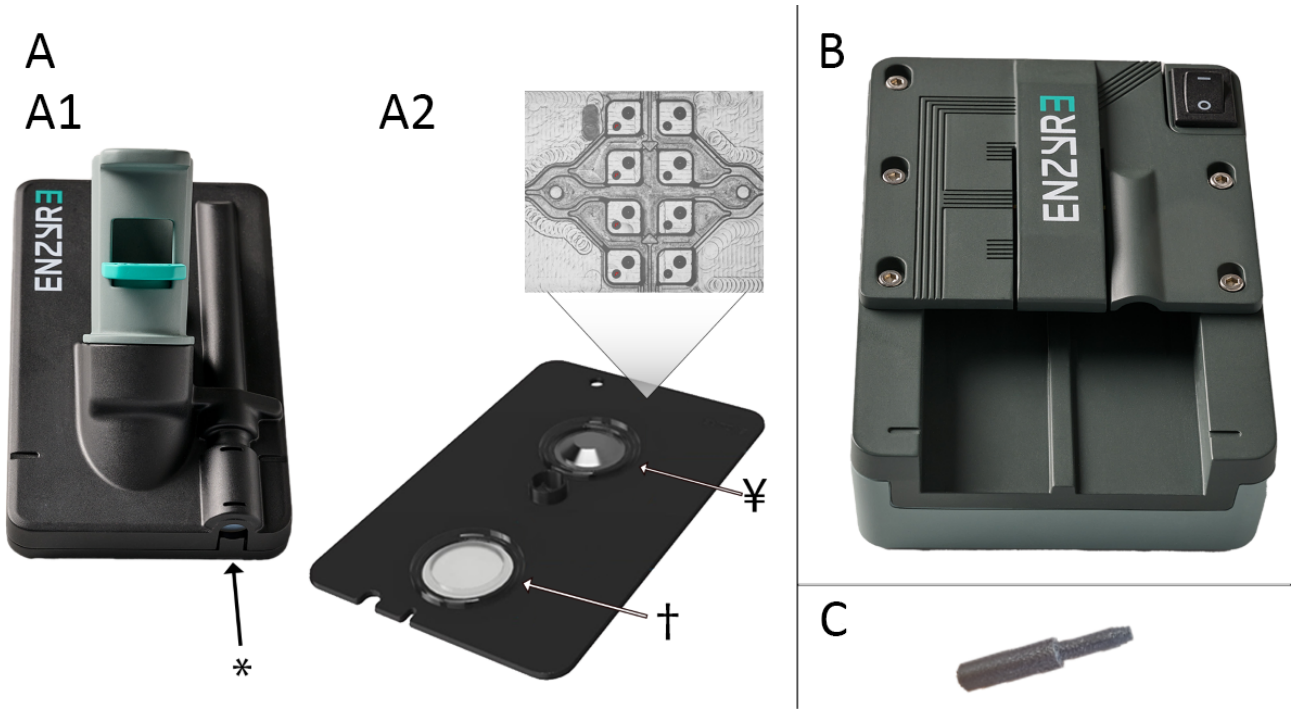

A nonfunctional mockup was used to enable efficient usability evaluation while avoiding known nonusability errors associated with the current prototype. Early usability evaluation with nonfunctional mockups is a well-established approach, supported by the Food and Drug Administration's human factors guidance [24]. Blood collection was excluded from the scenario, as this is a standard clinical procedure and the focus of the study was on platform-specific tasks.

Two included participants were deliberately tasked to perform the usability scenario without using the QRI to evaluate the intuitiveness of the EnzySystem HemA. They were purposefully selected based on their prior exposure to the platform in a focus group session on its potential use.

In addition to the EnzySystem HemA components, a standard 2.7-mL citrated vacutainer filled with water mixed with red dye, disposable nitrile hand gloves, and a standard medical waste container were used.

Study Procedures

All included participants attended an individual usability testing session on August 22 or 23, 2024, in a meeting room at the Indiana Hemophilia & Thrombosis Center. Each session lasted 1 hour and was moderated by 1 member of the research team (WT), assisted by the first author (ABA).

Participants were seated at a conference table with the testing materials placed before them. Baseline characteristics were gathered with a structured interview. The moderator outlined the study goals and the components available for use during the scenario, following a predetermined script (Multimedia Appendix 3). Participants were then instructed to perform an FVIII activity assay and to clean up afterward. Participants assigned to not use the QRI were instructed to open the EnzyApp as a first step. Participants were encouraged to talk out loud [25] and were informed that the waiting times during preheating (normally 3 min) and running the assay (normally up to 1 hour) were shortened. The actual waiting times were disclosed at the respective steps.

Following the scenario, a semistructured interview using a preconstructed interview guide (Multimedia Appendix 3) was conducted to assess the perceived acceptability, usability, effectiveness, and usefulness of the EnzySystem HemA.

Development of the Scenario and Interview Guide

The scenario and interview guides were constructed by the usability engineer in the research team (WT) based on medical input from the first author (AB). Both were refined based on 2 pilot sessions performed at the Radboudumc (Nijmegen, the Netherlands) with a nurse practitioner and a physician, both working in the field of hemophilia.

Data Acquisition and Analysis

Usability Session

The usability sessions were recorded with 2 cameras. The video recordings along with field notes represented the main data sources. Task performance was documented for each participant in a Microsoft Excel framework. An experienced usability engineer (WT) assessed each task as either:

- As intended: Completed correctly without observable difficulty
- As intended with difficulty: Completed correctly without support; however, repeated attempts were required, the participant demonstrated clear uncertainty about how to

XSL•FO
RenderX

perform the task or whether it was completed correctly, the task caused noticeable inconvenience, or, at the researchers' discretion, completing the task required substantial effort
- Not as intended: Not completed, completed incorrectly, or required assistance

The overall assay was performed *as intended* if none of the tasks were assessed as *not as intended*. Tasks relating to the disposal of the EnzyCard were evaluated separately.

Semistructured Interview

The interview data were analyzed with Braun and Clarke's [26] thematic analysis. The interviews were transcribed verbatim and coded inductively in Atlas TI version 23.4 by the first author (AB). Analytic field notes were recorded throughout the process. A coding book was kept during the data analysis (Multimedia Appendix 4). Interim findings and initial themes were discussed in research meetings with 3 of the investigators (AB, SEMS, and WVH). Codes concerning improvements for the EnzySystem HemA were counted to identify the most prevalent suggestions. Participants were not involved in checking the transcription and the final coding tree.

Results

Participants

A total of 110 HCPs were invited to participate. Twelve (11%) responded, of whom 5 (42%) were excluded: 1 was unable to attend physically, and 4 had professions that did not meet the inclusion criteria. Seven (58%) HCP participants were included in the study. Baseline characteristics are depicted in Table 1. None of the participants had used the platform before. However, 3 (43%) participated in a prior EnzySystem study involving focus group discussions on current coagulation testing and potential use cases.

Table . Baseline characteristics of participants.

	Age (years)	Gender	Profession	Participated in previous EnzySystem focus group discussions	Use of QRI[a]
HCP1[b]	39	Male	Pediatric nurse	No	Yes
HCP2	59	Female	Nurse practitioner, management role	No	Yes
HCP3	58	Female	Pediatric nurse practitioner	No	Yes
HCP4	38	Male	Nurse	Yes	Yes
HCP5	43	Female	Nurse practitioner	Yes	No
HCP6	34	Female	Pediatric hematologist	Yes	No
HCP7	61	Female	Nurse, management role	No	Yes

[a]QRI: quick reference instruction.

[b]HCP: health care provider.

Usability Scenario

An overview of the success rate per task is depicted in Figure 2. Of the 7 participants, 5 (71%) performed the assay as intended, including HCP6 who did not use the QRI. The mean time to obtain a test result was 359 (SD 148) seconds, including two 20-second waiting periods for warming up and running the assay. Two (29%) HCPs were unable to perform the test as intended. One (14%) applied insufficient force when inserting the blood tube, expressing uncertainty on how much force could be applied before damaging the tube. The other, who performed the test without using the QRI, needed support to turn on the EnzyPad and insert the blood tube. This participant did not use the plastic plug to depress the syringe and transfer the blood sample into the microfluidic cartridge. Three (43%) other participants had difficulties completing this last task, hesitating over where to insert the plastic plug.

XSL•FO
RenderX

Figure 2. Correct task performance per user. QRI: quick reference instruction.

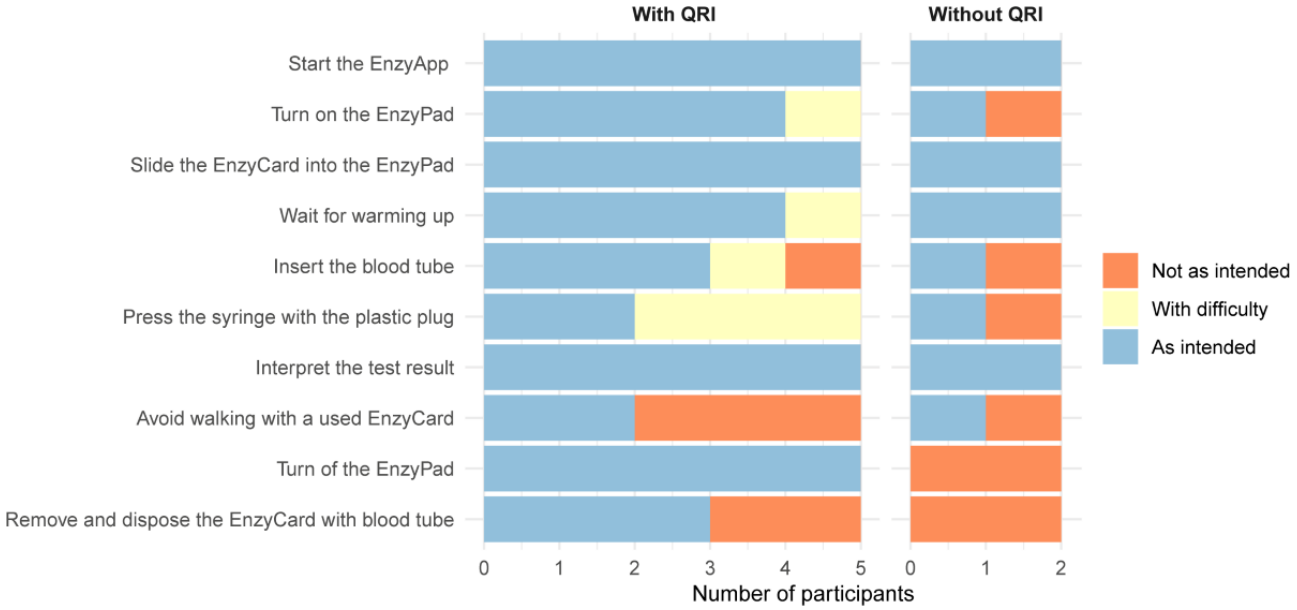

Only 2 of the 7 (29%) participants disposed of the EnzyCard as intended. Contrary to the instructions, 4 (57%) participants moved around with the used EnzyCard to discard it, and 4 (57%) removed the blood tube from the EnzyCard before disposal. The combined size of the EnzyCard and the blood tube constituted the main reason for difficulties with disposal, as together they did not fit into a standard US medical waste container.

Interview on User Experiences, Perceived Acceptability, and Usefulness

Five interrelated themes were generated from the postscenario interviews: *functional the way it is*, *utility through faster results and increased access*, *potential for patient engagement*, *financial investment and return*, and *ensuring accuracy* (Table 2).

XSL•FO
RenderX

Table . Illustrative quotes for generated themes.

Generated themes	Illustrative quotes
Functional the way it is	
1.1	"I think a lot of the feedback is more to make it nicer, but I do think it's functional the way it is." (HCP6[a])
1.2	"For clinic, I think it's, would be fairly easy for a nurse to go through the whole steps, collecting the specimen, and prepping." (HCP3)
1.3	"It's smooth. The platform kind of lends itself to not running into user error, [...] it's just a consistent chain. Even the program itself is pretty simple and straightforward. Everything goes together in one way." (HCP1)
1.4	"This is, you know, this felt kind of how I expected it to. Feels like when you're, when you're collecting, it's the same, same feeling popping the tube onto a needle, like when you're collecting blood. So it's a familiar feeling, to have that, and it's a familiar shape too, it's a vacutainer type of shape." (HCP4)
1.5	"That [preference for automatic EMR interface] being said, exporting to a PDF would be fine. It would be a little bit more of a process, but exporting to a PDF and then printing that, and then you can scan it in into the printer work just the same." (HCP1)
Utility through faster results and increased access	
2.1	"You don't have to do any further processing with the blood tube at all, or sending it someplace, waiting on the courier to pick it up." (HCP7)
2.2	"...after you do that [drop of the sample at the laboratory], you're watching for results. You don't know when they might populate, when they might not. So [with the EnzySystem] you would have the results readily available here to you, so you would not have to be sitting there watching every 10 to 15 minutes, clicking refresh to see if a result is coming through." (HCP4)
2.3	"When this is sort of presented as, like, a point of care machine, 60 minutes is a very, very long time" (HCP1)
2.4	"I mean, you look at what you're replacing. Like, if you told me you were running my blood sugar in 60 minutes, that would not be acceptable. But this is something that you're replacing a test that's trickier to get back on a timely manner." (HCP2)
2.5	"Well, in the OR you need...would need like an instant result. Otherwise you have a patient that is bleeding." (HCP5)
2.6	"And I can't think of a scenario where, like, you know, a 60 minute will be like, [a] life or death decision." (HCP6)
2.7	"If there is a bleed and I plan to give multiple factor infusions, it would be nice to know, like if the patient got replacement on one day...like got factor infusion on Monday and then the next day, you're deciding whether to or not...whether or not to give a second dose, to give a minor or major dose, then it would be nice. We usually do that just going based on clinic and symptoms, but this would be helpful information." (HCP6)
2.8	"Having this [EnzySystem] readily available for a pre-op patient would be great. Because I can check their pre level. I can run that while I am infusing. Collecting the post sample and by the time I'm done with everything this would be nearing completion on the first. And then I can run that. So by the time they are over at the hospital and I'm getting set up, everything's good to go. We know what the result is." (HCP1)
2.9	"Having this [the EnzySystem] available in the clinic, where we can get a result faster than what we can from the lab will help direct patient care faster, because sometimes they have to stay here until we get a result. If they can't self infuse, then they have to wait for that result to decide or not..." (HCP5)

XSL•FO
RenderX

Generated themes	Illustrative quotes
2.10	“And literally like to Kenya, we had all the people that we thought had hemophilia and we ran factor VIII tests first and then the people that didn’t were FVIII deficient, we ran factory IX tests. So it was a way, to kind of quickly do that. So I think if it was for...we were interested in a single test, [...] I don’t see really a downside to having it.” (HCP2)
Potential for patient engagement	
3.1	“My immediate thought with this is more the ability to use this at home. I think this would be a great tool for, for patients to have at home. So not all, but the majority of our patients, already have the ability to self infuse. So they stick themselves. They would be able to get a sample fairly easily. This is very user friendly. And again, with it being as small as it is, it’s not taking up a bunch of space. This would be great to have at the house if you are having breakthrough bleeding or something like that and if there is a concern that your prophylactic treatment is no longer working as it should, this would be a great way to provide some peace of mind and confirm whether or not levels are where they should be.” (HCP1)
3.2	“At home would be wonderful because that I, I know the, the families that I have with the CoaguChek monitors like it’s, it’s convenient because if they need to travel or anything like that, they can take the machine with them, and we get instant results, and I can tell them what to do with their medication instantly.” (HCP5)
3.3	“I think a child could [use the EnzySystem]...If you show how to do it. [...] So I think it helps them feel like they’re a part of their own care.” (HCP5)
3.4	“Some patients are not as reliable with regard to actually communicating information back, when they should. I guess you could consider only doing that [self testing] with patients that you would say, are more reliable, adherent to their treatment and that you know, are really going to be diligent about following whatever testing and ranges and things that you’ve set up.” (HCP7)
3.5	“If it’s [the blood tube] glass, breaking glass and causing potential injury to a layperson, and especially if you have kids are getting the test done, then proper disposal is going to be necessary.” (HCP3)
3.6	“It could be used outpatient home visits situation with our outreach nurses that actually go to the patient’s home and could do a test.” (HCP7)
3.7	“They would be in front of us and we would have that educational ability. And really, they would leave with a better understanding versus, you know, you get the results next day, then someone has to call them and talk to them. And so I think there’s a lot of information that the urgency, or that that immediate knowledge, when you’re in front of a provider would be, I mean, it would be very powerful.” (HCP2)
Financial investment and return	
4.1	“But even like blood sugar strips, you know, for diabetics are expensive. So I would anticipate this along with computer technology would not be inexpensive to use.” (HCP3)
4.2	“And that is the biggest struggle, is whether or not the insurance company will cover the device. I have some patients that still have to go to a lab because the insurance company, the insurance they carry won’t cover the device. So that’s the biggest burden.” (HCP5)
4.3	“Well, I think the hospital probably would not appreciate us bringing the device in to test our patients. Because that means they’re not getting a test, and they’re not getting the revenue for that test.” (HCP7)
4.4	“I think from the standpoint of resources, as far as thinking of factor as a resource, it’s very expensive. And if I can adjust it even a little bit and then get the same outcomes, I think that’s worth it.” (HCP2)
Ensuring accuracy	
5.1	“Yeah, I guess it [downsides of the EnzySystem] is just...it’s really determined because how accurate the point of care is.” (HCP3)

XSL•FO
RenderX

Generated themes	Illustrative quotes
5.2	"I, I think, seeing the research study information, I think would be important for ensuring that the test information from the point of care and the conventional way would be equivalent in the result." (HCP7)
5.3	"I would probably for the first time I may do this with a lab confirmed that, like they're compatible, and then if that's the case, then I will just use this device." (HCP6)
5.4	"So if we have a known concentration, we run that through. All as well in the world and great." (HCP1)
5.5	"Would there be something that you could put in here that would be like a little control, so I could see, like, their level, and then I'd see the control curve at the same time?" (HCP2)

[a]HCP: health care provider.

Theme 1: Functional the Way It Is

Although all participants suggested improvements (Table 3), the platform was considered usable in the current iteration (Quote 1.1, Table 2). The primary areas for improvement involved disposability, revising specific terminology in the QRI and graphical user interface, automating blood transfer into the EnzyCard, introducing feedback to verify task execution, and enabling automatic results export to the electronic medical record (EMR). Beyond these suggested improvements, the handling was considered sufficiently straightforward to be used in a clinic by HCPs or by patients (Quote 1.2, Table 2).

Participants attributed the experienced simplicity to the components fitting together in one way, making it intuitive and reducing error risk (Quote 1.3, Table 2). In addition, inserting the blood tube was perceived as analogous to using a regular connector during venipuncture (Quote 1.4, Table 2), which enhanced some participants' confidence in performing the task correctly.

In agreement with the participants' overall acceptance, most suggested improvements were framed as optimization rather than obligatory updates. The comments on potential connectivity were exemplary. All participants strongly preferred automatic EMR integration, but most would accept manual data import (Quote 1.5, Table 2). Notably, only 1 HCP deemed automatic export mandatory, as manual export would increase the risk of user error.

The participants highlighted the EnzySystem's advantages over current testing as a reason for its acceptance.

XSL•FO
RenderX

Table . Number of participants suggesting specific improvements and requirements for the EnzySystem HemA.

Area for improvement	Count, n	Illustrative quote (participant)
Improve ease of handling		
Improve disposability	5	"If this is not reusable, it just seems like it's going to be cumbersome for a lot of this disposal. So I don't know what is in there, but, I mean, it's big, if we do every test and we're throwing this away, that's a lot of pieces to be thrown out." (HCP3[a])
Streamline pressing the syringe[b]	4	"This [Plastic plug] didn't feel like it fit very well. So it seemed clunky. But other than that, no. It would be nice to have it kind of already in there, and then you just engage it because, like, what if I dropped this or lost it or whatever?" (HCP2)
Refine blood tube handling	2	"Taking out [was difficult] ... More out than in. But a lot of it... I wasn't sure how much to keep pushing." (HCP6)
Incorporate automatic continuation of the GUI[c]	2	"Like I have used point-of-care testing, [...] so normally, once you put the cartridge in, it automatically runs itself. So to know the click ready on the screen, I think I wouldn't want to do that..." (HCP5)
Improve cleanability	1	"Stainless or, you know, or something that that doesn't have a lot of, like, little cracks and that kind of thing. These little lines, I know are... make it look nice, but, you have to clean it." (HCP2)
Make sliding in the EnzyCard smoother	1	"I wish that the joining of the two, it felt just a little bit smoother." (HCP1)
Prevent use errors		
Enhance understandability and consistency of instructions	7	"It was a little bit weird understanding that this is... this was a EasyPad [sic]. [...] It doesn't say easy pad on it. And calling this an EasyCard seems a strange name to me because, it is not really cardlike." (HCP7)
Add feedback mechanisms	5	"Just a little positive feedback once you finished that [pressing the syringe], such as a click, once you finish pressing in the syringe." (HCP4)
Add handling instructions	4	"It felt firm going in [with plastic plug], so it'd be nice that... to make note that you will feel that firmness in there but continue on. Because I felt hesitant..." (HCP3)
Automate the blood volume transferred into the EnzyCard	2	"And also control that like, like blood that goes into the thing. [...] Because if I didn't push that all the way in, would that still be 100 microliters." (HCP6)
Label components	2	"What would be nice, because I know it's prototype, but add more color to that to be easy to identify where the pieces go." (HCP3)
Add a trigger release for sliding out the EnzyCard	1	"Given that you are handling a bodily fluid and maybe nice if it's more of a trigger release. So if you had that locking mechanism where it latched and something that you can press release that, so that you're not applying some arbitrary, force." (HCP1)

XSL•FO
RenderX

Area for improvement	Count, n	Illustrative quote (participant)
Invert the direction of the blood tube connector	1	"So you insert the card this way, right. You're applying an opposing force with the tube in the opposite direction. If it were possible, then might be nice if this were switched to the other direction." (HCP1)
Prevent blood tube removal	1	"...if you wanted to remove the element of potential splash, you could put something inside of this tube during the manufacturing process, so it's a one way." (HCP1)
Protect from damage		
Enhance resistance to damage	2	"...the risk for it being dropped and broken and then repair. So for you making sure it will sustain falling and children." (HCP3)
Prevent losing the plastic plug	2	"You could open this package up and potentially lose that little white thing and then you don't have it." (HCP7)
Enhance results interpretation		
Show past results	3	"What would be nice if this is this has been going on with the patient to see a trend and the numbers of past tests that we have." (HCP3)
Add axis information	2	"If the scales are added, that may be a little more benefit to a physician, than a nurse, where I'm just kind of looking at the numbers." (HCP4)
Add reference values	2	"I guess, if it was within range or. [.] Like, if it's within range, like, do we need to do anything if it's not within range, then an alert, a red alert.' (HCP5)
Add a countdown clock	1	"A countdown clock would be great." (HCP2)
Blood tube compatibility		
Prevent wasting unused sample volume	4	"In particular with babies, it would... might be hard to get that amount of blood that's required, for the 2.7 ml tubes." (HCP5)
Facilitate using other citrate concentrations	1	"If, for whatever reason, say, normal tubes were on backorder, we found a supplier that had a tube with a different concentration […]. Would the machine tolerate that?" (HCP1)
Data management and connectivity		
Enable automatic export to EMR[d,e]	6	"I mean it would need to interface into something, because I would hate to superimpose the numbers. And 45 is different than 54." (HCP2)
Add a wireless connection	3	"So, I see, I see that it's plugged in. It's obviously a wired device. I don't know if wireless is an option." (HCP2)
Couple results to the patient ID	3	"But no, actually, I would want this to be connected to a patient ID because, even if this is not going to connect to the EMR, at least if there is memory of, like this [is the] patient and these are the results." (HCP6)
Combine the results with treatment data	1	"...it would be nice if [...] patients could put in their last infusion and the amount [...], like WAPPS Hemo, the PK... […] If you could tie this to that somehow." (HCP2)
Other		

XSL•FO
RenderX

Area for improvement	Count, n	Illustrative quote (participant)
Add additional assays	3	"Probably would like factor nine [FIX] and VWD." (HCP6)

[a]HCP: health care provider.

[b]Also mentioned in preventing use errors; however, quotes in this category indicate that the usability could be more convenient rather than less prone to error.

[c]GUI: graphical user interface.

[d]EMR: electronic medical record.

[e]Some comments relate to preventing use errors; all comments about the automatic export of data were categorized under this code.

Theme 2: Utility Through Faster Results and Increased Access

Throughout the interviews, the primary reasons for the perceived utility of the EnzySystem were its potential to deliver faster and more accessible test results. The compatibility of the platform with whole blood combined with its near-patient design would enable streamlined and rapid testing (Quote 2.1, Table 2) and more predictable result availability (Quote 2.2, Table 2).

Interestingly, participants expected near-patient tests to be faster than the EnzySystem HemA (Quote 2.3, Table 2). However, their experiences with FVIII testing and understanding of its complexities increased acceptance of the required testing time (Quote 2.4, Table 2). Nonetheless, some participants considered shortening the testing time critical for use in a surgical setting and at the emergency department to guide treatment decisions (Quote 2.5, Table 2). However, others noted that the turnaround time would not limit clinical use in these situations because patients typically remain in a care facility for over an hour when receiving care (Quote 2.6, Table 2).

The most significant potential for the platform was envisioned in 3 types of situations. First, enhanced accessibility would enable personalization of treatment by performing testing in scenarios where it is currently not available (Quote 2.7, Table 2). This would be useful during at-home treatment for a bleed, before high-risk activities, to tailor prophylaxis, and after surgery. Second, the platform could streamline standard monitoring situations, such as presurgical assessments (Quote 2.8, Table 2) and regular clinic visits (Quote 2.9, Table 2). Finally, increased access could facilitate the diagnostic process in patients with suspected bleeding disorders, especially in low-resource countries (Quote 2.10, Table 2). However, some HCPs contended that the EnzySystem HemA might be unsuitable for diagnostic purposes, as additional laboratory tests beyond FVIII and TGA are often required. Consequently, blood samples would be sent to a laboratory regardless of using the EnzySystem.

Theme 3: Potential for Patient Engagement

Most participants anticipated opportunities for patient engagement through the EnzySystem HemA. First, home-based self-testing would allow for testing as needed and provide peace of mind for patients (Quote 3.1, Table 2) while eliminating the need for travel (Quote 3.2, Table 2). In addition, involving patients in testing was suggested to enhance their overall engagement with treatment (Quote 3.3, Table 2). However, self-testing might not be valid for all patients due to differences in patients' technical skills, availability of equipment, insurance coverage, and reliability (Quote 3.4, Table 2). The significance of these factors depends on the features of the final product. For example, automatic export to the EMR could reduce the criticality of patient reliability. While the EnzySystem HemA was considered easy to use, the need for clear instructions and operational safety was further emphasized for patient self-testing (Quote 3.5, Table 2).

Even if home testing would not be feasible for all patients, increased patient engagement could be achieved by equipping outreach nurses with the EnzySystem HemA (Quote 3.6, Table 2) and by reducing the time to result in the clinic (Quote 3.7, Table 2). Observing the test result emerge on a screen during appointments could enhance understanding of the result, particularly when reference ranges are indicated. In addition, having results available for discussion during appointments was considered advantageous compared with calling patients after they leave the clinic.

Theme 4: Financial Investment and Return

One factor that could influence the use of the EnzySystem HemA was the cost and insurance coverage of the platform. Based on experience with other POCT platforms and the utilization of computer technology, 1 participant expected the costs of testing to be high because of typically expensive disposables (Quote 4.1, Table 2). Two other participants also indicated that costs would influence the device's acceptability.

Arrangements regarding testing costs would impact implementation. Insurance companies have been reluctant to cover POCT costs, necessitating some patients to continue visiting laboratories for assays that could be performed using POCT platforms (Quote 4.2, Table 2). Furthermore, hospitals might prefer using their own conventional tests or acquiring an EnzySystem HemA themselves for in-house use, instead of using patient- or HTC-owned devices. This approach enables them to maintain testing revenues (Quote 4.3, Table 2).

While some participants raised questions about the platform and maintenance costs, others thought that increased access to testing would ultimately reduce overall health-related costs. Given the high costs of factor products used to treat hemophilia patients, reducing factor usage without increasing the risk of bleeding is highly desired. The 2 participants in management roles believed that adjusting treatment based on more frequent testing could prevent unnecessary administration of factor products and associated costs (Quote 4.4, Table 2).

XSL•FO
RenderX

Theme 5: Ensuring Accuracy

Across all interviews, the importance of ensuring accurate results was emphasized (Quote 5.1, Table 2). Participants indicated that they desired proof of accuracy before using the EnzySystem HemA (Quote 5.2, Table 2), as incorrect assay results could misdirect treatment or cause unnecessary concerns, such as fear of inhibitor development. Participants would study the results of clinical validation trials, which could offer sufficient proof of accuracy. However, some providers would still compare the EnzySystem Hem A's results with those obtained with conventional coagulation tests as independent accuracy confirmation (Quote 5.3, Table 2).

Providing control samples was suggested as an additional method for accuracy validation (Quote 5.4, Table 2). Several participants indicated that access to samples with known FVIII activity, which could be sent with a batch of EnzyCards, would be sufficient for periodic quality control. Similar quality control strategies are used for other currently available point-of-care devices. One HCP noted that ideally, an internal control would be performed automatically with each run to provide assurance of the run's validity (Quote 5.5, Table 2). This last suggestion would also provide confirmation of having correctly executed the assay. Several other participants suggested feedback mechanisms to ensure correct assay performance (Table 3). This consistent emphasis on validation highlights the necessity of ensuring result accuracy before implementing the EnzySystem HemA.

Discussion

General Results

In this first usability study of the EnzySystem HemA with HCPs, the overall usability and perceived utility of the platform were high. While the platform was considered usable, 2 participants were unable to use the EnzySystem as intended due to critical issues with the user interface. In addition, there is a uniform need to improve the disposability of the EnzyCard, as most participants did not dispose of it as intended. The platform's perceived benefits over existing systems mitigate the concerns about its shortcomings. However, evidence demonstrating the platform's reliability and accuracy is needed before implementation.

Comparison With Previous Findings

Studies assessing human factor aspects of point-of-care in vitro diagnostics are scarce [27]. To our knowledge, no previous studies have evaluated the usability of products comparable to the EnzySystem HemA. However, research on attitudes toward POCT in general provides relevant insights. Notably, concerns about accuracy are not unique to the EnzySystem HemA. In a systematic review of primary care providers' attitudes, uncertainty about accuracy was a key limitation of POCT [28]. Checking POCT results with conventional methods, as was suggested in this study, is also applicable to POCT in general [29]. Quinn et al [30] found that concerns about accuracy, data management, costs, and quality assurance were key barriers to POCT adoption in hospital settings, consistent with the findings of this study. However, regarding the EnzySystem HemA, 2 of the 7 (29%) participants suggested that implementation might reduce factor product use and related costs, aligning with broader discussions on the benefits of personalized medicine in hemophilia A [31-33].

The main envisioned benefits of the EnzySystem HemA also align with those described for POCT in general: shorter time to initiate treatment [29,34-37], improved communication with patients and patient engagement [29,34-37], and improved access to testing [35-37]. However, as no POCT is available for hemophilia A care at this time, the potential implications of these advantages, especially in this field, were described for the first time in this study.

Compared to most POCT platforms, a limitation of the current EnzySystem HemA is the need for a venipuncture, which limits the potential applicability for patient self-testing, as suggested in this study. An EnzySystem version compatible with a capillary blood collection system is planned for future development, but its usability for patients will require dedicated evaluation.

Methodological Reflection

An experiential, inductive thematic analysis informed by a constructionist epistemology was used. Rather than seeking reproducible or externally generalizable results, the aim of qualitative research is to gain an in-depth, contextual understanding of participants' experiences and the factors underlying the perceived acceptability and usability of the EnzySystem. This approach recognizes that both the participants' perspectives and the researchers' interpretations are influenced by subjective perspectives and that meaning is co-constructed throughout the research process.

To enhance the credibility of the analysis, quotes from participants were added to support the claims made. A codebook was maintained and is provided as supplementary material to communicate interpretative decisions and enhance methodological transparency (Multimedia Appendix 4). In addition, codes related to improvements of the EnzySystem Hem A were presented along with their frequency of mention (Table 3). While such counting techniques are uncommon in reflective thematic analysis—as frequency does not necessarily reflect the perceived importance of an issue [38]—they were included to highlight which improvements were most widely desired. Ultimately, however, the credibility of our findings rests on their ability to convincingly represent the participants' experiences with the EnzySystem HemA and its perceived utility.

An inductive coding strategy was chosen, as this was the first usability evaluation of the platform, and no comparable platforms are available. However, the themes generated in this study relate to those commonly found in usability frameworks. For example, facets of Morville's Usability Honeycomb [39], such as usability, usefulness, and credibility, relate strongly to themes 1, 2, and 5, respectively. As such, the generated themes do not exist in a theoretical vacuum but relate to the existing theoretical frameworks.

XSL•FO
RenderX

Reflexivity

In reflective thematic analysis, it is acknowledged that researchers inevitably bring their own norms, values, and interpretive frameworks into the processes of data collection, analysis, and reporting [40]. Reflexivity in this study focused on the first author (AB)—who performed the coding and theme generation and assisted during the use scenarios—and the human factor engineer who moderated the scenarios (WT).

AB is a White Dutch male, a medical resident in internal medicine, and a PhD candidate at the Radboud University Medical Center (Radboudumc) in Nijmegen. He is employed by the developer of the EnzySystem during his PhD trajectory but holds no financial interests in the platform's development. He received formal training in qualitative research, and his work focuses on personalized care in hemophilia A. He has no formal background in engineering or technological development.

WT is a White Dutch male senior product designer employed by Panton (Deventer, the Netherlands), a company specializing in medical innovation design. He has extensive experience in medical innovations, although this was his first project in the field of hemophilia A.

Both AB and WT had no prior relationship with any of the participants.

Study Limitations

Study Setup and Contextual Considerations

Several factors limit the transferability of this study's findings to real-world scenarios. First, the study was conducted in an office setting, which may differ from real-world clinical conditions. During use in practice, users must perform the assays while managing concurrent tasks, cope with environmental distractions typical of clinical settings (eg, maintaining communication with patients), and work under time pressure inherent to outpatient care. These factors may negatively impact user performance compared to the findings in this study. In addition to office use, several participants suggested that the EnzySystem HemA might be used in a laboratory environment, where the impact of unexamined environmental influences could be more pronounced.

In addition, a nonfunctional prototype of the EnzySystem HemA was used, and no real blood samples were collected. This approach was chosen to avoid invasive blood sampling and focus on usability aspects rather than technical validation, especially because the current prototype still has some known technical flaws. However, venipuncture is a critical step for performing the assay, and integrating it into the use scenario could uncover additional and significant usability issues.

Furthermore, as participants engaged in the use scenario only once, no information was obtained regarding learnability and memorability. As implementation of the EnzySystem HemA will involve repeated use by the same HCPs, these aspects should be explicitly addressed in future evaluations to provide a more comprehensive assessment of the platform's long-term usability.

Finally, the findings of this study are specific to the contextual factors of the Hemophilia Treatment Center involved. While generalization is not the primary aim of qualitative research, it is important to consider that results may differ in other settings due to variations in treatment access, testing availability, and the degree of care specialization.

Potential Sources of Bias

The convenience inclusion procedure in this study may have favored the inclusion of HCPs with comparatively greater openness to research participation and new technologies. In addition, several factors may have biased the perspectives of the participating HCPs. For practical reasons, the assay's waiting time was shortened. Although participants were informed of the actual waiting time, their experiences may have differed had the real waiting time been implemented.

Furthermore, participants were informed that this study was conducted by Enzyre, the company developing the EnzySystem HemA, and that a company-affiliated assistant was present during the study. While participants primarily interacted with the lead moderator and the assistant was formally trained in qualitative research—including methods to avoid suggestive or leading questions—some participants may still have adjusted their responses to be more socially desirable.

Regarding the interpretation of the results, it should be acknowledged that in qualitative studies, the influence of researchers on participants and interpretation cannot and should not be eliminated [40]. However, to mitigate the risk of conflation of company interests with the analysis, the primary analysis of the video recordings to assess task performance was performed by the unaffiliated scenario moderator. Furthermore, the full investigative team, including 5 members not affiliated with the developer, critically reviewed and approved the final analysis in the manuscript. Notably, the unaffiliated moderator had independently analyzed all use scenarios and interviews before reviewing the manuscript.

Sample Size

Finally, the sample size was relatively small with 7 participants, who were heterogeneous in terms of roles. The number of participants needed for usability studies is controversial. While some authors consider 5 participants sufficient due to the descriptive nature of these studies [41,42], more recent approaches advocate for larger sample sizes [43]. A reason for pursuing larger sample sizes is the nonhomogeneous probability of identifying specific use errors. Using the good-turning model modified by Hertzum and Jacobsen as outlined by Borsci et al [44], the adjusted estimated percentage of use errors discovered by an individual participant in our study was 23.8%. Following this, 85% of use errors are expected to be discovered with this study of 7 participants.

However, this estimation does not reflect saturation within the interview data. While the suitability of data saturation is controversial in the context of Braun and Clarke's [45] thematic analysis [46], it should be acknowledged that this study's design did not allow for ongoing assessment of saturation to guide further inclusion. Nevertheless, no unique codes were applied exclusively to the final interview transcript, suggesting that the

XSL•FO
RenderX

contribution of additional participants for the ongoing analysis was diminishing. More importantly, although including additional participants will always have the potential to reveal new insights or deepen existing ones, the study's findings provide actionable information to guide the further development of the EnzySystem, indicating data sufficiency.

Overall Conclusion and Future Directions

The current findings demonstrate high acceptance across HCPs for the EnzySystem HemA, despite some critical limitations of the user interface. To evaluate the effectiveness of potential modifications and ensure the platform's suitability for broader implementation, future usability evaluations should be conducted in realistic use environments and include participants from diverse care facilities. The universal demand among participants for faster results and improved testing to monitor treatment in people with hemophilia A underscores the urgency of these next steps, along with the need to assess the platform's usability for people with hemophilia A.

Acknowledgments

The authors are grateful to the following individuals for their assistance in this study: Sander Weernink contributed to the conceptualization of the study design; Dr Monique van Lier, Dr Tanja Hamacher, and Neelam Thukral assisted with obtaining institutional review board approval; and the research team of the Indiana Hemophilia & Thrombosis Center facilitated the execution of the study. Generative artificial intelligence (ChatGPT, OpenAI) was used solely for grammar, spelling, and readability checks of select sentences in this manuscript. Prompts such as "Please check the following sentence for a scientific publication, with a focus on grammar, spelling, and academic style" were used, followed by one or several sentences.

Data Availability

The raw data from this study are not publicly available. Deidentified data can be shared upon request and on a case-by-case basis.

Authors' Contributions

AB, WT, KD, and WVH contributed to the study design and obtained institutional review board approval. The usability scenario was moderated by WT with assistance from AB. Data analysis was conducted by WT and AB. Findings were repeatedly discussed with SEMS and WVH, who provided guidance for further analysis. The first draft of the manuscript was written by AB. The manuscript was iteratively refined based on critical revisions by all authors. The final version of the manuscript was approved by all authors.

Conflicts of Interest

AB is an employee of Enzyre BV and has received travel support from CAF-DCF. SEMS received honoraria for a one-time workshop and travel support from Kedrion. WT is an employee of Panton. Panton received consultancy fees from Enzyre BV. AS served on advisory boards for Novo Nordisk, Pfizer, Genentech/Roche, Sanofi-Genzyme/Bioverativ, Hema Biologics, Be Biopharma, and BioMarin. She served as a consultant for Novo Nordisk and Kedrion Biopharma and participated in the speakers' bureaus of Novo Nordisk, Genentech/Roche, Kedrion Biopharma, and Sanofi-Genzyme/Bioverativ. She received clinical research support from Novo Nordisk, Pfizer, Genentech/Roche, Kedrion Biopharma, Sanofi-Genzyme/Bioverativ, Regeneron, Takeda Pharmaceuticals, Hema Biologics, Pharmacosmos A/S, Hemab, and Be Biopharma. She is a board member of Innovative Hematology, the Novo Nordisk Hemophilia Foundation, and the Plasminogen Deficiency Foundation. All honoraria and income from these activities are paid to the Indiana Hemophilia & Thrombosis Center.

NMAB has no conflicts to declare. KD has no conflicts to declare. WVH is a founder and employee of Enzyre BV. He is also co-owner of Van Heerde Holding BV, a shareholder of Enzyre. He received travel grant support from SOBI and CAF-DCF. Enzyre is the company developing the EnzySystem. Enzyre has received funding for contract research from Takeda and Novo Nordisk. In addition, Enzyre holds 2 granted patents and multiple pending patent applications related to the technology described in this manuscript.

Multimedia Appendix 1
Graphical user interface wireframe.
[PNG File, 1982 KB - formative_v9i1e77621_app1.png]

Multimedia Appendix 2
Quick reference instructions.
[PDF File, 675 KB - formative_v9i1e77621_app2.pdf]

Multimedia Appendix 3
Interview guide.

XSL•FO
RenderX

[PDF File, 180 KB - formative_v9i1e77621_app3.pdf]

Multimedia Appendix 4
Codebook.
[XLSX File, 25 KB - formative_v9i1e77621_app4.xlsx]

Checklist 1
COREQ checklist.
[PDF File, 495 KB - formative_v9i1e77621_app5.pdf]

References

1. Trossaert M, Lienhart A, Nougier C, et al. Diagnosis and management challenges in patients with mild haemophilia A and discrepant FVIII measurements. Haemophilia 2014 Jul;20(4):550-558. [doi: 10.1111/hae.12381] [Medline: 24517184]
2. van Dijk K, Fischer K, van der Bom JG, Grobbee DE, van den Berg HM. Variability in clinical phenotype of severe haemophilia: the role of the first joint bleed. Haemophilia 2005 Sep;11(5):438-443. [doi: 10.1111/j.1365-2516.2005.01124.x] [Medline: 16128885]
3. den Uijl IEM, Fischer K, Van Der Bom JG, Grobbee DE, Rosendaal FR, Plug I. Analysis of low frequency bleeding data: the association of joint bleeds according to baseline FVIII activity levels. Haemophilia 2011 Jan;17(1):41-44. [doi: 10.1111/j.1365-2516.2010.02383.x] [Medline: 20825504]
4. Srivastava A, Santagostino E, Dougall A, et al. WFH Guidelines for the Management of Hemophilia, 3rd edition. Haemophilia 2020 Aug;26 Suppl 6(S6):1-158. [doi: 10.1111/hae.14046] [Medline: 32744769]
5. Fischer K, Collins PW, Ozelo MC, Srivastava A, Young G, Blanchette VS. When and how to start prophylaxis in boys with severe hemophilia without inhibitors: communication from the SSC of the ISTH. J Thromb Haemost 2016 May;14(5):1105-1109. [doi: 10.1111/jth.13298] [Medline: 27186714]
6. Dargaud Y, Béguin S, Lienhart A, et al. Evaluation of thrombin generating capacity in plasma from patients with haemophilia A and B. Thromb Haemost 2005 Mar;93(3):475-480. [doi: 10.1160/TH04-10-0706] [Medline: 15735797]
7. Dargaud Y, Negrier C, Rusen L, et al. Individual thrombin generation and spontaneous bleeding rate during personalized prophylaxis with Nuwiq® (human-cl rhFVIII) in previously treated patients with severe haemophilia A. Haemophilia 2018 Jul;24(4):619-627. [doi: 10.1111/hae.13493] [Medline: 29855112]
8. Delavenne X, Ollier E, Lienhart A, Dargaud Y. A new paradigm for personalized prophylaxis for patients with severe haemophilia A. Haemophilia 2020 Mar;26(2):228-235. [doi: 10.1111/hae.13935] [Medline: 32100950]
9. Peyvandi F, Kenet G, Pekrul I, Pruthi RK, Ramge P, Spannagl M. Laboratory testing in hemophilia: impact of factor and non-factor replacement therapy on coagulation assays. J Thromb Haemost 2020 Jun;18(6):1242-1255. [doi: 10.1111/jth.14784] [Medline: 32115865]
10. Lenting PJ. Laboratory monitoring of hemophilia A treatments: new challenges. Blood Adv 2020 May 12;4(9):2111-2118. [doi: 10.1182/bloodadvances.2019000849] [Medline: 32396619]
11. Kost GJ, editor. Principles & Practice of Point-of-Care Testing: Lippincott Williams & Wilkins; 2002.
12. Tudos AJ, Besselink GJ, Schasfoort RB. Trends in miniaturized total analysis systems for point-of-care testing in clinical chemistry. Lab Chip 2001 Dec;1(2):83-95. [doi: 10.1039/b106958f] [Medline: 15100865]
13. Lilly CM, Wang Z, Dunlap D, et al. 2021 patient preferences for point of care testing survey: more acceptance and less concern. J Appl Lab Med 2022 Oct 29;7(6):1302-1310. [doi: 10.1093/jalm/jfac070] [Medline: 36093730]
14. Lilly CM, Ensom E, Teebagy S, et al. Patient preferences for point-of-care testing: survey validation and results. Point Care 2020 Dec;19(4):112-115. [doi: 10.1097/poc.0000000000000214] [Medline: 34707464]
15. Price CP, Kricka LJ. Improving healthcare accessibility through point-of-care technologies. Clin Chem 2007 Sep;53(9):1665-1675. [doi: 10.1373/clinchem.2006.084707] [Medline: 17660275]
16. Varzgaliene L, Heerey A, Cox C, et al. Point-of-care testing in primary care: needs and attitudes of Irish GPs. BJGP Open 2018 Jan;1(4):bjgpopen17X101229. [doi: 10.3399/bjgpopen17X101229] [Medline: 30564692]
17. Mohammadi Aria M, Erten A, Yalcin O. Technology advancements in blood coagulation measurements for point-of-care diagnostic testing. Front Bioeng Biotechnol 2019;7:395. [doi: 10.3389/fbioe.2019.00395] [Medline: 31921804]
18. Sackmann EK, Fulton AL, Beebe DJ. The present and future role of microfluidics in biomedical research. Nature New Biol 2014 Mar 13;507(7491):181-189. [doi: 10.1038/nature13118] [Medline: 24622198]
19. Convery N, Gadegaard N. 30 years of microfluidics. Micro Nano Eng 2019 Mar;2:76-91. [doi: 10.1016/j.mne.2019.01.003]
20. Lyon AR, Aung T, Bruzios KE, Munson S. Human-centered design to enhance implementation and impact in health. Annu Rev Public Health 2025 Apr;46(1):467-485. [doi: 10.1146/annurev-publhealth-071823-122337] [Medline: 39656968]
21. Steeghs TJ, van Engelshoven L, Kuijpers SPL, et al. Novel microfluidic device for factor VIII quantification by chemiluminescence in hemophilia A patients. Biosens Bioelectron 2025 Aug 15;282:117469. [doi: 10.1016/j.bios.2025.117469] [Medline: 40286644]

XSL•FO
RenderX

22. Tong A, Sainsbury P, Craig J. Consolidated Criteria for Reporting Qualitative Research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
23. Kelley JF. An empirical methodology for writing user-friendly natural language computer applications. Presented at: CHI '83: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems; Dec 12-15, 1983; Boston, MA. [doi: 10.1145/800045.801609]
24. Applying human factors and usability engineering to medical devices: guidance for industry and Food and Drug Administration staff. US Food and Drug Administration. 2016 Feb 3. URL: https://www.fda.gov/media/80481/download [accessed 2025-10-07]
25. Cooke L. Assessing concurrent think-aloud protocol as a usability test method: a technical communication approach. IEEE Trans Profess Commun 2010;53(3):202-215. [doi: 10.1109/TPC.2010.2052859]
26. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
27. Borsci S, Buckle P, Hanna GB. Why you need to include human factors in clinical and empirical studies of in vitro point of care devices? Review and future perspectives. Expert Rev Med Devices 2016;13(4):405-416. [doi: 10.1586/17434440.2016.1154277] [Medline: 26878393]
28. Jones CHD, Howick J, Roberts NW, et al. Primary care clinicians' attitudes towards point-of-care blood testing: a systematic review of qualitative studies. BMC Fam Pract 2013 Aug 14;14(1):117. [doi: 10.1186/1471-2296-14-117] [Medline: 23945264]
29. Hardy V, Thompson M, Alto W, et al. Exploring the barriers and facilitators to use of point of care tests in family medicine clinics in the United States. BMC Fam Pract 2016 Nov 3;17(1):149. [doi: 10.1186/s12875-016-0549-1] [Medline: 27809865]
30. Quinn AD, Dixon D, Meenan BJ. Barriers to hospital-based clinical adoption of point-of-care testing (POCT): a systematic narrative review. Crit Rev Clin Lab Sci 2016;53(1):1-12. [doi: 10.3109/10408363.2015.1054984] [Medline: 26292075]
31. Iorio A. Using pharmacokinetics to individualize hemophilia therapy. Hematology Am Soc Hematol Educ Program 2017 Dec 8;2017(1):595-604. [doi: 10.1182/asheducation-2017.1.595] [Medline: 29222309]
32. Fernandes S, Carvalho M, Lopes M, Araújo F. Impact of an individualized prophylaxis approach on young adults with severe hemophilia. Semin Thromb Hemost 2014 Oct;40(7):785-789. [doi: 10.1055/s-0034-1394109] [Medline: 25313523]
33. Hazendonk H, Lock J, Mathôt RAA, et al. Perioperative treatment of hemophilia A patients: blood group O patients are at risk of bleeding complications. J Thromb Haemost 2016 Mar;14(3):468-478. [doi: 10.1111/jth.13242] [Medline: 26714028]
34. Huddy JR, Ni MZ, Barlow J, Hanna GB. Qualitative analysis of stakeholder interviews to identify the barriers and facilitators to the adoption of point-of-care diagnostic tests in the UK. BMJ Open 2021 Apr 13;11(4):e042944. [doi: 10.1136/bmjopen-2020-042944] [Medline: 33849848]
35. Dózsa C, Horváth K, Cserni I, Cseh B. Roadmap for large-scale implementation of point-of-care testing in primary care in Central and Eastern European countries: the Hungarian experience. Prim Health Care Res Dev 2022 Apr 21;23:e26. [doi: 10.1017/S1463423622000159] [Medline: 35445652]
36. Shephard M, Shephard A, Matthews S, Andrewartha K. The benefits and challenges of point-of-care testing in rural and remote primary care settings in Australia. Arch Pathol Lab Med 2020 Nov 1;144(11):1372-1380. [doi: 10.5858/arpa.2020-0105-RA] [Medline: 33106858]
37. Kuecuekbalaban P, Schmidt S, Kraft K, Hoffmann W, Muehlan H. Exploring risks and benefits of point-of-care tests for healthcare and self-tests for laypersons: an interview study assessing complementary expert perspectives on diagnostic lab-on-a-chip systems. Technol Health Care 2014;22(6):817-833. [doi: 10.3233/THC-140857] [Medline: 25318957]
38. Buetow S. Thematic analysis and its reconceptualization as "saliency analysis". J Health Serv Res Policy 2010 Apr;15(2):123-125. [doi: 10.1258/jhsrp.2009.009081] [Medline: 19762883]
39. User experience design. Semantic Studios. URL: https://semanticstudios.com/user_experience_design/ [accessed 2025-10-06]
40. Olmos-Vega FM, Stalmeijer RE, Varpio L, Kahlke R. A practical guide to reflexivity in qualitative research: AMEE Guide No. 149. Med Teach 2022 Apr 7:1-11. [doi: 10.1080/0142159X.2022.2057287] [Medline: 35389310]
41. Virzi RA. Refining the test phase of usability evaluation: how many subjects is enough? Hum Factors 1992 Aug;34(4):457-468. [doi: 10.1177/001872089203400407]
42. Nielsen J, Landauer TK. A mathematical model of the finding of usability problems. Presented at: INTERCHI93: Conference on Human Factors in Computing; Apr 24-29, 1993; Amsterdam, The Netherlands. [doi: 10.1145/169059.169166]
43. Faulkner L. Beyond the five-user assumption: benefits of increased sample sizes in usability testing. Behav Res Methods Instrum Comput 2003 Aug;35(3):379-383. [doi: 10.3758/bf03195514] [Medline: 14587545]
44. Borsci S, Macredie RD, Martin JL, Young T. How many testers are needed to assure the usability of medical devices? Expert Rev Med Devices 2014 Sep;11(5):513-525. [doi: 10.1586/17434440.2014.940312] [Medline: 25033757]
45. Braun V, Clarke V. To saturate or not to saturate? Questioning data saturation as a useful concept for thematic analysis and sample-size rationales. Qualitative Res Sport Exerc Health 2021 Mar 4;13(2):201-216. [doi: 10.1080/2159676X.2019.1704846]

XSL•FO
RenderX

46. Varpio L, Ajjawi R, Monrouxe LV, O'Brien BC, Rees CE. Shedding the cobra effect: problematising thematic emergence, triangulation, saturation and member checking. Med Educ 2017 Jan;51(1):40-50. [doi: 10.1111/medu.13124] [Medline: 27981658]

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Studies
EMR: electronic medical record
FVIII: factor VIII
HCP: health care provider
POCT: point-of-care testing
QRI: quick reference instructions
TGA: thrombin generation assay

Edited by A Mavragani, J Sarvestan; submitted 16.05.25; peer-reviewed by M Joseph John; revised version received 06.08.25; accepted 21.09.25; published 16.10.25.

Please cite as:
Bavinck A, Schols SEM, Teunissen W, Shapiro A, Blijlevens NMA, Davis K, van Heerde W
Measuring FVIII Activity and Thrombin Generation Simultaneously With a Novel Point of Care Platform (EnzySystem HemA): Qualitative Usability Evaluation
JMIR Form Res 2025;9:e77621
URL: https://formative.jmir.org/2025/1/e77621
doi:10.2196/77621

© Aernoud Bavinck, Saskia E M Schols, Wilfred Teunissen, Amy Shapiro, Nicole M A Blijlevens, Kyle Davis, Waander van Heerde. Originally published in JMIR Formative Research (https://formative.jmir.org), 16.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Feasibility of a Guided Web-Based Procrastination Intervention for College Students: Open Trial

Sevin Ozmen, MSc; Arpana Amarnath, MSc; Leonore de Wit, PhD; Chris van Klaveren, PhD; Pim Cuijpers, PhD; Annemieke van Straten, PhD; Sascha Struijs, PhD; Caring Universities Consortium

[1]Department of Clinical, Neuro and Developmental Psychology, Amsterdam Public Health Research Institute, Vrije Universiteit Amsterdam, De Boelelaan 1105, Amsterdam, The Netherlands

Corresponding Author:
Sevin Ozmen, MSc
Department of Clinical, Neuro and Developmental Psychology, Amsterdam Public Health Research Institute, Vrije Universiteit Amsterdam, De Boelelaan 1105, Amsterdam, The Netherlands

Abstract

Background: College students commonly struggle with procrastination, which is linked to mental health complaints and poor academic performance. Interventions based on cognitive behavioral therapy can be effective in reducing procrastination. Traditional face-to-face therapy and online interventions have shown promising outcomes, with the latter overcoming help-seeking barriers such as lengthy referral processes and waiting lists.

Objective: This study aims to examine the feasibility and acceptability of a new eHealth intervention targeting procrastination for college students ("GetStarted") with guidance by student e-coaches. This cognitive behavioral therapy–based intervention was designed specifically for and together with the target demographic of students studying in the Netherlands. Guidance was offered by trained clinical psychology students in the form of written motivational, supportive messages.

Methods: We conducted a single-arm study. The primary outcomes were satisfaction (8-item Client Satisfaction Questionnaire [CSQ-8]), usability (10-item System Usability Scale [SUS-10]), and adherence (completion rate). The secondary outcomes were changes to procrastination (Irrational Procrastination Scale [IPS]), depression (9-item Patient Health Questionnaire [PHQ-9]), stress (10-item Perceived Stress Scale [PSS-10]), quality of life (Mental Health Quality of Life Questionnaire [MHQoL]), and e-coaching satisfaction (Working Alliance Inventory for Guided Internet Interventions [WAI-I]).

Results: Of 734 participants who started the intervention, 335 (45.6%) completed the posttest. Students reported being satisfied with the intervention (CSQ-8: mean 23.48, SD 3.23) and found it very usable (SUS-10: mean 34.39, SD 4.52). Regarding adherence, participants completed 68.95% of the intervention on average, while 36.65% (n=269) of participants completed the full intervention. Participants showed a significant decrease in procrastination (IPS: mean decrease 35.39-32.56, Cohen *d*=0.63), depression (PHQ-9: mean decrease 9.27-7.73, Cohen *d*=0.35), and stress (PSS-10: mean decrease 20.79-19.02, Cohen *d*=0.31) as well as an increase in quality of life (MHQoL: mean increase 12.81-13.65, Cohen *d*=0.37) from baseline to posttest to follow-up. Participants reported a moderate-to-strong alliance with their e-coach (WAI-I: mean 45.26, SD 7.72).

Conclusions: The internet-based, student-guided intervention "GetStarted" targeting procrastination appears to be acceptable and feasible for college students in the Netherlands. However, high attrition rates and the lack of a control group mean that results must be interpreted with caution. To further examine intervention effectiveness, a randomized controlled trial needs to be conducted.

International Registered Report Identifier (IRRID): RR2-10.2196/44907

(JMIR Form Res 2025;9:e72896) doi:10.2196/72896

KEYWORDS

adherence; eHealth; guidance; online interventions; procrastination; satisfaction

Introduction

Procrastination, which refers to the voluntary delay of a necessary course of action despite expecting negative consequences [1], is pervasive among college or university students (henceforth referred to as "college students"). Academic environments, in particular, present unique challenges that could exacerbate procrastination, such as heavy workload, high autonomy, and competing priorities [2-4]. Consequently, studies suggest that 29%-67% of students [5-8] frequently procrastinate, making it a significant concern in higher education settings. In an academic setting, procrastination is associated with poor academic performance with low grades, increased dropout, and academic dishonesty among college students [9-12]. In addition,

XSL•FO
RenderX

procrastination has also been linked to several mental health concerns, such as depression, anxiety, stress, low self-esteem, and maladaptive coping [13-17]. Given these associations, it is not surprising that there would be long-term adverse effects of procrastination on overall health, interpersonal relationships, financial stability, and career opportunities [16,18,19].

It is therefore important to find and implement effective treatments that could help students understand and decrease their procrastination behaviors. Meta-analytical evidence shows that effective treatments for procrastination exist [20] and that cognitive behavioral therapy (CBT) was most effective in reducing procrastination in comparison with self-regulation training, strengths training, and acceptance-based behavior therapy [21].

However, accessing mental health care is a challenge for many young people. Unclear access points, lengthy referral processes, waiting lists, and concerns about privacy pose significant barriers [22]. In addition to this, some students who experience problems are reluctant to seek help for various reasons such as perceiving the problem as being temporary or minor, preferring to handle the problem alone or by talking to close ones, a perceived lack of time, fear of stigmatization, and limited awareness of available resources [23-26]. This highlights the need for more accessible, community-based support that is less stigmatized than traditional clinical interventions.

In recent years, eHealth interventions have gained popularity and proved well-suited to college students' dynamic lifestyles and diverse challenges [27-29]. From accessibility, convenience, cost-effectiveness, and anonymity to the possibilities of personalization, promotion of health literacy, feedback mechanisms, and cultural and linguistic inclusivity, eHealth interventions address many barriers to treatments commonly faced by students [27]. In line with this, some studies have shown that eHealth interventions effectively reduce procrastination among college students [30-32]. Compared to face-to-face, chat-based, and in-person group therapy, eHealth treatment has yielded comparable results in improving procrastination behaviors [33,34].

However, a common problem faced by most of these studies was that of nonadherence. Several eHealth interventions targeting different mental health concerns have had to face this issue of nonadherence with higher dropout rates [35-38] and participants not making optimal use of interventions [39]. This undermines the efficacy of the interventions and reduces their potential to deliver meaningful health outcomes [40]. To add to this, the very nature of procrastination could exacerbate nonadherence as these individuals could continuously put off participating in or completing the intervention [32]. There could be several reasons for nonadherence, such as lack of motivation, engagement [41], or understanding of the platform's utility [37]. A possible solution could be adding guidance to the intervention [42-44]. Providing structured guidance could enhance user engagement, accountability, and the perceived value of the intervention [45,46]. However, this solution is not always feasible due to the scarcity of mental health professionals [47].

Therefore, it is imperative to form alternate and effective forms of guidance. In recent years, past studies have indicated that eHealth interventions that utilized guidance provided by nonclinicians (eg, peers, research assistants, or other lay persons) were just as effective as those guided by clinicians [48]. This form of guidance can have added benefits by reducing the cost of implementation and improving the scalability, accessibility, and flexibility of the intervention [49]. In addition, laypersons could foster a connection and a positive working alliance by reducing stigma, as the interaction could feel less clinical compared to a professional setting [48]. An available resource that could provide this type of guidance is guidance provided by clinical psychology students. As clinical psychology students receive psychological training during their studies, they will likely provide optimal qualitative guidance. However, to the best of our knowledge, this type of guidance in eHealth interventions for procrastination has not yet been evaluated.

We developed a new online intervention targeting procrastination ("GetStarted") for and in collaboration with students studying in the Netherlands, addressing the lack of existing interventions specifically designed for this group. This allowed us to ensure that the contents of the intervention were relevant and appealing to this target demographic. The intervention was based on CBT, given meta-analytical evidence that this is the most effective type of intervention [21]. Specifically, the intervention utilized cognitive restructuring, which is a key mechanism of change in CBT [50].

Due to its novel nature, the intervention has not yet been examined in the scientific literature. Therefore, this study evaluates the feasibility and accessibility of a newly developed online intervention guided by psychology student e-coaches to reduce procrastination among college students. The following research questions (RQs) were addressed:

- Main RQs:
- RQ1: How satisfied are participants with the new online intervention "GetStarted" and its platform?
- RQ2: What are the adherence rates of the new intervention?
- Secondary RQs:
- RQ3: What are the differences between baseline to posttest to follow-up for procrastination, depression, perceived stress levels, and quality of life?
- RQ4: How satisfied are participants with the e-coaching provided by clinical psychology students?

Methods

Study Design

A single-arm open trial design was used to assess the feasibility and acceptability of a new eHealth intervention aimed at reducing procrastination among college students, which are the primary aims at this stage. This time- and cost-efficient approach allowed for initial insights into engagement and implementation before proceeding to a randomized controlled trial to better examine effectiveness. In this single-arm within-group design, assessments were administered at baseline (t0), post intervention (t2), and follow-up (t2), occurring at 4 weeks and 6 months post baseline, respectively.

XSL•FO
RenderX

Ethical Considerations

The ethical approval was granted by the VU Scientific and Ethical Review Board (reference 2020.088). Participants gave their informed consent, and their privacy and confidentiality were respected through the omission of any identifying details. Participants were not offered compensation for their participation. This study is conducted within the Caring Universities Consortium, which comprises 7 Dutch universities (VU Amsterdam, Leiden University, Maastricht University, Utrecht University, Erasmus University, University of Amsterdam, and Inholland University of Applied Sciences). This consortium is affiliated with the World Health Organization's World Mental Health International College Student Initiative (WMH-ICS) [51].

This study adheres to the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines. A completed checklist is provided in Checklist 1.

Sample Size

There is no standardized method for determining the sample size of an open feasibility study, and previous research has recommended 12 [52] to 35 or more participants [53]. Drawing on these guidelines and comparable studies [54,55], we estimated that 50 participants would be sufficient to address our primary objectives. However, the present intervention was available to the general student population for several years, and we expected that more than 50 participants would sign up. For quality assurance, a power analysis was conducted, and aiming for a conservative 2-tailed calculation with a small effect size (0.2), our statistical analyses (estimating differences between 2 dependent means) would require a total sample size of 199 participants [56]. In this study, we included all participants who had used the intervention and completed the posttest at the time of data analysis to be able to draw firmer conclusions.

Participants and Enrollment

Participants were students enrolled in any one of the 7 Dutch universities within the Caring Universities consortium. Recruitment took place in three ways from January 2021 to December 2023. First, through an annual online mental health survey of the WMH-ICS, participants who scored above 28 points on the Irrational Procrastination Scale (IPS) in the survey were invited to participate in the intervention. Second, through social media marketing as well as on-campus marketing via the distribution of posters, flyers, and other promotional materials. Third, college staff members such as student psychologists, study advisors, and lecturers could recommend the intervention to students who might benefit from it.

Eligibility Criteria

Participants were eligible if they were (1) aged 16 years or older, (2) enrolled as a student in one of the participating universities, (3) reported no active suicidal ideation, and (4) provided informed consent.

We identified participants with suicidal ideation with a series of questions. Participants who indicated that they had recently had thoughts of killing themselves and were somewhat likely or very likely to act on these thoughts were excluded from the intervention. These participants were provided with an extensive list of possible resources to turn to. This message was also sent to their email address.

Intervention

The intervention "GetStarted" was created in collaboration with a User Experience designer and followed the principles of an optimized user interface [57]. Based on the principles of CBT, the intervention consisted of 5 mandatory modules and 4 optional modules. The optional modules became available to participants upon completion of the second mandatory module. Participants could choose whether they wanted to do optional modules, which ones, and in which order. An overview of all the modules can be found in Table 1.

XSL•FO
RenderX

Table . Intervention content.

Module type	Module name	Contents
Mandatory	You, the monkey and the monster	Introduction to the program and its contents
Mandatory	Protection level: 5000	Understanding the psychological processes behind procrastination: procrastination as an avoidance mechanism
Mandatory	A chat with Procrastination Brain	Identifying the negative thoughts and feelings that accompany certain tasks, causing procrastination
Mandatory	I've got to break free	Cognitive restructuring to challenge and change negative thinking patterns
Mandatory	Self high-five	Reflecting on and celebrating one's progress, summary of all main modules
Optional	Bigger isn't always better	Breaking a big task into smaller parts and prioritizing or ordering these tasks
Optional	Panic Monster to the rescue	Estimating how much time a task will take, creating a timeline, methods to uphold the timeline
Optional	Focus booster pack	Ways to increase productivity (eg, preparing properly, eliminating distractions, working in time blocks)
Optional	The master of motivation	Motivational techniques (eg, goal setting, visualization, positive thinking)

Guidance

The intervention was guided by e-coaches. These e-coaches were trained (research) master's students in clinical psychology and third-year clinical psychology bachelor students. To be eligible for the position of e-coach, students had to have successfully completed courses on diagnostics and psychological interviewing. Selected candidates completed 6 hours of training before commencing their coaching activities.

When participants enrolled for the intervention, they were given the option to choose an e-coach based on a photo and a short biography of 3 random e-coaches. From then on, the e-coaches supported the participants and provided guidance through asynchronous, text-based feedback delivered via the platform. Feedback was written based on the participants' answers to the questions posed to them by the intervention. In the feedback, coaches showed empathy and understanding, encouraged the participant to apply the newly learned techniques, and asked reflective questions.

These coaches also attended weekly 1-hour intervision sessions under the supervision of the research team to enhance their coaching skills and maintain quality standards. During each intervision session, several e-coaches would share messages they had sent to their participants and would receive feedback from their peers and the supervisor. The supervisor also checked messages sent by e-coaches on the platform on a sample basis as a form of quality control.

Assessment Measures

Primary Outcomes

Satisfaction With the Intervention

The 8-item Client Satisfaction Questionnaire (CSQ-8) was used to evaluate participants' satisfaction with the intervention [58]. It consists of 8 questions on a 4-point Likert scale (sum score 8 - 32), and a higher score indicates greater satisfaction. The CSQ-8 has shown high reliability and validity for online interventions [59].

Usability

The usability of the intervention was assessed using the 10-item System Usability Scale (SUS-10) [60]. It consists of 10 questions on a 5-point Likert scale (sum score 0 - 40). Total scores are then multiplied by 2.5 to achieve a total score of 0-100, where a higher score indicates greater usability. The SUS-10 has shown good psychometric properties (reliability and validity) [61].

Adherence

Adherence is "the degree to which the user followed the program as it was designed" [62]. In this study, modules 1 - 4 were considered core modules as they contained the CBT elements. Therefore, participants who completed these 4 core modules were considered completers of the intervention. Additionally, average adherence across the total sample was calculated by dividing the number of core modules completed by the total number of core modules in the program and multiplying this by 100. The resulting percentage indicates the adherence rate.

Participants who did not complete a single module were excluded from the analyses as they did not start the treatment.

XSL•FO
RenderX

Secondary Outcomes

Procrastination Tendencies

Procrastination behavior was measured using the IPS [63,64]. It consists of 9 questions on a 5-point Likert scale (sum score 9 - 45), and a higher score indicates more procrastination behavior. The IPS showed good internal consistency of Cronbach α=0.91 [63]; a high level of reliability (point measure correlation of 0.58 - 0.74); and good content, structural, and substantive validity [65].

Depressive Symptoms

The 9-item Patient Health Questionnaire (PHQ-9) was used to measure depression [66]. It comprises 9 questions on a 4-point Likert scale (sum score 0 - 27), and a higher score indicates a higher level of depression. The PHQ-9 showed high sensitivity (0.71 - 0.84), specificity (0.90 - 0.97), internal consistency (Cronbach α=0.86 - 0.89), test-retest reliability (*r*=0.84), and validity [66,67].

Perceived Stress

The 10-item Perceived Stress Scale (PSS-10) was used to measure stress [68,69]. It comprises 10 questions on a 5-point Likert scale (sum score 0 - 40), where a higher score indicates higher perceived stress. The PSS-10 showed high validity, internal consistency (Cronbach α=0.74 - 0.91), and test-retest reliability (*r*=0.74 - 0.88) [69,70].

Quality of Life

The Mental Health Quality of Life Questionnaire (MHQoL) was used to assess quality of life [71]. It comprises 7 questions about different dimensions of life (eg, mood, relationships) on a 4-point Likert scale (sum score 0 - 21), where a higher score indicates better quality of life. The MHQoL showed good validity, test-retest reliability (*r*=0.85), and internal consistency (Cronbach α=0.85) [71].

E-Coach Evaluation

Participants' satisfaction with the e-coach was assessed using the Working Alliance Inventory for Guided Internet Interventions (WAI-I) [72]. It comprises 12 questions on a 5-point Likert scale (sum score 12 - 60), where higher scores indicate higher satisfaction. A score of 36 or higher indicates sufficient satisfaction, where a score of 48-60 means high to very high satisfaction with the e-coach. The psychometric characteristics of the WAI-I yielded adequate results [72].

Sociodemographic Information

The following sociodemographic information was collected: age, gender, marital status, nationality, attending university, faculty, education level (bachelor's, master's, or PhD), and whether psychotherapy, medication, or both were received. This information was collected, first, to gain insight into the characteristics of individuals who are interested in the intervention when offered access to it in a real-life setting. Second, we wanted to ensure that there were no major differences between participants who did and did not complete the study to account for high attrition rates.

Assessments

The overview of the assessments can be seen in Table 2.

Table . Measures and assessment points.

	Assessment points		
	T0[a]	T1[b]	T2[c]
Sociodemographics	✓		
8-item Client Satisfaction Questionnaire		✓	
10-item System Usability Scale		✓	
Working Alliance Inventory for Guided Internet Interventions		✓	
Irrational Procrastination Scale	✓	✓	✓
10-item Perceived Stress Scale	✓	✓	✓
9-item Patient Health Questionnaire	✓	✓	✓
Mental Health Quality of Life Questionnaire	✓	✓	✓

[a]T0: baseline.

[b]T1: posttest (4 wk after T0).

[c]T2: Follow-up (6 mo after T0).

Data Analysis

IBM SPSS version 27 was used for the data analyses.

First, we examined the baseline characteristics of the whole sample, study completers versus noncompleters (ie, participants who did vs did not complete the posttest assessment) and intervention completers versus noncompleters (ie, participants who did vs did not complete at least 4 main modules). Potential differences in baseline characteristics between these groups were examined using χ^2 tests and independent sample *t* tests.

For our primary outcomes of client satisfaction (CSQ-8) and usability (SUS-10), we conducted a complete case analysis and

XSL•FO
RenderX

analyzed only study completers. We calculated descriptive statistics for the whole sample and examined intervention completers versus noncompleters as an additional sensitivity analysis. For the primary outcome of adherence, we calculated how much of the program was completed on average, as well as the percentage of participants who completed the intervention versus did not complete the intervention.

For our secondary outcomes, which were of an exploratory nature, we conducted a complete case analysis and examined whether there was a statistical difference from baseline to posttest in reported procrastination behavior (IPS), depression (PHQ-9), stress (PSS-10), and quality of life (MHQoL). We conducted 2-tailed paired *t* tests using a significance level of α=.05 to assess these changes. To interpret the effect size, we calculated the Cohen *d*, interpreting the benchmarks of 0.2, 0.5, and 0.8 as small, moderate, and large, respectively [73].

Results

Participants

A total of 1746 students were assessed for eligibility for the intervention between January 2021 and May 2023. Of these, 657 participants did not receive access to the intervention because they dropped out during the baseline assessment (n=204), dropped out during account creation (n=359), or were excluded due to active suicidal ideation (n=94).

A total of 1089 participants were granted access to the intervention. Of these, 355 (32.6%) did not start the intervention. The reasons for not accessing the intervention were not activating their account from a link they received via email (n=121) or completing zero modules within the intervention and therefore not starting the treatment (n=234). There were no baseline differences between those who started the intervention and those who did not.

A total of 734 participants started the intervention. Our analyses on adherence were conducted based on this sample. Of these 734 participants, 335 completed the posttest (ie, study completers) and were included in the satisfaction and clinical outcome analyses. The details on the participant flow can be found in Figure 1.

XSL•FO
RenderX

Figure 1. Flow diagram of participants.

XSL•FO
RenderX

The average age of the total sample (n=734) was 23.6 (SD 4.15) years, and most participants were female (n=543, 74%). Most were from the Netherlands (n=446, 60.8%) or another country in Europe (28.6%). Almost half of the participants were currently doing their master's (n=340, 46.3%). A vast majority of participants (n=615, 83.8%) were currently receiving no medication or psychotherapy. We compared baseline characteristics of intervention completers and noncompleters as well as study completers and noncompleters. Age, gender, nationality, university, education level, marital status, current professional help, and baseline clinical characteristics (procrastination, depression, stress, and quality of life) were compared. We found no significant differences between intervention completers versus noncompleters. When comparing study completers versus noncompleters, the only significant difference we found was the current professional help, with study noncompleters reporting use of medication, psychotherapy, or both more frequently than study completers (P<.001). Full details on the baseline characteristics of the total sample as well as of study completers versus noncompleters and intervention completers versus noncompleters can be found in Multimedia Appendix 1.

Main RQs

RQ1: How Satisfied Are Participants With the New Online Intervention "GetStarted" and Its Platform?

Satisfaction

Satisfaction was assessed among participants who completed the posttest assessment (n=335). The average score on the CSQ-8 for the total sample was 23.48 (SD 3.23, scores ranging from 12 to 31). Intervention completers (n=175) had an average of 24.47 (SD 3.03), and intervention noncompleters (n=160) averaged 22.39 (SD 3.09). Intervention completers were significantly more satisfied than noncompleters (t_{333}=6.212; P<.001; Cohen d=0.68).

Usability

Usability was assessed among participants who completed the posttest assessment (n=335). The average score on the SUS-10 for the total sample was 34.39 (SD 4.52, scores ranging from 16 to 40). Intervention completers (n=175) had an average of 35.51 (SD 3.68), while intervention noncompleters (n=160) averaged 33.17 (SD 5.01). The SUS-10 multiplies these scores by 2.5 to achieve a total score of 0-100. The total scores end up being 85.98 (total sample), 88.8 (completers), and 82.9 (noncompleters). Intervention completers reported significantly higher usability than noncompleters (t_{333}=4.94; P<.001; Cohen d=0.54).

RQ2: What Are the Adherence Rates of the New Intervention?

Adherence was calculated based on all participants who started the intervention (ie, completed at least 1 module; n=734). In this study, completing 4 main modules meant that the participant completed the intervention, which was done by 269 of 734 (36.65%) participants. This would indicate a dropout rate of 63.35% (n=465); however, participants who dropped out still partially adhered to the intervention. To capture this partial adherence, we calculated that participants completed an average of 2.97 out of 4 main modules, corresponding to 68.95% of the overall intervention.

Many participants also chose to do one or several optional modules. The most popular topics were "productivity" and "motivation boost," which were completed by 26.7% (n=196) and 24.5% (n=180) of participants, respectively. The optional modules about "task management" and "time management" were completed by 19.5% and 14.9% of participants, respectively.

Secondary RQs

RQ3: What Are the Differences Between Baseline to Posttest to Follow-Up for Procrastination, Depression, Perceived Stress Levels, and Quality of Life?

Procrastination, Mood, Stress, and Quality of Life Post Intervention

A total of 335 participants completed the posttest and were included in the analyses. We conducted a 2-tailed paired t test comparing the baseline and posttest for each outcome measure. Participants showed improvements in procrastination, depression, stress, and quality of life: IPS, PHQ-9, and PSS-10 scores all decreased significantly, while MHQoL scores increased significantly (all P<.001). The effect remains significant when correcting for multiple comparisons (type 1 error) using a Bonferroni-corrected α (k=4, adjusted α=0.0125). The standardized mean difference in procrastination scores was moderate to large (Cohen d=0.63). Standardized mean difference scores for depression, stress, and quality of life were small to moderate (Cohen d=0.35, 0.32, and 0.37, respectively). Full details can be found in Table 3.

XSL•FO
RenderX

Table . Secondary outcomes changes from baseline to posttest (n=335).

Secondary outcome	Baseline scores, mean (SD)	Posttest scores, mean (SD)	*t* test (*df*)	*P* value (2-tailed)	Cohen *d*	Cohen *d* 95% CI
IPS[a]	35.39 (4.31)	32.56 (5.34)	11.57 (334)	<.001	0.63	0.51-0.75
PHQ-9[b]	9.27 (4.7)	7.73 (4.57)	6.36 (334)	<.001	0.35	0.24-0.46
PSS-10[c]	20.79 (6.02)	19.02 (6.3)	5.64 (334)	<.001	0.31	0.2-0.42
MHQoL[d]	12.81 (3.02)	13.65 (2.93)	6.75 (334)	<.001	0.37	0.26-0.48

[a]IPS: Irrational Procrastination Scale.

[b]PHQ-9: 9-item Patient Health Questionnaire.

[c]PSS-10: 10-item Perceived Stress Scale.

[d]MHQoL: Mental Health Quality of Life Questionnaire.

Procrastination, Mood, Stress, and Quality of Life at Follow-Up

A series of mixed-effects repeated measures models was conducted to evaluate changes in procrastination, depression symptoms, perceived stress, and mental health quality of life over time (n=734). The models revealed significant changes in all 4 outcomes across the 3 time points (all $P<.001$). Procrastination scores decreased by 2.86 points from preintervention to postintervention and by 5.38 points at follow-up ($\chi^2_2=400.89$). Depression symptoms were reduced by 1.62 points post intervention and by 2.58 points at follow-up ($\chi^2_2=109.04$). Perceived stress scores declined by 1.96 points post intervention and by 3.51 points at follow-up ($\chi^2_2=111.89$). Mental health quality of life improved over time, with increases of 0.87 points post intervention and 1.02 points at follow-up ($\chi^2_2=75.10$). The models accounted for individual differences in baseline scores across all outcomes.

RQ4: How Satisfied Are Participants with the E-Coaching Provided by Clinical Psychology Students?

E-coach satisfaction was assessed among participants who completed the posttest assessment (n=335). The average score on the WAI-I for the total sample was 45.26 (SD 7.72, scores ranging from 15 to 60). Participants who completed the intervention (n=175) reported higher satisfaction (mean 47.39, SD 6.91) than intervention noncompleters (n=160; mean 42.93, SD 0.62). The difference was significant ($t_{333}=5.52$; $P<.001$; Cohen $d=0.6$).

Discussion

Principal Results

This study demonstrates promising outcomes for the feasibility and acceptability of “GetStarted,” an eHealth intervention targeting procrastination specifically designed for college students guided by psychology students. Participants reported high satisfaction with the intervention and found the system usable. Adherence rates were acceptable given the nature of the intervention (self-help), the topic (procrastination), and the target group (students). While most participants did not finish the intervention, on average, they still completed a large part of the intervention. Participants showed a moderate-to-large decrease in procrastination, a small-to-moderate decrease in depression and stress, and a small-to-moderate increase in quality of life between pre- and posttest. However, these results must be interpreted with caution given the lack of a control group in the current study design. Lastly, participants were sufficiently satisfied with the guidance of student e-coaches.

Comparison With Prior Work

Our findings are in line with previous research on eHealth interventions among college students. Students report being satisfied with interventions for varying mental health problems such as stress, depression, and anxiety [54,74,75]. In this study, satisfaction scores were also found to be acceptable for the total sample and high for intervention completers. Internet interventions for mental health complaints are also generally found to have moderate-to-high usability [54,74,76,77]. Usability scores for the present intervention were high and corresponded with very good–to-excellent user experience [78].

Regarding intervention adherence, a meta-analysis reported dropout rates of up to 50.33% in guided and unguided internet interventions for college students [79]. More recent individual studies of guided internet interventions similar to ours reported dropout rates of 66.4% [74] and 48% [54]. In this study, 63.35% (n=288) of participants who started the intervention did not complete it. While this is on the high side, it is still in line with prior research. Additionally, on average, participants completed nearly 70% of the intervention and therefore were exposed to the larger part of the treatment. It is likely that participants benefited from this exposure to the treatment, even if they subsequently dropped out. We therefore deem our adherence rate on the low side, yet consider it acceptable in order to move forward with the “GetStarted” intervention.

This study also examined any changes to procrastination behavior, depression, stress, and quality of life in an exploratory fashion. While the absence of a control group means we cannot draw any conclusions on treatment effectiveness, we did see an improvement in all these areas. This is also in line with prior research, which has shown that internet interventions can be effective in treating different kinds of mental health problems in students, such as depression, anxiety, stress, and eating disorder symptoms [79] as well as procrastination [30-32]. The implication of our findings is 2-fold. First, this study provides preliminary indications that GetStarted may be a promising eHealth intervention to reduce procrastination behavior in

XSL•FO
RenderX

college students, as reflected by a moderate-to-large effect size. Second, our findings tentatively suggest that targeting one specific problem such as procrastination can also potentially reduce other mental health complaints such as depression or stress, even if the intervention was not aimed at these problems specifically. However, it must be noted that spontaneous remission or regression to the norm cannot be ruled out in our study, given our single-arm design. Further research utilizing a control group is needed to draw conclusions on intervention effectiveness.

Lastly, we examined whether guidance by trained psychology students was acceptable. Our findings indicate sufficient satisfaction with the guidance, which is comparable to other guided, internet-based, or face-to-face interventions [74,80,81].

Considering these findings, the new internet intervention "GetStarted" targeting procrastination in college students appears very promising.

Strengths and Limitations

A main strength of this study is the comprehensiveness of the assessment. When examining the acceptability and feasibility of new eHealth interventions, studies mainly focus on participant satisfaction and usability. While these were the main outcomes of this study, we also explored changes to procrastination and several other related mental health complaints. Given that procrastination is linked to mental health issues such as depression, social anxiety, stress, and low self-esteem [13,14,16,17], this study provides valuable insights into how treating procrastination may also impact mood, stress, and quality of life. Additionally, we employed posttest as well as follow-up measures. This allowed us to explore both short-term and longer term changes to these mental health complaints.

However, a major limitation of this study is that we cannot draw any conclusions about the effectiveness of the intervention on student mental health due to the lack of a control group. While this feasibility study was not designed to evaluate effectiveness, we did report pre-post change scores on secondary outcomes. While we found positive changes on all the measures, none of these can be directly attributed to the intervention given the lack of a control group. Regression to the mean or spontaneous remission offers alternative explanations for the improvements we found. Moreover, the observed effects may be inflated due to selective study attrition. Although a formal bias analysis falls outside the scope of this study, we acknowledge that the true change may be smaller than reported. Therefore, in future work, we plan to conduct a randomized controlled trial that includes a suitable comparison condition (eg, waitlist control, treatment-as-usual, or active placebo). This will allow us to more rigorously evaluate intervention effectiveness and account for nonspecific factors such as spontaneous remission and placebo effects.

A third limitation of this study regards the sample. During the recruitment process, all students of participating universities were offered access to the intervention. However, not all students are equally likely to be interested in or willing to sign up for the intervention, which means our findings are not directly generalizable to the total student population. For example, the majority of participants in this study were Dutch and female, which limits our understanding of intervention feasibility for other genders and nationalities. As of April 2025, 44.3% of university students in the Netherlands were male and 27.1% were international [82,83], compared to 25.2% (n=185) and 39.2% (n=288), respectively, in our sample. Our sample seems sufficiently diverse in terms of nationality; however, male students are underrepresented. A possible approach for future research is to first try to gain a deeper understanding of why male students in the Netherlands are less likely to make use of the intervention. Prior research has found that male students may perceive fewer benefits of and hold more stigma-related attitudes regarding mental health help-seeking [84]. Additional interviews with this target demographic could yield specific, actionable insights on how to increase the uptake of the "GetStarted" intervention. Additionally, targeted marketing efforts could be made. For example, colleges host many events for specific groups. These events can be a great opportunity to create awareness of the intervention among certain groups. Similarly, online marketing efforts such as social media posts could target certain demographics. A recommendation for future recruitment is to aim for a sample that is between 40% and 45% male participants. A more diverse study sample in further research on the intervention would greatly benefit the generalizability of future findings.

Lastly, the high attrition is a limitation in this study. Less than half of the participants who started the intervention completed the postmeasure, and about a third completed the 6-month follow-up. Though we found no differences in the baseline characteristics of study completers and noncompleters, we cannot rule out any unmeasured confounders. This results in a risk of attrition bias, meaning that "differences between people who leave a study and those that continue can be the reason for any observed effect and not the intervention itself" [85]. Future research should employ strategies to reduce attrition, such as calling participants as a reminder or offering a financial incentive, in order to avoid attrition bias. Both reminder-based approaches and financial incentives have been shown to improve retention in online intervention trials [86,87].

A strength of this study is our exploration of student e-coaches as the source of guidance. To the best of our knowledge, this study was the first to investigate the acceptability of guidance offered by trained psychology students. Prior research has shown that guided internet interventions are more effective than unguided treatment. However, mental health resources such as properly trained professionals are scarce. It is of paramount importance to find scalable and cost-effective alternatives to a licensed psychologist so that people in need can get timely mental health care. Clinical psychology students are required as well as eager to gain hands-on experience during their studies, which makes them a promising potential resource. This study shows that participants of our eHealth intervention are satisfied with the guidance offered by student coaches, and tapping into this more readily available resource means people in need can get more or quicker access to mental health support.

Finally, a last strength of this study is the scalability of our approach, starting with the guidance offered by e-coaches. In this study, e-coaches were involved as a part of their curriculum

XSL•FO
RenderX

(ie, a “mini-internship”), meaning they did receive study credits but not financial compensation. This greatly adds to the scalability of our approach. While supervising more e-coaches does require slightly more resources in terms of work hours, this investment is negligible compared to the costs involved with utilizing mental health professionals.

The digital platform used to deliver the intervention was specifically developed with scalability in mind. The platform was created by the Caring Universities Project, which is affiliated with the World Health Organization’s WMH-ICS [51]. We currently offer access to 11 digital interventions, including “GetStarted,” and adding additional interventions incurs minimal additional cost. Participating institutions pay a membership fee in exchange for unlimited access to all interventions. Importantly, there is no cap on the number of students who can enroll. Although there are some ongoing maintenance costs associated with platform upkeep, these remain relatively modest, especially in light of the wide range of interventions and the large number of students that can be reached. Additionally, given the noncommercial nature of Caring Universities, the per-institution fee decreases as additional universities join the initiative, while platform costs remain stable. This model offers substantial potential for large-scale dissemination and implementation.

Conclusions

The internet-based, student-guided intervention “GetStarted” targeting procrastination appears to be acceptable and feasible for college students in the Netherlands. The intervention could potentially also reduce complaints in the areas of procrastination, mood, and stress. Future research should determine possible intervention effectiveness using a control group, as well as explore strategies to address high study attrition.

Acknowledgments

This work was developed and conducted within the framework of the Caring Universities Consortium, a collaboration among several academic institutions, including Vrije Universiteit Amsterdam, Leiden University, Utrecht University, Maastricht University, Erasmus University Rotterdam, University of Amsterdam, InHolland University of Applied Sciences, Rotterdam University of Applied Sciences, and Avans University of Applied Sciences.

The group authors for the 'Caring Universities Consortium' author are: Sascha Struijs, Nadia Garnefski, Vivian Kraaij, Elske Salemink, Petra Hurks, Marilisa Boffo, Danielle Remmerswaal, Reinout Wiers, Claudia van der Heijde, Lisa Klinkenberg, Monique de Bruijn-Smolders & Jessica Nooij. Vrije Universiteit Amsterdam, Amsterdam, The Netherlands. Leiden University, Leiden, The Netherlands. Utrecht University, Utrecht, The Netherlands. Maastricht University, Maastricht, The Netherlands. Erasmus University Rotterdam, Rotterdam, The Netherlands. University of Amsterdam, Amsterdam, The Netherlands. InHolland University of Applied Sciences, Hoofddorp, The Netherlands. Rotterdam University of Applied Sciences, Rotterdam, The Netherlands. Avans University of Applied Sciences, Breda, The Netherlands.

The authors would also like to thank Michelle Donovan for her support in editing this manuscript and Ronald Haver and Pieter Overbeeke for their technical support.

Authors' Contributions

Conceptualization: PC, SO, SS
Data curation: AA, SO
Formal analysis: AA, CVK, SO, SS
Funding acquisition: LDW, PC, SS
Project administration: SO, SS
Software: SO
Supervision: AVS, LDW, PC, SS
Validation: SS
Writing – original draft: SO
Writing – review & editing: AA, AVS, CVK, LDW, PC, SO, SS

Conflicts of Interest

None declared.

Multimedia Appendix 1
Comparison of baseline characteristics.
[DOCX File, 21 KB - formative_v9i1e72896_app1.docx]

Checklist 1
STROBE checklist.
[PDF File, 326 KB - formative_v9i1e72896_app2.pdf]

XSL•FO
RenderX

References

1. Steel P. The nature of procrastination: a meta-analytic and theoretical review of quintessential self-regulatory failure. Psychol Bull 2007 Jan;133(1):65-94. [doi: 10.1037/0033-2909.133.1.65] [Medline: 17201571]
2. Aziz N, Zain Z, Mafuzi R, Mustapa AM, Najib NHM, Lah NFN. Prioritizing causes of procrastination among university students: an empirical analysis. J Telecommunication Electronic Computer Eng 2017;9:179-183 [FREE Full text]
3. Sulistia V, Widigdo AMN. Effect of workload, perceived organizational support, and time management on academic procrastination with work school conflict as a mediating variable. Eur J Business Manage Res 2023;8(1):206-212. [doi: 10.24018/ejbmr.2023.8.1.1797]
4. YeckehZaare I, Chen SS, Barghi T. Reducing procrastination without sacrificing students' autonomy through optional weekly presentations of student-generated content. Presented at: SIGCSE 2023; Mar 15-18, 2023; Toronto, ON p. 151-157. [doi: 10.1145/3545945.3569839]
5. Hayat AA, Jahanian M, Bazrafcan L, Shokrpour N. Prevalence of academic procrastination among medical students and its relationship with their academic achievement. Shiraz E-Med J 2020;21(7):e96049. [doi: 10.5812/semj.96049]
6. Mahasneh AM, Bataineh OT, Al-Zoubi ZH. The relationship between academic procrastination and parenting styles among Jordanian undergraduate university students. Open Psychol J 2016 May 24;9(1):25-34. [doi: 10.2174/1874350101609010025]
7. Onwuegbuzie AJ. Academic procrastination and statistics anxiety. Assess Eval High Educ 2004 Feb;29(1):3-19. [doi: 10.1080/0260293042000160384]
8. Özer BU, Demir A, Ferrari JR. Exploring academic procrastination among Turkish students: possible gender differences in prevalence and reasons. J Soc Psychol 2009 Apr;149(2):241-257. [doi: 10.3200/SOCP.149.2.241-257]
9. Aruğaslan E. Examining the relationship of academic dishonesty with academic procrastination, and time management in distance education. Heliyon 2024 Oct 15;10(19):e38827. [doi: 10.1016/j.heliyon.2024.e38827] [Medline: 39397943]
10. Bäulke L, Eckerlein N, Dresel M. Interrelations between motivational regulation, procrastination and college dropout intentions. Unterrichtswiss 2018 Dec;46(4):461-479. [doi: 10.1007/s42010-018-0029-5]
11. García-Ros R, Pérez-González F, Tomás JM, Sancho P. Effects of self-regulated learning and procrastination on academic stress, subjective well-being, and academic achievement in secondary education. Curr Psychol 2023 Oct;42(30):26602-26616. [doi: 10.1007/s12144-022-03759-8]
12. Kim KR, Seo EH. The relationship between procrastination and academic performance: a meta-analysis. Pers Individ Dif 2015 Aug;82:26-33. [doi: 10.1016/j.paid.2015.02.038]
13. Flett AL, Haghbin M, Pychyl TA. Procrastination and depression from a cognitive perspective: an exploration of the associations among procrastinatory automatic thoughts, rumination, and mindfulness. J Rat-Emo Cognitive-Behav Ther 2016 Sep;34(3):169-186. [doi: 10.1007/s10942-016-0235-1]
14. Schraw G, Wadkins T, Olafson L. Doing the things we do: a grounded theory of academic procrastination. J Educ Psychol 2007;99(1):12-25. [doi: 10.1037/0022-0663.99.1.12]
15. Sirois FM, Kitner R. Less adaptive or more maladaptive? A meta-analytic investigation of procrastination and coping. Eur J Pers 2015 Jul;29(4):433-444. [doi: 10.1002/per.1985]
16. Stead R, Shanahan MJ, Neufeld RWJ. "I'll go to therapy, eventually": procrastination, stress and mental health. Pers Individ Dif 2010 Aug;49(3):175-180. [doi: 10.1016/j.paid.2010.03.028]
17. Tan JF, Ma ZW, Li XT. Global self-esteem mediates the effect of general self-efficacy on Chinese undergraduates' general procrastination. Soc Behav Pers 2015 Sep 18;43(8):1265-1271. [doi: 10.2224/sbp.2015.43.8.1265]
18. Höcker A, Engberding M. Prokrastination Ein Manual Zur Behandlung Des Pathologischen Aufschiebens: Hogrefe; 2013, Vol. 70. URL: https://pub.uni-bielefeld.de/record/2619912 [accessed 2025-09-30]
19. Patrzek J, Grunschel C, Fries S. Academic procrastination: the perspective of university counsellors. Int J Adv Counselling 2012 Sep;34(3):185-201. [doi: 10.1007/s10447-012-9150-z]
20. Rozental A, Bennett S, Forsström D, et al. Targeting procrastination using psychological treatments: a systematic review and meta-analysis. Front Psychol 2018;9:1588. [doi: 10.3389/fpsyg.2018.01588] [Medline: 30214421]
21. van Eerde W, Klingsieck KB. Overcoming procrastination? A meta-analysis of intervention studies. Educ Res Rev 2018 Nov;25:73-85. [doi: 10.1016/j.edurev.2018.09.002]
22. Boonstra A, van Mastrigt G, Evers S, van Amelsvoort T, Leijdesdorff SMJ, @ease consortium. @ease peer - to - peer youth walk - in centres in the Netherlands: a protocol for evaluating longitudinal outcomes, follow - up results and cost - of - illness. Early Intervention Psych 2023 Sep;17(9):929-938. [doi: 10.1111/eip.13443]
23. Czyz EK, Horwitz AG, Eisenberg D, Kramer A, King CA. Self-reported barriers to professional help seeking among college students at elevated risk for suicide. J Am Coll Health 2013;61(7):398-406. [doi: 10.1080/07448481.2013.820731] [Medline: 24010494]
24. Ebert DD, Mortier P, Kaehlke F, et al. Barriers of mental health treatment utilization among first-year college students: first cross-national results from the WHO World Mental Health International College Student Initiative. Int J Methods Psychiatr Res 2019 Jun;28(2):e1782. [doi: 10.1002/mpr.1782] [Medline: 31069905]
25. Eisenberg D, Golberstein E, Gollust SE. Help-seeking and access to mental health care in a university student population. Med Care 2007 Jul;45(7):594-601. [doi: 10.1097/MLR.0b013e31803bb4c1] [Medline: 17571007]

XSL•FO
RenderX

26. Watkins DC, Hunt JB, Eisenberg D. Increased demand for mental health services on college campuses: perspectives from administrators. Qual Soc Work 2012 May;11(3):319-337. [doi: 10.1177/1473325011401468]
27. Liu XQ, Guo YX, Zhang XR, Zhang LX, Zhang YF. Digital interventions empowering mental health reconstruction among students after the COVID-19 pandemic. World J Psychiatry 2023 Jun 19;13(6):397-401. [doi: 10.5498/wjp.v13.i6.397] [Medline: 37383289]
28. Andersson G, Titov N. Advantages and limitations of Internet-based interventions for common mental disorders. World Psychiatry 2014 Feb;13(1):4-11. [doi: 10.1002/wps.20083] [Medline: 24497236]
29. Ebert DD, Cuijpers P, Muñoz RF, Baumeister H. Prevention of mental health disorders using internet- and mobile-based interventions: a narrative review and recommendations for future research. Front Psychiatry 2017;8:116. [doi: 10.3389/fpsyt.2017.00116] [Medline: 28848454]
30. Eckert M, Ebert DD, Lehr D, Sieland B, Berking M. Does SMS-support make a difference? Effectiveness of a two-week online-training to overcome procrastination. A randomized controlled trial. Front Psychol 2018;9:1103. [doi: 10.3389/fpsyg.2018.01103] [Medline: 30026713]
31. Lukas CA, Berking M. Reducing procrastination using a smartphone-based treatment program: a randomized controlled pilot study. Internet Interv 2018 Jun;12:83-90. [doi: 10.1016/j.invent.2017.07.002] [Medline: 30135772]
32. Rozental A, Forsell E, Svensson A, Andersson G, Carlbring P. Internet-based cognitive-behavior therapy for procrastination: a randomized controlled trial. J Consult Clin Psychol 2015 Aug;83(4):808-824. [doi: 10.1037/ccp0000023] [Medline: 25939016]
33. Gieselmann A, Pietrowsky R. Treating procrastination chat-based versus face-to-face: an RCT evaluating the role of self-disclosure and perceived counselor's characteristics. Comput Human Behav 2016 Jan;54:444-452. [doi: 10.1016/j.chb.2015.08.027]
34. Rozental A, Forsström D, Lindner P, et al. Treating procrastination using cognitive behavior therapy: a pragmatic randomized controlled trial comparing treatment delivered via the internet or in groups. Behav Ther 2018 Mar;49(2):180-197. [doi: 10.1016/j.beth.2017.08.002] [Medline: 29530258]
35. Cohen KA, Schleider JL. Adolescent dropout from brief digital mental health interventions within and beyond randomized trials. Internet Interv 2022 Mar;27:100496. [doi: 10.1016/j.invent.2022.100496] [Medline: 35257001]
36. Lillevoll KR, Vangberg HCB, Griffiths KM, Waterloo K, Eisemann MR. Uptake and adherence of a self-directed internet-based mental health intervention with tailored e-mail reminders in senior high schools in Norway. BMC Psychiatry 2014 Jan 21;14:14. [doi: 10.1186/1471-244X-14-14] [Medline: 24443820]
37. Oti O, Pitt I. Online mental health interventions designed for students in higher education: a user-centered perspective. Internet Interv 2021 Dec;26:100468. [doi: 10.1016/j.invent.2021.100468] [Medline: 34703772]
38. Shams F, Tai AMY, Kim J, et al. Adherence to e-health interventions for substance use and the factors influencing it: systematic review, meta-analysis, and meta-regression. Digit Health 2023;9:20552076231203876. [doi: 10.1177/20552076231203876] [Medline: 37780062]
39. Lin J, Faust B, Ebert DD, Krämer L, Baumeister H. A web-based acceptance-facilitating intervention for identifying patients' acceptance, uptake, and adherence of internet- and mobile-based pain interventions: randomized controlled trial. J Med Internet Res 2018 Aug 21;20(8):e244. [doi: 10.2196/jmir.9925] [Medline: 30131313]
40. Linardon J, Fuller-Tyszkiewicz M. Attrition and adherence in smartphone-delivered interventions for mental health problems: a systematic and meta-analytic review. J Consult Clin Psychol 2020 Jan;88(1):1-13. [doi: 10.1037/ccp0000459] [Medline: 31697093]
41. Torous J, Nicholas J, Larsen ME, Firth J, Christensen H. Clinical review of user engagement with mental health smartphone apps: evidence, theory and improvements. Evid Based Mental Health 2018 Aug;21(3):116-119. [doi: 10.1136/eb-2018-102891]
42. Achilles MR, Anderson M, Li SH, Subotic-Kerry M, Parker B, O'Dea B. Adherence to e-mental health among youth: considerations for intervention development and research design. Digit HEALTH 2020;6:2055207620926064. [doi: 10.1177/2055207620926064] [Medline: 32547775]
43. Domhardt M, Geßlein H, von Rezori RE, Baumeister H. Internet- and mobile-based interventions for anxiety disorders: a meta-analytic review of intervention components. Depress Anxiety 2019 Mar;36(3):213-224. [doi: 10.1002/da.22860] [Medline: 30450811]
44. Zarski AC, Lehr D, Berking M, Riper H, Cuijpers P, Ebert DD. Adherence to internet-based mobile-supported stress management: a pooled analysis of individual participant data from three randomized controlled trials. J Med Internet Res 2016 Jun 29;18(6):e146. [doi: 10.2196/jmir.4493] [Medline: 27357528]
45. Cantisano LM, Gonzalez-Soltero R, Blanco-Fernández A, Belando-Pedreño N. ePSICONUT: an e-health programme to improve emotional health and lifestyle in university students. Int J Environ Res Public Health 2022 Jul 28;19(15):9253. [doi: 10.3390/ijerph19159253] [Medline: 35954601]
46. Mamukashvili-Delau M, Koburger N, Dietrich S, Rummel-Kluge C. Efficacy of computer- and/or internet-based cognitive-behavioral guided self-management for depression in adults: a systematic review and meta-analysis of randomized controlled trials. BMC Psychiatry 2022 Nov 24;22(1):730. [doi: 10.1186/s12888-022-04325-z] [Medline: 36424570]

XSL•FO
RenderX

47. de Vroege L, van den Broek A. Post-pandemic self-reported mental health of mental healthcare professionals in the Netherlands compared to during the pandemic—an online longitudinal follow-up study. Front Public Health 2023;11:1221427. [doi: 10.3389/fpubh.2023.1221427] [Medline: 37465168]
48. Leung C, Pei J, Hudec K, Shams F, Munthali R, Vigo D. The effects of nonclinician guidance on effectiveness and process outcomes in digital mental health interventions: systematic review and meta-analysis. J Med Internet Res 2022 Jun 15;24(6):e36004. [doi: 10.2196/36004] [Medline: 35511463]
49. Erku D, Khatri R, Endalamaw A, et al. Digital health interventions to improve access to and quality of primary health care services: a scoping review. Int J Environ Res Public Health 2023 Sep 28;20(19):6854. [doi: 10.3390/ijerph20196854] [Medline: 37835125]
50. Beck AT. Cognitive Therapy and the Emotional Disorders: International Universities Press; 1976.
51. Cuijpers P, Auerbach RP, Benjet C, et al. The World Health Organization World Mental Health International College Student initiative: An overview. Int J Methods Psychiatr Res 2019 Jun;28(2):e1761. [doi: 10.1002/mpr.1761] [Medline: 30614123]
52. Julious SA. Sample size of 12 per group rule of thumb for a pilot study. Pharm Stat 2005 Oct;4(4):287-291. [doi: 10.1002/pst.185]
53. Teare MD, Dimairo M, Shephard N, Hayman A, Whitehead A, Walters SJ. Sample size requirements to estimate key design parameters from external pilot randomised controlled trials: a simulation study. Trials 2014 Jul 3;15:264. [doi: 10.1186/1745-6215-15-264] [Medline: 24993581]
54. Rahmadiana M, Karyotaki E, Schulte M, et al. Transdiagnostic internet intervention for Indonesian university students with depression and anxiety: evaluation of feasibility and acceptability. JMIR Ment Health 2021 Mar 5;8(3):e20036. [doi: 10.2196/20036] [Medline: 33666553]
55. Amanvermez Y, Karyotaki E, Cuijpers P, et al. Feasibility and acceptability of a guided internet-based stress management intervention for university students with high levels of stress: protocol for an open trial. Internet Interv 2021 Apr;24:100369. [doi: 10.1016/j.invent.2021.100369] [Medline: 33614413]
56. Faul F, Erdfelder E, Lang AG, Buchner A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods 2007 May;39(2):175-191. [doi: 10.3758/bf03193146] [Medline: 17695343]
57. Hentati A, Forsell E, Ljótsson B, Kaldo V, Lindefors N, Kraepelien M. The effect of user interface on treatment engagement in a self-guided digital problem-solving intervention: a randomized controlled trial. Internet Interv 2021 Dec;26:100448. [doi: 10.1016/j.invent.2021.100448]
58. Larsen DL, Attkisson CC, Hargreaves WA, Nguyen TD. Assessment of client/patient satisfaction: development of a general scale. Eval Program Plann 1979 Jan;2(3):197-207. [doi: 10.1016/0149-7189(79)90094-6]
59. Boß L, Lehr D, Reis D, et al. Reliability and validity of assessing user satisfaction with web-based health interventions. J Med Internet Res 2016 Aug 31;18(8):e234. [doi: 10.2196/jmir.5952] [Medline: 27582341]
60. Brooke J. SUS—a quick and dirty usability scale. ResearchGate. URL: https://www.researchgate.net/publication/319394819_SUS_--_a_quick_and_dirty_usability_scale [accessed 2025-09-30]
61. Lewis JR. The system usability scale: past, present, and future. Int J Hum Comput Interact 2018 Jul 3;34(7):577-590. [doi: 10.1080/10447318.2018.1455307]
62. Donkin L, Christensen H, Naismith SL, Neal B, Hickie IB, Glozier N. A systematic review of the impact of adherence on the effectiveness of e-therapies. J Med Internet Res 2011 Aug 5;13(3):e52. [doi: 10.2196/jmir.1772] [Medline: 21821503]
63. Steel P. Arousal, avoidant and decisional procrastinators: do they exist? Pers Individ Dif 2010 Jun;48(8):926-934. [doi: 10.1016/j.paid.2010.02.025]
64. Steel PDG. The Measurement and Nature of Procrastination: University of Minnesota; 2002.
65. Shaw A, Zhang JJ. A Rasch analysis of the irrational procrastination scale (IPS). Front Psychol 2020;11:615341. [doi: 10.3389/fpsyg.2020.615341] [Medline: 33510693]
66. Kroenke K, Spitzer RL, Williams JBW. The PHQ-9. J Gen Intern Med 2001 Sep;16(9):606-613. [doi: 10.1046/j.1525-1497.2001.016009606.x]
67. Wittkampf KA, Naeije L, Schene AH, Huyser J, van Weert HC. Diagnostic accuracy of the mood module of the Patient Health Questionnaire: a systematic review. Gen Hosp Psychiatry 2007;29(5):388-395. [doi: 10.1016/j.genhosppsych.2007.06.004] [Medline: 17888804]
68. Cohen S. Perceived stress in a probability sample of the United States. In: Spacapan S, Oskamp S, editors. The Social Psychology of Health: Sage Publications, Inc; 1988:31-67 URL: https://psycnet.apa.org/record/1988-98838-002 [accessed 2025-10-09]
69. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. J Health Soc Behav 1983 Dec;24(4):385-396. [Medline: 6668417]
70. Lee EH. Review of the psychometric evidence of the perceived stress scale. Asian Nurs Res (Korean Soc Nurs Sci) 2012 Dec;6(4):121-127. [doi: 10.1016/j.anr.2012.08.004] [Medline: 25031113]
71. van Krugten FCW, Busschbach JJV, Versteegh MM, Hakkaart-van Roijen L, Brouwer WBF. The Mental Health Quality of Life Questionnaire (MHQoL): development and first psychometric evaluation of a new measure to assess quality of life

XSL•FO
RenderX

in people with mental health problems. Qual Life Res 2022 Feb;31(2):633-643. [doi: 10.1007/s11136-021-02935-w] [Medline: 34241821]
72. Gómez Penedo JM, Berger T, Grosse Holtforth M, et al. The Working Alliance Inventory for guided Internet interventions (WAI-I). J Clin Psychol 2020 Jun;76(6):973-986. [doi: 10.1002/jclp.22823] [Medline: 31240727]
73. Cohen J. Statistical Power Analysis for the Behavioral Sciences, 2nd edition: Routledge; 2013. [doi: 10.4324/9780203771587]
74. Amanvermez Y, Rahmadiana M, Karyotaki E, et al. Stress management interventions for college students: a systematic review and meta-analysis. Clin Psychol: Sci Pract 2023;30(4):423-444. [doi: 10.1111/cpsp.12342]
75. Kählke F, Berger T, Schulz A, et al. Efficacy of an unguided internet‐based self‐help intervention for social anxiety disorder in university students: a randomized controlled trial. Int J Methods Psych Res 2019 Jun;28(2):e1766. [doi: 10.1002/mpr.1766]
76. Levin ME, Pistorello J, Seeley JR, Hayes SC. Feasibility of a prototype web-based acceptance and commitment therapy prevention program for college students. J Am Coll Health 2014 Jan 2;62(1):20-30. [doi: 10.1080/07448481.2013.843533]
77. Papadatou-Pastou M, Campbell-Thompson L, Barley E, et al. Exploring the feasibility and acceptability of the contents, design, and functionalities of an online intervention promoting mental health, wellbeing, and study skills in higher education students. Int J Ment Health Syst 2019;13:51. [doi: 10.1186/s13033-019-0308-5] [Medline: 31367229]
78. Bangor A, Kortum PT, Miller JT. An empirical evaluation of the system usability scale. Int J Hum Comput Interact 2008 Jul 29;24(6):574-594. [doi: 10.1080/10447310802205776]
79. Harrer M, Adam SH, Baumeister H, et al. Internet interventions for mental health in university students: a systematic review and meta-analysis. Int J Methods Psychiatr Res 2019 Jun;28(2):e1759. [doi: 10.1002/mpr.1759] [Medline: 30585363]
80. Howard R, Berry K, Haddock G. Therapeutic alliance in psychological therapy for posttraumatic stress disorder: a systematic review and meta‐analysis. Clin Psychology and Psychoth 2022 Mar;29(2):373-399. [doi: 10.1002/cpp.2642]
81. Schlicker S, Baumeister H, Buntrock C, et al. A Web- and mobile-based intervention for comorbid, recurrent depression in patients with chronic back pain on sick leave (Get.Back): pilot randomized controlled trial on feasibility, user satisfaction, and effectiveness. JMIR Ment Health 2020 Apr 15;7(4):e16398. [doi: 10.2196/16398] [Medline: 32293577]
82. Hoger onderwijs; eerstejaarsstudenten, vooropleiding en opleidingsfase (dataset 85422NED). CBS StatLine. 2025. URL: https://opendata.cbs.nl/statline/#/CBS/nl/dataset/85422NED/table?ts=1759250749804 [accessed 2025-9-30]
83. Hoger onderwijs; internationale studenten, onderwijskenmerken (dataset 85125NED). CBS StatLine. 2025. URL: https://opendata.cbs.nl/statline/#/CBS/nl/dataset/85125NED/table?ts=1759250787800 [accessed 2025-09-30]
84. Vidourek RA, King KA, Nabors LA, Merianos AL. Students' benefits and barriers to mental health help-seeking. Health Psychol Behav Med 2014 Jan 1;2(1):1009-1022. [doi: 10.1080/21642850.2014.963586] [Medline: 25750831]
85. Nunan D, Aronson J, Bankhead C. Catalogue of bias: attrition bias. BMJ Evid Based Med 2018 Feb;23(1):21-22. [doi: 10.1136/ebmed-2017-110883] [Medline: 29367321]
86. Booker CL, Harding S, Benzeval M. A systematic review of the effect of retention methods in population-based cohort studies. BMC Public Health 2011 Apr 19;11:249. [doi: 10.1186/1471-2458-11-249] [Medline: 21504610]
87. Brueton VC, Tierney JF, Stenning S, et al. Strategies to improve retention in randomised trials: a Cochrane systematic review and meta-analysis. BMJ Open 2014 Feb 4;4(2):e003821. [doi: 10.1136/bmjopen-2013-003821] [Medline: 24496696]

Abbreviations

CBT: cognitive behavioral therapy
CSQ-8: 8-item Client Satisfaction Questionnaire
IPS: Irrational Procrastination Scale
MHQoL: Mental Health Quality of Life Questionnaire
PHQ-9: 9-item Patient Health Questionnaire
PSS-10: 10-item Perceived Stress Scale
RQ: research question
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
SUS-10: 10-item System Usability Scale
WAI-I: Working Alliance Inventory for Guided Internet Interventions
WMH-ICS: World Mental Health International College Student Initiative

XSL•FO
RenderX

Edited by J Sarvestan; submitted 20.02.25; peer-reviewed by M Adnan, X Zheng; revised version received 12.08.25; accepted 13.08.25; published 16.10.25.

Please cite as:
Ozmen S, Amarnath A, de Wit L, van Klaveren C, Cuijpers P, van Straten A, Struijs S, Caring Universities Consortium Feasibility of a Guided Web-Based Procrastination Intervention for College Students: Open Trial
JMIR Form Res 2025;9:e72896
URL: https://formative.jmir.org/2025/1/e72896
doi:10.2196/72896

© Sevin Ozmen, Arpana Amarnath, Leonore de Wit, Chris van Klaveren, Pim Cuijpers, Annemieke van Straten, Sascha Struijs, Caring Universities Consortium. Originally published in JMIR Formative Research (https://formative.jmir.org), 16.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

A Machine Learning–Based Scoring System to Identify High Immunoactivity Microsatellite Stability Tumors by Quantifying Similarity to Microsatellite Instability-High Tumors in Colorectal Cancers: Development and Quantitative Study

Hongkai Yan[1,2,3*], MD; Li Jiang[4*], Dr rer nat; Yaqi Li[1,3], MD, PhD; Fengchong Wang[5], MSc; Shaobo Mo[1,3], MD, PhD; Weiqi Sheng[1,6,7], MD, PhD; Dan Huang[1,6,7*], MD, PhD; Junjie Peng[1,3*], MD, PhD

[1]Department of Oncology, Shanghai Medical College, Fudan University, Shanghai, China

[2]Department of Pediatric Cardiology, Xinhua Hospital Affiliated to Shanghai Jiao Tong University School of Medicine, Shanghai, China

[3]Department of Colorectal Surgery, Fudan University Shanghai Cancer Center, 270 Dong'An Road, Shanghai, China

[4]Weifang Key Laboratory of Collaborative Innovation of Intelligent Diagnosis and Treatment and Molecular Diseases, School of Basic Medical Sciences, Shandong Second Medical University, Weifang, China

[5]Weifang Ten Nanometer Biotechnology Co., Ltd., Weifang, China

[6]Department of Pathology, Fudan University Shanghai Cancer Center, Shanghai, China

[7]Institute of Pathology, Fudan University, Shanghai, China

[*]these authors contributed equally

Corresponding Author:
Junjie Peng, MD, PhD
Department of Oncology, Shanghai Medical College, Fudan University, Shanghai, China

Abstract

Background: Microsatellite stability (MSS) colorectal cancers (CRCs) have a limited response to immune checkpoint inhibitors (ICIs) compared to microsatellite instability-high (MSI-H) CRCs. Nevertheless, previous studies have shown that some MSS CRCs are sensitive to ICIs, although established criteria for treatment justification are still lacking.

Objective: This study aimed to test the tumor-infiltrating lymphocyte (TIL) features of MSS and develop a novel computational tool for the similarity prediction between MSS and MSI-H status in patients with CRC based on multiple factors.

Methods: We collected and analyzed data from 188 patients with CRC, including MSI status, immune cell distributions, clinical features, and gene mutations, using statistical methods and Cox regression. An ensemble machine learning–based MSI-H score was developed using stacked extreme gradient boosting classifiers to quantify the similarity of patient data to MSI-H data based on immune cell distributions, clinical features, and gene mutations. The model was robust and could address missing input data for immune cell distributions and gene mutations.

Results: The scorer performed well (mean Cohen κ of 0.40, SD 0.05, over 10 random seeds) in identifying MSI-H–like MSS samples with TIL distributions similar to genuine MSI-H CRCs. No significant difference was observed between the TIL features of MSI-H–like MSS CRCs and MSI-H CRCs. The disparity between MSI-H–like MSS CRCs and MSS CRCs potentially lies in the T regulatory cells (P=.09) and macrophage (P=.16) populations within the tumor stromal region.

Conclusions: Some patients with MSS CRC presented similar immune cell distributions with high immunoactivity compared to patients with MSI-H CRC. The MSI-H score serves as a metric to quantify the similarity of MSS CRCs to MSI-H CRCs and presents a promising avenue for more personalized and effective cancer immunotherapy treatment, offering a clinical reference for potential ICI targets in MSS CRCs.

(JMIR Form Res 2025;9:e66960) doi:10.2196/66960

KEYWORDS

colorectal tumor; machine learning; ML; immunotherapy; immunoactivity; tumors; mutation; genetics; cancer; colorectal; colorectal cancer; microsatellite stability; immune cell; gene mutation; Cox regression; regression analysis

XSL•FO
RenderX

Introduction

Colorectal cancer (CRC) is one of the leading causes of cancer death worldwide [1]. Microsatellite status divides CRCs into two subtypes: (1) deficient mismatch repair or microsatellite instability-high (MSI-H) tumors and (2) proficient mismatch repair or microsatellite stability (MSS) and microsatellite instability-low tumors [2]. These 2 subtypes are distinct in terms of clinicopathological factors, gene mutations, and the immune microenvironment [2,3]. One of the pivotal treatment modalities in the field of CRC is immunotherapy, especially the administration of immune checkpoint inhibitors (ICIs), including antiprogrammed cell death-1 (PD-1) and antiprogrammed cell death ligand 1 (PD-L1) antibodies [4]. A reliable predictor of immunotherapy response and immunoactivity is MSI-H status; notably, the Food and Drug Administration and the European Medicines Agency granted approval for the use of ICIs to treat MSI-H CRC in 2017 and 2021, respectively [5-7].

While ICIs offer an alternative to surgery and chemotherapy for MSI-H CRCs, the use of ICIs to treat MSS CRCs still lacks justification, with no guidelines for identifying high immunoactivity MSS CRCs; thus, a large population of patients with MSS CRC lacks effective treatment options [8]. However, MSS status is not an absolute marker for excluding immunotherapy. Part of MSS CRCs showed response to ICI therapies [9]. A meta-analysis provides evidence for the application of ICI therapies in nonmetastatic MSS CRCs and highlights its safety and the potential for organ preservation with this approach [10]. In addition, Motta et al [11] demonstrated that some MSS CRCs (up to 20%) harbor a similar profile, including immunological, genetic, pathological, and clinical characteristics, to MSI-H tumors. Therefore, identifying MSS CRCs with similar profiles to MSI-H CRCs could be a reasonable approach, and a strategy for achieving this is urgently needed. Tumor-infiltrating lymphocytes (TILs), a polymorphic group consisting primarily of effector T lymphocytes, regulatory T lymphocytes, natural killer cells, dendritic cells, and macrophages, are a critical feature of CRC immunology [12]. TILs are useful in immunotherapy and immunoactivity prediction [13]. Notably, the intratumoral spatial heterogeneity of TILs is an important factor for precisely stratifying prognostic immune subgroups of MSI-H CRC [14].

In this study, we developed a novel MSI-H score based on ensemble machine learning to quantify the degree of similarity of immunoactivity between patients with MSS CRC and patients with MSI-H CRC. A subgroup of patients with MSS CRC with high MSI-H scores was defined as patients with MSI-H–like MSS CRC, exhibiting MSI-H–like features in immune cell distributions, gene mutations, pathological reports, and clinical characteristics. This work paves the way for more personalized, accurate, and effective cancer immunotherapy treatments, delivering a clinical reference for identifying potential ICI targets and advancing patient care.

Methods

Recruitment

Data from 188 patients with stage II CRC and tissue samples were collected from the institutional database of the Fudan University Shanghai Cancer Center between 2013 and 2019. The American Joint Committee on Cancer staging system was used to determine each patient's stage [15]. Tested by next-generation sequencing, 24 patients were classified as MSI-H. None of the patients had radiation therapy, chemotherapy, or immunotherapy before tumor resection. Clinical and pathological data were obtained from patient records and postoperative pathology reports.

Multiplex Immunohistochemistry Staining

Sections (4 mm thick) were cut from formalin-fixed, paraffin-embedded CRC tissue and control tonsil tissue for multiplex immunohistochemistry (mIHC). The slides were dewaxed in xylene, rehydrated, and rinsed in graded ethanol solutions and tap water. Antibody diluent/block (72424205; PerkinElmer) was applied to block endogenous peroxidase. The slides were boiled in a Tris-EDTA buffer (pH: 9; 643901; Klinipath) and underwent microwave treatment (MWT) for antigen retrieval. Information on the primary antibodies and the corresponding fluorophores is provided in Table S1 in Multimedia Appendix 1, including 2 panels (Figure 1). One antigen required 1 round of labeling, including primary antibody incubation, secondary antibody incubation, and tyramide signal amplification (TSA) visualization, followed by labeling of the subsequent antibody. After incubation with the primary antibody for 1 hour at room temperature, the slides were incubated with Opal Polymer HRP Ms+Rb (2414515; PerkinElmer) at 37 ˚C for 10 minutes. TSA visualization was performed with the Opal 7-Color IHC Kit (NEL797B001KT; PerkinElmer) containing the fluorophores 4,6-diamidino-2-phenylindole (DAPI; Thermo Scientific) and the TSA Coumarin system (NEL703001KT; PerkinElmer). MWT was performed to remove the antibody-TSA complex with the Tris-EDTA buffer (pH: 9). TSA single-stained slides were finished with MWT, counterstained with DAPI for 5 minutes, and enclosed in Antifade mounting medium (I0052; NobleRyder).

XSL•FO
RenderX

Figure 1. Two panels of multiplex immunohistochemistry (mIHC). (A) Representative hematoxylin and eosin (HE) and mIHC staining images of panel 1: the upper line of images represents HE staining and the staining of CD8, CD45RO, and CD3; the lower line of images represents programmed cell death-1 (PD-1) staining, programmed cell death ligand 1 (PD-L1) staining, and the merge image. (B) Representative HE and mIHC staining images of panel 2: the upper line of images represents HE staining and the staining of CD4, FOXP3, and CD68; the lower line of images represents CD163 staining, PD-L1 staining, and the merge image.

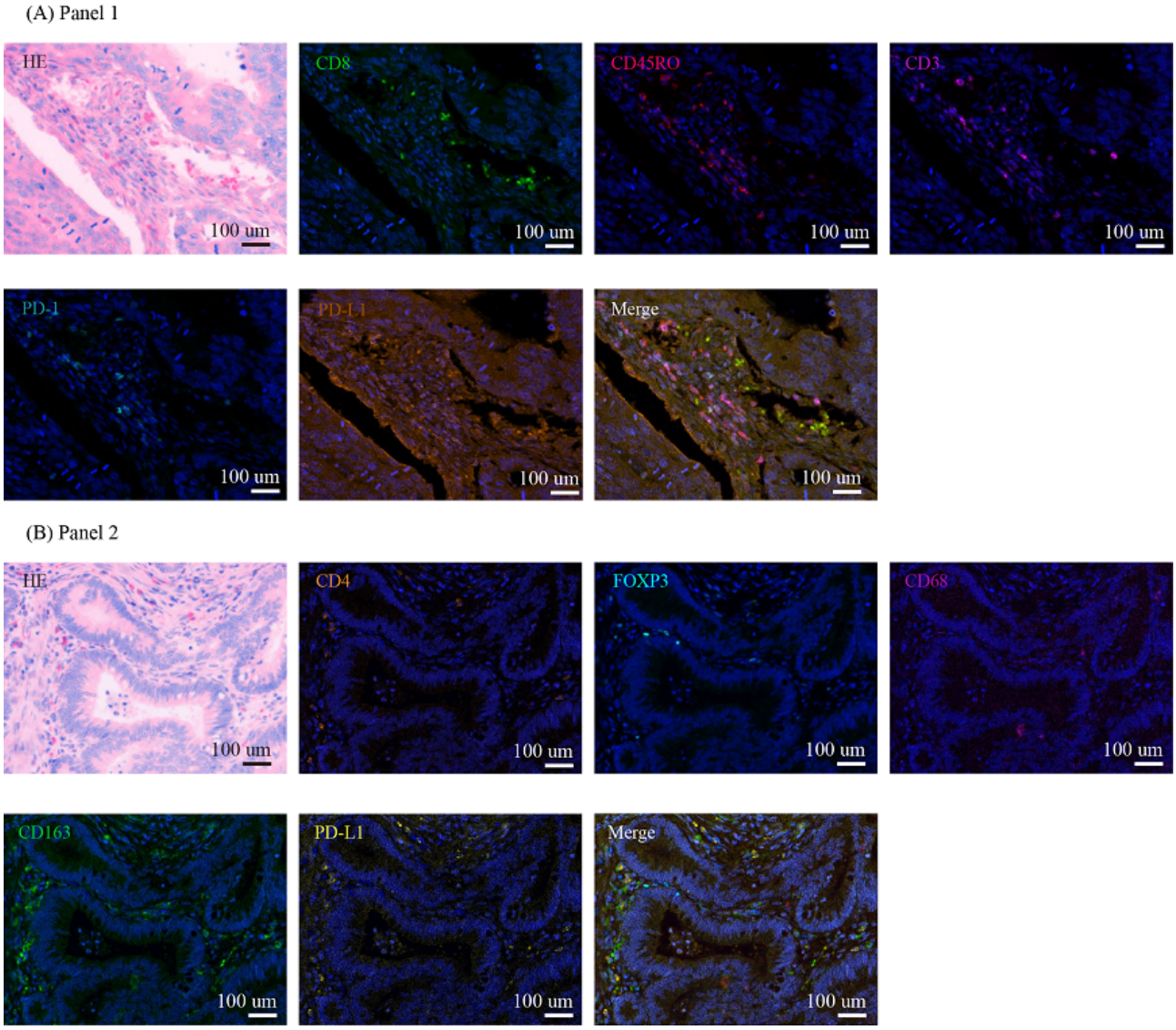

Image Acquisition and Analysis

Multiplexed and single-color control slides were scanned at an absolute magnification of 200× by the PerkinElmer Vectra automated multispectral microscope. Representative fields from the single-color slides were imaged, and a spectral library for unmixing was generated by inForm image analysis software (version 2.1; PerkinElmer). Index cases were stained using the multiplex method and then imaged. Channels were unmixed using the spectral library. All settings were saved within an algorithm to allow for batch analysis of multiple original multispectral images of the same tissue [16].

Quantification of Immune Cell Densities and Classification

The nuclear morphological features were based on DAPI staining. The numbers of immune cells in each image were scored as percent cellularity (number of positive cells/number of nucleated cells). Five representative fields at 200× magnification of tissue area were selected. The densities of immune cells were segmented independently by 2 pathologists. Immune variables were classified based on the patterns of fluorochrome intensity.

Patient Follow-Up

Patients were monitored every 3 - 6 months for 3 years, then every 6 - 12 months up to 5 years. Follow-ups included rectal exams, carcinoembryonic antigen (CEA) tests, annual radiological studies, and colonoscopies as needed.

Test of MSI and CRC-Relevant Mutations

The ColonCore panel (Burning Rock) is designed for simultaneous detection of MSI status and mutations in 37 CRC-related genes (Table S2 in Multimedia Appendix 1). The MSI detection method was a read-count-distribution–based approach, using the coverage ratio of a specific set of repeat lengths as the main characteristic of each microsatellite locus.

XSL•FO
RenderX

The MSI status of a sample was determined by the percentage of unstable loci in the given sample [17].

Statistical Tests and Survival Analysis

Statistical analysis was performed and visualized by R (version 3.4.3; R Foundation for Statistical Computing), SPSS software (version 25.0; IBM Corp), and GraphPad Prism 7 software (GraphPad Software Inc). All group-wise comparisons were conducted by the 2-sided unpaired Mann-Whitney *U* test, followed by the Bonferroni procedure. The Cox proportional hazards regression model was used to assess the hazard ratios, 95% CIs, and *P* values for univariate and multivariate analysis. Variables with $P<.10$ after adjusting for common clinicopathological parameters were included in the multivariate analysis. Survival times were compared using the log-rank test. A *P* value of <.05 was considered statistically significant, and all *P* values corresponded to 2-sided statistical tests.

Feature Engineering

The process from feature engineering to model evaluation is depicted in Multimedia Appendix 2. Categorical features were one-hot encoded as dummy variables. The mutation landscape was also one-hot encoded based on gene classes, with 2 classification stringencies, using the Database for Annotation, Visualization, and Integrated Discovery (DAVID) gene functional classification tool [18]. To further engineer the mutation landscape, we calculated the joint posterior mutation probability P(MSI|M) with the following equation:

$$P(MSI|M)=\frac{P(MSI)\prod_{gi\in M}P(gi|MSI)}{P(MSI)\prod_{gi\in M}P(gi|MSI)+P(MSS)\prod_{gi\in M}P(gi|MSS)} \quad (1)$$

where P(MSI|M) is the probability that a patient has MSI-high status given their mutation landscape and is based on previous probabilities and the frequencies of mutated genes in MSI-H and MSS populations, gi is a mutated gene in a patient's sample, M=gi is a set of all detected mutated genes in the sample, PMSI is the probability of MSI-H in a CRC sample (0.83), PMSS is the probability of MSS in a CRC sample (0.17), Pgi|MSI is the frequency of a mutated gene gi in a CRC MSI-H population, and Pgi|MSS is the frequency of a mutated gene gi in a CRC MSS population. Pgi|MSI and Pgi|MSS were based on the findings of Serebriiskii et al [19] and 2 datasets [20,21] on cBioPortal [22,23]. Our Bayesian-based metric can explicitly incorporate previous biological knowledge, including MSI-H/MSS prevalence in populations with CRC and microsatellite status–specific mutation frequencies, and probabilistic reasoning into the modeling process. Leveraging these priors potentially enhances the model's ability to distinguish MSI-H from MSS cases. This metric was included in the dataset along with the other features and used for model training.

Though no missing data were presented in the dataset, our model can handle missing input from users because we trained several models with varied complexities, as elaborated in the following section.

Model Training and Deployment

Sample microsatellite status was one-hot labeled (MSI-H as 1 and MSS as 0). Multiple extreme gradient boosting (XGBoost) models [24] were trained to identify MSI-H–like tumors with different combinations of features, that is, patient metainformation, mutational landscape–derived features, and mIHC results, such as PD-L1, CD163, and CD8 mIHC staining results. The predicted likeliness of MSI-H by the models was defined as MSI-H score. Specifically, 44 models, with different combinations of features shown in Table S3 in Multimedia Appendix 1, were trained. To address potential bias caused by class imbalance, the scale_pos_weight parameter was set to the class ratio during model training.

The models were then deployed as a public web interface. As users' data privacy is prioritized, users' input is never stored on our server. Each model ensemble consists of 10 submodels, each trained with a distinct random seed. The final prediction for any user input is the average of the middle 6 submodel outputs, excluding the extremes. One of the XGBoost tree models in pseudocode is shown in Multimedia Appendix 3.

Visualization and Clustering of TILs

To understand the feature importance in an unbiased and holistic way, a massive exploratory XGBoost model with all features was trained. These features include the features patient metainformation, mutational landscape–derived features, and all mIHC results. The massive model can capture the full spectrum of the MSI-H score variation and avoid the potential bias or noise introduced by the feature selection process. The feature importance was then computed using both the XGBoost built-in function and the Shapley additive explanations package [25].

Following classification by this model (threshold=0.3 defining MSI-H, chosen so that predicted MSI-H proportion approximates the epidemiologically documented prevalence of MSI-H CRC), we visualized all mIHC features of all samples by grouped box plots. For each cell type, we compared 4 groups (all MSS vs MSI-H, other MSS vs MSI-H, other MSS vs MSI-H–like MSS, and MSI-H–like MSS vs MSI-H) by 2-sided unpaired Mann-Whitney *U* test and Benjamini-Hochberg adjustment [26]. To cluster cells based on 4 comparisons, we projected each cell type into a 4D latent space using a formula measuring similarity between cell percentages of former and latter populations:

$$\text{Similarity}(\text{former},\text{latter})=\sum_{i=1}^{4}\left(\frac{r_{former}\cdot r_{latter}}{\sqrt{r_{former}^{2}\cdot r_{latter}^{2}}}\cdot\left(1-\left|p_{adj,i}(former)-p_{adj,i}(latter)\right|\right)\right) \quad (2)$$

where r_{former} and r_{latter} represent the percentages of a specific cell type (eg, CD8+ cells) measured in individual samples belonging to the "former" and "latter" groups, respectively. p_{adj} is the adjusted *P* value of a comparison. We then performed hierarchical clustering based on Euclidean distance in latent space with complete linkage [27].

Feature and Model Evaluation

Model generalizability was assessed by training models with identical hyperparameters through stratified 5-fold cross-validation. Cohen κ coefficients were computed on each hold-out fold, and the mean Cohen κ was computed based on the 5 κ's, with greater κ's indicating better model performance. This training and validation process was repeated 10 times with different random stratified splits and model initializations.

XSL•FO
RenderX

Ethical Considerations

Ethics approval was obtained from the Ethics Committee of Fudan University Shanghai Cancer Center, and informed consent was obtained from all participants (1808190 - 12). Neither the patients nor the public were involved in this study (ie, only database tissue samples and data from patient records and postoperative pathology reports were used). All patient data collected for this study were deidentified prior to analysis.

Results

Higher TIL Infiltration in the Stromal Region Than in the Tumor Region

TILs were analyzed using the mIHC method. Significant differences were found between the stromal region and the tumor region (Figure 2). The stromal region showed a higher prevalence of CD3+ T cells ($P<.001$), CD8+ T cells ($P<.001$), memory T cells ($P<.001$), CD8+ memory T cells ($P<.001$), CD3+ PD-1+ T cells ($P<.001$), CD4+ T cells ($P=.048$), regulatory T cells (Tregs; $P<.001$), macrophages ($P=.001$), M1 macrophages ($P=.003$), M2 macrophages ($P=.001$), and PD-L1+ cells ($P=.007$) than the tumor region. However, no significant difference was observed for CD8+ PD-1+ T cells and PD-L1+ macrophages between the stromal region and tumor region.

XSL•FO
RenderX

Figure 2. The difference in tumor-infiltrating lymphocytes between the stromal region and tumor region. Each cell type ratio of relevant samples is shown by a box plot. The cell percentage difference in the stromal region and tumor region was compared, and all *P* values were adjusted with the Bonferroni procedure and are shown on the right side. NS: not significant; PD-1: programmed cell death-1; PD-L1: programmed cell death ligand 1; Treg: regulatory T cell. *$P<.05$; **$P<.01$; ***$P<.001$.

XSL•FO
RenderX

Prognostic Impact of Clinical Characteristics, MSI Status, and Immune Cell Infiltration

Variables (Table 1) commonly collected in clinics and related to prognosis or with a *P* value <.10 in univariate analysis (Tables S4 and S5 in Multimedia Appendix 1) were analyzed using the Cox proportional hazards regression model (Table S6 in Multimedia Appendix 1). Microsatellite status was not linked to overall survival or disease-free survival in multivariate analysis. Significant overall survival predictors included age, CEA, M1 macrophage (CD68+ CD163–) infiltration in stromal region, and PD-1+ T cell (CD3+ PD-1+) infiltration in tumor region. Significant disease-free survival predictors included age, CEA, tumor differentiation, and CD8+ T cell (CD8+) infiltration in tumor region (Table S6 in Multimedia Appendix 1).

XSL•FO
RenderX

Table . Clinical characteristics related to microsatellite instability-high (MSI-H) and microsatellite stability (MSS) status.

Characteristic		Patients, n	MSS tumor (n=164), n (%)	MSI-H tumor (n=24), n (%)	*P* value
Sex					.82
	Male	106	93 (57)	13 (54)	
	Female	82	71 (43)	11 (46)	
Age (y)					.71
	<65	111	96 (59)	15 (63)	
	≥65	77	68 (41)	9 (38)	
Mucinous					.006
	No	152	138 (84)	14 (58)	
	Yes	36	26 (16)	10 (42)	
Differentiation					.002
	Poor	40	29 (18)	11 (50)	
	Moderate to well	140	129 (82)	11 (50)	
T stage					.59
	T3	88	78 (48)	10 (42)	
	T4	100	86 (52)	14 (58)	
Tumor site					<.001
	Right	52	36 (22)	16 (67)	
	Left	52	49 (30)	3 (13)	
	Rectum	83	78 (48)	5 (21)	
Lymphovascular invasion					.43
	No	149	128 (78)	21 (88)	
	Yes	39	36 (22)	3 (13)	
Perineural invasion					.86
	No	136	119 (73)	17 (71)	
	Yes	52	45 (27)	7 (29)	
CEA[a] (ng/ml)					.94
	<5	124	108 (66)	16 (67)	
	≥5	64	56 (34)	8 (33)	
Chemotherapy					.45
	No	76	68 (41)	8 (33)	
	Yes	112	96 (59)	16 (67)	
Radiotherapy					>.99
	No	163	142 (92)	21 (91)	
	Yes	15	13 (8)	2 (9)	

[a]CEA: carcinoembryonic antigen.

TIL Distribution in MSI-H CRCs, All MSS CRCs, MSI-H–Like MSS CRCs, and Other MSS CRCs

The mIHC experiment was performed to examine the TILs in all CRCs (Figure 3). MSI-H CRCs exhibited significantly higher infiltration of TILs compared to MSS CRCs in both the tumor region and stromal region. Specifically, MSI-H CRCs had a more abundant presence of PD-L1+ M2 macrophages (*P*=.001), CD163+ cells (*P*=.001), PD-L1+ macrophages (*P*=.01), M2 macrophages (*P*=.001), and macrophages (*P*=.03) in the stromal region, as well as M2 macrophages in the tumor region (*P*=.02).

XSL•FO
RenderX

Figure 3. Tumor-infiltrating lymphocyte (TIL) distributions in microsatellite instability-high (MSI-H), all microsatellite stability (MSS), MSI-H–like MSS, and other MSS colorectal cancers (CRCs). The box plot on the right displays the percentages of a cell type in relevant samples for each class. Cell percentage differences were compared, and *P* values were adjusted and are presented in the heat map (*P*<.001 when adjusted *P* values were <.001 due to rounding), along with the median comparison result between the 2 populations (+: former population median>latter population median; -: former population median<latter population median). FDR: false discovery rate; mIHC: multiplex immunohistochemistry; PD-1: antiprogrammed cell death-1; PD-L1: antiprogrammed cell death ligand 1; Treg: regulatory T cell.

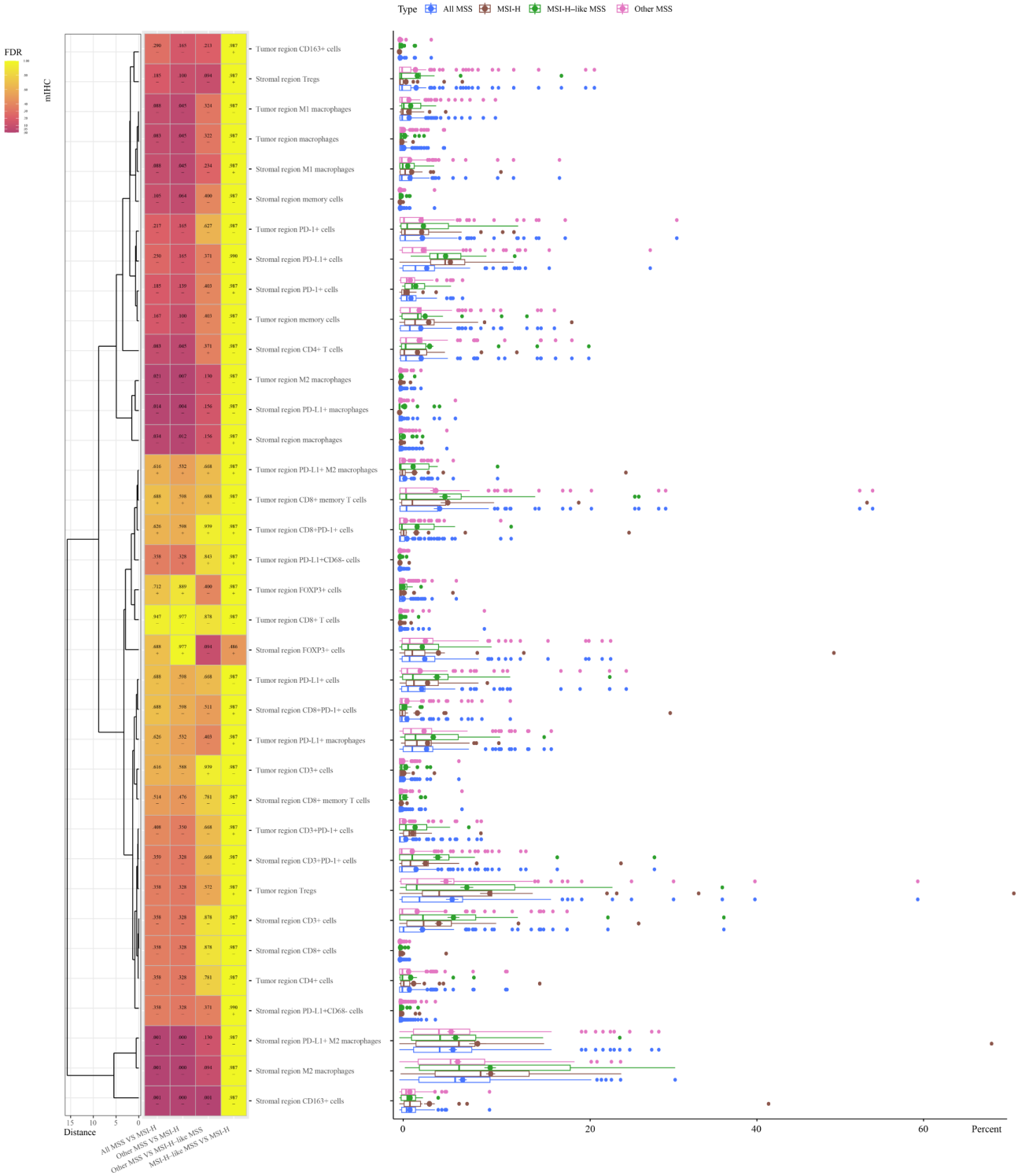

MSI-H–like MSS CRCs exhibited TIL infiltration patterns akin to MSI-H CRCs but distinct from other MSS CRCs. No significant difference was observed between MSI-H–like MSS and genuine MSI-H CRCs (the fourth column in Figure 3). Compared to other MSS CRCs, MSI-H–like MSS CRCs showed higher infiltration of CD163+ cells in the stromal region (*P*=.001; the box plot and the third column in Figure 3) and potentially increased levels of PD-L1+ M2 macrophages (*P*=.13), FOXP3+ cells (*P*=.09), Tregs (*P*=.09), PD-L1+ macrophages (*P*=.16), M2 macrophages (*P*=.09), and macrophages (*P*=.16) in the stromal region, as well as M2 macrophages (*P*=.13) in the tumor region. The distinct infiltration patterns of TILs indicate that heightened presence

XSL•FO
RenderX

of macrophages and Tregs are key factors in distinguishing MSI-H–like MSS CRCs from MSS CRCs.

Macrophages were also found to be significantly more abundant in genuine MSI-H CRCs than in other MSS CRCs (the second column in Figure 3). Specifically, in the stromal region, PD-L1+ M2 macrophages (P<.001), CD163+ cells (P<.001), PD-L1+ macrophages (P=.004), M2 macrophages (P<.001), M1 macrophages (P=.045), CD4+ T cells (P=.045), and macrophages (P=.01) were significantly more abundant in MSI-H CRCs than in other MSS CRCs. In the tumor region, M2 macrophages (P=.007), M1 macrophages (P=.045), and macrophages (P=.045) were found to be significantly increased in MSI-H CRCs compared with other MSS CRCs.

The TIL distribution shows that the model performed well. The scorer, which was trained and validated on only 3 types of lymphocytes, classified MSI-H–like MSS CRC samples with similar TIL distributions as MSI-H CRC samples (the fourth column in Figure 3) rather than MSS CRC samples (the third column in Figure 3), despite most features of TILs (15 other lymphocytes) being unknown to the model. In addition, as anticipated, the heat map in Figure 3 (second and third columns) revealed that other MSS CRC samples exhibited a slightly closer TIL distribution to MSI-H–like MSS CRC samples than to MSI-H CRC samples.

The feature importance in a large predictive model is described in Multimedia Appendix 4. TIL features from the whole or stromal region were more predictive than the tumor region alone for the MSI-H status. Top features for MSI-H score predictor included macrophage subtypes, mutational landscape, and immune cell distributions.

MSI-H Score Predictor Generalization Ability Affected by Feature Number and Type

Increasing the number of features generally enhanced $\bar{\kappa}$, indicating better generalization performance (Figure 4). However, models incorporating PD-L1 mIHC staining tended to exhibit lower $\bar{\kappa}$ compared to those without, likely due to noise in PD-L1 measurements, as evidenced by the high SD of $\bar{\kappa}$ for models 2 to 4. This noise effect was mitigated by increasing model complexity; for example, model 44 had a greater $\bar{\kappa}$ than model 41, despite including PD-L1. Feature importance analysis (Figure 5) revealed that while PD-L1 remained relevant, its importance diminished as models became more complex, suggesting that sophisticated models learned to filter out noise and extract useful information from PD-L1. The variable Spearman correlation matrix heat map is shown in Multimedia Appendix 5. In addition, an MSI-H scorer web interface is freely accessible [28].

Figure 4. Box plot of mean Cohen κ values evaluated through 5-fold cross-validation repeated over 10 random seeds for each model. In general, as model complexity increases, the model's ability to generalize tends to improve, as indicated by higher Cohen κ values.

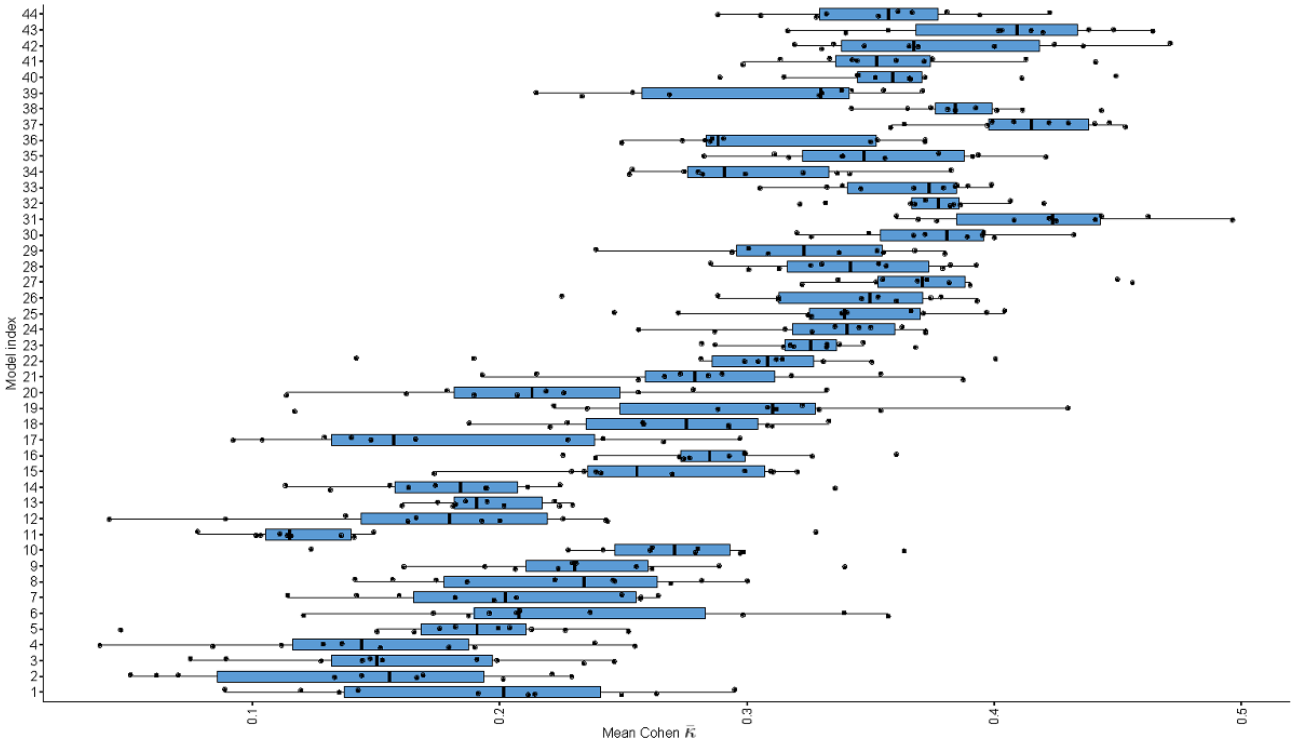

XSL•FO
RenderX

Figure 5. Bubble chart of feature importance. Each bubble represents a feature (y-axis) and its importance in the model (x-axis), as computed by using the native "gain" importance metric (built-in) from extreme gradient boosting (XGBoost; a darker bubble color indicates higher importance) and Shapley additive explanations (SHAP; a larger bubble area indicates higher importance). Mutation landscapes and tumor sites consistently have relatively dark, large bubbles, indicating their importance in the model. CEA: carcinoembryonic antigen.

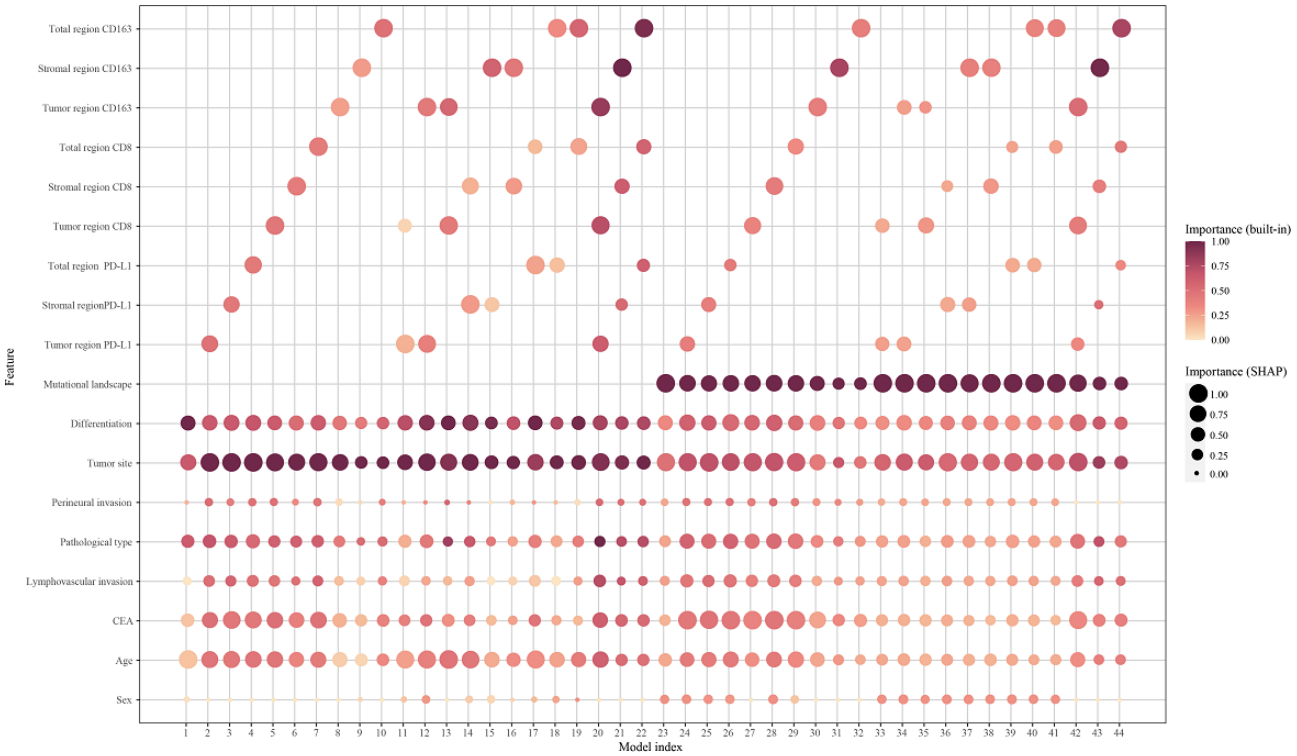

Discussion

Principal Findings

Immunotherapy has been successful for treating MSI-H CRCs but is not as effective in MSS CRCs, which comprise the majority of CRCs. Thus, we developed a machine learning–based MSI-H predictor to generate a robust and reliable score that can capture the complexity and heterogeneity of CRC and better target patients with MSS CRC who may benefit from immunotherapy. Our study also provides insights into the immune landscape of CRC and the role of immune cell distributions, clinical features, and gene mutations in influencing MSI status. This CRC prognostic study mostly agrees with our previous research [29] and with findings from other authors [30]. For example, according to our results, TIL infiltration, primarily by macrophages or CD163+ cells, was significantly higher in MSI-H CRCs than in MSS CRCs (Figure 3), consistent with previous studies [31].

We observed a higher abundance of TIL subsets in the stromal region than in the tumor region, indicating a more active immune response in the stromal region (Figure 2). Comparing the MSI predictive performance of models 42, 43, and 44 (Figure 4) also highlights the importance of stromal TILs. Regional disparities underscore the importance of analyzing the complete tumor region for comprehensive insights. Our scorer successfully identified MSI-H–like MSS samples with TIL distributions similar to genuine MSI-H CRCs (Figure 3). In addition, the balance of proinflammatory and anti-inflammatory scale is an important feature for immunological characters. Macrophages can be classified into 2 main subtypes: M1 macrophages with proinflammatory and antitumor functions and M2 macrophages with anti-inflammatory and protumor functions. The ratio of M1/M2 macrophages may influence immunotherapy outcomes, reflecting the balance between proinflammatory and anti-inflammatory signals in the tumor microenvironment [32]. Tregs are frequently known to be immunosuppressive and can predict both the host immune response and chemotherapeutic response [33]. Both macrophages and Tregs are important in the regulation of immunoactivity. As is shown in our results, the distribution of macrophages and Tregs appears to be important in differentiating MSI-H–like MSS CRCs from other MSS CRCs based on TIL infiltration patterns (Figure 3). By comparing κ^{-} variations within model sets and between specific model set pairs (Figure 4)—including model 5/6/7 versus 13/16/19, 27/28/29 versus 35/38/41, 2/3/4 versus 12/15/18, 24/25/26 versus 34/27/40, 11/14/17 versus 20/21/22, and 33/36/39 versus 42/43/44—we observed that CD163 mIHC result increased the predictive value for whole-tumor MSI scores but reduced it for tumor region scores. To better understand the differences in M2 macrophages and Tregs between MSS CRCs and MSI-H CRCs, further research on their function in CRCs is necessary.

Our analysis revealed that PD-L1+ M2 macrophages in the total region, mutational landscape, CD163+ cells in the stromal region, PD-L1+ M2 macrophages in the stromal region, and tumor site were the most important features for predicting MSI-H status (Multimedia Appendix 4), aligning with other research results [2,3]. Macrophages can express PD-L1 and interact with PD-1+ T cells, which may affect the response to immunotherapy [34]. However, PD-L1+ macrophages potentially indicate M1-like polarization profiles [35]. The

XSL•FO
RenderX

stroma is important because it can influence the extracellular matrix formation, angiogenesis, immune response, and therapeutic resistance of tumors [3]. The importance of the mutational landscape for prediction is widely known [36]. We did not find an obvious immunological explanation for why tumor site would impact the similarity between MSI-H and MSS. Further study is needed to clarify the underlying mechanisms. Moreover, we observed that feature number and type influenced the generalization ability of the MSI-H score prediction models (Figure 4 and Figure 5). This suggests that the omission of diverse variables requires specific computational models, and our machine learning scorer is adept at incorporating all such considerations, thereby highlighting our advantage.

On the basis of our results, we proposed a hypothesis regarding the changes that occur in MSI-H–like MSS CRCs compared to other MSS CRCs. MSI-H–like MSS CRCs foster an immunosuppressive microenvironment with M2 macrophages, Tregs, and PD-L1 that inhibits T cell responses [37]. However, there are enough T cells present that can be reactivated upon PD-1/PD-L1 blockade, leading to the sensitivity of MSI-H–like MSS CRCs to ICIs. The abundance of macrophages suggests that there may be some M1-like populations that, when disinhibited, promote antitumor immunity. Detailed differences in immune cell populations and their functions in MSI-H–like MSS CRCs and other MSS CRCs should be further investigated to understand the mechanisms underlying the differential response to immunotherapy. Furthermore, future clinical trials could be conducted to evaluate ICI treatment between patients with MSI-H–like MSS CRC and other patients with MSS CRC with low MSI-H scores.

Limitations

Limitations of our study include the lack of internal or external validation of the MSI-H score in patients with MSS CRC receiving immunotherapy and the absence of investigation into the underlying molecular mechanisms. Further research and clinical trials are needed to validate our MSI-H score and elucidate the associated mechanisms.

Conclusions

In conclusion, our study revealed significant variations in TIL distribution across tumor regions and MSI status. Integrating clinical, TIL, and mutational data, we developed a robust MSI-H scorer that captures CRC's complexity and heterogeneity. Macrophages, gene mutations, and tumor site emerged as key predictors. MSI-H–like MSS CRCs exhibited TIL infiltration patterns with high immunoactivity similar to MSI-H CRCs, distinctly different from other MSS CRCs. Our privacy-protected MSI-H score predictor is freely available on the web, enabling clinical and research applications.

Acknowledgments

This study was supported by grants from the National Natural Science Foundation of China (82473500 to JP and 82372974 to YL) and the Natural Science Foundation of Shandong Province (ZR2023QC282 to LJ). The sponsors had no role in the study design; data collection, analysis, or interpretation; writing the report; or the decision to submit the paper for publication.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

JP and DH contributed to the study conception and design. HY, YL, and LJ developed the methodology. LJ, HY, WS, SM, and FW performed the analysis and interpretation. YL, HY, and LJ drafted and revised the manuscript. JP and DH supervised the study. All authors read and approved the final manuscript.

Conflicts of Interest

FW is employed by Weifang Ten Nanometer Biotechnology Co, Ltd. All other authors declare no other conflicts of interest.

Multimedia Appendix 1
Detailed information on the experimental protocols, genetic panels, model specifications, and statistical analyses performed in this study, as provided by 6 supplementary tables.
[DOC File, 249 KB - formative_v9i1e66960_app1.doc]

Multimedia Appendix 2
The flowchart of feature engineering, model training, deployment, and feature and model evaluation.
[PNG File, 203 KB - formative_v9i1e66960_app2.png]

Multimedia Appendix 3
An example of one of the extreme gradient boosting tree models in pseudocode.
[DOC File, 92 KB - formative_v9i1e66960_app3.doc]

XSL•FO
RenderX

Multimedia Appendix 4
Feature importance in a large predictive model.
[PDF File, 894 KB - formative_v9i1e66960_app4.pdf]

Multimedia Appendix 5
Heat map of Spearman correlation coefficients between all pairs of variables (features and targets).
[PDF File, 534 KB - formative_v9i1e66960_app5.pdf]

References

1. Siegel RL, Wagle NS, Cercek A, Smith RA, Jemal A. Colorectal cancer statistics, 2023. CA Cancer J Clin 2023;73(3):233-254. [doi: 10.3322/caac.21772] [Medline: 36856579]
2. Li K, Luo H, Huang L, Luo H, Zhu X. Microsatellite instability: a review of what the oncologist should know. Cancer Cell Int 2020;20(1):16. [doi: 10.1186/s12935-019-1091-8] [Medline: 31956294]
3. Ganesh K, Stadler ZK, Cercek A, et al. Immunotherapy in colorectal cancer: rationale, challenges and potential. Nat Rev Gastroenterol Hepatol 2019 Jun;16(6):361-375. [doi: 10.1038/s41575-019-0126-x] [Medline: 30886395]
4. Wu X, Gu Z, Chen Y, et al. Application of PD-1 blockade in cancer immunotherapy. Comput Struct Biotechnol J 2019;17:661-674. [doi: 10.1016/j.csbj.2019.03.006] [Medline: 31205619]
5. Golshani G, Zhang Y. Advances in immunotherapy for colorectal cancer: a review. Therap Adv Gastroenterol 2020;13:1756284820917527. [doi: 10.1177/1756284820917527] [Medline: 32536977]
6. Casak SJ, Marcus L, Fashoyin-Aje L, et al. FDA approval summary: pembrolizumab for the first-line treatment of patients with MSI-H/dMMR advanced unresectable or metastatic colorectal carcinoma. Clin Cancer Res 2021 Sep 1;27(17):4680-4684. [doi: 10.1158/1078-0432.CCR-21-0557] [Medline: 33846198]
7. Trullas A, Delgado J, Genazzani A, et al. The EMA assessment of pembrolizumab as monotherapy for the first-line treatment of adult patients with metastatic microsatellite instability-high or mismatch repair deficient colorectal cancer. ESMO Open 2021 Jun;6(3):100145. [doi: 10.1016/j.esmoop.2021.100145] [Medline: 33940347]
8. Wu H, Deng M, Xue D, et al. PD-1/PD-L1 inhibitors for early and middle stage microsatellite high-instability and stable colorectal cancer: a review. Int J Colorectal Dis 2024 May 29;39(1):83. [doi: 10.1007/s00384-024-04654-3] [Medline: 38809459]
9. Guven DC, Kavgaci G, Erul E, et al. The efficacy of immune checkpoint inhibitors in microsatellite stable colorectal cancer: a systematic review. Oncologist 2024 May 3;29(5):e580-e600. [doi: 10.1093/oncolo/oyae013] [Medline: 38309719]
10. Zhang H, Huang J, Xu H, et al. Neoadjuvant immunotherapy for DNA mismatch repair proficient/microsatellite stable non-metastatic rectal cancer: a systematic review and meta-analysis. Front Immunol 2025;16:1523455. [doi: 10.3389/fimmu.2025.1523455]
11. Motta R, Cabezas-Camarero S, Torres-Mattos C, et al. Immunotherapy in microsatellite instability metastatic colorectal cancer: current status and future perspectives. J Clin Transl Res 2021 Aug 26;7(4):511-522. [Medline: 34541365]
12. Mantovani A, Allavena P, Sica A, Balkwill F. Cancer-related inflammation. Nature New Biol 2008 Jul 24;454(7203):436-444. [doi: 10.1038/nature07205] [Medline: 18650914]
13. Brummel K, Eerkens AL, de Bruyn M, Nijman HW. Tumour-infiltrating lymphocytes: from prognosis to treatment selection. Br J Cancer 2023 Feb;128(3):451-458. [doi: 10.1038/s41416-022-02119-4] [Medline: 36564565]
14. Jung M, Lee JA, Yoo SY, Bae JM, Kang GH, Kim JH. Intratumoral spatial heterogeneity of tumor-infiltrating lymphocytes is a significant factor for precisely stratifying prognostic immune subgroups of microsatellite instability-high colorectal carcinomas. Mod Pathol 2022 Dec;35(12):2011-2022. [doi: 10.1038/s41379-022-01137-0] [Medline: 35869301]
15. Weiser MR. AJCC 8th edition: colorectal cancer. Ann Surg Oncol 2018 Jun;25(6):1454-1455. [doi: 10.1245/s10434-018-6462-1] [Medline: 29616422]
16. Gorris MAJ, Halilovic A, Rabold K, et al. Eight-color multiplex immunohistochemistry for simultaneous detection of multiple immune checkpoint molecules within the tumor microenvironment. J Immunol 2018 Jan 1;200(1):347-354. [doi: 10.4049/jimmunol.1701262] [Medline: 29141863]
17. Zhu L, Huang Y, Fang X, et al. A novel and reliable method to detect microsatellite instability in colorectal cancer by next-generation sequencing. J Mol Diagn 2018 Mar;20(2):225-231. [doi: 10.1016/j.jmoldx.2017.11.007] [Medline: 29277635]
18. Huang DW, Sherman BT, Tan Q, et al. The DAVID gene functional classification tool: a novel biological module-centric algorithm to functionally analyze large gene lists. Genome Biol 2007;8(9):R183. [doi: 10.1186/gb-2007-8-9-r183] [Medline: 17784955]
19. Serebriiskii IG, Connelly C, Frampton G, et al. Comprehensive characterization of RAS mutations in colon and rectal cancers in old and young patients. Nat Commun 2019 Aug 19;10(1):3722. [doi: 10.1038/s41467-019-11530-0] [Medline: 31427573]
20. Mondaca S, Walch H, Nandakumar S, Chatila WK, Schultz N, Yaeger R. Specific mutations in APC, but not alterations in DNA damage response, associate with outcomes of patients with metastatic colorectal cancer. Gastroenterology 2020 Nov;159(5):1975-1978. [doi: 10.1053/j.gastro.2020.07.041] [Medline: 32730818]

XSL•FO
RenderX

21. Chatila WK, Kim JK, Walch H, et al. Genomic and transcriptomic determinants of response to neoadjuvant therapy in rectal cancer. Nat Med 2022 Aug;28(8):1646-1655. [doi: 10.1038/s41591-022-01930-z] [Medline: 35970919]
22. Cerami E, Gao J, Dogrusoz U, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov 2012 May;2(5):401-404. [doi: 10.1158/2159-8290.CD-12-0095] [Medline: 22588877]
23. Gao J, Aksoy BA, Dogrusoz U, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal 2013 Apr 2;6(269):pl1. [doi: 10.1126/scisignal.2004088] [Medline: 23550210]
24. Chen T, Guestrin C. Xgboost: a scalable tree boosting system. Presented at: Proceedings of the 22nd ACM SIGKKD International Conference on Knowledge Discovery and Data Mining; Aug 13, 2016 to Aug 17, 2026; San Francisco, CA p. 785-794.
25. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. arXiv. Preprint posted online on May 22, 2017. [doi: 10.48550/arXiv.1705.07874]
26. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser C Appl Stat 1995 Jan 1;57(1):289-300. [doi: 10.1111/j.2517-6161.1995.tb02031.x]
27. McQuitty LL. Hierarchical linkage analysis for the isolation of types. Educ Psychol Meas 1960 Apr;20(1):55-67. [doi: 10.1177/001316446002000106]
28. MSI-H score predictor. URL: https://www.jiangbioinfo.com/msi-score/ [accessed 2025-09-30]
29. Li Y, Liang L, Dai W, et al. Prognostic impact of programed cell death-1 (PD-1) and PD-ligand 1 (PD-L1) expression in cancer cells and tumor infiltrating lymphocytes in colorectal cancer. Mol Cancer 2016 Aug 24;15(1):55. [doi: 10.1186/s12943-016-0539-x] [Medline: 27552968]
30. Idos GE, Kwok J, Bonthala N, Kysh L, Gruber SB, Qu C. The prognostic implications of tumor infiltrating lymphocytes in colorectal cancer: a systematic review and meta-analysis. Sci Rep 2020 Feb 25;10(1):3360. [doi: 10.1038/s41598-020-60255-4] [Medline: 32099066]
31. Millen R, Hendry S, Narasimhan V, et al. CD8+ tumor-infiltrating lymphocytes within the primary tumor of patients with synchronous de novo metastatic colorectal carcinoma do not track with survival. Clin Transl Immunology 2020;9(7):e1155. [doi: 10.1002/cti2.1155] [Medline: 32953115]
32. Edin S, Wikberg ML, Dahlin AM, et al. The distribution of macrophages with a M1 or M2 phenotype in relation to prognosis and the molecular characteristics of colorectal cancer. PLoS One 2012;7(10):e47045. [doi: 10.1371/journal.pone.0047045] [Medline: 23077543]
33. Oshi M, Sarkar J, Wu R, et al. Intratumoral density of regulatory T cells is a predictor of host immune response and chemotherapy response in colorectal cancer. Am J Cancer Res 2022;12(2):490-503. [Medline: 35261782]
34. Wang H, Tian T, Zhang J. Tumor-associated macrophages (TAMs) in colorectal cancer (CRC): from mechanism to therapy and prognosis. IJMS 2021 Aug 6;22(16):8470. [doi: 10.3390/ijms22168470]
35. Elomaa H, Ahtiainen M, Väyrynen SA, et al. Spatially resolved multimarker evaluation of CD274 (PD-L1)/PDCD1 (PD-1) immune checkpoint expression and macrophage polarisation in colorectal cancer. Br J Cancer 2023 Jun;128(11):2104-2115. [doi: 10.1038/s41416-023-02238-6] [Medline: 37002343]
36. Li E, Hu Y, Han W, et al. The mutational landscape of MSI-H and MSS colorectal cancer. J Clin Oncol 2019 May 26;37(15_suppl):e15122. [doi: 10.1200/JCO.2019.37.15_suppl.e15122]
37. Zhang Y, Rajput A, Jin N, Wang J. Mechanisms of immunosuppression in colorectal cancer. Cancers (Basel) 2020 Dec 20;12(12):3850. [doi: 10.3390/cancers12123850] [Medline: 33419310]

Abbreviations

CEA: carcinoembryonic antigen
CRC: colorectal cancer
DAPI: 4,6-diamidino-2-phenylindole
DAVID: Database for Annotation, Visualization, and Integrated Discovery
ICI: immune checkpoint inhibitor
mIHC: multiplex immunohistochemistry
MSI-H: microsatellite instability-high
MSS: microsatellite stability
MWT: microwave treatment
PD-1: antiprogrammed cell death-1
PD-L1: antiprogrammed cell death ligand 1
TIL: tumor-infiltrating lymphocyte
Treg: regulatory T cell
TSA: tyramide signal amplification
XGBoost: extreme gradient boosting

XSL•FO
RenderX

Edited by J Sarvestan; submitted 02.10.24; peer-reviewed by J Lai, W Ma; revised version received 23.08.25; accepted 26.08.25; published 16.10.25.

Please cite as:
Yan H, Jiang L, Li Y, Wang F, Mo S, Sheng W, Huang D, Peng J
A Machine Learning–Based Scoring System to Identify High Immunoactivity Microsatellite Stability Tumors by Quantifying Similarity to Microsatellite Instability-High Tumors in Colorectal Cancers: Development and Quantitative Study
JMIR Form Res 2025;9:e66960
URL: https://formative.jmir.org/2025/1/e66960
doi:10.2196/66960

© Hongkai Yan, Li Jiang, Yaqi Li, Fengchong Wang, Shaobo Mo, Weiqi Sheng, Dan Huang, Junjie Peng. Originally published in JMIR Formative Research (https://formative.jmir.org), 16.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Clinical Usefulness of a Smartphone-Based 6-Minute Walk Test in a Hospital Outpatient Clinic Within the Constraints of the COVID-19 Pandemic: Mixed Methods Study

Dario Salvi[1], PhD; Carl Magnus Olsson[1], PhD; Jackson Molloy[2], MSc; Elizabeth Orchard[3], MD

[1]Department of Computer Science and Media Technology, Sustainable Digitalisation Research Centre, Malmö University, Nordenskiöldsgatan 1, Malmo, Sweden

[2]Cardiac Physiology, Oxford University Hospitals NHS Trust, Oxford, United Kingdom

[3]Department of Cardiology, Oxford University Hospitals NHS Trust, Oxford, United Kingdom

Corresponding Author:
Dario Salvi, PhD
Department of Computer Science and Media Technology, Sustainable Digitalisation Research Centre, Malmö University, Nordenskiöldsgatan 1, Malmo, Sweden

Abstract

Background: The 6-minute walk test (6MWT) measures exercise capacity in cardiorespiratory, neurological, and musculoskeletal conditions. It consists of observing how far a patient can walk in 6 minutes and is usually performed in a corridor in a clinic. During the COVID-19 pandemic, as health care systems cancelled nonurgent outpatient appointments, many tests were conducted online. At Oxford University Hospitals National Health Service Foundation Trust, patients followed up on by cardiovascular outpatient clinics were asked to use the open-source Timed Walk app to perform the 6MWT in their community as a substitute for the regular tests in the clinic.

Objective: This study aimed to assess the clinical usefulness of the app within the context of the pandemic.

Methods: Consented patients were invited to perform a 6MWT outdoors using the app at least once a month and report the results through periodic telephone calls and visits. Clinical decisions made for the same cohort were registered, with a focus on the effect of the app in supporting decision-making. Data collected through the app during the study period were compared with 6MWTs performed in the prepandemic period.

Results: This study was conducted between October 2021 and December 2022. A total of 55 participants consented (n=25, 45% female; mean age 44.80, SD 17.49 y). In total, 741 events were logged. A total of 51 medical decisions were made for 25 patients; in 41% (21/51) of the decisions, the app played a role, affecting 44% (11/25) of the patients. Between 2018 and 2022, a cohort of 49 patients for whom data were available performed 63 6MWTs in the clinic (18 in 2021), whereas the same patients performed 605 tests using the app in 2022 (ie, October 2021 to December 2022).

Conclusions: The use of the Timed Walk app for remote 6MWTs allowed clinicians to obtain frequent and objective indications of the status of the patients during the pandemic, compensating for the absence of regular clinic appointments and providing 33 times more tests than in the prepandemic period. These tests supported approximately half of the clinical decisions made regarding the consented patients, showing that the app is useful in clinical practice.

Trial Registration: ClinicalTrials.gov NCT05096819; https://www.clinicaltrials.gov/study/NCT05096819

(JMIR Form Res 2025;9:e70495) doi:10.2196/70495

KEYWORDS

Timed Walk app; 6-minute walk test; 6MWT; physical capacity; mobile health; technology acceptance; usability; mixed methods

Introduction

Background

The switch to home-based, remote follow-up during the COVID-19 pandemic quickly became crucial as outpatient appointments were cancelled to minimize exposure to infected persons [1] and reduce the burden on health care providers. This was especially important for vulnerable and high-risk patients and showed encouraging clinical outcomes [1-3]. In the context of respiratory and cardiac diseases, it was suggested that exercise tests could be delivered remotely in home settings instead of clinics [4,5]. Simple tests such as the 6-minute walk test (6MWT), the sit-to-stand test, the Timed Up and Go test, and the step test were particularly useful, given their reliability and strong correlation with general symptoms and the possibility of patients performing them independently owing to the use of technology such as accelerometers and smartphone apps.

XSL•FO
RenderX

The 6MWT in particular is a standard method for measuring exercise capacity in conditions such as pulmonary arterial hypertension, heart failure, and cardiorespiratory and musculoskeletal conditions. The test consists of measuring how far a patient can walk in 6 minutes. The measured 6-minute walk distance (6MWD) is indicative of the functional status of the patient and can be used for follow-up and for measuring response to rehabilitation and pharmacological interventions [6-8].

The test can be delivered through smartphone apps, as shown in a previous research project where a mobile phone app was designed, developed, and evaluated to allow patients with pulmonary hypertension to perform the 6MWT in the community [9,10]. This is particularly useful for patients who cannot attend regular visits at the clinic, as was the case during the COVID-19 pandemic, or to reduce costs associated with undertaking the test itself. Remote testing reduces the burden of traveling to clinics for patients and eliminates the need for health care professionals to receive the patients, guide the tests, and process the results [9]. Remote 6MWTs are, in fact, recommended in recent guidelines for cardiac and pulmonary rehabilitation [11,12].

Notwithstanding its promises, telemedicine still faces significant barriers, such as security and privacy concerns, lack of proven effectiveness, and poor design and fit into clinical workflows [13,14]. Understanding benefits of telemedicine as complementary to in-clinic visits for both clinicians and patients may be a reasonable postpandemic perspective for research. To address this need, our study evaluated the adoption of a remote mobile-based 6MWT among hospital outpatients from a clinical and patient perspective during the COVID-19 pandemic.

Objectives

This study had three specific aims: (1) to assess the clinical usefulness of a smartphone app for remote 6MWT in the context of the pandemic, (2) to assess user acceptance of the app, and (3) to validate and improve the algorithms that compute the walked distance from the sensor data collected by the phone. This paper, which is part of a series of 3 manuscripts, focuses on the first aim, concretely, the *clinical usefulness* of the app.

Methods

Overview

The goal of this study was to contribute to existing research on the adoption of remote mobile-based 6MWTs through a mixed methods approach where we assessed complementary quantitative and qualitative data to validate the use of an app in the context of the pandemic. As this first paper focuses on clinical usefulness, we report on the study design particularly focusing on this aim.

The app used in this study relies on the outdoor 6MWT algorithm developed in the studies by Salvi et al [9,10]. This was motivated not only by the satisfactory results shown in previous research but also by the important facts that the algorithm is published under an open license and is simple to implement. For this study, we reimplemented the same algorithm in a new smartphone app described in the following section.

The Timed Walk App

The Timed Walk app is available free of charge for both Android and iOS, and its source code is distributed with an open-source Massachusetts Institute of Technology license [15]. The app has 1 main functionality: to measure the distance walked in 6 minutes using the global navigation satellite system (GNSS) of the phone. The data collected by the app are stored locally on the user's phone. Users decide whether they want to share either the history of all past tests (containing date, distance, and duration) or the details of the last performed test (containing data from sensors including localization coordinates, speed, and acceleration) using a button that opens the standard dialogue of the operating system where the channel of communication (eg, email or SMS text message) can be selected.

The walked distance is computed using the information provided by the GNSS, which thus requires that the test be performed outdoors, where satellites are within reach. As GNSS data can be affected by noise, the app uses an algorithm to filter out unreliable estimated positions and computes a total walked distance as described and validated in the studies by Salvi et al [9,10] (under a different app name).

The app includes extensive instructions about its purpose and how to perform the test, with both text and voice instructions during the test following the American Thoracic Society guidelines [16]. Even if the test is considered generally safe, patients are instructed to stop walking if they do not feel well and to seek medical help. Screenshots of the app are shown in Figure 1.

XSL•FO
RenderX

Figure 1. Screenshots of the Timed Walk app showing the instructions for the test (A), the running test with a countdown timer (B), and the information after completion of the test (C).

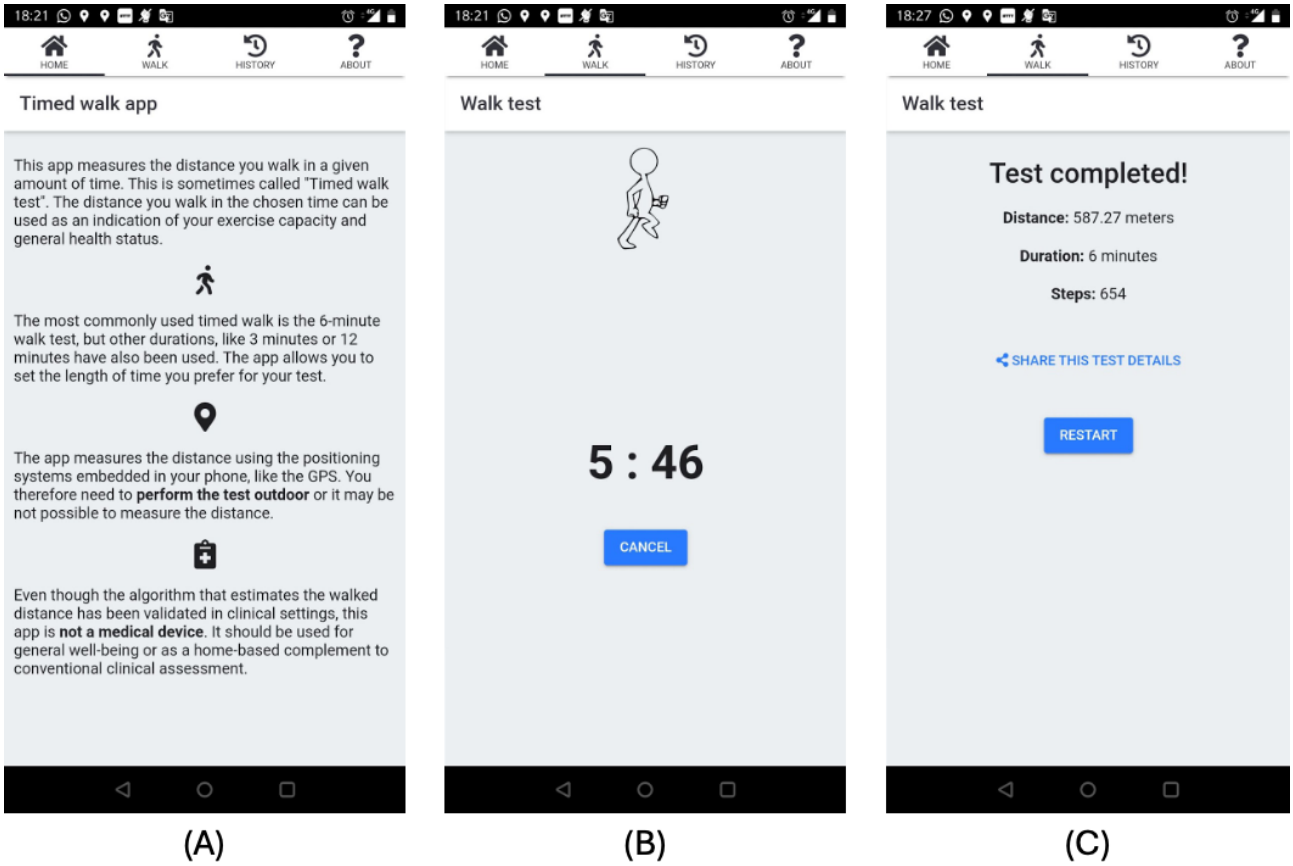

Clinical Usefulness

While the accuracy, reliability, and usability of apps for the 6MWT have been examined in research [9,10,17-19], little is known about the usefulness of app-based measurements in a clinical context. The usefulness of the 6MWT is commonly assessed retrospectively by investigating the relationship between walked distance and clinical variables, including hospitalizations, quality of life, or response to therapy [20]. While the improvement in these variables is an undoubtedly desirable goal, they are measurable only in large-scale, longitudinal studies where confounders can be mitigated. In the context of smartphone apps, frameworks and standards for evaluation exist, but they rarely describe which quantities or indicators to measure as these can vary greatly and are context dependent [21].

In our context, a simple and effective way to measure the clinical usefulness of an app is to measure *the effect of the app in supporting clinical decision-making* [22]. This is particularly suitable for the scenario in which a 6MWT app would provide frequent and objective measurements about patients, information that is "clinically actionable" [23] (ie, that is considered as a relevant input while making clinical decisions).

Measuring usefulness according to this definition (as support for clinical decision-making) can be done by recording clinical events such as visits, evaluations, and treatment changes, which allows for the examination of how clinical decisions are affected by the presence of certain information or workflows [24,25]. To the best of our knowledge, logging of clinical decisions has never been used to evaluate the clinical usefulness of apps despite providing the opportunity to measure the impact of adopting an app in an objective way within a specific context of use, such as the COVID-19 pandemic.

Study Design

Data were collected from patients attending the cardiology outpatient department at Oxford University Hospitals National Health Service Foundation Trust (OUH), Oxford, United Kingdom, during the restrictions put in place due to the COVID-19 pandemic. Patients were either under regular follow-up for their cardiac disease or had been referred to see clinicians in the outpatient setting for assessment and treatment.

Outpatients who were undergoing follow-up in the cardiovascular outpatient clinics (either face-to-face or remotely) in Oxford and were able to walk and use a smartphone were considered eligible. Some of these patients were already using the app before the start of the study, for example, in the case of the pulmonary hypertension clinic.

The inclusion criteria were as follows: (1) being enrolled in an OUH cardiovascular outpatient clinic, (2) owning or having access to a smartphone with either Android or iOS, (3) being able to use a smartphone app, and (4) being able to walk.

The exclusion criteria were as follows: (1) long-term oxygen therapy, (2) cognitive impairments, (3) inability to use a smartphone, (4) pregnancy, and (5) inability to complete a 6MWT.

XSL•FO
RenderX

We did not sample participants as the aim was to involve all patients in the cardiovascular clinics who required a 6MWT and were able to use the app. However, as restrictions started to be lifted in 2022 and practical reasons made the recruitment pace slower than desired (approximately 4 patients per month), a decision was made to stop recruitment after the 100th patient was selected.

Patients were screened by reviewing outpatient clinic lists for possible inclusion every week based on already scheduled clinic visits (either face-to-face or virtual). A list of candidates who met the inclusion criteria was prepared before their scheduled outpatient clinic. Prospective participants were approached by a member of the clinical team for consideration and recruited during their regular outpatient consultation either via phone call or during a visit to the clinic.

During the consultations, patients were asked about their access to and comfort with smartphones to verify compliance with the inclusion criteria. Those who were confirmed to comply with all inclusion criteria were invited to participate, signing a consent form as they were enrolled. Consent forms were delivered via post.

Participants in the study were asked to download the Timed Walk app and use it to perform a 6MWT in the community outdoors at least once a month and report the measured 6MWD to the clinicians during the outpatient consultations via phone or at the clinic.

Patients enrolled at the clinic were introduced to the app during the outpatient appointment and shown how to use it. This was different from the usual clinical care, as they would normally simply undertake a 6MWT. Patients enrolled via phone were instructed verbally. As the app is quite minimalistic, only minor recommendations about performing the test outdoors were needed.

To assess the usefulness of the app for the clinical management of consented patients, at each consultation (either through phone or face-to-face visit), the clinician who contacted the patient logged any study-related event on a spreadsheet marked with a code. Coding was used to simplify and automate later analysis and was defined in collaboration with the clinical professionals that were part of the study.

The spreadsheet contained the following columns:

- Patient ID: study identifier of the patient
- Date: date of the event
- Event type: using a custom code schema, these events were related to recruitment, withdrawal, home or clinic 6MWT, app use, technical problems, clinical investigations, multidisciplinary team discussions, and changes in treatments
- Comments: any additional comments (free text)
- Distance: if the event was related to a 6MWT, the walked distance in meters was reported in this column

In addition, the following demographic information was recorded for each patient: age, sex, main conditions, and study participation (declined participation, consented, completed the study, or withdrew).

Other aspects included in this study but not part of this paper were usability and acceptance questionnaires, interviews, and data quality analysis.

This study sits between the *appropriateness* and *feasibility* levels of the taxonomy of implementation science outcomes [26], where constructs such as perceived fit, relevance, usefulness, actual fit or utility, suitability for everyday use, and practicability are measured at an individual health care provider such as a clinic or organization. As suggested by Proctor et al [26], common measurement tools at this stage are surveys, interviews, and administrative data, which is compatible with our study design.

Data Analysis

For the purpose of assessing the clinical usefulness of the app within the context of the pandemic, we identified 2 objectives.

Objective 1: Assess Whether Differences in Distance Walked According to the App Altered the Management of Patients

This was estimated by counting the number of clinical investigations, multidisciplinary team discussions, and changes in treatments triggered by variations in app-based 6MWD. The spreadsheet where events were logged was used for this purpose, including data from all the patients who consented. We computed descriptive statistics, including the number of clinical decisions; the number of patients involved in those decisions; and, given that patients were involved in the study over a varying number of days, the number of decisions per patient and year.

Objective 2: Compare the Use of the 6MWT During the Pandemic With Prepandemic Use

This was done by comparing the total number of 6MWTs performed during the pandemic period compared with the prepandemic period. The number of in-clinic 6MWTs was retrieved from the electronic health record of the hospital, whereas the number of app-based 6MWTs was extracted from the event log. We considered only those patients who consented and whose data were available on the clinical health record over the entire period under analysis, comprising 2018 to 2022.

Ethical Considerations

This study was a collaboration between the OUH, United Kingdom, and Malmö University (MAU), Sweden. MAU is the main developer and maintainer of the Timed Walk app and provided engineering support in terms of both the app and data analysis, whereas OUH recruited patients and collected data. This study was conducted in accordance with the Declaration of Helsinki, with ethics approval from the UK National Health Service Health Research Authority (protocol reference 17/WM/0355). Written informed consent was obtained from all participants involved in the study, and no compensation was provided to them.

All data were sent by patients to the care team at the OUH through email or verbally shared via phone calls; data were subsequently pseudonymized with study participant IDs and

XSL•FO
RenderX

analyzed by staff at MAU. After analysis, data were archived following UK and Swedish regulations.

Results

Demographics

This study was conducted between October 2021 and December 2022. A total of 100 patients initially agreed to participate. Of these 100 patients, 88 (88%) were contacted over the phone, and 12 (12%) were unreachable. Of the 88 patients contacted, 77 (88%) confirmed their agreement to participate and received printed study information and informed consent forms via post. Of these 77 patients, 55 (71%) signed and returned the consent forms, and their data were used for analysis for objective 1. Of the 55 participants, 3 (5%) withdrew during the study but agreed to have their data used for analysis. Of the 55 consented patients, 49 (89%) had clinical health records at the hospital and were included for objective 2.

A CONSORT (Consolidated Standards of Reporting Trials)–like diagram is shown in Figure 2.

XSL•FO
RenderX

Figure 2. CONSORT (Consolidated Standards of Reporting Trials)–like diagram of the study.

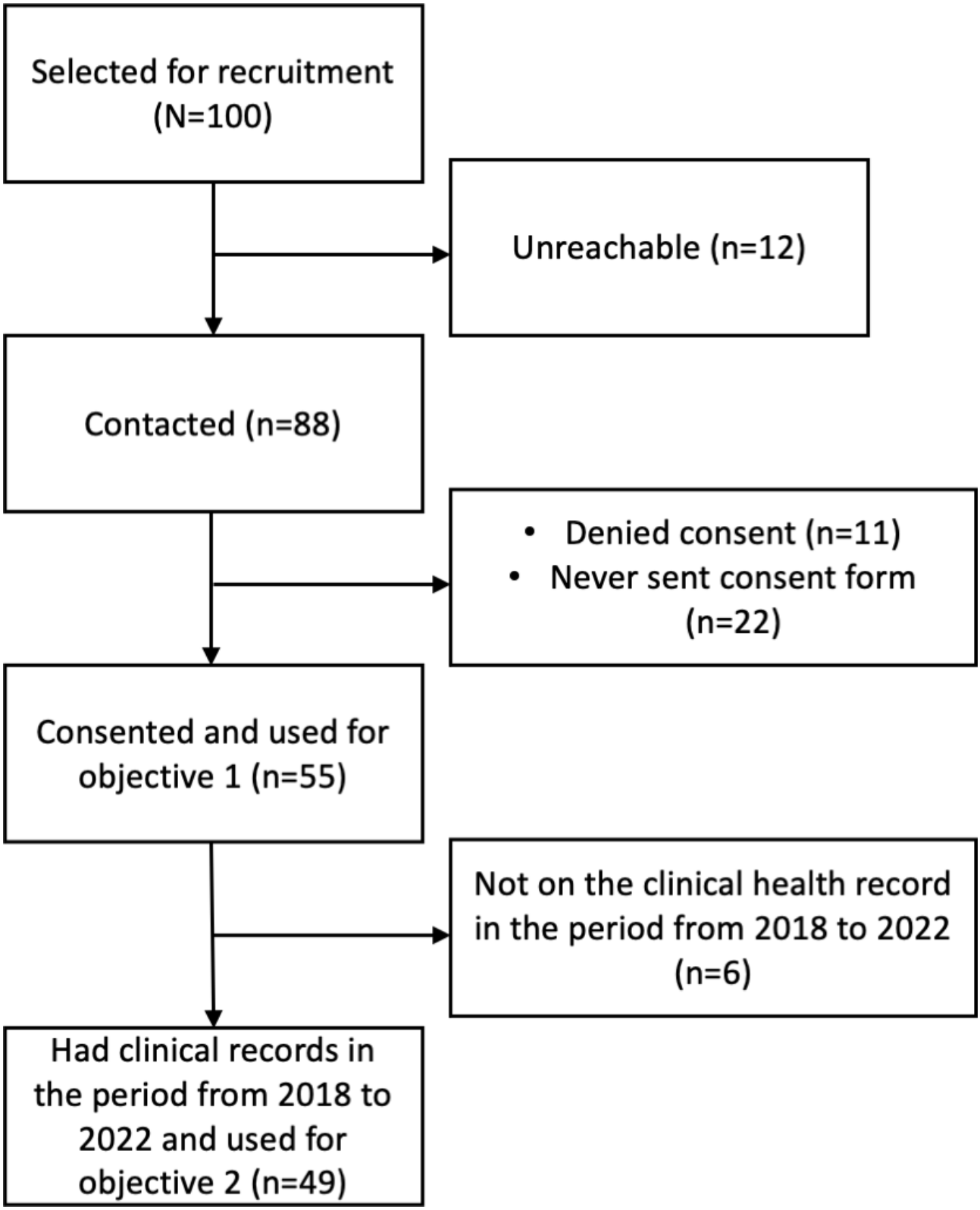

Of the 55 consented participants, 25 (45%) were female. The average age was 44.80 (SD 17.49) years. Patients were involved in the study for a varying number of days (mean 314, SD 93; range 62-441 days). In total, 741 events were logged between September 29, 2021 (the earliest event), and December 30, 2022 (the latest event). In total, 612 mobile-based 6MWTs were completed, with an average of 11.26 (SD 15.11; range 0-84) mobile 6MWTs per patient. The average distance measured in those tests was 471.20 (SD 125.51; range 95-989) m. In total, 13% (7/55) of the patients did not perform any mobile 6MWT.

XSL•FO
RenderX

Objective 1: Assess Whether Differences in Distance Walked According to the App Altered the Management of Patients

We used the data from the 55 patients who consented for the analysis, including the 3 (5%) patients who withdrew but were happy to have their data used.

From the logs, we computed descriptive statistics about the frequency of clinical decisions, including new evaluations and changes in treatment plans. The types of logged items are reported in Table 1 for each type. As patients were involved in the study over different time spans, we also report the average and SD of the number of events per year (365 days) per patient.

Table . Number of events, number of patients involved, and yearly average of times logged per event type per patient.

Logged item type	Number of events	Number of patients involved	Events per year per patient, mean (SD)
The patient agreed to join the study and provided consent.	54	54	1.31 (0.88)
The patient decided to withdraw from the study.	1	1	0.05 (0.34)
The patient declined consent and did not join the study.	0	0	0.0 (0.0)
The patient declared having the app installed on a phone they could use.	0	0	0.0 (0.0)
The patient declared that they had technical issues with the app.	0	0	0.0 (0.0)
The patient shared a 6MWT[a] not older than 2 weeks through either email or telephone.	612	48	13.73 (14.36)
Regular clinic 6MWT.	14	10	0.29 (0.64)
The patient needed further evaluation because of changed medical conditions. App-based 6MWT was irrelevant to the decision.	7	4	1.1 (0.52)
The patient needed further evaluation because of changed medical conditions. App-based 6MWT was partially involved in the decision.	7	7	0.22 (0.64)
The patient needed further evaluation because of changed medical conditions. App-based 6MWT was fundamental to the decision.	4	3	0.15 (0.71)
The patient needed a change in treatment. App-based 6MWT was irrelevant to the decision.	23	15	0.53 (1.11)
The patient needed a change in treatment. App-based 6MWT was partially involved in the decision.	10	5	0.22 (0.75)
The patient needed a change in treatment. App-based 6MWT was fundamental to the decision.	0	0	0.0 (0.0)
Any clinical decision.	51	25	1.25 (1.79)
Any clinical decision for which app-based 6MWT was partially involved or fundamental.	21	11	0.59 (1.27)

[a]6MWT: 6-minute walk test.

As shown in the table, most of the logged items (612/741, 82.6%) were app-based 6MWTs provided by 87% (48/55) of the patients. In comparison, only 1.9% (14/741) of clinic-based 6MWTs were logged in the same period for 18% (10/55) of the patients. A total of 6.9% (51/741) of the logs were related to clinical evaluations or changes in treatment, affecting 45% (25/55) of the patients. In 41% (21/51) of the decisions, affecting 44% (11/25) of the patients, the information provided by the app played a role. In 8% (4/51) of the decisions, affecting 12% (3/25) of the patients, the app played a fundamental role.

XSL•FO
RenderX

In terms of events per patient per year, we observed slightly more than one 6MWT per month (13.73, SD 14.36 per year), with some patients performing as many as 81 tests per year. Clinical decisions were less frequent, with a mean of 1.25 (SD 1.79) per year, approximately half of which involved the use of the 6MWT as measured through the app.

It is noteworthy that patient consent, withdrawal from study, app installation, use, and issues were not consistently recorded on the log but were also noted on a separate sheet.

Objective 2: Compare the Use of the 6MWT During the Pandemic With Prepandemic Use

We provide statistics regarding the number of 6MWTs performed by our cohort of patients in the clinic over the last 5 years and compare it with the number of 6MWTs performed during the study period. This is to identify the difference between the number of 6MWTs per patient per year before the pandemic (standard 6MWT in the clinic) and after the pandemic (using the app).

The number of 6MWTs conducted in the clinic for that cohort was 17 in 2018, 16 in 2019, 8 in 2020, 18 in 2021, and 13 in 2022. The total number of 6MWTs performed using the app during the study period (October 2021 to December 2022) was 605 for the same group of 49 participants.

The reduction in in-clinic tests in the year 2020 was due to the restrictions imposed by the pandemic, which was later compensated for by more tests in 2021. Of note is that the number of app-based tests during the study period was 33 times larger than the highest number of in-clinic tests for the analyzed cohort (18 in 2021). This study also yielded 8 times more tests than the number of in-clinic tests conducted over 5 years.

Discussion

Principal Findings

The use of the app allowed for the retrieval of frequent information about patients remotely. In fact, while only 14 6MWTs were performed in the clinic on our 55 participants, 612 additional 6MWTs were performed using the app and communicated to the involved clinicians. The information sent played a role in the clinical decisions made regarding those patients. In 41% (21/51) of the clinical decisions, the distances provided by the app were considered a contributing element, and in 8% (4/51) of the cases, they were considered even fundamental to the decision. These decisions affected one-fifth of the patients who used the app (11/55, 20%) and approximately half (11/25, 44%) of the patients for whom clinical decisions were made during the study period.

The 6MWT has a recognized prognostic value. The walked distance has been shown to be a predictor of all-cause mortality and hospitalization [27] and to be responsive to clinical change [6]; therefore, it was expected that the information retrieved through the app would be used in clinical decisions. Given the circumstances during the pandemic, the app may have been the only means to obtain objective information about some of the patients involved. This clearly indicates its usefulness as a tool that can augment the information that clinicians obtain either as a complement to in-clinic 6MWTs or as a substitute, especially if supervised 6MWTs are not readily available, for example, in remote areas.

Data show that, for the cohort of 49 patients for whom we had historical data, the number of 6MWTs in the clinic varied between 8 and 18 per year, with the lowest number in 2020 due to the pandemic. During the study period, which roughly encompassed the entire year 2022, the number of 6MWTs in the clinic was 13, which means that 26% (13/49) of those patients were able to perform a test in the clinic. In comparison, 87% (48/55) of our patients performed at least one 6MWT on the app in the same period, and on average, our patients performed one 6MWT per month on the app. This proves not only that the app could compensate for the regular 6MWTs in the clinic during the pandemic but also that the amount of data that the app allows to be retrieved is 33 times larger than that for regular visits, with 605 app-based tests in 2022 compared to 18 retrieved in the clinic in 2021.

These results have to be considered in the context of the fact that the distribution of the number of tests per patient was wide, with 7 participants performing no tests at all and 2 patients performing one-quarter (147/612, 24%) of all tests (one performed 66 tests per year, and the other performed 81 tests per year).

Limitations

We showed that 41% (21/51) of the clinical decisions made regarding our cohort during the study were at least partially based on the measurements provided by the app. However, this study did not use a control group; therefore, it is impossible to ascertain whether patients would have received the same level of care without the use of the app.

It is also important to observe that this study was conducted under special circumstances due to the restrictions imposed during the COVID-19 pandemic. During that period, patients may have been keener to use telemedicine to ensure that they were being followed up on. Therefore, the results obtained may not be generalizable outside of those circumstances.

Another limitation of our study is that our population was relatively young (mean age 44.80, SD 17.49 years) and healthy (mean 6MWD 471.20, SD 125.51 m). The distribution of age and 6MWD was wide but skewed, with 62% (34/55) of the patients being below 50 years of age and 72% of the median 6MWD per patient being above 400 m. Therefore, further studies targeting unhealthier individuals are advised to confirm usefulness in those populations rather than assume similar results to those we obtained.

Conclusions

This study was motivated by the need to follow up on patients during the restrictions imposed by the COVID-19 pandemic. The use of the Timed Walk app for remote 6MWTs was helpful, as it allowed clinicians to obtain objective measurements of the status of the patients remotely without the need to see patients in person at the clinic. These measurements were at least partially used in approximately half of the medical decisions, showing that the Timed Walk app is useful in clinical practice.

XSL•FO
RenderX

In the future, the usefulness of the app could also be investigated in a study in which patients at greater risk of worsening health are involved. In such a case, it would be important to design a protocol with periodical data review that leverages data automatically sent by the app. A randomized controlled trial would be needed to compare usual care versus app-aided care and assess differences. The study could be designed for any type of patient who is typically administered a 6MWT, such as patients with cardiology (heart failure or pulmonary hypertension), neurology (multiple sclerosis or Parkinson disease), or respiratory diseases (chronic obstructive pulmonary disease).

Acknowledgments

The authors would like to warmly thank all the participants in this study. Their contribution has been invaluable. This work has been partially supported by 2 grants from the Swedish Knowledge Foundation: the Internet of Things and People Research Profile and the Intelligent and Trustworthy IoT Systems Synergy project.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: DS, EO, CMO
Data curation: DS, JM
Formal analysis: DS, CMO
Funding acquisition: CMO, EO
Investigation: JM
Methodology: DS, EO, CMO
Project administration: EO
Resources: EO
Software: DS
Supervision: EO
Writing—original draft: DS
Writing—review and editing: JM, CMO, EO

Conflicts of Interest

None declared.

References

1. Colbert GB, Venegas-Vera AV, Lerma EV. Utility of telemedicine in the COVID-19 era. Rev Cardiovasc Med 2020 Dec 30;21(4):583-587. [doi: 10.31083/j.rcm.2020.04.188] [Medline: 33388003]
2. Biancuzzi H, Dal Mas F, Bidoli C, et al. Economic and performance evaluation of e-Health before and after the pandemic era: a literature review and future perspectives. Int J Environ Res Public Health 2023 Feb 24;20(5):4038. [doi: 10.3390/ijerph20054038] [Medline: 36901048]
3. Alkilany R, Tarabichi Y, Hong R. Telemedicine visits during COVID-19 improved clinic show rates. ACR Open Rheumatol 2022 Feb;4(2):136-141. [doi: 10.1002/acr2.11372] [Medline: 34786877]
4. Holland AE, Malaguti C, Hoffman M, et al. Home-based or remote exercise testing in chronic respiratory disease, during the COVID-19 pandemic and beyond: a rapid review. Chron Respir Dis 2020;17:1479973120952418. [doi: 10.1177/1479973120952418] [Medline: 32840385]
5. Hwang R, Fan T, Bowe R, et al. Home-based and remote functional exercise testing in cardiac conditions, during the COVID-19 pandemic and beyond: a systematic review. Physiotherapy 2022 Jun;115:27-35. [doi: 10.1016/j.physio.2021.12.004] [Medline: 35180642]
6. Bellet RN, Adams L, Morris NR. The 6-minute walk test in outpatient cardiac rehabilitation: validity, reliability and responsiveness--a systematic review. Physiotherapy 2012 Dec;98(4):277-286. [doi: 10.1016/j.physio.2011.11.003] [Medline: 23122432]
7. Hoeper MM, Badesch DB, Ghofrani HA, et al. Phase 3 trial of sotatercept for treatment of pulmonary arterial hypertension. N Engl J Med 2023 Apr 20;388(16):1478-1490. [doi: 10.1056/NEJMoa2213558] [Medline: 36877098]
8. Giannitsi S, Bougiakli M, Bechlioulis A, Kotsia A, Michalis LK, Naka KK. 6-minute walking test: a useful tool in the management of heart failure patients. Ther Adv Cardiovasc Dis 2019;13:1753944719870084. [doi: 10.1177/1753944719870084] [Medline: 31441375]
9. Salvi D, Poffley E, Tarassenko L, Orchard E. App-based versus standard six-minute walk test in pulmonary hypertension: mixed methods study. JMIR Mhealth Uhealth 2021 Jun 7;9(6):e22748. [doi: 10.2196/22748] [Medline: 34096876]

XSL•FO
RenderX

10. Salvi D, Poffley E, Orchard E, Tarassenko L. The mobile-based 6-minute walk test: usability study and algorithm development and validation. JMIR Mhealth Uhealth 2020 Jan 3;8(1):e13756. [doi: 10.2196/13756] [Medline: 31899457]
11. Psotka MA, Abraham WT, Fiuzat M, et al. Conduct of clinical trials in the era of COVID-19. J Am Coll Cardiol 2020 Nov;76(20):2368-2378. [doi: 10.1016/j.jacc.2020.09.544]
12. Scherrenberg M, Wilhelm M, Hansen D, et al. The future is now: a call for action for cardiac telerehabilitation in the COVID-19 pandemic from the secondary prevention and rehabilitation section of the European Association of Preventive Cardiology. Eur J Prev Cardiol 2021 May 14;28(5):524-540. [doi: 10.1177/2047487320939671]
13. Schreiweis B, Pobiruchin M, Strotbaum V, Suleder J, Wiesner M, Bergh B. Barriers and facilitators to the implementation of eHealth services: systematic literature analysis. J Med Internet Res 2019 Nov 22;21(11):e14197. [doi: 10.2196/14197] [Medline: 31755869]
14. Granja C, Janssen W, Johansen MA. Factors determining the success and failure of eHealth interventions: systematic review of the literature. J Med Internet Res 2018 May 1;20(5):e10235. [doi: 10.2196/10235] [Medline: 29716883]
15. Salvi D. Dariosalvi78/timedwalkapp. GitHub. 2024 Nov 2. URL: https://github.com/dariosalvi78/TimedWalkApp [accessed 2025-09-26]
16. Enright PL. The six-minute walk test. Respir Care 2003 Aug;48(8):783-785. [Medline: 12890299]
17. Brooks GC, Vittinghoff E, Iyer S, et al. Accuracy and usability of a self-administered 6-minute walk test smartphone application. Circ Heart Fail 2015 Sep;8(5):905-913. [doi: 10.1161/CIRCHEARTFAILURE.115.002062] [Medline: 26283292]
18. Capela NA, Lemaire ED, Baddour N. Novel algorithm for a smartphone-based 6-minute walk test application: algorithm, application development, and evaluation. J Neuroeng Rehabil 2015 Feb 20;12(1):19. [doi: 10.1186/s12984-015-0013-9] [Medline: 25889112]
19. Ziegl A, Rzepka A, Kastner P, et al. mHealth 6-minute walk test - accuracy for detecting clinically relevant differences in heart failure patients. Annu Int Conf IEEE Eng Med Biol Soc 2021 Nov;2021:7095-7098. [doi: 10.1109/EMBC46164.2021.9630118] [Medline: 34892736]
20. Acquistapace F, Piepoli MF. The walking test: use in clinical practice. Monaldi Arch Chest Dis 2009 Mar;72(1):3-9. [doi: 10.4081/monaldi.2009.336] [Medline: 19645206]
21. Mathews SC, McShea MJ, Hanley CL, Ravitz A, Labrique AB, Cohen AB. Digital health: a path to validation. NPJ Digit Med 2019;2(1):38. [doi: 10.1038/s41746-019-0111-3] [Medline: 31304384]
22. McLernon DJ, Giardiello D, Van Calster B, et al. Assessing performance and clinical usefulness in prediction models with survival outcomes: practical guidance for cox proportional hazards models. Ann Intern Med 2023 Jan;176(1):105-114. [doi: 10.7326/M22-0844] [Medline: 36571841]
23. Henson P, David G, Albright K, Torous J. Deriving a practical framework for the evaluation of health apps. Lancet Digit Health 2019 Jun;1(2):e52-e54. [doi: 10.1016/S2589-7500(19)30013-5] [Medline: 33323229]
24. Berkhout M, Smit K, Versendaal J. Decision discovery using clinical decision support system decision log data for supporting the nurse decision-making process. BMC Med Inform Decis Mak 2024 Apr 18;24(1):100. [doi: 10.1186/s12911-024-02486-3] [Medline: 38637792]
25. Shalom E, Shahar Y, Lunenfeld E. An architecture for a continuous, user-driven, and data-driven application of clinical guidelines and its evaluation. J Biomed Inform 2016 Feb;59:130-148. [doi: 10.1016/j.jbi.2015.11.006] [Medline: 26616284]
26. Proctor E, Silmere H, Raghavan R, et al. Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda. Adm Policy Ment Health 2011 Mar;38(2):65-76. [doi: 10.1007/s10488-010-0319-7] [Medline: 20957426]
27. Du H, Wonggom P, Tongpeth J, Clark RA. Six-minute walk test for assessing physical functional capacity in chronic heart failure. Curr Heart Fail Rep 2017 Jun;14(3):158-166. [doi: 10.1007/s11897-017-0330-3] [Medline: 28421409]

Abbreviations

6MWD: 6-minute walk distance
6MWT: 6-minute walk test
CONSORT: Consolidated Standards of Reporting Trials
GNSS: global navigation satellite system
MAU: Malmö University
OUH: Oxford University Hospitals National Health Service Foundation Trust

XSL•FO
RenderX

Edited by A Mavragani; submitted 23.12.24; peer-reviewed by KAGJ Romijnders, S Kling; revised version received 01.09.25; accepted 19.09.25; published 10.10.25.

Please cite as:
Salvi D, Olsson CM, Molloy J, Orchard E
Clinical Usefulness of a Smartphone-Based 6-Minute Walk Test in a Hospital Outpatient Clinic Within the Constraints of the COVID-19 Pandemic: Mixed Methods Study
JMIR Form Res 2025;9:e70495
URL: https://formative.jmir.org/2025/1/e70495
doi:10.2196/70495

© Dario Salvi, Carl Magnus Olsson, Jackson Molloy, Elizabeth Orchard. Originally published in JMIR Formative Research (https://formative.jmir.org), 10.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Exploring the Acceptability and Suitability of Synchronous Online Focus Groups for Health Research With Métis Nation of Ontario Citizens: An Internet-Based Survey

Abigail J Simms[1,2,3], BHSc, MPH; Noel Tsui[1,2,3], BSc, MPH; Robynn Sadler[1]; Amy Mersereau[1], MA; Cindi Rye[1]; C David Crenna[4], BA, MA, MS, PhD; Jeff Evenson[1,5]; Sarah A Edwards[1,2,3], HBSc, MHSc, PhD

[1]Housing and Infrastructure Branch, Métis Nation of Ontario, Ottawa, ON, Canada

[2]Data and Analytic Services, ICES (formerly the Institute for Clinical Evaluative Sciences), V1 06, 2075 Bayview Avenue, Toronto, ON, Canada

[3]Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada

[4]Priority Decision Data, Perth, ON, Canada

[5]Edgewood Associates, Toronto, ON, Canada

Corresponding Author:
Abigail J Simms, BHSc, MPH
Housing and Infrastructure Branch, Métis Nation of Ontario, Ottawa, ON, Canada

Abstract

Background: The COVID-19 pandemic forced many researchers to adjust research methods from in-person to online formats. This paper explores the acceptability and suitability of synchronous online focus groups used to explore housing and health with Métis Nation of Ontario (MNO) citizens, one of 3 constitutionally recognized Indigenous Peoples in Canada.

Objective: The objective of this internet-based survey was to understand the experiences of MNO citizens participating in synchronous online focus groups.

Methods: Only participants of the "Understanding Housing and Health" project were eligible to complete the survey and were recruited via a 'thank you' email. The survey asked respondents to rate their experience, satisfaction, and preference, as well as the feasibility and cultural appropriateness of the online focus group. An open textbox allowed respondents to share additional thoughts. Demographic and personal information (ie, age, gender, MNO region, and email) were collected. A total of 33/35 eligible participants completed the survey. A content analysis was conducted to generate themes from the open textbox responses and used to triangulate the results. The survey was developed collaboratively with MNO staff.

Results: Most respondents identified as women and were 45 - 65 years and older. All respondents had used Zoom before, and most (n=28, 85%) were either strongly or somewhat confident in their ability to use Zoom. One hundred percent of respondents strongly or somewhat agreed that they would participate in an online focus group in the future, and 86% of respondents strongly (n=22, 67%) or somewhat (n=7, 21%) agreed that an online focus group was culturally appropriate for Métis health research. A total of 82% (n=27) of respondents strongly or somewhat agreed that an online focus group was more feasible. Moreover, 58% (n=19) of respondents strongly or somewhat disagreed that they would have preferred to participate in an in-person focus group, while 27% (n=9) were neutral. Around 58% (n=19) of respondents could see other participants all of the time and did not experience lag at any point, while 25 (76%) could hear other participants all of the time, indicating fewer issues with audio. A total of 70% (n=23) of respondents felt they could connect with others all or most of the time, while 30% (n=10) felt they could do this some of the time or rarely. Content analysis of the open textbox responses generated 4 themes: internet and technology issues, accessibility, structure of the group discussion, and positive feedback.

Conclusions: The use of online focus groups for research with MNO citizens is acceptable; however, internet and technology issues can affect a participant's ability to fully engage. Considerations around cultural appropriateness and connecting with others should be made. This information will help inform method selection for future research work conducted in collaboration with the MNO.

(JMIR Form Res 2025;9:e70986) doi:10.2196/70986

KEYWORDS

Indigenous health; Métis health; acceptability; virtual research methods; synchronous virtual focus group

XSL•FO
RenderX

Introduction

The COVID-19 pandemic forced a pivot to online platforms for face-to-face research activities such as focus groups, which impacted research projects conducted by the Métis Nation of Ontario (MNO) [1-3]. The MNO represents the rights and interests of Métis in Ontario, Canada, and provides many social, justice, and education programs for its citizens and conducts relevant research on their behalf [4]. The MNO registry comprises more than 30,000 citizens and recently underwent a comprehensive review [4,5]. Using the national definition of Métis, the MNO registry determines citizenship for an applicant based on (1) self-identifying as Métis, (2) being of historic Métis Nation ancestry, (3) being distinct from other Indigenous Nations, (4) being accepted by the Métis Nation [5]. The Métis are one of 3 recognized Indigenous Nations with the section 35 rights under Canada's constitution. The Métis are descendants of First Nations and European settler ancestry that formed distinct communities via ethnogenesis along the waterways and lakes of Ontario, as well as the 3 prairie provinces, parts of British Columbia, the Northwest Territories, and Northern United States [6]. Métis have a distinct history, language, and culture, though there are regional differences [6]. The MNO pivoted to adapt to the COVID-19 pandemic. Online communication tools, such as Facebook (Meta), Instagram (Meta), X (formerly Twitter), the MNO website, and email list services, made up an important part of MNO communications with citizens throughout the COVID-19 pandemic. Online town halls were also used to connect with MNO citizens to share pandemic updates and to host question-and-answer sessions with Métis doctors on the SARS-CoV-2 virus and vaccines [7]. Other major community events also shifted online, such as the MNO's Annual General Assembly.

The need to rely on Zoom (Zoom Video Communications, Inc) to interact with Métis citizens for research purposes during the COVID-19 pandemic posed important questions around the acceptability and suitability of online data collection for this population since visiting with one another and spending time face-to-face is an important part of traditional ways of knowing and being for Métis people [8]. Visiting in a Métis context has even been adapted as a health research approach [9]. Interest in what may be considered a culturally appropriate data collection method for research with Indigenous Peoples led to a review of the literature conducted by Wright et al [10]. They identified photovoice, symbol-based reflection, circles, and story-telling oral-data collection methods as being rigorous and culturally appropriate for research with Indigenous Peoples in Canada [10]. This review specifically excluded interviews and focus groups due to concerns that they may not always be culturally appropriate for this population [10,11]. However, a systematic review by Hunt and Young [12], looking at studies using group-based qualitative methods for health research with Indigenous children, found that many studies incorporated aspects of Indigenous sharing circles into Western focus groups to produce a hybrid method that was culturally appropriate and safe for research with this population, which is in alignment with our focus group method. Globally, many other researchers have engaged focus groups, both online and in-person, in community-based participatory study designs with Indigenous Peoples exploring topics such as consent in research, housing, dementia, and traditional medicines, among others [13-17].

Although we were forced to use online tools over more traditional in-person, face-to-face ways of collecting information, online data collection takes advantage of technological advancements and offers a variety of benefits [2,3,18-21]. Our survey asked MNO citizen participants of online synchronous focus groups about their experience with the online data collection method, focusing on acceptability and suitability. Synchronous focus groups involve participants meeting on an online platform (ie, Zoom, Skype; Microsoft, etc) and having a facilitated discussion in real-time [22]. Online delivery for focus groups reduces costs for researchers since transportation and renting space are no longer required. The logistics of arranging meetings with several groups of participants are also simplified since it is easier to set up meetings online than in-person, which speeds up data collection and reduces participant burden [23-26]. There is also more flexibility around meeting times and around canceling and rescheduling for online delivery [18]. Participants living great distances apart can be brought together with the use of an online platform which may allow for the collection of more diverse information [27]. In addition, researcher bias can be reduced, and participants may feel less pressure to conform to other opinions expressed in the group since they are not in the same physical space [23,28]. Some richness of information may be lost, however, due to body language being harder to communicate and capture over online platforms [22].

More academic literature was produced on the use of online data collection methods during COVID-19. In 2 papers reflecting on online focus groups, it was determined that this method is feasible for different populations and that sufficient interaction can be achieved with a skilled moderator [29,30]. Coming out of the COVID-19 pandemic, Indigenous researchers are reimagining Indigenous data collection approaches for online spaces, such as Cooms et al's [31] exploration of online yarning using Facebook, an Indigenist research method belonging to First Nations in Australia. This survey adds more to the literature concerning the use of synchronous online focus groups from the perspective of Métis people, one that had yet to be explored.

Methods

Overview

This voluntary internet-based survey used a convenience sample and recruited respondents via a 'thank you' email sent to participants at the end of synchronous online focus groups conducted for the MNO "Understanding Housing and Health" project between August 2022 and February 2023.

Summary of the "Understanding Housing and Health Project"

The "Understanding Housing and Health" project was meant to better understand the relationship between housing and health for MNO citizens and to provide important feedback to improve the MNO's existing housing programs [32]. A purposive sample of participants was recruited using the MNO's social media

XSL•FO
RenderX

channels and website. Eligibility criteria included being (1) a registered MNO citizen, (2) a current resident within Ontario, and (3) 18 years of age or older. A total of 7 online focus groups were conducted, with 3‑9 participants per group and 35 participants overall. The average time for the focus groups was not reported. All focus groups were recorded using the Zoom record feature, with consent from participants, and transcript summaries were shared back for participant verification. Participants were emailed the questions and Zoom 101 instructions before their scheduled online focus group. Before starting the focus groups, an MNO Senator would share a prayer to start the group in a good way, and both AJS and NT would introduce themselves and the research, including the purpose and goals. The MNO Senator would also close the group with a prayer and offer their reflections on the discussion. The focus group guide was created and pilot tested within the research team.

Eligibility Criteria

Only participants who had completed an online focus group for the aforementioned study were eligible to complete this survey via Qualtrics (Qualtrics International Inc) [33].

Survey Design and Structure

Respondents were asked 20 questions total (15 Likert, 1 open textbox response, and 4 demographic and personal information). The Likert style questions were either 3- or 5-point and asked respondents to rate their experience, satisfaction, preference, feasibility, and the cultural appropriateness of the online focus groups using the Zoom Video Conference tool. An open textbox was included for respondents to share any other thoughts they had about their online focus group experience. The survey was planned to take 5‑10 minutes to complete. Around 1-5 questions were displayed per page over 7 pages (including a thank you message at the end). All questions were forced responses, except for the open textbox and nonresponse options were not provided. Adaptive questioning was not needed. The survey questions were based on a satisfaction survey distributed to Black sexual minority men after completing an online synchronous focus group on HIV prevention conducted by Dangerfield II et al [34] with additions made by MNO staff to account for cultural relevance for Métis citizens. Demographic and personal information (age, gender, MNO Region, and email) was also collected. The MNO has 9 regions that divide up the province of Ontario and represent administrative borders. It is the MNO's preference that the location be reported by the MNO Region. A copy of the survey is available in the Multimedia Appendix 1. The CHERRIES (Checklist for Reporting Results of Internet E-Surveys) checklist was used to report the findings of this survey, and the study design domain of COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist was used to summarize select information from the qualitative "Understanding Housing and Health" project [35,36].

Métis Nation of Ontario and ICES Data Sharing Agreement

This survey was created through collaboration between the MNO and ICES (formerly the Institute for Clinical and Evaluative Sciences) seconded staff. This research relationship works to advance the data sovereignty of MNO citizens and produces important information about citizens' health and wellness needs for the MNO [37]. As part of this agreement, 2 epidemiologists (AJS and NT) and a staff scientist (SAE) are seconded from ICES to the MNO and work exclusively on Métis health and wellness research at the direction of the MNO. The data from this project is not held at ICES and is kept on MNO secure servers. As the representative body of Métis in Ontario, the MNO conducts research on behalf of its citizens with the purpose of closing health and wellness knowledge and outcome gaps.

Data Analysis

Survey data were analyzed using descriptive statistics. An inductive thematic analysis of the open textbox responses was conducted by coders AJS and NT using Microsoft Excel. Both coders independently coded the responses and discussed the themes they observed. These themes were defined, and a codebook was created that was applied to all responses. There was a total of 14 open textbox responses. Three nonspecific answers, "no, all good," "nothing," and "N/A," were not included in the analysis, giving a total of 11 open textbox responses for the content analysis.

Reflexivity

The research team for the bigger "Understanding Housing and Health" project was comprised of Master of Public Health students, experienced PhD public health researchers, topic area specialists, community members, MNO staff, and MNO Senators. AJS and NT, who both identify as women, were Master of Public Health students during this project and took turns facilitating or taking notes for the focus groups. Before conducting the online focus groups, they closely reviewed the literature and sat in on a focus group to gain observational experience. AJS is also Métis and an MNO citizen.

Ethical Considerations

This survey was approved by the MNO and the University of Toronto Health Sciences Research Ethics Board (RIS Human Protocol Number 42320). One informed consent form was read and signed by all participants for both the online focus groups and this survey. This was done before completing a screening survey for the "Understanding Housing and Health" project. The informed consent form did not describe the length of the survey, but this information was shared during the online focus groups. The form also states that only deidentified information would be used for analysis and stored on a secure and confidential server. The primary investigator was identified, and a contact phone number was provided. Both the screening survey and this internet-based survey were delivered using Qualtrics survey software (Qualtrics International Inc) [33]. It was made clear in this consent process that respondents could stop participating at any time for any reason without consequence. Respondents who completed the survey received a CA$5 (US$3.49) gift card for their time.

XSL•FO
RenderX

Results

Participant Characteristics

Thirty-three of 35 eligible respondents completed the survey (94% completion rate) and 34 of 35 started the survey (97% participation rate). Only results from unique visitors (unique name and IP address) were included in the analysis. Three responses were deleted from the dataset because they were either multiple entries (2) or due to the lack of completeness (>20%; 1) (Figure 1). We did not collect or receive reasons why participants chose not to complete the survey.

Figure 1. Flow chart depicting the number of participants invited to complete the voluntary internet-based survey about their online focus group experience, the initial number of responses, the number of excluded responses, and the total number of responses.

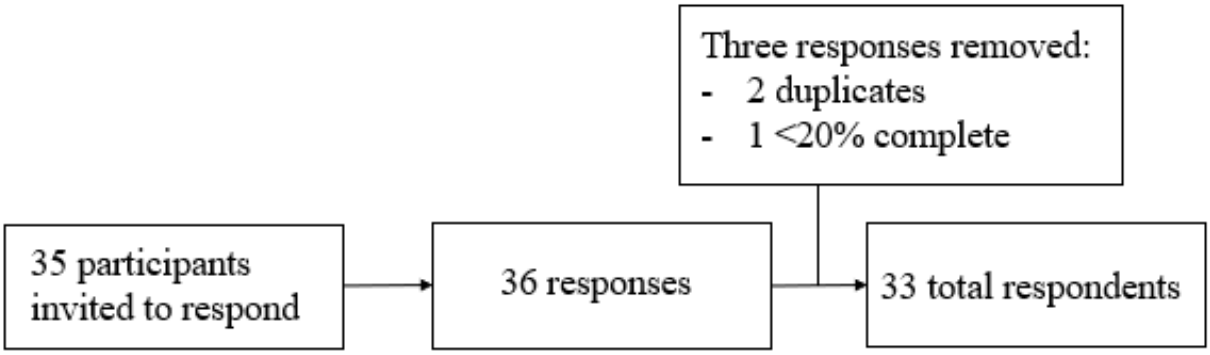

Most respondents (n=24, 75%) identified as cis-women, and 58% (n=19) were 45 - 65 years and older. Additional gender identities included in the survey included trans woman, trans man, genderqueer, genderfluid, androgynous, and nonbinary, but no respondents selected these options. Twenty-one percent (n=7) of respondents selected "identity not listed" and provided a response in an open textbox. One open textbox response was left blank, while other respondents answered either "woman," "man," or "female". These responses were added to the "cis-woman" or "cis-man" counts and indicate that the terms "cis-woman" and "cis-man" were not readily understood by respondents. There are 9 MNO Regions in Ontario. For the MNO Region, 21% (n=7) of respondents indicated they were from Region 7, and 18% (n=6) indicated they were from Region 6. Both regions are in Southern Ontario and Region 7 has the highest population of MNO citizens. A detailed breakdown of MNO Regions is not provided in Table 1 since fewer than 6 participants selected any of the other MNO Regions. Counts less than 5 are either not reportable or are suppressed to minimize the risk of reidentification. This is also why "cis-man" and Two-Spirit are combined in Table 1.

Table . This table provides demographic information on the self-reported age and gender by Métis online focus group respondents for our acceptability and suitability survey on the use of online focus groups for this project. Participant Characteristics.

Participant characteristics	Count (N=33), n (%)
Age (years)	
18 - 24	0 (0)
25 - 44	14 (42)
45 - 65+	19 (58)
Gender[a]	
Cis-woman	24 (75)
Cis-man and Two-Spirit	8 (25)

[a]One participant did not indicate gender

Participant Zoom Video Conference Tool Experience

All respondents had prior Zoom experience, and 70% (n=23) of respondents indicated that they used 'Zoom regularly for work, school, personal use, etc' (see Table 2). The majority (n=28, 85%) were also strongly confident 64% (n=21) or somewhat confident 21% (n=7) in their ability to use Zoom.

XSL•FO
RenderX

Table . This table summarizes responses to our survey questions exploring respondents' Zoom conference platform experience and confidence.

Question	Response option	Participants, n (%)
Please select the statement that best describes your use of the Zoom Video Conference Tool	• This was my first time using Zoom • I have used Zoom before, but I do not use Zoom regularly • I use Zoom regularly for work, school, personal use, etc	• 0 (0) • 10 (30) • 23 (70)
I am confident in my ability to use the Zoom Video Conference tool	• Strongly agree • Somewhat agree • Neither agree nor disagree • Somewhat disagree • Strongly disagree	• 21 (64) • 7 (21) • 1 (3) • 3 (9) • 1 (3)

Acceptability of Focus Groups

Seventy-six percent (n=25) of respondents strongly agreed that they were satisfied with participating in an online focus group (see Table 3). Fifty-two percent (n=17) strongly disagreed that they would have preferred to participate in an in-person focus group, although 27% (n=9) of respondents were neutral for this statement, and 15% (n=5) would have preferred an in-person focus group. All respondents strongly agreed 88% (n=29) or somewhat agreed 12% (n=4) that they would want to participate in an online focus group in the future. Seventy-three percent (n=24) of respondents strongly agreed that it was more feasible for them to participate in an online focus group than an in-person focus group. The majority of respondents strongly (n=22, 67%) or somewhat (n=7, 21%) agreed that it was culturally appropriate to conduct online focus groups with MNO citizens.

Eighty-eight percent (n=29) of respondents strongly agreed or somewhat agreed that they believed the research team would keep their information confidential, while 76% (n=25) of respondents strongly or somewhat agreed that other participants would keep their information confidential (see Table 3). Almost all 94% (n=31) of respondents strongly agreed and somewhat agreed that they believed other participants were in a private space for the online focus groups. Finally, just over three-quarters of respondents strongly agreed or somewhat agreed that they were satisfied with their compensation for participating in an online focus group.

XSL•FO
RenderX

Table . This table summarizes responses to our survey questions exploring participant satisfaction with the online focus group method, feasibility, cultural appropriateness, compensation, and participant privacy.

Question, total (N=33)	Response option, n (%)				
	Strongly agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Strongly disagree
I was satisfied with participating in a focus group online	25 (76)	6 (18)	1 (3)	0 (0)	1 (3)
I would have preferred to participate in the focus group in-person	0 (0)	5 (15)	9 (27)	2 (6)	17 (52)
In the future, I would like to participate in other online focus groups	29 (88)	4 (12)	0 (0)	0 (0)	0 (0)
It is more feasible for me to participate in an online focus group than in an in-person focus group	24 (73)	3 (9)	6 (18)	0 (0)	0 (0)
I believe it is culturally appropriate to use online focus groups for research involving citizens of the Métis Nation of Ontario	22 (67)	7 (21)	2 (6)	1 (3)	1 (3)
I felt that my information will be kept confidential by the research team	24 (73)	5 (15)	3 (9)	0 (0)	1 (3)
I believe other participants were in a private space	21 (64)	10 (30)	1 (3)	1 (3)	0 (0)
I believe my information will be kept confidential by the other people in the focus group	18 (55)	7 (21)	6 (18)	2 (6)	0 (0)
I was satisfied with the compensation I received for participating in the study	22 (67)	4 (12)	5 (15)	1 (3)	1 (3)

Suitability of Online Focus Groups using the Zoom Video Conference Tool

Only 58% (n=19) of respondents could see other participants all of the time (video quality). However, 76% (n=25) of respondents could hear other participants all of the time (audio quality), and 42% (n=14) did experience some lag in video or audio quality. Finally, just over half of the respondents felt they could connect with other participants during the focus group, while the other half felt less confident that they were connecting with others (see Table 4).

XSL•FO
RenderX

Table . This table summarizes respondents' technology and internet experiences (ie, audio, video, and lag connecting with other participants) to explore the suitability of the online focus group method.

Question, total N=33)	Response option, n (%)				
	All of the time	Most of the time	Some of the time	Rarely	None of the time
I was able to see other participants during the focus group discussion (video quality)	19 (58)	10 (30)	4 (12)	0 (0)	0 (0)
I was able to hear other participants during the focus group discussion (audio quality)	25 (76)	6 (18)	2 (6)	0 (0)	0 (0)
I was able to hear and see other participants without delays in the live feed (lag in audio and video)	19 (58)	11 (33)	3 (9)	0 (0)	0 (0)
I feel that I am able to connect with other participants in a meaningful way during the focus group discussion	17 (52)	6 (18)	9 (27)	1 (3)	0 (0)

Open Textbox Responses

There were 11 open textbox responses analyzed. Three themes were generated from this analysis: *Internet and technology issues*, *Accessibility*, *Structure of the group discussion*, and *Positive feedback*. *Internet and technology issues* included audio distortion, and losing connectivity or having poor connectivity. One respondent shared, "Zoom works best when people agree to be on camera and on audio. If people are just typing and few on video it doesn't make for great discussion."

As facilitators, we chose to allow participants to contribute their ideas in the Zoom chat, particularly if they were having internet or technology issues that prevented video and audio, as this respondent did, "I had to use the chat to communicate because of poor connection on my end. It worked." The previous quote is an example of the *Structure of the group discussion*, which included keeping participants more on topic, having a different opening that doesn't include the Lord's prayer, and enjoying the small group sizes and mixed MNO Regions. Having mixed MNO Regions reduced the likelihood of participants knowing each other based on living in the same MNO Region. *Accessibility* included a request to ensure closed captions are turned on and to supply some technology equipment to support joining a Zoom call. One respondent also shared that joining the focus group virtually is "…easier for me than in-person." A couple *Positive feedback* comments were shared like "great group of people."

Discussion

Principal Findings

To our knowledge, this is the first study to explore the acceptability and suitability of an online focus group data collection method for a Métis health and wellness study. Overall, most aspects of the online focus groups were rated highly for acceptability, with the lowest-rated aspects being around privacy, cultural appropriateness, and ability to connect with others. The only other study to explore methods acceptability with Métis research participants was for blood spot testing for HIV [38]. With the advancement of technology for both the medical field and for health and wellness research, it is important to understand what is and is not acceptable or suitable for Métis people. This question is particularly important in the MNO context, since Ontario is a large province with MNO citizens living in urban, rural, and remote regions, which produces costly and time-consuming barriers for research that all citizens have a right to participate in if they are eligible. Online ways to connect with MNO citizens for research purposes offer the potential to reduce the time, cost, and travel burden for both participants and researchers. This is an important consideration for projects with populations that are similarly spread out, have financial and time restrictions, or that are working with populations that have barriers to transportation [39].

During the COVID-19 pandemic, other groups conducting research with Indigenous Peoples shifted to online data collection methods with more consideration being given to translating Indigenist research approaches to online spaces [14,31]. Moreover, the use of online spaces and technology for data collection may help to better include historically excluded groups from research, like the Métis, as Folk et al [40] discuss in their reflection on using online focus groups to explore physical activity with Black and African American women and children. Similar to our open textbox themes, this group's main lessons learned included participant privacy, technological challenges and time burdens, and participant engagement. Other important lessons they shared include the importance of community-building, flexibility, and understanding social context when conducting online focus groups [40]. Similarly, Lathen and Laestadius [41] shared important insights comparing in-person versus online focus groups conducted during the COVID-19 pandemic from their work with low socio-economic status African American adults. Many of the benefits we see as

XSL•FO
RenderX

researchers for using online spaces for focus groups can create barriers for some participants, such as access to technology, incurring data costs to participate, and needing to spend money on transport to get to a computer or public Wi-Fi, among others. These researchers also reported that everyone in their online focus group (participants, facilitators, and assistants) had a more difficult time building rapport and connecting with others in the group, which is consistent with our survey results [41].

One hundred percent of respondents strongly or somewhat agreed that they would participate in an online focus group in the future, and most respondents strongly (n=22, 67%) or somewhat (n=7, 21%) agreed that an online focus group was culturally appropriate for Métis health research; however, fewer respondents strongly agreed with the latter statement in comparison to other statements of which most were strongly agreed upon. In addition, 82% (n=27) of respondents strongly or somewhat agreed that an online focus group was more feasible for them; however, results were more mixed for the statement asking respondents if they would have preferred to participate in an in-person focus group, with 52% (n=17) strongly disagreeing, 27% (n=9) being neutral, and 15% (n=5) somewhat agreeing. These results contrast somewhat with those found by Dangerfield II et al's [34] work assessing the acceptability of online synchronous focus groups among Black sexual minority men. In their work, all participants agreed or strongly agreed that they were satisfied with an online focus group, while only 94% (n=31) of our respondents were strongly or somewhat satisfied [34]. Moreover, 100% of their participants agreed or strongly agreed that their information would be kept safe by others in their group [34]. For our results, only 76% (n=25) (strongly and somewhat agreed) of participants felt their information would be kept confidential by the group, while 18% (n=6) were neutral, and 6% (n=2) disagreed.

The findings from our open textbox responses may give insight into why respondents felt this way about confidentiality. One respondent replied that they appreciated that the groups were made up of participants from different MNO Regions, so that it was less likely someone would know you. As part of the introduction to the focus group, we included ground rules that asked participants to respect the privacy and confidentiality of others in the group by refraining from repeating information outside of the group. However, we cannot control if this is adhered to, which may result in some distrust between participants in our focus groups and may affect confidence around privacy. Some participants were also having difficulty with audio and video and may have missed some of these instructions, which could lead them to believe they were not covered in enough detail. In the future, these ground rules will be emailed to participants with the questions we send ahead of time, in addition to being covered verbally before the group starts.

For suitability, only 58% (n=19) of participants could see other participants all of the time and did not experience lag at any point, while 76% (n=25) could hear other participants all of the time. Fifty-two percent (n=17) of respondents felt they were able to connect with other participants in a meaningful way. Some respondents elaborated on their technology issues in their open textbox responses. Due to connectivity issues, some participants had no choice but to keep their cameras off and use the chat function to participate during the online focus group. But as 1 participant responded, they felt it would have been a better experience if all participants had their cameras on and contributed to the discussion orally, which they felt would allow for more connection. As such, online focus groups may not be for everyone. Technology issues could be addressed by sending participants headphones ahead of time or arranging for them to use a computer at the nearest MNO office. Internet connectivity, especially in rural and remote areas of Ontario, is more difficult to address.

Similar to other studies, we found that online focus groups were very convenient to implement, since renting space and organizing transportation was replaced by a Zoom link and rescheduling was considerably easier [2,3,42]. The use of online focus groups for research with MNO citizens is acceptable; however, internet and technology issues can affect a participant's ability to fully engage. Considerations around the limitations of traditional ways of sharing and gathering knowledge for the Métis people in online spaces and the ability to create connections with and between participants should also be considered [41].

Limitations

This study has a few limitations. Self-selection bias of participants needs to be considered. Participants who are more accepting of online data collection methods may have been more likely to self-select into a study using them, influencing our survey results. Second, the study topic could also influence the acceptability of using an online data collection method. For example, some research topics can be triggering, such as those that involve discussions related to mental health, and it is more difficult to provide crisis intervention over an online platform, which may make certain conversations more appropriate in-person. Third, this research was conducted during the COVID-19 pandemic when it was not safe to gather in groups and was a driving factor for us choosing an online method in the first place. In Ontario, Canada, COVID-19 continues to circulate, although there are no longer any restrictions on gathering [43]. Participants' responses may have differed if the online focus groups were conducted during a stricter period of the COVID-19 pandemic. However, our research group continues to use online platforms for MNO health and wellness research, and citizens continue to participate. Finally, this survey explored participants' experiences at a high level and did not include an open textbox response for each question, which may have resulted in more detailed insights. In the future, if we select to use online focus groups, we may include time at the end to allow participants to give more detailed feedback about their experience with the method. Although there was repetitiveness in the open textbox responses that aligned with results from our other questions, it is unlikely that data saturation was reached due to the limited number and sparseness of the open textbox responses. Including an acceptability and feasibility survey for other projects that use online group methods will help us better tailor online approaches for research with MNO citizens.

XSL•FO
RenderX

Conclusions

This paper contributes to the growing literature on the use of online focus groups as a method to collect information and stories for health research. Our survey specifically explored the use of this method with the Métis people (MNO citizens) and concludes that it may be acceptable, but important considerations around technology and internet barriers, the ability of the group to connect with each other, and traditional knowledge sharing for the Métis people should be considered. This information is valuable for guiding the selection of methods for future research with the Métis and for other populations facing similar barriers.

Acknowledgments

We are extremely grateful to all Métis Nation of Ontario citizens who took part in this study and the Métis Nation of Ontario leadership and Senators who contributed at each step of the research process. We acknowledge the Canada Mortgage and Housing Corporation (CMHC) for financial support of the project. CMHC had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript. The views expressed herein do not necessarily represent the views of CMHC. We thank the Métis Nation of Ontario for in-kind funding. No part of this manuscript was generated by or with the use of ChatGPT or other AI tools.

This study was funded by the CMHC and in-kind funding from the Métis Nation of Ontario.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to the principles of data sovereignty and governance applied by the Métis Nation of Ontario but may be available from the corresponding author on reasonable request with approval required from the Métis Nation of Ontario.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Acceptability and suitability survey questions on the use of online focus groups for the qualitative “Understanding Housing and Health” project.
[DOCX File, 16 KB - formative_v9i1e70986_app1.docx]

References

1. Simms AJ, King KD, Tsui N, Edwards SA, Mecredy G, Métis Nation of Ontario. COVID-19 vaccine behaviour among citizens of the Métis Nation of Ontario: a qualitative study. Vaccine (Auckl) 2023 Aug 31;41(38):5640-5647. [doi: 10.1016/j.vaccine.2023.07.060] [Medline: 37550144]
2. Dodds S, Hess AC. Adapting research methodology during COVID-19: lessons for transformative service research. JOSM 2021 Feb 8;32(2):203-217. [doi: 10.1108/JOSM-05-2020-0153]
3. Eigege CY, Daundasekara SS, Gomez ML, Walton QL, Hernandez DC. Conducting research six feet apart: the feasibility of transitioning qualitative research to meet the emerging research needs during a pandemic. Int J Qual Methods 2022 Apr;21(6094069211069442):16094069211069442. [doi: 10.1177/16094069211069442]
4. The Métis Nation of Ontario. URL: https://www.metisnation.org/about-the-mno/?doing_wp_cron=1718295417.9019351005554199218750 [accessed 2024-06-06]
5. Métis Nation of Ontario – Fact Sheet: Implementation of the National Definition of Métis within Ontario and the MNO Registry Policy. Métis Nation of Ontario. URL: https://www.metisnation.org/wp-content/uploads/2020/12/fact-sheet-implementation-of-the-national-definition-in-ontario.pdf [accessed 2024-06-06]
6. Culture & heritage. Métis Nation of Ontario. URL: https://www.metisnation.org/culture-heritage/ [accessed 2024-06-06]
7. All MNO citizen online town hall. Métis Nation of Ontario. 2021 Jan 18. URL: https://www.metisnation.org/news/all-mno-citizen-online-town-hall/ [accessed 2024-06-06]
8. Macdougall B. National Collaborating Centre for Aboriginal Health. Land, family and identity: Contextualizing Métis health and well-being. URL: https://www.nccih.ca/docs/context/RPT-ContextualizingMetisHealth-Macdougall-EN.pdf [accessed 2025-07-16]
9. Gaudet JC. Keeoukaywin: the visiting way - fostering an Indigenous research methodology. aboriginal policy stud 2019;7(2). [doi: 10.5663/APX.V7I2.29336]
10. Wright A, Wahoush O, Ballantyne M, Gabel C, Jack S. Qualitative health research involving Indigenous Peoples: culturally appropriate data collection methods. TQR 2016:12. [doi: 10.46743/2160-3715/2016.2384]
11. Smith L. Decolonizing Methodologies: Research and Indigenous Peoples: New York, NY: St. Martin’s Press; 1999.

XSL•FO
RenderX

12. Hunt SC, Young NL. Blending Indigenous sharing circle and Western focus group methodologies for the study of Indigenous children's health: a systematic review. Int J Qual Methods 2021 Jan;20:1-16. [doi: 10.1177/16094069211015112]
13. Sitotaw R, Lulekal E, Abate D. Ethnomycological study of edible and medicinal mushrooms in Menge District, Asossa Zone, Benshangul Gumuz Region, Ethiopia. J Ethnobiol Ethnomed 2020 Mar 4;16(1):11. [doi: 10.1186/s13002-020-00361-9] [Medline: 32131860]
14. Lambrou NH, Blind MJ, Ketcher D, et al. Indigenous cultural understandings of Alzheimer's disease and related dementias – research and engagement (ICARE): relationships between cultural strengths, education and stigma. Alzheimers Dement 2022 Dec;18(S8):e066094. [doi: 10.1002/alz.066094]
15. Jacklin K, Ly A, Calam B, Green M, Walker L, Crowshoe L. An innovative sequential focus group method for investigating diabetes care experiences with Indigenous Peoples in Canada. Int J Qual Methods 2016 Dec 1;15(1):1-12. [doi: 10.1177/1609406916674965]
16. Agudelo-Hernández F, Coral-Vela LP, Pabuena-Yepes LH. Mental health risk communication and community participation among an Indigenous people in Colombia. Rev Panam Salud Publica 2025;49:e12. [doi: 10.26633/RPSP.2025.12] [Medline: 39811045]
17. Andersen MJ, Williamson AB, Fernando P, Redman S, Vincent F. "There's a housing crisis going on in Sydney for Aboriginal people": focus group accounts of housing and perceived associations with health. BMC Public Health 2016 May 24;16(1):429. [doi: 10.1186/s12889-016-3049-2] [Medline: 27220748]
18. Kite J, Phongsavan P. Insights for conducting real-time focus groups online using a web conferencing service. F1000Res 2017;6:122. [doi: 10.12688/f1000research.10427.2]
19. Gray LM, Wong-Wylie G, Rempel GR, Cook K. Expanding qualitative research interviewing strategies: Zoom Video Communications. TQR 2020;24(5):4212. [doi: 10.46743/2160-3715/2020.4212]
20. Halliday M, Mill D, Johnson J, Lee K. Let's talk virtual! Online focus group facilitation for the modern researcher. Res Social Adm Pharm 2021 Dec;17(12):2145-2150. [doi: 10.1016/j.sapharm.2021.02.003] [Medline: 33653681]
21. Menary J, Stetkiewicz S, Nair A, et al. Going virtual: adapting in-person interactive focus groups to the online environment. Emerald Open Research 2021(6). [doi: 10.35241/emeraldopenres.14163.1]
22. Liamputtong P. Chapter 9: virtual focus groups. In: Liamputtong P, editor. Focus Group Methodology: Principles and Practice, 1st edition: SAGE; 2011:149-164.
23. Gaiser TJ. Online focus groups. In: Feilding N, Lee RM, Blank G, editors. The Sage Handbook of Online Research Methods: London: Sage; 2008:290-307.
24. Mann C, Stewart F. Internet Communication and Qualitative Research: A Handbook for Researching Online, 1st edition: London: Sage; 2000.
25. Valaitis RK, Sword WA. Online discussions with pregnant and parenting adolescents: perspectives and possibilities. Health Promot Pract 2005 Oct;6(4):464-471. [doi: 10.1177/1524839904263897] [Medline: 16210689]
26. Liamputtong P. Health Research in Cyberspace: Methodological, Practical and Personal Issues: New York: Nova Science Publishers; 2006.
27. Matthews KL, Baird M, Duchesne G. Using online meeting software to facilitate geographically dispersed focus groups for health workforce research. Qual Health Res 2018 Aug;28(10):1621-1628. [doi: 10.1177/1049732318782167] [Medline: 29911490]
28. Reid DJ, Reid FJM. Online focus groups: an in-depth comparison of computer-mediated and conventional focus group discussions. Int J Market Res 2005 Mar;47(2):131-162. [doi: 10.1177/147078530504700204]
29. Henage CB, Ferreri SP, Schlusser C, et al. Transitioning focus group research to a videoconferencing environment: a descriptive analysis of interactivity. Pharmacy (Basel) 2021 Jun 24;9(3):117. [doi: 10.3390/pharmacy9030117] [Medline: 34202707]
30. Cordell A, Wilk C, Orsulic-Jeras S, Powers S, Ejaz F, Sanders L. The focus group must go on: lessons learned from conducting virtual focus groups. Innov Aging 2021 Dec 17;5(Supplement_1):837-838. [doi: 10.1093/geroni/igab046.3067]
31. Cooms S, Leroy-Dyer S, Muurlink O. The rise of virtual yarning: an Indigenist research method. Qual Res 2025 Feb;25(1):168-188. [doi: 10.1177/14687941241234303]
32. Tsui N, Simms AJ, Lacka C, et al. Exploring the relationship between housing conditions and Métis Nation of Ontario citizen's health: a qualitative study. BMC Public Health 2025 Jul 25;25(1):2545. [doi: 10.1186/s12889-025-23716-x] [Medline: 40713581]
33. Qualtrics.: Provo, UT: Qualtrics; 2023. URL: https://www.qualtrics.com [accessed 2025-09-24]
34. Dangerfield Ii DT, Wylie C, Anderson JN. Conducting virtual, synchronous focus groups among Black sexual minority men: qualitative study. JMIR Public Health Surveill 2021 Feb 8;7(2):e22980. [doi: 10.2196/22980] [Medline: 33427671]
35. Eysenbach G. Improving the quality of Web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34. [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
36. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]

XSL•FO
RenderX

37. ICES and Métis Nation of Ontario partner on new agreement. IC/ES. URL: https://www.ices.on.ca/announcements-and-events/ices-and-metis-nation-of-ontario-partner-on-new-agreement/ [accessed 2024-06-06]
38. Landy R, Atkinson D, Ogilvie K, et al. Assessing the acceptability of dried blood spot testing for HIV and STBBI among Métis people in a community driven pilot project in Alberta, Canada. BMC Health Serv Res 2022 Dec 8;22(1):1496. [doi: 10.1186/s12913-022-08763-z] [Medline: 36482470]
39. Mabragaña M, Carballo-Diéguez A, Giguere R. Young women's experience with using videoconferencing for the assessment of sexual behavior and microbicide use. Telemed J E Health 2013 Nov;19(11):866-871. [doi: 10.1089/tmj.2013.0008] [Medline: 24050616]
40. Folk AL, Grace SM, Urvig M, Barr-Anderson DJ. Virtual focus groups on Zoom: "Lessons Learned" from two physical activity studies among Black and African American women and children. Int J Soc Res Methodol 2025;28(2):209-220. [doi: 10.1080/13645579.2024.2330050] [Medline: 40330134]
41. Lathen L, Laestadius L. Reflections on online focus group research with low socio-economic status African American adults during COVID-19. Int J Qual Methods 2021 Jan;20:20. [doi: 10.1177/16094069211021713]
42. Irani E. The use of videoconferencing for qualitative interviewing: opportunities, challenges, and considerations. Clin Nurs Res 2019 Jan;28(1):3-8. [doi: 10.1177/1054773818803170] [Medline: 30470151]
43. Post-COVID-19 condition in canada: what we know, what we don't know, and a framework for action. Government of Canada. 2022 Dec. URL: https://science.gc.ca/site/science/en/office-chief-science-advisor/initiatives-covid-19/post-covid-19-condition-canada-what-we-know-what-we-dont-know-and-framework-action [accessed 2025-09-02]

Abbreviations

CHERRIES: Checklist for Reporting Results of Internet E-Surveys
COREQ: Consolidated Criteria for Reporting Qualitative Research
ICES: Institute for Clinical and Evaluative Sciences
MNO: Métis Nation of Ontario

Edited by A Mavragani; submitted 07.01.25; peer-reviewed by A Singleton; revised version received 06.08.25; accepted 11.08.25; published 09.10.25.

Please cite as:
Simms AJ, Tsui N, Sadler R, Mersereau A, Rye C, Crenna CD, Evenson J, Edwards SA
Exploring the Acceptability and Suitability of Synchronous Online Focus Groups for Health Research With Métis Nation of Ontario Citizens: An Internet-Based Survey
JMIR Form Res 2025;9:e70986
URL: https://formative.jmir.org/2025/1/e70986
doi:10.2196/70986

© Abigail J Simms, Noel Tsui, Robynn Sadler, Amy Mersereau, Cindi Rye, Charles David Crenna, Jeff Evenson, Sarah A Edwards. Originally published in JMIR Formative Research (https://formative.jmir.org), 9.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Noninvasive Multiparameter Monitoring for the Detection of Decompensated Heart Failure: Exploratory Study

Cyrille Herkert[1], MD, PhD; Mayke van Leunen[1], MD; Ignace Luc Johan De Lathauwer[1], MD; Valerie Albertina Antonetta van Es[2], MSc; Jialu Tang[3], MSc; Aaqib Saeed[3], PhD; Rudolph Ferdinand Spee[1], MD, PhD; Yuan Lu[3], PhD; Hareld Marijn Clemens Kemps[1,3], MD, Prof Dr

[1]Department of Cardiology, Máxima Medical Centre, Dominee Theodor Fliednerstraat 1, Eindhoven, The Netherlands

[2]MoMiLab Research Unit, IMT School for Advanced Studies Lucca, Lucca, Italy

[3]Department of Industrial Design, Eindhoven University of Technology, Eindhoven, The Netherlands

Corresponding Author:
Ignace Luc Johan De Lathauwer, MD
Department of Cardiology, Máxima Medical Centre, Dominee Theodor Fliednerstraat 1, Eindhoven, The Netherlands

Abstract

Background: Remote patient monitoring strategies in patients with heart failure (HF) are often based on manual readings and interpretation of various parameters by health care professionals. Automated multiparameter predictive models (MPMs) have the potential to improve early recognition of decompensated HF and to reduce the workload for both health care professionals and patients. To reduce costs and facilitate large-scale implementation, these models should preferably be based on noninvasive measurements, with user-friendly devices.

Objective: This exploratory study aimed to evaluate whether an MPM, using various parameters from a wrist-worn device supplied with a photoplethysmography sensor and a triaxial accelerometer, contributes to the detection of decompensated HF and death in patients with unstable HF.

Methods: Patients who were admitted to the hospital with acute decompensated HF, regardless of etiology or left ventricular ejection fraction, were instructed to wear a research-grade wrist-worn device from the moment of discharge. The device measured heart rate (HR), interbeat intervals (IBIs), respiration rate (RR), activity counts (AC), energy expenditure (EE), and sleep. Participants were instructed to wear the device 24 hours a day for 3 consecutive months. We evaluated 7 classifiers under four strategies for handling extreme class imbalance; the best model was then tested via leave-one-subject-out cross-validation on untouched data. The combined end point of interest was hospital readmission due to decompensated HF, decompensated HF treated at the outpatient clinic by increasing the loop diuretic dose, or death due to HF.

Results: A total of 17 patients participated in the study (median age 77, IQR 70‑84 y; n=9, 53% male). During follow-up, the device-wearing compliance was 78% (55%‑81%). The activity-related parameters (EE and AC) performed best with respect to data quality: 72% and 79% of the data were of high quality, respectively. Concerning HR, 46% of the data were of high quality, whereas only 29% of the IBI and 14% of the RR data were of high quality. Sleep data were lacking 99% of the time during follow-up, resulting in exclusion from training the classifier. The most optimal model for the detection of the combined end point of HF deterioration showed a specificity of 97.2% and a sensitivity of 5.3% in the 2 weeks prior to an event (area under the curve=0.59) after leave-one-subject-out cross-validation analysis.

Conclusions: An MPM using a noninvasive wrist-worn device, measuring HR, IBI, RR, AC, and EE, showed high specificity but low sensitivity for the prediction of decompensated HF and HF-related mortality. Low sensitivity likely reflects the extreme class imbalance and sequences with low data quality (especially HR, RR, and sleep), resulting in exclusion from training the MPM in our older, real-world HF cohort. Future studies should improve data fidelity and enroll larger cohorts to address class imbalance and enhance predictive performance.

Trial Registration: Netherlands Trial Register NL9038; https://www.onderzoekmetmensen.nl/en/trial/21559

(JMIR Form Res 2025;9:e59116) doi:10.2196/59116

KEYWORDS

heart failure; remote patient monitoring; predictive model; photoplethysmography; accelerometry

XSL•FO
RenderX

Introduction

Heart failure (HF) is a highly common clinical syndrome, affecting over 64 million people worldwide [1]. This amount is expected to increase by 46% between 2012 and 2030 [1,2], due to a combination of factors such as aging of the population and optimization of diagnostic and therapeutic management. Once being admitted to the hospital for decompensated HF, the 30-day all-cause readmission rate in patients with HF is up to 22% [3], accounting for the majority of the annual €25,000 (US $28,974) health care costs per patient with HF in the Western world [1].

In order to optimize HF care and to reduce the risk of readmission in this patient population, remote patient monitoring (RPM) strategies have been developed. It is generally conceived that monitoring the patients' vital parameters and symptoms in the home environment enables early detection of deterioration, timely up-titration of HF medication, and improvement of self-management skills. Nevertheless, meta-analyses on noninvasive RPM show contradicting results [4-9]. Besides heterogeneity in study populations and end points, this may be explained by a large variety in sensor technology and (lack of) algorithms that were used.

In order to improve noninvasive sensor technology for the prediction of HF decompensation, it should comprise several characteristics. First, to improve the chances of preventing hospitalization by increasing the time frame to adjust medical therapy, sensors should detect signs of HF decompensation as early as possible. The first pathophysiological changes that lead to decompensated HF appear around 3 to 4 weeks before hospitalization due to decompensated HF, starting with an increase in cardiac filling pressures, followed by autonomic changes, and later, intrathoracic impedance changes [10]. Traditional RPM parameters, such as body weight or symptoms, change only a few days to a week before hospitalization [10]. Although previous research demonstrated that combining these traditional parameters into more complex, patient-tailored algorithms already increases the accuracy of predicting HF-related events [11,12], further improvement may be achieved by adding parameters that change earlier in the pathophysiological cascade of HF deterioration. As such, heart rate variability (HRV), a marker of autonomic adaptation, was shown to decrease at least 2 weeks prior to hospitalization [10,13]. Moreover, as the autonomic nervous system is involved in regulating the sleep-wake system [14], an increase in autonomic imbalance starting 2 weeks prior to hospital admission might alter sleep duration and patterns. Monitoring of sleep behavior might therefore also improve early detection of HF deterioration. Additionally, respiration rate (RR) increases as a result of fluid retention, but more importantly, sleep-related respiration patterns are altered as a result of autonomic imbalance [15]. Therefore, RR might also add value in the early detection of decompensated HF.

Second, current RPM programs mostly rely on daily spot measurements of several parameters. Continuous monitoring of parameters such as RR or heart rate (HR) might improve early prediction of decompensated HF, because it increases the ability to observe trends in these parameters.

Third, the use of a single device that continuously measures multiple parameters without the patient needing to take additional actions, rather than the use of multiple devices, might increase adherence to an RPM intervention, which attributes to its success. Whereas previous studies using multiparameter predictive models (MPMs) based on a single invasive device showed good diagnostic accuracy [16-18], their practical utility is limited due to the invasive nature of these devices. The LINK-HF study demonstrated that an MPM (including HR, HRV, arrhythmia burden, RR, gross activity, sleep, body tilt, and posture) using a chest patch could accurately predict the risk of rehospitalization for HF [19]. Wrist-worn devices, enabling monitoring of similar parameters, could add to this concept as they are widely accepted, can provide the patient with direct feedback, and are sustainable. To our knowledge, there are no studies available that evaluated an MPM to predict decompensated HF, based on different parameters obtained from a wrist-worn device.

This study aims to explore the predictive value of continuous data from a wrist-worn device (HR-related data, RR, activity, and sleep) to predict decompensated HF and death in older patients with unstable HF.

Methods

Study Protocol

This study aimed to develop an MPM, using various parameters from a wrist-worn device (Philips Data Logger [PDL]; Philips Electronics Nederland B.V.), for the detection of a combined end point: hospital readmission due to decompensated HF, decompensated HF treated at the outpatient clinic by increasing the dose of loop diuretics, or death due to HF. Patients who were admitted to the hospital with acute decompensated HF were asked for study participation. All participants were instructed to wear a PDL for a 3-month period after hospital discharge.

Study Population

Adults (aged 18 y or older) with acute decompensated HF, who were admitted to the cardiology department of the Máxima Medical Center, Veldhoven, the Netherlands, were included in this study. Patients were included regardless of HF etiology or left ventricular ejection fraction (LVEF). Exclusion criteria were permanent atrial fibrillation (as sufficient data quality could not be assured in patients with permanent atrial fibrillation), not being able to perform daily physical activities (eg, due to an orthopedic or a neurological condition), not being able to wear the PDL due to a skin condition, and not having a Wi-Fi connection at home.

Ethical Considerations

All patients provided written informed consent. The study was reviewed by the local medical ethics committee but received a waiver that medical ethics approval was not required. The study was registered at the Netherlands Trial Register (NL9038) and

XSL•FO
RenderX

was performed in accordance with the principles of the Declaration of Helsinki.

Philips Data Logger

The PDL is a wrist-worn, research-grade monitoring device measuring photoplethysmography (PPG)- and accelerometry-derived data (Figure 1A). Participants were blinded to their recorded parameters and did not receive feedback based on the measurements. The PDL is rain and splash resistant, but cannot be worn during showering or swimming. Data were offloaded from the PDL via Bluetooth as soon as the device was within reach of a gateway (Monitoring Study boX [MSX]; Figure 1B). The data were uploaded automatically from the MSX to the Philips Research cloud server via a 4G network. Parameters measured by the PDL that were used for analysis were: activity counts (AC), energy expenditure (EE), HR-related data such as HR and interbeat intervals (IBIs), RR, and sleep data (sleep stages and quality). All incoming data were classified as low, medium, or high quality by the software of Philips. This classification was used to evaluate the quality of the collected data over a 3-month period within this study population.

Figure 1. (A) Philips Data Logger. (B) Monitoring Study box.

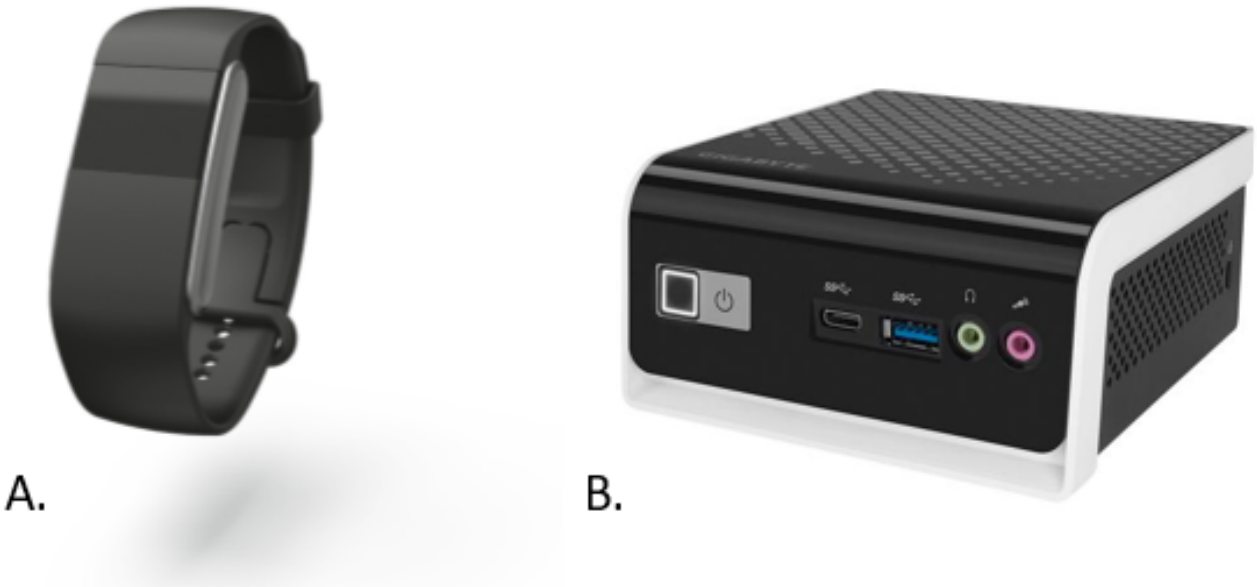

Study Procedures

At hospital discharge (ie, in the recompensated state), the participants received a PDL together with an MSX. Instructions for use were given by the local investigators (CH, ML, and ILJL). Participants were asked to wear the PDL 24 hours per day (except when taking a shower or swimming), on the same wrist, for 3 consecutive months after hospital discharge. For every participant, the location of the PDL was registered (ie, right or left wrist), as well as the handedness of the participant. When data were lacking for a period of 3 days, the local investigators received a warning and contacted the participant to discuss and solve potential problems. During study participation, the local investigators periodically checked the participants' medical records to report the events of interest, which were: a hospital readmission due to decompensated HF, an increase of loop diuretic dose at the outpatient clinic due to decompensated HF, and death, as expected to be the result of HF. An HF readmission is defined as a hospitalization resulting from an acute exacerbation of HF symptoms, primarily due to elevated ventricular filling pressures as a result of volume overload caused by the heart's impaired ability to effectively pump blood [20]. The study was completed after the participant had worn the PDL for 3 months after hospital discharge, regardless of the encounter of a recurrent episode of decompensated HF.

Data Preparation

In addition to the six physiological parameters collected by the PDL (AC, EE, HR, IBI, RR, and sleep), two extra parameters, related to wearing behavior, were included in the MPM. The first is "skin proximity," a parameter calculated by Philips' Software, revealing the distance between the skin and the device, varying from the device being worn properly to the device being off-skin. The second parameter, "fundamental," indicates the proportion of the recorded data when the device was both powered on and properly worn to the recorded data of the entire period the device was powered on.

Eventually, all the collected parameters can be grouped as follows: (1) the parameters related to wearing behavior: skin proximity and fundamental; (2) the physiological parameters: AC, EE, HR, IBI, RR, and sleep stages and quality. In particular, HRV was calculated using the SD of all IBIs per 24 hours [21].

Only data segments labeled as high quality by the Philips software were included in the MPM. All data were split into 14-day segments, resulting in a total of 1218 time series. For each parameter time series, the mean, SD, minimum, and maximum values were extracted, yielding a total of 32 features used for model training.

These features were labeled into two distinct classes:

XSL•FO
RenderX

- Clinically stable, representing instances without the events of interest
- HF events, denoting instances with the events of interest (eg, decompensated HF treated at the outpatient clinic or leading to a hospital readmission, or death due to HF).

After labeling, the dataset consisted of 1209 instances of patients being clinically stable and 9 instances with HF events, indicating a substantial class imbalance.

Multiparameter Predictive Model Training and Data Analysis

To address the extreme class imbalance in the dataset (1209:9, clinically stable:HF event), four modeling strategies were evaluated: baseline (unaltered data), random undersampling of the majority class, random oversampling of the minority class, and cost-sensitive learning that penalized false negatives more heavily. These strategies were applied across seven classifiers: decision tree, random forest (100 trees), gradient boosting (LogitBoost implementation), one-class support vector machine (RBF kernel), support vector classifier (RBF kernel), logistic regression (LASSO-regularized), and multilayer perceptron (single hidden layer with Rectified Linear Unit activation). Class weights or sampling adjustments were applied within each strategy depending on the model's compatibility. Model training and selection were performed using stratified 5-fold cross-validation. The most promising model configuration was further validated using leave-one-subject-out cross-validation (LOOCV), ensuring that only real, unaltered data from the left-out participant was used for testing [22].

Since the included study population was diverse, including patients with different HF etiologies and LVEF, and the relatively small dataset, LOOCV allowed us to use as much data as possible for training the model. This method can reduce the bias and variance of the test error compared to using a single test set. With LOOCV, we anticipated learning how the predictive model performed for each patient and whether there were outlier participants. To evaluate the model's performance, several key metrics were used, including the sensitivity, specificity, and area under the curve (AUC). Feature importance was evaluated using the out-of-bag permutation method for random forest, quantifying each feature's contribution to model accuracy. Feature importances were aggregated by physiological modality (HR, IBI, RR, EE, AC, and wearing behavior) and normalized to 100% for interpretability.

Descriptive statistics were used to describe the baseline characteristics of the participants, the follow-up duration, and the occurrence of events.

Mann-Whitney *U* tests were performed to analyze the differences in physiological parameters (HR, IBI, RR, EE, and AC) between time series during the stable phase and time series preceding an HF decompensation event. Time series analysis was performed to reveal trend changes in the physiological parameters over a 4-week period before a HF decompensation event. An overview of the study aims and methods is provided in Figure 2.

XSL•FO
RenderX

Figure 2. Overview of study aim and methods. ADHF: acute decompensated heart failure; AF: atrial fibrillation; EE: energy expenditure; HF: heart failure; HR: heart rate; HRV: heart rate variability; IBI: interbeat intervals; LOOCV: leave-one-out cross-validation; MPM: multiparameter prediction model; PA: physical activity; RR: respiration rate.

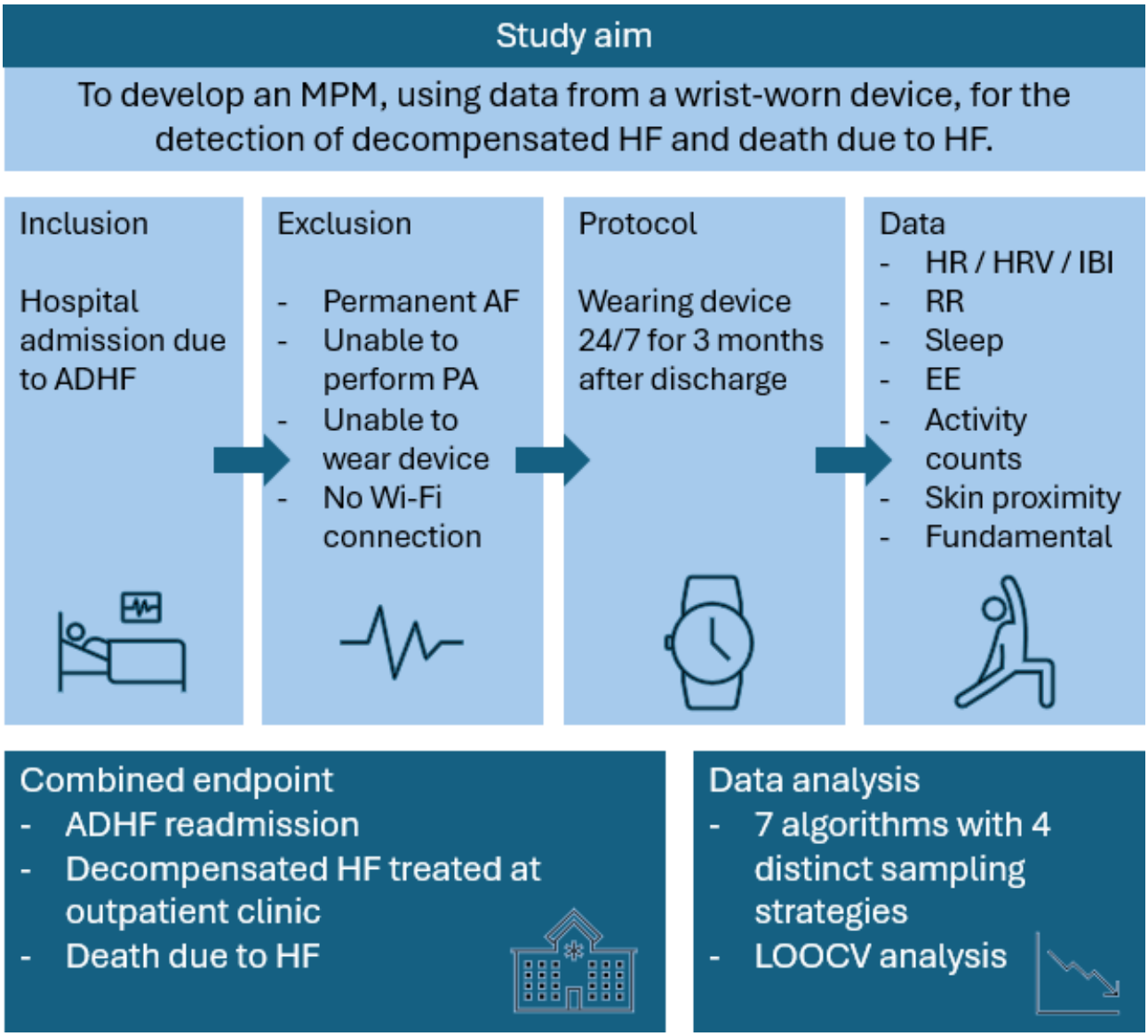

Results

Patient Characteristics

A total of 24 patients participated in the study; however, only 17 were included in the primary analysis. The reason for excluding seven participants was: one participant signed informed consent but withdrew from participation before hospital discharge; two participants experienced study participation as too demanding and decided to withdraw from participation shortly after inclusion; and in one participant, the data transfer was restricted due to technical problems, resulting in zero data transfer. Moreover, three additional patients were excluded from analysis, as they experienced an HF-related event (two hospital admissions due to decompensated HF and one HF decompensation treated at the outpatient clinic) within a week after inclusion and did not continue the study afterwards. As the aim was to observe trends in the data prior to an event, only patients with at least 2 weeks of data before a HF-related event were included. The patient characteristics of the remaining 17 participants are presented in Table 1.

XSL•FO
RenderX

Table . Patient characteristics.

Characteristics	Value (n=17)
Age (years), median (IQR)	77 (70 - 84)
Sex, n (%)	
Male	9 (53)
Female	8 (47)
Height (m), median (IQR)	1.69 (1.61 - 1.74)
Weight (kg), median (IQR)	80.6 (67.5 - 94.5)
BMI (kg/m^2), median (IQR)	30.7 (23.7 - 32.3)
Race, n (%)	
Caucasian	17 (100)
Non-Caucasian	0 (0)
LVEF[a] (%), median (IQR)	53 (43.5 - 60.0)
HF[b] type, n (%)	
HFrEF[c]	3 (17.6)
HFmrEF[d]	3 (17.6)
HFpEF[e]	11 (64.7)
Device therapy, n (%)	
Pacemaker	1 (6)
CRT(P/D)[f]	0 (0)
ICD[g]	0 (0)
Comorbidities, n (%)	
Paroxysmal AF[h]	9 (53)
Hypertension	11 (65)
Diabetes mellitus	7 (41)
Chronic kidney disease	5 (29)
Neurological history	5 (29)
Malignancy	4 (24)
Asthma or COPD[i]	2 (12)

[a]LVEF: left ventricular ejection fraction.
[b]HF: heart failure.
[c]HFrEF: heart failure with reduced ejection fraction.
[d]HFmrEF: heart failure with mildly reduced ejection fraction.
[e]HFpEF: heart failure with preserved ejection fraction.
[f]CRT(P/D): cardiac resynchronization therapy with either a pacemaker or a defibrillator.
[g]ICD: implantable cardioverter defibrillator.
[h]AF: atrial fibrillation.
[i]COPD: chronic obstructive pulmonary disease.

Follow-Up and Events

The median follow-up duration of the 17 included participants was 92 (IQR 88 - 93) days. Reasons for the premature ending of the study were: death (one participant), the start of palliative care resulting in the cancellation of further follow-up at the hospital (one participant), and skin irritation caused by the device (one participant). During follow-up, the compliance with wearing the device was 78% (55% - 81%). Nine events occurred during follow-up: seven times an increase of loop diuretic dose at the outpatient cardiology clinic due to decompensated HF, one hospital readmission due to decompensated HF, and one participant died due to worsening of HF. The median time from inclusion to the first event was 36 (IQR 22-76) days.

XSL•FO
RenderX

Data Quality

The accelerometry-related parameters (EE and AC) performed best with respect to data quality: averaged per day, 72% and 79% of the data were of high quality, respectively. Concerning HR, 46% of the daily data were of high quality; 29% of the IBI (based on which HRV was calculated) data were of high quality. Only 14% of the RR data were of high quality, and 65% of this data were lacking. Sleep data were lacking 99% of the time during follow-up. Therefore, sleep features were excluded from the analysis due to nearly complete data absence.

Performance Predictive Model

Table 2 presents the performance of 7 machine learning classifiers under four training strategies, evaluated using stratified 5-fold cross-validation. Ensemble models, particularly random forest and gradient boosting, achieved the highest AUC values (up to 0.99) when combined with oversampling or cost-sensitive training. Sensitivity ranged from 0% in baseline models such as logistic regression and multilayer perceptron to 96% in the best-performing model (random forest with oversampling), with an AUC of 0.96. Specificity remained high across most configurations, with values reaching up to 99.8%. One-class support vector machine and support vector classifier demonstrated limited sensitivity despite high specificity. These findings highlight the superiority of ensemble methods under imbalance-corrected training and the necessity of such strategies for improving predictive performance in small, event-sparse datasets.

XSL•FO
RenderX

Table . Comparative results of 5 different predictive models.

Strategy	Model	Sensitivity, mean (95% CI)	Specificity, mean (95% CI)	AUC[a], mean (95% CI)	NPV[b], mean (95% CI)	PPV[c], mean (95% CI)
Baseline	DecisionTree	7.14 (4.28-10.00)	98.1 (96.7-99.5)	0.50 (0.45-0.55)	99.59 (94.61-100)	1.7 (0.0-6.14)
Baseline	RandomForest	51.43 (47.86-55.00)	95.0 (94.0-96.0)	0.70 (0.66-0.74)	99.73 (94.74-100)	10.0 (9.0-11.0)
Baseline	GradientBoosting	55.0 (45.5-64.5)	98.0 (97.0-99.0)	0.72 (0.67-0.77)	99.73 (94.74-100)	10.0 (9.0-11.0)
Baseline	OneClassSVM	0.0 (0.0-0.0)	70.0 (69.0-71.0)	0.67 (0.62-0.72)	99.05 (94.10-100)	0.0 (0.0-0.0)
Baseline	MLP[d]	0.0 (0.0-0.0)	63.0 (62.0-64.0)	0.6 (0.57-0.63)	99.05 (94.10-100)	0.0 (0.0-0.0)
Baseline	LR[e]	0.0 (0.0-0.0)	63.0 (62.0-64.0)	0.6 (0.57-0.63)	99.05 (94.10-100)	0.0 (0.0-0.0)
Baseline	SVC[f]	0.0 (0.0-0.0)	80.0 (79.0-81.0)	0.98(0.93-1.00)	99.05 (94.10-100)	0.0 (0.0-0.0)
Undersample	DecisionTree	99.0 (98.0-100.0)	37.81 (35.92-39.70)	0.74 (0.70-0.78)	99.0 (98.0-100.0)	1.52 (1.44-1.60)
Undersample	RandomForest	99.0 (98.0-100.0)	43.97 (41.77-46.17)	0.91 (0.86-0.96)	99.0 (98.0-100.0)	1.68 (1.60-1.76)
Undersample	GradientBoosting	99.0 (98.0-100.0)	35.41 (33.64-37.18)	0.9 (0.85-0.95)	99.0 (88.0-100.0)	1.46 (1.39-1.53)
Undersample	OneClassSVM[g]	99.0 (98.0-100.0)	91.92 (87.32-96.52)	0.93 (0.88-0.98)	99.0 (98.0-100.0)	10.61 (10.08-11.14)
Undersample	MLP	0.0 (0.0-0.0)	99.0 (98.0-100.0)	0.6 (0.57-0.63)	99.05 (94.10-100.0)	0.0 (0.0-0.0)
Undersample	LR	0.0 (0.0-0.0)	99.0 (98.0-100.0)	0.6 (0.57-0.63)	99.05 (94.10-100)	0.0 (0.0-0.0)
Undersample	SVC	92.86 (88.22-97.50)	58.63 (55.70-61.56)	0.83 (0.79-0.87)	99.88 (94.89-100)	2.11 (2.00-2.22)
Oversample	DecisionTree	92.0 (88.5-95.5)	89.5 (85.3-93.7)	0.94 (0.90-0.98)	98.7 (97.6-99.8)	7.8 (5.5-10.1)
Oversample	RandomForest	96.0 (93.8-98.2)	91.8 (87.9-95.7)	0.96 (0.93-0.99)	99.2 (98.3-100)	11.3 (9.2-13.4)
Oversample	GradientBoosting	94.5 (91.6-97.4)	91.0 (86.9-95.1)	0.95 (0.92-0.98)	99.0 (98.0-100)	10.2 (8.2-12.2)
Oversample	OneClassSVM	88.0 (84.0-92.0)	85.5 (80.5-90.5)	0.89 (0.85-0.93)	97.8 (96.4-99.2)	6.4 (3.9-8.9)
Oversample	MLP	33.5 (27.5-39.5)	90.5 (86.2-94.8)	0.66 (0.62-0.70)	95.4 (92.9-97.9)	2.9 (1.3-4.5)
Oversample	LR	0.0 (0.0-0.0)	99.52 (94.54-100)	0.65 (0.62-0.68)	99.05 (94.10-100)	0.0 (0.0-0.0)
Oversample	SVC	85.0 (80.8-89.2)	88.2 (83.4-93.0)	0.90 (0.87-0.93)	97.2 (95.5-98.9)	5.6 (3.4-7.8)
CostSensitive	DecisionTree	64.29 (61.08-67.50)	99.8 (94.81-100)	0.98 (0.93-1.00)	99.66 (94.68-100)	5.0 (1.25-8.75)
CostSensitive	RandomForest	71.43 (67.86-75.00)	99.0 (98.0-100)	0.99 (0.94-1.00)	99.73 (94.74-100)	9.0 (8.0-10.0)
CostSensitive	GradientBoosting	71.43 (67.86-75.00)	99.0 (98.0-100)	0.99 (0.94-1.00)	99.73 (94.74-100)	9.0 (8.0-10.0)
CostSensitive	OneClassSVM	0.0 (0.0-0.0)	99.0 (98.0-100)	0.97 (0.92-1.00)	99.05 (94.10-100)	0.0 (0.0-0.0)
CostSensitive	MLP	0.0 (0.0-0.0)	98.29 (93.38-100)	0.59 (0.56-0.62)	99.03 (94.08-100)	0.0 (0.0-0.0)
CostSensitive	LR	0.0 (0.0-0.0)	99.0 (98.0-100)	0.6 (0.57-0.63)	99.05 (94.10-100)	0.0 (0.0-0.0)
CostSensitive	SVC	0.0 (0.0-0.0)	99.0 (98.0-100)	0.97 (0.92-1.00)	99.05 (94.10-100)	0.0 (0.0-0.0)

[a]AUC: area under the curve.
[b]NPV: negative predictive value.
[c]PPV: positive predictive value.
[d]MLP: multilayer perceptron.
[e]LR: logistic regression.
[f]SVC: support vector classifier.
[g]SVM: support vector machine.

LOOCV was applied using the best-performing model configuration (random forest with oversampling) to assess generalizability across individual participants. As anticipated, performance metrics were lower compared to stratified 5-fold cross-validation, reflecting the impact of increased variance and the challenge of predicting rare events in single-subject test folds. Pulled AUC across participants was 0.59(95% CI 0.54-0.64; Figure 3), with individual values ranging from 0.47 to 0.64. Specificity remained consistently high (97.2%, 95% CI 95.9-98.5), whereas sensitivity was notably limited (5.3%, 95% CI 0.0-15.5), underscoring the difficulty of reliably identifying decompensation events in highly imbalanced datasets (Table 3).

XSL•FO
RenderX

Figure 3. ROC curve for the random forest model with oversampling, evaluated by LOOCV. The solid blue line shows model performance across all thresholds, the dashed black diagonal represents chance level (AUC=0.5), and the text box reports the AUC (0.59, 95% CI 0.54-0.64). AUC: area under the curve; LOOCV: leave-one-subject-out cross-validation; ROC: receiver operating characteristic

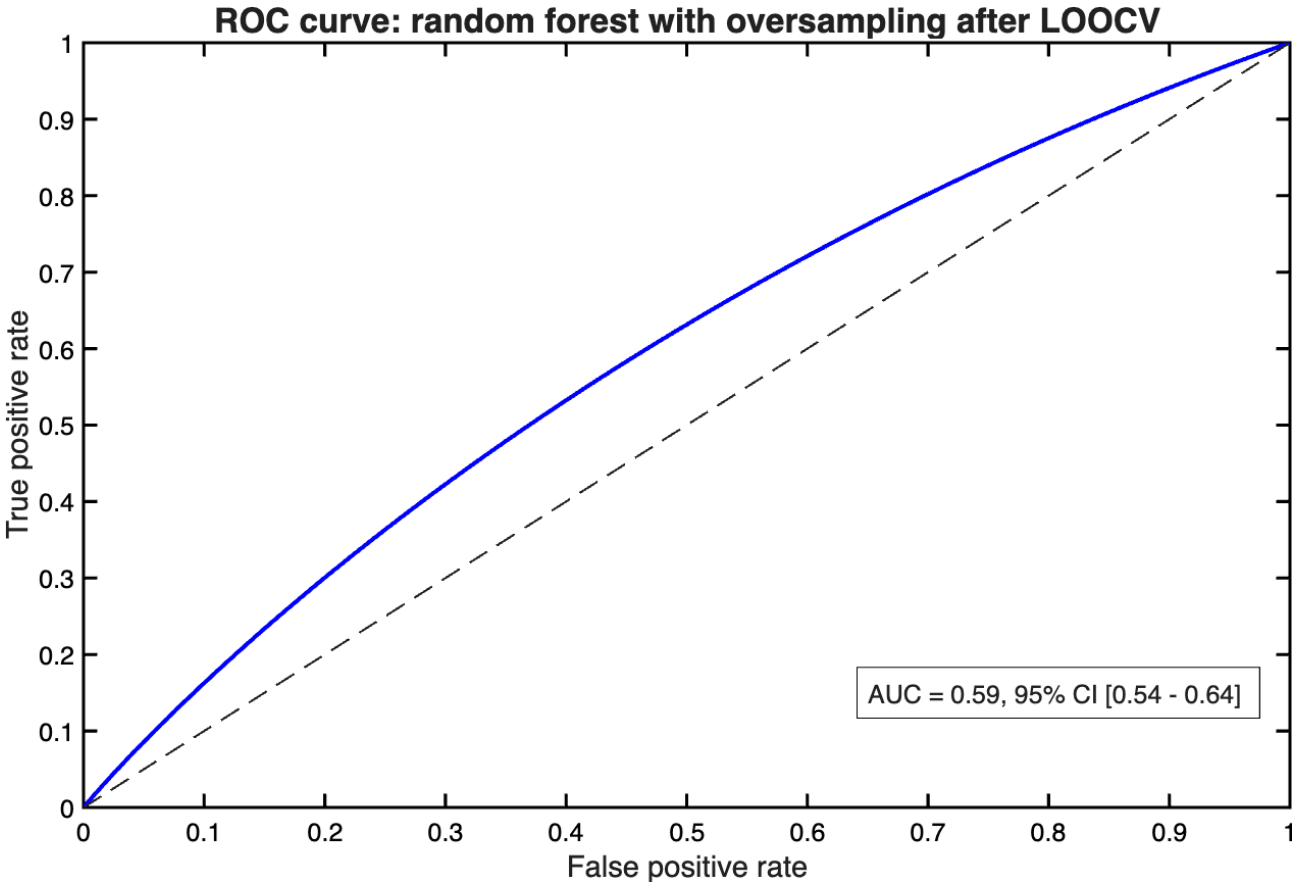

XSL•FO
RenderX

Table . Iterations of the LOOCV analysis.

Participant number	AUC[c] (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)
1	0.61 (0.56-0.66)	97.1 (95.8-98.4)	6.2 (0.0-17.4)
2	0.57 (0.51-0.63)	96.5 (94.5-98.5)	4.3 (0.0-14.8)
3	0.59 (0.55-0.63)	98.3 (97.2-99.4)	5.1 (0.0-15.1)
4	0.55 (0.52-0.58)	95.9 (94.4-97.4)	3.9 (0.0-13.3)
5	0.60 (0.55-0.65)	97.8 (96.6-99.0)	7.1 (0.0-18.1)
6	0.58 (0.52-0.64)	96.7 (95.3-98.1)	4.9 (0.0-15.7)
7	0.62 (0.57-0.67)	97.9 (96.8-99.0)	6.8 (0.0-17.7)
8	0.48 (0.47-0.49)	98.5 (97.7-99.3)	0.0 (0.0-0.0)
9	0.64 (0.60-0.68)	98.1 (97.1-99.1)	8.7 (0.0-20.0)
10	0.53 (0.49-0.57)	96.0 (94.0-98.0)	3.1 (0.0-12.6)
11	0.58 (0.53-0.63)	96.8 (95.5-98.1)	5.4 (0.0-15.8)
12	0.59 (0.54-0.64)	97.5 (96.3-98.7)	5.9 (0.0-16.1)
13	0.61 (0.56-0.66)	97.6 (96.6-98.6)	6.5 (0.0-17.1)
14	0.56 (0.51-0.61)	97.2 (95.9-98.5)	4.4 (0.0-14.2)
15	0.57 (0.52-0.62)	97.3 (96.2-98.4)	4.6 (0.0-14.5)
16	0.56 (0.51-0.61)	96.9 (95.7-98.1)	4.2 (0.0-13.9)
17	0.59 (0.54-0.64)	97.4 (96.3-98.5)	5.8 (0.0-15.8)
All participants	0.59 (0.54-0.64)	97.2 (95.9-98.5)	5.3 (0.0-15.5)

[a]LOOCV: leave-one-subject-out cross-validation.

[b]The table shows the AUC, specificity, and sensitivity when (eg, in row 1) participant 1 is left out of the analysis.

[c]AUC: area under the curve.

Feature importance analysis indicated that HR metrics contributed most to the model's predictions (36%), followed by IBIs (18%), AC (14%), RR (12%), and EE (11%). Wearing behavior (9%) also contributed modestly.

Trends in Physiological Parameters

Table 4 represents the mean and SD of the physiological parameters for both the stable phase and the phase preceding HF decompensation. Notably, 2 weeks prior to the HF decompensation event, HR significantly increased by 11 beats per minute. Additionally, the daily IBI, related to the HRV index, decreased significantly by 124 ms. All other parameters showed no significant differences.

The time series analysis (Figure 4) also revealed a trend in the physiological parameters over the 4-week period before the HF decompensation event. HR demonstrated an upward trajectory, while IBI displayed a downward trend, indicating an imbalance in autonomic regulation.

XSL•FO
RenderX

Table . Means and SDs of physiological parameters during the stable phase and decompensated heart failure phase.

	Stable phase, mean (SD)	ADHF[a] phase, mean (SD)	Δ	*P* value
HR[b] (bpm)	68 (7)	79 (7)	11 (0)	.03
IBI[c] (ms)	886 (91)	762 (60)	−124 (30)	.03
RR[d] (BrPM)	67 (8)	71 (6)	4 (2)	.22
EE[e] (kcal/day)	375 (129)	466 (304)	91 (175)	.69
AC[f] (counts/day)	13e+10 (9e+7)	85e+10 (3e+7)	72e+10 (6e+7)	.29

[a]ADHF: acute decompensated heart failure.
[b]HR: heart rate.
[c]IBI: interbeat interval.
[d]RR: respiratory rate.
[e]EE: energy expenditure.
[f]AC: activity counts.

Figure 4. Timeseries of the physiological parameters 4 weeks prior to HF decompensation events with the daily average values of the HR, IBI, RR, EE, and AC of an example patient. AC: activity count; EE: energy expenditure; HF: heart failure; HR: heart rate; IBI: interbeat interval; RR: respiration rate.

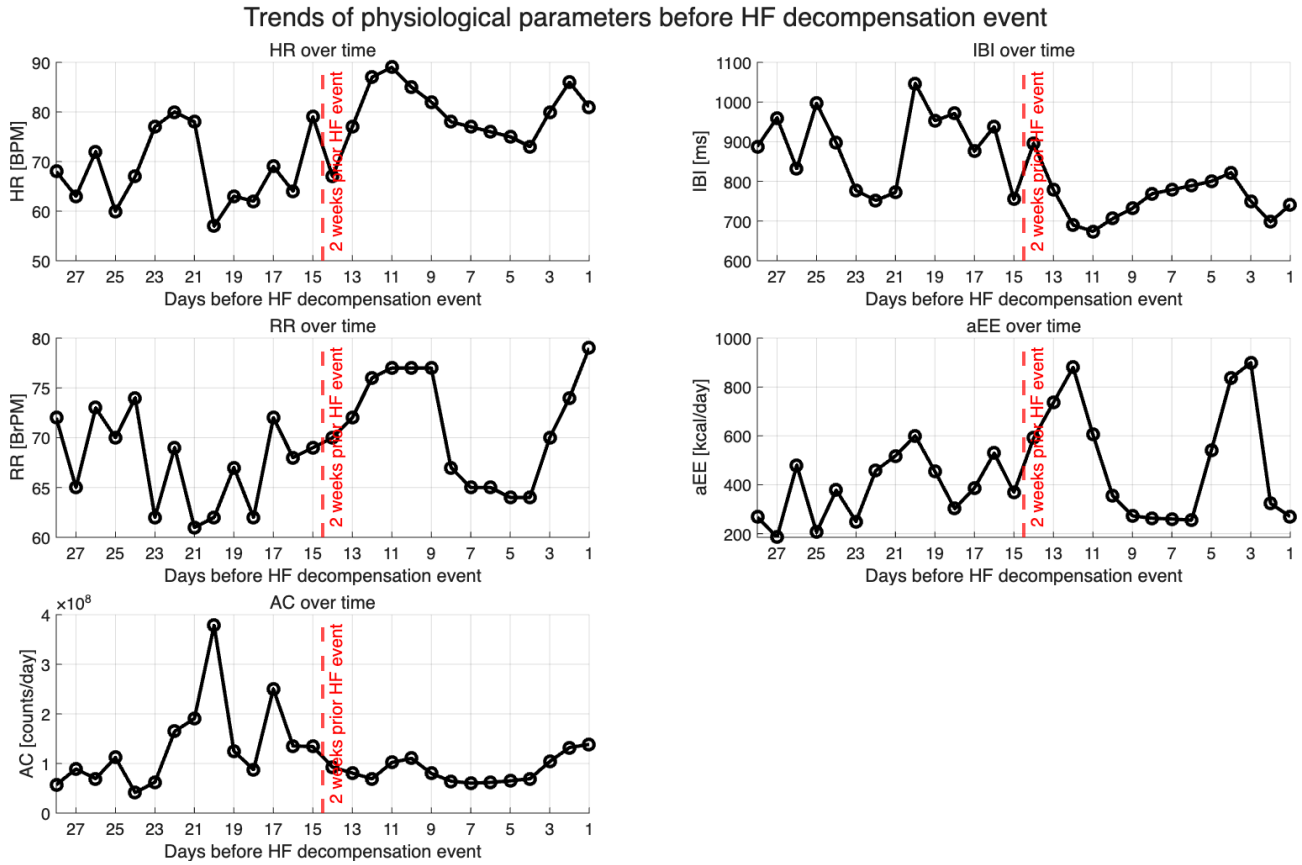

Discussion

Principal Findings

This exploratory study is one of the first that used noninvasive continuous data monitoring (rather than spot measurements) to predict HF-related events in an older HF population in a real-life setting. An MPM based on HR, IBI, RR, AC, and EE predicted the combined end point of decompensated HF and death with a high specificity but low sensitivity. This limited sensitivity likely reflects two key challenges: first, the poor signal quality observed from wrist-worn sensors in an older cohort, which resulted in the removal of data from several 14-day sequences; and second, the severe imbalance between stable-phase and decompensation events, which makes it difficult for even powerful classifiers to learn a reliable decision boundary without overfitting. Future work should therefore focus both on boosting the fidelity of wearable-device data in real-world, older populations and aiming to increase the number of monitored

XSL•FO
RenderX

events—either by extending follow-up or broadening inclusion criteria to mitigate extreme class imbalance.

HR-Related Data (HR and IBI) and RR

HR-related parameters (HR and IBI) were the most important contributors to the prediction model. These results are in line with previous studies showing that an increased HR is associated with a higher risk of hospital readmissions and mortality in stable chronic patients with HF with sinus rhythm, regardless of LVEF [23,24]. Moreover, the amount of evidence is growing that a higher HR at discharge after an episode of acute decompensated HF is associated with a worse prognosis [25-27]. When HR is followed over time, an increase in HR is associated with a worse prognosis and higher risk of hospital readmission [18,23]. In addition, previous studies in patients with implantable devices showed that HRV declines prior to HF decompensation as a result of autonomic imbalance [10,18]. In patients with HF with preserved ejection fraction (HFpEF), less is known about the prognostic role of HRV, although a reduced HRV seems to predict 1-year mortality in these patients [28]. To our knowledge, there is no data available on longitudinal follow-up of HRV in patients with HFpEF prior to an episode of decompensated HF. Nevertheless, as an increased autonomic imbalance precedes congestion [29], HRV is likely to change earlier in the cascade than traditional RPM parameters such as weight and symptoms, and should therefore be considered in future RPM studies. RR was the fourth most important contributor in the prediction model in this study. Previous research on RR data extracted from implantable cardioverter defibrillators/cardiac resynchronization therapy devices in New York Heart Association class III patients with HF showed that median, minimum, and maximum RR were significantly elevated prior to HF decompensation [30]. Nevertheless, in patients with HF with concomitant sleep-disordered breathing treated with adaptive servo-ventilation, nightly RR was significantly lower, and apnea-hypopnea index was significantly increased in patients with decompensated HF [31]. This indicates that the change in RR pattern is dependent on the presence of comorbidities and the timing of measurement. Studies about the accuracy of wrist-worn devices in the estimation of RR are scarce. A validation study in healthy individuals, in resting conditions, showed that a wrist-worn PPG-based device could accurately measure RR [32]. Moreover, a previous validation study with the Philips health watch showed that it was able to estimate RR in healthy individuals undergoing a laboratory protocol with sufficient accuracy [33]. Nightly RR in individuals undergoing polysomnography can also be accurately measured by a consumer wrist-worn device; however, accuracy is decreasing with worsening of sleep apnea [34]. In addition, in patients undergoing postabdominal surgery residing at the intensive care or postanesthesia unit, RR can be measured sufficiently accurately using a PPG wristband, although the detection of RR was easily corrupted by motion artifacts, leading to the exclusion of 66% of the measurements [35]. This highlights the need for validation of wrist-worn devices in the HF population in free-living conditions. The limited amount of high-quality RR data in this study might partially be attributed to external factors such as motion artifacts in free-living conditions and closeness of the device to the skin [36]. Individual factors, such as age or BMI, can also contribute to noise in the PPG signal. Particularly, in older people, the PPG signal may be negatively influenced by decreased arterial compliance and thinning of the skin [36]. Moreover, the BMI in our study population was high, which is associated with an increased cutaneous blood flow, a decreased capillary density, and an increased transepidermal water loss. These BMI-induced changes negatively affect the PPG signal [36].

Activity Monitoring

Activity-related parameters (AC and EE) also contributed to the best predictive model. This is consistent with a previous systematic review, which showed that noninvasive actigraphy-quantified physical activity showed a strong inverse relationship with intercurrent HF events such as hospitalizations [37]. However, most of these studies used a single physical activity measurement to quantify the level of physical activity in relation to the risk of HF-related events. Cakmak et al [38] showed that continuous, smartphone-based monitoring of accelerometry data, alone or together with other parameters, is helpful in predicting HF decompensation up to 6 days in advance of the encounter of the event. Moreover, Adamson et al [13] showed that invasively monitored physical activity, using cardiac resynchronization therapy data, significantly decreased prior to a HF hospitalization, although these changes were less sensitive than changes in HRV measured with the same device. The SELENE-HF study also used invasively monitored physical activity in an MPM to predict hospitalization in patients with HF, although its contribution to the model was relatively low as compared to HR [18]. An explanation for the limited predictive value of activity behavior in this study may be related to data quality and accuracy. Although a previous validation study with a Philips health watch showed that it accurately estimates EE in healthy people undergoing a laboratory protocol [33], these results might not be applicable to older patients with HF in free-living conditions. Indeed, we recently showed that the accuracy of commercially available wrist-worn devices in estimating EE in patients with HF performing a laboratory protocol was low [39]. More importantly, the responsiveness, that is, the ability to detect within-patient changes in EE within patients, was also poor [39]. This indicates the need for population-specific algorithms for monitoring activity in the home environment. Moreover, it was shown that hip placement of an accelerometer led to a more accurate estimation of metabolic equivalents as compared to wrist, ankle, upper arm, and thigh placement, in older adults performing a laboratory protocol [40]. To our knowledge, it is not known to what extent the accelerometer location is relevant in older patients with HF in free-living conditions.

Sleep Monitoring

Data on sleep stages and quality could not be accurately measured by the device used in this study, leading to the exclusion of sleep features from our prediction model. For this reason, sleep data were not included in the MPM. Nevertheless, sleep data are expected to be of particular value in future prediction models, as previous research has found an association between delayed bedtime or wake-up time and the development of congestive HF [41]. Moreover, congestive HF negatively

XSL•FO
RenderX

influences sleep quality [42,43]. As decompensated HF is often associated with orthopnea and nycturia, sleep duration and quality are expected to decrease upon fluid retention. Additionally, 50%‑75% of the patients with HF have concomitant sleep-disordered breathing [44], which can worsen during a HF exacerbation. This also contributes to poorer sleep quality during the night and somnolence during the day. However, it remains questionable how to accurately assess sleep data in patients with HF in the home environment using a patient-friendly device. Although commercially available sleep tracking devices tend to accurately detect sleep, they detect wake less accurately as compared to polysomnography in healthy individuals, whereas sleep stage assessments were inconsistent as well [45,46]. In the study of Chinoy et al [45], commercially available sleep trackers outperformed research-grade actigraphy devices, and notably, all devices performed worse during nights with disrupted or poor sleep. Other nonwrist-worn sleep tracking devices showed variable agreement with polysomnography in individuals without uncontrolled respiratory conditions [47]. To our knowledge, sleep-tracking devices have not been specifically validated in an HF population in which interrupted sleep is highly common. Therefore, more research is needed in this specific population.

Future Implications

PPG Signal-to-Noise Ratio

Although the individual parameters used in the algorithm (HR, IBI, AC, RR, and EE) are potentially indicative of HF decompensation, the combination of these parameters led to a model with high specificity but low sensitivity. This indicates that while the model is reliable in identifying clinically stable periods (ie, few false alarms), it fails to detect the majority of actual HF decompensation events. In other words, when an alarm is generated by the MPM, decompensated HF (or death due to HF) will likely occur in the 2-week period, but many events occur without any alarm being raised. Due to this high miss rate, the current model is not suitable for use in daily clinical practice. To further improve the prediction model, several issues should be addressed. First, there were far more "stable" data points than decompensation events, and even with oversampling and cost-sensitive methods, the model struggled to distinguish the rare events without overfitting to the majority class. To address this, future studies should increase the number of decompensation cases, either by extending the monitoring period, increasing the patient population, or stricter inclusion criteria, to balance the dataset at the source. Second, the question is whether the device used in this study provides accurate measurements in older patients with HF in free-living conditions and whether it accurately detects intraindividual changes. As patients with HF often have a disturbed respiration pattern, a disturbed sleep pattern, function in the low-intensity physical domain, and have ectopic heartbeats or intermittent heart rhythm disturbances, validation results of wrist-worn devices in healthy individuals cannot be extrapolated to the HF population. Ideally, the device should be first validated in the HF population, and population-specific algorithms should be developed. Third, several factors negatively influence the wrist-based PPG signal, such as closeness of the device to the skin, skin properties, and vascular phenomena related to older age. Therefore, other types of non-PPG devices, such as the thoracic patch, using a combination of electrocardiogram and impedance, which was used in the LINK-HF study [19], might deliver more accurate results when firmly attached to the skin. Wearing a patch might also lead to a higher adherence as compared to a wrist-worn device, as wrist-worn devices can easily be removed and forgotten afterwards. In a real-world setting, the expected adherence to a wrist-worn device might be even lower than was reported in this study, as in this study, the local investigators contacted participants when data were lacking for 3 consecutive days. Nevertheless, patches also have disadvantages as they are often disposable and might be costly when they have to be replaced repeatedly. In addition, as the skin of older people is thinner and easily disrupted, patches might lead to skin problems when prolonged wearing is necessary. Moreover, wrist-worn devices are generally accepted, also in the older population, and provide an opportunity for direct feedback based on the measured parameters. Moreover, wrist-worn devices are far less costly (ie, an estimated €150 [US $176]) as compared to invasive devices and are expected to be suitable for reuse. Whether they are also less costly as compared to other noninvasive devices should, however, be further evaluated. Wrist-worn devices with a higher signal-to-noise ratio that are tailored to an older population might therefore be preferred over other devices.

Workflow Methodologies

In order to obtain a greater amount of high-quality data, some alterations in the workflow may be considered. For example, asking the patient to perform daily recordings lasting a few minutes while sitting completely still. This might lead to a higher amount of high-quality data, which might improve the predictive value of the algorithm, although this hypothesis should be tested. High-quality sleep and activity data cannot be obtained using this strategy. Using a strap that connects the PPG sensor closer and more stable to the skin might lead to a higher data quality in all categories (ie, HR, HRV, RR, physical activity, and sleep). This concept, however, needs to be developed and tested before implementation.

Potential of Artificial Intelligence

Although the models that have been evaluated in this study did not lead to one with sufficient diagnostic accuracy, artificial intelligence has great potential to increase the efficiency of health care, particularly in the prediction of worsening of HF. Machine-learning and deep learning methods based on noninvasive collected parameters do show promising results in the early detection of decompensated HF, although the effects on clinical outcomes should be further evaluated [48,49].

Strengths and Limitations

This study is one of the first to develop and evaluate an MPM for the early detection of worsening HF, using a single, noninvasive, wrist-worn device. A major limitation of this study is the inclusion of a limited number of participants. Including more participants in the study, that is, increasing the dataset size, might have improved the model's performance and class balance. In this study, only nine events occurred, of which seven were episodes of decompensated HF treated at the outpatient clinic. Only one participant was readmitted to the hospital due

XSL•FO
RenderX

to decompensated HF. Another limitation was that the included population predominantly consisted of patients with HFpEF, and patients with permanent atrial fibrillation were excluded from study participation. Therefore, the results cannot be extrapolated to the entire HF population. Nevertheless, being an exploratory study, the results of this study have shed light on the future steps needed to optimize the use of wrist-worn wearables to predict HF events in older patients.

Conclusions

A noninvasive MPM using a single wrist-worn device measuring HR, IBI (HRV), RR, AC, and EE showed a high specificity but a low sensitivity in predicting decompensated HF and death. An important reason for the low sensitivity could be the extreme class imbalance and low data quality (specifically in IBI, RR, and sleep data), resulting in the loss of usable input for the MPM in this real-world HF population. Future studies are needed for validation and optimization of sensor technology in wrist-worn devices for this specific population.

Acknowledgments

This research did not receive a grant from any funding agency. The devices and technical support were provided by Philips Electronics Nederland B.V. Philips Electronics Nederland B.V. was not involved in the data analysis and interpretation, nor in the drafting of the manuscript.

Authors' Contributions

Conceptualization: CH, HMCK.
Data curation: CH, ML, ILJL.
Formal analysis: JT, VAAE.
Investigation: CH, ML, ILJL.
Methodology: JT, VAAE.
Project administration: CH.
Resources: HMCK.
Software: JT, VAAE.
Supervision: AS, HMCK.
Validation: VAAE.
Visualization: CH, ML, ILJL, VAAE, JT.
Writing original draft: CH, ML, ILJL, JT, AS, HMCK.
Writing review and editing: ILJL, VAAE, CH.

Conflicts of Interest

None declared.

References

1. Savarese G, Becher PM, Lund LH, Seferovic P, Rosano GMC, Coats AJS. Global burden of heart failure: a comprehensive and updated review of epidemiology. Cardiovasc Res 2023 Jan 18;118(17):3272-3287. [doi: 10.1093/cvr/cvac013]
2. Virani SS, Alonso A, Aparicio HJ, et al. Heart disease and stroke statistics-2021 update: a report from the American Heart Association. Circulation 2021 Feb 23;143(8):e254-e743. [doi: 10.1161/CIR.0000000000000950] [Medline: 33501848]
3. Lawson C, Crothers H, Remsing S, et al. Trends in 30-day readmissions following hospitalisation for heart failure by sex, socioeconomic status and ethnicity. eClinicalMedicine 2021 Aug;38:101008. [doi: 10.1016/j.eclinm.2021.101008]
4. Del Toro MR, Leaño NMH, Barahona-Correa JE, Velandia OMM, Ávila DGF, Peña Á. Effectiveness of mobile telemonitoring applications in heart failure patients: systematic review of literature and meta-analysis. Heart Fail Rev 2023 Mar;28(2):431-452. [doi: 10.1007/s10741-022-10291-1] [Medline: 36652096]
5. Scholte NTB, Gürgöze MT, Aydin D, et al. Telemonitoring for heart failure: a meta-analysis. Eur Heart J 2023 Aug 14;44(31):2911-2926. [doi: 10.1093/eurheartj/ehad280] [Medline: 37216272]
6. Ding H, Chen SH, Edwards I, et al. Effects of different telemonitoring strategies on chronic heart failure care: systematic review and subgroup meta-analysis. J Med Internet Res 2020;22(11):e20032. [doi: 10.2196/20032]
7. Zhu Y, Gu X, Xu C. Effectiveness of telemedicine systems for adults with heart failure: a meta-analysis of randomized controlled trials. Heart Fail Rev 2020 Mar;25(2):231-243. [doi: 10.1007/s10741-019-09801-5] [Medline: 31197564]
8. Drews TEI, Laukkanen J, Nieminen T. Non-invasive home telemonitoring in patients with decompensated heart failure: a systematic review and meta-analysis. ESC Heart Fail 2021 Oct;8(5):3696-3708. [doi: 10.1002/ehf2.13475] [Medline: 34165912]
9. De Lathauwer ILJ, Nieuwenhuys WW, Hafkamp F, et al. Remote patient monitoring in heart failure: a comprehensive meta-analysis of effective programme components for hospitalization and mortality reduction. Eur J Heart Fail 2025 Jan 20. [doi: 10.1002/ejhf.3568] [Medline: 39834044]

XSL•FO
RenderX

10. Adamson PB. Pathophysiology of the transition from chronic compensated and acute decompensated heart failure: new insights from continuous monitoring devices. Curr Heart Fail Rep 2009 Dec;6(4):287-292. [doi: 10.1007/s11897-009-0039-z] [Medline: 19948098]
11. Gontarska K, Wrazen W, Beilharz J, et al. Predicting medical interventions from vital parameters: towards a decision support system for remote patient monitoring. 2021 Presented at: 19th International Conference on Artificial Intelligence in Medicine, AIME 2021; Jun 15-18, 2021; Online.
12. Moazeni M, Numan L, Brons M, et al. Developing a personalized remote patient monitoring algorithm: a proof-of-concept in heart failure. Eur Heart J Digital Health 2023 Dec;4(6):455-463. [doi: 10.1093/ehjdh/ztad049] [Medline: 38045433]
13. Adamson PB, Smith AL, Abraham WT, et al. Continuous autonomic assessment in patients with symptomatic heart failure: prognostic value of heart rate variability measured by an implanted cardiac resynchronization device. Circulation 2004 Oct 19;110(16):2389-2394. [doi: 10.1161/01.CIR.0000139841.42454.78] [Medline: 15313946]
14. Kim H, Jung HR, Kim JB, Kim DJ. Autonomic dysfunction in sleep disorders: from neurobiological basis to potential therapeutic approaches. J Clin Neurol 2022 Mar;18(2):140-151. [doi: 10.3988/jcn.2022.18.2.140] [Medline: 35274834]
15. Valika A, Costanzo MR. Sleep-disordered breathing during congestive heart failure: to intervene or not to intervene? Card Fail Rev 2017 Nov;3(2):134-139. [doi: 10.15420/cfr.2017:7:1] [Medline: 29387467]
16. Whellan DJ, Ousdigian KT, Al-Khatib SM, et al. Combined heart failure device diagnostics identify patients at higher risk of subsequent heart failure hospitalizations: results from PARTNERS HF (Program to Access and Review Trending Information and Evaluate Correlation to Symptoms in Patients With Heart Failure) study. J Am Coll Cardiol 2010 Apr 27;55(17):1803-1810. [doi: 10.1016/j.jacc.2009.11.089] [Medline: 20413029]
17. Boehmer JP, Hariharan R, Devecchi FG, et al. A multisensor algorithm predicts heart failure events in patients with implanted devices: results from the MultiSENSE study. JACC Heart Fail 2017 Mar;5(3):216-225. [doi: 10.1016/j.jchf.2016.12.011] [Medline: 28254128]
18. D'Onofrio A, Solimene F, Calò L, et al. Combining home monitoring temporal trends from implanted defibrillators and baseline patient risk profile to predict heart failure hospitalizations: results from the SELENE HF study. Europace 2022 Feb 2;24(2):234-244. [doi: 10.1093/europace/euab170] [Medline: 34392336]
19. Stehlik J, Schmalfuss C, Bozkurt B, et al. Continuous wearable monitoring analytics predict heart failure hospitalization: the LINK-HF multicenter study. Circ Heart Fail 2020 Mar;13(3):e006513. [doi: 10.1161/CIRCHEARTFAILURE.119.006513] [Medline: 32093506]
20. Gheorghiade M, Zannad F, Sopko G, et al. Acute heart failure syndromes: current state and framework for future research. Circulation 2005 Dec 20;112(25):3958-3968. [doi: 10.1161/CIRCULATIONAHA.105.590091] [Medline: 16365214]
21. Shaffer F, Ginsberg JP. An overview of heart rate variability metrics and norms. Front Public Health 2017;5:258. [doi: 10.3389/fpubh.2017.00258] [Medline: 29034226]
22. Wong TT. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. Pattern Recognit DAGM 2015 Sep;48(9):2839-2846. [doi: 10.1016/j.patcog.2015.03.009]
23. Castagno D, Skali H, Takeuchi M, et al. Association of heart rate and outcomes in a broad spectrum of patients with chronic heart failure: results from the CHARM (Candesartan in Heart Failure: Assessment of Reduction in Mortality and morbidity) program. J Am Coll Cardiol 2012 May 15;59(20):1785-1795. [doi: 10.1016/j.jacc.2011.12.044] [Medline: 22575317]
24. Böhm M, Swedberg K, Komajda M, et al. Heart rate as a risk factor in chronic heart failure (SHIFT): the association between heart rate and outcomes in a randomised placebo-controlled trial. Lancet 2010 Sep 11;376(9744):886-894. [doi: 10.1016/S0140-6736(10)61259-7] [Medline: 20801495]
25. Faragli A, Tano GD, Carlini CD, et al. In-hospital heart rate reduction with beta blockers and ivabradine early after recovery in patients with acute decompensated heart failure reduces short-term mortality and rehospitalization. Front Cardiovasc Med 2021;8:665202. [doi: 10.3389/fcvm.2021.665202]
26. Laskey WK, Alomari I, Cox M, et al. Heart rate at hospital discharge in patients with heart failure is associated with mortality and rehospitalization. J Am Heart Assoc 2015 Apr 22;4(4):e001626. [doi: 10.1161/JAHA.114.001626] [Medline: 25904590]
27. Bahouth F, Elias A, Ghersin I, et al. The prognostic value of heart rate at discharge in acute decompensation of heart failure with reduced ejection fraction. ESC Heart Fail 2022 Feb;9(1):585-594. [doi: 10.1002/ehf2.13710] [Medline: 34821080]
28. Ksela J, Rupert L, Djordjevic A, Antonic M, Avbelj V, Jug B. Altered heart rate turbulence and variability parameters predict 1-year mortality in heart failure with preserved ejection fraction. J Cardiovasc Dev Dis 2022 Jul 2;9(7):213. [doi: 10.3390/jcdd9070213] [Medline: 35877575]
29. Arrigo M, Parissis JT, Akiyama E, Mebazaa A. Understanding acute heart failure: pathophysiology and diagnosis. Eur Heart J Suppl 2016 Nov;18(suppl G):G11-G18. [doi: 10.1093/eurheartj/suw044]
30. Goetze S, Zhang Y, An Q, et al. Ambulatory respiratory rate trends identify patients at higher risk of worsening heart failure in implantable cardioverter defibrillator and biventricular device recipients: a novel ambulatory parameter to optimize heart failure management. J Interv Card Electrophysiol 2015 Jun;43(1):21-29. [doi: 10.1007/s10840-015-9983-6]
31. Fox H, Rudolph V, Munt O, et al. Early identification of heart failure deterioration through respiratory monitoring with adaptive servo - ventilation. J Sleep Res 2023 Feb;32(1) [FREE Full text] [doi: 10.1111/jsr.13749]

XSL•FO
RenderX

32. Dur O, Rhoades C, Ng MS, Elsayed R, van Mourik R, Majmudar MD. Design rationale and performance evaluation of the wavelet health wristband: benchtop validation of a wrist-worn physiological signal recorder. JMIR Mhealth Uhealth 2018 Oct 16;6(10):e11040. [doi: 10.2196/11040] [Medline: 30327288]
33. Hendrikx J, Ruijs LS, Cox LG, Lemmens PM, Schuijers EG, Goris AH. Clinical evaluation of the measurement performance of the Philips health watch: a within-person comparative study. JMIR Mhealth Uhealth 2017 Feb 2;5(2):e10. [doi: 10.2196/mhealth.6893] [Medline: 28153815]
34. Jung H, Kim D, Choi J, Joo EY. Validating a consumer smartwatch for nocturnal respiratory rate measurements in sleep monitoring. Sensors (Basel) 2023 Sep 19;23(18):7976. [doi: 10.3390/s23187976] [Medline: 37766031]
35. van der Stam JA, Mestrom EHJ, Scheerhoorn J, et al. The accuracy of wrist-worn photoplethysmogram-measured heart and respiratory rates in abdominal surgery patients: observational prospective clinical validation study. JMIR Perioper Med 2023 Feb 20;6:e40474. [doi: 10.2196/40474] [Medline: 36804173]
36. Fine J, Branan KL, Rodriguez AJ, et al. Sources of Inaccuracy in photoplethysmography for continuous cardiovascular monitoring. Biosensors (Basel) 2021 Apr 16;11(4):126. [doi: 10.3390/bios11040126] [Medline: 33923469]
37. Tan MKH, Wong JKL, Bakrania K, et al. Can activity monitors predict outcomes in patients with heart failure? A systematic review. Eur Heart J Qual Care Clin Outcomes 2019 Jan 1;5(1):11-21. [doi: 10.1093/ehjqcco/qcy038] [Medline: 30215706]
38. Cakmak AS, Perez Alday EA, Densen S, et al. Passively captured interpersonal social interactions and motion from smartphones for predicting decompensation in heart failure: observational cohort study. JMIR Form Res 2022 Aug 24;6(8):e36972. [doi: 10.2196/36972] [Medline: 36001367]
39. Herkert C, Kraal JJ, van Loon EMA, van Hooff M, Kemps HMC. Usefulness of modern activity trackers for monitoring exercise behavior in chronic cardiac patients: validation study. JMIR Mhealth Uhealth 2019 Dec 19;7(12):e15045. [doi: 10.2196/15045] [Medline: 31855191]
40. Davoudi A, Mardini MT, Nelson D, et al. The effect of sensor placement and number on physical activity recognition and energy expenditure estimation in older adults: validation study. JMIR Mhealth Uhealth 2021 May 3;9(5):e23681. [doi: 10.2196/23681] [Medline: 33938809]
41. Yan B, Li R, Li J, et al. Sleep timing may predict congestive heart failure: a community-based cohort study. J Am Heart Assoc 2021 Mar 16;10(6):e018385. [doi: 10.1161/JAHA.120.018385] [Medline: 33666090]
42. Zuurbier LA, Luik AI, Leening MJG, et al. Associations of heart failure with sleep quality: the Rotterdam Study. J Clin Sleep Med 2015 Jan 15;11(2):117-121. [doi: 10.5664/jcsm.4454] [Medline: 25406270]
43. Redeker NS, Stein S. Characteristics of sleep in patients with stable heart failure versus a comparison group. Heart Lung 2006;35(4):252-261. [doi: 10.1016/j.hrtlng.2005.10.007] [Medline: 16863897]
44. Cowie MR, Gallagher AM. Sleep disordered breathing and heart failure: what does the future hold? JACC Heart Fail 2017 Oct;5(10):715-723. [doi: 10.1016/j.jchf.2017.06.016] [Medline: 28888522]
45. Chinoy ED, Cuellar JA, Huwa KE, et al. Performance of seven consumer sleep-tracking devices compared with polysomnography. Sleep 2021 May 14;44(5):zsaa291. [doi: 10.1093/sleep/zsaa291] [Medline: 33378539]
46. Miller DJ, Sargent C, Roach GD. A validation of six wearable devices for estimating sleep, heart rate and heart rate variability in healthy adults. Sensors (Basel) 2022 Aug 22;22(16):6317. [doi: 10.3390/s22166317] [Medline: 36016077]
47. Lee T, Cho Y, Cha KS, et al. Accuracy of 11 wearable, nearable, and airable consumer sleep trackers: prospective multicenter validation study. JMIR Mhealth Uhealth 2023 Nov 2;11:e50983. [doi: 10.2196/50983] [Medline: 37917155]
48. Wattanachayakul P, Kittipibul V, Salah HM, et al. Non-invasive heart failure monitoring: leveraging smart scales and digital biomarkers to improve heart failure outcomes. Heart Fail Rev 2024 Sep;29(5):1145-1156. [doi: 10.1007/s10741-024-10426-6] [Medline: 39039364]
49. Penso M, Solbiati S, Moccia S, Caiani EG. Decision support systems in HF based on deep learning technologies. Curr Heart Fail Rep 2022 Apr;19(2):38-51. [doi: 10.1007/s11897-022-00540-7] [Medline: 35142985]

Abbreviations

AC: activity counts
AUC: area under the curve
EE: energy expenditure
HF: heart failure
HFpEF: heart failure with preserved ejection fraction
HR: heart rate
HRV: heart rate variability
IBI: interbeat interval
LOOCV: leave-one-out cross-validation
LVEF: left ventricular ejection fraction
MPM: multiparameter predictive model
MSX: monitoring study box
PDL: Philips data logger

XSL•FO
RenderX

PPG: photoplethysmography
RPM: remote patient monitoring
RR: respiration rate

Edited by J Sarvestan; submitted 02.04.24; peer-reviewed by S Sarkar, V Giampatzis; revised version received 09.07.25; accepted 10.07.25; published 08.10.25.

Please cite as:
Herkert C, van Leunen M, De Lathauwer ILJ, van Es VAA, Tang J, Saeed A, Spee RF, Lu Y, Kemps HMC
Noninvasive Multiparameter Monitoring for the Detection of Decompensated Heart Failure: Exploratory Study
JMIR Form Res 2025;9:e59116
URL: https://formative.jmir.org/2025/1/e59116
doi:10.2196/59116

©Cyrille Herkert, Mayke van Leunen, Ignace Luc Johan De Lathauwer, Valerie Van Es, Jialu Tang, Aaqib Saeed, Rudolph Ferdinand Spee, Yuan Lu, Hareld Marijn Clemens Kemps. Originally published in JMIR Formative Research (https://formative.jmir.org), 8.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

A Digital, Self-Management Behavior Change Intervention for People With Chronic Obstructive Pulmonary Disease: Cohort Study

Liam Knox[1], PhD; Rachel Rahman[2], PhD; Gareth Norris[2], PhD; Carol-Anne Davies[3], MSc; Kimberley Littlemore[4], BA; Claire Hurlin[3], BSN; Sam Rice[3], MBBS; Keir Lewis[3], MBBS, PhD

[1]Sheffield Institute for Translational Neuroscience, University of Sheffield, 385a Glossop Road, Sheffield, United Kingdom

[2]Department of Psychology, Aberystwyth University, Aberystwyth, United Kingdom

[3]Respiratory Department, Hywel Dda University Health Board, Carmarthen, United Kingdom

[4]Creative Department, eHealth Digital Media, Swansea, United Kingdom

Corresponding Author:
Liam Knox, PhD
Sheffield Institute for Translational Neuroscience, University of Sheffield, 385a Glossop Road, Sheffield, United Kingdom

Abstract

Background: People with chronic obstructive pulmonary disease (COPD) experience a range of limitations, which have a significant effect on their health. Self-management and pulmonary rehabilitation (PR) are key treatments for people with COPD; however, barriers often limit their uptake and adherence.

Objective: To overcome these barriers, a digital self-management intervention called PocketMedic (PM) was developed and evaluated in people with COPD alongside, and in addition, to PR.

Methods: A total of 53 participants were recruited to 1 of 3 groups: PM and PR, PM, or PR. Data were collected at baseline and 7 weeks (after the interventions had finished). Questionnaires on health-related quality of life, self-management knowledge, and disease knowledge were collected. Multivariate analysis of variances and ANOVAs were used to analyze the data.

Results: The analyses found that the improvements in those receiving PM were not statistically significantly different from those receiving PR, indicating that PM may replicate the benefits underpinning self-management behaviors observed in those attending PR. However, there were no additional benefits when participants received PM and PR in combination.

Conclusions: PM may be a useful treatment to support COPD self-management, especially when barriers prevent people with COPD receiving traditional services such as PR. The quantitative results suggest that PM may be less beneficial when delivered alongside PR. Feedback from participants indicated that they would prefer to receive PM while they were on the waiting list for PR, to support them during this time and alleviate the apprehensions associated with attending PR. Implications, limitations, and suggestions for future research are discussed.

Trial Registration: ClinicalTrials.gov NCT03263754; https://clinicaltrials.gov/ct2/show/NCT03263754

(JMIR Form Res 2025;9:e75683) doi:10.2196/75683

KEYWORDS

chronic obstructive pulmonary disease; self-management; telehealth; pulmonary rehabilitation; PocketMedic

Introduction

People with chronic obstructive pulmonary disease (COPD) experience a range of daily-living limitations that have a significant effect on health-related quality of life (HRQoL) [1]. Regular symptoms include dyspnea, sputum production, coughing [2], as well as anxiety and depression [3]. Although people with COPD are extensive health care users [4], the majority of their time is spent outside of health services [5]. Therefore, knowing how to self-manage their condition is important for people with COPD; supporting regular self-management and increasing HRQoL have been identified as crucial objectives for health care providers [6].

For people with COPD, self-management behaviors include a range of activities such as regular exercise, taking medication, and attending health care appointments [7]; however, these behaviors are often neglected by people with COPD [1,8,9]. Self-management interventions combine educational components with behavior change techniques to engender long-term change [10] and have been associated with multiple beneficial outcomes, such as increased HRQoL and cost savings for health care providers [11-14]. Despite their effectiveness,

XSL•FO
RenderX

these programs are often time-intensive for both people with COPD and health care staff [12,14-16].

Outside of specialist programs, self-management training is also a central component alongside structured, supervised exercise within pulmonary rehabilitation (PR) [17,18]. PR is a multidisciplinary group program recognized as one of the most effective interventions available for people with COPD [2,17,18]. However, attendance and adherence to PR are historically low [19-21], with a disruption to the usual routine, distance to the course, and a lack of transportation identified as common barriers [20]. In addition, access to PR is also a traditional problem, with some areas not able to provide the program [21,22].

To overcome the time-intensive and accessibility problems associated with current self-management and PR programs, this study used a digital, film-based, self-management intervention called PocketMedic (PM) and a parallel-group design to investigate its effectiveness in increasing self-management compared with, and in addition to, standard PR. This intervention is not intended to replace PR but is considered a potentially beneficial adjunct to the program, providing materials that people with COPD can use outside of PR, or an alternative when barriers restrict PR attendance. A diabetes version of PM has been used as an adjunct to health care previously, finding positive results [23]. Two hypotheses were formed: (1) people with COPD receiving solely PM would have increases in self-management knowledge, disease knowledge, and HRQoL, comparable to those receiving solely PR; and (2) people with COPD receiving PM in addition to PR would have larger increases in self-management knowledge, disease knowledge, and HRQoL, compared to those receiving PR or PM.

Methods

Study Design

This study used a parallel-group, cohort design. The project had 3 conditions: (1) people with COPD received both PM and PR (PM+PR), (2) only PM, or (3) only PR.

This study also measured the ability of PM and PR to increase self-determined, self-management motivation, alongside semistructured interviews. These findings have been published separately [24]; however, where qualitative feedback regarding PM was provided, this will be used to support the interpretation of this study's findings within the discussion.

Participants

Participants were recruited between September 2016 and March 2018 from Hywel Dda University Health Board (HDUHB) Wales, a semirural health region serving 3 counties, Carmarthenshire, Ceredigion, and Pembrokeshire. Consenting participants from Carmarthenshire who were due to attend PR were allocated to PM+PR or PR cohorts. Consenting participants from Ceredigion were allocated to PM (because PR was not available in this county at the time of this study). No one was recruited from Pembrokeshire.

Participants had to be eligible to attend PR, even if they did not receive this as part of their intervention. To attend PR for COPD within HDUHB, everyone must conform to the Global Initiative for Chronic Obstructive Lung Disease definition of COPD [2]: the participants must be ≥40 years old, have a 10 pack-year smoking history, and have spirometry less than 80% predicted. There were no exclusion criteria for the study.

Of the 55 people with COPD who consented, 2 withdrew from the study due to mental health reasons. Table 1 displays the demographics for the remaining 53 participants in the clinical trial. Figure 1 shows the flow of participants through the study.

Table . Participant demographics for each of the 3 research conditions.

Characteristics	PM[a]+PR[b] (n=28)	PM (n=7)	PR (n=18)
Age (years), mean (SD)	66 (5.54)	68.1 (10.9)	65.3 (7.1)
Sex, n			
Male	15	6	11
Female	13	1	7
Smoking status, n			
Current smoker	3	2	1
Nonsmoker	3	0	0
Ex-smoker	15	5	14
Missing	7	0	3
Amount smoked per day (cigarettes), mean (SD)	20.3 (18.2)	26.4 (17)	25.4 (10.2)

[a]PM: PocketMedic.

[b]PR: pulmonary rehabilitation.

XSL•FO
RenderX

Figure 1. Number of participants at each stage of the study. PM: PocketMedic; PR: pulmonary rehabilitation.

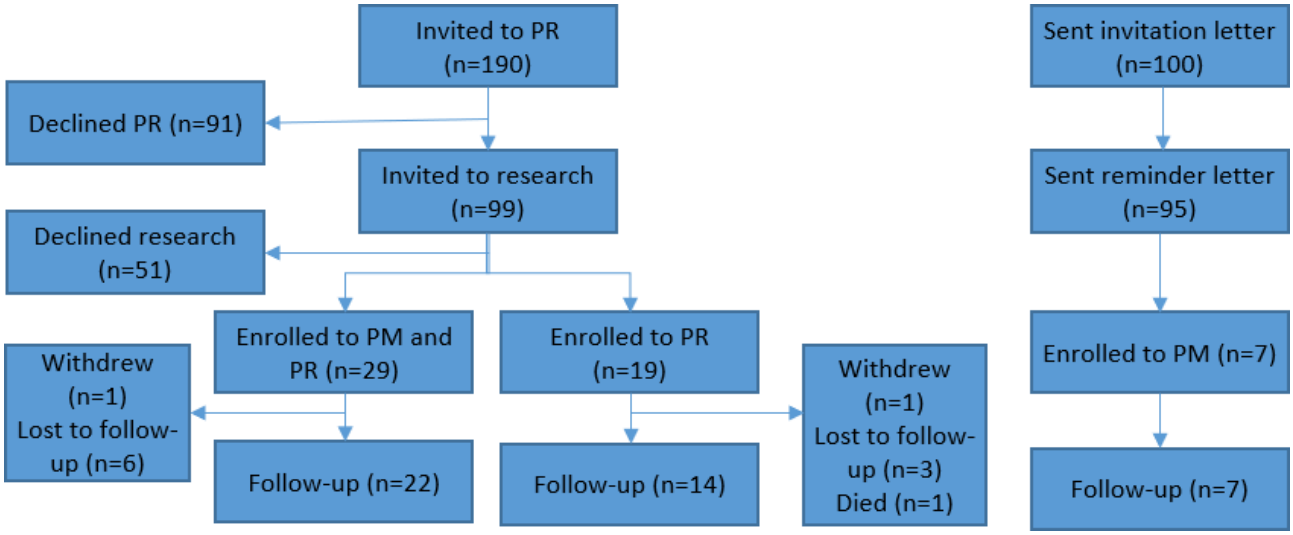

PocketMedic

The PM is a digital, film-based intervention designed with the intention of increasing health outcomes in people with COPD by improving self-management motivation and knowledge. This is accomplished by delivering educational components and applying the principles of self-determination theory (SDT) [25,26] to better support people in engaging with their self-management. Specifically, techniques within the films were designed with the intention of increasing the satisfaction of the 3 basic psychological needs (autonomy, competence, and relatedness) by providing participants with meaningful rationales and choice, addressing obstacles for change, and identifying available support. A recent study has provided support for the use of SDT to predict outcomes in people with COPD [27].

A total of 10 digital films, lasting approximately 5 minutes each, could be watched by participants using a smartphone, tablet, or computer, at a time and place that suited them. The films showed a mix of health care professionals and people with COPD discussing the illness and providing information on how to self-manage effectively. PM covers a range of topics, including breathlessness, physical activity, and well-being. The films were available to participants for 7 weeks and could be rewatched at any time if desired. Multimedia Appendix 1 provides a title and brief description of each of the films used in this study.

Pulmonary Rehabilitation

The study by Knox et al [28] describes in detail how HDUHB delivers PR, which is in line with best practice and current guidelines [17]. Briefly, approximately 7 - 10 participants (per cohort) attended twice-weekly sessions for 7 weeks. An experienced occupational therapist, physiotherapist, and assistant exercise instructor delivered the course, which consisted of structured exercise for 1 - 1.5 hours, followed by a 20 - to 40-minute educational talk, which included topics such as smoking cessation, mental health, and dietary education.

Data Collection

Participants completed questionnaires at baseline and at 7 weeks after the completion of PM+PR, PR, and PM. Questionnaires measured HRQoL, as assessed by the COPD assessment test[29]; self-management knowledge, as assessed by the Understanding COPD questionnaire (UCOPD) [30]; and disease knowledge, as assessed by the Bristol COPD Knowledge Questionnaire [31]. The COPD assessment test is negatively scored, with higher scores indicating a worse HRQoL. The UCOPD and Bristol COPD Knowledge Questionnaire are positively scored.

Adherence was operationalized as the number of PM films watched or PR sessions attended by participants. A maximum of 10 films could be viewed and 14 sessions attended.

All data were anonymized with participant-identifiable information replaced by a study number. Data were stored in a password-protected file, accessible only to members of the research team.

Analysis

All quantitative data were analyzed using SPSS (IBM Corp). ANOVAs and chi-square tests were used to identify any possible baseline differences between the 3 research conditions. A multivariate analysis of variance (MANOVA) was conducted to test the hypotheses using the test statistic that maximized statistical power due to the relatively small sample size. In case of a significant MANOVA, separate ANOVAs for each dependent variable were implemented. Pairwise *t* tests, using the Sidak or Games-Howell correction, were used for post hoc analyses. All tests used an α level of .05.

Approximately half of the participants allocated to receive PM did not watch any of the digital health films. Therefore, subgroup analyses were conducted focusing solely on those who engaged with the intervention to remove any concealing effects nonengaging participants may have had on the results. Similar techniques have been used in previous self-management intervention evaluations [32]. The subgroup analyses results are presented in the Results section below to avoid confusion.

Ethical Considerations

A favorable ethical opinion was provided by the Health Research Authority research ethics committee Wales 7 (16/WA/0130). The Declaration of Helsinki was followed throughout the study. All participants provided written informed consent. The

XSL•FO
RenderX

participants' privacy and confidentiality were maintained at all times. Funding was not available to compensate participants for their time.

Results

Overview

One-way ANOVAs and chi-square tests were conducted on participant demographics and baseline variables to identify any differences between participants in the 3 research conditions (Tables 1 and 2).

Table . Baseline descriptive statistics for each of the 3 research conditions.

Measures	PM[a]+PR[b] (n=28)	PM (n=7)	PR (n=18)
HRQoL[c], mean (SD)	27.7 (6.21)	28.6 (3.60)	27.1 (5.15)
Self-management knowledge, mean (SD)			
About COPD[d]	69.6 (14.8)	59 (24.1)	66.1 (14.3)
Managing symptoms of COPD	63.1 (14.2)	53.5 (21.7)	53.7 (22)
Accessing help and support	43.2 (18.5)	38.2 (20.8)	44.4 (18.9)
Total	61.2 (10.3)	52.2 (17)	56.5 (15.2)
Disease knowledge	47.9 (14.9)	35.7 (8.50) *[e]	50.3 (9.59) *[e]

[a]PM: PocketMedic.

[b]PR: pulmonary rehabilitation.

[c]HRQoL: health-related quality of life.

[d]COPD: chronic obstructive pulmonary disease.

[e]* denotes statistically significant differences at P=.04.

Only the variable disease knowledge showed a significant difference between groups at baseline ($F_{2,38}$=3.53; P=.04). Post hoc, pairwise t tests were implemented using the Games-Howell correction, finding statistically significant differences between participants receiving PM compared to those receiving PR (mean difference [MD] −14.5, 95% CI 0.51-28.6; P=.04). This indicates that, on average, those receiving PM alone had lower disease knowledge at baseline.

Table 3 displays average PM and PR, PM, or PR adherence for each of the 3 research conditions. As only 51.4% (18/35) of participants receiving PM+PR or PM watched one or more of the digital health films, average PM adherence (excluding nonengagers) was also calculated.

Table . Intervention adherence data for each of the 3 research conditions. Participants could watch a maximum of 10 PocketMedic films and attend 14 pulmonary rehabilitation sessions (depending on their research condition).

Intervention adherence	PM[a]+PR[b] (n=28)	PM (n=7)	PR (n=18)
PM adherence, mean (SD)	2.93 (4.06)	7 (4.12)	—[c]
PM adherence (only engagers)[d], mean (SD)	6.83 (3.35)	8.17 (2.99)	—
PR adherence, mean (SD)	10.33 (4.76)	—	9 (4.4)

[a]PM: PocketMedic.

[b]PR: pulmonary rehabilitation.

[c]Not available.

[d]For PM adherence (only engagers): n=12 for PM+PR; n=6 for PM.

To investigate whether access to PM was associated with increases in self-management knowledge, disease knowledge, and HRQoL, a repeated measures MANOVA was conducted. Using Pillai trace, there were no significant interaction effects between the 3 research conditions in any of the variables between baseline and follow-up (V=0.77, $F_{14,42}$=1.89; P=.06). However, when the research condition was held constant, there was a significant effect for the within-group variable of time (V=0.81, $F_{7,20}$=12.1; P<.001). The Bartlett test of sphericity was significant (χ^2_{27}=388; P<.001), and therefore the following ANOVAs used the Greenhouse-Geisser correction. The univariate tests found significant effects of time (between baseline and post) for all the included variables: HRQoL ($F_{1,26}$=20.1; P<.001), UCOPD: about COPD ($F_{1,26}$=18.2; P<.001), UCOPD: managing symptoms of COPD ($F_{1,26}$=10.3; P=.004), UCOPD: accessing help and support ($F_{1,26}$=21.6; P<.001), UCOPD: total ($F_{1,26}$=28; P<.001), and disease

XSL•FO
RenderX

knowledge ($F_{1,26}$=58.9; P<.001). All average scores at follow-up indicated more beneficial health states than those at baseline (Table 4).

Table . Baseline and follow-up descriptive statistics for all research measures for each of the 3 conditions.

Measures	PM[a]+PR[b]		PM		PR	
	Baseline	Follow-up	Baseline	Follow-up	Baseline	Follow-up
HRQoL[c], mean (SD)	27.7 (6.43)	23.6 (4.99)	28.2 (4.32)	21.4 (3.91)	26.2 (5.57)	24.7 (6.75)
Self-management knowledge, mean (SD)						
About COPD[d]	66.8 (13.9)	76.5 (14.6)	65.1 (23.9)	72.9 (21.4)	65.3 (13.6)	68.6 (7.77)
Managing symptoms of COPD	60.3 (15.2)	66.5 (24.6)	47.4 (23.4)	66.7 (23.6)	54.4 (12.59)	62.5 (13.2)
Accessing help and support	49.5 (16.4)	56.7 (21.9)	34 (23.9)	62.5 (20.4)	41.8 (14)	59.8 (9.89)
Total	60.4 (11.3)	68.1 (17.1)	51.3 (20.2)	68.5 (21.3)	55.8 (6.93)	64.3 (7.53)
Disease knowledge	46.8 (15.2)	60.8 (19)	38.4 (8.74)	70.9 (6.22)	52.2 (8.68)	60.6 (9.04)

[a]PM: PocketMedic.

[b]PR: pulmonary rehabilitation.

[c]HRQoL: health-related quality of life.

[d]COPD: chronic obstructive pulmonary disease.

Unexpectedly, despite the original MANOVA indicating that there were no significant differences between the conditions, the follow-up ANOVAs showed a significant interaction effect between the conditions and time points for the variable of disease knowledge ($F_{2,26}$=0.14; P=.003). However, pairwise, post hoc t tests using the Sidak correction found no significant differences between the research conditions (PM+PR vs PM: MD −0.88, 95% CI −17.2 to 15.4; P>.99; PM+PR vs PR: MD −2.63, 95% CI −15.6 to 10.3; P=.94; PM vs PR: MD −1.75, 95% CI −18.9 to 15.4; P=.99).

Subgroup Analysis

The subgroup analysis excluding participants who did not watch any of the PM digital health films was repeated using the same MANOVA and ANOVAs described above.

Using Roy largest root, significant main effects were found for both time (θ=9.50, $F_{7,12}$=16.3; P<.001) and research condition (θ=1.79, $F_{7,13}$=3.33; P=.03). The Bartlett test of sphericity was again significant (χ^2_{27}=257; P<.001); therefore, the following ANOVAs used the Greenhouse-Geisser correction. Univariate analyses found significant differences between the 2 time points for all the included variables: HRQoL ($F_{1,18}$=21.5; P<.001), UCOPD: about COPD ($F_{1,18}$=14.8; P<.001), UCOPD: managing symptoms of COPD ($F_{1,18}$=37; P<.001), UCOPD: accessing help and support ($F_{1,18}$=46.8; P<.001), UCOPD: total ($F_{1,18}$=71.2; P<.001), and disease knowledge ($F_{1,18}$=44.7, P<.001). All average scores at follow-up indicated more beneficial health states compared to those at baseline. Disease knowledge was the only variable to show a significant interaction effect between the 3 conditions and 2 time points ($F_{2,18}$=6.02; P=.01), with HRQoL and self-management knowledge total outside the required level of significance ($F_{2,18}$=3.31; P=.06 and $F_{2,18}$=3.37; P=.06, respectively). However, post hoc, pairwise t tests conducted on the variable of disease knowledge using the Sidak correction again found no significant differences between the 3 research conditions (PM+PR vs PM: MD 2.60, 95% CI −17.7 to 22.9; P=.98; PM+PR vs PR: MD 0.85, 95% CI −16.4 to 18.1; P>.99; PM vs PR: MD −1.75, 95% CI −20.1 to 16.6; P=.99).

Discussion

Principal Findings

This study aimed to investigate whether a digital, film-based, self-management intervention could improve self-management knowledge, disease knowledge, and HRQoL in people with COPD, by comparing it to, and alongside, a well-established PR program. The quantitative analyses indicated that the improvements from those receiving PM were not significantly different from those receiving PR, resulting in the first hypothesis being accepted. This indicates that PM may replicate some of the benefits underpinning self-management behaviors seen by those attending PR. However, there were no additional benefits when participants received PM+PR, resulting in the second hypothesis being rejected.

The lack of statistically significant differences between those receiving solely PM and those receiving solely PR does support the potential to deliver PM independently. Travel and transport, in addition to a disruption to the usual routine, have been highlighted as barriers to PR adherence [20]. Participant feedback indicated that a particular strength of the intervention was the ability to watch the health films whenever and wherever the participant wished, in addition to being able to rewatch them if something was missed. The film format was also considered engaging and could provide information in a clearer manner

XSL•FO
RenderX

than written leaflets. Therefore, the ability of a technology-delivered intervention to deliver significant benefits to self-management knowledge and understanding, potentially at par with PR, which can overcome these barriers, could prove beneficial [33]. This is supported by one of the key principles of the "A Healthier Wales" health care policy, being a reduction in geographical heterogeneity in the services provided [34]. Due to rurality and low staffing preventing PR from being delivered fully across HDUHB, PM may be a viable, effective, and suitable alternative, while recognizing the additional benefits that PR may offer through the structured, supervised exercise component.

There were some conflicting statistics associated with the variable of disease knowledge, with an original MANOVA in addition to post hoc *t* tests indicating no statistically significant differences between groups and an ANOVA reporting the reverse. Subsequently, the subgroup MANOVA and ANOVA found significant differences between the groups for disease knowledge, which were not replicated by the post hoc *t* tests. However, the analysis of baseline variables did indicate a significant difference between the groups for disease knowledge, where participants receiving PM had considerably lower levels of knowledge compared to the other 2 conditions. Therefore, it is possible that the contradictory statistics could be because of this baseline difference. Alternatively, the low sample size may have had an effect on the conflicting statistics; therefore, future research should replicate this study with a larger sample, in addition to using a randomized controlled trial design—the gold standard for evaluating the effectiveness of interventions.

The results of this study indicate that both PM and PR were effective at increasing self-management knowledge, disease knowledge, and HRQoL, corroborating previous literature regarding self-management interventions and PR [14,18,31,35]. However, the analyses showed that PM+PR did not result in a cumulative effect on the outcome measures. Although one possible interpretation is that there is a maximum level these variables can be improved in a 7-week period, collected feedback suggested that PM did not provide participants with anything additional to the PR learning, and, as a result, some participants stopped watching the health films altogether. Therefore, this feedback supports the interpretation that delivering PM alongside PR is not an effective way to distribute the health films. Several participants suggested that the interventions should be delivered independently of one another, with the consensus being that receiving PM before PR would be most beneficial. If delivered to people on a PR waiting list, this would enable people with COPD to receive self-management support at a time when they have limited treatment options, alleviating some of their concerns regarding the exercise components of PR. Fears of increased dyspnea due to exercise have been reported as a barrier to PR previously [36]; therefore, an intervention to reduce these fears may increase attendance. In addition, the UK's PR audit found that 31% of those invited to PR failed to attend the initial assessment [21,37], representing the highest dropout throughout the entire program. Therefore, prescribing PM before PR could decrease this non-uptake and result in positive benefits for people with COPD and health care services.

Figure 1 shows the number of participants throughout the project and illustrates the lack of response from those invited to PR. Although a 31% nonuptake rate has been reported previously [37], this study found that 48% of people with COPD invited to PR did not attend the initial assessment. HDUHB serves a wide, rural geographical area, with only 1 out of its 3 counties regularly delivering PR. Therefore, given that travel and transport are both negative predictors of PR attendance and adherence [20], this could explain the heightened nonattendance rate. The acceptance rate for the research was also relatively low. Overall, 28% (55/199) of those invited consented to the study; however, there was only a 4% response rate for the PM condition, which used a postal consent process. This is significantly worse than the average of 20% reported widely in other studies using similar methods [38]. Although several strategies were implemented to increase the acceptance rate across the 3 research conditions, such as a 6-month extension, the sample size remains a limitation of this project, and therefore, the authors do suggest care in interpreting these results. However, one of the justifications for conducting this research in people with COPD is that self-management motivation in this condition is poor, evidenced by low adherence to treatment [39,40]. Therefore, although unfortunate, it is not surprising that there were recruitment issues within this study, because such issues are also evident throughout other areas of care for people with COPD [41]. Future research should explore how best to engage people with COPD with low motivation in both research and clinical practice.

This poor uptake was also found toward PM, with only 51.4% (18/35) of all participants who received the intervention watching any of the health films. However, when looking at each group individually, participants receiving PM+PR had an uptake rate of 42.9% (12/28), compared to 85.7% (6/7) for those receiving PM. In addition, on average, people receiving PM watched 80% (8/10) of the films, which is a higher adherence rate compared to the PR adherence for people receiving PM+PR (10/14, 71%) and those receiving PR (n/N, 64%). Therefore, participants who had access to another intervention (ie, PR) were less likely to engage with PM, whereas those who only received the health films had a high uptake and adherence rate. Thus, if PM is delivered independent of PR, the uptake and adherence may meet standard published thresholds of acceptability [42].

Conclusions

In conclusion, this study provides novel contributions relevant to the care of people with COPD, such as the finding that although the interventions were effective separately, there were no cumulative benefits when the interventions were prescribed together. The lack of differences between the research groups does support the independent prescription of a technology-delivered, SDT-based, COPD self-management intervention, especially where environments prevent traditional programs from functioning fully. Participant feedback did support that prescribing PM to people with COPD before PR would be most beneficial. To the author's knowledge, no previous study has evaluated an intervention that combines self-management, digital delivery, SDT, and people with COPD. Future research should investigate the safety and effectiveness

XSL•FO
RenderX

of such an intervention within a larger sample size; however, this study does provide a proof of concept for increasing self-management knowledge, disease knowledge, and HRQoL in people with COPD through PM.

Acknowledgments

The authors would like to acknowledge and thank all the participants who gave their time to participate in this study. In addition, the Carmarthenshire Pulmonary Rehabilitation team was extremely welcoming and knowledgeable, helping to support the study throughout. We acknowledge the favorable ethical opinion from the NHS Research Ethics Committee. This research received no specific grant from a funding agency in the public, commercial, or not-for-profit sectors. The University of Sheffield Institutional Open Access Fund funded the open access of this article. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) license to any Author Accepted Manuscript version arising.

Data Availability

The datasets generated or analyzed during this study are not publicly available due

to the ethical conditions of this project that require that data are not shared.

Conflicts of Interest

KL is one of the founders of PocketMedic but was not involved in the collection or analysis of data.

Multimedia Appendix 1
PocketMedic film titles.
[DOCX File, 13 KB - formative_v9i1e75683_app1.docx]

References

1. Russell S, Ogunbayo OJ, Newham JJ, et al. Qualitative systematic review of barriers and facilitators to self-management of chronic obstructive pulmonary disease: views of patients and healthcare professionals. NPJ Prim Care Respir Med 2018 Jan 17;28(1):2. [doi: 10.1038/s41533-017-0069-z] [Medline: 29343739]
2. Singh D, Agusti A, Anzueto A, et al. Global strategy for the diagnosis, management, and prevention of chronic obstructive lung disease: the GOLD science committee report 2019. Eur Respir J 2019 May;53(5):53. [doi: 10.1183/13993003.00164-2019] [Medline: 30846476]
3. Phan T, Carter O, Waterer G, et al. Determinants for concomitant anxiety and depression in people living with chronic obstructive pulmonary disease. J Psychosom Res 2019 May;120:60-65. [doi: 10.1016/j.jpsychores.2019.03.004] [Medline: 30929709]
4. Dhamane AD, Moretz C, Zhou Y, et al. COPD exacerbation frequency and its association with health care resource utilization and costs. Int J Chron Obstruct Pulmon Dis 2015;10:2609-2618. [doi: 10.2147/COPD.S90148] [Medline: 26664109]
5. Five Year Forward View: National Health Service England; 2014. URL: https://www.england.nhs.uk/wp-content/uploads/2014/10/5yfv-web.pdf [accessed 2025-09-01]
6. Kielmann T, Huby G, Powell A, et al. From support to boundary: a qualitative study of the border between self-care and professional care. Patient Educ Couns 2010 Apr;79(1):55-61. [doi: 10.1016/j.pec.2009.07.015] [Medline: 19709844]
7. Chronic Obstructive Pulmonary Disease in over 16s: Diagnosis and Management: National Institute for Health and Care Excellence; 2018. URL: https://www.nice.org.uk/guidance/ng115/resources/chronic-obstructive-pulmonary-disease-in-over-16s-diagnosis-and-management-pdf-66141600098245 [accessed 2025-09-01]
8. Kantorowski A, Wan ES, Homsy D, Kadri R, Richardson CR, Moy ML. Determinants and outcomes of change in physical activity in COPD. ERJ Open Res 2018 Jul;4(3):00054-02018. [doi: 10.1183/23120541.00054-2018] [Medline: 30083551]
9. Korpershoek Y, Bos-Touwen ID, de Man-van Ginkel JM, Lammers JW, Schuurmans MJ, Trappenburg J. Determinants of activation for self-management in patients with COPD. Int J Chron Obstruct Pulmon Dis 2016;11:1757-1766. [doi: 10.2147/COPD.S109016] [Medline: 27536087]
10. Effing TW, Vercoulen JH, Bourbeau J, et al. Definition of a COPD self-management intervention: international expert group consensus. Eur Respir J 2016 Jul;48(1):46-54. [doi: 10.1183/13993003.00025-2016] [Medline: 27076595]
11. Benzo RP, Abascal-Bolado B, Dulohery MM. Self-management and quality of life in chronic obstructive pulmonary disease (COPD): the mediating effects of positive affect. Patient Educ Couns 2016 Apr;99(4):617-623. [doi: 10.1016/j.pec.2015.10.031] [Medline: 26632024]
12. Turner A, Anderson J, Wallace L, Kennedy-Williams P. Evaluation of a self-management programme for patients with chronic obstructive pulmonary disease. Chron Respir Dis 2014 Aug 1;11(3):163-172. [doi: 10.1177/1479972314539979] [Medline: 24980127]

XSL•FO
RenderX

13. Khdour MR, Agus AM, Kidney JC, Smyth BM, McElnay JC, Crealey GE. Cost-utility analysis of a pharmacy-led self-management programme for patients with COPD. Int J Clin Pharm 2011 Aug;33(4):665-673. [doi: 10.1007/s11096-011-9524-z] [Medline: 21643784]
14. Zwerink M, Brusse-Keizer M, van der Valk P, et al. Self management for patients with chronic obstructive pulmonary disease. Cochrane Database Syst Rev 2014 Mar 19;2014(3):CD002990. [doi: 10.1002/14651858.CD002990.pub3] [Medline: 24665053]
15. Baker E, Fatoye F. Patient perceived impact of nurse-led self-management interventions for COPD: a systematic review of qualitative research. Int J Nurs Stud 2019 Mar;91:22-34. [doi: 10.1016/j.ijnurstu.2018.12.004] [Medline: 30669076]
16. Richardson J, Loyola-Sanchez A, Sinclair S, et al. Self-management interventions for chronic disease: a systematic scoping review. Clin Rehabil 2014 Nov;28(11):1067-1077. [doi: 10.1177/0269215514532478] [Medline: 24784031]
17. Bolton CE, Bevan-Smith EF, Blakey JD, et al. British Thoracic Society guideline on pulmonary rehabilitation in adults. Thorax 2013 Sep;68 Suppl 2:ii1-i30. [doi: 10.1136/thoraxjnl-2013-203808] [Medline: 23880483]
18. McCarthy B, Casey D, Devane D, Murphy K, Murphy E, Lacasse Y. Pulmonary rehabilitation for chronic obstructive pulmonary disease. Cochrane Database Syst Rev 2015 Feb 23;2015(2):CD003793. [doi: 10.1002/14651858.CD003793.pub3] [Medline: 25705944]
19. Jongepier L, Barlow R. Pulmonary rehabilitation (PR) capacity, uptake and completion rate in the East of England. European Respiratory Journal 2016 Sep. [doi: 10.1183/13993003.congress-2016.PA3584]
20. Keating A, Lee A, Holland AE. What prevents people with chronic obstructive pulmonary disease from attending pulmonary rehabilitation? A systematic review. Chron Respir Dis 2011;8(2):89-99. [doi: 10.1177/1479972310393756] [Medline: 21596892]
21. Steiner M, Holzhauer-Barrie J, Lowe D, et al. Pulmonary rehabilitation: steps to breathe better. National Chronic Obstructive Pulmonary Disease (COPD) Audit Programme: clinical audit of pulmonary rehabilitation services in England and Wales 2015. : Royal College of Physicians, British Thoracic Society; 2016 URL: https://www.hqip.org.uk/wp-content/uploads/2018/02/national-chronic-obstructive-pulmonary-disease-copd-audit-full-pulmonary-rehabilitation-report-2015-published-2016.pdf [accessed 2025-09-01]
22. Yohannes AM, Connolly MJ. Pulmonary rehabilitation programmes in the UK: a national representative survey. Clin Rehabil 2004 Jun;18(4):444-449. [doi: 10.1191/0269215504cr736oa] [Medline: 15180129]
23. Rice S, Cranch H, Littlemore K, Mortimer J, Platts J, Stephens JW. A pilot service-evaluation examining change in HbA1c related to the prescription of internet-based education films for type 2 diabetes. Prim Care Diabetes 2017 Jun;11(3):305-308. [doi: 10.1016/j.pcd.2017.02.002] [Medline: 28291678]
24. Knox L, Rahman R, Norris G, et al. Understanding the impact of COPD and self-management interventions: a self-determination theory perspective. Psychol Health Med 2025 May 2:1-19. [doi: 10.1080/13548506.2025.2487226] [Medline: 40314456]
25. Deci EL, Ryan RM. Intrinsic Motivation and Self-Determination in Human Behavior: Springer; 1985. [doi: 10.1007/978-1-4899-2271-7]
26. Deci EL, Ryan RM. Motivation, personality, and development within embedded social contexts: an overview of self-determination theory. In: The Oxford Handbook of Human Motivation: Oxford University Press; 2012:85-107. [doi: 10.1093/oxfordhb/9780195399820.013.0006]
27. Knox L, Norris G, Lewis K, Rahman R. Using self-determination theory to predict self-management and HRQoL in moderate-to-severe COPD. Health Psychol Behav Med 2021 Jun 6;9(1):527-546. [doi: 10.1080/21642850.2021.1938073] [Medline: 34150367]
28. Knox L, Dunning M, Davies CA, et al. Safety, feasibility, and effectiveness of virtual pulmonary rehabilitation in the real world. Int J Chron Obstruct Pulmon Dis 2019;14:775-780. [doi: 10.2147/COPD.S193827] [Medline: 31040656]
29. Jones PW, Harding G, Berry P, Wiklund I, Chen WH, Kline Leidy N. Development and first validation of the COPD assessment test. Eur Respir J 2009 Sep;34(3):648-654. [doi: 10.1183/09031936.00102509] [Medline: 19720809]
30. O'Neill B, Cosgrove D, MacMahon J, McCrum-Gardner E, Bradley JM. Assessing education in pulmonary rehabilitation: the understanding COPD (UCOPD) questionnaire. COPD 2012 Apr;9(2):166-174. [doi: 10.3109/15412555.2011.644601] [Medline: 22409266]
31. White R, Walker P, Roberts S, Kalisky S, White P. Bristol COPD knowledge questionnaire (BCKQ): testing what we teach patients about COPD. Chron Respir Dis 2006;3(3):123-131. [doi: 10.1191/1479972306cd117oa] [Medline: 16916006]
32. Bucknall CE, Miller G, Lloyd SM, et al. Glasgow supported self-management trial (GSuST) for patients with moderate to severe COPD: randomised controlled trial. BMJ 2012 Mar 6;344:e1060. [doi: 10.1136/bmj.e1060] [Medline: 22395923]
33. Wootton R. Twenty years of telemedicine in chronic disease management--an evidence synthesis. J Telemed Telecare 2012 Jun;18(4):211-220. [doi: 10.1258/jtt.2012.120219] [Medline: 22674020]
34. A healthier Wales: long term plan for health and social care. : Welsh Government; 2018 URL: https://www.gov.wales/sites/default/files/publications/2021-09/a-healthier-wales-our-plan-for-health-and-social-care.pdf [accessed 2025-09-01]

XSL•FO
RenderX

35. Bringsvor HB, Langeland E, Oftedal BF, Skaug K, Assmus J, Bentsen SB. Effects of a COPD self-management support intervention: a randomized controlled trial. Int J Chron Obstruct Pulmon Dis 2018;13:3677-3688. [doi: 10.2147/COPD.S181005] [Medline: 30510410]
36. Harris D, Hayter M, Allender S. Improving the uptake of pulmonary rehabilitation in patients with COPD: qualitative study of experiences and attitudes. Br J Gen Pract 2008 Oct;58(555):703-710. [doi: 10.3399/bjgp08X342363] [Medline: 18826782]
37. Steiner M, Holzhauer-Barrie J, Lowe D, et al. Pulmonary rehabilitation: time to breathe better. National chronic obstructive pulmonary disease (COPD) audit programme: resources and organisation of pulmonary rehabilitation services in England and Wales. : Royal College of Physicians; 2015 URL: https://www.hqip.org.uk/wp-content/uploads/2018/02/pulmonary-rehabilitation-organisational-report-time-to-breathe-better.pdf [accessed 2025-09-01]
38. Roberts PJ, Roberts C, Sibbald B, Torgerson DJ. Increasing response rates to postal questionnaires. Effect of incentives on response rates must be considered. BMJ 2002 Aug 24;325(7361):444. [Medline: 12201284]
39. Bourbeau J, Bartlett SJ. Patient adherence in COPD. Thorax 2008 Sep;63(9):831-838. [doi: 10.1136/thx.2007.086041] [Medline: 18728206]
40. Sheldon KM, Williams G, Joiner T. Self-Determination Theory in the Clinic: Motivating Physical and Mental Health Self-Determination Theory in the Clinic: Motivating Physical and Mental Health: Yale University Press; 2003. [doi: 10.12987/yale/9780300095449.001.0001]
41. Restrepo RD, Alvarez MT, Wittnebel LD, et al. Medication adherence issues in patients treated for COPD. Int J Chron Obstruct Pulmon Dis 2008;3(3):371-384. [doi: 10.2147/copd.s3036] [Medline: 18990964]
42. Charlesworth G, Burnell K, Hoe J, Orrell M, Russell I. Acceptance checklist for clinical effectiveness pilot trials: a systematic approach. BMC Med Res Methodol 2013 Jun 13;13:78. [doi: 10.1186/1471-2288-13-78] [Medline: 23758922]

Abbreviations

COPD: chronic obstructive pulmonary disease
HDUHB: Hywel Dda University Health Board
HRQoL: health-related quality of life.
MANOVA: multivariate analysis of variance
PM: PocketMedic
PR: pulmonary rehabilitation
SDT: self-determination theory
UCOPD: Understanding COPD Questionnaire

Edited by A Bucher; submitted 08.04.25; peer-reviewed by J Ambros-Antemate, K Heslop-Marshall; revised version received 24.07.25; accepted 11.08.25; published 07.10.25.

Please cite as:
Knox L, Rahman R, Norris G, Davies CA, Littlemore K, Hurlin C, Rice S, Lewis K
A Digital, Self-Management Behavior Change Intervention for People With Chronic Obstructive Pulmonary Disease: Cohort Study
JMIR Form Res 2025;9:e75683
URL: https://formative.jmir.org/2025/1/e75683
doi:10.2196/75683

© Liam Knox, Rachel Rahman, Gareth Norris, Carol-Anne Davies, Kimberley Littlemore, Claire Hurlin, Sam Rice, Keir Lewis. Originally published in JMIR Formative Research (https://formative.jmir.org), 7.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Probing the Relationship Between Perioperative Complications in Patients With Valvular Heart Disease: Network Analysis Based on Bayesian Network

Wenyuan Lu, MS; Kun Zhu, MD; Zhiliang Gao, MS; Yuanming Li, MS; Hanwei Tang, MD; Cheng Sun, MD; Jianfeng Hou, MD

Fuwai Hospital, Chinese Academy of Medical Sciences, Peking Union Medical College, No. 167, North Lishi Street, Beijing, China

Corresponding Author:
Jianfeng Hou, MD
Fuwai Hospital, Chinese Academy of Medical Sciences, Peking Union Medical College, No. 167, North Lishi Street, Beijing, China

Abstract

Background: Heart valve surgery is associated with a high risk of perioperative complications. However, current approaches for predicting perioperative complications are all based on preoperative or intraoperative factors, without taking into account the fact that perioperative complications are multifactorial, dynamic, heterogeneous, and interdependent.

Objective: We aimed to construct and quantify the association network among multiple perioperative complications to elucidate the possible evolution trajectories.

Methods: This study used the data from China Cardiac Surgery Registry (CCSR), in which 37,285 patients were included in the analysis. A Bayesian network was used to analyze the associations among 12 complications. Score-based hill-climbing algorithms were used to build the structure and the association between them was quantified using conditional probabilities.

Results: We obtained the network of valve surgery complications. A total of 13 nodes represented complications or death, and 34 arcs with arrows represented the directly dependent relationship between them. We identified clusters of complications that were logically related and not related and quantified the associations. The correlation coefficient between complications increases with the severity of the complications, ranging from 0.01 to 0.41. Meanwhile, the probability of death when multiple complications occurred was calculated. Even mild complications, when progressing to multiple organ dysfunction syndrome, result in a mortality rate of over 90%.

Conclusions: Our network facilitates the identification of associations among specific complications, which help to develop targeted measures to halt the cascade of complications in patients undergoing the valve surgery.

(JMIR Form Res 2025;9:e68710) doi:10.2196/68710

KEYWORDS

perioperative complications; valvular heart disease; bayesian network; network analysis; probability inference

Introduction

Valvular heart disease (VHD) is a leading cause of cardiovascular morbidity and mortality and affects >2% of the population [1]. Its treatment mainly relies on valve repair and replacement. The incidence of perioperative complications is estimated to remain as high as 10.7% [2]. The complications translate into prolonged hospital stays, additional health care costs, and increased mortality. Therefore, many studies have evaluated predictors of perioperative complications. Current approaches for identifying the significant risk factors associated with perioperative complications are mainly through traditional machine learning models. The risk factors involved in these models are all preoperative factors, and the perioperative complications are regarded as independent events, which ignore the consequences of complications. There is growing evidence that perioperative complications are not independent, and the first complications could lead to secondary complications or even death [3,4], such as retained blood within the pericardium increasing the risk of postoperative atrial fibrillation (AF) [5,6] and acute renal failure related to cardiac arrest and death, similar to the domino effect [3]. The evidence suggested that most perioperative complications are multifactorial, dynamic, and context-specific. It is likely that perioperative complications form a complex system contributing to progressive physiological and biochemical deterioration. However, traditional approaches missed some differences in the trajectory of occurrence from initial complications to death [7] and failed to chart a comprehensive landscape of perioperative complications. A better understanding of the development pattern of perioperative complications is the focus of surgical quality improvement initiatives and an ongoing challenge.

XSL•FO
RenderX

The complex relationships among perioperative complications could provide strategies to improve prediction of further outcomes and inform the target and timing of interventions. Meanwhile, the difficulties in estimating the likelihood that a certain complication is secondary to other complications or that the selected treatment will result in a certain outcome make clinical decision-making a challenging task. Bayesian networks have been proven to be a strong tool to model the complex dependence between many factors [8] and capture the way an expert understands the relationships among all the features [9]. Bayesian network is characterized by several capabilities: (1) model complex problems with causal dependencies where a significant degree of uncertainty is involved, (2) combine multiple sources of information including data and experts' judgment, (3) present as an interpretable graphical structure, and (4) model interventions and reason both diagnostically and prognostically [10]. Bayesian network modeling is widely used in fields like clinical decision support [11], interactions between multiple diseases [12], and also in diagnostic diseases [13,14]. Previous research suggested that predictions from Bayesian network models perform better than competing approaches such as stepwise logistic regression [15]. Tevis et al [16] applied the Bayesian network to visualize the relationships among multiple perioperative complications to identify those complications most strongly associated with adjacent complications. Shen et al [7] also used the Bayesian network model to construct and quantify the association network among multiple perioperative complications and to elucidate the possible evolution trajectories. Cardiac surgery is characterized by the high risk of cardiovascular and other complications [17]. Reducing perioperative complications is key to improving the prognosis of patients undergoing heart valve surgery. However, few studies have revealed and quantified the possible progression trajectories of perioperative complications in patients with VHD.

Hence, the focus of our study was to construct and quantify the association network among multiple perioperative complications to investigate whether we could identify predictable patterns of secondary complication occurrences in the postoperative period after the first complication based on the Bayesian network model. We aimed to explore critical pathways in the network to support the development of targeted postoperative prevention strategies in patients with VHD.

Methods

Data Sources and Patients

The patients who underwent valvular surgery were identified from the Chinese Cardiac Surgery Registry (CCSR) database from January 2016 to December 2018. The CCSR database prospectively collected in-hospital and 30-day outcomes data on valvular surgery patients from 94 hospitals nationwide [18]. Patients were included in the study if they underwent valvular surgery procedures or valve surgery combined with coronary artery bypass graft surgery (CABG). Patients with missing information on perioperative complications and those who have not undergone cardiopulmonary bypass (CPB) were excluded from this study.

Perioperative Complications

We selected 12 perioperative complications, prolonged intensive care unit (ICU) stays (defined as exceeding the 75th percentile for ICU length of stay [19]), prolonged mechanical ventilation (defined as more than 24 h by the Society of Thoracic Surgeons [STS] guidelines [20]), increased chest tube drainage (defined as exceeding the 75th percentile for chest tube drainage), postoperative AF, postoperative myocardial infarction (postoperative myocardial infarction [MI], defined as new-onset pathologic Q-wave or left bundle branch block on the 12-lead electrocardiogram [21]), sternal wound infections, pericardial tamponade, reoperation, secondary tracheal intubation, stroke (defined as an acute episode of focal or global neurologic dysfunction caused by brain, spinal cord, or retinal vascular injury as a result of hemorrhage or infarction in which the neurologic dysfunction lasts for more than 24 h), postoperative renal failure (RF), multiple organ dysfunction syndrome (MODS). Among the 12 outcomes, 6 outcomes were found in the Society of Thoracic Surgeons database, including prolonged mechanical ventilation, postoperative AF, sternal wound infections, reoperation, stroke, postoperative RF [22]. In addition, prolonged ICU stays, increased chest tube drainage, postoperative MI, and pericardial tamponade have been demonstrated to be associated with increased perioperative morbidity and mortality after cardiac surgery [23-26]. Secondary tracheal intubation is a common but critical issue in the management of patients requiring mechanical ventilation in the ICU [27], which may further worsen the overall prognosis of patients. MODS is also the major cause of mortality after heart valve surgery [28]. Therefore, the types of complications we included are representative and reasonable. Death was defined as death occurring during the hospitalization and after discharge from the hospital but before the end of the 30th postoperative day.

Statistical Analysis

We calculated the proportion of common complications of valvular surgery. The number of deaths and the mortality rate for patients with each complication were also calculated to reflect the severity of different complications. The rate of co-occurring complications was defined as the percentage of patients that developed each complication and who also had other complications. The phi coefficient (phi) quantified the degree of concordance between any two complications. We used the rate of co-occurring complications and the phi to identify the co-occurrence relationships between complications. In addition, we compared the perioperative characteristics between patients with and without perioperative complications using the χ^2 test. R software (version 4.2.3; R Foundation for Statistical Computing) was used for statistical analyses.

Construction and Evaluation of a Complications Network

To investigate further interactions among perioperative complications, we applied a Bayesian network to patients with at least one complication. A Bayesian network is a graphical representation of a joint probability distribution. Learning a Bayesian network implies two tasks: (1) structure learning: identifying the topology of the Bayesian network and (2)

XSL•FO
RenderX

parametric learning: estimating numerical parameters (conditional probabilities) given a network topology [29].

During the structure learning stage, we constructed a directed acyclic graph (DAG) to describe the dependencies between these nodes. The nodes from which an arrow pointed were referred to as the "parent" nodes and the nodes pointed to by an arrow were referred to as "child" nodes. The state of the "parent" nodes affects their "child" nodes [30]. Taking into account the forbidden directions of the arcs (from death to other complications), we adopted the hill-climbing algorithm based on a score search to create a Bayesian network [31]. In addition, we used bootstrap resampling (1000 replicates, each with a resultant candidate network structure) to obtain a stable network structure. The arcs with strength probability >0.75 and direction probabilities ≥0.5 were considered finally.

Parameter values were obtained by using the maximum likelihood algorithm in the parameter learning phase [32]. The parameters are used to describe the probability distribution for each node, based on which we calculated the probability of a complication predicted using the Bayesian network. When one or more complications occurred, the probability of the occurrence of any complication was predicted, which we used to quantify the association between complications.

All analyses were performed using R software, including the *UpSet*, *bnlearn*, and *gRain* packages.

Ethical Considerations

The CCSR database was generated in 2013, including 87 large cardiac centers across the country with routine data submission, strict data audits, and instant feedback. These sites are the leading cardiac centers in local regions and have many features that are common among large cardiac care centers in China. The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board of Fuwai Hospital (protocol code 2021‐1477; date of approval is August 11, 2021). Informed consent was obtained from all subjects involved in the study. Data were collected and analyzed using a pseudonymous form. The participants did not receive any form of financial compensation for participating in this trial.

Results

Characteristics of the Participants

Based on inclusion and exclusion criteria, 37,285 patients were included in the analysis. Table 1 shows that we compared the perioperative characteristics of patients with at least one complication with those without. Nearly all assessed perioperative characteristics showed significant differences between groups ($P<.05$). Compared with patients who did not develop perioperative complications, patients with at least one complication were more likely to be male and aged 65 years and older (2724/18,700, 14.6% vs 4686/18,585, 25.2%), have more comorbidities, and have longer CPB time (7539/18,700, 40.3% vs 11,102/18,585, 59.7%) and aortic cross-clamp (ACC) time (7701/18,700, 41.2% vs 10,702/18,585, 57.6%).

XSL•FO
RenderX

Table . Baseline characteristics of the participants in each sample.

Characteristic	Total, n (%)	Without complication (N=18,700), n (%)	With complication (N=18,585), n (%)	*P* value
Demographics				
Age ≥65 years	7410 (19.9)	2724 (14.6)	4686 (25.2)	<.01
Female	17,640 (47.3)	9353 (50)	8287 (44.6)	<.01
BMI ≥30	1147 (3.1)	538 (2.9)	609 (3.3)	.03
Comorbidity				
Smoking	14,901 (40.0)	7766 (52.1)	7135 (38.4)	<.01
Diabetes mellitus	2494 (6.7)	972 (5.2)	1522 (8.2)	<.01
Hypertension	9875 (26.5)	4402 (23.5)	5473 (29.4)	<.01
CRF[a]	1140 (3.1)	605 (3.2)	535 (2.9)	.05
COPD[b]	342 (0.9)	114 (0.6)	228 (1.2)	<.01
Peripheral vascular disease	530 (1.4)	240 (1.3)	290 (1.6)	.03
Stroke	1605 (4.3)	705 (3.8)	900 (4.8)	<.01
Heart failure	2286 (6.1)	901 (4.8)	1385 (7.5)	<.01
Arrhythmia	8883 (23.8)	3898 (20.8)	4985 (26.8)	<.01
Myocardial infarction	758 (2)	173 (0.9)	585 (3.1)	<.01
Symptomatic coronary artery disease	2722 (7.3)	812 (4.3)	1910 (10.3)	<.01
Echocardiography results				
EF[c]<50	3870 (10.4)	1464 (7.8)	2406 (12.9)	<.01
LVEDD[d,e]	18,768 (50.3)	9181 (49.1)	9587 (51.6)	<.01
LAD[f,g]	28,319 (76)	13,884 (74.2)	14,435 (77.7)	<.01
TRV[h]>2.9 m/s	12,540 (33.6)	5893 (31.5)	6647 (35.8)	<.01
Biochemical indexes				
High creatinine[i]	5935 (15.9)	2474 (13.2)	3461 (18.6)	<.01
TC[j]>5.2 mmol/L	5670 (15.2)	2877 (15.4)	2793 (15)	.34
LDL[k]>3.4 mmol/L	5501 (14.8)	2894 (15.5)	2607 (14)	<.01
CPB[l]				
Prolonged CPB time[m]	18,641 (50)	7539 (40.3)	11,102 (59.7)	<.01
Prolonged ACC[n] time[o]	18,403 (49.4)	7701 (41.2)	10,702 (57.6)	<.01
Mortality	816 (2.2)	82 (0.4)	734 (3.9)	<.01

[a]CRF: chronic renal failure.
[b]COPD: chronic obstructive pulmonary disease.
[c]EF: ejection fraction.
[d]LVEDD: left ventricular end-diastolic diameter.
[e]LVEDD: for women, LVEDD >50; for men, >54.
[f]LAD: left atrial diameter.
[g]LAD: for women, LAD >37; for men, >39.
[h]TRV: tricuspid regurgitation velocity.
[i]High creatinine: for women, creatinine >84 mmol/L; for men, >104 mmol/L
[j]TC: total cholesterol.
[k]LDL: low-density lipoprotein.

XSL•FO
RenderX

[l]CPB: cardiopulmonary bypass time.

[m]Prolonged CPB time: Prolonged CPB time was defined as CPB time longer than the median of the overall cohort.

[n]ACC:

[o]Prolonged ACC time: Prolonged ACC time was defined as ACC time longer than the median of the overall cohort.

Complication Occurrence and Mortality

In total, 18,585 patients developed at least one complication. Table 2 shows the frequency and mortality for each complication. Prolonged ICU stays, prolonged mechanical ventilation, and increased chest tube drainage were the 3 most common complications, occurring in nearly half of patients, while patients with these complications had the lowest mortality rate, accounting for 4.8% (n=448), 4.9% (n=460), and 6.8% (n=579), respectively. Our results showed that sternal wound infections and stroke were least common in valvular disease patients undergoing cardiopulmonary bypass (0.5%), but were associated with higher mortality (20.2% [n=19] and 35.9% [n=33], respectively). The mortality rate for patients with different complications varied widely. Except for MODS, postoperative RF was the more common complication with the highest mortality rate (n=223, 49.2%).

Table . Frequency and mortality of individual complications.

Complications	Total patients with complication, n (%)	Deaths after complication, n (%)
Increased chest tube drainage	9312 (50.1)	448 (4.8)
Prolonged ICU[a] stays	9321 (50.2)	460 (4.9)
Prolonged mechanical ventilation	8489 (45.7)	579 (6.8)
Postoperative AF[b]	484 (2.6)	39 (8.1)
Postoperative MI[c]	181 (1)	15 (8.3)
Sternal wound infections	94 (0.5)	19 (20.2)
Pericardial tamponade	163 (0.9)	33 (20.2)
Reoperation	1068 (5.7)	174 (16.3)
Secondary tracheal intubation	537 (2.9)	202 (37.6)
Stroke	92 (0.5)	33 (35.9)
Postoperative RF[d]	453 (2.4)	223 (49.2)
MODS[e]	299 (1.6)	274 (91.6)

[a]ICU: intensive care unit.

[b]AF: atrial fibrillation.

[c]MI: myocardial infarction.

[d]RF: renal failure.

[e]MODS: multiple organ dysfunction syndrome.

Co-Occurrence of Complications

The UpSet plot in Figure 1 describes the co-occurrence of 12 complications in the dataset, with the rate of co-occurrence on the left side. More than 40% of the participants who experienced complications had 2 or more complications. Nearly 90% of patients who died had 2 or more complications. The rate of co-occurrence of nearly all complications was more than 60%. Especially for some serious complications, such as stroke, postoperative RF, sternal wound infections, and MODS, tended to occur in clusters. The rate of co-occurrence of these complications exceeded 90%. Similarly, the correlation analysis showed the strong associations between postoperative RF, sternal wound infections, and MODS (see Figure 2). The strongest correlation was between postoperative RF and MODS (phi=0.41). Although some complications with lower mortality were milder, such as prolonged ICU stays, prolonged mechanical ventilation, and increased chest tube drainage, the rate of co-occurrence of these complications was also higher (63.5%, 73.2%, and 49.7%, respectively). These complications appeared to be the underlying factors for subsequent complications.

XSL•FO
RenderX

Figure 1. UpSet plot of the co-occurrence of complications. The number of each complication is represented on the left bar plot. Possible co-occurrence between complications is represented by dark circles, and their number is shown on the top bar plot. The rate of co-occurrence is shown on the left side of the dark circles. ICU: intensive care unit; AF: atrial fibrillation; MI: myocardial infarction; RF: renal failure; MODS: multiple organ dysfunction syndrome.

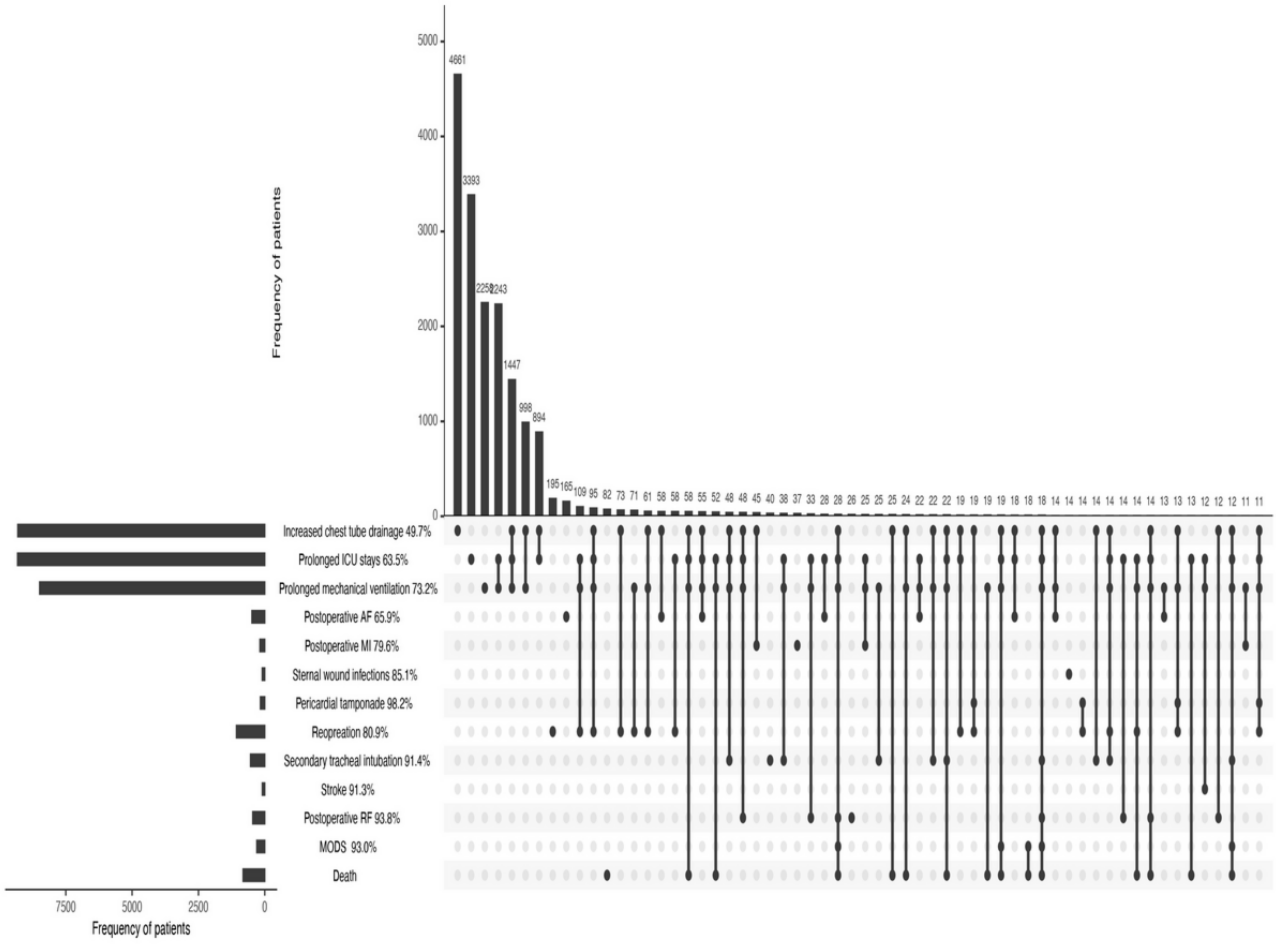

XSL•FO
RenderX

Figure 2. The heat map of the phi coefficients between complications. ICU: intensive care unit; AF: atrial fibrillation; MI: myocardial infarction; RF: renal failure; MODS: multiple organ dysfunction syndrome.

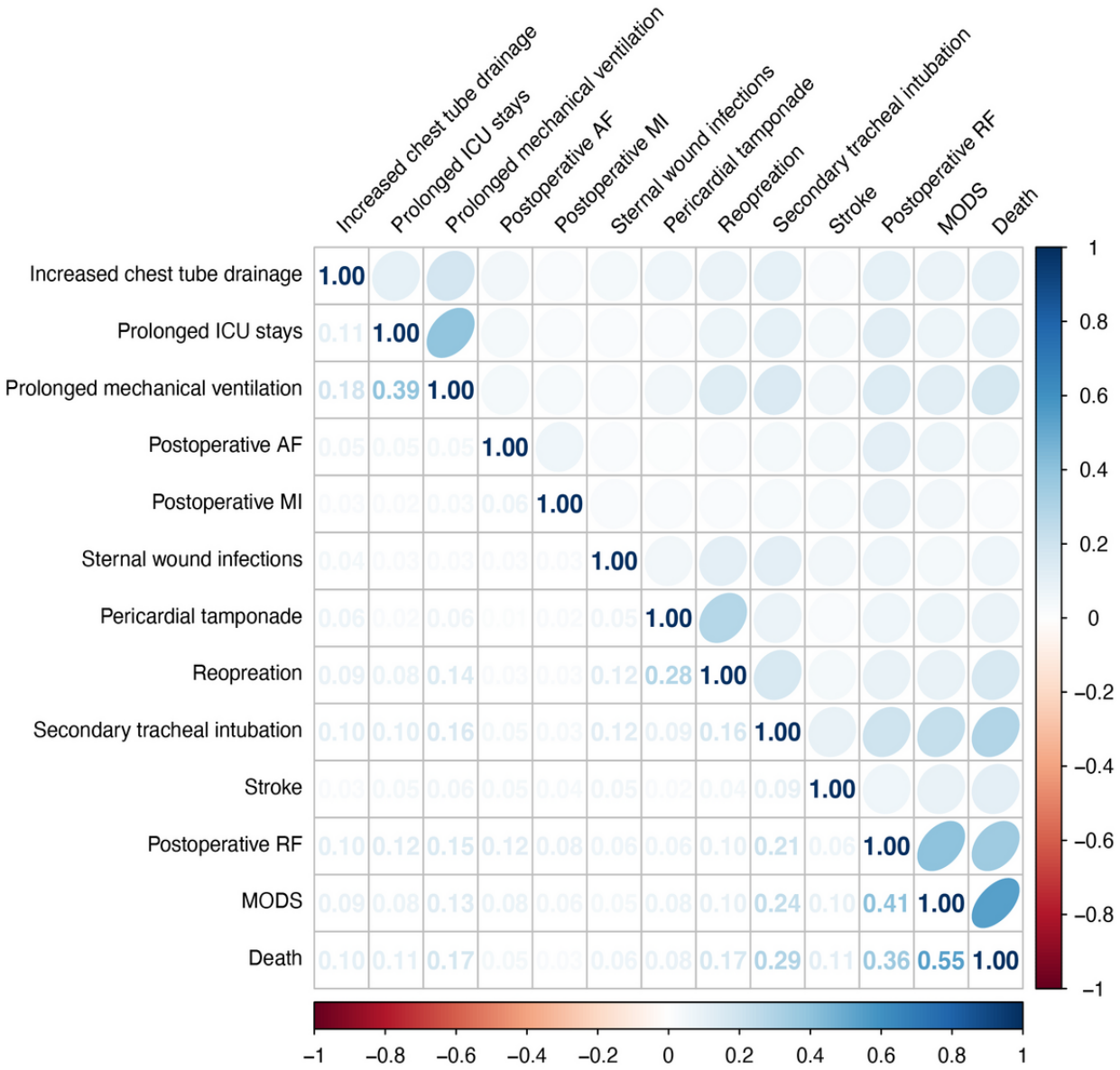

Construction of the Bayesian Network of Perioperative Complications

Figure 3 shows the Bayesian network modeling the relationships among complications in patients who developed complications. Each node in the figure represents a complication, and the arrows represent direct relationships among complications. There were 13 nodes and 34 directed arcs in the Bayesian network. All the arcs had a strength >0.85 and 30/34 arcs (88.2%) had a direction probability >0.70 (Multimedia Appendix 1). We identified complex relationships between multiple complications. In addition, by calculating the Markov blanket, we identified the critical complications of the network, namely secondary tracheal intubation, stroke, reoperation, and postoperative RF. Prolonged ICU stays, prolonged mechanical ventilation, and increased chest tube drainage appeared to be the starting points for these critical complications. By affecting the occurrence of postoperative RF, stroke, secondary tracheal intubation, and reoperation, prolonged ICU stays, prolonged mechanical ventilation, and increased chest tube drainage led indirectly to death, forming the path from primary complications to critical complications, ultimately leading to death. Postoperative AF, postoperative MI, sternal wound infections, and pericardial tamponade had no direct arc connected with death. In addition, there were more positively correlated arcs between serious complications than other complications, such as the strong relationships between the complications that could be connected to death and MODS, reflecting the cascading effect between complications.

XSL•FO
RenderX

Figure 3. Bayesian network of the complications and death. As the arc becomes thicker, the phi coefficient between the connected complications increases. ICU: intensive care unit; AF: atrial fibrillation; MI: myocardial infarction; RF: renal failure; MODS: multiple organ dysfunction syndrome.

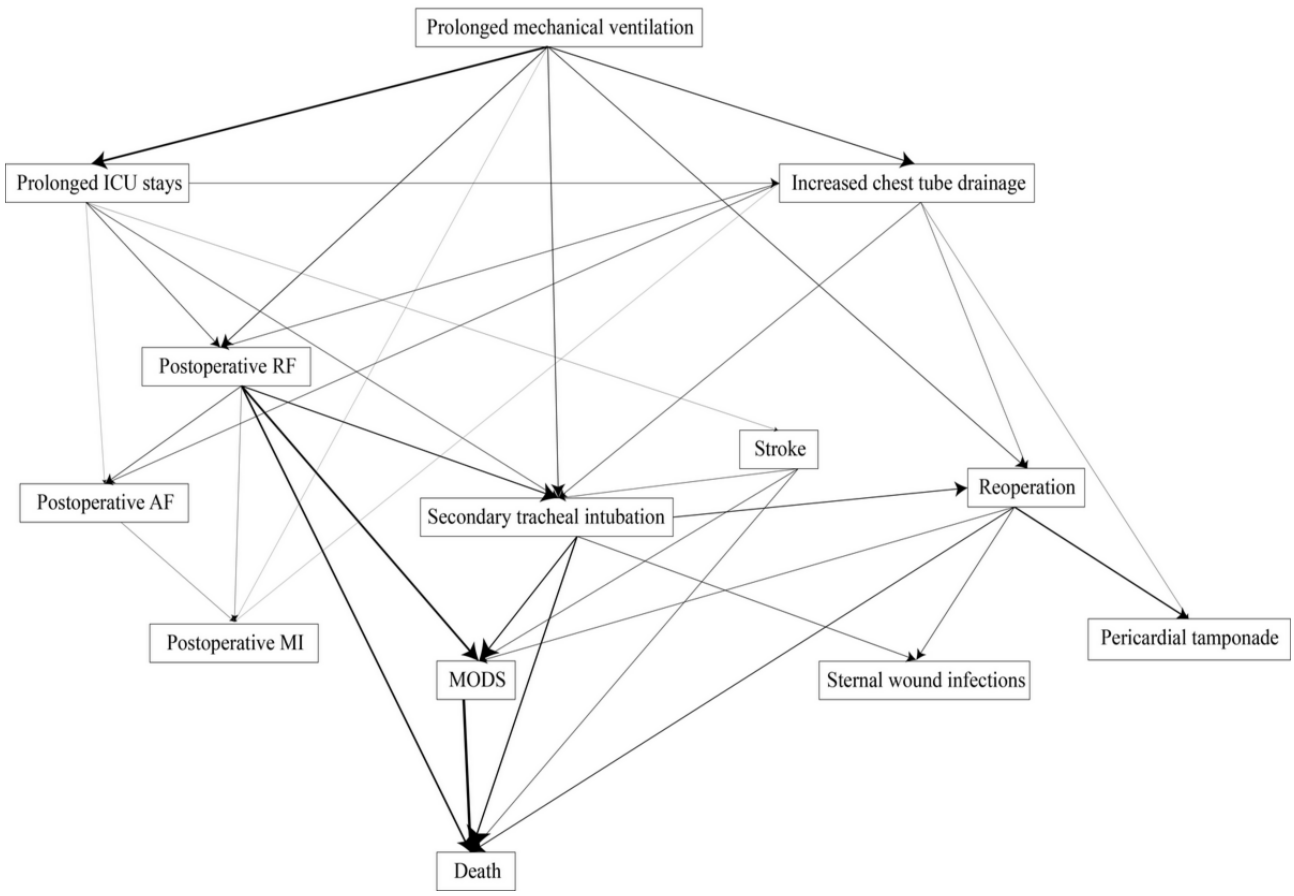

Probability Prediction Using Complications Network

The parameters (probability) of the Bayesian network were learned from the dataset containing 13 variables and 37,285 patients. When providing additional information regarding patients' complications, the probability of each complication occurring in the network can be automatically updated. Therefore, we calculated the probability of the occurrence of other complications and death, given the occurrence of 1 complication (see Table 3). The probability of death of more than 40% was the complications consisting of secondary tracheal intubation, stroke, postoperative RF, and MODS. Except for stroke, there is a close relationship between secondary tracheal intubation, postoperative RF, and MODS. When any one of these complications occurred, the calculated probability of the other complication is around 20% or more. In addition, we found that postoperative MI, sternal wound infections, and stroke were less likely to occur secondary to other complications. Given the occurrence of other complications, the probability of postoperative MI, sternal wound infections, and stroke was all below 10%. The occurrence of these complications appeared to be mainly associated with the preoperative or intraoperative factors. We also calculated the probability of death when assuming that more than 1 complication co-occurred (see Table 4). When one complication eventually caused MODS, the mortality rate exceeded 90%.

XSL•FO
RenderX

Table . The probability of the occurrence of the complications given the occurrence of one complication (%).

Interested complications	Evidence provided to the Bayesian network, %											
	Increased chest tube drainage	Prolonged ICU[a] stays	Prolonged mechanical ventilation	Postoperative AF[b]	Postoperative MI[c]	Sternal wound infections	Pericardial tamponade	Reoperation	Secondary tracheal intubation	Stroke	Postoperative RF[d]	MODS[e]
Increased chest tube drainage	__[f]	31.6	38.9	51.6	48	51.5	68.6	49.1	63	40	64.6	52.2
Prolonged ICU stays	32.1	—	55.4	48.4	38	51.5	45.1	48.4	59.1	66.7	66.1	55.6
Prolonged mechanical ventilation	36.3	51	—	40.6	50	60.6	51	62.6	76.2	52.4	74.8	62.2
Postoperative AF	3.3	3	2.8	—	16	4.3	5.9	5.1	4.4	0	14.3	7.3
Postoperative MI	1	0.8	1.1	5.2	—	0	2	1.5	1.1	0	7.1	2.3
Sternal wound infection	0.7	0.7	0.9	0.8	0	—	2	5.5	7.7	0	2.4	2.1
Pericardial tamponade	1.4	0.9	1.1	2.3	2.3	5.3	—	13.6	4.3	3.3	2.6	3.4
Reoperation	5.5	5.3	7.5	9	8	45.5	72.5	—	23.2	9.5	22.8	16.7
Secondary tracheal intubation	4.7	4.3	6.1	6.6	4.8	42.4	11.8	15.4	—	3.3	27.6	30
Stroke	0.5	0.6	0.5	0	0	0	2	0.7	0.7	—	0.7	5.6
Postoperative RF	3.4	3.4	4.2	12.9	18	9.1	7.8	10.6	19.3	4.8	—	63.3
MODS	1.9	2	2.5	0.8	3.3	10	6.5	5.5	19.1	16.7	37.7	—
Death	4.7	4.9	5.7	8.1	7	15	15.2	21.2	46.9	47.6	57	96

[a]ICU: intensive care unit.

[b]AF: atrial fibrillation.

[c]MI: myocardial infarction.

[d]RF: renal failure.

[e]MODS: multiple organ dysfunction syndrome.

[f]Not available.

XSL•FO
RenderX

Table . The probability of death given the occurrence of multiple complications.

Evidence provided to the Bayesian network			
Complication 1 (The probability of death when Complication 1 occurs, %)	Complication 2	Complication 3	The probability of death when complications 1 - 3 occur (%)
Prolonged ICU stays (4.9)	Stroke		42.9
	Stroke	Secondary tracheal intubation	58.6
	Stroke	MODS	99.5
	Secondary tracheal intubation		48.3
	Secondary tracheal intubation	Reoperation	57.1
	Secondary tracheal intubation	MODS	90.1
	Postoperative RF		57.0
	Postoperative RF	Secondary tracheal intubation	79.2
	Postoperative RF	MODS	93.3
Increased chest tube drainage (4.7)	Reoperation		26.4
	Secondary tracheal intubation		50.0
	Secondary tracheal intubation	Reoperation	53.3
	Secondary tracheal intubation	MODS	92.3
	Postoperative RF		61.9
	Postoperative RF	Secondary tracheal intubation	79.2
	Postoperative RF	MODS	97.7
Prolonged mechanical ventilation (5.7)	Reoperation		25.1
	Secondary tracheal intubation		46.7
	Secondary tracheal intubation	Reoperation	50.0
	Secondary tracheal intubation	MODS	93.3
	Postoperative RF		57.5
	Postoperative RF	Secondary tracheal intubation	78.1
	Postoperative RF	MODS	94.2

[a]ICU: intensive care unit.

[b]Not available.

[c]MODS: multiple organ dysfunction syndrome.

[d]RF: renal failure.

Discussion

Principal Findings

In this study, we constructed, for the first time, a new Bayesian network-based model of 12 perioperative complications and death from 37,285 VHD patients undergoing valve surgery in the CCSR. According to the network model, we demonstrated possible associations and visualized development paths between perioperative complications and quantified the pathways of complication occurrence and progression. Therefore, we identified critical pathways for the development of complications associated with valve surgery, which help to explore ways to halt the cascade of complications in patients undergoing the valve surgery in our future work. Different from previous studies, our finding provides an intuitive perspective that emphasizes the overall correlation for improving the prevention strategies of perioperative complications of VHD.

The complication network describes the dependencies between perioperative complications after valve surgery. The complication in the network may represent a point of intervention. The associations among complications provide the basis for the measures that will prevent or interrupt the occurrence of more serious complications. Our network analysis showed that increased chest tube drainage might contribute to the development of postoperative AF and postoperative RF, and prolonged mechanical ventilation increases the risk of acute kidney injury. The work is consistent with evidence for the role of higher drainage volume on the emergence of postoperative acute kidney injury and postoperative AF [33,34]. Kimura et al [35] also demonstrated that in pediatric cardiac patients, cardiac surgery-associated acute kidney injury was associated with prolonged mechanical ventilation. While discovering the logical

XSL•FO
RenderX

relationship between complications, we found stroke appears to be less related to other serious complications. Although previous studies have concluded that patients with postoperative AF have an increased risk of stroke, Rasmussen et al [36] demonstrated that there was no difference in the adjusted risk of postoperative stroke in postoperative AF versus non - postoperative AF patients, which agreed with our results. The distinct trajectory for stroke suggests that postoperative factors may not be directly associated with the occurrence of stroke. For the prevention of stroke, it may be more appropriate to intervene in preoperative or intraoperative factors. In addition, we found that the association between serious complications was stronger. When secondary tracheal intubation, postoperative RF, and MODS occurred, the probability of any other of these complications occurring was higher. Therefore, when rescuing a certain organ, joint prevention and intervention measures must be formulated to protect the functions of other organs and minimize the compounding effects of multiple complications [37].

We also explored the synergistic effect of multiple complications on mortality. An increase in the number of multiple complications shows a positive cumulative effect, suggesting the importance of special measures for serious complications such as real-time monitoring. In addition, we calculated mortality in the presence of multiple complications. Therefore, the network could be used to inform real-time risk prediction models that alert surgical teams to the changing risks of associated events for individual patients. For example, patients with postoperative RF or stroke may require careful selection of secondary tracheal intubation to avoid progression to more serious consequences. Our analysis suggests that for patients with valvular disease undergoing surgery, the prediction and prevention of complications may meaningfully extend into the postoperative period. Strategies based on the postoperative events could significantly improve our predictive capabilities and thus initiate strategies to reduce secondary complications.

In the past, it has been witnessed that some studies have applied network models to explore the complex relationships underlying a wide range of perioperative complications [16,38]. These studies initially found the relationship between perioperative complications. However, complication networks in the above studies are defined pairwise associations between diseases, which might miss important synergistic relationships in data. The uncertainties, including the progress of complications in individual patients, hospital work-flow management, and other factors, render an overall understanding hard to achieve. Now, our attempt provides a starting point to implement network theory to dissect the mechanistic basis of complex perioperative complications of cardiac surgery, from which to generate new discoveries traditional approaches fail to identify.

Our study had several limitations. First, the complications we included are not comprehensive. In the future, we will include more perioperative complications, such as paravalvular leakage and blood transfusion, to reveal the development process of perioperative complications. In addition, we will also include the information of postdischarge complications to improve completeness of research of complications. Second, our data lacks the ability to sequence complication, which prevents further confirmation of the causality. In the follow-up, we will collect the cohort recording the time of complication occurrence and further explore the associations of perioperative complication development using dynamic Bayesian networks. Third, high collinearity between variables will affect Bayesian network performance, including inaccurate parameter estimation, increased model complexity, unreliable inference results, and difficulty in structure learning. We evaluated the multicollinearity between variables, and the results showed that the Pearson correlation coefficients between all variables were below 0.6 (Multimedia Appendix 1). The highest Pearson correlation coefficient was 0.55, between MODS and mortality. In addition, we performed eigenvalue analysis and calculated the condition number (5.660998, with a condition number greater than 30 indicating severe multicollinearity). The results indicate that the variables are linearly independent, and multicollinearity is minimal. In any case, the role of the network model is predicated on a tremendous advance in profiling the complex complication relationship. The model may also find more promising applications in analyzing the comprehensive impact of preoperative impact on multiple complications and the design of more personalized prevention of perioperative complications.

Conclusions

In conclusion, we have created a Bayesian network that elucidates the underlying relationships and evolving progress of complications in valve surgery, and we have quantified the associations between 12 complications, which provides a basis for the development of targeted preventive measures to prevent further deterioration in patients undergoing valve surgery.

Acknowledgments

This study was conducted on behalf of the CCSR. We thank the colleagues from the CCSR coordinating board for their excellent work on data management and all members from the 94 participating hospitals. We thank our colleagues at Fuwai Hospital and the NCCD for their assistance with data audit and entry. This research was funded by the National Key R&D Program of China (2020YFC2008102) and National High Level Hospital Clinical Research Funding (2023-GSP-GG-12).

Data Availability

The data presented in this study are available upon request from the corresponding author. The data are not publicly available due to the No.717 Chinese Decree of the State Council, suggesting that the raw data of large samples cannot be shared globally.

XSL•FO
RenderX

Authors' Contributions

WL and JH contributed to conceptualization. WL, KZ, and JH contributed to data curation. WL, YL, and ZG contributed to formal analysis. JH handled funding acquisition. KZ contributed to the investigation. WL and KZ contributed to methodology. CS and JH contributed to project administration. JH handled resources. WL handled the software. HT, CS, and JH performed supervision. WL, YL, and ZG contributed to writing the original draft. All authors contributed to writing, review, and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary tables.
[DOCX File, 17 KB - formative_v9i1e68710_app1.docx]

References

1. Nkomo VT, Gardin JM, Skelton TN, Gottdiener JS, Scott CG, Enriquez-Sarano M. Burden of valvular heart diseases: a population-based study. Lancet 2006 Sep 16;368(9540):1005-1011. [doi: 10.1016/S0140-6736(06)69208-8] [Medline: 16980116]
2. Xu H, Liu Q, Cao K, et al. Distribution, characteristics, and management of older patients with valvular heart disease in China: China-DVD study. JACC Asia 2022 Jun;2(3):354-365. [doi: 10.1016/j.jacasi.2021.11.013] [Medline: 36338396]
3. Silber JH, Rosenbaum PR, Trudeau ME, et al. Changes in prognosis after the first postoperative complication. Med Care 2005 Feb;43(2):122-131. [doi: 10.1097/00005650-200502000-00005] [Medline: 15655425]
4. Wakeam E, Hyder JA, Jiang W, Lipsitz SA, Finlayson S. Risk and patterns of secondary complications in surgical inpatients. JAMA Surg 2015 Jan;150(1):65-73. [doi: 10.1001/jamasurg.2014.1795] [Medline: 25472013]
5. Gozdek M, Pawliszak W, Hagner W, et al. Systematic review and meta-analysis of randomized controlled trials assessing safety and efficacy of posterior pericardial drainage in patients undergoing heart surgery. J Thorac Cardiovasc Surg 2017 Apr;153(4):865-875. [doi: 10.1016/j.jtcvs.2016.11.057] [Medline: 28087110]
6. Butts B, Goeddel LA, George DJ, et al. Increased inflammation in pericardial fluid persists 48 hours after cardiac surgery. Circulation 2017 Dec 5;136(23):2284-2286. [doi: 10.1161/CIRCULATIONAHA.117.029589] [Medline: 29203568]
7. Shen Y, Zhang L, Wu P, et al. Construction and evaluation of networks among multiple postoperative complications. Comput Methods Programs Biomed 2023 Apr;232(107439):107439. [doi: 10.1016/j.cmpb.2023.107439] [Medline: 36870170]
8. Fuster-Parra P, García-Mas A, Ponseti FJ, Leo FM. Team performance and collective efficacy in the dynamic psychology of competitive team: a Bayesian network analysis. Hum Mov Sci 2015 Apr;40(98-118):98-118. [doi: 10.1016/j.humov.2014.12.005] [Medline: 25546263]
9. DeFelipe J, López-Cruz PL, Benavides-Piccione R, et al. New insights into the classification and nomenclature of cortical GABAergic interneurons. Nat Rev Neurosci 2013 Mar;14(3):202-216. [doi: 10.1038/nrn3444] [Medline: 23385869]
10. Kyrimi E, Dube K, Fenton N, et al. Bayesian networks in healthcare: what is preventing their adoption? Artif Intell Med 2021 Jun;116(102079):102079. [doi: 10.1016/j.artmed.2021.102079] [Medline: 34020755]
11. Sesen MB, Nicholson AE, Banares-Alcantara R, Kadir T, Brady M. Bayesian networks for clinical decision support in lung cancer care. PLoS One 2013;8(12):e82349. [doi: 10.1371/journal.pone.0082349] [Medline: 24324773]
12. Lappenschaar M, Hommersom A, Lucas PJF, Lagro J, Visscher S. Multilevel Bayesian networks for the analysis of hierarchical health care data. Artif Intell Med 2013 Mar;57(3):171-183. [doi: 10.1016/j.artmed.2012.12.007] [Medline: 23419697]
13. Charitos T, van der Gaag LC, Visscher S, Schurink KAM, Lucas PJF. A dynamic Bayesian network for diagnosing ventilator-associated pneumonia in ICU patients. Expert Syst Appl 2009 Mar;36(2):1249-1258. [doi: 10.1016/j.eswa.2007.11.065]
14. Maskery SM, Hu H, Hooke J, Shriver CD, Liebman MN. A Bayesian derived network of breast pathology co-occurrence. J Biomed Inform 2008 Apr;41(2):242-250. [doi: 10.1016/j.jbi.2007.12.005] [Medline: 18262472]
15. Miao M, Zeng H, Wang A, Zhao C, Liu F. Discriminative spatial-frequency-temporal feature extraction and classification of motor imagery EEG: an sparse regression and weighted naïve Bayesian classifier-based approach. J Neurosci Methods 2017 Feb 15;278(13-24):13-24. [doi: 10.1016/j.jneumeth.2016.12.010] [Medline: 28012854]
16. Tevis SE, Cobian AG, Truong HP, Craven MW, Kennedy GD. Implications of multiple complications on the postoperative recovery of general surgery patients. Ann Surg 2016 Jun;263(6):1213-1218. [doi: 10.1097/SLA.0000000000001390] [Medline: 27167563]
17. Ji F, Li Z, Nguyen H, et al. Perioperative dexmedetomidine improves outcomes of cardiac surgery. Circulation 2013 Apr 16;127(15):1576-1584. [doi: 10.1161/CIRCULATIONAHA.112.000936] [Medline: 23513068]
18. Rao C, Zhang H, Gao H, et al. The Chinese cardiac surgery registry: design and data audit. Ann Thorac Surg 2016 Apr;101(4):1514-1520. [doi: 10.1016/j.athoracsur.2015.09.038] [Medline: 26652141]

XSL•FO
RenderX

19. Martins RS, Waqar U, Raza HA, Memon MKY, Akhtar S. Assessing risk factors for prolonged intensive care unit stay after surgery for adult congenital heart disease: a study from a lower-middle-income country. Cureus 2023 Feb;15(2):e35606. [doi: 10.7759/cureus.35606] [Medline: 37007353]
20. Légaré JF, Hirsch GM, Buth KJ, MacDougall C, Sullivan JA. Preoperative prediction of prolonged mechanical ventilation following coronary artery bypass grafting. Eur J Cardiothorac Surg 2001 Nov;20(5):930-936. [doi: 10.1016/s1010-7940(01)00940-x] [Medline: 11675177]
21. Tiemuerniyazi X, Yang Z, Song Y, Xu F, Zhao W, Feng W. Coronary endarterectomy combined with coronary artery bypass grafting might decrease graft patency: a cohort study. Hellenic J Cardiol 2024;76:40-47. [doi: 10.1016/j.hjc.2023.07.001] [Medline: 37437778]
22. Bowdish ME, D'Agostino RS, Thourani VH, et al. STS Adult Cardiac Surgery Database: 2021 update on outcomes, quality, and research. Ann Thorac Surg 2021 Jun;111(6):1770-1780. [doi: 10.1016/j.athoracsur.2021.03.043] [Medline: 33794156]
23. Christensen MC, Dziewior F, Kempel A, von Heymann C. Increased chest tube drainage is independently associated with adverse outcome after cardiac surgery. J Cardiothorac Vasc Anesth 2012 Feb;26(1):46-51. [doi: 10.1053/j.jvca.2011.09.021] [Medline: 22100857]
24. Kuvin JT, Harati NA, Pandian NG, Bojar RM, Khabbaz KR. Postoperative cardiac tamponade in the modern surgical era. Ann Thorac Surg 2002 Oct;74(4):1148-1153. [doi: 10.1016/s0003-4975(02)03837-7] [Medline: 12400760]
25. Hassan A, Anderson C, Kypson A, et al. Clinical outcomes in patients with prolonged intensive care unit length of stay after cardiac surgical procedures. Ann Thorac Surg 2012 Feb;93(2):565-569. [doi: 10.1016/j.athoracsur.2011.10.024] [Medline: 22197534]
26. Paparella D, Guida P, Caparrotti S, et al. Myocardial damage influences short- and mid-term survival after valve surgery: a prospective multicenter study. J Thorac Cardiovasc Surg 2014 Nov;148(5):2373-2379. [doi: 10.1016/j.jtcvs.2013.10.061] [Medline: 24290712]
27. Gowardman JR, Huntington D, Whiting J. The effect of extubation failure on outcome in a multidisciplinary Australian intensive care unit. Crit Care Resusc 2006 Dec;8(4):328-333. [Medline: 17227270]
28. Husain-Syed F, Ricci Z, Brodie D, et al. Extracorporeal organ support (ECOS) in critical illness and acute kidney injury: from native to artificial organ crosstalk. Intensive Care Med 2018 Sep;44(9):1447-1459. [doi: 10.1007/s00134-018-5329-z] [Medline: 30043276]
29. Fuster-Parra P, Tauler P, Bennasar-Veny M, Ligęza A, López-González AA, Aguiló A. Bayesian network modeling: a case study of an epidemiologic system analysis of cardiovascular risk. Comput Methods Programs Biomed 2016 Apr;126(128-142):128-142. [doi: 10.1016/j.cmpb.2015.12.010] [Medline: 26777431]
30. Heckerman D, Geiger D, Chickering DM. Learning Bayesian networks: the combination of knowledge and statistical data. Mach Learn 1995 Sep;20(3):197-243. [doi: 10.1007/BF00994016]
31. García-Peña C, Ramírez-Aldana R, Parra-Rodriguez L, Gomez-Verjan JC, Pérez-Zepeda MU, Gutiérrez-Robledo LM. Network analysis of frailty and aging: empirical data from the Mexican health and aging study. Exp Gerontol 2019 Dec;128(110747):110747. [doi: 10.1016/j.exger.2019.110747] [Medline: 31665658]
32. Lauritzen SL. The EM algorithm for graphical association models with missing data. Comput Stat Data Anal 1995 Feb;19(2):191-201. [doi: 10.1016/0167-9473(93)E0056-A]
33. St-Onge S, Perrault LP, Demers P, et al. Pericardial blood as a trigger for postoperative atrial fibrillation after cardiac surgery. Ann Thorac Surg 2018 Jan;105(1):321-328. [doi: 10.1016/j.athoracsur.2017.07.045] [Medline: 29174782]
34. Wang Z, Ge M, Wang Z, et al. Identification of risk factors for postoperative stage 3 acute kidney injury in patients who received surgical repair for acute type A aortic dissection. BMC Surg 2022 Mar 2;22(1):75. [doi: 10.1186/s12893-022-01526-x] [Medline: 35236329]
35. Kimura S, Shimizu K, Iwasaki T, et al. Outcomes associated with unrecognized acute kidney injury in postoperative pediatric cardiac patients. Pediatr Nephrol 2023 Aug;38(8):2861-2871. [doi: 10.1007/s00467-023-05925-x] [Medline: 36929386]
36. Feilberg Rasmussen L, Andreasen JJ, Riahi S, et al. Risk and subtypes of stroke following new-onset postoperative atrial fibrillation in coronary bypass surgery: a population-based cohort study. J Am Heart Assoc 2022 Dec 20;11(24):e8032. [doi: 10.1161/JAHA.122.027010] [Medline: 36533595]
37. Kim M, Li G. Two-way interaction effects of perioperative complications on 30-day mortality in general surgery. World J Surg 2018 Jan;42(1):2-11. [doi: 10.1007/s00268-017-4156-7] [Medline: 28755257]
38. Alzubaidi AN, Karabayir I, Akbilgic O, Langham MR Jr. Network analysis of postoperative surgical complications in a cohort of children reported to the national surgical quality improvement program: pediatric. Ann Surg 2022 Jun 1;275(6):1194-1199. [doi: 10.1097/SLA.0000000000004234] [Medline: 33196492]

Abbreviations

AF: atrial fibrillation
CCSR: China Cardiac Surgery Registry
CPB: cardiopulmonary bypass
ICU: intensive care unit

XSL•FO
RenderX

MI: myocardial infarction
MODS: multiple organ dysfunction syndrome
RF: renal failure
VHD: valvular heart disease

Edited by A Mavragani; submitted 12.11.24; peer-reviewed by J Yang, MJ Khaledi; revised version received 29.03.25; accepted 02.04.25; published 07.10.25.

Please cite as:
Lu W, Zhu K, Gao Z, Li Y, Tang H, Sun C, Hou J
Probing the Relationship Between Perioperative Complications in Patients With Valvular Heart Disease: Network Analysis Based on Bayesian Network
JMIR Form Res 2025;9:e68710
URL: https://formative.jmir.org/2025/1/e68710
doi:10.2196/68710

©Wenyuan Lu, Kun Zhu, Zhiliang Gao, Yuanming Li, Hanwei Tang, Cheng Sun, Jianfeng Hou. Originally published in JMIR Formative Research (https://formative.jmir.org), 7.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Use of a Preliminary Artificial Intelligence-Based Laryngeal Cancer Screening Framework for Low-Resource Settings: Development and Validation Study

Shao Wei Sean Lam[1], MEng, PhD; Min Hun Lee[2], PhD; Michael Dorosan[1], MSc; Samuel Altonji[3], MD; Hiang Khoon Tan[4,5], MBBS, PhD; Walter T Lee[3], MD, MHS

[1]Health Services Research Centre, Singapore Health Services Pte Ltd, Ngee Ann Kongsi Discovery Tower Level 6, 20 College Road, Singapore, Singapore

[2]School of Computing and Information Systems, Singapore Management University, Singapore, Singapore

[3]Department of Head and Neck Surgery & Communication Sciences, Duke University Health System, Durham, NC, United States

[4]Division of Surgery and Surgical Oncology, Singapore General Hospital and National Cancer Centre Singapore, Singapore, Singapore

[5]SingHealth Duke-NUS Global Health Institute, Duke-National University of Singapore, Singapore, Singapore

Corresponding Author:
Shao Wei Sean Lam, MEng, PhD
Health Services Research Centre, Singapore Health Services Pte Ltd, Ngee Ann Kongsi Discovery Tower Level 6, 20 College Road, Singapore, Singapore

Abstract

Background: Early-stage diagnosis of laryngeal cancer significantly improves patient survival and quality of life. However, the scarcity of specialists in low-resource settings hinders the timely review of flexible nasopharyngoscopy (FNS) videos, which are essential for accurate triage of at-risk patients.

Objective: We introduce a preliminary AI-based screening framework to address this challenge for the triaging of at-risk patients in low-resource settings. This formative research addresses multiple challenges common in high-dimensional FNS videos: (1) selecting clear, informative images; (2) deriving regions within frames that show an anatomical landmark of interest; and (3) classifying patients into referral grades based on the FNS video frames.

Methods: The system includes an image quality model (IQM) to identify high-quality endoscopic images, which are then fed into a disease classification model (DCM) trained on efficient convolutional neural network (CNN) modules. To validate our approach, we curated a real-world dataset comprising 132 patients from an academic tertiary care center in the United States.

Results: Based on this dataset, we demonstrated that the IQM quality frame selection achieved an area under the receiver operating characteristic curve (AUROC) of 0.895 and an area under the precision-recall curve (AUPRC) of 0.878. When using all the image frames selected by the IQM, the DCM improved its performance by 38% considering the AUROC (from 0.60 to 0.83) and 8% considering the AUPRC (from 0.84 to 0.91). Through an ablation study, it was demonstrated that a minimum of 50 good-quality image frames was required to achieve the improvements. Additionally, an efficient CNN model can achieve 2.5-times-faster inference time than ResNet50.

Conclusions: This study demonstrated the feasibility of an AI-based screening framework designed for low-resource settings, showing its capability to triage patients for higher-level care efficiently. This approach promises substantial benefits for health care accessibility and patient outcomes in regions with limited specialist care in outpatient settings. This research provides necessary evidence to continue the development of a fully validated screening system for low-resource settings.

(JMIR Form Res 2025;9:e66110) doi:10.2196/66110

KEYWORDS

head and neck cancers; flexible nasopharyngoscopy; efficient neural nets; deep learning; cancer triage; machine learning; artificial intelligence

Introduction

Head and neck cancers (HNCs) are the 6th most common cancer worldwide, with a disproportionate growth in incidence and mortality in low- and middle-income countries (LMICs), particularly the West-Pacific and Southeast Asia regions [1-3]. Among HNCs, laryngeal cancer can be challenging to diagnose, with nonspecific and mild symptoms in the early stages. Early-stage diagnosis of laryngeal cancer is crucial to improve survival and quality of life [4]. Patients presenting with

XSL•FO
RenderX

early-stage cancers have a 60% - 90% chance of cure with local therapy, while those with late-stage cancers have a significantly reduced opportunity for remission [5,6]. In addition, patients with advanced cancers have worse quality of life due to their swallowing, verbal communication, and breathing dysfunctions [7].

The early detection of laryngeal cancer requires highly trained health care providers (eg, otolaryngologists) to visualize and interpret the relevant anatomical structures to detect anomalies. In addition, a definitive diagnosis requires downstream histopathological confirmation. Sophisticated endoscopic equipment, such as flexible nasopharyngoscopy (FNS), is necessary to examine the upper aerodigestive tract for abnormalities [8]. Experts who can perform this examination and interpret the endoscopic videos are limited in many low- and middle-income countries and in low-resource settings [9]. Limited access to specialty care in low- and middle-income countries is apparent for HNCs, with one study estimating the otolaryngologists-to-population ratio in some Asian countries to be as low as 1 per 2,146,000 [10]. This results in missed opportunities for early-stage diagnosis [2]. Technological advancements, particularly in fiberoptic flexible endoscopy and laser systems, have enabled the shift of many laryngological procedures from the operating room to outpatient clinics [11]. In high-volume outpatient settings, trained non-specialists may benefit from artificial intelligence (AI)-based clinical decision support systems encapsulating domain expert knowledge. Clinical decision support systems (CDSS) with embedded clinical practice guidelines, rules, and specialist knowledge may more effectively assess and triage the endoscopies performed by non-specialist health care workers while having the advantage of portability and accessibility [12].

AI, specifically machine learning and deep learning, is increasingly used to detect abnormalities in medical images and support cancer clinical decision-making, including screening, diagnosis, and prognosis [13-16]. The early application of deep learning and machine learning models in laryngeal cancer management has demonstrated the potential for detection capabilities comparable to human experts [16,17]. Deep convolutional neural networks (DCNNs) have been reported to deal with various data modalities for different use cases across the entire care chain [17,18]. These include real-time lesion detection [19-21] and segmentation [19,21], as well as screening, diagnosis [22], management, and prognosis of laryngeal cancers [16].

Deploying these AI models in the imaging practice presents several challenges. Imaging modalities use high-dimensional data. Applications that process single image frames account for the frames' pixel resolution and other features (eg, color channels). This "curse of dimensionality" effect, confronted by computational AI models, is compounded when considering video streams that capture many sequential image frames. For FNS procedures, the frame count ranges from hundreds to thousands, depending on the frame capture rate and the procedure's duration.

Furthermore, the high computational requirements of performant AI-based models must be considered in low-resource settings [23-28]. Another concern, especially for video-based procedures, such as screening for laryngeal cancer through FNS, is that the frames of interest may only lie within a range of non-blurry, contiguous frames that capture the anatomical landmark of interest (ie, the region of interest within each frame). This adds the challenge of localizing decision-making to a few clear and relevant regions and frames that best inform case escalation to more advanced diagnostic and treatment procedures. A previous study has proposed manually filtering frames to exclude low-quality frames (ie, blurry, noisy) before making an assessment [17]. Others have suggested various preprocessing steps to improve the quality of input images [29-32]; for instance, Huang et al [31] suggesting using the grayscale adaptive entropy value for setting the threshold to eliminate unclear images and recognize vocal fold disorders.

This formative study introduces an AI-based framework that denoises high-dimensional FNS videos, selects relevant frames, and suggests care escalation decisions through a referral grade classification task. To handle noisy real-world data and select relevant frames, our framework proposes an image quality module (IQM) that conducts a two-step procedure of filtering redundant images using a histogram of gradient-based threshold model and selecting good quality frames using supervised DCNN models. This IQM is used in conjunction with a disease classification module (DCM) that outputs a probability that a case should be escalated to appropriate downstream test and treat procedures. We aim to explore the use of efficient DCNN models and validate whether the proposed framework enhances the performance of correctly classifying cases to appropriate referral grades to address the resource constraints envisioned in less well-resourced settings [26,33].

Methods

Data Acquisition

Our study dataset has 132 full-color FNS videos of varying lengths collected from laryngoscopy procedures conducted in the Duke University Health System from December 2019 to December 2020. The shortest video was 5 seconds, while the longest was 165 seconds. The video clips were captured with various orientations, movements, and variable lighting and contrast conditions during the procedure. Some of these patients were healthy (no laryngeal pathology), some had benign disease processes, and some had laryngeal cancer. Patients were excluded if the videos were taken post-laryngectomy or if the larynx was not visualized on the video. Expert clinicians annotated the videos with medical conditions and referral levels for training classification models. The medical conditions were classified into three referral levels by a panel of 4 clinicians (two senior and two junior specialists): Grade 1, no referral required; Grade 2, non-urgent referral or close follow-up in 3 - 4 weeks; and Grade 3, urgent referral.

Ethical Considerations

All FNS videos were de-identified before analysis to protect patient privacy and confidentiality. The study was approved by the Duke University Health System Institutional Review Board (No. Pro00106209). The IRB granted a waiver of informed consent, as the study involved only de-identified data and posed

XSL•FO
RenderX

no risk to participants. No compensation was provided to participants. No identifiable individuals appear in any images or materials included in the manuscript or supplementary files.

AI-Based Framework for Laryngeal Cancer Screening

Overview

Figure 1 shows the proposed framework for screening patients receiving an FNS procedure. The framework includes two main components: the IQM and the DCM. The IQM filters low-quality and irrelevant images through a histogram of gradients-based threshold compared to an indexed 1st frame of each video. A U-Net model, trained with segmentation masks derived from the open-source benchmark for automatic glottis segmentation (BAGLS) dataset [30], was used to generate a labeled dataset for training the IQM to select relevant frames. The trained IQM network is then used to further refine the set of high-quality frames. Using selected frames, we train an efficient AI-based DCM to classify the referral grade. Figure 1 shows the schematic of the training and inference process based on the IQM and DCM.

Figure 1. Schematic of the proposed AI-based framework based on the IQM and DCM. AI: artificial intelligence; BAGLS: Benchmark for Automatic Glottis Segmentation; DCM: disease classification module; DHS: Duke University Health System; HOG: histogram of gradients; IQM: image quality module.

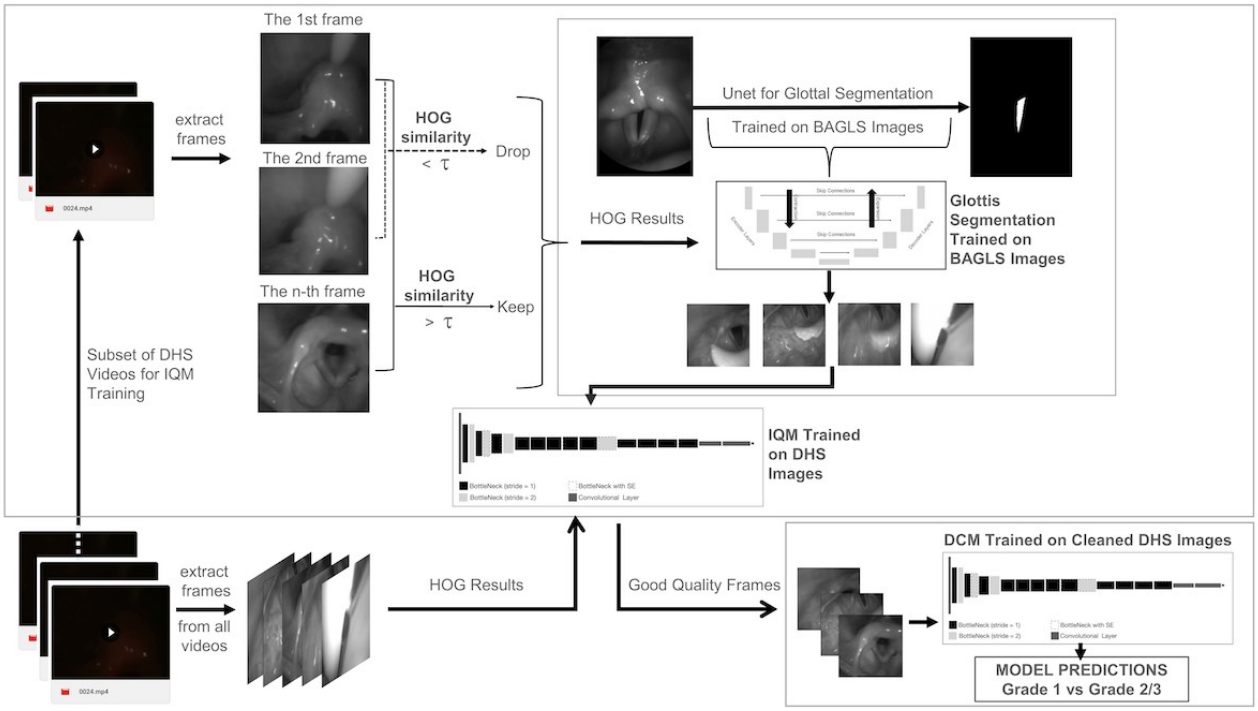

Image Quality Module

We developed a two-stage IQM based on (1) histogram of gradients-based similarity filtering, followed by (2) a U-Net-based DCNN module to identify a set of relevant, good-quality images. The correlation (ie, by cosine similarity) of features from a histogram of gradients [34] was used to evaluate the similarity of contiguous frames with the indexed frame (which is the first frame of relevance in the FNS process). After similarity filtering, the U-Net model was trained to identify the glottal region using the open-source BAGLS dataset [30]. Good-quality images are frames where the glottal area is entirely visible, regardless of the image's position and the glottis's size. We then assigned positive and negative quality labels to the Duke University Health System dataset using the U-Net model's predictions. Negative labels indicate poor quality due to obscuration and blurring by natural bodily secretions and movements or irrelevance, that is, frames not of the glottal region.

Disease Classification Module

The 132 unique patient videos were filtered into "good" and "poor" quality frames with the IQM. Sixteen videos were excluded due to insufficient good-quality image frames. The remaining 116 videos were used to train the DCM model using an 80 - 20 patient-level train-test split to avoid data leakage. The DCM classifies patients into binary referral grades: non-referral (Grade 1) versus referral (Grades 2 to 3).

To develop the DCM, we compared a baseline CNN model [35], a ResNet50-based model [36], a MobileNetV2-based model [33], and a GhostNet-based model [26] across six validation metrics. The baseline CNN model has six convolutional layers with maximum pooling and batch normalization, adding dropout [37] to the last two layers. ResNet50 employs deep residual learning with skip connections, enabling training of very deep neural networks without the challenge of vanishing gradients [38]. MobileNetV2 employs inverted residuals and depthwise separable convolutions for more efficient performance, intended for mobile and embedded devices. Lastly, GhostNet further enhances the efficiency of computations by generating more feature maps from efficient

XSL•FO
RenderX

operations; this results in an even more efficient DCNN suitable for lightweight applications.

Model Evaluation

The classification performance for IQM at the image level and DCM at the video (ie, patient) level was evaluated across six validation metrics. The train-test split was determined at the patient level to avoid data leakage. For the patient-level classification, we utilized the concept of bootstrap aggregation to evaluate the average classification probability of image frames within the same video. The primary metrics that describe the quality of the model predictions are accuracy, weighted F1 score, area under the receiver operating characteristic (AUROC), and area under the precision-recall curve (AUPRC). The secondary metrics that describe the efficiency of the models are the total number of floating-point operations in the order of 10^9 (GFLOPs) [39] and inference times. These outcome metrics describe the efficiency and effectiveness of the algorithm for model training, validation, and inference [23,26,28]. In addition, to address potential data imbalance that may hinder the classification model's ability to learn minor class patterns, the final selected model was further evaluated with training sample augmentation, and binary focal cross entropy loss [40]. To assess the impact of the IQM in the overall framework, we performed an ablation study [41] in which the DCM classifier was evaluated both with and without IQM-based preprocessing. Specifically, we trained and tested the DCM using input sequences that had undergone the IQM step. In addition, we systematically varied the number of high-quality frames provided to the DCM to examine the effect of input frame count on the classification performance.

Results

Comparison of the BAGLS Dataset and the Study Dataset

Table 1 summarizes the BAGLS and the study dataset. The BAGLS dataset has approximately 60% healthy patients, whereas our dataset has 30% healthy patients. The number of frames derived from the patient-level videos is roughly the same ratio. We used the entire BAGLS cohort, comprising 59,250 frames, to label informative frames. The raw dataset comprised 190,978 images derived from 132 patients.

Table . Summary of cohorts from the BAGLS[a] dataset and our dataset.

	BAGLS		Study dataset	
Disorder Status	Cohort size (%)	Patient count (%)	Frame count (%)	Cohort size (%)
Healthy (Grade 1)	35,400 (59.7)	382 (59.7)	49,282 (25.8)	40 (30.3)
Unhealthy (Grade 2/3)	23,850 (40.3)	258 (40.3)	141,696 (74.2)	92 (69.7)
Total	59,250 (100)	640 (100)	190,978 (100)	132 (100)

[a]BAGLS: Benchmark for Automatic Glottis Segmentation.

Performance of the Image Quality Module

Table 2 compares the test performance of the baseline CNN model [35], ResNet50-based model [36], and GhostNet-based model [26] for the IQM. Although the ResNet50-based model had the best accuracy of 0.833, the best F1 score of 0.832, and the best AUPRC of 0.957, the GhostNet model had comparable performance with the ResNet50 model and the best AUROC score of 0.895 with the fewest GFLOPs for computation.

The IQM model generated 20,040 good-quality frames from 116 patients in the study dataset cohort for the DCM training and test sets. Of these, 34/116 patients (29.3%) were classified as having Grade 1 disease, while the remainder 82/116 (70.7%) were classified as having Grade 2/3 disease. GhostNet resulted in the highest AUC-ROC and AUPRC, while being the most efficient, that is, the lowest GFLOPs.

Table . Comparison of the performance of the different deep convolutional neural network architectures used in the image quality module (IQM).

Model	Accuracy	F1 score	AUROC[a]	AUPRC[b]	GFLOPs[c]
Baseline convolutional neural network	0.699	0.673	0.724	0.729	50.0
ResNet50	0.833	0.832	0.746	0.957	245.0
GhostNet	0.829	0.827	0.895	0.878	8.7

[a]AUROC: area under the receiver operating characteristic curve.

[b]AUPRC: Area under the precision-recall curve.

[c]GFLOP: Number of floating-point operations in the order of 10^9.

Performance of the Disease Classification Module

The ResNet50 and GhostNet DCM achieved accuracy, optimal F1-scores, and AUPRC exceeding 80% at the video-level classification (Table 3). The ResNet50 model's inference time was 20.44 s, nearly 2.5 times slower than that of GhostNet (7.95 s per batch). Using an inference batch size of 64, 224-pixel-sized images (ie, height and width), ResNet50 had 245.0 GFLOPs, 40 times more than the GhostNet model with 8.7 GFLOPs.

XSL•FO
RenderX

Table . Performance comparison of different disease classification module (DCM) classifiers at the patient level.

Model	Accuracy	F1 score	AUROC[a]	AUPRC[b]	Inference time (s)	GFLOPs[c]
Convolutional neural network	0.652	0.624	0.595	0.805	8.09	50.0
ResNet50	0.739	0.697	0.667	0.850	16.71	245.0
MobileNetV2	0.696	0.629	0.611	0.833	8.62	20.3
GhostNet	0.870	0.863	0.833	0.912	7.95	8.7

[a]AUROC: area under the receiver operating characteristic curve.

[b]AUPRC: Area under the precision-recall curve.

[c]GFLOP: Number of floating-point operations in the order of 10^9.

Ablation Study

As the video-level prediction is based on bootstrap aggregation or bagging [42], the number of frames available to generate disease predictions (post-IQM) will be sensitive to the number of good-quality frames available per patient.

Table 4 shows the sensitivity of predictive quality across the number of high-quality frames. When using all the image frames selected by the IQM, the DCM improved its performance by 38% considering the AUROC (from 0.60 to 0.83) and 8% considering the AUPRC (from 0.84 to 0.91). Our results showed that 50 good-quality frames per patient video were required to outperform the model's base case without IQM.

Table . Ablation study results show GhostNet-based disease classification module performance at varying numbers (n) of good-quality frames per patient selected by the image quality module (IQM).

	Accuracy	F1 score	AUROC[a]	AUPRC[b]
Without IQM				
Original number	0.704	0.633	0.600	0.840
With IQM, n				
10	0.676	0.545	0.500	0.839
30	0.622	0.625	0.698	0.877
50	0.784	0.770	0.710	0.884
No limit[c]	0.870	0.863	0.833	0.912

[a]AUROC: area under the receiver operating characteristic curve.

[b]AUPRC: area under the precision-recall curve.

[c]All frames classified as good quality by the IQM are used.

Discussion

Principal Findings

This study showed the feasibility of an efficient AI-based screening framework incorporating an image quality filtering module to select high-quality and relevant image frames from FNS videos. Our ablation study demonstrated that the integration of IQM resulted in higher-quality DCM predictions across all the performance metrics at the patient level. Using a minimum of 50 high-quality frames, the DCM showed better predictive performance across all the metrics compared to the base model, where all the image frames were used without the IQM. Addressing the challenges of selecting informative image frames has been identified as a key impediment in developing laryngeal cancer screening algorithms [17]. Our formative research highlights the IQM's potential to enhance training and inference through effective frame selection.

We leveraged the efficient GhostNet architecture for our IQM and DCM as an alternative to the more resource-intensive ResNet50 model. GhostNet-based models have demonstrated performance comparable to those using less efficient architectures, such as ResNet50 [9]. In our study, the GhostNet-based DCM produced the best model across the validation metrics. The model achieved an accuracy of 87% and a high AUROC (0.833) and AUPRC (0.912) for classification at the patient level, with the optimal F1-score of 0.863 (Table 3). This level of performance, combined with the model's efficiency, makes it more suitable for integration into low-cost FNS facilities and screening equipment.

Our dataset, comprising 132 patients with 190,978 frames, is smaller than the dataset in a prior study [21], which trained and validated a segmentation model on data from 557 patients with 3933 frames and tested on two additional datasets. Nonetheless, limited patient datasets are common in this field. A related study [43] evaluated a CNN model on 100 patients with 170 images, while another study [44] used data from just 33 patients with 1320 images to assess machine learning algorithms.

Current state-of-the-art computer vision models use transformer-based AI models to classify images, segment pixels, or localize objects within images [45]. While achieving high

XSL•FO
RenderX

performance scores on established benchmarks, these models are computationally costly, with computational workloads exceeding those of the ResNets models [46-48]. Studies that compared traditional DCNNs, like those explored in this study, with transformer-based models highlighted greater computational costs and dependence on large training datasets [45,49,50]. Given these limitations, particularly in the context of deployment in low-resource clinical settings, there remains a strong case for exploring simpler, more efficient architectures. This study focused on efficient DCNNs to develop and validate the AI-based IQM-DCM screening framework for laryngeal cancer, emphasizing practical feasibility and predictive performance.

While our FNS videos reflect a realistic clinical setting, they may not fully represent the constraints of low-resource environments. This study serves as a preliminary step towards demonstrating the feasibility of the AI-based IQM-DCM screening framework. Acknowledging the limitations of our dataset, we are actively expanding data collection efforts with multiple partners to further enhance the framework's robustness and generalizability across diverse low-resourced clinical contexts [16,51]. Recent developments in efficient transformer network models will also be evaluated further to refine the dual-stage screening framework [52,53]. Cost-effectiveness analysis and implementation studies will also be conducted to achieve the envisioned system, which can support referral decisions in low-resource settings [16].

Conclusion

This study demonstrates the potential of the IQM-DCM framework to be embedded in an AI-based system to support early screening and triaging of patients at risk of laryngeal cancer. This preliminary work provides early evidence supporting the feasibility of this approach. Notably, the IQM-DCM framework, leveraged on lightweight neural network architectures, is shown to outperform conventional CNN models across various effectiveness and efficiency metrics. Future work will expand the dataset, incorporate recent advances in efficient network architectures, and validate the framework across more diverse populations to enhance its generalizability and real-world clinical applicability.

Acknowledgments

The study team would like to thank Ms. Zhenzhi Chen for managing this project and facilitating the necessary agreements that enabled its progress. This project is funded by the Duke/Duke-NUS Innovation Collaboration Pilot Grant (Duke/Duke-NUS/ICA(Pilot)/2020/0003). WTL wishes to acknowledge his joint appointment with the Duke-National University of Singapore.

Data Availability

Access to data used in this study is restricted to approved research collaborators due to concerns about privacy and institutional ethics policies.

Conflicts of Interest

None declared.

References

1. Gupta B, Johnson NW, Kumar N. Global epidemiology of head and neck cancers: a continuing challenge. Oncology (Williston Park, NY) 2016;91(1):13-23. [doi: 10.1159/000446117] [Medline: 27245686]
2. Patterson RH, Fischman VG, Wasserman I, et al. Global burden of head and neck cancer: economic consequences, health, and the role of surgery. Otolaryngol Head Neck Surg 2020 Mar;162(3):296-303. [doi: 10.1177/0194599819897265] [Medline: 31906785]
3. Schultz P. Vocal fold cancer. Eur Ann Otorhinolaryngol Head Neck Dis 2011 Dec;128(6):301-308. [doi: 10.1016/j.anorl.2011.04.004] [Medline: 21959270]
4. Unger J, Lohscheller J, Reiter M, Eder K, Betz CS, Schuster M. A noninvasive procedure for early-stage discrimination of malignant and precancerous vocal fold lesions based on laryngeal dynamics analysis. Cancer Res 2015 Jan 1;75(1):31-39. [doi: 10.1158/0008-5472.CAN-14-1458] [Medline: 25371410]
5. D'cruz A, Lin T, Anand AK, et al. Consensus recommendations for management of head and neck cancer in Asian countries: a review of international guidelines. Oral Oncol 2013 Sep;49(9):872-877. [doi: 10.1016/j.oraloncology.2013.05.010] [Medline: 23830839]
6. Chaturvedi P, Singhavi H, Malik A, Nair D. Outcome of head and neck squamous cell cancers in low-resource settings: challenges and opportunities. Otolaryngol Clin North Am 2018 Jun;51(3):619-629. [doi: 10.1016/j.otc.2018.01.008] [Medline: 29501327]
7. Hoare TJ, Thomson HG, Proops DW. Detection of laryngeal cancer--the case for early specialist assessment. J R Soc Med 1993 Jul;86(7):390-392. [doi: 10.1177/014107689308600707] [Medline: 8053995]
8. Alvi S, Harsha P. Flexible Nasopharyngoscopy: StatPearls Publishing; 2022. URL: https://www.ncbi.nlm.nih.gov/books/NBK539740/ [accessed 2025-09-23]

XSL•FO
RenderX

9. Araújo T, Santos CP, De Momi E, Moccia S. Learned and handcrafted features for early-stage laryngeal SCC diagnosis. Med Biol Eng Comput 2019 Dec;57(12):2683-2692. [doi: 10.1007/s11517-019-02051-5] [Medline: 31728933]
10. Smith JD, Tan KL. Workforce considerations, training, and diseases of the Asia-Pacific region. Otolaryngol Clin North Am 2018 Jun;51(3):659-665. [doi: 10.1016/j.otc.2018.01.010] [Medline: 29482921]
11. Huang JL, Khalid H, Alvaran KAB, Hey S, Watson N, Karagama Y. Improving laryngeal procedure workflow: moving from the operating room to the outpatient setting. Laryngoscope 2025 Mar;135(3):1132-1142. [doi: 10.1002/lary.31849] [Medline: 39460684]
12. Yue Y, Zeng X, Lin H, et al. A deep learning based smartphone application for early detection of nasopharyngeal carcinoma using endoscopic images. NPJ Digit Med 2024 Dec 31;7(1):384. [doi: 10.1038/s41746-024-01403-2] [Medline: 39738998]
13. Koh DM, Papanikolaou N, Bick U, et al. Artificial intelligence and machine learning in cancer imaging. Commun Med (Lond) 2022;2(1):133. [doi: 10.1038/s43856-022-00199-0] [Medline: 36310650]
14. Kourou K, Exarchos KP, Papaloukas C, Sakaloglou P, Exarchos T, Fotiadis DI. Applied machine learning in cancer research: a systematic review for patient diagnosis, classification and prognosis. Comput Struct Biotechnol J 2021;19:5546-5555. [doi: 10.1016/j.csbj.2021.10.006] [Medline: 34712399]
15. Kumar Y, Gupta S, Singla R, Hu YC. A systematic review of artificial intelligence techniques in cancer prediction and diagnosis. Arch Comput Methods Eng 2022;29(4):2043-2070. [doi: 10.1007/s11831-021-09648-w] [Medline: 34602811]
16. Bensoussan Y, Vanstrum EB, Johns MM, Rameau A. Artificial intelligence and laryngeal cancer: from screening to prognosis: a state of the art review. Otolaryngol Head Neck Surg 2023 Mar;168(3):319-329. [doi: 10.1177/01945998221110839] [Medline: 35787073]
17. Xiong H, Lin P, Yu JG, et al. Computer-aided diagnosis of laryngeal cancer via deep learning based on laryngoscopic images. EBioMedicine 2019 Oct;48:92-99. [doi: 10.1016/j.ebiom.2019.08.075] [Medline: 31594753]
18. Wang S, Chen Y, Chen S, Zhong Q, Zhang K. Hierarchical dynamic convolutional neural network for laryngeal disease classification. Sci Rep 2022;12(1):13914. [doi: 10.1038/s41598-022-18217-5]
19. He R, Jie P, Hou W, et al. Real-time artificial intelligence-assisted detection and segmentation of nasopharyngeal carcinoma using multimodal endoscopic data: a multi-center, prospective study. EClinicalMedicine 2025 Mar;81:103120. [doi: 10.1016/j.eclinm.2025.103120] [Medline: 40026832]
20. Baldini C, Migliorelli L, Berardini D, et al. Improving real-time detection of laryngeal lesions in endoscopic images using a decoupled super-resolution enhanced YOLO. Comput Methods Programs Biomed 2025 Mar;260:108539. [doi: 10.1016/j.cmpb.2024.108539] [Medline: 39689500]
21. Sampieri C, Azam MA, Ioppi A, et al. Real-time laryngeal cancer boundaries delineation on white light and narrow-band imaging laryngoscopy with deep learning. Laryngoscope 2024 Jun;134(6):2826-2834. [doi: 10.1002/lary.31255] [Medline: 38174772]
22. Xu ZH, Fan DG, Huang JQ, Wang JW, Wang Y, Li YZ. Computer-aided diagnosis of laryngeal cancer based on deep learning with laryngoscopic images. Diagnostics (Basel) 2023 Dec 14;13(24):3669. [doi: 10.3390/diagnostics13243669] [Medline: 38132254]
23. Liu S, Ha DS, Shen F, Yi Y. Efficient neural networks for edge devices. Comput Electr Eng 2021 Jun;92:107121. [doi: 10.1016/j.compeleceng.2021.107121]
24. Hu H, Peng R, Tai YW, Tang CK. Network trimming: a data-driven neuron pruning approach towards efficient deep architectures. arXiv. Preprint posted online on Jul 12, 2016. [doi: 10.48550/arXiv.1607.03250]
25. Jacob B, Kligys S, Chen B, et al. Quantization and training of neural networks for efficient integer-arithmetic-only inference. Presented at: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); Jun 18-23, 2018; Salt Lake City, UT. [doi: 10.1109/CVPR.2018.00286]
26. Han K, Wang Y, Tian Q, Guo J, Xu C, Xu C. GhostNet: more features from cheap operations. 2020 Presented at: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); Jun 13-19, 2020; Seattle, WA, USA p. 1577-1586. [doi: 10.1109/CVPR42600.2020.00165]
27. Zhang X, Zhou X, Lin M, Sun J. ShuffleNet: an extremely efficient convolutional neural network for mobile devices. Presented at: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); Jun 18-23, 2018; Salt Lake City, UT. [doi: 10.1109/CVPR.2018.00716]
28. Shi W, Cao J, Zhang Q, Li Y, Xu L. Edge computing: vision and challenges. IEEE Internet Things J 2016;3(5):637-646. [doi: 10.1109/JIOT.2016.2579198]
29. Dunham ME, Kong KA, McWhorter AJ, Adkins LK. Optical biopsy: automated classification of airway endoscopic findings using a convolutional neural network. Laryngoscope 2022 Feb;132 Suppl 4:S1-S8. [doi: 10.1002/lary.28708] [Medline: 32343434]
30. Gómez P, Kist AM, Schlegel P, et al. BAGLS, a multihospital benchmark for automatic glottis segmentation. Sci Data 2020 Jun 19;7(1):186. [doi: 10.1038/s41597-020-0526-3] [Medline: 32561845]
31. Huang CC, Leu YS, Kuo CFJ, Chu WL, Chu YH, Wu HC. Automatic recognizing of vocal fold disorders from glottis images. Proc Inst Mech Eng H 2014 Sep;228(9):952-961. [doi: 10.1177/0954411914551851] [Medline: 25313026]
32. Kuo CFJ, Lai WS, Barman J, Liu SC. Quantitative laryngoscopy with computer-aided diagnostic system for laryngeal lesions. Sci Rep 2021 May 12;11(1):10147. [doi: 10.1038/s41598-021-89680-9] [Medline: 33980940]

XSL•FO
RenderX

33. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. MobileNetV2: inverted residuals and linear bottlenecks. arXiv. Preprint posted online on Jan 13, 2018. [doi: 10.48550/arXiv.1801.04381]
34. Dalal N, Triggs B. Histograms of oriented gradients for human detection. Presented at: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05); Jun 20-25, 2025; San Diego, CA, USA p. 886-893. [doi: 10.1109/CVPR.2005.177]
35. Indolia S, Goswami AK, Mishra SP, Asopa P. Conceptual understanding of convolutional neural network- a deep learning approach. Procedia Comput Sci 2018;132:679-688. [doi: 10.1016/j.procs.2018.05.069]
36. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Presented at: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Jun 27-30, 2016; Las Vegas, NV, USA p. 770-778. [doi: 10.1109/CVPR.2016.90]
37. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 2014;15(56):1929-1958 [FREE Full text]
38. Hochreiter S. The vanishing gradient problem during learning recurrent neural nets and problem solutions. Int J Unc Fuzz Knowl Based Syst 1998 Apr;06(2):107-116. [doi: 10.1142/S0218488598000094]
39. Bangalore Vijayakumar S, Chitty-Venkata KT, Arya K, Somani AK. ConVision benchmark: a contemporary framework to benchmark CNN and ViT models. AI 2024;5(3):1132-1171. [doi: 10.3390/ai5030056]
40. Jadon S. A survey of loss functions for semantic segmentation. Presented at: 2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB); Oct 27-29, 2020; Via del Mar, Chile p. 1-7 URL: https://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=9276489 [accessed 2025-09-23] [doi: 10.1109/CIBCB48159.2020.9277638]
41. Meyes R, Lu M, de Puiseau CW, Meisen T. Ablation studies in artificial neural networks. arXiv. Preprint posted online on Jan 24, 2019. [doi: 10.48550/arXiv.1901.08644]
42. Breiman L. Bagging predictors. Mach Learn 1996 Aug;24(2):123-140. [doi: 10.1023/A:1018054314350]
43. Bengs M, Westermann S, Gessert N, et al. Spatio-spectral deep learning methods for in-vivo hyperspectral laryngeal cancer detection. Computer-Aided Diagnosis 2020:369-374. [doi: 10.1117/12.2549251]
44. Moccia S, De Momi E, Guarnaschelli M, Savazzi M, Laborai A. Confident texture-based laryngeal tissue classification for early stage diagnosis support. J Med Imag 2017 Sep;4(3):1. [doi: 10.1117/1.JMI.4.3.034502]
45. Takahashi S, Sakaguchi Y, Kouno N, et al. Comparison of vision transformers and convolutional neural networks in medical image analysis: a systematic review. J Med Syst 2024 Sep 12;48(1):84. [doi: 10.1007/s10916-024-02105-8] [Medline: 39264388]
46. Wang X, Zhang LL, Wang Y, Yang M. Towards efficient vision transformer inference: a first study of transformers on mobile devices. Presented at: Proceedings of the 23rd Annual International Workshop on Mobile Computing Systems and Applications; Mar 9-10, 2022; New York, NY, USA p. 1-7 URL: https://dl.acm.org/doi/10.1145/3508396 [accessed 2025-10-01] [doi: 10.1145/3508396.3512869]
47. Youn E, Prabhu SM, Chen S. Compressing vision transformers for low-resource visual learning. arXiv. Preprint posted online on Sep 5, 2023. [doi: 10.48550/arXiv.2309.02617]
48. Habib G, Saleem TJ, Lall B. Knowledge distillation in vision transformers: a critical review. arXiv. Preprint posted online on Feb 4, 2023. [doi: 10.48550/arXiv.2302.02108]
49. Maurício J, Domingues I, Bernardino J. Comparing vision transformers and convolutional neural networks for image classification: a literature review. Appl Sci (Basel) 2023 Jan;13(9):5521. [doi: 10.3390/app13095521]
50. Li J, Chen J, Tang Y, Wang C, Landman BA, Zhou SK. Transforming medical imaging with transformers? A comparative review of key properties, current progresses, and future perspectives. Med Image Anal 2023 Apr;85:102762. [doi: 10.1016/j.media.2023.102762] [Medline: 36738650]
51. Futoma J, Simons M, Panch T, Doshi-Velez F, Celi LA. The myth of generalisability in clinical research and machine learning in health care. Lancet Digit Health 2020 Sep;2(9):e489-e492. [doi: 10.1016/S2589-7500(20)30186-2] [Medline: 32864600]
52. Wang Y, Yang Y. Efficient visual transformer by learnable token merging. arXiv. Preprint posted online on Jul 21, 2024. [doi: 10.48550/arXiv.2407.15219]
53. Mehta S, Rastegari M. MobileViT: light-weight, general-purpose, and mobile-friendly vision transformer. arXiv. Preprint posted online on Oct 5, 2021. [doi: 10.48550/arXiv.2110.02178]

Abbreviations

AI: artificial intelligence
AUPRC: area under precision-recall curve
AUROC: area under the receiver operating characteristic curve
BAGLS: benchmark for automatic glottis segmentation
CNN: convolutional neural network
DCM: disease classification module

XSL•FO
RenderX

DCNN: deep convolutional neural network
FNS: flexible nasopharyngoscopy
GFLOP: number of floating-point operations in the order of 10^9
HNC: head and neck cancer
IQM: image quality module

Edited by A Mavragani; submitted 23.09.24; peer-reviewed by GK Gupta, PE Reynoso Aguirre, ST Arasteh; revised version received 31.07.25; accepted 05.08.25; published 07.10.25.

Please cite as:
Lam SWS, Lee MH, Dorosan M, Altonji S, Tan HK, Lee WT
Use of a Preliminary Artificial Intelligence-Based Laryngeal Cancer Screening Framework for Low-Resource Settings: Development and Validation Study
JMIR Form Res 2025;9:e66110
URL: https://formative.jmir.org/2025/1/e66110
doi:10.2196/66110

© Shao Wei Sean Lam, Min Hun Lee, Michael Dorosan, Samuel Altonji, Hiang Khoon Tan, Walter T Lee. Originally published in JMIR Formative Research (https://formative.jmir.org), 7.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Feasibility and Acceptability of a Positive Psychological Intervention for Patients With Metastatic Breast Cancer: Pre-Post Pilot Study

Claire C Conley[1], PhD; Elizabeth L Addington[2], PhD; Mikaela Velazquez-Sosa[1], BA; Brenna Mossman[1], PhD; Lesley Glenn[3], BS; Shontè Drakeford[4], MSN, CRNP, AGNP-C; Roxana Guerra[5]; Claudine Isaacs[1], MD; Ami Chitalia[6], MD; Christopher Gallagher[6], MD; Suzanne C O'Neill[1], PhD; Judith T Moskowitz[2], PhD

[1]Department of Oncology, Georgetown University, 2115 Wisconsin Ave NW, Suite 300, Washington, DC, United States

[2]Department of Medical Social Sciences, Feinberg School of Medicine, Northwestern University, Chicago, IL, United States

[3]Patient Author, Project Life MBC, Calabasas, CA, United States

[4]Patient Author, Upper Marlboro, MD, United States

[5]Patient Author, Washington, DC, United States

[6]MedStar Washington Hospital Center, Washington, DC, United States

Corresponding Author:
Claire C Conley, PhD
Department of Oncology, Georgetown University, 2115 Wisconsin Ave NW, Suite 300, Washington, DC, United States

Abstract

Background: Depression and anxiety are prevalent among patients with metastatic breast cancer (MBC), but there are few evidence-based psychological interventions specifically designed for this population.

Objective: This study aimed to assess the feasibility, acceptability, and clinical impact of a multicomponent positive psychological intervention, enhanced with an ecological momentary intervention for symptom management, for patients with MBC.

Methods: We recruited patients with MBC from a National Cancer Institute–designated comprehensive cancer center. Participants completed 5 weekly virtual individual sessions with a study counselor focused on positive emotion regulation skills. Participants also reported physical and psychological symptoms daily between sessions via SMS text messaging. Clinically elevated symptoms triggered a personalized coaching SMS text message tailored to the symptoms reported and the skills learned that week. Primary outcomes were intervention feasibility and acceptability. We also examined pre- to postintervention changes in depression, anxiety, positive affect, and positive emotion regulation skill use. Finally, a subset of participants completed qualitative exit interviews focusing on their experience in the study; interview data were analyzed using rapid qualitative analysis.

Results: We approached 20 patients with MBC, established contact with 15 (75%), received consent from 10 (67%), and retained 9 (90%) patients through the end of the study. Participants were 55 (SD 14.4, range 35-75) years old on average and identified as non-Hispanic White (5/10, 50%), non-Hispanic Black (4/10, 40%), or Latina (1/10, 10%). Participants attended 92% (46/50) of intervention sessions (mean 50, SD 9, range 36 - 71 min). On average, they completed 85% (SD 18%, range 46%-100%) of daily symptom assessments and received 23 (SD 5, range 13 - 31) coaching messages. Participants reported high perceived intervention feasibility (mean 4.81/5, SD 0.44), acceptability (mean 4.78/5, SD 0.33), and appropriateness for patients with MBC (mean 4.83/5, SD 0.35), above our a priori cutoff of ≥4. All 9 participants (n=9, 100%) recommended the intervention for other patients with MBC. We observed pre- to postintervention decreases in depression (d=−0.32) and anxiety (d=−0.27) and increases in positive affect (d=0.30) and positive emotion regulation skill use (d=0.99). Rapid qualitative analysis results demonstrate participants' positive experiences with the intervention, as well as suggestions for improvement.

Conclusions: This pilot study supports the feasibility of enrolling and retaining racially and ethnically diverse patients with MBC to this trial, the acceptability of the positive psychological intervention enhanced with ecological momentary intervention, and preliminary intervention impacts on depression, anxiety, positive affect, and positive emotion regulation skill use. A large-scale randomized controlled trial is needed to assess intervention efficacy for outcomes of interest.

(JMIR Form Res 2025;9:e77636) doi:10.2196/77636

KEYWORDS

advanced cancer; depression; anxiety; positive psychology; ecological momentary intervention

XSL•FO
RenderX

Introduction

Metastatic breast cancer (MBC) is breast cancer that has spread to other parts of the body; it is typically considered to be treatable, but incurable [1]. In the past decade, novel treatment approaches have resulted in significantly increased survival times following MBC diagnosis [2-4]. For this reason, the MBC population is anticipated to increase in the coming years; by 2030, an estimated 246,000 people in the United States will be living with MBC (a 59% increase from 2017) [5,6].

Living with MBC can affect psychological functioning. An estimated 20% to 52% of patients with MBC report clinically significant depression, and 36% to 60% report clinically significant anxiety [7-12]. However, most interventions to improve depression and anxiety in patients with breast cancer are designed for those with early-stage disease. The needs of patients with MBC are different from those with early-stage breast cancer [13], and resources that are not specific to MBC do not adequately meet the needs of this population [14]. Given their unique experiences, interventions addressing depression and anxiety, specifically for patients with MBC, are urgently needed.

There are few evidence-based interventions targeting depression and anxiety for patients with MBC. A 2023 systematic review identified 13 studies testing supportive care interventions designed to improve quality of life and symptom experience among patients with MBC [15]. Of the 7 studies that examined depression or anxiety as an outcome, only 2 demonstrated statistically significant improvement in these symptoms [16,17]. Thus, further research is needed to optimize interventions targeting depression and anxiety specifically for patients with MBC through the integration of novel intervention targets and methods.

Incorporating a focus on positive affect may be one way of optimizing psychological interventions for patients with MBC. Positive affect refers to pleasant feelings and moods, such as peace and joy [18]. Positive affect is *not* the same as absence of negative affect, as individuals can experience both positive and negative affect at the same time, and a high level of one does not necessitate a low level of the other [19,20]. Our prior work highlights the importance of positive affect among those with MBC and demonstrated that positive affect significantly predicted quality of life, above and beyond the effects of sociodemographic characteristics, clinical characteristics, and physical symptoms [21].

Positive psychological interventions (PPIs) are a promising approach to reduce depression and anxiety *and* increase positive affect among patients with MBC. PPIs focus on building practical skills designed to improve well-being [22,23]. Some PPIs focus on a single skill (eg, gratitude), while others teach several skills. In a prior pilot of a PPI among those with MBC, we demonstrated a reduction in depression but no improvement in positive affect [24]. Thus, prior pilot data suggested a need for further adaptations to the PPI to boost intervention effects on positive affect.

One potential explanation for these findings is the lack of PPI skill use during moments of higher negative affect when they are most needed. Nudge Theory suggests that behavior is susceptible to contextual factors (ie, how information is presented or structured) [25-28]. Thus, using nudges to influence choices is a potential strategy for behavior change [29]. Responsive behavioral nudges, triggered by participants' own symptom thresholds, engage patients when they are most receptive to help [30]. Compared to nudges that are not responsive (ie, daily reminders), responsive nudges can prevent habituation (ie, decreased response to an intervention over repeated exposures) [31]. Responsive interventions have been shown to improve positive affect in other populations and settings [32-34]. Thus, the previously observed null effects of the PPI on positive affect among patients with MBC might be addressed through responsive behavioral nudges.

In adapting the PPI in line with Nudge Theory, we incorporated an ecological momentary intervention (EMI). EMI delivers dynamic, individually tailored treatments to patients in real time and in natural settings [31,35]. EMI has several benefits over traditional "one-size-fits-all" interventions: maintaining participant engagement, sustaining continued behavior change for longer durations, and ultimately achieving greater intervention effects [36]. However, EMI has rarely been used in cancer survivorship research [37]. To our knowledge, there are no studies examining EMI in the context of advanced cancer.

To address these research gaps, we conducted a single-arm pilot study examining the feasibility and acceptability of a multicomponent PPI—enhanced with an EMI for symptom management—for patients with MBC. The aims of this study were 2-fold: (1) to demonstrate the feasibility and acceptability of the intervention for patients with MBC and (2) to examine pre- to postintervention changes in depression, anxiety, positive affect, and positive emotion regulation skill use.

Methods

Study Design

Following best practices for the development of behavioral interventions [38], we conducted a single-arm pilot study (see Checklist 1 for CONSORT [Consolidated Standards of Reporting Trials] checklist for reporting a pilot or feasibility trial [39]). Although this design does not allow us to assess intervention efficacy, we are able to assess intervention feasibility, acceptability, and clinical impact on outcomes of interest—the primary goals of this project. This research was designed and conducted with the active input of patient advocate collaborators to ensure the methodology was informed by the perspectives of individuals living with MBC. Qualitative data collection and analyses align with the COREQ (Consolidated Criteria for Reporting Qualitative Research; see Checklist 2) [40].

Participants and Procedures

Eligible participants were those diagnosed with MBC, assigned female at birth, aged ≥18 years, and able to communicate verbally in English. Participants were also required to have a working telephone capable of receiving SMS text messages and

XSL•FO
RenderX

accessing the internet to facilitate engagement with the EMI. We excluded individuals with a life expectancy of <6 months as assessed by their treating breast oncologist before study enrollment [41].

The target sample size for this feasibility study was 10 participants and was determined based on recommendations for pilot and feasibility studies in the palliative care setting [42]. Participants were recruited from a previous study cohort. Between January 2023 and April 2024, we conducted an observational study of quality of life among patients being treated for MBC at 5 medical oncology clinics in the mid-Atlantic region (N=125) [21]. A subset of participants (96/125, 77%) had agreed to be contacted for future studies. From this group, we selected individuals who had reported elevated levels of depression or anxiety (32/96, 33%), defined as a rating of 7 or higher on a 0 to 10 visual analog scale at any point throughout the study [43]. A member of the research team confirmed patients' vital status and eligibility with their treating oncologist; 27 (27/32, 84%) patients were alive, and we received oncologist permission to contact 22 (22/27, 81%) patients. Oncologists asked us not to contact 5 patients who had transferred care outside our health care system and were now under the care of another oncologist. Recruitment stopped upon achieving our target sample size of 10 consented participants.

From October to November 2024, a member of the research team contacted potential participants via telephone, introduced the study, and confirmed patient eligibility and interest. Following informed consent, participants completed a baseline (T0) survey assessing sociodemographic characteristics, depression, anxiety, positive affect, and positive emotion regulation skill use. On completion of the baseline survey, participants were scheduled for intervention sessions (see Intervention section below). Immediately post intervention (ie, after the fifth and final intervention session), participants completed a follow-up (T1) survey assessing primary and secondary outcomes (see Measures section below). Follow-up assessments took place between November and December 2024.

Finally, all participants were given the opportunity to complete an optional exit interview. A female research assistant with prior training and experience in qualitative interviewing (MVS) conducted the semistructured exit interviews. The interviewer did not have a relationship with the interviewees before the exit interview. At the start of the interview, the interviewer introduced herself and reminded participants of the purpose for the exit interview (ie, to obtain feedback on the PPI). In terms of positionality [44,45], the interviewer had foundational knowledge of MBC but not firsthand experience. She was also younger than the interview participants and from a different geographic region of the United States. These characteristics may introduce bias by affecting the interviewer's application of probing questions. Potential bias was managed through postinterview debriefing with the principal investigator (CCC).

Semistructured qualitative interviews explored participants' experiences with the intervention and suggestions for improvement (see Multimedia Appendix 1 for the interview guide). Interviewees were instructed to find a quiet, private location, without nonparticipants, for the duration of the interview. Interviews were conducted via telephone or video call (as preferred by the participant) and lasted 31 (SD 12, range 13-50) minutes on average. Interviews were audio recorded and transcribed verbatim. The interviewer took field notes during the interviews. As participants did not consent to be recontacted, we did not conduct repeat interviews, return transcripts to participants for review, or have participants provide feedback on the findings.

Intervention

Positive Emotion Skills Training (POET) is a PPI designed to bolster participants' ability to cope with major life stressors by increasing their experiences of positive affect [46]. Participants learned 8 evidence-based skills: noticing positive events, capitalizing (ie, maximizing positive experiences), gratitude, mindfulness, positive reappraisal, identifying personal strengths, creating attainable goals, and altruism and acts of kindness. POET includes 5 weekly individual sessions with a facilitator. In this study, sessions took place virtually (ie, via phone, 27/46, 59%, or video conference, 19/46, 41%, as preferred by the participant). Sessions lasted 50 (SD 9, range 36-71) minutes on average. Each session included (1) a brief check-in, (2) home practice review, (3) didactic portion in which the facilitator introduces the skills for that session, (4) interactive practice of the skills, and (5) discussion of home practice for the next week. Didactic content and home practice worksheets were provided as a paper intervention booklet; participants were encouraged to follow along with written content during the session and to document their home practice by writing in the booklet between sessions.

Participants also completed daily symptom reporting between POET sessions. Specifically, participants reported psychological symptoms (eg, depression, anxiety), physical symptoms (eg, pain, fatigue, gastrointestinal distress, nausea, reduced appetite, reduced libido), and positive experiences (eg, social connectedness, peace, joy) on a visual analog scale ranging from 0 to 10. Daily symptom reporting surveys were delivered via SMS text messages once per day, between 5 PM and 11 PM; to maximize the likelihood of survey completion, the exact time of survey delivery was individualized and based on participant preference.

Our intervention also incorporated an EMI, which was based on participants' responses to the daily symptom monitoring surveys. Specifically, high levels of symptoms on daily symptom monitoring surveys (scores≥7) triggered a next-day coaching message delivered by SMS text messages. Coaching messages were developed by adapting language from the existing intervention worksheets. Our multidisciplinary study team and patient advocates also provided input on the coaching messages to ensure their relevance to individuals living with MBC. The goal of the coaching messages was to encourage the use of positive emotion regulation skill(s) learned in POET sessions. Thus, coaching messages were specific to both the symptom(s) reported and the skill(s) that the participant learned in their most recent POET session. The coaching message was delivered the morning following the participant's evening symptom reporting survey, at approximately 10 AM.

XSL•FO
RenderX

Measures

Feasibility and Acceptability

At T1, participants completed the Feasibility of Intervention Measure, Acceptability of Intervention Measure, and Intervention Appropriateness Measure [47]. Each measure includes 4 items, which are rated on a 5-point Likert scale, ranging from "completely disagree" (1) to "completely agree" (5) and averaged to create scale scores ranging from 1 to 5 (higher=greater). Scores ≥4 indicate adequate levels of intervention feasibility, acceptability, and appropriateness.

In addition, participants provided feedback on the number of sessions in the POET intervention (too few, too many, or just right), the length of each POET session (too short, too long, or just right), preferred format for the POET intervention (eg, individual, group, self-guided website, in-person, virtual), and whether they would recommend the POET intervention for other people with MBC (yes or no).

Depression

Participants completed the Patient-Reported Outcomes Measurement Information System (PROMIS) Depression 4-item short form at T0 and T1 [48,49]. Items are rated on a 5-point Likert scale, ranging from "never" (1) to "always" (5), summed, and converted to T-scores, which enable comparison to the general population (mean 50, SD 10).

Anxiety

Participants completed the PROMIS Anxiety 4-item short form at T0 and T1 [48,50]. Items are rated on a 5-point Likert scale, ranging from "never" (1) to "always" (5), summed, and converted to T-scores, which enable comparison to the general population (mean 50, SD 10).

Positive Affect

Participants completed the positive affect subscale of the Modified Differential Emotions Scale (mDES) at T0 and T1 [51]. In total, 10 items are rated on a 5-point Likert scale, ranging from "never" (0) to "most of the time" (4), indicating the extent to which emotions have been experienced within the last 24 hours. Items are averaged to create a subscale score ranging from 0 to 4 (higher=more).

Positive Emotion Regulation Skill Use

As in prior work, we assessed participants' use of the positive emotion regulation skills taught in the POET intervention in the last week [24]. Participants respond to 11 items on a 5-point Likert scale, ranging from "definitely false" (1) to "definitely true" (5). Item scores are averaged to create a total score ranging from 1 to 5 (higher =greater).

Demographic Characteristics

At T0, participants self-reported their age, sex assigned at birth, gender identity, sexual orientation, race, ethnicity, education, employment status, annual household income, and insurance status.

Clinical Characteristics

Time since MBC diagnosis and disease subtype were abstracted from the electronic medical record.

Statistical Analysis

Quantitative

Descriptive statistics were used to characterize the sample and to examine measures of intervention feasibility, acceptability, and appropriateness. Student *t* tests were used to examine changes in depression, anxiety, positive affect, and positive emotion regulation skill use from T0 to T1. Analyses were conducted in SPSS (version 28; IBM Corp). All tests were 2-tailed, and significance was specified as $P<.05$. Cases with missing outcome data were deleted listwise.

Qualitative

Qualitative analysis was guided by a descriptive phenomenological approach, wherein the researchers sought to understand the unique lived experience of individuals by exploring the meaning of a phenomenon [52]. Exit interview transcripts were analyzed by 2 members of the study team (CCC, MVS) through a rapid qualitative analysis (RQA) procedure [53]. Main topics ("domains") were drawn from the interview guide, and a summary template was developed (see Multimedia Appendix 2 for the RQA template). Team members used the template to summarize interview transcripts. Summary templates were compiled into a single matrix reflecting the depth and breadth of information for each domain [54]. Although our sample size for qualitative exit interviews was limited to participants in the pilot intervention, we did obtain data saturation with 9 completed exit interviews [55].

Ethical Considerations

All study procedures were reviewed and approved by the institutional review board at Georgetown University (IRB #00008193). Study procedures ensured that participants' privacy and confidentiality were maintained. Patients provided written informed consent via an electronic informed consent form. Participants received a US $25 gift card on completion of each survey. Participants received a US $50 gift card on interview completion.

Results

Descriptive and Preliminary Analyses

We approached 20 patients with MBC and established contact with 15 (15/20, 75%) patients; 10 (10/15, 67%) patients provided consent (Figure 1). One participant withdrew due to prolonged hospitalization, and 9 (9/10, 90%) were retained through the study.

XSL•FO
RenderX

Figure 1. Study flowchart.

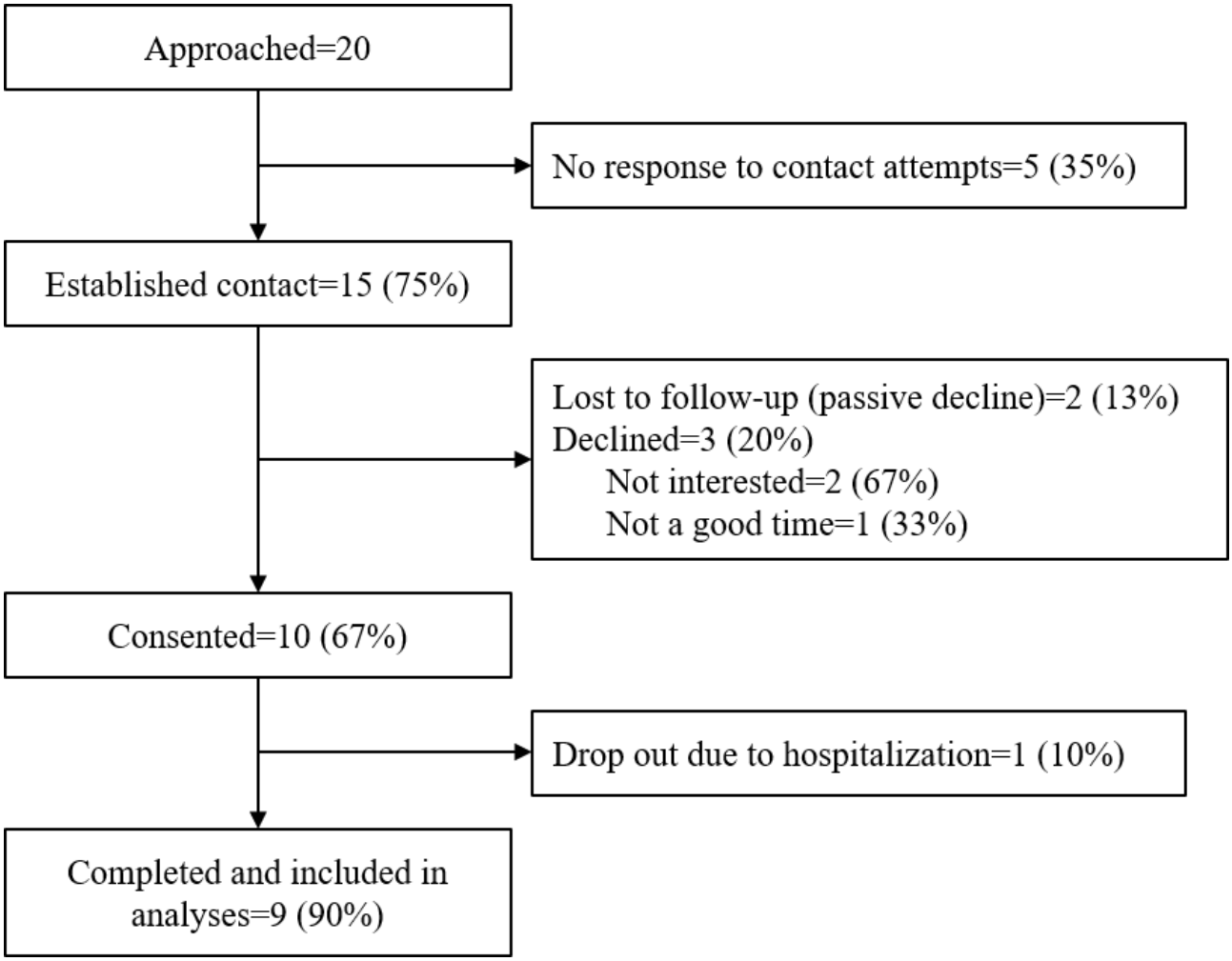

Table 1 provides a full description of the sample. All participants self-identified as cisgender heterosexual women. The average participant age was 55 (SD 14, range 35 - 85) years. The 10 participants identified as non-Hispanic White (n=5, 50%), non-Hispanic Black (n=4, 40%), or Latina (n=1, 10%). Most of the participants had a college degree or greater (n=9, 90%), were not working (n=5, 50%), had an annual household income greater than US $100,000 (n=5, 50%), and had private health insurance (n=9, 90%). The median time since MBC diagnosis was 3.7 years, but this ranged widely from a minimum of 1.5 years to a maximum of 18.1 years. Most participants (n=7, 70%) had hormone receptor–positive, human epidermal growth factor receptor 2–negative breast cancer.

XSL•FO
RenderX

Table . Participant demographic and clinical characteristics (N=10).

Characteristics	Values
Demographic characteristics	
Age (years), mean (SD)	55.2 (14.4)
Race and ethnicity, n (%)	
Non-Hispanic White	5 (50)
Non-Hispanic Black	4 (40)
Latina	1 (10)
Education, n (%)	
High school degree or equivalent	1 (10)
College degree	6 (60)
Postgraduate degree	3 (30)
Employment status, n (%)	
Working full time	1 (20)
Working part time	3 (30)
Retired	4 (40)
Unable to work	1 (10)
Annual household income, n (%)	
<US $100,000	4 (40)
≥US $100,000	5 (50)
Prefer not to answer	1 (10)
Health insurance type[a], n (%)	
Private	8 (80)
Public (eg, Medicare, Medicaid)	4 (40)
Military or veterans health care	1 (10)
Clinical characteristics	
Years since MBC[b] diagnosis, median (range)	3.7 (1.5 - 18.1)
Breast cancer subtype, n (%)	
HR[c] positive and HER2[d] positive	2 (20)
HR positive and HER2 negative	7 (70)
Triple negative	1 (10)

[a]As participants could select more than one, does not sum to 100%.

[b]MBC: metastatic breast cancer.

[c]HR: hormone receptor.

[d]HER2: Human Epidermal Growth Factor Receptor 2.

At baseline, participants demonstrated more depression (mean 55.4, SD 8.2) and anxiety (mean 57.1, SD 9.0) than the general population (mean 50.0, SD 10). These scores are indicative of mild depression and anxiety [56].

Feasibility and Acceptability

Participants attended 92% (46/50) of intervention sessions; specifically, 9 participants attended all 5 intervention sessions, and 1 participant attended only 1 intervention session before dropping out. Intervention sessions lasted 50 (SD 9, range 36-71) minutes on average.

For participants who were retained for the duration of the intervention, the average length of the intervention was 29 (SD 2, range 28 - 34) days. A total of 8 participants completed the 5 weekly intervention sessions in exactly 5 weeks, and 1 participant took an additional week after rescheduling due to competing demands. As symptom assessments were completed daily over the course of the intervention period, the possible number of daily symptom assessments ranged from 28 to 34, depending on the participant. Completion of daily symptom assessments ranged widely, from 46% (13/28) to 100% (28/28). On average, participants completed 85% (SD 18%) of daily

XSL•FO
RenderX

symptom assessments. The EMI was triggered for 93% (204/220) of daily symptom assessments. Participants received an average of 23 (SD 5, range 13 - 31) coaching messages, corresponding to 4 - 5 coaching messages per week in the intervention.

Participants reported high perceived intervention feasibility (mean 4.81, SD 0.44), acceptability (mean 4.78, SD 0.33), and appropriateness for patients with MBC (mean 4.83, SD 0.35). The average scores on all measures were above our a priori cutoff of ≥4, indicating adequate intervention feasibility, acceptability, and appropriateness. No harms or unintended intervention effects were identified.

Most of the 9 participants (n=6, 67%) felt that the number of POET sessions was "just right," while the remainder (n=3, 33%) felt that there were "too few" sessions. The length of the sessions was also rated by most as "just right" (n=7, 78%), with a few participants feeling the sessions were "too short" (n=1, 11%) or "too long" (n=1, 11%). In terms of intervention format, most participants (n=8, 89%) preferred individual meetings with the study counselor, either via telephone (n=5, 56%) or video call (n=4, 44%). Fewer participants (n=2, 22%) were interested in an in-person intervention. All participants (n=9, 100%) recommended the intervention for other patients with MBC.

Depression, Anxiety, Positive Affect, and Positive Emotion Regulation Skill Use

The average time between T0 and T1 was 43.6 (SD 8.7) days. Although there were no statistically significant changes in depression, anxiety, or positive affect from T0 to T1 (*P* values ≥.37; Table 2), we did observe small-to-moderate pre- to postintervention decreases in depression (d=−0.32) and anxiety (d=−0.27) and increases in positive affect (d=0.30). This corresponded to clinically significant reductions in depression and anxiety for 55% (5/9) and 33% (3/9) of participants, respectively [57]. We also observed a statistically significant increase in positive emotion regulation skill use (t_8=2.95; P=.02), corresponding to a large effect size (d=0.99).

Table . Pre- to postintervention changes in depression, anxiety, positive affect, and positive emotion regulation skill use (N=9).

Construct	T_0, mean (SD)	T_1, mean (SD)	*t* test (*df*)	*P* value	*d*	95% CI
PROMIS[a] depression	55.33 (8.73)	51.99 (7.86)	–0.95 (8)	.37	–0.32	–0.98 to 0.37
PROMIS anxiety	57.23 (9.53)	53.90 (8.48)	–0.81 (8)	.44	–0.27	–0.93 to 0.40
mDES[b] positive affect	2.04 (1.26)	2.31 (1.06)	0.91 (8)	.39	0.30	–0.38 to 0.97
Positive emotion regulation skill use	3.92 (0.86)	4.28 (0.67)	2.95 (8)	.02	0.99	0.16 to 1.77

[a]PROMIS: Patient-Reported Outcomes Measurement Information System.

[b]mDES: Modified Differential Emotions Scale.

RQA Results

Table 3 includes representative quotes from RQA, highlighting the feasibility, acceptability, and appropriateness of the intervention for people with MBC. All interviewees had positive experiences in the study, describing the POET program as useful, meaningful, or therapeutic. The weekly sessions helped many participants feel engaged or accountable; some participants expressed interest in adding additional sessions or having sessions every other week (to allow more time to practice POET skills). Interviewees also provided suggestions for improving the daily symptom reporting, including adding more emotions and open-ended questions to provide clarifications on their answers.

XSL•FO
RenderX

Table . Representative quotes from rapid qualitative analysis highlighting intervention feasibility, acceptability, and appropriateness.

Participant	Illustrative quote
No. 4, living with MBC[a] for 4 years	"People use [these skills] every day. They just don't know how to categorize them in a way you put on paper."
No. 5, living with MBC for 1.5 years	"And I loved the program, to be honest, I felt like it's a lot of stuff that a lot of people need... especially with cancer, you can become, 'Woe is me,' kind of quick. And so, it's a good way to keep it centered on the good in your life. And it's okay to have those bad days. But to have something to bring you back, I liked that a lot."
No. 7, living with MBC for 1.5 years	"If [a patient with MBC] were on the fence, that they should pursue the program, I would also say just go into it with an open mind... that the course content, that is actually beneficial."
No. 8, living with MBC for 3 years	"Mindfulness is one of those skills that helped me stop and realize that it's okay to put myself into the equation as well and focus on my own, which is something I often tend to ignore."
No. 9, living with MBC for 18 years	"I'm grateful for the opportunity that someone kind of feels like maybe we do need something different than the pink ribbon."
No. 10, living with MBC for 3 years	"[The coaching message] was cool. It cheered you up, like, 'Oh, you did this. Now, why don't you focus on this?' And it gives you a guide during the day on what to do instead of just being around your daily life and events and house chores and stuff like that. So that was good."

[a]MBC: metastatic breast cancer.

Most interviewees felt that all the skills covered in the POET program were helpful, applicable, or important. However, different participants preferred different skills; several individuals expressed interest in a tailored program, where they could "pick and choose" the skills that felt most important and relevant to them. Relatedly, many participants noted that they were already familiar with and practiced some of these skills. For some, the POET program was a helpful reminder of their skills, while others felt it was redundant to cover this material.

Interviewees generally liked the EMI component of the intervention and appreciated the structure that the coaching messages provided, reminding them to use the skills that they had learned. Some interviewees also reflected that it helped them to stay engaged in the POET program, rather than going a whole week between contacts. Interviewees felt that the frequency of the coaching messages was appropriate. However, 1 participant suggested that it would be nice to receive the coaching message daily, even if they did not complete the daily symptom assessment the day prior.

Finally, interviewees were asked about alternative delivery methods and formats for the POET program. While most participants were interested in the option of a self-paced web-based program, they still wanted to have weekly meetings with a facilitator. One participant suggested a "buddy program" between participants through the web-based platform. Regarding group sessions, interviewees often expressed interest with caution because of concerns about privacy and confidentiality. One participant noted that it would be important for group sessions to be structured and "different from a support group."

Discussion

Principal Findings

In this single-arm pilot study, we demonstrated that a multicomponent PPI incorporating EMI is feasible and acceptable among patients with MBC. Pilot study participants rated the POET intervention as highly feasible, acceptable, and appropriate for patients with MBC. High rates of recruitment and retention also demonstrate the feasibility of this trial. Preliminary data also suggest that the POET intervention may impact key outcomes of depression, anxiety, positive affect, and positive emotion regulation skill use. However, these findings are limited by the study design, including the small sample and lack of a control group, and must be replicated in a fully powered, randomized trial. Altogether, the results of this pilot study provide support for further examination of PPIs in the context of MBC.

Meta-analytic data demonstrate that PPIs have robust effects on the key MBC outcomes identified in our prior work: depression, anxiety, and positive affect [58]. However, in the oncology setting, PPIs have primarily been tested among patients with early-stage cancer [59-61]. A meta-analysis conducted by Casellas-Grau et al [59] identified 16 studies examining PPIs for patients with breast cancer, which they grouped into 5 categories: mindfulness-based approaches, expression of positive emotions, spiritual interventions, hope therapy, and meaning-making interventions. More than half of the included studies recruited only individuals with early-stage cancer; only 1 study focused on MBC specifically [62]. While additional trials have been published since this review was completed, to our knowledge, the initial pilot of POET is the only other study testing a PPI specifically for patients with MBC [24]. This study thus extends prior research in 2 key ways. First, we examine the feasibility and acceptability of a PPI among women with

XSL•FO
RenderX

MBC, who have been excluded from previous studies in this space. Second, our PPI was multicomponent and touched on several of the categories identified in the meta-analysis, including mindfulness, expression of positive emotions, hope, and meaning-making. By integrating multiple positive psychological skills, POET may be more appealing to potential participants and could have more robust effects on our outcomes of interest than single-component PPIs.

Results also highlight the potential added benefit of EMI on positive affect for patients with MBC. Despite specifically being designed to increase positive affect, an earlier version of the POET intervention did not demonstrate effects on positive affect among patients with MBC [24], indicating the need for adaptations. Although this small pilot study cannot provide evidence of intervention efficacy, we did observe pre- to postintervention increases in positive affect, possibly due to our consideration of Nudge Theory and the resulting inclusion of EMI. Nudges such as those included in this study can interrupt unwanted behaviors or experiences and prompt positive behavior change. For example, nudge-based interventions have successfully been used to disrupt sedentary time [63,64] and social media use [65]. In the context of PPIs, nudges represent an opportunity to interrupt negative emotions [66], prompt use of positive emotion skills [67], and reduce emotional spillover (when negative emotions experienced in one period continue into a later period) [68-70]. Thus, the responsive behavioral nudges that we incorporated in this intervention may have boosted the effects of POET on positive affect through these mechanisms.

Notably, Salsman et al [71] conducted a meta-analysis examining intervention effects on positive affect among cancer survivors and identified only 3 studies focusing on advanced cancer. Disappointingly, intervention effects on positive affect were stronger when trials focused on individuals with early-stage cancer (average effect size: g=0.42), compared to trials enrolling patients with advanced cancer (average effect size: g=−0.03). Salsman et al [71] noted that this may be due to differences in affective activation between individuals with early-stage and advanced disease. In other words, patients with advanced disease may tend to experience “low activation” positive affect (ie, peace, tranquility, and contentment), while individuals with early-stage disease may tend to experience “high activation” (ie, excitement). Our measure of positive affect (the mDES) included both low-activation and high-activation emotions. Although this small pilot study was not powered to examine changes in individual mDES items, future research should examine the impact of POET on low- versus high-activation emotions among women with MBC.

The results of this small pilot study also indicate the potential for the POET intervention to improve anxiety among patients with MBC. This is consistent with meta-analytic data demonstrating that PPIs ameliorate anxiety among medically ill patients [72]. However, specifically among people with cancer, 1 meta-analysis demonstrated high heterogeneity across studies in PPIs’ effects on anxiety, potentially resulting in less reliable effect size estimates [73]. The observed high heterogeneity in study findings could be due to differences in study populations, interventions, or outcome measurements. Thus, attention to these details through a rigorous, systematic program of research is needed to generate reliable estimates of the impact of PPIs on anxiety among women with MBC.

The generalizability of these findings is limited by the sociodemographic and clinical characteristics of the sample. While we were able to achieve diversity in terms of participant race and ethnicity, nearly all participants in this pilot study were highly educated and of high socioeconomic status. Prior literature has suggested that socioeconomic factors, such as education level, can moderate the effects of PPIs, particularly for behavioral outcomes (eg, antidepressant use) [74]. Future studies including larger, more socioeconomically diverse samples are needed to support the generalizability of these findings to the larger population of women with MBC. Recruiting outside of National Cancer Institute–designated comprehensive cancer centers may be one strategy to increase participant diversity, as patients presenting for care at these institutions are more likely to be non-Hispanic White, privately insured, and with high socioeconomic status [75]. Furthermore, all participants in this pilot study were able to communicate verbally in English. Prior research suggests that individuals with limited English proficiency have more barriers in accessing cancer care [76] and worse breast cancer outcomes than their English-speaking counterparts [77]. For example, Spanish-speaking Latina immigrant breast cancer survivors experience worse quality of life and emotional well-being than English-speaking survivors [78]. Expanding this PPI beyond English-speaking participants is a necessary next step to improve equity in psycho-oncology research. In terms of clinical characteristics, this study included only participants who reported elevated depression or anxiety in a prior study. While elevated depression and anxiety are common among patients with MBC [7-12], the present sample may not represent the broader MBC population, including those without elevated distress. It is possible that the POET intervention would be less feasible, acceptable, and impactful among patients with MBC and without elevated depression or anxiety.

Qualitative feedback from intervention participants can be used to guide future iterations of the POET program. Specifically, based on participant feedback, we will revise the daily symptom reporting procedure to include more emotions and open-ended questions. In addition, we are exploring the possibility of tailoring the POET program to individual participants based on their current affective experiences, prior experience with POET skills, or current use of the POET skills. Tailored or adaptive interventions that meet the specific needs and preferences of individual patients may be more acceptable, ultimately increasing participants’ engagement with and benefit from the intervention [31,79]. However, additional research would be needed to identify the decision points at which the POET intervention is tailored and test the effectiveness of the tailored intervention versus standard intervention delivery [80].

Strengths and Limitations

This pilot study has several strengths. Patients with metastatic cancer have unique psychosocial needs [13,14], yet are underrepresented in research [81,82]. Furthermore, this is among the first studies examining a PPI for patients with MBC; despite

XSL•FO
RenderX

the importance of positive experiences for this patient population, prior intervention trials do not focus on increasing positive affect [15,83,84]. Our incorporation of EMI is also a strength, as it may boost intervention results and increase engagement. We were able to recruit and retain a racially and ethnically diverse group of study participants, supporting the feasibility of doing so in a future, larger trial. In addition, participants had been living with MBC for varying lengths of time, providing unique perspectives on the disease and related psychosocial needs. We used robust and well-validated outcome measures with established cutoffs for clinically significant change, enabling us to explore the clinical impact of the POET intervention. Finally, the study was designed in collaboration with patient advocates, ensuring relevance to patients' lived experiences.

Nonetheless, the results must be interpreted in light of limitations. Most notably, this was a small, single-arm pilot study. While this design is appropriate for our primary goal of assessing intervention feasibility and acceptability [38], we are unable to determine whether improvements in our outcomes of interest are due to the intervention or other factors. A fully powered randomized clinical trial with long-term follow-up is needed to understand the effect of the POET intervention on depression, anxiety, positive affect, and positive emotion regulation skill use among patients with MBC. Moreover, from these pilot findings, we are unable to disentangle the relative effects of the POET sessions and the integrated EMI. A future dismantling study or factorial study would be needed to understand the individual and combined effects of different intervention components [85]. Our outcome measures were self-reported, which may introduce bias due to demand characteristics [86]. Furthermore, in an effort to minimize participant burden, we elected to use very brief measures of depression and anxiety; while they have strong psychometric properties, they may be less sensitive to change than longer versions [87]. Future studies might incorporate longer self-report measures or objective measures to further triangulate intervention effects. Participants in this pilot study were recruited from a prior trial, which included remote symptom reporting via ecological momentary assessment and required participants to own a phone capable of receiving SMS text messages and accessing the internet [21]. This may have resulted in a more "tech-savvy" and less representative sample, and limits generalizability. Only individuals with elevated depression or anxiety were selected, further narrowing the sample and limiting applicability to the broader MBC population. Furthermore, we only contacted individuals who had previously agreed to be contacted for future studies, which may have introduced self-selection bias (ie, individuals who agreed to be contacted for future studies may be more willing to participate). Recruitment rates might be lower among individuals who had not previously expressed interest in participating in research. Most participants in this study were highly educated and of high socioeconomic status, and all were able to communicate verbally in English and were receiving care at a National Cancer Institute–designated comprehensive cancer center. Thus, results may not generalize to the broader, socioeconomically and linguistically diverse population of people with MBC and those receiving care in community oncology settings [88,89]. Finally, our sample was relatively homogenous in terms of MBC subtype, with 70% having hormone receptor–positive Human Epidermal Growth Factor Receptor 2–negative disease. This has significant implications for participants' lived experience of MBC, including treatment regimen (ie, hormone therapy rather than chemotherapy), time toxicity, and quality of life. Relatedly, none of our participants were within a year of their MBC diagnosis and may therefore be in a different stage of adjustment to their disease. The effect of PPIs may vary for individuals closer to diagnosis and undergoing different types of treatments for MBC. Future studies with more clinically diverse samples are needed to understand if the intervention is more or less impactful for certain cohorts.

Conclusions

This pilot study demonstrates the promise of the POET program for patients with MBC. Specifically, the results support the feasibility of enrolling and retaining racially and ethnically diverse patients with MBC to this trial, the acceptability of the PPI enhanced with EMI, and potential intervention impacts on depression, anxiety, positive affect, and positive emotion regulation skill use. A large-scale randomized controlled trial is needed to assess intervention efficacy for outcomes of interest.

Acknowledgments

This work was supported by the National Cancer Institute (K08 CA270402, principal investigator [PI]: CCC; T32CA261787, PI: SCO, Adams-Campbell, Schwartz; P30 CA051008, PI: Weiner). The content presented here is solely the responsibility of the authors and does not represent the official views of the National Cancer Institute.

Data Availability

Deidentified quantitative data are publicly available on the Open Science Framework [90]. Sociodemographic data are omitted for privacy reasons and can be requested from the corresponding author (CCC). The qualitative datasets are not publicly available, as qualitative interview data cannot be completely anonymized, even when aliases are used to replace names. Qualitative data are available from the corresponding author (CCC) on reasonable request.

XSL•FO
RenderX

Authors' Contributions

All authors contributed to the study conception and design. Material preparation, data collection, and data analysis were performed by CCC and MVS. The first draft of the manuscript was written by CCC. All authors commented on subsequent versions of the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

CCC has received research support (to institution) from Pfizer. LG has participated in consulting activities for which she has received honoraria from Pfizer and Stemline/Menarini. RG has participated in consulting activities for which she has received honoraria from AstraZeneca. CI has participated in consulting activities for which she has received honoraria from AstraZeneca, Genentech, Seattle Genetics, Pfizer, Gilead, and Merck; received royalties from Wolters Kluwer (UptoDate) and McGraw Hill (Goodman and Gillman); and received research support (to institution) from Tesaro/GSK, Seattle Genetics, Pfizer, AztraZeneca, Bristol Myers Squibb, Genentech, and Novartis. SCO has received research support (to institution) from Pfizer and Gilead Sciences. No other authors have competing interests to disclose.

Multimedia Appendix 1
Exit interview guide.
[PDF File, 135 KB - formative_v9i1e77636_app1.pdf]

Multimedia Appendix 2
Rapid qualitative analysis transcript summary template.
[PDF File, 117 KB - formative_v9i1e77636_app2.pdf]

Checklist 1
CONSORT checklist.
[PDF File, 214 KB - formative_v9i1e77636_app3.pdf]

Checklist 2
COREQ checklist.
[PDF File, 137 KB - formative_v9i1e77636_app4.pdf]

References

1. Liang Y, Zhang H, Song X, Yang Q. Metastatic heterogeneity of breast cancer: molecular mechanism and potential therapeutic targets. Semin Cancer Biol 2020 Feb;60:14-27. [doi: 10.1016/j.semcancer.2019.08.012] [Medline: 31421262]
2. Caswell-Jin JL, Plevritis SK, Tian L, et al. Change in survival in metastatic breast cancer with treatment advances: meta-analysis and systematic review. JNCI Cancer Spectr 2018 Nov;2(4):pky062. [doi: 10.1093/jncics/pky062] [Medline: 30627694]
3. Santa-Maria CA, Gradishar WJ. Changing treatment paradigms in metastatic breast cancer: lessons learned. JAMA Oncol 2015 Jul;1(4):528-534. [doi: 10.1001/jamaoncol.2015.1198] [Medline: 26181262]
4. Sundquist M, Brudin L, Tejler G. Improved survival in metastatic breast cancer 1985-2016. Breast 2017 Feb;31:46-50. [doi: 10.1016/j.breast.2016.10.005] [Medline: 27810699]
5. Gogate A, Wheeler SB, Reeder-Hayes KE, et al. Projecting the prevalence and costs of metastatic breast cancer from 2015 through 2030. JNCI Cancer Spectr 2021 Aug;5(4):pkab063. [doi: 10.1093/jncics/pkab063] [Medline: 34409255]
6. Mariotto AB, Etzioni R, Hurlbert M, Penberthy L, Mayer M. Estimation of the number of women living with metastatic breast cancer in the United States. Cancer Epidemiol Biomarkers Prev 2017 Jun;26(6):809-815. [doi: 10.1158/1055-9965.EPI-16-0889] [Medline: 28522448]
7. Guo YQ, Ju QM, You M, Liu Y, Yusuf A, Soon LK. Depression, anxiety and stress among metastatic breast cancer patients on chemotherapy in China. BMC Nurs 2023 Feb 6;22(1):33. [doi: 10.1186/s12912-023-01184-1] [Medline: 36747213]
8. Park EM, Gelber S, Rosenberg SM, et al. Anxiety and depression in young women with metastatic breast cancer: a cross-sectional study. Psychosomatics 2018;59(3):251-258. [doi: 10.1016/j.psym.2018.01.007] [Medline: 29525523]
9. Islam N, Biswas J, Kowshik MM, et al. Depression, anxiety, and performance status among the women with metastatic breast cancer receiving palliative care in Bangladesh: a cross sectional study. Health Sci Rep 2022 Nov;5(6):e911. [doi: 10.1002/hsr2.911] [Medline: 36320658]
10. Low CA, Stanton AL. Activity disruption and depressive symptoms in women living with metastatic breast cancer. Health Psychol 2015 Jan;34(1):89-92. [doi: 10.1037/hea0000052] [Medline: 24447191]
11. Senkpeil RR, Olson JS, Fortune EE, Zaleta AK. Pain and nausea intensity, social function, and psychological well-being among women with metastatic breast cancer. J Patient Exp 2022;9:23743735221134733. [doi: 10.1177/23743735221134733] [Medline: 36339375]

XSL•FO
RenderX

12. Karibayeva I, Turdaliyeva B, Zainal NZ, Bagiyarova F. The prevalence of depression symptoms among advanced breast cancer patients: a systematic review and meta-analysis. Eur J Cancer 2020 Oct;138:S101. [doi: 10.1016/S0959-8049(20)30806-6]
13. Cardoso F, Spence D, Mertz S, et al. Global analysis of advanced/metastatic breast cancer: decade report (2005-2015). Breast 2018 Jun;39:131-138. [doi: 10.1016/j.breast.2018.03.002] [Medline: 29679849]
14. Oswald LB, Arredondo B, Kadono M, et al. A mixed-methods study of cyclin-dependent kinase 4 and 6 inhibitor symptom burden and quality of life among metastatic breast cancer patients and providers. Cancer Med 2021 Jul;10(14):4823-4831. [doi: 10.1002/cam4.4055] [Medline: 34165265]
15. Keane D, Phillips G, Mitchell N, Connolly RM, Hegarty J. Improving quality of life and symptom experience in patients with metastatic breast cancer: a systematic review of supportive care interventions. Psychooncology 2023 Aug;32(8):1192-1207. [doi: 10.1002/pon.6183] [Medline: 37434307]
16. Ye ZJ, Qiu HZ, Liang MZ, et al. Effect of a mentor-based, supportive-expressive program, Be Resilient to Breast Cancer, on survival in metastatic breast cancer: a randomised, controlled intervention trial. Br J Cancer 2017 Nov 7;117(10):1486-1494. [doi: 10.1038/bjc.2017.325] [Medline: 28926525]
17. Sheean P, Matthews L, Visotcky A, et al. Every day counts: a randomized pilot lifestyle intervention for women with metastatic breast cancer. Breast Cancer Res Treat 2021 Jun;187(3):729-741. [doi: 10.1007/s10549-021-06163-1] [Medline: 33742324]
18. Folkman S, Moskowitz JT. Positive affect and the other side of coping. Am Psychol 2000 Jun;55(6):647-654. [doi: 10.1037//0003-066x.55.6.647] [Medline: 10892207]
19. Diener E, Emmons RA. The independence of positive and negative affect. J Pers Soc Psychol 1984 Nov;47(5):1105-1117. [doi: 10.1037//0022-3514.47.5.1105] [Medline: 6520704]
20. Schmukle SC, Egloff B, Burns LR. The relationship between positive and negative affect in the positive and negative affect schedule. J Res Pers 2002 Oct;36(5):463-475. [doi: 10.1016/S0092-6566(02)00007-7]
21. Mossman B, Small BJ, Velazquez-Sosa MA, et al. Ecological momentary assessment of quality of life in women with metastatic breast cancer. Presented at: American Preventive Oncology Society Annual Meeting; Apr 6-8, 2025; Philadelphia, PA URL: https://aspo.org/wp-content/uploads/030-HP_ASPO-Abstracts-2025_8.5x11_FINAL.pdf [accessed 2025-10-02]
22. Moskowitz JT, Cheung EO, Freedman M, et al. Measuring positive emotion outcomes in positive psychology interventions: a literature review. Emot Rev 2021 Jan;13(1):60-73. [doi: 10.1177/1754073920950811]
23. Duan W, Klibert J, Schotanus-Dijkstra M, et al. Editorial: Positive psychological interventions: how, when and why they work: beyond WEIRD contexts. Front Psychol 2022;13:1021539. [doi: 10.3389/fpsyg.2022.1021539] [Medline: 36262443]
24. Cheung EO, Cohn MA, Dunn LB, et al. A randomized pilot trial of a positive affect skill intervention (lessons in linking affect and coping) for women with metastatic breast cancer. Psychooncology 2017 Dec;26(12):2101-2108. [doi: 10.1002/pon.4312] [Medline: 27862646]
25. Thaler RH, Sunstein CR, Balz JP. Choice architecture. In: Shafir E, editor. The Behavioral Foundations of Public Policy: Princeton University Press; 2014:428-439. [doi: 10.2307/j.ctv550cbm.31]
26. Payne JW, Bettman JR, Johnson EJ. Behavioral decision research: a constructive processing perspective. Annu Rev Psychol 1992 Jan;43(1):87-131. [doi: 10.1146/annurev.ps.43.020192.000511]
27. Gigerenzer G, Gaissmaier W. Heuristic decision making. Annu Rev Psychol 2011;62(1):451-482. [doi: 10.1146/annurev-psych-120709-145346] [Medline: 21126183]
28. Lichtenstein S, Slovic P. The Construction of Preference: Cambridge University Press; 2006.
29. Szaszi B, Palinkas A, Palfi B, Szollosi A, Aczel B. A systematic scoping review of the choice architecture movement: toward understanding when and why nudges work. Behav Decis Mak 2018 Jul;31(3):355-366. [doi: 10.1002/bdm.2035]
30. Tariman JD, Berry DL, Halpenny B, Wolpin S, Schepp K. Validation and testing of the acceptability E-scale for web-based patient-reported outcomes in cancer care. Appl Nurs Res 2011 Feb;24(1):53-58. [doi: 10.1016/j.apnr.2009.04.003] [Medline: 20974066]
31. Nahum-Shani I, Smith SN, Spring BJ, et al. Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. Ann Behav Med 2018 May 18;52(6):446-462. [doi: 10.1007/s12160-016-9830-8] [Medline: 27663578]
32. Shim Y, Scotney VS, Tay L. Conducting mobile-enabled ecological momentary intervention research in positive psychology: key considerations and recommended practices. J Posit Psychol 2022 Sep 3;17(5):708-717. [doi: 10.1080/17439760.2021.1913642]
33. LaFreniere LS, Newman MG. Upregulating positive emotion in generalized anxiety disorder: a randomized controlled trial of the SkillJoy ecological momentary intervention. J Consult Clin Psychol 2023 Jun;91(6):381-387. [doi: 10.1037/ccp0000794] [Medline: 36716146]
34. Sizemore KM, Gray S, Wolfer C, et al. A proof of concept pilot examining feasibility and acceptability of the positively healthy just-in-time adaptive, ecological momentary, intervention among a sample of sexual minority men living with HIV. J Happiness Stud 2022 Dec;23(8):4091-4118. [doi: 10.1007/s10902-022-00587-2]
35. Heron KE, Smyth JM. Ecological momentary interventions: incorporating mobile technology into psychosocial and health behaviour treatments. Br J Health Psychol 2010 Feb;15(Pt 1):1-39. [doi: 10.1348/135910709X466063] [Medline: 19646331]

XSL•FO
RenderX

36. Dao KP, De Cocker K, Tong HL, Kocaballi AB, Chow C, Laranjo L. Smartphone-delivered ecological momentary interventions based on ecological momentary assessments to promote health behaviors: systematic review and adapted checklist for reporting ecological momentary assessment and intervention studies. JMIR Mhealth Uhealth 2021 Nov 19;9(11):e22890. [doi: 10.2196/22890] [Medline: 34806995]
37. Thong MSY, Chan RJ, van den Hurk C, et al. Going beyond (electronic) patient-reported outcomes: harnessing the benefits of smart technology and ecological momentary assessment in cancer survivorship research. Support Care Cancer 2021 Jan;29(1):7-10. [doi: 10.1007/s00520-020-05648-x] [Medline: 32844316]
38. Czajkowski SM, Powell LH, Adler N, et al. From ideas to efficacy: the ORBIT model for developing behavioral treatments for chronic diseases. Health Psychol 2015 Oct;34(10):971-982. [doi: 10.1037/hea0000161] [Medline: 25642841]
39. Eldridge SM, Chan CL, Campbell MJ, et al. CONSORT 2010 statement: extension to randomised pilot and feasibility trials. BMJ 2016 Oct 24;355:i5239. [doi: 10.1136/bmj.i5239] [Medline: 27777223]
40. Tong A, Sainsbury P, Craig J. Consolidated Criteria for Reporting Qualitative Research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
41. White N, Kupeli N, Vickerstaff V, Stone P. How accurate is the "Surprise Question" at identifying patients at the end of life? A systematic review and meta-analysis. BMC Med 2017 Aug 2;15(1):139. [doi: 10.1186/s12916-017-0907-4] [Medline: 28764757]
42. Jones TA, Olds TS, Currow DC, Williams MT. Feasibility and pilot studies in palliative care research: a systematic review. J Pain Symptom Manage 2017 Jul;54(1):139-151. [doi: 10.1016/j.jpainsymman.2017.02.015] [Medline: 28450220]
43. Cutillo A, O'Hea E, Person S, Lessard D, Harralson T, Boudreaux E. The distress thermometer: cutoff points and clinical use. Oncol Nurs Forum 2017 May 1;44(3):329-336. [doi: 10.1188/17.ONF.329-336] [Medline: 29493167]
44. Darwin Holmes AG. Researcher positionality—a consideration of its influence and place in qualitative research—a new researcher guide. Education 2020;8(4):1-10. [doi: 10.34293/education.v8i4.3232]
45. Berger R. Now I see it, now I don't: researcher's position and reflexivity in qualitative research. Qual Res 2015 Apr;15(2):219-234. [doi: 10.1177/1468794112468475]
46. Moskowitz JT, Hult JR, Duncan LG, et al. A positive affect intervention for people experiencing health-related stress: development and non-randomized pilot test. J Health Psychol 2012 Jul;17(5):676-692. [doi: 10.1177/1359105311425275] [Medline: 22021272]
47. Weiner BJ, Lewis CC, Stanick C, et al. Psychometric assessment of three newly developed implementation outcome measures. Implementation Sci 2017 Dec;12(1):1-12. [doi: 10.1186/s13012-017-0635-3]
48. Schalet BD, Pilkonis PA, Yu L, et al. Clinical validity of PROMIS depression, anxiety, and anger across diverse clinical samples. J Clin Epidemiol 2016 May;73:119-127. [doi: 10.1016/j.jclinepi.2015.08.036] [Medline: 26931289]
49. Teresi JA, Ocepek-Welikson K, Kleinman M, Ramirez M, Kim G. Psychometric properties and performance of the Patient Reported Outcomes Measurement Information System® (PROMIS®) depression short forms in ethnically diverse groups. Psychol Test Assess Model 2016;58(1):141-181. [Medline: 28553573]
50. Teresi JA, Ocepek-Welikson K, Kleinman M, Ramirez M, Kim G. Measurement equivalence of the Patient Reported Outcomes Measurement Information System® (PROMIS®) anxiety short forms in ethnically diverse groups. Psychol Test Assess Model 2016;58(1):183-219. [Medline: 28649483]
51. Fredrickson BL, Tugade MM, Waugh CE, Larkin GR. What good are positive emotions in crises? A prospective study of resilience and emotions following the terrorist attacks on the United States on September 11th, 2001. J Pers Soc Psychol 2003 Feb;84(2):365-376. [doi: 10.1037//0022-3514.84.2.365] [Medline: 12585810]
52. Giorgi A. The Descriptive Phenomenological Method in Psychology: A Modified Husserlian Approach: Duquesne University Press; 2009.
53. Vindrola-Padros C, Johnson GA. Rapid techniques in qualitative research: a critical review of the literature. Qual Health Res 2020 Aug;30(10):1596-1604. [doi: 10.1177/1049732320921835] [Medline: 32667277]
54. Averill JB. Matrix analysis as a complementary analytic strategy in qualitative inquiry. Qual Health Res 2002 Jul;12(6):855-866. [doi: 10.1177/104973230201200611] [Medline: 12109729]
55. Hennink MM, Kaiser BN, Marconi VC. Code saturation versus meaning saturation: how many interviews are enough? Qual Health Res 2017 Mar;27(4):591-608. [doi: 10.1177/1049732316665344] [Medline: 27670770]
56. Cella D, Choi S, Garcia S, et al. Setting standards for severity of common symptoms in oncology using the PROMIS item banks and expert judgment. Qual Life Res 2014 Dec;23(10):2651-2661. [doi: 10.1007/s11136-014-0732-6] [Medline: 24938431]
57. Terwee CB, Peipert JD, Chapman R, et al. Minimal important change (MIC): a conceptual clarification and systematic review of MIC estimates of PROMIS measures. Qual Life Res 2021 Oct;30(10):2729-2754. [doi: 10.1007/s11136-021-02925-y] [Medline: 34247326]
58. Hendriks T, Schotanus-Dijkstra M, Hassankhan A, de Jong J, Bohlmeijer E. The efficacy of multi-component positive psychology interventions: a systematic review and meta-analysis of randomized controlled trials. J Happiness Stud 2020 Jan;21(1):357-390. [doi: 10.1007/s10902-019-00082-1]

XSL•FO
RenderX

59. Casellas-Grau A, Font A, Vives J. Positive psychology interventions in breast cancer. A systematic review. Psychooncology 2014 Jan;23(1):9-19. [doi: 10.1002/pon.3353] [Medline: 23897834]
60. Rouleau CR, Garland SN, Carlson LE. The impact of mindfulness-based interventions on symptom burden, positive psychological outcomes, and biomarkers in cancer patients. Cancer Manag Res 2015;7:121-131. [doi: 10.2147/CMAR.S64165] [Medline: 26064068]
61. Cillessen L, Johannsen M, Speckens AEM, Zachariae R. Mindfulness-based interventions for psychological and physical health outcomes in cancer patients and survivors: a systematic review and meta-analysis of randomized controlled trials. Psychooncology 2019 Dec;28(12):2257-2269. [doi: 10.1002/pon.5214] [Medline: 31464026]
62. Vilhauer RP. Perceived benefits of online support groups for women with metastatic breast cancer. Women Health 2009;49(5):381-404. [doi: 10.1080/03630240903238719] [Medline: 19851944]
63. Thomas JG, Bond DS. Behavioral response to a just-in-time adaptive intervention (JITAI) to reduce sedentary behavior in obese adults: Implications for JITAI optimization. Health Psychol 2015 Dec;34S:1261-1267. [doi: 10.1037/hea0000304] [Medline: 26651467]
64. Nazaret A, Sapiro G. A large-scale observational study of the causal effects of a behavioral health nudge. Sci Adv 2023 Sep 22;9(38):eadi1752. [doi: 10.1126/sciadv.adi1752] [Medline: 37738345]
65. Keller J, et al. A digital nudge-based intervention to interrupt Instagram usage: randomized controlled pilot study. Eur J Health Psychol 2024;31(3):128-140. [doi: 10.1027/2512-8442/a000150]
66. Bagchi D, Mukherjee M, Nag I. Exploring “Nudge” in mental health research: a systematic review. In: Rodriguez RV, K H, Majumdar A, Basu R, editors. 1st International Conference, ‘Resonance’: On Cognitive Approach, Social Ethics and Sustainability: 23 and 24th November, 2022 School Of Liberal Arts and Humanities, Woxsen University, India: Routledge; 2022.
67. Deges J, Hughes R, Tee D. Positive nudging: thematic analysis of the experience of participating in a signature strength behavioural nudging intervention. Int J Appl Posit Psychol 2024 Aug;9(2):1-22. [doi: 10.1007/s41042-023-00140-3]
68. Farmer AS, Kashdan TB. Social anxiety and emotion regulation in daily life: spillover effects on positive and negative social events. Cogn Behav Ther 2012;41(2):152-162. [doi: 10.1080/16506073.2012.666561] [Medline: 22428662]
69. Cohen LH, Gunthert KC, Butler AC, Parrish BP, Wenze SJ, Beck JS. Negative affective spillover from daily events predicts early response to cognitive therapy for depression. J Consult Clin Psychol 2008 Dec;76(6):955-965. [doi: 10.1037/a0014131] [Medline: 19045964]
70. Nezlek JB, Kuppens P. Regulating positive and negative emotions in daily life. J Pers 2008 Jun;76(3):561-580. [doi: 10.1111/j.1467-6494.2008.00496.x] [Medline: 18399953]
71. Salsman JM, Pustejovsky JE, Schueller SM, et al. Psychosocial interventions for cancer survivors: a meta-analysis of effects on positive affect. J Cancer Surviv 2019 Dec;13(6):943-955. [doi: 10.1007/s11764-019-00811-8] [Medline: 31741250]
72. Brown L, Ospina JP, Celano CM, Huffman JC. The effects of positive psychological interventions on medical patients’ anxiety: a meta-analysis. Psychosom Med 2019 Sep;81(7):595-602. [doi: 10.1097/PSY.0000000000000722] [Medline: 31232915]
73. Tian X, Zhou X, Sun M, et al. The effectiveness of positive psychological interventions for patients with cancer: a systematic review and meta-analysis. J Clin Nurs 2024 Aug;33(9):3752-3774. [doi: 10.1111/jocn.17358] [Medline: 38979929]
74. Addington EL, Cheung EO, Moskowitz JT. Who is most likely to benefit from a positive psychological intervention? Moderator analyses from a randomized trial in people newly diagnosed with HIV. J Posit Psychol 2020;15(5):605-612. [doi: 10.1080/17439760.2020.1789702] [Medline: 32905459]
75. Wolfson JA, Sun CL, Wyatt LP, Hurria A, Bhatia S. Impact of care at comprehensive cancer centers on outcome: results from a population-based study. Cancer 2015 Nov 1;121(21):3885-3893. [doi: 10.1002/cncr.29576] [Medline: 26218755]
76. Chen DW, Banerjee M, He X, et al. Hidden disparities: how language influences patients’ access to cancer care. J Natl Compr Canc Netw 2023 Sep;21(9):951-959. [doi: 10.6004/jnccn.2023.7037] [Medline: 37673110]
77. Spiegel D, Willcox J, Modest AM, Bhargava N, Recht A. Impact of limited English proficiency on breast cancer outcomes. Int J Radiat Oncol Biol Phys 2024 Oct;120(2):e67. [doi: 10.1016/j.ijrobp.2024.07.1926]
78. Janz NK, Mujahid MS, Hawley ST, et al. Racial/ethnic differences in quality of life after diagnosis of breast cancer. J Cancer Surviv 2009 Dec;3(4):212-222. [doi: 10.1007/s11764-009-0097-y] [Medline: 19760151]
79. Noar SM, Benac CN, Harris MS. Does tailoring matter? Meta-analytic review of tailored print health behavior change interventions. Psychol Bull 2007 Jul;133(4):673-693. [doi: 10.1037/0033-2909.133.4.673] [Medline: 17592961]
80. Klasnja P, Hekler EB, Shiffman S, et al. Microrandomized trials: an experimental design for developing just-in-time adaptive interventions. Health Psychol 2015 Dec;34S:1220-1228. [doi: 10.1037/hea0000305] [Medline: 26651463]
81. Flowers M, Birkey Reffey S, Mertz SA, Marc Hurlbert for the Metastatic Breast Cancer Alliance. Obstacles, opportunities and priorities for advancing metastatic breast cancer research. Cancer Res 2017 Jul 1;77(13):3386-3390. [doi: 10.1158/0008-5472.CAN-17-0232] [Medline: 28601794]
82. Mollica MA, Tesauro G, Tonorezos ES, Jacobsen PB, Smith AW, Gallicchio L. Current state of funded National Institutes of Health grants focused on individuals living with advanced and metastatic cancers: a portfolio analysis. J Cancer Surviv 2021 Jun;15(3):370-374. [doi: 10.1007/s11764-021-01008-8] [Medline: 33651327]

XSL•FO
RenderX

83. Beatty L, Kemp E, Butow P, et al. A systematic review of psychotherapeutic interventions for women with metastatic breast cancer: context matters. Psychooncology 2018 Jan;27(1):34-42. [doi: 10.1002/pon.4445] [Medline: 28432855]
84. Teo I, Krishnan A, Lee GL. Psychosocial interventions for advanced cancer patients: a systematic review. Psychooncology 2019 Jul;28(7):1394-1407. [doi: 10.1002/pon.5103] [Medline: 31077475]
85. Chakraborty B, Collins LM, Strecher VJ, Murphy SA. Developing multicomponent interventions using fractional factorial designs. Stat Med 2009 Sep 20;28(21):2687-2708. [doi: 10.1002/sim.3643] [Medline: 19575485]
86. Wenzel M, Winkler M, Lasi J, Rowland Z. Examining demand effects on direct and indirect affect measures in affect induction procedures. Emotion 2024 Oct;24(7):1753-1770. [doi: 10.1037/emo0001368] [Medline: 38900554]
87. Cella D, Choi SW, Condon DM, et al. PROMIS® adult health profiles: efficient short-form measures of seven health domains. Value Health 2019 May;22(5):537-544. [doi: 10.1016/j.jval.2019.02.004] [Medline: 31104731]
88. Gage-Bouchard EA, Rodriguez EM, Saad-Harfouche FG, Miller A, Erwin DO. Factors influencing patient pathways for receipt of cancer care at an NCI-designated comprehensive cancer center. PLoS ONE 2014;9(10):e110649. [doi: 10.1371/journal.pone.0110649] [Medline: 25329653]
89. Dickson-Witmer D, Bleznak AD, Kennedy JS, et al. Breast cancer care in the community: challenges, opportunities, and outcomes. Surg Oncol Clin N Am 2011 Jul;20(3):555-580. [doi: 10.1016/j.soc.2011.01.007] [Medline: 21640921]
90. Conley C. Lessons in linking affect and coping (LILAC2). Open Science Framework. 2025 Mar 5. URL: https://osf.io/ec7zh/ [accessed 2025-10-02]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
COREQ: Consolidated Criteria for Reporting Qualitative Research
EMI: ecological momentary intervention
MBC: metastatic breast cancer
mDES: Modified Differential Emotions Scale
POET: Positive Emotion Skills Training
PPI: positive psychological intervention
PROMIS: Patient-Reported Outcomes Measurement Information System
RQA: rapid qualitative analysis

Edited by J Sarvestan; submitted 16.05.25; peer-reviewed by M Pritchard, P Hebbard; revised version received 27.08.25; accepted 09.09.25; published 07.10.25.

Please cite as:
Conley CC, Addington EL, Velazquez-Sosa M, Mossman B, Glenn L, Drakeford S, Guerra R, Isaacs C, Chitalia A, Gallagher C, O'Neill SC, Moskowitz JT
Feasibility and Acceptability of a Positive Psychological Intervention for Patients With Metastatic Breast Cancer: Pre-Post Pilot Study
JMIR Form Res 2025;9:e77636
URL: https://formative.jmir.org/2025/1/e77636
doi:10.2196/77636

© Claire C Conley, Elizabeth L Addington, Mikaela Velazquez-Sosa, Brenna Mossman, Lesley Glenn, Shontè Drakeford, Roxana Guerra, Claudine Isaacs, Ami Chitalia, Christopher Gallagher, Suzanne C O'Neill, Judith T Moskowitz. Originally published in JMIR Formative Research (https://formative.jmir.org), 7.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Designing Digital Mental Health Tools to Support the Needs of Black Adults in the United States: Qualitative Analysis

Sarah Alexandra Popowski[1], BA; Jonah Meyerhoff[1], PhD; Olivia Marin Allen[1], BS; Theresa Nguyen[2], MSW; Terika McCall[3], PhD, MPH, MBA; Aderonke Bamgbose Pederson[4], MD; Madhu Reddy[5], PhD; David Mohr[1], PhD; Rachel Kornfield[1], PhD

[1]Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University, 750 N. Lake Shore Dr., 10th Floor, Chicago, IL, United States

[2]Mental Health America, Alexandria, United States

[3]Department of Biostatistics, Division of Health Informatics, Yale School of Public Health, Yale University, New Haven, United States

[4]Department of Psychiatry, Massachusetts General Hospital, Harvard Medical School, Boston, MA, United States

[5]Department of Informatics, University of California, Irvine, Irvine, United States

Corresponding Author:
Sarah Alexandra Popowski, BA
Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University, 750 N. Lake Shore Dr., 10th Floor, Chicago, IL, United States

Abstract

Background: Depression and anxiety are associated with excess morbidity and mortality, constituting a major health care challenge. The prevalence of these conditions is increasing. In the United States, the health-related burden of depression and anxiety may disproportionately affect Black adults, who face unique stressors impacting their mental health and barriers to accessing treatment, including but not limited to systemic racism, discrimination, underdiagnosis of common mental health concerns (ie, depression, anxiety), limited access to culturally sensitive care, and mental health stigma within and outside Black communities.

Objective: This study aimed to explore the mental health experiences of nontreatment-seeking Black adults, and how these experiences relate to their needs and preferences for the design of digital mental health (DMH) tools through user-centered design methods.

Methods: This study included 25 nontreatment-seeking Black adults (aged 18‑61 years) with experiences of depression or anxiety to share their perspectives on how DMH tools can meet their needs. Participants were recruited either through social media advertisements or depression and anxiety questionnaires. All participants engaged in an asynchronous online discussion group in which they discussed their past and current mental health experiences, distinct challenges faced by Black Americans, and perceptions of DMH tools, as well as how such tools can be tailored to meet their mental health needs. Participants also completed a technology probe in which they used an automated mental health self-management text messaging tool (Small Steps SMS; Audacious Software) for 18 days. They shared their perceptions of the tool and ideas for specific design improvements in the discussion group. A subset (n=6) completed follow-up interviews to elaborate on their online discussion group posts.

Results: All participants reported significant mental health concerns and difficulty managing related symptoms. A majority of participants (22/25, 88%) expressed that racism and mental health stigma severely impacted their mental health and limited opportunities to discuss their experiences within and outside Black communities. They were interested in the use of DMH tools for mental health self-management and nearly all participants (23/25, 92%) endorsed text messaging as a convenient way to introduce techniques for coping with symptoms of depression and anxiety; however, some participants strongly advocated for additional design features that they believed would improve the program, including the integration of content that centers the experiences of Black individuals, creating nonjudgmental spaces for discussing mental health experiences, and linking formal mental health treatment resources for those who want them.

Conclusions: These findings suggest that our participants hold generally favorable views toward DMH tools, which can provide psychoeducation, self-management support tailored to the needs of Black adults, and a safe environment to address mental health concerns. Furthermore, it is critical to consider the role of racial discrimination and mental health stigma when designing inclusive and culturally sensitive DMH tools.

(JMIR Form Res 2025;9:e73279) doi:10.2196/73279

XSL•FO
RenderX

KEYWORDS

technology; mHealth; SMS text messaging; mental health services; depression; anxiety; African Americans; health equity; race; mobile health

Introduction

The United States is facing a burgeoning mental health crisis, in which depression and anxiety are at the forefront. Depression and anxiety are pervasive mental health conditions that affect 8.3% and 19.1% of adults each year, respectively [1,2]. They are often chronic and debilitating, with adverse downstream effects on education, employment, and quality of life [1,2]. The clinical presentation of depression and anxiety varies from person to person and between groups [3,4]. Symptoms of these conditions are common in Black adults and occur at a similar rate to that observed in White adults [5]; however, Black adults' experiences of depression and anxiety are more likely to be chronic and severe [3,6]. Notably, in the United States, Black individuals' mental health challenges are compounded by experiences of structural racism, inequity, discrimination, and bias in society at large, as well as in health systems [7,8]. Black adults report higher numbers of chronic and acute stressors compared to White adults, including adverse childhood experiences, racial microaggressions, and both interpersonal and institutional maltreatment [9-11]. Furthermore, research has found that stressful life events are associated with poorer mental health outcomes and an elevated risk of depression, anxiety, and suicide, as well as higher overall severity and chronicity of mental health problems [11,12]. Previous research also suggests that the effects of institutional racism across multiple domains of social determinants of health (ie, economic stability, educational access and quality, health care access and quality, neighborhood and built environment, and social and community context) increase the likelihood of psychological distress and negative mental health outcomes [13].

Symptoms of depression and anxiety are less likely to be treated in Black people in the United States [14], who may be deterred from seeking treatment by costs and other structural barriers like the limited accessibility of services in their neighborhoods, limited mental health literacy, and attitudinal barriers such as stigma and beliefs that treatment is ineffective [15]. Cultural factors in Black communities also have the potential to heighten barriers to formal help-seeking. These include the belief that mental illness is a weakness or moral failing [16,17], as well as preferring faith-based management of mental health [18]. Estimates suggest that Black adults are less than half as likely to access mental health services compared to White adults [19].

Black adults who do seek and receive formal care may find that existing treatments fall short in addressing their needs. When given the choice, many Black individuals express a preference for providers who are the same race as them (ie, patient-provider race concordance) [20]; however, underrepresentation of providers of color has made it difficult for most Black adults to achieve such concordance [21,22], which may lead to challenges building rapport with their provider [23] and less satisfaction with treatment [24]. Furthermore, available treatments may overlook culturally relevant aspects of mental health, such as stigma, spirituality, and racial identity [20,25]. Many Black adults may therefore be looking for ways to manage mental health concerns outside of the formal care system.

Recent efforts have included consideration of digital mental health (DMH) tools to address unmet needs, bridge gaps in care, and provide culturally sensitive mental health services. For instance, studies have sought to understand Black adults' preferences for the design of chatbots to support managing chronic conditions [26], voice assistants to support older Black adults in seeking online health information [27,28], and DMH tools to better support Black women in the perinatal period [29]. However, limited work has focused on the needs of Black Americans in the context of scalable tools like automated text messaging programs for mental health self-management.

Text messaging is a particularly accessible method for delivering mental health support [30]. In the United States, cellphone ownership is almost universal, and high rates of text messaging extend across racial and ethnic groups and socioeconomic factors [31,32]. Furthermore, since text messaging is a built-in tool on the mobile phone, and among the most frequently used communication methods, it has extremely high reach and low barriers to entry [33]. Access can extend to individuals who may not own modern smartphones or reliably access a data plan [30]. Text messaging is also cost-effective in terms of program development, maintenance, and deployment [30]. Thus, it may constitute an appropriate modality for equitable, accessible DMH tools. Previous research suggests that Black adults send and receive text messages frequently and may be more likely to use text messaging overall compared to White adults [32]. Black adults also find text messaging to be a suitable approach for treating various health concerns [34].

Previous studies examining the use of text messaging tools to support health improvement across a wide range of medical conditions, including those targeting both specific diagnoses (eg, diabetes, hypertension, and asthma) and health-related behaviors (eg, medication adherence, smoking cessation, and weight loss), indicate positive effects on health outcomes [30]. The potential of text messaging has also been explored in mental health interventions targeting conditions such as depression, anxiety, suicidal thoughts and behaviors, and other mental health challenges, with studies demonstrating moderate yet significant positive effects. Thus, text messaging may offer an effective and scalable modality for delivering mental health support to individuals who otherwise may not receive care, although such programs also have possible risks and ethical considerations in the context of mental health treatment [35-37]. For example, concerns may include privacy and the risk of intrusiveness, management of sensitive health information, and the clinical appropriateness of automated responses to acute distress or crisis situations [35,38].

To address gaps in both the literature and the unmet needs of individuals facing barriers to traditional care, this study applies user-centered design methods to examine how Black adults who are not engaged in formal mental health treatment envision

XSL•FO
RenderX

using a text messaging intervention to manage depression and anxiety symptoms. We recruited Black adults in the United States who self-reported experiences of depression or anxiety and sought—through an online discussion group and parallel text messaging-based technology probe—to deepen our understanding of their perspectives related to mental health support and how DMH tools might help to meet their needs. We posed the following two research questions: (1) What types of mental health challenges are faced by Black adults who are not engaged in formal treatment? and (2) How should digital tools deliver support to users to address their challenges, particularly in the context of automated text messaging? Answering these questions can help align scalable DMH tools to the needs of Black people in the United States, with potential for broad impact in addressing their unmet mental health needs and advancing mental health care equity.

Methods

Overview

This section describes how we worked with Black adult participants to understand their needs and preferences for DMH tools.

Participants

Recruitment procedures for this research were fully remote, providing individuals across the United States the option to participate in the study from any convenient location and using their own devices. Participants were recruited through social media advertising (Instagram and Facebook) and through free online self-screening surveys for depression and anxiety—that is, the 9-item patient health questionnaire (PHQ-9) [39], or 7-item Generalized Anxiety Disorder scale (GAD-7) [40]—hosted by Mental Health America, a large nonprofit mental health advocacy organization. The advertisement text read, “You are invited to participate in a Northwestern online research project on Black and African American mental health,” followed by a link displayed in their social media feed or alongside their screening results. Interested individuals then were guided to a primary screening survey to assess study eligibility, with inclusion criteria specifying that participants should (1) identify as Black or African American, including individuals who also identify as mixed race; (2) self-report a history of depression or anxiety symptoms (whether or not they were professionally diagnosed); (3) reside in the United States; (4) be 18 years of age or older (or 19 years of age or older in Nebraska, reflecting the state’s age of majority); (5) have English language abilities sufficient to read, understand, and participate in study procedures; and (6) own and be willing to use a personal mobile phone. Individuals were excluded if they were currently seeing a therapist, counselor, or psychologist, or taking medications for their mental health symptoms. Individuals were also excluded if they reported a serious mental illness (eg, bipolar disorder and schizophrenia); if they had visual, voice, hearing, or motor impairments that would prevent completion of study procedures; or if they reported suicidal ideation with a current plan and intent to act. Furthermore, as the study involves receiving automated text messages, we confirmed that individuals provided the study team with a valid US mobile phone number before enrollment. Individuals who were deemed eligible after completing the primary screening survey were invited to complete a secondary screening survey, which evaluated the presence of suicidal thoughts, plan, and intent. Participants with a suicide plan or intent received risk assessment and resources and were not eligible to participate. Interested and eligible individuals were asked to sign a code of conduct stating participants would remain respectful to one another and refrain from sharing personally identifiable information (eg, name and address) within any of their responses. They were then asked to provide informed consent and were sent instructions to join the discussion group.

Consistent with guidance for research using asynchronous remote communities [41], we sought to recruit a small group size such that participants could participate actively and build rapport with one another. Of the 27 individuals who met eligibility criteria, provided consent, and enrolled in the discussion group, 2 did not respond to any prompts in the group and were not sent automated text messages or invited for follow-up interviews. Like recruitment procedures, all study activities were fully remote and allowed participants to engage from any convenient location using their own devices. Furthermore, study activities, apart from follow-up interviews, were available to participants asynchronously and could be completed at their leisure within the study timeframe. Data were collected between June 9, 2023, and July 27, 2023. The first discussion group prompt was posted on June 9, 2023, and participants were given the opportunity to engage in discussion group activities until July 3, 2023; however, some participants were interviewed through July 27, 2023, due to scheduling availability. Demographic characteristics for the 25 participants included are represented in Table 1.

XSL•FO
RenderX

Table . Demographics, symptom severity, recruitment source, academic achievement, and treatment history for US-based Black adults included in this qualitative analysis.

Characteristic	Discussion group (n=25)	Interview[a] (n=6)
Age (years), mean (SD)	34.76 (10.96)	27.17 (10.59)
Depression or anxiety symptoms, mean (SD)		
PHQ-9[b]	9.96 (5.70)	10.33 (6.02)
GAD-7[c]	10.36 (5.63)	13 (3.95)
Sex, n (%)		
Female	20 (80)	6 (100)
Male	5 (20)	0 (0)
Race, n (%)		
Black or African American	25 (100)	6 (100)
More than one race[d]	1 (4)	0 (0)
Recruitment source, n (%)		
Mental Health America	3 (12)	0 (0)
Social media	22 (88)	6 (100)
Highest level of education, n (%)		
Some high school or less	1 (4)	1 (17)
High school graduate	5 (20)	1 (17)
Some college	11 (44)	2 (33)
Associate's degree	4 (16)	1 (17)
Bachelor's degree	2 (8)	1 (17)
Post-bachelor's education	2 (8)	0 (0)
Mental health treatment history, n (%)		
Has seen a mental health professional	10 (40)	2 (33)
Previously prescribed medication for a mental health condition	7 (28)	0 (0)

[a]Interviewees were a subset of those who participated in the discussion group.

[b]PHQ-9: 9-item Patient Health Questionnaire.

[c]GAD-7: 7-item Generalized Anxiety Disorder scale.

[d]Participants who identify as Black or African American in addition to another race.

Procedure

Asynchronous Online Discussion Group

To increase comfort discussing issues related to racial identity and mental health, participants were informed that all other study participants identified as Black or African American and had experiences of depression or anxiety. Participants were asked to create a pseudonymous account on the study platform [42]. The platform was programmed to release a new prompt every 3 days for 24 days (8 total prompts). Prompts centered on understanding participants' mental health needs and management strategies, how these are impacted by racial identity, their perceptions and experiences of technologies as a method of mental health support, and ideas for how automated text messaging tools could be adapted to meet their needs. Each prompt posed a series of questions related to a specific topic (see Multimedia Appendix 1).

Participants were asked to provide substantive responses, defined as contributions that advance the discussion such as sharing personal stories or opinions, posing clarifying questions, or reflecting on others' posts. No minimum or maximum word count was imposed on participant responses. Participants were compensated based on the number of prompts to which they responded. They could earn additional compensation by replying to at least one other participant's comment on each prompt. Text-based responses were automatically recorded by the study platform [42] and downloaded at the study's end.

Text Messaging Technology Probe

Participants were enrolled in the Small Steps SMS program partway through the study (on day 6), allowing them the opportunity to engage in broader discussions of DMH tools and Black mental health before focusing on their experiences with and perceptions of the text messaging program. The program delivered support for self-management of depression and anxiety symptoms via daily interactive dialogs [43,44]. The content of

XSL•FO
RenderX

the messages was drawn from 11 evidence-based psychotherapy strategies and included a variety of formats such as psychoeducation messages that introduce each strategy and how it works, peer stories, opportunities to write or receive supportive messages, and skill-building exercises (see Multimedia Appendix 2). The peer stories included in the program were collected from a previous study that used group discussions and writing prompts to gather peer support narratives. Materials were subsequently edited and curated by research staff before being included; thus, participants were not involved in live exchanges with peers. Messages varied in length, reflecting the program's structure, which incorporated both brief, responsive interactions and more comprehensive psychoeducational messages. For example, some texts consisted of only a few words such as "Good morning" or "Thank you for sharing" and were included primarily to enhance the natural conversational tone of the program. In contrast, other messages were considerably longer, often consisting of several sentences and were intended to deliver valuable psychoeducational content. The number of messages sent each day varied based on participants' interactions with the program; participants who responded more frequently may have received a greater number of messages (eg, follow-up questions). Participants could engage with Small Steps SMS as much or as little as they wanted. They could also elect to end all messages by sending "STOP," although no participants did so. Participants were asked to reflect on their experiences with and impressions of Small Steps SMS in the discussion group.

Interviews

Participants were invited to complete optional one-on-one semistructured interviews, lasting about 20 - 25 minutes, at the end of discussion group activities. Interviews gave participants the opportunity to voice ideas or thoughts they did not share in the discussion group and provide additional feedback on Small Steps SMS, as well as ideas for DMH tools more broadly. In total, 6 participants completed interviews. All interviews were conducted via the Zoom teleconferencing program by a member of our research team. Participants were able to turn their video on or off based on their preferences. With participants' permission, interviews were audio recorded and transcribed.

Ethical Considerations

Study activities were approved by the Institutional Review Board (IRB) at Northwestern University (STU00211168). Informed consent was collected from all participants prior to enrollment in the study. Participants were compensated US $7 for responding to each online discussion group prompt and US $2 for replying to at least one other participant's response to each prompt, for a total possible compensation of US $72 for all discussion group activities. Participants who completed interviews were compensated an additional US $8. Our research team includes mental health clinicians and researchers, human-computer interaction researchers, and research staff. Our team includes individuals who are and are not Black, whereas all participants were Black and most were women. To enhance rapport and comfort based on a foundation of shared identity (ie, race and sex), all interviews were conducted by a Black female team member. Clinical psychologist team members provided input throughout the study to ensure participants' wellbeing and safety, which included advising on the study design and code of conduct, developing risk management protocols, and providing supervision of research staff in monitoring screening data, monitoring discussion posts for potential indicators of suicidal thoughts, behaviors, or nonsuicidal self-injury, and conducting risk assessments when necessary. All individuals who completed the screening process or participated in the study were given a list of resources for accessing 24/7 mental health support, if needed (eg, suicide prevention hotline and crisis text line). The website focusgroupit.com [42] was selected to host discussion group activities based on its security features and user-friendly design. The platform adheres to privacy regulations like General Data Protection Regulation and the EU-US Privacy Shield and does not sell user data. To further protect participant confidentiality, no personally identifiable information was stored on the focusgroupit.com website [42]; participants were identified through randomly generated codes linked to their study ID and were asked not to share identifying information. The discussion group was accessible only to enrolled participants. Privacy protocols were reviewed to ensure compliance with IRB requirements. Responses in the discussion group were monitored daily for compliance with the code of conduct. Across all activities, research staff had a risk management protocol in place, such that participants' sharing of any information signaling risk to themselves or others would prompt the research team to contact the participant and administer risk assessments and provide resources as needed. No such risks emerged, and therefore, no follow-up was conducted.

Data Analysis

Discussion group and interview data were subject to thematic analysis using an inductive, semantic approach, which allowed themes to emerge naturally from the data with a focus on the explicit meanings conveyed in participants' words [45]. A total of 3 coders, who are authors of this paper, first became immersed in the data by reading discussion group and interview transcripts. They performed open coding to identify preliminary codes and then met to discuss and prioritize these codes, guided by assessment of each code's relevance to the research questions. Prioritized codes and their definitions were captured in a shared codebook. The coders then applied the shared codebook to overlapping transcripts using the qualitative data analysis software Dedoose (SocioCultural Research Consultants) and subsequently met to discuss and resolve coding discrepancies and to refine and consolidate the codebook. After 4 overlapping coding rounds, discussions ceased to yield codebook revisions, at which point the coders divided the uncoded transcripts and applied the final codebook. Key themes were identified that encompassed the coded data, which were discussed with the rest of the research team for feedback. Excerpts were selected to illustrate each theme.

Results

Overview

All of the Black adults in our study reported facing profound challenges in managing their mental health concerns and were

XSL•FO
RenderX

receptive to using technology to help overcome these challenges. In the following sections, we describe the primary themes, which draw on responses from both the discussion group and follow-up interviews. Interview participants were a subset of discussion group participants, and their responses expanded on key points that emerged in the group discussion. First, we describe the unique mental health challenges participants faced related to navigating discrimination and stigma and perceptions of limited opportunities to share their mental health experiences with others due to concerns about racial bias. Next, we describe participants' receptivity to an automated text messaging program that allows them to address their mental health concerns independently through a convenient medium. Finally, we describe how participants envision that digital tools can be designed to meet the unique mental health needs of Black adults through integrating culturally appropriate content written by and for Black individuals, creating safe spaces for disclosure of mental health concerns, and supporting help-seeking beyond the DMH tool among interested users.

Mental Health Challenges of Black Adults

Black adults in our study overwhelmingly reported being distressed by their mental health symptoms, which included a lack of motivation, fatigue, reduced pleasure in daily activities, loneliness, a sense of overwhelm, and anxiety, among others. As the sections below describe, participants' everyday experiences of depression and anxiety were compounded by habitual negative experiences that relate to race, particularly experiences of discrimination, stigma, and invalidation.

The Role of Racism and Discrimination in Mental Health

When asked how mental health issues differ for Black individuals than for other groups, 88% (22/25) of participants shared that they face pervasive exposure to discrimination and racism across multiple domains of life, which negatively impact their mental wellbeing. A participant referred to these experiences as "anti-Blackness," sharing:

> *Black people unfortunately have to deal with everyday anti-Blackness and racism that exacerbates existing mental health issues.* [Participant 5, Discussion Group]

Participant 5 provided additional insight into the damaging effects of anti-Blackness, emphasizing that its harm can extend beyond direct targets through vicarious racism, such as when accounts of anti-Blackness are shared among Black community members or broadcast publicly on media channels.

> *If we are not encountering it in our day to day lives personally, then we are inundated with instances of racism in the news and social media.* [Participant 5, Discussion Group]

Participants described several specific effects that anti-Blackness had on their mental health, including increasing anxiety and decreasing motivation and energy level. Participant 6 felt that repeated instances of discrimination were linked to their anxiety, relaying that, "We deal with discrimination and injustices and fear and it affects our mental health because we are more prone to be on edge and anxious about certain things and situations" (Participant 6, Discussion Group). Another participant echoed this sentiment and provided further context, noting that encounters with authority figures, particularly law enforcement, can heighten anxiety and fear: "You have to always be on guard. When you're out in public you hope that you don't have any interactions with the police because that could be the end of your life" (Participant 10, Discussion Group). Concerns that interactions with law enforcement could reinforce experiences of discrimination and negatively impact mental health were shared by other participants, including Participant 12: "...we go through discrimination, blacks get arrested quicker than any other race. It does effect your mental health" (Participant 12, Discussion Group). Others connected anti-Blackness to "burn out" and withdrawal, such as Participant 25:

> *For most of us, we are surrounded by whiteness in school, work, and even socially, but there is a level of masking that can lead to feeling burned out, and then social anxiety about re-entering those spaces can pop up. I remember reading about Black workers dreading the return to the office due to the amount of microaggressions they had avoided while working from home.* [Participant 25, Discussion Group]

Participant 25's account highlights the extent to which anti-Blackness pervades participants' lives and identifies schools and workplaces as common settings for microaggressions. Participants emphasized the work that goes into navigating White spaces, observing that this continuous effort could exhaust them and hinder their success in both school and the workplace.

Participants further explained that the feeling of being inundated with anti-Blackness can also impact one's self-concept and self-worth. Participant 5 described:

> *I feel like sometimes being bombarded with images and videos of Black people being killed at the hands of police and other forms of state sanctioned violence, even the election of a white supremacist president to the highest office in the land, it can really feel as though my life doesn't matter, even though it does absolutely.* [Participant 5, Interview]

In this statement, Participant 5 linked exposure to what they described as "state-sanctioned violence" and feelings of worthlessness.

Participant 7 agreed, stating, "It's just like what did we do so wrong for people to hate us so much? Being treated like that makes a lot of people feel unworthy & they may want to harm themselves" (Participant 7, Discussion Group).

Their experiences suggest that feelings of worthlessness can contribute to self-harm ideation.

Disclosure and the Risk of Invalidation

Most participants (14/25, 56%) emphasized the importance of disclosure as a means of addressing experiences of anti-Blackness and processing the resulting impacts on their own mental health, as well as the broader mental health of the Black community. For instance, Participant 15 shared, "open and honest conversation can bring upon a wealth of relief" (Participant 15, Discussion Group), suggesting that disclosure

XSL•FO
RenderX

might aid in reducing mental health symptoms. Although most participants expressed interest in mental health disclosure, they perceived limited opportunities to discuss their experiences, particularly those related to anti-Blackness. In fact, 72% (18/25) of participants mentioned being wary to discuss these issues. They reported hesitation to discuss their experiences with Black and non-Black individuals alike, including health care providers. Participant 14 summarized this sentiment: "A lot of us really don't like talking to therapists, family or friends. So, we keep it bottled on the inside and try to deal with it our self" (Participant 14, Discussion Group). Participants overwhelmingly avoided disclosing how racism impacted their mental health and wellbeing to non-Black individuals, primarily citing fears of invalidation. Participant 5 described:

> *Dealing with discriminatory behavior can be so exhausting and depressing, not to mention the added frustration having our experiences of racism be downplayed or outright dismissed by non-Black people.* [Participant 5, Discussion Group]

This comment suggests that a dismissive response to one's disclosure can have a compounding effect on the initial instance of discrimination. Another participant feared that disclosure could activate damaging stereotypes, relaying that:

> *In my personal experience, it was and is still sometimes hard for me to be vocal without being labeled as the "angry black woman." I've been working on not dimming my own light because of someone's ignorance.* [Participant 1, Discussion Group]

Participants also noted that mental health disclosure could be perceived as a sign of weakness, with one person sharing, "Feel weak when we express our feelings" (Participant 2, Discussion Group). Another participant agreed with this sentiment, stating: "I agree, we express our feelings, but yet no one listens" (Participant 13, Discussion Group). Despite wanting to share their experiences broadly with confidants of other races, participants were attuned to potential negative repercussions that discouraged and often prevented disclosure.

Many participants (14/25, 56%) mentioned that therapy could hypothetically create a supportive environment for disclosing and processing their experiences. Yet, concerns about invalidation extended to mental health care providers. For example, Participant 18 reported:

> *It is hard for [therapy] to work, because not everyone can understand the feelings black people feel unless they experienced it themselves. It is important that the feelings of Black individuals are not invalidated.* [Participant 18, Discussion Group]

A preference for Black providers was discussed by 8 participants (32%). Participant 15 thought that having shared experience would allow them to disclose more candidly to a provider, describing that, "It makes it easier to be understood without shame or saying something that might be offensive to a non-black individual" (Participant 15, Discussion Group). Some also described that providers should understand the history of racial injustice in the United States. Participant 25 stated:

> *I think support systems must be well-versed in the cultural and historical foundation that plays out in our daily life. I think individual, as well as group therapy, led by those who have the training and cultural background, is a great start.* [Participant 25, Discussion Group]

While many participants thought they could benefit from working with providers who could understand their experiences, they expressed some skepticism that such providers would be available to them.

Finally, participants' concerns about disclosure often extended to other Black individuals, particularly when sharing details of their mental health issues, and 64% (16/25) mentioned that their mental health had been negatively impacted by stigma within the Black community. They feared that sharing their struggles with family, friends, and other community members could lead to judgment or ostracism. Participant 21 explained:

> *I feel like there's this stigma within the African American community, especially with surrounding mental health. I know I've been told by family members—not immediate family members, but kind of distant family members like, "Oh, therapy is not for our people; our type of people don't go to therapy." We don't broadcast our problems; we kinda internalize it.* [Participant 21, Interview]

A similar sentiment was relayed by Participant 15, who shared:

> *I have loved ones who think that talking about depression means I need to be placed in a mental facility. It is sad because some of them unknowingly need to seek therapy but most likely will not because they fear the stigma behind it.* [Participant 15, Discussion Group]

Some participants connected negative attitudes about mental health treatment to religious beliefs. Participant 11 stated:

> *African Americans deal with shame, and backlash for having depression and or anxiety. I grew up with a Christian mother and grandmother that tells you to pray about everything. Mental illness in general in the black community is a taboo thing.* [Participant 11, Discussion Group]

Participants expressed optimism that attitudes toward mental health treatment were shifting such that Black communities were becoming more accepting; however, Participant 25 reflected on both progress and disappointment, expressing pride in the growing openness around mental health within the Black community while simultaneously mourning the lack of support for mental health concerns historically:

> *I am proud of the evolution our community has experienced when it comes to [mental health], but I also feel saddened for all of us who suffered without support, and were forced to just "pray" it away.* [Participant 25, Discussion Group]

Overall, participants faced a dilemma. They felt a strong desire to have their experiences acknowledged and understood by others but perceived few appropriate outlets within which to

XSL•FO
RenderX

share, leading many to keep their mental health struggles private. As we describe in the next section, DMH tools were largely perceived positively since they provided a low-burden, private, and nonjudgmental method for receiving day-to-day support.

Receptivity Toward an Automated Text Messaging Program

Although participants in our study were not seeking formal mental health care, they largely saw the value of self-guided tools to help them manage their mental health privately, particularly when tools took advantage of convenient communication channels.

Introducing and Normalizing Self-Management

After they began receiving messages from Small Steps SMS in the technology probe, participants described the program positively, with one person suggesting that it aided in motivation, "I like that it keeps me motivated throughout the day" (Participant 8, Discussion Group). This perspective was reinforced by a second participant who responded, "Yes I feel the same way. It keeps me motivated and on track" (Participant 20, Discussion Group). Participant 16 believed the program empowered them to become more proactive in their own mental health. They described:

> *I think this program would be very beneficial for African Americans like myself. Especially when we are not ready to talk to someone face to face … I would definitely sign up for the program to help me cope with my anxiety and depression.* [Participant 16, Discussion Group]

Indeed, some participants reported they tried out the strategies supported by Small Steps SMS, such as Participant 15, who emphasized their success in challenging negative thoughts (one of the strategies the program supports). They described:

> *But reading those different tips and the different self-care things that were said, how to take that negative thought and kind of turn it into a positive one, that's been very helpful.* [Participant 15, Interview]

Some other strategies highlighted by participants included cognitive reappraisal, "I like receiving the messages. They have really given me things to think about in a different perspective" (Participant 4, Discussion Group), and relaxation techniques, "Also [Small Steps messages] help me deal my angry issues and help me calm me down. The meditation is really good. Just help you relax too" (Participant 22, Discussion Group). Participants also appreciated that the program felt more comfortable and approachable than face-to-face treatment, particularly because it was perceived to be private and judgment-free while allowing participants to feel anonymous, with Participant 16 describing:

> *I really enjoy getting the text messages. It really helps when I'm down and depressed and keeps my mind off things for that moment … it's like having your own personal counselor no judgment at all that makes me feel comfortable and to open up about my feelings more.* [Participant 16, Discussion Group]

For some participants, the nonjudgmental nature of the program may have fostered a sense of normalization that allowed them to be more accepting of and open about their mental health challenges. This normalization may have been achieved, in part, through the user-generated content integrated throughout Small Steps SMS in the form of peer stories and support messages. Participant 18 described:

> *Usually I don't like reading long messages, but if it was a story about how [a person] overcame something and the lessons that they learned from it, I actually loved reading those because it felt really personal and like something I could relate to* [Participant 18, Interview]

Similarly, Participant 5 reported:

> *I was not expecting to receive instances of people's personal experiences, and it was honestly my favorite part of the experience. I think having people share their experiences makes me feel less alone* [Participant 5, Discussion Group]

Although Small Steps SMS was automated, peer stories could provide a sense of relatability and social support that was otherwise lacking and were endorsed by 40% (10/25) of participants as a strongly favored feature.

Convenience of Text Messaging

Participants liked that digital tools like Small Steps SMS were convenient, addressing several barriers to traditional mental health treatment, including time and cost. Participant 15 reflected that:

> *This program is convenient as it does allow someone the comfort of seeking help without the hassle of an in-person visit. Programs like this one are time and money saving.* [Participant 15, Discussion Group]

The system was also perceived to be convenient since it operated via text messaging, a function of the mobile phone that participants all routinely used, and because messages were brief and required little active engagement. For example, Participant 24 described:

> *Well, I guess one thing for me in particular, since I'm very much a workaholic, was the reminders of little self-care things I thought was nice. And just the reminders to kinda take some time to just decompress, so to speak.* [Participant 24, Interview]

At the same time, participants appreciated that the system allowed 2-way communication. This was described by Participant 11, who shared: "I liked the fact that some of the messages were interactive and required some participation" (Participant 11, Discussion Group). A few participants mentioned potential challenges with receiving mental health support via text messages, including the frustration of receiving too many messages in one day, or too often, and finding them too lengthy. However, participants overall endorsed the low burden of receiving push messages and also recognized that asking the user for input and participation could add value.

XSL•FO
RenderX

Tailoring Digital Mental Health Tools for Black Adults

Participants described several ways the program, and other DMH tools, could be adapted to help them navigate the specific mental health challenges that emerge for Black adults.

Adapting Content to Reflect Black Individuals' Experiences

Most participants (15/25, 60%) stated that Small Steps SMS should directly address the unique challenges faced by Black adults with mental health concerns. Participant 15 described, "The program should list some common topics black people face on a daily basis such as racism, various traumas, and other systematic barriers" (Participant 15, Discussion Group). Participants also endorsed educating users on the mental health effects of racism. Participant 5 shared:

> *I think having topics such as systemic racism and the adverse effects on the mental health of Black Americans could be very beneficial since it serves to validate the daily injustices we face due to attitudinal racism* [Participant 5, Discussion Group]

In addition to facilitating understanding of bias and discrimination, participants wanted the program to provide actionable guidance for navigating these challenges. Participant 7 suggested, "support should help just by letting us know that everyone isn't racist or mean and teach us how to properly cope and deal with these things" (Participant 7, Discussion Group). Similarly, Participant 19 thought that the program should help connect users with resources around a range of issues faced by Black individuals:

> *Well racism and injustice is still alive and kicking. Social issues, high unemployment rates, policing problems … we deal with a lot … maybe provide information for different avenues to assist with resolving problems* [Participant 19, Discussion Group]

Participants felt that content addressing these issues should reflect Black voices. Participant 15 described, "anything that can be shared from someone that identifies as the same race with me, that's positive" (Participant 15, Interview). Similarly, Participant 17 stated, "especially if the content sent was authored and curated by Black Voices with Blackness in mind" (Participant 17, Discussion Group). Some participants expressed a particular interest in non-professional content written by other Black Americans. Participant 11 shared,

> *I would like to see relatable stories that highlight being African American, the stigma of mental health treatment/diagnosis in our community, juggling parenting, working, and on top of that battling mental health. The stories should vary in outcomes, successful and failed outcomes, [relatable] uplifting stories, funny stories, inspirational stories.* [Participant 11, Discussion Group]

Several participants also endorsed integrating quotes from famous Black individuals, including "well known people, like Oprah" (Participant 25, Discussion Group). Similarly, Participant 18 shared:

> *Sending inspirational quotes by famous Black people I think would help. Because they will see people who are like them who have been in situations that are stressful and have made it through the situations. So that would be encouraging.* [Participant 18, Discussion Group]

Participants saw value in integrating diverse Black voices into the content of a messaging program, including professionals, peers, and celebrities.

Constructing Safe Spaces for Social Sharing

Although the version of Small Steps SMS used in our technology probe featured user-generated content composed by other individuals with mental health concerns, some participants thought that the system should go farther in how it supports social interaction.

In fact, 36% (9/25) of participants suggested facilitating direct conversation between Black users of a DMH tool, often drawing inspiration from the discussion group through which we conducted the study, which allowed them to compose long-form messages and engage with one another while maintaining anonymity. These features led to a sense of safety, as Participant 16 described: "I definitely appreciate that I can be open and honest about my feelings without negative judgment. [This discussion group] has been very beneficial for me and my mental health" (Participant 16, Discussion Group). Participant 15 agreed, stating:

> *This group is nice, and I, too, would consider this a safe space. I feel that I can be open and honest and will get back the same in return. I do not feel worried about being judged, I can be completely who I am, and still feel valued.* [Participant 15, Discussion Group]

This comment suggests that a safe space is one where all participants can be uninhibited in expressing themselves, without fear of judgment.

A few participants (12%, 3/25) spoke to the importance of establishing a sense of safety by ensuring that forums were accessible only to Black individuals with mental health concerns. For example, Participant 25 suggested "making sure there is gatekeeping to protect the space, so that it's only for people who are Black American (or Biracial) and with a history of mental wellness challenges" (Participant 25, Discussion Group). Participant 23 shared this view, explaining that gatekeeping is essential given the scarcity of outlets for people like them: "I feel like it should be only catered to black Americans with mental issues because we don't have a lot of safe spaces to talk about these things at all" (Participant 23, Discussion Group). However, a participant who identified as multiracial relayed past experiences in which they had felt excluded from Black spaces and expressed that the program should also include adults with mixed racial backgrounds. They stated:

> *I think it'd be nice if you kind of opened it up to, maybe not people who were just fully African American, but that also identify as African American, but something else as well, because I feel like*

XSL•FO
RenderX

> *oftentimes, people try to take that away from people.* [Participant 21, Interview]

Participants also discussed how creating a safe environment for disclosure would require moderation, or the active involvement of individuals whose job is to establish and uphold community guidelines. Participants once again took inspiration from the study's discussion group, which was moderated by the study staff. Participant 17 described, "I think being a highly moderated and regulated platform allows people to speak about their mental health candidly while maintaining anonymity" (Participant 17, Discussion Group).

Finally, a sense of safety was one in which sharing was optional and could happen over time as participants became comfortable. A considerable number of participants (40%, 10/25) described that reading about others' experiences was valuable regardless of their own desire to share, providing a sense of reassurance and connection, or an understanding that they were not alone. Participant 9 reported, "I appreciate this group. It showed me I'm not losing my mind that there are other women going through the depression, anxiety etc" (Participant 9, Discussion Group).

Ultimately, participants expressed that feeling safe and supported in their environment would make them more likely to open up about their mental health experiences, with 24% (6/25) of participants mentioning more favorable attitudes toward disclosure following the conclusion of the focus group. Furthermore, they drew from their positive experiences in the discussion group to highlight the value of interaction, boundaries, and moderation, all of which allowed them to feel they could share freely, without the risk of invalidation that characterized many other spaces they inhabited in their day-to-day lives.

Connecting to Other Sources of Support

While participants in this study were not involved in formal care, many saw the potential value in it, and this increased for some over the course of the study. Participants described that a program like Small Steps SMS may initially appeal to them because of its focus on privately managing one's concerns, but that over time it might inspire individuals to seek additional formal or informal mental health support. For example, Participant 3 reported, "I think this opened my mind to possibly seeking mental health counseling" (Participant 3, Discussion Group). Similarly, Participant 5 described, "when I read stories of people reaching out to their pastors or their therapists, or a trusted friend, it just reminded me that I could seek out someone in my own circle" (Participant 5, Discussion Group).

Recognizing this possible evolution in attitudes around treatment, 32% (8/25) of participants mentioned that they wanted a DMH program to help users locate Black mental health care providers, who were perceived to be equipped to provide culturally sensitive support. For example, Participant 24 reported:

> *I think it would be great if the system had an option to see resources for POC [people of color] in their area. Maybe team up with black therapists on this as well. Another idea, would be the option to chat with a black therapist [real person] if necessary or if the AI alone isn't tailored to the individual enough* [Participant 24, Discussion Group]

Participants agreed that such features must be optional but saw value to integrating resources within a DMH tool to allow users flexibility in how to continue their mental health journeys.

Some participants also indicated that using the Small Steps SMS program and participating in the discussion group could inspire greater disclosure of their mental health challenges to important people in their lives. To support these behaviors, participants suggested that a DMH tool could provide guidance for navigating mental health stigma within Black communities. Participant 18 reflected that:

> *It is hard for [Black people] to accept that we need help. I think this program will give many people the push they need to get help or even talk to family or friends. especially if it introduces people to resources that helpful for mental health* [Participant 18, Discussion Group]

Thus, while many participants were initially receptive to Small Steps SMS because it could be used privately, a few (3/25, 12%) suggested that the program should support transitions in how individuals cared for their mental health, including greater willingness to engage in formal and informal help-seeking.

Discussion

Principal Findings

Our goal in this study was to understand the challenges faced by Black adults in the United States and to clarify their preferences and perspectives for the design of DMH tools that can be used outside formal care, including automated text messaging tools. Our research was guided by the following two research questions: (1) What types of mental health challenges are faced by Black adults who are not engaged in formal treatment? (2) How should digital tools deliver support to users to address their challenges, particularly in the context of automated text messaging? Our findings suggest that participants perceived anti-Black racism to affect their mental health daily, and that these experiences translated into a deep need for disclosure and social validation that they could not easily satisfy in their daily lives. They deemed automated text messaging and DMH interventions more broadly to be an acceptable approach to addressing mental health concerns through the introduction of mental health self-management strategies; however, they also clearly identified ways to better represent their experiences within the content of the Small Steps SMS program, as well as complementing pre-scripted content with safe opportunities for social sharing and pathways to access mental health support beyond the DMH tool. Below, we discuss the implications of our findings for understanding the mental health challenges of Black Americans and for designing DMH tools for this population.

XSL•FO
RenderX

Understanding Mental Health Challenges and Disclosure Needs of Black People

Our study suggests that participants made many connections between incidents of anti-Black racism and their affective states, thoughts, and behaviors, including linking exposures to racism to changes in their mood, anxiety level, productivity, and feelings of devaluation. They also emphasized that these effects are cumulative, reducing educational or career achievement and self-esteem and potentially leading to burnout or even suicide or self-harm-related thoughts and behaviors.

Participants in our study reported that they were affected not only by firsthand experiences of discrimination, but also by exposure to others' experiences as conveyed through social networking apps and mainstream news. Exposure to state-sanctioned violence was especially upsetting. The implication of this finding is significant given the highly publicized nature of events like the murder of George Floyd in 2020, with video footage of the murder disseminated widely both through news media and social media. These findings underscore a need to better understand the effects of vicarious experiences of anti-Black racism, including how these effects may be shaped by the type of event (eg, institutional discrimination, violence, and microaggressions), the channels through which experiences are shared (eg, news media, social media, or face-to-face), and accompanying commentary and contextualization. Additional research should examine strategies to mitigate the harms of vicarious trauma without compromising opportunities to access the benefits of sharing, such as building public understanding, spurring collective action, or coping with personal distress.

Findings from our study suggest that mental health disclosure, particularly in relation to instances of anti-Black racism, was viewed by participants as a pathway to coping. Conversely, inhibition was perceived as unhealthy, leading participants to feel the pressure of their "bottled up" distress. This view of disclosure as a potential solution for intense negative emotions is widely held in US culture [46,47]. In a model from clinical psychology, disclosure is compared to a fever as both a sign of an underlying destructive process and a restorative force [46]. There is some evidence for such views; Pennebaker, in his "expressive writing" paradigm, proposed that the act of writing about troubling thoughts and feelings could spur various benefits for writers, even if what they wrote was never shared with others [48]. The benefits of private writing may include relieving stress, facilitating the development of new perspectives, coping with illnesses and traumas, or improving general health and well-being [49]. However, the effectiveness of disclosure as a coping strategy, as well as its associated health benefits, may largely depend on cognitive processing of stressors, which plays a crucial role in race-based trauma recovery and adjustment [50,51]. In parallel, a body of literature has accumulated on the unique benefits that can come from social sharing of emotions, which may include eliciting social support, discovering alternative perspectives, prompting reciprocal disclosure, and potentially deepening relationships [52-54]. Despite its potential benefits, social sharing is complicated when disclosing identities or experiences that are stigmatized [55,56]. Concerns about how one's disclosure will be received are common, and anticipating negative reactions can lead to selective disclosure or nondisclosure [57,58], which is consistent with our findings.

Some work has examined how complexities of disclosure increase for individuals who have multiple stigmatized identities [59-61], reflecting that systems of oppression related to each identity can interact with one another to transform and sometimes magnify the experience of bias [62]. As described in our findings, participants found that identifying as a Black person with mental health concerns could compound the burden of racism and mental health stigma and reduce disclosure opportunities. Participants greatly desired to discuss the role racial discrimination played in their day-to-day mental health and well-being, but they were subject to judgment from non-Black individuals when sharing about anti-Blackness, and they anticipated stigma from Black individuals when sharing about their mental health struggles. The resulting unmet need for disclosure provides an important backdrop for introducing DMH tools for Black people.

Design Considerations for DMH Tools Catering to Black Americans

Our findings can help to inspire design and refinement of DMH tools to better match the priorities and needs of Black people in the United States. Below, we discuss opportunities to deliver self-management support and culturally tailored content through text messaging, to construct safe digital spaces for social sharing, and to provide pathways to other forms of care.

Representing Black Experiences and Voices in Text Messaging Programs

These findings are generally encouraging regarding automated text messaging programs for Black people in the United States. Following our technology probe, participants reported positive views of automated text messaging as a modality for mental health support. The text messaging program was recognized for its capacity to support private and convenient self-management, which allowed them to bypass formal disclosure or help-seeking. In a previous study investigating the acceptability of text messaging for facilitating one-on-one mental health support among Black women, participants voiced privacy and confidentiality concerns [63]; however, these concerns were not independently raised by our participants, who were encouraged to voice any potential concerns about DMH tools and text message-based tools like Small Steps, although they were not asked directly about privacy issues. The reason for this difference is unknown. It is possible that an automated program is less likely to raise concerns about social judgment or misrepresentation of identity relative to one where individuals discuss their mental health with professionals. Participants in our study also identified as nontreatment-seeking, expressing interest in alternative methods of mental health support (eg, automated mental health support) rather than traditional approaches involving trained professionals. Thus, participants may have been inclined to accept DMH tools and to perceive automated support more favorably. In addition, participants received messages from Small Steps in the context of a research study where they had learned about the program through recruitment and orientation procedures, as well as discussing DMH tools with peers in the discussion group, which could

XSL•FO
RenderX

have contributed to reduced concerns about the modality of text messaging. Future research should explore how privacy concerns may be thoughtfully addressed in texting programs, or other DMH programs, that are tailored to the needs of Black adults [64,65].

Many of our participants' concrete suggestions for adapting the text messaging program centered on the content. Participants advocated for representing issues related to race and elevating the voices of Black people. Suggested topics included racism and discrimination, stigma, and daily life stressors affecting Black communities (eg, unemployment and policing). Participants wanted the program to be clearly designed "for us" (ie, Black people with mental health concerns). A primary avenue to achieve this is by including Black voices. Efforts may extend to didactic content (eg, written by Black mental health professionals) and stories of peers managing similar challenges. Although the stories in our text messaging probe were not designed to represent the issues of Black people in particular, our findings suggest they were relatable based on shared mental health challenges. Relatability could likely be further enhanced by including stories explicitly written by Black Americans. These content-related suggestions are feasible to implement within a text-based program, including through crowdsourcing methods [66], or potentially soliciting stories directly from users of the DMH tool as part of their therapeutic disclosure process [64].

Integrating "Safe Spaces" Within DMH Tools

While some design ideas focused on content that could be integrated within the existing structure of Small Steps SMS, others pushed the boundaries of what an automated text messaging program can deliver. Although the discussion group was originally intended as a space for participants to share feedback on Small Steps SMS, participants unexpectedly drew on the experience as a model for future DMH tool design, emphasizing the importance of community and peer connection. In particular, they highlighted features that could create a "safe space" for sharing without the risk of judgment or invalidation. In other studies, asynchronous discussion groups have been used not only to establish understanding of users but as delivery tools for social support [67]. Our findings suggest that these groups may provide an appealing approach to support Black individuals' mental health. Participants indicated that the benefits of the discussion group stemmed from its interactive format, including threaded conversations, the group's moderation processes, and the boundaries of group membership. Although less frequently highlighted by participants, the group also featured anonymity, researcher-constructed discussion prompts, a text-based format, and asynchronicity. In what Walther [68] termed "hyperpersonal communication," asynchronicity can allow participants time to deliberately present themselves, while the reduction of nonverbal cues in text-based exchanges allows readers to focus on the content of the disclosures, potentially supporting idealization of one another and bonding that may even exceed what face-to-face communication allows.

There are various technologies that could potentially be applied to construct digital "safe spaces" for therapeutic disclosure. Our discussion group shared many similarities with online support groups, which individuals can use to connect with geographically dispersed strangers, and which are typically bounded by a shared experience or identity, creating a comfortable context for disclosure of sensitive issues [69]. Many online support groups include features like interactivity, anonymity or pseudonymity, and moderation [70,71]. Using SMS text messaging or other messaging apps could potentially allow for both interpersonal and programmatic support via the same technology. However, it is important to consider the specific context under which DMH tools facilitate the disclosure of distressing events, particularly traumas. Disclosure alone, if not accompanied by appropriate support and guidance, may increase emotional arousal, which has the potential to be harmful for individuals [50,72]. Thus, moderators may require specific training in responding to trauma disclosures and linking individuals to appropriate resources.

Regardless of the technology used, constructing supportive peer-to-peer communication environments is generally less scalable than delivering fully automated tools, particularly when forums are moderated and when boundaries are enforced around membership. In our study, participants' eligibility was vetted, and staff were trained in a code of conduct, risk management protocols, and moderation procedures. They also reviewed, daily, all new messages posted within the group and acted, if needed, to uphold the code of conduct and ensure safety. Such human labor is often the most complex or costly aspect of a DMH program [73,74], although efforts are underway to make moderation processes more scalable, such as through crowdsourcing [75] or automation [76,77], although these efforts must be complemented by appropriate human-in-the-loop review [78,79]. While challenging, providing moderated social spaces may be essential to address the paucity of disclosure opportunities Black people experience and their specific fears of invalidating responses.

Extending Openness Beyond the DMH Tool

Despite not being involved in formal mental health care, many participants were open to the possibility of seeking it in the future. The primary barrier they reported was their limited ability to find providers with shared racial backgrounds. Many participants described that they would connect to formal mental health services that are designed with Black people in mind. They suggested that a DMH tool should increase understanding and availability of culturally appropriate services, including by providing directories of Black clinicians and connecting users to immediate technology-delivered care from Black providers (eg, teletherapy). Thus, participants challenged the division of support into formal and informal programs, suggesting that these should be integrated.

Furthermore, some participants found that using the text messaging program and engaging in the discussion group changed how they felt about treatment. Some described a more general change in openness, such that they considered sharing about their mental health with close others to whom they had not previously disclosed. Others noted that their engagement with Small Steps SMS helped them recognize gaps in their existing support systems, prompting them to seek additional or

XSL•FO
RenderX

alternative forms of help. Designing for ongoing help-seeking and eventual termination would recognize the evolving nature of individuals' support needs and the ways that different forms of informal, formal, and interpersonal support may be complementary over the long term.

Comparison With Previous Work

Our findings add to a growing literature establishing the negative effects of pervasive racism, bias, and discrimination on the health of Black adults [7,80,81]. Past work has examined the effect of racism on outcomes like chronic stress, general mental and physical health, and cardiometabolic conditions [82,83]. For example, a study that used data from Gallup and the US Census before and after the 2020 death of George Floyd to examine anger, sadness, depression, and anxiety among Americans found that these emotions and symptoms spiked among Black people [84]. In addition, by leveraging mobile phone technologies, research has begun to examine the in-the-moment effects of experiences of discrimination, which may lead to longer-term outcomes [85,86].

Our work provides support for some established approaches to inclusive design of DMH tools. Past work suggests that story-based text messages (ie, narratives) can be a valuable form of mental health support and are well-received when they are perceived to be authentic, when they balance positivity with a realistic portrayal of the narrator's struggles, and when narrators are relatable to the receiver [66]. The relatability of narratives has also been enhanced by closely matching the circumstances of the story authors and receivers, such as when African American veterans compose narratives of managing hypertension for other African American veterans with the same condition [87]. Inclusive design may also involve creating a "safe space" wherein users can directly interact with those with shared identities and express themselves authentically [88]. Some work has sought to support conversation with peers via texting, including as one-to-one text chats. Work by Ybarra et al [89] facilitated one-to-one chats between sexual minority adolescent boys alongside pre-scripted automated messages that supported practicing safe sex and HIV prevention. Moderated group chats have also been applied to improve psychosocial functioning [90], increase preventive care [91], and support outpatient psychiatric care [92].

Furthermore, past work has emphasized the need for technologies to support transformations that occur, over time, in individuals' social needs as they use a DMH program. For example, individuals may move from needing informational support when considering and initiating behavior change to needing partnership and accountability when maintaining their change [93]. Our findings suggest a similar need for technologies to align with changes in openness to help-seeking. This may include framing help-seeking as an evolving process, demystifying it, and prompting periodic consideration of the sorts of help a person might want, as well as having external resources available to participants at any time. Decision-making tools may also help the user arrive at the appropriate treatment approach given their current priorities [94,95]. Such processes of deciding and committing to continue or leave treatment are built into some psychotherapies (ie, termination) [96], but largely have not been explored in relation to DMH tools. Such an approach might avoid a common problem in which users report that their disengagement with a tool, even if that tool has met their needs, becomes a source of guilt and stress that is counterproductive to their mental health [97].

Limitations

This study has limitations. First, although we sought to minimize barriers to participation, our sample represents those willing to discuss mental health issues and interact with researchers. Given our research focuses on DMH tools, our sample may also represent those who are interested in or comfortable with technologies. In addition, our study included web-based and SMS components, which limits our ability to speak to the needs of those who lack access to a mobile phone or the internet. Our sample overrepresents women. This is typical in DMH research but limits our ability to represent the needs of Black men. Mental health stigma can be especially high for Black men [17], leading some to argue for programs specifically targeted or tailored to Black men [98]. Our study also included anyone reporting their race as Black or African American, therefore encompassing individuals from diverse ethnic and cultural backgrounds, including individuals who may be African American, Black Caribbean, or African migrants and immigrants, or second-generation immigrants, among others. However, we did not collect data on these backgrounds or how they shaped participants' experiences and views. Future research should consider within-race ethnic differences when designing and tailoring DMH tools. Our inclusion criteria also allowed for multiracial participants (if they also identified as Black or African American), but their representation in our sample was low. The needs of multiracial individuals are important to consider, especially when tools are tailored toward a specific racial identity or background. In addition, our sample size limits the generalizability of our findings. Although this study encouraged participants to voice concerns regarding mental health support via text message, they were not explicitly prompted to reflect on specific issues like privacy concerns. Further research is necessary to understand how risks related to text messaging for health intervention are perceived by diverse populations with mental health concerns and to develop strategies that mitigate these risks while maximizing the benefits of digital tools.

Conclusions

Despite the potential for DMH tools to help address racial disparities in mental health care, this has yet to be achieved. To contribute to the design of tools that can reach and benefit Black adults outside of formal care, this paper examined how nontreatment-engaged Black adults perceived an existing automated text messaging tool and how they envisioned DMH tools being designed and refined to better reflect their experiences and meet their needs. Participants shared that the mental health challenges of Black individuals differ from other racial groups in that they are compounded by inequity, discrimination, and bias, and that these issues must be directly addressed within mental health tools. Participants endorsed text messaging as a medium for delivering convenient and nonstigmatizing self-management support but also endorsed

XSL•FO
RenderX

platforms that can facilitate moderated conversations between individuals with shared racial backgrounds, emphasizing that such opportunities are essential since they lack other "safe spaces" within which to discuss the role of race in their mental health. Thus, scalable DMH tools like Small Steps SMS can alleviate barriers to treatment that may disproportionately burden Black adults by providing a nonjudgmental space for mental health disclosure, which can help to ameliorate the extent to which experiences of racism, stigma, and bias limit Black adults' access to care. Future work should consider how text messaging can support social interaction between Black participants, as well as consider other platforms that can construct a safe social environment to complement self-management programs.

Acknowledgments

This work was supported by grants from the National Institute of Mental Health (K01MH125172, R34MH124960, and P50MH119029). In addition, a gift from the Microsoft AI for Accessibility program to the Center for Behavioral Intervention Technologies, in part, supported this work.

Data Availability

The datasets generated and analyzed during this study are not publicly available due to their sensitive nature and the risk of reidentification but are available from the corresponding author upon reasonable request.

Conflicts of Interest

JM has accepted consulting fees from Boehringer Ingelheim and Shirley Ryan Ability Lab. RK has accepted consulting fees and honoraria from CHESS Health and Jimini Health. DM has accepted honoraria and consulting fees from the National Association of State Mental Health Program Directors and Clinical Care Options, royalties from Oxford Press, and has an ownership interest in Adaptive Health, Inc. ABP has received financial compensation from Otsuka and Apple Inc for mental health-related educational activities. TM is an associate editor of the Journal of Medical Internet Research. None declared by the other authors.

Multimedia Appendix 1
Discussion group prompts analyzed in this study.
[DOCX File, 20 KB - formative_v9i1e73279_app1.docx]

Multimedia Appendix 2
Example participant SMS text messaging interactions in this study.
[DOCX File, 745 KB - formative_v9i1e73279_app2.docx]

References

1. Major depression. National Institute of Mental Health (NIMH). URL: https://www.nimh.nih.gov/health/statistics/major-depression [accessed 2024-02-13]
2. Any anxiety disorder. National Institute of Mental Health (NIMH). URL: https://www.nimh.nih.gov/health/statistics/any-anxiety-disorder [accessed 2024-08-05]
3. Bailey RK, Mokonogho J, Kumar A. Racial and ethnic differences in depression: current perspectives. Neuropsychiatr Dis Treat 2019;15:603-609. [doi: 10.2147/NDT.S128584] [Medline: 30863081]
4. Kirmayer LJ. Cultural variations in the clinical presentation of depression and anxiety: implications for diagnosis and treatment. J Clin Psychiatry 2001;62 Suppl 13:22-28. [Medline: 11434415]
5. Ettman CK, Koya SF, Fan AY, et al. More, less, or the same: A scoping review of studies that compare depression between Black and White U.S. adult populations. SSM - Mental Health 2022 Dec;2:100161. [doi: 10.1016/j.ssmmh.2022.100161]
6. Breslau J, Kendler KS, Su M, Gaxiola-Aguilar S, Kessler RC. Lifetime risk and persistence of psychiatric disorders across ethnic groups in the United States. Psychol Med 2005 Mar;35(3):317-327. [doi: 10.1017/s0033291704003514] [Medline: 15841868]
7. R. Williams D, Williams-Morris R. Racism and mental health: the African American experience. Ethn Health 2000 Aug;5(3-4):243-268. [doi: 10.1080/713667453]
8. Pieterse AL, Todd NR, Neville HA, Carter RT. Perceived racism and mental health among Black American adults: a meta-analytic review. J Couns Psychol 2012 Jan;59(1):1-9. [doi: 10.1037/a0026208] [Medline: 22059427]
9. Boardman JD. Health pessimism among black and white adults: the role of interpersonal and institutional maltreatment. Soc Sci Med 2004 Dec;59(12):2523-2533. [doi: 10.1016/j.socscimed.2004.04.014] [Medline: 15474206]
10. Mersky JP, Choi C, Plummer Lee C, Janczewski CE. Disparities in adverse childhood experiences by race/ethnicity, gender, and economic status: Intersectional analysis of a nationally representative sample. Child Abuse Negl 2021 Jul;117:105066. [doi: 10.1016/j.chiabu.2021.105066] [Medline: 33845239]

XSL•FO
RenderX

11. Nadal KL, Griffin KE, Wong Y, Hamit S, Rasmus M. The impact of racial microaggressions on mental health: counseling implications for clients of color. J Couns Dev 2014 Jan;92(1):57-66. [doi: 10.1002/j.1556-6676.2014.00130.x]
12. Howarth EJ, O'Connor DB, Panagioti M, Hodkinson A, Wilding S, Johnson J. Are stressful life events prospectively associated with increased suicidal ideation and behaviour? A systematic review and meta-analysis. J Affect Disord 2020 Apr 1;266:731-742. [doi: 10.1016/j.jad.2020.01.171] [Medline: 32217256]
13. Yelton B, Friedman DB, Noblet S, et al. Social determinants of health and depression among African American adults: a scoping review of current research. Int J Environ Res Public Health 2022 Jan 28;19(3):1498. [doi: 10.3390/ijerph19031498] [Medline: 35162519]
14. Cook BL, Trinh NH, Li Z, Hou SSY, Progovac AM. Trends in racial-ethnic disparities in access to mental health care, 2004-2012. Psychiatr Serv 2017 Jan 1;68(1):9-16. [doi: 10.1176/appi.ps.201500453] [Medline: 27476805]
15. Alang SM. Mental health care among blacks in America: confronting racism and constructing solutions. Health Serv Res 2019 Apr;54(2):346-355. [doi: 10.1111/1475-6773.13115] [Medline: 30687928]
16. Beauboeuf-Lafontant T. You have to show strength: an exploration of gender, race, and depression. Gend Soc 2007;21:28-51. [doi: 10.1177/0891243206294108]
17. Hudson DL, Eaton J, Banks A, Sewell W, Neighbors H. "Down in the Sewers": perceptions of depression and depression care among African American men. Am J Mens Health 2018 Jan;12(1):126-137. [doi: 10.1177/1557988316654864] [Medline: 27329141]
18. Chatters LM, Taylor RJ, Jackson JS, Lincoln KD. Religious coping among African Americans, Caribbean Blacks and Non-Hispanic Whites. J Community Psychol 2008 Apr;36(3):371-386. [doi: 10.1002/jcop.20202] [Medline: 21048887]
19. Rivera KJ, Zhang JY, Mohr DC, Wescott AB, Pederson AB. A narrative review of mental illness stigma reduction interventions among African Americans in the United States. J Ment Health Clin Psychol 2021;5(2):20-31. [doi: 10.29245/2578-2959/2021/2.1235] [Medline: 34632464]
20. Moore C, Coates E, Watson A, de Heer R, McLeod A, Prudhomme A. "It's Important to Work with People that Look Like Me": Black patients' preferences for patient-provider race concordance. J Racial and Ethnic Health Disparities 2023 Oct;10(5):2552-2564. [doi: 10.1007/s40615-022-01435-y]
21. Wyse R, Hwang WT, Ahmed AA, Richards E, Deville C. Diversity by race, ethnicity, and sex within the US psychiatry physician workforce. Acad Psychiatry 2020 Oct;44(5):523-530. [doi: 10.1007/s40596-020-01276-z] [Medline: 32705570]
22. Santiago CD, Miranda J. Progress in improving mental health services for racial-ethnic minority groups: a ten-year perspective. Psychiatr Serv 2014 Feb 1;65(2):180-185. [doi: 10.1176/appi.ps.201200517] [Medline: 24178249]
23. Nazione S, Perrault EK, Keating DM. Finding common ground: can provider-patient race concordance and self-disclosure bolster patient trust, perceptions, and intentions? J Racial and Ethnic Health Disparities 2019 Oct;6(5):962-972. [doi: 10.1007/s40615-019-00597-6]
24. Laveist TA, Nuru-Jeter A. Is doctor-patient race concordance associated with greater satisfaction with care? J Health Soc Behav 2002 Sep;43(3):296-306. [doi: 10.2307/3090205] [Medline: 12467254]
25. Thompson VLS, Bazile A, Akbar M. African Americans' perceptions of psychotherapy and psychotherapists. Prof Psychol Res Pract 2004;35(1):19-26. [doi: 10.1037/0735-7028.35.1.19]
26. Kim J, Muhic J, Robert LP, Park SY. Designing chatbots with black americans with chronic conditions: overcoming challenges against COVID-19. Presented at: CHI '22; Apr 30 to May 5, 2022; New Orleans, LA, USA. [doi: 10.1145/3491102.3502116]
27. Harrington CN, Egede L. Trust, comfort and relatability: understanding black older adults' perceptions of chatbot design for health information seeking. Presented at: CHI '23; Apr 23-28, 2023; Hamburg, Germany. [doi: 10.1145/3544548.3580719]
28. Harrington CN, Garg R, Woodward A, Williams D. It's kind of like code-switching": black older adults' experiences with a voice assistant for health information seeking. Proc SIGCHI Conf Hum Factor Comput Syst 2022 Apr;2022:1-15. [doi: 10.1145/3491102.3501995] [Medline: 35876765]
29. Oguamanam VO, Hernandez N, Chandler R, et al. An intersectional look at use of and satisfaction with digital mental health platforms: a survey of perinatal black women. Presented at: CHI '23; Apr 23-28, 2023; Hamburg Germany. [doi: 10.1145/3544548.3581475]
30. Hall AK, Cole-Lewis H, Bernhardt JM. Mobile text messaging for health: a systematic review of reviews. Annu Rev Public Health 2015 Mar 18;36(393–415):393-415. [doi: 10.1146/annurev-publhealth-031914-122855] [Medline: 25785892]
31. Mobile fact sheet. Pew Research Center. URL: https://www.pewresearch.org/internet/fact-sheet/mobile/ [accessed 2024-02-26]
32. Heron KE, Romano KA, Braitman AL. Mobile technology use and mHealth text message preferences: an examination of gender, racial, and ethnic differences among emerging adult college students. Mhealth 2019;5:2. [doi: 10.21037/mhealth.2019.01.01] [Medline: 30842950]
33. Smith A. How Americans use text messaging. Pew Research Center. 2011. URL: https://www.pewresearch.org/internet/2011/09/19/how-americans-use-text-messaging/ [accessed 2025-02-13]
34. Schoenthaler A, Leon M, Butler M, Steinhaeuser K, Wardzinski W. Development and evaluation of a tailored mobile health intervention to improve medication adherence in black patients with uncontrolled hypertension and type 2 diabetes: pilot randomized feasibility trial. JMIR Mhealth Uhealth 2020 Sep 23;8(9):e17135. [doi: 10.2196/17135] [Medline: 32965230]

XSL•FO
RenderX

35. Berrouiguet S, Baca-García E, Brandt S, Walter M, Courtet P. Fundamentals for future mobile-health (mHealth): a systematic review of mobile phone and web-based text messaging in mental health. J Med Internet Res 2016 Jun 10;18(6):e135. [doi: 10.2196/jmir.5066] [Medline: 27287668]
36. Rathbone AL, Prescott J. The use of mobile apps and SMS messaging as physical and mental health interventions: systematic review. J Med Internet Res 2017 Aug 24;19(8):e295. [doi: 10.2196/jmir.7740] [Medline: 28838887]
37. Meyerhoff J, Popowski SA, Lakhtakia T, et al. Automated digital safety planning interventions for young adults: qualitative study using online co-design methods. JMIR Form Res 2025 Feb 26;9(1):e69602. [doi: 10.2196/69602] [Medline: 40009840]
38. Bhattacharjee A, Williams JJ, Meyerhoff J, Kumar H, Mariakakis A, Kornfield R. Investigating the role of context in the delivery of text messages for supporting psychological wellbeing. Presented at: CHI '23; Apr 23-28, 2023; Hamburg, Germany. [doi: 10.1145/3544548.3580774]
39. Kroenke K, Spitzer RL, Williams JBW. The PHQ-9. J Gen Intern Med 2001 Sep;16(9):606-613. [doi: 10.1046/j.1525-1497.2001.016009606.x]
40. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
41. Maestre JF, MacLeod H, Connelly CL, et al. Defining through expansion: conducting asynchronous remote communities (ARC) research with stigmatized groups. Presented at: CHI '18: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems; Apr 21-26, 2018; Montreal, QC, Canada. [doi: 10.1145/3173574.3174131]
42. Free online focus group software. Focus Group It. URL: https://focusgroupit.com/ [accessed 2025-08-29]
43. Meyerhoff J, Beltzer M, Popowski S, et al. Small Steps over time: a longitudinal usability test of an automated interactive text messaging intervention to support self-management of depression and anxiety symptoms. J Affect Disord 2024 Jan 15;345:122-130. [doi: 10.1016/j.jad.2023.10.119] [Medline: 37866736]
44. Kornfield R, Stamatis CA, Bhattacharjee A, et al. A text messaging intervention to support the mental health of young adults: User engagement and feedback from a field trial of an intervention prototype. Internet Interv 2023 Dec;34:100667. [doi: 10.1016/j.invent.2023.100667] [Medline: 37746639]
45. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
46. Stiles WB. "I have to talk to somebody": a fever model of disclosure. In: Derlega VJ, Berg JH, editors. Self-Disclosure. Perspectives in Social Psychology.: Springer; 1987:257-282. [doi: 10.1007/978-1-4899-3523-6_12]
47. Illouz E. Saving the Modern Soul: Therapy, Emotions, and the Culture of Self-Help: University of California Press; 2008.
48. Pennebaker JW. Opening Up: The Healing Power of Expressing Emotions: Guilford Press; 2012.
49. Baikie KA, Wilhelm K. Emotional and physical health benefits of expressive writing. Adv psychiatr treat 2005 Sep;11(5):338-346. [doi: 10.1192/apt.11.5.338]
50. Comas-Díaz L, Hall GN, Neville HA. Racial trauma: Theory, research, and healing: Introduction to the special issue. Am Psychol 2019 Jan;74(1):1-5. [doi: 10.1037/amp0000442] [Medline: 30652895]
51. Lutgendorf SK, Ullrich P. Cognitive processing, disclosure, and health: psychological and physiological mechanisms. In: The Writing Cure: How Expressive Writing Promotes Health and Emotional Well-Being: American Psychological Association; 2002:177-196. [doi: 10.1037/10451-009]
52. Burke M, Develin M. Once more with feeling: supportive responses to social sharing on Facebook. Presented at: CSCW '16; Feb 27 to Mar 2, 2016; San Francisco, California, USA. [doi: 10.1145/2818048.2835199]
53. Rimé B. Emotion elicits the social sharing of emotion: theory and empirical review. Emot Rev 2009 Jan;1(1):60-85. [doi: 10.1177/1754073908097189]
54. Rimé B. Mental rumination, social sharing, and the recovery from emotional exposure. In: Emotion, Disclosure, and Health: American Psychological Association; 1995:271-291. [doi: 10.1037/10182-013]
55. Derlega VJ, Metts S, Petronio S, Self-disclosure MST. Self-Disclosure: Sage Publications; 1993.
56. Kim HS, Sherman DK, Taylor SE. Culture and social support. Am Psychol 2008 Sep;63(6):518-526. [doi: 10.1037/0003-066X] [Medline: 18793039]
57. Omarzu J. A disclosure decision model: determining how and when individuals will self-disclose. Pers Soc Psychol Rev 2000 May;4(2):174-185. [doi: 10.1207/S15327957PSPR0402_05]
58. Pachankis JE. The psychological implications of concealing a stigma: a cognitive-affective-behavioral model. Psychol Bull 2007 Mar;133(2):328-345. [doi: 10.1037/0033-2909.133.2.328] [Medline: 17338603]
59. Pyle C, Roosevelt L, Lacombe-Duncan A, Andalibi N. LGBTQ persons' pregnancy loss disclosures to known ties on social media: disclosure decisions and ideal disclosure environments. Presented at: CHI '21: CHI Conference on Human Factors in Computing Systems; May 8-13, 2021; Yokohama, Japan. [doi: 10.1145/3411764.3445331]
60. McDonald J, Harris KL, Ramirez J. Revealing and concealing difference: a critical approach to disclosure and an intersectional theory of "closeting". Commun Theory 2020 Feb 1;30(1):84-104. [doi: 10.1093/ct/qtz017]
61. Pinch A, Birnholtz J, Chaudary J, et al. Queerness and mental health in India: an intersectional approach to sensitive social media disclosures. Social Media + Society 2024 Oct;10(4):20563051241302145. [doi: 10.1177/20563051241302145]

XSL•FO
RenderX

62. Crenshaw K. Demarginalizing the intersection of race and sex: a black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. In: University of Chicago Legal Forum 1989. URL: https://scholarship.law.columbia.edu/faculty_scholarship/3007 [accessed 2025-08-29]
63. McCall T, Schwartz TA, Khairat S. The acceptability of text messaging to help African American women manage anxiety and depression: cross-sectional survey study. JMIR Ment Health 2020 Feb 3;7(2):e15801. [doi: 10.2196/15801] [Medline: 31909720]
64. Kornfield R, Mohr DC, Ranney R, et al. Involving crowdworkers with lived experience in content-development for push-based digital mental health tools: lessons learned from crowdsourcing mental health messages. Proc ACM Hum Comput Interact 2022 Apr;6(CSCW1):1-30. [doi: 10.1145/3512946] [Medline: 35529806]
65. Kornfield R, Meyerhoff J, Studd H, et al. Meeting users where they are: user-centered design of an automated text messaging tool to support the mental health of young adults. Presented at: CHI '22; Apr 29 to May 5, 2022; New Orleans, LA, USA. [doi: 10.1145/3491102.3502046]
66. Bhattacharjee A, Williams JJ, Chou K, et al. "I Kind of Bounce off It": translating mental health principles into real life through story-based text messages. Proc ACM Hum Comput Interact 2022 Nov;6(CSCW2):1-31. [doi: 10.1145/3555123] [Medline: 36816014]
67. Bhattacharya A, Nagar R, Jenness J, Munson SA, Kientz JA. Designing asynchronous remote support for behavioral activation in teenagers with depression: formative study. JMIR Form Res 2021 Jul 13;5(7):e20969. [doi: 10.2196/20969] [Medline: 34255665]
68. Walther JB. Computer-mediated communication: impersonal, interpersonal, and hyperpersonal interaction. Commun Res 1996;23:3-43. [doi: 10.1177/009365096023001001]
69. Tanis M. Health-related on-line forums: what's the big attraction? J Health Commun 2008;13(7):698-714. [doi: 10.1080/10810730802415316] [Medline: 18958781]
70. Chung JE. Social interaction in online support groups: Preference for online social interaction over offline social interaction. Comput Human Behav 2013 Jul;29(4):1408-1414. [doi: 10.1016/j.chb.2013.01.019]
71. Wadden D, August T, Li Q, Althoff T. The effect of moderation on online mental health conversations. ICWSM 2021 May 22;15:751-763. [doi: 10.1609/icwsm.v15i1.18100]
72. Marriott BR, Lewis CC, Gobin RL. Disclosing traumatic experiences: Correlates, context, and consequences. Psychol Trauma 2016 Mar;8(2):141-148. [doi: 10.1037/tra0000058] [Medline: 26010111]
73. Lattie EG, Stiles-Shields C, Graham AK. An overview of and recommendations for more accessible digital mental health services. Nat Rev Psychol 2022 Feb;1(2):87-100. [doi: 10.1038/s44159-021-00003-1] [Medline: 38515434]
74. Mohr DC, Azocar F, Bertagnolli A, et al. Banbury forum consensus statement on the path forward for digital mental health treatment. PS (Wash DC) 2021 Jun;72(6):677-683. [doi: 10.1176/appi.ps.202000561]
75. Iqbal W, Arshad MH, Tyson G, Castro I. Exploring crowdsourced content moderation through lens of Reddit during COVID-19. Presented at: AINTEC'22: The 17th Asian Internet Engineering Conference; Dec 19-21, 2022; Hiroshima, Japan. [doi: 10.1145/3570748.3570753]
76. Chandrasekharan E, Gandhi C, Mustelier MW, Gilbert E. Crossmod: a cross-community learning-based system to assist Reddit moderators. Proc ACM Hum-Comput Interact 2019 Nov 7;3(CSCW):1-30. [doi: 10.1145/3359276]
77. Kim S, Eun J, Oh C, Suh B, Lee J. Bot in the bunch: facilitating group chat discussion by improving efficiency and participation with a chatbot. Presented at: CHI '20; Apr 25-30, 2020; Honolulu, HI, USA. [doi: 10.1145/3313831.3376785]
78. Jhaver S, Birman I, Gilbert E, Bruckman A. Human-machine collaboration for content regulation: the case of Reddit automoderator. ACM Trans Comput-Hum Interact 2019 Oct 31;26(5):1-35. [doi: 10.1145/3338243]
79. He Q, Hong Y, Raghu TS. Platform governance with algorithm-based content moderation: an empirical study on Reddit. Information Systems Research 2025 Jun;36(2):1078-1095. [doi: 10.1287/isre.2021.0036]
80. Nelson A. Unequal treatment: confronting racial and ethnic disparities in health care. J Natl Med Assoc 2002 Aug;94(8):666-668. [Medline: 12152921]
81. Williams DR. Stress and the mental health of populations of color: advancing our understanding of race-related stressors. J Health Soc Behav 2018 Dec;59(4):466-485. [doi: 10.1177/0022146518814251] [Medline: 30484715]
82. Paradies Y, Ben J, Denson N, et al. Racism as a determinant of health: a systematic review and meta-analysis. In: Hills RK, editor. PLOS ONE 2015;10(9):e0138511. [doi: 10.1371/journal.pone.0138511] [Medline: 26398658]
83. Goosby BJ, Cheadle JE, Mitchell C. Stress-related biosocial mechanisms of discrimination and African American health inequities. Annu Rev Sociol 2018 Jul;44(1):319-340. [doi: 10.1146/annurev-soc-060116-053403] [Medline: 38078066]
84. Eichstaedt JC, Sherman GT, Giorgi S, et al. The emotional and mental health impact of the murder of George Floyd on the US population. Proc Natl Acad Sci U S A 2021 Sep 28;118(39):e2109139118. [doi: 10.1073/pnas.2109139118] [Medline: 34544875]
85. Brown KL, Graham AK, Perera RA, LaRose JG. Eating to cope: Advancing our understanding of the effects of exposure to racial discrimination on maladaptive eating behaviors. Int J Eat Disord 2022 Dec;55(12):1744-1752. [doi: 10.1002/eat.23820] [Medline: 36205358]
86. Sefidgar YS, Seo W, Kuehn KS, et al. Passively-sensed behavioral correlates of discrimination events in college students. Proc ACM Hum Comput Interact 2019 Nov;3(CSCW):1-29. [doi: 10.1145/3359216] [Medline: 34322658]

XSL•FO
RenderX

87. Buis L, Hirzel L, Dawood RM, et al. Text messaging to improve hypertension medication adherence in African Americans from primary care and emergency department settings: results from two randomized feasibility studies. JMIR Mhealth Uhealth 2017 Feb 1;5(2):e9. [doi: 10.2196/mhealth.6630] [Medline: 28148474]
88. Soubutts E, Shrestha P, Davidson BI, et al. Challenges and opportunities for the design of inclusive digital mental health tools: understanding culturally diverse young people's experiences. Presented at: CHI '24; May 11-16, 2024; Honolulu, HI, USA. [doi: 10.1145/3613904.3642641]
89. Ybarra ML, Prescott T, Mustanski B, Parsons J, Bull SS. Feasibility, acceptability, and process indicators for Guy2Guy, an mHealth HIV prevention program for sexual minority adolescent boys. J Adolesc Health 2019 Sep;65(3):417-422. [doi: 10.1016/j.jadohealth.2019.04.025] [Medline: 31277991]
90. Kaye LK, Quinn S. Psychosocial outcomes associated with engagement with online chat systems. Int J Hum Comput Interact 2020 Jan 20;36(2):190-198. [doi: 10.1080/10447318.2019.1620524]
91. Duong HT, Hopfer S. "Let's Chat": process evaluation of an intergenerational group chat intervention to increase cancer prevention screening among Vietnamese American families. Transl Behav Med 2021 Apr 7;11(3):891-900. [doi: 10.1093/tbm/ibaa120] [Medline: 33290557]
92. Scholl J, Kohls E, Görges F, et al. Acceptability and feasibility of the transfer of face-to-face group therapy to online group chats in a psychiatric outpatient setting during the COVID-19 pandemic: longitudinal observational study. JMIR Form Res 2021 Jul 23;5(7):e27865. [doi: 10.2196/27865] [Medline: 34161252]
93. Meyerhoff J, Kornfield R, Mohr DC, Reddy M. Meeting young adults' social support needs across the health behavior change journey: implications for digital mental health tools. Proc ACM Hum Comput Interact 2022 Nov;6(CSCW2):1-33. [doi: 10.1145/3555203] [Medline: 36387059]
94. Kruzan KP, Meyerhoff J, Nguyen T, Reddy M, Mohr DC, Kornfield R. "I Wanted to See How Bad it Was": online self-screening as a critical transition point among young adults with common mental health conditions. Presented at: CHI '22; Apr 29 to May 5, 2022; New Orleans, LA, USA. [doi: 10.1145/3491102.3501976]
95. Shah J, DePietro B, D'Adamo L, et al. Development and usability testing of a chatbot to promote mental health services use among individuals with eating disorders following screening. Int J Eat Disord 2022 Sep;55(9):1229-1244. [doi: 10.1002/eat.23798] [Medline: 36056648]
96. Joyce AS, Piper WE, Ogrodniczuk JS, Klein RH. Structure of therapy and termination. In: Termination in Psychotherapy: A Psychodynamic Model of Processes and Outcomes: American Psychological Association. URL: https://psycnet.apa.org/record/2006-21531-005 [doi: 10.1037/11545-005]
97. Beltzer ML, Meyerhoff J, Popowski SA, Mohr DC, Kornfield R. Mental health self-tracking preferences of young adults with depression and anxiety not engaged in treatment: qualitative analysis. JMIR Form Res 2023 Oct 6;7(1):e48152. [doi: 10.2196/48152] [Medline: 37801349]
98. Williams KDA, Dougherty SE, Utsey SO, LaRose JG, Carlyle KE. "Could Be Even Worse in College": social factors, anxiety, and depressive symptoms among Black men on a college campus. J Racial and Ethnic Health Disparities 2023 Jun;10(3):1165-1177. [doi: 10.1007/s40615-022-01302-w]

Edited by A Mavragani; submitted 03.03.25; peer-reviewed by D Frost, P Watson; revised version received 01.08.25; accepted 06.08.25; published 06.10.25.

Please cite as:
Popowski SA, Meyerhoff J, Allen OM, Nguyen T, McCall T, Pederson AB, Reddy M, Mohr D, Kornfield R
Designing Digital Mental Health Tools to Support the Needs of Black Adults in the United States: Qualitative Analysis
JMIR Form Res 2025;9:e73279
URL: https://formative.jmir.org/2025/1/e73279
doi:10.2196/73279

© Sarah Alexandra Popowski, Jonah Meyerhoff, Olivia Marin Allen, Theresa Nguyen, Terika McCall, Aderonke Bamgbose Pederson, Madhu Reddy, David Mohr, Rachel Kornfield. Originally published in JMIR Formative Research (https://formative.jmir.org), 6.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Impact of Dose in an mHealth Intervention to Support Parents and Carers Via Healthy Beginnings for Hunter New England Kids Program: Pragmatic Randomized Controlled Trial

Alison L Brown[1,2,3,4], BNutrDiet (Hons), PhD; Nayerra Hudson[1,2,3], BNutrDiet (Hons); Jessica Pinfold[1], BNutrDiet; Rebecca Sewter[1], BNutrDiet; Lynda Davies[1,3], MPH, MNutrDiet; Christophe Lecathelinais[1,3], DESS; Jacklyn K Jackson[1,2,3], PhD; Tessa Delaney[5], PhD; Sienna Kavalec[1,2], BNutrDiet; Rachel Sutherland[1,2,3,4], PhD

[1]Population Health, Hunter New England Local Health District, Booth Building, Longworth Avenue, Newcastle, New South Wales, Australia

[2]School of Medicine and Public Health, University of Newcastle, University Drive, Callaghan, Australia

[3]Hunter Medical Research Institute, Kookaburra Circuit, Newcastle, Australia

[4]National Centre of Implementation Science, Newcastle, Australia

[5]Centre for Population Health, New South Wales Ministry of Health, St Leonards, Australia

Corresponding Author:
Alison L Brown, BNutrDiet (Hons), PhD
Population Health, Hunter New England Local Health District, Booth Building, Longworth Avenue, Newcastle, New South Wales, Australia

Abstract

Background: The dose of mobile health (mHealth) interventions can influence participant engagement, acceptability, and overall impact. However, few mHealth interventions have explored this dose-response relationship.

Objective: This study aims to explore how dose influences the acceptability, engagement, cost, and impact on infant feeding status of a parent-targeted mHealth text messaging program which aims to enhance child health, including breastfeeding exclusivity and duration.

Methods: This pragmatic randomized controlled trial was conducted from October 2021 to May 2024. The Healthy Beginnings for Hunter New England Kids (HB4HNEKids) program provides- text messages aimed to support parents and carers and their children by providing evidence-based preventive health information across the first 2000 days. Participants were enrolled in HB4HNEKids from 5 Child and Family Health Services in the Hunter New England region of New South Wales, Australia, and randomized into either a high-dose or low-dose text message group for the first 2 years of the pilot program. Dose refers to the quantity and frequency of text messages sent to participants. Participants in the high-dose text message group received an average of 111 - 121 text messages, and the low-dose text message group received 80 - 82 text messages across the 2 years. Outcomes of interest included acceptability, engagement, cost, and infant feeding status in relation to dose. Engagement with the messages was determined using click rates and program opt-out rates. Participant acceptability was assessed via a brief survey. Impact on infant feeding status (ie, breastfeeding, formula feeding, or mixed feeding) was determined by participants reporting their feeding status at several time points across the program. Cost was determined by assessing the per participant and total cost of sending text messages for each dose group across the 2-year period.

Results: There were no statistically significant differences in click rates between high or low-dose text message groups. In the first 6 months, significantly more participants opted out of the high-dose text message group (191/2724; 7%) compared to the low-dose (108/2812; 3.8%; $P<.001$). In terms of program acceptability, 183 out of 214 (85.5%) participants of the high-dose and 228 out of 252 (90.5%) participants of the low-dose text message group were satisfied with the frequency of text messages. In addition, 188 out of 215 (87%) participants of high-dose and 220 out of 255 (86%) participants of low-dose text message group indicated they would recommend the program to other caregivers. The average per participant and total cost to the health service for sending messages was lower in the low-dose group (A$9.32 per participant and A$15,271.48 total; A$1 is approximately equal to US $0.68) compared to the high-dose text message group (A$12.96 per participant and A$21,241.44 total).

Conclusions: The HB4HNEKids program demonstrated positive outcomes including high acceptability across both groups and no impact on infant feeding status, irrespective of dose. Given the higher opt-out rates and message costs in the high-dose text message group, a lower dose is likely more scalable for future use.

Trial Registration: Australian New Zealand Clinical Trial Registry ACTRN12625000574448; https://anzctr.org.au/Trial/Registration/TrialReview.aspx?id=389640&isReview=true

XSL•FO
RenderX

(JMIR Form Res 2025;9:e70158) doi:10.2196/70158

KEYWORDS

child health; infants; mHealth; mobile health; digital health; text messaging; dose; engagement

Introduction

The first 2000 days of life (conception to 5 years) is a key period for child physical, cognitive, social, and emotional development and can have significant lifelong impacts [1]. Therefore, the first 2000 days is an important period to enhance a child's overall health. Accordingly, New South Wales (NSW) Health has developed a strategic policy to support families in the first 2000 days, stipulating the provision of care to all [1]. Currently, NSW Child and Family Health Services (CFHS) offer families routine health and development checks over their child's first 2000 days, via face-to-face home or clinic visits which target key age and stage related milestones. However, engagement with the health service by families is fragmented. Key factors that impact the provision of care by health services include limited staff capacity, while parent uptake of services is influenced by the need for convenient and accessible services that align with family routines and roles [2].

Mobile health (mHealth) is considered a promising service delivery model to support families during the first 2000 days [3,4]. A systematic review of mHealth interventions delivered via text messages or telephone to prevent early childhood obesity found that over 60% of the studies reported improvements in 1 or more key health-related behaviors, including breastfeeding, introduction of solids, physical activity, screen time, sleep, and tummy time [5]. This highlights the delivery of mHealth interventions as a promising population-level approach. mHealth text message-based interventions have the ability to provide evidence-based information and additional care pathways at scale, while being cost-effective [5-7]. mHealth infrastructure is currently being used by CFHS in the scheduling of appointments; however, it is not routinely used to support service delivery activities, including providing evidence-based information to parents and carers. As such, an mHealth program, Healthy Beginnings for Hunter New England Kids (HB4HNEKids), was developed that aims to support parents and carers with the provision of evidence-based age and stage-appropriate care, in the form of text messages, across the first 2000 days. The program was piloted across a 3-year period (October 2021-July 2024) to test the feasibility of program methods and test any adjustments required to improve the quality of the program. The program is now offered as part of usual care for CFHS and is currently being rolled out throughout the Hunter New England (HNE) region of NSW, Australia. However, as part of the pilot and ongoing rollout of HB4HNEKids, routine quality assurance data has been collected to monitor and improve program reach, engagement, and overall health outcomes.

Dose is an important aspect of any mHealth intervention, particularly long-term programs such as HB4HNEKids. The dose of a program can have a direct effect on participant engagement, and therefore the effectiveness of an intervention [8]. However, there are limited mHealth studies that have examined the dose-response relationship as part of program evaluation to assess its effect on acceptability, engagement, impact, and cost [9]. Furthermore, there is a scarcity of studies that have looked at the dose-response relationship in long-term studies [9]. Initial investigation into the acceptability of text message frequency for an intervention targeting preschool parents identified significant variability in what was an acceptable frequency of messages across the school year, with the study reporting 1 text message per week was too little and 5 text messages were excessive [10]. This highlights that dose can have a direct response on engagement with mHealth programs and must be a key consideration in implementation planning.

With minimal evidence regarding optimal dose in parent-targeted mHealth (text message) interventions, this study aims to explore how dose influences the acceptability, engagement, impact on infant feeding status, and cost of a parent-targeted mHealth text messaging program, HB4HNEKids.

Methods

Setting

This trial was a pragmatic randomized controlled trial. The program was offered as part of routine service delivery provided by CFHS in the HNE region of NSW, as part of a pilot feasibility trial. The HNE region covers 131,785 km^2 and has an estimated population of 973,653 people [11,12]. In 2020, there were 10,242 births within the HNE Local Health District, accounting for approximately 10% of births in NSW [13]. There are approximately 150 CFHS staff supporting parents and carers (hereafter referred to as parents) across HNE with the primary responsibility of comprehensively assessing and facilitating care pathways that best support the physical, developmental, psychological, and socio-economic health and well-being of families with children aged 0 - 5 years [14].

Participants and Recruitment

Child and Family Health Services

A sample of 5 CFHS in the HNE region was selected for inclusion in the pilot feasibility trial of HB4HNEKids. Recruitment of CFHS was identified through an expression of interest process, with services volunteering for study participation. To ensure representation across a variety of population groups and geographical areas, the research team sought to include CFHS that met at least 1 of each of the following eligibility criteria:

1. Aboriginal CFHS (n=2)
2. Rural or remote service CFHS (n=1; based on a population size of less than 50,000)
3. Regional CFHS (n=1; based on a population size between 50,000 - 100,000)

XSL•FO
RenderX

4. Metropolitan CFHS (n=1; based on a population size of more than 100,000)

Parents and Carers

As per existing health service procedures, CFHS receive a referral from public and private maternity services following the birth of a baby [14]. Upon discharge, staff from one of the 5 HNE CFHS contacted the parent (participants) via phone call or text message to offer a 1 - 4 week health and development check. Parents within the catchment area of the pilot CFHS sites were offered HB4HNEKids as part of routine care. During the initial contact to schedule families into a 1 - 4 week health and development check, CFHS staff confirmed feeding status (ie, if the child was breastfeeding, mixed feeding, formula feeding), child name, mother mobile number, and "offer of service" (ie, yes, client refused, or unable to contact after multiple attempts). If a staff member was unable to determine infant feeding status during the initial contact, they would locate the "infant feeding status" via the hospital discharge summary provided by maternity services. No other participant characteristics were collected during this study.

Healthy Beginnings for HNEKids Program

HB4HNEKids is a text message program (model of care) that targets barriers and enablers to recommended infant feeding practices, child nutrition, physical activity, small screen recreation, child development, parent well-being and primary health checks, and immunizations. One of the program's primary outcomes is to increase the duration of any breastfeeding, in alignment with international recommendations. The mHealth program was developed collaboratively with a multidisciplinary team including CFHS, multicultural health, Aboriginal partners, allied health professionals, and end-users (parents). The program has been adapted from the Healthy Beginnings CHAT study [15]. The optimized text messages used the behavior change wheel [16] and Theoretical Domains Framework [17] to develop message content and support link selection. The text messages aimed to provide brief content (mostly <160 characters) with clickable links to evidence-based "online" videos, fact sheets, websites, and relevant support services. Text message content is aligned to the age, stage, and developmental milestones of the child across the first 2000 days. Participants receive different text messages based on their reported infant feeding status (breastfeeding, mixed feeding, or formula feeding) that are tailored to support their child's needs. Example text messages can be found in Table 1. While the HB4HNEKids program aims to provide support across the first 2000 days, this study focuses specifically on the first 2 years. Outcome data on the effectiveness of HB4HNEKids will be reported elsewhere.

Table . Example text messages for Healthy Beginnings for Hunter New England Kids.

Days	Phases	Messages
Day 9	1 (day 3 to 6 months)	Ensuring bub is attached well helps reduce discomfort. Good attachment signs are mouth open wide, chin to breast, good suck & swallow.
Day 11	1 (day 3 to 6 months)	Breastfeeding takes practice. Give yourself time to try different positions & find what works best for you & #firstname#.
Day 15	1 (day 3 to 6 months)	There are simple ways to know if #firstname# is getting enough milk such as 5 - 6 wet nappies/day.
Day 47	1 (day 3 to 6 months)	Intending to express? Hand expressing breastmilk can be done anywhere. Practice at home so you're comfortable.
Day 192	2 (6 - 12 months)	There are 4 easy steps to healthy teeth & gums. Start brushing twice a day when #firstname# first teeth appear.
Day 334	2 (6-12 months)	Continuing to provide breastmilk beyond 12 months is great for you & #firstname#. If you're both enjoying it you don't have to stop, but if you are looking to introduce full cream milk you can from 12 months. For any questions contact your CFHN or GP.
Day 403	3 (1 - 2 years)	Managing your emotions is important so you can enjoy parenting & a happy home for you & your family.
Day 475	3 (1 - 2 years)	#firstname#does not need any screen time. Video chats with family & friends are ok. Try reading, singing, puzzles & craft instead.
Day 524	3 (1 - 2 years)	Remember, offering new foods many times (up to 15) gives #firstname# a chance to learn about their smell, touch & taste.

XSL•FO
RenderX

The program was broken up into four phases across the first 2 years of the program as follows: phase 0 (day 0-2; onboarding); phase 1 (day 3 to 6 months); phase 2 (6 - 12 months); and phase 3 (1 - 2 years).

Delivery of Messages

Participants were sent text messages to their nominated mobile number between the hours of 9 AM to 12 PM across the week according to the message schedule using the SMS platform, DirectSMS. Each text message unit (n<160 characters) had an associated cost to the health service of 13.5 cents (A$0.135; A$1 is approximately equal to US $0.68) per unit. Participants were provided a link to opt-out of the program if they wanted to on day 1, 51, 296, and 461.

Dose Randomization

Overview

Participants were randomized into one of 2 intervention groups (high or low dose text message group) in randomly sequenced blocks of 2 and 4 using a random number function set up by an independent statistician in SAS 9.3. As the intervention content varied by feeding status, the randomization was stratified by the feeding status (breastfed vs mixed fed vs formula fed) provided at the initial contact with CFHS. CFHS staff and participants were blinded to group allocation.

Low-Dose Text Message Group

The frequency and dose of text messages in the low-dose text message group were adapted from the "Healthy Beginnings" CHAT study [15]. The low-dose text message group received up to 2 messages per week in the first month and approximately 1 per week from 2 months to 2 years (Table 2).

Table . Number of messages sent by study phase, feeding status, and dose group.

Phases	Number of messages by group	
	High-dose	Low-dose
Phase 0 (0 - 2 days)		
Breastfeeding and mixed feeding	1	1
Formula feeding	1	1
Phase 1 (2 days to 6 months)		
Breastfeeding	40	24
Mixed feeding	38	23
Formula feeding	31	23
Phase 2[a] (6 - 12 months)		
Breastfeeding and mixed feeding	26	26
Formula feeding	25	25
Phase 3 (1 - 2 years)		
Breastfeeding and mixed feeding	54	31
Formula feeding	54	31
Total		
Breastfeeding	121	82
Mixed feeding	119	81
Formula feeding	111	80

[a]dose was not applied to phase 2 (6 - 12 months)

High-Dose Text Message Group

The high-dose text message group received up to 3 - 4 messages per week in the first month, 2 per week in the second month and 5 per month from 3 months to 2 years. The number of messages was determined by the number of identified barriers that were considered to impact behavioral outcomes.

Data Collection

Infant Feeding Status

As part of routine service delivery, participants were asked to respond to a message asking how they were feeding their child or children at days 21 (3 weeks), 97 (3 months), 169 (6 months), 275 (9 months), 360 (1 year), and 552 (18 months) to determine any change to feeding status. Participants who entered the study as formula feeding received the day 21 message prompt, and anyone flagged as formula feeding stopped receiving follow-up prompts. Possible feeding status responses included: breastmilk, mixed feeding, and formula. If the feeding status changed, the

XSL•FO
RenderX

message content was tailored according to their new feeding status. To calculate feeding rates and breastfeeding outcomes across the lifespan of the program, “assumed” feeding status was used whereby the participant’s feeding status was carried over to the next phase of the program unless participants provided an update to their feeding status.

Engagement

Engagement with the messages was determined using the click rates from weblinks provided in each message, in addition to opt-out rates from the program.

Click Rates

Weblinks provided in the text messages were given a unique weblink shortener. Each weblink shortener was coded with the associated feeding status and dose, to enable the reporting of click rates by feeding status and dose. Click rates were calculated based on the number of unique times a link had been clicked divided by the number of participants that had been sent the corresponding link.

Opt-Out Rates

Participants were sent a message at day 1, 51, 296, and 461 that contained an opt-out link that participants had to click and provide their mobile number in order to opt-out. The opt-out link remained live for the course of the program, so that participants could click on the link at any point in time if they chose to no longer receive the program. Participants could also opt out via contact details provided, or by notifying CFHS. Opt-out rates were calculated by feeding status and dose.

Acceptability

At day 169 (approx. 5 months), participants were provided a link to a brief online survey to determine acceptability of the program. Participants were asked the following on a 5-point Likert scale (strongly agree to strongly disagree): (1) I found the Healthy Beginnings for HNEKids program acceptable; (2) Since I started the program, I have been happy with how often I have received the text messages; and (3) I would recommend the HB4HNEKids program to other caregivers.

Cost

The average cost to the health service of sending 2 years of text messages per participant was collected for each group. This was calculated using the average number of text messages sent across the 3 infant feeding groups multiplied by the cost of sending a message via the SMS provider, DirectSMS. In addition, the total cost of sending the text messages was reported from the commencement of the program from October 2021 to May 2024, for each group.

Statistics

Data was analyzed using the statistical software package SAS (version 9.4; SAS Institute). Descriptive statistics were used to describe outcomes such as the sample, cost, and acceptability measures. An intention to treat approach was used in the analyses of all study outcomes, whereby analysis occurred according to the randomized treatment allocation (dose group). Between-group differences in percentages were examined using Chi-square tests for each categorical outcome. The comparisons of interest included feeding status versus dose after each message feeding prompt, click rates versus dose stratified by feeding status, and opt-out rates versus dose during each study phase. Analyses of the outcomes used a 2-tailed test with an alpha level of 5%.

Ethical Considerations

Ethical approval to undertake the study was obtained from the Hunter New England Human Research Ethics Committee (23/12/13/4.10) and the University of Newcastle Human Research Ethics Committee (H-2024‑0025). This research was conducted in compliance with informed consent guidelines and adhered to national law and regulations regarding the protection of personal information, privacy, and human rights. This trial has been retrospectively registered in Australian New Zealand Clinical Trial Registry (ACTRN12625000574448; registered June 3, 2025).

Results

Text Messages

From October 2021 to May 2024, 5783 participants were engaged in the program (Figure 1). In total, 307,726 messages were sent to participants with children between 0‑2 years old. Table 2 shows the number of messages sent to each parent, in each phase of the program by dose and feeding status.

XSL•FO
RenderX

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) flow diagram. BF: breastfeeding; MF: mixed feeding; FF: formula feeding.

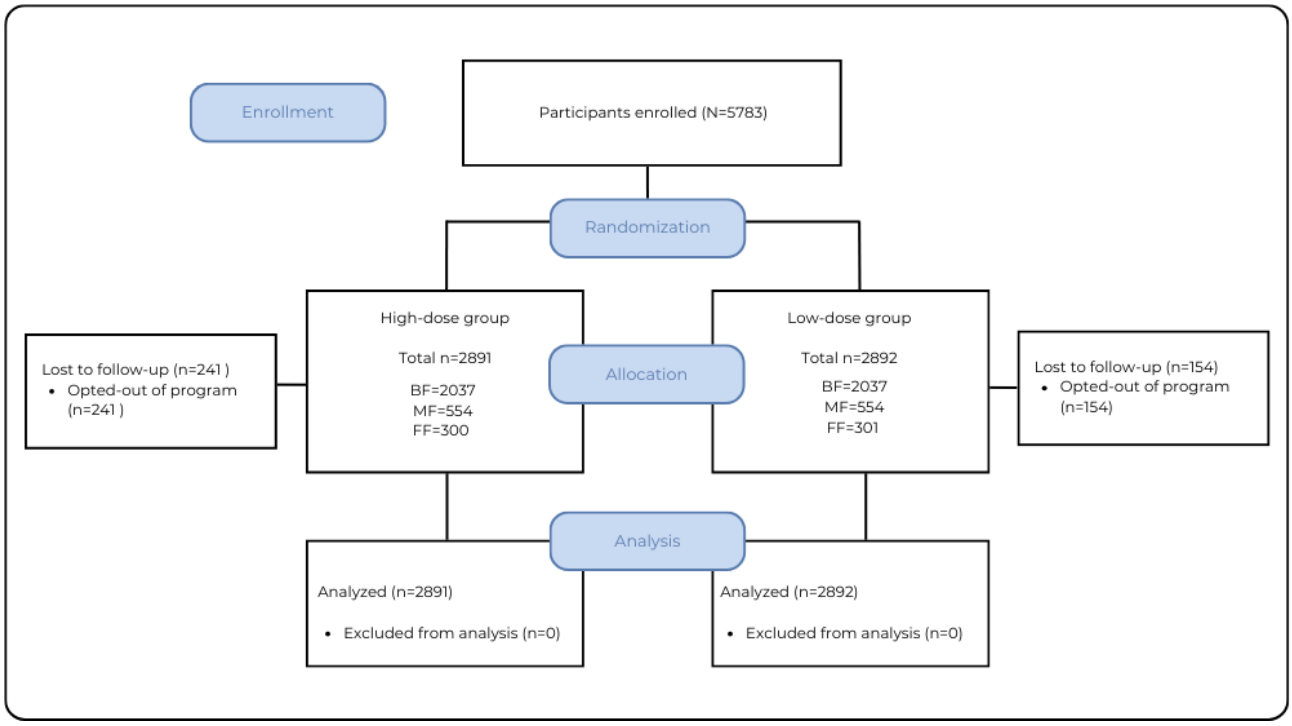

Feeding Status Responses

Of those that responded at least once with their feeding status, there were no statistically significant differences in feeding status between high and low dose text message groups, as seen in Table 3.

Table . Feeding status by dose.

Age of child (months) and feeding status	High-dose group, n (%)	Low-dose group, n (%)	Difference between groups (*P* value)
1 month			.74
Breastfeeding	1763 (64.8)	1760 (64.1)	
Mixed feeding	580 (21.3)	583 (21.3)	
Formula feeding	378 (13.9)	401 (14.6)	
3 months			.64
Breastfeeding	1280 (58.9)	1346 (59.4)	
Mixed feeding	414 (19.1)	447 (19.7)	
Formula feeding	478 (22)	474 (20.9)	
5 months			.29
Breastfeeding	1028 (54.5)	1061 (54.1)	
Mixed feeding	344 (18.2)	394 (20.1)	
Formula feeding	514 (27.3)	507 (25.8)	
9 months			.11
Breastfeeding	610 (46.6)	629 (46.2)	
Mixed feeding	219 (16.7)	267 (19.6)	
Formula feeding	481 (36.7)	465 (34.2)	
12 months			.10
Breastfeeding	502 (43.4)	510 (42.4)	
Mixed feeding	175 (15.1)	221 (18.4)	
Formula feeding	479 (41.4)	471 (39.2)	

XSL•FO
RenderX

Engagement

Click Rates

Table 4 outlines the click rates for the HB4HNEKids program. Overall, there was a total click rate on weblinks of 16.3% (47,369/290,771) across both doses. There were no statistically significant differences in click rates between high- or low-dose text message groups or via feeding status.

Table . Healthy Beginnings for Hunter New England Kids click rates by feeding status.

Feeding status	High-dose group, n (%)	Low-dose group, n (%)	Difference between groups (*P* value)
Breastfeeding	15,860 (16.2)	12,084 (17.8)	.90
Mixed feeding	5361 (17)	4519 (18.7)	.89
Formula feeding	5376 (13.4)	4169 (14.4)	.92
Total	26,597 (15.7)	20,772 (17.1)	.90

Opt-Out Rates

Table 5 describes the opt-out rates for HB4HNEKids by phase of the program. Overall, 395 participants out of 5783 (6.8%) opted out of HB4HNEKids. Of those in the high-dose text message group, significantly more participants (241/2891, 8.3%) opted out across all phases compared to the low-dose text message group (154/2892, 5.3%). In phase 1, significantly more participants opted out of the high-dose text message group (191/2724, 7.0%) compared to the low-dose text message group (108/2812, 3.8%; *P*<.001). There were no other statistically significant differences in opt-out rates by phase or dose. There were no significant differences in opt-out rates by feeding status.

Table . Healthy Beginnings for Hunter New England Kids opt-out rates by program phase.

Phases	High-dose group, n (%)	Low-dose group, n (%)	Differences between groups (*P* value)
Phase 0 (0 - 2 days)	8 (0.3)	6 (0.2)	.59
Phase 1 (2 days to 6 months)[a]	191 (7.0)	108 (3.8)	<.001
Phase 2 (6 - 12 months)	10 (0.6)	16 (0.9)	.28
Phase 3 (1 - 2 years)	32 (2.8)	24 (2.1)	.28
Total[a]	241 (8.3)	154 (5.3)	<.001

[a]Statistically significant difference in opt-out rates between high-dose and low-dose at *P*<.05.

Acceptability by Parents and Carers

Of those that received the acceptability online survey (n=1287), 479 participants completed the survey (37% response rate). When participants were asked about the acceptability of HB4HNEKids, 84% (181/215) of high-dose participants strongly agreed or agreed compared to 85% (218/255) in the low-dose. Of high-dose participants, 85.5% (183/214) strongly agreed or agree that they were happy with how often they received text messages compared to low-dose, whereby 90.5% (228/252) agreed or strongly agreed they were happy with the frequency of messages. Among the high-dose participants, 188 of 215 (87%) and 220 of 255 low-dose participants (86%) agreed or strongly agreed that they would recommend the program to other caregivers.

Cost

For high-dose, the average cost of sending text messages across all 3 feeding groups up to 2 years was A$12.9 per participant, whilst the average cost of sending messages for low-dose was A$9.32 per participant, making the cost difference per participant, per group A$3.64. The total cost of sending messages to families in the high-dose text message group was A$21,241.44, while the total equivalent cost for low-dose was A$15,271.48. Based on the number of participants, this is an estimated cost difference of A$5,965.96 between groups. Table 6 reports all costing of text messages per dose group.

Table . Healthy Beginnings for Hunter New England Kids text message cost per participant by dose.

Feeding groups	Cost per high-dose participant (A$)[a]	Cost per low-dose participant (A$)	Cost difference per group (A$)
Breastfeeding	13.50	9.45	4.05
Mixed feeding	13.23	9.32	3.91
Formula feeding	12.15	9.18	2.97
Average cost	12.96	9.32	3.64

[a]A$1 is approximately equal to US $0.68.

XSL•FO
RenderX

Discussion

Principal Findings

HB4HNEKids is an innovative mHealth program designed to provide theoretically developed and evidence-based age and stage appropriate information to parents in the first 2000 days. The findings of this study demonstrate the impact intervention dose has on acceptability, engagement, cost, and infant feeding outcomes in an mHealth text message–based program delivered to parents during the first 2000 days. Overall, infant feeding status and click rates did not differ by dose, with both dose groups showing high acceptability of the program. However, more participants opted out of the program from the high-dose text message group.

With feeding status not significantly changing by dose, this study demonstrated that a higher dose may not necessarily improve breastfeeding rates, which is one of the primary outcomes of the program. Click rates have been used in this study as a measure of engagement with the text messages. Of the total number of messages sent, 16% of links were clicked. With no statistically significant differences in click rates between high-dose and low-dose text message groups, it may be surmised that dose did not impact a participant's decision to engage with the online content. The click rate for HB4HNEKids is higher than reported in other mHealth interventions of a similar nature. For example, an mHealth intervention on infant feeding and growth reported an open rate of 8% of all notifications sent, which was notably lower than HB4HNEKids [18]. However, industry rates consider the average click rate to be between 20% and 35%, which is higher than the click rates for HB4HNEKids [19]. Few mHealth programs comprehensively evaluate engagement and click rates or consider it in relation to its primary outcomes [20], making it difficult to ascertain the engagement of HB4HNEKids in comparison.

Opt-out rates were highest in HB4HNEKids in the first 6 months of the program. The first 6 months are the most intensive, with participants initially receiving 2 messages weekly for low-dose and 3 - 4 messages for high. Subsequently, there was the highest level of opt-out rates in this age group, which may have been due to the intensity of the program during this phase. This pattern has also been reported in other mHealth interventions with a study on integrated health care systems reporting that those receiving more text messages were significantly more likely to opt out [21]. Acceptability of the program was considered high in both high-dose and low-dose text message groups, with the majority of participants reporting the program was acceptable and they would recommend the program to others. Interestingly, there were more participants in the low-dose text message group that considered the frequency of messages acceptable compared to those in the high-dose text message group, which needs to be considered in the context of the other study findings. When developing future content adaptations for HB4HNEKids, further consideration may need to be given if implementing higher dose messaging, in an effort to reduce opt-out rates, minimize costs, and improve acceptability of the frequency of messages.

It is important to consider this program of work in the context of its limitations. First, click rates of links within messages were used to measure engagement. Ideally, the use of open rates of text messages would be a more useful measure of engagement; however, this is a current limitation of SMS platforms. The use of web analytics may be important to explore in the future, to capture not only discrete open or click rates but to provide richer data, such as time spent on webpages, when information was accessed and whether there was repeated access. In the future, it may be useful to explore using a range of measures via a mixed-method approach to better understand the relationship between engagement and HB4HNEKids [22]. This study was conducted in the HNE region of NSW, which has an integrated public health system and experience delivering mHealth programs. These characteristics may have facilitated the implementation and engagement with the intervention. However, differences in health service organization, digital infrastructure, and population characteristics in other settings may influence the feasibility or effectiveness of similar interventions elsewhere. These contextual factors should be considered when applying the findings beyond this setting.

With infant feeding status and click rates being not significantly different based on dose and acceptability being high in both dose text message groups, the HB4HNEKids program has shown positive process outcomes, irrespective of dose. Considering opt-out rates and cost of sending messages was higher in the high-dose text message group and previous research supports the move to lowering the dose of mHealth programs long term [23], it is recommended that a lower dose of text messages may be more appropriate to use in further phases of the HB4HNEKids program as it expands across the 2 - 5 year age group, and must be considered when reviewing future iterations of the program.

Conclusions

With infant feeding status and click rates not significantly different based on dose and acceptability being high in both dose text message groups, the HB4HNEKids program has shown positive process outcomes, irrespective of dose. Considering opt-out rates and cost of sending messages was higher in the high-dose text message group, and previous research supports the move to lowering the dose of mHealth programs long term [23], it is recommended that a lower dose of text messages may be more appropriate to use in further phases of the HB4HNEKids program as it expands across the 2 - 5 year age group and must be considered when reviewing future iterations of the program.

XSL•FO
RenderX

Acknowledgments

We would like to acknowledge the Child and Family Health Services staff who contributed to enrolling participants into the program. This work was supported by the funding of Glencore Community Investment Grants (TPDD13). R Sutherland is supported by a Medical Research Future Fund Fellowship (APP1150661). The contents of this manuscript are the responsibility of authors and do not reflect the views of Glencore or Medical Research Future Fund.

Data Availability

The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

Authors' Contributions

R Sutherland, NH, and TD contributed to the concept and design. ALB, NH, R Sewter, LD, and CL had full access to the data in the study and took responsibility for the integrity of the data and accuracy of the data analysis. ALB, NH, JP, R Sewter, LD, CL, JKJ, TD, SK, and R Sutherland contributed to the acquisition, analysis, and interpretation of the data. ALB, NH, JKJ, SK, and R Sutherland contributed to the drafting of the manuscript. ALB and CL conducted the statistical analysis.

Conflicts of Interest

None declared.

Editorial Notice

This randomized study was initially undertaken as part of a routine health system quality assurance project; as such, this study was retrospectively registered. The editor granted an exception from the International Committee of Medical Journal Editors rules mandating prospective registration of randomized controlled trials, because the risk of bias appears low and the study was considered formative, guiding the development of the Healthy Beginnings for Hunter New England Kids program. However, readers are advised to carefully assess the validity of any potential explicit or implicit claims related to primary outcomes or effectiveness, as retrospective registration does not prevent authors from changing their outcome measures retrospectively.

Checklist 1
CONSORT 2025 checklist.
[PDF File, 150 KB - formative_v9i1e70158_app1.pdf]

References

1. The First 2000 Days Framework. : NSW Health; 2019 URL: https://www1.health.nsw.gov.au/pds/ActivePDSDocuments/PD2019_008.pdf [accessed 2025-09-05]
2. Love P, Laws R, Litterbach E, Campbell KJ. Factors influencing parental engagement in an early childhood obesity prevention program implemented at scale: the infant program. Nutrients 2018 Apr 19;10(4):509. [doi: 10.3390/nu10040509] [Medline: 29671815]
3. Long-term effects of breastfeeding compared with alternative food sources: a systematic review [Bachelor's thesis]. : Williams Honors College; 2019 URL: https://ideaexchange.uakron.edu/honors_research_projects/884/ [accessed 2025-09-14]
4. Manohar N, Hayen A, Bhole S, Arora A. Predictors of early introduction of core and discretionary foods in Australian infants-results from HSHK birth cohort study. Nutrients 2020 Jan 19;12(1):258. [doi: 10.3390/nu12010258] [Medline: 31963864]
5. Ekambareshwar M, Ekambareshwar S, Mihrshahi S, et al. Process evaluations of early childhood obesity prevention interventions delivered via telephone or text messages: a systematic review. Int J Behav Nutr Phys Act 2021 Jan 9;18(1):10. [doi: 10.1186/s12966-020-01074-8] [Medline: 33422066]
6. Gallegos D, Russell-Bennett R, Previte J, Parkinson J. Can a text message a week improve breastfeeding? BMC Pregnancy Childbirth 2014 Nov 6;14(1):374. [doi: 10.1186/s12884-014-0374-2] [Medline: 25369808]
7. Head KJ, Noar SM, Iannarino NT, Grant Harrington N. Efficacy of text messaging-based interventions for health promotion: a meta-analysis. Soc Sci Med 2013 Nov;97:41-48. [doi: 10.1016/j.socscimed.2013.08.003] [Medline: 24161087]
8. McVay MA, Bennett GG, Steinberg D, Voils CI. Dose-response research in digital health interventions: concepts, considerations, and challenges. Health Psychol 2019 Dec;38(12):1168-1174. [doi: 10.1037/hea0000805] [Medline: 31580127]
9. Li Y, Guo Y, Hong YA, et al. Dose-response effects of patient engagement on health outcomes in an mhealth intervention: secondary analysis of a randomized controlled trial. JMIR Mhealth Uhealth 2022 Jan 4;10(1):e25586. [doi: 10.2196/25586] [Medline: 34982724]
10. Cortes KE, Fricke H, Loeb S, Song DS. Too Little or Too Much? Actionable Advice in an Early Childhood Text Messaging Experiment: National Bureau of Economic Research; 2018. [doi: 10.3386/w24827]
11. Hunter New England. NSW Health. URL: https://www.health.nsw.gov.au/lhd/pages/hnelhd.aspx [accessed 2025-09-05]

XSL•FO
RenderX

12. Population estimates local health districts NSW. HealthStats NSW. 2024. URL: https://www.healthstats.nsw.gov.au/indicator-profiles/profile/LHD?view=Table&measure=Population&group=Age%20Yes&compare=Age%20Yes&dim1=LHD&dim2=New%20Entry%20LHD&dim3=Age%20Yes&dim4=Sex&dim5=Persons [accessed 2025-09-05]
13. Centre for Epidemiology and Evidence. New South Wales mothers and babies 2020. : NSW Ministry of Health; 2021 URL: https://www.health.nsw.gov.au/hsnsw/Publications/mothers-and-babies-2020.pdf [accessed 2025-09-05]
14. Child and family health nursing services model of care. NSW Health. 2021. URL: https://coilink.org/20.500.12592/5ou6cug [accessed 2025-09-05]
15. Wen LM, Rissel C, Baur LA, et al. A 3-Arm randomised controlled trial of communicating healthy beginnings advice by telephone (CHAT) to mothers with infants to prevent childhood obesity. BMC Public Health 2017 Jan 14;17(1):79. [doi: 10.1186/s12889-016-4005-x] [Medline: 28088203]
16. Michie S, van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. Implement Sci 2011 Apr 23;6:42. [doi: 10.1186/1748-5908-6-42] [Medline: 21513547]
17. Cane J, O'Connor D, Michie S. Validation of the theoretical domains framework for use in behaviour change and implementation research. Implement Sci 2012 Apr 24;7:37. [doi: 10.1186/1748-5908-7-37] [Medline: 22530986]
18. Laws RA, Denney-Wilson EA, Taki S, et al. Key lessons and impact of the growing healthy mHealth program on milk feeding, timing of introduction of solids, and infant growth: quasi-experimental study. JMIR Mhealth Uhealth 2018 Apr 19;6(4):e78. [doi: 10.2196/mhealth.9040] [Medline: 29674313]
19. Ellman K. SMS marketing by the numbers: 30+ texting statistics you need to know. Attentive. 2022. URL: https://www.attentive.com/blog/texting-statistics [accessed 2025-09-05]
20. Taki S, Lymer S, Russell CG, et al. Assessing user engagement of an mHealth intervention: development and implementation of the growing healthy app engagement index. JMIR Mhealth Uhealth 2017 Jun 29;5(6):e89. [doi: 10.2196/mhealth.7236] [Medline: 28663164]
21. Steiner JF, Zeng C, Comer AC, et al. Factors associated with opting out of automated text and telephone messages among adult members of an integrated health care system. JAMA Netw Open 2021 Mar 1;4(3):e213479. [doi: 10.1001/jamanetworkopen.2021.3479] [Medline: 33769509]
22. Delaney T, Mclaughlin M, Hall A, et al. Associations between digital health intervention engagement and dietary Intake: a systematic review. Nutrients 2021 Sep 20;13(9):3281. [doi: 10.3390/nu13093281] [Medline: 34579158]
23. Fjeldsoe BS, Goode AD, Job J, Eakin EG, Spilsbury KL, Winkler E. Dose and engagement during an extended contact physical activity and dietary behavior change intervention delivered via tailored text messaging: exploring relationships with behavioral outcomes. Int J Behav Nutr Phys Act 2021 Sep 7;18(1):119. [doi: 10.1186/s12966-021-01179-8] [Medline: 34493307]

Abbreviations

CFHS: Child and Family Health Services
HB4HNEKids: Healthy Beginnings for Hunter New England Kids
HNE: Hunter New England
mHealth: mobile health
NSW: New South Wales

Edited by N Cahill; submitted 17.12.24; peer-reviewed by J Job, K Crosby; revised version received 26.08.25; accepted 26.08.25; published 01.10.25.

Please cite as:
Brown AL, Hudson N, Pinfold J, Sewter R, Davies L, Lecathelinais C, Jackson JK, Delaney T, Kavalec S, Sutherland R
The Impact of Dose in an mHealth Intervention to Support Parents and Carers Via Healthy Beginnings for Hunter New England Kids Program: Pragmatic Randomized Controlled Trial
JMIR Form Res 2025;9:e70158
URL: https://formative.jmir.org/2025/1/e70158
doi:10.2196/70158

© Alison L Brown, Nayerra Hudson, Jessica Pinfold, Rebecca Sewter, Lynda Davies, Christophe Lecathelinais, Jacklyn K Jackson, Tessa Delaney, Sienna Kavalec, Rachel Sutherland. Originally published in JMIR Formative Research (https://formative.jmir.org), 1.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Multiuser Application for the Diagnosis and Treatment of Depression in Women's Self-Help Groups: Pilot Randomized Controlled Trial

Amritha Bhat[1], MD, MPH; Ruben Johnson-Pradeep[2,3], MBBS, MD; Bharat Kalidindi[4], BAMS, MPH; Dhinagaran Devadass[4], BDS, MBA; B Ramakrishna Goud[5], MBBS, MD; Tony Raj[4,6], MBBS, MD; Sumithra Selvam[7], PhD; Yesenia Navarro-Aguirre[1,8], BA; Pamela Y Collins[9,10], MD; Krishnamachari Srinivasan[2,3], MBBS, MD

[1]Department of Psychiatry and Behavioral Sciences, University of Washington, 1959 NE Pacific Street, Box 356560, Seattle, WA, United States
[10]Department of Psychiatry & Behavioral Sciences, The Johns Hopkins University School of Medicine, Baltimore, MD, United States
[2]Department of Psychiatry, St John's Medical College, Bengaluru, India
[3]Division of Mental Health & Neurosciences, St. John's Research Institute, Bengaluru, India
[4]Division of Medical Informatics, St. John's Research Institute, Bengaluru, India
[5]Department of Community Medicine, St John's Medical College, Bengaluru, India
[6]Department of Physiology, St. John's Medical College, Bengaluru, India
[7]Division of Epidemiology, Biostatistics and Population health, St. John's Research Institute, Bengaluru, India
[8]Department of Anthropology, University of Washington, Seattle, WA, United States
[9]Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, United States

Corresponding Author:
Amritha Bhat, MD, MPH
Department of Psychiatry and Behavioral Sciences, University of Washington, 1959 NE Pacific Street, Box 356560, Seattle, WA, United States

Abstract

Background: Depression in women results in elevated morbidity rates, functional impairment, diminished quality of life, and an increased risk of suicide. Numerous obstacles impede access to mental health treatment for women in India. Digital mental health solutions can bridge the treatment gap, but it is important to tailor these solutions to the context and to end-users.

Objective: We conducted a pilot randomized controlled trial to test the feasibility, acceptability, and preliminary effectiveness of a mental health app deployed in community-based organizations in improving depression outcomes.

Methods: The Multiuser Interactive Health Response Application (MITHRA) is a multiple-user mobile app used in community-based organizations for screening, tracking, and supporting stepped-care treatment for depression. MITHRA is based on the healthy activity program, a brief psychological intervention based on behavioral activation. It includes audio, video, and enhanced touchscreen capabilities to overcome the barrier of illiteracy and lack of access. It was developed in collaboration with a participatory design group consisting of primary and secondary end-users and is available on tablets installed in self-help groups (SHGs), which are community-based organizations in India. The SHGs were randomized to MITHRA (n=3) or enhanced usual care (EUC; n=3). During SHG meetings, women completed the Patient Health Questionnaire-9 (PHQ-9). Based on their PHQ-9 scores, they were assigned different modules. In the EUC SHGs, women viewed one module of education on symptoms of depression. Primary outcomes include feasibility and acceptability, and secondary outcomes include depressive symptoms and functioning. Repeated-measures ANOVA was performed to compare the change in the outcome scores over time between study groups. A P value of<.05 was considered statistically significant.

Results: MITHRA was found to be feasible and acceptable. A total of 96% of intervention arm participants completed at least half of their assigned modules. Although not powered for effectiveness outcomes, in this trial, we found that the change at 6 months from baseline in depressive symptoms (PHQ-9) were significantly different between MITHRA and EUC (P=.037), with greater improvement in the intervention group. Similarly, significant improvement in the World Health Organization Disability Assessment Scale score was noted in the MITHRA group (P=.005).

Conclusions: MITHRA is feasible and acceptable for use in women's SHGs. Larger studies should examine the effectiveness of this approach in identifying and treating depression.

Trial Registration: ClinicalTrials.gov NCT04480021; https://clinicaltrials.gov/study/NCT04480021

International Registered Report Identifier (IRRID): RR2-10.2196/42919

XSL•FO
RenderX

(JMIR Form Res 2025;9:e68052) doi:10.2196/68052

KEYWORDS

women; depression; community-based; mobile mental health; rural

Introduction

Mental disorders are among the top 10 leading causes of global disease burden [1]. Depression is the most common mental disorder and is twice as common in women than in men [2]. Women in general experience several sources of stress such as being the sole childrearing adult in a household, managing multiple roles, unequal power relations with men, a sense of powerlessness [3], and higher rates of poverty [4], all of which are associated with a higher risk of depression.

Women with depression, even of mild to moderate severity, experience high rates of morbidity, functional impairment, low quality of life [5], and higher risk of suicide [6]. The effects of depression in women impact families as well given the multidimensional roles that they play [7]. Yet, depression often goes undiagnosed and untreated in women in India due to stigma, childcare, gender inequality, and mental health workforce deficits, impacting mental health treatment access [8]. Only one woman attends an outpatient mental health appointment for every three men in India [8], and untreated depression among women is a major public health problem. While digital mental health solutions can help close this treatment gap, there is a need to adapt these digital mental health solutions to the context and to specific users. For example, while app-based depression screening and treatments are widely used and effective, issues such as shared ownership of phones, which is very common in India [9], may pose a barrier to the widespread use of digital mental health solutions. In general, access to mobile phones is lower among women in India, especially in rural India, as women have time allocation constraints on phone usage, limited digital skills, and often depend on men for phone ownership [10,11].

Given these contextual considerations, we used a participatory design to develop and test an app—Multiuser Interactive Health Response Application (MITHRA)—that screens for depression and provides a brief behavioral intervention [12]. This is a multiuser app deployed in self-help groups (SHGs), which are community-based organizations for women. The conceptual model for MITHRA is based on Fortney et al's [13] access to care model, and MITHRA targets actual and perceived barriers to access identified by patients. MITHRA uses nonencounter-based screening, tracking, and low-intensity interventions, addressing the barriers of stigma and travel time. MITHRA addresses the perceived need for care among women by providing education about depression. The process of app development has been described previously [14]. Briefly, women are presented with multimedia modules to view based on their Patient Health Questionnaire-9 (PHQ-9) score. Women who score <5 receive a prompt to watch general information on depression; women who score 5 or higher watch short (10 to 15 min) interactive multimedia-based modules based on the Healthy Activity Program [15], a brief behavioral intervention.

Here, we describe a pilot randomized controlled trial (RCT) designed to test the feasibility, acceptability, and preliminary effectiveness of MITHRA in improving depression outcomes.

Methods

Ethical Considerations

All study activities were approved by the Institutional Review Board at the University of Washington (no. STUDY00010415, approved 8 June 2020), and the Institutional Ethics Committee at St. John's Medical College (IEC Study Ref. no. 184/2020, approved 24 July 2020). We obtained informed consent from all participants; for those who could not read, we explained the study details and obtained consent by thumbprint and by recording verbal consent, witnessed, and signed by a neutral observer. Data for analyses were deidentified. Participants received INR 200 (US $ 2.50) for each assessment/interview

Study Sites

The study site was within Anekal taluk (subdistrict administrative block of a state in India) of Karnataka, which is approximately 25 miles from Bengaluru. We included 10 villages with functional SHGs with a population of 9724 people [9]. Focus groups, usability testing, and deployment of the tablet-based apps for the pilot RCT were all conducted in these SHGs. The methods for the pilot RCT are described below.

Randomization and Masking

We randomized three SHGs to use MITHRA and three SHGs to enhanced usual care (EUC). A simple randomization list was generated, and the eligible SHGs were allocated into MITHRA or EUC. Participating SHGs were in different villages, with no possibility of contamination or exchange of information about the intervention. Study investigators were blinded to the random allocation. At recruitment, 3 months, and 6 months, a research assistant administered outcome assessments over the telephone to avoid unmasking of the SHG randomization status (presence of tablets in the SHGs). The research assistant only had access to the name and mobile number of the participant, and no details on the participant's village of residence or SHG.

Intervention

MITHRA was available on tablets placed in an assigned private place in each SHG randomized to MITHRA. Women typically attend SHG meetings two to three times a month, and the use of MITHRA at every attendance was encouraged. They logged into MITHRA with a fingerprint secure single-user sign-on to complete depression screening using the Patient Health Questionnaire-9 (PHQ-9). On completion of the PHQ-9, each woman received a prompt based on her total score and viewed the recommended modules. Users could unlock virtual reward points/badges on completion of the required questionnaires, modules, and activities. Any woman who scored anything other than 0 on question 9 of the PHQ-9 (ie, the suicidal ideation

XSL•FO
RenderX

item) would be prompted to call the community health worker associated with that SHG. For these women, the app would also trigger an alert to the community health worker who would immediately contact the SHG administrator on site and call the patient to complete a suicide risk assessment [16]. Community health workers were trained in safety protocols.

In the EUC SHGs, women watched a module offering education regarding the symptoms of depression.

Data Analysis

Intent-to-treat analyses were performed. We used the Q-Q plot to test assumptions of normality. We used descriptive statistics to assess our primary outcome, the app usage rates. We used the mean and SD for normally distributed variables measured at baseline, 3 months, and 6 months. Variables measured included depressive symptoms measured on the Quick Inventory of Depressive Symptoms (QIDS) scale [17], functioning symptoms on the World Health Organization Disability Assessment Scale (WHODAS) [10], and behavioral activation on the Behavioral Activation Depression Scale (BADS) [11]. We used the χ^2 test or Fisher exact tests, as appropriate, to compare the categorical variables at baseline between the study groups. At each time of assessment, we compared the outcome parameters between study groups using independent *t* tests or Mann-Whitney *U* tests, as appropriate. The change in outcomes at 3 and 6 months from baseline was compared between the study groups using the Mann-Whitney *U* test. Within each study group, the McNemar χ^2 test was used to compare the change in the proportion of depression categories (minimal, and mild to moderate) from baseline. In addition, repeated-measures ANOVA was performed on log-transformed outcome variables to compare the change in the outcome scores over time between the study groups. A *P* value of <.05 was considered statistically significant. All analyses were performed using SPSS (version 26.0; IBM Corp.).

Results

The flowchart showing participant inclusion is shown in Figure 1.

Figure 1. Enrollment flow chart. SHGs: self-help groups.

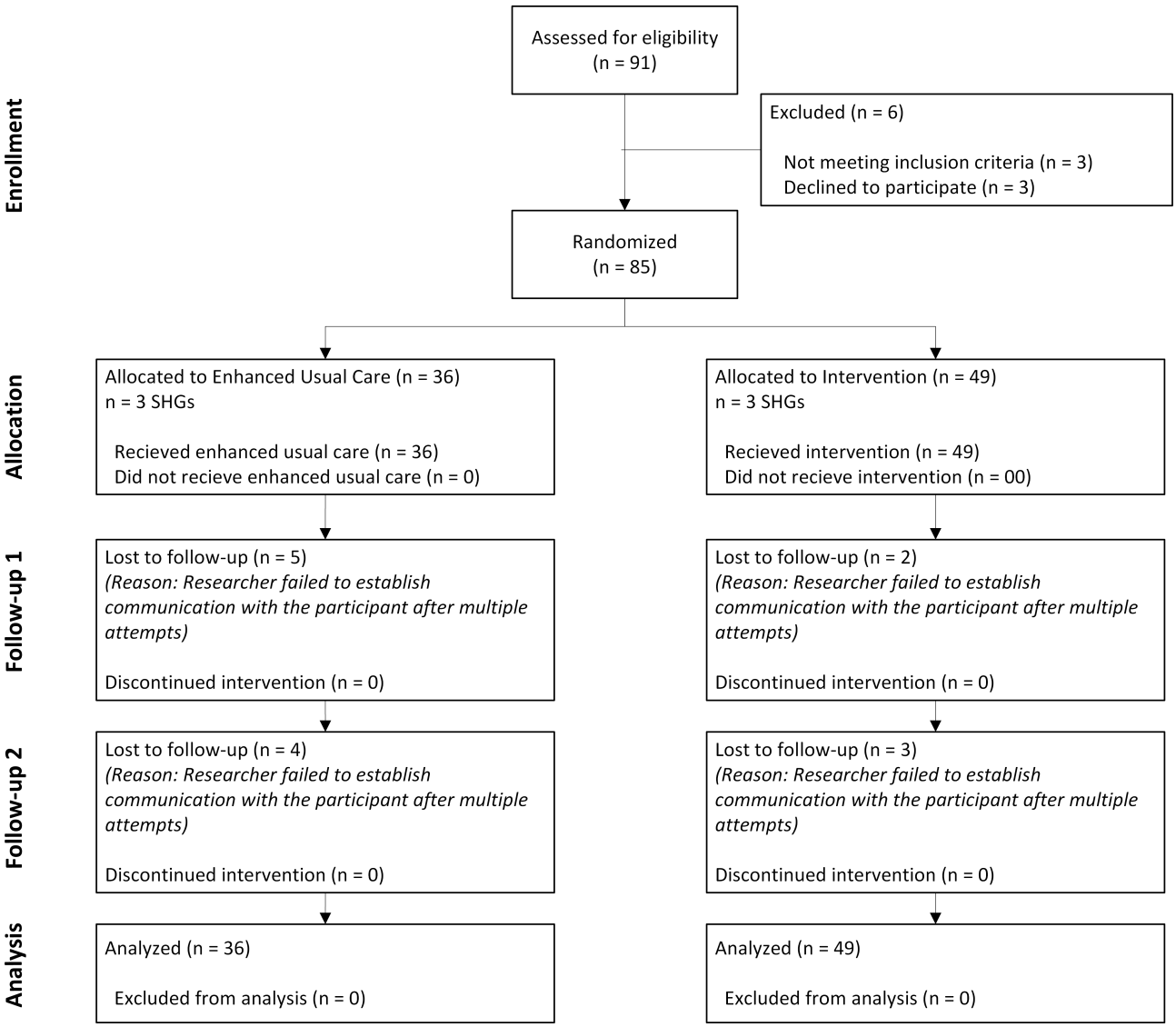

XSL•FO
RenderX

Participant Characteristics

The baseline characteristics of the study participants are provided in Table 1 and were comparable between study groups except for the type of family, with a significantly higher proportion of participants belonging to nuclear families in the intervention group compared to the EUC group. At baseline, the PHQ-9 and QIDS scale scores were significantly higher and the WHODAS score was significantly lower in the intervention group than in the EUC group.

Table . Baseline characteristics of participants.

Variable	MITHRA group (n=49)	EUC group (n=36)
Age, y; mean (SD); range, 18-63 y	41.6 (10.5)	40.7 (11.9)
Marital status		
Married	49 (100)	33 (91.7)
Never married/Widow	0	3 (8.3)
Education, n (%)		
No schooling	23 (46.9)	19 (52.8)
Primary–Middle school	11 (22.5)	2 (5.5)
Secondary-Higher Secondary	12 (24.5)	11 (30.6)
Graduation/Diploma	3 (6.1)	4 (11.1)
Type of family, n (%)[a]		
Nuclear	41 (83.7)	19 (52.8)
Joint	8 (16.3)	17 (47.2)
Occupation, n (%)		
Unskilled manual work	27 (55.1)	20 (55.6)
Skilled manual work	10 (20.4)	14 (38.9)
Homemaker	12 (24.5)	1 (2.8)
Working professional	0	1 (2.8)
Monthly family income, n (%)		
≤INR 10,000 (US $112.65)	22 (44.9)	10 (27.8)
> INR 10,000 (US $112.65)	27 (55.1)	26 (72.2)
Years of association with self-help groups, n (%)		
<5 years	22 (44.9)	10 (27.8)
5-10 years	9 (18.4)	10 (27.8)
>10 years	18 (36.7)	16 (44.4)
Baseline measures, mean (SD)		
PHQ-9[ab]	4.24 (3.35)	3.03 (3.40)
QIDS scale[ac]	4.18 (2.78)	3.25 (3.36)
WHODAS[ad]	27.78 (4.73)	29.92 (3.64)

[a]P<.05

[b]PHQ-9: Patient Health Questionnaire-9.

[c]QIDS: Quick Inventory of Depressive Symptoms.

[d]WHODAS: World Health Organization Disability Assessment Scale.

Feasibility

Feasibility of Randomization and Masking

Randomization was feasible, although some participant level characteristics were significantly different between the intervention and EUC groups. Patient data continued to be masked, as the research assistant was not aware of group assignment during the phone-based research assessments.

Retention

At the 3-month follow-up, we had 47/49 and 31/36 complete assessments for the intervention and EUC groups, respectively. At the 6-month follow-up, we had 46/49 and 32/36 complete

XSL•FO
RenderX

assessments for the intervention and EUC groups, respectively, with retention rates of 92% at both 3 and 6 months.

Safety

There were no adverse events, and no participants reported suicidal ideation or required a referral to the Primary Health Center per the protocol.

App Usage Rates

Only the MITHRA arm was considered, and completion of 50% of the modules was considered adequate usage. Of the 49 women assigned to MITHRA, 47 completed 50% of the modules.

Effectiveness of the Intervention

The change in outcome parameters at 3 and 6 months from baseline between the MITHRA and EUC groups is reported in Table 2. At 6 months, the change in the PHQ-9 score from baseline was significantly different between the study groups (*P*=.037), with greater improvement in the intervention group. Although statistically non-significant, the change in the QIDS scale score was greater in the intervention group than in the EUC group (*P*=.07). There was no significant difference in the BADS score at 6 months. The change in the WHODAS score was significantly different between the study groups (*P*=.005), with more improvement in the intervention group.

Table . Comparison of outcome parameters between study groups at baseline, 3, and 6 months.

Outcome parameter	MITHRA group (n=49)	EUC group (n=36)	*P* value[a]
PHQ–9[b] score, median (IQR)			
Baseline	4 (1 to 6)	2 (0.25 to 4.75)	.04
Change at 3 months	-2 (-4 to - 1)	-1 (-4 to 0)	.20
Change at 6 month	-3 (-5 to - 0.75)	-1 (-3 to 0)	.04
QIDS scale[c] score, median (IQR)			
Baseline	4 (2 to 5)	2.5 (0 to 4)	.04
Change at 3 months	0 (-2.0 to 1.0)	0 (-2.0 to 1.0)	.58
Change at 6 months	-1.0 (-3.0 to 1.0)	0 (-0.75 to 2.75)	.07
BADS[d] score, median (IQR)			
Baseline	30.0 (27.0 to 33.5)	34.0 (28.0 to 38.7)	.13
Change at 3 months	1 (-3.0 to 5.0)	0 (-4.0 to 3.0)	.33
Change at 6 months	1.50 (-3.0 to 6.0)	0 (-5.0 to 4.0)	.19
WHODAS[e] score, median (IQR)			
Baseline	27.0 (24.5 to 31.0)	29.5 (27.0 to 33.0)	.03
Change at 3 months	-1.0 (-4.0 to 1.0)	-1.0 (-4.0 to 0)	.48
Change at 6 months	0.50 (-3.0 to 4.0)	-3.0 (-5.0 to 0)	.005

[a]*P* values were obtained using the Mann-Whitney *U* test at each time point.

[b]PHQ-9: Patient Health Questionnaire-9.

[c]QIDS: Quick Inventory of Depressive Symptoms.

[d]BADS: Behavioral Activation Depression Scale.

[e]WHODAS: World Health Organization Disability Assessment Scale.

The PHQ-9 scores were categorized as minimal and mild-to-moderate and compared by study groups, as presented in Table 3. There was no significant difference in the proportion of participants with mild-to-moderate depression between the study groups. However, within-group analysis showed a significant reduction in the proportion of participants with mild-to-moderate depression from baseline to 6 months in the intervention group (*P*<.001).

XSL•FO
RenderX

Table . Comparison of Patient Health Questionnaire-9 (PHQ-9) scores between the study groups.

PHQ-9 score	MITHRA group (n=49)	EUC group (n=36)	*P* value
PHQ-9 score at baseline[a], mean (SD)			.09
Minimal depression	28 (57.1)	27 (75.0)	
Mild-to-moderate depression	21 (42.9)	9 (25.0)	
PHQ-9 score at 3 months, mean (SD)			.76
Minimal depression	46 (97.9)	30 (96.8)	
Mild-to-moderate depression	1 (2.1)	1 (3.2)	
PHQ-9 score at 6 months, mean (SD)			.19
Minimal depression	42 (91.3)	26 (81.3)	
Mild-to-moderate depression	4 (8.7)[b]	6 (18.8)	

[a]For within-group comparisons, the change from baseline to that at 6 months was calculated using the McNemar chi-square test.

[b]$P<.001$

The results of the additional analysis using repeated-measures ANOVA revealed that there was a significant time effect for the PHQ-9, BADS, and QIDS scale scores, indicating that the change in the outcome scores over time was statistically significant ($p<0.01$). However, there was no significant interaction between time and the study groups for any of the outcomes.

Discussion

This study demonstrates the feasibility and acceptability of deploying a culturally tailored app-based depression intervention (MITHRA) within women's SHGs in rural India. MITHRA is designed to address the large treatment gap for depression in women. It was developed in consultation with key end-users, it is culturally appropriate (settings and characters in the video reflect the users), available in the local language, and based on an evidence-based intervention.

We successfully completed a pilot cluster RCT comparing MITHRA to EUC to assess the feasibility of recruitment and randomization, acceptability, and preliminary effectiveness. In this RCT, we randomized participants to app-based depression screening and treatment in community-based settings. Participants in the control arm received EUC; they viewed one module of education on the symptoms of depression and how to seek help. Typical or usual care in SHGs includes no screening or education regarding depression. We found that it is feasible to recruit and randomize SHGs to app-based treatment or usual care, and to obtain consent from SHG members to participate in a clinical trial. Our pilot cluster RCT adds to a small but growing body of evidence supporting the use of digital mental health interventions for women in low-resource settings, particularly in South Asia, where access to mental health care remains limited. For example, a similar study in a low- or middle-income country highlights the high acceptability of mobile-based interventions among women with limited access to traditional mental health services [18].

Randomization was not balanced for patient-level characteristics such as socioeconomic status and depression, a limitation that can occur in cluster RCTs [19]. We controlled for these characteristics in our analysis. We had an excellent retention rate of 92% at both the time-points of 3 months and 6 months. Women in the intervention arm engaged well with the app, with 96% of the women completing at least 50% of the modules assigned to them. This suggests that carefully designed locally contextualized digital tools can be both feasible and acceptable in these settings. We also demonstrated the feasibility of collecting the required data and conducting analyses.

Preliminary analyses showed that symptom scores as measured by the PHQ-9 and QIDS scales decreased over time in women assigned to MITHRA compared to EUC. Our post-hoc analysis showed that in MITHRA SHGs, the proportion of women with mild-to-moderate depression symptoms as measured using the PHQ-9 was significantly lower at 6 months compared to that at baseline. This proportion did not reach statistical significance in the EUC arm. While the study was not powered to detect efficacy, these signals are promising and aligned with broader evidence, indicating that even lower-intensity digital psychological interventions can lead to meaningful improvements in mental health outcomes.

Our findings also contribute to the growing literature on stepped care approaches for depression in low- or middle-income countries. MITHRA incorporates the core principles of stepped care by providing low-intensity scalable interventions for women with mild-to-moderate symptoms while also including protocols for referral or treatment intensification for those with severe symptoms. Although no women in our study met the criteria for severe depression or required stepped-up care, the infrastructure we established for monitoring and referral are aligned with best practices to allocate resources efficiently while ensuring that individuals receive appropriate levels of care. The absence of referrals in our sample may reflect the fact that women with more severe depressive symptoms are less likely to participate in SHGs, an important gap that future work must

XSL•FO
RenderX

address. Similarly, no women in the study endorsed suicidal ideation; however, we did have appropriate treatment intensification protocols in place as part of the stepped care.

The limitations of our study include that our results may not be applicable to women with severe depressive symptoms who may not be attending SHG meetings. Additionally, as a feasibility study, our primary objective was to assess feasibility and acceptability, and we were not powered to assess effectiveness. The preliminary data generated will inform the design of a fully-powered RCT. In addition, the study was funded in 2020, and India was in a national lockdown for 2 months in early 2020 due to the COVID-19 pandemic. This delayed the study onset, as community gatherings were not allowed.

We found that it is feasible to conduct a pilot RCT of an app-based depression screening and brief behavioral intervention in community-based organizations in rural India. App-based depression screening and treatment such as that offered by MITHRA can help address barriers to mental health treatment access such as transportation and mental health workforce deficits. Deploying MITHRA in women's SHGs can also help address gender-specific stigma and access to treatment. Widespread implementation of community-based solutions such as MITHRA that focus on the prevention of depression onset or treatment of mild-to-moderate symptoms should always be coupled with a plan to intensify care as needed for those with severe symptoms. The next steps are to test the effectiveness of this approach in the treatment of depression in women in a larger RCT.

Acknowledgments

Funding was provided by the National Institute of Health (grant no. R21MH124073).

We are grateful for consultation from the Healthy Activities Program, and from Mr. Azaz Khan and Dr. John Naslund. The study would not have been possible without the support of the self-help groups, the governing organization, and the members of the participatory design group, as well as the women who participated in focus group discussions and in the pilot randomized controlled trial of using Multiuser Interactive Health Response Application (MITHRA).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Checklist 1
CONSORT checklist.
[PDF File, 221 KB - formative_v9i1e68052_app1.pdf]

References

1. GBD 2019 Mental Disorders Collaborators. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990–2019: a systematic analysis for the global burden of disease study 2019. Lancet Psychiatry 2022 Feb;9(2):137-150. [doi: 10.1016/S2215-0366(21)00395-3]
2. Salk RH, Hyde JS, Abramson LY. Gender differences in depression in representative national samples: meta-analyses of diagnoses and symptoms. Psychol Bull 2017 Aug;143(8):783-822. [doi: 10.1037/bul0000102] [Medline: 28447828]
3. Patel V, Kirkwood BR, Pednekar S, Weiss H, Mabey D. Risk factors for common mental disorders in women. Br J Psychiatry 2006 Dec;189(6):547-555. [doi: 10.1192/bjp.bp.106.022558]
4. World health organization. Global Health Observatory data repository. 2018. URL: http://apps.who.int/gho/data/node.main.MHSUICIDE?lang=en [accessed 2025-09-19]
5. Judd LL, Schettler PJ, Akiskal HS. The prevalence, clinical relevance, and public health significance of subthreshold depressions. Psychiatr Clin North Am 2002 Dec;25(4):685-698. [doi: 10.1016/s0193-953x(02)00026-6] [Medline: 12462855]
6. Sadek N, Bona J. Subsyndromal symptomatic depression: a new concept. Depress Anxiety 2000;12(1):30-39. [doi: 10.1002/1520-6394(2000)12:1<30::AID-DA4>3.0.CO;2-P] [Medline: 10999243]
7. Sonavane R, Kasthuri A. Knowledge, attitude and practice of first aid among women in a rural area. : Department of Community Health; 2008 URL: https://api.semanticscholar.org/CorpusID:74515111 [accessed 2025-09-19]
8. Malhotra S, Shah R. Women and mental health in India: an overview. Indian J Psychiatry 2015 Jul;57(Suppl 2):S205-S211. [doi: 10.4103/0019-5545.161479] [Medline: 26330636]
9. Bhat A, Goud BR, Pradeep JR, Jayaram G, Radhakrishnan R, Srinivasan K. Can mobile health improve depression treatment access and adherence among rural Indian women? A qualitative study. Cult Med Psychiatry 2020 Dec;44(4):461-478. [doi: 10.1007/s11013-019-09664-3] [Medline: 31916181]

XSL•FO
RenderX

10. Üstün TB, Kostanjsek N, Chatterji S, Rehm J. Measuring Health and Disability: Manual for WHO Disability Assessment Schedule WHODAS 20: World Health Organization; 2010. URL: https://iris.who.int/bitstream/handle/10665/43974/9789241547598_eng.pdf [accessed 2025-09-19]
11. Kanter JW, Mulick PS, Busch AM, Berlin KS, Martell CR. The behavioral activation for depression scale (BADS): psychometric properties and factor structure. J Psychopathol Behav Assess 2007 Jul 13;29(3):191-202. [doi: 10.1007/s10862-006-9038-5]
12. Bhat A, Goud BR, Kalidindi B, et al. Mobile mental health in women's community-based organizations: protocol for a pilot randomized controlled trial. JMIR Res Protoc 2023 Feb 8;12(1):e42919. [doi: 10.2196/42919] [Medline: 36753310]
13. Fortney JC, Burgess JF, Bosworth HB, Booth BM, Kaboli PJ. A re-conceptualization of access for 21st century healthcare. J Gen Intern Med 2011 Nov;26 Suppl 2(Suppl 2):639-647. [doi: 10.1007/s11606-011-1806-6] [Medline: 21989616]
14. Ruben JP, Devadass D, Goud BR, et al. Development of a multiuser interactive health response application (MITHRA) for depression in women from a community-based organisation in India. BJPsych Open 2025 Mar 25;11(2):e58. [doi: 10.1192/bjo.2025.8] [Medline: 40129275]
15. Chowdhary N, Anand A, Dimidjian S, et al. The healthy activity program lay counsellor delivered treatment for severe depression in India: systematic development and randomised evaluation. Br J Psychiatry 2016 Apr;208(4):381-388. [doi: 10.1192/bjp.bp.114.161075] [Medline: 26494875]
16. Murray LK, Skavenski S, Bass J, et al. Implementing evidence-based mental health care in low-resource settings: a focus on safety planning procedures. J Cogn Psychother 2014 Aug;28(3):168-185. [doi: 10.1891/0889-8391.28.3.168] [Medline: 31602096]
17. Rush AJ, Trivedi MH, Ibrahim HM, et al. The 16-Item quick inventory of depressive symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. Biol Psychiatry Cogn Neurosci Neuroimaging 2003 Sep;54(5):573-583. [doi: 10.1016/S0006-3223(02)01866-8]
18. Ben-Zeev D, Agorinya J, Beaulieu A, et al. Sexual trauma and interest in mobile health among women with mental illness in Ghana. Internet Interv 2025 Jun;40:100829. [doi: 10.1016/j.invent.2025.100829] [Medline: 40276094]
19. Puffer S, Torgerson DJ, Watson J. Cluster randomized controlled trials. J Eval Clin Pract 2005 Oct;11(5):479-483. [doi: 10.1111/j.1365-2753.2005.00568.x] [Medline: 16164589]

Abbreviations

BADS: Behavioral Activation Depression Scale
EUC: enhanced usual care
MITHRA: Multiuser Interactive Health Response Application
PHQ-9: Patient Health Questionnaire-9
QIDS: Quick Inventory of Depressive Symptoms
RCT: randomized controlled trial
SHG: self-help group
WHODAS: World Health Organization Disability Assessment Scale

Edited by A Mavragani; submitted 13.11.24; peer-reviewed by L Raymond Guo, S Mukherjee; revised version received 12.08.25; accepted 30.08.25; published 01.10.25.

Please cite as:
Bhat A, Johnson-Pradeep R, Kalidindi B, Devadass D, Goud BR, Raj T, Selvam S, Navarro-Aguirre Y, Collins PY, Srinivasan K Multiuser Application for the Diagnosis and Treatment of Depression in Women's Self-Help Groups: Pilot Randomized Controlled Trial
JMIR Form Res 2025;9:e68052
URL: https://formative.jmir.org/2025/1/e68052
doi:10.2196/68052

© Amritha Bhat, Ruben Johnson-Pradeep, Bharat Kalidindi, Dhinagaran Devadass, B Ramakrishna Goud, Tony Raj, Sumithra Selvam, Yesenia Navarro-Aguirre, Pamela Y Collins, Krishnamachari Srinivasan. Originally published in JMIR Formative Research (https://formative.jmir.org), 1.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Avatar Customization and Embodiment in Virtual Reality Self-Compassion Therapy for Depressive Symptoms: Three-Part Mixed Methods Study

Thomas C Elliott[1*], BSc, MIDes; Yanzhuo Yang[1*], BEng, MIDes; Jarrod Knibbe[1], MEng, PhD; Julie D Henry[2*], PhD; Nilufar Baghaei[1*], PhD

[1]School of Electrical Engineering and Computer Science, Faculty of Engineering, Architecture and Information Technology, The University of Queensland, General Purpose South (78), St Lucia QLD 4067 Level 4, Brisbane, Australia

[2]School of Psychology, Faculty of Health and Behavioural Sciences, The University of Queensland, Brisbane, Australia

[*]these authors contributed equally

Corresponding Author:
Thomas C Elliott, BSc, MIDes
School of Electrical Engineering and Computer Science, Faculty of Engineering, Architecture and Information Technology, The University of Queensland, General Purpose South (78), St Lucia QLD 4067 Level 4, Brisbane, Australia

Abstract

Background: As virtual reality technologies become more accessible, understanding how design features influence user experience (UX) and psychological benefit is critical, particularly for emotionally sensitive interventions. Thus, while prior studies support the use of self-compassion paradigms in immersive virtual reality (VR) environments, the effects of avatar stylization, customization, and mirrored self-representation on therapeutic outcomes are not well understood. For instance, while it is often assumed that increasingly realistic avatars are preferable to less realistic ones, this basic premise remains largely untested.

Objective: This study aimed to evaluate whether avatar appearance, customization features, and virtual mirrors affect UX and therapeutic outcomes in VR self-compassion therapy. Specifically, we examined whether stylized avatars, avatar customization, and virtual mirror feedback influenced user-rated self-compassion and depression symptoms.

Methods: Across three between-subjects studies (N=107 neurotypical adults), participants engaged in an immersive individualized VR therapy protocol based on a 2-phase compassion task. The conditions were (1) stylized avatars (n=20), (2) stylized customizable avatars (n=49), and (3) stylized customizable avatars with a virtual mirror (n=38). Participants completed the User Experience Questionnaire, the Self-Compassion Scale, and the 8-item Patient Health Questionnaire (PHQ-8). In study 3, presence was also assessed using the Slater-Usoh-Steed scale. Qualitative feedback was analyzed thematically. Between- and within-study comparisons used *t* tests and Mann-Whitney *U* tests.

Results: Avatar customization (study 2) led to a significant increase in self-compassion (Self-Compassion Scale: baseline mean 3.05, SD 0.98; follow-up mean 3.55, SD 1.16; t_{89}=2.22; *P*=.03; *d*=–0.47), though PHQ-8 scores remained unchanged. The virtual mirror condition (study 3) significantly improved depression scores (PHQ-8: *U*=477.5; *z*=2.53; *P*=.01; *r*=0.30) and UX across four User Experience Questionnaire categories, including attractiveness and dependability. However, self-compassion did not significantly change in study 3 (mean 3.88, SD 1.33 → mean 4.09, SD 1.05; t_{63}=0.71; *P*=.47; *d*=0.18). Presence scores in study 3 (mean 4.56, SD 1.58) were also comparable to real-world benchmarks. Qualitative feedback highlighted strong engagement with avatars and mirrors, and participants reported emotional safety and personalization benefits.

Conclusions: Stylized avatars, when paired with customization and mirrored embodiment, can support UX and therapeutic benefit in VR self-compassion therapy. These findings challenge the assumption that hyperrealistic avatars are superior and highlight the importance of emotionally congruent design choices. The combination of stylization, individualization, and visual feedback may offer a low-barrier, user-aligned strategy for future therapeutic VR applications.

(JMIR Form Res 2025;9:e71004) doi:10.2196/71004

KEYWORDS

avatar design; virtual reality; self-compassion; depression; mental health; user experience; uncanny valley

XSL•FO
RenderX

Introduction

Virtual reality (VR) is emerging as a promising tool in therapeutic interventions for depression [1,2]. One approach involves users embodying avatars to give and receive compassion, which has been shown to lead to positive self-compassion outcomes. Customization of avatars and environments may improve these effects [3], yet the impact of such design decisions has not been systematically tested. Consequently, their influence on therapeutic outcomes remains unclear.

Much is already known about embodiment-related constructs relevant to VR therapy, including body ownership, agency, presence, and immersion [4-11]. These concepts provided were important theoretical grounding in support of a similar VR therapy study [12] and are therefore reviewed here to provide theoretical grounding in the context of the present study. Embodiment has been defined as the sense that the properties of a virtual body are experienced as one's own [13]. Supporting conditions for agency, self-location, and ownership contribute to embodiment and can be facilitated by a first-person perspective and visuomotor synchrony.

Presence is critical to this study, enabling behavioral engagement in VR. It depends in part on embodiment—specifically on recognizing a virtual body as one's own [14]. Immersion, however, refers to the technical capacity of the system to produce realistic and interactive environments [11]. To support immersion, this study used the Meta Quest 2 head-mounted display to support high immersion through multisensory stimulation [15].

Beyond these foundational elements, this study is situated within the domain of individualized VR (iVR), where VR experiences are tailored to the user—often through customization options. iVR has been shown to improve both user experience (UX) and therapeutic outcomes [2,3]. Customization of VR features such as adjusting game difficulty, audio, and objects from virtual environments has also been beneficial for the user [16-19]. Within iVR, avatar customization is one of many possibilities that allow users to modify the avatar's face, body, and clothing. Prior studies have shown that customization of an avatar has benefits for users with respect to identity, learning, inclusion, and engagement [20-23]. Because users often perceive avatars as extensions of themselves, this can enhance their sense of presence, embodiment, and attachment [24]. Also, modifying avatar features to resemble the user improved embodiment, even in healthy populations [25].

Prior studies show that strong embodiment can occur even with dissimilar or stylized avatars as long as key conditions like visuomotor synchrony and first-person perspective are met [4,26-28]. Also, human familiarity with mirrors from early development supports the perception of mirrored avatars as self-representations [29,30]. Therefore, virtual mirrors can further enhance embodiment by allowing users to observe their avatars in motion, including facial expressions [31,32]. Apart from UX benefits, virtual mirrors influence psychological and behavioral outcomes in VR. The "Proteus effect" suggests that avatars can alter user behavior and attitudes in virtual settings [29,33]. However, the virtual mirror potential in VR therapy is unclear and remains underexplored.

Self-compassion is the ability to "soothe oneself with kindness and nonjudgmental understandings in times of difficulty" [34] and is the foundation of compassion-focused therapy (CFT) [35]. In VR therapy, avatars replace human actors to deliver CFT [1,12]. However, there are few established avatar design standards, which may disadvantage neurodivergent populations, such as those with depression, who may require accessible VR design considerations.

Although prior work has demonstrated the therapeutic value of avatar-based CFT in VR, it is unclear whether these benefits generalize to design factors such as avatar stylization, customization, and the use of virtual mirrors. Personalization has been linked to improved body ownership, agency, and immersion [25], while virtual mirrors have shown potential therapeutic benefits in VR counseling [31,32]. These findings motivate the current study, which investigates how avatar appearance and environmental design influence UX and self-compassion in VR therapy. Accordingly, we address two research questions:

- RQ1: Do avatar and environmental design features influence UX?
- RQ2: Do these features influence therapeutic benefit, measured as self-rated self-compassion?

The aim of this study is to examine whether avatar appearance, avatar customization, and the use of virtual mirrors influence UX and self-compassion in VR therapy. We conducted three studies to examine the effects of (1) stylized avatars, (2) customizable stylized avatars, and (3) customizable stylized avatars with a virtual mirror. Each study built upon prior work [3], focusing only on the specified experimental features. A total of 107 neurotypical participants were recruited.

Methods

Overview

This manuscript adheres to the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines. This study replicates and extends a prior VR-based self-compassion therapy [3], which serves as the control condition for study 1. The original iVR experience included an onboarding interface where participants selected avatars, environments, and companion avatars with emotional behaviors (crying or upset). The therapy consisted of two stages: delivering compassion and receiving compassion.

Three between-subject studies were conducted:

- Study 1: Manipulated avatar appearance
- Study 2: Introduced stylized avatar customization
- Study 3: Introduced a virtual mirror

Each study built on the previous, allowing for pairwise comparisons. Stylized avatars and new features were integrated by editing the original Unity project files.

XSL•FO
RenderX

Avatar Fidelity

Figure 1 compares the stylized avatar used in experimental conditions with the realistic avatar from the original system. Based on Weidner et al [36], stylized avatars feature simplified textures and nonhuman proportions, while realistic avatars maintain human morphology and detail. The avatars of the control study (Figure 1, right) had such realistic features and therefore aligned with generic realistic avatars as defined in Weidner et al [36].

Figure 1. Comparison of stylized and realistic avatars used in the experimental and control conditions of study 1. The stylized avatar (left) featured simplified textures and nonhuman proportions, while the realistic avatar (right) reflected detailed human morphology. Participants (n=20) completed a between-subject virtual reality self-compassion protocol targeting depressive symptoms at The University of Queensland extended reality laboratory (2023 - 2024).

For clarity, for the remainder of the paper, the term "realistic" refers to the generic realistic category [36].

Measures

Two key instruments were used:

- User Experience Questionnaire (UEQ): Assesses UX across six scales.
- Self-Compassion Scale (SCS) [37]: Measures self-compassion

From study 2 onward, the 8-item Patient Health Questionnaire (PHQ-8) [38], a validated measure for screening depression symptoms, was introduced to support downstream clinical effects of the iVR intervention. As this paper continues the work of Halim et al [3], it was considered prudent to include ongoing depression measurement.

Within study 3, presence was measured using the Slater-Usoh-Steed (SUS) questionnaire [8,39], a 6-item scale to support the analysis of UX. With respect to measurement, presence was prioritized over embodiment measures due to practical constraints; also, as introduced earlier, presence performance implies embodiment quality [13,14].

A qualitative questionnaire gathered open-ended feedback about the avatars (studies 1 and 2) and a virtual mirror (study 3). These questions targeted participants' likes and dislikes to probe for system improvements and aimed to elicit uncanny valley (UV) responses via an open-ended question in studies 1 and 2:

- What were the top three things that you liked about individualized VR?
- How do you think the next version can be improved?
- How do you feel about your experience with the avatars? (Only in studies 1 and 2.)

In study 3, participants were given questions pertaining to their experience with the virtual mirror:

- How do you feel about your experience with the mirror; what aspects of the mirror did you like or dislike? (Only in study 3.)

In the absence of eye tracking technology, these responses provided a means of validating mirror interactions. Thematic analysis was used to evaluate these responses [40].

Data Analysis

Two comparisons were made:

- Within-study: pre-post changes in SCS, PHQ-8, and UEQ
- Between-studies: differences across conditions at session completion

For each quantitative data set, the Shapiro-Wilk normality test was first applied. Given the normality of the data, either a Mann-Whitney U test or a 2-tailed t test was performed. The null hypothesis assumed no significant differences between conditions.

XSL•FO
RenderX

Onboarding Procedure

iVR Therapy Stage 0: Individualization

From a lobby environment, the participants began by selecting an avatar that resembled themselves, selecting a therapeutic environment (Figure 2), a companion avatar, and the avatar's behavior. During customization (study 2), participants created avatars that resembled themselves. Movement in the VR scene was fixed to maintain task focus.

Figure 2. Environment selection interface during the individualization stage of the individualized virtual reality self-compassion therapy. Participants (n=49) in study 2 selected a preferred virtual setting—living room, Victorian house, or park—as part of the avatar and scene customization process. The study targeted depressive symptoms and was conducted at The University of Queensland's extended reality laboratory (2023 - 2024).

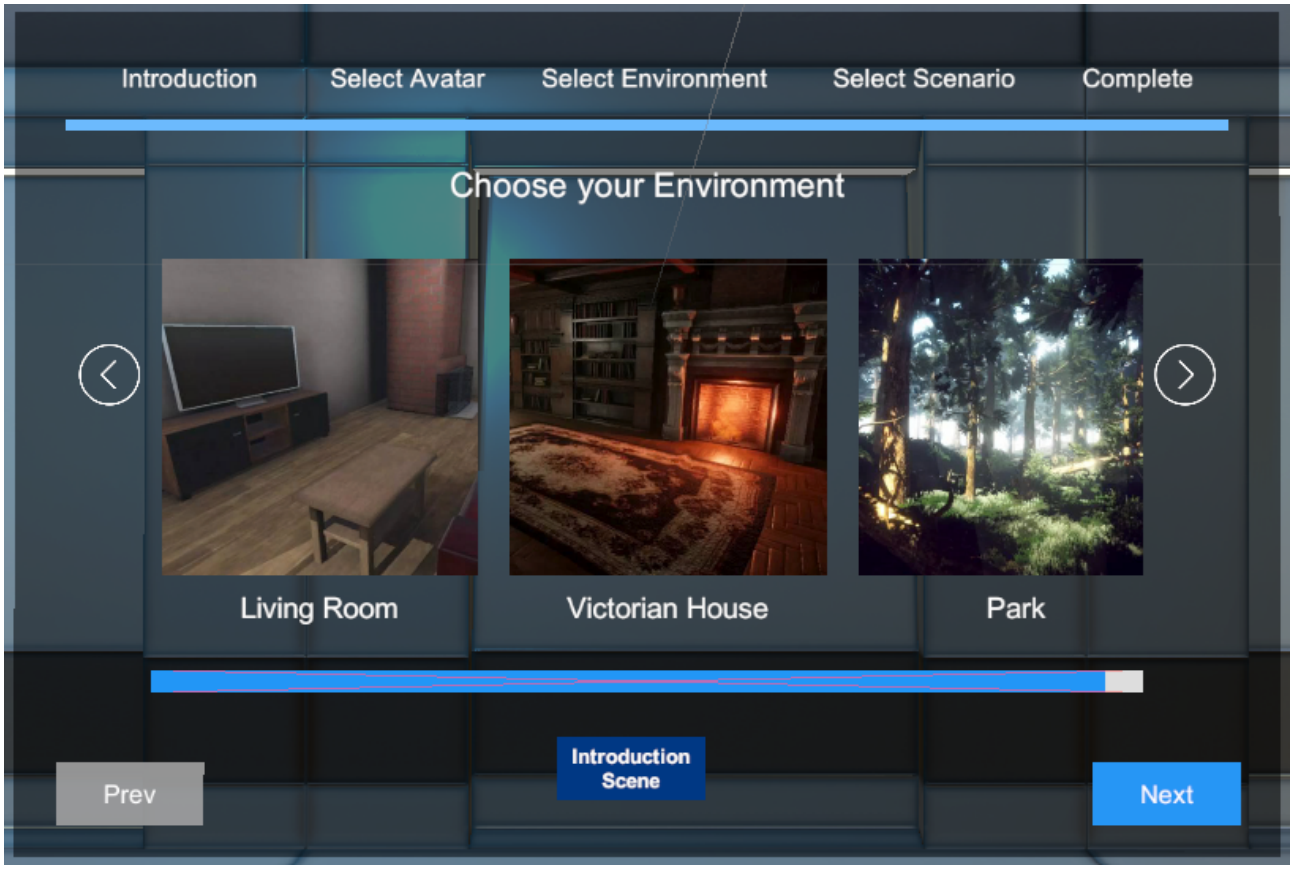

iVR Therapy Stage 1: Giving Compassion

Participants interacted with their selected companion, who acted out their emotional state before the user. An instruction panel guided the participants through three compassion strategies [1]:

- Validation: Acknowledge and accept the companion's distress
- Redirection of attention: Shift attention to something more positive
- Memory activation: Encourage recalling a comforting memory (eg, someone who is kind to them)

iVR Therapy Stage 2: Receiving Compassion

Participants reentered the same environment from the companion's perspective and observed their avatar delivering the recorded compassionate message from stage 1. This playback is intended to evoke a therapeutic experience and aligns with prior iVR and VR compassion interventions [3,12]. A second session occurred for participants in studies 2 and 3; these second sessions occurred 2 weeks after initial exposure to the intervention.

Experimental Procedure

Study 1: Stylized Avatars

To explore RQ1 and RQ2, study 1 replaced all avatars with stylized avatars purchased from SunBox Games [41]. Prior work suggests that highly realistic avatars do not necessarily enhance UX in immersive, stressful environments [42,43]. The virtual environment in this study has a unique human-avatar interaction in stage 2, where participants receive compassion from a virtual self—an essential component of self-compassion therapy.

According to Gisbergen et al [42], stylized avatars can potentially avoid UV effects [44], which may arise when avatars do not achieve full human realism. Therefore, by using lower-fidelity avatars, this study aimed to reduce user expectations for realism and improve predictability—key aspects of system dependability as measured by the UEQ. Although dependability is not a direct measure of UV, it serves as a relevant proxy in the context of UV evaluation.

A qualitative method was included to capture UV-related responses, similar to Becker-Asano et al [45]. UEQ and SCS were used for cross-study comparison. Figure 2 shows the avatars for participant selection via the lobby interface.

XSL•FO
RenderX

Study 2: Stylized Avatar Customization

Study 2 extended study 1 by adding customization options to stylized avatars (Figure 3). Participants adjusted the avatar using the following SunBox integrated tools [41]:

- Body: 8 parameters—height, fat, muscle, skin, nails, eye, eyelashes, and brow
- Features: hair and facial hair
- Face: 22 facial features (eg, ears, nose, mouth, chin)
- Clothing: 5 items—glasses, hat, tops (shirts, jumpers, etc), bottoms (pants, skirts, etc), and shoes

Figure 3. Enhanced avatar customization interface used in study 2 of the virtual reality self-compassion therapy protocol. Participants (n=49) modified a stylized avatar by adjusting body shape, facial features, and clothing using predefined editable parameters. Customization was mandatory and guided by the instruction to create an avatar resembling their self-image. The study targeted depressive symptoms and was conducted at The University of Queensland's extended reality laboratory (2023 - 2024).

Customization was required before proceeding. As a quality check, the facilitator encouraged participants to create avatars that resembled themselves within the limitations of the stylized appearance. Participants could not continue unless this criterion was satisfied by a visual check from the researcher supervising. While no formal metric was used, key features such as hairstyle, skin tone, and body shape were considered. All participants engaged fully in the customization process, and informal observation confirmed that stylized avatars generally resembled participants before continuing. Study 2 also added a second iVR session 2 weeks after the first, as this was required by the introduction of the PHQ-8.

Study 3: Introduction of Virtual Mirror

Study 3 extended study 2 by introducing a virtual mirror positioned to the side of the user during stage 1 (Figure 4). This design was intended to enhance embodiment and presence, as theorized in the Background subsection, in an effort to elicit improvements in UX and therapeutic outcomes using the current customized stylized avatar format. Participants observed real-time reflections of their avatars without interfering with the primary task.

XSL•FO
RenderX

Figure 4. Virtual mirror setup introduced in study 3 of the individualized virtual reality self-compassion therapy protocol. Participants (n=38) viewed real-time reflections of their customized stylized avatars through a virtual mirror positioned to their side during the interaction task. The setup was designed to enhance embodiment and therapeutic presence. The study targeted depressive symptoms and was conducted at The University of Queensland's extended reality laboratory (2023 - 2024). The participant shown is not identifiable and provided consent for inclusion.

Participant and Study Setting

Participants were recruited from the general community via posters, social media, university mailing lists, and word of mouth. Most were students or staff at The University of Queensland. Eligibility required being ≥18 years of age and capable of using immersive VR; those with severe motion sickness or conditions that might interfere with VR use were excluded.

Sessions were conducted one-on-one in a controlled room at The University of Queensland's extended reality laboratory (Brisbane, Australia), with a trained researcher present. Participants were recruited and tested between 2023 and 2024. Studies 2 and 3 included a 2-week in-person follow-up. Some individuals who registered did not attend their session and were not awarded course credit; no reasons were collected.

Sample sizes were determined pragmatically based on prior VR studies, available resources, and recruitment feasibility in a university setting. No formal power analysis was conducted due to the exploratory nature of the research.

Ethical Considerations

This study was approved by The University of Queensland Human Research Ethics Committee (approval #2023/HE000468). All participants provided written informed consent prior to participation. Data were de-identified before analysis to protect participant privacy and confidentiality. Participants received course credit through the university's research participation system. No images or materials in the manuscript or supplementary files contain identifiable individuals.

Results

Assessment Schedule

A flow diagram showing participant progression and the assessment schedule is presented in Figure 5.

XSL•FO
RenderX

Figure 5. Flow diagram showing participant progression and assessment schedule across three between-subjects studies conducted at The University of Queensland extended reality laboratory (2023 - 2024). Participants (N=107) were enrolled in study 1 (n=20), study 2 (n=49), or study 3 (n=38), and 7 were later excluded from the analysis. The studies targeted depressive symptoms and included pre- and posttest measures of self-compassion (SCS), depression (PHQ-8), presence (SUS), and qualitative feedback. Studies 2 and 3 included a second session at least 2 weeks later. Interventions varied by avatar customization and the presence of a mirror. PHQ-8: 8-item Patient Health Questionnaire; SCS: Self-Compassion Scale; SUS: Slater-Usoh-Steed.

XSL•FO
RenderX

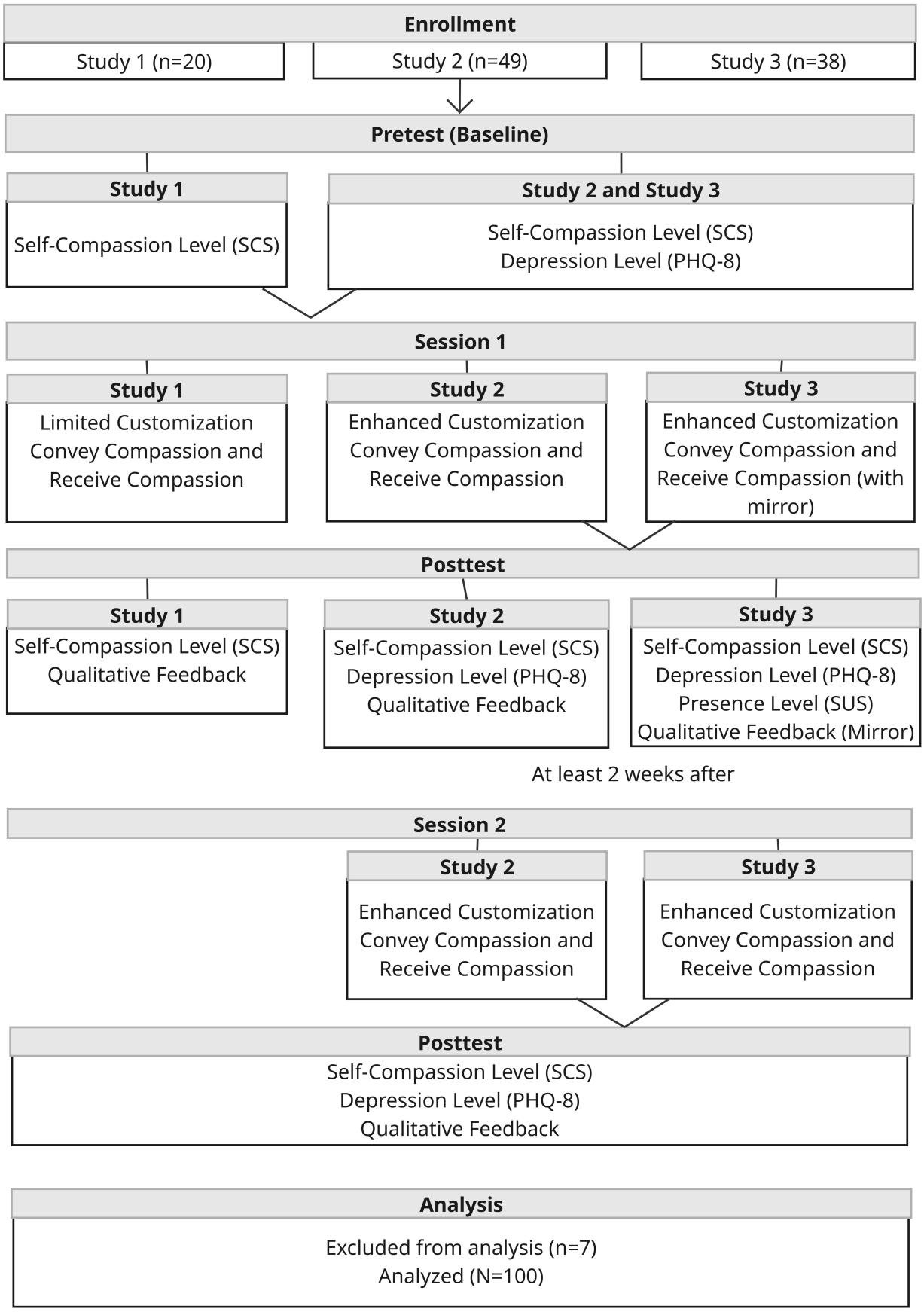

XSL•FO
RenderX

Study 1: Stylized Avatars

A total of 20 participants completed session 1 (n=12 female participants, n=6 male participants, and n=2 undisclosed).

Impact on User Experience (UEQ)

At the time of analysis, the UEQ benchmark included 21,175 users from 468 studies assessing software and digital products [46]. As shown in Figure 6, stylized avatars resulted in a category increase in perspicuity compared to the original study, as seen in Figure 7. All UEQ categories remained at least "above average," indicating no negative impact on the UX. Thus, in relation to RQ1, stylized avatars appear appropriate.

Figure 6. User Experience Questionnaire (UEQ) results from study 1, where participants (n=20; n=12 female, n=6 male, and n=2 undisclosed) interacted with stylized avatars during a single-session virtual reality self-compassion therapy targeting depressive symptoms. Participants completed the UEQ after session 1. Compared to the baseline results using realistic avatars (Figure 7), perspicuity scores improved while other dimensions remained at or above average. The study was conducted at The University of Queensland's extended reality laboratory (2023).

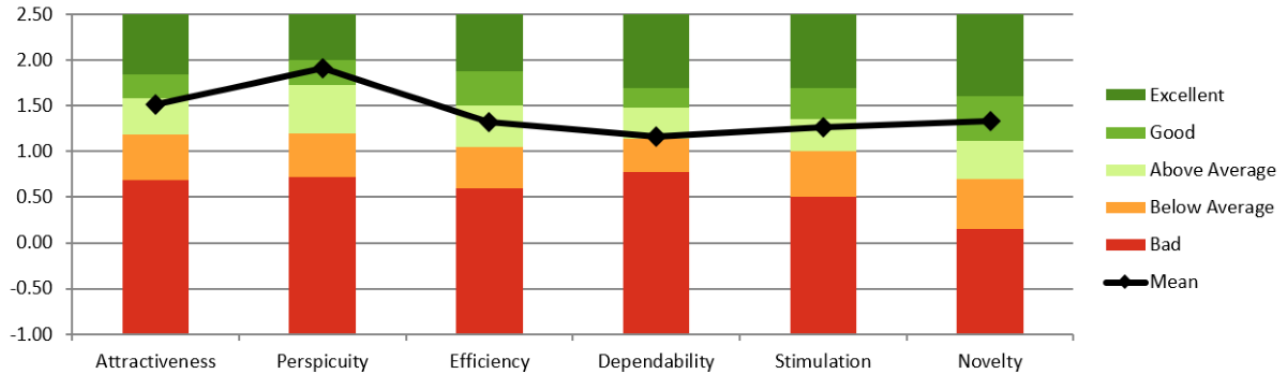

Figure 7. User Experience Questionnaire (UEQ) results from a prior study [3] using realistic avatars in a self-compassion virtual reality therapy protocol. This figure presents participant scores across six UEQ dimensions and serves as a baseline for evaluating the impact of avatar visual style in the current research. The original study targeted depressive symptoms and was conducted at The University of Queensland's extended reality laboratory (2022). Participants were neurotypical adults recruited from the university community.

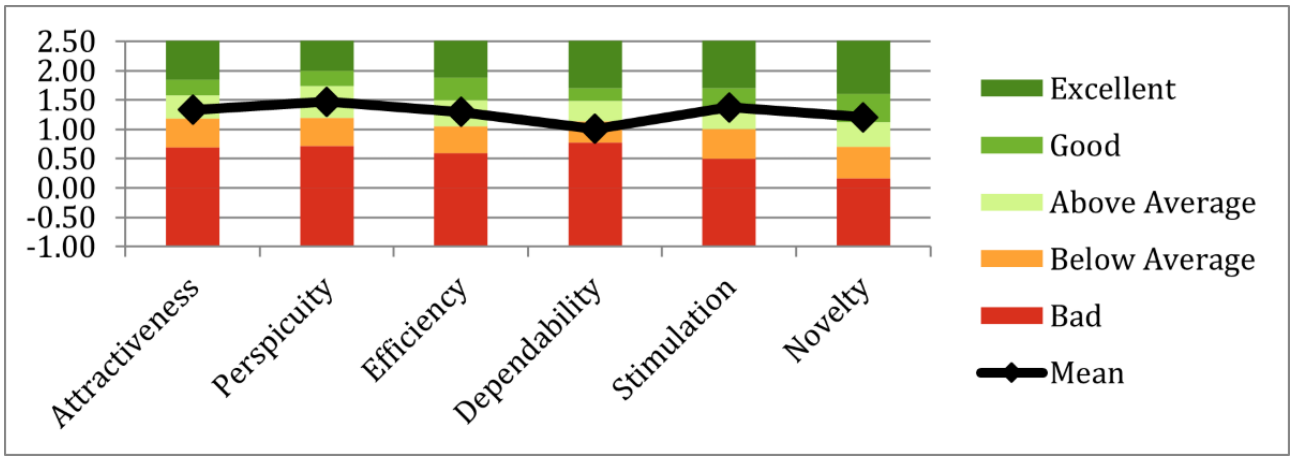

Impact of Self-Compassion (SCS)

The mean SCS score in study 1 was 3.11 (SD 1.14; Table S1 in Multimedia Appendix 1). Scores below 2.4 are low, between 2.4 and 3.6 are average, and above 3.6 are high [47]. Therefore, after one VR therapy session, participants reported average levels of self-compassion.

A Shapiro-Wilk test for normality confirmed that data were normally distributed for study 1. A between-studies comparison with Halim et al [3] using a 2-tailed *t* test found no significant difference in SCS between the control group (mean 3.07, SD 0.73) and study 1 (mean 3.11, SD 1.14; t_{53}=0.14; *P*=.89; Cohen *d*=0.04). These results indicate stylized avatars had no measurable impact on self-compassion (RQ2).

Study 2: Stylized Avatar Customization

A total of 49 participants completed session 1; 42 (86%) returned for session 2. The gender of the participants was as follows:

- Session 1: 22 female, 22 male, 5 undisclosed
- Session 2: 18 female, 21 male, 3 undisclosed

Impact on User Experience (UEQ)

Study 2 reused the same UEQ tool as study 1. As seen in Figure 8, scores declined in perspicuity, dependability, and novelty, with only dependability dropping below average. Given a between-studies comparison (Table S2 in Multimedia Appendix 1), UX differences were not statistically significant; therefore, with respect to RQ1, we did not find evidence that stylized avatar customization significantly influenced UX under the conditions tested.

XSL•FO
RenderX

Figure 8. User Experience Questionnaire (UEQ) results from study 2 (n=49; session 1), in which participants interacted with personalized stylized avatars during a virtual reality self-compassion therapy targeting depressive symptoms. Participants completed the UEQ after their first session. Compared to study 1, mean scores decreased in perspicuity, dependability, and novelty, with only dependability rated below the neutral benchmark. The study was conducted at The University of Queensland's extended reality laboratory (2023 - 2024).

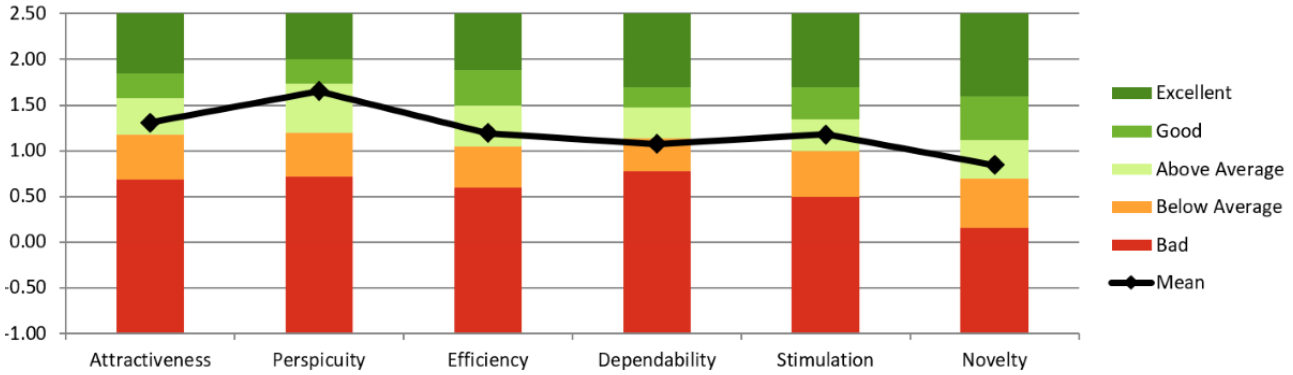

Figure 9. User Experience Questionnaire (UEQ) results from study 3 (n=38; session 1), where participants engaged in virtual reality self-compassion therapy using stylized avatars and a virtual mirror. Compared to study 2, scores significantly improved in attractiveness, perspicuity, dependability, and stimulation, with three dimensions rated as "excellent." Assessments were completed after the first session. The study was conducted at The University of Queensland's extended reality laboratory (2023 - 2024) and targeted depressive symptoms.

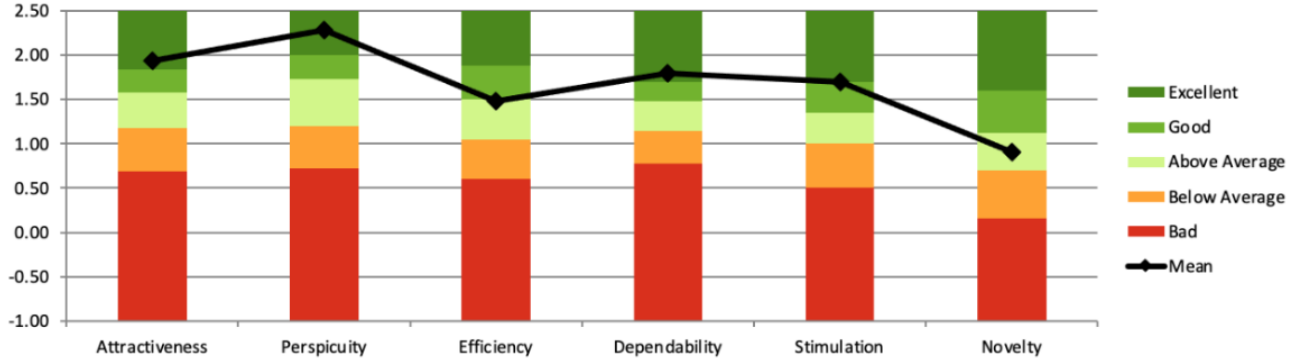

Impact of Self-Compassion (SCS)

For a within-study comparison, the Shapiro-Wilk test for normality confirmed that data were normally distributed. As seen in Table 1, a 2-tailed *t* test comparing baseline and session 2 SCS in study 2 revealed a statistically significant difference with a moderate effect size. These results provide evidence that customization positively influenced the therapeutic outcome (RQ2).

XSL•FO
RenderX

Table . Summary statistics and internal comparisons from study 2 of a three-part mixed methods study investigating virtual reality self-compassion therapy for depressive symptoms.[a]

Measures	Mean (SD) or mean rank	Mann-Whitney *U* test	*t* test (df) or *z*	*P* value (2-tailed)	Effect size (Cohen *d* or *r*)
UEQ[b]		—[c]	—	—	—
Attractiveness	1.31 (0.85)				
Perspicuity	1.65 (0.96)				
Efficiency	1.19 (0.79)				
Dependability	1.07 (0.81)				
Stimulation	1.18 (0.96)				
Novelty	0.85 (1.01)				
SCS[d]					
Baseline	3.05 (0.98)	—	—	—	—
Session 2	3.55 (1.16)	—	—	—	—
Comparison	—	—	2.219 (89)[e]	.03	−0.47
PHQ-8[f]					
Baseline	46.16[g]	—	—	—	—
Session 2	44.71[g]	—	—	—	—
Comparison	—	972	0.26[h]	.79	0.02[i]

[a]Participants (n=49 at session 1; n=42 at session 2) interacted with personalized stylized avatars. Measures included the UEQ, SCS, and PHQ-8. The study was conducted at The University of Queensland extended reality laboratory between 2023 and 2024 using a between-subjects design.

[b]UEQ: User Experience Questionnaire.

[c]Not applicable

[d]SCS: Self-Compassion Scale.

[e]*t* test value.

[f]PHQ-8: 8-item Patient Health Questionnaire.

[g]Mean rank value.

[h]*z* value.

[i]*r* value.

Impact on Depression (PHQ-8)

The Shapiro-Wilk test for normality revealed nonnormal distribution; thus, a Mann-Whitney *U* test was used. No significant difference was found between baseline and session 2 PHQ-8 scores, with a negligible effect size (see Table 1).

Study 3: Introduction of Virtual Mirror

A total of 38 participants completed session 1; 35 (92%) returned for Session 2. Gender of the participants was as follows:

- Session 1: 20 female, 18 male
- Session 2: 20 female, 15 male

Impact on User Experience (UEQ)

No changes to the UEQ tool have been made since study 2. As seen in Figures 8 and 9, study 3 scored higher in attractiveness, perspicuity, dependability, and stimulation compared to study 2. Three of six dimensions were rated “Excellent.” Given a between-studies comparison (Table S2 in Multimedia Appendix 1), UX differences were statistically significant; therefore, we found evidence that stylized avatar customization and virtual mirrors significantly influenced UX. These findings support RQ1, suggesting that stylized avatar customization and virtual mirrors significantly influenced UX.

Impact of Self-Compassion (SCS)

After removing five participants who did not complete the full SCS questionnaire, the Shapiro-Wilk test confirmed that data were normally distributed. A within-study, 2-tailed *t* test revealed no significant difference between baseline and session 2 SCS, with a small effect size (see Table 2). These results do not provide evidence that mirrors influenced self-compassion outcomes (RQ2).

XSL•FO
RenderX

Table . Summary statistics and internal comparisons from study 3 of a three-part mixed methods investigation into virtual reality self-compassion therapy for depressive symptoms.[a]

Measures	Mean (SD) or mean rank	Mann-Whitney *U* test	*t* test or *z*	*P* value (2-tailed)	Effect Size (Cohen *d* or *r*)
UEQ[b]		—[c]	—	—	—
Attractiveness	1.88 (0.92)				
Perspicuity	2.22 (0.65)				
Efficiency	1.42 (1.04)				
Dependability	1.76 (0.75)				
Stimulation	1.65 (1.02)				
Novelty	0.87 (1.04)				
SCS[d]					
Baseline	3.88 (1.33)	—	—	—	—
Session 2	4.09 (1.05)	—	—	—	—
Comparison	—	—	0.71 (63)[e]	.47	0.18
PHQ-8[f]					
Baseline	44.93[g]	—	—	—	—
Session 2	32.07[g]	—	—	—	—
Comparison	—	477.5	2.53[h]	.01	0.30[i]

[a]Participants (n=38 at session 1; n=35 at session 2) interacted with personalized stylized avatars and received visual feedback via a virtual mirror. Measures included the UEQ, SCS, and PHQ-8. The study was conducted at The University of Queensland extended reality laboratory between 2023 and 2024 using a between-subjects design.

[b]UEQ: User Experience Questionnaire.

[c]Not applicable.

[d]SCS: Self-Compassion Scale.

[e]*t* test value.

[f]PHQ-8: 8-item Patient Health Questionnaire.

[g]Mean rank value.

[h]*z* value.

[i]*r* value.

Impact on Depression (PHQ-8)

The data were not normally distributed; therefore, a Mann-Whitney *U* test found a significant reduction in PHQ-8 scores from baseline to session 2 (see Table 2). This result suggests a potential therapeutic benefit associated with the inclusion of stylized avatar customization and virtual mirrors.

Impact on Presence (SUS)

The mean SUS score in study 3 (mean 4.56, SD 1.58, 95% CI 4.02-5.10) overlapped with the benchmark SUS scores in Usoh et al [39], which reported a virtual (mean 3.8) and a real (mean 4.4) environment. This suggests that presence in our virtual mirror condition fell within a comparable range previously validated by Usoh et al[39] and did not significantly differ from real-world experience (t_{46}=0.30; P=.77; d=0.18). The lower bound of our CI exceeded the virtual benchmark, suggesting a relatively strong presence in the mirror condition.

Cross-Study Qualitative Analysis

Open-ended feedback was collected from 100 participants across all three studies; participants completed their assigned iVR therapy exposures prior to providing responses. Deductive and inductive thematic analysis was conducted on participants' open-ended responses. Participants were labeled P1-P20 (study 1), P21-P62 (study 2), and P63-P100 (study 3). The deductive thematic approach (Table S3 in Multimedia Appendix 1) includes codes corresponding to UEQ dimensions from Schrepp et al [46], which provided a structured lens through which to interpret qualitative feedback.

In parallel, an inductive analysis was performed to identify emergent themes not captured by the UEQ framework. These data-driven themes reflected patterns in participant experience that appeared to be consistent with experimental changes. The results of the inductive analysis are summarized in Table S4 in Multimedia Appendix 1.

XSL•FO
RenderX

Discussion

Overview

This study explored how stylized avatars, avatar customization, and virtual mirrors influenced UX (RQ1) and self-compassion outcomes (RQ2) across three iVR therapy conditions. In study 1, qualitative feedback indicated general acceptance of the stylized avatars. Study 2 showed a significant improvement in self-compassion outcomes compared to study 1. Study 3 produced significant improvements in four of the six UEQ dimensions, which suggests improved UX.

Summary of User Experience (RQ1)

For cross-study comparisons, see Table S2 in Multimedia Appendix 1. Study 3 yielded the highest attractiveness scores, with a statistically significant improvement compared to study 2 (P=.01). Qualitative data indicated this may be attributed to the enjoyment of avatar customization, which was introduced in study 2 and praised by participants in study 3. However, immersion-related feedback also emerged, such as "some of the environment designs feel too rigid…which breaks the immersion" (P78). Since Study 3 introduced virtual mirrors—known to support embodiment—this addition may explain the rise in UX.

On the perspicuity scale, study 3 again scored highest and significantly outperformed study 2 (P=.01). Participants highlighted the ease of learning, but across all studies, a lack of narrative context was noted. For example, participants reported difficulty understanding the avatars' emotional states, affecting their ability to engage in role-play scenarios.

Study 3 also significantly improved dependability scores (P<.001). The stylized avatars, rather than triggering discomfort, were perceived as appealing—"they're all cute" (P37)—and met participant expectations for a supportive environment. Thematic analysis suggested that the participants felt the system supported therapeutic outcomes: "I think this can make me feel relaxed" (P83).

Stimulation was significantly higher in study 3 than in study 2 (P=.04). Participants reported increased interest when viewing themselves delivering compassion, suggesting that the mirror feature enhanced engagement.

With respect to novelty, all studies received positive feedback—"innovative and unique" (P20)—but only study 1 reached an "above average" benchmark. Some participants reported reduced novelty due to restricted avatar movement, for example, "better if the users got a little more freedom during the simulation" (P56). This may indicate a design trade-off between structured therapeutic focus and exploratory freedom.

Interpreting User Experience

The UEQ served as the primary UX tool due to its alignment with iVR design goals for mental health. Though no significant differences were observed between studies 1 and 2, stylized avatars performed well compared to the UEQ benchmark, validating their continued exploration in therapeutic contexts. However, despite enhanced customization in study 2, UEQ scores did not improve significantly. One possible explanation may lie in elevated participant expectations (eg, limited skin tone options), which may have led to reduced satisfaction. This supports the idea that "less is more"; a limited but well-executed customization system may better serve therapeutic VR experiences.

Feedback across studies emphasized the importance of avatar individualization for user acceptance. Study 3's improvement in multiple UEQ dimensions suggests that mirrors—without being explicitly highlighted—supported user presence (and likely embodiment). "Attractiveness is a pure valence dimension" [46], that is, representative of the user's general experience. Attractiveness rose significantly in study 3, likely due to parallel improvements in perspicuity, dependability, and stimulation.

This connection is further supported by SUS findings. Significant improvements on items related to perceived realness—questions 2 and 3 of the SUS questionnaire [8]—and memory of the experience indicate that participants recalled the virtual mirror scenario as immersive and believable. Given the high scores (Table S5 in Multimedia Appendix 1), this kind of recall implies a strong sense of presence [39].

Taken together, study 3 suggests that the inclusion of virtual mirrors was a key turning point in enhancing iVR UX. While prior studies did not yield significant gains, mirrors may have "unlocked" the potential of the design. Whether these effects are solely due to mirrors or the cumulative impact of prior changes remains unclear, but no prior condition showed similar improvements.

Summary of Therapeutic Outcomes (RQ2)

In study 2, participants exposed to customizable stylized avatars over two sessions showed a significant increase in self-compassion (SCS). This supports the notion that stylistic, customized avatars in iVR interactions may enhance therapeutic benefit. However, these gains were not observed in study 3, despite the enhanced UX. This suggests that while mirrors improve UX and presence, they may not directly translate to better therapeutic outcomes.

Nevertheless, qualitative feedback highlighted that participants valued the ability to create avatars resembling themselves. It was noted that this customization improved their emotional connection to the experience in stage 2 of the experiment. For example, "customizing it to look like me did make the difference of how I perceived the audio replay" (P60). This statement implies that appearance alone was insufficient; users needed a visual feedback loop (eg, mirrors) to fully benefit from avatar customization.

Interestingly, mirror placement was not emphasized to participants yet still yielded a significant impact. However, incorrect or unnatural mirror placement disrupted immersion. For example, "[mirror] could be weird [to see] in the park environment" (P64) or "I see the mirror and then walk through it. I'm [a] ghost and that makes me a little bit scar[ed]" (P76). This underscores the importance of environmental congruence in design-embodied iVR therapy.

XSL•FO
RenderX

The Uncanny Valley

The appearance of stylized avatars was generally well received, supporting the idea that less realistic avatars can be effectively used in emotionally sensitive and immersive environments such as iVR self-compassion therapy, where avatar-owner resemblance is important. This finding aligns with prior research suggesting that stylized avatars may reduce the risk of UV effects while still fostering user connection [42,43,48]. Participants responded positively to avatars with cartoonish or humanoid qualities. “I don’t recommend realistic avatars as much. I think the more cartoonish appeal is better” (P41).

The UV model suggests that entities close to human-likeness [44], but imperfect, can cause discomfort. The use of stylized avatars in this study—located near “Humanoid Robot” on the affinity line—appears to have mitigated this risk. Mirrors further supported predictability and safety (dependability), suggesting that this combination may serve as a UX-enhancing strategy to avoid UV interference.

That said, UV was not the primary aim of this study but a theoretical background for design rationale. However, these reflections are offered as considerations for future iVR therapy design. It is important to note that violations of physical logic (eg, walking through mirrors, unnatural placement) also negatively impacted immersion and should be carefully addressed in future iterations.

Implications and Design Reflections

This study shows that stylized avatars, when paired with suitable UX enhancements such as virtual mirrors and avatar customization (along with other individualization concepts, environment selection, companion avatars, emotional behavior, etc), can offer stronger UX without losing therapeutic benefit, which might be an assumed consequence of selecting a lesser fidelity avatar. Pursuing hyperrealistic avatars without considering context may be counterproductive, as this ignores the Sisyphus-like limitation that avatars will risk falling back into the UV when realism is applied without careful design intent. This study gathers evidence to suggest that lower-fidelity avatars can be more appropriate in therapeutic settings but also that context matters when selecting avatar designs.

Limitation

This study used the SCS—a validated tool—but it has known psychometric limitations [49]. Additionally, participants were not filtered based on PHQ-8 scores; however, 74% had mild or higher range depression results. Future study iterations should include codesign and testing with clinical populations.

While the fidelity comparison was carefully controlled across studies, data were collected at different time points, which introduces the possibility of unmeasured societal or contextual differences. Future studies should aim to run control and experimental conditions concurrently to further minimize potential time-related confounds. Also, avatar fidelity was classified using a binary framework [36]. Future research may benefit from applying more continuous fidelity measures—such as user-rated realism scales or quantitative visual complexity metrics—to better capture subtle design differences.

Mirror engagement was inferred rather than directly measured, as the Meta Quest 2 headset lacks built-in eye tracking. However, mirrors were centrally positioned, and qualitative responses strongly suggest that participants did engage with them.

Finally, some tests were underpowered, which limited the ability to detect small-to-moderate effects. To help interpret these findings and provide additional insight, qualitative methods were included to support quantitative results.

Conclusion

If compassion means “to suffer together” [50], then does engaging in self-compassion mean suffering alone? Perhaps not with these avatars. In this iVR intervention, participants were comforted by their own stylized, customized avatars within immersive environments enhanced by virtual mirrors. Participants experienced self-compassion in a space tailored to their preferences, offering new insights into how avatar design and environmental cues shape therapeutic experiences.

This study sequentially investigated whether stylized avatars, avatar customization, and virtual mirrors influence UX and therapeutic outcomes in self-compassion VR therapy. Across three studies, results demonstrated that stylized avatars were well accepted and did not diminish self-compassion outcomes. Avatar customization improved self-compassion scores in one study, while virtual mirrors significantly improved UX across multiple dimensions. However, enhanced UX alone did not guarantee stronger therapeutic outcomes.

These findings suggest that individualized avatars and virtual mirrors can meaningfully enhance iVR experiences without triggering UV effects. More broadly, this study supports the feasibility of lower-fidelity avatars in mental health VR applications. Future research should explore these design strategies in clinical populations to better understand avatar-based therapy interventions.

Acknowledgments

The authors acknowledge Swaraj Vishwas Randhir and Vibhav Chitale for their contributions to the development of the individualized virtual reality environment and provision of the previous virtual environment. The authors would also like to acknowledge Lachlan Greig for providing consultation on psychological survey analysis and Chaitanya Dasi Luanzon for providing imagery. Generative artificial intelligence (ChatGPT; OpenAI) was used for grammar and language editing assistance during manuscript preparation. All content was reviewed and edited by the authors for accuracy and clarity. This research received no external funding. The authors conducted the study independently. A portion of the findings from studies 1 and 2 has been accepted for presentation at the Mixed/Augmented Reality for Mental Health workshop, part of the IEEE International Symposium on

XSL•FO
RenderX

Mixed and Augmented Reality–Adjunct, as "Customizing Stylized Avatars for Self-Compassion: Design Impacts in Virtual Reality Therapy" [50]. The present manuscript extends this work to include the complete set of three studies with expanded therapeutic framing, detailed methodology, and integrated qualitative findings.

Data Availability

Additional materials may be made available upon reasonable request to the corresponding author.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Tables presenting additional quantitative and qualitative outcomes across studies 1-3.
[DOCX File, 32 KB - formative_v9i1e71004_app1.docx]

References

1. Falconer CJ, Slater M, Rovira A, et al. Embodying compassion: a virtual reality paradigm for overcoming excessive self-criticism. PLoS ONE 2014;9(11):e111933. [doi: 10.1371/journal.pone.0111933] [Medline: 25389766]
2. Baghaei N, Stemmet L, Khaliq I, et al. Designing individualised virtual reality applications for supporting depression: a feasibility study. Presented at: EICS '21: Companion of the 2021 ACM SIGCHI Symposium on Engineering Interactive Computing Systems; Jun 8-11, 2021. [doi: 10.1145/3459926.3464761]
3. Halim I, Stemmet L, Hach S, et al. Individualized virtual reality for increasing self-compassion: evaluation study. JMIR Ment Health 2023 Oct 2;10:e47617. [doi: 10.2196/47617] [Medline: 37782537]
4. Slater M, Spanlang B, Sanchez-Vives MV, Blanke O. First person experience of body transfer in virtual reality. PLoS ONE 2010 May 12;5(5):e10564. [doi: 10.1371/journal.pone.0010564] [Medline: 20485681]
5. Blanke O, Metzinger T. Full-body illusions and minimal phenomenal selfhood. Trends Cogn Sci 2009 Jan;13(1):7-13. [doi: 10.1016/j.tics.2008.10.003] [Medline: 19058991]
6. Banakou D, Slater M. Embodiment in a virtual body that speaks produces agency over the speaking but does not necessarily influence subsequent real speaking. Sci Rep 2017 Oct 27;7(1):14227. [doi: 10.1038/s41598-017-14620-5] [Medline: 29079802]
7. Slater M. A note on presence terminology. UCL Computer Science. 2003. URL: http://www0.cs.ucl.ac.uk/research/vr/Projects/Presencia/ConsortiumPublications/ucl_cs_papers/presence-terminology.htm [accessed 2025-09-26]
8. Slater M, Usoh M, Steed A. Depth of presence in virtual environments. Presence Teleoperators Virtual Environ 1994 Jan;3(2):130-144. [doi: 10.1162/pres.1994.3.2.130]
9. Grassini S, Laumann K, Rasmussen Skogstad M. The use of virtual reality alone does not promote training performance (but sense of presence does). Front Psychol 2020;11:1743. [doi: 10.3389/fpsyg.2020.01743] [Medline: 32765384]
10. Wirth W, Hartmann T, Böcking S, et al. A process model of the formation of spatial presence experiences. Media Psychol 2007 May 15;9(3):493-525. [doi: 10.1080/15213260701283079]
11. Sanchez-Vives MV, Slater M. From presence to consciousness through virtual reality. Nat Rev Neurosci 2005 Apr;6(4):332-339. [doi: 10.1038/nrn1651] [Medline: 15803164]
12. Falconer CJ, Rovira A, King JA, et al. Embodying self-compassion within virtual reality and its effects on patients with depression. BJPsych Open 2016 Jan;2(1):74-80. [doi: 10.1192/bjpo.bp.115.002147] [Medline: 27703757]
13. Kilteni K, Groten R, Slater M. The sense of embodiment in virtual reality. Presence Teleoperators Virtual Environ 2012 Nov;21(4):373-387. [doi: 10.1162/PRES_a_00124]
14. Slater M, Wilbur S. A framework for immersive virtual environments (FIVE): speculations on the role of presence in virtual environments. Presence Teleoperators Virtual Environ 1997 Dec;6(6):603-616. [doi: 10.1162/pres.1997.6.6.603]
15. Wu B, Yu X, Gu X. Effectiveness of immersive virtual reality using head‐mounted displays on learning performance: a meta‐analysis. Brit J Educational Tech 2020 Nov;51(6):1991-2005. [doi: 10.1111/bjet.13023]
16. Du Q, Song Z, Jiang H, Wei X, Weng D, Fan M. LightSword: a customized virtual reality exergame for long-term cognitive inhibition training in older adults. Presented at: CHI '24: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems; May 11-16, 2024; Honolulu, HI. [doi: 10.1145/3613904.3642187]
17. Cuthbert R, Turkay S, Brown R. The effects of customisation on player experiences and motivation in a virtual reality game. Presented at: OZCHI'19: 31st Australian Conference on Human-Computer-Interaction; Dec 2-5, 2019; Fremantle, WA, Australia. [doi: 10.1145/3369457.3369475]
18. Kao D, Ratan R, Mousas C, Joshi A, Melcer EF. Audio matters too: how audial avatar customization enhances visual avatar customization. Presented at: CHI' 22: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems; Apr 29 to May 5, 2022; New Orleans, LA. [doi: 10.1145/3491102.3501848]

XSL•FO
RenderX

19. Altarteer S, Vassilis C, Harrison D. Product customisation: virtual reality and new opportunities for luxury brands online trading. Presented at: Web3D '16: The 21st International Conference on Web3D Technology; Jul 22-24, 2016; Anaheim, CA. [doi: 10.1145/2945292.2945317]
20. McArthur V. The UX of avatar customization. Presented at: CHI '17: CHI Conference on Human Factors in Computing Systems; May 6-11, 2017; Denver, CO. [doi: 10.1145/3025453.3026020]
21. Buck L, Young GW, McDonnell R. Avatar customization, personality, and the perception of work group inclusion in immersive virtual reality. Presented at: CSCW '23: Computer Supported Cooperative Work and Social Computing; Oct 14-18, 2023; Minneapolis, MN. [doi: 10.1145/3584931.3606992]
22. Birk MV, Mandryk RL. Combating attrition in digital self-improvement programs using avatar customization. Presented at: CHI '18: CHI Conference on Human Factors in Computing Systems; Apr 21-26, 2018; Montreal, QC. [doi: 10.1145/3173574.3174234]
23. Fitton I, Clarke C, Dalton J, Proulx MJ, Lutteroth C. Dancing with the avatars: minimal avatar customisation enhances learning in a psychomotor task. Presented at: CHI '23: CHI Conference on Human Factors in Computing Systems; Apr 23-28, 2023; Hamburg, Germany. [doi: 10.1145/3544548.3580944]
24. Freeman G, Zamanifard S, Maloney D, Adkins A. My body, my avatar: how people perceive their avatars in social virtual reality. Presented at: CHI '20: CHI Conference on Human Factors in Computing Systems; Apr 25-30, 2020; Honolulu, HI. [doi: 10.1145/3334480.3382923]
25. Waltemate T, Gall D, Roth D, Botsch M, Latoschik ME. The impact of avatar personalization and immersion on virtual body ownership, presence, and emotional response. IEEE Trans Vis Comput Graph 2018 Apr;24(4):1643-1652. [doi: 10.1109/TVCG.2018.2794629] [Medline: 29543180]
26. Tajadura-Jiménez A, Banakou D, Bianchi-Berthouze N, Slater M. Embodiment in a child-like talking virtual body influences object size perception, self-identification, and subsequent real speaking. Sci Rep 2017 Aug 29;7(1):9637. [doi: 10.1038/s41598-017-09497-3] [Medline: 28851953]
27. Sanchez-Vives MV, Spanlang B, Frisoli A, Bergamasco M, Slater M. Virtual hand illusion induced by visuomotor correlations. PLoS ONE 2010 Apr 29;5(4):e10381. [doi: 10.1371/journal.pone.0010381] [Medline: 20454463]
28. Tsakiris M, Haggard P. The rubber hand illusion revisited: visuotactile integration and self-attribution. J Exp Psychol Hum Percept Perform 2005 Feb;31(1):80-91. [doi: 10.1037/0096-1523.31.1.80] [Medline: 15709864]
29. Inoue Y, Kitazaki M. Virtual mirror and beyond: the psychological basis for avatar embodiment via a mirror. J Robotics Mechatronics 2021 Oct 20;33(5):1004-1012. [doi: 10.20965/jrm.2021.p1004]
30. Banakou D, Groten R, Slater M. Illusory ownership of a virtual child body causes overestimation of object sizes and implicit attitude changes. Proc Natl Acad Sci U S A 2013 Jul 30;110(31):12846-12851. [doi: 10.1073/pnas.1306779110] [Medline: 23858436]
31. Gonzalez-Franco M, Perez-Marcos D, Spanlang B, Slater M. The contribution of real-time mirror reflections of motor actions on virtual body ownership in an immersive virtual environment. Presented at: 2010 IEEE Virtual Reality Conference (VR); Mar 20-24, 2010; Boston, MA. [doi: 10.1109/VR.2010.5444805]
32. Slater M, Neyret S, Johnston T, et al. An experimental study of a virtual reality counselling paradigm using embodied self-dialogue. Sci Rep 2019 Jul 29;9(1):10903. [doi: 10.1038/s41598-019-46877-3] [Medline: 31358846]
33. Mal D, Wolf E, Dollinger N, Wienrich C, Latoschik ME. The impact of avatar and environment congruence on plausibility, embodiment, presence, and the proteus effect in virtual reality. IEEE Trans Vis Comput Graph 2023 May;29(5):2358-2368. [doi: 10.1109/TVCG.2023.3247089] [Medline: 37027615]
34. Wilson AC, Mackintosh K, Power K, Chan SWY. Effectiveness of self-compassion related therapies: a systematic review and meta-analysis. Mindfulness (N Y) 2019 Jun;10(6):979-995. [doi: 10.1007/s12671-018-1037-6] [Medline: 31354877]
35. Germer CK, Neff KD. Self-compassion in clinical practice. J Clin Psychol 2013 Aug;69(8):856-867. [doi: 10.1002/jclp.22021] [Medline: 23775511]
36. Weidner F, Boettcher G, Arboleda SA, et al. A systematic review on the visualization of avatars and agents in AR & VR displayed using head-mounted displays. IEEE Trans Vis Comput Graph 2023 May;29(5):2596-2606. [doi: 10.1109/TVCG.2023.3247072] [Medline: 37027741]
37. Neff KD. The self-compassion scale is a valid and theoretically coherent measure of self-compassion. Mindfulness (N Y) 2016 Feb;7(1):264-274. [doi: 10.1007/s12671-015-0479-3]
38. Kroenke K, Strine TW, Spitzer RL, Williams JBW, Berry JT, Mokdad AH. The PHQ-8 as a measure of current depression in the general population. J Affect Disord 2009 Apr;114(1-3):163-173. [doi: 10.1016/j.jad.2008.06.026] [Medline: 18752852]
39. Usoh M, Catena E, Arman S, Slater M. Using presence questionnaires in reality. Presence Teleoperators Virtual Environ 2000 Oct;9(5):497-503. [doi: 10.1162/105474600566989]
40. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
41. Stylized customizable avatars. Sunbox Games. URL: https://sunbox.games/asset/stylized-customizable-avatars/ [accessed 2025-09-17]

XSL•FO
RenderX

42. van Gisbergen MS, Sensagir I, Relouw J. How real do you see yourself in VR? The effect of user-avatar resemblance on virtual reality experiences and behaviour. In: Jung T, tom Dieck MC, Rauschnabel PA, editors. Augmented Reality and Virtual Reality: Changing Realities in a Dynamic World 2020:401-409. [doi: 10.1007/978-3-030-37869-1_32]
43. Lugrin JL, Latt J, Latoschik ME. Avatar anthropomorphism and illusion of body ownership in VR. Presented at: 2015 IEEE Virtual Reality (VR); Mar 23-27, 2015; Arles, France. [doi: 10.1109/VR.2015.7223379]
44. Mori M, MacDorman KF, Kageki N. The uncanny valley [from the field]. IEEE Robotics Automation Magazine 2012 Jun;19(2):98-100. [doi: 10.1109/MRA.2012.2192811]
45. Becker-Asano C, Ogawa K, Nishio S, Ishiguro H. Exploring the uncanny valley with geminoid HI-1 in a real-world application. Presented at: IADIS International Conference Interfaces and Human Computer Interaction; Jul 26-30, 2010; Freiburg, Germany URL: https://www.researchgate.net/profile/Christian_Becker-Asano/publication/229059888_Exploring_the_uncanny_valley_with_Geminoid_HI-1_in_a_real-world_application/links/0912f50b5da8f939c9000000.pdf [accessed 2025-09-26]
46. Schrepp M, Hinderks A, Thomaschewski J. Construction of a benchmark for the User Experience Questionnaire (UEQ). Int J Interactive Multimedia Artif Intelligence 2017;4(4):40. [doi: 10.9781/ijimai.2017.445]
47. Neff KD, Tóth-Király I. Self-Compassion Scale (SCS). In: Medvedev ON, Krägeloh CU, Siegert RJ, Singh NN, editors. Handbook of Assessment in Mindfulness Research: Springer; 2022:1-22. [doi: 10.1007/978-3-030-77644-2_36-1]
48. McMahan RP, Lai C, Pal SK. Interaction fidelity: the uncanny valley of virtual reality interactions. In: Lackey S, Shumaker R, editors. Virtual, Augmented and Mixed Reality: 8th International Conference, VAMR 2016, Held as Part of HCI International 2016, Toronto, Canada, July 17-22, 2016 Proceedings: Springer; 2016. [doi: 10.1007/978-3-319-39907-2_6]
49. Muris P, Otgaar H. The process of science: a critical evaluation of more than 15 years of research on self-compassion with the self-compassion scale. Mindfulness (N Y) 2020 Jun;11(6):1469-1482. [doi: 10.1007/s12671-020-01363-0]
50. Elliott TC, Yang Y, Knibbe J, Henry JD, Baghaei N. Customizing stylized avatars for self-compassion: design impacts in virtual reality therapy (forthcoming). Presented at: MARMH'25: Proceedings of the Mixed/Augmented Reality for Mental Health, Workshop, IEEE International Symposium on Mixed and Augmented Reality – Adjunct (ISMAR 2025); Oct 8-12, 2025; Daejeon, South Korea.

Abbreviations

CFT: compassion-focused therapy
iVR: individualized virtual reality
PHQ-8: 8-item Patient Health Questionnaire
SCS: Self-Compassion Scale
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
SUS: Slater-Usoh-Steed
UEQ: User Experience Questionnaire
UV: uncanny valley
UX: user experience
VR: virtual reality

Edited by A Mavragani; submitted 08.01.25; peer-reviewed by F Weidner, K Zempo; revised version received 19.08.25; accepted 30.08.25; published 02.10.25.

Please cite as:
Elliott TC, Yang Y, Knibbe J, Henry JD, Baghaei N
Avatar Customization and Embodiment in Virtual Reality Self-Compassion Therapy for Depressive Symptoms: Three-Part Mixed Methods Study
JMIR Form Res 2025;9:e71004
URL: https://formative.jmir.org/2025/1/e71004
doi:10.2196/71004

© Thomas C Elliott, Yanzhuo Yang, Jarrod Knibbe, Julie D Henry, Nilufar Baghaei. Originally published in JMIR Formative Research (https://formative.jmir.org), 2.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Internet-Delivered Psychological Treatment for Parents With Health Anxiety by Proxy: Replicated Randomized Single-Case Experimental Design

Katrine Ingeman[1,2], MSc, PhD; Ditte Hoffmann Frydendal[3], MSc, PhD; Lisbeth Frostholm[2,3], MSc, PhD; Ellen Bjerre-Nielsen[1], BA; Kaare Bro Wellnitz[3], MSc, PhD; Patrick Onghena[4], MSc, PhD; Kristi Wright[5], MSc, PhD; Charlotte Ulrikka Rask[1,2], MD, PhD, Prof Dr Med

[1]Department of Child and Adolescent Psychiatry, Research Unit, Aarhus University Hospital Psychiatry, Palle Juul-Jensens Boulevard 175, Aarhus, Denmark

[2]Department of Clinical Medicine, Faculty of Health, Aarhus University, Aarhus, Denmark

[3]Department of Functional Disorders, Aarhus University Hospital, Aarhus, Denmark

[4]Faculty of Psychology and Educational Sciences, KU Leuven, Leuven, Belgium

[5]Department of Psychology, Faculty of Arts, University of Regina, Regina, SK, Canada

Corresponding Author:
Katrine Ingeman, MSc, PhD
Department of Child and Adolescent Psychiatry, Research Unit, Aarhus University Hospital Psychiatry, Palle Juul-Jensens Boulevard 175, Aarhus, Denmark

Abstract

Background: Health anxiety by proxy is characterized by ruminations about severe illness in one's child that can cause severe distress in affected parents. Health anxiety by proxy may lead to repeated unnecessary medical consultations and checking the child's body for symptoms, as well as heightened attention to their child's behavior, sign of illness, and bodily symptoms. It has been hypothesized that health anxiety by proxy may pose a risk for transmission of maladaptive symptom coping and health anxiety from the parent to their child. In spite of this, no targeted treatment has previously been evaluated. Therefore, we developed an internet-delivered psychological treatment containing 8 modules based on cognitive behavioral therapy and acceptance and commitment therapy.

Objective: The objective of this study was to investigate the feasibility and effect of the internet-delivered treatment PROXY for parents with health anxiety by proxy.

Methods: A total of 4 participants with health anxiety by proxy entered a replicated randomized single-case experimental design. They were randomly allocated to a baseline period of 7 - 26 days before entering the 8-week treatment and 14 - 33 days follow-up phase. The primary outcomes were daily measures of anxiety, impact of anxiety, and value-based actions measured using 5 questions answered on a scale of 1 - 10 through a text-message link. The primary outcomes were analyzed using visual analysis and supplemented with statistical randomization tests. Secondary outcomes were standardized questionnaire measures of anxiety-related symptoms, experience of the treatment, and negative effects of the treatment reported using descriptive statistics for each participant individually.

Results: Visual and statistical analyses indicated that PROXY was an effective treatment for 2 participants as the primary outcomes changed in the preferable direction for both of them. The effect of PROXY was questionable for the remaining 2 participants, although visual analysis showed that the impact of anxiety decreased for one of them. The 2 participants with questionable effect also thought that the treatment was too short. All 4 participants were happy with the treatment, but 2 participants experienced that health anxiety for their own health deteriorated during treatment.

Conclusions: PROXY holds potential as a treatment for HA by proxy. However, more work is required to determine when and how PROXY should be introduced to parents with HA by proxy, particularly in relation to duration of treatment, possible comorbidities, and the need for findings to be replicated in larger groups.

Trial Registration: ClinicalTrials.gov NCT04830605; https://clinicaltrials.gov/study/NCT04830605

(JMIR Form Res 2025;9:e65396) doi:10.2196/65396

XSL•FO
RenderX

KEYWORDS

health anxiety by proxy; health anxiety; illness anxiety; internet-delivered treatment; single-case experimental design; parental worries; psychological treatment; internet-delivered; internet; parents; anxiety; depression; treatment; severe illness; child; youth; proxy; severe distress; distress; feasibility; anxiety-related symptoms; psychological; mental health

Introduction

Severe health anxiety (HA) is characterized by ruminative thoughts about harboring a serious disease [1]. It is classified as hypochondriasis in the ICD-10 (*International Statistical Classification of Diseases and Related Health Problems, Tenth Revision* [2]) and the *ICD-11* (*International Classification of Diseases, 11th Revision*) [3], or as Illness Anxiety Disorder or Somatic Symptom Disorder in the *DSM-5* (*Diagnostic and Statistical Manual of Mental Disorders* [Fifth Edition] [4]). HA is a debilitating disorder where illness worries are accompanied by a tendency to interpret natural bodily sensations as abnormal or signs of serious illness [5]. Excessive attention directed toward bodily symptoms is common and is typically accompanied by frequent medical examinations, which only provide temporary reassurance [6] and often lead to excessive health care use [7,8]. The disorder is common with a prevalence of 5.7% in the general adult population [9] and 9.5% in the primary sector [1].

Clinical observations and growing evidence suggest that parents can also suffer from HA on behalf of their children [10,11]. This has recently been referred to as HA by proxy [12]. Characteristic features are ruminations about and persistent fear that one's child has a serious illness that doctors are missing [12]. For some parents, this fear may lead to repeated medical consultations and checking the child's body for symptoms, as well as heightened attention to their child's behavior, signs of illness, and bodily symptoms [13,14]. While repeated medical consultations for a child may superficially resemble the clinical presentation of Munchausen syndrome by proxy [12], it is important to distinguish this from HA by proxy. Unlike Munchausen syndrome by proxy—now formally termed factitious disorder imposed on another—where a caregiver intentionally induces or fabricates illness in a child, HA by proxy does not involve deliberate deception or harm.

Research suggests that illness worries and illness behaviors can be transmitted from parents to their children [15-20]. The way parents model coping with their own symptoms [16-18] and how they respond to their children's health complaints can significantly shape the children's health attitudes and behaviors [15,21]. For example, a study on pain found that children report experiencing more pain when their parents focus more on their pain symptoms and thereby reinforces the child's negative illness perception (Levy [22]). Therefore, parental HA by proxy may contribute to the development of maladaptive illness behaviors, perceptions, and health-related worries in the child. In a previous qualitative study, we found that HA by proxy caused distress for the affected parents and impacted their daily life negatively as a result of attempts to control the anxiety or avoid situations that may trigger it [14]. Parents with HA by proxy reported struggling with beliefs that their worrying might be necessary in order to take proper responsibility for their child's health. At the same time, they realized that their worries led to maladaptive coping behavior, such as asking the child repeatedly about the child's well-being or looking for potential bodily illness signs such as bruises while bathing the child. Also, parents were concerned that their behavior could have negative consequences for their children, for example, that the children over time would develop health worries themselves [14].

To date, research on HA by proxy has been very limited, which means that the condition has not been systematically addressed in health care settings. One of the reasons for this is that until recently, no validated measure existed to address the construct [13,23]. Still, HA by proxy has been suggested as an important treatment target for parents with HA [24]. Currently, the evidence base for cognitive behavioral treatment of HA is strong [25-27]. In particular, acceptance and commitment therapy (ACT) and cognitive behavioral therapy (CBT) have also been investigated in web-based treatment formats with strong positive results for treating HA [28-31]. Given the conceptual and clinical overlap between HA and HA by proxy, internet-delivered treatment of HA by proxy based on these therapeutic approaches may also prove to be effective. However, some parents may only suffer from HA by proxy and not from HA in relation to themselves [11]. Furthermore, distinct cognitive, emotional, and behavioral processes—such as heightened vigilance toward a child's bodily symptoms, perceived responsibility for the child's health, and potential intergenerational transmission of HA—are not addressed in regular online HA treatment programs. Therefore, our research group has developed PROXY, the first internet-delivered psychological treatment program specifically designed for HA by proxy [32]. This study aims to conduct an initial assessment of the effect and feasibility of the new internet-delivered psychological treatment program PROXY [32] for parents with HA by proxy.

Methods

Design

Single-case experimental design (SCED) was used for this first testing of PROXY as the prevalence of HA by proxy is still unknown. This design allows for detailed, individualized analysis and is well-suited for exploring interventions in emerging or under-researched areas. It entails that (1) the intervention is experimentally manipulated across a series of separate phases and (2) the targeted behavior is measured repeatedly and frequently throughout the phases to investigate whether the changes occur as a function of the manipulated intervention [33].

This SCED study used a replicated randomized single-case phase design [34], meaning the procedure was replicated with more than a single individual. For each participant, the baseline length was computer-randomized to last between 7 and 26 days before the 8-week treatment period was initiated (56 d). Baseline

XSL•FO
RenderX

intervals were stratified into three categories: 7 - 11 days, 12 - 19 days, and 20 - 26 days, with an equal number of participants assigned to each stratum. Follow-up periods ranged from 14 to 33 days, depending on the length of the baseline phase, ensuring that the duration of each individual course was the same. Treatment outcomes were monitored daily by text messages throughout the study providing 96 measurement points in total (according to Kratochwill et al [35] the minimum number of data points are 5 per phase; see Figure 1). Detailed elaborations on design, measures, and intervention content have been reported elsewhere [32].

Figure 1. Study design. Adapted from Ingeman et al [32].

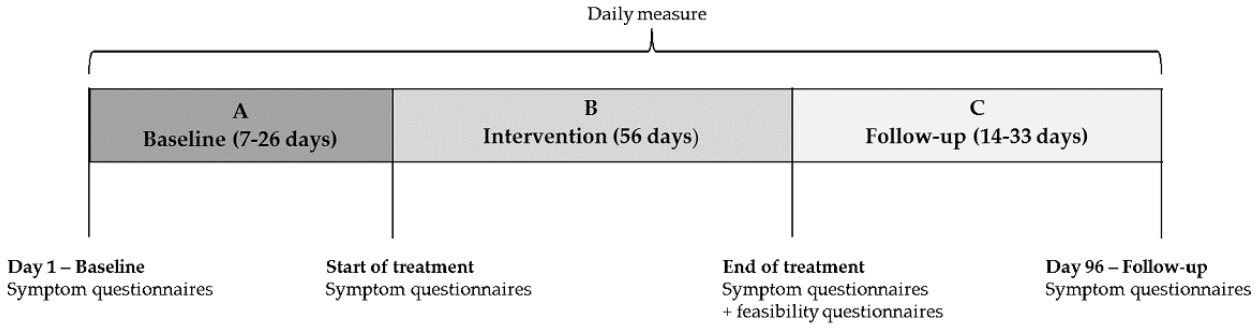

Daily measure includes 5 questions that participants answered via a link in a daily text message (see Textbox 1). Symptom questionnaires include the Health Anxiety by Proxy Scale, Adult Response to Children's Symptoms, the Pain Catastrophizing Scale—parent version, and the Whiteley Index 6 Revised. Feasibility questionnaires include the Internet Evaluation and Utility Questionnaire, the Negative Effects Questionnaire—20, and the Experience of Service Questionnaire.

Textbox 1. Items answered daily.

How well do the following statements describe your day?

1. I have had persistent worries about my child's health.
2. I have been worried that my child suffers from a serious physical illness.
3. I have experienced the need to seek reassurance by consulting a doctor, examining my child, searching for symptoms online, or something else.
4. My anxiety for my child's health has had an impact on my time together with my child.
5. I have done something in the past 24 hours that was important to me in spite of my anxiety.

Recruitment

Participants were recruited at the Research Clinic for Functional Disorders and Psychosomatics (the Research Clinic) at Aarhus University Hospital, Denmark. The recruitment period was from June 2022 until the end of January 2023.

Participants could self-refer to the treatment by sending a secured email to the Research Clinic. They would then receive a phone call and complete a questionnaire, including the Health Anxiety by Proxy Scale (HAPYS) [23], where the minimum score required to be considered for assessment was 26. This was a pragmatic cut-off score corresponding to the 75% percentile of the HAPYS score in parents without any severe somatic or psychiatric diagnoses. Participants could also be recruited at the research clinic in relation to the routine diagnostic assessments of referred patients for HA where clinical assessment of HA by proxy was included.

The inclusion criteria were: (1) clinically assessed to suffer from HA by proxy, which was defined as being severely impacted by persistent, recurring worries about one's child's health, fear that one's child has a serious illness that is being overlooked, and increased attention to bodily symptoms and physiological reactions that are misinterpreted as signs of serious illness in the child; (2) have children under 18 years without severe diagnoses or disabilities requiring regular hospital visits; and (3) be able to read, write, and speak Danish.

Exclusion criteria were: (1) comorbid diagnoses of substance abuse, bipolar disorder, psychotic disorders (ICD-10: F20-29) or autism spectrum disorders; (2) recently starting up medication for anxiety (within the last 2 months); and (3) suicidal risk.

Assessment

Clinical assessments were performed by psychologists at the Research Clinic using a short standardized diagnostic interview based on the Schedules for Clinical Assessment in Neuropsychiatry; World Health Organization [36,37] supplemented with assessment of HA by proxy using an interview version of the HAPYS (see section 2.5.2. for a detailed description of the HAPYS; see Ingeman et al [13] and Ingeman et al [23]).

The Intervention

The internet-delivered treatment program PROXY was developed by our research group and is a new therapist-assisted treatment targeting HA by proxy [32]. It is based on frameworks from CBT and ACT (for details on development procedure and content, see Ingeman et al [32]). PROXY consists of 8 modules that are made available weekly over 8 weeks (the first week, both module 1 and 2 were available). The modules focus on

XSL•FO
RenderX

providing psychoeducation about HA by proxy (module 1), identifying personal values (module 2), identifying control and avoidance behavior (module 3), acceptance of unpleasant inner states (module 4), exposure to anxiety-inducing situations (module 5), practicing self-compassion (module 6), seeking support from relatives (module 7), and knowledge about intergenerational transmission of health worries and preparation for continuing on their own and using what they have learned (module 8). The content of each module is presented by text, written exercises, audio files, and patient videos. In this study, the intervention was delivered over the internet in a specialized hospital setting supported by a psychologist who provided written individual feedback through an embedded text message system once a week. Phone calls were made if patients were inactive for more than a week or in need of specific help to move forward in the program.

Ethical Considerations

All participants received verbal and written information about the project before deciding to participate and gave their written consent to publish data individually and pseudo-anonymized. The participants did not receive any compensation. The project was approved by the Danish research ethics committee (1-10-72-296-20) and registered at the Danish data protection agency (1-16-02-921-17). Furthermore, the project was registered on ClinicalTrials.gov with the US National Library of Medicine on February 04, 2021 (NCT04830605).

Measures and Data Collection

All data was collected using the online data collection system Redcap [38,39].

Primary Outcome: Daily Measure

The participants received a daily text message at 7 PM containing a link to 3 items selected from the HAPYS to assess HA by proxy. The selection of these items was based on face validity and an indication of high sensitivity to change [13] (see items 1 - 3). The text message also contained an impact question (item 4) and a de novo formulated question about value-based action (item 5). Items were answered on a scale of 1 - 10 (1="does not describe it at all" and 10="describes it very well") but coded as 0 - 9 for analyses. Scores were summed for items 1 - 3 (range 0 - 27) while items 4 and 5 were analyzed individually (range 0 - 9). For item 5, the scoring was reversed, indicating that for this specific item alone, higher scores corresponded to a greater level of the desired behavior (see Textbox 1).

Secondary Outcomes

Symptom Questionnaires

The symptom questionnaires were answered 4 times during the study: at the beginning of the baseline period, at start of treatment, at end of treatment, and at follow-up except for feasibility measures, which were only answered at the end of treatment (see Figure 1).

The HAPYS assessed HA by proxy. It comprises 26 items rated on a 5-point scale (0=Not at all/never, 1=A little/rarely, 2=Some/sometimes, 3=Quite a lot/often and 4=A lot/most of the time; range 0 - 104, higher scores indicating more anxiety) and an impact section with additional six items rated on a 4-point rating scale (0=no, 1=yes, a little bit, 2=yes, quite a bit, and 3=yes, a great deal; range 0 - 18, higher scores indicating more impact) and two items providing information on time span for worries and if parents perceive their worries to be a problem [13]. In the evaluation of measurement properties, the HAPYS displayed 1-factor dimensionality as well as satisfactory convergent and known-groups validity along with test-retest reliability in a sample of parents with and without HA [23].

The Adult Response to Children's Symptoms consists of 4 subscales: Protect, Monitor, Minimize, and Distract, which together measure parental behavior when their child is having abdominal pain [40]. For this study, only the protect (13 items) and monitor (4 items) subscales were used, with "abdominal pain" replaced by "feels unwell" to accommodate broader use [24]. Both were scored 0=neverto 4=always (average total score: 0 - 4) with higher scores indicating more protective and monitoring parental behavior, respectively. The original ACRS was validated in a sample of parents of children with and without chronic pain. It displays satisfactory convergent and known-groups validity and a 4-factor structure corresponding to the mentioned subscales [41-43].

The Pain Catastrophizing Scale–parent version assesses parents' thoughts and feelings when their child has pain symptoms on 13 items (0=not at all to 4=extremely [range 0 - 52]; higher scores indicating more catastrophizing thoughts) [44]. The questionnaire has demonstrated a 3-factor model with good internal consistency, but the scale is still used with a one sum score [44]. The Danish version showed good face validity after cultural adaptation tested with parents of healthy children [45].

The Whiteley Index 6 Revised (WI-6-R) consists of 6 items to assess illness worries. In contrast to the original Whiteley-7 Index [46,47], WI-6-R excludes items concerning somatic symptoms and includes an item about obsessive illness rumination resulting in strengthened psychometric properties in a population-based sample [48]. Scores range from 0=not at all to 4=a great deal and are summed to a total score of 0 - 24 with higher scores representing higher levels of illness worry.

Feasibility Questionnaires

The feasibility of PROXY was assessed using three questionnaires that were answered at the end of treatment: (1) The Negative Effects Questionnaire-20 (NEQ), which measured factors related to negative effects of the treatment [49]; (2) the Internet Evaluation and Utility Questionnaire (IEUQ), which was used to evaluate experience of receiving internet-delivered treatment [50]; and (3) The Experience of Service Questionnaire (ESQ), which measures experience with treatment [51,52]. A modified version of the ESQ was used where 7 items were added from the Danish revised version developed by the Department of Psychology and Behavioral Sciences, Aarhus University, Denmark [53]. These additional questions evaluated participants' experience of how the intervention impacted their family and interaction with their child. A total of 3 questions related to physical settings were removed as they were irrelevant in relation to internet-based treatment. In the present study, all three feasibility measures are assessed and reported at

XSL•FO
RenderX

single-item level. Finally, the participants were asked to describe their experience of answering a text message every day.

Analyses

All graphs for visual analysis and further statistical analyses were performed using Stata 18 (StataCorp) [54]. Demographic characteristics were summarized for each participant individually. Only text messages answered on the same day as they were sent were considered valid and included in the analyses.

Data from the daily measures were submitted to structured visual analysis using the guideline from Lane and Gast [55]. For each participant, daily data was graphed for anxiety level, impact, and value-based action. In total, six main features were visually examined for each graph: (1) level change between phases; (2) variability of data points, both within and between phases (meaning how much the reported values changed and the minimum and maximum value in each phase); (3) trend in data; (4) immediacy of effect, which refers to whether the level changed when the intervention was introduced. Specifically, as a delayed effect was expected, we inspected whether the delayed effect was consistent across participants; (5) overlap of data points between phases, that is, the proportion of data points in the intervention phase that were not improved compared to baseline; and (6) consistency of data patterns across participants. The visual analysis was a collaborative effort by the first, second, third, and last authors.

The daily measure of anxiety was also submitted to statistical analyses: first, randomization tests [56-58] were conducted for each participant applying mean comparison of anxiety level before and after treatment entry as test statistic (MeanA–MeanBC=Test-Statistic). Phases B and C were combined because the follow-up phase did not include any new intervention, but still named B and C to indicate end of the treatment. The mean comparison was calculated for every possible randomization scenario for each participant. The *P* values calculated for each participant were combined using the Edgington additive method [59]. Second, the original study protocol indicated that response functions [60] would serve as a secondary randomization test statistic. However, this analysis was omitted due to the very poor fit visually estimated between observed trajectories and hypothesized response function (see Multimedia Appendix 1). Third, effect sizes were calculated for each participant using the nonoverlap method Tau-U, which controls for baseline trends [61]. Finally, data from the additional questionnaires were summarized using descriptive statistics for each participant individually. IEUQ and ESQ were analyzed and reported on item level, and all negative events reported by participants were summarized. Responses provided in free writing were documented as citations.

Results

Participants

A participant self-referred but withdrew the consent before clinical assessment. In addition, 2 participants self-referred and 6 patients were recruited at the research clinic. In total, 8 participants were assessed and met the inclusion criteria for participation. However, 3 participants were excluded due to their preference of receiving treatment for HA first or a change of mind regarding participation. Thus, 5 participants entered treatment. A participant dropped out due to technical issues with login (issues were unrelated to the technique at our platform) at the start of treatment, whereas 4 participants completed the treatment and data collection. None of these participants received any additional psychological treatment during their participation in the PROXY program. Table 1 shows the characteristics for the 4 included participants.

Table . Participant characteristics.

Participants	Age (years)	Sex	Biological children, n	Aged 10 years and younger, n	Education level	Recruitment
Participant 1	33	Male	2	2	>4 y	Clinic
Participant 2	51	Female	2	0	>4 y	Self-referral
Participant 3	34	Female	1	1	Skilled worker	Clinic
Participant 4	31	Female	3	3	>4 y	Self-referral

Daily measures

A total of 91 (23.7%) daily measures were missing equaling the text messages sent to participants.

Visual Analysis

Daily Anxiety

The effect on participants' anxiety levels did not occur immediately but was delayed. Particularly, toward the end of treatment and during follow-up (see part C in Figure 2), anxiety levels were lowest for all participants (see Figure 2). For Participants 1 and 4, there was a decrease in mean anxiety levels between baseline (see part A) and intervention (part B), and this effect was sustained during follow-up (part C) where anxiety leveled at zero.

XSL•FO
RenderX

Figure 2. Graphed data of the daily measures for participants 1 to 4 (P1-P4). Red line=mean value; black line=trend line; small vertical dotted line=module 5 that introduces exposure; large vertical dotted line=separates phases; A: baseline phase; B: intervention phase; and C: follow-up phase.

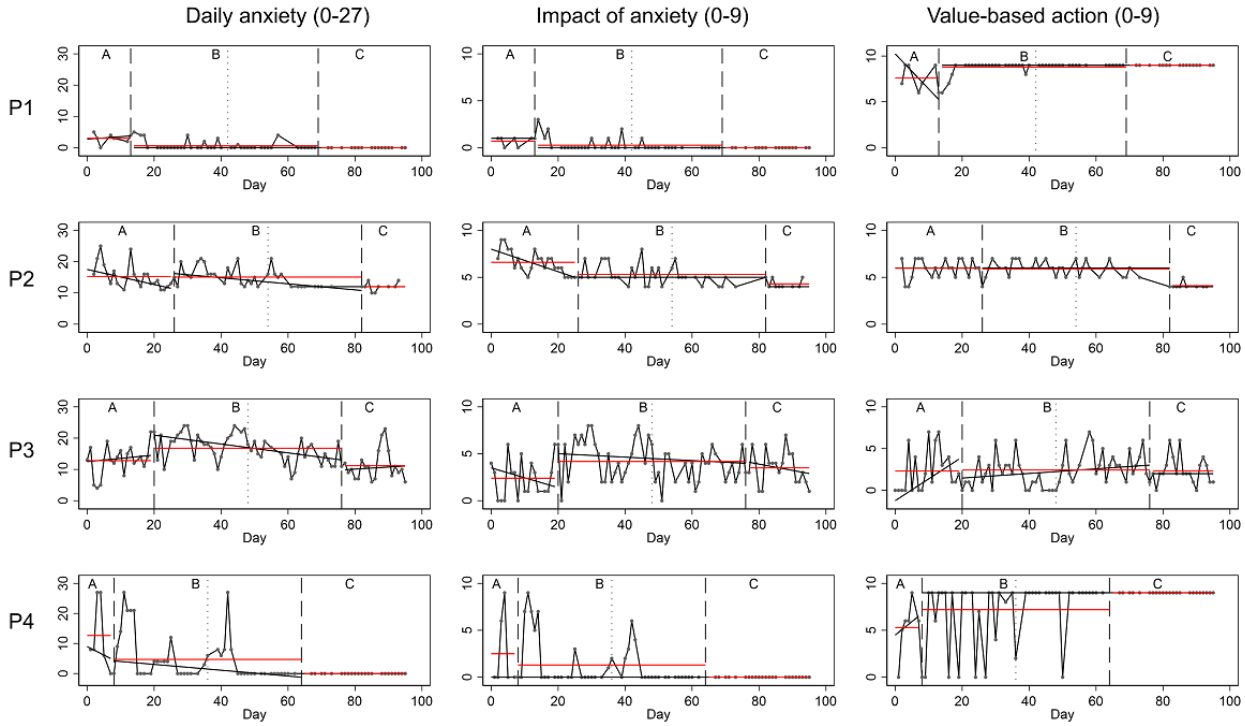

For participants 2 and 3, anxiety trend lines were decreasing during intervention (B), but there was no improvement of anxiety mean levels in the intervention period compared to their baseline mean levels (A). However, anxiety mean levels for participant 2 and participant 3 improved during follow-up (C) compared to intervention (B) as illustrated in the graph (see Figure 2) and demonstrated by the mean level changes between phases (see Table 2). These patterns are also reflected in the percentage of overlapping data. From phase A to B, all data points overlapped for participant 2 and participant 3, whereas only 22% and 75% of data points overlapped between phases B and C.

Table . Mean level change and percent of overlapping data for daily self-reported anxiety.

Participant	Mean level change		POD[a], %	
	A to B	B to C	A to B	B to C
Participant 1	−2.4	−0.6	19	100
Participant 2	−0.2	−3.1	100	22
Participant 3	4	−5.6	100	75
Participant 4	−8	−4.7	41	100

[a]POD: Percent of overlapping data. POD is the number of data points above the lowest value of the previous condition [55]. A perfect treatment effect has been achieved if POD equals 0% between phase A and B.

Considerable variability in data points was observed across the participants but also within participant 2, 3, and 4. This was particularly evident for participant 4 where the highest score was 27 and the lowest was zero for both phase A and phase B. The anxiety level was more stable for participant 1 where the lowest score was zero and the highest reported score was 5.

Daily Impact of Anxiety and Value-Based Actions

Data patterns for the impact of anxiety followed the patterns of anxiety for participant 1 and participant 4, with great variability for participant 4 and scores leveling at zero toward the end of treatment and during follow-up for both. Participant 1 and participant 4's reports of value-based actions also stabilized at the most extreme score of 9 towards the end of the study period.

For participant 2, the impact of anxiety improved during baseline (A) and continued to improve through intervention (B) and follow-up (C). Interestingly, this participant reported fewer value-based actions during follow-up (C) compared to baseline (A) and intervention (B).

The variability for impact of anxiety was very high for participant 3, and she did not report experiencing less impact after treatment (see Figure 2). Similarly, this participant did not report more value-based actions during treatment, and the level remained the same between phases, with high variability.

XSL•FO
RenderX

Statistical Analyses

Effect sizes and randomization tests using mean difference as test statistic are included for each participant in Table 3.

Table . Results of statistical analyses of daily measures of anxiety.

Participants	*P* value of randomization test[a] for mean difference	Effect size Tau-U[a]
Participant 1	.45	−0.71
Participant 2	≥.99	−0.01
Participant 3	.35	0.21
Participant 4	.50	−0.61
All combined	.69	__[b]

[a]Compares baseline to intervention and follow-up combined.

[b]Not available.

Additional Questionnaires

There were no missing data on the symptom or the feasibility questionnaires.

Symptom Questionnaires

Levels of HA by proxy and catastrophizing thoughts decreased for participants 1, 2, and 4 during treatment, while distinct improvements on impact of anxiety measured by the HAPYS were only evident for participant 1 and participant 4. It is less clear if levels of overprotective and monitoring behavior changed significantly for the participants. Levels of own HA were reduced by 50% for participant 4, but were reported unchanged by the remaining participants (see Table 4).

XSL•FO
RenderX

Table . Descriptive results for symptom questionnaires.

Questionnaire and participants	Time point			
	Baseline	Start[a]	EOT[b]	FU[c]
HA by proxy[d] 0 - 104				
Participant 1	58	51	20	27
Participant 2	74	79	56	48
Participant 3	79	60	74	64
Participant 4	73	72	10	12
Impact of HA by proxy[d] 0 - 18				
Participant 1	9	7	0	0
Participant 2	16	16	12	10
Participant 3	8	8	8	5
Participant 4	10	11	0	0
Overprotectiveness[e] 0 - 4				
Participant 1	1.7	1.0	0.6	0.7
Participant 2	2.7	2.6	2.0	1.8
Participant 3	2.3	2.0	1.5	2.2
Participant 4	2.1	1.8	1.3	2.4
Monitoring behavior[e] 0 - 4				
Participant 1	2.3	1.3	1.0	1.3
Participant 2	3.8	3.8	2.8	3.0
Participant 3	4.0	3.0	2.8	3.0
Participant 4	3.0	2.8	2.8	2.3
Pain catastrophizing thoughts[f] 0 - 52				
Participant 1	24	23	7	8
Participant 2	44	47	29	23
Participant 3	43	49	45	45
Participant 4	27	34	1	0
HA[g,h] 0 - 24				
Participant 1	11	18	11	17
Participant 2	19	22	18	22
Participant 3	21	23	24	22
Participant 4	24	24	24	12

[a]Start: Start of treatment.
[b]EOT: End of treatment.
[c]FU: Follow-up after 96 days from study entry.
[d]Health anxiety by proxy scale.
[e]Adult response to Child Symptoms—protect and monitor subscales.
[f]Pain catastrophizing scale-parent version.
[g]HA: health anxiety.
[h]Whiteley Index 6 items revised.

Participant Experience

Based on the ESQ, participants reported to be satisfied with the treatment and the help they had received (see Table 5). Participant 1 was the least satisfied, and participant 4, who had the most significant effect of the treatment, also had the highest

XSL•FO
RenderX

rating of PROXY on the ESQ. Citations from the ESQ's open-ended questions are reported in Multimedia Appendix 2.

Table . Results from the Experience of Service Questionnaire (answer 0=not true, 1=partly true, and 2=certainly true).

Item	Participant 1	Participant 2	Participant 3	Participant 4
The treatment helped me.	1	2	1	2
If a friend needed this sort of help, I would suggest to them to come here.	2	2	2	2
I have been given enough explanation about the help available here.	2	2	2	2
We feel better in our family now than before the treatment.	1	1	1	2
During the treatment, I became able to change my behavior toward my child in a positive way.	1	1	1	2
During treatment I gained a better understanding of my mental health.	2	1	2	2
I trusted my therapist.	2	2	2	2
The treatment made my child feel worse (reversed item).	0	0	0	0
The treatment made me feel worse (reversed item).	1	0	0	0
It was easy to talk to my therapist.	1	2	2	2
I was treated well by the therapist in charge of my treatment.	2	2	2	2
My views and worries were taken seriously.	2	2	2	2
I feel the people here know how to help me.	1	2	2	2
Overall, the help I have received here is good.	1	2	2	2

All participants reported on the IEUQ that the program was easy to use and that the content was useful and easy to understand. Participants stated that they liked receiving a text message every day and that they did not experience that the daily message affected their answers or their daily lives.

Negative Events

Negative events that occurred during the treatment period are listed in Multimedia Appendix 3. In free writing, participant 1 stated:

> *I think the treatment gave me a better understanding of my anxiety towards my children. But I feel like the treatment has not been good for my own anxiety towards myself. So I would have preferred to have targeted that first.* [Participant 1]

In addition, participant 2 said:

> *I feel like the treatment helped my anxiety towards my children. […] But my anxiety towards my own health has deteriorated.* [Participant 2]

Participant 3 stated:

> *I don't think that anything affected me negatively. I just think it has been tough but I believe you need to go through these feelings for it to get better.* [Participant 3]

Furthermore, participant 4 stated:

> *I had a flare-up of symptoms during the program. I think it was a result of lack of sleep and stress at home. It gave me more anxiety and a feeling of hopelessness, but I found the right track again. And then the symptoms disappeared.* [Participant 4]

XSL•FO
RenderX

Discussion

Principal Findings

To the best of our knowledge, this paper describes the only existing study exploring treatment specifically targeting HA by proxy. Using SCED, the aim of this study was to conduct an initial assessment of the effect and feasibility of the internet-based treatment program PROXY on parents' excessive worries about their children's health and to gather information about individual treatment courses. Overall, the results were not uniform as they indicated that PROXY may be an effective treatment of HA by proxy for two of the participants but questionable for the remaining 2 participants. Furthermore, feasibility results demonstrated general satisfaction with the treatment, but all the participants also reported at least one negative event caused by the treatment. A detailed discussion of the results will be presented in the following sections.

Based on the visual analysis, 2 of the participants (participants 1 and 4) improved on daily measures of anxiety, impact, and value-based action, which was supported by the large effect sizes for participant 1 (r=−0.71) and participant 4 (r=−0.61) on daily anxiety. This is in contrast to the randomization tests that indicated no treatment effect for either of the participants (P values from .35 to 1). This discrepancy might be explained by the choice of test statistics used in the randomization tests, as the mean difference is best suited for testing immediate effect of an intervention [34]. Nevertheless, the inconsistency between the results of the visual analysis and the randomization tests highlights the need for replication of the evaluation of PROXY as a treatment program for HA by proxy.

Based on the visual analysis and the effect size, it appears that participant 2 and participant 3 did not experience an effect of PROXY on their daily anxiety comparing the baseline to the intervention phase. However, participant 2 showed improvements in daily impact of anxiety during the intervention and a small reduction in impact scores on the HAPYS. Within the ACT framework, symptom reduction is not a goal in itself, and the experience of unpleasant inner states is only considered problematic if they impact living, values, and decision-making in a negative way [62]. Following this rationale, it could be argued that P2 did experience some positive effect of PROXY.

During the development of PROXY, we decided to introduce ACT components, such as values and acceptance early in module 2, in order to enhance motivation for performing the exposure exercises [32]. Previous studies have shown large effects of exposure-based CBT for HA [29,30,63]. Thus, in our study, it was expected that the introduction of exposure exercises would have a significant positive impact on participants' level of anxiety. However, this was not evident from the visual analysis. Exposure is most likely still effective in treating HA by proxy, but the effect might be more gradual. Similar results were found in a SCED study of exposure-based therapy for HA where the specific immediate effects of introducing exposure were also not evident from the visual analyses [64]. In future studies, it may be important to further investigate treatment courses in detail to understand the specific active treatment mechanisms in internet-based anxiety programs like PROXY.

Results from the secondary outcomes measuring symptoms of HA by proxy, catastrophizing thoughts, and overprotective and monitoring behaviors and own HA were also heterogeneous. All four participants improved on the HAPYS (0-104) with between 15 and 61 points. Similarly, patterns of positive change in catastrophizing thoughts were observed during the study period. In fact, the validation study of the HAPYS supported a strong correlation between catastrophizing thoughts and HA by proxy [23]. The present results provide additional support for a relationship between parental HA by proxy and catastrophizing, as they suggest that both seem to be targeted by PROXY. However, it is important to be careful not to generalize the results of these secondary outcomes as they are based on pre-post measures from just 4 participants.

All the participants described that they were happy with the treatment and stated that it had helped them with their distress caused by HA by proxy. However, participant 2 and participant 3 indicated that the treatment period of 8 weeks was too short (see Multimedia Appendix 2). In addition, their daily anxiety level did not improve until follow-up. We aimed for a brief treatment program when deciding for 8 weeks. In comparison, existing internet-delivered treatments for HA all had a duration of 12 weeks [28-31]. Durations of other internet-delivered treatments of anxiety ranged between 5 and 10 weeks [65]. Based on results and feedback from the patients, considerations about the duration of the treatment are important for future use of PROXY.

All participants reported at least one negative event caused by PROXY. Participant 2 in particular reported several negative events, including unpleasant feelings, dependence on treatment, and deterioration of anxiety. Other SCED studies investigating psychological treatment of generalized anxiety disorder and HA also report negative effects [64,66]. In general, temporary distress from psychological interventions is common as a result of confronting distressing thoughts and feelings [67-69]. However, for participant 2, the distress did not seem to be temporary, and she even reported in the NEQ that she was starting to think that her HA by proxy could not be made any better. At the same time, she stated in free text in the ESQ that she felt "seen," "understood," and "could recognize almost everything described." This exemplifies that patients can be satisfied with the treatment and still experience negative events caused by that same treatment. Clinically, we would expect some discomfort of participation in anxiety treatment, because patients often use maladaptive strategies to avoid or control their anxiety. In treatment, these strategies are targeted and altered, for example, through exposure. This will often increase anxiety symptoms temporarily. In accordance with this, an internet treatment of HA found increased anxiety as the most frequent negative effect during treatment; however, there was no association between the experience of negative effects of treatment and symptom deterioration [28]. Overall, this underlines the importance of assessing both negative events and patients' experience of receiving psychological treatment, as patients may be satisfied with treatment, experience effects, and still report negative effects.

Interestingly, both participant 1 and participant 2 stated that their own HA deteriorated during the treatment. Engaging in

XSL•FO
RenderX

psychological treatment in general can be quite stressful, and it cannot be excluded that engaging in a treatment where the patients were exposed to illness-related content may have triggered their own HA. In addition, the results seem to indicate that HA and HA by proxy could be seen as two separate anxiety conditions with symptoms that do not follow each other or respond equally to treatment. Furthermore, it raises important considerations regarding when and how to introduce PROXY when parents suffer from both HA and HA by proxy.

Strengths and Limitations

The reporting of this study followed the recommendation of the Single-Case Reporting Guideline In BEhavioural Interventions statement [70] and has a number of strengths. First, the visual analysis was supplemented by statistical analyses. Statistical analyses in SCED research have previously been considered controversial [55], but in recent years, several researchers have advocated for the supplement of statistical analyses [71-73]. Second, randomizing the start of the intervention—and thereby, the length of the baseline phase—for each participant is a key strength of the study, as it enhances internal validity [56,58]. Third, the combination of daily measures, feasibility measures, and questionnaires to obtain information on participants' experience as well as effect of PROXY was a considerable strength of the study, and finally, the participants underwent systematic and diagnostic assessments for other mental health disorders before inclusion in the project.

There are also some methodological limitations that warrant consideration. The first concerns the recruitment strategy, which entailed recruitment of participants who also suffered from HA. This overlap between HA and HA by proxy complicates the attribution of outcomes specifically to HA by proxy, as it is unclear whether observed changes (including the deterioration in own HA reported by two participants) reflect effects of the intervention on HA by proxy, on general HA, or on the interaction between the two. It is also possible that participants with high levels of own HA may have had less motivation or ability to focus on the PROXY treatment. Future studies with larger samples and stratified inclusion are needed to disentangle treatment effects on HA by proxy from those on general HA. The second limitation concerns the causal relationship between the changes we see in the patients' anxiety level and the intervention. Confounding factors that are time-related can be statistically controlled for by randomly determining the intervention start point (the baseline length) in a SCED [35]. It remains unknown whether the baseline in the present study was long enough to conclude that the changes seen were an effect of the intervention and not due to maturation or spontaneous improvements in symptoms, natural fluctuations, or other factors. In the present study, the participants were randomized to one of 20 possible treatment entries. If the observed value from the test statistic was the single most extreme value obtained out of all the randomization possibilities, the *P* value would be 1/20=.05. This means that in our study, the smallest possible *P* value was .05. Therefore, randomization tests require a large amount of randomization possibilities to obtain a small *P* value. This means that when applying the method to SCED for determining baseline lengths, it also becomes an ethical question of how long participants need to wait before they enter the treatment, and for how long it is feasible to obtain daily measures. In our study, the percentage of missing data was already at 23.7% which must be considered a limitation in itself. Further limitations pertain to the reliability and validity of the visual analysis of the daily measures. In general, researchers agree on the six features assessed in this study [74,75]. Nevertheless, visual analysis inherently involves an element of subjectivity. Some researchers argue that the key to visual analysis is to maintain structure [74,76]. Other researchers attempt to remove subjective influence by incorporating statistical calculations and rules in the visual analysis (eg, 55,75,77). A final limitation is the fact that the daily measures relied on self-report, which has been raised by several researchers as a potential issue in SCED studies because the validity of daily self-report measures has often not been investigated and because of the risk of response bias like social desirability [78-84]. The inclusion of clinician-based evaluations of anxiety levels, in addition to self-reports, would have strengthened the assessment of the primary outcome. Yet, the use of self-report measures in SCED studies of interventions for anxiety disorders is not uncommon [66,81,85]. In our study, the most suitable method for assessing HA by proxy was through self-report measures as we were interested in exploring the subjective experience of parents' daily anxiety level, impact of anxiety, and value-based actions. Summing up, when replicating the SCED study, it is important to consider how to strengthen the validity and reliability of the daily self-report measures; to supplement the visual analysis with further statistical analyses; and to increase the internal validity of the study. The latter can be achieved, for example, by having participants receive treatment concurrently in time and by extending the maximum baseline length [33].

Conclusions and Future Research Perspectives

To the best of our knowledge, PROXY is currently the only existing treatment for HA by proxy. While findings from this initial study are mixed, they suggest that the intervention may be feasible. Key questions remain regarding effectiveness, optimal treatment duration, timing, and its use alongside treatment for primary HA. Future research on treatment of HA by proxy should focus on recruiting a larger sample and introducing methodological refinements to enable stronger conclusions about treatment effects.

Acknowledgments

The project group would like to thank the participants and therapist Line Sørensen. The project was funded by Aarhus University, Central Denmark Region (A2614), the AP Moller Foundation (18-L-0180), and the Aase and Ejnar Danielsen Foundation (19-10-0030). No generative artificial intelligence was used in the preparations of this manuscript.

XSL•FO
RenderX

Authors' Contributions

Conceptualization: KL, CUR, LF, DHF
Methodology: KL, KBW, PO
Investigation: EB-N, DHF, LF, CUR
Formal analysis: KL, LF, DHF, CUR, KW
Writing - original draft: KL
Writing - review and editing: DHF, LF, EB-N, KBW, PO, KW, CUR
Funding acquisition: KL
Supervision: DHF, LF, PO, KW, CUR

Conflicts of Interest

None declared.

Multimedia Appendix 1
Response function for participants in PROXY.
[DOCX File, 110 KB - formative_v9i1e65396_app1.docx]

Multimedia Appendix 2
Experience of Service Questionnaire answers to open-ended questions at the end of treatment.
[DOCX File, 22 KB - formative_v9i1e65396_app2.docx]

Multimedia Appendix 3
List of reported negative events at the end of treatment. Participants 1, 2, 3, and 4 indicate the participants who reported having experienced the particular event.
[DOCX File, 13 KB - formative_v9i1e65396_app3.docx]

References

1. Fink P, Ørnbøl E, Toft T, Sparle KC, Frostholm L, Olesen F. A new, empirically established hypochondriasis diagnosis. Am J Psychiatry 2004 Sep;161(9):1680-1691. [doi: 10.1176/appi.ajp.161.9.1680] [Medline: 15337660]
2. Collaborating Centres for Classification of D, World Health Organization. ICD-10: International Statistical Classification of Diseases and Related Health Problems, 10th edition: WHO; 1992:3.
3. World Health Organization. International Classification of Diseases for Mortality and Morbidity Statistics (11th Revision) 2018.
4. American Psychiatric Association. Diagnostic and statistical manual of mental disorders. In: A American Psychiatric and DSMTF: American Psychiatric Association; 2013. [doi: 10.1176/appi.books.9780890425596]
5. Marcus DK, Gurley JR, Marchi MM, Bauer C. Cognitive and perceptual variables in hypochondriasis and health anxiety: a systematic review. Clin Psychol Rev 2007 Mar;27(2):127-139. [doi: 10.1016/j.cpr.2006.09.003] [Medline: 17084495]
6. Asmundson GJG, Abramowitz JS, Richter AA, Whedon M. Health anxiety: current perspectives and future directions. Curr Psychiatry Rep 2010 Aug;12(4):306-312. [doi: 10.1007/s11920-010-0123-9] [Medline: 20549396]
7. Tyrer P, Eilenberg T, Fink P, Hedman E, Tyrer H. Health anxiety: the silent, disabling epidemic. BMJ 2016 Apr 25;353:i2250. [doi: 10.1136/bmj.i2250] [Medline: 27112356]
8. Fink P, Ørnbøl E, Christensen KS. The outcome of health anxiety in primary care. A two-year follow-up study on health care costs and self-rated health. PLoS ONE 2010 Mar 24;5(3):e9873. [doi: 10.1371/journal.pone.0009873] [Medline: 20352043]
9. Sunderland M, Newby JM, Andrews G. Health anxiety in Australia: prevalence, comorbidity, disability and service use. Br J Psychiatry 2013 Jan;202(1):56-61. [doi: 10.1192/bjp.bp.111.103960] [Medline: 22500013]
10. Thorgaard MV, Frostholm L, Walker L, et al. Health anxiety by proxy in women with severe health anxiety: A case control study. J Anxiety Disord 2017 Dec;52:8-14. [doi: 10.1016/j.janxdis.2017.09.001] [Medline: 28950218]
11. Salama HZ, Alnajjar YA, AbuHeweila MO, Dukmak ON, Ikhmayyes I, Saadeh S. Health Anxiety by Proxy Disorder: A Case Report. Cureus 2023 Jan;15(1):e34058. [doi: 10.7759/cureus.34058] [Medline: 36824537]
12. Lockhart E. Health anxiety in children and parents. BMJ 2016 May 24;353:i2891. [doi: 10.1136/bmj.i2891] [Medline: 27222022]
13. Ingeman K, Frostholm L, Frydendal DH, et al. A new measure of excessive parental worries about children's health: Development of the Health Anxiety by Proxy Scale (HAPYS). Nord J Psychiatry 2021 Oct;75(7):523-531. [doi: 10.1080/08039488.2021.1900389] [Medline: 33724904]
14. Ingeman K, Hulgaard DR, Rask CU. Health anxiety by proxy - through the eyes of the parents. J Child Health Care 2024 Mar;28(1):22-36. [doi: 10.1177/13674935221095648] [Medline: 35510964]

XSL•FO
RenderX

15. Walker LS, Williams SE, Smith CA, Garber J, Van Slyke DA, Lipani TA. Parent attention versus distraction: Impact on symptom complaints by children with and without chronic functional abdominal pain. Pain 2006;122(1):43-52. [doi: 10.1016/j.pain.2005.12.020]
16. Köteles F, Freyler A, Kökönyei G, Bárdos G. Family background of modern health worries, somatosensory amplification, and health anxiety: A questionnaire study. J Health Psychol 2015 Dec;20(12):1549-1557. [doi: 10.1177/1359105313516661] [Medline: 24406331]
17. Wright KD, Reiser SJ, Delparte CA. The relationship between childhood health anxiety, parent health anxiety, and associated constructs. J Health Psychol 2017 Apr;22(5):617-626. [doi: 10.1177/1359105315610669] [Medline: 26466851]
18. Craig TKJ, Cox AD, Klein K. Intergenerational transmission of somatization behaviour: a study of chronic somatizers and their children. Psychol Med 2002 Jul;32(5):805-816. [doi: 10.1017/s0033291702005846] [Medline: 12171375]
19. Thorgaard MV, Frostholm L, Rask CU. Childhood and family factors in the development of health anxiety: A systematic review. Child Health Care 2018 Apr 3;47(2):198-238. [doi: 10.1080/02739615.2017.1318390]
20. Svestkova A, Kvardova N, Smahel D. Health Anxiety in Adolescents: The Roles of Online Health Information Seeking and Parental Health Anxiety. J Child Fam Stud 2024 Apr;33(4):1083-1094. [doi: 10.1007/s10826-023-02689-8]
21. Jungmann SM, Witthöft M. Medically unexplained symptoms in children and adolescents: Illness-related self-concept and parental symptom evaluations. J Behav Ther Exp Psychiatry 2020 Sep;68:101565. [doi: 10.1016/j.jbtep.2020.101565] [Medline: 32171996]
22. Levy RL. Exploring the intergenerational transmission of illness behavior: from observations to experimental intervention. Ann Behav Med 2011 Apr;41(2):174-182. [doi: 10.1007/s12160-010-9254-9] [Medline: 21170690]
23. Ingeman K, Wright KD, Frostholm L, Frydendal DH, Ørnbøl E, Rask CU. Measurement properties of the Health Anxiety by Proxy Scale (HAPYS): A new questionnaire to assess parents' worries about their child's health. J Psychosom Res 2024 Jan;176:111555. [doi: 10.1016/j.jpsychores.2023.111555] [Medline: 38039620]
24. Thorgaard MV. Health anxiety and illness behaviour in children of mothers with severe health anxiety. Dan Med J 2017 May;64(5):B5365. [Medline: 28552100]
25. Cooper K, Gregory JD, Walker I, Lambe S, Salkovskis PM. Cognitive behaviour therapy for health anxiety: A systematic review and meta-analysis. Behav Cogn Psychother 2017 Mar;45(2):110-123. [doi: 10.1017/S1352465816000527] [Medline: 28229805]
26. Olatunji BO, Kauffman BY, Meltzer S, Davis ML, Smits JAJ, Powers MB. Cognitive-behavioral therapy for hypochondriasis/health anxiety: a meta-analysis of treatment outcome and moderators. Behav Res Ther 2014 Jul;58:65-74. [doi: 10.1016/j.brat.2014.05.002] [Medline: 24954212]
27. Maass U, et al. Psychological Interventions for Health Anxiety and Somatic Symptoms: A Systematic Review and Meta-Analysis. Zeitschrift für Psychologie 2020;228(2):68-80. [doi: 10.1027/2151-2604/a000400]
28. Hoffmann D. Efficacy of internet-delivered acceptance and commitment therapy for severe health anxiety: results from a randomized, controlled trial. Psychol Med 2020:1-11.
29. Hedman E, Andersson G, Andersson E, et al. Internet-based cognitive-behavioural therapy for severe health anxiety: randomised controlled trial. Br J Psychiatry 2011 Mar;198(3):230-236. [doi: 10.1192/bjp.bp.110.086843] [Medline: 21357882]
30. Hedman E, Axelsson E, Görling A, et al. Internet-delivered exposure-based cognitive-behavioural therapy and behavioural stress management for severe health anxiety: randomised controlled trial. Br J Psychiatry 2014 Oct;205(4):307-314. [doi: 10.1192/bjp.bp.113.140913] [Medline: 25104835]
31. Newby JM, Smith J, Uppal S, Mason E, Mahoney AEJ, Andrews G. Internet-based cognitive behavioral therapy versus psychoeducation control for illness anxiety disorder and somatic symptom disorder: A randomized controlled trial. J Consult Clin Psychol 2018 Jan;86(1):89-98. [doi: 10.1037/ccp0000248] [Medline: 29172593]
32. Ingeman K, Frostholm L, Wellnitz KB, et al. Internet-Delivered Therapy for Parents With Health Anxiety by Proxy: Protocol for a Single-Case Experimental Design Study. JMIR Res Protoc 2023 Nov 24;12:e46927. [doi: 10.2196/46927] [Medline: 37999936]
33. Tate RL, Perdices M, Rosenkoetter U, et al. The Single-Case Reporting Guideline In BEhavioural Interventions (SCRIBE) 2016: Explanation and elaboration. Arch Sci Psychol 2016;4(1):10-31. [doi: 10.1037/arc0000027]
34. Michiels B, Onghena P. Randomized single-case AB phase designs: Prospects and pitfalls. Behav Res Methods 2019 Dec;51(6):2454-2476. [doi: 10.3758/s13428-018-1084-x] [Medline: 30022457]
35. Kratochwill TR, et al. Single-case designs technical documentation. In: What Works Clearinghouse 2010.
36. World Health Organization. Schedules for clinical assessment in neuropsychiatry: version 2. In: D.oM Health: American Psychiatric Press; 1994.
37. Petersen MW, Schröder A, Jørgensen T, et al. RIFD - A brief clinical research interview for functional somatic disorders and health anxiety. J Psychosom Res 2019 Jul;122:104-111. [doi: 10.1016/j.jpsychores.2019.04.001] [Medline: 30955913]
38. Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: Building an international community of software platform partners. J Biomed Inform 2019 Jul;95:103208. [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]

XSL•FO
RenderX

39. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381. [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
40. Van Slyke DA, Walker LS. Mothers' responses to children's pain. Clin J Pain 2006 May;22(4):387-391. [doi: 10.1097/01.ajp.0000205257.80044.01] [Medline: 16691093]
41. Walker LS, Levy RL, Whitehead WE. Validation of a measure of protective parent responses to children's pain. Clin J Pain 2006 Oct;22(8):712-716. [doi: 10.1097/01.ajp.0000210916.18536.44] [Medline: 16988567]
42. Noel M, Palermo TM, Essner B, et al. A developmental analysis of the factorial validity of the parent-report version of the Adult Responses to Children's Symptoms in children versus adolescents with chronic pain or pain-related chronic illness. J Pain 2015 Jan;16(1):31-41. [doi: 10.1016/j.jpain.2014.10.005] [Medline: 25451623]
43. Claar RL, Guite JW, Kaczynski KJ, Logan DE. Factor structure of the Adult Responses to Children's Symptoms: validation in children and adolescents with diverse chronic pain conditions. Clin J Pain 2010 Jun;26(5):410-417. [doi: 10.1097/AJP.0b013e3181cf5706] [Medline: 20473048]
44. Goubert L, Eccleston C, Vervoort T, Jordan A, Crombez G. Parental catastrophizing about their child's pain. The parent version of the Pain Catastrophizing Scale (PCS-P): a preliminary validation. Pain 2006 Aug;123(3):254-263. [doi: 10.1016/j.pain.2006.02.035] [Medline: 16644128]
45. Bjerre-Nielsen E, et al. The pain catastrophizing scale for parents" – et spørgeskema til at vurdere de tanker og følelser, forældre kan have, når deres barn har smerter. : P - Psykologernes Fagmagasin; 2020.
46. Pilowsky I. Dimensions of hypochondriasis. Br J Psychiatry 1967 Jan;113(494):89-93. [doi: 10.1192/bjp.113.494.89] [Medline: 6029373]
47. Fink P, Ewald H, Jensen J, et al. Screening for somatization and hypochondriasis in primary care and neurological in-patients: a seven-item scale for hypochondriasis and somatization. J Psychosom Res 1999 Mar;46(3):261-273. [doi: 10.1016/s0022-3999(98)00092-0] [Medline: 10193917]
48. Carstensen TBW, Ørnbøl E, Fink P, et al. Detection of illness worry in the general population: A specific item on illness rumination improves the Whiteley Index. J Psychosom Res 2020 Nov;138:110245. [doi: 10.1016/j.jpsychores.2020.110245] [Medline: 32950761]
49. Rozental A, Kottorp A, Forsström D, et al. The Negative Effects Questionnaire: psychometric properties of an instrument for assessing negative effects in psychological treatments. Behav Cogn Psychother 2019 Sep;47(5):559-572. [doi: 10.1017/S1352465819000018] [Medline: 30871650]
50. Ritterband LM, Ardalan K, Thorndike FP, et al. Real world use of an Internet intervention for pediatric encopresis. J Med Internet Res 2008 Jun 30;10(2):e16. [doi: 10.2196/jmir.1081] [Medline: 18653440]
51. Brown A, Ford T, Deighton J, Wolpert M. Satisfaction in child and adolescent mental health services: translating users' feedback into measurement. Adm Policy Ment Health 2014 Jul;41(4):434-446. [doi: 10.1007/s10488-012-0433-9] [Medline: 22829193]
52. Attride-Stirling J. Development of methods to capture users' views of CAMHS in clinical governance reviews. : corc. uk; 2002 URL: https://www.corc. uk. net/media/1215/chi_projectevaluationreport.pdf [accessed 2025-09-26]
53. Stjerneklar S, Hougaard E, Nielsen AD, Gaardsvig MM, Thastum M. Internet-based cognitive behavioral therapy for adolescents with anxiety disorders: A feasibility study. Internet Interv 2018 Mar;11(C):30-40. [doi: 10.1016/j.invent.2018.01.001] [Medline: 30135757]
54. StataCorp. Stata statistical software: release 18. : StataCorp LLC; 2023.
55. Lane JD, Gast DL. Visual analysis in single case experimental design studies: brief review and guidelines. Neuropsychol Rehabil 2014;24(3-4):445-463. [doi: 10.1080/09602011.2013.815636] [Medline: 23883189]
56. Edgington ES. Randomized single-subject experimental designs. Behav Res Ther 1996 Jul;34(7):567-574. [doi: 10.1016/0005-7967(96)00012-5] [Medline: 8826764]
57. Edgington ES. Randomization tests. In: Statistics, Textbooks and Monographs, 3rd edition 1995, Vol. 147.
58. Heyvaert M, Onghena P. Randomization tests for single-case experiments: State of the art, state of the science, and state of the application. J Contextual Behav Sci 2014 Jan;3(1):51-64. [doi: 10.1016/j.jcbs.2013.10.002]
59. Onghena P, Edgington ES. Customization of pain treatments: single-case design and analysis. Clin J Pain 2005;21(1):56-68. [doi: 10.1097/00002508-200501000-00007] [Medline: 15599132]
60. Manolov R, Onghena P. Testing delayed, gradual, and temporary treatment effects in randomized single-case experiments: A general response function framework. Behav Res Methods 2024 Apr;56(4):3915-3936. [doi: 10.3758/s13428-023-02230-1] [Medline: 37749426]
61. Parker RI, Vannest KJ, Davis JL, Sauber SB. Combining nonoverlap and trend for single-case research: Tau-U. Behav Ther 2011 Jun;42(2):284-299. [doi: 10.1016/j.beth.2010.08.006] [Medline: 21496513]
62. Hayes SC, Strosahl K, Wilson KG. Acceptance and Commitment Therapy: The Process and Practice of Mindful Change, 2nd edition: Guilford Press; 2012:402.
63. Hedman E, Axelsson E, Andersson E, Lekander M, Ljótsson B. Exposure-based cognitive-behavioural therapy via the internet and as bibliotherapy for somatic symptom disorder and illness anxiety disorder: randomised controlled trial. Br J Psychiatry 2016 Nov;209(5):407-413. [doi: 10.1192/bjp.bp.116.181396] [Medline: 27491531]

XSL•FO
RenderX

64. Sauer KS, Bailer J, Jungmann SM, Witthöft M. Inhibitory Learning-Based Exposure Therapy for Patients With Pathological Health Anxiety: Results From a Single Case Series Study. Cogn Behav Pract 2023 Nov;30(4):610-634. [doi: 10.1016/j.cbpra.2022.08.001]
65. Cuijpers P, Marks IM, van Straten A, Cavanagh K, Gega L, Andersson G. Computer-aided psychotherapy for anxiety disorders: a meta-analytic review. Cogn Behav Ther 2009;38(2):66-82. [doi: 10.1080/16506070802694776] [Medline: 20183688]
66. Miller CB, Gu J, Henry AL, et al. Feasibility and efficacy of a digital CBT intervention for symptoms of Generalized Anxiety Disorder: A randomized multiple-baseline study. J Behav Ther Exp Psychiatry 2021 Mar;70:101609. [doi: 10.1016/j.jbtep.2020.101609] [Medline: 32950939]
67. Gullickson KM, Hadjistavropoulos HD, Dear BF, Titov N. Negative effects associated with internet-delivered cognitive behaviour therapy: An analysis of client emails. Internet Interv 2019 Dec;18:100278. [doi: 10.1016/j.invent.2019.100278] [Medline: 31890627]
68. Schermuly-Haupt ML, Linden M, Rush AJ. Unwanted Events and Side Effects in Cognitive Behavior Therapy. Cogn Ther Res 2018 Jun;42(3):219-229. [doi: 10.1007/s10608-018-9904-y]
69. Strauss B, Gawlytta R, Schleu A, Frenzl D. Negative effects of psychotherapy: estimating the prevalence in a random national sample. BJPsych open 2021 Nov;7(6). [doi: 10.1192/bjo.2021.1025] [Medline: 34956647]
70. Tate RL, Perdices M, Rosenkoetter U, et al. The Single-Case Reporting Guideline In BEhavioural Interventions (SCRIBE) 2016 statement. Arch Sci Psychol 2016;4(1):1-9.
71. Manolov R, Moeyaert M. Recommendations for Choosing Single-Case Data Analytical Techniques. Behav Ther 2017 Jan;48(1):97-114. [doi: 10.1016/j.beth.2016.04.008] [Medline: 28077224]
72. Vannest KJ, Ninci J. Evaluating Intervention Effects in Single‐Case Research Designs. Jour of Counseling & Develop 2015 Oct;93(4):403-411. [doi: 10.1002/jcad.12038]
73. Vannest KJ, Peltier C, Haas A. Results reporting in single case experiments and single case meta-analysis. Res Dev Disabil 2018 Aug;79:10-18. [doi: 10.1016/j.ridd.2018.04.029] [Medline: 29960830]
74. Kazdin AE. Single-case experimental designs: Characteristics, changes, and challenges. J Exp Anal Behav 2021 Jan;115(1):56-85. [doi: 10.1002/jeab.638] [Medline: 33205436]
75. Ledford JR, Lane JD, Severini KE. Systematic Use of Visual Analysis for Assessing Outcomes in Single Case Design Studies. Brain Impair 2018 Mar;19(1):4-17. [doi: 10.1017/BrImp.2017.16]
76. Bulté I, Onghena P. When the Truth Hits You Between the Eyes: A Software Tool for the Visual Analysis of Single-Case Experimental Data. Methodology (Gott) 2011;8(3):104-114. [doi: 10.1027/1614-2241/a000042]
77. Manolov R, Vannest KJ. A Visual Aid and Objective Rule Encompassing the Data Features of Visual Analysis. Behav Modif 2023 Nov;47(6):1345-1376. [doi: 10.1177/0145445519854323] [Medline: 31165621]
78. Vlaeyen JWS, de Jong J, Geilen M, Heuts P, van Breukelen G. Graded exposure in vivo in the treatment of pain-related fear: a replicated single-case experimental design in four patients with chronic low back pain. Behav Res Ther 2001 Feb;39(2):151-166. [doi: 10.1016/s0005-7967(99)00174-6] [Medline: 11153970]
79. Kemani MK, Olsson GL, Holmström L, Wicksell RK. Exploring Changes in Valued Action in the Presence of Chronic Debilitating Pain in Acceptance and Commitment Therapy for Youth - A Single-Subject Design Study. Front Psychol 2016;7:1984. [doi: 10.3389/fpsyg.2016.01984] [Medline: 28082933]
80. Ruiz FJ, Flórez CL, García-Martín MB, et al. A multiple-baseline evaluation of a brief acceptance and commitment therapy protocol focused on repetitive negative thinking for moderate emotional disorders. J Contextual Behav Sci 2018 Jul;9:1-14. [doi: 10.1016/j.jcbs.2018.04.004]
81. Marchand L, Marchand A, Landry P, Letarte A, Labrecque J. Efficacy of two cognitive-behavioral treatment modalities for panic disorder with nocturnal panic attacks. Behav Modif 2013 Sep;37(5):680-704. [doi: 10.1177/0145445513492792] [Medline: 23821775]
82. Smith JD. Single-case experimental designs: a systematic review of published research and current standards. Psychol Methods 2012 Dec;17(4):510-550. [doi: 10.1037/a0029312] [Medline: 22845874]
83. Tate RL, Perdices M, Rosenkoetter U, et al. Revision of a method quality rating scale for single-case experimental designs and n-of-1 trials: the 15-item Risk of Bias in N-of-1 Trials (RoBiNT) Scale. Neuropsychol Rehabil 2013;23(5):619-638. [doi: 10.1080/09602011.2013.824383] [Medline: 24050810]
84. What Works Clearinghouse™ Standards Handbook, Version 4.1: What Works Clearinghouse; 2020.
85. Sighvatsson MB, Salkovskis PM, Sigurdsson E, Valdimarsdottir HB, Thorsdottir F, Sigurdsson JF. Evaluation of mechanism of change in transdiagnostic cognitive behaviour therapy using single case experimental design. J Behav Ther Exp Psychiatry 2021 Jun;71:101634. [doi: 10.1016/j.jbtep.2020.101634] [Medline: 33388441]

Abbreviations

ACT: acceptance and commitment therapy
CBT: cognitive behavioral therapy
***DSM-5*:** *Diagnostic and Statistical Manual of Mental Disorders (Fifth Edition)*

XSL•FO
RenderX

ESQ: Experience of Service Questionnaire
HA: health anxiety
HAPYS: Health Anxiety by Proxy Scale
***ICD-10*:** *International Statistical Classification of Diseases and Related Health Problems, Tenth Revision*
***ICD-11*:** *International Classification of Diseases, 11th Revision*
IEUQ: Internet Evaluation and Utility Questionnaire
NEQ: Negative Effects Questionnaire-20
SCED: single-case experimental design
WI-6-R: Whiteley Index 6 Revised

Edited by J Sarvestan; submitted 14.08.24; peer-reviewed by S Alves; revised version received 24.08.25; accepted 25.08.25; published 02.10.25.

Please cite as:
Ingeman K, Frydendal DH, Frostholm L, Bjerre-Nielsen E, Wellnitz KB, Onghena P, Wright K, Rask CU
Internet-Delivered Psychological Treatment for Parents With Health Anxiety by Proxy: Replicated Randomized Single-Case Experimental Design
JMIR Form Res 2025;9:e65396
URL: https://formative.jmir.org/2025/1/e65396
doi:10.2196/65396

© Katrine Ingeman, Ditte Hoffmann Frydendal, Lisbeth Frostholm, Ellen Bjerre-Nielsen, Kaare Bro Wellnitz, Patrick Onghena, Kristi Wright, Charlotte Ulrikka Rask. Originally published in JMIR Formative Research (https://formative.jmir.org), 2.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Dietary Fiber Estimate of DialBetesPlus App Users: Secondary Analysis of Data From a Randomized Controlled Trial

Wei Thing Sze[1,2], MSc, BPharm; Kayo Waki[1,2,3], MD, MPH, PhD; Ryohei Nakada[2], BEng; Toshimasa Yamauchi[3], MD, PhD; Masaomi Nangaku[4], MD, PhD; Kazuhiko Ohe[1,2], MD, PhD

[1]Department of Clinical Information Engineering, Graduate School of Medicine, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan
[2]Department of Planning, Information and Management, The University of Tokyo Hospital, Tokyo, Japan
[3]Department of Diabetes and Metabolic Diseases, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan
[4]Division of Nephrology and Endocrinology, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan

Corresponding Author:
Kayo Waki, MD, MPH, PhD
Department of Clinical Information Engineering, Graduate School of Medicine, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan

Abstract

Background: Despite the importance of dietary fiber in regulating glycemic control, the reported intake among patients with type 2 diabetes (T2D) in Japan was around 12 - 16 g, well below the local official recommended intake of 20 g and above. Recent data is lacking, with the most recent available estimates collected between 2014 and 2019. Most mHealth dietary intervention apps for T2D focus on calorie and carbohydrate outcomes, with limited evidence on fiber intake. Fiber data was collected in a recent 12-month trial of DialBetesPlus, a multimodal diabetes mHealth self-management system that supports dietary behavior change by allowing users to record their meals and provides timely and detailed information on users' nutrient intake.

Objective: This study aimed to assess the pattern of dietary fiber intake among DialBetesPlus intervention users to provide recent data. As a secondary objective, the study explored factors that may influence dietary fiber intake among participants.

Methods: Meal records were extracted using the DialBetesPlus app developer dashboard. Dietary fiber intake was measured for all intervention participants. The analysis included only data from participants with complete (breakfast, lunch, and dinner) meal records for at least 7 days. We calculated the average dietary fiber intake and fiber density per day by first averaging across all participants for each day, then averaging these daily values over 1 year. We averaged fiber intake per meal type across all days with available data for each participant, without excluding incomplete meal record days.

Results: Out of 66 participants from the intervention group who were assigned to the DialBetesPlus intervention, 47 (71.2%) had at least 7 days of complete meal records and were included in the analysis. A 1-year trend analysis revealed a slight upward trend of daily fiber intake with the rolling average consistently below 18 g. The average fiber intake was 17.1 g/day, with a corresponding mean fiber density of 10.5 g/1000 kcal. The overall mean fiber intake was 17.1 g/day. Separate analysis by meal types revealed that the highest fiber intake was during dinner (6.7 g), followed by lunch (4.8 g), breakfast (4.4 g), and snacks (1.5 g), while fiber density was lowest for snacks (7.8 g/1000 kcal), followed by dinner (10.2 g/1000 kcal), lunch (10.5 g/1000 kcal), and breakfast (10.8 g/1000 kcal). No significant correlations were observed between average fiber intake and participant characteristics such as age, sex, BMI, hemoglobin A_{1c}, blood pressure, and frequency of meal logging.

Conclusions: Despite using a general diabetes self-management mHealth app (DiabetesPlus) that included dietary self-monitoring and basic nutritional feedback, users consumed less than the recommended 20 g/day of dietary fiber on average over a 1-year period. This study highlights the need to explore alternative mHealth strategies to further enhance dietary fiber intake among patients with T2D.

(JMIR Form Res 2025;9:e69340) doi:10.2196/69340

KEYWORDS

type 2 diabetes; dietary fiber; mobile health; meal monitoring; diabetes self-management app

Introduction

Dietary fiber is a macronutrient of importance in the dietary self-management of individuals with diabetes. Dietary fiber improves glycemic control by decreasing postprandial hyperglycemia, enhancing peripheral insulin sensitivity, and activating incretin secretion via short-chain fatty acids produced by fermentation of fiber in the intestines [1-3]. It also enhances satiety, which leads to a reduction in body weight [4].

XSL•FO
RenderX

Despite the importance of dietary fiber in regulating glycemic control, the reported dietary fiber intake among patients with type 2 diabetes (T2D) in Japan was around 12 - 16 g, below the minimum intake of 20 g recommended by the Japan Diabetes Society [5-8]. Higher fiber intake (>25 g) is recommended for patients with T2D to prevent cardiovascular disease, reduce premature mortality, and to achieve meaningful hemoglobin A_{1c} (HbA_{1c}) reduction [7,9]. Higher fiber intake was also reported to be associated with lower obesity risk and increased physical activity among patients with T2D in Japan [8].

Although many mHealth dietary self-monitoring apps target patients with T2D, most of them focus on calorie and carbohydrate intake, with limited evidence addressing fiber intake [10]. Furthermore, recent evidence on the fiber consumption of Japanese patients with T2D is limited, with the most recent available estimates collected between 2014 and 2019 [8]. Therefore, we performed a secondary analysis of the dietary data collected from a mHealth self-management support system that used the DialBetesPlus app to assess dietary fiber intake of the intervention participants. DialBetesPlus allows patients to self-monitor blood glucose, blood pressure, body weight, daily step counts, and diet. The effectiveness of DialBetesPlus on improving albuminuria and glycemic control in patients with diabetic kidney disease when added to standard care was demonstrated in a 12-month clinical trial conducted between 2018 and 2021 [11].

For diet self-monitoring, DialbetesPlus intervention participants recorded the content and quantity of their meals and snacks on the DialBetesPlus app via text input as well as registering a photograph of their meal to assist with the input. The recorded data was sent to the DialBetesPlus server and immediately evaluated using the FoodDataBank database (SARAH Co, Ltd). Based on the calculated consumed calories and nutrient intake, the server generated feedback related to nutritional balances and dietary habits, which were then sent to the patients through the app via text messages. DialBetesPlus strengthens perceived and actual control of dietary behavior modification by giving patients timely and detailed information about their nutrient intake, such as protein, fiber, and fat, as well as general tips and suggestions to eat healthier [11].

The aim of this secondary analysis is to explore patterns of dietary fiber intake of DialBetesPlus intervention users using meal records captured on the DialBetesPlus app, specifically on the overall average, trend over 1 year, and the average intake by meal type. As a secondary objective, we examined the association between total dietary fiber intake and participant characteristics such as age, sex, BMI, HbA_{1c}, blood pressure as well as engagement parameters such as frequency of meal logging.

Methods

Overview

We conducted this secondary analysis using meal log data from a 12-month randomized controlled trial evaluating the effectiveness of DialBetesPlus, a self-management support system for diabetic kidney disease. The primary study randomized 132 participants from 8 clinical sites in Japan, 66 intervention and 60 control group between July 2018 and August 2019. Eligibility criteria included diagnosis of T2D with moderately increased albuminuria (urine albumin-creatinine ratio: 30 - 299 mg/g creatinine), HbA_{1c} of 6.5% or more, and between 20 and 75 years of age [11]. We randomized participants to either an intervention group using DialBetesPlus, a self-management support system allowing patients to monitor exercise, blood glucose, diet, blood pressure, and body weight via a smartphone app or a control group receiving standard care.

After logging each meal in the DialBetesPlus app, participants received immediate feedback on the nutrient composition of the meal, including total energy (kcal), protein, fat, carbohydrate, dietary fiber, and salt content. For example, after logging breakfast, the app displayed the total caloric content of the meal along with a breakdown by individual food items. The app evaluated nutrient intake as a percentage of the recommended daily intake and provided personalized advice on each nutrient intake accordingly (Figure 1).

XSL•FO
RenderX

Figure 1. Example of nutrient assessment and feedback provided by the DialBetesPlus app after a meal log. English translation: left panel: meal type: breakfast; total energy: 1365 kcal; meal items: open sandwich – 570 kcal, milk pudding – 131 kcal, yogurt – 67 kcal, iced coffee – 47 kcal (1 glass). Note (bottom bar): "energy, protein, fat, carbohydrate, salt intake was excessive." Right Panel (breakfast advice): percentage of each nutrient relative to daily recommended intake: calories, protein, fat, carbohydrate, dietary fiber, and salt. For textual advice, please refer to Textbox 1.

Textbox 1. Textual Advice

- You are consuming too much energy. Try to reduce your energy intake by choosing low-calorie dishes and eating lots of vegetables. Seaweed and vegetables are low in calories and help you feel full easily.
- You are consuming too much fat. Limit your intake of cholesterol (found only in animal-based foods such as eggs, full-fat dairy products, and meat), which increases the risk of arterial disease.
- You are consuming too little dietary fiber. Dietary fiber is found not only in vegetables, but also in mushrooms, seaweed, and beans.
- You are consuming too much salt. Heavily seasoned foods contain a lot of salt. For your health, try to get used to less salty foods.

We extracted both meal records and associated nutrient data including calorie, carbohydrate, fiber, salt content, and lipid from the DialBetesPlus app developer dashboard. We defined a complete meal record for a given patient and day as a record that had complete logged meal occasions for breakfast, lunch, and dinner in a day. Meal types (eg, “breakfast”) were selected by participants during the time of meal submission. We defined an incomplete meal record as one that was missing a meal record for any of breakfast, lunch, or dinner in a day, and days with an incomplete or missing meal log were excluded. To enhance the accuracy of dietary fiber intake estimates, we limited analysis to participants who recorded complete meals for at least 7 days, given that fiber intake is more accurately assessed from complete meal data.

XSL•FO
RenderX

Statistical Analysis

We calculated the means and standard deviations for continuous variables, and frequencies and proportions for categorical variables, to summarize the baseline demographic and clinical characteristics of the study population. To assess whether there were any systematic differences between the analyzed and excluded groups at baseline, we conducted group comparisons using Welch *t* test or Mann-Whitney *U* test for continuous variables and chi-square test for categorical variables.

The average dietary fiber intake for each day was calculated by averaging the intake across all participants. To visualize the trend over time, a 30-day rolling average was applied to these daily averages. Then, we calculated the average fiber intake per day by taking the average of these daily values over 366 days. Similarly, we calculated fiber density (g/1000 kcal) each day by dividing each participant's total fiber intake by their total energy intake, and then averaged the daily values across participants. We included days with at least 3 main meals (breakfast, lunch, and dinner) recorded in the analysis, regardless of whether snacks were logged. We took this approach because snack consumption tends to be more variable and inconsistent across individuals and days, making it a less reliable indicator of daily fiber intake.

To find mean dietary fiber intake for each meal type (breakfast, lunch, dinner, or snacks), the recorded dietary fiber intake for each participant was averaged per meal type across all days, and these averages were aggregated across all participants. Fiber density was also calculated for each meal type by dividing the fiber intake by the corresponding energy intake, standardized per 1000 kilocalories.

To explore the relationship between average fiber intake and individual characteristics, we calculated Spearman correlation. We explored the correlations between the average fiber intake per day and age, sex, baseline BMI, BMI change postintervention, baseline systolic blood pressure (SBP), baseline diastolic blood pressure (DBP), SBP and DBP change post intervention, baseline HbA_{1c}, and HbA_{1c} change postintervention.

Python 3.12.4 (Python Software Foundation) was used for the data analysis.

Ethical Considerations

The Institutional Review Board of the University of Tokyo School of Medicine approved this study (approval number 2021084NI-(2)). The approval includes permission for secondary analyses without requiring additional consent. We deidentified all study data to maintain confidentiality and protect the privacy of research participants. Participants in the original study received compensation as approved by the relevant ethics committee. No additional compensation was provided for this secondary analysis.

Results

Out of 66 participants from the intervention group who used the DialBetesPlus app, 47 (71.2%) participants had at least 7 days with complete meal records and thus were included in the analysis (referred to as "analyzed participants" in this text). Participants were mostly male (72.3%, 34/47), late middle age or older individuals having a mean age of 60.2 (SD 8.6) years, with an average BMI of 28.1 (SD 5.3) kg/m^2 (Table 1). Among the participants included in the analysis, the average percentage of days with complete meal records during the 1-year intervention period was 55.5% (203/366 days). There were no statistically significant differences in baseline characteristics such as age, sex, BMI, baseline blood pressure, and baseline HbA_{1c} between participants included in the analysis and those excluded (Multimedia Appendix 1 and 2).

Table . Demographic and clinical characteristics of the analyzed participants (N=47) in a 12-month randomized controlled trial investigating DialBetesPlus intervention for individuals with type 2 diabetes in Japan.

Demographic and clinical characteristics	Values (N=47)
Age (years), mean (SD)	60.2 (8.6)
Sex, n (%)	
Male	34 (72.3)
Female	13 (27.6)
Baseline BMI (kg/m^2), mean (SD)	28.1 (5.3)
Changes in BMI (kg/m^2) post intervention (n=46)	−0.7 (1.2)
Baseline HbA_{1c}[a], mean (SD)	7.8 (1.1)
Changes in HbA_{1c} post intervention	–0.2 (0.9)
Baseline blood pressure (SBP[b]/DBP[c], mm Hg), mean (SD), (n=46)	135/81 (15/11)
Changes in blood pressure (SBP[b]/DBP[c], mm Hg)	−3/−3 (18/13)

[a]HbA_{1c}: hemoglobin A_{1c}.

[b]SBP: systolic blood pressure.

[c]DBP: diastolic blood pressure.

XSL•FO
RenderX

Participants consumed an average of 17.1 (SD 1.4, SE 0.1) g/day of dietary fiber, with the 95% CI values ranging from 16.9 to 17.2 g/day. When adjusted for energy intake, the mean fiber density was 10.5 (SD 0.7, SE 0.04) g per 1000 kcal, with a 95% CI value of 10.5 to 10.6 g/1000 kcal. Visual inspection of Q-Q plots indicated that both daily fiber intake and fiber density were approximately normally distributed (Multimedia Appendix 3). The 12-month rolling average trend graph reveals that there was a gradual increase of approximately 1.5 - 2 g over the first 6 months, followed by a slight decline of 0.5 g over the last 3 months, with the 30-day rolling average staying below 18 g most of the time (Figure 2). As for fiber density, the 30-day rolling average fiber density stayed relatively stable between 10 and 11 g/1000 kcal over the year with minor fluctuations (Figure 3). Analysis by meal types reveals that the highest fiber intake was during dinner (6.7 g), followed by lunch (5.8 g), breakfast (4.4 g), and snacks (1.5 g), while fiber density was lowest for snacks (7.8 g/1000 kcal), followed by dinner (10.2 g/1000 kcal), lunch (10.5 g/1000 kcal), and breakfast (10.8 g/1000 kcal; Table 2).

Figure 2. Thirty-day rolling average of daily dietary fiber intake over 366 days among analyzed participants (N=47) in a 12-month randomized controlled trial investigating DialBetesPlus intervention for individuals with type 2 diabetes in Japan.

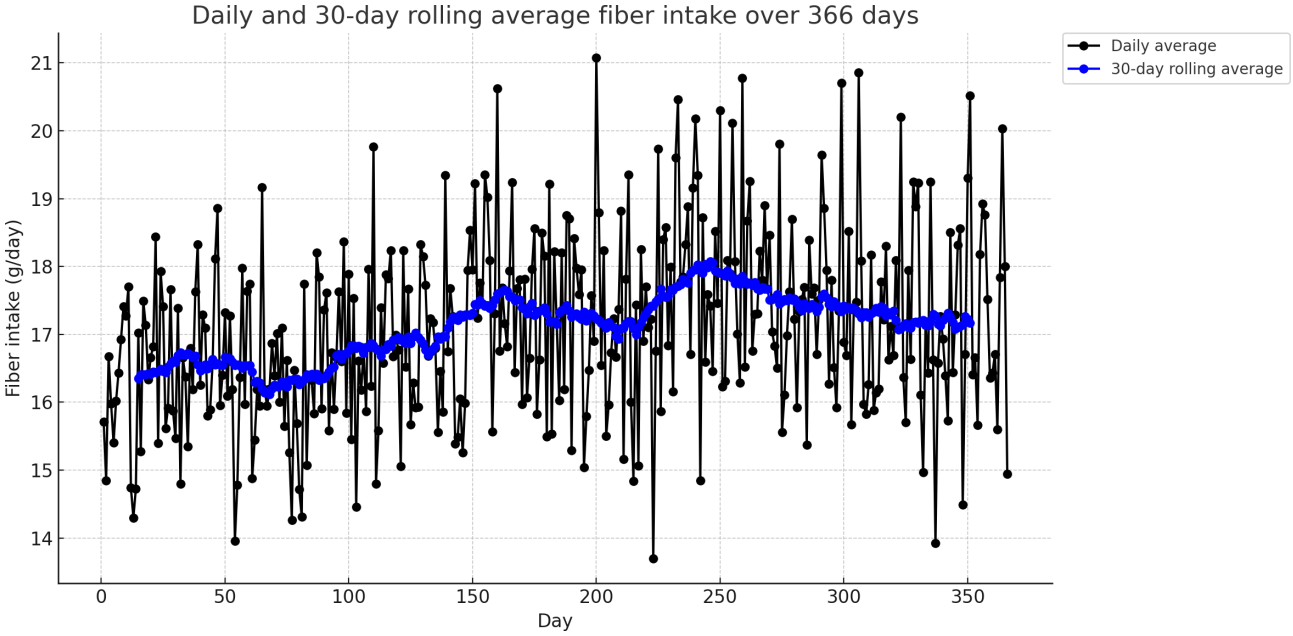

Figure 3. Thirty-day rolling average of fiber density over 366 days among analyzed participants (N=47) in a 12-month randomized controlled trial investigating DialBetesPlus intervention for individuals with type 2 diabetes in Japan.

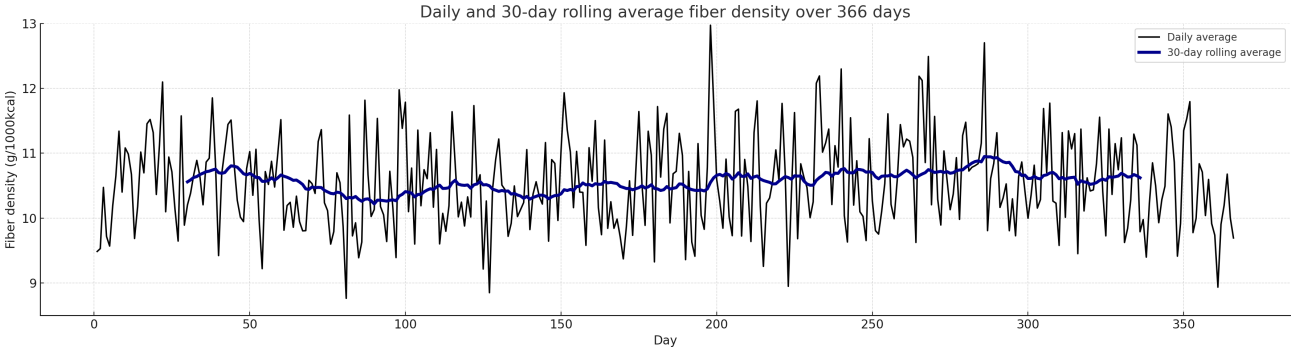

XSL•FO
RenderX

Table . Average dietary fiber and fiber density during breakfast, lunch, dinner, and snacks among analyzed participants (N=47) in a 12-month randomized controlled trial investigating DialBetesPlus intervention for individuals with type 2 diabetes in Japan. Absolute fiber intake (g).

Meal type	Mean (SD)	SE[a] (95% CI)
Breakfast	4.3 (1.8)	0.3 (3.8-4.9)
Lunch	5.8 (1.9)	0.3 (5.3-6.4)
Dinner	6.7 (2.9)	0.4 (5.9-7.5)
Snacks	1.5 (0.7)	0.1 (1.3-1.7)
Fiber density (g/1000 kcal)		
Breakfast	10.8 (4.4)	0.6 (9.5-12.1)
Lunch	10.5 (3.6)	0.5 (9.4-11.5)
Dinner	10.2 (3.5)	0.5 (9.2-11.2)
Snacks	7.8 (4.5)	0.6 (6.5-9.2)

[a]SE: standard error.

Spearman correlation analysis revealed no statistically significant associations between the average fiber intake per day and participant characteristics including age, sex, baseline BMI, baseline HbA_{1c}, baseline SBP, and baseline DBP, as well as changes in BMI, HbA_{1c}, and blood pressure post intervention (all P>.05) (Multimedia Appendix 4). In addition, Spearman correlation analysis was conducted to examine the relationship between logging frequency and average daily fiber intake. Meal logging frequency was defined as the total number of meal logs logged over the study period. The analysis revealed no significant association between meal logging frequency and fiber intake (P=.03 and .82, respectively), suggesting that the meal logging frequency was not meaningfully related to the participants' dietary fiber consumption.

Discussion

Principal Findings

In this secondary analysis, average dietary fiber intake among DialBetesPlus users was 17.1 g/day. Fiber intake gradually increased during the first 6 months of the intervention before declining slightly in the final 3 months. Fiber intake was highest at dinner and lowest from snacks. No significant associations were observed between total fiber intake and participant characteristics, including age, sex, baseline BMI, baseline HbA_{1c}, or baseline blood pressure. Similarly, meal logging frequency showed no meaningful relationship with average fiber intake.

Our secondary analysis demonstrated that, even among this group of highly compliant and engaged app users, the average dietary fiber intake remained below the clinical recommendation of the Japan Diabetes Society, and the average fiber density fell well short of the general recommendation of 14 g/1000 kcal [12,13]. The DialBetesPlus app users performed diet self-monitoring and received feedback on nutrient intake and general dietary advice on the app over 1 year. These findings indicate that, despite the use of a sophisticated tool and high participant motivation, meeting fiber intake goals remains difficult, suggesting the need for more tailored or intensive support. Furthermore, patients might not be fully aware of the importance of eating more fiber or how to incorporate more fiber in their diet. Previous studies showed that various factors influence fiber intake, such as lack of knowledge on the benefits of fiber, beliefs on high-fiber food as unpalatable, the higher price of fiber-rich food, and lack of skills in recognizing sources of dietary fiber [14,15].

The average fiber intake per day of DialBetesPlus intervention users was generally higher than what was reported among the Japanese population with T2D in non-mHealth studies (17.1 g vs 12 - 16 g) [5-7]. The trend graph shows a small and gradual increase of mean dietary fiber intake over the first 6 months (1.5 - 2 g), which may be due to the influence of the intervention to help patients eat a healthier diet. The observed plateau in the fiber increase after 6 months is consistent with findings from previous studies showing that the effect size of mHealth intervention typically diminishes over time, possibly due to declining user engagement [11,16].

Breakfast was reported to contain the lowest fiber intake compared to lunch and dinner. Analysis of diet quality through Nutrient-Rich Food Index 9.3 showed that, on average, the percentage contribution of key nutrients (protein, dietary fiber, vitamins, and minerals) to the total dietary intake of Japanese was highest at dinner, followed by lunch, breakfast, and snacks [17]. The lower fiber intake observed at breakfast compared to lunch and dinner may be attributed to the simpler composition of morning meals, such as rice, plain bread, and miso soup, while lunch and dinner often featured a wider variety of vegetable side dishes [18]. Nevertheless, fiber density at breakfast was higher than lunch and dinner, suggesting it was the most fiber-efficient meal relative to energy consumed. A study reported that dinner accounted for the largest share of total daily energy intake (40%) among Japanese adults, which helps explain the lower fiber density observed for dinner in this study [17]. Snacks provided the lowest fiber (1.5 g) and density (7.8 g/1000 kcal) among all the meal types, consistent with previous reports indicating that snack consumption in Japan is largely composed of low-fiber food such as confectioneries and sugar-sweetened beverages [17]. These results suggest a need for practical strategies to motivate patients with T2D in Japan

XSL•FO
RenderX

to increase fiber intake for breakfast and snacks, as well as to increase overall fiber density in all meals.

The absence of associations between fiber intake and participant characteristics such as age, sex, BMI, blood pressure, HbA_{1c}, as well as meal logging frequency, suggests that these factors and the extent of meal self-monitoring may not strongly influence dietary fiber intake among patients with T2D in Japan. Other factors such as knowledge, attitude, culture, psychosocial factors, and socioeconomic status such as household income and education level may play a more prominent role in shaping dietary fiber intake behaviors in this population [19-22]. Future studies should investigate other behavioral and sociocultural determinants to better understand what drives dietary fiber intake and to inform the design of more effective interventions among patients with T2D in Japan.

Strengths and Limitations

In this secondary analysis, 1 year of meal record data was used, providing a comprehensive assessment of dietary fiber intake that captured long-term eating habits and seasonal variations, as compared to assessments from short-term diet records.

We excluded days with incomplete meal records from the analysis, which may have favored inclusion of more engaged and health-motivated individuals. Participants who consistently logged all 3 meals may differ from those who did not, particularly in terms of health behaviors and motivation [23]. Therefore, the reported mean fiber intake of 17.1 g/day may overestimate the true average for the intervention users and population with T2D using similar app at large. Nevertheless, this method allowed uniform representation of dietary fiber intake assessment from each meal type and each participant and ensured that daily fiber estimates were not disproportionately influenced by partial or unbalanced records. It also provides insight into the dietary patterns of participants who were more adherent to meal logging as they are most likely to benefit from or respond to such digital dietary interventions. Moreover, the similarity in baseline characteristics between analyzed and excluded participants suggests that the analyzed sample remains representative of the broader study population, minimizing concerns about selection bias.

Due to dependence on self-reporting, reporting errors such as underreporting of food items and inaccurate portion size estimation may have occurred. We reduced the likelihood of reporting errors by using image-assisted food recording, which assisted participants in recording their meals accurately [24].

Conclusions

Japanese patients with T2D who performed diet self-monitoring through the DialBetesPlus app consumed less than the recommended intake of 20 g and above of dietary fiber per day, on average over 1 year. Our findings suggest that even among highly motivated and digitally engaged participants, diet self-monitoring via the app alone did not lead to participants meeting the clinically recommended dietary fiber intake. This highlights the need to complement app-based tracking with additional evidence-based intervention features such as the use of fiber-focused intervention, goal setting, personalized feedback, action planning, and coping planning [25-29]. Exploring patients' beliefs on and perceptions of dietary fiber intake is important to build user-informed dietary mHealth interventions [30].

Acknowledgments

We thank Daniel Lane for his support in manuscript editing and scientific discussions.

Data Availability

Data from this study are available from the corresponding author upon reasonable request for noncommercial use.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Demographic and clinical characteristics of the total study population and the subgroup of participants excluded from analysis in a 12-month randomized controlled trial evaluating DialBetesPlus intervention for individuals with type 2 diabetes in Japan.
[DOCX File, 16 KB - formative_v9i1e69340_app1.docx]

Multimedia Appendix 2
Comparison of baseline demographic and clinical characteristics between analyzed (N=47) and excluded (n=19) participants in a 12-month randomized controlled trial of the DialBetesPlus intervention for individuals with type 2 diabetes in Japan.
[DOCX File, 14 KB - formative_v9i1e69340_app2.docx]

Multimedia Appendix 3
Q-Q plots evaluating the distribution of daily dietary fiber intake (S1A) and fiber density (S1B) among analyzed participants in a 12-month randomized controlled trial investigating DialbetesPlus intervention for individuals with type 2 diabetes in Japan.
[DOCX File, 224 KB - formative_v9i1e69340_app3.docx]

XSL•FO
RenderX

Multimedia Appendix 4
Spearman correlation coefficients and corresponding *P* values between total dietary fiber intake and participant characteristics.
[DOCX File, 15 KB - formative_v9i1e69340_app4.docx]

References

1. Koloverou E, Panagiotakos DB. Macronutrient composition and management of non-insulin-dependent diabetes mellitus (NIDDM): a new paradigm for individualized nutritional therapy in diabetes patients. Rev Diabet Stud 2016;13(1):6-16. [doi: 10.1900/RDS.2016.13.6] [Medline: 27563693]
2. Mazhar M, Zhu Y, Qin L. The interplay of dietary fibers and intestinal microbiota affects type 2 diabetes by generating short-chain fatty acids. Foods 2023 Feb 28;12(5):1023. [doi: 10.3390/foods12051023] [Medline: 36900540]
3. Alahmari LA. Dietary fiber influence on overall health, with an emphasis on CVD, diabetes, obesity, colon cancer, and inflammation. Front Nutr 2024;11:1510564. [doi: 10.3389/fnut.2024.1510564] [Medline: 39734671]
4. Burton-Freeman B. Dietary fiber and energy regulation. J Nutr 2000 Feb;130(2S Suppl):272S-275S. [doi: 10.1093/jn/130.2.272S] [Medline: 10721886]
5. Fujii H, Iwase M, Ohkuma T, et al. Impact of dietary fiber intake on glycemic control, cardiovascular risk factors and chronic kidney disease in Japanese patients with type 2 diabetes mellitus: the Fukuoka Diabetes Registry. Nutr J 2013 Dec 11;12(1):159. [doi: 10.1186/1475-2891-12-159] [Medline: 24330576]
6. Kobayakawa A, Suzuki T, Ikami T, Saito M, Yabe D, Seino Y. Improvement of fasting plasma glucose level after ingesting moderate amount of dietary fiber in Japanese men with mild hyperglycemia and visceral fat obesity. J Diet Suppl 2013 Jun;10(2):129-141. [doi: 10.3109/19390211.2013.790335] [Medline: 23725526]
7. Tanaka S, Yoshimura Y, Kamada C, et al. Intakes of dietary fiber, vegetables, and fruits and incidence of cardiovascular disease in Japanese patients with type 2 diabetes. Diabetes Care 2013 Dec;36(12):3916-3922. [doi: 10.2337/dc13-0654] [Medline: 24170762]
8. Ferreira ED, Hatta M, Laymon K, et al. Demographic and dietary determinants of the association between dietary fibre intake and obesity in Japanese adults with type 2 diabetes: a cross-sectional study (JDDM 78). Public Health Nutr 2025 Feb 4;28(1):e49. [doi: 10.1017/S136898002500014X] [Medline: 39901432]
9. Reynolds AN, Akerman AP, Mann J. Dietary fibre and whole grains in diabetes management: systematic review and meta-analyses. In: Ma RCW, editor. PLoS Med 2020 Mar;17(3):e1003053. [doi: 10.1371/journal.pmed.1003053] [Medline: 32142510]
10. Zheng Y, Campbell Rice B, Melkus GD, et al. Dietary self-management using mobile health technology for adults with type 2 diabetes: a scoping review. J Diabetes Sci Technol 2023 Sep;17(5):1212-1225. [doi: 10.1177/19322968231174038] [Medline: 37162011]
11. Waki K, Nara M, Enomoto S, et al. Effectiveness of DialBetesPlus, a self-management support system for diabetic kidney disease: randomized controlled trial. NPJ Digit Med 2024 Apr 27;7(1):104. [doi: 10.1038/s41746-024-01114-8] [Medline: 38678094]
12. Araki E, Goto A, Kondo T, et al. Japanese clinical practice guideline for diabetes 2019. Diabetol Int 2020 Jul;11(3):165-223. [doi: 10.1007/s13340-020-00439-5] [Medline: 32802702]
13. Dahl WJ, Stewart ML. Position of the Academy of Nutrition and Dietetics: health implications of dietary fiber. J Acad Nutr Diet 2015 Nov;115(11):1861-1870. [doi: 10.1016/j.jand.2015.09.003] [Medline: 26514720]
14. Jezewska-Zychowicz M, Królak M. The choice of bread: the association between consumers' awareness of dietary fiber and declared intentions to eat. Nutrients 2020 Jan 30;12(2):360. [doi: 10.3390/nu12020360] [Medline: 32019124]
15. Alfawaz H, Khan N, Alhuthayli H, et al. Awareness and knowledge regarding the consumption of dietary fiber and its relation to self-reported health status in an adult Arab population: a cross-sectional study. Int J Environ Res Public Health 2020 Jun 13;17(12):4226. [doi: 10.3390/ijerph17124226] [Medline: 32545755]
16. Mönninghoff A, Kramer JN, Hess AJ, et al. Long-term effectiveness of mHealth physical activity interventions: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2021 Apr 30;23(4):e26699. [doi: 10.2196/26699] [Medline: 33811021]
17. Murakami K, Shinozaki N, Livingstone MBE, et al. Characterisation of breakfast, lunch, dinner and snacks in the Japanese context: an exploratory cross-sectional analysis. Public Health Nutr 2022 Mar;25(3):689-701. [doi: 10.1017/S1368980020004310] [Medline: 33168120]
18. Shinozaki N, Murakami K, Asakura K, Masayasu S, Sasaki S. Identification of dish-based dietary patterns for breakfast, lunch, and dinner and their diet quality in Japanese adults. Nutrients 2020 Dec 28;13(1):67. [doi: 10.3390/nu13010067] [Medline: 33379273]
19. Daud NM, Fadzil NI, Yan LK, et al. Knowledge, attitude and practice regarding dietary fibre intake among Malaysian rural and urban adolescents. Mal J Nutr 2018;24(1):77-88 [FREE Full text]
20. James DCS. Factors influencing food choices, dietary intake, and nutrition-related attitudes among African Americans: application of a culturally sensitive model. Ethn Health 2004 Nov;9(4):349-367. [doi: 10.1080/1355785042000285375] [Medline: 15570680]

XSL•FO
RenderX

21. Wang DH, Kogashiwa M, Mori N, et al. Psychosocial determinants of fruit and vegetable consumption in a Japanese population. Int J Environ Res Public Health 2016 Aug 5;13(8):786. [doi: 10.3390/ijerph13080786] [Medline: 27527198]
22. Nakamura S, Inayama T, Hata K, et al. Association of household income and education with eating behaviors in Japanese adults: a cross-sectional study. BMC Public Health 2016 Jan 22;16:61. [doi: 10.1186/s12889-016-2748-z] [Medline: 26800891]
23. Reed JR, Struwe L, Bice MR, Yates BC. The impact of self-monitoring food intake on motivation, physical activity and weight loss in rural adults. Appl Nurs Res 2017 Jun;35:36-41. [doi: 10.1016/j.apnr.2017.02.008] [Medline: 28532724]
24. Gemming L, Utter J, Ni Mhurchu C. Image-assisted dietary assessment: a systematic review of the evidence. J Acad Nutr Diet 2015 Jan;115(1):64-77. [doi: 10.1016/j.jand.2014.09.015] [Medline: 25441955]
25. Giang NH, Hung NT, Hien VTT, et al. Fiber-focused nutrition counseling through nutrition software improved HbA1c of Vietnamese type 2 diabetes mellitus patients. Asian J Diet 2020 [FREE Full text]
26. Herman A, Hand L, Sullivan D, Hull H. A single-goal fiber intervention does not compromise nutritional intake during pregnancy. Current Developments in Nutrition 2020 Jun;4(Suppl 2):nzaa054. [doi: 10.1093/cdn/nzaa054_072]
27. Adachi M, Yamaoka K, Watanabe M, et al. Effects of lifestyle education program for type 2 diabetes patients in clinics: a cluster randomized controlled trial. BMC Public Health 2013 May 14;13(1):467. [doi: 10.1186/1471-2458-13-467] [Medline: 23672733]
28. Godinho CA, Alvarez MJ, Lima ML, Schwarzer R. Will is not enough: coping planning and action control as mediators in the prediction of fruit and vegetable intake. Br J Health Psychol 2014 Nov;19(4):856-870. [doi: 10.1111/bjhp.12084] [Medline: 24308823]
29. Parkinson J, Hannan T, McDonald N, Moriarty S, Nguyen TM, Hamilton K. Health action process approach: promoting physical activity, and fruit and vegetable intake among Australian adults. Health Promot Int 2023 Aug 1;38(4):daad095. [doi: 10.1093/heapro/daad095] [Medline: 37647521]
30. Mummah SA, King AC, Gardner CD, Sutton S. Iterative development of vegethon: a theory-based mobile app intervention to increase vegetable consumption. Int J Behav Nutr Phys Act 2016 Aug 8;13(1):90. [doi: 10.1186/s12966-016-0400-z] [Medline: 27501724]

Abbreviations

DBP: diastolic blood pressure
HbA_{1c}: hemoglobin A_{1c}
SBP: systolic blood pressure
T2D: type 2 diabetes

Edited by A Mavragani; submitted 27.11.24; peer-reviewed by AK Jansen, M Rahimlou; revised version received 19.08.25; accepted 19.08.25; published 02.10.25.

Please cite as:
Sze WT, Waki K, Nakada R, Yamauchi T, Nangaku M, Ohe K
Dietary Fiber Estimate of DialBetesPlus App Users: Secondary Analysis of Data From a Randomized Controlled Trial
JMIR Form Res 2025;9:e69340
URL: https://formative.jmir.org/2025/1/e69340
doi:10.2196/69340

© Wei Thing Sze, Kayo Waki, Ryohei Nakada, Toshimasa Yamauchi, Masaomi Nangaku, Kazuhiko Ohe. Originally published in JMIR Formative Research (https://formative.jmir.org), 2.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Motives for Cannabis Use and Readiness to Change Among Users of the "Stop-Cannabis" Mobile App: Cluster Analysis

Milena Wegener[1], MMed; Stéphane Rothen[2], MSc, PhD; Elise Dan-Glauser[3], PhD; Tania Lecomte[4,5], PhD; Stéphane Potvin[5,6], PhD; Lucien Rochat[2], PhD; Marissa Sjöblom, MMed; Germano Vera Cruz[7], PhD; Jean-François Etter[8], MPH, PhD; Yasser Khazaal[1,5,6,9], MD

[1]Faculty of Biology and Medicine, University of Lausanne, Lausanne, Switzerland
[2]Department of Mental Health and Psychiatry, Geneva University Hospitals, Geneva, Switzerland
[3]Institute of Psychology, University of Lausanne, Lausanne, Switzerland
[4]Department of Psychology, University of Montreal, Montreal, QC, Canada
[5]Research Center of the Institut Universitaire en Santé Mentale de Montréal, Montreal, QC, Canada
[6]Department of Psychiatry and Addictology, University of Montreal, Montreal, QC, Canada
[7]Department of Psychology, University of Picardie Jules Verne, University of Picardie Jules Verne, Amiens, France
[8]Institute of Global Health, University of Geneva, Geneva, Switzerland
[9]Service de médecine des addictions, University Hospital of Lausanne, Rue du Bugnon 23, Lausanne, Switzerland

Corresponding Author:
Yasser Khazaal, MD
Faculty of Biology and Medicine, University of Lausanne, Lausanne, Switzerland

Abstract

Background: Cannabis use is widespread and driven by diverse motives, ranging from recreational purposes to coping with psychological distress. Understanding the underlying reasons for cannabis use, their distribution across different subgroups of people who use cannabis, and how they relate to possible behavior change is essential for developing effective prevention and intervention strategies such as smartphone apps designed to support change.

Objective: The primary objective of the study was to determine whether analyzing profiles on the "Stop-cannabis" app (Institute of Global Health, University of Geneva, Switzerland) could reveal subgroups based on motives for cannabis use and readiness to change. A secondary objective was to explore differences among these subgroups in terms of problematic use and other indicators of change readiness.

Methods: This study analyzed data from 2578 individuals using the "Stop-cannabis app", a mobile app developed in Switzerland to support those seeking to manage their cannabis use. Participants completed validated questionnaires assessing motives for use (Marijuana Motives Measure [MMM]), readiness to change (Stages of Change Readiness and Treatment Eagerness Scale [SOCRATES]), and risk of problematic use (Alcohol, Smoking, and Substance Involvement Screening Test [ASSIST]). They also self-rated their "readiness for action," the "importance of change," and their "confidence in their ability to change." These assessments were part of the app's intervention model, with personalized feedback delivered based on participants' responses; no external incentives were offered. Cluster analysis was conducted to identify subgroups based on MMM and SOCRATES scores.

Results: In total, 3 distinct profiles emerged: the "individually coping users" (ICU), the "social and coping users" (SCU), and the "enhancement-seeking users" (ESU). ICU and SCU scored higher on coping motives compared with ESU, along with greater ambivalence and stronger recognition of problematic use, as measured by SOCRATES. They also scored higher on the ASSIST (indicating greater risk of problematic cannabis use), placed more importance on making behavioral changes, yet reported lower confidence in their ability to enact those changes. By contrast, ESU primarily used cannabis for recreational reasons and had low recognition of problematic use, despite being at moderate risk.

Conclusions: This research highlights that while motives for cannabis use are varied and individually nuanced, distinct subgroups can be identified, each with specific challenges. The findings align with previous research emphasizing the importance of coping motives in behavior change. Tailoring app content to reflect the unique profiles and needs of each subgroup may improve intervention outcomes. For instance, SCU and ICU may benefit from strategies targeting emotion regulation and alternative coping mechanisms, whereas ESU may respond better to brief motivational feedback and harm reduction strategies. Such tailored approaches can enhance the effectiveness of digital tools in promoting meaningful and long-term behavior change.

(JMIR Form Res 2025;9:e70849) doi:10.2196/70849

XSL•FO
RenderX

KEYWORDS

cannabis; motives; coping; cannabis problematic use; app; mHealth; mobile phone; mobile health

Introduction

Background

With approximately 200 million cannabis users worldwide [1], cannabis is a widely used substance with recognized harmful effects [2], particularly when it is smoked [3]. In Switzerland [4], 3.1% of the population were identified as current users (defined as past-month use), with 20% classified as heavy users engaging in daily or near-daily consumption. Despite this widespread use, few people with cannabis use disorders seek treatment due to various intrinsic and structural barriers [5].

The widespread accessibility of mobile health (mHealth) platforms presents a potential for addressing the treatment gap identified in substance use disorder [6]. The recent trend of developing mental health apps [7] has led to the emergence of a small number of apps focusing on cannabis use [8]. In Switzerland, the "Stop-cannabis" app was developed to be a stand-alone intervention [9]. It provides users with information about cannabis, gives out supportive messages to help them change their cannabis use, helps schedule a cannabis weaning plan, and allows access to an online discussion forum. Encouragingly, preliminary studies indicate users find this app useful [9].

While using the app, the consumers are prompted to complete questionnaires for a better assessment and understanding of their consumption. One of the questionnaires, the Marijuana Motives Measure (MMM) [10], is the most widely used scale to investigate motives of cannabis use. These motives are important because they vary between consumers and are associated with various factors related to the severity of cannabis-related issues [11] as well as different treatment outcomes [12,13].

Readiness to change is a pivotal element in the process of behavioral change, as it mirrors a person's recognition of their problematic cannabis use and their willingness to modify it. The app evaluates it through the Stages of Change Readiness and Treatment Eagerness Scale (SOCRATES) [14], which is a self-reporting questionnaire. The items the tool explores are articulated around the transtheoretical model of change. This model conceptualizes change as a progression through 5 set stages, extending from the precontemplation of change to its maintenance once change is achieved [15]

The World Health Organization recommends motivational interviewing to support patients through this process of behavioral change. This provenly effective [16] intervention is based on nonjudgmental dialogue [17] that helps individuals find the internal motivation to change their behavior. The "Stop-cannabis" app offers a "Consider Change" self-assessment survey that includes 3 important questions, frequently used as prompts during motivational interviews, asking users to rate their "readiness for action," the "importance of change," and their "confidence in [their] ability to change" cannabis use habits.

However, despite the growing availability of digital tools aiming to support cannabis-use–related behavior change, there remains a limited understanding, particularly in naturalistic settings, of who engages with these platforms, what drives their cannabis use, and how ready they are to change. Most existing research focuses either on app efficacy or usage metrics, without sufficiently exploring the heterogeneity of app users in terms of psychological drivers and change readiness. This gap limits the potential for delivering nuanced, individualized interventions through mental health apps. Characterizing users according to their motives for cannabis use and their readiness to change is thus essential. It may open avenues for developing more tailored interventions that could improve outcomes in real-world settings.

Objective

The primary objective of the study was to determine if analyzing cannabis users' profiles on the "Stop-cannabis" app could reveal subgroups of consumers based on their motives for cannabis use and their readiness to change said use. The second objective was then to study the characteristics of these subgroups, such as problematic cannabis use risk, while exploring their potential for change through factors such as "readiness for action," perceived "importance of change," and "confidence in one's ability to change."

Methods

Recruitment

Participants in the study were individuals who chose to download the app spontaneously. The app can be obtained at no cost from the Google Play Store (Android) and the App Store (iOS), and it does not include any commercial advertisement. Once downloaded, users were invited to anonymously provide their age and gender, after which they were prompted to use the Alcohol, Smoking, and Substance Involvement Screening Test (ASSIST) tool to assess their cannabis use. This enabled them to get personalized feedback and to customize the app. Users could also get feedback through an automated motivational interview feature, which used the MMM and SOCRATES surveys, as well as the "Consider Change" self-rating survey.

In this study, we performed a secondary analysis of data that had been previously used in a published study about the motives of cannabis use in cannabis consumers [18].

Participants

Between September 20, 2010, and January 13, 2020, 4077 user profiles for individuals from French-speaking countries (Switzerland, France, Canada, Belgium, and Luxembourg) were created on the app, and these individuals completed questionnaires provided in the app. Upon analysis, 95 duplicate user profiles and 450 underage (<18 y old) user profiles were excluded, leaving 3532 individuals' user profiles as the study sample. Of these, 2578 (73%) fully completed the ASSIST, SOCRATES, MMM, and "Consider Change" surveys and were selected as the analyzed sample (N). We did not consider user age and gender in this study. Users completed these

XSL•FO
RenderX

questionnaires as part of their natural engagement with the app's functionalities. They were not offered any incentives and could complete the questionnaires at any time after downloading the app, at their own discretion and pace.

Measures-Questionnaires

MMM

The MMM is a 25-item self-report questionnaire that is divided into 5 subscales. It was developed in 1998 [10] and subsequently validated in different samples of young adult marijuana users [19,20]. Each of the 5 subscales explores a different motive for cannabis use: "coping," "enhancement," "social," "conformity," and "expansion." Coping refers to using cannabis to relieve negative affective states (eg, "to forget my problems or my worries"). Enhancement means the substance is used to provide positive sensations or emotions (eg, "because it's fun," "I like the feeling"). Social motive indicates cannabis use together with peers to enhance a positive situation (eg, "because it improves parties or social gatherings"). Conformity alludes to its use to bond and avoid social rejection (eg, "to fit in with the group"). Expansion refers to the use of cannabis as a form of mind expansion (eg, "to expand my awareness," "to understand things differently"). Items are rated on a 5-point scale ranging from 1 (almost never or never) to 5 (almost always or always), the higher scores indicating a greater agreement with the motive.

SOCRATES

This tool was initially developed based on a cohort of adults with alcohol use disorders [14] but was adapted to other drugs in a later version (designated 8D). It assesses both preparedness for change in drug use habits and motivation to partake in treatment. In its current form, it consists of 19 self-reporting items, rated from 1 (strongly disagree) to 5 (strongly agree). The items explore 3 dimensions: "recognition," "ambivalence," and "taking steps." "Recognition" refers to the cognitive awareness of the problems caused by the substance use (eg, "I have serious problems with cannabis."). "Ambivalence" represents an individual's level of certainty about having a substance use problem (eg, "There are times when I wonder if I use too much cannabis."). "Taking steps" evaluates already implemented behavioral changes (eg, "I am actively doing things now to cut down or stop cannabis use."). The resulting scores are expressed in deciles, from 10 (very low) to 90 (very high).

ASSIST

This tool [21-23] was developed by the World Health Organization to screen for problematic substance use. It includes subscales for different substances, with the marijuana subscale examining use in the last 3 months. The severity of problematic cannabis use is assessed over 7 questions. Frequency of use, presence of craving, impact on health, and negative social consequences (eg, legal and financial problems) are rated on a 5-point Likert scale ranging from 0 ("never in the past 3 months") to 4 ("daily or almost daily"). Failure to perform daily responsibilities, concern expressed by relatives, and failure to control, cut down, or stop cannabis use are rated on a 3-point Likert scale (with 0="no, never," 1="yes in the past 3 months," and 2="yes but not in the past 3 months"). The compounded score ranges from 0 to 39, with a score above 27 indicating a high risk of dependence and of adverse health or social outcomes from the substance use.

"Consider Change" Survey

A total of 3 questions that are frequently asked in motivational interviews were combined to form a self-assessment tool. "Readiness for action," the "importance of change," and "confidence in one's ability to change" were rated on a visual scale and then converted into scores from 0 to 100.

Ethical Considerations

All participants provided informed online consent before data collection, explicitly agreeing that their anonymized data could be stored and used for research purposes. Only data from consenting participants were included in the study. Completion of the surveys was an integral part of the app intervention system and provided users with automated personalized feedback on their cannabis use. In addition, the data served to monitor app quality, as required by the funding source (Canton of Geneva).

All assessments were collected and reported anonymously. According to the Swiss Human Research Act (Chapter 1, Section 1, Article 2, Scope: 2c) [24], studies involving the collection of fully anonymized health-related personal data are exempt from ethical approval requirements. Therefore, this survey did not require submission to an ethics committee. The study was conducted in accordance with the principles of the Declaration of Helsinki.

Statistical Analysis

We conducted a cluster analysis on the 5 MMM subscales and the 3 SOCRATES subscales. Initially, we assessed the suitability of the data for clustering analysis using the Hopkins statistic [25]. Cluster analysis, in contrast to traditional linear models like regression, highlights individual differences [26] and enables the evaluation of complex, nonlinear interactions, thus offering enhanced ecological validity. Indeed, linear models might overlook significant relationships that are relevant for some individuals but not for all [27].

Traditionally, cluster analysis has often been conducted using a single method, typically the k-means method. However, with the diverse range of clustering methods available, it is essential to explore multiple approaches as no single method may be universally optimal [26]. Hence, five clustering methods were used: (1) k-means (with 50 starting points to mitigate the risk of local minima), (2) partitioning around medoids, (3) single linkage hierarchical clustering, (4) complete linkage hierarchical clustering, and (5) average linkage hierarchical clustering. The range of clusters considered for each method was set from 2 to 10, resulting in a total of 45 clusterings. Subsequently, we chose the optimal solution.

To determine the optimal number of clusters and clustering method for the dataset, we used a composite index based on 4 key characteristics of clustering [28]. These characteristics include (1) average distance within clusters, assessing the similarity of observations within each cluster; (2) separation index, ensuring distinctiveness between clusters; (3) Pearson-Gamma, measuring clustering quality in terms of dissimilarity structure approximation [29]; and (4) entropy of

XSL•FO
RenderX

cluster membership distribution, aiming for clusters of equivalent size [30]. We used the R function clusterbenchstats to generate standardized validity index values for comparison across indexes.

Statistical analysis was done on R-Studio 2023.12.1.402 [31] using R 4.3.3 [32]. We computed the composite index with the clusterbenchstats function from the fpc library [33]. All data and R scripts are available on request to the corresponding author.

Results

Cluster Model

Hopkins statistic values above 0.75 indicate a strong clustering tendency [34]. In this sample, the Hopkins statistic was 0.99, which confirms that cluster analysis is suitable in this sample. Based on the composite index, the optimal clustering solution was found for the k-means methods with 3 or 6 clusters. We selected the 3-cluster solution because the 6-cluster solution was not clinically meaningful, with an absence of significant difference between the clusters and a small cluster composed of only 130 individuals (Multimedia Appendix 1). In the 3-cluster solution, cluster 1 is composed of 43.87% (1131/2578) of user profiles, cluster 2 of 26.1% (674/2578), and cluster 3 of 29.9% (773/2578) profiles (Multimedia Appendix 2).

Clusters' Characteristics

Description of the 3 Clusters Regarding MMM 5 Factors' Scores

Cluster 1 (1131/2578, 43.87% of the sample) is characterized by low social (median 8/25, IQR 6-11/25) and expansion (median 7/25, IQR 6-9/25) motive scores compared with clusters 2 and 3, with a high score on coping (median 13/20, IQR 10-16/20) compared with cluster 3 (Figure 1, Table 1). Enhancement, while high (median 11/15, IQR 9-13/15), is slightly lower than the 2 following clusters, and conformity is low (median 4/20, IQR 4-4/20).

The individuals within cluster 1 can be described as "individually coping users" (ICU). They resort to cannabis consumption principally for the self-regulation of negative emotions, with a high score in coping, but a low score for social, conformity, and expansion motives of use.

Cluster 2 (674/2578, 26.1%) is also characterized by a high score on coping (median 15/20, IQR 12-18/20) and enhancement (median 14/15, IQR 12-15/15), but the users differ from the ones of cluster 1 as they present higher scores on social (median 16/25, IQR 13-20/25) and expansion motives (median 14/25, IQR 10-18/25). While minimal, the conformity motive score has a higher tendency (median 4/20, IQR 4-6/20) than the other 2 groups.

Cluster 2 is more heterogeneous and comprises individuals who can be described as "social and coping users" (SCU). Their cannabis use predominantly revolves around social aspects, enhancing social interactions (highest score in social motive) and reinforcing social cohesion (relatively high conformity), while in parallel serving as a pivotal coping mechanism (highest coping scores). Their motives for cannabis use are complex as they also score the highest in enhancement and expansion.

Cluster 3 (773/2578, 29.9%) includes users with low coping scores (median 8/20, IQR 6-11/20) and high enhancement scores (median 13/15, IQR 11-14/15). The social motive scores are moderate (median 12/25, IQR 9-16/25), higher than cluster 1 but lower than cluster 2. Expansion (median 9/25, IQR 6-12/25) and conformity (median 4/20, IQR 4-4/20) scores are low.

The consumers in cluster 3 can be described as "enhancement-seeking users" (ESU), their most salient motive being recreational (enhancement of positive activities). Coping is only a minor reason for use, with social and expansion components also being of secondary interest and conformity not being a concern.

XSL•FO
RenderX

Figure 1. Representation of MMM 5 factors' scores for the 3 clusters. MMM: Marijuana Motives Measure.

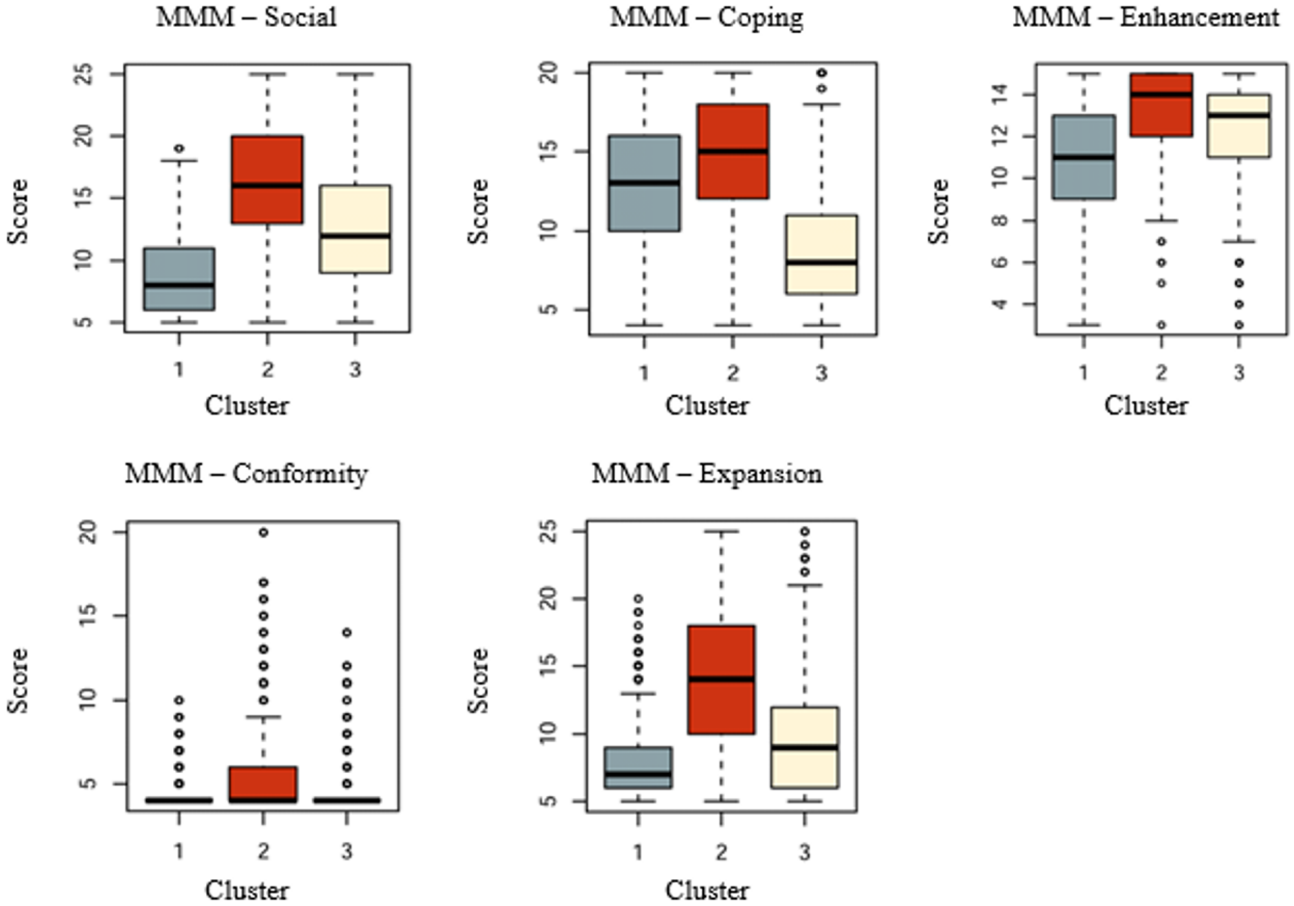

Table . Marijuana Motives Measure (MMM) 5 factors' scores for the 3 clusters and comparison.

MMM factor	Cluster 1 (ICU[a]; n=1131), median (IQR)	Cluster 2 (SCU[b]; n=674), median (IQR)	Cluster 3 (ESU[c]; n=773), median (IQR)	*P* value[d]
Social	8 (6-11)	16 (13 - 20)	12 (9 - 16)	<.001[e,f,g]
Coping	13 (10 - 16)	15 (12 - 18)	8 (6-11)	<.001[e,f,g]
Enhancement	11 (9 - 13)	14 (12 - 15)	13 (11 - 14)	<.001[e,f,g]
Conformity	4 (4-4)	4 (4-6)	4 (4-4)	<.001;[e,g] .15[f]
Expansion	7 (6-9)	14 (10 - 18)	9 (6-12)	<.001[e,f,g]

[a]ICU: individually coping users.

[b]SCU: social and coping users.

[c]ESU: enhancement-seeking users.

[d]Kruskal-Wallis rank sum test.

[e]Post hoc test comparing cluster 1 and cluster 2, Bonferroni adjusted *P* value <.05.

[f]Post hoc test comparing cluster 1 and cluster 3, Bonferroni adjusted *P* value <.05.

[g]Post hoc test comparing cluster 2 and cluster 3, Bonferroni adjusted *P* value <.05.

Description of the 3 Clusters Regarding SOCRATES Scores

Cluster 1 (ICU) is characterized by high recognition score (median 30/35, IQR 27-33/35) and ambivalence (median 16/20, IQR 15-18/20) scores, compared with cluster 3 (Figure 2, Table 2). "Taking steps" score (median 25/40, IQR 20-30/40) was moderate.

Cluster 2 (SCU) follows a similar pattern, with high recognition (median 29/35, IQR 25-32/35) and ambivalence (median 16/20, IQR 14-18/20) scores, while "taking steps" score (median 25/40, IQR 20-31/40) was moderate.

Cluster 3 (ESU) differs in presenting with the lowest recognition (median 20/35, IQR 16-23/35) and ambivalence (median 12/20, IQR 10-14/20) scores, and slightly lower "taking steps" scores

XSL•FO
RenderX

(median 20/40, IQR 16-25/40) compared with the other 2 groups.

Figure 2. Representation of SOCRATES subscales between the 3 clusters. SOCRATES: Stage of Change Readiness and Treatment Eagerness Scale.

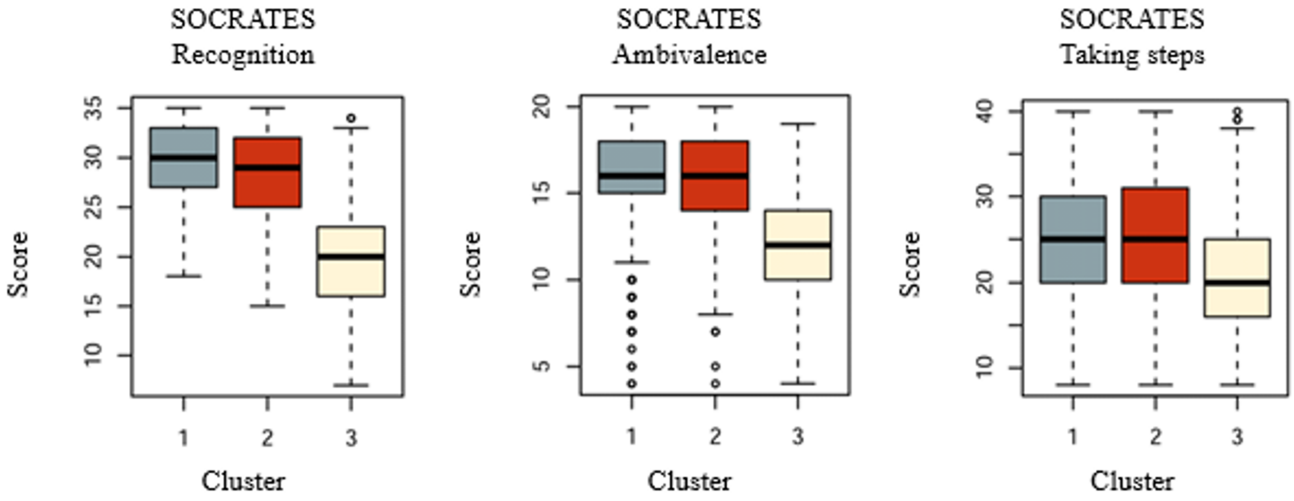

Table . Stage of Change Readiness and Treatment Eagerness Scale (SOCRATES) subscale scores for the 3 clusters and comparison.

SOCRATES subscale	Cluster 1 (ICU[a]; n=1131), median (IQR)	Cluster 2 (SCU[b]; n=674), median (IQR)	Cluster 3 (ESU[c]; n=773), median (IQR)	*P* value[d]
Recognition	30 (27 - 33)	29 (25 - 32)	20 (16 - 23)	<.001[e, f, g]
Ambivalence	16 (15 - 18)	16 (14 - 18)	12 (10 - 14)	<.001[f, g]; >.99[e]
Taking steps	25 (20 - 30)	25 (20 - 31)	20 (16 - 25)	<.001[f, g]; .50[e]

[a]ICU: individually coping users.

[b]SCU: social and coping users.

[c]ESU: enhancement-seeking users.

[d]Kruskal-Wallis rank sum test.

[e]Post hoc test comparing Cluster 1 and Cluster 2, Bonferroni adjusted *P* value <.05.

[f]Post hoc test comparing Cluster 1 and Cluster 3, Bonferroni adjusted *P* value <.05.

[g]Post hoc test comparing Cluster 2 and Cluster 3, Bonferroni adjusted *P* value <.05.

Description of the Other Clinical Measures in the 3 Clusters

Significant effects of cluster belonging were obtained on all other clinical measures (ASSIST and "Consider Change" survey), between clusters 1 and 2 compared with cluster 3. Between clusters 1 and 2, only "importance of change" was significantly different (Table 3).

Most participants (Multimedia Appendix 3) scored into the moderate (score between 4 and 26) and high-risk (score more than 27) cannabis use groups, according to ASSIST scores [14]. The median ASSIST score for the whole study sample was 27 (IQR 21 - 32), which is the cutoff between moderate and high-risk groups.

XSL•FO
RenderX

Table . Association between clusters and other clinical measures.

Clinical measure	Cluster 1 (ICU[a]; n=1131), median (IQR)	Cluster 2 (SCU[b]; n=674), median (IQR)	Cluster 3 (ESU[c]; n=773), median (IQR)	*P* value[d]
ASSIST[e]	30 (25 - 33)	30 (25 - 34)	21 (15 - 26)	<.001[g, h]
"Consider Change"				
Importance of change	99 (80 - 99)	90 (74 - 99)	66 (46 - 80)	<.001[f, g, h]
Readiness for action	80 (62 - 99)	80 (64 - 99)	68 (40 - 81)	<.001[g, h]
Confidence in one's ability to change	43 (24 - 70)	41 (22 - 68)	52 (30 - 80)	<.001[g, h]

[a]ICU: individually coping users.

[b]SCU: social and coping users.

[c]ESU: enhancement-seeking users.

[d]Kruskal-Wallis rank sum test.

[e]ASSIST: Alcohol, Smoking and Substance Involvement Screening Test.

[f]Post hoc test comparing cluster 1 and cluster 3, Bonferroni adjusted *P* value <.05.

[g]Post hoc test comparing cluster 1 and cluster 2, Bonferroni adjusted *P* value <.05.

[h]Post hoc test comparing cluster 2 and cluster 3, Bonferroni adjusted *P* value <.05.

ICU and SCU shared identical median ASSIST scores (median 30, IQR 25 - 34), which were significantly higher than those of ESU (median 21, IQR 15 - 26).

Cluster 1 (ICU) scored the highest when rating the "importance of change," with a significantly higher score than cluster 2 (SCU). Both groups also scored significantly higher than cluster 3 (ESU).

Regarding "readiness for action," clusters 1 (ICU) and 2 (SCU) scored similarly but significantly higher than cluster 3 (ESU).

Cluster 3 (ESU) achieved the highest scores in "confidence in one's ability to change," significantly surpassing clusters 1 (ICU) and 2 (SCU).

Discussion

Results Analysis

The aim of this study was to identify profiles of cannabis consumers based on their motives for use and their stage of change readiness and treatment eagerness. Both statistically significant and clinically relevant consumer groups were identified: ICU, SCU, and ESU.

Both ICU and SCU clusters disclosed largely using cannabis as a coping mechanism, which in cannabis users has been linked with poorer mental health, significant psychopathological symptoms, and increased psychological stress compared with noncannabis users [35]. Both groups also reported high ASSIST scores, indicating a high risk of problematic cannabis use, but with relatively high levels of problem recognition. This recognition, combined with high willingness to act scores and great importance given to change, offers opportunities for targeted health interventions.

Within this perspective, understanding the underlying coping motives may play a critical role. For instance, in this study, coping motives are positively associated with problematic use—as measured by the ASSIST—and this overlap was reflected in our clustering results. The assessment of motives provides, however, conceptually and clinically distinct information. The ASSIST captures the severity and consequences of substance use, whereas coping motives reflect why individuals engage in cannabis use in the first place. This distinction is critical because motives can serve as predictive and explanatory variables, informing not just current use severity but also potential relapse risk, underlying psychological vulnerabilities (eg, emotion regulation difficulties), and targets for intervention. In particular, coping motives have consistently emerged as one of the strongest and most reliable psychological predictors of both cannabis use frequency and cannabis-related problems [11,18]. Importantly, 2 individuals with similar ASSIST scores might differ in their underlying motives—one using cannabis primarily to cope with distress, and another for social or enhancement reasons—which may call for different intervention strategies. Thus, including motivational profiles allows for a more personalized and mechanism-based understanding of users, which is particularly relevant in the context of digital health tools aiming to offer tailored, scalable intervention.

As it stands, "taking steps" scores, according to the SOCRATES scores, remained moderate in ICU and SCU clusters, probably hampered by high ambivalence and low self-confidence to change. These last 2 factors may reflect pre-existing psychopathological symptoms or psychosocial distress, with cannabis being used to cope with and alleviate these symptoms [35-40]. This dimension should probably be specifically addressed for better therapeutic efficacy. In that respect, research has shown that reducing coping as a motive for cannabis use is a consistently positive predictive factor for a decrease in cannabis-related problems over the next months during treatment [13].

The ESU cluster reported the lowest ASSIST scores, which still translate into a moderate risk of problematic cannabis use. While low coping scores hint at a lower prevalence or burden of psychological comorbidity, individuals are still at risk for the

XSL•FO
RenderX

negative physical consequences of smoking cannabis [3]. In other studies, enhancement motives remained predictive of psychosocial problems associated with cannabis use [41]. While enhancement was a highly endorsed motive across all clusters, it still holds discriminative value, as the clusters differed significantly on this dimension—cluster 2 scored highest, cluster 1 lowest, and cluster 3 intermediate. Notably, in cluster 3, enhancement was the most prominent motive relative to other MMM dimensions, justifying its label as the "enhancement-seeking" cluster.

Recognition of the problem was the lowest in the ESU cluster among the 3 groups, and in regard to this, low ambivalence scores should be interpreted as reflecting low awareness of the problem. These results could suggest that such individuals might be more likely to return to the pre-contemplation stage of change despite having downloaded an app aimed at reducing or stopping their use of cannabis. In this case, motivational dialogue might help increase problem recognition and the overcoming of resistance to change. ESU individuals have higher self-confidence in changing their cannabis use, which may be related to lower reliance on cannabis as a coping mechanism for psychological distress.

The SCU cluster was the most intricate, with several salient motives for cannabis use, rendering prognosis analyses and planning treatment interventions more complex. In another study, young adults with social motives for use reported similar levels of distress to nonusers [35]. Notwithstanding these results, evidence suggests that multiplicity of motives for use correlates with increased consumption and a higher likelihood of experiencing adverse consequences of cannabis [42]. SCU also exhibited high rates of use for expansion motives. A decrease in expansion as a reason for cannabis use following treatment is surprisingly associated with poorer treatment outcomes [12].

The distinct features of the 3 main clusters clearly illustrate the multifaceted nature of cannabis use, highlighting diverse motives for use and hence the need for tailored targets of intervention. Targeted therapeutic interventions rely on such multidimensional phenotypic assessments [43] to enhance their efficiency. Motives of use offer a good treatment target, as different motives are associated with either favorable or unfavorable treatment outcomes [12]. As an example, cognitive behavioral therapy, which is sometimes dispensed through computer-based training, could be enhanced by the adjunction of specific modules designed to change motives of cannabis use [44].

Treatment strategies may be effectively tailored according to user profiles identified through clustering. For example, individuals in the ICU cluster appear to use cannabis primarily to cope with psychological distress. They show high recognition of the problem and a strong willingness to change, but report low self-confidence and only moderate levels of action-taking. For this group, tailored interventions could focus on emotion regulation, stress management, and the development of alternative coping strategies. In addition, motivational enhancement techniques may be used to strengthen self-efficacy and support-maintained behavior change.

Individuals in the ESU cluster tend to use cannabis for excitement, exhibit low recognition of the problem, low ambivalence, but express high confidence in their ability to change. In this case, brief interventions—such as personalized assessments and normative feedback—may be appropriate. These should help participants explore the discrepancy between their current behavior and personal goals. Harm reduction strategies (eg, promoting safer cannabis use) should also be offered to support goal-congruent outcomes.

Members of the SCU cluster demonstrate multiple motivations for use (coping, social, and expansion), combined with a high willingness to change but low self-confidence. A multicomponent treatment plan should be considered, including social skills training (to address socially motivated use) and emotion regulation techniques. Motivational interviewing, harm reduction strategies, and interventions aimed at enhancing self-confidence are also recommended. Peer support groups may further provide meaningful sources of social validation and connection.

These cluster-specific considerations could inform the development of future app-based interventions, enabling more precise and multimodal tailoring of treatment. Such proposals merit empirical evaluation in future studies.

Limitations

The app collected no data on the overall chronology of an individual's cannabis use, both regarding its overall duration (eg, to distinguish long-term chronic users) and possible time-dependent fluctuations in it. Consumer age was also not considered in the study design. Motives for cannabis use might therefore vary depending on consumer age (eg, phases of life) and their life-long history of use (new users vs chronic users), potentially providing an alternative explanation for the significant differences between the clusters. Fundamentally, this does not influence targets for intervention but might alter their modalities. Gender was not included as a variable in this study design and should be investigated further, as women tend to have higher coping scores and lower social motives scores than men [18].

A further limitation of this study is the absence of precise information regarding when, during their engagement with the app, users completed the questionnaires. As survey completion was anonymous, voluntary, and untied to specific stages of app use, we cannot determine whether responses were given at the outset, during active use of behavior change tools, or following exposure to features such as the within-app social network or psychoeducation content. These interactions may have already influenced the answers of the participants. Nevertheless, the findings revealed meaningful differences between user profiles, underscoring the heterogeneity of app users and suggesting varied needs that could be addressed through more tailored interventions.

This cross-sectional study yielded promising results, but further study, including studies with longitudinal designs, is warranted. In addition, due to the design of the study, participants were spontaneous users of a mobile app, and it is therefore unclear if the results can be extrapolated to the general population.

In addition, the choice of a more robust 3-cluster model rather than a model with a higher number of clusters can conceal other

XSL•FO
RenderX

subtypes of consumers. The SCU cluster might be a composite subgroup based on its intrinsic complexity and diversity of motives.

Finally, the very salient coping motive should also be further explored, with more detailed exploration of subgroups of anxiety-coping, depression-coping, and boredom-coping [45]. Numerous studies have shown both the clinical significance and the heterogeneity of coping as a motive [38-40]. Cannabis use as a therapeutic intervention for sleep disorders [40,46] or medical conditions (eg, chronic pain) was also not explored as a possible motive. Be it medically prescribed or self-administered, cannabis use in this setting has also been linked to problematic use [47]. However, given the increasing global use of cannabis for medicinal purposes, including in some of the countries from where the participants come from (ie, Canada and Switzerland), it is possible that coping-related items capture not only emotional coping but also elements of self-medication for physical or psychological symptoms. In the absence of explicit questions related to medical cannabis use, this study cannot disentangle therapeutic motives from other coping motives.

Conclusions

In this study, we identified 3 distinct profiles of individuals engaging with a cannabis behavior change app: the ICU, the SCU, and the ESU.

ICU and SCU demonstrated higher coping motive scores compared with ESU, along with greater ambivalence and stronger recognition of their problematic use, as measured by the SOCRATES questionnaires. In addition, they scored higher on the ASSIST, placed greater importance on making behavioral changes, yet reported lower confidence specifically in their ability to enact those changes. These findings align with existing research emphasizing the critical role of coping motives in behavior change and highlight the need to tailor the content and design of digital interventions to better address the unique needs and characteristics of these distinct user subgroups. By considering these clusters, digital tools can become more effective in supporting behavior change and achieving long-term success for diverse populations seeking help.

The heterogeneity of cannabis use is a challenge for researchers and practitioners alike, and this study showed that a cluster approach is a useful method for discovering different profiles of users. mHealth approaches are proving useful for research as they facilitate the collection of standardized data in high numbers.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Six-cluster solution.
[PNG File, 53 KB - formative_v9i1e70849_app1.png]

Multimedia Appendix 2
Three-cluster model repartition.
[PNG File, 6 KB - formative_v9i1e70849_app2.png]

Multimedia Appendix 3
ASSIST scores and corresponding risk levels in the sample. ASSIST: Alcohol, Smoking, and Substance Involvement Screening Test.
[PNG File, 190 KB - formative_v9i1e70849_app3.png]

References

1. Connor JP, Stjepanović D, Le Foll B, Hoch E, Budney AJ, Hall WD. Cannabis use and cannabis use disorder. Nat Rev Dis Primers 2021 Feb 25;7(1):16. [doi: 10.1038/s41572-021-00247-4] [Medline: 33627670]
2. Cohen K, Weizman A, Weinstein A. Positive and negative effects of cannabis and cannabinoids on health. Clin Pharma and Therapeutics 2019 May;105(5):1139-1147. [doi: 10.1002/cpt.1381]
3. Archie SR, Cucullo L. Harmful effects of smoking cannabis: a cerebrovascular and neurological perspective. Front Pharmacol 2019 Dec 6;10(1481):1481. [doi: 10.3389/fphar.2019.01481] [Medline: 31920665]
4. Gmel G, Kuendig H, Notari L, Gmel C. Addiction monitoring Switzerland- consumption of alcohol, tobacco, and illegal drugs in Switzerland in 2016 [Report in German]. : Addiction Switzerland; 2017 Jul URL: https://www.suchtmonitoring.ch/docs/library/gmel_5lbj5rqv9y5i.pdf [accessed 2025-09-18]
5. Gates P, Copeland J, Swift W, Martin G. Barriers and facilitators to cannabis treatment. Drug Alcohol Rev 2012 May;31(3):311-319. [doi: 10.1111/j.1465-3362.2011.00313.x] [Medline: 21521384]

XSL•FO
RenderX

6. Evans-Lacko S, Aguilar-Gaxiola S, Al-Hamzawi A, et al. Socio-economic variations in the mental health treatment gap for people with anxiety, mood, and substance use disorders: results from the WHO World Mental Health (WMH) surveys. Psychol Med 2018 Jul;48(9):1560-1571. [doi: 10.1017/S0033291717003336] [Medline: 29173244]
7. Lecomte T, Potvin S, Corbière M, et al. Mobile apps for mental health issues: meta-review of meta-analyses. JMIR Mhealth Uhealth 2020 May 29;8(5):e17458. [doi: 10.2196/17458] [Medline: 32348289]
8. Ramo DE, Popova L, Grana R, Zhao S, Chavez K. Cannabis mobile apps: a content analysis. JMIR Mhealth Uhealth 2015 Aug 12;3(3):e81. [doi: 10.2196/mhealth.4405] [Medline: 26268634]
9. Monney G, Penzenstadler L, Dupraz O, Etter JF, Khazaal Y. mHealth app for cannabis users: satisfaction and perceived usefulness. Front Psychiatry 2015;6(120):120. [doi: 10.3389/fpsyt.2015.00120] [Medline: 26379561]
10. Simons J, Correia CJ, Carey KB, Borsari BE. Validating a five-factor Marijuana Motives Measure: relations with use, problems, and alcohol motives. J Couns Psychol 1998;45(3):265-273. [doi: 10.1037/0022-0167.45.3.265]
11. Bresin K, Mekawi Y. Do marijuana use motives matter? Meta-analytic associations with marijuana use frequency and problems. Addict Behav 2019 Dec;99:106102. [doi: 10.1016/j.addbeh.2019.106102] [Medline: 31473571]
12. Banes KE, Stephens RS, Blevins CE, Walker DD, Roffman RA. Changing motives for use: outcomes from a cognitive-behavioral intervention for marijuana-dependent adults. Drug Alcohol Depend 2014 Jun 1;139(41–46):41-46. [doi: 10.1016/j.drugalcdep.2014.02.706] [Medline: 24685559]
13. Blevins CE, Banes KE, Stephens RS, Walker DD, Roffman RA. Change in motives among frequent cannabis-using adolescents: predicting treatment outcomes. Drug Alcohol Depend 2016 Oct 1;167(175–181):175-181. [doi: 10.1016/j.drugalcdep.2016.08.018] [Medline: 27577862]
14. Miller WR, Tonigan JS. Assessing drinkers' motivation for change: the Stages of Change Readiness and Treatment Eagerness Scale (SOCRATES). Psychol Addict Behav 1996;10(2):81-89. [doi: 10.1037/0893-164X.10.2.81]
15. Prochaska JO, DiClemente CC. Stages and processes of self-change of smoking: toward an integrative model of change. J Consult Clin Psychol 1983 Jun;51(3):390-395. [doi: 10.1037//0022-006x.51.3.390] [Medline: 6863699]
16. Calomarde-Gómez C, Jiménez-Fernández B, Balcells-Oliveró M, Gual A, López-Pelayo H. Motivational interviewing for cannabis use disorders: a systematic review and meta-analysis. Eur Addict Res 2021;27(6):413-427. [doi: 10.1159/000515667] [Medline: 33965941]
17. Lundahl B, Droubay BA, Burke B, et al. Motivational interviewing adherence tools: a scoping review investigating content validity. Patient Educ Couns 2019 Dec;102(12):2145-2155. [doi: 10.1016/j.pec.2019.07.003] [Medline: 31514978]
18. Rochat L, Rothen S, Edel Y, et al. Measurement invariance of the Marijuana Motives Measure among men and women using Stop Cannabis app. Addict Behav 2024 Jan;148(107866):107866. [doi: 10.1016/j.addbeh.2023.107866] [Medline: 37748226]
19. Benschop A, Liebregts N, van der Pol P, et al. Reliability and validity of the Marijuana Motives Measure among young adult frequent cannabis users and associations with cannabis dependence. Addict Behav 2015 Jan;40:91-95. [doi: 10.1016/j.addbeh.2014.09.003] [Medline: 25240105]
20. Zvolensky MJ, Vujanovic AA, Bernstein A, Bonn-Miller MO, Marshall EC, Leyro TM. Marijuana use motives: a confirmatory test and evaluation among young adult marijuana users. Addict Behav 2007 Dec;32(12):3122-3130. [doi: 10.1016/j.addbeh.2007.06.010] [Medline: 17602842]
21. Khan R, Chatton A, Thorens G, et al. Validation of the French version of the Alcohol, Smoking and Substance Involvement Screening Test (ASSIST) in the elderly. Subst Abuse Treat Prev Policy 2012 Jun 20;7:14. [doi: 10.1186/1747-597X-7-14] [Medline: 22538114]
22. Humeniuk R, Ali R, Babor TF, et al. Validation of the Alcohol, Smoking And Substance Involvement Screening Test (ASSIST). Addiction 2008 Jun;103(6):1039-1047. [doi: 10.1111/j.1360-0443.2007.02114.x] [Medline: 18373724]
23. Khazaal Y, Chatton A, Monney G, et al. Internal consistency and measurement equivalence of the cannabis screening questions on the paper-and-pencil face-to-face ASSIST versus the online instrument. Subst Abuse Treat Prev Policy 2015 Mar 8;10:8. [doi: 10.1186/s13011-015-0002-9] [Medline: 25886462]
24. The Swiss Human Research Act; Chapter 1, Section 1, Article 2, Scope: 2c. Fedlex. URL: https://www.fedlex.admin.ch/eli/cc/2013/617/en [accessed 2025-09-17]
25. Lawson RG, Jurs PC. New index for clustering tendency and its application to chemical problems. J Chem Inf Comput Sci 1990 Feb 1;30(1):36-41. [doi: 10.1021/ci00065a010]
26. Rapkin BD, Luke DA. Cluster analysis in community research: epistemology and practice. American J of Comm Psychol 1993 Apr;21(2):247-277. [doi: 10.1007/BF00941623]
27. von Eye A, Bogat GA. Person-oriented and variable-oriented research: concepts, results, and development. Merrill Palmer Q (Wayne State Univ Press) 2006;52(3):390-420. [doi: 10.1353/mpq.2006.0032]
28. Akhanli SE, Hennig C. Comparing clusterings and numbers of clusters by aggregation of calibrated clustering validity indexes. Stat Comput 2020 Sep;30(5):1523-1544. [doi: 10.1007/s11222-020-09958-2]
29. Halkidi M, Batistakis Y, Vazirgiannis M. On clustering validation techniques. J Intell Inf Syst 2001 Dec;17(2-3):107-145. [doi: 10.1023/A:1012801612483]
30. Meilă M. Comparing clusterings—an information based distance. J Multivar Anal 2007 May;98(5):873-895. [doi: 10.1016/j.jmva.2006.11.013]

XSL•FO
RenderX

31. Posit. URL: http://www.posit.co [accessed 2025-09-17]
32. The R Project for Statistical Computing. URL: https://www.R-project.org/ [accessed 2025-09-17]
33. Hennig C. Fpc: flexible procedures for clustering. The Comprehensive R Archive Network. 2023. URL: https://CRAN.R-project.org/package=fpc [accessed 2025-09-17]
34. Drab K, Daszykowski M. Clustering in analytical chemistry. J AOAC Int 2014;97(1):29-38. [doi: 10.5740/jaoacint.sgedrab] [Medline: 24672857]
35. Brodbeck J, Matter M, Page J, Moggi F. Motives for cannabis use as a moderator variable of distress among young adults. Addict Behav 2007 Aug;32(8):1537-1545. [doi: 10.1016/j.addbeh.2006.11.012] [Medline: 17178197]
36. Bonn-Miller MO, Zvolensky MJ, Bernstein A. Marijuana use motives: concurrent relations to frequency of past 30-day use and anxiety sensitivity among young adult marijuana smokers. Addict Behav 2007 Jan;32(1):49-62. [doi: 10.1016/j.addbeh.2006.03.018] [Medline: 16647822]
37. Bujarski SJ, Norberg MM, Copeland J. The association between distress tolerance and cannabis use-related problems: the mediating and moderating roles of coping motives and gender. Addict Behav 2012 Oct;37(10):1181-1184. [doi: 10.1016/j.addbeh.2012.05.014] [Medline: 22698896]
38. Glodosky NC, Cuttler C. Motives matter: cannabis use motives moderate the associations between stress and negative affect. Addict Behav 2020 Mar;102:106188. [doi: 10.1016/j.addbeh.2019.106188] [Medline: 31706141]
39. Scarfe ML, Muir C, Rowa K, Balodis I, MacKillop J. Getting high or getting by? An examination of cannabis motives, cannabis misuse, and concurrent psychopathology in a sample of general community adults. Subst Abuse 2022;16:11782218221119070. [doi: 10.1177/11782218221119070] [Medline: 36051509]
40. Livingston NR, Hetelekides E, Bravo AJ, Looby A, Norms S. Negative affect regulation and marijuana use in college students: evaluating the mediating roles of coping and sleep motives. J Psychoactive Drugs 2023;55(2):203-212. [doi: 10.1080/02791072.2022.2054747] [Medline: 35341474]
41. Schultz NR, Bassett DT, Messina BG, Correia CJ. Differential role of cannabis use motives in predicting impairment across three measures. J Stud Alcohol Drugs 2019 Jan;80(1):26-31. [doi: 10.15288/jsad.2019.80.26] [Medline: 30807271]
42. Stevens AK, Drohan MM, Boyle HK, White HR, Jackson KM. More reasons, more use and problems? Examining the influence of number of motives on consumption and consequences across alcohol-only, cannabis-only, and simultaneous-use days. J Stud Alcohol Drugs 2021 Nov;82(6):782-791. [doi: 10.15288/jsad.2021.82.782] [Medline: 34762038]
43. Litten RZ, Ryan ML, Falk DE, Reilly M, Fertig JB, Koob GF. Heterogeneity of alcohol use disorder: understanding mechanisms to advance personalized treatment. Alcohol Clin Exp Res 2015 Apr;39(4):579-584. [doi: 10.1111/acer.12669] [Medline: 25833016]
44. Carroll KM. The profound heterogeneity of substance use disorders: implications for treatment development. Curr Dir Psychol Sci 2021 Aug;30(4):358-364. [doi: 10.1177/09637214211026984] [Medline: 34483503]
45. Biolcati R, Passini S. Development of the Substance Use Motives Measure (SUMM): a comprehensive eight-factor model for alcohol/drugs consumption. Addict Behav Rep 2019 Dec;10:100199. [doi: 10.1016/j.abrep.2019.100199] [Medline: 31517018]
46. Chabrol H, Ducongé E, Casas C, Roura C, Carey KB. Relations between cannabis use and dependence, motives for cannabis use and anxious, depressive and borderline symptomatology. Addict Behav 2005 May;30(4):829-840. [doi: 10.1016/j.addbeh.2004.08.027] [Medline: 15833585]
47. Metrik J, Jackson K, Bassett SS, Zvolensky MJ, Seal K, Borsari B. The mediating roles of coping, sleep, and anxiety motives in cannabis use and problems among returning veterans with PTSD and MDD. Psychol Addict Behav 2016 Nov;30(7):743-754. [doi: 10.1037/adb0000210] [Medline: 27786514]

Abbreviations

ASSIST: Alcohol, Smoking, and Substance Involvement Screening Test
ESU: enhancement-seeking users
ICU: individually coping users
mHealth: mobile health
MMM: Marijuana Motives Measure
SCU: social and coping users
SOCRATES: Stages of Change Readiness and Treatment Eagerness Scale

XSL•FO
RenderX

Edited by A Mavragani; submitted 03.01.25; peer-reviewed by J Thrul; revised version received 29.05.25; accepted 14.06.25; published 03.10.25.

Please cite as:
Wegener M, Rothen S, Dan-Glauser E, Lecomte T, Potvin S, Rochat L, Sjöblom M, Vera Cruz G, Etter JF, Khazaal Y
Motives for Cannabis Use and Readiness to Change Among Users of the "Stop-Cannabis" Mobile App: Cluster Analysis
JMIR Form Res 2025;9:e70849
URL: https://formative.jmir.org/2025/1/e70849
doi:10.2196/70849

© Milena Wegener, Stéphane Rothen, Elise Dan-Glauser, Tania Lecomte, Stéphane Potvin, Lucien Rochat, Marissa Sjöblom, Germano Vera Cruz, Jean-François Etter, Yasser Khazaal. Originally published in JMIR Formative Research (https://formative.jmir.org), 3.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Professional Support Through a Tailor-Made Mobile App to Reduce Stress and Depressive Symptoms Among Family Caregivers of People With Dementia: Mixed Methods Pilot Study

Aber Sharon Kagwa[1], RN, MSc; Jessica Longhini[2], RN, PhD; Muhammed Nazmul Islam[3], MSc; Sofia Vikström[4], OT, PhD; Åsa Dorell[1], RN, PhD; Hanne Konradsen[1,5,6], RN, PhD; Zarina Nahar Kabir[1], PhD

[1]Division of Nursing, Department of Neurobiology, Care Sciences and Society, Karolinska Institute, Alfred Nobels Allé 23, Huddinge, Stockholm, Sweden

[2]Laboratory of Studies and Evidence-based Nursing, Department of Cardiac, Thoracic, Vascular Sciences and Public Health, University of Padova, Padua, Italy

[3]Department of Research and Evaluation, Sajida Foundation, Dhaka, Bangladesh

[4]Division of Occupational Therapy, Department of Neurobiology, Care Sciences and Society, Karolinska Institutet, Huddinge, Stockholm, Sweden

[5]Department of Geriatrics, Hvidovre Hospital, Hvidovre, Denmark

[6]Steno Diabetes Center, Copenhagen, Denmark

Corresponding Author:
Aber Sharon Kagwa, RN, MSc
Division of Nursing, Department of Neurobiology, Care Sciences and Society, Karolinska Institute, Alfred Nobels Allé 23, Huddinge, Stockholm, Sweden

Abstract

Background: Providing informal care to people with dementia living at home can be challenging and may cause caregiver stress and depression. Interventions delivered through mobile apps provide innovative solutions for community-based social care professionals to address the increasing support needs of family caregivers (FCs) of people with dementia.

Objective: This study aimed to examine, among FCs of people with dementia living at home, (1) the potential association between professional support provided through a mobile app and caregiver stress and depressive symptoms, (2) types of support provided through chat interactions between FCs and social care professionals, and (3) how support provided through a mobile app relates to changes in caregiver stress and depressive symptoms.

Methods: A mixed methods pilot study integrated quantitative pre- and postintervention data with qualitative logged chat data. FCs of people with dementia living at home (n=35) were recruited to test a tailor-made mobile app over 8 weeks. The primary and secondary outcome measures were caregiver stress and depressive symptoms, respectively. Descriptive statistics were used to summarize sociodemographic factors; inferential statistics were used to analyze mean differences in outcomes pre- and postintervention. FCs were divided into 3 groups based on changes in caregiver stress scores between pre- and postintervention. Generalized linear model analyses determined the association between participation in the intervention and caregiver stress and depressive symptoms, adjusting for age, gender, and relationship to the person with dementia. Logged chat data were analyzed using summative content analysis to identify types of support provided and received. Changes in caregiver stress were integrated with chat data to determine patterns in types of support received.

Results: The mean age of FCs was 69.4 (SD 11.9) years, with most being women (28/35, 80%), partners (24/35, 68.6%), and living with the person with dementia (26/35, 74%). The mean score of caregiver stress was marginally higher postintervention (24.1, SD 9.3) than preintervention (23.9, SD 9.2), whereas the mean score of depressive symptoms decreased from pre- (6.5, SD 5.1) to postintervention (6.2, SD 5.2). These differences were not statistically significant. Regression analyses showed that participation in the intervention was not statistically significantly associated with caregiver stress (β=0.171, α=.05; P=.86) or depressive symptoms (β=−0.293, α=.05; P=.75) after adjusting for age, gender, and relationship to the person with dementia. However, mixed methods analysis at the subgroup level suggested that frequent tailored support by social care professionals delivered through a mobile app may reduce caregiver stress among FCs of people with dementia living at home.

Conclusions: The study highlights the importance of providing frequent and individualized support to meet the needs of FCs of people with dementia. Findings from this study may help community-based social care providers plan and organize digital support content provided to FCs of people with dementia living at home.

International Registered Report Identifier (IRRID): RR2-10.1186/s12877-020-01712-7

(JMIR Form Res 2025;9:e75113) doi:10.2196/75113

XSL•FO
RenderX

KEYWORDS

caregiver stress; dementia; dementia care; depressive symptoms; eHealth; mHealth; mobile app; social care professional; support; mobile health; informal caregiver

Introduction

Dementia

Dementia is a progressive condition with no current cure [1]. The condition is considered a global public health challenge and is the seventh leading cause of death worldwide [2]. It is characterized by impairments such as cognitive, functional, and behavioral [1] affecting the individual's memory, behavior, and daily functioning [2]. In 2020, the number of people living with dementia worldwide was 55 million, with a projection of 139 million in 2050. In Sweden, 150,000 people lived with dementia in 2019, with a projection of 250,000 by 2050 [3].

Being a Family Caregiver to a Person With Dementia

Most of the care provided to people with dementia in communities is provided through informal care by unpaid family caregivers (FCs) such as family members and friends [4]. Therefore, FCs play an integral part in the health care system, providing everyday care that is complex, challenging, and prolonged [1]. In 2019, the World Health Organization (WHO) estimated that FCs, on average, provided care and supervision to a person with dementia for 8 hours per day [5]. Most FCs of people with dementia in Sweden are older partners (including spouses and cohabiting partners) and women [6], and a recent study reported that about half of spouses of people with dementia provide care for ≥30 hours per week on average [7].

Positive aspects of informal caregiving include feeling satisfaction and pride in contributing to the quality of life of the person with dementia [8]. However, FCs often face challenges managing the person with dementia's complex and challenging symptoms of dementia such as restlessness, agitation, and physiological distress [9]. The increased dependency of the person with dementia as the condition deteriorates poses another challenge for FCs, often leaving them overwhelmed [8]. The added responsibilities of FCs occur gradually and tend to increase over time as the condition progresses. This may lead to constraints in their daily lives, such as having to leave work [10]. The challenge of informal caregiving may lead to caregiver stress, which is multifaceted and related to dimensions such as psychological, physical, and social burden due to factors including limited time resources, required vigilance, and supervision [11]. The challenges of informal caregiving can also lead to caregiver depression. The risk factors of depression among FCs of people with dementia include being a woman, being a spouse, and cohabitating with the person with dementia. Caregiver depression among FCs of people with dementia is also related to functional decline and mortality [12].

Support for FCs to People With Dementia in Sweden

The formal support provided to FCs in Sweden is most commonly indirect, in that the support is aimed toward the care recipient (eg, home care). Direct support to FCs is often provided by social care professionals in the municipalities. The educational backgrounds of social care professionals vary and include nurses, social workers, and nurse assistants. Social care professionals have an essential role in the development and coordination of support for FCs, including providing counseling through individual and group support and through education (eg, coping strategies) [13,14]. This type of support provided to FCs in Sweden [15] is often divided into three types of resources: (1) instrumental (ie, material aids and help with daily tasks), (2) informational (ie, relevant information, education, and provision of knowledge), and (3) emotional (ie, empathy, care, and validation) [15].

The National Board of Health and Welfare, through the Social Services Act (SOSFS; 2001:453), obligates municipalities to provide support to informal caregivers [16]. Despite this, most FCs are reported to be unaware of the available support services and who to turn to for support [6]. A study including FCs, people with dementia, and professionals conducted in eight European countries, including Sweden, explored the barriers and facilitators to accessing and using formal dementia care. The study found that a lack of knowledge about dementia and available services among FCs of people with dementia contributed to them not seeking needed support [17]. FCs of people with dementia reported retrospectively that they had waited too long to make use of professional support [17]. In line with this, research shows that FCs often delay seeking formal support services and neglect their personal needs [18] until experiencing a crisis [17,18], such as a sudden inability of the FCs to provide care to the person with dementia or behavioral symptoms of the person with dementia [17]. External support is also often sought when FCs no longer perceive they can function effectively due to increased responsibilities [10]. The competencies of social care professionals, including specific knowledge of dementia and having a key contact person, are important for access to formal care services [17]. In addition, a recent systematic review [19] aimed at identifying the needs of FCs of people with dementia living at home found that receiving tailor-made information to meet one's own needs as FCs was important [19].

Mobile Health–Based Support

The increasing need to support FCs of people with dementia has brought forth the development of innovative interventions delivered through mobile health (mHealth) technology, such as mobile app [20]. mHealth is defined by the WHO [21] as a medical and public health practice supported by mobile devices, such as mobile phones and other wireless devices. The benefits of mHealth technologies are that they are accessible and therefore reported to be easily incorporated into the daily lives of FCs of people with dementia when time permits [22]. The accessibility of mHealth technologies also increases the potential of reaching people who live long distances from health care services and those who are unable to meet in person [23]. Another benefit of mHealth technologies is their potential to provide cost-effective support that is personalized and needs-based by facilitating the connection between the health care provider and the health care recipient in real time [23].

XSL•FO
RenderX

However, a recent meta-analysis found that there was a limited number of interventions using mobile apps targeting FCs of people with dementia living at home [24]. There are also limited available apps that are directed toward the personal needs of FCs, such as stress management [25]. Furthermore, there are limited studies that assess the effectiveness of interventions through mobile apps in terms of health-related outcomes such as caregiver stress and depression of FCs of people with dementia living at home [24].

The Intervention Tool STAV (STöd till AnhörigVårdare)

Considering the aforementioned gaps and the potential use of mHealth technologies in bridging these needs, we developed a tailor-made mobile app named STAV (STöd till AnhörigVårdare: support to FCs) with Karolinska Institutet eHealth Core Facility and byBrick Elevate [26]. The mobile app STAV consisted of (1) a chat feature that enabled FCs and social care professionals to interact, (2) a digital diary that enabled FCs to note any stressful or upsetting events such as behavioral changes in the person with dementia, (3) a mindfulness feature offering access to 3 guided sessions, each averaging 5 minutes in length, (4) a contact feature for registering relevant contact information, and (5) a web links feature with a curated collection of updated and relevant information regarding caregiving and dementia. The app was available on both iOS (Apple Inc) and Android (Google) digital platforms [26]. STAV was developed based on ideas and insights from user perspectives following individual interviews with FCs of people with dementia living at home and a focus group interview with community-based social care professionals specialized in dementia care [27]. The benefits of receiving support through the app for FCs have previously been reported, including offering a foundation for inner calm and relevant information in one place [28]. To gain a more comprehensive understanding of the potential of professional support delivered through the mobile app STAV, a mixed methods research design was used [29]. This approach enabled us to gain insights into how changes in the quantitative outcomes were related to the type of support provided by social care professionals as recorded in chat conversations, an approach that has rarely been used [30]. This study aimed to examine, among FCs of people with dementia living at home, (1) the potential association between professional support provided through STAV and caregiver stress and depressive symptoms, (2) the types of support provided through chat interaction between FCs and social care professionals, and (3) how provision of support through a mobile app relates to changes in caregiver stress and depressive symptoms.

Methods

Study Design

An explanatory sequential mixed methods design was used in this study. This method is useful when applying qualitative data to explain quantitative results [29]. The pilot study adhered to the Good Reporting of A Mixed Methods Study (GRAMMS) checklist [31].

Setting and Participants

The study was conducted among FCs of people with dementia living at home in municipalities located in three regions of Sweden: Stockholm County, Västerbotten County, and Skåne County. The social care professionals who provided support through the chat feature of STAV were purposively sampled and worked in a municipality in Sweden, providing community-based support to FCs of people with dementia living at home. The social care professionals who participated in the study consisted of registered nurses specialized in dementia care, dementia coordinators, social workers, social aid officers, and nurse assistants specialized in dementia care, and worked in the municipalities or senior day care centers for people with dementia. The FCs of people with dementia were purposively recruited through the networks of social care professionals. Other channels of recruitment were local health care centers, websites, newsletters, and social media channels of the municipalities, dementia-related associations, and FCs' networks. The eligibility criteria for the FCs included being an adult aged older than 18 years, who provided care for a person with dementia living at home for a minimum of 6 months, could read and write Swedish, possessed a smartphone or tablet or both, and had access to the internet at their own expense [26]. STAV was available on both iOS and Android devices. The participants received a brief introduction to STAV and, when needed, technical support by a member of the research team either via the telephone, in person, or both. Forty-six FCs were recruited for the intervention, all of whom completed the intervention, including the preintervention assessments. Of the 46 FCs, 11 (23.9%) FCs were lost to follow-up. Therefore, a total of 35 FCs were included in this study. This sample size is considered sufficient for a pilot study [32,33].

The Intervention

The 8-week intervention in terms of professional support was delivered by community-based social care professionals through the chat feature of STAV. The support was delivered through synchronous and asynchronous chat conversations tailored to the individual needs of the FCs [26]. The other support features of STAV—the contact, the diary, web links, and mindfulness feature—were also available for the FCs to use at any time. Before the study commenced, social care professionals received an information package containing details about the intervention and the intervention tool STAV [26]. The social care professionals also participated in an introductory meeting, either in person or online, with members of the research team, to become familiar with the project. The intervention was implemented within the participants' existing social care context and was tailored and delivered based on the FCs' individual needs. As such, no additional training or retraining on the support content was provided beyond the initial introduction. However, regular contact was maintained between a member of the research team and the social care professionals throughout the intervention period. FCs were also given a brief introduction to the intervention tool STAV. In addition, the backend server where the logged chat was stored was monitored regularly. Members of the research team were available to provide technical support to both the social care professionals and the FCs when needed.

XSL•FO
RenderX

Data Collection

The 8-week intervention had 2 data collection time points for the quantitative element (preintervention and immediately after the intervention). Data were collected through self‐administered questionnaires. Preintervention data were collected from the FCs through the mobile app STAV after obtaining informed consent and after they had downloaded the app. Follow-up data were collected through postal questionnaires.

Data for the qualitative element were collected using logged chat data based on chat conversations between the FCs and the social care professionals during the 8-week intervention. The logged chat data were stored in a back-end server of STAV and hosted by Karolinska Institutet.

The Data

Quantitative Background Variables

The background variables included sociodemographic characteristics such as age (in years), sex (female or male), relationship to the person with dementia (partner [including spouse and cohabitating partner] or others [ie, adult children]), as well as living arrangements of the person with dementia (living with a family member or living alone).

Quantitative Outcome Variables

Caregiver Stress

Caregiver stress, the primary outcome, was measured using the validated Swedish version of the 12-item Zarit Burden Interview (ZBI-12) Short form [34], which captures the subjective feelings of burden derived from the caregiving role. The ZBI includes 12 items that measure the multidimensional aspects of burden, such as interpersonal relationships, FCs' health, and emotional well-being. The tool includes a five-point response scale with the options (0=never; 1=rarely; 2=sometimes; 3=frequently; 4=or nearly always) [35]. A summary score ranging from 0 to 48 was calculated, where higher scores indicate higher burden (0‐10=no to mild burden, 10‐20=mild to moderate burden, >20=high burden) [36].

Depressive Symptom

The Swedish version of the Patient Health Questionnaire (PHQ-9) was used to measure the secondary outcome of depressive symptoms [37]. The PHQ-9 consists of 9 criteria on which the diagnosis of depressive disorders is based in the *DSM-IV* (*Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition*) [38]. The validated tool includes a 4-point response scale ranging between 0=not at all, 1=several days, 2=more than half the days, and 3=nearly every day. A summary score ranging from 0 to 27 was calculated, with higher scores indicating greater symptom severity: 0-4=none, 5‐9=mild, 10‐14=moderate, 15‐19=moderately severe, and 20‐27=severe [37].

Qualitative Logged Chat Data

The chat conversations through the app between the FCs and the social care professionals were extracted from the back end of STAV into a Microsoft Excel sheet. The chat entries were matched using participants' identification numbers for both the FCs and the social care professionals. The data consisted of text and, in some cases, emojis. The length of the text in each chat entry ranged from a single word or emoji to 102 words for the FCs and 103 words for the social care professionals.

Data Analysis

Quantitative Analysis

There were no missing values for the preintervention assessments. For the postassessments, there were missing values for caregiver stress (3/35, 8.6%) and depressive symptoms (2/35, 5.7%). The missing values were replaced using the individual mean imputation method, in which the imputed value was based on the calculated mean of the participants' complete responses to other questions for the specific instrument [39]. Descriptive statistics using frequencies, means, and SDs were calculated to describe the study sample, background characteristics, and outcome variables. A paired sample *t* test was used to compare the means of caregiver stress (ZBI-12) and depressive symptoms (PHQ-9) between pre- and post-intervention.

The Shapiro-Wilk test was not significant for caregiver stress at pre- and post-intervention, suggesting that the pairwise differences were normally distributed. The data were not normally distributed for depressive symptoms at pre- and post-intervention; therefore, a related-samples nonparametric test, the Wilcoxon signed-rank test, was conducted. The data were analyzed using SPSS (version 29; IBM Corp) and Stata (version 18.0; StataCorp).

Linear Regression

Thereafter, generalized linear multivariable regression analyses were conducted after checking for multicollinearity. These analyses examined the associations of the outcome variables, caregiver stress and depressive symptoms, with participation in the intervention (pre- vs post-intervention), controlling for age, sex, and relationship to the person with dementia.

Qualitative Analysis

The data were analyzed using summative content analysis [40]. In the initial step, the chat entries were counted by each participant. Thereafter, the logged chat data were read through repeatedly to familiarize with the data. On reading through the chat contents, patterns related to support were observed. The observed patterns of support were then analyzed against the types of support identified in a 2019 national FC mapping [15], namely instrumental, emotional, and informational support. In the analysis process, some of the chat messages were merged or split to accurately reflect the type of support when analyzing the data. To increase trustworthiness, the data were coded according to the type of support and described based on what each type of support included, following discussion and consensus among coauthors [41]. Trustworthiness was further enhanced by using Microsoft Excel to manage and track the coding process. This included detailed documentation of how chat message entries were counted and how different types of support were identified within the logged chat data. This structured approach contributed to both dependability and confirmability by providing a transparent and organized method

XSL•FO
RenderX

for data handling, thereby reinforcing the credibility of the findings [41].

Mixed Methods Analysis

The quantitative and qualitative data were integrated by merging the chat log data with the quantitative primary outcome, caregiver stress, through the caregiver's identification numbers. First, the FCs were divided into 3 groups based on the change in caregiver stress scores from pre- to post-intervention. Group 1 included those with decreased scores in caregiver stress, group 2 showed no change, and group 3 included those with increased scores. Each group was further examined in terms of change in depressive symptom scores and the content of the type of support as identified in the logged chat data. The patterns that emerged in each group are narratively described. Finally, a joint display (Multimedia Appendix 1) was created to show the expansion meta-inferences drawn from the merged quantitative and qualitative data.

Ethical Considerations

The study was approved by the Swedish Ethical Review Authority (Dnr: 2019‐01632 and Dnr: 2020‐06882) and conducted according to the Declaration of Helsinki [42]. Informed written consent was obtained from the participants before the study. The participants were informed that their participation was voluntary, including consent to retrieve the logged chat data for research analysis. All the data were pseudonymized and kept confidential. No identifiable information was linked to the participants in the presentation of the results. The research data collected through the mobile app were securely stored on a backend server at Karolinska Institutet.

Results

Characteristics of Participants

Thirty-five FCs of people with dementia living at home completed the pre- and post-intervention evaluations. The FCs were aged 42-85 years (mean 69.4, SD 11.9 years) and most (28/35, 80%) were women and partners, including spouses and cohabitating partners (24/35, 68.6%), who lived together with the person with dementia (26/35, 74%). Almost 86% (30/35) of the people with dementia received respite care (data not shown). The baseline characteristics are summarized in Table 1.

Table . Baseline characteristics of the study sample (N=35).

Variable	Participants (N=35)
Age (years), mean (SD; range)	69.4 (11.9; 42-85)
Sex, n (%)	
Female	28 (80)
Male	7 (20)
Relationship to the person with dementia, n (%)	
Partner (spouse and cohabitating partner)	24 (68.6)
Others (ie, adult children)	11 (31.4)
Living arrangement of the person with dementia, n (%)	
Living with a family member	26 (74.3)
Living alone	9 (25.7)

Caregivers' Stress and Depressive Symptoms

The mean score of FCs' caregiver stress was marginally higher postintervention (mean 24.1, SD 9.3) compared to preintervention (mean 23.9, SD 9.2), although this difference was not statistically significant. The mean score of FCs' depressive symptoms declined between preintervention (mean 6.5, SD 5.1) and postintervention (mean 6.2, SD 5.2), but was also not found to be statistically significant.

Association Between Participation in the Intervention and Caregiver Stress

The generalized multivariable regression analyses (Table 2) showed that the intervention did not have a significant association between the participation in the intervention on caregiver stress (dependent variable) after adjusting for age, sex, and relationship to the person with dementia (β=0.171, α=.05; P=.86).

XSL•FO
RenderX

Table . Generalized multivariable association between family caregiver's participation in the intervention and i) caregiver stress and ii) depressive symptoms, analyzed using a generalized linear model with a Gaussian distribution and log link function (number of observations: n=70; number of groups=35).

Characteristics	Caregiver stress		Depressive symptoms	
	β coefficient (95% CI)	*P* value	β coefficient (95% CI)	*P* value
Age (in years)	−0.133 (−0.471 to 0.205)	.44	−0.118 (−0.278 to 0.042)	.15
Sex: female (reference group: male)	−5.72 (−12.50 to 1.05)	.10	−2.44 (−5.364 to 0.75)	.13
Relationship to the person with dementia: partner (reference group: others)	5.06 (−3.56 to 13.68)	.25	5.28 (1.201 to 9.349)	.01
The intervention: postintervention (reference group: preintervention)	0.171 (−1.76 to 2.10)	.86	−0.293 (−2.057 to 1.471)	.75

Association Between Participation in the Intervention and Depressive Symptoms

The generalized multivariable regression analyses (Table 2) showed that the intervention did not have an association with depressive symptoms (dependent variable) after adjusting for age, sex, and relationship to the person with dementia (β=−0.293, α=.05; *P*=.75). However, a significant association was found between the relationship to the person with dementia and depressive symptoms, as partners, including spouses and cohabitating partnerFCs (β=5.28, α=.05; *P*=.01), were more likely to report higher levels of depressive symptoms than others (ie, adult children).

Type of Support in the Chat Log Data

The most common types of support that emerged between the FCs and the social care professionals in the logged chat data in terms of chat entries were informational (278/587, 47.4%) and emotional (192/587, 32.7%), followed by instrumental support (26/587, 4.4%). The other types of chat entries (91/587, 15.5%) mostly included logistics and information concerning the intervention (ie, confirming received messages and conveying start and end dates for the intervention).

Informational Support

Informational support was the most common type of support discussed by the FCs and the social care professionals in the chat. The social care professionals used the chat feature to guide the FCs to available offline and online information (ie, information about legal advice and dementia) and support services (ie, group support). Social care professionals also used the chat feature to direct the FCs to the web link feature of STAV and provide information regarding upcoming events (ie, lectures about yoga and mindfulness) and instrumental support (ie, home care and memory aid for the person with dementia). These messages went beyond daily greetings and check-ins, often consisting of probing questions to the FCs, such as asking them to name the diagnoses of the person with dementia or reflect on the symptoms and causes of their recent behaviors. The FCs often used the chat to update the social care professionals about the current state of the person with dementia that they cared for, which included physical and behavioral changes of the person with dementia, such as exhibiting signs of restlessness, wandering, withdrawal, apathy, and memory loss. The social care professionals and the FCs also used the chat feature to confirm new symptoms of the person with dementia with one another or to discuss conflicting accounts about the daily activities of the person with dementia at the daycare center who was experiencing memory loss.

The FCs often asked the social care professionals for coping strategies, such as keeping the person with dementia activated while waiting to return to respite care and setting boundaries with the person with dementia when they exhibited challenging behaviors (ie, suspicion, restlessness, repeating themselves, and anger). The social care professionals responded by educating the FCs on common symptoms of dementia and providing advice on how best to manage them, which included advising the FCs to change the topic of discussion or divert the conversation as a coping strategy. However, some complex conversations were challenging for the social care professionals and the FCs to discuss in the form of chat. In those cases, they reverted to other communication channels, such as the telephone or in-person meetings. Finally, as the intervention period was ending, most social care professionals made sure to provide the FCs with their contact information and navigated them to other support services, such as individual and group support for FCs provided by the municipality.

Emotional Support

Emotional support was the second most common type of support that emerged in the chat entries between the FCs and the social care professionals. The social care professionals provided support ranging from regular check-ins with the FCs, conveying their availability, or sending brief greetings, to more in-depth queries regarding the well-being of the FCs and the person with dementia they cared for. The chat feature provided the FCs an outlet to vent and share their negative emotions, such as feeling overwhelmed, frustration, fatigue, and guilt related to caregiving tasks (eg, medication responsibilities and coordination of health care services). The chat feature enabled the social care professionals to convey empathy and care to the FCs by sending messages that validated and reassured the caregivers regarding their negative emotions related to caregiving.

XSL•FO
RenderX

The social care professionals also used the chat to send self-care reminders, encouraging the FCs to prioritize their needs and accept support (ie, home care provided by the municipalities). The self-care reminders also focused on the FCs' physical well-being, reminding them to breathe, rest, stay hydrated, and engage in stress-reducing activities. These prompts encouraged the FCs to reflect, and some caregivers responded that they met their self-care needs by engaging in activities such as spending time with loved ones (ie, children and grandchildren), baking, or exercising. Some social care professionals also encouraged the FCs to use the mindfulness and diary features of STAV for destressing and offloading purposes.

Instrumental Support

Instrumental support was the least common type of support found in the chat entries. Instrumental support mostly related to coordinating care, such as arranging walking aids to facilitate the person with dementia during transportation to and from medical appointments. The chat feature of STAV also enabled the FCs and the social care professionals to coordinate personal care (ie, foot care) for the person with dementia while attending respite care. The FCs' chat conversations regarding instrumental support also included practical support, such as needing help to search for an item that the person with dementia had lost or misplaced at the senior daycare center. The chat feature enabled the social care professionals to accommodate and respond to the changing needs of the FC and the person with dementia. This could include rescheduling transportation or respite care of the person with dementia due to conflicting leisure activities (ie, vacation plans for the FCs and the person with dementia) and other respite activities of the person with dementia.

Mixed Methods Analysis Results

Group 1: FCs With Decreased Caregiver Stress After Intervention (17/35, 48.6%)

FCs with decreased caregiver stress scores after the intervention generally had a higher number of chat message entries (mean 6.5) directed toward social care professionals, indicating regular engagement and active help-seeking behaviors. These FCs also received a relatively higher number of chat message responses (mean 10) from social care professionals, indicating timely and reciprocal engagement between the users to meet the FCs' needs. Most FCs in this group with decreased caregiver stress also showed improvements (n=8) or no change (n=3) in depressive symptoms.

Through the chat message entries, the FCs vented and shared their negative experiences and emotions related to caregiving with the social care professional, such as feeling frustrated, guilty, and overwhelmed with caregiving tasks (eg, having responsibilities regarding medication administration, coordination of health care services, and respite care).

> *That is just how I feel; guilt and duty, and a mother (the person with dementia) that is demanding in the relationship…* [FC 1057]
>
> *I feel like a hotel hostess with medication responsibilities.* [FC 1040]

These FCs often sought and received tailor-made emotional, instrumental, and informational support, such as coping advice and information about relevant services to address caregiving challenges. Responses from social care professionals included giving relevant information by entering direct chat messages or providing relevant web links.

> *From a psychological perspective, the disease progression can be divided into different phases (provides web link).* [Social care professional 1010]

The social care professionals provided emotional support by validating the FCs' emotions, offering empathy and reassurance, and encouraging the caregivers to adopt self-care behaviors and activities. One social care professional also directed the FCs to other support features of STAV (ie, the diary feature) for offloading purposes.

> *…Remember to breathe and take care of yourself as well!* [Social care professional 1001]

The chat message entries also focused on information and provision of instrumental support, including coordinating care for the person with dementia regarding aids (eg, memory aids for the person with dementia), personal care, respite care, and visits to health care services.

> *Hi…I have arranged foot care for (the person with dementia) on Tuesday. Regards.* [Social care professional 1011]

Group 2: FCs With No Changes in Caregiver Stress After Intervention (5/35, 14.3%)

For the FCs with no change in caregiver stress after the intervention, all except one provided (n=0‑2 messages) and received a low frequency of chat message entries (n=1‑5 messages). None of the chat messages addressed instrumental support, and only a few of the messages by the FCs conveyed the need for information about, for example, caregiver support and coping advice, as well as emotional support (ie, frustration related to home care services for the person with dementia).

> *Hello, in contact with home health care, it is sometimes difficult as an FC to be understood. I feel frustrated that my concerns are not being heard or responded to…* [FC 1042]

The messages from the social care professionals in this group generally consisted of conveying their availability or reminders for the FCs to stay in touch; some provided fixed time slots, while others adjusted their availability to meet the needs of the FCs. The chat entries by the social care professionals also consisted of generic holiday greetings or acknowledgments (eg, "That's great!") regarding updates concerning the person with dementia, which included daily activities and respite care. Thus, these messages were less individualized than the messages sent for the FCs in group 1 and often lacked deeper engagement in terms of reassurance and validation, which were conveyed less frequently.

> *Welcome! Now you can reach out to me; I'm available here.* [Social care professional 1009]

XSL•FO
RenderX

Although this group did not have any change in caregiver stress, all but one of the FCs showed improvements (n=3) or no change (n=1) in depressive symptoms.

Group 3: FCs With Increased Caregiver Stress After Intervention (13/35, 37.1%)

The FCs with an increase in caregiver stress postintervention generally had frequent engagements (mean 6) in the chat and expressed the need for emotional and informational support to a greater extent than the FCs in group 2 (no change in caregiver stress), which was conveyed by asking the social care professionals for coping advice on how to activate the person with dementia and on managing challenging behaviors of the person with dementia, such as, frustration and memory loss.

> *Perhaps (I) can ask if there is anything I can activate (the person with dementia) with now when we do not have (access) the senior daycare centre.* [FC 1051]

For this group, the social care professionals generally provided a higher number of chat message responses (mean 9). The social care professionals also conveyed their availability in the chat and supported the FCs by validating and checking in with them. The provided support was generally less generic than the support identified for group 2. For instance, one social care professional who worked in a senior daycare center updated an FC about the current emotional state of the person with dementia they cared for. However, the emotional support focused less on reassuring the FCs about their expressed emotions related to caregiving (ie, frustration) as identified in the support provided for group 1.

> *Hi [FC], how nice that you respond [in the chat], yes of course it's nice to get out without having to have a bad conscience and that he (the person with dementia) is not alone.* [Social care professional 1008]

Furthermore, the social care professionals also guided the FCs to additional support services through the chat by sending information about upcoming lectures and relevant web links to educate the FCs about dementia. The chat was also used to inform the FCs about available support services and to direct them to social aid officers in the municipalities for additional support (eg, respite care for people with dementia) to relieve the FCs in their role as caregivers.

> *Hi (FC) …call the social aid officer for information about potential services that can support your partner (the person with dementia) when you need to get away from the house to run your errands, (your) interests.* [Social care professional 1007]

A few of the FCs in this group also needed direct instrumental support, including assistance to book or reschedule transport for the person with dementia to and from activities (ie, health care visits) or help with searching for misplaced items belonging to the person with dementia at the senior daycare center.

> *Hi, mum (the person with dementia) has misplaced her house keys. Could you look for them at [the senior day care centre]?* [FC 1055]

While the FCs in this group showed an increase in caregiver stress postintervention, most of the FCs showed improvements (n=5) or no change (n=3) in depressive symptoms. However, no differences were found in the chat entries in terms of support needs expressed by the FCs and provided by the social care professionals between the FCs who showed improvements or no changes and those who showed a deterioration in depressive symptoms.

Discussion

Principal Findings

Participation in the intervention did not have a significant association with caregiver stress and depressive symptoms among the FCs of people with dementia in this study. However, the effectiveness of internet-based or mobile app interventions for FCs of older adults has previously been shown for psychological outcomes, including reduction in depression and caregiver stress [43]. The results showed that informational support, followed by emotional support, was the most mentioned support type. A review [44] reports that some RCTs based on multicomponent interventions, including the provision of informational support, have shown positive impacts on outcomes such as perceived stress and depression. However, to be considered useful, the information provided must be tailor-made to the needs of the FCs [44]. This impact was potentially indicated among the subgroup of FCs with decreased stress who received support in the chat by the social care professionals, tailored to their needs (ie, reassurance, validation, and information) to a greater extent than those FCs with no change or increase in caregiver stress postintervention. However, it is important to note that the impact of informational support is difficult to assess since it is often integrated as a component of a broader intervention [44], as for the mobile app STAV, where web links to available resources are included. Although these results are not generalizable, the mixed methods analysis gave important insights into how the changes in the primary outcome, that is, caregivers' stress, were related to specific types of support provided by the social care professionals through chat conversations with FCs.

The heterogeneity of FCs of people with dementia also needs consideration when designing online-based interventions. One of the main implementation challenges includes the condition of the care recipient (ie, the variation in the severity, type, and stage of dementia) [45], which was not considered in this study. However, the FCs with no change in stress between pre- and post-intervention generally had low engagement in the chat, which might be indicative of a low support need related to caring for a family member with early-stage dementia, as reported in a previous qualitative evaluation of the intervention [46]. Another important factor that needs consideration is the relationship between the FCs and the person with dementia. High caregiver stress is more prevalent among FCs who cohabitate with the person with dementia than among noncohabitating FCs [47]. In this study, a significant association was found, as partners, including spouses and cohabitating partners, were more likely to report higher levels of depressive symptoms than others (ie, adult children). These results are in line with a recent Swedish study where spouses reported significantly higher mean scores in negative outcomes of

XSL•FO
RenderX

informal caregiving, including "physical and psychological stress" and "trouble finding time to spend with friends," than nonspouse FCs [7].

The interaction between the FCs and the social care professionals through the chat feature of STAV could also have impacted the outcome of the intervention on caregiver stress and depressive symptoms. A systematic review [44] reports that interacting directly with professionals (ie, nurses, social workers, or occupational therapists) through internet-based interventions provides FCs of people with dementia living at home with a positive experience. Furthermore, receiving tailor-made emotional and practical advice can decrease feelings of isolation for FCs [44]. Research [45] further suggests that higher engagement and lower intervention dropout rates occur when human contact (eg, health care professionals, research teams, and coaches) is incorporated in the intervention, as opposed to an intervention with no human contact, depending on external sources of interaction (ie, providing a social media page or a telephone number) [45]. Nurses described in a Norwegian study the importance of offering FCs of people with dementia the opportunity to vent their frustrations remotely through telecare using a web camera and a web forum [48]. Similarly, the intent of the chat feature of STAV in this study was to provide support remotely. The results potentially indicate that receiving individualized emotional support in terms of validation and self-care prompts, as opposed to generic validation, had a greater impact on caregiver stress. This was identified for the group of FCs with decreased caregiver stress postintervention, who also mostly showed improvement in depressive symptoms.

There were varied levels of engagement between the FCs and the social care professional through the chat. The FCs with decreased and increased stress had relatively frequent interaction, while FCs with no change generally had low engagement. A previous qualitative evaluation of the intervention [46] reported that some of the FCs and the social care professionals were unfamiliar with each other before partaking in the intervention. Some of the FCs also received supplementary forms of caregiver support during the intervention period (ie, face-to-face support groups). This variance between the interplay of interacting through the chat and meeting in person needs consideration, since text-based platforms, including the mobile app STAV, lack nonverbal cues, which can be expressed through in-person conversations (ie, touch and facial expressions) [49]. This could have contributed to the engagement levels of the FCs and the social care professionals.

Recent research has shown that digital literacy skills and education needs of FCs of people with dementia are significantly associated with the use of mHealth apps [50]. The fact that most of the FCs in this study were older partners, including spouses and cohabitating partners, of people with dementia could have contributed to some challenges with digital literacy. The motivation, privacy concerns, training, and education of the FC of people with dementia further determine eHealth technology implementation [51]. Therefore, it is suggested [51] to support FCs in the use of technology before implementation. It is also important to provide the FCs with adequate time for practice and provide in-person assistance in case of technical issues [51]. Most of the FCs received a brief introduction to the mobile app STAV used in this intervention. Members of the research team also assisted with technical issues when needed. However, this was mostly conducted via telephone conversations due to the COVID-19 pandemic and at the wish of the FCs. Thus, conducting in-person training sessions with the FCs could have resulted in increased uptake of the intervention for the FCs who had challenges with digital literacy. The duration of the intervention, consisting of 8 weeks, could also have impacted the intervention outcomes because a lack of digital literacy can increase the need for additional effort and time to adapt to the new technology. Subsequently, some of the FCs in this study might have needed an onboarding process where enough time was provided to familiarize themselves with the new technology at their own pace [52]. Finally, other factors that could have affected the associations of the intervention were not considered, such as the education and financial level of the FCs and whether the caregivers lived with or had support from other family members.

Limitations and Future Research

The participants of this study consisted mostly of women. Even though the sex distribution is representative of FCs of people with dementia in Sweden [6], a larger proportion of male FCs is needed to explain sex differences in the association between the intervention and the outcome measures. The 8-week duration of the intervention might be too short to bring about significant changes in caregiver stress and depressive symptoms. Digital interventions such as STAV have the potential to reduce selection bias by reaching individuals who are typically underrepresented, such as those living in remote settings [53]. It is possible that the study primarily attracted FCs who were motivated to receive the intervention and were digitally literate. As a result, FCs who needed support but found it challenging to use a mobile app as a mode of communication could have been excluded, potentially leading to a selection bias among FCs of people with dementia. Therefore, the use of the mobile app should serve as a complement, rather than a replacement for traditional means of support. Future studies should also consider the heterogeneity of FCs to the person with dementia, such as the stage of dementia of the care recipient. Future studies could integrate outcome measures with logged analytics data to further examine associations between the digital intervention and the outcome measures.

Conclusions

Patterns observed in the mixed methods analysis at the subgroup level suggest that tailored support provided through asynchronous and synchronous chat delivered through a mobile app may be associated with a potential decline in caregiver stress among FCs of people with dementia living at home. However, a full-scale study with a larger sample size is necessary to confirm these findings and generalize the results of this pilot study to a broader population. These findings highlight the importance of providing frequent and tailor-made informational, emotional, and instrumental support to meet the individual needs of FCs of people with dementia. The findings from this study can help community-based social care providers

XSL•FO
RenderX

plan and organize the content of digital support provided for FCs of people with dementia living at home.

Acknowledgments

We would like to thank the family caregivers and social care professionals who participated in this study. We would also like to thank the Kamprad Foundation, Karolinska Institutet (Engagement Grant and Verify for Impact Grant), Strategic Research Area Health Care Science (SFO-V), and Division of Nursing, Karolinska Institutet, for their financial contribution.

Authors' Contributions

ASK conceptualized the study, was involved in project administration, data curation, participant recruitment, formal analyses, and wrote the original draft. JL conceptualized the study and was involved in the formal analyses. MNI was involved in the formal quantitative data analysis. SV was involved in participant recruitment and the formal qualitative data analysis. ÅD was involved in participant recruitment and the formal qualitative data analysis. HK researched literature, conceptualized the study, was involved in protocol development, and conducted the formal qualitative data analysis. ZNK researched literature, conceptualized the study, was involved in protocol development, was responsible for supervision, project administration, provision of resources, funding acquisition, obtaining ethical approval, data curation, participant recruitment, and was involved in formal analyses.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Joint display: integration of quantitative and qualitative results.
[PDF File, 82 KB - formative_v9i1e75113_app1.pdf]

Checklist 1
GRAMMS checklist.
[DOCX File, 16 KB - formative_v9i1e75113_app2.docx]

References

1. Alzheimer's disease facts and figures. Alzheimer's Association. 2025. URL: https://www.alz.org/alzheimers-dementia/facts-figures [accessed 2025-09-06]
2. Dementia. World Health Organization. 2023. URL: https://www.who.int/news-room/fact-sheets/detail/dementia [accessed 2025-09-06]
3. From plan to impact v. Alzheimer's Disease International. 2022 URL: https://www.alzint.org/u/From-Plan-to-Impact-V.pdf
4. Global estimates of informal care. Alzheimer's Disease International. 2018 URL: https://www.alzint.org/resource/global-estimates-of-informal-care/ [accessed 2025-09-11]
5. Global status report on the public health response to dementia. World Health Organization. 2021. URL: https://www.who.int/publications/i/item/9789240033245 [accessed 2025-09-06]
6. Anhöriga som ger omsorg till närstående fördjupad studie av omfattning och konsekvenser. Caregivers who provide care to relatives: In-depth study of scope and consequences. : Socialstyrelsen (The National Board of Health and Welfare); 2014 URL: https://www.socialstyrelsen.se/globalassets/sharepoint-dokument/artikelkatalog/ovrigt/2014-10-10.pdf [accessed 2025-09-06]
7. Johansson MF, McKee KJ, Dahlberg L, et al. A comparison of spouse and non-spouse carers of people with dementia: a descriptive analysis of Swedish national survey data. BMC Geriatr 2021 Jun 2;21(1):338. [doi: 10.1186/s12877-021-02264-0] [Medline: 34078292]
8. Daley S, Murray J, Farina N, et al. Understanding the quality of life of family carers of people with dementia: development of a new conceptual framework. Int J Geriatr Psychiatry 2019 Jan;34(1):79-86. [doi: 10.1002/gps.4990] [Medline: 30251443]
9. Carrarini C, Russo M, Dono F, et al. Agitation and dementia: prevention and treatment strategies in acute and chronic conditions. Front Neurol 2021;12:644317. [doi: 10.3389/fneur.2021.644317] [Medline: 33935943]
10. Hazzan AA, Dauenhauer J, Follansbee P, Hazzan JO, Allen K, Omobepade I. Family caregiver quality of life and the care provided to older people living with dementia: qualitative analyses of caregiver interviews. BMC Geriatr 2022 Jan 31;22(1):86. [doi: 10.1186/s12877-022-02787-0] [Medline: 35101000]
11. Krutter S, Schaffler-Schaden D, Essl-Maurer R, et al. Comparing perspectives of family caregivers and healthcare professionals regarding caregiver burden in dementia care: results of a mixed methods study in a rural setting. Age Ageing 2020 Feb 27;49(2):199-207. [doi: 10.1093/ageing/afz165] [Medline: 31875879]

XSL•FO
RenderX

12. Huang SS. Depression among caregivers of patients with dementia: associative factors and management approaches. World J Psychiatry 2022 Jan 19;12(1):59-76. [doi: 10.5498/wjp.v12.i1.59] [Medline: 35111579]
13. Alftberg Å. Anhörigkonsulenters uppdrag och arbete: rapport från en forskningscirkel om stöd till anhöriga och kunskapsbehov. The mission and work of social care professionals: report from a research circle on support for family caregivers and knowledge needs. : Malmö Universitet (Malmö University); 2020 URL: https://www.diva-portal.org/smash/get/diva2:1503667/FULLTEXT01.pdf [accessed 2025-09-06]
14. Winqvist M. Anhörigkonsulentens arbete och yrkesroll - resultat från en enkätundersökning. the social care professionals' work and professional role - results from a questionnaire survey. : Nationellt komptenscentrum anhöriga (The Swedish Family Care Competence Centre); 2014 URL: https://anhoriga.se/globalassets/media/dokument/publicerat/rapporter/webb_rapport_2014-1_winqvist_m.pdf [accessed 2025-09-06]
15. Takter M. Anhörigperspektiv - en möjlighet till utveckling? Nationell kartläggning av kommunernas stöd till anhöriga. Family caregiver perspectives – an opportunity for development? National mapping of municipalities support for family caregivers. : Anhörigas Riksförbund (Carers Sweden); 2019 URL: https://anhorigasriksforbund.se/site/wp-content/uploads/Anho¨rigas_Riksfo¨rbund_Rapport_Anho¨rigperspektiv_M_Takter_4nov.pdf [accessed 2025-09-06]
16. Socialtjänstlag (2001:453). The Social Services act (2001:453). : Sveriges Riksdag; 2001 URL: https://www.riksdagen.se/sv/dokument-och-lagar/dokument/svensk-forfattningssamling/socialtjanstlag-2001453_sfs-2001-453 [accessed 2025-09-06]
17. Stephan A, Bieber A, Hopper L, et al. Barriers and facilitators to the access to and use of formal dementia care: findings of a focus group study with people with dementia, informal carers and health and social care professionals in eight European countries. BMC Geriatr 2018 Jun 4;18(1):131. [doi: 10.1186/s12877-018-0816-1] [Medline: 29866102]
18. Oliveira D, Zarit SH, Orrell M. Health-promoting self-care in family caregivers of people with dementia: the views of multiple stakeholders. Gerontologist 2019 Sep 17;59(5):e501-e511. [doi: 10.1093/geront/gnz029] [Medline: 30953585]
19. Bressan V, Visintini C, Palese A. What do family caregivers of people with dementia need? A mixed-method systematic review. Health Soc Care Community 2020 Nov;28(6):1942-1960. [doi: 10.1111/hsc.13048] [Medline: 32542963]
20. Christie HL, Bartels SL, Boots LMM, Tange HJ, Verhey FJJ, de Vugt ME. A systematic review on the implementation of eHealth interventions for informal caregivers of people with dementia. Internet Interv 2018 Sep;13:51-59. [doi: 10.1016/j.invent.2018.07.002] [Medline: 30206519]
21. Ryu S. Book Review: mHealth: new horizons for health through mobile technologies: based on the findings of the Second Global Survey on eHealth: World Health Organization; 2012, Vol. 3:231-233. [doi: 10.4258/hir.2012.18.3.231]
22. Wang AH, Newman K, Martin LS, Lapum J. Beyond instrumental support: mobile application use by family caregivers of persons living with dementia. Dementia (London) 2022 Jul;21(5):1488-1510. [doi: 10.1177/14713012211073440] [Medline: 35414298]
23. Brown EL, Ruggiano N, Li J, Clarke PJ, Kay ES, Hristidis V. Smartphone-based health technologies for dementia care: opportunities, challenges, and current practices. J Appl Gerontol 2019 Jan;38(1):73-91. [doi: 10.1177/0733464817723088] [Medline: 28774215]
24. Shin YH, Kim SK, Kim Y, Go Y. Effects of app-based mobile interventions for dementia family caregivers: a systematic review and meta-analysis. Dement Geriatr Cogn Disord 2022;51(3):203-213. [doi: 10.1159/000524780] [Medline: 35609526]
25. Castillo LIR, Tran V, Hadjistavropoulos T. Are mobile apps meeting the needs of caregivers of people living with dementia? An evaluation of existing apps for caregivers. Aging Ment Health 2024 Apr;28(4):577-586. [doi: 10.1080/13607863.2023.2177832] [Medline: 36775643]
26. Kabir ZN, Leung AYM, Grundberg Å, et al. Care of family caregivers of persons with dementia (CaFCa) through a tailor-made mobile app: study protocol of a complex intervention study. BMC Geriatr 2020 Aug 26;20(1):305. [doi: 10.1186/s12877-020-01712-7] [Medline: 32847495]
27. Moberg C, Grundberg Å, Konradsen H, Kallström AP, Leung AY, Kabir ZN. Meeting own needs and supporting ability to care: family caregivers' and health care professionals' perspectives on professional support provided through a potential mobile application. Dementia (London) 2022 May;21(4):1120-1134. [doi: 10.1177/14713012211065313] [Medline: 35300535]
28. Dorell Å, Konradsen H, Kallström AP, Kabir ZN. "A friend during troubled times": experiences of family caregivers to persons with dementia when receiving professional support via a mobile app. PLoS ONE 2022;17(8):e0271972. [doi: 10.1371/journal.pone.0271972] [Medline: 35917295]
29. Creswell JW, Plano Clark VL. Designing and Conducting Mixed Methods Research, 3rd edition: SAGE Publications; 2017.
30. Gunawan J, Marzilli C, Aungsuroch Y. Online "chatting" interviews: an acceptable method for qualitative data collection. Belitung Nurs J 2022;8(4):277-279. [doi: 10.33546/bnj.2252] [Medline: 37546501]
31. O'Cathain A, Murphy E, Nicholl J. The quality of mixed methods studies in health services research. J Health Serv Res Policy 2008 Apr;13(2):92-98. [doi: 10.1258/jhsrp.2007.007074] [Medline: 18416914]
32. Kunselman AR. A brief overview of pilot studies and their sample size justification. Fertil Steril 2024 Jun;121(6):899-901. [doi: 10.1016/j.fertnstert.2024.01.040] [Medline: 38331310]
33. Julious S. Pilot study sample size rules of thumb. In: PASS Sample Size Software: NCSS Statistical Software; 2021:684-681.
34. Zarit Burden Interview (ZBI). eProvide by Mapi Research Trust. URL: https://eprovide.mapi-trust.org/instruments/zarit-burden-interview [accessed 2025-09-06]

XSL•FO
RenderX

35. Bédard M, Molloy DW, Squire L, Dubois S, Lever JA, O'Donnell M. The Zarit Burden Interview: a new short version and screening version. Gerontologist 2001 Oct;41(5):652-657. [doi: 10.1093/geront/41.5.652] [Medline: 11574710]
36. Zarit burden interview assessing caregiver burden. A courtesy of L'Orech Yomim/Center for Healthy Living. 2011. URL: https://wai.wisc.edu/wp-content/uploads/sites/1129/2021/11/Zarit-Caregiver-Burden-Assessment-Instruments.pdf [accessed 2025-09-06]
37. Depressionsenkät (PHQ-9) (Patient Health Questionnaire). : Viss.nu; 2005 URL: https://viss.nu/download/18.668ca5111744c9a9eed82aaf/1610623941602/PHQ-9_svensk.pdf [accessed 2025-09-06]
38. Kroenke K, Spitzer RL, Williams JBW. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613. [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
39. Shrive FM, Stuart H, Quan H, Ghali WA. Dealing with missing data in a multi-question depression scale: a comparison of imputation methods. BMC Med Res Methodol 2006 Dec 13;6:1-10. [doi: 10.1186/1471-2288-6-57] [Medline: 17166270]
40. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005 Nov;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
41. Lincoln YS, Guba EG. Naturalistic Inquiry: SAGE Publications Inc; 1985.
42. Cook R, Dickens B, Fathalla M. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human participants. In: Reproductive Health and Human Rights: Integrating Medicine, Ethics, and Law 2003:428-432. [doi: 10.1093/acprof:oso/9780199241323.003.0025]
43. Guay C, Auger C, Demers L, et al. Components and outcomes of internet-based interventions for caregivers of older adults: systematic review. J Med Internet Res 2017 Sep 19;19(9):e313. [doi: 10.2196/jmir.7896] [Medline: 28928109]
44. Hopwood J, Walker N, McDonagh L, et al. Internet-based interventions aimed at supporting family caregivers of people with dementia: systematic review. J Med Internet Res 2018 Jun 12;20(6):e216. [doi: 10.2196/jmir.9548] [Medline: 29895512]
45. Atefi GL, Koh WQ, Kohl G, et al. Adherence to online interventions for family caregivers of people with dementia: a meta-analysis and systematic review. Am J Geriatr Psychiatry 2024 Oct;32(10):1271-1291. [doi: 10.1016/j.jagp.2024.04.008] [Medline: 38735829]
46. Kagwa AS, Dorell Å, Konradsen H, Vikström S, Kabir ZN. Providing and receiving support through a tailor-made mobile app: a qualitative study on experience of professionals and family caregivers to persons with dementia. BMC Geriatr 2024 Jun 25;24(1):554. [doi: 10.1186/s12877-024-05151-6] [Medline: 38918728]
47. Ekström H, Auoja NL, Elmståhl S, Sandin Wranker L. High burden among older family caregivers is associated with high prevalence of symptoms: data from the Swedish study "Good Aging in Skåne (GÅS)". J Aging Res 2020;2020:5272130. [doi: 10.1155/2020/5272130] [Medline: 32774920]
48. Solli H, Hvalvik S. Nurses striving to provide caregiver with excellent support and care at a distance: a qualitative study. BMC Health Serv Res 2019 Nov 27;19(1):893. [doi: 10.1186/s12913-019-4740-7] [Medline: 31771566]
49. Lieberman A, Schroeder J. Two social lives: How differences between online and offline interaction influence social outcomes. Curr Opin Psychol 2020 Feb;31:16-21. [doi: 10.1016/j.copsyc.2019.06.022] [Medline: 31386968]
50. Rathnayake S, Moyle W, Jones C, Calleja P. Family carers' needs related to management of functional disability in dementia care and use of mHealth applications in health information seeking: an online survey. Collegian 2020 Jun;27(3):288-297. [doi: 10.1016/j.colegn.2019.09.001]
51. Bastoni S, Wrede C, da Silva MC, et al. Factors influencing implementation of eHealth technologies to support informal dementia care: umbrella review. JMIR Aging 2021 Oct 8;4(4):e30841. [doi: 10.2196/30841] [Medline: 34623314]
52. Guisado-Fernández E, Giunti G, Mackey LM, Blake C, Caulfield BM. Factors influencing the adoption of smart health technologies for people with dementia and their informal caregivers: scoping review and design framework. JMIR Aging 2019 Apr 30;2(1):e12192. [doi: 10.2196/12192] [Medline: 31518262]
53. Drake AL, Rothschild C, Jiang W, Ronen K, Unger JA. Utility of short message service (SMS) for remote data collection for HIV in low- and middle-income countries. Curr HIV/AIDS Rep 2020 Dec;17(6):654-662. [doi: 10.1007/s11904-020-00534-x] [Medline: 33010003]

Abbreviations

***DSM-IV*:** *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition*
FC: family caregiver
GRAMMS: Good Reporting of a Mixed Methods Study
mHealth: mobile health
PHQ-9: Patient Health Questionnaire
STAV: STöd till AnhörigVårdare
WHO: World Health Organization
ZBI-12: 12-item Zarit Burden Interview

XSL•FO
RenderX

Edited by J Sarvestan; submitted 07.04.25; peer-reviewed by K Meyer; revised version received 30.07.25; accepted 31.07.25; published 30.09.25.

Please cite as:
Kagwa AS, Longhini J, Islam MN, Vikström S, Dorell Å, Konradsen H, Kabir ZN
Professional Support Through a Tailor-Made Mobile App to Reduce Stress and Depressive Symptoms Among Family Caregivers of People With Dementia: Mixed Methods Pilot Study
JMIR Form Res 2025;9:e75113
URL: https://formative.jmir.org/2025/1/e75113
doi:10.2196/75113

© Aber Sharon Kagwa, Jessica Longhini, Muhammed Nazmul Islam, Sofia Vikström, Åsa Dorell, Hanne Konradsen, Zarina Nahar Kabir. Originally published in JMIR Formative Research (https://formative.jmir.org), 30.9.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Validation and Accuracy of Wearable Heart Rate Trackers in Children With Heart Disease: Prospective Cohort Study

Hidde J Hardon, MSc; Yara N Van Kerkhof, MSc; Beatrijs Bartelds, MD, PhD; Janneke A E Kammeraad, MD, PhD; Arend W Van Deutekom, MD, PhD

Division of Pediatric Cardiology, Department of Pediatrics, Erasmus MC, Dr Molewaterplein 40, Rotterdam, The Netherlands

Corresponding Author:
Hidde J Hardon, MSc
Division of Pediatric Cardiology, Department of Pediatrics, Erasmus MC, Dr Molewaterplein 40, Rotterdam, The Netherlands

Abstract

Background: Wearables are increasingly used in pediatric cardiology for heart rate (HR) monitoring due to advantages over traditional HR monitoring, such as prolonged monitoring time, increased patient comfort, and ease of use. However, their validation in this population is limited.

Objective: The objective of this paper was to assess HR accuracy and validity from 2 wearables, the Corsano CardioWatch bracelet and the Hexoskin smart shirt, in children attending the pediatric cardiology outpatient clinic, exploring factors that influence accuracy, the Hexoskin shirt's arrhythmia detection efficacy, and patient satisfaction.

Methods: Children with an indication for 24-hour Holter monitoring were equipped with a 24-hour Holter electrocardiogram (ECG; gold standard), together with both wearables. HR accuracy was defined as the percentage of HRs within 10% of Holter values, and agreement was assessed using Bland-Altman analysis. Subgroup analyses were conducted based on BMI, age, and time of wearing, among other factors. The association between accelerometry (expressed in gravitational units, g) and HR accuracy was analyzed to assess the impact of bodily movement on measurement accuracy. A blinded pediatric cardiologist analyzed Hexoskin shirt data for rhythm classification. Patient satisfaction was measured using a 5-point Likert scale questionnaire.

Results: A total of 31 participants (mean age 13.2, SD 3.6 y; n=14, 45% female) and 36 participants (mean age 13.3, SD 3.9 y) were included for the CardioWatch and Hexoskin measurements, respectively. Mean accuracy was 84.8% (SD 8.7%) for the CardioWatch and 87.4% (SD 11%) for the Hexoskin shirt. Hexoskin shirt accuracy was notably higher in the first 12 hours (94.9%, SD 7.4%) compared to the latter 12 (80%, SD 16.7%; $P<.001$). Higher accuracy was observed at lower HRs (low vs high HR: CardioWatch: 90.9%, SD 9.3% vs 79%, SD 10.6%; $P<.001$ and Hexoskin shirt: 90.6%, SD 14% vs 84.5, SD 11.8%; $P<.001$). Both wearables demonstrated good agreement in their HR measurement with Holter readings (CardioWatch bias: –1.4 beats per minute [BPM]; 95% limits of agreement [LoA] –18.8 to 16.0. Hexoskin shirt bias: –1.1 BPM; 95% LoA −19.5 to 17.4). HR measurement accuracy declined during more intense bodily movements. Correct classification of the Hexoskin's shirt rhythm recordings was achieved in 86% (31/36) of cases. Patient satisfaction scores were significantly higher for both the CardioWatch (median 3.8, range 3.5 - 4.3; $P<.001$) and Hexoskin shirt (median 3.7, range 3.0 - 4.0; $P<.001$) compared to the Holter (median 2.6, range 2.1 - 3.2).

Conclusions: The Corsano CardioWatch and Hexoskin shirt demonstrate good accuracy in pediatric HR monitoring and provide higher patient comfort than conventional monitoring. Both wearables show good agreement in relation to the gold standard device. However, measurement accuracy declines with increasing bodily movement and higher heart rates. More research is needed to explore the underlying causes for these inaccuracies and how to counteract them. The Hexoskin shirt also shows potential in arrhythmia detection. While further development is warranted, these wearables show promise in enhancing diagnostics, therapeutic monitoring, and patient safety in pediatric cardiology.

(JMIR Form Res 2025;9:e70835) doi:10.2196/70835

KEYWORDS

wearable technology; heart rate monitoring; ECG; patient satisfaction; electrocardiogram; validation

Introduction

The Holter electrocardiogram (ECG; Spacelabs Healthcare) is widely recognized as the gold standard for ambulatory monitoring of heart rate and rhythm in children [1,2]. However, its bulkiness, the potential for skin irritation caused by plaster electrodes during prolonged use, and the loose wires make it a cumbersome diagnostic tool for extended monitoring, particularly in children. In addition, infrequently appearing symptoms may result in a low diagnostic yield [3,4] and often necessitate extended or repeated monitoring [5]. The multiple

XSL•FO
RenderX

hospital visits that are often needed for cardiac monitoring further affect the child's quality of life [6]. This highlights the need for innovative technologies that can monitor heart rate in children reliably over extended periods of time, are comfortable to wear, and that can preferably be used at home. Ideally, these tools should also allow for remote data access, which could enable timely medical interventions when disease deterioration is detected or reassurance for the patient and family.

Recently, wearable devices, defined as biosensors that can be worn on the body to monitor various health parameters [7], have become increasingly popular and show potential for the follow-up care of patients with heart conditions [8]. Some devices have already been explored in cardiovascular care, particularly for heart rhythm diagnostics [9], such as atrial fibrillation [10], and may offer significant benefits in medical settings, especially for patients with (congenital) heart disease [11,12]. However, most data on wearable validity comes from studies including only adult patients [13,14]. This validity cannot be extrapolated to children, as children have distinct physiological and behavioral properties. For example, differences in wearable fit, children's high-intensity activity patterns, and higher, more variable heart rates—often exceeding 200 beats per minute (BPM)—may impact measurement quality and accuracy [15,16]. This should all be considered when aiming to use wearables in the pediatric setting and highlights the need for more validated research that includes the younger population.

Therefore, this study investigates the validity of 2 recently developed wearable devices in children: a wristband that measures heart rate and an ECG shirt that records heart rhythm and heart rate. Validation is performed in a cohort of children with congenital heart disease (CHD) or (suspected) arrhythmias, during a 24-hour free-living period, in comparison with the gold standard Holter ECG. In addition, this study aims to explore factors that influence measurement accuracy, such as bodily movement, and evaluates patient comfort and satisfaction with the wearables during the 24-hour measuring period.

Methods

Recruitment

Patients between 6 and 18 years of age, attending the outpatient clinic from the pediatric cardiology department of Erasmus MC Sophia Children's Hospital, with an indication for 24-hour or 48-hour Holter ECG monitoring were invited to participate in this study. Before the scheduled Holter monitoring, the participant's health care provider contacted the patient's family to inform them about the research. An information letter was sent out via email. During their hospital visit, the child and family were asked to participate. In case of a positive answer, informed consent forms (see the Ethical Considerations section) were signed, and the patients were outfitted with the wearables. Children younger than 6 years of age were excluded because, at the time of inclusion, wearable sizes for younger children were not available. Another exclusion criterion was the inability to cooperate.

Ethical Considerations

The study was approved by the Institutional Review Board of the Erasmus MC (MEC-2022 - 0576 and MEC-2022 - 0499) and all participants aged 12 years and older and the legal guardians gave informed consent before participation. Children aged 12 years and younger provided assent. The informed consent forms were adapted to the level of understanding of the children. Data were pseudo-anonymized with the key table only accessible for members of the research team. Participants did not receive financial compensation, as the study procedures were integrated into routine clinical care and imposed only minimal additional burden.

Research Devices

The Holter ECG device (Spacelabs Healthcare), hereafter "Holter," is a medical-grade 3-lead ECG device known as the gold standard in ambulatory heart rate and rhythm measuring [2] and was therefore used as a criterion measure to investigate the validity of the heart rate measured by the Corsano CardioWatch 287-2B, hereafter "CardioWatch," and the validity of the heart rate and heart rhythm registered by the Hexoskin Pro shirt (Carré Technologies Inc), hereafter "Hexoskin shirt." The CardioWatch is a CE medically certified multisensor wristband for continuous, in-hospital, and remote patient monitoring to detect abnormal heart rates [17]. The CardioWatch uses reflective photoplethysmogram (PPG) signals to measure heart rate, R-R intervals, and respiration rate.

The Hexoskin shirt is a commercially available smart garment that measures ECG signals using 3 sensors that are woven into the fabric [18]. The sensors, with 2 positioned on either side at the level of the serratus anterior muscles and 1 on the right side of the abdominal muscles, register a single-lead ECG, capturing heart rhythm, heart rate, and heart rate variability.

In addition, both wearables are equipped with a built-in accelerometer that measures bodily movements.

Procedure

Both wearables were charged before the measurement and were set up with a new patient study account. After obtaining informed consent, the participant's demographic and anthropometric characteristics and their clinical indication for Holter monitoring were recorded, and both wearables and Holter were placed on the participant's body. The Holter electrodes were placed by a certified nurse, following the usual protocol. To prevent overlap between the Holter electrodes and the electrodes integrated in the Hexoskin shirt, the Holter electrodes, which could block the measurement of the Hexoskin shirt, were placed slightly more toward the sternum. Next, the participants were outfitted with both the CardioWatch wristband and Hexoskin shirt. First, the CardioWatch was fitted tightly, but comfortably around the nondominant wrist and was subsequently connected via Bluetooth to the researcher's smartphone. Once connected, a new recording session was started in the manufacturer's proprietary smartphone app. The smartphone was given to the participant or legal guardians with the instruction to keep it close, within 10 meters, to the CardioWatch for the 24-hour measuring period to allow for good data synchronization. Finally, the Hexoskin shirt was put on, with

XSL•FO
RenderX

the appropriate shirt size based on the participant's chest circumference. After visually checking for interference between the Holter and the Hexoskin shirt's electrodes, transmission gel was applied to the Hexoskin shirt's electrodes to allow for better signal conduction, as recommended by the manufacturer. The Hexoskin shirt was then connected with the Hexoskin pro device, which initialized recording of data. The data were visually checked in the manufacturer's smartphone app on the researcher's smartphone. No further calibration steps were needed before measuring. Participants were encouraged to live their normal daily routine but refrain from showering and swimming.

In line with the standard protocol for the 24- and 48-hour Holter measurements, the child was asked to fill in a diary with information about activities, the appearance of any symptoms, and bedtimes. All devices were removed by the patient or legal guardians 24 hours after the start of the measurement and returned to the hospital together with the smartphone and diary. For participants undergoing 48-hour Holter monitoring, only the wearable devices were removed after 24 hours, while the Holter measurement continued for the full 48-hour period. To investigate the perceived comfort of both wearables and the Holter device, patients filled in a Likert-scale questionnaire with questions related to comfort, ease of use, and impact on sleep (self-developed, see Multimedia Appendix 1 for the questionnaire [Dutch] or Multimedia Appendix 2 for the overview of questions [English]).

Data Acquisition

After synchronization, data obtained by the CardioWatch (firmware version 4.51) were automatically uploaded to the manufacturer's web application (Corsano Portal; version 1.0.3) via the accompanied smartphone. The manufacturer's proprietary algorithm calculates heart rate in BPM from raw PPG signals. These data in BPM were then downloaded from the online portal for further analysis.

In a similar way, BPM data from the Hexoskin shirt were downloaded from the Hexoskin web dashboard after synchronization of the pro device—which stored the measurement data—with the manufacturer's browser-based app (version 5.0.12.348; OneSync).

To investigate the association between bodily movement and measurement accuracy, raw accelerometry data of both wearables were downloaded. For the CardioWatch, the raw triaxial accelerometry data (acc_x, acc_y, and acc_z) is captured in unit 1/512 g. The intensity of the movement in g's was derived from the raw accelerometry data divided by 512, and the following equation [19]:

$$d_accX[i] = acc_x[i] - acc_x[i-1]$$
$$d_accY[i] = acc_y[i] - acc_y[i-1]$$
$$d_accZ[i] = acc_z[i] - acc_z[i-1]$$
$$Norm_of_d_acc[i] = \sqrt{d_accX[i]^2 + d_accY[i]^2 + d_accZ[i]^2}$$

Here, *d_accX, d_accY, and d_accZ* represent the change in acceleration along the X, Y, and Z axes between two consecutive time points, while *acc_x[i], acc_y[i], and acc_z[i]* represent the raw acceleration values at time point *i* along each respective axis. *Norm_of_d_acc* reflects the magnitude of the change in acceleration across all directions.

The resulting accelerometry intensity in g was used for further analysis. For Hexoskin, the accelerometry data were already calculated in unit g, as recorded by the Hexoskin pro-device.

Holter data were analyzed, following standard procedure, by a Holter specialist and were subsequently downloaded from the Erasmus MC database using Sentinel software (version 11.5.5.14400; Spacelabs Healthcare). Heart rate was derived from the R-R intervals and calculated as a moving average over the previous 16 beats to mitigate the impact of artifacts and noise. To adjust for differences in the devices' time settings, the optimal offset was calculated as the highest cross-correlation between the Holter and wearable device within 60 seconds and adjusted using a Python script (Python Software Foundation). The CardioWatch and Hexoskin shirt data were interpolated to the criterium measurement data and filtered for further analysis. BPM data of all three devices were used for analysis.

Missing Data

Missing heart rate data of the CardioWatch and Hexoskin shirt were considered underdetected, missing heart rate data of the Holter were considered erroneous due to malfunctioning of the criterion measure. To allow for accurate comparison of data, erroneous data were excluded from the analysis.

ECG Rhythm Registration

Holter ECG registrations were analyzed by a Holter specialist, using software to detect clinically relevant parts of the recording. These mostly include normal sinus rhythm (SR), supraventricular tachycardia (SVT), ventricular tachycardia (VT), bradycardia, and single extra beats, such as premature atrial contractions (PAC) or premature ventricular contractions (PVC), among others. To see if the Hexoskin shirt could detect these arrhythmias, its ECG rhythm recordings were presented to a pediatric cardiologist. The rhythm recordings made by the Hexoskin shirt were accessed online through the Hexoskin dashboard. After analysis of the Holter device by the Holter specialist, clinically relevant ECG rhythms were selected and compared to the readings from the Hexoskin shirt. In total, 36 of these signals were then presented to a blinded pediatric cardiologist for interpretation. For each Hexoskin recording, the corresponding Holter rhythm from the same timeframe was included, resulting in a total of 72 rhythms for review—36 from the Hexoskin shirt and 36 from the Holter monitor. The Holter data were presented without previous interpretation from the Holter specialist, ensuring unbiased classification. Hexoskin and Holter recordings were presented in a random order.

Statistical Analysis

Overview

Descriptives are shown as mean (SD) or as median IQR when appropriate. Differences between subgroups were analyzed using independent sample *t* tests or ANOVA tests in the case of more than 2 groups. Within-group effects were analyzed using paired sample *t* tests. Normality of distributions of data was checked both visually and with the Shapiro-Wilk test. In case of any violation of normality, the data were analyzed using nonparametric equivalent tests. The significance level was set at $P<.05$ (2-tailed). All statistical analyses were performed using

XSL•FO
RenderX

SPSS (version 28.0.1.0(142); IBM Corp) and Python (version 3.10.9).

Accuracy

Measurements within 10% of criterion measurement values were defined as accurate, which is in line with the Association for the Advancement of Medical Instrumentations (AAMI) guidelines [20].

Mean Absolute Error and Mean Absolute Percentage Error

Mean absolute error (MAE) was calculated as the MAE between wearable and criterion measure heart rate measurements. Mean absolute percentage error (MAPE) was calculated as the mean absolute error between wearable and criterion measure heart rate data, divided by the criterion measure's value and multiplied by 100. Individual MAPEs were averaged to obtain an overall MAPE for each wearable. The absolute error between the criterion measure and the wearable measure is used to calculate MAPE, as this approach accounts for both overestimation and underestimation of the measurements. The equations to calculate these measures of error are shown in Multimedia Appendix 3. As described by Nelson et al [21], a MAPE score of <10% was defined as accurate.

Bland-Altman Analysis

The Bland-Altman analysis was used to analyze the agreement between wearable devices and Holter measurements. The mean differences are reported with 95% limits of agreements (LoA). Bland-Altman plots were created to visualize the agreement.

Lin Concordance Correlation Coefficient (CCC)

Correlation between both wearables and Holter was assessed separately using Lin concordance correlation coefficient (CCC) and averaged over all participants to form a total CCC for each wearable [22]. Agreement strength was scored as weak (CCC<0.5), moderate (CCC=0.5 - 0.7), and strong (CCC>0.7), as in line with previous research [23].

Patient Satisfaction Scores

Patient satisfaction was measured using a 5-point Likert scale questionnaire, which was developed for the purpose of this study (see Multimedia Appendix 1 [Dutch] or Multimedia Appendix 2 [English]). The questionnaire outcomes were analyzed using nonparametric Wilcoxon tests. Wear time was included as a quantitative usability metric.

Accelerometry Analysis

To investigate the association between bodily movement and the accuracy of HR measurements, the accelerometry data of both the CardioWatch and the Hexoskin shirt were analyzed. As there is no clear consensus on physical activity intensity thresholds (in gravitational units) in children, and both wearables measure different bodily movements (ie, wrist-based for the CardioWatch and torso-based for the Hexoskin), the accelerometry analysis was based on individual thresholds of both devices. Thereby, the HR measurement accuracy was calculated based on the quantiles of the participants' own accelerometry data, where lower quantiles correspond to fewer bodily movements. Mean accuracy scores per quantile were computed across all participants to create an overall accuracy score per quantile for each wearable.

Subgroup Analysis

Based on our pilot study, which showed a relationship between wrist circumference and accuracy of the CardioWatch, wrist circumference was included as a covariable. Next to that, the effect of sex, age, BMI, skin type, time of day, heart rate frequency, and measuring period on the accuracy of heart rate measurements was investigated. Subgroups were created with the following thresholds: age (≤12 and >12 y), BMI (≤19 and >19 kg/m^2), wrist circumference (≤15.5 and >15.5cm, for CardioWatch only), time of day (wake time and sleep time), heart rate (above or below median heart rate), and measuring period (first 12 h and second 12 h of measuring). The subject's median heart rate was chosen as the threshold between lower and higher heart rates instead of the mean heart rate, as the mean heart rate is influenced by outliers. Skin type differentiation (Fitzpatrick I and II, Fitzpatrick III, and Fitzpatrick IV, V, and VI, for CardioWatch only) was included as darker skin is known to influence the optical PPG measurement [24,25]. To see if the patient's diagnosis had an influence on heart rate and rhythm accuracy, participants were grouped according to the indication for their Holter measurement: CHD, known arrhythmia, or symptoms (eg, palpitations or fainting). To accurately delineate the transition between wakefulness and sleep, the 60 minutes around reported sleep times were not included in sleep vs wake accuracy analysis. The diary information was used to differentiate between wake and sleep times.

Results

Baseline Characteristics

A total of 39 participants took part in this study. Due to the malfunctioning of the criterion measure, data from 2 participants were excluded before analysis. For the CardioWatch, data of 6 participants were excluded from analysis due to either early loss of battery power (n=1), an accidental reset of the CardioWatch during the measurement (n=1), or an accidental termination of the smartphone app (n=4), which all resulted in the loss of data. The data of the remaining 31 participants were included for analysis. For the Hexoskin shirt, the data of 1 participant was excluded because the shirt was taken off after 2 hours of measuring. The data of the remaining 36 participants were included for analysis. In total, 38 out of the 39 participants who completed the 24-hour measurement filled in the questionnaire. All were included in the questionnaire analysis. Table 1 shows the participants' characteristics.

XSL•FO
RenderX

Table . Participant characteristics. Wrist circumference and Fitzpatrick outcomes were not used for the Hexoskin shirt analysis.

Participants characteristics	CardioWatch (n=31)	Hexoskin shirt (n=36)
Characteristics		
Sex: female, n (%)	14 (45)	16 (44)
Age (years), mean (SD)	13.2 (3.6)	13.3 (3.9)
Height (m), mean (SD)	1.60 (0.18)	1.59 (16.5)
Weight (kg), mean (SD)	52.2 (16.8)	50.8 (16.0)
BMI (kg/m^2), mean (SD)	19.9 (3.7)	19.5 (3.6)
Wrist circumference (cm), mean (SD)	15.4 (1.7)	__[a]
Indication for Holter, n (%)		
Congenital heart disease	11 (36)	13 (36)
Arrhythmia	14 (45)	16 (44)
Symptoms	6 (19)	7 (19)
Fitzpatrick score, n (%)		
Type I	0 (0)	—
Type II	10 (34)	—
Type III	13 (45)	—
Type IV	4 (14)	—
Type V	2 (7)	—
Type VI	0 (0)	—

[a]Not applicable.

Missing Data

Underdetected data or missing data of the CardioWatch were calculated as the total missing timepoints and occurrences of zeros in the data. Due to a bug in the CardioWatch software, data of the last 9 participants showed a shift in sample rate at night from once per second to one measurement per minute. This resulted in missing time points and therefore a high percentage of missing data (36% - 40%). These participants were included in accuracy measurements but excluded from the calculation of missing data and the accelerometry analysis. Mean percentage of underdetected data for the CardioWatch was 2.5% (SD 2.1%), ranging from 0.02% to 10.6%. Erroneous data, which was defined as missing datapoints in Holter data, was significantly lower with a mean of 1.3% (2.2%), ranging from 0.01% to 10.2%. The percentage of missing data for all participants is shown in Multimedia Appendix 4. The missing data percentage of the Hexoskin shirt could not be determined because a heartbeat of 70, or the last known heartbeat, is shown for missing datapoints.

Accuracy of Heart Rate

For the CardioWatch, the overall accuracy of the 24-hour heart rate measurement ranged between 67.1% and 98.7% with a mean of 84.8% (SD 8.7%). For the Hexoskin shirt, the accuracy of the heart rate measurements was captured between 60.3% and 99.4% with a mean of 87.4% (SD 11%). Both wearables, therefore, demonstrate good accuracy in heart rate measurements (see Table 2), although with high interindividual variability (see Multimedia Appendix 5 for the individual accuracy scores per participant).

XSL•FO
RenderX

Table . Main study parameter outcomes for both the CardioWatch and the Hexoskin.

Parameters	CardioWatch (n=31)	Hexoskin (n=36)
Accuracy (%), mean (SD)	84.8 (8.7)	87.4 (11)
MAE[a] (BPM[b]), mean (SD)	4.9 (1.8)	4.3 (3.2)
MAPE[c] (%), mean (SD)	5.5 (2.2)	4.9 (3.8)
Agreement (BPM), median (IQR)	–1.4 (–18.8 to 16.0)	–1.1 (–19.5 to 17.4)
CCC[d] (95% CI)	0.872 (0.869 - 0.874)	0.845 (0.842 - 0.848)
Comfort score (n=38), median (IQR)	3.8 (3.5 - 4.3)	3.7 (3.0 - 4.0)

[a]MAE: mean absolute error.

[b]BPM: beats per minute.

[c]MAPE: mean absolute percentage error.

[d]CCC: Lin concordance correlation coefficient.

Error

The CardioWatch had an overall MAE and MAPE of 4.9 (SD1.8) BPM and 5.5% (SD 2.2%), respectively. The Hexoskin shirt showed an overall MAE and MAPE of 4.3 (SD 3.2) BPM and 4.9% (SD 3.8%). These results indicate that both wearables had a comparable average absolute error relative to the criterion measure. Similarly, both wearables showed a MAPE of around 5%, meaning that the average difference between measurements from the wearables and the Holter was around 5% of Holter measurement values. This error margin is well below the 10% threshold described by Nelson and colleagues [21] and indicates a small error and accurate measurement performance. An overview of all the individual scores is presented in Multimedia Appendix 6.

Agreement

The CardioWatch had a slight underestimation of the HR measurements compared to the Holter, with a mean difference of −1.4 BPM, and 95% LoA ranging from −18.8 to 16.0 (see Figure 1). For the Hexoskin shirt, the Bland-Altman analysis showed a small underestimation of −1.1 BPM and 95% LoA ranging from −19.5 to 17.4 (see Figure 2).

Figure 1. Bland-Altman plot CardioWatch. bpm: beats per minute; LOA: limits of agreement.

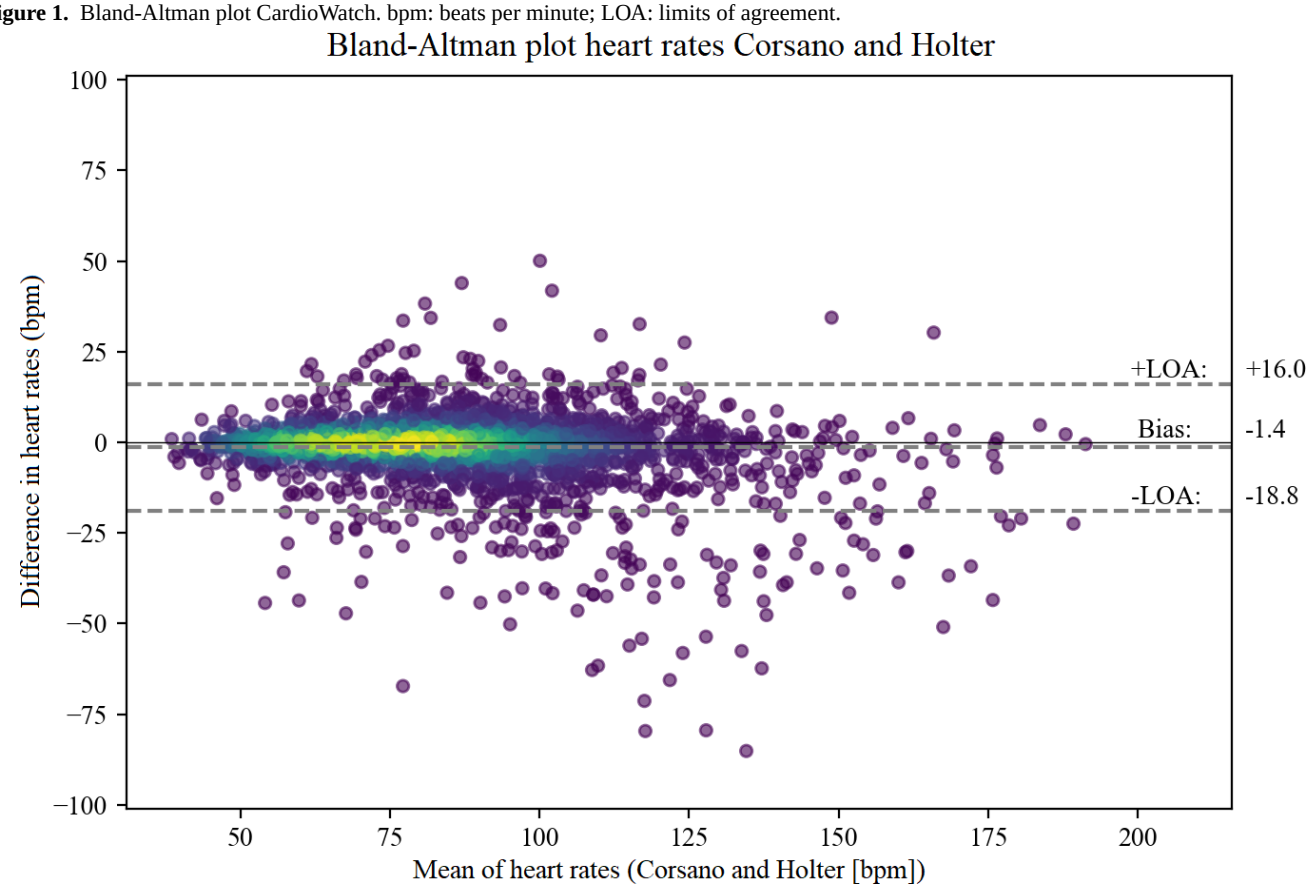

XSL•FO
RenderX

Figure 2. Bland-Altman plot Hexoskin. bpm: beats per minute; LOA: limits of agreement.

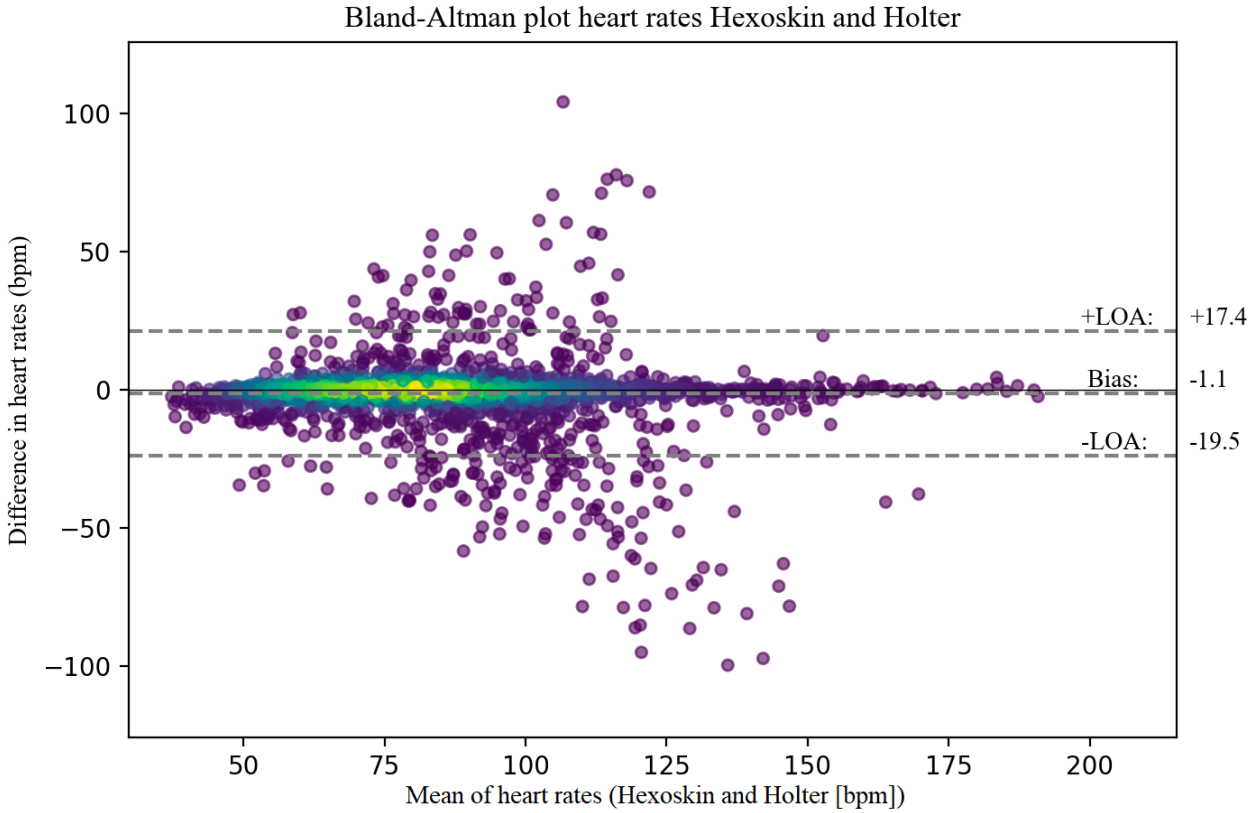

Concordance Correlation Coefficient

The CardioWatch had a CCC of 0.872 (95% CI 0.869 - 0.874) and the Hexoskin shirt showed a CCC of 0.845 (95% 0.842 - 0.848), both indicating strong agreement in relation to the criterion measure. Individual scores are shown in Multimedia Appendix 7.

Accuracy of Heart Rhythm

Figures 3-6 show examples of the ECG rhythms recorded by the Holter monitor and the Hexoskin shirt. No significant rhythm abnormalities, such as SVTs or VTs, were detected during the 24-hour monitoring period for any of the participants. Minor abnormalities, including isolated extra beats, did occur. All 36 Holter recordings were accurately classified by the pediatric cardiologist. The Hexoskin shirt's ECG recordings were correctly classified in 31 out of 36 cases. The remaining 5 recordings could not be classified with certainty due to artifacts and noise in the Hexoskin shirt's ECG signal. The 2 examples of clear registrations are shown in Figures 3 and 4. In addition, 2 examples of rhythms with noise, which impeded accurate classification, are shown in Figures 5 and 6. None of the recordings were incorrectly classified as normal or abnormal. Multimedia Appendix 8 shows an overview of all the arrhythmias that were included in the rhythm recordings that were presented to the pediatric cardiologist.

XSL•FO
RenderX

Figure 3. Example of a clear electrocardiogram (ECG) rhythm. Hexoskin recordings are depicted above and Holter ECG rhythms below.

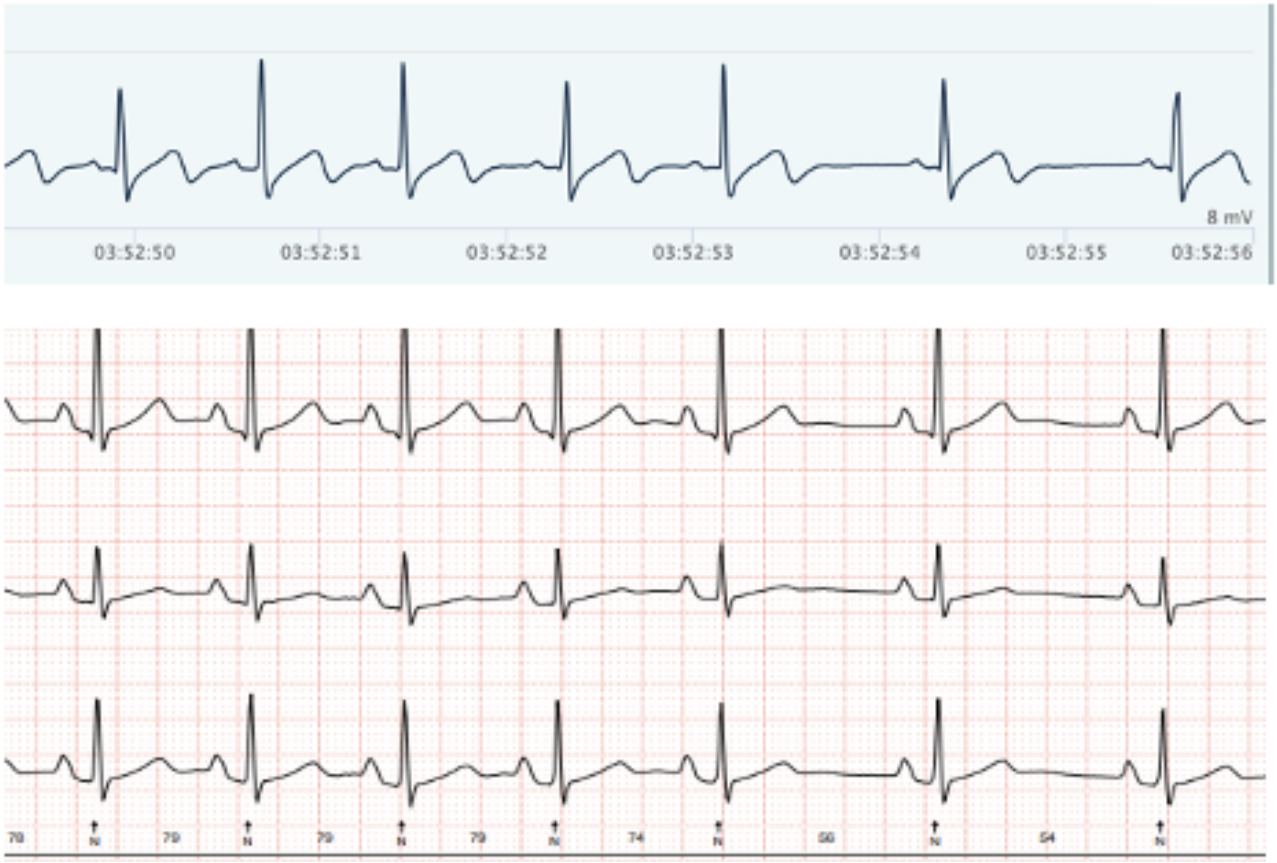

Figure 4. Example of a clear electrocardiogram (ECG) rhythm. Hexoskin recordings are depicted above and Holter ECG rhythms below.

XSL•FO
RenderX

Figure 5. Example of electrocardiogram (ECG) rhythm with noise. Hexoskin recordings are depicted above and Holter ECG rhythms below.

Figure 6. Example of electrocardiogram (ECG) rhythm with noise. Hexoskin recordings are depicted above and Holter ECG rhythms below.

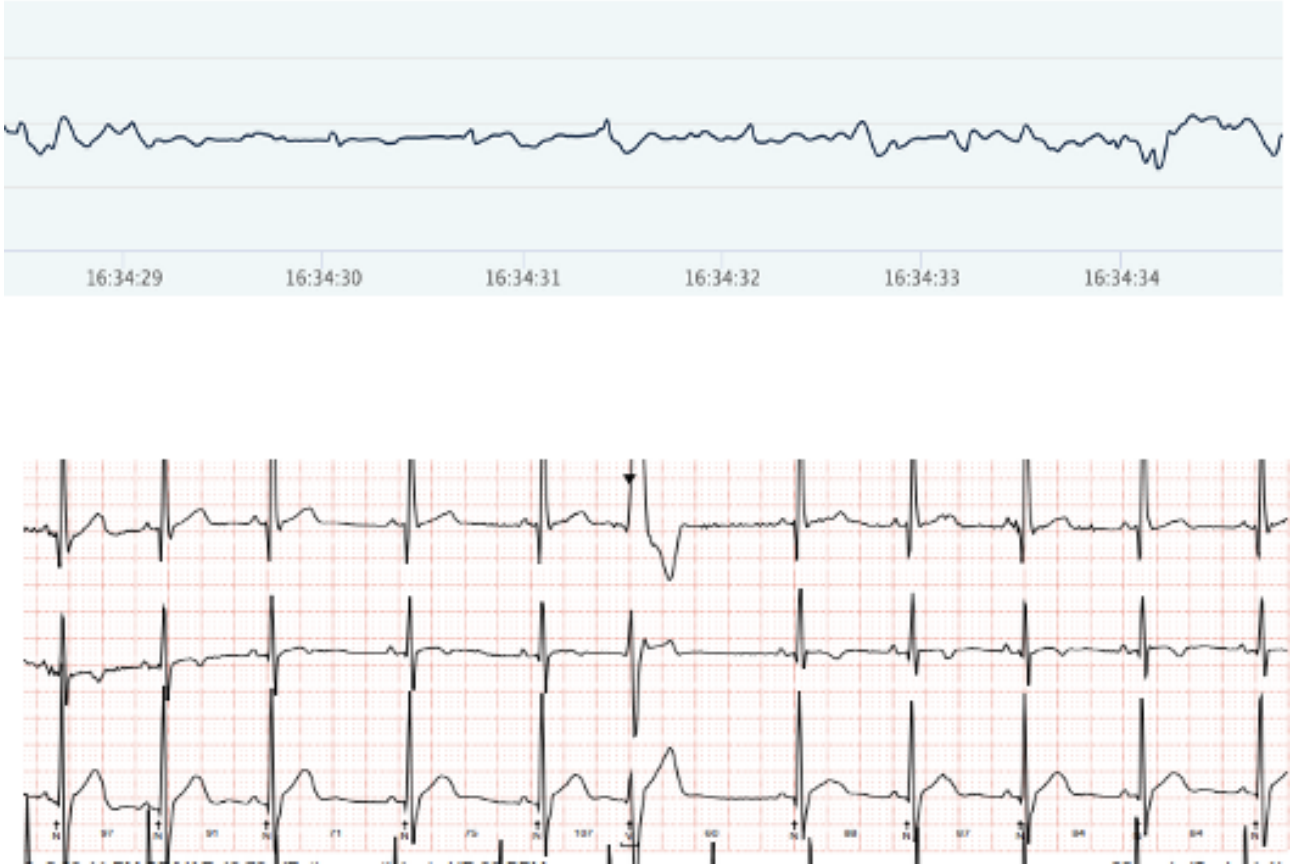

XSL•FO
RenderX

Questionnaire Results

The outcomes of the comfort questionnaire are shown in Table 2. Both wearables scored significantly higher on all questions compared to the Holter, reflecting better patient acceptance. The median scores were 3.8 (IQR 3.5 - 4.3) for the CardioWatch and median 3.7 (IQR 3.0 - 4.0) for the Hexoskin shirt, in contrast to median 2.5 (IQR 1.8 - 3.2) for the Holter. This indicates that the CardioWatch and Hexoskin shirt were perceived as much more comfortable and pleasant than the Holter. Notably, all participants expressed a preference for wearing the CardioWatch and Hexoskin shirt instead of the Holter monitor. A participant removed the Hexoskin shirt prematurely during the study period. See Figure 7 for a visual representation of the comfort scores. An overview of all the question scores is provided in Multimedia Appendix 2.

Figure 7. Boxplots questionnaire comfort scores.

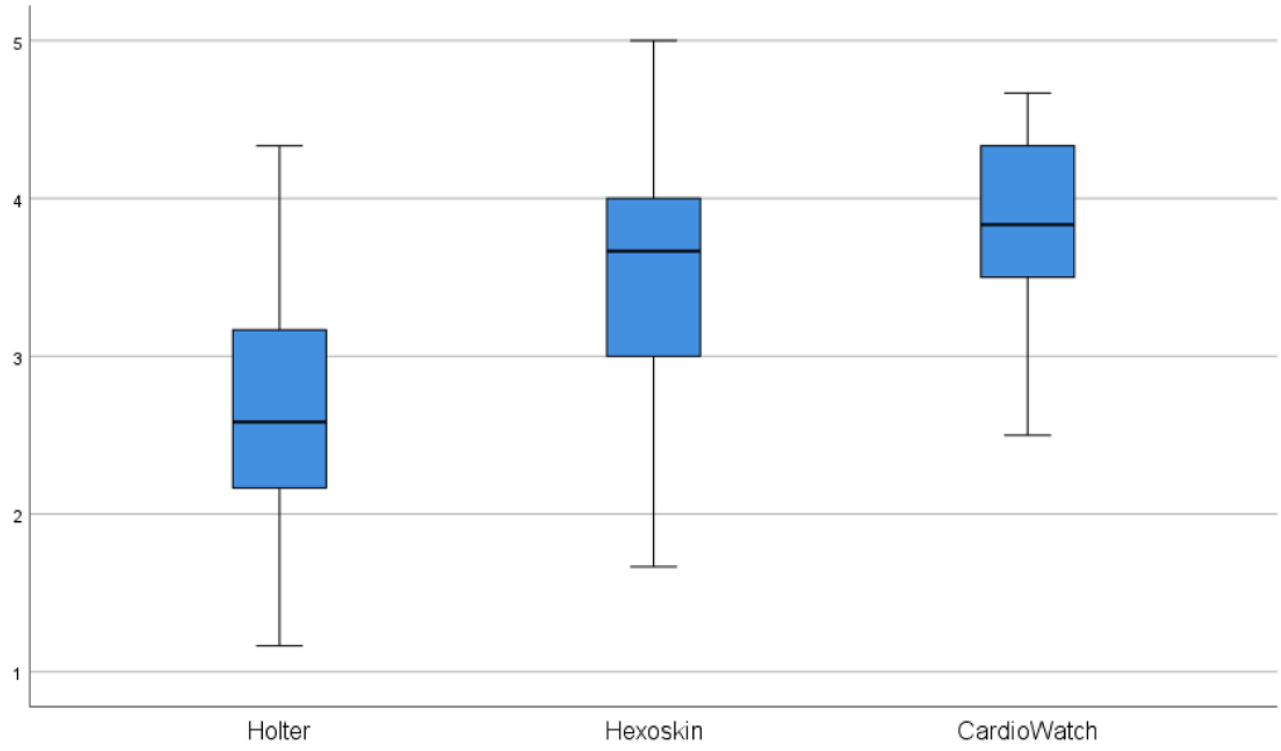

Accelerometry Analysis

Table 3 summarizes the mean accuracy of HR measurements across quantiles of movement intensity derived from participants' accelerometry data, averaged over all participants. For both wearables, the accuracy of HR measurements decreased at higher quantiles, reflecting greater bodily movement intensity.

To further investigate potential age-related differences in activity and their impact on HR accuracy, a linear regression was performed with heart rate accuracy as the outcome, and age group and accelerometry as predictors, including an interaction term age group x accelerometry. There was no statistically significant interaction effect, which means that the relation between acceleration and accuracy did not differ between the two age groups. Figure 8 shows a visual representation of the effect of bodily movement on HR measurement accuracy for both wearables, where the difference in HR measurement between both the CardioWatch (depicted in green) and Hexoskin shirt (depicted in red) and Holter device are plotted in correspondence with the accelerometry data (orange and blue, respectively).

The individual accuracy of HR measurements in relation to the participants' measures of accelerometry is shown in Multimedia Appendix 9.

Table . Accelerometry quantile-based accuracy measurement.

Quantile	CardioWatch (%), mean (SD)	Hexoskin (%), mean (SD)
First quantile	89.1 (8.5)	88.0 (10.7)
Second quantile	82.3 (9.3)	85.4 (11.1)
Third quantile	77.6 (11.4)	80.5 (13.8)
Fourth quantile	74.0 (16.9)	76.7 (17.5)

XSL•FO
RenderX

Figure 8. Differences in heart rate (HR) measurement versus accelerometry for Hexoskin (red and blue graph) and CardioWatch (green and yellow graph).

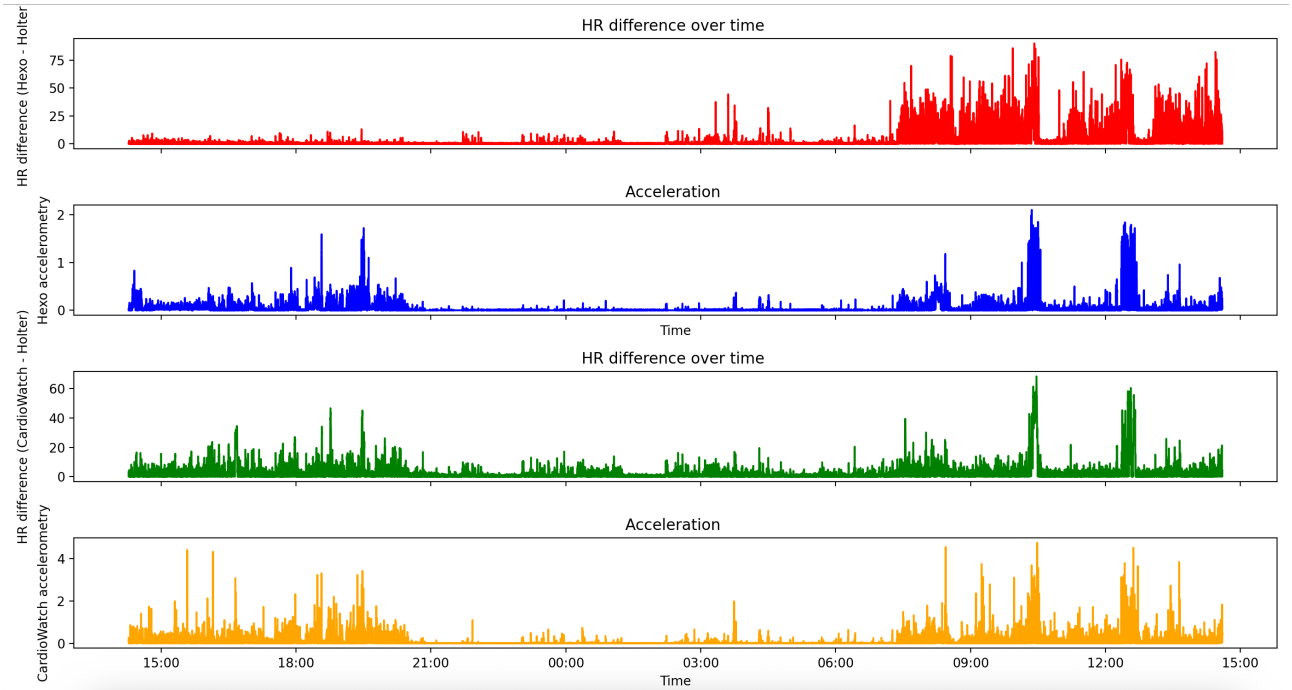

Subgroup Analysis

Results of the subgroup analyses are shown in Table 4 and Multimedia Appendix 10. For the CardioWatch, the accuracy of HR measurements below the median heart rate was significantly higher than that of measurements above the median (90.9, SD 9.3% vs 79, SD 10.6%; $P<.001$). In addition, the accuracy of the CardioWatch was significantly higher during sleep than during wake time (90.1, SD 10.3% vs 82.1, SD 9.8%; $P<.001$).

Statistical analysis of HR measurements of the Hexoskin shirt showed a significantly greater accuracy in older children, in children with higher BMI, during higher HRs and during measurements at night. The accuracy scores per participant for the variables “Heart rate,” “Time of day,” and “Measuring period” are shown in Multimedia Appendices 11 and 12. The error outcomes for subgroup scores are shown in Table 5.

XSL•FO
RenderX

Table . Overview of subgroup analysis outcomes on accuracy for the CardioWatch and Hexoskin.

Accuracy	CardioWatch		Hexoskin	
	Mean (SD)	*P* value	Mean (SD)	*P* value
Sex		.71		.82
Female	85.4 (9.7)		86.8 (11.9)	
Male	84.3 (8.0)		87.9 (10.4)	
Age (years)		.76		.03
≤12	85.4 (7.4)		82.2 (11.5)	
>12	84.4 (9.6)		90.4 (9.7)	
BMI		.06		.002
≤19 kg/m2	81.3 (6.6)		82.0 (11.5)	
>19 kg/m2	87.3 (9.3)		92.3 (8.0)	
Diagnosis group		.88		.97
CHD[a]	83.7 (9.7)		86.8 (11.6)	
Arrhythmia	85.2 (8.4)		87.7 (10.6)	
Symptom	85.8 (7.6)		88.0 (12.2)	
Heart rate		<.001		<.001
≤Median heart rate	90.9 (9.3)		90.6 (14.0)	
>Median heart rate	79.0 (10.6)		84.5 (11.8)	
Time of day		<.001		<.001
Wake time	82.1 (9.8)		86.1 (10.9)	
Sleep time	90.1 (10.3)		90.8 (15.2)	
Wrist circumference		.92		—[b]
≤15.5 cm	84.9 (7.5)		—	
>15.5 cm	84.6 (10.7)		—	
Skin type (Fitzpatrick score)		.92		—
I and II	85.4 (7.8)		—	
III and IV	84.0 (8.3)		—	
V and VI	85.8 (7.6)		—	
Measuring period		.25		<.001
1st 12 hours	83.9 (10.7)		94.9 (7.4)	
2nd 12 hours	85.7 (8.4)		80.0 (16.7)	

[a]CHD: congenital heart disease.

[b]Not applicable.

XSL•FO
RenderX

Table . Error outcomes for the CardioWatch and Hexoskin.

Error	CardioWatch		Hexoskin	
	MAE[a] (BPM[b]), mean (SD)	MAPE[c] (%)	MAE (BPM), mean (SD)	MAPE (%)
Overall	4.9 (1.8)	5.5	4.3 (3.2)	4.9
HR[d]				
<Median HR	2.7 (1.4)	4.1	2.9 (3.2)	4.4
>Median HR	7.1 (2.8)	6.8	5.7 (4.1)	5.3
Time of day				
Wake time	6.0 (2.3)	6.2	5.0 (3.5)	5.1
Sleep time	2.9 (1.8)	4.1	2.7 (3.2)	4.1

[a]MAE: mean absolute error.

[b]BPM: beats per minute.

[c]MAPE: mean absolute percentage error.

[d]HR: heart rate.

Discussion

Principal Findings

This study assessed the accuracy and validity of HR measurements from the CardioWatch in 31 children and the accuracy and validity of both HR and rhythm measurements from the Hexoskin shirt in 36 children attending a pediatric cardiology outpatient clinic. Both the CardioWatch and Hexoskin shirt demonstrated high measurement accuracy and strong agreement with the Holter for HR assessment. The wearables showed minimal bias and reliable performance, suggesting they are valid tools for measuring HR in children with CHD or suspected arrhythmias. Nevertheless, Bland-Altman analyses indicated notable variability in measurement agreement, with accuracy decreasing at higher HRs. Consequently, both devices performed better during nighttime periods characterized by lower HRs compared to daytime periods, when activity levels and heart rates were higher. In addition, the accelerometry analysis demonstrated a clear inverse relationship between physical activity intensity and HR measurement accuracy. Taken together, these findings imply that the tested wearables may be particularly useful for continuous monitoring during rest or sleep rather than during vigorous physical activities, such as sports.

The Hexoskin shirt shows promising results for arrhythmia detection. However, noise and artifacts sometimes make it impossible to accurately classify rhythm recordings. This noise is caused by a combination of poor conduction due to the drying of conductive gel and shifting electrode placement. Future research should focus on these two factors to increase measurement accuracy. It is thereby important to realize that the Hexoskin shirt forms a 1-lead ECG signal, in contrast to the 3-lead ECG signal of the Holter device. The interpretation of the 1 lead signal will therefore always be inferior to that of the Holter device, as the latter usually provides at least one lead that allows for reliable rhythm interpretation.

Regarding the main study parameter, namely HR accuracy and validity, a previous study by Blok et al [26] reported higher accuracy (94.6%) and agreement (mean bias 0.06, LoA −3.89 to 3.77) for the CardioWatch compared to the findings in this study. They investigated the accuracy of the CardioWatch using a 12-lead ECG in older adults under resting conditions [26]. The difference in accuracy between the study by Blok et al [26] and ours can be explained by several factors, including their use of a signal qualifier and the absence of physical movement during their measurements. This lack of movement is known to influence PPG accuracy positively. Activity and motion artifacts can disturb the optical measurement of blood volume changes, leading to missing beats or inaccurate readings [8,25,27-32]. These factors likely explain the higher accuracy during nighttime and at lower heart rates in this study, as illustrated in Multimedia Appendices 11 and 12. The findings in this study are consistent with those of Hermans et al [28], who reported similar discrepancies in PPG accuracy in patients with atrial fibrillation. They attributed their higher observed accuracy during nighttime to factors such as reduced respiratory arrhythmia and fewer stimuli during the night [28]. To fully understand the effect of bodily movement on measurement accuracy in this study, the measures of accelerometry were analyzed in relation to HR accuracy. Unfortunately, not all CardioWatch data could be included in the analysis, as some data were missing. However, the outcomes confirmed that bodily movement negatively influences HR measurement accuracy. Multimedia Appendix 13 shows examples of 24-hour HR data with corresponding measures of accelerometry. As most wearables nowadays have built-in accelerometers, future research should include this accelerometry data to better understand the impact of bodily movements on measurement validity. In addition, improved algorithms that better filter out motion artifacts may help to enhance measurement accuracy [33].

The subgroup analysis revealed 2 notable findings. First, a positive but non-significant effect of BMI on CardioWatch performance was observed, contrasting with previous studies where higher BMI negatively impacted PPG accuracy [34,35]. Participants with a BMI >19 kg/m^2 demonstrated higher accuracy, potentially due to a better fit from increased wrist fat

XSL•FO
RenderX

tissue, similar to findings by Blok et al [26]. Likewise, the Hexoskin shirt showed higher HR accuracy in children with higher BMI, likely due to a tighter fit and subsequently less shifting of sensors, which could have improved ECG signal conduction. However, the BMI cutoff of 19 kg/m² may have oversimplified body composition differences, suggesting future research should use more detailed metrics to better understand their impact on accuracy. Second, Hexoskin accuracy declined over time, being nearly excellent during the first 12 hours—even showing resilience to bodily movements, as seen in Figure 8—but showed a significant reduction in the second 12-hour period. This decline likely resulted from the drying of conductive gel that was initially applied to the electrodes to improve signal conduction. As a result, noise increased, which was also visually observed in the signal during the second 12-hour measuring period. It is possible that the limited perspiration in children played a role too, as reduced moisture could have negatively affected signal transmission. To ensure consistent signal quality during long-term use, reapplying conductive gel every 12 hours is therefore recommended.

Besides HR monitoring, the Hexoskin shirt captures ECG rhythms. Of the 36 rhythm strips assessed, 31 were successfully classified by the blinded pediatric cardiologist, indicating that the Hexoskin can be useful for detecting minor arrhythmias, such as PVCs, PACs, and rhythms like ventricular trigeminy. However, further research is needed to evaluate the device's accuracy in detecting more clinically relevant arrhythmias, such as SVT and bradycardias, as they did not occur during the measurements of this study. If validated, the Hexoskin could serve as a comfortable alternative to the Holter monitor for prolonged cardiac monitoring in children, offering a more practical and patient-friendly diagnostic tool for patients with infrequent appearing symptoms.

Although patients were instructed not to interfere with the smartphone, some of them closed and terminated the app. As a result, 4 measurements were discontinued and could not be used for analysis. This should be considered when using wearables for extended monitoring periods. Children are curious by nature and tend to explore how things work, which can lead to failed monitoring once they accidentally terminate a measurement. It may be good use to implement child locks or parental control on devices used for monitoring in children. Another contributor to the loss of CardioWatch data in this research was the shift to a different measurement frequency at night, from 1/s to 1/min, caused by a bug in the firmware. This shift occurred during the last 9 measurements. This shows that the firmware must be reliable, especially when monitoring a critical patient group, where missing data could have detrimental effects.

Inaccurate measurements can also occur due to poor wearable fit, resulting in a loss of skin contact. For optimal performance, especially during more intense activities, the CardioWatch must not be placed too loose to allow shifting or too tight to impede circulation [27]. The same holds for the Hexoskin shirt. Although the researchers visually checked the fitting of both devices at the start of the 24-hour measurement period, it is possible that the fit shifted during monitoring. Regarding wrist size, no differences in accuracy were observed between participants with smaller wrists (≤15.5 cm) and those with larger wrist circumferences (>15 cm). This may be attributed to the stretchable fabric used in the CardioWatch, which likely ensured a secure fit regardless of the wrist size. Manufacturers should consider using more adaptable materials to maintain consistent skin contact and reduce sensor errors.

Currently, there is no clear consensus on the methodology for determining heart rate measurement validity and accuracy. The MAPE is often adhered to as a measure of error which defines measurement validity and accuracy, but the threshold percentage of MAPE varies amongst studies. Some studies use a MAPE threshold of <5% to define validity and accuracy [36-38], where others use a MAPE threshold of <10% [21,29,39]. Besides MAPE, various studies measure validity and accuracy by calculating the percentage of measurements within 10% of the criterion measure values [21,40-43], which is in line with recommendations of the American National Standards Institute (ANSI) and Association for the Advancement of Medical Instrumentation guidelines (AAMI) [20]. Although the MAPE, ANSI, and AAMI methodologies may appear similar, they represent fundamentally different concepts and should not be conflated. To provide a comprehensive assessment of measurement validity and accuracy, this study reported both the MAPE and the accuracy percentage measured following the ANSI and AAMI guidelines. The lack of consensus on validity and accuracy guidelines complicates comparisons between studies, and methodological differences between wearable validation studies highlight the need for universal standards [16].

Strengths and Limitations

This is one of the first studies that measured the validity of wearables in a cohort of children with CHD or (suspected) arrhythmias, while monitoring for prolonged duration under normal living conditions. Conducting this study outside the clinical setting enhanced its ecological validity by reflecting real-world conditions more accurately, which increased the generalizability of findings [26,44,45]. Another strength of this study is the included analysis of accelerometry, which shows the effect of bodily movement on measurement accuracy. On the other hand, a limitation of this study is the inability to fully validate the Hexoskin shirt's diagnostic accuracy for measuring heart rhythm abnormalities, due to the small number of children with arrhythmias present in this study, and the fact that only minor abnormalities were observed. To determine its validity, it is essential that the Hexoskin shirt is also tested in detecting clinically relevant rhythm abnormalities, such as brady- and tachycardias, which were not observed during the measurements of this study. Nonetheless, the arrhythmias that were present were accurately recorded. In addition, the study's cohort was relatively homogeneous, especially regarding BMI and skin type, which are known factors to influence accuracy. A more diverse study population would have made it possible to more accurately research the influence of these variables on measurement accuracy. Therefore, it is suggested that future studies adopt a multicenter study approach or include patients from different pediatric subgroups. Next to that, it is suggested to implement longer measuring periods to gain understanding of the technologies' long-term performance and durability. Although this study included a relatively small sample size, the

XSL•FO
RenderX

24-hour measurement recorded over 100,000 repeated measurements, which ensured adequate statistical power. Future research should, in addition to addressing the aforementioned considerations, focus on enhancing measurement accuracy during increased bodily movements and minimizing missing data. With these valuable insights, new algorithms could be introduced that could improve measurement accuracy by reducing noise that is caused by physical movements.

Conclusion

The CardioWatch and the Hexoskin shirt show good accuracy and strong correlation with the gold standard Holter ECG for monitoring heart rate in children attending the pediatric cardiology outpatient clinic. Both wearables offer improved comfort compared to the Holter, making them more suitable for prolonged use. The Hexoskin shirt's ECG recordings show potential for effective ambulant rhythm monitoring. However, further research is needed to address the decline in accuracy during periods of bodily movement and higher heart rates, reduce missing data, and to validate the Hexoskin shirt's capability to detect clinically significant arrhythmias in children. Conclusively, the wearables currently lack the validity required to replace clinical tools but could still contribute to valuable insights into heart rate and rhythm monitoring in noncritical situations, preferably during rest or sleep.

Acknowledgments

This research received funding from the nonprofit organization, Stichting Hartekind, which commits to increase survival chances and improve quality of life of children with congenital heart disease [46]. BB is supported by a grant from the Dutch Heart Foundation, Clinical Established Investigator (grant 03-001-2021-T105).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Questionnaire.
[DOCX File, 614 KB - formative_v9i1e70835_app1.docx]

Multimedia Appendix 2
Patient satisfaction questionnaire outcomes.
[DOCX File, 17 KB - formative_v9i1e70835_app2.docx]

Multimedia Appendix 3
Mean absolute error (MAE) and mean absolute percentage error (MAPE) equations.
[DOCX File, 37 KB - formative_v9i1e70835_app3.docx]

Multimedia Appendix 4
Overview of percentage missing data for all participants.
[DOCX File, 17 KB - formative_v9i1e70835_app4.docx]

Multimedia Appendix 5
Accuracy scores for the total 24-hour measurement period for all participants.
[DOCX File, 16 KB - formative_v9i1e70835_app5.docx]

Multimedia Appendix 6
Error scores (mean absolute error and mean absolute percentage error) for the total 24-hour measurement period for all participants.
[DOCX File, 18 KB - formative_v9i1e70835_app6.docx]

Multimedia Appendix 7
Concordance correlation coefficient scores for the total 24-hour measurement period for all participants.
[DOCX File, 20 KB - formative_v9i1e70835_app7.docx]

Multimedia Appendix 8

XSL•FO
RenderX

Frequency of included arrhythmias.
[DOCX File, 15 KB - formative_v9i1e70835_app8.docx]

Multimedia Appendix 9
Hexoskin heart rate measurement accuracy (%) per quantile of participants' accelerometry data.
[DOCX File, 23 KB - formative_v9i1e70835_app9.docx]

Multimedia Appendix 10
Concordance correlation coefficient outcomes for CardioWatch and Hexoskin.
[DOCX File, 16 KB - formative_v9i1e70835_app10.docx]

Multimedia Appendix 11
Accuracy scores subgroup analysis for CardioWatch.
[DOCX File, 46 KB - formative_v9i1e70835_app11.docx]

Multimedia Appendix 12
Accuracy scores subgroup analysis for Hexoskin.
[DOCX File, 46 KB - formative_v9i1e70835_app12.docx]

Multimedia Appendix 13
Recordings of heart rate difference including measures of bodily movement.
[DOCX File, 469 KB - formative_v9i1e70835_app13.docx]

References

1. Zimetbaum P, Goldman A. Ambulatory arrhythmia monitoring: choosing the right device. Circulation 2010 Oct 19;122(16):1629-1636. [doi: 10.1161/CIRCULATIONAHA.109.925610] [Medline: 20956237]
2. Holter NJ. New method for heart studies. Science 1961 Oct 20;134(3486):1214-1220. [doi: 10.1126/science.134.3486.1214] [Medline: 13908591]
3. Bonato P. Wearable sensors and systems. From enabling technology to clinical applications. IEEE Eng Med Biol Mag 2010;29(3):25-36. [doi: 10.1109/MEMB.2010.936554] [Medline: 20659855]
4. Reiffel JA, Schwarzberg R, Murry M. Comparison of autotriggered memory loop recorders versus standard loop recorders versus 24-hour Holter monitors for arrhythmia detection. Am J Cardiol 2005 May 1;95(9):1055-1059. [doi: 10.1016/j.amjcard.2005.01.025] [Medline: 15842970]
5. Giada F, Raviele A. Clinical approach to patients with palpitations. Card Electrophysiol Clin 2018 Jun;10(2):387-396. [doi: 10.1016/j.ccep.2018.02.010] [Medline: 29784490]
6. Mellion K, Uzark K, Cassedy A, et al. Health-related quality of life outcomes in children and adolescents with congenital heart disease. J Pediatr 2014 Apr;164(4):781-788. [doi: 10.1016/j.jpeds.2013.11.066] [Medline: 24412135]
7. Guk K, Han G, Lim J, et al. Evolution of wearable devices with real-time disease monitoring for personalized healthcare. Nanomaterials (Basel) 2019 May 29;9(6):813. [doi: 10.3390/nano9060813] [Medline: 31146479]
8. Sana F, Isselbacher EM, Singh JP, Heist EK, Pathik B, Armoundas AA. Wearable devices for ambulatory cardiac monitoring: JACC state-of-the-art review. J Am Coll Cardiol 2020 Apr 7;75(13):1582-1592. [doi: 10.1016/j.jacc.2020.01.046] [Medline: 32241375]
9. Bayoumy K, Gaber M, Elshafeey A, et al. Smart wearable devices in cardiovascular care: where we are and how to move forward. Nat Rev Cardiol 2021 Aug;18(8):581-599. [doi: 10.1038/s41569-021-00522-7] [Medline: 33664502]
10. Singh B, Chastin S, Miatke A, et al. Real-world accuracy of wearable activity trackers for detecting medical conditions: systematic review and meta-analysis. JMIR Mhealth Uhealth 2024 Aug 30;12:e56972. [doi: 10.2196/56972] [Medline: 39213525]
11. Smuck M, Odonkor CA, Wilt JK, Schmidt N, Swiernik MA. The emerging clinical role of wearables: factors for successful implementation in healthcare. NPJ Digit Med 2021 Mar 10;4(1):45. [doi: 10.1038/s41746-021-00418-3] [Medline: 33692479]
12. Tandon A, Nguyen HH, Avula S, et al. Wearable biosensors in congenital heart disease: needs to advance the field. JACC Adv 2023 Mar;2(2):100267. [doi: 10.1016/j.jacadv.2023.100267] [Medline: 37152621]
13. Behere SP, Janson CM. Smart wearables in pediatric heart health. J Pediatr 2023 Feb;253:1-7. [doi: 10.1016/j.jpeds.2022.08.009] [Medline: 36162539]
14. Zhang W, Xiong K, Zhu C, Evans R, Zhou L, Podrini C. Promoting child and adolescent health through wearable technology: A systematic review. Digit Health 2024;10:20552076241260507. [doi: 10.1177/20552076241260507] [Medline: 38868368]
15. Lambrechts L, Fourie B. How to interpret an electrocardiogram in children. BJA Educ 2020 Aug;20(8):266-277. [doi: 10.1016/j.bjae.2020.03.009] [Medline: 33456960]

XSL•FO
RenderX

16. Sartor F, Papini G, Cox LGE, Cleland J. Methodological shortcomings of wrist-worn heart rate monitors validations. J Med Internet Res 2018 Jul 2;20(7):e10108. [doi: 10.2196/10108] [Medline: 29967000]
17. The CardioWatch bracelet. Corsano. 2025. URL: https://corsano.com/products/bracelet-2 [accessed 2024-06-01]
18. Hexoskin. URL: https://hexoskin.com [accessed 2024-06-01]
19. Acc data explained. Corano. URL: https://developer.corsano.com/android/bracelet/acc_data [accessed 2025-05-20]
20. ANSI. Cardiac Monitors, Heart Rate Meters, and Alarms: AAMI; 2022.
21. Nelson BW, Allen NB. Accuracy of consumer wearable heart rate measurement during an ecologically valid 24-hour period: intraindividual validation study. JMIR Mhealth Uhealth 2019 Mar 11;7(3):e10828. [doi: 10.2196/10828] [Medline: 30855232]
22. Watson PF, Petrie A. Method agreement analysis: a review of correct methodology. Theriogenology 2010 Jun;73(9):1167-1179. [doi: 10.1016/j.theriogenology.2010.01.003] [Medline: 20138353]
23. Wallen MP, Gomersall SR, Keating SE, Wisløff U, Coombes JS. Accuracy of heart rate watches: implications for weight management. PLoS One 2016;11(5):e0154420. [doi: 10.1371/journal.pone.0154420] [Medline: 27232714]
24. Gupta V, Sharma VK. Skin typing: Fitzpatrick grading and others. Clin Dermatol 2019;37(5):430-436. [doi: 10.1016/j.clindermatol.2019.07.010] [Medline: 31896400]
25. Bent B, Goldstein BA, Kibbe WA, Dunn JP. Investigating sources of inaccuracy in wearable optical heart rate sensors. NPJ Digit Med 2020;3:18. [doi: 10.1038/s41746-020-0226-6] [Medline: 32047863]
26. Blok S, Piek MA, Tulevski II, Somsen GA, Winter MM. The accuracy of heartbeat detection using photoplethysmography technology in cardiac patients. J Electrocardiol 2021;67:148-157. [doi: 10.1016/j.jelectrocard.2021.06.009] [Medline: 34256184]
27. Merschel S, Reinhardt L. Analyzability of photoplethysmographic smartwatch data by the preventicus heartbeats algorithm during everyday life: feasibility study. JMIR Form Res 2022 Mar 28;6(3):e29479. [doi: 10.2196/29479] [Medline: 35343902]
28. Hermans ANL, Isaksen JL, Gawalko M, et al. Accuracy of continuous photoplethysmography-based 1 min mean heart rate assessment during atrial fibrillation. Europace 2023 Mar 30;25(3):835-844. [doi: 10.1093/europace/euad011] [Medline: 36748247]
29. Giggins OM, Doyle J, Smith S, Crabtree DR, Fraser M. Measurement of heart rate using the withings scanwatch device during free-living activities: validation study. JMIR Form Res 2022 Sep 1;6(9):e34280. [doi: 10.2196/34280] [Medline: 36048505]
30. Schäfer A, Vagedes J. How accurate is pulse rate variability as an estimate of heart rate variability? A review on studies comparing photoplethysmographic technology with an electrocardiogram. Int J Cardiol 2013 Jun 5;166(1):15-29. [doi: 10.1016/j.ijcard.2012.03.119] [Medline: 22809539]
31. Haddad M, Hermassi S, Aganovic Z, et al. Ecological validation and reliability of Hexoskin wearable body metrics tool in measuring pre-exercise and peak heart rate during shuttle run test in professional handball players. Front Physiol 2020;11:957. [doi: 10.3389/fphys.2020.00957] [Medline: 32848879]
32. Kinnunen H, Rantanen A, Kenttä T, Koskimäki H. Feasible assessment of recovery and cardiovascular health: accuracy of nocturnal HR and HRV assessed via ring PPG in comparison to medical grade ECG. Physiol Meas 2020 May 7;41(4):04nt01. [doi: 10.1088/1361-6579/ab840a] [Medline: 32217820]
33. Pollreisz D, TaheriNejad N. Detection and removal of motion artifacts in PPG signals. Mobile Netw Appl 2022 Apr;27(2):728-738. [doi: 10.1007/s11036-019-01323-6]
34. Boonya-Ananta T, Rodriguez AJ, Du Le VN, Ramella-Roman JC. Monte Carlo analysis of optical heart rate sensors in commercial wearables: the effect of skin tone and obesity on the photoplethysmography (PPG) signal. Biomed Opt Express 2021;12(12):7445. [doi: 10.1364/BOE.439893]
35. Boonya-Ananta T, Rodriguez AJ, Ajmal A, et al. Synthetic photoplethysmography (PPG) of the radial artery through parallelized Monte Carlo and its correlation to body mass index (BMI). Sci Rep 2021 Jan 28;11(1):2570. [doi: 10.1038/s41598-021-82124-4] [Medline: 33510428]
36. Shcherbina A, Mattsson CM, Waggott D, et al. Accuracy in wrist-worn, sensor-based measurements of heart rate and energy expenditure in a diverse cohort. J Pers Med 2017 May 24;7(2):3. [doi: 10.3390/jpm7020003] [Medline: 28538708]
37. Navalta JW, Montes J, Bodell NG, Salatto RW, Manning JW, DeBeliso M. Concurrent heart rate validity of wearable technology devices during trail running. PLoS One 2020;15(8):e0238569. [doi: 10.1371/journal.pone.0238569] [Medline: 32866216]
38. Navalta JW, Ramirez GG, Maxwell C, Radzak KN, McGinnis GR. Validity and reliability of three commercially available smart sports bras during treadmill walking and running. Sci Rep 2020 Apr 30;10(1):7397. [doi: 10.1038/s41598-020-64185-z] [Medline: 32355249]
39. Germini F, Noronha N, Borg Debono V, et al. Accuracy and acceptability of wrist-wearable activity-tracking devices: systematic review of the literature. J Med Internet Res 2022 Jan 21;24(1):e30791. [doi: 10.2196/30791] [Medline: 35060915]
40. Helmer P, Hottenrott S, Rodemers P, et al. Accuracy and systematic biases of heart rate measurements by consumer-grade fitness trackers in postoperative patients: prospective clinical trial. J Med Internet Res 2022 Dec 30;24(12):e42359. [doi: 10.2196/42359] [Medline: 36583938]

XSL•FO
RenderX

41. Fuller D, Colwell E, Low J, et al. Reliability and validity of commercially available wearable devices for measuring steps, energy expenditure, and heart rate: systematic review. JMIR Mhealth Uhealth 2020 Sep 8;8(9):e18694. [doi: 10.2196/18694] [Medline: 32897239]
42. Bai Y, Hibbing P, Mantis C, Welk GJ. Comparative evaluation of heart rate-based monitors: Apple Watch vs Fitbit Charge HR. J Sports Sci 2018 Aug;36(15):1734-1741. [doi: 10.1080/02640414.2017.1412235] [Medline: 29210326]
43. Hahnen C, Freeman CG, Haldar N, et al. Accuracy of vital signs measurements by a smartwatch and a portable health device: validation study. JMIR Mhealth Uhealth 2020 Feb 12;8(2):e16811. [doi: 10.2196/16811] [Medline: 32049066]
44. Villar R, Beltrame T, Hughson RL. Validation of the Hexoskin wearable vest during lying, sitting, standing, and walking activities. Appl Physiol Nutr Metab 2015 Oct;40(10):1019-1024. [doi: 10.1139/apnm-2015-0140] [Medline: 26360814]
45. Elliot CA, Hamlin MJ, Lizamore CA. Validity and reliability of the Hexoskin wearable biometric vest during maximal aerobic power testing in elite cyclists. J Strength Cond Res 2019 May;33(5):1437-1444. [doi: 10.1519/JSC.0000000000002005] [Medline: 28759538]
46. Stichting Hartekind. URL: https://www.hartekind.nl [accessed 2024-04-19]

Abbreviations

AAMI: Association for the Advancement of Medical Instrumentations
ANSI: American National Standards Institute
BPM: beats per minute
CCC: Lin concordance correlation coefficient
CHD: congenital heart disease
ECG: electrocardiography
HR: heart rate
LoA: limits of agreements
MAE: mean absolute error
MAPE: mean absolute percentage error
PAC: premature atrial contraction
PPG: photoplethysmography
PVC: premature ventricular contraction
SVT: supraventricular tachycardia
VT: ventricular tachycardia

Edited by A Mavragani; submitted 03.01.25; peer-reviewed by S Huang; revised version received 13.06.25; accepted 14.06.25; published 30.09.25.

Please cite as:
Hardon HJ, Van Kerkhof YN, Bartelds B, Kammeraad JAE, Van Deutekom AW
The Validation and Accuracy of Wearable Heart Rate Trackers in Children With Heart Disease: Prospective Cohort Study
JMIR Form Res 2025;9:e70835
URL: https://formative.jmir.org/2025/1/e70835
doi:10.2196/70835

© Hidde J Hardon, Yara N Van Kerkhof, Beatrijs Bartelds, Janneke A E Kammeraad, Arend W Van Deutekom. Originally published in JMIR Formative Research (https://formative.jmir.org), 30.9.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

AI-Enabled, Text-Based Health Coaching and Navigation for Employees to Support Health Outcomes: Pre-Post Observational Study

Paula Wilbourne[1,2*], PhD; Susan Mirch-Kretschmann[3*], PhD; Denise Walker[4], PhD; Michael Varghese[1], MBA; Roberto Arnetoli[1], MSc

[1]Sibly, Inc, 400 Concar Dr., San Mateo, CA, United States
[2]RecoveryWorksAI, Pacifica, CA, United States
[3]VA Palo Alto Health Care, U.S. Department of Veterans Affairs, Palo Alto, CA, United States
[4]School of Social Work, University of Washington, Seattle, WA, United States
[*]these authors contributed equally

Corresponding Author:
Paula Wilbourne, PhD
Sibly, Inc, 400 Concar Dr., San Mateo, CA, United States

Abstract

Background: Limited, timely access to quality mental health treatment harms well-being and quality of life while costing individuals and organizations millions in increased medical spending and reduced productivity. Too few qualified professionals, inconsistent quality, and stigma thwart traditional solutions, creating the need for scalable, science-based solutions.

Objective: This report provides an overview of a novel digital health coaching service that consists of artificial intelligence (AI)–assisted, human-delivered, text-based health coaching. This report provides data evaluating the efficacy of this service for delivering mental health support, improving well-being, and enhancing workplace productivity.

Methods: This observational study analyzed operational and self-reported health data from employees of subscribing organizations who used Sibly's digital health coaching service. Data included response times, changes in expressed member sentiment, conversation topics, and adherence to motivational interviewing. A subset of members (n=38) provided pre-post self-reported assessment measures of distress, unhealthy days, and presenteeism, having engaged in at least 4 coaching conversations over a minimum of 14 days. Sentiment was evaluated using a natural language processing tool.

Results: Sibly provided quick access to interactive human coaching, with a median response time of 132 seconds. Sentiment analysis showed that 57% (878/1540) of conversations increased in positive emotions. The coaches maintained strong fidelity to the techniques of motivational interviewing, with adherence exceeding 90% (387/430). The proportion of users reporting severe distress declined from 33.3% (10/30) at baseline to 6.7% (2/30) at follow-up, representing a 79% relative reduction ($P<.001$). Participants also reported a reduction in the number of unhealthy days per month, decreasing from 19.57 to 15.87 per month ($P=.02$). Self-reported productivity improved by 18% during the study period ($P<.001$). Additionally, 61% (47/77) of users who received referrals to additional employer-sponsored benefits engaged with those resources, demonstrating effective care navigation to relevant support services.

Conclusions: This report provides an overview of novel mental health support and navigation services that use AI-enabled, text-based health coaching and care navigation. Data suggest that the services provide effective, scalable mental health support in workplace settings. The platform helped reduce distress, improve well-being, and boost productivity by offering immediate access to trained coaches and personalized guidance. These results are consistent with existing research on digital mental health services. They highlight the potential of AI-assisted coaching to improve access to care. Future research should include larger, diverse populations and more rigorous randomized controlled trials. This formative report provides data that describes and demonstrates a proof of concept for an innovative technology-enabled service that addresses the problems of scalability, access, quality, and stigma that challenge the provision of traditional mental health services.

(JMIR Form Res 2025;9:e64553) doi:10.2196/64553

KEYWORDS

digital mental health; machine learning; health coaching; behavioral; artificial intelligence; quality of life; interviewing; proof of concept; scalability; access

XSL•FO
RenderX

Introduction

Mental Health Service Need

Untreated mental health problems are costly, having been linked to physical health problems, lower treatment adherence, and shorter lifespans [1,2]. Even before the pandemic, rates of emergency room visits for people with mental health diagnoses and comorbidities increased nationwide [3]. Finally, employees distressed at work impact culture, lower productivity of those around them, and burden human resource services [4].

In the years since the pandemic, the growing demand for mental health treatment has collided with the longstanding limitations in our capacity to provide it. Mental health provider shortages predated the pandemic and have worsened, with wait times averaging 5‑6 weeks [5-7]. Unfortunately, patients do not respond well to treatment delays with longer wait times, resulting in greater no-show rates [8]. The pandemic has worsened mental health, straining an already burdened system. Only 20% of adults who needed emotional support received help from a mental health professional [8]. Issues such as location, timing, and provider capacity (45%); costs (39%); and not knowing how to find help (21%) prevented access to care. Other barriers included privacy concerns, not thinking their problems are serious enough, concerns about stigma, or other people finding out [9]. Traditional entry to care is designed for more severe conditions, while 76‑90% of distressed populations do not want or need higher levels of care [10,11]. Evaluations of stepped-care models find that only about 20% of those seeking help needed medical treatment [11]. This mismatch wastes resources, can be off-putting to treatment seekers, and creates inefficiencies for efficacious treatment.

A large body of research identifies effective, efficient tools that hold promise in the face of the current mental health crisis. Despite strong evidence, traditional mental health and medical services have not taken robust advantage of this research. We will briefly review stepped care, the use of trained coaches and paraprofessional staff, digital interventions, text-based interventions, empirically based brief interventions, and measurement-based interventions as a foundation for the program that follows.

Quality often refers to evidence-based practices and the outcomes associated with those interventions [12]. Effective solutions to the mental health crisis must measure outcomes and ensure fidelity to empirically supported models to deliver the promised track record of efficacy. Using text-based interventions and real-time natural language processing (NLP) provides an opportunity to ensure that the services deliver quality in the form of adherence to evidence-based protocols. To be successful, quality, accessibility, and efficacy must be accomplished with fewer staff members, at a low cost, and promptly to meaningfully address the problems of existing solutions.

Stepped-care models provide lower levels of high-quality assistance, improve outcomes, and expand the capacity and accessibility of current mental health resources. Timely assistance (eg, self-help, digital tools, and coaches) can lead to lasting improvements [13-22]. Those who have to wait for assistance often continue to experience symptoms, get worse, and may never reach the same degree of improvement as those who get assistance quickly [13-18]. The advantages of self-help, digital resources, and nonmedical assistance, as initial steps in a stepped-care model, have been demonstrated in multiple studies of depression, substance use, anxiety, insomnia, and overall mental health symptoms [14-17,21-29]. Early initial steps of care prevent more severe mental health problems [13,14,23,30-33].

Trained coaches and nonprofessional staff provide another science-based option to address barriers to effective mental health treatment. Well-trained and supervised unlicensed staff can achieve similar outcomes as licensed professionals [34-36] when delivering well-specified interventions like motivational interviewing and cognitive-behavioral interventions [32,35-38]. Trained coaches and unlicensed staff have demonstrated both treatment fidelity and efficacy in the treatment of anxiety and depression [39] and in facilitating health behavior change [34,36-38]. The World Health Organization concluded that specialized staff are not required to deliver mental health intervention [40].

Fortunately, flexible options for delivering assistance, such as text and asynchronous support, go beyond traditional face-to-face or even video telehealth. SMS text messaging reduces the barriers to mental health treatment noted above, including access, location, timing, stigma, and privacy [41,42]. Digital content materials and SMS text messaging are effective tools [16,18,24,42,43]. Digital text messages written by trained coaches expand accessibility and lower the cost of mental health services with outcomes comparable to telephone coaching [39,42-49]. Text has become a ubiquitous form of communication and is the preferred method of communication for adults under 50 years of age. Over two-thirds of adults value SMS text messages from health care providers, and more than 41% are “constantly online” [50-52].

The potential of artificial intelligence (AI) in mental health care represents a revolutionary shift in managing mental health problems. Interactive and supportive chatbots make use of large language models, NLP, and machine learning (ML) to provide real-time interaction, creating a safe environment for users and offering immediate coping strategies, which may have the potential to address the shortage of mental health treatment resources [53,54]. NLP detects sentiment and language features linked to mental health symptoms. NLP can also evaluate the quality of care and adherence to evidence-based interventions demonstrated in clinical interactions. In general, using digital interventions and AI allows for real-time optimization of mental health interventions [55,56].

One aspect of NLP, sentiment analysis, has been applied to mental health content in the published literature. Sentiment analysis measures the attitudes, sentiments, evaluations, and emotions of a speaker or writer based on the computational treatment of subjectivity in a text [55]. It is generally performed using rule-based tools and lexicons to calculate the semantic orientation of words and phrases in a text. A sentiment score between −1 and +1 indicates generally negative or positive

XSL•FO
RenderX

sentiment, respectively (0=neutral sentiment). Sentiment analysis related to mental health concerns has been conducted on large text datasets, including those drawn from social media, electronic health records, narrative writing, and less frequently on smaller samples of the content of interviews or narrative writing samples [57-59].

Sibly: An Innovative Approach to Health Behavior Change

Sibly provides 24/7/365 digital coaching via text with trained human coaches to organizations that pay a fee to provide the service to their members. The platform integrates with the health and wellness resources of these organizations, allowing health coaches to navigate members to timely, effective referrals as needed during coaching. The platform gathers member and coach data to enhance service delivery.

Members maintain anonymity through display names, creating a secure channel for receiving coaching and educational materials related to health behaviors, life challenges, and mental health. The mobile-friendly platform eliminates barriers of location, waitlists, stigma, and timing, offering on-demand support.

Members use a messaging app and are assigned to a small team of health coaches, who respond collectively as "Sibly." Trained health coaches, using evidence-based practices such as motivational interviewing, mindfulness, coping skills training, and cognitive-behavioral tools, help members set goals and take action. In addition, Sibly's health coaches guide members to use self-help materials, sponsored benefits, and community resources that are recommended based on each member's individual needs.

Members receive unlimited coaching and progress reports. Baseline and follow-up well-being assessments are conducted. Personally identifiable information (ie, name, address, phone number, and date of birth) is collected and used only for emergencies (danger to self, others, or vulnerable persons) by PhD-level experts who respond in less than 30 minutes.

Protocols on the scope and development of text-based coaching proposed 5 domains: selection and training of coaches, specific coaching techniques, how to structure communication with those being coached, monitoring adherence to guidelines, and quality of coaching [39]. Sibly uses strategic recruitment and in-house, science-based coach training to address provider shortages and scale services. Successful applicants must have a bachelor's degree and are behaviorally evaluated for empathy, coachability, and professionalism. Qualified candidates are invited to participate in a paid training program.

Training consists of 240 hours of competency-based instruction in listening skills, motivational interviewing, cognitive-behavioral tools, mindfulness, and crisis response, grounding the service in rigorous, science-based interventions. New health coaches demonstrate competence during observed training cases. To date, 98 health coaches have entered training, with 94.9% (93/98) successfully completing it. Training each coach costs approximately US $4800. Approximately 40% of these costs are offset by the services they provide to members during training.

Quality assurance ensures health coaches retain and improve their skills. Sibly health coaches participate in quarterly continuing education classes and receive monthly feedback on their coaching skills from expert trainers using the adapted Motivational Interviewing Treatment Integrity (MITI) coding system [60]. A proprietary tool using ML and AI detects adherence to training skills and provides real-time feedback. Ninety-five percent of coded quality assurance conversations meet competency guidelines similar to those specified by the MITI. Health coaches not meeting competency standards participate in weekly conversation reviews until they improve. Those unable to improve are disqualified from continuing their work as health coaches.

AI and ML play a crucial role in supporting Sibly's human health coaches and structuring communication with members. AI suggests optimal next steps, identifies service improvement opportunities, enhances engagement, and improves the member experience. ML provides the health coaches with insights in real time that might not occur to the coach, or that may be based on previous conversations. Sibly's ML tools reliably detect 40 specific topics, measure sentiment fluctuations, identify the start and end of a text conversation, and suggest optimal next steps to the health coaches. The topic models were designed to reduce bias by accurately representing topics related to race, gender, and sexual orientation that are less frequent in the population and typically overlooked in models that are not specifically trained to detect them.

The platform provides health coaches with digital notifications or "nudges." Nudges assist the health coaches by highlighting information and recommending the next steps. Nudges personalize recommendations based on member priorities and goals. For example, if a member is interested in weight loss, nudges suggest relevant content and employer-sponsored benefits related to weight loss. If a member is concerned about their mood, nudges can suggest screening measures, self-help materials, and employer-sponsored mental health benefits related to this member's concern. The health coaches refine AI nudges through real-time feedback to the model.

The text-based communication platform allows for rigorous quality assurance and data analytics. AI and ML offer real-time feedback, reducing skill drift that can occur with episodic training. Through AI-assisted health coaching, Sibly is designed to provide personalized, high-quality, human-led coaching, resource navigation, and crisis response that is science-based, data-informed, efficient, consistent across the health coaches and, therefore, more scalable.

The current report includes observational data that describes the structure, service level, and impact of a novel digital health platform. We report on a subset of members who provided pre-post self-report questionnaires on mental and physical well-being.

Methods

Sample

We report on two data samples. First, over a 5-month period, all new members were asked to complete a baseline assessment

XSL•FO
RenderX

after completion of their first coaching conversation. After 2 weeks and completion of the fourth coaching conversation, these members were asked to complete a follow-up assessment. Thirty-eight members who completed a baseline assessment went on to complete 3 additional coaching conversations and were asked to complete a follow-up questionnaire. Of the 38 eligible participants, 30 (78.9%) agreed to complete the follow-up.

All 38 members meeting these eligibility criteria were detected using ML-enabled tools that prompted the coach to send the follow-up survey within 15 minutes of the end of the fourth coaching conversation. On average, the conversations of the original 38 members completing the baseline assessment involved the exchange of 25 messages, 11 of which were from the member. There were no significant differences in the length or timing of conversations between those completing a follow-up assessment and those who did not.

Sample selection reflects a combination of both theoretical and practical considerations. Previous product analyses suggest that members are more likely to complete a questionnaire after exchanging 10 messages with a coach; therefore, the baseline symptom assessment was sent after the end of the first conversation. The follow-up questionnaire was sent after members completed their fourth conversation to allow credible exposure to the service in the eyes of our customers. We limited the window of time to 14 days to reduce the chances that members might be lost to follow-up.

Second, we report data on our larger sample of operational data for the entire population of members using the service. The number of participants available for analysis is specified in each analysis below. All data available were included. No participants were selectively omitted from these analyses.

Ethical Considerations

During registration, users confirmed that they were at least 18 years old and consented to the use of their deidentified aggregate data. This study received a retrospective exemption from the institutional review board of the University of Washington Human Subjects Division (STUDY00023309) and was not considered human subjects research due to the observational and deidentified nature of the data collected.

All data management and analysis adhered to our terms of service and privacy policies, to which members agreed at enrollment. Messaging, operational, and survey data analyzed for ML were anonymous and processed in a separate system from member-identifying information. Data were reported only in aggregate and encrypted for privacy. Three data sources are reported here, including pre-post self-reported health data, operational data, and an ML analysis of the sentiment measured in participant conversations.

Quantitative Variables and Operational Data

Response time was measured as the seconds between a member's message and the coach's reply. Sequential member messages were counted as one.

The digital platform and phone app allowed us to measure, analyze, and optimize a number of product dimensions: response rates, time of day, sentiment, coaches per conversation, topics, and adherence to motivational interviewing. These data provide insight into the performance of the product and ways in which members used the service, which will be reported below.

Self-Reported Health-Related Questionnaire Data

Self-report measures were administered within the app. Pre-post data were then analyzed using a paired samples *t* test to determine the significance of changes made by participants using the service. Self-reported measures included the Lam Employment Absence and Productivity Scale (LEAPS) [61,62], health-related quality of life as measured by the Center for Disease Control and Prevention's Healthy Days [63], a 2-item measure of mental anguish and distress [64,65], and a single item assessing the member's evaluation of the change in mental well-being as a result of their use of the service.

The LEAPS is a 10-item scale with documented internal consistency and external validity. It took 3‑5 minutes to complete and was administered as a measure of the degree to which a participant's work performance was impacted by their mental health symptoms [60,61]. With the author's permission, distress was measured using 2 items assessing mental anguish that was first published in 2013 and later developed into a 10-item scale in 2019 [62,63]. The questions and response options are as follows: How much are you suffering emotionally (mental anguish, not pain, or discomfort in your body)? 1=absent; 2=very mild or occasionally; 3=mild, comes in moments and goes away; 4=moderate, steady and in specific moments; 5=marked, hurts all the time and does not get better; 6=severe, unbearable; 7=extreme, feels like you want to die; and 8=skip. A face-valid question assessing a member's change in mental well-being attributable to Sibly—"To what degree has your mental well-being changed as a result of working with Sibly?"—with response options that included 1=much better, at least 50% better, or improved on most days; 2=better, at least 25% better, or improved 2‑3 days per week; 3=a little better, some improvement, or better on 1 day per week; 4=no change; 5=a little worse, somewhat worse on at least 1 day per week; 6=worse, at least 25% worse, or worse 2‑3 days per week; 7=much worse, at least 50% worse, or worse on most days.

Sentiment Analysis

To assess the impact of coaching on sentiment, we analyzed the change in sentiment or sentiment shift. First, within a conversation, sentiment shift was defined by the change in sentiment from the first 5 messages of the conversation to the last 5 messages of the conversation. Second, sentiment shift was measured from the start of each member's first conversation to the end of the conversation, occurring immediately prior to their completion of the follow-up questionnaire. We used VADER (Valence Aware Dictionary and Sentiment Reasoner), a sentiment analysis tool designed to process informal text, including slang, emojis, and punctuation. VADER is part of the Natural Language Toolkit [66,67], a leading suite of Python libraries and programs for NLP. A sample of our first 1512 conversations with B2B (business-to-business or participants whose membership was paid for through an employer) members was analyzed to look at the number of coaches per conversation. These reflect the total number of conversations with employee

XSL•FO
RenderX

members that had taken place at the time of this analysis. No conversations or members were omitted. A chi-square test was used to look at the relationship between the length of the conversations and the sentiment of the conversation.

Coaches per Conversation

We analyzed 1549 B2B conversations to look at the number of coaches per conversation. These reflect the total number of conversations with employee members that had taken place at the time of this analysis. No conversations or members were omitted. A chi-square test was used to look at the relationship between the number of coaches per conversation (1-7) and members' experience of those conversations as reflected by the percentage of conversations with a positive sentiment shift.

Topics

We used an unsupervised topic modeling technique, Latent Dirichlet Allocation (LDA), to identify conversation topics. LDA is a generative statistical model used in NLP and ML to identify thematic structures within large text datasets. Overall, we used 3 trained health coaches as coders to label the topics detected by the models. We then asked the coders to rate the degree to which the proposed models captured the topics discussed in 100 randomly selected conversations. The output and evaluation of this process are reported below.

Results

Response Time

Coaches responded to member messages with a median response time of 132 seconds and an average of 197 seconds.

Optimizing Questionnaire Response Rates

Before December 2021, approximately 29% (2795/9565) of our requests to members to complete surveys were acted on. To improve this response rate, we examined optimal points in the member relationship (number of messages sent) and the timing in the conversation at which members were most likely to complete a questionnaire or survey (within 15 minutes of the end of a conversation). Optimizing survey administration along these dimensions increased our response rate to 48.6% (2041/4201; χ^2_1=485.4, n=13,766, P<.001).

Time of Day and Work Hours

Fifty-six percent (8938/15,960) of our member messages were sent outside of the hours of 9 AM to 5 PM, suggesting that most employees are accessing assistance without disruption to their workday and outside the hours that most mental health services would be available to them (see Figure 1). During work hours, members take twice as long to respond to coaches as they do outside of work hours, suggesting the ability to fit coaching conversations amid other demands on their time.

Figure 1. Distribution of employee messages by time of day, 2021 - 2023.

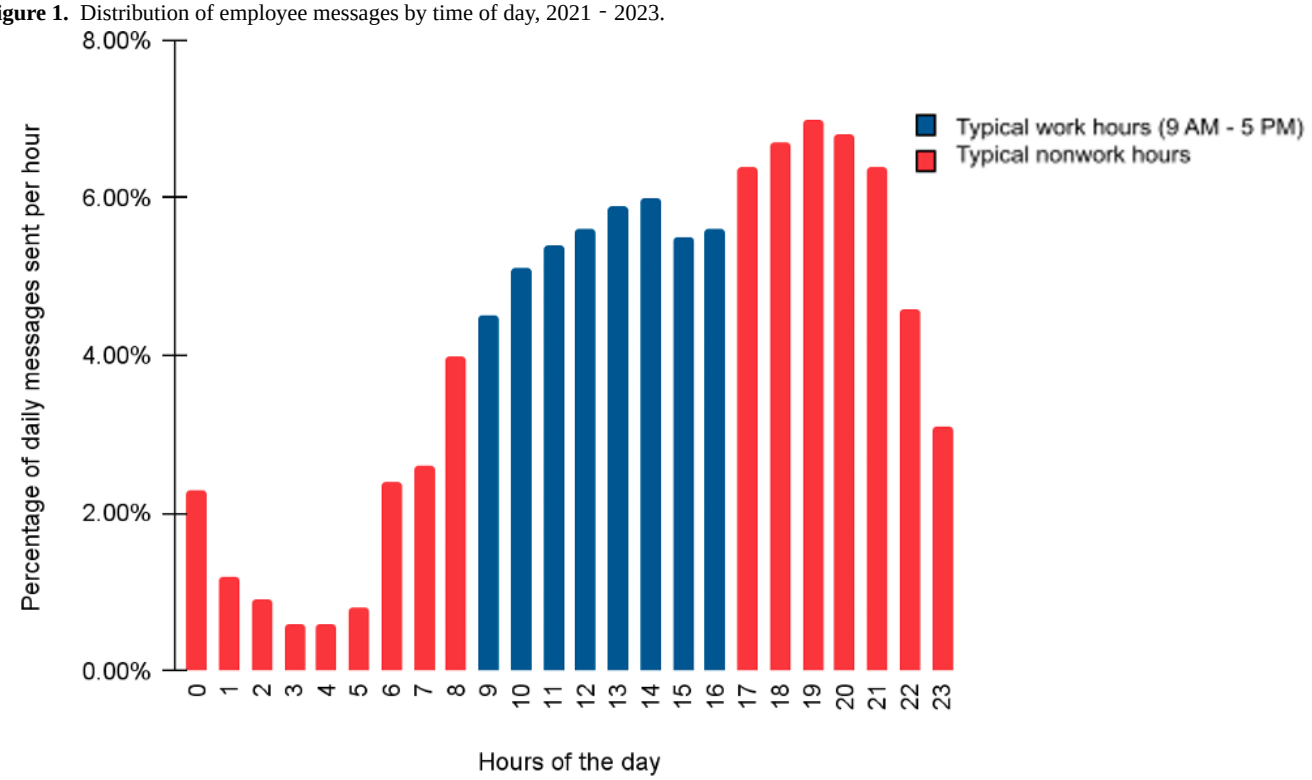

Sentiment

Over each member's relationship with Sibly, the slope of the sentiment line between the beginning of the first conversation and the end of the conversation occurring before the follow-up assessment was positive for a majority of the participants. Twenty-one (70%) of the sample of 30 members completing pre-post measures demonstrated a positive shift in sentiment, while 9 demonstrated a negative shift in sentiment. Within individual conversations, an average of 57% (878/1540) demonstrated a positive sentiment shift. This information provided insight into the impact of coaching conversations on member sentiment that occurred during the member conversations and over the initial coaching conversations. In

XSL•FO
RenderX

our larger member population, there was a positive correlation between the length of member conversations with sentiment shift suggesting that the longer a member talks with a coach, the better they feel or that the better a person is feeling, the longer they are likely to continue talking to Sibly (χ^2_5=58.01, n=1512, *P*<.001).

Coaches per Conversation

In a larger sample of 1549 member conversations, 70% (1084/1549) of conversations were completed by a single coach, 23% (356/1549) of conversations included responses from 2 coaches, while 7% (108/1549) of conversations included 3 to 7 coaches. Sentiment analysis did not detect a decrease in sentiment associated with conversations that switched between one coach and another. In fact, conversations with more than one coach were longer and demonstrated a greater improvement in sentiment than those in which a member spoke to fewer health coaches during the conversation (χ^2_3=17.701, n=1549, *P*<.001). Generally, members are not aware that they have spoken to more than one coach during a single conversation. However, in <2% (23/1549) of conversations, members express concern or dissatisfaction about working with more than one coach or the perceived change in coach during a conversation.

Topics

Using an unsupervised topic modeling technique, we developed and evaluated 7 possible models with 30, 40, or 50 topic clusters. Topics were labeled, and labels were validated by 3 independent coders. We then asked the coders to rate the fit of these models on a 5-point Likert scale for 100 randomly selected conversations. Models A-F were rated as “best reflecting the conversations” as follows: A (45/100, 45%), B (21/100, 21%), C (17/100, 17%), D (10/100, 10%), E (10/100, 10%), F (2/100, 2%), and G (1/100, 1%) of the time. Model A, with 40 topics identified, was selected as the one best reflecting the conversations 45% (45/100) of the time, with the topics detected in model A evaluated as fair to very good 84% (84/100) of the time.

Applying the selected model, we found that members came to the first conversation with an average of 5.6 topics, where over 76% (12,129/15,960) presented with at least 3 interconnected topics of conversation. Thirty-one topics of conversation were detected across all member conversations in this sample. Negative emotions, work, family, relationships, and mental health were the most common topics of conversation, occurring in more than 20% of conversations. Self-organization, self-improvement, employer-sponsored benefits, health behaviors, and love were discussed in more than 10% of conversations. Twenty-one additional topics occurred in fewer than 10% of the conversations. See Table 1 for all topics.

XSL•FO
RenderX

Table . Machine learning–detected topics from member-initiated messages (n=15,960).

Topic	Values, n (%)
Negative emotions	5108 (32.0)
Work	5152 (32.3)
Family	4739 (29.7)
Relationship	3783 (23.7)
Mental health	3377 (21.2)
Self-organization	2626 (16.5)
Self-improvement	2472 (15.5)
Discussing benefits	2457 (15.4)
Health behaviors	2417 (15.1)
Love	2319 (14.5)
Sleep	1385 (8.7)
Emotions: hope	1252 (7.8)
Living situation	1227 (7.7)
Hobbies	1188 (7.4)
Coping strategies	1126 (7.1)
Divorce	1115 (7.0)
Finance	1115 (7.0)
Medical problem	1090 (6.8)
Education	1050 (6.6)
COVID	1025 (6.4)
Goals focusing	875 (5.5)
Religion	666 (4.2)
Happy fun	610 (3.8)
Relaxation	548 (3.4)
Racism	337 (2.1)
Legal	272 (1.7)
LGBTQ+	265 (1.7)
Friendship	257 (1.6)
Disability	252 (1.6)
Politics	113 (0.7)
Gender	40 (0.3)

Evidence-Based Practice Adherence

During the period in which data were collected, 2 conversations per coach per month were coded for adherence to motivational interviewing using the Sibly quality assurance manual, which was based on the MITI scale [60]. This resulted in 430 coded conversations, of which 387 (90%) were determined to be adherent to motivational interviewing.

Distress

Participants were asked to rate their levels falling into categories of “1=mild or none” for those indicating “absent or occasionally,” “2=moderate” for those indicating “comes and goes or steady,” and “3=severe” for those indicating that it “hurts all the time, unbearable, and feels like they want to die.” These groupings were chosen to aid in the interpretability when presenting to nonclinical audiences. These were referred to triage with PhD, per previously mentioned procedure, earning scores 1 - 3, respectively, at both baseline and follow-up. At baseline, participants reported an average distress level of 2.27, with 33.3% (10/30) reporting severe distress. After using the service, average distress decreased to 1.87, with an 80% decrease in those reporting severe distress and a 200% increase in those reporting mild or no distress. A paired samples 2-tailed *t* test found a significant improvement in average distress levels, $t_{29}=-3.89$ ($P<.001$), and a medium effect size $d=0.71$.

XSL•FO
RenderX

Unhealthy Days

Participants were queried as to the number of unhealthy days in the past month attributable to mental health and then physical health. These numbers were added together up to a maximum of 30 days, consistent with the scoring recommendations for this measure by the Centers for Disease Control. Participants also reported an average of 19.57 unhealthy days per month at baseline, with 60% (18/30) experiencing poor health (16+ unhealthy days/month). At follow-up, unhealthy days decreased to 15.87 (P=.02, d=0.42), and the proportion of members in poor health dropped to 36.7% (11/30).

Productivity or Presenteeism

Analyzing data from the LEAPS questionnaire found that at baseline assessment, participants scored an 8.5, which falls in the mild range for productivity impairment due to mental health symptoms at work. At follow-up, participants reported ~18% (1.5/8.5) decrease with a mean score of 7. A paired samples 2-tailed t test found a significant improvement in productivity, t_{29}=–3.99 (P<.001), and a medium effect size d=0.71.

Benefit Referral and Engagement

During the evaluation period, the health coaches recommended 18 benefit programs to 14 participants. Members engaged with 11 (61.1%) of the benefit services recommended to them. Additionally, among a sample of 77 participants who were referred to their EAP program, 46 of these participants enrolled in that benefit program. For comparison, a third-party comparison of Sibly to more traditional benefit navigators found that Sibly's referral success rate was 3 times higher than that of more traditional navigation services.

Discussion

This paper describes a novel AI-enabled, text-based health coaching platform that supports mental health via immediate, accessible assistance from live human coaches. Analysis found that the health coaches respond to members in less than 3 minutes. Objective quality assurance ratings found that more than 90% (387/430) of conversations were adherent to motivational interviewing and that 56% (8938/15,960) of member messages were sent outside of traditional clinical service hours of 9 AM to 5 PM. While 70% (1084/1549) of member conversations were conducted by a single coach, sentiment did not decline when multiple coaches participated; instead, conversations with multiple coaches were longer, with a greater improvement in sentiment. Over time, a majority of members showed an increase in positive sentiment between their first and last conversations and within the majority of individual conversations. The most common topics of conversation included negative emotions, work, family, relationships, and mental health. Other common topics included self-organization, self-improvement, employer benefits, health behaviors, and love. A subset of members providing pre-post self-report data reported an 80% decrease in severe distress, a 19% decrease in unhealthy days, and an 18% increase in productivity. Sixty-one percent of the time that a health coach referred a member to an additional employer-sponsored benefit, the referral was successful. The results demonstrate the potential of AI-supported text coaching as an efficient, scalable, and effective workplace mental health solution.

There are important implications for the key findings that highlight the speed of access, rapid response times, consistent use of empirically based tools, and improvements in distress, productivity, and engagement with employer-sponsored benefits. The analyses presented indicate that the service increases accessibility for individuals who may not want or cannot schedule therapy, reduces delays in care, and holds a credible promise to supplement traditional services, especially in those who need immediate support.

The service uses a one-to-many model, a novel and efficient way to scale the human coaching relationship. Each member is assigned to a small team of coaches who are selected for empathy and are trained in empirically based tools, allowing the coaches to speak with a similar voice and deliver a consistent intervention. The analysis reported shows that coach transitions within a single conversation do not impact member sentiment. This finding supports the credibility of the one-to-many model, where each member is assigned to a small team of coaches who respond as a single entity. Sibly's digital platform and AI support enable a seamless collaboration among providers, which is difficult to achieve in traditional care.

Members most frequently discussed topics of negative emotion, work, family, relationships, and mental health and aligned with previous findings that work-related stress, personal relationships, and emotional distress significantly impact mental health challenges. The fact that work stress ranked second underscores the importance of workplace mental health solutions, and these findings suggest that text-based coaching may improve mental well-being and workplace functioning. Previous research has found that mental health interventions enhance productivity and reduce absenteeism at work [68].

These findings are consistent with prior research on digital mental health interventions, which have been shown to improve access to care and support well-being in a scalable way [69]. Stepped-care models suggest that lower-intensity interventions, such as text-based coaching, can be highly effective for individuals who do not require or want more traditional therapy or clinical treatment [60]. By integrating benefit navigation, Sibly addresses a common challenge seen in both mental health support and employer-sponsored benefits—helping individuals connect with appropriate resources when they are needed. ML also played a role in increasing engagement and response rates, helping to optimize interactions in ways that traditional benefit navigation methods cannot do. This suggests that AI-assisted coaching could enhance accessibility and service quality while also supporting employee well-being and workplace productivity.

Despite these encouraging results, this study has several limitations. Because it was observational and lacked a control group, we cannot conclude with certainty that Sibly was the direct cause of the observed improvements. While members reported significant reductions in distress and unhealthy days, external factors could have influenced these outcomes. Additionally, the sample size was relatively small and composed of employees with access to employer-sponsored benefits,

XSL•FO
RenderX

meaning the results may not generalize to broader populations. Furthermore, employees self-selected to use the service, which may introduce selection bias, as those who engaged with the platform might differ systematically from those who did not. As a result, the findings may not fully reflect the experiences or needs of the larger employee population. Future research should aim to include larger and more diverse samples, as well as implement randomized controlled trials to establish causality. While sentiment analysis and engagement metrics provide useful insights into user experiences, more qualitative research is needed to understand how members perceive and engage with the platform.

The broader implications of these findings highlight the increasing role of digital health coaching in expanding access to mental health support and reducing the burden on traditional clinical services. AI-enabled platforms like Sibly offer a promising, scalable, and cost-effective solution to the growing need for workplace mental health solutions. Digital services have the potential to shift mental health care from a reactive, appointment-based medical model to a proactive, on-demand system. As AI and NLP continue to advance, these services will likely become more precise, responsive, and personalized to individual users. Additional research and transparent evaluation will be crucial in refining these models and ensuring that digital mental health solutions continue to meet the evolving needs of individuals and organizations alike.

Acknowledgments

The authors wish to thank Cindy Levin Eaton, PhD, Vice President of Coaching, Sibly, Inc, and Nguy n Th Minh Tâm, ML Coding Supervisor, for their contributions to this project. We also thank Dr Raymond Lam and Dr Eliana Tossani for their permission to use their published measures in this project. ChatGPT was used to identify content for potential deletion to shorten the manuscript and to provide feedback on adherence to the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines. The authors reviewed suggestions and revised the manuscript based on their judgment.

Data Availability

Data analyzed in this study are not publicly available due to the proprietary nature of the data generated by a commercial organization.

Authors' Contributions

Conceptualization: PW (lead), SM-K (equal), DW (supporting)
Data curation: PW, MV (lead), RA
Formal analysis: MV (lead), RA (supporting), PW (supporting), DW (supporting)
Funding acquisition: PW
Investigation: PW, SM-K, DW, MV, RA
Methodology: PW, SM-K, DW, MV, RA
Project administration: DW (lead), PW (supporting)
Resources: PW
Supervision: PW
Validation: PW, SM-K, DW, MV, RA
Visualization: MV (lead), RA (supporting)
Writing – original draft: SM-K (lead), PW (equal)
Writing – review & editing: DW (lead), PW (supporting), SM-K (supporting), RA (supporting)

Conflicts of Interest

PW, MV, and RA each have a small amount of stock in the company that purchased Sibly. PW was the founder and chief science officer at Sibly, and SM-K, MV, and RA are consultants with Sibly. No other conflicts of interest exist.

References

1. Understanding the link between chronic disease and depression [NIH Publication No. 24-MH-8015]. National Institute of Mental Health. URL: https://www.nimh.nih.gov/health/publications/chronic-illness-mental-health [accessed 2025-08-27]
2. Valiant GE. Natural history of male psychologic health – effects of mental health on physical health. N Engl J Med 1979;301:1249-1254. [doi: 10.1056/NEJM197912063012302]
3. Capp R, Hardy R, Lindrooth R, Wiler J. National trends in emergency department visits by adults with mental health disorders. J Emerg Med 2016 Aug;51(2):131-135. [doi: 10.1016/j.jemermed.2016.05.002] [Medline: 27614303]
4. Hemp P. Presenteeism: at work—but out of it. Harvard Business Review. 2004 Oct. URL: https://hbr.org/2004/10/presenteeism-at-work-but-out-of-it [accessed 2025-08-27]
5. SAMHSA. URL: https://www.samhsa.gov/about/careers/behavioral-health-workforce [accessed 2022-05-25]

XSL•FO
RenderX

6. Gruber J, Prinstein MJ, Clark LA, et al. Mental health and clinical psychological science in the time of COVID-19: challenges, opportunities, and a call to action. Am Psychol 2021 Apr;76(3):409-426. [doi: 10.1037/amp0000707] [Medline: 32772538]
7. Dembosky A. Americans can wait many weeks to see a therapist: California law aims to fix that. NPR. 2021 Nov 18. URL: https://www.npr.org/sections/health-shots/2021/11/18/1053566020/americans-can-wait-many-weeks-to-see-a-therapist-California-law-aims-to-fix-that [accessed 2025-08-27]
8. Gallucci G, Swartz W, Hackerman F. Impact of the wait for an initial appointment on the rate of kept appointments at a mental health center. Psychiatr Serv 2005 Mar;56(3):344-346. [doi: 10.1176/appi.ps.56.3.344]
9. Conroy J, Lin L, Ghaness A. Why people aren't getting the care they need. Monitor on Psychology. 2020 Jul 1 URL: https://www.apa.org/monitor/2020/07/datapoint-care [accessed 2025-08-27]
10. Nicholas J, Ringland KE, Graham AK, et al. Stepping up: predictors of "stepping" within an iCBT stepped-care intervention for depression. Int J Environ Res Public Health 2019 Nov 25;16(23):4689. [doi: 10.3390/ijerph16234689] [Medline: 31775297]
11. Richards DA, Bower P, Pagel C, et al. Delivering stepped care: an analysis of implementation in routine practice. Implement Sci 2012 Jan 16;7(1):3. [doi: 10.1186/1748-5908-7-3] [Medline: 22248385]
12. Kilbourne AM, Beck K, Spaeth-Rublee B, et al. Measuring and improving the quality of mental health care: a global perspective. World Psychiatry 2018 Feb;17(1):30-38. [doi: 10.1002/wps.20482] [Medline: 29352529]
13. Harris KB, Miller WR. Behavioral self-control training for problem drinkers: components of efficacy. Psychol Addict Behav 1990;4(2):82-90. [doi: 10.1037/0893-164X.4.2.82]
14. Miller WR, Baca LM. Two-year follow-up of bibliotherapy and therapist-directed controlled drinking training for problem drinkers. Behav Ther 1983 Jun;14(3):441-448. [doi: 10.1016/S0005-7894(83)80107-5]
15. Edwards G, Orford J, Egert S, et al. Alcoholism: a controlled trial of "treatment" and "advice". J Stud Alcohol 1977 May;38(5):1004-1031. [doi: 10.15288/jsa.1977.38.1004] [Medline: 881837]
16. Lancee J, van den Bout J, van Straten A, Spoormaker VI. Internet-delivered or mailed self-help treatment for insomnia? A randomized waiting-list controlled trial. Behav Res Ther 2012 Jan;50(1):22-29. [doi: 10.1016/j.brat.2011.09.012] [Medline: 22055281]
17. Schmidt MM, Miller WR. Amount of therapist contact and outcome in a multidimensional depression treatment program. Acta Psychiatr Scand 1983 May;67(5):319-332. [doi: 10.1111/j.1600-0447.1983.tb00349.x] [Medline: 6869040]
18. Moberg C, Niles A, Beermann D. Guided self-help works: randomized waitlist controlled trial of Pacifica, a mobile app integrating cognitive behavioral therapy and mindfulness for stress, anxiety, and depression. J Med Internet Res 2019 Jun 8;21(6):e12556. [doi: 10.2196/12556] [Medline: 31199319]
19. Knapstad M, Smith ORF. Social anxiety and agoraphobia symptoms effectively treated by prompt mental health care versus TAU at 6- and 12-month follow-up: secondary analysis from a randomized controlled trial. Depress Anxiety 2021 Mar;38(3):351-360. [doi: 10.1002/da.23132] [Medline: 33393688]
20. Knapstad M, Lervik LV, Sæther SMM, Aarø LE, Smith ORF. Effectiveness of prompt mental health care, the Norwegian version of improving access to psychological therapies: a randomized controlled trial. Psychother Psychosom 2020;89(2):90-105. [doi: 10.1159/000504453] [Medline: 31794968]
21. Edwards G, Guthrie S. A controlled trial of inpatient and outpatient treatment of alcohol dependency. Lancet 1967 Mar;289(7489):555-559. [doi: 10.1016/S0140-6736(67)92129-0]
22. Miller WR, Gribskov CJ, Mortell RL. Effectiveness of a self-control manual for problem drinkers with and without therapist contact. Int J Addict 1981 Oct;16(7):1247-1254. [doi: 10.3109/10826088109039178] [Medline: 7327785]
23. Behavioral self-control training. Williammiller.net. URL: https://williamrmiller.net/behavioral-self-control-training/ [accessed 2025-08-27]
24. Andrews G, Cuijpers P, Craske MG, McEvoy P, Titov N. Computer therapy for the anxiety and depressive disorders is effective, acceptable and practical health care: a meta-analysis. PLoS ONE 2010 Oct 13;5(10):e13196. [doi: 10.1371/journal.pone.0013196] [Medline: 20967242]
25. Andrews G, Tolkien II Team. Tolkien II: a needs-based, costed, stepped-care model for mental health services, clinical pathways, treatment flowcharts, costing structures. : World Health Organization Collaborating Centre for Classification in Mental Health; 2006 URL: https://researchers.mq.edu.au/en/publications/tolkien-ii-a-needs-based-costed-stepped-care-model-for-mental-hea [accessed 2025-08-27]
26. Cuijpers P, Donker T, van Straten A, Li J, Andersson G. Is guided self-help as effective as face-to-face psychotherapy for depression and anxiety disorders? A systematic review and meta-analysis of comparative outcome studies. Psychol Med 2010 Dec;40(12):1943-1957. [doi: 10.1017/S0033291710000772] [Medline: 20406528]
27. Ho FYY, Yeung WF, Ng THY, Chan CS. The efficacy and cost-effectiveness of stepped care prevention and treatment for depressive and/or anxiety disorders: a systematic review and meta-analysis. Sci Rep 2016 Jul 5;6:29281. [doi: 10.1038/srep29281] [Medline: 27377429]
28. Richards D, Richardson T. Computer-based psychological treatments for depression: a systematic review and meta-analysis. Clin Psychol Rev 2012 Jun;32(4):329-342. [doi: 10.1016/j.cpr.2012.02.004] [Medline: 22466510]
29. Rivero-Santana A, Perestelo-Perez L, Alvarez-Perez Y, et al. Stepped care for the treatment of depression: a systematic review and meta-analysis. J Affect Disord 2021 Nov 1;294:391-409. [doi: 10.1016/j.jad.2021.07.008] [Medline: 34315101]

XSL•FO
RenderX

30. Salomonsson S, Santoft F, Lindsäter E, et al. Stepped care in primary care–guided self-help and face-to-face cognitive behavioural therapy for common mental disorders: a randomized controlled trial. Psychol Med 2018 Jul;48(10):1644-1654. [doi: 10.1017/S0033291717003129] [Medline: 29095133]
31. van't Veer-Tazelaar PJ, van Marwijk HWJ, van Oppen P, et al. Stepped-care prevention of anxiety and depression in late life: a randomized controlled trial. Arch Gen Psychiatry 2009 Mar;66(3):297-304. [doi: 10.1001/archgenpsychiatry.2008.555] [Medline: 19255379]
32. American Psychological Association. Guidelines for prevention in psychology. American Psychologist 2014;69(3):285-296. [doi: 10.1037/a0034569]
33. National Research Council and Institute of Medicine. Preventing Mental, Emotional, and Behavioral Disorders Among Young People: Progress and Possibilities: National Academies Press; 2009. URL: https://www.fredla.org/wp-content/uploads/2022/08/Preventing-Mental-Emotional-and-Behavioral-Disorders-Among-Young-People.pdf [accessed 2025-08-27]
34. Diebold A, Ciolino JD, Johnson JK, Yeh C, Gollan JK, Tandon SD. Comparing fidelity outcomes of paraprofessional and professional delivery of a perinatal depression preventive intervention. Adm Policy Ment Health 2020 Jul;47(4):597-605. [doi: 10.1007/s10488-020-01022-5] [Medline: 32086657]
35. Valentine SE, Ahles EM, Dixon De Silva LE, et al. Community-based implementation of a paraprofessional-delivered cognitive behavioral therapy program for youth involved with the criminal justice system. J Health Care Poor Underserved 2019;30(2):841-865. [doi: 10.1353/hpu.2019.0059] [Medline: 31130554]
36. Hattie JA, Sharpley CF, Rogers HJ. Comparative effectiveness of professional and paraprofessional helpers. Psychol Bull 1984 May;95(3):534-541 [FREE Full text] [Medline: 6399757]
37. Montgomery EC, Kunik ME, Wilson N, Stanley MA, Weiss B. Can paraprofessionals deliver cognitive-behavioral therapy to treat anxiety and depressive symptoms? Bull Menninger Clin 2010;74(1):45-62. [doi: 10.1521/bumc.2010.74.1.45] [Medline: 20235623]
38. Rubak S, Sandbaek A, Lauritzen T, Christensen B. Motivational interviewing: a systematic review and meta-analysis. Br J Gen Pract 2005 Apr;55(513):305-312. [Medline: 15826439]
39. Lattie EG, Graham AK, Hadjistavropoulos HD, Dear BF, Titov N, Mohr DC. Guidance on defining the scope and development of text-based coaching protocols for digital mental health interventions. Digit Health 2019;5. [doi: 10.1177/2055207619896145] [Medline: 31897306]
40. World Health Organization. mhGAP intervention guide for mental, neurological and substance use disorders in non-specialized health settings: mental health gap action programme (mhGAP), version 2.0. 2016. URL: https://apps.who.int/iris/handle/10665/250239 [accessed 2025-08-27]
41. Muñoz RF, Chavira DA, Himle JA, et al. Digital apothecaries: a vision for making health care interventions accessible worldwide. Mhealth 2018;4:18. [doi: 10.21037/mhealth.2018.05.04] [Medline: 30050914]
42. Rathbone AL, Prescott J. The use of mobile apps and SMS messaging as physical and mental health interventions: systematic review. J Med Internet Res 2017 Aug 24;19(8):e295. [doi: 10.2196/jmir.7740] [Medline: 28838887]
43. Cho SMJ, Lee JH, Shim JS, et al. Effect of smartphone-based lifestyle coaching app on community-dwelling population with moderate metabolic abnormalities: randomized controlled trial. J Med Internet Res 2020 Oct 9;22(10):e17435. [doi: 10.2196/17435] [Medline: 33034564]
44. Markert C, Sasangohar F, Mortazavi BJ, Fields S. The use of telehealth technology to support health coaching for older adults: literature review. JMIR Hum Factors 2021 Jan 29;8(1):e23796. [doi: 10.2196/23796] [Medline: 33512322]
45. Lindner P, Olsson EL, Johnsson A, Dahlin M, Andersson G, Carlbring P. The impact of telephone versus e-mail therapist guidance on treatment outcomes, therapeutic alliance and treatment engagement in Internet-delivered CBT for depression: a randomised pilot trial. Internet Interv 2014 Oct;1(4):182-187. [doi: 10.1016/j.invent.2014.09.001]
46. Gupta I, Di Eugenio B, Ziebart B, et al. Human-human health coaching via text messages: corpus, annotation, and analysis. In: Proceedings of the 21st Annual Meeting of the Special Interest Group on Discourse and Dialogue: Association for Computational Linguistics; 2020:246-256. [doi: 10.18653/v1/2020.sigdial-1.30]
47. Dol J, Aston M, Grant A, McMillan D, Tomblin Murphy G, Campbell-Yeo M. Effectiveness of the "essential coaching for every mother" postpartum text message program on maternal psychosocial outcomes: a randomized controlled trial. Digit Health 2022;8. [doi: 10.1177/20552076221107886] [Medline: 35720618]
48. Gell NM, Grover KW, Savard L, Dittus K, Mace E. Outcomes of a text message, Fitbit, and coaching intervention on physical activity maintenance among cancer survivors: a randomized control pilot trial. J Cancer Surviv 2020 Feb;14(1):80-88. [doi: 10.1007/s11764-019-00831-4] [Medline: 31776849]
49. Oreopoulos P, Petronijevic U, Logel C, Beattie G. Improving non-academic student outcomes using online and text-message coaching. J Econ Behav Organ 2020 Mar;171:342-360. [doi: 10.1016/j.jebo.2020.01.009]
50. Gelles-Watnick R. Americans' use of mobile technology and home broadband. Pew Research Center. 2024 Jan 31. URL: https://www.pewresearch.org/internet/2024/01/31/home-broadband-mobile-acknowledgments/ [accessed 2025-08-27]
51. Rainie L, Zickuhr K. Americans' views on mobile etiquette. Pew Research Center. 2015 Aug 26. URL: https://www.pewresearch.org/internet/2015/08/26/americans-views-on-mobile-etiquette/ [accessed 2025-08-27]

XSL•FO
RenderX

52. Campbell KJ, Blackburn BE, Erickson JA, et al. Evaluating the utility of using text messages to communicate with patients during the COVID-19 pandemic. J Am Acad Orthop Surg Glob Res Rev 2021 Jun 15;5(6):e21.00042. [doi: 10.5435/JAAOSGlobal-D-21-00042] [Medline: 34128913]
53. Minerva F, Giubilini A. Is AI the future of mental healthcare? Topoi (Dordr) 2023;42:809-817. [doi: 10.1007/s11245-023-09932-3] [Medline: 37361723]
54. Vaidyam AN, Wisniewski H, Halamka JD, Kashavan MS, Torous JB. Chatbots and conversational agents in mental health: a review of the psychiatric landscape. Can J Psychiatry 2019 Jul;64(7):456-464. [doi: 10.1177/0706743719828977] [Medline: 30897957]
55. Sadeh-Sharvit S, Hollon SD. Leveraging the power of nondisruptive technologies to optimize mental health treatment: case study. JMIR Ment Health 2020 Nov 26;7(11):e20646. [doi: 10.2196/20646] [Medline: 33242025]
56. Olawade DB, Wada OZ, Odetayo A, David-Olawade AC, Asaolu F, Eberhardt J. Enhancing mental health with artificial intelligence: current trends and future prospects. J Med Surg Public Health 2024 Aug;3:100099. [doi: 10.1016/j.glmedi.2024.100099]
57. Liu B. Sentiment Analysis and Opinion Mining: Morgan & Claypool Publishers; 2012. [doi: 10.1007/978-3-031-02145-9]
58. Zhang T, Schoene AM, Ji S, Ananiadou S. Natural language processing applied to mental illness detection: a narrative review. NPJ Digit Med 2022 Apr 8;5(1):46. [doi: 10.1038/s41746-022-00589-7] [Medline: 35396451]
59. Zunic A, Corcoran P, Spasic I. Sentiment analysis in health and well-being: systematic review. JMIR Med Inform 2020 Jan 28;8(1):e16023. [doi: 10.2196/16023] [Medline: 32012057]
60. Moyers TB, Manuel JK, Ernst D. Motivational interviewing treatment integrity (MITI) coding manual 4.1. : University of New Mexico; 2014 URL: https://motivationalinterviewing.org/sites/default/files/miti4_2.pdf [accessed 2025-08-27]
61. Lam RW, Michalak EE, Yatham LN. A new clinical rating scale for work absence and productivity: validation in patients with major depressive disorder. BMC Psychiatry 2009 Dec 3;9:78. [doi: 10.1186/1471-244X-9-78] [Medline: 19958540]
62. The Lam Employment Absence and Productivity Scale (LEAPS). : Department of Psychiatry, University of British Columbia URL: https://med-fom-psychiatry-wwd.sites.olt.ubc.ca/files/2012/07/LEAPS-description-and-scale.pdf [accessed 2025-08-27]
63. Measuring healthy days: population assessment of health-related quality of life. : Centers for Disease Control and Prevention; 2000 Nov URL: https://www.cdc.gov/hrqol/pdfs/mhd.pdf [accessed 2025-08-27]
64. Tossani E. The concept of mental pain. Psychother Psychosom 2013;82(2):67-73. [doi: 10.1159/000343003] [Medline: 23295405]
65. Fava GA, Tomba E, Brakemeier EL, et al. Mental pain as a transdiagnostic patient-reported outcome measure. Psychother Psychosom 2019 Nov 27;88(6):341-349. [doi: 10.1159/000504024]
66. Hutto C, Gilbert E. VADER: a parsimonious rule-based model for sentiment analysis of social media text. Proc Int AAAI Conf Web Soc Media 2014;8(1):216-225. [doi: 10.1609/icwsm.v8i1.14550]
67. NLTK. URL: https://www.nltk.org/ [accessed 2025-8-27]
68. Burton WN, Schultz AB, Chen C, Edington DW. The association of worker productivity and mental health: a review of the literature. Int J Workplace Health Manag 2008 Jun 27;1(2):78-94. [doi: 10.1108/17538350810893883]
69. Bond RR, Mulvenna MD, Potts C, O'Neill S, Ennis E, Torous J. Digital transformation of mental health services. npj Mental Health Res 2023;2(1). [doi: 10.1038/s44184-023-00033-y]

Abbreviations

AI: artificial intelligence
B2B: business to business
LDA: Latent Dirichlet Allocation
LEAPS: Lam Employment Absence and Productivity Scale
MITI: Motivational Interviewing Treatment Integrity
ML: machine learning
NLP: natural language processing
VADER: Valence Aware Dictionary and Sentiment Reasoner

Edited by A Mavragani; submitted 19.07.24; peer-reviewed by A Gabarda, Y Yang; revised version received 01.07.25; accepted 02.07.25; published 30.09.25.

Please cite as:
Wilbourne P, Mirch-Kretschmann S, Walker D, Varghese M, Arnetoli R
AI-Enabled, Text-Based Health Coaching and Navigation for Employees to Support Health Outcomes: Pre-Post Observational Study
JMIR Form Res 2025;9:e64553
URL: https://formative.jmir.org/2025/1/e64553
doi:10.2196/64553

XSL•FO
RenderX

© Paula Wilbourne, Susan Mirch-Kretschmann, Denise Walker, Michael Varghese, Roberto Arnetoli. Originally published in JMIR Formative Research (https://formative.jmir.org), 30.9.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Bedtime App–Guided Mindfulness Meditation in Patients With Insomnia: Mixed Methods Feasibility and Acceptability Pilot Study

Yan Ma[1], MPH, MD; Peter M Wayne[1], PhD; Janet M Mullington[2], PhD; Suzanne Bertisch[3], MPH, MD; Gloria Y Yeh[4], MPH, MD

[1]Osher Center for Integrative Health, Division of Preventive Medicine, Brigham and Women's Hospital, Harvard Medical School, 75 Francis St, Boston, MA, United States

[2]Department of Neurology, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, United States

[3]Division of Sleep and Circadian Disorders, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, United States

[4]Division of General Medicine, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, United States

Corresponding Author:
Yan Ma, MPH, MD
Osher Center for Integrative Health, Division of Preventive Medicine, Brigham and Women's Hospital, Harvard Medical School, 75 Francis St, Boston, MA, United States

Abstract

Background: While mindfulness meditation (MM) apps have gained popularity as a tool for promoting sleep, research focusing on bedtime mindfulness practice and app usage is limited.

Objective: As the first step toward understanding the efficacy and mechanisms of such bedtime practice and to inform future investigations, the goal of this pilot study was to explore the feasibility of app-guided bedtime MM practice with both in-lab and at-home physiological and self-report sleep remote assessments.

Methods: We conducted a single-arm, prospective mixed methods pilot study that included both standard in-lab sleep studies and remote at-home assessments of individuals with insomnia disorder with self-reported difficulty falling asleep. Participants practiced MM guided by a commercially available smartphone app at bedtime for 4 weeks. Pre-post assessments included a battery of sleep-related and psychological health questionnaires, objective physiological sleep measures (polysomnography and actigraphy), and daily sleep logs. We also conducted qualitative exit interviews to further assess feasibility and acceptability. Transcripts were analyzed for dominant themes using inductive and deductive qualitative methods.

Results: We recruited 13 participants with chronic insomnia (symptoms ≥3 nights weekly for ≥3 months) to complete the study protocol within 8 months (retention rate 77%). We were able to collect analyzable physiological and psychometric data with overall completion rates of more than 90%. The study was deemed feasible, meeting a priori benchmarks including recruitment, retention, completion, and adherence. The 10 participants retained in the program had excellent engagement (95% completion of in-lab studies, 100% completion of questionnaires, and 91% compliance with use of the app). Our preliminary analysis of subjective measures indicated improvement in sleep quality, insomnia severity, and presleep arousal, including Pittsburgh Sleep Quality Index change of −3.7 (95% CI −6.7 to −0.7), Insomnia Severity Index change of −4.5 (95% CI −7.7 to −1.4), Pre-Sleep Arousal Scale change of −7.7 (95% CI −13.1 to −2.3), and trend toward improvement in the Ford Insomnia Response to Stress Test indicated by a change of −2.5 (95% CI −5.9 to 0.9). From qualitative data, we identified domains that inform the feasibility and acceptability of the study, including (1) barriers to sleep prior to the study, (2) benefits and skills imparted by mindfulness, and (3) feedback on app use. Benefits and skills imparted by mindfulness included decreased catastrophizing, acceptance and nonreactivity, body awareness and relaxation, self-kindness, awareness of sleep hygiene and bedtime routine, earlier defusing of stress, increased focus and presence, and calm throughout the day.

Conclusions: Bedtime app-guided MM as an intervention in patients with insomnia and the hybrid study design with in-lab and at-home assessments are feasible and acceptable. This study informs the design of future clinical and mechanistic research examining app-guided MM to impact insomnia severity and presleep arousal.

Trial Registration: ClinicalTrials.gov NCT04242771; https://clinicaltrials.gov/study/NCT04242771

(JMIR Form Res 2025;9:e67366) doi:10.2196/67366

KEYWORDS

sleep; insomnia; mindfulness; meditation; smartphone; app; feasibility; remote delivery; bedtime; mixed methods

XSL•FO
RenderX

Introduction

Insomnia affects up to 30% of the population with a total public health cost of US $107 billion annually in the United States [1-4]. Insomnia disorder accounts for more than 5.5 million visits to physicians each year [5] and is associated with a wide range of deleterious health consequences including cognitive difficulties, poor work performance, and decreased quality of life, as well as increased risk of hypertension, diabetes, obesity, anxiety, depression, heart disease, stroke, and all-cause mortality [6,7]. Sleep medications can produce adverse effects and have risks, including fall risks, dependence, and potential abuse, which vary by the classes of medicine [8-11]. Clinical guidelines [12-14], in fact, recommend nonpharmacological therapy (cognitive behavioral therapy for insomnia) as first-line therapy (Grade 2B); however, this is underutilized and often not available. Even with emerging cognitive behavioral therapy for insomnia apps, limited availability, overgeneralized advice, usability, limited scope of interventions, insufficient clinical evidence, patients' engagement, and self-discipline are common issues. Results from the National Health Interview Survey showed that 49% of responders reported sleep troubles (n=26,742 adults), and mindfulness meditation was one of the most commonly used mind-body approaches [15]. Recent studies support mindfulness-based interventions as a promising nonpharmacological intervention for reliable and durable improvement of sleep quality, including among patients with insomnia disorder [16-20]. However, the existing evidence on sleep improvement is limited due to heterogeneity, lack of rigor across studies, and poor understanding of underlying mechanisms [19,21,22].

Models of insomnia often include cognitive arousal, hyperarousal, rumination, and worry. In clinical settings, however, many patients with insomnia commonly complain of inability to fall asleep due to "racing thoughts" [23-25]. This more colloquial term has been described as a subjective acceleration and overproduction of thoughts, which is highly associated with affective dysregulation (ie, cyclothymic temperament traits) and hyperarousal [26,27]. "Racing thoughts" has been shown to be increased in the evening and at bedtime in patients with sleep-onset insomnia. Importantly, "racing thoughts" at bedtime, but not rumination and worry, was associated with insomnia severity [28]. Under this high mental load, the sense of urgency about falling asleep adversely affects sleep onset latency [29]. Well-established evidence has shown that people with presleep hyperarousal are prone to difficulty initiating and maintaining sleep at night. Mindfulness meditation (MM) may be a promising nonpharmacological treatment for improving sleep quality by reducing bedtime stress and arousal [30], while barriers to adoption and scalability are also issues.

Technology-based approaches facilitate remote delivery of mindfulness-based interventions to be more affordable, accessible, and scalable. The increased popularity of mobile apps has the potential to increase access, assist home practice, and enhance care after formal in-person treatment has concluded. In practice, there has been a surge in the use of commercially available mindfulness-based smartphone apps as alternatives to formal interventions. In the case of mindfulness and insomnia, a survey from 12,151 paying subscribers of a commercially available mindfulness app Calm showed that 90% of subscribers have sleep difficulties, 77% of users download the app primarily for improving sleep, and the most common reasons for sleep disturbances were racing thoughts (7084/8604, 82%) and stress or anxiety (6307/8604, 73%) [31-33].

While MM apps have gained popularity as a tool for promoting sleep, there are no established guidelines on how best or when to use these tools. While the number of people who tend to use apps at bedtime is rapidly increasing [31,32], specific research focusing on bedtime app usage is limited, and the efficacy and mechanisms of such bedtime practice are unknown. In fact, bedtime screen use is often discouraged with typical sleep hygiene strategies. Many apps now target sleep difficulties and have begun to incorporate sleep-specific content but the research is deficient. Among the small number of clinical trials using remote-delivered mindfulness for sleep, there are several gaps. For example, most study the use of these programs during the daytime with no studies explicitly examining bedtime practice. Within this context, we tested the feasibility of a clinical trial involving an app-based bedtime MM intervention and both in-lab and at-home assessments. This pilot study lays the foundation for future studies to better understand bedtime app-guided MM and explore temporal correlations of bedtime practice and subsequent sleep quality.

Methods

Study Design

We used a single-arm prospective study design investigating a 4-week app-based bedtime mindfulness intervention in individuals with chronic insomnia disorder with self-reported difficulty falling asleep that included pre-post measurements, both standard in-lab and remote at-home assessments. We have followed reporting guidelines from the CONSORT (Consolidated Standards of Reporting Trials) pilot extension except for those criteria specific to randomized controlled trials (RCTs).

Ethical Considerations

This study was registered at ClinicalTrials.gov (NCT04242771). The Institutional Review Board at Beth Israel Deaconess Medical Center approved this study. Participants were provided with a copy of the consent form by email before the consent discussion. Initial screening and informed consent procedures were done remotely. Signatures were obtained using REDCap (Research Electronic Data Capture) e-signature, prior to the conduct of any research procedures. To guard against concerns of signature legitimacy, the signature will be obtained live during the consent appointment. A PDF of the electronically signed consent form was provided to the participant by email and was retained in the study research file. Only the primary investigator and a key co-investigator have access to participant information. All identifiers were transferred to a password-protected database. Our study database did not include data items that uniquely identify participants such as subject name, address, phone number, social security number and medical record number, but will be indexed by a study ID number. After all interviews were complete, the recordings were

XSL•FO
RenderX

transcribed verbatim and stored in a password-protected file. The transcriptions only had coded information without any participant identifiers. Participants were compensated up to US $200 for study participation (US $50 after the first visit and US $150 after completion of the study). Participants were also reimbursed for 1-year Calm app charges and parking expenses for each visit.

Eligibility Criteria

We included adult participants who met the following inclusion criteria: (1) age 20 - 50 years; (2) chronic insomnia (based on common key criteria from *Diagnostic and Statistical Manual of Mental Disorders* [Fifth Edition] [*DSM-V*] and International Classification of Sleep Disorders [ICSD-3]) with any complaints of difficulty initiating sleep, difficulty maintaining sleep, early morning awakening, and the frequency (≥3 nights per week) and duration (≥3 months) despite adequate sleep opportunity and an appropriate sleep environment, they had self-reported sleep onset latency >20 minutes; (3) speak and understand English; and (4) have a smart device for mobile app installation. Exclusion criteria included (1) those with sleep disorders other than insomnia (eg, narcolepsy, restless leg syndrome, insufficient sleep syndrome, jet lag, etc), (2) shift workers or having routine night shifts, (3) women who are pregnant or breastfeeding, (4) those with a history of head trauma or surgery, (5) those with regular (defined as twice a week or more) practice of mind-body interventions, (6) those with neurological disorders (eg, epilepsy, dementia, stroke, Parkinson disease, neuroinfections, brain tumors, etc), and (7) those with current use of neurologic, psychiatric, or cardiovascular medications that have expected neuropsychiatric or cardiovascular effects (ie, β-blockers, hypnotics, and antidepressants).

Recruitment and Consent

We recruited participants through fliers posted at local libraries and community centers in and around Boston and through ResearchMatch.org, a public-facing institutional review board–approved web-based recruitment tool. Interested individuals contacted us via email or phone. Eligibility was then assessed over a phone screen. Screening interviews were conducted by a research fellow trained in sleep medicine with both clinical and research experience working with patients with insomnia. Potential participants were explicitly informed that participation in the research study is voluntary. We obtained full informed consent during a consent appointment via phone, and signatures were obtained using REDCap e-signature in real time. A PDF of the electronically signed consent form was provided to the participant via email with REDCap secure link and code.

Interventions

Participants were asked to download a commonly used and commercially available mobile app (Calm), which was available for both Android and iOS systems. We instructed participants to practice MM using the specific sleep programming available within the app every night at bedtime for 4 weeks (weeks 1 - 4). In the first week of intervention (week 1), participants followed an introductory program, “Seven Days of Sleep,” which included 7 soundtracks (12 - 15 minutes each) for guided MM practice. This practice involves techniques such as breathing exercises, mental imagery, awareness of body and mind, and muscle and body relaxation. During weeks 2 - 4, participants repeated the same sessions or chose their preferred programs under “Meditation” within the “Sleep” section. Participants could choose from a list of 28 programs with a total of more than a hundred audio tracks of various time lengths (from about 5 minutes to more than 1 hour). Participants were instructed to practice a minimum of 10 minutes but no more than 30 minutes at bedtime. We also provided a series of instructions to minimize the negative impacts from light exposure and the use of digital devices (eg, dim light on the screen, auto-stop playing after a set time duration, no-disturb mode, audio only without screen exposure, etc). There were no limitations for daytime practice or for use of other meditations available on the app.

Questionnaires

Self-report measures were collected during participants’ baseline and postintervention visits. The battery of sleep-related measures included the following.

Pittsburgh Sleep Quality Index (PSQI) is a subjective sleep assessment, which includes multiple sleep-related variables over the preceding month. The PSQI yields 7 component scores: subjective sleep quality, sleep latency, sleep duration, habitual sleep efficiency, sleep disturbances, sleep medication, and daytime dysfunction. Component scores range from 0 to 3 and are summed to obtain a global score, which ranges from 0 to 21. A score of >5 suggests poor sleep quality, and higher scores suggest greater sleep disturbance [34].

Insomnia Severity Index (ISI) is a brief instrument that assesses insomnia according to the criteria from the *Diagnostic and Statistical Manual of Mental Disorders* (Fourth Edition) (*DSM-IV*) and the International Classification of Sleep Disorders. Its reliability, validity, and sensitivity to treatment response have been documented in the general population and with patients with insomnia [35,36]. A 5-point Likert scale is used to rate each item (eg, 0=no problem; 4=very severe problem), yielding a total score ranging from 0 to 28. The total score is interpreted as follows: absence of insomnia (0 - 7), subthreshold insomnia (8-14), moderate insomnia (15-21), and severe insomnia (22-28).

Ford Insomnia Response to Stress Test (FIRST) questionnaire is a 9-item self-report measure of trait vulnerability to sleep reactivity [37,38]. The FIRST has shown acceptable reliability and validity, is predictive of future sleep dysfunction, and has been related to objective measures of sleep reactivity. The total score ranges from 9 to 36. High scores on the FIRST indicate greater vulnerability to sleep disruption.

Pre-Sleep Arousal Scale (PSAS) is a widely used measure to evaluate presleep arousal [39]. It is a 16-item scale that measures 2 distinct domain subscores of cognitive and somatic arousal before falling asleep (range 8 - 40, higher score indicates greater arousal).

Dysfunctional Beliefs and Attitudes about Sleep (DBAS) was developed to evaluate sleep-disruptive cognitions and purported maladaptive beliefs in insomnia. The validated brief version (DBAS-16) has a 10-point Likert scale ranging from 0 to 10.

XSL•FO
RenderX

For each of the 16 beliefs, the number corresponding to the degree of belief (ie, 0=strongly disagree, 10=strongly agree) is circled. The DBAS-16 index score is a mean item score (ie, item scores are summed and divided by 16), and a higher score reflects greater dysfunctional beliefs about sleep [40]. DBAS-16 has been shown sensitive to several indices of changes with insomnia treatment and has been found reliable for discriminating between self-defined good and poor sleepers in both younger and older adults.

Generalized Anxiety Disorder 7-item Scale is a widely used diagnostic self-report scale for screening, diagnosis, and severity assessment of anxiety disorder. It measures anxiety symptom severity as a predictor of anxiety disorders (range 0 - 21, higher score worse), 0 - 4 indicating minimal anxiety, 5 - 9 indicating mild anxiety, 10 - 14 indicating moderate anxiety, and scores greater than 15 indicating severe anxiety [41,42].

Patient Health Questionnaire–9 (Depression Scale) is a validated measure of depressive symptoms. The 9 items are based directly on the 9 diagnostic criteria for major depressive disorder in the *DSM-IV*, with ratings from “0” (not at all) to “3” (nearly every day). The total score ranges from 0 to 27 and suggests the level of depression: 0 - 4 minimal, 5 - 9 mild, 10 - 14 moderate, 15 - 19 moderately severe, and 20 - 27 severe depression [43,44].

Mindful Attention Awareness Scale (MAAS) is a commonly measured scale for measuring mindfulness in positive psychology. The MAAS measures an individual’s tendency to enter a state of mindfulness through the individual’s frequency of having certain experiences related to mindfulness and mindlessness. It includes 15 statements that respondents rate on how frequently they engage in the activities described, on a scale from 1=almost always to 6=almost never. Scores on these statements are then combined to create an overall score of mindlessness or mindfulness, with higher agreement indicating a lesser tendency to enter into a mindful state, while a lower score indicates a greater tendency toward mindfulness.

Daily Logs

After the baseline visit, participants were asked to complete a sleep diary and MM practice log each morning to recall the previous night. The logs were completed daily during the 4 weeks of intervention. The sleep logs documented time to bed and waking up, total sleep time, number of overnight awakenings, overall score of sleep quality, and stimulant intake (eg, any coffee, alcohol, etc). MM practice adherence during the 4-week intervention period was tracked with a daily log format similar to the sleep logs, including frequency and duration of practice, strategies used (eg, breathing exercise, body scan, etc), and the names of the selected app programs or recordings.

Qualitative Interview

All participants were interviewed postintervention using a semistructured interview to gather information about each participant’s experience with sleep prior to and during the study, as well as study and intervention feedback. Questions included both simple response (eg, Yes/No) questions focused on app acceptability and open-ended questions on feasibility or acceptability, barriers or preferences to practice, specific app feedback, and overall study experience. We further explored helpful components of the intervention and their relation to sleep. To ensure consistency, all the interviews were conducted by a single research team member (YM) who was trained in qualitative research and conducting interviews. Interviews were audio-recorded and transcribed verbatim. Two reviewers (YM and GY) independently reviewed the transcripts. Responses to Yes/No questions are reported as frequencies. For open-ended responses, we used a combined approach with both inductive and deductive methods to describe the data and extract emergent themes. With the deductive (theory-driven) approach, we used existing frameworks and hypotheses about how the intervention might help patients to seek support for specific themes within the data. With the inductive (data-driven) approach, we also allowed completely new patterns and themes to emerge from the data de novo.

Objective Sleep Assessments

As part of our hybrid design, we conducted both in-lab and at-home remote objective sleep assessments. This included (1) standard, attended in-lab polysomnography (Natus SleepWorks; Natus Medical Incorporated) at baseline and then postintervention (week 5) at the Clinical Research Center at Beth Israel Deaconess Medical Center; (2) ambulatory heart rate monitoring 1 night per week on a typical night to include bedtime routine, MM practice, and overnight during sleep; and (3) actigraphy (Actiwatch 2s; Respironics Inc) for 2 consecutive weeks starting at the baseline visit, week 0 as baseline assessment, and week 1 as the first week of intervention. Participants were instructed to wear the watch continuously on the nondominant wrist. To minimize discrepancies between sleep log and actigraphy data, participants were asked to push the event marker just prior to falling asleep and upon first awaking every morning. All raw data of physiological signals were extracted. For these objective sleep assessments, this manuscript focuses only on feasibility and adherence metrics. Given the scope and depth of sleep physiological signal analyses, additional findings are planned in a separate manuscript that explores the mechanism of bedtime MM and association between physiological changes and subsequent night sleep quality.

Feasibility and Acceptability Measures

This is a proof-of-concept pilot feasibility study. Primary analysis of data was limited to descriptive feasibility statistics. Assessment completion rate was defined as the number of valid data recordings divided by the total planned nights, that is, 5 nights of heart rate recording (1 in each week) and 14 nights of actigraph data. Compliance rate was calculated as the total number of nights of practice during the 4 weeks divided by 28. We hypothesized that (1) we would be able to recruit and enroll 12 patients with insomnia within 12 months, (2) we would retain >75% of participants in the program, (3) they would exhibit >70% compliance with use of the app, (4) participants would find the intervention acceptable based on qualitative interviews, and (5) we would be able to collect clean analyzable physiological and psychometric data from at least 80% of the participants.

XSL•FO
RenderX

Statistical Analyses

Per guidelines for pilot trials set forth by National Center for Complementary and Integrative Health and others [45], we did not power this study for testing intervention efficacy or pre-post differences and thus emphasize descriptive statistics and data visuals here. The planned target of 12 participants was based on study needs and pragmatics (including time and budget considerations). Based on other pilot studies our group and others have conducted, this is a sufficient sample to evaluate feasibility metrics [46]. Statistical analyses were performed using IBM SPSS Statistics (version 25; IBM Corp). Continuous data are reported in text and tables with means and SDs for normally distributed data and with medians and IQRs for skewed data. Discrete data are reported as absolute numbers and percentages.

Results

Feasibility Outcomes

Recruitment started in December 2022, and data collection was completed in July 2023 (Figure 1). During the 8 months, we screened 37 interested individuals and were able to identify 13 eligible participants who consented to the study. The resulting consent-to-screening ratio was 13/37 (35%). Of these 13, 1 participant withdrew prior to any study procedures or testing (changed her mind regarding overnight sleep study). Two additional participants withdrew due to personal reasons (unemployment and schedule changes) prior to the start of intervention (with incomplete and complete baseline assessments, respectively). Ten participants completed all procedures including intervention and assessments (overall retention of 10/13, 77%).

Figure 1. Modified CONSORT (Consolidated Standards of Reporting Trials) flowchart (single-arm nonrandomized study).

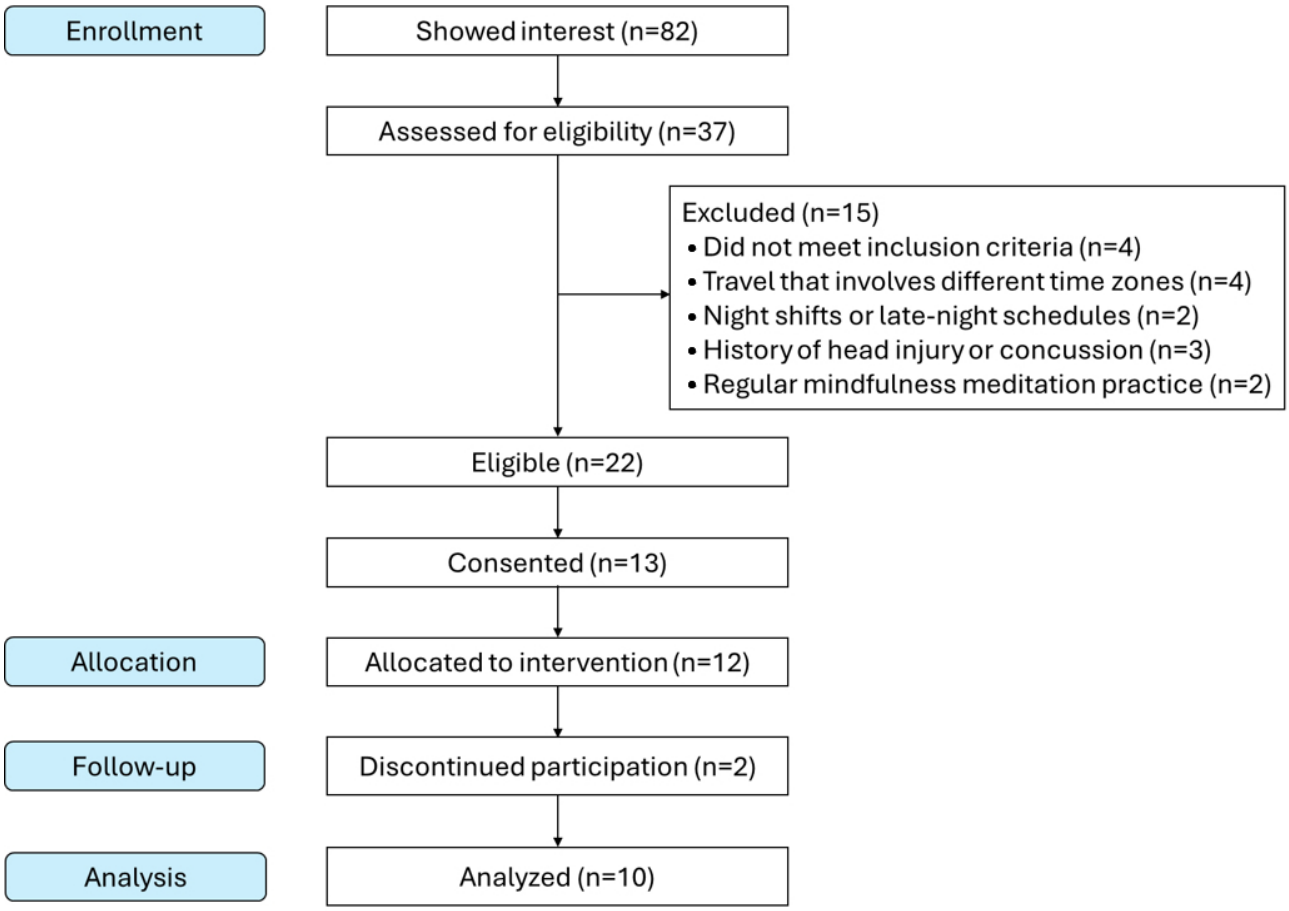

Demographic Characteristics

Full baseline demographic characteristics are shown in Table 1. The 12 participants for whom we have baseline data ranged in age from 22 to 50 years (mean age 30.42, SD 7.77years). There were no differences between those who enrolled (n=12) and those who completed the study (n=10) with respect to demographic characteristics (Table 1). The final sample of participants who completed the study were more women (7/10, 70%), relatively well educated, having insomnia for more than a year (7/10, 70%), full-time employed (7/10, 70%), and single (7/10, 70%).

XSL•FO
RenderX

Table . Baseline demographic characteristics of participants with chronic insomnia[a].

Characteristics	All enrolled (n=12)	All retained and completed (n=10)
Age (years)	30.42 (7.77)	30.46 (8.15)
BMI	22.60 (2.35)	22.49 (2.45)
Sex, n (%)		
Female	7 (58.3)	7 (70.0)
Male	5 (41.7)	3 (30.0)
Race or ethnicity, n (%)		
White	5 (41.7)	4 (40.0)
Asian	6 (50.0)	5 (50.0)
Other	1 (8.3)	1 (10.0)
Education levels, n (%)		
Bachelor's degree or below	7 (58.3)	6 (60.0)
Master's degree or above	5 (41.7)	4 (40.0)
Insomnia duration, n (%)		
9 months to 1 year	3 (25.0)	3 (30.0)
<5 years	5 (41.7)	3 (30.0)
>5 years	4 (33.3)	4 (40.0)
Employment status, n (%)		
Full-time	7 (58.3)	7 (70.0)
Part-time	5 (41.7)	3 (30.0)
Marital status, n (%)		
Single	9 (75.0)	7 (70.0)
Married	3 (25.0)	3 (30.0)
Subjective sleep measures, mean (SD)		
Usual sleep-onset latency (minutes)	N/A[b]	78.75 (48.1)
Usual sleep duration (hours)	N/A	363.0 (43.5)
Objective sleep measures		
Sleep-onset latency (minutes), mean (SD)	N/A	44.0 (44.2)
Total sleep time (minutes), mean (SD)	N/A	357.8 (80.9)
Sleep efficiency, n (%)	N/A	80.5 (7.7)
Wake after sleep onset (minutes), mean (SD)	N/A	90.3 (48.7)

[a]Continuous data (age and BMI) are presented as mean and SD of the mean. Categorical data are presented as n (%).

[b]N/A: not applicable.

Participants reported an average of 4.3 nights per week with difficulty falling asleep and 4 mornings per week waking up early and still feeling tired or unrested. At baseline, on average, participants reported 101 minutes to fall asleep, 92 minutes staying awake at night, and 24.5 minutes to fall back to sleep after awakening at night. Most of the participants (7/10) had tried melatonin for sleep but did not find it helpful.

At baseline, all participants had poor sleep quality, indicated by PSQI scores ranging from 7 to 16. Insomnia severity, indicated by the ISI score, ranged from 8 to 22, with 6 (60%) participants having subthreshold insomnia, 3 (30%) participants having moderate insomnia, and 1 (10%) participant having severe insomnia. Based on the Generalized Anxiety Disorder 7-item Scale scores (ranged 3 - 8 at baseline), participants had no (4/10, 40%) or mild anxiety (6/10, 60%). The Patient Health Questionnaire–9 scores (ranged 3 - 8 at baseline) indicated that 3 (30%) participants had minimal depression and 7 (70%) participants had mild depression.

XSL•FO
RenderX

Assessments Completion and Intervention Compliance Rates

All the 12 participants completed baseline in-lab sleep study. The 10 participants retained in the program exhibit an overall 95% completion of in-lab studies, 100% completion of questionnaires, and 91.4% compliance with use of the app (weekly compliance analysis shown in Table 2). On average, the completion rate of daily sleep logs was 100% at baseline week and 89.6% during the 4 weeks of intervention. Remote data collection by actigraph device and heart rate monitor showed overall 100% and 92% completion rates, respectively. At the end of the study, 9 of 10 participants completed their second in-lab study. One participant missed his postintervention in-lab sleep study due to extended travel out of the state. All of the 10 participants completed pre- and postquestionnaires and exit interviews.

Table . Assessment completion rates by study week.

Variables	Baseline (week 0)	Week 1	Week 2	Week 3	Week 4	Postintervention
In-lab sleep study	100%	N/A[a]	N/A	N/A	N/A	90%
Questionnaires	100%	N/A	N/A	N/A	N/A	100%
Bedtime practice	N/A	95.7%	90%	90%	90%	N/A
Sleep logs	100%	100%	100%	100%	100%	N/A
Heart rate monitoring	90%	90%	90%	90%	100%	N/A
Actigraphy	100%	100%	N/A	N/A	N/A	N/A
Exit interview	N/A	N/A	N/A	N/A	N/A	100%

[a]N/A: not applicable.

Table 2 shows weekly participant completion rate of study assessments. In-lab sleep study and questionnaires were conducted at baseline and postintervention. Sleep logs and heart rate monitoring were conducted during the 4-week intervention. The actigraphy watch was to be worn for baseline week 0 and week 1 of the intervention. An exit interview was conducted postintervention.

Simple Response Feedback on Acceptability

Regarding the intervention, 70% (7/10) of the participants reported no difficulties to engage in practice, 90% (9/10) of the participants reported that the app was easy to use, and all of them reported that the audios were easy to follow. Regarding the preferred practice time, most of the participants (8/10) preferred shorter practice (less than 20 minutes, with 10 - 15 minutes most favored). Most of the participants would like to keep using the app or to practice what they learned at bedtime after they finished the study (Multimedia Appendix 1). The bar graphs present the percentage of participants' simple answer responses on individual acceptability questions during the qualitative exit interview.

Open-Ended Qualitative Data

All participants responded to all questions during the exit interviews. We identified multiple themes within 3 major domains. Textboxes 1-3 show representative quotes from participants according to emergent themes. Multiple quotes within each category are from unique individuals.

XSL•FO
RenderX

Textbox 1. Representative quotes within qualitative domain 1.

Domain 1: Thoughts or focus of attention during the presleep period as a barrier to sleep

Theme 1.1. Racing, uncontrolled thoughts

"I didn't have much control over my thoughts. I have an issue where my thoughts would be just coming and passing through my head. I would feel, for example, very sleepy before I go to bed. But then, when I would go to bed, all of a sudden, I would feel awake and then I just kept having racing thoughts, for a very long time to be about work. But then, I would have a lot of random things that would just keep popping in my head."

"...I would say sometimes I get carried away with thinking about different things. So then it ends up taking longer to fall asleep in general."

"Just like wandering minds, wandering thoughts...My attention is like wandering very quickly through a lot of different thoughts...With all of those different thoughts, I just can't get myself to fall asleep."

Theme 1.2. Loss of control leading to increased stress

"...[I] would think about all the things that happened during the day and would also think about all the things I need to do tomorrow and sometimes I would like to take out my phone and write more things down of like, oh, I'm gonna do this tomorrow. I created to-do things like that...I sometimes try to force myself to not think about anything...but I don't think I was very successful...sometimes it would get to a point where if I'm thinking too much about something and still awake then that itself becomes stressful."

"Focus my attention on my phone...My mind definitely goes through stressful things. I definitely think about work, I think about life situations, and I do definitely get stressed...Sometimes I try to change my mind to make it focus on other stuff. The other times I would just let it run wild and free and let myself worry about those things. I 100% think that where my mind goes kind of dictates how I sleep."

Theme 1.3. Worry, rumination

"If I'm worrying about something that happens during the day or I'm depressed by something that happens during the day, or I'm worried about something that's going to happen in the next day, I would find it difficult to fall asleep on that night."

"...once I go to sleep, I start worrying about not sleeping early and how that will impact the following day. And that sort of makes my heart rate of my brain to be more active."

XSL•FO
RenderX

Textbox 2. Representative quotes within qualitative domain 2.

Domain 2: Benefits and skills imparted by mindfulness

Category 2A: Mindfulness specifically at bedtime for sleep

Theme 2A.1. Decreased catastrophizing

"So it made me feel that I shouldn't be anxious, or wondering, if I'm like, 'oh, I have to go to sleep or am I going to sleep well or not?' I just learned that I should let it be. And that's like I'll be fine the next day no matter what. Versus how I felt before, like where I thought that like, 'oh, if I don't sleep well, I'm not going to perform well at work. I'm going to be fatigued. I'm going to feel like this my whole day is going to be ruined.' So that's only my perspective about my sleep, sleeping pattern, or just like if I sleep good or not, I just know that I'll be good the next day."

Theme 2A.2. Acceptance, nonreactivity, decreased anxiety

" I don't have to wait any longer. I can just turn off my lights and turn on the meditation app and it doesn't matter when I'm going to fall asleep...Instead of like, 'oh, I need to go and sleep.' It is just like I'll go to bed and I'll meditate and I'll see where that's going to get me. If I'm kind of stressed, I would start feeling more relaxed about thinking about my bedtime."

"Well, I'm trying to tell myself that it's like, you know, it's OK if I don't fall asleep right away. And like if I'm resting, then that's still good. So I'm like, that's something I'm trying to do."

"The app was teaching [me] not worrying about sleep per se. I don't know how to explain this one, but it was just teaching that you shouldn't really worry about sleep. You should just be lying on your bed and the sleep would come naturally if you're not worried about it. And it just gave a lot of practical advice about how to deal with, I don't know, sleep anxiety is a term. But that feeling of me not having to force myself to sleep and that I can be on bed until I naturally fall to sleep, it seems to be very helpful in like putting me to sleep earlier."

"I think some of the mindfulness meditation is gonna actually help me find my relationship with sleep in terms of something that you don't have to continuously struggle or grapple with...Okay, if I fall asleep then okay. If I don't then that's okay too. But try to find ways to settle yourself, settle your mind, think about it."

" I definitely found that I do fall asleep better, fall asleep easier. And even if I wake up in the middle of the night, I feel less anxious about the next day. So overall, my anxiety about not sleeping well is way, way, way less...and I think that quality of my sleep got improved."

Theme 2A.3. Body awareness and relaxation

"...and it definitely helped me out, the meditation helped me out just by relaxing my body a lot. And that's something, that is already making you feel good when you know that you can really reduce the stress from your body that has been collected maybe in certain parts of your body."

"The breathing techniques that were taught in the app, I felt they were very helpful and they gave an instant feeling of relaxation...And then there was another aspect of trying to focus attention towards specific muscles or specific parts of the body. And I think that's also very, very helpful...That just somehow made me feel relaxed. I don't know how, but it made me feel more relaxed and more likely to fall asleep."

"I think some of the directions given around breath work are quite helpful. Like taking deep breaths, holding them longer and then releasing."

"...it [the practice] definitely made me more aware of sort of where I was holding unnecessary tension. Yeah. And so sometimes it would make me notice that I wasn't relaxed and then I relax."

"...But focusing on the body, I think it does help because sometimes when they say specifically like relaxing your feet, relaxing your legs, and I would switch my focus to those parts of the body and then try to relax them."

Theme 2A.4. Self-kindness

"...And I think just being told like, it's some of the common things being common and how like normalizing certain things but then also not like reinforcing certain behaviors but kind of like giving some grace and peace so that you can move beyond some of these normal behaviors that are not great for sleep hygiene."

"...That's something that if you try to force it more and more then it becomes harder to fall asleep. So you just, kind of, you have a little bit of peace and face to yourself..."

Theme 2A.5. Focus on the present

"...for example, when I started a new job, I was very, very focused about the next day at work. But that started being less and less throughout the time. So like I was thinking less about work. But I think that the same anxiety pattern played in my head that I think is where my brain kind of had that pattern, like, 'oh, we go to bed. It's time to start worrying about something. It's time to start worrying about job tomorrow.' Even though my job was not like stressful anymore, my brain would look up for something else to be worried about. So that's, I think, that is the part that I am now able to change with the mindfulness practice."

"...as time goes on these days, I don't think about stressful things as much at night before I sleep. I think my coping strategy is, oh, you did enough today. The rest is tomorrow's problem. And I'll just battle the issues tomorrow. So like, I would probably briefly think about it. But then I Yeah, but then I would tell myself, okay, we'll think about tomorrow. Let's sleep."

"I think over four or five weeks, [I was] definitely just paying more attention to sleep habits and [my thoughts were like] 'don't do this, this can wait till tomorrow.'"

Category 2B: Mindfulness beyond the presleep period impacting sleep

Theme 2B.1. More awareness of sleep habits and sleep hygiene

XSL•FO
RenderX

"...So it definitely has made a difference in terms of sleep habits, too. No more caffeine after 2 p.m. or 2 p.m. to 3 p.m. No more caffeine, like coffee, green tea. So that's been a big difference there, too. Yeah, I think, in general, I think the mindfulness really helped just relaxing with the music. And so it definitely has improved...so I'm always more conscious and mindful about my sleep habits. And so, yeah, it's made a difference during those five weeks compared to before."

"It makes me think that sleep is, it is a lifestyle. Sleep, a good quality sleep, not just the quantity of, you know, how many hours, but the quality of sleep, which is going to affect your other lifestyle."

Theme 2B.2. Improved bedtime routine

"...also I mentioned about just managing my time better...[I] make sure I brush my teeth, wash my face, do this, do that. So it changes my routine to do that earlier, to not do too many things at night, to also turn off the phone and the laptop, for example, don't check emails, don't do this, don't do that on the computer...It definitely has changed my routine. It is also, a lot less eating, obviously, later in the night. So those are the routines that have changed."

"So one thing I also wanted to mention is I started to have a more structured and calm approach towards, you know, the end of the day, an hour or two before I sleep...So I go to a different world where I'm reading about something completely different, and then by the time I come to sleep, it seems easier."

Theme 2B.3. To defuse earlier, not let stress build up to impact sleep at night

"...part of the course did focus on how some of the stresses that happened during the day could be carried over during the night, whether we know it or not. So what I began to do is, and this was also suggested in the course, is to keep an eye on your own, on your own psychological health, let's say during the day. So if you're getting too worked up on something or too stressed, it's nice to kind of go away, relax a little bit, so that the stresses don't accumulate on top of each other and cause an issue during the night. So it's always good to refresh or empty it out or flush it out, I think, as the course didn't use the exact word, but just flush it out before you come back, you know, to working on the same thing, be it work or part of your personal life. So I'm kind of keeping an eye on that at the back of my mind."

Category 2C: Global effects of mindfulness

Theme 2C.1. Increased focus, presence

"I noticed that [the practice] not only affected my focus on my body or breathing during the meditation, but it also enhanced my focus during the day. I noticed that part as well that I am more aware of myself, of the reality of my body during the day. And I can control [my thoughts] easily. If I have a little anxiety going on or anything like that, I'm more aware of being in tune with my body."

"I want to be consistent with the mindfulness practice because I do see a lot of benefits already from it,...my overall awareness did improve a lot. So I can say that my focus not only on work is better, but like focus on everything that I'm doing consciously. It's like there's more unconscious stuff that I'm doing. It's like I can be part of more in tune in my everyday life."

"...I think that maybe even a few minutes of doing that would already change a lot of things throughout my day. So I just like shorter mindfulness practices during the day, I started changing. I think I've got more control over many things and then became more aware in everyday situations."

Theme 2C.2. Calm throughout the day

"The experience itself was quite positive because I think it like brings more like relaxation and like calm to like a certain time of the day and I think it can like last the whole rest of the day or at least last a couple of hours after like doing a little bit of work on this. So even like a couple of minutes, I've done it."

"...So I think the mindfulness technique helps, just to take more deep breaths, you know, to really even closing your eyes, like listening to midday music. So it definitely, it definitely has helped me, how to just be more, more focused on just, you know, just like, okay, you know, just to really practicing more mindfulness, calming down, reducing the level of stress during the day."

Theme 2C.3. Incorporating mindfulness in activities during day

"The Calm app has definitely helped focusing on mindfulness, relaxing more with the mindfulness. I think it also helps even outside of the Calm app, because I have practiced during the day at certain times...I think I practice it more, the mindfulness, I even listen to it during the daytime,...try to combine the mindfulness music with exercise, such as walking. I think it also helps because at night, it makes me more relaxed, you know, just listening to the Calm app."

"I think it [mindfulness meditation] allowed me to also slow things down during the day to be more present and to be more mindful about what's happening in front of me."

XSL•FO
RenderX

Textbox 3. Representative quotes within qualitative domain 3.

Domain 3: Feedback on app use

Category 3A: Preferences and likes

Theme 3A.1. App content

"There are some good things that they talk about, you know, how to reduce stress and anxiety, how to cope with whatever stressful things in life, and how to fall asleep faster, how to calm yourself down. So these are some really good things."

"I do enjoy it [the app] and I'm definitely going to keep using it. I think that I'll be using it more."

Theme 3A.2. Multiple options

"I think what I liked is that there was a breadth of content and there's multiple different narrators talking about multiple different things. So there's that variability that can be exercise, maybe to overcome any feeling of, you know, boredom."

"I think I like the app because there are a lot of options for me to try for the mindfulness practice. And as I was trying different meditations, I also noticed that some meditation might not be very helpful for me and some others are a lot more helpful, I think more useful to me."

"Overall, I love that I could choose actually, like, every night, a different meditation."

Theme 3A.3. Diverse voices

"I love the diversity in terms of voices...The app that I had before, I think it was like, only one guy who would take you through meditation. And it could be the same way every time. But this one, [with audios] from different people, you would get different experiences and they would start on a different way and finish on different way or tell you different things. So, you can always learn something new from each one of them. So that's amazing."

Theme 3A.4. Varied preferences on the length of soundtracks

"I love the app very much. The only thing, maybe I would add a little bit more options for sleep, maybe shorter options."

"...Like this is just a five minute meditation rather than a full course, which would be, you know, in 10 installments or something...I think five to 10 minutes min is preferred for me."

"So I liked the audios themselves...I liked the 'Seven Days of Sleep' because I feel like they were a good length for me. I feel like around 15 minutes min, maybe 20, was like my favorite length. But I do also feel like a lot of their meditation are longer, like 30 or 40 minutes min, and I wish there were more variety of the ones that were a little bit on the shorter side. Just so I could have tried those out as well..."

"I guess one of the other things I liked is that you could kind of choose how much time you wanted to put into it. And I tried a bunch of different ones...I did like that it was a little shorter. Just because, I mean, the meditation was sort of adding a step into my established bedtime routine. And I didn't really like doing the super long ones. But it was nice that you could kind of choose."

"There's the vast majority of them are like 30 minutes min at least. And I wish I could make it 20 minutes min instead of 30 or 40."

"I like it to be longer so that I just doze off to sleep while I'm listening to it."

"...Sometimes it takes me longer to fall asleep and I need longer meditation and I never know how long I need and how short I need."

Category 3B: Challenges and suggestions for improvement

Theme 3B.1. Areas for app improvement

"...I would say that maybe I would sometimes expect more breath work at the beginning of certain meditation..."

"Sometimes...[the practice] would bring me to the point where I can just feel great body sensations and slowly start drifting into sleep. But then I would hear a voice again and I would snap out of sleep."

"I feel like the user interface was fine, but it felt a little bit cluttered sometimes when you're like first at the homepage. I feel like they make the sleep tab more relaxed looking, but I feel like when you first open it up, it's like, okay, there's all these different things to look at, and it's a little bit annoying to have to sort of like, okay, first discover and then sleep and then scroll through to find what you're looking for."

"Fairly easy to use if you know which program you're going directly for, I would have liked it to be easier to find individual programs."

"I think the main thing that I did find to be distracting was when they would take long breaks in the audio where it would, you know, say, okay, now we're going to leave you to focus on that practice. And sometimes if you were in a sleepy state, you wouldn't know whether or not it had ended. And then they would start talking again and you would kind of, you know, wake up a little bit. So I found that distracting as opposed to if it had a background music or something to let you know, still going."

Theme 3B.2. Initial investment for successful app engagement

"I would say it depends on which practice that you choose. Like some of them, perhaps I don't like that voice that much, or the content that I just did not enjoy that much. And then yeah, at that time when I chose those, I would feel like it's hard to follow. But after like two or three weeks after I eie,xplore some, [a good] amounts, I found some of those that I think it's really helpful for me. So then I would just use those repeatedly."

XSL•FO
RenderX

"I would say...before using the Calm app, maybe someone who has experienced using Calm app can kind of talk...more about the Calm app,...maybe someone who has experienced using it recently and how that has helped him or her or them, I think would be good before using the app. I think overall I've been very, very satisfied. One thing I can think about is maybe talk more about the benefits of the Calm app, how to use it. Because a lot of it seems to be, you have to try it on your own. You have to, you know, really learn more on your own. I think it would help to have someone kind of talk about it before starting using it. To have a session, have a mini session maybe, [from] someone with experience using it."

"...probably if I wasn't obligated to do that [mindfulness meditation], I might have given up in the beginning because it still is an effort. But once you get it into the routine and that barrier is no past, then it's easy."

"I think it was difficult to start. I think it got easier over time because it became like a routine, became something that was just part of it, but initially, the first week I think was [difficult]."

Theme 3B.3. Adherence or maintaining engagement

"At the beginning, no [not difficulty to follow], because it was something new. And, you know, my brain was focused on it and I would listen to every word of it and implement it. But over the, you know, the coming weeks, particularly in the last week, because I know exactly what type of routine I'm going to go through. I don't know, I might be feeling a little bit bored or a little bit, you know, so I'm not so engaged. And I did find my mind started to wander around during the mindfulness practice. So it wasn't, it wasn't so easy in the last, in the last week. In the first week, it was much easier."

"...the difficult part is when it's so easy to follow and I know what the presenter, even though it's a new presenter or a new narration, I kind of have an idea of where they're heading towards in terms of the breathing and then focusing on different parts of the body. So it's nice. It's easy to follow, but because it's so easy and I know what's coming, it might get boring. And if it gets boring, my mind will wander around."

The first domain is Thoughts or Focus of Attention During the Presleep Period as a Barrier to Sleep. Themes within this domain included the following: 1.1. Racing, uncontrolled thoughts; 1.2. Loss of control leading to increased stress; and 1.3. Worry, rumination (Textbox 1).

The second domain focuses on Benefits and Skills Imparted by Mindfulness. Within this domain, there are 3 categories (Textbox 2). In the first category (2A) Mindfulness Specifically at Bedtime for Sleep, themes included the following: 2A.1. Decreased catastrophizing; 2A.2. Acceptance, nonreactivity, and decreased anxiety; 2A.3. Body awareness and relaxation; 2A.4. Self-kindness; and 2A.5. Focus on the present. Within the second category (2B) Mindfulness Beyond the Presleep Period Impacting Sleep, themes included the following: 2B.1. More awareness of sleep habits and sleep hygiene, 2B.2. Improved bedtime routine, and 2B.3. Defusing stress earlier in the day to avoid impacting sleep. Within the third category (2C) Global Effects of Mindfulness, themes included the following: 2C.1. Increased focus, presence, 2C.2. Calm throughout the day, and 2C.3. Incorporating mindfulness in activities during the day.

Within the third domain Feedback on App Use, there were 2 categories (Textbox 3). In the first category (3A) Preferences and Likes, themes included the following: 3A.1. App content, 3A.2. Multiple options, 3A.3. Diverse voices, and 3A.4. Varied preferences on the length of soundtracks. Within the second category (3B) Challenges and Suggestions for Improvement, themes included the following: 3B.1. Areas for app improvement, 3B.2. Initial investment for successful app engagement, and 3B.3. Adherence or maintaining engagement.

Exploratory Results From Self-Reported Outcomes

Based on the self-reported data collected at baseline and postintervention, we observed overall improvement on most of the questionnaire-based outcomes (Multimedia Appendix 1 and Table 3). PSQI scores decreased from 10.6 to 6.9, with a postintervention change of −3.7 (95% CI −6.7 to −0.7), and ISI scores decreased from 14.4 to 9.9, with a postintervention change of −4.5 (95% CI −7.7 to −1.4). Improvement in presleep arousal was indicated by the postintervention change of −7.7 in PSAS scores (95% CI −13.1 to −2.3). Mean scores on anxiety and depression were also decreased, but MAAS scores remained relatively unchanged at the end of the study (Figure 2 and Table 3).

XSL•FO
RenderX

Table . Group-based descriptive statistics of self-reported outcomes at each visit and pre-post changes.

Variables	Baseline (visit 1)		Postintervention (visit 2)		Pre-post changes	
	Mean	SD	Mean	SD	Mean	95% CI
PSQI[a]	10.6	2.6	6.9	2.3	−3.7	−6.7 to −0.7
ISI[b]	14.4	3.8	9.9	4.3	−4.5	−7.7 to −1.4
FIRST[c]	22.2	5.7	19.7	6.4	−2.5	−5.9 to 0.9
PSAS[d]	39.8	7.4	32.1	6.0	−7.7	−13.1 to −2.3
Somatic	13.8	4.2	10.5	2.1	−3.3	−6.0 to −0.6
Cognitive	26.0	4.2	21.6	5.1	−4.4	−7.5 to −1.3
DBAS-16[e]	5.6	1.9	4.4	2.0	−1.1	−2.3 to 0
GAD-7[f]	5.0	1.8	3.5	2.0	−1.5	−3.3 to 0.3
PHQ-9[g]	5.9	1.9	3.8	1.5	−2.1	−3.6 to −0.7
MAAS[h]	4.21	0.77	4.16	0.94	−0.05	−0.43 to 0.32

[a]PSQI: Pittsburgh Sleep Quality Index.

[b]ISI: Insomnia Severity Index.

[c]FIRST: Ford Insomnia Response to Stress Test.

[d]PSAS: Pre-Sleep Arousal Scale.

[e]DBAS-16: Dysfunctional Beliefs and Attitudes about Sleep.

[f]GAD-7: Generalized Anxiety Disorder 7-item.

[g]PHQ-9: Patient Health Questionnaire–9.

[h]MAAS: Mindful Attention Awareness Scale.

Figure 2. Individual changes on self-reported outcomes at each visit. DBAS-16: Dysfunctional Beliefs and Attitudes about Sleep; FIRST: Ford Insomnia Response to Stress Test; GAD-7: Generalized Anxiety Disorder 7-item; ISI: Insomnia Severity Index; MAAS: Mindful Attention Awareness Scale; PHQ-9: Patient Health Questionnaire–9; PSAS: Pre-Sleep Arousal Scale; PSQI: Pittsburgh Sleep Quality Index.

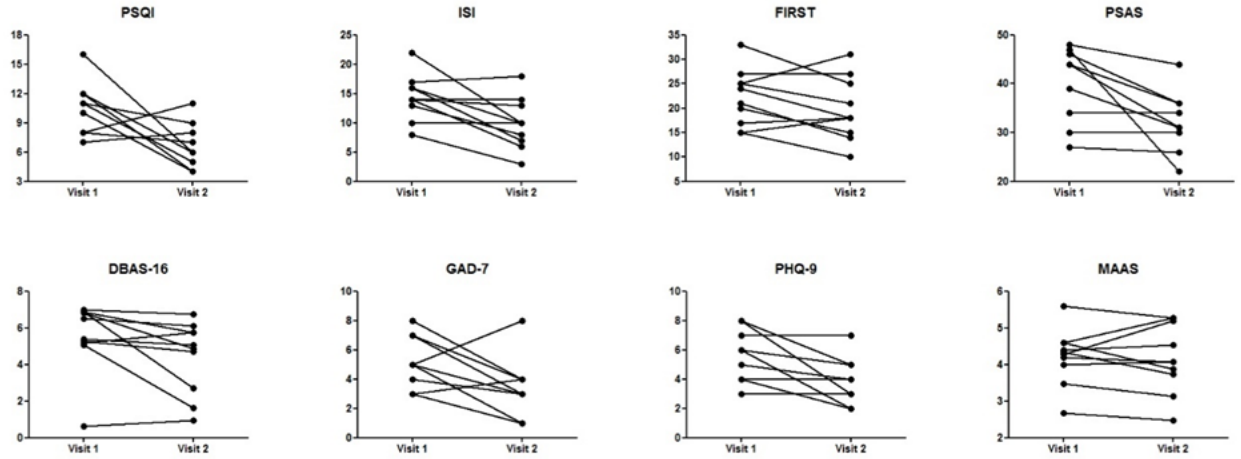

Discussion

Principal Findings

As the first step toward understanding the effectiveness and mechanisms of bedtime MM practice for insomnia, this study was focused on feasibility and acceptability of app-guided bedtime MM practice with in-lab and at-home sleep assessments. We recruited 13 participants and retained 10 (77%) to complete the study protocols within 8 months. We achieved our target rates (90% - 100%) with respect to completion of questionnaires, daily sleep log, heart rate and actigraphy data collection, daily practice adherence, and exit interview supporting the feasibility of a larger study with these procedures [47].

Overall, most participants reported that the app was easy to use and app-guided MM was helpful. Bedtime practice not only helped with sleep but skills learned also translated to and were helpful at other times during the day. Multiple cognitive-emotional processes were mentioned, including decreased catastrophizing, increased acceptance and nonreactivity, overall body awareness, decreased anxiety and letting go of worry, normalizing, and focusing on the present.

XSL•FO
RenderX

We also gained valuable feedback from participants regarding future adaptations that might be considered. Future iterations may consider preferred lengths of bedtime MM practice, instructions around use of ear pods, more user-friendly app interface, and customizations based on other comorbidities such as chronic pain.

We observed clinically meaningful improvements in sleep quality (indicated by PSQI score), insomnia severity (indicated by ISI score), presleep arousal (indicated by PSAS score), vulnerability to sleep disruption (indicated by FIRST), and dysfunctional beliefs about sleep (indicated by DBAS score). Interestingly, we only found very small changes in MAAS score after the intervention. Several possible explanations include heterogeneity of the study subjects and small sample size, short duration of the intervention, and a relevance mismatch between questions on MAAS and our intervention. Other studies have also reported an insensitivity of MAAS in response to mindfulness exposures (lots of references on this). Future studies might consider other measures of mindfulness, for example, the Multidimensional Assessment of Interoceptive Awareness [48,49].

Telehealth and technology-based approaches can improve accessibility to health care. Given the increasing number of studies with remotely delivered mindfulness-based interventions [50-57], we designed this study to include remote intervention and home assessments. Although data collection through wearables can decrease participants' burden associated with scheduling, travel, costs, and wait time, structured demonstration and detailed instructions are important to ensure successful data collection. In our study, we found that active investigator demonstration with the participant prior to leaving the visit was critical to success, even in a relatively young, technology-literate population. The acceptability and feasibility of such app-based intervention and assessment with wearable technology in other populations, for example, older populations, or otherwise less tech-savvy or lower literacy, remain unclear.

We identified gaps and needs to be addressed for future larger-scale studies. For example, while we acknowledge that using electronic devices in bed is contradictory to current standard sleep hygiene recommendations, a large number of patients already follow standard recommendations but still greatly experience difficulty falling asleep. While our study developed targeted instructions for the app-based intervention (eg, use only audios, no or minimal screen exposure, dim light, do-not-disturb mode, and auto-stop playing), optimal app use regarding the timing and duration of practice, as well as the choice of audios to meet individual preferences, should be further studied.

Exploratory analysis of pre- to postsubjective measures supports the potential role for MM in the improvement of sleep, insomnia severity, and presleep arousal [22,58-62]. Recent RCTs in varied populations support community dissemination of the Calm app for sleep disturbance [47]. [63], Several RCTs have shown that the use of the Calm app decreased fatigue, daytime sleepiness, and cognitive and somatic arousal and improved depression and anxiety symptoms [47,63,64]. Our line of inquiry is the first to probe a deeper understanding of efficacy and mechanisms of app-based MM practice specifically at bedtime.

Limitations and Future Research

One limitation of this study is that population composition may not reflect sociodemographic profiles of the general population. Our final participant sample was relatively younger (aged 22 - 50 years), healthy, well-educated, and with chronic insomnia, albeit less severe upon baseline assessment. While chronic insomnia disorder is commonly diagnosed in women, we excluded breastfeeding women or those currently with a young baby, due to presumed irregular circadian rhythms common in this population. Previous research also demonstrates differences between children or adolescents and adults with respect to circadian rhythmicity, causes of insomnia, wake-sleep schedules, and scoring rules for sleep studies; thus, we excluded children. Similarly, as we age, the structure of sleep and the functionality of brain regions change drastically, and physiological changes associated with mindfulness practice in older adults may be different from younger adults. Thus, older adults were also not included in this protocol. In addition, this was a proof-of-concept uncontrolled, single-arm study in a relatively younger and healthier population. The feasibility of retaining a control arm, implementing such a study in other populations (eg, older adults or with comorbid conditions), remains unclear.

In this feasibility pilot study, we prioritized evaluating engagement and adherence rather than enforcing strict adherence to a bedtime-only meditation schedule. While we instructed participants to practice at bedtime, we allowed them to explore the app at other times of the day to understand the naturalistic use of the app, which is relevant for real-world implementation. Although the majority of participants practiced only during bedtime, we acknowledge that the flexibility introduces variability in exposure, making it challenging to attribute outcome changes solely to bedtime mindfulness practice.

It is also important to emphasize that we were explicitly aware of the impact of digital devices and light exposure at bedtime and therefore carefully developed instructions and tips to minimize risks but acknowledge that further refinement of instructions is needed. This is especially critical for potential future RCTs with a larger sample size.

To our knowledge, this is the first study to primarily target bedtime mindfulness practice using remote intervention and to collect data from comprehensive assessments including self-reported outcomes, in-lab sleep study, at-home assessment, and qualitative data from interviews. Despite the limitations, this feasibility study informs and inspires future investigation to better understand neurophysiological mechanisms of bedtime mindfulness practice. In addition, future studies are encouraged to refine instructions for optimal intervention, to collect comprehensive subjective and objective data, and to include a diverse group of patients with insomnia with broader eligibility criteria.

Conclusions

The feasibility outcomes met our preset criteria for success; therefore, a study with bedtime app-guided MM as an

XSL•FO
RenderX

intervention in patients with insomnia and the hybrid design with in-lab and at-home assessments is feasible. Our exploratory analysis of subjective measures indicated improvement in sleep, insomnia severity, and presleep arousal.

Acknowledgments

Dr. Ma acknowledges the support from the National Institutes of Health (National Center for Complementary and Integrative Health K23AT012475). This work was supported by the National Institutes of Health (National Center for Complementary and Integrative Health T32AT000051, K24AT009465, and K24AT009282) and a pilot research grant from the Osher Center for Integrative Health of Harvard Medical School and Brigham and Women's Hospital. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Participants' feedback on bedtime app–based mindfulness intervention acceptability.
[PNG File, 30 KB - formative_v9i1e67366_app1.png]

References

1. Mack LJ, Rybarczyk BD. Behavioral treatment of insomnia: a proposal for a stepped-care approach to promote public health. Nat Sci Sleep 2011;3:87-99. [doi: 10.2147/NSS.S12975] [Medline: 23616720]
2. Maness DL, Khan M. Nonpharmacologic management of chronic insomnia. Am Fam Physician 2015 Dec 15;92(12):1058-1064 [FREE Full text] [Medline: 26760592]
3. Ringdahl EN, Pereira SL, Delzell JE Jr. Treatment of primary insomnia. J Am Board Fam Pract 2004;17(3):212-219. [doi: 10.3122/jabfm.17.3.212] [Medline: 15226287]
4. Roth T. Insomnia: definition, prevalence, etiology, and consequences. J Clin Sleep Med 2007 Aug 15;3(5 Suppl):S7-10. [doi: 10.5664/jcsm.26929] [Medline: 17824495]
5. Ford ES, Wheaton AG, Cunningham TJ, Giles WH, Chapman DP, Croft JB. Trends in outpatient visits for insomnia, sleep apnea, and prescriptions for sleep medications among US adults: findings from the National Ambulatory Medical Care survey 1999-2010. Sleep 2014 Aug 1;37(8):1283-1293. [doi: 10.5665/sleep.3914] [Medline: 25083008]
6. Drake CL, Roehrs T, Roth T. Insomnia causes, consequences, and therapeutics: an overview. Depress Anxiety 2003;18(4):163-176. [doi: 10.1002/da.10151] [Medline: 14661186]
7. Sawadogo W, Adera T. Insomnia symptoms and increased risk of all-cause mortality by age and sex. J Clin Sleep Med 2024 Oct 1;20(10):1585-1593. [doi: 10.5664/jcsm.11200] [Medline: 38722281]
8. Fitzgerald T, Vietri J. Residual effects of sleep medications are commonly reported and associated with impaired patient-reported outcomes among insomnia patients in the United States. Sleep Disord 2015;2015:607148. [doi: 10.1155/2015/607148] [Medline: 26783470]
9. Belleville G, Morin CM. Hypnotic discontinuation in chronic insomnia: impact of psychological distress, readiness to change, and self-efficacy. Health Psychol 2008 Mar;27(2):239-248. [doi: 10.1037/0278-6133.27.2.239] [Medline: 18377143]
10. Hintze JP, Edinger JD. Hypnotic discontinuation in chronic insomnia. Sleep Med Clin 2018 Jun;13(2):263-270. [doi: 10.1016/j.jsmc.2018.02.008] [Medline: 29759276]
11. Hintze JP, Edinger JD. Hypnotic discontinuation in chronic insomnia. Sleep Med Clin 2020 Jun;15(2):147-154. [doi: 10.1016/j.jsmc.2020.02.003] [Medline: 32386690]
12. NIH State-of-the-Science Conference Statement on manifestations and management of chronic insomnia in adults. NIH Consens State Sci Statements 2005;22(2):1-30. [Medline: 17308547]
13. Morgenthaler T, Kramer M, Alessi C, et al. Practice parameters for the psychological and behavioral treatment of insomnia: an update. Am Acad Sleep Med Rep 2006;29(11):1415-1419. [doi: 10.1093/sleep/29.11.1415]
14. Wilson SJ, Nutt DJ, Alford C, et al. British Association for Psychopharmacology consensus statement on evidence-based treatment of insomnia, parasomnias and circadian rhythm disorders. J Psychopharmacol 2010 Nov;24(11):1577-1601. [doi: 10.1177/0269881110379307] [Medline: 20813762]
15. Voiß P, Höxtermann MD, Dobos G, Cramer H. The use of mind-body medicine among US individuals with sleep problems: analysis of the 2017 National Health Interview Survey data. Sleep Med 2019 Apr;56:151-156. [doi: 10.1016/j.sleep.2019.01.008] [Medline: 30799256]

XSL•FO
RenderX

16. Hazlett-Stevens H, Singer J, Chong A. Mindfulness-based stress reduction and mindfulness-based cognitive therapy with older adults: a qualitative review of randomized controlled outcome research. Clin Gerontol 2019;42(4):347-358. [doi: 10.1080/07317115.2018.1518282] [Medline: 30204557]
17. Rash JA, Kavanagh VAJ, Garland SN. A meta-analysis of mindfulness-based therapies for insomnia and sleep disturbance: moving towards processes of change. Sleep Med Clin 2019 Jun;14(2):209-233. [doi: 10.1016/j.jsmc.2019.01.004] [Medline: 31029188]
18. Rusch HL, Rosario M, Levison LM, et al. The effect of mindfulness meditation on sleep quality: a systematic review and meta-analysis of randomized controlled trials. Ann N Y Acad Sci 2019 Jun;1445(1):5-16. [doi: 10.1111/nyas.13996] [Medline: 30575050]
19. Shallcross AJ, Visvanathan PD, Sperber SH, Duberstein ZT. Waking up to the problem of sleep: can mindfulness help? A review of theory and evidence for the effects of mindfulness for sleep. Curr Opin Psychol 2019 Aug;28:37-41. [doi: 10.1016/j.copsyc.2018.10.005] [Medline: 30390479]
20. Suh HW, Jeong HY, Hong S, et al. The mindfulness-based stress reduction program for improving sleep quality in cancer survivors: a systematic review and meta-analysis. Complement Ther Med 2021 Mar;57:102667. [doi: 10.1016/j.ctim.2021.102667] [Medline: 33460741]
21. Ji X, Ivers H, Beaulieu-Bonneau S, Morin CM. Complementary and alternative treatments for insomnia/insomnia-depression-anxiety symptom cluster: meta-analysis of English and Chinese literature. Sleep Med Rev 2021 Aug;58:101445. [doi: 10.1016/j.smrv.2021.101445] [Medline: 33582583]
22. Ong JC, Moore C. What do we really know about mindfulness and sleep health? Curr Opin Psychol 2020 Aug;34:18-22. [doi: 10.1016/j.copsyc.2019.08.020] [Medline: 31539830]
23. Espie CA, Brooks DN, Lindsay WR. An evaluation of tailored psychological treatment of insomnia. J Behav Ther Exp Psychiatry 1989 Jun;20(2):143-153. [doi: 10.1016/0005-7916(89)90047-5] [Medline: 2685045]
24. Harvey AG. Pre-sleep cognitive activity: a comparison of sleep-onset insomniacs and good sleepers. Br J Clin Psychol 2000 Sep;39(3):275-286. [doi: 10.1348/014466500163284] [Medline: 11033749]
25. Lichstein KL, Rosenthal TL. Insomniacs' perceptions of cognitive versus somatic determinants of sleep disturbance. J Abnorm Psychol 1980 Feb;89(1):105-107. [doi: 10.1037//0021-843x.89.1.105] [Medline: 7365114]
26. Weiner L, Weibel S, de Sousa Gurgel W, et al. Measuring racing thoughts in healthy individuals: the Racing and Crowded Thoughts Questionnaire (RCTQ). Compr Psychiatry 2018 Apr;82:37-44. [doi: 10.1016/j.comppsych.2018.01.006] [Medline: 29407357]
27. Bertschy G, Weibel S, Giersch A, Weiner L. Racing and crowded thoughts in mood disorders: a data-oriented theoretical reappraisal. Encephale 2020 Jun;46(3):202-208. [doi: 10.1016/j.encep.2020.01.007] [Medline: 32151457]
28. Weiner L, Martz E, Kilic-Huck Ü, et al. Investigating racing thoughts in insomnia: a neglected piece of the mood-sleep puzzle? Compr Psychiatry 2021 Nov;111:152271. [doi: 10.1016/j.comppsych.2021.152271] [Medline: 34555554]
29. Ansfield ME, Wegner DM, Bowser R. Ironic effects of sleep urgency. Behav Res Ther 1996 Jul;34(7):523-531. [doi: 10.1016/0005-7967(96)00031-9] [Medline: 8826759]
30. Ma Y, Dong M, Mita C, Sun S, Peng CK, Yang AC. Publication analysis on insomnia: how much has been done in the past two decades? Sleep Med 2015 Jul;16(7):820-826. [doi: 10.1016/j.sleep.2014.12.028] [Medline: 25979182]
31. Huberty J, Green J, Glissmann C, Larkey L, Puzia M, Lee C. Efficacy of the mindfulness meditation mobile app "Calm" to reduce stress among college students: randomized controlled trial. JMIR Mhealth Uhealth 2019 Jun 25;7(6):e14273. [doi: 10.2196/14273] [Medline: 31237569]
32. Huberty J, Vranceanu AM, Carney C, Breus M, Gordon M, Puzia ME. Characteristics and usage patterns among 12,151 paid subscribers of the calm meditation app: cross-sectional survey. JMIR Mhealth Uhealth 2019 Nov 3;7(11):e15648. [doi: 10.2196/15648] [Medline: 31682582]
33. Huberty J, Puzia ME, Larkey L, Irwin MR, Vranceanu AM. Use of the consumer-based meditation app calm for sleep disturbances: cross-sectional survey study. JMIR Form Res 2020 Nov 13;4(11):e19508. [doi: 10.2196/19508] [Medline: 33185552]
34. Buysse DJ, Reynolds CF III, Monk TH, Berman SR, Kupfer DJ. The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research. Psychiatry Res 1989 May;28(2):193-213. [doi: 10.1016/0165-1781(89)90047-4]
35. Morin CM, Belleville G, Bélanger L, Ivers H. The Insomnia Severity Index: psychometric indicators to detect insomnia cases and evaluate treatment response. Sleep 2011 May 1;34(5):601-608. [doi: 10.1093/sleep/34.5.601] [Medline: 21532953]
36. Bastien CH, Vallières A, Morin CM. Validation of the Insomnia Severity Index as an outcome measure for insomnia research. Sleep Med 2001 Jul;2(4):297-307. [doi: 10.1016/s1389-9457(00)00065-4] [Medline: 11438246]
37. Jarrin DC, Chen IY, Ivers H, Drake CL, Morin CM. Temporal stability of the Ford Insomnia Response to Stress Test (FIRST). J Clin Sleep Med 2016 Oct 15;12(10):1373-1378. [doi: 10.5664/jcsm.6192] [Medline: 27568895]
38. Kalmbach DA, Pillai V, Arnedt JT, Drake CL. Identifying at-risk individuals for insomnia using the Ford Insomnia Response to Stress Test. Sleep 2016 Feb 1;39(2):449-456. [doi: 10.5665/sleep.5462]
39. Nicassio PM, Mendlowitz DR, Fussell JJ, Petras L. The phenomenology of the pre-sleep state: the development of the Pre-Sleep Arousal Scale. Behav Res Ther 1985;23(3):263-271. [doi: 10.1016/0005-7967(85)90004-x] [Medline: 4004706]

XSL•FO
RenderX

40. Morin CM, Vallières A, Ivers H. Dysfunctional Beliefs and Attitudes about Sleep (DBAS): validation of a Brief Version (DBAS-16). Sleep 2007 Nov;30(11):1547-1554. [doi: 10.1093/sleep/30.11.1547]
41. Kroenke K, Spitzer RL, Williams JBW, Monahan PO, Löwe B. Anxiety disorders in primary care: prevalence, impairment, comorbidity, and detection. Ann Intern Med 2007 Mar 6;146(5):317-325. [doi: 10.7326/0003-4819-146-5-200703060-00004] [Medline: 17339617]
42. Plummer F, Manea L, Trepel D, McMillan D. Screening for anxiety disorders with the GAD-7 and GAD-2: a systematic review and diagnostic metaanalysis. Gen Hosp Psychiatry 2016;39:24-31. [doi: 10.1016/j.genhosppsych.2015.11.005] [Medline: 26719105]
43. Cameron IM, Crawford JR, Lawton K, Reid IC. Psychometric comparison of PHQ-9 and HADS for measuring depression severity in primary care. Br J Gen Pract 2008 Jan;58(546):32-36. [doi: 10.3399/bjgp08X263794] [Medline: 18186994]
44. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613. [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
45. Leon AC, Davis LL, Kraemer HC. The role and interpretation of pilot studies in clinical research. J Psychiatr Res 2011 May;45(5):626-629. [doi: 10.1016/j.jpsychires.2010.10.008] [Medline: 21035130]
46. Ma Y, Yang H, Vazquez M, et al. Dismantling the component-specific effects of yogic breathing: feasibility of a fully remote three-arm RCT with virtual laboratory visits and wearable physiology. Int J Environ Res Public Health 2023 Feb 11;20(4):3180. [doi: 10.3390/ijerph20043180] [Medline: 36833875]
47. Pilot studies: common uses and misuses. National Center for Complementary and Integrative Health. URL: https://www.nccih.nih.gov/grants/pilot-studies-common-uses-and-misuses [accessed 2023-02-05]
48. Mehling WE, Price C, Daubenmier JJ, Acree M, Bartmess E, Stewart A. The Multidimensional Assessment of Interoceptive Awareness (MAIA). PLoS One 2012;7(11):e48230. [doi: 10.1371/journal.pone.0048230] [Medline: 23133619]
49. Mehling WE, Acree M, Stewart A, Silas J, Jones A. The Multidimensional Assessment of Interoceptive Awareness, Version 2 (MAIA-2). PLoS One 2018;13(12):e0208034. [doi: 10.1371/journal.pone.0208034] [Medline: 30513087]
50. Ma Y, Kraemer KM, Lyu J, Yeh GY. Randomized controlled trials of mindfulness and acceptance-based interventions over the past two decades: a bibliometric analysis. J Altern Complement Med 2021 Nov;27(11):930-939. [doi: 10.1089/acm.2020.0548] [Medline: 34252294]
51. Tymofiyeva O, Hu MY, Sipes BS, et al. A feasibility study of a remotely-delivered mindfulness-based training for adolescents during the COVID-19 pandemic. Front Psychiatry 2022;13:838694. [doi: 10.3389/fpsyt.2022.838694] [Medline: 35633797]
52. Wang C, Darbari I, Tolaymat M, Quezada S, Allen J, Cross RK. Randomized controlled pilot study of feasibility and effectiveness of peer led remote mindfulness-based art workshops on stress, anxiety, and depression in medical students. In: Psychology in the Schools: Wiley; 2023:1744-1759. [doi: 10.1002/pits.22809]
53. Farewell CV, Walls J, Powers JN, et al. Feasibility of a perinatal mindfulness-based intervention delivered remotely due to COVID-19. OBM Integr Complement Med 2021 Jun 30;06(3):1-1. [doi: 10.21926/obm.icm.2103028]
54. Moulton-Perkins A, Moulton D, Cavanagh K, Jozavi A, Strauss C. Systematic review of mindfulness-based cognitive therapy and mindfulness-based stress reduction via group videoconferencing: feasibility, acceptability, safety, and efficacy. J Psychother Integr 2022;32(1):110-130. [doi: 10.1037/int0000216]
55. Mirabito G, Verhaeghen P. Remote delivery of a Koru Mindfulness intervention for college students during the COVID-19 pandemic. J Am Coll Health 2024 Apr;72(3):897-904. [doi: 10.1080/07448481.2022.2060708] [Medline: 35427456]
56. Hanson P, Villarreal M, Khan M, Dale J, Sankar S. Effect of an online mindfulness course for hospital doctors during COVID-19 pandemic on resilience and coping. J Prim Care Community Health 2022;13:21501319221138425. [doi: 10.1177/21501319221138425] [Medline: 36448091]
57. Boettcher J, Aström V, Påhlsson D, Schenström O, Andersson G, Carlbring P. Internet-based mindfulness treatment for anxiety disorders: a randomized controlled trial. Behav Ther 2014 Mar;45(2):241-253. [doi: 10.1016/j.beth.2013.11.003] [Medline: 24491199]
58. Ong J, Sholtes D. A mindfulness-based approach to the treatment of insomnia. J Clin Psychol 2010 Nov;66(11):1175-1184. [doi: 10.1002/jclp.20736] [Medline: 20853441]
59. Ong JC, Manber R, Segal Z, Xia Y, Shapiro S, Wyatt JK. A randomized controlled trial of mindfulness meditation for chronic insomnia. Sleep 2014 Sep 1;37(9):1553-1563. [doi: 10.5665/sleep.4010] [Medline: 25142566]
60. Ong JC, Shapiro SL, Manber R. Mindfulness meditation and cognitive behavioral therapy for insomnia: a naturalistic 12-month follow-up. Explore (NY) 2009;5(1):30-36. [doi: 10.1016/j.explore.2008.10.004] [Medline: 19114261]
61. Ong JC, Xia Y, Smith-Mason CE, Manber R. A randomized controlled trial of mindfulness meditation for chronic insomnia: effects on daytime symptoms and cognitive-emotional arousal. Mindfulness (N Y) 2018 Dec;9(6):1702-1712. [doi: 10.1007/s12671-018-0911-6]
62. Perini F, Wong KF, Lin J, et al. Mindfulness-based therapy for insomnia for older adults with sleep difficulties: a randomized clinical trial. Psychol Med 2023 Feb;53(3):1038-1048. [doi: 10.1017/S0033291721002476] [Medline: 34193328]
63. Huberty JL, Green J, Puzia ME, et al. Testing a mindfulness meditation mobile app for the treatment of sleep-related symptoms in adults with sleep disturbance: a randomized controlled trial. PLoS One 2021;16(1):e0244717. [doi: 10.1371/journal.pone.0244717] [Medline: 33411779]

XSL•FO
RenderX

64. Huberty J, Puzia ME, Green J, et al. A mindfulness meditation mobile app improves depression and anxiety in adults with sleep disturbance: analysis from a randomized controlled trial. Gen Hosp Psychiatry 2021;73:30-37. [doi: 10.1016/j.genhosppsych.2021.09.004] [Medline: 34537477]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
DBAS-16: Dysfunctional Beliefs and Attitudes about Sleep
***DSM-IV*:** *Diagnostic and Statistical Manual of Mental Disorders* (Fourth Edition)
***DSM-V*:** *Diagnostic and Statistical Manual of Mental Disorders* (Fifth Edition)
FIRST: Ford Insomnia Response to Stress Test
ISI: Insomnia Severity Index
MAAS: Mindful Attention Awareness Scale
MM: mindfulness meditation
PSAS: Pre-Sleep Arousal Scale
PSG: polysomnograph
PSQI: Pittsburgh Sleep Quality Index
RCT: randomized controlled trial
REDCap: Research Electronic Data Capture

Edited by A Mavragani; submitted 09.10.24; peer-reviewed by L Paulos-Guarnieri; revised version received 24.04.25; accepted 30.04.25; published 30.09.25.

Please cite as:
Ma Y, Wayne PM, Mullington JM, Bertisch S, Yeh GY
Bedtime App–Guided Mindfulness Meditation in Patients With Insomnia: Mixed Methods Feasibility and Acceptability Pilot Study
JMIR Form Res 2025;9:e67366
URL: https://formative.jmir.org/2025/1/e67366
doi:10.2196/67366

© Yan Ma, Peter M Wayne, Janet M Mullington, Suzanne Bertisch, Gloria Y Yeh. Originally published in JMIR Formative Research (https://formative.jmir.org), 30.9.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

A Series of Personalized Melatonin Supplement Interventions for Poor Sleep: Feasibility Randomized Crossover Trial for Personalized N-of-1 Treatment

Mark J Butler[1,2], PhD; Thevaa Chandereng[1,2], PhD; Heejoon Ahn[1,2], MS; Stefani Slotnick[1,2], MS; Danielle Miller[1,2], MPH; Alexandra Perrin[1,2], BA; Jordyn Rodillas[1,2], BA; Ciaran P Friel[1,2], EdD; Ashley M Goodwin[1,2], PhD; Ying Kuen Cheung[3], PhD; Karina W Davidson[1,2], PhD

[1]Northwell, 2000 Marcus Ave, Suite 300, New Hyde Park, NY, United States

[2]Institute of Health System Science, Feinstein Institutes For Medical Research, Manhasset, NY, United States

[3]Mailman School of Public Health, Columbia University, New York, NY, United States

Corresponding Author:
Mark J Butler, PhD
Northwell, 2000 Marcus Ave, Suite 300, New Hyde Park, NY, United States

Abstract

Background: Poor sleep (defined by short sleep duration or poor quality) is a common condition with potential serious health consequences. Exogenous melatonin supplements have been found to effectively improve poor sleep but have also been shown to have heterogeneity of treatment effects (HTEs) between individuals. Personalized N-of-1 trials, in which each participant is the unit of analysis, are ideal for identifying whether a treatment with high HTE is beneficial for each individual patient.

Objective: This study aimed to identify the feasibility, acceptability, and effectiveness of a series of personalized N-of-1 trials of melatonin for poor sleep.

Methods: This study consisted of 60 digital, personalized N-of-1 crossover trials comparing the effects of 3.0 mg and 0.5 mg of melatonin versus placebo for poor sleep with randomization to 1 of 2 orders. The trial comprised a 2-week baseline period and a 12-week intervention period. The primary outcomes were usability of the personalized trial system (measured using the System Usability Scale [SUS]) and participant satisfaction with the trial. Effectiveness outcomes included sleep duration (measured using a Fitbit activity tracker [Google]) and sleep quality (measured using the consensus sleep diary).

Results: Participants rated the usability of the personalized trial as acceptable (average SUS score 76.3, SD 17.1), and 96% (55/57) of those who completed satisfaction surveys stated that they would recommend the trial to others. Importantly, indices of HTE were low for 3.0 mg and 0.5 mg doses of melatonin, indicating that the effect of these treatments on sleep duration and sleep quality did not substantially vary between participants and that averaged treatment responses are appropriate. Averaged participant sleep duration did not significantly differ between the 3.0 mg (P=.70) and 0.5 mg (P=.90) melatonin intervention periods and the baseline period. In addition, regression models did not show differences between different levels of melatonin and placebo periods for sleep duration or quality.

Conclusions: Participant ratings of the usability of and satisfaction with this series of personalized N-of-1 trials of melatonin for sleep suggest these trials are both feasible and acceptable. However, our results show that melatonin supplements did not significantly improve sleep duration or sleep quality. Furthermore, the treatment effects' lack of heterogeneity among participants suggests that future use of N-of-1 trials of melatonin for poor sleep is not needed.

Trial Registration: ClinicalTrials.gov NCT05349188; https://clinicaltrials.gov/study/NCT05349188

(JMIR Form Res 2025;9:e58192) doi:10.2196/58192

KEYWORDS

personalized trial; personalized; N-of-1; sleep quality; poor sleep; sleep duration; virtual; melatonin; placebo; feasibility; intervention; randomized crossover trial; heterogeneity; effectiveness; sleep health; Fitbit; wearable

Introduction

Sleep health is a multidimensional construct that can be difficult to define. Healthy sleep can be defined in terms of sleep duration, sleep quality, and the subjective experience of restful sleep, among other aspects [1]. Although most sleep research focuses on discrete sleep disorders [1,2], difficulties with sleep are extremely common [3,4]. Recent research has shown that short sleep duration (≤6 per night) is prevalent within 37.6% of US workers while poor sleep quality is prevalent among 19.2%

XSL•FO
RenderX

[5]. Poor sleep (as defined by short duration or poor quality sleep) [6] has been associated with adverse emotional (eg, stress and depression), cognitive (eg, memory and cognitive performance), and physiological (eg, pain) symptoms in previous research [7,8]. Poor sleep has also been associated with obesity [9], diabetes [10], hypertension [11], cardiovascular disease [8,11], reduced quality of life [12], depression [12], and mortality [13]. The broad-ranging consequences of poor sleep show the need for interventions to improve sleep quality and duration.

Exogenous melatonin supplements are a widely available and commonly used intervention to treat poor sleep [14]. Several clinical trials have been conducted demonstrating the efficacy of exogenous melatonin in treating sleep disorders regardless of the etiology of the disorder [15-18]. However, clinical trials of melatonin for sleep have found significant heterogeneity in the effectiveness of melatonin interventions [14,19,20]. This is supported by a systematic review that assessed melatonin's effect on several dimensions of sleep across 23 clinical trials and identified significant between-study heterogeneity [20]. These findings suggest that not all participants benefit equally from treatment with melatonin.

This heterogeneity in treatment response for individuals using exogenous melatonin suggests that participants who wish to improve their sleep may benefit from personalized N-of-1 trials [21,22]. Personalized N-of-1 trials are a single-case research approach designed to help patients select the treatment that works best for them as an individual. A personalized N-of-1 trial consists of multiple crossover intervention periods alternating between treatment(s) and control with continuous data collection on outcomes of interest [23,24]. Using single-case statistical methods, participants in personalized N-of-1 trials can evaluate which treatment (or treatments) works best for them on the outcome (or outcomes) that are most important to them. Personalized trials are designed to provide patients with high-integrity, evidence-based information uniquely relevant to the outcomes and values important to them so they may identify efficacious treatments [25]. Furthermore, personalized N-of-1 trials present a significant opportunity to determine the heterogeneity of treatment effects (HTEs) for a particular treatment. Although heterogeneity in treatment response has often been found for treatments (including melatonin), researchers are seldom able to calculate formal HTEs by comparing within and between participant treatment effects [26]. A series of personalized N-of-1 trials presents an ideal opportunity to formally measure the HTE of particular doses of melatonin [26,27]. High levels of HTE for doses of melatonin will indicate that future personalized N-of-1 trials of melatonin may be useful in a clinical context, while low levels of HTE suggest that the magnitude of treatment response does not vary greatly among participants.

Preliminary interviews suggest that personalized trials to improve poor sleep are a high priority among both patients and clinicians [23,28,29]. Despite the potential benefit of personalized N-of-1 trials, they are not common in clinical practice [23,28-30] due to the effort, participant burden, and cost required for implementation [31]. In order for personalized trials of melatonin for sleep to be more widely used, researchers need to identify a way to administer personalized N-of-1 trials in a manner that is feasible and acceptable to study participants. Some potential methods include the use of digital recruitment, remote intervention delivery, and online survey assessment. These methods aligned with increased emphasis on personalized medicine approaches for the improvement of sleep health [32,33].

This study evaluates the feasibility, acceptability, and effectiveness of a series of personalized N-of-1 trials of 2 doses of melatonin (3.0 mg and 0.5 mg, respectively) and a placebo supplement among 60 participants with self-reported poor sleep. Using commercially available supplements, medication adherence devices, an online survey assessment, and wearable technologies, we examined the overall effect of 2 doses of melatonin relative to usual care (ie, placebo) for each participant and aggregated among all participants. Furthermore, using statistical methods and implementation procedures developed in previous personalized N-of-1 trials [27], we identified the HTE for 3.0 mg and 0.5 mg of melatonin. Results from this study will determine whether digital delivery of these interventions is feasible and acceptable for participants with self-reported poor sleep and will allow clinicians to identify whether digital delivery of melatonin can effectively allow patients to manage their sleep.

Methods

Study Design

This study consisted of 60 digital, personalized N-of-1 crossover trials comparing the effects of 3.0 mg and 0.5 mg of melatonin versus placebo for poor sleep. The trial was comprised of a 2-week baseline assessment period and a 12-week intervention period. Before beginning the study, participants were provided with the necessary supplements, 3 Nomi by SMRxT smart pill bottles to contain the supplements, and a Fitbit Charge 5 (Google) activity tracker. The Nomi smart pill bottle is a medication adherence device that uses a weight sensor and cellular signals to measure and transmit medication adherence data in real-time.

During the 2-week baseline assessment period, participants did not take any melatonin or placebo pills and were asked to manage their sleep as they usually would. Participants were also encouraged to wear their Fitbit activity tracker and complete ecological momentary assessment (EMA) measures of fatigue, pain, concentration, confidence, mood, and stress delivered by text message 3 times daily, at randomized times. Each morning, participants completed an additional text-message survey assessing their sleep quality from the previous night.

Following baseline, participants were randomized into 1 of 2 treatment orders comprised of six 2-week intervention blocks of 3.0 mg melatonin, 0.5 mg melatonin, or placebo pills. Following the intervention, each participant was provided with a report summarizing the results of their personalized N-of-1 trial. At the end of the study, participants completed a survey containing satisfaction measures and the System Usability Scale (SUS). An example of the participant timeline can be found in Figure S1 in Multimedia Appendix 1. Study recruitment began

XSL•FO
RenderX

in April 2022 and was completed in May 2023. Additional details about the study design can be found in the trial protocol [34].

Study Population and Recruitment

This was a digital, remote clinical trial. Participants were recruited digitally from several states within the United States. Recruitment methods included internet and social media posting such as Facebook (Meta) and Google advertisements. In addition, online posts to communities on platforms such as Reddit, Craigslist, and LinkedIn were used to appeal to individuals with poor sleep quality. With 93% of the US adults reporting using the internet, 69% reporting ever using Facebook, and 18% reporting using Reddit, usage of online recruitment methods significantly broadened the potential span of our outreach [35]. Emails and advertisements were also distributed through the Northwell Health system, consisting of approximately 85,000 employees [36].

All potential participants identified as having poor sleep for at least the past month using the Insomnia Symptom Questionnaire [37]. Poor sleep for this study was defined by self-reported difficulty falling asleep, staying asleep, and/or feeling refreshed by sleep 3 to 7 times per week, with 1 or more of those symptoms lasting for a period of at least 4 weeks. Participants were also required to report that 1 or more poor sleep quality symptoms bother them “quite a bit” or “always” to meet eligibility criteria.

Ethical Considerations

Ethical Review

This study protocol was approved by the ethics committee and Northwell Health Institutional Review Board (protocol 21-0239). Only IRB-approved team members were allowed access to identifiable data. Participant data were stored in IRB-approved secure databases.

Consent

Study recruitment materials directed participants to an information screen with details about the study and digital screening measures containing questions regarding study inclusion and exclusion criteria. If study inclusion criteria were met, eligible individuals were provided with a link to an electronic consent form and a short video explaining details of the study protocol and consent form. Consent for study participation was obtained electronically. Participants who had questions regarding the study details or consent process were offered a 30-minute phone call with study staff. All participants were aware the study intervention used 2 doses of melatonin and a placebo. Contact information for the study team was provided to potential participants throughout the consent process.

Compensation

After completing all components of the study (ie, successful completion of a research team-approved intervention period, submission of a satisfaction survey, and returning the Nomi devices), study participants were mailed a US $100 payment card (ClinCard). In addition, as a thanks to study participants for their participation, they were allowed to keep their Fitbit Charge 5 (a value of US $150).

Interventions

Participants each received 3 Nomi pill bottles labeled A, B, and C. Bottles labeled with “A” contained 28 pills of 3.0 mg melatonin, bottles with “B” contained 28 pills of 0.5 mg melatonin, and bottles with “C” contained 28 pills of a placebo. All melatonin and placebo doses were provided for this study by Pure Encapsulations LLC, a National Sanitation Foundation– and Good Manufacturing Practices–registered laboratory with a long history of contributions to clinical research [38-40]. The placebo was 100% cellulose powder contained in a capsule, and the capsules of each placebo or supplement were color-matched and designed to be identical. Participants were blind to which capsule was stored in each bottle. The authors selected 3.0 mg and 0.5 mg doses to compare higher and lower doses of melatonin. Previous research has identified 0.3 to 1.0 mg as a low dose of melatonin capable of improving sleep [41]. Doses ranging between 2.0 and 6.0 mg of melatonin have also been found to be effective for improving sleep [42,43]. The doses of 3.0 mg, 0.5 mg, and placebo were selected in this study as we were able to obtain identical capsules for all 3 pills.

Participants were assigned 1 of 2 treatment orders: ABCCBA and CBAABC. Participants received an intervention schedule indicating which bottle labeled “A,” “B,” or “C” they should use for which weeks in the trial. During the intervention, participants received nightly reminder text messages instructing them to take a pill from the appropriate bottle (eg, “Take a pill from Bottle A”). All participants were discouraged from taking other melatonin products while enrolled in the trial. In total, 30 participants were assigned to each treatment order with study randomization conducted by the statistical team using Python to help randomize a list of each treatment order of 30. Randomization allocation was concealed from the study coordinators and investigators, and treatment orders were assigned by the statistical team only after minimum protocol adherence thresholds were achieved by each individual participant during baseline. The participant timeline in Figure S1 in Multimedia Appendix 1 shows these 2 treatment orders. During all study periods, participants were encouraged to wear the Fitbit device 24 hours a day and answer survey measures.

Adherence

During the baseline period, the study team assessed participant adherence to the protocol. Adherence was evaluated using the Fitbit tracker and survey completion. Participants who wore the Fitbit for 12 or more hours per day and for at least 180 minutes while sleeping were defined as adherent to Fitbit wear [44,45]. Submission of a survey before the end of the 2-week baseline period was defined as adherent to survey completion. Participants who did not demonstrate adherence to 80% of study measures during baseline did not progress to the intervention.

Participant Report

Participants who completed the trial received a 15-page personalized report summarizing their trial results by showing the impact of 3.0 mg melatonin, 0.5 mg melatonin, and placebo on their Fitbit-recorded sleep duration. The participant report

XSL•FO
RenderX

also presented summaries and results for other sleep- and wellness-related outcomes including fatigue, stress, pain, and physical activity that may help the participant identify whether melatonin is useful to help their sleep or other well-being goals. Participants were also provided with summaries of their adherence to study measures and to the intervention. All reports were generated automatically but were reviewed by multiple study team members before being sent to participants. A similar version of the participant report in a series of personalized trials for chronic lower back pain is described and displayed in a study by D'Angelo et al [46].

Primary Outcome

The primary outcome of this study was feasibility measured using the 10-item SUS [47] and acceptability as measured by an 18-item satisfaction survey. The SUS is a validated questionnaire that assesses the usability of a system with 10 items on a 5-point Likert scale rated from "strongly disagree" (1) to "strongly agree" (5) [5,48,49]. To obtain a total score, item scores are multiplied by 2.5 and summed to generate a total score ranging from 0 to 100, with higher scores indicating a greater level of usability. The SUS has been validated in multiple contexts [48,49] and has also been used for previous series of personalized N-of-1 trials [45,50]. As participant usability of a personalized trial is essential for a personalized N-of-1 design to be successful, the SUS is used as a feasibility measure. Analyses of the SUS include mean total scores and individual item scores with SDs. Mean SUS scores of ≥68 have been identified as "acceptable" levels of usability while scores ≥80 have been identified as "excellent" levels of usability [51]. We compared participant scores from this study to these established standards as well as SUS scores for other comparable trials to identify the feasibility of this series of personalized trials.

Participant satisfaction with the trial was measured using a 6-item survey rated on a scale of 1 ("not very satisfied") to 5 ("very satisfied"). This survey measures participant satisfaction with specific elements of the trial (eg, the Fitbit activity tracker, SMS text messaging for reminders and surveys, and the personalized trial design). Participants also received a 7-item survey assessing their experience with the trial overall (eg, trial onboarding and informational videos) rated on a scale of 1 ("strongly disagree") to 5 ("strongly agree"). Means and SDs for each item were represented, and frequencies of each participant response were depicted graphically. All satisfaction survey measures were structured similarly to previous series of personalized N-of-1 trials [44,45] but had the specific text altered to fit this trial design.

Secondary Outcomes

The secondary outcomes included sleep duration measured using the Fitbit wearable activity tracker and sleep quality measured using a modified version of the consensus sleep diary (CSD). Nightly sleep duration was calculated using Fitbit's proprietary algorithm with participant information from the Fitbit Charge 5 device. Participant Fitbit activity tracker data were retrieved using the online portals N1thrive and Fitabase. These portals allow for extraction of minute-to-minute activity data as well as daily summaries of outcomes including sleep duration and steps count. Although previous meta-analyses indicate that Fitbit activity tracking devices may overestimate sleep duration [52], the personalized N-of-1 design allows for within-participant comparisons of sleep durations between intervention periods. The CSD is a validated survey instrument used to assess and track intervention effects on sleep symptoms, quality, and satisfaction [53]. This study used a modified version of the CSD to assess perceived sleep quality using a 5-point Likert scale rated from "very poor" to "very good," with higher scores indicating better sleep. The CSD was delivered daily via text message each morning within an hour of the participant's self-reported typical wake time. Means and SDs of participant sleep duration and sleep quality were reported for baseline and each of the intervention periods (3.0 mg melatonin, 0.5 mg melatonin, and placebo). Paired samples *t* tests were used to compare mean participant sleep duration and sleep quality during the intervention periods to baseline. For analyses of self-reported sleep quality, participant responses were change log-transformed to comply with the assumptions for a normal distribution in a paired samples *t* test.

Additional secondary outcomes measured included self-reported EMA fatigue ratings and self-reported EMA stress ratings. EMA fatigue and stress were assessed using single-item measures adapted from the numeric pain rating scale [54]. These EMA measures were administered 3 times daily via SMS text message asking participants to rate their fatigue and stress in the current moment on a scale of 0-10. The EMA assessments were delivered randomly throughout the participant's self-reported sleep and wake times. Paired samples *t* tests were used to compare ratings of EMA fatigue and EMA stress between baseline and each of the intervention periods (3.0 mg melatonin, 0.5 mg melatonin, and placebo).

The effect of each intervention (3.0 mg melatonin, 0.5 mg melatonin, and placebo) on sleep duration, EMA fatigue, and EMA stress was examined using generalized linear mixed models with an autoregressive model, which was used to account for possible linear trends and autocorrelation between outcome ratings over time. These models provided estimates of the effects of each treatment on each secondary outcome across all participants. We conducted additional analyses at the individual participant level to identify the effects of each intervention for each participant. These regressions also used generalized least squares analyses with autoregressive models. Using these regressions, we were able to generate estimates of the effect of each intervention on each outcome for every participant individually. Due to low levels of variation in self-reported sleep quality, participant ratings were averaged over each week, and a linear mixed model was conducted using these weekly averages for the overall sample and for each individual participant.

HTEs

To support the utility of conducting a series of N-of-1 trials of melatonin for poor sleep, we examined the HTE for both 3.0 mg and 0.5 mg of melatonin. The HTE for both doses of melatonin was examined for 4 outcomes (ie, sleep duration, sleep quality, EMA fatigue, and EMA stress) using methodology previously developed for analyses of N-of-1 trials [27]. Briefly,

XSL•FO
RenderX

we modeled the effects of 3.0 mg and 0.5 mg of melatonin relative to placebo on outcomes using linear mixed models with first-order autocorrelation autoregressive model. We then compared 2 models of the effects for 3.0 mg and 0.5 mg of melatonin versus placebo (1 with a random intercept only and 1 with random slopes) using a likelihood ratio test. The random slope model accounts for HTE, whereas the random intercept model does not. The results of this likelihood ratio test were then used to determine whether HTE exists for a particular treatment on a particular outcome. We also calculated an index of heterogeneity to quantify the HTE for each treatment on each outcome [27,55]. This index is used to interpret the utility of N-of-1 trials; higher values indicate that an N-of-1 trial may be more useful for examining the effects of a particular treatment on a particular outcome.

Sample Size Calculation

The sample size of 60 participants was chosen to ensure a sufficient number to obtain a preliminary assessment of the feasibility of this series of personalized trials of melatonin therapy for sleep. With 60 participants receiving the intervention and use of a single sample binomial test with α set to 2.5% significance, we will have approximately 90% power if 70% (42/60) of participants complete the trial. With 57 participants completing the outcome, we have greater than 90% power.

Results

Enrollment and Sample Characteristics

In total, 654 participants were screened for the study, yielding an enrolled sample of 60 participants. Characteristics of the enrolled sample can be found in Table 1. Full details regarding study screening and enrollment can be found in the CONSORT (Consolidated Standards of Reporting Trials) diagram (Figure 1). Of this sample, 57 participants completed the primary outcome of the SUS. In this sample of 57 individuals, the mean age was 41.6 (SD 12.8) years, sex was 53% (n=30) female, 58% (n=33) White, and 18% (n=10) were Hispanic. Participant characteristics did not differ between the 2 treatment orders. Full characteristics of this sample can be found in Table S1 in Multimedia Appendix 1.

Table . Baseline characteristics of the sample (N=60).

Variable	Total sample	Treatment order 1 (n=30)[a]	Treatment order 2 (n=30)[b]	*P* value[c]
Age (y), mean (SD)	41.1 (12.7)	42.9 (13.1)	39.4 (12.3)	.29
Sex, n (%)				.30
Female	32 (53)	18 (60)	14 (47)	
Male	27 (45)	11 (37)	16 (53)	
Other	1 (2)	1 (3)	0 (0)	
Race, n (%)				.97
Asian	7 (12)	4 (13)	3 (10)	
Black	7 (12)	4 (13)	3 (10)	
Mixed	3 (5)	2 (7)	1 (3)	
Other	7 (12)	3 (10)	4 (13)	
White	34 (57)	16 (53)	18 (60)	
Unknown	2 (3)	1 (3)	1 (3)	
Ethnicity, n (%)				.75
Hispanic	12 (20)	7 (23)	5 (17)	
Non-Hispanic	47 (78)	23 (77)	24 (80)	
Unknown	1 (2)	0 (0)	1 (3)	

[a]Treatment order 1: 3 mg melatonin, 0.5 mg melatonin, placebo, placebo, 0.5 mg melatonin, 3 mg melatonin.

[b]Treatment order 2: placebo, 0.5 mg melatonin, 3 mg melatonin, 3 mg melatonin, 0.5 mg melatonin, placebo.

[c]*P* values for comparisons of participant characteristics between treatment orders were obtained.

XSL•FO
RenderX

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) extension for reporting N-of-1 (CENT) flow diagram.

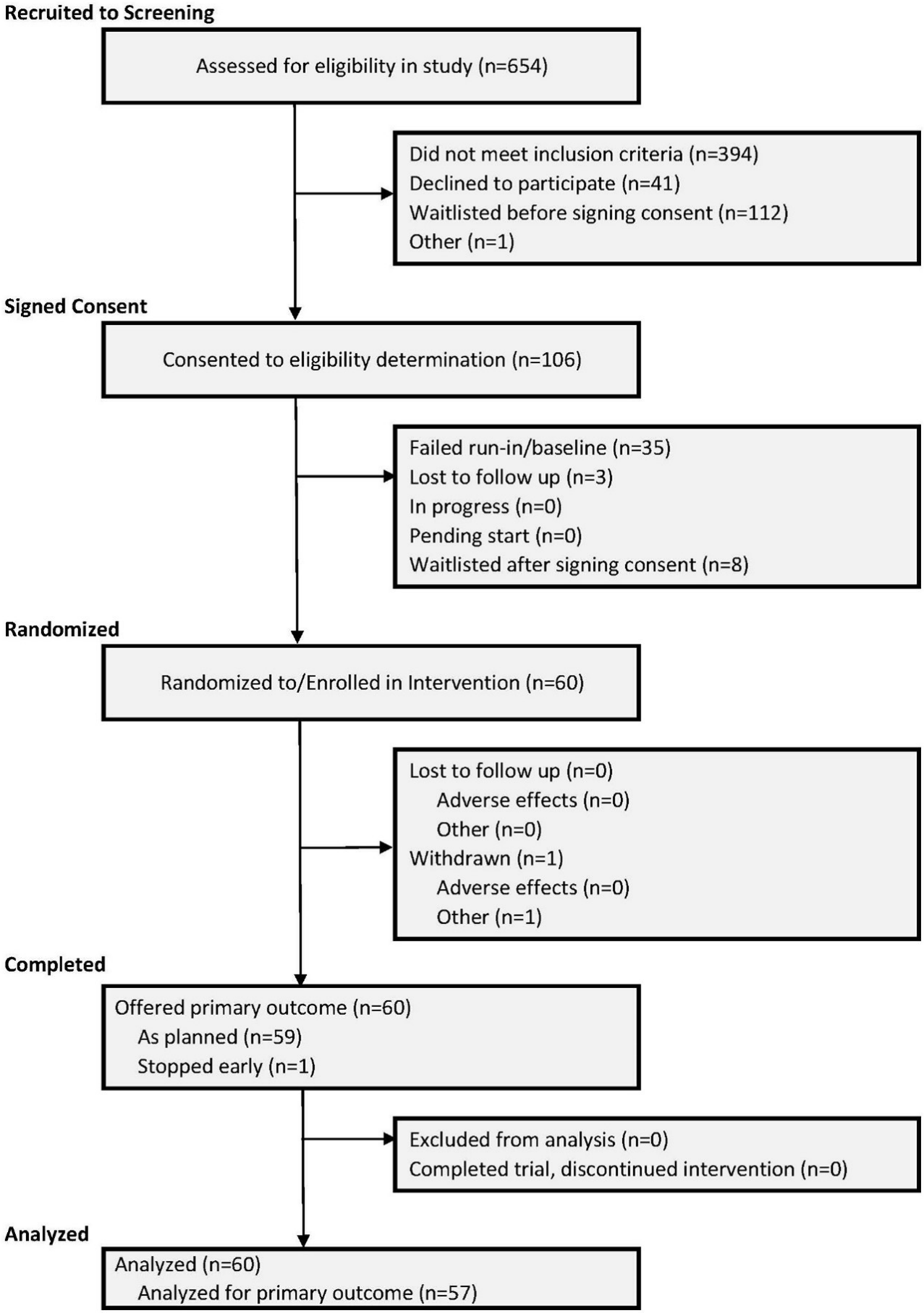

XSL•FO
RenderX

SUS

A total of 57 participants completed the primary outcome of the SUS. The mean SUS score across all participants was 76.27 (SD 17.09, Table 2), indicating that this trial had an acceptable level of usability (SUS total score≥70) [51]. Scores on the SUS ranged from 30 to 100 (Figure 2), indicating that several participants (13/57, 22.8%) found the usability and learnability of this series of personalized trials to be less than acceptable, while others (22/57, 38.6%) rated the usability of this series of personalized trials to be excellent. The most positively rated items from the SUS were "I do not think that I would need the support of a technical person to be able to use this system" (mean 3.28, SD 0.94) and "I felt very confident using the system" (mean 3.25, SD 0.81, Table 2). The least positively rated items were "I found the various functions in this system were well integrated" (mean 2.88, SD 0.76) and "I did not think there was too much inconsistency in this system" (mean 2.81, SD 0.97). Full distributions of responses for each item on the SUS can be found in Figure 3.

Table . System Usability Scale scores (N=57).

Measure	Values, mean (SD)	Range
System Usability Scale overall score	76.27 (17.09)	30-100
System Usability Scale individual items[a]		
I think that I would like to use this system frequently.	2.75 (1.07)	0-4
I did not find the system unnecessarily complex.[b]	1.02 (1.01)	0-3
I thought the system was easy to use.	3.14 (0.95)	0-4
I do not think that I would need the support of a technical person to be able to use this system.[b]	0.72 (0.94)	0-3
I found the various functions in this system were well integrated.	2.88 (0.76)	1-4
I did not think there was too much inconsistency in this system.	1.19 (0.97)	0-3
I would imagine that most people would learn to use this system very quickly.	3.14 (0.79)	1-4
I did not find the system very awkward to use.[b]	0.93 (0.88)	0-3
I felt very confident using the system.	3.25 (0.81)	1-4
I did not need to learn a lot of things before I could get going with this system.[b]	0.79 (0.86)	0-3

[a]Questions rated on a 5-point Likert scale from 0 "strongly disagree" to 4 to "strongly agree."

[b]Items were initially reverse-coded but have been re-coded to be on the same scale as other items. The text of these questions has been revised from the original items to reduce confusion.

XSL•FO
RenderX

Figure 2. System Usability Scale (SUS) scores by participant (N=57*)*. This figure displays each participant's response to the SUS administered at the end of the trial. Ranges for excellent, acceptable, and less than acceptable ratings are based on the established literature for interpreting the SUS.

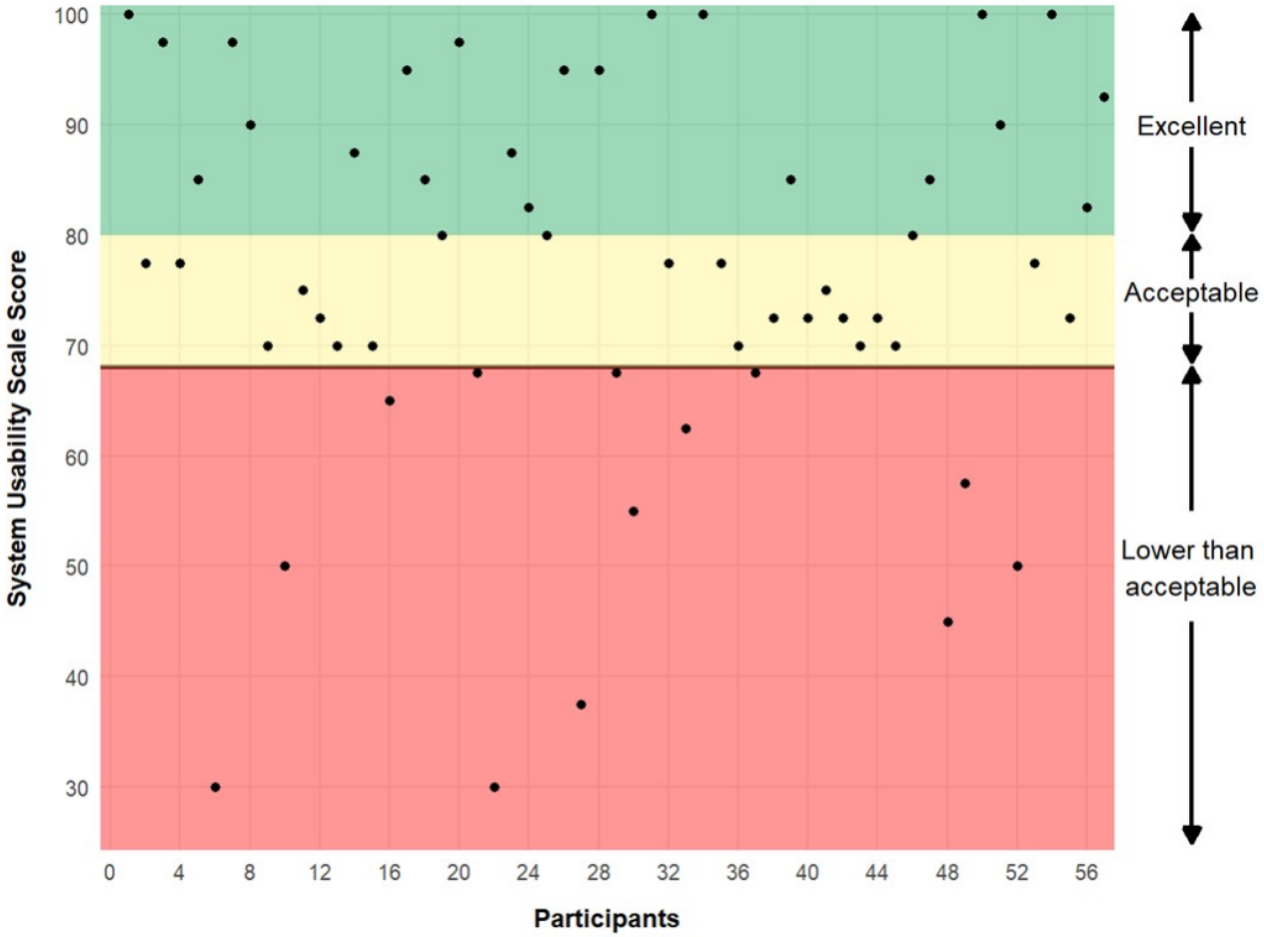

XSL•FO
RenderX

Figure 3. Distribution of responses for individual items on the System Usability Scale.

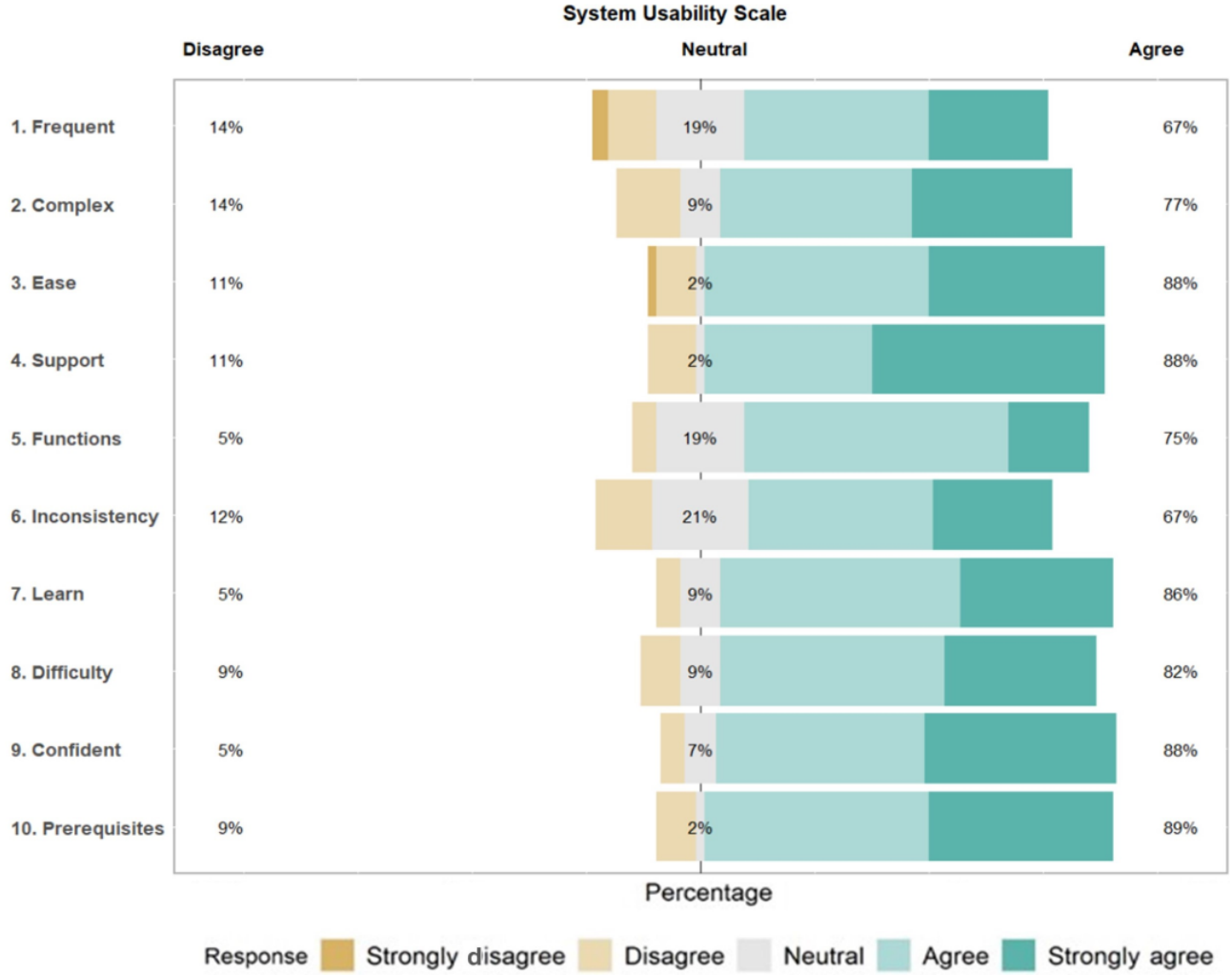

Satisfaction

Participant satisfaction ratings of the trial were generally high on a 5-point scale (Figures 4 and 5). The most positively rated items on the satisfaction survey were "use of the Fitbit to track your activity and sleep" (mean 4.58, SD 0.94) and "presentation of your results" (mean 4.51, 0.89). The least positively rated items on the satisfaction survey were "I enjoyed receiving daily text message prompts and surveys on my cell phone" (mean 3.12, SD 1.23) and "text messaging for survey questions" (mean 3.47, SD 1.48).

XSL•FO
RenderX

Figure 4. Participant satisfaction with elements of the personalized trial.

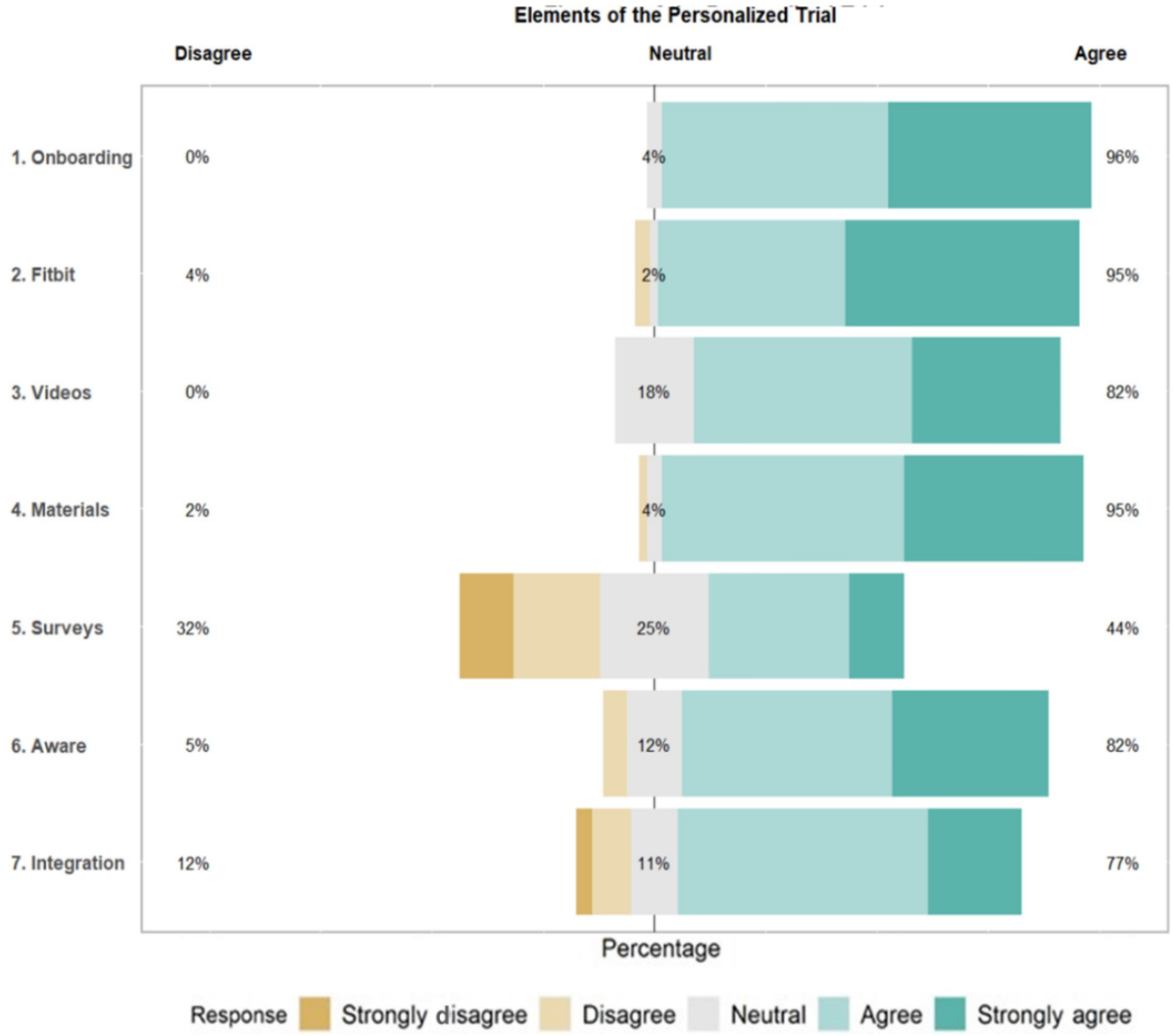

XSL•FO
RenderX

Figure 5. Participant satisfaction with components of the trial.

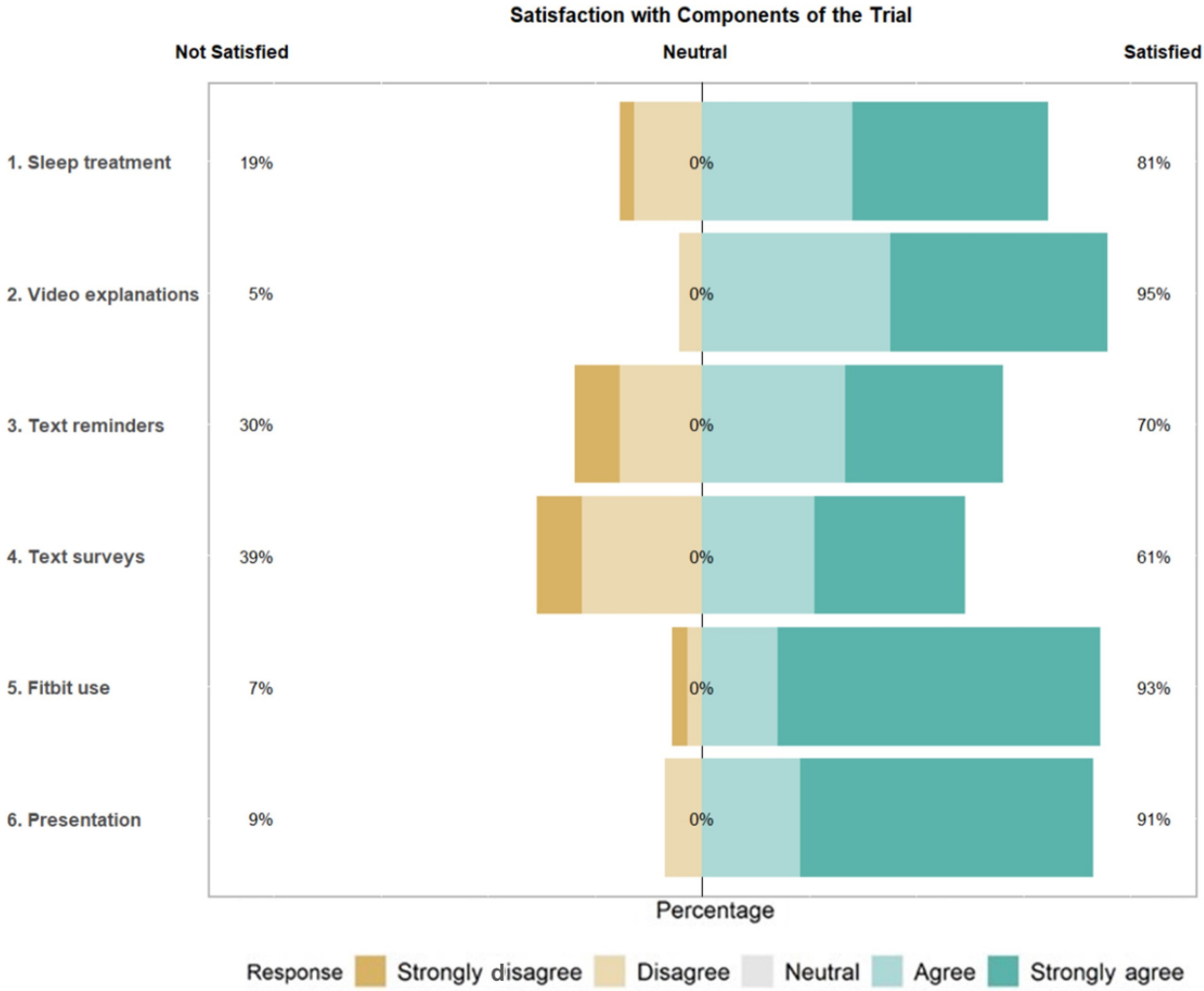

In addition, the majority of participants (36/57, 63%) stated they would "strongly recommend" the trial to others (Multimedia Appendix 2). Only 3.5% (2/57) of participants stated they "would not recommend" the personalized trial to others. Participants mostly felt the trial was helpful; 12% (7/57) found the trial "extremely" helpful, 40% (23/57) rated the trial "very much" helpful, and 39% (22/57) found the trial "somewhat" helpful (Multimedia Appendix 3).

Fitbit Sleep Duration

At baseline, participants had 399 (SD 53) minutes of sleep per night, indicating participants started with roughly 6-and-a-half hours of sleep per night. Participant sleep durations were comparable during placebo (mean 398, SD 59 min), 0.5 mg melatonin (mean 398, SD 54 min), and 3.0 mg melatonin intervention periods (mean 402, SD 61 minutes) (Table 3). Paired samples *t* tests also showed no significant difference in mean sleep duration from baseline between placebo (*P*=.80), 0.5 mg melatonin (*P*=.90), and 3.0 mg melatonin (*P*=.70) intervention periods. Regression models in the pooled sample of participants showed no significant overall effects of 3.0 mg melatonin versus placebo (B=4.73, 95% CI −6.85 to 16.30), 0.5 mg melatonin versus placebo (B=1.23, 95% CI −8.28 to 10.75), and 3.0 mg melatonin versus 0.5 mg melatonin (B=3.40, 95% CI −6.24 to 13.03) (Table 4). Regression estimates for these treatment effects with 95% CIs can be found for overall sample and for each individual participant in Figure 6.

XSL•FO
RenderX

Table . Participant-reported outcomes by intervention period (N=55).

Outcome	Values, mean (SD)				Values, mean difference (95% CI)		
	Baseline	Placebo	0.5 mg melatonin	3 mg melatonin	Baseline versus placebo	0.5 mg versus baseline	3 mg versus baseline
Fitbit sleep duration (min/night)	399 (53)	398 (59)	398 (54)	402 (61)	−1.5 (−14.00 to 11.00)	−0.78 (−12.00 to 11.00)	2.6 (−9.60 to 15.00)
CSD[a] sleep quality	3.13 (0.64)	3.38 (0.53)	3.44 (0.55)	3.35 (0.56)	0.25[b] (0.11 to 0.39)	0.31[b] (0.17 to 0.45)	0.21[c] (0.07 to 0.35)
EMA[d] fatigue	3.85 (1.95)	3.44 (2.00)	3.49 (2.07)	3.58 (1.98)	−0.41[c] (−0.77 to −0.04)	−0.36[c] (−0.72 to −0.01)	−0.29 (−0.61 to 0.03)
EMA stress	3.19 (1.86)	3.33 (1.96)	3.29 (1.86)	3.34 (1.78)	0.14 (−0.15 to 0.43)	0.10 (−0.14 to 0.33)	0.11 (−0.16 to 0.38)

[a]CSD: consensus sleep diary; measured on a scale of 1 to 5 with higher scores indicating better quality.

[b]*P* values for paired samples *t* tests: *P*<.001.

[c]*P* values for paired samples *t* tests: *P*<.05.

[d]EMA: ecological momentary assessment; measured on a scale of 1 to 10 with higher scores indicating higher levels (eg, a 10 would be the most stressed).

Table . Participant-reported outcomes by intervention period regression results.

Outcome	Values, B (95% CI)		
	3 mg melatonin versus placebo	0.5 mg melatonin versus placebo	3 mg melatonin versus 0.5 mg melatonin
Fitbit sleep duration (minutes/ night)	4.73 (−6.85 to 16.30)	1.23 (−8.28 to 10.75)	3.40 (−6.24 to 13.03)
CSD[a] sleep quality	−0.04 (−0.13 to 0.06)	0.04 (−0.05 to 0.13)	−0.09 (−0.18 to −0.01)[b]
EMA[c] fatigue	0.08 (−0.11 to 0.26)	0.04 (−0.12 to 0.20)	0.02 (−0.12 to 0.17)
EMA stress	−0.07 (−0.25 to 0.11)	−0.04 (−0.17 to 0.10)	−0.04 (−0.21 to 0.14)

[a]CSD: consensus sleep diary; measured on a scale of 1 to 5 with higher scores indicating better quality.

[b]*P*<.05 in an autoregressive model

[c]EMA: ecological momentary assessment; measured on a scale of 1 to 10 with higher scores indicating higher levels.

XSL•FO
RenderX

Figure 6. Difference in Fitbit-measured sleep duration by intervention period. These forest plots depict the results of autoregressive models of order 1 for each individual participant as well as pooled across all participants. CIs highlighted in green are for estimates that are statistically significant with $P<.05$.

XSL•FO
RenderX

Sleep Quality

During baseline, participants reported that their sleep quality was average (mean 3.13, SD 0.64 on a scale of 1 to 5, Table 3). Paired samples *t* tests also showed significant increases in mean sleep quality from baseline between placebo ($P<.001$), 0.5 mg melatonin ($P<.001$), and 3.0 mg melatonin ($P=.005$) intervention periods. However, regression models in the pooled sample of participants showed no significant overall effects for 3.0 mg melatonin versus placebo (B=−0.04, 95% CI −0.13 to 0.06) and 0.5 mg melatonin versus placebo (B=0.04, 95% CI −0.05 to 0.13). There was a small but statistically significant difference between 3.0 mg melatonin versus 0.5 mg melatonin (B=−0.09, 95% CI −0.18 to −0.01; Table 4). In Figure 7, regression estimates for these treatment effects with 95% CIs can be found for the overall sample and for each individual participant.

XSL•FO
RenderX

Figure 7. Difference in self-reported sleep quality by intervention period. These forest plots depict the results of autoregressive models of order 1 for each individual participant as well as pooled across all participants. CIs highlighted in green are for estimates that are statistically significant with $P<.05$.

XSL•FO
RenderX

EMA Measures of Fatigue and Stress

During baseline (mean 3.85, SD 1.95), participants also reported relatively low levels of fatigue (Table 3). Mean participant levels of fatigue decreased during placebo (mean 3.44, SD 2.00), 3.0 mg melatonin (mean 3.58, SD 1.98), and 0.5 mg melatonin periods (mean 3.49, SD 2.07). In addition, paired sample *t* tests showed small but statistically significant reductions in EMA fatigue between baseline and both 0.5 mg melatonin (mean difference −0.36, 95% CI −0.72 to −0.01; *P*=.045) and placebo periods (mean difference −0.41, 95% CI −0.77 to −0.04; *P*=.03). No significant differences were found between baseline and 3.0 mg melatonin periods in paired samples *t* tests (*P*=.08; Table 3). Regression estimates comparing EMA fatigue values between 3.0 mg melatonin (B=0.08, 95% CI −0.11 to 0.26) and 0.5 mg melatonin (B=0.04, 95% CI −0.12 to 0.20) versus placebo periods were also not statistically significant (Table 4).

Participants reported relatively low levels of stress during baseline (mean 3.19, SD 1.86), placebo (mean 3.33, SD 1.96), 0.5 mg melatonin (mean 3.29, SD 1.86), and 3.0 mg melatonin periods (mean 3.34, SD 1.78). Results from paired samples *t* tests did not show any significant difference between baseline and intervention periods (Table 3). Furthermore, there was no difference between treatment periods in the autoregression models conducted (Table 4). Figures 8 and 9 present regression estimates for these fatigue and stress, respectively, with 95% CIs for the overall sample and for each individual participant.

XSL•FO
RenderX

Figure 8. Difference in EMA fatigue. These forest plots depict the results of autoregressive models of order 1 for each individual participant as well as pooled across all participants. CIs highlighted in green are for estimated whichestimates that are statistically significant with $P<.05$. EMA: ecological momentary assessment.

XSL•FO
RenderX

Figure 9. Difference in EMA stress. These forest plots depict the results of autoregressive models of order 1 for each individual participant as well as pooled across all participants. CIs highlighted in green are for estimated whichestimates that are statistically significant with $P<.05$. EMA: ecological momentary assessment.

XSL•FO
RenderX

HTEs

Analyses of the HTE for placebo, 0.5 mg melatonin, and 3.0 mg melatonin showed that a random slope model was a better fit for the data than a random intercept for modeling the outcomes of sleep duration, EMA fatigue, and EMA stress (Table 5; likelihood ratio *P* value<.001). The SD of the treatment effect for 3.0 mg versus placebo and 0.5 mg melatonin versus placebo was also larger than the treatment effects. However, the indices of HTE for 3.0 mg versus placebo and 0.5 mg versus placebo were low for sleep duration, EMA fatigue, and EMA stress. These indices reflect the potential power for future personalized N-of-1 trials of each intervention for each outcome [27]. Based on these indices, future series of personalized N-of-1 trials will have statistical power ranging between 7% and 23% depending on the intervention and outcome (Table 5). This suggests that although a random slope model is more useful than a random intercept model for examining the effects of each dose of melatonin on these 3 outcomes, participants did not have a heterogeneous treatment response. Based on these preliminary findings, further N-of-1 trials of melatonin for sleep duration, EMA fatigue, and EMA stress are unlikely to be necessary in this population.

Table . Heterogeneity of treatment effect for secondary outcomes.

Outcomes	3 mg melatonin			0.5 mg melatonin			Overall		
	Treatment effect (3 mg melatonin versus placebo)	SD of treatment effect (3 mg melatonin)	HTE[a] index for 3 mg melatonin	Treatment effect (0.5 mg melatonin versus placebo)	SD of treatment effect (0.5 mg melatonin)	HTE[a] index for 0.5 mg melatonin	Between-subject SD of outcome	Within-subject SD of outcome	Likelihood ratio test *P* value[b]
Fitbit sleep duration (min/night)	4.52	25.85	0.23	0.96	10.60	0.07	53.04	110.25	<.001
EMA[c] fatigue	0.08	0.45	0.17	0.05	0.33	0.10	1.97	1.42	<.001
EMA[c] stress	−0.07	0.44	0.18	−0.03	0.20	0.07	1.92	1.23	<.001

[a]HTE: heterogeneity of treatment effect; this index ranges from 0.00 to 1.00 with higher scores indicating greater levels of heterogeneity.

[b]Results for this likelihood ratio test reflect an omnibus test comparison of a random intercept model of the treatment effect on the outcome versus a random slope model of the same treatment effect. A *P* value <.05 reflects a comparison in which the random slope model has a better fit.

[c]EMA: ecological momentary assessment; measured on a scale of 1 to 10 with higher scores indicating higher levels.

Analyses of the HTE for self-reported sleep quality were unable to be completed due to a lack of variance in participant responses. While a formal analysis of the HTE was not conducted for sleep quality, this lack of variance in participant responses also suggests that treatment effects of melatonin on sleep quality were not heterogeneous.

Discussion

Principal Findings

The results from this trial show that digital, remote personalized N-of-1 trials of melatonin for poor sleep are both feasible and acceptable to participants. However, the melatonin interventions were not significantly associated with changes in measures of poor sleep (including device-measured sleep duration and self-reported sleep quality). This suggests that, although this system of personalized N-of-1 trials can be successfully implemented among individuals with poor sleep, melatonin may not be an ideal intervention for a personalized N-of-1 trial design.

Participants rated the feasibility (defined using the SUS) as being acceptable. Although the mean SUS scores were lower than the previous series of personalized trials conducted by the authors [45,50], participant ratings were still higher than many other interventions with sleep as the primary outcome. In fact, usability ratings were comparable or better than previous interventions to improve sleep, including blue light therapy for sleep among patients with Parkinson disease (SUS median70, IQR 57.5-80.0) [56], a digital sleep diary intervention (SUS mean 74.6, SD 9.4) [57], a mobile health (mHealth) intervention to improve sleep among shift workers (SUS mean 62.7, SD 12.7) [58], an mHealth app to support the use of continuous positive airway pressure (CPAP) among patients with obstructive sleep apnea (SUS mean 79.4, SD 19.3) [59], an mHealth sleep monitoring device for management of sleep symptoms among older individuals (SUS mean 75.5, SD 10.2) [60], and a web-based decision aid for identifying treatment for obstructive sleep apnea (SUS mean 78.2, SD 15.1) [61]. Interestingly, Fitbit wearable activity trackers tend to have acceptable levels of usability on the SUS in previous research [62,63].

What is perhaps more interesting is the fact that both 3.0 mg and 0.5 mg doses of melatonin did not differ significantly from placebo in terms of Fitbit device-recorded sleep duration, self-reported sleep quality, and EMA measures of fatigue and stress. Furthermore, we did not find heterogeneity in the treatment response among participants. Notably, however, there were small but statistically significant differences in sleep quality and fatigue between the baseline and intervention periods. In addition, meta-analyses of the effect of melatonin on sleep duration have found an average increase of 12 minutes [14] and a significant improvement in sleep quality [20]. Both of these meta-analytic studies found heterogeneity in the treatment effects of melatonin [14,20]. Although this

XSL•FO
RenderX

heterogeneity was identified in both meta-analyses using Cochran Q statistic (a method that has been called into question in recent literature) [64], our study found homogeneous, nonsignificant effects of melatonin versus placebo for many of our outcomes. This is especially surprising since the authors have previously found interventions used in series of N-of-1 trials to treat chronic lower back pain and fatigue had small overall effects but high levels of heterogeneity in treatment response for some interventions and outcomes [65,66].

While this finding is surprising, it is not unprecedented. Previous research has found that melatonin is most effective among individuals dealing with sleep restriction or altered sleep schedules (eg, individuals with jet lag and shift workers) [67]. This is logical given melatonin's role in the regulation of circadian rhythm. However, the causes of poor sleep can be variable and complex. Lifestyle factors (eg, smoking, diet, and physical activity) [4,13], individual characteristics (eg, sex, BMI, socioeconomic status, and medical comorbidity) [4,13,68,69], and environment factors (eg, having children, marital status, irregular work hours, and noise) [4,70-72] can all influence sleep quality and duration. While melatonin may address how some of these factors impact circadian rhythm, melatonin supplements' ability to address how environmental factors, such as noise, influence sleep will be difficult. However, we would hope that some participants in this trial would have poor sleep that was linked to a disruption of circadian rhythm. The lack of heterogeneity in our effects suggests that melatonin may not be a useful intervention for personalized N-of-1 trials that broadly recruit individuals with self-reported poor sleep.

Strengths and Limitations

This study has several notable strengths. First, this series of personalized trials was implemented by a research team with extensive experience in the design and conduct of personalized N-of-1 research. The statistical team supporting the work of this trial has also developed methods for specifically analyzing personalized N-of-1 data and calculating HTE [27,73]. We believe this expertise has contributed to a well-designed and well-conducted series of personalized N-of-1 trials. Second, the use of digital, remote interventions, data collection, and dissemination of findings also means that the design of this study may be easily replicable. Third, feasibility and participant satisfaction with the trial were assessed using multiple metrics, including the SUS, a validated measure of usability used to evaluate other digital interventions. This measure was supplemented by satisfaction surveys that were tailored specifically to the design of this trial. These measures have allowed us to compare our series of personalized N-of-1 trials with other digital interventions for sleep while also collecting feedback specific to the design of this study.

Although this trial does have many strengths, there are also some limitations to consider. First, the outcomes used in this trial are all self-reported measures or measurements from a commercially available device. Both sleep duration [52] and sleep quality [74] can be measured using more accurate and precise measures to better capture dimensions of poor sleep. However, this trial uses readily available measures and devices designed to reduce participant burden. By using accessible and lower-cost tools to capture dimensions of poor sleep, the goal was to increase the feasibility of personalized N-of-1 trials for sleep. Regardless, more accurate measures of sleep duration and sleep quality may capture nuances of poor sleep that the measures in this trial do not. Second, participants entered the trial with generally high levels of self-reported sleep duration and sleep quality. The mean baseline sleep duration for the sample was 399 minutes. Although this duration is lower than the 7 - 10 hours of sleep duration recommended for adults [75], sleep durations of ≤5 hours per night are mostly associated with adverse outcomes [76]. However, this average sleep duration falls within the definition of "short sleep" [76,77]. In addition, participants reported average levels of sleep quality at baseline (mean 3.13, SD 0.64). Individuals with poorer levels of sleep during baseline may have benefitted more from melatonin during the intervention period. Third, this trial only examined 2 doses of commercially available melatonin (3.0 mg and 0.5 mg). As discussed previously, doses of melatonin ranging from 0.3 mg to 6.0 mg have been used to successfully improve poor sleep [41-43]. The doses used in this trial may not have been optimal melatonin levels to influence poor sleep. Another limitation with commercially available melatonin is that some evidence suggests that the actual dose in melatonin supplements may vary significantly between lots [78]. We aimed to minimize this limitation by purchasing supplements from a single manufacturer in a single lot. A fourth and final limitation is that this trial was not designed to enroll a representative sample of individuals with poor sleep. Although there was a fairly diverse sample, to definitively say that the results from this series of personalized trials will generalize to all populations is impossible.

Conclusion

The findings of this study show that personalized N-of-1 trials for poor sleep are highly feasible and acceptable. However, our results show that melatonin supplements did not significantly improve sleep duration or sleep quality. Furthermore, these treatment effects were not heterogeneous between participants. This suggests that, despite the increased prevalence of melatonin use in the United States [79], melatonin may not be broadly applicable as a treatment for poor sleep, especially in personalized N-of-1 trials.

Acknowledgments

We are grateful for the contributions and support of the National Library of Medicine, Northwell Health System, Pure Encapsulations, and Nomi by SMRxT for providing the investigators with the support and tools for the development and implementation of this trial.

XSL•FO
RenderX

This work was supported by the National Library of Medicine of the National Institutes of Health (grant R01LM012836). The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication. The views expressed in this paper are those of the authors and do not represent the views of the National Institutes of Health, the US Department of Health and Human Services, or any other government entity.

Data Availability

The datasets generated or analyzed during this study are available in the Open Science Framework repository [80].

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary figures and tables, including the participant timeline and baseline characteristics of those who completed the System Usability Scale.
[DOCX File, 87 KB - formative_v9i1e58192_app1.docx]

Multimedia Appendix 2
Descriptive statistics for satisfaction measures (N=57). All questions were assessed using a satisfaction survey at the end of the trial.
[DOCX File, 16 KB - formative_v9i1e58192_app2.docx]

Multimedia Appendix 3
Participant ratings of the helpfulness of the trial (N=57).
[DOCX File, 15 KB - formative_v9i1e58192_app3.docx]

Checklist 1
CONSORT extension for N-of-1 checklist.
[PDF File, 203 KB - formative_v9i1e58192_app4.pdf]

References

1. Buysse DJ. Sleep health: can we define it? Does it matter? Sleep 2014 Jan 1;37(1):9-17. [doi: 10.5665/sleep.3298] [Medline: 24470692]
2. Czeisler CA. Impact of sleepiness and sleep deficiency on public health--utility of biomarkers. J Clin Sleep Med 2011 Oct 15;7(5 Suppl):S6-S8. [doi: 10.5664/JCSM.1340] [Medline: 22003334]
3. Hisler GC, Muranovic D, Krizan Z. Changes in sleep difficulties among the U.S. population from 2013 to 2017: results from the National Health Interview Survey. Sleep Health 2019 Dec;5(6):615-620. [doi: 10.1016/j.sleh.2019.08.008] [Medline: 31685440]
4. Bixler E. Sleep and society: an epidemiological perspective. Sleep Med 2009 Sep;10 Suppl 1:S3-S6. [doi: 10.1016/j.sleep.2009.07.005] [Medline: 19660985]
5. Yong LC, Li J, Calvert GM. Sleep-related problems in the US working population: prevalence and association with shiftwork status. Occup Environ Med 2017 Feb;74(2):93-104. [doi: 10.1136/oemed-2016-103638] [Medline: 27609325]
6. Fernandez-Mendoza J, Vgontzas AN, Bixler EO, et al. Clinical and polysomnographic predictors of the natural history of poor sleep in the general population. Sleep 2012 May 1;35(5):689-697. [doi: 10.5665/sleep.1832] [Medline: 22547895]
7. Medic G, Wille M, Hemels ME. Short- and long-term health consequences of sleep disruption. Nat Sci Sleep 2017;9(151-61):151-161. [doi: 10.2147/NSS.S134864] [Medline: 28579842]
8. Gottesman RF, Lutsey PL, Benveniste H, et al. Impact of sleep disorders and disturbed sleep on brain health: a scientific statement from the American Heart Association. Stroke 2024 Mar;55(3):e61-e76. [doi: 10.1161/STR.0000000000000453] [Medline: 38235581]
9. Rahe C, Czira ME, Teismann H, Berger K. Associations between poor sleep quality and different measures of obesity. Sleep Med 2015 Oct;16(10):1225-1228. [doi: 10.1016/j.sleep.2015.05.023] [Medline: 26429750]
10. Lou P, Chen P, Zhang L, et al. Relation of sleep quality and sleep duration to type 2 diabetes: a population-based cross-sectional survey. BMJ Open 2012;2(4):e000956. [doi: 10.1136/bmjopen-2012-000956] [Medline: 22872722]
11. Malhotra A, Loscalzo J. Sleep and cardiovascular disease: an overview. Prog Cardiovasc Dis 2009;51(4):279-284. [doi: 10.1016/j.pcad.2008.10.004] [Medline: 19110129]

XSL•FO
RenderX

12. Lee S, Kim JH, Chung JH. The association between sleep quality and quality of life: a population-based study. Sleep Med 2021 Aug;84:121-126. [doi: 10.1016/j.sleep.2021.05.022] [Medline: 34147026]
13. Hafner M, Stepanek M, Taylor J, Troxel WM, van Stolk C. Why sleep matters-the economic costs of insufficient sleep: a cross-country comparative analysis. Rand Health Q 2017 Jan;6(4):11. [Medline: 28983434]
14. Brzezinski A, Vangel MG, Wurtman RJ, et al. Effects of exogenous melatonin on sleep: a meta-analysis. Sleep Med Rev 2005 Feb;9(1):41-50. [doi: 10.1016/j.smrv.2004.06.004] [Medline: 15649737]
15. Tseti IK. Melatonin as a food supplement for sleep disorders. In: Melatonin-The Hormone of Darkness and Its Therapeutic Potential and Perspectives: IntechOpen; 2020.
16. Brzezinski A. Melatonin in humans. N Engl J Med 1997 Jan 16;336(3):186-195. [doi: 10.1056/NEJM199701163360306] [Medline: 8988899]
17. Andersen LPH, Werner MU, Rosenberg J, Gögenur I. A systematic review of peri-operative melatonin. Anaesthesia 2014 Oct;69(10):1163-1171. [doi: 10.1111/anae.12717] [Medline: 24835540]
18. Bourne RS, Mills GH, Minelli C. Melatonin therapy to improve nocturnal sleep in critically ill patients: encouraging results from a small randomised controlled trial. Crit Care 2008;12(2):R52. [doi: 10.1186/cc6871] [Medline: 18423009]
19. Ferracioli-Oda E, Qawasmi A, Bloch MH. Meta-analysis: melatonin for the treatment of primary sleep disorders. PLoS ONE 2013;8(5):e63773. [doi: 10.1371/journal.pone.0063773] [Medline: 23691095]
20. Fatemeh G, Sajjad M, Niloufar R, Neda S, Leila S, Khadijeh M. Effect of melatonin supplementation on sleep quality: a systematic review and meta-analysis of randomized controlled trials. J Neurol 2022 Jan;269(1):205-216. [doi: 10.1007/s00415-020-10381-w] [Medline: 33417003]
21. Suls JM, Alfano C, Yap C. Personalized (N-of-1) trials for patient-centered treatments of multimorbidity. Harv Data Sci Rev 2022;4(SI3). [doi: 10.1162/99608f92.d99e6ff5] [Medline: 38009131]
22. Davidson KW, Silverstein M, Cheung K, Paluch RA, Epstein LH. Experimental designs to optimize treatments for individuals: personalized N-of-1 trials. JAMA Pediatr 2021 Apr 1;175(4):404-409. [doi: 10.1001/jamapediatrics.2020.5801] [Medline: 33587109]
23. Guyatt G. N of 1 randomized trials: a commentary. J Clin Epidemiol 2016 Aug;76(4-5):4-5. [doi: 10.1016/j.jclinepi.2015.09.020]
24. Guyatt GH, Keller JL, Jaeschke R, Rosenbloom D, Adachi JD, Newhouse MT. The n-of-1 randomized controlled trial: clinical usefulness. Our three-year experience. Ann Intern Med 1990 Feb 15;112(4):293-299. [doi: 10.7326/0003-4819-112-4-293] [Medline: 2297206]
25. Duan N, Kravitz RL, Schmid CH. Single-patient (n-of-1) trials: a pragmatic clinical decision methodology for patient-centered comparative effectiveness research. J Clin Epidemiol 2013 Aug;66(8 Suppl):S21-S28. [doi: 10.1016/j.jclinepi.2013.04.006] [Medline: 23849149]
26. Raman G, Balk EM, Lai L, et al. Evaluation of person-level heterogeneity of treatment effects in published multiperson N-of-1 studies: systematic review and reanalysis. BMJ Open 2018 May 26;8(5):e017641. [doi: 10.1136/bmjopen-2017-017641] [Medline: 29804057]
27. Cheung K, Mitsumoto H. Evaluating personalized (N-of-1) trials in rare diseases: how much experimentation is enough? Harv Data Sci Rev 2022;2022(Spec Iss 3). [doi: 10.1162/99608f92.e11adff0] [Medline: 38283317]
28. Gabler NB, Duan N, Vohra S, Kravitz RL. N-of-1 trials in the medical literature: a systematic review. Med Care 2011 Aug;49(8):761-768. [doi: 10.1097/MLR.0b013e318215d90d] [Medline: 21478771]
29. Kravitz RL, Duan N, Niedzinski EJ, Hay MC, Subramanian SK, Weisner TS. What ever happened to N-of-1 trials? Insiders' perspectives and a look to the future. Milbank Q 2008 Dec;86(4):533-555. [doi: 10.1111/j.1468-0009.2008.00533.x] [Medline: 19120979]
30. Task Force for the Development of Practice Recommendations for the Provision of Humanistic Psychosocial Services. Recommended principles and practices for the provision of humanistic psychosocial services: alternative to mandated practice and treatment guidelines. The Humanistic Psychologist 2004;32(1):3-75. [doi: 10.1080/08873267.2004.9961745]
31. Kwasnicka D, Inauen J, Nieuwenboom W, et al. Challenges and solutions for N-of-1 design studies in health psychology. Health Psychol Rev 2019 Jun;13(2):163-178. [doi: 10.1080/17437199.2018.1564627] [Medline: 30626274]
32. Garbarino S, Bragazzi NL. Revolutionizing sleep health: the emergence and impact of personalized sleep medicine. J Pers Med 2024 Jun 4;14(6):598. [doi: 10.3390/jpm14060598] [Medline: 38929819]
33. Ji KH, Yun CH. Emerging technologies to track and improve sleep health. Sleep Med Clin 2025 Mar;20(1):47-55. [doi: 10.1016/j.jsmc.2024.10.011] [Medline: 39894598]
34. Butler M, D'Angelo S, Perrin A, et al. A series of remote melatonin supplement interventions for poor sleep: protocol for a feasibility pilot study for a series of personalized (N-of-1) trials. JMIR Res Protoc 2023 Aug 3;12(1):e45313. [doi: 10.2196/45313] [Medline: 37535419]
35. Internet, broadband fact sheet. Pew Research Center. 2021. URL: https://www.pewresearch.org/internet/fact-sheet/internet-broadband [accessed 2023-12-15]
36. About Northwell. Northwell Health. 2022. URL: https://www.northwell.edu/about-northwell#:~:text=At%20Northwell%20Health%2C%20our%2078%2C000,of%20health%20care%20for%20all [accessed 2025-12-15]

XSL•FO
RenderX

37. Okun ML, Kravitz HM, Sowers MF, Moul DE, Buysse DJ, Hall M. Psychometric evaluation of the insomnia symptom questionnaire: a self-report measure to identify chronic insomnia. J Clin Sleep Med 2009 Feb 15;5(1):41-51. [Medline: 19317380]
38. Epstein LJ, Cai A, Klerman EB, Czeisler CA. Resolving delayed sleep-wake phase disorder with a pandemic: two case reports. J Clin Sleep Med 2022 Jan 1;18(1):315-318. [doi: 10.5664/jcsm.9526] [Medline: 34254935]
39. Nishihara T, Hashimoto S, Ito K, et al. Oral melatonin supplementation improves oocyte and embryo quality in women undergoing in vitro fertilization-embryo transfer. Gynecol Endocrinol 2014 May;30(5):359-362. [doi: 10.3109/09513590.2013.879856] [Medline: 24628045]
40. Sletten TL, Magee M, Murray JM, et al. Efficacy of melatonin with behavioural sleep-wake scheduling for delayed sleep-wake phase disorder: a double-blind, randomised clinical trial. PLoS Med 2018 Jun;15(6):e1002587. [doi: 10.1371/journal.pmed.1002587] [Medline: 29912983]
41. Zhdanova IV, Wurtman RJ, Lynch HJ, et al. Sleep-inducing effects of low doses of melatonin ingested in the evening*. Clin Pharmacol Ther 1995 May;57(5):552-558. [doi: 10.1016/0009-9236(95)90040-3]
42. Cardinali DP, Srinivasan V, Brzezinski A, Brown GM. Melatonin and its analogs in insomnia and depression. J Pineal Res 2012 May;52(4):365-375. [doi: 10.1111/j.1600-079X.2011.00962.x] [Medline: 21951153]
43. Mendelson WB. A critical evaluation of the hypnotic efficacy of melatonin. Sleep 1997 Oct;20(10):916-919. [doi: 10.1093/sleep/20.10.916] [Medline: 9415955]
44. Butler M, D'Angelo S, Lewis C, et al. Series of virtual light therapy interventions for fatigue: a feasibility pilot study protocol for a series of personalised (N-of-1) trials. BMJ Open 2022 Oct 25;12(10):e055518. [doi: 10.1136/bmjopen-2021-055518] [Medline: 36283748]
45. Butler M, D'Angelo S, Kaplan M, et al. A series of virtual interventions for chronic lower back pain: a feasibility pilot study for a series of personalized (N-of-1) trials. Harv Data Sci Rev 2022;4(SI3). [doi: 10.1162/99608f92.72cd8432] [Medline: 37609556]
46. D'Angelo S, Ahn H, Miller D, Monane R, Butler M. Personalized feedback for personalized trials: construction of summary reports for participants in a series of personalized trials for chronic lower back pain. Harv Data Sci Rev 2022;4(SI3). [doi: 10.1162/99608f92.d5b57784] [Medline: 38009134]
47. Brooke J. Sus: a "quick and dirty" usability. In: Usability Evaluation in Industry, 1st edition: CRC Press. [doi: 10.1201/9781498710411-35]
48. Brooke J. SUS: a retrospective. J Usability Stud 2013;8(2):29-40 [FREE Full text]
49. Lewis JR. The system usability scale: past, present, and future. International Journal of Human–Computer Interaction 2018 Jul 3;34(7):577-590. [doi: 10.1080/10447318.2018.1455307]
50. Butler M, D'Angelo S, Ahn H, et al. A series of personalized virtual light therapy interventions for fatigue: feasibility randomized crossover trial for N-of-1 treatment. JMIR Form Res 2023 Sep 18;7(7):e45510. [doi: 10.2196/45510] [Medline: 37721795]
51. Bangor A, Kortum PT, Miller JT. An empirical evaluation of the system usability scale. Int J Hum Comput Interact 2008 Jul 29;24(6):574-594. [doi: 10.1080/10447310802205776]
52. Haghayegh S, Khoshnevis S, Smolensky MH, Diller KR, Castriotta RJ. Accuracy of wristband Fitbit models in assessing sleep: systematic review and meta-analysis. J Med Internet Res 2019 Nov 28;21(11):e16273. [doi: 10.2196/16273] [Medline: 31778122]
53. Carney CE, Buysse DJ, Ancoli-Israel S, et al. The consensus sleep diary: standardizing prospective sleep self-monitoring. Sleep 2012 Feb 1;35(2):287-302. [doi: 10.5665/sleep.1642] [Medline: 22294820]
54. Jensen MP, Turner JA, Romano JM, Fisher LD. Comparative reliability and validity of chronic pain intensity measures. Pain 1999 Nov;83(2):157-162. [doi: 10.1016/s0304-3959(99)00101-3] [Medline: 10534586]
55. Personalized (N-of-1) trial design tools. roadmap2health. 2022. URL: https://roadmap2health.io/cmi/ [accessed 2023-12-15]
56. Smilowska K, van Wamelen DJ, Schoutens AMC, Meinders MJ, Bloem BR. Blue light therapy glasses in Parkinson's disease: patients' experience. Parkinsons Dis 2019;2019:1906271. [doi: 10.1155/2019/1906271] [Medline: 31316746]
57. Văcăre u T, Batalas N, Erten-Uyumaz B, van Gilst M, Overeem S, Markopoulos P. Subjective sleep quality monitoring with the hypnos digital sleep diary: evaluation of usability and user experience. Presented at: Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies; Feb 22-24, 2019; Prague, Czech Republic. [doi: 10.5220/0007407601130122]
58. Oftedal S, Burrows T, Fenton S, Murawski B, Rayward AB, Duncan MJ. Feasibility and preliminary efficacy of an m-health intervention targeting physical activity, diet, and sleep quality in shift-workers. Int J Environ Res Public Health 2019 Oct 10;16(20):3810. [doi: 10.3390/ijerph16203810] [Medline: 31658624]
59. Suarez-Giron M, Garmendia O, Lugo V, et al. Mobile health application to support CPAP therapy in obstructive sleep apnoea: design, feasibility and perspectives. ERJ Open Res 2020 Jan;6(1):00220-2019. [doi: 10.1183/23120541.00220-2019] [Medline: 32055630]
60. Marquard JL, Howard J, LeBlanc R. Using novel data visualization methods to understand mobile health usability: exemplar from a technology-enabled sleep self-monitoring intervention. Comput Inform Nurs 2023 Jan 1;41(1):1-5. [doi: 10.1097/CIN.0000000000000970] [Medline: 36634231]

XSL•FO
RenderX

61. Trenaman L, Munro S, Almeida F, Ayas N, Hicklin J, Bansback N. Development of a patient decision aid prototype for adults with obstructive sleep apnea. Sleep Breath 2016 May;20(2):653-661. [doi: 10.1007/s11325-015-1269-9] [Medline: 26449553]
62. Aupetit S, Dubroca G, Escaich S, Cabon P. A qualitative study of sleep trackers usage: evidence of orthosomnia. CIEHF 2019:1-7 [FREE Full text]
63. Liang J, Xian D, Liu X, et al. Usability study of mainstream wearable fitness devices: feature analysis and system usability scale evaluation. JMIR Mhealth Uhealth 2018 Nov 8;6(11):e11066. [doi: 10.2196/11066] [Medline: 30409767]
64. Hoaglin DC. Misunderstandings about Q and "Cochran's Q test" in meta-analysis. Stat Med 2016 Feb 20;35(4):485-495. [doi: 10.1002/sim.6632] [Medline: 26303773]
65. Butler M, D'Angelo S, Ahn H, et al. A series of personalized virtual light therapy interventions for fatigue: feasibility randomized crossover trial for N-of-1 treatment. JMIR Form Res 2023 Sep 18;7(1):e45510. [doi: 10.2196/45510] [Medline: 37721795]
66. Butler MJ, Arader L, Chandereng T, d'Angelo S, et al. Heterogeneity in treatment response to yoga and massage interventions for chronic lower back pain: results from a series of personalized N-of-1 trials (under review). Appl Psychophysiol Biofeedback 2023. [doi: 10.1162/99608f92.72cd8432]
67. Costello RB, Lentino CV, Boyd CC, et al. The effectiveness of melatonin for promoting healthy sleep: a rapid evidence assessment of the literature. Nutr J 2014 Nov 7;13:106. [doi: 10.1186/1475-2891-13-106] [Medline: 25380732]
68. Patel SR. Social and demographic factors related to sleep duration. Sleep 2007 Sep;30(9):1077-1078. [doi: 10.1093/sleep/30.9.1077] [Medline: 17910376]
69. Parish JM. Sleep-related problems in common medical conditions. Chest 2009 Feb;135(2):563-572. [doi: 10.1378/chest.08-0934] [Medline: 19201722]
70. Halperin D. Environmental noise and sleep disturbances: a threat to health? Sleep Sci 2014 Dec;7(4):209-212. [doi: 10.1016/j.slsci.2014.11.003] [Medline: 26483931]
71. Hale L. Who has time to sleep? J Public Health (Bangkok) 2005 Jun 1;27(2):205-211. [doi: 10.1093/pubmed/fdi004]
72. Johnson DA, Billings ME, Hale L. Environmental determinants of insufficient sleep and sleep disorders: implications for population health. Curr Epidemiol Rep 2018 Jun;5(2):61-69. [doi: 10.1007/s40471-018-0139-y] [Medline: 29984131]
73. Chandereng T, Liao Z, D'Angelo S, Butler M, Davidson KW, Cheung YK. Role of digital healthcare approaches in the analysis of personalized (n-of-1) trials. In: Personal Health Informatics: Patient Participation in Precision: Springer; 2022:131-146. [doi: 10.1007/978-3-031-07696-1_7]
74. Krystal AD, Edinger JD. Measuring sleep quality. Sleep Med 2008 Sep;9 Suppl 1:S10-S17. [doi: 10.1016/S1389-9457(08)70011-X] [Medline: 18929313]
75. Hirshkowitz M, Whiton K, Albert SM, et al. National Sleep Foundation's updated sleep duration recommendations: final report. Sleep Health 2015 Dec;1(4):233-243. [doi: 10.1016/j.sleh.2015.10.004] [Medline: 29073398]
76. Grandner MA, Hale L, Moore M, Patel NP. Mortality associated with short sleep duration: the evidence, the possible mechanisms, and the future. Sleep Med Rev 2010 Jun;14(3):191-203. [doi: 10.1016/j.smrv.2009.07.006] [Medline: 19932976]
77. Centers for Disease Control and Prevention (CDC). Effect of short sleep duration on daily activities--United States, 2005-2008. MMWR Morb Mortal Wkly Rep 2011 Mar 4;60(8):239-242. [Medline: 21368739]
78. Erland LAE, Saxena PK. Melatonin natural health products and supplements: presence of serotonin and significant variability of melatonin content. J Clin Sleep Med 2017 Feb 15;13(2):275-281. [doi: 10.5664/jcsm.6462] [Medline: 27855744]
79. Li J, Somers VK, Xu H, Lopez-Jimenez F, Covassin N. Trends in use of melatonin supplements among US adults, 1999-2018. JAMA 2022 Feb 1;327(5):483-485. [doi: 10.1001/jama.2021.23652] [Medline: 35103775]
80. Personalized trial to improve sleep quality. OSF. URL: https://osf.io/sk59n/ [accessed 2025-09-11]

ABBREVIATIONS

CONSORT: Consolidated Standards of Reporting Trials
CSD: consensus sleep diary
EMA: ecological momentary assessment
HTE: heterogeneity of treatment effects
mHealth: mobile health
SUS: System Usability Scale

XSL•FO
RenderX

Edited by A Mavragani; submitted 11.03.24; peer-reviewed by A Hassan, B Cahill, J Mistry, S Ranjbaran; revised version received 24.04.25; accepted 05.05.25; published 26.09.25.

Please cite as:
Butler MJ, Chandereng T, Ahn H, Slotnick S, Miller D, Perrin A, Rodillas J, Friel CP, Goodwin AM, Cheung YK, Davidson KW
A Series of Personalized Melatonin Supplement Interventions for Poor Sleep: Feasibility Randomized Crossover Trial for Personalized N-of-1 Treatment
JMIR Form Res 2025;9:e58192
URL: https://formative.jmir.org/2025/1/e58192
doi:10.2196/58192

© Mark J Butler, Thevaa Chandereng, Heejoon Ahn, Stefani Slotnick, Danielle Miller, Alexandra Perrin, Jordyn Rodillas, Ciaran P Friel, Ashley M Goodwin, Ying Kuen Cheung, Karina W Davidson. Originally published in JMIR Formative Research (https://formative.jmir.org), 26.9.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Health Care Content and Engagement in Chronic Illness Instagram Posts: Content Analysis

Lindsey D Daon[1,2], MSc, MD; Fu-Shiuan Whitney Lee[2], MPH, MD; Olga Saynina[3], MBA; C Jason Wang[2,3], MD, PhD

[1]Division of General Pediatrics and Adolescent Medicine, Department of Pediatrics and Adolescent Medicine, Mayo Clinic, 200 First Street SW, Rochester, MN, United States

[2]Department of Pediatrics, Stanford University School of Medicine, Stanford, CA, United States

[3]Department of Health Policy, Stanford University, Stanford, CA, United States

Corresponding Author:
Lindsey D Daon, MSc, MD
Division of General Pediatrics and Adolescent Medicine, Department of Pediatrics and Adolescent Medicine, Mayo Clinic, 200 First Street SW, Rochester, MN, United States

Abstract

Background: Instagram and other social media platforms provide a unique environment for people with chronic illnesses to share experiences, but posts with higher engagement may also shape behavior. The hashtag #ChronicIllness appears in over 5 million posts, reflecting the large digital community where users seek validation, connection, and support. Frameworks such as social cognitive theory, self-presentation theory, and illness identity theory suggest that highly engaging content can shift social norms and drive behavior change via observational learning. Despite the strong theoretical basis for this behavioral impact, little is known about what chronic illness–related content is the most engaging.

Objective: The aim of this study is to identify the content of Instagram posts related to chronic illness and determine if health care content is associated with higher engagement.

Methods: This study is a mixed methods content analysis of 279 publicly available Instagram posts tagged with #chronicillness, #chronicallyill, or #spoonie. Posts were selected via convenience sampling and included if they featured original, nonvideo content. Photos, hashtags, and captions were coded for themes including location, medical equipment, health care experience, and illness identity. Quantitative metrics, such as likes, comments, and overperforming scores (a normalized metric of engagement), were extracted using CrowdTangle. Multivariate analyses assessed if health care content (posts featuring health care experiences or photos in a medical setting or with medical equipment) was associated with a higher odds of overperforming.

Results: Posts had a median of 25 (IQR 0-14,936) likes, 3 (IQR 0-525) comments, and 20 (IQR 1-31) hashtags. A total of 222 (80%) posts were created by women, and 110 (40%) were overperforming. Photo analysis (260 posts with 406 photos) showed 27 (10%) in health care settings, and 49 (19%) included medical equipment, with 10 (4%) featuring invasive devices (eg, intravenous lines and feeding tubes), which were strongly associated with higher engagement. Hashtag analysis revealed that 243 (87%) posts referenced a medical condition, most commonly chronic pain (n=101, 36%), fibromyalgia (n=56, 20%), and Ehlers-Danlos syndrome (n=38, 14%), while 57 (20%) included medical interventions. Captions reflected 4 main themes: medical experience, illness journey, connection, and nonillness experiences. In multivariate regression analysis, longer captions (odds ratio [OR] 2.44, 95% CI 1.05 - 5.67), health care content (OR 1.85, 95% CI 1.00 - 3.42), and invasive medical equipment (OR 6.19, 95% CI 1.16 - 32.99) were independently associated with overperforming.

Conclusions: Posts featuring health care content and invasive medical equipment were associated with significantly more engagement, suggesting that medicalized portrayals of illness may be amplified on Instagram. This visibility may offer support but also risks reinforcing illness-centered identities and overmedicalization through the influence of observational learning and identity formation. Medical professionals must be aware of these trends and promote balanced, evidence-based content. Future research should explore how social media shapes health behaviors, identity, and utilization to mitigate potential harms while preserving support.

(JMIR Form Res 2025;9:e57523) doi:10.2196/57523

KEYWORDS

social media; chronic illness; content analysis; Instagram; infodemics; web-based communities; digital health; social presentation theory; social cognitive theory; illness identity theory

XSL•FO
RenderX

Introduction

#Chronicillness is trending on social media, with over 5 million posts on Instagram alone [1]. Social media platforms, like Instagram, TikTok, and Facebook, serve as a space for individuals to share personal health experiences, connect with others, and seek validation. While these platforms offer support and self-empowerment opportunities, they also shape how chronic illness is perceived and presented, reinforcing certain narratives over others and influencing illness identity and behaviors. Three theoretical frameworks can be used to understand social media's influence on behavior: social cognitive theory (SCT), self-presentation theory (SPT), and illness identity theory (IIT).

For viewers of chronic illness content, SCT describes how social media serves as a catalyst for behavior change. This occurs through reciprocal determinism, the interaction between personal characteristics, behavioral patterns, and environmental influences, particularly through observational learning [2]. This process, which involves attention, retention, reproduction, and reinforcement, shapes how viewers engage with health-related digital content [3]. Social media users span all age groups, with many individuals turning to these platforms for health-related information [4-6]. As web-based health communities grow, they provide both valuable peer support and potential pathways for symptom reinforcement and behavior change [7-9]. Ziebland and Wyke [10] raise concerns that the information people learn from social media could be used to overexaggerate symptoms to manipulate health care encounters, underscoring the need to understand the content shared on social media and its potential influence on health behavior.

For content creators, SPT can be used to understand how social media influences their behavior by shaping the way they curate and share their experiences. SPT explains how individuals attempt to control their public image through the selective presentation of behaviors, appearance, and communication [11]. Social media amplifies this effect as creators can add, remove, or edit content at will. This allows creators to present themselves in ways that maximize engagement and positive feedback, reinforcing certain narratives over others [12]. This narrative reinforcement can have cascading effects, as Masur et al [13] found that repeated exposure to behaviors on social media can change perceived norms and influence social media viewers' self-disclosure patterns as well.

Both frameworks help explain how both creators and viewers may become more deeply embedded in an illness identity. According to IIT, chronic illness can become a core part of one's self-concept, behavior, and self-perception across four dimensions [14]. The dimensions of acceptance and enrichment are associated with better health outcomes, while rejection and engulfment are associated with greater distress and maladaptive coping [15,16]. A study by Gabarron et al [17] highlights the concern about social media's influence on illness identity, finding that users are least engaged with content that fosters empowerment, such as health education and research, and most with personal narratives. This suggests that users are more drawn to relatable, emotionally compelling content that may reinforce illness-centered narratives rather than recovery-oriented perspectives. The content of such narratives becomes important in the context of social media algorithms, which boost the most engaging content, further shaping chronic illness narratives [18].

The emergence of illness-centered digital identities is not new. In 2013, the term "Spoonie" was popularized through a viral blog post, becoming a widely recognized identity for those with chronic conditions [19]. While this sense of community can be empowering, it also raises concerns about the reinforcement of illness identity. Media reports have highlighted cases in which individuals exaggerate symptoms or seek unnecessary medical interventions in pursuit of validation [20,21]. Beyond anecdotal reports, research has documented the psychogenic spread of illness through social media, such as with functional tics and dissociative identity disorder, where symptoms increased among social media users following viral exposure [22,23]. Similarly, studies have found a rise in self-diagnosis trends for mental health conditions [24,25] and the digital amplification of various symptoms [26-28].

With these rising concerns, it is essential to examine the types of chronic illness–related content shared on web-based platforms and which narratives receive the most attention, represented by user engagement. Attention is a key driver of observational learning and self-presentation, reinforcing certain behaviors and narratives. Highly visible content is more likely to be seen, remembered, replicated, and incorporated into one's identity. Understanding existing content and its engagement patterns is crucial to recognizing the behaviors that people who self-identify as having chronic illness are exposed to and may adopt. This study aims to analyze Instagram posts related to chronic illness, examining both the content being posted and that which receives the most user engagement.

Methods

Study Design

This was a mixed methods study, combining qualitative and quantitative methods, to conduct a multimodal content analysis of Instagram posts. The analysis involved the qualitative examination of photos, hashtags, captions, and quantitative metrics, such as overperforming scores, likes, comments, and followers. Figure 1 provides a visual representation of the components of an Instagram post.

XSL•FO
RenderX

Figure 1. Recreated Instagram post using a stock photo, with the key components labeled (eg, caption, hashtags, photo, likes, and comments).

Sampling and Data Collection

The units of analysis for this study are Instagram posts, which were selected using convenience sampling. Convenience sampling was necessary due to the large volume of Instagram posts and the desire to focus on more recent posts as content on social media changes over time. Three search terms, that is, hashtags, were selected for their relevance to individuals with chronic illness, their broad representation of chronic illness rather than specific diseases, and the substantial number of posts associated with each hashtag: #chronicillness, #chronicallyill, and #spoonie.

To minimize the influence of algorithms on postselection, two independent reviewers (LDD and FSWL) conducted separate searches on an arbitrarily chosen date (October 26, 2021) and chose qualifying posts from the "Most Recent" section of the search results. Posts were included if they were publicly available and appeared to be original content, such as photos, artwork, or products being promoted. Posts were excluded if they contained videos or the photos were nonoriginal content, such as memes or reposted content from another person's account. Each reviewer selected posts in reverse chronological order until 100 posts for each hashtag were included, resulting in a total of 300 posts per reviewer. The reviewers cross-compared the selected posts to ensure that only common posts were included, removed duplicates, and resolved inclusion or exclusion inconsistencies, resulting in a final sample of 279 posts.

The data for analysis, including caption text, hashtags, number of likes, and overperforming score, were downloaded from CrowdTangle and cleaned using Excel (Microsoft Corp), 10 days after identifying the posts. This 10-day delay allowed the posts to receive views, the bulk of which occurred within the first few days after being posted. During data cleaning, the number of followers and country location for each account were verified by reviewing the associated account profiles and posts. These data were used for the descriptive statistics (eg, likes, comments, total interactions, followers, and number of photos) for all 279 posts. Photos were not included in the CSV file and were analyzed directly on the CrowdTangle website in March 2022. By the time photo analysis occurred, 19 posts had been removed or made private; thus, only 260 posts underwent photo analysis and inclusion in the multivariate analysis.

Qualitative Analysis

The qualitative analysis involved multimodal content analysis of Instagram posts, including photos, hashtags, and captions. Multiple elements of the posts were used to determine gender and identify the presence of advertising content. Gender identification involved visually inspecting the photos and assessing the captions or hashtags for gender-specific language. Advertising content was defined as any promotional material related to businesses or products that have the potential to generate income, such as product shops, coaching services, podcasts, or blogs. Two coders (LDD and FSWL) collaborated to analyze the content of the posts.

Photos for each post underwent visual content analysis simultaneously by the two coders. Each photo was coded for

XSL•FO
RenderX

location, presence of medical equipment, presence of a support animal, and number of people. Location was coded by medical (eg, hospital and clinic) or nonmedical (eg, yard, neighborhood, and house). Medical equipment was coded by invasive (eg, intravenous lines [IVs], peripherally inserted central catheters [PICCs], and feeding tubes) or noninvasive (eg, medications, nasal cannula, and mobility aids). When multiple photos were present in a post, if any photo within a post contained the theme of interest, the entire post was considered to have that theme for analysis purposes.

Hashtags underwent quantitative content analysis. They were coded for mentions of specific conditions and medical interventions (Table 1). A hashtag was considered a condition if it included a specific diagnosis (eg, chronic fatigue syndrome) but not a general symptom (eg, fatigue). If a hashtag could be considered either a diagnosis or a symptom (eg, anemia, headaches, gastroparesis, and anxiety), it was coded as a diagnosis. Medical interventions included procedures, testing, and treatments.

XSL•FO
RenderX

Table . Summary of key findings from photo and hashtag content analysis[a].

	Value, n (%)
Photo analysis (n=260)	
Location of photo	
Health care setting (ie, clinic or hospital)	27 (10.4)
Non–health care setting (ie, home and outdoors[b])	199 (76.5)
Location indeterminate (ie, art and slides)	34 (13.1)
Medical equipment	49 (18.8)
Invasive medical equipment[c]	10 (3.8)
Noninvasive medical equipment[d]	39 (15)
Hashtag analysis (n=279)	
Condition[e]	243 (87.1)
Chronic pain	101 (36.2)
Fibromyalgia	56 (20.1)
Autoimmunity	38 (13.6)
Ehlers-Danlos syndrome (EDS)	38 (13.6)
Postural orthostatic tachycardia syndrome (POTS)	36 (12.9)
Anxiety	31 (11.1)
Myalgic encephalitis/chronic fatigue syndrome (ME/CFS)	27 (9.6)
Dysautonomia	25 (9)
Depression	22 (7.9)
Endometriosis or adenomyosis	21 (7.5)
Gastroparesis	20 (7.2)
Inflammatory bowel disease (IBD)	20 (7.2)
Arthritis	15 (5.4)
Headache	14 (5)
Lupus	13 (4.7)
Multiple sclerosis (MS)	12 (4.3)
Post traumatic stress disorder (PTSD)	11 (3.9)
Lyme disease	10 (3.6)
Medical Intervention	57 (20.4)
Mobility aids	25 (9)
Medications	16 (5.7)
Invasive procedures	15 (5.4)
Feeding tubes	11 (3.9)

[a]Photos from 260 posts were included in the analysis. Hashtags from all 279 posts were included in the analysis.

[b]Including nature, yards, neighborhoods, and cities.

[c]Including peripheral intravenous (PIV) lines, peripherally inserted central catheter (PICC) lines, ports, and feeding tubes.

[d]Including medications, bandages, blood vials, mobility aids (canes, wheelchairs, walkers), compression socks, hospital beds, hospital gowns, telemetry wires, nasal cannulas, oxygen tanks, diagnostic images, and visual symptoms of illness.

[e]107 different conditions were identified. Conditions that were mentioned in 10 or more posts are listed.

Captions underwent content analysis using an iterative and inductive process. The initial categories and codes were based on a first pass review of the included post captions, illness identity framework, and the research question, and included health care–related, disease-related, personal feelings, interactions with others, life events or beliefs, and illness identity. The two reviewers then simultaneously coded the first 20 posts to ensure mutual agreement and understanding of the

XSL•FO
RenderX

categories and emerging codes. The remaining posts were divided between the two team members for independent coding. As new codes emerged or at least every 50 posts, the reviewers met frequently to review each other's work, refine categories and codes, and resolve inconsistencies. If a consensus was unable to be reached, a third team member (CJW) was used as a tiebreaker. The final list of categories and codes can be seen in Figure 2.

Figure 2. The post captions were independently coded by 2 reviewers using an iterative and inductive process, resulting in 4 major themes with associated subthemes. Coding discrepancies were discussed and resolved through consensus.

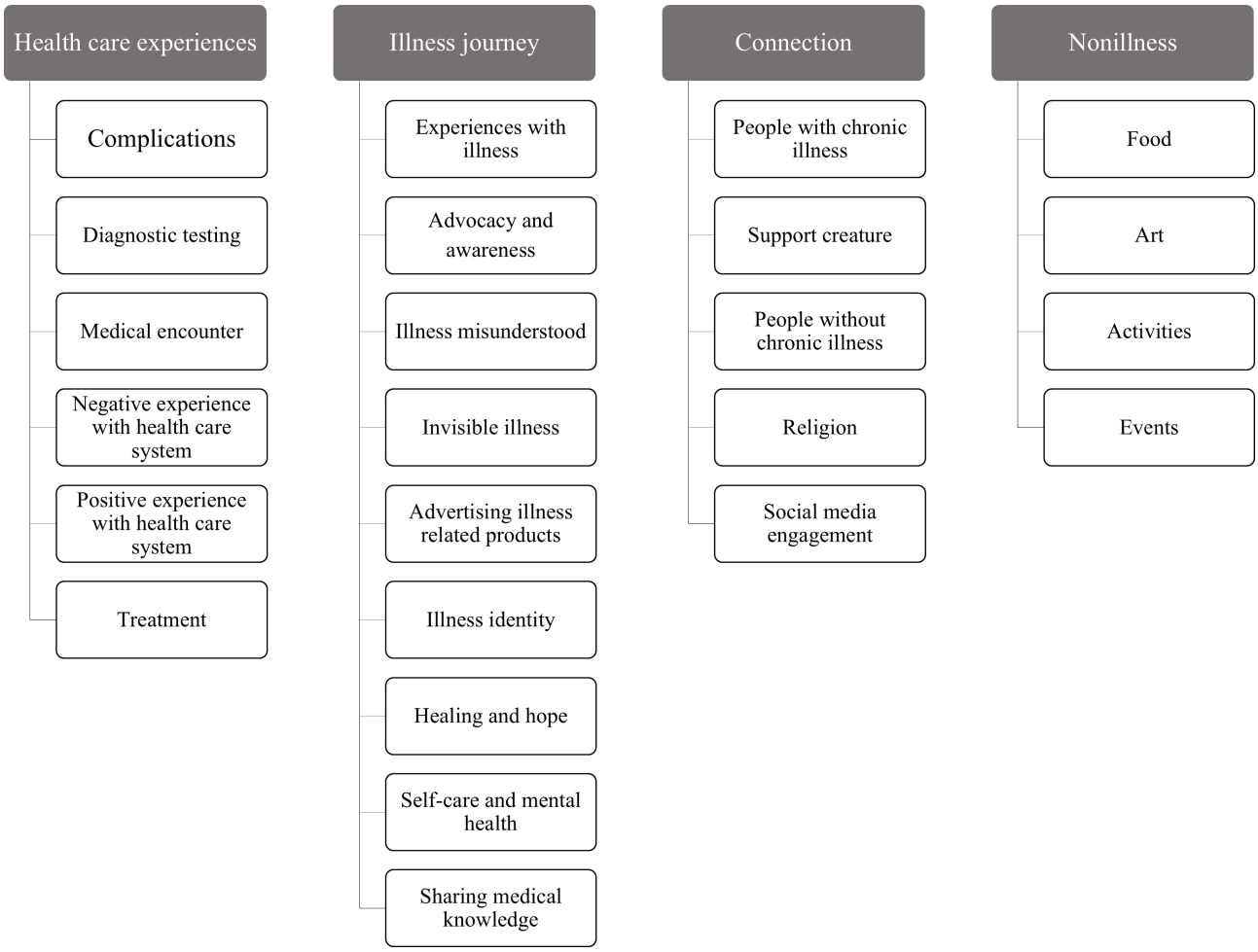

Quantitative Analysis

The overperforming score [29], calculated by CrowdTangle, is a metric that compares a post's engagement (likes, comments, and shares) to posts from the same account. This allows for normalization of engagement across accounts, reducing the advantage of accounts with more followers or frequent posts. A score greater than 1 indicates overperformance, a score less than 0 suggests underperformance, and scores between 0 and 1 are considered indeterminate. Posts that overperform suggest higher than typical engagement on that post, highlighting content that resonates more with audiences.

For this study, overperforming was used as a proxy for user engagement, as it accounts for both absolute engagement and relative performance within an account's posting history. This approach helps mitigate the effects of larger accounts naturally receiving more engagement and allows for a more meaningful comparison of content performance.

The overperforming score was used as the primary dependent variable to investigate the association between health care content and engagement. Health care content is a summary variable that counted posts that included photos with medical equipment (ie, peripherally inserted central catheter lines, feeding tubes, IVs, mobility assistive devices, and medications), medical location (ie, hospital and clinic), or a description of a specific health care encounter (ie, clinic visit, emergency department, and starting a new medication).

Bivariate and multivariate models were conducted by the data analyst (OS) to examine the dependent variable (overperforming score) in relation to several independent variables, outlined in Table 2. For the multivariate analysis, the overperforming score was transformed into a binary variable, categorizing a score of 1 or greater as "overperforming" and scores less than 1 as "not overperforming."

XSL•FO
RenderX

Table . OR[a] bivariate and multivariate analyses.

Variable	Bivariate, OR (95% CI)	Model 1, OR (95% CI)	Model 2, OR (95% CI)	Model 3, OR (95% CI)
People present in photo	1.511 (0.9-2.537)	1.208 (0.655-2.227)	1.181 (0.645-2.162)	1.195 (0.647-2.207)
Type of post (photo vs album)	1.653 (0.957-2.856)	1.47 (0.763-2.831)	1.496 (0.778-2.875)	1.315 (0.673-2.566)
Hashtag: quartile 2 versus 1	1.333 (0.68-2.612)	1.527 (0.663-3.518)	1.503 (0.655-3.448)	1.874 (0.789-4.446)
Hashtag: quartile 3 versus 1	1.379 (0.706-2.693)	2.156 (0.918-5.066)	1.955 (0.844-4.528)	2.296 (0.961-5.484)
Hashtag: Quartile 4 versus 1	1.24 (0.62-2.481)	1.136 (0.464-2.782)	1.055 (0.435-2.558)	1.24 (0.494-3.113)
Caption length: quartile 2 versus 1	1.304 (0.637-2.668)	1.186 (0.496-2.836)	1.167 (0.493-2.767)	1.137 (0.475-2.721)
Caption length: quartile 3 versus 1	1.875 (0.929-3.783)	2.048 (0.868-4.829)	2.112 (0.902-4.948)	2.145 (0.909-5.064)
Caption length: quartile 4 versus 1	2.727 (1.352-5.499)[b]	2.02 (0.846-4.824)	2.438 (1.047-5.674)[b]	2.179 (0.925-5.135)
Number of condition: 3 - 4 versus 1 - 2	0.874 (0.479-1.596)	0.591 (0.288-1.211)	0.655 (0.324-1.325)	0.628 (0.307-1.284)
Number of condition: 5+ versus 1 - 2	0.876 (0.451-1.703)	0.646 (0.298-1.404)	0.748 (0.352-1.591)	0.731 (0.34-1.571)
Gender: male versus female	0.267 (0.057-1.249)	0.292 (0.057-1.502)	0.27 (0.054-1.354)	0.279 (0.056-1.391)
Gender: other versus female	0.543 (0.27-1.091)	0.404 (0.127-1.285)	0.39 (0.123-1.239)	0.425 (0.134-1.345)
Number of followers: quartile 2 versus 1	1.061 (0.541-2.083)	0.955 (0.416-2.191)	0.983 (0.432-2.236)	1.051 (0.456-2.422)
Number of followers: quartile 3 versus 1	0.942 (0.478-1.856)	0.852 (0.366-1.984)	0.813 (0.352-1.88)	0.868 (0.371-2.034)
Number of followers: quartile 4 versus 1	0.907 (0.458-1.795)	0.741 (0.309-1.781)	0.727 (0.305-1.733)	0.788 (0.327-1.899)
Health care content summary[c]	2.025 (1.224-3.349)[b]	1.854 (1.004-3.421)[b]	—[d]	—
Photo with noninvasive medical equipment	1.213 (0.609-2.414)	—	1.005 (0.474-2.134)[b]	—
Photo with invasive medical equipment	6.526 (1.357-31.382)[b]	—	—	6.187 (1.16-32.991)[b]
Photo in medical location	1.248 (0.559-2.788)	—	—	—
Medical experience	1.831 (1.078-3.11)[b]	—	—	—
Condition discussed	1.737 (0.977-3.091)	—	—	—
Support object	0.697 (0.324-1.498)	—	—	—
Nonillness content[e]	0.664 (0.393-1.121)	—	—	—
Advertisement	0.7 (0.37-1.323)	—	—	—

[a]OR: odds ratio.

[b]Statistically significant odds ratio with 95% CI.

[c]Summary variable that includes photo with medical equipment (invasive or noninvasive), photo in a medical location, or a caption containing the category Health Care Experience.

[d]Indicates a controlled-for variable in each model.

[e]Nonillness content if photo or caption did not include illness journey or health care content.

Ethical Considerations

Per Stanford University's institutional review board, this study, which examines publicly available Instagram posts, was exempt from the institutional review board approval and informed consent as it does not meet the criteria for human participant research. No direct interaction with users occurred, and all data were publicly accessible and deidentified during the cleaning process for quantitative analysis. While photos could not be fully deidentified due to their role in visual coding, all coded

XSL•FO
RenderX

data were anonymized and stored without user identifiers. To protect privacy in visual presentation, all illustrative images used in figures were recreated using licensed Shutterstock photos and adapted to mimic the structure and content of actual posts; no original user photos were published.

Results

Post Characteristics

A total of 279 posts were included in the descriptive statistics analysis (Table 3) and the caption and hashtag content analysis. Among these, only 260 posts containing 406 photos were used for the visual analysis and multivariate analysis (Figure 3). The median number of followers at the time of data collection was 661 (IQR 326-1624). Posts received a median of 25 (IQR 12-53) likes and 3 (IQR 1-8) comments. The median number of hashtags per post was 20 (IQR 12-28). Of the analyzed posts, 110 (40%) were considered overperforming, and 52 (20%) contained marketing content. Most posts were authored by women (n=222, 80%) and featured a single photo (n=209, 75%). Age information was unavailable.

Table . Characteristics of the 279 posts included in the analysis, summarizing the metadata, creator demographics, and content features.

Characteristic of post	Value (n=279)
Hashtags (total), median (IQR)	20 (12-28)
Caption length (in characters), median (IQR)	596 (396-1053)
Conditions (total in hashtags), median (IQR)	2 (1-4)
Followers (total), median (IQR)	661 (326–1624)
Likes (total), median (IQR)	25 (12-53)
Comments (total), median (IQR)	3 (1-8)
Interactions (total), median (IQR)	29 (14-63)
Gender, n (%)	
Women	222 (79.6)
Men	12 (4.3)
Other or unknown	45 (16.1)
Country of post, n (%)	
United States	51 (18.2)
United Kingdom	16 (5.7)
Others[a]	18 (6.4)
Unknown	194 (69.5)
Content features, n (%)	
Overperforming status	110 (39.4)
Health care content present	97 (34.7)
Advertising present[b]	52 (18.6)
Type of post—album[c]	70 (25.1)

[a]Canada, Mexico, Ireland, Sweden, the Netherlands, South Africa, India, Australia, and New Zealand.

[b]Advertisements included podcasts, blogs, supplements, hyperbaric oxygen therapy, health coaching services, money donation accounts, and products from digital shops such as “tubie” supplies and chronic illness pins or stickers or graphic tees or books.

[c]Albums are posts that contain multiple photos.

XSL•FO
RenderX

Figure 3. Post selection process flowchart. An initial search of the 3 hashtags on October 16, 2021, revealed millions of posts per search term. The two reviewers used convenience sampling, from newest to oldest, to select posts for inclusion, resulting in a total of 300 posts per reviewer. After cross-comparison, duplicate removal, and exclusion resolution, 279 posts were included in the final dataset. Photos were analyzed separately in March 2022, with 260 posts available at that time for visual analysis and inclusion in multivariate analysis.

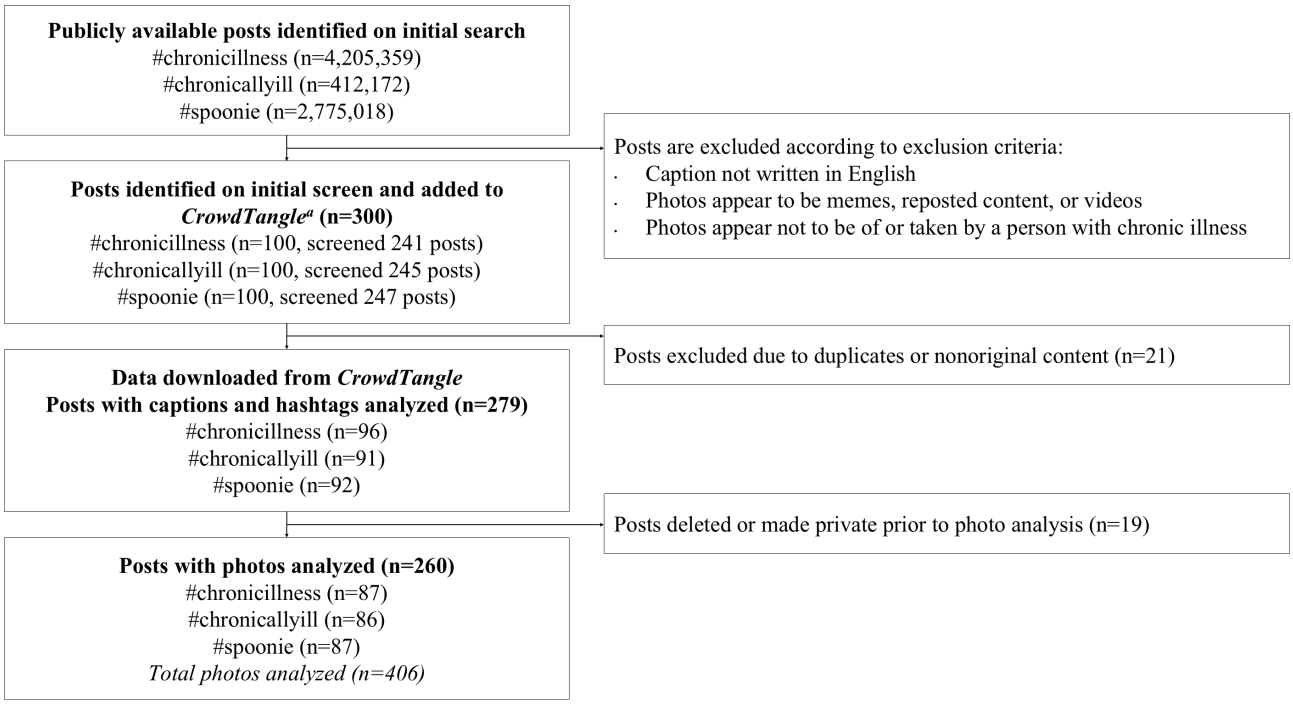

a: *CrowdTangle:* a tool created by *Meta* to track and analyze publicly available social media data

Qualitative Analysis

Photos

Out of the 260 posts with photos analyzed (Table 1), 27 (10%) were in a health care setting, and 49 (19%) contained medical equipment. Among the posts featuring medical equipment, 10 posts depicted invasive devices, primarily IVs and feeding tubes, while 39 posts showed noninvasive equipment, mainly mobility aids. Support animals, such as dogs, cats, and stuffed animals, were present in 34 (13%) posts, often with captions written from the perspective of the support animal. The number of people in the photos ranged from 0 to 5, with a median of 1 (IQR 0-1).

Hashtags

A total of 5454 hashtags across the 279 posts were analyzed for conditions and medical interventions (Table 1, conditions and interventions mentioned in 10 or more posts; Table S1 in Multimedia Appendix 1, full list of hashtagged conditions and interventions). Among the analyzed posts, 243 (87%) posts contained hashtags referencing specific medical conditions, with 107 different conditions identified.

Many conditions are those that have a susceptibility to self-diagnosis, including fibromyalgia (56 posts) [30], anxiety (31 posts) [25], depression (22 posts) [31], Lyme disease (10 posts) [32], mast cell activation syndrome (5 posts) [33], and dissociative identity disorder (3 posts) [24]. Other frequently mentioned conditions are functional disorders such as chronic pain (101 posts), myalgic encephalitis/chronic fatigue syndrome (27 posts), headache (14 posts), and irritable bowel syndrome (7 posts) [34]. In contrast, it was uncommon to see posts referring to conditions that are less likely to be self-diagnosed or functional, such as inflammatory bowel disease (20 posts), multiple sclerosis (12 posts), diabetes (8 posts), and cancer (6 posts). Additionally, 57 (20%) posts included hashtags referencing medical interventions, with 28 different interventions identified.

Captions

Content analysis of the captions revealed four major themes: (1) health care experiences (n=97, 35%), (2) illness journey (n=182, 65%), (3) connection (n=155, 55%), and (4) nonillness experiences (n=91, 33%; Table 4). The health care experiences theme encompassed content related to specific health care encounters, diagnostic testing, and medical treatments. The illness journey theme explored the impact of chronic illness on daily life, including the ways in which other people misunderstood their conditions, the concept of "invisible illness," and the promotion of products related to their illnesses. Invisible illness is a phrase used to raise awareness and highlight the unique struggles faced by individuals whose conditions may not be immediately apparent to others. The connection theme highlighted relationships established among individuals with chronic illnesses, the support received from both human and nonhuman sources, and advice for those without chronic illnesses to foster empathy and understanding. Finally, the nonillness experiences theme includes food, art, and activities. Excerpts from the full list of categories and codes are provided in Table S2 in Multimedia Appendix 1. An in-depth exploration of the qualitative codes and representative quotes will be the focus of future work.

XSL•FO
RenderX

Table . Summary of the 4 major caption content themes with the number of posts that included the theme (the same post may include multiple themes).

Caption theme	Posts, n (%)	Caption excerpt
Illness journey	182 (65)	"I can pretend I'm 'healthy' but I know deep down, that it's all fake. I have a pretty negative self image. I see a broken person. I struggle with exhaustion so heavy I feel drunk tired."
Connection	155 (55)	"Thank you to everyone who checked in with me these past 2 months. You have no idea how much I appreciate it!"
Health care experience	97 (35)	"Throw back to when I had my PICC Line placed in my arm 3 years ago. I sure don't miss having a PICC Line... This was when my illnesses became much more visible... Having a tube in my arm for 7.5 months says a lot about my health. I'm glad I had a port surgically implanted 2 years & 5 months ago & have been PICC Line free for 2 years & 5 months now…"
Nonillness experience	91 (33)	"Spent yesterday decorating for Halloween! Can't wait to show you all"

Quantitative Analysis

The bivariate logistic regression (Table 2) showed that posts had a statistically significantly higher odds of overperforming, or receiving more user engagement if they contained a longer caption (odds ratio [OR] 2.727, 95% CI 1.352, 5.499), health care content (OR 2.025, 95% CI 1.224-3.349), a photo with invasive medical equipment (OR 6.526, 95% CI 1.357-31.382), or a health care experience (OR 1.831, 95% CI 1.078-3.11). Health care content is a summary variable that includes posts with a photo of medical equipment, a photo in a medical location, or a caption containing the theme health care experience. During multivariate regression analysis, the calculations were controlled for the presence of people in the photo, type of post, number of hashtags, caption length, number of conditions hashtagged, gender, and number of followers. After controlling for these variables, caption length (OR 2.438, 95% CI 1.047-5.674), health care content (OR 1.854, 95% CI 1.004-3.421), and photos with invasive medical equipment (OR 6.187, 95% CI 1.16-32.991) remained statistically significant. Figure 4 provides examples of posts containing overperforming content and advertisements.

XSL•FO
RenderX

Figure 4. Example posts created from Shutterstock images of visual themes identified during photo content analysis, including health care setting, invasive medical equipment, noninvasive medical equipment, and advertisements.

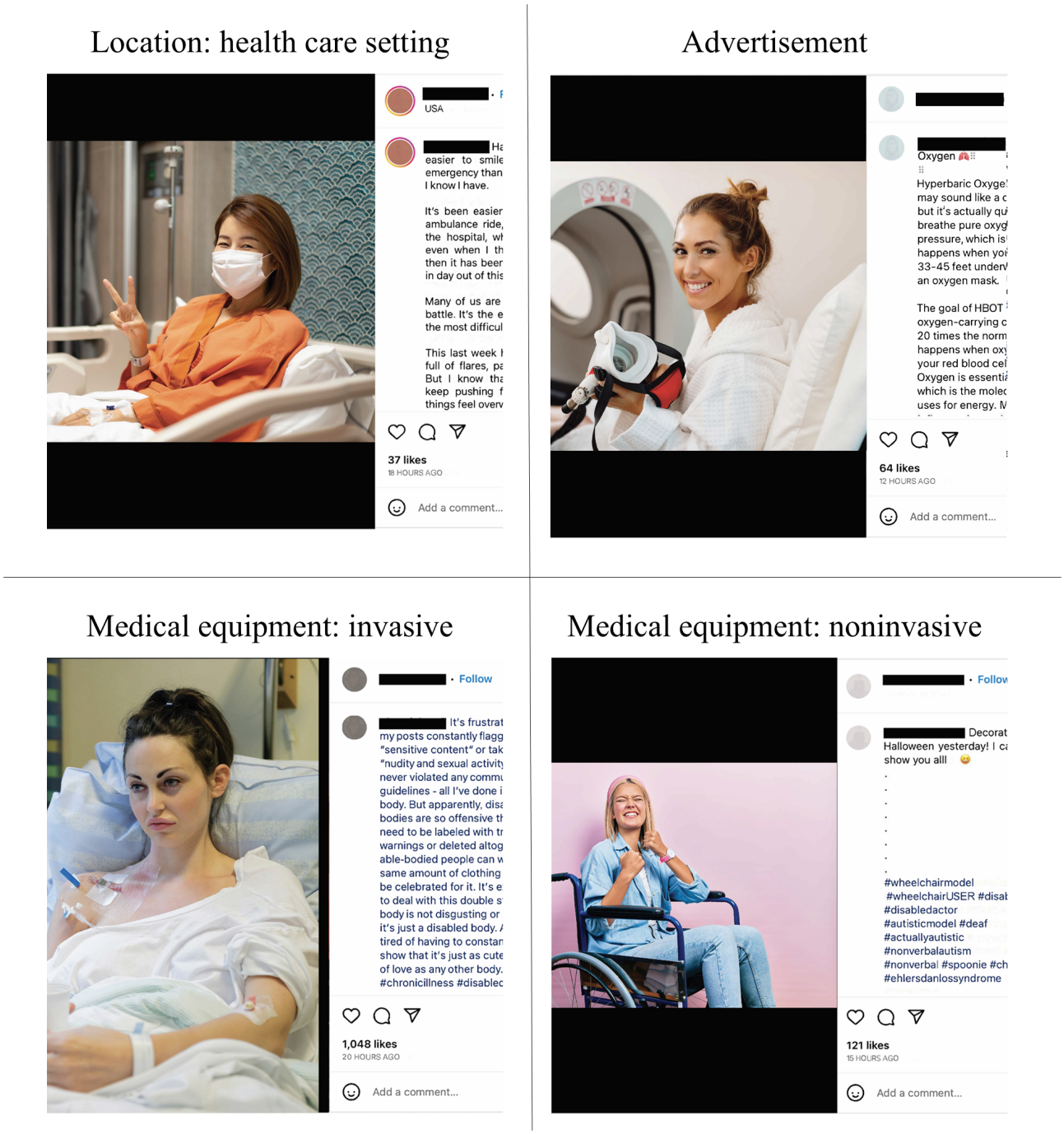

Discussion

Principal Findings

This study identified the content of Instagram posts related to chronic illness and assessed if health care content received higher engagement. The captions frequently included personal illness journey (n=182, 65%), while 35% (n=97) featured overt references to health care experiences. The quantitative analysis revealed that posts with longer captions (OR 2.44, 95% CI 1.05-5.67), health care content (OR 1.85, 95% CI 1.00-3.42), and images showing invasive medical equipment (OR 6.19, 95% CI 1.16-32.99) were more likely to overperform, even when controlling for other posts' characteristics. These findings suggest that more descriptive and medicalized portrayals of illness receive greater user engagement.

These findings support prior work demonstrating that social media is used to share illness narratives and medical experiences [5,17,35]. Notably, the study adds that health care content, especially posts with images showing invasive medical equipment, garners remarkably high engagement. Medicalized portrayals of chronic illness may be amplified by social media algorithms, resulting in implications for both creators and viewers, as it may reinforce a narrow, medicalized view of chronic illness. Repeated exposure to algorithmically amplified content increases visibility and creates perceived social norms, shaping how users interact with and internalize illness-related

XSL•FO
RenderX

information. Through curated self-presentation, observational learning, and identity reinforcement, social media becomes not just a space for sharing experiences, but a place for shaping illness identity and behavior [3,36,37].

For content creators, higher engagement provides positive reinforcement, encourages certain behaviors, and motivates the production of similar posts. SPT suggests that creators will curate their content based on their desired identity and feedback [12]. These results suggest posts featuring medicalized portrayals of illness receive more engagement, which may reinforce the sick role and increase focus on health care encounters and medical interventions. This not only deepens their connection to their illness but also increases their perceived self-efficacy, which can improve their confidence in managing their condition and ability to self-advocate for their health [38]. However, it can also lead them to believe they have expert knowledge about their conditions even above medical providers, as suggested by posts that downplay physician knowledge and promote commercial self-help content.

Commercial interests further complicate this dynamic. Approximately one in five posts advertised products, ranging from alternative therapies (ie, oxygen therapy and supplements) to personal Etsy shops selling illness-related merchandise. This is consistent with the broader role of social media influencers in product marketing [39] and further showcases social media's enhancement of self-efficacy and autonomy. It can provide accessible income opportunities for individuals who may be underserved by traditional employment. However, the potential for financial gain may further incentivize creators to medicalize their digital identity and promote content that aligns with the marketable aspects of chronic illness. As a result, authentic self-expression becomes enmeshed with commercial motivation.

For viewers, social media platforms quickly tailor content based on engagement patterns, further exposing viewers to posts like those previously viewed. An investigation by the Wall Street Journal suggests that some platforms' algorithms use individual user engagement patterns to quickly adjust post recommendations, leading viewers to less monitored and more extreme content [40]. This is concerning through the lens of SCT's observational learning, where individuals adopt behaviors they repeatedly see and that are reinforced through social approval. Content that receives greater attention is not only more visible and memorable but is also algorithmically promoted, increasing exposure and perceived social norms. Repeated presentation of health care encounters, and even more concerningly, invasive medical equipment in overperforming posts may normalize these interventions as a required part of the chronic illness journey. This could lead to an exaggerated emphasis on the importance of medicalization in the role of illness validation, solidifying the illness identity.

From an IIT perspective, validation through post attention may solidify an illness-centered identity. IIT describes how chronic illness becomes integrated into an individual's identity via acceptance, enrichment, rejection, and engulfment. While acceptance and enrichment have improved health outcomes, engulfment and rejection are associated with higher medical utilization and worse outcomes [14-16]. Social media has the potential to increase enrichment through advocacy and shared learning, but one study found that viewers were actually less likely to engage with enriching content [5]. Higher engagement with medicalized content may push users toward engulfment, where illness becomes a more consuming part of identity.

The focus on medicalized content and its ability to integrate illness into one's identity is concerning for viewers, regardless of current chronic illness diagnosis. Posts often reference self-diagnosable conditions, such as chronic pain (n=101, 36.2%), anxiety (n=31, 11.1%), and chronic fatigue (n=27, 9.6%) [25,26,31]. These experiences, which are common in the general population, are highly susceptible to somatic hypervigilance, where excessive attention to bodily sensations leads to symptom amplification and perceived illness. In web-based communities, this may lead to the pursuit of popularized diagnoses, such as postural orthostatic tachycardia syndrome (POTS) or fibromyalgia, that can result in a prolonged diagnostic odyssey seeking validation. Self-efficacy narratives compound this process by encouraging individuals to trust their self-perception over the results of clinical evaluations. Viewers are told not to give up until they receive the diagnosis they believe they have, further reinforcing the cycle of symptom monitoring, medical validation-seeking, and identity reinforcement.

For both content creators and viewers, repeated engagement with chronic illness–related content can improve self-efficacy, increasing confidence in their understanding and management of illness, even when these beliefs differ from medical expertise. Clinically, this aligns with the authors' observations of patients requesting evaluation for rare diagnoses or requesting inappropriate treatments based on social media content. Prior research confirms that social media exposure can change behavior, with Kankova et al [36] demonstrating that influencers shape health-related attitudes and behaviors in both positive and negative ways. Other studies highlight how social media influencers drive chronic illness conversations, shape public perceptions of health, and may influence health care–seeking behaviors [17,37,41].

Posts with health care content may attract more attention due to perceived authenticity, emotional appeal, and unique insight into the lived experience of chronic illness. These narratives may provide visibility and validation for individuals who feel misunderstood by doctors and healthy individuals. While social media provides invaluable support and visibility for individuals with chronic illness, it also plays a powerful role in shaping illness identity. Increased engagement with medicalized content may inadvertently shift individuals toward an illness-centered identity, where symptoms become entwined with both one's sense of self and social belonging. Once this identity is established, the transition to wellness can be difficult, as recovery may feel like a loss of identity and community support. The benefits of connection and validation on social media must come with recognition of the potential harms of reinforcing maladaptive illness narratives.

Limitations

There are several limitations in this study. First, data collection was conducted on a single day, which could have skewed results

XSL•FO
RenderX

toward time-sensitive events such as awareness weeks. However, our goal was not to generalize findings about specific conditions, but rather to identify broader themes in chronic illness content. Second, convenience sampling was necessary due to the large volume of posts and continuous content creation. While this approach may introduce bias, this study mitigated this by using a "most recent" search strategy to ensure that included posts reflected current discussions. Third, demographic details were not available for all posts, limiting insights into the characteristics of content creators. However, the study's primary focus was on content rather than user demographics. Finally, while private accounts and posts were inaccessible, the study captured perspectives from the digital chronic illness community that are publicly available.

Conclusions and Future Work

This study demonstrates that social media engagement reinforces medicalized portrayals of chronic illness, particularly through overperforming content featuring health care encounters and invasive medical equipment. By integrating the SPT, SCT, and IIT frameworks, it becomes clear that creators curate content to maximize engagement, viewers internalize and normalize behaviors through observational learning, and both adapt their illness identities based on social feedback and validation. While platforms like Instagram provide validation, support, and advocacy opportunities for people with chronic illness, they also present risks of symptom amplification, overmedicalization, and the spread of psychogenic symptoms. The commercialization of chronic illness content complicates things further by blurring the line between authentic self-expression and financially motivated narratives. These, along with the anecdotal reports of patients requesting specific diagnoses or treatments based on social media content, highlight the urgent need for health care providers to be informed about digital health trends. Medical professionals actively engaging with chronic illness content is vital to ensure accurate and evidence-based information is available to counter misinformation.

These findings underscore the importance of integrating social media literacy into clinical education and underscore the need for social media platforms and policymakers to consider how engagement metrics shape public health discourse. Future research should explore the effects of chronic illness content on both individuals and at the public health level. In particular, understanding how repeated exposure to chronic illness content influences health behaviors, self-diagnosis trends, and health care utilization is of vital importance. Further examination of social media influencers' role in shaping chronic illness narratives is also warranted. As digital platforms evolve, a balanced approach is needed, one that leverages the benefits of web-based health communities while mitigating the risks associated with misinformation, commercial influence, and illness identity reinforcement. Health care providers can play a critical role by providing accessible, evidence-based content that empowers patients to make informed decisions. Ensuring that social media remains a space for support, education, and empowerment is essential to preventing unintended consequences of chronic illness identification and overmedicalization.

Acknowledgments

No funding was obtained for this study. Thank you to Drs Bonnie Halpern-Felsher and Lee Sanders for their mentorship and assistance in idea development.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

LDD and FSWL conceptualized and designed the study, collected data, carried out the initial analyses, drafted the initial manuscript, and critically reviewed and revised the manuscript. OS carried out the quantitative analyses and critically reviewed and revised the manuscript. CJW conceptualized and designed the study, coordinated and supervised data collection, and critically reviewed and revised the manuscript for important intellectual content. Portions of the manuscript were drafted and revised with the assistance of OpenAI's ChatGPT, under the supervision of the lead author. All authors approved the final manuscript as submitted and agreed to be accountable for all aspects of the work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Full list of hashtag conditions and interventions and full list of categories and codes with excerpts from post.
[DOCX File, 30 KB - formative_v9i1e57523_app1.docx]

References

1. #chronicillness. Instagram. URL: https://www.instagram.com/explore/search/keyword/?q=%23chronicillness [accessed 2025-02-26]

XSL•FO
RenderX

2. Bandura A. Human agency in social cognitive theory. Am Psychol 1989 Sep;44(9):1175-1184. [doi: 10.1037/0003-066x.44.9.1175] [Medline: 2782727]
3. Turner G, Ferguson AM, Katiyar T, Palminteri S, Orben A. Old strategies, new environments: reinforcement learning on social media. Biol Psychiatry 2025 May 15;97(10):989-1001. [doi: 10.1016/j.biopsych.2024.12.012] [Medline: 39725300]
4. Patel R, Chang T, Greysen SR, Chopra V. Social media use in chronic disease: a systematic review and novel taxonomy. Am J Med 2015 Dec;128(12):1335-1350. [doi: 10.1016/j.amjmed.2015.06.015] [Medline: 26159633]
5. Hale TM, Pathipati AS, Zan S, Jethwani K. Representation of health conditions on Facebook: content analysis and evaluation of user engagement. J Med Internet Res 2014 Aug 4;16(8):e182. [doi: 10.2196/jmir.3275] [Medline: 25092386]
6. Benetoli A, Chen TF, Aslani P. Consumer health-related activities on social media: exploratory study. J Med Internet Res 2017 Oct 13;19(10):e352. [doi: 10.2196/jmir.7656] [Medline: 29030326]
7. Lin X, Kishore R. Social media-enabled healthcare: a conceptual model of social media affordances, online social support, and health behaviors and outcomes. Technol Forecast Soc Change 2021 May;166:120574. [doi: 10.1016/j.techfore.2021.120574]
8. Farsi D, Martinez-Menchaca HR, Ahmed M, Farsi N. Social media and health care (part II): narrative review of social media use by patients. J Med Internet Res 2022 Jan 7;24(1):e30379. [doi: 10.2196/30379] [Medline: 34994706]
9. Allen C, Vassilev I, Kennedy A, Rogers A. Long-term condition self-management support in online communities: a meta-synthesis of qualitative papers. J Med Internet Res 2016 Mar 10;18(3):e61. [doi: 10.2196/jmir.5260] [Medline: 26965990]
10. Ziebland S, Wyke S. Health and illness in a connected world: how might sharing experiences on the internet affect people's health? Milbank Q 2012 Jun;90(2):219-249. [doi: 10.1111/j.1468-0009.2012.00662.x] [Medline: 22709387]
11. Raffel S. The everyday life of the self: reworking early Goffman. J Class Sociol 2013 Feb;13(1):163-178. [doi: 10.1177/1468795X12474055]
12. Shi J, Lai KK, Chen G. Examining retweeting behavior on social networking sites from the perspective of self-presentation. PLoS One 2023;18(5):e0286135. [doi: 10.1371/journal.pone.0286135] [Medline: 37216346]
13. Masur PK, DiFranzo D, Bazarova NN. Behavioral contagion on social media: effects of social norms, design interventions, and critical media literacy on self-disclosure. PLoS One 2021;16(7):e0254670. [doi: 10.1371/journal.pone.0254670]
14. Charmaz K. The self as habit: the reconstruction of self in chronic illness. Occup Ther J Res 2002 Jan;22(1_suppl):31S-41S. [doi: 10.1177/153944920202205105]
15. Rassart J, Van Wanseele C, Debrun L, et al. Illness identity in inflammatory bowel disease. Int J Behav Med 2023 Feb;30(1):77-88. [doi: 10.1007/s12529-022-10072-y] [Medline: 35257307]
16. Oris L, Rassart J, Prikken S, et al. Illness identity in adolescents and emerging adults with type 1 diabetes: introducing the illness identity questionnaire. Diabetes Care 2016 May;39(5):757-763. [doi: 10.2337/dc15-2559] [Medline: 26989179]
17. Gabarron E, Larbi D, Dorronzoro E, Hasvold PE, Wynn R, Årsand E. Factors engaging users of diabetes social media channels on Facebook, Twitter, and Instagram: observational study. J Med Internet Res 2020 Sep 29;22(9):e21204. [doi: 10.2196/21204] [Medline: 32990632]
18. Harriger JA, Evans JA, Thompson JK, Tylka TL. The dangers of the rabbit hole: reflections on social media as a portal into a distorted world of edited bodies and eating disorder risk and the role of algorithms. Body Image 2022 Jun;41:292-297. [doi: 10.1016/j.bodyim.2022.03.007] [Medline: 35378338]
19. The Spoon Theory written by Christine Miserandino. But You Dont Look Sick?. 2013. URL: https://butyoudontlooksick.com/articles/written-by-christine/the-spoon-theory/ [accessed 2025-02-27]
20. Sickness and lies. BBC News. URL: https://www.bbc.co.uk/programmes/p09r3w40 [accessed 2025-02-27]
21. Mulcahy L. Dying for attention: faking illness becomes an online epidemic. WebMD. URL: https://www.webmd.com/mental-health/features/dying-attention-faking-illness-becomes-online-epidemic [accessed 2025-03-02]
22. Müller-Vahl KR, Pisarenko A, Jakubovski E, Fremer C. Stop that! It's not Tourette's but a new type of mass sociogenic illness. Brain (Bacau) 2022 Apr 18;145(2):476-480. [doi: 10.1093/brain/awab316] [Medline: 34424292]
23. Olvera C, Stebbins GT, Goetz CG, Kompoliti K. TikTok tics: a pandemic within a pandemic. Mov Disord Clin Pract 2021 Nov;8(8):1200-1205. [doi: 10.1002/mdc3.13316] [Medline: 34765687]
24. Salter M, Brand BL, Robinson M, Loewenstein R, Silberg J, Korzekwa M. Self-diagnosed cases of dissociative identity disorder on social media: conceptualization, assessment, and treatment. Harv Rev Psychiatry 2025;33(1):41-48. [doi: 10.1097/HRP.0000000000000416] [Medline: 39761444]
25. Hasan F, Foster MM, Cho H. Normalizing anxiety on social media increases self-diagnosis of anxiety: the mediating effect of identification (but not stigma). J Health Commun 2023 Sep 2;28(9):563-572. [doi: 10.1080/10810730.2023.2235563] [Medline: 37448221]
26. Tan W, Pickup B, Faasse K, Colagiuri B, Barnes K. Peer-to-peer: the social transmission of symptoms online. Ann Behav Med 2023 Jun 30;57(7):551-560. [doi: 10.1093/abm/kaac081]
27. Giedinghagen A. The tic in TikTok and (where) all systems go: mass social media induced illness and Munchausen's by internet as explanatory models for social media associated abnormal illness behavior. Clin Child Psychol Psychiatry 2023 Jan;28(1):270-278. [doi: 10.1177/13591045221098522] [Medline: 35473358]

XSL•FO
RenderX

28. Haltigan JD, Pringsheim TM, Rajkumar G. Social media as an incubator of personality and behavioral psychopathology: symptom and disorder authenticity or psychosomatic social contagion? Compr Psychiatry 2023 Feb;121:152362. [doi: 10.1016/j.comppsych.2022.152362] [Medline: 36571927]
29. How do you calculate overperforming scores? CrowdTangle Help Center. URL: http://help.crowdtangle.com/en/articles/2013937-how-do-you-calculate-overperforming-scores [accessed 2023-06-14]
30. Walitt B, Katz RS, Bergman MJ, Wolfe F. Three-quarters of persons in the US population reporting a clinical diagnosis of fibromyalgia do not satisfy fibromyalgia criteria: the 2012 National Health Interview Survey. PLoS One 2016;11(6):e0157235. [doi: 10.1371/journal.pone.0157235] [Medline: 27281286]
31. Monteith S, Glenn T, Geddes JR, Whybrow PC, Achtyes ED, Bauer M. Implications of online self-diagnosis in psychiatry. Pharmacopsychiatry 2024 Mar;57(2):45-52. [doi: 10.1055/a-2268-5441] [Medline: 38471511]
32. Kobayashi T, Higgins Y, Melia MT, Auwaerter PG. Mistaken identity: many diagnoses are frequently misattributed to Lyme disease. Am J Med 2022 Apr;135(4):503-511. [doi: 10.1016/j.amjmed.2021.10.040] [Medline: 34861197]
33. Buttgereit T, Gu S, Carneiro-Leão L, Gutsche A, Maurer M, Siebenhaar F. Idiopathic mast cell activation syndrome is more often suspected than diagnosed—a prospective real-life study. Allergy 2022 Sep;77(9):2794-2802. [doi: 10.1111/all.15304] [Medline: 35364617]
34. Fall EA, Chen Y, Lin JMS, et al. Chronic overlapping pain conditions in people with myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS): a sample from the multi-site clinical assessment of ME/CFS (MCAM) study. BMC Neurol 2024 Oct 18;24(1):399. [doi: 10.1186/s12883-024-03872-0] [Medline: 39425035]
35. Foufi V, Timakum T, Gaudet-Blavignac C, Lovis C, Song M. Mining of textual health information from Reddit: analysis of chronic diseases with extracted entities and their relations. J Med Internet Res 2019 Jun 13;21(6):e12876. [doi: 10.2196/12876] [Medline: 31199327]
36. Kaňková J, Binder A, Matthes J. Health-related communication of social media influencers: a scoping review. Health Commun 2025 Jun;40(7):1300-1313. [doi: 10.1080/10410236.2024.2397268] [Medline: 39258728]
37. Myrick JG, Willoughby JF, Francis DB, Noar SM. The impact of celebrity and influencer illness disclosures. Health Commun 2024 Dec;39(14):3577-3580. [doi: 10.1080/10410236.2024.2326261] [Medline: 38594789]
38. Green BM, Van Horn KT, Gupte K, Evans M, Hayes S, Bhowmick A. Assessment of adaptive engagement and support model for people with chronic health conditions in online health communities: combined content analysis. J Med Internet Res 2020 Jul 7;22(7):e17338. [doi: 10.2196/17338] [Medline: 32492651]
39. Jamil K, Dunnan L, Gul RF, Shehzad MU, Gillani SHM, Awan FH. Role of social media marketing activities in influencing customer intentions: a perspective of a new emerging era. Front Psychol 2021;12:808525. [doi: 10.3389/fpsyg.2021.808525] [Medline: 35111111]
40. Investigation: how TikTok's algorithm figures out your deepest desires. WSJ. URL: https://tinyurl.com/muke5yme [accessed 2025-03-02]
41. Necaise A, Amon MJ. Peer support for chronic pain in online health communities: quantitative study on the dynamics of social interactions in a chronic pain forum. J Med Internet Res 2024 Sep 5;26:e45858. [doi: 10.2196/45858] [Medline: 39235845]

Abbreviations

IIT: illness identity theory
IV: intravenous line
OR: odds ratio
SCT: social cognitive theory
SPT: self-presentation theory

Edited by A Mavragani; submitted 19.02.24; peer-reviewed by A Hassan, A Markiewitz; revised version received 30.06.25; accepted 02.07.25; published 26.09.25.

Please cite as:
Daon LD, Lee FSW, Saynina O, Wang CJ
Health Care Content and Engagement in Chronic Illness Instagram Posts: Content Analysis
JMIR Form Res 2025;9:e57523
URL: https://formative.jmir.org/2025/1/e57523
doi:10.2196/57523

© Lindsey D Daon, Fu-Shiuan Whitney Lee, Olga Saynina, C Jason Wang. Originally published in JMIR Formative Research (https://formative.jmir.org), 26.9.2025. This is an open-access article distributed under the terms of the Creative Commons

XSL•FO
RenderX

Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Disease Activity Assessment Frequency in Rheumatoid Arthritis: A Retrospective Observational Study of the Medical Support System for Rheumatoid Arthritis System Implementation

Mari Yamamoto[1,2,3], MD; Yuki Kataoka[2,4,5,6], MD, MPH, DrPH; Yasushi Tsujimoto[2,7,8,9], MD, MPH, PhD; Fumika Nagase[1], MD; Keita Iwasaki[1], MD; Yuuki Ito[1], MD; Hiroki Ikai[1], MD; Tsuyoshi Watanabe[1,3], MD, PhD; Waka Yokoyama-Kokuryo[1], MD, PhD; Naoho Takizawa[1], MD; Yoshiro Fujita[1,3], MD

[1]Department of Rheumatology and Nephrology, Chubu Rosai Hospital, 1-10-6 Komei, Minato-ku, Nagoya, Aichi, Japan

[2]Scientific Research WorkS Peer Support Group (SRWS-PSG), Osaka, Japan

[3]Infection Control Team, Chubu Rosai Hospital, 1-10-6 Komei, Minato-ku, Nagoya, Aichi, Japan

[4]Department of Internal Medicine, Kyoto Min-iren Asukai Hospital, Kyoto, Japan

[5]Department of International and Community Oral Health, Tohoku University Graduate School of Dentistry, Miyagi, Japan

[6]Department of Healthcare Epidemiology, Kyoto University Graduate School of Medicine/Public Health, Kyoto, Japan

[7]Oku Medical Clinic, Osaka, Japan

[8]Department of Health Promotion and Human Behavior, Kyoto University Graduate School of Medicine/School of Public Health, Kyoto University, Kyoto, Japan

[9]Division of Rheumatology, Department of Internal Medicine, Showa University School of Medicine, Shinagawa-ku, Tokyo, Japan

Corresponding Author:
Mari Yamamoto, MD
Department of Rheumatology and Nephrology, Chubu Rosai Hospital, 1-10-6 Komei, Minato-ku, Nagoya, Aichi, Japan

Abstract

Background: Rheumatoid arthritis (RA) is a global health concern with increasing prevalence. Despite recommendations for regular disease activity assessments, their implementation in routine clinical practice remains challenging. The Medical Support System for Rheumatoid Arthritis (MiRAi) is an offline, semi-automated system that calculates disease activity indices by integrating patient and clinician inputs from electronic health records (EHRs).

Objective: This study evaluated the association between MiRAi implementation and the frequency of disease activity assessments in patients with RA.

Methods: We conducted a retrospective cohort study of patients with RA treated at a tertiary hospital in Japan between April 2022 and March 2023. We included all adult outpatients (aged ≥18 years) with RA diagnosed according to the 2010 ACR/EULAR [American College of Rheumatology/European Alliance of Associations for Rheumatology] classification criteria. The hospital introduced MiRAi in June 2022 and achieved full deployment by October 2022. MiRAi calculated the clinical disease activity index (CDAI) and the modified health assessment questionnaire (mHAQ) through automated extraction of joint counts, patient global assessment, and functional status from structured EHR fields. Primary outcomes included the frequency of CDAI and mHAQ assessments. We administered a structured post-implementation survey to assess rheumatologists' perceptions of MiRAi.

Results: Physicians used MiRAi for 236/884 (26.7%) patients with RA. Patients with documented CDAI and mHAQ scores increased from 29 (5.9%) in June 2022 to 81 (19.0%) in November 2022, representing a 3.2-fold increase. Among surveyed rheumatologists (n=10), 5 (50%) reported the regular use of MiRAi. Physicians who regularly used MiRAi (n=5) cited improved accuracy in disease assessment and enhanced treatment decision-making. Non-users and occasional users (n=5) identified three primary barriers: limited familiarity with MiRAi, time constraints, and discrepancies between clinical judgment and MiRAi-generated outputs. Despite MiRAi's availability, only 168 (19%) patients underwent quantitative disease activity assessment by the study end. Among 15 patients with high disease activity (CDAI >22), physicians recorded 3 treatment modifications and 2 intra-articular steroid injections.

Conclusions: MiRAi implementation increased disease activity assessment frequency by 3.2-fold over 6 months; however, physician adoption remained at 26.7%, below the 80% target for routine care. Future implementation strategies should address identified barriers through system integration, structured user training, and workflow optimization to achieve guideline-concordant care for patients with RA.

(JMIR Form Res 2025;9:e74222) doi:10.2196/74222

XSL•FO
RenderX

KEYWORDS

disease management; electronic health records; quality improvement; rheumatoid arthritis; Medical Support System for Rheumatoid Arthritis; MiRAi

Introduction

Rheumatoid arthritis (RA) is a major global health concern. A 2021 meta-analysis, which included data from 67 facilities across 41 countries, reported a pooled RA prevalence of 0.46% (95% CI 0.37% - 0.57%) based on studies conducted between 1986 and 2014 [1]. A 2020 national survey in Japan estimated the prevalence of RA to be 0.75% [2]. The prevalence of RA varies across countries and is influenced by socioeconomic conditions, environmental exposures, and rheumatologist availability, as demonstrated in a global systematic review [3]. Despite regional differences, the global prevalence of RA continues to increase annually [4]. The assessment of RA activity requires both clinical and biological evaluations. Standard assessment tools for RA, such as the Clinical Disease Activity Index (CDAI) and Disease Activity Score in 28 Joints (DAS28), include subjective assessments [5]. Treatment adjustments based on disease activity are widely recommended [6].

DAS28 requires complex calculations, limiting its practicality in routine clinical settings, whereas CDAI is simpler to compute and demonstrates high utility [7]; however, published reports on their real-world application remain scarce and limited. A 2015 survey revealed that 59% of Japanese rheumatologists aimed to achieve clinical remission; however, only 45% used a composite measure of disease activity [8]. An online assessment tool has been developed for RA; however, it cannot be used if electronic health records (EHRs) are not connected to the Internet because of security concerns [9].

The treat to target (T2T) strategy—characterized by shared decision-making between patients and clinicians—effectively reduces disease activity and functional impairment [10]. However, a 2024 cross-sectional survey reported that most of the 166 rheumatologists perceived discordance between composite scores (eg, DAS28, CDAI) and clinical judgment, especially in cases with conflicting patient-reported outcomes [11]. This perceived misalignment and lack of consistent monitoring remain significant real-world barriers to T2T implementation. These concerns were echoed at the 2024 American College of Rheumatology and U.S. Food and Drug Administration joint summit, where experts highlighted critical limitations in RA disease activity assessments—including inconsistent definitions, non-standardized evaluation frequency, and difficulty in leveraging EHRs and insurance claims data for longitudinal monitoring [12]. The summit emphasized the need for the harmonization of assessment frameworks across real-world clinical environments.

The Medical Support System for Rheumatoid Arthritis (MiRAi) is an automated offline tool that integrates data from electronic health records to calculate objective assessment scores for collagen diseases. Since its development in 2009, the MiRAi has been implemented in 16 facilities across Japan. The use of MiRAi for RA disease activity assessment has been shown to improve patient satisfaction [13]; however, its impact on the frequency of assessment remains unclear.

We aimed to evaluate whether the implementation of MiRAi increased the frequency of assessment of disease activity in patients with RA.

Methods

Study Design and Participants

This retrospective cohort study was conducted at the Chubu Rosai Hospital, a tertiary-care facility in Japan, and included patients with RA who visited the Rheumatology Department between April 2022 and March 2023. We identified patients from the outpatient database and included them consecutively based on diagnostic codes to reflect routine clinical practice. The primary physician confirmed each RA diagnosis and enrolled the patients. No additional interventions were introduced during the observation period. In this institute, the operational use of MiRAi began in June 2022. After integrating MiRAi into the EHR system, we began using it in an outpatient setting. Initially, a few physicians used the MiRAi on a trial basis. In October 2022, all 10 staff members were informed about the system, and MiRAi was fully implemented.

The MiRAi System

MiRAi, developed by a Japanese medical institution in 2009, is an electronic system that automatically provides RA activity indicators after patient and physician assessments are entered into an attached tablet and EHR. The MiRAi system operates as follows. First, the patient enters the patient visual analog scale (VAS) and modified health assessment questionnaire (mHAQ) on a tablet in an outpatient setting. Next, during the consultation, the physician clicks on the MiRAi tab, and the values entered by the patient are reflected in the EHR. The physician then enters the doctor's VAS, C-reactive protein level, and erythrocyte sedimentation rate, following which the DAS28 and CDAI are automatically calculated. The locations of the swollen and tender joints are entered by clicking on the diagram, and the previously recorded data can also be reviewed.

Outcomes

To verify the changes in disease activity frequency following the introduction of MiRAi, we assessed the following factors: the proportion of CDAI and mHAQ evaluations among unique patients with RA per month; for patients with active RA (CDAI >22), the percentage of patients who changed or added targeted synthetic disease-modifying antirheumatic drugs, biological DMARDs, or conventional synthetic DMARDs; and the proportion of patients who received steroid joint injections during the follow-up period.

Questionnaire Survey and Data Analysis

In April 2024, we conducted a questionnaire survey based on the behavior change stage model [14]. The survey was administered to physicians who provided care at the outpatient

XSL•FO
RenderX

clinic of the Rheumatology Department. The survey included multiple-choice questions regarding the disease activity indicators used for RA and the use of MiRAi. Information regarding the advantages and disadvantages of MiRAi was collected from physicians who reported using it, while non-users were requested to state their reasons for not adopting MiRAi. We performed descriptive statistics to analyze the data using R Studio (Posit PBC, Boston, MA, USA) [15].

Ethical Considerations

The study adhered to the World Medical Association Declaration of Helsinki ethical guidelines [16] and the Ethical Guidelines for Medical and Biological Research Involving Human Subjects [17]. The Chubu Rosai Hospital Ethics Committee approved the study protocol (approval number: 202102‐06). We obtained written informed consent from all participants before enrollment. To ensure confidentiality, we de-identified all data before analysis. Participants did not receive any compensation or incentives. As the manuscript does not include any images or supplementary materials containing identifiable information, we did not require additional consent for image publication.

Results

Patient Characteristics

During the study period, 884 patients with RA visited our outpatient clinic, with an average of approximately 480 visits per month. MiRAi was used in 236 patients, accounting for 26.7% of all patients with RA. This retrospective observational study evaluated changes in disease activity assessment frequency following MiRAi implementation in routine clinical practice. MiRAi was introduced in June 2022, and full implementation began in October 2022.

Evaluation of Disease Activity and Therapeutic Interventions for Patients

Before the introduction of MiRAi, no quantitative evaluations of disease activity were performed at our facility. The number of CDAI and mHAQ evaluations increased from 29 (5.9%) in June 2022 during the trial period to 81 (19%) in November 2022 after the full implementation of MiRAi (Figure 1). Additionally, the percentage distribution of MiRAi ratings assigned to each physician was analyzed (Figure 2). We performed treatment evaluations in patients with high RA activity and those with RA remission. The number of events did not exceed 5 (Figure 3).

XSL•FO
RenderX

Figure 1. Monthly proportion of patients with RA who underwent CDAI and mHAQ assessments before and after implementation of the MiRAi system at Chubu Rosai hospital (April 2022–March 2023). RA: rheumatoid arthritis; CDAI: clinical disease activity index; mHAQ: modified health assessment questionnaire; MiRAI: Medical Support System for Rheumatoid Arthritis.

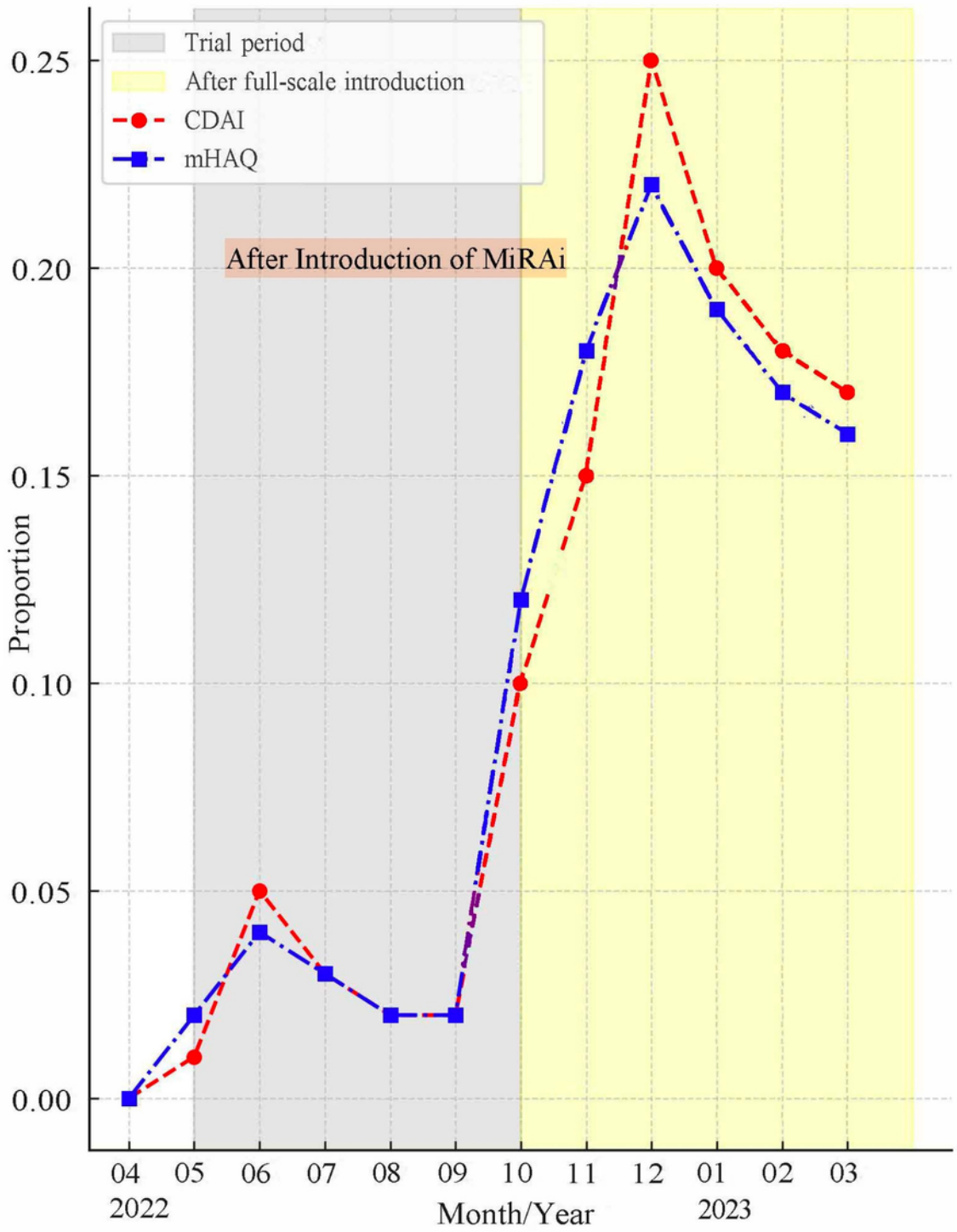

XSL•FO
RenderX

Figure 2. Proportion of patients with RA who underwent CDAI assessment by individual physicians at Chubu Rosai hospital during the post-implementation period of MiRAi (June to November 2022). This figure illustrates the variability in MiRAi usage across physicians. Differences may reflect variations in familiarity with the system, time constraints during outpatient visits, and attitudes identified in the post-implementation physician survey. The study was conducted as a retrospective observational analysis in a single tertiary-care facility in Japan. RA: rheumatoid arthritis; CDAI: clinical disease activity index; MiRAI: Medical Support System for Rheumatoid Arthritis.

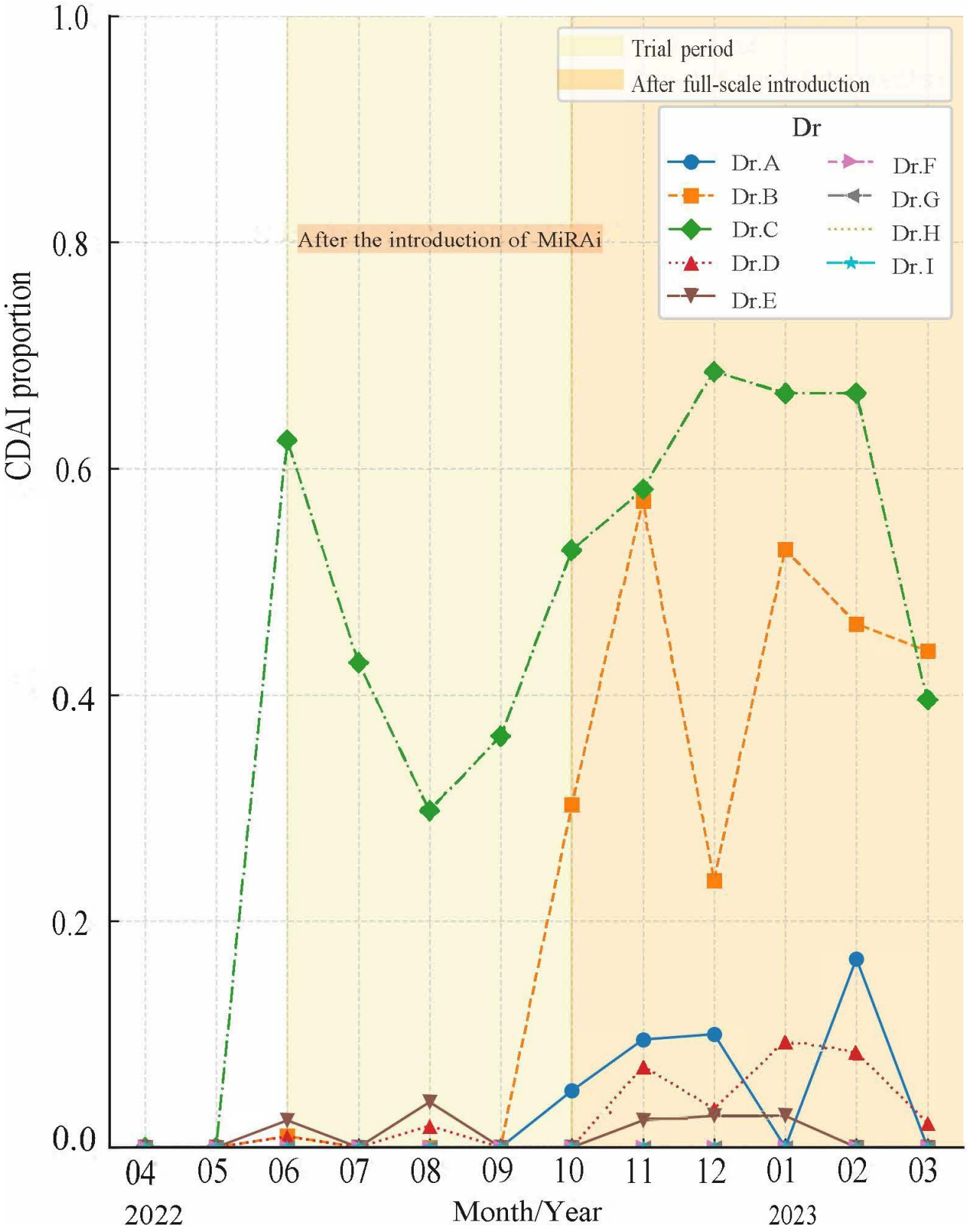

XSL•FO
RenderX

Figure 3. Proportion of patients with active rheumatoid arthritis (RA) (CDAI >22) who underwent treatment modifications—such as changes or additions of targeted synthetic disease-modifying antirheumatic drugs (tsDMARDs), biological DMARDs (bDMARDs), or conventional synthetic DMARDs (csDMARDs)—and those who received intra-articular steroid injections during the follow-up period (April 2022–March 2023). This figure reflects treatment response trends among patients with high disease activity in the context of MiRAi implementation at a tertiary-care hospital in Japan. Due to the limited number of high-activity cases, the number of events was small.

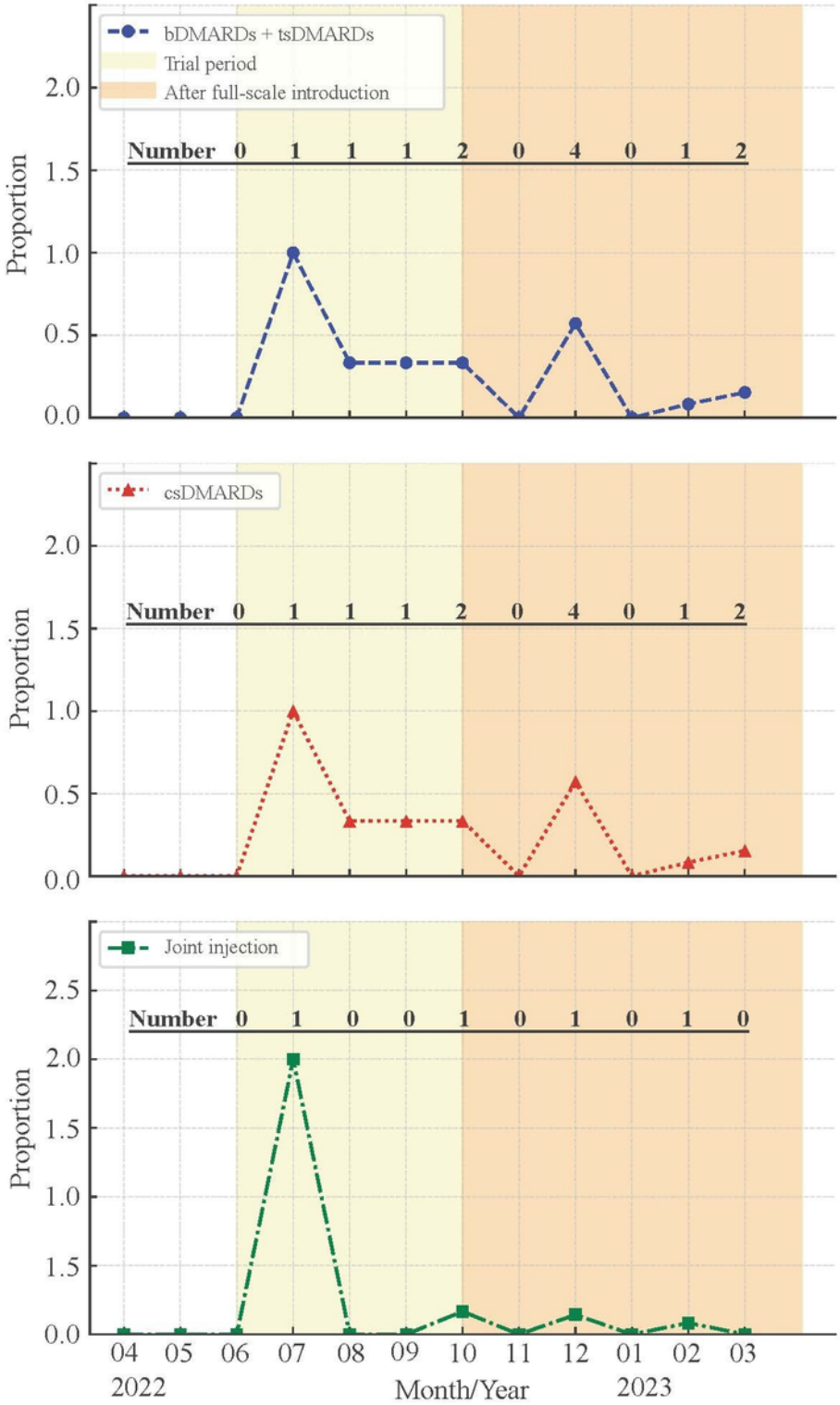

Questionnaire Survey

We mailed the survey to 10 rheumatologists affiliated with the Collagen Disease Department at Chubu Rosai Hospital, and all 10 responded. Among them, 70% (7/10) had over 4 years of experience in treating collagen diseases, while 30% (3/10) had less than 3 years of experience. Half of the rheumatologists believed that their disease activity assessments were appropriate (5/10), 40% (4/10) were not confident, and 10% (1/10) were unsure. Additionally, 50% (5/10) of the physicians reported regular use of MiRAi. The results of the content analysis are presented in Table 1.

XSL•FO
RenderX

Table . Post-implementation survey results.

Category	Negative comments	Positive comments
System	• I do not understand how to use the system. • Using MiRAi[a] takes time. • I want to be able to easily attach joint diagrams to medical records. • I want to be able to easily make reservations for equipment. • I want detailed explanations on how to use it.	• It was convenient after using it.
Evaluation method	• I think other evaluation methods are sufficient. • There is a discrepancy between MiRAi's evaluation and my own activity assessment. • It is hard to grasp the degree of joint swelling.	• I thought about improving the RA[b] evaluation method. • It allows for more accurate assessments of disease activity. • Deciding on treatment plans became easier. • Communication with patients became smoother.
Information sharing	• I want a mechanism to explain, distribute, and introduce MiRAi for all patients with RA. • I do not have time to explain the introduction of MiRAi to patients.	• It became easier to share information with other medical staff.

[a]MiRAi: Medical Support System for Rheumatoid Arthritis.

[b]RA: rheumatoid arthritis.

Discussion

Principal Findings

In this retrospective cohort of 884 patients with RA at Chubu Rosai Hospital (April 2022–March 2023), MiRAi use rose to 26.7%, and the combined CDAI/mHAQ assessment proportion increased from 5.9% (June 2022) to 19% (November 2022). Among those with high activity (CDAI >22), targeted DMARD adjustments and steroid injections were recorded, though infrequently. In a survey of all 10 rheumatologists, 50% regularly used MiRAi—citing its convenience, assessment accuracy, and smoother patient communication—while noting time demands, usability hurdles, and training needs. These results suggest that MiRAi integration boosts RA monitoring but requires refinements to optimize clinician uptake. These results warrant further interpretation in the context of existing RA literature.

Despite system-wide availability, MiRAi was used in only 19% of patients during the study period. A post-observation survey of 10 rheumatologists revealed that three main barriers limited adoption: unfamiliarity with the system, time constraints, and perceived discrepancies between clinical judgment and MiRAi outputs. Conversely, half of the physicians reported the regular use of MiRAi and identified three major benefits: improved clarity in disease activity assessment, enhanced support for treatment decisions, and more effective patient communication.

Interpretation and Comparison With Prior Works

To our knowledge, this is the first study to examine the integration of MiRAi, a semi-automated system, into routine clinical practice. We investigated whether its use is associated with changes in the frequency of RA disease activity assessment. In a clinical setting with outpatient visits every 1 - 3 months, approximately 20% of patients underwent quantitative assessments within 1 year of MiRAi implementation. This finding contrasts with those of previous studies that reported less frequent assessments. A multinational registry of routine RA care (METEOR) includes prospectively collected data from both academic and community rheumatology practices across 23 countries. It revealed that 85.9% of patients had at least one documented disease activity assessment over 4 years [9]. Additionally, a nationwide analysis of U.S. insurance claims data from commercially insured and Medicare Advantage patients with newly diagnosed RA highlighted substantial gaps in care quality. Only 44.1% of patients visited a rheumatologist within 1 year of diagnosis, and 31.3% met none of the recommended quality care indicators—such as annual clinical assessments and routine laboratory testing [18]. These results suggest that, particularly in routine care settings, routine disease activity assessments remain suboptimal. The semi-automated nature of MiRAi might improve the frequency of disease activity assessments compared with previous findings.

Clinical and Research Implications

The use of MiRAi might be associated with increased frequency of disease activity assessments, which in turn could support more standardized and objective monitoring in routine RA care. However, its clinical impact on treatment decisions and patient outcomes remains uncertain. In our study, MiRAi was used in only 81 patients (19%) during the observation period, and the number of patients with high disease activity or remission was small. As a result, we were not able to evaluate whether increased assessment frequency led to treatment modifications or changes in clinical outcomes.

XSL•FO
RenderX

Given these observations, further research is warranted. Future multicenter studies with larger sample sizes and longer follow-up are needed to clarify whether continued use of MiRAi contributes to timely treatment adjustments, better disease control, and improved long-term outcomes in patients with RA. In addition, barriers to adoption—such as usability issues, time constraints, and perceived discrepancies between clinician judgment and system outputs—must be addressed. Similar challenges have been reported in other clinical decision support system implementations, where limited integration into clinical workflows, lack of transparency, and insufficient personalization have been identified as key factors impeding adoption [19]. These insights underscore the broader need for thoughtful integration of digital tools into clinical workflows, ensuring usability, trust, and alignment with clinical needs to maximize their potential impact.

Limitations

Our study had some limitations. After the introduction of the system, only half of the physicians conducting outpatient consultations adopted MiRAi. Furthermore, the extent of improvement in disease assessment frequency among physicians using the system varied significantly. To investigate the reasons for this limited adoption, we conducted a post-observation survey of physicians and performed content analysis, which revealed several key findings. First, some physicians initially did not know how to use MiRAi. This indicates that the explanation provided during the introduction of the system was insufficient, underscoring the need for accessible user manuals. Second, many physicians rely on swollen joint counts, tender joint counts, and C-reactive protein levels for evaluation [20]. In contrast, half the physicians admitted that they lacked confidence in their disease activity assessments during consultations. This may be due to the absence of clear indicators of disease activity in routine clinical practice. Some physicians noted the benefits of MiRAi, including more accurate disease activity assessments and improved decision-making regarding treatments. These findings suggest that MiRAi can serve as a valuable tool for conducting appropriate disease activity evaluations. Many physicians who did not regularly use MiRAi cited time constraints, indicating that integrating the new system within a limited consultation time may have been burdensome for outpatient doctors. Therefore, it is necessary to automate the system to seamlessly incorporate objective evaluations into daily clinical practice. Additionally, tablet-based input by all patients with RA, without requiring explanations from attending physicians, could help reduce the clinicians' workload.

In addition, the single-center design and limited MiRAi adoption (19% of patients with RA) constrain generalizability. Institutional factors such as staffing, EHR integration, and patient characteristics may affect implementation in other settings. These findings should be interpreted cautiously, and future multicenter studies are planned to evaluate MiRAi's utility, adoption, and clinical impact across diverse care environments. Lastly, although MiRAi is designed to support both clinicians and patients, this study did not include an evaluation of patient perspectives. Understanding patient satisfaction, perceived usefulness, and the system's impact on shared decision-making is essential to fully assess its clinical value. We acknowledge this as a limitation and are currently planning a follow-up study that incorporates a patient questionnaire to address this important aspect.

Conclusions and Broader Implications

The implementation of MiRAi might increase the frequency of RA disease activity assessment. Future multicenter studies with sufficient sample sizes using MiRAi are warranted to verify whether the continued quantitative assessment of patients with RA leads to appropriate treatment changes.

Acknowledgments

We would like to extend our heartfelt gratitude to Shiro Ohshima, Kota Shimada, and Naoto Yokogawa for their invaluable assistance in conducting the statistical analyses for this research. We utilized Editage for professional English editing services, which were funded personally by our research team. Portions of the manuscript were edited using GPT-4o (OpenAI, San Francisco, CA, USA) to improve clarity and language; all content was subsequently reviewed and approved by the authors. This study was funded by participation in the Hospital Function Improvement Research operated by the Japan Labor Health and Welfare Organization, as well as research funds from the Rheumatology Department of Chubu Rosai Hospital. The Hospital Function Improvement Research was applied for as a two-year research project in the 2020 fiscal year, with the necessary details submitted to the Clinical Research Conflict of Interest Committee, and its review and approval obtained.

Data Availability

Datasets generated and analyzed during this study are not publicly available due to restrictions under the Japanese Act on the Protection of Personal Information (APPI) and institutional privacy regulations. De-identified individual-level data may be made available upon reasonable request to the corresponding author. In accordance with the Ethical Guidelines for Medical and Biological Research Involving Human Subjects [17], data sharing requires approval from the Director of Chubu Rosai Hospital and is contingent upon protocol review and approval by the ethics committee of the requesting institution (approval number: 202102‐06). The data supporting this study's findings are available from the corresponding author, Mari Yamamoto, upon reasonable request.

XSL•FO
RenderX

Authors' Contributions

All authors contributed to the study's conception and design. Mari Yamamoto, Fumika Nagase, Keita Iwasaki, Yuuki Ito, Hiroki Ikai, Tsuyoshi Watanabe, Waka Yokoyama-Kokuryo, Naoho Takizawa, and Yoshiro Fujita participated in the research process. Mari Yamamoto, Yuki Kataoka, and Yasushi Tsujimoto conducted documentation, data collection, and analysis. Mari Yamamoto drafted the initial manuscript, and all authors provided critical revisions. All authors reviewed and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Almutairi K, Nossent J, Preen D, Keen H, Inderjeeth C. The global prevalence of rheumatoid arthritis: a meta-analysis based on a systematic review. Rheumatol Int 2021 May;41(5):863-877. [doi: 10.1007/s00296-020-04731-0] [Medline: 33175207]
2. Kojima M, Nakayama T, Tsutani K, et al. Epidemiological characteristics of rheumatoid arthritis in Japan: prevalence estimates using a nationwide population-based questionnaire survey. Mod Rheumatol 2020 Nov;30(6):941-947. [doi: 10.1080/14397595.2019.1682776] [Medline: 31625435]
3. Almutairi KB, Nossent JC, Preen DB, Keen HI, Inderjeeth CA. The prevalence of rheumatoid arthritis: a systematic review of population-based studies. J Rheumatol 2021 May;48(5):669-676. [doi: 10.3899/jrheum.200367] [Medline: 33060323]
4. Finckh A, Gilbert B, Hodkinson B, et al. Global epidemiology of rheumatoid arthritis. Nat Rev Rheumatol 2022 Oct;18(10):591-602. [doi: 10.1038/s41584-022-00827-y] [Medline: 36068354]
5. Bombardier C, Mian S. Quality indicators in rheumatoid arthritis care: using measurement to promote quality improvement. Ann Rheum Dis 2013 Apr;72 Suppl 2:ii128-ii131. [doi: 10.1136/annrheumdis-2012-202259] [Medline: 23253925]
6. Romano MJ, Stafford RS. Electronic health records and clinical decision support systems: impact on national ambulatory care quality. Arch Intern Med 2011 May 23;171(10):897-903. [doi: 10.1001/archinternmed.2010.527] [Medline: 21263077]
7. Aletaha D, Bécède M, Smolen JS. Information technology concerning SDAI and CDAI. Clin Exp Rheumatol 2016;34(5 Suppl 101):S45-S48. [Medline: 27762194]
8. Kaneko Y, Koike T, Oda H, et al. Obstacles to the implementation of the treat-to-target strategy for rheumatoid arthritis in clinical practice in Japan. Mod Rheumatol 2015 Jan;25(1):43-49. [doi: 10.3109/14397595.2014.926607] [Medline: 24950169]
9. Navarro-Compán V, Smolen JS, Huizinga TWJ, et al. Quality indicators in rheumatoid arthritis: results from the METEOR database. Rheumatology (Oxford) 2015 Sep;54(9):1630-1639. [doi: 10.1093/rheumatology/kev108] [Medline: 25877910]
10. Smolen JS, Aletaha D, Bijlsma JWJ, et al. Treating rheumatoid arthritis to target: recommendations of an international task force. Ann Rheum Dis 2010 Apr;69(4):631-637. [doi: 10.1136/ard.2009.123919] [Medline: 20215140]
11. Kakehasi AM, Duarte A, Brenol CV, et al. Challenges in implementing treat-to-target in rheumatoid arthritis: a perspective from Brazilian rheumatologists. Adv Rheumatol 2024 Aug 26;64(1):63. [doi: 10.1186/s42358-024-00403-w] [Medline: 39187901]
12. Kay J, Nikolov NP, Weisman MH, American College of Rheumatology Committee on Research. American college of rheumatology and food and drug administration summit: summary of the meeting, May 17-18, 2022. Arthritis Rheumatol 2024 Sep;76(9):1323-1333. [doi: 10.1002/art.42864] [Medline: 38622107]
13. Kunii A, Yokogawa N, Oguma T, et al. AB0162 Shared decision making between patient and rheumatologist using a novel touch-panel system for treating rheumatoid arthritis to target. Ann Rheum Dis 2013 Jun;72:A835. [doi: 10.1136/annrheumdis-2013-eular.2485]
14. Prochaska JO, DiClemente CC. Stages and processes of self-change of smoking: toward an integrative model of change. J Consult Clin Psychol 1983 Jun;51(3):390-395. [doi: 10.1037//0022-006x.51.3.390] [Medline: 6863699]
15. Allaire JJ. RStudio: integrated development environment for r. Presented at: Proceedings of useR! 2011,The R User Conference; Aug 16-18, 2011; Koblenz, Germany URL: https://www.r-project.org/conferences/useR-2011/abstracts/180111-allairejj.pdf [accessed 2025-08-24]
16. World Medical Association. World medical association declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA 2013 Nov 27;310(20):2191-2194. [doi: 10.1001/jama.2013.281053] [Medline: 24141714]
17. Ministry of Education, Culture, Sports, Science and Technology, Ministry of Health, Labour and Welfare, Ministry of Economy, Trade and Industry. Ethical Guidelines for Medical and Biological Research Involving Human Subjects. 2021 Mar 23 URL: https://www.mext.go.jp/content/20250325-mxt_life-000035486-01.pdf [accessed 2025-09-10]
18. Seyferth AV, Cichocki MN, Wang CW, et al. Factors associated with quality care among adults with rheumatoid arthritis. JAMA Netw Open 2022 Dec 1;5(12):e2246299. [doi: 10.1001/jamanetworkopen.2022.46299] [Medline: 36508216]
19. Zheng R, Jiang X, Shen L, et al. Investigating clinicians' intentions and influencing factors for using an intelligence-enabled diagnostic clinical decision support system in health care systems: cross-sectional survey. J Med Internet Res 2025 Apr 7;27:e62732. [doi: 10.2196/62732] [Medline: 40194276]
20. Studenic P, Aletaha D, de Wit M, et al. American College of Rheumatology/EULAR remission criteria for rheumatoid arthritis: 2022 revision. Ann Rheum Dis 2023 Jan;82(1):74-80. [doi: 10.1136/ard-2022-223413] [Medline: 36280238]

XSL•FO
RenderX

Abbreviations

ACR/EULAR: American College of Rheumatology/European Alliance of Associations for Rheumatology
bDMARDs: biological disease-modifying antirheumatic drugs
CDAI: Clinical Disease Activity Index cs
DAS28: Disease Activity Score in 28 Joints
DMARDs: conventional synthetic disease-modifying antirheumatic drugs
DMARDs: targeted synthetic disease-modifying antirheumatic drugs
EHR: electronic health record
mHAQ: modified Health Assessment Questionnaire
MiRAi: Medical Support System for Rheumatoid Arthritis
RA: rheumatoid arthritis ts
T2T: treat to target
VAS: visual analog scale

Edited by A Mavragani; submitted 04.04.25; peer-reviewed by A Adekola, O Fatunbi; revised version received 30.07.25; accepted 11.08.25; published 23.09.25.

Please cite as:
Yamamoto M, Kataoka Y, Tsujimoto Y, Nagase F, Iwasaki K, Ito Y, Ikai H, Watanabe T, Yokoyama-Kokuryo W, Takizawa N, Fujita Y
Disease Activity Assessment Frequency in Rheumatoid Arthritis: A Retrospective Observational Study of the Medical Support System for Rheumatoid Arthritis System Implementation
JMIR Form Res 2025;9:e74222
URL: https://formative.jmir.org/2025/1/e74222
doi:10.2196/74222

© Mari Yamamoto, Yuki Kataoka, Yasushi Tsujimoto, Fumika Nagase, Keita Iwasaki, Yuuki Ito, Hiroki Ikai, Tsuyoshi Watanabe, Waka Yokoyama-Kokuryo, Naoho Takizawa, Yoshiro Fujita. Originally published in JMIR Formative Research (https://formative.jmir.org), 23.9.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

A Behavioral Science-Informed Agentic Workflow for Personalized Nutrition Coaching: Development and Validation Study

Eric Yang, MBI; Tomas Garcia, MPA; Hannah G Williams, d.MBA; Bhawesh Kumar, MS; Martin Ramé, MS; Eileen Rivera, MS; Yiran Ma, MS; Jonathan Amar, PhD; Caricia Catalani, PhD; Yugang Jia, MPH, PhD

Verily Life Sciences, 2999 Olympus Blvd, Ste 1000, Dallas, TX, United States

Corresponding Author:
Eric Yang, MBI
Verily Life Sciences, 2999 Olympus Blvd, Ste 1000, Dallas, TX, United States

Abstract

Background: Effective management of cardiometabolic conditions requires sustained positive nutrition habits, often hindered by complex and individualized barriers. Direct human management is simply not scalable, and deterministic automated approaches to nutrition coaching may lack the personalization needed to address these diverse challenges.

Objective: We report the development and validation of a novel large language model (LLM)-powered agentic workflow designed to provide personalized nutrition coaching by directly identifying and mitigating patient-specific barriers.

Methods: We used behavioral science principles to create a comprehensive workflow that can map nutrition-related barriers to corresponding evidence-based strategies. First, a specialized LLM agent to intentionally probe for and identify root causes of a patient's dietary struggles. Subsequently, a separate LLM agent to deliver tailored tactics that were designed to overcome those specific barriers. We conducted a user study with individuals with cardiometabolic conditions (N=16) to inform our workflow design and then validated our approach through an additional user study (n=6). We also conducted a large-scale simulation study, grounding on real patient vignettes and expert-validated metrics, where human experts evaluated the system's performance across multiple scenarios and domains.

Results: In our user study, the system accurately identified barriers and provided personalized guidance. Five out of 6 participants agreed that the LLM agent helped them recognize obstacles preventing them from being healthier, and all participants strongly agreed that the advice felt personalized to their situation. In our simulation study, experts agreed that the LLM agent accurately identified primary barriers in more than 90% of cases (27 or 28/30). Additionally, experts determined that the workflow delivered personalized and actionable tactics empathetically, with average ratings of 4.17 - 4.79 on a 5-point Likert scale.

Conclusions: Our findings demonstrate the potential of this LLM-powered agentic workflow to improve nutrition coaching by providing personalized, scalable, and behaviorally informed interventions.

(JMIR Form Res 2025;9:e75421) doi:10.2196/75421

KEYWORDS

behavioral science; large language model; LLM; nutritional coaching

Introduction

Background

The increasing prevalence of cardiometabolic diseases such as diabetes mellitus and hypertension poses a significant global health challenge [1]. Effective management of these conditions hinges on sustained lifestyle modifications, with nutrition playing a central role [2,3]. Digital health interventions, including mobile apps and online platforms, have emerged as readily accessible solutions. These platforms offer nutrition guidance, meal planning tools, and sometimes one-on-one interactions with human experts who aim to promote long-term adherence [4-6]. However, traditional coaching models often face barriers like limited accessibility, high costs, and difficulties in scaling hyper-personalization.

The integration of large language models (LLMs) into digital health nutrition coaching presents an exciting opportunity. LLMs can revolutionize personalized support, optimize intervention effectiveness, and improve access to care. These models can engage in natural language conversations, enabling them to provide dynamic interactions beyond traditional static information delivery. Notably, their conversational ability allows LLMs to clarify user intent, answer nutrition-related questions with contextual nuance, provide tailored recommendations that align with individual preferences and needs, and adapt their approach based on user feedback [7-9]. This personalized conversational approach, combined with the accessibility and scalability of digital platforms, holds immense promise for transforming nutrition coaching and improving the health of individuals around the world. In addition, recent research has showcased LLMs' potential in gaining user acceptance and

XSL•FO
RenderX

delivering accurate domain-specific knowledge, specifically in nutrition applications addressing cardiometabolic conditions [10-13].

Yet, simply delivering nutrition information through general foundation models is unlikely to achieve sustained behavior change. Leveraging the principles of behavioral science is critical for designing sustainable interventions that address the physical, psychological, and social factors influencing dietary choices. Frameworks such as the capability-opportunity-motivation-behavior (COM-B) model are instrumental in understanding the multidimensional factors that contribute to current behavioral patterns [14]. Once the factors preventing behavioral change are better understood, additional frameworks such as the Behavioral Change Wheel, the Behavioral Change Taxonomy (BCT) of 93 hierarchically clustered techniques, and the Easy, Attractive, Social, and Timely (EAST) framework can provide the right approach and tactics on how to promote change [14-16]. These frameworks offer insights into designing interventions that leverage psychological principles to enhance engagement, simplify behavior, make actions rewarding, and trigger positive associations. Combining these complementary frameworks, addressing both the root causes and offering solutions, is crucial to realize LLMs' potential in delivering sustainable behavioral change.

Related Work

The application of LLMs specifically to nutrition is a rapidly emerging field. Foundational work has demonstrated their utility for core nutritional analysis tasks, such as generating personalized meal plans based on caloric targets and decomposing compound ingredients for precise analysis [1,2,17,18]. Building on these capabilities, a second stream of research focuses on developing interactive, conversational nutrition assistants. Systems like ChatDiet, for example, use an orchestrator with personal and population data models to deliver highly personalized food recommendations [3,19]. Other work explores multimodality, such as the Purrfessor chatbot which integrates visual meal analysis with conversational advice to enhance user engagement [4,20]. However, while these systems show increasing sophistication in processing nutritional data, their design often overlooks the principles required for sustained behavior change.

Recent research has also explored the potential of LLMs for digital coaching in health behavior change [21-26]. Some literature has reported on interventions that often rely on high-level motivational strategies, utilizing LLMs to deliver advice primarily through broad motivational interviewing techniques or empathetic tones [21,23]. Others have laid groundwork to infuse behavior science principles to the guidance provided by LLMs, incorporating frameworks like COM-B to identify barriers, but their reliance on single conversational turns for barrier classification may limit the depth of their assessments [22]. A separate group of studies has explored LLMs' ability to leverage external tools effectively to foster user engagement and self-reflection [24,25], or has investigated LLM capabilities to segment users in order to set up different action courses [26].

These separate points of focus in the relevant literature highlight one-by-one the multiple components required for effective full coaching strategies, but several opportunities remain to enhance the integration of behavioral science principles with LLM capabilities, in order to gather individualized user insights and ultimately inform targeted, actionable behavior change tactics. There may be benefits in framing the coaching strategy as a succession of focus points, starting with quantifying barriers from a behavioral science perspective using iterative motivational probing processes, followed by offering specific tactics tailored to individual needs. Overall, there is significant potential for a nuanced integration of behavioral science expertise in the design and development of LLM-powered digital coaching interventions to support more effective and sustainable behavior change.

Our aim in this study was to introduce a novel approach in leveraging LLM-powered nutrition coaching for cardiometabolic condition management by advancing the integration of behavioral science principles and developing a comprehensive, scalable, and expert- and user-validated framework. Our research directly addressed current limitations in the field by creating a multi-agent conversational workflow powered by a deep understanding of nutrition-related barriers and corresponding strategies that directly mitigated the barriers. Through motivational probing, our approach directly identified the root causes of dietary behavior rather than addressing the surface-level symptoms. It fostered a personalized coaching experience, moving beyond high-level motivational techniques and offering targeted tactics supported by behavioral science research. Our approach also enabled a learning system, where barrier-strategy mappings could be tailored to specific individual context, habits, and tactic adoption. We summarize the steps in our study as follows.

Study Overview

User Research and Literature Review for Comprehensive Barrier Identification and Strategy Mapping

We conducted a user research study to identify common nutrition-related barriers experienced by cardiometabolic patients. In addition, we performed a comprehensive literature review of academic research papers and internal reports. We then synthesized these barriers into an overarching set of main barriers encompassing all aspects of nutrition goal achievement. Furthermore, we mapped these barriers to a comprehensive set of strategies and behavioral science tactics, enabling a tailored approach to addressing barrier-specific challenges. This went beyond existing research by providing a more holistic and nuanced understanding of nutrition barriers and their corresponding solutions, paving the way for targeted interventions.

Multi-Agent Workflow Design for Personalized Coaching

Building on our comprehensive barrier and strategy mapping and insights from user research, we designed a multi-agent workflow where a specialist LLM agent was tasked with probing and classifying barriers through conversation, while another specialist LLM agent carried out strategies and offered specific tactics. We opted for a multi-agent approach after preceding

XSL•FO
RenderX

work (E Yang, MBI, unpublished data, September 2024) showed suboptimal results using a single agent across all the coaching tasks (barrier identification and strategy execution). The present results show that this multi-agent approach improved upon other existing single-agent strategies.

Real-World Validation: Collecting Impressions From Cardiometabolic Populations

To ensure practical application and effectiveness, we validated our workflow with participants representative of the cardiometabolic populations we aim to serve. This direct validation provided evidence for the efficacy of our system in addressing real-world nutrition challenges, enhancing the credibility and impact of our research. The real-world validation set our work apart from purely simulation frameworks.

Benchmarking and Expert Annotation

Along with behavioral science experts, we used a granular benchmark for evaluating the performance of our LLM agents in various stages, including barrier identification, tactic delivery, and overall conversational attributes. Expert annotations on these metrics provided a rigorous evaluation of our system's performance.

Simulation-Based Evaluation at Scale

We curated real patient vignettes from our user study to generate large-scale simulated conversations of various realistic barrier situations. These data were then evaluated by LLM auto-evaluators, allowing for a scalable and accurate assessment of our system's performance across a diverse range of scenarios. This simulation-based evaluation approach provided a scalable method for evaluating the generalizability of our LLM-powered coaching system, setting a standard for evaluating similar artificial intelligence (AI)-driven interventions.

Methods

Ethical Considerations

Participants provided informed consent and were compensated at the rate of US $100 per hour. Verily Life Sciences research committee performed the ethical review and approved the study plan, declaring the research and resulting dataset exempt. For privacy and confidentiality protection, data were deidentified for analysis and subsequent publication.

User Research Study

Research Objectives

Conducting a user research study was crucial for designing a coaching workflow with the specific needs of individuals with cardiometabolic conditions in mind. By directly involving the target population, we aimed to gain a deeper understanding of realistic nutrition challenges and preferences, ensuring that the system is designed to provide relevant, practical, and effective guidance.

This research study aimed to uncover key user insights necessary for building a patient-centered AI coaching workflow. Specifically, we sought to answer 2 main research questions: (1) What user motivations, characteristics, and challenges must our workflow understand and adapt to in order to be effective? (2) How can an AI coach's character and conversational patterns inspire trust and engagement while delivering behavioral science strategies?

Participants

A total of 16 participants were recruited using third-party services, and following a purposeful sampling strategy, selecting for recent cardiometabolic diagnosis, and otherwise diversity in demographics, US geographic regions, health conditions, attitudes toward health care and technology, and familiarity and comfort with AI. Participants were recruited for sessions across a 12-week period.

Procedure

Each participant took part in one-on-one, semi-structured qualitative interviews led by the research team. These interviews focused on participants' medical histories, the barriers they encountered when working toward their health goals, and their responses to different conversational approaches, as well as their familiarity and general perceptions about AI. Participants shared their experiences and reflected on the specific challenges they faced in managing their health, while also discussing strategies they had found useful. The first step in our research plan, before addressing the second research objective, consisted of participants engaging in interactive exercises where they conversed with 2 types of LLM agents: a supportive agent and an assertive agent (Multimedia Appendix 1, section A). The supportive agent facilitated conversation by treating the user as the expert on their body and experience. This agent encouraged self-reflection and demonstrated high levels of compassion, curiosity, and affirmation. Its tone was plain-spoken, easygoing, and patient, kindly taking direction from the user. In contrast, the assertive agent adopted a directive approach, positioning itself as the resident expert with high energy, authoritative knowledge, and a strategic mindset. This agent aimed to empower the user by being assertive and eager to motivate change. Both agents are powered by Gemini-1.5 Pro, a commercially available model known for its conversational and reasoning capabilities [27]. Their character and tone were guided by specific phrasing and conversational styles in single prompts. From this exercise, we selected the agent type to be carried forward to assemble our agentic workflow. The sessions were conducted on a web-based chat interface, under the observation of user-experience researchers; AI agents were described to participants as "genies" that could work autonomously or in collaboration with clinicians.

To extract output of these study user sessions, analysts used a modified-grounded approach to qualitative analysis, using prior behavior change theories and user experience expertise in the data analysis process. The process began with qualitative coding where analysts labeled sections of data with short words and phrases, capturing what was being communicated. Then, similar codes were grouped together into broader categories that represent recurring patterns. These themes included promoters and detractors of trust and engagement, attitudes toward AI, needs, and expectations. Then, analysts reviewed and refined each theme to ensure fit with matching data and accuracy. Assessing each theme and groups of themes enabled the

XSL•FO
RenderX

development of key insights on preferred conversational style. Overall, this approach allowed for a focused exploration of specific aspects of AI desirability and behavior change motivation, while grounding observations on the data itself.

Behavioral Science Agentic Workflow Design

Curation of Barriers and Behavioral Science Strategies

Recognizing that the barriers described by the participants in the user study may not be all-encompassing, we additionally researched the challenges faced by people with cardiometabolic conditions in the literature and marketplace. We analyzed numerous research papers and reports, covering areas like nutrition, medication adherence, exercise, and goal setting. This helped us identify over 100 total barriers individuals experience. Next, we used affinity mapping to organize these barriers, grouping them based on common themes and patterns. By prioritizing the most frequent and impactful challenges, we narrowed our focus to 28 key barriers (Multimedia Appendix 1 section B).

To uncover effective strategies and tactics for addressing these barriers, we conducted a comprehensive review of existing behavioral science frameworks and developed a solutions repository. First, we examined established frameworks, seeking strategies and tactics with proven efficacy in overcoming similar barriers within the health domain. Our review encompassed well-known models like the COM-B model, the BCT Taxonomy, the EAST framework, and other behavioral change models [14-16] (Multimedia Appendix 1 section B). Following this extensive mapping exercise, we curated a final selection of a repository containing over 50 strategies and 100 tactics that are mapped to the 28 barriers. This repository comprehensively linked each identified barrier to a range of potential solutions, offering guidance on optimal implementation to maximize impact. There are popular frameworks and pre-defined sequences for intervention design, which strictly connect the COM-B model to the Behavior Change Wheel to guide the selection of "intervention functions" and "policy domains." Although we recognize the value of such approaches, we deliberately chose not to explicitly link our workflow to any singular pathways at this stage. We found that the incorporation of other tools, like the 93 Behavioral Change Techniques Taxonomy v1 (BCTv1) and the EAST framework, could be not only suitable but even more appropriate for our population of interest and for our agent capabilities. Ultimately, this decision allowed for greater flexibility and innovation to optimize and personalize the way our agent responded to our population of interest. We could explore the full potential of our future capabilities, ensuring no constraints by any predetermined intervention pathways. A few examples of barriers, strategies, and tactics mapping are shown in Table 1.

Table . Examples of mapping of barriers, strategies, and tactics.

Barriers	Strategies[a]	Tactics[b]
Decision fatigue: The mental exhaustion from making too many choices can lead to poor, sub-optimal future decisions. For example, "As a working parent I have to deal with so many things that I don't want to think about cooking when I get home." [28]	Heuristics: Mental heuristics are cognitive shortcuts our brains use to simplify complex situations and make quick decisions [29].	• Rules of thumb: Offer the user a set of actionable principles to help them automate tasks. For example, "Always, fill one third of your plate with lean protein." • Default: Encourage the user to pick an option and use it as a default to save time and mental effort. For example, "Let's set Tuesday as a kale salad day." [29]
Present bias: It is the user tendency to overvalue immediate rewards over future, larger rewards. For example, "It's the weekend! Let's start with dieting next week." [30]	Future self. Future self is a vivid and emotional connection with a future version of ourselves that affects our intention to engage in a future behavior [31].	• Mental rehearsal of successful performance: Encourage the user to practice visualizing themselves successfully performing the desired behavior in realistic scenarios. For example, suggest they vividly imagine eating greens and feeling light to continue with the day [15].

[a]Strategy: a high-level behavioral science concept aimed to mitigate a barrier.

[b]Tactic: concrete, tangible, and actionable step to execute on a given strategy.

Multi-Agent Workflow Description

This section details the architecture of the multi-agent LLM coaching workflow, aiming to provide tailored nutrition guidance to cardiometabolic patients by identifying and addressing their unique barriers to achieving their nutrition goals. As shown in Figure 1, the system operated through 2 core agents, seamlessly integrated to offer a unified user experience: the barrier identification agent and the strategy execution agent. Both agents were powered by Gemini-1.5 Pro and were prompted to converse in the supportive manner preferred by participants in the user study. We developed the system in Python 3.9 using the Vertex AI API for Gemini. The specific instructions for each agent were defined in system prompts passed during the model's instantiation, and we used a chat session object to maintain the conversational context.

XSL•FO
RenderX

Figure 1. Summary of study steps. (A) The multi-agent artificial intelligence (AI) coach workflow was infused with behavioral science principles to address patients' barriers toward their nutrition goals. The workflow consisted of 2 core agents, the barrier identification agent and the strategy execution agent. (B) To assess the performance of our workflow at scale, we developed patient simulators that portray validated nutrition vignettes drawn from real cardiometabolic patient profiles. (C) Real cardiometabolic patients and patient simulators interacted with our multi-agent AI coach workflow. The conversation experiences were evaluated via survey, expert annotation of dialogs, and were additionally compared to a baseline large language model (LLM) via auto-evaluation.

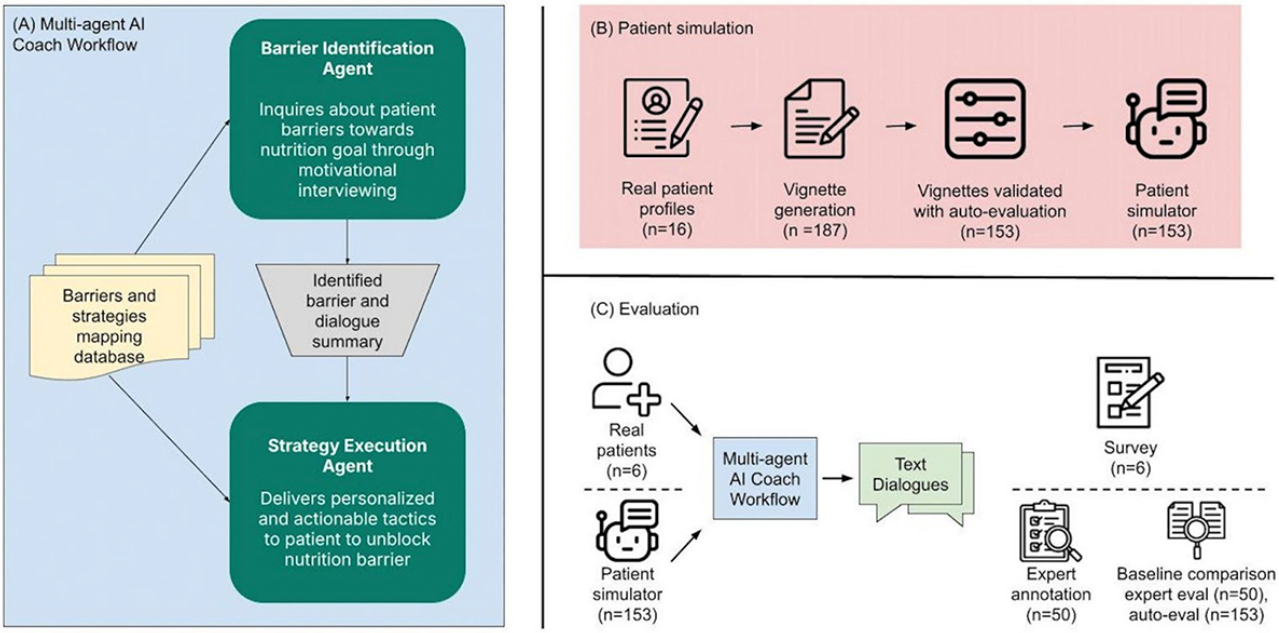

The barrier identification agent initiated conversations by inquiring about the user's current nutrition goals and progress. It then employed motivational interviewing techniques to explore the specifics of the user's nutrition habits, with a focus on identifying the most prominent barrier that may be hindering progress toward their stated goal. The agent was equipped with a predefined taxonomy of 28 barrier concepts, provided within the prompt alongside detailed descriptions and examples. Through iterative dialog, the agent analyzed the user's responses to classify the identified struggles into one of the barrier concepts. While a user could exhibit multiple barrier concepts, the agent was instructed to prioritize and focus on the most prominent one displayed by the user. By tasking the agent with the identification of a primary barrier (as opposed to relying on users themselves), we circumvented burdening users with triaging technical barrier concepts with which they may be unfamiliar, or sorting through voluminous associated tactic lists that may be overwhelming. Once the agent determined that sufficient information had been gathered for barrier classification, it generated an internal summary of the conversation and the identified barrier concept, which was then relayed to the strategy execution agent. The prompt used by the barrier identification agent can be found in Multimedia Appendix 1, section C.1.

The strategy execution agent received the conversation summary and the identified barrier concept from the barrier identification agent. The agent relied on a structure where each barrier was mapped to a strategy (a high-level behavioral science concept aimed to mitigate a barrier) and, ultimately, to a list of tactics (concrete, tangible, and actionable steps to execute on that strategy) to be carried out by the strategy execution agent. Therefore, each barrier concept eventually mapped to a set of potential tactics, each with examples and an associated execution sequence outlining mandatory and optional tactics, as well as their prioritization. This way, once the agent received the input about the barrier, it retrieved the corresponding tactics and execution sequences from a predefined table. The agent then engaged in further conversation with the user, drawing upon the preceding conversation summary and the prescribed tactics. For the optional tactics, the agent dynamically adapted its offerings based on user responses and perceived acceptability and effectiveness of the previously deployed tactics. The conversation concluded when the agent determined that the user was sufficiently equipped with the necessary tools to overcome their identified barrier and progress toward their nutrition goals. The prompt used by the strategy execution agent can be found in Multimedia Appendix 1, section C.2.

Agent orchestration was defined as when the back-end functionality relied on 2 core distinct agents, the user interacted with the system as a single, continuous AI-powered nutrition coach. This seamless transition between barrier identification and strategy execution was facilitated by the orchestration of information exchange between agents. Crucially, conversation summaries and key outputs, including identified barriers related to nutrition, were explicitly transferred between the agents, ensuring the preservation of all necessary context while excluding irrelevant banter. This modular design allowed each agent to specialize in its respective task, enabling more focused model improvements for both barrier identification and strategy execution. An example of the agent orchestration can be found in Multimedia Appendix 1, section D.2.

XSL•FO
RenderX

Evaluation of User Impressions of Workflow

Research Objectives

Having implemented the AI coaching workflow grounded on the insights from the initial user research, we conducted further sessions to evaluate users' impressions of the workflow through direct interactions. This phase aimed to test the real-world applicability and accuracy of the system's barrier identification and strategy execution capabilities, providing a critical assessment of how well the workflow translates theoretical principles into practical outcomes. Engaging with users also helped us uncover any usability or communication insights that may not have been apparent in agentic workflow and prompt development. Key questions guiding this evaluation were as follows:

Effectiveness of Behavioral Science-Informed Workflow

How did users who are managing health concerns respond to our workflow infused with behavior science frameworks? How effectively did our workflow help users identify obstacles that prevent them from being healthier? Were the strategies and tactics offered by our workflow easy for users to put into action? How did our workflow impact users' motivation and confidence to make positive changes in their health?

Building Trust and Engagement Through AI Interactions

Did our AI coach's character and personality inspire trust and engagement? To what degree did users feel supported by our coaching workflow?

Participants

We selected a subset of 6 participants out of the 16 in the user research study described in the prior section, for direct conversational interaction with our LLM coaching workflow in a single session. This scaled-down sample allowed us to conduct the study in a timely fashion. The subset was selected on the basis of diversity with respect to AI comfort levels and stated motivation levels. The participants were primed to think of challenges they might face achieving a nutrition goal they had set, and to channel that while interacting with our coaching workflow. After the experience, the participants were asked to reflect on their interactions through qualitative interviews and to complete a customized survey for this study. Analysts used a modified-grounded approach to qualitative analysis.

Simulation Study

Patient Vignettes

To assess the performance of our multi-agent LLM coaching workflow at a larger scale, we conducted a simulation study with the high-level workflow illustrated in Figure 1. Leveraging patient profiles curated from the user research study that contained patient lifestyle context and medical history, we identified the prominent nutrition barriers for each profile, aligning them with our predefined taxonomy of 28 barrier concepts. This involved a manual process of translating patients' self-reported nutrition struggles into corresponding barrier concepts. Subsequently, for each identified barrier within each profile, we crafted a detailed patient nutrition vignette paragraph with a separate base Gemini model. These vignettes aimed to vividly depict the specific manifestation of that barrier within the context of the individual's profile. Providing prewritten nutrition vignettes allowed us to:

Focus the Patient Simulator

Our approach enabled the downstream patient simulator to concentrate on portraying the specific barrier and engaging in a meaningful conversation, without the need to generate complex nutrition narratives dynamically.

Enhance Control and Consistency

We enabled greater control over the simulated conversations, minimizing the risk of inconsistencies or incoherent narratives that might arise from on-the-fly story generation.

Isolate Barrier Identification Impact

Overview

Our approach allowed us to isolate and assess the system's ability to accurately identify and address specific barriers, independent of other confounding factors within a patient's profile.

In total, we simulated a wide range of scenarios across diverse patient profiles, nutrition goals, and barriers, resulting in 187 total simulated vignettes. We provide a sample vignette generated from a barrier concept for a patient profile in Multimedia Appendix 1, section D.1.

To ensure the quality and relevance of the generated vignettes with respect to the simulated behavioral barrier, we leveraged OpenAI's GPT-4o (gpt-4o-2024-08-06) model in an LLM-as-judges framework, also known as auto-evaluation [32]. The GPT-4o model was intentionally chosen as the vignettes were generated by a Gemini model. The auto-evaluator received two inputs: the target barrier to be simulated and the generated patient vignette. It then evaluated the vignette across four dimensions: evidence, realism, completeness, and leakage.

Evidence (High/Medium/Low)

The extent to which the vignette provided clear indications that the patient's behavior or thoughts were influenced by the target barrier.

Realism (High/Medium/Low)

The plausibility and believability of the depiction of the target barrier, reflecting how it might manifest in a real person's life.

Completeness (High/Medium/Low)

The sufficiency of details provided in the vignette to fully understand the impact of the target barrier on the patient's ability to achieve their nutrition goals.

Leakage (Yes/No)

Whether the vignette directly mentioned the technical term of the target barrier. It was important that vignettes do not explicitly contain the technical terms, which may skew the barrier identification task down the line.

Conversation Simulation

Having established a robust set of patient vignettes, we proceeded to simulate dialogs between a Gemini-powered

XSL•FO
RenderX

patient simulator and our LLM coaching workflow for each of the 153 higher quality vignettes. The patient simulator was provided with two key inputs: the generated vignette from the corresponding patient profile and barrier, and the communication style observed for that patient during the user research study. Notably, the patient simulator's nutrition barriers were only portrayed via the vignette without explicit mentioning of any technical behavioral science barrier terms. The patient simulator, guided by these inputs, engaged in a conversation with the AI coach. The dialog continued until the AI coach determined, based on its internal logic and the patient's responses, that the patient had reached a higher level of preparedness to address their nutrition-related challenges. This endpoint represented a complete coaching interaction within the simulation framework. An example of simulated conversations can be found in Multimedia Appendix 1, section D.2.

Expert Assessment

Overview

Evaluating the quality and effectiveness of the simulated coaching conversations was a crucial step in assessing the overall performance of our multi-agent LLM workflow. We leaned on our behavioral science expertise to develop an evaluation rubric encompassing 5 key dimensions: barrier identification accuracy, tactic comprehensiveness, tactic personalization, tactic actionability, and conversation empathy. These dimensions were chosen to reflect the core competencies required for effective digital coaching in the context of motivating dietary behavioral change.

Barrier Identification Accuracy

Given the list of 28 nutrition barrier concepts, the AI coach identified the correct patient barrier. Accurate identification of the patient's primary barrier is paramount for delivering tailored and effective interventions. Rating options: "Yes" or "No."

Tactic Comprehensiveness

The AI coach delivered all the instructed primary tactics to the patient via clear descriptions or relevant examples in an understandable way. Ensuring the delivery of all intended coaching tactics is crucial for maximizing the potential impact of the intervention. Rating options: "Yes" or "No."

Tactic Personalization

For the tactics that were delivered, the AI coach made the tactics personalized to the patient's unique context. Personalizing coaching tactics to the individual's unique context and circumstances enhances engagement and promotes behavior change. Rating options: 5-point Likert scale.

Tactic Actionability

The AI coach discussed actionable steps to help patients overcome their barriers toward their nutrition goals. Providing clear, actionable steps empowers patients to translate intentions into concrete behaviors. Rating options: 5-point Likert scale.

Conversation Empathy

The AI coach provided encouragement and motivation to the patient empathetically. Expressing empathy and providing emotional support fosters a positive and trusting coach-patient relationship. Rating options: 5-point Likert scale.

The corresponding conversations that were simulated for the 50 randomly selected generated vignettes were assessed by human experts with extensive academic training and professional experience in behavioral science. Specifically, two behavioral science experts independently labeled 30 simulated conversations each, with 10 overlapping conversations used to assess inter-rater reliability.

Comparative Study

To evaluate the advantages of our multi-agent LLM workflow infused with behavioral science principles, we conducted a comparative study against a single base Gemini model that lacked explicit behavioral science knowledge infusion and did not employ a multi-agent approach. The base model was instructed to assist patients in overcoming their nutrition challenges by leveraging its inherent capabilities to apply motivational interviewing and behavioral science tactics without structured guidance. We simulated coaching conversations with the same patient simulator used previously.

Then, a GPT-4o auto-evaluator was employed to determine preference between the 2 conversation sets based on behavioral science criteria. To ensure a fair comparison, we measured the conversation lengths across both conversation sets, ensuring that they contained a comparable number of characters to control for LLM preference for longer contexts. The order in which the conversations were presented to the evaluator was alternated to additionally control for any position bias.

To generate an additional reference data point, the same two behavioral science experts from the Expert Assessment exercise independently and blindly evaluated a randomly shuffled subset of 30 pairs (10 overlapping) of these simulated conversations, namely, conversations from our multi-agent LLM workflow and from the basic Gemini model.

Results

User Research Study

We enrolled N=16 participants in this portion of our study (Table 2).

XSL•FO
RenderX

Table . Characteristics of the study participants (N=16).

Characteristics	Value, n (%)
Sex	
Female	8 (50)
Male	8 (50)
Age (y)	
<45	5 (31.3)
45 - 65	9 (56.3)
>65	2 (12.5)
Race/ethnicity	
American Indian/Alaskan Native	1 (6.3)
Black/African American	6 (37.5)
Hispanic/Latino	2 (12.5)
White	7 (43.8)
Recent[a] cardiometabolic diagnoses	
Hypertension	9 (56.3)
Prediabetes	4 (25)
Type 2 diabetes	10 (62.5)
Hyperlipidemia	2 (12.5)
Obesity	15 (93.8)

[a]Diagnosis within the last 6 months, n=13; diagnosis more than 6 months ago, n=3.

The user research study revealed crucial insights into the barriers and motivations that should be integrated into the design of the AI coaching workflow. Participants shared detailed accounts of their physical, psychological, and social challenges in achieving their health goals. Common themes that emerged included balancing competing priorities, dealing with physical limitations, and struggling with low self-efficacy. For example, participants frequently expressed feeling overwhelmed by the demands of daily life, which made it difficult to sustain motivation to work on their health over time.

Additionally, participants highlighted the importance of feeling understood and empowered by a coach. A preference emerged for the supportive agent, as most participants felt more comfortable and engaged with its empathetic approach. They noted that the supportive agent fostered trust and provided a space for self-reflection. Although a few participants appreciated the assertive agent's high-energy style, the majority found it to be less aligned with their desire for a collaborative and self-empowering experience. This feedback was instrumental in refining the AI coaching workflow, which ultimately prioritized the compassionate, user-driven style of the supportive agent to promote long-term engagement and behavior change.

Informed by these findings, we established two guiding principles for the agentic workflow design. First, the workflow must move beyond generic advice to probe and understand the root causes of a user's barriers, ensuring that subsequent strategies are practical, incremental, and tailored to their personal context. Second, every interaction must be delivered through the supportive, conversational style that participants confirmed was essential for fostering trust and engagement.

User Impressions of Workflow

Effectiveness of Behavioral Science-Informed Workflow

The AI coaching workflow, grounded in behavioral science principles, was found to be effective in addressing barriers users have toward healthier habits. First, the AI coach was successful in helping users identify specific barriers to their health. Five out of 6 participants agreed that the assistant helped them recognize obstacles that prevented them from being healthier. Participants also indicated that their interactions with the AI coach were informative. When asked whether they learned something new about their health habits, 4 out of 6 participants strongly agreed, and the remaining somewhat agreed with the sentiment. As a participant reflected, the agent's approach of leading with additional questions and building from user experiences enabled "more refined, bite-sized" understanding, enhancing quality of the nutrition motivations and barriers insights collected. Furthermore, participants responded positively to the strategies and tactics provided by the AI coach. All participants strongly agreed that the AI coach's advice felt personalized to their situation, reflecting its ability to adapt to users' unique needs. Notably, all participants agreed that the AI coach's advice was easy to put into action, highlighting its practicality. Participants appreciated the focus on manageable, small changes, with one stating, "Making small changes makes a big difference," highlighting the perceived effectiveness of the behavior change strategies recommended by the AI coach. The AI coach's ability to offer novel suggestions tailored to

XSL•FO
RenderX

individual preferences was highlighted by participants, such as when one participant appreciated the recommendation to swap quinoa porridge for oatmeal, and another found value in being advised on alternatives to high-sugar protein drinks. The impact of these strategies on user motivation was significant, with all participants agreeing that the AI coach increased their confidence to make positive changes in their health. Each participant also reported feeling more motivated after their conversations due to excitement for having a plan to implement manageable, novel, small changes.

Building Engagement Through AI Interactions

Overall, the AI coaching workflow fostered relationship building and engagement through their conversational style and empathetic personality. All participants strongly agreed that they felt supported by the AI coach and were interested in having future conversations about their health, indicating a high level of engagement. One participant noted, “I could sit there and talk to that thing all day.” Furthermore, participants frequently described the AI coach as having a “friendly” and “human-like” demeanor, with one participant mentioning, “It did feel like you’d spoken with somebody, like you had an actual conversation with somebody.” As one participant noted, the AI coach “seemed to be friendly, kind of human-like,” creating a sense of connection and understanding that facilitated deeper discussions about their nutrition habits. The AI coach’s ability to create a supportive environment was evident, with all participants agreeing that they felt comfortable discussing their health issues with them. This sense of comfort was further supported by the finding that the great majority of participants strongly agreed that their conversations were engaging. The AI coach’s empathetic and person-centered communication style not only facilitated open dialog but also encouraged continued user interaction, which is crucial for sustained engagement and positive health outcomes. The full survey results can be found in Multimedia Appendix 1, section E.

Simulation Study

Assessment of the Patient Vignette Auto-Evaluator

To calibrate the auto-evaluator’s performance, we conducted an adversarial analysis. We evaluated the auto-evaluator on a patient vignette paired with a randomly chosen behavioral barrier different from the one simulated in the vignette. This mismatch was designed to assess the auto-evaluator’s sensitivity to inconsistencies between the intended barrier and the generated narrative. We hypothesized that scores for these adversarial cases would be significantly lower, reflecting the lack of alignment between the vignette and the incorrect barrier. The results of this analysis are presented in Table 3.

Table . Auto-evaluator performance on matched (N=187) and mismatched vignettes (N=187).

	Matched vignette, n (%)			Mismatched vignette, n (%)		
Dimension	High	Medium	Low	High	Medium	Low
Evidence	184 (98.4)	3 (1.6)	0	36 (19.3)	32 (17.1)	119 (63.6)
Realism	187 (100)	0	0	84 (44.9)	94 (50.3)	9 (4.8)
Completeness	156 (83.4)	31 (16.6)	0	19 (10.2)	46 (24.6)	122 (65.2)

Of the 187 simulated vignettes, 153 of them received high marks across evidence, realism, and completeness dimensions by the auto-evaluator. These higher quality vignettes were selected for downstream conversation simulation. As hypothesized, the auto-evaluator assigned significantly lower scores to the mismatched vignettes across all dimensions. This indicates the auto-evaluator’s ability to discern between accurately and inaccurately represented barriers, supporting its validity for assessing vignette quality. In addition, there was no barrier term concept leakage across all cases. Our vignette generation and validation processes provided a robust foundation for our simulation study, ensuring realistic and diverse scenarios for evaluating the effectiveness of our multi-agent LLM coaching workflow.

Expert Assessment

The evaluation of the simulated coaching conversations by human experts provided important insights into the performance of our multi-agent LLM workflow. Full results are shown in Table 4. For Barrier Identification Accuracy, the experts agreed that the AI coach accurately identified the primary patient barrier in greater than 90% of cases (in 27/30 cases for 1 reviewer and 28/30 for the other), and the cases with low accuracy score were due to another barrier being deemed more prominent over the one selected by the agent, not to an inability of the agent to identify a barrier. For Tactic Comprehensiveness, the AI coach successfully delivered all instructed primary tactics in 70% of conversations (21/30) labeled by expert 1% and 90% of conversations (27/30) labeled by expert 2. The inter-rater reliability for both dimensions was high, with agreement percentages of 80%, indicating strong consistency between the experts’ assessments. In addition, the agentic workflow received high scores across the remaining dimensions. For Tactic Personalization, the AI coach demonstrated a strong ability to tailor its coaching tactics to the individual’s unique context, with average ratings of 4.38 and 4.79 on a 5-point Likert scale. Tactic Actionability was also rated highly, with average scores of 4.17 and 4.59, reflecting the clarity and feasibility of the steps recommended by the AI coach. Finally, Conversation Empathy received high marks, with average scores of 4.58 and 4.76, indicating that the AI coach was perceived as empathetic and supportive, effectively fostering a positive coaching relationship. While Expert 2 gave slightly higher ratings on average, the absolute difference in ratings for overlapping cases was minimal, reinforcing the reliability of the evaluations. These results highlight the AI’s coach capacity to deliver personalized, actionable, and empathetic coaching aligned with core behavioral science principles.

XSL•FO
RenderX

Table . Expert evaluation on simulated conversations.

	Expert 1 average rating (n=30)	Expert 2 average rating (n=30)	Interrater reliability (%) and average absolute difference in rating
Dimension (Yes=0; No=1)			
Barrier identification accuracy	0.93	0.90	80[a]
Tactic comprehensiveness	0.70	0.90	80[a]
Dimension (5-point Likert), mean (SD)			
Tactic personalization	4.38 (0.94)	4.79 (0.49)	0.78[b]
Tactic actionability	4.17 (1.10)	4.59 (0.63)	0.89[b]
Conversation empathy	4.58 (0.73)	4.76 (0.44)	0.56[b]

[a]Values correspond to interrater reliability.

[b]Values correspond to average absolute difference in rating.

Comparative Study

As seen in Table 5, the results of the comparative study demonstrated a preference for the behavioral science-informed workflow. The GPT-4o auto-evaluator preferred the conversations generated by the multi-agent workflow in 102 out of 153 (66.7%) cases, compared to 51 out of 153 (33.3%) cases for the single-agent base model. The blinded human expert evaluation of a subset of these conversations produced consistent results with the auto-evaluator, with both experts showing preference for the multi-agent workflow (in 66%-73% of cases) as compared with the single-agent (26%-33% of cases). To ensure a fair comparison, we report the conversation lengths, with average character counts of 3825 for the multi-agent workflow and 3904 for the base model. In addition to alternating the order presented to the auto-evaluator, the mitigation of length bias was important for ensuring that the auto-evaluator's preferences were based on the quality of behavioral science content rather than superficial factors.

Table . Comparative evaluation on simulated conversations with behavioral science agentic workflow and base Gemini.

	No. of conversations preferred, n (%)		Conversations avg. Char. length (N=153), mean (SD)
	Autoevaluation (N=153)	Human expert review (N=30)	
Behavioral science agentic workflow	102 (66.7)	• Expert 1: 22 (73.3) • Expert 2: 20 (66.7)	3825 (1678)
Base Gemini	51 (33.3)	• Expert 1: 8 (26.7) • Expert 2: 10 (33.3)	3904 (2056)

Anecdotally, our review of the conversations revealed that the base model frequently provided more generic advice, often suggesting alternative goals or broad solutions rather than leveraging nuanced behavioral science tactics tailored to overcome the patient's original nutrition goals. In contrast, the multi-agent workflow generated more specific and contextually relevant strategies, highlighting the benefits of structured behavioral science integration in digital coaching conversations.

Discussion

Overall Learnings

Our findings offer important implications for the integration of behavioral science principles into multi-agent AI coaching workflows, particularly for managing cardiometabolic conditions. The results demonstrate that our novel workflow, which utilized comprehensive barrier identification and strategy mapping, has the potential to significantly enhance the effectiveness of digital coaching interventions. The real-world user study confirmed that the workflow provides relevant, personalized support that resonates with users, fostering trust and engagement. Additionally, the introduction of structured patient simulators allowed us to systematically evaluate AI's performance across diverse scenarios, providing a scalable method to refine and validate the system's approach. By moving beyond surface-level motivational techniques, our approach directly targeted the root causes of nutrition-related behaviors, offering personalized and tailored coaching experiences. The strong preference for our workflow, as evidenced by both human expert evaluations and auto-evaluation, underscores the potential of structured, behaviorally informed AI coaching systems to deliver more nuanced and relevant guidance. This study sets a precedent for the development of agentic workflows that can effectively adapt to individual patient contexts, fostering sustained engagement and positive health outcomes through a deep understanding of patient-specific barriers and strategies.

XSL•FO
RenderX

System Design Implications Beyond This Research Work

The implications of our findings extend beyond the cardiometabolic nutrition domain, suggesting that personalized AI coaching systems have the potential to play a transformative role in the future of digital health interventions. By aligning coaching systems with individual patient needs, these AI models can improve patient engagement, which is often a major barrier in digital health tools. Our results reinforce the growing importance of integrating behavioral science into AI systems to ensure that the interventions proactively help patients navigate their health goals in a structured, effective way. This also opens the door for AI to contribute more meaningfully to other areas of health management, such as physical activity management, mental health support, and preventive care.

Limitations and Future Research Opportunities

Our study has several limitations that highlight opportunities for future research. First, the sample size for both the user research study and human expert evaluation was relatively small, which may limit the generalizability of the findings. Future research should replicate these studies with larger, more diverse populations over time to confirm the effectiveness of the multi-agent LLM workflow in broader contexts. Future research could also refine barrier identification by introducing direct questioning by the agent to the user about whether they feel equipped to overcome their barriers, rather than relying on conversational inferences. In addition, although the use of auto-evaluation enables scalable assessment, there are known biases when using LLMs as judges despite our best efforts to mitigate them [33-36]. Incorporating a diverse set of human evaluators could provide richer insights into the system's real-world applicability and user experience. Moreover, considering feedback mechanisms and information retrieval approaches will create additional opportunities to improve the workflow. For instance, user feedback loops can enable reinforcement learning techniques to improve model performance, and retrieval-augmented generation approaches can ensure that the agent's knowledge base is up-to-date and based on latest evidence [37-39]. This optimization research could focus on the long-term impact of AI-driven coaching, measuring (and testing strategies for improvement of) long-term adherence and durability of healthy nutrition habits.

Another layer of future research during real-world deployment could focus on the need for ongoing oversight to ensure safety, efficacy, and ethical compliance, with the goal of determining whether (or to which extent) it would be possible for the system to function with minimal supervision [36,40,41]. In this regard, studies that monitor specific use cases and potential risks, particularly the occurrence of hallucinations, and investigate mitigation strategies are warranted. Finally, real-world operational challenges related to scalability, integration into health care systems, and ensuring equitable access must be carefully evaluated to maximize the positive impact of AI-driven coaching on diverse populations [37,42].

Conclusions

In conclusion, this study presents a novel multi-agent LLM workflow that leverages behavioral science principles to enhance digital coaching for nutrition management among individuals with cardiometabolic conditions. Our approach, validated through expert assessments, real-world user studies, and large-scale simulation-based evaluations, provides strong evidence for the effectiveness of AI-driven, personalized coaching systems that go beyond generic advice to offer tailored, actionable, and empathetic guidance. Although further research is needed to validate these findings across larger and more diverse populations, our results pave the way for the development of scalable, behaviorally informed AI systems that can support meaningful and sustained health behavior change, addressing both current limitations and future possibilities in digital health interventions.

Acknowledgments

This study was sponsored by Verily Life Sciences. Verily Life Sciences was responsible for data collection. Authors were fully responsible for the data analysis and interpretation presented herein and the writing of this article. EY, TG, HW, BK, MR, ER, and JA had access to the raw data. The authors had access to the full dataset for the study and reviewed and approved the final manuscript for submission.

All authors report employment and equity ownership in Verily Life Sciences.

Data Availability

The data supporting this study are not accessible for sharing.

Authors' Contributions

Conceptualization: CC, EY, HGW, TG, YJ
Investigation: CC, EY, HW, TG, YM
Data curation: CC, EY, HW, TG, YM
Formal analysis: BK, ER, EY, HGW, JA, MR, TG
Supervision: YJ
Writing - original draft: EY
Writing - review and editing: BK, CC, ER, EY, HGW, JA, MR, TG, YM, YJ

XSL•FO
RenderX

Conflicts of Interest

All authors report employment and equity ownership in Verily Life Sciences.

Multimedia Appendix 1
Supplementary methods and results.
[DOCX File, 30 KB - formative_v9i1e75421_app1.docx]

References

1. Becker T, Majmundar MK, Harris KM. High and Rising Mortality Rates among Working Age Adults: National Academies of Sciences. [doi: 10.17226/25976]
2. Mozaffarian D, Aspry KE, Garfield K, et al. "Food is medicine" strategies for nutrition security and cardiometabolic health equity: JACC state-of-the-art review. J Am Coll Cardiol 2024 Feb 27;83(8):843-864. [doi: 10.1016/j.jacc.2023.12.023] [Medline: 38383100]
3. Bermingham KM, Linenberg I, Polidori L, et al. Effects of a personalized nutrition program on cardiometabolic health: a randomized controlled trial. Nat Med 2024 Jul;30(7):1888-1897. [doi: 10.1038/s41591-024-02951-6] [Medline: 38714898]
4. Majithia AR, Kusiak CM, Armento Lee A, et al. Glycemic outcomes in adults with type 2 diabetes participating in a continuous glucose monitor-driven virtual diabetes clinic: prospective trial. J Med Internet Res 2020 Aug 28;22(8):e21778. [doi: 10.2196/21778] [Medline: 32856597]
5. Azelton KR, Crowley AP, Vence N, et al. Digital health coaching for type 2 diabetes: randomized controlled trial of healthy at home. Front Digit Health 2021;3(764735):34901926. [doi: 10.3389/fdgth.2021.764735]
6. Lim WX, Fook-Chong S, Lim JW, Gan WH. The outcomes of app-based health coaching to improve dietary behavior among nurses in a tertiary hospital: pilot intervention study. JMIR Nurs 2022 Jul 15;5(1):e36811. [doi: 10.2196/36811] [Medline: 35838811]
7. Tu T, Palepu A, Schaekermann M, et al. Towards conversational diagnostic AI. ArXiv. Preprint posted online on Jan 11, 2024. [doi: 10.48550/arXiv.2401.05654]
8. Peng C, Yang X, Chen A, et al. A study of generative large language model for medical research and healthcare. NPJ Digit Med 2023;6(1):37973919. [doi: 10.1038/s41746-023-00958-w]
9. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature New Biol 2023 Aug;620(7972):172-180. [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
10. Huang Z, Berry MP, Chwyl C, Hsieh G, Wei J, Forman EM. Comparing large language model AI and human-generated coaching messages for behavioral weight loss. J Technol Behav Sci 2025. [doi: 10.1007/s41347-025-00491-5]
11. Ong QC, Ang CS, Chee DZY, et al. Advancing health coaching: a comparative study of large language model and health coaches. Artif Intell Med 2024 Nov;157:103004. [doi: 10.1016/j.artmed.2024.103004] [Medline: 39454500]
12. Szymanski A, Wimer BL, Anuyah O, Eicher-Miller HA, Metoyer RA. Integrating expertise in LLMs: crafting a customized nutrition assistant with refined template instructions. 2024 May 11 Presented at: CHI '24; Honolulu HI USA p. 1-22. [doi: 10.1145/3613904.3641924]
13. Sun H, Zhang K, Lan W, et al. An AI dietitian for type 2 diabetes mellitus management based on large language and image recognition models: preclinical concept validation study. J Med Internet Res 2023 Nov 9;25:e51300. [doi: 10.2196/51300] [Medline: 37943581]
14. Michie S, van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. Implement Sci 2011 Apr 23;6:42. [doi: 10.1186/1748-5908-6-42] [Medline: 21513547]
15. Michie S, Richardson M, Johnston M, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013 Aug;46(1):81-95. [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
16. BIT. East: Four simple ways to apply behavioural insights. 2014. URL: https://www.bi.team/publications/east-four-simple-ways-to-apply-behavioural-insights [accessed 2024-08-21]
17. Kopitar L, Bedrač L, Strath LJ, Bian J, Stiglic G. Improving personalized meal planning with large language models: identifying and decomposing compound ingredients. Nutrients 2025 Apr 29;17(9):1492. [doi: 10.3390/nu17091492] [Medline: 40362801]
18. Khamesian S, Arefeen A, Carpenter SM, NutriGen GH. Personalized meal plan generator leveraging large language models to enhance dietary and nutritional adherence. arXiv. Preprint posted online on Feb 28, 2025. [doi: 10.48550/arXiv.2502.20601]
19. Yang Z, Khatibi E, Nagesh N, et al. ChatDiet: empowering personalized nutrition-oriented food recommender chatbots through an LLM-augmented framework. Smart Health (2014) 2024 Jun;32:100465. [doi: 10.1016/j.smhl.2024.100465]
20. Lu L, Deng Y, Tian C, Yang S, Shah D. Purrfessor: a fine-tuned multimodal llava diet health chatbot. arXiv. Preprint posted online on Nov 22, 2024. [doi: 10.48550/arXiv.2411.14925]
21. Jörke M, Sapkota S, Warkenthien L, et al. Supporting physical activity behavior change with LLM-based conversational agents. ArXiv. Preprint posted online on May 9, 2024. [doi: 10.48550/arXiv.2405.06061]

XSL•FO
RenderX

22. Hegde N, Vardhan M, Nathani D, et al. Infusing behavior science into large language models for activity coaching. PLOS Digit Health 2024 Apr;3(4):e0000431. [doi: 10.1371/journal.pdig.0000431] [Medline: 38564502]
23. Meywirth S. Designing a large language model-based coaching intervention for lifestyle behavior change. In: Mandviwalla M, Söllner M, Tuunanen T, editors. Lecture Notes in Computer Science: Springer; 2024, Vol. 14621. [doi: 10.1007/978-3-031-61175-9_6]
24. Kumar H, Yoo S, Bernuy A, et al. Large language model agents for improving engagement with behavior change interventions: application to digital mindfulness. ArXiv. Preprint posted online on Jul 3, 2024. [doi: 10.48550/arXiv.2407.13067]
25. Dao D, Teo JYC, Wang W, Nguyen HD. LLM-powered multimodal AI conversations for diabetes prevention. 2024 Jun 10 Presented at: ICMR '24; Phuket Thailand p. 1-6. [doi: 10.1145/3643479.3662049]
26. Bak M, Chin J. The potential and limitations of large language models in identification of the states of motivations for facilitating health behavior change. J Am Med Inform Assoc 2024 Sep 1;31(9):2047-2053. [doi: 10.1093/jamia/ocae057] [Medline: 38527272]
27. Team G, Anil R, Borgeaud S, Alayrac JB, Yu J, Soricut R, et al. Gemini: a family of highly capable multimodal models. ArXiv. Preprint posted online on Dec 19, 2023. [doi: 10.48550/arXiv.2312.11805]
28. Pignatiello GA, Martin RJ, Hickman RL Jr. Decision fatigue: a conceptual analysis. J Health Psychol 2020 Jan;25(1):123-135. [doi: 10.1177/1359105318763510]
29. Moorhouse A. Decision fatigue: less is more when making choices with patients. Br J Gen Pract 2020 Aug;70(697):399. [doi: 10.3399/bjgp20X711989] [Medline: 32732206]
30. Chakraborty A. Present bias. ECTA 2021;89(4):1921-1961. [doi: 10.3982/ECTA16467]
31. Hershfield HE, Goldstein DG, Sharpe WF, et al. Increasing saving behavior through age-progressed renderings of the future self. J Mark Res 2011 Nov;48:S23-S37. [doi: 10.1509/jmkr.48.SPL.S23] [Medline: 24634544]
32. OpenAI. Hello gpt-4o. 2024. URL: https://openai.com/index/hello-gpt-4o [accessed 2024-08-27]
33. Wang P, Li L, Chen L, et al. Large language models are not fair evaluators. ArXiv. Preprint posted online on May 29, 2023. [doi: 10.48550/arXiv.2305.17926]
34. Wei H, He S, Xia T, Wong A, Lin J, Han M. Systematic evaluation of LLM-as-a-judge in LLM alignment tasks: explainable metrics and diverse prompt templates. ArXiv. Preprint posted online on Aug 23, 2024. [doi: 10.48550/arXiv.2408.13006]
35. Thakur AS, Choudhary K, Ramayapally VS, Vaidyanathan S, Hupkes D. Judging the judges: evaluating alignment and vulnerabilities in llms-as-judges. ArXiv. Preprint posted online on Jun 18, 2024. [doi: 10.48550/arXiv.2406.12624]
36. Zack T, Lehman E, Suzgun M, et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. Lancet Digit Health 2024 Jan;6(1):e12-e22. [doi: 10.1016/S2589-7500(23)00225-X] [Medline: 38123252]
37. Chen D, Liu Y, Guo Y, Zhang Y. The revolution of generative artificial intelligence in psychology: the interweaving of behavior, consciousness, and ethics. Acta Psychol (Amst) 2024 Nov;251:104593. [doi: 10.1016/j.actpsy.2024.104593] [Medline: 39522296]
38. Yang E, Amar J, Lee JH, Kumar B, Jia Y. The geometry of queries: query-based innovations in retrieval-augmented generation. ArXiv. Preprint posted online on Jul 25, 2024. [doi: 10.48550/arXiv.2407.18044]
39. Anisuzzaman DM, Malins JG, Friedman PA, Attia ZI. Fine-tuning large language models for specialized use cases. Mayo Clin Proc Digit Health 2025 Mar;3(1):100184. [doi: 10.1016/j.mcpdig.2024.11.005] [Medline: 40206998]
40. Pfohl SR, Cole-Lewis H, Sayres R, et al. A toolbox for surfacing health equity harms and biases in large language models. Nat Med 2024 Dec;30(12):3590-3600. [doi: 10.1038/s41591-024-03258-2]
41. Ayoub NF, Balakrishnan K, Ayoub MS, Barrett TF, David AP, Gray ST. Inherent bias in large language models: a random sampling analysis. Mayo Clin Proc Digit Health 2024 Jun;2(2):186-191. [doi: 10.1016/j.mcpdig.2024.03.003] [Medline: 40207170]
42. Freyer O, Wiest IC, Kather JN, Gilbert S. A future role for health applications of large language models depends on regulators enforcing safety standards. Lancet Digit Health 2024 Sep;6(9):e662-e672. [doi: 10.1016/S2589-7500(24)00124-9] [Medline: 39179311]

Abbreviations

AI: artificial intelligence
BCT: Behavioral Change Taxonomy
COM-B: capability-opportunity-motivation-behavior
EAST: Easy, Attractive, Social, and Timely
LLM: large language models

XSL•FO
RenderX

Edited by A Bucher; submitted 03.04.25; peer-reviewed by A Martins, S Clynes; revised version received 25.07.25; accepted 14.08.25; published 24.09.25.

Please cite as:
Yang E, Garcia T, Williams HG, Kumar B, Ramé M, Rivera E, Ma Y, Amar J, Catalani C, Jia Y
A Behavioral Science-Informed Agentic Workflow for Personalized Nutrition Coaching: Development and Validation Study
JMIR Form Res 2025;9:e75421
URL: https://formative.jmir.org/2025/1/e75421
doi:10.2196/75421

© Eric Yang, Tomas Garcia, Hannah G Williams, Bhawesh Kumar, Martin Ramé, Eileen Rivera, Yiran Ma, Jonathan Amar, Caricia Catalani, Yugang Jia. Originally published in JMIR Formative Research (https://formative.jmir.org), 24.9.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Assessing the User Experience of the EU Mobile App for Cancer Prevention: Mixed Methods Study

Furqan Ahmed[1], BDS, MPH, PhD; Silvana Romero Saletti[2], PhD; Erica D'Souza[3], PhD; Carolina Espina[3], PhD; David Ritchie[3], PhD; Ana Molina Barceló[4], PhD; Marina Pinto Carbó[4], PhD; Paula Romeo Cervera[4,5], MSc; Teresa Seum[5], MSc; Hermann Brenner[5], PhD; Stephan Van den Broucke[2], PhD; Maria Krini[6], MSc; Cristiana Fonseca[7], MSc; Patricia Pinto[7], MSc; Diana Krivic[8], MSc; Helena Ros Comesana[9], MSc; Wendy Yared[9], DrPH; Hajo Zeeb[1,10], PhD; Tilman Brand[1], PhD

[1]Department of Prevention and Evaluation, Leibniz Institute for Prevention Research and Epidemiology, Achterstraße 30, Bremen, Germany
[10]Health Sciences Bremen, University of Bremen, Bremen, Germany
[2]Psychological Sciences Research Institute, Université Catholique de Louvain, Louvain-la-Neuve, Belgium
[3]International Agency for Research on Cancer, World Health Organization, Lyon, France
[4]Cancer and Public Health Research Unit, The Foundation for the Promotion of Health and Biomedical Research of Valencia Region, Valencia, Spain
[5]Division of Clinical Epidemiology and Aging Research, German Cancer Research Center, Heidelberg, Germany
[6]Pagkyprios Syndesmos Karkinopathon Kai Filon, Paralimni, Cyprus
[7]Liga Portuguesa Contra o Cancro, Lisbon, Portugal
[8]Association of Slovenian Cancer Societies, Ljubljana, Slovenia
[9]Association of European Cancer Leagues, Brussels, Belgium

Corresponding Author:
Furqan Ahmed, BDS, MPH, PhD
Department of Prevention and Evaluation, Leibniz Institute for Prevention Research and Epidemiology, Achterstraße 30, Bremen, Germany

Abstract

Background: In 2022, nearly 20 million new cancer cases and 9.7 million deaths occurred globally. Europe, comprising under 10% of the world's population, accounted for over 22% of cases and 20% of deaths, reflecting an aging population, lifestyle risk factors, and extensive screening. With 40% of cancers preventable through modifiable risk factor interventions, effective prevention is essential. The European Code Against Cancer provides evidence-based guidelines that drive health initiatives across Europe. Supported by Europe's Beating Cancer Plan and the EU4Health program, the EU Mobile App for Cancer Prevention was developed to disseminate these recommendations. However, its effectiveness depends on usability across populations with varying digital and health literacy; this study evaluates the app's usability among diverse European populations.

Objective: This study aimed to identify enablers, barriers, and user requirements for the use and maintenance of the English version of the EU Mobile App for Cancer Prevention, focusing on how usability varied across individuals with different levels of digital health literacy and diverse sociodemographic backgrounds. In addition, user feedback on mock wireframes—visual representations of the app's interface and functionality—was gathered to evaluate usability and ease of use, providing insights for tailoring the app design to a broader population.

Methods: We conducted a mixed methods study in 7 European countries with 76 adults aged 19‑84 years recruited via purposive quota sampling. Participants completed quantitative usability testing using mock wireframes to perform 10 predefined tasks simulating core app functionalities (eg, profile setup and health goal tracking). We recorded task completion time, success rates, self-reported confidence, and perceived difficulty. Digital health literacy was assessed using the eHealth Literacy Scale (eHEALS) scale. Qualitative data were collected through focus group discussions guided by a semistructured interview guide, and transcripts were analyzed via thematic content analysis. Statistical analyses included descriptive statistics and 1-way ANOVA to explore group differences.

Results: Overall task completion rates ranged from 75% to 98%, with a median of 86%, indicating general usability. However, usability varied by age, education, and digital health literacy: younger participants and those with higher education and literacy levels reported greater confidence and lower difficulty, whereas older adults and lower-literacy users experienced more challenges. Qualitative analysis identified key themes affecting usability: the need for accessibility (multilingual support and simple language), user-centric design (age-friendly interfaces and intuitive navigation), ethical concerns (data privacy and security), and motivational features (gamification and personalized health goals).

XSL•FO
RenderX

Conclusion: The app is generally usable across diverse populations but requires streamlined interfaces and design adaptations to accommodate varying digital health literacy. Ensuring robust data privacy practices is essential for fostering user trust, and integrating motivational elements may enhance sustained engagement. Future work will involve piloting the finalized app to evaluate its real-world uptake and impact on cancer prevention behaviors.

(JMIR Form Res 2025;9:e73844) doi:10.2196/73844

KEYWORDS

cancer prevention; mobile application; usability testing; digital health literacy; European Code Against Cancer

Introduction

According to the most recent Global Cancer Observatory database (GLOBOCAN) 2022 estimates, nearly 20 million new cancer cases (including nonmelanoma skin cancers) and approximately 9.7 million cancer deaths were recorded worldwide in 2022 [1]. Although Europe comprises less than 10% of the global population, it accounts for over 22% of new cancer cases and about 20% of cancer deaths driven largely by its aging demographic, lifestyle risk factors, and extensive screening protocols [1]. Notably, lung cancer remains the leading cause of cancer morbidity and mortality, responsible for roughly 12.4% of new cases and 18.7% of cancer deaths globally. With demographic projections indicating that annual new cancer cases could surge to 35 million by 2050—a 77% increase from 2022—there is an urgent need for comprehensive prevention strategies, equitable health policies, and rigorous research to reduce future cancer burdens globally [1].

An estimated 40% of cancer cases are preventable through interventions targeting modifiable risk factors, reinforcing the importance of strong preventive measures [2]. The European Code Against Cancer (ECAC), an initiative launched by the European Commission over 3 decades ago and coordinated by the International Agency for Research on Cancer (IARC), provides a framework of evidence-based recommendations to reduce cancer risk [3]. The ECAC fourth edition, introduced in 2014, outlines 12 actionable guidelines that range from limiting exposure to recognized carcinogens to embracing health-promoting behaviors and participating in organized screening [3]. These recommendations guide health promotion initiatives across Europe, supported by civil society organizations, national and regional cancer leagues, and public health authorities [4].

In 2021, the European Commission launched Europe's Beating Cancer Plan, emphasizing the critical need to support cancer-related health literacy [5]. Funding from the EU4Health program supports the implementation of ECAC guidelines, particularly among vulnerable and underserved communities [5]. As part of these efforts, the Commission financed the EU Mobile App for Cancer Prevention to broaden ECAC's reach [6]. In recent years, mobile health apps have emerged as promising tools for cancer prevention, offering interactive platforms to deliver lifestyle interventions and educational content. For example, a systematic review reported that such mobile app interventions are being used to support cancer prevention and early detection in diverse global contexts, including low- and middle-income countries [7]. Within Europe, researchers have also tailored apps for specific populations; a notable case is an adolescent-focused app based on the ECAC recommendations, which demonstrated high user engagement through gamified modules and provided valuable insights for the development of the EU Mobile App [8].

Health literacy—the ability to access, understand, and apply health information to make informed decisions—has expanded in scope to include digital health literacy (DHL), which encompasses the skills needed to navigate, evaluate, and act on health information through digital platforms [9,10]. Together, these competencies are critical determinants of cancer prevention outcomes, shaping an individual's capacity to adopt preventive behaviors, interpret risk factors, and engage with screening programs. Digital tools, such as mobile apps, offer opportunities to democratize access to evidence-based health information across diverse populations. However, without purposeful design, these tools risk perpetuating or deepening existing inequities, as disparities in health literacy and DHL intersect with socioeconomic, age-related, and technological barriers. Within the EU, approximately 10% of the population struggles with inadequate health literacy [9], while far more face challenges in DHL, particularly older adults, rural communities, and marginalized groups. Limited health literacy is associated with delayed cancer screenings, poor adherence to preventive guidelines, and higher mortality rates [10]. Low DHL can limit access to digital resources, spread misinformation, and reduce trust in online health content [11,12]. Improving both health literacy and DHL is thus crucial for equitable cancer prevention.

In response to these challenges, the Boosting the Usability of the EU Mobile App for Cancer Prevention (BUMPER) project was initiated in November 2022 [6]. In this paper, we describe the pilot usability testing of the EU Mobile App for Cancer Prevention. The study focused on app usability among individuals with varying levels of DHL, seeking to enhance the app's appropriateness in supporting cancer prevention across heterogeneous communities. Participants were asked to complete 10 tasks using mock wireframes of the app, followed by focus group discussions (FGDs) that provided extensive insights into enablers, barriers, and user requirements.

Methods

Aim of the Study

This study aimed to identify enablers, barriers, and user requirements associated with the use and maintenance of the EU Mobile App for Cancer Prevention (English version). Specifically, the investigation explored how app usability varied across individuals with different levels of DHL and diverse sociodemographic backgrounds. In addition, user feedback on

XSL•FO
RenderX

mock wireframes (visual representations of the app's interface and functionality) was gathered to evaluate usability and ease of use.

Study Design

A mixed‐methods design was implemented, combining quantitative pilot testing of wireframes with qualitative FGDs to assess app usability [13]. This approach enabled the collection of both measurable usability metrics and in‐depth qualitative insights into user perspectives [14,15]. We adhered to the STAndards for Reporting of Evaluation studies in Health Informatics (STARE-HI) reporting guidelines for health informatics studies; the completed STARE-HI checklist is provided in Multimedia Appendix 1.

Study Setting

The study was conducted in 7 European countries: Cyprus, Finland, Germany, Hungary, Portugal, Slovenia, and Spain. Data collection included in-person sessions with local Cancer Leagues in Cyprus, Finland, Hungary, Portugal, Slovenia, and Spain. In Germany, the Leibniz Living Lab Bremen (part of the Leibniz Institute for Prevention Research and Epidemiology, BIPS) supported the pilot study [16,17]. In Cyprus, limited space ruled out group meetings, so data was collected through individual interviews. A total of 2 FGDs were conducted per country, with individual interviews and pilot testing in Cyprus. Study sites and countries were chosen based on the presence of BUMPER consortium partners; accordingly, in Germany, all data were collected at the Living Lab, rather than a cancer league, since only project partners conducted evaluations.

Participant Eligibility Criteria

Participants were included if they (1) were consenting adults (aged 18 y and older) residing in one of the participating countries, (2) owned a smartphone or tablet with internet access (required for engaging with the wireframes), and (3) could communicate in English or the local language.

Sampling and Recruitment

A total of 2 FGDs, each comprising 8‐12 participants, were planned per country. A quota sampling strategy was used to ensure representation across sociodemographic backgrounds, education levels, sex, and age groups [18]. Participants were stratified into three predefined categories: (1) age (18‐40, 41‐60, and ≥61 y), (2) gender, and (3) educational attainment (low: ≤12 y of schooling and high: university degree or higher). Recruitment was facilitated through collaborations with local organizations, including national cancer leagues, to ensure diverse participation across all 7 countries. Further details of the sampling framework are provided in Multimedia Appendix 1.

Data Collection

Moderators received training on both FGD facilitation and wireframe usability testing in collaboration with the app's technical developer [13,19]. Detailed training manuals provided questionnaires, wireframe tasks, and reporting templates. A total of 3 training sessions were conducted in October 2023: 2 covering FGDs (led by BIPS) and 1 led by the app developer on wireframe implementation and evaluation.

Data were collected from October 2023 to November 2023. Before usability testing, participants completed a sociodemographic questionnaire and the eHealth Literacy Scale (eHEALS) to measure DHL [20]. They then completed 10 predefined tasks using an English-language mock-up of the app (see Table 1). These tasks addressed essential functionalities such as user profile setup, health goal tracking, and accessing cancer prevention resources. Participants scanned QR codes and followed on-screen instructions for each task. Afterward, they rated confidence and perceived difficulty on Likert scales, while trained observers noted usability issues or challenges [19]. Usability metrics included task completion time and success or failure status, as recorded by moderators.

Wireframes were fully interactive app prototypes on participants' smartphones, mirroring the look, feel, and navigation of the final app, and thus, task completion times provide a valid indicator of actual interaction difficulty and cognitive load. Reporting time on task alongside confidence and difficulty ratings enriched our understanding of usability by quantifying how long users take to locate features under realistic conditions. The EU Mobile App for Cancer Prevention was organized into 4 main tabs, "Dashboard," "Goals," "Learning," and "Profile," each accessed via a bottom navigation bar. The "Dashboard" greeted users by name and visualized real-time progress on personalized prevention goals through simple charts and daily 1-tap logs (eg, tobacco use and physical activity). In "Goals," users browsed suggested targets, monitored active objectives, and reviewed completed milestones via clear icons and progress bars. The "Learning" tab featured a "Discover" feed of evidence-based topics (eg, smoking cessation and healthy weight) directly linked to the European Code Against Cancer's "12 Ways" guidelines. Finally, "Profile" allowed users to input age, sex, and health data to tailor content delivery and reminders.

XSL•FO
RenderX

Table . Task details and success criteria.

Task	Description	Success criteria (Expected outcome)	Completion status	Additional details
Task 1: Profile set up	Download the app and set up your personal profile to receive personalized recommendations.	Land on the profile setup screen.	Pass or fail	Time tracked; trainer notes available.
Task 2: Set initial goal	Set a goal for smoking reduction based on app-suggested goals.	Land on the goal-setting screen.	Pass or fail	Time tracked; trainer notes available.
Task 3: Goal progress	Find an active goal and a successfully completed goal.	Identify "Smoking reduction" or "Weight loss" as goals.	Pass or fail	Time tracked; trainer notes available.
Task 4: Daily tracking	Track your daily progress toward active goals.	Locate the recommendation "Avoid being outside."	Pass or fail	Time tracked; trainer notes available.
Task 5: UV Index	Check if conditions are ideal for outdoor activities.	Land on the UV Index screen and click "SAVE."	Pass or fail	Time tracked; trainer notes available.
Task 6: Goal management	Find tips on achieving the "Smoking reduction" goal.	Locate the "Practical tips" section.	Pass or fail	Time tracked; trainer notes available.
Task 7: Challenge a friend	Challenge a friend to the "Smoking reduction" goal.	Land on the challenge screen and click "SEND."	Pass or fail	Time tracked; trainer notes available.
Task 8: Reminders	Find upcoming reminders and create a new one for breast screening.	Land on the reminder screen and set a new reminder.	Pass or fail	Time tracked; trainer notes available.
Task 9: Learning	Find articles about smoking with the World Health Organization as the source.	Select "Smoking" and "World Health Organization" under the categories.	Pass or fail	Time tracked; trainer notes available.
Task 10: Awards	Find the criteria or steps needed to earn the "Smoke Warrior" award.	Identify "Be smoke-free for 12 months" as the criteria.	Pass or fail	Time tracked; trainer notes available.

Following usability testing, participants joined FGDs to explore their experiences, challenges, and perceptions of the app. A semistructured interview guide, based on the Leibniz ScienceCampus Digital Public Health framework, ensured methodological consistency throughout the sessions [17,21]. All discussions were audio-recorded to ensure comprehensive data capture [22]. The interview guide is provided in Multimedia Appendix 1.

Data Analysis: Quantitative Analysis

Descriptive statistics summarized categorical variables (country, biological sex, education level, disability status, and task completion) and continuous variables (age, eHEALS scores, and task completion times), along with Likert scale responses. The eHEALS comprises 8 items scored on a 5-point Likert scale (1=strongly disagree and 5=strongly agree), total scores from 8 to 40 (higher scores represent higher self-rated DHL). Key variables were categorized as follows: education level (low: no formal education, primary, lower, or upper secondary and high: trade, technical or vocational, bachelor's degree, master's degree, or doctorate); median-split DHL scores (low vs high DHL), using an eHEALS median of 30.5 as the cutoff; and 4 age groups (young adults: 18 - 40 y, middle-aged adults: 41 - 60 y, and older adults: ≥61 y). We dichotomized eHEALS scores at the sample median (30.5) because no validated threshold exists for "high" versus "low" digital health literacy, yielding 2 equal - sized groups for robust 1 - way ANOVA comparisons. Furthermore, 1-way ANOVAs assessed the association of these factors on continuous outcomes (eg, task completion time, confidence, perceived difficulty, and likelihood of future use) [20]. A series of 1-way ANOVAs using ordinary least squares determined statistical significance at $P<.05$. Group means were computed for interpretation. To validate the ANOVA findings, assumptions were checked: the Shapiro-Wilk and Q-Q plots for normality, the Levene test for homogeneity of variance, and the Cook distance (<1) for outlier detection [23]. Statistical analyses were performed using Python (version 3.11.8; Python Software Foundation) with *Pandas*, *NumPy*, *SciPy*, *Statsmodels*, and *Seaborn* packages [24,25].

Data Analysis: Qualitative Analysis

All FGDs and interviews were audio-recorded and later transcribed verbatim by professional transcription services. Artificial intelligence (AI)–assisted software supported translations into English, with local teams verifying translation accuracy. Thematic content analysis followed Bengtsson's framework [26], encompassing decontextualization, recontextualization, categorization, and compilation. Transcripts were imported into ATLAS.ti software (version 9.22.0–2025-08-26; ATLAS.ti Scientific Software Development GmbH), which used AI-driven natural language processing to expedite coding [26-28]. The final coding scheme emerged from a combination of open coding and research question–driven coding, with 2 reviewers reconciling any discrepancies to ensure

XSL•FO
RenderX

reliability. Multiple researchers subsequently evaluated the findings to increase validity and capture nuanced interpretations [26-28].

Ethical Considerations

The BUMPER pilot study obtained ethical clearance from the Ethics Commission at the University of Bremen (Application 2023 - 10). All participants provided informed consent before enrolling, with assurance of confidentiality and anonymity in published findings. No participants are identifiable in any images or results presented, and participants received no monetary compensation beyond reimbursement of travel expenses (only by some countries).

Results

Demographic Characteristics

A total of 76 participants from 7 European countries (Germany: n=13, 17%; Finland, Hungary, and Spain: n=12, 16% each; Portugal and Slovenia: n=10, 13% each; and Cyprus: n=7, 9%) were enrolled, with a near-balanced gender distribution (women: n=40, 53% and men: n=36, 47%). Age groups spanned 18 - 84 years (mean 46, SD 17.3 y), though representation skewed toward younger adults (18 - 40 y: n=30, 39%) compared to middle-aged (41 - 60 y: n=25, 33%) and older adults (≥61 y: n=21, 28%). Education levels reflected stratification into “low” (≤12 years: n=33, 43% [primary: n=1, 1%; lower secondary: n=16, 21%; and upper secondary: n=15, 20%]) and “high” (postsecondary: n=43, 56% [vocational: n=7, 9%; bachelor’s: n=18, 24%; master’s: n=17, 22%; and doctorate: n=1, 1%]), with minimal representation of no formal education (n=1, 1%). Disabilities were reported by n=15 (20%) participants, primarily chronic illnesses (n=8, 11%), while n=61 (80%) reported no disability. Despite logistical constraints necessitating convenience sampling, the cohort broadly reflected the intended diversity in age, gender, and education. Table 2 shows these demographic characteristics.

XSL•FO
RenderX

Table . Participant demographics by country, gender, education level, and disability status.

Characteristic	Frequency, n (%)[a]
Age group category	
Young adults (18 - 40 y)	30 (39)
Middle-aged adults (41 - 60 y)	25 (33)
Older adults (≥61 y)	21 (28)
Country	
Germany	13 (17)
Finland	12 (16)
Hungary	12 (16)
Spain	12 (16)
Portugal	10 (13)
Slovenia	10 (13)
Cyprus	7 (9)
Gender	
Women	40 (53)
Men	36 (47)
Other	0 (0)
Education level	
Doctorate degree	1 (1)
Master's degree	17 (22)
Bachelor's degree	18 (24)
Trade, technical, or vocational training	7 (9)
Upper secondary education	15 (20)
Lower secondary education	16 (21)
Primary education	1 (1)
No formal education	1 (1)
Disability status	
No disability	61 (80)
Chronic illness or autoimmune disorder (eg, diabetes or multiple sclerosis)	8 (11)
Visual impairment (eg, low vision)	3 (4)
Hearing impairment (eg, deafness or hard of hearing)	1 (1)
Physical disability (eg, mobility impairment or amputation)	1 (1)
Physical disability, mental health conditions, chronic illness, or autoimmune disease	1 (1)
Mental health condition (eg, anxiety, depression, or bipolar disorder)	1 (1)

[a]All percentages are based on a total of 76 observations per category. Rounding may cause totals not to sum exactly to 100%.

Digital Health Literacy

Participants' DHL was assessed using the eHEALS [20]. As depicted in Figure 1, DHL scores varied across sociodemographic groups. The overall DHL score averaged 29.4 (SD 7.0), with a median split at 30.5 categorizing participants into low and high DHL groups. Overall, the DHL scores show differences by age, sex, education, and country (see Figure 1). The analysis shows that younger adults (n=30) have a median DHL score of 31.5 (range 17-40; mean 31.0, SD 6.1), whereas older adults (n=21) have a lower median of 29.0 (range 8-40; mean 26.1, SD 8.1). Women exhibit a slightly higher median of 31.0 (n=40; range 11-39; mean 30.2, SD 5.7) compared to 29.5 median for men (n=36; range 8-40; mean 28.4, SD 8.1). Participants with higher education (n=43) demonstrate a median of 32.0 (range 21-40; mean 30.6, SD 6.5)

XSL•FO
RenderX

relative to 30.0 for those with lower education (n=33; range 11-40; mean 27.8, SD 7.3). Portugal (n=10) among countries had the highest median DHL score at 35.5 (range 29-40; mean 35.3, SD 4.5), whereas Spain (n=12; median 27.0; range 15-35; mean 27.7, SD 6.9) and Cyprus (n=7; median 29.0; range 11-33; mean 25.7, SD 7.9) ranked lower.

Figure 1. This figure presents the distribution of digital health literacy (DHL) scores across different sociodemographic categories, represented in 4 boxplots: (A) illustrates the DHL scores for 4 distinct age categories: young adults (18 - 40 years), middle-aged adults (41 - 60 years), and older adults (≥61 years); (B) shows the scores by sex, comparing women and men; (C) displays DHL scores based on education level education level (low: no formal education, primary, lower, or upper secondary; high: trade, technical or vocational, bachelor's degree, master's degree, or doctorate); and (D) compares the DHL scores by country, reflecting the various countries represented in the dataset.

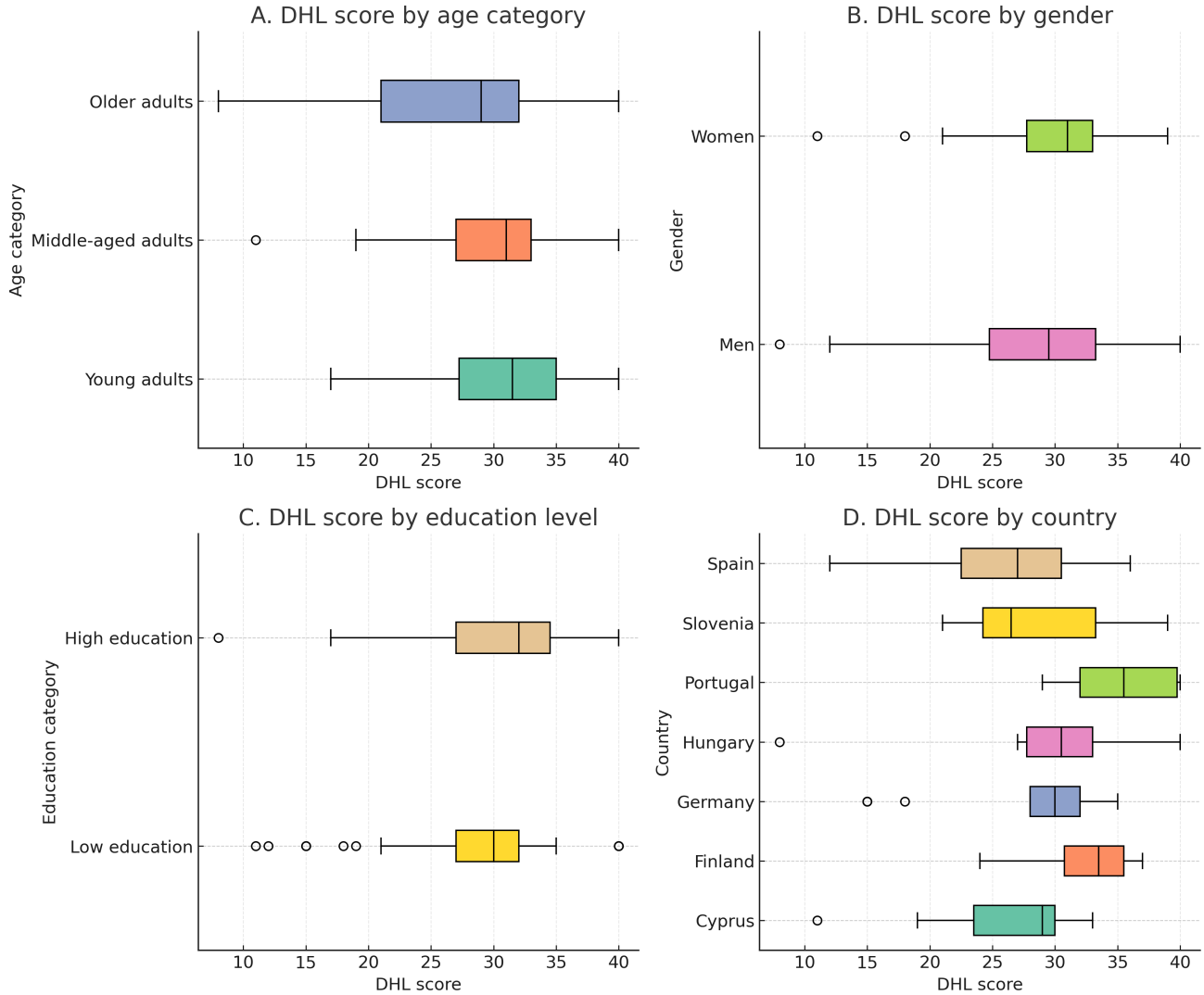

Time Spent on Tasks, Completion Status, Confidence, Difficulty Ratings, and Likelihood of Use

Mean task completion times ranged from 54 seconds (Task 5) to 200 seconds (Task 1). Task 7 took an average of 108 seconds, with a maximum of 420 seconds, suggesting variability in perceived complexity. Task pass rates ranged from 75% (Task 7) to 98% (Tasks 1, 2, and 5), with a median completion rate of 86%. Mean confidence ratings (on a 5-point Likert scale) were generally high (3.6 - 4.4), while mean difficulty ratings spanned 1.8 - 3.2. Task 7 had both the lowest pass rate and the highest difficulty. Participants rated their likelihood of future app use at a moderate level (mean 3.5, SD 1.0). Figure 2 shows task completion percentages, time taken, confidence, and difficulty ratings for each task, along with the overall averages.

XSL•FO
RenderX

Figure 2. This figure presents task completion percentages, completion times, confidence, and difficulty ratings for each task, together with the overall averages. Each horizontal bar shows the proportion of participants who passed (green) or failed (red). Red dots on the bars represent the mean time to complete each task, with black whiskers indicating the SD of completion time. To the right, confidence and difficulty ratings are displayed with their respective SDs. Confidence ratings are shown in green bars (higher values=greater confidence), while difficulty ratings are shown in red bars (higher values=greater perceived difficulty). Dashed vertical lines mark the overall averages for confidence and difficulty, with the corresponding average values annotated. The average pass rate across all tasks is also indicated by a dashed line in the completion panel. Avg conf: average confidence; Avg diff: average difficulty.

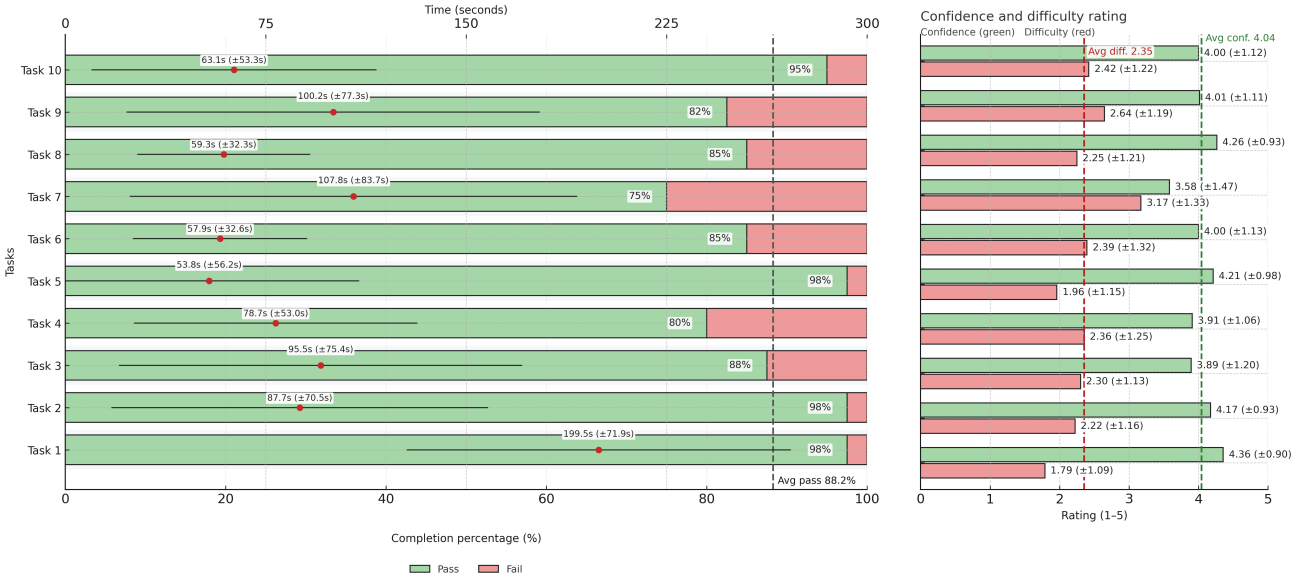

ANOVA

A 1-way ANOVA examined the relationships between age, education level, or DHL and task-related outcomes (confidence and perceived difficulty). Assumptions of normality (Shapiro-Wilk test and Q-Q plots), homogeneity of variances (the Levene test), and absence of influential outliers (the Cook distance <1) were confirmed for all analyses unless otherwise specified.

Age Category

Average Time

Age category was significantly associated with average time spent per task ($F_{2,72}$=3.6; P=.04). Young adults spent significantly less time (mean 75.4, SD 34.6) compared to middle-aged adults (mean 97.0, SD 40.7) and older adults (mean 111.2, SD 21.7).

Confidence in Tasks

Age category was significantly associated with confidence ($F_{2,72}$=10.2; P<.001). Young adults (mean 4.4, SD 0.5) reported significantly higher confidence compared to middle-aged adults (mean 4.0, SD 0.6) and older adults (mean 3.6, SD 0.7).

Perceived Difficulty

Age category was significantly associated with perceived difficulty ($F_{2,72}$=3.7; P=.03). Young adults reported significantly less difficulty (mean 2.1, SD 0.8) compared to middle-aged adults (mean 2.4, SD 0.7) and older adults (mean 2.7, SD 0.6).

Education Level

Education was categorized as low (≤12 y of schooling: primary, lower, or upper secondary) or high (trade, technical, or vocational training or postsecondary degree). Participants with high education reported greater confidence ratings (mean 4.2, SD 0.6) compared to those with low education (mean 3.8, SD 0.7; $F_{1,74}$=8.2; P=.01).

Digital Health Literacy (DHL)

Confidence in Tasks

DHL was significantly associated with confidence ($F_{1,74}$=5.7; P=.02). Participants with high DHL (mean 4.2, SD 0.6) reported higher confidence than those with low DHL (mean 3.9, SD 0.7). Although the Shapiro-Wilk test suggested a minor normality violation (P=.04), Q-Q plots showed only slight deviation, the Levene test confirmed homogeneity of variance, and no outliers exceeded the Cook distance of 1.

Perceived Difficulty

DHL was significantly associated with perceived difficulty ($F_{1,74}$=7.9, P=.01). Participants with high DHL reported less difficulty (mean 2.1, SD 0.8) compared to those with low DHL (mean 2.6, SD 0.7).

Qualitative Data Analysis

FGDs and interviews revealed 5 major themes relevant to the usability and acceptance of the EU Mobile App for Cancer Prevention: accessibility, core features, ethical concerns, motivating and engaging features, and user-friendly design.

Accessibility

Participants highlighted the importance of making the app broadly accessible across different cultural, linguistic, and technical contexts. Many emphasized that the app should be available in multiple languages to serve diverse user groups. For example, a participant noted, “It depends on the country... In Cyprus, I think it should have the choice of Greek, English, and Russian because we have many” (Participant 8, Cyprus). Along with multilanguage support, simplicity in language was

XSL•FO
RenderX

stressed to accommodate varying literacy levels. As an interviewee stated, “Simple. The simpler the better (Language)” (Participant 6, Cyprus), underscoring the need for plain, nontechnical language. Furthermore, educational content should aim for a 6th-8th grade reading level using plain language guidelines to improve comprehension across literacy levels. In addition, participants were concerned about connectivity issues; offline functionality was deemed critical so that users could continue to access key features even without an internet connection. A participant remarked, “Logging should work even without the Internet” (Participant 1, Slovenia). Together, these insights indicate that ensuring both linguistic inclusivity and offline access are essential for equitable use of the app.

Core Features

Participants stressed the necessity of incorporating comprehensive and reliable information about various cancers, including causes, symptoms, and prevention strategies. A participant requested, “Tell me in detail what cancer is in simple words and various symptoms” (Participant 9, Cyprus), reflecting the demand for clear, evidence-based health education. In addition, practical features that aid early detection were highlighted. For instance, an interviewee suggested, “Include screening advice... when you’re consulting screening programs that may be in your community” (Participant 1, Spain), emphasizing the importance of guiding users through regular screenings and self-examinations. Furthermore, resources that support healthier lifestyles, such as tools for tracking physical activity, diet, and other health metrics, were identified as key motivators. A participant observed, “Seeing progress happening. that would motivate me to go back there the next day” (Participant 2, Portugal), which illustrates how integrated tracking features can reinforce healthy behaviors.

Ethical Concerns

Data protection and privacy emerged as primary ethical issues. Participants highlighted the need for robust security measures to protect personal information and ensure compliance with privacy regulations, such as the General Data Protection Regulation. As a participant explained, “You have to keep all information you get from everyone well-guarded to comply with the European regulation” (Participant 4, Cyprus). This emphasis on data security reflects broader concerns about maintaining user trust. Some participants also expressed that without stringent privacy measures, users might be reluctant to engage with the app, thus impeding its adoption and long-term success.

Motivating and Engaging Features

Beyond the provision of core health information, participants emphasized that the app should include features that actively motivate and engage users. A visually appealing, user-friendly design with gamified elements was frequently recommended. A participant noted, “The instructions should be as simple as possible and there should be steps to follow them easily……Very likely with photos, bright colors which are very likely to attract more (users)” (Participant 7, Cyprus), highlighting the role of visual appeal and clear guidance in sustaining user interest. In addition, interactive features that encourage social competition and goal setting were suggested. For instance, a participant commented, “We compete (with friends/family) almost a bit on which one will do several kilometers a month” (Participant 2, Slovenia), demonstrating that competitive and social elements can enhance engagement and adherence to healthy behaviors.

User-Friendly Design

Designing the app to be user-friendly across different age groups was seen as essential. For older users, a simplified interface with larger fonts and clear instructions is crucial, while younger users may favor more interactive and gamified elements. A participant stated, “Making it a bit more playful can be beneficial” (Participant 2, Portugal), suggesting that a touch of playfulness could improve usability for younger demographics. In addition, the inclusion of onboarding tutorials was noted as a key feature for easing new users into the app. As another participant remarked, “There are also applications that... with the first use they make you like a small tutorial” (Participant 1, Spain). This reflects the expectation that an intuitive interface with clear navigation, large buttons, and customizable settings will reduce barriers to use and enhance overall accessibility.

Discussion

Principal Findings

The study aimed to understand the usability and barriers experienced by different user groups while interacting with a mock version of the EU Mobile App for Cancer Prevention as part of an iterative app development process. Quantitative findings indicated generally high task completion rates and confidence levels; however, significant variations were associated with age, education, and DHL. Qualitative insights further highlighted user preferences and barriers, emphasizing the critical importance of accessibility, user-centered design, ethical considerations, and motivational features. These findings highlight the need for tailored solutions to address the varying needs of diverse user groups.

Associations of Age, Education, and Digital Health Literacy

User confidence and perceived task difficulty varied by age. Younger adults exhibited higher confidence and reported lower difficulty levels, consistent with literature suggesting greater digital competence among younger individuals [29,30]. In contrast, older adults faced higher difficulty, underscoring the need for age-friendly design elements, such as simplified interfaces, larger fonts, and clear instructions [31]. Importantly, because eHEALS measures self-perceived ability to find, evaluate, and use online health information, lower DHL scores in older participants may reflect reduced confidence rather than objective skill deficits. These findings imply that a one-size-fits-all approach to app design may not be suitable, particularly for older populations who require additional support and resources [32].

Education level and DHL were also linked to user confidence and task difficulty; participants with higher education and DHL scores reported greater confidence and found tasks less challenging. This aligns with studies linking higher digital

XSL•FO
RenderX

literacy to better navigation of health tools [30,33]. However, as health literacy data were not collected, we emphasize that educational attainment and DHL are interrelated factors contributing to digital divides [34]. Addressing inequities in DHL (eg, through guided tutorials and simplified navigation) is critical to ensuring equitable access to digital health tools, as these inequities often reflect systemic barriers such as unequal access to technology [35].

Enhancing Accessibility and Inclusivity

Multilingual support is important to ensure accessibility for non-English speakers and to accommodate the region's linguistic diversity. Offering the app in multiple languages with high-quality, culturally sensitive translations can significantly enhance comprehension and engagement, as demonstrated in studies of digital health interventions in multicultural settings [29,36]. This approach acknowledges the diverse linguistic backgrounds of European users, making the app more inclusive. Simplified language and clear, jargon-free explanations are essential for users with varying literacy levels, helping to reduce potential barriers to understanding and engagement.

In addition, implementing offline functionality allows users in areas with limited internet access to benefit from the intervention without requiring constant connectivity. This feature is particularly relevant for rural or underserved regions, where reliable internet may not be available [35]. Improving accessibility also involves incorporating features such as adjustable text sizes, voice commands, and compatibility with assistive technologies to accommodate users with disabilities. Incorporating features such as color contrast options, customizable display settings, and alternative text for images can further enhance accessibility for individuals with visual impairments [34]. The goal is to create a flexible and adaptable user interface that caters to the diverse needs of all users, regardless of their physical or cognitive abilities. Addressing these aspects can help reduce the digital divide and promote equitable access to digital health interventions.

Building Trust Through Ethical Considerations

Robust data protection measures and transparent communication about privacy policies can enhance user trust and willingness to engage with digital health tools. Compliance with regulations like the General Data Protection Regulation (GDPR) is imperative to ensure data security [37]. However, a key dilemma lies in balancing transparency (eg, detailed privacy policies) with user-friendly onboarding processes [38-41]. For instance, lengthy consent forms may deter users, while overly simplified policies risk concealing critical details. Transparency regarding data collection, storage, and usage practices is fundamental to user adoption and engagement, as shown in studies of health app adoption [38-41]. Users need to feel confident that their personal data is being handled responsibly, particularly in the context of sensitive health information. Building trust also involves providing users with control over their data, including options to manage consent and review data-sharing practices. Ensuring that privacy policies are communicated in a clear and understandable manner can help clarify data practices and address user concerns [34].

Boosting Engagement With Motivational Features

Incorporating motivational and engaging features, such as gamification, social support options, and personalized goal-setting, can promote sustained user engagement and adherence to health-promoting behaviors. Gamification elements, such as rewards, badges, and progress tracking, can make the experience more enjoyable and motivate users to continue using the app, as evidenced in digital health studies [42]. Users particularly appreciated features that actively motivated them, including progress tracking and social interaction elements. Social support features leverage influence and support networks to enhance motivation. Providing options for users to connect with peers, share progress, and receive encouragement can foster a supportive environment that encourages sustained engagement. Simplifying social interaction features and providing guided tutorials or clearer prompts assist users in navigating these aspects effectively, especially those who may not be familiar with social media or online communities [34].

Providing Educational Content and User-Centric Design

Offering evidence-based, easily understandable information on health-related topics, such as prevention, early detection, and healthy lifestyle practices, is crucial. Educational content must be accurate and sourced from reputable organizations to establish credibility and effectively inform users. In addition to providing general health information, incorporating interactive educational modules (eg, quizzes and videos) can enhance learning and retention, as shown in studies of DHL [42]. Participants specifically requested tutorials and simplified explanations, underscoring the need for iterative, user-centered design. Adopting a culturally sensitive design by involving diverse users throughout the development process allows for feedback and continuous improvements toward enhancing accessibility and relevance. For developers, being mindful of cultural differences in health beliefs and practices enhances the relevance and acceptance of digital health tools across different populations, fostering greater engagement and effectiveness. In addition, iterative pilot testing and refinement based on user feedback are critical to creating a responsive and user-friendly app. Regular updates that incorporate user suggestions can help maintain engagement and ensure that the app continues to meet users' needs over time. The inclusion of culturally relevant imagery, language, and examples can improve the app's appeal and effectiveness for diverse user groups.

Strengths and Limitations of the Study

The sample size was of moderate size for a study using predominantly qualitative methods (n=76) and is likely not representative of the broader European population. However, including participants from 7 European countries and using a mixed methods pilot approach provided valuable preliminary insights. The sampling approach may have introduced selection bias, potentially affecting the generalizability of the findings. The use of mock wireframes instead of a fully functional app could not fully capture user interactions and potential usability issues present in the final app. To enhance methodological rigor, we used the validated eHEALS instrument to assess DHL. We

XSL•FO
RenderX

acknowledge that the eHEALS captures self-rated confidence in locating, appraising, and applying online health information rather than objectively evaluating digital health skills. Although it is the most widely validated instrument for perceived electronic health literacy, its reliance on subjective self-assessments may not fully correspond to actual proficiency in using digital health resources. In addition, the cross-sectional design limits the ability to assess long-term engagement and behavior change resulting from app use. Future studies should consider larger, more diverse samples to improve generalizability. Using a longitudinal design could also provide insights into how user engagement and behavior change evolve over time. Furthermore, incorporating a fully functional app in future usability testing will allow more accurate assessment of real-world user interactions and barriers.

Future Directions

This initial usability study provides a foundation for further development and refinement of the app. Following this pilot, we delivered a comprehensive summary of identified enablers and barriers to the app development team and conducted multiple collaborative iterative meetings. As a result, the next prototype incorporated larger and more intuitive navigation elements, additional language support, an interactive onboarding tutorial, and a settings screen, ensuring that user feedback directly shaped the app's evolution before its live testing in the next phase. The next phase will involve piloting the developed app by distributing it to participants for installation on their personal devices [43]. This real-world testing will allow users to interact with the app and provide feedback on functionality, engagement, and impact on health behaviors before making it available to EU citizens [43]. Future research should focus on strategies to enhance DHL among users, further improving engagement and effectiveness. This could include incorporating educational modules that improve users' digital competence and offering community support programs to assist users in effectively using digital health tools [34].

Integration with health care systems presents another promising direction. Collaborating with health care providers to align digital health tools with existing health services could enhance their usability and impact [44]. Including features that allow health care professionals to provide input or feedback may increase the credibility and effectiveness of digital interventions. Features such as secure messaging, cancer screening appointment scheduling, and health data sharing could foster greater integration with traditional health care services. Engaging with policymakers to promote supportive regulations and standardization efforts can facilitate broader adoption and integration of digital health tools. Policymaker engagement is also critical to addressing systemic barriers to digital health adoption, such as inadequate infrastructure and lack of digital literacy programs [34].

Conclusion

The pilot usability testing of the EU Mobile App for Cancer Prevention indicates that while the app is generally user-friendly, optimizing its effectiveness across diverse populations requires careful consideration of age, educational attainment, and DHL. Incorporating user feedback—particularly on accessibility features, ethical considerations, and engaging content—is crucial for enhancing adoption and promoting equitable access to cancer prevention resources. By focusing on social factors and equity, developers can create a cancer prevention digital health intervention that is more effective, inclusive, and responsive to the diverse needs of the population. Addressing the barriers identified in this study can significantly strengthen the app's role in supporting Europe's Beating Cancer Plan and reducing the cancer burden across the European Union.

Acknowledgments

We acknowledge the contributions of all study participants and our partners in the Cancer Leagues across Europe for their assistance with recruitment and data collection. We thank the BUMPER (Boosting the Usability of the EU Mobile App for Cancer Prevention) consortium for their collaboration and feedback throughout the project and the Leibniz Living Lab Bremen for their invaluable support in coordinating on-site data.

We acknowledge that GitHub Copilot assisted in developing Python scripts for quantitative data analysis and the artificial intelligence (AI) features of ATLAS.ti facilitated thematic coding in our qualitative analysis. These tools supplemented, but did not replace, the authors' expertise and interpretation. Finally, we acknowledge the funding provided by the EU4Health program and thank the European Commission for its support in advancing cancer prevention research.

The authors are identified as personnel of the International Agency for Research on Cancer/World Health Organization. The authors alone are responsible for the views expressed in this article, and they do not necessarily represent the decisions, policy, or views of the International Agency for Research on Cancer/World Health Organization.

Data Availability

The deidentified quantitative datasets and analysis code generated during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: FA (lead), SRS, ED, CE, DR, AMB, MPC, PRC, TS, HB, SVdB, MK, CF, PP, DK, HRC, WY, HZ, TB (equal)

XSL•FO
RenderX

Methodology: FA (lead), SRS, ED, HRC, WY, HZ, TB (equal)
Software: FA (lead), SRS, ED, TB (supporting)
Validation: FA (lead), SRS, ED, CE, DR, AMB, MPC, PRC, TS, HB, SVdB, MK, CF, PP, DK, HRC, WY, HZ, TB (equal)
Formal analysis: FA (lead), SRS, ED, HRC, TB (equal).
Investigation: FA (lead), SRS, ED, AMB, MPC, PRC, TS, HB, SVdB, MK, CF, PP, DK, HZ, TB (equal)
Resources: FA; CE, DR, AMB, MPC, PRC, TS, HB, SVdB, MK, CF, PP, DK, HRC, WY, HZ, TB
Data curation: FA (lead), AMB, MPC, PRC, TS, HB, SVdB, MK, CF, PP, DK, HZ, TB (equal).
Writing – original draft: FA (lead), SRS, ED, HRC, WY, HZ, TB (supporting)
Writing – review & editing: FA (lead), SRS, ED, CE, DR, AMB, MPC, PRC, TS, HB, SVdB, MK, CF, PP, DK, HRC, WY, HZ, TB (equal)
Visualization: FA (lead), HRC, WY, HZ, TB (equal)
Supervision: FA, DR, HB, HRC, WY, HZ (lead), TB (equal)
Project administration: FA, HRC (equal), WY (lead), HZ, TB (equal)
Funding acquisition: DR (lead), HB (equal), WY (equal), HZ (equal), TB (equal)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Recruitment strategy and interview guide.
[DOCX File, 22 KB - formative_v9i1e73844_app1.docx]

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018 Nov;68(6):394-424. [doi: 10.3322/caac.21492] [Medline: 30207593]
2. Wild E, CP W, Stewart BW. World cancer report: cancer research for cancer prevention. : International Agency for Research on Cancer; 2020 URL: https://publications.iarc.fr/586 [accessed 2025-09-10]
3. Schüz J, Espina C, Villain P, et al. European Code against Cancer 4th Edition: 12 ways to reduce your cancer risk. Cancer Epidemiol 2015 Dec;39 Suppl 1:S1-10. [doi: 10.1016/j.canep.2015.05.009] [Medline: 26164654]
4. Leon ME, Peruga A, McNeill A, et al. European Code against Cancer, 4th Edition: Tobacco and cancer. Cancer Epidemiol 2015 Dec;39 Suppl 1:S20-S33. [doi: 10.1016/j.canep.2015.06.001] [Medline: 26272517]
5. Europe's beating cancer plan. European Commission. 2021. URL: https://op.europa.eu/en/publication-detail/-/publication/f6d84a1a-6aac-11eb-aeb5-01aa75ed71a1 [accessed 2025-09-10]
6. BUMPER – Boosting the Use of the EU Mobile App for Cancer Prevention. European Commission. URL: https://health.ec.europa.eu/non-communicable-diseases/cancer/europes-beating-cancer-plan-eu4health-financed-projects/projects/bumper_en [accessed 2025-09-10]
7. Donkor A, Ayitey JA, Adotey PN, et al. Mobile-based application interventions to enhance cancer control and care in low- and middle-income countries: a systematic review. Int J Public Health 2023;68(1):1606413. [doi: 10.3389/ijph.2023.1606413] [Medline: 38125709]
8. Mallafré-Larrosa M, Papi G, Trilla A, Ritchie D. Development and promotion of an mHealth app for adolescents based on the European code against cancer: retrospective cohort study. JMIR Cancer 2023 Nov 28;9(1):e48040. [doi: 10.2196/48040] [Medline: 38015612]
9. Sørensen K, Van den Broucke S, Pelikan JM, et al. Measuring health literacy in populations: illuminating the design and development process of the European Health Literacy Survey Questionnaire (HLS-EU-Q). BMC Public Health 2013 Oct 10;13:948. [doi: 10.1186/1471-2458-13-948] [Medline: 24112855]
10. Berkman ND, Sheridan SL, Donahue KE, Halpern DJ, Crotty K. Low health literacy and health outcomes: an updated systematic review. Ann Intern Med 2011 Jul 19;155(2):97-107. [doi: 10.7326/0003-4819-155-2-201107190-00005] [Medline: 21768583]
11. Sturman N, Ostini R, Choy M, Cyphers G. Digital health access, uptake, literacy and trust in people with experience of homelessness. Aust J Gen Pract 2024 Apr;53(4):220-226. [doi: 10.31128/AJGP-2-23-6716] [Medline: 38575543]
12. Bin Naeem S, Kamel Boulos MN. COVID-19 misinformation online and health literacy: a brief overview. Int J Environ Res Public Health 2021 Jul 30;18(15):8091. [doi: 10.3390/ijerph18158091] [Medline: 34360384]
13. Fetters MD, Curry LA, Creswell JW. Achieving integration in mixed methods designs-principles and practices. Health Serv Res 2013 Dec;48(6 Pt 2):2134-2156. [doi: 10.1111/1475-6773.12117] [Medline: 24279835]
14. Creswell JW, Plano Clark VL. Designing and conducting mixed methods research. Aust N Z J Public Health 2007. [doi: 10.1111/j.1753-6405.2007.00096.x]

XSL•FO
RenderX

15. Kitzinger J. Qualitative research. Introducing focus groups. BMJ 1995 Jul 29;311(7000):299-302. [doi: 10.1136/bmj.311.7000.299] [Medline: 7633241]
16. Sutton C. Developing and evaluating complex interventions. Matern Child Nutr 2014 Apr;10(2):163-165. [doi: 10.1111/mcn.12127] [Medline: 24661462]
17. Developing and assessing digital public health interventions: a comprehensive framework [Report in German]. : Leibniz wissenschaftscampus - DIPH; 2023 URL: https://www.lsc-digital-public-health.de/forschung/framework.html [accessed 2025-03-12]
18. Patton MQ. Qualitative Research & Evaluation Methods, 4th edition: SAGE Publications; 2015.
19. Brooke J. SUS: a “quick and dirty” usability scale. In: Jordan B, McClelland IL, Weerdmeester B, PW T, editors. Usability Evaluation in Industry: Taylor & Francis; 1996:189-194.
20. Norman CD, Skinner HA. eHEALS: The eHealth Literacy Scale. J Med Internet Res 2006 Nov 14;8(4):e27. [doi: 10.2196/jmir.8.4.e27] [Medline: 17213046]
21. Craig P, Dieppe P, Macintyre S, et al. Developing and evaluating complex interventions: the new Medical Research Council guidance. BMJ 2008 Sep 29;337:a1655. [doi: 10.1136/bmj.a1655] [Medline: 18824488]
22. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
23. Sullivan GM, Feinn R. Using effect size-or why the P value is not enough. J Grad Med Educ 2012 Sep;4(3):279-282. [doi: 10.4300/JGME-D-12-00156.1] [Medline: 23997866]
24. Ghasemi A, Zahediasl S. Normality tests for statistical analysis: a guide for non-statisticians. Int J Endocrinol Metab 2012;10(2):486-489. [doi: 10.5812/ijem.3505] [Medline: 23843808]
25. McKinney W. Data structures for statistical computing in Python. Presented at: Python in Science Conference; Jun 28 to Jul 3, 2010; Austin, Texas. [doi: 10.25080/Majora-92bf1922-00a]
26. Bengtsson M. How to plan and perform a qualitative study using content analysis. NursingPlus Open 2016;2:8-14. [doi: 10.1016/j.npls.2016.01.001]
27. Friese S. Qualitative Data Analysis with ATLASTi, 2nd edition: Sage Publications; 2022. [doi: 10.4135/9781529799590]
28. Barrera B, Feliu A, Espina C, Brand T, Zeeb H, Ahmed F. Leveraging AI to enhance qualitative research: experiences and recommendations from case studies in cancer prevention literacy across the European Union. Int J Qual Methods 2025 Jul;24. [doi: 10.1177/16094069251365766]
29. Zarling A, Kim J, Russell D, Cutrona C. Increasing older adults’ social connectedness: development and implementation of a web-assisted acceptance and commitment therapy-based intervention. JMIR Aging 2024 Apr 22;7:e47943. [doi: 10.2196/47943] [Medline: 38647321]
30. Karami B, Ostad-Taghizadeh A, Rashidian A, Tajvar M. Developing a conceptual framework for an age-friendly health system: a scoping review. Int J Health Policy Manag 2023;12:7342. [doi: 10.34172/ijhpm.2023.7342] [Medline: 37579375]
31. Naderbagi A, Loblay V, Zahed IUM, et al. Cultural and contextual adaptation of digital health interventions: narrative review. J Med Internet Res 2024 Jul 9;26:e55130. [doi: 10.2196/55130] [Medline: 38980719]
32. Nimmanterdwong Z, Boonviriya S, Tangkijvanich P. Human-centered design of mobile health apps for older adults: systematic review and narrative synthesis. JMIR Mhealth Uhealth 2022 Jan 14;10(1):e29512. [doi: 10.2196/29512] [Medline: 35029535]
33. Dadaczynski K, Okan O, Messer M, et al. Digital health literacy and web-based information-seeking behaviors of university students in Germany during the COVID-19 pandemic: cross-sectional survey study. J Med Internet Res 2021 Jan 15;23(1):e24097. [doi: 10.2196/24097] [Medline: 33395396]
34. Yuen E, Winter N, Savira F, et al. Digital health literacy and its association with sociodemographic characteristics, health resource use, and health outcomes: rapid review. Interact J Med Res 2024 Jul 26;13:e46888. [doi: 10.2196/46888] [Medline: 39059006]
35. Global Strategy on Digital Health 2020-2025: World Health Organization; 2021. URL: http://apps.who.int/bookorders.%0Ahttp://apps.who.int/bookorders.%0Ahttp://apps.who.int/bookorders.%0Ahttps://www.who.int/docs/default-source/documents/gs4dhdaa2a9f352b0445bafbc79ca799dce4d.pdf [accessed 2025-09-13]
36. Maita KC, Maniaci MJ, Haider CR, et al. The impact of digital health solutions on bridging the health care gap in rural areas: a scoping review. Perm J 2024 Sep 16;28(3):130-143. [doi: 10.7812/TPP/23.134] [Medline: 39135461]
37. Regulation (EU) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (general data protection regulation). Official Journal of the European Union 2016;(L119/1). URL: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN [accessed 2025-03-12]
38. Katusiime J, Pinkwart N. A review of privacy and usability issues in mobile health systems: Role of external factors. Health Informatics J 2019 Sep;25(3):935-950. [doi: 10.1177/1460458217733121] [Medline: 28992749]
39. Kolasa K, Mazzi F, Leszczuk-Czubkowska E, Zrubka Z, Péntek M. State of the art in adoption of contact tracing apps and recommendations regarding privacy protection and public health: systematic review. JMIR Mhealth Uhealth 2021 Jun 10;9(6):e23250. [doi: 10.2196/23250] [Medline: 34033581]

XSL•FO
RenderX

40. Seto E, Challa P, Ware P. Adoption of COVID-19 contact tracing apps: a balance between privacy and effectiveness. J Med Internet Res 2021 Mar 4;23(3):e25726. [doi: 10.2196/25726] [Medline: 33617459]
41. Zegers CML, Witteveen A, Schulte MHJ, et al. Mind your data: privacy and legal matters in eHealth. JMIR Form Res 2021 Mar 17;5(3):e17456. [doi: 10.2196/17456] [Medline: 33729163]
42. Gentry SV, Gauthier A, L'Estrade Ehrstrom B, et al. Serious gaming and gamification education in health professions: systematic review. J Med Internet Res 2019 Mar 28;21(3):e12994. [doi: 10.2196/12994] [Medline: 30920375]
43. Ahmed F, Brand T, Zeeb H. Piloting a cancer awareness app across six European countries: a pre-post study. Front Public Health Forthcoming 2025. [doi: 10.3389/fpubh.2025.1648212]
44. Tsai CH, Eghdam A, Davoody N, Wright G, Flowerday S, Koch S. Effects of electronic health record implementation and barriers to adoption and use: a scoping review and qualitative analysis of the content. Life (Basel) 2020 Dec 4;10(12):327. [doi: 10.3390/life10120327] [Medline: 33291615]

Abbreviations

AI: artificial intelligence
BUMPER: Boosting the Usability of the EU Mobile App for Cancer Prevention
DHL: digital health literacy
ECAC: European Code Against Cancer
eHEALS: eHealth Literacy Scale
FGD: focus group discussion

Edited by A Mavragani; submitted 13.03.25; peer-reviewed by A Noblin, N Ribeiro; revised version received 15.07.25; accepted 24.07.25; published 19.09.25.

Please cite as:
Ahmed F, Romero Saletti S, D'Souza E, Espina C, Ritchie D, Barceló AM, Carbó MP, Cervera PR, Seum T, Brenner H, Van den Broucke S, Krini M, Fonseca C, Pinto P, Krivic D, Comesana HR, Yared W, Zeeb H, Brand T
Assessing the User Experience of the EU Mobile App for Cancer Prevention: Mixed Methods Study
JMIR Form Res 2025;9:e73844
URL: https://formative.jmir.org/2025/1/e73844
doi:10.2196/73844

© Furqan Ahmed, Silvana Romero Saletti, Erica D'Souza, Carolina Espina, David Ritchie, Ana Molina Barceló, Marina Pinto Carbó, Paula Romeo Cervera, Teresa Seum, Hermann Brenner, Stephan Van den Broucke, Maria Krini, Cristiana Fonseca, Patricia Pinto, Diana Krivic, Helena Ros Comesana, Wendy Yared, Hajo Zeeb, Tilman Brand. Originally published in JMIR Formative Research (https://formative.jmir.org), 19.9.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Knowledge and Perceptions of AI Among Medical Students in Morocco: Cross-Sectional Study

Imad Chakri, MD; Otmane El Khayali, MD; Laila Lahlou, MD

Faculty of Medicine and Pharmacy Agadir, University Ibn Zohr, Quartier Tilila Bp 7519 Agence Abb Agadir Al Fidia, Agadir, Morocco

Corresponding Author:
Imad Chakri, MD
Faculty of Medicine and Pharmacy Agadir, University Ibn Zohr, Quartier Tilila Bp 7519 Agence Abb Agadir Al Fidia, Agadir, Morocco

Abstract

Background: Artificial intelligence (AI) is rapidly transforming medical practice by enhancing diagnostic accuracy, streamlining workflows, and supporting clinical decision-making. However, the integration of AI into health care largely depends on the preparedness and acceptance of future physicians. Therefore, assessing their knowledge and perceptions of AI is crucial. Notably, no study has yet evaluated these factors among medical students in Morocco.

Objective: The aim of this study was to describe Moroccan medical students' knowledge and perception of AI.

Methods: A cross-sectional, observational study was conducted from February to May 2023 at the Faculty of Medicine and Pharmacy, Agadir, Morocco. All undergraduate medical students from the first to the seventh year were eligible, excluding graduate students. A snowball sampling method was used, with a calculated minimum sample size of 385. To account for potential missing data, and given the target population size of 1150, the sample size was increased by 50%. Data were collected through a validated online questionnaire and analyzed using JAMOVI 2.6.2, with significance set at $P<.05$.

Results: A total of 580 medical students (female n=363, 62.6%; mean age 21.3, SD 2.13 years; response rate 50.4%) participated. While 96% (n=557) had heard of AI, 73.1% (n=424) were unfamiliar with key AI terminologies, only 11% (n=64) understood AI functioning, and 14.8% (n=86) were familiar with everyday AI applications. Objectively, 88.1% (n=511) correctly identified deep learning as a method for automated pattern recognition, with 71.5% (n=415) acknowledging its interpretability challenges. First-cycle students demonstrated significantly higher familiarity with AI terms (83/156, 53.2% vs 51/156, 32.7% vs 22/156, 14.1%; $P<.001$). In terms of perception, 83% (n=482) viewed AI as a collaborative tool, 84.1% (n=488) anticipated a transformative impact on medicine, 39% (n=227) expected noninterventional medicine to be replaced within a decade, and 57.1% (n=331) believed certain specialties could be supplanted by AI. Regarding AI in medical education, 90% (n=522) supported its integration into the curriculum and 94% (n=546) expected enhanced learning conditions, but only 48.1% (n=279) felt ready to use AI tools upon graduation. Additionally, gender and technology familiarity significantly influenced specific perceptions, with technology-savvy students reporting greater readiness ($P<.001$) and women more likely to view AI as revolutionary (315/488, 64.5% vs 173/488, 35.5%; $P=.02$).

Conclusions: Medical students' knowledge of AI is still limited, but their awareness of the potential impact of this technology on future practice and their openness to its integration into the medical curriculum constitute a promising basis for the successful implementation of these new concepts in our health care system.

(JMIR Form Res 2025;9:e66156) doi:10.2196/66156

KEYWORDS

knowledge; perception; medical education; medical knowledge; medical training; medical students; machine learning; ML; artificial intelligence; AI; large language models; LLM; natural language processing; NLP; deep learning; cross-sectional studies; observational studies; Morocco

Introduction

Artificial intelligence (AI) is commonly defined as the capability of machines to perform tasks that typically require human intelligence, such as reasoning, learning, problem-solving, and decision-making [1]. AI encompasses various subfields, including machine learning (ML), which enables systems to identify patterns and improve performance through experience without explicit programming, and deep learning (DL), a specialized branch of ML that utilizes artificial neural networks to analyze complex data structures with high accuracy [2].

Recent advancements in computational power, big data analytics, and algorithmic efficiency have accelerated AI's integration across multiple domains, including finance, education, and transportation [3-5]. In health care, AI has

XSL•FO
RenderX

demonstrated remarkable potential in diagnostics, predictive modeling, medical imaging, and personalized treatment, paving the way for a paradigm shift in clinical practice and medical education [6-8].

However, despite its transformative potential, AI also introduces critical challenges related to ethics, data security, and regulatory compliance, necessitating careful oversight to ensure responsible and equitable implementation in health care settings. Concerns regarding patient data privacy, algorithmic bias, and the interpretability of AI-driven decisions highlight the need for stringent regulatory frameworks and ethical guidelines to govern AI applications in medicine [9-12]. Additionally, AI adoption in health care requires well-trained professionals who can critically evaluate and effectively utilize AI-driven tools while ensuring patient safety and maintaining ethical standards.

Given the increasing role of AI in modern health care, medical education must evolve to equip future physicians with the necessary knowledge and skills to integrate AI into clinical practice. Various international organizations, including the World Medical Association, have advocated for incorporating AI-related topics into medical curricula to enhance AI literacy among health care professionals [13]. Despite these recommendations, AI education remains largely absent from undergraduate medical training in many countries, including Morocco, leaving medical students with limited exposure to AI concepts, applications, and ethical implications.

This study aimed to describe the knowledge and perceptions of medical students in Morocco regarding AI. It seeks to assess their familiarity with AI applications in medicine, their perceptions of AI's role in medical practice, and the potential impact of AI on their career choices.

Methods

Design

A cross-sectional study was conducted among undergraduate medical students at the Faculty of Medicine and Pharmacy, Agadir (FMPA), Morocco, from February 1, 2023, to May 1, 2023. An online self-administered questionnaire served as the principal instrument for data collection. All undergraduate students were included. Graduate students were excluded because, at the time of the study, FMPA had not yet had any graduates. Students from other medical faculties were also excluded. There were no exclusion criteria based on age or gender.

Sampling

A snowball sampling method was employed. The minimum sample size was calculated with a 5% margin of error, 5% precision, 95% confidence level [14,15], and an estimated AI knowledge prevalence of 50% among medical students. This estimate is based on a study in Saudi Arabia [16], which found that approximately 50% of students believed they had a good understanding of AI; however, when knowledge of AI was tested, only 22% of the questions were answered correctly. The required sample size was determined to be at least 385 students. To account for potential missing data, and given that the target population size was 1150, we increased the sample size by 50%.

Questionnaire Development and Validation

The questionnaire was created using commonly used questions from previous studies [17-19] and reviewed by the authors to ensure clarity, completeness, and relevance, with specific improvements made based on feedback. To validate the clarity of the questions, a small pilot group of 10 students completed the survey.

The survey is structured into 32 questions across 5 sections:

- Demographics (4 questions): This section collects essential demographic information, including age, gender, educational level, and study location.
- Knowledge of AI (12 questions): Participants are asked about their sources of information on AI, familiarity with AI terminology, understanding of AI applications in daily life, awareness of AI investments in health care, and interest in AI among medical professionals.
- Perceptions of AI (7 questions): This section explores participants' views on whether AI will collaborate with or compete against health care providers, its potential impact on medicine, concerns about AI replacing doctors, and overall attitudes toward AI development.
- AI in medical education (6 questions): Participants are asked about the perceived benefits of AI in medical education, potential changes in teaching methods due to AI, and related considerations.
- Influence of AI on specialty choice (3 questions): This section examines participants' opinions on which medical specialties will be most affected by AI, concerns about AI replacing doctors in certain specialties, and whether AI advancements might influence their choice of medical specialty.

Data Collection

Data were collected using a self-administered online questionnaire developed on the Google Forms platform and written in French. The questionnaire employed straightforward and comprehensible language, disseminated through social media channels, particularly on WhatsApp groups comprising students from the first to the seventh year at FMPA.

Ethical Considerations

Ethics Approval and Review

This study was conducted in accordance with the ethical principles outlined in the Declaration of Helsinki (1964) and its subsequent amendments [20]. As there is currently no institutional review board or ethics committee at our faculty, the study was conducted following institutional ethical guidelines and best practices for research involving human participants.

Informed Consent

Participation in this study was entirely voluntary. All participants were informed about the study's objectives, methods, and their rights, including the right to withdraw at any time without any consequences. Informed consent was obtained from each participant before data collection. Since the study involved secondary data analysis, consent for the original data collection

XSL•FO
RenderX

was obtained, and its use for this research was conducted in line with ethical guidelines.

Privacy and Confidentiality

The study ensured complete anonymity and confidentiality of participants. No identifiable personal data were collected, and all information was deidentified prior to analysis. Data were securely stored and accessible only to authorized research personnel to prevent unauthorized access.

Compensation

Participants did not receive any financial or material compensation for their participation in this study.

Participant Identifiability in Images or Supplementary Materials

This study does not include any images or supplementary materials that could lead to the identification of individual participants. If any identifiable images were to be included, explicit consent would be obtained, and the relevant consent forms would be submitted as supplementary material.

Data Analysis

Data analysis was conducted using JAMOVI (version 2.6.2) [21,22]. Descriptive statistics summarized qualitative variables in terms of frequencies and percentages, while the quantitative variable (age) was described using mean (SD). Comparative analyses were performed using the Student *t* test for independent samples when comparing two groups and 1-way ANOVA when comparing 3 or more groups. The choice between the *t* test and ANOVA was based on the number of groups being compared.

For categorical variables, the χ^2 test or Fisher exact test was applied depending on the expected frequency of the data. This allowed assessment of students' knowledge of AI in medical education, the influence of AI on specialty choice, and predictors of perceptions (including demographics such as gender, grade, and technology experience). A P value <.05 was considered statistically significant.

Results

Demographic Characteristics

Figure 1 illustrates the participant flow, including the inclusion and exclusion criteria applied during the sampling process. A total of 580 medical students participated in our survey, yielding a response rate of 50.4%. Most of the participants were female, comprising 363 (62.6%) individuals. The mean age of the participants was 21.3 (SD 2.13) years, with an age range from 18 to 26 years.

XSL•FO
RenderX

Figure 1. Flowchart of participant inclusion and exclusion criteria. FMPA: Faculty of Medicine and Pharmacy, Agadir.

XSL•FO
RenderX

Participants were distributed across various academic years as follows: 19.1% (n=111) were first-year students, 17.6% (n=102) second-year, 13.8% (n=80) third-year, 13.8% (n=80) fourth-year, 11.7% (n=68) fifth-year, 11.9% (n=69) sixth-year, and 12.1% (n=70) seventh-year students. Regarding academic cycles, 36.7% (n=213) of students were enrolled in the first cycle (first and second years), 39.3% (n=228) of students in the second cycle (third, fourth, and fifth years), and 24% (n=139) of students in the third cycle (sixth and seventh years). The demographics of the included participants are shown in Table 1.

Table . Demographic characteristics of the medical students (n=580).

Characteristics	Values
Gender, n (%)	
Female	363 (63.6)
Male	217 (37.4)
Age (years)	
Mean (SD)	21.3 (2.13)
Range	18 - 26
Academic year, n (%)	
First year	111 (19.1)
Second year	102 (17.6)
Third year	80 (13.8)
Fourth year	80 (13.8)
Fifth year	68 (11.7)
Sixth year	69 (11.9)
Seventh year	70 (12.1)
Academic cycle, n (%)	
First cycle (first and second years)	213 (36.7)
Second cycle (third, fourth, and fifth years)	228 (39.3)
Third cycle (sixth and seventh years)	139 (24.0)

Knowledge of AI

Regarding knowledge, 249 (42.9%) of the 580 participating medical students declared themselves knowledgeable about technology. Almost all students (n=557, 96%) had heard of AI, with social media being the primary source for 545 students, followed by traditional media such as television, radio, and journals for 350 students; scientific journals for 236 students; and universities for 150 students. However, a notable 73% (n=423) of participants were unfamiliar with basic AI terminology such as machine learning, deep learning, and neural networks. Only 11% (n=64) of students claimed to have an understanding of how AI functions. Additionally, a mere 14.8% (n=86) reported familiarity with everyday AI applications like speech recognition and antispam filters.

The association of knowledge of AI with different variables is given in Table 2. Female participants showed higher self-perceived familiarity with technology (131/248, 52.8% vs 117/248, 47.2% male participants; $P<.001$) and better recognition of specific AI terms (81/156, 51.9% female vs 75/156, 48.1% male; $P=.001$). However, there was no statistical difference in understanding underlying AI technologies between genders (61.4% female vs 38.6% male; $P=.50$). Male participants demonstrated greater awareness of everyday AI applications like speech recognition and antispam filters (48/86, 55.8% vs 38/86, 44.2% female participants; $P<.001$). First-cycle students reported greater familiarity with technology than those in the second and third cycles (116/248, 46.8% vs 81/248, 32.7% and 51/248, 20.6%, respectively; $P<.001$). Additionally, first-cycle students demonstrated a better understanding of AI terms such as deep learning compared to their peers (83/156, 53.2% vs 51/156, 32.7% and 22/156, 14.1%, respectively; $P<.001$).

XSL•FO
RenderX

Table . Comparison of artificial intelligence (AI) knowledge among medical students by gender, academic cycle, and familiarity with technology.

Question and response	Gender, n (%)[a]		*P* value	Academic cycle, n (%)[a]			*P* value	Familiarity with technology, n (%)[a]		*P* value
	Female (n=363)	Male (n=217)		First (n=213)	Second (n=228)	Third (n=139)		Knowledgeable (n=332)	Not knowledgeable (n=248)	
Knowledgeable in technology?			*<.001* [b]				*<.001*	—[c]	—	—
Yes	131 (52.8)	117 (47.2)		116 (46.8)	81 (32.7)	51 (20.6)				
No	232 (69.9)	100 (30.1)		97 (29.2)	147 (44.3)	88 (26.5)				
Familiar with AI terms?			*.001*				*<.001*			*<.001*
Yes	81 (51.9)	75 (48.1)		83 (53.2)	51 (32.7)	22 (14.1)		109 (69.9)	47 (30.1)	
No	282 (66.5)	142 (33.5)		130 (30.7)	177 (41.7)	117 (27.6)		139 (32.8)	285 (67.2)	
Understanding of AI and deep learning?			*<.001*				.15			*<.001*
Great extent	25 (39.1)	39 (60.9)		31 (48.4)	21 (32.8)	12 (18.8)		54 (84.4)	10 (15.6)	
Low extent	177 (66.8)	88 (33.2)		94 (35.5)	99 (37.4)	72 (27.1)		138 (52.1)	127 (47.9)	
Not at all	161 (64.1)	90 (35.9)		88 (35.1)	108 (43)	55 (21.9)		56 (22.3)	195 (77.7)	
Aware of AI in everyday applications?			*<.001*				.39			*<.001*
Great extent	38 (44.2)	48 (55.8)		32 (37.2)	35 (40.7)	19 (22.1)		71 (82.6)	15 (17.4)	
Low extent	217 (66.2)	111 (33.8)		126 (38.4)	118 (36)	84 (25.6)		148 (45.1)	180 (54.9)	
Not at all	108 (65.1)	58 (34.9)		55 (33.1)	75 (45.2)	36 (21.7)		29 (17.5)	137 (82.5)	
Aware of AI in medical research?			.14				.11			*<.001*
Yes	193 (59.9)	129 (40.1)		130 (40.4)	117 (36.3)	75 (23.3)		193 (59.9)	129 (40.1)	
No	170 (65.9)	88 (34.1)		83 (32.2)	111 (43.0)	64 (24.8)		55 (21.3)	203 (78.7)	
Great extent	46 (54.8)	38 (45.2)		28 (33.3)	36 (42.9)	20 (23.8)		56 (66.7)	28 (33.3)	
Aware of AI as a topic of debate in the medical community?			.23				.16			*<.001*
Low extent	191 (62.8)	38 (37.2)		123 (40.5)	105 (34.5)	76 (25.0)		150 (49.3)	154 (50.7)	
Not at all	126 (65.6)	66 (34.4)		62 (32.3)	87 (45.3)	43 (22.4)		42 (21.9)	150 (78.1)	

[a]Percentages are calculated based on the total number in each row per category.

[b]Italicized values indicate statistically significant differences between subgroups based on the corresponding *P* values (<.05).

[c]Not applicable.

Figure 2 illustrates the statistically significant findings from Table 2, highlighting key differences in AI knowledge across gender, academic cycle, and technology familiarity.

XSL•FO
RenderX

Figure 2. Statistically significant findings in AI knowledge among medical students. (A) Gender differences in technology knowledge, (B) academic cycle impact on AI familiarity, (C) technology familiarity impact on AI understanding, and (D) technology familiarity impact on awareness of everyday AI applications. AI: artificial intelligence.

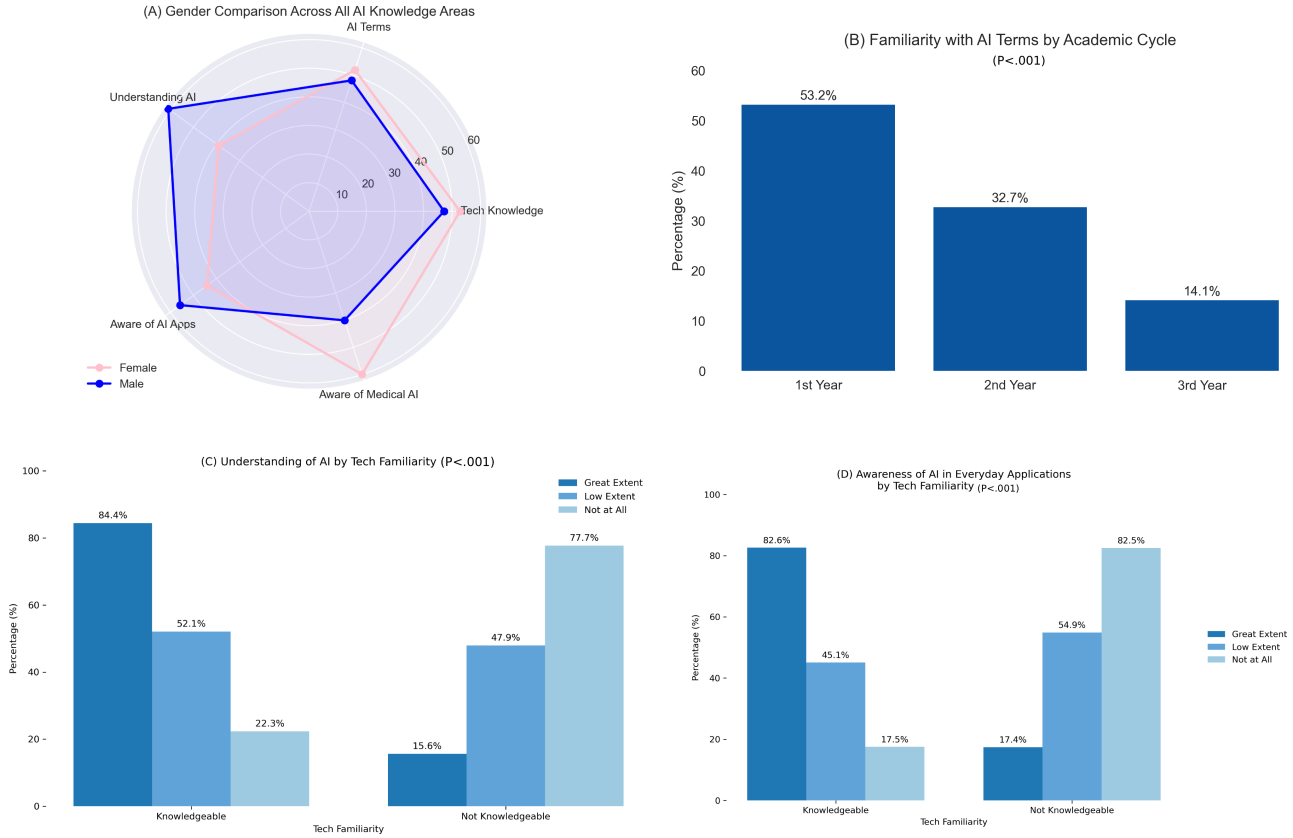

To objectively assess students' knowledge of AI concepts, a four-item questionnaire was administered. A substantial majority (n=511, 88%) accurately identified DL as a method for automated pattern recognition essential for biomedical image analysis. Similarly, 88% (n=511) recognized the importance of extensive annotated medical image databases for applying DL in radiology. A notable proportion (n=415, 71.5%) acknowledged the challenges associated with interpreting the decision-making processes of DL models. Additionally, 68% (n=394) understood the limitations of DL in deductive reasoning despite its strengths in pattern recognition. Objective comparison of AI knowledge across demographic variables revealed no significant gender differences. Students in earlier academic cycles (first and second) demonstrated a marginally better comprehension of the complexities underlying DL models compared to their counterparts in the third cycle (*P*=.009). However, this difference was not observed in other AI-related concepts. Regarding technology proficiency, an objective analysis indicated that participants with prior technology exposure were significantly more likely to accurately identify the need for large annotated medical image datasets (*P*=.01). Conversely, those without such experience exhibited a stronger understanding of DL's limitations in deductive reasoning (*P*=.03).

Perception of AI Among Medical Students

A majority of students (n=482, 83.1%) viewed AI as a collaborative tool rather than a competitor, while a smaller proportion (n=98, 16.9%) saw it as a potential rival to physicians. When considering the future of noninterventional medicine, 227 (39.1%) anticipated its replacement within a decade, whereas 353 (60.9%) disagreed. Overall, 81% of students believed AI would not supplant human physicians.

However, 488 (84.1%) anticipated AI significantly transforming medicine, while 92 (15.9%) held a different view. Student attitudes toward AI development were divided: 271 (46.7%) expressed concern, while 309 (53.3%) were optimistic. Moreover, 315 (54.3%) were enthusiastic about AI, compared to 265 (45.7%) who were less excited.

Regarding AI's perceived impact, opinions ranged from negligible (n=22, 3.8%) to substantial (n=207, 35.7%), with most students (n=302, 52.1%) perceiving a moderate influence. When examining these perceptions by gender, a significant majority of both male and female respondents viewed AI as a partner in medicine rather than a competitor (*P*=.94). However, women were more likely to perceive AI as a revolutionary force in medicine (315/488, 64.5%) compared to men (173/488, 35.5%), which was a statistically significant difference (*P*=.02).

While both genders equally anticipated the replacement of noninterventional doctors by AI within a decade (*P*=.99), women expressed greater concern about AI rendering human doctors obsolete (60/111, 54.1%) than men (51/111, 45.9%) a statistically significant difference (*P*=.04). Reactions to AI

XSL•FO
RenderX

developments were similar across genders, with no significant differences in fear (*P*=.56) or excitement (*P*=.31).

Perceptions also varied based on technology familiarity. While there was no significant difference in viewing AI as a partner (*P*=.67), students less familiar with technology were more likely to believe AI will revolutionize medicine (260/488, 53.3% vs 228/488, 46.7%; *P*<.001) and make human doctors unnecessary (74/111, 66.7% vs 37/111, 33.3%; *P*=.03). Nonexperts also exhibited greater fear (172/271, 63.5% vs 99/271, 36.5%; *P*=.005) but less enthusiasm (138/315, 43.8% vs 177/315, 56.2%; *P*<.001) for AI developments compared to technology-savvy students. The association of perception of AI with different variables is given in Table 3.

Table . Comparative analysis of artificial intelligence (AI) perceptions among medical students by gender, familiarity with technology, and academic cycle.

Question and response	Gender, n (%)[a]		*P* value	Familiarity with technology, n (%)[a]		*P* value	Academic cycle, n (%)[a]			*P* value
	Male (n=217)	Female (n=363)		Knowledgeable (n=248)	Not knowledgeable (n=305)		First (n=213)	Second (n=228)	Third (n=139)	
I perceive AI in medicine as a partner rather than a competitor.			.94			.67				.14
Yes	180 (37.3)	302 (62.7)		208 (43.2)	247 (56.8)		172 (35.7)	187 (38.8)	123 (25.5)	
No	37 (37.8)	61 (62.2)		40 (40.8)	58 (59.2)		41 (41.8)	41 (41.8)	16 (16.3)	
AI will revolutionize medicine in general.			*.02*[b]			*<.001*				*.002*
Yes	173 (35.5)	315 (64.5)		228 (46.7)	260 (53.3)		173 (35.5)	185 (37.9)	130 (26.6)	
No	44 (47.8)	48 (52.2)		20 (21.7)	72 (78.3)		40 (43.5)	43 (46.7)	9 (9.8)	
Noninterventional physicians will be replaced within a decade by AI.			.99			.72				*.006*
Yes	85 (37.4)	142 (62.6)		95 (41.9)	132 (58.1)		101 (44.5)	82 (36.1)	44 (19.4)	
No	132 (37.4)	221 (62.6)		153 (43.3)	200 (56.7)		112 (31.7)	146 (41.4)	95 (26.9)	
AI will render human physicians obsolete.			*.04*			*.03*				.16
Yes	51 (45.9)	60 (54.1)		37 (33.3)	74 (66.7)		43 (38.7)	49 (44.1)	19 (17.1)	
No	166 (35.4)	303 (64.6)		211 (45.0)	258 (55.0)		170 (36.2)	179 (38.2)	120 (25.6)	
AI developments frighten me.			.56			*.005*				*<.001*
Yes	98 (36.2)	173 (63.8)		99 (36.5)	172 (63.5)		121 (44.6)	(96 (35.4)	54 (19.9)	
No	119 (38.5)	190 (61.5)		149 (48.2)	160 (51.8)		92 (29.8)	132 (42.7)	85 (27.5)	
AI developments make medicine more exciting for me.			.31			*<.001*				*.005*
Yes	112 (35.6)	203 (64.4)		177 (56.2)	138 (43.8)		111 (35.2)	112 (35.6)	92 (29.2)	
No	105 (39.6)	160 (60.4)		71 (26.8)	194 (73.2)		102 (38.5)	116 (43.8)	47 (17.7)	

[a]Percentages are calculated based on the total number in each row per category.

[b]Italicized values indicate statistically significant differences between subgroups based on the corresponding *P* values (<.05).

Notably, women showed a higher perception of AI revolutionizing medicine (315/488, 64.5% vs 173/488, 35.5%; *P*=.02), while technological familiarity significantly impacted perspectives. Students less familiar with technology were more likely to believe AI would revolutionize medicine (260/488, 53.3% vs 228/488, 46.7%; *P*<.001) and potentially render physicians obsolete (74/107, 66.7% vs 33/107, 33.3%; *P*=.03). Academic cycle progression revealed declining enthusiasm, with first-cycle students showing 43.5% (40/92) positive perception, decreasing to 9.8% (9/92) by the third cycle. Technology-naive students demonstrated greater fear of AI developments (172/271, 63.5% vs 99/271, 36.5%; *P*=.005) and less excitement about AI in medicine (138/315, 43.8% vs 177/315, 56.2%; *P*<.001), highlighting the complex relationship between technological exposure and perception.

AI in Medical Education

In exploring the integration of AI in medical education among Moroccan students, findings revealed strong enthusiasm and perceived benefits. A significant majority (n=524, 90.3%) expressed support for incorporating AI into their curriculum, while an even higher percentage (n=539, 92.9%) believed AI education would positively impact their careers. Regarding its impact on learning conditions, nearly all students (n=546,

XSL•FO
RenderX

94.1%) anticipated AI to enhance their educational experiences. However, when assessing readiness to utilize AI tools upon graduation, opinions were divided: 279 (48.1%) felt capable, whereas 301 (51.9%) did not.

Regarding awareness of AI in medical practice, only 168 (29%) considered themselves well-informed, with the majority (n=412, 71%) feeling less knowledgeable. Examples cited by informed students included AI's role in medical imaging assistance, robotic surgery, and medical simulation.

Comparative analysis showed that gender did not significantly influence perceptions as responses were similar between men and women across all questions, with *P* values ranging from .15 to .91.

Differences emerged when analyzing responses by academic cycle. While students from all cycles were generally positive about incorporating AI into medical education (*P*=.23), their perceptions varied significantly regarding AI's impact on medical training (*P*=.04) and its influence on career prospects (*P*=.048). Additionally, there were notable differences in readiness to use AI tools upon graduation (*P*=.03). Specifically, students in the third cycle showed lower confidence (76/279, 27.2%) compared to those in the second cycle (115/279, 41.2%) and first cycle (88/279, 31.5%).

Familiarity with technology also played a role in shaping perceptions. Technology-savvy students were more likely to support the inclusion of AI in the curriculum (*P*=.16), believe in its career benefits (*P*=.26), and expect it to improve learning conditions (*P*=.84). Notably, students familiar with technology felt more prepared to use AI tools upon graduation (*P*<.001), with 56.6% (158/279) of technology-savvy students expressing confidence compared to 29.9% (90/301) of those less familiar with technology. Table 4 presents the results of the comparative analysis.

Table . Comparative analysis of perception of artificial intelligence (AI) in medical education by gender, academic cycle, and familiarity with technology.

Question and response	By gender, n (%)[a]		*P* value	By academic cycle, n (%)[a]			*P* value	By familiarity with technology, n (%)[a]		*P* value
	Female (n=363)	Male (n=217)		First (n=213)	Second (n=228)	Third (n=139)		Knowledgeable (n=248)	Not knowledgeable (n=332)	
AI should be part of medical education.			.39				.23			.16
Yes	325 (62.0)	199 (38.0)		193 (35.8)	216 (40.1)	130 (24.1)		229 (43.7)	295 (56.3)	
No	38 (67.9)	18 (32.1)		20 (48.8)	12 (29.3)	9 (22.0)		19 (33.9)	37 (66.1)	
AI education will benefit my career.			.91				*.04* [b]			.26
Yes	337 (62.5)	202 (37.5)		184 (35.1)	210 (40.1)	130 (24.8)		227 (42.1)	312 (57.9)	
No	26 (63.4)	15 (36.6)		29 (51.8)	18 (32.1)	9 (16.1)		21 (51.2)	20 (48.8)	
AI will improve learning conditions.			.64				*.048*			.85
Yes	343 (62.8)	203 (37.2)		194 (35.5)	220 (40.3)	132 (24.2)		234 (42.9)	312 (57.1)	
No	20 (58.8)	14 (41.2)		19 (55.9)	8 (23.5)	7 (22.6)		14 (41.2)	20 (58.8)	
I will be able to use AI tools in health care after graduation.			.15				*.03*			*<.001*
Yes	183 (65.6)	96 (34.4)		88 (31.5)	115 (41.2)	76 (27.2)		158 (56.6)	121 (43.4)	
No	180 (59.8)	121 (40.2)		125 (41.5)	113 (37.5)	63 (20.9)		90 (29.9)	211 (70.1)	

[a]Percentages are calculated based on the total number in each row per category.

[b]Italicized values in the table indicate statistically significant differences between subgroups based on the corresponding *P* values (<.05).

Table 4 reveals the evolution of medical students' perceptions about AI in medical education across academic cycles and technological familiarity. There was a declining belief that AI education would benefit careers, decreasing from 35.1% (184/524) in the first cycle to 24.8% (130/524) in the third cycle (*P*=.04). Similarly, there was a reduction in perceiving AI's potential to improve learning conditions, from 35.5% (194/546) in the first cycle to 24.2% (132/546) in the third cycle (*P*=.048). A progressive decrease was also noted in confidence about using AI tools post graduation, dropping from 31.5% (88/279) in the first cycle to 27.2% (76/279) in the third cycle (*P*=.03). There was also a significant impact of technological familiarity on confidence in using AI tools after graduation, with knowledgeable students showing 56.6% (158/279) confidence compared to 43.4% (121/279) among those less familiar (*P*<.001).

The Influence of AI on Students' Perceptions and Specialty Choices

The impact of AI on students' perceptions and specialty choices in medical education reveals several noteworthy findings. First, radiology was overwhelmingly viewed as the specialty most likely to be affected by AI, with 63% (n=365) of students selecting it as the top choice, followed by surgery at 17% (n=99), and anatomical pathology at 11% (n=62). Concerning the replacement of medical specialties by AI, 57% (n=331) of

XSL•FO
RenderX

students believed certain specialties would be substituted, while 43% (249) held differing views. Regarding the influence of AI advancements on specialty selection, a significant majority (n=327, 56%) indicated they were not inclined to choose specialties heavily influenced by AI progress, contrasting with 43% (n=249) who felt otherwise.

Analysis by gender showed no substantial influence on specialty choices. Similarly, technological familiarity did not significantly impact the perception of AI replacing medical specialties (*P*=.35). However, significant differences emerged when considering whether students would avoid AI-affected specialties, with 48% (119/327) of technology-savvy students versus 62.7% (208/327) of nonsavvy students (*P*<.001) indicating potential avoidance. Additionally, while there was no significant difference among different academic cycles regarding the replacement of medical specialties by AI (*P*=.70), a statistically significant difference was observed (*P*<.001) concerning the avoidance of AI-impacted specialties across academic stages. Table 5 presents the results of the comparative analysis.

Table . Influence of artificial intelligence (AI) on students' perceptions and specialty choices by gender, technology familiarity, and academic cycle.

Question and response	By gender, n (%)[a]		*P* value	By academic cycle, n (%)[a]			*P* value	By familiarity with technology, n (%)[a]		*P* value
	Female (n=363)	Male (n=217)		First (n=213)	Second (n=228)	Third (n=139)		Knowledgeable (n=248)	Not knowledgeable (n=332)	
Certain specialties will be replaced by AI.			.16				.23			.35
Yes	199 (60.1)	132 (39.9)		126 (38.1)	129 (39.0)	76 (23.0)		136 (41.1)	195 (58.9)	
No	164 (65.9)	85 (34.1)		87 (34.9)	99 (39.8)	63 (25.3)		112 (45.0)	137 (55.0)	
I might not choose a specialty like radiology because of AI.			.82				*<.001* [b]			*<.001*
Yes	206 (63.0)	121 (37.0)		138 (42.2)	128 (39.1)	61 (18.7)		119 (48.0)	208 (62.7)	
No	157 (62.1)	96 (37.9)		75 (29.6)	100 (39.5)	78 (30.8)		129 (52.0)	124 (37.3)	

[a]Percentages are calculated based on the total number in each row per category.

[b]Italicized values indicate statistically significant differences between subgroups based on the corresponding *P* values (<.05).

Discussion

This study aimed to assess the knowledge and perceptions of AI among medical students in a medical school in Morocco. The findings reveal a generally low level of AI knowledge, with significant variations based on gender and year of study, yet students maintain a positive perception of AI, particularly regarding its potential to revolutionize medicine and its integration into medical education.

The study identified a female majority among the participants, comprising 63% of the sample. This gender distribution aligns with national data from the Haut-Commissariat au Plan, which reported a 67.8% female presence in medical faculties [23]. Similar trends are observed internationally, as evidenced by studies conducted by Bisdas et al [24] (66.5%), Pinto Dos Santos et al [17] (63%), and Buabbas et al [25] (88.9%). The consistent female predominance across different contexts suggests a broader trend in the medical education landscape.

Our study shows that although a large majority of students (96%) have been exposed to AI-related terms, there is a significant gap in understanding its core concepts. Specifically, 73% of students are unfamiliar with terms like machine learning and neural networks. This finding aligns with a study in Syria, where 70% of participants had a basic understanding of AI, but only 34.7% knew about ML and DL, and just 23.7% understood AI applications [26]. An additional study from India reported that 73.6% of respondents felt unknowledgeable about AI's basic principles and applications, with 81.1% lacking understanding of AI's limitations [27]. Similarly, a German study found that while 83.3% of students understood the term "artificial intelligence," familiarity with more specific terms was lower: 65.9% knew about "machine learning," 42.3% about "neural networks," and only 18.7% about "deep learning" [28]. These findings suggest that despite widespread awareness of AI, students' knowledge is often superficial, likely influenced more by popular media than by academic sources.

The gender differences in AI knowledge observed in our study, where female participants reported higher familiarity with technology and AI terms but male participants were more aware of AI applications, reflect patterns seen in other research. For instance, a UK study found that male students were more aware of AI applications despite both genders having similar understanding of AI principles [19]. This aligns with our study, suggesting that while women may be more comfortable with technology and AI terminology, men may have a greater exposure to AI applications in their daily lives. Additionally, research in Germany showed that male students had higher AI knowledge than female students [17]. This discrepancy between self-reported familiarity and actual knowledge underscores the need to distinguish between perceived and objective understanding of AI concepts, as both genders may have similar objective knowledge but different experiences and exposures.

Despite limited knowledge, students had a generally positive perception of AI. Most of them (83%) saw AI as a partner rather

XSL•FO
RenderX

than a competitor, which is consistent with the 72.2% reported in the Bisdas et al [24] study. Additionally, 84% believed AI would revolutionize medicine, similar to findings from Bisdas et al [24] (83.9%) and Buabbas et al [25] (99.1%). However, 61% thought AI might replace noninterventionist physicians, a concern more prevalent in Morocco compared to other countries (eg, 32.5% in Bisdas et al's [24] study and 6.5% in Pinto Dos Santos et al's [17] study). This reflects a nuanced understanding among students, recognizing AI's potential and its limitations in clinical practice.

Our study also found that 56% of students would avoid specialties heavily influenced by AI, with this sentiment more pronounced among those less familiar with technology (62.7%) compared to technology-savvy students (48%). This avoidance behavior has been documented in other studies as well. For example, a survey of radiology students in Canada found that 45% were concerned about AI replacing their future roles, leading to hesitation in choosing radiology as a specialty [29].

Enthusiasm for AI in medical education was high, with 90% of students supporting its inclusion, exceeding the 75% reported among UK medical students and 69.4% among Malaysian students [19,30]. However, only 48% felt ready-to-use AI tools upon graduation, with confidence declining to 27.2% in the third cycle. This aligns with findings from Malaysia, where 44.5% of students felt prepared to apply AI in practice, and a Qatari study, where 43.8% stated they would consistently use AI in decision-making [30,31].

This study highlights key implications for medical education in Morocco. The significant knowledge gap in AI among students signals the need for urgent curricular reform, particularly by introducing AI-specific modules early in medical training. The positive perception of AI as a partner suggests that future doctors are ready to embrace AI tools in clinical practice, potentially enhancing patient care. Additionally, comparing these findings with global studies underscores the importance of aligning Moroccan medical education with international standards.

However, this study has limitations. The use of snowball sampling may have introduced selection bias, affecting the generalizability of the results. Also, focusing on a specific region might not fully represent the broader context of medical education in Morocco. Finally, reliance on self-reported data could lead to inaccuracies as students might over- or underestimate their knowledge and perceptions of AI.

In conclusion, this study sheds light on the current state of AI knowledge and perceptions among medical students in Morocco, revealing both challenges and opportunities. While there is a general enthusiasm for AI, the significant knowledge gap underscores the need for curricular reform and targeted educational interventions. The positive perceptions of AI suggest that students are ready to engage with these technologies, paving the way for a future where AI plays a central role in medical practice.

Acknowledgments

The authors gratefully acknowledge JMIR Publications for granting an article processing fee waiver for this manuscript. We also extend our sincere thanks to all the medical students who participated in this study for their valuable time and insights.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

References

1. Michalski RS, Carbonell JG, Mitchell TM. Machine Learning: An Artificial Intelligence Approach: Elsevier Reference Monographs; 2014:63089.
2. Choi RY, Coyner AS, Kalpathy-Cramer J, Chiang MF, Campbell JP. Introduction to machine learning, neural networks, and deep learning. Transl Vis Sci Technol 2020 Feb 27;9(2):14. [doi: 10.1167/tvst.9.2.14] [Medline: 32704420]
3. Abirami S, Pethuraj M, Uthayakumar M, Chitra P. A systematic survey on big data and artificial intelligence algorithms for intelligent transportation system. Case Studies on Transport Policy 2024 Sep;17:101247. [doi: 10.1016/j.cstp.2024.101247]
4. Xu X. AI optimization algorithms enhance higher education management and personalized teaching through empirical analysis. Sci Rep 2025 Mar 24;15(1):10157. [doi: 10.1038/s41598-025-94481-5] [Medline: 40128297]
5. Babatunde Adeyeri T. Economic impacts of AI-driven automation in financial services. Int J Sci Res Manage 2024 Jul 6;12(7):6779-6791. [doi: 10.18535/ijsrm/v12i07.em07]
6. Alowais SA, Alghamdi SS, Alsuhebany N, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ 2023 Sep 22;23(1):689. [doi: 10.1186/s12909-023-04698-z] [Medline: 37740191]
7. Pinto-Coelho L. How artificial intelligence is shaping medical imaging technology: a survey of innovations and applications. Bioengineering (Basel) 2023 Dec 18;10(12):1435. [doi: 10.3390/bioengineering10121435] [Medline: 38136026]
8. Alum EU, Ugwu OPC. Artificial intelligence in personalized medicine: transforming diagnosis and treatment. Discov Appl Sci 2025 Mar 1;7(3):193. [doi: 10.1007/s42452-025-06625-x]

XSL•FO
RenderX

9. Harishbhai Tilala M, Kumar Chenchala P, Choppadandi A, et al. Ethical considerations in the use of artificial intelligence and machine learning in health care: a comprehensive review. Cureus 2024 Jun;16(6):e62443. [doi: 10.7759/cureus.62443] [Medline: 39011215]
10. Bouderhem R. Shaping the future of AI in healthcare through ethics and governance. Humanit Soc Sci Commun 2024 Mar 15;11(1):1-12. [doi: 10.1057/s41599-024-02894-w]
11. Hanna MG, Pantanowitz L, Jackson B, et al. Ethical and bias considerations in artificial intelligence/machine learning. Mod Pathol 2025 Mar;38(3):100686. [doi: 10.1016/j.modpat.2024.100686] [Medline: 39694331]
12. Murdoch B. Privacy and artificial intelligence: challenges for protecting health information in a new era. BMC Med Ethics 2021 Sep 15;22(1):122. [doi: 10.1186/s12910-021-00687-3] [Medline: 34525993]
13. WMA statement on augmented intelligence in medical care. World Medical Association. 2019 Oct. URL: https://www.wma.net/policies-post/wma-statement-on-augmented-intelligence-in-medical-care/ [accessed 2025-04-27]
14. Cochran WG. Sampling Techniques, 3rd edition: John Wiley & Sons; 1977.
15. Barlett JE, Kotrlik J, Higgins C. Organizational research: determining appropriate sample size in survey research. Inf Technol Learn Perform J 2001 Jan 1;19(1) [FREE Full text]
16. Bin Dahmash A, Alabdulkareem M, Alfutais A, et al. Artificial intelligence in radiology: does it impact medical students preference for radiology as their future career? BJR Open 2020;2(1):20200037. [doi: 10.1259/bjro.20200037] [Medline: 33367198]
17. Pinto Dos Santos D, Giese D, Brodehl S, et al. Medical students' attitude towards artificial intelligence: a multicentre survey. Eur Radiol 2019 Apr;29(4):1640-1646. [doi: 10.1007/s00330-018-5601-1] [Medline: 29980928]
18. Al Saad MM, Shehadeh A, Alanazi S, et al. Medical students' knowledge and attitude towards artificial intelligence: an online survey. Open Public Health J 2022;15(e187494452203290). [doi: 10.2174/18749445-v15-e2203290]
19. Sit C, Srinivasan R, Amlani A, et al. Attitudes and perceptions of UK medical students towards artificial intelligence and radiology: a multicentre survey. Insights Imaging 2020 Feb 5;11(1):14. [doi: 10.1186/s13244-019-0830-7] [Medline: 32025951]
20. World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA 2013 Nov 27;310(20):2191-2194. [doi: 10.1001/jama.2013.281053] [Medline: 24141714]
21. The Jamovi Project. Jamovi. 2024. URL: https://www.jamaovi.org [accessed 2025-04-27]
22. R Core Team. R: A Language and environment for statistical computing. 2024. URL: https://cran.r-project.org [accessed 2025-04-27]
23. Gauthier C. Le haut-commissariat au plan publie « la femme marocaine en chiffres: 20 ans de progrès » [Article in French]. Site Institutionnel Du Haut-Commissariat Au Plan Du Royaume Du Maroc. URL: https://www.hcp.ma/Le-haut-commissariat-au-plan-publie-la-femme-marocaine-en-chiffres-20-ans-de-progres_a2897.html [accessed 2025-04-27]
24. Bisdas S, Topriceanu CC, Zakrzewska Z, et al. Artificial intelligence in medicine: a multinational multi-center survey on the medical and dental students' perception. Front Public Health 2021;9:795284. [doi: 10.3389/fpubh.2021.795284] [Medline: 35004598]
25. Buabbas AJ, Miskin B, Alnaqi AA, et al. Investigating students' perceptions towards artificial intelligence in medical education. Healthcare (Basel) 2023 May 1;11(9):1298. [doi: 10.3390/healthcare11091298] [Medline: 37174840]
26. Swed S, Alibrahim H, Elkalagi NKH, et al. Knowledge, attitude, and practice of artificial intelligence among doctors and medical students in Syria: a cross-sectional online survey. Front Artif Intell 2022;5:1011524. [doi: 10.3389/frai.2022.1011524] [Medline: 36248622]
27. Kansal R, Bawa A, Bansal A, et al. Differences in knowledge and perspectives on the usage of artificial intelligence among doctors and medical students of a developing country: a cross-sectional study. Cureus 2022 Jan;14(1):e21434. [doi: 10.7759/cureus.21434] [Medline: 35223222]
28. Mehta N, Harish V, Bilimoria K, et al. Knowledge and attitudes on artificial intelligence in healthcare: a provincial survey study of medical students. MedEdPublish 2021;10(75). [doi: 10.15694/mep.2021.000075.1]
29. Gong B, Nugent JP, Guest W, et al. Influence of artificial intelligence on Canadian medical students' preference for radiology specialty: a national survey study. Acad Radiol 2019 Apr;26(4):566-577. [doi: 10.1016/j.acra.2018.10.007] [Medline: 30424998]
30. Tung AYZ, Dong LW. Malaysian medical students' attitudes and readiness toward AI (artificial intelligence): a cross-sectional study. J Med Educ Curric Dev 2023;10:23821205231201164. [doi: 10.1177/23821205231201164] [Medline: 37719325]
31. Ahmad MN, Abdallah SA, Abbasi SA, Abdallah AM. Student perspectives on the integration of artificial intelligence into healthcare services. Digit Health 2023;9:20552076231174095. [doi: 10.1177/20552076231174095] [Medline: 37312954]

Abbreviations

AI: artificial intelligence
DL: deep learning
FMPA: Faculty of Medicine and Pharmacy, Agadir
ML: machine learning

XSL•FO
RenderX

Edited by A Mavragani; submitted 05.09.24; peer-reviewed by KM Giannakopoulou, ST Arasteh; revised version received 27.04.25; accepted 02.05.25; published 19.09.25.

Please cite as:
Chakri I, El Khayali O, Lahlou L
Knowledge and Perceptions of AI Among Medical Students in Morocco: Cross-Sectional Study
JMIR Form Res 2025;9:e66156
URL: https://formative.jmir.org/2025/1/e66156
doi:10.2196/66156

© Imad Chakri, Otmane El Khayali, Laila Lahlou. Originally published in JMIR Formative Research (https://formative.jmir.org), 19.9.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Understanding Health Care Workers' Attitudes and Preferences Toward Digital Patient Monitoring Platforms: Cross-Country Survey Study

Costanza Tortù[1], PhD; Chiara Seghieri[1], PhD, Prof Dr; Ditila Doracaj[2], MD, PhD; Natalya Usheva[3], MD, PhD; Natalia Giménez-Legarre[4,5,6,7], PhD; Yannis Manios[8], MMSc, MPhil, PhD

[1]Institute of Management, L'EMbeDS, Sant'Anna School of Advanced Studies, Piazza Martiri della Libertà 33, Pisa, Italy
[2]Department of Internal Medicine, University of Medicine of Tirana, Tirana, Albania
[3]Department of Social Medicine and Health Care Organization, Medical University of Varna, Varna, Bulgaria
[4]GENUD (Growth, Exercise, Nutrition and Development) Research Group, University of Zaragoza, Zaragoza, Spain
[5]Institute for Health Research Aragón (IIS Aragón), Zaragoza, Spain
[6]Biomedical Research Centre on Obesity Physiopathology and Nutrition (CIBERObn), Carlos III Health Institute, Madrid, Spain
[7]Aragon Agroalimentary Institute (IA2), University of Zaragoza-CITA, Zaragoza, Spain
[8]Department of Nutrition and Dietetics, School of Health Science and Education, Harokopio University, Athens, Greece

Corresponding Author:
Chiara Seghieri, PhD, Prof Dr
Institute of Management, L'EMbeDS, Sant'Anna School of Advanced Studies, Piazza Martiri della Libertà 33, Pisa, Italy

Abstract

Background: The integration of digital health tools into the routines of health care workers (HCWs) holds potential to enhance health care delivery. In particular, digital tools for patient data monitoring allow HCWs to quickly access patient health information and detect early warning signs of potential issues. However, while interest in these tools, such as telemedicine and mobile health, has rapidly grown in recent years, limited research has explored HCWs' attitudes toward digital innovations or their feature preferences.

Objective: This study aims to (1) assess HCWs' attitudes toward digital health tools for patient data monitoring, (2) identify socioeconomic factors influencing these attitudes, (3) determine HCWs' preferences for features of a novel digital health platform for patient data monitoring, and (4) examine whether HCWs' baseline attitudes impact their feature preferences.

Methods: This study uses an integrated approach combining item response theory (IRT) and discrete choice experiment to evaluate the attitudes and preferences of HCWs. Data come from a web conjoint survey distributed to an international cohort of HCWs across the following 4 European countries: Spain, Albania, Bulgaria, and Greece.

Results: Survey respondents comprise 260 HCWs from the 4 countries. The findings indicate that HCWs generally hold a positive attitude toward technological devices (all the IRT coefficients are statistically significant with $P<.05$). Socioeconomic characteristics, including factors such as gender (P=0.05), professional role (P=0.01), and educational background (P=0.01), significantly influence these attitudes. Results show that highly educated female HCWs are those who are mostly inclined to use technologies. In addition, the specific features of a digital health platform for patient data monitoring highly impact HCWs' willingness to incorporate such a tool into their daily practice (all coefficients related to the attributes' effects in the models for the discrete choice experiment results are significant (all P=0.01 except the data looking attribute which has P=0.03) . The findings suggest that an ideal digital health platform for patient data monitoring should offer intuitive graphs, comparative statistics against standards, and include patients' family clinical history. In addition, health workers should receive instructor-led group training to effectively use the platform.

Conclusions: This study shows that health workers generally support the use of digital health tools, which have the potential to improve health care efficiency and patient outcomes through enhanced monitoring and timely interventions. To facilitate adoption, policymakers should strengthen infrastructure, enact supportive legislation, and tailor interventions for groups less inclined to use these tools. Aligning digital health platform features with HCW preferences is crucial, as it directly impacts HCWs' willingness to integrate these tools into daily routines, ultimately benefiting patients. Future research should examine additional factors influencing HCW adoption and address organizational and infrastructural barriers to optimize implementation of digital health platform and improve patient care.

(JMIR Form Res 2025;9:e67142) doi:10.2196/67142

XSL•FO
RenderX

KEYWORDS

health care workers; health technology; discrete choice experiments; preferences; survey study

Introduction

In recent years, there has been a growing interest in the scientific debate on telemedicine [1-3]. The field of telemedicine has become increasingly prominent, as advancements in technology and communication have opened up new possibilities for delivering health care services remotely [4]. This evolving trend has sparked discussions among researchers, health care professionals, and policymakers about the potential benefits and challenges associated with the widespread adoption of telemedicine [5,6]. The term "telemedicine" refers to the adoption of technological tools to provide medical information and services from a distance. This can include online consultations [7,8], remote monitoring of patients [9,10], exchange of medical data through digital means, and the usage of other digital technologies [11]. The convenience and accessibility offered by telemedicine have garnered attention, particularly in addressing issues of health care accessibility in remote areas [4,12,13].

The integration of telemedicine technologies can not only enhance the efficiency of health care delivery but also empower both patients and health care providers by fostering continuous and proactive health care management. By embracing these tools, health care workers (HCWs) can offer personalized and responsive care, improving the overall patient outcomes and contributing to the broader evolution of health care toward a more patient-centered and technologically advanced scheme [14-16]. In particular, the use of technological tools for remote monitoring of patient data allows HCWs to gather real-time insights into a patient's health status, facilitating early detection of potential issues and enabling timely interventions [16].

Following the growing interest in the impact of telemedicine and, particularly, of remote monitoring of patients on the quality of health care, there are several contributions that explore HCW's perspectives about these new tools [17-22]. However, detailed information on the HCW preferences on how to read and monitor patients' data in a technological platform is still scarce, being mostly gathered from the patient's perspective [23,24]. Understanding the attitudes of health workers toward technology, particularly their preferences regarding the usage of digital health, such as mobile health (mHealth) to support health care for reading and monitoring patients' data and feedback, is of paramount importance. Understanding HCWs' receptiveness toward these tools and their preferences in terms of user interfaces, data visualization, and ease of integration into existing workflows can significantly impact the efficiency and accuracy of patient data interpretation. This basically ensures that technological solutions align seamlessly with their needs, facilitating a smoother transition to telemedicine. On the other hand, identifying potential concerns or barriers HCWs may have regarding the use of digital health tools for patient data monitoring is essential for targeted training and support [21,25].

In summary, exploring health workers' attitudes toward digital health tools for reading and monitoring patients' data is a pivotal step in enhancing the overall effectiveness of telemedicine and in encouraging a technologically empowered health care system. Existing contributions in the field of social research on HCWs' attitudes and preferences toward digital health tools mainly focus on gathering their opinions based on specific experiences with these tools by using different methodologies.

Systematic reviews in this theme have pointed out that there are technical, social, and organizational factors that may influence the adoption of digital health applications by the health care staff [19,20]. About the technical factors, some existing works have explored the digital health literacy of these professionals [26], while others have collected and evaluated their general attitude and opinions on using such digital health technologies [19-21,26-29]. These analyses encompass the social and organizational elements of HCWs' preferences and perspectives about digital health tools by either using qualitative methodology [18,19,30,31], mixed methods approaches [23,32], quantitative methodologies [33-36] or discrete choice experiments (DCEs) [37-40], which implicitly elicit health workers' barriers and facilitators toward digital health technologies.

However, recent contributions state that there remains considerable room for improvement in encouraging health workers to embrace such technologies and enhancing the overall effectiveness of telemedicine even in high-income countries [18,20,22,41]. In particular, the ongoing digitalization of health care causes a change in the traditional role of the health care worker, which may come with specific barriers and challenges [18]. Moreover, it would be important to jointly consider these aspects and to assess whether a different attitude toward the mHealth tools impacts the preferences on the usage of such tools. This integrated approach provides a deeper understanding and enhances the capacity to develop tailored interventions that succeed across diverse contexts, ultimately contributing to the optimization of telemedicine practices globally.

This study proposes an integrated approach aimed at evaluating the receptivity of HCWs toward adopting digital health technologies, while also identifying potential barriers or facilitators that may influence their willingness to use such devices for patient data monitoring. The investigation is planned throughout the EU-funded DigiCare4You project, an innovative intersectoral initiative using digital tools for the prevention and management of type 2 diabetes and hypertension. The project spans across 2 high-income European countries, Spain and Greece, as well as middle-income countries, Bulgaria and Albania.

The digital tools developed within the project are designed for both patients with chronic diseases and HCWs. An mHealth system has been developed consisting of a novel mobile app for patients in which they are encouraged to input data related to their clinical parameters, dietary habits, and exercise routines, and a digital health platform that HCWs are asked to introduce

XSL•FO
RenderX

in their workday routine to monitor the data of their patients. Before implementing these digital tools, the project partners conducted a comprehensive analysis to gain insights into the attitudes of both patients and HCWs regarding the use of digital technologies and to understand their preferences about these tools, thus laying the groundwork for the successful implementation and acceptance of the proposed digital solutions.

To explore attitudes and preferences of HCWs toward the platform for patient data monitoring, the project partners planned, developed, and disseminated a Web Conjoint Survey (WCS) across all 4 implementation countries. The WCS comprises three key sections: (1) capturing basic sociodemographic characteristics, (2) delving into the HCWs' baseline attitude toward technologies, and (3) using a DCE [42] to elicit their preferences about a novel digital health platform to be introduced in their workday routine for patients' data monitoring.

In particular, the second section, which explores the HCWs' attitude toward technology, includes some statements with respect to which respondents are asked to declare their level of agreement. These statements describe behaviors or beliefs pertaining to the use of technological devices in their daily work routines and to the integration of technologies in the management of lifestyle habits. The responses from HCWs to these statements are analyzed using item response theory (IRT) models [43,44]. These models enable the estimation of their overall latent propensities to incorporate technological devices into their working activities and into the management of lifestyle habits.

In the third section, HCWs undergo a DCE, designed to elicit their preferences about different specific characteristics of a digital health platform for patients' data monitoring. They are asked to choose their preferred option in a series of pairwise competing scenarios, each comparing 2 platforms for patients' data monitoring with different attributes. The attributes considered as potential obstacles or facilitators in platform usage are derived from a thorough literature review and targeted focus groups implemented in the 4 implementation countries involving health care professionals and health care policymakers. The selection of competing scenarios is randomized using an appropriate randomization design.

In analyzing the DCE results using conditional regression models, we also include as a covariate in the regression the information about the predicted latent traits toward the usage of technologies (the estimated attitude toward technology) resulting from the IRT analysis. This integrated analytical approach combining IRT and DCE significantly enhances our understanding of HCWs' inclinations toward the usage of technological tools and provides valuable insights into their preferences on digital tools for patient data monitoring. Furthermore, this method enables an assessment of whether an interaction exists between HCWs' baseline attitudes toward technology and their preferences for the platform.

Given that the analysis spans an international cohort of HCWs situated in multiple European countries, the results hold the potential for generalization across diverse contexts.

Methods

Overview

This study analyzes the data coming from a web conjoint survey administered to a sample of HCWs in Europe. The web survey was implemented through the online software platform Qualtrics Experience Management. The survey was launched in April 2022 and ended in November 2023. To get access to the survey, the respondent required only an anonymous survey link, which could be easily shared via email, social networks, QR codes, or private messages, ensuring a user-friendly and widespread dissemination. Within the DigiCare4You Project, local partners from the 4 implementation countries were in charge of distributing the survey link to HWs within their network, and those who received the link were asked to further circulate it among their peers. By actively encouraging participants to share the survey link within their working networks, we seek to foster a widespread reach, initiating a dynamic snowball sampling mechanism [45]. While acknowledging that the resultant sample is nonprobabilistic, this strategic approach plays a pivotal role in attracting a diverse range of respondents and enriching the data with a multitude of perspectives.

The WCS is structured across three main sections: (1) the first section includes demographic details of respondents, encompassing age, employment, and years of education; (2) the second section evaluates the baseline individual attitude toward technology. In particular, the goal is to understand the propensity of the respondents to use technological devices in their working life and in the managing of their lifestyle habits, by looking at their level of agreement with some statements; (3) the third section provides the DCE experiment to elicit the HCWs preferences on the novel platform to monitor the health status of their patients. The survey was implemented in 5 languages, which include the 4 official languages in the 4 implementation countries of the DigiCare4You project plus the English language. The survey was completely anonymous and does not collect personally identifiable information. The complete survey (in English) and the CHERRIES (Checklist for Reporting Results of Internet E-Surveys) checklist are available as Multimedia Appendix 1 and Checklist 1, respectively.

Item Response Theory

Considering previous evidence pointing out that the usage of technologies in working activities and in daily life may influence the adoption of digital tools by the health workers [20], the second section of the survey explores the HCWs' baseline attitudes toward technology. Participants were asked to indicate, on a scale ranging from 1-10, their level of agreement with 10 statements. These statements were designed to capture attitudes and beliefs related to the use of technological devices, with 5 statements focusing on working activities and the remaining 5 statements on the integration of technology into lifestyle management. In this scale, a rating of 1 signifies "totally disagree," while a rating of 10 indicates "totally agree." The list of statements is presented in Table 1.

XSL•FO
RenderX

Table . Survey items regarding the health care workers' attitude toward using technology both in working activities and in managing lifestyle habits.

Item	Statement
Usage of technologies in working activities	
Work 1	"I think that using health technologies (e.g electronic health records, mHealth, telehealth) in my job would enable me to accomplish tasks more quickly."
Work 2	"I think that using health technologies (e.g electronic health records, mHealth, telehealth) would make it easier to do my job."
Work 3	"I would find health technologies (e.g electronic health records, mHealth, telehealth) useful in my job."
Work 4	"I think that using health technologies (e.g electronic health records, mHealth, telehealth) would improve the quality of the work I do."
Work 5	"I think that using health technologies (e.g electronic health records, mHealth, telehealth) in my job would increase my productivity."
Usage of technologies in managing lifestyle habits	
Habits1	"I used to / I would like to track some health parameters (e.g hearth rate, Oxygen saturation, quality of sleep) through a technological device."
Habits 2	"I like to track my lifestyle habits through a mobile app."
Habits 3	"I like to monitor what i eat and drink through a mobile app."
Habits 4	"I like to track the statistics regarding some activity parameters (steps taken, stairs climbed, minutes of workout performed) through a technological device."
Habits 5	"I like to write on a virtual notebook some notes about what I do in a day and how I feel."

Using their responses to such statements, we intend to quantify the HCWs' baseline attitude toward technology. In particular, we intend to estimate their latent attitude in using technological tools for working activities and lifestyle habits using IRT models. These models provide a robust framework for estimating an individual's latent trait, based on their responses to a set of items [44]. This methodology offers valuable insights into abilities or latent characteristics by considering individual responses across a range of assessment items. IRTs are widely used in educational [46,47], psychological [48], and clinical [49,50] measurement. Using an IRT model, we can estimate, through the maximum likelihood approach, the values assumed by the latent trait of each respondent.

In this study, our primary goal was to assess the attitudes of HCWs toward technology usage, distinguishing between their perspectives on professional tasks and leisure pursuits. To achieve this, we analyzed their responses to distinct sets of items tailored to each aspect, with each item rated on a scale from 1 to 10. To accurately capture and interpret these ordinal responses, we used a specialized form of IRT known as the Graded Response Model (GRM) [51,52]. This model is well suited for assessing attitudes expressed through ordered categories, providing a nuanced understanding of the varying degrees of agreement or disagreement with the given statements related to technology use. In this setting, we estimated 2 separate GRMs to predict the individual attitude of health workers to use technological tools in their working activities and in the management of their lifestyle habits. We denoted these predicted traits as $\hat{\theta}_W$ and $\hat{\theta}_H$, respectively. In addition, to enhance the robustness of our analysis, we also estimated the overall latent HCWs' attitude toward technology $\hat{\theta}_{WH}$ by applying GRM to all items pertaining to statements on technology usage in both professional tasks and daily routine. Before fitting GRMs, we noted that there were a few missing entries in the responses to the items. Since, upon exploring the data, we did not observe relevant patterns that rule the missingness in the items, we can impute them through Predictive Mean Matching (PMM) [53,54]. All the descriptives related to the missing data in the items and the performance of the PMM algorithm in imputing missing entries are available in the Multimedia Appendix 2. Finally, after estimating the latent traits, we identified the determinants influencing these traits by means of linear regression models with the latent traits $\hat{\theta}_W$, $\hat{\theta}_H$ and $\hat{\theta}_{WH}$ as the outcome variables and individual health worker's characteristics as regressors. The independent variables in these models encompassed the health worker's gender, the country of residence, the binary variable less than 30 (indicating whether the health worker is less than 30 years old), the binary variable high education (equal to 1 if the health worker has more than 14 years of education), the variable electronic health record (indicating the health worker's experience with electronic health records), and the binary variable doctor (taking the value 1 for general practitioners and specialized doctors and 0 for the other health workers' employment positions).

Discrete Choice Experiments

The third section of the survey provided a DCE [42] to assess the HCW's preferences on the digital health platform they might use to monitor the health status of their patients. DCEs are widely used in several fields of social sciences [55,56] and health care applications [42] as they are flexible and easy to

XSL•FO
RenderX

interpret. They enable researchers to elicit the user preferences on a specific issue without explicitly asking them, as respondents face some pairwise comparisons and choose their preferred option between 2 competing scenarios, whose characteristics (attributes) provide different options (levels).

In this setting, HCWs indicated the preferred platform for patient data monitoring between 2 competing platforms that differed in the characteristics of some of their aspects. These aspects have been identified based on a literature review of studies using DCE in similar contexts. In addition, insights were gathered through focused group discussions involving health workers and health care policymakers across the 4 implementation countries (details about the participants and the output of the focus groups are available in the Multimedia Appendix 3).

The key aspects considered for the web platform are as follows:

1. Data visualization: Determines whether health workers prefer data presented through tables or graphs;
2. Statistical content: Focuses on the type of statistics preferred by health workers (patient-specific statistics or statistics comparing patients to a standard, such as a national average or app users' average);
3. Additional content: Explores the extra data health workers desire in the platform beyond clinical parameters and lifestyle habits (eg, patients' family clinical history or emotional status)
4. Platform usage training: Evaluates the preferred type of training for using the platform (group training with an instructor or self-training with tutorials).

Table 2 describes the attributes and levels that have been used in the DCE. In the proposed DCE, respondents are asked to pick their preferred platform for patient data monitoring between two competing platforms, whose main characteristics –summarized by the attributes- provide different options –represented by the levels. The DCE data were analyzed using alternative-specific conditional logit regression models, in which the outcome was HCW's preference (ie, platform A or platform B).

Table . Attributes and levels of the Discrete Choice Experiment.

Attributes	Level 1	Level 2
Data looking	Tables	Graphs
Statistical content	Statistics of my patients	Statistics of a given patient against a standard
Additional contents	Patient's family clinical history	Patients' emotional status
Training	Group training with instructor	Self-training with a tutorial

Ethical Considerations

The study was carried out in compliance and accordance with the General Data Protection Regulation (2016/679) and the Italian Legislative Decree No. 196/2003 ("Personal Data Protection Code ").

The data used in the study were collected from a web survey, which is completely anonymous and does not collect personally identifiable information. In particular, a link to a survey was sent to individual email addresses or through a QR code. Responses were submitted directly to a survey software package (not returned via email). The survey included a disclosure letter describing the project and the purpose of the survey, the section of the questionnaire, the length of time of the survey, and that participation and consent were totally voluntary. The questionnaire did not gather any personally identifiable information for any purpose, or a combination of identifiers that might make it more likely to identify an individual. The individual respondents' responses and data could not be linked to their email. No incentives were offered.

Since the data do not interfere with the rights and freedoms of the respondents, an evaluation by the Ethics Committee is not necessary. According to the Declaration of Helsinki on "ethical principles for medical research involving human participants" (Preamble, Art. 1), such evaluation is required only when the medical research involves humanly identifiable material or data. In addition, the DigiCare4You project received the ethical approval number 34-19/11/2021 from the Harokopio University Research Ethics Committee.

Results

Overview

In this section, we present some preliminary descriptive evidence based on our dataset, (additional descriptive results are available in the Multimedia Appendix 4). The final dataset comprises responses from 260 HCWs situated in the 4 implementation countries of the DigiCare4You project; 102 respondents from Albania, 62 from Bulgaria, 68 from Greece, and 28 respondents from Spain. As more extensively discussed in the Multimedia Appendix 2, data notably exhibit some missing entries, as respondents had the option to skip questions they chose not to answer. Specifically, 3 respondents abstained from answering over half of the available questions; hence, we made the decision to exclude them from the analysis.

The distribution of sociodemographic characteristics across the 4 implementation countries is depicted in Table 3.

The reports include both the overall distribution of the variables and the stratified distribution by countries. Reported data include both absolute and relative frequencies, expressed in percentages. Upon examination of the table, it becomes evident that the majority of respondents were female (204/260, 78%), relatively young (171/260, 66% aged between 18 and 44 years old), and highly educated, with 200/260 respondents reporting having more than 16 years of education. Notably, 116/260 (45%)

XSL•FO
RenderX

respondents were employed as medical specialists, followed by 62/260 (24%) respondents employed as general practitioners. Differences emerge when the 4 countries are compared. For instance, Albanian HCWs tend to be younger compared to their counterparts, while respondents from Spain exhibit the highest percentage of higher-educated respondents (26/28, 93% of the respondents declared having more than 15 years of education). Greece exhibits the highest percentage of respondents employed as dietitians or nutritionists (19/68, 28%) and health visitors (13/68, 18%)—qualified nurses who have additional training in public health and who offer support to families with young children. Albania has the highest percentage of respondents employed as general practitioners (40/102, 39%), followed by Spain and Bulgaria (7/28, 25% and 15/62, 24%, respectively).

Table . Sample sociodemographic characteristics overall and by country.

Characteristics	Albania	Bulgaria	Greece	Spain	All countries
Respondents, n (%)	102 (39)	62 (24)	68 (26)	28 (11)	260 (100)
Gender, n (%)					
Women	81 (79)	46 (74)	53 (78)	24 (86)	204 (78)
Men	20 (20)	15 (24)	15 (22)	4 (14)	54 (21)
Prefer not to answer	1 (1)	1 (2)	0 (0)	0 (0)	2 (1)
Age groups, n (%)					
18-29	52 (51)	15 (24)	23 (34)	4 (14)	94 (36)
30-44	30 (29)	10 (16)	26 (38)	11 (39)	77 (30)
45-59	19 (19)	30 (48)	15 (22)	11 (39)	75 (29)
60-75	1 (1)	7 (11)	4 (6)	2 (7)	14 (5)
Years of education, n (%)					
10-12	8 (8)	7 (11)	3 (4)	1 (4)	19 (7)
13-14	1 (1)	1 (2)	3 (4)	0 (0)	5 (2)
15-16	12 (12)	3 (5)	13 (19)	1 (4)	29 (11)
>16	81 (79)	46 (74)	47 (69)	26 (93)	200 (77)
Prefer not to answer	0 (0)	5 (8)	2 (2)	0 (0)	7 (3)
Employment, n (%)					
Dietitian or nutritionist	0 (0)	0 (0)	19 (28)	6 (21)	25 (10)
Health visitor	2 (2)	0 (0)	13 (19)	0 (0)	15 (6)
Medical specialist	52 (51)	23 (37)	30 (44)	11 (39)	116 (45)
General practitioner	40 (39)	15 (24)	0 (0)	7 (25)	62 (24)
Nurse	0 (0)	1 (2)	3 (4)	4 (14)	8 (3)
Other health allied professions	1 (1)	11 (18)	2 (3)	0 (0)	5 (14)
Prefer not to answer	7 (1)	12 (19)	1 (1)	0 (0)	20 (14)

IRT Outcomes

As we have previously hinted throughout the paper, we used the information about the level of assessment of the 10 statements related to the HCWs' attitude toward technology to predict their latent attitude in using such technologies. Specifically, we separately used the items related to the usage of digital technologies in working activities and to the usage of technological devices in the management of daily habits. The IRT models fit the data well, as all parameters resulted in being statistically significant, and all items contributed to explaining the latent traits. The complete estimation and fitting graphs are available in the Multimedia Appendix 5.

Once these traits were estimated, we explored whether socioeconomic factors have a role in determining their variability. Figure 1 graphically reports the main results of the 3 linear regression results in which the outcome variables were θw^θH^ and θ^WH the latent attitudes of HCW toward using the technology during working activities, daily habits, and both, respectively. Complete tables are available in the Multimedia Appendix 5. In particular, Figure 1 provides an overview of the estimated coefficients in the 3 models, together with their corresponding 95% CIs. Results suggest that (1) males exhibit a less pronounced attitude in using technologies compared to women, and this effect is particularly significant when estimating the latent trait targeted at working activities θ^W and the comprehensive latent trait θ^WH; (2) HCWs in Spain and Bulgaria demonstrate a lower propensity to use technologies

XSL•FO
RenderX

compared to their Albanian counterparts; (3) being highly educated leads to a higher inclination to use technologies; and (4) doctors exhibit a lower propensity to use technologies compared to HCWs referred to other employment categories.

Figure 1. The determinants of the individual attitude toward using technology in the working activities, daily habits, and both. The figure reports estimates regarding the linear regression coefficient, together with their corresponding 95% CIs, in the 3 different models. EHR: electronic health record.

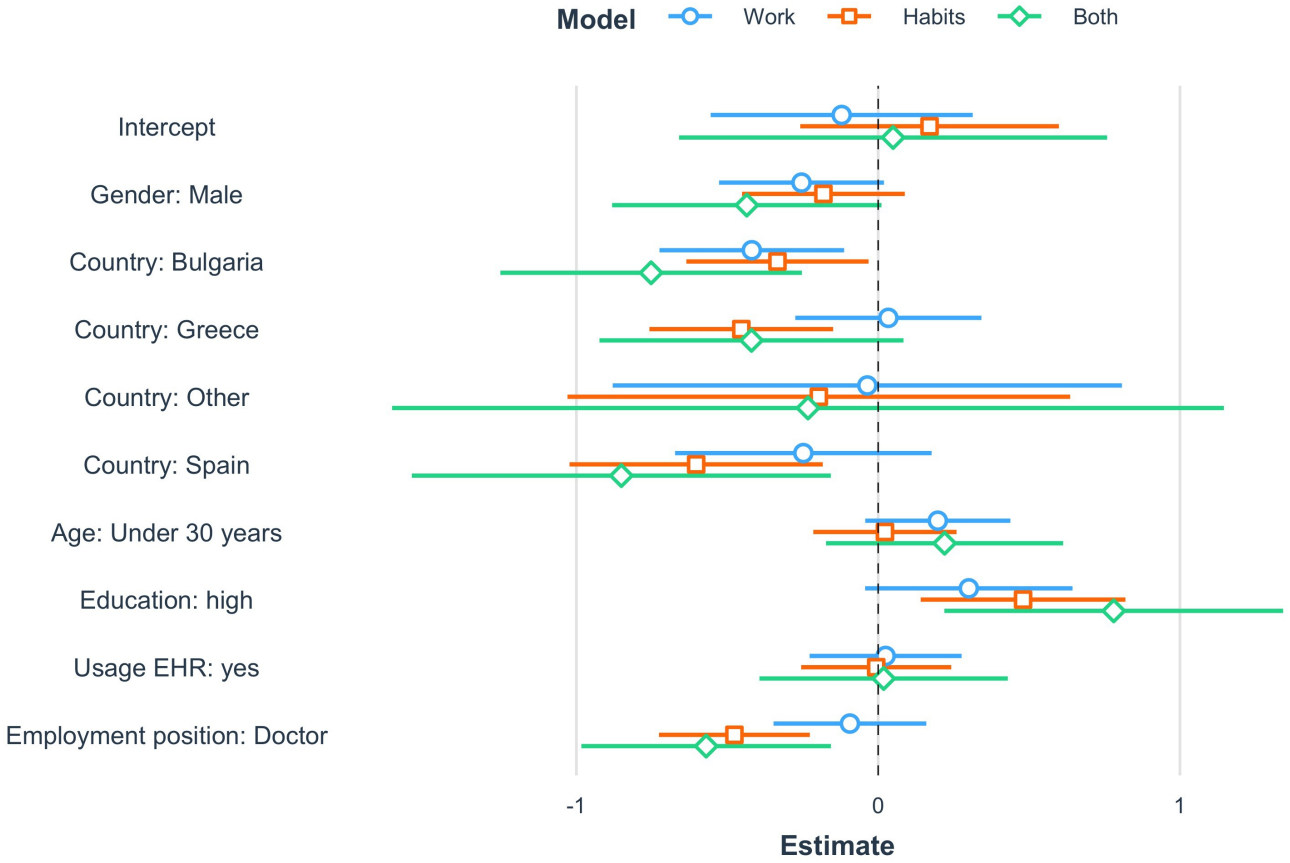

DCE Outcomes

In this section, we reported the DCE results regarding the features that represent potential barriers or facilitators that may encourage or discourage HCWs in integrating a given digital health platform for patients' data monitoring in their working routine. Moreover, we investigate whether their attitude toward the usage of technologies in working activities and in the management of their habits shapes their preferences about the ideal mHealth platform for patients' data monitoring.

In the DCE used in this setting, HCWs faced 6 pairwise comparisons (named Choice Sets), each presenting 2 alternative characterizations of the platform, which differed in the levels of their attributes. Figure 2 reports the main results of the DCE analysis, obtained by applying conditional logit models [57], where the response variable is the choice variable, and the regressors are the 4 attributes of the alternatives.

XSL•FO
RenderX

Figure 2. Discrete choice experiment results: odds ratios and corresponding 95% CIs of 3 conditional logit models.

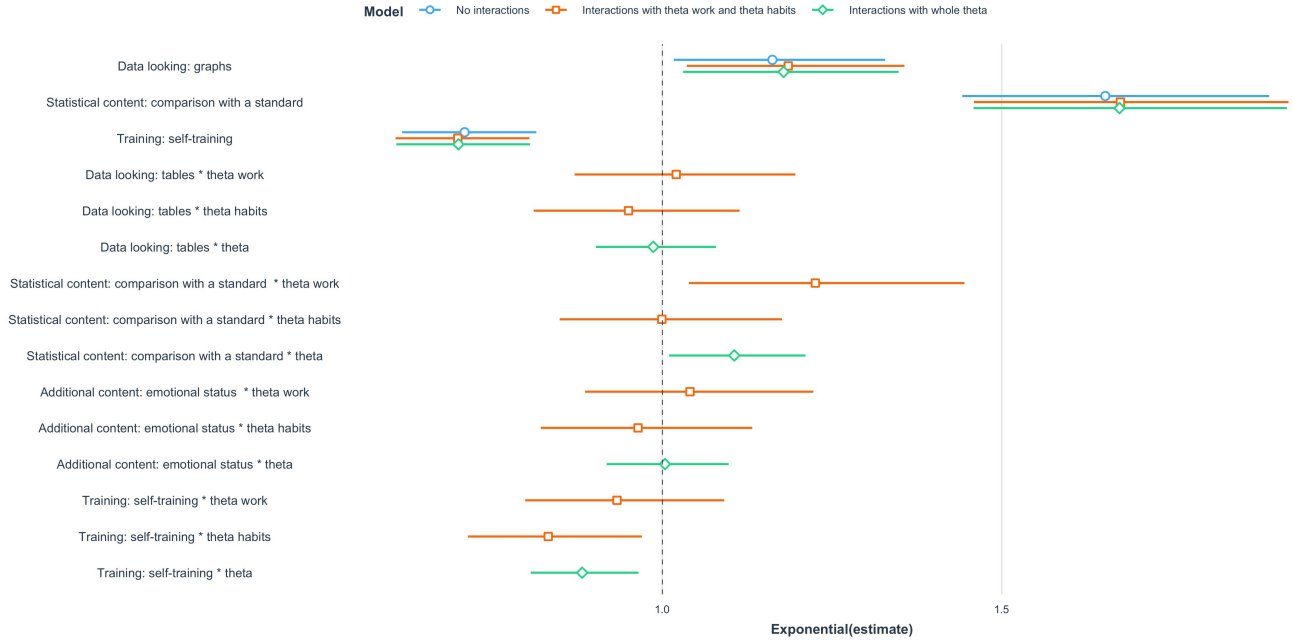

Specifically, to analyze data from the DCE, we run three conditional logit models. Figure 2 reports the estimated odds ratios, together with their corresponding 95% CIs of each model: (1) the first one (blue lines) provides the inclusion as independent variables of the 4 attributes only, (2) the second one (orange lines) also includes pairwise interactions between the 4 attributes and the 2 estimated latent traits θ^W and θ^H, and (3) the third one (green lines) still includes interactions but with respect to the comprehensive latent trait θ^WH. Results here are presented in terms of odds ratios for ease of interpretation (odds ratios greater than 1 signal that the presence of a given feature encourages the respondents to choose that alternative). The coefficients related to the attributes that affect the HCWs' choice of the ideal platform for patient data monitoring are highly significant.

By looking at the figure, we can definitely state that HCWs (1) prefer to analyze data through graphs rather than through tables, (2) find it very useful to have comparative statistics that relate data of their patients to a standard, rather than simply observing data of their patients, (3) prefer to have information on the patients' clinical history rather than on their emotional status, and (4) prefer to receive an active group training on the platform rather than completing self-tutorials.

By looking at the interactions, we can also state that (1) having a higher attitude in using technologies in working activities (θ^W) and having a high comprehensive attitude toward technology (θ^WH) significantly increases the preference of having statistical data about the comparison of patients against a standard, instead of having data related to the given patient only; and (2) an increased attitude toward the usage of technological devices in managing lifestyle habits (θ^H) and toward technologies as a whole (θ^WH) decreases the preference of having a self-training program to learn the platform, instead of a group training.

As a consequence, the ideal digital health platform for patient data monitoring should provide intuitive graphs, statistics that compare patients' data against a standard, information about the patients' family clinical history, and health workers should be trained for the usage of that platform through a group training led by an instructor. The characteristics of the ideal platform chosen by HCW for patient data monitoring are represented in Figure 3.

XSL•FO
RenderX

Figure 3. Discrete choice experiment results: ideal characteristics of the web platform.

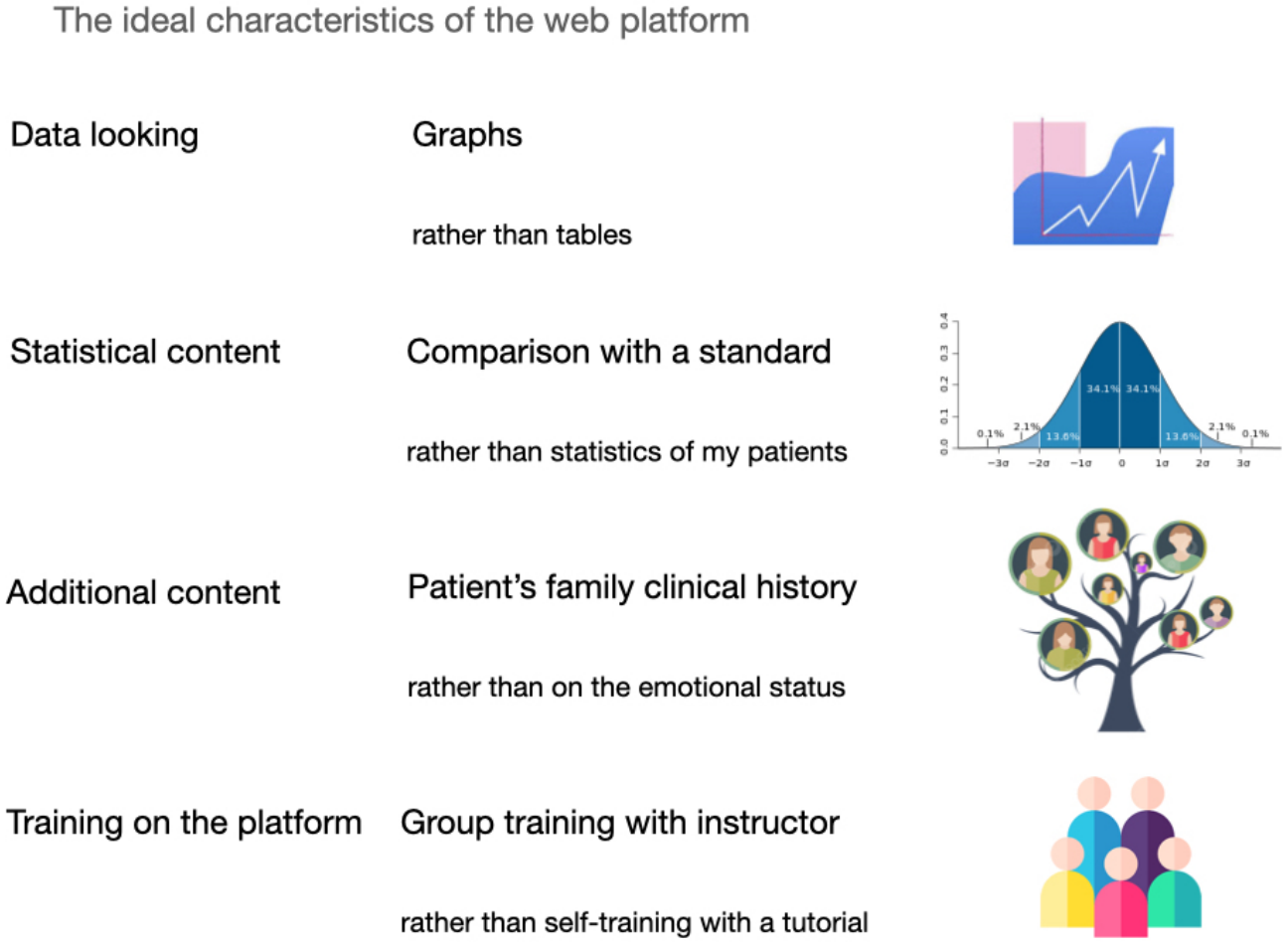

Discussion

Principal Findings

This contribution proposes an integrated approach based on IRT and DCE (1) to evaluate the HCWs' attitude toward the usage of technological devices in their working activities and in managing their lifestyle habits, (2) to explore the role of sociodemographic characteristics in shaping these attitudes, and (3) to appraise whether their attitude toward technology shapes their preferences about a novel digital health platform for patient data monitoring. First, results suggest that health workers show an overall positive attitude toward the usage of technological devices both in their working activities and in their lifestyle habits. This propensity is particularly strong for women and highly educated professionals, while it appears to be weaker for doctors compared to other health care professionals. Similarly to the study by Muiruri et al [17], the employment position plays a role in determining the attitude toward technology.

The different country of the respondents also affects the attitude toward technology, as Albanian HCWs have a stronger propensity to integrate technological tools within their daily routines compared to the other countries. These findings are in line with the review done by Borges et al [26], who identified that country-specific technologies can affect the adoption of digital health devices. Thus, statistically different results may be hypothesized as the respondents being a proxy of their countries' readiness to adopt these technologies.

In some studies [17], sociodemographic variables, such as gender and education, do not play a role in determining the HCWs' willingness to adopt technologies, while in our contribution, they result in significantly impacting the HCWs' attitude toward technology, as already pointed out in Borges do Nascimento et al [26]. Second, the DCE results suggest that the type of training, the way data are presented, and the information provided by the platform play a pivotal role in shaping the HCWs' willingness to adopt a digital platform for patients' data monitoring. The inclination for personal training instead of self-guided tutorials was already pointed out before [4,19,20], besides the findings of [21,26], indicating a concern with the quality of the training received by the staff.

The preferences regarding data looking in the format of graphs and statistics compared to a standard hint to preferences aligned with more intuitive and user-friendly interfaces, which are also previously mentioned in other studies [4,14,20,21] as facilitators of the adoption of digital health tools by the HCW.

To the best of our knowledge, there are no existing studies that jointly explore the HCWs' attitude toward technology and their preferences about such digital technologies. Our findings suggest that there are relevant interactions between these 2 aforementioned topics, and this issue should be further explored, as it provides a global understanding of the approach of HCWs toward digital health systems.

XSL•FO
RenderX

Strengths and Limitations

This work has some limitations. The main limitation is that the sample of individuals asked to complete the WCS might be affected by a selection bias due to the snowball sampling nonprobabilistic design. The sampling technique was chosen based on the availability of resources in the 4 countries under study. An additional limitation of the study is that the sample size in each of the 4 countries, even if it is relatively large, does not allow us to conduct a stratified analysis, where we separately implement the integrated analytical framework in each country and then we inspect differences among countries. The relatively low response rates and participation, especially in some countries like Spain, can be due to people's fatigue with invitations and participation in various surveys (oversurveying), especially after the COVID-19 period of dominance of e-communication. However, this work also offers valuable contributions. First, it proposes an integrated analytical approach based on IRT and DCE to simultaneously investigate the HCWs' attitude toward the usage of technologies, the HCWs' preferences about potential digital tools to be integrated within their working routine, and potential interactions between the aforementioned topics. This approach empowers us to delve into the intricate dynamics between HCWs' fundamental attitudes toward technology and their specific inclinations toward the novel platform. This exploration adds a layer of depth to our understanding, allowing for a more comprehensive assessment of the factors influencing their preferences. Second, this study covers a wide international cohort of HCWs enrolled in multiple European countries. Therefore, it does not focus on specific experiences that refer to empirical contexts that are likely to be narrow both from the spatial and the temporal perspective but attempts to obtain findings that can be potentially generalized.

Conclusions

The analytical framework presented in this work, merging IRT and DCE, elevates our comprehension of HCWs' propensities regarding the integration of digital health tools. The breadth of the analysis, encompassing a wide international cohort of HCWs, places the results as not just informative but also highly applicable across a spectrum of contexts. This potential generalizability underscores the significance of the findings, making them a relevant resource for shaping strategies and targeting interventions in various health care settings.

The findings from this study have significant policy implications for the integration of digital systems in health care systems. First, the study reveals that HCWs generally hold a positive attitude toward the use of technological devices, emphasizing the potential benefits of these tools in enhancing health care efficiency. It follows that political authorities should push for the progressive integration of digital health systems, integrating patient-generated data into the daily workflows of health care workers to enhance care delivery by improving the available infrastructures through which digital tools are implemented and promoting laws and programs to encourage the gradual development of a technologically empowered health care system. Second, the research highlights the influence of socioeconomic factors in shaping health workers' attitudes toward digital health tools. Policymakers should take into account these factors when designing strategies for the adoption of such technologies, targeting specific support to those categories that have a weaker propensity for adopting these technologies. Furthermore, the study underscores the importance of considering specific characteristics of platforms for patients' data monitoring, as these features significantly impact HCWs' willingness to integrate them into their working routines. Despite the relevance of our results, the attitude of HCWs toward digital tools and their preferences toward the ideal characteristics of these tools should be further explored in future studies to enhance the global understanding of this topic and to provide even stronger evidence for policymakers. Future works may consider wide international samples. In addition, researchers should deeply understand which are the organizational and infrastructural barriers that still limit the development of digital health systems and which are the possible approaches to enhance the HCWs' attitude toward these technologies.

Acknowledgments

This study is conducted within the DigiCare4You Project, which has received funding from the European Union's Horizon 2020 research and innovation program (grant 945246). The content of this paper reflects only the authors' views, and the European Union is not liable for any use that may be made of the information contained therein. We thank the collaborators of the DigiCare4You Consortium: Coordinator: Yannis Manios, Harokopio University (Greece); Steering Committee: Yannis Manios, Chiara Seghieri, Djordje Djokic, Konstantinos Votis, Konstantinos Makrilakis, Lieven Annemans/Ruben Willems, Rajesh Vedanthan, Sabine Dupont; Baker Heart and Diabetes Institute (Australia): Brian Oldenburg; Center for Research & Technology Hellas (Greece): Konstantinos Votis, Andreas Triantafyllidis, Sofia Segkouli, Anastasios Alexiadis, Aggeliki Spyrou, Panagiotis Petridis, Katerina Skoumpourdi; Ghent University (Belgium): Lieven Annemans, Ruben Willems, Delfien Gryspeerdt; Harokopio University (Greece): Yannis Manios, Eva Karaglani, Costas Anastasiou, Konstantinos Tserpes, Christos Diou, Constantine Tsigos, Christina Mavrogianni, Agathi Ntzouvani, Katerina Kontochristopoulou, Rafaela Makri, Matzourana Argyropoulou, Amalia Filippou, Maria Chrysini, Katerina Vervesou, Renos Roussos, Niovi Chrysanthakopoulou, Thalia Lampidi, Irene Rifioti, Olga Voureka, Renia Terzaki, Christina Pelekanou, Panayiota Kostarellou, Christina Patmiou; International Diabetes Federation European Region (Belgium): Sabine Dupont, Elizabeth Dupont, Izabell Thunström; International Foundation for Integrated Care (UK): Alejandro Gil-Salmerón, Niamh Daly Day; La Trobe University (Australia): George Moschonis, George Siopis; Meteda (Italy): Sandro Girolami, Giacomo Vespasiani, Teresa Almonti, Nazzareno Pierantozzi, Eric Notts, Giorgia Stoppaccioli, Claudia D'Antonio; Monash University (Australia): Helen Skouteris, Melissa Savaglio; National and Kapodistrian University of Athens (Greece): Konstantinos Makrilakis, George Stergiou, Stavros Liatis, George Karamanakos, Chrysi Koliaki, Anastasios Kollias,

XSL•FO
RenderX

Eva Zikou, Haris Dimosthenopoulos; New York University Grossman School of Medicine (USA): Rajesh Vedanthan, Keng-Yen Huang, Samrachana Adhikari, Jessica Gjonaj; Privanova (France): Farhan Sahito, Dusan Pavlovic, Djordje Djokic, Arzoo Sahito, Gisella Battalova; Sant'Anna School of Advanced Studies (Italy): Chiara Seghieri, Costanza Tortu, Tallys Feldens; Medical University of Varna (Bulgaria): Violeta Iotova, Natalya Usheva, Tanya Zlateva, Kaloyan Tzochev, Yoto Yotov, Rositsa Chamova, Anna Kozhuharova, Vesela Todorova, Alexander Mitev, Ralitsa Marinova, Donika Kandilarova, Sonya Nedelcheva, Petya Boncheva, Martin Mirchev; Universidad De Zaragoza (Spain): Luis Moreno Aznar, Rosa Magallón Botaya, Gloria Bueno Lozano, Natalia Giménez-Legarre, Ivie Maneschy, María L Miguel-Berges, Paloma Flores-Barrantes, Esther Ma Gonzalez-Gil, Pilar De Miguel-Etayo, Bárbara Oliván Blázquez, Susana Pérez; and University of Medicine, Tirana (Albania): Florian Toti, Ditila Doracaj, Ergita Nelaj, Skerdi Prifti, Luftime Bruka, Adriana Lapardhaja, Ornela Laze.

Data Availability

The survey data generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

CT contributed to literature search, data collection, data analysis, writing – original draft, data interpretation, conceptualization, validation. CS was involved in data collection, data analysis, data interpretation, conceptualization, review & editing, validation, and supervision. NU handled data collection, validation, review & editing. DD managed data collection, validation, review & editing. NGL was responsible for data collection, validation, review & editing. YM contributed to data collection, validation, review & editing, supervision, and project administration.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The Web Conjoint Survey.
[DOCX File, 18 KB - formative_v9i1e67142_app1.docx]

Multimedia Appendix 2
Analysis of missing data.
[DOCX File, 4466 KB - formative_v9i1e67142_app2.docx]

Multimedia Appendix 3
Focus groups.
[DOCX File, 25 KB - formative_v9i1e67142_app3.docx]

Multimedia Appendix 4
Additional descriptives.
[DOCX File, 1300 KB - formative_v9i1e67142_app4.docx]

Multimedia Appendix 5
Item response theory.
[DOCX File, 6215 KB - formative_v9i1e67142_app5.docx]

Checklist 1
CHERRIES checklist.
[PDF File, 146 KB - formative_v9i1e67142_app6.pdf]

References

1. AlQudah AA, Al-Emran M, Shaalan K. Technology acceptance in healthcare: a systematic review. Appl Sci (Basel) 2021;11(22):10537. [doi: 10.3390/app112210537]
2. Kario K, Harada N, Okura A. Digital therapeutics in hypertension: evidence and perspectives. Hypertension 2022 Oct;79(10):2148-2158. [doi: 10.1161/HYPERTENSIONAHA.122.19414] [Medline: 35726619]
3. Hengst TM, Lechner L, Dohmen D, Bolman CA. The facilitators and barriers of mHealth adoption and use among people with a low socio-economic position: a scoping review. Digit Health 2023;9:20552076231198702. [doi: 10.1177/20552076231198702] [Medline: 37691766]

XSL•FO
RenderX

4. Neves AL, Burgers J. Digital technologies in primary care: implications for patient care and future research. Eur J Gen Pract 2022 Dec;28(1):203-208. [doi: 10.1080/13814788.2022.2052041] [Medline: 35815445]
5. Manteghinejad A, Javanmard SH. Challenges and opportunities of digital health in a post-COVID19 world. J Res Med Sci 2021;26(1):11. [doi: 10.4103/jrms.JRMS_1255_20] [Medline: 34084190]
6. Li X, Huang L, Zhang H, Liang Z. Enabling telemedicine from the system-level perspective: scoping review. J Med Internet Res 2025 Mar 5;27:e65932. [doi: 10.2196/65932] [Medline: 40053725]
7. Quinn LM, Davies MJ, Hadjiconstantinou M. Virtual consultations and the role of technology during the COVID-19 pandemic for people with type 2 diabetes: the UK perspective. J Med Internet Res 2020 Aug 28;22(8):e21609. [doi: 10.2196/21609] [Medline: 32716898]
8. Campbell K, Greenfield G, Li E, et al. The impact of virtual consultations on the quality of primary care: systematic review. J Med Internet Res 2023 Aug 30;25:e48920. [doi: 10.2196/48920] [Medline: 37647117]
9. Alvarez P, Sianis A, Brown J, Ali A, Briasoulis A. Chronic disease management in heart failure: focus on telemedicine and remote monitoring. Rev Cardiovasc Med 2021 Jun 30;22(2):403-413. [doi: 10.31083/j.rcm2202046] [Medline: 34258907]
10. Rhoden PA, Bonilha H, Harvey J. Patient satisfaction of telemedicine remote patient monitoring: a systematic review. Telemed J E Health 2022 Sep;28(9):1332-1341. [doi: 10.1089/tmj.2021.0434] [Medline: 35041549]
11. Iyamu I, Xu AXT, Gómez-Ramírez O, et al. Defining digital public health and the role of digitization, digitalization, and digital transformation: scoping review. JMIR Public Health Surveill 2021 Nov 26;7(11):e30399. [doi: 10.2196/30399] [Medline: 34842555]
12. Mohammadzadeh N, Rezayi S, Saeedi S. Telemedicine for patient management in remote areas and underserved populations. Disaster Med Public Health Prep 2022 May 19;17:e167. [doi: 10.1017/dmp.2022.76] [Medline: 35586911]
13. Allan J, Webster E, Chambers B, Nott S. “This is streets ahead of what we used to do”: staff perceptions of virtual clinical pharmacy services in rural and remote Australian hospitals. BMC Health Serv Res 2021 Dec 4;21(1):1306. [doi: 10.1186/s12913-021-07328-w] [Medline: 34863164]
14. Hood M, Wilson R, Corsica J, Bradley L, Chirinos D, Vivo A. What do we know about mobile applications for diabetes self-management? A review of reviews. J Behav Med 2016 Dec;39(6):981-994. [doi: 10.1007/s10865-016-9765-3] [Medline: 27412774]
15. Record JD, Ziegelstein RC, Christmas C, Rand CS, Hanyok LA. Delivering personalized care at a distance: how telemedicine can foster getting to know the patient as a person. J Pers Med 2021 Feb 17;11(2):137. [doi: 10.3390/jpm11020137] [Medline: 33671324]
16. Battineni G, Sagaro GG, Chintalapudi N, Amenta F. The benefits of telemedicine in personalized prevention of cardiovascular diseases(CVD): a systematic review. J Pers Med 2021 Jul 14;11(7):658. [doi: 10.3390/jpm11070658] [Medline: 34357125]
17. Muiruri K, King’utu K, Gitahi MW, Otieno GO. Determinants of point-of-care technology use among health care workers in comprehensive care centres, a case of central Kenya. East Afr J Inf Technol 2023;6(1):107-118. [doi: 10.37284/eajit.6.1.1330]
18. Öberg U, Orre CJ, Isaksson U, Schimmer R, Larsson H, Hörnsten Å. Swedish primary healthcare nurses’ perceptions of using digital eHealth services in support of patient self-management. Scand J Caring Sci 2018 Jun;32(2):961-970. [doi: 10.1111/scs.12534] [Medline: 28960451]
19. Koivunen M, Saranto K. Nursing professionals’ experiences of the facilitators and barriers to the use of telehealth applications: a systematic review of qualitative studies. Scand J Caring Sci 2018 Mar;32(1):24-44. [doi: 10.1111/scs.12445] [Medline: 28771752]
20. Jacob C, Sanchez-Vazquez A, Ivory C. Social, organizational, and technological factors impacting clinicians’ adoption of mobile health tools: systematic literature review. JMIR Mhealth Uhealth 2020 Feb 20;8(2):e15935. [doi: 10.2196/15935] [Medline: 32130167]
21. Early J, Gonzalez C, Gordon-Dseagu V, Robles-Calderon L. Use of mobile health (mHealth) technologies and interventions among community health workers globally: a scoping review. Health Promot Pract 2019 Nov;20(6):805-817. [doi: 10.1177/1524839919855391] [Medline: 31179777]
22. Odnoletkova I, Buysse H, Nobels F, et al. Patient and provider acceptance of telecoaching in type 2 diabetes: a mixed-method study embedded in a randomised clinical trial. BMC Med Inform Decis Mak 2016 Nov 9;16(1):142. [doi: 10.1186/s12911-016-0383-3] [Medline: 27825340]
23. Chan G, Nwagu C, Odenigbo I, Alslaity A, Orji R. The shape of mobile health: a systematic review of health visualization on mobile devices. Int J Hum-Comput Interact 2025 Jan 17;41(2):1154-1172. [doi: 10.1080/10447318.2024.2313282]
24. Alshehhi YA, Abdelrazek M, Philip BJ, Bonti A. Understanding user perspectives on data visualization in mHealth apps: a survey study. IEEE Access 2023;11:84200-84213. [doi: 10.1109/ACCESS.2023.3302325]
25. Vaughan EM, Naik AD, Lewis CM, Foreyt JP, Samson SL, Hyman DJ. Telemedicine training and support for community health workers: improving knowledge of diabetes. Telemed J E Health 2020 Feb;26(2):244-250. [doi: 10.1089/tmj.2018.0313] [Medline: 30839244]
26. Borges do Nascimento IJ, Abdulazeem H, Vasanthan LT, et al. Barriers and facilitators to utilizing digital health technologies by healthcare professionals. NPJ Digit Med 2023 Sep 18;6(1):161. [doi: 10.1038/s41746-023-00899-4] [Medline: 37723240]

XSL•FO
RenderX

27. Alipour J, Payandeh A. Assessing the level of digital health literacy among healthcare workers of teaching hospitals in the southeast of Iran. Inform Med Unlocked 2022;29:100868. [doi: 10.1016/j.imu.2022.100868]
28. Schoen J, Mallett JW, Grossman-Kahn R, Brentani A, Kaselitz E, Heisler M. Perspectives and experiences of community health workers in Brazilian primary care centers using m-health tools in home visits with community members. Hum Resour Health 2017 Sep 29;15(1):71. [doi: 10.1186/s12960-017-0245-9] [Medline: 28962569]
29. Coetzee B, Kohrman H, Tomlinson M, Mbewu N, Le Roux I, Adam M. Community health workers' experiences of using video teaching tools during home visits-A pilot study. Health Soc Care Community 2018 Mar;26(2):167-175. [doi: 10.1111/hsc.12488] [Medline: 28872210]
30. Odendaal WA, Anstey Watkins J, Leon N, et al. Health workers' perceptions and experiences of using mHealth technologies to deliver primary healthcare services: a qualitative evidence synthesis. Cochrane Database Syst Rev 2020 Mar 26;3(3):CD011942. [doi: 10.1002/14651858.CD011942.pub2] [Medline: 32216074]
31. Gopalakrishnan L, Buback L, Fernald L, Walker D, Diamond-Smith N, in addition to The CAS Evaluation Consortium. Using mHealth to improve health care delivery in India: a qualitative examination of the perspectives of community health workers and beneficiaries. PLoS ONE 2020;15(1):e0227451. [doi: 10.1371/journal.pone.0227451] [Medline: 31940326]
32. Wagaba MT, Musoke D, Bagonza A, Ddamulira JB, Nalwadda CK, Orach CG. Does mHealth influence community health worker performance in vulnerable populations? A mixed methods study in a multinational refugee settlement in Uganda. PLOS Glob Public Health 2023;3(12):e0002741. [doi: 10.1371/journal.pgph.0002741] [Medline: 38157328]
33. Della Vecchia C, Leroy T, Bauquier C, et al. Willingness of French general practitioners to prescribe mHealth apps and devices: quantitative study. JMIR Mhealth Uhealth 2022 Feb 11;10(2):e28372. [doi: 10.2196/28372] [Medline: 35147508]
34. Ravi RK, Shiva A, Jacob J, Baby P, Pareek B, V KB. Exploring the factors influencing the intention to use mHealth applications in resource scare settings; a SEM analysis among future nurses. Glob Transit 2025;7:199-210. [doi: 10.1016/j.glt.2025.04.002]
35. Anderson K, Francis T, Ibanez-Carrasco F, Globerman J. Physician's perceptions of telemedicine in HIV care provision: a cross-sectional web-based survey. JMIR Public Health Surveill 2017 May 30;3(2):e31. [doi: 10.2196/publichealth.6896] [Medline: 28559226]
36. Hao H, Padman R. An empirical study of opinion leader effects on mobile technology implementation by physicians in an American community health system. Health Informatics J 2018 Sep;24(3):323-333. [doi: 10.1177/1460458216675499] [Medline: 30068266]
37. Leigh S, Ashall-Payne L, Andrews T. Barriers and facilitators to the adoption of mobile health among health care professionals from the United Kingdom: discrete choice experiment. JMIR Mhealth Uhealth 2020 Jul 6;8(7):e17704. [doi: 10.2196/17704] [Medline: 32628118]
38. Jiang S, Xiong M, Rao X, et al. Medical personnel behavior preferences for providing mHealth service in China: a discrete choice experiment. Risk Manag Healthc Policy 2023;16:2405-2418. [doi: 10.2147/RMHP.S430636] [Medline: 38024491]
39. Vo LK, Allen MJ, Cunich M, et al. Stakeholders' preferences for the design and delivery of virtual care services: a systematic review of discrete choice experiments. Soc Sci Med 2024 Jan;340:116459. [doi: 10.1016/j.socscimed.2023.116459] [Medline: 38048738]
40. Houwen T, Vugts MAP, Lansink KWW, et al. Developing mHealth to the context and valuation of injured patients and professionals in hospital trauma care: qualitative and quantitative formative evaluations. JMIR Hum Factors 2022 Jun 20;9(2):e35342. [doi: 10.2196/35342] [Medline: 35723928]
41. Parretti C, La Regina M, Tortu C, Candido G, Tartaglia R. Telemedicine in Italy, the starting point. Intern Emerg Med 2023 Apr;18(3):949-951. [doi: 10.1007/s11739-022-03176-6] [Medline: 36539605]
42. de Bekker-Grob EW, Ryan M, Gerard K. Discrete choice experiments in health economics: a review of the literature. Health Econ 2012 Feb;21(2):145-172. [doi: 10.1002/hec.1697] [Medline: 22223558]
43. Kean J, Reilly J. Item response theory. In: Handbook for Clinical Research: Design, Statistics, and Implementation: Springer; 2014:195-198. [doi: 10.1891/9781617050992.0049]
44. Harvey RJ, Hammer AL. Item Response Theory. Couns Psychol 1999 May;27(3):353-383. [doi: 10.1177/0011000099273004]
45. Goodman LA. Snowball Sampling. Ann Math Statist 1961 Mar;32(1):148-170. [doi: 10.1214/aoms/1177705148]
46. Ackerman TA, Gierl MJ, Walker CM. Using multidimensional item response theory to evaluate educational and psychological tests. Educational Measurement 2003 Sep;22(3):37-51. [doi: 10.1111/j.1745-3992.2003.tb00136.x]
47. De Champlain AF. A primer on classical test theory and item response theory for assessments in medical education. Med Educ 2010 Jan;44(1):109-117. [doi: 10.1111/j.1365-2923.2009.03425.x] [Medline: 20078762]
48. Schatschneider C, Francis DJ, Foorman BR, Fletcher JM, Mehta P. The dimensionality of phonological awareness: an application of item response theory. J Educ Psychol 1999;91(3):439-449. [doi: 10.1037//0022-0663.91.3.439]
49. Thomas ML. Advances in applications of item response theory to clinical assessment. Psychol Assess 2019 Dec;31(12):1442-1455. [doi: 10.1037/pas0000597] [Medline: 30869966]
50. Reise SP, Waller NG. Item response theory and clinical measurement. Annu Rev Clin Psychol 2009;5(1):27-48. [doi: 10.1146/annurev.clinpsy.032408.153553] [Medline: 18976138]

XSL•FO
RenderX

51. Samejima F. The general graded response model. In: Nering ML, Ostini R, editors. Handbook of Polytomous Item Response Theory Models: Taylor & Francis US; 2010:87-118 URL: https://www.taylorfrancis.com/chapters/edit/10.4324/9780203861264-11/general-graded-response-model-fumiko-samejima [accessed 2025-09-09]
52. LaHuis DM, Clark P, O'Brien E. An examination of item response theory item fit indices for the graded response model. Organ Res Methods 2011 Jan;14(1):10-23. [doi: 10.1177/1094428109350930]
53. Landerman LR, Land KC, Pieper CF. An empirical evaluation of the predictive mean matching method for imputing missing values. Sociol Methods Res 1997 Aug;26(1):3-33. [doi: 10.1177/0049124197026001001]
54. Vink G, Frank LE, Pannekoek J, van Buuren S. Predictive mean matching imputation of semicontinuous variables. Stat Neerl 2014 Feb;68(1):61-90. [doi: 10.1111/stan.12023]
55. Ryan M, Netten A, Skåtun D, Smith P. Using discrete choice experiments to estimate a preference-based measure of outcome—An application to social care for older people. J Health Econ 2006 Sep;25(5):927-944. [doi: 10.1016/j.jhealeco.2006.01.001]
56. Luyten J, Kessels R, Atkins KE, Jit M, van Hoek AJ. Quantifying the public's view on social value judgments in vaccine decision-making: a discrete choice experiment. Soc Sci Med 2019 May;228:181-193. [doi: 10.1016/j.socscimed.2019.03.025] [Medline: 30925392]
57. McFadden D. Conditional logit analysis of qualitative choice behavior. In: Frontiers in Econometrics: Academic Press; 1971:105-142 URL: https://eml.berkeley.edu/reprints/mcfadden/zarembka.pdf [accessed 2025-09-09]

Abbreviations

DCE: discrete choice experiment
GRM: Graded Response Model
HCW: health care worker
IRT: item response theory
PMM: Predictive Mean Matching
WCS: Web Conjoint Survey

Edited by A Mavragani; submitted 03.10.24; peer-reviewed by EVD Velde, KL Du; revised version received 17.07.25; accepted 18.07.25; published 23.09.25.

Please cite as:
Tortù C, Seghieri C, Doracaj D, Usheva N, Giménez-Legarre N, Manios Y
Understanding Health Care Workers' Attitudes and Preferences Toward Digital Patient Monitoring Platforms: Cross-Country Survey Study
JMIR Form Res 2025;9:e67142
URL: https://formative.jmir.org/2025/1/e67142
doi:10.2196/67142

© Costanza Tortù, Chiara Seghieri, Ditila Doracaj, Natalya Usheva, Natalia Giménez-Legarre, Yannis Manios. Originally published in JMIR Formative Research (https://formative.jmir.org), 23.9.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Typology and Ethical Considerations of Digital Health Promotion Tools for Youth in Sub-Saharan Africa: Review of Examples From Ghana, Kenya, and South Africa

Agata Ferretti[1], PhD; Shannon Hubbs[1], MA; Richard Mawutor Dzikunu[2], MA; Keymanthri Moodley[3], PhD; Frederick Murunga Wekesah[4], PhD; Jonty Wright[5]; Effy Vayena[1], PhD

[1]Department of Health Sciences and Technology, ETH Zurich, Hottinherstrasse 10, Zurich, Switzerland

[2]YIELD Hub, Rutgers International, Utrecht, The Netherlands

[3]Division for Medical Ethics and Law, Faculty of Medicine and Health Sciences, Stellenbosch University, Western Cape, South Africa

[4]African Population and Health Research Center, Nairobi, Kenya

[5]Faculty of Medicine and Health Sciences, Stellenbosch University, Western Cape, South Africa

Corresponding Author:
Effy Vayena, PhD
Department of Health Sciences and Technology, ETH Zurich, Hottinherstrasse 10, Zurich, Switzerland

Abstract

Background: Digital technologies for health promotion have proliferated over the past decade, with uptake increasing steadily among young people, including those in low- and middle-income countries (LMICs). Youth increasingly rely on digital tools for health information, and the early influence of this digital technology can have an impact throughout the lifespan. While there is a growing body of literature on the opportunities and challenges of digital health promotion (DHP) for young people, a gap remains in research that closely examines the characteristics of digital health strategies developed specifically for youth in LMICs.

Objective: In this paper, we investigate and compare selected examples of DHP tools from 3 countries in Sub-Saharan Africa, namely Ghana, Kenya, and South Africa. Our aim is to create a multidimensional descriptive typology of DHP tools developed specifically to promote the health of adolescents and young adults in these countries.

Methods: To select the tools, we conducted systematic internet-based searches using relevant keywords, incorporating the expertise of local professionals to ensure a thorough search. Included solutions originated from one of the 3 countries of focus and could take any number of forms such as apps, websites, chatbots, or social media initiatives. We thereafter deductively created a typology describing selected features of each tool, including the health area of focus, key stakeholders, type of service, and ethical values explicitly referenced within the tool. While such high-level features of interest were selected based on the existing literature in the field, the detailed descriptive categories were identified through an inductive analysis of the tools.

Results: A total of 31 DHP tools were identified. Sexual and reproductive health was the most common health area of focus for DHP services, which were primarily funded and supported by local non-governmental organizations, foundations, and international organizations. The assessed tools were predominantly web-based and social media-based, with the overarching goal and core value of expanding health knowledge and offering access to health promotion services to young people.

Conclusions: With sustained investment, DHP can improve the health of young people while relieving pressure on health care services. The areas of mental health, as well as substance use prevention and nutrition, stand out with clear potential for health gains through investment in DHP. Addressing ethical concerns such as privacy, transparency, equity, and inclusiveness is essential to the safety, usefulness, and fairness of DHP. To achieve the greatest benefit, local youth perspectives and priorities should be included in DHP development. Local initiatives have the potential to be the most agile, flexible, and relevant for the target audience of young people, with the overall goal of early intervention and greater health quality throughout the lifespan, and more efficient use of health care resources.

(JMIR Form Res 2025;9:e54472) doi:10.2196/54472

KEYWORDS

youth; young people; low- and middle-income countries; LMICs; health promotion; digital health; Sub-Saharan Africa; taxonomy; digital health; tool; African; technology; digital tool; health information; digital health information

XSL•FO
RenderX

Introduction

The use of technology to promote the health of youth is gaining popularity due to the increasing availability of digital tools worldwide. Three-quarters of the world's population owns a smartphone, and in 2022, 66% of the world's population had access to the internet [1]. In low- and middle-income countries (LMICs), the uptake of digital technologies increased over the past decade, offering marginalized and underresourced population groups the opportunity to access services (eg, health care) remotely [2].

Young individuals are among those most engaged with technology, increasingly turning to the internet, adopting digital tools, and asking questions to artificial intelligence–powered interfaces. While rates and reliability of access vary significantly between and within individual countries, recent data indicate that internet accessibility among young people in LMICs stands at approximately 67% and reaches 95% in upper-middle-income economies [3]. However, it is important to note that a reliable and consistent internet connection, as well as access to personal devices, remains a challenge due to infrastructural limitations and socioeconomic disparities [4].

Youth tend to rely on technology (such as apps, social media, podcasts, websites, and forums) to obtain health-related information and to be supported in maintaining a healthy lifestyle [5]. Such use of digital technology by youth can contribute to the mitigation of health risk factors and ultimately reduce the burden of diseases [6,7]. Notably, adolescence and young adulthood, as defined by the World Health Organization to include people of the ages 10 - 24, represent a pivotal period for health promotion interventions [8]. Such interventions can positively impact lifelong health by modifying lifestyle habits and raising awareness about health topics during this transitional phase [9]. Novel digital apps can encourage new possibilities for awareness of one's own health and create new means of caring for one's personal health needs [10].

Effective digital health promotion (DHP) for youth is one valuable means of intervention to combat health inequalities that can emerge early in life, and whose effects can persist throughout the lifespan [11,12]. Although for young people, health inequalities are often related to social inequalities and factors such as family income, stability of parent employment, housing, or ethnicity, active engagement in healthy behaviors can play a role in improving and maintaining well-being [13]. DHP holds potential to positively impact young people in LMICs in particular, as in these regions, technology offers one solution to overcome the limitations of public health systems that grapple with resource constraints. Digital means can reach youth, irrespective of their physical proximity to health care facilities, the presence of national health campaigns, or the availability of health practitioners, thereby not adding to the burden on these systems.

In 2017, Ippoliti and L'Engle [14] reviewed health projects in LMICs that used mobile phones as a means for improving adolescent health, noting that the appeal of mobile technology for health apps lies partly in its affordability and confidentiality. Health promotion was the purpose of a majority of the examined initiatives, which used such methods as text messaging, hotlines, and social media. While this work did not specifically address ethical considerations of mobile health technology for adolescents, it is an example of an earlier effort to investigate and describe available digital health promotion tools for young people in LMICs.

Despite the growing availability of literature on the opportunities and challenges of DHP for young people [15,16], research closely examining the characteristics of DHP strategies specifically developed for and in LMICs and targeting youth in these countries is still missing. Therefore, the objective of this paper is to identify existing DHP strategies that have been developed by different entities for youth in sub-Saharan Africa (SSA), and to establish a typology for these tools, including highlights of their key features. We focused our investigation on 3 English-speaking countries in SSA: Ghana, Kenya, and South Africa.

In these 3 countries, youth face significant challenges to their well-being, including a high burden of HIV and sexually transmitted diseases, road injuries, interpersonal violence, malnutrition, and mental health conditions [8]. Primary risk factors affecting this demographic encompass unsafe sexual practices, limited knowledge of sexual and mental health, unaddressed existing mental health conditions, insufficient physical activity, unhealthy dietary habits, overweight concerns, and substance use (smoking, binge drinking, and drug usage) [17]. These areas of health present opportunities for intervention through enhanced health awareness and individualized reinforcement of positive health behavior strategies, which stand at the core of DHP.

Methods

Overview

In May 2023, we aimed to identify examples of active digital health promotion tools and strategies targeting youth in the 3 countries of interest (Ghana, Kenya, and South Africa). Given the absence of a centralized repository for DHP tools, we identified initiatives through systematic online searches using English-language keywords related to youth, health promotion, and digital media. In addition, to ensure a comprehensive survey of available tools, we relied on the expertise of local professionals in each of the included countries (RMD, FW, JW, and KM) who work in the field of youth health and digital technologies and were knowledgeable about relevant examples. The list of tools we compiled reflects the specific moment in time when the search was conducted and will likely evolve as new tools emerge and others become inactive or unavailable. However, this represents the first effort in this direction to collect and analyze digital health promotion tools for youth in SSA, with the aim that future research will build upon.

To be included in our study, DHP initiatives needed to explicitly focus on young people, adolescents, or young adults as the primary users or stakeholders. Furthermore, they should originate from one of the 3 countries of interest and be developed and sponsored by local non-governmental organizations (NGOs), foundations, entrepreneurs, private sector

XSL•FO
RenderX

entities, research institutions, or governments. Therefore, we did not consider apps and strategies that were created in other parts of the world, or that were introduced and promoted by global brands or multinational corporations, in order to maintain a regional focus. Included DHP tools took various forms, such as websites, social media interventions, texting services (eg, WhatsApp and chatbots), mobile health and fitness apps, and audio services (eg, podcasts and calls) [18].

Once the tools were identified, we established a typology based on deductively selected descriptive variables. We derived these variables from existing literature on the subject, emphasizing key characteristics of DHP [18-22]. Specifically, we examined the tool's health focus area, the key stakeholders and actors involved in its development and deployment, the format and type of services provided, and the ethical principles and values explicitly promoted within the DHP or in its content. The subcategories within each feature were identified inductively. Two researchers independently reviewed the online content of each tool, comparing their analyses and discussing any disagreements to reach a consensus. Importantly, we only analyzed primary information available directly on the websites of the tools and did not use secondary data about these tools (eg, articles published about such tools).

In compiling and describing various typologies of DHP in Ghana, Kenya, and South Africa, we acknowledge that our efforts do not encompass all available tools, as some may not have been detected in our search, and as new tools are continuously being developed. Nonetheless, we aimed to maintain a systematic approach.

It is important to note that inclusion of these tools in our analysis does not imply endorsement of these specific initiatives and products. Rather, we would like to showcase examples of existing DHP services that are presently available to any young user, and to report on the variety of features offered by these tools.

Ethical Considerations

Data used in our study consisted of information that is publicly available on digital health promotion initiatives online. Therefore, no anonymization or deidentification was carried out in the course of this study.

Results

Overview

We identified a total of 31 DHP tools, including 16 in Kenya, 13 in South Africa, and 10 in Ghana (Figure 1). All tools originated in one of the 3 countries, with some tools available in multiple countries. The majority of these tools were designed for all young users, with few designed specifically for women, and notably, only 1 out of 31 targeting men.

Figure 1. Typology of examples of digital health promotion tools for young people from Ghana, Kenya, and South Africa, as identified via systematic online search in May 2023. Search criteria included English-language keywords related to youth, health promotion, and digital media, with additional consultation with local professionals to gather any further relevant tools available at that time. Analyzed characteristics include country, target population, target health area, key actors, format, service, and ethical principles and values. DHP: digital health promotion.

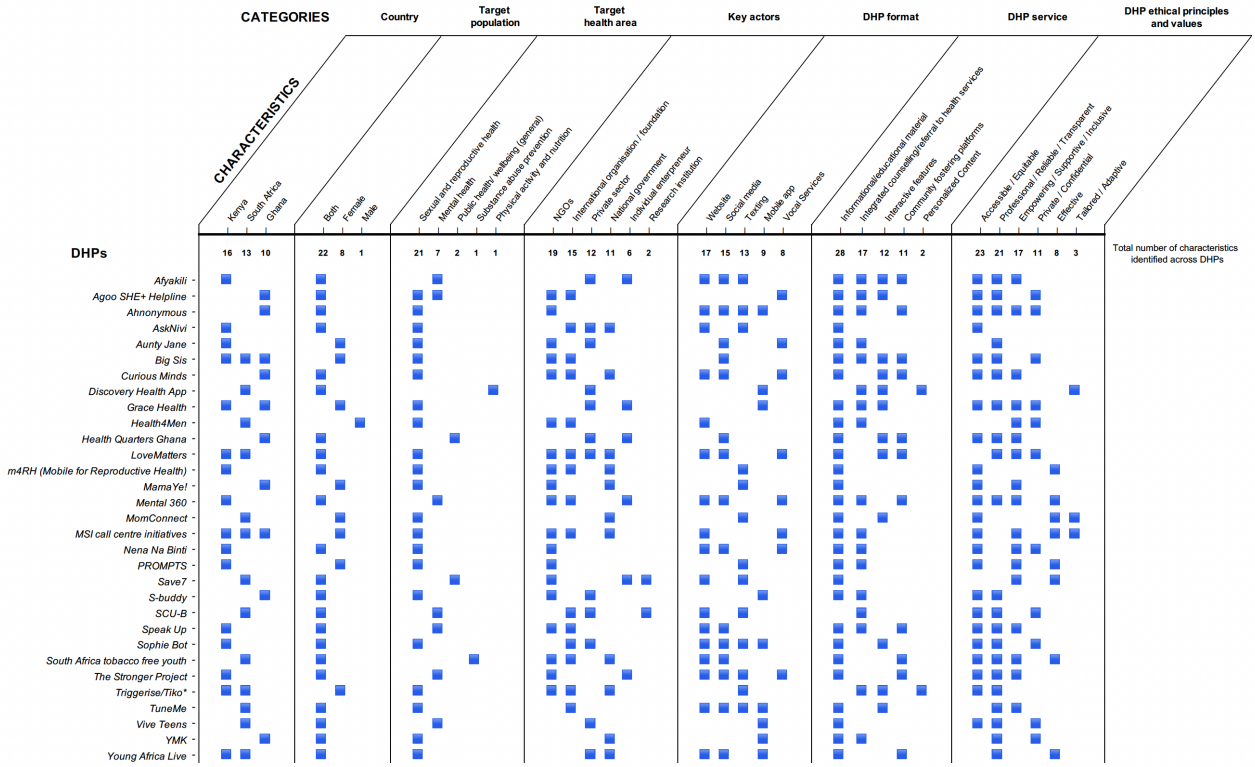

XSL•FO
RenderX

DHP Area of Health Focus

The majority of DHP tools concentrated on sexual and reproductive health topics (21/31), which is understandable given the prevalence of underage pregnancies, HIV, and other sexually transmitted infections in the 3 countries, and the associated maternal health risks. We found a limited number of DHP solutions addressing youth mental health (7/31), an area that is steadily gaining attention in the region. Notably, we identified a scarcity of locally developed DHP targeting substance abuse prevention or promoting healthy lifestyle habits [23].

Key Stakeholders and Actors Involved in DHP Development and Deployment

The development and promotion of most DHPs was facilitated by local NGOs (19/31) and foundations and international organizations (15/31), with the private sector (eg, health insurance and tech companies) also offering substantial support (12/31). National governments also contributed to the development of existing DHPs (11/31), supporting broad initiatives often made available at the regional or country level. The role of individual entrepreneurs (6/31) and research institutions (2/31) was relatively minor among the examples we identified.

DHP Formats and Type of Services Provided

The majority of the identified DHP tools were primarily accessible through websites (17/31), followed by availability on social media platforms (15/31) and texting solutions, including messaging and chatbots (9/31). Only a small number of DHP solutions consisted of dedicated mobile apps (9/31).

The majority of the reviewed DHP solutions (28/31) offered informational and educational materials to users. In addition, integrated counseling services, the option to engage with experts, and referrals to health services were provided by 17 out of 31 DHP services. This finding appears to be in contrast with what young people have commonly cited as an appeal of DHP, namely, the possibility to avoid discussions with biased adults or health care professionals regarding sensitive or potentially taboo topics [18]. However, the inclusion of such services in countries such as Ghana, Kenya, and South Africa can potentially be explained by the cultural value of involving others in decision-making and seeking the input of medical professionals [4].

Interactive design (12/31), platforms fostering community interaction (11/31), and personalized content (2/31) were less commonly featured in the analyzed tools. This aspect is noteworthy, as it seems to conflict with what youth report to be the most important DHP features. Research shows that young people tend to engage more and stay committed to using digital solutions when they find them aesthetically appealing and functional [24]. They also appreciate content tailored specifically to their needs [25,26]. Moreover, young people value opportunities to share their concerns with peers and receive feedback from those who have faced similar experiences, as such peer dynamics enable a sense of support and understanding [22].

Ethical Principles and Values Explicitly Endorsed by DHP

DHP tools frequently emphasized ethical values and considerations among their stated goals and missions. Most often cited (23/31) was increasing accessibility of health information and health promotion services, thereby expanding opportunities to a greater number of youths and complementing the work of traditional health systems. Low-budget (if not in most cases, free) DHP strategies are presented as a solution for young people wishing to take care of their own health, reaching them via digital tools regardless of socioeconomic, cultural, or location barriers. In addition, by leveraging design and aesthetics, as well as audio and visual features, these health promotion strategies engage young people at all levels of education and health literacy, making DHP a more equitable tool.

Many of the DHP initiatives analyzed also emphasized values of transparency, truthfulness, and reliability of the information provided to users (21/31). In particular, these values are realized by combating data misinterpretation and misinformation. This is accomplished by providing users with scientifically valid and verifiable information that is supported by sound sources of data, integrating the ability to contact health experts on demand, combating fake news, and being transparent about which data will be used by the DHP tool. Inclusiveness and empowerment also recurred frequently in ethics and value statements (17/31). Notably, DHP emphasizes being user-centered (youth-focused and often youth-driven), supporting individuals (and indirectly communities) in behaviors that will help their health and specific well-being to flourish, and including all individuals, rejecting stigma or cultural barriers.

In contrast, ethical issues related to privacy (11/31), effectiveness (8/31), and adaptability of DHP featured less among the mission and value statements of the DHP tools we identified.

Discussion

Principal Findings and Comparison With Previous Works

Our analysis and creation of a multidimensional typology of DHP strategies identifies recurrent features, as well as potential gaps and areas for further development, among identified active DHP tools in Ghana, Kenya, and South Africa. Youth in these countries face a variety of unique health challenges and risks. According to our results, presently available solutions offer a great deal to young people as health promotion resources, but do not fully address these health needs, suggesting a significant untapped opportunity for investment in DHP. Our results indicate that this is particularly true in the area of youth mental health, as well as for substance abuse prevention and nutrition. Apps developed in Western contexts frequently cover these topics, but may be less relevant for the specific health needs of youth in SSA. For example, DHP apps that recommend outdoor physical activity or specific dietary habits may be impractical in areas where it is unsafe to be outside or where food options

XSL•FO
RenderX

are limited. Thus, DHP developed locally with input from health experts can enrich the landscape of fit-for-the-setting solutions.

Our findings demonstrate that many DHP initiatives available to youth in SSA are already formed and operate locally, with agile development processes that are not reliant on bureaucratic systems or formal government programs for approval. Such smaller-scale projects run by regional NGOs offer flexibility in tailoring tools and content to the needs of specific youth populations and vulnerable groups. Such solutions can not only take into account local barriers to the implementation of DHP, but can also integrate local values, such as relational privacy or group solidarity. This agility, together with the potential for valuable local data collection, provides further rationale for private sector investment in DHP.

Our findings suggest that currently available DHP tools may only partially address the health needs and risk factors most prevalent among youth in these 3 SSA countries, for example, regarding substance abuse prevention, mental health, and nutrition. This omission could be attributed to factors such as potential stigma around mental health and substance use issues, and their absence from cultural dialogue (hindering investment or cultural support); greater incentive to invest in DHP strategies for issues considered more critical (eg, HIV); or lack of data and understanding of the significance of these areas of adolescent health, among developers or government bodies. In our survey of DHP tools, only one aimed to address more than one area of health. This could be due to the need to be effective and valuable to users while standing out in a competitive market. To succeed, tools must offer content tailored to the needs of youth, and therefore diversifying across multiple health areas could be counterproductive, diluting their focus and reducing their impact.

Our data highlight the prevailing preference in LMICs for cost-effective solutions such as websites, with developers and investors often opting to leverage existing infrastructure such as social media and texting platforms, instead of developing entirely new apps that demand additional resources. In addition, if the youth population is receptive to tools developed in Western countries and sponsored by large companies, it can be more challenging for local providers and developers to secure market space.

DHP presents an opportunity to intervene early in unhealthy behaviors, thereby fostering health and reducing the number of individuals who may require acute or chronic care as adults. DHP encourages individuals to take an active role in managing their own health rather than being passive recipients of medical advice. These factors make a strong case for these national governments to continue to invest in DHP projects. Supporting tools that strengthen health education and promote healthy behavior can lead to long-term benefits, ultimately reducing the burden on health care systems. Involving multiple investors with reliable funding can enhance the sustainability of DHP programs beyond one-time strategies and short-term campaigns.

Interdisciplinary teams of experts and community leaders who focus on technical aspects and also ensure ethical alignment of DHP could be valuable. This analysis confirms that crucial ethical issues such as individual and group privacy, equity, inclusiveness, and transparency should not be overlooked, especially in settings where access to health care and health literacy is limited. While our analysis suggests a strong focus among tool developers on the reliability of health information, the literature points to the frequent inaccuracy of information shared by DHP tools. For the time being, beyond checks instituted by developers themselves, there are no independent review mechanisms assessing information provided to users. In the absence of external oversight, a holistic approach to the development and implementation of DHP, beginning with interdisciplinary teams, can mitigate the dissemination of unreliable and potentially harmful information to users. Referencing sources, being transparent about the criteria used to develop recommendations, introducing feedback systems, and developing open-source digital health software repositories are further steps to ensure that DHP solutions are more ethically aligned and safe.

Reflecting on the ethical principles specifically endorsed by DHP tools, research indicates that privacy and effectiveness are not among young people's top priorities when considering DHP, which may in turn lead developers and deployers not to prioritize these values [16,18,27]. In SSA contexts, Western ideas of privacy and technology tailored to individual needs give way to more fundamental interpretations based on relational aspects and group dynamics. However, glossing over these values could signal inadequate recognition of their significance. Further reflection is warranted concerning implications for privacy breaches, tool effectiveness, and tailoring content to user needs.

As new DHP initiatives are developed regularly, our review captures the tools that the authors were able to identify at the time of our search. Thus, our analysis examines a selection of DHP tools that is not comprehensive. While 2 reviewers examined all information accessible online about each tool and reached agreement in our findings, our analysis was limited to the data currently provided about each initiative.

This analysis points to areas of misalignment between the formats and features of DHP tools and the preferences of young people as described in the literature, raising questions about the process of conception and development of initiatives intended for young people. It suggests that if youth have more meaningful involvement in developing DHP tools for their use, the result will be even better suited to their needs. Our analysis also demonstrates that existing literature reports predominantly on research in high-income countries, reflecting the perspectives of youth in these countries. This ambiguity underscores the need for greater engagement and meaningful consultation with youth in LMICs, in planning and developing DHP tools, and in research on digital health solutions.

Conclusions

This paper presents and compares selected examples of digital tools and solutions for youth health promotion in SSA, particularly in Ghana, Kenya, and South Africa. As DHP becomes increasingly available to youth, it is important to analyze their characteristics to better understand the strengths and pitfalls present in different settings. Our work to create a typology of DHP characteristics in these countries aims to highlight recurring patterns among the tools, as well as possible

XSL•FO
RenderX

gaps that should be addressed to ensure ethically aligned technologies. In addition to expanding access to DHP in LMICs by increasing investment in the health sector and empowering local developers to create context-specific digital programs, our analysis shows the need to incorporate a number of ethical principles in the development and deployment of such tools. Ethical considerations of inclusiveness, equity, transparency, effectiveness, and personalization cannot be neglected, if not to the detriment of the quality and safety of DHP. Only by meaningfully engaging young people and knowing their concerns is it possible to create tools that effectively meet users' needs and that are sustainable over time. As research on youth perspectives on digital health progresses, technology development should adapt and incorporate such knowledge and feedback.

Acknowledgments

This work was supported by Fondation Botnar (grant number OOG-20-024).

Disclaimer

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data Availability

Data are supplied in supporting files available for download along with the published manuscript in the multimedia appendix.

Authors' Contributions

AF and EV conceptualized the study. AF, RMD, FW, and JW identified the DHP tools and collected the data. AF and SH conducted the data analysis, drafted, and edited the manuscript. EV and KM were the senior authors who revised and edited the manuscript. EF acquired the grant which funded this project. All authors contributed to the intellectual content of the manuscript and approved the final version.

Conflicts of Interest

None declared.

References

1. Facts and figures 2022 - internet use. ITU. 2022. URL: https://www.itu.int/itu-d/reports/statistics/2022/11/24/ff22-internet-use [accessed 2023-09-10]
2. The state of mobile internet connectivity 2022. GSMA. 2022. URL: https://www.gsma.com/r/wp-content/uploads/2022/12/The-State-of-Mobile-Internet-Connectivity-Report-2022.pdf [accessed 2023-09-10]
3. Facts and figures 2022 - youth internet use. ITU. 2022. URL: https://www.itu.int/itu-d/reports/statistics/2022/11/24/ff22-youth-internet-use [accessed 2023-09-10]
4. Third A, Moody L. Our rights in the digital world: a report on the children's consultations to inform UNCRC general comment 25. : 5Rights Foundation and Western Sydney University; 2021 URL: https://5rightsfoundation.com/wp-content/uploads/2024/09/OurRIghtsinaDigitalWorld-FullReport.pdf [accessed 2025-09-10]
5. Governing Health Futures Commission. Digital health futures: insights into young people's use and opinions of digital health technologies. Summary report of a 2020 U-Report poll. 2021 URL: https://www.governinghealthfutures2030.org/wp-content/uploads/2021/09/Digital-health-futures-U-Report.pdf [accessed 2025-09-10]
6. Kickbusch I, Piselli D, Agrawal A, et al. The Lancet and Financial Times Commission on governing health futures 2030: growing up in a digital world. The Lancet 2021 Nov;398(10312):1727-1776. [doi: 10.1016/S0140-6736(21)01824-9]
7. Oh C, Carducci B, Vaivada T, Bhutta ZA. Digital interventions for universal health promotion in children and adolescents: a systematic review. Pediatrics 2022 May 1;149(Suppl 5):e2021053852H. [doi: 10.1542/peds.2021-053852H] [Medline: 35503335]
8. Adolescent and young adult health. WHO. 2023. URL: https://www.who.int/news-room/fact-sheets/detail/adolescents-health-risks-and-solutions [accessed 2023-09-10]
9. Global accelerated action for the health of adolescents (AA-HA!): guidance to support country implementation. WHO. 2017. URL: https://iris.who.int/bitstream/handle/10665/255415/9789241512343-eng.pdf?sequence=1 [accessed 2023-09-10]
10. Chaney S, Mechael P. Digital self-care: a framework for design, implementation & evaluation.: Self-Care Trailblazer Group; 2020. URL: https://media.psi.org/wp-content/uploads/2020/10/31000255/Digital-Self-Care.pdf [accessed 2025-06-22]
11. Oliver S, Kavanagh J, Caird J, Lorenc T, Oliver K, Harden A, et al. Health promotion, inequalities and young people's health: a systematic review of research. : ePPI Centre; 2008 URL: https://eppi.ioe.ac.uk/cms/Portals/0/PDF%20reviews%20and%20summaries/Inequalities%20Young%20People%20R2008Oliver.pdf [accessed 2025-09-10]
12. Ferretti A, Vayena E, Blasimme A. Unlock digital health promotion in LMICs to benefit the youth. PLOS Digit Health 2023 Aug;2(8):e0000315. [doi: 10.1371/journal.pdig.0000315] [Medline: 37540713]

XSL•FO
RenderX

13. Wong BLH, Smith RD, Siepmann I, Hasse A, Tandon S. Youth engagement in digital health: a critical perspective towards meaningful youth agency in governance. Medicus Mundi Schweiz. 2021. URL: https://www.medicusmundi.ch/de/advocacy/publikationen/mms-bulletin/digital-health-a-new-era-of-global-health/kapitel-3/youth-engagement-in-digital-health-a-critical [accessed 2023-09-10]
14. Ippoliti NB, L'Engle K. Meet us on the phone: mobile phone programs for adolescent sexual and reproductive health in low-to-middle income countries. Reprod Health 2017 Jan 17;14(1):11. [doi: 10.1186/s12978-016-0276-z] [Medline: 28095855]
15. Holly L, Smith RD, Ndili N, Franz C, Stevens EAG. A review of digital health strategies in 10 countries with young populations: do they serve the health and wellbeing of children and youth in a digital age? Front Digit Health 2022;4:817810. [doi: 10.3389/fdgth.2022.817810] [Medline: 35373182]
16. Lupton D. Young people's use of digital health technologies in the global north: narrative review. J Med Internet Res 2021 Jan 11;23(1):e18286. [doi: 10.2196/18286] [Medline: 33427684]
17. Mokdad AH, Forouzanfar MH, Daoud F, et al. Global burden of diseases, injuries, and risk factors for young people's health during 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. The Lancet 2016 Jun;387(10036):2383-2401. [doi: 10.1016/S0140-6736(16)00648-6]
18. Ferretti A, Hubbs S, Vayena E. Global youth perspectives on digital health promotion: a scoping review. BMC Digit Health 2023;1(1):25. [doi: 10.1186/s44247-023-00025-0]
19. Koh A, Swanepoel DW, Ling A, Ho BL, Tan SY, Lim J. Digital health promotion: promise and peril. Health Promot Int 2021 Dec 13;36(Supplement_1):i70-i80. [doi: 10.1093/heapro/daab134] [Medline: 34897444]
20. Clarke AM, Kuosmanen T, Barry MM. A systematic review of online youth mental health promotion and prevention interventions. J Youth Adolesc 2015 Jan;44(1):90-113. [doi: 10.1007/s10964-014-0165-0] [Medline: 25115460]
21. Bailey J, Mann S, Wayal S, et al. Sexual health promotion for young people delivered via digital media: a scoping review. Public Health Res 2015;3(13):1-120. [doi: 10.3310/phr03130]
22. Wong CA, Madanay F, Ozer EM, et al. Digital health technology to enhance adolescent and young adult clinical preventive services: affordances and challenges. J Adolesc Health 2020 Aug;67(2S):S24-S33. [doi: 10.1016/j.jadohealth.2019.10.018] [Medline: 32718511]
23. Mabrouk A, Mbithi G, Chongwo E, et al. Mental health interventions for adolescents in sub-Saharan Africa: a scoping review. Front Psychiatry 2022;13:937723. [doi: 10.3389/fpsyt.2022.937723] [Medline: 36061286]
24. Plaisime M, Robertson-James C, Mejia L, Núñez A, Wolf J, Reels S. Social media and teens: a needs assessment exploring the potential role of social media in promoting health. Social Media + Society 2020 Jan;6(1):2056305119886025. [doi: 10.1177/2056305119886025]
25. Schwarz AF, Huertas-Delgado FJ, Cardon G, DeSmet A. Design features associated with user engagement in digital games for healthy lifestyle promotion in youth: a systematic review of qualitative and quantitative studies. Games Health J 2020 Jun;9(3):150-163. [doi: 10.1089/g4h.2019.0058] [Medline: 31923363]
26. Michel T, Tachtler F, Slovak P, Fitzpatrick G. A review of youth mental health promotion apps towards their fit with youth media preferences. EAI Endorsed Trans Perv Health Tech 2019;5(17):e2. [doi: 10.4108/eai.13-7-2018.161419]
27. Livingstone S, Stoilova M, Nandagiri R. Children's data and privacy online: growing up in a digital age: an evidence review. : London School of Economics and Political Science; 2019 URL: https://eprints.lse.ac.uk/101283/1/Livingstone_childrens_data_and_privacy_online_evidence_review_published.pdf [accessed 2025-09-10]

Abbreviations

DHP: digital health promotion
LMIC: low- and middle-income countries
NGO: non-governmental organization
SSA: sub-Saharan Africa

Edited by A Mavragani; submitted 13.11.23; peer-reviewed by T Zan; revised version received 03.03.25; accepted 04.03.25; published 15.09.25.

Please cite as:
Ferretti A, Hubbs S, Dzikunu RM, Moodley K, Wekesah FM, Wright J, Vayena E
Typology and Ethical Considerations of Digital Health Promotion Tools for Youth in Sub-Saharan Africa: Review of Examples From Ghana, Kenya, and South Africa
JMIR Form Res 2025;9:e54472
URL: https://formative.jmir.org/2025/1/e54472
doi:10.2196/54472

XSL•FO
RenderX

© Agata Ferretti, Shannon Hubbs, Richard Mawutor Dzikunu, Keymanthri Moodley, Frederick Murunga Wekesah, Jonty Wright, Effy Vayena. Originally published in JMIR Formative Research (https://formative.jmir.org), 15.9.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Personality-Driven Variations in Fitness App Affordance Actualization Among Adults: Quantitative Survey Study

Moayad Alshawmar[1], PhD; Bengisu Tulu[2], PhD; E Vance Wilson[2], PhD; Adrienne Hall-Phillips[2], PhD

[1]College of Business, Imam Mohammad Ibn Saud Islamic University (IMSIU), Al Thoumamah Rd, Riyadh, Saudi Arabia

[2]Business School, Worcester Polytechnic Institute, Worcester, MA, United States

Corresponding Author:
Moayad Alshawmar, PhD
College of Business, Imam Mohammad Ibn Saud Islamic University (IMSIU), Al Thoumamah Rd, Riyadh, Saudi Arabia

Abstract

Background: Fitness apps aim to advance individuals' health and wellness by encouraging consistent healthy habits. Despite their widespread use, sustaining user engagement remains a challenge. Research studies on fitness apps have identified app affordances as one of the key factors that influence user engagement. Some affordances, such as exercise guidance and activity status updates, are shown to support users in achieving their health goals if the users actualize them. However, these affordances need to be actualized by the users to seize these benefits. While identifying these app affordances can deepen our insight into user-app interactions, the impact of personality traits on the actualization of these affordances remains underexplored.

Objective: This study aims to examine the influence of personality traits on the actualization of fitness app affordances.

Methods: Building on affordance actualization theory and the Big Five personality framework, we hypothesized about certain personality traits influencing the actualization of certain app affordances. We tested these hypotheses using a survey of adult Fitbit app (Google LLC) users (N=442). We used validated measures from the literature to assess these variables. We analyzed the survey data using covariance-based structural equation modeling.

Results: Our findings reveal distinct affordance actualization patterns based on users' personality traits. Users with the conscientious personality trait primarily actualize the updating affordance (β=0.136, *P*=.01), while the influence of the conscientious trait on actualization of rewards (β=–0.154, *P*=.06), competing (β=−0.118, *P*=.18), comparing (β=–0.084, *P*=.33), reminding (β=−0.060, *P*=.44), or guidance (β=−0.006, *P*=.95) affordances was not significant. The openness to experience trait showed a significant positive effect on actualization of updating affordances (β=0.227, *P*=.001), but did not significantly influence actualization of searching (β=−0.172, *P*=.11), watching others (β=−0.077, *P*=.50), or guidance (β=−0.005, *P*=.96) affordances. Users with the agreeableness trait actualized comparison (β=0.213, *P*=.02), guidance (β=0.259, *P*=.003), and encouragement (β=0.244, *P*=.01) affordances, while the effect of the agreeableness trait on actualization of watching others was not significant (β=0.143, *P*=.13). Extravert users actualized recognition (β=0.191, *P*<.001), self-presentation (β=0.165, *P*=.002), and watching others (β=0.167, *P*=.003) affordances, but did not actualize updating affordances (β=0.001, *P*=.98). Finally, a lower emotional stability trait did not significantly influence any of the hypothesized affordances, with nonsignificant effects on guidance (β=−0.083, *P*=.30), reminding (β=−0.093, *P*=.21), and updates (β=−0.036, *P*=.49).

Conclusions: Our study shows that certain personality traits are associated with the actualization of specific affordances. These findings underscore the need to tailor fitness app affordances to individual differences, rather than relying on a one-size-fits-all approach. Designing fitness app functionality that aligns with various personality traits may promote deeper and more sustained user engagement. Further research is needed to investigate the relationship between personality traits and app affordance actualization.

(JMIR Form Res 2025;9:e72691) doi:10.2196/72691

KEYWORDS

fitness apps; personality traits; affordance actualization; user interaction; mHealth app; mobile health

Introduction

Fitness apps encourage users to engage in regular physical activity by offering personalized guidance, progress tracking, and social support [1]. When used consistently, these apps can enhance physical and mental well-being, reduce the risk of chronic disease, and improve memory [2,3]. Despite these benefits, many users discontinue use shortly after downloading [4,5]. Understanding the factors that influence sustained engagement is therefore critical.

One way to explain this discontinuation is by examining the affordances that users perceive and actualize, opportunities for

XSL•FO
RenderX

action provided by the app. Studies have identified common affordances in fitness apps, such as exercise guidance, social comparison, rewards, watching others, and reminders [6-10]. These affordances are shaped by users' goals and how they interact with the app, rather than by design alone. As Leonardi [11] notes, individuals' goals determine what they perceive technology features as enabling. Strong et al [12] distinguish affordances as "potentials for action" and actualization of affordances as "actions taken by individuals to realize those potentials" through the use of technology. Research shows that these goals influence the affordances users actualize; for example, those with extrinsic exercise goals may avoid social affordances [9], while older users may value health progress updates [13].

While users' goals influence which affordances they actualize, these goals themselves may be shaped by deeper, stable individual characteristics, such as their personality traits. Personality theory suggests that life pursuits are expressions of underlying dispositions [14]. In the personality traits literature, the Five-Factor Model, also known as Big Five traits, is the most widely accepted among the many approaches to identifying personality traits [15]. The Big Five traits characterize human personality into 5 major characteristics while accepting that an individual will tend to have some of the traits more than others [16]. The Big Five traits, agreeableness, conscientiousness, extraversion, neuroticism, and openness capture broad individual differences in tendencies such as cooperation, organization, sociability, emotional stability, and curiosity [17].

Prior work links these traits to technology use behaviors, including fitness apps [15,18-20], yet it remains unclear whether personality influences which affordances users actualize. Accordingly, this study examines how personality traits influence the actualization of specific fitness app affordances. We hypothesize that variations in the Big Five traits will lead to differences in which affordances users actualize.

People high in conscientiousness are organized, efficient, and achievement-oriented [21]. They excel at following physical activity guidelines [22] and perform tasks more effectively with reminder technologies [15]. In fitness apps, conscientious users rate competitive strategies as strong motivators [23]. Given their drive to accomplish goals, we hypothesize (H1) that conscientious Fitbit (Google LLC) users will actualize affordances related to achievement and exercise effectiveness, such as guidance, reminding, updating, competing, rewards, and comparing, allowing them to complete exercises correctly and on time and to track their progress.

People high in agreeableness are cooperative, compassionate, and warm [21]. They thrive in collaborative settings [24] and engage with technologies that support teamwork [25]. In fitness apps, they are drawn to features that foster social connection, such as viewing others' progress, challenging others, and exercising in live workout sessions [19,25,26]. Agreeable people are also caring in nature and expect others to repay such care [27]. Therefore, we hypothesize (H2) that because agreeable people are more likely to engage in collaborative work and care more about others, they will actualize comparing, guidelines, watching others, and encouragement affordances.

Emotional stability reflects individuals' control over their emotions [28]. Those with lower stability tend to be anxious and insecure [28] and often report lower satisfaction with work and technology use [16,27,29-31]. However, when faced with health concerns, they may become highly motivated to improve their health [32-34]. One study reported that such individuals smoked fewer cigarettes, drank less alcohol, and quit smoking immediately when they faced negative health conditions [35]. Another study reported a positive impact of users' having lower emotional stability on how they rate the fitness apps' usefulness [36]. Given their focus on managing health, we hypothesize (H3) that Fitbit users with lower emotional stability will actualize affordances such as updating, guidance, and reminding that help them monitor health status, perform exercises correctly, and stay on schedule.

People high in openness are imaginative, curious, and receptive to new ideas [21]. They embrace change, explore novel activities such as mobile health apps, and are intrinsically motivated by enjoyment rather than external rewards [37-41]. Given their curiosity and desire to avoid exercise errors, we hypothesize (H4) that open Fitbit users will actualize affordances that support exploration and accurate performance, such as searching, watching others, guidance, and updating.

People high in extraversion are sociable, outgoing, and energetic [38]. Their desire for social connection and maintaining a positive image drives engagement in interactive activities [25], both in real life and on social media [42,43]. In fitness apps, extraverts are likely to be drawn to socially oriented affordances that involve sharing and interacting with others. We hypothesize (H5) that extraverted Fitbit users will actualize affordances such as self-presentation, social recognition, watching others, and updating.

Methods

Study Design

This study used a cross-sectional, web-based survey design, and its findings were reported in accordance with the CHERRIES (Checklist for Reporting Results of Internet E-Surveys) checklist [44]. The survey aimed to investigate how users' personality traits influence their actualization of fitness app affordances. The instrument included validated measures drawn from prior research, covering fitness app affordances and personality traits [7,8,35]. A pilot study was conducted prior to the main data collection to validate the survey instrument. Fitbit, identified as the most downloaded fitness app [36], was selected as the focal platform, and current and previous Fitbit users were recruited as participants.

Measures

To evaluate the research hypotheses, we adopted the measures and scales from previous information systems, mobile health and fitness app studies that call attention to fitness app affordances [7,8] along with a widely used short version (20 items) measure of the Big Five personality traits [35]. The survey instrument consisted of 4 sections: (1) Big Five personality traits questions (4 items for each trait) [35], (2) 11 fitness app affordances questions (4 items for each affordance)

XSL•FO
RenderX

[7,8], (3) user usage status, and (4) demographic information. All items were measured on a 5-point Likert scale with relevant anchors, as shown in Multimedia Appendices 1 and 2 (the latter was adapted from [9,10]).

Participants and Recruitment Procedures

Prior to the main study, a pilot study was conducted with participants recruited between June 1 and July 5, 2022 via Sona system participant pool at a university and via social media platforms (Facebook [Meta Platforms Inc] and Reddit [Reddit Inc]). A total of 403 participants completed the pilot study. After excluding incomplete responses, those failing validation checks, or showing signs of inattentive responding (eg, straightlining or unusually short completion time), 81 valid responses remained. These 81 participants were between 18‑52 years old, with 79% (64/81) of the sample 25 years or older. All scales had acceptable reliability with a Cronbach α of at least 0.651, as shown in Multimedia Appendix 3 [45]. The pilot study helped us validate the survey instrument and identified recruitment challenges that were addressed to improve the main study.

For the main study, the target sample size was calculated based on the formula suggested by Gaskin [46]. Given the limitations of social media recruitment (eg, large percentage of invalid responses) identified during the pilot study, for the main study, we recruited participants through Prolific (Prolific Academic Ltd), a crowdsourcing platform known for high-quality data [47]. We first distributed a prescreening survey to 1300 participants to identify current adult US-based Fitbit app users engaging in exercise. We identified 622 users who were then invited to participate in the main study. A total of 506 participants completed the survey. After removing the responses of participants who failed attention checks, 442 valid responses (71%) remained.

Ethical Considerations

Both the pilot and main studies received approval from the Institutional Review Board (IRB-22‑0351) of Worcester Polytechnic Institute and were conducted in accordance with ethical guidelines for human participant research. Participants were provided informed consent electronically at the beginning of each survey, where they were informed about the study's purpose, the voluntary nature of participation, and their right to withdraw at any time. The consent form also noted that data would be stored securely and accessed only by the research team. All responses were collected anonymously.

For the pilot study, student participants recruited through the university's local participant pool received 0.5 course credit as an incentive, whereas participants recruited through social media received an opportunity to enter a raffle for one of 10 US $10 prizes as an incentive. The distribution of incentives was managed by one of the authors. For the main study, all recruitment and distribution of incentives were managed by Prolific. Participants recruited for the prescreening survey received a flat rate of US $0.18 as an incentive. Among these participants, those who were eligible to participate in the actual study (current Fitbit users who engage in exercise) received a flat rate of US $6 as an additional incentive.

Results

Descriptive Statistics

The main study sample demographics are illustrated in Table 1. The study sample was balanced by gender (50% male, 48% female, and 2% transgender and nonbinary). Most of the participants were between the ages of 25 and 44 (64%), White (81%), and employed (84%).

XSL•FO
RenderX

Table . Demographics of the main study sample (N=442), based on an online survey of US-based Fitbit users engaging in regular exercise; source: authors' own work.

	Participants, % (N=442)
Gender	
Men	50
Women	48
Transgender	1
Nonbinary	1
Age (years)	
18 - 24	7
25 - 34	33
35 - 44	31
45 - 54	13
55 - 64	11
65 - 74	4
75 or older	1
Ethnicity	
White	81
Black	9
Asian	8
Other	2
Employment status	
Full-time	70
Part-time	14
Not in paid work (eg, homemaker, retired, or disabled)	9
Unemployed (and seeking a job)	4
Other	3

All participants were current Fitbit users with varying usage patterns as illustrated in Table 2. Among our sample, 74% (327/442) had used the app for more than a year, 80% (354/442) had tried premium subscription, 44% (194/442) used the app daily, and only 4% (18/442) used the app once a week or less. Only 41% (181/442) reported using Fitbit wearable device. Walking was the most commonly engaged activity tracked by the Fitbit app (274/442, 62%), followed by running (106/442, 24%) and biking (22/442, 5%).

XSL•FO
RenderX

Table . Usage patterns of the main study sample (N=442), based on an online survey of US-based Fitbit users engaging in regular exercise; source: authors' own work.

	Participants, % (N=442)
Frequency of Fitbit app use	
Daily	44
4 - 6 times a week	31
2 - 3 times a week	21
Once a week	2
Less than once a week	2
Duration of Fitbit app use	
Less than 1 month	0
Between 1 week and 1 month	0
1 - 3 months	4
4 - 6 months	9
7 - 9 months	5
10 - 12 months	8
More than 12 months	74
Duration of premium use	
Never	20
Up to 3 months or free trial	25
4 - 6 months	10
7 - 12 months	13
More than 12 months	32
Device used	
Phone	40
Fitbit wearable device	41
Smart watch (such as a Samsung watch and Apple Watch)	11
Laptop/desktop computer	8
Exercise engaged in	
Running	24
Walking	62
Biking	5
Swimming	1
Spinning	1
Hiking	3
Yoga	1
Other	3

Measurement Model Evaluation

To ensure that the observed variables in our survey accurately represent the underlying constructs in our model, we used SmartPLS 4 software (SmartPLS GmbH) and confirmatory factor analysis to assess internal reliability, convergent validity, discriminant validity, outer loadings, and model fit. As illustrated in Table 3, internal reliability was assessed using Cronbach α, with values above 0.8 indicating satisfactory reliability [45]. Convergent validity was measured using the average variance extracted (AVE), with acceptable values being 0.5 or higher [45]. Outer loadings results exceeded 0.7, confirming their contribution to their constructs. A few outer loadings were slightly lower (0.59 and above), but these were considered acceptable, especially considering the adequate scores of Cronbach α and AVE [45]. Thus, all items were retained in the model.

XSL•FO
RenderX

Table . Construct reliability, validity measures (Cronbach α and average variance extracted), and factor loadings for the main study constructs; source: authors' own work.

	Cronbach α	AVE[a]	Item loading				
Construct			1	2	3	4	5
Comparing (Affordance)	0.977	0.894	0.950	0.964	0.969	0.927	0.917
Competing (Affordance)	0.974	0.907	0.914	0.967	0.978	0.949	—[b]
Encourage (Affordance)	0.972	0.897	0.948	0.959	0.927	0.954	—
Guidance (Affordance)	0.911	0.728	0.889	0.924	0.857	0.730	—
Self-presentation (Affordance)	0.973	0.899	0.916	0.953	0.966	0.957	—
Recognize (Affordance)	0.977	0.914	0.941	0.952	0.955	0.975	—
Reminding (Affordance)	0.973	0.901	0.926	0.934	0.968	0.969	—
Rewards (Affordance)	0.965	0.861	0.877	0.894	0.969	0.968	—
Searching (Affordance)	0.940	0.798	0.885	0.879	0.908	0.901	—
Updating (Affordance)	0.857	0.610	0.681	0.714	0.884	0.827	—
Watching others (Affordance)	0.961	0.864	0.936	0.957	0.945	0.877	—
Agreeableness (Big Five)	0.870	0.619	0.882	0.844	0.687	0.717	—
Conscientiousness (Big Five)	0.843	0.584	0.675	0.595	0.887	0.861	—
Emotional stability (Big Five)	0.828	0.552	0.701	0.626	0.870	0.753	—
Extraversion (Big Five)	0.913	0.726	0.778	0.870	0.881	0.874	—
Openness (Big Five)	0.826	0.548	0.631	0.773	0.762	0.784	—

[a]AVE: average variance extracted.

[b]Not applicable.

Discriminant validity was evaluated using the Fornell and Larcker [48] criterion. The results, illustrated in Table 4, showed that the AVE values of each construct were higher than the correlations with other constructs, indicating satisfactory discriminant validity. Model fit indices also supported a good model fit: the comparative fit index was 0.929, Tucker-Lewis index was 0.922, root mean square error of approximation was 0.054, and standardized root mean square residual was 0.051, all within acceptable ranges [45].

XSL•FO
RenderX

Table . Discriminant validity assessment using the Fornell-Larcker criterion for main study constructs; source: authors' own work.

Construct	Agree	Compare	Compete	Cons	Extra	Emot	Openn	Present	Remind	Search	Encourage	Guidance	Recognize	Reward	Update	Watch
Agree	0.787	0.091	0.036	0.003	0.222	0.072	0.322	0.057	0.107	0.122	0.103	0.145	0.082	0.051	0.145	0.07
Compare	0.091	0.946	0.786	−0.041	0.201	−0.017	0.012	0.523	0.251	0.383	0.701	0.419	0.662	0.498	0.153	0.791
Compete	0.036	0.786	0.952	−0.061	0.181	−0.011	−0.004	0.512	0.161	0.388	0.764	0.38	0.716	0.535	0.099	0.797
Cons	0.003	−0.041	−0.061	0.764	0.003	0.401	0.135	0.064	0.017	0.137	−0.069	0.029	−0.064	−0.095	0.16	−0.053
Extra	0.222	0.201	0.181	0.003	0.852	0.27	0.237	0.138	0.05	0.072	0.206	0.124	0.167	0.079	0.023	0.147
Emot	0.072	−0.017	−0.011	0.401	0.27	0.743	0.208	0.05	−0.057	0.058	−0.046	0.077	−0.037	−0.051	0.049	−0.058
Openn	0.322	0.012	−0.004	0.135	0.237	0.208	0.74	0.016	0.066	0.074	−0.013	0.024	0.02	−0.046	0.2	0.001
Present	0.057	0.523	0.512	0.064	0.138	0.05	0.016	0.948	0.259	0.362	0.57	0.436	0.587	0.506	0.2	0.528
Remind	0.107	0.251	0.161	0.017	0.05	−0.057	0.066	0.259	0.949	0.297	0.316	0.405	0.256	0.28	0.443	0.258
Search	0.122	0.383	0.388	0.137	0.072	0.058	0.074	0.362	0.297	0.893	0.423	0.745	0.421	0.371	0.38	0.49
Encourage	0.103	0.701	0.764	−0.069	0.206	−0.046	−0.013	0.57	0.316	0.423	0.947	0.473	0.848	0.564	0.195	0.771
Guidance	0.145	0.419	0.38	0.029	0.124	0.077	0.024	0.436	0.405	0.745	0.473	0.853	0.443	0.451	0.386	0.507
Recognize	0.082	0.662	0.716	−0.064	0.167	−0.037	0.02	0.587	0.256	0.421	0.848	0.443	0.956	0.557	0.126	0.727
Reward	0.051	0.498	0.535	−0.095	0.079	−0.051	−0.046	0.506	0.28	0.371	0.564	0.451	0.557	0.928	0.119	0.548
Update	0.145	0.153	0.099	0.16	0.023	0.049	0.2	0.2	0.443	0.38	0.195	0.386	0.126	0.119	0.781	0.15
Watch	0.07	0.791	0.797	−0.053	0.147	−0.058	0.001	0.528	0.258	0.49	0.771	0.507	0.727	0.548	0.15	0.929

Hypothesis Testing

The model was initially evaluated using the coefficient of determination (R^2) to assess how well the independent variables explained the variance in different affordances related to fitness app usage. The R^2 values indicated that the model explained a small to moderate amount of variance across the different affordances: 1.7% for comparing, 0.5% for competing, 1.8% for encouragement, 3.1% for guidance, 3.5% for recognition, 0.7% for reminding, 1.0% for rewards, 0.8% for searching, 2.4% for self-presentation, 6.3% for updates, and 3.3% for watching others. These findings suggest that the included variables provide limited explanatory power for these constructs, highlighting the need for further exploration of additional factors that might influence user interaction with fitness apps.

The results are presented in detail in Figure 1.

XSL•FO
RenderX

Figure 1. Study model results showing the relationships between Big Five personality traits and fitness app affordance actualizations; source: authors' own work.

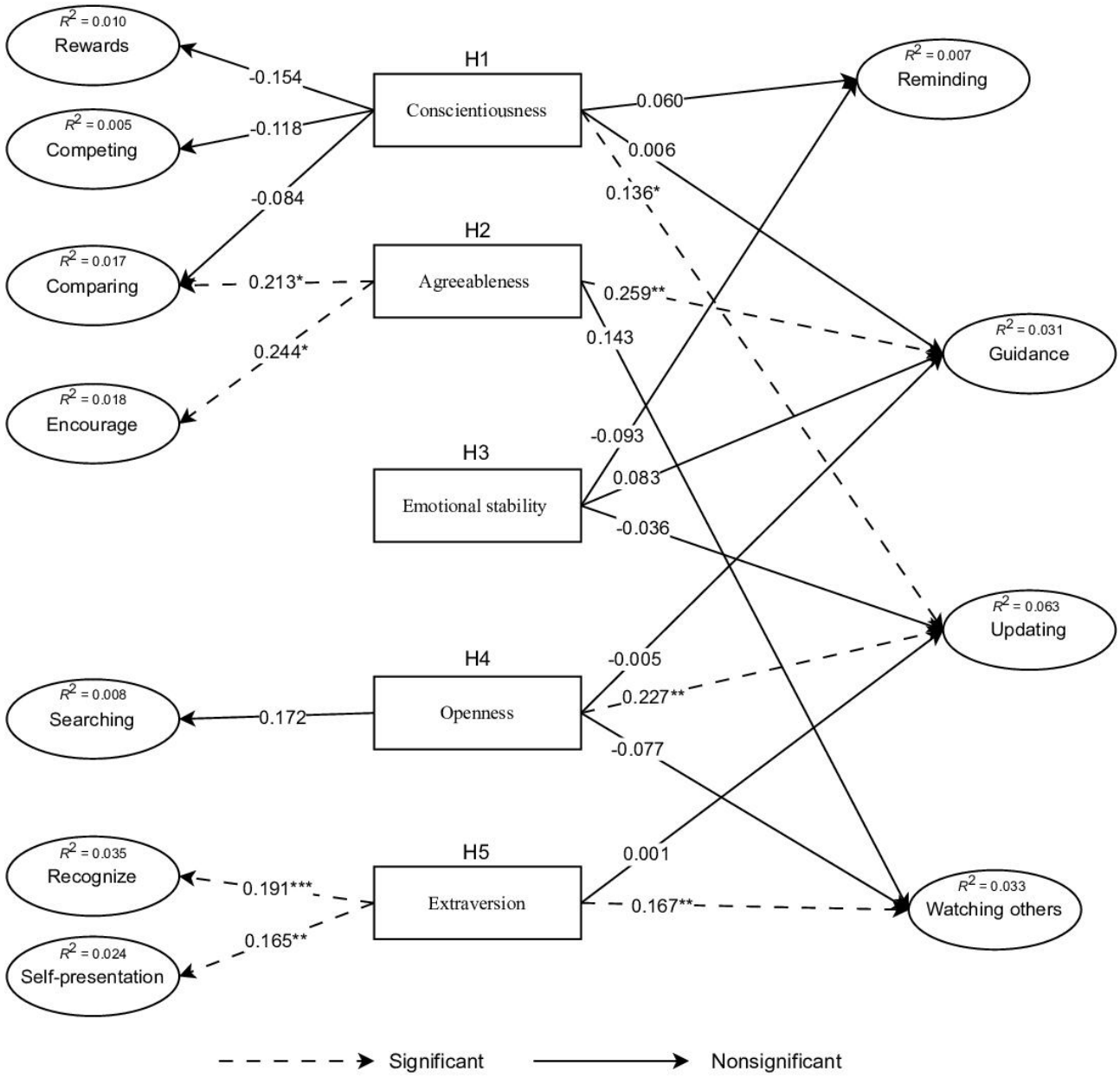

Conscientiousness significantly influenced updates (β=0.136, *P*=.01). The influence on rewards was not. Likewise, conscientiousness did not significantly influence competing, comparing, reminding, or guidance. While the hypotheses suggested a positive relationship between conscientiousness and these affordances, only update was shown to have a significant positive impact. Agreeableness had significant positive effects on guidance (β=0.259, *P*=.003), encouragement (β=0.244, *P*=.01), and comparing (β=0.213, *P*=.02). However, its effect on watching others was not significant. The significant positive results support the hypothesized relationship between agreeableness and the actualization of these affordances. Emotional stability has a nonsignificant effect on guidance, reminding, and updates. The results reject the hypothesized positive relationship between low emotional stability and actualizing these affordances. Openness had a significant positive effect on updates (β=0.227, *P*=.001). However, it did not significantly influence searching, watching others, or guidance. The results support the study hypothesis on the relationship between openness and actualizing update affordances and reject all the other hypotheses on the relationship between openness and actualizing the affordances. Extraversion showed significant positive effects on recognize (β=0.191, *P*<.001), self-presentation (β=0.165, *P*=.002), and watching others (β=0.167, *P*=.003). However, the effect on updates was not significant. The results support the study hypothesis on the relationship between extraversion and actualizing these affordances, except for the update affordance. The results are presented in Table 5.

XSL•FO
RenderX

Table . Results for the hypotheses; source: authors' own work.

Hypothesis[a]	Path coefficient	*P* value	Hypothesis support
H1: Conscientiousness – Guidance	0.006	0.98	Not supported
H1: Conscientiousness – Reminding	0.060	0.44	Not supported
H1: Conscientiousness – Updating	0.136	0.01	Supported
H1: Conscientiousness – Competing	−0.118	0.18	Not supported
H1: Conscientiousness – Comparing	−0.084	0.33	Not supported
H1: Conscientiousness – Rewards	−0.154	0.06	Not supported
H2: Agreeableness – Comparing	0.213	0.02	Supported
H2: Agreeableness – Guidance	0.259	0.003	Supported
H2: Agreeableness – Watching others	0.143	0.13	Not supported
H2: Agreeableness – Encourage	0.244	0.01	Supported
H3: Emotional Stability – Updating	−0.036	0.49	Not supported
H3: Emotional stability – Guidance	−0.083	0.30	Not supported
H3: Emotional stability – Reminding	−0.093	0.29	Not supported
H4: Openness – Searching	0.172	0.11	Not supported
H4: Openness – Watching others	−0.077	0.50	Not supported
H4: Openness – Guidance	−0.005	0.96	Not supported
H4: Openness – Updating	0.227	0.001	Supported
H5: Extraversion – Self-presentation	0.165	0.002	Supported
H5: Extraversion – Recognize	0.191	0.000	Supported
H5: Extraversion – Watching others	0.167	0.003	Supported
H5: Extraversion – Updating	0.001	0.98	Not supported

[a]Model fit: CFI=.929, TLI=.922, RSMEA=.054, SRMR=.051.

Discussion

Principal Findings

Based on the affordance actualization theory [12,49] and empirical research on the impact of personality traits on behavior [25,38], this study suggests that users' personality traits play a role in the actualization of affordances. The findings of the study support the connection between personality traits and affordance actualization, providing valuable insights. While not all predicted relationships were confirmed, each personality trait influenced the actualization of specific affordances, confirming some of the hypotheses. Specifically, agreeableness has a positive impact on the actualization of comparing, guidance, and encouragement affordances; extraversion influences recognition, self-presentation, and watching others; and conscientiousness and openness affect updating.

Theoretical Contribution

This research contributes to the existing literature on affordance theory. Previous studies have found that people's perceptions and actualizations of an object's affordances are influenced by factors such as ability, age, social context, preference, situation, policies, procedures, and rules [10,12,50]. Furthermore, this study emphasizes the role of personality traits as a crucial factor that impacts users' perceptions and actualizations of affordances.

This study also contributes to the literature on user interaction with technology functionality and design [19,51]. The findings highlight that personality traits extend beyond technology functionality to affect how users perceive and actualize what these technologies afford to them. In personality research, conscientious individuals are described as organized, efficient, and goal-oriented [21]. Previous studies have shown that conscientious individuals are more likely to use technologies that support task completion and collaboration, especially when these technologies provide reminders to perform specific actions [15,22]. Thus, conscientious users will be more likely to actualize affordances that support goal achievement and self-monitoring, such as updating, guidance, reminding, competing, comparing, and rewards. Our results indicated that among the hypothesized affordances, only updating was significantly actualized by conscientious users. Conscientious users actualize updating affordances using Fitbit to stay on track with their fitness goals, reflecting their organized and disciplined nature.

The absence of actualization of other affordances among conscientious users could be due to user preferences for specific

XSL•FO
RenderX

Fitbit features. For example, some users may prefer alternative features to actualize guidance instead of the features offered by Fitbit. Therefore, focusing solely on Fitbit may limit the variety of features that some users prefer over those provided by Fitbit to actualize these affordances. Individuals who score high in agreeableness are described as being cooperative, considerate, and good-natured [21]. Previous research suggests that agreeable individuals are more likely to use technologies that support collaboration and social interaction [25]. For instance, agreeable students find university bulletin boards and collaborative classroom systems especially useful [25,26]. Moreover, a study on gamification apps indicated a positive correlation between agreeableness and engagement in challenges with others [19]. In the context of fitness apps, we predict that agreeable users will actualize affordances that promote social interaction and collaboration. Our study corroborates these findings, demonstrating that agreeable users effectively actualize comparison, guidance, and encouragement affordances, underscoring their inclination toward seeking and appreciating social interactions and collaborative activities in their fitness app use.

People who are high in openness to experience tend to be imaginative, curious, and open-minded [21]. Studies suggest that individuals with high levels of openness are more likely to use authoritative strategies to achieve their fitness goals [23].

Our research indicates that openness affects the actualization of updating affordance in fitness apps, but not guidance affordance. This aligns partly with existing literature showing that open individuals prefer helpful strategies but may not necessarily seek guidance. We also expected that people high in openness would be more inclined to seek new information to improve their fitness. However, our results did not support this assumption. The lack of interest in seeking new information may be due to how this feature is presented. Often, the available information is shared by users in posts, which, although useful, may lack reliable sources, making the information less trustworthy despite its potential accuracy.

Individuals high in extraversion are sociable, outgoing, and energetic [21]. Previous research suggests that people who are extraverted and use social media platforms such as Facebook tend to be more socially active. They frequently upload photos, update their statuses, and have larger social networks [42]. Their social nature motivates them to continue using social media [43]. Based on this, we hypothesized that extraversion has a positive influence on social aspects of apps, such as recognition, self-presentation, and observing others' actions in fitness apps. Our results support this hypothesis by indicating that extraverted users are likely to seek social interaction and recognition, using the app to share their achievements and follow others' activities.

People with low emotional stability are prone to stress, anxiety, and emotional fluctuations [21]. Previous studies have found that these individuals are less likely to accept and actively use technology. In this study, we focused on current Fitbit users who have voluntarily accepted and used the device, likely motivated by health concerns. Therefore, we hypothesized that these users are primarily driven by health improvement rather than curiosity, enjoyment, collaboration, or external motivations. We expected them to use fitness app features related to health, such as updates, guidance, and reminders, to monitor their health status, perform exercises correctly, and maintain regular exercise routines. However, our results did not show any influence on the expected features. This lack of significant findings might be due to the characteristics of our study sample. A homogeneous sample with slight variation in personality traits may not reveal significant relationships, whereas a more diverse sample might highlight stronger influences.

Practical Contribution

The study findings also have practical implications for the design of fitness apps. Since personality traits influence users' actualization of an app's affordances, it becomes imperative to consider the potential implications of user personality traits when designing and implementing affordances in fitness apps. Considering the results, removing certain features that facilitate actualizing some undesirable affordances for users with particular personality traits, such as conscientious users, could be a viable strategy to prevent the possible negative actualization of some affordances. For example, the challenges feature in the Fitbit app can play a prominent role in actualizing reward affordances. By excluding this feature for users with the conscientiousness traits, conscientious users would not be bothered with less-needed affordances.

Limitation and Future Research

This study is focused on characterizing users based on the Big Five personality traits, to gain an abstract understanding of users' motives in using specific affordances. Future research could expand user characterization by considering additional factors such as culture, economy, situation, circumstances, age, and gender to achieve a more precise understanding of how users interact with fitness apps.

The study primarily examined the context of Fitbit; however, it is important to acknowledge that different fitness apps may have distinct designs and feature sets. These differences in technology design can influence users' perceptions of technology affordances. Therefore, future research should explore a broader range of fitness apps, considering the differences in their technological designs and assessing how these variations impact users' actions.

Conclusion

This study investigates the influence of users' personality traits on affordance actualization within the context of fitness apps. It highlights how users' characterizations (personality traits) shape their interactions with technology, leading them to actualize various affordances based on these traits. These findings enhance our understanding of the factors impacting IT interaction and provide valuable insights for designing user-centered technologies that promote greater user interaction.

XSL•FO
RenderX

Acknowledgments

This work was supported and funded by the Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University (IMSIU) (grant number IMSIU-DDRSP2501).

Disclaimer

Artificial intelligence tools were used exclusively to improve the English language and enhance readability.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to institutional review board approval but are available from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Survey items used to measure Big Five personality traits using the prompt "Indicate how accurately each statement describes you" and responses on a 5-point Likert Scale (1=very inaccurate, 5=very accurate); source: authors' own work.
[PDF File, 56 KB - formative_v9i1e72691_app1.pdf]

Multimedia Appendix 2
Survey items used to measure fitness app affordances using the prompt "I use Fitbit for [user's selected exercise, eg, walking, is inserted here] to" and responses on a 5-point Likert Scale (1=strongly disagree to 5=strongly agree); source: authors' own work.
[PDF File, 69 KB - formative_v9i1e72691_app2.pdf]

Multimedia Appendix 3
Cronbach α scores for the pilot study; source: authors' own work.
[PDF File, 133 KB - formative_v9i1e72691_app3.pdf]

References

1. Vaghefi I, Tulu B. The continued use of mobile health apps: insights from a longitudinal study. JMIR mHealth uHealth 2019 Aug 29;7(8):e12983. [doi: 10.2196/12983] [Medline: 31469081]
2. Physical activity. World Health Organization. 2020. URL: https://www.who.int/news-room/fact-sheets/detail/physical-activity [accessed 2025-08-29]
3. Sousa Basto P, Ferreira P. Mobile applications, physical activity, and health promotion. BMC Health Serv Res 2025 Mar 10;25(1):359. [doi: 10.1186/s12913-025-12489-z] [Medline: 40065339]
4. Chiu W, Cho H, Chi CG. Consumers' continuance intention to use fitness and health apps: an integration of the expectation–confirmation model and investment model. Inf Technol People 2021 May 24;34(3):978-998. [doi: 10.1108/ITP-09-2019-0463]
5. Alshawmar M, Tulu B, Hall-Phillips A. Influence of personality traits on the continued use of fitness apps. 2022 Presented at: Hawaii International Conference on System Sciences; Jan 4-7, 2022; Online. [doi: 10.24251/HICSS.2022.478]
6. Alshawmar M, Mombini H, Tulu B, Vaghefi I. Investigating the affordances of wellness mHealth apps. 2021 Presented at: Hawaii International Conference on System Sciences; Jan 4-9, 2021; Online. [doi: 10.24251/HICSS.2021.462] [Medline: 34804627]
7. James TL, Wallace L, Deane JK. Using organismic integration theory to explore the associations between users' exercise motivations and fitness technology feature set use. MIS Q 2019 Mar 1;43(1):287-312. [doi: 10.25300/MISQ/2019/14128]
8. Rockmann R, Gewald H. Activity tracking affordances: identification and instrument development. 2018 Presented at: PACIS 2018 Proceedings; Jun 26-30, 2018; Yokohama, Japan p. 232.
9. James TL, Deane JK, Wallace L. An application of goal content theory to examine how desired exercise outcomes impact fitness technology feature set selection. Inf Syst J 2019 Sep;29(5):1010-1039. [doi: 10.1111/isj.12233]
10. Alshawmar M. A review of the applications of affordance theory in mhealth app research. 2021 Presented at: Hawaii International Conference on System Sciences; Jan 4-9, 2021; Online. [doi: 10.24251/HICSS.2021.437] [Medline: 34804627]
11. Leonardi PM. When does technology use enable network change in organizations? A comparative study of feature use and shared affordances. MIS Q 2013 Sep 1;37(3):749-775. [doi: 10.25300/MISQ/2013/37.3.04]
12. Strong D, Volkoff O, Johnson S, et al. A theory of organization-EHR affordance actualization. J Assoc Inf Syst 2014;15(2):53-85. [doi: 10.17705/1jais.00353]

XSL•FO
RenderX

13. Suh A, Li M. How the use of mobile fitness technology influences older adults' physical and psychological well-being. Comput Human Behav 2022 Jun;131:107205. [doi: 10.1016/j.chb.2022.107205]
14. Roberts BW, O'Donnell M, Robins RW. Goal and personality trait development in emerging adulthood. J Pers Soc Psychol 2004 Oct;87(4):541-550. [doi: 10.1037/0022-3514.87.4.541] [Medline: 15491277]
15. Cruz-Maya A, Tapus A. Influence of user's personality on task execution when reminded by a robot. 2016 Presented at: 8th International Conference, ICSR 2016; Nov 1-3, 2016; Kansas City, MO. [doi: 10.1007/978-3-319-47437-3_81]
16. Mouakket S. The role of personality traits in motivating users' continuance intention towards Facebook: gender differences. J High Technol Manag Res 2018;29(1):124-140. [doi: 10.1016/j.hitech.2016.10.003]
17. Digman JM. Personality structure: emergence of the five-factor model. Annu Rev Psychol 1990 Jan;41(1):417-440. [doi: 10.1146/annurev.ps.41.020190.002221]
18. Kim S, Park JY, Chung K. The relationship between the Big Five personality traits and the Theory of Planned Behavior in using mindfulness mobile apps: cross-sectional survey. J Med Internet Res 2022 Nov 30;24(11):e39501. [doi: 10.2196/39501] [Medline: 36449344]
19. Jia Y, Xu B, Karanam Y, Voida S. Personality-targeted gamification: a survey study on personality traits and motivational affordances. 2016 Presented at: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems; May 7-12, 2016; San Jose, CA. [doi: 10.1145/2858036.2858515]
20. Weik L, Fehring L, Mortsiefer A, Meister S. Big 5 personality traits and individual- and practice-related characteristics as influencing factors of digital maturity in general practices: quantitative web-based survey study. J Med Internet Res 2024 Jan 22;26:e52085. [doi: 10.2196/52085] [Medline: 38252468]
21. Costa PT, McCrae RR, Dye DA. Facet scales for agreeableness and conscientiousness: a revision of the NEO personality inventory. Pers Individ Dif 1991;12(9):887-898. [doi: 10.1016/0191-8869(91)90177-D]
22. McDowell CP, Wilson KE, Monroe DC, McCrory C, Kenny RA, Herring MP. Physical activity partially mediates associations between "Big" personality traits and incident generalized anxiety disorder: findings from the Irish longitudinal study on ageing. J Affect Disord 2020 Dec 1;277:46-52. [doi: 10.1016/j.jad.2020.07.124] [Medline: 32795714]
23. Halko S, Kientz JA. Personality and persuasive technology: an exploratory study on health-promoting mobile applications. 2010 Presented at: International Conference on Persuasive Technology; Jun 7-10, 2010; Copenhagen, Denmark. [doi: 10.1007/978-3-642-13226-1_16]
24. Barrick MR, Mount MK, Judge TA. Personality and performance at the beginning of the new millennium: what do we know and where do we go next? Int J Selection Assessment 2001 Mar;9(1-2):9-30. [doi: 10.1111/1468-2389.00160]
25. Devaraj S, Easley RF, Crant JM. Research note —how does personality matter? Relating the five-factor model to technology acceptance and use. Inf Syst Res 2008 Mar;19(1):93-105. [doi: 10.1287/isre.1070.0153]
26. Lin MYC, Ong CS. Understanding information systems continuance intention: a five-factor model of personality perspective. 2010 Presented at: .
27. Anthony LM, Clarke MC, Anderson SJ. Technophobia and personality subtypes in a sample of South African university students. Comput Human Behav 2000 Jan;16(1):31-44. [doi: 10.1016/S0747-5632(99)00050-3]
28. Carver CS, Connor-Smith J. Personality and coping. Annu Rev Psychol 2010;61:679-704. [doi: 10.1146/annurev.psych.093008.100352] [Medline: 19572784]
29. Smith CA, Organ DW, Near JP. Organizational citizenship behavior: its nature and antecedents. J Appl Psychol 1983;68(4):653-663. [doi: 10.1037/0021-9010.68.4.653]
30. Assadi V, Hassanein K. Continuance intention to use high maintenance information systems: the role of perceived maintenance effort. 2010 Presented at: ECIS 2010 Proceedings; Jun 6-9, 2010; Pretoria, South Africa.
31. Seibert SE, Kraimer ML. The five-factor model of personality and career success. J Vocat Behav 2001 Feb;58(1):1-21. [doi: 10.1006/jvbe.2000.1757]
32. Friedman HS. Long-term relations of personality and health: dynamisms, mechanisms, tropisms. J Pers 2000 Dec;68(6):1089-1107. [doi: 10.1111/1467-6494.00127] [Medline: 11130733]
33. Vollrath M, Torgersen S. Who takes health risks? A probe into eight personality types. Pers Individ Dif 2002 May;32(7):1185-1197. [doi: 10.1016/S0191-8869(01)00080-0]
34. Weston SJ, Jackson JJ. The role of vigilance in the relationship between neuroticism and health: a registered report. J Res Pers 2018 Apr;73:27-34. [doi: 10.1016/j.jrp.2017.10.005] [Medline: 32394989]
35. Donnellan MB, Oswald FL, Baird BM, Lucas RE. The mini-IPIP scales: tiny-yet-effective measures of the Big Five factors of personality. Psychol Assess 2006 Jun;18(2):192-203. [doi: 10.1037/1040-3590.18.2.192] [Medline: 16768595]
36. Ceci L. Leading health and fitness apps in the US 2018. Statista. 2022. URL: https://tinyurl.com/32yazws9 [accessed 2025-08-29]
37. Breil B, Kremer L, Hennemann S, Apolinário-Hagen J. Acceptance of mHealth apps for self-management among people with hypertension. Stud Health Technol Inform 2019 Sep 3;267:282-288. [doi: 10.3233/SHTI190839] [Medline: 31483283]
38. Goldberg LR. The development of markers for the Big-Five factor structure. Psychol Assess 1992;4(1):26-42. [doi: 10.1037/1040-3590.4.1.26]
39. Prabhu V, Sutton C, Sauser W. Creativity and certain personality traits: understanding the mediating effect of intrinsic motivation. Creat Res J 2008 Feb 6;20(1):53-66. [doi: 10.1080/10400410701841955]

XSL•FO
RenderX

40. Su J, Dugas M, Guo X, Gao GG. Influence of personality on mHealth use in patients with diabetes: prospective pilot study. JMIR mHealth uHealth 2020 Aug 10;8(8):e17709. [doi: 10.2196/17709] [Medline: 32773382]
41. Mikolasek M, Witt CM, Barth J. Adherence to a mindfulness and relaxation self-care app for cancer patients: mixed-methods feasibility study. JMIR mHealth uHealth 2018 Dec 6;6(12):e11271. [doi: 10.2196/11271] [Medline: 30522990]
42. Lee E, Ahn J, Kim YJ. Personality traits and self-presentation at Facebook. Pers Individ Dif 2014 Oct;69:162-167. [doi: 10.1016/j.paid.2014.05.020]
43. Deng S, Liu Y, Li H, Hu F. How does personality matter? An investigation of the impact of extraversion on individuals' SNS use. Cyberpsychol Behav Soc Netw 2013 Aug;16(8):575-581. [doi: 10.1089/cyber.2012.0383] [Medline: 23697532]
44. Eysenbach G. Improving the quality of web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34. [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
45. Hair JF, Ringle CM, Sarstedt M. PLS-SEM: indeed a silver bullet. J Mark Theory Pract 2011 Apr;19(2):139-152. [doi: 10.2753/MTP1069-6679190202]
46. Gaskin J. Causal models. Gaskination's StatWiki. 2016. URL: https://statwiki.gaskination.com/index.php/Causal_Models [accessed 2025-08-29]
47. Peer E, Brandimarte L, Samat S, Acquisti A. Beyond the Turk: alternative platforms for crowdsourcing behavioral research. J Exp Soc Psychol 2017 May;70:153-163. [doi: 10.1016/j.jesp.2017.01.006]
48. Fornell C, Larcker DF. Evaluating structural equation models with unobservable variables and measurement error. J Mark Res 1981 Feb;18(1):39-50. [doi: 10.1177/002224378101800104]
49. McGrenere J, Ho W. Affordances: clarifying and evolving a concept. 2000 Presented at: Graphics interface 2000; May 15-17, 2000; Montréal, QC URL: https://www.researchgate.net/publication/221474920_Affordances_Clarifying_and_Evolving_a_Concep [accessed 2025-09-09]
50. Kappen DL, Mirza-Babaei P, Nacke LE. Gamification through the application of motivational affordances for physical activity technology. 2017 Oct 15 Presented at: Proceedings of the Annual Symposium on Computer-Human Interaction in Play; Amsterdam, Netherlands. [doi: 10.1145/3116595.3116604]
51. Benz C, Riefle L, Satzger G, Karlsruhe Service Research and Innovation Hub, Institute of Information Systems and Marketing, Karlsruhe Institute of Technology. User engagement and beyond: a conceptual framework for engagement in information systems research. Commun Assoc Inf Sys 2024;54(1):331-359. [doi: 10.17705/1CAIS.05412]

Abbreviations

AVE: average variance extracted
CHERRIES: Checklist for Reporting Results of Internet E-Surveys

Edited by A Mavragani; submitted 15.02.25; peer-reviewed by M Fraser; revised version received 13.08.25; accepted 14.08.25; published 12.09.25.

Please cite as:
Alshawmar M, Tulu B, Wilson EV, Hall-Phillips A
Personality-Driven Variations in Fitness App Affordance Actualization Among Adults: Quantitative Survey Study
JMIR Form Res 2025;9:e72691
URL: https://formative.jmir.org/2025/1/e72691
doi:10.2196/72691

© Moayad Alshawmar, Bengisu Tulu, E Vance Wilson, Adrienne Hall-Phillips. Originally published in JMIR Formative Research (https://formative.jmir.org), 12.9.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Design of an Automated Mobile Phone-Based Reminder and Incentive System: Application in a Quasi-Randomized Controlled Trial to Improve the Timeliness of Childhood Vaccinations in Tanzania

Marco van Zwetselaar[1*], MSc; Jan Ostermann[2*], PhD; Melkiory Beti[3], BBIT; Joy Noel Baumgartner[4], MSSW, PhD; Sayoki Mfinanga[5], MD, PhD; Esther Ngadaya[5], MD, PhD; Lavanya Vasudevan[6], MPH, CPH, PhD; Nathan Thielman[7], MD, MPH

[1]Zwets IT, Harskamp, The Netherlands

[2]Department of Health Services Policy and Management, Arnold School of Public Health, University of South Carolina, 915 Greene St, Columbia, SC, United States

[3]Kilimanjaro Clinical Research Institute, Moshi, United Republic of Tanzania

[4]School of Social Work, University North Carolina, Chapel Hill, NC, United States

[5]National Institute for Medical Research, Dar es Salaam, United Republic of Tanzania

[6]Hubert Department of Global Health, Rollins School of Public Health, Emory University, Atlanta, GA, United States

[7]Department of Medicine, Duke Global Health Institute, Duke University, Durham, NC, United States

[*]these authors contributed equally

Corresponding Author:
Jan Ostermann, PhD
Department of Health Services Policy and Management, Arnold School of Public Health, University of South Carolina, 915 Greene St, Columbia, SC, United States

Abstract

Background: The global penetration of mobile phones has offered novel opportunities for communicating health-related information to individuals. A low-cost system that facilitates autonomous communication with individuals via mobile phones holds potential for expanding the reach of health messaging in settings with human resource and infrastructure limitations.

Objective: We sought to design a flexible, low-code system using open-source software that could be adapted to different contexts and technical environments and accommodate a wide range of automation needs. We report on key details of the mobile phone–based appointment reminder and incentive system (mParis), document its use, review implementation challenges and adaptations to address these challenges in the context of a quasi-randomized trial of mobile phone–based reminders and incentives as means of increasing the timeliness of childhood vaccinations in Tanzania, and outline other use cases that highlight the versatility of the system.

Methods: The mParis instance described in this paper, which is hosted in Tanzania, sent automated, individualized vaccination reminders in the form of SMS text messages to the mobile phones of mothers of young children. Process workflows, based on the national vaccination schedule of Tanzania, were programmed into mParis. Reminders for vaccinations due at ages 6, 10, and 14 weeks were sent 7 days and 1 day before and 14 days after each vaccination due date. A subset of messages included financial incentive offers to mothers for the timely vaccination of their children. We report on implementation outcomes, challenges, and adaptations to address these challenges.

Results: Between August and December 2017, a total of 412 pregnant women were enrolled in the trial. After mothers reported the birth of their children, individualized vaccination reminder messages were sent for vaccination due dates between January and July 2018. From March 2018, messages contained financial incentive offers. Of 1397 messages sent, 1122 (80.3%) messages were recorded as delivered, 249 (18.8%) as expired and resent; 23 (1.6%) as failed, and 3 (0.2%) as sent but lacking a delivery confirmation. In total, 633 (45.3%) messages contained incentive offers. Of 173 women who received at least 1 message, 67 (38.7%) were sent reminders only; 106 (61.3%) women were sent at least 1 incentivized message. Numerous challenges were encountered during the system's implementation, despite its deliberate design to accommodate basic problems, such as intermittent internet access and power failures. Continuous adaptation to increase the resilience of the system resulted in a successful deployment.

Conclusions: mParis' open-source nature, auditability, and ability to autonomously execute algorithms in a low-resource setting with frequent infrastructure challenges suggest favorable prospects to automate health communication in a wide range of settings.

XSL•FO
RenderX

mParis' use in other applications, including enrollment and follow-up for health-related research studies, demonstrates its versatility and ability to accommodate diverse challenges that may be encountered.

Trial Registration: ClinicalTrials.gov NCT03252288; https://clinicaltrials.gov/study/NCT03252288

International Registered Report Identifier (IRRID): RR2-10.1186/s13063-019-3430-4

(JMIR Form Res 2025;9:e65150) doi:10.2196/65150

KEYWORDS

mHealth; automation; childhood vaccinations; SMS; Tanzania; open-source software; vaccination timeliness; digital health intervention; formative research; mobile phones

Introduction

The global, nearly universal penetration of mobile phones has offered novel opportunities for engaging and retaining individuals in health-related interventions via electronic communication [1]. However, the large digital divide [2] between high- and low-resource settings has resulted in engagement and retention approaches that tend to either depend on robust and reliable technology infrastructure that does not exist in low-resource settings, or on costly human resources that are not scalable in high-resource settings [3]. Open-source systems that autonomously and flexibly execute algorithms even in the face of unreliable infrastructure, and that depend on human resources only to back-stop residual gaps, hold great potential for bridging the digital divide and are key to sustainability and scale-up [4,5].

The World Health Organization (WHO) global guidelines on digital interventions for health system strengthening recommend targeted client communication strategies using mobile phone communication modalities, such as text messaging [6]. In a systematic review [7] conducted to inform the guidelines, evidence was identified to support the use of targeted client communication in diverse health behaviors [8], including use of oral contraception, adherence to antiretroviral medication, and for supporting childhood vaccinations. In parallel, several challenges were identified related to equity in reach of mobile phone–based strategies, including gaps in mobile network coverage, lower mobile phone ownership among women, and gaps in literacy and digital health literacy. Combining targeted client communication strategies with digital tracking and clinical decision support may be beneficial for improving quality of care through better data documentation and use, although this may require greater resources to implement [9]. Thoughtful design with attention to scalability and equity is necessary to maximize the reach of mobile phone–based strategies to vulnerable populations, including women and those living in rural and remote areas.

Mobile phone–based strategies to support childhood vaccinations have seen a steady interest, with evidence for text message–based interventions dating back over a decade [10-13]. A systematic review by Mekonnen et al [14] found that SMS reminders significantly improved childhood vaccination coverage, demonstrating their effectiveness in diverse low- and middle-income country contexts. In Guatemala, Domek et al [15] conducted a randomized controlled trial showing that SMS reminders reduced delays in infant immunization visits and were well received by caregivers. Such digital strategies support the goals of the WHO Immunization Agenda 2030, which advocates for leveraging digital innovations to improve equitable vaccine coverage and strengthen health systems globally.

In previous research in Tanzania, we identified 3 primary challenges with the implementation of vaccination programs in rural settings: gaps in knowledge of vaccination recommendations, geographic barriers (lack of transportation options and costs of transportation) to accessing vaccination services, and gaps in cold chain infrastructure affecting vaccination service delivery [16]. In response to these challenges, we sought to examine the use of sending automated text messages that could address 2 of the 3 challenges: we sought to use personalized text messages to send reminders about upcoming vaccinations and deliver conditional incentives for timely vaccination to overcome geographic barriers [17].

The objectives of this study are to provide details on the formation, design, functionality, and technical features of the mobile phone–based appointment reminder and incentive system (mParis), describe the design and implementation of mParis to support timely childhood vaccinations in Tanzania, review implementation challenges and subsequent adaptations to address these challenges, and outline other use cases that highlight the versatility of the system to support research and intervention studies in both high- and low-resource settings.

Methods

Overview

We report here key details of mParis in accordance with guidelines for the reporting of health interventions using mobile phones: mERA (mHealth evidence reporting and assessment) checklist (Checklist 1) as well as the TiDieR (Template for Intervention Description and Replication) (Checklist 2) [18,19]. We illustrate the system's versatility by describing its use in a quasi-randomized trial of mobile phone–based reminders and incentives as means of increasing the timeliness of routine childhood vaccinations in Tanzania [17]. The design of the system was guided by the study protocol and the national vaccination schedule in Tanzania. The process-driven system is particularly suitable to this application because it tracks the mother's journey through the vaccination schedule of her child, which may include missed or delayed vaccinations that result in gaps in the vaccination schedule. We briefly discuss the broader applicability of the system to any scheduled events, especially when conditionality results in alternative process flows (eg, due to maternal or child death), and to other use cases,

XSL•FO
RenderX

such as engaging and retaining participants in research studies. A parallel instance of mParis was used to send reminder messages to individuals participating in a research study focusing on HIV testing [20], and a third application of mParis currently delivers confidential, mobile phone–based HIV testing invitations to social network contacts of individuals seeking HIV testing or treatment [21].

Technical Specifications of mParis

Software

mParis is a task scheduling, execution, and event tracking application built on Flowable [22], an open-source, Java-based workflow engine. The scheduling and execution of tasks are governed by intuitive, well-defined process definitions. Figure 1 illustrates this concept by showing a process definition for sending an SMS text message via mParis. The "Send SMS" process sends an SMS text message based on parameters that are specified by a parent process. In this example, the parent process prescribes the time at which the message should be sent, the recipient's phone number, and the message text. After the SMS text message is sent, the process awaits an acknowledgement from the recipient's SMS center, which can be either a confirmation that the message was sent, a delivery confirmation, or a send failure notification. Depending on the acknowledgment received, the process retries a failed message or continues to wait for a delivery confirmation. Finally, the process reports the status of the SMS text message (delivered, failed, sent, and unclear) to the parent process.

Figure 1 illustrates this concept by showing a process definition for sending an SMS text message via mParis. The "Send SMS" process sends a message based on parameters that are specified by a parent process. In this example, the parent process prescribes the time at which the SMS text message should be sent, the recipient's phone number, and the message text. After the message is sent, the process awaits an acknowledgement from the recipient's SMS center, which can be either a confirmation that the message was sent, a delivery confirmation, or a send failure notification. Depending on the acknowledgment received, the process retries a failed SMS text message or continues to wait for a delivery confirmation. Finally, the process reports the status of the SMS text message (delivered, failed, sent, and unclear) to the parent process.

Figure 1. Process definition for the "Send SMS" process within mParis. This process diagram describes the sending of an SMS text message and results in either a "failed," "delivered," "sent," or "unclear" status of the message. The process is started by a parent process. "x" indicates "or."

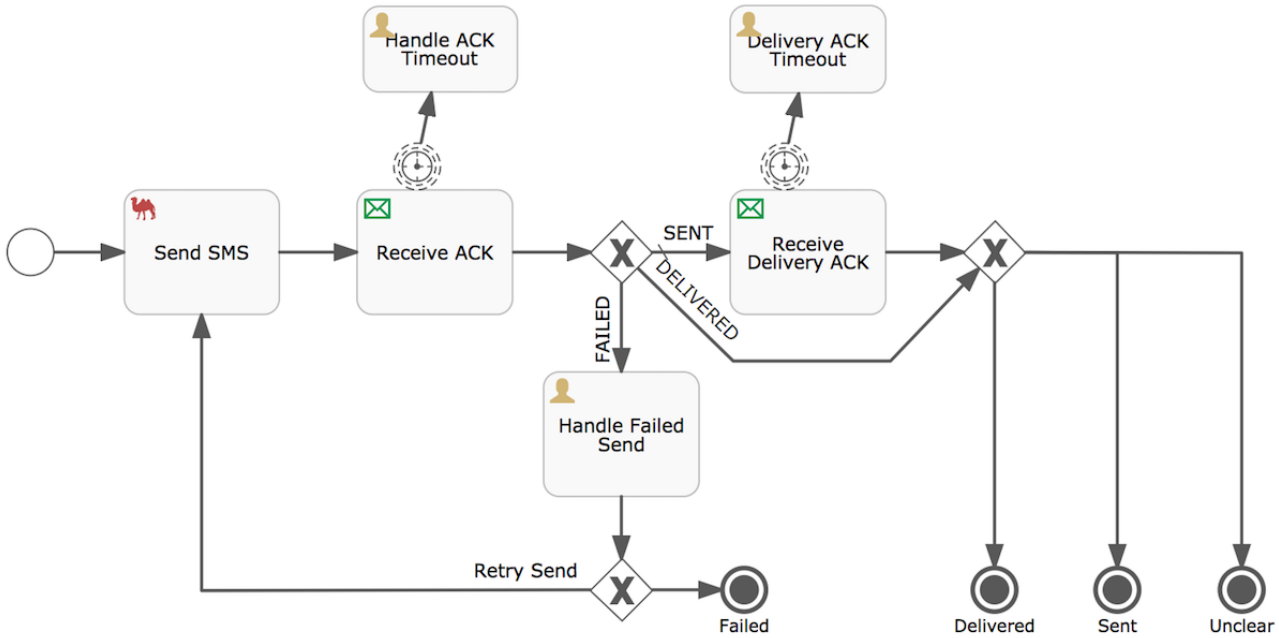

Processes may trigger tasks, await task completion, or both. Tasks can be either robotic or human. Robotic tasks include, for example, the sending of SMS text messages or email messages, or the generation of reports. Human tasks include, for example, the handling of failure events, phone calls, or the completion of forms. Human tasks can be assigned to and claimed by users with prespecified roles and permissions. The system integrates events (eg, incoming messages) into the processes and tracks task completion.

Notably, Flowable is a relatively code-free ("low-code") environment: process modeling can be done in an intuitive visual modeling environment using the standardized Business Process Modeling Notation (BPMN; version 2.0). Models execute directly in the Flowable framework. Depending on the application, front-end user screens (forms) may be added for user interactions. Processes are readily integrated with diverse back-ends, such as an SMS or email server. Data persistence is managed by the framework, which stores the values of process variables as they change throughout a process execution. Thus, there is no need to develop an underlying database.

Technology

Details of the technology comprising mParis are summarized in Table 1.

XSL•FO
RenderX

Table . Technology comprising the mParis instance used in a quasi-randomized trial of automated mobile phone–based vaccination reminders and incentive offers to mothers of newborn children (Tanzania, 2018).

Component	Details
Workflow engine	Flowable 6.3
Web proxy	Apache/2.4.7 and HTTPD server: https to public internet (Apache Software Foundation)
Application server	Apache Tomcat 8.5: mParis app, Flowable runtime framework and process modeling environment
Database	PostgreSQL 10 (PostgreSQL Global Development Group; process model and process instance storage)
Message broker	ActiveMQ 5.14.1 (Apache Software Foundation; guaranteed once-and-only-once event delivery between application server and SMS server)
SMS server	SMSTools3 using serial USB dongles for multiple providers
SIM card dongles	Airtel Tanzania SIM card
Redundancy	Redundant network providers, Uninterruptible power supply (UPS), offsite encrypted backup

Testing

Flowable facilitates testing of processes using unit tests and simulation, with support for accelerated timelines and mock events. mParis testing included, for example, the temporary removal of time-outs, and the use of keywords in mock SMS text messages to simulate specific events (eg, message failures, incoming messages, and missing delivery confirmations) or to trigger activities (eg, human tasks). A test server was set up to facilitate testing of the system from the perspective of all users. Reporting templates were developed to facilitate process evaluation in near real-time.

Security, Access, and Auditability

The instance of mParis described in this paper is hosted in the data center of the Kilimanjaro Clinical Research Institute (KCRI) in Moshi, Tanzania. Servers are hosted in server rooms with controlled physical access. Back-up to encrypted secondary (off-site) storage occurs daily. Another instance of mParis is hosted on a virtual server at Tanzania's National Internet Data Center [23]. mParis can be accessed remotely over authenticated and encrypted network connections.

User roles are defined in the process modeler and assigned to user accounts. Roles determine the processes the user can initiate, and the tasks they can select from "team inboxes." Users with administrator roles control account and role assignment. All events, including every change to process data, are fully auditable.

Cost

Flowable is open-source software freely available under the Apache License (version 2.0; Apache Software Foundation). Flowable's low-code development allows for system adaptation and maintenance with limited programming. With the purchase of monthly SMS text message bundles, mParis can send large numbers of messages at minimal cost to any Tanzanian mobile phone number. Because SMS text message receipt is free in Tanzania, no cost is incurred by message recipients.

Application: Reminders and Incentives for Timely Vaccinations

The first application of mParis was to support a quasi-randomized controlled trial in Mtwara Region in southern Tanzania; the study protocol and the characteristics of study participants, including their mobile phone use, have been previously described [17,24]. In this formative research study, we provide details on the implementation and adaptation of mParis in relation to the objectives of the trial, as well as the system's functionality and extended use cases.

Study Context

In 2023, 21 million infants missed out on one or more life-saving vaccines [25], with the majority of "zero-dose" children living in the African region [26]. In Tanzania, only 53% of children aged 12 - 23 months are estimated to be fully vaccinated against all basic antigens [27], and rates of vaccination timeliness are low [28,29]. Strategies that leverage digital health technologies (eg, vaccination appointment reminders) [14] have been shown to be effective at increasing vaccination coverage and timeliness. Our quasi-randomized trial aimed to evaluate the feasibility and potential efficacy of automated mobile phone–based vaccination reminders and conditional financial incentive offers for improving the timeliness of routine childhood vaccinations. The study was conducted in Mtwara Region in southern Tanzania, a region with large numbers of un- or undervaccinated children. We previously reported on high rates of mobile phone access and use in the study area [17]. At the time of the study, Tanzania reported 75 mobile phone subscriptions per 100 people; by 2022, the number had reached 92 subscriptions per 100 people [30].

Study Design

The trial compared three study arms: (1) standard of care (no reminders or incentives), (2) mobile phone–based reminders, and (3) mobile phone–based reminders with conditional financial incentive offers. The assignment to study arms was implemented at the level of the vaccination event. The primary study outcome was the timeliness of routine vaccinations for the newborn children, defined as vaccination receipt within 28 days of the

XSL•FO
RenderX

scheduled vaccination due date, for vaccinations recommended at ages 6, 10, and 14 weeks.

Rationale

Reminders and incentives were evaluated as potentially scalable means of increasing the timeliness of routine childhood vaccinations. Reminders served as nudges [31], whereas conditional financial incentive offers served to offset transportation and other barriers that may prevent mothers from acting on their intention to get their children vaccinated on time [32].

Recruitment and Enrollment

To participate in the study, women had to be at least 16 years of age, in the third trimester of pregnancy, have access to a mobile phone, and provide informed consent. The target enrollment was 400 women from at least 4 urban and at least 8 rural health facilities and the surrounding communities. Health facilities within a 30 km radius of the regional capital that routinely provide childhood vaccinations were considered eligible venues for the recruitment of study participants.

A combination of purposive and snowball sampling strategies was used for recruitment. Eligible women presenting for antenatal care at participating facilities were approached by trained study personnel and offered enrollment in the study. Participating women and local community leaders were asked to identify other pregnant women in their community who, if eligible, were also offered enrollment in the study.

Use of mParis

Pregnant women aged 16 years or older were enrolled into the study in their third trimester of pregnancy. Mobile phone numbers were documented for the women or anyone designated by the woman as a vaccine advocate, that is, someone who could relay information about upcoming vaccination appointments to the woman. The woman's name, mobile phone number, and expected date of delivery were batch-uploaded into mParis, thereby triggering the mother process (Figure 2).

Figure 2. Process definition for the "Mother process" from study enrollment in pregnancy through the completion of the child's 14-week vaccination. This process diagram describes an enrolled mother's progression from pregnancy through vaccinations recommended at birth and ages 6, 10, and 14 weeks. The process is started by the enrollment process and invokes the pregnancy and vaccination sub-processes. "+" indicates "all," while "x" indicates "or". Participants were enrolled in Mtwara Region, Tanzania. The short message service-based intervention lasted from January to July 2018.

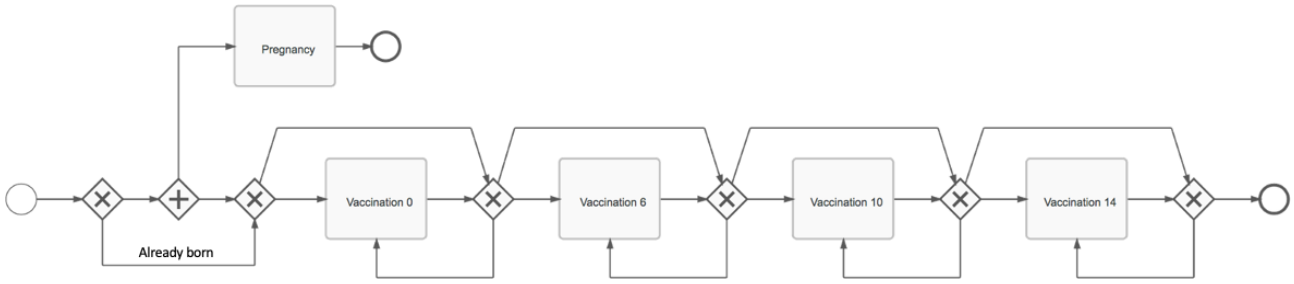

mParis sent reminders and incentive offers in the form of individualized SMS text messages (Table 2). Separate message templates were used for reminders before and after a vaccination due date, and for mothers and designees (ie, vaccine advocates). Reminders and incentive offers in the form of SMS text messages were found to be feasible and acceptable in this setting [24].

XSL•FO
RenderX

Table . Message templates for short message service messages sent via mParis during a quasi-randomized trial of automated mobile phone-based vaccination reminders and incentive offers to mothers of newborn children, by study arm and recipient type (Tanzania, 2018).

Arm, timing, and recipient	SMS message text[a]
Arm 2 (reminders)	
Before the due date	
Mother	Your child is due to receive a vaccination on [*due_date*].
Designee	Please remind [*mother_name*] that her child is due to be vaccinated on [*due_date*].
After the due date	
Mother	Has your child received the vaccination scheduled for [*due_date*]? If s/he has not been vaccinated, please bring her/him before [*end_date*].
Designee	Has [*mother_name*] brought her child to be vaccinated on [*due_date*]? If not, she should do so before [*end_date*].
Arm 3 (reminders and incentive offers)	
Before the due date	
Mother	Your child is due to receive a vaccination on [*due_date*]. If s/he receives the vaccination on time you will receive Sh2000.
Designee	Please remind [*mother_name*] that her child is due to be vaccinated on [*due_date*]. For a timely vaccination she will receive Sh2000.
After the due date	
Mother	Has your child received the vaccination scheduled for [*due_date*]? If s/he has not been vaccinated, please bring her/him before [*end_date*] to receive Sh1000.
Designee	Has [*mother_name*] brought her child to be vaccinated on [*due_date*]? If not, she should do so before [*end_date*] to receive Sh1000.

[a]Italicized content in square brackets represents a placeholder for message-specific information. All messages also contained the following information: "If s/he already received the vaccination call [*phone number*]." Message templates were translated to Kiswahili.

Recipients were study participants or their designees; messages were sent for routine childhood vaccinations with due dates between January and July 2018.

Process Definitions

After the report of a live birth, SMS reminder messages, in Kiswahili, were sent 1 week and 1 day before each vaccination due date (at 6, 10, and 14 weeks of age), and if no vaccination was reported, 14 days after the due date. For arm 3, reminders included a financial incentive offer for each timely vaccination.

A mother's progression in the trial from enrollment during pregnancy through birth of the child, and the child's vaccinations at birth and at ages 6, 10, and 14 weeks was described by the mother process (Figure 2). The mother process was started by the enrollment of a pregnant mother and invoked vaccination and other nested processes. Delays in the 6- and 10-week vaccination processes automatically shifted subsequent vaccination processes, that is, vaccination due dates for the 10- and 14-week vaccinations were set to 28 days after the previously completed vaccination.

The "vaccination process" (Figure 3) illustrates the versatility of the workflow approach, and thus the mParis system. Study-arm specific vaccination events required a process definition that flexibly supported study arm–specific invocations of reminder, call-back (phone call), and incentive disbursement processes. The result of the vaccination process is reported back to the mother process and results either in the restart of the vaccination process (in the case of a rescheduled vaccination), progression to the next vaccination (in the case of a completed vaccination), or the end of the mother process (in the case of a completed 14-week vaccination).

XSL•FO
RenderX

Figure 3. Process definition for the "Vaccination process" as a function of the study arm of the vaccination event. Notes: The vaccination process awaits notification of a received or rescheduled vaccination and delegates the study-arm specific sending of SMS messages, phone calls, and incentive payouts to child processes. "+" indicates "all," while "x" indicates "or."

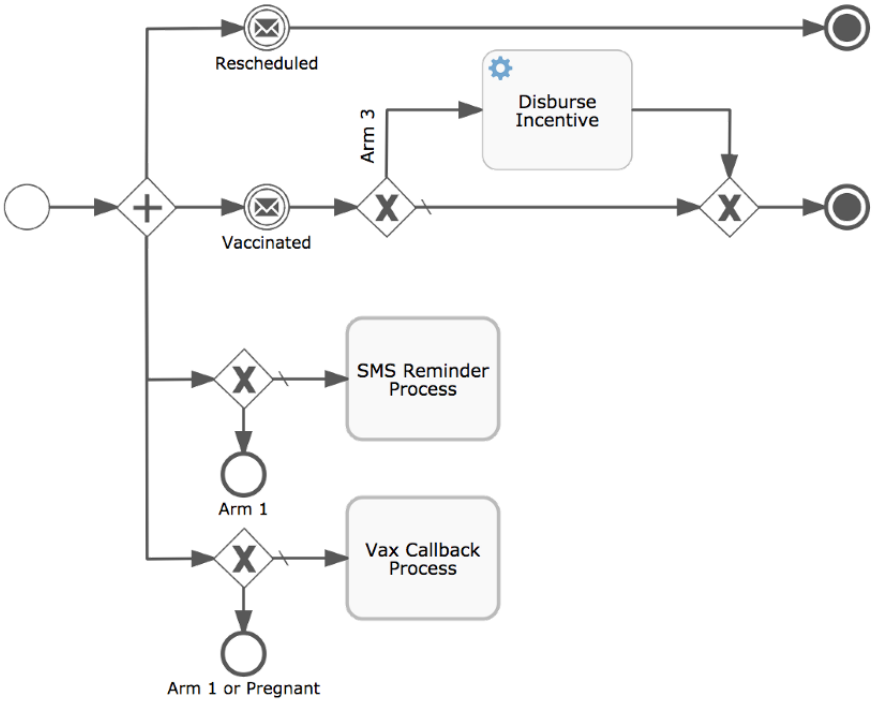

Implementation Details

The content of the messages was pilot tested with 32 women, and iterative changes were made to optimize comprehension, content, and clarity. Message templates (Table 2) were stored in a properties file, with placeholders that allowed for the individualization of message content to each recipient (eg, mother's name and vaccination due date).

Children's birth dates, dates of completed vaccinations, and scheduled and rescheduled vaccination due dates were ascertained by research assistants who regularly contacted mothers via mobile phone–based follow-up. These dates were used to iteratively update active mParis processes until process completion, which occurred upon completion of the 14-week vaccination or the end of the follow-up interview, whichever came first.

Ethical Considerations

The protocol was approved by the institutional review boards at Duke University (protocol 2017‐0591) and the University of South Carolina (facilitated review, Pro00051213), United States, and the National Institute for Medical Research (NIMR) in Tanzania (NIMR/HQ/R.8a/Vol. IX/2194). The trial was registered with ClinicalTrials.gov (NCT03252288). Written informed consent was obtained from all study participants. Participants received a compensation of TSH 3000 after the completion of each survey. At the time of the study, US $1 was worth approximately TSH 2220. Password-protected tablet devices were used to collect survey data; analytic data are stored on encrypted, password-protected devices. Identifying information was collected, transferred, and stored separately from survey data. The security of the mParis system is described above.

Results

Implementation

Between August 15, 2017, and December 8, 2017, a total of 412 pregnant women in their last trimester of pregnancy were enrolled into the study. Neither reminders nor incentive offers were sent for vaccination due dates through January 11, 2018 (arm 1; "control" phase). During the control phase, 140 women (34% of the sample) were confirmed to have completed the vaccination process; 14 (3.4%) were withdrawn due to maternal or child death; 79 (19.2%) women were lost to follow-up.

Vaccination reminders were automatically sent for vaccination due dates between January 12, 2018, to July 10, 2018. In total, during the 6-month implementation period, 1397 SMS text messages were scheduled to be sent through mParis. Of these, 1122 (80.3%) messages were recorded as delivered, 249 (18.8%) as expired (these were automatically resent), 23 (1.6%) as failed, and 3 (0.2%) as sent but lacking a delivery confirmation.

Women with known vaccination due dates between January 12, 2018, and March 5, 2018, were sent reminder messages (arm 2). Women with known vaccination due dates between March 6, 2018, and July 10, 2018, received reminder messages with tiered incentive offers for timely vaccinations (arm 3). Of 173 women who received at least 1 message, 67 (38.7%) women were sent reminders only, whereas 106 (61.3%) women were sent at least 1 incentivized message. In total, 633 (45.3%) messages contained a financial incentive offer for the timely vaccination of the child.

Challenges, Adaptations, and Resolutions

Multiple challenges were encountered during the implementation of mParis, despite the system's deliberate design to a priori

XSL•FO
RenderX

accommodate basic problems, such as intermittent internet access and power failures. Key challenges are summarized in Table 3, with illustrative sample quotes from communications between the study team and the system developer. Formative, continuous adaptation of the system to increase the resilience of the system, complemented by limited manualized backstops during trial implementation, enabled the successful deployment of a robust system capable of accommodating diverse challenges that could be encountered in varied settings.

Table . Challenges encountered in the implementation of mParis in the context of the quasi-randomized trial (Tanzania, 2018).

Challenge	Description and sample quotes from communication between the study team and the system developer	Resolutions
Unreliable electricity	Heavy rains cause disruptions in electricity supply and internet connectivity. "When there are longer power cuts in Moshi, [network provider's] network machinery goes down and we lose the connection. Very frustrating, as it's outside of our control, and very frequent now with the heavy rains."	• Server is plugged into an uninterruptible power supply (UPS) unit, UPS is backed by a generator. • SMS text messages are received and sent from a local server, with no dependence on internet connectivity.
Accessing mParis via the internet	Disruptions in the server's internet connectivity led to inaccessibility of mParis via the internet. "The issue with mPARIS is the connection between us and the internet exchange downtown." "Or... again ... someone forgot to pay the bill." "… when I tried to switch over to our dedicated DSL line it turned out that [internet provider] did not assign the promised static IP but gave us a dynamic IP instead."	• Redundant internet connections: "I'm thinking about alternatives (wireless broadband over VPN to an [external vendor] endpoint)" • Batch export of human tasks: "I have a fairly good redundancy set up now -- each day I pull the status reports from mParis, process them, and send a spreadsheet with priority callbacks … so they can work online as well as offline."
Flowable versioning	"I decided to go with the latest Flowable release (6.3.0 just out), but they've overhauled the way they do configuration."	• Additional testing after version changes.
Lost SMS text messages	Unannounced change in provider-imposed daily limit on the number of SMS text messages that could be sent from a single SIM card.	• Allocate nontime-sensitive messages to different days; distribute load over multiple modems. Shorten message templates to stay within the SMS text message character limit. Alternatively, contract with multiple SMS providers.
Support for multiple mobile phone carriers	Not all modems support Linux. "I got a [network provider] SIM this weekend, with (sigh) a cheap imitation dongle because they ran out of the proper ones. Sure enough, despite what it says on the box it doesn't work on Linux or MacOS…"	• Testing of diverse modems, including "legacy dongles". • "Luckily I found my sturdy old [mobile phone provider] dongle (dated 2009) and it works fine ... I'll look for two more oldies."
A large number of processes to be launched	The large number of study participants and unreliable internet connectivity made the individualized registration of participants in mParis and the individualized launch of the respective "mother processes" inefficient.	Added functionality for batch uploads. "For the initial entry, I think it'll be easy for me to convert a CSV table to REST calls and shoot those at the system from here (Flowable exposes pretty much all functionality it has also over REST)."
Time sensitivity of SMS reminders	Late reports of prior vaccinations and unsent or failed reminder messages required the dynamic and highly time-sensitive rescheduling of SMS reminder messages in accordance with the study protocol.	Added functionality allowed for the instantaneous sending of relevant SMS reminder messages for individual study participants.

[a] SMS-short message service. Source: email record between the system developer and the study team.

Other mParis Use Cases

Since 2018, 2 other instances of mParis have supported research studies in Tanzania, and a third use case is currently being developed. Between 2018 and 2020, mParis was used to "batch-send" recruitment reminders to support enrollment into a randomized controlled trial aimed at improving uptake of HIV testing in northern Tanzania [3]. During the recruitment of potential participants at off-site recruitment venues, eligible individuals provided contact phone numbers to members of the

XSL•FO
RenderX

research team. mParis sent follow-up messages to these individuals, inviting them to come to the study's research offices for consent and enrollment. Subsequently, for participants enrolled in the study, mParis sent scheduled SMS reminders to return to the research offices for follow-up. mParis also sent reminder messages to participants who were given HIV testing invitation cards to get tested for HIV as described on the cards.

mParis is currently supporting a randomized trial to evaluate the acceptability, efficacy, and cost-effectiveness of confidential SMS-based social network referrals for HIV testing as a means of reaching high-risk individuals and prompting them to test for HIV [21]. Clients presenting for HIV testing or treatment are offered the opportunity to send confidential SMS-based HIV testing invitations to social network contacts (eg, friends, family members, and coworkers) who could benefit from HIV testing. Inviters can choose from a menu of message options; mParis autonomously schedules invitation messages to be sent to invitees' mobile phones after 1 and 7 days. The system ensures that an inviter cannot be identified by their invitee, that no invitee phone number receives more than 2 messages regardless of how often they are invited, and that failed messages are automatically resent.

mParis is presently being adapted to support a cluster-randomized trial of an intervention aimed at improving the timeliness of routine childhood vaccinations in rural western Tanzania. mParis will be integrated with Dimagi's CommCare system [33] to support community health care workers to remind mothers of upcoming vaccination due dates, notify mothers of vaccine stock-outs, and mitigate vaccination-related knowledge gaps. These applications exemplify the versatility of the system, which can be adapted to support other health interventions (eg, messages to support medication adherence), extended beyond SMS technology (eg, to WhatsApp, Meta, or email communication), and even to support mobile phone–based survey research (eg, via Unstructured Supplementary Service Data [USSD]).

Discussion

Principal Findings

In this study, we describe the design of a highly versatile mobile phone–based appointment reminder and incentive system ("mParis") to facilitate autonomous communication with individuals. We provide a detailed description of a specific instance implemented in a low-resource setting to support the timely vaccination of children. mParis autonomously sent more than 1300 individualized SMS text messages to Tanzanian mothers participating in a research study aimed at improving the timeliness of routine childhood vaccinations. We detail the technical architecture, demonstrate mParis' technical feasibility, assess implementation challenges, and explore its broader applicability to public health and research contexts.

mParis, built on a business process modeling platform, combines several features [34] that increase its potential applicability to other contexts. First, the system is built in Flowable, an open-source, low-code modeling environment. Intuitive and highly flexible process definitions allow for the specification of diverse algorithms with minimal programming requirements. The low-code nature of the system allows for workflow logic to be visually specified, with no "hand coding" needed, and to be executed and supervised by the platform. This ensures that the system can be adapted with limited programming expertise and that processes are executed in full alignment with their design. Second, the system maintains consistency, that is, every process instance is in a well-defined state at each moment (comparable with the way databases guarantee data consistency) and is fully auditable (ie, it keeps track of every event and action that occurs in a process). These features support operational efficiency and minimize the need for coordination between stakeholders. For example, if multiple team members jointly process events (such as call-backs to mothers), there is instantaneous tracking of events, minimizing the need for additional event-related communication. Finally, the system is open source, with many integration components available to connect to external services. For example, while mParis' primary task in the sample application comprised the algorithmic sending of SMS text messages, the system may be integrated with other popular communication platforms, such as WhatsApp and Facebook Messenger (Meta), or WeChat (Tencent). Similarly, mParis is readily integrable with existing health information systems; its robust security infrastructure and full auditability suggest that an appropriate integration can be fully compliant with HIPAA (Health Insurance Portability and Accountability Act) rules and regulations. The mParis instance and application described in this paper was subject to several limitations. Some of these limitations were addressed, as described. Others are being addressed in ongoing use cases or may be addressed in future use cases.

Most notably, the mParis application described here relied on a dongle to send SMS text messages. A larger-scale implementation of the system requires additional capacity for sending messages and would benefit from a dedicated short message peer-to-peer connection, though there are no fundamental limits to the scale of the existing system.

Second, reminders and incentive offers were sent using SMS text messages. SMS technology is not secure and, due to restrictions in message type and length, may not be optimal for many use cases. mParis allows for the integration of other modes of communication, such as emails or WhatsApp messages. These were not considered feasible in the application described here. Mobile-cellular subscription rates are high in Tanzania at 92 per 100 people [30]; however, rural populations have lower smartphone access than urban populations, and a majority with phones rely on basic communication strategies, such as phone calls and SMS text messages [24,35].

Third, while at enrollment into the trial, more than 98% of women enrolled in this study had used a mobile phone to make phone calls, only two thirds had used mobile phones to send or receive SMS text messages, and two thirds of women reported using a mobile phone less than once a week [24]. Thus, despite its reliance on basic SMS-based communication, mParis may not have been able to overcome all access barriers and may need manualized back-stops, such as phone calls, to reach all intended recipients within a timely manner. However, by fully automating reminders and incentive offers for those with mobile phone

XSL•FO
RenderX

access, limited human resources can be targeted toward back-stopping residual gaps. This highlights the broad usage potential of this highly versatile system for autonomously engaging individuals across a variety of contexts.

In summary, mParis was successfully deployed and used in a quasi-randomized trial of SMS reminders and incentives as means of increasing the timeliness of childhood vaccinations in Tanzania. Iterative adaptations to mParis in response to stochastic environmental and technical implementation barriers resulted in the formation of a robust mHealth system that is locally hosted in a low-resource setting, but readily extensible to other geographic and topical contexts, including high-resource settings.

Conclusion

mParis' open-source nature, auditability, and ability to autonomously execute algorithms in the face of diverse challenges suggest favorable prospects for the automation of critical health communication and research-related tasks in a wide range of settings. These characteristics are key to the sustainability and scale-up of mHealth systems in both high- and low-resource settings.

Acknowledgments

The authors thank the women who participated in this study and acknowledge the research assistants from the National Institute for Medical Research Muhimbili Medical Research Centre who supported data collection activities, Ayubu Masasi for inputs on algorithms and their implementation, and the regional and district medical officers of Mtwara and their staff for their support of the study. The authors would also like to acknowledge administrative support from the staff of Tanzania's National Institute for Medical Research, the Kilimanjaro Clinical Research Institute, the University of South Carolina's Arnold School of Public Health, the Duke Global Health Institute, and Duke University's Center for Health Policy and Inequalities Research. This study was supported by grants from the Fogarty International Center of the National Institutes of Health (award number R21TW010262) to Duke University, and the Maternal, Adolescent and Child Health (MACH) working group of the Duke Global Health Institute. LV was funded by the National Center for Advancing Translational Sciences of the National Institutes of Health under (award number 1KL2TR002554). The development of the mobile phone–based appointment reminder and incentive system (mParis) was supported by the National Institute of Mental Health under (award number R01MH106388). The funding bodies have no role in the design of the study; the collection, analysis, and interpretation of data; or the writing of the manuscript. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Generative AI was not used in any portion of the manuscript writing.

Data Availability

The data described in this manuscript are available from the authors in a deidentified format. The dataset comprises 1397 records, one record for each message sent. Variables include (a) a scrambled participant ID; (b) a sequential message ID within participant; (c) the text of the message template (Table 2); and (d) the status of the message. Source code for the secure and reliable scheduled delivery of SMS text messages is available on Github [36].

Conflicts of Interest

None declared.

Checklist 1
mERA (mHealth Evidence Reporting and Assessment) checklist.
[DOCX File, 18 KB - formative_v9i1e65150_app1.docx]

Checklist 2
TiDieR (Template for Intervention Description and Replication) checklist.
[DOCX File, 31 KB - formative_v9i1e65150_app2.docx]

References

1. Okano JT, Ponce J, Krönke M, Blower S. Lack of ownership of mobile phones could hinder the rollout of mHealth interventions in Africa. Elife 2022 Oct 18;11:e79615. [doi: 10.7554/eLife.79615] [Medline: 36255055]
2. Heeks R. Digital inequality beyond the digital divide: conceptualizing adverse digital incorporation in the global South. Inf Technol Dev 2022 Oct 2;28(4):688-704. [doi: 10.1080/02681102.2022.2068492]
3. Badr J, Motulsky A, Denis JL. Digital health technologies and inequalities: a scoping review of potential impacts and policy recommendations. Health Policy 2024 Aug;146:105122. [doi: 10.1016/j.healthpol.2024.105122] [Medline: 38986333]
4. Pscheidt M, Weide TP. Bridging the digital divide by open source: a theoretical model of best practice. In: Druicã E, editor. Digital Economy Innovations and Impacts on Society: IGI Global; 2012:151-167. [doi: 10.4018/978-1-4666-1556-4.ch010]

XSL•FO
RenderX

5. Saa-Dittoh F, Akkermans H. Bridging the digital divide. In: Werthner H, Ghezzi C, Kramer J, editors. Introduction to Digital Humanism: A Textbook: Springer; 2024:283-298. [doi: 10.1007/978-3-031-45304-5_19]
6. Recommendations on digital interventions for health system strengthening – research considerations. World Health Organization. 2019. URL: https://www.who.int/publications/i/item/WHO-RHR-19.9 [accessed 2025-07-18]
7. Palmer MJ, Henschke N, Villanueva G, et al. Targeted client communication via mobile devices for improving sexual and reproductive health. Cochrane Database Syst Rev 2020 Jul 14;8(8):CD013680. [doi: 10.1002/14651858.CD013680] [Medline: 32779730]
8. Recommendations on digital interventions for health system strengthening-evidence and recommendations. World Health Organization. 2019. URL: https://iris.who.int/bitstream/handle/10665/311980/WHO-RHR-19.10-eng.pdf?sequence=1 [accessed 2025-07-18]
9. Agarwal S, Chin WY, Vasudevan L, et al. Digital tracking, provider decision support systems, and targeted client communication via mobile devices to improve primary health care. Cochrane Database Syst Rev 2025 Apr 7;4(4):CD012925. [doi: 10.1002/14651858.CD012925.pub2] [Medline: 40193137]
10. Uddin M, Shamsuzzaman M, Horng L, et al. Use of mobile phones for improving vaccination coverage among children living in rural hard-to-reach areas and urban streets of Bangladesh. Vaccine (Auckl) 2016 Jan;34(2):276-283. [doi: 10.1016/j.vaccine.2015.11.024]
11. Chen L, Du X, Zhang L, et al. Effectiveness of a smartphone app on improving immunization of children in rural Sichuan Province, China: a cluster randomized controlled trial. BMC Public Health 2016 Aug 31;16(1):909. [doi: 10.1186/s12889-016-3549-0] [Medline: 27581655]
12. Onigbogi O, Ojo OY, Kinnunen UM, Saranto K. Mobile health interventions on vaccination coverage among children under 5 years of age in Low and Middle-Income countries; a scoping review. Front Public Health 2025;13:1392709. [doi: 10.3389/fpubh.2025.1392709] [Medline: 39935874]
13. Mechael P, Gilani S, Ahmad A, et al. Evaluating the “Zindagi Mehfooz” electronic immunization registry and suite of digital health interventions to improve the coverage and timeliness of immunization services in Sindh, Pakistan: mixed methods study. J Med Internet Res 2024 Oct 11;26:e52792. [doi: 10.2196/52792] [Medline: 39162666]
14. Mekonnen ZA, Gelaye KA, Were MC, Gashu KD, Tilahun BC. Effect of mobile text message reminders on routine childhood vaccination: a systematic review and meta-analysis. Syst Rev 2019 Jun 28;8(1):154. [doi: 10.1186/s13643-019-1054-0] [Medline: 31253186]
15. Domek GJ, Contreras-Roldan IL, Bull S, et al. Text message reminders to improve infant immunization in Guatemala: a randomized clinical trial. Vaccine (Auckl) 2019 Sep 30;37(42):6192-6200 [FREE Full text] [doi: 10.1016/j.vaccine.2019.08.046] [Medline: 31492475]
16. Vasudevan L, Baumgartner JN, Moses S, Ngadaya E, Mfinanga SG, Ostermann J. Parental concerns and uptake of childhood vaccines in rural Tanzania - a mixed methods study. BMC Public Health 2020 Oct 20;20(1):1573. [doi: 10.1186/s12889-020-09598-1] [Medline: 33081744]
17. Ostermann J, Vasudevan L, Baumgartner JN, Ngadaya E, Mfinanga SG. Do mobile phone-based reminders and conditional financial transfers improve the timeliness of childhood vaccinations in Tanzania? Study protocol for a quasi-randomized controlled trial. Trials 2019 Jul 4;20(1):397. [doi: 10.1186/s13063-019-3430-4] [Medline: 31272487]
18. Agarwal S, LeFevre AE, Lee J, et al. Guidelines for reporting of health interventions using mobile phones: mobile health (mHealth) evidence reporting and assessment (mERA) checklist. BMJ 2016 Mar 17;352:i1174. [doi: 10.1136/bmj.i1174] [Medline: 26988021]
19. Hoffmann TC, Glasziou PP, Boutron I, et al. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. BMJ 2014 Mar 7;348(mar07 3):g1687. [doi: 10.1136/bmj.g1687] [Medline: 24609605]
20. Ostermann J, Njau B, Hobbie A, et al. Using discrete choice experiments to design interventions for heterogeneous preferences: protocol for a pragmatic randomised controlled trial of a preference-informed, heterogeneity-focused, HIV testing offer for high-risk populations. BMJ Open 2020 Nov 6;10(11):e039313. [doi: 10.1136/bmjopen-2020-039313] [Medline: 33158826]
21. Ostermann J, Njau B, van Zwetselaar M, et al. Mobile Phone-Based Confidential Social Network Referrals for HIV Testing (CONSORT): protocol for a randomized controlled trial. JMIR Res Protoc 2024 May 30;13:e55068. [doi: 10.2196/55068] [Medline: 38814692]
22. Intelligent, open, low-code business automation. Flowable. URL: https://www.flowable.org/ [accessed 2024-08-06]
23. National Internet Data Center. National Internet Data Center. 2025. URL: https://nidc.co.tz/ [accessed 2025-05-19]
24. Vasudevan L, Ostermann J, Moses SM, Ngadaya E, Mfinanga SG. Patterns of mobile phone ownership and use among pregnant women in Southern Tanzania: cross-sectional survey. JMIR Mhealth Uhealth 2020 Apr 8;8(4):e17122. [doi: 10.2196/17122] [Medline: 32267240]
25. Immunization coverage. World Health Organization. 2024 Jul 15. URL: https://www.who.int/news-room/fact-sheets/detail/immunization-coverage [accessed 2025-07-18]
26. Kaur G, Danovaro-Holliday MC, Mwinnyaa G, et al. Routine vaccination coverage - worldwide, 2022. MMWR Morb Mortal Wkly Rep 2023 Oct 27;72(43):1155-1161. [doi: 10.15585/mmwr.mm7243a1] [Medline: 37883326]

XSL•FO
RenderX

27. Ministry of Health T, National Bureau of Statistics - NBS/Tanzania. 2022 tanzania demographic and health survey and malaria indicator survey (2022 TDHS-MIS). : Tanzania National Bureau of Statistics; 2022 URL: https://dhsprogram.com/pubs/pdf/OF50/OF50.pdf [accessed 2025-08-12]
28. Yelverton V, Hair NL, Ghosh SH, et al. Beyond coverage: Rural-urban disparities in the timeliness of childhood vaccinations in Tanzania. Vaccine (Auckl) 2022 Sep 2;40(37):5483-5493. [doi: 10.1016/j.vaccine.2022.07.020] [Medline: 35961796]
29. Elderly and Children - MoHCDGEC/Tanzania Mainland, Ministry of Health - MoH/Zanzibar, National Bureau of Statistics - NBS/Tanzania, Office of Chief Government Statistician - OCGS/Zanzibar, ICF. Tanzania demographic and health survey and malaria indicator survey (TDHS-MIS) 2015-16. The DHS Program. 2016. URL: https://dhsprogram.com/publications/publication-FR321-DHS-Final-Reports.cfm [accessed 2021-06-17]
30. Mobile cellular subscriptions (per 100 people). World Bank Group. 2019 Oct 23. URL: http://data.worldbank.org/indicator/IT.CEL.SETS.P2 [accessed 2019-10-23]
31. Thaler R, Sunstein C. Nudge: Improving Decisions About Health, Wealth, and Happiness, th edition: Penguin Books; 2009.
32. Ostermann J, Hair NL, Moses S, et al. Is the intention to vaccinate enough? Systematic variation in the value of timely vaccinations and preferences for monetary vs non-monetary incentives among pregnant women in southern Tanzania. Vaccine X 2023 Apr;13:100266. [doi: 10.1016/j.jvacx.2023.100266] [Medline: 36814594]
33. CommCare. Dimag. 2025. URL: https://dimagi.com/commcare/ [accessed 2025-05-19]
34. Santiago D. Key benefits of business process modeling. Floable. 2017 May 20. URL: https://www.flowable.com/blog/business/key-benefits-of-business-process-modeling [accessed 2024-09-11]
35. Schroeder K, Maiarana J, Gisiri M, et al. Caregiver acceptability of mobile phone use for pediatric cancer Care in Tanzania: cross-sectional questionnaire study. JMIR Pediatr Parent 2021 Dec 8;4(4):e27988. [doi: 10.2196/27988] [Medline: 34889763]
36. Zwets/sms-scheduler. GitHub. URL: https://github.com/zwets/sms-scheduler [accessed 2025-08-12]

Abbreviations

BPMN: Business Process Modeling Notation
HIPAA: Health Insurance Portability and Accountability Act
KCRI: Kilimanjaro Clinical Research Institute
mERA: mHealth evidence reporting and assessment
mParis: mobile phone-based appointment reminder and incentive system
NIMR: National Institute for Medical Research
TiDieR: Template for Intervention Description and Replication
USSD: Unstructured Supplementary Service Data
WHO: World Health Organization

Edited by A Mavragani; submitted 12.09.24; peer-reviewed by W Tumuhimbise; revised version received 06.06.25; accepted 09.06.25; published 10.09.25.

Please cite as:
van Zwetselaar M, Ostermann J, Beti M, Baumgartner JN, Mfinanga S, Ngadaya E, Vasudevan L, Thielman N
Design of an Automated Mobile Phone-Based Reminder and Incentive System: Application in a Quasi-Randomized Controlled Trial to Improve the Timeliness of Childhood Vaccinations in Tanzania
JMIR Form Res 2025;9:e65150
URL: https://formative.jmir.org/2025/1/e65150
doi:10.2196/65150

© Marco van Zwetselaar, Jan Ostermann, Melkiory Beti, Joy Noel Baumgartner, Sayoki Mfinanga, Esther Ngadaya, Lavanya Vasudevan, Nathan Thielman. Originally published in JMIR Formative Research (https://formative.jmir.org), 10.9.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Implementation of Fully Automated AI-Integrated System for Body Composition Assessment on Computed Tomography for Opportunistic Sarcopenia Screening: Multicenter Prospective Study

Bushra Urooj[1,2*], BS; Yousun Ko[1,2*], PhD; Seongwon Na[1,2], PhD; In-One Kim[3], MD; Eun-Hee Lee[3], PhD; Seon Cho[3], MS; Heeryeol Jeong[4], PhD; Seungwoo Khang[4], PhD; Jeongjin Lee[4], PhD; Kyung Won Kim[1,2], MD, PhD

[1]Department of Medical Science, Asan Medical Institute of Convergence Science and Technology, University of Ulsan College of Medicine, 88 Olympic-ro 43-gil, Asan Medical Center, Seoul, Republic of Korea

[2]Department of Radiology, Asan Medical Center, University of Ulsan College of Medicine, 88 Olympic-ro 43-gil, Asan Medical Center, Seoul, Republic of Korea

[3]Medicheck Research Institute, Korea Association of Health Promotion, Seoul, Republic of Korea

[4]School of Computer Science and Engineering, Soongsil University, Seoul, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Kyung Won Kim, MD, PhD
Department of Medical Science, Asan Medical Institute of Convergence Science and Technology, University of Ulsan College of Medicine, 88 Olympic-ro 43-gil, Asan Medical Center, Seoul, Republic of Korea

Abstract

Background: Opportunistic computed tomography (CT) screening for the evaluation of sarcopenia and myosteatosis has been gaining emphasis. A fully automated artificial intelligence (AI)–integrated system for body composition assessment on CT scans is a prerequisite for effective opportunistic screening. However, no study has evaluated the implementation of fully automated AI systems for opportunistic screening in real-world clinical practice for routine health check-ups.

Objective: The aim of this study is to evaluate the performance and clinical utility of a fully automated AI-integrated system for body composition assessment on opportunistic CT during routine health check-ups.

Methods: This prospective multicenter study included 537 patients who underwent routine health check-ups across 3 institutions. Our AI algorithm models are composed of selecting L3 slice and segmenting muscle and fat area in an end-to-end manner. The AI models were integrated into the Picture Archiving and Communication System (PACS) at each institution. Technical success rate, processing time, and segmentation accuracy in Dice similarity coefficient were assessed. Body composition metrics were analyzed across age and sex groups.

Results: The fully automated AI-integrated system successfully retrieved anonymized CT images from the PACS, performed L3 selection and segmentation, and provided body composition metrics, including muscle quality maps and muscle age. The technical success rate was 100% without any failed cases requiring manual adjustment. The mean processing time from CT acquisition to report generation was 4.12 seconds. Segmentation accuracy comparing AI results and human expert results was 97.4%. Significant age-related declines in skeletal muscle area and normal-attenuation muscle area were observed, alongside increases in low-attenuation muscle area and intramuscular adipose tissue.

Conclusions: Implementation of the fully automated AI-integrated system significantly enhanced opportunistic sarcopenia screening, achieving excellent technical success and high segmentation accuracy without manual intervention. This system has the potential to transform routine health check-ups by providing rapid and accurate assessments of body composition.

(JMIR Form Res 2025;9:e69940) doi:10.2196/69940

KEYWORDS

artificial intelligence; body composition; sarcopenia; myosteatosis; opportunistic screening; health check-up

XSL•FO
RenderX

Introduction

Sarcopenia, characterized by the progressive loss of skeletal muscle mass and function, and myosteatosis, defined as excessive fat accumulation within muscle tissue, are critical health issues associated with functional impairment, frailty, and increased mortality [1]. With the global population aging, the prevalence of these conditions is expected to rise significantly [2]. Early diagnosis is essential to mitigate these risks by enabling timely management [3]. Thus, routine health screening for sarcopenia and myosteatosis is important for preventive health care.

The imaging diagnosis of sarcopenia relies on assessing both muscle mass and quality. While magnetic resonance imaging and whole-body dual-energy X-ray absorptiometry provide accurate evaluations, their limited accessibility and longer imaging times pose significant barriers to widespread clinical adoption [3]. In contrast, computed tomography (CT), widely used for routine diagnostic purposes, has emerged as a practical imaging modality for body composition assessment [4].

Especially, opportunistic CT screening, where CT scans performed for other clinical indications are repurposed for secondary analyses, represents an innovative approach to diagnosing sarcopenia without additional costs or radiation exposure [5]. The development of automated body composition assessment methods powered by artificial intelligence (AI) has further streamlined this process, making opportunistic CT screening a practical option in daily clinical practice [6]. Integration of such AI solutions into clinical workflows, particularly through Picture Archiving and Communication Systems (PACS), is essential for establishing standardized processes for fully automated opportunistic CT screening.

Although some studies have explored opportunistic CT screening for sarcopenia [7,8], no study has prospectively evaluated the seamless workflow integration, technical success, and processing efficiency of fully AI-integrated systems in routine health check-up settings. Therefore, this study aims to implement and evaluate the fully automated AI-integrated opportunistic CT screening system for routine health check-ups. By assessing its technical success, processing efficiency, and diagnostic outcomes across multiple health check-up institutions, this study seeks to demonstrate the system's scalability, robustness, and clinical utility.

Methods

Recruitment

Designed as a multicenter prospective observational study, it included patients who underwent health check-ups at 3 health institutions affiliated with the Korea Association of Health Promotion (KAHP) between December 2022 and December 2023. Consecutive recruitment was implemented according to inclusion and exclusion criteria. Inclusion criteria were adults aged 20 years or older who had undergone abdominal CT scans as part of their routine health check-ups. Exclusion criteria were patients with canceled or incomplete CT scans, or those with significant artifacts on CT images. No specific screening for comorbidities such as diabetes or prior malignancies was performed. We included all patients who underwent abdominal CT, regardless of contrast enhancement.

CT Acquisition

CT scans were performed using Revolution Discovery (GE Healthcare), Somatom Definition AS+, and Somatom Definition Edge (Siemens Healthineers) scanners, using parameters as follows: tube voltage, 120 kVp; effective tube current, 200 reference mAs (Care Dose 4D; Siemens Healthineers) or 100 - 400 mA (AutomA or SmartmA; GE Healthcare); field of view, 30 - 40 cm; collimation, 0.31 - 0.75 mm; and pitch, 0.98 - 1.00. Images were reconstructed using the filtered backprojection technique with the soft tissue reconstruction algorithm at a section thickness of 5 mm with no interslice gap. To ensure that variations between scanners did not affect segmentation accuracy or Hounsfield Unit (HU) measurements, we standardized acquisition protocols with the same kVp, reconstruction technique, and slice thickness.

AI Algorithm Model

The AI solution (Aid-U, iAID Inc.), a fully automatic deep learning solution for both L3 selection and body composition assessment, was used for the AI algorithm model (Video S1 in Multimedia Appendix 1). The software uses a You Only Look Once v3 (YOLOv3)–based algorithm for automatic L3 inferior endplate level selection and a fully convolution network for segmentation of abdominal muscle area and fat areas, as illustrated in Figure 1 [9]. Selected L3 level CT images were automatically segmented to generate the boundary of total abdominal muscles and measured the abdominal muscle and fat area [10]. Based on predefined HU thresholds, our system then classified the segmented muscle into normal-attenuation muscle area (NAMA, +30 HU to +150 HU), low-attenuation muscle area (LAMA, −29 HU to +29 HU), and inter- and intramuscular adipose tissue (IMAT, −190 HU to −30 HU). The distinction between NAMA, LAMA, and IMAT was made to differentiate healthy muscle from fatty infiltration and to assess myosteatosis. This approach allows a more detailed assessment of muscle quality [11,12]. We defined a skeletal muscle area (SMA) by combining NAMA and LAMA. The BMI-adjusted SMA is defined by SMA/BMI. The visceral fat area (VFAT) and the subcutaneous fat area (SFAT) were demarcated using fat tissue thresholds of −190 HU to −30 HU information.

XSL•FO
RenderX

Figure 1. AI algorithm models. CT: computed tomography; FCN: fully convolution network; IMAT: inter/intramuscular adipose tissue; LAMA: low-attenuation muscle area; NAMA: normal-attenuation muscle area; SFAT: subcutaneous fat area; SMA: skeletal muscle area; VFAT: visceral fat area; YOLOv3: You Only Look Once v3.

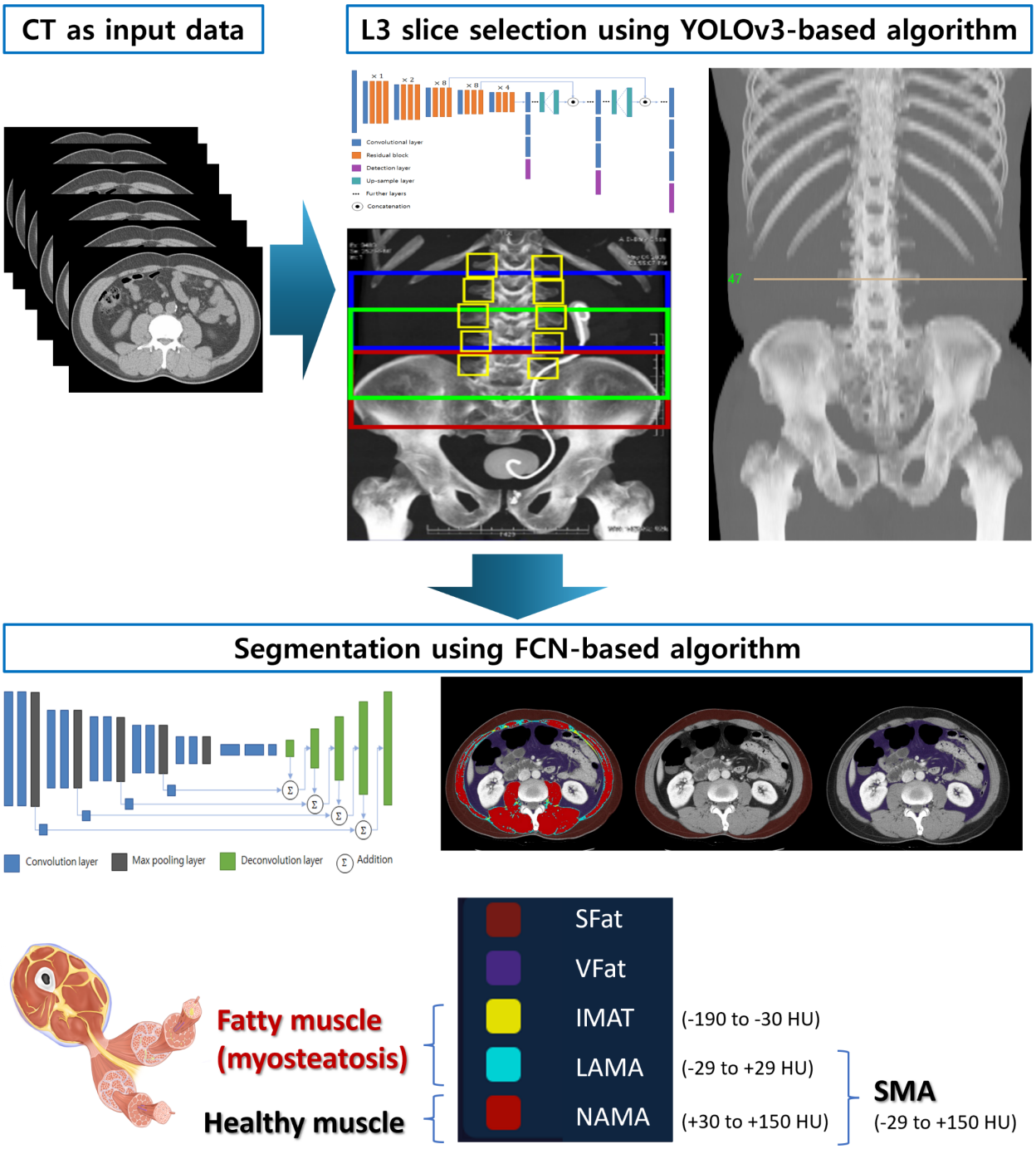

Setup for Fully Automated AI-Integrated System

To ensure seamless compatibility with existing clinical workflows and institutional imaging infrastructures, the AI model (Aid-UTM, iAID Inc.) was directly integrated with the PACS (INFINITT PACS, Infinitt Co.) at the participating health institutions. Processes for the fully automated AI-integrated opportunistic CT screening system were established in a step-by-step manner, following the most recent implementation guidelines, as illustrated in Figure 2 [13].

XSL•FO
RenderX

Figure 2. Step-by-step AI implementation and integration in clinical practice. AI: artificial intelligence; PACS: Picture Archiving and Communication System.

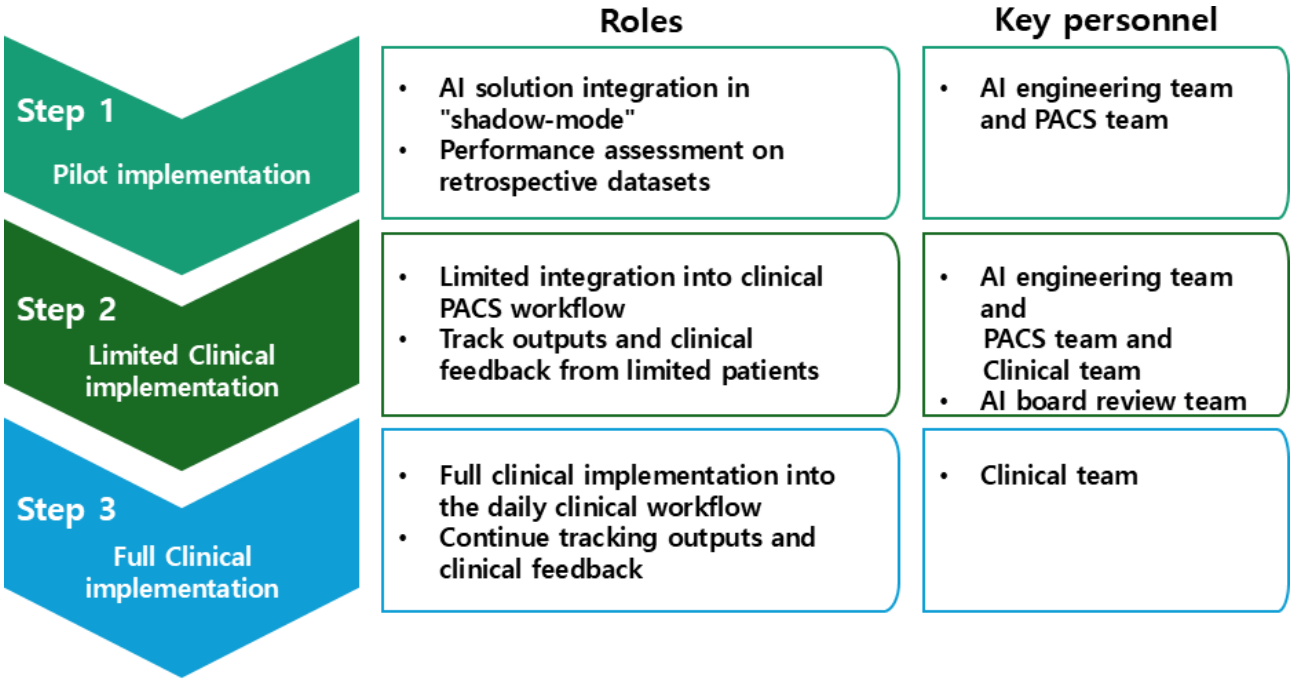

The first step was pilot implementation. A project kick-off meeting was conducted with the AI engineering team, the clinical team from KABP, and the PACS team to ensure seamless coordination. During this phase, the AI solution servers were integrated with the PACS infrastructure. Retrospective testing was performed using test CT datasets (n=10) obtained from the open data repository of the 2022 Body Morphometry AI Segmentation Challenge. The second step involved limited clinical deployment. CT scans from the first 10 patients enrolled in this study were used to evaluate real-time AI processing in a prospective manner. An internal monitoring tool was deployed alongside this step to track CT data transfer and AI processing performance, ensuring robustness and transparency. The final step was full clinical deployment, enabling the system for long-term use in routine health check-up settings. This phase marked the integration of the AI system into daily clinical workflows for opportunistic CT screening, ensuring scalability and sustained performance.

Statistical Analysis

The technical success rate of the AI system was calculated as the proportion of CT scans successfully processed without requiring manual adjustments. The overall processing time was reported as the mean time interval from the completion of CT scan acquisition to the availability of the AI-generated report. Segmentation accuracy was assessed by comparing AI-generated results with manual annotations by radiologists, with agreement quantified using the Dice similarity coefficient (DSC).

To evaluate age-related and gender-specific differences in body composition metrics, 1-way ANOVA was conducted across age groups. $P<.05$ was considered statistically significant.

Our study proposed the concept of muscle age, defined as the estimated age at which the participant's muscle mass corresponds to the average muscle mass of that age group. Muscle age is calculated using data from Kim et al's study [14], which includes 11,845 healthy Korean adults who underwent abdominal CT scans during routine health check-ups. Participants with known malignancy, chronic diseases, or metabolic disorders were excluded. SMA and SMA/BMI values were stratified by sex and age. Participants aged 35 to 65 years, representing the 5th to 95th percentile of the healthy adult population, were used to generate sex-specific regression models for estimating muscle age. First, we regressed SMA/BMI on a constant and age in a sample of participants aged between 35 and 65 years, representing approximately the 5th and 95th percentiles for age in both sexes. Next, we derived the inverse of the regression equation by solving for age as a function of SMA/BMI. The SE of the estimated age was computed using the delta method, based on the original variance-covariance matrix adjusted for heteroscedasticity. Finally, separate regression analyses were performed for men and women, resulting in distinct muscle age functions for each sex. The statistical analyses were conducted using R (version 4.3.0; R Foundation for Statistical Computing).

Ethical Considerations

This study was approved by the Asan Medical Center institutional review board (approval code IRB number 2022-1049; approval date August 4, 2022). The requirement for patient consent was waived.

Results

Participants

A total of 537 consecutive patients (mean age 46.9, SD 13.2 y; 345 males and 192 females) were recruited for the study. During the recruitment period, no participants were excluded. Most patients underwent abdominal CT scans for initial evaluation or follow-up of incidental lesions detected on routine abdominal ultrasonography such as pancreatic cysts, gallbladder polyps,

XSL•FO
RenderX

and focal lesions in the liver, kidney, or spleen. Some patients underwent abdominal CT scans due to increased levels of blood tumor markers. The demographic characteristics of the participants, including height, weight, and BMI, are listed in Table 1.

Table . Demographic characteristics of the study participants by age groups.

Characteristics	Age groups (years)					
	20 - 29	30 - 39	40 - 49	50 - 59	60 - 69	70+
Sex, n						
M[a]	20	98	109	68	35	15
F[b]	20	34	42	51	30	15
Height (cm), mean (SD)						
M	174.6 (5.6)	174.3 (5.3)	174.9 (6.3)	171 (4.4)	169.6 (5.5)	165.5 (6.2)
F	160.4 (5.1)	162.9 (5.8)	161.3 (5.5)	159.1 (4.9)	154.4 (5.1)	154.5 (4.5)
Weight (kg), mean (SD)						
M	80.2 (17.3)	81.6 (14.2)	80.7 (11.9)	74.7 (9.6)	74.6 (8.1)	72.2 (9.5)
F	58.8 (10.4)	59.4 (9.1)	59.6 (8.7)	60 (8.2)	58 (7.3)	59.5 (9.6)
BMI (kg/m^2), mean (SD)						
M	26.2 (5)	26.8 (3.9)	26.3 (3.1)	25.5 (2.8)	26 (2.7)	26.3 (2.6)
F	23 (4.5)	22.4 (3.2)	22.9 (3.2)	23.8 (3.2)	24.3 (3.2)	24.2 (3.9)

[a]M: male.

[b]F: female.

Fully Automated AI-Integrated Opportunistic CT Screening System

The AI algorithm model was implemented on on-premise servers which were successfully integrated into the PACS in 3 institutions. As illustrated in Figure 3, to establish fully automated AI-integrated opportunistic CT screening system, the following processes were established:

1. Automatic retrieval: the system automatically retrieved anonymized CT images from the PACS without requiring manual intervention. Height and weight information was also extracted and integrated into the Digital Imaging and Communications in Medicine (DICOM) metadata.
2. Real-time AI processing: images were processed in real-time to generate detailed body composition metrics, including measurements of skeletal muscle and adipose tissue.
3. Result storage: the processed results were stored back into the PACS in compliance with DICOM standards.

XSL•FO
RenderX

Figure 3. Processes of fully automated AI-integrated opportunistic CT screening system. AI: artificial intelligence; CT: computed tomography; IMAT: inter/intramuscular adipose tissue; LAMA: low-attenuation muscle area; NAMA: normal-attenuation muscle area; PACS: Picture Archiving and Communication System; SFAT: subcutaneous fat area; SMA: skeletal muscle area; VFAT: visceral fat area.

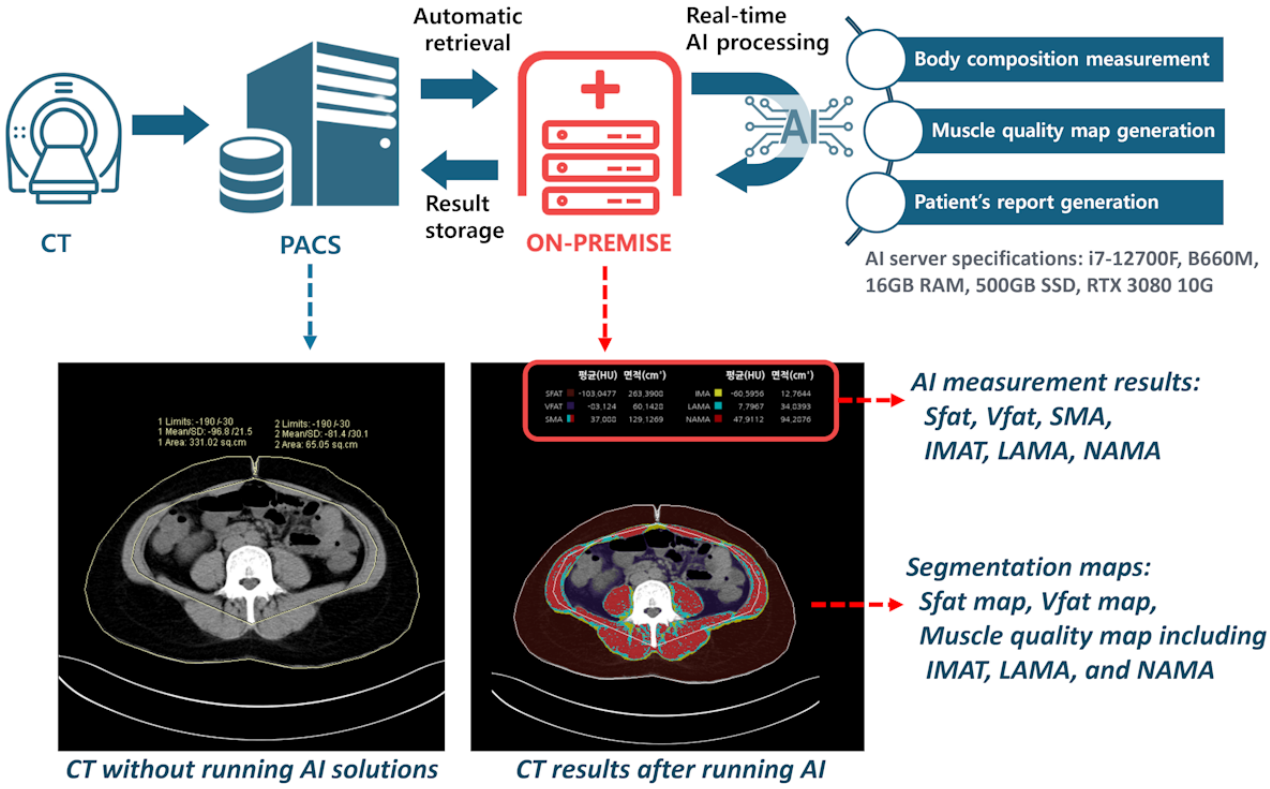

The fully automated AI-integrated system provided detailed body composition metrics, including muscle quality maps and estimated muscle age, as part of the health check-up reports (Figure 4). The integration of the serial analysis function was added for tracking longitudinal changes. In addition, a muscle age calculator was developed so that clinical users can calculate muscle age manually if DICOM metadata does not contain height and weight information (Video S2 in Multimedia Appendix 2; assessed on the web [15]).

XSL•FO
RenderX

Figure 4. Report form of opportunistic sarcopenia screening results. AI: artificial intelligence; CT: computed tomography; HT: height; SFAT: subcutaneous fat area; SMA: skeletal muscle area; VFAT: visceral fat area; WT: weight.

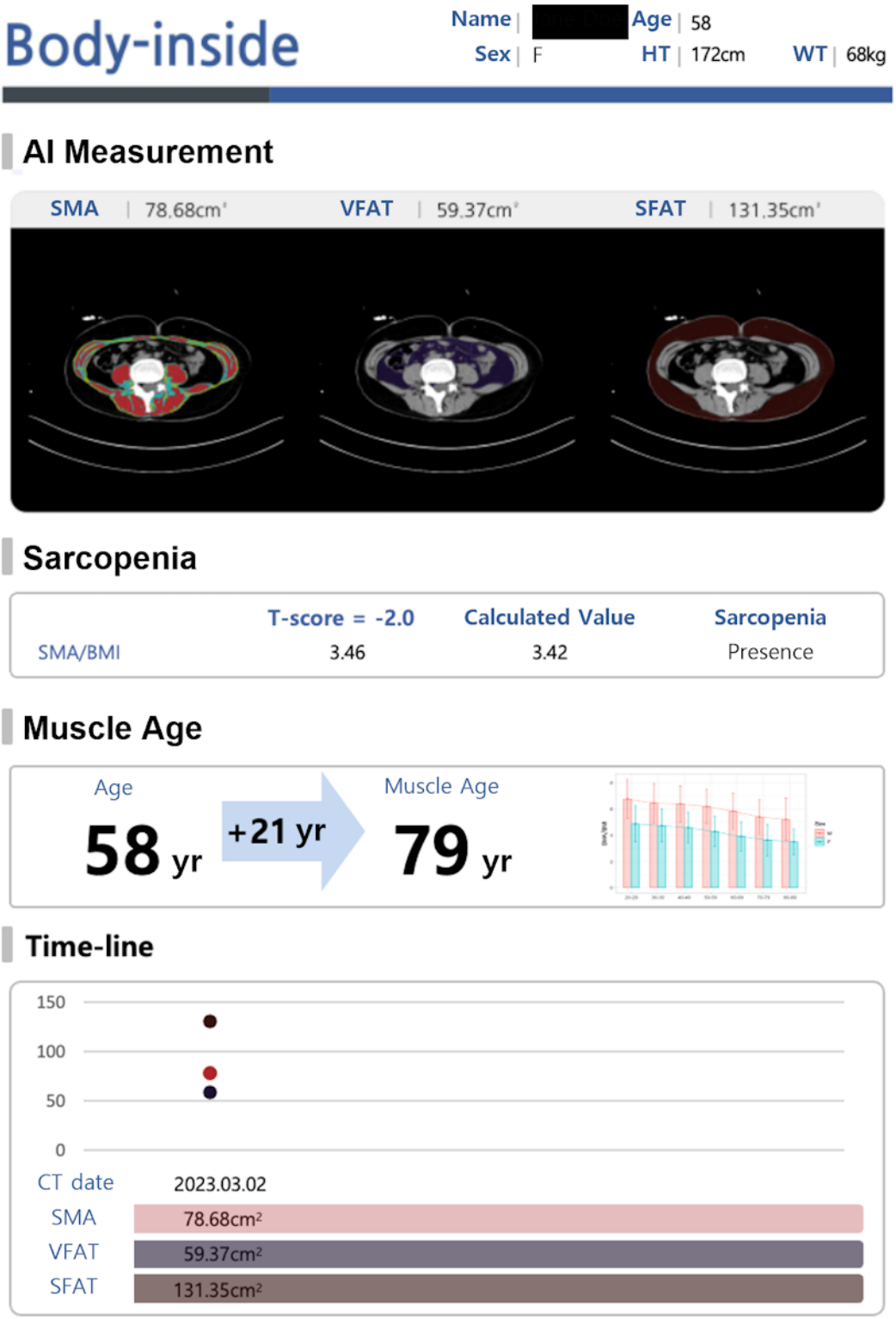

AI System Performance

The AI system demonstrated robust performance across all evaluated metrics. Among the 537 abdominal CT scans analyzed, the technical success rate was 100%, with no cases requiring manual adjustment or failing due to technical errors.

The mean overall processing time from the completion of CT scan acquisition to the generation of AI results was 4.12 (SD 0.15) seconds per scan, confirming the system's real-time applicability.

Segmentation accuracy of the AI algorithm, as assessed by the DSC, was 0.974 (SD 0.011), indicating excellent agreement with manual radiologist annotations.

XSL•FO
RenderX

Opportunistic Screening Results

Results of opportunistic screening for body composition are presented in Table 2. In terms of sarcopenia screening, age-related changes in muscle mass and quality were evident in both male and female participants, as NAMA and SMA, indicators of healthy muscle and total muscle, respectively, consistently declined with age. In contrast, LAMA and IMAT, which reflect fat infiltration and reduced muscle quality, showed significant increases across age groups (Figure 5). These trends align with the natural progression of sarcopenia and myosteatosis, emphasizing the impact of aging on muscle health.

Figure 5. Representative cases with healthy muscles (**A**) and sarcopenia (**B**). (**A**) A 58-year-old female (172 cm, 55 kg) with an SMA/BMI ratio of 6.1, indicating healthy muscle quality, corresponding to the representative SMA/BMI value of a 20-year-old female. (**B**) A 50-year-old female (169 cm, 58 kg) with an SMA/BMI ratio of 4.0, suggesting sarcopenia, corresponding to the representative SMA/BMI value of a 61-year-old female. SFAT: subcutaneous fat area; SMA: skeletal muscle area; VFAT: visceral fat area.

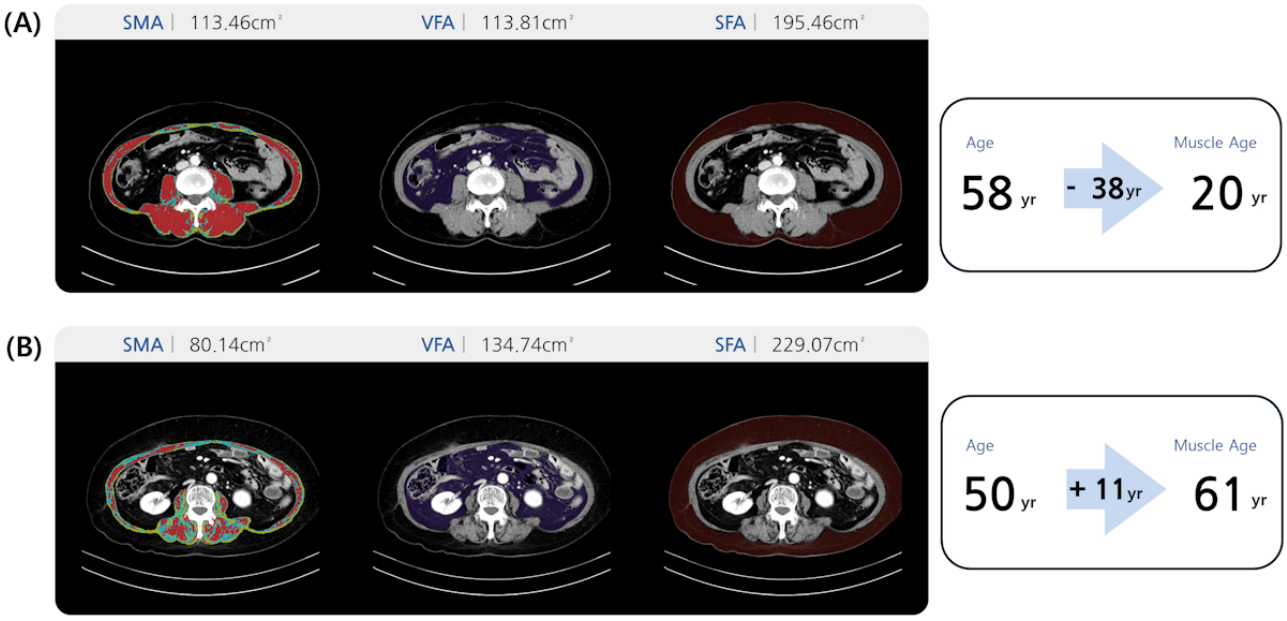

XSL•FO
RenderX

Table . Mean (SD) body composition results in male and female participants across age groups.

	Age groups (years)						
Male	20 - 29	30 - 39	40 - 49	50 - 59	60 - 69	70+	*P* value[a]
NAMA[b]	139.1 (19.8)	132.7 (22.7)	125.7 (20.1)	117.2 (23.2)	106.3 (22)	75.5 (26)	<.001
LAMA[c]	28.8 (15.9)	37.4 (17.5)	40.7 (17)	40.3 (15.5)	46.4 (18.2)	59.7 (14.5)	<.001
SMA[d]	168 (21.7)	170.1 (25.4)	166.4 (22.9)	157.5 (21.7)	152.6 (18.8)	135.3 (23.2)	<.001
SMA/BMI	6.5 (0.9)	6.4 (0.6)	6.3 (0.7)	6.2 (0.7)	5.9 (0.7)	5.1 (0.6)	<.001
IMA[e]	13.3 (7.6)	16.7 (6.9)	16.9 (5.7)	17.2 (5.4)	19 (5.6)	28.4 (7)	<.001
SFAT[f]	200.2 (126.8)	220.6 (90.1)	192.3 (59)	158.9 (52.2)	158.4 (47.5)	168.5 (47)	<.001
VFAT[g]	92.2 (58)	125.5 (56.6)	152.8 (54.2)	146.5 (56.1)	184 (74.1)	213.9 (73.3)	<.001
Female							
NAMA	81.2 (13.1)	85 (11.7)	73.2 (12.7)	69 (15)	57.4 (14.4)	55.3 (13.9)	<.001
LAMA	23.2 (10.8)	24.8 (6.6)	28.6 (8.6)	33.7 (9.6)	40.1 (8.8)	45.9 (10.9)	<.001
SMA	104.5 (12.2)	110 (11.7)	102.7 (12.2)	102.7 (14.6)	97.6 (11.4)	101.5 (9.5)	<.01
SMA/BMI	4.6 (0.7)	5 (0.7)	4.5 (0.5)	4.3 (0.5)	4.1 (0.6)	4.2 (0.5)	<.001
IMA	10.7 (5.3)	12.4 (4.2)	15.7 (6.6)	17.5 (5.4)	19.9 (5.1)	23.9 (7.1)	<.001
SFAT	192.4 (121.7)	183.1 (70.1)	173.2 (67.9)	180.3 (57.1)	176.2 (60.3)	185.4 (60)	.94
VFAT	45 (23.8)	53.1 (29)	68 (33.7)	89 (41.4)	107 (47.6)	132.5 (61.3)	<.001

[a]*P* values from 1-way ANOVA.

[b]NAMA: normal-attenuation muscle area.

[c]LAMA: low-attenuation muscle area.

[d]SMA: skeletal muscle area.

[e]IMA: intramuscular adipose tissue.

[f]SFAT: subcutaneous fat area.

[g]VFAT: visceral fat area.

Regarding obesity screening, VFAT increases with age in both genders, highlighting the trend of age-related visceral fat accumulation, a known risk factor for metabolic disorders. SFAT, however, exhibited differing trends between males and females. In males, SFAT decreased in older age groups, while in females, it remained relatively stable across all ages, suggesting a relative preservation of subcutaneous fat. These differences may point to sex-specific mechanisms in fat storage and distribution.

Discussion

Principal Results

This prospective multicenter study evaluated the implementation of a fully automated AI-integrated system of body composition assessment for opportunistic CT screening during routine health check-ups. The AI system demonstrated robust performance, with a 100% technical success rate, an average processing time of 4.12 seconds per scan, and high segmentation accuracy (DSC: mean 0.9740.011). The AI system provided comprehensive body composition metrics, including muscle quality maps and estimated muscle age, seamlessly integrated into health check-up reports. These findings validate the reliability, efficiency, and scalability of the AI system for opportunistic sarcopenia screening.

The study revealed significant age-related changes in muscle quality and fat distribution, consistent with the progression of sarcopenia and myosteatosis [11,12,16]. In both males and females, NAMA and SMA declined with advancing age, reflecting a reduction in muscle mass and quality. In contrast, LAMA and IMA, markers of muscle fat infiltration, increased significantly with age. These results underscore the capability of the AI system to detect and quantify meaningful variations in body composition, aligning with physiological expectations and clinical relevance [1].

We used BMI-adjusted skeletal muscle index (SMA/BMI) in our study. Although height-adjusted skeletal muscle index (SMA/height^2) is widely used in Western populations, previous research in a large Korean cohort demonstrated that the BMI-adjusted index (SMA/BMI) better reflects age-related muscle decline. Therefore, we adopted SMA/BMI in this study [14].

Notably, our study proposed the concept of muscle age, which provides an individualized estimate of physiological muscle health by projecting the participant's muscle mass onto the age group with a similar average muscle mass. This metric allows for direct comparison between an individual's muscle mass and population norms, enabling more nuanced assessment beyond

XSL•FO
RenderX

traditional metrics such as chronological age or raw muscle mass, like the concept of biological age [17].

Comparison With Prior Work

There have been numerous retrospective studies which have assessed AI systems for body composition analysis in terms of segmentation accuracy, technical adequacy, and prognostic utility [11,12,18-20]. Retrospective studies, though valuable, often lack the operational insights necessary to understand the challenges and benefits of integrating AI into real-time clinical workflows. To date, only a few prospective studies have explored the use of AI-based systems in specific clinical contexts, such as routine abdominal CT scans for pancreas cancer follow-up or lung cancer screening programs [21,22].

To the best of our knowledge, this is the first prospective multicenter study to implement a fully automated AI-integrated system for opportunistic screening of body composition in health check-up institutions. By evaluating the system's performance across diverse institutions, this study bridges the gap between research applications and practical clinical implementation, highlighting the readiness of AI technologies to transition from the research setting to routine health care practice [13]. The successful integration of the AI system into clinical workflows demonstrates its potential use in routine health check-ups by automating time-intensive tasks and report generation. Furthermore, the system's ability to detect early signs of sarcopenia and myosteatosis offers significant clinical value, as timely interventions can prevent progression to frailty and associated complications [6]. Despite these promising results, integrating AI into institutional PACS presented several challenges, including ensuring seamless data transfer, maintaining compliance with DICOM standards, and obtaining clinical team acceptance. A phased deployment approach and early collaboration between AI engineers, radiologists, and PACS teams were essential to overcoming these barriers. For real-world implementation, we are preparing regulatory approval as a software as a medical device and ensuring compliance with regulations such as the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR).

Limitations

This study has several limitations. First, the study population consisted of participants undergoing routine health check-ups, which may limit the generalizability of the findings to other clinical settings, such as acute care or cancer follow-up. Second, while the AI system demonstrated high segmentation accuracy and efficiency, the study relied on standardized imaging protocols across the participating institutions. Variations in CT acquisition protocols or differences in scanner types at other facilities might affect the system's performance, necessitating additional validation in diverse imaging environments. Finally, the study focused on technical performance, without incorporating long-term patient outcomes or clinical decision-making for patient care. Despite these limitations, this study represents an important step toward integrating fully automated AI systems into routine clinical practice, offering a foundation for future research to address these gaps.

Conclusions

This study demonstrated the successful implementation of a fully automated AI-integrated system for body composition assessment using opportunistic CT screening during routine health check-ups, achieving high accuracy, efficiency, and scalability. Significant age-related changes in muscle quality and fat distribution were identified, highlighting the system's clinical utility in detecting sarcopenia and myosteatosis.

Acknowledgments

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant (RS-2025 - 02219599) and Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI) grant (RS-2018-KH049509).

Data Availability

The datasets generated and analyzed during the current study are not publicly available due to the security of the data but are available from the corresponding author upon reasonable request.

Authors' Contributions

KWK, YK, IOK, EHL, and SC designed this study. YK and BU wrote the initial draft of this manuscript. YK, SN, HJ, and SK were involved in data collection. BU, YK, and HK performed the statistical analysis. KWK, BU, YK, SN, IOK, EHL, SC, HK, HJ, SK, and JL contributed to the interpretation and revision of this article and jointly completed the final draft.

Conflicts of Interest

KWK, JL, HJ, and SK are patent holders.

Multimedia Appendix 1
Video S1: Demo video of the AI solution for automatic L3 level selection and body composition assessment.
[MP4 File, 81561 KB - formative_v9i1e69940_app1.mp4]

XSL•FO
RenderX

Multimedia Appendix 2
Video S2: Demo video of the web muscle age calculator, which can be assessed on the web [15].
[MP4 File, 506 KB - formative_v9i1e69940_app2.mp4]

References

1. Kim KW, Huh J, Urooj B, et al. Artificial intelligence in gastric cancer imaging with emphasis on diagnostic imaging and body morphometry. J Gastric Cancer 2023 Jul;23(3):388-399. [doi: 10.5230/jgc.2023.23.e30] [Medline: 37553127]
2. Cruz-Jentoft AJ, Landi F, Schneider SM, et al. Prevalence of and interventions for sarcopenia in ageing adults: a systematic review. Report of the International Sarcopenia Initiative (EWGSOP and IWGS). Age Ageing 2014 Nov;43(6):748-759. [doi: 10.1093/ageing/afu115] [Medline: 25241753]
3. Lee K, Shin Y, Huh J, et al. Recent issues on body composition imaging for sarcopenia evaluation. Korean J Radiol 2019 Feb;20(2):205-217. [doi: 10.3348/kjr.2018.0479] [Medline: 30672160]
4. Kim DW, Kim KW, Ko Y, et al. Effects of contrast phases on automated measurements of muscle quantity and quality using CT. Korean J Radiol 2021 Nov;22(11):1909-1917. [doi: 10.3348/kjr.2021.0105] [Medline: 34431247]
5. Pickhardt PJ, Summers RM, Garrett JW. Automated CT-based body composition analysis: a golden opportunity. Korean J Radiol 2021 Dec;22(12):1934-1937. [doi: 10.3348/kjr.2021.0775] [Medline: 34719894]
6. Pickhardt PJ, Graffy PM, Perez AA, Lubner MG, Elton DC, Summers RM. Opportunistic screening at abdominal CT: use of automated body composition biomarkers for added cardiometabolic value. Radiographics 2021;41(2):524-542. [doi: 10.1148/rg.2021200056] [Medline: 33646902]
7. Graffy PM, Liu J, Pickhardt PJ, Burns JE, Yao J, Summers RM. Deep learning-based muscle segmentation and quantification at abdominal CT: application to a longitudinal adult screening cohort for sarcopenia assessment. Br J Radiol 2019 Aug;92(1100):20190327. [doi: 10.1259/bjr.20190327] [Medline: 31199670]
8. Lee SJ, Liu J, Yao J, Kanarek A, Summers RM, Pickhardt PJ. Fully automated segmentation and quantification of visceral and subcutaneous fat at abdominal CT: application to a longitudinal adult screening cohort. Br J Radiol 2018 Sep;91(1089):20170968. [doi: 10.1259/bjr.20170968] [Medline: 29557216]
9. Park HJ, Shin Y, Park J, et al. Development and validation of a deep learning system for segmentation of abdominal muscle and fat on computed tomography. Korean J Radiol 2020 Jan;21(1):88-100. [doi: 10.3348/kjr.2019.0470] [Medline: 31920032]
10. Ha J, Park T, Kim HK, et al. Development of a fully automatic deep learning system for L3 selection and body composition assessment on computed tomography. Sci Rep 2021 Nov 4;11(1):21656. [doi: 10.1038/s41598-021-00161-5] [Medline: 34737340]
11. Kim DW, Kim KW, Ko Y, et al. Assessment of myosteatosis on computed tomography by automatic generation of a muscle quality map using a web-based toolkit: feasibility study. JMIR Med Inform 2020 Oct 19;8(10):e23049. [doi: 10.2196/23049] [Medline: 33074159]
12. Ahn H, Kim DW, Ko Y, et al. Updated systematic review and meta-analysis on diagnostic issues and the prognostic impact of myosteatosis: a new paradigm beyond sarcopenia. Ageing Res Rev 2021 Sep;70:101398. [doi: 10.1016/j.arr.2021.101398] [Medline: 34214642]
13. Daye D, Wiggins WF, Lungren MP, et al. Implementation of clinical artificial intelligence in radiology: who decides and how? Radiology 2022 Dec;305(3):555-563. [doi: 10.1148/radiol.212151] [Medline: 35916673]
14. Kim EH, Kim KW, Shin Y, et al. Reference data and T-Scores of lumbar skeletal muscle area and its skeletal muscle indices measured by CT scan in a healthy Korean population. J Gerontol A Biol Sci Med Sci 2021 Jan 18;76(2):265-271. [doi: 10.1093/gerona/glaa065] [Medline: 32179888]
15. How Old Are You MuscleSize? - iAID. iAID. URL: https://iaidimage.co.kr/service_muscleage [accessed 2025-07-22]
16. Jang M, Park HW, Huh J, et al. Predictive value of sarcopenia and visceral obesity for postoperative pancreatic fistula after pancreaticoduodenectomy analyzed on clinically acquired CT and MRI. Eur Radiol 2019 May;29(5):2417-2425. [doi: 10.1007/s00330-018-5790-7] [Medline: 30406311]
17. Li Z, Zhang W, Duan Y, et al. Progress in biological age research. Front Public Health 2023;11(1074274):1074274. [doi: 10.3389/fpubh.2023.1074274] [Medline: 37124811]
18. Pooler BD, Garrett JW, Southard AM, Summers RM, Pickhardt PJ. Technical adequacy of fully automated artificial intelligence body composition tools: assessment in a heterogeneous sample of external CT examinations. Am J Roentgenol 2023 Jul;221(1):124-134. [doi: 10.2214/AJR.22.28745] [Medline: 37095663]
19. Bates DDB, Pickhardt PJ. CT-derived body composition assessment as a prognostic tool in oncologic patients: from opportunistic research to artificial intelligence-based clinical implementation. Am J Roentgenol 2022 Oct;219(4):671-680. [doi: 10.2214/AJR.22.27749] [Medline: 35642760]
20. Kim DW, Ahn H, Kim KW, et al. Prognostic value of sarcopenia and myosteatosis in patients with resectable pancreatic ductal adenocarcinoma. Korean J Radiol 2022 Nov;23(11):1055-1066. [doi: 10.3348/kjr.2022.0277] [Medline: 36098341]
21. Hsu TMH, Schawkat K, Berkowitz SJ, et al. Artificial intelligence to assess body composition on routine abdominal CT scans and predict mortality in pancreatic cancer- a recipe for your local application. Eur J Radiol 2021 Sep;142:109834. [doi: 10.1016/j.ejrad.2021.109834] [Medline: 34252866]

XSL•FO
RenderX

22. Ledda RE, Sabia F, Valsecchi C, et al. The added value of an AI-based body composition analysis in a lung cancer screening population: preliminary results. Nutr Metab Cardiovasc Dis 2025 Feb;35(2):103696. [doi: 10.1016/j.numecd.2024.07.013] [Medline: 39278738]

Abbreviations

AI: artificial intelligence
CT: computed tomography
DICOM: Digital Imaging and Communications in Medicine
DSC: Dice similarity coefficient
GDPR: General Data Protection Regulation
HIPAA: Health Insurance Portability and Accountability Act
HU: Hounsfield Unit
IMAT: inter- and intramuscular adipose tissue
KAHP: Korea Association of Health Promotion
LAMA: low-attenuation muscle area
NAMA: normal-attenuation muscle area
PACS: Picture Archiving and Communication System
SFAT: subcutaneous fat area
SMA: skeletal muscle area
VFAT: visceral fat area
YOLOv3: You Only Look Once v3

Edited by A Mavragani; submitted 11.12.24; peer-reviewed by MYK Wei, R Brecheisen; revised version received 30.05.25; accepted 02.06.25; published 05.09.25.

Please cite as:
Urooj B, Ko Y, Na S, Kim IO, Lee EH, Cho S, Jeong H, Khang S, Lee J, Kim KW
Implementation of Fully Automated AI-Integrated System for Body Composition Assessment on Computed Tomography for Opportunistic Sarcopenia Screening: Multicenter Prospective Study
JMIR Form Res 2025;9:e69940
URL: https://formative.jmir.org/2025/1/e69940
doi:10.2196/69940

© Bushra Urooj, Yousun Ko, Seongwon Na, In-One Kim, Eun-Hee Lee, Seon Cho, Heeryeol Jeong, Seungwoo Khang, Jeongjin Lee, Kyung Won Kim. Originally published in JMIR Formative Research (https://formative.jmir.org), 5.9.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

AI Model Based on Diaphragm Ultrasound to Improve the Predictive Performance of Invasive Mechanical Ventilation Weaning: Prospective Cohort Study

Feier Song[1*], MD; Huazhang Liu[2*], MS; Huan Ma[3], MD; Xuanhui Chen[2], MS; Shouhong Wang[4], MD; Tiehe Qin[4], MD; Huiying Liang[2], PhD; Daozheng Huang[4,5], MD

[1]Department of Emergency and Intensive Care Unit, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, Guangzhou, China

[2]Medical Big Data Center, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, Number 106, Zhongshaner Road, Guangzhou, China

[3]Department of Cardiology, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, Guangdong Provincial Cardiovascular Institute, Guangzhou, China

[4]Department of Critical Care Medicine, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, Guangdong Provincial Geriatrics Institute, No. 106, Zhongshaner Rd, Guangzhou, China

[5]Office of Organ Procurement Organizations, Medical Department, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, Guangzhou, China

[*]these authors contributed equally

Corresponding Author:
Daozheng Huang, MD
Department of Critical Care Medicine, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, Guangdong Provincial Geriatrics Institute, No. 106, Zhongshaner Rd, Guangzhou, China

Abstract

Background: Point-of-care ultrasonography has become a valuable tool for assessing diaphragmatic function in critically ill patients receiving invasive mechanical ventilation. However, conventional diaphragm ultrasound assessment remains highly operator-dependent and subjective. Previous research introduced automatic measurement of diaphragmatic excursion and velocity using 2D speckle-tracking technology.

Objective: This study aimed to develop an artificial intelligence–multimodal learning framework to improve the prediction of weaning failure and guide individualized weaning strategies.

Methods: This prospective study enrolled critically ill patients older than 18 years who received mechanical ventilation for more than 48 hours and were eligible for a spontaneous breathing trial in 2 intensive care units in Guangzhou, China. Before the spontaneous breathing trial, diaphragm ultrasound videos were collected using a standardized protocol, and automatic measurements of excursion and velocity were obtained. A total of 88 patients were included, with 50 successfully weaned and 38 experiencing weaning failure. Each patient record included 27 clinical and 6 diaphragmatic indicators, selected based on previous literature and phenotyping studies. Clinical variables were preprocessed using OneHotEncoder, normalization, and scaling. Ultrasound videos were interpolated to a uniform resolution of 224×224×96. Artificial intelligence–multimodal learning based on clinical characteristics, laboratory parameters, and diaphragm ultrasonic videos was established. Four experiments were conducted in an ablation setting to evaluate model performance using different combinations of input data: (1) diaphragmatic excursion only, (2) clinical and diaphragmatic indicators, (3) ultrasound videos only, and (4) all modalities combined (multimodal). Metrics for evaluation included classification accuracy, area under the receiver operating characteristic curve (AUC), average precision in the precision-recall curve, and calibration curve. Variable importance was assessed using SHAP (Shapley Additive Explanation) to interpret feature contributions and understand model predictions.

Results: The multimodal co-learning model outperformed all single-modal approaches. The accuracy improved when predicted through diaphragm ultrasound video data using Video Vision Transformer (accuracy=0.8095, AUC=0.852), clinical or ultrasound indicators (accuracy=0.7381, AUC=0.746), and the multimodal co-learning (accuracy=0.8331, AUC=0.894). The proposed co-learning model achieved the highest score (average precision=0.91) among the 4 experiments. Furthermore, calibration curve analysis demonstrated that the proposed colearning model was well calibrated, as the curve was closest to the perfectly calibrated line.

Conclusions: Combining ultrasound and clinical data for colearning improved the accuracy of the weaning outcome prediction. Multimodal learning based on automatic measurement of point-of-care ultrasonography and automated collection of objective

XSL•FO
RenderX

clinical indicators greatly enhanced the practical operability and user-friendliness of the system. The proposed model offered promising potential for widespread clinical application in intensive care settings.

(JMIR Form Res 2025;9:e72482) doi:10.2196/72482

KEYWORDS

deep learning; multimodal learning; point-of-care ultrasound; POCUS; critical care

Introduction

Invasive mechanical ventilation (MV) is a life-saving therapeutic strategy in critically ill patients, yet 20% - 40% of invasive MV patients endure difficult weaning to varying extents [1]. WEAN-SAFE study, a prospective cohort of 10,232 invasive MV patients in 481 intensive care units (ICU) among 50 countries, showed that the incidence of difficult weaning, prolonged weaning, and weaning failure reached up to 35.3% [2]. Diaphragm dysfunction is a major concern in critically ill patients due to its potential association with prolonged MV dependency [3,4], as the diaphragm plays a central role in respiratory mechanics.

The use of point-of-care ultrasonography (POCUS) to assess the diaphragm has seen rapid growth within the ICU because it is readily available, portable, noninvasive, visual, easily repeatable, and free of ionizing radiation [5]. Previous systematic reviews have examined the accuracy and usefulness of diaphragmatic ultrasound indicators in predicting weaning outcomes, mainly focusing on diaphragm thickening fraction (DTF), excursion, and rapid shallow breathing index [6-10]. Encouragingly, our serial research, published in January 2023, developed an automatic measurement of diaphragm excursion and velocity based on 2D speckle-tracking technology, showing that the mean diaphragm excursion and velocity had a high diagnostic value on the weaning outcomes [11].

In parallel, multimodal artificial intelligence (AI), or multimodal learning, enabling the AI/machine learning (ML) model to learn from and process multiple modes and types of data (image, text, audio, and video) rather than just one, is a rising trend and has the potential to reshape both the AI landscape and clinical research for its novel implications and benefits. By integrating AI into the evaluation process, clinicians can receive more accurate and consistent predictions, allowing for better decision-making. Furthermore, AI-guided prediction models can adapt to the individual characteristics of patients, offering personalized insights that improve the safety and efficacy of the weaning process [12,13].

To further extend our findings, this study aims to develop and validate a multimodal AI model that integrates clinical characteristics, laboratory parameters, and diaphragm ultrasound videos from the abovementioned cohort. This might provide multisystem and multiorgan morphological and functional status information, which helps predict weaning failure and guide the personalized, accurate weaning strategies in the ICU.

Methods

Study Design

Patients aged older than 18 years who received MV for more than 48 hours, suitable for a spontaneous breathing trial (SBT), were prospectively enrolled in ICUs within 2 hospitals in Guangzhou City, China. Ultrasound video and diaphragm excursion and velocity were acquired, followed by a standardized protocol using the TE7 Diagnostic Ultrasound System (C5-2 array probe, Shenzhen Mindray Biomedical) in a supine or semirecumbent position before SBT by a well-trained expert. Of the total 88 critically ill patients, 50 (57%) patients maintained spontaneous breathing for ≥48 h with no need for any level of ventilator support after extubating, while 38 (43%) patients failed to wean. Clinical indicators, diaphragm ultrasound video, and diaphragmatic indicators (using automatic speckle-tracking measurement) were collected to construct the clinical/diaphragmatic dataset and diaphragm ultrasound dataset.

Indicators Selection

Among the clinical or diaphragmatic indicators, each case contained 27 clinical and 6 diaphragmatic indicators (Table S1 of Multimedia Appendix 1). In this study, we selected height, weight, BMI, tidal volume, positive end-expiratory pressure, albumin, hemoglobin, age, temperature, systolic blood pressure, diastolic blood pressure, heart rate, respiratory rate, FiO2, PaO2, PaCO2, pH, serum sodium, serum potassium, white blood cell count, serum creatinine, serum calcium, serum magnesium, serum phosphorus, APACHE II score, SOFA (Sequential Organ Failure Assessment), and N-terminal pro-brain natriuretic peptide. In our previous retrospective study published in 2023 [14], clinical phenotypes of ventilated ICU patients to predict outcomes were derived from the eICU (electronic intensive care unit) Collaborative Research Database cohort and validated in the MIMIC-IV (Medical Information Mart for Intensive Care) cohort. Four clinical phenotypes were identified, which correlate with different clinical characteristics and outcomes, including 28-day mortality and extubation success. A total of 24 clinical variables were selected, including patient demographics, vital signs, blood gas analysis, laboratory measurements, scores, and MV parameters for phenotyping. Further, 18 of 24 variables in the abovementioned study were collected in the present cohort. Demographics were collected as baseline data upon ICU admission. Arterial blood gas analysis was performed at the end of the SBT. Ventilator parameters were recorded during the ultrasound examination. Additionally, a systematic review identified 21 factors associated with increased risk for reintubation, including higher APACHE II scores, low hemoglobin, low albumin, and high brain natriuretic peptide [15]. A case was reported in 2022 of iatrogenic hypermagnesemia, which presented as respiratory depression,

XSL•FO
RenderX

preventing weaning from MV following cardiac surgery [16]. Hypophosphatemia was associated with respiratory muscular weakness, leading to weaning failure [17,18]. SOFA score during the first 24 hours of ICU admission represented significant independent predictors of diaphragm dysfunction and thus led to difficulties weaning from MV [19]. Laboratory markers, APACHE II, and SOFA scores were collected within 24 h before extubation. Furthermore, 6 diaphragmatic indicators were calculated using automatic speckle-tracking measurements based on the diaphragm ultrasound video. The detailed method and material were published previously [11].

Data

For the diaphragm ultrasound videos, they were recorded with a high resolution of 900*1100 pixels and the length ranged from 48 to 556 frames. The percentage of the videos of 48 - 150 frames is 30% and videos of 350 - 556 frames is 30%. The overview of the patients and the details of the ultrasound video frame were presented in Table 1. During data preprocessing, the numeric variables and the categorical variables within the clinical indicators were preprocessed separately using 1-hot encoding, scaling, and normalization methods. The ultrasound videos were interpolated to a uniform size of 224*224*96. The size was unified considering the diaphragmatic excursion cycle and the model structure. The interpolation method was adapted from the bilinear interpolation algorithm in the Python (Python Software Foundation) package *SciPy*. Additionally, the data were normalized to a specific range of (0,1) before being encoded by the model.

Table . Overview of demographics, outcomes, and the characteristics of ultrasound video (N=88).

		Values
Age, mean (SD)		73 (14)
BMI, mean (SD)		22.40 (3.12)
Gender, n (%)		
	Male	60 (68)
	Female	28 (32)
Weaning outcome, n (%)		
	Success	50 (57)
	Failure	38 (43)
Ultrasound length (frame), n (%)		
	0 - 50	8 (9)
	50 - 150	17 (21)
	150 - 250	16 (25)
	250 - 350	28 (44)
	350 - 450	13 (21)
	450 - 560	6 (10)

Multimodal Colearning Model

A multimodal colearning model was then constructed (Figure 1), which simultaneously encoded both the video and clinical data and subsequently predicted weaning success from the integrated features. In this case, features would come from both the ultrasound videos (eg, image-based features) and the clinical indicators. The co-learning model had 3 components. First, Video Vision Transformer (ViViT) [20] was run as the ultrasound data encoder (Figure 2), aiming at extracting the ultrasound features. It included a spatial information encoder and a temporal information encoder based on the attention mechanism. Moreover, the ViViT captured temporal and spatial information at the same time, and for long videos, it also realized long-range sequence interaction through an attention mechanism. A small-scale grid search was used for pretraining. The learning rate for ViViT was between 1e-5 and 1e-4. The dropout rate was 0.5 to 0.7. We experimented with batch sizes of 4, 8, and 16 and set weight decay between 1e-4 and 1e-2. We trained the model for 50 to 100 epochs and applied an early stopping strategy (patience) with a range of 10 to 15 epochs.

Second, for the clinical and diaphragmatic indicators, a multilayer perceptron model, consisting of a sequence of fully connected layers and dropout layers, was adopted for clinical data encoding. Finally, in the fusion module, the ultrasound and clinical features were integrated. A multilayer perceptron classifier was used to obtain the final weaning outcome. In the multilayer perceptron classifier, 2 dropout layers were applied to avoid overfitting.

In the 5-fold cross-validation process, the 88 cases were randomly divided into training sets (n=60) and testing sets (n=18) in the ratio of 8:2. In each fold, 60 cases were randomly chosen as the training set, while the test set was the remaining 18 cases. In each fold, the model was trained on the training set and tested on the test set. The final result was calculated using the average results in the 5 folds. In the training process, the early stopping method was managed to avoid overfitting and improve model generalization performance. Receiver operating

XSL•FO
RenderX

characteristic–area under the curve (AUC), precision-recall curve, and calibration curve for deep learning models were used to evaluate the results of the variant models. The model was developed and fit using Keras (Google LLC). The hyperparameters were optimized in the training process, with balanced class weights, binary cross-entropy loss, and 0.005 learning rate using the Adam optimizer.

Variable importance analysis was conducted to identify the key features influencing the prediction of weaning outcomes and to assess their relative impact. In deep learning models, currently, no universally accepted sensitivity analysis method fits all types of models. Therefore, we used the SHAP (Shapley Additive Explanation) method, a powerful technique for explaining individual predictions made by complex models. SHAP was applied to the classification of clinical indicators. Categorical features were processed using OneHotEncoder, and continuous features were normalized. SHAP values provide a measure of the contribution of each feature to a particular prediction, quantifying both the magnitude and direction (positive or negative) of their influence on the model's output. SHAP values enabled a deeper understanding of the model's behavior and the sensitivity of predictions to changes in individual indicators [21]. It should be noted that in this study, we incorporated SHAP to understand the importance of different features and enhance the interpretability, rather than to investigate causal relationships in clinical scenarios. SHAP plots generally do not directly show statistical significance (such as *P* values). To observe whether there is statistical significance, it needs to be verified through statistical tests. SHAP analysis can serve as a computer-aided diagnosis tool for clinicians, and its application requires integration with clinicians' domain expertise to achieve better diagnostic outcomes.

Figure 1. Overview of the proposed multimodal co-learning model. The model included 3 components: ViViT (ultrasound encoder), multilayer perceptron (indicator encoder), and Fusion module. MLP: Multilayer Perceptron; NT-proBNP: N-terminal pro–B-type natriuretic peptide; SOFA: Sequential Organ Failure Assessment; ViViT: Video Vision Transformer.

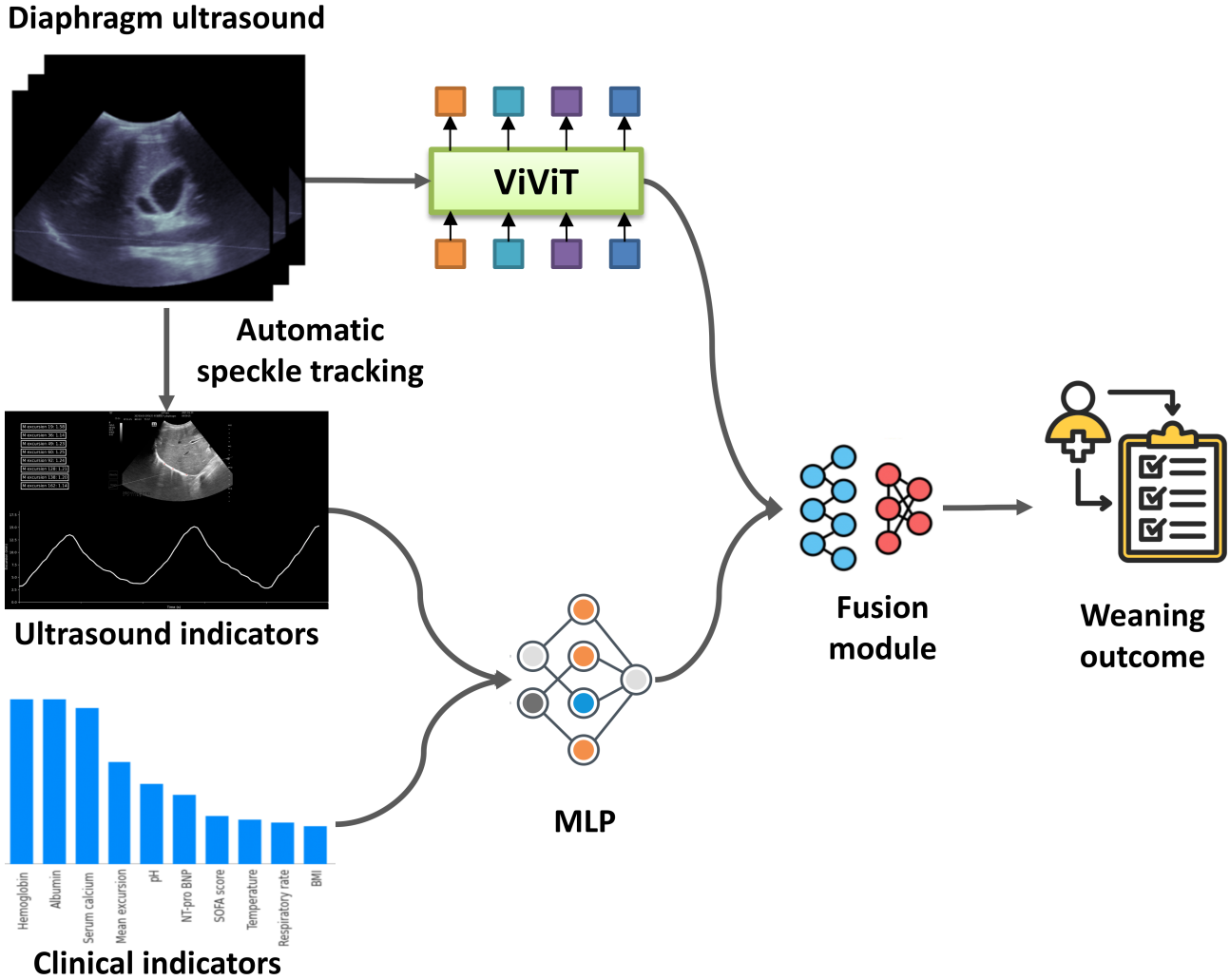

XSL•FO
RenderX

Figure 2. Structure of Video Vision Transformer. Ultrasound data were encoded into spatial feature (si) and temporal feature (ti) and processed by a spatial-temporal encoder. MLP: Multilayer Perceptron.

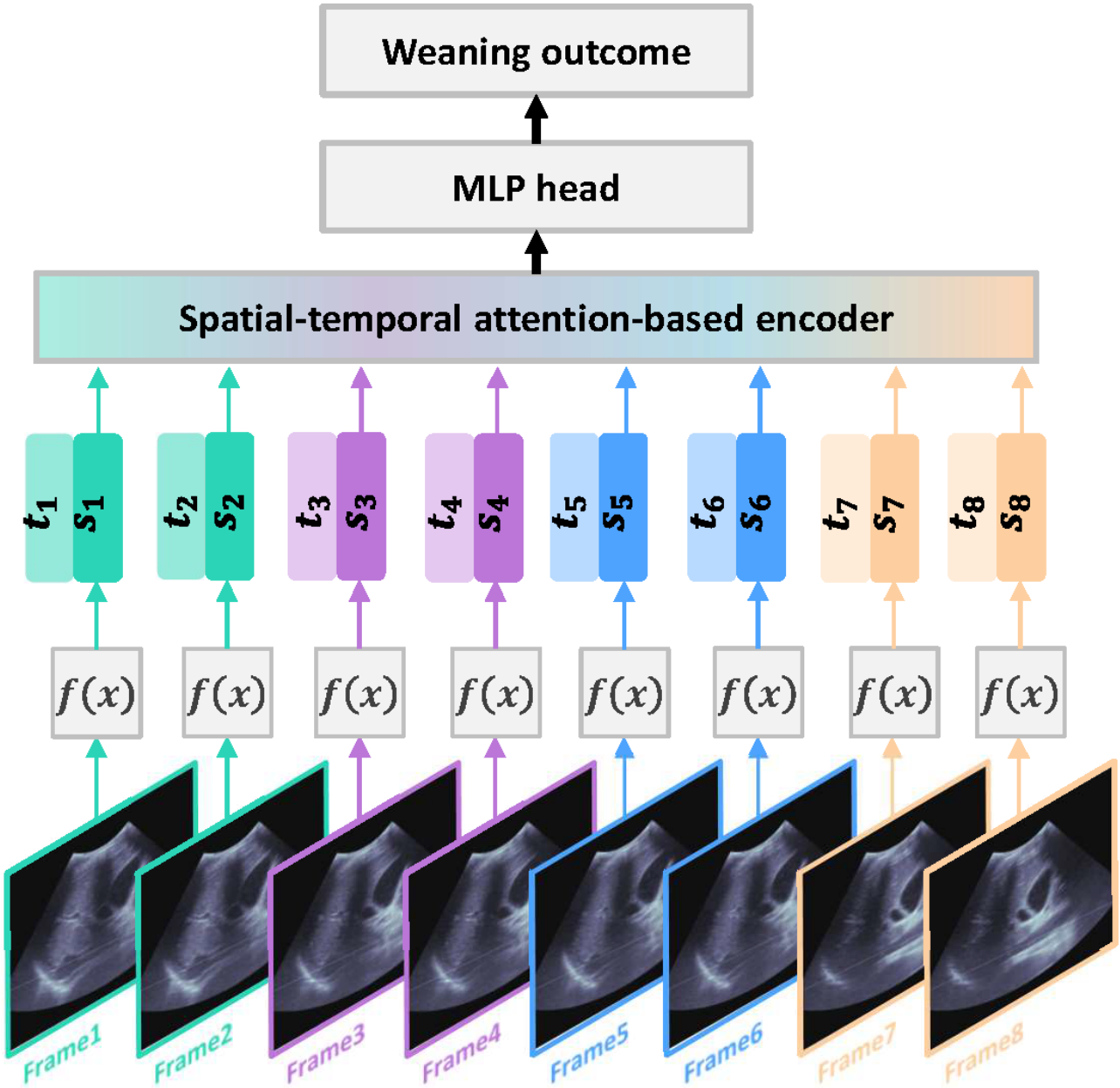

Ethical Considerations

All subjects' families provided written informed consent. The study was performed following the approval of the ethics committee of Guangdong Provincial People's Hospital (No. 2020-246H-1). All procedures were followed per the ethical standards of the responsible committee on human experimentation (institutional or regional) and with the Helsinki Declaration. Measures were taken to ensure participant privacy, including data anonymization and secure data storage. No financial or material compensation was provided to the participants or their families for participation in this study.

Results

Four experiments were conducted on the prediction of weaning failure in an ablation setting, testing each component of the colearning model with corresponding data. To demonstrate the advantages of the proposed co-learning model in terms of modal interaction, the prediction performance was compared in a multimodal setting with 3 single-modal settings (diaphragm excursion, clinical or diaphragmatic indicators, and diaphragm ultrasound video). The 88 cases of ultrasound and clinical data were divided into training sets and test sets at a ratio of 8:2, and a 5-fold cross-validation was performed (Table 2). The results of the 4 experiments were demonstrated by classification accuracy, receiver operating characteristic curves, precision-recall curve, and calibration curve.

Initially, using automatic mean excursion as the sole predictor yielded a moderate performance with an AUC score of 0.659 [11], while the use of diaphragm ultrasound video data, processed with the ViViT model, resulted in an enhanced prediction performance, achieving a notable accuracy of 0.8095 and an AUC of 0.852. The prediction model based on a combination of clinical and ultrasound indicators provided a

XSL•FO
RenderX

respectable performance (accuracy=0.7381, AUC=0.746). The most accurate and effective model was the multimodal co-learning approach. This model achieved an accuracy of 0.8331 and an AUC of 0.894, reflecting a significant improvement over the 3 single-modality approaches (Figure 3).

Moreover, in the precision-recall curve, the average precision score was calculated to summarize a precision-recall curve as the weighted mean of precisions achieved at each threshold. The multimodal co-learning model ran the best compared to the other 3 single modalities. The proposed co-learning model achieved the highest score (average precision=0.91) among the 4 experiments (Figure 4).

At last, the calibration curve was presented as a measurement of the confidence level of a proposed model and provided support for probability prediction (Figure 5). The proposed co-learning model was demonstrated to be well calibrated as the pink curve was closest to the perfectly calibrated line (the diagonal line at a 45-degree angle), showing that the predicted probabilities were highly consistent with the true outcomes.

Table . Results of 5-fold cross-validation.

Data	Colearning (AUC)[a]	ViViT[b] (AUC)
1st-fold (n=18)	0.857	0.889
2nd-fold (n=18)	0.843	0.778
3rd-fold (n=18)	0.892	0.861
4th-fold (n=18)	0.924	0.843
5th-fold (n=18)	0.956	0.889

[a]AUC: area under curve.

[b]ViViT: Video Vision Transformer.

XSL•FO
RenderX

Figure 3. ROC curve of weaning failure prediction in 4 experiments. Pink curve: the multimodal co-learning (accuracy=0.8331, AUC=0.894), green curve: diaphragm ultrasound via ViViT (accuracy=0.8095, AUC=0.852), yellow curve: clinical or ultrasound indicators (accuracy=0.7381, AUC=0.746), and blue curve: automatic mean excursion via speckle-tracking measurement (AUC=0.659). AUC: area under curve; ROC: receiver operating characteristic; ViViT: Video Vision Transformer.

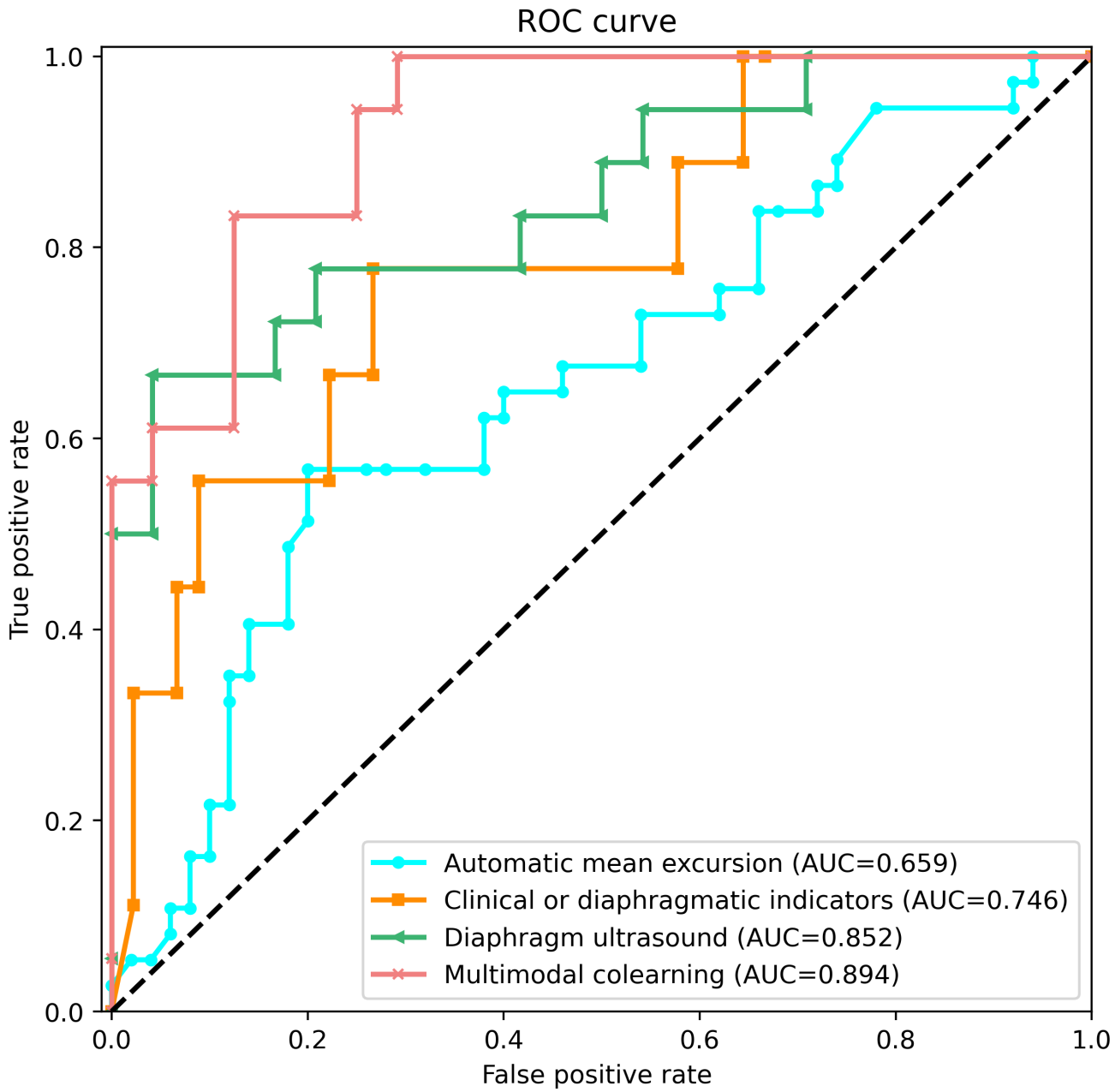

XSL•FO
RenderX

Figure 4. Precision-recall curve of weaning failure prediction in 4 experiments. Pink curve: the multimodal co-learning (AP=0.91), green curve: diaphragm ultrasound via ViViT (AP=0.86), yellow curve: clinical or ultrasound indicators (AP=0.47), and blue curve: automatic mean excursion via speckle-tracking measurement (AP=0.58). AP: average precision; ViViT: Video Vision Transformer.

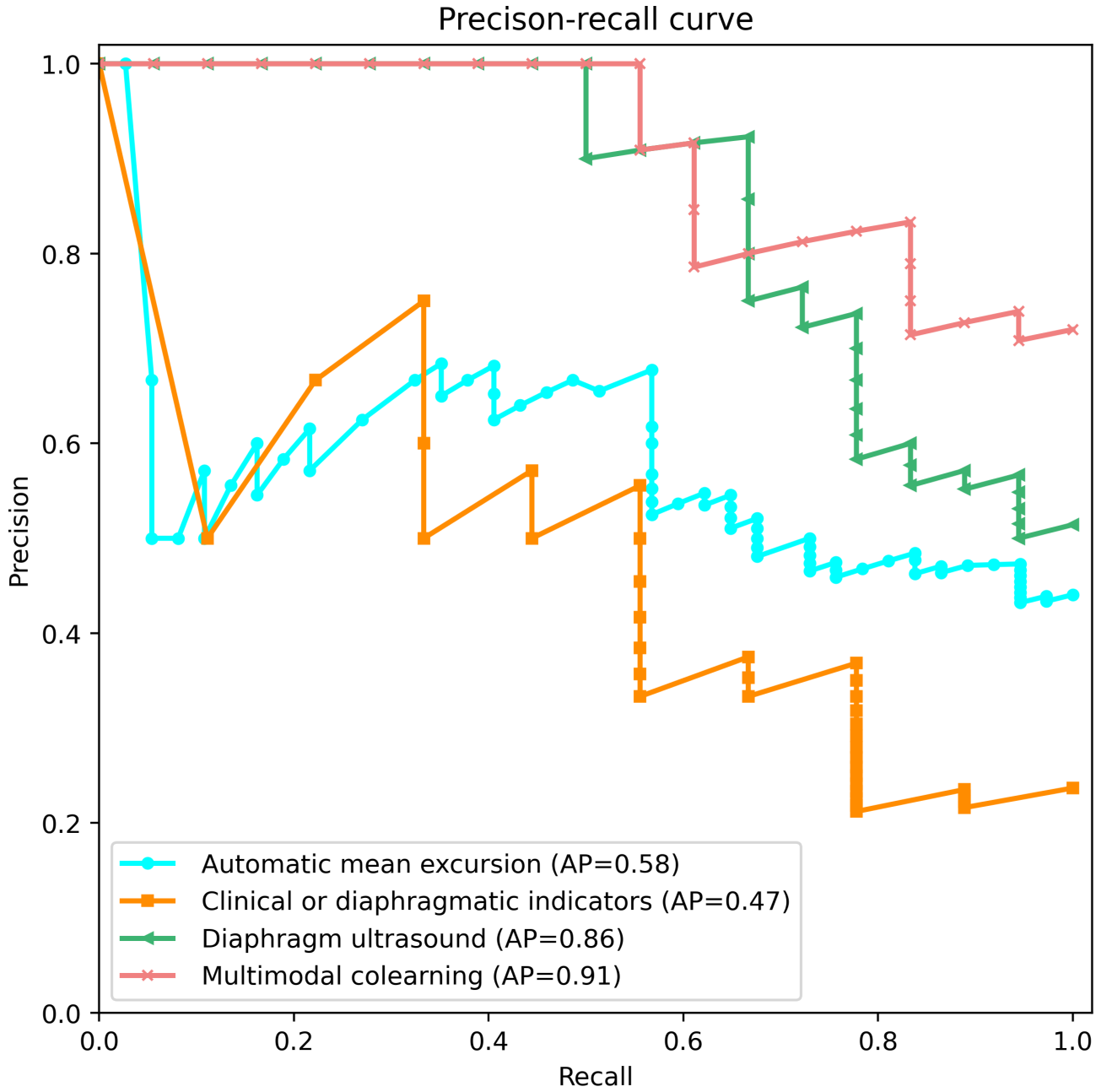

XSL•FO
RenderX

Figure 5. Calibration curve of weaning failure prediction in 4 experiments. Pink curve: the multimodal co-learning, green curve: diaphragm ultrasound via ViViT, yellow curve: clinical or ultrasound indicators, and blue curve: automatic mean excursion via speckle-tracking measurement. ViViT: Video Vision Transformer.

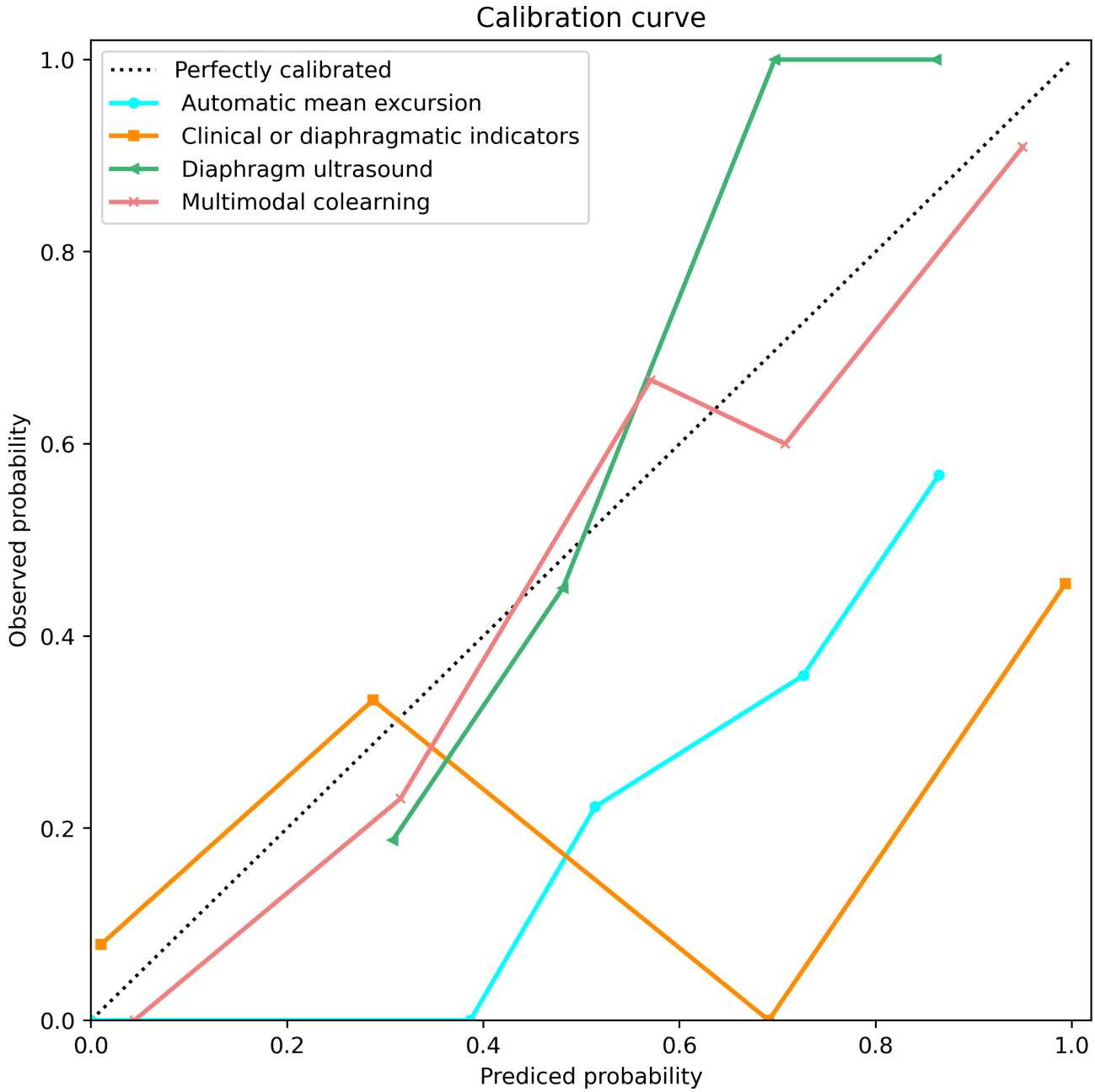

The experiments showed that the proposed co-learning model fully exploited the single-modal features and multimodal complementary features and achieved feature interaction between the modalities. Using both ultrasound and clinical data for colearning improved the accuracy of the weaning outcome prediction.

The SHAP values for the features were shown in Figure 6. The SHAP plot ranks features by their importance, with the highest importance at the top. The vertical axis ranked the features by the sum of the SHAP values across all samples, while the horizontal axis represented the SHAP value for each sample. Each dot corresponded to an individual sample, with the sample size indicated by vertical stacking of dots. On one hand, the color of each dot reflected the feature's value, with redder colors corresponding to higher values and bluer colors to lower values. For example, tidal volume had the highest overall SHAP value. High tidal volumes (in red) clustered on the negative side of the SHAP axis (located to the left of the SHAP=0 line), suggesting that in this cohort, it negatively impacted weaning success. Furthermore, hemoglobin level was identified as a significant predictor, with higher hemoglobin values being associated with a greater likelihood of successful weaning. Albumin also emerged as an important factor in predicting weaning outcomes. The red dots were primarily concentrated on the right side of the SHAP=0 line, indicating that higher albumin levels were associated with successful weaning. However, lower albumin levels were linked to a higher probability of weaning failure, as evidenced by the blue dots located on the leftmost side of the albumin row. On the other hand, the importance of each feature was determined by the

XSL•FO
RenderX

sum of its SHAP values, and the features with the highest sum of SHAP values were ranked at the top. Tidal volume, with the largest total SHAP value, emerged as the most influential feature for accurately predicting weaning outcomes. It should be noted that in this study, we incorporated SHAP to understand the importance of different features and enhance the interpretability, rather than to investigate causal relationships in clinical scenarios. SHAP plots generally do not directly show statistical significance (such as *P* values), which needs to be verified through statistical tests.

XSL•FO
RenderX

Figure 6. The SHAP values spot the impact on weaning outcome prediction. (1) Speckle-tracking automatic measurement of the diaphragm excursion, (2) N-terminal-pro-B-type natriuretic peptide, (3) SOFA score, and (4) speckle-tracking automatic measurement of the diaphragm excursion velocity. APACHE II: Acute Physiology and Chronic Health Evaluation II; PEEP: positive end-expiratory pressure; NT-pro BNP: N-terminal pro-brain natriuretic peptide; SHAP: Shapley Additive Explanation; SOFA: Sequential Organ Failure Assessment.

XSL•FO
RenderX

Discussion

The Clinical Challenge of Weaning and the Role of Diaphragm Dysfunction

About 26% - 39% of patients have trouble weaning [22,23] after resolution of the cause of respiratory failure and passing an SBT. Diaphragm dysfunction, resulting from weakness, atrophy, neurological damage, or mechanical disruption, is linked with prolonged MV duration, higher risk of weaning failure, longer length of ICU stay, and poor clinical outcomes [24]. Unsurprisingly, there is growing interest in investigating noninvasive bedside tools to evaluate the diaphragm and detect diaphragm dysfunction accurately. This is an important consideration when selecting which multiple parameters may be clinically useful to assess outcomes and for the prognostication of patients who may benefit from targeted interventions. Our serial research indicated that the incidence of diaphragm dysfunction in severe patients was as high as 75%, and diaphragm dysfunction was an important cause of weaning failure [25]. Over the past 10 years, diaphragm ultrasound has gradually become an extensively used tool for assessing diaphragm function [26]. As demonstrated in our previous studies, the ability to detect diaphragm dysfunction might become an important part of predicting weaning failure or prolonged weaning and holds the potential for determining weaning techniques. However, different studies have different predictors in addition to diaphragmatic indicators. Peer colleagues have also pointed out that the optimal prediction model has not been established, which is worth further exploration [27].

Existing AI-Based Weaning Prediction Studies

This report discusses the potential of POCUS assessment of the diaphragm using automatic speckle-tracking measurement plus other clinical characteristics in predicting weaning outcomes. Integrated with diaphragmatic indicators and 27 clinical indexes, the multimodal model was constructed to achieve accurate and intelligent prediction of weaning outcomes (accuracy=0.8331, AUC=0.894). The results reflected a significant improvement over the other 3 single-modality approaches. The high AUC value indicated that the multimodal model had a strong ability to correctly classify the outcomes, suggesting that the fusion of diverse data types (clinical or ultrasound video) generated from the critical care scenarios allowed the model to capture a more comprehensive and complex set of patterns for better predictions. The use of AI algorithms or ML to build predictive models for the successful weaning process was promising. Derivation of different clinical phenotypes of ventilated patients from the MIMIC-IV cohort was associated with extubation success. A study conducted at Chi Mei Medical Center selected 26 feature variables to build the predictive models with AI/ML algorithms, with a high AUC ranging from 0.792 to 0.868 [28]. Cheng Hsin General Hospital collected 28 variables and created a simple AI model for predicting weaning time [29]. A retrospective cohort study in Central Taiwan showed that explainable ML could predict successful weaning among patients requiring prolonged MV (AUC=0.908) [30]. Moreover, the 2-stage AI prediction models effectively and precisely predicted the optimal timing to wean intubated patients, with the AUCs of optimal models ranging from 0.889 to 0.944 [13].

Future Optimization 1: AI Integrates Multisystem POCUS Assessment

Of note, the result of this study was compromised due to a small sample. The AI model required a prospective cohort of a larger sample to validate its efficacy. Multiple risk factors were associated with weaning failure, such as cardiopulmonary dysfunction and older adults with severe comorbidities. It was urgent to optimize the decision-making process of weaning based on merely curtailing a few indicators. POCUS provides the dynamic assessment of multiorgan function during the weaning process. Even though high operator dependency limited the implementation and promotion of multisystem ultrasound monitoring in invasive MV patients, AI-integrated multisystem ultrasonic parameters rapidly and accurately assess, including for heart, lung, pleural cavity, diaphragm, and volume assessment.

Future Optimization 2: Diaphragm-Protective Ventilation Strategy

For the concern raised by Sabourin et al [31], suggesting that diaphragm ultrasonography should be done in patients with spontaneous breathing without pressure support, we hope to refine the objective parameters such as laboratory indexes and MV settings to improve the prediction in a clinical scenario. We hypothesize that, with the assistance of AI, the multimodal model might potentially distinguish between MV support and the patient's spontaneous breathing efforts, resembling clinical patterns seen in pressure support ventilation or synchronized intermittent mandatory ventilation mode. In our previous study, the excursion did not fully reflect whether the diaphragm movement was active or passive. Moreover, the value of diaphragmatic indicators was not only appreciated merely through excursion, but also in combination with the automatic measurement of diaphragm thickness and DTF. The abovementioned integration would become a strategy for diaphragm-protective ventilation.

Limitations

First, an automatic measurement for DTF, a reliable indicator for the weaning outcome and supposedly less impacted by pressure support ventilation [31], was under development. Moreover, the success of the proposed model relied on the quality of diaphragm ultrasound videos. In clinical practice, variations in equipment quality and ICU doctors' experience in performing POCUS could affect the reliability of these measurements. Future algorithms, integrating DTF and AI, would be explored to enhance accuracy and improve inter- and intraoperator reproducibility. Second, while the co-learning model achieved high predictive performance, the underlying decision-making process of the model was not fully interpretable. The translation of the findings of this study into clinical strategies still requires further refinement of the algorithm. Third, the relatively small sample size was not formally calculated, which limited the generalizability of the results to larger populations. A larger, multicenter study was needed to confirm the robustness and applicability of our

XSL•FO
RenderX

findings across diverse patient cohorts. Recent literature, including 2 reviews published in 2022 [32] and 2024 [33], lacked evidence that combined ultrasound videos with AI for predicting weaning outcomes. As such, our proposed model from 2022 [20], integrating diaphragm ultrasound videos and AI techniques, represented its novelty in this specific context. In conclusion, the weaning outcome was closely related to multiple dynamic factors and the function of the diaphragm and even various organs. The integration of multimodal AI, using automatic measurement from POCUS and automated collection of objective clinical and laboratory indicators, greatly enhanced the prediction of weaning outcomes. This approach held promise for broader clinical application, a personalized tool that supported decision-making and was conducive to clinical promotion.

Acknowledgments

This research was supported by the self-funded project within the Scientific Research Institute of Guangdong Provincial People's Hospital（No.8237080090, No.8247080256), and the Guangdong Basic and Applied Basic Research Foundation (No. 2024A1515220052). To reflect appropriate contributors to the study, we thank Mingliu Zhu, PhD; Jiaming Jiao, PhD; and Shuo Liu, PhD, who are the system research engineers from the ultrasound imaging system development department of Shenzhen Mindray Bio-Medical Electronics Co, Ltd, for technical support.

Data Availability

The data generated or analyzed during this study are not publicly available to protect the privacy of participants, but are available from the corresponding author upon reasonable request.

Authors' Contributions

FS investigated and wrote the original draft. H Liu worked on the formal analysis. HM reviewed and edited the writing. XC curated the data. SW handled the project administration. TQ collected the resources. H Liang supervised this study. DH conceptualized and supervised. All authors read and approved the final paper.

Authors DH (hdzdoctor@139.com) and H Liang (lianghuiying@hotmail.com) are co-corresponding authors for this article

Conflicts of Interest

None declared.

Multimedia Appendix 1
Table of clinical or diaphragm ultrasound indicators used in the model.
[DOCX File, 27 KB - formative_v9i1e72482_app1.docx]

References

1. McConville JF, Kress JP. Weaning patients from the ventilator. N Engl J Med 2012 Dec 6;367(23):2233-2239. [doi: 10.1056/NEJMra1203367] [Medline: 23215559]
2. Pham T, Heunks L, Bellani G, et al. Weaning from mechanical ventilation in intensive care units across 50 countries (WEAN SAFE): a multicentre, prospective, observational cohort study. Lancet Respir Med 2023 May;11(5):465-476. [doi: 10.1016/S2213-2600(22)00449-0] [Medline: 36693401]
3. Dres M, Dubé BP, Mayaux J, et al. Coexistence and impact of limb muscle and diaphragm weakness at time of liberation from mechanical ventilation in medical intensive care unit patients. Am J Respir Crit Care Med 2017 Jan 1;195(1):57-66. [doi: 10.1164/rccm.201602-0367OC] [Medline: 27310484]
4. Goligher EC, Dres M, Fan E, et al. Mechanical ventilation–induced diaphragm atrophy strongly impacts clinical outcomes. Am J Respir Crit Care Med 2018 Jan 15;197(2):204-213. [doi: 10.1164/rccm.201703-0536OC]
5. Lee J, Geyer B, Naraghi L, et al. Advanced imaging use in intensive care units has decreased, resulting in lower charges without negative effects on patient outcomes. J Crit Care 2015 Jun;30(3):460-464. [doi: 10.1016/j.jcrc.2014.12.012] [Medline: 25596998]
6. Truong D, Abo S, Whish-Wilson GA, et al. Methodological and clinimetric evaluation of inspiratory respiratory muscle ultrasound in the critical care setting: a systematic review and meta-analysis. Crit Care Med 2023 Feb 1;51(2):e24-e36. [doi: 10.1097/CCM.0000000000005739] [Medline: 36661463]
7. Tashiro N, Hasegawa T, Nishiwaki H, et al. Clinical utility of diaphragmatic ultrasonography for mechanical ventilator weaning in adults: a study protocol for systematic review and meta-analysis. Health Sci Rep 2023 Jul;6(7):e1378. [doi: 10.1002/hsr2.1378] [Medline: 37455705]
8. Llamas-Álvarez AM, Tenza-Lozano EM, Latour-Pérez J. Diaphragm and lung ultrasound to predict weaning outcome: systematic review and meta-analysis. Chest 2017 Dec;152(6):1140-1150. [doi: 10.1016/j.chest.2017.08.028] [Medline: 28864053]

XSL•FO
RenderX

9. Mahmoodpoor A, Fouladi S, Ramouz A, Shadvar K, Ostadi Z, Soleimanpour H. Diaphragm ultrasound to predict weaning outcome: systematic review and meta-analysis. Anaesthesiol Intensive Ther 2022;54(2):164-174. [doi: 10.5114/ait.2022.117273] [Medline: 35792111]
10. Le Neindre A, Philippart F, Luperto M, et al. Diagnostic accuracy of diaphragm ultrasound to predict weaning outcome: a systematic review and meta-analysis. Int J Nurs Stud 2021 May;117:103890. [doi: 10.1016/j.ijnurstu.2021.103890] [Medline: 33588324]
11. Huang D, Song F, Luo B, et al. Using automatic speckle tracking imaging to measure diaphragm excursion and predict the outcome of mechanical ventilation weaning. Crit Care 2023 Jan 14;27(1):18. [doi: 10.1186/s13054-022-04288-3] [Medline: 36639710]
12. Jia Y, Kaul C, Lawton T, Murray-Smith R, Habli I. Prediction of weaning from mechanical ventilation using convolutional neural networks. Artif Intell Med 2021 Jul;117:102087. [doi: 10.1016/j.artmed.2021.102087] [Medline: 34127233]
13. Liu CF, Hung CM, Ko SC, et al. An artificial intelligence system to predict the optimal timing for mechanical ventilation weaning for intensive care unit patients: a two-stage prediction approach. Front Med (Lausanne) 2022;9:935366. [doi: 10.3389/fmed.2022.935366] [Medline: 36465940]
14. Chen X, Li J, Liu G, et al. Identification of distinct clinical phenotypes of heterogeneous mechanically ventilated ICU patients using cluster analysis. J Clin Med 2023 Feb 14;12(4):1499. [doi: 10.3390/jcm12041499] [Medline: 36836034]
15. Li W, Zhang Y, Wang Z, et al. The risk factors of reintubation in intensive care unit patients on mechanical ventilation: a systematic review and meta-analysis. Intensive Crit Care Nurs 2023 Feb;74:103340. [doi: 10.1016/j.iccn.2022.103340] [Medline: 36369190]
16. Premkumar CR, Parthasarathy R, Kumar S, Vijaykumar S. Postoperative respiratory depression caused by iatrogenic hypermagnesaemia. Anaesth Rep 2022;10(1):e12154. [doi: 10.1002/anr3.12154] [Medline: 35360363]
17. Talakoub R, Bahrami M, Honarmand A, Abbasi S, Gerami H. The predicting ability of serum phosphorus to assess the duration of mechanical ventilation in critically ill patients. Adv Biomed Res 2017;6:51. [doi: 10.4103/2277-9175.205192] [Medline: 28620595]
18. Zhao Y, Li Z, Shi Y, et al. Effect of hypophosphatemia on the withdrawal of mechanical ventilation in patients with acute exacerbations of chronic obstructive pulmonary disease. Biomed Rep 2016 Apr;4(4):413-416. [doi: 10.3892/br.2016.605] [Medline: 27073623]
19. Pu H, Doig GS, Lv Y, et al. Modifiable risk factors for ventilator associated diaphragmatic dysfunction: a multicenter observational study. BMC Pulm Med 2023 Sep 13;23(1):343. [doi: 10.1186/s12890-023-02633-y] [Medline: 37700263]
20. Arnab A, Dehghani M, Heigold G, Sun C, Lucic M, Schmid C. ViViT: a Video Vision Transformer. . Preprint posted online on Mar 29, 2021 URL: https://ui.adsabs.harvard.edu/abs/2021arXiv210315691A [accessed 2025-07-08] [doi: 10.1109/ICCV48922.2021.00676]
21. Li J, Liu S, Hu Y, Zhu L, Mao Y, Liu J. Predicting mortality in intensive care unit patients with heart failure using an interpretable machine learning model: retrospective cohort study. J Med Internet Res 2022 Aug 9;24(8):e38082. [doi: 10.2196/38082] [Medline: 35943767]
22. Funk GC, Anders S, Breyer MK, et al. Incidence and outcome of weaning from mechanical ventilation according to new categories. Eur Respir J 2010 Jan;35(1):88-94. [doi: 10.1183/09031936.00056909] [Medline: 19541716]
23. Peñuelas O, Frutos-Vivar F, Fernández C, et al. Characteristics and outcomes of ventilated patients according to time to liberation from mechanical ventilation. Am J Respir Crit Care Med 2011 Aug 15;184(4):430-437. [doi: 10.1164/rccm.201011-1887OC] [Medline: 21616997]
24. Purro A, Appendini L, De Gaetano A, Gudjonsdottir M, Donner CF, Rossi A. Physiologic determinants of ventilator dependence in long-term mechanically ventilated patients. Am J Respir Crit Care Med 2000 Apr;161(4 Pt 1):1115-1123. [doi: 10.1164/ajrccm.161.4.9812160] [Medline: 10764299]
25. Huang D, Ma H, Zhong W, et al. Using M-mode ultrasonography to assess diaphragm dysfunction and predict the success of mechanical ventilation weaning in elderly patients. J Thorac Dis 2017 Sep;9(9):3177-3186. [doi: 10.21037/jtd.2017.08.16] [Medline: 29221294]
26. Haaksma ME, Smit JM, Boussuges A, et al. Expert Consensus On Diaphragm Ultrasonography in the Critically Ill (EXODUS): a Delphi consensus statement on the measurement of diaphragm ultrasound-derived parameters in a critical care setting. Crit Care 2022 Apr 8;26(1):99. [doi: 10.1186/s13054-022-03975-5] [Medline: 35395861]
27. Wang L, Muhetaer Y, Zhu L, Wang Q, Wu W. Is it reasonable to predict weaning by measuring diaphragm activity under ultrasound especially its reduction of excursion? Crit Care 2023;27(1):309. [doi: 10.1186/s13054-023-04585-5]
28. Liao KM, Ko SC, Liu CF, et al. Development of an interactive AI system for the optimal timing prediction of successful weaning from mechanical ventilation for patients in respiratory care centers. Diagnostics (Basel) 2022 Apr 13;12(4):975. [doi: 10.3390/diagnostics12040975] [Medline: 35454023]
29. Chen WT, Huang HL, Ko PS, Su W, Kao CC, Su SL. A simple algorithm using ventilator parameters to predict successfully rapid weaning program in cardiac intensive care unit patients. J Pers Med 2022 Mar 21;12(3):501. [doi: 10.3390/jpm12030501] [Medline: 35330500]

XSL•FO
RenderX

30. Lin MY, Li CC, Lin PH, et al. Explainable machine learning to predict successful weaning among patients requiring prolonged mechanical ventilation: a retrospective cohort study in central Taiwan. Front Med (Lausanne) 2021;8:663739. [doi: 10.3389/fmed.2021.663739] [Medline: 33968967]
31. Sabourin E, Carpentier C, Lai C, Monnet X, Pham T. “Under pressure”: should we use diaphragm excursion to predict weaning success in patients receiving pressure support ventilation? Crit Care 2023 Jun 15;27(1):238. [doi: 10.1186/s13054-023-04504-8] [Medline: 37322500]
32. Gallifant J, Zhang J, Del Pilar Arias Lopez M, et al. Artificial intelligence for mechanical ventilation: systematic review of design, reporting standards, and bias. Br J Anaesth 2022 Feb;128(2):343-351. [doi: 10.1016/j.bja.2021.09.025] [Medline: 34772497]
33. Stivi T, Padawer D, Dirini N, Nachshon A, Batzofin BM, Ledot S. Using artificial intelligence to predict mechanical ventilation weaning success in patients with respiratory failure, including those with acute respiratory distress syndrome. J Clin Med 2024 Mar 5;13(5):1505. [doi: 10.3390/jcm13051505] [Medline: 38592696]

Abbreviations

AI: artificial intelligence
AUC: area under the curve
DTF: diaphragm thickening fraction
eICU: electronic intensive care unit
ICU: intensive care unit
MIMIC-IV: Medical Information Mart for Intensive Care
ML: machine learning
MV: mechanical ventilation
POCUS: point-of-care ultrasonography
SBT: spontaneous breathing trial
SHAP: Shapley Additive Explanation
SOFA: Sequential Organ Failure Assessment
ViViT: Video Vision Transformer

Edited by J Sarvestan; submitted 11.02.25; peer-reviewed by AN Elshaer, G Tu; revised version received 01.06.25; accepted 04.06.25; published 08.09.25.

Please cite as:
Song F, Liu H, Ma H, Chen X, Wang S, Qin T, Liang H, Huang D
AI Model Based on Diaphragm Ultrasound to Improve the Predictive Performance of Invasive Mechanical Ventilation Weaning: Prospective Cohort Study
JMIR Form Res 2025;9:e72482
URL: https://formative.jmir.org/2025/1/e72482
doi:10.2196/72482

© Feier Song, Huazhang Liu, Huan Ma, Xuanhui Chen, Shouhong Wang, Tiehe Qin, Huiying Liang, Daozheng Huang. Originally published in JMIR Formative Research (https://formative.jmir.org), 8.9.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Recognizing Skateboard and Kickboard Commuting Behaviors Using Activity Trackers: Feasibility Study Using Machine Learning Approaches

Nathanael Aubert-Kato[1,2], PhD; Hitomi Hatori[3], MSc; Arisa Orihara[3], BSc; Takashi Nakagata[4,5], PhD; Yuji Ohta[3], DEng; Julien Tripette[2,3], PhD

[1]Department of Computer Science, Ochanomizu University, Tokyo, Japan

[2]Center for Interdisciplinary AI and Data Science, Ochanomizu University, 2-1-1 Otsuka, Bunkyo District, Tokyo, Japan

[3]Department of Human-Environmental Sciences, Ochanomizu University, Tokyo, Japan

[4]Center for Physical Activity Research, National Institutes of Biomedical Innovation, Health and Nutrition, Settsu, Japan

[5]Laboratory of Gut Microbiome for Health, Microbial Research Center for Health and Medicine, National Institutes of Biomedical Innovation, Health and Nutrition, Settsu, Japan

Corresponding Author:
Julien Tripette, PhD
Center for Interdisciplinary AI and Data Science, Ochanomizu University, 2-1-1 Otsuka, Bunkyo District, Tokyo, Japan

Abstract

Background: Active commuting, such as skateboarding and kickboarding, is gaining popularity as an alternative to traditional modes of transportation such as walking and cycling. However, current activity trackers and smartphones, which rely on accelerometer data, are primarily designed to recognize symmetrical locomotive activities (eg, walking and running) and may struggle to accurately identify the unique push-push-glide motion patterns of skateboarding and kickboarding.

Objective: The primary objective of this study was to evaluate the feasibility of classifying skateboard and kickboard commuting behaviors using data from wearable sensors and smartphones. A secondary objective was to identify the most important sensor-derived features for accurate activity recognition.

Methods: Ten participants (4 women and 6 men; aged 12 - 55 y) performed 9 activities, including skateboarding, kickboarding, walking, running, bicycling, ascending stairs, descending stairs, sitting, and standing. Data were collected using wearable sensors (accelerometer, gyroscope, and barometer) placed on the wrist and the hip, as well as in the pocket to replicate the sensing characteristics of commercial activity trackers and smartphones. The signal processing approach included the extraction of 211 features from 10- and 20-second sliding windows. Random forest classifiers were trained to perform multiclass and binary classifications, including distinguishing skateboarding and kickboarding from other activities.

Results: Wrist-worn sensor configurations achieved the highest balanced accuracies for multiclass classification (range 84% - 88%). Skateboarding and kickboarding were identified with high sensitivity, ranging from 93% to 99% and 97% to 99%, respectively. Hip and pocket sensor configurations showed lower performance, particularly in distinguishing skateboarding (range 49% - 58% sensitivity) from kickboarding (78% sensitivity). Binary classification models grouping skateboarding and kickboarding into a push-push-glide superclass achieved high accuracies (range 91% - 95%). Key features for classification included low- and high-frequency accelerometer signals, as well as roll-pitch-yaw angles.

Conclusions: This study demonstrates the feasibility of recognizing skateboard and kickboard commuting behaviors using wearable sensors, particularly wrist-worn devices. While hip and pocket sensors showed limitations in differentiating these activities, the broader push-push-glide classification achieved acceptable accuracy, suggesting its potential for integration into activity tracker software. Future research should explore sensor fusion approaches to further enhance recognition performance and address the question of energy expenditure estimation.

(JMIR Form Res 2025;9:e71969) doi:10.2196/71969

KEYWORDS

physical activity; activity recognition; wearable sensor; skateboard; machine learning

XSL•FO
RenderX

Introduction

Background

Active commuting is increasingly recognized as an effective way to encourage healthy habits among individuals with limited opportunities for regular physical activity (PA) [1,2]. According to a systematic review and meta-analysis by Dinu et al [3], active commuters experience an 8% reduction in mortality risk and a 9% to 30% lower risk of various noncommunicable diseases.

While walking and cycling remain the most common forms of active commuting [4], alternative nonmotorized transportation modes such as skateboarding and kickboarding are gaining popularity [5]. Often described as "fun, faster than walking, and more convenient than a bike" [6], skateboards and similar devices are involved in >30,000 multimodal trips per day in Los Angeles, California [5]. In Portland, Oregon, at least 1 skateboarder passes through 79% of the intersections of the city. Skateboarding is particularly popular among young adults, with mode shares reaching up to 7% on university campuses in California. Outside the United States, Ward et al [7] reported that 7% of surveyed teenagers in New Zealand used skateboarding as a mode of transport. Skateboarding, specifically in a cruising context, can reach intensities as high as 8.5 metabolic equivalents of task (METs), as compared to 3.5 METs for walking, 6.8 - 9.3 METs for bicycling, and 6.8 METs for e-bicycling [8,9]. The similar motion pattern—characterized by several successive kicking or pushing actions followed by a gliding phase, hereafter referred to as the "push-push-glide" pattern—suggests that the intensity of kickboarding may be comparable to that of skateboarding. In a treadmill-based laboratory study, Kijima et al [10] found that kickboarding elicited energy expenditures ranging from 3.9 to 5.0 METs. Collectively, these findings indicate that kickboard and skateboard commuting may be viable strategies for individuals seeking to increase their daily PA.

Activity trackers are wearable devices equipped with sensors that evaluate PA and provide users with quantitative feedback, such as daily energy expenditure and step counts. These devices are commonly worn on the wrist and the hip or placed in the pocket. Nowadays, many smartphone apps leverage the sensing capabilities of smartphones to perform similar PA assessment. While modern activity trackers and smartphones may feature a variety of advanced sensing technologies, their core evaluation method typically relies on processing accelerometer data to estimate PA intensity [11].

Although algorithms developed by consumer-grade activity tracker manufacturers are often proprietary, several published studies using devices that provide access to raw data have aimed to develop accurate software tools for estimating energy expenditure, shedding light on the data processing flow of activity tracker algorithms [12-18]. They generally use sliding analytic windows, ranging in length from 1 to 60 seconds, and involve 4 successive operations. First, signals are smoothed to filter out noise-unrelated human movement, typically within the 0.25 - 10.00 Hz frequency range. Second, signal features are extracted in both the time and frequency domains [16]. Third, the extracted data are classified into categories based on their characteristics [19,20]. Finally, class-specific regressions are applied to derive accurate estimates of PA intensity, sometimes expressed as METs. These estimates are then integrated with time- and user-specific characteristics to predict energy expenditure.

For instance, 2 studies have documented the algorithmic framework used by OMRON (Omron Corporation). At the third processing stage, the OMRON algorithm categorizes accelerometer data into 3 activity types: locomotive, nonlocomotive, or mixed activities [14,20]. However, the locomotive activities considered in its development were limited to walking and running tasks, which involve regular, cyclic alternations of stance and swing phases between the left and right legs.

A Google Scholar search for studies focusing on locomotive behavior recognition using wearable accelerometer sensors highlights a gap in research on skateboard and kickboard activities. This suggests that locomotive activities, characterized by irregular or asymmetric motion patterns, are largely absent in existing algorithms, potentially leading to errors in energy expenditure estimates. Moreover, the unique push-push-glide motion patterns of skateboarding and kickboarding often span several seconds, highlighting the need for longer analytic windows to accurately recognize these behaviors.

Objectives

While inaccurate feedback from activity trackers or smartphones can discourage users from engaging in active lifestyles [21], this study aims to evaluate the ability of these devices to recognize skateboard and kickboard transportation activities, thereby promoting active commuting habits. The primary objective of this study is to build models that classify sensor data from devices akin to contemporary activity trackers—waist-worn research-grade sensors, wrist-worn consumer-grade devices, and smartphones—by accurately identifying the push-push-glide dynamics of skateboard and kickboard use. The secondary objective is to identify data features that contribute most to the accuracy of the classification models.

Methods

Overview of the Study Protocol

A heterogeneous, convenient sample of 10 participants was recruited from the campus of Ochanomizu University and local skate parks to capture a range of skateboarding cruising. Details of participant characteristics are presented in Table 1. Exclusion criteria included being aged <11 years, having any lower limb injury that impaired normal gait, or lacking prior experience with skateboard cruising. Participants followed a structured protocol that involved 9 sedentary and locomotive activities: sitting (while chatting or scrolling a smartphone), standing (while chatting or scrolling a smartphone), bicycling, walking, running, kickboarding, skateboarding, ascending stairs, and descending stairs. Each activity was designed to last at least 10 minutes. This duration was selected to provide enough data samples (approximately 30 nonoverlapping windows of 20 s),

XSL•FO
RenderX

while ensuring the entire protocol could be completed within half a day. Some activities required discontinuous data collection due to environment constraints. For instance, stair activities were conducted in an 8-story building, requiring multiple bouts of 2 - 3 minutes to achieve a total of 10 minutes. Participants were permitted to stop any activity before reaching the 10-minute mark upon request, to prevent excessive fatigue. All activities were conducted on the campus of Ochanomizu University in Tokyo and were performed in a randomized order to avoid systematic metabolic interferences. Inertial data were collected using wearable sensors, and participants also wore a chest-mounted action camera (Virb 360, Garmin) to record the entire experiment.

Table . Participant characteristics.

Participant ID	Sex	Age (years)	Skateboarding skill level[a]	Stance[b]	Weight (kg)
001	Woman	12	Expert	Regular	38
002	Woman	26	Beginner	Goofy	42
003	Man	40	Expert	Regular	83
004	Woman	27	Expert	Regular	45
005	Man	22	Beginner	Regular	54
006	Man	22	Expert	Goofy	65
007	Man	13	Expert	Regular	50
008	Man	25	Expert	Goofy	80
009	Man	22	Expert	Regular	65
010	Man	55	Beginner	Regular	65

[a]Participants categorized as “beginners” reported at least 1 prior experience with skateboarding but were not consistently able to perform the sharp turns required during the experiment.

[b]Participants with a regular stance naturally position their left foot at the front of the board and kick with their right foot during push-push-glide activities, such as skateboarding. Conversely, participants with a goofy stance naturally place their right foot at the front and kick with their left foot.

Ethical Considerations

The experimental protocol was approved by the Ochanomizu University research ethics committee (#2021 - 10). All participants gave their written consent. To ensure privacy, all data were anonymized before analysis. Some participants did not receive compensation, while others were compensated at an hourly rate ranging from ¥1050 (US $7.10) to ¥1250 (US $8.50), depending on their age and university degree status.

Data Collection and Preprocessing

Raw accelerometer, gyroscope, and barometer data were collected using 3 MetaMotionS wearable devices (Mbient Lab) worn on the left wrist and the right hip and placed in the right back pocket of the trousers (Figure 1A). These placements replicated the typical positions for activity trackers and smartphones. The accelerometer and gyroscope sensors were triaxial devices and configured with a sampling rate of 100 Hz, while the barometer sensor sampled at approximately 1 Hz.

The collected data streams were segmented based on the start and end times of each activity. To minimize boundaries effects, the first and last 5 seconds of each activity were removed. The preprocessed data were stored as independent CSV files, resulting in a final dataset comprising 196 files corresponding to the 100 activities performed by the 10 participants. Each file contained 7 independent time series: 6 corresponding to the x-, y-, and z-axis of the accelerometer and gyroscope sensors and 1 corresponding to the barometer sensor. Signal samples for the 9 activities are presented in Multimedia Appendix 1.

XSL•FO
RenderX

Figure 1. Overview of the materials and protocol. (A) Three MetaMotionS sensors were worn on the left wrist and the right hip and placed in the right back pocket of the trousers. The sensor in the back pocket was clipped onto a piece of cardboard and inserted in a small pouch matching the dimension of a smartphone. An action camera (Virb 360) was mounted on the chest. (B) The participants completed 9 different activities, including 2 sedentary behaviors (light blue); 5 typical locomotive behaviors with symmetrical motion patterns (navy blue); and 2 push-push-glide locomotive activities with asymmetrical motion patterns, that is, skateboard and kickboard cruising (light red). Each activity lasted approximately 10 minutes. Stair activities were performed in an 8-story building and completed in several bouts of approximately 2 minutes each, totaling 10 minutes.

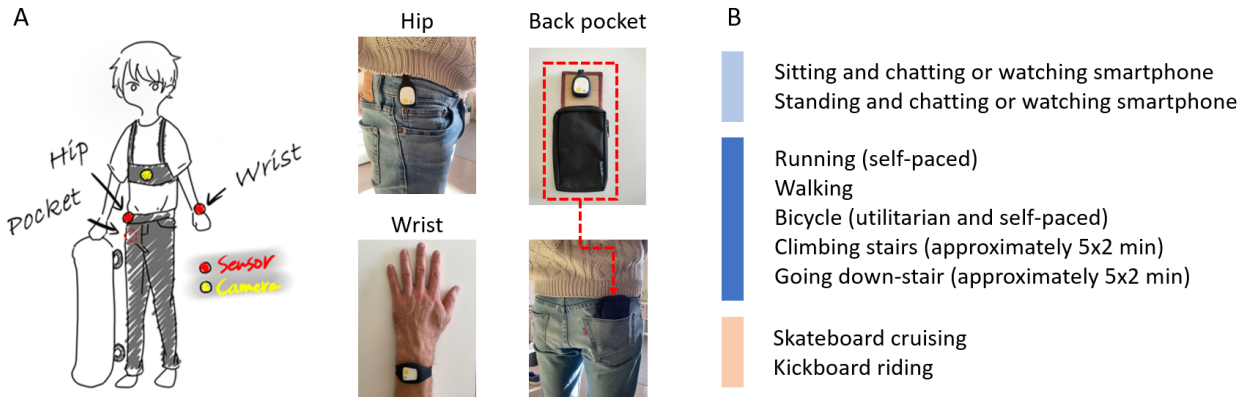

Feature Extraction

Each CSV file was analyzed using sliding analytic windows of 10 or 20 seconds. For each window, 211 data features were extracted from the accelerometer, gyroscope, and barometer data, as follows:

- Accelerometer features*:* 112 features were derived, primarily based on the approach by Willetts et al [16]. These included 29 parameters from the general statistical analysis in the time domain, 12 parameters from the roll-pitch-yaw analysis, and 71 parameters from the fast Fourier transform (FFT) analysis [22,23].
- Gyroscope features*:* 90 features were extracted. These included 30 parameters from general statistical analysis in the time domain and 60 parameters from the FFT analysis.
- Combined features*:* 8 features combining information from the accelerometer and gyroscope sensor were extracted to examine synchrony between the acceleration generated by the push motion and angular velocity during the push-to-glide phases of skateboarding.
- Barometer feature*:* 1 feature was extracted to examine the slope of movements.

Further details on the computation of these features can be found in the data and code repository (refer to “data_preparation.py” for the exact code) [24].

Configurations

The various combinations of hardware and software used for the analysis replicated the characteristics of various types of commercial devices, as follows:

- Hip sensor*:* analysis of accelerometer data only, reflecting the capabilities of waist-worn research-grade activity trackers [11].
- Pocket sensor: analysis of accelerometer, gyroscope, and barometer data, simulating the capabilities of modern smartphones.
- Wrist sensor (low end): analysis of accelerometer data only, replicating the capabilities of entry-level wrist-worn consumer-grade activity trackers.
- Wrist sensor (high end)*:* analysis of accelerometer, gyroscope, and barometer data simulating the capabilities of high-end, wrist-worn, consumer-grade activity trackers.

Activity List and Recognition Models

Activity recognition was performed using random forest classifiers across four analyses:

1. Multiclass classification*:* differentiating all 9 activities (shown in Figure 1B).
2. Kickboard versus *rest* and skateboard versus *rest* binary classifications: extracting the kickboard and skateboard estimators trained together using the 1 versus *rest* approach [25] along the full list of activities and applying them to the test samples.
3. Push-push-glide versus *rest* binary classification: grouping kickboard and skateboard activities into a push-push-glide superclass and distinguishing from other activities.
4. Kickboard versus skateboard binary classification: differentiating kickboard and skateboard activities.

Random forests models were configured with 100 trees. For training, data from 8 participants were used, while data from the remaining 2 were reserved for performance evaluation. Due to the data collection protocol, activities had slightly different durations, resulting in small discrepancies in the number of data points available. The training used class balancing to avoid related biases, that is, the weight of each data point during training was normalized by the number of data points available for the corresponding activity. To ensure robustness, the recognition performance was averaged across 100 independent runs for each 8:2 split. The overall performance for each activity list and each sensor configuration (eg, multiclass classification with the pocket sensor) was obtained by averaging the performances across all possible splits. The activity recognition analysis was conducted using *Scikit-learn* [26].

Feature Importance Analysis

Feature importance quantifies the contribution of each feature. The importance was extracted from the models using the built-in method provided by *Scikit-learn* [26]. This metric is calculated based on the decrease of Gini impurity, which evaluates how

XSL•FO
RenderX

effectively each feature separates the different classes [27]. The final importance value for a given feature was calculated by averaging the importance scores obtained across all training repetitions for a specific configuration.

Results

Participant Characteristics

As shown in Table 1, the group of participants included 4 women and 6 men, aged between 12 and 55 (mean 26, SD 12) years, with weights ranging from 38 to 83 (mean 59, SD 15) kg. Seven participants were comfortable riding a skateboard, while 3 were classified as beginners, as they were unable to perform sharp turns and occasionally needed to step off and manually adjust the direction of the board. All participants were comfortable using a kickboard.

Multiclass Classification Models

As shown in Table 2, the balanced accuracies for the multiclass classification models ranged from 84% (SD 3%; wrist sensor using accelerometer data with 10-s analytic windows) to 88% (SD 3%; wrist sensor using accelerometer, gyroscope, and barometer data with 10- or 20-s analytic windows). The size of the analytic window did not affect the overall performance of the models. The confusion matrices presented in Figure 2 highlight lower performance of models using hip sensor accelerometer data or the pocket sensor accelerometer, gyroscope, and barometer data for recognizing skateboarding and kickboarding. Skateboarding was correctly classified in only 49% (SD 23%; across train-test splits) of cases by the hip sensor models (Figure 2A) and in 58% (SD 32%; across train-test splits) of cases by the pocket sensor models (Figure 2B). Kickboarding was accurately classified in 78% (SD 20%) of cases by the hip sensor and the pocket sensor models. Most classification errors involving these 2 activities were due to mutual confusion. Figure 2 shows the results for the 20-second window analysis only. The 10-second window analysis resulted in similar scores, which are presented in Multimedia Appendix 1.

Table . Balance accuracies for each device, configuration, and classification model (10-s and 20-s window results).

Classification task		Hip sensor and features from accelerometer data[a] (%), mean (SD)	Pocket sensor and features from accelerometer, gyroscope, and barometer data[b] (%), mean (SD)	Wrist sensor and features from accelerometer data[c] (%), mean (SD)	Wrist sensor and features from accelerometer, gyroscope, and barometer data[d] (%), mean (SD)
Multiclass					
	10-s window	86 (5)	86 (5)	84 (3)	88 (3)
	20-s window	87 (5)	87 (5)	85 (4)	88 (3)
Kickboard versus *rest*					
	10-s window	70 (9)	71 (13)	89 (10)	91 (8)
	20-s window	71 (9)	71 (13)	87 (10)	90 (9)
Skateboard versus *rest*					
	10-s window	56 (7)	65 (12)	84 (9)	83 (11)
	20-s window	55 (7)	67 (13)	83 (10)	83 (11)
Push-push-glide versus *rest*					
	10-s window	92 (6)	95 (3)	91 (7)	91 (8)
	20-s window	93 (7)	95 (4)	91 (6)	92 (8)
Kickboard versus skateboard					
	10-s window	65 (12)	65 (18)	100 (1)	100 (0)
	20-s window	66 (12)	68 (15)	100 (0)	100 (0)

[a]The configuration of the hip sensor and accelerometer-derived features replicates that of waist-worn research-grade activity trackers.

[b]The configuration of the pocket sensor and accelerometer-, gyroscope-, and barometer-derived features replicates that of contemporary smartphone devices.

[c]The configuration of the wrist sensor and accelerometer-derived features replicates that of entry-level wrist-worn activity trackers.

[d]The configuration of the wrist sensor and accelerometer-, gyroscope-, and barometer-derived features replicates that of high-end, wrist-worn activity trackers.

XSL•FO
RenderX

Figure 2. Confusion matrices for each device and configuration and for the multiactivity classification models (20-s window results). For each activity, values refer to the normalized sensitivity. (**A**) Hip sensor and accelerometer-derived features; replicates the configuration of waist-worn research-grade activity trackers. (**B**) Pocket sensor and accelerometer-, gyroscope-, and barometer-derived features; replicates the configuration of contemporary smartphone devices. (**C**) Wrist sensor and accelerometer-derived features; replicates the configuration of entry-level wrist-worn activity trackers. (**D**) Wrist sensor and accelerometer-, gyroscope-, and barometer-derived features; replicates the configuration of high-end, wrist-worn activity trackers. "Downstairs" denotes descending stairs, "Kick" denotes kickboard, "Skate" denotes skateboard, and "Upstairs" denotes ascending stairs.

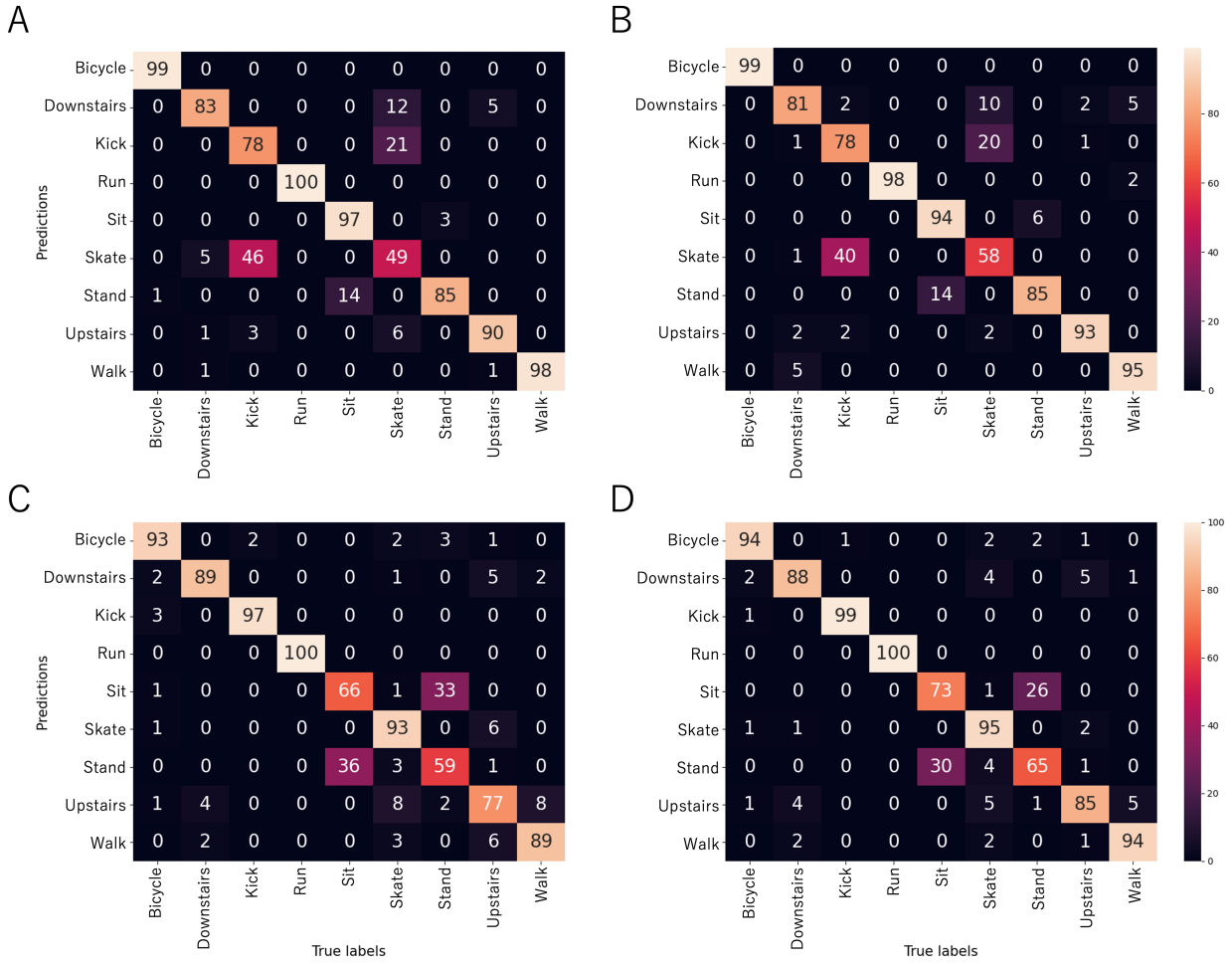

Binary Classification Models

Detailed results are shown in Figure 3. The balanced accuracies for the kickboard versus *rest* classification ranged from 70% (SD 9%) to 91% (SD 8%). Higher accuracies were observed for the wrist sensor models, regardless of whether only accelerometer data or a combination of accelerometer, gyroscope, and barometer data was used. The skateboard versus *rest* classification models demonstrated lower balanced accuracies, ranging from 55% (SD 7%) to 84% (SD 9%). Similar to kickboard versus *rest*, higher accuracies were observed for the wrist sensor models.

The push-push-glide versus *rest* classification showed consistently high balanced accuracies, ranging from 91% (SD 8%) to 95% (SD 4%). In contrast, the kickboard versus skateboard classification models achieved balanced accuracies ranging from 65% (SD 12%; pocket sensor using accelerometer, gyroscope, and barometer data with 10 s analytic windows) to 100% (SD 0%; wrist sensor using accelerometer, gyroscope, and barometer data and with 10-s analytic windows or the wrist sensor with any configuration using the 20-s analytic windows).

All balanced accuracy scores are presented in Table 2. The performances remained consistent across models, regardless of whether 10- and 20-second analytic windows were used.

XSL•FO
RenderX

Figure 3. Confusion matrices for each device, configuration, and binary classification model (20-s window results). For each activity, values refer to the normalized sensitivity. "Kick" denotes kickboard, "Pushpush" denotes push-push-glide activities, and "Skate" denotes skateboard.

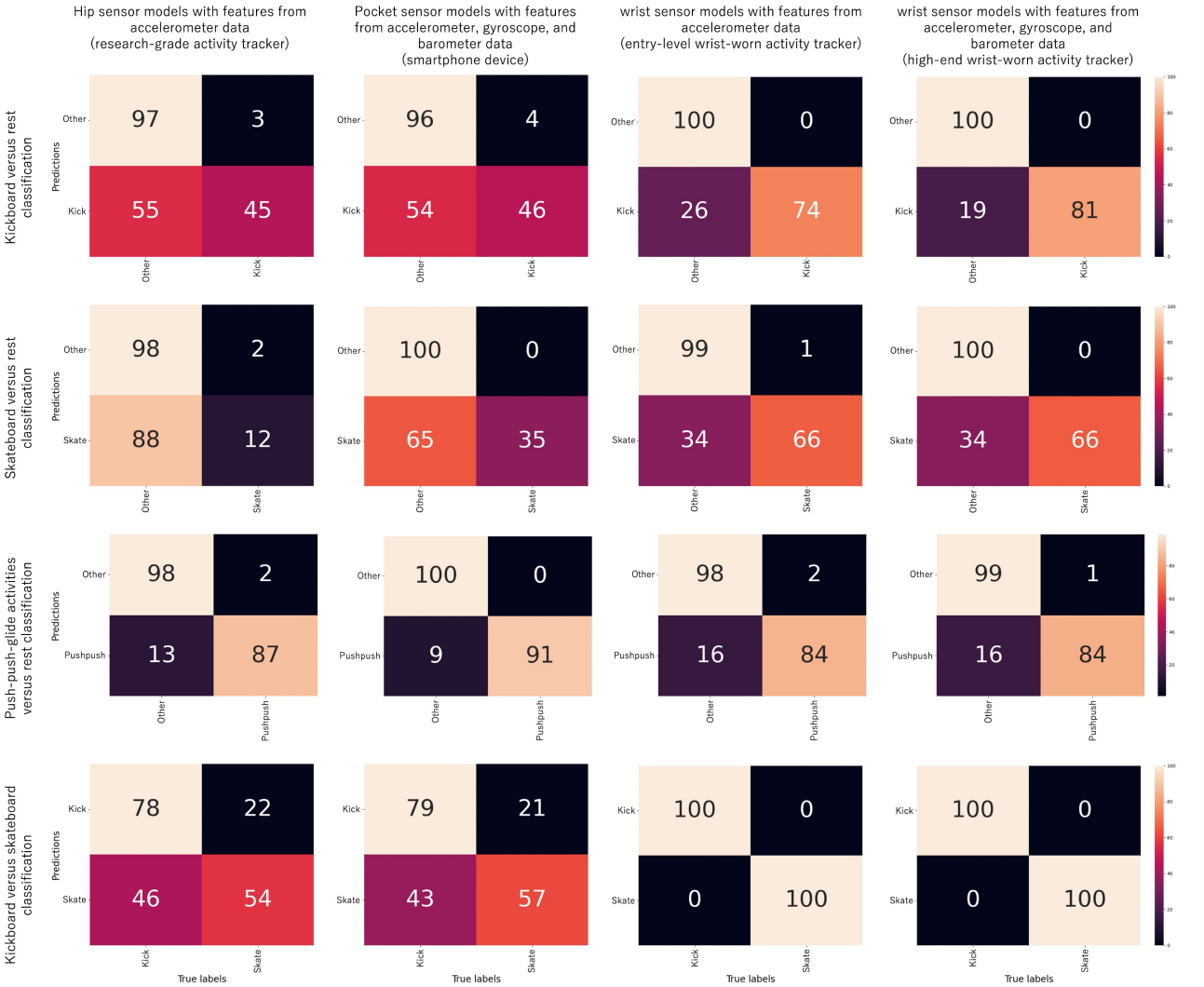

Feature Importance

A list of the 10 most important features used in the binary classifications is shown in Multimedia Appendix 2. A large number of the most common features are shared across configurations and classification targets (kickboard vs *rest*, skateboard vs *rest*, both vs *rest*, and kickboard vs skateboard). This is particularly the case for statistics features based on the norm of the acceleration vector, such as the skewness (9 uses), kurtosis (7 uses), or its FFT, in particular low frequencies (8 uses for 1 - 2 Hz and 7 uses for 0 - 1 Hz) and ratio of maximum power to total power (6 uses). In addition, the pitch and yaw of the acceleration vector relative to the gravity vector are among the most used features (7 and 6 uses, respectively). Beside the norm of the acceleration vector, the average of its y-axis and z-axis components is also strongly represented (6 uses each).

The remaining features tend to have a low use count but similar interpretations. For wrist sensor configurations, we have a large number of FFT features related to high frequencies (10 - 15 Hz and 20 combined uses).

Discussion

Main Findings

In recent years, kickboarding and skateboarding have become increasingly common modes of transportation [5-7]. However, there is a lack of algorithms capable of accurately recognizing these behaviors, which impacts the accuracy of daily PA estimates from activity trackers. Such inconsistencies may discourage users from engaging in active transportation behaviors [21].

In this study, accelerometer data—alone or combined with gyroscope and barometer data—collected from devices mimicking contemporary activity trackers or smartphone sensing capabilities were used to develop classification models capable of recognizing kickboarding and skateboarding commuting behaviors. A total of 112 accelerometer-derived features were extracted, as described in previous studies [22,23]. In addition, 99 features based on gyroscope and barometer data were extracted, some of which were specifically designed to capture the irregular and asymmetric motion patterns of kickboarding or skateboarding. These features were expected to capture the unique characteristics of the push-push-glide motion. Feature

XSL•FO
RenderX

extraction was conducted using 10- and 20-second sliding window analyses. Random forest algorithms using subject-wise training-test assignments were used to develop multiclass and binary classification activity recognition models across four sensor configurations: (1) hip sensor with accelerometer data only; (2) pocket sensor with the accelerometer, gyroscope, and barometer data; (3) wrist sensor with accelerometer data only; and (4) wrist sensor with the accelerometer, gyroscope, and barometer data. The wrist sensor multiclass classification models achieved high sensitivities for recognizing kickboarding (range 97% - 99%) and skateboarding (range 93% - 95%). In contrast, the hip and pocket sensor multiclass models exhibited lower performance for these activities, with sensitivities of 78% for kickboarding and 49% to 58% for skateboarding. Most classification errors emanated from confusion between kickboarding and skateboarding. In addition, the low number of participants with a goofy stance may have affected model training, as their movement patterns are inherently opposite to those of regular riders. Future studies would benefit from a higher number of goofy-stance participants to improve the generalizability of the results. Binary classification models distinguishing push-push-glide activities from others significantly improve recognition accuracy.

Kickboard and Skateboard Recognition by the Hip and Pocket Sensor Configurations

The multiclass classification models achieved balanced accuracies ranging from 84% to 87%, comparable to those obtained with the hip and pocket sensor configurations (Table 2). The confusion matrices shown in Figures 2A and 2B indicated high accuracy of the estimators for the 7 activities unrelated to kickboarding or skateboarding. These findings are consistent with those of Suker et al [28], who reported classification rates of approximately 90% for recognizing sedentary and locomotive activities (excluding kickboarding and skateboarding) using accelerometer data from a smartphone sensor placed in the pocket. The recognition of kickboarding and skateboarding by the multiclass classification models proved more challenging. In particular, skateboarding yielded sensitivities of 49% with the hip sensor using accelerometer data and 58% with the pocket sensor using the accelerometer, gyroscope, and barometer data. Most classification errors arose from confusions between these 2 activities. This observation led to the testing of binary classification models considering a push-push-glide activity class versus all other activities. These models achieved balanced accuracies ranging from 92% to 95%. However, binary models distinguishing between kickboarding and skateboarding did not outperform the multiclass classification models, discouraging the implementation of a multilevel algorithmic procedure involving the initial recognition of push-push-glide activities followed by differentiation between kickboarding and skateboarding. Given the motion similarities between the 2 activities in the context of commuting, software applications designed to evaluate energy expenditure for these activities might consider halting the classification process at the push-push-glide activity recognition stage. Further studies investigating the energy expenditure associated with kickboarding and skateboarding in commuting contexts are necessary to validate this approach.

Kickboard and Skateboard Recognition by the Wrist Sensor Configurations

The performances of the multiclass classification models were promising, with balanced accuracies of 84% for wrist sensor with accelerometer data only and 88% for the wrist sensor with the accelerometer, gyroscope, and barometer data (Table 2). Interestingly, all locomotive activities including skateboarding and kickboarding achieved sensitivity scores of approximately ≥90%, except for the ascending stairs activity, which exhibited confusions with various activities (Figures 2C and 2D). Such values are on par with or above performances found in the literature for random forest models [29]. Confusions among sedentary activities, such as sitting or standing while chatting or scrolling a smartphone, also contributed to lower balanced accuracies in the multiclass classification models. These results should not overshadow the main objectives of this study. First, the primary aim was to evaluate the feasibility of recognizing kickboarding and skateboarding commuting behaviors. Second, distinguishing between sedentary and locomotive behaviors has been a standard feature in activity tracker software for over a decade [14,20]. It may be suggested that specific methods be developed to further improve the recognition of sitting and standing, but this was beyond the scope of this study.

With regard to the primary aim, kickboarding and skateboarding achieved sensitivity scores of ≥93% in the multiclass models. Binary classification models (eg, kickboard vs *rest*, skateboard vs *rest*, push-push-glide activities vs *rest*, and kickboard vs skateboard) were tested to improve performance. However, in all cases, binary models performed worse than the multiclass models. This discrepancy may stem from the lack of concurrent classification processes in binary models. For instance, even if a data point is moderately likely to belong to one activity class (eg, kickboarding), it might be far more likely to belong to another activity class (eg, skateboarding), which is captured more effectively in multiclass models. In those scenarios, eliminating 1 activity during evaluation yielded improved results (Multimedia Appendix 1), consistent with findings from models classifying push-push-glide activities versus all other activities. While binary classification models distinguishing kickboarding from skateboarding achieved 100% accuracies, confusion matrices for the push-push-glide activities versus *rest* revealed that 16% (SD 0.2%) of kickboarding or skateboarding activities were misclassified.

This suggests that a multilevel classification model, beginning with the recognition of push-push-glide activities and followed by differentiation between kickboarding and skateboarding, is unlikely to outperform direct multiclass classification models. Further studies conducted in less controlled environments will be necessary to identify the optimal approach for recognizing kickboarding and skateboarding behaviors.

Features

The analysis of feature importance enhances model explainability by providing insights into the classification process. First, data shown in Multimedia Appendix 2 highlight features extracted from roll-pitch-yaw analysis as recurrent variables in the classification models for the hip and pocket sensor configurations. These features likely capture the leg

XSL•FO
RenderX

extension during the push movement, which is more pronounced in kickboarding and skateboarding compared to other locomotive activities such as walking, running, and cycling. Features capturing low-frequency information are also highly represented, as the glide phase has a long duration, and the push (or propulsion) phase is extended compared to walking or running. Regarding wrist sensor configurations, FFT feature analysis revealed a prominence of signals in higher-frequency bands (ie, >10 Hz), likely due to vibrations from ground surface imperfections. Activity tracker software designed to recognize kickboarding and skateboarding may need to analyze signals in these higher frequencies—an approach that may not currently be standard, given that human movement typically occurs at lower-frequency ranges. Second, the analysis of the top features used to classify kickboard versus skateboard helps explain the difficulty in distinguishing these activities in the kickboard versus *rest* and skateboard versus *rest* models. Only the wrist sensor configurations could strongly differentiate the 2 activities. The feature named "accelerometer, z-axis, mean" captures the fact that the hands remain on the handlebar while kickboarding, whereas yaw measurement provides information about the wrist orientation, further aiding the differentiation. Third, the distribution of feature values across kickboarding, skateboarding, or other activities (Multimedia Appendix 1) revealed distinct activity-based patterns but also substantial overlaps. These overlaps cover a broad region of the feature space, supporting the hypothesis that treating *rest* activities as a single category increases the classification difficulty.

Beyond improving model interpretability, the analysis of feature importance has practical implications for industrial applications, particularly in optimizing memory and computational efficiency. For instance, in the push-push-glide versus *rest* classification task, the wrist sensor configuration using gyroscope data relied on a different set of accelerometer-based features compared to the wrist sensor configuration using accelerometer data only. Given that both configurations achieved comparable classification performance, this suggests redundancy in the extracted information, as expected. Consequently, the number of features could likely be reduced without compromising classification accuracy, thereby streamlining computational or hardware requirements for integration into commercial activity trackers.

This study used features previously suggested in the literature for recognizing daily life activities [16,22,23]. Such studies usually focus on common locomotive activities such as walking, running, and cycling. To address the challenge of recognizing kickboarding and skateboarding behaviors, a set of features was designed to capture the specific motion frequencies of these activities, as well as the body rotation specific to skateboarding. However, additional research may be needed to develop further features tailored to these activities.

Conclusions

This study demonstrates the feasibility of developing activity tracker software capable of classifying commuting behaviors such as kickboarding and skateboarding. Despite a relatively small sample size of 10 participants, the heterogeneity of the population (age, sex, riding skills, and stance) provided enough information for the machine learning algorithm to perform well, validating the approach. The wrist sensor configuration showed promising results, achieving high accuracy across locomotive activities, including kickboarding and skateboarding. In contrast, classification performances for hip and pocket sensor configurations were more variable, particularly in distinguishing between kickboarding and skateboarding, likely due to the similar lower limb motion patterns involved. Nevertheless, the recognition of the broader push-push-glide activity category achieved acceptable accuracies (range 87% - 91%), suggesting its potential as a practical intermediate classification step. Accurate classification would enable activity tracker software to implement intraclass regression algorithms for more precise predictions of energy expenditure associated with kickboarding and skateboarding commuting behaviors. However, future studies using a larger dataset are required to train models that could be deployed in existing commercial activity trackers. Finally, considering the respective strengths of wrist sensor and hip and pocket sensor configurations, models using sensor fusion approaches could further enhance classification performance.

Acknowledgments

The authors sincerely thank all study participants. JT, NAK, and YO received research funding from the Japan Society for the Promotion of Science (21K11335). During the preparation of this manuscript, the authors used OpenAI GPT-3.5 and DeepSeek to enhance language and readability. All content was subsequently reviewed and edited by the authors, who take full responsibility for the final publication.

Data Availability

Data and code are available on Zenedo's website [24].

Authors' Contributions

NAK contributed to study design, funding acquisition, statistical analysis, coding, data interpretation, and drafting the manuscript. HH was involved in data cleaning and processing, statistical analysis, coding, and manuscript drafting. AO contributed to data collection, cleaning, processing, and statistical analysis. TN participated in study design, funding acquisition, ethical oversight, and manuscript drafting. YO contributed to study design, funding acquisition, ethical oversight, and manuscript drafting. JT was responsible for study design, funding acquisition, ethical oversight, data acquisition, cleaning, processing, statistical analysis, coding, data interpretation, and drafting the manuscript.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional results.
[PDF File, 1999 KB - formative_v9i1e71969_app1.pdf]

Multimedia Appendix 2
Identification of the most important features.
[XLSX File, 13 KB - formative_v9i1e71969_app2.xlsx]

References

1. Rojas-Rueda D, de Nazelle A, Andersen ZJ, et al. Health impacts of active transportation in Europe. PLoS One 2016;11(3):e0149990. [doi: 10.1371/journal.pone.0149990] [Medline: 26930213]
2. Kuwahara K, Sawada SS, Noma H, et al. Commuter cycling and risk of type 2 diabetes: a cohort study in Japan. Diabetes Care 2022 Dec 1;45(12):e179-e180. [doi: 10.2337/dc22-1267] [Medline: 36251985]
3. Dinu M, Pagliai G, Macchi C, Sofi F. Active commuting and multiple health outcomes: a systematic review and meta-analysis. Sports Med 2019 Mar;49(3):437-452. [doi: 10.1007/s40279-018-1023-0] [Medline: 30446905]
4. Mueller N, Rojas-Rueda D, Cole-Hunter T, et al. Health impact assessment of active transportation: a systematic review. Prev Med 2015 Jul;76:103-114. [doi: 10.1016/j.ypmed.2015.04.010] [Medline: 25900805]
5. Fang K, Handy S. Skateboarding for transportation: exploring the factors behind an unconventional mode choice among university skateboard commuters. Transportation (Amst) 2019 Feb;46(1):263-283. [doi: 10.1007/s11116-017-9796-9]
6. Rainey S, Heffelman L, Parvin S, Ward AL. “Fun, faster than walking, and more convenient than a bike”: skateboarding in a college town. J Transp Health 2024 Sep;38:101888. [doi: 10.1016/j.jth.2024.101888]
7. Ward AL, McGee R, Gendall PJ. Strengths and vulnerabilities of teenagers who skateboard for transport in New Zealand. J Transp Health 2021 Mar;20:100947. [doi: 10.1016/j.jth.2020.100947]
8. Board WJ, Browning RC. Self-selected speeds and metabolic cost of longboard skateboarding. Eur J Appl Physiol 2014 Nov;114(11):2381-2386. [doi: 10.1007/s00421-014-2959-x] [Medline: 25085605]
9. Herrmann SD, Willis EA, Ainsworth BE, et al. 2024 Adult Compendium of Physical Activities: a third update of the energy costs of human activities. J Sport Health Sci 2024 Jan;13(1):6-12. [doi: 10.1016/j.jshs.2023.10.010] [Medline: 38242596]
10. Kijima A, Arimoto M, Muramatsu S. Metabolic equivalents during scooter exercise. J Physiol Anthropol 2007 Jun;26(4):495-499. [doi: 10.2114/jpa2.26.495] [Medline: 17704628]
11. Chen KY, Bassett DR Jr. The technology of accelerometry-based activity monitors: current and future. Med Sci Sports Exerc 2005 Nov;37(11 Suppl):S490-S500. [doi: 10.1249/01.mss.0000185571.49104.82] [Medline: 16294112]
12. Welk GJ. Principles of design and analyses for the calibration of accelerometry-based activity monitors. Med Sci Sports Exerc 2005 Nov;37(11 Suppl):S501-S511. [doi: 10.1249/01.mss.0000185660.38335.de] [Medline: 16294113]
13. Staudenmayer J, Pober D, Crouter S, Bassett D, Freedson P. An artificial neural network to estimate physical activity energy expenditure and identify physical activity type from an accelerometer. J Appl Physiol (1985) 2009 Oct;107(4):1300-1307. [doi: 10.1152/japplphysiol.00465.2009] [Medline: 19644028]
14. Ohkawara K, Oshima Y, Hikihara Y, Ishikawa-Takata K, Tabata I, Tanaka S. Real-time estimation of daily physical activity intensity by a triaxial accelerometer and a gravity-removal classification algorithm. Br J Nutr 2011 Jun;105(11):1681-1691. [doi: 10.1017/S0007114510005441] [Medline: 21262061]
15. Bassett DR Jr, Rowlands A, Trost SG. Calibration and validation of wearable monitors. Med Sci Sports Exerc 2012 Jan;44(1 Suppl 1):S32-S38. [doi: 10.1249/MSS.0b013e3182399cf7] [Medline: 22157772]
16. Willetts M, Hollowell S, Aslett L, Holmes C, Doherty A. Statistical machine learning of sleep and physical activity phenotypes from sensor data in 96,220 UK Biobank participants. Sci Rep 2018 May 21;8(1):7961. [doi: 10.1038/s41598-018-26174-1] [Medline: 29784928]
17. Csizmadia G, Liszkai-Peres K, Ferdinandy B, Miklósi Á, Konok V. Human activity recognition of children with wearable devices using LightGBM machine learning. Sci Rep 2022 Mar 31;12(1):5472. [doi: 10.1038/s41598-022-09521-1] [Medline: 35361854]
18. Chen D, Du Y, Liu Y, et al. Development and validation of a smartwatch algorithm for differentiating physical activity intensity in health monitoring. Sci Rep 2024 Apr 25;14(1):9530. [doi: 10.1038/s41598-024-59602-6] [Medline: 38664457]
19. Crouter SE, Kuffel E, Haas JD, Frongillo EA, Bassett DR Jr. Refined two-regression model for the ActiGraph accelerometer. Med Sci Sports Exerc 2010 May;42(5):1029-1037. [doi: 10.1249/MSS.0b013e3181c37458] [Medline: 20400882]
20. Oshima Y, Kawaguchi K, Tanaka S, et al. Classifying household and locomotive activities using a triaxial accelerometer. Gait Posture 2010 Mar;31(3):370-374. [doi: 10.1016/j.gaitpost.2010.01.005] [Medline: 20138524]
21. Shih PC, Han K, Poole ES, Rosson MB, Carroll JM. Use and adoption challenges of wearable activity trackers. Presented at: IConference 2015; Mar 24-27, 2015; Newport Beach, CA, USA.

XSL•FO
RenderX

22. Zhang S, Rowlands AV, Murray P, Hurst TL. Physical activity classification using the GENEA wrist-worn accelerometer. Med Sci Sports Exerc 2012 Apr;44(4):742-748. [doi: 10.1249/MSS.0b013e31823bf95c] [Medline: 21988935]
23. Ellis K, Kerr J, Godbole S, Staudenmayer J, Lanckriet G. Hip and wrist accelerometer algorithms for free-living behavior classification. Med Sci Sports Exerc 2016 May;48(5):933-940. [doi: 10.1249/MSS.0000000000000840] [Medline: 26673126]
24. Aubert-Kato N, Hatori H, Orihara A, Nakagata T, Ohta Y, Tripette J. Data and code "Recognizing skateboard and kickboard commuting behaviors using activity trackers: a feasibility study using machine learning approaches". Zenodo. URL: https://zenodo.org/records/14769798 [accessed 2025-08-06]
25. Bishop CM. Pattern Recognition and Machine Learning: Springer; 2006.
26. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res 2011;12(85):2825-2830.
27. Breiman L. Random forests. Mach Learn 2001 Oct;45(1):5-32. [doi: 10.1023/A:1010933404324]
28. Sukor ASA, Zakaria A, Rahim NA. Activity recognition using accelerometer sensor and machine learning classifiers. Presented at: CSPA 2018; Mar 9-10, 2018; Batu Feringghi, Malaysia. [doi: 10.1109/CSPA.2018.8368718]
29. Chong J, Tjurin P, Niemelä M, Jämsä T, Farrahi V. Machine-learning models for activity class prediction: a comparative study of feature selection and classification algorithms. Gait Posture 2021 Sep;89:45-53. [doi: 10.1016/j.gaitpost.2021.06.017] [Medline: 34225240]

Abbreviations

FFT: fast Fourier transform
MET: metabolic equivalent of task
PA: physical activity

Edited by J Sarvestan; submitted 02.02.25; peer-reviewed by M Fraser; revised version received 19.05.25; accepted 20.05.25; published 29.08.25.

Please cite as:
Aubert-Kato N, Hatori H, Orihara A, Nakagata T, Ohta Y, Tripette J
Recognizing Skateboard and Kickboard Commuting Behaviors Using Activity Trackers: Feasibility Study Using Machine Learning Approaches
JMIR Form Res 2025;9:e71969
URL: https://formative.jmir.org/2025/1/e71969
doi:10.2196/71969

© Nathanael Aubert-Kato, Hitomi Hatori, Arisa Orihara, Takashi Nakagata, Yuji Ohta, Julien Tripette. Originally published in JMIR Formative Research (https://formative.jmir.org), 29.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Web-Based Video Platforms as Sources of Information on Body Image Dissatisfaction in Adolescents: Content and Quality Analysis of a Cross-Sectional Study

Li Liu[1,2], BSc; Jianning Yang[3], ME; Fengmei Tan[2], MS; Huan Luo[4], BS; Yanhua Chen[4], PhD; Xiaolei Zhao[4], PhD

[1]School of Nursing, Southwest Medical University, Luzhou, China

[2]Department of Hematopathology, The Affiliated Hospital, Southwest Medical University, Luzhou, China

[3]Department of Nephrology, The Affiliated Hospital, Southwest Medical University, Luzhou, China

[4]Nursing Department, The Affiliated Hospital, Southwest Medical University, No 1, Section 1, Xianglin Road, Longmatan District, Luzhou, China

Corresponding Author:
Xiaolei Zhao, PhD
Nursing Department, The Affiliated Hospital, Southwest Medical University, No 1, Section 1, Xianglin Road, Longmatan District, Luzhou, China

Abstract

Background: Body image dissatisfaction among children and adolescents is a significant public health concern and is associated with numerous physical and mental problems. Social media platforms, including TikTok, BiliBili, and YouTube, have become popular sources of health information. However, the quality and reliability of content related to body image dissatisfaction have not been comprehensively evaluated.

Objective: The primary goal of this study was to examine the quality and reliability of videos related to body image dissatisfaction on TikTok, BiliBili, and YouTube.

Methods: The keywords “body image dissatisfaction” were searched on YouTube, TikTok, and BiliBili in November 2024. Videos were collected based on platform-specific sort filters, including the filter of “Most liked” on TikTok and the filter of “Most viewed” on BiliBili and YouTube. The top 100 videos on each platform were reviewed and screened in the study. After excluding videos that were (1) not in English or Chinese, (2) duplicates, (3) irrelevant, (4) no audio or visual, (5) contained advertisements, and (6) with a Global Quality Scale (GQS) score of 1, the final sample consisted of 64 videos, which formed the basis of our research and subsequent findings. Two reviewers (LL and JNY) screened, selected, extracted data, and evaluated all videos using the GQS, the Modified DISCERN (mDISCERN) scores, and the Modified Journal of the American Medical Association (mJAMA) benchmark criteria. Statistical analysis was performed using SPSS (version 28.0; IBM Corp).

Results: In total, 64 videos were analyzed in the study, including 20 from TikTok, 13 from BiliBili, and 31 from YouTube. The median duration of the involved videos was 3.01 (IQR 1.00-5.94) minutes on TikTok, 3.52 (IQR 2.36-5.63) minutes on BiliBili, and 4.86 (IQR 3.10-6.93) minutes on YouTube. Compared with the other 2 platforms, BiliBili videos received higher likes and more comments. The majority of the videos (n=40, 62%) were uploaded by self-media. The quality of the videos on YouTube shows the highest overall scores. Videos uploaded by professional authors had significantly higher GQS, mDISCERN, and mJAMA scores compared to those uploaded by nonprofessionals. There was no significant correlation between video quality and the number of views or likes. However, the number of views and likes were significantly positively correlated. Furthermore, a significant correlation was found between the mJAMA, mDISCERN, and GQS scores.

Conclusions: Web-based video platforms have become an important source for adolescents to access health information. However, the lack of a significant correlation between video quality and the number of likes and comments poses a challenge for users seeking reliable health information. It is suggested that the quality of the videos on health information would be taken into consideration in the recommendation algorithm on web-based video platforms.

(JMIR Form Res 2025;9:e71652) doi:10.2196/71652

KEYWORDS

adolescents; body image; video; quality analysis; TikTok; BiliBili; YouTube; GQS; Global Quality Scale; mDISCERN; Modified DISCERN; mJAMA; Modified Journal of the American Medical Association

XSL•FO
RenderX

Introduction

Body image is defined as a person's perceptions, emotions, and thoughts about his or her body [1]. Body image dissatisfaction is the difference between how an individual perceives their physical appearance characteristics and their desire to be [2]. There is a relatively high incidence of body image dissatisfaction among adolescents who undergo psychological and physical changes during puberty and attempt to fit in with cultural expectations of a slim body shape. A web-based international survey with 10,765 youths aged 16 to 25 years from Australia, Canada, New Zealand, the United Kingdom, and the United States showed a body image dissatisfaction rate as high as 75.19% [3]. Another study in Southwest China indicated that 55.2% of young women experienced body image dissatisfaction in a cross-sectional study with 4149 college students in Sichuan province and Chongqing city [4]. Adolescents with body image dissatisfaction are associated with psychological and mental problems such as low self-esteem [5,6], anxiety [7,8], depression [9,10], and chronic dysphoria [11]. It can also lead to a series of serious physical and behavioral issues, such as eating disorders [12], extreme weight loss behaviors [13], alcohol [14], drug abuse [15], excessive pursuit of cosmetic surgeries [16], and suicidality [17,18].

The internet has become an indispensable part of adolescents' lives [19]. Web-based video platforms have gradually become crucial for adolescents to obtain health information and interact with others [20]. According to a study in Austria, 79% of respondents get their health information through websites [21], which was more prevalent than the doctors. TikTok, BiliBili, and YouTube, because of their ease of use and anonymous accessibility, are preferred for adolescents to obtain health information. Video platforms have become the primary medium for adolescents to seek health information since video content is more vivid and efficient than traditional text-based information. However, it is difficult for users to access the reliability and quality of the video content because of the lack of peer review and guidelines.

Studies examined the quality of educational videos concerning physical diseases on YouTube, covering topics like pain relief [22], kidney transplantation [23], and COVID-19 [24]. It was demonstrated that the quality of the educational videos varied [25]. No studies have analyzed the accuracy and quality of the videos related to body image dissatisfaction among adolescents on web-based video platforms. Therefore, this study aims to assess the accuracy, adequacy, quality, and reliability of the video content on body image dissatisfaction available on TikTok, BiliBili, and YouTube.

Methods

Ethical Considerations

This study was conducted without the use of clinical data, human specimens, or laboratory animals. All relevant information was derived from publicly available platforms, including TikTok, BiliBili, and YouTube, and none of the data entailed personal privacy concerns, as they were deidentified. Therefore, an ethics review was deemed unnecessary.

Data Collection

The keyword "body image dissatisfaction" was searched on YouTube, TikTok, and BiliBili in November 2024. New user accounts were created on each of the 3 web-based video platforms to ensure that the data collected were unbiased, without being influenced by previous search history. Videos were collected based on platform-specific sort filters, including the filter of "Most liked" on TikTok and the filter of "Most viewed" on BiliBili and YouTube. The top 100 videos on each platform were reviewed and screened in the study. The 2 researchers (LL and JNY) reviewed the videos independently. The videos in relation to causes, symptoms, or coping strategies of body image dissatisfaction were included. After excluding videos that were (1) not in English or Chinese, (2) duplicates, (3) no audio or visual, (4) contained advertisements, and (5) with a GQS score of 1, the final sample consisted of 64 videos, which formed the basis of our research and subsequent findings.

The following parameters were documented, including uploaders, upload date, platform, views, likes, comments, shares, saves, video duration, and the quality of the video (Multimedia Appendix 1). However, the following data were unavailable: (1) views on TikTok and (2) shares and saves on YouTube. The uploaders could be divided into individuals (self-media and medical workers), professional organizations, and for-profit corporations, such as Dove. Self-media uploaders were regarded as nonprofessionals, while others were regarded as professionals.

Video Review

In November 2024, 2 reviewers (LL and JNY) independently reviewed the videos using the 3 quality assessment tools. The videos about causes, symptoms, and coping strategies of body image dissatisfaction were involved. Videos not covering these contents were deemed irrelevant and should be excluded. Discrepancies were resolved through discussion or consultation with the third author (XLZ). A video could have multiple types of content, so the ratios and percentages were reported as higher than the total number of videos.

Video Quality Assessments

Three widely adopted standardized scales, namely the Global Quality Scale (GQS) [26], Modified DISCERN (mDISCERN) [27], and Modified Journal of the American Medical Association (mJAMA) [28] were selected to assess the quality of the videos.

The GQS is a 5-point Likert-type scale used to measure a video's overall quality, flow, and usefulness, ranging from 1 to 5 according to quality (a higher score indicates higher quality). Generally, scores of 4 - 5 points indicated high quality, 3 points indicated moderate quality, and 1 - 2 points indicated low quality [26] (Multimedia Appendix 2).

The mDISCERN, a revised version of the original DISCERN, is more suitable for evaluating video-based content. It consists of 5 standards to assess the reliability of video content, including clarity, relevance, fairness, traceability, and robustness. Each of the above questions was scored 1 point for "yes" and 0 point for "no," and the cumulative score was calculated as 0 - 5 points. Higher scores signified superior reliability and quality (Multimedia Appendix 3).

XSL•FO
RenderX

The mJAMA criteria is a well-known quality assessment tool used to evaluate information from health-related websites. The mJAMA score awards 1 point for the following 4 standards: authorship, attribution, disclosure, and currency. A score of 4 represents the highest quality (Multimedia Appendix 4).

Statistical Analysis

Statistical analysis was performed using SPSS (version 28.0; IBM Corp). Categorical variables were presented as frequencies and percentages. Continuous variables were expressed as mean and SD or median (IQR). The Shapiro-Wilk tests were used to test the normality of the data distribution. For data with a normal distribution, an independent sample 2-tailed *t* test was used. The Mann-Whitney *U* test was conducted to compare the 2 groups of quantitative data that were not normally distributed. The Kruskal-Wallis test was used to compare 3 or more groups. The Spearman test was conducted to examine the correlation between the scores and video characteristics. The Spearman correlation coefficient was low ($\rho<0.40$), moderate ($\rho=0.40$ - 0.59), and high ($\rho>0.60$). Cohen κ was used to measure the agreement for categorical data between the 2 raters, while an intraclass correlation coefficient (ICC) was used to assess interrater reliability on quantitative data. The κ values were interpreted as follows: $\kappa>0.8$ indicated outstanding consistency, $0.6<\kappa\leq0.8$ suggested good agreement, $0.4<\kappa\leq0.6$ signified moderate agreement, and $\kappa\leq0.4$ was indicative of poor agreement [29]. The ICC values were interpreted as follows: ICC>0.9 indicated outstanding consistency, 0.75<ICC≤0.9 suggested good agreement, 0.5<ICC≤0.75 signified moderate agreement, and ICC≤0.5 was indicative of poor agreement [30]. A 2-tailed *P* value of <.05 indicates statistical significance. For variables with missing data (eg, TikTok views and YouTube shares and saves), listwise deletion was applied.

Results

Overview

The ICC values showing the level of agreement between the 2 independent raters (LL and JNY) assessing the videos are presented with 95% CIs. There was excellent statistical agreement among the raters, the ICC for GQS was 0.87 (95% CI 0.78-0.92), the ICC for mDISCERN was 0.84 (95% CI 0.76-0.90), and the ICC for the mJAMA score was 0.86 (95% CI 0.78-0.91), with all *P* values <.001.

Video Characteristics

This study evaluated 100 videos on 3 platforms, respectively. A flowchart of the video selection process is reported in Figure 1. The 236 videos were excluded due to being duplicates (n=28), having a duration exceeding 30 minutes (n=27), having no audio or audio only (n=10), or being unrelated to body image (n=113). A refined collection of 20 videos from TikTok, 13 from BiliBili, and 31 from YouTube was included for the subsequent analysis. Among the 64 videos included, the uploaders of 40 videos were self-media uploaders, 12 were from professional organizations, 6 were from medical workers, and another 6 were from for-profit corporations. The average duration of the videos included was 4.53 (SD 3.01; range 0.53-14.16) minutes. There were videos with as few as 946 views and those with more than 10,000,000 views. The number of likes varied from 10 to 270,000, with a median of 907 (IQR 10 - 270,000). However, views on TikTok and share and save on YouTube are unavailable. The median GQS of these videos was 4 (IQR 2-5), the median mDISCERN score was 3 (IQR 1-5), and the median mJAMA score was 3 (IQR 1-4). For general information about the videos, such as uploaders, likes, time since upload, and duration, see Table 1.

XSL•FO
RenderX

Figure 1. Search strategy for videos on body image dissatisfaction.

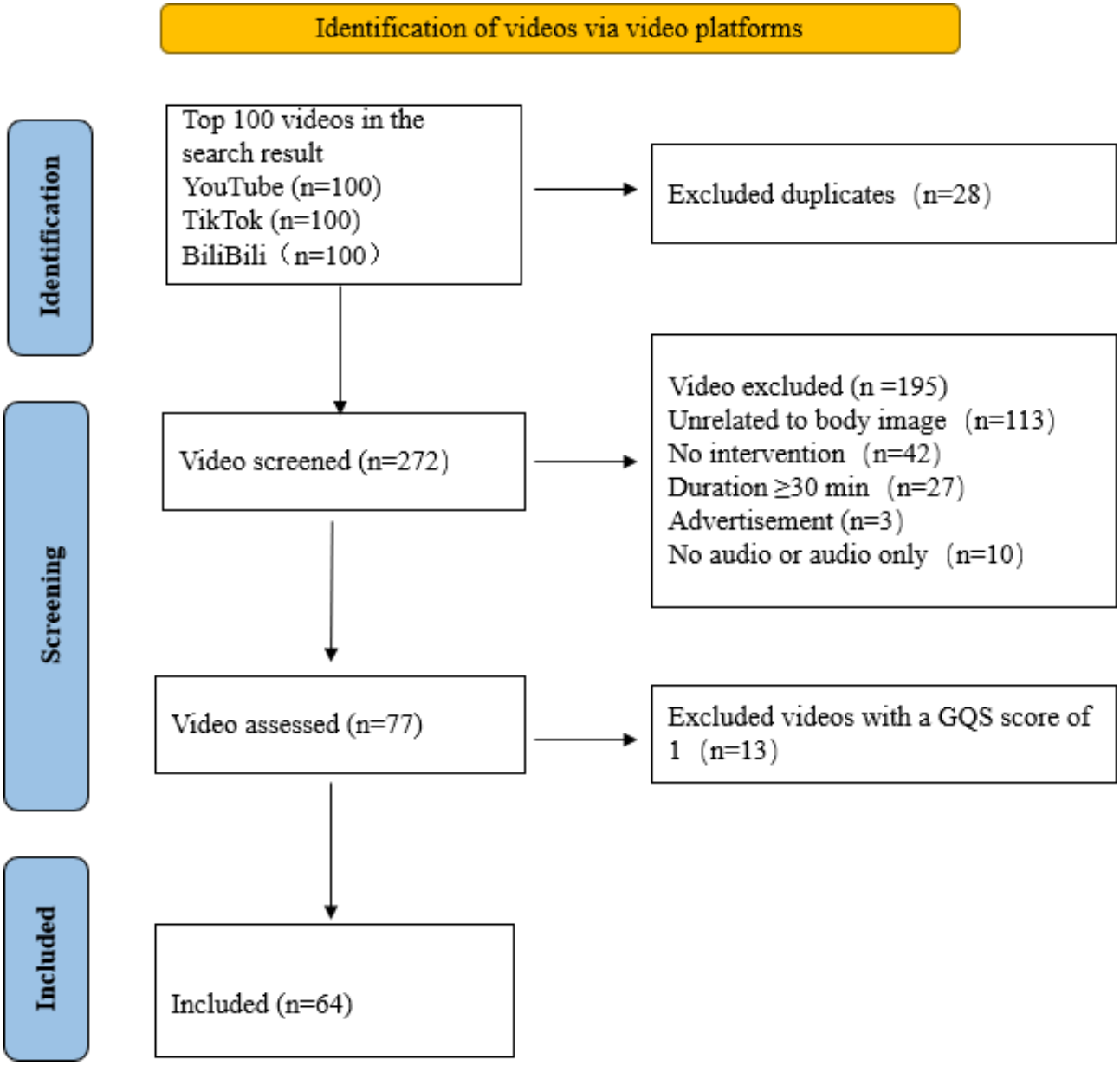

XSL•FO
RenderX

Table . Detailed characteristics of the videos (n=64).

Parameters	Values, median (IQR)
Duration (minutes)	4.13 (0.53-14.16)
Views[a]	59,000 (946-10,400,000)
Likes	907 (10 - 270,000)
Shares[b]	475 (1 - 86,000)
Saves[b]	341 (4 - 75,000)
Times since video upload (months)	28.5 (1-142)
GQS[c] score	4 (2-5)
mDISCERN[d] score	3 (1-5)
mJAMA[e] score	3 (1-4)

[a]Data from BiliBili and YouTube.
[b]Data from TikTok and BiliBili.
[c]GQS: Global Quality Scale.
[d]mDISCERN: Modified DISCERN.
[e]mJAMA: Modified Journal of the American Medical Association.

Video Content Analysis

Content analysis was conducted in the involved videos. Coping strategies were the most frequently discussed topics, with 84% (n=54) of the videos providing suggestions for self-acceptance, refusal to make comparisons, social support seeking, and life adjustment. Concerning the treatments of body image dissatisfaction, cognitive behavioral therapy and meditation are the most widely applied, as shown in Textbox 1. In addition, causes and symptoms of body image dissatisfaction were also topics of public concern, with 23% (n=15) and 52% (n=33) of videos providing full explanations, respectively.

XSL•FO
RenderX

Textbox 1. Recommended therapies and tips for body image dissatisfaction.

Ten recommended therapies for body image dissatisfaction

1. Cognitive behavioral therapy (n=5)
2. Meditation (n=5)
3. Selective serotonin reuptake inhibitors (n=3)
4. Mirror therapy (n=2)
5. Sports participation (n=2)
6. Mindfulness (n=1)
7. Dance therapy (n=1)
8. Hypnotherapy (n=1)
9. Talk therapy (n=1)
10. Point percussion therapy (n=1)

Ten tips for building a positive body image

1. Self-acceptance (n=14)
2. Social support seeking (n=8)
3. Stay away from negative comments (n=7)
4. Focus on your strengths (n=6)
5. Express gratitude to your body (n=6)
6. Stop comparing (n=5)
7. Define beauty by yourself (n=3)
8. Positive self-affirmation (n=2)
9. Take part in social activities (n=2)
10. Wear clothes that make you feel good and truly suit you (n=1)

Characteristics Comparison of Different Platforms

The characteristics and the quality of the involved videos on 3 web-based video platforms are presented in Table 2. The median duration of the included video sources was 3.01 (IQR 0.53 - 8.75) minutes on TikTok, which was shorter than that on BiliBili (median 3.52, IQR 1.18 - 14.16 minutes) and YouTube (median 4.86, IQR 1 - 13.15 minutes). There was a significant difference in video duration between TikTok and YouTube. The median time since the included videos uploaded was 18 (IQR 1-38) months on TikTok, which was significantly shorter than those of the videos posted on BiliBili (median 30, IQR 13-91 months) and YouTube (median 35, IQR 3-142 months). For the video characteristics, there were no statistics differences in views, likes, comments, or saves among the 3 groups. However, the median number of likes and comments was 2169 (IQR 39 - 193,000) and 60 (IQR 3 - 2527) on BiliBili, respectively, which were higher than that on TikTok and YouTube. The median number of views on YouTube was higher than those videos on BiliBili, and the median number of shares and saves on BiliBili was higher than those videos on TikTok. In terms of video quality, the GQS, mDISCERN, and mJAMA scores of the included videos on YouTube were significantly better than those of videos posted on TikTok and BiliBili.

XSL•FO
RenderX

Table . Characteristics of video platforms about body image dissatisfaction.

Characteristic	Platforms			*P* value[a]		
	TikTok (n=20), median (IQR)	BiliBili (n=13), median (IQR)	YouTube (n=31), median (IQR)	$P_{T\text{-}B}$ [b]	$P_{B\text{-}Y}$ [c]	$P_{T\text{-}Y}$ [d]
Time since video upload (months)	18 (1 - 38)	30 (13 - 91)	35 (3 - 142)	.04[e]	.10	<.001[e]
View	__[f]	29,000 (946 - 1,323,000)	117,437 (2020 - 10,400,000)	—	.29	—
Likes	537 (13 - 238,000)	2169 (39 - 193,000)	1502 (0 - 270,000)	.54	.92	.58
Comments	37.5 (0 - 15,000)	60 (3 - 2527)	36 (0 - 16,634)	.47	.76	.99
Share	396 (1 - 86,000)	475 (34 - 23,000)	—	.84	—	—
Save	237 (4 - 39,000)	718 (24 - 75,000)	—	.11	—	—
Duration (minutes)	3.01 (0.53 - 8.75)	3.52 (1.18 - 14.16)	4.86 (1 - 13.15)	.31	.28	.03[e]
GQS[g]	3 (2 - 4)	3.0 (2 - 5)	4 (2 - 5)	.98	.02[e]	.01[e]
mDISCERN[h]	2.5 (1 - 4)	2 (1 - 5)	4 (2 - 5)	.50	<.001[e]	<.001[e]
mJAMA[i]	2 (2 - 3)	2.0 (1 - 3)	4 (2 - 4)	.52	<.001[e]	<.001[e]

[a]All the *P* values were obtained from Mann-Whitney *U* test.

[b]$P_{T\text{-}B}$: TikTok versus BiliBili.

[c]$P_{B\text{-}Y}$: BiliBili versus YouTube.

[d]$P_{T\text{-}Y}$: TikTok versus YouTube.

[e]*P* value <.05.

[f]Not available.

[g]GQS: Global Quality Scale.

[h]mDISCERN: Modified DISCERN.

[i]mJAMA: Modified Journal of the American Medical Association.

Characteristics Comparison of Different Video Sources

Table 3 shows the characteristics of video uploaders about body image dissatisfaction on the 3 web-based video platforms. The result showed that most videos (n=40, 63%) were uploaded by self-media. The median duration of the videos posted by medical workers was 6.56 (IQR 2.53 - 13.15) minutes, which was longer than the videos from other video sources. The median number of likes was 2313 (IQR 10 - 35,000) for videos posted by professional organizations, which was higher than the videos posted by self-media uploaders (median 1384, IQR 13 - 27,000), medical workers (median 1204, IQR 0 - 5611), and commercial corporations (median 456, IQR 50 - 81,000). The median number of comments was 49.5 (IQR 0 - 16,634) for videos posted by self-medias, which was more than those videos from other video sources. Regarding the video quality, videos uploaded by professionals had significantly higher GQS, mDISCERN, and mJAMA scores compared with those uploaded by nonprofessionals.

XSL•FO
RenderX

Table . Characteristics of video uploaders about body image dissatisfaction on TikTok or BiliBili or YouTube.

Characteristics	Uploaders				*P* value					
	Self-medias (n=40), median (IQR)	Professional organizations (n=12), median (IQR)	Medical workers (n=6), median (IQR)	For-profit corporations (n=6), median (IQR)	$P_{s\text{-}p}$ [a]	$P_{p\text{-}m}$ [b]	$P_{m\text{-}c}$ [c]	$P_{s\text{-}c}$ [d]	$P_{s\text{-}m}$ [e]	$P_{p\text{-}c}$ [f]
Time since video upload (months)	26 (3 - 142)	30 (1 - 110)	38.5 (31 - 95)	24 (3 - 64)	.62	.20	.09	.55	.03[g]	.38
Views	__[h]	—	34,500 (8858 - 190,000)	203,409 (15,971 - 3,087,713)	—	—	.24	—	—	—
Likes	1384 (13 - 27,000)	2313 (10 - 35,000)	1204 (0 - 5611)	456 (50 - 81,000)	.98	.49	.99	.57	.47	.55
Comments	49.5 (0 - 16,634)	37.5 (0 - 2334)	37 (0 - 667)	0 (0 - 415)	.68	.82	.13	.01[g]	.64	.08
Duration (minutes)	4.13 (0.53 - 14.16)	4.05 (1.67 - 7.68)	6.56 (2.53 - 13.15)	3.29 (1 - 7.36)	.82	.10	.13	.81	.09	.89
GQS[i]	3.5 (2 - 5)	3 (3 - 4)	4 (2 - 5)	4 (3 - 5)	.99	.49	.93	.31	.49	.29
mDISCERN[k]	3 (1 - 5)	4 (1 - 5)	4 (3 - 5)	4.5 (4 - 5)	.03[g]	.49	.48	.01[g]	.08	.18
mJAMA[j]	2 (1 - 4)	4 (2 - 4)	4 (3 - 4)	4 (4 - 4)	<.001[g]	.75	.39	<.001[g]	<.001[g]	.18

[a]$P_{s\text{-}p}$: self-medias versus professional organizations.

[b]$P_{p\text{-}m}$: professional organizations versus medical workers.

[c]$P_{m\text{-}c}$: medical workers versus corporations.

[d]$P_{s\text{-}c}$: self-medias versus corporations.

[e]$P_{s\text{-}m}$: self-medias versus medical workers.

[f]$P_{p\text{-}c}$: professional organizations versus corporations.

[g]*P* value <.05.

[h]Not available.

[i]GQS: Global Quality Scale.

[j]mJAMA: Modified Journal of the American Medical Association.

[k]mDISCERN: Modified DISCERN.

Correlation Analysis

A correlation analysis was performed between different video characteristics and different scores in Table 4. A strong correlation was found between the uploaders and platforms (r=0.641; P<.001), as well as a moderate correlation between platforms and the time since videos were uploaded (r=0.497; P<.001), and a low correlation between platforms and video duration (r=0.28; P=.02). Moderate to strong correlations were observed between views and likes (r=0.880; P<.001) and views and comments (r=0.561; P<.001). A strong correlation was found between mDISCERN and mJAMA (r=0.67; P<.001), a moderate correlation between GQS and mDISCERN (r=0.558; P<.001), and a weak correlation between GQS and mJAMA (r=0.306; P=.01). Additionally, low to strong correlations were found between platforms and mJAMA (r=0.782; P<.001), mDISCERN (r=0.594; P<.001), and GQS (r=0.39; P<.001). Low to moderate correlations were observed between the time since videos were uploaded and mDISCERN (r=0.454; P<.001), GQS (r=0.405; P<.001), and mJAMA (r=0.368; P<.001).

XSL•FO
RenderX

Table . Correlation analyses[a].

		Uploaders	Time since video upload	Platform	Views	Likes	Comments	Duration	GQS[b]	mDISCERN[c]	mJAMA[d]
Uploaders											
	r	__[e]	0.106	0.641	0.097	0.092	−0.23	0.007	0.133	0.518	0.777
	P value	—	.41	<.001[f]	.54	.47	.07	.58	.29	<.001[f]	<.001[f]
Time since video upload											
	r	0.106	—	0.497	0.186	0.11	0.188	0.175	0.405	0.454	0.369
	P value	.41	—	<.001[f]	.23	.38	.14	.17	<.001[f]	<.001[f]	<.001[f]
Platform											
	r	0.641	0.497	—	0.192	0.072	0.02	0.282	0.393	0.594	0.782
	P value	<.001[f]	<.001[f]	—	.21	.57	.99	.02[f]	<.001[f]	<.001[f]	<.001[f]
Views											
	r	0.097	0.186	0.192	—	0.880	0.561	0.011	−0.033	0.128	0.201
	P value	.54	.23	.21	—	<.001[f]	<.001[f]	.94	.83	.41	.19
Likes											
	r	0.092	0.11	0.072	0.880	—	0.778	−0.084	−0.056	0.007	0.030
	P value	.47	.38	.57	<.001[f]	—	<.001[f]	.51	.66	.95	.81
Comments											
	r	−0.23	0.188	0.02	0.561	0.778	—	0.089	−0.048	−0.048	−0.079
	P value	.07	.14	.99	<.001[f]	<.001[f]	—	.49	.71	.71	.54
Duration											
	r	0.07	0.175	0.282	0.011	−0.084	0.089	—	0.463	0.317	0.201
	P value	.58	.17	.02[f]	.94	.51	.49	—	<.001[f]	.01[f]	.11
GQS											
	r	0.133	0.405	0.393	−0.033	−0.056	−0.048	0.463	—	0.558	0.306
	P value	.29	<.001[f]	<.001[f]	.83	.66	.71	<.001[f]	—	<.001[f]	.01[f]
mDISCERN											
	r	0.518	0.454	0.594	0.128	0.007	−0.048	0.317	0.558	—	0.674
	P value	<.001[f]	<.001[f]	<.001[f]	.41	.95	.70	.01[f]	<.001[f]	—	<.001[f]
mJAMA											
	r	0.777	0.369	0.782	0.201	0.030	−0.079	0.204	0.306	0.674	—
	P value	<.001[f]	<.001[f]	<.001[f]	.19	.81	.54	.11	.01[f]	<.001[f]	—

[a]All the *P* values were obtained from Spearman ρ.

[b]GQS: Global Quality Scale.

[c]mDISCERN: Modified DISCERN.

[d]mJAMA: Modified Journal of the American Medical Association.

[e]Not available.

[f]*P* value <.05.

XSL•FO
RenderX

Discussion

Principal Findings

Web-based video platforms provide unprecedented opportunities for viewers to swiftly obtain and disseminate health information [31], including allowing health communication for adolescents with body image dissatisfaction. The length of the videos on TikTok was shorter than that of the videos on BiliBili and YouTube (*P*=.03), and the time since videos were uploaded was shortest on TikTok. The audience engagement metrics (median likes or comments) were highest on BiliBili. The majority of videos (n=40, 63%) were uploaded from self-media. Videos uploaded by medical workers, organizations, and corporations demonstrated significantly higher GQS, mDISCERN, and mJAMA scores compared with those uploaded by self-media. An important finding was the lack of significant correlation between the objective video quality metrics (GQS, mDISCERN, and mJAMA) and user engagement indicators (likes and comments), indicating that the popularity of videos would not reflect their quality.

This study found a moderate correlation between platforms and the time since videos were uploaded. The median time since videos were uploaded was 18 (IQR 1-38) months on TikTok, which was significantly shorter than those of BiliBili (median 30, IQR 13-91 months) and YouTube (median 35, IQR 3-142 months), consistent with previous research [32]. This may reflect that platform algorithms prioritize novelty on TikTok. The study also showed that BiliBili had higher likes and comments than TikTok and YouTube. This could be attributed to BiliBili's unique features, such as bullet-screen comments, which may enhance interactivity among viewers. However, differences in audience demographics (eg, age and cultural background) and uploader popularity could also contribute to engagement disparities. Additionally, among the included videos, BiliBili focused more on storytelling. These findings align with the narrative transportation theory [33], in which storytelling facilitates emotional immersion and reduces counterarguing, thereby enhancing perceived authenticity and trust [34]. This may explain why videos uploaded by professionals, despite higher quality scores, received fewer comments. Besides, there was no correlation between video quality and audience engagement in digital literacy, and users may prioritize emotional resonance over source credibility in this study. This highlights an urgent need for platforms to integrate trust cues, such as verified author badges and quality ratings, into interfaces to guide users to access reliable videos [35].

Comparison to Prior Work

In terms of video quality, the videos about body image dissatisfaction on YouTube were significantly better than those on TikTok and BiliBili. This result was inconsistent with previous research that mentioned the quality was better on BiliBili [36]. In this study, over 93% (n=31) of the videos on TikTok and BiliBili were uploaded by self-media with more poor content about body image dissatisfaction. On YouTube, a considerable proportion of the videos was uploaded by medical workers and profit corporations. According to the mDISCERN and mJAMA scores, videos uploaded by professionals were significantly better than those uploaded by nonprofessionals, which is consistent with the results of previous studies [37,38]. Additionally, the included videos from profit corporations received fewer comments. This may be because these videos were mostly from Dove with high-quality animation productions, but these videos were silent animations lack of textual explanations.

Low-to-moderate positive correlations were observed between the time since the videos were uploaded and the scores on mDISCERN, GQS, and mJAMA, indicating that higher scores on these assessments may take a considerable amount of time to develop. The video quality score of BiliBili and YouTube was better than TikTok because they promote deeper engagement through in-depth discussion, which requires a longer time to be fully presented and a relatively longer production cycle. Low to strong correlations were found between GQS, mDISCERN, and mJAMA scores. This is consistent with results reported in prior studies [39], indicating the mutual verifiability among video quality assessment tools. There was no significant correlation between video quality and the number of likes or comments. This result, consistent with previous similar studies [29], may indicate that viewers often do not care about the quality of health-related videos.

Strengths and Limitations

Strengths

The quality assessment using GQS, mDISCERN, and mJAMA scales provided comprehensive reliability. Cross-platform comparison (TikTok, BiliBili, and YouTube) in English and Chinese improved the generalizability of the result. The interrater reliability with κ value of 0.767 indicated a good agreement between the raters.

Limitations

This study has several limitations. First, the sample size was relatively small, with only 64 videos included, which could not represent all the videos about body image dissatisfaction on the 3 video platforms. Second, the videos were restricted to Chinese and English, which would overlook potentially valuable body image dissatisfaction videos in other languages. This also may lead to the omission of important information about body image dissatisfaction in different cultural backgrounds. Third, this study only analyzed the latest 100 videos in the search results, and some high-quality videos uploaded previously may have been missed. Fourth, the unavailability of TikTok views and YouTube shares and saves may introduce bias in correlation analysis. Fifth, the study did not account for potential confounders such as audience demographics, uploader popularity when exploring the causes of differences in video engagement, and quality. In the future, the researchers will focus on the development of high-quality body image dissatisfaction videos by following scientific processes. It is suggested that the quality of the videos on health information would be taken into consideration in the recommendation algorithm.

Future Directions

Future research should include videos with more languages, cultures, and video types. Longitudinal analysis tracking video

XSL•FO
RenderX

quality and engagement over time is needed to understand how algorithms impact content quality. Platforms can integrate quality criteria into algorithms by collaborating with medical experts. A specific bullet could be designed by the platforms to indicate the quality of the videos by using GQS, mDISCERN, and mJAMA scales.

Conclusions

This study offers a valuable exploration into the landscape of videos regarding body image dissatisfaction on TikTok, BiliBili, and YouTube. We discovered that web-based video platforms have indeed emerged as significant sources of health information for the public. However, the lack of a significant correlation between video quality and the number of likes and comments poses a challenge for users seeking reliable health information. The quality of the videos on health information would be taken into consideration in the recommendation algorithm on web-based video platforms in the future.

Acknowledgments

The authors would like to thank Professor Roger Watson for his help with the review statement and English language editing. This research received a Science Popularization Works Creation Project of Southwest Medical University (project 2024ZP005).

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

LL, HL, and FT conceived and designed the study. LL and JY searched the top 300 videos, collected characteristics, and performed reviewing, classifying, and assessing of the videos. LL analyzed the data and drafted the original manuscript. YC and XZ reviewed and edited the manuscript. All the authors contributed to manuscript writing and editing and approved the final draft for submission.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Parameters.
[XLSX File, 15 KB - formative_v9i1e71652_app1.xlsx]

Multimedia Appendix 2
Global Quality Scale.
[DOCX File, 20 KB - formative_v9i1e71652_app2.docx]

Multimedia Appendix 3
Modified DISCERN.
[DOCX File, 20 KB - formative_v9i1e71652_app3.docx]

Multimedia Appendix 4
Modified Journal of the American Medical Association.
[DOCX File, 20 KB - formative_v9i1e71652_app4.docx]

References

1. Gonzaga I, Ribovski M, Claumann GS, et al. Secular trends in body image dissatisfaction and associated factors among adolescents (2007-2017/2018). PLoS ONE 2023;18(1):e0280520. [doi: 10.1371/journal.pone.0280520] [Medline: 36656894]
2. Jiménez Flores P, Jiménez Cruz A, Bacardi Gascón M. Body-image dissatisfaction in children and adolescents: a systematic review. Nutr Hosp 2017 Mar 30;34(2):479-489. [doi: 10.20960/nh.455] [Medline: 28421808]
3. Milton A, Hambleton A, Roberts A, et al. Body Image Distress and Its Associations From an International Sample of Men and Women Across the Adult Life Span: Web-Based Survey Study. JMIR Form Res 2021 Nov 4;5(11):e25329. [doi: 10.2196/25329] [Medline: 34734831]
4. Yunyu Z, Yu L, Jackson T, Yang W, Hong C. Towards evaluating an internet-based dissonance intervention to reduce body dissatisfaction among young Chinese women: some preliminary findings (Article in Chinese). Commun Psychol Res 2016;2(1):150-167 [FREE Full text]
5. Morán C, Victoriano M, Parra J, et al. Body dissatisfaction, self-esteem, depressive symptoms, and nutritional status in adolescents. Andes Pediatr 2024 Feb;95(1):69-76. [doi: 10.32641/andespediatr.v95i1.4779] [Medline: 38587346]

XSL•FO
RenderX

6. Davison TE, McCabe MP. Adolescent body image and psychosocial functioning. J Soc Psychol 2006 Feb;146(1):15-30. [doi: 10.3200/SOCP.146.1.15-30] [Medline: 16480119]
7. Nazmi S, Nikbakht HA, Behmanesh F, Gholamnia-Shirvani Z, Azizi A. Body image concern and demographic characteristics as predictors of anxiety in adolescent girls. Int J Adolesc Med Health 2024 Aug 1;36(4):409-417. [doi: 10.1515/ijamh-2024-0070] [Medline: 38899865]
8. Pritchard M, Brasil K, McDermott R, Holdiman A. Untangling the associations between generalized anxiety and body dissatisfaction: the mediating effects of social physique anxiety among collegiate men and women. Body Image 2021 Dec;39:266-275. [doi: 10.1016/j.bodyim.2021.10.002] [Medline: 34695680]
9. Blundell E, De Stavola BL, Kellock MD, et al. Longitudinal pathways between childhood BMI, body dissatisfaction, and adolescent depression: an observational study using the UK Millenium Cohort Study. Lancet Psychiatry 2024 Jan;11(1):47-55. [doi: 10.1016/S2215-0366(23)00365-6] [Medline: 38101872]
10. Flores-Cornejo F, Kamego-Tome M, Zapata-Pachas MA, Alvarado GF. Association between body image dissatisfaction and depressive symptoms in adolescents. Braz J Psychiatry 2017;39(4):316-322. [doi: 10.1590/1516-4446-2016-1947] [Medline: 28355343]
11. Rosenström T, Jokela M, Hintsanen M, et al. Body-image dissatisfaction is strongly associated with chronic dysphoria. J Affect Disord 2013 Sep 5;150(2):253-260. [doi: 10.1016/j.jad.2013.04.003] [Medline: 23639211]
12. Foster L, Lundh LG, Daukantaité D. Disordered eating in a 10-year perspective from adolescence to young adulthood: stability, change, and body dissatisfaction as a predictor. Scand J Psychol 2024 Feb;65(1):32-41. [doi: 10.1111/sjop.12950] [Medline: 37491950]
13. Neumark-Sztainer D, Paxton SJ, Hannan PJ, Haines J, Story M. Does body satisfaction matter? Five-year longitudinal associations between body satisfaction and health behaviors in adolescent females and males. J Adolesc Health 2006 Aug;39(2):244-251. [doi: 10.1016/j.jadohealth.2005.12.001] [Medline: 16857537]
14. Holzhauer CG, Zenner A, Wulfert E. Poor body image and alcohol use in women. Psychol Addict Behav 2016 Feb;30(1):122-127. [doi: 10.1037/adb0000115] [Medline: 26348220]
15. Copeland J, Peters R, Dillon P. Anabolic-androgenic steroid use disorders among a sample of Australian competitive and recreational users. Drug Alcohol Depend 2000 Jul 1;60(1):91-96. [doi: 10.1016/s0376-8716(99)00141-6] [Medline: 10821993]
16. Sharp G, Tiggemann M, Mattiske J. The role of media and peer influences in Australian women's attitudes towards cosmetic surgery. Body Image 2014 Sep;11(4):482-487. [doi: 10.1016/j.bodyim.2014.07.009] [Medline: 25129686]
17. Perniciaro K, Moore MJ, Zeglin RJ, Terrell KR. Body dissatisfaction and health risk behaviors among middle school girls. J Sch Health 2024 May;94(5):453-461. [doi: 10.1111/josh.13440] [Medline: 38287684]
18. Kim DS. Body image dissatisfaction as an important contributor to suicidal ideation in Korean adolescents: gender difference and mediation of parent and peer relationships. J Psychosom Res 2009 Apr;66(4):297-303. [doi: 10.1016/j.jpsychores.2008.08.005] [Medline: 19302886]
19. Amante DJ, Hogan TP, Pagoto SL, English TM, Lapane KL. Access to care and use of the Internet to search for health information: results from the US National Health Interview Survey. J Med Internet Res 2015 Apr 29;17(4):e106. [doi: 10.2196/jmir.4126] [Medline: 25925943]
20. Mohamed F, Shoufan A. Users' experience with health-related content on YouTube: an exploratory study. BMC Public Health 2024 Jan 3;24(1):86. [doi: 10.1186/s12889-023-17585-5] [Medline: 38172765]
21. Haluza D, Naszay M, Stockinger A, Jungwirth D. Digital natives versus digital immigrants: influence of online health information seeking on the doctor-patient relationship. Health Commun 2017 Nov;32(11):1342-1349. [doi: 10.1080/10410236.2016.1220044] [Medline: 27710132]
22. Kara M, Ozduran E, Mercan Kara M, Hanci V, Erkin Y. Assessing the quality and reliability of YouTube videos as a source of information on inflammatory back pain. PeerJ 2024;12:e17215. [doi: 10.7717/peerj.17215] [Medline: 38618560]
23. Haberal HB, Piana A, Pecoraro A, et al. Decoding YouTube: an in-depth analysis of living donor kidney transplantation videos and their implications for patient education. Eur Urol Open Sci 2024 Dec;70:64-69. [doi: 10.1016/j.euros.2024.10.005] [Medline: 39497749]
24. Andika R, Kao CT, Williams C, Lee YJ, Al-Battah H, Alweis R. YouTube as a source of information on the COVID-19 pandemic. J Community Hosp Intern Med Perspect 2021 Jan 26;11(1):39-41. [doi: 10.1080/20009666.2020.1837412] [Medline: 33552412]
25. Morahan-Martin JM. How internet users find, evaluate, and use online health information: a cross-cultural review. Cyberpsychol Behav 2004 Oct;7(5):497-510. [doi: 10.1089/cpb.2004.7.497] [Medline: 15667044]
26. Bernard A, Langille M, Hughes S, Rose C, Leddin D, Veldhuyzen van Zanten S. A systematic review of patient inflammatory bowel disease information resources on the World Wide Web. Am J Gastroenterol 2007 Sep;102(9):2070-2077. [doi: 10.1111/j.1572-0241.2007.01325.x] [Medline: 17511753]
27. Singh AG, Singh S, Singh PP. YouTube for information on rheumatoid arthritis--a wakeup call? J Rheumatol 2012 May;39(5):899-903. [doi: 10.3899/jrheum.111114] [Medline: 22467934]

XSL•FO
RenderX

28. Silberg WM, Lundberg GD, Musacchio RA. Assessing, controlling, and assuring the quality of medical information on the Internet: Caveant lector et viewor—let the reader and viewer beware. JAMA 1997 Apr 16;277(15):1244-1245. [Medline: 9103351]
29. Liu Z, Chen Y, Lin Y, et al. YouTube/ Bilibili/ TikTok videos as sources of medical information on laryngeal carcinoma: cross-sectional content analysis study. BMC Public Health 2024;24(1):1594. [doi: 10.1186/s12889-024-19077-6]
30. Çoşkun N, Demir E. Analysis of YouTube videos on circumcision: evaluating reliability and quality for patients and parents. Eur J Ther 2024;30(5):626-637 [FREE Full text] [doi: 10.58600/eurjther2250]
31. Bujnowska-Fedak MM. Trends in the use of the Internet for health purposes in Poland. BMC Public Health 2015 Feb 27;15(1):194. [doi: 10.1186/s12889-015-1473-3] [Medline: 25886280]
32. Du RC, Zhang Y, Wang MH, Lu NH, Hu Y. TikTok and Bilibili as sources of information on Helicobacter pylori in China: a content and quality analysis. Helicobacter 2023 Oct;28(5):e13007. [doi: 10.1111/hel.13007] [Medline: 37452727]
33. Green MC, Brock TC. The role of transportation in the persuasiveness of public narratives. J Pers Soc Psychol 2000 Nov;79(5):701-721. [doi: 10.1037//0022-3514.79.5.701] [Medline: 11079236]
34. Henry SG, Feng B, Verba S, Kravitz RL, Iosif AM. The story vs the storyteller: factors associated with the effectiveness of brief video-recorded patient stories for promoting opioid tapering. Health Expect 2021 Jun;24(3):991-999. [doi: 10.1111/hex.13243] [Medline: 33835644]
35. Metzger MJ, Flanagin AJ. Credibility and trust of information in online environments: the use of cognitive heuristics. J Pragmat 2013 Dec;59:210-220. [doi: 10.1016/j.pragma.2013.07.012]
36. Song S, Xue X, Zhao YC, Li J, Zhu Q, Zhao M. Short-video apps as a health information source for chronic obstructive pulmonary disease: information quality assessment of TikTok videos. J Med Internet Res 2021 Dec 20;23(12):e28318. [doi: 10.2196/28318] [Medline: 34931996]
37. Liu H, Peng J, Li L, et al. Assessment of the reliability and quality of breast cancer related videos on TikTok and Bilibili: cross-sectional study in China. Front Public Health 2023;11:1296386. [doi: 10.3389/fpubh.2023.1296386]
38. Zheng S, Tong X, Wan D, Hu C, Hu Q, Ke Q. Quality and reliability of liver cancer-related short Chinese videos on TikTok and Bilibili: cross-sectional content analysis study. J Med Internet Res 2023 Jul 5;25:e47210. [doi: 10.2196/47210] [Medline: 37405825]
39. Ergenç M, Abdullayev R. YouTube as a source of education on tunneled catheter insertion: content and quality analysis. BMC Med Educ 2024 Nov 15;24(1):1318. [doi: 10.1186/s12909-024-06330-0] [Medline: 39548462]

Abbreviations

GQS: Global Quality Scale
ICC: intraclass correlation coefficient
mDISCERN: modified DISCERN
mJAMA: Modified Journal of the American Medical Association

Edited by J Sarvestan; submitted 23.01.25; peer-reviewed by E Demir; revised version received 01.07.25; accepted 04.07.25; published 02.09.25.

Please cite as:
Liu L, Yang J, Tan F, Luo H, Chen Y, Zhao X
Web-Based Video Platforms as Sources of Information on Body Image Dissatisfaction in Adolescents: Content and Quality Analysis of a Cross-Sectional Study
JMIR Form Res 2025;9:e71652
URL: https://formative.jmir.org/2025/1/e71652
doi:10.2196/71652

© Li Liu, Jianning Yang, Fengmei Tan, Huan Luo, Yanhua Chen, Xiaolei Zhao. Originally published in JMIR Formative Research (https://formative.jmir.org), 2.9.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Evaluating ChatGPT's Utility in Biologic Therapy for Systemic Lupus Erythematosus: Comparative Study of ChatGPT and Google Web Search

Kai Li[1], PhD; Yunfei Peng[1*], BA; Luyi Li[1*], BA; Bo Liu[2], MM; Zhijian Huang[2], MM

[1]School of Journalism and Communication, Guangxi University, No.100, Daxue East Road, Nanning, China

[2]Department of Rheumatology and Immunology, The Second Affiliated Hospital of Guangxi Medical University, Nanning, China

[*]these authors contributed equally

Corresponding Author:
Kai Li, PhD
School of Journalism and Communication, Guangxi University, No.100, Daxue East Road, Nanning, China

Abstract

Background: Systemic lupus erythematosus (SLE) is a life-threatening, multisystem autoimmune disease. Biologic therapy is a promising treatment for SLE. However, public understanding of this therapy is still insufficient, and the quality of related information on the internet varies, which affects patients' acceptance of this treatment. The effectiveness of artificial intelligence technologies, such as ChatGPT (OpenAI), in knowledge dissemination within the health care field has attracted significant attention. Research on ChatGPT's utility in answering questions regarding biologic therapy for SLE could promote the dissemination of this treatment.

Objective: This study aimed to evaluate ChatGPT's utility as a tool for users to obtain health information about biologic therapy for SLE.

Methods: This study extracted 20 common questions related to biologic therapy for SLE, their corresponding answers, and the sources of these answers from both Google Web Search and ChatGPT-4o (OpenAI). Then, based on Rothwell's classification, the questions were categorized into 3 main types: fact, policy, and value. The sources of the answers were classified into 5 categories: commercial, academic, medical practice, government, and social media. The accuracy and completeness of the answers were assessed using Likert scales. The readability of the answers was evaluated using the Flesch Reading Ease and Flesch-Kincaid Grade Level (FKGL) scores.

Results: The study found that, in terms of question types, ChatGPT-4o had the highest proportion of fact questions (10/20), followed by policy (7/20) and value (3/20). Google Web Search had the highest proportion of fact questions (12/20), followed by value (5/20) and policy (3/20). In terms of website sources, ChatGPT-4o's answers were sourced from 48 sources, with the majority coming from academic sources (29/48). Google Web Search provided answers from 20 sources, with an even distribution across all 5 categories. For accuracy, ChatGPT-4o's mean score of 5.83 (SD 0.49) was higher than that of Google Web Search (mean 4.75, SD 0.94), with a mean difference of 1.08 (95% CI 0.61 - 1.54). For completeness, ChatGPT-4o's mean score of 2.88 (SD 0.32) was higher than that of Google Web Search (mean 1.68, SD 0.69), with a mean difference of 1.2 (95% CI 0.96 - 1.44). For readability, the Flesch Reading Ease and Flesch-Kincaid Grade Level scores for ChatGPT-4o and Google Web Search were 11.7 and 14.9, and 16.2 and 20, respectively, indicating that both texts were of high reading difficulty, requiring readers to have a college graduate–level reading proficiency. When asking ChatGPT to respond at a sixth-grade level, the readability of the answers significantly improved.

Conclusion: ChatGPT's answers are characterized by accuracy, rigor, comprehensiveness, and professional supporting materials, and demonstrate humanistic care. However, the readability of the provided text is low, requiring users to have a college education background. Given the study's limitations in question scope, comparison dimensions, research perspectives, and language types, further in-depth comparative research is recommended.

(JMIR Form Res 2025;9:e76458) doi:10.2196/76458

KEYWORDS

biologic therapy; SLE; ChatGPT; google web search; health information; systemic lupus erythematosus

XSL•FO
RenderX

Introduction

ChatGPT (OpenAI) has gained significant attention for its robust capabilities in content generation since its launch. It can respond to user inquiries by providing relevant, accurate, and evidence-based information in a logically coherent and easily understandable manner [1]. ChatGPT-4o (OpenAI) and its subsequent versions have more natural human-computer interaction capabilities, enabling them to respond to user commands more efficiently [2]. Although ChatGPT was originally developed neither for the health care domain [3] nor explicitly for answering medical questions [4], its content generation potential in the health care field is particularly noteworthy. Studies have found that ChatGPT plays a positive role in helping users gain health knowledge and answering medical inquiries [5,6].

Ayik et al [7] and Javaid et al [8] have indicated that ChatGPT can assist users in answering common questions in the health care domain. Temsah et al [9] noted that ChatGPT offers high-quality health care information, thereby bridging the gap in health care access. According to Ruksakulpiwat et al [10], the medical information provided by ChatGPT can facilitate communication between patients and health care providers, enhance patients' self-care abilities, and improve treatment outcomes. Ayoub et al [5] are optimistic about ChatGPT's potential for knowledge acquisition in the health care domain, suggesting that in certain situations, ChatGPT could enhance or even replace the commonly used Google Web Search, becoming an important information source for patients.

Systemic lupus erythematosus (SLE) is a multisystem autoimmune disease with a prevalence that varies across populations, with a global average prevalence ranging from 12 to 39 per 100,000 individuals [11]. China ranks among the highest in the world in terms of the burden of SLE [12], with a prevalence ranging from 30.13 to 70.41 per 100,000 individuals. The disease is most commonly observed in women of reproductive age, particularly those between 20 and 40 years old [11]. SLE leads to severe organ complications and may even be life-threatening [13], requiring urgent medical attention. In recent years, with in-depth research on the pathogenesis of SLE, targeted therapy has been proposed and has gradually gained attention. Some new biologics have been developed and introduced into clinical use. According to the 2023 annual "Patient Chart Dynamix: SLE (US)" and "Patient Chart Dynamix: SLE (EU5)" published by Spherix Global Insights, an increasing number of patients with SLE in the United States and 5 European countries are receiving biologic therapy, with the proportion of patients receiving such treatment rising by 32% compared with 2022 [14]. Currently, belimumab and anifrolumab have been successively approved by the Food and Drug Administration for the treatment of SLE [15]. In China, belimumab, telitacicept, and rituximab are the 3 biologic agents currently used in clinical practice [16].

Although biologic agents play a crucial role in SLE treatment, Liu et al [17] found that many patients with SLE lack knowledge about treatment options and disease management. This not only significantly affects their attitudes, treatment practices, and willingness to choose biologic therapy but also hinders their active participation in disease management, ultimately impacting their prognosis [17]. A study by Waldron et al [18] found that when health care providers fail to provide clear and consistent information during diagnosis, patients with SLE experience distress and confusion. In such situations, patients with SLE are highly likely to use the internet to meet their information needs [19]. According to the data from the Pew Research Center, 77% of health information seekers reported that their search began with applications such as Google, Bing, or Yahoo [20]. Jacquemart and Zweigenbaum [21] found that Google is the most effective search engine for retrieving answers to clinical questions. Lopes and Ribeiro [22] also found that Google significantly outperforms other search engines in accessing health information. Nowadays, with the emergence of large language models (LLMs), such as ChatGPT, some users are shifting from traditional web searches to these models as a primary source of health care information [5,6,23]. Traditional search engines, such as Google Bard and Microsoft Copilot, have also integrated LLMs into their platforms [24]. Hopkins et al [25] believed that LLMs (eg, ChatGPT) will fundamentally transform the way patients seek health care information.

The readability of health care texts directly impacts how effectively readers acquire health information from them. According to the readability assessment standards set by the National Institutes of Health and the American Medical Association, presenting health information on websites at a sixth-grade reading level or lower may help improve patients' understanding of their condition and the treatment plans recommended by health care providers [26]. A study by Daraz et al [27] in the United States and Canada found that the vast majority of health information on the internet has poor readability, making it unsuitable for the general public. Furthermore, it may lead to misinformation and have a detrimental effect on health. Massie et al [28] noted that the readability of the American Society of Metabolic Surgery's Patient Information Publications is poor and does not meet the standard of reading levels at or below the sixth grade. However, a study by Moons and Van Bulck [29] found that ChatGPT is a valuable tool for improving the readability and comprehensibility of health information. It can generate health care information suitable for the general public and easy to understand, based on user instructions, thereby addressing the issue of poor readability in health information on the internet [29].

Previous studies have evaluated the effectiveness of ChatGPT-3.5 (OpenAI) in responding to SLE-related questions and found that, in response to certain questions, its answers were not comprehensive or accurate and were sometimes unrelated to the questions [30]. ChatGPT-4o is widely recognized as a relatively new artificial intelligence model with a high user adoption rate [31]. However, no studies have yet evaluated the effectiveness of the new version, ChatGPT-4o, in responding to common SLE-related questions. Biologic therapy is a relatively new and important treatment for SLE and is regarded with high expectations. However, public awareness of this therapeutic approach remains insufficient [17]. Furthermore, the quality of SLE-related information on the

XSL•FO
RenderX

internet varies, with some content containing advertisements and misleading information. This significantly hinders patients' acceptance of biologic therapy [32], ultimately causing a significant negative impact on the clinical outcomes of patients with SLE. Therefore, this study aims to replicate users' search behavior on Google Web Search to evaluate the utility of ChatGPT-4o in responding to common questions about biologic therapy for SLE. Specifically, we compare the following three aspects: (1) the most common questions related to biologic therapy for SLE, categorized by question types and topics; (2) the answers to these questions; and (3) the sources of these answers.

Methods

Overview

This study extracted 20 common questions related to biologic therapy for SLE, their corresponding answers, and the sources of these answers from both Google Web Search and ChatGPT-4o [33]. First, this study used the "People also ask" function in Google Web Search by entering the search terms "biologic therapy" and "systemic lupus erythematosus." We selected the 20 most common questions about biologic therapy for SLE and recorded their answers and sources. Second, we instructed ChatGPT-4o to execute the command "Do a Google search with the search term 'biologic therapy,' 'systemic lupus erythematosus,' and record 20 common questions related to the search term." Similarly, we recorded the questions, their answers, and the sources of those answers. Third, the obtained questions and the sources of answers were classified. Fourth, the accuracy, completeness, and readability of the answers were evaluated. In total, 2 rheumatology physicians (BL and ZH) were involved throughout the process of question selection and answer evaluation. In the event of a disagreement, a third physician was included in the discussion to reach a consensus.

Google Web Search Operation

On July 17, 2024, a new installation of the Google Chrome browser was performed, and all browsing history, cookies, site data, cached images, files, sponsored websites, and advertisements were cleared before initiating the search. The keywords "biologic therapy" and "systemic lupus erythematosus" were entered simultaneously, and questions were identified using the "People Also Ask" function in Google Web Search. Additional related questions were obtained by clicking on each question.

We conducted 10 rounds of searches. During the selection of relevant questions, we found that once the sample size exceeded 30, the system started generating irrelevant or duplicate questions. Therefore, to ensure a representative sample, we selected only the first 30 questions from each round, yielding a total of 300 questions. After merging similar questions and analyzing their frequencies, 2 professional physicians (BL and ZH) assigned importance levels to each question according to their clinical experience. Finally, 20 questions related to biologic therapy for SLE were selected, along with their corresponding answers and the website information sources.

ChatGPT Operation Execution

The following command was entered into ChatGPT-4o: "Do a Google search with the search term 'biologic therapy,' 'systemic lupus erythematosus,' and record 30 common questions related to the search term." Then, each question was individually entered into ChatGPT-4o, and the corresponding answers and their sources were recorded. We conducted 10 rounds of searches, selecting 30 questions per round, which resulted in a total of 300 questions. We then used the same selection process for questions with Google Web Search, ultimately extracting 20 questions related to biologic therapy for SLE, along with their corresponding answers and website information sources.

Notably, unlike Google Web Search, ChatGPT-4o provided 2 or more sources for its answers. We used a 2-step verification process to assess whether hallucinations exist in its responses: First, each link attached to the answers was manually accessed and verified to check whether the link was valid and if the page content was consistent with the information provided in the answer. If the link was invalid or the page content was irrelevant to the answer, it was considered a potential indicator of hallucination. Second, based on the accuracy ratings given by 2 professional physicians (BL and ZH; refer to Accuracy and Completeness Evaluation section), if the average score of any answer was below 3, it was considered a hallucination, indicating that the answer contained factual errors or that the main information could not be substantiated [34]. No hallucination was found in the links or answers provided by ChatGPT after the 2-step verification. After removing duplicate links, we compiled statistics on the final website information sources.

Inclusion and Exclusion Criteria

This study invited 2 professional physicians to screen and extract questions that are highly relevant to the topic and of significant concern to patients based on their clinical experience. The selected questions and answers were required to be closely related to biologic therapy for SLE. The inclusion criteria involved any question related to the terms "biologic therapy," "biological therapy," "biologics therapy," "biologic treatment," "rituximab," "belimumab," "anifrolumab," and "systemic lupus erythematosus," "lupus," or "SLE." The exclusion criteria included duplicate questions and questions unrelated to biologic therapy for SLE (eg, "What is the enemy of lupus?").

Ethical Considerations

This study analyzes publicly available data from Google Web Search and ChatGPT-4o. No personal or identifiable patient data were used. According to the Declaration of Helsinki established by the World Medical Association [35], this study does not involve human participants or personal data. Therefore, it is exempt from institutional review board review.

Comparison Analysis

The common questions related to biologic therapy for SLE in this study were categorized using a modification of the Rothwell system [36]. According to this system, questions can be classified into 3 main types (fact, policy, and value), as well as 10 subtypes, that are education, accessibility, timeline of effectiveness, technical details, cost, indications or management,

XSL•FO
RenderX

risks or complications, future prospects and challenges, longevity, and evaluation of treatment (refer to Table 1 for details). The website information sources were categorized into 5 types—commercial, academic, medical practice, government, and social media [33,37] (refer to Table 2 for details). Furthermore, 2 rheumatology physicians (BL and ZH) conducted question classification. In the event of disagreements, a third physician was brought in to achieve a consensus through discussion.

Table . Question types categorized by Rothwell's classification [36]. The questions are categorized into 3 main types and subsequent subtypes.

Categorization	Explanation
Fact	Asks whether something is true and to what extent, objective information
Education	Educational information about biologic therapies for SLE[a], not fitting into other subcategories
Accessibility	Where and how patients can get biologic treatment for SLE
Timeline of effectiveness	Length of time for the biologic therapy to show effectiveness in SLE, rather than just recovery milestones
Technical details	Details of how biologic therapies work for SLE, including dosing, method of delivery, and so on
Cost	Cost of treatment or rehabilitation
Policy	Asks whether a specific course of action should be taken to solve a problem
Indications or management	Therapy indications and alternatives, timing of biologic therapy for SLE, and posttreatment monitoring and management
Risks and complications	Risks and complications before, during, or after biologic treatment, including the rehabilitation period for SLE
Future prospects and challenges	The progress and difficulty in the development of biologic therapy for SLE
Value	Asks for the evaluation of an idea, object, or event
Longevity	Longevity and sustained effect of biologic therapy for patients with SLE
Evaluation of treatment	Successfulness, seriousness, or invasiveness of biologic therapy for SLE

[a]SLE: systemic lupus erythematosus.

Table . Categorization of website information sources.

Categorization	Explanation
Commercial	Organizations that provide public health information, including medical device, manufacturing, or pharmaceutical companies and news outlets
Academic	Universities, academic medical centers, or academic societies
Medical practice	Local hospitals or medical groups without clear academic affiliation
Government	Websites maintained by a national government
Social media	Blog, internet forums, support groups, and nonmedical organizations designed for information and video sharing

Accuracy and Completeness Evaluation

The accuracy and completeness of responses generated by Google Web Search and ChatGPT-4o were evaluated by 2 professional physicians based on the guidelines published by the American College of Rheumatology [38]. Accuracy was assessed using a 6-point Likert scale (with 1=completely incorrect, 2=more incorrect than correct, 3=approximately equal correct and incorrect, 4=more correct than incorrect, 5=nearly entirely correct, and 6=completely correct). Completeness was evaluated using a 3-point Likert scale: 1 indicated "incomplete" (the response addressed some aspects of the question but omitted important information), 2 indicated "adequate" (the response addressed all aspects of the question with sufficient information for completeness), and 3 indicated "comprehensive" (the response addressed all aspects and provided additional information or context beyond expectations) [39]. Statistical analysis was performed using IBM SPSS Statistics 27, using descriptive statistics (mean and SD) and the Cohen κ test to assess the interrater reliability.

Readability Measurement

Readability is defined as the ease of understanding or comprehension attributed to the writing style. Although many different readability tests exist, their measurement basically relies on two elements: (1) vocabulary and (2) word or sentence complexity (ie, the number of characters per space, syllables per word, or words per sentence) [40]. This study used the

XSL•FO
RenderX

assessment tool "Readable" to evaluate the readability of the texts generated by Google Web Search and ChatGPT-4o [41]. Readability was assessed using the Flesch Reading Ease (FRE) and Flesch-Kincaid Grade Level (FKGL), both of which are commonly used to evaluate the readability of health-related literature on the internet [27,28]. The FRE score ranges from 0 to 100, with lower scores indicating higher reading difficulty [42]. The FKGL ranges from 0 to 18, representing the number of years of education required to comprehend the text. A lower score indicates easier readability [43].

Results

Question Analysis

The classification results for the 20 most common questions obtained from Google Web Search and ChatGPT-4o are presented in Table 3. Regarding the distribution of main question types, questions categorized under fact accounted for 12 from Google Web Search and 10 from ChatGPT-4o, those under policy for 3 and 7, and those under value for 5 and 3, respectively. Fact-category questions accounted for the highest proportion in both sources.

XSL•FO
RenderX

Table . 20 most common questions for biologic therapy for systemic lupus erythematosus from Google Web Search and ChatGPT-4o. The main types and subtypes of questions, categorized by Rothwell's classification [36], are shown in parentheses.

	20 most common questions from Google Web Search	20 most common questions from ChatGPT-4o
1	What is the safest biologic for lupus? (Value, evaluation of treatment)	What is biologic therapy for SLE[a]? (Fact, education)
2	What is the new biologic for lupus? (Fact, education)	How do biologic therapies work in treating SLE? (Fact, technical details)
3	Is rituximab good for lupus? (Value, evaluation of treatment)	What are the most common biologic therapies used for SLE? (Fact, education)
4	What is the new drug for lupus approved by the FDA[b]? (Fact, accessibility)	What are the potential side effects of biologic therapies for SLE? (Policy, risks, and complications)
5	Can biologics cause lupus? (Policy, risks, and complications)	Are there any new biologic therapies being developed for SLE? (Fact, future prospects, and challenges)
6	How long does it take rituximab to work for lupus? (Fact, timeline of Effectiveness)	How does biologic therapy compare to traditional treatments for SLE? (Value, evaluation of treatment)
7	How effective is rituximab for lupus?(Value, evaluation of treatment)	Can biologic therapies be used in combination with other SLE treatments? (Policy, indications, or management)
8	What is the biologic treatment for systemic lupus erythematosus? (Fact, education)	How is the effectiveness of biologic therapies in SLE measured in clinical trials? (Fact, technical details)
9	Is rituximab FDA approved for SLE? (Fact, accessibility)	What are the long-term outcomes for SLE patients on biologic therapies? (Value, longevity)
10	What is the new treatment for lupus in the UK? (Fact, accessibility)	What is the cost and insurance coverage situation for biologic therapies for SLE? (Fact, cost)
11	Can rituximab cure lupus? (Value, evaluation of treatment)	How long does it take to see improvements with biologic therapy in lupus? (Fact, timeline of effectiveness)
12	How does rituximab work for SLE? (Fact, technical details)	What are the criteria for a patient with SLE to be eligible for biologic therapy? (Policy, indications, or management)
13	Is rituximab used to treat SLE? (Policy, indications, or management)	How do biologic therapies impact the overall quality of life for SLE patients? (Value, evaluation of treatment)
14	How do biologics help lupus? (Fact, technical details)	How often are biologic therapies administered for SLE patients? (Fact, technical details)
15	What is the best infusion for lupus? (Value, evaluation of treatment)	What is the role of Benlysta (belimumab) in SLE treatment? (Fact, technical details)
16	What is the self-injection for lupus? (Fact, technical details)	Are there any dietary or lifestyle changes recommended for SLE patients on biologic therapies? (Policy, indications, or management)
17	How often is Rituxan given for lupus?(Fact, technical details)	What are the approved biologic therapies for SLE? (Fact, education)
18	What are the indications for rituximab in SLE? (Policy, indications or management)	How are patients monitored while on biologic treatments for SLE? (Policy, indications, or management)
19	When will anifrolumab be FDA approved?(Fact, accessibility)	Are there any specific biologic therapies recommended for lupus nephritis? (Policy, indications or management)
20	Can adalimumab cause lupus? (Fact, risks and complications)	Can biologic therapies be used during pregnancy in SLE patients? (Policy, indications or management)

XSL•FO
RenderX

[a]SLE: systemic lupus erythematosus.

[b]FDA: Food and Drug Administration.

The distribution of question subtypes is illustrated in Multimedia Appendix 1. Questions from Google Web Search fell into 7 subtypes, with the highest proportions in the evaluation of treatment (n=5), followed by accessibility (n=4) and technical details (n=4). Other subtypes included education (n=2), risks and complications (n=2), indications or management (n=2), and timeline of effectiveness (n=1). Questions generated by ChatGPT-4o covered 9 subtypes, with indications or management (n=6) having the highest proportion, followed by technical details (n=4), education (n=3), evaluation of treatment (n=2). The remaining subtypes, including timeline of effectiveness, cost, risks and complications, future prospects and challenges, and longevity, were each represented by 1 question.

Answer Analysis

Multimedia Appendix 2 presents the responses generated by ChatGPT-4o and Google Web Search. In terms of accuracy, the accuracy scores indicate that ChatGPT-4o achieved a higher mean score (mean 5.83, SD 0.49) compared with Google Web Search responses (mean 4.75, SD 0.94), with a mean difference of 1.08 (95% CI 0.61 - 1.54). The ratings provided by the 2 physicians demonstrated good interrater reliability, with Cohen κ values of 0.72 for ChatGPT-4o and 0.78 for Google Web Search.

For completeness, ChatGPT-4o's mean score was 2.88 (SD 0.32), significantly higher than Google Web Search's 1.68 (SD 0.69), with a mean difference of 1.2 (95% CI 0.96 - 1.44). The ratings also showed strong interrater reliability, with Cohen κ values of 0.94 for ChatGPT-4o and 0.88 for Google Web Search.

In terms of readability, the FRE and FKGL scores for Google Web Search and ChatGPT-4o were 16.2 and 20, and 11.7 and 14.9, respectively (Table 4). These results indicate that both texts have a high level of reading difficulty, requiring readers to have a college graduate-level reading proficiency [41]. In terms of response length, answers from Google Web Search were more concise, while those from ChatGPT-4o were more detailed and comprehensive.

Table . Accuracy, completeness, and readability of Google Web Search and ChatGPT-4o answers on biologic therapy for systemic lupus erythematosus.

	Item	Google Web Search	ChatGPT-4o
1	Accuracy, mean (SD)	5.83 (0.49)	4.75 (0.94)
2	Completeness, mean (SD)	2.88 (0.32)	1.68 (0.69)
3	Readability		
	FKGL[a]	16.2	11.7
	FRE[b]	20	17.3

[a]FKGL: Flesch-Kincaid Grade Level.

[b]FRE: Flesch Reading Ease.

Furthermore, 3 of the 20 questions obtained from each source were similar (Table 5), but their answers differed significantly. The answers provided by Google Web Search were not comprehensive enough, lacking explanations of drug mechanisms and missing information on new medications (eg, Saphnelo) and emerging therapies (eg, chimeric antigen receptor–T cell therapy).

Table . Similar questions on biologic therapy for systemic lupus erythematosus of Google web search and ChatGPT-4o.

Question	Google Web Search	ChatGPT-4o
1	What is the new biologic for lupus?	Are there any new biologic therapies being developed for SLE[a]?
2	What is the biologic treatment for systemic lupus erythematosus?	What is biologic therapy for SLE?
3	How do biologics help lupus?	How do biologic therapies work in treating SLE?

[a]SLE: systemic lupus erythematosus.

Analysis of Website Information Source

The distribution of website information sources differed notably between Google Web Search and ChatGPT-4o (Multimedia Appendix 3). Google Web Search provided answers from 20 sources: 2 commercial, 4 academic, 5 medical practice, 4 government, and 5 social media. ChatGPT-4o's answers were cited from 48 sources, including 9 commercial, 29 academic, 1 medical practice, 2 government, and 7 social media sources. Notably, ChatGPT-4o relied more heavily on academic sources, whereas the source types of Google Web Search were more evenly distributed.

XSL•FO
RenderX

Discussion

Principal Findings

This study aimed to evaluate the utility of ChatGPT in providing users with health information on biologic therapy for SLE. The key findings are described further.

Accuracy and Rigor of Questions and Answers Provided by ChatGPT

On the one hand, the study found that ChatGPT-4o provided answers with higher accuracy than Google Web Search, which is consistent with the findings of Cohen et al [44]. On the other hand, among the 20 questions provided by Google Web Search, 9 were related to rituximab treatment for SLE, covering topics such as its accessibility, technical details, indications, and treatment evaluation. In contrast, the questions provided by ChatGPT-4o did not focus on rituximab treatment for SLE. There are significant differences between the questions regarding biologic therapy for SLE, which are of common concern to users, presented by both platforms. This finding is consistent with the findings of Dubin et al [33]. The reasons could be that Google provides an effective marketing channel for some pharmaceutical and health websites, enabling them to gain more exposure through bidding. Sponsored links are prioritized and displayed at the top or in prominent positions on the search results page [45-47]. However, many of these links may not only lack accurate clinical evidence but may also contain concerning misinformation [48]. In addition, search engine optimization techniques, through the optimization of website elements and the building of high-quality interlinks, effectively enhance a website's ranking and visibility in search engines, making it a key channel for increasing organic traffic [49]. However, this could also mean that high-quality, unoptimized medical content pages may be pushed down in rankings, replaced by lower-quality pages [50]. ChatGPT-4o did not generate questions related to rituximab treatment for SLE. This may be because although textbooks such as "Harrison's Principles of Internal Medicine" [51] and clinical guidelines such as "American College of Rheumatology Guidelines for Screening" [38] recommend the use of rituximab for treating SLE and its clinical application is widespread, rituximab has never been officially approved for SLE treatment at the legal level, nor is SLE listed as an indication in the drug's prescribing information. The "Usage Policies" published by OpenAI under "Building with the OpenAI API Platform" explicitly state that ChatGPT-4o must not provide tailored medical or health advice without review by a qualified professional and without disclosing the use of artificial intelligence assistance and its potential limitations, as such actions may significantly impair the safety, well-being, or rights of others [52].

Comprehensive and Professional Nature of the Supporting Materials for the Answers Provided by ChatGPT

The study found that ChatGPT-4o provided answers with higher completeness than Google Web Search and provided information drawn from professional information sources, such as peer-reviewed journals, open-access academic publications from IntechOpen, and clinical practice guidelines from Kidney Disease: Improving Global Outcomes. This finding is consistent with the findings of Oeding et al [53]. This may be because ChatGPT's training data primarily consists of a large volume of journal articles, books, reports, and other resources related to biomedical and health sciences, sourced from databases such as PubMed [54-56]. When responding to user queries, ChatGPT can quickly retrieve and integrate relevant information from its knowledge base. Based on this, compared with the previous study by Chen et al [57], we believe that ChatGPT can provide the general public with more professional answers regarding biologic therapy for SLE and help users better filter out false and misleading information from the internet.

Aspects of Humanistic Care in ChatGPT's Answers

Based on content evaluation by 2 physicians, the study found that in ChatGPT-4o's responses to 20 questions about biologic therapy for SLE, 50% reflected aspects of humanistic care, such as medication risk warnings, expressions of concern for patients' physical and mental well-being, reminders about lifestyle adjustments, and considerations of medication affordability. This finding aligns with the research of Sorin et al [58] and Howe et al [59]. When providing answers regarding drug dosage and patient self-management of diseases, ChatGPT-4o includes recommendations at the end to seek guidance from health care providers. This reflects the rigor with which ChatGPT-4o delivers information to users. It also highlights the advantage of ChatGPT-4o, based on natural language processing technology, in effectively simulating human language and interaction patterns [60,61]. Compared with information on websites, which may contain logical inconsistencies or unclear statements, the way in which ChatGPT presents information is likely more accessible and comprehensible to the public. This helps to reduce the likelihood of patients being misled by vague or exaggerated information on advertising websites [57].

Low Readability of ChatGPT's Answers

The study found that the texts generated by ChatGPT-4o regarding biologic therapy for SLE follow a general-specific-general structure, which aligns with human reading habits. However, the content's readability was low. Based on the FRE and FKGL [27,28], the content provided by ChatGPT-4o regarding biologic therapy for SLE required a reading level equivalent to that of a college graduate [62], which is consistent with the findings of Ghanem et al [41]. The FRE score of Google Web Search was higher than that of ChatGPT-4o, which is consistent with the findings of Shen et al [63]. Possible reasons are as follows. First, ChatGPT was initially designed for a reading level comparable to that of high school students and above [29]. Second, ChatGPT's training data and cited sources include high-quality medical literature, ensuring the accuracy of the medical terminology used in its responses. However, the use of medical jargon may reduce the readability of the text [64]. Third, ChatGPT generates responses based on patterns learned from large-scale text data, which typically requires readers to have a reading proficiency above the sixth-grade level [29]. However, when given explicit instructions to respond at a sixth-grade reading level, ChatGPT's terminology density decreases, resulting in a significant

XSL•FO
RenderX

improvement in the text's readability. Specifically, the Flesch Kincaid Grade Level score decreased from 11.7 to 10.5, while the FRE score increased substantially from 17.3 to 51.9, indicating a marked enhancement in readability. This finding is consistent with that of Shen et al [63] and Mohammadi et al [65]. This capability enables ChatGPT to provide responses tailored to different educational levels, which may have significant implications for improving health care information accessibility and reducing health care disparities [66].

Implications

First, this study compares ChatGPT-4o and Google Web Search in providing information on biologic therapy for SLE, offering a reference for the future dissemination of SLE treatment knowledge based on LLMs.

Second, this study conducts an in-depth analysis of ChatGPT-4o's utility in providing information on biologic therapy for SLE, deepening understanding in both academia and industry regarding the potential of LLMs in delivering medical information and treatment recommendations.

Third, this study provides a reference for how users can use LLMs to access information on biologic therapy for SLE. It also offers insights into the effective use of LLMs for obtaining medical information to meet health care information needs.

Limitations and Future Prospects

A few limitations should be noted in this study.

First, this study primarily focuses on common questions regarding biologic therapy for SLE, whereas in practice, users have a much broader range of concerns. Future studies can expand the scope of evaluated questions to gain a deeper understanding of ChatGPT's utility in providing information on biologic therapy for SLE.

Second, this study only compares the basic characteristics of questions provided by Google and ChatGPT, including question types, answer sources, the accuracy, completeness, and readability of answers. Future research could broaden the comparison dimensions by including more factors.

Third, this study lacks a comparative analysis of ChatGPT-4o's response characteristics and utility across different language contexts. Future research could focus on comparative studies across multiple languages.

Fourth, this study focuses on general internet users and does not account for differences across user groups. Future research could investigate the applicability and effectiveness of ChatGPT in meeting the information needs of various user groups.

Fifth, this study primarily focuses on evaluating information from a physician's perspective, without reflecting patient or user concerns and experiences. Future research could explore the effectiveness of ChatGPT in addressing the information needs of users or patients through surveys or interviews, providing a deeper understanding of how well ChatGPT-generated content meets these needs.

Conclusion

By comparing the utility of ChatGPT-4o and Google Web Search in providing information on biologic therapy for SLE, this study finds that ChatGPT has the potential to be more effective in delivering accurate medical knowledge and supporting patient education. However, given the limitations of this study in terms of question scope, comparison dimensions, language types, and research perspectives, further in-depth comparative research is recommended.

Acknowledgments

We appreciate the valuable comments provided by all reviewers. This study used the generative artificial intelligence tool of ChatGPT by OpenAI to generate common questions and answers related to biologic therapy for systemic lupus erythematosus. The method and process used to generate the responses are described in the relevant sections of the study.

Data Availability

The datasets are available from the corresponding author upon reasonable request.

Authors' Contributions

KL and BL are co-corresponding authors of this paper. Conceptualization was contributed by KL and BL. Methodology was contributed by KL and YP. Resources was handled by KL. Supervision was managed by KL and BL. Validation was contributed by KL, BL, and ZH. Investigation and data curation was handled by YP and LL. Formal analysis was contributed by YP, LL, BL, and ZH. VisualizationVisualization was contributed by LL. Writing—original draft was contributed by KL, YP, and LL. Writing—review and editing was contributed by KL, BL, YP, and LL.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Subtype distribution of 20 most common questions on biologic therapy for systemic lupus erythematosus from Google Web Search and ChatGPT-4o.

XSL•FO
RenderX

[PNG File, 6 KB - formative_v9i1e76458_app1.png]

Multimedia Appendix 2
The responses of ChatGPT-4o and Google Web Search.
[XLSX File, 52 KB - formative_v9i1e76458_app2.xlsx]

Multimedia Appendix 3
tableDistribution of website information sources for Google Web Search and ChatGPT-4o answers on biologic therapy for systemic lupus erythematosus.
[PNG File, 193 KB - formative_v9i1e76458_app3.png]

References

1. Cox A, Seth I, Xie Y, Hunter-Smith DJ, Rozen WM. Utilizing ChatGPT-4 for providing medical information on blepharoplasties to patients. Aesthet Surg J 2023 Jul 15;43(8):NP658-NP662. [doi: 10.1093/asj/sjad096] [Medline: 37032521]
2. Hello GPT-4o. OpenAI. 2024. URL: https://openai.com/index/hello-gpt-4o/ [accessed 2024-06-13]
3. Choudhury A, Shamszare H. Investigating the impact of user trust on the adoption and use of ChatGPT: survey analysis. J Med Internet Res 2023 Jun 14;25:e47184. [doi: 10.2196/47184] [Medline: 37314848]
4. Liu JL, Wang CY, Liu SR. Utility of ChatGPT in clinical practice. J Med Internet Res 2023 Jun;25:e48568. [doi: 10.2196/48568]
5. Ayoub NF, Lee YJ, Grimm D, Divi V. Head-to-head comparison of ChatGPT versus Google Search for medical knowledge acquisition. Otolaryngol Head Neck Surg 2024 Jun;170(6):1484-1491. [doi: 10.1002/ohn.465] [Medline: 37529853]
6. Fernández-Pichel M, Pichel JC, Losada DE. Evaluating search engines and large language models for answering health questions. NPJ Digit Med 2025 Mar 10;8(1):153. [doi: 10.1038/s41746-025-01546-w] [Medline: 40065094]
7. Ayık G, Ercan N, Demirtaş Y, Yıldırım T, Çakmak G. Evaluation of ChatGPT-4o's answers to questions about hip arthroscopy from the patient perspective. Jt Dis Relat Surg 2025 Jan 2;36(1):193-199. [doi: 10.52312/jdrs.2025.1961] [Medline: 39719917]
8. Javaid M, Haleem A, Singh RP. ChatGPT for healthcare services: an emerging stage for an innovative perspective. BenchCouncil Transactions on Benchmarks, Standards and Evaluations 2023 Feb;3(1):100105. [doi: 10.1016/j.tbench.2023.100105]
9. Temsah MH, Jamal A, Alhasan K, et al. Transforming virtual healthcare: the potentials of ChatGPT-4omni in telemedicine. Cureus 2024 May;16(5):e61377. [doi: 10.7759/cureus.61377] [Medline: 38817799]
10. Ruksakulpiwat S, Kumar A, Ajibade A. Using ChatGPT in medical research: current status and future directions. J Multidiscip Healthc 2023;16:1513-1520. [doi: 10.2147/JMDH.S413470] [Medline: 37274428]
11. Zhong N, Lu Z, Ge J, et al. Internal Medicine: People's Medical Publishing House; 2024.
12. Tian J, Zhang D, Yao X, Huang Y, Lu Q. Global epidemiology of systemic lupus erythematosus: a comprehensive systematic analysis and modelling study. Ann Rheum Dis 2023 Mar;82(3):351-356. [doi: 10.1136/ard-2022-223035] [Medline: 36241363]
13. Carter EE, Barr SG, Clarke AE. The global burden of SLE: prevalence, health disparities and socioeconomic impact. Nat Rev Rheumatol 2016 Oct;12(10):605-620. [doi: 10.1038/nrrheum.2016.137] [Medline: 27558659]
14. In a Push to Limit Use of Glucocorticoids, US and EU5 Rheumatologists are Increasingly Turning to GSK's Benlysta and AstraZeneca's Saphnelo for the Treatment of Moderate-to-Severe Systemic Lupus Erythematosus, According to Spherix Global In. Insights SG. URL: https://www.globenewswire.com/news-release/2023/10/04/2754325/0/en/In-a-Push-to-Limit-Use-of-Glucocorticoids-US-and-EU5-Rheumatologists-are-Increasingly-Turning-to-GSK-s-Benlysta-and-AstraZeneca-s-Saphnelo-for-the-Treatment-of-Moderate-to-Severe-S.html [accessed 2023-06-15]
15. Kirou KA, Dall Era M, Aranow C, Anders HJ. Belimumab or anifrolumab for systemic lupus erythematosus? A risk-benefit assessment. Front Immunol 2022;13:980079. [doi: 10.3389/fimmu.2022.980079] [Medline: 36119023]
16. Geng Y, Wu L, Xie Q, et al. 2024 Chinese expert consensus on the use of biologic therapy for systemic lupus erythematosus. Chinese Journal of Rheumatology 2024;28(2):78-92.
17. Liu X, Song Y, Wan L, Du R. Knowledge, attitudes, and practices among patients with systemic lupus erythematosus toward disease management and biologic therapy. J Multidiscip Healthc 2024;17:937-947. [doi: 10.2147/JMDH.S444619] [Medline: 38455274]
18. Waldron N, Brown S, Hewlett S, Elliott B, McHugh N, McCabe C. "It's more scary not to know": a qualitative study exploring the information needs of patients with systemic lupus erythematosus at the time of diagnosis. Musculoskeletal Care 2011 Dec;9(4):228-238. [doi: 10.1002/msc.221] [Medline: 21993962]
19. Ichikawa T, Kishida D, Shimojima Y, et al. Impact of online health information-seeking behavior on shared decision-making in patients with systemic lupus erythematosus: The TRUMP2-SLE project. Lupus (Los Angel) 2023 Oct;32(11):1258-1266. [doi: 10.1177/09612033231200104] [Medline: 37651318]

XSL•FO
RenderX

20. Majority of Adults Look Online for Health Information. Pew Research Center. 2013. URL: https://www.pewresearch.org/short-reads/2013/02/01/majority-of-adults-look-online-for-health-information/ [accessed 2024-06-15]
21. Jacquemart P, Zweigenbaum P. Towards a medical question-answering system: a feasibility study. Stud Health Technol Inform 2003;95:463-468. [Medline: 14664030]
22. Teixeira Lopes C, Ribeiro C. Comparative evaluation of web search engines in health information retrieval. Online Information Review 2011 Nov 29;35(6):869-892. [doi: 10.1108/14684521111193175]
23. Huang Y, Huang JX. Exploring ChatGPT for next-generation information retrieval: opportunities and challenges. WEB 2024;22(1):31-44. [doi: 10.3233/WEB-230363]
24. Giannakopoulos K, Kavadella A, Aaqel Salim A, Stamatopoulos V, Kaklamanos EG. Evaluation of the performance of generative AI large language models ChatGPT, Google Bard, and Microsoft Bing Chat in supporting evidence-based dentistry: comparative mixed methods study. J Med Internet Res 2023 Dec 28;25:e51580. [doi: 10.2196/51580] [Medline: 38009003]
25. Hopkins AM, Logan JM, Kichenadasse G, Sorich MJ. Artificial intelligence chatbots will revolutionize how cancer patients access information: ChatGPT represents a paradigm-shift. JNCI Cancer Spectr 2023 Aug 31;7(2):pkad010. [doi: 10.1093/jncics/pkad010] [Medline: 36808255]
26. Eltorai AEM, Han A, Truntzer J, Daniels AH. Readability of patient education materials on the American Orthopaedic Society for Sports Medicine website. Phys Sportsmed 2014 Nov;42(4):125-130. [doi: 10.3810/psm.2014.11.2099] [Medline: 25419896]
27. Daraz L, Morrow AS, Ponce OJ, et al. Readability of online health information: a meta-narrative systematic review. Am J Med Qual 2018;33(5):487-492. [doi: 10.1177/1062860617751639] [Medline: 29345143]
28. Massie PL, Arshad SA, Auyang ED. Readability of American Society of Metabolic Surgery's patient information publications. J Surg Res 2024 Jan;293:727-732. [doi: 10.1016/j.jss.2023.09.018] [Medline: 37862852]
29. Moons P, Van Bulck L. Using ChatGPT and Google Bard to improve the readability of written patient information: a proof of concept. Eur J Cardiovasc Nurs 2024 Mar 12;23(2):122-126. [doi: 10.1093/eurjcn/zvad087] [Medline: 37603843]
30. Huang C, Hong D, Chen L, Chen X. Assess the precision of ChatGPT's responses regarding systemic lupus erythematosus (SLE) inquiries. Skin Res Technol 2023 Oct;29(10):e13500. [doi: 10.1111/srt.13500] [Medline: 37881050]
31. How can i access GPT-4, GPT-4 turbo, GPT-4o, and GPT-4o mini? OpenAI. 2024. URL: https://help.openai.com/en/articles/7102672-how-can-i-access-gpt-4-gpt-4-turbo-gpt-4o-and-gpt-4o-mini?utm_medium=email&utm_source=transaction [accessed 2024-06-15]
32. O'Mathúna DP. How should clinicians engage with online health information? AMA J Ethics 2018 Nov 1;20(11):E1059-E1066. [doi: 10.1001/amajethics.2018.1059] [Medline: 30499435]
33. Dubin JA, Bains SS, Chen Z, et al. Using a Google Web Search analysis to assess the utility of ChatGPT in total joint arthroplasty. J Arthroplasty 2023 Jul;38(7):1195-1202. [doi: 10.1016/j.arth.2023.04.007] [Medline: 37040823]
34. Chelli M, Descamps J, Lavoué V, et al. Hallucination rates and reference accuracy of ChatGPT and Bard for systematic reviews: comparative analysis. J Med Internet Res 2024 May 22;26:e53164. [doi: 10.2196/53164] [Medline: 38776130]
35. WMA declaration of helsinki - ethical principles for medical research involving human participants. World Medical Association. 2025. URL: https://www.wma.net/policies-post/wma-declaration-of-helsinki/ [accessed 2025-04-11]
36. Rothwell JD. Mixed Company: Communication in Small Groups: Wadsworth Cengage Learning; 2013.
37. López-Jornet P, Camacho-Alonso F. The quality of internet sites providing information relating to oral cancer. Oral Oncol 2009 Sep;45(9):e95-e98. [doi: 10.1016/j.oraloncology.2009.03.017] [Medline: 19457707]
38. Hahn BH, McMahon MA, Wilkinson A, et al. American College of Rheumatology guidelines for screening, treatment, and management of lupus nephritis. Arthritis Care Res (Hoboken) 2012 Jun;64(6):797-808. [doi: 10.1002/acr.21664]
39. Goodman RS, Patrinely JR, Stone CA Jr, et al. Accuracy and reliability of chatbot responses to physician questions. JAMA Netw Open 2023 Oct 2;6(10):e2336483. [doi: 10.1001/jamanetworkopen.2023.36483] [Medline: 37782499]
40. Morris LA, Myers A, Thilman DG. Application of the readability concept to patient-oriented drug information. Am J Hosp Pharm 1980 Nov;37(11):1504-1509. [Medline: 7211852]
41. Ghanem YK, Rouhi AD, Al-Houssan A, et al. Dr. Google to Dr. ChatGPT: assessing the content and quality of artificial intelligence-generated medical information on appendicitis. Surg Endosc 2024 May;38(5):2887-2893. [doi: 10.1007/s00464-024-10739-5] [Medline: 38443499]
42. Jindal P, MacDermid JC. Assessing reading levels of health information: uses and limitations of Flesch formula. Educ Health (Abingdon) 2017;30(1):84-88. [doi: 10.4103/1357-6283.210517] [Medline: 28707643]
43. Solnyshkina M, Zamaletdinov R, Gorodetskaya L, Gabitov A. Evaluating text complexity and Flesch-Kincaid grade level. J Soc Stud Educ Res 2017;8(3):238-248 [FREE Full text]
44. Cohen SA, Brant A, Fisher AC, Pershing S, Do D, Pan C. Dr. Google vs. Dr. ChatGPT: exploring the use of artificial intelligence in ophthalmology by comparing the accuracy, safety, and readability of responses to frequently asked patient questions regarding cataracts and cataract surgery. Semin Ophthalmol 2024 Aug;39(6):472-479. [doi: 10.1080/08820538.2024.2326058] [Medline: 38516983]
45. Srinivasan D. Why Google dominates advertising markets. Stanf Technol Law Rev 2020;24(1):55 [FREE Full text]

XSL•FO
RenderX

46. Bojazar R, Do TP, Hansen JM, Dodick DW, Ashina M. Googling migraine: a study of Google as an information resource of migraine management. Cephalalgia 2020 Dec;40(14):1633-1644. [doi: 10.1177/0333102420942241] [Medline: 32703015]
47. Zenone M, Schalkwyk M, Maani N, et al. Google Allows Snake-Oil Advertisers to Target Cancer Patient Queries. Brewminate: A Bold Blend of News and Ideas. 2025. URL: https://brewminate.com/google-allows-snake-oil-advertisers-to-target-cancer-patient-queries/ [accessed 2025-06-07]
48. Pleasants E, Guendelman S, Weidert K, Prata N. Quality of top webpages providing abortion pill information for Google searches in the USA: an evidence-based webpage quality assessment. PLoS ONE 2021;16(1):e0240664. [doi: 10.1371/journal.pone.0240664] [Medline: 33476340]
49. Wijaya D, Daniawan B, Gunawan Y. Search engine optimization (SEO) as a promotional media on Google Search. bit-Tech 2021;4(1):31-39 [FREE Full text] [doi: 10.32877/bt.v4i1.237]
50. Lewandowski D, Sünkler S, Yagci N. The influence of search engine optimization on google's results: a multi-dimensional approach for detecting SEO. 2021 Presented at: WebSci '21: Proceedings of the 13th ACM Web Science Conference 2021; Jun 21-25, 2021; Virtual Event United Kingdom. [doi: 10.1145/3447535.3462479]
51. Kasper D, Fauci A, Hauser S, Longo D, Jameson J, Loscalzo J. Harrison's Principles of Internal Medicine, 19e: Mcgraw-hill New York, NY, USA; 2015.
52. Usage policies. OpenAI. URL: https://openai.com/policies/usage-policies/ [accessed 2025-01-07]
53. Oeding JF, Lu AZ, Mazzucco M, et al. ChatGPT-4 performs clinical information retrieval tasks using consistently more trustworthy resources than does Google Search for queries concerning the latarjet procedure. Arthroscopy: The Journal of Arthroscopic & Related Surgery 2025 Mar;41(3):588-597. [doi: 10.1016/j.arthro.2024.05.025]
54. PubMed. National Institutes of Health. 2023. URL: https://pubmed.ncbi.nlm.nih.gov/about/ [accessed 2025-08-16]
55. Tian S, Jin Q, Yeganova L, et al. Opportunities and challenges for ChatGPT and large language models in biomedicine and health. Brief Bioinformatics 2023 Nov 22;25(1). [doi: 10.1093/bib/bbad493]
56. AlSamhori ARF, AlSamhori JF, AlSamhori AF. ChatGPT role in a medical survey. HYMR 2023;1(2). [doi: 10.59707/hymrTFFP5435]
57. Chen L, Li H, Su Y, et al. Using a Google Web Search analysis to assess the utility of ChatGPT in stem cell therapy. Stem Cells Transl Med 2024 Jan 12;13(1):60-68. [doi: 10.1093/stcltm/szad074] [Medline: 37936506]
58. Sorin V, Brin D, Barash Y, et al. Large language models and empathy: systematic review. J Med Internet Res 2024 Dec 11;26:e52597. [doi: 10.2196/52597] [Medline: 39661968]
59. Howe PDL, Fay N, Saletta M, Hovy E. ChatGPT's advice is perceived as better than that of professional advice columnists. Front Psychol 2023;14:1281255. [doi: 10.3389/fpsyg.2023.1281255] [Medline: 38078232]
60. Praveen SV, Vajrobol V. Can ChatGPT be trusted for consulting? Uncovering doctor's perceptions using deep learning techniques. Ann Biomed Eng 2023 Oct;51(10):2116-2119. [doi: 10.1007/s10439-023-03245-7] [Medline: 37208451]
61. Shahriar S, Lund BD, Mannuru NR, et al. Putting GPT-4o to the sword: a comprehensive evaluation of language, vision, speech, and multimodal proficiency. Appl Sci (Basel) 2024;14(17):7782. [doi: 10.3390/app14177782]
62. Doğan L, Özçakmakcı GB, Yılmaz İ. The performance of chatbots and the AAPOS website as a tool for amblyopia education. J Pediatr Ophthalmol Strabismus 2024;61(5):325-331. [doi: 10.3928/01913913-20240409-01] [Medline: 38661309]
63. Shen SA, Perez-Heydrich CA, Xie DX, Nellis JC. ChatGPT vs. web search for patient questions: what does ChatGPT do better? Eur Arch Otorhinolaryngol 2024 Jun;281(6):3219-3225. [doi: 10.1007/s00405-024-08524-0] [Medline: 38416195]
64. Lang IA, King A, Boddy K, et al. Jargon and readability in plain language summaries of health research: cross-sectional observational study. J Med Internet Res 2025 Jan 13;27:e50862. [doi: 10.2196/50862] [Medline: 39805102]
65. Mohammadi SS, Khatri A, Jain T, et al. Evaluation of the appropriateness and readability of ChatGPT-4 responses to patient queries on uveitis. Ophthalmol Sci 2025;5(1):100594. [doi: 10.1016/j.xops.2024.100594] [Medline: 39435137]
66. Van Bulck L, Moons P. What if your patient switches from Dr. Google to Dr. ChatGPT? A vignette-based survey of the trustworthiness, value, and danger of ChatGPT-generated responses to health questions. Eur J Cardiovasc Nurs 2024 Jan 12;23(1):95-98. [doi: 10.1093/eurjcn/zvad038] [Medline: 37094282]

Abbreviations

LLM: large language model
SLE: systemic lupus erythematosus
FRE: Flesch Reading Ease
FKGL: Flesch-Kincaid Grade Level

XSL•FO
RenderX

Edited by A Mavragani; submitted 24.04.25; peer-reviewed by M Eldakar, Z Huang; revised version received 18.06.25; accepted 03.07.25; published 28.08.25.

Please cite as:
Li K, Peng Y, Li L, Liu B, Huang Z
Evaluating ChatGPT's Utility in Biologic Therapy for Systemic Lupus Erythematosus: Comparative Study of ChatGPT and Google Web Search
JMIR Form Res 2025;9:e76458
URL: https://formative.jmir.org/2025/1/e76458
doi:10.2196/76458

© Kai Li, Yunfei Peng, Luyi Li, Bo Liu, Zhijian Huang. Originally published in JMIR Formative Research (https://formative.jmir.org), 28.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Exploring the Utility of Virtual Clinics for Neurosurgical Patient Consults: Cohort Study to Assess Feasibility

Hassan A Khayat[1], MD; Radwan Takroni[1], MD; Majid Aljoghaiman[1,2], MD; Jessy Moore[1], MSc; Mohamed Alhantoobi[1], MD; Oscar Obiga[1], MD; Marcos Ezequiel Yasuda[1], MD; Bill H Wang[1], MD; Almunder Algird[1], MD; Kesava Reddy[1], MD

[1]Division of Neurosurgery, McMaster University, 237 Barton St. East, Hamilton, ON, Canada

[2]Department of Surgery, King Faisal University, Hofuf, Al-Ahsa, Saudi Arabia

Corresponding Author:
Hassan A Khayat, MD
Division of Neurosurgery, McMaster University, 237 Barton St. East, Hamilton, ON, Canada

Abstract

Background: The popularity of virtual clinics has increased in many settings, especially during and following the COVID-19 pandemic. However, their applicability in neurosurgical care remains understudied.

Objective: The primary goal of this study was to assess the feasibility of conducting a larger, more definitive study at our hospital site in the future. We assessed participant enrollment rate, ability to complete a neurosurgical consult virtually, need for a third party to be present, and participant satisfaction rates. Preliminary evidence on the utility of virtual examination substitutes, compared to currently used in-person assessments, was also explored in our sample of neurosurgical patients.

Methods: In this feasibility study, a cohort of neurosurgery patients, consisting of both new referrals and follow-up visits, was evaluated. Each patient participated in a virtual neurological assessment via Zoom and subsequently in an in-person assessment. Both visits were completed by the same physician. We compared clinical findings and treatment decisions (surgical vs conservative management) between the 2 settings and recorded patient satisfaction with the virtual consultation.

Results: A total of 95 patients were deemed eligible for the study, and of the 52 patients contacted, 35 provided verbal consent and were enrolled. Both the virtual and in-person assessments were completed by 30 participants (86%) with an average length of 3.25 days between visits, which was within the required 2-week period outlined in the study protocol. No barriers were noted from participants (n=6; 20%) who required a third party to be present and this individual was present at both visits. Participant satisfaction rate with the virtual consults exceeded 90%. Clinical decisions were consistent between both visits in 28 cases, and in the 2 visits where decisions differed, it was noted to be a result of inconclusive findings during the virtual consultation. Comparison of individual examination components between the virtual and in-person consultations revealed exam findings to be consistent 77% of the time, and importantly, none of these discrepancies led to a change in clinical decision. No single examination component was noted to be inconsistent more than twice.

Conclusions: These findings support the applicability of the proposed study design to a larger-scale project. No major obstacles or methodological challenges were encountered in achieving the goals of this feasibility study within the target timeframe. This study provides preliminary evidence to support further exploration of the use of virtual consults to help inform clinical decisions in a neurosurgical population.

(JMIR Form Res 2025;9:e69372) doi:10.2196/69372

KEYWORDS

virtual clinic; neurosurgery; Zoom; telehealth; consultation; neurosurgical assessment; consultation

Introduction

The concept of the “virtual clinic” first emerged in orthopedic practice in 2011 as a solution to reduce patient wait times effectively [1]. Since then, this model has expanded to other specialties, including urology [2], ophthalmology [3], rehabilitation [4], and orthopedics [5,6]. Virtual clinics in neurosurgical care remain largely unexplored, with limited literature on their application in this field [7-9] and no studies that directly compare findings from virtual to in-person consultations.

The term “virtual clinic” lacks a universally accepted definition in clinical practice, leading to varied interpretations in the literature. In some instances, a virtual clinic may involve patient evaluations by nonphysician health professionals [3]. In other cases, it could refer to follow-ups conducted primarily through questionnaires [5]. In addition, telecommunication technologies, including real-time audio and video calls, are sometimes used

XSL•FO
RenderX

to assess patients remotely [3]. A combination of these approaches has also been described in different clinical settings [5].

Virtual care offers several demonstrated benefits, such as cost savings and environmental advantages [2]. In Canada, virtual care has been shown to reduce wait times, enhance access, decrease missed appointments, and reduce hospital admissions due to unmanaged conditions. These efficiencies help to lower system costs and complement existing care options [10]. During the COVID-19 pandemic, virtual primary care visits in Canada surged from 4% to 60%, driven by the need to maintain isolation measures. According to the 2021 Canadian Digital Health Survey, virtual care saved Canadians CAD $5.9 (USD $4.3) billion, along with 89 million hours previously spent waiting in clinics or taking time off work to attend appointments [11].

The neurosurgery service at Hamilton Health Sciences (HHS), one of Canada's largest health care systems with a catchment area of approximately 2.3 million people [12], has faced longstanding challenges with wait times. These challenges intensified during the pandemic, as COVID-19 restrictions further limited health care access. In response, virtual clinics emerged as a viable solution to alleviate pandemic-related backlogs and facilitate safe, contactless access to care. Within HHS, various models were implemented during the pandemic to establish virtual clinics, including a telemedicine system for neurosurgical patient follow-ups. Patients were contacted to discuss current symptoms and complaints, review recent imaging, and conduct basic examinations. From this experience, we explored several methods for assessing neurosurgical patients remotely, noting that virtual assessments could serve as an effective alternative to routine in-person visits.

This study aims to evaluate the feasibility and practicality of running a larger, more definitive study at our site in the future and to provide data that could be used for a sample size estimate. As well, we hope to assess preliminary evidence of the effectiveness of a virtual clinic model at informing a clinical decision for outpatient neurosurgical consultations compared to traditional in-person visits. Using an innovative approach, we compared each patient's virtual and in-person assessments. To our knowledge, this methodology has not been used in neurosurgical care. The findings from this study may support the expanded use of virtual clinics within our organization, potentially attracting funding to enhance virtual care infrastructure. Ultimately, we aim to improve patient access to neurosurgical services in a high-demand, large-catchment setting while providing a foundation for conducting reliable neurological examinations in virtual formats.

Methods

Patient Sampling and Recruitment

This feasibility study enrolled 35 outpatients (both new and follow-up) recruited from the Hamilton General Hospital. From July 2023 to April 2024, 95 patients were identified by the treating surgeon or senior investigator as eligible for the study. Patients were considered eligible if they had been referred for a neurosurgical consult at our hospital, were of age to provide consent (18+years), had a laptop or tablet with a webcam, microphone and working internet, and had available imaging evidence of a probable neurosurgical diagnosis. Patients who did not have a working phone number on file, did not answer, or did not return messages (after 2 attempts) were not included in the reported number of eligible participants for this study. A clinical team member contacted candidates by phone to request permission for study representatives to reach out. A research team member then discussed the study with potential participants, and they were given sufficient time to make a decision about study participation and ask any questions they had.

Of the 52 patients contacted by the research team, 35 provided verbal consent to participate, and 17 were not interested in participating, canceled their neurosurgical consult appointment, or declined participation because they did not have available technology to complete the virtual consult. All patients were capable of providing their own consent. Verbal consent was obtained prior to scheduling a virtual visit via Zoom, and all participants received a signed copy of the consent form via email. Virtual Zoom calls were scheduled no more than 2 weeks prior to the patients' in-person clinic visit. To maintain anonymity, each patient was assigned a random code once the data were collected. Each patient was assessed by a physician representative of the study team, and importantly, the same physician conducted both the virtual and in-person assessments with the goal of avoiding interrater bias. Based on our study aims and resource considerations, the sample size is considered adequate to achieve our feasibility objectives and provide valuable insights to guide the design of a future definitive trial.

Consultative Methods

The table below (Table 1) outlines how the virtual examination was conducted through real-time video calls, using virtual substitutes for each component of the neurological examination. The table also compares these substitutes with the standard in-person examination methods. For standardization, the same examination techniques were applied to each study participant.

XSL•FO
RenderX

Table . Explanation of the virtual examination (Zoom) substitute used for each standard, in-person consultative method in a cohort of patients referred for a neurosurgical consult.

Exam component	Standard method (in-person)	Virtual substitute
Orientation	Ask location, date, and time	Same as standard
Mental status	Run MMSE[a] questionnaire by the patient	Send the patient link to complete the exam
Extraocular muscles	Ask patient to follow object in H shape. Report diplopia	Ask the patient to look in 4 main directions (up, down, right, and left) and report any diplopia
Visual field	Confrontation test	Ask the patient to look across the room to a single point on the wall with one hand covering one eye. Move the other hand to enter the visual field through each quadrant; report visibility.
Facial sensory	Touch patient's face with a cloth over dermatomes, ask for comparison	Show patient how to touch the same areas with a cloth or tissue for comparison via screen
Facial motor	Ask patient to mimic eye closure, blow cheeks, and give a big smile	Same
Hearing	Rub fingers next to each ear simultaneously	Same
Palate	Ask patient to say "Aaaa" and observe palate elevation	Same (ask patient to provide light source if needed)
Shoulder	Ask patient to shrug shoulders	Same
Tongue	Ask patient to protrude tongue and move side to side	Same
Pronator drift	Ask patient to outstretch arms and close eyes	Same
Upper limb motor	Physician examines each muscle group	Ask patient to move against gravity, using a 0.5 L water bottle as shown by the physician. Examine each muscle group
Lower limb motor	Physician examines each muscle group	Ask patient to stretch legs out while seated or to stand from chair; assess walking on toes and heels
Sensory	Light and sharp touch on all dermatomes	Show the patient how to touch areas with a cloth and compare sensation
Articulation	Say name, count 1 - 5	Same
Nystagmus	Physician closely examines	Ask patient to look directly at the camera, changing gaze to periphery if needed
Dysmetria	Finger-to-nose test	Ask patient to alternate finger between nose and a chosen nearby point
Rapid alternating	Ask patient to tap one hand over the other with palm and dorsum	Same
Gait	Take steps across the room, turn, and return	Same

[a]MMSE: mini mental state examination.

Virtual Telemedicine Assessment

Patients were contacted through a real-time video call (via Zoom). The presence of a relative or nurse was allowed, similar to a face-to-face setting. Assistance was provided to patients as needed for certain examination maneuvers. The examination here is done through proposed techniques that enable the physician and the patient to conduct the examination maneuvers virtually. Patients can use their phones, laptops, or office personal computers as long as they have a working camera. Table 1 shows the virtual substitutes compared to the standard examination techniques.

Routine Face-To-Face Assessment

Following the virtual clinic assessment, each patient was seen in person at the clinic within a 2-week period.

Study Outcomes

The primary goal of this study was to assess the feasibility of conducting a larger-scale project in the future. Secondary outcomes included evaluating: (1) whether clinical decisions (surgical vs conservative) changed following the in-person visit compared to the virtual visit, (2) patient satisfaction with the assessments, (3) consistency of clinical examination findings between the 2 visits, and (4) the distance patients traveled to the clinic.

XSL•FO
RenderX

To assess protocol feasibility, several performance metrics were observed to ensure reproducibility and effectiveness, including (1) the number of participants who were approached about the study and agreed, (2) the number of participants who completed both visits, (3) the completion of both visits within the 2-week period, (4) satisfaction rates of patients with virtual consult visits, and (5) the need for a third party to be present.

Data Management and Analysis

Data from all study participants were collected using electronic forms via REDCap (Vanderbilt University). Each patient was assigned a unique code to deidentify their data. All patient information was kept strictly confidential and securely stored by the research team. Electronic data were password-protected and housed behind a firewall (Citrix and REDCap), with no paper-based forms used in this study. Usual progress notes from everyday clinical practice were used as reference in cases of conflict, discrepancy, or missing data during the study.

Collected data from all participants were pooled and analyzed using IBM SPSS Statistics for Windows (version 20.0; IBM Corp). Categorical variables were expressed as frequencies and percentages, while continuous variables were expressed as means and SD. Sensitivity and specificity were calculated to compare the ability of a virtual consultation on Zoom to help physicians make the correct clinical decision regarding patient outcome (ie, surgical vs conservative management). This was compared to in-person visits as the current "gold standard" for making a clinical decision. Sensitivity is interpreted as "the sensitivity of the virtual consult tool to correctly identify patients with a surgical outcome." Specificity is interpreted as "the specificity of the virtual consult tool to correctly identify the number of patients who should undergo conservative management."

Ethical Considerations

This study is a single-center feasibility study that has been approved by the Hamilton Integrated Research Ethics Board (HiREB approval number 13408). This study was conducted following the ethical principles of the Declaration of Helsinki. All participants were contacted by the research coordinator, provided with the study consent form, and given time to review the form and ask any questions they had. Study procedures and the measures taken to ensure participant privacy were explained to all study participants prior to consent being obtained. All included participants provided voluntary verbal consent to the research coordinator. Following data collection, all data were anonymized using a random ID code prior to data analysis. Participants were not provided compensation for their study participation.

Results

Participant Demographics

Data were collected over the study recruitment period from October 2023 to April 2024. A total of 52 patients were contacted about participating in this study, 35 patients consented and were enrolled, and all 35 attended a virtual consultation. Of these, 30 participants met with the same physician during both their virtual and in-person consultations (Figure 1). Five patients attended a virtual consultation and later a surgical clinic visit but were seen by different physicians for each and thus are excluded from the main analysis of this study.

Figure 1. Schematic displaying the flow of patients through the study protocol and the number of individuals included at each stage of the study.

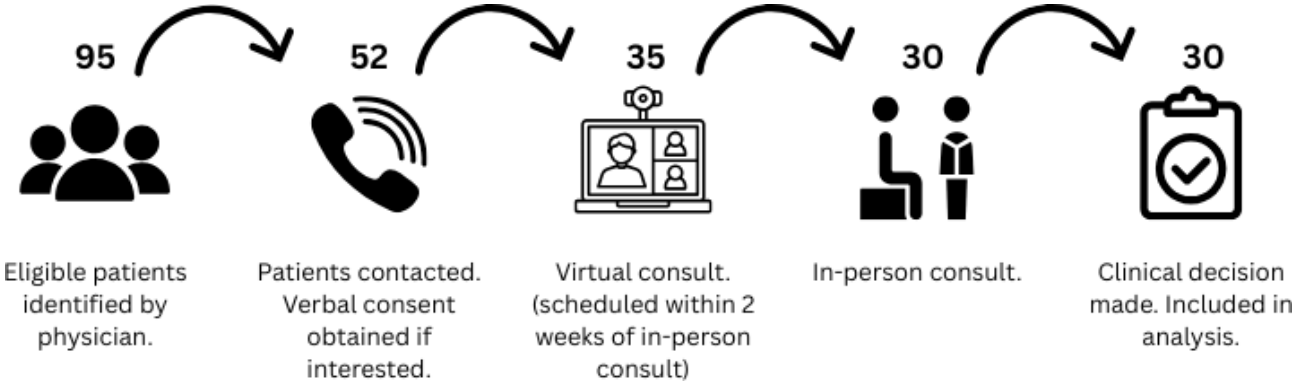

The average age of participants was 57.32 years (±12.75 years) and 56.7% of the sample were male. Among the 30 included patients, 7 had sellar tumors, 7 had lumbar spine degeneration or stenosis, 4 had cervical radiculopathy, 3 had cranial meningiomas, 3 had gliomas, 3 had cerebellopontine angle tumors, 1 had spine tumor, 2 had idiopathic intracranial hypertension, and a sacral cyst. Of the 30 included participants, 11 were scheduled for surgery.

Virtual appointments took place via Zoom and lasted approximately 20 - 25 minutes, which is comparable to the average in-person clinic visit at our hospital. The average interval between the virtual and in-person consultations was 3.25 days (minimum 1 day; maximum 5 days). Twenty-one patients (70%) were seen for new consultations and 9 patients (30%) were follow-up visits. Six participants (20%) had a third party present during both their virtual and in-person consultations (Table 2).

XSL•FO
RenderX

Table . Participant demographics and descriptive variables (N=30) in a cohort of patients referred for neurosurgical consultation at the Hamilton General Hospital between July 2023 and April 2024.

Variable	Value
Age, years, mean (SD)	57.32 (12.45)
Sex, n (%)	
Male	17 (56.7)
Female	13 (43.3)
Visit type, n (%)	
New	21 (70)
Follow-up	9 (30)
Third party present, n (%)	
Yes	6 (20)
No	24 (80)
Discrepancy in findings, n (%)	
Yes	7 (23.3)
No	23 (76.7)
Number of discrepancies, n (%)	
0	23 (76.6)
1 - 3	6 (20.0)
3 - 6	0
>6	1 (3.4)
Distance traveled, n (%)	
0 - 20 Km	10 (33.3)
21 - 40kmm	4 (13.3)
41 - 60 km	10 (33.3)
61 - 80 km	4 (13.3)
>100 km	2 (6.7)
Patient satisfaction (%), mean (SD)	91.1 (11.8)

Approximately one-third of the sample lived within 20 km of the clinic, while 53.4% resided more than 40 km away, including 2 participants who traveled over 100 km. Overall satisfaction with the virtual consultations was high, with a mean satisfaction score of 91.1% (±11.8%) when participants were asked “how satisfied were you with using a Zoom appointment for your visit on a scale of 0 - 100” (Table 2).

Of the 30 patients recruited, only 2 (7%) were noted to have a change in the treatment plan between the virtual and in-person visits, and this was noted to be a result of inconclusive findings during the virtual consultation. The treatment decision was consistent across both visits for the remaining participants (28/30; 93%). Comparisons of neurosurgical examination components between the virtual and in-person consultations revealed that 5 patients had 1 discrepancy between findings, 1 patient had 2 discrepancies, and 1 patient had 7 discrepancies (Table 2). Importantly, none of these discrepancies led to a change in clinical decision.

The virtual consult demonstrated a sensitivity of 90% and a specificity of 98% when compared to the in-person consult, the current gold standard, for making a clinical decision for neurosurgical patients. In addition, the positive predictive value was 78% and the negative predictive value was 99%.

Upon examining individual examination components across both consultative methods, no single component accounted for more than 2 discrepancies. Discrepancies were noted in visual acuity, lower limb motor function, sensory body assessments, and gait for 2 participants, although not all were observed in the same individuals (Table 3). Differences in ophthalmologic examinations were observed but were comparatively minor. None of these differences in findings led to changes in clinical decisions by the physician.

XSL•FO
RenderX

Table . Comparison of findings for each examination component between in-person consults and virtual consults on Zoom in a cohort of patients referred to a neurosurgeon at the Hamilton General Hospital.

Examination component	Virtual consult	In-person consult	Difference noted
Orientation	0	0	0
Mental status	1	1	0
Extraocular muscles	2	1	1
Visual field	5	4	1
Visual acuity	7	8	1
Facial sensory	2	2	0
Facial motor	1	1	0
Hearing	3	3	0
Palate movement	1	0	1
Shoulder	0	0	0
Tongue	0	0	0
Pronator Drift	0	0	0
Upper limb motor	5	4	1
Lower limb motor	5	3	2
Sensory whole body	9	7	2
Nystagmus	0	0	0
Dysmetria	1	1	0
Rapid alternating	1	1	0
Gait	2	4	2

In Table 3, values refer to the number of patients with a reported finding that was abnormal for each examination component.

Discussion

Principal Findings and Comparison With Previous Works

This study assessed the feasibility of conducting the proposed study design in a larger-scale, more definitive trial. The design, ultimately, was structured to assess the utility of virtual consultations as an alternative to in-person consultations for neurosurgical patients, with in-person assessments currently used as the gold standard. To maintain consistency, the same health care provider conducted both assessments for each patient, allowing for direct comparisons in terms of treatment decisions (surgical vs conservative management) and clinical findings. The study findings indicate that the proposed protocol was implemented without significant barriers and met the designated timeframe (8 - 10 months). Of the 52 eligible patients that were successfully contacted, 35 were enrolled in the study, giving an enrollment rate of 67%. We feel this is a satisfactory rate given the number of physicians that were involved and supports the feasibility to run a larger study at our site. Furthermore, 86% of patients completed both virtual and in-person visits and were included in our final analysis, and 100% of patients were able to be scheduled for both visits within a 2-week window. Overall, 20% of participants required a third party to be present during both consults, but this led to no barriers.

A review of the literature reveals that this particular methodology—comparing virtual and in-person consultations in neurosurgery—has not been previously applied. Several studies conducted during the COVID-19 pandemic assessing a neurosurgical population have shown high levels of satisfaction with telemedicine consultations among both patients and surgeons [7-9]. One study highlighted the perceived benefits of telemedicine from both perspectives [9]. Patients found virtual consultations to offer enhanced efficiency and improved communication, while identifying safety concerns, as well as technological and administrative issues, as drawbacks. Neurosurgeons, on the other hand, noted the benefits of increased efficiency and reduced COVID-19 exposure, while the drawbacks included challenges with physical examinations, technological issues, and concerns for patients. It was suggested that virtual consultations may be particularly suitable for postoperative reviews and discussing scan results, while new patient consultations, neuro-oncology cases, and patients with new-onset neurological deficits might be contraindicated for virtual assessments [9]. Nonetheless, in line with Chaganty et al (2023), virtual consultations may not benefit the entire neurosurgical patient population, and each patient should be evaluated individually. Importantly, to our knowledge, no other studies have directly compared virtual consultations with in-person consultations within the same patient population.

XSL•FO
RenderX

In a related study by Mahmoud et al [13], the applicability of telemedicine was prospectively evaluated in pediatric perioperative care against a historical cohort of in-person assessments. This study concluded that telemedicine could effectively bridge the communication gap between patients and physicians, with a satisfaction rate of 94% [13], close to our reported rate of 91%. While numerous studies highlight the benefits of virtual care, including cost-effectiveness, reduced wait times, and decreased no-show rates, challenges remain. Implementing a virtual care program requires advanced technical capabilities, cybersecurity, and regulatory compliance [14, 15].

Most studies on virtual care focus on metrics such as cost, efficiency, and accessibility rather than on the safety, consistency, and reliability of clinical assessments. A potential reason for this gap is the lack of tools specifically designed to evaluate these aspects. Conducting a thorough clinical examination remotely presents unique challenges, especially in neurosurgery, where traditional assessments rely on direct physical contact between patient and provider, which can significantly impact treatment decisions. These challenges hinder the broader implementation of virtual care in neurosurgical practice.

One of the main challenges we encountered in this study was developing a consistent neurological examination protocol for telemedicine. To address this, we created “virtual substitutes” for each component of the neurological examination. With these substitutes, a comprehensive neurological examination could be adapted to the virtual setting. Patients were also asked to complete previously validated questionnaires that assess specific neural functions. Following the virtual assessment, patients underwent an in-person visit, allowing for a direct comparison between the 2 modes. Based on the results of this study, the virtual substitutes we used yielded findings that were consistent 77% of the time with in-person examinations. In a future larger, more definitive study, we plan to evaluate the safety and reliability of the virtual clinic for neurosurgical consultations.

Data analysis revealed that in 93% of cases, the treating physician upheld the same treatment decision and plan (operative vs conservative management) across both virtual and in-person assessments. It is important to note that in the 2 cases (7%) where the treatment decision varied, it was a result of inconclusive findings during the virtual consultation. Both of these participants had cervical radiculopathy, and while conclusions cannot be drawn directly from this feasibility study, this may represent a diagnosis that is not appropriate to assess virtually. Clinical examination findings were consistent in 77% of cases, and no discrepancies in examination findings led to a change in clinical decision. To our knowledge, this is the first study to directly compare virtual and in-person consultations in neurosurgical patients, demonstrating a sensitivity of 90% and specificity of 98% for virtual consultations as a clinical decision-making tool in relation to in-person assessments. We report these findings with caution, acknowledging that this is a feasibility study with a small sample size. Nevertheless, we felt it was important to report these preliminary findings which support a future study to explore the use of virtual visits in neurosurgical consultations.

More than 50% of the patients in our study resided farther than 40 km from the hospital. Offering a broader range of patients, the option for virtual neurosurgical consultations could contribute to reduced carbon emissions from transportation and may alleviate the burden of travel time and costs for patients. Assuming an estimated 251 g of CO_2 are emitted per kilometer by light-duty vehicles in Canada (based on 2017 - 2021 models) [16], and considering only the 16 participants who live more than 40 km from Hamilton General Hospital, the use of virtual consultations in place of in-person visits could have potentially reduced CO_2 emissions by at least 421,680 g. This represents a preliminary estimate and highlights a possible area for further exploration. Previous research highlights several demonstrated benefits of virtual care including cost savings, environmental advantages, reduced patient wait times, enhanced access to care, fewer missed appointments, and fewer hospital admissions due to unmanaged conditions [2,10,11]. Future research should aim to develop guidelines outlining which patient populations are suitable for virtual consultations and identify scenarios where this approach may not be feasible or safe.

Limitations

The limitations of this study include the small sample size and the lack of prior validation for the virtual examination method. With a larger sample size, future research aims to validate the virtual examination methods used in this study and assess the safety and reliability of virtual consults in a neurosurgical population. More specific insights could be made regarding appropriate patient populations for virtual consults, and a larger study may provide more generalizable results. Conclusions from this study cannot be generalized to all virtual consults or to all neurosurgical patients, and while we aimed to minimize interrater bias by having the same physician meet with each participant both virtually and in person, a future study may include a subset of patients who are seen by a different physician at each visit to explore the potential influence of intrarater bias.

Expected Outcome

Most existing literature highlights the cost-effectiveness [4,10] and efficiency [11,12] of virtual care. While the primary aim of this study was to assess feasibility at our site, it also offers preliminary evidence on the potential utility of virtual examination substitutes in the assessment of neurosurgical patients. Although the reported sensitivity and specificity should be interpreted with caution, they nonetheless provide meaningful insight into the effectiveness of these virtual tools in informing clinical decision-making. The data generated from this project contribute to the growing body of literature and can assist stakeholders in directing funding and resources toward improving access to care through safe and reliable virtual modalities. Our findings support the need for a larger study to evaluate the safety and reliability of virtual consultations and to produce more generalizable results. Ultimately, the overarching goal of this research is to enhance access to care and explore whether virtual consultations can serve as a viable alternative for neurosurgical patients.

XSL•FO
RenderX

Conclusions

Collaborative efforts and a well-organized research team with clear objectives and timelines enabled the successful execution of this feasibility study. No major obstacles were encountered, and the project adhered to the predetermined timeframe. Preliminary data from this study show that it may be feasible to plan a larger future study and provide the data to perform a power analysis for such a study. We examined the consistency in clinical assessments and treatment plans and showed that no discrepancies in examination components between virtual and in-person consultations led to a change in clinical decision. A more statistically robust study is required to substantiate these findings and provide more generalizable results for a population of neurosurgical patients.

Acknowledgments

HK, KR, RT, AA, and BW were responsible for the study concept. HK, MA, MEY, and OO performed patient consultations. JM contacted patients and scheduled virtual consultations. HK and JM performed data analysis. HK, MA, and JM prepared the manuscript. All authors have reviewed and approved the manuscript for publication. This project used the resources and facilities of outpatient clinics at Hamilton Health Sciences, along with associated telecommunication platforms for data collection. Data management and storage were conducted using research tools provided by McMaster University. No specific grants or funding were received for this project.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

References

1. Vardy J, Jenkins PJ, Clark K, et al. Effect of a redesigned fracture management pathway and “virtual” fracture clinic on ED performance. BMJ Open 2014 Jun 13;4(6):e005282. [doi: 10.1136/bmjopen-2014-005282] [Medline: 24928593]
2. Miah S, Dunford C, Edison M, et al. A prospective clinical, cost and environmental analysis of a clinician-led urology virtual clinic. Ann R Coll Surg Engl 2019 Jan;101(1):30-34. [doi: 10.1308/rcsann.2018.0151] [Medline: 30286648]
3. Kern C, Kortuem K, Hamilton R, et al. Clinical outcomes of a hospital-based teleophthalmology service: what happens to patients in a virtual clinic? Ophthalmol Retina 2019 May;3(5):422-428. [doi: 10.1016/j.oret.2019.01.011] [Medline: 31044734]
4. Worthen-Chaudhari L. Effectiveness, usability, and cost-benefit of a virtual reality-based telerehabilitation program for balance recovery after stroke: a randomized controlled trial. Arch Phys Med Rehabil 2015 Aug;96(8):1544. [doi: 10.1016/j.apmr.2015.03.025] [Medline: 26216402]
5. Preston N, McHugh GA, Hensor EMA, et al. Developing a standardized approach to virtual clinic follow-up of hip and knee arthroplasty. Bone Joint J 2019 Aug;101-B(8):951-959. [doi: 10.1302/0301-620X.101B8.BJJ-2018-1566.R1]
6. Lovelock T, O’Brien M, Young I, Broughton N. Two and a half years on: data and experiences establishing a “Virtual Clinic” for joint replacement follow up. ANZ J Surg 2018 Jun 27;2018(7):707-712. [doi: 10.1111/ans.14752] [Medline: 29952097]
7. De Biase G, Freeman WD, Bydon M, et al. Telemedicine utilization in neurosurgery during the COVID-19 pandemic: a glimpse into the future? Mayo Clin Proc Innov Qual Outcomes 2020 Dec;4(6):736-744. [doi: 10.1016/j.mayocpiqo.2020.07.013] [Medline: 33324948]
8. Sharma A, Winkelman RD, Schlenk RP, Rasmussen PA, Angelov L, Benzil DL. The utility of remote video technology in continuing neurosurgical care in the COVID-19 era: reflections from the past year. World Neurosurg 2021 Dec;156:43-52. [doi: 10.1016/j.wneu.2021.08.145] [Medline: 34509681]
9. Chaganty SS, Walter H, Smith F, Sharma H. Evaluating patient and neurosurgeon perspectives on virtualisation of neurosurgery clinics in the COVID-19 era: a prospective mirror survey study. Br J Neurosurg 2023 Apr;37(2):142-147. [doi: 10.1080/02688697.2021.1979192] [Medline: 34546116]
10. Covid19 policy brief-realizing the full potential of virtual care in ontario. Ontario Chamber of Commerce. URL: https://occ.ca/wp-content/uploads/COVID19-Policy-Brief-Virtual-Care-final.pdf [accessed 2025-08-18]
11. Environmental and financial benefits of virtual care in canada. Canada Health Infoway. URL: https://www.infoway-inforoute.ca/en/component/edocman/3885-environmental-and-financial-benefits-of-virtual-care-in-canada/ [accessed 2025-08-05]
12. MacIsaac R. A healthcare vision for hamilton’s future. Hamilton Health Sciences. 2018. URL: https://www.hamiltonhealthsciences.ca/share/healthcare-vision/ [accessed 2025-08-05]

XSL•FO
RenderX

13. Mahmoud MA, Daboos M, Gouda S, et al. Telemedicine (virtual clinic) effectively delivers the required healthcare service for pediatric ambulatory surgical patients during the current era of COVID-19 pandemic: a mixed descriptive study. J Pediatr Surg 2022 Apr;57(4):630-636. [doi: 10.1016/j.jpedsurg.2021.11.018] [Medline: 34953564]
14. Ravitz N. The economics of a telehealth visit: a time-based study at penn medicine. hfma. URL: https://www.hfma.org/topics/financial-sustainability/article/the-economics-of-a-telehealth-visit--a-time-based-study-at-penn-.html [accessed 2025-08-05]
15. McMaster T, Wright T, Mori K, Stelmach W, To H. Current and future use of telemedicine in surgical clinics during and beyond COVID-19: a narrative review. Ann Med Surg (Lond) 2021 Jun;66(66):102378. [doi: 10.1016/j.amsu.2021.102378] [Medline: 33996071]
16. Hien NLH, Kor AL. Analysis and prediction model of fuel consumption and carbon dioxide emissions of light-duty vehicles. Appl Sci (Basel) 2022;12(2):803. [doi: 10.3390/app12020803]

Abbreviations

HHS: Hamilton Health Sciences
HiREB: Hamilton Integrated Research Ethics Board

Edited by A Mavragani; submitted 04.12.24; peer-reviewed by AA Arafat, S Yordanov; revised version received 06.07.25; accepted 22.07.25; published 28.08.25.

Please cite as:
Khayat HA, Takroni R, Aljoghaiman M, Moore J, Alhantoobi M, Obiga O, Yasuda ME, Wang BH, Algird A, Reddy K
Exploring the Utility of Virtual Clinics for Neurosurgical Patient Consults: Cohort Study to Assess Feasibility
JMIR Form Res 2025;9:e69372
URL: https://formative.jmir.org/2025/1/e69372
doi:10.2196/69372

© Hassan A Khayat, Radwan Takroni, Majid Aljoghaiman, Jessy Moore, Mohamed Alhantoobi, Oscar Obiga, Marcos Ezequiel Yasuda, Bill H Wang, Almunder Algird, Kesava Reddy. Originally published in JMIR Formative Research (https://formative.jmir.org), 28.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Usability and Acceptability of an App-Based Approach to Treat Low Back Pain: Preplanned Secondary Analysis of a Randomized Controlled Trial

Svenja Kaczorowski[1,2], MSc; Katharina Trompeter[2], PhD; Carolin Kramer[2], MSc; Christian Grüneberg[2], PhD; Christian Thiel[2], PhD; Lars Donath[1], PhD; Daniel L Belavy[2], PhD

[1]Department of Intervention Research in Exercise Training, German Sport University Cologne, Cologne, Germany

[2]Division of Physiotherapy, Department of Nursing, Midwifery and Therapeutic Sciences, Bochum University of Applied Sciences, Gesundheitscampus 6-8, Bochum, Germany

Corresponding Author:
Svenja Kaczorowski, MSc
Department of Intervention Research in Exercise Training, German Sport University Cologne, Cologne, Germany

Abstract

Background: Low back pain (LBP) is a major cause of disability worldwide. To tackle issues such as long wait times and limited access to conventional care, telemedicine is emerging as a viable alternative. It offers benefits such as reduced travel and increased flexibility, with evidence showing comparable effectiveness to in-person care. However, usability remains a key challenge, impacting patient compliance.

Objectives: In a 3-arm randomized controlled trial, our preplanned secondary analysis aimed to assess the usability and patient perceptions toward an autonomous app-based intervention ("NOLA") for LBP to improve clinical practice of telemedicine interventions. Objectives included evaluating app usability, assessing perceptions toward telemedicine, and exploring app usage, adherence, and motivation.

Methods: Patients with LBP were recruited from May to August 2022 and randomized into App, Physio+App, or Physio groups. App and Physio+App groups were included in this subanalysis. Intervention duration was 6 weeks. Data on baseline characteristics, System Usability Scale, Telemedicine Perception Questionnaire, app usage, adherence, and motivation were collected via web-based questionnaires.

Results: A total of 64 participants were randomized to use the app with available data for 38 participants. The mean age of participants who completed was 49.9 (SD 13.6) years, with 78% (29/38) experiencing LBP for more than 2 years. Usability scores (0 - 100) were good (Physio+App: median 78, IQR 58-92, app: median 86, IQR 65-91). Positive telemedicine perceptions were noted, with 84% (15/20) rating it an adequate addition to usual care. App usage varied, with 43% (16/35) using it 3 to 5 days per week, and 64% (21/33) reported motivation to use the app. The dropout rates (App: 14/32, 44%; Physio+App: 12/32, 38%; Physio: 11/29, 38%) were similar, but participants who dropped out had statistically significantly less pain (completers: mean 3.9, SD 2.0; dropouts: mean 3.0, SD 2.0; P=.02). Reasons for dropout were mostly not reported.

Conclusions: The app "NOLA" demonstrated good usability, and participants expressed positive perceptions toward telemedicine in those who completed the intervention. Despite concerns about the lack of physical contact, the majority considered telemedicine a convenient form of health care delivery. App usage and motivation were favorable, emphasizing the potential of app-based interventions in managing LBP.

Trial Registration: German Clinical Trial Registry DRKS00029099; https://drks.de/search/en/trial/DRKS00029099

(JMIR Form Res 2025;9:e59866) doi:10.2196/59866

KEYWORDS

telemedicine; mHealth; low back pain; app-based intervention; usability; acceptability; app; randomized controlled trial; patient compliance; app usage; pain; mobile phone

Introduction

Low back pain (LBP) is the most common cause for disability worldwide [1] with a high economic burden of health care costs, especially in developed countries [2], and creates up to US $108 billion of health care costs per year [3]. Recommendations for the treatment of patients with LBP include an active approach that also takes education into account [4]. Treatment of patients with LBP usually takes place in an in-person setting. Waiting times, insufficient infrastructure, the recent COVID-19

XSL•FO
RenderX

pandemic, and access to therapy emphasize the need for the integration of alternative approaches into LBP care. The importance of telemedicine as an alternative or adjunctive to conventional physiotherapy has increased rapidly in the last decade [5,6]. Telemedicine shows several advantages compared to standard in-person care, such as reduced travel times [7], increased time flexibility [8,9], and self-management [10]. Several pathways exist to deliver telemedicine, such as websites, apps, videoconferencing, or phone calls. The current evidence shows that telemedicine tends to be effective for patients with LBP in improving pain, physical function, and quality of life compared to conventional care [11,12].

While the effectiveness of an intervention is an important factor, it is also essential to consider patients' perceptions and attitudes toward the intervention. With the increasing use of telemedicine, it became clear that it is generally accepted by patients seeking help for musculoskeletal conditions [13-15]. However, a recent umbrella review revealed that 1 main barrier in the implementation of telemedicine is the usability of a digital intervention, which is linked to adherence to the intervention [16]. Due to a variety of possibilities for how telemedicine can be delivered, factors that influence usability are specific for each mode of delivery [17]. The systematic review of Thurnheer et al [18] showed that mobile apps in pain management have, in general, good usability, but details that improve usability remain unclear. To support the implementation of mobile apps into clinical practice, it is of crucial importance to investigate factors that influence acceptance and usability.

This paper aimed to evaluate the usability of the "NOLA" app, which integrates exercise and education for patients with LBP, as well as patient perceptions of telemedicine in patients who received an app-only intervention compared to those who received in-person physiotherapy in addition to the app. Further, this study aimed to evaluate the frequency of use of the app, compliance with the telemedicine intervention, and participants' motivation to use the app. These factors are important in shaping future clinical trials and improving the current clinical practice of telemedicine interventions to enhance patient satisfaction.

Methods

Study Design

This study is a preplanned secondary analysis of a 3-arm parallel randomized controlled trial. The main analysis on effectiveness, adherence, and sustainability is to be reported elsewhere by the principal investigator. The trial has been conducted per the Declaration of Helsinki and followed recommendations of the CONSORT (Consolidated Standards of Reporting Trials)–eHealth checklist [19]. The trial took place in 2 physiotherapy practices in the Ruhr area in Germany.

Ethical Considerations

The trial was approved by the local ethical committee of the Hochschule für Gesundheit (University of Applied Sciences; ID: 220317) and registered at the German Clinical Trial Registry (DRKS00029099). Participants who consented to participate in the main trial agreed to the use of their data for this secondary analysis. Data have been anonymized for the analysis. The participants did not receive any direct financial compensation but could use the app for free during the time of this trial.

Patient and Public Involvement Statement

No patient or public involvement occurred in this study.

Recruitment

Participants were recruited from 9 orthopedic practices in the Ruhr area in Germany from May to August 2022. Additionally, recruitment was done via social media (Instagram [Instagram from Meta] and Facebook [Meta]) and local newspaper advertisements. In case of interest, potential participants were contacted via telephone and informed about the benefits and risks of this study. If they were still interested, they were invited to 1 of this study's sites to review the eligibility for this study. All participants signed an informed consent document before being randomized into 1 of the 3 groups. To control for expectation bias, it was explained to the participants that it was unknown which intervention was superior. This study presents a preplanned exploratory analysis of secondary outcomes of a randomized controlled trial. The sample size was determined for the preplanned primary outcome "pain intensity" of the main trial, which is not included in the present analysis. Formal sample size calculations were not performed for the current secondary outcomes to avoid unnecessarily inflating this trial's size.

Eligibility Criteria

Individuals older than 18 years were included if they had a diagnosis of LBP (*ICD-10* [*International Statistical Classification of Diseases, Tenth Revision*] M54.5) with pain lasting for at least 6 weeks. Participants were required to have access to a smartphone or tablet and be able to use it. Further inclusion criteria were sufficient knowledge of German, willingness to travel to the treatment centers, and cognitive capacity to understand this study's information and give informed consent. Exclusion criteria included patients with LBP originating from a serious pathology (eg, fractures, spinal stenosis, or tumors), patients with complaints in other body parts that made participation in exercise training not feasible, and patients who were pregnant.

Randomization and Blinding

Participants were randomized using permuted block randomization with random block sizes. A randomization list was generated online (Sealed Envelope Ltd 2021, 2022) with only subject ID assignment. The allocation ratio between groups was 1:1:1. Randomization was performed after clinical baseline tests for the main analysis were completed. Group allocation was communicated verbally to the participants, as this allowed any unanswered questions about the respective intervention to be discussed directly. Participants and physiotherapists could not be blinded to randomization because they were performing and administering the active intervention.

Intervention

Overview

Participants were randomized into 1 of the following 3 groups: App, Physio+App, or Physio. Participants in the Physio+App

XSL•FO
RenderX

group also received access to the app "NOLA" in addition to standard physiotherapy. Only the first 2 groups (App, Physio+App) were included in this preplanned secondary analysis. The intervention duration was 6 weeks for both groups.

App Intervention

Participants in the App group were given access to the app "NOLA" (MyLetics GmbH), which has been developed by physiotherapists, medical doctors, and information technology experts. It is stand-alone software per the European Medical Device Directive MDD (2007/47/EC), which can be set up on the smartphone without any additional devices (Figure 1). The training program integrated in Nola can be used autonomously. First, each user completes an anamnesis in which questions about age, sex, current complaints, and treatment goals are answered. The anamnesis takes into account medical, pain-physiological, and training-scientific parameters. On this basis, a personalized therapy and training plan is drawn up automatically, which is adapted to the individual's training progress. In addition to training, the multimodal therapy plan includes everyday exercises and knowledge lessons. On the day of the initial medical examination, a short introduction to the app was provided. Participants were advised to complete a minimum of 2 sessions per week, while no maximum threshold was set. Reminders to exercise were sent out via push messages regularly. Each session ranged from 8 to 25 minutes, depending on the current pain status, and contained exercises that were demonstrated in video tutorials and could be performed without additional equipment.

Figure 1. Screenshots of the NOLA app that participants in both included groups used on their mobile devices. The app contains an exercise program as well as educational content.

Physiotherapy Intervention

Participants received the physiotherapy intervention in the practice "Schwerpunktpraxis für Human Movement" in Essen, Germany. The content of the physiotherapy intervention was at the discretion of the therapist and consisted of one or more of the following: manual therapy, strength training, stabilization exercises, or stretching. The app usage was integrated into the therapeutic process so that exercises could be explained and supervised directly by the therapist. All 4 physiotherapists were trained and experienced in the integration of the app in the therapeutic process. A standardized introduction to the use of the app was achieved through internal training of the physiotherapists. Participants were advised to complete a minimum of 2 sessions per week, including the physiotherapy session. To reflect the reality of care in Germany, it was possible for participants to attend 6 to 12 physiotherapy sessions in person overall with a duration of 20 minutes each.

Outcomes

Baseline characteristics are reported for demographic variables (sex, age, height, and weight), pain duration, medication use, and current sick leave, as well as for pain intensity (numeric rating scale: 0 - 10) and disability (German Oswestry Disability Index: 0 - 32 [20]). For this preplanned secondary analysis, we collected data on the usability of the NOLA app and perceptions of patients toward telemedicine in general. We used a German translation of the System Usability Scale [21] (SUS) and adapted the wording accordingly to the use of an app (Multimedia Appendix 1). The overall score ranges from 0 (worst) to 100 (best). Further, we used the German translation of the Telemedicine Perception Questionnaire [22] (TMPQ) and

XSL•FO
RenderX

adapted it to only use meaningful questions regarding the app-based intervention (Multimedia Appendix 2). As we only used a selection of questions from the TMPQ, the overall score ranged from 0 (worst) to 60 (best). Furthermore, the self-reported frequency of app usage, the adherence, motivation, and perceptions toward the future use of the app were evaluated by multiple-choice questions, which were piloted before data collection started. In case of a dropout after the start of the intervention, we asked for the reasons for the dropout. All data used in our analysis were collected online via "SoSci Survey" (SoSci Survey GmbH) at the end of the intervention. There were no changes to trial outcomes after this trial commenced.

Deviations From the Registered Trial Protocol

We recruited fewer participants than planned for the main analysis on pain intensity. Importantly, this secondary analysis was not part of the original power calculation and should therefore be considered exploratory in nature.

Statistical Analyses

Frequencies of answers for each question in the SUS, TMPQ, adherence, and motivation questionnaire were calculated for each group, as well as the overall score as median and its IQR for the SUS and TMPQ. Between-group differences of the overall scores have been compared with the Wilcoxon rank sum test. To evaluate a potential association between the improvement in pain intensity and the TMPQ score, we used a simple linear regression model. Baseline demographics (age, pain intensity, and disability) have been compared between those who dropped out and those who completed the intervention with the Wilcoxon rank sum test. The level of statistical significance was set to 5% for all analyses. The analysis has been conducted with RStudio (R Foundation, version 4.3.1). The code is available on Open Science Framework [23].

Results

Patient Characteristics

Of 191 initially screened potential participants, 64 were randomized equally into the Physio+App and App group. Additionally, 29 participants were randomized into the Physio group, which was not included in this secondary analysis. From those randomized, 20 participants (12 female) from the Physio+App group and 18 (11 female) from the App group were followed up and included in this analysis. Figure 2 shows a flowchart of the enrollment process.

Figure 2. CONSORT flow diagram showing the flow of participants through this study. Only the "App only" and the "App+Physiotherapy" groups are included in this secondary analysis. The "Physiotherapy" group is included in the main analysis. The analysis was conducted based on outcomes that were collected at the end of the intervention. CONSORT: Consolidated Standards of Reporting Trials.

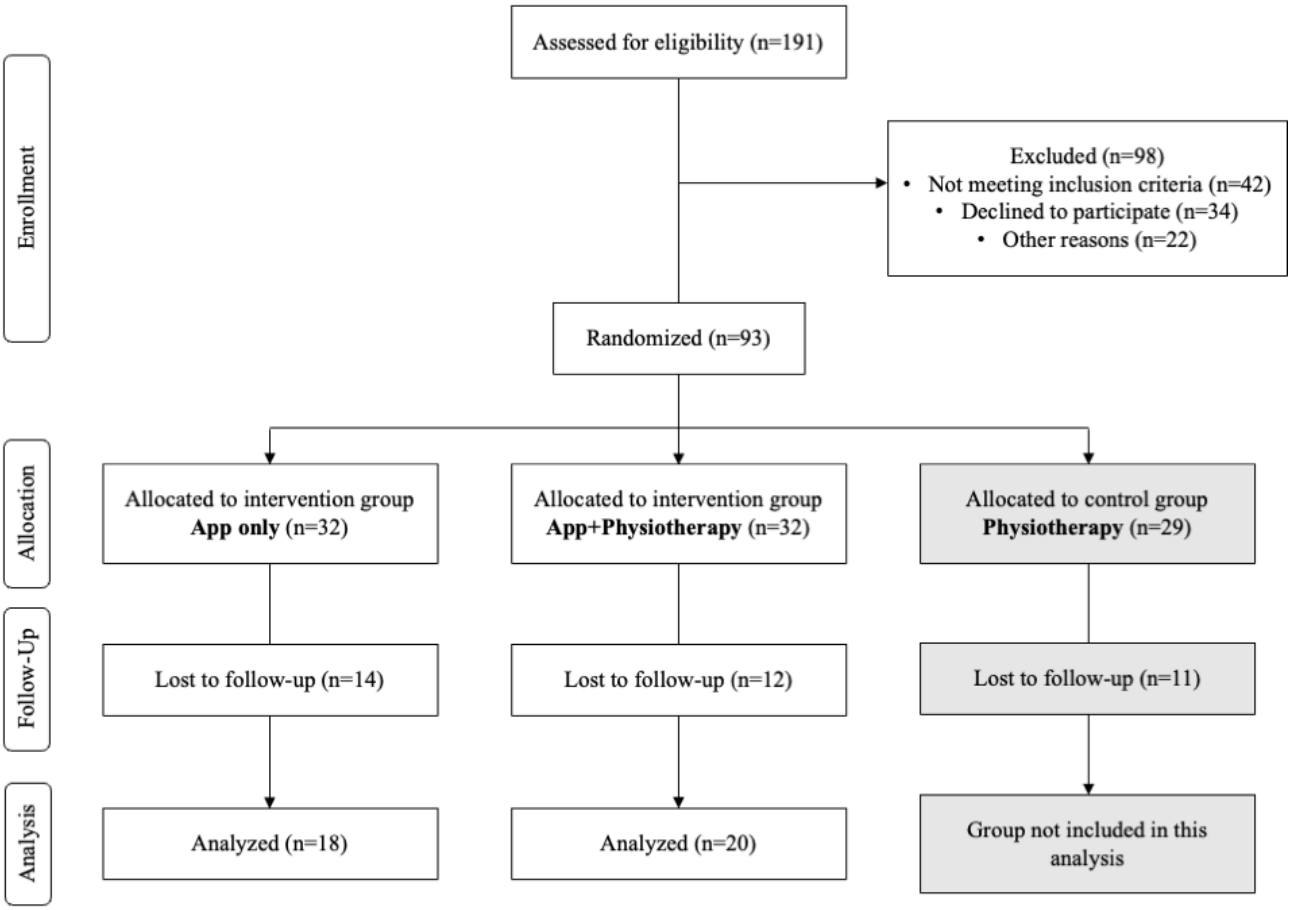

The mean age of all participants from those 2 groups who completed was 49.9 (SD 13.6) years. Most of the participants (29/38, 78%) had LBP for more than 2 years. Of the 38 participants, 15 (39%) participants currently used pain medication and 7 (18%) were on sick leave. The average pain intensity at baseline was 3.9 (SD 2.1) and the mean Oswestry Disability Index score 24.6 (SD 9.6). Baseline characteristics for each group are displayed in Table 1.

XSL•FO
RenderX

Table . Baseline characteristics obtained from the sample that was included in the secondary analysis on usability and acceptability of an app-based intervention. All data were self-reported and collected in web-based questionnaires.

Variable	Overall, N=38	App only, n=18	Physio+App, n=20
Gender, n (%)			
Female	23 (61)	11 (61)	12 (60)
Male	15 (39)	7 (39)	8 (40)
Age, mean (SD)	49.87 (13.55)	51.83 (14.09)	48.10 (13.15)
Height, mean (SD)	174.79 (10.59)	172.72 (10.09)	176.65 (10.95)
Weight, mean (SD)	84.34 (20.63)	82.28 (20.56)	86.20 (21.05)
Pain duration, n (%)			
1 month to 6 months	1 (2.7)	0 (0)	1 (5)
6 months to 1 year	4 (11)	2 (12)	2 (10)
1 to 2 years	3 (8.1)	3 (18)	0 (0)
2 to 5 years	10 (27)	3 (18)	7 (35)
More than 5 years	19 (51)	9 (53)	10 (50)
Medication use, n (%)	15 (39)	10 (56)	5 (25)
Current sick leave, n (%)	7 (18)	4 (22)	3 (15)
Pain intensity (NRS[a] 0 - 10), mean (SD)	3.92 (2.05)	3.67 (2.00)	4.15 (2.11)
Oswestry Disability Index, mean (SD)	24.58 (9.57)	24.43 (10.32)	24.72 (9.11)

[a]NRS: numeric rating scale.

Usability

The median score of the SUS among both groups was 82.5 (IQR 62.5 - 91.9), which indicates a good usability [24]. The between-group comparison of medians revealed no differences (Physio+App: 77.5, IQR 58.1 - 91.9; App: 86.3, IQR 65 - 90.6; *P*=.57). Table 2 displays the frequency of each answer on the Likert scale per question for each group. The inspection of single items revealed that most patients (20/38, 53%) strongly agreed on item 3 ("I thought the system was easy to use."). Furthermore, 24/38 (63%) strongly disagreed on item 4 ("I think that I would need the support of a technical person to be able to use this system.").

XSL•FO
RenderX

Table . Frequency of answers in the System Usability Scale that evaluates the usability of a mobile app, specifically the app "NOLA" used in this trial, which combines exercise and education for patients with low back pain. Each item could be answered with a numeric value between 0 ("do not agree") to 4 ("fully agree"). Between-group differences were evaluated using the Wilcoxon rank sum test.

		Between-group comparison		
Item[a]	Overall, N=38	App only, n=18	Physio+App, n=20	*P* value[b]
I think that I would like to use this system frequently, n (%)				.49
0	1 (2.6)	1 (5.6)	0 (0)	
1	4 (11)	1 (5.6)	3 (15)	
2	9 (24)	5 (28)	4 (20)	
3	12 (32)	7 (39)	5 (25)	
4	12 (32)	4 (22)	8 (40)	
I found the system unnecessarily complex, n (%)				.50
0	19 (50)	8 (44)	11 (55)	
1	6 (16)	3 (17)	3 (15)	
2	10 (26)	5 (28)	5 (25)	
3	2 (5.3)	2 (11)	0 (0)	
4	1 (2.6)	0 (0)	1 (5)	
I thought the system was easy to use, n (%)				.33
0	0 (0)	0 (0)	0 (0)	
1	1 (2.6)	1 (5.6)	0 (0)	
2	6 (16)	3 (17)	3 (15)	
3	11 (29)	6 (33)	5 (25)	
4	20 (53)	8 (44)	12 (60)	
I think that I would need the support of a technical person to be able to use this system, n (%)				.33
0	24 (63)	10 (56)	14 (70)	
1	5 (13)	3 (17)	2 (10)	
2	7 (18)	3 (17)	4 (20)	
3	1 (2.6)	1 (5.6)	0 (0)	
4	1 (2.6)	1 (5.6)	0 (0)	
I found the various functions in this system were well integrated, n (%)				.69
0	0 (0)	0 (0)	0 (0)	
1	0 (0)	0 (0)	0 (0)	
2	9 (24)	5 (28)	4 (20)	
3	18 (47)	8 (44)	10 (50)	
4	11 (29)	5 (28)	6 (30)	
I thought there was too much inconsistency in this system, n (%)				.59
0	14 (37)	6 (33)	8 (40)	
1	12 (32)	8 (44)	4 (20)	
2	9 (24)	4 (22)	5 (25)	
3	3 (7.9)	0 (0)	3 (15)	
4	0 (0)	0 (0)	0 (0)	
I would imagine that most people would learn to use this system very quickly, n (%)				.37
0	0 (0)	0 (0)	0 (0)	
1	1 (2.6)	1 (5.6)	0 (0)	

XSL•FO
RenderX

Item[a]	Overall, N=38	Between-group comparison		
		App only, n=18	Physio+App, n=20	*P* value[b]
2	5 (13)	2 (11)	3 (15)	
3	16 (42)	9 (50)	7 (35)	
4	16 (42)	6 (33)	10 (50)	
I found the system very cumbersome to use, n (%)				.35
0	18 (47)	7 (39)	11 (55)	
1	13 (34)	7 (39)	6 (30)	
2	5 (13)	3 (17)	2 (10)	
3	2 (5.3)	1 (5.6)	1 (5)	
4	0 (0)	0 (0)	0 (0)	
I felt very confident using the system, n (%)				.51
0	1 (2.6)	0 (0)	1 (5)	
1	3 (7.9)	3 (17)	0 (0)	
2	6 (16)	2 (11)	4 (20)	
3	13 (34)	7 (39)	6 (30)	
4	15 (39)	6 (33)	9 (45)	
I needed to learn a lot of things before I could get going with this system, n (%)				.96
0	17 (45)	8 (44)	9 (45)	
1	9 (24)	4 (22)	5 (25)	
2	6 (16)	4 (22)	2 (10)	
3	6 (16)	2 (11)	4 (20)	
4	0 (0)	0 (0)	0 (0)	

[a]0 = do not agree to 4=fully agree.

[b]Wilcoxon rank sum test.

Perceptions Toward Telemedicine

The median overall score in the TMPQ (0 - 60) was 45 (IQR 36.3 - 53.8) in the Physio+App group and 43.5 (IQR 36.5 - 50.5) in the App group (*P*=.25; Multimedia Appendix 3). Simple linear regression with within-group change of pain intensity (numeric rating scale) as an independent variable revealed a positive association with the TMPQ score ($F_{1,35}$=7.5, *P*=.01), but change of pain intensity could only explain 15.4% of the variance in the TMPQ score.

Figure 3 shows the frequency of each answer per question of the TMPQ. Participants in the Physio+App group were indecisive ("no opinion": 9/20, 45%) about whether they could always trust the technical equipment to work or not, while 9/18 (50%) participants in the App group agreed or strongly agreed with that statement. Further, 16/20 (80%) participants in the Physio+App group and 12/18 (67%) in the App group did not find the use of the necessary equipment difficult. A total of 6/20 (30%) participants in the Physio+App group and 11/18 (61.6%) in the App group agreed or strongly agreed to the statement that they do not like that telemedicine does not include physical contact. In line with this, only 3/20 (15%) participants in the Physio+App group and 2/18 (11%) in the App group agreed that telemedicine is as satisfying as a personal conversation with the therapist. Despite this, most participants (Physio+App: 12/20, 60%; App: 10/18, 56%) agreed that telemedicine is a convenient form of health care delivery, and 17/20 (85%) participants in the Physio+App group and 15/18 (84%) in the App group rated telemedicine as an adequate addition to usual care. Regarding the advantages of telemedicine, most participants (Physio+App: 15/20, 75%; App: 12/18, 67%) think that telemedicine can improve the general health, reduce health care costs (Physio+App: 16/20, 80%; App: 15/18, 83%), and save time (Physio+App: 17/20, 75%; App: 14/18, 78%). No tendency was found regarding the reduction of personal costs in both groups. Most participants (Physio+App: 17/20, 85%; App: 15/18, 84%) were not concerned about a violation of their privacy through a telemedicine app. Finally, 11/20 (55%) participants in the Physio+App group, but only 5/18 (28%) in the App group, agreed on the statement that telemedicine will be a standard way of health care delivery in the future.

XSL•FO
RenderX

Figure 3. Answers in the Telemedicine Perception Questionnaire per group. The questionnaire evaluates the perceptions of an individual toward telemedicine in general. The questionnaire was adapted to include only meaningful questions related to the intervention under evaluation in this trial. Participants could answer with "strongly disagree," "disagree," "no opinion," "agree," or "strongly agree" to each item. Answers are displayed as colored bars as well as percentages for each direction.

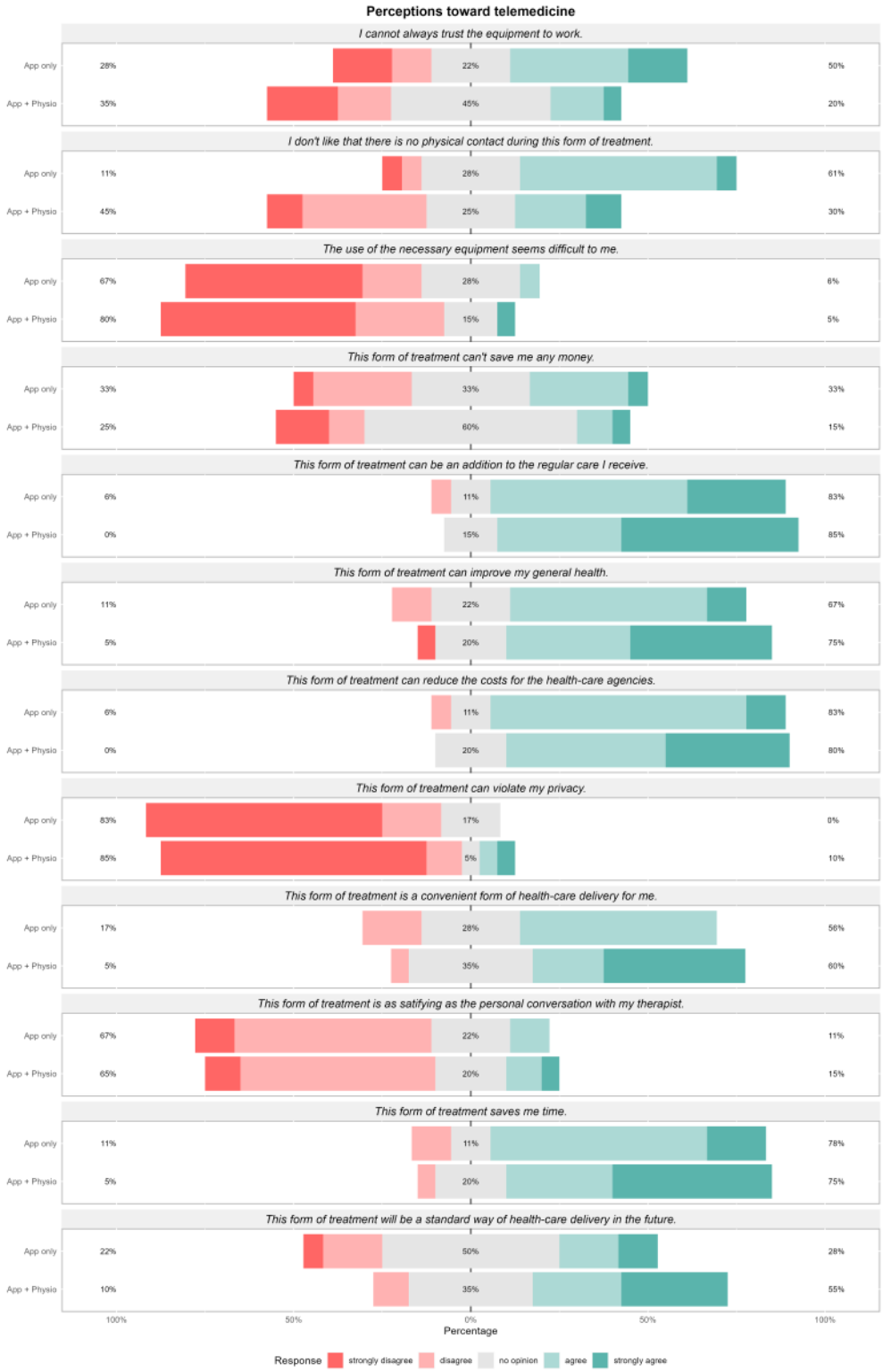

App Usage, Adherence, and Motivation

Thirty-seven participants self-reported their app usage. In the App group, 15/18 (84%) participants, and in the Physio+App group, 17/19 (90%) participants reached the goal of using the app at least 2 times per week (Table 3). Participants in the Physio+App group attended, on average, 5.3 (SD 1.2) sessions of physiotherapy.

XSL•FO
RenderX

Table . Adherence, app usage, and motivation to use the app "NOLA" in the context of the randomized controlled trial. As not all participants answered all questions, the number of valid answers per question is displayed.

Variable	Valid answers	App only, n=18	Physio+App, n=20	*P* Value[a]
Adherence, n (%)	38			.16
I was able to use the app as planned for 6 weeks at a time		7 (39)	12 (60)	
I had to interrupt app use due to illness		5 (28)	4 (20)	
I had to interrupt app use due to holidays		4 (22)	0 (0)	
I had to interrupt app use due to scheduling problems		3 (17)	0 (0)	
I had to interrupt app use due to an exacerbation of pain in my back		1 (5.6)	0 (0)	
I had to interrupt app use because of pain in another part of my body		0 (0)	2 (10)	
I had to interrupt app use due to other reasons		3 (17)	4 (20)	
How often have you trained your personal workout (back) with the app? n (%)	37			.25
Less than 2 days per week		3 (17)	2 (11)	
2 days per week		5 (28)	2 (11)	
3 - 5 days per week		7 (39)	9 (47)	
Nearly every day		3 (17)	6 (32)	
Think about the last 6 weeks: How motivated were you to open the app independently and train with it? n (%)	33			.80
I was not motivated at all, but thought of the results of the study		2 (13)	0 (0)	
I have found it very difficult to open the app regularly		0 (0)	0 (0)	
It took me a little effort to open the app at times		3 (19)	3 (18)	
I was mostly motivated		7 (44)	14 (82)	
I was always highly motivated to do the training frequently and regularly		4 (25)	0 (0)	
How would you rate the usefulness of the NOLA app in school grades? n (%)	35			.23
1		2 (13)	6 (32)	
2		10 (63)	10 (53)	
3		3 (19)	2 (11)	
4		1 (6.3)	0 (0)	
5		0 (0)	0 (0)	
6		0 (0)	1 (5.3)	

XSL•FO
RenderX

Variable	Valid answers	App only, n=18	Physio+App, n=20	*P* Value[a]
How strong is your desire to continue training with the app? n (%)	38			.23
No more need		2 (11)	0 (0)	
I'm still thinking about it, but the trend is more towards habit again (gym, sports club, rehab sports).		2 (11)	3 (15)	
I am still indecisive		7 (39)	6 (30)	
I would give it a chance, even if the app is chargeable		5 (28)	6 (30)	
I am very enthusiastic. I will definitely continue to use the app, as I see great progress		2 (11)	5 (25)	

[a]Wilcoxon rank sum test.

Of 38 participants, 12/20 (60%) in the Physio+App group and 7/18 (39%) in the App group indicated that they were able to use the app as planned for 6 weeks at a time, showing a higher compliance in the Physio+App group. The reasons for an interruption of the app use are displayed in Table 3.

Most participants (Physio+App: 14/16, 82%; App: 7/17, 44%) indicated they were mostly motivated to train with the app, while the motivation was higher in the Physio+App group. Although the usefulness of the app was rated in both groups with a median of 2 (median absolute deviation: 0, n=35), 6/20 (30%) participants in the Physio+App group and 7/18 (39%) in the App group were indecisive about the desire to train with the app in the future (Table 3).

Drop-Out Analysis

The group of participants who dropped out after randomization (n=26; 40%) showed statistically significant lower pain (dropouts: mean 3.0, SD 2.0, completers: mean 3.9, SD 2.0, *P*=.02) values at baseline compared to those who completed the intervention (n=38, 60%), but did not differ in disability and age (Table 4). From those participants who dropped out of the study (n=26), only 5 reported the reasons. Two participants dropped out for family reasons, 1 had an accident, 1 participant from the Physio+App group had no time for the physiotherapy sessions, and 1 indicated that the usage of the app was too complicated.

Table . Comparison of baseline values between those who dropped out and those who completed the intervention.

Variable	Dropouts (n=26)	Completers (n=38)	*P* value[a]
Age, mean (SD)	52.9 (12.0)	49.9 (13.5)	.35
Pain, mean (SD)	3.0 (2.0)	3.9 (2.0)	.02
Disability, mean (SD)	19.0 (9.4)	24.6 (9.6)	.09

[a]Wilcoxon rank sum test.

Discussion

Principal Findings

This preplanned secondary analysis shows that an app that delivered an exercise intervention with education showed good usability. It was also shown that participants with LBP have a generally positive perception of telemedicine, but it should be used as an adjunct to in-person therapy with physical contact. Although there was a tendency for participants who received the app as an adjunct to standard physiotherapy to perceive usability and telemedicine in general as better, there was no clear difference between the 2 groups. Motivation to use the app during the intervention phase was mostly high, with most patients using the app on average 3 to 5 days per week. However, most patients were indecisive about using the app in the future. Only a few patients reported the reasons for a dropout from the study, which might bias the results.

While several studies investigated patients' perceptions toward a specific telemedicine modality, the perception toward telemedicine in general of patients with LBP has not yet been evaluated extensively. Our study adds this knowledge about perceptions toward telemedicine in general after using a telemedicine approach alone and in addition to in-person physiotherapy. Barton et al [13] evaluated patients' perceptions regarding the use of physiotherapy via telemedicine approaches during the COVID-19 pandemic in patients with musculoskeletal pain. The results indicated that patients could accept telemedicine when there is an option to have regular in-person appointments with the therapist. This is in line with results from the TMPQ in our study, which showed that although participants were, in general, open toward telemedicine, they were missing the physical contact with a therapist. Similar findings are identifiable from a systematic review of qualitative studies for the population of chronic pain regarding perceptions toward a specific telemedicine intervention [10]. Our data showed that

XSL•FO
RenderX

participants who received only the app intervention were more likely to agree that they missed the physical contact in telemedicine interventions. Having physical contact regularly is also a critical facilitator for therapists [25]. Participants in our study believed that telemedicine can reduce health care costs and save time. This is in line with findings from qualitative studies on user experiences of telemedicine interventions in patients with LBP [26,27] and musculoskeletal pain [13].

Two qualitative studies showed that technical issues with an app-based intervention for LBP are an important factor that negatively influences the user experience [26,28]. The good usability rated in the SUS indicates that participants in our analysis did not have relevant technical issues, which could partly explain why perceptions toward telemedicine were mainly positive. Further, participants who received physiotherapy in addition to the app intervention, and thus had technical support throughout the whole intervention, trusted the equipment to work more than those who received only the app intervention. The usefulness of the app was rated with a median of 2 (IQR 2 - 2, on a scale from 1 to 6, with 1 being the best), which leads to the conclusion that the use of the app was well accepted among the participants who completed the intervention. This is in line with findings from Hasenöhrl et al [29] who evaluated the acceptance of therapeutic exercises delivered via an app.

Svendsen et al [26] found that decreased or increased pain could lead to not using the app as often as planned. In line with this, our dropout analysis showed that participants who dropped out showed lower levels of pain compared to those who completed. It can be assumed that participants dropped out because the burden of their LBP was not high enough to continue with the intervention.

Our study shows some limitations that must be considered. The dropout rate was relatively high in our study, but only a few patients reported reasons for dropout. It cannot be ruled out that some participants dropped out due to dissatisfaction with the app, which could bias results and does not allow for the clear conclusion that acceptability was high. We evaluated patients' perceptions with questionnaires and not with interviews, which could give a better insight into the user experience. As outcome measurements from the primary analysis were already time-consuming, we did not perform additional interviews in order to keep the motivation to participate in follow-up measurements high. Although the sample was relatively small, the included patients showed an age range where the prevalence of back pain is at its highest [30]. Regarding the adherence to the intervention and the app usage, we were not able to retrieve usage data directly from the app due to data protection reasons. As a consequence, we had to rely on self-reported data about adherence and app usage. It is known from other medical conditions that self-reported measures of adherence may overestimate the real effect [31].

The results of our study have implications for clinical practice. Patients seem to accept telemedicine approaches in general, but their use in combination with an in-person setting should be considered. Furthermore, it is important to ensure that technical issues are reduced and technical support is available to enhance the user experience and thus increase patient satisfaction.

Future research should focus on patients' perceptions regarding different delivery pathways, such as apps, websites, or videoconferencing, in order to obtain knowledge about preferences regarding specific forms of treatment. Additionally, it has to be investigated which factors influence adherence to telemedicine in a real-life setting, as well as which subgroups may show a higher acceptance and could thus benefit more from telemedicine interventions.

Conclusions

Our randomized controlled trial indicated good usability of the app "NOLA" regarding need for support, complexity, and training, as well as positive patient perceptions toward telemedicine in general in those patients who completed the intervention. This was independent of whether participants received only the app intervention or physiotherapy in addition to the app intervention. Despite concerns about the lack of physical contact, participants recognized telemedicine as a convenient addition to usual care. App usage and participant motivation were high, emphasizing the potential of such interventions in managing LBP. However, it has to be kept in mind that the high dropout rate might have been due to dissatisfaction with the intervention, which could bias our results. The findings from our study contribute insights to the evolving field of telemedicine, suggesting that well-designed apps could enhance patient engagement and satisfaction in patients with LBP. Further exploration and optimization of app-based interventions are needed for the effective integration into routine clinical practice.

Acknowledgments

We would like to thank Dr. Christopher Kaminski for providing access to the app and the participating physiotherapists from the practice "Schwerpunktpraxis für Human Movement" for conducting the physiotherapy intervention. Further, we would like to thank Dr. Sam Vickery for cross-checking the statistical code. No funding was received for this trial. Generative artificial intelligence was not used in any portion of this paper's writing.

Data Availability

The datasets generated or analyzed during this study are not publicly available as we have no consent to publish patient data, but are available from the corresponding author on reasonable, ethically conforming request. The code is available on the Open Science Framework [23].

XSL•FO
RenderX

Authors' Contributions

KT, CK, SK, DLB, CT, and CG contributed to the conceptualization of this study. KT, CK, SK, and CT developed the methodology. SK was responsible for software development, formal analysis, data curation, drafting the original paper, and creating visualizations. Sam Vickery, PhD, validated the results and outputs of this study. KT, CK, and SK carried out the investigation. SK, KT, CK, CG, CT, LD, and DLB reviewed and edited the writing. DLB and CT provided supervision. KT managed project administration. No external funding was acquired for this study.

Conflicts of Interest

None declared.

Multimedia Appendix 1
System Usability Scale - German adapted version.
[DOCX File, 19 KB - formative_v9i1e59866_app1.docx]

Multimedia Appendix 2
TMPQ – Telemedicine Perception Questionnaire – German adapted version. TMPQ: Telemedicine Perception Questionnaire.
[DOCX File, 20 KB - formative_v9i1e59866_app2.docx]

Multimedia Appendix 3
Answers in the Telemedicine Perception Questionnaire per item.
[DOCX File, 23 KB - formative_v9i1e59866_app3.docx]

Checklist 1
CONSORT-eHealth checklist (V 1.6.1).
[PDF File, 430 KB - formative_v9i1e59866_app4.pdf]

References

1. GBD 2021 Low Back Pain Collaborators. Global, regional, and national burden of low back pain, 1990–2020, its attributable risk factors, and projections to 2050: a systematic analysis of the global burden of disease study 2021. Lancet Rheumatol 2023 May:e316-e329. [doi: 10.1016/S2665-9913(23)00098-X]
2. Dagenais S, Hayflinger DC, Mayer JM. Economic evaluation of an extended telehealth worksite exercise intervention to reduce lost work time from low back pain in career firefighters. J Occup Rehabil 2021 Jun;31(2):431-443. [doi: 10.1007/s10926-020-09933-8] [Medline: 33394268]
3. Martin BI, Deyo RA, Mirza SK, et al. Expenditures and health status among adults with back and neck problems. JAMA 2008 Feb 13;299(6):656-664. [doi: 10.1001/jama.299.6.656] [Medline: 18270354]
4. Lin I, Wiles L, Waller R, et al. What does best practice care for musculoskeletal pain look like? Eleven consistent recommendations from high-quality clinical practice guidelines: systematic review. Br J Sports Med 2020 Jan;54(2):79-86. [doi: 10.1136/bjsports-2018-099878] [Medline: 30826805]
5. Lewis J, Mc Auliffe S, O'Sullivan K, O'Sullivan P, Whiteley R. Musculoskeletal physical therapy after COVID-19: time for a new "normal". J Orthop Sports Phys Ther 2021 Jan;51(1):5-7. [doi: 10.2519/jospt.2021.0102] [Medline: 33383997]
6. Fritz JM, Davis AF, Burgess DJ, et al. Pivoting to virtual delivery for managing chronic pain with nonpharmacological treatments: implications for pragmatic research. Pain 2021 Jun 1;162(6):1591-1596. [doi: 10.1097/j.pain.0000000000002139] [Medline: 33156148]
7. Lord Ferguson S. Is the end of the pandemic the end of telerehabilitation? Phys Ther 2022 Apr 1;102(4):pzac004. [doi: 10.1093/ptj/pzac004] [Medline: 35079821]
8. Report of the world physiotherapy/INPTRA digital physical therapy practice task force. World Physiotherapy. 2019. URL: https://world.physio/sites/default/files/2021-06/digital-practice-report-2021-FINAL.pdf
9. Kruse CS, Karem P, Shifflett K, Vegi L, Ravi K, Brooks M. Evaluating barriers to adopting telemedicine worldwide: a systematic review. J Telemed Telecare 2018 Jan;24(1):4-12. [doi: 10.1177/1357633X16674087] [Medline: 29320966]
10. Fernandes LG, Devan H, Fioratti I, Kamper SJ, Williams CM, Saragiotto BT. At my own pace, space, and place: a systematic review of qualitative studies of enablers and barriers to telehealth interventions for people with chronic pain. Pain 2022 Feb 1;163(2):e165-e181. [doi: 10.1097/j.pain.0000000000002364] [Medline: 34433776]
11. Kaczorowski S, Donath L, Owen PJ, et al. Telemedicine for patients with musculoskeletal pain lacks high-quality evidence on delivery modes and effectiveness: an umbrella review. Telemed J E Health 2024 May;30(5):1221-1238. [doi: 10.1089/tmj.2023.0255] [Medline: 38117672]

XSL•FO
RenderX

12. Moreno-Ligero M, Moral-Munoz JA, Salazar A, Failde I. mHealth intervention for improving pain, quality of life, and functional disability in patients with chronic pain: systematic review. JMIR mHealth uHealth 2023 Feb 2;11:e40844. [doi: 10.2196/40844] [Medline: 36729570]
13. Barton CJ, Ezzat AM, Merolli M, et al. “It’s second best”: a mixed-methods evaluation of the experiences and attitudes of people with musculoskeletal pain towards physiotherapist delivered telehealth during the COVID-19 pandemic. Musculoskelet Sci Pract 2022 Apr;58:102500. [doi: 10.1016/j.msksp.2021.102500] [Medline: 35074694]
14. Amin J, Ahmad B, Amin S, Siddiqui AA, Alam MK. Rehabilitation professional and patient satisfaction with telerehabilitation of musculoskeletal disorders: a systematic review. Biomed Res Int 2022;2022:7366063. [doi: 10.1155/2022/7366063] [Medline: 35958819]
15. Cottrell MA, Hill AJ, O’Leary SP, Raymer ME, Russell TG. Patients are willing to use telehealth for the multidisciplinary management of chronic musculoskeletal conditions: a cross-sectional survey. J Telemed Telecare 2018 Aug;24(7):445-452. [doi: 10.1177/1357633X17706605] [Medline: 28449620]
16. Eze ND, Mateus C, Hashiguchi TCO. Telemedicine in the OECD: an umbrella review of clinical and cost-effectiveness, patient experience and implementation. PLoS ONE 2020;15(8):e0237585. [doi: 10.1371/journal.pone.0237585] [Medline: 32790752]
17. Svendsen MJ, Wood KW, Kyle J, et al. Barriers and facilitators to patient uptake and utilisation of digital interventions for the self-management of low back pain: a systematic review of qualitative studies. BMJ Open 2020 Dec 12;10(12):e038800. [doi: 10.1136/bmjopen-2020-038800] [Medline: 33310794]
18. Thurnheer SE, Gravestock I, Pichierri G, Steurer J, Burgstaller JM. Benefits of mobile apps in pain management: systematic review. JMIR mHealth uHealth 2018 Oct 22;6(10):e11231. [doi: 10.2196/11231] [Medline: 30348633]
19. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126. [doi: 10.2196/jmir.1923] [Medline: 22209829]
20. Mannion AF, Junge A, Fairbank JCT, Dvorak J, Grob D. Development of a German version of the Oswestry Disability Index. Part 1: cross-cultural adaptation, reliability, and validity. Eur Spine J 2006 Jan;15(1):55-65. [doi: 10.1007/s00586-004-0815-0] [Medline: 15856341]
21. Brooke J. SUS: a quick and dirty usability scale. 1995 URL: https://digital.ahrq.gov/sites/default/files/docs/survey/systemusabilityscale%2528sus%2529_comp%255B1%255D.pdf [accessed 2025-08-05]
22. Altmann P, Ivkic D, Ponleitner M, et al. Individual perception of telehealth: validation of a German translation of the Telemedicine Perception Questionnaire and a derived short version. Int J Environ Res Public Health 2022 Jan 14;19(2):902. [doi: 10.3390/ijerph19020902] [Medline: 35055724]
23. Evaluation of the usability and acceptability of an app-based approach to treat low back pain: a pre-planned secondary analysis of a randomized controlled trial. OSF Home. URL: https://osf.io/eqzru/?view_only=ba0c7543870147f58e2db6e914bbfbaf [accessed 2025-08-05]
24. Bangor A, Kortum P, Miller J. Determining what individual SUS scores mean: adding an adjective rating scale. J Usability Stud 2009;4(3):114-123 [FREE Full text]
25. Weber F, Kloek C, Arntz A, Grüneberg C, Veenhof C. Blended care in patients with knee and hip osteoarthritis in physical therapy: Delphi study on needs and preconditions. JMIR Rehabil Assist Technol 2023 Jul 7;10:e43813. [doi: 10.2196/43813] [Medline: 37418301]
26. Svendsen MJ, Nicholl BI, Mair FS, Wood K, Rasmussen CDN, Stochkendahl MJ. One size does not fit all: participants’ experiences of the selfBACK app to support self-management of low back pain-a qualitative interview study. Chiropr Man Therap 2022 Oct 3;30(1):41. [doi: 10.1186/s12998-022-00452-2] [Medline: 36192724]
27. Skolasky RL, Kimball ER, Galyean P, et al. Identifying perceptions, experiences, and recommendations of telehealth physical therapy for patients with chronic low back pain: a mixed methods survey. Arch Phys Med Rehabil 2022 Oct;103(10):1935-1943. [doi: 10.1016/j.apmr.2022.06.006] [Medline: 35803329]
28. Nordstoga AL, Bach K, Sani S, et al. Usability and acceptability of an app (SELFBACK) to support self-management of low back pain: mixed methods study. JMIR Rehabil Assist Technol 2020 Sep 9;7(2):e18729. [doi: 10.2196/18729] [Medline: 32902393]
29. Hasenöhrl T, Windschnurer T, Dorotka R, Ambrozy C, Crevenna R. Prescription of individual therapeutic exercises via smartphone app for patients suffering from non-specific back pain: a qualitative feasibility and quantitative pilot study. Wien Klin Wochenschr 2020 Mar;132(5-6):115-123. [doi: 10.1007/s00508-020-01616-x] [Medline: 32060724]
30. Hoy D, Bain C, Williams G, et al. A systematic review of the global prevalence of low back pain. Arthritis Rheum 2012 Jun;64(6):2028-2037. [doi: 10.1002/art.34347] [Medline: 22231424]
31. Guedes VG, Corrente JE, Farrugia A, Thomas S, Wachholz PA, de Oliveira Vidal EI. Comparing objective and self-reported measures of adherence in haemophilia. Haemophilia 2019 Sep;25(5):821-830. [doi: 10.1111/hae.13811] [Medline: 31322311]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials

XSL•FO
RenderX

ICD-10**:** *International Statistical Classification of Diseases, Tenth Revision*
LBP: low back pain
SUS: System Usability Scale
TMPQ: Telemedicine Perception Questionnaire

Edited by A Mavragani; submitted 24.04.24; peer-reviewed by C Sitges, JB Lechauve; revised version received 10.06.25; accepted 14.06.25; published 25.08.25.

Please cite as:
Kaczorowski S, Trompeter K, Kramer C, Grüneberg C, Thiel C, Donath L, Belavy DL
Usability and Acceptability of an App-Based Approach to Treat Low Back Pain: Preplanned Secondary Analysis of a Randomized Controlled Trial
JMIR Form Res 2025;9:e59866
URL: https://formative.jmir.org/2025/1/e59866
doi:10.2196/59866

© Svenja Kaczorowski, Katharina Trompeter, Carolin Kramer, Christian Grüneberg, Christian Thiel, Lars Donath, Daniel L Belavy. Originally published in JMIR Formative Research (https://formative.jmir.org), 25.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Implementation Status and Usability of Digital Health Interventions Among Health Care Workers and End Users at the Primary Health Care Level in Chandigarh, North India: Cross-Sectional Study

Munita Jat*, BDS, MPH; Madhu Gupta*, MBBS, MD, PhD; Shankar Prinja, MD, DNB, MSc

Department of Community Medicine and School of Public Health, Post Graduate Institute of Medical Education and Research, RN Dogra Block, Sector 12, Chandigarh, India

*these authors contributed equally

Corresponding Author:
Madhu Gupta, MBBS, MD, PhD
Department of Community Medicine and School of Public Health, Post Graduate Institute of Medical Education and Research, RN Dogra Block, Sector 12, Chandigarh, India

Abstract

Background: Digital health interventions (DHIs) refer to the use of information and communication technologies to support or facilitate the achievement of health objectives. The Government of India has launched various DHIs at the primary health care level to improve health services and health-seeking behaviors. However, there is a paucity of evidence on the effectiveness of implementing these interventions and the user response from target end-users within the government health system setting.

Objective: This study aimed to assess the implementation status of DHIs and the user response of target end users, that is, the general population and health care workers (HCWs), in health and wellness centers (HWCs) in Chandigarh, India.

Methods: A cross-sectional study was conducted to assess the implementation status of 9 DHIs: the Electronic Vaccine Intelligence Network (eVIN), Reproductive and Child Health (RCH), Health Management Information System (HMIS), HWC portal, Comprehensive Primary Health Care–Noncommunicable Disease (CPHC-NCD), Family Planning–Logistics Management Information System (FP-LMIS), eSanjeevani, Integrated Disease Surveillance Program–Integrated Health Information Program (IDSP-IHIP) portal, Aarogya Setu, and the COVID-19 Vaccine Intelligence Network (CoWIN) app. Data were collected from 4 purposively selected HWCs using a pretested data extraction form and observation checklist from June to September 2022. The implementation status of these DHIs was evaluated by categorizing indicators into input, process, and output components and estimating cumulative percentage scores using a score-based logic model framework. Pretested interview schedules were used to assess awareness and user response of DHIs among 120 target end users (clients visiting HWCs) and 120 HCWs (auxiliary nurse midwives, data entry operators, and medical officers). The prevalence of user response was then estimated.

Results: The implementation status scores of the eVIN and RCH portals ranged from 70% to 90%. The HMIS portal, HWC portal, CPHC-NCD portal, and FP-LMIS scored between 25% and 50%, while eSanjeevani and the IDSP-IHIP portal scored between 51% and 70%. Community awareness of DHIs was poor, ranging from 1% to 18.3%, except for Aarogya Setu (94/120, 78.3%) and the CoWIN app (43/120, 35.8%), despite 86.7% (104/120) of participants having access to a mobile phone. Low awareness of DHIs was significantly associated with lower socioeconomic status (P=.02) and lower education levels (P=.04). In total, 66% (80/120) of HCWs reported that working with DHIs was easy; however, 89.2% (107/120) stated that dual data entry increased their workload. Frequent technical glitches were most commonly reported for the Auxiliary Nurse Midwife OnLine app (78/80, 97%) by HCWs. Help desk or feedback options in DHIs were rarely used by auxiliary nurse midwives/multipurpose workers (0% - 3.8%).

Conclusions: The RCH and eVIN portals were effectively implemented, eSanjeevani was moderately implemented, while the HMIS, HWC portal, CPHC, and FP-LMIS were poorly implemented. Community awareness of DHIs was low, except for the Aarogya Setu and CoWIN apps. Although HCWs found DHIs easy to use, increased workload due to dual data entry and frequent technical issues was a key concern.

(JMIR Form Res 2025;9:e69824) doi:10.2196/69824

KEYWORDS

digital health interventions; DHIs; health care workers; implementation; usability

XSL•FO
RenderX

Introduction

Primary health care is essential health care made universally accessible to individuals and families in the community by means acceptable to them, through their full participation, and at a cost that the community and country can afford, as per the World Health Organization (WHO) [1]. One of the 4 principles of primary health care is the use of appropriate technology, defined by WHO as "methods, procedures, technologies, and equipment which are scientifically validated, adapted to local needs, acceptable to users as well as recipients, and maintainable with local resources" [1]. The term digital health was coined as a "broad comprehensive term including eHealth, that is, the application of information and communication technology in health and health-related field, mobile health (the use of mobile and wireless technologies for medical care), as well as emerging areas such as the application of advanced computing sciences in "big data and artificial intelligence" [2]. Digital health interventions (DHIs) are services delivered electronically through formal or informal care, such as electronic medical health records, health management information systems (HMISs), mobile apps, and telemedicine. [3] In low- and middle-income countries, including India, the imbalance between demand and supply of health care services can be addressed through DHIs, as they can act as catalysts to provide affordable, accessible, and equitable health care services. Recognizing this, the Government of India embarked on the journey of digital health, as evident in the National Health Policy (NHP), National eHealth Authority (NeHA), and the National Digital Health Blueprint (NDHB). A flagship program, the National Digital Health Mission (NDHM), was launched on August 15, 2020, with a vision of establishing a digital ecosystem to achieve universal health coverage [4,5]. Over the past decade, the Ministry of Health and Family Welfare (MoHFW) has taken various digital initiatives for health information management, tracking of beneficiaries, establishing online inventory, and supply chain management. These also include platforms or mobile apps made to seek patient feedback, improve health literacy, and enhance the use of health services [6,7]. Recently, due to COVID-19, there was a major boost in the adoption of telemedicine services such as eSanjeevani OPD [8]. The Aarogya Setu mobile app, launched during the pandemic, also achieved the remarkable milestone of reaching 50 million downloads in just 13 days, making it the fastest-growing app globally [9]. As reported in various studies, the use of DHIs in primary health care has shown promising results in areas such as mental health issues, tobacco cessation programs, maintaining the vaccine supply chain, and hypertension management [10-12]. However, the implementation status of these DHIs at the primary health care level remains uncertain. A systematic review of previous studies reported barriers such as infrastructure and technical limitations, psychological and personal factors, and concerns regarding increasing working hours, which challenge the successful adoption of DHIs [13,14]. DHIs, as a tool to achieve universal health coverage, can be leveraged only if there is greater demand and use. This depends on factors such as the level of awareness among the community [15], clients' or health care providers' interaction with DHIs, and their belief that using technology will positively affect health [16]. However, there is little evidence on end-user perception (eg, satisfaction and motivation) of DHIs at the primary health care level, or awareness among the population in Chandigarh. The majority of available studies have evaluated either an independent intervention for a single condition or were part of nongovernmental organization or isolated projects. There is a paucity of evidence in the existing literature regarding the status of government-launched DHIs' implementation at the primary care level and the usability of these interventions among the community and health care workers (HCWs) in Chandigarh, India. Hence, this study was planned with the objectives to assess the implementation status of existing government-launched DHIs in public primary health centers and to ascertain the usability, in terms of user response, of DHIs by target end users (general population and clients) and HCWs while delivering health service in primary health centers in Chandigarh. The results of this study might guide program managers regarding the status and scope for improvement of the digitization of health interventions at the primary care level.

Methods

Operational Definitions

Operational definitions as per the WHO guidelines for monitoring and evaluation of DHIs are given below [16]:

1. Digital health: "The use of digital, mobile, and wireless technologies to support the achievement of health objectives. Digital health describes the general use of information and communication technologies for health and is inclusive of both mHealth and eHealth."
2. Usability: "The degree to which a product or system can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use." In this study, only part of usability was evaluated in terms of user response.
3. Users or target end users: "The individuals who directly employ the technology using their mobile phones or wireless devices either to deliver health services or to receive services."

Study Design and Settings

A cross-sectional study was conducted in Chandigarh, a union territory with an estimated population of 1,364,000 (2024). Approximately 97% of the population resides in urban areas. Chandigarh has 2 tertiary health care centers, 1 district hospital, 1 subdivisional hospital, 2 urban community health centers, and 34 health and wellness centers (HWCs; including 5 Ayurveda, Yoga and Naturopathy, Unani, Siddha, and Homeopathy centers) [17]. To assess the implementation status of existing government-launched DHIs, public primary health centers were selected. To ascertain the usability of DHIs in terms of user response by target end users, face-to-face interviews were conducted, and questions were administered in person to members of the general population visiting the study HWCs, as well as to the HCWs delivering health services in primary health centers in Chandigarh.

XSL•FO
RenderX

Study Health Facilities

To assess the implementation of DHIs, study health facilities were HWCs. Under the Ayushman Bharat Scheme, existing civil dispensaries and subcenters were upgraded to HWCs (recently renamed as Ayushman Arogya Mandir) [18]. These were established to provide universally accessible, free-of-cost, comprehensive primary health care. The primary focus is on promoting well-being and offering an extended range of services within the community. The comprehensive range of services now includes care in pregnancy and childbirth, neonatal and infant health care, adolescent health care, family planning, reproductive health care, management of communicable and noncommunicable diseases through National Health Programs, general outpatient care for acute simple illnesses, basic oral health care, care for common ophthalmic and ENT problems, elderly and palliative health care, emergency medical services, and screening with basic management of mental health ailments [17,18].

Study DHIs

Through a preliminary exercise, we enlisted health apps or portals that were implemented at HCWs and used by HCWs. For the general population, we enlisted health apps and portals that were launched for them and were functional. DHIs (apps or portals) that were included in the study were (1) Comprehensive Primary Health Care–Noncommunicable Disease (CPHC-NCD) portal; (2) Reproductive, and Child Health (RCH) portal; (3) Auxiliary Nurse Midwife OnLine (ANMOL) app; (4) Electronic Vaccine Intelligence Network (eVIN); (5) Family Planning–Logistics Management Information System (FP-LMIS); (6) eSanjeevani OPD; (7) Integrated Disease Surveillance Program–Integrated Health Information Program (IDSP-IHIP) portal; (8) Ayushman Bharat–Health and Wellness portal (AB-HWC portal); (9) HMIS portal; (10) Ni-kshay portal or app, (11) COVID-19 Vaccine Intelligence Network (CoWIN); (12) Arogya Setu app; (13) National Health Portal; and (14) My Health Records. The list of DHIs to assess the implementation and HCWs' and clients' responses is given in Table S1 in Multimedia Appendix 1.

Study Population

To assess the response of target end users, clients aged 18 years or older, irrespective of gender, who visited the study HWCs either seeking care for illness or as attendants, were contacted. HCWs, including data entry operators (DEOs), medical officers, auxiliary nurse midwives (ANMs), multipurpose workers (MPWs), and pharmacists, who were directly involved in the use and maintenance of DHIs, were enrolled in the study.

Sample Size

The sample size for assessing usability among both target end users and HCWs was calculated using the formula: $n=4pq/e^2$, where p=prevalence of usability of DHIs, which was assumed to be 50% (since there was a paucity of sufficient evidence in the existing literature on the prevalence of usability given the heterogeneity of selected DHIs, it was assumed to be 50% to estimate the sample size), q=(1p), and e=margin of error, assumed to be 10%. Considering a nonresponse rate of 20%, the final sample size was calculated as 120 for each group (target-end users and HCWs).

Sampling Technique

The criteria for the selection of HWCs were the mean number of patients seen per day. A list of HWCs with a mean number of patients seen per day was made, and 2 HWCs with high, 1 with moderate, and 1 with low patient footfall were randomly selected (a total of 4 HWCs). For target end-user enrollment, the catchment area of 1 HWC with moderate footfall was selected. Beneficiaries or persons who were visiting the facility and who consented were interviewed in person consecutively until the sample size was achieved. Similarly, HCWs in HWCs who were available and consented were interviewed in person consecutively. For this, we visited randomly selected 11 HWCs of Chandigarh.

Data Collection Tools and Methods

The data collection tools included a pretested observation checklist and data extraction form (Tool 1) to assess the implementation status of the DHIs, and pretested interview schedules (paper-based) for target end users (Tool 2) and HCWs (Tool 3) to assess the usability of DHIs. Tool 1 dealt with information on the functionality of DHIs, including technical factors such as connectivity, power, maintenance, data entry frequency, use, and management of currently operational DHIs. Portal-specific real-time reporting status, number of stock-out events, and number of clients enrolled under DHIs in the health facilities from April to June 2022 were collected. Data quality was checked by comparing the information recorded in the register (hard copies) with the corresponding data entries on the portal (DHI). The interview schedule for clients (Tool 2) was pretested in a health center in Haryana. During pretesting, it was found that there was low awareness of DHIs in the population except for the Aarogya Setu app, so questions were modified to first assess the awareness regarding DHIs. Tool 2 dealt with collecting information on the sociodemographic status of clients, language literacy, source of seeking health information, awareness of DHIs (ever heard of or aware of the mentioned DHIs), and whether they have visited portals or downloaded any mobile app from the listed DHIs. Data were collected using paper-based forms and entered into Epicollect 5 [19]. Tool 3 was used to collect information about the background characteristics of HCWs, the digital ecosystem in the working area, language literacy, training on DHIs, and reported responses on the usability of DHI with options such as "easy to use," "motivated/intending to use DHIs," "satisfaction with DHIs," and "several minutes/hours spent daily on the portal." Responses from Tool 3 were entered in an internet-based format (Google Forms).

Data Analysis

Data were entered in Microsoft Excel, and analysis was done using SPSS (trial version 25.0, IBM Corp) [20]. Descriptive statistics such as numbers and frequencies were calculated for awareness, apps downloaded, language proficiency, and user response in terms of ease of use, user satisfaction, and motivation. The chi-square test was applied to assess the association between sociodemographic factors and awareness

XSL•FO
RenderX

of DHIs among clients. A $P<.05$ was considered significant. A logical framework model having input, process, and output indicators was used to assess the implementation status of the DHIs. Care was taken while deciding on these indicators from the review of literature, respective DHI guidelines, and public health expert opinions. A score-based system was developed to categorize the implementation status of DHIs. The scoring criteria for each indicator using the logic model are given in Table S2 in Multimedia Appendix 1. After scoring each indicator, cumulative scores of DHIs were calculated in all 3 categories—that is, input, process, and output indicators—and measured in percentages. Equal weight was given to each category. The status of DHI implementation, as per percentage cumulative scores obtained (ie, cumulative scores obtained divided by maximum cumulative scores×100), was categorized as follows: "not implemented=0% - 24%," "poorly implemented=25% - 50%," "fairly implemented=51% - 70%," "effectively implemented=71% - 90%," and "very effectively implemented=100% - 91%."

Ethical Considerations

For this observational cross-sectional study, ethical clearance was obtained from the Institute's Ethics Committee (IEC), Postgraduate Institute of Medical Education and Research (PGIMER), on March 17, 2022 (reference no IEC-INT/2022/MPH-59). Permission was obtained from the Chandigarh administration for the participation of HCWs. All ethical criteria were fulfilled before, during, and after data collection. Informed consent was obtained from participants before the interviews using a written informed consent form. A participant information sheet was given to all the participants before participation in the study. Personal information of participants was kept confidential, and there was no breach of confidentiality. The study data were anonymous. As this was an observational study, no compensation was provided, as per IEC guidelines. No identification of individual participants or users in any images in the paper or Multimedia Appendix was possible.

Results

Implementation Status of DHIs at the Primary Care Level

Regarding the implementation status of DHIs, eVIN and RCH were found to be effectively implemented (70% - 90%), whereas ANMOL was not implemented. HMIS, HWC, CPHC-NCD app, and FP-LMIS were found to be poorly implemented (25% - 50%). eSanjeevani and IDSP-IHIP portal fall in the category of fairly implemented (51% - 70%). HWC and HMIS were not operational in 2 HWCs because DEOs were not positioned at these centers. FP-LMIS and CPHC-NCD were in the poorly implemented category, as they were not operational due to technical issues or lack of training. Although inputs were available for the ANMOL app, it was not operational in any HWC because of technical issues. Regarding maturity level, all DHIs were in the mature category except the FP-LMIS app, eSanjeevani, and CPHC-NCD app, which were at the scale-up level (Table 1).

XSL•FO
RenderX

Table . Overall implementation and maturity status of 9 DHIs[a] at 4 health and wellness centers, obtained through data extraction sheet and observation checklist at the primary care level in Chandigarh, 2022.

Name of DHI	Total scores obtained per HWCs[b]				Sum of total score obtained/sum of maximum score	Cumulative score, %	Maturity stage	Inference
	HWC 1	HWC 2	HWC 3	HWC 4				
RCH[c] portal	24	21	24	24	93/116	80	6	Matured and effectively implemented
ANMOL[d]	3	3	3	3	12/64	18	6	Matured but not implemented
eVIN[e]	23	21	24	22	90/104	86	6	Matured and effectively implemented
IDSP-IHIP portal[f]	14	13	14	19	60/86	69	6	Matured and fairly implemented
FP-LMIS[g] app	10	1	6	2	19/52	36	5	Scale-up and poorly implemented
HMIS portal[h]	20	21	0	0	41/96	42	6	Matured and poorly implemented
HWC portal	19	18	0	0	37/96	38	6	Matured and poorly implemented
eSanjeevani	7	7	8	9	31/52	59	5	Scale-up and fairly implemented
CPHC-NCD app[i]	6	3	10	14	33/84	39	5	Scale-up and poorly implemented

[a]DHI: digital health intervention.
[b]HWC: health and wellness center.
[c]RCH: Reproductive and Child Health.
[d]ANMOL: Auxiliary Nurse Midwife Online.
[e]eVIN: Electronic Vaccine Intelligence Network.
[f]IDSP-IHIP: Integrated Disease Surveillance Program–Integrated Health Information Platform.
[g]FP-LMIS: Family Planning–Logistics Management Information System.
[h]HMIS: Health Management Information System.
[i]CPHC-NCD: Comprehensive Primary Health Care–Noncommunicable Disease.

Awareness and Users' Response Among Target End Users

Out of 120 target end users, 33.3% (40/120) were in the age group of 18 - 30 years and 30 - 45 years, and 86.7% (104/120) of participants owned a smartphone. Around 90% (108/115) of those with access reported having an adequate internet connection. There was low awareness (ever heard of DHIs or knew something about the respective DHI) of DHIs (range: 1%-18.3%; 1/120 to 22/120), except for Aarogya Setu (94/120, 78.3%) and the CoWIN app (43/120, 35.8%) in the community. Out of 94 participants, 56 (46.7%) had downloaded the Aarogya Setu app (Table 2). Clients with low socioeconomic status and low education were significantly more likely to be unaware of DHIs (P=.02 and P=.04, respectively). There was no statistically significant association between age, sex, and awareness of DHIs (Table 3). Since the awareness of DHIs was very low, we limited our questions to awareness only, and user response was not explored.

XSL•FO
RenderX

Table . Background characteristics of clients (target end users) visiting health and wellness centers, Chandigarh, 2022.

Parameter	Frequency (N=120), n (%)
Age (years)	
18 - 30	40 (33.3)
31 - 45	40 (33.3)
46 - 60	15 (12.5)
>60	25 (20.8)
Sex	
Male	71 (59.16)
Female	49 (40.83)
Socioeconomic status	
Upper	11 (9.2)
Upper middle	57 (47.5)
Lower middle	32 (26.7)
Upper lower	20 (16.7)
Education	
Illiterate or up to primary school	10 (8.3)
Up to high school or a diploma	55 (45.5)
Graduate or above	55 (45.5)
Area of residence	
Urban	78 (65.0)
Rural	3 (2.5)
Slum or slum rehabilitation	39 (32.5)
Occupation	
Unemployed or elementary work	75 (62.0)
Craft, trade-related work, or clerks	24 (19.8)
Professionals or managers	21 (17.4)
Experiencing from health condition (multiple responses, n=131)	
Hypertension	27 (22.5)
Tuberculosis	3 (2.5)
DM[a]	19 (15.8)
None of the above	79 (65.8)
Others	3 (2.5)
Language easily understood for health information (multiple responses, n=258)	
English	81 (67.5)
Hindi	120 (100.0)
Punjabi	57 (47.5)
Sources to receive health information (multiple responses, n=235)	
Mass media (including social media, Newspapers, and Television)	122 (51.9)
Public health facility	92 (39.1)
Web browsing	19 (8.1)
Others	2 (0.9)
Have smartphones that they use	104 (86.7)
Adequate internet connection (n=115)	108 (90.0)

XSL•FO
RenderX

Parameter	Frequency (N=120), n (%)
Awareness of DHIs[b]	
CoWIN[c]	43 (35.8)
Aarogya Setu	94 (78.3)
eSanjeevani OPD	22 (18.3)
National Health Portal	12 (10.0)
My Health Records	1 (0.8)
None of the above	19 (15.8)
Downloaded app	
CoWIN	6 (5)
Aarogya Setu	56 (46.7)
eSanjeevani OPD	3 (2.5)
Visited portal	
NHP[d] portal	3 (2.5)

[a]DM: diabetes mellitus.
[b]DHI: digital health intervention.
[c]CoWIN: COVID-19 Vaccine Intelligence Network.
[d]NHP: National Health Portal.

XSL•FO
RenderX

Table . Awareness of DHIs[a] among clients (target end users) by age, sex, education, and socioeconomic status at health and wellness centers, Chandigarh, 2022.

Parameter	Awareness of DHIs (aware of any DHI), n (%)			*P* value
	Yes (n=101)	No (n=19)	Total (N=120)	
Age group (years)				.85
18 - 30	33 (82.5)	7 (17.5)	40 (100)	
31 - 45	35 (87.5)	5 (12.5)	40 (100)	
46 - 60	13 (86.7)	2(13.3)	15 (100)	
>60	20 (80.0)	5 (20.0)	25 (100)	
Sex				.25
Male	62 (87.3)	9 (12.7)	71 (100)	
Female	39 (79.6)	10 (20.4)	49 (100)	
Area of residence				.07
Urban	70 (89.7)	8 (10.3)	78 (100.0)	
Rural	2 (66.7)	1 (33.3)	3 (100.0)	
Slum or slum rehabilitation	29 (74.4)	10 (25.6)	39 (100.0)	
Education				.04
Illiterate or up to primary school	6 (60.0)	4 (40.0)	10 (100.0)	
Up to middle school or diploma	45 (81.8)	10 (18.2)	55 (100.0)	
Graduate or above	50 (90.9)	5 (9.1)	55 (100.0)	
Socioeconomic status				.02
Upper	8 (72.7)	3 (27.3)	11(100.0)	
Upper middle	54 (94.7)	3 (5.3)	57 (100.0)	
Lower middle	25 (78.1)	7 (21.9)	32 (100.0)	
Upper lower	14 (70.0)	6 (30.0)	20 (100.0)	

[a]DHI: digital health intervention.

Status of HCWs' Usability

Nearly 49.2% (59/120) of HCWs were aged 31 to 40 years, 87.5% (105/120) were females, 66.6% (80/120) were working as ANMs and MPWs, 5.8% (7/120) as DEOs, and 10% (12/120) as medical officers (Table 4).

XSL•FO
RenderX

Table . Background characteristics of health care workers and reported or perceived digital ecosystem at primary health care level, Chandigarh, 2022.

Parameter	Number (N=120), n (%)
Age group (years)	
21 - 30	27 (22.5)
31 - 40	59 (49.2)
41 - 50	21 (17.5)
51 - 60	13 (10.8)
Sex	
Male	15 (12.5)
Female	105 (87.5)
Education level	
High school	14 (11.7)
Graduate or diploma	75 (62.5)
Postgraduation	31 (25.8)
Designation	
ANM[a]	67 (55.8)
MPW[b]	13 (10.8)
Staff nurse or LHV[c]	1 (0.8)
DEO[d]	7 (5.8)
Medical officer	12 (10)
STS[e] or NTEP[f]	8 (6.7)
Others	12 (10)
Language literacy (multiple responses, n=291)	
English	120 (100)
Hindi	107 (89.2)
Punjabi	64 (53.3)
DHIs[g] used by HCWs[h] (multiple responses, n=503)	
RCH[i] portal	82 (68.3)
ANMOL[j] portal	80 (66.7)
Ni-kshay	9 (7.5)
IDSP IHIP[k] portal	87 (72.5)
eVIN[l]	48 (40.0)
FP-LMIS[m]	42 (35.0)
HWC[n] portal	11 (9.2)
CoWIN[o]	7 (5.8)
CPHC-NCD[p]	71 (59.2)
HMIS[q]	11 (9.2)
eSanjeevani	15 (12.5)
Other (PMMVY[r] and eHospital)	40 (33.3)
Perceived adequacy: adequate number of computers/tablets at the health facility center	
Yes	102 (85.0)

XSL•FO
RenderX

Parameter	Number (N=120), n (%)
Strength or speed of internet at health facility	
No connectivity	4 (3.3)
Very poor	22 (18.3)
Poor	38 (31.7)
Satisfactory	49 (40.8)
Strong	7 (5.8)
Very strong	0 (0)
Ease of working on digital device as compared with paper system	
Very difficult	0 (0)
Difficult	23 (19.2)
Equal	17 (14.2)
Easy	74 (61.7)
Very easy	6 (5)
Agree that dual data entry has increased your workload	
Strongly agree	94 (78.3)
Agree	13 (10.8)
Neutral	2 (1.7)
Disagree	3 (2.5)
Strongly disagree	1 (0.8)
Preferred language for web portals or mobile health apps	
English	115 (95.8)
Hindi	5 (4.2)
Punjabi	0 (0)

[a]ANM: auxiliary nurse midwife.
[b]MPW: multipurpose worker.
[c]LHV: lady health visitor.
[d]DEO: data entry operator.
[e]STS: Senior Treatment Supervisor.
[f]NTEP: National Tuberculosis Elimination Program.
[g]DHI: digital health intervention.
[h]HCW: health care worker.
[i]RCH: Reproductive and Child Health.
[j]ANMOL: Auxiliary Nurse Midwife OnLine.
[k]IDSP IHIP: Integrated Disease Surveillance Platform–Integrated Health Information Platform.
[l]eVIN: Electronic Vaccine Intelligence Network.
[m]FP-LMIS: Family Planning–Logistics Management Information System.
[n]HWC: health and wellness center.
[o]CoWIN: COVID-19 Vaccine Intelligence Network.
[p]CPHC-NCD: Comprehensive Primary Health Care–Noncommunicable Disease.
[q]HMIS: Health Management Information System.
[r]PMMVY: Pradhan Mantri Matru Vandana Yojana.

In terms of the digital ecosystem, 85% (102/120) of participants reported an adequate number of computers or tablets at the workplace (perceived adequacy), and 41%(49/120) reported having satisfactory internet strength. In total, 66% (80/120) of HCWs rated working on DHIs as “easy,” and 89.2% (107/120) agreed that dual data entry has increased their workload. English (115/120, 95%) was the most preferred language while working on DHIs (Table 5).

XSL•FO
RenderX

Table . Usability of digital health interventions by health care workers at health and wellness centers (HWCs), Chandigarh, Union Territory, 2022.

Parameter	RCH[a] portal	ANMOL[b]	Ni-kshay	eVIN[c]	IDSP-IHIP[d]	CPHC-NCD[e] app	eSanjeevani	FP-LMIS[f]	HMIS[g]	HWC portal
Using it on daily or weekly basis, n (%)	85 (100)	80 (100)	9 (100)	82 (100)	101 (100)	94 (100)	15 (100)	82 (100)	__[h]	12 (100)
Yes, n (%)	82 (96.5)	17 (21.3)	9 (100)	47 (57.3)	82 (81.2)	70 (74.5)	15 (100)	40 (48.7)	10 (100)	8 (66.7)
Easy to use, n	82	78	9	47	82	70	15	40	10	8
Very difficult, n (%)	0 (0)	59 (75.6)	0 (0)	0 (0)	1 (1.2)	2 (2.9)	0 (0)	0 (0)	0 (0)	0 (0)
Difficult, n (%)	6 (7.3)	11 (14.1)	1 (11.1)	1 (2.12)	1 (1.2)	10 (14.3)	3 (20)	0 (0)	2 (20)	0 (0)
Equal, n (%)	7 (8.5)	4 (5.1)	1 (11.1)	1 (2.12)	6 (7.3)	22 (31.4)	3 (20)	0 (0)	3 (30)	0 (0)
Easy, n (%)	52 (63.4)	4 (5.1)	3 (33.3)	5 (10.6)	31 (37.8)	28 (40)	7 (46.7)	6 (15)	3 (30)	2 (25)
Very easy, n (%)	17 (20.7)	0 (0)	4 (44.4)	41 (87.2)	43 (52.4)	8 (11.4)	2 (13.3)	34 (85)	2 (20)	8 (75)
Easy to navigate different pages or task, n	82	78	9	47	82	70	15	40	10	8
Strongly disagree, n (%)	3 (3.7)	35 (53)	0 (0)	0 (0)	1 (1.2)	0 (0)	1 (6.7)	0 (0)	0 (0)	0 (0)
Disagree, n (%)	9 (11.0)	11 (16.7)	1 (11.1)	0 (0)	4 (4.9)	11 (15.7)	4 (26.7)	0 (0)	2 (20)	0 (0)
Neutral, n (%)	6 (7.3)	8 (12.1)	3 (33.3)	0 (0)	6 (7.3)	29 (41.4)	3 (20)	1 (2.5)	1 (10)	0 (0)
Agree, n (%)	56 (68.3)	10 (15.2)	3 (33.3)	12 (25.5)	38 (46.3)	23 (32.9)	4 (26.7)	11 (27.5)	6 (60)	4 (50)
Strongly agree, n (%)	8 (9.8)	2 (3)	2 (22.2)	35 (74.5)	33 (40.2)	7 (10)	3 (20)	28 (70)	1 (10)	4 (50)
Total minutes spent on data entry/working on portal in a day, n	82	17	9	47	82	70	15	40	10	8
0 - 30 minutes, n (%)	0 (0)	0 (0)	3 (33.3)	47 (100)	82 (100)	5 (7.1)	0 (0)	40 (100)	7 (70)	8 (100)
30 - 60 minutes, n (%)	25 (30.5)	7 (41.2)	4 (44.4)	0 (0)	0 (0)	25 (35.7)	8 (53.3)	0 (0)	0 (0)	0 (0)
60 - 90 minutes, n (%)	19 (23.2)	4 (23.5)	2 (22.2)	0 (0)	0 (0)	8 (11.4)	3 (20)	0 (0)	0 (0)	0 (0)
>90 minutes, n (%)	38 (46.3)	6 (35.3)	0 (0)	0 (0)	0 (0)	32 (45.7)	4 (26.7)	0 (0)	3 (30)	0 (0)

XSL•FO
RenderX

Parameter	RCH[a] portal	ANMOL[b]	Ni-kshay	eVIN[c]	IDSP-IHIP[d]	CPHC-NCD[e] app	eSanjeevani	FP-LMIS[f]	HMIS[g]	HWC portal
Received training, n[i]	82	78	9	47	82	70	15	40	10	8
Yes, n (%)	66 (80.5)	78 (97.5)	8 (88.9)	45 (80.5)	79 (78.2)	70 (74.5)	8 (53.3)	37 (74.5)	7 (70)	5 (62.5)
No of days received training, n	66	78	8	45	79	70	8	37	6	5
1 - 2 days, n (%)	64 (97)	75 (96.2)	6 (75)	45 (100)	77 (97)	70 (100)	8 (100)	37 (100)	6 (85.7)	5 (100)
3 - 4 days, n (%)	2 (3)	3 (3.8)	2	0 (0)	2 (3)	0 (0)	0 (0)	0 (0)	1 (14.3)	0 (0)
1 week, n (%)	0 (0)	0 (0)	2 (25)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Used help desk or feedback option, n	82	78	9	47	82	70	15	40	10	8
Yes, n (%)	0 (0)	3 (3.8)	4 (44.4)	3 (6.4)	2 (2.4)	1 (1.4)	0 (0)	0 (0)	0 (0)	0 (0)
No, n (%)	82 (100)	76 (96.2)	5 (55.5)	44 (93.6)	80 (97.6)	69 (98.6)	15 (100)	40 (100)	10 (100)	8 (100)
Technical glitches while using portal, n	82	78	9	47	82	70	15	40	10	8
Never, n (%)	0 (0)	0 (0)	0 (0)	34 (70.8)	20 (24.4)	5 (7.1)	0 (0)	29 (72.5)	0 (0)	2 (25)
Rarely, n (%)	15 (18.3)	0 (0)	2 (22.2)	14 (29.2)	27 (32.9)	21 (30)	3 (20)	10 (25)	2 (20)	0 (0)
Sometimes, n (%)	47 (57.3)	2 (2.5)	4 (44.4)	0 (0)	18 (22)	23 (32.9)	7 (46.7)	1 (2.5)	5 (50)	2 (25)
Most of the time, n (%)	18 (22.0)	4 (5.0)	3 (33.3)	0 (0)	13 (15.9)	21 (30)	5 (33.3)	0 (0)	3 (30)	3 (37.5)
Always, n (%)	2 (2.4)	74 (92.5)	0 (0)	0 (0)	4 (4.9)	0 (0)	0 (0)	0	0 (0)	1 (12.5)
Clearly understand the language of portal, n	82	78	9	47	82	70	15	40	10	8
Yes, n (%)	81 (98.8)	76 (97.4)	8 (88.9)	46 (97.9)	80 (97.6)	70 (100)	15 (100)	40 (100)	5 (50)	8 (100)
Few words difficult to understand, n (%)	1 (1.2)	0 (0)	1 (11.1)	1 (2.1)	2 (2.4)	0 (0)	0 (0)	0 (0)	5 (50)	0 (0)
Perceive portal as useful										
Yes, n (%)	75 (91.5)	5 (6.3)	8 (88.9)	44 (93.6)	29 (35.4)	25 (35.7)	12 (80)	35 (87.5)	5 (50)	8 (100)
Not sure, n (%)	5 (6.1)	0 (0)	1 (11.1)	2 (4.3)	38 (46.3)	37 (52.9)	1 (6.7)	4 (10)	2 (20)	0 (0)

XSL•FO
RenderX

Parameter	RCH[a] portal	ANMOL[b]	Ni-kshay	eVIN[c]	IDSP-IHIP[d]	CPHC-NCD[e] app	eSanjeevani	FP-LMIS[f]	HMIS[g]	HWC portal
No, n (%)	2 (2.4)	75 (93.8)	0 (0)	1 (2.1)	15 (18.3)	8 (11.4)	2 (13.3)	1 (2.5)	3 (30)	0 (0)
Satisfied with functioning of portal, n	82	78	9	7	82	70	15	40	10	8
Yes, n (%)	56 (68.3)	1 (1.3)	5 (55.6)	44 (93.6)	56 (68.3)	25 (35.7)	5 (33.3)	37 (92.5)	5 (50)	5 (62.5)
Not sure, n (%)	17 (20.7)	61 (78.2)	2 (22.2)	3 (6.4)	14 (17.1)	28 (40)	7 (46.7)	3 (7.5)	4 (40)	2 (25)
No, n (%)	9 (11.0)	16 (20.5)	2 (22.2)	0 (0)	12 (14.6)	17 (24.3)	3 (20)	0 (0)	1 (10)	1 (12.5)
Motivated to use portal										
Yes, n (%)	77 (93.9)	1 (1.3)	9 (100)	44 (93.6)	26 (31.7)	15 (21.4)	6 (40)	32 (80)	6 (60)	8 (100)
Not sure, n (%)	2 (2.4)	1 (1.3)	0 (0)	3 (6.4)	26 (31.7)	34 (48.6)	5 (33.3)	8 (20)	3 (30)	0 (0)
No, n (%)	3 (3.7)	78 (97.5)	0 (0)	0 (0)	30 (36.6)	21 (30)	4 (26.7)	0 (0)	1(10)	0 (0)
Has helped in adhering to treatment protocol, n	82	78	9	47	82	70	15	40	10	8
Yes, n (%)	58 (70.7)	5 (6.3)	9 (100)	41 (87.2)	20 (24.4)	30 (42.9)	11 (73.3)	32 (80)	5 (50)	3 (37.5)
Sometimes, n (%)	19 (23.2)	11 (13.8)	0 (0)	2 (4.3)	46 (56.1)	22 (31.4)	3 (20)	5 (12.5)	5 (50)	5 (62.5)
No, n (%)	5 (6.1)	64 (80.0)	0 (0)	4 (8.5)	16 (19.5)	18 (25.7)	1 (6.7)	3 (7.5)	0 (0)	0 (0)

[a]RCH: Reproductive and Child Health.

[b]ANMOL: Auxiliary Nurse Midwife Online.

[c]eVIN: Electronic Vaccine Intelligence Network.

[d]IDSP-IHIP: Integrated Health Information Platform.

[e]CPHC-NCD: Comprehensive Primary Health Care–Noncommunicable Disease.

[f]FP-LMIS: Family Planning–Logistics Management Information System.

[g]HMIS: Health Management Information System.

[h]Not available.

[i]Where n is the total number of participants who received training for the app or portal among the total users.

ANMs and MPWs reported disagreement in most of the parameters for the ANMOL portal (78/80, 97%), mainly due to technical glitches. It was reported that even after entering data on the app, the data were not synced with the RCH portal. Therefore, they had to maintain similar details of one beneficiary at three places, that is, the RCH register, the RCH portal, and the ANMOL app. Triple data entry was the reason for lack of satisfaction (1/80, 1.3%), low motivation (1/80, 1.3%), and low perceived usefulness (5/80, 6.3%) of the ANMOL portal among ANMs. In total, 46.3% (38/85) of HCWs reported spending >90 minutes per day on entering data on the RCH portal. eSanjeevani, currently in the early phase of implementation, was used by medical officers (n=15). In addition, 60% (9/15) of them rated it as “easy to use,” but 33.3% (5/15) reported having technical issues (slow internet speed) most of the time. Help desk or feedback options were rarely used by health care providers for any of the DHIs. The most often cited reason for limited use of the feedback or help desk option was delayed responses. It was informed that verbal communication with superiors was the most common way to raise queries or provide feedback on DHIs. Technical issues were reported with all of the DHIs except for eVIN (34/82, 70% never; 14/82, 30% rarely). HCWs clearly understood the language of DHIs except for HMIS (5/10, 50%). ANMs or MPWs reported the IDPS portal (S form) as “easy to use” (74/80, 92%), and 66% (53/80) were satisfied with the functioning of the portal, albeit only

XSL•FO
RenderX

27% (23/80) of ANMs were motivated to use it. The primary reason was the belief among ANMs or MPWs that their primary duty was to provide MCH services, and that activities related to disease surveillance should be delegated to other staff (Table 5).

Discussion

Principal Findings

The results of this study highlighted that the RCH portal and eVIN app were effectively implemented; the IDSP-IHIP portal and eSanjeevani were fairly implemented, the HMIS portal, HWC portal, CPHC-NCD portal or app, and FP-LMIS were weakly implemented; and the ANMOL app was not operational in the study HWCs. There was very low awareness among clients (general population and beneficiaries) regarding DHIs launched by the Government of India to improve health literacy and health-seeking behavior (except for Aarogya Setu and the CoWIN app). This low awareness was found to be significantly associated with low socioeconomic status and low education level. The majority of the HCWs agreed on having an adequate digital ecosystem (number of computers, tablets, and connectivity) and considered operating data online to be easy. However, they reported increased workload due to dual data entry. Despite having a similar ecosystem at health care facilities, health care providers' responses to each DHI varied depending upon technical issues, the time consumed to enter data, and individual factors.

Effective implementation of eVIN at the primary health care level is similar to a recent study that reported improvement in vaccines cold chain management after implementation of eVIN [21]. This could be because of the better user interface of eVIN or the smaller number of data elements to be entered on the eVIN portal. The RCH portal, although effectively implemented, needed improvement in its output indicators. Although the RCH portal was supposed to be a cloud-based system, data interoperability was up to the state level only, while its usage was at the primary care level. A beneficiary in the second or third trimester migrating to a UT needed to reregister herself again on the portal. This was one of the reasons for the inability to achieve full registration within the first trimester. A study conducted in Uttar Pradesh and Rajasthan on the assessment of an earlier version of the RCH portal, that is, the Mother and Child Tracking System (MCTS) portal, reported poor data quality assessment (overall performance numbers were 34% for pregnant women and 33% for children in Rajasthan, while performance numbers were 18% for pregnant women and 25% for children in Uttar Pradesh) [22]. However, a study on the MCTS portal in the Ambala district of Haryana reported that HCWs perceived that MCTS has ensured better data quality [23]. Although HMIS has evolved into a robust system for health information management up to the primary level, due to a lack of dedicated DEOs, reports were sent in paper-based format to the district health administration at 2 HWCs. A study on the implementation status of HMIS in hospitals of central Ethiopia reported overall implementation as poor with a 56% overall percentage score [24]. It was even lower, that is, 42% in our study. The parameters used in the Ethiopia study (2021) were availability, compliance, and use of HMIS in tertiary care hospitals.

At the community level, this study reported low awareness of DHIs among clients despite having a smartphone and adequate internet connection, except for the Aarogya Setu app and CoWIN. This could be because of the enhanced awareness during the COVID-19 pandemic in the community due to widespread mass marketing of the Aarogya Setu app and regulations mandating its download while traveling during the pandemic [25]. Around 46.7% (56/120) of participants had downloaded the Aarogya Setu app in this study. Responses of users on usefulness (53%), easy to use (70%), and satisfaction with functionality (50%) were similar to a review study by Kodali et al [26], which reported higher satisfaction level (>50%), acceptance rate (80%), and perceived usefulness (72.8%) as important parameters among clients for DHI usability. A study conducted in Singapore on awareness of the COVID-19 pandemic and adoption of digital health services in urban Asian populations reported a high level of awareness of COVID-19 infection but low acceptance of digital health services. They also reported that older and lower-income participants were less likely to use digital health services [27]. This finding is similar to our study regarding lower awareness among people from low socioeconomic status; however, no significant association was noted with age. These results were also consistent with a systematic literature review by Reiners et al [28], which explored sociodemographic factors affecting the use of eHealth. It was reported that people with lower income, less education, and those living in rural areas were less likely to use eHealth. Tailored delivery of eHealth was recommended to improve usage [28]. Hence, to increase awareness and empower Indian citizens with DHIs, a push in terms of awareness camps and counseling at health facilities is needed.

Use of feedback or help desk option for the HCWs was limited despite facing several technical issues in running the portals, such as the ANMOL app and the RCH portal. This might be due to a defective or underperforming feedback resolution mechanism, leading to HCWs rarely using it in the future. Despite the majority of participants perceiving data entry into the digital system as easy, HCWs reported (89%, 107/120) an increase in workload due to dual data entry (both in manual and digital form) and considered working on DHIs to be extremely time-consuming. A study on the usage of a health information portal at the primary health care level in Haryana also reported an increased workload of health workers as a result of additional time spent on entering physical records. There were limited opportunities for using the data produced at the primary health care level in web-based platforms [29]. Another study conducted in Uttar Pradesh and Rajasthan on the assessment of the MCTS portal [22], and in Ambala district, Haryana [23], reported overburdened HCWs due to data documentation, which was identified as a barrier to implementation of MCTS.

Strengths

This is the first study to assess the implementation status of DHIs at the primary health care level using a logic framework approach and WHO guidelines on monitoring and evaluation

XSL•FO
RenderX

of DHIs in Chandigarh, India. Using input, process, and output parameters helped in identifying implementation gaps at these levels and providing specific recommendations that will strengthen the implementation of the DHIs at the primary care level. This study has also looked into the awareness and usability of digital health initiatives from the clients' perspective to improve health literacy, awareness, and health service usage. The sample size and sampling technique for assessing the usability of the HCWs are representative of Chandigarh city, and hence those results can be generalizable to the city.

Limitations

There might be limitations in deciding the input, process, and output indicators, which might influence the implementation status of the DHIs. However, care was taken while deciding on these indicators from the review of literature, respective DHIs' guidelines, and public health experts' opinions. Due to time constraints and feasibility issues, the implementation status of DHIs was assessed in 4 centers; hence, selection bias cannot be ruled out, which may limit the external validity. Each of these DHIs needs further exploration on implementation, barriers, and facilitators specific to a program by conducting a qualitative study and including a greater number of HWCs for the generalizability of results. The generalizability of the results pertaining to the clients' usability and awareness regarding DHIs is limited to the study area only. Due to feasibility and time constraints, we aimed only for descriptive analysis. The usability of DHIs by health care providers and associated factors can be further assessed more objectively in future studies.

Conclusions

Effective implementation and use are important for the sustainability and positive impact of DHIs. It is important to assess and monitor the currently operational DHIs in terms of inputs, processes, and outputs to identify gaps in implementation and ways to improve them. Also, how routinely HCWs use DHIs and their response to use will eventually affect the output, outcome, and impact of DHIs. As the Government of India is moving forward to a digitized health care system, it is important to know the awareness of DHIs at the community level, as they are the end-target users and receivers of digital health services. This study assessed the perspective of target-end users of DHIs. Awareness, acceptance, and demand for DHIs among marginalized people are of utmost importance for reducing health inequality and improving their health status. That will eventually help to achieve the Sustainable Development Goals and universal health coverage. The results of this study might guide policy makers to identify gaps and improve the implementation of DHIs by conducting enhanced training and timely refresher training sessions on the use or incentivization of HCWs, developing targeted awareness campaigns, or technical support initiatives. Further, adopting technological innovations such as automating data entry or improving system interfaces, along with a robust feedback or issues-resolution system, can minimize technological barriers. We recommend that there is a need for further in-depth assessment of barriers to implementation, awareness, and adoption of poorly performing government-launched DHIs, as well as evaluating long-term outcomes such as sustained adoption rates and improvement in health care delivery. This study also suggests the need to develop behavior-change communication interventions to increase the demand and uptake of DHIs in the community.

Acknowledgments

The authors are grateful to Ligia Paina, Associate Professor, Johns Hopkins Bloomberg School of Public Health, Baltimore, United States, for her permission and support in covering the article processing charges for this manuscript. They are also grateful to Dr Suman Singh, Director of Health Services, Chandigarh Administration, and Dr Charu, Nodal Officer, National Health Mission, Chandigarh Administration, UT Chandigarh, for providing permissions in collecting the data and supporting the study in Chandigarh. They are thankful to Dr Adarsh Bansal, PhD student at PGIMER, Chandigarh, for his support in data collection. Partial funding for the publication of this manuscript was provided by Johns Hopkins Bloomberg School of Public Health, Baltimore, United States, under the project titled "Common Good for Health: Reducing the Health Equity Gap." All authors declared that they had insufficient funding to support open access publication of this manuscript, including from affiliated organizations or institutions, funding agencies, or other organizations. JMIR Publications provided article processing fee (APF) support for the publication of this article.

Data Availability

The raw data generated during this study are not publicly available to maintain the privacy and confidentiality of the study participants.

Authors' Contributions

MJ contributed to conceptualization, data curation, formal analysis, investigation, methodology, project administration, visualization, and writing—original draft. MG contributed to conceptualization, formal analysis, methodology, resources, supervision, validation, and writing—review and editing. SP contributed to formal analysis, validation, and writing—review and editing.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
List of digital health interventions (DHIs) included in the study to assess implementation, health care worker, and client responses, and the indicators with corresponding scores used to assess the implementation status of DHIs in primary health care facilities. [DOCX File, 25 KB - formative_v9i1e69824_app1.docx]

References

1. World health organization (WHO). Alma-ata declaration, USSR. Presented at: International Conference on Primary Health Care, Alma-Ata, USSR; Sep 6-12, 1978; Alma-Ata, USSR URL: http://psycnet.apa.org/record/1969-15018-001
2. Recommendation on digital intervention for health system strengthening. World Health Organization. 2019. URL: https://www.who.int/publications/i/item/9789241550505 [accessed 2025-08-18]
3. Soobiah C, Cooper M, Kishimoto V, et al. Identifying optimal frameworks to implement or evaluate digital health interventions: a scoping review protocol. BMJ Open 2020 Aug 13;10(8):e037643. [doi: 10.1136/bmjopen-2020-037643] [Medline: 32792444]
4. National health portal. National Portal of India. URL: https://www.india.gov.in/national-health-portal [accessed 2025-08-17]
5. Gudi N, Lakiang T, Pattanshetty S, Sarbadhikari SN, John O. Challenges and prospects in india's digital health journey. Indian J Public Health 2021;65(2):209-212. [doi: 10.4103/ijph.IJPH_1446_20] [Medline: 34135195]
6. E-health. Ministry of Health and Family Welfare. URL: https://www.mohfw.gov.in/?q=en/Organisation/departments-health-and-family-welfare/e-Health-Telemedicine [accessed 2025-08-17]
7. An e-booklet on IT initiatives under national health mission. Ministry of Health and Family Welfare Government of India. 2020. URL: https://nhm.gov.in/uhc-day/Session 2/eBooklet on IT Initiatives of NHM.pdf [accessed 2025-08-19]
8. Health ministry's flagship telemedicine service - "eSanjeevani" records 3 crore tele-consultations. Press Information Bureau, Government of India. URL: https://pib.gov.in/PressReleasePage.aspx?PRID=1809569 [accessed 2025-08-18]
9. Aarogya Setu becomes world's fastest app to reach 50 mn downloads. The Hindu. URL: https://www.thehindu.com/sci-tech/technology/aarogya-setu-becomes-worlds-fastest-app-to-reach-50-mn-downloads-niti-aayog-ceo/article31347807.ece [accessed 2025-08-18]
10. Siopis G, Moschonis G, Eweka E, et al. Effectiveness, reach, uptake, and feasibility of digital health interventions for adults with hypertension: a systematic review and meta-analysis of randomised controlled trials. The Lancet Digital Health 2023 Mar;5(3):e144-e159. [doi: 10.1016/S2589-7500(23)00002-X]
11. Faujdar DS, Sahay S, Singh T, Kaur M, Kumar R. Field testing of a digital health information system for primary health care: a quasi-experimental study from India. Int J Med Inform 2020 Sep;141:104235. [doi: 10.1016/j.ijmedinf.2020.104235] [Medline: 32688290]
12. Electronic vaccine intelligence network (evin) has ensured essential immunization services during the COVID pandemic. Press Information Bureau, Government of India. 2020. URL: https://pib.gov.in/PressReleasePage.aspx?PRID=1643172 [accessed 2025-08-18]
13. Borges do Nascimento IJ, Abdulazeem H, Vasanthan LT, et al. Barriers and facilitators to utilizing digital health technologies by healthcare professionals. NPJ Digit Med 2023 Sep 18;6(1):161. [doi: 10.1038/s41746-023-00899-4] [Medline: 37723240]
14. Kaboré SS, Ngangue P, Soubeiga D, et al. Barriers and facilitators for the sustainability of digital health interventions in low and middle-income countries: a systematic review. Front Digit Health 2022;4:1014375. [doi: 10.3389/fdgth.2022.1014375] [Medline: 36518563]
15. Scott K, Ummer O, Chamberlain S, et al. At the frontlines of digitisation: a qualitative study on the challenges and opportunities in maintaining accurate, complete and timely digital health records in India's government health system. BMJ Open 2022 Feb 9;12(2):e051193. [doi: 10.1136/bmjopen-2021-051193] [Medline: 35140145]
16. Monitoring and evaluating digital health interventions. World Health Organization. 2016. URL: http://who.int/publications/i/item/9789241511766 [accessed 2025-08-18]
17. National Health Mission Chandigarh Administration. URL: http://nrhmchd.gov.in/ [accessed 2025-08-18]
18. Ayushman Arogya Mandir. URL: https://aam.mohfw.gov.in/ [accessed 2025-08-18]
19. Epicollect5. URL: https://five.epicollect.net/ [accessed 2025-08-18]
20. IBM SPSS software. IBM. URL: https://www.ibm.com/in-en/spss [accessed 2025-08-18]
21. Gurnani V, Singh P, Haldar P, et al. Programmatic assessment of electronic Vaccine Intelligence Network (eVIN). PLoS ONE 2020;15(11):e0241369. [doi: 10.1371/journal.pone.0241369] [Medline: 33151951]
22. Gera R, Muthusamy N, Bahulekar A, et al. An in-depth assessment of India's Mother and Child Tracking System (MCTS) in Rajasthan and Uttar Pradesh. BMC Health Serv Res 2015 Aug 11;15(1):315. [doi: 10.1186/s12913-015-0920-2] [Medline: 26259836]
23. Nagarajan P, Tripathy JP, Goel S. Is mother and child tracking system (MCTS) on the right track? An experience from a northern state of India. Indian J Public Health 2016;60(1):34-39. [doi: 10.4103/0019-557X.177298] [Medline: 26911215]

XSL•FO
RenderX

24. Bogale A. Implementation status of health management information system in hospitals of south west shoa zone, Oromia, Central Ethiopia. CEOR 2021;Volume 13:1-8. [doi: 10.2147/CEOR.S288998]
25. Aarogya Setu mandatory: who all must download the app right away. The Indian Express. URL: https://indianexpress.com/article/technology/social/aarogya-setu-app-mandatory-contact-tracing-app-6389284/ [accessed 2025-08-18]
26. Kodali PB, Hense S, Kopparty S, Kalapala GR, Haloi B. How Indians responded to the Arogya Setu app? Indian J Public Health 2020 Jun;64(6):228-230. [doi: 10.4103/ijph.IJPH_499_20] [Medline: 32496261]
27. Teo CL, Chee ML, Koh KH, et al. COVID-19 awareness, knowledge and perception towards digital health in an urban multi-ethnic Asian population. Sci Rep 2021 May 24;11(1):10795. [doi: 10.1038/s41598-021-90098-6] [Medline: 34031469]
28. Reiners F, Sturm J, Bouw LJW, Wouters EJM. Sociodemographic factors influencing the use of eHealth in people with chronic diseases. Int J Environ Res Public Health 2019 Feb 21;16(4):645. [doi: 10.3390/ijerph16040645] [Medline: 30795623]
29. Gupta S, Ramadass S, Ballabgarh Teaching and Research Group, et al. Use of web-based health information portals in primary health care: experience from a rural primary health centre in Haryana. J Family Med Prim Care 2021 Aug;10(8):3144-3150. [doi: 10.4103/jfmpc.jfmpc_2352_20] [Medline: 34660460]

Abbreviations

ANM: auxiliary nurse midwife
ANMOL: Auxiliary Nurse Midwife OnLine
CoWIN: COVID-19 Vaccine Intelligence Network
CPHC-NCD: Comprehensive Primary Health Care–Noncommunicable Disease
DEO: data entry operator
DHI: digital health intervention
eVIN: Electronic Vaccine Intelligence Network
FP-LMIS: Family Planning–Logistics Management Information System
HCW: health care worker
HMIS: Health Management Information System
HWC: health and wellness center
IDSP-IHIP: Integrated Disease Surveillance Program–Integrated Health Information Platform
MCTS: Mother and Child Tracking System
MPW: multipurpose worker
NDHB: National Digital Health Blueprint
NDHM: National Digital Health Mission
NeHA: National eHealth Authority
NHP: National Health Policy
RCH: Reproductive and Child Health
WHO: World Health Organization

Edited by A Mavragani; submitted 09.12.24; peer-reviewed by A Jamal, E Hannah; revised version received 29.04.25; accepted 15.05.25; published 25.08.25.

Please cite as:
Jat M, Gupta M, Prinja S
Implementation Status and Usability of Digital Health Interventions Among Health Care Workers and End Users at the Primary Health Care Level in Chandigarh, North India: Cross-Sectional Study
JMIR Form Res 2025;9:e69824
URL: https://formative.jmir.org/2025/1/e69824
doi:10.2196/69824

© Munita Jat, Madhu Gupta, Shankar Prinja. Originally published in JMIR Formative Research (https://formative.jmir.org), 25.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Encouraging General Practitioners to Refer Patients With Insomnia to a Digital Therapeutic (Sleepio): Feasibility Repeated-Measures Intervention Study

Ohoud Alkhaldi[1], BSc, MSc; Brian McMillan[2], BSSc, MBChB, PhD; John Ainsworth[1,3], MSc, PhD

[1]Division of Informatics, Imaging and Data Sciences, School of Health Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Vaughan House, Portsmouth St, Manchester, United Kingdom

[2]Centre for Primary Care and Health Services Research, Division of Population Health, Health Services Research and Primary Care, School of Health Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, United Kingdom

[3]National Institute of Health and Care Research (NIHR) Manchester Biomedical Research Centre, Manchester University Hospitals National Health Service (NHS) Foundation Trust, Manchester Academic Health Science Centre, Manchester, United Kingdom

Corresponding Author:
Ohoud Alkhaldi, BSc, MSc
Division of Informatics, Imaging and Data Sciences, School of Health Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Vaughan House, Portsmouth St, Manchester, United Kingdom

Abstract

Background: Sleepio, a digital therapeutic offering digital cognitive behavioral therapy for insomnia, has been recommended by the National Institute for Health and Care Excellence in the United Kingdom as an alternative to offering sleep hygiene or sleeping pills. However, understanding of the referral behavior of general practitioners (GPs) regarding Sleepio is lacking.

Objective: The aim of this study was to investigate the feasibility of using an intervention targeting GPs in Scotland to increase referrals of patients with insomnia to Sleepio.

Methods: GPs working in primary care in Scotland were invited to join the study. GPs were recruited through the Primary Care Research Network in Scotland from June 10, 2024, to October 13, 2024. The behavior change wheel (BCW) was used to inform the design of the intervention. During the intervention, GPs reviewed an orientation on using Sleepio and received a visual reminder midway through the intervention. The primary outcome was the number of Sleepio referrals every 2 weeks over 2 months. The secondary outcome was the change in the GPs' reported confidence level that Sleepio will be successful in reducing patients' insomnia symptoms, and confidence in recommending Sleepio to patients.

Results: Of the 23 GPs who joined the study, 16 completed all stages. Overall, 68.8% (11/16) of participants were females, and the mean age was 42 (SD 8) years. The total number of Sleepio referrals in 2 months was 96 for all 16 GPs. In the first 2 weeks of the intervention, the mean referral rate to Sleepio was 22.4% for all 16 GPs, but this rate increased to 45% by the end of week 8. A repeated measures analysis indicated there was no statistically significant difference in GPs' referral rates across 4 data points. GPs' reported confidence level in recommending Sleepio increased significantly ($z=-3.436$; $P<.001$), from a mean of 5.44 (SD 1.7; somewhat confident) to 8.13 (SD 2; very confident).

Conclusions: This study explored the feasibility and impact of an intervention aimed at supporting GPs to refer patients with insomnia to the digital therapeutic, Sleepio. Improvements were seen in GP-reported confidence levels at recommending Sleepio. A large-scale intervention and a longer study duration could provide useful information concerning how long the intervention effect on GPs' behavior toward Sleepio referrals might be maintained.

(JMIR Form Res 2025;9:e75359) doi:10.2196/75359

KEYWORDS

digital therapeutics; mobile app; insomnia; behavior change; primary care

Introduction

Background

The use of mobile apps and digital technologies in the medical field has become common [1]. Today, health care professionals (HCPs) use apps for educational purposes, especially for medical students [2,3] but also for other purposes, such as the delivery of patient care [4]. For example, mobile health apps have helped patients to engage with their health care by enabling them to track their physical activities, as well as diet, sleep, and other health metrics [5,6]. These apps have empowered patients to take control of their health, with potential beneficial results, including better health outcomes and lower health care costs. For these reasons, many countries are now increasingly

XSL•FO
RenderX

approving the use of mobile health apps and digital therapeutics [7,8]

A key difference between "health apps" and "digital therapeutics" is that the latter refers to evidence-based interventions that are clinically validated, often regulated, and designed to treat, manage, or prevent diseases [9].

Sleepio is a digital therapeutic using digital cognitive behavioral therapy for insomnia (CBTi) and has been recommended by the United Kingdom's National Institute for Health and Care Excellence (NICE) for treating patients with insomnia and insomnia symptoms [10]. The clinical evidence for Sleepio indicates that it is an effective treatment for insomnia and can improve insomnia symptoms and poor sleep [11]. NICE has concluded in-year cost savings over usual insomnia treatments in primary care [10]. If a medication is thought to be required to help with trouble sleeping, a brief course of benzodiazepines or Z drugs may be provided (treatment for insomnia). However, CBTi may be preferable to some medications that may be addictive [12]. According to the guidelines, CBTi is the first-line recommended treatment for insomnia, whereas sleep medications are not [13,14]. Therefore, as part of the Scottish Government's National Digital Mental Health Programme, Sleepio is currently available free to all adults across Scotland until 2030, through which patients can access the treatment directly by self-referral or can be signposted to it by their GPs [15].

Nevertheless, general practitioners (GPs) have been slow to adopt digital therapeutics. They have reported that some of the barriers that dissuade them from prescribing such treatments include a lack of awareness and knowledge, concerns about app quality and regulations, and limited training and support [16-18]. Consequently, several studies have been conducted to support GPs to incorporate digital therapeutics into their practices. A systematic review of interventions aimed at helping HCPs overcome barriers to prescribing digital therapeutics revealed that most interventions were successful in changing HCPs' behavior toward app prescriptions [19]. The most commonly reported outcome measures were changes in the numbers of prescriptions, changes in the levels of knowledge about health apps, or changes in the reported level of confidence about prescribing digital therapies. Restructuring the GPs' environment, such as offering technical resources [3,20] or providing education and training, was most likely to produce changes in HCPs' prescribing behavior [21,22].

Similar barriers have been reported in a survey of GPs in Scotland regarding referring their patients with insomnia to Sleepio [23]. This survey study showed that GPs would recommend Sleepio to their patients if they had better knowledge regarding the evidence confirming the benefits of Sleepio. GPs predicted that they would develop a habit of referring patients to Sleepio if they knew how to determine if someone would benefit from it. This interventional study was based on this existing literature, but it adopts a theory-driven approach that incorporates the behavior change wheel (BCW) to address some of the behavioral barriers to the adoption of Sleepio [24]. The BCW is an effective framework to systematically understand the target behavior and identify the barriers and enablers of the behavior. It consists of 8 steps to guide the design of a behavior change intervention. It starts with addressing Capability, Opportunity, and Motivation sources of Behavior (COM-B) model that represents the hub of the BCW. Then, it goes through the second and outer layer of the wheel by identifying intervention functions and policy categories. Once details from the BCW have been obtained, the appropriate behavior change techniques can be identified. The application of BCW could assist in the integration of Sleepio into clinical practice, thereby improving the uptake of NICE guidance and the benefits to patients and cost savings to the National Health Service identified therein.

Objective

This study aimed to explore the feasibility of using an intervention to support GPs to refer patients to a digital therapeutic (Sleepio) for the treatment of insomnia.

Methods

Study Participants

Inclusion criteria were currently practicing GPs (full time or part time) who worked in Scotland and provided informed consent. No exclusion criteria were applied. To facilitate participation, an internet-based study was conducted from June 10, 2024, to October 13, 2024. This decision was made because GPs have busy schedules, which were taken into account when designing the intervention by applying the 8 steps of the BCW. The mode of delivery was therefore chosen to be internet-based, at the individual level, and accessible via computers at a time convenient for the participants. This approach was intended to encourage GPs to participate by accommodating their workload and time constraints. Participants were recruited through the Primary Care Research Network in Scotland [25]. Interested GPs were asked to contact the research team via email.

Based on the practical considerations of recruiting viability and available resources, a target sample size of 12 participants was chosen based on recommendations in the literature for feasibility studies [26]. However, we exceeded this target and successfully recruited 16 GPs. This number was more likely to offer reliable information and enhance the robustness of the findings.

Study Intervention

The BCW framework was used to systematically understand the target behavior and inform intervention design in terms of targeting the COM-B system [27]. Our prior research followed the 8 steps of BCW to design this intervention [23]. It involved a survey of GPs working in primary care settings, identified their attitudes toward prescribing CBTi, and analyzed the sources of behavior using the COM-B self-evaluation questionnaire. Capability was addressed by introducing Sleepio to GPs and providing them with instructions for recommending and using Sleepio. Opportunity was targeted by addressing technical issues and solving accessibility problems with Sleepio. Motivation was addressed by supporting GPs to develop habits by asking them to report the number of referrals they made every 2 weeks and by sending them visual email reminders.

XSL•FO
RenderX

A preintervention questionnaire was sent to all study participants to collect demographic information and to identify GPs' current behavior and confidence toward referring patients to Sleepio (Multimedia Appendix 1).

At the start of the intervention, GPs were asked to review an orientation presentation on their own time regarding using the Sleepio digital therapeutic in patient care (Multimedia Appendix 2). The duration of the orientation presentation was approximately 11 minutes. The objective of the orientation was to enhance the GPs' confidence in referring patients to Sleepio by ensuring that they understood how Sleepio treats insomnia, the clinical credibility Sleepio has as a digital treatment, and how patients can have instant access to Sleepio.

The influence of the training on each GP's self-reported referrals to Sleepio was assessed every 2 weeks over 4 time periods (2, 4, 6, and 8 weeks) (Multimedia Appendix 3).

The participants received visual reminders to recommend Sleepio to their patients at 4 weeks after the beginning of the study (Multimedia Appendix 4). Ongoing technical support was available for the GPs to solve any technical issues during the intervention. The study period was 2 months.

Outcome Measures

The primary outcome measure was the change in self-reported Sleepio referrals over time, at 2, 4, 6, and 8 weeks after reviewing the orientation presentation about the use of Sleepio in patient care. The number of Sleepio referrals by each GP was collected every 2 weeks via short internet-based forms using Qualtrics [18].

The secondary outcome was the change in GPs' confidence in Sleepio as a successful treatment for insomnia and their confidence in recommending Sleepio to their patients before and after completing the intervention.

Data Analysis

Data were analyzed using SPSS Statistics (IBM Corp). Descriptive statistics, including mean and SD or median and IQR for continuous data and percentages for categorical data, were used. The primary outcome, which is the change in Sleepio referrals at 4 data points, was analyzed using the Friedman test. Changes in Likert-scale questions about confidence before and after the intervention were analyzed using the paired *t* test (2-tailed) and Wilcoxon signed-rank test.

Ethical Considerations

Ethics approval was obtained from the University of Manchester research committee (reference numbers 2023-18524-32221; 24/NRS/0015) and Health Research Authority (reference number 326797). All participants gave their informed consent for inclusion before they participated in the study. Compensation included £97 (US $124) for participating in the study. Data were anonymized for analysis to help preserve privacy and confidentiality. Participants were informed that participation in the research was voluntary and that they could withdraw their consent at any time without giving a reason.

Results

Overview

A total of 23 participants signed the consent form and joined the study. Out of 23 participants, 6 GPs dropped out before the start of the study, and another dropped out after the first data collection. Overall, 16 GPs completed the full intervention and became the participants of the statistical analysis. The mean age was 42 (SD 7.9) years, and 68.8% (11/16) of the GPs were female (Table 1). The participating GPs spent an average of 6 sessions in their practice per week. A session is the time GP spent on the provision of care [28].

XSL•FO
RenderX

Table . Demographics of general practitioners who completed the study (N=16).

Variables	Values
Gender, n (%)	
Male	5 (31.3)
Female	11 (68.8)
Age (years), mean (SD)	42.88 (7.9)
Number of sessions, mean (SD)	5.94 (1.8)
Size of the general practice, n (%)	
Small (<3000 patients)	4 (25)
Medium (3000 - 10,000 patients)	10 (62.5)
Large (>10,000 patients)	2 (12.5)
Patients presented with insomnia per week, n (%)	
<5	9 (56.3)
5 - 10	7 (43.8)
Prescribed digital CBTi[a] to treat insomnia in the past 12 months, n (%)	
Daily	0 (0)
Weekly	5 (31.3)
Monthly	4 (25)
Rarely	5 (31.3)
Never	2 (12.5)
Aware that Sleepio has been recommended by NICE[b], n (%)	
Yes	14 (87.5)
No	2 (12.5)
Any specific training on Sleepio, n (%)	
No	16 (100)

[a]CBTi: cognitive behavioral therapy for insomnia.

[b]NICE: National Institute for Health and Care Excellence.

Overall, 87% (14/16) of GPs reported that they were aware that NICE has recommended Sleepio as an alternative treatment to the usual insomnia treatments. None of the GPs had received any previous training about Sleepio.

Referrals

The total number of referrals made by GPs over the 2-month study period was 96. The number of referrals to Sleepio per GP was determined at each data collection point (ie, at 2, 4, 6, and 8 weeks; Figure 1). The mean referral rate to Sleepio at week 2 of the intervention was 22.4%, and this doubled to 47.14% at week 4. A drop in referral rates to almost 36% was noted in week 6. The last 2 weeks of the intervention had higher Sleepio referral rates, at 45% (Figure 2).

The Friedman test was conducted to determine if any significant differences were present in the number of referrals across the 4 time points (week 2, week 4, week 6, and week 8) during the intervention. The results indicated no statistically significant differences in the referral rates across the time points (χ^2_3=6.024; P=.11). This suggests that the intervention did not result in significant changes in referral numbers over the measured weeks.

XSL•FO
RenderX

Figure 1. Frequency of Sleepio referral per GP. GP: general practitioner.

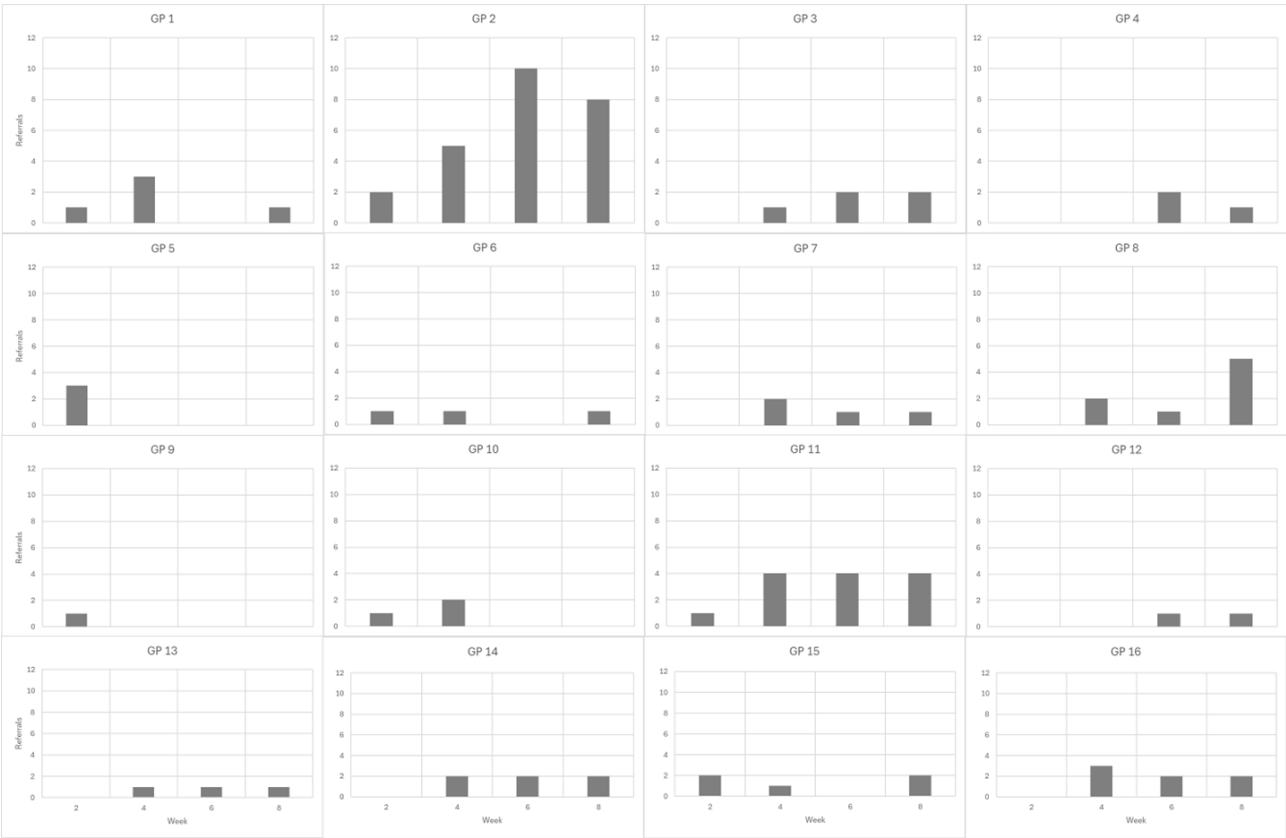

Figure 2. General practitioners' referral rates to Sleepio.

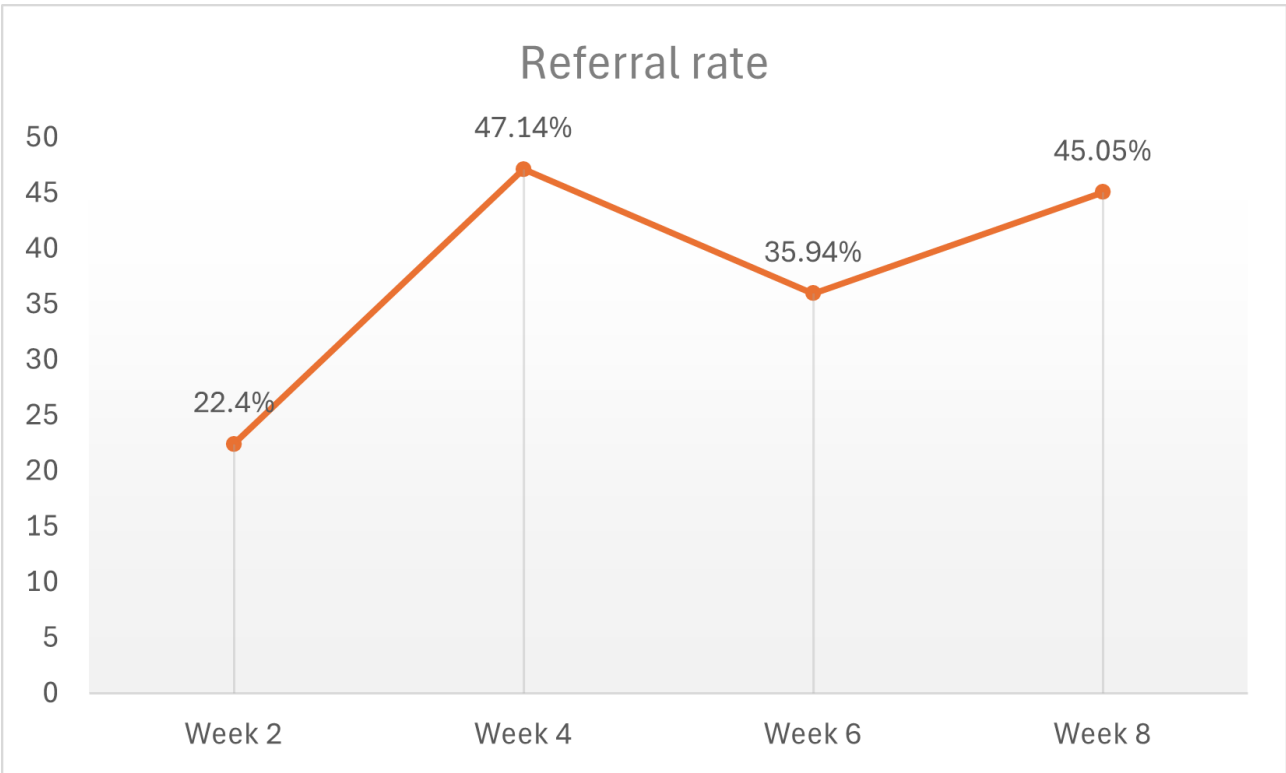

Confidence

The GPs showed significantly improved confidence in referring patients to Sleepio (Table 2). They reported that they were confident that Sleepio was a successful treatment for reducing insomnia symptoms. To measure the change in GPs' confidence level before and after the intervention, the paired *t* test was used. It revealed a significant increase in confidence level following the intervention, with a mean difference of 1.44 (SD 2, 95% CI 0.37-2.50; *P*=.01).

XSL•FO
RenderX

Regarding the GPs' reported confidence in recommending Sleepio to their patients, the Wilcoxon signed-ranks test showed a significant increase in confidence levels following the intervention (z=−3.436; P<.001), with 15 participants showing improved scores and 1 showing no change.

Table . General practitioners' confidence level before and after the intervention.

Confidence level	Baseline rating	Postintervention	*P* value
Digital CBTi[a] (Sleepio) is successful in reducing patients' insomnia symptoms, mean (SD)	5.06 (1.7)	6.50 (1.5)	.01
Recommending digital CBTi (Sleepio) to a patient[b]	5.44 (1.7)	8 (2)	<.001

[a]CBTi: cognitive behavioral therapy for insomnia.

[b]Baseline rating for recommending digital CBTi (Sleepio) to a patient is mean (SD); postintervention for recommending digital CBTi (Sleepio) to a patient is median (IQR).

Discussion

Principal Findings

The purpose of this study was to explore a feasibility intervention aimed at increasing the use of the Sleepio digital therapeutic in primary care. The GPs received orientation material at the start of the intervention about the use of Sleepio as a treatment for patients with insomnia, and they were asked to report the number of Sleepio referrals they made every 2 weeks.

Over the 2-month study, a total of 96 Sleepio referrals were made by the 16 participating GPs. This number was similar to some previously reported values [29] but lower than the number reported for other interventions in the literature [30,31]. The repeated self-reported Sleepio referrals revealed no significant differences between the 4 time point measurements. The observed variation in the fortnightly Sleepio referrals could be because of the timing during the year, as seasonal fluctuations in different factors, such as GP availability and patient consultation rates, could influence the patterns of referral to Sleepio [32]. Moreover, the behavior of patients with insomnia toward sleeping issues and barriers encountered when seeking treatment could also influence GPs' referrals to Sleepio [33.

Nevertheless, the intervention was clearly effective in improving the GPs' confidence in the Sleepio. This is promising, as the lack of confidence among GPs in discussing and recommending nonpharmacological treatments remains a significant barrier to the adoption of these treatments [34]. This reluctance originates from the GPs' limited knowledge or training in CBTi but is exacerbated by accessibility issues, as some GPs may not know how to refer patients to CBTi services. Therefore, many GPs may tend to prescribe short-term pharmacological treatments for their patients with insomnia rather than endeavoring to treat the underlying behavioral causes.

The discrepancy between the significant improvement in GPs' confidence in recommending Sleepio and a lack of statistically significant change in referral rates might be an indication of a typical time delay or lag between changes in attitudes and real behavioral changes in clinical practice. Although greater confidence is a sign that one is prepared to suggest digital CBTi, it sometimes can take more time and repeated reinforcement for behavioral changes to become completely incorporated in daily routine. In addition, it is likely that the improved confidence resulted in more suitable and selective referrals rather than increased frequency overall, especially if GPs were better able to identify which patients would benefit the most.

This study was built on our previous study that used the BCW to design a tailored intervention aimed at overcoming GPs' perceived barriers to referring patients with insomnia to Sleepio [23]. A key objective of this design was to enhance GPs' knowledge and confidence in making these referrals by incorporating targeted educational and practical components. The findings of this study demonstrate a significant improvement in the GPs' confidence levels, thereby validating the strategies proposed during the design phase. This alignment between the design objectives and the intervention outcomes underscores the utility of adopting a theory-driven approach, such as the BCW, to address specific behavioral barriers27.

Strengths and Limitations

This study is the first to test the feasibility of an intervention aimed at encouraging GPs in Scotland to refer patients with insomnia to Sleepio digital therapeutics. The design of the intervention was built on a behavior-based framework, which helped (1) to understand GPs' behavior toward using and recommending Sleepio and (2) to address GPs' reported barriers to Sleepio referrals. The study findings demonstrate the value of using the BCW to develop tailored interventions with GPs in primary care.

The limitations of this study include its small sample size and short duration, which limited the measurement of the long-term impact of the intervention. A potential Hawthorne effect may also have existed, as participating GPs may have been motivated to refer patients to Sleepio because they needed to report the number of referrals they made every 2 weeks. A follow-up assessment study to observe whether the impact of the intervention has been sustained would provide useful information.

GP recruitment and retention in interventional studies are challenging. To promote GPs' participation in this study, incentives were provided in return for their involvement in all research activities. While this approach may have helped recruitment, it also introduced a potential source of selection

XSL•FO
RenderX

bias as some participants may have been motivated more by the financial compensation rather than their interest in the research topic. Originally, the study design did not include compensation or monetary incentives for participants. However, the Primary Care Research Network required that GPs be reimbursed for their time and contributions as a condition for promoting the study through their network. As a result, compensation was added to the protocol. Researchers designing similar studies should consider alternative forms of incentives or nonmonetary compensation to minimize selection bias while still supporting GP engagement.

Another source of self-selection bias in this study is GPs' voluntary participation. Given their busy schedules and availability, GPs were invited to join the study voluntarily. While this approach facilitated recruitment, it has introduced self-selection bias, which means that those who joined the study may differ systematically from those who did not. For future research, randomization should be used to minimize this bias.

The primary outcome measure was based on the participants' self-reported data. Subjective measures such as these tend to have recall issues or social desirability bias. Collecting referral data from objective tools, such as the Sleepio website or a GP record system, was not possible in this study, because of the complexity of obtaining approval at the health board level.

A control group is necessary to provide a baseline for comparisons between GPs who would get the interventions and those who would not. As a result, the true impact of the intervention can be measured and compared between the 2 groups in terms of Sleepio referrals and confidence levels.

Conclusion

This study showed the feasibility of implementing an intervention that targeted GPs and aimed to improve their patient referrals to Sleepio and their confidence in using digital tools. The current approach of addressing GP-reported barriers to using digital therapeutics by providing educational materials and sending reminders was effective in changing the GPs' referral behavior. Future studies should endeavor to observe long-term changes in GPs' behavior toward Sleepio. The inclusion of patients to assess their perspective would also be beneficial.

Acknowledgments

This research is part of a PhD study sponsored by the Ministry of Education in Saudi Arabia. JA is funded by the National Institute for Health and Care Research (NIHR) Manchester Biomedical Research Centre. BM is funded by an NIHR Advanced Fellowship (reference number NIHR300887). The views expressed are those of the authors and not necessarily those of the National Health Service, the NIHR, or the Department of Health and Social Care.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Preintervention questionnaire assessing general practitioner demographics, insomnia management, and experience with Sleepio.
[DOCX File, 17 KB - formative_v9i1e75359_app1.docx]

Multimedia Appendix 2
Orientation material provided to general practitioners introducing Sleepio and its use in clinical practice.
[DOCX File, 641 KB - formative_v9i1e75359_app2.docx]

Multimedia Appendix 3
Self-report form completed every 2 weeks, recording the number of patients seen with insomnia and the number referred to Sleepio by general practitioners.
[DOCX File, 14 KB - formative_v9i1e75359_app3.docx]

Multimedia Appendix 4
Visual reminder sent to general practitioners midway through the intervention to encourage Sleepio referrals.
[DOCX File, 1551 KB - formative_v9i1e75359_app4.docx]

References

1. Ventola CL. Mobile devices and apps for health care professionals: uses and benefits. P T 2014 May;39(5):356-364. [Medline: 24883008]

XSL•FO
RenderX

2. Ellaway RH, Fink P, Graves L, Campbell A. Left to their own devices: medical learners' use of mobile technologies. Med Teach 2014 Feb;36(2):130-138. [doi: 10.3109/0142159X.2013.849800] [Medline: 24195445]
3. Rodder SG, Kindratt TB, Xiao C, et al. Teaching and evaluating smartphone applications: the effectiveness of a curriculum expansion. Educ Health (Abingdon) 2018;31(2):95-102. [doi: 10.4103/efh.EfH_322_16] [Medline: 30531051]
4. Wangler J, Jansky M. How can primary care benefit from digital health applications? – a quantitative, explorative survey on attitudes and experiences of general practitioners in Germany. BMC Digit Health 2024;2(1):14. [doi: 10.1186/s44247-024-00068-x]
5. Wang Q, Egelandsdal B, Amdam GV, Almli VL, Oostindjer M. Diet and physical activity apps: perceived effectiveness by app users. JMIR Mhealth Uhealth 2016 Apr 7;4(2):e33. [doi: 10.2196/mhealth.5114] [Medline: 27056639]
6. Tong HL, Maher C, Parker K, et al. The use of mobile apps and fitness trackers to promote healthy behaviors during COVID-19: a cross-sectional survey. PLOS Digit Health 2022 Aug;1(8):e0000087. [doi: 10.1371/journal.pdig.0000087] [Medline: 36812578]
7. Spahn: "digitale lösungen können den patienten-alltag konkret verbessern" [Web page in German]. Bundesministerium für Gesundheit. 2019. URL: https://www.bundesgesundheitsministerium.de/presse/pressemitteilungen/dvg-2/3-lesung.html [accessed 2025-01-11]
8. Iribarren SJ, Akande TO, Kamp KJ, Barry D, Kader YG, Suelzer E. Effectiveness of mobile apps to promote health and manage disease: systematic review and meta-analysis of randomized controlled trials. JMIR Mhealth Uhealth 2021 Jan 11;9(1):e21563. [doi: 10.2196/21563] [Medline: 33427672]
9. Lougheed T. How "digital therapeutics" differ from traditional health and wellness apps. CMAJ 2019 Oct 28;191(43):E1200-E1201. [doi: 10.1503/cmaj.1095801] [Medline: 31659065]
10. Sleepio to treat insomnia and insomnia symptoms. NICE. 2022. URL: https://www.nice.org.uk/guidance/mtg70 [accessed 2025-07-30]
11. Kyle SD, Hurry MED, Emsley R, et al. The effects of digital cognitive behavioral therapy for insomnia on cognitive function: a randomized controlled trial. Sleep 2020 Sep 14;43(9):zsaa034. [doi: 10.1093/sleep/zsaa034] [Medline: 32128593]
12. Schifano F, Chiappini S, Corkery JM, Guirguis A. An insight into Z-Drug abuse and dependence: an examination of reports to the European medicines agency database of suspected adverse drug reactions. Int J Neuropsychopharmacol 2019 Apr 1;22(4):270-277. [doi: 10.1093/ijnp/pyz007] [Medline: 30722037]
13. Riemann D, Espie CA, Altena E, et al. The European insomnia guideline: an update on the diagnosis and treatment of insomnia 2023. J Sleep Res 2023 Dec;32(6):e14035. [doi: 10.1111/jsr.14035] [Medline: 38016484]
14. Wilson S, Anderson K, Baldwin D, et al. British Association for Psychopharmacology consensus statement on evidence-based treatment of insomnia, parasomnias and circadian rhythm disorders: an update. J Psychopharmacol 2019 Aug;33(8):923-947. [doi: 10.1177/0269881119855343] [Medline: 31271339]
15. Scotland is first country to offer free digital therapies to every adult. Big Health. 2021. URL: https://www.bighealth.co.uk/news/scotland-is-first-country-to-offer-free-digital-therapies-to-every-adult [accessed 2025-07-30]
16. Sarradon-Eck A, Bouchez T, Auroy L, Schuers M, Darmon D. Attitudes of general practitioners toward prescription of mobile health apps: qualitative study. JMIR Mhealth Uhealth 2021 Mar 4;9(3):e21795. [doi: 10.2196/21795] [Medline: 33661123]
17. Giebel GD, Speckemeier C, Abels C, et al. Problems and barriers related to the use of digital health applications: scoping review. J Med Internet Res 2023 May 12;25:e43808. [doi: 10.2196/43808] [Medline: 37171838]
18. Gordon WJ, Landman A, Zhang H, Bates DW. Beyond validation: getting health apps into clinical practice. NPJ Digit Med 2020;3(1):14. [doi: 10.1038/s41746-019-0212-z] [Medline: 32047860]
19. Alkhaldi O, McMillan B, Maddah N, Ainsworth J. Interventions aimed at enhancing health care providers' behavior toward the prescription of mobile health apps: systematic review. JMIR Mhealth Uhealth 2023 Feb 27;11:e43561. [doi: 10.2196/43561] [Medline: 36848202]
20. Barnett J, Harricharan M, Fletcher D, Gilchrist B, Coughlan J. myPace: an integrative health platform for supporting weight loss and maintenance behaviors. IEEE J Biomed Health Inform 2015 Jan;19(1):109-116. [doi: 10.1109/JBHI.2014.2366832] [Medline: 25438329]
21. Chen J, Allman-Farinelli M. Impact of training and integration of apps into dietetic practice on dietitians' self-efficacy with using mobile health apps and patient satisfaction. JMIR Mhealth Uhealth 2019 Mar 4;7(3):e12349. [doi: 10.2196/12349] [Medline: 30829577]
22. PhD AC, Ciulla R, Edwards-Stewart A, Hoyt T, Bush N. Best practices of mobile health in clinical care: the development and evaluation of a competency-based provider training program. Prof Psychol Res Pract 2018 Oct 1;49:355-363. [doi: 10.1037/pro0000194]
23. Alkhaldi O, McMillan B, Ainsworth J. Using the behaviour change wheel to develop a tailored intervention to overcome general practitioners' perceived barriers to referring insomnia patients to digital therapeutic sleepio. BMC Health Serv Res 2024 Aug 22;24(1):967. [doi: 10.1186/s12913-024-11384-3] [Medline: 39174968]
24. The Behaviour Change Wheel: A Guide to Designing Interventions: Silverback; 2014. URL: https://www.behaviourchangewheel.com/ [accessed 2025-08-10]

XSL•FO
RenderX

25. Primary care. NHS Research Scotland. 2023. URL: https://www.nhsresearchscotland.org.uk/research-areas/primary-care [accessed 2023-12-18]
26. Julious SA. Sample size of 12 per group rule of thumb for a pilot study. Pharm Stat 2005 Oct;4(4):287-291. [doi: 10.1002/pst.185]
27. Michie S, van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. Implement Sci 2011 Apr 23;6(1):42. [doi: 10.1186/1748-5908-6-42] [Medline: 21513547]
28. GP sessions: a guide for locum and salaried GPs. GP World. 2023. URL: https://www.gpworld.co.uk/news/gp-sessions-a-guide-for-locum-and-salaried-gps/12/ [accessed 2025-02-09]
29. Lopez Segui F, Pratdepadua Bufill C, Abdon Gimenez N, Martinez Roldan J, Garcia Cuyas F. The prescription of mobile apps by primary care teams: a pilot project in Catalonia. JMIR Mhealth Uhealth 2018 Jun 21;6(6):e10701. [doi: 10.2196/10701] [Medline: 29929948]
30. Byambasuren O, Beller E, Hoffmann T, Glasziou P. mHealth app prescription in Australian general practice: pre-post study. JMIR Mhealth Uhealth 2020 Jun 1;8(6):e16497. [doi: 10.2196/16497] [Medline: 32478660]
31. Makhni S, Zlatopolsky R, Fasihuddin F, Aponte M, Rogers J, Atreja A. Usability and learnability of RxUniverse, an enterprise-wide app prescribing platform used in an academic tertiary care hospital. AMIA Annu Symp Proc 2017;2017:1225-1232. [Medline: 29854191]
32. How practices can plan for summer cover. Management in Practice. 2024. URL: https://managementinpractice.com/practice-intelligence/how-practices-can-plan-for-summer-cover/?utm_source=chatgpt.com [accessed 2025-01-12]
33. Stinson K, Tang NKY, Harvey AG. Barriers to treatment seeking in primary insomnia in the United Kingdom: a cross-sectional perspective. Sleep 2006 Dec;29(12):1643-1646. [doi: 10.1093/sleep/29.12.1643] [Medline: 17252896]
34. Everitt H, McDermott L, Leydon G, Yules H, Baldwin D, Little P. GPs' management strategies for patients with insomnia: a survey and qualitative interview study. Br J Gen Pract 2014 Feb;64(619):e112-e119. [doi: 10.3399/bjgp14X677176] [Medline: 24567616]

Abbreviations

BCW: behavior change wheel
CBTi: cognitive behavioral therapy for insomnia
COM-B: Capability, Opportunity, and Motivation sources of Behavior
GP: general practitioner
HCP: health care professional
NICE: National Institute for Health and Care Excellence

Edited by A Mavragani; submitted 01.04.25; peer-reviewed by D Sato, HC Moungui; revised version received 13.06.25; accepted 14.06.25; published 25.08.25.

Please cite as:
Alkhaldi O, McMillan B, Ainsworth J
Encouraging General Practitioners to Refer Patients With Insomnia to a Digital Therapeutic (Sleepio): Feasibility Repeated-Measures Intervention Study
JMIR Form Res 2025;9:e75359
URL: https://formative.jmir.org/2025/1/e75359
doi:10.2196/75359

© Ohoud Alkhaldi, Brian McMillan, John Ainsworth. Originally published in JMIR Formative Research (https://formative.jmir.org), 25.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Optimizing Intervention Components of a Preventive Stress Management mHealth Intervention for Health Care Workers: Experimental Factorial Study

Leo Kowalski[1], MS; Anna Finnes[2,3], PhD; Sabine Koch[1], PhD; Aleksandra Bujacz[1], PhD

[1]Health Informatics Centre, Department of Learning, Informatics, Management, and Ethics, Karolinska Institutet, Tomtebodavägen 18a, Stockholm, Sweden

[2]Division of Psychology, Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden

[3]Academic Primary Healthcare Centre, Region Stockholm, Stockholm, Sweden

Corresponding Author:
Leo Kowalski, MS
Health Informatics Centre, Department of Learning, Informatics, Management, and Ethics, Karolinska Institutet, Tomtebodavägen 18a, Stockholm, Sweden

Abstract

Background: Work stress is a prevalent risk factor for mental health problems, such as burnout and depression. Health care workers, especially during the COVID-19 pandemic, face high levels of work stress that make them a vulnerable population in need of support. Digital interventions are a promising way to combat this issue, offering the possibility of scalable programs that are easily accessible. While a wide range of stress management techniques can be incorporated into digital interventions, applying the multiphase optimization strategy enables systematic evaluation of what specific content most contributes to preventing the negative effects of work stress.

Objective: The primary aim of this research was to identify which digital intervention components and combinations of components are most effective at preventing symptoms of stress-related health problems. These insights are valuable to inform future intervention development for optimizing intervention design.

Methods: This study tested 5 digital intervention components aimed at improving stress management among workers. Engagement and Demands components allow participants to self-reflect on their work engagement and work challenges, while the Control component aids a more action-oriented process considering job crafting and detachment strategies. The Journaling component encourages a deeper reflection, and the Psychoeducation component provides evidence-based strategies for managing stress. In an experimental factorial study, Swedish health care workers (n=297) tested different versions of the intervention containing all possible combinations of these 5 components. Stress-related health outcomes, such as burnout, anxiety, and depression, were measured using questionnaires immediately before, immediately after, and 1 month after the end of the intervention.

Results: The most promising intervention effects were observed when the Demands and Control components were present together in the intervention. Including these components led to an increase in social support (β=0.68; $P<.001$) and job crafting (β=0.41; $P=.06$) during the intervention, as well as a decrease in symptoms of emotional exhaustion (β=−0.50; $P=.005$), burnout (β=−0.54; $P=.004$), and anxiety (β=−0.44; $P=.04$) after the intervention. Notably, including one of the components without the other made outcomes worse than including neither of these 2 components. Furthermore, mindfulness after the intervention increased when both the Engagement and Demands components (β=0.72; $P=.001$) were included as well as when the Journaling and Psychoeducation components were included (β=0.46; $P=.04$).

Conclusions: Results indicate that components aiding self-insight should be integrated with components providing actionable advice to achieve optimal intervention effects. Results from this optimization study may inform the development of preventive digital stress management interventions to be tested in future randomized controlled trials.

Trial Registration: ClinicalTrials.gov NCT04719351; https://clinicaltrials.gov/study/NCT04719351

(JMIR Form Res 2025;9:e71032) doi:10.2196/71032

KEYWORDS

digital behavior change intervention; mHealth; work stress; factorial experimental design; stress management; mental health; formative assessment; multiphase optimization strategy

XSL•FO
RenderX

Introduction

Work Stress

Stress has been termed the epidemic of the 21st century, reflecting both its widespread prevalence and adverse health effects [1,2]. Work is a common source of stress, with work-related stress defined as "the harmful physical and emotional responses that occur when the requirements of a job do not match the capabilities, resources, or needs of the worker" by the National Institute for Occupational Safety and Health [3]. This mismatch may have severe implications, as chronic exposure to work stress is a well-known risk factor for mental health problems, such as burnout and depression [4-6].

Alarmingly, work stress is very prevalent: a 2023 national survey by the American Psychological Association indicated that 77% of US workers experienced work stress in the previous month, with 57% reporting negative impacts because of this [7]. Beyond the detrimental effects of work stress on the individual level, stress-related health problems have a substantial economic impact as they are linked with increased sick leave, turnover rates, and productivity loss [8]. A 2024 report by the European Trade Union Institute estimates that costs due to depression attributable to psychosocial work exposures are upwards of US $114 billion in the European Union [9].

Health care workers are particularly vulnerable to work-related stress, with high levels of burnout and stress-related health problems within this population [10]. This issue was further exacerbated during the COVID-19 pandemic when health care workers faced unusually high work demands, such as increased workload and lack of equipment, leading to increases in mental health issues [11,12]. Proactive initiatives that support employees through primary prevention, intervening before work stress leads to health problems, are urgently needed to decrease individual issues and social costs.

Digital Interventions

Preventive interventions targeting health promotion may improve employee health and well-being, and by extension provide beneficial effects for organizations in terms of increased productivity, job satisfaction, and higher work engagement [13-15]. Digital interventions are among the most promising options for providing accessible and scalable support to workers and may constitute a valuable complement to other efforts. Compared with conventional on-site interventions, such as support groups and counseling, digital tools offer considerable advantages for standardizing interventions [16,17]. These solutions have been found effective for stress-related health outcomes, such as perceived stress, depressive symptoms, and emotional exhaustion within an organizational context [18-20].

Digital Behavior Change Interventions (DBCI) may be well-suited for addressing work stress. These interventions support users in adopting new behaviors and can be effective in a range of behavioral and health outcomes, including physical activity, dietary behavior, and improved mental well-being [21-23]. DBCIs may help workers develop a habit of preventive stress management techniques that mitigate the negative consequences of stress, thereby decreasing the risk of developing stress-related health problems [24,25].

Intervention Components

Overview

To develop a successful DBCI, it is necessary to choose what specific components to include, that is, distinct sets of content or functionality that support intervention aims. Common intervention approaches are often insufficient for determining what set of components provides optimal intervention effects. Complex interventions, for instance, cognitive behavioral therapy (CBT), integrate several components as part of the intervention, making it difficult to tease apart the effect of each individual component [26]. Other approaches, such as Mindfulness-Based Stress Reduction, function primarily through one component (mindfulness) without exploring the benefits of integrating other stress management techniques as part of the intervention [27]. While the following suggested components are evidence-based strategies for stress prevention, it is necessary to investigate the effectiveness of their combinations.

Stress Management for Work Stress

As a first step, it is necessary to choose a set of intervention components to consider. Given the intended intervention aims, components should serve to improve stress management for preventing the onset of stress-related health problems. While there is a wide range of evidence-based stress management techniques in academic literature known to alleviate stress reactions and improve mental health (eg, physical activity, mindfulness, and social support), some techniques are especially appropriate in the context of a digital intervention for workers [27-29].

First, it may be important for workers to self-monitor job demands, which are defined as aspects of the job that require sustained physical, cognitive, or emotional effort, to better understand what aspects of the work environment affect their stress levels and health [30]. Job demands are a well-known contributor to work stress, and self-monitoring may be an effective way to increase awareness of these. In addition, interventions targeting work engagement, a positive, energized, and focused state of mind, have been found to be effective in a range of relevant outcomes, including job satisfaction and well-being [15,31]. Combining work engagement elements with awareness of job demands could also be fruitful—this is similar to positive work reflection interventions, which can improve mental health in a working context [32,33].

Another strategy that may be effective is making employees more aware of how to exert control in their lives, both at work and during their free time [34]. Job crafting interventions, that is, programs that help employees shape their work to better fit their needs, are widely considered effective for combating work stress [35]. Likewise, learning to detach from work-related thoughts during leisure time is important for effective recovery [36]. This control component can potentially be combined with raising awareness of job demands. For instance, an intervention teaching job crafting strategies in relation to job demands had positive effects on need satisfaction and work engagement [37].

XSL•FO
RenderX

Journaling, a self-reflection practice in which one expresses thoughts and emotions in writing, is another technique that can have positive effects on mental health for workers. Dimitroff et al [38] found evidence that journaling may reduce burnout and compassion fatigue, while MacIsaac et al [39] suggest that journaling may increase self-reflection, which mediates improved well-being. In addition, incorporating psychoeducation, the structured provision of relevant evidence-based information, regarding recovery strategies may be helpful, as these behaviors are known to relieve short-term stress reactions [25,40]. As an example, a behavioral stress recovery intervention indicated a decrease in perceived stress and burnout symptoms [41]. Combining Psychoeducation with a Journaling component is common in therapeutic practices, such as CBT, and may yield a fruitful integration. Asplund et al [42] found positive effects on mental health and well-being from an intervention, including a stress diary and psychoeducation regarding recovery.

Multiphase Optimization Strategy

Research is needed to determine what components, Demands, Engagement, Control, Journaling, and Psychoeducation, are most effective in an mHealth intervention for work stress among health care workers. To guide decisions on what specific components to include, the Multiphase Optimization Strategy (MOST) offers a systematic and iterative approach to building and evaluating multicomponent interventions [43]. In the optimization phase of this framework, multiple components and their interactions are tested to better understand their effect on selected outcomes. Factorial experimental designs play a key role in helping to identify which components and combinations of components are most effective. Using these designs, it is possible to test the effect of multiple intervention components in the same trial rather than conducting separate experiments for each component [44].

Aims and Objectives

By identifying the most effective intervention components for stress management, we may improve digital interventions for work stress and thereby decrease stress-related health problems in the working population. Thus, the aim of the study was to investigate which intervention components and their combinations are most effective in preventing the negative effects of work stress.

1. What single components or specific combinations of components are most effective at preventing symptoms of stress-related health problems?
2. What is the effect of intervention components during as compared to after the intervention?

Methods

Participants

This study was conducted as part of a larger research project examining the impact of COVID-19 and psychological support initiatives on Swedish health care staff during the pandemic. The particular organization and timeframe were determined through a research partnership with a local government organization, following the urgent need for developing scalable psychological support solutions during the pandemic.

For recruitment, employees were initially informed by managers about the opportunity to participate and subsequently received an invitation email from the research team. As part of a related data collection in December 2020-January 2021, 1267 health care workers were asked in an electronic survey whether they wanted to take part in this study, to which 297 employees agreed. We used broad inclusion criteria to ensure representation across diverse roles and settings. Participants were eligible if they were currently employed in any capacity within the health care sector. Table 1 shows demographic characteristics of all participants.

XSL•FO
RenderX

Table . Participant demographics of Swedish health care workers immediately before a 1-month intervention (Preintervention), immediately after the intervention (Postintervention), and 1 month after the end of the intervention (Follow-up).[a] Occupation category "Other" included, for example, ambulance drivers, technicians, counselors, and physiotherapists. Burnout was measured by the Oldenburg Burnout Inventory. Depression was measured by the Patient Health Questionnaire. Posttraumatic stress disorder was measured by selected items from the Posttraumatic Stress Disorder Checklist. Anxiety was measured by Generalized Anxiety Disorder. Stress was measured by the Perceived Stress Scale. Mindfulness was measured by the Mindful Attention Awareness Scale. Job crafting was measured by Recovery Experience Questionnaire and Job Crafting Questionnaire. Social support was measured by the Job Crafting Questionnaire. Recovery was measured by the Recovery Experience Questionnaire. Emotional exhaustion was measured by the Shirom-Melamed Burnout Questionnaire.

Group	Preintervention measure (n=283)	Postintervention measure (n=218)	Follow-up measure (n=189)
Gender, n (%)			
Women	230 (82.4)	176 (81.1)	154 (81.9)
Men	49 (17.6)	41 (18.9)	34 (18.1)
Occupation, n (%)			
Assistant nurse	16 (5.7)	11 (5.1)	9 (4.8)
Nurse	74 (26.4)	47 (21.8)	43 (22.9)
Doctor	46 (16.4)	45 (20.8)	37 (19.7)
Admin	50 (17.9)	38 (17.6)	33 (17.6)
Other	94 (33.6)	75 (34.7)	66 (35.2)
Tenure (years)			
0 - 1	8 (3.0)	6 (2.9)	2 (1.1)
2 - 5	62 (23.0)	41 (19.6)	35 (19.3)
6 - 10	47 (17.4)	36 (17.2)	30 (16.7)
more than 11	153 (56.6)	126 (60.3)	114 (63.0)
Age (years), mean (SD)	45.4 (12.1)	46.8 (12.0)	47.3 (11.7)
Mental health (score range), n (%)			
Burnout (1-4)	2.42 (0.68)	2.36 (0.68)	2.25 (0.68)
Depression (1-4)	1.72 (0.74)	1.69 (0.73)	1.59 (0.70)
PTSD[b] (0 - 4)	2.06 (0.93)	1.97 (0.89)	1.79 (0.80)
Anxiety (1-4)	1.67 (0.61)	1.63 (0.55)	1.50 (0.48)
Stress (1-5)	3.00 (0.35)	2.99 (0.36)	2.94 (0.34)
Mindfulness (1-6)	2.97 (1.25)	3.11 (1.19)	3.00 (1.126)
Job crafting (1-7)	4.68 (0.91)	4.51 (1.01)	4.70 (0.97)
Social support (1-7)	4.23 (1.05)	4.37 (1.05)	4.28 (1.10)
Recovery (1-7)	4.41 (1.09)	4.58 (1.01)	4.66 (1.05)
Emotional exhaustion (1-7)	3.73 (1.44)	3.60 (1.44)	3.45 (1.48)

[a]Percentages are calculated based on the participants who actually responded to each category.

[b]PTSD: posttraumatic stress disorder.

Daily Intervention for Active Recovery Intervention

Daily Intervention for Active Recovery is a preventive mobile health intervention designed to improve employee stress management [45]. Daily Intervention for Active Recovery is developed by the research team, with this study being part of an iterative development process. During 4 weeks, users are prompted once daily to engage in a few minutes of self-reflection and psychoeducation. By building a habit of daily reflection and recovery, the intervention aims to support employees in shaping new and effective behavioral responses that mitigate the negative consequences of work stress and thereby prevent stress-related mental health problems.

The design of the Daily Intervention for Active Recovery is rooted in the science of behavior change and makes use of specific strategies to form new habits, such as those outlined in the Behavior Change Technique taxonomy [46]. For instance, self-monitoring of relevant behaviors and emotions serves as a tool for self-insight, which contributes to behavior change. Daily reminders in the app help to maintain engagement and are aligned with habit formation principles. Knowledge-shaping provides users with concrete behavioral or cognitive tools they

XSL•FO
RenderX

can use to improve daily recovery from work stress. This combination of self-monitoring, knowledge-shaping, and daily structure is intended to support the formation of new stress management behaviors.

This study investigates the effect of 5 intervention components (Engagement, Demands, Control, Journaling, and Psychoeducation) as well as 3 selected component combinations (Engagement and Demands, Demands and Control, Journaling and Psychoeducation). Table 2 shows each of the 5 components of Daily Intervention for Active Recovery and how they were operationalized, providing an example text used in the intervention.

Table . Description of Daily Intervention for Active Recovery intervention components including specific examples.

Intervention component	Description	Example text used in Daily Intervention for Active Recovery
Engagement	Prompts participants to reflect on 3 positive emotions they experienced during the last 24 hours related to engagement—asking whether they felt relaxed, engaged, and focused.	During the last 24 h, to what extent have you felt engaged? (1 – Not at all, 6 – Very much)
Demands	Prompts participants to reflect on various kinds of job demands they experienced during that day.	To what extent did you have to work under time pressure today? (1 – Not at all, 6 – Very much).
Control	Prompts participants to consider the ways in which they can exert control in their daily lives, both at work and during their free time through job crafting and detachment.	To what extent could you choose your pace of working today? (1 – Not at all, 6 – Very much)
Journaling	Allows participants to journal their thoughts and emotions in free-text form based on prompts related to work-related stress and recovery.	What do you do at work when you feel overwhelmed? (Free-text response)
Psychoeducation	Provides participants with evidence-based information regarding stress and recovery as well as suggesting specific behavioral strategies they may use to recover from daily stress more effectively. Strategies are sampled from the overall literature on stress management, including physical activity, mindfulness, and psychological detachment [24,26,27].	"Micro-breaks. Another way to recover from work stress is to take breaks during the day – moments of relaxation when you completely let go of work demands. However, when there is lots going on and we feel stressed, it can be difficult to find the time for longer breaks. Perhaps there are no clear opportunities for resting in between tasks. In these cases, it is especially important to do short interruptions – micro-breaks – to sit down, close your eyes, and breathe deeply for a minute."

Participants were randomized to experimental conditions using blocked randomization with a block size of 32 participants, ensuring equal distribution of participants across all conditions. The randomization sequence was computer-generated. All aspects of the randomization process, participant enrollment, sequence generation, and condition assignment were conducted by the first author. Participants were blinded to their experimental condition, not knowing what intervention version they had received.

Study Design

The study followed the preregistered design as described in ClinicalTrials.gov (registration NCT04719351) [47]. Data were collected between March 2021 and June 2021, with 3 identical surveys sent out during this time period (as depicted in Figure 1) : (1) preintervention measured immediately before the intervention, (2) postintervention measured immediately after the intervention, and (3) follow-up measured 1 month after finishing the intervention. All data were collected through the mobile app Daily Intervention for Active Recovery.

Figure 1. Flowchart showing the timeline of measures and intervention. Note: Some dates overlap or suggest a longer time interval due to participants starting the study at slightly different time points. There were 297 participants who took part in the study, where 283 completed the full preintervention measure.

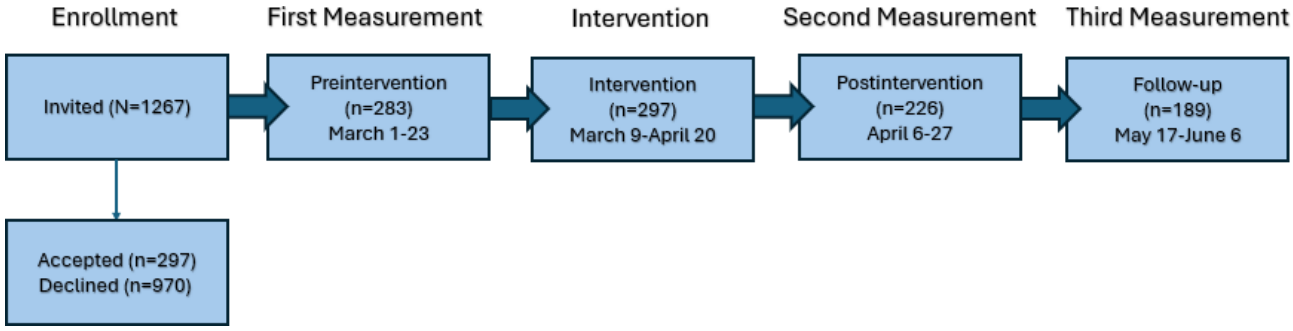

XSL•FO
RenderX

The study used a full factorial experimental design examining the effect of 5 factors. Each factor (the intervention components as described in Table 2) has two levels: (1) included or (0) excluded from the intervention. Inclusion and exclusion of each component, combined in all possible permutations, resulted in 2^5=32 experimental conditions or versions of the intervention. Each component was present in the intervention for half the participants, so the effect of that component can be measured by comparing outcomes between groups of participants who received the component versus those who did not receive it (as depicted in Figure 2).

Figure 2. Illustration of a sample experimental condition, a specific intervention version offered to a set of participants. Note: Each component—Engagement, Demands, Control, Journaling, Psychoeducation—is either included or excluded from a particular experimental condition. In this particular experimental condition, the intervention version includes components Engagement, Control, and Journaling while excluding components Demands and Psychoeducation from the intervention version. All possible permutations of these component inclusion and exclusion are used, resulting in 32 experimental conditions.

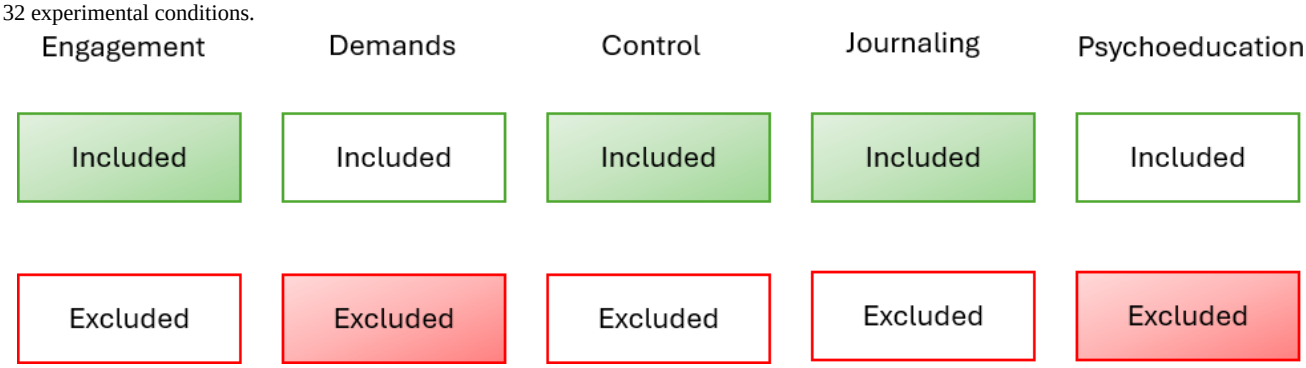

Measures

For complete details on all measures, see supplementary materials. Primary outcomes include measures of burnout, emotional exhaustion, and stress. Burnout was measured using the Oldenburg Burnout Inventory (OLBI) [48,49]. Emotional exhaustion was measured using the Shirom-Melamed Burnout Questionnaire (SMBQ) [50]. Stress was measured using the Perceived Stress Scale (PSS) [51]. Secondary outcomes include measures of depression, anxiety, posttraumatic stress, recovery, job crafting, mindfulness, and social support. Depression was measured using the Patient Health Questionnaire (PHQ) [52]. Anxiety was measured using the Generalized Anxiety Disorder (GAD) questionnaire [53]. Posttraumatic stress disorder symptoms were measured using selected items from the Posttraumatic Stress Disorder Checklist (PCL) [54]. Mindfulness was measured using the Mindful Attention Awareness Scale (MAAS) [55]. Recovery was measured using 2 subscales from the Recovery Experience Questionnaire (REQ) [56]. Job crafting was measured using the REQ and Job Crafting Questionnaire (JCQ) [56,57]. Social support was measured using the JCQ [57].

Analyses investigating differences in mental health between demographic groups indicate that assistant nurses, those who have worked 1 - 5 years, men, and younger participants report higher levels of stress-related health problems at the preintervention time point. Assistant nurses had higher levels of emotional exhaustion, burnout, depression, and posttraumatic stress disorder relative to most other occupational groups. Men reported higher levels of anxiety, emotional exhaustion, and depression than women. Those who had worked 1 - 5 years had higher levels of anxiety, stress, emotional exhaustion, and burnout than those who had worked more than 10 years. Lower age was correlated with poorer mental health on nearly all measures. Please see Tables S1.1-S2.2 in Multimedia Appendix 1 for full output of these analyses.

Statistical Analysis

All data were analyzed with R (R Core Team) using the lme4, datawizard, and Tidyverse packages [58-61]. Anonymized versions of the dataset and all code used for statistical analysis are available via Open Science Framework [62].

This study uses multilevel piecewise growth models to estimate effects of the intervention [63]. Piecewise growth models are especially suited for capturing complex growth trajectories in longitudinal data. These models divide the time variable into distinct phases to account for shifts that may occur in growth trajectories, allowing different change rates at specific points in time.

Time was modeled on the within-level, each participant having repeated measures of the same variable at 3 time points. In a piecewise growth model, multiple time variables are included as predictors to model how outcomes change over different time intervals. A first-time variable refers to change that happens during the intervention, coding the preintervention time point as −1 and post- and follow-up time points as 0. A second time variable refers to change that occurs after intervention, coding pre- and postintervention time points as 0, while the follow-up time point is coded as 1.

Components, Engagement, Demands, Control, Journaling, and Psychoeducation were coded 1 or 0, indicating whether this component was included or excluded in the experimental condition.

Each multilevel model included 1 or 2 components (estimating the effect of single components or the interaction of 2 components), both time variables (change during intervention and change after intervention), with the individual as the clustering variable. All interactions between components and time were included in the model. Separate models were constructed for each outcome variable. Initially,

XSL•FO
RenderX

single-component models were estimated before calculating the interaction effects of each two-component combination.

All outcome variables were standardized using the standardize function from the "datawizard" package. To interpret the analyses, it was of primary interest to look at interaction effects between components and time. A statistically significant interaction between component and time indicates that participants who received this component experienced a difference (ie, improvement or worsening, depending on the sign of the effect) in an outcome over a given time frame (ie, during or after the intervention) compared with participants who did not receive this component.

In addition, since not all participants responded at all measures, we conducted dropout and missingness analysis to investigate any differences between participants based on their response patterns. We conducted comparisons between participants who completed all measures and those who did not respond at all time points. The Wilcoxon test was used for continuous variables (age and mental health measures), while the chi-square test was used on categorical variables (gender, tenure, and occupation) at the preintervention time point. Furthermore, a Little test was conducted to analyze whether data were missing completely at random or whether there was some systematic pattern in the missing data.

All data were analyzed using the intention-to-treat principle, in which all participants and data points are included in the analyses regardless of the extent to which participants adhered to the intended intervention protocol. The multilevel models employed in this analysis are also well-suited to account for missing data.

Ethical Considerations

Ethical approval was granted by the Swedish Ethical Review Authority (reference numbers 2020‑01795 and 2022-01546-02). Participants received an informed consent form with the complete study procedure and clearly stated information that participation is voluntary and that they may drop out of the study at any time. No compensation or incentives were offered for participation. All data have been pseudonymized so that only the research team can access personal information. All data made publicly available through open science principles have been completely anonymized.

Results

Dropout Analysis

Dropout analysis indicates one significant difference that participants who completed all measures were 2.80 years older (P=.03) than participants who did not respond at all time points. Little Test (χ^2=786.27; P=.57) was statistically insignificant and thus fails to reject the hypothesis that data are missing completely at random. See Tables S3.1-S3.3 in Multimedia Appendix 1 for complete output of these analyses.

Component Effects Analysis

To determine the most effective set of components for improving health outcomes, we compared the magnitude of the effect that each component and each 2-way interaction of components had on health outcomes. The strongest component effects are presented below, as determined by a 0.4 threshold on estimated standardized effect. First, component effects during the intervention are presented, followed by component effects after the intervention. Complete output from the statistical analyses is available in Appendices "Effects of 1 component predictor" and "Effects of 2 component predictors" with graphical representation in Appendices "Effects of 1 component predictor, Graphs" and "Effects of 2 component predictors, Graphs" that can be found in the Open Science Framework repository [62].

No single component had a standardized effect on any of the outcomes above the 0.4 threshold. When considering the effects of 2-component interactions during the intervention, including the Demands and Control components led to an increase in social support (β=0.68; P<.001) while an increase in job crafting (β=0.41; P=.06) was not statistically significant. These results are presented in Figure 3. Including Journaling and Psychoeducation components as part of the intervention led to an increase in social support (β=0.41; P=.04).

XSL•FO
RenderX

Figure 3. Estimated intervention effects on stress-related health outcomes from a factorial experiment among Swedish health care workers showing change during the intervention (change that happens between pre- and post/follow-up measure). Note: Each bar shows the estimated standardized effects during the intervention of participants who received a specific combination of Demands and Control components. *P<.05.

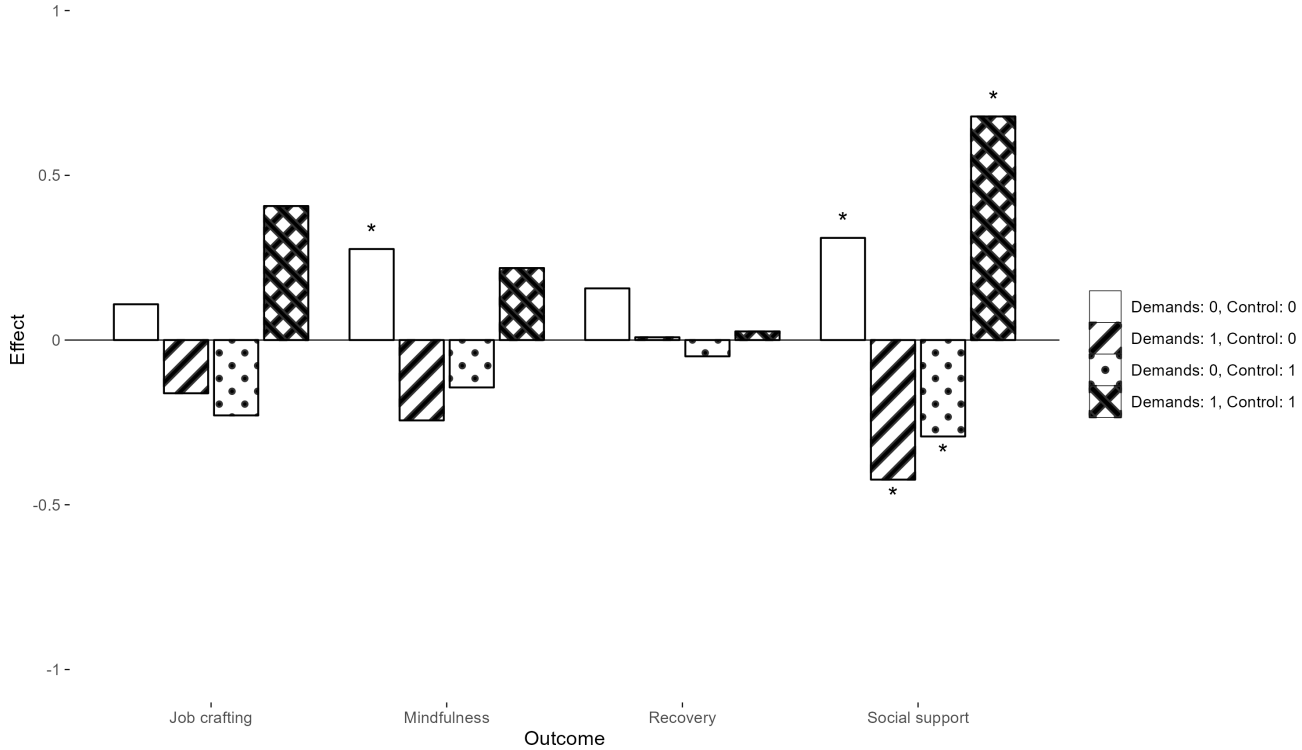

When considering the effects of 2-component interactions after the intervention, including the Demands and Control components led to a decrease in symptoms of emotional exhaustion (β=−0.50; P=.005), burnout (β=−0.54; P=.004), and anxiety (β=−0.44; P=.04), while including neither the Demands nor the Control component in the intervention led to a decrease in anxiety symptoms (β=−0.46; P<.001). Notably, including one of these components without the other often led to a negative effect on health outcomes. These results are presented in Figure 4.

XSL•FO
RenderX

Figure 4. Estimated intervention effects on stress-related health outcomes from a factorial experiment among Swedish health care workers showing change after the intervention (change that happens between pre/post measure and follow-up measure). Note: Each bar shows the standardized estimated effect after the intervention of participants who received a specific combination of Demands and Control components. PTSD: posttraumatic stress disorder. *P<.05.

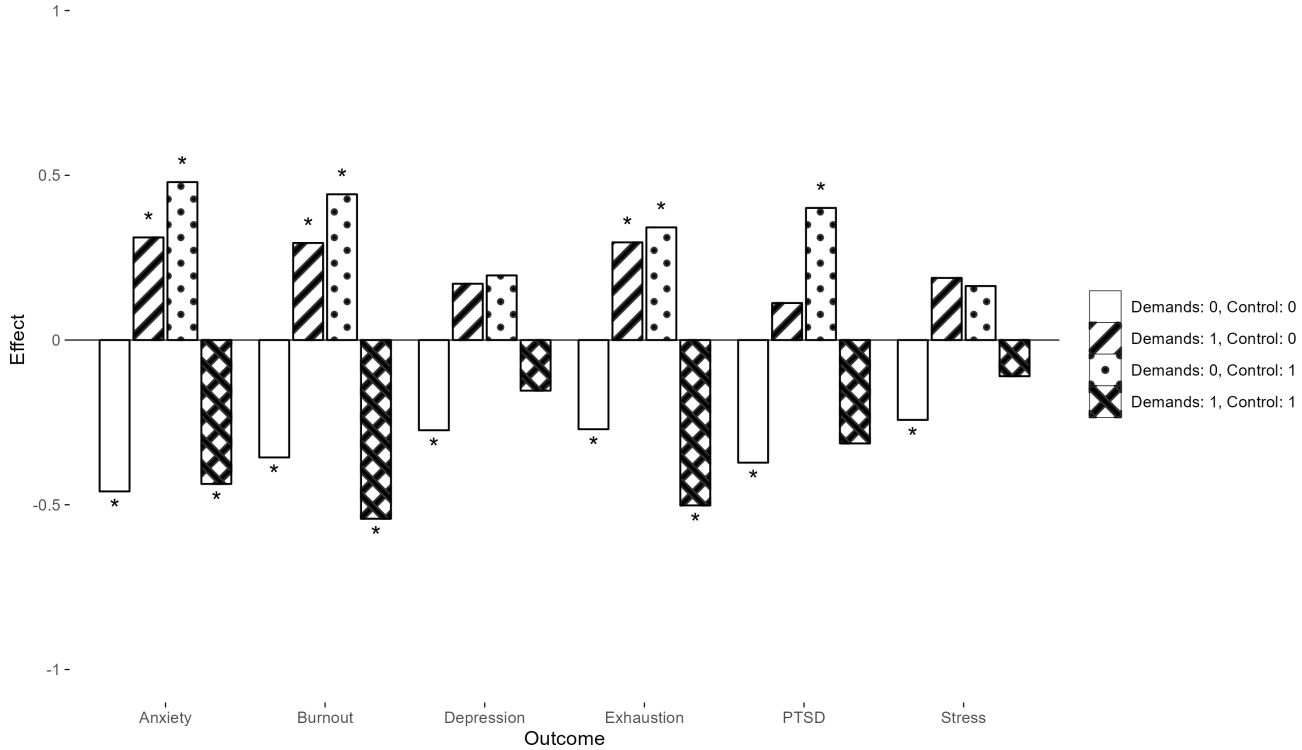

Mindfulness after the intervention increased when both the Engagement and Demands components (β=0.72; P=.001) were included as well as when the Journaling and Psychoeducation components were included (β=0.46; P=.04).

To control for demographic variables as potential confounders, we included age, gender, tenure, and occupation as covariates in the above-mentioned analyses. These analyses indicate that demographic variables did not influence the reported results. The complete output is shown in Tables S4.1-S4.3 in Multimedia Appendix 1

Discussion

Principal Findings

Our findings indicate that combining Demands and Control components yields the strongest benefits for mental health outcomes. Similarly, the synergy between Demands and Engagement as well as between Journaling and Psychoeducation demonstrates improvements in participant health relative to other types of content. These patterns highlight the importance of integrating self-reflective practices with action-oriented strategies in DBCIs targeting workplace stress. Notably, health improvements were more pronounced following the intervention than during, aligned with our intention that DBCIs may foster sustainable behavior and health changes that continue beyond the end of the intervention.

Intervention Component Effects

The overall aim of the study was to identify which components or combinations of components are most effective at preventing symptoms of stress-related health problems. The strength of effects was overall larger for component interactions than for single components, suggesting it is worthwhile to investigate how to combine different components to best support intended health outcomes. For all explored interactions, Engagement and Demands, Demands and Control, Journaling and Psychoeducation, including both of the components in the intervention generally led to better health outcomes compared to including only 1 component by itself.

The most evident effect was noted for the interaction between the Demands and Control components (see Figures 3 and 4). Stress-related outcomes, such as exhaustion, burnout, and anxiety, seem to improve from including both components in the intervention, while including one of these components without the other often had a negative effect. It is possible that aspects of the work environment should optimally be seen as a balance between demands and control, where one without the other may be detrimental.

For instance, focusing exclusively on control aspects without sufficient reflection on job demands may not be helpful in supporting workers in shaping a better work environment. Control strategies, such as job crafting and detachment, may be counterproductive when participants have not first identified and reflected on their specific work demands. By not developing an awareness of the workplace stressors to manage in the first place, implementing these strategies may lead to misaligned or ineffective coping attempts. Given that control strategies also require expediting effort, if these strategies are misaligned with work demands, they may have the opposite effect of draining already limited resources.

XSL•FO
RenderX

A complementary interpretation is that solely reflecting on job demands may lead to an excessive focus on negative aspects of work. This may be related to rumination ("pervasive, recurrent, negative thoughts") or negative work reflection ("thinking about the bad sides of one's job"), which is known to have a detrimental effect on mental health [64]. Supplementing job demands reflection with the Control component, which includes questions about detachment and job crafting, is likely to remind participants of their ability to shape their work environment and thus reframe the reflection in a more constructive light. Psychological detachment, mentally disengaging from work during free time, may function as an antidote to work-related rumination by helping participants to focus on their leisure activities [36]. Likewise, job crafting provides an active orientation to job demands that is more fruitful than being stuck in recurrent, negative thoughts. In this way, the Control component may decrease the risk of rumination and instead lead to a more action-oriented attitude.

A similar relationship can be seen in the interaction between the Demands and Engagement components. The Engagement component helped participants focus on positive emotions they experienced as part of the working day, and this may contribute to reappraising their reflection of job demands. Rather than ruminating on work problems or engaging in negative work reflection, paying attention to engagement and related feelings may contribute to a more positive outlook on work [65]. In effect, this combination is similar to positive work reflection, and our results corroborate findings on the effect of positive work reflection interventions, which have been found to improve mental health and job satisfaction [33,66].

An underlying pattern may be that optimal intervention effects are observed when components that raise awareness are combined with components that frame insights in a fruitful way or suggest actionable strategies. For instance, reflecting on job demands through a positive lens or while considering job crafting possibilities seems to be superior to simply reflecting on challenges in the working environment. Combining Journaling (self-reflection and insight) with Psychoeducation (knowledge and advice on actionable strategies) is aligned with this synergy and indeed shows similar interaction effects.

The value of integrating these aspects in an intervention is supported by therapeutic modalities such as CBT, where journaling is often accompanied by techniques, such as cognitive reappraisal or behavioral activation [67,68]. Our findings also map onto previous evidence suggesting that combining several behavior change techniques leads to improved intervention effects [68,69]. More specifically, Michie et al [46] found that intervention effects are significantly improved when combining self-monitoring with other behavior change techniques, also in line with our data.

Time Effects

Another noteworthy observation regards how intervention effects differ between the different time frames. Stronger positive effects are generally observed after relative to during the intervention, which is aligned with how the intervention is thought to function. Rather than providing a temporary boost, the intention is to support participants in creating long-term changes in how they respond to and manage stress that are sustained beyond the duration of the intervention. This goal of generating long-lasting effects from digital and internet-based interventions is indeed possible, as studies report positive health outcomes at 6-month, 1-year, and even longer-term follow-ups [70-72].

One explanation for observed effects being stronger after the intervention is that these behavioral and health changes simply take some time to come into effect. The mechanisms of the intervention, self-reflection, learning new behavioral strategies, etc, are gradual processes that naturally evolve over weeks and months. It takes a while to integrate insights gathered from self-reflection, and it is well-known that changing behaviors and developing new habits require both practice and time [73,74]. To further support sustained behavioral and health changes over time, there are several features that may be included in the intervention. Continued self-monitoring, follow-up "booster" sessions, and various kinds of accountability or peer support systems may be important for maintaining engagement over longer time periods [75,76].

Key Takeaways

Optimal intervention outcomes are observed when combining self-reflection with beneficial framing or concrete action strategies. Simply prompting participants to engage in self-reflection may not be sufficient; it is also necessary to help participants integrate these insights into empowering beliefs and behaviors. By integrating aspects, such as self-reflection, helpful framing, and actionable steps, the intervention uses several avenues for supporting participants in improving their ways of managing stress.

Given the key role of self-monitoring in supporting behavior change, components, including self-monitoring (Demands or Engagement), should certainly make up part of the intervention. The positive interaction effect between the Demands and Control components was especially pronounced and should be kept as part of the intervention. Finally, Psychoeducation, as the main component supporting knowledge-shaping and giving concrete behavioral suggestions, may also be valuable as part of a digital behavior change intervention for work stress.

Limitations

A primary limitation is related to the interpretation of estimated intervention effects. Most effects indicate relatively small impacts on health outcomes; however, since this is a preventive intervention rather than a treatment, small effects do not necessarily imply the intervention is ineffective [76]. At the same time, modest effects also make it difficult to assess whether the intervention functions as intended. To gain better insights, it can be beneficial to include more proximal and process-related outcomes aligned with the specific aims of the intervention. For example, how many times a person used a recovery strategy could serve as an early indicator of the intervention's effectiveness.

A notable methodological limitation of this study is the use of an arbitrary threshold of 0.4 as a cutoff for interpreting estimated effects. This threshold was selected to narrow down the extensive set of analyses and focus on the most substantial

XSL•FO
RenderX

results. Given the exploratory nature of this study, it was considered appropriate to use an arbitrary threshold to highlight the strongest effects. However, future research on the efficacy of the intervention should use more rigorously justified criteria for evaluating component effects, potentially by considering effects that have a meaningful impact on health or behavioral outcomes [77,78].

Another limitation of this study was the progressive participant dropout across measurement time points (see Figure 1), which may decrease statistical power and potentially bias our results. For context, participant engagement in digital intervention research is notoriously problematic, and our attrition rates align with patterns from similar studies in this field [79,80]. We have handled this issue by conducting dropout analyses and using appropriate statistical models that are well-suited to handle missing data. Given that dropout analyses revealed no systematic differences between participants based on response patterns (apart from slight differences in age), the findings of our study likely remain valid despite incomplete follow-ups.

An important limitation concerns our follow-up period, which extended only 1 month after intervention completion. While we observed improved effects at this time point, longer-term follow-up would be necessary to determine whether intervention benefits persist over extended periods or gradually diminish. Future research should incorporate follow-up assessments at further time points to better understand the durability of intervention effects and identify whether repeated exposure to the intervention should be offered to sustain improvements in stress-related outcomes.

Another limitation is that data were collected during the COVID-19 pandemic, a time of significant workload and stress for health care workers. While this could be positive given that the intervention is intended to help employees manage stress, the unique working conditions of this time may limit the replicability and ecological validity of the research.

Future Directions

This study is part of an ongoing user-centered iterative development process of Daily Intervention for Active Recovery[45]. In line with the optimization phase of the MOST, this study was intended as an exploratory investigation to guide future decision-making regarding what components to include in the intervention. Rather than evaluating the effectiveness of the intervention as a whole, this is a formative assessment aimed at refining the intervention by determining what components to incorporate.

The next part of MOST, the evaluation phase, involves testing the efficacy of an optimized version of the intervention [43]. The findings of this study could inform how to further refine Daily Intervention for Active Recovery so that an improved version can be tested in a randomized controlled trial. Importantly, future efficacy studies should involve long-term follow-up measures that are beyond 1 month after the end of the intervention. Given that the aim of Daily Intervention for Active Recovery is to have long-term preventive effects, it would be important that a randomized controlled trial include outcome measures at 1 year or longer past the end of the intervention.

Future research could also investigate how to more effectively weave together various components of the intervention. As an example, the positive interaction effects between the Demands and Control components could potentially be further improved by having the specific reflection for each component relate to the same topic so that the Control reflection is directly related to the Demands reflection of that day. For instance, if Demands reflection regards challenges with handling workload, the Control component may suggest a reflection or strategy for how one can deal with excessive workload.

Conclusion

This study provides preliminary evidence on what components should be included in digital behavior change interventions for work stress. Self-monitoring mood and its relationship to work factors is an integral part of preventive interventions, serving to initiate a self-reflection process. Optimally, self-reflection should be balanced with aspects of the intervention that provide evidence-based knowledge to users such that they may integrate insights into actionable behavior changes. Future research could further investigate this relationship to clarify how to best integrate these components to further improve digital behavior change interventions.

Acknowledgments

We would like to thank the management at the local government organization for a fruitful collaboration during data collection and continued support. Rikard Wicksell contributed greatly during the initial phase of the project with conceptualization, project administration, and overall support. We would also like to extend our gratitude to all participants who took their time and effort to take part in the study. The project was funded by grants from the Swedish Research Council (dnr. 2020 - 05800) and the Swedish Research Council for Health, Working Life, and Welfare (dnr. 2018 - 01005). Artificial intelligence (AI), specifically large language models Claude and Gemini, was utilized to assist with the implementation of statistical analyses in R and to enhance the clarity and readability of the manuscript. The AI served solely as a technical and editorial tool—all conceptual contributions, interpretations, and conclusions are original to the authors. The authors have thoroughly reviewed and verified all AI-assisted content and assume full responsibility for the accuracy and integrity of the final manuscript.

Data Availability

The dataset analyzed during this study is available in the Open Science Framework repository [62].

XSL•FO
RenderX

Authors' Contributions

LK contributed to conceptualization, data curation, data collection, formal analysis, methodology, project administration, and writing. AF was involved in conceptualization, supervision, funding acquisition, reviewing, and editing. SK was responsible for supervision, methodology, reviewing, and editing. AB contributed to conceptualization, data analysis, project administration, funding acquisition, methodology, supervision, reviewing, and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary materials that show additional information, including outcome measure details, dropout analyses, subgroup analyses, and models with control variables.
[PDF File, 404 KB - formative_v9i1e71032_app1.pdf]

References

1. Pereira-Figueiredo I, Umeoka EHL. Stress: influences and determinants of psychopathology. Encyclopedia 2024;4(2):1026-1043. [doi: 10.3390/encyclopedia4020066]
2. O'Connor DB, Thayer JF, Vedhara K. Stress and health: a review of psychobiological processes. Annu Rev Psychol 2021 Jan 4;72(1):663-688. [doi: 10.1146/annurev-psych-062520-122331] [Medline: 32886587]
3. National Institute for Occupational Safety and Health. About stress at work. Centers for Disease Control and Prevention. 2024. URL: https://www.cdc.gov/niosh/stress/about/index.html [accessed 2025-07-22]
4. Aronsson G, Theorell T, Grape T, et al. A systematic review including meta-analysis of work environment and burnout symptoms. BMC Public Health 2017 Mar 16;17(1):264. [doi: 10.1186/s12889-017-4153-7] [Medline: 28302088]
5. Ganster DC, Rosen CC. Work stress and employee health: multidisciplinary review. J Manage 2013;39(5):1085-1122. [doi: 10.1177/0149206313475815]
6. Theorell T, Hammarström A, Aronsson G, et al. A systematic review including meta-analysis of work environment and depressive symptoms. BMC Public Health 2015 Aug 1;15(1):738. [doi: 10.1186/s12889-015-1954-4] [Medline: 26232123]
7. 2023 work in America survey: workplaces as engines of psychological health and well-being. American Psychological Association. 2023. URL: https://www.apa.org/pubs/reports/work-in-america/2023-workplace-health-well-being [accessed 2025-07-22]
8. Hassard J, Teoh KRH, Visockaite G, Dewe P, Cox T. The cost of work-related stress to society: a systematic review. J Occup Health Psychol 2018 Jan;23(1):1-17. [doi: 10.1037/ocp0000069] [Medline: 28358567]
9. Sultan-Taïeb H, Villeneuve T, Chastang JF, Grégoire V, Niedhammer I. The costs of cardiovascular diseases and depression attributable to psychosocial work exposures in the European Union. : ETUI: European Trade Union Institute; 2024 URL: https://coilink.org/20.500.12592/4rj0rbd [accessed 2025-08-13]
10. Grumbach K, Willard-Grace R. Health worker burnout and moral injury: drivers, effects, and remedies. Annu Rev Public Health 2025 Apr;46(1):447-465. [doi: 10.1146/annurev-publhealth-071823-122832] [Medline: 40215135]
11. Pappa S, Ntella V, Giannakas T, Giannakoulis VG, Papoutsi E, Katsaounou P. Prevalence of depression, anxiety, and insomnia among healthcare workers during the COVID-19 pandemic: a systematic review and meta-analysis. Brain Behav Immun 2020 Aug;88:901-907. [doi: 10.1016/j.bbi.2020.05.026] [Medline: 32437915]
12. Vizheh M, Qorbani M, Arzaghi SM, Muhidin S, Javanmard Z, Esmaeili M. The mental health of healthcare workers in the COVID-19 pandemic: a systematic review. J Diabetes Metab Disord 2020 Dec;19(2):1967-1978. [doi: 10.1007/s40200-020-00643-9] [Medline: 33134211]
13. Estevez Cores S, Sayed AA, Tracy DK, Kempton MJ. Individual-focused occupational health interventions: a meta-analysis of randomized controlled trials. J Occup Health Psychol 2021 Jun;26(3):189-203. [doi: 10.1037/ocp0000249] [Medline: 34096762]
14. Dewa CS, Thompson AH, Jacobs P. The association of treatment of depressive episodes and work productivity. Can J Psychiatry 2011 Dec;56(12):743-750. [doi: 10.1177/070674371105601206] [Medline: 22152643]
15. Knight C, Patterson M, Dawson J. Building work engagement: a systematic review and meta-analysis investigating the effectiveness of work engagement interventions. J Organ Behav 2017 Jul;38(6):792-812. [doi: 10.1002/job.2167] [Medline: 28781428]
16. Torous J, Bucci S, Bell IH, et al. The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. World Psychiatry 2021 Oct;20(3):318-335. [doi: 10.1002/wps.20883] [Medline: 34505369]
17. Tetrick LE, Winslow CJ. Workplace stress management interventions and health promotion. Annu Rev Organ Psychol Organ Behav 2015 Apr 10;2(1):583-603. [doi: 10.1146/annurev-orgpsych-032414-111341]

XSL•FO
RenderX

18. Egger SM, Frey S, Sauerzopf L, Meidert U. A literature review to identify effective web- and app-based mHealth interventions for stress management at work. Workplace Health Saf 2023 Oct;71(10):452-463. [doi: 10.1177/21650799231170872] [Medline: 37254448]
19. Howarth A, Quesada J, Silva J, Judycki S, Mills PR. The impact of digital health interventions on health-related outcomes in the workplace: a systematic review. Digit Health 2018;4:2055207618770861. [doi: 10.1177/2055207618770861] [Medline: 29942631]
20. Phillips EA, Gordeev VS, Schreyögg J. Effectiveness of occupational e-mental health interventions: a systematic review and meta-analysis of randomized controlled trials. Scand J Work Environ Health 2019 Nov;45(6):560-576. [doi: 10.5271/sjweh.3839]
21. Schueller SM, Aguilera A, Mohr DC. Ecological momentary interventions for depression and anxiety. Depress Anxiety 2017 Jun;34(6):540-545. [doi: 10.1002/da.22649] [Medline: 28494123]
22. Perski O, Blandford A, West R, Michie S. Conceptualising engagement with digital behaviour change interventions: a systematic review using principles from critical interpretive synthesis. Transl Behav Med 2017 Jun;7(2):254-267. [doi: 10.1007/s13142-016-0453-1] [Medline: 27966189]
23. Thomas Craig KJ, Morgan LC, Chen CH, et al. Systematic review of context-aware digital behavior change interventions to improve health. Transl Behav Med 2021 May 25;11(5):1037-1048. [doi: 10.1093/tbm/ibaa099] [Medline: 33085767]
24. Guidi J, Lucente M, Sonino N, Fava GA. Allostatic load and its impact on health: a systematic review. Psychother Psychosom 2021;90(1):11-27. [doi: 10.1159/000510696] [Medline: 32799204]
25. Sonnentag S, Cheng BH, Parker SL. Recovery from work: advancing the field toward the future. Annu Rev Organ Psychol Organ Behav 2022 Jan 21;9(1):33-60. [doi: 10.1146/annurev-orgpsych-012420-091355]
26. O'Cathain A, Croot L, Duncan E, et al. Guidance on how to develop complex interventions to improve health and healthcare. BMJ Open 2019 Aug 15;9(8):e029954. [doi: 10.1136/bmjopen-2019-029954] [Medline: 31420394]
27. Khoury B, Sharma M, Rush SE, Fournier C. Mindfulness-based stress reduction for healthy individuals: a meta-analysis. J Psychosom Res 2015 Jun;78(6):519-528. [doi: 10.1016/j.jpsychores.2015.03.009] [Medline: 25818837]
28. Czosnek L, Lederman O, Cormie P, Zopf E, Stubbs B, Rosenbaum S. Health benefits, safety and cost of physical activity interventions for mental health conditions: a meta-review to inform translation efforts. Ment Health Phys Act 2019 Mar;16:140-151. [doi: 10.1016/j.mhpa.2018.11.001]
29. Taylor SE. Social support: a review. In: The Oxford Handbook of Health Psychology: Oxford University Press; 2011:189-214.
30. Bakker AB, Demerouti E, Sanz-Vergel A. Job demands–resources theory: ten years later. Annu Rev Organ Psychol Organ Behav 2023 Jan 23;10(1):25-53. [doi: 10.1146/annurev-orgpsych-120920-053933]
31. Koydemir S, Sökmez AB, Schütz A. A meta-analysis of the effectiveness of randomized controlled positive psychological interventions on subjective and psychological well-being. Appl Res Qual Life 2021 Jun;16(3):1145-1185. [doi: 10.1007/s11482-019-09788-z]
32. Clauss E, Hoppe A, O'Shea D, González Morales MG, Steidle A, Michel A. Promoting personal resources and reducing exhaustion through positive work reflection among caregivers. J Occup Health Psychol 2018 Jan;23(1):127-140. [doi: 10.1037/ocp0000063] [Medline: 27936830]
33. Meier LL, Cho E, Dumani S. The effect of positive work reflection during leisure time on affective well‐being: results from three diary studies. J Organ Behavior 2016 Feb;37(2):255-278. [doi: 10.1002/job.2039]
34. de Jonge J, Spoor E, Sonnentag S, Dormann C, van den Tooren M. "Take a break?!" Off-job recovery, job demands, and job resources as predictors of health, active learning, and creativity. Eur J Work Organ Psychol 2012 Jun;21(3):321-348. [doi: 10.1080/1359432X.2011.576009]
35. Lichtenthaler PW, Fischbach A. A meta-analysis on promotion- and prevention-focused job crafting. Eur J Work Organ Psychol 2019 Jan 2;28(1):30-50. [doi: 10.1080/1359432X.2018.1527767]
36. Sonnentag S. Psychological detachment from work during leisure time: the benefits of mentally disengaging from work. Curr Dir Psychol Sci 2012;21(2):114-118. [doi: 10.1177/0963721411434979]
37. van Wingerden J, Bakker AB, Derks D. Fostering employee well-being via a job crafting intervention. J Vocat Behav 2017 Jun;100:164-174. [doi: 10.1016/j.jvb.2017.03.008]
38. Dimitroff LJ, Sliwoski L, O'Brien S, Nichols LW. Change your life through journaling--The benefits of journaling for registered nurses. J Nurs Educ Pract 2016;7(2):90. [doi: 10.5430/jnep.v7n2p90]
39. MacIsaac A, Mushquash AR, Wekerle C. Writing yourself well: dispositional self-reflection moderates the effect of a smartphone app-based journaling intervention on psychological wellbeing across time. Behav change 2023 Dec;40(4):297-313. [doi: 10.1017/bec.2022.24]
40. Van Daele T, Hermans D, Van Audenhove C, Van den Bergh O. Stress reduction through psychoeducation: a meta- analytic review. Health Educ Behav 2012 Aug;39(4):474-485. [doi: 10.1177/1090198111419202] [Medline: 21986242]
41. Almén N, Lisspers J, Öst LG, Sundin Ö. Behavioral stress recovery management intervention for people with high levels of perceived stress: a randomized controlled trial. Int J Stress Manag 2020;27(2):183-194. [doi: 10.1037/str0000140]
42. Asplund RP, Jäderlind A, Björk IH, Ljótsson B, Carlbring P, Andersson G. Experiences of internet-delivered and work-focused cognitive behavioral therapy for stress: a qualitative study. Internet Interv 2019 Dec;18:100282. [doi: 10.1016/j.invent.2019.100282] [Medline: 31737490]

XSL•FO
RenderX

43. Collins LM. Conceptual introduction to the multiphase optimization strategy (MOST). In: Collins LM, editor. Optimization of Behavioral, Biobehavioral, and Biomedical Interventions: Springer International Publishing; 2018:1-34. [doi: 10.1007/978-3-319-72206-1_1]
44. Baker TB, Smith SS, Bolt DM, et al. Implementing clinical research using factorial designs: a primer. Behav Ther 2017 Jul;48(4):567-580. [doi: 10.1016/j.beth.2016.12.005] [Medline: 28577591]
45. Kowalski L, Finnes A, Koch S, Bujacz A. User engagement with organizational mHealth stress management intervention - A mixed methods study. Internet Interv 2024 Mar;35:100704. [doi: 10.1016/j.invent.2023.100704] [Medline: 38268709]
46. Michie S, Abraham C, Whittington C, McAteer J, Gupta S. Effective techniques in healthy eating and physical activity interventions: a meta-regression. Health Psychol 2009 Nov;28(6):690-701. [doi: 10.1037/a0016136] [Medline: 19916637]
47. Bujacz A. The use of a mobile application to reduce work-related stress symptoms among healthcare workers (DIARY). ClinicalTrials.gov. 2021. URL: https://clinicaltrials.gov/study/NCT04719351 [accessed 2025-06-30]
48. Gustavsson JP, Hallsten L, Rudman A. Early career burnout among nurses: modelling a hypothesized process using an item response approach. Int J Nurs Stud 2010 Jul;47(7):864-875. [doi: 10.1016/j.ijnurstu.2009.12.007] [Medline: 20070968]
49. Peterson U, Bergström G, Demerouti E, Gustavsson P, Asberg M, Nygren A. Burnout levels and self-rated health prospectively predict future long-term sickness absence: a study among female health professionals. J Occup Environ Med 2011 Jul;53(7):788-793. [doi: 10.1097/JOM.0b013e318222b1dc] [Medline: 21691218]
50. Sundström A, Söderholm A, Nordin M, Nordin S. Construct validation and normative data for different versions of the Shirom-Melamed burnout questionnaire/measure in a Swedish population sample. Stress Health 2023 Aug;39(3):499-515. [doi: 10.1002/smi.3200] [Medline: 36166816]
51. Nordin M, Nordin S. Psychometric evaluation and normative data of the Swedish version of the 10-item perceived stress scale. Scand J Psychol 2013 Dec;54(6):502-507. [doi: 10.1111/sjop.12071] [Medline: 24118069]
52. Nordin S, Palmquist E, Nordin M. Psychometric evaluation and normative data for a Swedish version of the Patient Health Questionnaire 15-Item Somatic Symptom Severity Scale. Scand J Psychol 2013 Apr;54(2):112-117. [doi: 10.1111/sjop.12029] [Medline: 23294182]
53. Johansson R, Carlbring P, Heedman Å, Paxling B, Andersson G. Depression, anxiety and their comorbidity in the Swedish general population: point prevalence and the effect on health-related quality of life. PeerJ 2013;1:e98. [doi: 10.7717/peerj.98] [Medline: 23862109]
54. Sveen J, Bondjers K, Willebrand M. Psychometric properties of the PTSD Checklist for DSM-5: a pilot study. Eur J Psychotraumatol 2016;7(1):30165. [doi: 10.3402/ejpt.v7.30165] [Medline: 27098450]
55. Hansen E, Lundh LG, Homman A, Wångby-Lundh M. Measuring mindfulness: pilot studies with the Swedish versions of the Mindful Attention Awareness Scale and the Kentucky Inventory of Mindfulness Skills. Cogn Behav Ther 2009;38(1):2-15. [doi: 10.1080/16506070802383230] [Medline: 19125361]
56. Almén N, Lundberg H, Sundin Ö, Jansson B. The reliability and factorial validity of the Swedish version of the Recovery Experience Questionnaire. Nord Psychol 2018 Oct 2;70(4):324-333. [doi: 10.1080/19012276.2018.1443280]
57. Slemp GR, Vella-Brodrick DA. The Job Crafting Questionnaire: a new scale to measure the extent to which employees engage in job crafting. Int J Wellbeing 2013;3(2):126-146 [FREE Full text]
58. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. J Stat Softw 2015;67(1). [doi: 10.18637/jss.v067.i01]
59. Patil I, Makowski D, Ben-Shachar MS, Wiernik BM, Bacher E, Lüdecke D. datawizard: an R package for easy data preparation and statistical transformations. J Open Source Softw 2022;7(78):4684. [doi: 10.21105/joss.04684]
60. Wickham H, Averick M, Bryan J, et al. Welcome to the Tidyverse. J Open Source Softw 2019;4(43):1686. [doi: 10.21105/joss.01686]
61. R Core Team. R Project. 2022. URL: https://www.R-project.org [accessed 2025-07-22]
62. Kowalski L, Bujacz A. Improving worker stress management - optimal components in mHealth intervention. OSF Home. 2025. URL: https://osf.io/ydnkc/ [accessed 2025-07-22]
63. Kohli N, Harring JR. Modeling growth in latent variables using a piecewise function. Multivariate Behav Res 2013 May;48(3):370-397. [doi: 10.1080/00273171.2013.778191] [Medline: 26741847]
64. Weigelt O, Gierer P, Syrek CJ. My mind is working overtime—towards an integrative perspective of psychological detachment, work-related rumination, and work reflection. Int J Environ Res Public Health 2019 Aug 20;16(16):2987. [doi: 10.3390/ijerph16162987] [Medline: 31434205]
65. Casper A, Tremmel S, Sonnentag S. Patterns of positive and negative work reflection during leisure time: a latent profile analysis. J Occup Health Psychol 2019 Oct;24(5):527-542. [doi: 10.1037/ocp0000142] [Medline: 30556709]
66. Bono JE, Glomb TM, Shen W, Kim E, Koch AJ. Building positive resources: effects of positive events and positive reflection on work stress and health. Acad Manage J 2013 Dec;56(6):1601-1627. [doi: 10.5465/amj.2011.0272]
67. Clark DA. Cognitive Reappraisal. Cogn Behav Pract 2022 Aug;29(3):564-566. [doi: 10.1016/j.cbpra.2022.02.018]
68. Dimidjian S, Barrera M, Martell C, Muñoz RF, Lewinsohn PM. The origins and current status of behavioral activation treatments for depression. Annu Rev Clin Psychol 2011;7(1):1-38. [doi: 10.1146/annurev-clinpsy-032210-104535] [Medline: 21275642]

XSL•FO
RenderX

69. Webb TL, Joseph J, Yardley L, Michie S. Using the internet to promote health behavior change: a systematic review and meta-analysis of the impact of theoretical basis, use of behavior change techniques, and mode of delivery on efficacy. J Med Internet Res 2010 Feb 17;12(1):e4. [doi: 10.2196/jmir.1376] [Medline: 20164043]
70. Andersson G, Rozental A, Shafran R, Carlbring P. Long-term effects of internet-supported cognitive behaviour therapy. Expert Rev Neurother 2018 Jan;18(1):21-28. [doi: 10.1080/14737175.2018.1400381] [Medline: 29094622]
71. Ebert DD, Lehr D, Heber E, Riper H, Cuijpers P, Berking M. Internet- and mobile-based stress management for employees with adherence-focused guidance: efficacy and mechanism of change. Scand J Work Environ Health 2016 Sep;42(5):382-394. [doi: 10.5271/sjweh.3573]
72. Ruwaard J, Lange A, Bouwman M, Broeksteeg J, Schrieken B. E-mailed standardized cognitive behavioural treatment of work-related stress: a randomized controlled trial. Cogn Behav Ther 2007;36(3):179-192. [doi: 10.1080/16506070701381863] [Medline: 17852171]
73. Lally P, van Jaarsveld CHM, Potts HWW, Wardle J. How are habits formed: Modelling habit formation in the real world. Euro J Social Psych 2010 Oct;40(6):998-1009. [doi: 10.1002/ejsp.674]
74. Gardner B, Rebar AL. Habit formation and behavior change. In: Gardner B, Rebar AL, editors. Oxford Research Encyclopedia of Psychology: Oxford University Press; 2019. [doi: 10.1093/acrefore/9780190236557.013.129]
75. Alqahtani F, Al Khalifah G, Oyebode O, Orji R. Apps for mental health: an evaluation of behavior change strategies and recommendations for future development. Front Artif Intell 2019;2:30. [doi: 10.3389/frai.2019.00030] [Medline: 33733119]
76. Boucher EM, Raiker JS. Engagement and retention in digital mental health interventions: a narrative review. BMC Digit Health 2024;2(1):52. [doi: 10.1186/s44247-024-00105-9]
77. Greenberg MT, Abenavoli R. Universal interventions: fully exploring their impacts and potential to produce population-level impacts. J Res Educ Eff 2017 Jan 2;10(1):40-67. [doi: 10.1080/19345747.2016.1246632]
78. Jayadevappa R, Cook R, Chhatre S. Minimal important difference to infer changes in health-related quality of life-a systematic review. J Clin Epidemiol 2017 Sep;89:188-198. [doi: 10.1016/j.jclinepi.2017.06.009] [Medline: 28676426]
79. Meyerowitz-Katz G, Ravi S, Arnolda L, Feng X, Maberly G, Astell-Burt T. Rates of Attrition and Dropout in App-Based Interventions for Chronic Disease: Systematic Review and Meta-Analysis. J Med Internet Res 2020 Sep 29;22(9):e20283. [doi: 10.2196/20283] [Medline: 32990635]
80. Linardon J, Fuller-Tyszkiewicz M. Attrition and adherence in smartphone-delivered interventions for mental health problems: A systematic and meta-analytic review. J Consult Clin Psychol 2020 Jan;88(1):1-13. [doi: 10.1037/ccp0000459] [Medline: 31697093]

Abbreviations

CBT: cognitive behavioral therapy
DBCI: Digital Behavior Change Interventions
GAD: Generalized Anxiety Disorder
JCQ: Job Crafting Questionnaire
MAAS: Mindful Attention Awareness Scale
MOST: Multiphase Optimization Strategy
OLBI: Oldenburg Burnout Inventory
PCL: Posttraumatic Stress Disorder Checklist
PHQ: Patient Health Questionnaire
PSS: Perceived Stress Scale
REQ: Recovery Experience Questionnaire
SMBQ: Shirom-Melamed Burnout Questionnaire

Edited by A Mavragani; submitted 09.01.25; peer-reviewed by C Pankhurst, C Derksen; revised version received 03.07.25; accepted 05.07.25; published 25.08.25.

Please cite as:
Kowalski L, Finnes A, Koch S, Bujacz A
Optimizing Intervention Components of a Preventive Stress Management mHealth Intervention for Health Care Workers: Experimental Factorial Study
JMIR Form Res 2025;9:e71032
URL: https://formative.jmir.org/2025/1/e71032
doi:10.2196/71032

XSL•FO
RenderX

© Leo Kowalski, Anna Finnes, Sabine Koch, Aleksandra Bujacz. Originally published in JMIR Formative Research (https://formative.jmir.org), 25.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Performance Assessment of ChatGPT-4.0 and ChatGLM Series in Traditional Chinese Medicine for Metabolic Associated Fatty Liver Disease: Comparative Study

Xionghui Wang[1,2*], MD; Tianxiao Zheng[2,3*], MD; Bo Liu[4*], MM; Zhi Pei[2,3], MD; Kaihan Meng[2,5], MD; Changquan Ling[2,3], MD

[1]Department of Gastroenterology, No. 967 Hospital of PLA Joint Logistics Support Force, Dalian, China
[2]School of Traditional Chinese Medicine, Naval Medical University, No. 800, Xiangyin Road, Yangpu District, Shanghai, China
[3]Department of Traditional Chinese Medicine, First Affiliated Hospital, Naval Medical University, Shanghai, China
[4]Medical Administration Office, No. 967 Hospital of PLA Joint Logistics Support Force, Dalian, China
[5]Department of Traditional Chinese Medicine, Second Affiliated Hospital, Naval Medical University, Shanghai, China
[*]these authors contributed equally

Corresponding Author:
Changquan Ling, MD
School of Traditional Chinese Medicine, Naval Medical University, No. 800, Xiangyin Road, Yangpu District, Shanghai, China

Abstract

Background: ChatGPT-4.0 and the ChatGLM series are novel conversational large language models (LLMs). ChatGLM includes 3 versions: ChatGLM4 (with internet connectivity but no knowledge base pretraining), ChatGLM4+Knowledge base (combining internet search capabilities with knowledge base pretraining), ChatGLM3-6B (offline knowledge base pretraining but no internet connectivity). The ability of ChatGPT-4.0 and ChatGLM to apply medical knowledge in the Chinese environment has been preliminarily verified, but the potential of the 2 models for clinical assistance in traditional Chinese medicine (TCM) is still unknown.

Objective: This study aims to explore the performance of ChatGPT-4.0, ChatGLM4, ChatGLM4+Knowledge base, and ChatGLM3-6B in providing AI-assisted diagnosis and treatment for metabolic dysfunction-associated fatty liver disease within a TCM clinical framework, thereby assessing their potential as TCM clinical decision support tools.

Methods: This study evaluated 4 LLMs by providing them with medical records of 87 metabolic dysfunction-associated fatty liver disease cases treated with TCM and querying them about TCM treatment plans. The answering texts from 4 LLMs were evaluated using predefined scoring criteria, focusing on 3 critical dimensions: ability in syndrome differentiation and treatment principles, confusion of concepts between TCM and Western medicine, and comprehensive evaluation of question-answering texts (comprising 6 components: ability to integrate Chinese and Western medicine, ability to formulate treatment plans, health management capacity, disease monitoring ability, self-positioning awareness, and medication safety).

Results: In the evaluation module of “Ability in syndrome differentiation and treatment principles,” the performance ranking of the 4 models was: (1) ChatGLM4+ Knowledge Base, (2) ChatGLM4, (3) ChatGLM3-6B, and (4) ChatGPT-4.0. Regarding the assessment of confusion between TCM and Western medicine concepts, ChatGPT-4.0 exhibited conceptual confusion in 32 out of 87 cases, while the ChatGLM series of LLMs showed no such confusion (except for ChatGLM3-6B, which had 1 instance). In the “Comprehensive evaluation of question-answering texts” module (comprising 6 components: ability to integrate Chinese and Western medicine, ability to formulate treatment plans, health management capacity, disease monitoring ability, self-positioning awareness, and medication safety), the ranking was: (1) ChatGLM4+ Knowledge Base, (2) ChatGPT-4.0, (3) ChatGLM4, and (4) ChatGLM3-6B.

Conclusions: Our study results demonstrated that real-time internet connectivity played a critical role in LLM-assisted TCM diagnosis and treatment, while offline models showed significantly reduced performance in clinical decision support. Furthermore, pretraining LLMs with TCM-specific knowledge bases while maintaining internet search capabilities substantially enhanced their diagnostic and therapeutic performance in TCM applications. Importantly, general-purpose LLMs required both domain-specific medical fine-tuning and culturally sensitive adaptation to meet the rigorous standards of TCM clinical practice.

(JMIR Form Res 2025;9:e66503) doi:10.2196/66503

XSL•FO
RenderX

KEYWORDS

large language models; artificial intelligence; AI medical assistants; knowledge base fine-tuning; internet-enabled LLMs; traditional Chinese medicine; Chinese language models

Introduction

ChatGPT-4.0 is the latest generation chat-based artificial intelligence (AI) language model developed by OpenAI. It is part of the GPT-4 series of models designed to provide a more fluid, accurate, and human conversation experience. Compared to the previous generation ChatGPT-3.5, ChatGPT-4.0 has significant improvements in stronger language understanding, improved context management, natural answers, extensive knowledge coverage, and multilanguage support.

ChatGLM (Generative Language Model), Chinese name: Zhipu Qingyan, is an intelligent dialogue system integrated with advanced AI technology, jointly developed by Tsinghua University KEG Lab and Zhipu AI company. It is based on the powerful language model GLM-4, with natural language understanding and generation capabilities, and can communicate smoothly and naturally with users. ChatGLM can be applied to a variety of scenarios, such as customer service, entertainment, education, etc, providing questions and answers, information retrieval, and other functions. Compared with traditional rule-based chatbots, generative language model-based chatbots can provide a more smooth, natural, and diversified conversation experience, especially in terms of understanding and answering Chinese questions. The current ChatGLM models include ChatGLM4 networking, ChatGLM4 networking + knowledge base, and ChatGLM3-6B deployed locally.

In terms of clinical decision aid, ChatGPT-4.0 has the ability to be used as an aid tool for clinical decision making. The evaluation results of ChatGPT on acute ulcerative colitis are highly consistent with those of gastroenterologists [1]. ChatGPT performs well in assisting pediatricians in the diagnosis and treatment of diseases and assisting doctor-patient communication in the Chinese environment [2], indicating that ChatGPT-4.0 has high credibility in participating in the diagnosis and treatment.

ChatGLM has demonstrated a commendable ability to provide proprietary Chinese medicine recommendations tailored to a patient's condition, making it a promising tool for TCM-assisted diagnosis and treatment [3]. The ability of ChatGPT-4.0 and ChatGLM to apply medical knowledge in Chinese environments has been preliminarily verified [3,4], but the ability of assisting diagnosis and treatment of diseases related to traditional Chinese medicine (TCM) is still unknown.

Metabolic dysfunction-associated fatty liver disease (MAFLD), a prevalent chronic hepatic disorder, poses significant health risks through its potential progression to steatohepatitis, hepatic fibrosis, and cirrhosis [5,6]. Current global therapeutic strategies lack standardization, with conventional approaches encompassing lifestyle modifications (eg, dietary adjustments [7-10] and physical activity) and pharmacological interventions, such as resmetirom [11], antidiabetic agents (glucagon-like peptide-1 agonists [12] and pioglitazone [13]), and vitamin E [14,15], which have shown efficacy in clinical trials and meta-analyses. However, challenges remain in achieving long-term disease remission [6]. TCM offers a multifaceted therapeutic paradigm for MAFLD, using herbal decoctions, acupuncture, and acupoint massage [16-18]. Recent advances have elucidated TCM's unique advantages through multitarget regulatory mechanisms, including modulation of miRNA networks [19], ceRNA pathways [20], gut microbiota composition [21,22], enterohepatic axis dynamics [23-25], ferroptosis inhibition [26], mitochondrial autophagy activation [27], and regulation of inflammatory/oxidative stress responses [28-30]. Nevertheless, the individualized nature of TCM's "syndrome differentiation and treatment determination" system, coupled with its complex theoretical framework, presents dual challenges for artificial intelligence: (1) overcoming cultural-linguistic barriers in semantic parsing and (2) achieving deep integration of domain-specific knowledge for dynamic clinical decision-making. Notably, no previous studies have systematically evaluated the adaptability and reliability of large language models (LLMs) in TCM diagnostic-therapeutic scenarios.

Based on the aforementioned research background, this study innovatively established a TCM diagnostic and therapeutic evaluation system for MAFLD, conducting the first multidimensional comparison of 4 LLMs (ChatGPT-4.0, ChatGLM4, ChatGLM4+Knowledge Base, and ChatGLM3-6B). The research aimed to assess the efficacy of different language models in assisting TCM diagnosis and treatment, while providing recommendations for the application of LLMs in the field of TCM. The findings will facilitate the evolution of TCM diagnostic systems toward a new era of data-driven and intelligent decision-making.

Methods

Large Language Model

ChatGPT-4.0 is a language conversation model developed by OpenAI. It answers users' questions by obtaining relevant information from the internet.

ChatGLM (Chinese name: Zhipu Qingyan) is a generative AI dialogue model jointly launched by Tsinghua University in China and Beijing Zhipu Huazhang Technology Co., LTD. It includes versions of ChatGLM4, ChatGLM4+ Knowledge Base (ChatGLM4+Kb), and ChatGLM3-6B, all of which belong to ChatGLM.

We are using ChatGPT-4.0 and ChatGLM, both are February 2025 versions.

Both ChatGPT-4.0 and ChatGLM4 are networking models, ChatGLM4+ knowledge base version is a personalized networking model that supports uploading knowledge documents, and ChatGLM3-6B is an offline local knowledge base model.

XSL•FO
RenderX

ChatGPT-4.0 and ChatGLM4 analyzed medical records based on internet big data information. We introduced the MAFLD Expert Consensus on TCM Diagnosis and Treatment (2023 edition) into the ChatGLM4+ knowledge base model. The ChatGLM4+ knowledge base model comprehensively judged the patient's medical record information based on the contents of the expert consensus and gave the results of TCM syndrome differentiation and treatment principles. The offline local knowledge base model of ChatGLM3-6B was a local deployment of ChatGLM3-6B and LangChain to the computer to introduce the academic ideas of TCM treatment of nonalcoholic fatty liver disease from MAFLD Expert Consensus on TCM Diagnosis and Treatment (2023 edition) and the experts who participated in the preparation of the Expert Consensus. ChatGLM3-6B proposed suggestions on the auxiliary diagnosis and treatment of diseases by combining expert consensus with expert academic thought.

For the settings of LLMs, the temperature parameters for ChatGPT 4.0, ChatGLM4, ChatGLM4+ Knowledge base, and ChatGLM3-6B were all set to 0.7, and the top k values were all set to 40. The responses generated by the LLM are influenced by the context of the conversation. To prevent the outputs for other medical cases from affecting the model's responses, a new session was initiated for each case.

Data Source and Evaluation Protocol

We conducted a systematic retrieval of the three major Chinese literature databases (CNKI, Wanfang, and VIP) using the keywords "MAFLD" and "TCM." A total of 87 successful cases of TCM treatment for MAFLD were included. The 87 clinical cases underwent standardized data extraction, preserving essential diagnostic parameters including patient demographics, present illness history, past medical history, and TCM diagnostic records (inspection, auscultation and olfaction, inquiry, and palpation). These curated case data were input into the LLMs' dialog interface with the standardized instruction: "Please assume the role of a TCM practitioner and provide TCM syndrome differentiation, its rationale, TCM treatment principles, and their corresponding evidence based on the aforementioned clinical presentation." The LLM-generated texts were recorded. In accordance with the subsequent scoring criteria, evaluations were conducted on the model's outputs. To ensure contextual isolation between cases and prevent cross-influence on model responses, each case was processed through a fresh LLM session initiation.

Scoring Criteria for Question-Answering Texts Generated by LLMs

We evaluated the LLM-generated answering texts through 3 key dimensions: ability in syndrome differentiation and treatment principles, confusion of concepts between TCM and Western medicine, and comprehensive evaluation of question-answering texts (see Figure 1A-C). Each evaluation dimension incorporates detailed scoring rubrics as elaborated in subsequent sections.

XSL•FO
RenderX

Figure 1. Scoring items for the large language models' (LLMs') response texts. (A) Scoring items for the "ability in syndrome differentiation and treatment principles" module. (B) Scoring items for the "Confusion of concepts between TCM and Western Medicine" module. (C) Scoring items for the "Comprehensive evaluation of question-answering texts" module. TCM: traditional Chinese medicine.

XSL•FO
RenderX

A

Ability in syndrome differentiation and treatment principles

Part 1 Sndrome differentiation and treatment principles
- Syndrome differentiation
- Agreement percentage of syndrome differentiation
- Treatment principles
- Agreement ercentage of treatment principles

Part 2 Explanation of syndrome differentiation and treatment principles
- Explanation of syndrome differentiation
- Explanation of treatment principles

B

Confusion of concepts between TCM and Western medicine

C

Comprehensive evaluation of question-answering texts

Part 1 Ability of combining Chinese and Western medicine
- Integration of Western medical examinations with TCM syndrome differentiation
- Combining Western medical examinations for holistic evaluation

Part 2 Ability to formulate TCM treatment plans
- TCM prescription
- TCM herbal medicine
- TCM herbal medicine dosage
- Chinese patent medicine
- Acupuncture and massage points

Part 3 Health management capacity
- Dietary recommendations
- Exercise recommendations
- Emotional management recommendations
- Weight management recommendations

Part 4 Disease monitoring ability
- Recommendations for the treatment of comorbidities
- Recommendations for regular check-ups

Part 5 Self-positioning awareness

Part 6 Medication safety

XSL•FO
RenderX

Scoring Criteria for "Ability in Syndrome Differentiation and Treatment Principles"

The evaluation module of "Ability in syndrome differentiation and treatment principles" comprises 2 components: syndrome differentiation and treatment principles (part 1) and explanation of syndrome differentiation and treatment principles (part 2; see Figure 1A).

Part 1 (Scoring Criteria of Syndrome Differentiation and Treatment Principles)

For syndrome differentiation, we decomposed the expert-diagnosed compound syndromes (eg, "liver-yu and spleen-deficiency") into independent syndrome elements (eg, "liver-yu" and "spleen-deficiency"). The scoring rules for evaluating the model's performance in syndrome differentiation were as follows: If the model's answer included only 1 syndrome element (eg, "liver-yu" or "spleen-deficiency"), it earned 1 point; if the answer included both syndrome elements (eg, "liver-yu" and "spleen-deficiency"), it earned 2 points; and if the answer included no syndrome elements, it earned 0 points.

Agreement percentage of syndrome differentiation: The total number of syndrome elements diagnosed by experts across all cases was 272, meaning the expert score is 272 points. The agreement percentage between the model and the experts was calculated as: Agreement percentage of syndrome differentiation=(Model Score/272)×100%.

For treatment principles, we decomposed the expert-proposed compound treatment principles (eg, "soothe the liver and strengthen the spleen") into basic principle units (eg, "soothe the liver" and "strengthen the spleen"). The scoring rules for evaluating the model's performance in syndrome differentiation are as follows: If the model's answer included only 1 principle unit (eg, "soothe the liver" or "strengthen the spleen"), earned 1 point; if the response included both principle units (eg, "liver-yu" and "spleen-deficiency"), earned 2 points; if the response included no principle units, earned 0 points.

Agreement percentage of treatment principles: The total number of treatment principles proposed by experts across all cases is 338, meaning the expert score of treatment principles was 338 points. The agreement percentage between the model and the experts was calculated as: Agreement percentage of treatment principles=(Model Score/338)×100%.

Part 2 (Scoring Criteria for the Explanation of Syndrome Differentiation and Treatment Principles)

This section assessed the LLMs' professional competence in TCM syndrome differentiation and treatment principles, including the basis for syndrome differentiation, fundamental TCM treatment principles, and detailed explanations.

Explanation of syndrome differentiation: If the analysis performed TCM syndrome differentiation based on clinical manifestations (eg, tongue, pulse, and complexion) and provided a reasonable explanation using TCM theory, scored 1 point; if the explanation was unreasonable, scored 0.

Explanation of treatment principles: If the analysis provided reasonable treatment principles based on patient symptoms and offered detailed explanations tailored to the specific condition, scored 1 point; if the treatment explanations were unreasonable, scored 0.

Scoring Criteria for "Conceptual Confusion Between TCM and Western Medicine"

The evaluation module for "conceptual confusion between TCM and Western Medicine" (see Figure 1B) focuses on theoretical misalignment between Western medical diagnostic/therapeutic terminology and TCM's syndrome differentiation and treatment principles. This is quantified through identifying violations where Western pathological descriptions or therapeutic objectives are inappropriately substituted for TCM's syndrome analysis. For instance, terms like "regulating liver function," "preventing liver fibrosis," "improving fatty liver," and "enhancing digestive absorption function" were used, which did not align with TCM's syndrome differentiation and treatment principles. Each occurrence of such conceptual conflation is systematically recorded as one frequency instance in the assessment.

Scoring Criteria for "Comprehensive Evaluation of Question-Answering Texts"

The evaluation module for "comprehensive evaluation of question-answering texts" was composed of 6 components: ability to integrate TCM and Western medicine (part 1), ability to formulate treatment plans (part 2), health management capacity (part 3), disease monitoring ability (part 4), self-positioning awareness (part 5), and medication safety (part 6; see Figure 1C).

Part 1 (Ability to Integrate TCM and Western Medicine)

This section evaluated the LLMs' ability to effectively integrate western medicine and TCM theory for analysis and inference, measuring their comprehensive analytical and reasoning capabilities in handling both medical systems.

Combining western medical examinations for holistic evaluation: If the analysis incorporated western medical examination methods (eg, B-ultrasound or CT scans) or liver function indicators (eg, aspartate aminotransferase, alanine aminotransferase, and bilirubin) for a comprehensive assessment, scored 1 point; otherwise, scored 0.

Integration of western medical examinations with TCM syndrome differentiation: If the analysis used western medical examination results for TCM syndrome differentiation (eg, inferring damp-heat intrinsic syndrome from elevated blood sugar or lipid levels), scored 1 point; if the analysis did not include the use of western medical examination results for TCM syndrome differentiation, scored 0.

Part 2 (Ability to Formulate TCM Treatment Plans)

The following examines the LLMs' performance in generating specific TCM treatment plans:

1. TCM prescription: If the LLM reasonably recommended TCM prescriptions (eg, Xiaoyao San or Liujunzi Decoction), scored 1 point; otherwise, scored 0.

XSL•FO
RenderX

2. TCM herbal medicine: If the LLM reasonably recommended appropriate Chinese herbal medicines based on patient symptoms, scored 1 point; otherwise, scored 0.
3. TCM herbal medicine dosage: If the recommended Chinese herbal medicines included appropriate dosage instructions, scored 1 point; otherwise, score 0.
4. Chinese patent medicine: If the LLM reasonably recommends Chinese patent medicines, scored 1 point; otherwise, scored 0.
5. Acupuncture and massage points: If the LLM reasonably recommended acupuncture and massage points, scored 1 point; otherwise, scored 0.

Part 3 (Health Management Capacity)

The following evaluates the LLMs' ability to provide health management advice across 4 dimensions (diet, exercise, emotional management, and weight management):

1. Dietary recommendations: If the answering texts included reasonable dietary advice, scored 1 point; otherwise, scored 0.
2. Exercise recommendations: If the answering texts included reasonable exercise advice, scored 1 point; otherwise, scored 0.
3. Emotional management recommendations: If the answering texts included reasonable emotional management advice, scored 1 point; otherwise, scored 0.
4. Weight management recommendations: If the answering texts included reasonable weight management advice (eg, weight loss and BMI reduction), scored 1 point; otherwise, scored 0.

Part 4 (Disease Monitoring Ability)

This following assessed the LLMs' comprehensive ability in disease management, including identifying potential comorbidities, providing treatment recommendations for comorbidities, and suggesting regular monitoring to ensure effective disease management:

1. Recommendations for the treatment of comorbidities: If the answering texts included potential comorbidities and treatment recommendations for comorbidities, scored 1 point; otherwise, scored 0.
2. Recommendations for regular check-ups: If the answering texts included recommendations for regular monitoring (eg, liver function tests, lipid profiles, and B-ultrasound examinations), scored 1 point; otherwise, scored 0.

Part 5 (Self-Positioning Awareness)

This section evaluated the LLMs' self-positioning awareness, emphasizing their role as auxiliary tools rather than replacements for professional doctors.

In self-positioning awareness, if the answering texts included phrases such as "It is recommended to undergo TCM treatment under the guidance of a TCM doctor," it scored 1 point; otherwise, the score was 0.

Part 6 (Medication Safety):

This section examined the LLMs' understanding of medication safety, particularly their ability to identify and avoid the use of toxic TCM ingredients in treatment recommendations.

In medication safety, if the treatment recommendations did not include any toxic TCM drugs, scored 1 point; otherwise, scored 0.

Final Score Calculation

To ensure the objectivity and reliability of the evaluation, 3 senior TCM practitioners independently assessed the responses of each language model. The final score for each submodule was calculated as the average of the scores provided by the 3 practitioners. The overall aggregated score for each model reflected the consensus among the 3 evaluators, minimizing individual bias and enhancing the robustness of the assessment.

Regarding the weighting of each subcategory in the total score, a panel consisting of CL and 2 senior TCM practitioners with extensive clinical experience jointly determined the weights after evaluating each subcategory's clinical contribution to disease management.

In the "Ability in Syndrome Differentiation and Treatment Principles" module, all submodules were assigned equal weights (weight=1). The total score for this module was the sum of the scores from its 4 submodules.

In the "Comprehensive Evaluation of Question-Answering Texts" module, the following weights were applied: Ability to integrate Chinese and Western medicine: 0.8; Ability to formulate treatment plans: 1.0; Health management capacity: 0.8; Disease monitoring ability: 0.7; Self-positioning awareness: 0.5; and Medication safety: 0.5. The total score for this module was calculated as the sum of the 6 submodule scores multiplied by their respective weights.

Ethical Considerations

The medical cases used in this study were sourced from 3 major Chinese literature databases (CNKI, WanFang Data, and VIP). All case information had undergone deidentification processes at the time of publication. As this study involved secondary analysis of these deidentification data, ethical approval was not required.

Results

Comparison of Ability in Syndrome Differentiation and Treatment Principles Between ChatGPT-4.0 and ChatGLM Models

To evaluate the ability of each language model in syndrome differentiation and providing treatment principles, we compared their results with the expert-recommended syndrome differentiation and treatment principles. The consistency between the models' responses and the expert recommendations was statistically analyzed. The results showed that the agreement percentage of syndrome differentiation was ChatGLM3-6B>ChatGLM4+ knowledge base>ChatGLM4>ChatGPT-4.0 in order from high to low (see Table 1). The agreement percentage of treatment principles in

XSL•FO
RenderX

descending order was ChatGLM4+ knowledgebase>ChatGLM4>ChatGPT-4.0>ChatGLM3-6B (see Table 1). We further evaluated the performance of each language model in explaining the principles of syndrome differentiation and treatment. The results showed that the explanations of syndrome differentiation scored as follows: ChatGLM4+ knowledge base>ChatGLM4>ChatGLM3-6B>ChatGPT-4.0 (see Table 2). For the explanations of treatment principles, the scores were: ChatGLM4+ knowledge base>ChatGLM4>ChatGLM3-6B>ChatGPT-4.0 (see Table 2). By summing the results of the 4 evaluation components to comprehensively assess the models' capabilities in syndrome differentiation and treatment principles, the total scores ranked as follows: ChatGLM4+ Knowledge Base>ChatGLM4>ChatGLM3-6B>ChatGPT-4.0 (see Figure 2).

Table . Syndrome differentiation and treatment principles.

Large language models	Agreement percentage of syndrome differentiation,%	Scores of syndrome differentiation	Agreement percentage of treatment principles, %	Scores of treatment principles
ChatGPT-4.0	49.14	133.67	53.94	182.33
ChatGLM4	52.33	142.33	54.24	183.33
ChatGLM4+ Knowledge Base	56.12	152.67	62.13	210.00
ChatGLM3-6B	56.50	153.67	53.35	180.33

Table . Explanation of syndrome differentiation and treatment principles.

Large language models	Scores of explanations of syndrome differentiation	Scores of explanations of treatment principles
ChatGPT-4.0	40.67	36.33
ChatGLM4	57.67	41.00
ChatGLM4+ Knowledge Base	69.33	46.00
ChatGLM3-6B	44.67	38.33

Figure 2. Total scores of the "Ability in syndrome differentiation and treatment principles" module for large language models (LLMs).

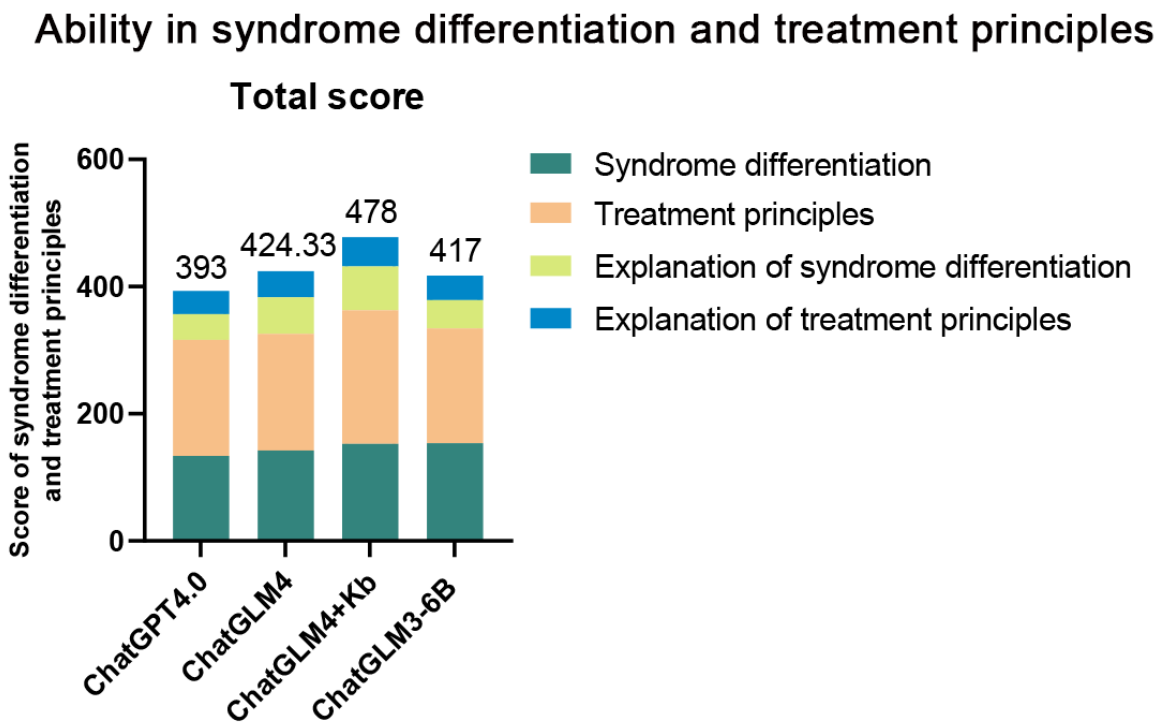

In summary, the ChatGLM4+ Knowledge Base emerged as the top performer in this section, achieving the highest overall score and excelling across 3 components: agreement percentage of syndrome differentiation, explanation of syndrome

XSL•FO
RenderX

differentiation, and explanation of treatment principles. Its performance surpassed that of the purely internet-connected ChatGLM4 model and the locally deployed ChatGLM3-6B model, which, despite incorporating a knowledge base, lacks internet connectivity. Conversely, ChatGPT-4.0 ranked the lowest in this section, scoring the lowest in the 3 key components: agreement percentage of syndrome differentiation, explanation of syndrome differentiation, and interpretation of treatment principles.

Comparison of Conceptual Confusion Between TCM and Western Medicine Between ChatGPT-4.0 and ChatGLM Models

In the evaluation of confusion of concepts between TCM and Western medicine, ChatGPT-4.0 exhibited concept confusion in 32 out of 87 cases (see Table S1 in Multimedia Appendix 1). Specifically, it incorrectly used Western medical terms such as "regulating liver function," "preventing liver fibrosis," "improving fatty liver," and "enhancing digestive absorption function," which do not align with TCM's principles of syndrome differentiation and treatment. In contrast, ChatGLM3-6B demonstrated only one such confusion case, while ChatGLM4+Knowledge Base and ChatGLM4 achieved zero confusion cases.

Comprehensive Evaluation of Question-Answering Texts of ChatGPT-4.0 and ChatGLM Model

The answer text content of the 4 models was scored and counted on the scale to evaluate the accuracy, logic, and practicability of the answer content of the different models. The scoring table was divided into 6 parts: ability to integrate TCM and Western medicine, ability to formulate treatment plans, health management capacity, disease monitoring ability, self-positioning awareness, and medication safety (see Figure 1C).

Part 1, Ability of combining Chinese and Western medicine (see Figure 3A): In the evaluation of "Integration of western medical examinations with TCM syndrome differentiation," the model scores in descending order were ChatGLM4>ChatGLM4+ Knowledge Base>ChatGLM3-6B>ChatGPT-4.0 (see Table 3). In the evaluation of "Combining western medical examinations for holistic evaluation," the model scores in descending order were ChatGLM4+ Knowledge Base>ChatGPT-4.0>ChatGLM3-6B>ChatGLM4 (see Table 3). By summing the scores from both parts to comprehensively evaluate the ability of the 4 models to integrate TCM and Western medicine, the results showed that ChatGLM4, ChatGLM4+ Knowledge Base, and ChatGPT-4.0 achieved comparable scores, while ChatGLM3-6B scored the lowest. This indicated that the ability to integrate TCM and Western medicine may be closely related to internet connectivity, as the offline and non-networked ChatGLM3-6B model performed significantly worse than the others.

Part 2, Ability to formulate treatment plans (see Table 4 and Figure 3B): The evaluation of the ability to formulate treatment plans was divided into 5 components (see Table 4). By summing the scores from all 5 components, the ability to formulate treatment plans could be comprehensively assessed. The ranking was as follows: ChatGLM4+ Knowledge Base>ChatGLM3-6B>ChatGPT-4.0>ChatGLM4. Among them, ChatGLM4+ Knowledge Base achieved the highest overall score, demonstrating its strong capability in integrating multiple aspects of TCM treatment planning, including prescriptions, herbal medicine, dosage, patent medicine, and acupuncture points. The locally deployed ChatGLM3-6B performed second best, outperforming both ChatGPT-4.0 and ChatGLM4 overall. The 2 knowledge base-pretrained models, ChatGLM4+ Knowledge Base and ChatGLM3-6B, scored the highest, suggesting that formulating treatment plans may rely more heavily on knowledge base pretraining. If internet resources were further integrated on the basis of knowledge base pretraining (as in ChatGLM4+ Knowledge Base), the model's performance could be even better enhanced.

Part 3, Health management capacity (see Figure 3C): The evaluation of health management capacity was divided into 4 components: dietary recommendations, exercise recommendations, emotional management recommendations, and weight management recommendations (see Table 5). The results indicated that ChatGLM4 achieved the highest overall performance in health management ability, excelling particularly in dietary, exercise, and emotional management recommendations. ChatGLM4+ Knowledge Base also performed well, especially in exercise and weight management recommendations, though it slightly lagged behind ChatGLM4 in emotional management. ChatGPT-4.0 showed moderate performance, with strengths in dietary recommendations but weaknesses in exercise and weight management. ChatGLM3-6B scored the lowest across all components, particularly in weight management and emotional management.

XSL•FO
RenderX

Figure 3. Detailed scoring results of the "comprehensive evaluation of question-answering texts" module. (A) Total scores of the "Ability to integrate TCM and Western medicine" submodule. (B) Total scores of the "Ability to formulate treatment plans" submodule. "C" Total scores of the "health management capacity" submodule. "D" Total scores of the "disease monitoring ability" submodule. "E" Total scores of the "Comprehensive evaluation of question-answering texts" module. TCM: traditional Chinese medicine.

XSL•FO
RenderX

Table . Ability of combining Chinese and Western medicine.

Large language models	Scores of integration of western medical examinations with TCM[a] syndrome differentiation	Scores of combining western medical examinations for holistic evaluation
ChatGPT-4.0	29.67	35.00
ChatGLM4	42.67	23.67
ChatGLM4+ Knowledge Base	31.33	35.33
ChatGLM3-6B	13.00	33.67

[a]TCM: traditional Chinese medicine.

Table . Ability to formulate traditional Chinese medicine (TCM) treatment plans.

Large language models	TCM[a] prescription	TCM herbal medicine	TCM herbal medicine dosage	Chinese patent medicine	Acupuncture and massage points
ChatGPT-4.0	82	71	0	0	0
ChatGLM4	65	44	6	0	7
ChatGLM4+ Knowledge Base	74	60	11	27	24
ChatGLM3-6B	86	70	4	12	0

[a]TCM: traditional Chinese medicine.

Table . Health management capacity.

Large language models	Dietary recommendations	Exercise recommendations	Emotional management recommendations	Weight management recommendations
ChatGPT-4.0	71	36	27	4
ChatGLM4	75	68	45	8
ChatGLM4+Knowledge Base	67	64	32	9
ChatGLM3-6B	65	34	23	2

Part 4, Disease monitoring ability (see Figure 3D): ChatGLM4 + Knowledge Base performed the best in recommendations for the treatment of comorbidities, scoring 16 points, significantly outperforming the other models (see Table S2 in Multimedia Appendix 1). In recommendations for regular check-ups, ChatGLM3-6B achieved the highest score of 49 points (see Table S2 in Multimedia Appendix 1). In the total score ranking, ChatGLM3-6B ranked first, which may be attributed to the comprehensive health monitoring guidelines included in its local knowledge base. In contrast, although ChatGLM4+ Knowledge Base also integrates a knowledge base, its potential limitations in combining local knowledge with internet resources may have led to its third-place ranking in total score. ChatGPT-4.0, likely benefiting from its powerful web searching capabilities, ranked second in total score.

Part 5, Self-positioning awareness (see Table S3 in Multimedia Appendix 1): At present, the diagnosis and treatment ability of AI language models was still in the preliminary exploration stage, and the responsibility of AI was to assist doctors in diagnosis and treatment, rather than replace the role of doctors. In the evaluation of self-positioning awareness, ChatGLM4 scored the highest (67 points), significantly outperforming the other models, followed by ChatGLM4+ Knowledge Base (47 points) and ChatGPT-4.0 (27 points), ChatGLM3-6B scored the lowest (10 points). These results indicated that large language models with networking capabilities had a clearer sense of self-positioning. In contrast, the offline and non-networked ChatGLM3-6B demonstrated weaker self-positioning awareness, likely due to its lack of dynamic updates and real-time interaction capabilities.

Part 6, Medication safety (see Table S4 in Multimedia Appendix 1): In all of the AI medical record analysis answers, no toxic Chinese medicine was found, indicating that LLMs possessed a certain level of awareness regarding medication safety when recommending TCM.

Part 7, Total score ranking (see Figure 3E): Based on the respective weights of the 6 modules, calculate the total score of these 6 sections to comprehensively analyze the answer text content of the 4 models. The results showed that ChatGLM4+ Knowledge Base ranked first with a total score of 488.93 points, significantly outperforming the other models. ChatGLM4 ranked second with 429.16 points, followed by ChatGPT-4.0 with 409.23 points in third place, while ChatGLM3-6B ranked fourth with 396.93 points, scoring significantly lower than the other three models.

XSL•FO
RenderX

AI-Assisted TCM Doctor Diagnosis and Treatment Flow Chart

In order to standardize the AI-assisted diagnosis and treatment in the future, we have conceived the process of assisted diagnosis and treatment, as shown in Figure 4, which is mainly divided into the following steps:

First, we collect the basic information of the patient, including current history, past history, auxiliary examination results, and key data such as tongue and pulse. This information is then analyzed using AI to identify physical conditions, assess symptoms and syndromes, and make treatment recommendations accordingly. This patient information will be used for further research on assisted diagnosis and treatment of diseases, new drug research and development, medical education, etc, to enhance AI's ability in assisted diagnosis and treatment and medical education.

Figure 4. Artificial intelligence (AI) language model for integrated Chinese and Western medicine assisted diagnosis and treatment concept flow chart.

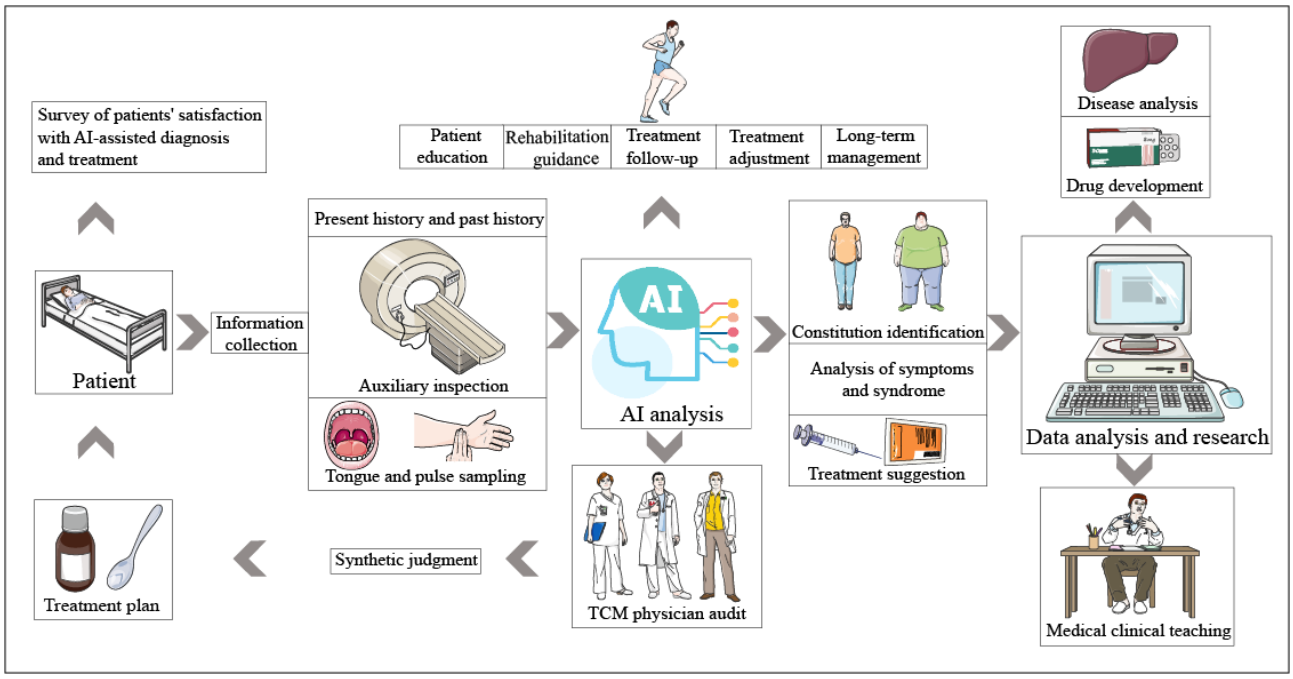

Subsequently, TCM doctors carefully review the diagnosis and treatment recommendations put forward by AI and make comprehensive judgments based on their own professional experience, and finally tailor personalized treatment plans for patients. In order to ensure the quality of service and patient satisfaction, patients are regularly followed up.

In addition, the AI system can also provide health education, rehabilitation guidance, and treatment tracking services according to the specific situation of the patient, while giving adjustment suggestions to the doctor's treatment plan and long-term management of patients with chronic diseases. Through this process, we aim to improve the efficiency and accuracy of care, reduce the workload of clinicians, and ultimately provide patients with a better service experience.

Discussion

Principal Findings

This study systematically evaluated the application performance of LLMs in clinical practice of TCM, yielding the following key findings: First, real-time internet connectivity significantly enhances the models' TCM-assisted diagnosis and treatment capabilities. Second, a notable synergistic effect exists between real-time internet connectivity and TCM domain knowledge pretraining, with their combination more effectively improving the overall performance of LLMs in TCM clinical applications. Third, large language models must undergo specialized TCM-oriented fine-tuning, which not only meets the professional requirements of TCM clinical practice but also effectively prevents the confusion of concepts between Chinese and Western medicine, ensuring the professionalism and accuracy of diagnostic recommendations. These findings provide important theoretical foundations and practical guidance for the in-depth application of AI in the field of TCM, holding significant value for promoting the intelligent development of TCM.

Analysis and Discussion of "Ability in Syndrome Differentiation and Treatment Principles" and "Conceptual Confusion Between TCM and Western Medicine"

Syndrome differentiation and treatment constitute the core component of the theoretical system of TCM, forming the foundational framework for TCM diagnosis and therapy. In clinical TCM practice, physicians conduct comprehensive analysis and judgment based on a patient's specific symptoms, physical signs, and other multifaceted factors to determine the type, nature, and stage of the disease. This process not only embodies the personalized treatment philosophy of TCM but also ensures the specificity and effectiveness of the therapeutic approach. Therefore, we first analyzed the question-answering texts of large language models regarding syndrome differentiation and treatment principles. The ranking of scores

XSL•FO
RenderX

in this section was (1) ChatGLM4+ Knowledge Base, (2) ChatGLM4, (3) ChatGLM3-6B, and (4) ChatGPT-4.0. The ChatGLM4+ Knowledge Base model performed the best in this section (ranking first with a total score of 486), achieving the highest scores in all 3 subcomponents: consistency in syndrome differentiation, explanation of syndrome differentiation, and explanation of treatment principles. This highlights the significant advantage of integrating domain-specific knowledge documents into AI systems, enabling more accurate and reliable responses in specialized fields such as TCM. The performance of the ChatGLM4+ Knowledge Base model surpassed that of the purely internet-connected ChatGLM4 model and the locally deployed ChatGLM3-6B model, which, despite incorporating a knowledge base, lacks internet connectivity. This points to a promising direction for the future development of AI-assisted diagnosis and treatment: a hybrid model that combines pretrained knowledge bases with real-time internet search capabilities.

It was worth noting that ChatGPT-4.0 scored the lowest among the 4 models. In addition, in the evaluation of confusion between TCM and Western medicine concepts, ChatGPT-4.0 made conceptual errors in 32 out of 87 cases. Specifically, it incorrectly used terms such as “regulating liver function,” “preventing liver fibrosis,” and “enhancing digestive absorption function,” which do not align with TCM’s principles of syndrome differentiation and treatment. This may be attributed to its training data being predominantly in English, with relatively limited quality and quantity of Chinese data, leading to its poor performance in providing syndrome differentiation and treatment principles and its frequent confusion between TCM and Western medicine concepts. To improve its performance in TCM applications, ChatGPT-4.0 should expand its Chinese and TCM-specific training data while incorporating language- and culture-specific fine-tuning, thereby enhancing its accuracy in delivering proper syndrome differentiation and treatment principles and preventing conceptual confusion between TCM and Western medicine.

Comprehensive Evaluation of Question-Answering Texts From LLMs

By analyzing the answering texts across 6 dimensions (ability to integrate TCM and Western medicine, ability to formulate treatment plans, health management capacity, disease monitoring ability, self-positioning awareness, and medication safety), we systematically evaluated the accuracy, logic, and practicality of the texts generated by different models. The results showed that ChatGLM4+ Knowledge Base (combining internet search capabilities with knowledge base pretraining) ranked first in total score, significantly outperforming ChatGLM4 and ChatGPT-4.0, which relied solely on internet search capabilities. These findings indicated that the combination of internet search functionality and knowledge base pretraining could significantly enhance the accuracy, logic, and practicality of medical response texts. The future development trend of AI language-assisted diagnosis and treatment should focus on the “knowledge base pre-training + internet search” model. By integrating domain-specific expertise with real-time data retrieval, this approach can further improve the accuracy, logic, and practicality of medical texts, providing more reliable support for AI applications in the medical field. It was noteworthy that the locally deployed ChatGLM3-6B model, which only incorporated knowledge base pretraining, ranked fourth with a score of 380.4 points, performing significantly worse than the other 3 models equipped with internet connectivity. This result highlighted the limitations of relying solely on knowledge base pretraining, particularly in handling dynamic and complex medical scenarios, where the lack of real-time data integration capability became a critical weakness.

Advantages and Challenges of LLMs in Medical-Assisted Diagnosis

By analyzing 87 TCM medical records of nonalcoholic fatty liver disease using ChatGPT-4.0 and ChatGLM models, the generated responses were manually scored and evaluated. Each language model demonstrated unique strengths. Overall, ChatGLM4+ Knowledge Base (combining internet search capabilities with knowledge base pretraining) significantly outperformed ChatGLM4 (with internet connectivity but no knowledge base pretraining) and ChatGLM3-6B (offline knowledge base pretraining but no internet connectivity). Compared to its predecessor, ChatGPT3.5, ChatGPT-4.0 showed significant improvements in analyzing and addressing medical issues within a Chinese context. However, from the perspective of TCM knowledge in this study, ChatGPT-4.0 still has room for improvement in its reasoning and analytical capabilities for clinical problems in a Chinese context, lagging behind ChatGLM4+ Knowledge Base, which was specifically optimized for Chinese contexts and incorporates knowledge base pretraining. Previous reports have also suggested that the balance between generalization and accuracy in GPT-based models like ChatGPT may need adjustment to better adapt to the conditions of handling Chinese medical issues [4].

LLM has shown significant practical significance in the field of medical assisted diagnosis and treatment. It can assist doctors in improving diagnosis and treatment efficiency, achieving more accurate diagnosis, and making diagnosis and treatment plans more efficiently by efficiently analyzing and processing patients’ medical records, symptoms, examination reports, and other information. It can also extend high-quality medical resources to remote and economically underdeveloped areas to narrow the gap between regional medical resources. Provide diagnosis and treatment recommendations based on the latest medical guidelines and research results, and promote the standardization of medical services. It provides support for scientific research and medical education, such as assisting in disease trend analysis, drug research and development, etc, and can also be used as a tool for medical education to help medical students and doctors to learn cases and train skills.

However, the disadvantages should not be ignored. The practice of medicine relies on experienced doctors and evidence-based decision-making, which requires a high degree of transparency. When applying AI in health care, the lack of transparency of the model can hinder the trust of clinicians and patients, thus limiting its scope of application [31]. To gain widespread application in key areas such as health care, AI models must overcome these challenges and improve their transparency and

XSL•FO
RenderX

interpretability to ensure they can be trusted and accepted by health care professionals and patients.

Over-reliance on AI assistance may lead to a decline in doctors' own clinical experience and judgment. Medical data involves patient privacy, and connecting models to the internet may increase the risk of data breaches. The decision of the model is based on the existing data, and if the data is biased or the model is not perfect, it may lead to misdiagnosis. When errors occur in AI-assisted diagnosis and treatment, the attribution of responsibility may become complicated, increasing the difficulty of handling doctor-patient disputes. At present, AI technology has not reached the stage of full maturity, especially in dealing with complex diseases and rare diseases; there are still technical limitations. The development and maintenance of advanced AI-assisted diagnosis and treatment systems will require significant investment, which may increase health care costs.

To sum up, AI large language models have great potential in the field of medically assisted diagnosis and treatment, but at the same time, it is necessary to be careful about the problems they may bring. Reasonable regulation, continued advances in technology, and collaboration between doctors and AI will be key to promoting the healthy development of AI in the medical field.

Limitations

Despite the integration of specialized knowledge bases such as the Expert Consensus on TCM Diagnosis and Treatment of Metabolic-Associated Fatty Liver Disease (2023 Edition), the dynamic and heterogeneous nature of online medical information remains a potential source of interference for the optimal-performing LLM in this study. For example, internet search results may encompass fragmented knowledge that lacks peer review (eg, patient forum experiences); misunderstandings of TCM syndrome concepts within some Western medical literature might inadvertently be incorporated into the model through cross-validation; and real-time retrieval of drug monographs might lag behind the most recent clinical guidelines. Future directions for AI-assisted TCM diagnosis and treatment can encompass strengthening the screening and filtering of online information, consistently integrating the latest therapeutic evidence and guidelines, and augmenting the model's adaptability and learning capabilities, ultimately enhancing the accuracy and reliability of AI-assisted TCM diagnosis and treatment methods.

Another limitation concerns the sample size. Although the 87 cases included in this study were rigorously selected from authoritative sources, the relatively small sample size may restrict the generalizability of the findings. Future research should aim to expand the sample scale and enhance clinical diversity to further validate and extend the discoveries of this study. As an exploratory investigation, this paper establishes a methodological foundation for subsequent large-scale surveys.

Because ChatGPT and ChatGLM models are constantly updated, the model versions we use for validation may not accurately reflect the latest performance. Since clinical case reports on TCM treatment for MAFLD are primarily published in Chinese medical journals, with no relevant cases retrievable from PubMed or other English databases, the absence of English-language TCM medical records for MAFLD makes it particularly challenging to accurately assess LLMs' capability in assisting TCM diagnosis and treatment within English-language contexts. At present, our validation is only based on TCM treatment of MAFLD, and in the future, we can consider adding more diseases for adjuvant diagnosis and treatment tests to assess the model's capabilities more comprehensively. It should be noted that it is not enough to evaluate the ability of AI in TCM clinical diagnosis and treatment based on the performance of a single disease. This case analysis is limited to a single disease, and the number of records is limited. To validate the model more fully, a comprehensive assessment using a wider variety of diseases and a larger number of medical records is needed. In order to improve the personalized level of TCM prescription, we can consider fine-tuning the model to generate TCM prescriptions that are more in line with patients' specific conditions, rather than just relying on standard TCM prescription templates.

When evaluating the text responses generated by the LLMs, we found that these models tended to provide fixed TCM prescription templates and failed to flexibly adjust the use of herbs according to the specific conditions of patients, which is inconsistent with the personalized treatment concept of TCM. In the future, with further pretraining and fine-tuning, we expect that the LLM can generate personalized TCM-assisted diagnosis and treatment dialogue models for different diseases and populations.

Due to resource and cost constraints, we have only locally deployed the ChatGLM3-6B model at this time and have not yet made further fine-tuning. The size of the underlying model parameters and the amount of data used in this study are relatively small, which may lead to some errors, such as the inclusion of English characters in the generated Chinese text or the appearance of information illusion. To address this, we plan to experiment with larger scale base models and use manual evaluation methods to ensure higher quality.

In addition, we aim to expand the corpus not only to TCM but also to other drug classes to improve the model's performance on diverse data and make it applicable to a wider range of medical conditions. In terms of model training, we plan to incorporate image information from medically assisted examinations, such as the patient's tongue image or auxiliary examination images, to realize multimodal medical queries and further improve the accuracy and safety of diagnosis.

Comparison With Previous Work

At present, there is no detailed public report or research literature on the specific application progress of LLMs, such as ChatGPT-4.0 and ChatGLM, in the field of TCM. Our study preliminarily confirmed the potential of LLM in TCM-assisted diagnosis and treatment.

Potential Future Research Directions

On the basis of the previous research, we can speculate some possible application directions according to the characteristics of LLM and the needs of the TCM field.

XSL•FO
RenderX

Chinese Medicine Question and Answer

LLMs can be used as a Chinese medicine question-and-answer system to help users obtain information about Chinese medicine, prescriptions, treatment methods, etc.

TCM Intelligent Diagnosis and Decision Making

Using the natural language processing ability of LLM, tools can be developed to assist TCM diagnosis, collect disease information through natural language dialogue with patients, preliminarily judge the disease by analyzing the symptoms of patients, and provide corresponding personalized TCM diagnosis suggestions in combination with the specific illness and physique of patients, and assist doctors in formulating treatment plans.

TCM college education and TCM knowledge popularization at the grassroots level: LLM can be used as an educational tool for TCM learners' self-study or teachers' teaching assistance, providing online learning and training resources for TCM students and practitioners, providing TCM-related knowledge points explanation, case analysis, and other functions to help them master the basic knowledge and skills of TCM. In addition, it can also provide relevant knowledge of Chinese medicine to the public, including the efficacy, usage, contraindications, etc, to help the public better understand and use Chinese medicine.

Research and interpretation of classical Chinese medicine literature: LLMs can assist researchers in quickly searching and analyzing classical Chinese medicine-related literature, such as Huangdi Neijing and Compendium of Materia Medica, extracting and summarizing medical knowledge in them, and improving research efficiency.

TCM intelligent health consultation: In the field of health consultation, LLM can provide TCM health care and conditioning suggestions for individual health conditions, which has positive significance for reducing medical burden and improving people's quality of life.

It is worth noting that these applications are still in the stage of continuous development and improvement, and although LLM is excellent in processing natural language, its application in the field of TCM still needs to be combined with the experience and judgment of professional physicians and cannot rely solely on the advice provided by robots. Although ChatGPT-4.0 lacks the ability to understand Chinese compared with ChatGLM, ChatGPT-4.0's powerful English comprehension and reasoning ability is obvious to all. With the continuous development and optimization of technology, ChatGPT-4.0 can be systematically trained on Chinese medicine knowledge in the future to form a Chinese medicine English question-and-answer model. It will promote the spread and exchange of Chinese medicine in English-speaking countries. At present, with the development and optimization of artificial intelligence technology and the accumulation of TCM data, ChatGPT-4.0 and ChatGLM will have more breakthroughs and progress in the field of TCM, and their applications will be more and more extensive, providing more support for the development and popularization of TCM.

Conclusions

The superior performance of ChatGLM4+ Knowledge Base across both evaluation modules—"Ability in syndrome differentiation and treatment principles" and "Comprehensive evaluation of question-answering texts"— demonstrated the significant advantages of integrating domain-specific knowledge bases with real-time internet search capabilities in AI-assisted TCM diagnosis and treatment. Furthermore, the integration of real-time internet search functionality proved essential for AI applications in TCM practice, with offline-only models showing significantly poorer performance in clinical decision support. It should be noted that general-purpose LLMs require extensive domain-specific fine-tuning and cultural adaptation for effective TCM applications.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

XW contributed to data curation, formal analysis, investigation, writing the original draft, and writing – review and editing. TZ contributed to data curation, formal analysis, investigation, and writing the original draft. BL contributed to the investigation and methodology and handled resources. ZP and KM contributed to the investigation. CL contributed to conceptualization, funding acquisition, methodology, resources, supervision, and writing – review and editing. CL and BL provided guidance.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary tables.
[DOCX File, 17 KB - formative_v9i1e66503_app1.docx]

References

XSL•FO
RenderX

1. Levartovsky A, Ben-Horin S, Kopylov U, Klang E, Barash Y. Towards AI-augmented clinical decision-making: an examination of ChatGPT's utility in acute ulcerative colitis presentations. Am J Gastroenterol 2023 Dec 1;118(12):2283-2289. [doi: 10.14309/ajg.0000000000002483] [Medline: 37611254]
2. FENG W, REN C, LUO P, et al. Preliminary exploration of ChatGPT-assisted pediatric diagnosis, treatment and doctor-patient communication. West China Med J 2024 Sep 3:1-4 [FREE Full text]
3. Liu C, Sun K, Zhou Q, et al. CPMI-ChatGLM: parameter-efficient fine-tuning ChatGLM with Chinese patent medicine instructions. Sci Rep 2024 Mar 16;14(1):6403. [doi: 10.1038/s41598-024-56874-w] [Medline: 38493251]
4. Wang H, Wu W, Dou Z, He L, Yang L. Performance and exploration of ChatGPT in medical examination, records and education in Chinese: pave the way for medical AI. Int J Med Inform 2023 Sep;177:105173. [doi: 10.1016/j.ijmedinf.2023.105173] [Medline: 37549499]
5. Gofton C, Upendran Y, Zheng MH, George J. MAFLD: how is it different from NAFLD? Clin Mol Hepatol 2023 Feb;29(Suppl):S17-S31. [doi: 10.3350/cmh.2022.0367] [Medline: 36443926]
6. Ali SMJ, Lai M. Metabolic dysfunction-associated steatotic liver disease. Ann Intern Med 2025 Jan;178(1):ITC1-ITC16. [doi: 10.7326/ANNALS-24-02933] [Medline: 39805112]
7. Sung KC, Ryu S, Lee JY, Kim JY, Wild SH, Byrne CD. Effect of exercise on the development of new fatty liver and the resolution of existing fatty liver. J Hepatol 2016 Oct;65(4):791-797. [doi: 10.1016/j.jhep.2016.05.026] [Medline: 27255583]
8. Tsunoda K, Kitano N, Kai Y, Jindo T, Uchida K, Arao T. Dose-response relationships of accelerometer-measured sedentary behaviour and physical activity with non-alcoholic fatty liver disease. Aliment Pharmacol Ther 2021 Nov;54(10):1330-1339. [doi: 10.1111/apt.16631] [Medline: 34633105]
9. Abdelbasset WK, Tantawy SA, Kamel DM, Alqahtani BA, Soliman GS. A randomized controlled trial on the effectiveness of 8-week high-intensity interval exercise on intrahepatic triglycerides, visceral lipids, and health-related quality of life in diabetic obese patients with nonalcoholic fatty liver disease. Medicine (Baltimore) 2019 Mar;98(12):e14918. [doi: 10.1097/MD.0000000000014918] [Medline: 30896648]
10. van Kleef LA, Hofman A, Voortman T, de Knegt RJ. Objectively measured physical activity is inversely associated with nonalcoholic fatty liver disease: the Rotterdam study. Am J Gastroenterol 2022 Feb 1;117(2):311-318. [doi: 10.14309/ajg.0000000000001584] [Medline: 34904966]
11. Harrison SA, Bedossa P, Guy CD, et al. A phase 3, randomized, controlled trial of resmetirom in NASH with liver fibrosis. N Engl J Med 2024 Feb 8;390(6):497-509. [doi: 10.1056/NEJMoa2309000] [Medline: 38324483]
12. Armstrong MJ, Gaunt P, Aithal GP, et al. Liraglutide safety and efficacy in patients with non-alcoholic steatohepatitis (LEAN): a multicentre, double-blind, randomised, placebo-controlled phase 2 study. Lancet 2016 Feb;387(10019):679-690. [doi: 10.1016/S0140-6736(15)00803-X]
13. Musso G, Cassader M, Paschetta E, Gambino R. Thiazolidinediones and advanced liver fibrosis in nonalcoholic steatohepatitis: a meta-analysis. JAMA Intern Med 2017 May 1;177(5):633-640. [doi: 10.1001/jamainternmed.2016.9607] [Medline: 28241279]
14. Sanyal AJ, Chalasani N, Kowdley KV, et al. Pioglitazone, vitamin E, or placebo for nonalcoholic steatohepatitis. N Engl J Med 2010 May 6;362(18):1675-1685. [doi: 10.1056/NEJMoa0907929] [Medline: 20427778]
15. Klein EA, Thompson IM Jr, Tangen CM, et al. Vitamin E and the risk of prostate cancer: the Selenium and Vitamin E Cancer Prevention Trial (SELECT). JAMA 2011 Oct 12;306(14):1549-1556. [doi: 10.1001/jama.2011.1437] [Medline: 21990298]
16. Bi Y, Yin B, Fan G, et al. Effects of acupoint therapy on nonalcoholic fatty liver disease: a systematic review and meta-analysis. Complement Ther Clin Pract 2021 May;43:101376. [doi: 10.1016/j.ctcp.2021.101376] [Medline: 33813176]
17. Wang L, Gong B, Zhu LL, et al. Influence of electroacupuncture stimulation on serum and hepatic interleukin-18 expression in nonalcoholic fatty liver rats. Zhen Ci Yan Jiu 2013 Jun;38(3):208-213. [Medline: 24006666]
18. Dai L, Ooi VV, Zhou W, Ji G. Acupoint embedding therapy improves nonalcoholic fatty liver disease with abnormal transaminase: a PRISMA-compliant systematic review and meta-analysis. Medicine (Baltimore) 2020 Jan;99(3):e18775. [doi: 10.1097/MD.0000000000018775] [Medline: 32011470]
19. Zheng C, Nie H, Pan M, et al. Chaihu shugan powder influences nonalcoholic fatty liver disease in rats in remodeling microRNAome and decreasing fatty acid synthesis. J Ethnopharmacol 2024 Jan 10;318(Pt A):116967. [doi: 10.1016/j.jep.2023.116967] [Medline: 37506783]
20. Zhu M, Li M, Zhou W, et al. Qianggan extract improved nonalcoholic steatohepatitis by modulating lncRNA/circRNA immune ceRNA networks. BMC Complement Altern Med 2019 Jul 3;19(1):156. [doi: 10.1186/s12906-019-2577-6] [Medline: 31269941]
21. Qiu J, Chen L, Zhang L, et al. Xie Zhuo Tiao Zhi formula modulates intestinal microbiota and liver purine metabolism to suppress hepatic steatosis and pyroptosis in NAFLD therapy. Phytomedicine 2023 Dec;121:155111. [doi: 10.1016/j.phymed.2023.155111] [Medline: 37804819]
22. Tan YY, Yue SR, Lu AP, et al. The improvement of nonalcoholic steatohepatitis by Poria cocos polysaccharides associated with gut microbiota and NF-κB/CCL3/CCR1 axis. Phytomedicine 2022 Aug;103:154208. [doi: 10.1016/j.phymed.2022.154208]

XSL•FO
RenderX

23. Xu Y, Wang N, Tan HY, Li S, Zhang C, Feng Y. Gut-liver axis modulation of Panax notoginseng saponins in nonalcoholic fatty liver disease. Hepatol Int 2021 Apr;15(2):350-365. [doi: 10.1007/s12072-021-10138-1] [Medline: 33656663]
24. Li L, Lan X, Peng X, et al. Polysaccharide from salviae miltiorrhizae radix et rhizoma attenuates the progress of obesity-induced non-alcoholic fatty liver disease through modulating intestinal microbiota-related gut–liver axis. IJMS 2022 Sep 13;23(18):10620. [doi: 10.3390/ijms231810620]
25. Yang S, Duan Z, Zhang S, et al. Ginsenoside Rh4 improves hepatic lipid metabolism and inflammation in a model of NAFLD by targeting the gut liver axis and modulating the FXR signaling pathway. Foods 2023 Jun 26;12(13):2492. [doi: 10.3390/foods12132492] [Medline: 37444230]
26. Yang M, Xia L, Song J, et al. Puerarin ameliorates metabolic dysfunction-associated fatty liver disease by inhibiting ferroptosis and inflammation. Lipids Health Dis 2023 Nov 24;22(1):202. [doi: 10.1186/s12944-023-01969-y] [Medline: 38001459]
27. Gong LL, Yang S, Zhang W, et al. Akebia saponin D alleviates hepatic steatosis through BNip3 induced mitophagy. J Pharmacol Sci 2018 Apr;136(4):189-195. [doi: 10.1016/j.jphs.2017.11.007] [Medline: 29609842]
28. Jiang G, Chen D, Li W, Liu C, Liu J, Guo Y. Effects of wogonoside on the inflammatory response and oxidative stress in mice with nonalcoholic fatty liver disease. Pharm Biol 2020 Dec;58(1):1177-1183. [doi: 10.1080/13880209.2020.1845747] [Medline: 33253604]
29. Lai YS, Nguyen HT, Salmanida FP, Chang KT. MERTK+/hi M2c macrophages induced by baicalin alleviate non-alcoholic fatty liver disease. Int J Mol Sci 2021 Sep 30;22(19):10604. [doi: 10.3390/ijms221910604] [Medline: 34638941]
30. Zheng Y, Fang D, Huang C, et al. Gentiana scabra restrains hepatic pro-inflammatory macrophages to ameliorate non-alcoholic fatty liver disease. Front Pharmacol 2021;12:816032. [doi: 10.3389/fphar.2021.816032] [Medline: 35115947]
31. Miller KC, Geyer B, Alexopoulos AS, Moylan CA, Pagidipati N. Disparities in metabolic dysfunction-associated steatotic liver disease prevalence, diagnosis, treatment, and outcomes: a narrative review. Dig Dis Sci 2025 Jan;70(1):154-167. [doi: 10.1007/s10620-024-08722-0] [Medline: 39560808]

Abbreviations

AI: artificial intelligence
ChatGLM: Chat Generative Language Model
LLM: large language model
MAFLD: metabolic dysfunction-associated fatty liver disease
TCM: traditional Chinese medicine

Edited by J Sarvestan; submitted 15.09.24; peer-reviewed by D Chrimes, L Raymond Guo; revised version received 02.06.25; accepted 04.06.25; published 25.08.25.

Please cite as:
Wang X, Zheng T, Liu B, Pei Z, Meng K, Ling C
Performance Assessment of ChatGPT-4.0 and ChatGLM Series in Traditional Chinese Medicine for Metabolic Associated Fatty Liver Disease: Comparative Study
JMIR Form Res 2025;9:e66503
URL: https://formative.jmir.org/2025/1/e66503
doi:10.2196/66503

© Xionghui Wang, Tianxiao Zheng, Bo Liu, Zhi Pei, Kaihan Meng, Changquan Ling. Originally published in JMIR Formative Research (https://formative.jmir.org), 25.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

External Exposome Factors and Adverse Heart Failure Outcomes in the OneFlorida+ Network: Retrospective Cohort Study

Wenxi Huang[1], MS; Stephen E Kimmel[2], MD, MSCE; Mustafa Ahmed[3], MD; Steven M Smith[1,3], PharmD; Yao An Lee[1], MS; Lanting Yang[4], PhD; Inmaculada Hernandez[4], PhD; Jiang Bian[5,6,7,8], PhD; Jingchuan Guo[1], MD, PhD

[1]Department of Pharmaceutical Outcomes and Policy, College of Pharmacy, University of Florida, 1889 Museum Road, DSIT 6004, Gainesville, FL, United States

[2]Department of Epidemiology, College of Public Health and Health Professions and College of Medicine, University of Florida, Gainesville, FL, United States

[3]Department of Medicine, College of Medicine, University of Florida, Gainesville, FL, United States

[4]Division of Clinical Pharmacy, Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, La Jolla, CA, United States

[5]Department of Biostatistics and Health Data Science, School of Medicine, Indiana University, Indianapolis, IN, United States

[6]Center for Biomedical Informatics, Regenstrief Institute, Indianapolis, IN, United States

[7]Department of Medicine, Melvin and Bren Simon Comprehensive Cancer Center, Indiana University, Indianapolis, IN, United States

[8]Indiana University Health, Indianapolis, IN, United States

Corresponding Author:
Jingchuan Guo, MD, PhD
Department of Pharmaceutical Outcomes and Policy, College of Pharmacy, University of Florida, 1889 Museum Road, DSIT 6004, Gainesville, FL, United States

Abstract

Background: Heart failure (HF) readmission rates vary across geographic regions in the United States, yet the impact of external exposome factors, such as contextual-level social determinants of health (SDoH), on adverse HF outcomes is not well understood.

Objective: This study aims to examine the association between external exposome factors and the risk of HF readmission and all-cause mortality using a data-driven approach.

Methods: We conducted a retrospective cohort study using electronic health record (EHR) data from the OneFlorida+ Network, including patients hospitalized for HF (HHF) from 2016 to 2022. A total of 1308 external exposome factors, covering a wide range of SDoH (eg, economic stability, education, health care access, natural and built environments, and social context), were linked to patients' EHR data based on their county-level residential location. Patients were followed for 1 year after their first HHF to capture readmission and mortality events. We applied the least absolute shrinkage and selection operator regularization to preselect candidate variables, followed by a 2-phase external exposome-wide association study using mixed-effects logistic regression to identify key factors associated with the composite outcome of 1-year HF readmission and mortality.

Results: Among 63,940 patients with HF (n=30,475, 48% women; mean age 65, SD 14 years), higher maximum temperature in May was significantly associated with increased risk of the composite outcome (adjusted odds ratio [aOR] 1.04, 95% CI 1.02-1.06; $P<.001$). Subgroup analyses showed consistent associations across age, sex, race, socioeconomic status, and rural or urban strata.

Conclusions: Using a data-driven approach, we found that elevated maximum temperature in May (late spring) was significantly associated with HF readmission and mortality in Florida. Further investigations are warranted to uncover the intricate mechanisms through which extreme heat potentially influences HF outcomes.

(JMIR Form Res 2025;9:e71595) doi:10.2196/71595

KEYWORDS

machine learning; social determinants of health; heart failure; electronic health record; climate change

Introduction

Background

Heart failure (HF) is a significant public health challenge in the United States, being the second most common reason for hospital admissions. It accounts for over 4.4 million cases leading to hospitalization annually, with health care costs surpassing US $31 billion [1]. Despite advances in treatment and management strategies, HF remains a chronic and progressive condition with limited prospects for a cure [2]. For

XSL•FO
RenderX

patients hospitalized with HF, readmission rates and mortality serve as key quality indicators, reflecting the efficacy of both inpatient treatment and postdischarge care [3,4]. Approximately 55% of adult patients with HF are readmitted within one year, and the one-year mortality rate is about 25% [5-7].

Large geographic variation exists in the prevalence of HF and the incidence of complications [8], suggesting that the environment may play a critical role in the disease course. Extensive research has delineated clinical predictors of HF hospitalization trajectories; however, less clearly defined is the role of contextual-level socioeconomic, built, and natural environments as modifiable upstream factors impacting patients' hospitalization patterns and mortality risk [9-13]. Such contextual factors are increasingly recognized to influence health outcomes and serve as a critical source of information to develop policy interventions designed to improve population health management and value-based care [14,15]. While previous studies have identified multiple contextual exposures associated with HF outcomes in the natural, built, and social environments [9-13], these studies only focused on a subset of preselected environmental factors. In addition, these studies assessed these factors in isolation without considering the totality of the environment or the external exposome [16]. Significant knowledge gaps remain regarding the combined effects of multiple environmental exposures on cardiovascular health, and there is a need for a more comprehensive approach that considers the full spectrum of external factors an individual is exposed to over a lifetime.

The external exposome-wide association study (ExWAS) framework addresses this need by providing a systematic and efficient approach to screen hundreds of environmental exposures simultaneously, offering an opportunity to identify novel environmental factors associated with HF outcomes [17,18].

Objective

In this study, we applied the ExWAS framework to identify novel external exposome factors associated with HF readmission and mortality [16,18-20]. We used electronic health records (EHRs) to examine the association between the external exposome and the risk of HF admission and mortality. Specifically, we integrated data on a wide range of external exposome factors, including economic stability, access to and quality of education, health care access, the natural environment, and the neighborhood and built environment surrounding patients who were hospitalized for heart failure (HHF). A more nuanced understanding of which external exposome factors influence adverse HF outcomes is critical to tailor care coordination, transitional support, and preventive services to patients, ultimately helping clinicians and health care systems reduce the HF burden.

Methods

Data Source and Study Population

We conducted a retrospective cohort study using 2016 - 2022 EHRs from the OneFlorida+ Clinical Research Network [21]. OneFlorida+ Clinical Research Network is one of the largest statewide clinical data repositories, encompassing data from 13 unique health care systems that collectively serve approximately half of Florida's population (~17.2 million individuals), covering all 67 Florida counties [22]. OneFlorida+ data are racially and ethnically diverse, with a mixture of rural-urban populations, reflecting national demographic changes, thus enhancing the generalizability of our study [22,23]. In addition, OneFlorida+ is linked with the National Death Index, allowing for mortality status tracking of its patients [22]. These longitudinal, patient-level data were chosen to enable a comprehensive analysis of HF adverse outcomes by leveraging extensive clinical and demographic information across a large, diverse population in Florida.

The study cohort included patients aged ≥18 years who had at least one HHF between January 1, 2016, and January 31, 2021. We excluded individuals missing a residential history (county), patients residing outside of Florida, and those with less than one year of health care encounter history before the first observed HHF. HHF was defined as a primary admission diagnosis with HF, identified using *ICD-10* (*International Classification of Diseases, Tenth Revision*) codes (I50.x, I11.0, I13.0, I13.2, I97.13, and I09.81) [24]. This algorithm was previously validated against discharge summary or medical record by Thygesen et al [24], with a positive predictive value of 100% (95% CI 92.9 - 100) based on primary diagnoses.

The index date was defined as the date of the patient's first HHF recorded in the EHR during the study period. We followed each patient for one year after their index hospitalization until death or the end of the study period, whichever occurred first.

Study Outcome

Our primary outcome was a composite of HF readmission or all-cause mortality within one year after the index date, defined as HF readmission or death, whichever occurred first. We included death as part of the composite outcome to account for competing risks. In HF studies, patients who die before a potential readmission are no longer at risk for readmission, which could bias readmission risk estimates. By incorporating death into the composite outcome, we provide a more accurate estimate of the true burden of HF, capturing both readmission and mortality risks. HF readmission was identified using *International Classification of Diseases, Ninth Revision: Clinical Modification* (*ICD-9-CM*) and *International Classification of Diseases, Tenth Revision: Clinical Modification* (*ICD-10-CM*) codes as mentioned in the "Data Source and Study Population" section. All-cause mortality was identified by death records in OneFlorida+ EHR, which was linked with the National Death Index. Our secondary outcome was HF readmission alone within one year after the index date.

Exposures of Interest

Our exposures of interest included 1308 contextual-level exposome factors obtained from 11 well-validated sources, including the Agency for Healthcare Research and Quality [25], County Health Roadmap [26], Local Area Unemployment Statistics [27], economic resilience [28], US Cancer Statistics [29], Dartmouth Health [30], Air Quality Index [31], Air Quality System [31], Local Area Unemployment Statistics [27], religion

XSL•FO
RenderX

[32], water [33], and social capital [34]. We categorized these factors into 6 domains, inspired by the Healthy People 2030 framework [35] (refer to Table 1) social and community context (eg, food access), neighborhood and built environment (eg, hospital density), natural environment (eg, temperature), health care access and quality (eg, insurance), education access and quality (eg, graduation rate), and economic stability (eg, income). These contextual-level exposome factors were temporospatially linked to each patient's EHR data based on their residential histories at the county level, using Federal Information Processing Standard codes.

Table . Summary of external exposome measures included in our retrospective cohort, exposome-wide analysis.

Category and data source[a]	Time period	Number of variables
Economic stability		
Agency for Healthcare Research and Quality	2009 - 2019	88
County Health Roadmap	2009 - 2019	10
Local Area Unemployment Statistics	2019	4
Economic resilience	2015	3
Education access and quality		
Agency for Healthcare Research and Quality	2009 - 2019	19
County Health Roadmap	2009 - 2019	2
Health care access and quality		
US Cancer Statistics	2018	8
Agency for Healthcare Research and Quality	2009 - 2019	451
County Health Roadmap	2009 - 2019	54
Dartmouth Health	2009 - 2019	8
Natural environment		
Agency for Healthcare Research and Quality	2009 - 2019	97
Neighborhood and built environment		
Agency for Healthcare Research and Quality	2009 - 2019	117
Air Quality Index	2009 - 2019	15
Air Quality System	2009 - 2019	14
County Health Roadmap	2009 - 2019	18
Local Area Unemployment Statistics	2009 - 2019	2
Religion	2010	3
Water	2017	242
Social and community context		
Social capital	2009 - 2014	13
Agency for Healthcare Research and Quality	2009 - 2019	138
County Health Roadmap	2009 - 2019	2

[a]All data sources are spatially scaled at the county level.

Covariates

Covariates were measured during the lookback period (1 year before the index HHF). Covariates were selected based on previous studies [36-39] and clinical experience. These included age, race/ethnicity (non-Hispanic White, non-Hispanic Black, Hispanic, and others), sex, HF subtype (HF with reduced vs preserved ejection fraction), and comorbidities (chronic obstructive pulmonary disease [COPD], myocardial infarction, anemia, diabetes, cancer, and the use of an implantable cardioverter-defibrillator [ICD], cardiac resynchronization therapy [CRT], or both), as well as medication use (sacubitril/valsartan, sodium-glucose cotransporter-2 inhibitors, nitrates [isosorbide mononitrate, nitroglycerin, isosorbide dinitrate], loop diuretics, and angiotensin-converting enzyme inhibitors). Comorbidities were identified using *ICD-9-CM* or *ICD-10-CM* diagnosis and procedure codes (Table S1 in Multimedia Appendix 1) [9,40-43]. Medication use was identified from patient prescription records.

XSL•FO
RenderX

Area Deprivation Index

The Area Deprivation Index (ADI) is a metric developed to assess the level of socioeconomic disadvantage in a given geographic area [44]. It was initially created by researchers at the University of Wisconsin-Madison and has become widely used in health research to understand how neighborhood-level socioeconomic factors influence health outcomes. The ADI provides a composite score based on several indicators, capturing various dimensions of deprivation, such as education, employment, housing quality, and income [44]. The ADI is a continuous value, with higher scores indicating areas with greater socioeconomic disadvantage. We converted it to a dichotomized variable, with the top 50% ADI defined as high deprivation. Each patient's ADI was linked at the patient level, based on their Federal Information Processing Standard code.

Statistical Analysis

In the initial stage of our analysis, we conducted data preprocessing for the contextual-level external exposome factors (Figure S1 in Multimedia Appendix 1). We performed normalization transformations on all continuous variables using the bestNormalize package [45] in R statistical software (version 3.6.1; R Development Core Team), which supports various transformation methods, including logarithmic, square root, exponential, arcsinh, Box-Cox, and Yeo-Johnson transformations. The optimal transformation method was selected based on the Pearson *P* statistics. Next, we examined the correlation among variables and removed highly correlated predictors (correlation >0.9) to reduce redundancy. We then used the least absolute shrinkage and selection operator (LASSO) for variable selection to identify key contextual-level variables [46]. The LASSO model was fit using the glmnet package [47], with cross-validation to determine the optimal regularization parameter (λ). To retain more variables, we used Elastic Net regularization, combining Ridge and LASSO penalties (α=.01). Cross-validation identified λ.min, the lambda value that yielded the minimum cross-validated error, which was then used to refit the model. We retained variables with nonzero coefficients in the final model, indicating significant predictors identified by LASSO. This approach effectively reduced multicollinearity while preserving model interpretability. The final set of selected variables was stored for subsequent analysis. Details of the selected transformations and parameters for preselected external exposome variables are provided in Table S2 in Multimedia Appendix 1. Furthermore, we standardized all continuous variables to a *z* score (mean 0, SD 1). The missingness rate for all contextual-level exposome factors and covariates of interest in our model was restricted to <50% [48]. Variables with >50% missingness were excluded for further analysis to avoid biased results and maintain reliability, as extensive missing data can limit the effectiveness of imputation and compromise study validity [49-52]. We imputed missing data for all external exposome factors and covariates using either the median or mode value, depending on the variable type [53].

We then used a 2-phase ExWAS approach, as presented in Figure S1 in Multimedia Appendix 1, to identify key contextual-level factors associated with 1-year HF readmission or mortality [20,54]. In phase 1, we randomly divided the entire dataset into discovery (50%) and replication (50%) sets. We then tested the association of each external exposome factor with our composite outcome using mixed-effects logistic regression models, adjusting for covariates in both sets. Mixed-effect models were fitted for each external exposome factor, adjusting for all potential confounders (age, race, sex, history of heart failure with preserved ejection fraction (HFpEF), myocardial infarction, COPD, anemia, diabetes, cancer, and the presence of ICD and/or CRT-D; prescription history of sacubitril/valsartan, sodium-glucose cotransporter-2 inhibitors, nitrates, loop diuretics, and angiotensin-converting enzyme inhibitors), with population size set as the offset and a random intercept by county. To address the issue of multiple testing, we applied the Benjamini-Hochberg procedure to restrict the false discovery rate at a 5% level [55]. Variables with false discovery rate–adjusted *P* values (or q values) <.05 in both the discovery and replication sets were deemed significant in phase 1. We then assessed the pairwise Pearson correlations among these variables to further reduce collinearity. For variable pairs with absolute correlation coefficients exceeding 0.7, we retained the variable with the lower mean absolute correlation across all variables [56]. In phase 2, we constructed a mixed-effects logistic regression model incorporating all remaining significant variables identified in phase 1, along with all covariates of interest, including age, race, sex, comorbidities, and medication history. We reported the results of this model using adjusted odds ratios (aORs) with corresponding 95% CI for variables that remained significant (q<0.05) in phase 2.

All analyses were conducted using the R statistical software (version 3.6.1; R Development Core Team). Subgroup analyses were performed based on race (ie, White, Black, and so on), sex (female and male), age (≥65 and <65 years), residential region (rural and urban), ADI level (low deprivation and high deprivation), and COVID-19 period (prepandemic: January 2016 to February 2020; pandemic: March 2020 to January 2021).

Ethical Considerations

We obtained ethics approval from the University of Florida Institutional Review Board (202201080) before abstracting data. This study used retrospective, deidentified patient EHR data from the OneFlorida+ Network. Informed consent was waived by the review boards due to the retrospective nature of this research. No patients were directly recruited, and no individual compensation was provided. All data used complied with relevant patient privacy regulations, and no identifiable individuals are depicted in any figures or other study materials.

Results

This study is reported following the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines for cohort studies. Our study cohort consisted of a total of 63,937 patients with an index HHF, of whom 30,475 (47.7%) were women, with a mean age of 65 (SD 14) years (Figure 1, Table 2, Checklist 1).

XSL•FO
RenderX

Figure 1. Patient flowchart. dx: diagnosis.

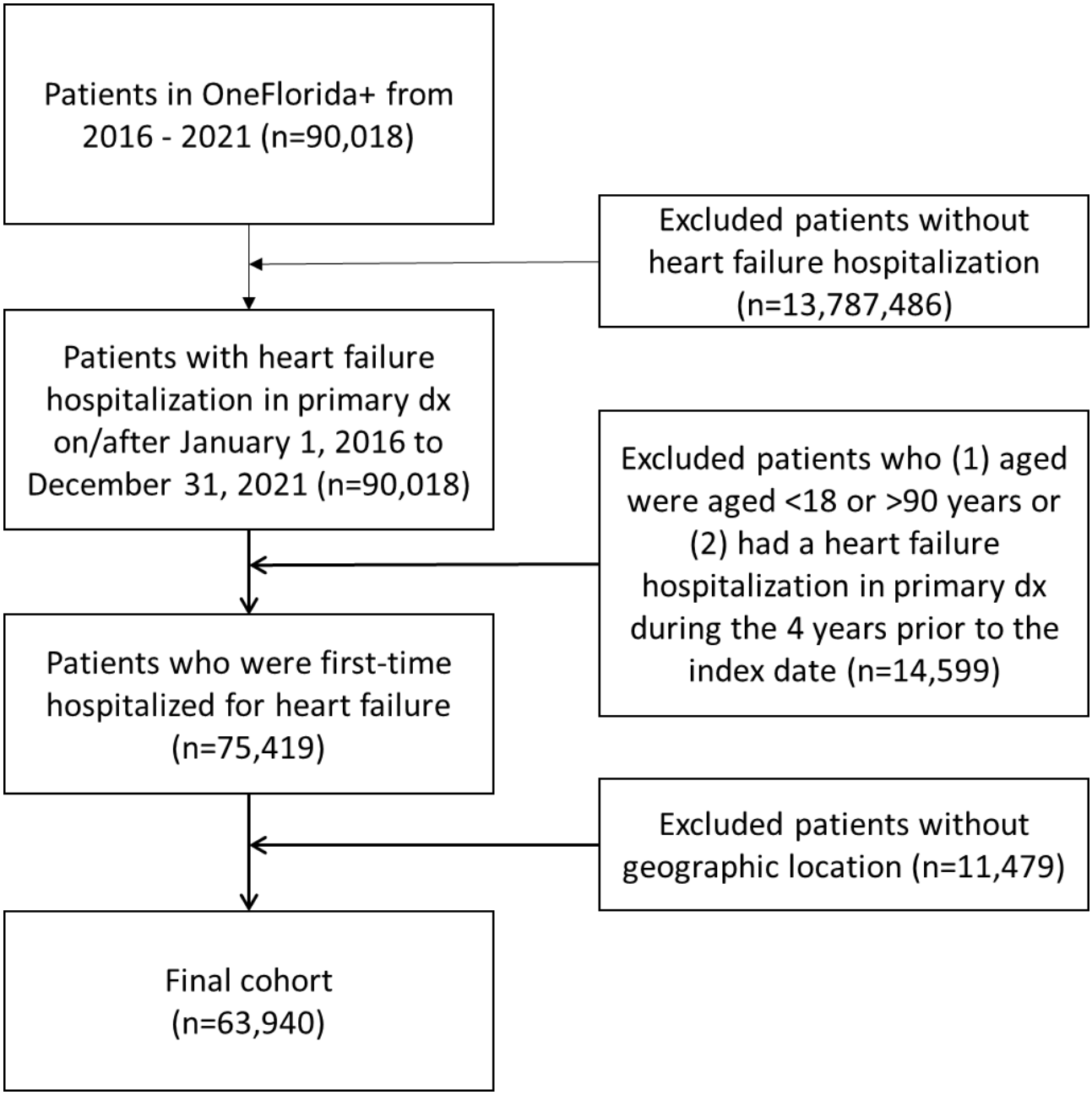

XSL•FO
RenderX

Table . Patient characteristics, by race and ethnicity, among patients hospitalized with heart failure in the OneFlorida+ Network.

Variable	Overall (n=63,940)[a]	Hispanic (n=12,380)[a]	NHB[b] (n=17,853)[a]	NHW[c] (n=27,484)[a]	Other (n=6223)[a]	*P* value[d]
Age (years), mean (SD)	65 (14)	68 (15)	61 (15)	68 (13)	65 (14)	<.001
Female, n (%)	30,475 (47.7)	5872 (47.4)	9024 (50.5)	12,766 (46.4)	2813 (45.2)	<.001
OneFlorida+ Network, n (%)						<.001
A	22,816 (35.7)	3535 (28.6)	5139 (28.8)	9834 (35.8)	4308 (69.2)	
B	11,558 (18.1)	2811 (22.7)	2353 (13.2)	5599 (20.4)	795 (12.8)	
C	10,119 (15.8)	387 (3.1)	3851 (21.6)	5538 (20.2)	343 (5.5)	
D	7807 (12.2)	3942 (31.9)	2710 (15.2)	1036 (3.8)	119 (1.9)	
E	6078 (9.5)	941 (7.6)	1791 (10.0)	3092 (11.2)	254 (4.1)	
Other	5562 (8.7)	764 (6.2)	2009 (11.2)	2385 (8.7)	404 (6.5)	
Anemia, n (%)	30,547 (47.8)	6169 (49.8)	9212 (51.6)	12,505 (45.5)	2661 (42.8)	<.001
COPD[e], n (%)	23,309 (36.4)	4269 (34.5)	5636 (31.6)	11,395 (41.5)	2009 (32.3)	<.001
Diabetes, n (%)	32,615 (51.0)	7194 (58.1)	9655 (54.1)	12,817 (46.6)	2949 (47.4)	<.001
HFpEF[f], n (%)	33,407 (52.2)	6424 (51.9)	9128 (51.1)	14,695 (53.5)	3160 (50.7)	<.001
History of myocardial infarction, n (%)	10,356 (16.2)	2070 (16.7)	2494 (14.0)	4831 (17.6)	961 (15.4)	<.001
ICD and/or CRT-D[g], n (%)	5114 (8.0)	1072 (8.7)	1397 (7.8)	2228 (8.1)	417 (6.7)	<.001
Mechanical ventilation, n (%)	5863 (9.2)	1533 (12.4)	1911 (10.7)	1917 (7.0)	502 (8.1)	<.001
Cancer, n (%)	7419 (11.6)	1472 (11.9)	1859 (10.4)	3518 (12.8)	570 (9.2)	<.001
HF[h] readmission, n (%)	21,442 (33.5)	4418 (35.7)	6847 (38.3)	8194 (29.8)	1983 (31.9)	<.001
HF readmission or all-cause mortality, n (%)	25,078 (39.2)	4977 (40.2)	7897 (44.2)	10,063 (36.6)	2141 (34.4)	<.001

[a]Values are means (SDs) for continuous variables and counts (%) for categorical variables. Values of polytomous variables may not sum to 100% due to rounding.

[b]NHB: non-Hispanic Black.

[c]NHW: non-Hispanic White.

[d]*P* values are based on the Kruskal-Wallis rank sum test or Pearson chi-square test, as appropriate.

[e]COPD: chronic obstructive pulmonary disease.

[f]HFpEF: heart failure with preserved ejection fraction.

[g]ICD and/or CRT-D: implantable cardioverter defibrillator and/or cardiac resynchronization therapy defibrillator within 4 years before the index date.

[h]HF: heart failure.

A total of 33,407 out of the study cohort (52.2%) had HFpEF. Patient characteristics by race and ethnicity are detailed in Table 2. Notably, non-Hispanic Black patients were younger and more likely to have diabetes, anemia, and require mechanical ventilation compared to non-Hispanic White patients. During the one-year follow-up, 33.5% of patients were readmitted for heart failure, and an additional 5.7% died without a prior heart failure readmission, resulting in 39.2% experiencing either heart failure readmission or all-cause mortality. In other words, patients who were readmitted did not overlap with those who died within the follow-up period, indicating that each patient contributed to only one outcome category. After data preprocessing and LASSO variable selection, 96 of the initial 1308 external exposome factors were retained for the ExWAS analysis (refer to Table S2 in Multimedia Appendix 1). In phase 1 of the ExWAS analysis, we identified 3 variables (median income of grandparent householder and/or spouse responsible for grandchildren younger than 18 years, percentage of households not receiving food stamps/Supplemental Nutrition Assistance Program (SNAP) with income below the poverty level, and maximum temperature in May) that were significantly associated with the composite outcome of HHF and death in both the discovery and replication sets (Table 3).

XSL•FO
RenderX

Table . Results from the external exposome-wide association study (ExWAS) in phase 1.

Exposure variable	Category	Transformation	Phase 1, discovery set		Phase 1, replication set		Phase 2	
			OR[a] (95% CI)	q-value	OR[a] (95% CI)	q-value	OR[a] (95% CI)	q-value
Monthly (May) maximum temperature (°F)	Neighborhood and built environment	No transform	1.06 (1.03-1.10)	2.94E-03	1.04 (1.02-1.06)	3.67E-05	1.04 (1.02-1.06)	6.57E-05
Median income of grandparent householder and/or spouse responsible for grandchildren <18 years (US dollars, inflation-adjusted to data file year)	Economic stability	sqrt_x	0.95 (0.92-0.98)	3.86E-02	0.96 (0.94-0.98)	5.42E-03	0.98 (0.96-1.00)	7.64E-02
Percentage of households not receiving food stamps/SNAP[b] with income below the poverty level	Economic stability	Yeo-Johnson	1.05 (1.02-1.07)	4.77E-02	1.05 (1.03-1.07)	3.67E-05	1.02 (1.00-1.04)	8.84E-02

[a]OR: odds ratio.

[b]SNAP: Supplemental Nutrition Assistance Program.

A visual summary of the phase 1 results is presented in Figure S3 in Multimedia Appendix 1 using a volcano plot. This figure illustrates the significance and aOR of the tested variables, with a solid dot indicating the variables that met the significance threshold in both sets. Some categories had zero significant factors, indicating no statistical significance for those exposures in both sets. To address the collinearity among the identified temperature-related variables, we calculated the mean absolute correlation between each pair of variables (Figure S4 in Multimedia Appendix 1).

In phase 2, all significant variables from phase 1 (median income of grandparent householder and/or spouse responsible for grandchildren younger than 18 years, percentage of households not receiving food stamps/SNAP with income below the poverty level, and maximum temperature in May) were incorporated into a multivariable mixed-effect binomial regression model, adjusting for all covariates (age, race, sex, history of HFpEF, myocardial infarction, COPD, anemia, diabetes, cancer, and the presence of ICD and/or CRT-D). Only the maximum temperature recorded in May for the county where the patient resides maintained its statistical significance in the phase 2 analysis, with an aOR of 1.04 (95% CI 1.02-1.06) (Table 3).

Although we only retained the May maximum temperature in the final model, we found that some other monthly temperature variables were also associated with adverse HF outcomes (Table S3 in Multimedia Appendix 1). Our final model retained a single temperature variable based on LASSO regularization, which prioritized predictive strength while mitigating multicollinearity. As such, the broader dynamics of temperature patterns and cumulative heat exposure warrant further exploration in future studies.

Subgroup analyses were conducted to examine the association between May maximum temperature and the composite outcome of 1-year HF readmission and mortality, stratified by race, sex, age group, COVID-19 pandemic period, rurality, and ADI. In our subgroup analysis of the association between May maximum temperature and the outcome, we observed consistent trends across most subgroups, although not all were statistically significant. Among racial groups, the aORs were statistically significant for White (aOR 1.07, 95% CI 1.04-1.09) and Black patients (aOR 1.08, 95% CI 1.03-1.14), indicating a meaningful association in these populations. However, the association was not statistically significant for Hispanic (aOR 1.04, 95% CI 0.99-1.10) or other race groups (aOR 1.02, 95% CI 0.96-1.08). Both female (aOR 1.06, 95% CI 1.03-1.09) and male (aOR 1.06, 95% CI 1.03-1.09) subgroups showed statistically significant associations with the outcome, as did both age groups: individuals aged <65 years (aOR 1.06, 95% CI 1.02-1.09) and those aged ≥65 years (aOR 1.06, 95% CI 1.04-1.09) showed similar effect sizes. Regional differences were also significant, with both urban (aOR 1.06, 95% CI 1.03-1.08) and rural regions (aOR 1.08, 95% CI 1.02-1.15) displaying associations with the outcome. For ADI, individuals in areas of low deprivation showed a significant association (aOR 1.14, 95% CI 1.04-1.26), and those in areas of high deprivation also exhibited a statistically significant association (aOR 1.07, 95% CI

XSL•FO
RenderX

1.03-1.12). In addition, the COVID-19 period analysis indicated a significant association both before the pandemic (aOR 1.06, 95% CI 1.04-1.09) and after the pandemic began (aOR 1.05, 95% CI 1.02-1.09). This detailed breakdown suggests that a higher maximum temperature in May was consistently linked to an increased risk of the adverse composite outcomes for patients with HF, independent of their race, age, or socioeconomic status (SES) (Table 4; Figure 2).

Table . Results from multivariate logistic regression and subgroup analysis for monthly (May) maximum temperature (Fahrenheit).

Variable and subgroup	Adjusted OR[a] (95% CI)
Overall	1.04 (1.02-1.06)
Race	
White	1.07 (1.04-1.09)
Black	1.08 (1.03-1.14)
Hispanic	1.04 (0.99-1.10)
Other	1.02 (0.96-1.08)
Sex	
Female	1.06 (1.03-1.09)
Male	1.06 (1.03-1.09)
Age (years)	
<65	1.06 (1.02-1.09)
≥65	1.06 (1.04-1.09)
Region	
Urban	1.06 (1.03-1.08)
Rural	1.08 (1.02-1.15)
ADI[b]	
Low deprivation	1.14 (1.04-1.26)
High deprivation	1.07 (1.03-1.12)
COVID-19	
After pandemic	1.05 (1.02-1.09)
Before pandemic	1.06 (1.04-1.09)

[a]OR: odds ratio.

[b]ADI: Area Deprivation Index.

XSL•FO
RenderX

Figure 2. Results for multivariate logistic regressions and subgroup analysis for monthly (May) maximum temperature (Fahrenheit). ADI: Area Deprivation Index; OR: odds ratio.

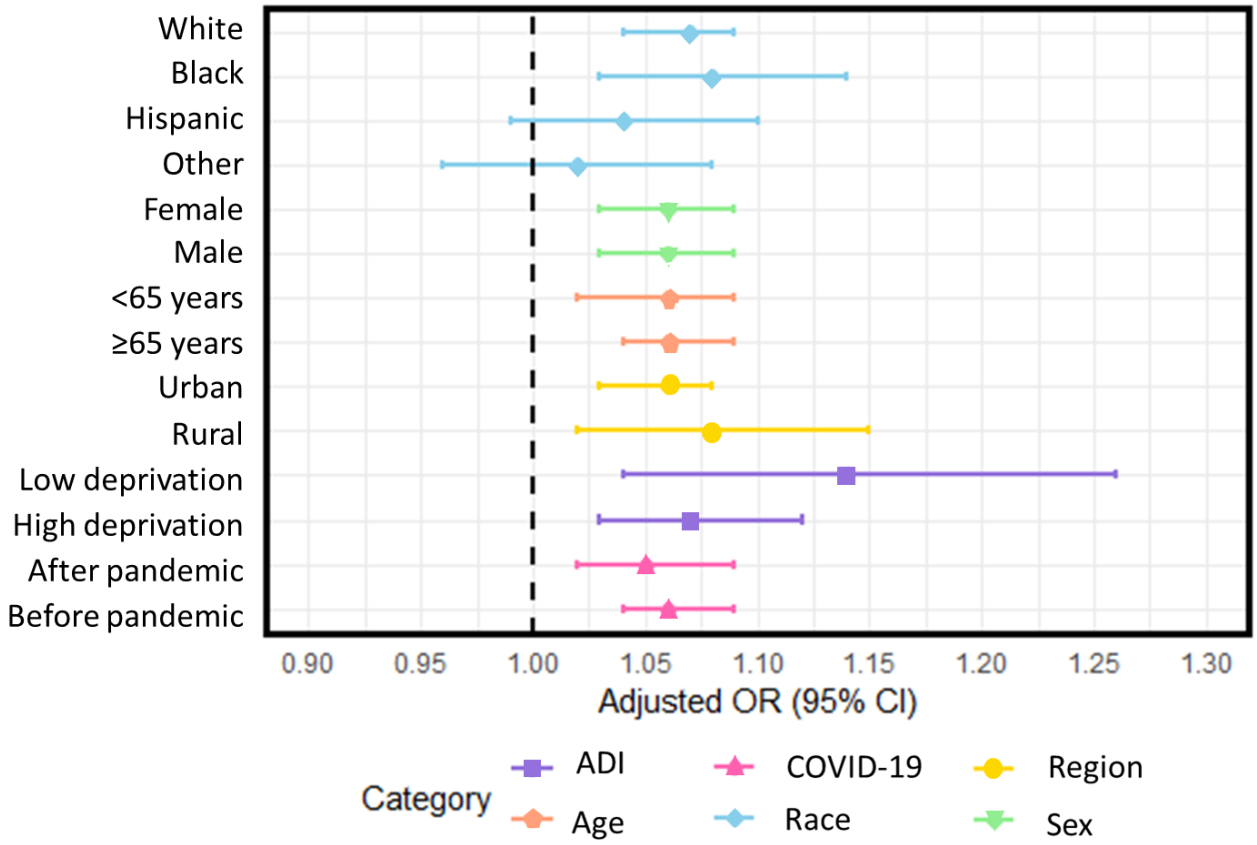

Furthermore, we tested the association between May maximum temperature and the 1-year readmission outcome and observed similar trends, with an aOR of 1.05 (95% CI 1.03-1.07). These results suggest a consistent association between elevated May temperatures and adverse outcomes in patients with HF, reinforcing the robustness of elevated temperature as a positive risk factor for higher HF-related adverse outcomes.

Discussion

Principal Findings

Using the standard 2-phase EWAS-MLR approach, we assessed the association of 1308 external exposome variables in relation to 1-year HF readmission and mortality, leveraging real-world data. We discovered that an elevated maximum May temperature was significantly associated with an increased risk of HF readmission and mortality in Florida.

Concerns have been raised about the interaction between extreme temperatures and factors such as SES, urban-rural differences, and greenspace availability. Populations in higher-SES areas may experience less impact from high temperatures due to better access to air conditioning and health care, mitigating some of the heat's effects [57]. To explore this, we conducted subgroup analyses by SES (using the ADI) and by urban versus rural areas. Our results showed that none of the subgroup variables, including sex, rurality, SES, and COVID-19 period, significantly modified the association between temperature and HF outcomes. This consistency suggests that temperature exerts a direct, uniform impact on HF one-year readmission and mortality, regardless of socioeconomic differences. Additional subgroup analysis revealed that higher maximum temperatures in May were consistently associated with an increased risk of HF 1-year readmission and mortality across most groups, with significant associations observed among White and Black patients, both sexes, and age groups. This consistency suggests that elevated May temperatures pose a broad risk to patients with HF, largely independent of demographic or socioeconomic factors.

Comparison With Prior Work

To our knowledge, this retrospective cohort study is the first to evaluate the impact of a comprehensive array of contextual-level external exposome factors, derived from patients' residential histories, on HF 1-year readmission and mortality among patients with HHF. Our findings align with previous studies on the impact of temperature and climate on cardiovascular disease (CVD) outcomes, which have consistently highlighted a complex relationship, with both extreme heat and cold temperatures being associated with increased risks of various CVD adverse outcomes [58,59]. For example, a study by Lopez et al [58] found that mortality in chronic HF exhibits a seasonal pattern, with a 13.9% higher all-cause mortality rate in winter than in summer. On the other hand, Chan et al [60] found that an average 1°C increase in daily mean temperature above 28.2°C was associated with a 1.8% increase in mortality, with higher

XSL•FO
RenderX

vulnerability observed among women, men younger than 75 years, individuals in low-SES districts, those with unknown residence, married people, and deaths from cardiovascular and respiratory infections. However, differences emerge in the magnitude of risk associated with specific temperature thresholds, which can vary by geographic location, population acclimatization, and study methodologies [59,61-63]. Our findings align with previous studies that an increasing level of heat significantly elevates the risk of CVD-related mortality and morbidity, with pronounced risks for residents of tropical climates [62]. Considering Florida's subtropical environment, which experiences relatively mild winters, the impact of low temperatures on HF outcomes in this setting may be limited.

The potential biological plausibility behind the impact of heat on adverse outcomes of patients with HF can be attributed to several mechanisms. Heat exposure can lead to dehydration and increased blood viscosity, placing additional strain on the cardiovascular system [64]. It can also induce systemic inflammation and oxidative stress, which are known contributors to atherosclerosis and other CVDs [64,65]. Furthermore, heat can exacerbate the body's stress response, increasing heart rate and blood pressure, thereby imposing further risk on individuals with pre-existing cardiovascular conditions [66]. However, previous research by Nam et al [67] has highlighted that increased potassium use during summer months showed a protective effect against all-cause mortality increases with higher temperatures. Understanding these mechanisms is crucial for developing strategies to mitigate adverse outcomes in patients with HHF residing in high-temperature environments.

Limitations

This study has limitations. First, the generalizability of our findings may be limited due to the regional focus of the study. The majority of our cohort population resided in Florida, and as such, our observations may not readily extend to other states with different climatic conditions. Our study did not reveal a notable impact of cold temperatures, as prior studies have, which is likely attributable to Florida's lack of low-temperature winters [63,68]. Future studies using nationally representative data sources should investigate the consistency of these associations across more diverse environmental settings. Second, the EHR data are limited to individuals who made health care encounters within the participating networks, so there would be a chance that we failed to capture HF readmission or mortality that occurred outside of the participating health care systems, which would lead to misclassification in the nonoutcome group. To minimize loss to follow-up for longer-term outcomes, we limited the study follow-up period to one year. Considering that the observed association remains significant despite potential misclassification in the nonoutcome group, this suggests that the effect of maximum temperature in May on adverse HF outcomes may be even stronger than reported, underscoring its potential impact on HR-related adverse outcome risk. Third, we acknowledge that using county-level exposures may overlook intracounty variations that could provide more localized environmental and social insights into HR outcomes. Unfortunately, our data were not available at a finer spatial granularity, such as ZIP code or census tract, which may have limited our ability to detect these localized effects. Future studies with access to more detailed geographic data could offer deeper insights into the impact of environmental exposures on HR, allowing for targeted interventions at the community level. Fourth, our selection of external exposome factors may not fully encompass the wide range of social determinants of health that impact adverse HF outcomes. Future studies could enhance understanding by incorporating additional environmental variables, such as walkability and data from the EPA's Smart Location Database, to provide a more comprehensive view of how the built environment influences health outcomes. Furthermore, our analysis was limited to exposure data available only through 2019, which prevented us from including pandemic-era data (2020‑2022) for variables such as employment, economic resilience, and air quality. Due to a typical 2‑5-year delay in accessing contextual social risk data, we believe these contextual social factors likely remained relatively stable over this short period, minimizing potential effects on our findings. In addition, our effect of the exposome is not quantified as aggregated exposure. While an aggregate exposure approach—combining individual- and contextual-level measures—would ideally capture a more comprehensive picture of the factors affecting cardiovascular outcomes, implementing such a method typically requires a preselected study cohort and prospective data collection. This process can be costly, time-intensive, and may restrict both the sample size and the generalizability of findings. Moreover, measuring temperature by the maximum temperature in May limits our ability to capture the day-to-day and month-to-month temperature fluctuations, which could affect HF outcomes. As a result, our study may not fully reflect the nuanced relationship between short-term temperature variation and adverse health outcomes. Furthermore, as an observational study, our analysis is inherently limited in its ability to rule out unmeasured confounding factors, which may influence the observed associations. Although we have applied rigorous statistical methods and adjusted for known confounders, the possibility of residual confounding remains. Future research, including randomized controlled trials or studies with more comprehensive confounder adjustment, could provide stronger causal evidence.

Conclusions

In summary, in a cohort of HHF patients from OneFlorida+ Network, we identified that elevated maximum temperatures in May were consistently associated with an increased risk of 1-year HF readmission and mortality. This association remained significant across various subgroups, including race, sex, age, SES, and urban-rural residence, suggesting that high temperatures in late spring pose a uniform risk to patients with HF, regardless of demographic or socioeconomic background. Our study has important implications for the impact of global warming and environmental factors on HF outcomes. Given the projected increases in global temperatures, further investigations are warranted to uncover the intricate mechanisms through which extreme heat potentially influences adverse HF outcomes and to develop targeted strategies to mitigate these risks in affected populations. Clinicians should be aware of temperature-related risk, particularly in regions with pronounced seasonal temperature variations, and consider environmental conditions when planning HF management and follow-up care.

XSL•FO
RenderX

Acknowledgments

This research was supported in part by the American Foundation for Pharmaceutical Education (AFPE) through the Pre-Doctoral Fellowship in Health Outcomes Disparities. The authors thank AFPE for its support in advancing this work.

Data Availability

The datasets generated or analyzed during this study are not publicly available, due to the OneFlorida+ electronic health record data containing protected health information, but are available from the corresponding author on reasonable request. Access is granted under a data use agreement and requires institutional review board approval.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary tables and figures.
[DOCX File, 212 KB - formative_v9i1e71595_app1.docx]

Checklist 1
STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) checklist.
[PNG File, 13 KB - formative_v9i1e71595_app2.png]

References

1. Heidenreich PA, Albert NM, Allen LA, et al. Forecasting the impact of heart failure in the United States: a policy statement from the American Heart Association. Circ Heart Fail 2013 May;6(3):606-619. [doi: 10.1161/HHF.0b013e318291329a] [Medline: 23616602]
2. What is heart failure? National Heart, Lung, and Blood Institute (NHLBI). 2022. URL: https://www.nhlbi.nih.gov/health/heart-failure [accessed 2024-02-12]
3. Rasmussen M, Prado A, Hominal MA, et al. Global variations in heart failure etiology, management, and outcomes. JAMA 2023 May 16;329(19):1650. [doi: 10.1001/jama.2023.5942]
4. Al-Omary MS, Khan AA, Davies AJ, et al. Outcomes following heart failure hospitalization in a regional Australian setting between 2005 and 2014. ESC Heart Fail 2018 Apr;5(2):271-278. [doi: 10.1002/ehf2.12239] [Medline: 29265710]
5. Al-Omary MS, Davies AJ, Evans TJ, et al. Mortality and readmission following hospitalisation for heart failure in Australia: a systematic review and meta-analysis. Heart Lung Circ 2018 Aug;27(8):917-927. [doi: 10.1016/j.hlc.2018.01.009] [Medline: 29519691]
6. Lan T, Liao YH, Zhang J, et al. Mortality and readmission rates after heart failure: a systematic review and meta-analysis. TCRM 2021 Dec;17:1307-1320. [doi: 10.2147/TCRM.S340587]
7. Parizo JT, Kohsaka S, Sandhu AT, Patel J, Heidenreich PA. Trends in readmission and mortality rates following heart failure hospitalization in the veterans affairs health care system from 2007 to 2017. JAMA Cardiol 2020 Sep 1;5(9):1042-1047. [doi: 10.1001/jamacardio.2020.2028] [Medline: 32936253]
8. Casper M, Nwaise I, Croft JB, Hong Y, Fang J, Greer S. Geographic disparities in heart failure hospitalization rates among medicare beneficiaries. J Am Coll Cardiol 2010 Jan 26;55(4):294-299. [doi: 10.1016/j.jacc.2009.10.021] [Medline: 20117432]
9. Su A, Al'Aref SJ, Beecy AN, Min JK, Karas MG. Clinical and socioeconomic predictors of heart failure readmissions: a review of contemporary literature. Mayo Clin Proc 2019 Jul;94(7):1304-1320. [doi: 10.1016/j.mayocp.2019.01.017] [Medline: 31272573]
10. Foraker RE, Rose KM, Suchindran CM, Chang PP, McNeill AM, Rosamond WD. Socioeconomic status, medicaid coverage, clinical comorbidity, and rehospitalization or death after an incident heart failure hospitalization. Circ: Heart Failure 2011 May;4(3):308-316. [doi: 10.1161/CIRCHEARTFAILURE.110.959031]
11. Philbin EF, Dec GW, Jenkins PL, DiSalvo TG. Socioeconomic status as an independent risk factor for hospital readmission for heart failure. Am J Cardiol 2001 Jun 15;87(12):1367-1371. [doi: 10.1016/s0002-9149(01)01554-5] [Medline: 11397355]
12. Greenberg BH, Abraham WT, Albert NM, et al. Influence of diabetes on characteristics and outcomes in patients hospitalized with heart failure: a report from the Organized Program to Initiate Lifesaving Treatment in Hospitalized Patients with Heart Failure (OPTIMIZE-HF). Am Heart J 2007 Aug;154(2):277. [doi: 10.1016/j.ahj.2007.05.001] [Medline: 17643576]
13. Sarma S, Mentz RJ, Kwasny MJ, et al. Association between diabetes mellitus and post‐discharge outcomes in patients hospitalized with heart failure: findings from the EVEREST trial. European J of Heart Fail 2013 Feb;15(2):194-202. [doi: 10.1093/eurjhf/hfs153]

XSL•FO
RenderX

14. Zinman B, Wanner C, Lachin JM, et al. Empagliflozin, cardiovascular outcomes, and mortality in type 2 diabetes. N Engl J Med 2015 Nov 26;373(22):2117-2128. [doi: 10.1056/NEJMoa1504720] [Medline: 26378978]
15. Inzucchi SE, Zinman B, Fitchett D, et al. How does empagliflozin reduce cardiovascular mortality? Insights from a mediation analysis of the EMPA-REG OUTCOME trial. Diabetes Care 2018 Feb;41(2):356-363. [doi: 10.2337/dc17-1096] [Medline: 29203583]
16. Wild CP. The exposome: from concept to utility. Int J Epidemiol 2012 Feb;41(1):24-32. [doi: 10.1093/ije/dyr236] [Medline: 22296988]
17. Li Y, Hu H, Zheng Y, et al. Impact of contextual-level social determinants of health on newer antidiabetic drug adoption in patients with type 2 diabetes. IJERPH 2023;20(5):4036. [doi: 10.3390/ijerph20054036] [Medline: 33896999]
18. Hu H, Zheng Y, Wen X, et al. An external exposome-wide association study of COVID-19 mortality in the United States. Sci Total Environ 2021 May;768(144832):144832. [doi: 10.1016/j.scitotenv.2020.144832]
19. Niedzwiecki MM, Walker DI, Vermeulen R, Chadeau-Hyam M, Jones DP, Miller GW. The exposome: molecules to populations. Annu Rev Pharmacol Toxicol 2019 Jan 6;59:107-127. [doi: 10.1146/annurev-pharmtox-010818-021315] [Medline: 30095351]
20. Hu H, Zhao J, Savitz DA, Prosperi M, Zheng Y, Pearson TA. An external exposome-wide association study of hypertensive disorders of pregnancy. Environ Int 2020 Aug;141:105797. [doi: 10.1016/j.envint.2020.105797] [Medline: 32413622]
21. Data. OneFlorida+ Clinical Research Network. URL: https://onefloridaconsortium.org/data/ [accessed 2023-10-04]
22. Hogan WR, Shenkman EA, Robinson T, et al. The OneFlorida Data Trust: a centralized, translational research data infrastructure of statewide scope. J Am Med Inform Assoc 2022 Mar 15;29(4):686-693. [doi: 10.1093/jamia/ocab221] [Medline: 34664656]
23. Integrated data repository research services, clinical and translational science institute. University of Florida Health. URL: https://idr.ufhealth.org/ [accessed 2024-02-28]
24. Thygesen SK, Christiansen CF, Christensen S, Lash TL, Sørensen HT. The predictive value of ICD-10 diagnostic coding used to assess Charlson comorbidity index conditions in the population-based Danish National Registry of Patients. BMC Med Res Methodol 2011 May 28;11(1):83. [doi: 10.1186/1471-2288-11-83] [Medline: 21619668]
25. Social determinants of health database. Agency for Healthcare Research and Quality. URL: https://www.ahrq.gov/sdoh/data-analytics/sdoh-data.html [accessed 2024-10-04]
26. Building power for health and equity. County Health Rankings & Roadmaps. URL: www.countyhealthrankings.org [accessed 2025-07-18]
27. Tables and maps. U.S. Bureau of Labor Statistics. URL: https://www.bls.gov/lau/tables.htm [accessed 2025-07-18]
28. Economic resilience of US counties during the great recession. The Northeast Regional Center for Rural Development. URL: https://aese.psu.edu/nercrd/community/economicresilience [accessed 2023-10-04]
29. Download data tables. United States Cancer Statistics, CDC. 2023. URL: https://www.cdc.gov/cancer/uscs/dataviz/download_data.htm [accessed 2023-10-04]
30. Dartmouth Atlas DATA. URL: https://data.dartmouthatlas.org/ [accessed 2023-10-04]
31. Air Quality Statistics Report. United States Environmental Protection Agency. URL: https://www.epa.gov/outdoor-air-quality-data/air-quality-statistics-report [accessed 2023-10-04]
32. US Religion Census. URL: https://www.usreligioncensus.org/ [accessed 2023-10-04]
33. Pesticide National Synthesis Project. US Geological Survey (USGS). URL: https://water.usgs.gov/nawqa/pnsp/data/ [accessed 2023-10-04]
34. County-level measure of social capital. The Northeast Regional Center for Rural Development. URL: https://aese.psu.edu/nercrd/community/social-capital-resources [accessed 2023-10-04]
35. Social Determinants of Health. Office of Disease Prevention and Health Promotion. URL: https://health.gov/healthypeople/priority-areas/social-determinants-health [accessed 2023-10-18]
36. Ponce SG, Norris J, Dodendorf D, Martinez M, Cox B, Laskey W. Impact of ethnicity, sex, and socio-economic status on the risk for heart failure readmission: the importance of context. Ethn Dis 2018;28(2):99-104. [doi: 10.18865/ed.28.2.99] [Medline: 29725194]
37. Lick D, Mulhem E. Mortality and hospital readmission rates for heart failure among patients of middle eastern ancestry compared to non-Middle Eastern Whites in Southeast Michigan. J Immigrant Minority Health 2023 Aug;25(4):835-840. [doi: 10.1007/s10903-023-01473-6]
38. Roshanghalb A, Mazzali C, Lettieri E. Multi-level models for heart failure patients' 30-day mortality and readmission rates: the relation between patient and hospital factors in administrative data. BMC Health Serv Res 2019 Dec 30;19(1):1012. [doi: 10.1186/s12913-019-4818-2] [Medline: 31888610]
39. Curtis LH, Greiner MA, Hammill BG, et al. Early and long-term outcomes of heart failure in elderly persons, 2001-2005. Arch Intern Med 2008 Dec 8;168(22):2481-2488. [doi: 10.1001/archinte.168.22.2481] [Medline: 19064833]
40. Bristow MR, Saxon LA, Boehmer J, et al. Cardiac-resynchronization therapy with or without an implantable defibrillator in advanced chronic heart failure. N Engl J Med 2004 May 20;350(21):2140-2150. [doi: 10.1056/NEJMoa032423] [Medline: 15152059]

XSL•FO
RenderX

41. Cardiovascular. Abbott. URL: https://www.cardiovascular.abbott/content/dam/bss/divisionalsites/cv/cv-live-site/hcp/reimbursement/crm/Pacemaker%20CRT-P_Procedures_v2.pdf [accessed 2024-01-02]
42. Cohen SS, Roger VL, Weston SA, et al. Evaluation of claims - based computable phenotypes to identify heart failure patients with preserved ejection fraction. Pharmacology Res & Perspec 2020 Dec;8(6):e00676. [doi: 10.1002/prp2.676]
43. Chronic conditions data warehouse. Chronic Conditions Data Warehouse. URL: https://www2.ccwdata.org/web/guest/data-dictionaries [accessed 2024-01-05]
44. Neighborhood Atlas. URL: https://www.neighborhoodatlas.medicine.wisc.edu/ [accessed 2021-05-25]
45. Peterson RA. Finding optimal normalizing transformations via bestNormalize. R J 2021;13(1):310. [doi: 10.32614/RJ-2021-041]
46. Tibshirani R. Regression shrinkage and selection via the Lasso. J R Stat Soc Series B Stat Methodol 1996 Jan 1;58(1):267-288. [doi: 10.1111/j.2517-6161.1996.tb02080.x]
47. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate escent. J Stat Softw 2010;33(1):1-22. [doi: 10.18637/jss.v033.i01] [Medline: 20808728]
48. Segar MW, Hall JL, Jhund PS, et al. Machine learning-based models incorporating social determinants of health vs traditional models for predicting in-hospital mortality in patients with heart failure. JAMA Cardiol 2022 Aug 1;7(8):844-854. [doi: 10.1001/jamacardio.2022.1900] [Medline: 35793094]
49. Little RJA, Rubin DB. Statistical Analysis with Missing Data, 1st edition: Wiley; 2002. [doi: 10.1002/9781119013563]
50. Pedersen AB, Mikkelsen EM, Cronin-Fenton D, et al. Missing data and multiple imputation in clinical epidemiological research. Clin Epidemiol 2017;9(157–166):157-166. [doi: 10.2147/CLEP.S129785] [Medline: 28352203]
51. Enders CK. In: Littlee TD, editor. Applied Missing Data Analysis: Guilford Press; 2010.
52. Graham JW. Missing data analysis: making it work in the real world. Annu Rev Psychol 2009;60(549–576):549-576. [doi: 10.1146/annurev.psych.58.110405.085530] [Medline: 18652544]
53. Sun M, Engelhard MM, Bedoya AD, Goldstein BA. Incorporating informatively collected laboratory data from EHR in clinical prediction models. BMC Med Inform Decis Mak 2024 Jul 24;24(1):206. [doi: 10.1186/s12911-024-02612-1] [Medline: 39049049]
54. Agier L, Portengen L, Chadeau-Hyam M, et al. A systematic comparison of linear regression–based statistical methods to assess exposome-health associations. Environ Health Perspect 2016 Dec;124(12):1848-1856. [doi: 10.1289/EHP172] [Medline: 27219331]
55. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Stat Methodol 1995 Jan 1;57(1):289-300. [doi: 10.1111/j.2517-6161.1995.tb02031.x]
56. Vatcheva KP, Lee M, McCormick JB, Rahbar MH. Multicollinearity in regression analyses conducted in epidemiologic studies. Epidemiology (Sunnyvale) 2016 Apr;6(2):227. [doi: 10.4172/2161-1165.1000227] [Medline: 27274911]
57. Motairek I, Makhlouf MHE, Rajagopalan S, Al-Kindi S. The exposome and cardiovascular health. Can J Cardiol 2023 Sep;39(9):1191-1203. [doi: 10.1016/j.cjca.2023.05.020] [Medline: 37290538]
58. Lopez PD, Cativo-Calderon EH, Otero D, Rashid M, Atlas S, Rosendorff C. The impact of environmental factors on the mortality of patients with chronic heart failure. Am J Cardiol 2021 May 1;146(48–55):48-55. [doi: 10.1016/j.amjcard.2021.01.019] [Medline: 33577810]
59. Moghadamnia MT, Ardalan A, Mesdaghinia A, Keshtkar A, Naddafi K, Yekaninejad MS. Ambient temperature and cardiovascular mortality: a systematic review and meta-analysis. PeerJ 2017 Aug 4;5:e3574. [doi: 10.7717/peerj.3574]
60. Chan EYY, Goggins WB, Kim JJ, Griffiths SM. A study of intracity variation of temperature-related mortality and socioeconomic status among the Chinese population in Hong Kong. J Epidemiol Community Health 2012 Apr;66(4):322-327. [doi: 10.1136/jech.2008.085167] [Medline: 20974839]
61. Liu C, Yavar Z, Sun Q. Cardiovascular response to thermoregulatory challenges. Am J Physiol Heart Circ Physiol 2015 Dec 1;309(11):H1793-H1812. [doi: 10.1152/ajpheart.00199.2015] [Medline: 26432837]
62. Liu J, Varghese BM, Hansen A, et al. Heat exposure and cardiovascular health outcomes: a systematic review and meta-analysis. Lancet Planet Health 2022 Jun;6(6):e484-e495. [doi: 10.1016/S2542-5196(22)00117-6] [Medline: 35709806]
63. Escolar V, Lozano A, Larburu N, et al. Impact of environmental factors on heart failure decompensations. ESC Heart Fail 2019 Dec;6(6):1226-1232. [doi: 10.1002/ehf2.12506] [Medline: 31483570]
64. Gostimirovic M, Novakovic R, Rajkovic J, et al. The influence of climate change on human cardiovascular function. Arch Environ Occup Health 2020 Oct 2;75(7):406-414. [doi: 10.1080/19338244.2020.1742079]
65. Sverdlov AL, Figtree GA, Horowitz JD, Ngo DTM. Interplay between oxidative stress and inflammation in cardiometabolic syndrome. Mediators Inflamm 2016;2016:8254590. [doi: 10.1155/2016/8254590] [Medline: 27563174]
66. Crandall CG, Wilson TE. Human cardiovascular responses to passive heat stress. Compr Physiol 2015 Jan;5(1):17-43. [doi: 10.1002/cphy.c140015] [Medline: 25589263]
67. Nam YH, Bilker WB, Leonard CE, Bell ML, Hennessy S. Outdoor temperature and survival benefit of empiric potassium in users of furosemide in US Medicaid enrollees: a cohort study. BMJ Open 2019 Feb 18;9(2):e023809. [doi: 10.1136/bmjopen-2018-023809] [Medline: 30777859]

XSL•FO
RenderX

68. Danesh Yazdi M, Wei Y, Di Q, et al. The effect of long-term exposure to air pollution and seasonal temperature on hospital admissions with cardiovascular and respiratory disease in the United States: a difference-in-differences analysis. Sci Total Environ 2022 Oct 15;843(156855):156855. [doi: 10.1016/j.scitotenv.2022.156855] [Medline: 35750164]

Abbreviations

ADI: Area Deprivation Index
aOR: adjusted odds ratio
COPD: chronic obstructive pulmonary disease
CRT: cardiac resynchronization therapy
CVD: cardiovascular disease
EHR: electronic health record
ExWAS: external exposome-wide association study
HF: heart failure
HFpEF: heart failure with preserved ejection fraction
HHF: hospitalized for heart failure
ICD: implantable cardioverter-defibrillator
***ICD-10*:** *International Classification of Diseases, Tenth Revision*
***ICD-10-CM*:** *International Classification of Diseases: Tenth Revision: Clinical Modification*
***ICD-9-CM*:** *International Classification of Diseases: Ninth Revision: Clinical Modification*
LASSO: least absolute shrinkage and selection operator
SES: socioeconomic status
SNAP: Supplemental Nutrition Assistance Program
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by A Mavragani; submitted 22.01.25; peer-reviewed by D Miric; revised version received 24.05.25; accepted 25.05.25; published 25.08.25.

Please cite as:
Huang W, Kimmel SE, Ahmed M, Smith SM, Lee YA, Yang L, Hernandez I, Bian J, Guo J
External Exposome Factors and Adverse Heart Failure Outcomes in the OneFlorida+ Network: Retrospective Cohort Study
JMIR Form Res 2025;9:e71595
URL: https://formative.jmir.org/2025/1/e71595
doi:10.2196/71595

© Wenxi Huang, Stephen E Kimmel, Mustafa Ahmed, Steven M Smith, Yao An Lee, Lanting Yang, Inmaculada Hernandez, Jiang Bian, Jingchuan Guo. Originally published in JMIR Formative Research (https://formative.jmir.org), 25.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Development of a Theory-Based mHealth App for Fatigue Management in Lupus: Human-Centered Design Approach

Anna Deck[1], BS; Kiran Singh[2], MBBS; Paula Caras[1], BS; Amy LeClair[3], PhD; Monique Gore-Massy[4], BA; Faye Chiu[5], BA; Lucas Ogura Dantas[6], PhD; Timothy McAlindon[7], MD, MPH; Sara Folta[8], PhD; Shanthini Kasturi[2], MD, MS

[1]School of Medicine, Tufts University, Boston, MA, United States
[2]Division of Rheumatology, Department of Medicine, Tufts Medical Center, 800 Washington Street, Boston, MA, United States
[3]Department of Medicine, Tufts Medical Center, Boston, MA, United States
[4]Patient Partner, West Orange, NJ, United States
[5]Patient Partner, New York, NY, United States
[6]Ambulomics, Sao Paolo, Brazil
[7]Division of Rheumatology, University of Massachusetts Chan Medical School, Worcester, MA, United States
[8]Friedman School of Nutrition, Tufts University, Boston, MA, United States

Corresponding Author:
Shanthini Kasturi, MD, MS
Division of Rheumatology, Department of Medicine, Tufts Medical Center, 800 Washington Street, Boston, MA, United States

Abstract

Background: Fatigue is a highly prevalent and debilitating symptom of systemic lupus erythematosus (SLE), significantly affecting the quality of life and employment of those living with the disease. Nonpharmacologic approaches, such as physical activity interventions, have shown promise in reducing fatigue but are often resource-intensive and lack grounding in behavior change theory. Mobile health (mHealth) technology offers a scalable approach to delivering behavioral interventions.

Objective: This study describes the development of an mHealth app, grounded in behavior change theory, to support fatigue self-management in individuals with SLE by promoting physical activity.

Methods: We used a human-centered design (HCD) approach to develop an mHealth app grounded in the self-determination theory of motivation and the social cognitive theory of behavior change. The process included two phases: (1) inspiration and (2) ideation. In the inspiration phase, key user needs were identified from focus groups of adults with SLE. During the ideation phase, a prototype was developed and iteratively refined based on feedback from additional individuals with SLE who participated in multiple rounds of semistructured interviews and online feedback surveys.

Results: In the inspiration phase, 12 individuals participated in 2 focus groups and identified key priorities for the mHealth intervention, including symptom tracking (fatigue, pain, sleep, and physical activity), reliable educational content, social connection, and reminders. In the ideation phase, a prototype was developed based on these findings and refined through 2 rounds of user feedback interviews with 12 additional adults with SLE. Participants rated the features and format of the prototype favorably, with average scores ranging from 1.4 to 2.1 on a 5-point Likert scale (1=highest rating), and 78% (7/9) of interviewees reported they were likely or highly likely to use the app. Several themes around preferences for the app emerged from the interviews, including the importance of: (1) community and social connection, (2) accessibility and inclusion, (3) options for customization, (4) integration of the app with existing digital health tools, and (4) notifications for reminders and motivational messages. Based on this feedback, the prototype was refined, and a digital messaging feature was created. A library of 154 reminders and motivational messages was developed with input from 10 individuals with SLE who took part in a third round of interviews.

Conclusions: Using an HCD approach, we developed an mHealth app tailored to the needs of individuals with SLE, integrating behavior change theory to support fatigue self-management. Through engagement with end users, we iteratively refined the app to address key priorities and enhance usability. This study demonstrates the feasibility of using HCD to develop an mHealth app grounded in behavior change theory and provides a model for creating rigorous digital health interventions for individuals with SLE and other chronic conditions.

(JMIR Form Res 2025;9:e75399) doi:10.2196/75399

KEYWORDS

systemic lupus erythematosus; fatigue; quality of life; mHealth; digital health; behavioral intervention; behavior change theory; human-centered design

XSL•FO
RenderX

Introduction

Fatigue is the most common symptom of systemic lupus erythematosus (SLE) and affects 80% - 90% of individuals with this chronic autoimmune inflammatory disease [1]. Fatigue significantly impacts morbidity and is a leading cause of withdrawal from the workforce, with unemployment rates ranging from 40% to 60% among individuals with SLE [1-8]. People living with SLE consistently cite fatigue as their greatest unmet need; however, the relationship between SLE disease activity and fatigue is complex [9-15]. Studies suggest SLE-related fatigue is of multifactorial etiology, with the greatest contributors often being behavioral factors, namely poor sleep quality and physical inactivity; psychosocial factors, such as social isolation, and comorbid conditions, for example, depression and fibromyalgia [1,2,13-19].

Despite the significant burden of fatigue in individuals with SLE, there are limited effective treatment options available. Pharmacotherapy modifies fatigue mainly in those with highly active disease [20]. Nonpharmacologic approaches, such as physical activity interventions and education, have shown promise in reducing fatigue in individuals with SLE [21-29]. While a number of such interventions have been evaluated in SLE, these interventions often lack a grounding in and rigorous approach to behavior change theory [30-35]. In addition, such interventions are often resource-intensive or delivered in person, which can limit access and scalability.

Mobile health (mHealth) technology offers a platform to deliver behavioral interventions that lower barriers to participation and engagement. mHealth apps delivered through smartphones can provide tools for self-monitoring behavior, such as physical activity, by enabling real-time data capture and automated personalized feedback provided in individuals' native context [36-38]. Interventions using mHealth-enabled behavior change strategies, such as self-monitoring and feedback via messaging, have been shown to facilitate physical activity, improve diet, and encourage other healthy behaviors across several medical conditions [39-43]. In addition, mHealth apps can overcome the logistical barriers of time and geography and offer resource and cost-effective alternatives to traditional in-person interventions.

Currently available mHealth apps for individuals with SLE are of limited quality and functionality. A 2020 systematic review of mHealth apps and technologies for SLE found that the majority of available apps scored poorly on the Mobile App Rating Scale, a validated tool for assessing the quality of apps based on engagement, functionality, aesthetics, information quality, and overall quality [44]. Existing apps were not developed with patient input and lacked evidence-based information and functionalities desired by patients, such as symptom and medication trackers and discussion groups and community features [44]. Of note, of the 20 apps identified in this review, many are no longer available or lack functional updates to be compatible with current technology. Individuals with SLE have a desire to use mHealth technology to aid with disease management, but there are few studies of mHealth interventions to improve health outcomes in SLE [45-47]. As such, there is a demonstrated need for mHealth tools for people with SLE that provide high-quality educational information, disease monitoring and management support, and social connection.

To address these gaps, we aimed to develop an mHealth-enabled behavioral intervention to reduce fatigue in individuals living with SLE by promoting physical activity. Rather than adapting an existing app, we sought to create a novel intervention from the ground up to address the unique behavioral and psychosocial contributors to fatigue in SLE—factors not targeted by mHealth tools designed for general populations. Recognizing the limitations of existing mHealth apps, our approach integrates behavior change theory and follows a human-centered design (HCD) process to ensure usability and alignment with patient needs. By leveraging mHealth technology, this intervention seeks to overcome traditional barriers to access, providing a scalable solution for fatigue management in SLE. Here, we outline the development process of this app.

Methods

Study Design

Theoretical Framework

The mHealth intervention is grounded in the self-determination theory (SDT) of motivation and social cognitive theory (SCT; Figure 1). SDT posits that successful behavior change hinges on addressing the key needs of relatedness, autonomy, and competence, while SCT emphasizes that behavior change relies on both environmental factors, such as social support, and cognitive factors, particularly self-efficacy [48-50]. Both SDT and SCT have been applied to a variety of health behaviors and provide a robust theoretical framework for promoting adherence to physical activity, the central behavior to be targeted in the mHealth intervention [51-53].

XSL•FO
RenderX

Figure 1. Theoretical framework for the intervention: self-determination and social cognitive theories. mHealth: mobile health.

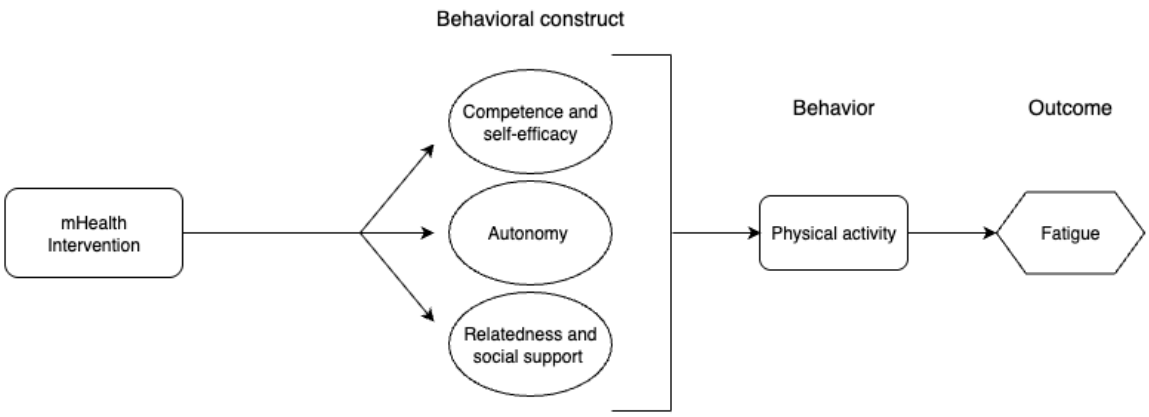

HCD Approach

A HCD approach was used to identify key stakeholder needs and develop the mHealth app for fatigue self-management. HCD places end users, in this case individuals living with SLE, at the center of the design process to create products that address key user needs and emphasize the whole user experience. This contrasts with traditional design methods, which typically develop products based primarily on expert opinion [54]. HCD consists of 3 stages: inspiration, ideation, and implementation. In the inspiration phase, key user needs are identified through immersive research, such as interviews and observation, and engagement with users. In the ideation phase, a prototype is developed and iteratively refined based on user feedback. During implementation, the final stage of HCD, the developed product is launched, integrated, and scaled in real-world settings [55,56]. This paper presents the findings from the first 2 stages, while the third phase will be reported separately following the completion of a pilot clinical trial of the intervention.

In addition to gathering feedback from end users through HCD methods, 2 patient partners (FC and MGM) were fully integrated into the research and development team. They actively contributed to every stage of development of the mHealth intervention, from conceptualization to implementation. These members of the research team participated in regular team meetings, ensuring that patient voices were heard throughout the process. They provided input on desired features, functionality, and format for the intervention, assisted with the development of feedback guides, and provided context to interpret the feedback provided by other end users. In cases of conflicting feedback, the patient partners shared their perspectives to support the decision-making process.

HCD Phase 1: Inspiration

The inspiration phase of HCD focuses on understanding the needs, challenges, and preferences of the end user with the goal of developing a design solution that is human-centered. Our design challenge, to create an mHealth app to help people living with SLE manage their fatigue, emerged from formative and immersive work, including literature review, targeted conversations, and focus groups with patients with SLE.

Study Population

Participants were English-speaking adults (≥18 y of age) who had a diagnosis of SLE based on 2019 European League Against Rheumatism/American College of Rheumatology classification criteria and owned a smartphone [57]. Participants were recruited from the rheumatology clinic at Tufts Medical Center, an academic medical center in Boston, Massachusetts. Electronic health records were queried for patients with *ICD-9* (*International Classification of Diseases, Ninth Revision*) and/or *ICD-10* (*International Statistical Classification of Diseases, Tenth Revision*) diagnosis codes of SLE. Patients meeting inclusion criteria were contacted by telephone or in person at the time of a clinical visit and invited to participate in the study. Interested individuals were purposively sampled based on demographics (age, race, ethnicity, and insurance type) and disease duration to capture a diverse range of perspectives.

Data Collection

Two focus groups were conducted to elicit patient needs and preferences for the mHealth intervention for fatigue in SLE. A total of 12 individuals participated in these focus groups. Conversations were guided by a focus group guide (Multimedia Appendix 1), which probed desired content and features for the mHealth intervention, such as symptom and activity trackers, essential educational content, and practical strategies. Focus groups were led by an experienced moderator (AL) and were conducted in person. All focus groups were audio-recorded, transcribed, and deidentified.

Data Analysis

Participant demographics were analyzed using descriptive statistics. Focus group transcripts were independently reviewed by 2 investigators (AD and SK) using an integrated approach in which inductive codes emerging from line-by-line reading of the data were used alongside a priori codes generated from the theoretical frameworks of SDT and SCT. Codes were then

XSL•FO
RenderX

reviewed together by both investigators and revised via a comparison and consensus approach, and themes were identified.

HCD Phase 2: Ideation

The ideation phase of HCD aims to generate diverse ideas and potential solutions to the design challenge with the input of a multidisciplinary team of end users, designers/developers, and other stakeholders. Our development team consisted of individuals living with SLE, an app developer, a behavioral interventionalist, medical students, and a practicing rheumatologist. During this phase, app prototypes were generated and educational content for the intervention was developed. Here we discuss the development of the app prototype; development of the educational content is reported elsewhere [58].

Prototype Development

Potential features for the mHealth app identified during phase 1 were mapped to relevant behavioral constructs from SDT and SCT. A prototype of the app with desired features was developed using Figma (Figma Inc), a web-based collaborative design tool for creating and testing user interfaces and experiences for mobile apps. The prototype was reviewed and revised with input from the development team. The prototype was then iteratively refined through several rounds of feedback interviews with individuals with SLE.

Study Population

As in phase 1, participants in phase 2 were English-speaking adults diagnosed with SLE who owned smartphones. Participants were recruited once again through purposive sampling from the rheumatology clinic at Tufts Medical Center and additionally through the networks of our patient partners. There was no overlap of participants between phases 1 and 2.

Data Collection

A total of 12 individuals with SLE participated in 3 rounds of semistructured interviews to provide feedback on app prototypes. Conversations were led by an experienced facilitator (SK), using an interview guide (Multimedia Appendix 2) and modeled after think-aloud sessions [59,60]. Screenshots of the mHealth app from Figma were shared using Zoom (Zoom Video Communications) screen sharing capabilities. Each page of the app was displayed, and the functionality was explained. Interview participants were asked to share their impressions, including features they liked, aspects they would change, and additional ideas for improvement. The prototype was then modified based on feedback from the interviews, and a second round of interviews was conducted to evaluate the updated prototype. Interview participants also completed surveys rating the format and features of the prototypes (layout, color scheme, font, and graphics) and their likelihood to use the app using a numeric rating scale (1-5). A third round of interviews was conducted to evaluate app-based motivational and educational messaging that was developed in response to feedback from the first 2 cycles of interviews. Participants reviewed and rated the usefulness of the messages as part of the fatigue self-management intervention and provided open-ended feedback. Feedback was used to modify the messaging library and generate several new messages to address topics participants felt were missing. All surveys were administered using REDCap (Research Electronic Data Capture; Vanderbilt University). Interviews were conducted via Zoom web-based teleconference and were audio-recorded, transcribed, and deidentified.

Data Analysis

Participant demographics and survey responses were analyzed using descriptive statistics. Interview transcripts were independently reviewed by 2 investigators (AD and SK) who identified dominant themes guided by the theoretical frameworks of SDT and SCT. Themes were then reviewed and revised via a comparison and consensus approach. Transcripts of prototype feedback interviews were analyzed using a rapid analysis protocol and feedback capture grid [61,62].

Ethical Considerations

The study protocol was reviewed and approved by the Tufts Health Sciences institutional review board (IRB number 13159 and number 00002106). Informed consent was obtained from all individuals participating in interviews or focus group discussions. Participants were compensated US $30 for each focus group or interview session.

Results

HCD Phase 1: Inspiration

Two focus groups including a total of 12 individuals with SLE (Table 1) were conducted.

Focus groups were about 30 minutes in duration and identified 4 key needs for the mHealth intervention (Table 2): social connection, educational content, tracking features, and reminders.

Participants expressed a desire for social connection and suggested ideas, including peer support features, chat functions, and connecting with health coaches. Participants emphasized the feeling of validation that came from participating in focus groups and felt that a peer support component in the app could offer similar validation. Participants also identified a need for reliable, evidence-based educational content within the app. Participants wanted general information about their SLE diagnosis and lupus-related fatigue, along with strategies for managing fatigue and improving sleep. They suggested content on physical activity, nutrition, and mental health and wellness. Finally, participants desired tracking features (such as pain, fatigue, activity, and sleep trackers) and reminders/notifications (such as prompts to be physically active) to help with SLE and fatigue self-management.

XSL•FO
RenderX

Table . Characteristics of focus group and semistructured interview participants (N=24).

Characteristic	Focus group participants (n=12)	Interview participants (n=12)	All participants (n=24)
Age (years), mean (SD)	44.3 (13.3)	42.4 (8.2)	43.4 (10.8)
SLE disease duration (years), mean (SD)	10.1 (8.8)	14.1 (7.2)	12.1 (8.1)
Sex, n (%)			
Female	12 (100)	10 (83)	22 (92)
Male	0 (0)	2 (17)	2 (8)
Race, n (%)			
White	3 (25)	4 (33)	7 (29)
Black or African-American	5 (42)	4 (33)	9 (38)
Asian	2 (17)	2 (17)	4 (17)
More than one race	2 (17)	2 (17)	4 (17)
Ethnicity, n (%)			
Hispanic	2 (17)	4 (33)	6 (25)
Non-Hispanic	10 (83)	8 (67)	18 (75)
Insurance type, n (%)			
Medicaid	6 (50)	1 (8)	7 (29)
Medicare	1 (8)	2 (17)	3 (13)
Third party/private	5 (42)	8 (67)	13 (54)
None	0 (0)	1 (8)	1 (4)
Education, n (%)			
High school graduate	2 (17)	2 (17)	4 (17)
Associate degree	3 (25)	1 (8)	4 (17)
Bachelor’s degree	4 (33)	6 (50)	10 (42)
Graduate degree	3 (25)	3 (25)	6 (25)
Employment, n (%)			
Part time	0 (0)	2 (17)	2 (8)
Full time	7 (58)	5 (42)	12 (50)
Self-Employed	1 (8)	2 (17)	3 (13)
On disability	2 (17)	3 (25)	5 (21)
Retired	1 (8)	0 (0)	1 (4)
Looking for work	1 (8)	0 (0)	1 (4)

XSL•FO
RenderX

Table . Inspiration phase: themes from focus groups.

User identified needs	Desired app features
Social connection	• Peer support • Chat with other app users • Access to health coach

Educational content	• Information about:
	  • Lupus
	  • Fatigue and Lupus
	  • Tips to manage fatigue
	  • Sleep
	  • Nutrition
	  • Mental health
	• Reliable (with references)
	• Accessible (appealing format and appropriate reading level)

Trackers	• Pain • Fatigue • Sleep (quality and quantity) • Physical activity
Reminders	• Prompts to exercise

HCD Phase 2: Ideation

Prototype Development

Drawing from the literature review and findings from qualitative work in phase 1, the following features were identified as essential components of the mHealth app: (1) physical activity goal-setting logs and trackers, (2) educational modules with self-assessment questions, (3) sleep trackers, (4) symptom trackers (fatigue, pain, and anxiety), (5) peer coach support, and (6) notifications. An initial prototype was developed with these functionalities. On review by the development team, the patient partners offered insight into visual accessibility (font size and color) and suggested several improvements. They recommended adding a daily "check-in" screen where users could rate how they were feeling, using weather pictograms, and their energy level, using spoons, a common analogy for describing energy in chronic disease [63]. They suggested adding free text journal features to this screen and the physical activity logs. Finally, they proposed reformatting the educational modules from text documents to infographics. These changes were incorporated into a second iteration of the prototype. Once the core components of the mHealth app were finalized, each feature was mapped to the relevant behavioral construct from SDT or SCT (Table 3).

XSL•FO
RenderX

Table . mHealth intervention behavioral construct mapping.

Behavioral construct from self-determination and social cognitive theories	mHealth[a] intervention component
Competence/Self-Efficacy	
Knowledge	• Educational modules (information on causes of fatigue, treatments for fatigue, benefits of physical activity in people living with lupus)
Behavioral skills/Self-regulation	• Educational modules (information on how to be physically active, pacing/ "listen to your body") • Weekly physical activity goal setting • Physical activity logs and progress visualization (self-monitoring) • Tracker data (self-monitoring) • Energy and movement journals • Messages and notifications (feedback) • Achievement page (feedback)
Outcome expectations	• Educational modules ("what to expect when you start moving")
Autonomy	• Weekly physical activity goals (choice of type and amount of activity) • Messages and notifications (autonomy supportive language) • Educational modules ("find your why"; autonomy supportive language)
Social support/relatedness	• Peer coach • Group meetings via Zoom • Participant chat group • Program buddy
Barriers and opportunities	• Messages and notifications (reminders)
Observational learning/normative beliefs	• Educational modules (participant testimonials)

[a]mHealth: mobile health.

Prototype Iteration

A total of 12 individuals with SLE reviewed the prototype in 3 rounds of feedback interviews and surveys (Table 1). Participants gave high ratings to prototype features and format, and 78% (7/9) of participants reported being likely or highly likely to use the app (Table 4).

Interviews averaged 20 - 25 minutes in length. The first round of interviews was analyzed using a rapid analysis protocol and feedback capture grids to categorize feedback in four major areas: (1) like (positive feedback), (2) impact (potential impact of the app), (3) change (suggestions for improvement), and (4) ideas (ideas for novel app features; Table 5).

Table . Prototype ratings Individuals rated prototype features using a 5-point Likert scale (1=Loveit, 5=Hateit).

Features	Prototype 1 (n=9)	Prototype 2 (n=7)
Features, mean (SD)		
Education	1.6 (0.7)	Not rated
Self-Assessment quizzes	2.1 (1.2)	Not rated
Goal setting and logs	1.6 (1.0)	Not rated
Trackers (steps/sleep/mood)	1.7 (1.0)	Not rated
Reminders	1.4 (0.7)	Not rated
Community/Social connection	1.4 (0.5)	Not rated
Format, mean (SD)		
Color scheme	2.1 (0.8)	2.1 (1.1)
Layout	2.0 (0.9)	1.1 (0.4)
Text	2.0 (1.0)	1.9 (0.9)
Graphics/Images	1.9 (0.9)	1.9 (0.9)
Likeliness to use app	1.7 (0.9)	1.7 (0.8)

XSL•FO
RenderX

Table . mHealth app prototype feedback.

Feedback category	App feature/characteristic	Theme	Exemplar quotation
Like	• Educational modules • Graphics • Messages/ Notifications • Interface • Trackers	• Reliable educational information • Validation · • Self-management • Usability · • Clinical care; self-management	• "I think that every lupus patient should have educational materials that give them important information and insights into how they can manage their fatigue." • "What I like especially, is the person in a wheelchair. Because when you say activities, usually people think of someone who is already standing up and who is able to do something. But, you have a wheelchair and you have exercise. It gives motivation to people that are in the same situation." • "Reminders are great for lupus fog!!! Also great if there's some kind of encouragement involved in the message." • "It doesn't feel like it'd be a huge learning gap if someone is going from another health tracker to this one. It doesn't feel disorienting if they've used a health tracker before. This feels exactly the same, and to me that's a good thing." • "The ability to utilize, or be able to integrate this information, all this data, to show or to report to the physician. How helpful is this going to be for the patient to manage, or to help the doctor understand how active or not active they've been or how they've been feeling."
Impact	Trackers	• Self-management • Self-management • Clinical care	• "This notion of really checking on your quality of life and your mood is such an important part of it, because I think that's sort of the underserved aspect or component of this journey." • "This is not only going to help me, it is going to push me to do better, it is going to show me [to recognize] what I need to work on more." • "This is something that is going to help on my doctor visits because my doctor asked me questions like how you've been feeling, have you been sleeping normally."
Impact	Journal	• Validation	• "Knowing that I have my [app] best friend can be like 'hello, [name] how are you feeling today.' Sometimes we just need that."

XSL•FO
RenderX

Feedback category	App feature/characteristic	Theme	Exemplar quotation
Change	Trackers	• Usability • Gamification • Usability	• "A lot of it requires patient self-report, patient logging. There's a lot of effort needed on the patient side." • "It's just not worthwhile if there's no real time feedback or incentive for me to continue logging in or putting in my information. It can get tedious." • "I was hoping that it communicated with and was able to pull that data from apple health or fitness."
Change	Graphics	• Validation	• "I want to feel connected with some of the images...I'm the type of person who loves to see real people. I want to feel that connection."
Change	Notifications	• Customizability	• "People want to be reminded when they want to be reminded. Not that someone reminds them at any given time."
Ideas	Trackers	• Self-management • Self-management	• "A food log...trying to help people find that correlation to like 'Oh, I felt bad because I may have eaten a food like a nightshade or something." • "Let's say you want to upload a video of your activity...I can actually take video of the progress like actually show the physical progress in video format."
Ideas	Educational Modules	• Gamification	• "When a module is completed or something you get a badge and there's emojis coming out. Like you reached this level."
Ideas	Messages/Notifications	• Self-management • Self-management • Self-management	• "Maybe there's a reminder piece: 'Hey it's nine o'clock. Take your medicine." • "A pop-up reminder...don't forget to pace yourself." • "Little words of affirmation. Something to encourage you."
Ideas	Community	• Social support • Social support	• "It would be kind of cool to have it all-inclusive to people that have lupus and maybe a parent guardian or friend...where I can invite my mom to be a part of this app." • "If there was some sort of way that...you could friend people, almost like it is its own social network."

Key themes and insights that emerged from the interviews in each of these 4 areas are summarized below.

- Like: Overall, participants liked the educational modules, graphics, messages and notifications, trackers, and app interface. The educational modules were felt to be useful and easy to understand, while the trackers were recognized

XSL•FO
RenderX

as important tools for monitoring symptoms. The graphics were considered motivating, especially for their inclusivity in depicting people in wheelchairs and engaging in diverse activities. The app interface and functionality were perceived to be intuitive.

- Impact: Participants highlighted the trackers and journal features as particularly impactful. They felt that the trackers allowed for self-motivation, reflection, and sharing data with their physicians. The journal feature, which allows individuals to annotate their mood, energy, and activity logs, was felt to be validating and personalized.
- Users identified several key areas for potential change. They felt that the tracking features were burdensome, requiring users to actively log activity and symptoms. They noted this burden could be reduced with integration of the app with preexisting health apps, including Apple Health (Apple). They felt the app could be more engaging and use gamification to reward milestones. While some users found the graphics appealing, as noted in the "like" section, others felt that using real images of people would create a stronger connection. Finally, users desired app customizability, such as the ability to choose which symptoms to track and the option to select if and when to receive notifications.
- Ideas: Participants generated ideas for several novel app functionalities, including dietary trackers or food logs and video logs or journaling. They desired notifications, especially reminders to pace physical activity and take medications, but also messages of encouragement and affirmation. Several cited a need for gamification related to completing educational modules and achieving activity goals. In addition, users wanted expanded social support options, such as "friending" capabilities and live chat within the app, and the ability to invite family members to join the app "community."

This feedback was used to iteratively refine the prototype. Key updates included the addition of a banner to congratulate users upon completion of educational modules and enhanced app functionality, such as integration with Apple Health or Google Fit to automatically track movement, steps, and sleep data. Of note, several suggested features, including food logs, medication reminders, and a social network/community with live chat, were not incorporated into the mHealth intervention given the scope of our design challenge and focus on physical activity.

The refined prototype was shared with participants in a second round of feedback interviews, which focused on evaluating the format and usability of the app. Ratings of the second prototype were generally favorable (Table 4). The following changes were made based on this round of interviews: (1) font size was increased to improve readability and accessibility, (2) the mood question was clarified, and (3) a graphic was added to the activity log to enhance representation. The screenshots of early and final prototypes of the app are illustrated in Figures S1 and S2 in Multimedia Appendix 3.

A recurring theme in feedback interviews was the user desire for app-based notifications, including reminders and educational and motivational messaging around physical activity and fatigue management. To meet this need, we undertook the creation of a digital messaging feature as part of the app. A literature search was performed to find existing messaging libraries that promoted physical activity and/or addressed fatigue management and were grounded in behavior change theory. A published text messaging library created to support a diabetes prevention program was identified [64,65]. These messages were developed to promote healthy diet and physical activity and were mapped to specific behavior change techniques from the Behavior Change Wheel framework, a comprehensive model that synthesizes multiple behavior change theories [66]. This library was selected as the foundation for developing messages for our mHealth intervention because of its focus on promoting physical activity and rigorous incorporation of behavior change theory.

The 124 messages in the published library were reviewed, and messages specific to the diabetes program or content related to diet were removed as these were not relevant for our intervention. The remaining 69 messages were adapted to be more specific to people with SLE. Forty-nine new messages about physical activity were generated based on feedback from the first round of interviews and the educational modules developed for the mHealth intervention. Message content included positive self-talk, reminders to pace yourself, and reminders to hydrate. The newly developed physical activity messages were mapped to relevant behavior change theories by 2 investigators working independently (AD and PC), and coding was then compared until a final coding was assigned for each message. Codes were reviewed by 2 other investigators (SK and SF), and messages were edited to improve autonomy-supportive language in keeping with a central concept of SDT. Thirty new messages related to fatigue were also created and included testimonials from individuals living with lupus, reminders to gauge energy levels, and strategies to manage fatigue. All messages were revised to ensure an eighth-grade or lower reading level using Readability.io and were limited to 160 characters to fit within a standard SMS text message.

Interview participants were then invited to review the messages and provide feedback in a third round of interviews. Ten individuals participated and rated the messages highly, with an average message rating of 1.3 out of 5 (where 1 is highest). Four messages were removed from the library based on low ratings by participants. Interviewees shared open-ended comments, which were used to revise the messages. Twenty-seven messages were modified to minimize "lupus warrior" phrasing, which received mixed feedback, and to remove "work-out," "exercise," "strength-based," or "aerobic" as these terms were felt to be more intimidating than "movement" or "physical activity." Several messages were edited to use more autonomy-supportive language. Eleven new messages were generated based on feedback about content missing from the library, including "lupus life hacks," managing pain, and evolving while living with lupus. The complete messaging library consisted of 154 messages (Table S1 in Multimedia Appendix 4).

Discussion

Principal Findings

Our study used an HCD process to develop a novel mHealth app, grounded in behavior change theory, for self-management of fatigue in SLE. We describe the first 2 phases of

XSL•FO
RenderX

development: inspiration and ideation. In the inspiration phase, we identified key user needs, including high-quality educational resources, symptom and activity tracking features, social connection, and reminders. In the ideation phase, we created an app prototype using an iterative process informed by feedback from people with SLE at every stage of the design process. One key point of iteration was the user desire for a wide range of notifications and messaging. Based on this feedback, we generated a novel messaging library anchored in behavior change techniques for use with the mHealth app. Using this approach, we were able to create an mHealth app tailored specifically to stakeholder needs that users reported they were highly likely to use.

Our app offers a unique tool for the self-management of fatigue in SLE. Prior apps designed for this patient population often lacked reliable educational information and had limited functionality, with little involvement from people with SLE in the design process [44,67]. Existing mHealth apps for the self-management of SLE have shown potential to improve quality of life metrics but could benefit from a stronger HCD and a more rigorous approach to behavior change theory [68]. By applying an HCD approach and engaging stakeholders throughout development, we identified essential mHealth functions valued by individuals with SLE. Participants in our study expressed a need for symptom trackers, specifically for fatigue, pain, sleep, and physical function, which aligns with previous findings [69]. In addition, we identified key priorities for mHealth tools among those with SLE, including social connection, accessibility and inclusion, customizability, and integration with existing digital health tools for disease self-management. These aspects have received limited attention in the literature to date.

mHealth apps have been successfully developed for the self-management of various chronic conditions, including chronic pain, diabetes, rheumatoid arthritis, and chronic lung disease [70-73]. Many of these apps have incorporated symptom tracking, medication reminders, lifestyle monitoring, and education, while also integrating behavior change theory and stakeholder feedback into their design [74]. However, to our knowledge, this process has not yet been systematically applied to mHealth interventions for SLE.

Our mHealth app, with its symptom and physical activity tracking and evidence-based education, shares features with these previously developed self-management apps. However, it was specifically designed to address the unique needs of individuals with SLE, particularly in the context of fatigue management. While our prototype includes core features such as symptom tracking, lifestyle monitoring through physical activity tracking, and high-quality educational modules, we were unable to incorporate medication tracking, reminders, or food logs due to design constraints. Notably, these features were also identified as important by participants and represent potential areas for future development.

Through its grounding in behavior change theory and its emphasis on user-centered design, our mHealth app fills a critical gap in existing digital health tools for SLE. Future research is planned to evaluate the mHealth app in the third and final stage of the HCD process, implementation. We plan to pilot test the usability and effectiveness of the app in a clinical trial of a peer coaching intervention for SLE-related fatigue. Findings from this study will inform further refinements and implementation of the app.

Limitations of this study included a small sample size and design constraints. Each round of interviews had a small sample size, which could have biased the results of this study. While participants were purposively sampled based on demographics and disease characteristics to represent a variety of perspectives, it is possible that the needs identified in the focus groups and interviews do not represent those of the entire population of people with SLE. In addition, the app development was restricted in scope by our design challenge to address fatigue in SLE and by budgetary constraints. This limited our ability to create an mHealth app that addressed every feature requested by interview participants.

Conclusion

In conclusion, this study presents a structured methodology for developing an mHealth app to support fatigue self-management in individuals with SLE. By using an HCD approach and integrating behavior change theory, we created a tool that directly addresses the needs of people with SLE—a group for whom digital health technologies have been limited. Our iterative design process, which actively engaged end users, demonstrates the feasibility of applying HCD in this setting and underscores the value of user involvement in enhancing adoption and usability. This approach offers a rigorous framework for mHealth app development that can be adapted to further expand digital health solutions for the self-management of SLE and other chronic conditions.

Acknowledgments

The authors acknowledge Stephanie Tate, Thomas Farrar, Sanjana Murthy, and Maria Blanco for their contributions in developing the educational content for the app, Andre Dantas for technical expertise in developing the app, and all of the patients who participated in the formative work for the study. The authors attest that there was no use of generative artificial intelligence (AI) technology in the generation of text, figures, or other informational content of this manuscript. This work was supported by grant K23AR078177 from the National Institute of Arthritis and Musculoskeletal and Skin Diseases (SK), a medical student preceptorship award from the Rheumatology Research Foundation (AD, SK) and a Marc R. Chevrier, MD, PhD, FACR, Lupus Research Memorial Fund medical student preceptorship award from the Rheumatology Research Foundation (PC, SK). The project was also supported by the National Center for Advancing Translational Sciences award number UL1TR002544. The funding bodies were not involved in the study design, data collection, analysis, data interpretation, or manuscript development.

XSL•FO
RenderX

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

SK conceived the study idea and secured funding. SK, LD, TEM, FC, and MGM were involved with app design and development. SK, KS, AL, AD, PC, FC, MGM, and SF contributed to conduct and analysis of the focus groups and interviews. AD, PC, SK, and SF developed the app messaging library. AD and SK drafted the manuscript and prepared figures, tables, and supplemental materials.

Conflicts of Interest

LD and TEM report ownership interests in Ambulomics, which developed the app. MGM reports being a patient consult/advocate/speaker for BMS, a patient advocate for J&J, and a patient experience researcher/advocate for Boehringer Ingelheim. The other authors declare no conflicts of interest.

Multimedia Appendix 1
Focus group guide.
[DOCX File, 25 KB - formative_v9i1e75399_app1.docx]

Multimedia Appendix 2
Interview guide.
[DOCX File, 20 KB - formative_v9i1e75399_app2.docx]

Multimedia Appendix 3
Prototype screenshots.
[DOCX File, 1355 KB - formative_v9i1e75399_app3.docx]

Multimedia Appendix 4
App messages.
[DOCX File, 94 KB - formative_v9i1e75399_app4.docx]

References

1. Burgos PI, Alarcón GS, McGwin G, Crews KQ, Reveille JD, Vilá LM. Disease activity and damage are not associated with increased levels of fatigue in systemic lupus erythematosus patients from a multiethnic cohort: LXVII. Arthritis Rheum 2009 Sep 15;61(9):1179-1186. [doi: 10.1002/art.24649] [Medline: 19714612]
2. Tench CM, McCurdie I, White PD, D'Cruz DP. The prevalence and associations of fatigue in systemic lupus erythematosus. Rheumatology (Oxford) 2000 Nov;39(11):1249-1254. [doi: 10.1093/rheumatology/39.11.1249] [Medline: 11085805]
3. Baker K, Pope J. Employment and work disability in systemic lupus erythematosus: a systematic review. Rheumatology (Oxford) 2009 Mar;48(3):281-284. [doi: 10.1093/rheumatology/ken477] [Medline: 19153144]
4. Booth S, Price E, Walker E. Fluctuation, invisibility, fatigue - the barriers to maintaining employment with systemic lupus erythematosus: results of an online survey. Lupus (Los Angel) 2018 Dec;27(14):2284-2291. [doi: 10.1177/0961203318808593] [Medline: 30451638]
5. Agarwal N, Kumar V. Burden of lupus on work: Issues in the employment of individuals with lupus. Work 2016 Oct 17;55(2):429-439. [doi: 10.3233/WOR-162398] [Medline: 27689581]
6. Garris C, Oglesby A, Sulcs E, Lee M. Impact of systemic lupus erythematosus on burden of illness and work productivity in the United States. Lupus (Los Angel) 2013 Sep;22(10):1077-1086. [doi: 10.1177/0961203313498795] [Medline: 23920376]
7. Bultink IEM, Turkstra F, Dijkmans BAC, Voskuyl AE. High prevalence of unemployment in patients with systemic lupus erythematosus: association with organ damage and health-related quality of life. J Rheumatol 2008 Jun;35(6):1053-1057. [Medline: 18381792]
8. Yelin E, Trupin L, Katz P, et al. Work dynamics among persons with systemic lupus erythematosus. Arthritis Rheum 2007 Feb 15;57(1):56-63. [doi: 10.1002/art.22481] [Medline: 17266065]
9. Moses N, Wiggers J, Nicholas C, Cockburn J. Prevalence and correlates of perceived unmet needs of people with systemic lupus erythematosus. Patient Educ Couns 2005 Apr;57(1):30-38. [doi: 10.1016/j.pec.2004.03.015] [Medline: 15797150]
10. Danoff-Burg S, Friedberg F. Unmet needs of patients with systemic lupus erythematosus. Behav Med 2009;35(1):5-13. [doi: 10.3200/BMED.35.1.5-13] [Medline: 19297299]
11. Zirkzee EJM, Steup-Beekman GM, Schouffoer AA, et al. Health care in systemic lupus erythematosus (SLE): the patient's perspective. Clin Rheumatol 2014 Sep;33(9):1279-1287. [doi: 10.1007/s10067-014-2595-1] [Medline: 24744153]

XSL•FO
RenderX

12. Kasturi S, Ahearn EL, Batterman A, et al. Measuring what matters: a qualitative study of the relevance and clinical utility of PROMIS surveys in systemic lupus erythematosus. J Rheumatol 2023 Jul 1:jrheum.2023-0184. [doi: 10.3899/jrheum.2023-0184] [Medline: 37399463]
13. Bruce IN, Mak VC, Hallett DC, Gladman DD, Urowitz MB. Factors associated with fatigue in patients with systemic lupus erythematosus. Ann Rheum Dis 1999 Jun;58(6):379-381. [doi: 10.1136/ard.58.6.379] [Medline: 10340963]
14. Jump RL, Robinson ME, Armstrong AE, Barnes EV, Kilbourn KM, Richards HB. Fatigue in systemic lupus erythematosus: contributions of disease activity, pain, depression, and perceived social support. J Rheumatol 2005 Sep;32(9):1699-1705. [Medline: 16142863]
15. Schmeding A, Schneider M. Fatigue, health-related quality of life and other patient-reported outcomes in systemic lupus erythematosus. Best Pract Res Clin Rheumatol 2013 Jun;27(3):363-375. [doi: 10.1016/j.berh.2013.07.009] [Medline: 24238693]
16. Da Costa D, Dritsa M, Bernatsky S, et al. Dimensions of fatigue in systemic lupus erythematosus: relationship to disease status and behavioral and psychosocial factors. J Rheumatol 2006 Jul;33(7):1282-1288. [Medline: 16758508]
17. Zonana-Nacach A, Roseman JM, McGwin G, Friedman AW, Baethge BA, Reveille JD. Systemic lupus erythematosus in three ethnic groups. VI: factors associated with fatigue within 5 years of criteria diagnosis. Lupus (Los Angel) 2000 Feb;9(2):101-109. [doi: 10.1191/096120300678828046]
18. Iaboni A, Ibanez D, Gladman DD, Urowitz MB, Moldofsky H. Fatigue in systemic lupus erythematosus: contributions of disordered sleep, sleepiness, and depression. J Rheumatol 2006 Dec;33(12):2453-2457. [Medline: 17143980]
19. Dobkin PL, Da Costa D, Fortin PR, et al. Living with lupus: a prospective pan-Canadian study. J Rheumatol 2001 Nov;28(11):2442-2448. [Medline: 11708416]
20. Arnaud L, Mertz P, Amoura Z, et al. Patterns of fatigue and association with disease activity and clinical manifestations in systemic lupus erythematosus. Rheumatology (Oxford) 2021 Jun 18;60(6):2672-2677. [doi: 10.1093/rheumatology/keaa671] [Medline: 33175957]
21. Neill J, Belan I, Ried K. Effectiveness of non-pharmacological interventions for fatigue in adults with multiple sclerosis, rheumatoid arthritis, or systemic lupus erythematosus: a systematic review. J Adv Nurs 2006 Dec;56(6):617-635. [doi: 10.1111/j.1365-2648.2006.04054.x] [Medline: 17118041]
22. del Pino-Sedeño T, Trujillo-Martín MM, Ruiz-Irastorza G, et al. Effectiveness of nonpharmacologic interventions for decreasing fatigue in adults With systemic lupus erythematosus: a systematic review. Arthritis Care Res (Hoboken) 2016 Jan;68(1):141-148. [doi: 10.1002/acr.22675] [Medline: 26238554]
23. Fangtham M, Kasturi S, Bannuru RR, Nash JL, Wang C. Non-pharmacologic therapies for systemic lupus erythematosus. Lupus (Los Angel) 2019 May;28(6):703-712. [doi: 10.1177/0961203319841435] [Medline: 30961418]
24. Robb-Nicholson LC, Daltroy L, Eaton H, et al. Effects of aerobic conditioning in lupus fatigue: a pilot study. Br J Rheumatol 1989 Dec;28(6):500-505. [doi: 10.1093/rheumatology/28.6.500] [Medline: 2590802]
25. Ramsey-Goldman R, Schilling EM, Dunlop D, et al. A pilot study on the effects of exercise in patients with systemic lupus erythematosus. Arthritis Care Res 2000;13(5):262-269. [doi: 10.1002/1529-0131(200010)13:5<262::aid-anr4>3.0.co;2-8]
26. Tench CM, McCarthy J, McCurdie I, White PD, D'Cruz DP. Fatigue in systemic lupus erythematosus: a randomized controlled trial of exercise. Rheumatology (Oxford) 2003 Sep;42(9):1050-1054. [doi: 10.1093/rheumatology/keg289] [Medline: 12730519]
27. Karlson EW, Liang MH, Eaton H, et al. A randomized clinical trial of a psychoeducational intervention to improve outcomes in systemic lupus erythematosus. Arthritis Rheum 2004 Jun;50(6):1832-1841. [doi: 10.1002/art.20279] [Medline: 15188360]
28. Sohng KY. Effects of a self-management course for patients with systemic lupus erythematosus. J Adv Nurs 2003 Jun;42(5):479-486. [doi: 10.1046/j.1365-2648.2003.02647.x] [Medline: 12752868]
29. Carvalho MD, Sato EI, Tebexreni AS, Heidecher RTC, Schenkman S, Neto TLB. Effects of supervised cardiovascular training program on exercise tolerance, aerobic capacity, and quality of life in patients with systemic lupus erythematosus. Arthritis Rheum 2005 Dec 15;53(6):838-844 [FREE Full text] [doi: 10.1002/art.21605]
30. Frade S, O'Neill S, Walsh S, et al. Telehealth-supervised exercise in systemic lupus erythematosus: a pilot study. Lupus (Los Angel) 2023 Apr;32(4):508-520. [doi: 10.1177/09612033231157073] [Medline: 36803286]
31. Boström C, Elfving B, Dupré B, Opava CH, Lundberg IE, Jansson E. Effects of a one-year physical activity programme for women with systemic lupus erythematosus - a randomized controlled study. Lupus (Los Angel) 2016 May;25(6):602-616. [doi: 10.1177/0961203315622817] [Medline: 26768748]
32. Li LC, Feehan LM, Xie H, et al. Efficacy of a physical activity counseling program with use of a wearable tracker in people with inflammatory arthritis: a randomized controlled trial. Arthritis Care Res (Hoboken) 2020 Dec;72(12):1755-1765. [doi: 10.1002/acr.24199] [Medline: 32248626]
33. Avaux M, Hoellinger P, Nieuwland-Husson S, Fraselle V, Depresseux G, Houssiau FA. Effects of two different exercise programs on chronic fatigue in lupus patients. Acta Clin Belg 2016 Dec;71(6):403-406. [doi: 10.1080/17843286.2016.1200824] [Medline: 27377292]
34. Lopes-Souza P, Dionello CF, Bernardes-Oliveira CL, et al. Effects of 12-week whole-body vibration exercise on fatigue, functional ability and quality of life in women with systemic lupus erythematosus: a randomized controlled trial. J Bodyw Mov Ther 2021 Jul;27:191-199. [doi: 10.1016/j.jbmt.2021.01.015] [Medline: 34391233]

XSL•FO
RenderX

35. Sheikh SZ, Kaufman K, Gordon BB, et al. Evaluation of the self-directed format of Walk With Ease in patients with systemic lupus erythematosus: the Walk-SLE Pilot Study. Lupus (Los Angel) 2019 May;28(6):764-770. [doi: 10.1177/0961203319846387] [Medline: 31042128]
36. Moller AC, Merchant G, Conroy DE, et al. Applying and advancing behavior change theories and techniques in the context of a digital health revolution: proposals for more effectively realizing untapped potential. J Behav Med 2017 Feb;40(1):85-98. [doi: 10.1007/s10865-016-9818-7] [Medline: 28058516]
37. Aguilera A, Figueroa CA, Hernandez-Ramos R, et al. mHealth app using machine learning to increase physical activity in diabetes and depression: clinical trial protocol for the DIAMANTE Study. BMJ Open 2020 Aug 20;10(8):e034723. [doi: 10.1136/bmjopen-2019-034723] [Medline: 32819981]
38. Nahum-Shani I, Smith SN, Spring BJ, et al. Just-in-Time Adaptive Interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. Ann Behav Med 2018 May 18;52(6):446-462. [doi: 10.1007/s12160-016-9830-8] [Medline: 27663578]
39. Zhao J, Freeman B, Li M. Can mobile phone apps influence people's health behavior change? An evidence review. J Med Internet Res 2016 Oct 31;18(11):e287. [doi: 10.2196/jmir.5692] [Medline: 27806926]
40. Payne HE, Lister C, West JH, Bernhardt JM. Behavioral functionality of mobile apps in health interventions: a systematic review of the literature. JMIR Mhealth Uhealth 2015 Feb 26;3(1):e20. [doi: 10.2196/mhealth.3335] [Medline: 25803705]
41. Schoeppe S, Alley S, Van Lippevelde W, et al. Efficacy of interventions that use apps to improve diet, physical activity and sedentary behaviour: a systematic review. Int J Behav Nutr Phys Act 2016 Dec 7;13(1):127. [doi: 10.1186/s12966-016-0454-y] [Medline: 27927218]
42. Gal R, May AM, van Overmeeren EJ, Simons M, Monninkhof EM. The effect of physical activity interventions comprising wearables and smartphone applications on physical activity: a systematic review and meta-analysis. Sports Med Open 2018 Sep 3;4(1):42. [doi: 10.1186/s40798-018-0157-9] [Medline: 30178072]
43. Mönninghoff A, Kramer JN, Hess AJ, et al. Long-term effectiveness of mHealth physical activity interventions: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2021 Apr 30;23(4):e26699. [doi: 10.2196/26699] [Medline: 33811021]
44. Dantas LO, Weber S, Osani MC, Bannuru RR, McAlindon TE, Kasturi S. Mobile health technologies for the management of systemic lupus erythematosus: a systematic review. Lupus (Los Angel) 2020 Feb;29(2):144-156. [doi: 10.1177/0961203319897139] [Medline: 31924145]
45. Tani C, Trieste L, Lorenzoni V, Cannizzo S, Turchetti G, Mosca M. Health information technologies in systemic lupus erythematosus: focus on patient assessment. Clin Exp Rheumatol 2016;34(5 Suppl 101):S54-S56. [Medline: 27762204]
46. Khan F, Granville N, Malkani R, Chathampally Y. Health-related quality of life improvements in systemic lupus erythematosus derived from a digital therapeutic plus tele-health coaching intervention: randomized controlled pilot trial. J Med Internet Res 2020 Oct 20;22(10):e23868. [doi: 10.2196/23868] [Medline: 33079070]
47. Pasyar N, Sam A, Rivaz M, Nazarinia M. A smartphone-based supportive counseling on health anxiety and acceptance of disability in systemic lupus erythematosus patients: a randomized clinical trial. Patient Educ Couns 2023 May;110:107676. [doi: 10.1016/j.pec.2023.107676] [Medline: 36841083]
48. Ryan RM, Deci EL. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. Am Psychol 2000 Jan;55(1):68-78. [doi: 10.1037//0003-066x.55.1.68] [Medline: 11392867]
49. Bandura A. Health promotion by social cognitive means. Health Educ Behav 2004 Apr;31(2):143-164. [doi: 10.1177/1090198104263660] [Medline: 15090118]
50. Bandura A. Self-efficacy: toward a unifying theory of behavioral change. Psychol Rev 1977 Mar;84(2):191-215. [doi: 10.1037//0033-295x.84.2.191] [Medline: 847061]
51. González-Cutre D, Megías Á, Beltrán-Carrillo VJ, Cervelló E, Spray CM. Effects of a physical activity program on post-bariatric patients: a qualitative study from a self-determination theory perspective. J Health Psychol 2020 Sep;25(10-11):1743-1754. [doi: 10.1177/1359105318770729] [Medline: 29722281]
52. Patrick H, Canevello A. Methodological overview of a self-determination theory-based computerized intervention to promote leisure-time physical activity. Psychol Sport Exerc 2011 Jan;12(1):13-19. [doi: 10.1016/j.psychsport.2010.04.011] [Medline: 21103069]
53. Fortier MS, Duda JL, Guerin E, Teixeira PJ. Promoting physical activity: development and testing of self-determination theory-based interventions. Int J Behav Nutr Phys Act 2012 Mar 2;9(1):20. [doi: 10.1186/1479-5868-9-20] [Medline: 22385751]
54. Holeman I, Kane D. Human-centered design for global health equity. Inf Technol Dev 2019 Sep 29;26(3):477-505. [doi: 10.1080/02681102.2019.1667289] [Medline: 32982007]
55. Nijagal MA, Patel D, Lyles C, et al. Using human centered design to identify opportunities for reducing inequities in perinatal care. BMC Health Serv Res 2021 Jul 20;21(1):714. [doi: 10.1186/s12913-021-06609-8] [Medline: 34284758]
56. Field guide to human centered design. Ideo. 2015. URL: https://www.ideo.com/post/design-kit [accessed 2025-07-08]
57. Aringer M, Costenbader K, Daikh D, et al. 2019 European League Against Rheumatism/American College of Rheumatology Classification Criteria for Systemic Lupus Erythematosus. Arthritis Rheumatol 2019 Sep;71(9):1400-1412. [doi: 10.1002/art.40930] [Medline: 31385462]

XSL•FO
RenderX

58. Kasturi S, Tate S, Farrar T, et al. Development of a theory-based educational intervention for fatigue in systemic lupus erythematosus (SLE) using human-centered design methodology. Arthritis Rheumatol 2022;74 [FREE Full text]
59. van SM, Barnard YF, Sandberg JAC. The Think Aloud Method: A Practical Guide to Modelling Cognitive Processes: Academic Press; 1994.
60. Jaspers MWM, Steen T, Bos C, Geenen M. The think aloud method: a guide to user interface design. Int J Med Inform 2004 Nov;73(11-12):781-795. [doi: 10.1016/j.ijmedinf.2004.08.003]
61. Hamilton AB, Finley EP. Qualitative methods in implementation research: an introduction. Psychiatry Res 2019 Oct;280:112516. [doi: 10.1016/j.psychres.2019.112516] [Medline: 31437661]
62. Feedback capture grid. Interaction design foundation. 2016. URL: https://www.interaction-design.org/literature/topics/feedback-capture-grid [accessed 2025-07-08]
63. Miserandino C. The spoon theory but you don't look sick. But You Dont Look Sick. 2003. URL: https://butyoudontlooksick.com/articles/written-by-christine/the-spoon-theory [accessed 2025-07-08]
64. MacPherson MM, Cranston KD, Locke SR, Bourne JE, Jung ME. Using the behavior change wheel to develop text messages to promote diet and physical activity adherence following a diabetes prevention program. Transl Behav Med 2021 Aug 13;11(8):1585-1595. [doi: 10.1093/tbm/ibab058]
65. MacPherson M, Cranston K, Johnston C, Locke S, Jung ME. Evaluation and refinement of a bank of SMS text messages to promote behavior change adherence following a diabetes prevention program: survey study. JMIR Form Res 2021 Aug 27;5(8):e28163. [doi: 10.2196/28163] [Medline: 34448713]
66. Michie S, Atkins L, West R. The Behavior Change Wheel: A Guide to Designing Interventions, 1st edition: Silverback Publishing; 2014.
67. Ramasamy A, Dugyala P, Mohan C. Mobile health apps for systemic lupus erythematosus and lupus nephritis: a critical appraisal. Arthritis Res Ther 2022;24:110. [doi: 10.1186/s13075-022-02791-0]
68. Haglo H, Wang E, Berg OK, Hoff J, Helgerud J. Smartphone-assisted high-intensity interval training in inflammatory rheumatic disease patients: randomized controlled trial. JMIR Mhealth Uhealth 2021 Oct 21;9(10):e28124. [doi: 10.2196/28124] [Medline: 34673536]
69. Nowell WB, Gavigan K, Kannowski CL, et al. Which patient-reported outcomes do rheumatology patients find important to track digitally? A real-world longitudinal study in ArthritisPower. Arthritis Res Ther 2021 Feb 10;23(1):53. [doi: 10.1186/s13075-021-02430-0] [Medline: 33568191]
70. Moreno-Ligero M, Moral-Munoz JA, Salazar A, Failde I. mHealth intervention for improving pain, quality of life, and functional disability in patients with chronic pain: systematic review. JMIR Mhealth Uhealth 2023 Feb 2;11:e40844. [doi: 10.2196/40844] [Medline: 36729570]
71. Quach S, Michaelchuk W, Benoit A, et al. Mobile heath applications for self-management in chronic lung disease: a systematic review. Netw Model Anal Health Inform Bioinform 2023;12(1):25. [doi: 10.1007/s13721-023-00419-0] [Medline: 37305790]
72. Mollard E, Michaud K. Self-management of rheumatoid arthritis: mobile applications. Curr Rheumatol Rep 2020 Nov 26;23(1):2. [doi: 10.1007/s11926-020-00968-7] [Medline: 33244621]
73. Brzan PP, Rotman E, Pajnkihar M, Klanjsek P. Mobile applications for control and self management of diabetes: a systematic review. J Med Syst 2016 Sep;40(9):210. [doi: 10.1007/s10916-016-0564-8] [Medline: 27520615]
74. Song T, Yu P, Bliokas V, et al. A clinician-led, experience-based co-design approach for developing mHealth services to support the patient self-management of chronic conditions: development study and design case. JMIR Mhealth Uhealth 2021 Jul 20;9(7):e20650. [doi: 10.2196/20650] [Medline: 34283030]

Abbreviations

HCD: human-centered design
***ICD-10*:** *International Classification of Diseases, Tenth Revision*
***ICD-9*:** *International Classification of Diseases, Ninth Revision*
mHealth: mobile health
REDCap: Research Electronic Data Capture
SCT: social cognitive theory
SDT: self-determination theory
SLE: systemic lupus erythematosus

XSL•FO
RenderX

Edited by A Mavragani; submitted 02.04.25; peer-reviewed by E Ross, E Jupe, LV Costa, RL Randell, S Elghreeb; revised version received 06.06.25; accepted 10.06.25; published 26.08.25.

Please cite as:
Deck A, Singh K, Caras P, LeClair A, Gore-Massy M, Chiu F, Dantas LO, McAlindon T, Folta S, Kasturi S
Development of a Theory-Based mHealth App for Fatigue Management in Lupus: Human-Centered Design Approach
JMIR Form Res 2025;9:e75399
URL: https://formative.jmir.org/2025/1/e75399
doi:10.2196/75399

© Anna Deck, Kiran Singh, Paula Caras, Amy LeClair, Monique Gore-Massy, Faye Chiu, Lucas Ogura Dantas, Timothy McAlindon, Sara Folta, Shanthini Kasturi. Originally published in JMIR Formative Research (https://formative.jmir.org), 26.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Transforming Patient Feedback Into Actionable Insights Through Natural Language Processing: Knowledge Discovery and Action Research Study

Ravi Shankar[1], PhD; Alexander Yip[2], MBBS

[1]Medical Affairs – Research Innovation & Enterprise, Alexandra Hospital, National University Health System, 378 Alexandra Rd, Singapore, Singapore
[2]Alexandra Research Centre for Healthcare in a Virtual Environment, Alexandra Hospital, National University Health System, Singapore, Singapore

Corresponding Author:
Ravi Shankar, PhD
Medical Affairs – Research Innovation & Enterprise, Alexandra Hospital, National University Health System, 378 Alexandra Rd, Singapore, Singapore

Abstract

Background: Patient feedback has emerged as a critical measure of health care quality and a key driver of organizational performance. Traditional manual analysis of unstructured patient feedback presents significant challenges as data volumes grow, making it difficult to extract meaningful patterns and actionable insights efficiently.

Objective: We attempted to develop and evaluate a comprehensive methodology for analyzing patient feedback data using natural language processing and Knowledge Discovery in Databases (KDD) approaches, aiming to identify key patterns, themes, and variations in patient experiences across different demographic groups and care settings, and to translate these insights into actionable improvements in health care delivery.

Methods: This study applied an integrated KDD-action research framework to analyze 126,134 patient feedback entries collected at Alexandra Hospital, Singapore, in 2023. A comprehensive suite of text mining techniques, including sentiment analysis, topic modeling, emotion detection, and aspect-based sentiment analysis, was employed to uncover patterns in patient-reported experiences. The dataset included 92,578 (73.4%) entries containing free-text comments, comprising 1,568,932 tokens with an average comment length of 16.9 words. Multiple analytical techniques were used to ensure the validity and reliability of the findings. Stakeholder engagement throughout the research process facilitated the translation of analytical insights into practical improvements. The study was granted an exemption from ethical review by the National Healthcare Group Domain Specific Review Board (number: NUS-IRB-2025-087E), with a waiver of informed consent granted for this retrospective analysis of deidentified patient feedback data.

Results: Text mining analysis revealed a moderately positive overall sentiment across the feedback corpus (average polarity score: 0.42), with 68.8% (63,685/92,578) of comments classified as positive, 25.4% (23,515/92,578) as neutral, and 5.8% (5378/92,578) as negative. Topic modeling identified 10 distinct topics, including staff attitude and service (9443/92,578, 10.2%), health care staff professionalism (9350/92,578, 10.1%), hospital environment (9258/92,578, 10.0%), and waiting time (9258/92,578, 10.0%). Aspect-based sentiment analysis highlighted nurse attitude (sentiment score: 0.65), staff helpfulness (0.61), and doctor expertise (0.58) as the most positive aspects, while waiting time (−0.42) and billing transparency (−0.28) emerged as the most negative. Demographic segmentation revealed significant variations in patient priorities, with younger patients (<35 years) expressing 37% more concerns about digital accessibility and efficiency than older patients who valued face-to-face interactions 42% more highly. Implementation of targeted interventions based on these findings resulted in measurable improvements, including an 18% increase in waiting time satisfaction, a 15% improvement in doctor-patient communication ratings, and a 23% reduction in billing-related complaints.

Conclusions: The integration of natural language processing techniques with KDD and action research principles provides a powerful framework for transforming unstructured patient feedback into actionable insights for health care improvement. This approach enables health care organizations to understand the complex patterns and drivers of patient experience, identify targeted improvement opportunities, and implement evidence-based initiatives that enhance care quality and patient-centeredness.

(JMIR Form Res 2025;9:e69699) doi:10.2196/69699

KEYWORDS

natural language processing; text mining; health care quality improvement; knowledge discovery in databases; action research; sentiment analysis; artificial intelligence; AI

XSL•FO
RenderX

Introduction

In the rapidly evolving landscape of health care, patient experience has emerged as a critical measure of care quality and a key driver of organizational performance [1]. As health care systems shift toward patient-centeredness and value-based care, the ability to understand and enhance the entire patient journey has become a strategic imperative [2]. At the heart of this transformation lies the effective collection, analysis, and application of patient feedback data [3].

Traditionally, health care organizations have relied on structured surveys to assess patient experiences. While providing standardized metrics, surveys often fail to capture the full richness and complexity of patients' perceptions and emotions [4]. Open-ended comments, where patients express their thoughts in their own words, offer valuable, complementary insights into satisfaction drivers [5]. However, the unstructured nature of free-text feedback presents significant challenges for traditional manual analysis, especially as data volumes grow [6]. The patterns and themes within may not be readily apparent, limiting the depth of insights [7].

Recent advancements in text mining and natural language processing (NLP) have opened new possibilities for automated analysis of free-text patient feedback [8]. Machine learning algorithms and computational linguistic methods allow for efficiently processing large volumes of unstructured data to extract meaningful knowledge [9]. Techniques like sentiment analysis, topic modeling, and aspect-based sentiment analysis have been successfully applied to patient experience data in various contexts, uncovering insights into factors shaping perceptions and behaviors [10-14].

Despite these advances, significant gaps remain in the literature regarding the systematic integration of multiple text mining techniques within established frameworks. Previous studies have typically employed single analytical methods (eg, only sentiment analysis or topic modeling), which can limit the depth and breadth of insights [9,15]. Moreover, most existing research focuses solely on the technical aspects of analysis without addressing the critical challenge of translating computational findings into practical health care improvements [16]. There is a particular lack of frameworks that combine rigorous data mining methodologies with stakeholder engagement processes to ensure that insights are both technically sound and practically actionable.

This study addresses these gaps by developing and evaluating a comprehensive framework that (1) integrates multiple NLP techniques within the Knowledge Discovery in Databases (KDD) process to provide multidimensional analysis of patient feedback; (2) incorporates action research principles to ensure stakeholder engagement throughout the analysis process; and (3) demonstrates how analytical insights can be systematically translated into measurable health care improvements.

The primary objective of this study is to develop and validate an integrated methodology that combines NLP, knowledge discovery, and action research principles to transform unstructured patient feedback into actionable insights for health care quality improvement.

Methods

Study Design and Theoretical Framework

This study employed a mixed methods approach integrating the KDD process [17] with action research principles [18] to analyze patient feedback data. The KDD process provided a systematic framework for extracting patterns from large-scale unstructured data through 5 stages: selection, preprocessing, transformation, data mining, and interpretation/evaluation. Action research complemented this technical approach by ensuring stakeholder engagement and practical relevance through iterative cycles of planning, action, observation, and reflection [19]. This hybrid framework enabled both rigorous computational analysis and contextually grounded interpretation of the findings (Figure 1).

Figure 1. Research framework. KDD: Knowledge Discovery in Databases.

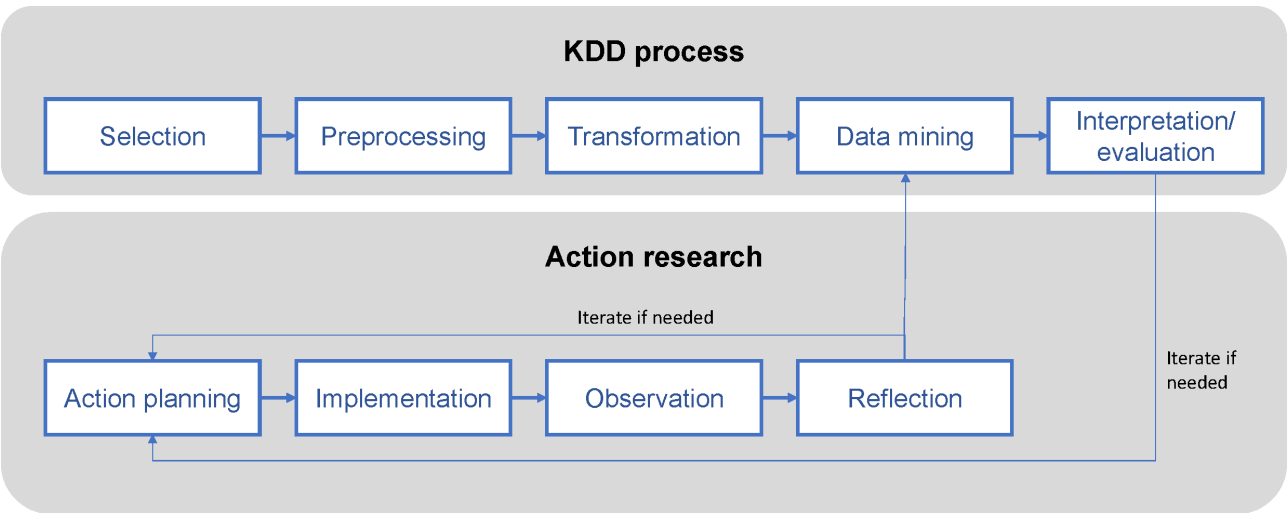

XSL•FO
RenderX

Data Selection

The dataset consists of 126,134 patient feedback entries collected through the hospital's patient experience portal in 2023, including structured metadata fields (eg, patient demographics and visit type) and unstructured free-text comments. The dataset was extracted and anonymized prior to analysis.

Data Preprocessing

The free-text comment fields underwent systematic preprocessing using a standardized pipeline implemented in Python 3.8 with the Natural Language Toolkit (NLTK) library version 3.6 [20] (detailed data preprocessing steps are provided in Section S1 in Multimedia Appendix 1).

The preprocessed corpus retained 1,568,932 tokens from the original 92,578 comments, representing an average reduction of 23% in token count while preserving semantic content.

Ethical Considerations

This study was reviewed by the National Healthcare Group Institutional Review Board and granted exemption (number: NUS-IRB-2025-087E) from full ethics review as it involved retrospective analysis of fully deidentified patient feedback data from routine hospital operations. All personally identifiable information was removed prior to analysis, including names, identification numbers, and contact details. Data were stored on secure, password-protected servers, with access restricted to authorized research personnel only.

Data Transformation

The preprocessed text data were transformed using techniques like term frequency-inverse document frequency (TF-IDF), n-gram extraction, and part-of-speech (POS) tagging to prepare it for mining (detailed data transformation steps are provided in Section S2 in Multimedia Appendix 1)

Multiple complementary text mining techniques were applied to extract insights from different dimensions of the patient feedback data:

1. Sentiment analysis: TextBlob (version 0.15.3; Loria S) was used to calculate sentiment polarity scores ranging from −1 (most negative) to +1 (most positive). Comments were classified into positive (>0.1), neutral (−0.1 to 0.1), and negative (<−0.1) categories.
2. Topic modeling: Latent Dirichlet allocation was implemented using Gensim (version 4.1.2; Řehůřek R) with 10 topics, α=.01, and β=.01 parameters. Model performance was evaluated using coherence scores and perplexity metrics.
3. Aspect-based sentiment analysis: A hybrid approach combining rule-based aspect extraction with machine learning–based sentiment classification was developed. Aspects were identified using dependency parsing and predefined health care–specific patterns.
4. Emotion detection: The NRC Emotion Lexicon was used to detect 8 basic emotions (joy, trust, fear, surprise, sadness, anger, disgust, and anticipation) in patient comments.
5. Named entity recognition: A combination of spaCy's pretrained models and custom health care entity patterns was used to identify mentions of health care professionals, departments, procedures, and facilities.

Interpretation/Evaluation

The text mining and segmentation results were interpreted and evaluated collaboratively with key stakeholders, following the action research approach, including insight synthesis, stakeholder validation, recommendation development, action planning, implementation and monitoring, and reflection and learning.

Three stakeholder workshops were conducted with health care professionals (n=15), administrators (n=8), and patient representatives (n=5) to validate the findings and co-develop improvement strategies. Workshop discussions were recorded, transcribed, and analyzed thematically to ensure comprehensive integration of stakeholder perspectives.

Results

Data Selection and Preprocessing Results

The initial data selection and preprocessing resulted in a clean, structured dataset of 126,134 feedback entries, with 92,578 (73.4%) containing at least one free-text comment. The preprocessed corpus consisted of 1,568,932 tokens, with an average comment length of 16.9 words.

Data Transformation Results

The transformation steps, including TF-IDF weighting, n-gram extraction, and POS tagging, generated rich features for the text mining analyses. The TF-IDF scores highlighted potentially significant terms like "nurse," "doctor," "care," "staff," "time," and "wait." N-gram analysis revealed frequent phrases, providing more context-specific insights into patient experience domains, such as "friendly staff," "long waiting time," "excellent care," and "room condition." POS tagging showed that nouns and adjectives were most common, indicating a focus on describing aspects, entities, and qualities of care experience. These transformed features enabled a more nuanced exploration of the feedback content and structure.

Text Mining Results

Applying the suite of text mining techniques yielded multidimensional insights into the patient feedback data.

Word Cloud

Word cloud analysis provided an engaging visual overview of the most salient terms in the patient feedback corpus. The dominant words like "nurse," "doctor," "staff," "care," "time," and "wait" indicated the centrality of human interactions, clinical processes, and environmental factors in shaping patient experiences (Multimedia Appendix 2).

N-Gram Analysis

N-gram analysis uncovered the most frequent word sequences, offering a more granular view of the key themes and topics. The top bigrams and trigrams, such as "friendly staff," "long waiting," "good care," "clean room," and "discharge process," revealed specific aspects frequently mentioned by patients, both

XSL•FO
RenderX

positively and negatively, providing valuable context and nuance to the single-word insights from the word cloud (Figure 2).

Figure 2. Top 20 bigrams (A) and trigrams (B) in patient feedback.

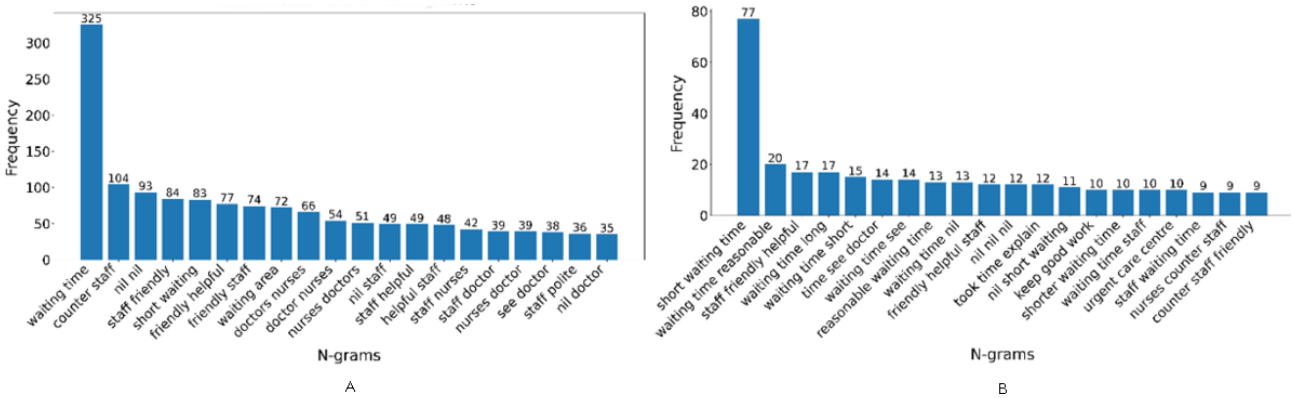

Sentiment Analysis

Sentiment analysis using the TextBlob model revealed a moderately positive overall sentiment across the feedback corpus, with an average polarity score of 0.42 (on a scale from −1 to +1). The distribution of sentiment categories showed that 68.8% (63,685/92,578) of comments were positive, 25.4% (23,515/92,578) were neutral, and 5.8% (5378/92,578) were negative. A more detailed analysis of sentiment polarity and subjectivity scores for individual comments revealed significant variation across different patient subgroups (Figure 3). This sentiment distribution underscores the importance of examining sentiment at a granular level to identify specific issues and opportunities for improvement, rather than relying solely on overall average scores.

Figure 3. Sentiment analysis of patient feedback.

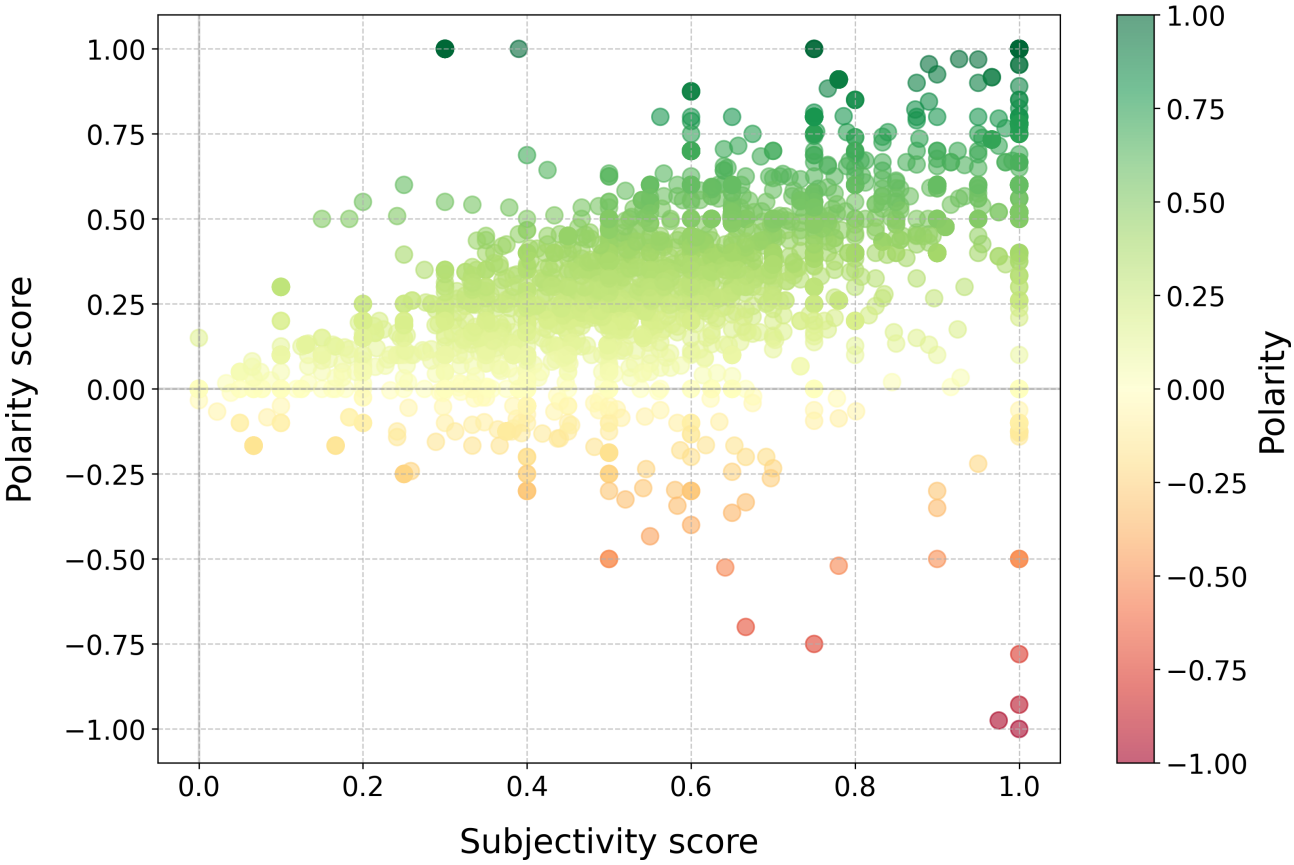

Emotion Detection

Emotion detection using the EmoLex lexicon revealed a diverse range of emotions expressed in the patient comments. The most prevalent emotions were joy (30,180/92,578, 32.6%), trust (26,200/92,578, 28.3%), anticipation (22,311/92,578, 24.1%), and surprise (18,886/92,578, 20.4%), reflecting the generally positive sentiment. However, high proportions of comments also contained negative emotions, such as sadness (17,312/92,578, 18.7%), anger (14,072/92,578, 15.2%), fear (12,776/92,578, 13.8%), and disgust (8795/92,578, 9.5%), highlighting areas of patient dissatisfaction and concern (Multimedia Appendix 3).

XSL•FO
RenderX

Linguistic Inquiry and Word Count Analysis

Linguistic Inquiry and Word Count analysis provided a comprehensive set of linguistic and psychological metrics for the feedback comments. The comments had a relatively high percentage of personal pronouns (13,146/92,578, 14.2%), positive emotion words (4814/92,578, 5.2%), social words (11,480/92,578, 12.4%), and cognitive process words (9443/92,578, 10.2%), indicating a personal, subjective, interpersonal, and reflective narrative style. Health-related words were moderately frequent (4259/92,578, 4.6%). Moreover, 5.8% (5378/92,578) of comments involved negative emotion words, representing patient dissatisfaction and concerns (Multimedia Appendix 4).

Text Classification

Text classification using a hybrid rule-based and machine learning approach achieved an overall accuracy of 82.4% in categorizing the comments by their primary content and intent. The most frequent categories were compliments (38,142/92,578, 41.2%), suggestions (22,774/92,578, 24.6%), complaints (17,405/92,578, 18.8%), and inquiries (14,257/92,578, 15.4%) (Table 1).

Table . Text classification results.

Comment category	Value (N=92,578), n (%)
Compliments	38,142 (41.2)
Suggestions	22,774 (24.6)
Complaints	17,405 (18.8)
Inquiries	14,257 (15.4)
Other categories	0 (0)

Topic Modeling

Topic modeling using the latent Dirichlet allocation algorithm identified 10 main topics in the feedback comments, capturing key themes, such as staff attitude and service, health care staff professionalism, hospital environment, patient care and communication, and waiting time and efficiency. The topic weights were relatively evenly distributed, ranging from 9.9% to 10.2%, indicating a balanced representation of various aspects of patient experience in the feedback data (Figure 4).

Figure 4. Topic modeling results. Visualization showing topic distribution and relationships. (A) Intertopic distance map (via multidimensional scaling). (B) Top 30 most relevant terms for topic 1 (30.9% of tokens). *saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t [21], **relevance (term w | topic t) = λ * p(w | t) + (1 – λ) * p(w | t)/p(w) [22].

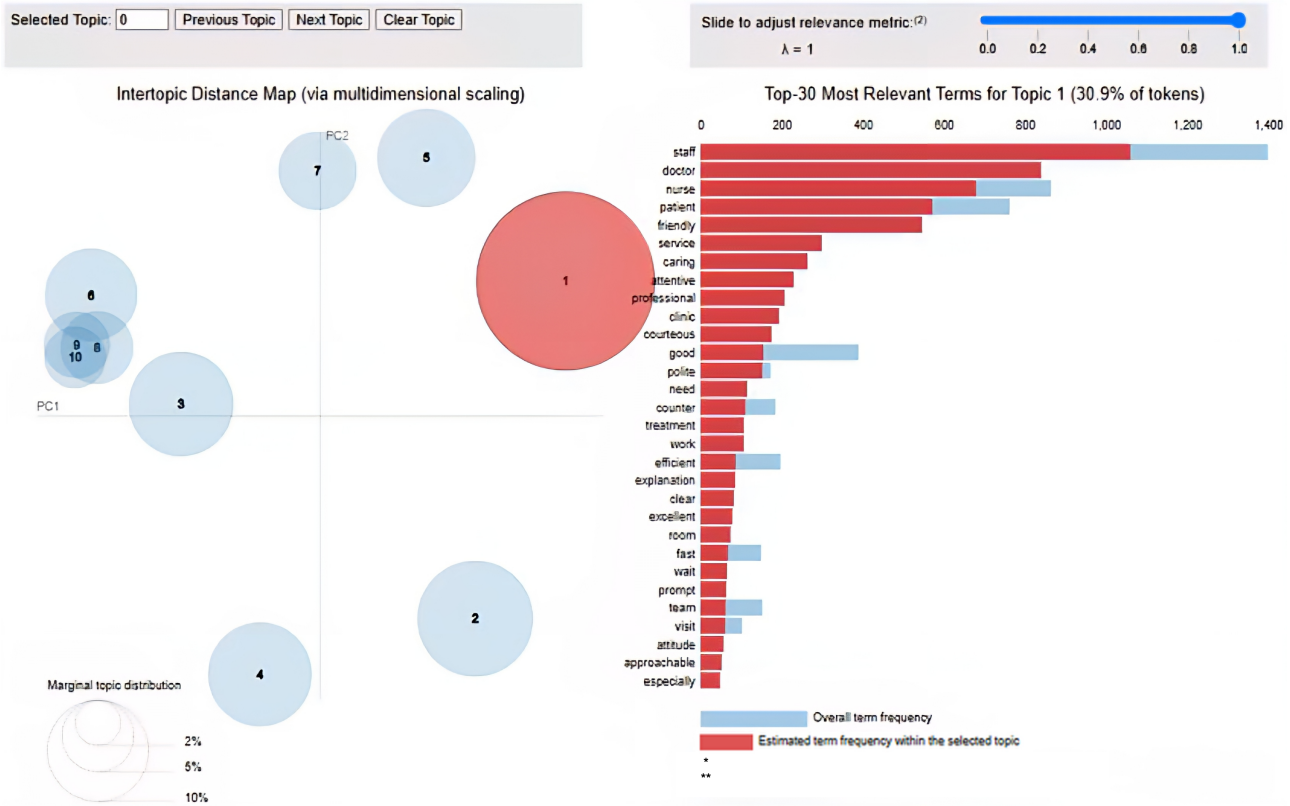

To gain further insights into the semantic composition of each topic, the top 10 words associated with each topic were examined, along with their corresponding topic-word scores. These scores indicate the importance or weight of each word in defining the topic. For example, in the staff attitude and service topic (topic 0), the top words included "friendly,"

XSL•FO
RenderX

"helpful," "courteous," "polite," and "caring," highlighting the centrality of positive staff-patient interactions. Similarly, the waiting time and efficiency topic (topic 6) was characterized by words, such as "time," "waiting," "long," "short," and "reasonable," capturing common patient frustrations with lengthy waiting periods. The topic-word scores provide a more nuanced understanding of the key concepts, attributes, and sentiments associated with each topic, enabling a richer interpretation of the patient feedback themes (Figure 5).

Figure 5. Topic word scores for the top 10 topics. Horizontal bar charts showing word importance for each topic. Topic 0: Staff attitude and service; Topic 1: Doctor-patient interaction; Topic 2: Hospital environment; Topic 3: Clinical communication and information; Topic 4: Physical facilities and comfort; Topic 5: Health care staff professionalism; Topic 6: Waiting time and efficiency; Topic 7: Administrative processes; Topic 8: Amenities and comfort; Topic 9: Patient care and communication.

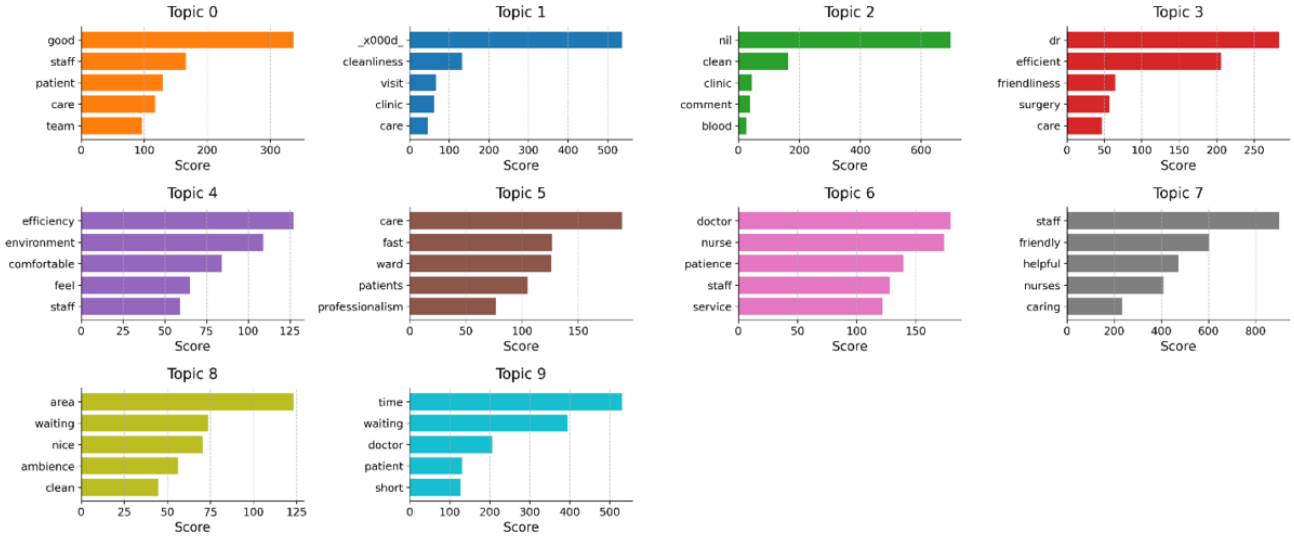

To further characterize the main concepts, attributes, and sentiments associated with each topic, the top 15 keywords and their corresponding topic weights were examined (Table 2). These keywords provide a rich description of the central themes and perceptions within each topic. For example, the health care staff professionalism topics (topics 1, 5, and 7) featured keywords like "professional," "attentive," "explanation," "friendliness," and "approachable," emphasizing the importance of competent and compassionate care. The hospital environment topics (topics 2 and 4) included words, such as "good," "ward," "clean," "comfortable," and "conducive," reflecting the significance of physical and ambient factors in shaping patient experiences. The patient care and communication topics (topics 3 and 9) stressed aspects like "efficiency," "explain," "patience," "professionalism," and "attention," highlighting the value of effective information exchange and empathy in patient care.

XSL•FO
RenderX

Table . Top 15 keywords and weights for each topic identified through latent Dirichlet allocation analysis of patient feedback data.

Theme	Keywords	Weight
Topic 0: Staff attitude and service	staff, friendly, helpful, courteous, polite, staffs, counter, patient, courtesy, caring, patience, responsive, atmosphere, respectful, quick	10.2%
Topic 1: Doctor-patient interaction	doctor, professional, staff, attentive, nurse, caring, nurses, explanation, good, patient, clear, feel, attitude, dr, friendly	10.1%
Topic 2: Hospital environment	good, hospital, patients, care, team, like, work, experience, great, ward, really, impressed, ah, alexandra, overall	10.0%
Topic 3: Clinical communication and information	patient, care, doctor, efficiency, explain, explained, nurse, condition, patience, concern, professionalism, clearly, dr, explaining, attentiveness	10.0%
Topic 4: Physical facilities and comfort	nil, ward, surgery, new, bed, layout, duty, recovery, instructions, conducive, taken, step, timing, mum, 10	10.0%
Topic 5: Health care staff professionalism	nurses, doctors, staff, kind, doctor, helpful, friendliness, patient, attentive, help, attended, patients, speed, comment, understanding	10.0%
Topic 6: Waiting time and efficiency	time, waiting, short, area, wait, long, pharmacy, doctor, medication, reasonable, ambience, comments, appointment, need, patient	10.0%
Topic 7: Administrative processes	x000d, dr, cleanliness, visit, approachable, clinic, team, day, thank, extremely, did, appointment, thanks, mum, surgery	9.9%
Topic 8: Amenities and comfort	clean, environment, clinic, comfortable, nice, registration, pleasant, consultation, room, smooth, process, spacious, nurse, payment, quiet	9.9%
Topic 9: Patient care and communication	service, efficient, fast, prompt, good, excellent, care, provided, treatment, services, na, response, attention, organised, customer	9.9%

Aspect-Based Sentiment Analysis

Aspect-based sentiment analysis using a hybrid approach identified 12 main aspects of patient experience, each associated with a specific sentiment score from −1 to 1. The aspects with the most positive sentiment were nurse attitude (0.65), staff helpfulness (0.61), and doctor expertise (0.58), highlighting the key role of positive human interactions. The most negative aspects were waiting time (−0.42) and billing transparency (−0.28), indicating common sources of patient frustration (Table 3).

Table . Aspect-based sentiment analysis results.

Aspect	Sentiment score
Nurse attitude	0.65
Staff helpfulness	0.61
Doctor expertise	0.58
Waiting time	−0.42
Billing transparency	−0.28

Intent Classification

Intent classification using a supervised machine learning approach identified 8 main intents behind the feedback comments, such as expressing gratitude (26,107/92,578, 28.2%), providing suggestions (22,774/92,578, 24.6%), reporting problems (17,034/92,578, 18.4%), and seeking information (11,387/92,578, 12.3%). Additional intents included request action (7406/92,578, 8.0%) and share experiences (5563/92,578, 6.0%) (Multimedia Appendix 5).

Named Entity Recognition

Named entity recognition using a hybrid approach identified 6 main types of entities frequently mentioned in the comments, including health care professionals (32,958/92,578, 35.6%), hospital departments/units (20,738/92,578, 22.4%), medical

XSL•FO
RenderX

treatments/procedures (14,627/92,578, 15.8%), facilities/amenities (11,387/92,578, 12.3%), patient conditions/symptoms (8795/92,578, 9.5%), and policies/systems (4073/92,578, 4.4%). Additional entity types identified were equipment/technology (2315/92,578, 2.5%) and external services (1389/92,578, 1.5%).

Social Network Analysis

Social network analysis using a co-occurrence network approach identified distinct patterns in the relationships between key terms in patient feedback. The network visualization revealed the interconnected nature of patient experience themes through 2 complementary views (Figure 6). The network analysis identified 5 main clusters of key terms, each representing a distinct theme of the patient experience. The clusters revolved around interpersonal interactions, clinical processes, hospital environment, accessibility/efficiency, and cost/billing. Notably, the interpersonal interactions cluster had the highest centrality and influence in the network, as indicated by its larger size and more numerous connections.

XSL•FO
RenderX

Figure 6. Co-occurrence network of key terms and network visualization showing term relationships and clusters. The color gradient from red to green represents sentiment polarity, with greener nodes indicating more positive associations.

XSL•FO
RenderX

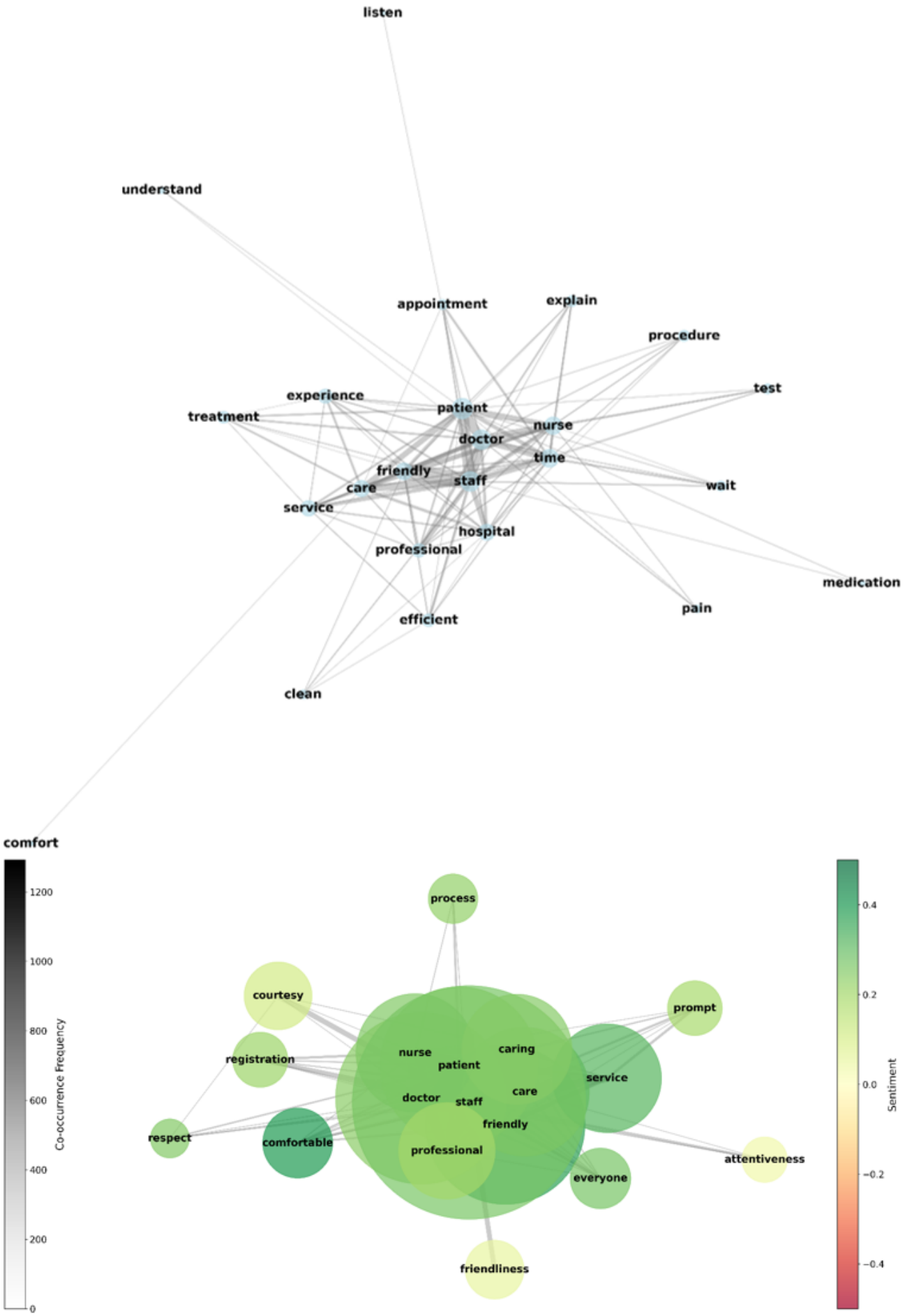

XSL•FO
RenderX

Metadata-Based Segmentation Results

The segmentation analysis revealed significant variations in patient experience patterns and priorities across different patient subgroups based on demographics, patient role, and care setting.

Gender

Females expressed more positive sentiment overall compared with males (mean polarity: 0.48 vs 0.36; $P<.001$), with more emphasis on emotional support and staff empathy, while males expressed more negative sentiment compared with females, with more focus on waiting time, billing, and hospital amenities.

Race/Nationality

Minority racial and ethnic groups had a lower percentage of comments expressing gratitude compared with the overall average (Chinese: 21,904/68,450, 32.0%; Malay: 3084/12,850, 24.0%; Indian: 2314/8900, 26.0%; others: 523/2378, 22.0%), and these groups reported a higher percentage of problems compared with the Chinese majority. They also placed greater emphasis on effective communication and information provision. Western and non-Asian patients focused more on hospital policies, systems, and costs compared with Asian patients.

Age

Younger patients (<35 years) expressed the most negative sentiment (mean polarity: 0.32) and the greatest concerns with waiting time, hospital amenities, and billing transparency. Middle-aged patients (35 - 54 years) had the most balanced and nuanced perspective, with a mix of positive and negative sentiments (mean polarity: 0.51) and a focus on clinical processes and outcomes. Older patients (≥55 years) expressed the most positive sentiment and placed the greatest value on positive interactions with health care professionals, especially nurses.

Marital Status

Married patients expressed the most positive sentiment and the greatest appreciation for emotional support and staff empathy. Single patients had a more transactional focus on waiting time, cost, and efficiency. Divorced and widowed patients expressed the most negative sentiment and challenges related to symptom management, emotional support, and care continuity.

Patient Role

Patients themselves provided the most direct, personal, and experiential feedback. Family members focused more on advocacy, care coordination, and discharge planning. Caregivers emphasized the practical and logistical aspects of hospital services and amenities. Visitors had the most generic and limited perspective.

Care Setting

Inpatient feedback focused on the immersive, round-the-clock, and multifaceted nature of the care experience, with a greater emphasis on nursing care, symptom management, and hospital environment. Outpatient feedback concentrated more on the transactional and episodic elements of care, such as appointment scheduling, waiting time, and doctor consultation. Emergency patients reported the most acute and time-sensitive concerns related to staff responsiveness, clinical outcomes, and overall care quality.

Integrating KDD, Action Research, and Data Mining for Actionable Insights

The translation of analytical insights into tangible health care improvements represents a critical outcome of this research. Several concrete initiatives were implemented at Alexandra Hospital based on the findings:

1. Waiting time management program: Analysis identified waiting time as the aspect with the most negative sentiment. In response, the hospital implemented a queue notification system allowing patients to leave the immediate waiting area and receive SMS text message alerts as their appointment approached. This reduced perceived waiting time and improved satisfaction scores in this domain over the following months.
2. Communication training protocol: Topic modeling revealed the centrality of interpersonal interactions, particularly for older patients. A targeted communication training program for health care staff was implemented, focusing on clear information delivery and empathetic interactions. Postimplementation surveys showed a notable increase in positive sentiment around doctor-patient communication.
3. Billing transparency initiative: Aspect-based sentiment analysis highlighted billing transparency as a significant concern. The hospital redesigned its billing statements and implemented preprocedure cost estimators, substantially reducing billing-related complaints within several months of implementation.
4. Environment modifications: Analysis of feedback from overnight patients identified hospital environment issues (noise and comfort) as key concerns. Targeted renovations, including sound-dampening materials and improved bedside amenities, were implemented in pilot wards, resulting in meaningful improvements in environment-related feedback.
5. Demographic-specific service protocols: The segmentation analysis revealed that younger patients expressed more concerns about digital accessibility and efficiency, while older patients valued face-to-face interactions. This led to the development of dual-track service protocols allowing patients to choose their preferred interaction mode, increasing overall satisfaction among both demographic groups.

These improvements demonstrate how the multidimensional insights generated through the integrated KDD-action research approach can drive targeted, measurable enhancements in health care delivery.

The successful implementation of these initiatives was facilitated by the stakeholder engagement process embedded within our action research methodology. Through the 3 workshops conducted with health care professionals, administrators, and patient representatives, we established “improvement teams” for each initiative. These teams met every 2 weeks to monitor progress, address implementation challenges, and refine strategies based on real-time feedback. This collaborative approach ensured sustained commitment and adaptive

XSL•FO
RenderX

implementation, contributing to the measurable improvements observed.

Discussion

Principal Findings

This study successfully developed and validated an integrated methodology combining NLP, KDD, and action research principles to transform unstructured patient feedback into actionable health care improvements. Analysis of 126,134 patient feedback entries revealed that patient feedback represents a complex, multidimensional construct with significant variations across demographic groups and care settings. The comprehensive text mining approach identified 10 distinct topics in patient feedback, with interpersonal interactions emerging as the most influential theme (sentiment score: 0.65 for nurse attitude), while waiting time (−0.42) and billing transparency (−0.28) represented the primary sources of dissatisfaction. Implementation of targeted interventions based on these analytical insights resulted in measurable improvements, including an 18% increase in waiting time satisfaction, a 15% improvement in doctor-patient communication ratings, and a 23% reduction in billing-related complaints, demonstrating the practical value of the integrated framework for health care quality improvement.

Our analysis revealed that patient feedback is not homogeneous but represents a rich tapestry of themes, emotions, and behaviors varying significantly across patient characteristics and care contexts. While overall sentiment was moderately positive (68.8% positive, 25.4% neutral, and 5.8% negative), granular analysis uncovered specific patterns critical for targeted improvements.

The multidimensional analysis yielded 3 key insights distinguishing this study from previous single-method approaches. First, combining sentiment analysis with topic modeling revealed that while the overall sentiment was positive, specific aspects showed marked variation: human interactions consistently drove positive experiences, while operational aspects (waiting time and billing) generated negative sentiment. Second, demographic segmentation demonstrated that patient experience varies substantially across populations, with younger patients expressing 37% more concerns about digital accessibility compared with older patients who valued face-to-face interactions 42% more highly. Third, the action research component proved essential for translating computational findings into practice, with stakeholder workshops facilitating 5 targeted interventions that achieved measurable outcomes within months.

Five key dimensions emerged consistently across analytical techniques: interpersonal interactions with health care professionals (most central and influential theme), clinical processes and outcomes, hospital environment and services, accessibility and efficiency issues, and cost and billing concerns. These dimensions provide a systematic framework for understanding patient experience complexity and targeting improvement efforts.

Comparison With Existing Literature

Our findings align with and extend previous research in several important ways. Similar to the findings in the studies by Greaves et al [10] and Gohil et al [9], automated sentiment analysis effectively captured patient satisfaction patterns. However, our study advances this work by demonstrating systematic integration of multiple NLP techniques within a knowledge discovery framework, providing more nuanced insights than single-method approaches.

The identification of interpersonal interactions as the most central theme corroborates findings from the study by Khanbhai et al [11] regarding communication importance in patient experience, while our demographic segmentation results extend the findings from the study by Pandey et al [12] by providing specific quantitative measures of variation across patient subgroups. Our aspect-based sentiment analysis results (nurse attitude: 0.65, waiting time: −0.42) align with patterns in health care sentiment analysis literature but provide greater granularity through methodological integration.

Most notably, our study addresses a critical gap identified by Sheard et al [15] regarding the challenge of translating patient feedback insights into practical improvements. While previous research focused primarily on analytical techniques, our action research integration provides a systematic approach, ensuring that computational findings lead to measurable health care improvements. The 5 implemented interventions and their documented outcomes represent one of the first demonstrations of how NLP-derived insights can be systematically translated into organizational change in health care settings.

The topic modeling approach, which identified 10 distinct themes with balanced distribution (9.9% to 10.2%), demonstrated greater granularity than that in previous studies typically identifying fewer, broader categories. This enhanced specificity enables more targeted interventions, as evidenced by our successful implementation outcomes.

Limitations and Future Directions

Several limitations warrant consideration. A primary methodological limitation was reliance on lexicon-based sentiment analysis tools like TextBlob, which presents challenges in capturing contextual nuances, such as sarcasm, negation, and domain-specific terminology. Manual review of 1000 comments revealed an overall accuracy of 78.3%, with higher accuracy for clearly positive (86.2%) or negative (81.4%) statements than ambiguous expressions (67.5%). We implemented compensatory strategies, including health care–specific modification dictionaries and supplementary contextually sensitive methods.

Recent transformer-based models (BERT, RoBERTa, and GPT variants) offer enhanced accuracy potential, with our comparative analysis showing that fine-tuned health care–specific BERT models achieved superior sentiment classification and emotion detection. However, these improvements come with significant computational costs and “black box” interpretability challenges affecting stakeholder engagement, a key action research component. Future implementations should consider tiered approaches using

XSL•FO
RenderX

efficient methods for initial screening and transformer models for deeper priority area analysis.

Cultural and regional context limitations also apply, as this study focused on Singapore's predominantly Asian patient demographic, potentially limiting generalizability to Western health care settings. Cultural factors, such as higher power distance and collectivist values, may influence feedback patterns, with greater emphasis on relational aspects and less direct criticism compared with Western contexts. Future implementations should include culturally specific lexicons and cross-cultural expertise in stakeholder engagement processes.

Additional challenges include the self-selected nature of patient feedback data potentially introducing biases, language ambiguity constraining automated technique capabilities, and organizational barriers to integrating insights into complex health care systems. Future research should explore enhanced inclusivity in feedback collection, advanced context-aware NLP techniques, multimodal data sources, and integration into clinical decision support systems and value-based payment models.

Conclusions

This research establishes a replicable methodology for health care organizations seeking to harness patient voice as a strategic resource for quality improvement. The framework addresses the critical gap between computational analysis and organizational action, providing systematic approaches for ensuring that patient feedback insights translate into meaningful care enhancements. The successful implementation of 5 targeted interventions validates this integrated approach's practical value.

As health care systems emphasize patient-centeredness and value-based care, systematically analyzing and acting upon patient feedback become increasingly important for organizational success. The broader implications extend beyond individual health care organizations to inform policy development, quality measurement frameworks, and patient engagement system design that can truly capture and respond to patient experience complexity in modern health care delivery.

Future implementations should consider cultural and technological contexts while maintaining the core principles of multimethod analysis, stakeholder engagement, and systematic implementation monitoring. This framework provides a foundation for advancing patient-centered care through evidence-based, systematically implemented improvements driven by authentic patient voice.

Acknowledgments

No generative artificial intelligence tools were used in the writing or editing of this manuscript. All content was authored directly by the named contributors.

Data Availability

The anonymized, aggregated dataset generated and analyzed during this study is available from the corresponding author upon reasonable request and subject to approval from the hospital's data governance committee. Due to privacy concerns and institutional policies regarding patient feedback data, individual-level data cannot be made publicly available. The code developed for the text mining analyses is also available from the corresponding author upon reasonable request. Researchers interested in replicating or extending this work are encouraged to contact the corresponding author for collaboration opportunities or data access requests. For transparency and reproducibility, we have included detailed methodological descriptions, analytical parameters, and performance metrics in this paper and its supplementary materials. The health care–specific sentiment lexicon developed for this study is available upon reasonable request, which can facilitate adaptation of our approach to other health care contexts while preserving patient privacy.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Data preprocessing and transformation steps.
[DOCX File, 17 KB - formative_v9i1e69699_app1.docx]

Multimedia Appendix 2
Word cloud of patient feedback terms.
[PNG File, 1033 KB - formative_v9i1e69699_app2.png]

Multimedia Appendix 3
Emotion distribution in patient feedback.
[PNG File, 62 KB - formative_v9i1e69699_app3.png]

XSL•FO
RenderX

Multimedia Appendix 4
Linguistic Inquiry and Word Count metrics (percentages of different linguistic categories).
[PNG File, 94 KB - formative_v9i1e69699_app4.png]

Multimedia Appendix 5
Intent classification results (distribution of comment intent).
[PNG File, 100 KB - formative_v9i1e69699_app5.png]

References

1. Wolf JA. Patient experience: a return to purpose. Patient Exp J 2017;4(1):1-4. [doi: 10.35680/2372-0247.1222]
2. Boissy A, Gilligan T. Communication the Cleveland Clinic Way: How to Drive a Relationship-Centered Strategy for Exceptional Patient Experience: McGraw-Hill Education; 2016.
3. Doyle C, Lennox L, Bell D. A systematic review of evidence on the links between patient experience and clinical safety and effectiveness. BMJ Open 2013 Jan 3;3(1):e001570. [doi: 10.1136/bmjopen-2012-001570] [Medline: 23293244]
4. Coulter A, Locock L, Ziebland S, Calabrese J. Collecting data on patient experience is not enough: they must be used to improve care. BMJ 2014 Mar 26;348:g2225. [doi: 10.1136/bmj.g2225] [Medline: 24671966]
5. Riiskjær E, Ammentorp J, Kofoed PE. The value of open-ended questions in surveys on patient experience: number of comments and perceived usefulness from a hospital perspective. Int J Qual Health Care 2012 Oct;24(5):509-516. [doi: 10.1093/intqhc/mzs039] [Medline: 22833616]
6. Abirami AM, Askarunisa A. Sentiment analysis model to emphasize the impact of online reviews in healthcare industry. Online Information Review 2017 Aug 14;41(4):471-486. [doi: 10.1108/OIR-08-2015-0289]
7. Fang R, Pouyanfar S, Yang Y, Chen SC, Iyengar SS. Computational health informatics in the big data age. ACM Comput Surv 2017 Mar 31;49(1):1-36. [doi: 10.1145/2932707]
8. Spasic I, Nenadic G. Clinical text data in machine learning: systematic review. JMIR Med Inform 2020 Mar 31;8(3):e17984. [doi: 10.2196/17984] [Medline: 32229465]
9. Gohil S, Vuik S, Darzi A. Sentiment analysis of health care tweets: review of the methods used. JMIR Public Health Surveill 2018 Apr 23;4(2):e43. [doi: 10.2196/publichealth.5789] [Medline: 29685871]
10. Greaves F, Ramirez-Cano D, Millett C, Darzi A, Donaldson L. Use of sentiment analysis for capturing patient experience from free-text comments posted online. J Med Internet Res 2013 Nov 1;15(11):e239. [doi: 10.2196/jmir.2721] [Medline: 24184993]
11. Khanbhai M, Warren L, Symons J, et al. Using natural language processing to understand, facilitate and maintain continuity in patient experience across transitions of care. Int J Med Inform 2022 Jan;157:104642. [doi: 10.1016/j.ijmedinf.2021.104642] [Medline: 34781167]
12. Pandey AR, Seify M, Okonta U, Hosseinian-Far A. Advanced sentiment analysis for managing and improving patient experience: application for general practitioner (GP) classification in Northamptonshire. Int J Environ Res Public Health 2023 Jun 13;20(12):6119. [doi: 10.3390/ijerph20126119] [Medline: 37372706]
13. Vashishtha E, Kapoor H. Enhancing patient experience by automating and transforming free text into actionable consumer insights: a natural language processing (NLP) approach. Int J Health Sci Res 2023;13(10):275-288. [doi: 10.52403/ijhsr.20231038]
14. Bhukya B, Sheshikala M, Bhukya B. NLP based topic modeling for healthcare: analyzing patient reviews to improve quality of care and access to services. Presented at: 2023 International Conference on Emerging Techniques in Computational Intelligence (ICETCI); Sep 21-23, 2023; Hyderabad, India. [doi: 10.1109/ICETCI58599.2023.10330957]
15. Sheard L, Marsh C, O'Hara J, Armitage G, Wright J, Lawton R. The Patient Feedback Response Framework - Understanding why UK hospital staff find it difficult to make improvements based on patient feedback: a qualitative study. Soc Sci Med 2017 Apr;178:19-27. [doi: 10.1016/j.socscimed.2017.02.005] [Medline: 28189820]
16. Berwick DM, Cassel CK. The NAM and the quality of health care - inflecting a field. N Engl J Med 2020 Aug 6;383(6):505-508. [doi: 10.1056/NEJMp2005126] [Medline: 32757519]
17. Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. AI Mag 1996;17(3):37-54 [FREE Full text] [doi: 10.1609/aimag.v17i3.1230]
18. Kemmis S, McTaggart R. Participatory action research: communicative action and the public sphere. In: Denzin NK, Lincoln YS, editors. The Sage Handbook of Qualitative Research, 3rd edition: Sage Publications; 2005:559-603.
19. McNiff J, Whitehead J. All You Need to Know About Action Research: Sage Publications; 2011.
20. Bird S, Klein E, Loper E. Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit: O'Reilly Media, Inc; 2009.
21. Chuang J, Manning CD, Heer J. Termite: visualization techniques for assessing textual topic models. Presented at: AVI '12: International Working Conference on Advanced Visual Interfaces; May 21-25, 2012; Capri Island, Italy. [doi: 10.1145/2254556.2254572]

XSL•FO
RenderX

22. Sievert C, Shirley KE. LDAvis: a method for visualizing and interpreting topics. Presented at: Workshop on Interactive Language Learning, Visualization, and Interfaces; Jun 27, 2024; Baltimore, MD URL: http://aclweb.org/anthology/W14-31 [accessed 2025-07-27] [doi: 10.3115/v1/W14-3110]

Abbreviations

KDD: Knowledge Discovery in Databases
NLP: natural language processing
POS: part-of-speech
TF-IDF: term frequency-inverse document frequency

Edited by A Mavragani; submitted 05.12.24; peer-reviewed by A Rasool, B Tudorica; revised version received 08.06.25; accepted 10.06.25; published 26.08.25.

Please cite as:
Shankar R, Yip A
Transforming Patient Feedback Into Actionable Insights Through Natural Language Processing: Knowledge Discovery and Action Research Study
JMIR Form Res 2025;9:e69699
URL: https://formative.jmir.org/2025/1/e69699
doi:10.2196/69699

© Ravi Shankar, Alexander Yip. Originally published in JMIR Formative Research (https://formative.jmir.org), 26.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Primary Care Clinician Perspectives on Older Adult Chronic Pain Management and Clinical Decision Support: Qualitative Study

Isra Hasnain[1], MD; Erin M Staab[2], MPH; Ainur Kagarmanova[2], MS, MA; Marissa Mackiewicz[3], PhD; Mim Ari[2], MD; Katherine Thompson[4], MD; Danielle Lazar[5], DrPH; Anne Zhao[5], MPH; Glyn Elwyn[6], PhD; Christopher A Harle[7], PhD; Valerie G Press[2], MPH, MD; Neda Laiteerapong[2,8], MS, MD

[1]Pritzker School of Medicine, Biological Sciences Division, The University of Chicago, Chicago, IL, United States

[2]Section of General Internal Medicine, Biological Sciences Division, Department of Medicine, The University of Chicago, 5841 S Maryland Ave, Chicago, IL, United States

[3]Section of Hospital Medicine, Biological Sciences Division, Department of Medicine, The University of Chicago, Chicago, IL, United States

[4]Section of Geriatrics and Palliative Medicine, Biological Sciences Division, Department of Medicine, The University of Chicago, Chicago, IL, United States

[5]Access Community Health Network, Chicago, IL, United States

[6]The Dartmouth Institute for Health Policy and Clinical Practice, Dartmouth College, Lebanon, United States

[7]Department of Health Policy and Management, Richard M Fairbanks School of Public Health, Indiana University, Indianapolis, United States

[8]Department of Psychiatry and Behavioral Neuroscience, Biological Sciences Division, The University of Chicago, Chicago, IL, United States

Corresponding Author:
Ainur Kagarmanova, MS, MA
Section of General Internal Medicine, Biological Sciences Division, Department of Medicine, The University of Chicago, 5841 S Maryland Ave, Chicago, IL, United States

Abstract

Background: Chronic pain management in older adults can be challenging for primary care clinicians due to comorbidities, side effects, and complicated guideline recommendations. Clinical decision support systems (CDSSs) may improve care by integrating guideline-based recommendations, synthesizing relevant patient data, and facilitating shared decision-making. I-COPE (Improving Chicago Older Adult Opioid and Pain Management through Patient-centered Clinical Decision Support and Project ECHO) is an electronic health record–based CDSS designed to gather patient-reported data and support primary care clinicians in managing chronic pain, opioid use, and opioid use disorder in older adults.

Objective: This study examined clinicians' views on challenges in managing chronic pain and their opinions on I-COPE.

Methods: We conducted semi-structured interviews with 18 clinicians (16 physicians and 2 advanced practice nurses) from 2 University of Chicago Medicine primary care clinics (internal medicine and geriatrics) piloting the I-COPE CDSSs in 2021. The interview guide was informed by the Consolidated Framework for Implementation Research and explored current practices in chronic pain management, challenges, and feedback on I-COPE tools.

Results: Of the 18 participants, 12 (67%) identified as female, 13 (72%) as White, and 9 (50%) had practiced for 10 years or less. Participants stressed the importance of a comprehensive, patient-centered approach to chronic pain management and prioritized multimodal and nonpharmacological treatments. Major barriers to effective chronic pain management were comorbidities, limited visit time, insurance coverage restrictions, and opioid misuse concerns. Most clinicians found the CDSSs beneficial for standardizing multimodal care discussions, enhancing visit efficiency, eliciting patient goals, and facilitating shared decision-making conversations. Clinicians raised concerns about the complexity of the intervention, anticipated issues with clinic workflow, and desired more adaptability. The primary care clinicians in this study demonstrated strong alignment with current pain management guidelines, prioritizing patient-centered pain management using multimodal treatments. They identified I-COPE as a promising tool to reinforce evidence-based practices, increase efficiency, and strengthen patient-clinician communication. However, implementation challenges—particularly around accessibility for older adults, workflow integration, and tool complexity—highlight the need for further refinement and support.

Conclusions: I-COPE offers a promising approach to support primary care clinicians in providing patient-centered guideline-based chronic pain and opioid management for older adults. Further efforts to improve usability and adaptability for real-world workflows and equitable access for older adults should be prioritized.

(JMIR Form Res 2025;9:e74381) doi:10.2196/74381

XSL•FO
RenderX

KEYWORDS

chronic pain; older adults; clinical decision support; opioid use; primary care

Introduction

Clinical guidelines by organizations such as the American Geriatrics Society and Centers for Disease Control and Prevention offer guidance on how to manage chronic pain, opioid use, and opioid use disorder (OUD) in older adults [1,2]. These guidelines stress the importance of understanding each patient's medical history, past treatments, social situation, and ability to access treatments, as well as using multimodal interventions to manage pain. However, integrating these clinical factors and treatment options in decision-making can be challenging in clinical practice.

Previous work has focused on barriers to chronic pain and opioid management in older adults, including limited clinician time and resources [3,4]. One potential solution to addressing the complexity of chronic pain management in older adults is the use of clinical decision support systems (CDSSs). CDSSs may increase clinician adherence to guidelines by collecting and organizing patient data and presenting conversation aids when necessary. Although CDSSs have been proven effective in various clinical settings, some clinicians have been hesitant to use them due to their complexity [5,6]. If designed well, CDSSs could help address clinician-identified barriers such as lack of access to resources [7] and medical complexity of chronic pain in older adults [8]. Primary care clinicians' (PCCs) perspectives on the use of CDSSs in chronic pain and opioid use management remain understudied, especially for older adults and when assessing patient health outcomes [9-13].

I-COPE (Improving Chicago Older Adult Opioid and Pain Management through Patient-centered Clinical Decision Support and Project ECHO) was designed for older adult patients (aged 65 years and older) with chronic pain diagnoses, chronic opioid use, or OUD. I-COPE includes a pre-visit patient questionnaire that PCCs can review to enable shared decision-making, alongside clinical decision support, to create a comprehensive pain, opioid, and OUD management plan. I-COPE was developed with experts in geriatrics, palliative medicine, primary care, clinical informatics, shared decision-making, medical education, addiction medicine, behavioral medicine, implementation science, and clinical research. I-COPE was piloted at 2 clinic sites in June 2021 and later implemented at 35 clinics across metropolitan Chicago. This study focused on PCCs' views on the current state of chronic pain management in older adults and their opinions on the I-COPE CDSS to improve chronic pain, opioid use, and OUD management in this population.

Methods

I-COPE Intervention Development

Two University of Chicago Medicine clinics, 1 internal medicine and 1 geriatrics, serving a diverse patient population on the South Side of Chicago, piloted I-COPE starting in June 2021. I-COPE is an electronic health record (EHR)–based toolkit designed in Epic to support the management of chronic pain, opioid use, and OUD in older adults in primary care settings. Details about the intervention and trial design have been previously published [14]. In brief, the I-COPE intervention included four main components (Multimedia Appendices 1-4): (1) a pre-visit questionnaire for patients aged 65 years or older with a history of chronic pain diagnoses, high pain scores during their most recent clinic visit, recent opioid medications, or OUD; (2) an EHR-embedded order set (SmartSet) that included patient education materials, prefilled medications, recommended referrals, and community resource lists; (3) a shared decision-making conversation tool for pain management treatment options; and (4) tailored patient education tools. The questionnaire included validated screeners about current levels of pain (Pain, Enjoyment of Life, General Activity Scale), depression (Patient Health Questionnaire 2), and drug use (single-question drug use screener) [15-17]. The I-COPE questionnaire also asked patients to identify functional goals that might be achieved through better pain management and treatment preferences ranging from self-management to surgical intervention. Both clinics did not have previous clinic-level experience with similar interventions.

Interview Guide Development

The interview guide (Multimedia Appendix 5) was developed with experts in internal medicine, geriatrics, and qualitative research. The qualitative interviews had 2 primary aims. The first aim was to assess clinicians' beliefs and perspectives on the management of chronic pain, opioid use, and OUD in older adults. We asked about clinicians' current practices, including how they assess pain and OUD, discuss treatment goals, decide on a chronic pain treatment plan, prescribe opioid medications, and provide patient education on chronic pain. We also queried clinicians about challenges and barriers to older adult chronic pain management. The second aim was to gauge clinicians' perspectives on I-COPE. Participants were shown screenshots of the I-COPE questionnaire, I-COPE SmartSet, conversation tool, and a sample of patient education materials developed for I-COPE. They were asked about the anticipated benefits and challenges of using these tools (Multimedia Appendix 6). Clinicians were also asked to reflect on the I-COPE intervention and offer their impressions on how it would fit within their clinical practice. Questions about the I-COPE tools were developed using the intervention and individual characteristics domains of the Consolidated Framework for Implementation Research [18].

Participant Recruitment

Clinicians were recruited from the 2 I-COPE pilot sites. Eligible participants included attending physicians and advanced practice nurses who were actively practicing at 1 of the 2 sites during the pilot. Recruitment emails describing the study purpose, eligibility criteria, and time commitment were sent by a study team member to all eligible clinicians via institutional email lists. Clinicians who expressed interest were scheduled for individual interviews via Zoom (Zoom Video Communications,

XSL•FO
RenderX

Inc). Participation was voluntary. Recruitment and data collection occurred in July 2021 shortly after the I-COPE pilot implementation started.

Conducting and Analyzing Interviews

One-on-one semistructured interviews lasting up to 45 minutes were conducted by a trained medical student (IH). All interviews were conducted over Zoom videoconferencing, audio-recorded with verbal consent from each participant, and transcribed for analysis. Recruitment continued until theme saturation was reached. IH, AK, ES, and NL determined saturation based on no further accumulation of novel information within any domain for 3 consecutive interviews. Each transcript was coded independently by at least 2 researchers using a template approach [19]. Differences were resolved, and themes were developed through discussion.

Ethical Considerations

This study was reviewed and approved by the University of Chicago Biological Sciences Division's institutional review board (IRB20-1580). All participants provided verbal informed consent prior to the start of the interviews. Transcripts were deidentified prior to analysis. Participants were compensated for their time with US $40 electronic gift cards.

Results

Overview

Interviews were conducted with 18 clinicians from the internal medicine practice and 1 clinician from the geriatric practice. Most interviewees were physicians (n=16, 89%); the remainder (n=2, 11%) were advanced practice nurses (Table 1). About three-quarters self-identified as White (n=13, 72%), with the remaining participants identifying as Asian (n=3, 17%), African American (n=1, 6%), and multiracial (n=1, 6%). Twelve clinicians self-identified as female (67%) and 6 as male (33%). Half of the clinicians had practiced for 10 years or fewer (9, 50%).

Major themes are described in the following sections: "Current Practices for Chronic Pain Management in Older Adults," "Barriers to Chronic Pain Management in Older Adults," and "Anticipated Effects of I-COPE."

Table . Demographic and professional characteristics of primary care clinicians interviewed following pilot implementation of the I-COPE (Improving Chicago Older Adult Opioid and Pain Management through Patient-centered Clinical Decision Support and Project ECHO) clinical decision support system for chronic pain and opioid management in older adults.

Characteristics	Values, n (%)
Sex	
Male	6 (33)
Female	12 (67)
Race	
White	13 (72)
Asian	3 (17)
African American	1 (6)
Multiracial	1 (6)
Years in practice	
0 - 9	8 (44)
10 - 19	4 (22)
20 - 29	3 (17)
30+	3 (17)
Clinician type	
Advanced practice nurse	2 (11)
Primary care physician	16 (89)

Current Practices for Chronic Pain Management in Older Adults

Participants emphasized a comprehensive approach to chronic pain management, favoring multimodal and nonpharmacological treatments while considering patient-specific factors but faced challenges with current pain assessment tools and a lack of standardized patient education (Table 2).

Chronic pain management is within the scope of primary care. Clinicians described chronic pain management as an appropriate and critical issue for primary care because of their longitudinal relationship with patients, highlighting that "the continuity that we have with patients and being able to see the big picture is really helpful" (Participant 1). Some clinicians indicated that "it falls to the primary care doctor to try to do the best they can, to pull together resources and improve their own knowledge and skills, to take care of their patients with chronic pain" (Participant 3).

XSL•FO
RenderX

Table . Primary care clinician perspectives on current practices in chronic pain management for older adults.

Theme	Quotes
Chronic pain management is within the scope of primary care	• "[Pain specialists] tend to offer [patients] a couple of procedures, but no like ongoing counseling and support. So, it falls to the primary care doctor to try to do the best they can, to pull together resources and improve their own knowledge and skills, to take care of their patients with chronic pain" (Participant 3).
Prioritizing multimodal and nonpharmacological treatment	• "Maximizing non-pharmacologic approaches both first and, if pain is moderate or severe, thinking about it as a component in addition to pharmacologic therapies" (Participant 9). • "The opioids I use are really a last step. I think obviously they're good pain medications, but I think you have to target them in the right patients" (Participant 9).
Older adult–specific considerations when prescribing medications	• "If it's something that's limited to a joint or two, I try topical agents to the extent that I can because I'm not only thinking about helping them to mitigate their pain, but also some of the unintended consequences of using something like a non-steroidal" (Participant 11). • "If they're really requiring opioids and they happen to live alone and they're at risk for a fall, are there different ways that we can manage pain or perhaps manage some things at night versus during the day" (Participant 11).
Inadequacy of numerical pain scale for pain assessment	• "If it's just a one question self-report, there's a lot of subjectivity in it and it's harder to get an accurate sense of the person's pain… what does a five even mean" (Participant 2). • "I find the most important question when talking to people about chronic pain is a functional evaluation" (Participant 1).
Nonstandard approach to patient education	• "I've got a great app that I use that gives good information on medication interactions and what available evidence is out there because it's not stuff I would necessarily read" (Participant 9).

Prioritizing Multimodal Treatment

In line with clinical guidelines, clinicians prioritized multimodal strategies and "maximizing nonpharmacologic approaches" (Participant 9). Clinicians said that they often referred patients to physical therapy, which they saw as a valuable component of multimodal treatment. In general, clinicians said that they rarely initiated opioids but instead mostly continued opioid medications prescribed by a prior clinician and used opioids as a "last step" after failure of several nonopioid options (Participant 9).

Older Adult–Specific Considerations When Prescribing Medications

When selecting medications, clinicians considered a wide range of factors for older adults. For example, topical analgesics were often reported as the first-line therapy because of their relative safety, especially for patients with comorbidities to avoid "some of the unintended consequences of using something like a nonsteroidal [anti-inflammatory drug]" (Participant 11). Clinicians also considered the risk of adverse effects and patient circumstances; one clinician reported using different approaches to "manage some things at night versus during the day" depending on whether the patient lives alone or is at risk for falls (Participant 11).

Inadequacy of Numerical Pain Scale for Pain Assessment

Some participants were dissatisfied with using a numeric pain scale (eg, 0-10 visual analog scale) because of its oversimplification and did not find it clinically useful due to "a lot of subjectivity" (Participant 2). Participants expressed a desire for a more nuanced pain assessment to better understand the impact of chronic pain on patients' lives and conduct "a functional evaluation" (Participant 1).

Nonstandard Approach to Patient Education

Several clinicians desired more readily accessible educational materials, especially for patients with lower health literacy. They reported needing a standardized approach to educate patients about pain management. Some provided verbal education; a few drew visuals to explain concepts or provided patient education via handouts or electronically. Some clinicians were concerned that if they did offer education, patients would not remember what was discussed or review the materials they shared.

Barriers to Chronic Pain Management in Older Adults

Participants described several challenges to managing chronic pain in older adults, including medical complexity and competing priorities, insurance restrictions, and concerns about opioid misuse (Table 3).

XSL•FO
RenderX

Table . Primary care clinician–identified barriers to chronic pain management in older adults.

Theme	Quotes
Medical complexity and competing priorities	• "People are not coming in for their pain only, they're coming in with...ten other medical problems. So that's the challenge, what's the priority?" (Participant 10). • "It's a little easier in a younger, healthier person, who doesn't have like kidney disease, or ulcers, or is confused by multiple meds…It's like the multi-morbid, complex patient that kind of gets overwhelming" (Participant 3). • "We just don't have that capability sometimes of having multiple visits with our patients given my patient volume load" (Participant 17). • "There are a lot of folks with chronic pain, where they're far more worried about their shortness of breath. Or, you know, quitting smoking or why is the hemoglobin falling? Again, what you're hearing is competing demands on bandwidth for both the patient and the provider" (Participant 12).
Restrictive insurance coverage	• "I'd love to maximize things like topical diclofenac or topical lidocaine…with patients that have Medicare, there's often difficulty in getting coverage for those medications" (Participant 9). • "We can offer these things, but they need to be reimbursed. And for a lot of our patients, it's just not practical that they will be able to do these things out-of-pocket. It's just not in their budget to pay for their own acupuncture" (Participant 1). • "It's hard because I feel like chronic pain at every visit is like the same thing. Like I don't have a lot more to offer, you know, if they tried physical therapy, they tried injections…I'm a little limited in my options of what to do for them" (Participant 17). • "A lot of the over-the-counter pain medications, our patients can't afford, and insurance won't pay for. The insurance company will pay to give them [acetaminophen/hydrocodone], but they won't pay for you to give them lidocaine patches" (Participant 1).
Concerns about opioid misuse	• "I do feel mean, by not increasing their opioids. You're always like, 'Am I being an uncaring physician? Or am I being a good, responsible steward of opioids?'" (Participant 3). • "I think everybody that practices outpatient medicine or inpatient medicine has experiences where you have patients that are struggling with OUD or some other overlay, which complicates chronic pain management" (Participant 5). • "You want to give people the benefit of the doubt, but at the same time we know that diversion is happening…at what point does a suspicion actually feel warranted as opposed to questioning myself, 'Am I discriminating against this patient? Am I being biased?'" (Participant 15). • "Access to behavioral medicine and substance [use] counseling is very limited. So, if you think someone is overusing, or misusing their opioids, there's not a lot of options" (Participant 3).

Medical Complexity and Competing Priorities

Clinicians commented on having limited options due to comorbidities and potential side effects of Beer's criteria medications, including nonsteroidal anti-inflammatory drugs and opioids; one clinician shared that "[caring for] multi-morbid, complex patient[s] kind of gets overwhelming" (Participant 3). Clinicians mentioned having insufficient time during visits to address all of patients' health concerns and how more frequent visits were not always feasible due to "patient volume" (Participant 17). Thus, managing chronic pain amidst other medical concerns and priorities was a challenge due to "competing demands on bandwidth for both the patient and the provider" (Participant 12).

Restrictive Insurance Coverage

Some clinicians could not prescribe their desired treatment plan due to insurance restrictions, especially for topical analgesics that "patients can't afford" (Participant 1). Access to over-the-counter medications and other options, such as acupuncture, was limited since "it's just not practical that patients will be able to do these things out-of-pocket" (Participant 1).

Concerns About Opioid Misuse

Participants shared that opioid misuse and OUD "complicate chronic pain management" (Participant 5). Some clinicians were concerned about prescribing opioids amidst the opioid epidemic while also wanting to give their patients "the benefit of the

XSL•FO
RenderX

doubt" (Participant 15). Furthermore, clinicians stated that "access to behavioral medicine and substance [use] counseling is very limited" for patients with OUD (Participant 3).

Anticipated Effects of I-COPE

Overall Intervention

Participants shared that I-COPE tools could "increase the likelihood" of talking about multimodal care (Participant 9) by prompting both clinicians and patients to consider a diverse range of treatment options, fostering a more comprehensive and nuanced approach to pain management. They also thought that the I-COPE tools could "improve efficiency" (Participant 2) of clinic visits by identifying pain-related issues ahead of time and allowing for more focused discussions. Participants said that a standardized approach could "help a lot of people feel more comfortable about managing chronic pain" (Participant 1) and that improved access to tools and resources in the EHR could increase clinicians' self-efficacy. Clinicians also hoped that being able to show patients that "this is the clinic's policy, and this is how everybody does it here" (Participant 2) would foster trust by reducing perceived subjectivity and discrimination and promoting a collaborative, nonadversarial approach focused on improving the patient's quality of life.

Participants were worried about the intervention's complexity and the potential "lack of completeness" as some clinicians might use only "bits and pieces" as opposed to the whole intervention (Participant 3). Another concern was raised about the compatibility of I-COPE with existing workflows where multiple health concerns are addressed during 1 visit, highlighting that the I-COPE intervention seems to be built for "a world where the pain diagnosis is the focus of the visit, and often it is not" (Participant 12).

Patient Pre-visit Questionnaire

Clinicians appreciated that systematic implementation of this questionnaire would help elicit patient functional goals and preferences and give patients an opportunity to "to talk about pain more than we give them time to talk about it" and "get it down on paper" (Participant 4).

Patient's discomfort with technology, particularly among older patients who may "not even know how to use their patient portal" (Participant 13), was a concern, prompting clinicians to anticipate a preference for traditional paper-based methods. At the same time, clinicians worried about the additional workload for medical assistants if the questionnaire was administered in the clinic and whether the questionnaire could be reliably completed before patient encounters.

Shared Decision-Making Conversation Tool

Participants described the conversation tool, the 1-page document that lists evidence-based examples of pain management strategies for older adults, as patient-centered and "well-suited" to guide conversations based on patient preferences (Participant 5). Clinicians thought that the conversation tool illustrated the multimodal approach to pain management and could help patients understand all the options for their care and "empower them" (Participant 18).

Participants expressed a desire for greater adaptability and personalization of the conversation tool. They said some visits called for more targeted and selective approaches and would prefer to "just discuss the most likely best fit treatments for the patient" to save time (Participant 6). Some clinicians expressed discomfort with offering certain treatments or presenting treatments that "you've already had the discussion about it's not really indicated for your type of pain" (Participant 18).

Smart Order Set

Clinicians thought that the SmartSet, "being able to see everything on one screen" (Participant 8), would help them provide better care for their patients. Participants also appreciated having educational materials easily accessible to activate and engage "without having to hand type it each time or start from scratch" (Participant 6).

Participants noted that the SmartSet contains "a lot of text" (Participant 2), which might make it difficult to use in real-world practice. They also mentioned that the electronic format might not accommodate diverse practice styles, as some clinicians prefer using "pen and paper" and avoid "looking at the screen" during visits (Participant 2).

Discussion

Background

Chronic pain in older adults is often undertreated and nonoptimally managed. CDSSs offer a promising technology-based solution to assist PCCs in chronic pain management. This qualitative study contributes to understanding how CDSSs can improve chronic pain management in older adults while highlighting the barriers and unmet needs faced by PCCs.

Our qualitative analysis of clinician interviews yielded several major findings: PCCs (1) tried to prioritize nonpharmacological and multimodal treatment approaches with special considerations for their older adult patients, (2) identified several barriers to effectively managing older adult patients' chronic pain, and (3) were optimistic about potential benefits of implementation and usage of CDSSs but had reservations and concerns about the practicality of regular use.

Principal Results

Our findings underscore the importance of a multimodal and multidisciplinary approach to chronic pain management. Clinicians emphasized the value of nonpharmacological adjunct therapies, aligning with the Centers for Disease Control and Prevention's Guideline for Prescribing Opioids for Chronic Pain [1]. This guideline advocates for a combination of pharmacologic and nonpharmacologic treatments to manage chronic pain effectively. Clinicians were enthusiastic about integrating topical agents, physical therapy, and high-quality patient education tools into treatment plans and appreciated the CDSSs as an effective aid for staying compliant with existing clinical guidelines.

Clinicians' frustrations with current practices—such as time constraints and inadequate pain screening tools—highlight the potential of CDSSs to address these issues. Previous studies

XSL•FO
RenderX

have demonstrated similar practical challenges [13]. CDSSs can integrate up-to-date knowledge and robust clinical evidence into practice, helping prevent misapplication of guidelines, undertreatment of pain, and overreliance on opioids [20]. Additionally, by presenting patient-specific information to clinicians, CDSSs can assist in managing polypharmacy, highlighting more and less safe medication options and facilitating individualized patient care. To address limited time, one potential strategy is to have clinic visits specifically focused on pain management, rather than appending pain management to preventive and chronic care management visits. Although these visits could leverage the full potential of CDSSs, this approach may not be feasible due to high patient volumes and reimbursement limitations.

Clinicians viewed I-COPE as a valuable tool for improving pain management in older adults through its structured approach to assessment and treatment. This is in line with previous research suggesting that standardized, multimodal interventions can improve clinical efficiency, clinician confidence, and patient outcomes by facilitating more consistent and collaborative care [21,22]. The benefits anticipated by clinicians reflect the broader evidence that well-implemented chronic care models can significantly enhance the quality of chronic pain management. Our findings suggest that the use of pain-focused CDSSs, such as I-COPE, could be enhanced by integrating mechanisms to identify patients' desire to prioritize pain management ahead of their visits, such as through pre-visit questionnaires or scheduling prompts.

However, clinicians also expressed concerns about the challenges of integrating I-COPE into routine practice, particularly regarding workflow disruption, increased workload, and data completeness. These challenges are well documented in the literature on CDSSs, where multicomponent interventions often face resistance due to their complexity [23]. Future innovations in CDSSs, such as through the use of artificial intelligence, may help synthesize complex patient information and deliver more pointed recommendations to clinicians while preserving clinicians' decision-making autonomy [24]. Additionally, issues related to technology accessibility among older adults highlight the need for careful planning to ensure equitable implementation, as older populations often encounter barriers to using electronic health tools [25]. Addressing these challenges is crucial for I-COPE, or any CDSS, to achieve its full potential. Technologies that read questionnaires aloud to patients and allow them to respond verbally are increasingly usable and may help reduce this barrier in the future [26].

While the results of this study provide a valuable contribution to the literature around CDSSs and pain management in older adults in particular, future work is needed. Future research should explore strategies for streamlining the user interface, minimizing documentation burden, and optimizing integration with clinical encounters. Additionally, studies should evaluate how pre-visit tools can be adapted to better accommodate patients with limited digital literacy.

Limitations

Our study includes several limitations. As is the nature of qualitative analysis, our results are influenced by the researchers' perceptions. In addition, our sample is limited to a single health system and geographic location. PCCs self-selected to participate in interviews. It is possible that clinicians who participated were more interested in chronic pain care or more engaged with the intervention than those who did not participate, and their experiences, barriers, and facilitators may differ. The small sample size and specificity of the content also make our findings difficult to generalize.

Conclusions

The findings underscore clinician perspectives on current pain management in older adults and the potential of the I-COPE intervention to address existing barriers. Successful implementation will hinge upon addressing challenges related to workflow integration, clinician and patient adaptation, and accessibility. These insights are crucial for optimizing the delivery of comprehensive pain care and enhancing patient outcomes within primary care settings for older adults.

Acknowledgments

Agency for Healthcare Research and Quality provided funding but was not involved in any component of the design, methods, participant recruitment, data collections, analysis, or preparation of paper.

Disclaimer

No generative artificial intelligence or large language models were used in preparation of this manuscript.

Data Availability

Data are not publicly available due to the sensitive nature of qualitative interview content and the risk of participant reidentification. However, descriptive summary tables with representative deidentified participant quotes are included in the manuscript to support key findings. Additional data may be made available from the corresponding author upon reasonable request and with appropriate institutional approvals.

Authors' Contributions

IH, EMS, AK, MA, KT, DL, and NL made substantial contributions to the conception or design of the work. IH acquired the data for the work; IH, EMS, and AK analyzed the data and drafted the work; and MM, MA, KT, DL, AZ, GE, CAH, VGP, and

XSL•FO
RenderX

NL reviewed the work critically for important intellectual content. All authors interpreted data for the work, gave final approval of the version to be published, and agreed to the accuracy of the work.

Conflicts of Interest

VG Press reports receiving funding from National Institutes of Health and Agency for Healthcare Research and Quality and receiving consulting fees from Humana. All other authors report no conflicts of interest.

Multimedia Appendix 1
I-COPE (Improving Chicago Older Adult Opioid and Pain Management through Patient-centered Clinical Decision Support and Project ECHO) pre-visit patient questionnaire.
[DOCX File, 288 KB - formative_v9i1e74381_app1.docx]

Multimedia Appendix 2
I-COPE (Improving Chicago Older Adult Opioid and Pain Management through Patient-centered Clinical Decision Support and Project ECHO) electronic order set.
[DOCX File, 2322 KB - formative_v9i1e74381_app2.docx]

Multimedia Appendix 3
I-COPE (Improving Chicago Older Adult Opioid and Pain Management through Patient-centered Clinical Decision Support and Project ECHO) conversation tool.
[DOCX File, 153 KB - formative_v9i1e74381_app3.docx]

Multimedia Appendix 4
Sample of I-COPE (Improving Chicago Older Adult Opioid and Pain Management through Patient-centered Clinical Decision Support and Project ECHO) patient education materials.
[DOCX File, 126 KB - formative_v9i1e74381_app4.docx]

Multimedia Appendix 5
Interview guide.
[DOCX File, 27 KB - formative_v9i1e74381_app5.docx]

Multimedia Appendix 6
Anticipated benefits and challenges of using I-COPE (Improving Chicago Older Adult Opioid and Pain Management through Patient-centered Clinical Decision Support and Project ECHO) intervention.
[DOCX File, 26 KB - formative_v9i1e74381_app6.docx]

References

1. Dowell D, Ragan KR, Jones CM, Baldwin GT, Chou R. CDC clinical practice guideline for prescribing opioids for pain—United States, 2022. MMWR Recomm Rep 2022 Nov 4;71(3):1-95. [doi: 10.15585/mmwr.rr7103a1] [Medline: 36327391]
2. American Geriatrics Society Panel on the Pharmacological Management of Persistent Pain in Older Persons. Pharmacological Management of Persistent Pain in Older Persons. J Am Geriatr Soc 2009 Aug;57(8):1331-1346. [doi: 10.1111/j.1532-5415.2009.02376.x]
3. Zomahoun HTV, Visca R, George N, Ahmed S. Effectiveness and harms of clinical decision support systems for referral within chronic pain practice: protocol for a systematic review and meta-analysis. Syst Rev 2021 Feb 9;10(1):53. [doi: 10.1186/s13643-021-01596-7] [Medline: 33563328]
4. Dahlhamer J, Lucas J, Zelaya C, et al. Prevalence of chronic pain and high-impact chronic pain among adults—United States, 2016. MMWR Morb Mortal Wkly Rep 2018 Sep 14;67(36):1001-1006. [doi: 10.15585/mmwr.mm6736a2] [Medline: 30212442]
5. Huang KTL, Blazey-Martin D, Chandler D, Wurcel A, Gillis J, Tishler J. A multicomponent intervention to improve adherence to opioid prescribing and monitoring guidelines in primary care. J Opioid Manag 2019;15(6):445-453. [doi: 10.5055/jom.2019.0535] [Medline: 31850506]
6. Wang EJ, Helgesen R, Johr CR, Lacko HS, Ashburn MA, Merkel PA. Targeted program in an academic rheumatology practice to improve compliance with opioid prescribing guidelines for the treatment of chronic pain. Arthritis Care Res (Hoboken) 2021 Oct;73(10):1425-1429. [doi: 10.1002/acr.24354] [Medline: 32558375]

XSL•FO
RenderX

7. Frost MC, Soyer EM, Achtmeyer CE, et al. Treating opioid use disorder in veterans with co-occurring substance use: a qualitative study with buprenorphine providers in primary care, mental health, and pain settings. Addict Sci Clin Pract 2023 May 4;18(1):26. [doi: 10.1186/s13722-023-00382-1] [Medline: 37143162]
8. Kaseweter K, Nazemi M, Gregoire N, Louw WF, Walsh Z, Holtzman S. Physician perspectives on chronic pain management: barriers and the use of eHealth in the COVID-19 era. BMC Health Serv Res 2023 Oct 20;23(1):1131. [doi: 10.1186/s12913-023-10157-8] [Medline: 37864210]
9. Spithoff S, Mathieson S, Sullivan F, et al. Clinical decision support systems for opioid prescribing for chronic non-cancer pain in primary care: a scoping review. J Am Board Fam Med 2020;33(4):529-540. [doi: 10.3122/jabfm.2020.04.190199] [Medline: 32675264]
10. Nair KM, Malaeekeh R, Schabort I, Taenzer P, Radhakrishnan A, Guenter D. A clinical decision support system for chronic pain management in primary care: usability testing and its relevance. J Innov Health Inform 2015;22(3):329-332. [doi: 10.14236/jhi.v22i3.149]
11. Harle CA, Apathy NC, Cook RL, et al. Information needs and requirements for decision support in primary care: an analysis of chronic pain care. AMIA Annu Symp Proc 2018;2018:527-534. [Medline: 30815093]
12. Allen KS, Danielson EC, Downs SM, et al. Evaluating a prototype clinical decision support tool for chronic pain treatment in primary care. Appl Clin Inform 2022 May;13(3):602-611. [doi: 10.1055/s-0042-1749332] [Medline: 35649500]
13. Cuadros P, McCord E, McDonnell C, et al. Barriers, facilitators, and recommendations to increase the use of a clinical decision support tool for managing chronic pain in primary care. Int J Med Inform 2024 Dec;192:105649. [doi: 10.1016/j.ijmedinf.2024.105649] [Medline: 39427385]
14. Kagarmanova A, Sparkman H, Laiteerapong N, et al. Improving the management of chronic pain, opioid use, and opioid use disorder in older adults: study protocol for I-COPE study. Trials 2022 Jul 27;23(1):602. [doi: 10.1186/s13063-022-06537-w] [Medline: 35897111]
15. Elder JW, Wu EF, Chenoweth JA, et al. Emergency department screening for unhealthy alcohol and drug use with a brief tablet-based questionnaire. Emerg Med Int 2020;2020:8275386. [doi: 10.1155/2020/8275386] [Medline: 32724677]
16. Krebs EE, Gravely A, Nugent S, et al. Effect of opioid vs nonopioid medications on pain-related function in patients with chronic back pain or hip or knee osteoarthritis pain: the SPACE Randomized Clinical Trial. JAMA 2018 Mar 6;319(9):872-882. [doi: 10.1001/jama.2018.0899] [Medline: 29509867]
17. Löwe B, Kroenke K, Gräfe K. Detecting and monitoring depression with a two-item questionnaire (PHQ-2). J Psychosom Res 2005 Feb;58(2):163-171. [doi: 10.1016/j.jpsychores.2004.09.006] [Medline: 15820844]
18. Damschroder LJ, Reardon CM, Widerquist MAO, Lowery J. The updated Consolidated Framework for Implementation Research based on user feedback. Implement Sci 2022 Oct 29;17(1):75. [doi: 10.1186/s13012-022-01245-0] [Medline: 36309746]
19. King N. Using templates in the thematic analysis of text. In: Cassell C, Symon G, editors. Essential Guide to Qualitative Methods in Organizational Research: SAGE Publications Ltd:257-270 URL: https://sk.sagepub.com/dict/edvol/essential-guide-to-qualitative-methods-in-organizational-research/chpt/using-templates-the-thematic-analysis-text [accessed 2025-08-10] [doi: 10.4135/9781446280119.n21]
20. Rindal DB, Pasumarthi DP, Thirumalai V, et al. Clinical decision support to reduce opioid prescriptions for dental extractions using SMART on FHIR: implementation report. JMIR Med Inform 2023 Nov 7;11:e45636. [doi: 10.2196/45636] [Medline: 37934572]
21. Dobscha SK, Corson K, Perrin NA, et al. Collaborative care for chronic pain in primary care: a cluster randomized trial. JAMA 2009 Mar 25;301(12):1242-1252. [doi: 10.1001/jama.2009.377] [Medline: 19318652]
22. Kroenke K, Krebs EE, Wu J, Yu Z, Chumbler NR, Bair MJ. Telecare collaborative management of chronic pain in primary care: a randomized clinical trial. JAMA 2014 Jul 16;312(3):240-248. [doi: 10.1001/jama.2014.7689] [Medline: 25027139]
23. Buntin MB, Burke MF, Hoaglin MC, Blumenthal D. The benefits of health information technology: a review of the recent literature shows predominantly positive results. Health Aff (Millwood) 2011 Mar;30(3):464-471. [doi: 10.1377/hlthaff.2011.0178] [Medline: 21383365]
24. Granviken F, Meisingset I, Bach K, et al. Personalised decision support in the management of patients with musculoskeletal pain in primary physiotherapy care: a cluster randomised controlled trial (the SupportPrim project). Pain 2025 May 1;166(5):1167-1178. [doi: 10.1097/j.pain.0000000000003456] [Medline: 39432806]
25. Garcia Reyes EP, Kelly R, Buchanan G, Waycott J. Understanding older adults' experiences with technologies for health self-management: interview study. JMIR Aging 2023 Mar 21;6:e43197. [doi: 10.2196/43197] [Medline: 36943333]
26. Azad CL, Beres LK, Wu AW, Fong A, Giladi AM. Developing a multimedia patient-reported outcomes measure for low literacy patients with a human-centered design approach. PLoS One 2024;19(6):e0304351. [doi: 10.1371/journal.pone.0304351] [Medline: 38838037]

Abbreviations

CDSS: clinical decision support system
EHR: electronic health record

XSL•FO
RenderX

I-COPE: Improving Chicago Older Adult Opioid and Pain Management through Patient-centered Clinical Decision Support and Project ECHO
OUD: opioid use disorder
PCC: primary care clinician

Edited by A Mavragani; submitted 23.03.25; peer-reviewed by F Yeh, L Bai; revised version received 16.06.25; accepted 18.06.25; published 26.08.25.

Please cite as:
Hasnain I, Staab EM, Kagarmanova A, Mackiewicz M, Ari M, Thompson K, Lazar D, Zhao A, Elwyn G, Harle CA, Press VG, Laiteerapong N
Primary Care Clinician Perspectives on Older Adult Chronic Pain Management and Clinical Decision Support: Qualitative Study
JMIR Form Res 2025;9:e74381
URL: https://formative.jmir.org/2025/1/e74381
doi:10.2196/74381

© Isra Hasnain, Erin M Staab, Ainur Kagarmanova, Marissa Mackiewicz, Mim Ari, Katherine Thompson, Danielle Lazar, Anne Zhao, Glyn Elwyn, Christopher A Harle, Valerie G Press, Neda Laiteerapong. Originally published in JMIR Formative Research (https://formative.jmir.org), 26.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Technical Description of eHealth Tools in an Environmental Exposure Chamber: Implementation Study

Irene Garcia-Gutierrez[1,2,3*], MD; Irene Dueñas-Requena[4,5*], BME; Violeta Sanchez-Garcia[4,6], BME; Cristiana Solorzano-Zepeda[3,4,6], MD; Samuel Santos-Benito[6], HD; Belen de la Hoz-Caballer[3,4,6], MD, PhD; Dario Antolin-Amerigo[3,4,6], MD, PhD

[1]Allergy Department, Marqués de Valdecilla University Hospital, Santander, Spain
[2]Instituto de Investigación Marqués de Valdecilla, Santander, Spain
[3]Faculty of Medicine, Universidad de Alcalá, Alcalá de Henares, Spain
[4]Allergy Department, Hospital Universitario Ramón y Cajal, M-607, km. 9, 100, Madrid, Spain
[5]Technical School of Telecommunication Engineering, Universidad Politécnica de Madrid, Madrid, Spain
[6]Instituto Ramón y Cajal de Investigación Sanitaria, Madrid, Spain
[*]these authors contributed equally

Corresponding Author:
Cristiana Solorzano-Zepeda, MD
Faculty of Medicine, Universidad de Alcalá, Alcalá de Henares, Spain

Abstract

Background: Environmental exposure chambers (EECs) provide a controlled and reproducible setting for studying allergic rhinitis, allowing standardized exposure to aeroallergens. However, real-time symptom monitoring remains challenging, particularly as patients must be observed remotely while inside the chamber.

Objective: This study aimed to design, implement, and evaluate an integrated eHealth system that leverages digital tools for data collection and real-time clinical monitoring during EEC exposures. The goal was to streamline workflow, enhance data reliability, and improve patient safety during allergen challenges.

Methods: We conducted a prospective pilot study in a validated EEC at Ramón y Cajal University Hospital. A total of 34 provocation tests were conducted: 27 in patients with confirmed grass pollen allergy and 7 in nonallergic controls. All exposures took place outside the pollen season using standardized *Phleum pratense* (grass pollen) at concentrations of 940±100 grains/m³. Clinical data were collected using REDCap (Research Electronic Data Capture; Vanderbilt University), a secure electronic data capture platform. Surveys were automatically scheduled and included standardized instruments: Visual Analog Scale (VAS), Total Nasal Symptom Score (TNSS), Total Ocular Symptom Score (TOSS), asthma symptom questionnaires, peak expiratory flow (PEF), and peak nasal inspiratory flow (PNIF). Data were visualized in real time using Power BI dashboards (Microsoft Corp).

Results: Among allergic patients, 85% had a positive response to the allergen exposure, with a median TNSS of 6 (IQR 6‑7), while controls showed no positive reactions. Mean exposure time for positive cases was 48.9 minutes (SD 28.3), as patients were withdrawn early upon reaching symptom thresholds. The REDCap system was configured to collect symptom surveys at 15-minute intervals, allowing up to 6 entries per patient during exposure, depending on clinical tolerance. This setup enabled consistent symptom monitoring and integration with Power BI for real-time visualization. All records were successfully synchronized between REDCap and Power BI. Informal feedback from medical staff highlighted improved workflow and usability.

Conclusions: The integration of REDCap with Power BI enabled automated, real-time tracking of symptoms during EEC exposures. This eHealth solution enhanced clinical oversight, supported timely decision-making, and ensured patient safety. The platform provides a flexible and reproducible model for use in allergen exposure studies and other clinical settings requiring structured remote monitoring. Clinical response data are reported in a separate validation study.

(JMIR Form Res 2025;9:e71276) doi:10.2196/71276

KEYWORDS

telemedicine; allergen challenge; respiratory allergy; remote symptom monitoring; REDCap; Power BI; allergic rhinitis; asthma; digital health; clinical data integration

XSL•FO
RenderX

Introduction

Allergic rhinitis affects 1 in 5 people worldwide. Among other things, allergic rhinitis is associated with a decrease in overall quality of life and productivity loss. Although there are various treatment options, some patients remain undertreated or affected by drug side effects. Rhinoconjunctivitis and asthma in patients who are polysensitized to different aeroallergens constitute a challenge for their diagnosis and treatment, since there is a need to identify the relationship between the load of a specific allergen and patients' symptoms, considering that, in some cases, pollination takes place at the same time of the year for different species. For this reason, a greater understanding of etiological agents and their pathological mechanisms may be effective in guiding a more precise treatment [1].

Environmental exposure chambers (EECs) are generally recognized as safe, effective, and suitable for conducting randomized clinical trials of new allergy treatments [2]. Since their introduction, EECs have provided a controlled setting for allergen exposure, ensuring reproducibility and safety. They have become increasingly important in clinical research, particularly in studies of allergic rhinitis, as they provide a fixed allergen concentration in a tightly controlled environment [3], and outcomes observed within EECs are comparable to those obtained in natural exposure settings, suggesting the usefulness of EECs in allergic rhinitis studies [4].

Mobile health (mHealth) has emerged as a transformative force in the health care industry, leveraging the power of mobile technology to enhance medical services and improve overall health outcomes. These devices have become an indispensable part of modern life, offering a variety of health care–related apps and services [5]. With the widespread adoption of smartphones and other portable devices, mHealth has enabled individuals to access health care resources, monitor their health, and receive medical guidance. Digital solutions are the new challenge of today's fast-paced, changing, and volatile society.

Due to the characteristics of EECs, all monitoring of patient symptoms must be done remotely, as the patient remains inside the sealed environment during exposure. Real-time access to clinical data—such as symptom scores, positivity thresholds, and environmental variables—is essential to ensure both patient safety and the validity of the provocation. Without continuous monitoring, symptom escalation could go undetected, potentially compromising both participant well-being and study reliability.

Methods

Study Design

The exposures within the EEC were conducted for periods up to 90 minutes, scheduled outside the grass pollen season, to minimize the risk of priming phenomena. The design ensured that the responses observed during chamber exposures were attributable solely to the controlled allergen administration.

Environmental Exposure Chamber

The EEC used in this study is a fully controlled, state-of-the-art facility located at the Ramón y Cajal University Hospital (RyCUH). The EEC has a total area of 15.60 m^2 and is equipped to maintain stable temperature, relative humidity, and air pressure, ensuring homogeneous distribution of allergen particles through a calibrated air circulation system. The allergen administered was standardized grass pollen (*Phleum pratense*), aerosolized using a precision nebulizer at concentrations ranging between 940±100 grains/m^3.

The EEC has undergone formal validation [6] for use in allergic respiratory studies, including verification of allergen distribution uniformity, reproducibility of symptom induction, and environmental stability, as described in García-Gutiérrez et al [6]. This validation process ensures that the chamber provides consistent and controlled allergen exposure conditions across sessions. Such standardization is essential for both patient safety and scientific reproducibility in allergen challenge studies.

Study Population

The study population comprised 31 volunteers, including 25 patients with confirmed grass pollen allergy (cases) and 6 non–grass-allergic controls. Participants had a mean age of 31 (SD 10.6) years, and the sex distribution was 22 women and 9 men. Among the allergic individuals, 8 (32%) had mild asthma as a comorbidity. All cases were sensitized to *Phleum pratense*, confirmed by skin prick tests (mean wheal diameter 8.1 mm, SD 3.4 mm) or specific immunoglobulin E levels (median 20.3 kU/L, IQR 4.9 - 42.4 kU/L). Controls included 3 individuals without any sensitization and 3 with sensitization to other aeroallergens (*Olea europaea* and *Cupressaceae* or dust mites). None of the control participants developed symptoms during pollen exposure. All participants were free of allergy-specific treatment for at least 4 weeks before exposure. Exclusion criteria included significant comorbidities, uncontrolled asthma, or contraindications to EEC exposure. Relevant demographic and clinical characteristics were recorded at baseline using REDCap (Research Electronic Data Capture; Vanderbilt University).

In total, 34 grass pollen provocations were conducted, as 3 participants (1 control and 2 allergic individuals) underwent 2 separate exposures with a washout period of at least 4 weeks. In addition, 7 placebo exposures were performed (in both controls and allergic individuals) under identical chamber conditions but without allergen administration. No symptoms were observed during these sessions, supporting the specificity of the protocol.

As this was a pilot feasibility and implementation study, no formal sample size calculation was performed. The study was not statistically powered to test hypotheses, but rather designed to assess the usability and integration of the system in a real-world clinical setting. Consequently, randomization was not applied, as the focus was not on clinical outcome comparison but on technical validation.

Data Collection and Management

REDCap

REDCap, developed by Vanderbilt University, is a secure, web-based app designed to support data capture for research studies [7]. In this study, REDCap was used to build the project infrastructure required for collecting and storing

XSL•FO
RenderX

patient-provided data in a structured and accessible manner. Its dual-interface design offers functionalities that are tailored to both health care professionals (HCP) and patients, facilitating seamless data entry, remote symptom tracking, and centralized monitoring.

REDCap facilitates the creation of 2 types of forms essential for collecting clinical data. First, forms are completed by HCPs to document patient information during the EEC exposure process. Second, surveys are completed by patients to record their symptoms throughout the process, without requiring access to the project or account creation.

Patients completed 3 types of surveys—before, during, and after exposure. These surveys included standardized questionnaires:Visual Analog Scale (VAS) for nasal and general symptoms, Total Nasal Symptom Score (TNSS) and Total Ocular Symptom Score (TOSS), along with specific asthma questionnaires [8]. In addition, participants recorded objective respiratory measurements, including peak expiratory flow (PEF; Mini Wright, Clement Clarke International Ltd) and peak nasal inspiratory flow (PNIF; In-check, Clement Clarke International Ltd) [9,10].

At the beginning of the process, each patient was assigned a unique record number to identify all associated forms and surveys. Surveys completed multiple times were distinguished by an instance number (eg, record number “1” will have instance numbers “1,” “2,” and “3” for 3 survey completions). Both forms and surveys can be edited by HCP.

The data derived from patients, aimed at determining results and obtaining clinically useful conclusions, align with the concept of “Patient Reported Outcomes” (PROMs). These outcomes represent the patient’s direct perspective on their symptoms, quality of life, and response to treatment, without external interpretation. As part of the study, we included a work package that incorporated clinically validated questionnaires specifically created to identify unmet needs and improve patient health outcomes.

Data collection was organized into “events” grouping forms to be completed by both HCP and patients (via surveys) throughout the EEC exposure process. The process was divided into 7 events to collect and group information before, during, and after exposure. This structure allowed for automated survey distribution to patients at specified times and frequencies.

Reports were configured in REDCap to facilitate the review and export of information collected at each stage of the process. Users can filter data by record number, allowing targeted access to specific patient sessions or time points. The platform supports export in a wide range of formats, including CSV, SPSS (IBM Corp), Stata, SAS (SAS Institute Inc), R, and XML, making the data accessible for various types of statistical and programming tools.

Power BI Integration

To enhance data accessibility for medical staff, the collected data from REDCap were integrated into Power BI, a suite of business analytics tool developed by Microsoft for data visualization and sharing insights [11]. This integration enabled the creation of a live “monitor” displaying key variables—such as VAS, TNSS, and TOSS—through customizable filters and a color-coded system that highlights symptom severity. By presenting information in a user-friendly interface, even nontechnical staff could interpret patient status at a glance, improving clinical responsiveness and workflow efficiency during allergen exposure.

To ensure data integrity and reliability, several quality control measures were implemented throughout the digital workflow. REDCap forms were configured with field validation rules and required fields to minimize missing or inconsistent entries. During data transfer, synchronization routines were programmed to flag incomplete or anomalous data, which were subsequently verified against the original REDCap entries. In addition, during real-time monitoring in the EEC, the system was set up to alert health care staff if survey responses were missing or out of the expected range, enabling immediate review and resolution when necessary.

Ethical Considerations

The study was approved by the Bioethics Committee of RyCUH, with record number 445, dated December 20, 2022. The procedures followed were in accordance with the World Medical Association’s Declaration of Helsinki. Informed consent was obtained from each participant.

Results

REDCap Data Collection

The REDCap project was structured to align with the procedural stages of patient exposure within the EEC. Seven distinct events were defined, each corresponding to specific phases of the patient’s process, such as clinical interview, entry into the chamber, and postexposure follow-up. Each event was linked to dedicated forms and surveys, ensuring comprehensive and time-aligned data collection, as depicted in Figure 1.

XSL•FO
RenderX

Figure 1. Events in REDCap (Research Electronic Data Capture; Vanderbilt University) and their associated forms. In the left column, all the forms designed to collect data on environmental exposure chamber exposure are represented. The first row contains the names of the events that make up the process. Therefore, the table indicates which forms belong to each event. EEC: environmental exposure chamber.

Data entry form	1.interview (1)	2.pre_eec (2)	3.entrance_eec (3)	4.eec_forms (4)	5.exit_eec (5)	6.discharge (6)	7.post_eec (7)
First_Form_EEC (survey)		✔					
EEC_Forms (survey)				✔			
PostEEC_Form (survey)							✔
Form_PostEEC_End (survey)							✔
Inclusion_Criteria	✔						
Exclusion_Criteria	✔						
Inclusion_Patient	✔						
Background	✔						
Sociodemographic variables	✔						
PreEEC_Treatments		✔					
PreEEC_Tests		✔					
Entrance_EEC			✔				
EEC_Check				✔			
EEC_Medical_Form				✔			
PostEEC_Tests					✔		
Discharge						✔	

The initial event, “1.interview,” encompasses forms for clinical interview, inclusion and exclusion criteria, patient personal data, background, and sociodemographic variables. These forms are completed by HCPs to assess candidate eligibility for EEC exposure and to collect pertinent clinical and contact information. It is important to note that the patient’s personal data are excluded from data exports to maintain the patient’s anonymity.

The subsequent event, “2.pre_eec,” is completed on the day of the patient’s potential exposure, as it will serve to record treatments or tests carried out during the last 2 weeks that may contraindicate EEC entry. Upon confirming eligibility [6], a baseline symptom survey is emailed to the patient. At this point, the questions included in the survey are explained to the patient, since these will be the same questions that they will answer in subsequent surveys during the exposure. This survey includes assessments such as VAS for general and allergic rhinitis symptoms, TNSS, TOSS, and an asthma symptom questionnaire, as well as measurements of PEF and PNIF. An open-ended section allows patients to report additional symptoms.

Event “3.entrance_eec” records the patient’s entry time into the EEC. This timestamp triggers the automated dispatch of surveys every 15 minutes during the exposure period. This setup enables the collection of up to 6 time-stamped responses per session, allowing for detailed temporal tracking of symptom progression.

Event “4.eec_forms” includes several critical components. It contains patient-completed questionnaires, a form documenting EEC environmental conditions at the beginning and end of the exposure (eg, temperature, pressure, humidity, and airflow), and an optional form activated if severe symptoms require intervention. Completion of this form halts further survey distribution if the exposure lasts less than the planned 90 minutes.

Upon EEC exit, event “5.exit_eec” is completed by HCP to document postexposure assessments and treatments. These include spirometry, fractional exhaled nitric oxide (FeNO; NIOX VERO, Circassia AB), vital signs (blood pressure, heart rate, respiratory rate, and temperature), oxygen saturation, and clinical examinations, with the possibility of recording the results of several spirometries, if necessary. The form also records administered treatments and what it was (eg, antihistamines, corticosteroids, vasoconstrictors, and adrenaline) and determines the provocation outcome in terms of positive or negative for rhinitis, conjunctivitis, or asthma based on symptom scores and predefined thresholds.

Event “6.discharge” logs the patient’s discharge time. This timestamp prompts REDCap to send follow-up surveys every 2 hours on the day of exposure. A final survey is dispatched the following morning, thereby completing the postexposure data collection sequence.

As shown in Figure 2, the Record Status Console in REDCap provides a comprehensive overview of all data collection instruments used throughout the patient exposure process. It visually displays the forms and surveys linked to each event and indicates their completion status. This interface allows study staff to monitor progress and ensure that no required documentation is overlooked at any stage.

XSL•FO
RenderX

Figure 2. Record Status Console in REDCap (Research Electronic Data Capture, Vanderbilt University). Displays collected records, initial patient exposure events, associated forms or surveys, and their completion status. EEC: environmental exposure chamber.

Patient Monitoring Using Power BI

REDCap data were successfully integrated into Power BI for comprehensive visualization and monitoring. The primary dashboard (Figure 3) provides a detailed overview of patient status during exposure. A filtering panel on the left allows the study team to select data by record (assigned per EEC entry) and instance number (corresponding to each survey completion). For example, if only the registration number is selected and not the instance number, the evolution of the patient's symptoms can be seen in the graphs on the right throughout all the registered instances.

XSL•FO
RenderX

Figure 3. Power BI main dashboard. Illustrates the progression of patient record number 1 across 6 survey instances during environmental exposure chamber exposure. Default values represent the most severe readings, as recorded in instance number 6. EEC: environmental exposure chamber; PEF: peak expiratory flow; PNIF: peak nasal inspiratory flow; TNSS: Total Nasal Symptom Score; TOSS: Total Ocular Symptom Score; VAS: Visual Analog Scale.

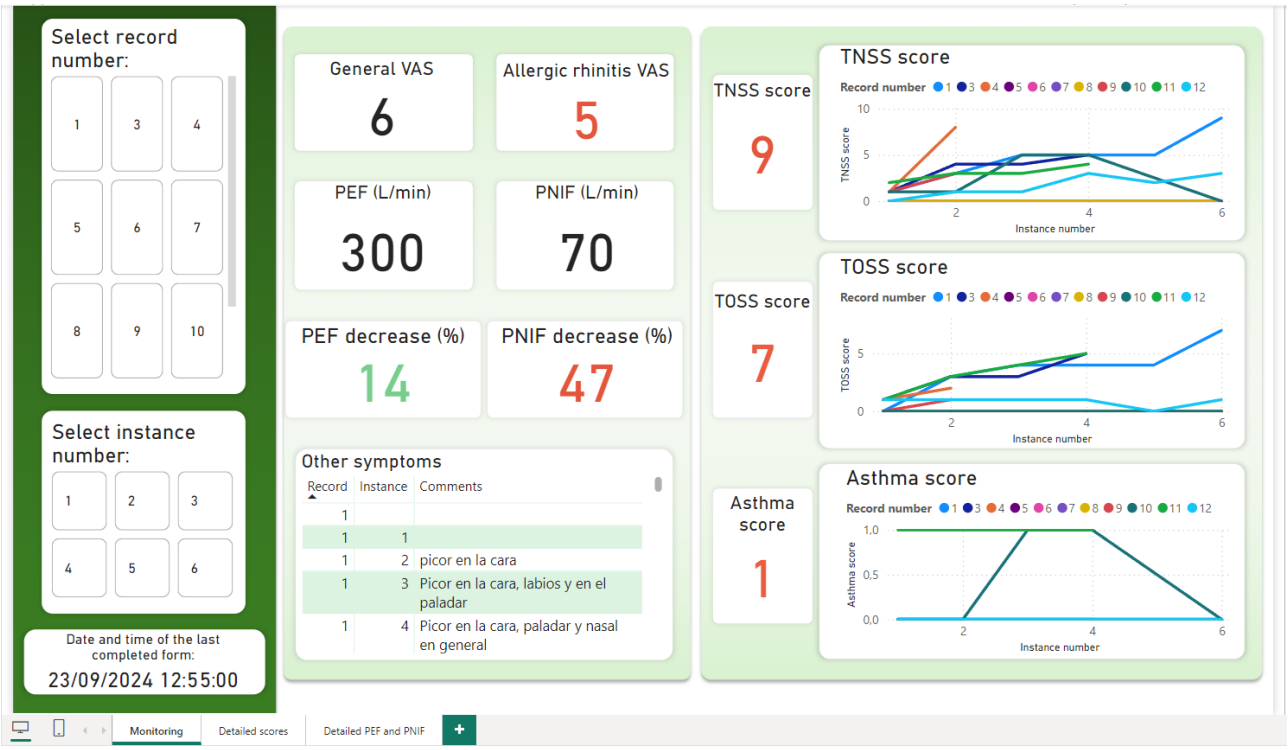

The values shown in the labels (“TNSS score,” “General VAS,” “PEF decrease,” and so on) reflect the most severe values recorded within the selected filters. For PEF and PNIF, the lowest values are highlighted, as they indicate the most critical respiratory function readings. The dashboard supports multirecord and multi-instance selections to compare severity across different exposures. In addition, the dashboard indicates the timestamp of the last completed form for each patient. This information is consistently displayed across all dashboard pages.

Certain metrics in the Power BI dashboard are color-coded to indicate the severity of reported symptoms. Acceptable values appear in black or green, while moderate but noncritical readings are highlighted in orange. When values reach a threshold considered severe or potentially hazardous, they are displayed in red, indicating the need for immediate clinical attention.

The second dashboard page (Figure 4) offers a general view of patient responses to the TNSS, TOSS, and questionnaires related to asthma symptoms. These results are displayed in detailed tables, each showing record number, instance number, individual questions, and corresponding answers. This structure facilitates precise analysis of symptom patterns and enables comparisons across different time points or patients.

XSL•FO
RenderX

Figure 4. Second page of the Power BI dashboard. Presents raw, detailed responses to the Total Nasal Symptom Score, Total Ocular Symptom Score, and asthma symptoms questionnaire (Spanish is used as the language of the included population) can be observed for each test in each of the tables. Tables include record numbers, instance numbers, questions, and patient responses.TNSS: Total Nasal Symptom Score; TOSS: Total Ocular Symptom Score.

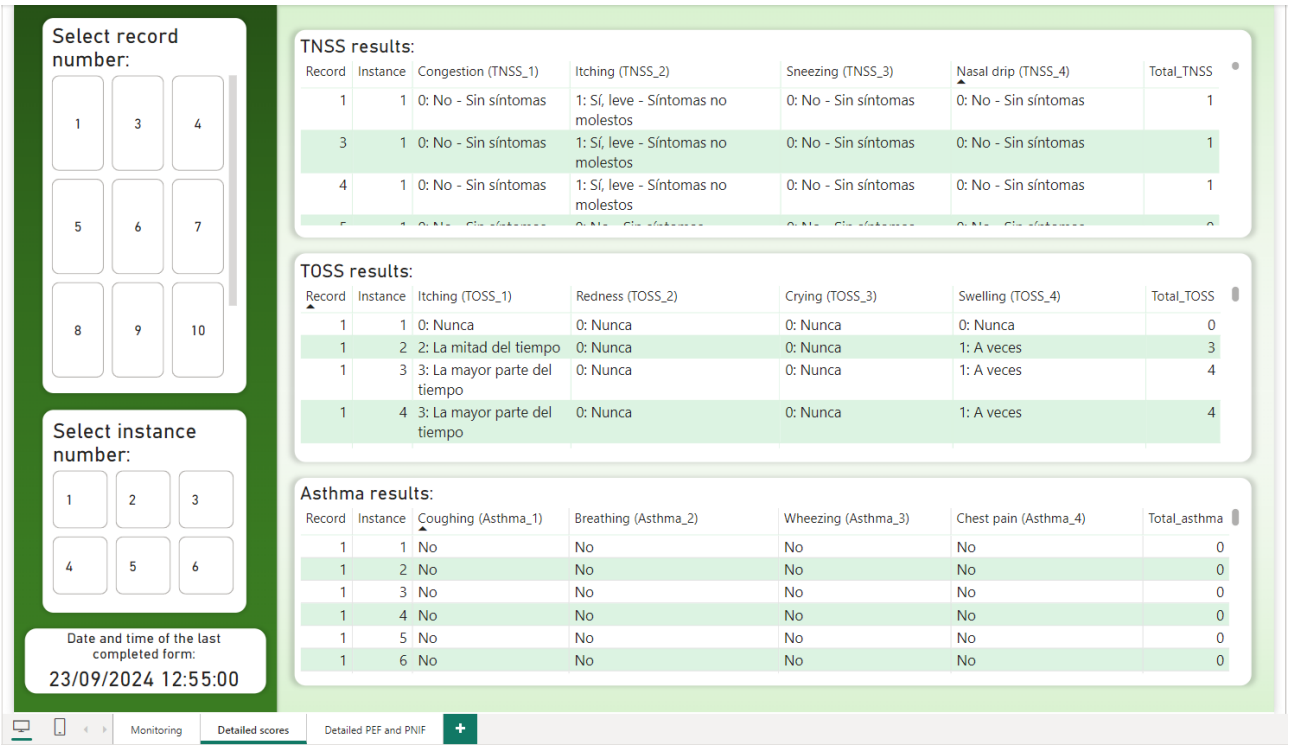

The third dashboard (Figure 5) focuses on PEF and PNIF results. It allows the research team to compare multiple recordings within the same instance or across different moments in the exposure process. This facilitates a more accurate evaluation of the patient's respiratory function and helps identify potential patterns or deterioration over time. A short video has been added as Multimedia Appendix 1 showing the operation of the EEC using this system.

Figure 5. Third and last page of the Power BI dashboard. Displays peak expiratory flow and peak nasal inspiratory flow test results across all instances for record number 1, including baseline values recorded during the initial pre-exposure survey. PEF: peak expiratory flow; PNIF: peak nasal inspiratory flow.

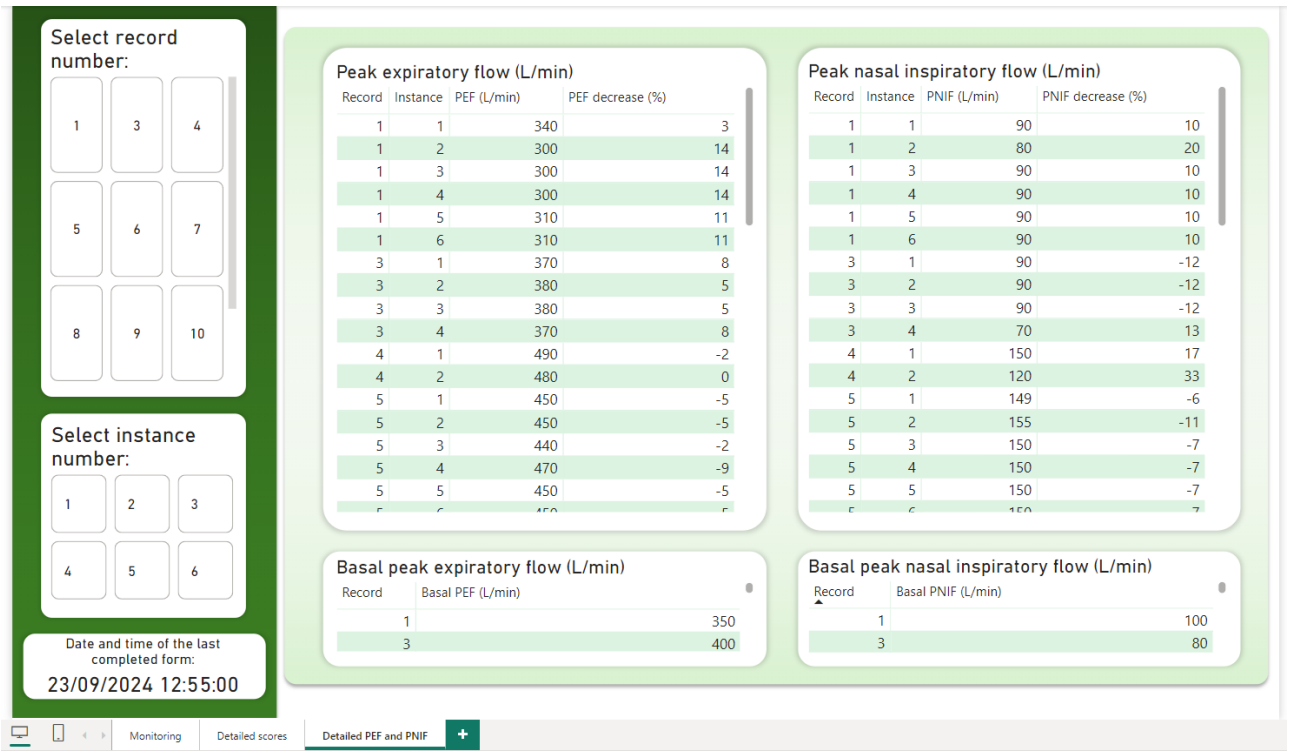

XSL•FO
RenderX

Discussion

Principal Findings

In this study, we developed and implemented a digital monitoring system that integrates REDCap and Power BI to enable structured data collection and real-time visualization during controlled allergen exposures in an EEC. The platform allowed for continuous remote tracking of PROMs, symptom progression, and PEF and PNIF values, facilitating timely clinical decision-making. These results support the feasibility and usefulness of applying digital tools to improve patient monitoring and operational efficiency in allergen challenge studies.

Telemedicine and remote care have emerged as revolutionary forces in the health care sector, transforming the way medical services are delivered and accessed in various specialties, including allergology. These tools have enabled more flexible, scalable, and timely interactions between HCPs and patients, particularly in chronic disease management. With the advancements in technology and the growing demand for efficient health care, telemedicine has become an essential resource for allergists, other HCPs, and patients alike [5].

The specialty of allergology has witnessed significant advances in telemedicine in recent years. Telemedicine in this medical specialty in Spain involves the use of telecommunications and digital technologies to provide internet-based consultations, monitor patients remotely, offer allergy advice, and facilitate follow-up care. This approach has proven to be particularly beneficial for Spanish patients with allergy who often require ongoing management and support [12].

Telemedicine in Spain has emerged as particularly effective in managing asthma and allergic rhinitis, 2 of the most common allergic conditions. Patients can consult with allergists through video calls, receive personalized treatment plans, and learn about environmental control measures to mitigate allergen exposure, or even benefit from the use of connected devices to monitor their lung function and share the data with health care providers [13]. This allows for early detection of asthma exacerbations and timely intervention.

Telemedicine and remote care have had a profound impact on allergy care in Spain, bringing specialized allergy care within reach of patients across the country. However, in certain cases, face-to-face consultations are still required for comprehensive diagnosis and treatment. Integrating telemedicine with traditional health care practices [14] is essential to ensure continuity of care, comprehensive allergy diagnostics and treatment, and to ensure the best disease control. Building trust through internet-based consultations may require extra effort on the part of HCP [15].

Moreover, all forms of telehealth overcome common barriers of time and transportation, making them convenient for patients and HCPs, subsequently trained [5,16,17].

Allergists and other HCPs must remain informed about the rapidly evolving landscape of regulations and reimbursement policies. This is particularly important as telemedicine continues to expand and integrate into mainstream clinical practice. Ensuring compliance with local, regional, national, and international legal frameworks is essential for the safe, ethical, and effective delivery of online care [17].

Telemedicine plays a crucial role in EEC studies, where real-time clinical monitoring is essential and physical interaction is limited. However, there is currently no standardized method for collecting clinical data during EEC provocations [3]. The system presented in this study addresses this gap by offering a versatile and structured digital tool with several practical features (Figure 6): (1) Ease of data recording [5]: the data recording process is structured in different forms, organized in events. This way, medical or research staff have a guide to follow throughout the entire process. As for patients, no training is required to record the data. (2) Automated sending of surveys: medical or research staff do not need to manually configure the sending of any survey; it is enough to record the times of entry to the EEC and discharge of the patient. (3) Automatic determination and recording of positives: for each exposure to the EEC, the form to be filled out with the tests (and possible treatments) of the patient upon leaving the EEC will automatically determine whether the patient presented symptoms of rhinitis, conjunctivitis, or asthma, so that medical staff can fill out the rest of the form later without needing to review the results of the surveys. This point has been especially important during the provocations carried out so far in the EEC in order to avoid errors and to be able to provide the most complete care to the patient [6]. (4) Downloading data in different formats [18]: the collected data can be downloaded in different formats, useful for statistical analysis programs such as SPSS or statistical programming languages such as SAS or R programming language. This is one of the reasons why this study is considered to promote interdisciplinarity, since it offers results that can be analyzed by other professionals related to the medical field. (5) Presentation of the data on a dashboard: as previously mentioned, it was considered necessary to transfer the data to another program to facilitate patient follow-up within the EEC, since the REDCap interface can be complicated for nonspecialized personnel. Thanks to the dashboard that was developed, it was possible to include a color code to indicate the severity of the patient's symptoms, a visual and striking way for medical or research personnel to identify a possible dangerous situation for the patient. (6) Manual or automatic update of the dashboard: the dashboard can be used in its desktop version, so that the update would be manual (refreshing the dashboard), or it can be uploaded to a server and a time interval configured before the automatic update. (7) Possibility of continuing to update this tool: in case changes are required (eg, more exclusion or inclusion criteria to be checked), both the project in REDCap and the data displayed by the dashboard in Power BI can continue to be updated, which gives the project greater versatility.

XSL•FO
RenderX

Figure 6. . Characteristics of the innovative eHealth tool for environmental exposure chamber studies.

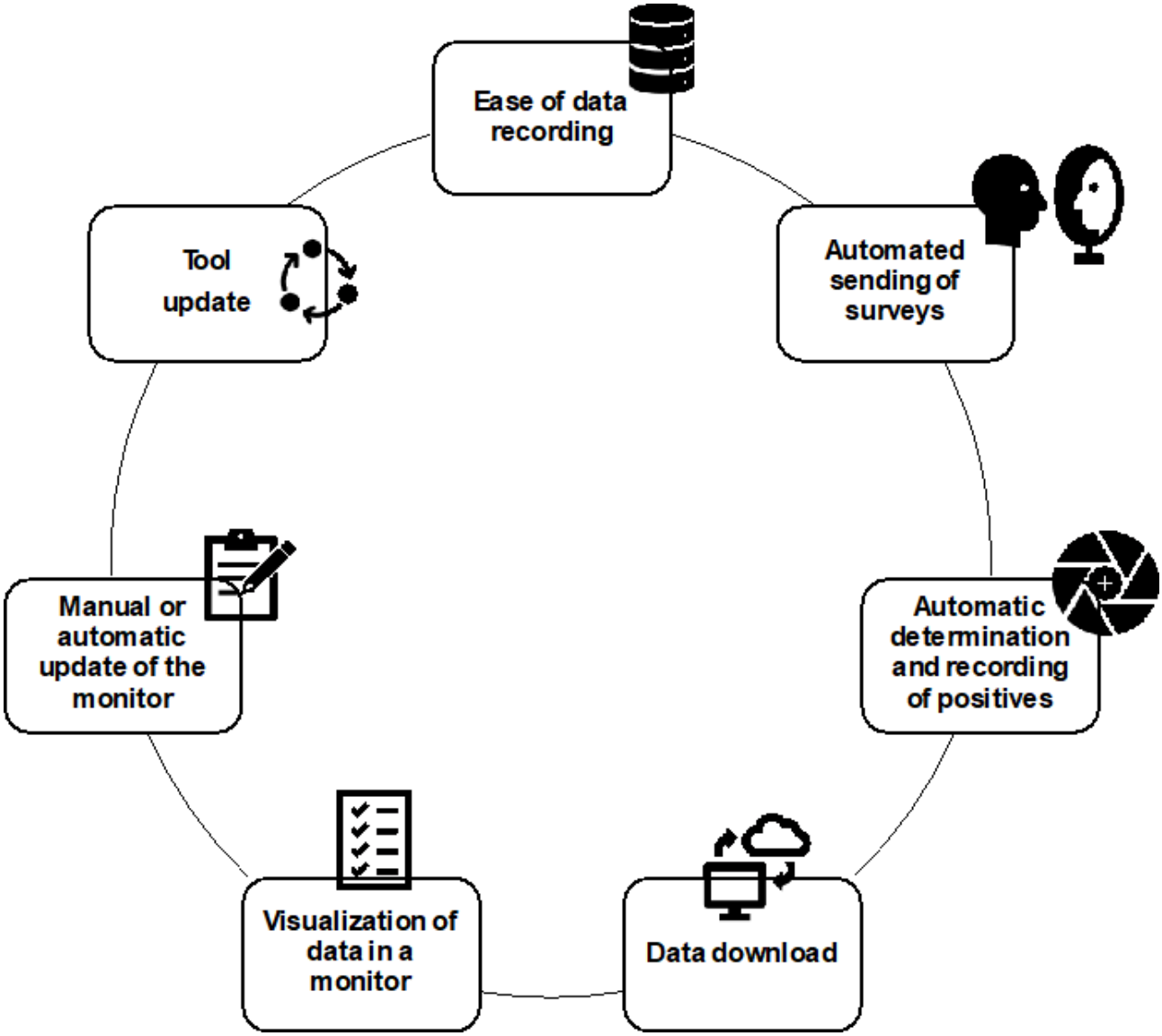

The symptoms' appearance and the measure of scores (TOSS, TNSS, etc) are the main outcomes in the trials searching drug onset of action performed in the EECs [19]. Although it is not yet standardized how the clinical data of patients should be collected in the EEC provocations, using telemedicine is critical in this field [3,20]. All these features of this new infrastructure for data collection during the EEC provocations allow for improved patient follow-up through accurate symptom monitoring, and it is paramount to carry out reliable and accurate EEC studies [6].

Informal feedback from the medical and research staff who participated in the implementation phase has been positive. Users highlighted the clarity and structure provided by the REDCap event design, as well as the real-time visualization in Power BI, which facilitated patient monitoring and reduced the cognitive load during allergen exposure. Several team members expressed that the integration of both tools significantly improved workflow efficiency and confidence in managing symptom progression inside the EEC.

This study has some limitations. It was designed as a pilot feasibility and technical implementation study; therefore, no formal power analysis was performed, and the sample size was not intended for hypothesis testing. The primary objective was to validate the integration and real-time use of the digital health monitoring system in a controlled exposure setting. Furthermore, the study focused exclusively on grass pollen allergy, and the tool has not yet been validated with other allergens or in different EECs. Future studies with larger and more diverse populations are needed to assess broader usability and clinical outcomes.

Conclusions

This work presents a valuable digital solution for both patients and HCPs involved in allergen exposure studies. By integrating clinical workflows with real-time data monitoring, the system enables comprehensive tracking of PROMs in a controlled environment. Designed collaboratively by experts in clinical care and biomedical research, this tool aims to enhance the safety, precision, and effectiveness of the follow-up of patients with respiratory allergies.

XSL•FO
RenderX

Acknowledgments

We thank Diego Cárdenas Cuadrado, Jorge Lázaro Bailón, and Sergio Olmos Piñero (Data Science, IRYCIS Hospital Universitario Ramón y Cajal) for their knowledge and help in developing the REDCap project, as well as its maintenance. This work has received 2 financial grants from the Spanish Society of Allergology and Clinical Immunology (SEAIC), as follows: - Kinetics of biomarkers and metabolomics in patients with respiratory allergy due to grass pollen in an environmental exposure chamber. - Technical and clinical validation of an innovative environmental exposure chamber.

Data Availability

The datasets generated during this study are available in the GitHub repository [21]. The repository includes anonymized .xml configuration files and example dashboards to support reuse, reproducibility, and potential adaptation in similar clinical or research settings. These resources are provided to promote transparency and encourage collaboration within the scientific and medical informatics communities.

Authors' Contributions

IG-G was responsible for conceptualization, writing the original draft, reviewing and editing the manuscript, and project administration. ID-R contributed to methodology, software development, formal analysis, data curation, original draft writing, and visualization. VS-G was involved in investigation, data curation, and visualization. CS-Z contributed to conceptualization, methodology, validation, investigation, resources, and manuscript review and editing. SS-B was responsible for software development and data curation. BHC contributed to conceptualization, methodology, manuscript review and editing, supervision, and project administration. DA-A contributed to conceptualization, methodology, manuscript review and editing, supervision, and project administration.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Short video showing the operation of the implemented system described in the article.
[MP4 File, 63091 KB - formative_v9i1e71276_app1.mp4]

References

1. Silver JD, Spriggs K, Haberle S, Katelaris CH, Newbigin EJ, Lampugnani ER. Crowd-sourced allergic rhinitis symptom data: the influence of environmental and demographic factors. Sci Total Environ 2020 Feb 25;705:135147. [doi: 10.1016/j.scitotenv.2019.135147] [Medline: 31841904]
2. Horak F, Jager S. The Vienna challenge chamber- a new method of allergy exposure tests. Wien Klin Wochenschr 1986(55):509-510.
3. Pfaar O, Bergmann KC, Bonini S, et al. Technical standards in allergen exposure chambers worldwide - an EAACI Task Force Report. Allergy 2021 Dec;76(12):3589-3612. [doi: 10.1111/all.14957] [Medline: 34028057]
4. Gherasim A, Dietsch F, Beck M, Domis N, de Blay F. Birch-induced allergic rhinitis: results of exposure during nasal allergen challenge, environmental chamber, and pollen season. World Allergy Organ J 2023 Jul;16(7):100801. [doi: 10.1016/j.waojou.2023.100801] [Medline: 37520615]
5. Matricardi PM, Dramburg S, Alvarez-Perea A, et al. The role of mobile health technologies in allergy care: An EAACI position paper. Allergy 2020 Feb;75(2):259-272. [doi: 10.1111/all.13953] [Medline: 31230373]
6. García-Gutiérrez I, Solórzano-Zepeda C, Sánchez-García V, et al. Validation of an environmental exposure chamber for assessment of allergy to grass pollen. J Investig Allergol Clin Immunol 2025 Feb 13:0. [doi: 10.18176/jiaci.1036] [Medline: 39946169]
7. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381. [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
8. GEMA. Spanish Guidelines on the Management of Asthma (GEMA 54), 5th edition: Luzán; 2024. URL: https://www.gemasma.com [accessed 2025-08-19]
9. Singh NK, Nagpure PS, Yadav MK, Chavan S, Manpe S, Ganeshkar R. Effectiveness of oral antihistamines and intranasal steroid spray in relieving ocular symptoms in allergic rhinitis using Total Ocular Symptom Score. Int J Otorhinolaryngol Clin 2016 Aug;8(2):45-50. [doi: 10.5005/jp-journals-10003-1226]
10. Augé J, Vent J, Agache I, et al. EAACI position paper on the standardization of nasal allergen challenges. Allergy 2018 Aug;73(8):1597-1608. [doi: 10.1111/all.13416] [Medline: 29377177]

XSL•FO
RenderX

11. What is power BI? Microsoft. 2024. URL: https://learn.microsoft.com/en-us/power-bi/fundamentals/power-bi-overview [accessed 2025-07-29]
12. Alvarez-Perea A, Dimov V, Popescu FD, Zubeldia JM. The applications of eHealth technologies in the management of asthma and allergic diseases. Clin Transl Allergy 2021 Aug;11(7):e12061. [doi: 10.1002/clt2.12061] [Medline: 34504682]
13. Antolín-Amérigo D, Alvarez-Perea A, Popescu FD. Chapter 20 - Patient-friendly HIT tools and the advent of crowdsourcing clinical trials. IPMBPCA 2019:135-144. [doi: 10.1016/B978-0-12-813471-9.00023-2]
14. Smolinska S, Popescu FD, Izquierdo E, et al. Telemedicine with special focus on allergic diseases and asthma-status 2022: an EAACI position paper. Allergy 2024 Apr;79(4):777-792. [doi: 10.1111/all.15964] [Medline: 38041429]
15. Jácome C, Pereira AM, Amaral R, et al. The use of remote care during the Coronavirus disease 2019 pandemic: a perspective of Portuguese and Spanish physicians. Eur Ann Allergy Clin Immunol 2022 Jan;54(1):25-29. [doi: 10.23822/EurAnnACI.1764-1489.184] [Medline: 33354962]
16. Portnoy JM, Pandya A, Waller M, Elliott T. Telemedicine and emerging technologies for health care in allergy/immunology. J Allergy Clin Immunol 2020 Feb;145(2):445-454. [doi: 10.1016/j.jaci.2019.12.903] [Medline: 32035604]
17. Bajowala SS, Shih J, Varshney P, Elliott T. The future of telehealth for allergic disease. J Allergy Clin Immunol Pract 2022 Oct;10(10):2514-2523. [doi: 10.1016/j.jaip.2022.08.022] [Medline: 36038132]
18. Alvarez-Perea A, Sánchez-García S, Muñoz Cano R, Antolín-Amérigo D, Tsilochristou O, Stukus DR. Impact of "eHealth" in allergic diseases and allergic patients. J Investig Allergol Clin Immunol 2019;29(2):94-102. [doi: 10.18176/jiaci.0354] [Medline: 30457105]
19. Bousquet J, Meltzer EO, Couroux P, et al. Onset of action of the fixed combination intranasal azelastine-fluticasone propionate in an allergen exposure chamber. J Allergy Clin Immunol Pract 2018;6(5):1726-1732. [doi: 10.1016/j.jaip.2018.01.031] [Medline: 29425904]
20. Zemelka-Wiacek M, Kosowska A, Jutel M. Symptom assessment of patients with allergic rhinitis using an allergen exposure chamber. J Vis Exp 2023 Mar 3(193). [doi: 10.3791/64801] [Medline: 36939263]
21. Irene-dr/cea_forms. GitHub. URL: https://github.com/irene-dr/CEA_forms [accessed 2025-08-19]

Abbreviations:

EEC: environmental exposure chamber
FeNO: fractional exhaled nitric oxide
HCP: health care professional
mHealth: mobile health
PEF: peak expiratory flow
PNIF: peak nasal inspiratory flow
PROM: patient-reported outcome
REDCap: Research Electronic Data Capture
RyCUH: Ramón y Cajal University Hospital
TNSS: Total Nasal Symptom Score
TOSS: Total Ocular Symptom Score
VAS: Visual Analog Scale

Edited by J Sarvestan; submitted 14.01.25; peer-reviewed by M Zemelka-Wiacek, M Koeberle; revised version received 18.06.25; accepted 20.06.25; published 26.08.25.

Please cite as:
Garcia-Gutierrez I, Dueñas-Requena I, Sanchez-Garcia V, Solorzano-Zepeda C, Santos-Benito S, de la Hoz-Caballer B, Antolin-Amerigo D
Technical Description of eHealth Tools in an Environmental Exposure Chamber: Implementation Study
JMIR Form Res 2025;9:e71276
URL: https://formative.jmir.org/2025/1/e71276
doi:10.2196/71276

© Irene Garcia-Gutierrez, Irene Dueñas-Requena, Violeta Sanchez-Garcia, Cristiana Solorzano-Zepeda, Samuel Santos-Benito, Belen de la Hoz-Caballer, Dario Antolin-Amerigo. Originally published in JMIR Formative Research (https://formative.jmir.org), 26.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Automated Data Harmonization in Clinical Research: Natural Language Processing Approach

Pratheek Mallya[1], MS; Ricardo Henao[2], PhD; Chuan Hong[2], PhD; Daniel Wojdyla[3], MS; Tony Schibler[3], MPA; Vihaan Manchanda[1], BS; Michael Pencina[2], PhD; Jennifer Hall[1], PhD; Juan Zhao[1], PhD

[1]American Heart Association, 7272 Greenville Ave, Dallas, TX, United States

[2]Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, United States

[3]Duke Clinical Research Institute, Durham, NC, United States

Corresponding Author:
Juan Zhao, PhD
American Heart Association, 7272 Greenville Ave, Dallas, TX, United States

Abstract

Background: Integrating data is essential for advancing clinical and epidemiological research. However, because datasets often describe variables (eg, demographic and health conditions) in diverse ways, the process of integrating and harmonizing variables from research studies remains a major bottleneck.

Objective: The objective was to assess a natural language processing–based method to automate variable harmonization to achieve a scalable approach to integration of multiple datasets.

Methods: We developed a fully connected neural network (FCN) method, enhanced with contrastive learning, using domain-specific embeddings from the Bidirectional Encoder Representations from Transformers for Biomedical Text Mining language representation model, using 3 cardiovascular datasets: the Atherosclerosis Risk in Communities study, the Framingham Heart Study, and the Multi-Ethnic Study of Atherosclerosis. We used metadata variable descriptions and curated harmonized concepts as ground truth. We framed the problem as a paired sentence classification task. The accuracy of this method was compared with a logistic regression baseline method. To assess the generalizability of the trained models, we also evaluated their performance by separating the 3 datasets when preparing the training and validation sets.

Results: The newly developed FCN achieved a top-5 accuracy of 98.95% (95% CI 98.31% - 99.47%) and an area under the receiver operating characteristic (AUC) of 0.99 (95% CI 0.98 - 0.99), outperforming the standard logistic regression model, which exhibited a top-5 accuracy of 22.23% (95% CI 19.91% - 24.87%) and an AUC of 0.82 (95% CI 0.81 - 0.83). The contrastive learning enhancement also outperformed the logistic regression model, although slightly below the base FCN model, exhibiting a top-5 accuracy of 89.88% (95% CI 87.88% - 91.68%) and an AUC of 0.98 (95% CI 0.97 - 0.98).

Conclusions: This novel approach provides a scalable solution for harmonizing metadata across large-scale cohort studies. The proposed method significantly enhances the performance over the baseline method by using learned representations to categorize harmonized concepts more accurately for cohorts in cardiovascular disease and stroke.

(JMIR Form Res 2025;9:e75608) doi:10.2196/75608

KEYWORDS

harmonization; natural language processing; cardiovascular research; neural networks; multi-cohort studies

Introduction

The advent of large language models (LLMs), artificial intelligence, and computational power has the ability to transform our understanding of health and disease. One example is in developing predictive risk models for cardiovascular disease prevention, such as stroke [1]. Machine learning–based stroke risk prediction models enable the inclusion of a wide variety of factors (socioeconomic, behavioral, etc) to assess stroke risk [2]. To fully leverage these approaches and technology, datasets need to be integrated [3,4]. However, integration of datasets is challenging, given inconsistent variable names, column headers, and textual descriptions used to denote clinical or demographic measures [5,6].

These metadata variables, which are the textual labels describing data elements, often differ across studies, even when referring to the same underlying concept (eg, “Systolic_BP” vs “SBP_visit1”). In cardiovascular research, cohort datasets such as the Framingham Heart Study (FHS), the Multi-Ethnic Study of Atherosclerosis (MESA), and the Atherosclerosis Risk in Communities (ARIC) study include thousands of such variables, each with custom naming conventions and sparse documentation. This lack of standardization poses a major

XSL•FO
RenderX

challenge for dataset interoperability, phenotyping, and cross-cohort analyses [7].

Data harmonization is the process involving the standardization of disparate variables across multiple datasets into a cohesive and unified format [8,9]. This technique also increases the statistical power of a dataset to solve problems that could not be addressed when using data only from a single study [10]. Traditional harmonization approaches depend heavily on manual mapping by domain experts to map disparate variable descriptions into unified medical concepts, which is time-consuming, error-prone, and difficult to scale [10,11]. Standard vocabularies like Systematized Nomenclature of Medicine–Clinical Terms [12], Logical Observation Identifiers Names and Codes [13], *ICD* (*International Classification of Diseases*) codes [14,15], Current Procedural Terminology [16], Clinical Classifications Software [17], Normalized Names for Clinical Drugs [18], and National Drug Code [19] support structured data harmonization in electronic health records, but are not designed for the free-text, loosely formatted metadata descriptions found in cohort datasets. Recent advances in natural language processing (NLP), including the use of Bidirectional Encoder Representations from Transformers (BERT) models [20,21], knowledge network [22], and other semantic learning methods [23,24], offer promising opportunities to automate the process. Pretrained language models like Bidirectional Encoder Representations from Transformers for Biomedical Text Mining (BioBERT) and semantic embedding techniques can be adapted to understand and categorize medical text [21]. However, these models have not been widely applied to the harmonization of variable-level metadata in observational research settings. Our work addresses this gap.

The goal is to develop and evaluate an NLP-based method for harmonizing variable-level metadata across multiple biomedical datasets. Specifically, we aim to classify free-text variable names and descriptions into harmonized medical concepts that enable integration and analysis across multiple studies.

Methods

Overview

The goal of this approach was to combine different datasets by variable definitions into a harmonized variable defined as a medical concept—a term that describes information in a patient's medical record, such as a diagnosis, a prescription, or a measurement.

To do this, we treated the automation of harmonization as the following steps: (1) to select a list of predefined data harmonization biomedical concepts, and (2) to train a classifier to classify whether a variable belongs to a certain medical concept or not. We used 3 large-scale cardiovascular research cohort studies (ie, FHS, MESA, and ARIC) to harmonize cardiovascular disease risk variables.

For the second step, we used BioBERT embeddings with a fully connected neural network (FCN). BioBERT, a transformer language representation model pretrained on biomedical corpora, generates embeddings for variable descriptions, capturing their semantic relationships [21]. The FCN then classifies these embeddings into predefined harmonized concepts. To address the relatively low number of labeled samples, we also separately augmented the FCN using contrastive learning, a self-supervised representation learning method that is particularly effective in scenarios where training data is limited [25]. The process workflow for this approach is outlined in Multimedia Appendix 1.

Data Sources

We used the metadata from 3 research cohort datasets—FHS, MESA, and ARIC [7]. The metadata includes variable names and descriptions. In total, we extracted 885 variable descriptions categorized into 64 concepts (spread across 7 concept groups) through manual annotation by 3 independent reviewers, who adapted a preselected list of stroke-related concepts that were illustrated in our previous work [26]. The breakdown of each cohort dataset across cohorts and concept groups is provided in Table 1. The complete list of variable descriptions and their corresponding concepts is detailed in Multimedia Appendix 1. We used this labeled dataset for training and validation.

XSL•FO
RenderX

Table . Breakdown of the number of variables for each concept group across the 3 study cohorts: Framingham Heart Study, Multi-Ethnic Study of Atherosclerosis, and Atherosclerosis Risk in Communities. The 885 variable descriptions are categorized into 64 concepts across 7 concept groups via manual annotation.

Study	Variables		
	ARIC[a]	MESA[b]	FHS[c]
Total variable descriptions, n	315	161	409
Variables under each category of concept, n (%)			
Sociodemographics	12 (3.8)	11 (6.8)	13 (3.2)
Vitals	18 (5.7)	10 (6.2)	63 (15.4)
Comorbidities	59 (18.7)	76 (47.2)	98 (24)
Laboratories	32 (10.2)	16 (9.9)	49 (12)
Medications	131 (41.6)	30 (18.7)	91 (22.2)
Diet	42 (13.3)	1 (0.6)	74 (18.1)
Other	21 (6.7)	17 (10.6)	21 (5.1)

[a]ARIC: Atherosclerosis Risk in Communities.

[b]MESA: Multi-Ethnic Study of Atherosclerosis.

[c]FHS: Framingham Heart Study.

BioBERT Embeddings

We used a pretrained BioBERT model to convert the variable descriptions into embedding vectors. BioBERT is a transformer-based model specifically pretrained on large-scale biomedical corpora, including PubMed abstracts and PubMed Central articles [21]. Derived from a general-purpose model known as BERT [20], BioBERT has shown superior performance over BERT for biomedical-related tasks such as Named Entity Recognition [27], Relation Extraction [28], and Question Answering [29]. Particularly, for short-length sequences in the biomedical domain, with pretrained domain knowledge, BioBERT can capture domain-specific semantics and relationships better than a general-purpose model. Given its proven effectiveness in biomedical NLP tasks, BioBERT is an ideal choice for analyzing short-text sequences in the biomedical domain. In this study, we converted each variable description using BioBERT into a 768-dimensional embedding vector for downstream classification.

Paired Sentences for Classification

We framed the task as a binary classification problem using pairs of variable descriptions (x_1, x_2). Each pair was labeled as either belonging to the same concept or not. We calculated cosine similarity for each pair, and these similarity scores were used to train a supervised classifier to distinguish between matched and nonmatched pairs [30,31].

During inference, for a given variable description, the model compared it against all known concepts. The model calculated similarity scores for each pairing and assigned the description to the concept with the highest similarity score.

Data Preparation

The dataset was prepared as (1) matching pairs (for every concept, all combinations of variable descriptions belonging to that concept were generated as matched pairs) and (2) nonmatching pairs (for each variable description in a concept, a random sample of descriptions from other concepts was used to generate nonmatching pairs).

To balance the training dataset, we maintained a 1:3 ratio of matching to nonmatching pairs. This ensured sufficient representation of both types of data while maximizing training examples.

Models

We used the logistic regression model as a baseline classifier. The input was the cosine similarity between BioBERT embedding vectors of paired descriptions [32]. The model was trained using the cross-entropy loss function, and the output was a probabilistic score, which indicates whether the pair represented the same concept (eg, a matched pair or nonmatched pair).

The proposed FCN model consisted of 2 hidden layers, with the first hidden layer having a rectified linear unit activation function [33], and the second layer using a cosine similarity function, rescaled with a weight and a bias parameter, followed by a sigmoid activation function. The framework of the FCN model is outlined in Multimedia Appendix 1. The network was trained using binary cross-entropy loss [34]. The Adam optimizer with early stopping on the validation set was used for optimization [35]. During inference, given a new input variable description, the model calculated similarity scores between the embedding vectors for an input description and each known concept. The concept with the highest score was assigned to the variable description.

Contrastive Learning

To address the challenges of limited labeled training data, we used a contrastive learning approach [36]. The model was trained to minimize Noise-Contrastive Estimation loss, which improves the representation of variable descriptions by learning

XSL•FO
RenderX

from matched and nonmatched pairs [37]. For each variable description, we applied random permutations of embeddings to create augmented pairs. This method further optimized the FCN by leveraging noisy but informative examples. During inference, we used the same methodology as described for the FCN model to categorize an input variable description to a concept.

Evaluation

To assess the performance, applicability, and generalizability of the method, we used 2 strategies—a combined cohort approach and a separated cohort approach. In the combined cohort approach, we used data from all 3 cohort datasets and randomly split it into training, validation, and testing with an approximate ratio of 4:1:1. For the separated cohort approach, we trained and validated each model on 2 cohorts and used the remaining cohort for testing to assess generalizability across datasets.

We used the area under the receiver operating characteristic (AUC) as our primary performance measure distinguishing matched and nonmatched pairs [38,39]. To evaluate how often the correct concept ranks within the top-K predictions, we used top-1 and top-5 accuracy [40,41]. We used bootstrapping to obtain CIs for both the AUC and accuracy scores [42].

All models were developed using Python (v3.11.5) and PyTorch (v2.2.1). The code and trained models used are found on our GitHub repository: duke-harmonization.

Ethical Considerations

This study was approved by the Duke University Health System institutional review board (Pro00106364).

For the primary data collections, participants in the original studies provided informed consent, which included provisions for data sharing and secondary use. The datasets used in this study were accessed in accordance with those provisions, and no additional consent was required for this secondary analysis.

All datasets used in this study were fully deidentified and contained no direct or indirect identifiers. The analyses relied exclusively on aggregated metadata, with no linkage to individual-level information. Accordingly, participant confidentiality was maintained throughout.

No participants were directly involved or recruited for this secondary analysis; therefore, no compensation was provided. This paper does not include any images or materials that could lead to the identification of individual participants.

Results

We extracted a total of 885 variables from 3 datasets, including the FHS, MESA, and ARIC. We precategorized these variables into 64 harmonized concepts and generated 58,890 sentence pairs. In the combined cohort evaluation strategy, we split this dataset into training, validation, and test datasets in a 4:1:1 ratio. The FCN model outperformed the baseline logistic regression model, achieving an AUC of 0.99 (95% CI 0.98 - 0.99), compared with the baseline's AUC of 0.82 (95% CI 0.81 - 0.83). The contrastive learning approach achieved an AUC of 0.98 (95% CI 0.97 - 0.98), which also outperformed the baseline logistic regression model (Figure 1). For the top-K accuracy, the FCN model achieved a top-1 accuracy of 80.51% (95% CI 78.08% - 83.03%) and a top-5 accuracy of 98.95% (95% CI 98.31% - 99.47%), significantly outperforming the baseline model which achieved top-1 accuracy of 12.12% (95% CI 10.22% - 14.12%) and top-5 accuracy of 22.23% (95% CI 19.91% - 24.87%). The contrastive learning approach achieved a moderate top-1 accuracy score of 63.65% (95% CI 60.59% - 66.81%) and achieved a top-5 accuracy score of 89.88% (95% CI 87.88% - 91.67%; Table 2.

XSL•FO
RenderX

Figure 1. Receiver operating characteristic curves for each of the trained fully connected neural network models and the baseline logistic regression model for the combined cohort approach. In this setting, the variables from all three datasets (Atherosclerosis Risk in Communities, Multi-Ethnic Study of Atherosclerosis, and Framingham Heart Study) were pre-categorized into harmonized concepts. The area under the curve is directly proportional to the model's performance in distinguishing between matches and nonmatches for a given pair of variable descriptions. The data used to generate the receiver operating characteristic curves consisted of 11,880 pairs of variable descriptions that were absent from the training data, when evaluated on all the cohorts. AUC: area under the receiver operating characteristic; FCN: fully connected neural network.

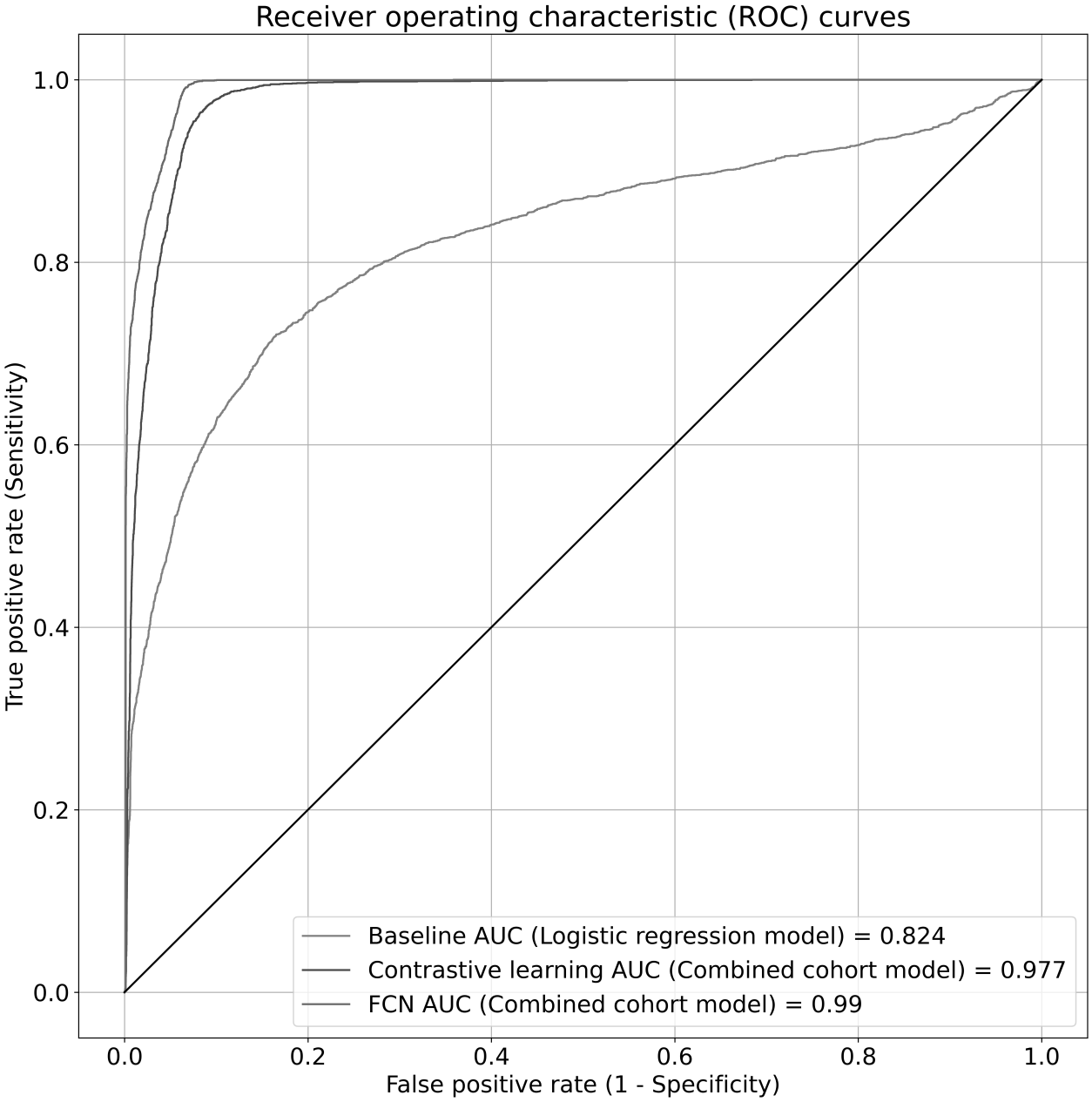

Table . Top-1 and top-5 accuracy with 95% CIs for baseline logistic regression model, the fully connected neural network model, and the fully connected neural network model with contrastive learning. The evaluation was performed under the combined cohort strategy, where the variables from all 3 cohorts (Atherosclerosis Risk in Communities, Multi-Ethnic Study of Atherosclerosis, and Framingham Heart Study) were precategorized into harmonized concepts.

Model	Top-1 accuracy, % (95% CI)	Top-5 accuracy, % (95% CI)	AUC[a] (95% CI)
Logistic regression	12.12 (10.22 - 14.12)	22.23 (19.91 - 24.87)	0.82 (0.81 - 0.83)
FCN[b] (combined cohort)	80.51 (78.08 - 83.03)	98.95 (98.31 - 99.47)	0.99 (0.98 - 0.99)
Contrastive learning	63.65 (60.59 - 66.81)	89.88 (87.88 - 91.67)	0.98 (0.97 - 0.98)

[a]AUC: area under the receiver operating characteristic.

[b]FCN: fully connected neural network.

XSL•FO
RenderX

We assessed the robustness of FCN in the separated cohort evaluation. The FCN model trained with the ARIC-Framingham model achieved an AUC of 0.78 (95% CI 0.73 - 0.83) on the MESA dataset. The MESA-ARIC model, evaluated on the Framingham dataset, achieved the highest AUC of 0.85 (95% CI 0.83 - 0.87). The Framingham-MESA model, evaluated on the ARIC dataset, achieved an AUC of 0.83 (95% CI 0.81 - 0.85). The ROC curves for the separated cohort models are shown in Figure 2. For the top-K metrics, the ARIC-Framingham model performed best with a top-1 accuracy of 49.33% (95% CI 43.11% - 55.11%) and a top-5 accuracy of 64% (95% CI 57.78% - 69.34%). The MESA-ARIC model performed slightly worse, with a top-1 accuracy of 39.32% (95% CI 35.09% - 43.76%) and a top-5 accuracy of 59.62% (95% CI 55.39% - 64.06%). The Framingham-MESA model exhibited the lowest accuracy performance, with a top-1 accuracy of 32.98% (95% CI 28.23% - 37.47%) and a top-5 accuracy of 48.81% (95% CI 43.79% - 53.56%), which were likely due to greater variability in the ARIC dataset (Table 3).

Figure 2. Receiver operating characteristic curves for each of the trained fully connected neural network models for the separated cohort approach. In this setting, the variables from all 3 datasets (Atherosclerosis Risk in Communities, Multi-Ethnic Study of Atherosclerosis, and Framingham Heart Study) were initially precategorized into harmonized concepts. The models were then trained and validated on 2 of the cohorts and then tested on the remaining cohort to assess generalizability of the model across different datasets. The area under the curve is directly proportional to the model's performance in distinguishing between matches and nonmatches for a given pair of variable descriptions. The receiver operating characteristic curves for each model were obtained by evaluating the model on the subset of the test dataset containing only data from the cohort excluded during training. ARIC: Atherosclerosis Risk in Communities; AUC: area under the receiver operating characteristic; FCN: fully connected neural network; MESA: Multi-Ethnic Study of Atherosclerosis.

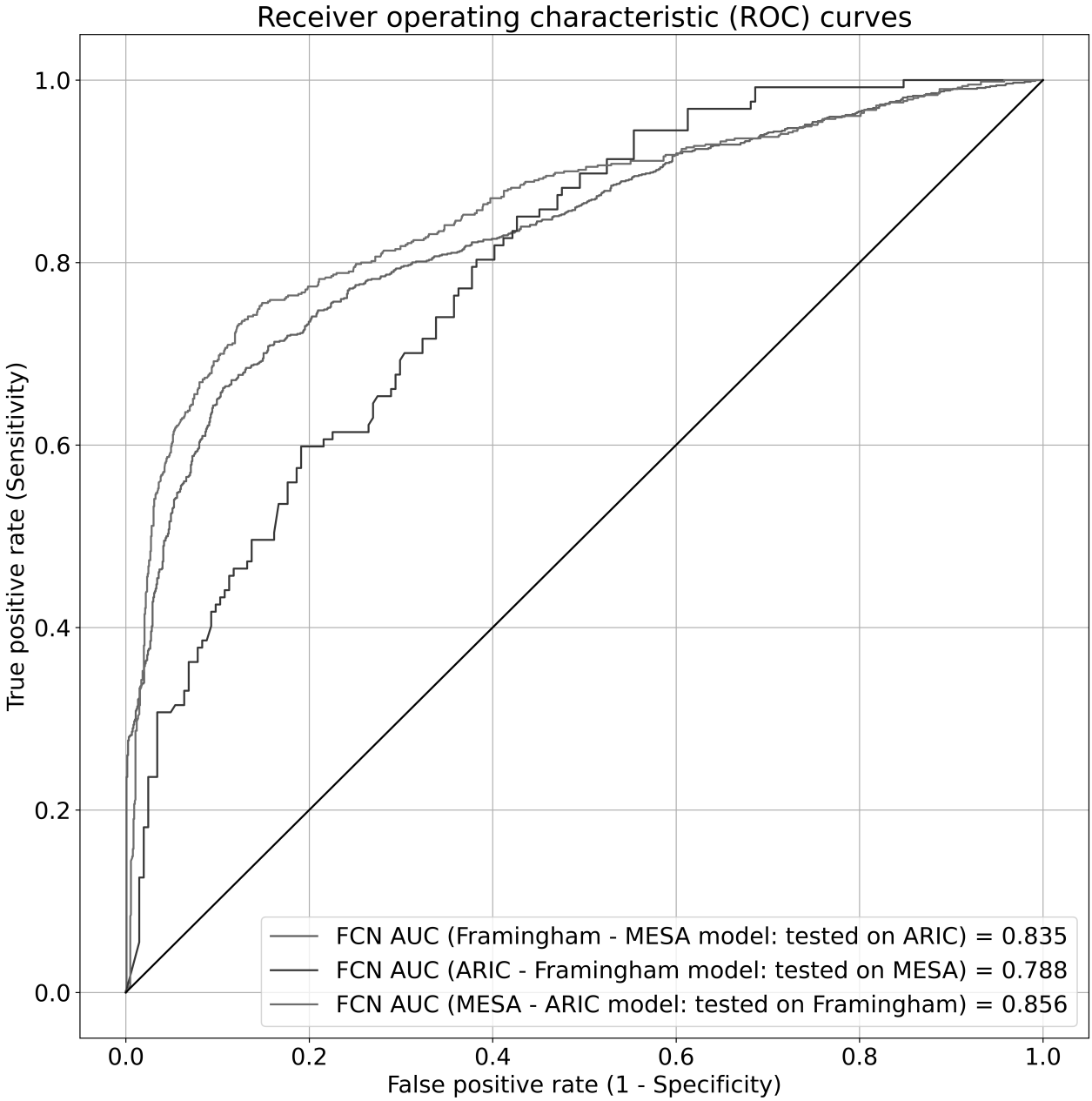

XSL•FO
RenderX

Table . Top-1 and top-5 accuracy with 95% CIs for the 3 cohort-specific fully connected neural network models. The evaluation was performed using the separated cohort evaluation strategy, where the variables from all three cohorts (Atherosclerosis Risk in Communities, Multi-Ethnic Study of Atherosclerosis, and Framingham Heart Study) were initially precategorized into harmonized concepts, and the models were then trained and validated on 2 cohorts and tested on the remaining cohort to assess generalizability of the model across different datasets.

Model	Top-1 accuracy, % (95% CI)	Top-5 accuracy, % (95% CI)	AUC[a] (95% CI)
FCN[b] (Framingham-MESA[c]), tested on ARIC	32.98 (28.23 - 37.47)	48.81 (43.79 - 53.56)	0.83 (0.81 - 0.85)
FCN (MESA-ARIC[d]) tested on Framingham	39.32 (35.09 - 43.76)	59.62 (55.39 - 64.06)	0.85 (0.83 - 0.87)
FCN (ARIC-Framingham) tested on MESA	49.33 (43.11 - 55.11)	64.0 (57.78 - 69.34)	0.78 (0.73 - 0.83)

[a]AUC: area under the receiver operating characteristic.

[b]FCN: fully connected neural network.

[c]MESA: Multi-Ethnic Study of Atherosclerosis.

[d]ARIC: Atherosclerosis Risk in Communities.

We plotted the distribution of the predicted score for matches and nonmatches across different concept groups using the baseline method and the FCN model, illustrated in Multimedia Appendix 1. The results indicated that the FCN model generally demonstrates narrower IQRs and more distinct separation between median probabilities for matches and nonmatches, particularly in the diet and sociodemographics categories, which achieved a perfect AUC of 1.0, indicating superior predictive performance compared with the baseline model. The AUC for each model setting when evaluated on a per-concept level is detailed in Multimedia Appendix 1.

We also computed the positive predictive value, negative predictive value, true positive rate, and false positive rate for each of the concepts when using the top-1 predicted concept from the FCN model on the test dataset. The mean positive predictive value across all concepts was 0.78 (SD 0.25), the mean negative predictive value was 0.99 (SD 0.01), the mean true positive rate was 0.85 (SD 0.21), and the mean false positive rate was 0.01 (SD 0.01). The metrics for all concepts are detailed in Multimedia Appendix 1.

Discussion

Overview

Harmonizing multiple diverse research cohort datasets can enlarge the data power for training and validating risk prediction models. However, traditional data harmonization techniques need manual comparison, which is time-consuming and barely scalable. This study presents an automated and scalable approach for variable harmonization by leveraging domain-specific NLP and machine learning applied to metadata. We implemented and evaluated the method using the metadata-level variable descriptions from the 3 National Institutes of Health research cohort studies. By reframing variable harmonization as a sentence-pair classification problem, our approach achieves accurate mapping between free-text variable descriptions and standardized concepts, even in the absence of patient-level data. This methodology addresses the common challenges of short text length, sparse annotation, and class imbalance in harmonization tasks.

Principal Results and Comparison With Previous Work

Our results showed that the FCN model trained on sentence pairs significantly outperformed the baseline logistic regression model. Specifically, both the basic FCN method and the enhanced version using contrastive learning achieved high AUC, top-1, and top-5 accuracy scores, surpassing the logistic regression method. The basic FCN model performed slightly better than the contrastive learning approach. We further assessed the generalizability of our model by separating cohorts for evaluation. Model performance was generally lower and varied, which is expected, due to different variable distributions across different research cohort datasets. The ARIC-Framingham model performed the best in terms of top-K accuracy, suggesting that the MESA dataset shared the most common metadata features with the other two. The Framingham-MESA model performed the worst, possibly because the ARIC metadata has more unique characteristics and models could not effectively learn due to its absence from the training data.

Similar to earlier manual harmonization efforts, our approach began with expert-curated categorization of variables into predefined concepts, which is a foundational step that was essential for the success of the automated classification process, as described in our previous work [26]. However, unlike traditional methods that rely heavily on manual effort throughout, our system automates the subsequent classification, significantly reducing the time and human effort required. While manual harmonization provides expert-driven accuracy, our findings suggest that the automated method can achieve comparable mapping quality with substantially less human input. This framework aligns with practices seen in other harmonization studies, where domain experts played a key role in defining variable concepts [9,43], and other automated harmonization studies [44,45].

Our proposed approach for automated variable harmonization used pretrained embeddings to learn the representations from variable descriptions. Similarly, Yang et al [45] used semantic embeddings and patient-level data to harmonize continuous variables. However, their approach excluded categorical variables and those with missing data. In contrast, our approach

XSL•FO
RenderX

uses only variable metadata, thus enabling harmonizing a broader spectrum of variables including both continuous and categorical. Since we use only metadata, this approach also allows harmonization of datasets with or without missing data—thus offering wider applicability for real-world cohort integration.

With the recent advancements in LLMs, Li et al [44] introduced a framework for variable matching using embeddings from general-purpose LLMs. We acknowledge this emerging direction in the field; however, the use of large models often requires fine-tuning on domain-specific data and incurs substantial computational costs, which may limit their practical applicability in resource-constrained settings [46]. By leveraging embeddings from domain-specific LLMs, such as BioBERT, we present a cost-effective approach, requiring fewer computational resources for training and implementation [47].

Implications for Research and Practice

Our experimental results suggest that our method achieves accurate harmonization for variables across different cardiovascular cohort studies by evaluating contextual similarity across disparate variable descriptions. For example, the descriptions of "diabetes mellitus status" are inconsistent across ARIC, MESA, and Framingham datasets. In the ARIC study, the description varies by visit or exam, such as "diabetes with fasting glucose cutpoint<126" or "diabetes using lower cutpoint 126 mg/dL." In contrast, Framingham and MESA use descriptions like "diabetes mellitus status, exam 1." Traditionally, aligning these variables to a SNOMED concept for the condition "diabetes mellitus" requires manual effort and domain expertise, which is difficult to scale across multiple cohorts. Our automated framework significantly reduces this burden, achieving consistent, accurate mapping in a fraction of the time. In practical settings, this approach enables researchers to integrate datasets for cross-cohort analyses, which are essential for predictive modeling and other data-driven applications.

Limitations

Despite these advancements, we acknowledge that several limitations and challenges remain. First, our proposed framework focused on metadata and did not include patient-level data. However, incorporating patient-level data could help resolve ambiguities in variable definitions. A hybrid approach that leverages patient-level data alongside learned representations from the metadata may help in verifying the automated harmonization results [22]. Another limitation is that we did not address the challenges remaining in the harmonization of different units for laboratory values given that our focus was on metadata and variable descriptions. Incorporating comparisons of variable distributions from patient-level data, in addition to the semantic representations of the variable descriptions, could help alleviate this problem [48]. Future work should explore hybrid methods to combine harmonized variable descriptions with patient-level data to create a more comprehensive and robust framework for cohort integration.

While our study focused on cardiovascular datasets, we acknowledge that the generalizability of the proposed harmonization method to other disease domains or datasets with differing data structures remains unproven. BioBERT is pretrained on large-scale biomedical corpora and thus has potential applicability beyond cardiovascular disease [49], but we recommend validating this approach in other domains, such as oncology, infectious disease, and mental health, where vocabulary, annotation practices, and data sparsity may vary. To improve robustness and portability, we recommend curating preharmonized benchmark datasets for external validation. In addition, future work could explore the integration of lightweight transformers, few-shot learning, or domain-adaptive transfer learning to handle limited labeled data and further extend the applicability of contrastive learning in diverse biomedical settings [50-52].

Third, we did not use more sophisticated models for sequential data such as recurrent neural networks [53], or Long Short-Term Memory networks [54], nor LLMs such as Generative Pre-trained Transformers [55], Pathways Language Model (Google AI) [56], or Large Language Model Meta AI [57], due to the sparse number of labeled examples present in our training data. Application of the contrastive learning approach in tandem with the advanced language models may prove effective in the scalability of the automated harmonization process. Using more complex batch selection methods may also lead to better results via contrastive learning [58].

Conclusions

In this study, we developed a scalable and automated method for variable harmonization using only metadata from research cohorts. By applying domain-specific language models and framing the task as a sentence-pair classification problem, our approach can accurately map variable descriptions to standardized concepts without needing patient-level data. This reduces the time and effort required for harmonization and is especially useful when access to detailed data is limited. Although we tested the method on cardiovascular datasets, it can potentially be used in other areas like cancer or mental health research. This work provides a foundation for faster and more efficient data integration, which is important for large-scale studies and real-world health research.

Acknowledgments

This work was funded by the National Institute of Neurological Disorders and Stroke (NINDS; grant R61/R33NS120246).

The Framingham Heart Study (FHS) is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with Boston University (Contract N01-HC-25195 and HHSN268201500001I). This manuscript was not prepared in collaboration with investigators of the FHS and does not necessarily reflect the opinions or views of the FHS, Boston University, or NHLBI. The original metadata used in this work can be found at dbGaP, using the dbGaP accession number phs000007.v32.

XSL•FO
RenderX

Multi-Ethnic Study of Atherosclerosis (MESA) and the MESA SHARe project are conducted and supported by the NHLBI in collaboration with MESA investigators. Support for MESA is provided by contracts HHSN268201500003I, N01-HC-95159, N01-HC-95160, N01-HC-95161, N01-HC-95162, N01-HC-95163, N01-HC95164, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169, UL1-TR-001079, UL1-TR000040, UL1-TR-001420, UL1-TR-001881, DK063491 and CTSA UL1-RR-024156. The original metadata used in this work can be found at dbGaP using the dbGaP accession phs000209.v13.

The Atherosclerosis Risk in Communities (ARIC) study has been funded in whole or in part with Federal funds from the NHLBI, National Institute of Health, Department of Health and Human Services, under contracts (HHSN268201700001I, HHSN268201700002I, HHSN268201700003I, HHSN268201700004I, and HHSN268201700005I). The authors thank the staff members and participants of the ARIC study for their important contributions. The original metadata used in this work can be found at the database of Genotypes and Phenotypes (dbGaP) using the dbGaP accession phs000280.v7.

The metadata for FHS, MESA, and ARIC can also be obtained from the NHLBI Biologic Specimen and Data Repository Information Coordinating Center (BioLINCC). BioLINCC does not necessarily reflect the opinions or views of the FHS, MESA, ARIC, or NHLBI. This work uses only the metadata from the FHS, MESA, and ARIC studies.

Data Availability

The datasets generated are not publicly available as we are unable to share the harmonized data as we do not have the right to grant access to individual-level data. This requires a signed data-use agreement from the data cohorts (FHS, MESA, and ARIC) via the database of Genotypes and Phenotypes (dbGaP) or via the Biologic Specimen and Data Repository Information Coordinating Center (BioLINCC). If you are interested in acquiring the data, please contact dbGaP or BioLINCC. The code used to create the concepts for manual harmonization of the cohorts can be found on our GitHub repository. The training metadata dataset containing the variable descriptions and their assigned concepts can be found in the supplementary materials (Multimedia Appendix 1). The code and the trained models used for our proposed automated harmonization method are also found on our GitHub repository.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional material.
[DOCX File, 505 KB - formative_v9i1e75608_app1.docx]

References

1. Tsao CW, Aday AW, Almarzooq ZI, et al. Heart disease and stroke statistics-2023 update: a report from the American Heart Association. Circulation 2023 Feb 21;147(8):e93-e621. [doi: 10.1161/CIR.0000000000001123] [Medline: 36695182]
2. Jamthikar A, Gupta D, Saba L, et al. Cardiovascular/stroke risk predictive calculators: a comparison between statistical and machine learning models. Cardiovasc Diagn Ther 2020 Aug;10(4):919-938. [doi: 10.21037/cdt.2020.01.07] [Medline: 32968651]
3. Zhao J, Feng Q, Wei WQ. Integration of omics and phenotypic data for precision medicine. In: Bai JPF, Hur J, editors. Methods Mol Biol 2022;2486:19-35. [doi: 10.1007/978-1-0716-2265-0_2] [Medline: 35437716]
4. Johnson KB, Wei W, Weeraratne D, et al. Precision medicine, AI, and the future of personalized health care. Clinical Translational Sci 2021 Jan;14(1):86-93 [FREE Full text] [doi: 10.1111/cts.12884]
5. Gurugubelli VS, Fang H, Shikany JM, et al. A review of harmonization methods for studying dietary patterns. Smart Health (2014) 2022 Mar;23:100263. [doi: 10.1016/j.smhl.2021.100263]
6. Peng Y, Bathelt F, Gebler R, et al. Use of metadata-driven approaches for data harmonization in the medical domain: scoping review. JMIR Med Inform 2024 Feb 14;12:e52967. [doi: 10.2196/52967] [Medline: 38354027]
7. Mallya P, Stevens LM, Zhao J, et al. Facilitating harmonization of variables in Framingham, MESA, ARIC, and REGARDS studies through a metadata repository. Circ: Cardiovascular Quality and Outcomes 2023 Nov;16(11):11. [doi: 10.1161/CIRCOUTCOMES.123.009938]
8. Cheng C, Messerschmidt L, Bravo I, et al. A general primer for data harmonization. Sci Data 2024 Jan 31;11(1):152. [doi: 10.1038/s41597-024-02956-3] [Medline: 38297013]
9. Pan K, Bazzano LA, Betha K, et al. Large-scale data harmonization across prospective studies. Am J Epidemiol 2023 Nov 10;192(12):2033-2049. [doi: 10.1093/aje/kwad153] [Medline: 37403415]
10. Adhikari K, Patten SB, Patel AB, et al. Data harmonization and data pooling from cohort studies: a practical approach for data management. Int J Popul Data Sci 2021;6(1):1680. [doi: 10.23889/ijpds.v6i1.1680] [Medline: 34888420]
11. Sony P. Concept-based electronic health record retrieval system in healthcare IOT. In: Cognitive Informatics and Soft Computing: Springer; 2019, Vol. 768:175-188. [doi: 10.1007/978-981-13-0617-4_17]

XSL•FO
RenderX

12. Ruch P, Gobeill J, Lovis C, Geissbühler A. Automatic medical encoding with SNOMED categories. BMC Med Inform Decis Mak 2008 Oct 27;8 Suppl 1(Suppl 1):S6. [doi: 10.1186/1472-6947-8-S1-S6] [Medline: 19007443]
13. McDonald CJ, Huff SM, Suico JG, et al. LOINC, a universal standard for identifying laboratory observations: a 5-year update. Clin Chem 2003 Apr;49(4):624-633. [doi: 10.1373/49.4.624] [Medline: 12651816]
14. W. H. Organization and others, International classification of diseases. Basic Tabulation List with Alphabetic Index: World Health Organization; 1978.
15. W. H. Organization, International Statistical Classification of Diseases and related health problems. Alphabetical Index: World Health Organization; 2004, Vol. 3.
16. Abraham M, Ahlman JT, Boudreau AJ, et al. CPT 2011 Standard Edition, 4th edition: American Medical Association; 2010.
17. Clinical Classifications Software (CCS) for ICD-9-CM. URL: https://www.oit.va.gov/Services/TRM/ToolPage.aspx?tid=7602 [accessed 2024-12-09]
18. Bennett CC. Utilizing RxNorm to support practical computing applications: capturing medication history in live electronic health records. J Biomed Inform 2012 Aug;45(4):634-641. [doi: 10.1016/j.jbi.2012.02.011] [Medline: 22426081]
19. National Drug Code Directory. U.S. Food and Drug Association. URL: https://www.fda.gov/drugs/drug-approvals-and-databases/national-drug-code-directory [accessed 2024-12-09]
20. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv. Preprint posted online on May 24, 2019 URL: http://arxiv.org/abs/1810.04805 [accessed 2025-08-18] [doi: 10.48550/arXiv.1810.04805]
21. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb 15;36(4):1234-1240. [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
22. Hong C, Rush E, Liu M, et al. Clinical knowledge extraction via sparse embedding regression (KESER) with multi-center large scale electronic health record data. NPJ Digit Med 2021 Oct 27;4(1):151. [doi: 10.1038/s41746-021-00519-z] [Medline: 34707226]
23. Kartchner D, Christensen T, Humpherys J, Wade S. Code2Vec: embedding and clustering medical diagnosis data. Presented at: 2017 IEEE International Conference on Healthcare Informatics (ICHI); Aug 23-26, 2017; Park City, UT, USA p. 386-390. [doi: 10.1109/ICHI.2017.94]
24. Choi E, Bahadori MT, Searles E, et al. Multi-layer representation learning for medical concepts. : ACM Presented at: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Aug 13, 2016; San Francisco California USA p. 1495-1504. [doi: 10.1145/2939672.2939823]
25. Hénaff OJ, et al. Data-efficient image recognition with contrastive predictive coding. . Preprint posted online on 2019. [doi: 10.48550/ARXIV.1905.09272]
26. Hong C, Pencina MJ, Wojdyla DM, et al. Predictive accuracy of stroke risk prediction models across Black and White race, sex, and age groups. JAMA 2023 Jan 24;329(4):306-317. [doi: 10.1001/jama.2022.24683] [Medline: 36692561]
27. Habibi M, Weber L, Neves M, Wiegandt DL, Leser U. Deep learning with word embeddings improves biomedical named entity recognition. Bioinformatics 2017 Jul 15;33(14):i37-i48. [doi: 10.1093/bioinformatics/btx228] [Medline: 28881963]
28. Lin C, Miller T, Dligach D, Bethard S, Savova G. A BERT-based universal model for both within-and cross-sentence clinical temporal relation extraction. Presented at: Proceedings of the 2nd Clinical Natural Language Processing Workshop; Jun 2019; Minneapolis, Minnesota, USA p. 65-71. [doi: 10.18653/v1/W19-1908]
29. Wiese G, Weissenborn D, Neves M. Neural domain adaptation for biomedical question answering. Presented at: Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017); 2017; Vancouver, Canada. [doi: 10.18653/v1/K17-1029]
30. Sutton A, Cristianini N. On the learnability of concepts: with applications to comparing word embedding algorithms. In: Maglogiannis I, Iliadis L, Pimenidis E, editors. Artificial Intelligence Applications and Innovations: Springer International Publishing; 2020, Vol. 584:420-432. [doi: 10.1007/978-3-030-49186-4_35]
31. Mueller J, Thyagarajan A. Siamese recurrent architectures for learning sentence similarity. Presented at: Proceedings of the AAAI conference on artificial intelligence; 2016. [doi: 10.1609/aaai.v30i1.10350]
32. Corley C, Mihalcea R. Measuring the semantic similarity of texts. In: Dolan B, Dagan I, editors. Presented at: Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment; Jun 18, 2005; Ann Arbor, Michigan. [doi: 10.3115/1631862.1631865]
33. Nair V, Hinton GE. Rectified linear units improve restricted boltzmann machines. Presented at: Proceedings of the 27th International Conference on Machine Learning (ICML-10); 2010; Madison, Wisconsin, USA p. 807-814 URL: https://dl.acm.org/doi/10.5555/3104322.3104425 [accessed 2025-08-18]
34. Zhang Z, Sabuncu MR. Generalized cross entropy loss for training deep neural networks with noisy labels. Adv Neural Inf Process Syst 2018 Dec;32:8792-8802. [Medline: 39839708]
35. Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv. Preprint posted online on 2014.
36. Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. arXiv. Preprint posted online on Feb 13, 2020 URL: http://arxiv.org/abs/2002.05709 [accessed 2025-08-18] [doi: 10.48550/arXiv.2002.05709]

XSL•FO
RenderX

37. Li Y, Vinyals O. Representation learning with contrastive predictive coding. . Preprint posted online on 2018. [doi: 10.48550/ARXIV.1807.03748]
38. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit DAGM 1997 Jul;30(7):1145-1159. [doi: 10.1016/S0031-3203(96)00142-2]
39. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982 Apr;143(1):29-36. [doi: 10.1148/radiology.143.1.7063747] [Medline: 7063747]
40. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. J Mach Learn Res 2011;12:2825-2830 [FREE Full text]
41. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Adv Neural Inf Process Syst 2012, Vol. 25.
42. Efron B, Tibshirani RJ. An introduction to the bootstrap. In: Chapman and Hall/CRC 1994. [doi: 10.1201/9780429246593]
43. Spjuth O, Krestyaninova M, Hastings J, et al. Harmonising and linking biomedical and clinical data across disparate data archives to enable integrative cross-biobank research. Eur J Hum Genet 2016 Apr;24(4):521-528. [doi: 10.1038/ejhg.2015.165] [Medline: 26306643]
44. Li Z, Prabhu SP, Popp ZT, et al. A natural language processing approach to support biomedical data harmonization: leveraging large language models. PLoS ONE 2025;20(7):e0328262. [doi: 10.1371/journal.pone.0328262] [Medline: 40705832]
45. Yang D, Zhou D, Cai S, et al. Robust automated harmonization of heterogeneous data through ensemble machine learning: algorithm development and validation study. JMIR Med Inform 2025 Jan 22;13:e54133. [doi: 10.2196/54133] [Medline: 39844378]
46. Gu Y, Tinn R, Cheng H, et al. Domain-specific language model pretraining for biomedical natural language processing. ACM Trans Comput Healthcare 2022 Jan 31;3(1):1-23. [doi: 10.1145/3458754]
47. Peng Y, Yan S, Lu Z. Transfer learning in biomedical natural language processing: an evaluation of BERT and elmo on ten benchmarking datasets. : Association for Computational Linguistics Presented at: Proceedings of the 18th BioNLP Workshop and Shared Task; 2019; Florence, Italy p. 58-65. [doi: 10.18653/v1/W19-5006]
48. Bradwell KR, Wooldridge JT, Amor B, et al. Harmonizing units and values of quantitative data elements in a very large nationally pooled electronic health record (EHR) dataset. J Am Med Inform Assoc 2022 Jun 14;29(7):1172-1182. [doi: 10.1093/jamia/ocac054] [Medline: 35435957]
49. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature New Biol 2023 Aug;620(7972):172-180. [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
50. Song Y, Wang T, Cai P, Mondal SK, Sahoo JP. A comprehensive survey of few-shot learning: evolution, applications, challenges, and opportunities. ACM Comput Surv 2023 Dec 31;55(13s):1-40. [doi: 10.1145/3582688]
51. Zhuang F, Qi Z, Duan K, et al. A comprehensive survey on transfer learning. Proc IEEE 2021 Jan;109(1):43-76. [doi: 10.1109/JPROC.2020.3004555]
52. Rohanian O, Nouriborji M, Kouchaki S, Clifton DA. On the effectiveness of compact biomedical transformers. Bioinformatics 2023 Mar 1;39(3):btad103. [doi: 10.1093/bioinformatics/btad103] [Medline: 36825820]
53. Graves A, Mohamed AR, Hinton G, Mohamed AR. Speech recognition with deep recurrent neural networks. Presented at: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing; May 26-31, 2013; Vancouver, BC, Canada p. 6645-6649. [doi: 10.1109/ICASSP.2013.6638947]
54. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997 Nov 15;9(8):1735-1780. [doi: 10.1162/neco.1997.9.8.1735] [Medline: 9377276]
55. Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners. arXiv. Preprint posted online on 2020. [doi: 10.48550/ARXIV.2005.14165]
56. Chowdhery A, Narang S, Devlin J, et al. PaLM: scaling language modeling with pathways. J Mach Learn Res 2023 Aug:1-113 [FREE Full text]
57. Touvron H, Lavril T, Izacard G, et al. LLaMA: open and efficient foundation language models. arXiv. Preprint posted online on Feb 27, 2023. [doi: 10.48550/ARXIV.2302.13971]
58. Kanakarajan KR, Kundumani B, Abraham A, Sankarasubbu M. BioSimCSE: biomedical sentence embeddings using contrastive learning. Presented at: Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI); 2022; Abu Dhabi, United Arab Emirates (Hybrid p. 81-86. [doi: 10.18653/v1/2022.louhi-1.10]

Abbreviations

ARIC: Atherosclerosis Risk in Communities
AUC: area under the receiver operating characteristic
BERT: Bidirectional Encoder Representations from Transformers
BioBERT: Bidirectional Encoder Representations from Transformers for Biomedical Text Mining
FCN: fully connected neural network
FHS: Framingham Heart Study

XSL•FO
RenderX

ICD*: *International Classification of Diseases
LLM: large language model
MESA: Multi-Ethnic Study of Atherosclerosis
NLP: natural language processing

Edited by A Mavragani; submitted 07.04.25; peer-reviewed by AA Arafat, SM Shaffi; revised version received 12.06.25; accepted 14.06.25; published 27.08.25.

Please cite as:
Mallya P, Henao R, Hong C, Wojdyla D, Schibler T, Manchanda V, Pencina M, Hall J, Zhao J
Automated Data Harmonization in Clinical Research: Natural Language Processing Approach
JMIR Form Res 2025;9:e75608
URL: https://formative.jmir.org/2025/1/e75608
doi:10.2196/75608

© Pratheek Mallya, Ricardo Henao, Chuan Hong, Daniel Wojdyla, Tony Schibler, Vihaan Manchanda, Michael Pencina, Jennifer Hall, Juan Zhao. Originally published in JMIR Formative Research (https://formative.jmir.org), 27.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

The MyRelief Digital Educational Self-Management Program for Persistent Low Back Pain: Feasibility Uncontrolled Trial

Caroline Larsson[1], PhD; Joanne Marley[2], PhD; Flavia Piccinini[3], MSc; Sarah Howes[4], PhD; Elisa Casoni[3], MSc; Vincenzo Aschettino[5], BSc; Carlos Vaz de Carvalho[6], PhD; Suzanne McDonough[7], PhD

[1]Department of Health Sciences, Lund University, Sölvegatan 19, Lund, Sweden
[2]School of Health Sciences, Ulster University, Belfast, United Kingdom
[3]IRCCS INRCA - Istituto Nazionale Ricovero e Cura Anziani (National Institute of Health and Science on Ageing), Ancona, Italy
[4]School of Health Sciences, Ulster University, Derry-Londonderry, United Kingdom
[5]Tech4Care srl, Ancona, Italy
[6]Instituto Superior de Engenharia do Porto, Porto, Portugal
[7]RCSI University of Medicine and Health Sciences, Dublin, Ireland

Corresponding Author:
Caroline Larsson, PhD
Department of Health Sciences, Lund University, Sölvegatan 19, Lund, Sweden

Abstract

Background: Low back pain (LBP) is a leading cause of work absence globally. Digital interventions have the potential to increase access to self-management support for individuals with persistent LBP.

Objective: This study aims to evaluate the feasibility, usability, and acceptability of a digital educational program (MyRelief) designed to support self-management strategies for people with persistent LBP.

Methods: A prospective uncontrolled feasibility study was conducted across 4 countries (Italy, Portugal, Sweden, and the United Kingdom) between 2020 and 2021. Adults in employment with nonspecific persistent LBP (>3 mo) with access to the internet were eligible to participate. Participants were given access to MyRelief, an 8-unit evidence-based educational self-management program. The feasibility of the MyRelief program was assessed using recruitment rates, an a priori success threshold of >70% of the target sample (50 participants), and a retention <35% dropout rate. Pre- and postintervention measures of functional disability were assessed using the Oswestry Disability Index (ODI), and health-related quality of life using the 5-level EuroQol questionnaire. Additional postintervention measures included the Patient Enablement Instrument and the System Usability Scale. Quantitative data were analyzed descriptively, and qualitative feedback was analyzed using a reflexive analytical approach.

Results: The recruitment feasibility threshold was met, and 40/50 (80%) participants (19 male and 21 female; mean age 57 years) were enrolled in the study. A total of 17 participants (11 male and 6 female) completed both the baseline and 12-week follow-up questionnaires. This represented a retention rate of 42.5% (17/40) and a dropout rate of 57.5%, which did not meet the a priori criteria of <35% dropouts. Approximately half of the participants presented with low baseline disability scores (mean ODI 24.0; 95% CI 18 - 31) with no significant change at follow-up (mean ODI 23.9; 95% CI 16 - 31). The 5-level EuroQol questionnaire scores improved from 0.68 (95% CI 0.608 - 0.76) to 0.72 (95% CI 0.66 - 0.79), indicating a clinically significant change. Patient Enablement Instrument scores postintervention were high (mean 5.31), indicating good perceived enablement. The mean System Usability Scale score was 72.4 (95% CI 67.5 - 73.3), indicating a good level of perceived ease-of-use. Overall, the quality of outcome measure completion was high (100%). Qualitative feedback indicated areas for improvement relating to challenges around access and navigation within the website.

Conclusions: The MyRelief study demonstrated feasibility in terms of recruitment but not retention. However, low baseline disability levels are not representative of the wider persistent LBP population. Future studies should broaden recruitment strategies, in particular, by recruiting from health care settings to improve representativeness. Although usability met industry standards, qualitative feedback suggests that navigation and accessibility require further optimization to better align with end user preferences for digital health interventions.

Trial Registration: ClinicalTrial.gov NCT04673773; https://clinicaltrials.gov/study/NCT04673773

International Registered Report Identifier (IRRID): RR2-10.1186/s40814-023-01382-3

(JMIR Form Res 2025;9:e74948) doi:10.2196/74948

XSL•FO
RenderX

KEYWORDS

low back pain; feasibility; self-management; online education; mooc; massive open online course

Introduction

Overview

Persistent low back pain (LBP) is among the most prevalent, disabling, and expensive conditions to treat [1]. The 2023 Global Burden of Disease study estimates that 619 million people lived with LBP in 2020, with numbers projected to rise to 843 million by 2050—nearly 39% of the associated disability can be attributed to occupational ergonomic exposures, smoking, and high BMI [2]. It can affect an individual's ability to work, maintain relationships, and function in normal day-to-day life [1]. Recent European data confirm LBP's labor-market impact: in a nationwide Swedish cohort of more than 830,000 workers, any episode of LBP-related sickness absence tripled the likelihood of subsequent disability retirement (hazard ratio 3.2) [3]. The UK evidence from the Whitehall II data similarly reports that recurrent back pain is associated with higher odds of exiting from paid work for health reasons, particularly in those in lower and middle grade jobs [3].

As the incidence of LBP is becoming more prevalent globally, with the steepest relative increases projected for low- and middle-income countries [4], where resources may be limited, it is important that accessible and effective interventions are available. The World Health Organization's 2023 Digital Health Strategy cites musculoskeletal conditions as a priority area for scalable eHealth solutions in resource-constrained settings. Since most LBP is not considered to be related to a specific identifiable spinal pathology, digital educational self-management programs are appealing as they encourage and facilitate wider participation. Digital delivery, including web-based and mobile formats, expands reach, allows asynchronous access, and can reduce per-capita costs by leveraging economies of scale [5].

Digital educational self-management programs have demonstrated effectiveness in a range of chronic disorders such as cardiovascular disease [4], chronic obstructive pulmonary disease [6], and diabetes [5]. For LBP specifically, a 2025 systematic review of digital interventions for employed adults with LBP identified small-to-moderate improvements in pain, disability, and work ability across 18 trials, although study quality was mixed and attrition was high [7]. The following are the findings of 2 recent randomized controlled trials: the UK SupportBack 2 trial (n=1201) found that an internet-delivered program with or without brief clinician support was safe, cost-effective, and produced modest reductions in disability at 12 months [8]; the Scandinavian selfBACK trial (n=346) reported a statistically significant reduction in Roland-Morris disability scores that persisted to 9 months among workers with physically demanding jobs [9]. However, digital interventions to date appear to largely focus on individuals seeking care within health systems, limiting accessibility. An option for improving this and scalability is by using a "MOOC" (massive open online course) format. MOOCs are a type of digital course designed using principles of educational technology. For employees with persistent LBP, a MOOC offers flexible, asynchronous learning that can be fitted around working hours and accessed from any location with an internet connection. While MOOCs have shown promise during COVID in helping health workers through the provision of health education [9], the feasibility of using this format to facilitate an educational self-management program for working individuals with back pain has not been investigated.

The MyRelief MOOC was developed specifically to support individuals in employment with LBP. The materials were derived from a synthesis of consistent recommendations across 6 clinical guidelines [10-15]. The key focus of the MyRelief program was to provide a summary of evidence-based information in user-friendly formats.

The aim of this study was to determine the feasibility and acceptability of the MyRelief MOOC.

Aim and Objectives

Aim

This study aimed to explore the feasibility, usability, and acceptability of a digital education program for people with persistent LBP.

Objectives

The objectives of the study are to evaluate the feasibility of the digital education program based on recruitment, retention, and completion of outcome measures; to assess the usability of the intervention using an industry standard tool to understand participants' ease of interaction with the digital education program; and to evaluate the acceptability of the content and format of the digital education program through qualitative feedback.

Methods

Overview

The MyRelief study protocol, providing the rationale and description of the study, has been published elsewhere [16]. This paper was prepared in accordance with the Consolidated Standards of Reporting Trials (CONSORT) guidelines for reporting of a pilot or feasibility trial, and a CONSORT checklist has been included as Checklist 1.

Study Design

The MyRelief study was a multicenter, multinational, uncontrolled feasibility study in Italy, Portugal, Sweden, and the United Kingdom. A fifth partner, Lithuania, contributed as a technology specialist but was not directly involved in the feasibility study. A separate paper was published [17] focusing on the technological aspects of the study. Ethics approval was obtained from the relevant ethics committee in each country. This study was funded by the European Commission's Erasmus+ Program.

XSL•FO
RenderX

Participants and Recruitment

Recruitment of adult workers with nonspecific, persistent LBP, present for at least 3 months, took place at the 4 study sites in Italy, Portugal, Sweden, and the United Kingdom between 2020 and 2021. Recruitment took place via several methods depending on each country's ethical approval, for example, via waiting list for outpatient rehabilitation services (Italy), email circulated to all staff and students at University (Sweden and the United Kingdom) or patients at a pain clinic (Portugal) or Social Media (Italy, Sweden, the United Kingdom, and Portugal). For the latter approach, a Facebook page was created with information about the project. Links to this page were posted on Instagram, Twitter, and other relevant forums. Individuals with unexplained symptoms, underlying pathology, or specific LBP were excluded, as were those who were not interested in using a digital intervention delivered via computer, laptop, tablet, or mobile. A full list of inclusion/exclusion criteria is in our study protocol [16].

Regardless of the method of recruitment, potential participants accessed the study information via a link provided in the recruitment material. If interested, they completed a web-based inclusion/exclusion questionnaire. Those meeting the inclusion criteria were given access to a digital consent form. After completing the consent form, they received access to the MyRelief study website to complete the baseline assessment and access the digital education program.

All participants were also asked to enter their email address and phone number to enable reminders and contacts with those interested in participating in the qualitative interview.

Various strategies to enhance recruitment were adopted in the different countries. In Sweden and the United Kingdom, interested participants responded via email to a member of the research team who checked eligibility and gave an overview of the study, answering any queries, prior to providing the study web link. The researcher arranged email, telephone, or video call support, depending on participant preference, to support them to get started with the system and checked in to resolve any technical issues. In Italy, multiple routes (email, telephone, and in-person recruitment meetings) were used to recruit participants, who were staff at the research hospital (nurses), local university (employees and administrative staff), and the general public. In Portugal, participants were recruited by direct contact with private clinics dealing with the issue.

Intervention

The MyRelief digital education program was designed to be accessed on a computer or a mobile device. The MyRelief content was informed by focus groups with adult workers with persistent LBP, LBP guidelines, and the clinical expertise of the research team. It comprised 8 modules which covered the following topics: understanding LBP; physical activity and exercise in relation to LBP; psychological factors; sleep/nutrition; management of LBP in the workplace; communication with health care providers; and other issues related to LBP (eg, medication, scans, and alternative therapies). Each module was designed to be short enough to complete during a work break, that is, <20 minutes, and was made up of short written and video content, as well as knowledge quizzes. Participants received 4 weeks' access to the digital education program prior to being invited to complete the postintervention assessment. Thus, completion of 2 online modules per week was encouraged to facilitate completion of the course within the timeframe.

Feasibility

The feasibility of the study was assessed as the ability to recruit participants and retain them in the study, and completion of functional disability and quality of life measures. Recruitment was assessed as the proportion of the recruitment target met during the timeframe. A record was kept of the number of individuals who entered the system, completed the screening questions, and provided consent. Retention was assessed by calculating the proportion of participants who completed outcome assessments at both baseline and postintervention timepoints. Completion of outcome measures at baseline and postintervention timepoints was recorded.

Usability

Participant views on usability and satisfaction with the system were assessed postintervention using the System Usability Scale (SUS), which is an industry standard measure with validity to measure a variety of aspects of usability, such as the need for support, training, and complexity [18].

Acceptability

Qualitative feedback was assessed through semistructured telephone interviews with a subgroup of study participants (Italy) and comments provided through email correspondence between participants and the study teams (Sweden and the United Kingdom). Semistructured interview questions included feedback on participants' experience using the digital education program, likes and dislikes, perceived usefulness, problems experienced, and suggestions for improvement.

Patient Characteristics

Participant characteristics (demographics, history of LBP, type of occupation, etc) were collected at baseline via the study website. Type of occupation was defined as sedentary (mostly sitting, eg, office worker, driver, and professional), light manual (mostly standing or walking, eg, retail worker, health care worker, and cook), heavy manual (heavy labor, eg, farming, construction, and cleaning), unsure, or other.

The following outcome measures were collected at baseline and postintervention via the study website:

1. Functional disability was measured using the Oswestry disability questionnaire (ODQ) v2.1a, a valid and reliable measure of pain and physical function in people with LBP [19,20]. A minimum important change of between 10 and 12 points over time, or an improvement from baseline of between 20% and 30% for an individual, has been recommended [21].

2. Health-related quality of life was measured using the EuroQol questionnaire (EQ-5D-3L), which has been recommended for use in LBP research [22]. A minimum important change of 0.03 points for the categorical scales of the EQ-5D and 10.5 points for the EQ-5D Visual Analog Scale has been suggested [22].

XSL•FO
RenderX

The following outcomes were measured at postintervention only:

3. Change in understanding and coping with their LBP was assessed via the study website using the Patient Enablement Index (adapted to reflect enablement after engaging with the system), which is considered "the gold standard" for measuring enablement [23,24]. This measure is calculated as a sum score ranging between –6 to 10 points. A score greater than 6 points is suggested to reflect "high" enablement [25].

4. SUS: Scores above 70 can be interpreted as good usability, while scores below 50 indicate unacceptably low usability according to Bangor et al [18].

5. Semistructured interview (see above).

Statistical Analysis

Statistical analysis was performed using Microsoft Excel. Participant characteristics and feasibility outcomes for recruitment and retention were presented as frequency and percentage (%), mean and SD, or median and IQR based on the distribution of the data.

The following feasibility criteria were applied: observed recruitment rate falls short of 70% of that anticipated, overall drop out of over 35%, no apparent change in the outcomes with confidence intervals that include large negative values, feedback from participants that they were unable to complete (or lack of engagement with) MyRelief, feedback from participants that suggests that the MyRelief content or format was not acceptable or usable.

Clinical and usability outcomes were reported descriptively as means and 95% CIs; significance tests were not performed in this feasibility study.

Changes From Study Protocol

A number of pragmatic changes were implemented during study delivery. Initially, study recruitment was entirely online via a web link. Engagement was poor with this automated recruitment method, so additional methods were implemented to support recruitment. Changes related to researcher-supported recruitment are described above in the "Participants and Recruitment" section. In addition to these changes, participants also received emails (Sweden, the United Kingdom, and Portugal) or telephone reminders (Italy) to complete the follow-up questionnaire.

The study protocol outlined plans to collect qualitative data from participants related to usability and acceptability at all study sites; although 3 follow-up telephone interviews were successfully conducted in Italy, Sweden, and the United Kingdom, informal participant feedback was collated via telephone and email correspondence due to challenges with recruitment.

Ethical Considerations

Ethics approval for this study has been obtained from the relevant ethics committee at each intervention site (Swedish Ethical Review Authority Dnr 2020-03536). The study was classified as human research involving minimal risk. The trial has been registered on ClinicalTrials.gov (NCT03575091). Regardless of the method of recruitment, potential participants accessed the study information via a link provided in the recruitment material. If interested, they completed a web-based inclusion/exclusion questionnaire. Those meeting the inclusion criteria were given access to a digital consent form. The consent covered both primary data collection and potential secondary analysis of deidentified data. Participants were informed of their right to withdraw from the study at any time without any consequences. All data were deidentified prior to analysis, and all materials are presented to ensure complete anonymity of the study participants. Participants were assigned unique study codes, and no personally identifiable information was stored with the research data. All digital files were stored on secure, password-protected servers in accordance with institutional data protection guidelines. Participants did not receive any compensation for their contribution to the study.

Results

Participant Baseline Characteristics

A total of 40 participants (19 male and 21 female; mean age 57 years) were recruited. Baseline demographics are reported in Table 1. Mean pain duration was 11.5 years and differed between countries from 7 years (Portugal) to 16.3 years (Sweden), and among individuals, it ranged from 1 year to 38 years. A total of 28 (70%) of participants reported a sedentary occupation. Among the participants in Portugal and Italy, these levels were lower, 40% and 54%, respectively (Table 1).

XSL•FO
RenderX

Table . Baseline demographics of participants by country and in total in the MyRelief feasibility study of a digital self-management program for persistent low back pain. The study included participants recruited from multiple European countries.

	Italy (n=13)	Northern Ireland (n=13)	Portugal (n=5)	Sweden (n=9)	Total (n=40)
Age (years)					
Mean (SD)	57.2 (2.8)	56.3 (2.0)	59.8 (3.4)	57.4 (7.7)	57.3 (4.2)
Range	55 - 65	55 - 61	55 - 64	48 - 76	48 - 76
Sex, n (%)					
Female	8 (62)	6 (46)	0 (0)	7 (78)	21 (53)
Male	5 (38)	7 (54)	5 (100)	2 (22)	19 (47)
Pain duration in years					
Mean (SD)	12.6 (10.4)	8.8 (9.5)	7.0 (3.9)	16.3 (13.3)	11.5 (10.5)
Range	2 - 30	1 - 30	1 - 11	2 - 38	1 - 38
Occupation category, n (%)					
Sedentary	7 (54)	11 (85)	2 (40)	8 (89)	28 (70)
Light manual	6 (46)	2 (15)	2 (40)	0 (0)	10 (25)
Heavy manual	0 (0)	0 (0)	1 (20)	1 (11)	2 (5)

Feasibility

Recruitment and Retention

The flow diagram in Figure 1 shows the participant flow in the study. A total of 100 individuals (44 male and 56 female) created a MyRelief account and reached the study landing page. Creating an account generated a system "login" but did not in itself constitute study enrollment. Of these 100 account holders, 40 completed the electronic consent form and baseline questionnaires and were therefore classified as recruited participants. Our a priori feasibility target was to enroll 50 participants [16]; we therefore achieved 80% (40/50) of the planned sample size.

Figure 1. Participant flow diagram for the MyRelief digital self-management program feasibility study. Out of 100 individuals who signed up and gave consent via the study website, 40 participants completed the baseline questionnaire. Of these, 17 completed both baseline and 12-week follow-up assessments, while 23 did not complete follow-up (dropouts). An additional 5 participants completed only the follow-up questionnaire.

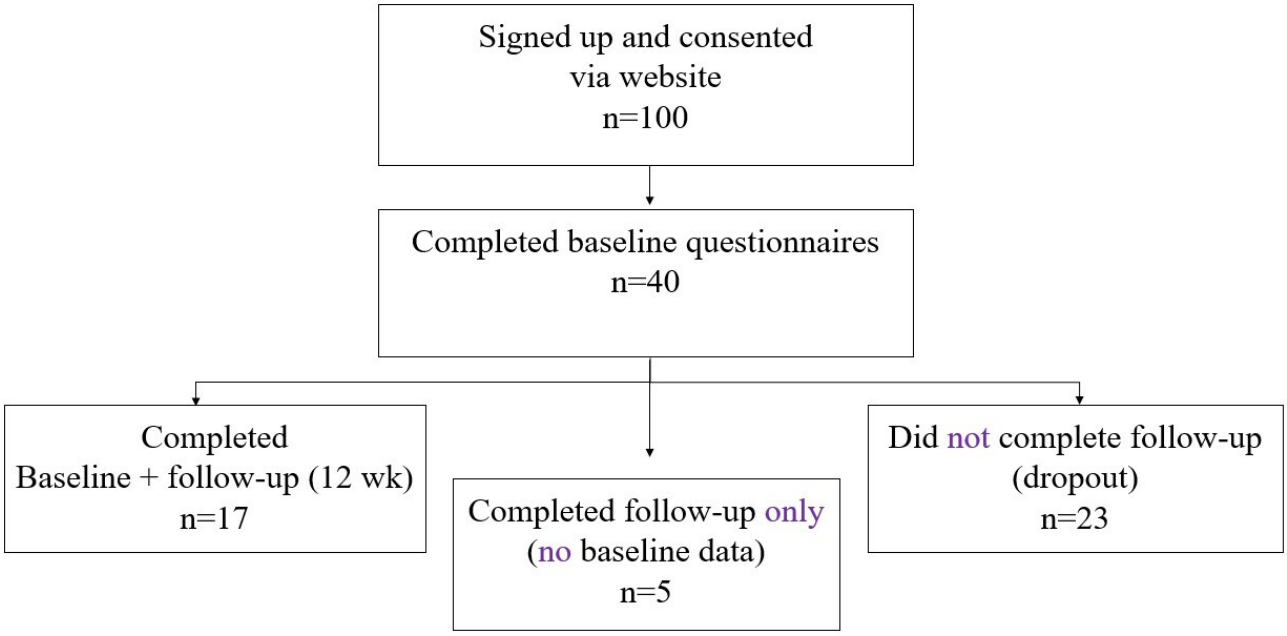

Recruitment and retention follow-up questionnaires were completed by 17 (11 male and 6 female) of the 40 recruited participants, yielding a retention rate of 42.5% and a corresponding dropout rate of 57.5%. Retention is expressed relative to the 40 recruited participants. However, an additional 5 people completed only the follow-up questionnaire, increasing the total follow-up responses to 22.

XSL•FO
RenderX

Completion of Outcome Measures

Quality of outcome measure completion was good, with valid and usable data obtained for all clinical outcomes from all participants (n=17). The outcome measures are described in Table 2.

Table . Outcome measures for participants (n=17) with complete data at both baseline and 12-week follow-up in a digital self-management program for persistent low back pain.

	Baseline	Follow-up
ODQ[a], mean (95% CI)	24.0 (18-31)	23.9 (16-31)
ODQ categories, n (%)		
>21 (moderate)	8 (47.0)	7 (41.2)
<21 (low)	9 (53.0)	10 (58.8)
Pain intensity, mean (95% CI)[b]	1.58 (1.07 - 2.10)	1.50 (1.07 - 1.93)
EQ-5D-3L, mean (95% CI)[c]	0.678 (0.598 - 0.759)	0.723 (0.662 - 0.785)
EuroQol Visual Analog Scale, mean (95% CI)[c]	67.76 (60.71 - 74.82)	70.59 (65.14 - 76.03)

[a]ODQ: Oswestry disability questionnaire (0 - 100), a higher score indicates more disability. A minimal clinically important difference between 10 points over time, or an improvement from baseline of between 20% and 30% for an individual has been recommended.

[b]Pain intensity: responds to the first item of ODQ "What is your pain intensity at the moment," scores ranging between 0 (no pain at the moment) to 5 (the pain is the worst imaginable at the moment).

[c]EuroQol questionnaire (EQ-5D-3L) Weighted Health Index, TTO (–0.59 to 1.0) a higher score indicates better quality of life. A change of 0.03 points for the categorical scales of the EQ-5D-3L and 10.5 points for the EuroQol Visual Analog Scale can be categorized as clinical important.

Across all participants, the mean baseline ODQ was 24.0 (95% CI 18 - 31), which would correspond to moderate disability [19,20]. Mean pain intensity at baseline was 1.58 out of 5 and ranged between 0 and 2, except for one participant who had scored 4. The mean pain intensity showed a positive correlation to the total mean ODQ (r=0.71). No change was found for the total ODQ mean at follow-up (23.9, 95% CI 16 - 31). However, 4 individual participants achieved change between baseline and follow-up corresponding to the minimal clinically important difference of the ODQ (a score >10%), where 2 participants improved from moderate to low disability, one participant deteriorated from low to moderate disability, and one individual showed an increase in disability but remained within the range of moderate disability (Figure 2).

Figure 2. Change in ODQ scores from baseline to 12-week follow-up in participants taking part in a digital education program for persistent low back pain (n=17). Scores below and above the gray line represent mild and moderate disability, respectively. *Indicates a change >10 points, which represents the minimal clinically important difference (MCID) for ODQ. ODQ: Oswestry Disability Questionnaire.

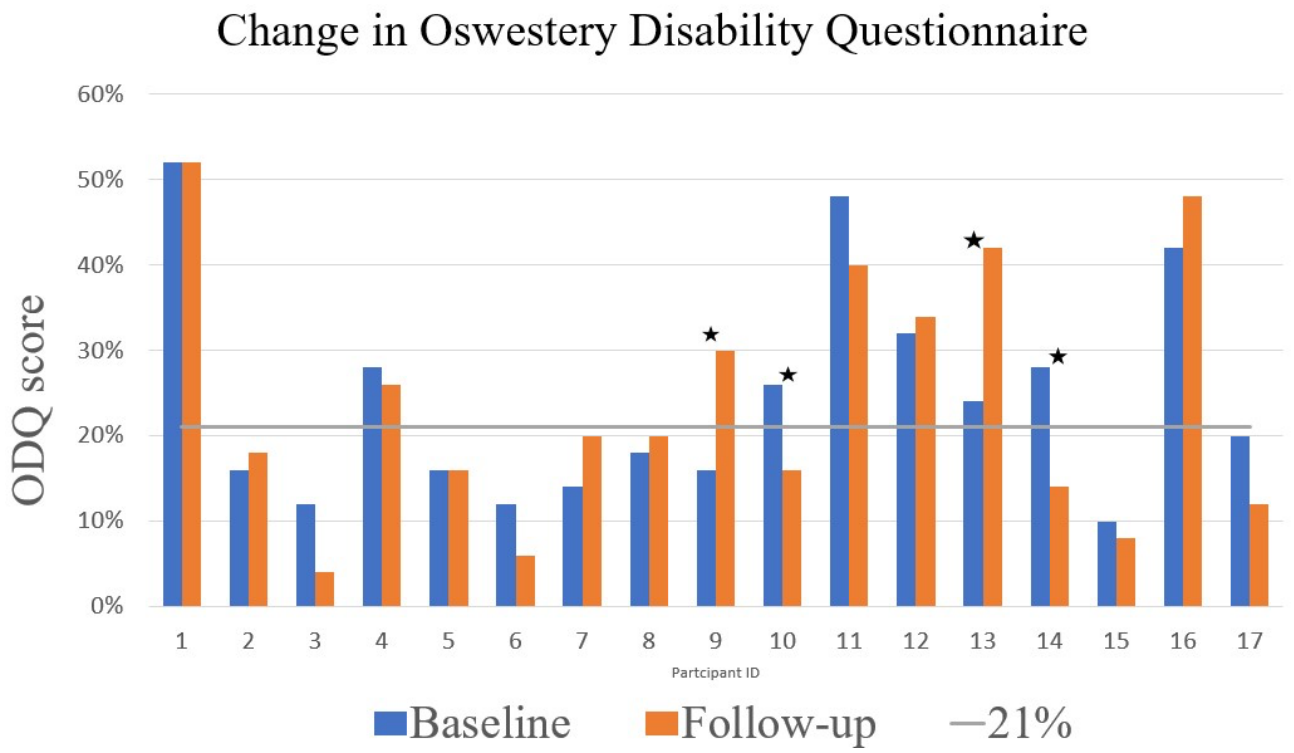

XSL•FO
RenderX

A clinically important increase was found for the mean EQ-5D Weighted Health Index between baseline (0.678, 95% CI 0.598 - 0.759) and follow-up (0.723, 95% CI 0.662 - 0.785). The EQ-5D visual analog scale also showed an increase from 67.76 (95% CI 60.71 - 74.82) to 70.59 (95% CI 65.14 - 76.03); this was not a clinically important change. At the individual level, the majority of participants increased or maintained their EQ-5D scores at follow-up (Table 2).

A total mean for the Patient Enablement Index was calculated as 5.31 (95% CI 3.94 - 6.70) at follow-up. The median value was 5.5 points. A Patient Enablement Instrument (PEI) greater than 6 points is suggested to reflect "high" enablement [25]. Two participants (9%) reached the highest score (12 points), and 2 participants (9%) scored the lowest score (1 point). Thus, no floor or ceiling effects were found for the total score.

Usability

Overview

A total of 22 participants responded to the SUS. The total mean SUS score was 72.4 (95% CI 67.5 - 73.3), which would be interpreted as good [18]. The median score was 75. The score covered the range from 52.5 (1 person) to 95.0 (1 person), with 13 out of 22 participants scoring above 70. Figure 3 describes the participant scores of SUS. Qualitative feedback confirmed some challenges with usability in terms of access to and navigation within the website (Textbox 1).

Figure 3. Participant scores of the SUS at 12-week follow-up in participants taking part in a digital education program for persistent low back pain (n=22). A higher score indicates better usability. Scores above 70 indicate good usability, while scores below 50 indicate unacceptably low usability. SUS: System Usability Scale.

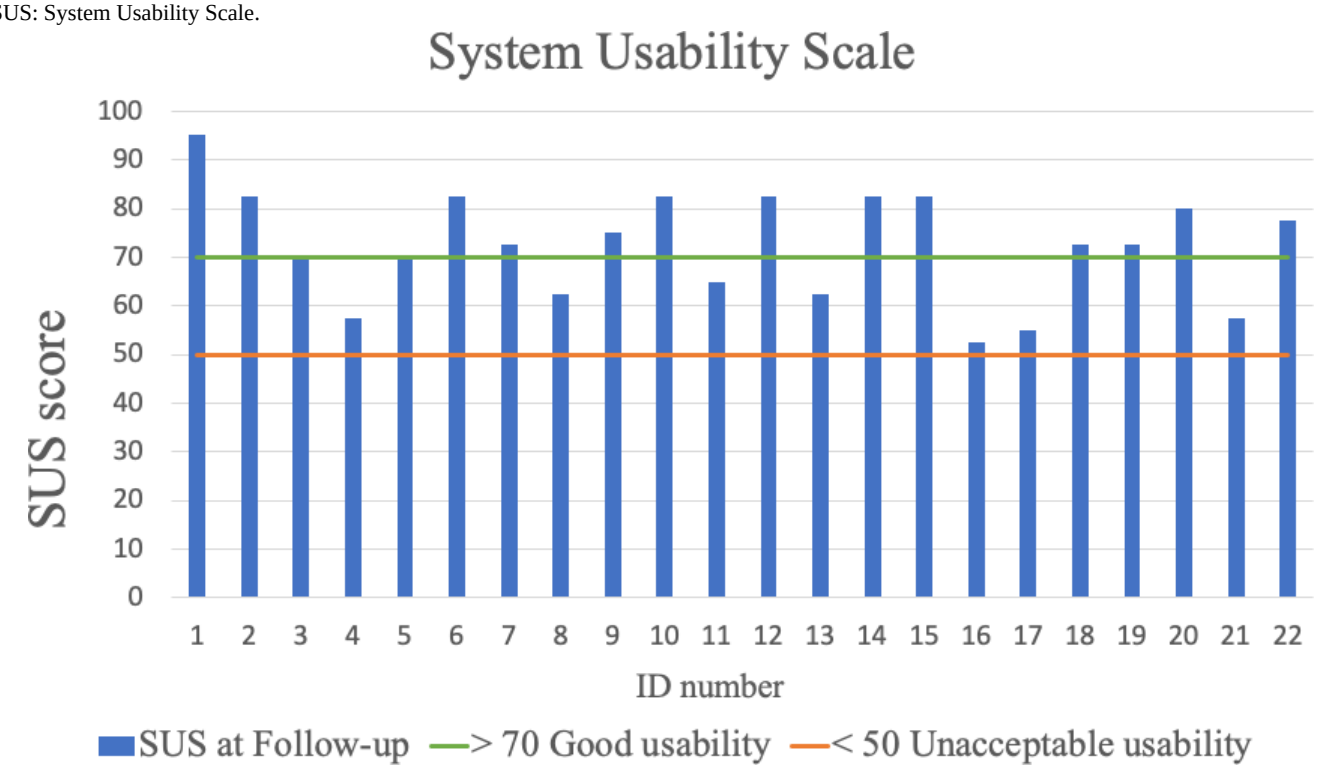

XSL•FO
RenderX

Textbox 1. Qualitative feedback on the acceptability of the MyRelief educational web page. Qualitative feedback was gathered via interviews (Italy) and email (Sweden and the United Kingdom), focusing on participants' experiences with the digital education program.

Pros

The international experts together. The fact of having input from experts from different countries was very good.

It was interesting, I really enjoyed it.

There were things I did not know that came in handy. Like the talk about stress can lead to spinal problems. It was very useful.

Cons

Access was a bit complicated at first, but once understood, everything went well.

Perhaps I expected a little more from the ergonomics point of view. Ergonomics applied to everyday life (eg, posture in the workplace).

The exercise program was very difficult. The exercises are tiring and difficult.

A slightly more graduated/modulated exercise program would be needed.

It is informative but not sufficient to support a change in lifestyle.

It lacks the possibility of face-to-face interaction and should be integrated with other interventions in order to work.

Not all speakers had the same communication skills.

All physiotherapists were female, while the doctor was male!

Unfortunately, some of the lessons were not in Italian, I had difficulties.

Difficult to navigate at the web page.

Some quiz questions were perceived as provocative.

Difficult to access the web page.

Acceptability

Qualitative feedback also made suggestions on the content of the website that were both positive and negative, for example, "having input from experts from different countries was very good" and "the exercise programme was very difficult" (Textbox 1).

Discussion

Principal Findings

This study explored the feasibility, usability, and acceptability of a digital education program for people living and working with persistent LBP. Recruitment of study participants reached the a priori feasibility criteria (>70% recruitment). However, retention was challenging, and the a priori feasibility criteria were not met (<35% dropouts). The quality of outcome measure completion was good. The SUS indicates that the MyRelief website met minimal industry standards, but qualitative feedback confirmed some challenges with usability in terms of access to and navigation within the website. Although mean functional disability was moderate, actually, half of our sample only had mild functional disability and low pain intensity. A clinically important increase between baseline and follow-up was found for EQ-5D Weighted Health Index, suggesting a meaningful improvement in patients' perceived health status. This aligns with the idea that participation in patient education sessions may play a positive role in enhancing quality of life. While the observed improvement is promising, further exploration is needed to determine whether this alone is sufficient to sustain long-term improvements in quality of life.

Feasibility

Recruitment and Retention

While the idea of an open, globally accessible, evidence-based self-management intervention for LBP is appealing, this multi-country study also revealed several challenges. Although we met our a priori recruitment target of 50 participants, it is worth noting that despite 100 individuals signing up and providing consent to participate, only 40 completed the baseline questionnaire, corresponding to a 40% response rate among those registered.

Based on the a priori recruitment target of 50 completed baseline assessments, the study achieved 80% of its goal, thereby meeting the predefined feasibility threshold (>70%).

Once recruited, we encountered a second significant challenge in terms of retention of the participants who have signed into the study website, as retention at follow-up was more limited: 17 participants completed both baseline and follow-up assessments, yielding a retention rate of 42.5% and a dropout rate of 57.5%, which did not meet the preset criterion (<35% dropouts). These difficulties reflect a broader challenge in trial recruitment, as shown in a comprehensive review of 151 randomized controlled trials, where only 56% met their final recruitment target, highlighting the ongoing complexity in recruiting for randomized controlled trials and the need for more effective planning [26]. It could also be postulated that using a completely web-based recruitment strategy is not advisable. A recent meta-analysis found web-based recruitment strategies superior to offline recruitment initiatives when measuring recruitment effectiveness by recruitment rate and cost-effectiveness. However, this study found that offline recruitment outperforms web-based recruitment when converting

XSL•FO
RenderX

potential participants to actual enrollees [27]. Engaging with participants through screening and providing an opportunity for discussion appeared to enhance recruitment in our study. These findings highlight important challenges in sustaining engagement in fully digital, self-directed interventions. Factors such as lack of personal contact, time constraints, or limited perceived relevance may have contributed to lower retention. Future studies may benefit from incorporating hybrid delivery models, tailored reminders, or additional onboarding support to improve participant engagement.

Completion of Outcome Measures

The baseline mean ODQ score of 24.0 (95% CI 18 - 31) was notably lower than in comparable studies examining disability levels among individuals with persistent LBP recruited via health care practitioners [28,29] but similar to studies, like ours, that recruited directly from the community [30]. This may explain in part the marginal improvements we observed in terms of numbers meeting the minimal important change criteria [21]. These findings suggest that the cohort under study might not be the most responsive to the intervention being evaluated. It is likely that a digital program like MyRelief could have a greater impact on individuals with more severe LBP. Given these insights, it seems prudent to target patients awaiting care in primary health care settings, as they may represent a more suitable demographic for intervention. Consequently, future recruitment efforts could be optimized by engaging health care providers in the process.

The mean score for patient enablement reported was 5.31, a level that corresponds to previously published studies reporting on patient enablement [31-33]. Patient enablement emphasizes patients' involvement in their care, understanding of their health, and participation in decision-making. The original PEI assesses change in patients' ability to cope and their understanding of their disease, following a consultation [34]. We used a modified PEI to assess the impact of MyRelief on knowledge acquisition for self-management strategies. Even though our study's relatively high enablement level may stem from MyRelief's focus on knowledge acquisition rather than solving complex issues typical of health care consultations. Our findings suggest that MyRelief has the potential to influence patient enablement, although the specific components contributing to this improvement require additional exploration.

Usability

The SUS score of 72.4 suggests that our digital education program meets acceptable usability standards and can be considered to offer good usability [18], indicating that MyRelief digital education has the prerequisites required for further development. In comparison to other digital interventions to support and facilitate self-management of LBP, previous mean scores have ranged from 64.7 to 70.5 [17,35]. However, comparability is limited as previous studies have mostly focused on the usability of mobile applications. Nonetheless, it is promising that the current digital program was perceived as usable by MyRelief study participants. The current platform did not support detailed use analytics, which limited our ability to assess which modules were most engaging or where participants may have disengaged. Future versions should include integrated tracking of module completion, time spent, and quiz interaction to better understand user patterns and adapt content accordingly. There was a lack of qualitative feedback on usability in our study, but the feedback that was collected primarily highlighted issues around accessibility to the website. It has been reported that a clear and intuitive navigation system can significantly enhance accessibility, ensuring easy access to important information [36]. Improved understanding regarding accessibility and usability preferences from the end users' perspective is required in order to improve the utility of the digital education format for health-related programs.

Acceptability

Also, the acceptability of the MyRelief digital education program was assessed through qualitative feedback collected via semistructured interviews and email correspondence. Participants generally appreciated the relevance of the content and its focus on self-management, with several highlighting the usefulness of practical examples. However, feedback also revealed that some participants found the material too basic in relation to their existing knowledge or needs. This suggests that although the education program was acceptable to many, others desired greater depth or personalization.

While the overall design and delivery format were well received, some users expressed preferences for shorter modules or more interactive elements. These comments provide important insights for future refinement of the education program. Acceptability could potentially be improved by tailoring content to different levels of prior knowledge or by offering optional advanced sections. Although qualitative feedback was limited in volume, it was important in identifying strengths and specific areas for improvement.

Strengths and Limitations

We conducted this study in accordance with CONSORT guidelines and followed a prospectively registered protocol [16]. A key strength and a novelty of this trial is the transnational approach in developing evidence-based material that could facilitate health education, providing an infrastructure that can be upscaled to reach more learners, leveraging digital tools and materials. However, the transnational development process also revealed that the cultural context of health differed between countries. A lot of the workload within the project was attributed to developing materials that required multiple translations and heavy editing to ensure the advice and information were appropriate for and sensitive to different cultural contexts. To maximize the utility of interventions delivered using a digital education format, cultural contexts of health need to be considered. The multi-country nature of this study revealed valuable insights into the importance of cultural and contextual adaptation. Although no formal subgroup comparisons were made, observations from the project teams indicated differences in how participants across sites perceived the program's content and usability. Some adjustments—such as modifying terminology, examples, and health care system references—were made during translation to improve relevance. Nonetheless, variations in language availability and local health literacy may have influenced both engagement and acceptability, highlighting the need for cocreation [37-39] and iterative user testing in

XSL•FO
RenderX

future implementations. Although we conducted initial focus groups with individuals with LBP, further iterative review opportunities with patient representatives and clinicians were not feasible within the project timeline. For future studies, more focus should be dedicated to structured cocreation and the involvement of health care providers in this process, including repeated feedback loops from patients in each participating country. In addition, the diversity of learners' own goals, motivations, and interests needs to be satisfied through the design, as the learning experience will benefit from personalization [40]. Another key consideration for future development is the importance of using an administratively simple and flexible course system. The current platform lacked integrated tracking features, which limited our ability to monitor progress and engagement. To minimize challenges with technology, interdisciplinary collaboration with developers and clinician academics is crucial. For digital education programs to be scalable and sustainable, they must also be supported by systems that allow easy management of participants, automated data collection (eg, module completion, quiz performance), and user support. A streamlined and user-friendly backend interface is essential for both researchers and educators to efficiently deliver content, monitor engagement, and provide timely feedback.

Conclusions

This study was feasible in terms of recruitment but not retention. About half of our sample had low levels of functional disability, and in the future, recruitment via health care providers may be more appropriate, and recruitment strategies should focus on efforts to recruit individuals from heavy manual and repetitive task jobs, as they are more representative of the wider persistent LBP population. Our usability scores suggest our website meets minimal industry standards. Completion of outcomes was good, apart from ODQ. Future studies should explore culturally adapted versions of the program, developed in collaboration with patients and health care providers. Earlier and more frequent end user input may help refine language and relevance, thereby improving both recruitment and retention.

Acknowledgments

The study has been cofunded by the Erasmus program of the European Union (project ID: 2018–1-SE01-KA204-039144) and Ann-Mari och Ragnar Hemborgs Minnesfond. The authors wish to acknowledge and express our sincere gratitude to all research participants for their valuable contributions to this study. The authors also acknowledge all coworkers at Virtual Campus (Portugal), Tech4Care (Italy), and the Lithuanian team for their dedicated work in developing the MyRelief Digital Education and their collaboration within the project.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request. All other materials related to the MyRelief massive open online course can be found online at the course page at Udemy [41].

Authors' Contributions

SMD developed the research design, methods, and study materials in collaboration with all authors (JM, CL, FP, SH, and EC). This manuscript was prepared by all authors. All authors reviewed the article for important intellectual content. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Checklist 1
CONSORT (Consolidated Standards of Reporting Trials) checklist.
[PDF File, 230 KB - formative_v9i1e74948_app1.pdf]

References

1. Wu A, March L, Zheng X, et al. Global low back pain prevalence and years lived with disability from 1990 to 2017: estimates from the global burden of disease study 2017. Ann Transl Med 2020 Mar;8(6):299. [doi: 10.21037/atm.2020.02.175] [Medline: 32355743]
2. GBD 2021 Low Back Pain Collaborators. Global, regional, and national burden of low back pain, 1990-2020, its attributable risk factors, and projections to 2050. Lancet Rheumatol 2021;5(6):e316-e329. [doi: 10.1016/S2665-9913(23)00098-X]
3. d'Errico A, Falkstedt D, Almroth M, Badarin K, Hemmingsson T, Kjellberg K. Long-term sick leave for back pain, exposure to physical workload and psychosocial factors at work, and risk of disability and early-age retirement among aged Swedish workers. Int Arch Occup Environ Health 2022 Sep;95(7):1521-1535. [doi: 10.1007/s00420-022-01862-8]
4. Global strategy on digital health 2020-2025. World Health Organization.: WHO URL: https://www.who.int/docs/default-source/documents/gs4dhdaa2a9f352b0445bafbc79ca799dce4d.pdf [accessed 2025-07-25]

XSL•FO
RenderX

5. Golovaty I, Wadhwa S, Fisher L, et al. Reach, engagement and effectiveness of in-person and online lifestyle change programs to prevent diabetes. BMC Public Health 2021 Dec;21(1):34225674. [doi: 10.1186/s12889-021-11378-4]
6. Geraghty AWA, Becque T, Roberts LC, et al. Supporting self-management of low back pain with an internet intervention with and without telephone support in primary care (SupportBack 2): a randomised controlled trial of clinical and cost-effectiveness. Lancet Rheumatol 2024 Jul;6(7):e424-e437. [doi: 10.1016/S2665-9913(24)00086-9]
7. Chen M, Sparkes V, Sheeran L. Systematic review of digital health interventions to support self-management of low back pain in the workplace. Digit Health 2025 May;11:20552076251336281. [doi: 10.1177/20552076251336281]
8. Geraghty AWA, Becque T, Roberts LC, et al. Supporting self-management with an internet intervention for low back pain in primary care: a RCT (SupportBack 2). Health Technol Assess 2025;29(7):1-90. [doi: 10.3310/GDPS2418]
9. Rasmussen CDN, et al. Effect of a smartphone self-management support system for low back pain (selfBACK) among workers with high physical demands. Scand J Work Environ Health 2024;50(8):613-621. [doi: 10.5271/sjweh.4186]
10. Qaseem A, Wilt TJ, McLean RM, et al. Noninvasive treatments for acute, subacute, and chronic low back pain: a clinical practice guideline from the American college of physicians. Ann Intern Med 2017 Apr 4;166(7):514-530. [doi: 10.7326/M16-2367] [Medline: 28192789]
11. Pangarkar SS, Kang DG, Sandbrink F, et al. VA/DoD clinical practice guideline: diagnosis and treatment of low back pain. J Gen Intern Med 2019 Nov;34(11):2620-2629. [doi: 10.1007/s11606-019-05086-4] [Medline: 31529375]
12. Chenot JF, Greitemann B, Kladny B, Petzke F, Pfingsten M, Schorr SG. Non-specific low back pain. Dtsch Arztebl Int 2017 Dec 25;114(51-52):883-890. [doi: 10.3238/arztebl.2017.0883] [Medline: 29321099]
13. Ailliet L, Berquin A, et al. Low Back Pain and Radicular Pain: Assessment and Management: Good Clinical Practice (GCP) Brussels: Belgian Health Care Knowledge Centre (KCE); 2017.
14. Evidence-informed primary care management of low back pain: clinical practice guideline. Alberta Medical Association. 2015. URL: http://www.topalbertadoctors.org/cpgs/885801 [accessed 2025-08-13]
15. NICE. Low back pain and sciatica in over 16s: assessment and management. 2016.
16. Marley J, Larsson C, Piccinini F, et al. Study protocol for a feasibility study of an online educational programme for people working and living with persistent low back pain. Pilot Feasibility Stud 2023 Sep 4;9(1):154. [doi: 10.1186/s40814-023-01382-3] [Medline: 37667407]
17. Maskeliūnas R, Damaševičius R, Kulikajevas A, Marley J, Larsson C. Evaluation of MyRelief serious game for better self-management of health behaviour strategies on chronic low-back pain. Informatics (MDPI) 2022;9(2):40. [doi: 10.3390/informatics9020040]
18. Bangor A, Kortum PT, Miller JT. An empirical evaluation of the system usability scale. Int J Hum Comput Interact 2008 Jul 29;24(6):574-594. [doi: 10.1080/10447310802205776]
19. Fairbank JC, Pynsent PB. The Oswestry disability index. Spine (Phila Pa 1976) 2000 Nov 15;25(22):2940-2952. [doi: 10.1097/00007632-200011150-00017] [Medline: 11074683]
20. Roland M, Fairbank J. The Roland-Morris disability questionnaire and the Oswestry disability questionnaire. Spine (Phila Pa 1976) 2000 Dec 15;25(24):3115-3124. [doi: 10.1097/00007632-200012150-00006] [Medline: 11124727]
21. Ostelo RW, Deyo RA, Stratford P, Waddell G, Croft P, Korff M, et al. Interpreting change scores for pain and functional status in low back pain: towards international consensus regarding minimal important change. Spine (Phila Pa 1986) 2008 Jan 1;33:18165753. [doi: 10.1097/BRS.0b013e31815e3a10]
22. Soer R, Reneman MF, Speijer BLGN, Coppes MH, Vroomen PCAJ. Clinimetric properties of the EuroQol-5D in patients with chronic low back pain. Spine J 2012 Nov;12(11):1035-1039. [doi: 10.1016/j.spinee.2012.10.030] [Medline: 23199409]
23. Nielsen J, Donald N. Usability on the web isn't a luxury. : InformationWeek; 2000.
24. Tolvanen E, Koskela TH, Kosunen E. Comparison of the patient enablement instrument (PEI) with two single-item measures among Finnish health care centre patients. BMC Health Serv Res 2019 Jun 13;19(1):376. [doi: 10.1186/s12913-019-4182-2] [Medline: 31196088]
25. Howie JG, Heaney DJ, Maxwell M. Measuring quality in general practice. pilot study of a needs, process and outcome measure. Occas Pap R Coll Gen Pract 1997 Feb(75):i-xii. [Medline: 9141884]
26. Walters SJ, Bonacho Dos Anjos Henriques-Cadby I, Bortolami O, et al. Recruitment and retention of participants in randomised controlled trials: a review of trials funded and published by the United Kingdom health technology assessment programme. BMJ Open 2017 Mar 20;7(3):e015276. [doi: 10.1136/bmjopen-2016-015276] [Medline: 28320800]
27. Brøgger-Mikkelsen M, Ali Z, Zibert JR, Andersen AD, Thomsen SF. Online patient recruitment in clinical trials: systematic review and meta-analysis. J Med Internet Res 2020 Nov 4;22(11):e22179. [doi: 10.2196/22179]
28. Hurley DA, Tully MA, Lonsdale C, et al. Supervised walking in comparison with fitness training for chronic back pain in physiotherapy: results of the SWIFT single-blinded randomized controlled trial (ISRCTN17592092). Pain 2015 Jan;156(1):131-147. [doi: 10.1016/j.pain.0000000000000013] [Medline: 25599309]
29. McDonough SM, Tully MA, Boyd A, et al. Pedometer-driven walking for chronic low back pain: a feasibility randomized controlled trial. Clin J Pain 2013 Nov;29(11):972-981. [doi: 10.1097/AJP.0b013e31827f9d81] [Medline: 23446066]
30. Lang AE, Hendrick PA, Clay L, et al. A randomized controlled trial investigating effects of an individualized pedometer driven walking program on chronic low back pain. BMC Musculoskelet Disord 2021 Feb 19;22(1):206. [doi: 10.1186/s12891-021-04060-8] [Medline: 33607979]

XSL•FO
RenderX

31. Skarbalienė A, Jurgutis A, Strandberg EL, Pawlikowska T. Patient involvement in assessing consultation quality: validation of patient enablement instrument (PEI) in Lithuanian general practice. BMC Fam Pract 2019 Dec 3;20(1):167. [doi: 10.1186/s12875-019-1061-1] [Medline: 31795934]
32. Röst M, Zielinski A, Petersson C, Strandberg EL. Reliability and applicability of the patient enablement instrument (PEI) in a Swedish general practice setting. BMC Fam Pract 2015 Mar 4;16(1):31. [doi: 10.1186/s12875-015-0242-9] [Medline: 25879632]
33. Ozvacić Adzić Z, Katić M, Kern J, Lazić D, Cerovecki Nekić V, Soldo D. Patient, physician, and practice characteristics related to patient enablement in general practice in Croatia: cross-sectional survey study. Croat Med J 2008 Dec;49(6):813-823. [doi: 10.3325/cmj.2008.49.813] [Medline: 19090607]
34. Howie JG, Heaney DJ, Maxwell M, Walker JJ. A comparison of a patient enablement instrument (PEI) against two established satisfaction scales as an outcome measure of primary care consultations. Fam Pract 1998 Apr;15(2):165-171. [doi: 10.1093/fampra/15.2.165] [Medline: 9613486]
35. Nordstoga AL, Bach K, Sani S, et al. Usability and acceptability of an app (SELFBACK) to support self-management of low back pain: mixed methods study. JMIR Rehabil Assist Technol 2020 Sep 9;7(2):e18729. [doi: 10.2196/18729] [Medline: 32902393]
36. Shi A, Huo F, Han D. Role of interface design: a comparison of different online learning system designs. Front Psychol 2021;12(681756):681756. [doi: 10.3389/fpsyg.2021.681756] [Medline: 34531786]
37. Ageberg E, Bunke S, Linnéll J, Moesch K. Co-creating holistic injury prevention training for youth handball: development of an intervention targeting end-users at the individual, team, and organizational levels. BMC Sports Sci Med Rehabil 2024 Jan 8;16(1):10. [doi: 10.1186/s13102-023-00800-6] [Medline: 38191495]
38. Leask CF, Sandlund M, Skelton DA, et al. Framework, principles and recommendations for utilising participatory methodologies in the co-creation and evaluation of public health interventions. Res Involv Engagem 2019;5(2):2. [doi: 10.1186/s40900-018-0136-9] [Medline: 30652027]
39. Vargas C, Whelan J, Brimblecombe J, Allender S. Co-creation, co-design, co-production for public health – a perspective on definitions and distinctions. Public Health Res Pract 2022;32(2):e3222211. [doi: 10.17061/phrp3222211):e2022]
40. Bucher A. The patient experience of the future is personalized: using technology to scale an N of 1 approach. J Patient Exp 2023;10(23743735231167975):23743735231167975. [doi: 10.1177/23743735231167975] [Medline: 37051113]
41. Self management strategies for people with low back pain. Udemy. URL: https://www.udemy.com/course/self-management-strategies-for-people-with-low-back-pain/ [accessed 2025-07-25]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
LBP: low back pain
MOOC: massive open online course
ODI: Oswestry Disability Index
ODQ: Oswestry Disability Questionnaire
PEI: Patient Enablement Instrument
SUS: System Usability Scale

Edited by A Mavragani; submitted 04.04.25; peer-reviewed by K Dunleavy; revised version received 02.07.25; accepted 03.07.25; published 27.08.25.

Please cite as:
Larsson C, Marley J, Piccinini F, Howes S, Casoni E, Aschettino V, Vaz de Carvalho C, McDonough S
The MyRelief Digital Educational Self-Management Program for Persistent Low Back Pain: Feasibility Uncontrolled Trial
JMIR Form Res 2025;9:e74948
URL: https://formative.jmir.org/2025/1/e74948
doi:10.2196/74948

© Caroline Larsson, Joanne Marley, Flavia Piccinini, Sarah Howes, Elisa Casoni, Vincenzo Aschettino, Carlos Vaz de Carvalho, Suzanne McDonough. Originally published in JMIR Formative Research (https://formative.jmir.org), 27.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Evaluating the Quality and Understandability of Radiology Report Summaries Generated by ChatGPT: Survey Study

Alexis Sunshine[1], MD; Grace H Honce[2], MPH; Andrew L Callen[1], MD; David A Zander[1], MD; Jody L Tanabe[1], MD; Samantha L Pisani Petrucci[1], MD; Chen-Tan Lin[3], MD; Justin M Honce[1], MD

[1]Department of Radiology, University of Colorado Anschutz Medical Campus, 19th Ave. Mail Stop C278, Aurora, CO, United States

[2]Hartway Evaluation Group, Denver, CO, United States

[3]Department of Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO, United States

Corresponding Author:
Justin M Honce, MD
Department of Radiology, University of Colorado Anschutz Medical Campus, 19th Ave. Mail Stop C278, Aurora, CO, United States

Abstract

Background: Radiology reports convey critical medical information to health care providers and patients. Unfortunately, they are often difficult for patients to comprehend, causing confusion and anxiety, thereby limiting patient engagement in health care decision-making. Large language models (LLMs) like ChatGPT (OpenAI) can create simplified, patient-friendly report summaries to increase accessibility, albeit with errors.

Objective: We evaluated the accuracy and clarity of ChatGPT-generated summaries compared to original radiologist-assessed radiology reports, assessed patients' understanding and satisfaction with the summaries compared to the original reports, and compared the readability of the original reports and summaries using validated readability metrics.

Methods: We anonymized 30 radiology reports created by neuroradiologists at our institution (6 brain magnetic resonance imaging, 6 brain computed tomography, 6 head and neck computed tomography angiography, 6 neck computed tomography, and 6 spine computed tomography). These anonymized reports were processed by ChatGPT to produce patient-centric summaries. Four board-certified neuroradiologists evaluated the ChatGPT-generated summaries on quality and accuracy compared to the original reports, and 4 patient volunteers separately evaluated the reports and summaries on perceived understandability and satisfaction. Readability was assessed using word count and validated readability scales.

Results: After reading the summary, patient confidence in understanding (98%, 116/118 vs 26%, 31/118) and satisfaction regarding the level of jargon/terminology (91%, 107/118 vs 8%, 9/118) and time taken to understand the content (97%, 115/118 vs 23%, 27/118) substantially improved. Ninety-two percent (108/118) of responses indicated the summary clarified patients' questions about the report, and 98% (116/118) of responses indicated patients would use the summary if available, with 67% (79/118) of responses indicating they would want access to both the report and summary, while 26% (31/118) of responses indicated only wanting the summary. Eighty-three percent (100/120) of radiologist responses indicated the summary represented the original report "extremely well" or "very well," with only 5% (6/120) of responses indicating it did so "slightly well" or "not well at all." Five percent (6/120) of responses indicated there was missing relevant medical information in the summary, 12% (14/120) reported instances of overemphasis of nonsignificant findings, and 18% (22/120) reported instances of underemphasis of significant findings. No fabricated findings were identified. Overall, 83% (99/120) of responses indicated that the summary would definitely/probably not lead patients to incorrect conclusions about the original report, with 10% (12/120) of responses indicating the summaries may do so.

Conclusions: ChatGPT-generated summaries could significantly improve perceived comprehension and satisfaction while accurately reflecting most key information from original radiology reports. Instances of minor omissions and under-/overemphasis were noted in some summaries, underscoring the need for ongoing validation and oversight. Overall, these artificial intelligence–generated, patient-centric summaries hold promise for enhancing patient-centered communication in radiology.

(JMIR Form Res 2025;9:e76097) doi:10.2196/76097

KEYWORDS

large language models; artificial intelligence; ChatGPT; radiology reports; patient-centered care; neuroradiology; report summarization

XSL•FO
RenderX

Introduction

The 21st Century Cures Act increased patients' access to their medical information [1,2]. This change reflected a growing desire among patients to engage with their care, a desire that is exaggerated in the field of radiology, where patients want access to reports and images, and access these materials at higher rates than general patient portals or clinical notes [3-6]. Unfortunately, radiology reports, which are primarily written for referring clinicians, use complex medical terminology that is well above the established guidelines for patient-friendly material [1-4,7-9]. Limited comprehension of these reports can leave patients feeling disempowered, anxious, and uncertain about their diagnoses and treatment plans [4]. This is particularly worrisome given that patient engagement with and comprehension of their health care information correlates with improved patient experience, improved outcomes, and decreased costs [1,3,8,10-12]. This evolving medical landscape necessitates urgent adaptations within radiology to provide easily understandable, patient-friendly reports [1,2].

There have been several attempts at improving the readability of radiology reports. For example, the University of Pennsylvania's PORTER platform is a structured reporting system that provides automated lay-language translation of reports, and University of California, San Diego, has implemented a consumer health vocabulary automated translation and embedded hyperlinks for defining key terms [4,13,14]. However, there is a need to find ways to simplify language and avoid jargon to facilitate clear patient communication [4,8,9,9,15,16]. Radiologists face increasing workloads, making the manual preparation of simplified reports for every patient impractical [17]. Recent improvements in artificial intelligence (AI), particularly large language models (LLMs) such as ChatGPT (OpenAI), have emerged as a promising solution for generating accessible and context-appropriate textual summaries [2]. These models show promise in efficiently processing complex medical information and producing coherent summaries that minimize technical jargon while preserving essential clinical content [18,19].

The ability of LLMs to generate patient-centric summaries may help address critical barriers to patient engagement, specifically low health literacy levels and limited health care professional time for conversational patient education [8,12]. Automating this process with an LLM offers a potential solution that could be integrated into clinical workflows while preserving radiologists' time for critical interpretive tasks. However, important concerns persist regarding the completeness and accuracy of LLM-generated outputs, biases in the underlying training data, and the consequences of generating misleading clinical information. Therefore, effective deployment of such technology requires rigorous assessment of, and balance between, the quality and simplicity of these summaries, as well as their actual benefit to patients [20,21].

This study evaluates ChatGPT's capacity to produce patient-centered radiology report summaries through three primary aims: (1) assessing the technical accuracy and completeness of the summaries by a radiologist, (2) measuring improvements in patient comprehension and satisfaction, and (3) quantifying changes in readability using validated metrics. Through this evaluation, we aimed to provide insights into how LLM-driven text generation could advance patient-centered radiology while maintaining clinical accuracy, ultimately contributing to broader discussions about the role of AI in health care communication.

Methods

Radiology Report Collection and Anonymization

The electronic medical record was queried for radiology reports performed at our institution in September 2023. A total of 30 radiology reports were selected for analysis (6 brain magnetic resonance imaging, 6 brain computed tomography, 6 head and neck computed tomography angiography, 6 neck computed tomography, and 6 lumbar spine computed tomography) to reflect a broad range of imaging types and pathologies encountered in our neuroradiology section. Each radiology report was anonymized to remove all identifying information, including dates of comparison studies, if referenced.

Patient-Centric Summary Creation

Each radiology report was processed by ChatGPT-4.0 (OpenAI), accessed via the web interface in October 2023, within a new chat dialogue to ensure consistent performance of the model across each report. No custom instructions or memory features were activated; thus, the prompt template for each report conversion followed a zero-shot prompt format, as follows:

> *Please rewrite the following radiology report as a patient-friendly summary between 100 and 175 words. Start with a simple overall impression statement using natural language that is easy to understand. Focus on concisely summarizing the most relevant findings, conclusions, and recommended next steps. Minimize medical jargon, but clearly define any complex technical terms needed for accuracy. Provide enough details and context so that a layperson understands the key results and implications of the radiology report. Accurately reflect any potential abnormalities rather than omitting them. The summary should strike a balance between brevity and inclusion of significant information from the original report. The goal is an accessible and easy to understand summary that avoids oversimplification and remains clinically aligned with objective findings. Ensure the significance and severity of any abnormalities are accurately represented without downplaying or exaggerating. Only include next steps if you are highly confident they are appropriate recommendations for the patient.*

Recruitment

Members of the study team (JMH and AS) visited our health system's Patient and Family Advisory Council regularly scheduled meetings to outline the purpose and structure of our study and its focus on evaluating AI-generated summaries of radiology reports, and to ask for volunteers to participate. Four patients subsequently volunteered to complete survey questions

XSL•FO
RenderX

about the reports and summaries. Four board-certified neuroradiologists were directly recruited by the study team (JMH) to complete survey questions on the quality and accuracy of the summaries from a single institution.

Survey Design

The patient survey was a closed survey designed to separately evaluate patients' understanding, satisfaction, and perceived utility of the radiology reports and summaries. The radiologist survey was a closed survey designed to assess the technical accuracy, clinical completeness, clarity, and potential errors that may have been introduced into the summary. Both surveys included 5-point Likert-scale questions and open-ended responses. Surveys were created and administered electronically to all study participants using Qualtrics software, with each participant being sent a unique link to prevent multiple submissions. In the survey, each report and summary was clearly labeled as "Original Radiology Report" and "ChatGPT Patient-Centric Summary," respectively. The usability of the survey instruments was tested by the study team before deployment using Qualtrics. Complete survey instruments are provided in Multimedia Appendix 1.

Word Count and Readability Analysis

Textual characteristics of the original radiology reports and patient-centric summaries were analyzed via word count and validated readability metrics. These included the Flesch-Kincaid Grade Level, Flesch Reading Ease score, Gunning-Fog Index, and SMOG (Simple Measure of Gobbledygook) Index, which provide estimates of the complexity and accessibility of the reports and summaries [22-24].

Statistical Analysis

Differences in word count and readability scores between the reports and summaries were calculated with paired *t* tests using the Microsoft Excel Analysis ToolPak. For the survey questions, closed-ended responses were rated on Likert scales, with scores reported as percentages, and reflected as top 2 (or bottom 2) box scores as appropriate, aggregating responses from the Likert scale. Each patient and radiologist's response for each report and summary was considered an individual data point.

Data from the open-ended questions were analyzed using a multifaceted qualitative approach. To capture a comprehensive understanding of patient and provider perspectives, the method of analysis was based on the structure of each question. During inductive analysis, open-ended responses were coded manually, sorting responses by found themes. To ensure that information was not lost or assumptions about implied importance, each response could be coded to multiple themes.

To determine how well patients understood the original radiology report and summary, their responses to the question "What do you believe to be the most important/main point that the radiologist was trying to convey?" were directly compared for congruence with answers supplied by the radiologist for each report/summary pair. Comparisons were then categorized as congruent or incongruent with the radiologist's interpretation, which was considered the gold standard.

In addition, patient responses after reading the summary were compared to their answers after reading only the report. If the answers from the summary matched the answer from the report, responses were considered congruent (ie, the summary did not change the patient's subjective understanding). If the answers did not match, responses were considered incongruent (ie, the summary did change the patient's subjective understanding), with an additional positive or negative change modifier. A positive change modifier was assigned when responses included additional details or demonstrated closer alignment with the radiologist's responses. Conversely, a negative change modifier was assigned when responses contained fewer details or diverged further from the radiologist's responses.

Ethical Considerations

This study evaluated the quality and understandability of ChatGPT-generated radiology report summaries through both quantitative and qualitative analyses via surveys. Ethics approval was granted by the Colorado Multiple institutional review board (23 - 0731). All study participants were fully informed about this study, potential risks, and their right to withdraw at any time. All participants provided electronic informed consent. No compensation was provided to participants. All data were deidentified, and all analyses were conducted in accordance with data privacy guidelines.

Results

Readability

Patient-centric summaries demonstrated significant improvements in readability compared to original radiology reports across all measured metrics (Table 1). The summaries achieved a mean Flesch-Kincaid Grade Level of 7.2, representing a middle school reading level, compared to 12.5 for the original reports ($P<.001$). Similar improvements were observed in Flesch Reading Ease score, Gunning-Fog Index, and SMOG Index, consistently indicating enhanced accessibility of the summary content.

XSL•FO
RenderX

Table . Validated readability metric scores for the original radiology reports and ChatGPT-generated summaries.

Metric	Original report, mean (SD)	Summaries, mean (SD)	*P* value
Word count	241.0 (130.5)	189.0 (34.7)	.02
Flesch-Kincaid Grade Level	12.5 (1.63)	7.2 (1.14)	<.001
Flesch Reading Ease score	21.5 (11.39)	69.8 (6.2)	<.001
Gunning-Fog Index	16.4 (2.38)	10.3 (1.41)	<.001
SMOG[a] Index	12.7 (1.55)	10.4 (1.05)	<.001

[a]SMOG: Simple Measure of Gobbledygook.

Radiologist Survey

Four board-certified neuroradiologists completed surveys for each of the 30 radiology report/summary pairs, resulting in a total of 120 responses per question. The neuroradiologists had a mean of 11.6 (SD 9.33, range 3.5 - 25) years of experience in neuroradiology after completing a fellowship.

Radiologist assessments of the summaries were generally favorable (Table 2). In 83% (100/120) of responses, radiologists indicated that the summaries represented the original reports extremely well/very well, and 90% (108/120) were satisfied with the balance of medical jargon to patient-centric language. Regarding accuracy, 5% (6/120) of responses indicated missing information, 12% (14/120) indicated overemphasis of findings, and 18% (22/120) indicated underemphasis. Notably, no hallucinations or fabricated findings were identified in any summary. Regarding the potential for patient misunderstanding, 82% (99/120) of radiologist responses indicated the summaries would likely not lead to incorrect conclusions, while 10% (12/120) indicated a perceived risk of misinterpretation.

Table . Radiologist assessment of the quality and accuracy of ChatGPT-generated summaries (N=120).

Measure and response	Value, n (%)
Representation of original findings	
Extremely well/very well	100 (83)
Neutral	14 (12)
Slightly well/not well at all	6 (5)
Satisfaction with the balance of medical jargon and patient-centric language	
Very satisfied/somewhat satisfied	108 (90)
Neutral	5 (4)
Somewhat dissatisfied/very dissatisfied	7 (6)
Error analysis	
Missing relevant information	6 (5)
Overemphasis of findings	14 (12)
Underemphasis of findings	22 (18)
Risk of leading patients to incorrect conclusions	
Definitely yes/probably yes	12 (10)
Neutral	9 (8)
Probably not/definitely not	99 (82)

In their free-text responses, neuroradiologists indicated several issues with the summaries. They noted an overemphasis on minor findings (eg, mild sinus disease, white matter changes, and renal cysts) and perceived some follow-up recommendations as unnecessary. Radiologists also reported disagreement with specific terminology choices in some cases (eg, using “mildly enlarged” to describe a “stable to incrementally” enlarged collection). They further noted that summaries underemphasized some critical findings, such as arterial occlusions, perfusion deficits, and nerve root compression, by not adequately conveying their clinical significance. In addition, radiologists expressed concerns that oversimplified vocabulary diminished the importance of certain findings, such as describing a vertebral artery as “a smaller artery in your neck” or referring to a normalized brain volume at the 6th percentile as “a bit below average.”

Qualitative analysis of radiologist feedback revealed several themes regarding desired improvements to the summaries. Fifty-four percent (66/122) of free-text responses indicated no improvements were needed. In 38% (46/122), radiologists expressed discomfort with the perceived oversimplification of language (eg, “lump” instead of “nodule,” “thinning” instead of “osteopenia,” “squishing” instead of “height loss”), and

XSL•FO
RenderX

generally found the summaries could have been more concise. In 4% (5/122) of responses, radiologists indicated a preference for more detail, including more clearly defined terminology and more explicit localization of relevant findings (eg, the level of spinal stenoses).

Patient Survey

Four patient volunteers completed surveys for each of the 30 report-summary pairs, resulting in a total of 120 responses across all 30 reports and summaries for each question. A total of 118 responses were valid, with the remaining 2 responses left blank. Patients' reported age ranged between 55 and 84 years: 1 (55-64 years), 2 (65-74 years), and 1 (75-84 years). Patients' reported educational attainment included high school graduate (n=1), college degree (n=2), and master's degree (n=1).

Patient survey responses indicated substantially improved subjective comprehension and satisfaction with summaries compared to the original reports (Table 3). There were large, highly significant improvements in patient ratings when reading the summary versus the original report (P<.001). While only 26% (31/118) of responses indicated confidence in understanding the original reports, this increased to 98% (116/118) of responses for summaries. Satisfaction with medical terminology showed similar improvement, rising from 8% (9/118) for original reports to 91% (107/118) for summaries. Time required for comprehension was a significant concern with original reports, with only 23% (27/118) satisfied with the time needed for understanding, compared to 97% (115/118) satisfaction with summaries.

Table . Patient confidence and satisfaction with the original radiology reports and ChatGPT-generated summaries (N=118).

Measure and response	Original report, n (%)	Summaries, n (%)	*P* value
Confidence in understanding			<.001
Very confident/somewhat confident	31 (26)	116 (98)	
Neutral	14 (12)	1 (1)	
Very unconfident/somewhat unconfident	73 (62)	1 (1)	
Satisfaction with medical jargon/terminology			<.001
Very satisfied/somewhat satisfied	9 (8)	107 (91)	
Neutral	18 (15)	9 (8)	
Very dissatisfied/somewhat dissatisfied	91 (77)	2 (2)	
Satisfaction with level of detail			<.001
Very satisfied/somewhat satisfied	51 (43)	116 (98)	
Neutral	43 (36)	1 (1)	
Very dissatisfied/somewhat dissatisfied	24 (20)	1 (1)	
Satisfaction with time to comprehension			<.001
Very satisfied/somewhat satisfied	27 (23)	115 (97)	
Neutral	31 (26)	2 (2)	
Very dissatisfied/somewhat dissatisfied	60 (51)	1 (1)	

The summaries demonstrated high utility, with 92% (108/118) of responses indicating they clarified questions about the original report and improved patients' understanding of clinical implications. When asked about format preferences, 67% (79/118) of responses indicated a preference for having access to both the summary and original report, while 26% (31/118) of responses preferred to have the summary alone. Only 1% (1/118) preferred the original report exclusively, with the remaining 6% (7/120) indicating a preference for alternative formats.

Qualitative analysis of patient feedback revealed several themes regarding desired improvements to the summaries. Of the free-text responses, 50% (53/107) indicated no improvements were needed. The remaining responses suggested enhancing the professional tone (13%, 14/107), providing more specific follow-up recommendations (14%, 15/107), and including additional detail about medical terminology and physiological processes relevant to the findings (23%, 25/107).

Effect of the Summary on Patient Understanding

Patient understanding was assessed by comparing each patient's identification of the main point of the report/summary against the radiologist's response regarding the main point of the report. When reading the original radiologist report, 60% (71/119) of patient responses aligned with radiologist interpretations, while 40% (48/119) were incongruent. Patient understanding improved with the summary, with 74% (88/119) of responses aligning with radiologist interpretations, while 27% (32/119) remained incongruent.

XSL•FO
RenderX

When comparing patient responses regarding the main point the radiologist was trying to convey after reading only the report with their response after reading the summary, answers did not change in 52% (62/119) (ie, responses were congruent). In 29% (34/119), the responses after reading the summary were different (ie, the summary changed a patient's understanding). Specifically, in 23% (27/119), responses after reading the summary contained additional relevant details or demonstrated closer alignment with radiologists' responses, suggesting an increase in understanding as a result of the summary. In 6% (7/119), responses after reading the summary contained fewer details or diverged further from radiologists' responses, suggesting decreased understanding as a result of the summary.

Discussion

Principal Findings

This study demonstrates that LLMs can successfully generate neuroradiology reports that may be more understandable to patients while maintaining clinical accuracy. The AI-generated summaries significantly improved readability metrics across all measured scales, reducing the required reading level from college (grade 12.5) to middle school (grade 7.2), while substantially increasing reading ease scores [25,26]. This improvement in accessibility was achieved while maintaining high accuracy, with radiologists confirming that 83% (100/120) of summaries represented the original findings "extremely" or "very well." Similar studies have shown that such summaries maintain high accuracy [27,28].

A unique strength of this study is the inclusion of patients in the evaluation of reports and AI-generated summaries, as the vast majority of research focuses on evaluation by radiologists and nonradiologist physicians [2,29-31]. Patient response data revealed a striking contrast between comprehension of traditional reports and AI-generated summaries, with only 26% (31/118) of patients reporting confidence in understanding the original reports and 98% (116/118) reporting confidence in understanding the summaries. In addition, the high satisfaction rates with the summaries' level of detail (98%, 116/118) and medical terminology (91%, 107/118) indicate that the AI successfully struck a balance between accessibility and informativeness.

The dramatic improvement in patient confidence was accompanied by objective improvements in understanding, with congruence between patient and radiologist interpretations increasing from 60% (71/119) on the original reports to 74% (88/119) after reading the summaries. In addition, 23% (27/119) of responses after reading the summary contained additional relevant details or demonstrated closer alignment with radiologist responses, consistent with an increased understanding as a result of the summary. This suggests that AI-generated summaries not only increase patient confidence but also lead to a more accurate understanding of radiological findings. This aligns with prior evidence that patient-friendly summaries can enhance patient understanding; for example, nonphysician readers in one study had comprehension scores of 4.69/6 for AI-generated summaries versus 2.71/5 for original reports [27].

Particularly noteworthy was the strong patient preference for having access to both the original report and the summary (67%, 79/118 of responses), with 98% (116/118) of responses indicating they would use the summary if available. This suggests that patients value the combination of detailed technical information with accessible explanations, rather than seeing the summary as a replacement for the full report. These findings align with the shifting preference among the general public toward greater transparency in health care [3,4,7,8].

Importantly, no frank hallucinations were observed in the summaries, which is crucial for clinical implementation. However, a small percentage (10%, 12/120) of radiologist responses indicated that summaries may be misleading. Radiologists noted instances of omission and both over- and underemphasis of findings (12%, 14/120 and 18%, 22/120 of responses, respectively), particularly regarding incidental findings and critical abnormalities. The concern about misleading summaries was corroborated by the 6% (7/119) of patient responses that contained fewer details or diverged further from radiologist responses after reading the summary. Misleading information has the potential to lead to a variety of negative consequences for the patient, further emphasizing the need for comprehensive standards surrounding LLM-generated resources [20].

In addition, in some instances, language was deemed too casual by radiologists (eg, "squished" instead of "compressed"), highlighting the need for more sophisticated prompt engineering to maintain professional medical communication standards while improving accessibility. An area of improvement identified by the patient reviewers was highlighted by the 14% (15/107) of patient responses indicating a desire for more explicit guidance on next steps and implications of findings.

Limitations

This study has several limitations. Principally, its focus on neuroradiology reports from a single institution potentially limits the generalizability of our findings to other radiologic subspecialties and other institutional reporting patterns. In addition, the study included a relatively small number of patient and radiologist reviewers. We hope that this work encourages other investigators to replicate this study with larger cohorts of reviewers to assess its applicability in subspecialties outside of neuroradiology and in larger and more diverse patient populations. Patient volunteers were relatively older and had a relatively higher average education level than the general population and were members of our health institution's Patient Family Advisory Council. This likely indicates a higher level of health literacy and reading ability compared to the average patient. Overall, we anticipate that if anything, this biased the results toward greater understanding of the original radiology report, rather than less. Finally, volunteers were not reading their own reports, which removes the emotional salience and may alter the experience.

Considerations for Clinical Deployment

Clinical implementation of AI-generated summaries of radiology reports raises a number of ethical and practical considerations that must be addressed. In reference to ChatGPT, OpenAI policy

XSL•FO
RenderX

explicitly requires that AI-generated medical content undergo review by qualified professionals due to the risk of AI errors, which would necessitate radiologist or ordering provider oversight of all summaries before patients can access them [32]. The choice of model is incredibly important, as different models perform differently on such summarization tasks [28]. Clinical deployment requires HIPAA (Health Insurance Portability and Accountability Act)-compliant platforms (many are already available) and integration into existing electronic health record workflows without excessive burden on providers. Given the extraordinary burden already placed on radiologists to complete standard dictations amid increasing imaging volumes, adding review of an AI-generated summary to the workflow after final signature may not be feasible. It is important to note, however, that patients are already using consumer AI tools/chatbots to interpret their medical records without professional oversight, potentially receiving inaccurate or anxiety-provoking information [33]. Institution-supervised AI summaries could provide a safer alternative to this unregulated landscape while ensuring some degree of professional oversight and quality assurance. Future implementations must balance the demonstrated benefits of improved patient comprehension with robust safeguards.

Future Research

AI-generated, patient-friendly summaries appear to represent a promising tool for improving patient understanding of radiology reports. The summaries had high accuracy rates and dramatically improved readability and patient comprehension while maintaining clinical precision. The perceived benefit to patients in conjunction with the objective increase in comprehension suggests potential benefit in terms of patient experience and outcomes, though further investigation is needed using patients' own scans and reports (which we are pursuing) [1,3,8,10-12]. Future work should also investigate how prompt optimization and custom model development may improve the quality of outputs. Integration of these summaries into existing radiology workflows and patient portals should be studied, with particular attention to the ultimate impact on radiologist workload, patient experience, and outcomes. Conscientious integration of this technology is required, given the extraction of sensitive patient data and associated concerns [34].

Future research should explore options to continue refining the clinical accuracy of these reports. Though LLMs lack the ability to independently reason, there are several ways to optimize their output, including optimization of prompt engineering, incorporating structured templates for critical findings, and using standardized language for common pathologies [2,30,31]. In addition, providing the LLM with more clinical context (eg, recent note from an ordering provider, problem list, patient class [IP/OP/ED]) may improve the quality of the summary and recommendations provided [31,35].

Conclusions

This study demonstrates that ChatGPT-generated summaries can help bridge the communication gap between complex radiology reports and patient understanding while maintaining accuracy. While promising, our findings highlight the need for continued refinement to address instances of overemphasis or underemphasis, which can mislead patients in certain circumstances. As such, it is critically important to establish robust quality assurance frameworks before clinical implementation.

Data Availability

The datasets generalized and analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Complete patient and radiology survey instruments.
[DOCX File, 22 KB - formative_v9i1e76097_app1.docx]

References

1. Amin K, Khosla P, Doshi R, Chheang S, Forman HP. Artificial intelligence to improve patient understanding of radiology reports. Yale J Biol Med 2023 Sep;96(3):407-417. [doi: 10.59249/NKOY5498] [Medline: 37780992]
2. Doshi R, Amin KS, Khosla P, Bajaj SS, Chheang S, Forman HP. Quantitative evaluation of large language models to streamline radiology report impressions: a multimodal retrospective analysis. Radiology 2024 Mar;310(3):e231593. [doi: 10.1148/radiol.231593] [Medline: 38530171]
3. Alarifi M, Patrick T, Jabour A, Wu M, Luo J. Full radiology report through patient web portal: a literature review. Int J Environ Res Public Health 2020 May 22;17(10):3673. [doi: 10.3390/ijerph17103673] [Medline: 32456099]
4. Lee CI, Langlotz CP, Elmore JG. Implications of direct patient online access to radiology reports through patient web portals. J Am Coll Radiol 2016 Dec;13(12 Pt B):1608-1614. [doi: 10.1016/j.jacr.2016.09.007] [Medline: 27888949]
5. Halaska C, Sachs P, Sanfilippo K, Lin CT. Patient attitudes about viewing their radiology images online: preintervention survey. J Med Internet Res 2019 Jul 18;21(7):e12595. [doi: 10.2196/12595] [Medline: 31322124]
6. Kemp J, Short R, Bryant S, Sample L, Befera N. Patient-friendly radiology reporting-implementation and outcomes. J Am Coll Radiol 2022 Feb;19(2 Pt B):377-383. [doi: 10.1016/j.jacr.2021.10.008] [Medline: 35152963]

XSL•FO
RenderX

7. Farmer CI, Bourne AM, O'Connor D, Jarvik JG, Buchbinder R. Enhancing clinician and patient understanding of radiology reports: a scoping review of international guidelines. Insights Imaging 2020 May 5;11(1):62. [doi: 10.1186/s13244-020-00864-9] [Medline: 32372369]
8. Hansberry DR, Donovan AL, Prabhu AV, Agarwal N, Cox M, Flanders AE. Enhancing the radiologist-patient relationship through improved communication: a quantitative readability analysis in spine radiology. AJNR Am J Neuroradiol 2017 Jun;38(6):1252-1256. [doi: 10.3174/ajnr.A5151] [Medline: 28385883]
9. Yi PH, Golden SK, Harringa JB, Kliewer MA. Readability of lumbar spine MRI reports: will patients understand? AJR Am J Roentgenol 2019 Mar;212(3):602-606. [doi: 10.2214/AJR.18.20197] [Medline: 30620671]
10. Johnson AM, Brimhall AS, Johnson ET, et al. A systematic review of the effectiveness of patient education through patient portals. JAMIA Open 2023 Apr;6(1):ooac085. [doi: 10.1093/jamiaopen/ooac085] [Medline: 36686972]
11. Alturkistani A, Qavi A, Anyanwu PE, Greenfield G, Greaves F, Costelloe C. Patient portal functionalities and patient outcomes among patients with diabetes: systematic review. J Med Internet Res 2020 Sep 22;22(9):e18976. [doi: 10.2196/18976] [Medline: 32960182]
12. Bange M, Huh E, Novin SA, Hui FK, Yi PH. Readability of patient education materials from RadiologyInfo.org: has there been progress over the past 5 years? AJR Am J Roentgenol 2019 Oct;213(4):875-879. [doi: 10.2214/AJR.18.21047] [Medline: 31386570]
13. Qenam B, Kim TY, Carroll MJ, Hogarth M. Text simplification using consumer health vocabulary to generate patient-centered radiology reporting: translation and evaluation. J Med Internet Res 2017 Dec 18;19(12):e417. [doi: 10.2196/jmir.8536] [Medline: 29254915]
14. Schwartz LH, Panicek DM, Berk AR, Li Y, Hricak H. Improving communication of diagnostic radiology findings through structured reporting. Radiology 2011 Jul;260(1):174-181. [doi: 10.1148/radiol.11101913] [Medline: 21518775]
15. Norris EC, Halaska C, Sachs PB, Lin CT, Sanfilippo K, Honce JM. Understanding patient experiences, opinions, and actions taken after viewing their own radiology images online: web-based survey. JMIR Form Res 2022 Apr 25;6(4):e29496. [doi: 10.2196/29496] [Medline: 35468086]
16. McEnteggart GE, Naeem M, Skierkowski D, Baird GL, Ahn SH, Soares G. Readability of online patient education materials related to IR. J Vasc Interv Radiol 2015 Aug;26(8):1164-1168. [doi: 10.1016/j.jvir.2015.03.019] [Medline: 25935147]
17. Recht MP, Donoso-Bach L, Brkljačić B, Chandarana H, Jankharia B, Mahoney MC. Patient-centered radiology: a roadmap for outpatient imaging. Eur Radiol 2024 Jul;34(7):4331-4340. [doi: 10.1007/s00330-023-10370-3] [Medline: 38047974]
18. Liebrenz M, Schleifer R, Buadze A, Bhugra D, Smith A. Generating scholarly content with ChatGPT: ethical challenges for medical publishing. Lancet Digit Health 2023 Mar;5(3):e105-e106. [doi: 10.1016/S2589-7500(23)00019-5] [Medline: 36754725]
19. Yeo YH, Samaan JS, Ng WH, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. Clin Mol Hepatol 2023 Jul;29(3):721-732. [doi: 10.3350/cmh.2023.0089] [Medline: 35964695]
20. Goodman KE, Yi PH, Morgan DJ. AI-generated clinical summaries require more than accuracy. JAMA 2024 Feb 27;331(8):637-638. [doi: 10.1001/jama.2024.0555] [Medline: 38285439]
21. Parillo M, Vaccarino F, Beomonte Zobel B, Mallio CA. ChatGPT and radiology report: potential applications and limitations. Radiol med ;129(12):1849-1863. [doi: 10.1007/s11547-024-01915-7]
22. Wang LW, Miller MJ, Schmitt MR, Wen FK. Assessing readability formula differences with written health information materials: application, results, and recommendations. Res Social Adm Pharm 2013;9(5):503-516. [doi: 10.1016/j.sapharm.2012.05.009] [Medline: 22835706]
23. Flesch R. A new readability yardstick. Journal of Applied Psychology 1948 Jun;32(3):221-233. [doi: 10.1037/h0057532] [Medline: 18867058]
24. Sabharwal S, Badarudeen S, Unes Kunju S. Readability of online patient education materials from the AAOS web site. Clin Orthop Relat Res 2008 May;466(5):1245-1250. [doi: 10.1007/s11999-008-0193-8] [Medline: 18324452]
25. Haver HL, Gupta AK, Ambinder EB, et al. Evaluating the use of ChatGPT to accurately simplify patient-centered information about breast cancer prevention and screening. Radiol Imaging Cancer 2024 Mar;6(2):e230086. [doi: 10.1148/rycan.230086] [Medline: 38305716]
26. Chung EM, Zhang SC, Nguyen AT, Atkins KM, Sandler HM, Kamrava M. Feasibility and acceptability of ChatGPT generated radiology report summaries for cancer patients. Digit Health 2023;9:20552076231221620. [doi: 10.1177/20552076231221620] [Medline: 38130802]
27. Park J, Oh K, Han K, Lee YH. Patient-centered radiology reports with generative artificial intelligence: adding value to radiology reporting. Sci Rep 2024;14(1):13218. [doi: 10.1038/s41598-024-63824-z]
28. Tepe M, Emekli E. Decoding medical jargon: the use of AI language models (ChatGPT-4, BARD, microsoft copilot) in radiology reports. Patient Educ Couns 2024 Sep;126:108307. [doi: 10.1016/j.pec.2024.108307] [Medline: 38743965]
29. Kuckelman IJ, Wetley K, Yi PH, Ross AB. Translating musculoskeletal radiology reports into patient-friendly summaries using ChatGPT-4. Skeletal Radiol 2024 Aug;53(8):1621-1624. [doi: 10.1007/s00256-024-04599-2] [Medline: 38270616]
30. Jeblick K, Schachtner B, Dexl J, et al. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. Eur Radiol 2024 May;34(5):2817-2825. [doi: 10.1007/s00330-023-10213-1] [Medline: 37794249]

XSL•FO
RenderX

31. Lyu Q, Tan J, Zapadka ME, et al. Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: results, limitations, and potential. Vis Comput Ind Biomed Art 2023 May 18;6(1):9. [doi: 10.1186/s42492-023-00136-5] [Medline: 37198498]
32. Usage policies. OpenAI. URL: https://openai.com/policies/usage-policies/ [accessed 2025-06-20]
33. Consumer perceptions of AI in healthcare. UserTesting. URL: https://www.usertesting.com/resources/reports/consumer-perceptions-ai-healthcare [accessed 2025-06-20]
34. Patel SB, Lam K. ChatGPT: the future of discharge summaries? Lancet Digit Health 2023 Mar;5(3):e107-e108. [doi: 10.1016/S2589-7500(23)00021-3] [Medline: 36754724]
35. Campbell DJ, Estephan LE. ChatGPT, obstructive sleep apnea, and patient education. J Clin Sleep Med 2023;19(12):2133 [FREE Full text] [doi: 10.5664/jcsm.10808]

Abbreviations

AI: artificial intelligence
HIPAA: Health Insurance Portability and Accountability Act
LLM: large language model
SMOG: Simple Measure of Gobbledygook

Edited by A Mavragani; submitted 16.04.25; peer-reviewed by E Emekli, J Hershenhouse; revised version received 20.06.25; accepted 27.06.25; published 27.08.25.

Please cite as:
Sunshine A, Honce GH, Callen AL, Zander DA, Tanabe JL, Pisani Petrucci SL, Lin CT, Honce JM
Evaluating the Quality and Understandability of Radiology Report Summaries Generated by ChatGPT: Survey Study
JMIR Form Res 2025;9:e76097
URL: https://formative.jmir.org/2025/1/e76097
doi:10.2196/76097

© Alexis Sunshine, Grace H Honce, Andrew L Callen, David A Zander, Jody L Tanabe, Samantha L Pisani Petrucci, Chen-Tan Lin, Justin M Honce. Originally published in JMIR Formative Research (https://formative.jmir.org), 27.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Design and Evaluation of a Digital Health App (SingaporeWALK) for Active Aging: Pre-Post Intervention Study

Huanyu Bao[1], PhD; Sowmiya Meena Siva Subramanian[2], BTECH; Sai Ganesh Sarvotham Pai[3], PhD; Navrag B Singh[3], PhD; Kai Zhe Tan[3], PhD; Tan Phat Pham[2], MBA; Feihong Pan[1], MA; Yin-Leng Theng[2], PhD; Edmund Lee[1,4], PhD

[1]Department of Media & Communication, City University of Hong Kong, M5010, 5/F Run Run Shaw Creative Media Centre, 18 Tat Hong Avenue, Hong Kong, China (Hong Kong)

[2]Wee Kim Wee School of Communication and Information, Nanyang Technological University, Singapore, Singapore, Singapore

[3]Singapore-ETH Centre, Future Health Technologies Programme, CREATE Campus, Singapore, Singapore, Singapore

[4]Institute of Digital Medicine, City University of Hong Kong, Hong Kong, China (Hong Kong)

Corresponding Author:
Edmund Lee, PhD
Department of Media & Communication, City University of Hong Kong, M5010, 5/F Run Run Shaw Creative Media Centre, 18 Tat Hong Avenue, Hong Kong, China (Hong Kong)

Abstract

Background: The global trend toward population aging poses significant challenges for maintaining older adults' health and well-being, particularly in multicultural urban environments like Singapore. Despite the potential of digital health interventions, older adults face substantial barriers to technology adoption, including complex interfaces and culturally inappropriate content. Existing mobile health apps often fail to integrate physical, nutritional, and mental health components or accommodate the needs of multicultural older adult populations.

Objective: To address gaps in mobile app design for older adults and bridge the digital divide, this research aimed to create and evaluate the SingaporeWALK (SGWALK) app, a culturally inclusive digital health solution promoting active aging through community-based interventions among Singapore's older adults.

Methods: The SGWALK app was developed using participatory design methodology involving iterative testing with older adults to ensure appropriateness and usability. The app integrates 3 core components: exergames (Fruit Ninja, Piano Step, and Arctic Punch) aligned with Singapore's exercise guidelines for older adults, nutrition tracking based on local dietary recommendations, and mental well-being assessment using the Mental Health Continuum–Short Form. Following development, a 4-week pre-post intervention study was conducted with 48 participants (aged 60 - 85 y) randomly allocated to 4 conditions: conventional exercise, exergames only, exergames with health coach support, or exergames with peer support. In total, 5 wearable inertial measurement unit sensors captured movement data during weekly 30-minute supervised sessions at community centers. Primary outcomes included changes in physical activity metrics, technology acceptance, and mental well-being measured through pre- and postintervention assessments.

Results: The 4-week intervention demonstrated significant improvements across multiple health domains. Physical activity measures showed a 6.5% increase in maximum acceleration (t_{47}=3.82, P<.001), while nutritional tracking revealed steady improvements in healthy eating patterns throughout the intervention period. Mental health assessments indicated that participants classified as "mentally well" consistently outperformed the "moderate" group across physical activity measures. Technology acceptance showed substantial enhancement, with willingness to use health apps increasing from mean 3.18 (SD .79) to mean 3.95 (SD .82; t_{29}=−3.63, P<.001), and perceived ease of use improving from mean 3.01 (SD .70) to mean 3.76 (SD .68; t_{29}=−4.08, P<.001). Additionally, participants developed more efficient movement patterns over time and formed supportive social relationships during the community-based implementation.

Conclusions: The SGWALK app shows promise for promoting active aging and reducing technology barriers among Singapore's older adults. The community-based implementation model, bilingual interface, integrated health monitoring approach, and sensor-based movement tracking offer potential advantages over existing solutions. These findings provide useful insights for researchers and practitioners developing digital health interventions for older adult populations in multicultural urban settings.

(JMIR Form Res 2025;9:e68937) doi:10.2196/68937

KEYWORDS

mHealth; exergames; digital health technologies; wearables; digital divide; usability testing

XSL•FO
RenderX

Introduction

Background

Singapore's demographic landscape is undergoing a profound transformation, with citizens aged 65 years and older projected to constitute nearly a quarter of the population by 2030 [1]. This rapid aging presents both opportunities and challenges, necessitating innovative approaches to safeguard the health and autonomy of older adults. The concept of active aging—a multidimensional framework encompassing physical activity, nutrition, and mental well-being—has gained prominence as a strategy to mitigate age-related decline [2]. Research indicates that regular physical activity in older adults can significantly reduce the risk of chronic diseases, improve mental health, and enhance cognitive function [3]. Studies have highlighted the critical role of physical exercise in preventing cardiovascular diseases and promoting musculoskeletal health, which is essential for maintaining mobility and independence in later life [4,5]. However, translating these evidence-based benefits into practice remains fraught with barriers, particularly in technology adoption among older adults.

Despite the promise of digital health solutions, a critical obstacle is the pervasive digital divide affecting older adults. In Singapore, many older adults report hesitancy toward health technologies, citing concerns about privacy, usability, or perceived irrelevance to their needs [6]. While mobile health (mHealth) applications show potential, such as exergaming platforms to enhance physical activity [7] or wearables for mobility monitoring [8], most fail to accommodate the unique requirements of older adult users. Common shortcomings include cognitively demanding interfaces, lack of multilingual support, and fragmented features that neglect the synergistic relationship between physical, nutritional, and psychological health [9]. This gap in design inclusivity risks excluding the very population these tools aim to empower.

To address these limitations, we introduce SingaporeWALK (SGWALK), a mobile application co-designed with older adults using participatory methods to promote active aging within community settings. SGWALK integrates 3 core functionalities: (a) physical activity promotion through exergames such as Fruit Ninja and Piano Step, aligned with Singapore's national exercise guidelines for older adults; (b) nutrition tracking with goal-setting features focused on hydration, dairy intake, and balanced meals; and (c) mental well-being monitoring using validated psychological assessment tools. Accessibility is central to SGWALK's design, with bilingual (English–Chinese) interfaces, enlarged visual elements, and seamless integration with inertial measurement unit (IMU) sensors for real-time movement analysis.

By anchoring the intervention in community centers—important social hubs for Singaporean seniors—this study contributes to the field of gerontechnology in several key ways. First, it offers a replicable framework for co-designing digital health interventions with older adults, highlighting the importance of cultural relevance and usability. Second, it demonstrates the practical integration of exergaming, sensor-based mobility tracking, and holistic health monitoring in real-world community environments. Third, it provides empirical insights into addressing adoption barriers among older users, with broader implications for multicultural, urban aging populations.

Literature Review: Active Aging and Digital Health Interventions

The concept of active aging has evolved beyond mere physical health to encompass holistic well-being, including nutritional status and mental health [2]. Research demonstrates that technology-enabled interventions can effectively support these dimensions: physical activity apps reduce sedentary behavior in older adults [10], while nutrition-tracking tools improve dietary adherence in populations with diabetes [11,12]. However, adoption rates remain low among older adults, with studies attributing this to complex interfaces, privacy concerns [6], and lack of cultural adaptation [13]. In Singapore, where many older adults speak non-English languages at home, these barriers are exacerbated by linguistic mismatches in existing apps [6].

Among digital health interventions, exergaming has emerged as a particularly promising strategy to engage older adults in physical activity. Studies show Kinect-based games improve balance and upper-body mobility [7], while wearable sensors enable precise tracking of gait and fall risk [8]. However, most solutions focus on individual use, neglecting the social dynamics of community settings—a critical gap given that group activities increase adherence compared to solo interventions [10]. Similarly, nutrition apps for seniors often fail to account for age-related needs or local dietary habits [13], while mental health apps using validated questionnaires rarely integrate data with physical activity metrics [14].

To overcome these limitations, effective digital health tools for older adults must address several key design challenges. Accessibility remains paramount, with studies showing the importance of interfaces tailored to older adults' specific needs [9]. Simplified navigation is equally critical, as multi-step workflows increase abandonment rates compared to single-task interfaces [15]. Motivation presents the third challenge, where personalized feedback boosts engagement more than generic prompts [16].

Despite these known requirements, current active aging technologies exhibit persistent limitations that this study addresses. First, many existing apps treat physical, nutritional, and mental health as isolated domains [13], contrary to established active aging frameworks. Second, while existing apps claim multilingual support, few adapt content to local contexts like Singapore's food culture [6]. Third, most solutions prioritize individual use, despite evidence that group settings enhance adherence [10]. These gaps highlight the need for integrated, culturally adapted digital health solutions that can effectively engage older adults in community settings.

SingaporeWALK App Development

Overview

To address these identified limitations in existing digital health technologies, we developed the SGWALK app using a comprehensive, integrated approach (Figure 1). The architecture

XSL•FO
RenderX

of the SGWALK app establishes a connection between wearables and the mobile application via Bluetooth connectivity, thereby establishing a user-friendly platform for monitoring and managing health and mobility. At its essence, the SGWALK app serves as the main interface for end users, offering user-friendly functionalities to access activity data, establish personal goals, and participate in gamified exercises. To effectively capture and monitor users' physical activity, the SGWALK app interfaces with wearable IMU sensors. These sensors capture movement-related data such as accelerations and angular velocities, facilitating real-time data synchronization between the app and wearables to ensure accurate tracking of the users' mobility and health metrics. The architecture further incorporates a cloud-hosted Firebase database, serving as the backbone for storing and managing collected data.

Figure 1. Architecture diagram of SingaporeWALK (SGWALK) app showing data flow between wearables, mobile app, and cloud database. SDK: software development kit.

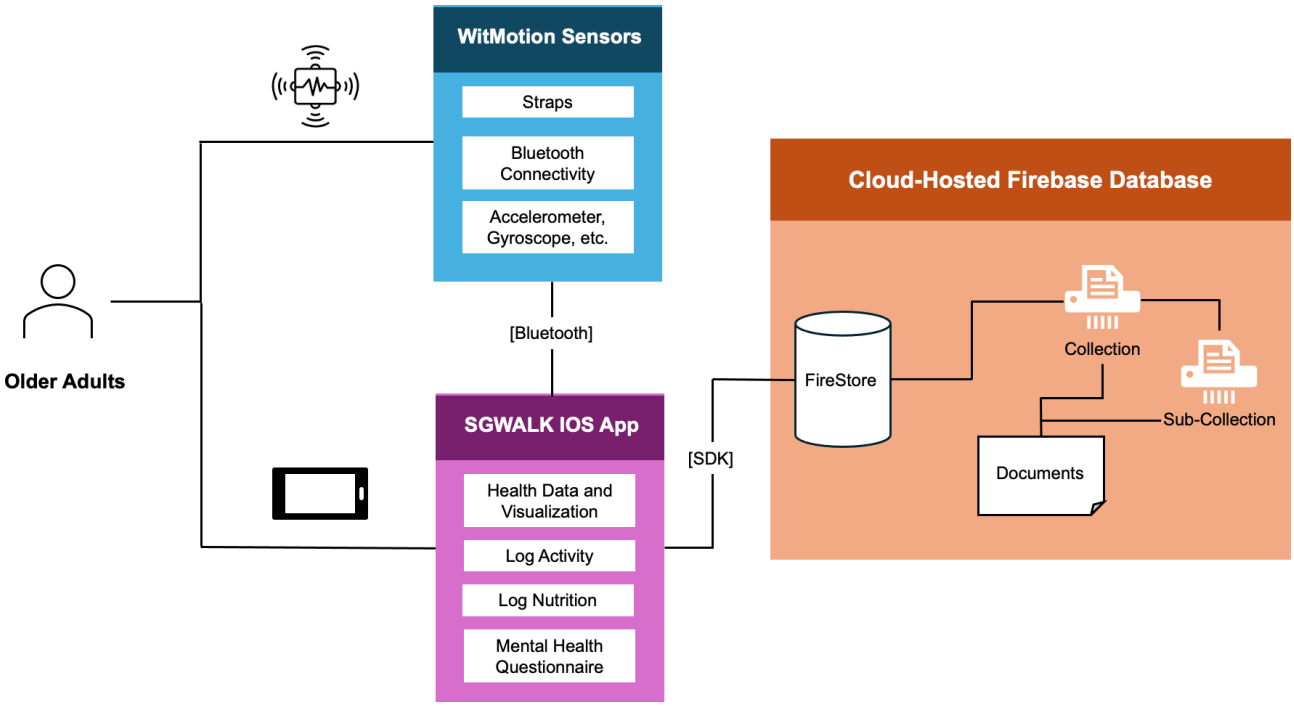

Main Components and Functionalities

The SGWALK app provides a comprehensive digital platform for older adults to track and manage their health across multiple domains. The system integrates 3 core components: wearable IMU sensors for motion capture, the mobile application interface, and cloud-based data storage.

The app incorporates three Xbox Kinect exergames specifically selected to align with Singapore's recommended exercises for older adults: (1) Fruit Ninja: players use arm movements to slice digital fruits, promoting upper body mobility and agility; (2) Piano Step: a rhythm-based game requiring synchronized leg movements to hit virtual piano keys, enhancing lower body coordination and balance; and (3) Arctic Punch: players punch ice cube targets to improve arm strength, precision, and coordination. These games were systematically selected based on their movement compatibility with conventional exercises while incorporating gamification elements to ensure sustained participation.

Beyond physical activity tracking, the app captures mental well-being data through standardized questionnaires and nutrition information via user-friendly input interfaces. All data is securely stored in a cloud-hosted Firebase database that enables real-time synchronization and long-term tracking across sessions. The app provides immediate feedback through data visualizations and supports goal-setting features to encourage sustained engagement.

User Interface Design Framework

The SGWALK interface was developed using a participatory design methodology involving 3 cycles of iterative testing with 32 older adults (mean age 69.4, SD 6.1 y). Grounded in the Fogg behavior model [16], the design emphasizes 3 key elements: motivation through gamification, ability through accessibility features, and triggers via context-aware prompts. This approach specifically addressed the digital disconnect identified in our needs assessment, where 78% of seniors reported abandoning health apps within 2 weeks due to frustration with complex interfaces.

The summary tab (Figure 2) presents integrated health metrics using visualizations refined through eye-tracking studies with older adults. Radial progress bars display daily activity goals, demonstrating 87% recognition accuracy versus 52% for numerical displays in usability testing, while sparklines depict weekly trends for mental well-being scores. The nutrition panel uses culturally adapted icons specifically designed for Singapore's context—a rice bowl for carbohydrates and a fish silhouette for protein—which improved comprehension by 41% compared to standard USDA icons during validation testing.

XSL•FO
RenderX

Navigation follows a strict 3-tap rule validated in gerontechnology literature, with the bottom-aligned menu bar persisting across all screens for consistent accessibility. Tactile feedback was implemented using Apple's haptic feedback API with medium intensity, a setting preferred by 83% of users in A/B testing. The log activity interface incorporates temporal landmarks ("After breakfast walk") rather than military time, reducing scheduling errors by 63% in field trials.

The mental health assessment module uses a 14-item adaptation of the Mental Health Continuum–Short Form (MHC-SF) with modified Likert scales using emoji anchors rather than text-only responses. This configuration increased completion rates from 58% to 89% compared to traditional text-only versions, significantly improving user engagement with mental health tracking. For nutrition tracking, the interface implements predictive text entry with Singaporean dish names and portion estimation using familiar hand-size references, specifically addressing the calorie underreporting commonly observed in older adult populations.

Three technical adaptations warrant particular attention for their impact on older adult usability. First, the dynamic text sizing system automatically adjusts layout elements when font sizes exceed 20 pt, preventing interface element overlap that commonly frustrates older adult users. Second, the context-sensitive "Help" button triggers contextual video tutorials demonstrating the current screen's functionality, providing immediate assistance without navigation complexity. Third, a specialized color correction algorithm adjusts interface hues based on ambient light sensor data to maintain minimum 4.5:1 contrast ratios under varying lighting conditions, ensuring consistent readability across different environments.

XSL•FO
RenderX

Figure 2. Summary of physical activity metrics (top), nutritional intake (middle), and mental well-being scores based on questionnaire responses (bottom).

XSL•FO
RenderX

Methods

Study Design and Participants

This study used a pre-post experimental design to evaluate the effectiveness of the SGWALK app in enhancing physical and mental well-being among older adults. This design was selected for several reasons: it allows for direct comparison of outcome variables before and after the intervention, provides preliminary evidence on usability and effectiveness for innovative technology, and accommodates practical constraints typical in early stages of health technology development [17-19].

Participants were Singaporean citizens or permanent residents aged over 60 years who were physically capable of standing and walking independently and proficient in either English or Chinese. Following written informed consent procedures, 51 participants were randomly allocated to one of four experimental conditions: (1) conventional exercise group, (2) exergames-only group, (3) exergames with health coach support group, or (4) exergames with peer support group. In total, 3 participants discontinued due to personal reasons, yielding a final sample of 48 participants (aged 60 - 85 y) who completed the full 4-week intervention protocol. The sample size was determined based on recruitment challenges for older adults in 4-week interventions, the exploratory nature of testing newly developed health technologies, and limited resources for large-scale studies [20,21]. Each participant received an honorarium of S$100 upon study completion.

Intervention Protocol

The 4-week intervention consisted of weekly 30-minute supervised sessions conducted at community centers with participants allocated to 4 conditions. The conventional exercise group participated in Singapore's evidence-based 7 Simple Exercises program targeting muscular strength and power enhancement. Exergaming groups used the SGWALK app, which integrates 3 core functionalities. Physical activity promotion was achieved through 3 Xbox Kinect exergames: Fruit Ninja (upper body slicing motions), Piano Step (rhythm-based stepping for lower body engagement), and Arctic Punch (full-body boxing for coordination). These games were selected to align with Singapore's recommended exercises for older adults. The app featured nutrition tracking with goal-setting for hydration, dairy intake, and balanced meals based on Singapore Health Promotion Board guidelines, and mental well-being assessment using the standardized MHC-SF [14]. Accessibility features included bilingual (English-Chinese) interfaces, enlarged visual elements, and real-time feedback through data visualizations.

Exergaming groups engaged with each game for 10 minutes per session, supervised by trained facilitators. The health coach-supported condition included real-time professional guidance for movement technique correction, performance feedback, motivational support, and safety monitoring. The peer-supported condition introduced collaborative gameplay to foster social connectedness and enhance motivation through peer modeling while maintaining equivalent physical exertion levels across all conditions.

Data Collection and Measures

Data collection involved multiple assessment methods across the 4-week intervention period. Movement kinematics were continuously monitored using 5 WitMotion IMUs with standardized placement: 2 sensors on the distal forearms (2 cm proximal to the ulnar styloid process), 1 at the L4 vertebral level on the waist, and 2 on the dorsal aspect of both feet. Sensors were secured using medical-grade adjustable straps and hypoallergenic adhesive tape to ensure consistent positioning without restricting natural movement.

Participants completed 20-minute pre- and postintervention questionnaires assessing self-reported physical health, mental well-being, and attitudes toward digital health technologies. Throughout the intervention, participants logged daily nutrition intake and completed weekly mental well-being assessments via the SGWALK app using the MHC-SF. The app provided real-time feedback on physical activity metrics, including action count and maximum acceleration, through line chart visualizations. For analysis purposes, mental health classification divided participants into "mentally well" and "moderate" groups based on MHC-SF scores [14]. Participants were classified as mentally well if they reported experiencing at least 1 of 3 hedonic signs and at least 6 of 11 eudemonic signs "every day" or "5 - 6 times a week."

Data Analysis

Raw motion data underwent processing to extract meaningful physical activity metrics using established techniques [22-24]. Key measures included:

- *z* score peak detection: This technique identifies significant activity changes by detecting peaks in the accelerometer data using the following formula: $z=(X - \mu)/\sigma$, where X is the data point, μ is the mean, and σ is the standard deviation.
- Action count: It is calculated as the normalized peak count for each game and date range using: Action count=(number of peak counts)/(number of data points)
- Maximum acceleration: The highest acceleration recorded during a specified date range: Maximum Acceleration=max(a), where a represents the acceleration data.
- Root mean square (RMS) acceleration: This provides a measure of overall movement intensity, calculated as: RMSAcceleration=1N∑i=1Nai2, where ai represents the acceleration data points and n is the total number of data points.
- Aggregate calorie count calculation: This estimates the total calorie expenditure for each game based on the accelerometer data. It involves calculating the metabolic equivalent of task (MET) from the RMS acceleration using the following formula: METs=1.8×RMS acceleration−15. Energy expenditure (in kcal/h)=1.05×METs×D× W, where D is the duration of activity and W is the participant's weight.

Statistical analyses used 2-tailed paired *t*-tests to compare pre- and postintervention measures. Physical activity trends were analyzed using temporal comparisons across the 4-week period, with separate analyses for mentally well and moderate

XSL•FO
RenderX

participant groups. Nutritional intake patterns and well-being metrics were examined through descriptive trend analysis. All analyses were conducted with a significance level set to identify statistically significant results (P<.05).

Ethical Considerations

Ethical approval was obtained from the Nanyang Technological University Institutional Review Board (IRB-2022‐734). Written informed consent was obtained from all participants after a full briefing on procedures, potential risks, and the right to withdraw without penalty. Participants received compensation of SGD $100 (US $77.73) for their involvement in the study. Data privacy and confidentiality were ensured through deidentification procedures, with participants registering usernames rather than real names in the app. All data were securely stored in the Firebase cloud database for research purposes only.

Results

During the study period, the SGWALK app provided real-time feedback on participants' physical activity metrics through the app interface, displaying key metrics such as action count and maximum acceleration. These metrics were presented as line charts for each game, enabling participants to track performance changes over time. Health coaches supplemented this technical data with interpretative guidance to help participants understand their performance data.

Postintervention data analysis examined how participants in the defined mental wellness categories ("mentally well" and "moderate") performed on physical activity metrics. The trend analysis (Figure 3) revealed that mentally well participants consistently demonstrated higher performance compared to the moderate group across most metrics throughout the study period. An important pattern emerged when examining temporal changes: while action count and calories burned gradually decreased over the 4 weeks for both groups, acceleration measurements showed a contrasting pattern with gradual increases over time. This inverse relationship suggests participants may have developed more efficient movement patterns as they became more familiar with the games, enabling them to achieve higher acceleration with fewer discrete actions.

Physical activity levels showed improvement across the 4-week intervention period. Maximum acceleration increased for both groups, with the mentally well group showing an increase from 18.7 to 19.1 (2.14% improvement) and the moderate group progressing from 17.7 to 19.4 (9.60% improvement), representing an overall 6.5% enhancement (t_{47}=3.82, P<.001). Although the moderate group showed a greater percentage increase, the mentally well participants maintained higher absolute levels throughout the study period. The heatmap analysis (Figure 4) confirmed consistent trends across all participants. While action counts and calories burned decreased over the 4-week period, maximum acceleration values remained stable or slightly increased, supporting the observation of more efficient movement patterns developing over time.

Figure 3. Trend of physical activity metrics over 4 weeks of study.

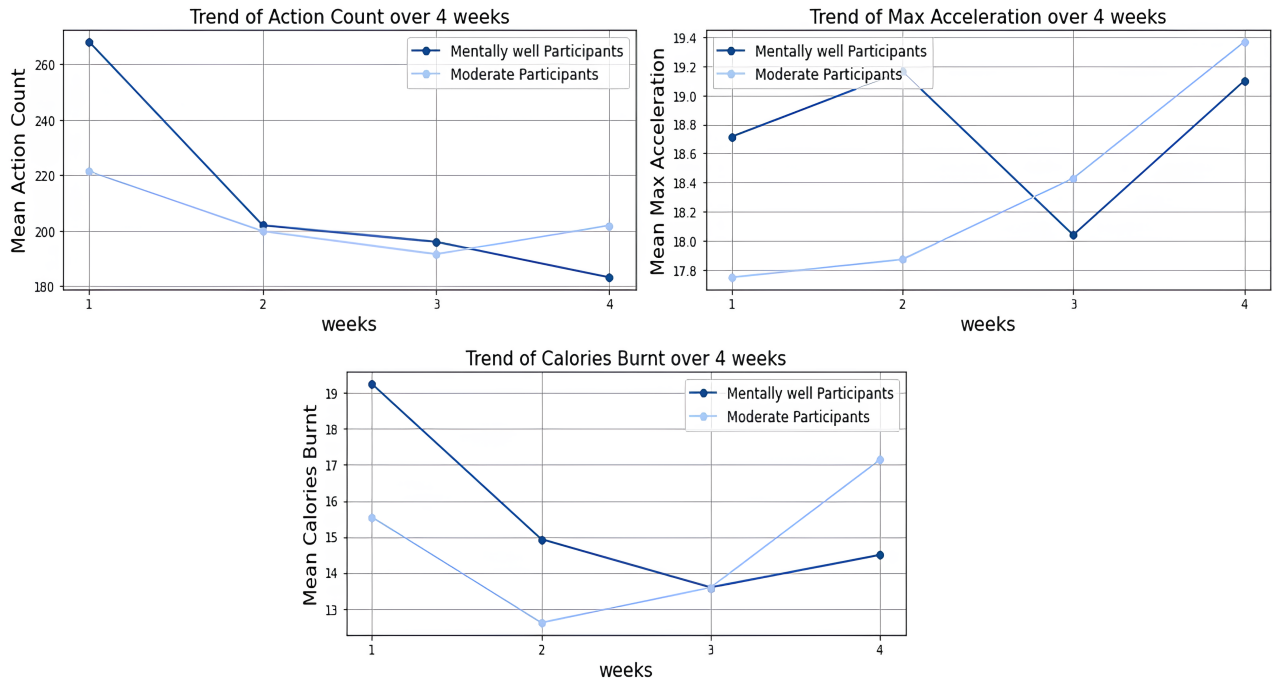

XSL•FO
RenderX

Figure 4. Heatmap of physical activity metrics over 4 weeks of study for all participants.

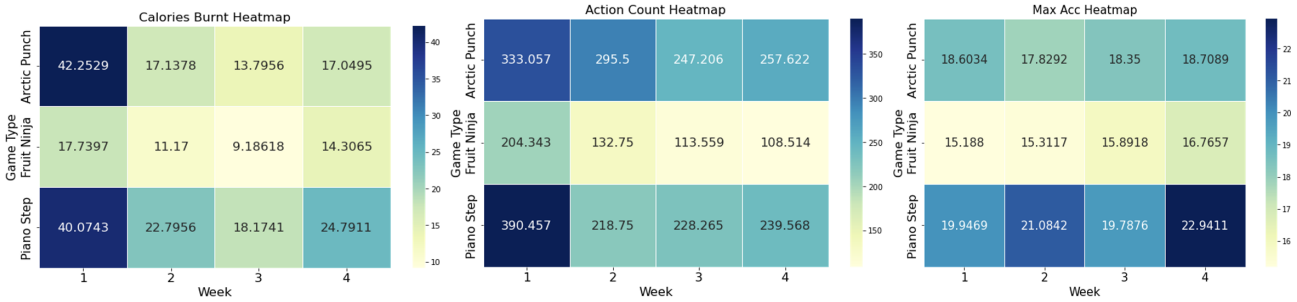

Nutritional intake data demonstrated positive trends throughout the four-week intervention. The analysis revealed that participants maintained a remarkably steady milk intake, indicating successful incorporation of dairy products into their dietary regimen. Additionally, water consumption showed a gradual yet consistent increase, suggesting improved hydration habits as participants progressed through the program. The most significant change was observed in healthy plate intake, which exhibited substantial increases over the four weeks. This upward trend reflects participants' growing commitment to balanced dietary practices and highlights their dedication to nutritional wellness during the intervention. Overall, these findings provide compelling evidence that the SGWALK app intervention effectively facilitated meaningful improvements in participants' dietary habits, potentially fostering patterns that could extend beyond the study period.

Analysis of emotional, psychological, and social well-being trends (Figure 5) revealed consistent differences between participant groups. Mentally well participants consistently reported higher mean scores across all 3 well-being dimensions compared to moderate participants throughout the intervention period. Emotional well-being scores maintained consistency for both groups throughout the intervention. In contrast, psychological and social well-being scores exhibited greater variability across weeks, particularly among moderate participants, suggesting these dimensions may be more responsive to external influences.

Figure 5.

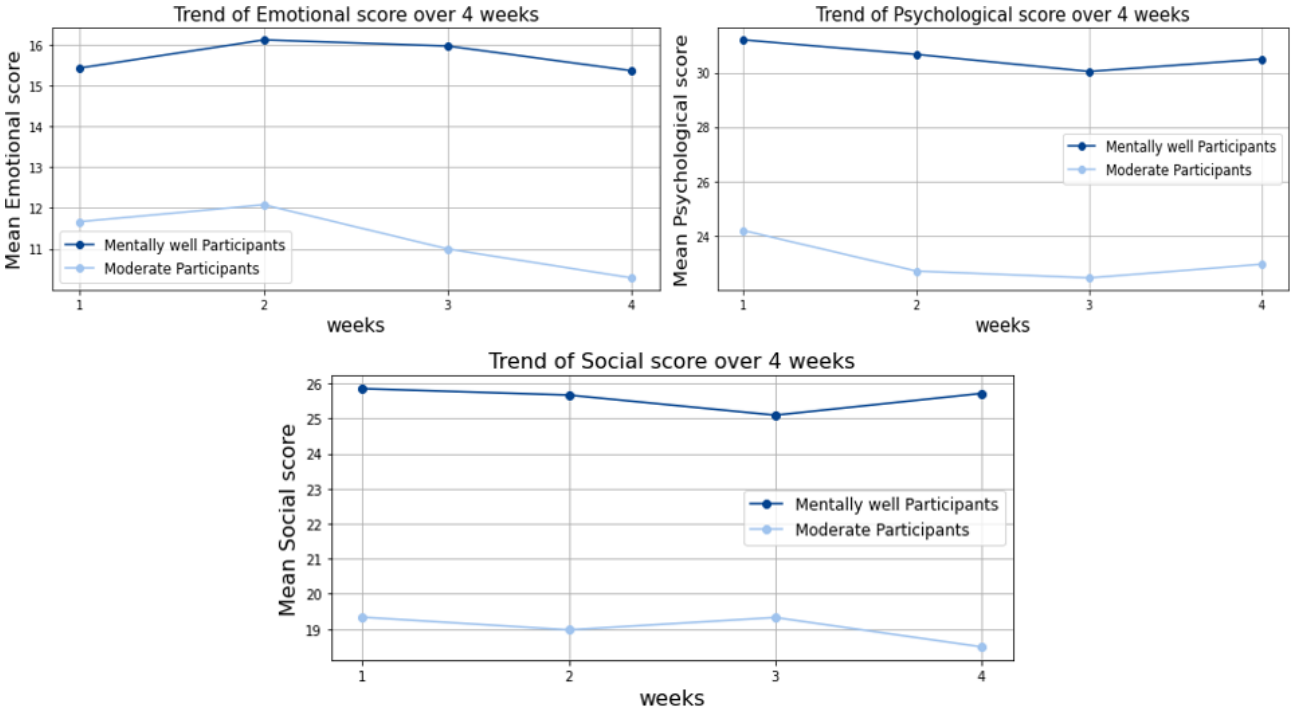

Analysis of *t* tests on questionnaire data revealed significant improvements across 6 of the 7 dimensions measured (Table 1). The most substantial improvements were observed in participants' perception of ease of use for health apps or wearables (t_{29}=−4.08, *P*<.001, *d*=1.08) and willingness to use health apps (t_{29}=−3.63, *P*<.001, *d*=0.96), both representing large effect sizes. Participants' attitudes toward using health apps or wearables also improved significantly from pre- to postintervention (mean 3.34-3.78, t_{29}=−2.35, *P*=.02, *d*=0.64). Additional significant improvements with medium effect sizes were found in participants' perception of how beneficial health technology has been in their lives (t_{29}=−2.70, *P*=.01, *d*=0.72) and their perceived availability of resources and knowledge about health apps and wearables (t_{29}=−2.53, *P*=.01, *d*=0.67). Participants also reported significantly better self-rated health over the past week following the intervention (t_{29}=−2.10, *P*=.04, *d*=0.55). Only one measure showed no significant change: frequency of feeling happy or satisfied with life over the past

XSL•FO
RenderX

week (t_{29}=−1.03, P=.31, d=0.28). Overall, these findings suggest the SGWALK intervention was effective in enhancing multiple aspects of health technology acceptance and self-perceived health among older adults.

Table . Paired *t*-test comparison between pre- and postsurvey responses. All tests were 2-tailed paired *t*-tests. Effect sizes (Cohen *d*) calculated using pooled standard deviation.

Item	Survey question	*t* test (*df*)	*P* value	Presurvey, mean (SD)	Postsurvey, mean (SD)	Cohen *d*
Q1	Rating own health over the past week	−2.10 (29)	.04	3.07 (.78)	3.48 (.70)	0.55
Q2	Frequency of feeling happy or satisfied with life over the past week	−1.03 (29)	.31	4.31 (1.05)	4.57 (.79)	0.28
Q3	Willingness to use health apps	−3.63 (29)	<.001	3.18 (.79)	3.95 (.82)	0.96
Q4	Perception of how beneficial health wearables or apps have been in their life	−2.70 (29)	.01	3.34 (.77)	3.86 (.67)	0.72
Q5	Perception of the ease of use of health apps or wearables	−4.08 (29)	<.001	3.01 (.70)	3.76 (.68)	1.08
Q6	Availability of resources or knowledge about health apps and wearables	−2.53 (29)	.01	2.92 (.84)	3.44 (.73)	0.67
Q7	Attitude toward using health apps or wearables	−2.35 (29)	.02	3.34 (.72)	3.78 (.66)	0.64

Discussion

Overview

The SGWALK intervention shows encouraging results for promoting active aging and digital health engagement among Singapore's older adult population. Statistical analyses revealed significant improvements across several domains, including participants' willingness to use health apps, perceived ease of use of health technologies, and self-reported health status. These findings suggest that carefully designed digital health solutions may help address some of the challenges associated with aging, particularly within Singapore's multicultural context.

The 4-week intervention was associated with improvements in participants' attitudes toward health technologies and self-perceived health status. T-test analyses indicated enhanced willingness to engage with health applications and increased perceived ease of use. These changes suggest that structured, supportive interventions may help reduce some older adults' initial hesitancy toward digital tools, though individual variation in response should be acknowledged. The app's integration of gamified exercises with visual feedback appeared to contribute to engagement, consistent with previous research demonstrating that gamified elements can enhance motivation in older adult populations [25,26]. The visual-based nutrition tracking features may have supported participants' reflection on dietary habits, though the mechanisms underlying these changes warrant further investigation [11].

An important finding was the relationship between mental well-being and physical activity performance observed throughout the intervention. Participants classified as "mentally well" consistently outperformed those in the "moderate" category across multiple physical activity indicators, including movement intensity, exercise completion rates, and overall engagement metrics. This pattern suggests a meaningful association between psychological well-being and physical performance capacity among older adults. This relationship aligns with existing literature linking flourishing mental health states to healthy lifestyle behaviors and physical activity engagement [27]. The observed correlation may reflect several underlying mechanisms: individuals with better mental health may possess greater motivation and self-efficacy for physical activities, while conversely, those experiencing moderate mental health challenges may face barriers such as reduced energy, lower confidence, or decreased intrinsic motivation for exercise participation. The bidirectional nature of this relationship is particularly relevant for intervention design. While causality cannot be definitively established from this study design, the findings suggest that mental well-being and physical activity may mutually reinforce each other in older adults. This interplay underscores the importance of holistic approaches when designing digital health interventions, as improvements in one domain may positively influence the other. The improvements

XSL•FO
RenderX

in technology acceptance observed are consistent with established theoretical frameworks [28,29], which emphasize perceived usefulness and ease of use as critical determinants of technology adoption. Our results suggest that structured, supportive interventions can shift perceptions even among individuals who may initially be resistant to technology use, reinforcing the value of targeted design and facilitation strategies in overcoming common barriers to digital health tool adoption among older adults.

When compared to existing active aging technologies such as the National Institute on Aging's Go4Life app [30] and SilverSneakers GO [31], SGWALK offers several distinct advantages. While these existing applications typically focus on isolated health dimensions, primarily exercise routines or fitness programs, SGWALK's integrated approach simultaneously addresses physical activity, nutrition, and mental well-being assessment. This comprehensive design enables a more holistic understanding of health status and provides users with a unified platform for health management. Several key innovations distinguish the SGWALK intervention from conventional approaches. First, the community-based implementation model goes beyond traditional individual-focused frameworks by embedding digital health engagement within group interaction settings. This strategy not only enhances social connectedness but also leverages peer dynamics to foster engagement and long-term use. Second, SGWALK's bilingual interface and culturally tailored design elements meet the diverse needs of Singapore's multicultural older population, ensuring inclusivity and accessibility. Third, by integrating multiple health domains, the app allows for cross-dimensional analyses, for example, how mental well-being relates to physical performance. Lastly, the use of IMU sensors in conjunction with exergames enables more precise and comprehensive tracking of movement compared to conventional step counters, offering richer insights into exercise quality and intensity.

Beyond the anticipated outcomes, the study revealed several unexpected yet important observations. While action counts and calories burned decreased over time, acceleration metrics increased, suggesting participants may have developed more efficient movement patterns as they became accustomed to the activities. This trend offers an interesting direction for future research into motor learning and exercise efficiency among older adults. Implementing SGWALK in community centers also highlighted the importance of social context in technology adoption. Participants showed higher engagement during group sessions, likely influenced by familiar environments and social interaction. Although these observations were qualitative, they point to the potential benefits of socially anchored digital interventions. The significant improvements in technology acceptance further support the idea that trusted, supportive environments, when paired with thoughtful exposure, can help overcome common barriers to adoption among older adults.

In total, 2 key design lessons emerged from the implementation process. First, striking the right balance between technological sophistication and ease of use is critical. While features like real-time feedback and activity tracking enhanced engagement for some, others found overly complex navigation or cognitive load challenging. Simplified interfaces, intuitive icons, and clear visual cues proved instrumental in facilitating uptake. Second, embedding the app within Singapore's established network of older adult activity centers was pivotal in supporting sustained participation. These familiar, supportive settings reduced apprehension, fostered peer motivation, and encouraged continuous engagement.

Limitations and Future Directions

While the SGWALK intervention shows promising results, several limitations should be acknowledged. The sample size of 48 participants, though sufficient for initial evaluation, may not fully represent the diversity of Singapore's older adult population. Future research should prioritize larger-scale studies with more representative sampling. The relatively brief 4-week intervention duration limits our understanding of long-term engagement patterns and sustained impact. Longitudinal studies spanning several months would provide valuable insights into the temporal evolution of usage patterns and determine whether observed benefits persist beyond the initial engagement period. Future development of the SGWALK platform should focus on several key areas. Enhancing cross-cultural adaptability would extend the app's applicability to different cultural contexts. Implementing personalization algorithms could tailor content based on individual user characteristics and preferences. Exploring integration with additional health monitoring technologies could expand the platform's health management capabilities. To support user engagement, future iterations should consider implementing additional gamification elements, more personalized feedback mechanisms, and features that enable social interaction within the app. As user numbers grow, robust data management infrastructure will become increasingly important to ensure data security while maintaining user privacy.

Conclusions

The SGWALK app represents a significant advancement in digital health interventions for older adults, offering a culturally responsive, holistic approach to promoting active aging through accessible technology. Our findings provide practical frameworks for adapting health technologies to multicultural older adult populations, implementing wearable sensor technology in community settings, and integrating physical activity tracking with mental well-being assessment. As global populations continue to age, interventions that successfully combine technological innovation with cultural sensitivity will become increasingly important. Digital health solutions that prioritize both effectiveness and accessibility can contribute to enhancing quality of life and promoting independence among older adults worldwide.

XSL•FO
RenderX

Acknowledgments

We want to express our heartfelt gratitude to our interns Ming Kiang Chua, Nigel Tan, Lim Boon Hian, and George Ong who played a crucial role in the development of the SingaporeWALK app.

Data Availability

The datasets generated and analyzed during the current study are not publicly available due to privacy concerns and ethical restrictions related to potentially identifiable information of older adult participants.

Conflicts of Interest

None declared.

References

1. Population trends. Department of Statistics Singapore. 2023. URL: https://www.singstat.gov.sg/publications/population/population-trends [accessed 2025-08-14]
2. Seah B, Kowitlawakul Y, Jiang Y, Ang E, Chokkanathan S, Wang W. A review on healthy ageing interventions addressing physical, mental and social health of independent community-dwelling older adults. Geriatr Nurs 2019;40(1):37-50. [doi: 10.1016/j.gerinurse.2018.06.002] [Medline: 29909022]
3. King AC, King DK. Physical activity for an aging population. Public Health Rev 2010 Dec;32(2):401-426. [doi: 10.1007/BF03391609]
4. McPhee JS, French DP, Jackson D, Nazroo J, Pendleton N, Degens H. Physical activity in older age: perspectives for healthy ageing and frailty. Biogerontology 2016 Jun;17(3):567-580. [doi: 10.1007/s10522-016-9641-0] [Medline: 26936444]
5. Bherer L, Erickson KI, Liu-Ambrose T. A review of the effects of physical activity and exercise on cognitive and brain functions in older adults. J Aging Res 2013;2013:657508. [doi: 10.1155/2013/657508] [Medline: 24102028]
6. Low STH, Sakhardande PG, Lai YF, Long ADS, Kaur-Gill S. Attitudes and perceptions toward healthcare technology adoption among older adults in singapore: a qualitative study. Front Public Health 2021;9:588590. [doi: 10.3389/fpubh.2021.588590] [Medline: 33659232]
7. Kappen DL, Mirza-Babaei P, Nacke LE. Older adults' motivation for physical activity using gamified technology: an eight-week experimental study. In: International Conference on Human-Computer Interaction: Springer; 2019:292-309. [doi: 10.1007/978-3-030-29387-1_22]
8. Kristoffersson A, Lindén M. A systematic review of wearable sensors for monitoring physical activity. Sensors (Basel) 2022 Jan 12;22(2):573. [doi: 10.3390/s22020573] [Medline: 35062531]
9. Czaja SJ, Charness N, Fisk AD, et al. Factors predicting the use of technology: findings from the center for research and education on aging and technology enhancement (CREATE). Psychol Aging 2015;21(2):333-352. [doi: 10.1037/0882-7974.21.2.333] [Medline: 16768579]
10. Burke SM, Carron AV, Eys MA, Ntoumanis N, Estabrooks PA. Group versus individual approach? A meta-analysis of the effectiveness of group-based physical activity interventions. Sport Exerc Psychol Rev 2006;2(1):13-29. [doi: 10.53841/bpssepr.2006.2.1.13]
11. Chen J, Gemming L, Hanning R, Allman-Farinelli M. Smartphone apps and the nutrition care process: current perspectives and future considerations. Patient Educ Couns 2018 Apr;101(4):750-757. [doi: 10.1016/j.pec.2017.11.011] [Medline: 29169863]
12. van Wijngaarden JP, Wojzischke J, van den Berg C, et al. Effects of nutritional interventions on nutritional and functional outcomes in geriatric rehabilitation patients: a systematic review and meta-analysis. J Am Med Dir Assoc 2020 Sep;21(9):1207-1215. [doi: 10.1016/j.jamda.2020.04.012] [Medline: 32723538]
13. Iancu I, Iancu B. Designing mobile technology for elderly. A theoretical overview. Technol Forecast Soc Change 2020 Jun;155:119977. [doi: 10.1016/j.techfore.2020.119977]
14. Lamers SMA, Westerhof GJ, Bohlmeijer ET, ten Klooster PM, Keyes CLM. Evaluating the psychometric properties of the Mental Health Continuum-Short Form (MHC-SF). J Clin Psychol 2011 Jan;67(1):99-110. [doi: 10.1002/jclp.20741] [Medline: 20973032]
15. Morris JM. User interface design for older adults. Interact Comput 1994 Dec;6(4):373-393. [doi: 10.1016/0953-5438(94)90009-4]
16. Fogg BJ. A behavior model for persuasive design. Presented at: Proceedings of the 4th International Conference on Persuasive Technology; Apr 26-29, 2009; Claremont, CA, USA. [doi: 10.1145/1541948.1541999]
17. Chow PI, Showalter SL, Gerber M, et al. Use of mental health apps by patients with breast cancer in the United States: pilot pre-post study. JMIR Cancer 2020 Apr 15;6(1):e16476. [doi: 10.2196/16476] [Medline: 32293570]
18. Chow PI, Showalter SL, Gerber MS, et al. Use of mental health apps by breast cancer patients and their caregivers in the United States: protocol for a pilot pre-post study. JMIR Res Protoc 2019 Jan 14;8(1):e11452. [doi: 10.2196/11452] [Medline: 31344674]

XSL•FO
RenderX

19. Byambasuren O, Beller E, Hoffmann T, Glasziou P. mHealth app prescription in Australian general practice: pre-post study. JMIR Mhealth Uhealth 2020 Jun 1;8(6):e16497. [doi: 10.2196/16497] [Medline: 32478660]
20. El Ayadi AM, Duggal M, Bagga R, et al. A mobile education and social support group intervention for improving postpartum health in northern India: development and usability study. JMIR Form Res 2022 Jun 29;6(6):e34087. [doi: 10.2196/34087] [Medline: 35767348]
21. Hendrickx R, van der Avoird T, Pilot P, Kerkhoffs G, Schotanus M. Exergaming as a functional test battery in patients who received arthroscopic ankle arthrodesis: cross-sectional pilot study. JMIR Rehabil Assist Technol 2021 May 5;8(2):e21924. [doi: 10.2196/21924] [Medline: 33949311]
22. Kang X, Huang B, Qi G. A Novel Walking Detection and Step Counting Algorithm Using Unconstrained Smartphones. Sensors (Basel) 2018 Jan 19;18(1):297. [doi: 10.3390/s18010297] [Medline: 29351223]
23. Carneiro S, Silva J, Aguiar B, et al. Accelerometer-based methods for energy expenditure using the smartphone. Presented at: 2015 IEEE International Symposium on Medical Measurements and Applications (MeMeA); May 7-9, 2015; Turin, Italy. [doi: 10.1109/MeMeA.2015.7145190]
24. Ainsworth BE, Haskell WL, Herrmann SD, et al. 2011 Compendium of physical activities: a second update of codes and MET values. Med Sci Sports Exerc 2011 Aug;43(8):1575-1581. [doi: 10.1249/MSS.0b013e31821ece12] [Medline: 21681120]
25. Aschentrup L, Steimer PA, Dadaczynski K, Mc Call T, Fischer F, Wrona KJ. Effectiveness of gamified digital interventions in mental health prevention and health promotion among adults: a scoping review. BMC Public Health 2024 Jan 2;24(1):69. [doi: 10.1186/s12889-023-17517-3] [Medline: 38167010]
26. Wong CK, Kwok RW. The effect of gamified mhealth app on exercise motivation and physical activity. Presented at: PACIS 2016 Proceedings; Jun 27 to Jul 1, 2016; Chiayi, Taiwan.
27. Hone LC, Jarden A, Duncan S, Schofield GM. Flourishing in New Zealand workers: associations with lifestyle behaviors, physical health, psychosocial, and work-related indicators. J Occup Environ Med 2015 Sep;57(9):973-983. [doi: 10.1097/JOM.0000000000000508] [Medline: 26340286]
28. Steele R, Lo A, Secombe C, Wong YK. Elderly persons' perception and acceptance of using wireless sensor networks to assist healthcare. Int J Med Inform 2009 Dec;78(12):788-801. [doi: 10.1016/j.ijmedinf.2009.08.001] [Medline: 19717335]
29. Kim D. Can healthcare apps and smart speakers improve the health behavior and depression of older adults? A quasi-experimental study. Front Digit Health 2023;5:1117280. [doi: 10.3389/fdgth.2023.1117280] [Medline: 36910571]
30. Herman SW. Go4life: http://go4life.nia.nih.gov. J Consum Health Internet 2014;18:271-278. [doi: 10.1080/15398285.2014.932184]
31. Hamar B, Coberley CR, Pope JE, Rula EY. Impact of a senior fitness program on measures of physical and emotional health and functioning. Popul Health Manag 2013 Dec;16(6):364-372. [doi: 10.1089/pop.2012.0111] [Medline: 23560492]

Abbreviations

IMU: inertial measurement unit
MET: metabolic equivalent of task
MHC-SF : Mental Health Continuum–Short Form
mHealth: mobile health
RMS: root mean square
SGWALK: SingaporeWALK (Wearables and Apps for community Living and Knowledge)

Edited by A Mavragani; submitted 18.11.24; peer-reviewed by R Tieben; revised version received 27.05.25; accepted 28.05.25; published 22.08.25.

Please cite as:
Bao H, Siva Subramanian SM, Pai SGS, Singh NB, Tan KZ, Pham TP, Pan F, Theng YL, Lee E
Design and Evaluation of a Digital Health App (SingaporeWALK) for Active Aging: Pre-Post Intervention Study
JMIR Form Res 2025;9:e68937
URL: https://formative.jmir.org/2025/1/e68937
doi:10.2196/68937

© Huanyu Bao, Sowmiya Meena Siva Subramanian, Sai Ganesh Sarvotham Pai, Navrag B Singh, Kai Zhe Tan, Tan Phat Pham, Feihong Pan, Yin-Leng Theng, Edmund Lee. Originally published in JMIR Formative Research (https://formative.jmir.org), 22.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Balance Improvement and Fall Risk Reduction in Stroke Survivors After Treatment With a Wearable Home-Use Gait Device: Single-Arm Longitudinal Study With 1-Year Follow-Up

Brianne Darcy[1], PT, DPT, DHSc; Lauren Rashford[1], PT, DPT; David Huizenga[1], PhD; Kyle B Reed[2], PhD; Stacy J M Bamberg[1], ScD

[1]Moterum Technologies, Inc., 2757 S 300 W Ste F, Salt Lake City, UT, United States

[2]Department of Mechanical Engineering, University of South Florida, Tampa, FL, United States

Corresponding Author:
Brianne Darcy, PT, DPT, DHSc
Moterum Technologies, Inc., 2757 S 300 W Ste F, Salt Lake City, UT, United States

Abstract

Background: Falls are a common and serious problem after stroke, often leading to injuries, loss of independence, and increased health care usage. Functional balance, a primary risk factor for falls, is frequently impaired in individuals with hemiparetic gait impairments. Previous research with the iStride gait device (Moterum Technologies, Inc) showed that functional balance improved immediately following 4 weeks of treatment. However, the long-term retention of these effects remains unknown and could improve the management of balance and mobility impairments after stroke.

Objective: This study aimed to determine the long-term functional balance effects of treatment with the gait device for individuals with hemiparetic gait impairments from stroke.

Methods: Eighteen individuals with chronic stroke (9 male, 9 female, mean age 57 years, and 60 months post stroke) participated in twelve 30-minute treatment sessions with the gait device. During each treatment session, the device was worn on the less affected lower extremity during overground ambulation in the participant's home. All treatment and assessments were overseen by licensed physical therapists. Functional balance was evaluated using the Berg Balance Scale (BBS), the Timed Up and Go (TUG) test, and the Functional Gait Assessment (FGA) at baseline and 5 posttreatment follow-ups: 1 week, 1 month, 3 months, 6 months, and 12 months after treatment. Balance improvement was analyzed using repeated-measures ANOVA from baseline to each follow-up time frame, correlation analysis, comparison to each outcome's minimal detectable change (MDC) value, evaluation of fall risk classification changes, and subjective questionnaires.

Results: Participants retained statistically significant improvements on the BBS, TUG, and FGA compared with baseline at all posttreatment time frames ($P<.05$). All participants initially identified as being at risk for falls reduced their fall risk on at least one outcome during one or more follow-up assessments. At 12 months post treatment, the average improvement on all 3 outcomes remained above their respective MDC thresholds, demonstrated by a 5.9-point improvement on the BBS, a 4.9-second improvement on the TUG, and a 34.6% (3.8-point) improvement on the FGA. At least 72% of participants exceeded the MDC of BBS, at least 44% exceeded the MDC of TUG, and at least 66% exceeded the MDC of FGA at every posttreatment time point. Subjective questionnaire responses indicated that 88% of participants perceived functional balance improvement following treatment with the gait device.

Conclusions: The findings of this study indicate that treatment with the gait device may result in long-term functional balance improvement for individuals with hemiparetic gait impairments from stroke. Larger, controlled studies are recommended to confirm these findings.

Trial Registration: ClinicalTrials.gov NCT03649217; https://www.clinicaltrials.gov/study/NCT03649217

(JMIR Form Res 2025;9:e67297) doi:10.2196/67297

KEYWORDS

stroke; falls; balance; gait device

Introduction

Falls are a common and serious problem after stroke. In the first year following a stroke, approximately 70% of individuals fall [1-3], and the risk of falls remains substantially elevated throughout the stroke survivor's lifespan [2]. The physical consequences of falls can be devastating, often leading to hospitalizations, serious injuries, and even death [4]. Stroke

XSL•FO
RenderX

survivors who fall are also especially vulnerable to adverse outcomes from falls, such as a 4-fold increase in fracture risk compared with healthy controls [5], and more severe fracture consequences [2].

In addition to physical complications, the psychological consequences of falls after stroke can be severe. Compromised mobility confidence and fear of falling are experienced by up to 88% of stroke survivors who have fallen, and 44% report further restricting their mobility activities after a fall [2]. Given the documented low activity levels of stroke survivors in general [6], further activity restriction reinforces limitations with functional independence. Furthermore, the psychological consequences of falls are associated with depression [3,7], reduced community participation [8], social deprivation [2], and decreased quality of life [4,9], all of which can be detrimental to a stroke survivor's well-being.

For individuals with stroke, gait and balance impairments are widely regarded as primary fall risk factors [2,10]. Impairments to muscle strength, sensation, proprioception, vision, cognition, and attention [2,11] can further compound this risk and present additional barriers to normalizing mobility and fall risk. As primary risk factors, balance and gait-focused activities are commonly prescribed treatments to reduce fall risk post stroke. However, there is a dearth of high-quality evidence to support the efficacy of specific balance and gait-focused activities [12-14]. As improved medical management continues to extend the life expectancy of stroke survivors [15], successful treatments for balance improvement and fall risk management are essential to improve the health and safety of these individuals.

The iStride gait device [16] (Moterum Technologies, Inc.) was developed to support the rehabilitation needs of the ambulatory stroke population. Specifically, the device targets the asymmetric movement patterns of individuals experiencing hemiparesis, a notable contributor to gait and balance impairments in neurologic populations [17,18]. The device is worn over the shoe on the foot of the nonparetic limb during ambulation. Four rotating, kinetic wheels [19] on the device generate a steady backward translation of this limb during weight bearing in the stance phase of gait. Given the typical reliance or "favoring" of the nonparetic limb by individuals with hemiparesis during ambulation [20], this stance phase motion induces a relative instability, prompting a subtle weight transfer to the affected limb. Resembling principles of lower extremity constraint-induced movement therapy [21], this weight transfer is designed to increase usage of the paretic limb during the gait cycle. In addition, the backward motion of the wheels increases the stride length of the affected limb. These exaggerated mechanics trigger the user to correct this pattern (similar to the aftereffect experienced with split-belt treadmill training [22]) once the device is removed and the individual continues overground ambulation without the device. Further details of the device's development and mechanisms of action are explained in previously published studies [19,23-25]. Figure 1 shows the gait device and its motion during the gait cycle.

The documented effects of treatment with the gait device include improvements to gait symmetry, specifically step length symmetry and double support symmetry [24,26]. Gait speed has also been shown to increase immediately posttreatment [24,27,28] with effects largely unchanged one year after treatment completion [29]. Additional benefits have been noted in the realm of functional balance. Specifically, a study conducted with participants using the gait device in their home environment found statistically significant improvements on balance-focused functional outcomes following 12 treatment sessions with the gait device. Moreover, fall risk reduction was experienced by 80% of participants, as determined by comparison to fall risk thresholds on one or more outcome measures [27]. While these posttreatment outcomes are promising for an immediate benefit of gait device treatment, retention of these improvements remains unknown and could provide further guidance on the long-term utility of this treatment and its role in managing poststroke balance and mobility impairments. Therefore, the objective of this study was to investigate the long-term effects of gait device treatment on functional balance and risk for falls for individuals with hemiparesis caused by stroke.

XSL•FO
RenderX

Figure 1. The iStride gait device and mechanism of action. (**A**) The iStride gait device. (**B**) The gait device motion. As the user bears weight into the device during ambulation, the device's rotating wheels push the nonparetic foot backward during stance. This backward motion promotes usage of the paretic leg by slightly destabilizing the nonparetic leg. To adjust for the added height of the device, a similar height but stationary platform is worn on the foot of the paretic limb.

Methods

Participant Recruitment

Community-dwelling individuals with chronic stroke were recruited to participate in the study using convenience sampling. Inclusion criteria specified: (1) age of 21 - 80 years, (2) one or more cerebral strokes (all on the same side), (3) stroke occurred at least 6 months previously, (4) gait asymmetry but can walk independently with or without a cane, (5) no evidence of severe cognitive impairment that would interfere with understanding instructions, (6) not currently receiving physical therapy, (7) no evidence of 1-sided neglect affecting ambulation, (8) adequate walking space within the home, and (9) body weight

XSL•FO
RenderX

less than 250 pounds. Individuals were excluded from the study if they had: (1) uncontrolled seizures, (2) pregnancy, (3) metal implants (stents), (4) chronic obstructive pulmonary disease, (5) uncontrolled high blood pressure, (6) myocardial infarction within the last 180 days, (7) head injury within the last 180 days, or (8) a history of a neurologic disorder other than stroke.

Ethical Considerations

The study was confirmed to meet ethical standards for research with human participants by the Western institutional review board (IRB), now named Western-Copernicus Group IRB. Participants signed a consent form that was approved by Western IRB before participating in the study. Data were coded without personally identifying information and stored using HIPAA (Health Insurance Portability and Accountability Act) compliant software. Compensation for participation in study activities included US $6 for each study visit, and US $50, US $75, and US $250 for follow-up assessment sessions after the treatment period.

Experimental Setup

The study followed a single group, a before-after design, with multiple follow-ups after the treatment period. Determination of eligibility included an initial phone screen by the principal investigator, followed by a home visit to verify adherence with eligibility criteria and to assess that the home environment met the spatial needs for gait device treatment (at least 25 feet of walking space).

Each participant's functional balance was measured at baseline (approximately 1 week before starting device treatment). After baseline assessments, the participants completed approximately 4 weeks of gait device treatment followed by 5 follow-up balance assessments: 1 week, 1 month, 3 months, 6 months, and 12 months after treatment. At the 12-month follow-up session, participants were provided a questionnaire regarding their clinical trial experience, fall history before and after treatment with the gait device, and subjectively observed balance changes. All aspects of this study were performed in each participant's home environment and were overseen by licensed, nonemployee physical therapists hired as contractors for clinical trial data collection. No other treatment or physical therapy services were provided to the participants as a part of this study during or after the treatment period.

Balance Assessments

Various outcome measures have been validated to assess unique aspects of functional balance [30], and using several balance-focused outcomes is commonly engaged in research and clinical practice to assess balance ability comprehensively [30,31]. The Berg Balance Scale (BBS), the Timed Up and Go (TUG) test, and the Functional Gait Assessment (FGA) were selected in the present study due to their high clinical usage [32,33], validation and recommended usage in the chronic stroke population [34,35], and ability to be performed in the home environment. In addition, each of these outcome measures provides a clinically relevant threshold value to differentiate high and low risk of falls, an objective of this study, as well as criteria depicting the necessary magnitude of improvement to exceed measurement error (ie, the minimal detectable change [MDC]). Each of these 3 outcomes has also been identified as having high-quality (Level 1) evidence to support usage in neurologic populations, and the BBS and FGA are recommended as core outcomes in clinical practice guidelines published by the Academy of Neurologic Physical Therapy [35].

The BBS [36] is a 14-item measure that assesses static balance, dynamic balance, and fall risk. This commonly used and widely studied outcome has various published fall risk and MDC threshold values. For our study, we selected a fall risk threshold value of 44 based on a 2011 study by Simpson et al [37], which studied a similar population of ambulatory, community-dwelling individuals with stroke. Similarly, an MDC score of 2.5 points [38] was selected for our comparison from several alternative thresholds [34,39,40] as their sample provides the most comparable "time since stroke" of approximately 46 months.

The TUG test [41] assesses dynamic balance and walking ability, and this outcome measure has been shown to differentiate individuals with stroke from healthy individuals [42]. A score above 14 seconds is used for predicting falls after stroke [43], and a threshold MDC value of −3.2 seconds indicates a change beyond measurement error for individuals with chronic stroke [34].

The FGA is used to assess postural stability during walking [44]. The FGA does not have a specific fall risk cutoff for the chronic stroke population. Therefore, we selected a related threshold of 15 points [45], used for individuals with Parkinson disease, another neurologic disorder affecting gait and balance. A threshold value of 14.1% or 4.2 points, determined from a sample of individuals with stroke [46], is available for MDC comparison.

Treatment Sessions

Following the baseline balance assessment, participants underwent treatment with the gait device 3 times per week over 4 weeks, totaling 12 treatment sessions. During each treatment session, the device was worn on the participant's nonparetic foot, and an approximately height-matched shoe was worn on the paretic foot. Participants walked on the gait device for a target of 30 minutes during each session with rest breaks at scheduled 5-minute intervals, or more frequently if needed. Licensed physical therapists supervised all sessions and provided the level of mobility assistance needed for participant safety and comfort while walking on the gait device.

Statistical Methods

The sample size for this long-term follow-up study is based on the number of individuals who participated in the 4-week treatment and who continued in the study for the year following treatment. The sample size for the short-term study [27] was derived using power analyses from 2 previous studies with the gait device [23,24]. In the first study [23], the analysis focused on the comparison between pretreatment and posttreatment data in healthy individuals without mobility impairments. It revealed an effect size of 0.68 for the difference in step length asymmetry. These calculations estimated that a sample size of 18 participants would obtain a power of 0.8. The second study [24], based on a pilot in-clinic study using the device with individuals with stroke, calculated an effect size of 0.71 for gait speed. A power

XSL•FO
RenderX

analysis based on gait speed showed that 21 participants would obtain a power of 0.85. In the present study, 21 participants completed the treatment. After the treatment period, 3 participants did not complete all remaining follow-up sessions. Since we are reporting repeated-measures statistical tests, this study reports results from the 18 participants who completed all outcome assessments at all time periods.

Statistical tests included both parametric (repeated-measures ANOVA) and nonparametric (Friedman) tests to determine the statistical significance of outcome measure score changes between baseline and each of the 5 follow-up time frames. The Friedman test was used when the data violated the assumptions of an ANOVA test. When statistical significance was found, post hoc pairwise comparisons were performed using Tukey honestly significant difference or the Wilcoxon signed-rank test for nonparametric (Friedman) tests. The Spearman rank-order correlation coefficient was used for correlation analysis. All statistical analyses were calculated using SPSS Version 26 software (IBM Corp).

To evaluate the risk for falls, we compared each participant's outcome measure score at each time frame to the related fall risk threshold of each outcome measure. In addition, to assess the magnitude of outcome measure change compared with the measurement error of each test, we compared both our study group average and each individual's score changes to the threshold MDC value for each test. Finally, survey responses were manually tabulated for the reported number of falls and the percentage of positive or negative responses to questionnaire items.

Results

Participant Characteristics

Figure 2 shows the study activities and the number of participants in each study phase from recruitment through follow-up. Recruitment occurred between July 2018 and September 2018. Treatment occurred between July 2018 and December 2018, and all study-related follow-up was completed in December 2019. The results and analyses of this study include the 18 study participants who completed all assessments through the 12-month follow-up. The demographics of the study participants are shown in Table 1.

Figure 2. Number of study participants at each stage of the research study. Key study activities are listed within each phase, and reasons for nonparticipation are available to the right of each phase heading. BBS: Berg Balance Scale; FGA: Functional Gait Assessment; TUG: Timed Up and Go.

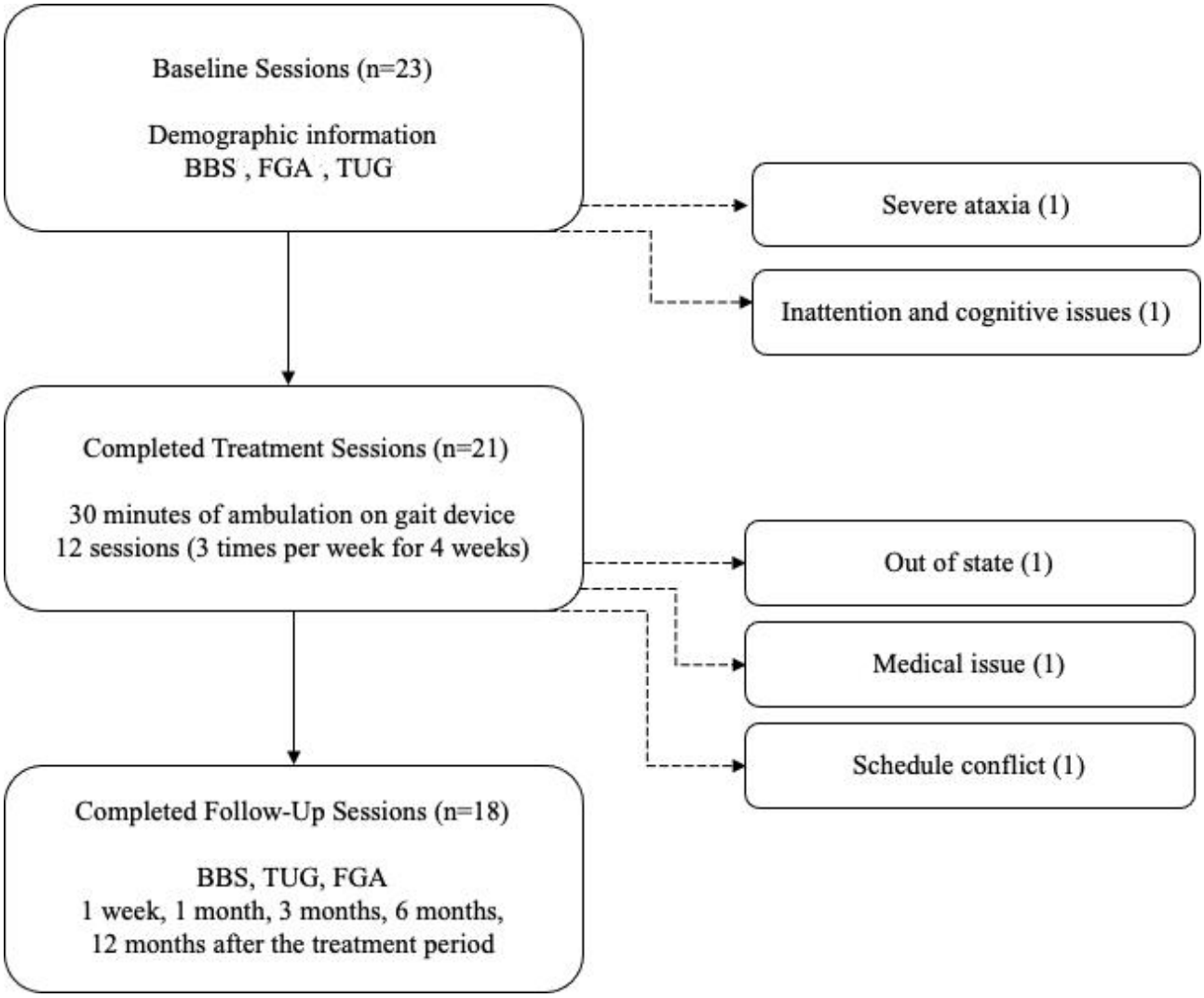

XSL•FO
RenderX

Table . Participant demographics.

ID	Sex	Weight (lbs)	Age (years)	Time since stroke (months)	Side of hemiparesis	AFO[a]	Treatment duration (minutes)
A	M[b]	160	53	24	Left	No	295
B	M	198	55	13	Right	Yes	355
C	F[c]	200	44	14	Left	No	270
D	M	240	51	50	Left	Partial[d]	360
E	M	235	52	130	Right	Yes	360
F	F	134	46	308	Left	No	360
G	M	220	61	22	Left	Yes	290
H	M	190	63	53	Left	Yes	285
I	F	250	58	92	Left	No	360
J	F	222	69	89	Left	No	330
K	F	189	47	28	Right	Partial[d]	215
L	F	172	53	32	Left	No	360
M	F	183	50	80	Left	Yes	210
N	M	134	61	46	Left	Partial[d]	165
O	F	150	54	21	Right	No	280
P	F	180	77	15	Left	No	170
Q	M	175	64	30	Right	Yes	260
R	M	188	62	28	Left	No	295
Total	18 (9 M and 9 F)	Mean (SD) 190 (34)	Mean (SD) 57 (9)	Mean (SD) 60 (70)	5 Right 13 Left	6 Yes 9 No 3 Partial	Mean (SD) 290 (66)

[a]AFO: Ankle-Foot Orthosis.

[b]M: male.

[c]F: female.

[d]Partial:participantss who used an AFO occasionally at baseline, but did not use during the study.

Statistical Analysis

Mauchly Test of Sphericity showed significance, indicating a violation of one of the assumptions of a repeated-measures ANOVA for the TUG and BBS (but not the FGA); therefore, Greenhouse-Geisser corrections were applied. Further, TUG did not show normally distributed results; therefore, we used the Friedman test for the TUG (BBS and FGA were normally distributed for all except one time frame). Using repeated-measures ANOVA, our findings showed statistical significance from baseline to each follow-up time frame for BBS ($F_{2.5,42.8}$=19.5; P<.001) and FGA ($F_{5,85}$=18.3; P<.001). Friedman test also showed statistical significance from baseline to each follow-up time frame for TUG, (χ^2_5=33.0; P<.001). Table 2 shows the outcome measure means at each assessment session, outcome measure score changes relative to baseline, and associated P values. Respective MDC values and fall risk thresholds are provided in the table for reference. Note that the percentage difference calculation for the FGA is the average of each participant's percent change score.

The relationship between the duration of treatment with the gait device (in min) and balance assessment scores at each posttreatment time period was assessed using Spearman rank-order correlation coefficient. Table 3 shows that statistically significant moderate to strong correlations were found between treatment minutes and the majority of balance scores at each of the 5 follow-up sessions. Statistically significant relationships are indicated in the table below.

XSL•FO
RenderX

Table . Statistical analysis, from baseline to each posttreatment time frame.

Values	Baseline	1 wk post treatment	1 mo post treatment	3 mo post treatment	6 mo post treatment	12 mo post treatment
BBS[a] (pts[b]) MDC[c]=2.5 pts [38]; fall risk≤44 pts [37]						
Mean (SD)	42.6 (6.0)	46.7 (4.8)	48.5 (4.8)	49.3 (5.2)	48.8 (5.4)	48.5 (5.8)
Difference	__[d]	4.1	5.9	6.7	6.2	5.9
P value	—	.01[e]	<.001[e]	[e]<.001[f]	<.001[e]	.001[e]
TUG[f] (sec) MDC=−3.2 sec [34]; fall risk≥14 sec [43]						
Mean (SD)	19.7 (8.6)	14.9 (6.1)	14.4 (6.0)	14.4 (5.8)	14.1 (6.6)	14.8 (6.9)
Difference	—	−4.8	−5.3	−5.3	−5.6	−4.9
P value	—	<.001[e]	.001[e]	.001[e]	.004[e]	.05[e]
FGA[g] (pts)						
Mean (SD)	15.1 (5.1)	20.1 (4.6)	20.7 (4.3)	21.7 (4.5)	20.4 (4.5)	18.9 (4.5)
MDC=4.2 pts or 14.1% [46]						
Difference	—	5.0	5.6	6.6	5.3	3.8
Fall risk≤15 pts [45]						
Difference	—	41.4	48.2	56.5	45.4	34.6
P value	—	<.001[e]	<.001[e]	<.001[e]	<.001[e]	.005[e]

[a]BBS: Berg Balance Scale.
[b]pts: points.
[c]MDC: minimal detectable change.
[d]Not applicable.
[e]Statistically significant, $P<.05$.
[f]TUG: Timed Up and Go.
[g]FGA: Functional Gait Assessment.

XSL•FO
RenderX

Table . Correlation analysis between balance scores at each time period and minutes of treatment.

Outcome and balance measurement	Minutes of treatment	
	Spearman ρ	*P* value
BBS[a]		
BBS 1 wk post treatment	0.574[b]	.01
BBS 1 mo post treatment	0.624[b]	.006
BBS 3 mo post treatment	0.634[b]	.005
BBS 6 mo post treatment	0.442	.07
BBS 12-mo post treatment	0.400	.10
TUG[c]		
TUG 1 wk post treatment	−0.549[b]	.02
TUG 1 mo post treatment	−0.613[b]	.007
TUG 3 mo post treatment	−0.640[b]	.004
TUG 6 mo post treatment	−0.560[b]	.02
TUG 12 mo post treatment	−0.749[b]	<.001
FGA[d]		
FGA 1 wk post treatment	0.118	.64
FGA 1 mo post treatment	−0.036	.89
FGA 3 mo post treatment	0.157	.53
FGA 6 mo post treatment	0.252	.31
FGA 12 mo post treatment	0.485[b]	.04

[a]BBS: Berg Balance Scale.

[b]Statistically significant correlation, *P*<.05.

[c]TUG: Timed Up and Go.

[d]FGA: Functional Gait Assessment

Balance Improvement

Overview

The average score on each balance measure at each time period is shown in Figures 3A-3C, with the respective threshold fall risk range indicated with gray shading for comparison. For the BBS, the average score improved from below the fall risk threshold (ie, high risk for falls) at baseline to above the fall risk threshold (ie, low risk for falls) at each of the 5 follow-ups. For the TUG, the average score at each time period improved by more than 4 seconds and approached the fall risk threshold score of 14 seconds [43]. Finally, the average FGA score of 15.1 approximates the fall threshold score of 15 points [45] at baseline but improves over 3 points beyond the threshold at each posttreatment assessment.

Figure 3. Average outcome measure score across all study participants (n=18) at each time period. (**A**) Average Berg Balance Scale score at each study time period. (**B**) Average Timed Up and Go score at each study time period. (**C**) Average Functional Gait Assessment score at each study time period. Error bars represent 95% CIs, and shading corresponds to each outcome's fall risk threshold. BBS: Berg Balance Scale; FGA: Functional Gait Assessment; TUG: Timed Up and Go. Post: Posttreatment.

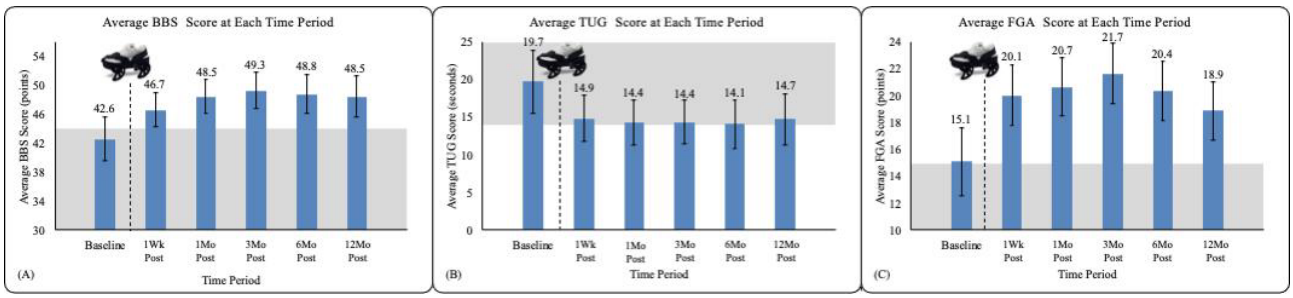

XSL•FO
RenderX

Improvement Relative to the Minimal Detectable Change

The MDC determines the smallest amount of change that is not likely due to chance variation in measurement [47]. Therefore, we sought to determine if the magnitude of posttreatment measurement changes (for each participant at each time frame) exceeded this metric. Table 4 shows a comparison to the MDC. Note that some individuals scored highly on the balance measures at baseline (such as participant R, who scored 8.8 s on the TUG, 52 on the BBS, and 26 on the FGA at baseline). Therefore, exceeding an MDC threshold may not have been expected for some participants. Nonetheless, findings show that at least 72% (13/18) of participants exceeded the MDC of BBS, at least 44% (8/18) exceeded the MDC of TUG, and at least 66% (12/18) exceeded the MDC of FGA at every posttreatment time point.

Table . Improvement relative to the minimal detectable change (✓ indicates improvement beyond minimal detectable change, and — indicates a change less than minimal detectable change).

ID	1 wk post treatment			1 mo post treatment			3 mo post treatment			6 mo post treatment			12 mo post treatment			%[a]
	BBS	TUG	FGA	BBS	TUG	FGA	BBS	TUG	FGA	BBS	TUG	FGA	BBS[b]	TUG[c]	FGA[d]	
A	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	100
B	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	100
C	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	—	✓	93
D	✓	✓	✓	✓	✓	✓	✓	✓	—	✓	✓	✓	✓	✓	✓	93
E	✓	—	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	93
F	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	—	✓	✓	—	✓	87
G	✓	✓	✓	✓	—	✓	✓	✓	✓	✓	✓	✓	✓	✓	—	87
H	✓	✓	—	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	—	87
I	✓	—	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	—	✓	✓	87
J	✓	—	✓	✓	✓	✓	✓	—	✓	✓	✓	✓	—	—	✓	73
K	—	✓	✓	—	✓	✓	—	—	✓	✓	✓	—	✓	✓	✓	67
L	✓	—	✓	✓	—	✓	✓	—	✓	✓	—	✓	✓	—	✓	67
M	✓	—	✓	✓	—	✓	✓	—	✓	✓	—	✓	✓	—	✓	67
N	—	✓	✓	—	✓	✓	—	✓	✓	✓	—	✓	—	—	✓	60
O	—	✓	—	✓	✓	—	✓	✓	—	✓	✓	—	✓	—	—	53
P	✓	—	✓	✓	✓	✓	✓	—	—	—	✓	✓	—	—	—	53
Q	—	—	✓	—	—	✓	—	—	✓	—	—	✓	✓	—	—	33
R	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0
%[e]	72.2	55.6	83.3	77.8	72.2	88.9	77.8	61.1	77.8	83.3	66.7	83.3	72.2	44.4	66.7	

[a]Percentage of time periods each individual participant exceeded the MDC.

[b]BBS: Berg Balance Scale.

[c]TUG: Timed Up and Go.

[d]FGA: Functional Gait Assessment.

[e]Percentage of participants that improved greater than the MDC at each time period.

Fall Risk Classification

The number of participants classified as high fall risk by each assessment's criteria is shown in Figure 4. For the BBS, 11 (61%) participants were categorized as a high fall risk at baseline [38], but only 2-4 participants remained at high fall risk after treatment. Thirteen (72%) participants scored as high fall risk using TUG criteria at baseline, and this number was reduced to 4-9 participants at each posttreatment time period. Similarly, while 11 (61%) participants scored at high fall risk at baseline on the FGA, only 1-3 participants remained at high risk for falls after treatment with the gait device. Of the 15 participants who were identified as being at risk for falls at baseline, 15 out of 15 (100%) reduced their risk for falls on at least one outcome (during at least one posttreatment measurement).

XSL•FO
RenderX

Figure 4. Number of participants at high risk for falls, as determined by the Berg Balance Scale, the Timed Up and Go test, and the Functional Gait Assessment, at each study time period. The Functional Gait Assessment uses the Parkinson disease fall risk threshold. Post: posttreatment.

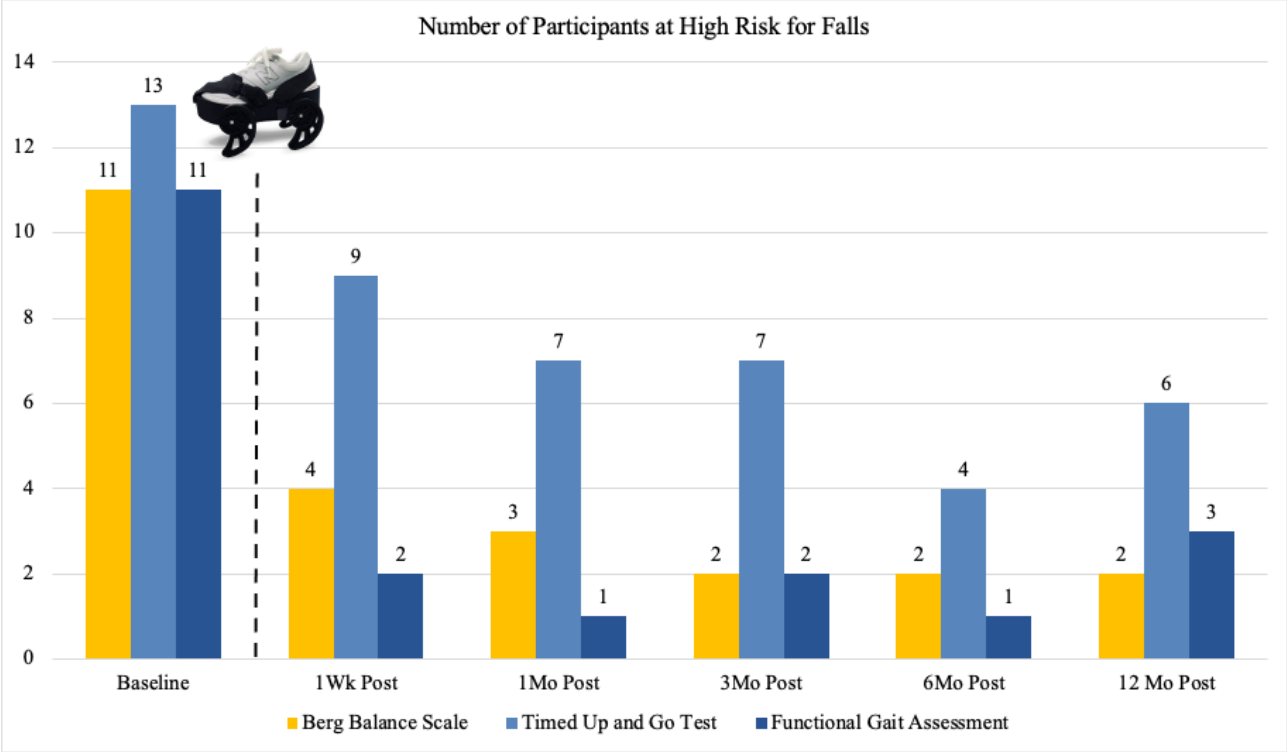

Outcome Measure Agreement

Fall risk classifications on all 3 functional balance assessments are combined into Figure 5. Findings show a sustained and progressive increase in the number of participants who scored as low risk for falls on all measures through the 6-month follow-up. At 12 months post treatment, the number of participants at low risk for falls remains substantially above baseline.

Figure 5. Number of participants at low risk for fall on all 3 outcome measures (the Berg Balance Scale, the Timed Up and Go test, and the Functional Gait Assessment), at each study time period. Post: posttreatment.

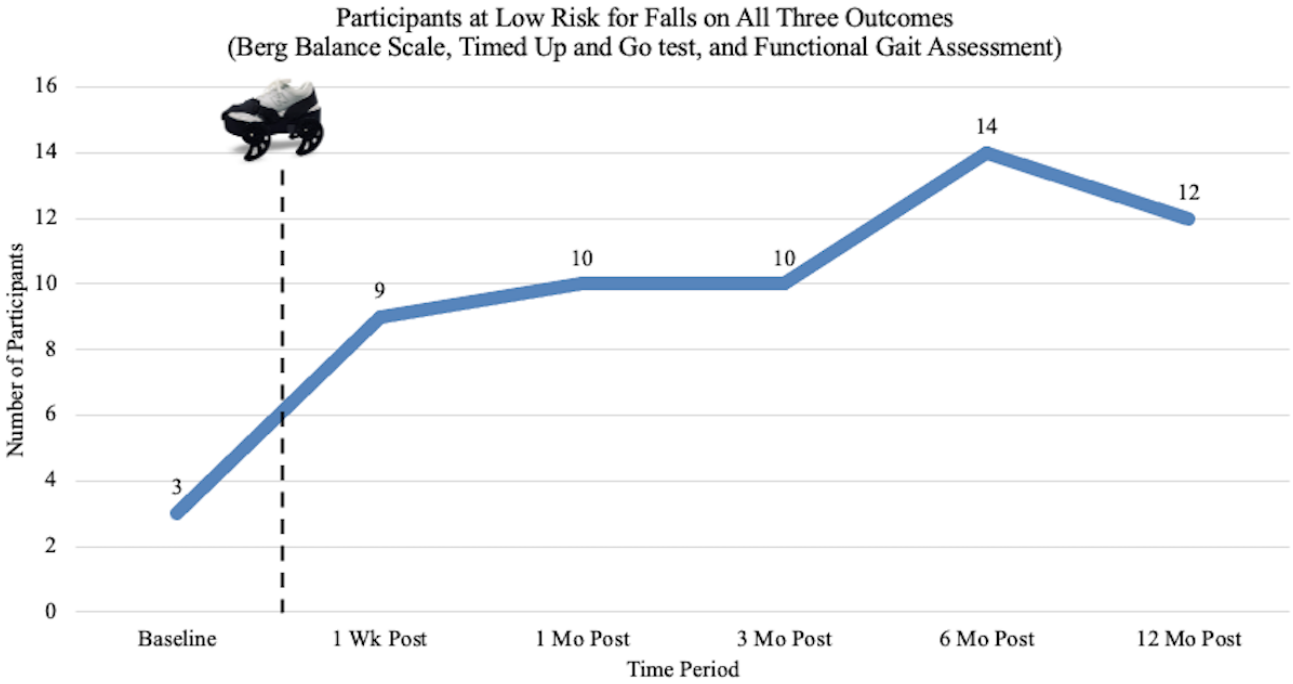

Questionnaire Results

Participants were questioned on the frequency of falls since their stroke and in the 12 months after being treated with the gait device. Questionnaire answer responses for the number of falls included the following ranges: (1) no falls, (2) 1 - 5 falls, (3) 5 - 10 falls, and (4)>10 falls. In addition, participants were asked if the gait device improved their balance. Questionnaire

XSL•FO
RenderX

results are provided in Figure 6. Note that since the number of falls was provided in a numerical range and the duration of time between fall reference durations is variable (ie, the duration of time since each participant's stroke and the duration of time since their treatment with the gait device), sparklines are used as a visual representation of the data trends. In addition, 1 participant did not complete the questionnaire (indicated by "n/a" in the table).

Figure 6. Questionnaire results. Sparkline thickness correlates with the number of sustained falls, with thicker lines indicating a higher number of falls. Months since stroke and age are provided for further context of the results. A downward-pointing arrow represents a trend toward fewer falls, an upward-pointing arrow represents a trend toward more falls, and a horizontal arrow represents no change in the fall trend.

ID	Months Since Stroke	Age (years)	Falls Since Stroke?		Falls Since Gait Treatment?		Falls Trend	Subjective Balance Improvement?
A	24	53	Yes			No	↘	Yes
B	13	55	No			No	→	Yes
C	14	44	n/a			n/a	n/a	n/a
D	50	51	Yes			Yes	↘	Yes
E	130	52	No			No	→	Yes
F	308	46	Yes			No	↘	Yes
G	22	61	Yes			Yes	↘	Yes
H	53	63	Yes			Yes	→	Not sure
I	92	58	Yes			Yes	↘	Yes
J	89	69	Yes			Yes	↘	Yes
K	28	47	Yes			Yes	↘	Yes
L	32	53	Yes			No	↘	Yes
M	80	50	Yes			No	↘	Yes
N	46	61	Yes			No	↘	Not sure
O	21	54	No			No	→	Yes
P	15	77	Yes			Yes	↗	Yes
Q	30	64	Yes			Yes	↘	Yes
R	28	62	Yes			Yes	↘	Yes

Discussion

Principal Findings

In this study, we measured participants' functional balance using the BBS, the TUG test, and the FGA before and several times after treatment with the gait device. Results reveal statistically significant improvements on each outcome at all 5 posttreatment time frames ($P<.05$) and a substantial reduction in the number of individuals categorized as high risk for falls by each of these outcome measures.

In addition to improved functional balance scores, the retention of balance improvement is a key finding of this study. Twelve months posttreatment, scores on each functional outcome remained significantly improved compared with baseline measurements. Moreover, using thresholds reported in the literature to classify fall risk, 12 out of 18 (67%) participants scored as low risk for falls on each of the three outcomes 12 months post treatment, an improvement from 3 out of 18 (17%) before treatment. A 2010 systematic review by Lubetzky-Vilnai and Kartin [48] analyzed the effect of balance training on balance performance for poststroke populations of varying chronicity. In comparison to the eight cited studies focused on the chronic stroke population with a comparable outcome measure [49-56], the majority of studies (5/8, 63%) report an immediate post treatment BBS improvement of lesser magnitude than the present study [49,51,52,55,56], although most of these studies demonstrated a more robust study design. Three studies reported greater improvement on the BBS compared with our study; however, these studies had longer or more intense treatment regimens [50,53,54]. In addition, none of these studies demonstrate retention of the reported improvements beyond 3 months after treatment.

The specific factors determining the retention of balance-related treatment effects post stroke lack substantial discussion in rehabilitation literature. However, research suggests that achieving improved performance of an affected limb or improved mobility levels after stroke may further enable improvement-sustaining activity participation [57-59]. In our study, a notable improvement in functional balance, in combination with other functional parameters targeted by the gait device (ie, asymmetry and gait speed [24,27,29]) could mitigate some of the barriers that limit poststroke mobility participation, subsequently sustaining the treatment effects. Future studies could clarify the exact mechanisms that translate to higher balance measure scores, and especially the retention of these effects, after treatment with the gait device.

While functional balance measures are routinely used in research and clinical practice to reflect risk for falls and fall risk reduction following treatment, it is important to correlate results on these measures with real-world fall rates. In this study, a posttreatment questionnaire regarding the participant's fall history (from the date of their stroke event until the start of the clinical trial and

XSL•FO
RenderX

in the 12 months following gait device treatment) was used to capture this information. Fourteen of 17 (82%) participants classified themselves as experiencing one or more falls before the start of the clinical trial. While this percentage appears high compared with rehabilitation literature [2], it is in line with studies reporting that those most likely to fall have moderate function (ie, impaired ability to maintain balance, yet still reasonably mobile) as found in our study population [60]. In the year following treatment with the gait device, the number of participants experiencing falls was reduced to 9 of 17 (53%), and some of the remaining 8 participants appeared to reduce their fall frequency. While the relative time duration of the pre- and posttreatment periods is not consistent for most participants (months since stroke is included in Figure 6 to provide context), a beneficial trend in real-world falls likely occurred. Importantly, the 5 oldest study participants (ages 62 - 77 years) remained fallers posttreatment, and 67% of those who fell after the treatment period were older than 60 years. These results highlight the well-known additional impact of aging on fall risk, including for those with stroke [61]. Moreover, the continued occurrence of falls despite improved balance function highlights the complexity surrounding falls management and the need for comprehensive assessments and treatments, in addition to gait and balance treatment, for individuals post stroke.

Limitations

Our study has several limitations. First, without a control group, we cannot exclude the possibility of confounding variables influencing our results. This limitation reduces our ability to draw definitive causal conclusions about the effectiveness of the intervention and limits generalizability to a broader poststroke populations or circumstances. Therefore, these results should be regarded as preliminary and in need of further confirmation with larger, controlled studies. A second limitation is attaining a falls history using a self-report questionnaire, which may be unreliable and subject to bias. In addition, the use of categorical measures for fall rates limits direct pre-post fall comparisons. These results could be strengthened by monitoring fall occurrence rates before, during, and after the treatment period. In addition, as the therapist assessors also oversaw the treatment, examiner bias is possible, and testing effects can occur with repeated outcome measurement.

Further, as noted in our inclusion criteria, the clinical trial participants did not receive additional physical therapy treatment from the clinical trial physical therapists or external physical therapists during the treatment period. However, given the extended duration of our study (ie, greater than 1 year), we did not exclude participants for obtaining additional therapy after the 1-week follow-up assessment was completed. Six of the 18 participants (B, D, H, M, Q, and R) did receive some form of therapy treatment between the 1-week and 12-month follow-up. Of these 6 participants, 3 had minimal treatment (6 or fewer therapy sessions), and of the remaining 3 participants, 2 received services focusing on upper extremity function (with some full-body therapy included). Importantly, two-thirds of our participants (12/18 participants) did not receive any additional physical therapy treatment throughout the study duration. Also, reviewing our results without the data from these 6 individuals yields largely unchanged findings (and even stronger trends in the percentages of participants exceeding MDC values). These results are shown in Multimedia Appendix 1.

Finally, in our results, we compare the participants' improvement posttreatment to each outcome measure's corresponding MDC value. While MDC provides a threshold to exceed the measurement error of each test, it does not provide context on the meaningfulness of the improvement (ie, minimal clinically important difference [MCID]), arguably a more important attribute post treatment. Unfortunately, MCID thresholds were not available for the selected outcomes in the population of chronic stroke. Several populations have been studied to obtain MCID values for these outcomes, including multiple sclerosis and the BBS [62] (MCID of 3 points compared with our used MDC value of 2.5 points), community-dwelling older adults and individuals with vestibular disorders for the FGA [63,64] (MCID of 4 points compared our used MDC value of 4.2 points or 14.1%), and patients undergoing lumbar surgery for the TUG test [65] (MCID of −3.4 s compared with our used MDC value of −3.2 s). While not directly generalizable, the similarities of these values to the MDC values for the chronic stroke population may suggest the meaningfulness of the improvement noted in this study.

Conclusions

The results of this study indicate that treatment with the gait device has the potential to result in long-term functional balance improvement for individuals with chronic stroke. Our findings reveal significantly improved scores on 3 validated balance measures, with improvements retained one year after treatment. In addition, score changes reflect a reduced risk of falls in the year following treatment with the gait device. Larger, controlled studies could confirm these results and verify the reduction in fall rates through pre- and posttreatment monitoring.

Acknowledgments

The authors would like to thank the study participants and physical therapists for their time and efforts associated with participation in this study. Funding for this study was provided by Moterum Technologies, Inc.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Authors' Contributions

LR, DH, and KBR designed the study. LR screened the participants, obtained informed consent, and oversaw the study protocol. BD, DH, KBR, and SJMB analyzed the data. BD wrote the manuscript with critical feedback from all authors.

Conflicts of Interest

Author KBR has a patent (US 9,295,302) related to this work. A management plan has been implemented and followed to reduce any effects of these conflicts of interest. The gait device is related to a commercial product by Moterum Technologies, Inc. Authors BD, LR, DH, and SJMB previously received a salary from Moterum Technologies. All authors have stock options in Moterum Technologies, Inc.

Multimedia Appendix 1
Improvement relative to the minimal detectable change with participants who received some additional therapy during posttreatment follow-ups removed from analysis (the plus symbol indicates improvement beyond MDC, and the dot symbol indicates a change less than MDC). With individuals who received some additional therapy during posttreatment follow-ups removed from analysis, the percentage of participants who improved greater than the MDC at each time period increased for 12 of the 15 measurements compared with Table 4.
[DOCX File, 21 KB - formative_v9i1e67297_app1.docx]

References

1. Sackley C, Brittle N, Patel S, et al. The prevalence of joint contractures, pressure sores, painful shoulder, other pain, falls, and depression in the year after a severely disabling stroke. Stroke 2008 Dec;39(12):3329-3334. [doi: 10.1161/STROKEAHA.108.518563] [Medline: 18787199]
2. Weerdesteyn V, de Niet M, van Duijnhoven HJR, Geurts ACH. Falls in individuals with stroke. J Rehabil Res Dev 2008;45(8):1195-1213. [Medline: 19235120]
3. Forster A, Young J. Incidence and consequences of falls due to stroke: a systematic inquiry. BMJ 1995 Jul 8;311(6997):83-86. [doi: 10.1136/bmj.311.6997.83] [Medline: 7613406]
4. Batchelor FA, Mackintosh SF, Said CM, Hill KD. Falls after stroke. Int J Stroke 2012 Aug;7(6):482-490. [doi: 10.1111/j.1747-4949.2012.00796.x] [Medline: 22494388]
5. Kapral MK, Fang J, Alibhai SMH, et al. Risk of fractures after stroke: results from the Ontario stroke registry. Neurology (ECronicon) 2017 Jan 3;88(1):57-64. [doi: 10.1212/WNL.0000000000003457] [Medline: 27881629]
6. Field MJ, Gebruers N, Shanmuga Sundaram T, Nicholson S, Mead G. Physical activity after stroke: a systematic review and meta-analysis. ISRN Stroke 2013 Nov 7;2013:1-13. [doi: 10.1155/2013/464176]
7. Jørgensen L, Engstad T, Jacobsen BK. Higher incidence of falls in long-term stroke survivors than in population controls. Stroke 2002 Feb;33(2):542-547. [doi: 10.1161/hs0202.102375]
8. Schmid AA, Van Puymbroeck M, Altenburger PA, et al. Balance and balance self-efficacy are associated with activity and participation after stroke: a cross-sectional study in people with chronic stroke. Arch Phys Med Rehabil 2012 Jun;93(6):1101-1107. [doi: 10.1016/j.apmr.2012.01.020] [Medline: 22502804]
9. Hong E. Health-related quality of life of community-dwelling stroke survivors: a comparison of fallers and non-fallers. J Phys Ther Sci 2015 Oct;27(10):3045-3047. [doi: 10.1589/jpts.27.3045] [Medline: 26644640]
10. Lamb SE, Ferrucci L, Volapto S, Fried LP, Guralnik JM. Risk factors for falling in home-dwelling older women with stroke. Stroke 2003 Feb;34(2):494-501. [doi: 10.1161/01.STR.0000053444.00582.B7]
11. Verheyden G, Weerdesteyn V, Pickering RM, et al. Interventions for preventing falls in people after stroke. Cochrane Database Syst Rev 2013 May 31;2013(5):CD008728. [doi: 10.1002/14651858.CD008728.pub2] [Medline: 23728680]
12. Batchelor F, Hill K, Mackintosh S, Said C. What works in falls prevention after stroke? Stroke 2010 Aug;41(8):1715-1722. [doi: 10.1161/STROKEAHA.109.570390]
13. Denissen S, Staring W, Kunkel D, et al. Interventions for preventing falls in people after stroke. Cochrane Database Syst Rev 2019 Oct 1;10(10):CD008728. [doi: 10.1002/14651858.CD008728.pub3] [Medline: 31573069]
14. Arienti C, Lazzarini SG, Pollock A, Negrini S. Rehabilitation interventions for improving balance following stroke: an overview of systematic reviews. PLoS ONE 2019;14(7):e0219781. [doi: 10.1371/journal.pone.0219781] [Medline: 31323068]
15. Lackland DT, Roccella EJ, Deutsch AF, et al. Factors influencing the decline in stroke mortality: a statement from the American Heart Association/American Stroke Association. Stroke 2014 Jan;45(1):315-353. [doi: 10.1161/01.str.0000437068.30550.cf] [Medline: 24309587]
16. Reed KB, Handzic I, University of South Florida. Gait altering shoes. 2016 Mar 29 inventorsassignee.
17. Patterson KK, Parafianowicz I, Danells CJ, et al. Gait asymmetry in community-ambulating stroke survivors. Arch Phys Med Rehabil 2008 Feb;89(2):304-310. [doi: 10.1016/j.apmr.2007.08.142] [Medline: 18226655]
18. An CM, Son YL, Park YH, Moon SJ. Relationship between dynamic balance and spatiotemporal gait symmetry in hemiplegic patients with chronic stroke. Hong Kong Physiother J 2017 Dec;37:19-24. [doi: 10.1016/j.hkpj.2017.01.002] [Medline: 30931042]

XSL•FO
RenderX

19. Handzˇić I, Reed KB. Kinetic shapes: analysis, verification, and applications. J Mech Des N Y 2014 Jun;136(6):0610051-0610058. [doi: 10.1115/1.4027168] [Medline: 25053871]
20. Olney SJ, Richards C. Hemiparetic gait following stroke. Part I: characteristics. Gait Posture 1996 Apr;4(2):136-148. [doi: 10.1016/0966-6362(96)01063-6]
21. Dos Anjos S, Morris D, Taub E. Constraint-induced movement therapy for lower extremity function: describing the LE-CIMT protocol. Phys Ther 2020 Apr 17;100(4):698-707. [doi: 10.1093/ptj/pzz191] [Medline: 31899495]
22. Reisman DS, McLean H, Keller J, Danks KA, Bastian AJ. Repeated split-belt treadmill training improves poststroke step length asymmetry. Neurorehabil Neural Repair 2013 Jun;27(5):460-468. [doi: 10.1177/1545968312474118] [Medline: 23392918]
23. Handzic I, Barno EM, Vasudevan EV, Reed KB. Design and pilot study of a gait enhancing mobile shoe. Paladyn 2011 Dec 1;2(4):24371521. [doi: 10.2478/s13230-012-0010-7] [Medline: 24371521]
24. Kim SH, Huizenga DE, Handzic I, et al. Relearning functional and symmetric walking after stroke using a wearable device: a feasibility study. J Neuroeng Rehabil 2019 Aug 28;16(1):106. [doi: 10.1186/s12984-019-0569-x] [Medline: 31455358]
25. Handzic I, Vasudevan E, Reed KB. Developing a gait enhancing mobile shoe to alter over-ground walking coordination. IEEE Int Conf Robot Autom 2012 May;2012:4124-4129. [doi: 10.1109/ICRA.2012.6225346] [Medline: 23484067]
26. Kim SH HI H, Edgeworth R, Lazinski M, Ramakrishnan T, Reed KB. Inital results of the gait enhancing mobile shoe on individuals with stroke. Presented at: 38th Annual International Conference of the IEEE Engineering in Biology and Medicine Society (EMBC); Aug 16-20, 2016; Orlando, FL.
27. Huizenga D, Rashford L, Darcy B, et al. Wearable gait device for stroke gait rehabilitation at home. Top Stroke Rehabil 2021 Sep;28(6):443-455. [doi: 10.1080/10749357.2020.1834272] [Medline: 33261520]
28. Darcy B, Rashford L, Shultz ST, et al. Gait device treatment using telehealth for individuals with stroke during the COVID-19 pandemic: nonrandomized pilot feasibility study. JMIR Form Res 2023 May 19;7:e43008. [doi: 10.2196/43008] [Medline: 37204830]
29. Darcy B, Rashford L, Tsai NT, Huizenga D, Reed KB, Bamberg SJM. One-year retention of gait speed improvement in stroke survivors after treatment with a wearable home-use gait device. Front Neurol 2023;14:1089083. [doi: 10.3389/fneur.2023.1089083] [Medline: 38274885]
30. Mancini M, Horak FB. The relevance of clinical balance assessment tools to differentiate balance deficits. Eur J Phys Rehabil Med 2010 Jun;46(2):239-248. [Medline: 20485226]
31. Bower K, Thilarajah S, Pua YH, et al. Dynamic balance and instrumented gait variables are independent predictors of falls following stroke. J Neuroeng Rehabil 2019 Jan 7;16(1):3. [doi: 10.1186/s12984-018-0478-4] [Medline: 30612584]
32. Sibley KM, Straus SE, Inness EL, Salbach NM, Jaglal SB. Balance assessment practices and use of standardized balance measures among Ontario physical therapists. Phys Ther 2011 Nov;91(11):1583-1591. [doi: 10.2522/ptj.20110063] [Medline: 21868613]
33. Bushnell C, Bettger JP, Cockroft KM, et al. Chronic Stroke Outcome Measures for Motor Function Intervention Trials: Expert Panel Recommendations. Circ Cardiovasc Qual Outcomes 2015 Oct;8(6 Suppl 3):S163-S169. [doi: 10.1161/CIRCOUTCOMES.115.002098] [Medline: 26515205]
34. Alghadir AH, Al-Eisa ES, Anwer S, Sarkar B. Reliability, validity, and responsiveness of three scales for measuring balance in patients with chronic stroke. BMC Neurol 2018 Sep 13;18(1):141. [doi: 10.1186/s12883-018-1146-9] [Medline: 30213258]
35. Moore JL, Potter K, Blankshain K, Kaplan SL, O Dwyer LC, Sullivan JE. A core set of outcome measures for adults with neurologic conditions undergoing rehabilitation: A CLINICAL PRACTICE GUIDELINE. J Neurol Phys Ther 2018 Jul;42(3):174-220. [doi: 10.1097/NPT.0000000000000229] [Medline: 29901487]
36. Berg KO, Wood-Dauphinee SL, Williams JI, Maki B. Measuring balance in the elderly: validation of an instrument. Can J Public Health 1992;83 Suppl 2:S7-11. [Medline: 1468055]
37. Simpson LA, Miller WC, Eng JJ. Effect of stroke on fall rate, location and predictors: a prospective comparison of older adults with and without stroke. PLoS One 2011 Apr 29;6(4):e19431. [doi: 10.1371/journal.pone.0019431] [Medline: 21559367]
38. Liston RA, Brouwer BJ. Reliability and validity of measures obtained from stroke patients using the Balance Master. Arch Phys Med Rehabil 1996 May;77(5):425-430. [doi: 10.1016/s0003-9993(96)90028-3] [Medline: 8629916]
39. Flansbjer UB, Blom J, Brogårdh C. The reproducibility of Berg Balance Scale and the Single-leg Stance in chronic stroke and the relationship between the two tests. PM R 2012 Mar;4(3):165-170. [doi: 10.1016/j.pmrj.2011.11.004] [Medline: 22306324]
40. Hiengkaew V, Jitaree K, Chaiyawat P. Minimal detectable changes of the Berg Balance Scale, Fugl-Meyer Assessment Scale, Timed “Up & Go” Test, gait speeds, and 2-minute walk test in individuals with chronic stroke with different degrees of ankle plantarflexor tone. Arch Phys Med Rehabil 2012 Jul;93(7):1201-1208. [doi: 10.1016/j.apmr.2012.01.014] [Medline: 22502805]
41. Podsiadlo D, Richardson S. The timed “Up & Go”: a test of basic functional mobility for frail elderly persons. J Am Geriatr Soc 1991 Feb;39(2):142-148. [doi: 10.1111/j.1532-5415.1991.tb01616.x] [Medline: 1991946]

XSL•FO
RenderX

42. Ng SS, Hui-Chan CW. The timed up & go test: its reliability and association with lower-limb impairments and locomotor capacities in people with chronic stroke. Arch Phys Med Rehabil 2005 Aug;86(8):1641-1647. [doi: 10.1016/j.apmr.2005.01.011] [Medline: 16084820]
43. Andersson AG, Kamwendo K, Seiger A, Appelros P. How to identify potential fallers in a stroke unit: validity indexes of 4 test methods. J Rehabil Med 2006 May;38(3):186-191. [doi: 10.1080/16501970500478023] [Medline: 16702086]
44. Wrisley DM, Marchetti GF, Kuharsky DK, Whitney SL. Reliability, internal consistency, and validity of data obtained with the functional gait assessment. Phys Ther 2004 Oct;84(10):906-918. [Medline: 15449976]
45. Leddy AL, Crowner BE, Earhart GM. Functional gait assessment and balance evaluation system test: reliability, validity, sensitivity, and specificity for identifying individuals with Parkinson disease who fall. Phys Ther 2011 Jan;91(1):102-113. [doi: 10.2522/ptj.20100113] [Medline: 21071506]
46. Lin JH, Hsu MJ, Hsu HW, Wu HC, Hsieh CL. Psychometric comparisons of 3 functional ambulation measures for patients with stroke. Stroke 2010 Sep;41(9):2021-2025. [doi: 10.1161/STROKEAHA.110.589739] [Medline: 20671244]
47. Haley SM, Fragala-Pinkham MA. Interpreting change scores of tests and measures used in physical therapy. Phys Ther 2006 May;86(5):735-743. [doi: 10.1093/ptj/86.5.735] [Medline: 16649896]
48. Lubetzky-Vilnai A, Kartin D. The effect of balance training on balance performance in individuals poststroke: a systematic review. J Neurol Phys Ther 2010 Sep;34(3):127-137. [doi: 10.1097/NPT.0b013e3181ef764d] [Medline: 20716987]
49. Olawale OA, Ogunmakin OS. The effect of exercise training on balance in adult patients with post-stroke hemiplegia. Int J Ther Rehabil 2006 Jul;13(7):318-322. [doi: 10.12968/ijtr.2006.13.7.21408]
50. Fritz SL, Pittman AL, Robinson AC, Orton SC, Rivers ED. An intense intervention for improving gait, balance, and mobility for individuals with chronic stroke: a pilot study. J Neurol Phys Ther 2007 Jun;31(2):71-76. [doi: 10.1097/NPT.0b013e3180674a3c] [Medline: 17558360]
51. Yen CL, Wang RY, Liao KK, Huang CC, Yang YR. Gait training induced change in corticomotor excitability in patients with chronic stroke. Neurorehabil Neural Repair 2008;22(1):22-30. [doi: 10.1177/1545968307301875] [Medline: 17507641]
52. Leroux A, Pinet H, Nadeau S. Task-oriented intervention in chronic stroke: changes in clinical and laboratory measures of balance and mobility. Am J Phys Med Rehabil 2006 Oct;85(10):820-830. [doi: 10.1097/01.phm.0000233179.64769.8c] [Medline: 16998429]
53. Macko RF, Benvenuti F, Stanhope S, et al. Adaptive physical activity improves mobility function and quality of life in chronic hemiparesis. J Rehabil Res Dev 2008;45(2):323-328. [doi: 10.1682/jrrd.2007.02.0025] [Medline: 18566949]
54. Michael K, Goldberg AP, Treuth MS, Beans J, Normandt P, Macko RF. Progressive adaptive physical activity in stroke improves balance, gait, and fitness: preliminary results. Top Stroke Rehabil 2009;16(2):133-139. [doi: 10.1310/tsr1602-133] [Medline: 19581199]
55. Stuart M, Benvenuti F, Macko R, et al. Community-based adaptive physical activity program for chronic stroke: feasibility, safety, and efficacy of the Empoli model. Neurorehabil Neural Repair 2009 Sep;23(7):726-734. [doi: 10.1177/1545968309332734] [Medline: 19318465]
56. Huijbregts MPJ, McEwen S, Taylor D. Exploring the feasibility and efficacy of a telehealth stroke self-management programme: a pilot study. Physiother Can 2009;61(4):210-220. [doi: 10.3138/physio.61.4.210] [Medline: 20808482]
57. Wolf SL, Winstein CJ, Miller JP, et al. Retention of upper limb function in stroke survivors who have received constraint-induced movement therapy: the EXCITE randomised trial. Lancet Neurol 2008 Jan;7(1):33-40. [doi: 10.1016/S1474-4422(07)70294-6] [Medline: 18077218]
58. Wang C, Winstein C, D'Argenio DZ, Schweighofer N. The efficiency, efficacy, and retention of task practice in chronic stroke. Neurorehabil Neural Repair 2020 Oct;34(10):881-890. [doi: 10.1177/1545968320948609] [Medline: 32830617]
59. Madhavan S, Lim H, Sivaramakrishnan A, Iyer P. Effects of high intensity speed-based treadmill training on ambulatory function in people with chronic stroke: A preliminary study with long-term follow-up. Sci Rep 2019 Feb 13;9(1):1985. [doi: 10.1038/s41598-018-37982-w] [Medline: 30760772]
60. Wei WE, De Silva DA, Chang HM, et al. Post-stroke patients with moderate function have the greatest risk of falls: a national cohort study. BMC Geriatr 2019 Dec 26;19(1):373. [doi: 10.1186/s12877-019-1377-7] [Medline: 31878876]
61. Ugur C, Gücüyener D, Uzuner N, Ozkan S, Ozdemir G. Characteristics of falling in patients with stroke. J Neurol Neurosurg Psychiatry 2000 Nov;69(5):649-651. [doi: 10.1136/jnnp.69.5.649] [Medline: 11032620]
62. Gervasoni E, Jonsdottir J, Montesano A, Cattaneo D. Minimal clinically important difference of Berg Balance Scale in people with multiple sclerosis. Arch Phys Med Rehabil 2017 Feb;98(2):337-340. [doi: 10.1016/j.apmr.2016.09.128] [Medline: 27789239]
63. Beninato M, Fernandes A, Plummer LS. Minimal clinically important difference of the functional gait assessment in older adults. Phys Ther 2014 Nov;94(11):1594-1603. [doi: 10.2522/ptj.20130596] [Medline: 24947198]
64. Wellons RD, Duhe SE, MacDowell SG, Hodge A, Oxborough S, Levitzky EE. Estimating the minimal clinically important difference for balance and gait outcome measures in individuals with vestibular disorders. J Vestib Res 2022;32(3):223-233. [doi: 10.3233/VES-201630] [Medline: 35147571]
65. Gautschi OP, Stienen MN, Corniola MV, et al. Assessment of the minimum clinically important difference in the Timed Up and Go test after surgery for lumbar degenerative disc disease. Neurosurgery 2017 Mar 1;80(3):380-385. [doi: 10.1227/NEU.0000000000001320] [Medline: 27352275]

XSL•FO
RenderX

Abbreviations

BBS: Berg Balance Scale
FGA: Functional Gait Assessment
HIPAA: Health Insurance Portability and Accountability Act
IRB: institutional review board
MCID: minimal clinically important difference
MDC: minimal detectable change
TUG: Timed Up and Go
WCG: Western-Copernicus Group

Edited by A Mavragani; submitted 08.10.24; peer-reviewed by L Chan, S Sas, VS Mercer; revised version received 05.05.25; accepted 08.05.25; published 15.08.25.

Please cite as:
Darcy B, Rashford L, Huizenga D, Reed KB, Bamberg SJM
Balance Improvement and Fall Risk Reduction in Stroke Survivors After Treatment With a Wearable Home-Use Gait Device: Single-Arm Longitudinal Study With 1-Year Follow-Up
JMIR Form Res 2025;9:e67297
URL: https://formative.jmir.org/2025/1/e67297
doi:10.2196/67297

© Brianne Darcy, Lauren Rashford, David Huizenga, Kyle B Reed, Stacy J M Bamberg. Originally published in JMIR Formative Research (https://formative.jmir.org), 15.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Feasibility and Efficacy of Commercial-Off-the-Shelf Virtual Reality Applications for Managing Chronic Pain and Enhancing Well-Being Among Older Adults in the Community: Mixed Methods Pilot Study

Rachel Yim Fong Leung, MSc; Megan Zichen Ye, MSc; Flora Yaqian Zhang, MSc; Tyrone Tai-On Kwok, PhD; Yuying Sun, MSc, PhD; Agnes Yuen Kwan Lai, DN, PhD; Juming Jiang, PhD; Mimi Mun Yee Tse, RN, PhD

School of Nursing and Health Sciences, Hong Kong Metropolitan University, 11th Floor, Jockey Club Institute of Healthcare, 1 Sheung Shing Street, Homantin, Kowloon, Hong Kong, China

Corresponding Author:
Tyrone Tai-On Kwok, PhD
School of Nursing and Health Sciences, Hong Kong Metropolitan University, 11th Floor, Jockey Club Institute of Healthcare, 1 Sheung Shing Street, Homantin, Kowloon, Hong Kong, China

Abstract

Background: Older adults may experience chronic pain as they age, which can affect their physical and psychological well-being. Virtual reality (VR) is emerging as a novel and nonpharmacological intervention that offers pain relief and mood enhancement through immersive experiences. However, the feasibility and effectiveness of using nonlocalized and commercial VR applications for chronic pain relief and mood enhancement among community-dwelling older adults remain underexplored.

Objective: The main objectives of this study were to (1) evaluate the feasibility of using commercial-off-the-shelf VR applications for managing chronic pain among older adults, (2) assess the efficacy of VR in alleviating chronic pain, and (3) examine its impact on the well-being of older adults in a community setting.

Methods: The study was a single-arm mixed methods pilot study. It was divided into two stages, including preparation and implementation. A total of 13 older adults (8 with chronic pain and 5 without) were recruited to participate in a 3-week VR intervention. Participants engaged in VR sessions that followed a step-by-step adaptation process. Each session included 360-degree relaxation videos and VR-based boxing exercises via Les Mills BODYCOMBAT, lasting 15 minutes. Pain intensity, pain self-efficacy, well-being, and mood were measured pre- and post-intervention using the Numeric Pain Rating Scale (NPRS), a pain self-efficacy question, the World Health Organization–Five Well-Being Index (WHO-5), and the Mood Assessment Scale (Mood). VR feasibility was evaluated based on completion rates, adverse outcomes, and qualitative feedback from semistructured interviews.

Results: Of the 13 participants recruited, 11 completed the intervention (84.6% completion rate). The mean age was 79.2 (SD 9.2) years. The study found a statistically significant 16.32% improvement in the mean mood score, with a mean increase of 2.64 (SD 1.45) points and a large effect size ($P<.001$; Cohen $d=1.82$). The median pain self-efficacy score decreased from 3.0 (IQR 1.5-3.0) to 1.0 (IQR 1.0-2.0) ($z=-2.236$; $P=.03$). However, no significant changes were observed in pain intensity or overall well-being. The study demonstrated the high feasibility of commercial-off-the-shelf VR technology for older adults. Minor adverse effects were reported, including back pain and headset discomfort. In addition, 90.9% of participants enjoyed the VR experience, and all were willing to join future sessions.

Conclusions: The pilot study demonstrated that commercial-off-the-shelf VR applications can effectively enhance mood and pain self-efficacy. Additional procedures, such as prebriefing, real-time interpretation, and a gradual adaptation process, were essential to overcoming barriers such as language, cultural nuances, and the digital literacy of older adults. Despite the lack of significant changes in pain intensity or overall well-being, the psychological benefits suggest that VR could be a valuable adjunct tool in chronic pain management. Future research should focus on larger sample sizes, longer intervention durations, randomized controlled trials, and the development of localized commercial VR applications to further explore their efficacy.

(JMIR Form Res 2025;9:e67765) doi:10.2196/67765

KEYWORDS

virtual reality; VR; commercial VR; off-the-shelf; non-localized; pain; older adults; well-being; NPRS; pain self-efficacy; WHO-5; mood; feasibility; pilot study; World Health Organization-Five Well-Being Index

XSL•FO
RenderX

Introduction

Background

Since Ivan Sutherland and Bob Sproull created the first virtual reality (VR) headset in 1968, VR technology has evolved significantly [1]. VR involves experiencing digital simulations, ranging from realistic to fantastical, in 3D spaces using computer technology [2]. Users can experience VR in immersive and nonimmersive forms. Today, VR technology is widely used in gaming, entertainment, education, and therapy.

In the medical field, VR has gained recognition as a nonpharmacological tool for psychosocial and physical therapy. Research has consistently shown VR's capacity to enhance mental health and cognitive abilities while diminishing symptoms of depression [3-6]. VR can assist in physical recovery and facilitate postsurgical rehabilitation in various clinical settings [7-9]. With the aging population and the cost of chronic pain potentially reaching up to US $635 billion annually [10], VR has also been developed to relieve persistent pain. Studies have validated its efficacy in pain reduction using techniques such as distraction, relaxation, graded exposure therapy, and pain management education [11-14]. Various VR applications have been designed and built specifically for chronic pain therapy. These include, but are not limited to, applications for lower back pain [15], neck pain [16,17], migraine headaches [18], knee osteoarthritis [19], and regional pain [20-22]. These VR applications are often used under the guidance of health care professionals such as physiotherapists or occupational therapists [16-18,21,22].

On the other hand, as chronic pain can affect various body parts, different studies have indicated the potential benefits of commercial-off-the-shelf VR applications for individuals with persistent pain [23-26]. One study explored the efficacy of using 12 commercial VR applications to provide veterans with chronic pain therapy, ranging from low-intensity distraction to high-intensity movement-based exposure. Benefits included reducing kinesiophobia, improving patient-specific functioning, and delivering pain distraction [23]. Another study aimed to desensitize a patient to visually evoked pain by using the Vacation Simulator (Owlchemy Labs), a commercial VR application. The results showed that the intervention could nearly extinguish the visually evoked pain by the end of the program [25]. Commercial VR applications may benefit individuals with chronic or multiple pain conditions, particularly those needing regular physical exercise [26-28]. Immersive commercial VR games can motivate people to work out to improve their health.

Locally, since 2005, the Hong Kong Hospital Authority has implemented VR for functional rehabilitation in walking, upper and lower limb movement, fall prevention, cognitive ability, and driving reaction, in which therapists supervise nonimmersive VR training to ensure patient safety [7]. However, research on VR's efficacy for chronic pain remains limited. Two systematic reviews found VR to be effective in reducing chronic pain levels [29,30], but one study highlighted that its impact might depend on the methodology and strategies used in the research [29]. Moreover, 5 clinical trials in Hong Kong explored VR's use in pain alleviation, but most trials primarily focused on acute pain relief in clinical settings such as peripheral intravenous cannulation [31], venipuncture [32], palliative care [33], and elective surgery [34]. Only one recent clinical trial demonstrated chronic pain reduction through VRiKnee [35], a designated VR application for knee osteoarthritis.

The challenges outlined above highlight the need to explore the use of commercial VR applications for older adults in the community in a local setting, without relying on high-cost, high-turnaround-time, tailor-made VR applications. However, the number of localized and commercial VR applications is limited. While international studies have explored commercial VR applications to manage chronic pain or well-being, many required participants to understand English. This often excluded older adults with limited English proficiency, as these applications were primarily in English. In addition, concerns regarding cultural nuances and perceived limited technological literacy were factors for exclusion in many studies [36-38]. Researchers often either specifically designed VR applications or used native-language and commercial-off-the-shelf VR applications for older adults in the community. There has been little research exploring the feasibility and effectiveness of applying nonlocalized, commercial VR applications to manage chronic pain and well-being in the population.

Research Objectives

The main objectives of this study were to (1) evaluate the feasibility of using commercial-off-the-shelf VR applications for managing chronic pain among older adults, (2) assess the efficacy of VR in alleviating chronic pain, and (3) examine its impact on the well-being of older adults in a community setting.

Methods

Study Design

This study was a single-arm, mixed methods pilot study conducted between January and March 2024 at an elderly day care center in Hong Kong. The study was approved by the Ethical Committee of the Hong Kong Metropolitan University (reference no HE-RD/2023/1.15). Participants signed an informed consent form and completed a questionnaire that included demographic information, the Numeric Pain Rating Scale (NPRS), pain self-efficacy, medical history queries, and their VR experience for initial screening. The study comprised two major stages: preparation and implementation. This paper is reported under the CONSORT (Consolidated Standards of Reporting Trials) 2010 extension for pilot and feasibility studies (items 1a-12b, as applicable to a nonrandomized, single-arm design).

This exploratory feasibility pilot followed Julious' [39] rule-of-thumb of 12 participants for pilot work. Accounting for possible attrition, we aimed for 14 enrollments and achieved 13 complete datasets. Using the large pre-post mood effect observed here (Cohen d=1.82), a conventional two-arm superiority trial would need roughly 34 participants in total (17 per arm) to provide 80 % power at α=.05, calculated with G*Power 3.1 (Heinrich Heine University Düsseldorf).

XSL•FO
RenderX

Preparation Stage

This stage consisted of four key steps before the formal study: theory setting, game selection, recruitment, and a pre-pilot study.

Theories Setting

The study was grounded in 3 theoretical frameworks. The biopsychosocial model suggests that health and disease are influenced by biological, psychological, and social factors [40-42]. The commercial VR applications selected should positively impact all 3 dimensions. The second theory is the gate control theory of pain. It suggests that pain perception is both a direct result of injury and brain modulation. VR can provide relaxation and distraction, closing the gate to pain signals and reducing the pain experience [43,44]. The last framework is self-determination theory, which emphasizes the importance of autonomy, competence, and relatedness in motivating behavior [45]. The VR games selected should motivate older adults and enhance their well-being.

Game Selection

Two commercial VR applications were selected for the intervention: YouTube VR (Odders Lab) [46] and Les Mills BODYCOMBAT (Phillip Mills) [47]. The applications fulfilled specific criteria: supporting both standing and seated positions, offering adjustable difficulty levels, accommodating varying fitness levels, and featuring highly visual, intuitive content that minimized reliance on text or voice instructions. Many studies have used YouTube VR as a medium to manage patients with chronic pain in conditions such as cancer [31] and haemophilic ankle arthropathy [48]. These studies demonstrated the capabilities of immersive 360-degree videos to alleviate pain and improve well-being [33,48,49]. Videos were chosen based on 4K or higher resolution and ease of visualization, accommodating the visual requirements of older adults. The chosen videos included diverse themes (refer to Multimedia Appendix 1). As for VR boxing exercises from Les Mills BODYCOMBAT, it incorporates elements from martial arts to promote physical activity.

Recruitment

Participants were recruited from an elderly day care center in Hong Kong. The center provides day care services, rehabilitation exercises, and social activities for older adults aged 60 years or older, including those with cognitive impairment, who were assessed by the Standardized Care Need Assessment Mechanism for Elderly Services (SCNAMES). We included both pain and nonpain older adults to assess VR's effects on general well-being, not just pain management. This also allowed us to compare the effectiveness and feasibility of VR between the two groups, making the findings more broadly applicable. The inclusion criteria were (1) aged 60 years or older, (2) residing in their own homes and not in a long-term care facility (such as a nursing home) within the last year, (3) willingness to participate in the VR intervention, (4) ability to understand Chinese and speak Cantonese, and (5) self-reported ability to see clearly for daily tasks (with habitual glasses if worn) and absence of diplopia. Participants were excluded if they had (1) an inability to communicate, (2) moderate to severe dementia in stage 4 or above under the Global Deterioration Scale (GDS), and (3) visual impairment or visual surgery within the past 1 year.

Prepilot Study

A prepilot study was conducted to evaluate the environment, process flow, and initial acceptance of VR by older adults. The prepilot also evaluated the eligibility of participants with visual impairments. Five older adults participated, revealing issues such as headset discomfort and latency in high-quality video playback. These issues were addressed by using a portable 5G router and refining written instructions to simplify VR usage. The feedback gathered helped refine the formal pilot study.

Implementation Stage

Intervention

Eligible participants engaged in VR sessions once weekly for 3 consecutive weeks. Each session comprised two parts—a 10-minute YouTube VR immersive experience to introduce the VR environment and a 5-minute VR boxing exercise using Les Mills BODYCOMBAT for higher-intensity interaction. For safety, participants were seated during all sessions. The study used the Meta Quest 3, an immersive VR device with a head-mounted display and wireless handheld controllers. To ensure smooth adaptation to the digital environment, participants followed a sequential process of learning, familiarization, and interaction during each VR session, which was as follows:

Learning

Before each VR session, we conducted a 10-minute briefing with written instructions on how to use the controllers and what to expect during the session (Figures 1 and 2).

Mood Assessment Scale (Mood) scores and potential discomfort were also assessed before starting. One-on-one assistance was provided throughout the intervention to help participants smoothly engage with the VR experience.

XSL•FO
RenderX

Figure 1. Guiding the participant to use the controllers.

XSL•FO
RenderX

Figure 2. Individual briefing of the scenes to be seen.

Familiarization

Sessions began with YouTube VR to help participants acclimate to the VR environment. Real-time translation was provided for the English-language content. Videos typically lasted 3-5 minutes to maximize the participants' VR experience.

Interaction

After becoming familiar with the VR environment, the participants engaged in high movement-based activities in Les Mills BODYCOMBAT (Figures 3 and 4). The VR classes, accessible in both standing and seated modes, included physical movements and cognitive challenges such as target hitting based on color and directional cues. Participants could begin with simpler, low-level classes and progress to more advanced levels involving punches and lateral movements for obstacle avoidance, provided they completed the lower levels successfully.

XSL•FO
RenderX

Figure 3. Participant engaging in Les Mills BODYCOMBAT, a 5-minute virtual reality boxing exercise.

Figure 4. Scene from the Les Mills BODYCOMBAT game session.

Data Collection

We gathered two categories of data: demographic and outcome data. Demographic data were collected at baseline when the potential participants had signed the consent form expressing their interest in joining the VR intervention.

The outcome data consisted of several measures: the World Health Organization–Five Well-Being Index (WHO-5), the NPRS, and assessments of pain self-efficacy, the Mood, and the feasibility of the VR intervention. In addition, we collected qualitative feedback from the participants after the last VR session.

To have a general picture of the effectiveness of VR in managing chronic pain and the well-being of the participants, the three outcome variables—NPRS, Pain Self-Efficacy, and WHO-5—were assessed before the first VR session (BT1) and after the end of the last VR session (AT3). To differentiate any potential change in mood before and after each VR intervention, the outcome variable “mood” was collected every time before

XSL•FO
RenderX

(BT1, BT2, and BT3) and after each VR session (AT1, AT2, and AT3). In addition, the participants were checked to see if they had any discomfort immediately after every VR session. If the participants reported discomfort, they were assessed for VR sickness to understand the specific areas of discomfort.

The NPRS

The NPRS is a widely used tool in clinical and research settings to measure pain intensity. The NPRS uses a 0-10 scale, where 0 indicates "no pain" and 10 denotes "the most severe pain imaginable" [50]. The NPRS is generally considered reliable and has high reproducibility for measuring pain intensity across different patient populations and settings. Research has shown that the NPRS provides consistent results when used in repeated measures and correlates well with other pain scales such as the Visual Analog Scale (VAS) [51,52].

Pain Self-Efficacy

Pain self-efficacy was a belief held by people with chronic pain that certain activities can be carried out despite the pain [53]. In this pilot study, to minimize the cognitive load for older adults, we used a single question to examine the extent to which the pain affected their daily lives. Participants rated their responses to the item using a 5-point Likert scale, ranging from 0 ("none at all") to 5 ("extremely").

The WHO-5

The WHO-5 is a widely recognized and validated scale for use across many ethnicities, including the Hong Kong population [54-56]. There are 5 short and direct questions. All items are scored from 0 ("none of the time") to 5 ("all the time"). Scores closer to the maximum reflect greater well-being. The raw score is obtained by summing the single response values (0 - 5).

Mood Assessment Scale

There are four questions to capture the mood states of participants: (1) "I feel cheerful and in good spirits," (2) "I feel relaxed," (3) "I feel irritable," and (4) "I feel scared." The questions were inspired by the WHO-5 [54] and the Positive and Negative Affect Schedule (PANAS) [57]. Participants rated their current feelings using a 5-point Likert scale, with responses ranging from 1 ("not at all") to 5 ("extremely").

To calculate mood scores, the responses to the first two positive affect questions—feeling cheerful and relaxed—were scored from 1 to 5 points, directly corresponding to the intensity of the feeling. Conversely, responses to the negative affect questions—feeling irritable and scared—were inversely scored, with 5 points for "not at all" and 1 point for "extremely." The scores for each session were summed and divided by the number of VR intervention sessions (3 in total) to calculate the average preintervention and postintervention mood scores. The methodology for calculating this composite mood score was tailored to accommodate the specific conditions and constraints of our study, including session count and participant demographics. It was also intended to minimize the possibility that outliers or particular incidents would disproportionately skew the results. Internal consistency of the 4-item scale in this sample was acceptable (Cronbach α=0.84, computed from 11 participants × 3 presession assessments, n=33).

Feasibility

The feasibility of the VR intervention was assessed through three primary measures: completion rate, adverse outcomes, and sentiment analysis of the feedback from the semistructured questionnaires. The completion rate was defined as the proportion of participants who attended all 3 VR intervention sessions compared to the number of participants present at the start of the first session.

Adverse outcomes refer to the number and percentage of instances recorded during the entire VR session based on VR sickness items. The VR sickness items were specially tailored to the context of our study, drawing from standard scales but modified to enhance clarity and relevance for our participants. This decision was informed by preliminary findings from a prepilot phase, where traditional VR sickness descriptors such as "fullness of the head" in the Virtual Reality Sickness Questionnaire (VRSQ) [58] were not well understood by participants. Instead, we included nine specific aspects that better matched our participants' experiences and feedback: general discomfort, fatigue, headache, eye strain, difficulty focusing, nausea, blurred vision, dizziness, and vertigo. The severity of each symptom was rated using a 5-point Likert scale, with responses ranging from 1 ("not at all") to 5 ("extremely"). Higher scores indicated greater symptom severity.

Qualitative Data Analysis

The final measure was based on the sentiment analysis of the feedback from the semistructured questionnaire. The number and percentage of positive and negative responses were categorized into 4 indicators for analysis. Participants' feedback was analyzed using inductive thematic content analysis via Microsoft Word. No formal audit trail was kept. Reflective team discussions helped ensure coding consistency. Due to the small sample size and pilot nature, data saturation was not formally evaluated. Trustworthiness was addressed by clarifying responses during interviews and transparently describing methods, following Braun and Clarke's [59] guidance on credibility, dependability, confirmability, and transferability.

Semistructured Interviews

Following the conclusive VR session, participants were invited to answer eight open-ended questions: (1) Do you enjoy the VR experience? (2) Does operating the VR equipment affect your pain level? If yes, in what way? (3) Do you find the VR system easy or difficult to operate? Please explain. (4) Are there any aspects of playing VR that you like the most? (5) Are there any aspects of playing VR that you dislike? (6) Do you prefer VR to other daily activities such as games, sports, or exercise? Please explain. (7) Would you be willing to participate in future VR experiences? (8) Do you have any other suggestions or comments about the VR experience?

Statistical Analysis

Data analysis was performed using IBM SPSS Statistics (version 29.0). Descriptive statistics, including counts, SDs, means, and proportions, were calculated for demographic data and feasibility factors.

XSL•FO
RenderX

The Shapiro-Wilk test was used to assess data normality for outcome variables, including NPRS, Pain Self-Efficacy, WHO-5, and Mood. In addition, Les Mills BODYCOMBAT performance scores, postintervention WHO-5 scores, and average mood scores after VR sessions were evaluated separately for group analysis.

Mann-Whitney U tests and independent sample *t* tests were used to compare group differences based on data normality. Cohen *d* was calculated to assess the magnitude of VR's impact on parametric variables. Statistical significance was set at a 2-tailed $P<.05$. Responses to the final open-ended questions were transcribed and analyzed using a combination of traditional and inductive content analysis methods, as recommended by Vaismoradi et al [60].

All hypothesis tests were exploratory; no formal adjustment for multiplicity was applied, so results should be interpreted with caution regarding potential type I error inflation.

Because multiple outcomes were analyzed without formal adjustment, *P* values near the threshold (eg, pain self-efficacy, $P=.03$) should be viewed as exploratory due to potential type I error inflation.

Ethical Considerations

The study was conducted following the Declaration of Helsinki and approved by the Research Ethics Committee of Hong Kong Metropolitan University (HE-RD/2023/1.15). All participants provided written informed consent permitting both primary data collection and secondary analysis of deidentified data. Datasets were pseudoanonymized with numeric identifiers, and the data analyst did not have access to the reidentification key. This key was securely stored on an encrypted, password-protected device accessible only to senior members of the research team not involved in data analysis. Participants received HK$150 (ie, US $19.11) supermarket coupons as compensation for their participation. Illustrative photographs do not allow identification because VR headsets and surgical masks obscure the face; a visible coauthor granted written permission for publication.

Results

Participant Progression

As shown in Figure 5, a total of 18 participants were recruited from the day care center. Of these, 5 were excluded based on eligibility criteria, resulting in a recruitment rate of 72.2%. During the 3 VR intervention sessions, 2 participants withdrew: 1 reported fatigue, and the other preferred not to participate further as it interfered with his usual mahjong activities. In total, 11 participants completed the intervention and were included in the final analysis.

XSL•FO
RenderX

Figure 5. Flow diagram of participant progress throughout the intervention.

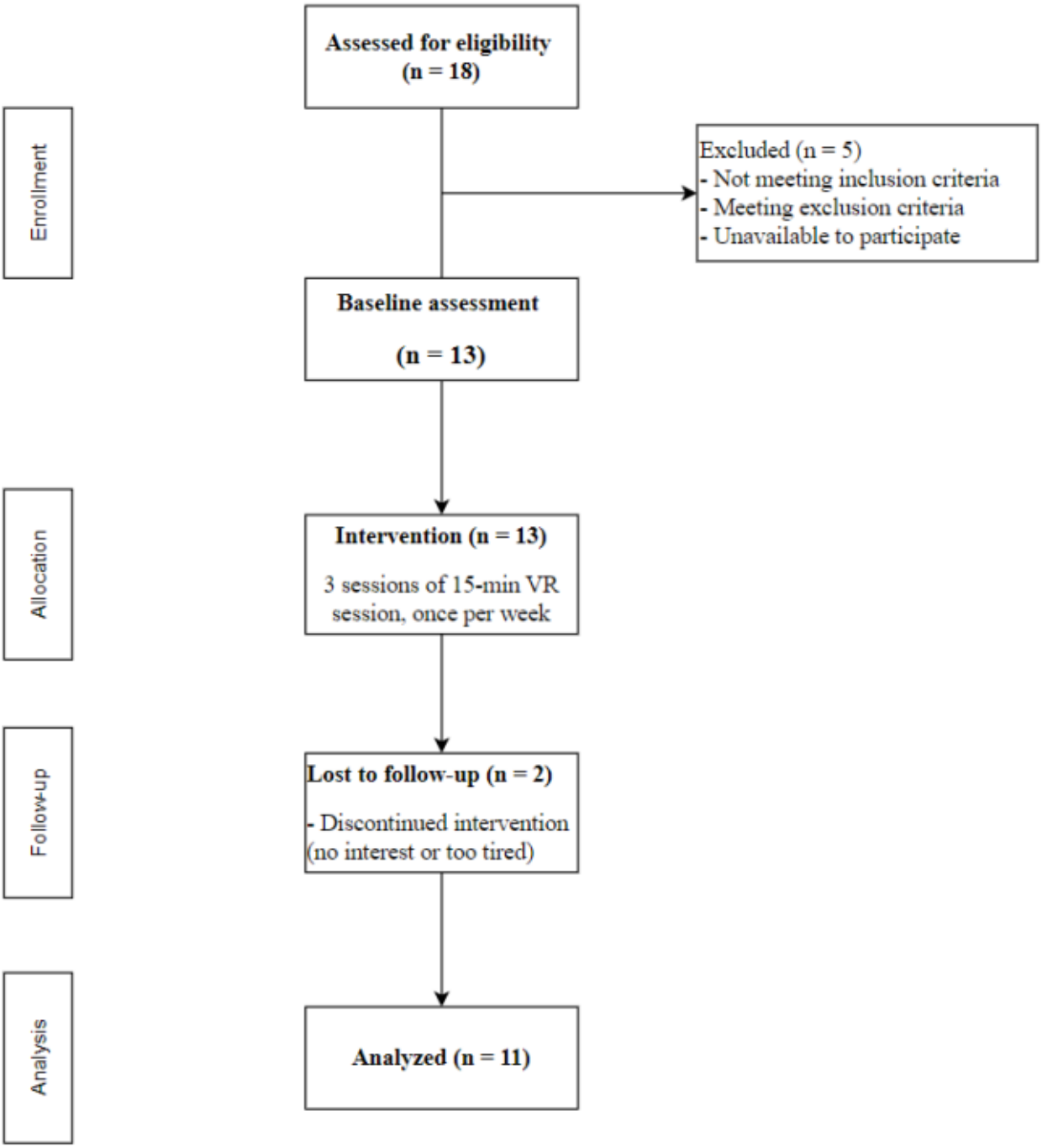

Demographic Characteristics

Table 1 summarizes the baseline characteristics of the 13 participants. The participants had a mean age of 79.23 (SD 9.221). Eight participants reported having chronic pain for over 3 years, while the remaining 5 were nonpainful older adults. Most of the participants had attended primary school or below, lived with others, had no VR experience, and had no glaucoma or cataract. Among the participants, 9 (69.2%) of the participants used smart devices in daily life.

XSL•FO
RenderX

Table . Baseline characteristics of the participants (N=13).

Characteristic	VR[a] participant
Mean age (SD)	79.23 (9.221)
Sex, n (%)	
Female	7 (46.2)
Male	6 (53.8)
Education level, n (%)	
Primary school or below	11 (84.6)
Secondary school	2 (15.4)
Marital status, n (%)	
Single (widowed, divorced, or unmarried)	6 (46.2)
Married	7 (53.8)
Living arrangement, n (%)	
Living with others (spouse, children, or maid)	12 (92.3)
Alone	1 (7.7)
Employment status, n (%)	
Retired	13 (100)
Income level, n (%)	
<HK$ 5000 (ie, US $636.95)	11 (84.6)
HK$5001-HK$ 10,000 (ie, US $637.07 to US $1273.89)	1 (7.7)
HK$10,001-HK$ 20,000 (ie, US $1274.02 to US $2547.79)	1 (7.7)
Chronic pain status, n (%)	
No	5 (38.5)
Yes	8 (61.5)
Duration of pain, n (%)	
>3 y	8 (100)
With glaucoma or cataract, n (%)	
No	9 (69.2)
Yes	4 (30.8)
VR experience before, n (%)	
No	11 (84.6)
Yes	2 (15.4)
Number of smart devices used (smartphone, iPad, laptop, and smartwatches), n (%)	
Not use	4 (30.8)
In use (≤2)	9 (69.2)

[a]VR: virtual reality.

Primary Outcomes

Preliminary Signals

Table 2 shows the efficacy of the VR intervention through a comparative analysis of scores across various outcome measures using appropriate parametric and nonparametric statistical tests.

For pain intensity measured by NPRS, despite a 20.04% decrease in the mean after the intervention, the paired samples *t* test revealed no statistically significant difference, with a small effect size (P=.53; Cohen d=−0.25).

In the case of pain self-efficacy, a median improvement of −2.0 was observed. This change was statistically significant, as shown by the Wilcoxon signed-rank test (z=−2.236; P=.03). No effect size was computed due to the nonparametric nature of the test.

For the WHO-5 well-being index, no increase in the median score was found, and the test indicated no statistical significance

XSL•FO
RenderX

following the intervention (z=1.427; *P*=.15). The effect size was not calculated for this nonparametric test.

Finally, a significant difference was detected in mood scores. The mean score was increased by 2.64 points (16.32%), supported by a large effect size (Cohen *d*=1.82) and statistically significant (*P*<.001).

Table . Summary of primary outcomes of the virtual reality intervention.Study design: a single-arm, mixed methods pilot study. Study population: N=13 (nonpain: n=5; pain: n=8). Study time frame: virtual reality sessions were held once weekly for 3 consecutive weeks. Study location: an elderly day care center in Hong Kong.

Outcome variable	Preintervention, mean (SD)	Postintervention, mean (SD)	Preintervention, median (IQR)	Postintervention, median (IQR)	Mean difference (SD; %) or median (IQR)	z-value	*P* value	Cohen *d* (95% CI)
NPRS[a]	4.29 (2.984)	3.43 (3.69)	N/A[b]	N/A[b]	−0.86 (SD 3.44; −20.04%)	N/A[b]	.53	−0.25 (95% CI −1.00 to 0.50)
Pain self-efficacy[c]	N/A[b]	N/A[b]	3.0 (1.5 - 3.0)	1.0 (1.0 - 2.0)	−2.0 (−2.0 to 0.0)	−2.236	.03	N/A[b]
WHO-5[d]	N/A[b]	N/A[b]	16.0 (12-17)	16.0 (14-18)	2.0 (−2.0 to 6.0)	1.427	.15	N/A[b]
Mood[e]	16.18 (1.319)	18.82 (0.9)	N/A[b]	N/A[b]	2.64 (SD 1.45; 16.32%)	N/A[b]	<.001	1.82 (95% CI 0.86 to 2.78)

[a]NPRS: Numeric Pain Rating Scale.

[b]N/A: not applicable.

[c]Pain self-efficacy: a belief of chronic pain sufferers that they can perform specific activities despite pain, measured by a 5-point Likert scale from 0 (none at all) to 5 (extremely).

[d]WHO-5: World Health Organization–Five Well-Being Index.

[e]Mood: a 5-point Likert scale with 4 questions to capture the current mood states of participants.

Feasibility

The acceptability and applicability of the commercial-off-the-shelf VR applications by the participants were determined by three criteria:

1. Intervention completion rate: descriptively, 7 of 8 participants with pain (87.5%) and 4 of 5 without pain (80%) completed all 3 sessions.
2. Adverse outcome: the second criterion concerns VR sickness. Most participants did not have any issues with VR sickness. Out of 36 participation instances, two participants reported general discomfort, one related to the eye mask and the other to existing back pain (Table 3). In addition, one participant experienced nervousness during the game and hand fatigue afterward.
3. Sentiment analysis: the last measure of VR feasibility was based on the sentiment analysis of participants' feedback during the conclusive interview. Positive and negative responses to 4 questions in a semistructured questionnaire were analyzed using content analysis.

XSL•FO
RenderX

Table . Summary table of adverse outcomes.

VR[a] sickness item	Number of incidence, N=36[b] (%)	Remark
General discomfort	2 (5.6%)	• The eye mask applied a slight pressure on the participant's nose bridge. • The participant hurt his back that morning and felt uncomfortable while playing VR.
Fatigue	1 (2.8%)	• The participant felt a bit nervous while playing the VR boxing game, but she felt happy and had hand fatigue after the match.
Headache	0	—
Eye strain	0	—
Difficulty focusing	0	—
Nausea	0	—
Blurred vision	0	—
Dizziness	0	—
Vertigo	0	—

[a]VR: virtual reality.

[b]The denominator is the sum of 13 participants in T1, 12 participants in T2, and 11 participants in T3. Each participant has the chance to have VR sickness, and therefore, their total number of participation is the denominator.

In Multimedia Appendix 2, 90.9% of participants (10/11) reported enjoying the VR experience. About 72.73% (8/11) indicated that VR did not induce pain, 81.82% (9/11) found VR easy to operate, and 100% (11/11) stated they would participate again. Among those with pain, approximately 27.27% mentioned a negative impact. The high percentage of positive feedback demonstrates the practicality and acceptance of VR among older adults.

Analysis by Pain Group

This analysis evaluates potential differences in outcomes between the pain and nonpain groups. It aims to analyze their overall well-being, immediate mood, and game performance in terms of the three outcome variables: WHO-5, mood, and score (the game score in Les Mills BODYCOMBAT) after the intervention.

As shown in Table 4, there was no significant difference between the two groups in their overall well-being and immediate mood, as measured by WHO-5 and mood scores after the intervention (P>.05). The lack of significance was supported by the relatively small Cohen d effect sizes for both variables.

For the distribution of game performance scores, the nonpain group had a median of 46,513.5 (IQR 73,674.25; range 10,676-93,648), while the pain group had a median of 8005 (IQR 39,460; range 2380-42,441). Despite this observed difference, the U value of 5 and the P value of .11 indicate that the difference was not statistically significant. No effect size was reported due to the nonparametric nature of the Mann-Whitney U test.

XSL•FO
RenderX

Table . Group analysis for pain and nonpain participants.

Variable and group	Mean (SD)	Median (IQR)	Test statistic (U)	*t* test (*df*)	*P* value	Cohen *d* (95% CI)
AT3_WHO5[a]				−0.21 (9)	.84	−0.13 (95% CI −1.36 to 1.10)
Pain	16.14 (3.08)	N/A[b]	N/A[b]	—	—	—
Nonpain	16.75 (6.65)	N/A[b]	N/A[b]	—	—	—
Ave_After_Mood[c]				0.64 (9)	.54	0.4 (95% CI −0.84 to 1.64)
Pain	18.95 (0.83)	N/A[b]	N/A[b]	—	—	—
Nonpain	18.58 (1.1)	N/A[b]	N/A[b]	—	—	—
AT3_Scores[d]				N/A[b]	.11	N/A[b]
Pain	N/A[b]	8005 (4461.5 - 32089.5)	5	—	—	—
Nonpain	N/A[b]	46513.5 (17149.5 - 81526)	N/A[b]	—	—	—

[a]AT3_WHO5: the postintervention World Health Organization–Five Well-Being Index (WHO-5) score after the third (last) virtual reality session.

[b]N/A: not applicable.

[c]Ave_After_Mood: the average postintervention mood scores.

[d]AT3_Scores: the scores obtained in Les Mills BODYCOMBAT after the third (last) virtual reality session.

Qualitative Findings

In addition to the sentiment analysis, we further coded the participants' responses into three themes. Participant feedback is detailed in Multimedia Appendix 3.

Theme 1: Enjoyment of VR Experience Despite Existing Conditions

Most of the participants expressed a strong appreciation for the overall VR experience, especially the 360-degree videos. These videos provided a close-to-reality feeling of traveling to places they had never been or could not visit due to varying factors such as inaccessibility, age, and physical limitations.

Theme 2: Physical and Psychological Impact of VR

According to the biopsychosocial model, VR and exercise had a positive influence on pain relief, as well as physical and psychological well-being [40-42]. Similar comments were found in the feedback from the participants.

Theme 3: Suggestions for Improvement in VR Settings and Design

Despite the overall positive feedback and the steps to adapt the nonnative commercial VR applications for older adults, one participant still raised a language barrier due to the English interface and reported comfort issues with the VR headset. Some participants also requested longer VR sessions and more varied content. These findings provide valuable insight to improve the research design and methods in the future.

Discussion

Principal Findings

This pilot study is one of the first to explore the feasibility and preliminary signals of nonlocalized, commercial-off-the-shelf VR applications for managing chronic pain and the well-being of community-dwelling older adults. The current study aimed to address three primary research questions: (1) evaluate the feasibility of using commercial-off-the-shelf VR applications for managing chronic pain among older adults, (2) assess the efficacy of VR in alleviating chronic pain, and (3) examine its impact on the well-being of older adults in a community setting. We found that nonlocalized commercial VR applications demonstrate good feasibility and acceptability in managing chronic pain among older adults in the community. However, their direct effects on pain intensity and overall well-being still need further exploration. These results will be discussed in detail in the following paragraphs.

Comparison With Prior Work

First, regarding the feasibility of using commercial-off-the-shelf VR applications, VR applications have been developed specifically for chronic pain therapy or rehabilitation in clinical settings [15-22,32]. However, the high completion rates, low adverse outcomes, and positive qualitative feedback demonstrated good feasibility and acceptance of market-ready VR applications among older adults in the community. Unlike previous studies that highlighted limited English proficiency, digital literacy, and cultural relevance as challenges to VR use among older adults [36-38], our study found that a structured approach to explain key aspects to older adults and warm-up

XSL•FO
RenderX

sessions before the VR intervention could effectively alleviate potential anxiety.

In addition, the practical implementation of commercial VR applications in real-world settings demonstrates considerable potential for scalability. For instance, the Meta Quest 3 is priced at approximately US $500, making it a relatively accessible technology. In contrast to other studies where VR interventions are typically administered by a qualified physiotherapist [26], a trained facilitator can also effectively oversee groups of 5-6 seated older adults after a single one-to-one orientation session. This ensures the safe and efficient use of VR applications. In addition, the option to conduct sessions at home eliminates transportation challenges, with informal caregivers able to lend support. This finding is consistent with previous studies demonstrating home-based VR therapy as a viable option for patients with chronic pain [61]. These practical factors collectively highlight the affordability, accessibility, and sustained engagement potential of commercial VR for older adults living in community settings.

Next, about the efficacy of VR for managing chronic pain and well-being, the pilot study yielded mixed results. The VR intervention showed a significant impact on emotional (immediate mood) and psychological aspects of pain management. Our qualitative findings also aligned with theoretical models [40-45], demonstrating that VR technology and exercise can provide psychological benefits. This result is consistent with the opinions of other researchers [62-64]. For instance, research showed that after receiving virtual reality hypnosis treatment, patients spent less time thinking about pain, and the reported intensity of the worst pain, anxiety, and pain discomfort were significantly reduced [63]. Distracting the participants' attention during VR intervention may be an important psychological mechanism for pain relief [62-64]. However, no statistically significant change was detected in pain intensity or overall well-being (NPRS and WHO-5), possibly because self-efficacy might be cognitively mediated and may improve before nociceptive intensity responds [65]. This pattern aligns with contemporary pain models that emphasize affective–cognitive appraisal [66]. In small pilots, these outcomes, therefore, might not move in tandem. Larger, controlled trials with longer follow-ups are required to determine whether the psychological gains translate into clinically meaningful pain relief.

In addition, we detected no statistically significant differences in postsession mood, WHO-5 scores, or boxing-game performance between pain and nonpain participants. Although the small sample size limits broad conclusions, our results tentatively indicate that both subgroups could effectively use and enjoy the VR protocol. A larger, controlled trial is needed to confirm these findings.

Limitations

This pilot study had several limitations that should be considered when interpreting the results. First, the study included a small number of participants in one center. The limited number of participants reduces statistical power, increasing the risk of not detecting true effects. Hence, a larger sample size from multiple centers would have provided higher statistical power for the analysis and broader applicability of the findings (refer to Multimedia Appendix 4 for the use of ChatGPT to improve the readability). For instance, based on the large mood improvement observed in this study (Cohen d=1.8), a two-arm trial would require approximately 34 participants per group to achieve 80% power at a 0.05 significance level.

Second, this single-arm design made it difficult to draw causal inferences about the VR intervention's effectiveness. Because this was an uncontrolled, one-arm pilot, any pre-post change could equally be attributable to placebo effects, Hawthorne reactivity, or simple regression to the mean rather than to the VR content itself. To address these threats, we are currently planning a randomized controlled trial that will compare the same VR protocol with an attention-matched control condition.

Third, the short duration of each VR session might not have been sufficient to observe significant changes in chronic pain and well-being. Previous studies with longer and more intensive VR interventions have shown better outcomes [29]. This pilot study only collected immediate postsession outcomes; our forthcoming randomized trial will obtain 1- and 3-month follow-up data to establish the durability of effects. Future research should extend the duration and intensity of the intervention and align with physical exercise guidelines to enhance its impact. These trials should test longer interventions with a 1 - to 3-month follow-up period to help determine if short-term psychological benefits lead to lasting reductions in pain and improvements in overall well-being.

As commercial-off-the-shelf VR applications may not fully support the local language, the research team tried to translate 360-degree videos and provided real-time interpretation for participants. However, this may still have limited participant engagement and enjoyment, as the experience was not fully immersive in their native language. While this served as a temporary solution, it highlights the need for developing native-language VR content. Future studies should prioritize VR applications in participants' native languages whenever available (refer to Multimedia Appendix 4 for the use of ChatGPT to improve the readability). Furthermore, trials comparing language-localized and nonlocalized VR applications could show whether native-language content boosts engagement and adherence in older adults.

In addition, the reliance on self-reported measures introduced potential biases such as social desirability and recall errors. Future studies should incorporate objective measures such as the Medical Research Council (MRC) Scale for muscle strength or activities of daily living assessments to provide a more comprehensive evaluation. Performance scores from VR exercises, such as scores in Les Mills BODYCOMBAT, could also be used to objectively track progress and motivate participants, in line with self-determination theory. Future research should include objective measures of physical function, such as timed-up-and-go tests, Medical Research Council muscle grading, or assessments of daily activities, to complement self-reported outcomes.

Finally, though the participants were recruited from the same day care center, there was significant variability in their pain levels, visual abilities, and experience with VR technology.

XSL•FO
RenderX

This heterogeneity could influence the results. It made it challenging to attribute outcomes solely to the VR intervention. Future research should use stratified analyses or larger samples from various locations to control for these variables. This would provide clearer insights into the intervention's effects.

Beyond the limitations above, several practical questions also require exploration. Research in various community settings is essential to evaluate scalability. These studies should consider the current headset cost (approximately US $500) and examine the facilitator-to-participant ratio used in this study. Here, one trained staff member supported 5-6 seated users after initial training. Cost-effectiveness studies and mixed methods research on implementation would further guide real-world use. By tackling these gaps, future work can build a strong evidence base for using commercial virtual reality applications with older adults living in the community.

Conclusions

This pilot study represents a pioneering exploration into the use of nonlocalized and commercial-off-the-shelf VR applications to manage chronic pain and well-being among older adults in community settings. This pilot demonstrated that despite language barriers, insufficient digital literacy, and cultural differences, these VR applications demonstrated good feasibility and acceptability among older adults through structured programs. Specifically, high completion rates, low adverse outcomes, and positive qualitative feedback attested to their promising prospects for application. In this uncontrolled sample, VR sessions were associated with large, within-group improvements in mood and modest gains in pain self-efficacy. While causal inferences cannot be drawn due to limitations in methodology, this study lays the groundwork for more rigorous future exploration. Through continuous research, development, and validation, VR technology is expected to become an economical, convenient, and effective nonpharmacological intervention for chronic pain management in older adults, thereby promoting aging in place and comprehensively improving their quality of life.

Acknowledgments

The work described in this paper was partially supported by the Hong Kong Metropolitan University Research (RD/2023/1.15). This work was initiated as a capstone project for the Master of Science in Smart Ageing and Gerontology program, a dual master's degree program offered jointly by Hong Kong Metropolitan University and Lingnan University. The authors used OpenAI's ChatGPT-4 and ChatGPT-4o versions for improving the language and enhancing readability. RLYF, MZY, and TTOK reviewed, edited, and fact-checked the outputs and take full responsibility for the final content.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The scene of 360-degree videos.
[DOCX File, 1845 KB - formative_v9i1e67765_app1.docx]

Multimedia Appendix 2
Sentiment analysis of the semistructured questionnaire.
[DOCX File, 20 KB - formative_v9i1e67765_app2.docx]

Multimedia Appendix 3
Qualitative analysis of the conclusive interview data.
[DOCX File, 20 KB - formative_v9i1e67765_app3.docx]

Multimedia Appendix 4
An example of a ChatGPT prompt-response pair used for language editing support.
[PDF File, 2504 KB - formative_v9i1e67765_app4.pdf]

References

1. Sutherland I, Sproull B. First virtual reality (VR) headset. Guinness World Records Limited 2025. URL: https://www.guinnessworldrecords.com/world-records/515907-first-virtual-reality-vr-headset [accessed 2024-09-24]
2. Hamad A, Jia B. How virtual reality technology has changed our lives: an overview of the current and potential applications and limitations. Int J Environ Res Public Health 2022 Sep 8;19(18):11278. [doi: 10.3390/ijerph191811278] [Medline: 36141551]
3. Maggio MG, Latella D, Maresca G, et al. Virtual reality and cognitive rehabilitation in people with stroke: an overview. J Neurosci Nurs 2019 Apr;51(2):101-105. [doi: 10.1097/JNN.0000000000000423] [Medline: 30649091]

XSL•FO
RenderX

4. Chan JYC, Chan TK, Wong MPF, Cheung RSM, Yiu KKL, Tsoi KKF. Effects of virtual reality on moods in community older adults. A multicenter randomized controlled trial. Int J Geriatr Psychiatry 2020 Aug;35(8):926-933. [doi: 10.1002/gps.5314] [Medline: 32346896]
5. Huang LC, Yang YH. The long-term effects of immersive virtual reality reminiscence in people with dementia: longitudinal observational study. JMIR Serious Games 2022 Jul 25;10(3):e36720. [doi: 10.2196/36720] [Medline: 35877169]
6. Kwan RYC, Ng F, Lam LCW, Yung RC, Sin OSK, Chan S. The effects of therapeutic virtual reality experience to promote mental well-being in older people living with physical disabilities in long-term care facilities. Trials 2023 Aug 26;24(1):558. [doi: 10.1186/s13063-023-07592-7] [Medline: 37633916]
7. Ng E, Shu L, Yeung Y, Kwan C, Lau J. When healthcare meets virtual reality. HASLink. 2019. URL: https://www3.ha.org.hk/ehaslink/issue102/HASLink_1.pdf [accessed 2024-09-24]
8. Pandrangi VC, Shah SN, Bruening JD, et al. Effect of virtual reality on pain management and opioid use among hospitalized patients after head and neck surgery: a randomized clinical trial. JAMA Otolaryngol Head Neck Surg 2022 Aug 1;148(8):724-730. [doi: 10.1001/jamaoto.2022.1121] [Medline: 35679057]
9. Specht BJ, Buse CR, Phelps JR, et al. Virtual reality after surgery-a method to decrease pain after surgery in pediatric patients. Am Surg 2023 Apr;89(4):596-602. [doi: 10.1177/00031348211032204] [Medline: 34275343]
10. Smith TJ, Hillner BE. The cost of pain. JAMA Netw Open 2019 Apr 5;2(4):e191532. [doi: 10.1001/jamanetworkopen.2019.1532] [Medline: 30951152]
11. Darnall BD, Krishnamurthy P, Tsuei J, Minor JD. Self-administered skills-based virtual reality intervention for chronic pain: randomized controlled pilot study. JMIR Form Res 2020 Jul 7;4(7):e17293. [doi: 10.2196/17293] [Medline: 32374272]
12. Jones T, Moore T, Choo J. The impact of virtual reality on chronic pain. PLoS ONE 2016;11(12):e0167523. [doi: 10.1371/journal.pone.0167523] [Medline: 27997539]
13. McConnell R, Lane E, Webb G, LaPeze D, Grillo H, Fritz J. A multicenter feasibility randomized controlled trial using a virtual reality application of pain neuroscience education for adults with chronic low back pain. Ann Med 2024 Dec;56(1):2311846. [doi: 10.1080/07853890.2024.2311846] [Medline: 38354690]
14. Tack C. Virtual reality and chronic low back pain. Disabil Rehabil Assist Technol 2021 Aug;16(6):637-645. [doi: 10.1080/17483107.2019.1688399] [Medline: 31746250]
15. France CR, Thomas JS. Virtual immersive gaming to optimize recovery (VIGOR) in low back pain: a phase II randomized controlled trial. Contemp Clin Trials 2018 Jun;69:83-91. [doi: 10.1016/j.cct.2018.05.001] [Medline: 29730393]
16. Rezaei I, Razeghi M, Ebrahimi S, Kayedi S, Rezaeian Zadeh A. A novel virtual reality technique (Cervigame®) compared to conventional proprioceptive training to treat neck pain: a randomized controlled trial. J Biomed Phys Eng 2019;9(3Jun):355-366. [doi: 10.31661/jbpe.v9i3Jun.556] [Medline: 32039103]
17. Guo Q, Zhang L, Han LL, et al. Effects of virtual reality therapy combined with conventional rehabilitation on pain, kinematic function, and disability in patients with chronic neck pain: randomized controlled trial. JMIR Serious Games 2024 Apr 24;12:e42829. [doi: 10.2196/42829] [Medline: 38656775]
18. Cuneo A, Yang R, Zhou H, et al. The utility of a novel, combined biofeedback-virtual reality device as add-on treatment for chronic migraine: a randomized pilot study. Clin J Pain 2023 Jun 1;39(6):286-296. [doi: 10.1097/AJP.0000000000001114] [Medline: 37026763]
19. Özlü A, Ünver G, Tuna H, Menekşeoğlu AK. The effect of a virtual reality-mediated gamified rehabilitation program on pain, disability, function, and balance in knee osteoarthritis: a prospective randomized controlled study. Games Health J 2023 Apr;12(2):118-124. [doi: 10.1089/g4h.2022.0130] [Medline: 36603100]
20. Won AS, Barreau AC, Gaertner M, et al. Assessing the feasibility of an open-source virtual reality mirror visual feedback module for complex regional pain syndrome: pilot usability study. J Med Internet Res 2021 May 26;23(5):e16536. [doi: 10.2196/16536] [Medline: 34037530]
21. Chau B, Phelan I, Ta P, et al. Immersive virtual reality for pain relief in upper limb complex regional pain syndrome: a pilot study. Innov Clin Neurosci 2020 Apr 1;17(4-6):47-52. [Medline: 32802594]
22. Lewis JS, Newport R, Taylor G, Smith M, McCabe CS. Visual illusions modulate body perception disturbance and pain in complex regional pain syndrome: a randomized trial. Eur J Pain 2021 Aug;25(7):1551-1563. [doi: 10.1002/ejp.1766] [Medline: 33759278]
23. Fowler CA, Ballistrea LM, Mazzone KE, et al. Virtual reality as a therapy adjunct for fear of movement in veterans with chronic pain: single-arm feasibility study. JMIR Form Res 2019 Oct 30;3(4):e11266. [doi: 10.2196/11266] [Medline: 31670696]
24. Tejera DM, Beltran-Alacreu H, Cano-de-la-Cuerda R, et al. Effects of virtual reality versus exercise on pain, functional, somatosensory and psychosocial outcomes in patients with non-specific chronic neck pain: a randomized clinical trial. Int J Environ Res Public Health 2020 Aug 16;17(16):5950. [doi: 10.3390/ijerph17165950] [Medline: 32824394]
25. Harvie DS, Stanton TR, Kennedy H, Coppieters MW. Visually evoked pain and its extinction using virtual reality in a patient with complex regional pain syndrome type II. Pain 2022 Oct 1;163(10):1874-1878. [doi: 10.1097/j.pain.0000000000002605] [Medline: 35297803]
26. Tuck N, Pollard C, Good C, et al. Active virtual reality for chronic primary pain: mixed methods randomized pilot study. JMIR Form Res 2022 Jul 13;6(7):e38366. [doi: 10.2196/38366] [Medline: 35830224]

XSL•FO
RenderX

27. Booth J, Moseley GL, Schiltenwolf M, Cashin A, Davies M, Hübscher M. Exercise for chronic musculoskeletal pain: a biopsychosocial approach. Musculoskeletal Care 2017 Dec;15(4):413-421. [doi: 10.1002/msc.1191] [Medline: 28371175]
28. Ambrose KR, Golightly YM. Physical exercise as non-pharmacological treatment of chronic pain: why and when. Best Pract Res Clin Rheumatol 2015 Feb;29(1):120-130. [doi: 10.1016/j.berh.2015.04.022] [Medline: 26267006]
29. Wong KP, Tse MMY, Qin J. Effectiveness of virtual reality-based interventions for managing chronic pain on pain reduction, anxiety, depression and mood: a systematic review. Healthcare (Basel) 2022 Oct 17;10(10):2047. [doi: 10.3390/healthcare10102047] [Medline: 36292493]
30. Lo HHM, Zhu M, Zou Z, et al. Immersive and nonimmersive virtual reality-assisted active training in chronic musculoskeletal pain: systematic review and meta-analysis. J Med Internet Res 2024 Aug 19;26(1):e48787. [doi: 10.2196/48787] [Medline: 39159449]
31. Wong CL, Li CK, Chan CWH, et al. Virtual reality intervention targeting pain and anxiety among pediatric cancer patients undergoing peripheral intravenous cannulation: a randomized controlled trial. Cancer Nurs 2021;44(6):435-442. [doi: 10.1097/NCC.0000000000000844] [Medline: 32511154]
32. Wong CL, Choi KC. Effects of an immersive virtual reality intervention on pain and anxiety among pediatric patients undergoing venipuncture: a randomized clinical trial. JAMA Netw Open 2023 Feb 1;6(2):e230001. [doi: 10.1001/jamanetworkopen.2023.0001] [Medline: 36795410]
33. Woo OKL, Lee AM. Case report: therapeutic potential of Flourishing-Life-Of-Wish Virtual Reality Therapy on Relaxation (FLOW-VRT-Relaxation)-a novel personalized relaxation in palliative care. Front Digit Health 2023;5:1228781. [doi: 10.3389/fdgth.2023.1228781] [Medline: 37674510]
34. Chiu PL, Li H, Yap KYL, Lam KMC, Yip PLR, Wong CL. Virtual reality-based intervention to reduce preoperative anxiety in adults undergoing elective surgery: a randomized clinical trial. JAMA Netw Open 2023 Oct 2;6(10):e2340588. [doi: 10.1001/jamanetworkopen.2023.40588] [Medline: 37906193]
35. Lo HHM, Ng M, Fong PYH, et al. Examining the feasibility, acceptability, and preliminary efficacy of an immersive virtual reality-assisted lower limb strength training for knee osteoarthritis: mixed methods pilot randomized controlled trial. JMIR Serious Games 2024 Sep 27;12(1):e52563. [doi: 10.2196/52563] [Medline: 39331525]
36. Dy M, Olazo K, Lyles CR, et al. Usability and acceptability of virtual reality for chronic pain management among diverse patients in a safety-net setting: a qualitative analysis. JAMIA Open 2023 Oct;6(3):ooad050. [doi: 10.1093/jamiaopen/ooad050] [Medline: 37449058]
37. Sarkar U, Lee JE, Nguyen KH, Lisker S, Lyles CR. Barriers and facilitators to the implementation of virtual reality as a pain management modality in academic, community, and safety-net settings: qualitative analysis. J Med Internet Res 2021 Sep 22;23(9):e26623. [doi: 10.2196/26623] [Medline: 34550074]
38. Birckhead B, Eberlein S, Alvarez G, et al. Home-based virtual reality for chronic pain: protocol for an NIH-supported randomised-controlled trial. BMJ Open 2021 Jun 15;11(6):e050545. [doi: 10.1136/bmjopen-2021-050545] [Medline: 34130965]
39. Julious SA. Sample size of 12 per group rule of thumb for a pilot study. Pharm Stat 2005 Oct;4(4):287-291. [doi: 10.1002/pst.185]
40. Engel GL. The need for a new medical model: a challenge for biomedicine. Science 1977 Apr 8;196(4286):129-136. [doi: 10.1126/science.847460] [Medline: 847460]
41. Miaskowski C, Blyth F, Nicosia F, et al. A biopsychosocial model of chronic pain for older adults. Pain Med 2020 Sep 1;21(9):1793-1805. [doi: 10.1093/pm/pnz329] [Medline: 31846035]
42. Meints SM, Edwards RR. Evaluating psychosocial contributions to chronic pain outcomes. Prog Neuropsychopharmacol Biol Psychiatry 2018 Dec 20;87(Pt B):168-182. [doi: 10.1016/j.pnpbp.2018.01.017] [Medline: 29408484]
43. Trachsel LA, Munakomi S, Cascella M. Pain Theory: StatPearls Publishing; 2023. URL: https://www.ncbi.nlm.nih.gov/books/NBK545194/ [accessed 2025-07-25]
44. Campbell TS, Johnson JA, Zernicke KA. Gate control theory of pain. In: Encyclopedia of Behavioral Medicine: Springer International Publishing; 2020:914-916 URL: https://doi.org/10.1007/978-3-030-39903-0_1134 [doi: 10.1007/978-3-030-39903-0_1134]
45. Navas-Otero A, Calvache-Mateo A, Martín-Núñez J, et al. The effectiveness of combined exercise and self-determination theory programmes on chronic low back pain: a systematic review and metanalysis. Healthcare (Basel) 2024 Feb 1;12(3):382. [doi: 10.3390/healthcare12030382] [Medline: 38338267]
46. Vr AL. YouTube VR. Altlabvr.com. URL: https://www.altlabvr.com/youtube-vr [accessed 2024-10-17]
47. Bka Digital Outfitters. LES MILLS BODYCOMBAT. Les Mills. URL: https://www.lesmills.com/workouts/bodycombat-xr/ [accessed 2024-10-17]
48. Ucero-Lozano R, Pérez-Llanes R, López-Pina JA, Cuesta-Barriuso R. 180-degree immersive VR motion visualization in the treatment of haemophilic ankle arthropathy. Haemophilia 2023 Jan;29(1):282-289. [doi: 10.1111/hae.14683] [Medline: 36261396]
49. Restout J, Bernache-Assollant I, Morizio C, et al. Fully immersive virtual reality using 360° videos to manage well-being in older adults: a scoping review. J Am Med Dir Assoc 2023 Apr;24(4):564-572. [doi: 10.1016/j.jamda.2022.12.026] [Medline: 36758621]

XSL•FO
RenderX

50. Farrar JT, Young JP Jr, LaMoreaux L, Werth JL, Poole MR. Clinical importance of changes in chronic pain intensity measured on an 11-point numerical pain rating scale. Pain 2001;94(2):149-158. [doi: 10.1016/S0304-3959(01)00349-9]
51. Bielewicz J, Daniluk B, Kamieniak P. VAS and NRS, aame or different? Are visual analog scale values and numerical rating scale equally viable tools for assessing patients after microdiscectomy? Pain Res Manag 2022;2022:5337483. [doi: 10.1155/2022/5337483] [Medline: 35391853]
52. Karcioglu O, Topacoglu H, Dikme O, Dikme O. A systematic review of the pain scales in adults: which to use? Am J Emerg Med 2018 Apr;36(4):707-714. [doi: 10.1016/j.ajem.2018.01.008] [Medline: 29321111]
53. Nicholas MK. The pain self-efficacy questionnaire: taking pain into account. Eur J Pain 2007 Feb;11(2):153-163. [doi: 10.1016/j.ejpain.2005.12.008] [Medline: 16446108]
54. Chan L, Liu RKW, Lam TP, Chen JY, Tipoe GL, Ganotice FA. Validation of the World Health Organization Well-Being Index (WHO-5) among medical educators in Hong Kong: a confirmatory factor analysis. Med Educ Online 2022 Dec;27(1):2044635. [doi: 10.1080/10872981.2022.2044635] [Medline: 35275804]
55. Kong CL, Lee CC, Ip YC, Chow LP, Leung CH, Lam YC. Validation of the Hong Kong Cantonese version of World Health Organization Five Well-Being Index for people with severe mental illness. East Asian Arch Psychiatry 2016 Mar;26(1):18-21. [Medline: 27086756]
56. Sischka PE, Costa AP, Steffgen G, Schmidt AF. The WHO-5 well-being index – validation based on item response theory and the analysis of measurement invariance across 35 countries. Journal of Affective Disorders Reports 2020 Dec;1:100020. [doi: 10.1016/j.jadr.2020.100020]
57. Roemer A. Positive and negative affect schedule (PANAS). In: Handbook of Assessment in Mindfulness Research: Springer International Publishing; 2023:1-11. [doi: 10.1007/978-3-030-77644-2_86-1]
58. Kim HK, Park J, Choi Y, Choe M. Virtual reality sickness questionnaire (VRSQ): motion sickness measurement index in a virtual reality environment. Appl Ergon 2018 May;69:66-73. [doi: 10.1016/j.apergo.2017.12.016] [Medline: 29477332]
59. Braun V, Clarke V. Supporting best practice in reflexive thematic analysis reporting in Palliative Medicine: a review of published research and introduction to the Reflexive Thematic Analysis Reporting Guidelines (RTARG). Palliat Med 2024 Jun;38(6):608-616. [doi: 10.1177/02692163241234800] [Medline: 38469804]
60. Vaismoradi M, Turunen H, Bondas T. Content analysis and thematic analysis: implications for conducting a qualitative descriptive study. Nurs Health Sci 2013 Sep;15(3):398-405. [doi: 10.1111/nhs.12048] [Medline: 23480423]
61. Garrett B, Taverner T, McDade P. Virtual reality as an adjunct home therapy in chronic pain management: an exploratory study. JMIR Med Inform 2017 May 11;5(2):e11. [doi: 10.2196/medinform.7271] [Medline: 28495661]
62. Sander Wint S, Eshelman D, Steele J, Guzzetta CE. Effects of distraction using virtual reality glasses during lumbar punctures in adolescents with cancer. Oncol Nurs Forum 2002;29(1):E8-E15. [doi: 10.1188/02.ONF.E8-E15] [Medline: 11845217]
63. Soltani M, Teeley AM, Wiechman SA, Jensen MP, Sharar SR, Patterson DR. Virtual reality hypnosis for pain control in a patient with gluteal hidradenitis: a case report(). Contemp Hypn Integr Ther 2011;28(2):142-147. [Medline: 23205274]
64. Schneider SM, Prince-Paul M, Allen MJ, Silverman P, Talaba D. Virtual reality as a distraction intervention for women receiving chemotherapy. Oncol Nurs Forum 2004;31(1):81-88. [doi: 10.1188/04.ONF.81-88] [Medline: 14722591]
65. Cieslak R, Benight CC, Caden Lehman V. Coping self-efficacy mediates the effects of negative cognitions on posttraumatic distress. Behav Res Ther 2008 Jul;46(7):788-798. [doi: 10.1016/j.brat.2008.03.007] [Medline: 18456241]
66. Eccleston C, Crombez G. Advancing psychological therapies for chronic pain. F1000Res 2017;6:461. [doi: 10.12688/f1000research.10612.1] [Medline: 28413627]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
GDS: Global Deterioration Scale
Mood: Mood Assessment Scale
MRC: Medical Research Council
NPRS: Numeric Pain Rating Scale
PANAS: Positive and Negative Affect Schedule
SCNAMES: Standardized Care Need Assessment Mechanism for Elderly Services
VAS: Visual Analog Scale
VR: virtual reality
VRSQ: Virtual Reality Sickness Questionnaire
WHO-5: World Health Organization–Five Well-Being Index

XSL•FO
RenderX

Edited by A Mavragani; submitted 21.10.24; peer-reviewed by E Maslowski, PK Prof; revised version received 02.06.25; accepted 03.06.25; published 18.08.25.

Please cite as:
Leung RYF, Ye MZ, Zhang FY, Kwok TTO, Sun Y, Lai AYK, Jiang J, Tse MMY
Feasibility and Efficacy of Commercial-Off-the-Shelf Virtual Reality Applications for Managing Chronic Pain and Enhancing Well-Being Among Older Adults in the Community: Mixed Methods Pilot Study
JMIR Form Res 2025;9:e67765
URL: https://formative.jmir.org/2025/1/e67765
doi:10.2196/67765

© Rachel Yim Fong Leung, Megan Zichen Ye, Flora Yaqian Zhang, Tyrone Tai-On Kwok, Yuying Sun, Agnes Yuen Kwan Lai, Juming Jiang, Mimi Mun Yee Tse. Originally published in JMIR Formative Research (https://formative.jmir.org), 18.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Clinical Management of Medication-Assisted Treatment for Opioid Use Disorder Using a Mobile Health App Within a Primary Care Clinic: Quasi-Experimental Study

Allison D Rosen[1], PhD; Steven J Shoptaw[1,2], PhD; Li Li[2], PhD; Bengisu Tulu[3], PhD; Omar Nieto[1], BS; Steven Jenkins[4]; Mariah M Kalmin[5], PhD

[1]Department of Family Medicine, David Geffen School of Medicine, University of California Los Angeles, Suite 1800, 10880 Wilshire Blvd, Los Angeles, CA, United States

[2]Department of Psychiatry and Biobehavioral Sciences, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, United States

[3]The Business School, Worcester Polytechnic Institute, Worcester, MA, United States

[4]Q2i, Boston, MA, United States

[5]Department of Behavioral and Policy Sciences, RAND Corporation, Santa Monica, CA, United States

Corresponding Author:
Allison D Rosen, PhD
Department of Family Medicine, David Geffen School of Medicine, University of California Los Angeles, Suite 1800, 10880 Wilshire Blvd, Los Angeles, CA, United States

Abstract

Background: Medication-assisted treatment (MAT) is an effective strategy for treating opioid use disorder and reducing opioid-related overdose deaths, yet retention in treatment remains low. Mobile health (mHealth) platforms may be a useful tool for increasing long-term engagement in MAT programs, but evaluation studies of such platforms are limited.

Objective: This study aimed to determine whether the use of the Opioid Addiction Recovery Support (OARS) software platform increased MAT engagement for patients with opioid use disorder.

Methods: The Technology Improving Success of Medication-Assisted Treatment in Primary Care Study was a quasi-experimental study conducted at a primary care clinic in the United States between January 2021 and February 2022. OARS is a software platform and mobile app (Q2i, LLC) that includes a dashboard of real-time appointment attendance, urine toxicology (UTOX) results, and educational content as well as messaging and journaling features. All patients who were invited to use OARS and had available data across the study were included in the analysis. The primary outcomes were engagement in treatment, defined as no more than a 35-day gap in appointment attendance, and UTOX. Changes in treatment engagement between the treatment as usual (TAU) period and OARS intervention period were assessed using the effect size (Cohen *g*) and McNemar chi-square test of discordant pairs.

Results: Among 205 patients invited to use OARS, 123 had available data and were thus included in the analysis. The median age was 37 (IQR 31 - 42.5) years, 61% (75/123) identified as men, and 95.1% (117/123) identified as non-Hispanic White. There were no statistically significant differences in demographic characteristics for patients who used OARS on more than 1 day compared to patients who used OARS on 0 or 1 day, or patients who did versus did not have available data. Among all patients, 20% (25/123) were engaged in appointment attendance during TAU only compared to 27% (33/123) during OARS only (g=0.07; P=.36), and 13% (16/123) were engaged in UTOX during TAU only and 33% (41/123) during OARS only (g=0.21; $P\leq.01$). Among a subsample of 52 patients who used OARS on more than 1 day, 17% (9/52) were engaged in appointment attendance during TAU only compared to 23% (12/52) during OARS only (g=0.07, P=.67), and 13% (7/52) were engaged in UTOX during TAU only and 35% (18/52) during OARS only (g=0.22, P=.05).

Conclusions: Introduction of OARS in a primary care setting may be associated with a moderate change in MAT engagement as measured by UTOX, but not appointment attendance. While barriers to implementation and adoption, including difficulty fully integrating OARS with the clinic's electronic health record, may have attenuated the potential effect of the intervention, this study provides evidence that mHealth interventions, such as OARS, are a promising addition to the MAT treatment landscape.

(JMIR Form Res 2025;9:e63526) doi:10.2196/63526

XSL•FO
RenderX

KEYWORDS

mobile health; medication-assisted treatment; medication for opioid use disorder; mHealth; opioid use; primary care; mobile health app; software; assistance; United States; American; behavioral therapy; opioid use disorder

Introduction

The United States is facing an unprecedented and growing opioid overdose epidemic, resulting in more than 80,000 deaths in 2022 [1]. Medication-assisted treatment (MAT), which involves the use of medications such as buprenorphine or methadone in combination with behavioral therapy and counseling, has proven effective for treating opioid use disorder (OUD) and reducing overdose deaths [2-4]. However, many patients do not remain engaged in MAT long-term due to a variety of psychological, social, and economic barriers [5-8].

Software and mobile health (mHealth) platforms offer a complementary service delivery modality for enhancing clinical management of OUD, increasing long-term engagement in MAT, and improving access to care [9-11]. A 2020 review did not find any evaluations of OUD-related mHealth apps [12]. Since then, multiple mobile apps have been identified as feasible and acceptable, including one that presents educational content and recent research (uMAT-r), one that facilitates patient-provider secure messaging (HOPE [Heal. Overcome. Persist. Endure.]), one that allows patients to track symptoms such as mood and cravings (KIOS), and an automated contingency management app that provides immediate monetary incentives for reaching treatment goals (PROCare Recovery) [13-17]. A 2024 scoping review of qualitative studies of OUD-related mHealth also reported high acceptability and perceived benefits by patients [18].

Research that builds on the feasibility and acceptability of these interventions by evaluating their effectiveness in improving OUD treatment outcomes remains limited, and findings are mixed. reSET-0, which uses neurobehavioral therapy techniques, was shown to reduce emergency and inpatient clinical encounters and increase treatment engagement, whereas the Addiction - Comprehensive Health Enhancement Support System (A-CHESS; CHESS Health) smartphone intervention was not associated with increased abstinence when added to medication treatment for OUD [19-21].

Given inconclusive findings from limited research on the effect of mHealth apps in improving OUD treatment outcomes, this analysis aims to contribute to the literature through a quasi-experimental pilot study evaluating the effect of the use of the Opioid Addiction Recovery Support (OARS) software on OUD treatment outcomes in a primary care setting. OARS is a software platform and mobile app (Q2i, LLC) that receives electronic health record (EHR) data to display a provider dashboard of real-time appointment attendance and urine toxicology (UTOX) results. Patients access the OARS mobile app on any smartphone where they can view their treatment records, message providers, interact with educational content including YouTube videos from the Substance Abuse and Mental Health Services Administration, and journal in a daily log. In addition to having access to a patient dashboard, providers can directly interact with patients via messaging and by viewing daily logs in order to inform treatment plans.

This study aimed to determine whether the use of the OARS software platform increased MAT engagement for patients receiving care at a primary care clinic in the United States.

Methods

Study Design

As part of a Small Business Technology Transfer Research Award, a partnership between a small business technology enterprise and an academic institution with research expertise, the Technology Improving Success of Medication-Assisted Treatment in Primary Care Study was conducted between January 2021 and February 2022 [22]. This quasi-experimental study observed patients receiving MAT at a primary care clinic in northeast Pennsylvania for 4 months (treatment as usual [TAU] period: January 1, 2021-April 30, 2021) before implementing the OARS platform clinic-wide for 10 months (OARS period: May 1, 2021-February 28, 2022).

No research staff were on site, and all interactions concerning the study occurred between patients and clinical staff (clinicians, nurses, case managers, etc). All patients receiving MAT at the clinic during the study period were eligible to participate, and the clinic's 28 providers were asked to invite their MAT patients to sign up for OARS; patients were eligible to be included if they were an adult receiving care for OUD, had a smartphone, and agreed to download the OARS app and sign up for an account. The clinic extracted EHR data for all patients who initiated OARS accounts and delivered those data to Q2i during the TAU and OARS periods. Q2i then linked EHR data to OARS usage data, deidentified all data files, and provided the combined data to the research analytical team. All patients who signed up for OARS accounts and had EHR data provided by the clinic to Q2i throughout the study period were included in this analysis.

Measures

The 2 primary outcomes of interest were consistent MAT engagement, as measured by (1) appointment attendance, defined as having no more than a 35-day gap between appointments, and (2) UTOX, defined as having no more than a 35-day gap between UTOX tests, with all UTOX tests being negative for opioids positive for buprenorphine, or both. A window of 35 days was chosen to allow for a 1-week grace period beyond the typical 28-day visit frequency.

Statistical Analysis

Demographic characteristics including age, gender, and race or ethnicity were described using median and IQR values for continuous variables and counts and percentages for categorical variables. Due to implementation challenges that led to lower-than-expected rates of OARS usage, differences in demographic characteristics were compared for patients who

XSL•FO
RenderX

used OARS on more than 1 day and patients who used OARS for 1 day or less, using Wilcoxon signed-rank tests for continuous variables and chi-square tests for categorical variables. A *P* value less than .05 was considered statistically significant.

The 2 outcomes of interest, engagement in MAT as measured by appointment attendance and engagement in MAT as measured by UTOX, were compared for the full sample as well as in a subsample of patients who used OARS regularly (ie, on more than 1 day). Each outcome was compared between the TAU and OARS periods using McNemar chi-square test of discordant pairs to account for paired data; a *P* value less than .03 was considered statistically significant as a Bonferroni adjustment for the 2 comparisons. Effect sizes were calculated for each outcome where *b* equals the number of patients engaged in MAT during the TAU period only and *c* equals the number of patients engaged in MAT during the OARS period only. Cohen *g* was calculated by computing $P-.5$, where *P* was defined as either $b/(b+c)$ or $c/(b+c)$, whichever was larger ($g<0.15$ was considered small, $0.15<g\leq0.25$ was considered moderate, and $g>0.25$ was considered large) [23]. All analyses were conducted in R (version 4.2.1; The R Foundation for Statistical Computing).

Ethical Considerations

This study was approved by the University of California, Los Angeles Institutional Review Board with a waiver of informed consent (#20‐001516). All data were deidentified before being shared with the research team in order to maintain participant privacy. Participants were not compensated.

Results

Among 205 patients invited to use OARS, the clinic provided appointment attendance and UTOX data during both the TAU and OARS periods for 123 patients (Figure 1). There were no differences in demographic characteristics or OARS usage for the 123 patients for whom the clinic provided data and the 82 for whom the clinic did not (Table 1).

Among the 123 patients analyzed, the median age was 37 (IQR 31‐42.5) years, 61% (75/123) identified as men, and 95.1% (117/123) identified as non-Hispanic White. Overall, 96 patients used OARS at least once: 99% (95/96) viewed test results at least once, 71.9% (69/96) viewed their treatment progress at least once, 67.7% (65/96) sent at least one message to their provider and averaged one message sent during the study, 66.7% (64/96) viewed educational content at least once, and 8.3% (8/96) created at least one daily log. There were no statistically significant differences in demographic characteristics for patients who used OARS on more than 1 day compared to patients who used OARS on 0 or 1 day (Table 2).

Figure 1. Flow diagram for the technology improving success of medication-assisted treatment in primary care study. OARS: Opioid Addiction Recovery Support; TAU: treatment as usual.

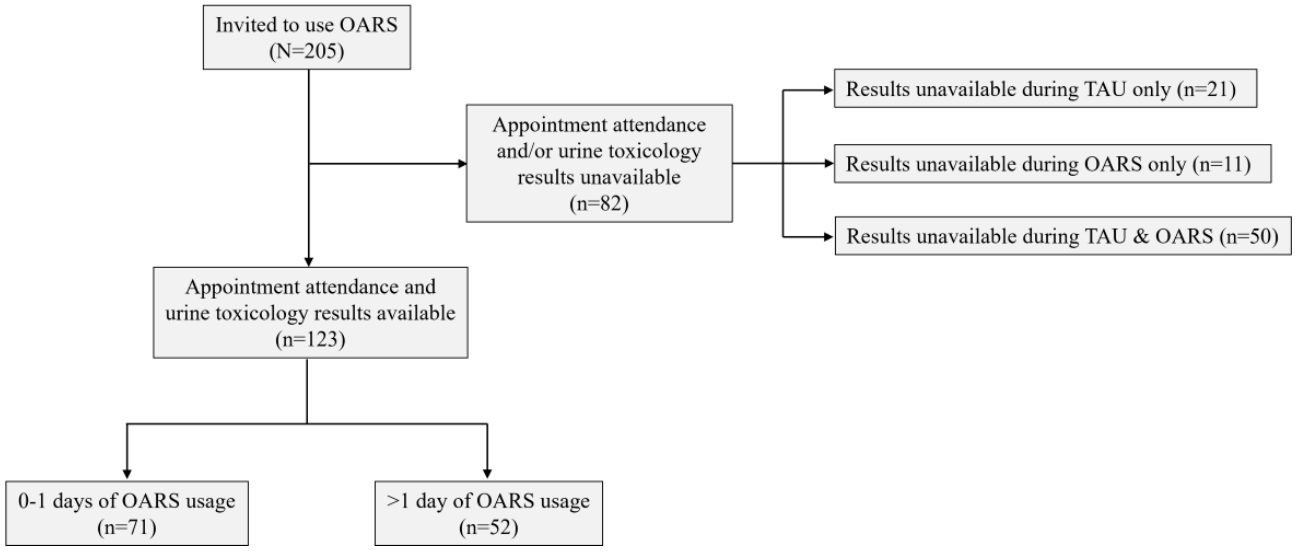

XSL•FO
RenderX

Table . Comparison of demographic characteristics and Opioid Addiction Recovery Support usage among patients for whom the clinic provided appointment and urine toxicology data (included in analysis) and those for whom the clinic did not (excluded from analysis).

Characteristics	Total (N=205)	Included in analysis (n=123)	Excluded from analysis (n=82)	*P* value
Age (years), median (IQR)	37 (31 - 44)	37 (31 - 42.5)	37 (32 - 45.0)	.24
Gender, n (%)				
Male	130 (63.4)	75 (61)	55 (67.1)	.46
Race or ethnicity, n (%)				
Non-Hispanic White	193 (94.1)	117 (95.1)	76 (92.7)	.55
>1 day of OARS[a] usage, n (%)	82 (40)	52 (42.3)	30 (36.6)	.50

[a]OARS: opioid addiction recovery support.

Table . Baseline characteristics of technology improving success of medication-assisted treatment in primary care study participants.

Characteristics	Total (N=123)	0 - 1 days of OARS[a] usage (n=71)	>1 days of OARS[a] usage (n=52)	*P* value
Age (years), median (IQR)	37 (31 - 42.5)	35 (29.5 - 42.5)	37.5 (33.0 - 42.3)	.23
Gender, n (%)				
Male	75 (61)	43 (60.6)	32 (61.5)	.91
Race or ethnicity, n (%)				
Non-Hispanic White	117 (95.1)	69 (95.8)	48 (92.3)	.37

[aa]OARS: opioid addiction recovery support

Among the full sample of 123 patients, 20% (25/123) were engaged in appointment attendance during TAU only, compared to 27% (33/123) during OARS only, and 13% (16/123) were engaged in UTOX during TAU only and 33% (40/123) during OARS only; 7 participants (6%) tested positive for opioids over the course of the study, making the majority of UTOX engagement based on missed urine drug screens. There was no statistically significant difference in engagement in appointment attendance between the TAU and OARS periods, and the effect size was small (g=0.07, 95% CI −.06 - 0.19; P=.36). There was greater UTOX engagement in the OARS period compared to the TAU period with a moderate effect size (g=0.21, 95% CI 0.09 - 0.33; P<.01) (Figure 2).

Among 52 patients who used OARS on more than 1 day, 17% (9/52) were engaged in appointment attendance during TAU only compared to 23% (12/52) during OARS only, and 13% (7/52) were engaged in UTOX during TAU only and 35%(18/52) during OARS only. There was no statistically significant difference in engagement in appointment attendance between the TAU and OARS periods, and the effect size was small (g=0.07, 95% CI −0.15 - to 0.29; P=.67). There was also no statistically significant difference in UTOX engagement between the TAU and OARS periods, and the effect size was moderate (g=0.22, 95% CI 0.02 - 0.39; P=.05,) (Figure 2).

XSL•FO
RenderX

Figure 2. Medication-assisted treatment engagement during the study period. OARS: Opioid Addiction Recovery Support; MAT: medication-assisted treatment; TAU: treatment as usual.

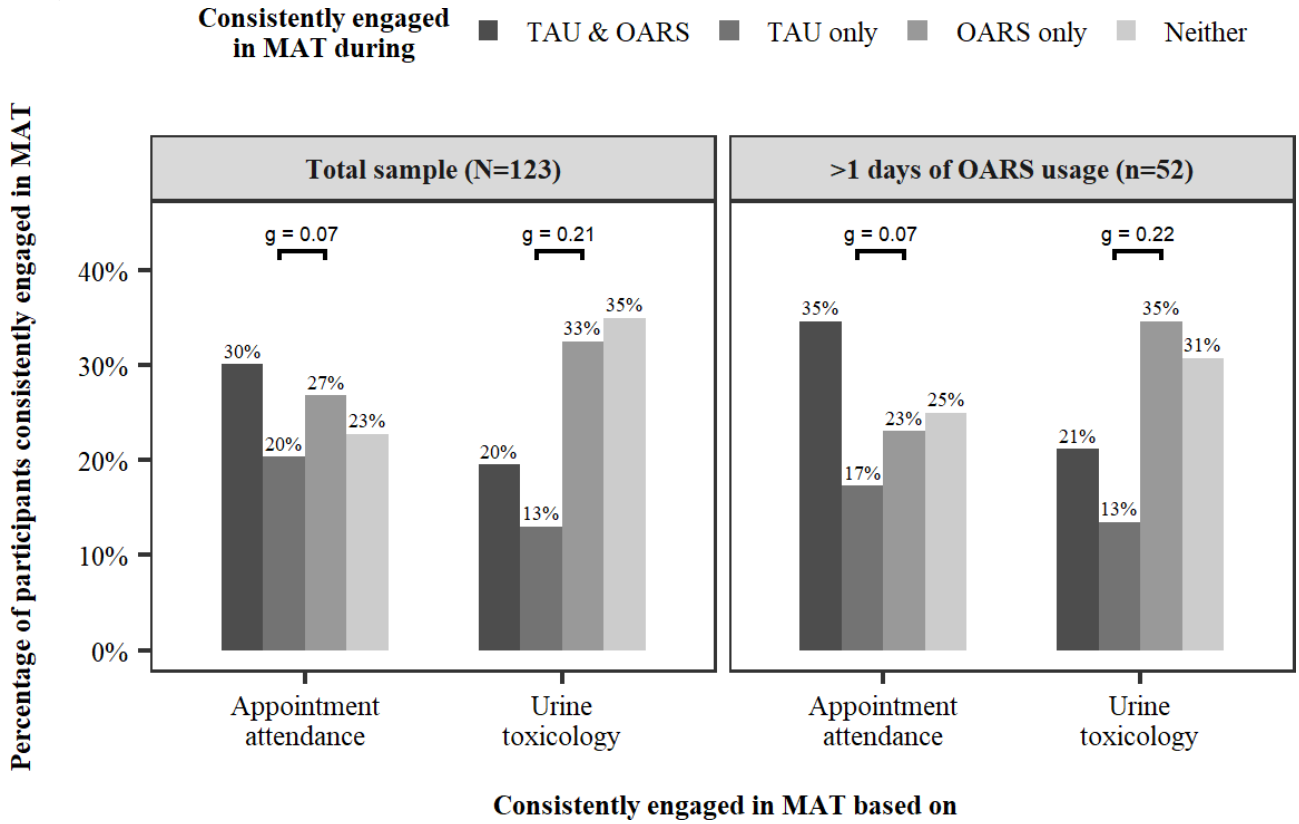

Discussion

Principal Findings

This study found that the introduction of the OARS platform within a primary care clinic may be associated with increased MAT engagement as measured by UTOX, but not with increased MAT engagement as measured by appointment attendance. Moderate effect sizes were observed for changes in UTOX, and the difference was statistically significant in the full sample. The difference in the effect of OARS on engagement as measured by UTOX compared to appointment attendance warrants further exploration. This study did not collect additional information that could provide insight into this variation. Nonetheless, UTOX is a valid method to determine treatment engagement in the primary care setting and is therefore an appropriate outcome to assess [24].

Further investigation is needed to determine whether the discordance between the statistical tests in the full sample and the subsample of patients who used OARS on more than 1 day represents a pertinent difference or reflects a data issue, such as bias induced by data availability (Figure 1) or insufficient statistical power in the subsample. In addition, a 2021 evaluation of the reSET-O app (Pear Therapeutics) found increased engagement in service use (ie, case management, behavioral and mental health services) and decreased rates of drug testing, and a 2024 randomized clinical trial found that adding the A-CHESS smartphone intervention was not associated with increased abstinence from opioids when included as part of an MAT program [20,21].

While some of the seemingly opposite results from this study and previous studies may be explained by different focuses of the mHealth interventions as well as different treatment program structures and patient populations, the ways in which mHealth interventions for OUD may increase engagement in particular aspects of MAT programs must be studied further. In particular, because the current mHealth interventions for OUD vary widely in their methodology (ie, focus on patient-provider communication vs incentives vs neurobehavioral therapy), future studies should compare the acceptability, feasibility, and efficacy of existing platforms to determine how to create mHealth interventions for OUD that consistently improve treatment outcomes for patients.

The findings from this study may be explained by its limitations. This study experienced multiple implementation challenges, including unexpected inability to fully integrate OARS with the clinic's EHR. This resulted in providers having to manually input information into OARS for patients instead of the data transferring automatically from the EHR, which in turn may have affected rates of usage by patients [22]. In addition, the majority of patients at this clinic were established maintenance patients who may have had low interest in and less need for a new recovery support tool. Furthermore, information was not available on patients who chose not to sign up for OARS, and it may be that more established patients who had high interest in care engagement were more likely to sign up. The sample size and limited background factors assessed in this study also made it unfeasible to account for confounding in analyses. Finally, the clinic was not able to provide EHR data for 82 participants. While measured demographic characteristics did not differ between these participants and those whose data were

XSL•FO
RenderX

available, it is possible that unmeasured differences between the included and excluded groups could still be influencing the findings of this study.

Conclusions

To our knowledge, this study is one of very few evaluations of mHealth platforms for the treatment of OUD in real-world settings that go beyond qualitative measures of acceptability and feasibility. While barriers to implementation and adoption may have impacted the ability to detect a strong effect of this intervention, this study provided the opportunity to introduce a first-of-its-kind app and provided valuable lessons for future implementation and platform development. The findings from this study reveal that OARS has the potential to be a promising contribution to the MAT treatment landscape, especially among patients who are new to MAT, and that mHealth should be further studied in the context of treatment for OUD. Future evaluations should consider ways to address implementation challenges as well as focus on patients who are in an earlier stage of treatment.

Acknowledgments

The authors wish to thank the staff and patients at the clinical site of this study who made this work possible. This work was supported by the National Institute on Drug Abuse (R42DA050398).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

SJ is the CEO of Q2i, the company that developed the software platform used in this study.

References

1. Drug overdose deaths: facts and figures. National Institute on Drug Abuse. 2023. URL: https://nida.nih.gov/research-topics/trends-statistics/overdose-death-rates [accessed 2025-07-22]
2. Lagisetty P, Klasa K, Bush C, Heisler M, Chopra V, Bohnert A. Primary care models for treating opioid use disorders: what actually works? A systematic review. PLoS ONE 2017;12(10):e0186315. [doi: 10.1371/journal.pone.0186315] [Medline: 29040331]
3. Sordo L, Barrio G, Bravo MJ, et al. Mortality risk during and after opioid substitution treatment: systematic review and meta-analysis of cohort studies. BMJ 2017 Apr 26;357:j1550. [doi: 10.1136/bmj.j1550] [Medline: 28446428]
4. Herring AA, Rosen AD, Samuels EA, et al. Emergency department access to buprenorphine for opioid use disorder. JAMA Netw Open 2024 Jan 2;7(1):e2353771. [doi: 10.1001/jamanetworkopen.2023.53771] [Medline: 38285444]
5. Samples H, Williams AR, Olfson M, Crystal S. Risk factors for discontinuation of buprenorphine treatment for opioid use disorders in a multi-state sample of Medicaid enrollees. J Subst Abuse Treat 2018 Dec;95:9-17. [doi: 10.1016/j.jsat.2018.09.001] [Medline: 30352671]
6. Villamil VI, Underwood N, Cremer LJ, Rooks-Peck CR, Jiang X, Guy GP. Barriers to retention in medications for opioid use disorder treatment in real-world practice. J Subst Use Addict Treat 2024 May;160:209310. [doi: 10.1016/j.josat.2024.209310] [Medline: 38331319]
7. Timko C, Schultz NR, Cucciare MA, Vittorio L, Garrison-Diehn C. Retention in medication-assisted treatment for opiate dependence: a systematic review. J Addict Dis 2016;35(1):22-35. [doi: 10.1080/10550887.2016.1100960] [Medline: 26467975]
8. Hall NY, Le L, Majmudar I, Mihalopoulos C. Barriers to accessing opioid substitution treatment for opioid use disorder: a systematic review from the client perspective. Drug Alcohol Depend 2021 Apr 1;221:108651. [doi: 10.1016/j.drugalcdep.2021.108651] [Medline: 33667783]
9. Ganesh R, Rao R, Deb KS, Bhad R, Yadav D. Digital capacity and interest in mHealth interventions among individuals on opioid agonist maintenance treatment: a cross-sectional community-based study. Indian J Psychol Med 2022 Jul;44(4):354-358. [doi: 10.1177/02537176211027239] [Medline: 35949629]
10. Free C, Phillips G, Galli L, et al. The effectiveness of mobile-health technology-based health behaviour change or disease management interventions for health care consumers: a systematic review. PLoS Med 2013;10(1):e1001362. [doi: 10.1371/journal.pmed.1001362] [Medline: 23349621]
11. mHealth: new horizons for health through mobile technologies. World Health Organization African Region. 2011. URL: https://www.afro.who.int/publications/mhealth-new-horizons-health-through-mobile-technologie [accessed 2025-07-22]
12. Nuamah J, Mehta R, Sasangohar F. Technologies for opioid use disorder management: mobile app search and scoping review. JMIR Mhealth Uhealth 2020 Jun 5;8(6):e15752. [doi: 10.2196/15752] [Medline: 32501273]
13. Cavazos-Rehg PA, Krauss MJ, Costello SJ, et al. Delivering information about medication assisted treatment to individuals who misuse opioids through a mobile app: a pilot study. J Public Health (Oxf) 2020 Feb 28;42(1):149-154. [doi: 10.1093/pubmed/fdy207] [Medline: 30445639]

XSL•FO
RenderX

14. Filiatreau LM, Szlyk H, Ramsey AT, et al. Sociodemographic differences in logins and engagement with the electronic health coach messaging feature of a mobile app to support opioid and stimulant use recovery: results from a 1-month observational study. JMIR Mhealth Uhealth 2025 Apr 10;13:e54753. [doi: 10.2196/54753] [Medline: 40210205]
15. Waselewski ME, Flickinger TE, Canan C, et al. A mobile health app to support patients receiving medication-assisted treatment for opioid use disorder: development and feasibility study. JMIR Form Res 2021 Feb 23;5(2):e24561. [doi: 10.2196/24561] [Medline: 33620324]
16. King VL, Siegel G, Priesmeyer HR, Siegel LH, Potter JS. Development and evaluation of a digital app for patient self-management of opioid use disorder: usability, acceptability, and utility study. JMIR Form Res 2024 Apr 1;8:e48068. [doi: 10.2196/48068] [Medline: 38557501]
17. Proctor SL, Rigg KK, Tien AY. Acceptability and usability of a reward-based mobile app for opioid treatment settings: mixed methods pilot study. JMIR Form Res 2022 Oct 5;6(10):e37474. [doi: 10.2196/37474] [Medline: 36197705]
18. Lyzwinski LN, Elgendi M, Menon C. Users' acceptability and perceived efficacy of mHealth for opioid use disorder: scoping review. JMIR Mhealth Uhealth 2024 Apr 11;12:e49751. [doi: 10.2196/49751] [Medline: 38602751]
19. Christensen DR, Landes RD, Jackson L, et al. Adding an internet-delivered treatment to an efficacious treatment package for opioid dependence. J Consult Clin Psychol 2014 Dec;82(6):964-972. [doi: 10.1037/a0037496] [Medline: 25090043]
20. Velez FF, Colman S, Kauffman L, Ruetsch C, Anastassopoulos K. Real-world reduction in healthcare resource utilization following treatment of opioid use disorder with reSET-O, a novel prescription digital therapeutic. Expert Rev Pharmacoecon Outcomes Res 2021 Feb;21(1):69-76. [doi: 10.1080/14737167.2021.1840357] [Medline: 33146558]
21. Gustafson DH, Landucci G, Vjorn OJ, et al. Effects of bundling medication for opioid use disorder with an mHealth intervention targeting addiction: a randomized clinical trial. Am J Psychiatry 2024 Feb 1;181(2):115-124. [doi: 10.1176/appi.ajp.20230055] [Medline: 37789744]
22. Nieto O, Rosen AD, Kalmin MM, et al. Adoption of a digital health tool for opioid use disorder treatment in primary care: facilitators and challenges (preprint). Journal of Medical Internet Research. Preprint posted online on May 2, 2025. [doi: 10.2196/preprints.69953]
23. Cohen J. Statistical Power Analysis for the Behavioral Sciences, 2nd edition: L. Erlbaum Associates; 1988.
24. Biondi BE, Zheng X, Frank CA, Petrakis I, Springer SA. A literature review examining primary outcomes of medication treatment studies for opioid use disorder: what outcome should be used to measure opioid treatment success? Am J Addict 2020 Jul;29(4):249-267. [doi: 10.1111/ajad.13051] [Medline: 32346932]

Abbreviations

EHR: electronic health record
MAT: medication-assisted treatment
mHealth: mobile health
OARS: Opioid Addiction Recovery Support
OUD: opioid use disorder
TAU: treatment as usual
UTOX: urine toxicology

Edited by A Mavragani; submitted 22.06.24; peer-reviewed by A Eisingerich, A King, R MacLean; revised version received 16.05.25; accepted 20.05.25; published 18.08.25.

Please cite as:
Rosen AD, Shoptaw SJ, Li L, Tulu B, Nieto O, Jenkins S, Kalmin MM
Clinical Management of Medication-Assisted Treatment for Opioid Use Disorder Using a Mobile Health App Within a Primary Care Clinic: Quasi-Experimental Study
JMIR Form Res 2025;9:e63526
URL: https://formative.jmir.org/2025/1/e63526
doi:10.2196/63526

© Allison D Rosen, Steven J Shoptaw, Li Li, Bengisu Tulu, Omar Nieto, Steven Jenkins, Mariah M Kalmin. Originally published in JMIR Formative Research (https://formative.jmir.org), 18.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Optimizing the Postcataract Patient Journey Using AI-Driven Teleconsultation: Prospective Case Study

Joukje C Wanten[1], MD; Noël J C Bauer[1], MD, PhD; Mohita Chowdhury[2], MSc; Aisling Higham[2], MBBS, MSc; Nick de Pennington[2], BM, BCh, MA, MBA; Frank J H M van den Biggelaar[1], PhD; Rudy M M A Nuijts[1], MD, PhD

[1]University Eye Clinic Maastricht, Maastricht University Medical Centre+, P Debyelaan 25, Maastricht, The Netherlands

[2]Ufonia Limited, Oxford, United Kingdom

Corresponding Author:
Joukje C Wanten, MD
University Eye Clinic Maastricht, Maastricht University Medical Centre+, P Debyelaan 25, Maastricht, The Netherlands

Abstract

Background: Given the increasing global demand for ophthalmologic care and the anticipated shortage of ophthalmology professionals, innovative solutions are essential for optimizing health care delivery. Digital health technologies offer promising opportunities to efficiently manage high patient volumes. Cataract surgery, with its established safety profile and routine postoperative care, provides an ideal setting for implementing such innovations. Structured clinical questions have proven effective in identifying patients requiring further assessment, supporting the feasibility of follow-up through telephone consultations. To further extend this approach, artificial intelligence–based follow-up systems may offer an opportunity to automate these interactions, reduce clinician workload, and streamline care pathways.

Objective: The aim of the study is to assess the clinical safety and effectiveness of an artificial intelligence–based follow-up call system (Dora-NL1) in identifying patients who require further assessment after cataract surgery in the Netherlands.

Methods: This prospective single-center study was conducted at the University Eye Clinic Maastricht, the Netherlands. Adult patients who underwent uncomplicated cataract surgery were eligible to participate. All patients received a Dora-NL1 follow-up telephone call at 1 and 4 weeks postoperatively in addition to standard care (a clinician-led telephone consultation at week 1 and an in-person hospital visit at week 4). The Dora-NL1 calls used a standard conversational flow to evaluate symptoms and recommend a clinical outcome. The recommended outcomes of Dora-NL1 were based on the symptoms reported by the patient. Clinical safety and accuracy were assessed by comparing Dora-NL1 outcomes to blinded clinician assessments of recorded calls and to standard postoperative care. Patient-reported usability was measured using the Telehealth Usability Questionnaire and Net Promoter Score.

Results: A total of 105 patients with a mean age of 72 (SD 7) years were included in the analysis. Dora-NL1 demonstrated high agreement with clinician-supervised calls, with symptom evaluation accuracy ranging from 89% to 99% (κ=0.390 - 0.947) and care management decision accuracy between 83% and 88% (κ=0.640 - 0.753). At week 1, Dora-NL1 showed a sensitivity of 100% and a specificity of 42% compared to standard clinician-led telephone consultations with no missed clinical concerns. At week 4, compared to the in-person follow-up, Dora-NL1 failed to identify 4 (4.1%) patients who required unexpected management changes, including 3 with asymptomatic complications detected only via slit lamp examination and 1 with complaints in the nonoperated eye. Patients rated Dora-NL1 positively, with Net Promoter Scores of +13.5 and +12.6 at week 1 and 4, respectively. The Telehealth Usability Questionnaire was completed by 98 patients, yielding a mean score of 3.19 (SD 1.13) on a 5-point scale, highlighting its simplicity, ease of use, and audibility.

Conclusions: Dora-NL1 is a safe and effective tool for automated postoperative screening following cataract surgery. It offers a safe alternative to clinician-led telephone consultations in routine cases but cannot fully replace in-person examinations.

(JMIR Form Res 2025;9:e72574) doi:10.2196/72574

KEYWORDS

algorithm-based screening; remote patient monitoring; postoperative follow-up; telemedicine; digital health; eHealth intervention; prospective clinical study

Introduction

Our health care systems worldwide face increasing challenges, driven in large part by demographic shifts such as population aging and the rising prevalence of chronic diseases. These developments contribute to a growing mismatch between demand for and supply of health care services. In ophthalmology, a field characterized by high patient volumes

XSL•FO
RenderX

and an aging patient population, it is expected that, by 2035, the workforce will be insufficient to meet the growing demand for care [1-3]. This trend is expected to intensify the burden on our health care systems. Moreover, the strain extends beyond ophthalmology, with similar challenges observed in most specialties including cardiology, urology, and neurology, underscoring the need for innovative, scalable solutions to maintain the quality and accessibility of care [2]. To mitigate this growing pressure, efficient management of high-volume workloads is essential. eHealth solutions offer promising opportunities to optimize health care delivery, enhance accessibility, and reduce the burden on clinical personnel. The COVID-19 pandemic further accelerated the adoption of such technologies, demonstrating their potential to ensure continuity of care and support remote follow-up [4,5]. In ophthalmology, advances in digital or remote health technologies have facilitated developments in triage, diagnosis, and patient monitoring [4-6]. Cataract care, in particular, is well suited to such innovations, as it is among the most frequently performed procedures worldwide and associated with a high safety profile and predictable recovery [7,8]. Previous studies have shown that structured sets of postoperative clinical questions can reliably identify patients who require further assessment within the first week after surgery. This has enabled safe and feasible alternatives to in-person follow-up during the early postoperative period, such as nurse-led telephone consultations [9,10]. Building on these developments, automated systems offer an additional opportunity to streamline care pathways. One such innovation is the voice-based automated conversational assistant, “Dora” (Ufonia Limited), which uses a machine learning–driven conversational model to conduct follow-up after cataract surgery via telephone. The system applies a standardized script of clinically validated questions to screen for symptoms that may indicate postoperative complications. Only patients who give responses requiring further attention are flagged for clinical review, potentially reducing the need for in-person or clinician-led consultations. Dora has already been validated for use in the United Kingdom, where it has demonstrated safety, feasibility, and high patient acceptability while also reducing the burden on outpatient services [11]. The English Dora system uses a discriminative dialogue mechanism that combines rule-based natural language processing with machine learning components. It uses a Dual Intent and Entity Transformer classifier for intent recognition [12]. This model interprets patient responses by mapping them to predefined “intents,” which in turn determine the next action in the dialogue flow. These “intents” were carefully designed to capture relevant aspects of postoperative symptomatology and are linked to a rule-based system that ensures that the complete sequence of questions is asked in a clinically appropriate order. Both the conversational flows and response content were developed in close collaboration with specialist clinicians and have been previously described in detail [11].

In this study, our aim is to validate and evaluate the clinical safety and effectiveness of the first Dutch version (Dora-NL1) in identifying patients who require further assessment after cataract surgery in the Netherlands. We hypothesize that the Dora-NL1 will be a safe and effective tool for early complication screening and a reliable alternative to human-led telephone consultation in routine cataract care.

Methods

Recruitment

Adult patients (≥21 years) who underwent uncomplicated cataract surgery on either 1 or both eyes, with implantation of a monofocal (including toric) intraocular lens (IOL) targeted for emmetropia, were recruited for this study at the University Eye Clinic of the Maastricht University Medical Center+. Patients were excluded if they had serious ocular comorbidities requiring face-to-face postoperative visits, underwent combined procedures (eg, glaucoma or retinal surgeries), needed follow-up visits within 1 week due to complications, were not able to speak Dutch, or had significant hearing or cognitive impairments.

Ethical Considerations

All patients were informed about the study and signed an informed consent before enrollment. This prospective validation study was approved by the local medical ethics committee (Medisch-ethische toetsingscommissie azM/UM, Maastricht, the Netherlands; approval METC 2022-3258) and executed in accordance with the Declaration of Helsinki. Patient data were anonymized to ensure confidentiality. Participants did not receive financial compensation for their involvement.

Dutch Model Development and Validation

The model for the autonomous cataract follow-up call system, Dora, was originally developed for English-speaking patients. In this study, the model was adapted for use in Dutch, with the initial Dutch system developed from the pretrained English version. To facilitate this adaptation, an intermediary translation function was integrated into the Dora-NL1 software, translating captured Dutch answers into English, formulating a response, and then converting it back to Dutch for speech output, as presented in Figure 1. While this is a crucial step, the translation process introduces complexity and the risk of altering the original message. The Dora model has a finite number of potential utterances; thus, it was possible to confirm that these were correctly translated into Dutch. In addition to the translation layer, the system was further adapted to improve its handling of Dutch-language input. Dutch utterances that were not recognized by the English Dora system were translated into English and manually reviewed to assign them to the appropriate intent category. These additions were incorporated into the training data to enhance the model’s ability to interpret Dutch responses. All translations and validation of the Dutch utterances were carried out by a bilingual clinician-researcher who was independent of the Dora development team. To ensure the reliability of the Dora-NL1 software, we ran a preliminary investigation before starting the study by looking at the agreement between Dutch (Dora-NL1) and English (Dora-R2) calls. In total, 12 different postoperative scenarios were created, reflecting both the common and most severe outcomes of patients who underwent cataract surgery. These scenarios were simulated by the bilingual clinician-researcher in both languages using a predefined strict script that underwent forward and

XSL•FO
RenderX

back-translation and independent expert review to ensure accuracy and consistency. The responses from both models were compared to assess their consistency, defined as acceptable when no statistically significant differences were found in symptom assessment and care management decisions. The questions and responses used for model validation were similar to those used in the trial version of the Dora-NL1. Furthermore, before starting the trial, we recruited volunteers to evaluate the different regional accents and variations in phrasing.

Figure 1. Schematic workflow of the Dora-NL1 system.

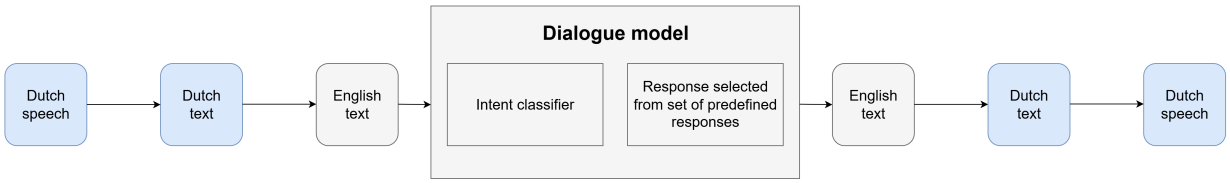

Postoperative Follow-Up

Patients undergoing bilateral cataract surgery underwent either immediate sequential bilateral or delayed sequential bilateral cataract surgery. For those undergoing unilateral surgery, the postoperative follow-up schedule aligned with that of immediate sequential bilateral cataract surgery. All patients received regular postoperative care in addition to the Dora-NL1 calls, as shown in Figure 2. One week after the first eye surgery, patients received both a Dora-NL1 call and a regular postoperative telephone consultation with a clinician. During the telephone consultations, a standardized set of questions was used to evaluate the patient's symptoms and identify any necessary management changes. Management changes included deviations from the standard eye drop tapering schedule, the addition of any drops (excluding artificial tears), the need for an additional hospital visit, or the implementation of further interventions (excluding suture removal). Clinicians conducting the telephone consultations were masked to the results of the Dora-NL1 calls. For patients receiving a toric IOL, a 1-week follow-up hospital visit was scheduled instead of a telephone consultation. At 4 weeks postoperatively, another Dora-NL1 call was conducted within 24 hours before the regular outpatient clinic visit. The final postoperative visit at the outpatient clinic consisted of a routine ophthalmological assessment, including evaluation of the refractive status and visual acuity as well as slit lamp and retinal examination.

Figure 2. Patient pathway. DSBCS: delayed sequential bilateral cataract surgery; ISBCS: immediate sequential bilateral cataract surgery; TUQ: Telehealth Usability Questionnaire.

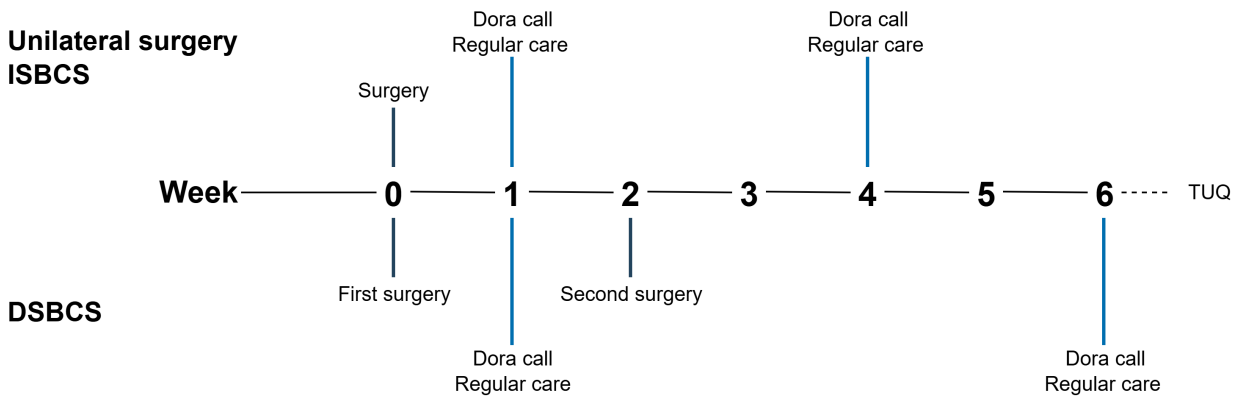

Dora-NL1 Call

The postoperative Dora-NL1 call followed a structured conversational flow: greeting and introduction, patient identification, symptom evaluation, patient queries, care management decision-making, user-acceptability question, and call closure. The key symptoms that were assessed included redness, pain, reduced vision, flashing lights, and floaters. At the core of this, Dora decision-making is an intent recognition system, a Dual Intent and Entity Transformer–based classifier that classifies the patient responses into predefined categories of intents (such as a "red eye" or "poor vision") [12-14]. These intent categories have been developed based on 2 prior clinical studies, and real-world use of Dora in the United Kingdom, with the input of clinicians and system developers [11,15]. For each patient utterance, the classifier assigns a confidence score reflecting the system's certainty in its interpretation. If the confidence score for a response falls below the preset threshold, Dora-NL1 prompts the patient to repeat or rephrase their response—allowing up to 2 attempts—to improve the accuracy of symptom detection.

Dora-NL1 classified each reported symptom as either absent, present without clinical significance, present with clinical significance, or inconclusive due to insufficient information.

XSL•FO
RenderX

The thresholds used to flag "potential concerns" were intentionally set to err on the side of caution. These were based on the framework established for the English version and were informed by clinical expertise and relevant literature on symptomatology, with a focus on sensitivity to minimize the risk of missing significant complications. If there was incomplete information regarding a symptom or if the patient asked a question that Dora-NL1 could not address, the overall call outcome defaulted to "potential clinical concerns."

Based on the overall symptom evaluation, Dora-NL1 provided an overall outcome of either "no clinical concerns" or "potential clinical concerns." The decision-making flowchart can be found in Multimedia Appendix 1. Depending on the overall call outcome, Dora-NL1 generated a corresponding care management recommendation: either "no review recommended" or "clinician review recommended." All Dora-NL1 calls were recorded and subsequently reviewed by a clinician, who independently assessed symptoms and made clinical decisions based on the audio recordings while masked to Dora-NL1's outputs and the regular care outcomes.

User Acceptability

At the end of each call, Dora asked patients to rate their likelihood of recommending the automated system to others on a scale of 1 to 10. The Net Promoter Score (NPS) was calculated by subtracting the "%-detractors" (scores of 1 to 6) from the "%-promoters" (scores of 9 or 10) [16]. Additionally, after the final postoperative visit, patients were invited to complete the Telehealth Usability Questionnaire (TUQ), translated to Dutch. The TUQ consists of 20 questions that can be answered on a 5-point scale ranging from 1=fully disagree to 5=fully agree [17].

Sample Size and Statistical Analysis

For this pilot study, a sample size of 100 participants who completed the 1-week call was used. Data analysis was conducted using SPSS (version 28.0; IBM Corp) and Microsoft Excel. Qualitative variables were summarized as frequencies and percentages, while descriptive statistics, including mean and SD, were calculated for quantitative variables. The agreement between the English and Dutch models in the developmental phase was tested using chi-square or Fisher exact tests. The interobserver reliability was calculated using the Cohen κ coefficient. The accuracy of the tool was defined in accuracy, specificity, and sensitivity, including 95% CIs of these proportions calculated using the Wilson score interval. A significance level of ≤.05 was applied.

Results

Preliminary Study Results

The agreement between the Dutch (Dora-NL1) and English (Dora-R2) versions used 12 scenarios. The accuracy in terms of detecting symptoms, including redness, pain, floaters, and flashing lights, was 100%. With respect to reduced vision, the accuracy was 83.3%, affected by 2 missing data points where Dora was unable to collect or understand the patient's response (1 from the English model and 1 from the Dutch model). The accuracy of the care management decisions was 91.7%, with the decrease attributed to these missing data points. There was no statistically significant difference found between the 2 language models (P=.32). Detailed responses for each scenario are provided in Multimedia Appendix 2.

Study Results

Between July 2023 and April 2024, a total of 182 patients who were scheduled for cataract surgery were screened for eligibility. Of these, 168 were suitable for participation, and 111 patients provided informed consent. In total, 105 patients (190 eyes) with a mean age of 72 (SD 7; range 50 - 84) years were included in the analysis; 100 patients completed the 1-week follow-up with Dora-NL1, and 98 patients completed the 4-week follow-up. The patient flowchart can be found in Figure 3. The baseline characteristics are summarized in Table 1.

XSL•FO
RenderX

Figure 3. Flowchart study. TUQ: Telehealth Usability Questionnaire.

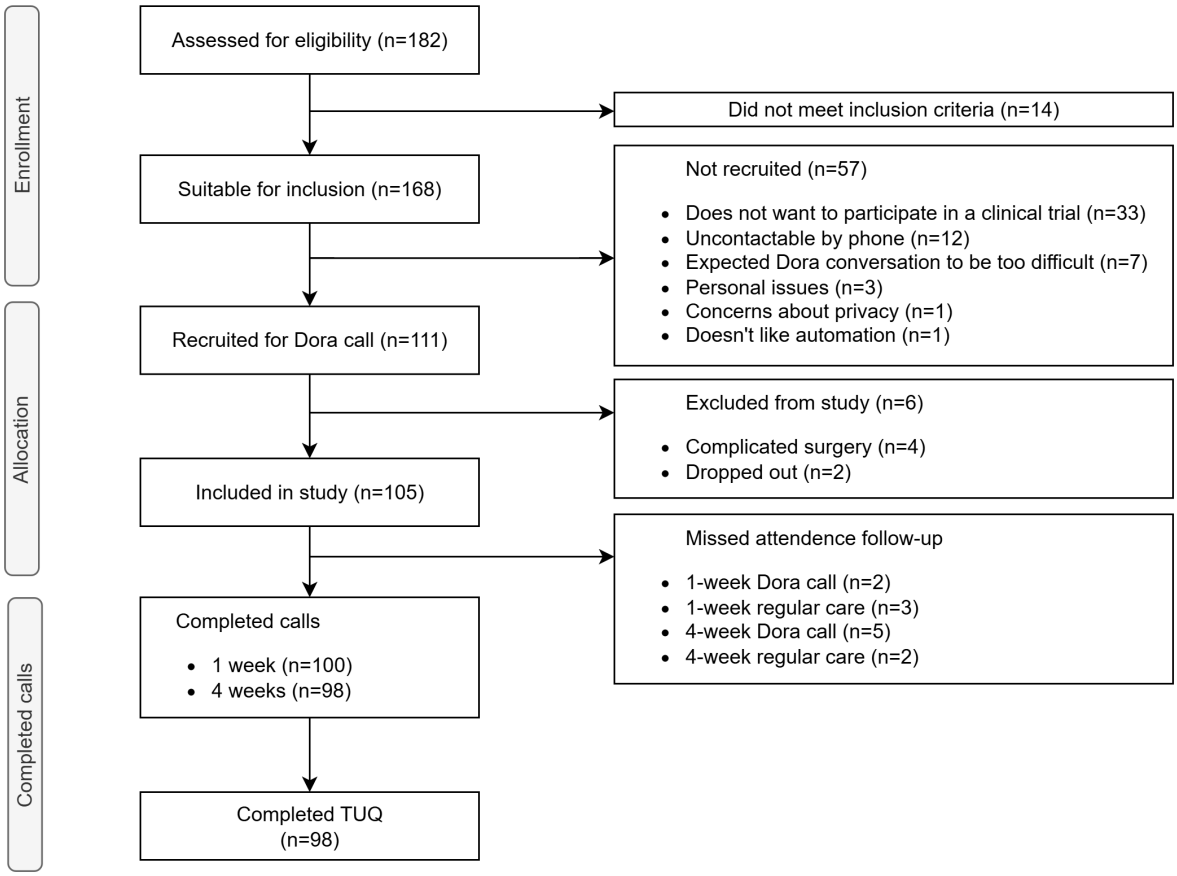

XSL•FO
RenderX

Table . Baseline characteristics.

Demographic variables	Participants (N=105), n (%)
Sex, n (%)	
Male	44 (41.9)
Female	61 (58.1)
Surgical risk factors, n (%)	
High myopia	3 (2.9)
High hyperopia	3 (2.9)
Glaucoma	2 (1.9)
Corneal pathology	7 (6.7)
Diabetic retinopathy	1 (1.0)
Previous vitrectomy	3 (2.9)
Type of surgery, n (%)	
Unilateral	20 (19.0)
DSBCS[a]	29 (27.6)
ISBCS[b]	56 (53.3)
IOL[c] characteristics (n=190)	
Type of IOL, n (%)	
Monofocal nontoric	169 (88.9)
Monofocal toric	21 (11.1)
IOL power (diopter), mean (SD)	20.0 (3.8)
Refractive target (diopter), mean (SD)	−0.03 (0.12)

[a]DSBCS: delayed sequential bilateral cataract surgery.
[b]ISBCS: immediate sequential bilateral cataract surgery.
[c]IOL: intraocular lens.

Postoperative Week 1

The agreement between the assessments made by Dora-NL1 at week 1 and those made by the clinician reviewing the Dora-NL1 call is shown in Table 2. The sensitivity ranged from 66.7% to 100%, and specificity ranged from 78.6% to 97.4% in identifying key symptoms and making care management decisions, with κ values ranging from 0.390 to 0.947. Only the detection of new floaters at week 1 showed a sensitivity of 66.67% and a specificity of 96.81%. Sensitivity was reduced because Dora failed to capture 2 patients' initial reports of floaters. In line with this protocol to repeat the question when the input is unclear or not captured by the system, Dora reasked the patients whether they had floaters. On the second attempt, both responded that they did not, resulting in an incorrect final classification. The agreement between the automated care management decisions made by Dora-NL1 and the regular care is shown in Table 3. At week 1, only patients who received regular telephone consultations were analyzed, excluding those with toric IOLs (n=13) who were scheduled for routine postoperative hospital visits. At this time, Dora-NL1 demonstrated 100% sensitivity with no false negatives and a specificity of 41.7%, yielding a false positive rate of 57.5%.

XSL•FO
RenderX

Table . Symptom and outcome accuracy of automated Dora versus supervised Dora call at week 1.

Decision (n=100)	Accuracy[a] (%) (95% CI)	Sensitivity (%) (95% CI)	Specificity (%) (95% CI)	κ (95% CI)	*P* value
Redness[b]	96.0 (90.12 to 98.4)	100 (56.6 to 100.0)	95.79 (89.7 to 98.4)	0.695 (0.416 to 0.973)	<.001
Pain[c]	97.0 (91.6 to 99.0)	100 (20.7 to 100.0)	96.97 (91.5 to 99.0)	0.390 (−0.149 to 0.929)	<.001
Vision issue[c]	94.0 (87.5 to 97.2)	100 (70.1 to 100.0)	93.41 (86.4 to 96.9)	0.718 (0.508 to 0.928)	<.001
New floaters	95.0 (88.8 to 97.9)	66.7 (30.0 to 90.3)	96.81 (91.0 to 98.9)	0.589 (0.262 to 0.916)	<.001
Flashing lights	98.0 (93.0 to 99.5)	100 (86.2 to 100.0)	97.37 (90.9 to 99.3)	0.947 (0.874 to 1.000)	<.001
Outcome	88.0 (80.2 to 93.0)	100 (92.0 to 100.0)	78.57 (66.2 to 87.3)	0.763 (0.641 to 0.885)	<.001

[a]Accuracy=(true positive+true negative)/(true positive+true negative+false positive+false negative).

[b]Amount of cases with “incomplete information”: n=4.

[c]Amount of cases with “incomplete information”: n=1.

Table . Accuracy in clinical decision-making between automated Dora versus regular telephone consultation at week 1[a].

Dora decision	Decision during the regular telephone consultation		Total, n (%)
	Potential concerns, n (%)	No clinical concerns, n (%)	
Potential concerns	1 (1.1)	50 (57.5)	51 58.6)
No clinical concerns	0 (0)	36 (41.4)	36 (41.4)
Total	1 (1.1)	86 (98.8)	87 (100)

[a]Sensitivity: 100% (95% CI 20.7-100.0), specificity: 41.9% (95% CI 32.0-52.4), κ: 0.016 (95% CI 0.00-0.398); *P*=.39.

Postoperative Week 4

Table 4 summarizes the agreement between the Dora-NL1 assessments at week 4 and those by the clinician reviewing the Dora-NL1 call, with a 100% sensitivity and specificity ranging from 88.3% to 99% for symptom evaluation. Despite this agreement, 3 asymptomatic patients who were identified as having “no clinical concerns” required an unexpected management change at the 4-week postoperative clinic visit, and 1 patient who had been asymptomatic of visual concerns at the week 1 Dora and clinician calls reported binocular diplopia. The 3 asymptomatic patients all had inflammatory cells detected during slit lamp examination, with 1 of these patients also having a residual lens fragment in the anterior chamber. They received topical corticosteroids and required additional follow-up visits. The fourth patient who reported binocular diplopia was evaluated by an orthoptist, who identified monocular diplopia in the contralateral eye due to postoperative changes after macular pucker peeling.

Table . Symptom and outcome accuracy of automated Dora versus supervised Dora call at week 4.

Decision (n=98)	Accuracy[a] (%) (95% CI)	Sensitivity (%) (95% CI)	Specificity (%) (95% CI)	κ (95% CI)	*P* value
Redness[b]	99.0 (94.4-99.8)	100 (34.2-100.0)	99.0 (94.3-99.8)	0.795 (0.403-1.000)	<.001
Pain[c]	98.0 (92.9-99.4)	100 (34.2-100.0)	97.9 (92.7-99.4)	0.657 (0.216-1.000)	<.001
Vision issue	96.9 (91.4-99.0)	100 (56.6-100.0)	96.8 (90.9-98.9)	0.754 (0.487-1.000)	<.001
New floaters[d]	88.8 (81.0-93.6)	100 (51.0-100.0)	88.3 (80.3-93.3)	0.823 (0.656-0.990)	<.001
Flashing lights	94.9 (88.6-97.8)	100 (77.2-100.0)	94.1 (87.0-97.5)	0.809 (0.648-0.970)	<.001
Outcome	82.7 (74.0-88.9)	93.6 (79.3-100.0)	77.6 (66.3-85.9)	0.640 (0.491-0.789)	<.001

[a]Accuracy = (true positive+true negative) /(true positive+true negative+false positive+false negative).

[b]Amount of cases with “incomplete information”: n=1.

[c]Amount of cases with “incomplete information”: n=3.

[d]Amount of cases with “incomplete information”: n=2.

Acceptability and Patient Experiences

Patients rated their likelihood of recommending the Dora-NL1 calls, with mean scores of 7.59 (SD 1.77) at week 1 and 7.38 (SD 2.35) at week 4, shown in Table 5. The corresponding NPSs for these calls were +13.5 and +12.6, respectively. The TUQ (n=98) had an overall mean score of 3.19 (SD 1.13). High scores were given for aspects such as time savings due to reduced

XSL•FO
RenderX

travel (mean 3.80, SD 1.11), simplicity (mean 3.89, SD 0.96), ease of use (mean 3.94, SD 0.88), and audibility (mean 3.97, SD 0.91). However, the user experience of Dora-NL1 was not deemed equivalent to an in-person consultation with a clinician, with a mean score of 2.46 (SD 1.07). Detailed questionnaire results are available in Multimedia Appendix 3.

Table . Net Promoter Scores of Dora-NL calls at 1 and 4 weeks postoperatively.

Net Promoter Score	Week 1 (n=96), n (%)	Week 4 (n=95), n (%)
1	1 (1)	3 (3.2)
2	1 (1)	3 (3.2)
3	1 (1)	3 (3.2)
4	4 (4.2)	3 (3.2)
5	5 (5.2)	6 (6.3)
6	6 (6.3)	7 (7.4)
7	17 (17.7)	13 (13.7)
8	30 (31.3)	20 (21.1)
9	23 (24)	22 (23.2)
10	8 (8.3)	15 (15.8)

Discussion

Principal Findings

The purpose of this study was to evaluate the clinical safety and effectiveness of the Dutch Dora-NL1 model in identifying patients who require further assessment following uncomplicated cataract surgery. Our findings demonstrated high overall accuracy of Dora-NL1's symptom evaluation and care management decision-making at both 1 and 4 weeks after surgery when compared to the clinician-supervised Dora-NL1 calls. At week 1, Dora-NL1 exhibited a perfect sensitivity of 100% for care management decision-making when compared to a regular telephone consultation. However, at week 4, clinical issues in 4 asymptomatic patients were identified during face-to-face consultations, which were not detected by Dora or by the clinician reviewing the calls. In terms of user acceptability, most patients expressed a positive attitude toward the tool, highlighting its simplicity, ease of use, and audibility.

The Dora-NL1 model was designed to err on the side of caution, thereby ensuring that patients classified as having "no concerns" by Dora are unlikely to have underlying clinical issues. Additionally, in cases of "incomplete information," patients were assumed to have potential concerns, which may have increased the rate of false positives and, with that, lowered the specificity. The sensitivity and specificity of the automated Dora-NL1 outcomes compared to the supervised Dora-NL1 calls were high both at 1 and 4 weeks. Only the sensitivity for floaters at week 1 was reduced, as Dora misunderstood 2 patients. In addition, the κ agreement for the symptom "pain" at week 1 was low due to the low prevalence of pain symptoms in this cohort (n=1). When comparing Dora-NL1 to regular care at week 1, no clinical issues were missed by Dora-NL1, resulting in 100% sensitivity, while the specificity was 41.86% due to the high rate of false positives.

Comparison With Prior Work

Previous studies have demonstrated that telephone follow-ups are a safe method for identifying significant issues after cataract surgery [9,18,19]. In this study, 3 patients had inflammatory cells, which required a management change during their 4-week hospital visit. These complications were detected during slit lamp examination, highlighting a limitation of telephone consultations, which may fail to identify such issues in asymptomatic patients. Postoperative inflammatory cells usually peak at day 1 after surgery and may take up to 3 months to resolve [20]. Studies have demonstrated a correlation between prolonged intraocular inflammation and an increased likelihood of developing cystoid macular edema (CME), suggesting a threshold beyond which extended corticosteroid treatment is advisable [21]. However, randomized controlled trials comparing corticosteroid treatments versus placebo for inflammation after cataract surgery have shown that only 2.7% to 3.8% of patients in the placebo groups developed CME within 3 months after surgery versus 1.4% to 3.8% in the corticosteroid-treated groups [22,23]. These findings pose the question of clinical relevance of the presence of inflammatory cells in asymptomatic patients who have no symptoms of red eye or ocular pain or decreased visual acuity due to clinically significant CME. Additionally, another patient in our cohort who required a management change at week 4 erroneously reported experiencing binocular diplopia while there appeared to be monocular diplopia in the contralateral eye. Notably, this symptom was also missed by the clinician-led call at week 1. The cases missed by Dora-NL1 at 4 weeks postoperatively highlight the limitations of this tool in detecting subclinical or asymptomatic complications. While some complications, such as mild inflammation, may resolve without intervention, others, such as retained lens material, could lead to more serious consequences, including chronic inflammation and secondary glaucoma, or corneal decompensation. It should, however, be noted that the Dora call provides safety net advice for the patient to seek review with new or worsening symptoms. Hence, this

XSL•FO
RenderX

tool should be viewed as a complement to, rather than a replacement for, in-person follow-up visits.

The English Dora-R1 model has been investigated in the United Kingdom, showing it to be a safe alternative for cataract surgery follow-up, with automated calls conducted 3‑4 weeks postoperatively and reviewed by an ophthalmologist [11,15]. The sensitivity of care management decision-making was 93.75% (95% CI 84.76-98.27), and the specificity was 86.26% (95% CI 79.16-91.56), with an accuracy of 88.72% (95% CI 93.42-92.79) [11]. These results are consistent with our findings in Tables 2 and 4. In the UK cohort, 10 patients who had been discharged following a Dora-R1 call—based on both Dora's and the supervising clinician's decision—experienced unexpected management changes at their 4-week routine hospital visit. These patients were either asymptomatic or exhibited mild symptoms that were not reported during the Dora-R1 call. Among them, 3 were diagnosed with CME, 5 presented with inflammatory cells, 1 experienced a refractive surprise, and 1 developed early posterior capsular opacification [11]. These findings are consistent with our results, as the majority of cases not detected via the telephone call were in asymptomatic patients with inflammatory cells in the anterior chamber.

The NPS for Dora-R1 ranged from +45 to +51, with mean scores of 8.55 (SD 1.67) and 8.59 (SD 2.05), respectively. The overall TUQ mean score was 3.96 (SD 0.47; 95% CI 3.76-4.16) [11,15]. The mean NPS at 3 weeks for the Dora-R1 was compared with that at 4 weeks postoperatively, yielding a statistically significant mean difference of −1.21 (SD 0.27; $P<.0001$). Compared to UK studies, both NPS and TUQ scores in our Dutch cohort were lower. Several factors may account for this difference. First, the study design differed: while the UK participants received either only a Dora-R1 call or a Dora-R1 call along with a hospital visit at week 4, all Dutch participants received automated Dora-NL1 calls in addition to standard postoperative care, including telephone consultations and hospital visits. Consequently, patient expectations may have been shaped by direct comparisons between Dora and conventional care, potentially leading to more critical evaluations. Additionally, the Dora-NL1 model was adapted from the pretrained English version but relied on more limited Dutch natural language processing resources at the time of development. This may have affected the performance of both the text-to-speech and speech-to-text functions, resulting in less fluent and less natural user interactions [24]. Furthermore, cultural attitudes toward artificial intelligence (AI) and digital communication in health care may have contributed. Dutch patients may place a higher value on personal interaction and may be more reserved in expressing satisfaction with automated systems, particularly when these are perceived as less empathetic or less competent communicators. Future cross-cultural studies using harmonized protocols are needed to clarify the extent to which these differences are attributable to cultural, technical, or methodological factors.

Limitations

Limitations of this study mainly included the potential selection bias. There were eligible patients who decided not to participate due to the expected difficulties of the Dora-NL1 call or who strongly opposed eHealth solutions in general. This population may not represent the broader cataract population, limiting generalizability. In addition, this is a mono-center study, which reduces the applicability to other health care facilities in the Netherlands, as there is a high variability in practice patterns of the postoperative cataract care pathway. Additionally, as all participants received both the Dora-NL1 and standard follow-up (clinician phone calls and in-person visits), the study is subject to performance bias. The awareness of being monitored by multiple systems may have influenced patient behavior, which may have influenced the outcomes.

Moreover, symptom-based systems like Dora-NL1 cannot detect structural or subclinical complications, which may require in-person assessment. Therefore, this tool is suitable as a complementary screening tool, replacing regular telephonic consultations for complication screening, but not as a substitute for in-person postoperative assessments.

Future Implementation

Our findings indicated that at the 1-week follow-up, the Dora-NL1 model accurately identified 41.4% (n=36) of patients as having "no clinical concerns." This suggests that approximately 4 of 10 patients could avoid clinician callbacks, potentially enhancing health care efficiency and reducing costs. Telephone reviews have been shown to be a safe time- and cost-saving alternative to short-term hospital visits after routine surgeries, with high patient satisfaction [18,19,25]. Moreover, automated calls allow clinicians to save time and focus on other care tasks while ensuring a standardized approach to symptom evaluation and decision-making. This consistency reduces variability, improves the quality of patient care, and facilitates the collection and comparison of patient outcomes. In the UK cohort, the average staff cost savings amount to US $47.78 per patient compared to regular care [11]. Further research is necessary to evaluate the potential cost-effectiveness in the Netherlands.

To implement automated calls in routine postoperative cataract care, several practical issues must be considered, particularly the significant variability in postoperative care across clinics and countries [26]. In some clinics, patients may not receive follow-up visits, as research indicates that for those undergoing uncomplicated cataract surgery without pre-existing ocular comorbidities, follow-up can be safely deferred to 2 to 4 weeks after surgery without significant differences in visual outcomes [27,28]. However, automated phone calls for interim check-ups could enhance patients' confidence and reassurance during their recovery without straining health care resources. Despite this potential benefit, the current Dutch cataract guidelines do not yet recommend remote care for the final 4-week postoperative follow-up due to the limited evidence regarding its safety and the necessity of visual acuity and refraction assessments for the mandatory quality registry of cataract surgeries in the Netherlands [29]. Further development of the Dora-NL1 model is expected to improve its accuracy and expand its applicability.

Conclusions

Based on the results of this study, we can conclude that the Dutch AI-based cataract follow-up call system, Dora-NL1, is

XSL•FO
RenderX

a safe alternative for telephone consultations in the postoperative cataract care pathway and is suitable as a screening tool for postoperative complications during the follow-up of cataract surgery. However, as this tool only captures subjective information, it is not able to fully replace final 4-week in-person hospital visits after cataract surgery in the Netherlands. Future research is necessary to investigate the practical implications and cost-effectiveness of this AI-driven teleconsultation system.

Acknowledgments

This study was funded by a research grant from UitZicht and the Bayer Ophthalmic Care Award (BOCA). UitZicht is a cooperation of several Dutch funds for people with visual impairment to fund scientific research into eye diseases. The BOCA is an award for health care professionals and researchers in ophthalmology for health care innovations.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

All authors contributed to the conceptualization and methodology of the study. Data curation, formal analysis, investigation, and project administration were carried out by JCW. Supervision was provided by NJCB and RMMAN. JCW was responsible for visualization and drafting the original manuscript. The manuscript was critically reviewed and edited by NJCB, AH, MC, FJHMB, NP, and RMMAN. All authors read and approved the final version of the manuscript.

Conflicts of Interest

JCW has no financial disclosures. NJCB is a consultant for Alcon, has received lecture fees from Teleon Surgical BV, and has received research support of Teleon Surgical BV and Alcon. AH and MC are employees at Ufonia Limited. NP is the director and shareholder of Ufonia Limited. FJHMB has no financial disclosures. RMMAN is a consultant for Alcon, Johnson&Johnson, and Carl Zeiss Meditec, has received lecture fees from Alcon and Ophtec, and has received research support from Alcon, Teleon Surgical BV, and Carl Zeiss Meditec.

Multimedia Appendix 1
Dora decision algorithm and symptom classification.
[PDF File, 63 KB - formative_v9i1e72574_app1.pdf]

Multimedia Appendix 2
Agreement between Dutch and English models across different patient scenarios.
[DOCX File, 14 KB - formative_v9i1e72574_app2.docx]

Multimedia Appendix 3
Outcomes of Telehealth Usability Questionnaire.
[DOCX File, 13 KB - formative_v9i1e72574_app3.docx]

References

1. Berkowitz ST, Finn AP, Parikh R, Kuriyan AE, Patel S. Ophthalmology Workforce Projections in the United States, 2020 to 2035. Ophthalmology 2024 Feb;131(2):133-139. [doi: 10.1016/j.ophtha.2023.09.018] [Medline: 37739231]
2. Dall TM, Gallo PD, Chakrabarti R, West T, Semilla AP, Storm MV. An aging population and growing disease burden will require a large and specialized health care workforce by 2025. Health Aff (Millwood) 2013 Nov;32(11):2013-2020. [doi: 10.1377/hlthaff.2013.0714] [Medline: 24191094]
3. Murphy A, McElnea E, Byrne S. Health technology assessment: a primer for ophthalmology. Eur J Ophthalmol 2018 Jul;28(4):358-364. [doi: 10.1177/1120672117747040] [Medline: 29973070]
4. Li JPO, Liu H, Ting DSJ, et al. Digital technology, tele-medicine and artificial intelligence in ophthalmology: a global perspective. Prog Retin Eye Res 2021 May;82(100900):100900. [doi: 10.1016/j.preteyeres.2020.100900] [Medline: 32898686]
5. Tognetto D, Brézin AP, Cummings AB, et al. Rethinking elective cataract surgery diagnostics, assessments, and tools after the COVID-19 pandemic experience and beyond: insights from the EUROCOVCAT Group. Diagnostics (Basel) ;10(12):1035. [doi: 10.3390/diagnostics10121035]
6. Tognetto D, Giglio R, Vinciguerra AL, et al. Artificial intelligence applications and cataract management: a systematic review. Surv Ophthalmol 2022;67(3):817-829. [doi: 10.1016/j.survophthal.2021.09.004] [Medline: 34606818]

XSL•FO
RenderX

7. Lindegger DJ, Wawrzynski J, Saleh GM. Evolution and applications of artificial intelligence to cataract surgery. Ophthalmol Sci 2022 Sep;2(3):100164. [doi: 10.1016/j.xops.2022.100164] [Medline: 36245750]
8. Stein JD. Serious adverse events after cataract surgery. Curr Opin Ophthalmol 2012 May;23(3):219-225. [doi: 10.1097/ICU.0b013e3283524068] [Medline: 22450221]
9. Lim JS, Tan SS, Yeo YL, et al. Replacing the postoperative week 1 visit after routine phacoemulsification with a telephone consult. Can J Ophthalmol 2024 Dec;59(6):386-393. [doi: 10.1016/j.jcjo.2024.03.015] [Medline: 38604239]
10. Moustafa GA, Borkar DS, Borboli-Gerogiannis S, et al. Optimization of cataract surgery follow-up: a standard set of questions can predict unexpected management changes at postoperative week one. PLoS One 2019;14(9):e0221243. [doi: 10.1371/journal.pone.0221243] [Medline: 31536500]
11. Meinert E, Milne-Ives M, Lim E, et al. Accuracy and safety of an autonomous artificial intelligence clinical assistant conducting telemedicine follow-up assessment for cataract surgery. EClinicalMedicine 2024 Jul;73:102692. [doi: 10.1016/j.eclinm.2024.102692] [Medline: 39050586]
12. Bunk TV, Varshneya D, Vlasov V, Nichol A. DIET: lightweight language understanding for dialogue systems. arXiv. Preprint posted online on May 11, 2020. [doi: 10.48550/arXiv.2004.09936]
13. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv. Preprint posted online on May 24, 2019. [doi: 10.48550/arXiv.1810.04805]
14. Gardiner OW, Chowdhury M, Lim E, Higham A, Pennington N. Can deep learning models understand natural language descriptions of patient symptoms following cataract surgery. Invest Ophthalmol Vis Sci 2022;63(7):F0012-F1694.
15. Khavandi S, Lim E, Higham A, et al. User-acceptability of an automated telephone call for post-operative follow-up after uncomplicated cataract surgery. Eye (Lond) 2023 Jul;37(10):2069-2076. [doi: 10.1038/s41433-022-02289-8] [Medline: 36274084]
16. Krol MW, de Boer D, Delnoij DM, Rademakers JJDJM. The Net Promoter Score—an asset to patient experience surveys? Health Expect 2015 Dec;18(6):3099-3109. [doi: 10.1111/hex.12297] [Medline: 25345554]
17. Parmanto B, Lewis AN, Graham KM, Bertolet MH. Development of the Telehealth Usability Questionnaire (TUQ). Int J Telerehabil 2016;8(1):3-10. [doi: 10.5195/ijt.2016.6196] [Medline: 27563386]
18. Al-Ani HH, Li S, Niederer RL. Telephone follow-up one day post-cataract surgery. Clin Exp Optom 2023 Sep;106(7):741-745. [doi: 10.1080/08164622.2022.2146482] [Medline: 36464321]
19. Ruiss M, Pai V, Pilwachs C, Bayer N, Palkovits S, Findl O. Quality assurance via telephone interviews after cataract surgery: an explorative study. PLoS ONE 2024;19(3):e0298149. [doi: 10.1371/journal.pone.0298149] [Medline: 38451971]
20. De Maria M, Iannetta D, Cimino L, Coassin M, Fontana L. Measuring anterior chamber inflammation after cataract surgery: a review of the literature focusing on the correlation with cystoid macular edema. Clin Ophthalmol 2020;14(41-52):41-52. [doi: 10.2147/OPTH.S237405] [Medline: 32021067]
21. Aaronson A, Taipale C, Achiron A, Aaltonen V, Grzybowski A, Tuuminen R. Relationship between prolonged intraocular inflammation and macular edema after cataract surgery. Transl Vis Sci Technol 2021 Jun 1;10(7):15. [doi: 10.1167/tvst.10.7.15] [Medline: 34125145]
22. Tyson SL, Bafna S, Gira JP, et al. Multicenter randomized phase 3 study of a sustained-release intracanalicular dexamethasone insert for treatment of ocular inflammation and pain after cataract surgery. J Cataract Refract Surg 2019 Feb;45(2):204-212. [doi: 10.1016/j.jcrs.2018.09.023] [Medline: 30367938]
23. Donnenfeld E, Holland E. Dexamethasone intracameral drug-delivery suspension for inflammation associated with cataract surgery: a randomized, placebo-controlled, phase III trial. Ophthalmology 2018 Jun;125(6):799-806. [doi: 10.1016/j.ophtha.2017.12.029] [Medline: 29397189]
24. Chowdhury M, Lim E, Higham A, et al. Can large language models safely address patient questions following cataract surgery. Presented at: The Proceedings of the 5th Clinical Natural Language Processing Workshop (ClinicalNLP 2023); Jul 14, 2023; Toronto, Canada p. 131-137.
25. Soegaard Ballester JM, Scott MF, Owei L, Neylan C, Hanson CW, Morris JB. Patient preference for time-saving telehealth postoperative visits after routine surgery in an urban setting. Surgery 2018 Apr;163(4):672-679. [doi: 10.1016/j.surg.2017.08.015] [Medline: 29398042]
26. Rossi T, Romano MR, Iannetta D, et al. Cataract surgery practice patterns worldwide: a survey. BMJ Open Ophthalmol 2021;6(1):e000464. [doi: 10.1136/bmjophth-2020-000464] [Medline: 33501377]
27. Deshpande R, Vora U, Mangiraj V, Dole K, Deshpande M. Can the postoperative follow-up visit be deferred up to four weeks after an uneventful cataract surgery?—A randomized controlled trial. Indian J Ophthalmol 2021 Jun;69(6):1409-1413. [doi: 10.4103/ijo.IJO_2390_20] [Medline: 34011710]
28. Kessel L, Andresen J, Erngaard D, Flesner P, Tendal B, Hjortdal J. Safety of deferring review after uneventful cataract surgery until 2 weeks postoperatively. J Cataract Refract Surg 2015 Dec;41(12):2755-2764. [doi: 10.1016/j.jcrs.2015.11.010] [Medline: 26796457]
29. Richtlijn cataract—zorg op afstand bij cataract [Article in Dutch]. Nederlands Oogheelkundig Gezelschap (NOG). 2023. URL: https://richtlijnendatabase.nl/richtlijn/cataract/startpagina_-_cataract.html [accessed 2025-07-24]

XSL•FO
RenderX

Abbreviations

AI: artificial intelligence
CME: cystoid macular edema
IOL: intraocular lens
NPS: Net Promoter Score
TUQ: Telehealth Usability Questionnaire

Edited by J Sarvestan; submitted 24.02.25; peer-reviewed by G Hadimani, TK Yoo; revised version received 28.05.25; accepted 04.06.25; published 18.08.25.

Please cite as:
Wanten JC, Bauer NJC, Chowdhury M, Higham A, de Pennington N, van den Biggelaar FJHM, Nuijts RMMA
Optimizing the Postcataract Patient Journey Using AI-Driven Teleconsultation: Prospective Case Study
JMIR Form Res 2025;9:e72574
URL: https://formative.jmir.org/2025/1/e72574
doi:10.2196/72574

© Joukje C Wanten, Noël J C Bauer, Mohita Chowdhury, Aisling Higham, Nick de Pennington, Frank J H M van den Biggelaar, Rudy M M A Nuijts. Originally published in JMIR Formative Research (https://formative.jmir.org), 18.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

AI-Enhanced 3D Models in Global Virtual Reality Case Conferences for Surgical Care in a Low-Income Country: Exploratory Study

Miriam Obst[1,2*]; Jan Arensmeyer[2,3*], Dr med; Henrik Bonsmann[2], MSc; Andreas Kolbinger[4], MD; Joel Kigenyi[5], MA; Francis Oneka[5], MBchB; Benard Owere[5], MBchB; Joachim Schmidt[2,3,6], Prof Dr Med; Philipp Feodorovici[2,3*], MD; Jan Wynands[2,7*], Dr med

[1]Faculty of Medicine, University of Bonn, Bonn, Germany
[2]Bonn Surgical Technology Centre (BOSTER), University Hospital Bonn, Joseph-Schumpert-Allee 1, Bonn, Germany
[3]Division of Thoracic Surgery, Department of Surgery, University Hospital Bonn, Bonn, Germany
[4]Medical Practice of Reconstructive and Aesthetic Surgery Pfaffenhofen, Pfaffenhofen, Germany
[5]Lamu Medical Centre for Reconstructive and Global Surgery, Jinja, Uganda
[6]Department of Thoracic Surgery, Helios Hsopital Bonn/Rhein-Sieg, Bonn, Germany
[7]Section Global Health, Institute of Hygiene and Public Health, University Hospital Bonn, Bonn, Germany
[*]these authors contributed equally

Corresponding Author:
Jan Arensmeyer, Dr med
Bonn Surgical Technology Centre (BOSTER), University Hospital Bonn, Joseph-Schumpert-Allee 1, Bonn, Germany

Abstract

Background: Approximately 5 billion people worldwide lack adequate access to surgical care, primarily in the Global South. Especially in crisis regions and war zones, telemedical applications may enhance health services. This study explores the feasibility of using artificial intelligence (AI)-enhanced 3D imaging and extended reality (XR) technologies for intercontinental surgical case conferences in a low-resource scenario in Uganda. Our pilot study aims to assess the value of these technologies to address the lack of surgical resources and multilateral knowledge exchange.

Objective: This study intends to determine the feasibility of using new AI-enhanced image modeling technology within an immersive spatial XR scenario to collaboratively and remotely assess reconstructive patient cases in the resource-limited country of Uganda.

Methods: Within a surgical camp at Lamu Medical Centre, Uganda, 3D models of patients' conditions were created using a smartphone app. Digital models were generated from photographs taken on-site and processed into 3D formats to be visualized in virtual case conferences. Here, surgeons from Uganda and Germany used virtual reality (VR) headsets to collaboratively discuss case strategies while marking surgical approaches on each digital patient model.

Results: The study included 15 patients requiring reconstructive surgery, with a diverse range of conditions. The use of XR technology facilitated detailed visualization and discussion of surgical strategies. The process was time-efficient, with a total of under 8 minutes per case for data acquisition and model creation, and resource-efficient with surgeons reporting sufficient quality of smartphone-derived models. Valuable user experience and precise interaction during the VR case processing were found, underlining its potential to improve surgical planning and patient care in resource-limited settings.

Conclusions: The findings indicate that AI-enhanced 3D imaging and immersive virtual communication platforms are valuable tools for integrative surgical case assessments. The cost-effectiveness of the used consumer solutions should be especially beneficial for low-resource environments. While the study demonstrates the feasibility of this approach, further research is needed to explore a broader application and impact of these technologies in global health. The study highlights the potential of XR to enhance training and surgical precision, contributing to better health care outcomes in underserved regions.

(JMIR Form Res 2025;9:e69300) doi:10.2196/69300

KEYWORDS

3D scanning; artificial intelligence; virtual reality; extended reality; metaverse; spatial computing; global surgery; reconstructive surgery

XSL•FO
RenderX

Introduction

Background

A modeling study has demonstrated that approximately 5 billion people worldwide lack adequate access to surgical care [1]. This issue particularly affects regions in the Global South, including Sub-Saharan Africa, South America, and Asia. To target this issue, "Global Surgery" has been initiated and prioritized in the past decade by the World Health Assembly as a concept to improve and equalize surgical care across international health systems [2]. The Lancet Commission of Global Surgery particularly emphasizes the role of innovation and mobile health to address the accessibility of surgical expertise [3]. Rural surgeons in low-income countries, who must provide broad surgical skills, play a critical role in advancing Global Surgery [4]. In crisis regions and war zones, remote medical assistance, such as telemedical consulting, can enhance health service delivery, as currently demonstrated in the Gaza Strip [5].

Limited funding for medical technology is a key burden holding back the evolution of treatment practice as well as medical education in low-income countries [6]. In high-income countries, an increasing variety of high-end technologies in health care systems is available, whereas limited resources in low-income countries could especially necessitate frugal designs [7]. Uganda is an East African country with a population of 45.5 million people [8]. The World Bank classifies Uganda as a low-income country per its income classification [9]. The country has a severe shortage of qualified surgeons, with fewer than one per 100,000 people, exacerbating challenges such as geographic isolation, low health literacy, limited financial resources, and insufficient health care infrastructure [10-13]. In many urgent cases, for example, in frequent road traffic accidents, acid attacks, or fire burns, patients cannot receive the necessary immediate treatment. Individuals surviving the initial trauma later often become affected by preventable contractures and immobility, eventually necessitating plastic and reconstructive surgery [14]. A recent review analyzes the Ugandan health care system and emphasizes the need for investments and enhancements in both new and existing infrastructure as well as in the further education of health care workers [15].

In reconstructive surgery, a 3D understanding of individual pathologies located in surface anatomy is pivotal. This applies to treatment approaches for both traumatically induced injuries, such as burns, accident- or war-related wounds, as well as benign space-occupying lesions or oncological entities.

Modern software for creating 3D models from 2D images, such as photogrammetry, has made significant strides in recent years. This technique analyzes multiple 2D images taken from various angles to generate accurate 3D models. AI-driven 3D reconstruction technologies further improve model creation, enhancing both speed and quality [16,17]. These technologies allow high-quality 3D model creation even from frugal sources like smartphone cameras.

The application of extended reality (XR), which includes virtual reality (VR), augmented reality (AR), and mixed reality (MR), is expanding rapidly in the medical field [18,19]. However, access to these technologies is not equally distributed from a global perspective. In surgery, the technology allows preoperative planning using patient-specific imaging in 3D environments and the training of surgical procedures through immersive simulations [20-22]. Where available, XR has a positive impact on surgical training [23-25]. Computed tomographic (CT) and magnetic resonance imaging (MRI) scans can be reconstructed into a 3D image and displayed with the use of head-mounted displays (HMDs), commonly known as VR headsets. Novel explorations even demonstrate intraoperative holographic overlays of 3D reconstructed images [26]. For the user, visualizing the patient data in a 3D stereoscopic view reduces the cognitive work needed to assess the 3D information out of 2D scan slices and similar imaging technologies [27-29]. The use of 3D-reconstructed medical imaging has been analyzed in several studies, often in the context of addressing challenges in surgical strategies for complex oncological scenarios [26,30-32].

Therefore, in lower-income regions without access to radiological imaging, photogrammetric modeling could be an additional tool for patient data acquisition. Besides novel possibilities of 3D image representation, the assembly of health care professionals in immersive virtual spaces from different locations opens new approaches in future medical conferencing [33].

The employment of XR and AI-supported smartphone apps facilitates surface visualization in a manner that is uncomplicated, expeditious, cost-effective, and radiation-free, thereby offering a distinct advantage over radiological imaging methods. Our study emphasizes frugal and robust technological solutions by using widely available, consumer-grade hardware that is well-suited to low-resource and rural settings, where access to surgical infrastructure and education remains limited. By validating the feasibility, we propose a potential pathway for the broader adoption of surgical training and assistance tools, ultimately contributing to more equitable access to surgical care worldwide.

No original empirical data or scientific work that shows the integration of XR technology with AI-enhanced image reconstruction for the purpose of global collaborative surgical planning in the Global South has been identified. Thus, an explorative study design was carried out to provide firsthand and original data [34]. To the best of our knowledge, we describe a pioneering explorative study that uses collaborative XR for case conferencing and assessment using AI-enhanced 3D surface models of patients from rural Uganda in reconstructive surgery while executed only on affordable consumer-based hardware platforms.

Objective

This study intends to determine the feasibility of using new AI-enhanced image modeling technology within an immersive spatial XR scenario to collaboratively and remotely assess reconstructive patient cases in the resource-limited country of Uganda.

XSL•FO
RenderX

Methods

Study Cohort

Surgeons participating in the surgical camp by INTERPLAST-Germany e.V. at Lamu Medical Centre for Reconstructive and Global Surgery (Lamu) in Jinja, Uganda (October 22 until November 4, 2023) were invited to be involved in this exploratory study. Lamu is a private medical facility that regularly hosts international surgical camps. During this period, patients with a variety of potential reconstructive surgical conditions were included based on the following criteria.

Inclusion criteria:

- Having a condition that requires evaluation for reconstructive surgery (eg, fire or acid burns, amputations, scars, or skin and soft tissue masses or tumors).
- Being over 18 years of age, legally competent, and having provided written informed consent.

Exclusion criteria:

- Having an underlying condition that does not allow for anesthesia or surgery.
- Testing positive for COVID-19 or exhibiting symptoms of a COVID-19 infection.

Patient eligibility was assessed through a comprehensive anamnesis and physical examination relevant to the presenting pathology. When available, medical records were also reviewed.

Workflow

After participants' enrollment, data acquisition and processing were conducted. Newly created datasets were prepared on-site while VR case conferences were scheduled. Figure 1 illustrates the study's workflow, which encompasses 2 phases: the 3D model acquisition and the virtual reality case conference. Initially, the images were collected and then processed into a 3D reconstructed model using an artificial intelligence-enhanced smartphone app. Following the successful setup, the virtual reality case conference was conducted, during which each patient's individual pathology and surgical strategy, including the marking of potential incision lines and flap approaches, were discussed.

Figure 1. The workflow of this study encompasses 2 major phases.

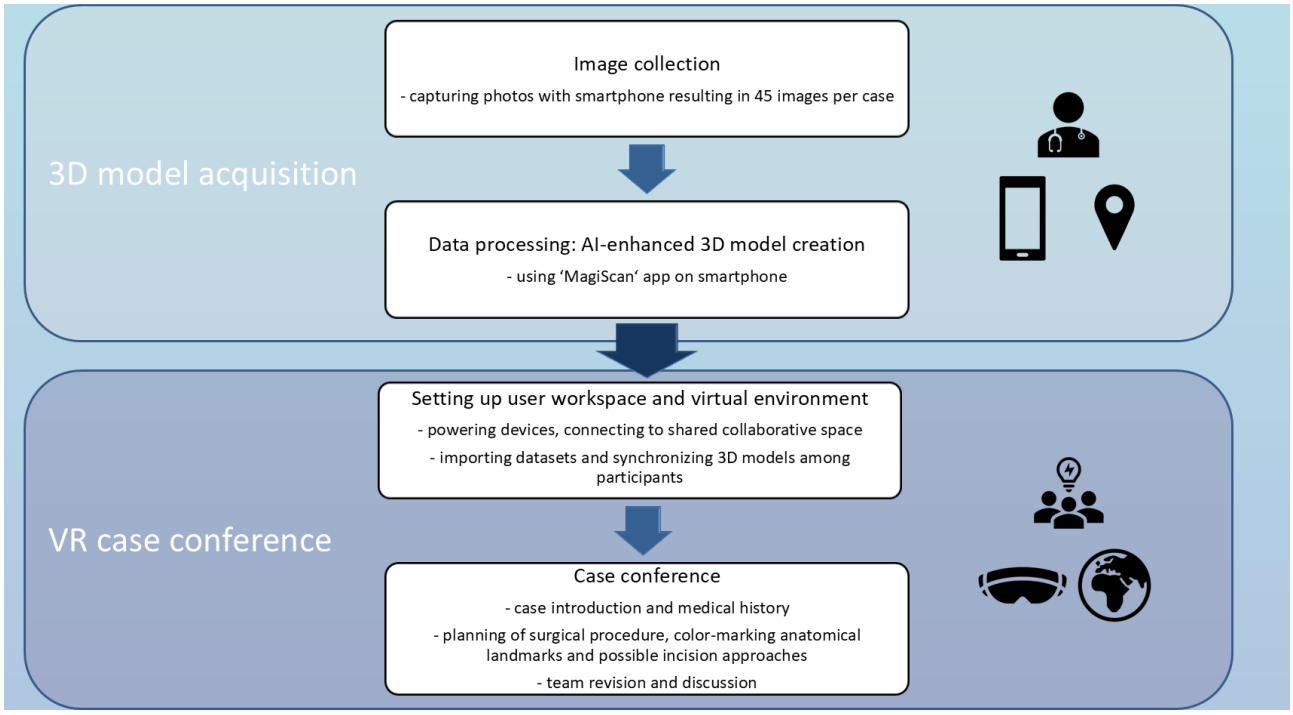

Creation of 3D Model

The 3D models of the participants' conditions were created using the AI-enhanced smartphone app MagiScan (App version 1.8), developed by MagiScan Inc. A Google Pixel 7a smartphone, Google Inc., equipped with the Android operating system (version 13), was used for scanning. Each patient underwent a photographic session. A total of 45 images were taken from various angles and orientations as the photographer executed a 360° circumferential movement around the patient. The patient was instructed to maintain a stable position, either standing or sitting, to ensure a comprehensive visualization of the pathology from all relevant angles. Inadequate lighting conditions are the most common cause of artifacts, such as inaccurate surface discontinuities. It is essential to avoid spotlights and instead use indirect lighting to ensure optimal accuracy. These photos were subsequently uploaded into the app and autonomously processed on a dedicated secure storage in a German-based data center (Frankfurt, Germany), and maintain guidelines of the EU General Data Protection Regulation (GDPR). The patient data was anonymized by assigning a numeric pseudonym.

Raw photo data were transformed into 3D models, the resulting digital representation of the patient was exported in Graphics Library Transmission Format (GLTF) to be further imported into the VR platform.

XSL•FO
RenderX

VR Hardware and Software

Two different types of VR headsets (HMDs) were used to mitigate the risk of insufficient functionality in one of the headsets and to ensure flexibility in setup by including both wired and wireless options. The Meta Quest 2 operated wirelessly as a standalone device, whereas the Oculus Rift S was wired to a workstation. Manual interaction was facilitated through VR controllers wirelessly connected to the HMDs. Both headsets enabled a functional setup. The headsets provide a similar technical performance and user experience.

The used software was XR Dissection Master (Version V0.14.3) by Medicalholodeck, installed as a standalone device.

Virtual Case Conference

The virtual case conference initiated with a visual and audio check-up. In each setting, a Ugandan surgeon from Lamu (n=2) and a remotely integrated surgeon affiliated with INTERPLAST-Germany e.V. (n=2) were equipped with a set of VR glasses and controllers, allowing them to participate in the shared digital environment. All sessions were accompanied by the research assistant and principal investigator, who ensured sufficient setup and data acquisition. The 3D datasets were uploaded onto the XR platform and automatically synchronized among participants. Case discussions commenced with an introduction of the participating surgeons. Thereafter, the patient's medical history and ongoing diagnosis were presented by the on-site surgeon. Possible procedures and tailored surgical approaches for each case were discussed. Throughout the case discussion, the app allowed users to zoom in and out of the 3D model, as well as rotate and move it as required. A toolbar provided a pen with various colors and stroke widths for highlighting and annotating specific areas of the object. During the interaction, the surgeons were represented by individually colored avatars, including virtual hands. After the mutual exchange, an individual treatment plan was proposed, and a final team revision was conducted for each case.

The case conferences were recorded on a server based at the University Hospital Bonn to later assess the durations of conference time sequences. The surgeons' feedback was collected through semistructured interviews, including Likert scale items on user experience and content quality.

Ethical Considerations

The study was conducted in accordance with the Declaration of Helsinki and received ethical approval from the institutional review board of the Mildmay Uganda Research Ethics Committee (approved September 20, 2023; Ref. MUREC-2023 - 299) and the ethics committee of the University of Bonn (approved August 15, 2023; Ref. 235/23-EP). Administrative clearance was granted by the Ugandan Ministry of Health on October 11, 2023. The study was registered with the Ugandan Council for Science and Technology (UNCST), which issued a research permit on October 23, 2023 (Ref. HS3198ES). Written informed consent was obtained from all participating patients and surgeons after they received comprehensive written and verbal information about the study. The participants were assured their right to withdraw from this study at any time. To safeguard anonymity and prevent any possibility of direct identification, a numerical pseudonym was assigned. Data access was limited exclusively to authorized members of the research team. No compensation was given to the participants.

Results

A total of 15 patients from the surgical camp conducted by INTERPLAST-Germany e.V. were included in this study (n=15). Four categories of working diagnosis were established according to the patient's surgical condition. Each patient's medical history and surgical procedure were discussed in a total of 3 intercontinental virtual case conferences. The meeting within an immersive virtual room facilitated real-time communication and interactive engagement with the 3D models of the patients. A recording of the virtual case conference is available as Multimedia Appendix 1.

The cohort consisted of 5 patients who endured burn contractures resulting from fire, hot water, or acid. Seven patients exhibited a variety of surface tumors, including lipoma, ganglion, cysts, or unknown lesions of the soft tissue. Two patients presented with chronic wounds. One patient showed soft tissue damage following trauma. The entities appeared in different locations that have been arranged into 3 distinct categories: Limb (8 patients), Head and Neck (7 patients), Trunk (1 patient). The summarized data of the cohort are presented in Table 1.

The median time for virtual case conferences was 3 minutes and 1 second, with a range from 1 minute and 33 seconds to 8 minutes and 14 seconds. Case introduction and medical history took a median of 50 seconds, while planning the surgical procedure had a median time of 1 minute and 20 seconds. Discussion and team revision took a median of 44 seconds. All time metrics related to data acquisition and 3D model creation are summarized in Table 2. The time metrics for the virtual case conferences are presented in Table 3. In some instances, team revision was not conducted if no additional questions or ambiguity arose following the discussion of surgical approaches and interacting with the 3D-reconstructed model.

Two general surgeons from Uganda and 2 reconstructive surgeons from Germany, affiliated with Interplast, conducted the XR case evaluations. During the virtual intercontinental case conferences, the surgeons actively engaged the app's tools to outline anatomical landmarks and potential incision lines. Possible flap mobilizations and the placement of skin grafts were collaboratively marked on the patient models within the immersive setting to explore potential surgical strategies. The final surgical techniques included a jumping man flap, a supraclavicular flap, and a latissimus dorsi flap. All approaches were outlined and discussed.

XSL•FO
RenderX

Table . Overview of patient cohort.

Variable	Value
Overall group (count), N (%)	15 (100)
Sex, n (%)	
Female	5 (33.33)
Male	10 (66.66)
Age (years)	
Mean (SD)	40.1(15.88)
Range (years)	18 - 75
Working diagnosis, n (%)	
Burn (fire, acid, hot water)	5 (33)
Trauma with soft tissue injury	1 (7)
Soft tissue space occupation or tumor	7 (47)
Chronic wound	2 (13)
Location, n (%)	
Head and neck	6 (40)
Trunk	1 (7)
Limb	8 (53)

Table . Metrics for 3D model reconstruction.

Data acquisition time (n=7)	Time (min: sec)
Photographic session	
Median time per case	02:33
Range	01:56 – 02:56
Processing data into 3D-reconstructed model	
Median time per case	03:10
Range	02:12 – 04:14
Total time per case	
Median	05:43
Range	05:08 – 07:13

XSL•FO
RenderX

Table . Metrics for virtual case conference.

Virtual case conference time sequences (n=15)	Time (min: sec)
Case introduction and medical history	
Median time per case	00:50
Range	00:30 – 02:11
Planning of surgical procedure	
Median time per case	01:20
Range	00:24 – 06:40
Discussion and team revision	
Median time per case	00:44
Range	00:00 – 01:13
Total virtual case conference	
Median time per case	03:01
Range	01:33 – 08:14

Figure 2 demonstrates a VR image of a patient with postburn contractures. Major anatomical landmarks are highlighted on the 3D model in Figure 2A, vascular structures are marked in Figure 2B, and key vessel axes along with their origins are shown in Figure 2C.

Figure 3A presents a VR image of another patient with postburn contractures. The "jumping man flap," proposed as a therapeutic option, is marked in Figure 3B, and further areas of tissue mobilization are highlighted in Figure 3C.

Figure 4A depicts a patient with a large lipoma, along with the proposed incision line for surgical treatment Figure 4B.

The objective of the surgeons' evaluation was to assess their detailed perception of the 3D-reconstructed patient models and their integration into a VR-based case conference. Specifically, the evaluation focused on the quality and visibility of the 3D-reconstructed models, communication features, and overall feasibility. Surgeons' feedback was collected through Likert-scale items. The frequency distribution of the response options was analyzed to compare the responses. Likert-scale results of the surgeon's feedback demonstrated that all 4 participating surgeons (4/4, 100%) strongly agreed that the 3D reconstructed models were of high quality. Regarding the visualization of large entities, all surgeons (4/4, 100%) found it satisfactory, while 2 surgeons (2/4, 50%) expressed satisfaction with the visibility of smaller entities, such as ganglions and finger contractures. All surgeons (4/4, 100%) agreed that the audio and visual communication tools were sufficient. No surgeons (0/4, 0%) reported any significant motion sickness, with only 1 surgeon reporting to have experienced slight motion sickness. However, at no point did this lead to the termination of the VR session or the need for a pause. Furthermore, all surgeons (4/4, 100%) affirmed that the technology could enhance surgical care in low-resource settings and could be integrated into routine practice for reconstructive surgery case discussions. The detailed results are provided in Multimedia Appendix 2.

XSL•FO
RenderX

Figure 2. Virtual reality images of a patient with postburn contractures of both axillae. 3D model showing the patient's back with postburn contracture of the left axilla, marking the borders of the latissimus dorsi muscle in red color. (**A**) Placing anatomical landmarks (inferior scapula, iliac crest, mid-axillary line, and dorsal median line) in purple color, drawing the muscle's main vessel and its branches (thoracodorsal artery) as well as its origin (subscapular artery) in red color, marking the skin paddle to be integrated in the flap (musculocutaneous flap) in yellow color. (**B**) Enlarged picture with landmarks and skin paddle, focusing on the vessel axis (red color), giving particular attention to its origin (yellow arrow) and the course of branches (green and blue arrows). (**C**) The enlarged picture allows a clearer imaging of the skin texture (eg, scar tissue close to the vascular origin); Images A and B show the participating surgeons as avatars; red hands and glasses: presenting surgeon, blue hands: attending surgeon.

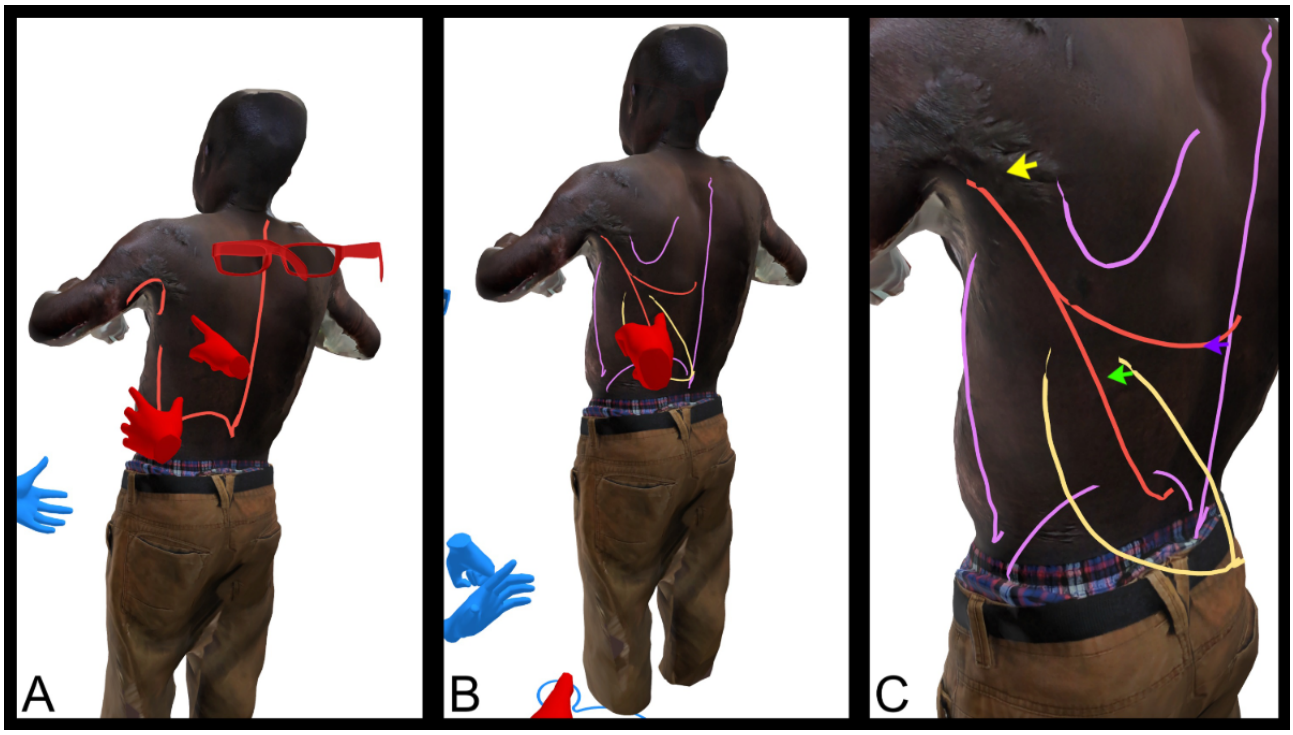

Figure 3. Point-of-View virtual reality images of a patient with postburn contractures of the right elbow. (A) 3D model showing the patient's ventral aspect of the right elbow with the planned incision line and local "jumping man" flap design marked in yellow. (B) Extensive scarring in the area of the cubital groove, as well as the neighboring tissue in the latero-ventral area of the distal upper and proximal lower arm, is marked in red. (C) Detailed surface analysis and discussion of the adjacent scar tissue (shaded area in red) by attending surgeon (purple hands) and presenting surgeon (red hands and glasses).

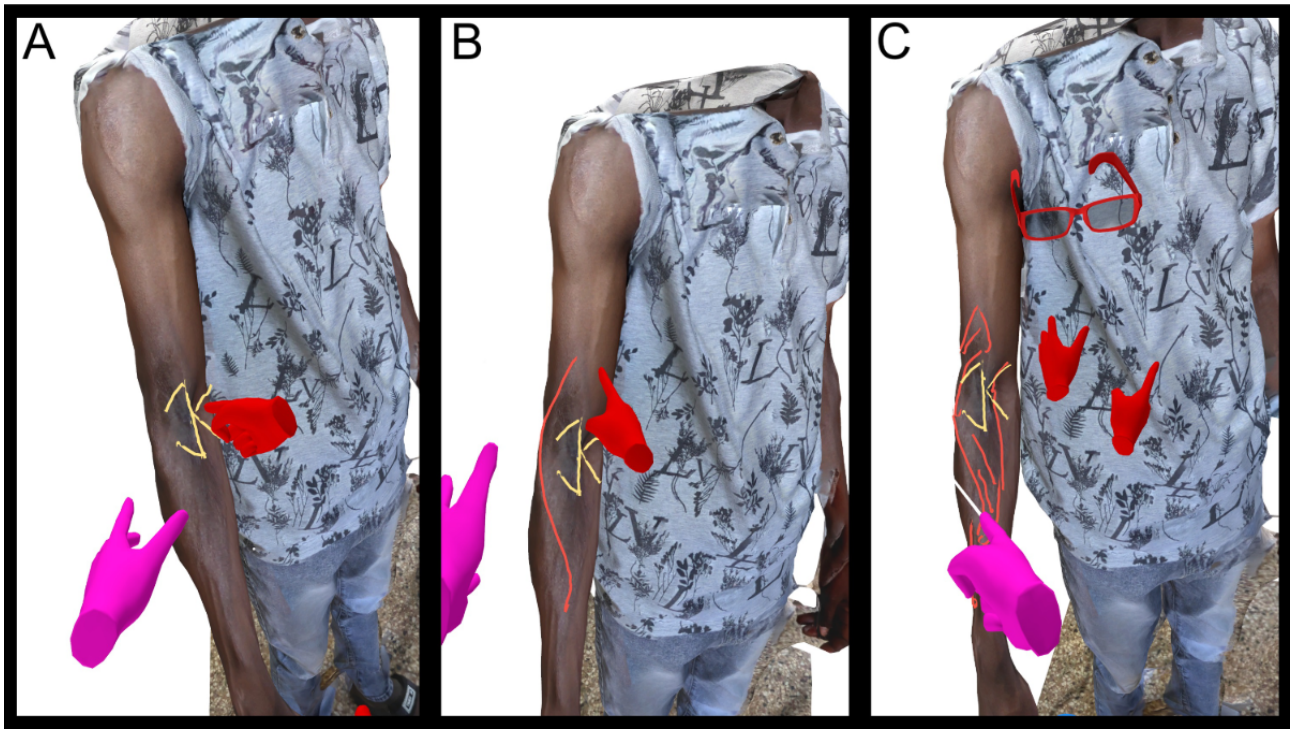

XSL•FO
RenderX

Figure 4. Patient model presenting with a painful lipoma of the lateral aspect of the left upper arm. (**A**) Outlined boundaries of the lipoma in red for an improved understanding of volume and dimension. (**B**) Marked preferred posterior incision line in yellow (dotted line) and anatomical landmarks (central axis) in purple color. Discussion and surgical planning by attending surgeon (purple hands) and presenting surgeon (red hands and glasses).

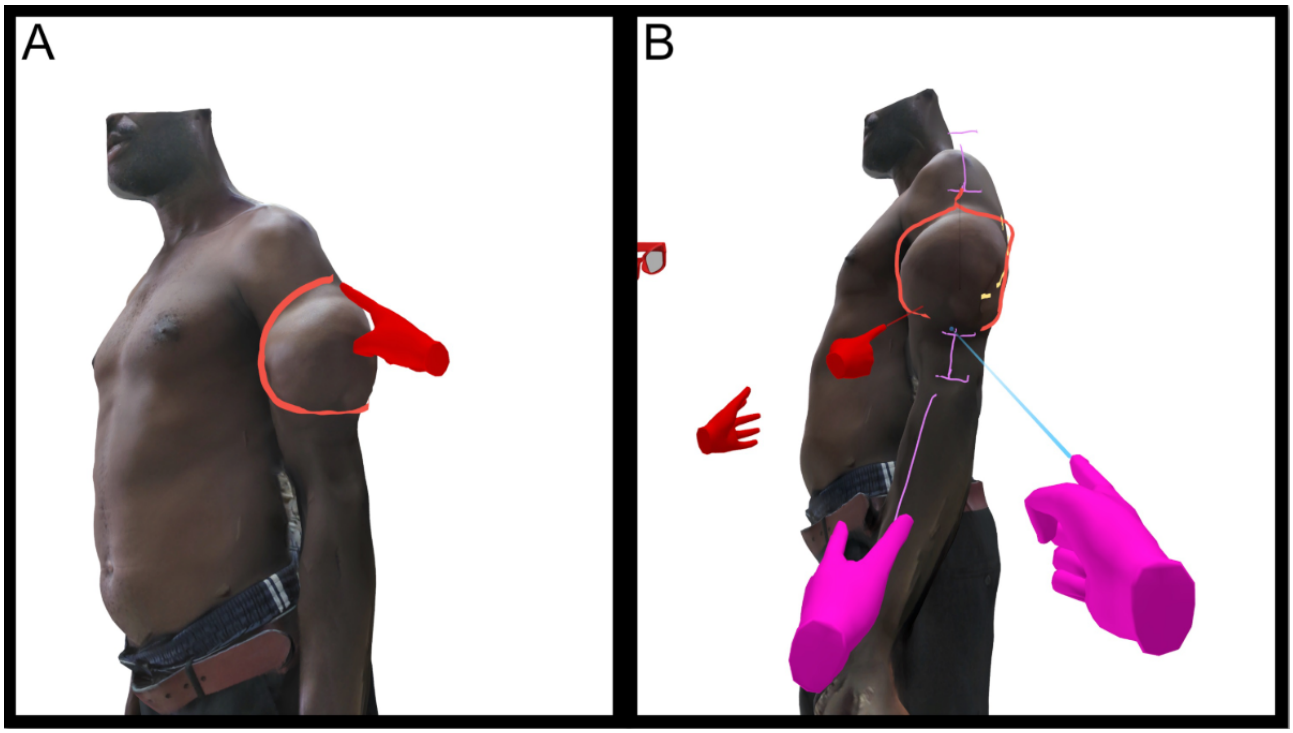

Discussion

Principal Results

Our findings demonstrated a novel, intercontinental immersive surgical health care approach. We initiated and explored a solution for remote case conferencing and interactive assessments of surgical strategies on digital patient models, providing insights into the requirements and effectiveness of AI-enhanced modeling and XR setups.

The VR-based case assessment process was time efficient, as reflected in the median metrics, enabling multiple surgeons to conduct remote, simultaneous, and detailed evaluations without being physically colocated. This collaborative setup supports improved resource allocation. Various surgical entities and conditions were assessed, with surgeons outlining individualized strategies for each case. The dynamic structure of the case conferences led to variations in timing across different assessment phases.

The accuracy of 3D reconstruction and representation of smaller lesions, such as small lipomas and ganglions, was limited, whereas larger entities were depicted with great accuracy. The smartphone-based 3D reconstructions were acquired in a reasonable timeframe but occasionally showed minor artifacts. Compared with conventional photogrammetry, AI-enhanced 3D reconstruction methods require significantly less time and computational resources [35], making them particularly valuable for rapid and cost-efficient 3D modeling in clinical settings. Ongoing developments in consumer mobile devices, particularly integrated light detection and ranging (LIDAR) technology, show promise for further applications and exploration in surgical practice, offering highly accurate 3D scanning capabilities that are already being used in the medical field [36].

Immersive Communication and Surgical Collaboration

Intercontinental communication in the immersive environment was effective during the real-time conferences in both visual gestures and speech. After their initial involvement in virtual case conferences, all participating surgeons recognized the high potential of 3D representation in XR to improve surgical care in the Global South. This approach may even serve as a future standard of care in remote planning of reconstructive operations. This underlines the quality of the patient models as well as the effectiveness of the VR communication platform.

Although no directly comparable studies currently exist, related literature underscores the influence of XR in surgery and digital communication. Numerous studies have demonstrated the potential benefits of XR in surgical planning, navigation, intraoperative guidance, and training [19,37]. In addition, the impact of digital communication solutions on case preparation and collaboration further emphasizes the potential of these technologies in surgical practice [38].

Limitations

This study aims to gain initial insights into the value and usability of the technology by implementing it during a temporal surgical camp to establish a standardized protocol. Consequently, the sample selection and composition may limit the representativeness of the patient sample compared to broader populations in public hospitals.

In addition, the small study population of this exploratory proof of principle limits the ability to draw meaningful conclusions about the technology's impact on specific surgical pathologies

XSL•FO
RenderX

or subentities. Furthermore, due to this design, the number of participating surgeons was limited, leaving room for a potential influence of individual bias.

Opportunities of XR for Surgical Care in Low-Income Regions

This study confirms the feasibility of XR-based case assessments for reconstructive surgery in a low-income country using frugal, off-the-shelf hardware. Multidirectional, immersive, and interactive communication between rural surgeons, specialists, and the broader community in the field of reconstructive surgery can reduce barriers to effective case discussions, facilitating global knowledge exchange based on real patient data [39,40]. Our study has shown that these goals are achievable with frugal equipment and limited training, with hardware costs kept below 1000 US$, making the setup accessible and scalable [41]. However, consistent, low-latency internet connectivity is essential for effective real-time VR use. XR has already proven valuable in surgical education, planning, increased precision, reduced operative times, and optimized outcomes while minimizing resource demands [42]. XR presents a significant opportunity to further educate health care professionals in anatomy and surgical skills, especially in resource-limited settings where extensive training on cadavers or live models is often unavailable. Particularly, in plastic and reconstructive surgery, XR training solutions can play a valuable role [43].

Conclusion and Outlook

Despite constrained access to modern medical technologies, this study demonstrates, for the first time, the feasibility of AI-enhanced 3D object reconstruction in conjunction with XR for reconstructive surgical assessment in a low-income country. Implementing such technologies in resource-poor settings across the Global South could significantly improve global surgery and reduce barriers to access to surgical care. It is suggested that medical AI applications need to be implemented equitably and inclusively from a global perspective [44].

While the affordability of VR headsets and AI-enabled mobile scanning continues to improve, financial barriers and infrastructural limitations, particularly internet access and initial investment, remain critical challenges [7]. Nonetheless, facilitating access to these technologies for rural health care providers could significantly expand their impact and adoption.

To scale this approach, larger prospective studies are needed to evaluate its effects on patient outcomes, cost- and time-effectiveness, and treatment workflows. Priorities for future research include exploring the implications of immersive planning on surgical decision-making, patient access, and rural outreach as patients often must travel long distances. As emphasized by the Lancet Commission, international collaboration and global visibility are essential to advance this solution beyond its pilot stage and realize its full potential in global surgery.

Acknowledgments

Parts of this study were previously presented at the “Chirurgische Forschungstage” (Surgical Research Days) in Tübingen, Germany, on September 26, 2024, and at the “Deutscher Chirurgie Kongress” (German Surgery Congress) in Munich, Germany, on March 26, 2025.

During the creation and revision process of this manuscript, we employed a language model (ChatGPT 3.5, 4.0, 4.0-mini) for grammar correction.

MO and JA are joint first authors. JW and PF are joint last authors.

This study was supported by ANDO Modular Aid e.V., Königswinter, Germany (field work, travel costs, fees for Institutional Review Board, research permit), MagiScan Inc., Warsaw Poland (application services and secure data storage), Medicalholodeck, Zurich, Switzerland (hardware, software license). The publication was supported by the Open Access Publication Fund of the University of Bonn.

Data Availability

Anonymized data are available on request from the corresponding author. Identifying patient data is not publicly available due to privacy regulations.

Authors' Contributions

JW, JK, JA, MO, and PF contributed to the conceptualization and development of the study design and methodology, under the supervision of JS. Participant recruitment was facilitated by JK, FO, BO, and MO. Data collection was conducted by FO, JK, and MO. Case conferencing and validation of the data involved JW, FO, BO, AK, and MO. Formal data analysis was performed by MO, HB, JA, and PF, with additional oversight provided by HB, JA, MO, and JW. MO, HB, JA, and JW were responsible for drafting and writing the original manuscript. JS, JW, and PF contributed to reviewing and editing the final manuscript. JW and MO also contributed to project administration.

XSL•FO
RenderX

Conflicts of Interest

The authors JA and PF declare that they hold a minority interest in Medicalholodeck AG. JA and PF received travel support from Medtronic Germany GmbH, Medicalholodeck AG and Distalmotion SA. JA declares that he has received a speaker's honorarium from Medicalholodeck AG and Chiesi GmbH. PF and JA received advisory fees from Richard Wolf GmbH. MO, JA, JW and PF hold interest in Aesthetic Vision Inc. All other authors declare that they have no conflicting interests.

Multimedia Appendix 1
Recording of the virtual case conference.
[MP4 File, 36228 KB - formative_v9i1e69300_app1.mp4]

Multimedia Appendix 2
Likert scale evaluation.
[DOCX File, 17 KB - formative_v9i1e69300_app2.docx]

References

1. Alkire BC, Raykar NP, Shrime MG, et al. Global access to surgical care: a modelling study. Lancet Glob Health 2015 Jun;3(6):e316-e323. [doi: 10.1016/S2214-109X(15)70115-4] [Medline: 25926087]
2. Price R, Makasa E, Hollands M. World Health Assembly Resolution WHA68.15: "Strengthening Emergency and Essential Surgical Care and Anesthesia as a Component of Universal Health Coverage"—addressing the public health gaps arising from lack of safe, affordable and accessible surgical and anesthetic services. World J Surg 2015 Sep;39(9):2115-2125. [doi: 10.1007/s00268-015-3153-y] [Medline: 26239773]
3. Meara JG, Leather AJM, Hagander L, et al. Global surgery 2030: evidence and solutions for achieving health, welfare, and economic development. Lancet 2015 Aug 8;386(9993):569-624. [doi: 10.1016/S0140-6736(15)60160-X] [Medline: 25924834]
4. Kim EK, Dutta R, Roy N, Raykar N. Rural surgery as global surgery before global surgery. BMJ Glob Health 2022 Mar;7(3):e008222. [doi: 10.1136/bmjgh-2021-008222] [Medline: 35318263]
5. Alser K, Mallah SI, El-Oun YRA, et al. Trauma care supported through a global telemedicine initiative during the 2023-24 military assault on the Gaza Strip, occupied Palestinian territory: a case series. Lancet 2024 Aug 31;404(10455):874-886. [doi: 10.1016/S0140-6736(24)01170-X] [Medline: 39216977]
6. Clifford GD. E-health in low to middle income countries. J Med Eng Technol 2016;40(7-8):336-341. [doi: 10.1080/03091902.2016.1256081] [Medline: 27885920]
7. Howitt P, Darzi A, Yang GZ, et al. Technologies for global health. Lancet 2012 Aug 4;380(9840):507-535. [doi: 10.1016/S0140-6736(12)61127-1] [Medline: 22857974]
8. Uganda Profile. Uganda Bureau of Statistics. 2023. URL: https://www.ubos.org/uganda-profile/ [accessed 2024-06-10]
9. World Bank Country Classifications by Income Level (Uganda). World Bank. 2022. URL: https://tinyurl.com/5cup6hrd [accessed 2024-06-10]
10. Nwanna-Nzewunwa OC, Ajiko MM, Kirya F, et al. Barriers and facilitators of surgical care in rural Uganda: a mixed methods study. J Surg Res 2016 Jul;204(1):242-250. [doi: 10.1016/j.jss.2016.04.051] [Medline: 27451893]
11. Butler EK, Tran TM, Nagarajan N, et al. Epidemiology of pediatric surgical needs in low-income countries. PLOS ONE 2017;12(3):e0170968. [doi: 10.1371/journal.pone.0170968] [Medline: 28257418]
12. Fuller AT, Corley J, Tran TM, et al. Prevalence of surgically untreated face, head, and neck conditions in Uganda: a cross-sectional nationwide household survey. World Neurosurg 2018 Feb;110:e747-e754. [doi: 10.1016/j.wneu.2017.11.099] [Medline: 29180091]
13. Davé DR, Nagarjan N, Canner JK, Kushner AL, Wong GB, SOSAS4 Research Group. Global burden of craniofacial disorders: where should volunteering plastic surgeons and governments focus their care? J Craniofac Surg 2020;31(1):121-124. [doi: 10.1097/SCS.0000000000005936] [Medline: 31821210]
14. Hodges S, Wilson J, Hodges A. Plastic and reconstructive surgery in Uganda--10 years experience. Paediatr Anaesth 2009 Jan;19(1):12-18. [doi: 10.1111/j.1460-9592.2008.02839.x] [Medline: 19076496]
15. Turyamureba M, Yawe B, Oryema JB. Health care delivery system in Uganda: a review. Tanzan J Health Res 2023;24(2):57-64. [doi: 10.4314/thrb.v24i2]
16. Müller T, Evans A, Schied C, Keller A. Instant neural graphics primitives with a multiresolution hash encoding. ACM Trans Graph 2022 Jul;41(4):1-15. [doi: 10.1145/3528223.3530127]
17. Miyake K. Evaluating the reliability of three-dimensional models constructed photogrammetry software 3DF Zephyr by measuring joint angles of fingers: a comparison to a conventional goniometer. J Plast Reconstr Surg 2024 Jan 27;3(1):34-38. [doi: 10.53045/jprs.2022-0051] [Medline: 40104417]
18. Verhey JT, Haglin JM, Verhey EM, Hartigan DE. Virtual, augmented, and mixed reality applications in orthopedic surgery. Int J Med Robot Comput Assist Surg 2020 Apr;16(2):e2067. [doi: 10.1002/rcs.2067]

XSL•FO
RenderX

19. Zhang J, Lu V, Khanduja V. The impact of extended reality on surgery: a scoping review. Int Orthop 2023 Mar;47(3):611-621. [doi: 10.1007/s00264-022-05663-z] [Medline: 36645474]
20. Sun P, Zhao Y, Men J, et al. Application of virtual and augmented reality technology in hip surgery: systematic review. J Med Internet Res 2023 Mar 10;25:e37599. [doi: 10.2196/37599] [Medline: 36651587]
21. Sayadi LR, Naides A, Eng M, et al. The new frontier: a review of augmented reality and virtual reality in plastic surgery. Aesthet Surg J 2019 Aug 22;39(9):1007-1016. [doi: 10.1093/asj/sjz043] [Medline: 30753313]
22. Yammine K, Violato C. A meta-analysis of the educational effectiveness of three-dimensional visualization technologies in teaching anatomy. Anat Sci Educ 2015;8(6):525-538. [doi: 10.1002/ase.1510] [Medline: 25557582]
23. Mao RQ, Lan L, Kay J, et al. Immersive virtual reality for surgical training: a systematic review. J Surg Res 2021 Dec;268:40-58. [doi: 10.1016/j.jss.2021.06.045] [Medline: 34284320]
24. McKinney B, Dbeis A, Lamb A, Frousiakis P, Sweet S. Virtual reality training in unicompartmental knee arthroplasty: a randomized, blinded trial. J Surg Educ 2022;79(6):1526-1535. [doi: 10.1016/j.jsurg.2022.06.008] [Medline: 35863958]
25. Sánchez-Margallo JA, Plaza de Miguel C, Fernández Anzules RA, Sánchez-Margallo FM. Application of mixed reality in medical training and surgical planning focused on minimally invasive surgery. Front Virtual Real 2021;2. [doi: 10.3389/frvir.2021.692641]
26. Arensmeyer J, Bedetti B, Schnorr P, et al. A system for mixed-reality holographic overlays of real-time rendered 3D-reconstructed imaging using a video pass-through head-mounted display-a pathway to future navigation in chest wall surgery. J Clin Med 2024 Apr 3;13(7):2080. [doi: 10.3390/jcm13072080] [Medline: 38610849]
27. Yeo CT, MacDonald A, Ungi T, et al. Utility of 3D reconstruction of 2D liver computed tomography/magnetic resonance images as a surgical planning tool for residents in liver resection surgery. J Surg Educ 2018;75(3):792-797. [doi: 10.1016/j.jsurg.2017.07.031] [Medline: 28822820]
28. Wong KC, Sun EY, Wong IOL, Kumta SM. Mixed reality improves 3D visualization and spatial awareness of bone tumors for surgical planning in orthopaedic oncology: a proof of concept study. Orthop Res Rev 2023;15:139-149. [doi: 10.2147/ORR.S421077] [Medline: 37546697]
29. Dey A, Chatburn A, Billinghurst M. Exploration of an EEG-based cognitively adaptive training system in virtual reality. Presented at: 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR); Mar 23-27, 2019; Osaka, Japan p. 220-226. [doi: 10.1109/VR.2019.8797840]
30. Feodorovici P, Schnorr P, Bedetti B, Zalepugas D, Schmidt J, Arensmeyer JC. Collaborative virtual reality real-time 3D image editing for chest wall resections and reconstruction planning. Innovations (Phila) 2023;18(6):525-530. [doi: 10.1177/15569845231217072] [Medline: 38073259]
31. Thumerel M, Belaroussi Y, Prisciandaro E, et al. Immersive three-dimensional omputed tomography to plan chest wall resection for lung cancer. Ann Thorac Surg 2022 Dec;114(6):2379-2382. [doi: 10.1016/j.athoracsur.2022.06.059] [Medline: 35963442]
32. Shirk JD, Thiel DD, Wallen EM, et al. Effect of 3-dimensional virtual reality models for surgical planning of robotic-assisted partial nephrectomy on surgical outcomes: a randomized clinical trial. JAMA Netw Open 2019 Sep 4;2(9):e1911598. [doi: 10.1001/jamanetworkopen.2019.11598] [Medline: 31532520]
33. Feodorovici P, Arensmeyer J, Schnorr P, Schmidt J. [Extended Reality (XR) - Applications in Thoracic Surgery]. Zentralblatt fuer Chirurgie 2023;148(4):367-375. [doi: 10.1055/a-2121-6478]
34. Sandelowski M. What's in a name? Qualitative description revisited. Res Nurs Health 2010 Feb;33(1):77-84. [doi: 10.1002/nur.20362]
35. Kingsland K. Comparative analysis of digital photogrammetry software for cultural heritage. Digit Appl Archaeol Cult Herit 2020 Sep;18:e00157. [doi: 10.1016/j.daach.2020.e00157]
36. Rudy HL, Wake N, Yee J, Garfein ES, Tepper OM. Three-dimensional facial scanning at the fingertips of patients and surgeons: accuracy and precision testing of iPhone X three-dimensional scanner. Plast Reconstr Surg 2020 Dec;146(6):1407-1417. [doi: 10.1097/PRS.0000000000007387] [Medline: 33234980]
37. Isikay I, Cekic E, Baylarov B, Tunc O, Hanalioglu S. Narrative review of patient-specific 3D visualization and reality technologies in skull base neurosurgery: enhancements in surgical training, planning, and navigation. Front Surg 2024;11:1427844. [doi: 10.3389/fsurg.2024.1427844] [Medline: 39081485]
38. Hammer RD, Fowler D, Sheets LR, Siadimas A, Guo C, Prime MS. Digital tumor board solutions have significant impact on case preparation. JCO Clin Cancer Inform 2020 Aug;4:757-768. [doi: 10.1200/CCI.20.00029] [Medline: 32816529]
39. Diaka J, Van Damme W, Sere F, Benova L, van de Put W, Serneels S. Leveraging smart glasses for telemedicine to improve primary healthcare services and referrals in a remote rural district, Kingandu, DRC, 2019-2020. Glob Health Action 2021 Dec 6;14(1):2004729. [doi: 10.1080/16549716.2021.2004729] [Medline: 34889718]
40. Necker FN, Cholok DJ, Shaheen MS, et al. The reconstructive metaverse - collaboration in real-time shared mixed reality environments for microsurgical reconstruction. Surg Innov 2024 Oct;31(5):563-566. [doi: 10.1177/15533506241262946] [Medline: 38905568]
41. Alsop T. Comparison of virtual reality (VR) headsets worldwide in 2024, by price (in US dollars). Statista. 2024. URL: https://www.statista.com/statistics/1337123/vr-headset-comparison-by-price/ [accessed 2024-08-14]

XSL•FO
RenderX

42. Sullivan J, Skladman R, Varagur K, et al. From augmented to virtual reality in plastic surgery: blazing the trail to a new frontier. J Reconstr Microsurg 2024 Jun;40(5):398-406. [doi: 10.1055/a-2199-3870] [Medline: 37884060]
43. Bielsa VF. Virtual reality simulation in plastic surgery training. Literature review. J Plast Reconstr Aesthet Surg 2021 Sep;74(9):2372-2378. [doi: 10.1016/j.bjps.2021.03.066]
44. Liebrenz M, Bhugra D, Alibudbud R, Ventriglio A, Smith A. AI in health care and the fragile pursuit of equity and social justice. Lancet 2024 Aug 31;404(10455):843. [doi: 10.1016/S0140-6736(24)01604-0] [Medline: 39216963]

Abbreviations

AI: artificial intelligence
AR: augmented reality
CT: computed tomographic
GDPR: General Data Protection Regulation
GLTF: Graphics Library Transmission Format
HMD: head-mounted displays
LIDAR: light detection and ranging
MR: mixed reality
MRI: magnetic resonance imaging
UNCST: Ugandan Council for Science and Technology
VR: virtual reality
XR: extended reality

Edited by A Mavragani; submitted 26.11.24; peer-reviewed by A Edalatpour, J Kolcz; revised version received 11.04.25; accepted 23.04.25; published 18.08.25.

Please cite as:
Obst M, Arensmeyer J, Bonsmann H, Kolbinger A, Kigenyi J, Oneka F, Owere B, Schmidt J, Feodorovici P, Wynands J
AI-Enhanced 3D Models in Global Virtual Reality Case Conferences for Surgical Care in a Low-Income Country: Exploratory Study
JMIR Form Res 2025;9:e69300
URL: https://formative.jmir.org/2025/1/e69300
doi:10.2196/69300

© Miriam Obst, Jan Arensmeyer, Henrik Bonsmann, Andreas Kolbinger, Joel Kigenyi, Francis Oneka, Benard Owere, Joachim Schmidt, Philipp Feodorovici, Jan Wynands. Originally published in JMIR Formative Research (https://formative.jmir.org), 18.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Comparability of Daily-Life Walking Speed Measured by Smartphone GPS and Ankle-Band Accelerometer: Cross-Sectional Study

Hisashi Kawai, PhD; Keigo Imamura, PT, PhD; Rui Gong, PhD; Manami Ejiri, PhD; Shuichi Obuchi, PT, PhD

Research Team for Human Care, Tokyo Metropolitan Institute for Geriatrics and Gerontology, 35-2 , Sakae-cho, Itabashi-ku, Tokyo, Japan

Corresponding Author:
Hisashi Kawai, PhD
Research Team for Human Care, Tokyo Metropolitan Institute for Geriatrics and Gerontology, 35-2 , Sakae-cho, Itabashi-ku, Tokyo, Japan

Abstract

Background: Daily-life walking speed (DWS), which is a critical health indicator in older adults, can be measured using smartphone GPS technology. Although this method is becoming more widely accessible**,** it is restricted to outdoor walking. In contrast, accelerometers can capture walking speed indoors; however**,** there is limited information on the comparability of DWS data between these two techniques.

Objective: In this study, we aimed to investigate the agreement and systematic error between DWS measured using the built-in GPS of a smartphone and an ankle-band accelerometer.

Methods: Participants were recruited from a previously selected cohort of community-dwelling older adults. Their DWS was assessed using both a smartphone app and ankle band accelerometer. Data from the two devices were matched based on simultaneous recordings, and agreement between the walking speeds was evaluated using the intraclass correlation coefficient (ICC) and a Bland–Altman plot.

Results: A total of 99 participants (38 men, 61 women; mean [SD] age 71.5 [4.9] y) were included, yielding 3652 paired data points. The mean (SD) DWS as measured by GPS was 1.30 (0.19) m/s, and this was significantly higher than the value of 1.12 (0.23) m/s measured by the accelerometer (P<.001). The ICC(2, 1) (95% CI) value was 0.523 (−0.022 to 0.765), and the ICC(2, k) value was 0.687 (−0.045 to 0.867). The Bland–Altman plot revealed a fixed error of 0.18 m/s with 95% limits of agreement ranging from −0.49 m/s to 0.13 m/s in the GPS-measured walking speed compared to that measured by the ankle-band accelerometer.

Conclusions: The GPS app consistently recorded a faster walking speed than the ankle-band accelerometer. The agreement between the measurements of the two devices was poor. The data suggest that a correction is necessary when comparing the DWS between these two devices.

(JMIR Form Res 2025;9:e73722) doi:10.2196/73722

KEYWORDS

daily-life walking speed; accelerometer; GPS; smartphone; wearable health technology; ankle-band accelerometer; walking speed assessment

Introduction

Walking speed is a crucial indicator of various health outcomes in older adults, including instrumental activities of daily living disability [1], hospitalization [2], and mortality [3]. Previous studies have measured walking speed in controlled laboratory settings, typically on a walkway. However, recent advancements in wearable sensor technology have facilitated the measurement of daily-life walking speed (DWS) [4-9]. Unlike laboratory-based walking assessments, DWS captured by wearable sensors reflects natural, unintentional walking behavior, which makes it a valuable tool for daily health monitoring. A meta-analysis indicated that the use of activity trackers increased the number of steps taken per day and the moderate-to-vigorous physical activity duration [10]. Their results demonstrated that providing participants with daily feedback on their walking may encourage behavioral changes that promote better health.

We previously reported that the test-retest reliability of 1 week of DWS measured via smartphone GPS was adequate [7], that 1 month of DWS data was significantly associated with frailty [6,11], and that DWS distribution over a certain period followed a unimodal normal distribution [12]. These observations highlighted the potential utility of GPS in preventing frailty. In Japan, the smartphone ownership rate among people aged ≥70 years is 60.6% [13]. Therefore, GPS-based DWS measurement

XSL•FO
RenderX

has the potential for widespread adoption. However, a key limitation of GPS technology is that it only captures outdoor walking data. As frailty progresses and they spend more time indoors, assessing indoor walking speed will become increasingly important.

Accelerometers provide an alternate method for measuring indoor walking speed. Previous studies have used wrist- [4,5,14], waist- [15], and ankle-band accelerometers [16,17] for this purpose. Ankle-band accelerometers directly measure foot movements during walking and are more accurate than waist-band accelerometers for classifying walking speeds [16]. The DWS distributions measured by accelerometers have exhibited a bimodal pattern that allows for the differentiation between indoor and outdoor walking speeds [15]. In contrast, DWS distribution captured by smartphone GPS typically yields a unimodal distribution. These output differences demonstrate the need for comparative studies to compare the DWS measured by each device. Despite the importance of such comparisons, no such research exists.

This study aimed to investigate the agreement and systematic error between DWS measured using the built-in GPS of a smartphone and an ankle-band accelerometer, which is considered to be the most accurate accelerator-based method for measuring walking speed. We also examined the comparability of DWS measurements obtained using these two devices.

Methods

Participants

Participants were recruited from the "Otassha Study 2011" cohort during the 2022 survey as part of a life-log monitoring project using wearable sensors [18,19]. This cohort study recruited residents of nine areas within Itabashi Ward, Tokyo, and conducted annual comprehensive health assessments. A total of 623 individuals participated in the 2022 survey, of whom 276 consented to 1 year of life-log monitoring.

This life-log monitoring project, conducted as a research project by the Tokyo Metropolitan Government, is officially known as "Smart Watch Innovation for Next Geriatrics and Gerontology (SWING-Japan)." In this project, the walking, sleep, and physical activity of participants were continuously monitored over 1 year using a smartwatch, an ankle-band accelerometer, and a smartphone application. Participants were instructed to wear these devices consistently during their daily activities.

In the follow-up cohort survey conducted 1 year later, participants in the present study consented to use data collected via the smartphone app and ankle-band accelerometer.

Ethical Considerations

The purpose and procedures of the study were explained to the participants both orally and in writing, and written informed consent was obtained from all individuals. This study was approved by the ethics committee of the Tokyo Metropolitan Institute for Geriatrics and Gerontology (Approval number: R22-094). The data were de-identified. No compensation was provided to the participants.

DWS Measurement

Walking speed, step length, and cadence during daily life were measured using a smartphone app (Kenko Choujyu App, TMIG, Tokyo, Japan) and an ankle-band accelerometer (WALK X, ACOS Co., Ltd., Nagano, Japan). The participants used their personal smartphones in this study. The app determined the location using the built-in GPS of the smartphone. Walking activity was detected using a built-in step counter app programming interface. The app began measuring walking speed when stable outdoor walking was detected and continued recording until the end of the session. The GPS measured the walking status and position every second, and the walking speed was calculated from the distance and time during stable walking. The number of steps taken during stable walking was obtained from the step counter, and cadence and step length were also calculated. The test-retest reliability and validity of the walking speed measurements obtained via this app were verified in a previous study [7]. Although a variety of communication environments such as Wifi, 5G, and 4G were available, the smartphone app only used mobile data when uploading the required information, and there was no impact on the measured values.

An ankle-band accelerometer was securely attached to the upper lateral malleolus of each participant. This device used a patented algorithm (Patent number: 6843332) to calculate the walking speed, step length, and cadence for each walking cycle based on the height of the participant and signals from a three-axis acceleration sensor. The accelerometer was connected to a smartphone app via Bluetooth, and the data were transferred to a cloud-based storage with timestamps that used mobile data from the smartphone. The accelerometer calculated walking speed for each walking cycle, and the average value for 1 min was recorded. The 1-minute average values were transmitted and maintained in cloud-based storage during long-term monitoring. The validity of this ankle-band accelerometer, compared to stopwatch measurements, was confirmed in a previous study [17]. Daily walking data from the smartphone app and the ankle-band accelerometer were stored in the cloud, along with corresponding timestamps.

Data Matching

Daily-life walking data that had been recorded between November 1, 2022, and June 1, 2023 were obtained from the cloud for this study. The smartphone app data were converted to minute-by-minute datasets by truncating the seconds from each timestamp. These adjusted data were then synchronized with the corresponding minute-by-minute data measured using the ankle-band accelerometer.

Statistical Analysis

Participant characteristics such as age, height, weight, normal walking speed in a laboratory environment, Mini-Mental State Examination-Japanese (MMSE-J) score, and prevalence of hypertension, stroke, heart disease, and diabetes were summarized as means and SD. These parameters were obtained from the cohort survey.

Normal walking speed was assessed by measuring the time required to walk the middle 5 m of an 11 m walkway (3 m

XSL•FO
RenderX

acceleration + middle 5 m + 3 m deceleration). The measurement was performed twice, and the better of the two results was used. The MMSE-J was administered by a trained examiner, and the presence of chronic diseases was determined through nurse-conducted interviews.

The differences in the daily-life walking parameters measured by the two devices were assessed using paired *t* tests. Agreement between devices was measured using intraclass correlation coefficients (ICCs) ICC(2,1) and ICC(2,k). As the DWS is often represented by the average value for a week or month [7,11], we also examined the ICC(2, k) to determine the reliability across multiple measurements. Based on the 95% CI of the ICC estimate, we assessed values <0.5, 0.5 - 0.75, 0.75 - 0.9, and >0.9 as indicative of poor, moderate, good, and excellent reliability, respectively [20]. Distributional patterns were visualized using violin plots, and additional examinations that were performed after walking bouts <100 m were excluded. The DWS was assessed for systematic errors using Bland–Altman Plots. The mean percentage error (MPE) and mean absolute percentage error (MAPE) for DWS were also calculated as follows using the formulae from previous research [21]:

$$MPE = \frac{acclerometerDWS - (GPSDWS)}{accelerometerDWS} \times 100$$

$$MAPE = \frac{acclerometerDWS - (GPSDWS)}{accelerometerDWS} \times 100$$

All statistical analyses were conducted using SPSS Statistics 27 (IBM Japan, Ltd., Tokyo) and R version 4.4.1. We did not use generative AI to produce any portion of this work.

Results

A total of 99 participants, whose walking data matched those from the smartphone app and ankle-band accelerometer, were included in the final analysis (Figure 1). The study included 99 participants (38 men and 61 women), with a mean (SD) age of 71.5 (4.9) years. The mean normal walking speed was 1.5 (0.3) m/s, and the mean MMSE-J score was 28.8 (1.5). The prevalence rates of hypertension, stroke, heart disease, and diabetes were 41.4%, 7.1%, 21.2%, and 10.1%, respectively (Table 1).

Figure 1. Flowchart of the study participants.

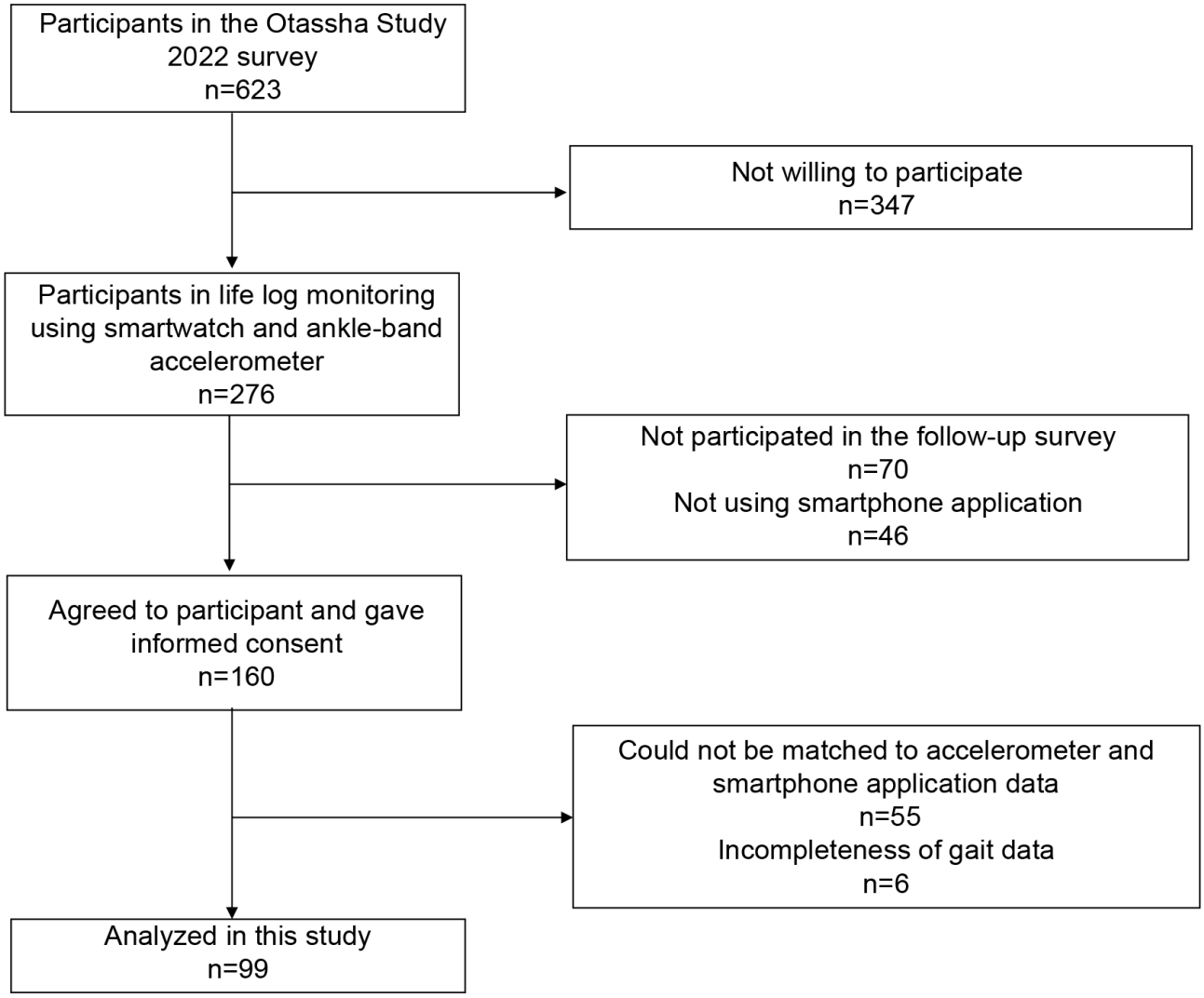

XSL•FO
RenderX

Table . Participant characteristics.

Characteristics	Men (n=38)	Women (n=61)	Overall (n=99)
Age (years), mean (SD)	71.5 (4.9)	71.5 (5.0)	71.5 (4.9)
Height (cm), mean (SD)	167.7 (5.4)	153.4 (5.6)	158.9 (8.9)
Weight (kg), mean (SD)	66.9 (10.4)	53.3 (9.6)	58.5 (11.9)
Normal walking speed (m/s), mean (SD)	1.5 (0.3)	1.5 (0.2)	1.5 (0.3)
MMSE-J[a] score, mean (SD)	28.3 (2.0)	29 (1.1)	28.8 (1.5)
Hypertension, n (%)	18 (47.4)	23 (37.7)	41 (41.4)
Stroke, n (%)	3 (7.9)	4 (6.6)	7 (7.1)
Heart disease, n (%)	9 (23.7)	12 (19.7)	21 (21.2)
Diabetes, n (%)	7 (18.4)	3 (4.9)	10 (10.1)

[a]MMSE-J: Mini-Mental State Examination-Japanese.

The average (SD) measurement period was 121.1 (68.2) days. The smartphone app recorded 6527 minute-by-minute walking data points over the measurement period; however, only 3652 data points could be matched with the accelerometer data. The mean (SD) DWS measured by GPS was 1.30 (0.19) m/s, which was significantly higher than the value measured by the accelerometer of 1.12 (0.23) m/s (P<.001; Table 2). The step length and cadence recorded by the GPS were also significantly greater than those measured by the accelerometer (P<.001). The ICC(2,1) [95% CI] values for DWS, step length, and cadence measured using the two devices were 0.523 [−0.022 to 0.765], 0.297 [−0.068 to 0.557], and 0.440 [0.385 to 0.490], respectively, and the ICC(2,k) [95% CI] values were 0.687 [−0.045 to 0.867], 0.458 [−0.147 to 0.715], and 0.611 [0.556 to 0.657], respectively (Table 2).

Table . Comparisons of daily-life walking speed, step length, and cadence measured by the GPS and ankle-band accelerometer.

Variables	GPS		ACC[a]	*P* value	ICC[b] (2,1)	95% CI	ICC (2,k)	95% CI
	Observation	Mean (SD)	Mean (SD)					
Walking speed (m/s)	3652	1.30 (0.19)	1.12 (0.23)	<.001	0.523	−0.022 to 0.765	0.687	−0.045 to 0.867
Step length (m)	3652	0.67 (0.08)	0.57 (0.09)	<.001	0.297	−0.068 to 0.557	0.458	−0.147 to 0.715
Cadence (step/min)	3652	116.72 (13.13)	111.80 (21.09)	<.001	0.440	0.385 to 0.490	0.611	0.556 to 0.657

[a]ACC: ankle-band accelerometer.

[b]ICC: intra-class correlation coefficient.

Violin plots of daily-life walking parameters measured using the two devices demonstrated that GPS-measured walking speeds were consistently higher than those recorded using the accelerometer (Figure 2). The discrepancy in the distribution was more pronounced for step length than for cadence. These trends persisted even when walking bouts of <100 m were excluded (Supplementary Figure 1 in Multimedia Appendix 1).

XSL•FO
RenderX

Figure 2. Violin plots of daily walking speed, step length, and cadence measured using GPS and an ankle-band accelerometer (ACC) (observations=3652).

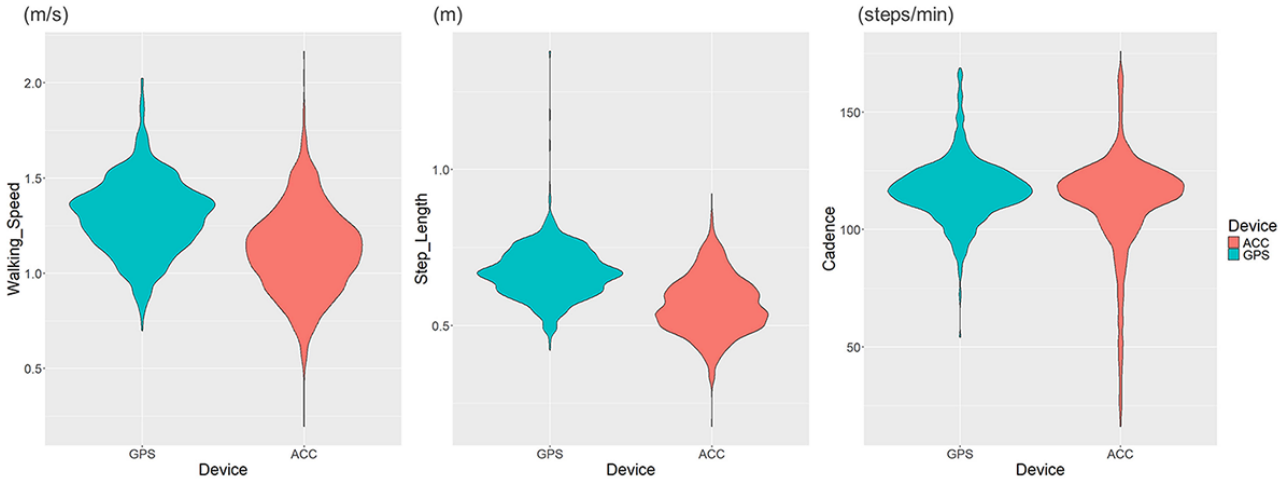

A Bland–Altman plot, which was used to depict the differences between DWS measurements obtained by the accelerometer and GPS on the vertical axis, revealed a negatively skewed distribution (Figure 3). The mean difference in DWS between the two devices was −0.18 m/s, with 95% limits of agreement ranging from −0.49 m/s to 0.13 m/s. The MPE and MAPE for DWS (SD) were −18.5% (19.2) and 19.9% (17.8), respectively.

Figure 3. Bland-Altman plots of comparing daily walking speed measurements obtained from GPS and an ankle-band accelerometer (ACC) (Observations=3652). Dotted lines: 95% limit of agreement.

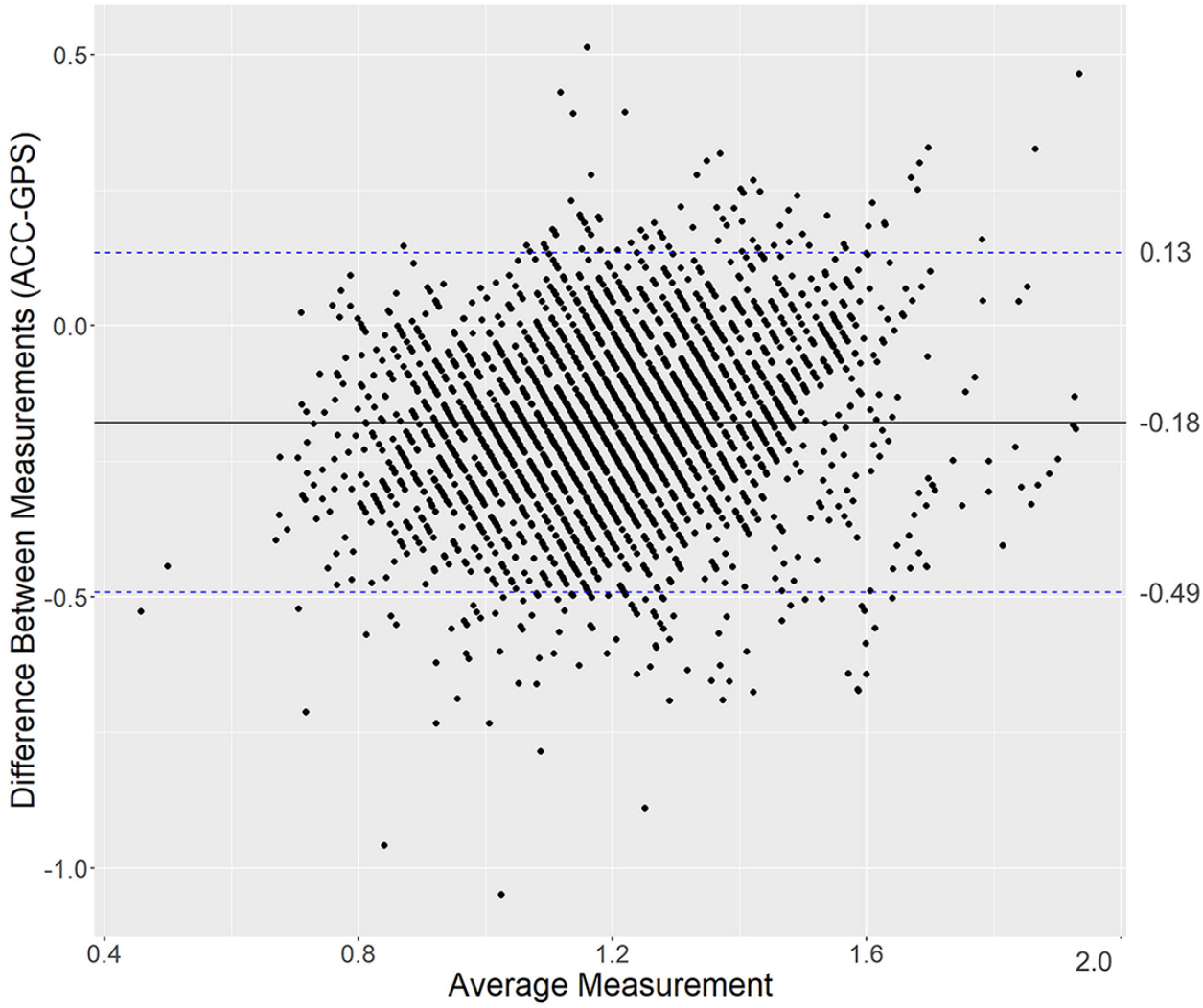

XSL•FO
RenderX

Discussion

Principal Results

This study compared DWS measurements of community-dwelling older adults obtained from a smartphone GPS and an ankle-band accelerometer. The normal walking speed measured in the cohort survey was 1.5 m/s, which was faster than the 1.25 - 1.29 m/s reported in the Japanese integral cohort study [22]. Additionally, the prevalence of hypertension, stroke, heart disease, and diabetes among participants tended to be lower in the present study. These findings suggest that individuals who participated in long-term measurements using wearable sensors were healthier than the representative cohort of community-dwelling older people in Japan.

Although we included data spanning up to 7 months in the analyses, the mean number of paired data points per participant was small (36.9). Some participants had only a few paired data points despite long measurement periods, indicating that the thresholds for detecting walking differed between the two devices.

The agreement between the two devices was poor, except for the ICC(2,k) of cadence, and not compatible. The range of 95% limits of agreement for DWS was wide (from −0.49 to 0.13 m/s). Although the 95% minimum detection change for walking speed is approximately 0.1 - 0.2 m/s [23], in the present study, the negative side, in particular, was significantly larger, suggesting that the two measurements were not comparable. The MPE and MAPE suggested that the two measurements exhibited unacceptable percentages of error. The violin plots showed that the GPS had a narrower measurement range than the accelerometer, which shifted the distribution toward a faster walking speed. This observation suggested that the accelerometer was more sensitive to slower walking speeds.

Comparison With Prior Work

A previous study reported that an ankle-band accelerometer had greater classification accuracy for walking speed than accelerometers attached to other body locations [16]. In the present study, the ankle-band accelerometer might have provided a more accurate measurement of slower walking speeds over short outdoor walking bouts. A previous study on DWS that included indoor walking showed that short walking bouts of 5 - 10 steps (equivalent to 3 - 7 m) accounted for more than 40% of all walking [24]. However, in the present study, which was limited to outdoor walking, the accelerometer detected walking for distances <100 m in 47.3% of cases, whereas only 9.4% were detected by the GPS. However, excluding walking bouts <100 m did not alter the overall distribution pattern, which indicated that the discrepancy between devices was not solely attributable to short-duration walks. Walking distances of >500 m accounted for 17.7% of the total GPS-detected walking bouts, but only 1.1% of those were detected by the ankle-band accelerator. This information suggested that the criteria for determining stable walking interruptions differed between the two devices. These results indicate that the GPS-based tracking does not detect short walking bouts as effectively as the ankle-band accelerator, but is more likely to capture longer walking bouts. However, since previous studies have demonstrated the reliability and validity of GPS-based measurements [7,11,12], this characteristic does not indicate that GPS is inherently less accurate in measuring DWS.

Limitations

A key limitation of this study was that it used a single accelerometer and GPS product. Therefore, the extent to which these findings apply to other wearable devices remains uncertain. Nevertheless, the results reflected the high classification accuracy of the ankle-band accelerometer in detecting walking speeds and they may be generalizable to some extent. As ankle edema was not examined in this study, its effect on the accuracy of ankle-band accelerometers is unknown. Additionally, the degree of agreement between devices might have varied among older adults with frailty or specific medical conditions [25]. This study also analyzed DWS exclusively during outdoor walking. Future research should examine the relationship between indoor and outdoor DWS.

Potential participants for this study were 230 smartphone users who participated in the life-log monitoring project; however, 70 participants did not participate in the 1-year follow-up survey or did not consent to this study. Additionally, 61 participants who could not be matched with data dropped out, resulting in a high dropout rate of 57.0%. Approximately 30% of participants stopped life-log monitoring within 1 year. Moreover, maintaining the long-term wearing rate of wearable devices among older adults is challenging.

Conclusions

This study compared DWS measurements between a smartphone GPS and an ankle-band accelerometer in community-dwelling older adults. The GPS consistently recorded faster walking speeds than the ankle-band accelerometer, with a fixed unacceptable error of 0.18 m/s. Correction factors should be applied when comparing data between these two devices. Further research is necessary to investigate the comparability of DWS measurements across different populations and walking environments.

Acknowledgments

This study was supported by the Japan Society for Promotion of Science (JSPS) KAKENHI (Grant Numbers 23K11959 and 23K06919). We would like to express our gratitude to all the residents and staff members who participated in the Otassha study. We also extend our appreciation to the members of the SmartWatch Innovation for Next Geriatrics and Gerontology (SWING-Japan) team for their support.

XSL•FO
RenderX

Data Availability

The data from the Otassha Study contain sensitive participant information and cannot be publicly released due to the ethical and legal restrictions imposed by the Ethics Committee of the Tokyo Metropolitan Institute for Geriatrics and Gerontology.

Authors' Contributions

HK and SO contributed to the study's design and conceptualization. HK, KI, and RG acquired data. HK, KI, and RG analyzed and interpreted the data and drafted the manuscript. KI, ME, and SO revised the manuscript for intellectual content.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Violin plots of daily walking speed, step length, and cadence measured by GPS and an ankle-band accelerometer (ACC) for walking distances greater than 100 m (Observations = 1,921).
[PNG File, 121 KB - formative_v9i1e73722_app1.png]

References

1. Perera S, Patel KV, Rosano C, et al. Gait speed predicts incident disability: a pooled analysis. J Gerontol A Biol Sci Med Sci 2016 Jan;71(1):63-71. [doi: 10.1093/gerona/glv126] [Medline: 26297942]
2. Fritz S, Lusardi M. White paper: "walking speed: the sixth vital sign". J Geriatr Phys Ther 2009;32(2):46-49. [Medline: 20039582]
3. Studenski S, Perera S, Patel K, et al. Gait speed and survival in older adults. JAMA 2011 Jan 5;305(1):50-58. [doi: 10.1001/jama.2010.1923] [Medline: 21205966]
4. Chan LLY, Espinoza Cerda MT, Brodie MA, Lord SR, Taylor ME. Daily-life walking speed, running duration and bedtime from wrist-worn sensors predict incident dementia: A watch walk - UK biobank study. Int Psychogeriatr 2025 Jun;37(3):100031. [doi: 10.1016/j.inpsyc.2024.100031] [Medline: 39827009]
5. Chan LLY, Lord SR, Brodie MA. Is maximal or usual walking speed from large scale wrist sensor data better at predicting dementia, depression and death?*. Presented at: 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); Jul 24-27, 2023; Sydney, Australia p. 1-4. [doi: 10.1109/EMBC40787.2023.10340255]
6. Kawai H, Obuchi S, Ejiri M, Ito K. Association between daily life walking speed and frailty measured by a smartphone application: a cross-sectional study. BMJ Open 2023 Jan 6;13(1):e065098. [doi: 10.1136/bmjopen-2022-065098] [Medline: 36609327]
7. Obuchi SP, Tsuchiya S, Kawai H. Test-retest reliability of daily life gait speed as measured by smartphone global positioning system. Gait Posture 2018 Mar;61:282-286. [doi: 10.1016/j.gaitpost.2018.01.029] [Medline: 29413798]
8. Takayanagi N, Sudo M, Yamashiro Y, et al. Predictivity of daily gait speed using tri-axial accelerometers for two-year incident disability among Japanese older adults. Sci Rep 2022 Jun 16;12(1):10067. [doi: 10.1038/s41598-022-14304-9] [Medline: 35710722]
9. Van Ancum JM, van Schooten KS, Jonkman NH, et al. Gait speed assessed by a 4-m walk test is not representative of daily-life gait speed in community-dwelling adults. Maturitas 2019 Mar;121:28-34. [doi: 10.1016/j.maturitas.2018.12.008] [Medline: 30704562]
10. Ferguson T, Olds T, Curtis R, et al. Effectiveness of wearable activity trackers to increase physical activity and improve health: a systematic review of systematic reviews and meta-analyses. Lancet Digit Health 2022 Aug;4(8):e615-e626. [doi: 10.1016/S2589-7500(22)00111-X] [Medline: 35868813]
11. Kawai H, Obuchi S, Watanabe Y, et al. Association between daily living walking speed and walking speed in laboratory settings in healthy older adults. Int J Environ Res Public Health 2020 Apr 15;17(8):2707. [doi: 10.3390/ijerph17082707] [Medline: 32326419]
12. Obuchi SP, Kawai H, Murakawa K. Reference value on daily living walking parameters among Japanese adults. Geriatr Gerontol Int 2020 Jul;20(7):664-669. [doi: 10.1111/ggi.13931] [Medline: 32378318]
13. Results of the 2023 communication usage trend survey. Ministry of Internal Affairs and Communications, Japan. 2023. URL: https://www.soumu.go.jp/main_sosiki/joho_tsusin/eng/pressrelease/2024/6/7_1.html [accessed 2025-08-07]
14. Soltani A, Abolhassani N, Marques-Vidal P, Aminian K, Vollenweider P, Paraschiv-Ionescu A. Real-world gait speed estimation, frailty and handgrip strength: a cohort-based study. Sci Rep 2021 Sep 23;11(1):18966. [doi: 10.1038/s41598-021-98359-0] [Medline: 34556721]
15. Rojer AGM, Coni A, Mellone S, et al. Robustness of in-laboratory and daily-life gait speed measures over one year in high functioning 61- to 70-year-old adults. Gerontology 2021;67(6):650-659. [doi: 10.1159/000514150] [Medline: 33752214]
16. Moulaee Conradsson D, Bezuidenhout LJR. Establishing accelerometer cut-points to classify walking speed in people post stroke. Sensors (Basel) 2022 May 27;22(11):4080. [doi: 10.3390/s22114080] [Medline: 35684697]

XSL•FO
RenderX

17. Imamura K, Kawai H, Ejiri M, et al. Validity of walking speed measurements using an ankle accelerometer in older individuals. Aging Med Healthc (Forthcoming) 2025.
18. Fujiwara Y, Suzuki H, Kawai H, et al. Physical and sociopsychological characteristics of older community residents with mild cognitive impairment as assessed by the Japanese version of the Montreal cognitive assessment. J Geriatr Psychiatry Neurol 2013 Dec;26(4):209-220. [doi: 10.1177/0891988713497096] [Medline: 23920040]
19. Kera T, Kawai H, Yoshida H, et al. Classification of frailty using the Kihon checklist: a cluster analysis of older adults in urban areas. Geriatr Gerontol Int 2017 Jan;17(1):69-77. [doi: 10.1111/ggi.12676] [Medline: 26792435]
20. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016 Jun;15(2):155-163. [doi: 10.1016/j.jcm.2016.02.012] [Medline: 27330520]
21. Bianchini E, Caliò B, Alborghetti M, et al. Step-counting accuracy of a commercial smartwatch in mild-to-moderate PD patients and effect of spatiotemporal gait parameters, laterality of symptoms, pharmacological state, and clinical variables. Sensors (Basel) 2022 Dec 25;23(1):214. [doi: 10.3390/s23010214] [Medline: 36616812]
22. Seino S, Shinkai S, Fujiwara Y, et al. Reference values and age and sex differences in physical performance measures for community-dwelling older Japanese: a pooled analysis of six cohort studies. PLoS ONE 2014;9(6):e99487. [doi: 10.1371/journal.pone.0099487] [Medline: 24923425]
23. Middleton A, Fritz SL, Lusardi M. Walking speed: the functional vital sign. J Aging Phys Act 2015 Apr;23(2):314-322. [doi: 10.1123/japa.2013-0236] [Medline: 24812254]
24. Mueller A, Hoefling HA, Muaremi A, et al. Continuous digital monitoring of walking speed in frail elderly patients: noninterventional validation study and longitudinal clinical trial. JMIR Mhealth Uhealth 2019 Nov 27;7(11):e15191. [doi: 10.2196/15191] [Medline: 31774406]
25. Kirk C, Küderle A, Micó-Amigo ME, et al. Mobilise-D insights to estimate real-world walking speed in multiple conditions with a wearable device. Sci Rep 2024 Jan 19;14(1):1754. [doi: 10.1038/s41598-024-51766-5] [Medline: 38243008]

Abbreviations

DWS: daily-life walking speed
ICC: intraclass correlation coefficient
MAPE: mean absolute percentage error
MPE: mean percentage error

Edited by A Mavragani; submitted 14.03.25; peer-reviewed by E Ross, E Bianchini; revised version received 16.06.25; accepted 23.06.25; published 18.08.25.

Please cite as:
Kawai H, Imamura K, Gong R, Ejiri M, Obuchi S
Comparability of Daily-Life Walking Speed Measured by Smartphone GPS and Ankle-Band Accelerometer: Cross-Sectional Study
JMIR Form Res 2025;9:e73722
URL: https://formative.jmir.org/2025/1/e73722
doi:10.2196/73722

© Hisashi Kawai, Keigo Imamura, Rui Gong, Manami Ejiri, Shuichi Obuchi. Originally published in JMIR Formative Research (https://formative.jmir.org), 18.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Refining a Video and Text Message Intervention (STAR, Skills Training in Active Recovery) to Prevent the Onset or Escalation of Posttraumatic Stress and Opioid Misuse Among Recent Sexual Assault Survivors: Community Engaged Study

Kate Walsh[1], PhD; Anne Marie Schipani-McLaughlin[2], MPH, PhD; Cynthia Stappenbeck[3], PhD; Shahzarin Khan[4], MSc; Anna Woodworth[5], MS; Sanika Panwalkar[1], BS; Ronald Acierno[6], PhD; Amanda K Gilmore[2], PhD

[1]Department of Psychology, University of Wisconsin–Madison, 1202 W Johnson Street, Madison, WI, United States

[2]Department of Health Policy and Behavioral Sciences, School of Public Health, Georgia State University, Atlanta, GA, United States

[3]Department of Psychology, Georgia State University, Atlanta, GA, United States

[4]Department of Applied Health Science, School of Public Health–Bloomington, Indiana University, Bloomington, IN, United States

[5]Anthropology Division, SUNY Polytechnic Institute, Utica, NY, United States

[6]Department of Psychiatry and Behavioral Sciences, McGovern Medical School, UTHealth Houston, Houston, TX, United States

Corresponding Author:
Kate Walsh, PhD
Department of Psychology, University of Wisconsin–Madison, 1202 W Johnson Street, Madison, WI, United States

Abstract

Background: Sexual violence is prevalent, and the consequences can be chronic and impairing. However, few interventions exist to prevent the onset or escalation of posttraumatic stress symptoms and opioid misuse among recent sexual violence survivors.

Objective: This study describes a collaborative process of updating an integrated postsexual assault video and developing an SMS text messaging intervention program with a community advisory board (CAB) of sexual assault survivors.

Methods: Research team members met virtually for six 60‑90-minute meetings with a 5-member CAB of sexual assault survivors with diverse racial and gender identities located throughout the United States. CAB members provided feedback on written documents detailing an adapted video script and newly developed text intervention to address the risk of substance misuse and posttraumatic stress disorder symptoms following sexual assault. CAB members also received SMS text messages to provide feedback from the end-user perspective.

Results: We identified overarching themes to improve relatability (destigmatize and increase awareness of support, reduce technical language, and increase representation in actors), content (increase social support, include substance-related assault, and suggest activities), and wording (normalize different terms for sexual assault and reduce insensitive language) for the video intervention. For the text intervention, we identified themes relating to acceptability (timing, frequency, and format of texts), relatability (having an avatar introduce the program and identifying the study name in messages), content (messaging), and wording (increasing clarity).

Conclusions: Findings reinforce the importance of including community members' perspectives and suggestions to improve the acceptability and relatability of interventions, including the video and SMS text message intervention described here.

(JMIR Form Res 2025;9:e72095) doi:10.2196/72095

KEYWORDS

intervention development; community advisory board; sexual assault; early intervention; text message; survivor

Introduction

Sexual assault (SA) is a significant public health problem that affects more than 1 in 2 women, nearly one-third of men, and 2 in 5 trans people in the United States [1,2]. SA survivors experience numerous deleterious consequences, ranging from physical health conditions such as headaches, chronic pain, and irritable bowel syndrome [3], to mental health conditions including posttraumatic stress disorder (PTSD), depression, suicidality, and substance misuse and disorders [4]. Medical, mental health, lost work productivity, criminal legal, and property damage costs associated with SA amount to more than US $3.1 trillion (2014 US dollars) over survivors' lifetimes in the United States [5].

PTSD and substance misuse are particularly pernicious and frequently co-occurring difficulties following SA [6-8]. Indeed,

XSL•FO
RenderX

meta-analyses suggest that 75% of SA survivors meet criteria for PTSD 1 month after assault and 40% meet criteria 12 months after assault [9]. Opioid misuse correlates strongly with both SA and PTSD [10,11]. Among women seeking a sexual assault medical forensic exam (SAMFE), 21% reported nonmedical use of prescription opioids in the year prior to the assault, and 15.6% reported using them nonmedically in the 6 weeks after the assault [12]. However, few interventions target PTSD and opioid misuse, and none are tailored for the immediate post-SA period.

Early interventions that are perceived positively by SA survivors are key to preventing chronic and impairing symptoms [13]. The Prevention of Post-Rape Stress (PPRS) video is a 9-minute intervention that has been used in 3 randomized controlled trials (RCTs) with SA survivors presenting for a SAMFE [14-17]. The original PPRS video was developed by clinician-researcher experts in SA interventions that drew on three frameworks relevant to trauma, substance misuse, and depression recovery: (1) emotional processing theory [18], which posits that fear responses to trauma-related stimuli can be extinguished through exposure; (2) substance use relapse prevention approaches [19] that encourage avoiding substance misuse activators and instead focusing on coping in ways that are incompatible with substance use; and (3) behavioral activation approaches [20] that encourage engaging in healthy coping activities like spending time with supportive people and taking walks. Therapeutic exposure, relapse prevention, and behavioral activation have strong evidence for efficacy in reducing PTSD [21], substance misuse [22], and depression [23,24], respectively. The PPRS includes a single narrator describing possible reactions to SA, including posttraumatic stress symptoms, substance misuse, and depression, along with ways to address these reactions by recognizing and preventing trauma-related avoidance (ie, therapeutic exposure), recognizing substance misuse activators, and engaging in activities incompatible with substance use such as spending time with people who do not use substances or engaging in activities to fight depression like exercising.

In 2 RCTs, the PPRS was associated with reductions in PTSD and depression for survivors with a prior history of SA [15,25] and reductions in cannabis use among those with preassault use [16,17] when compared to treatment as usual (TAU). TAU consists of the standard provision of a SAMFE that can include injury treatment, pregnancy and sexually transmitted infection prophylaxis, evidence collection, or connection to other referrals and resources. The PPRS video has been associated with reductions in cigarette smoking [26], alcohol use [17], and suicidality [27] compared to TAU. The latter 2 findings have been for survivors with preassault alcohol use and those with heightened distress at the SAMFE.

In a third RCT, a mindfulness video was delivered as an active control condition in comparison to the PPRS and TAU. Compared to TAU, the mindfulness video was effective at reducing prescription opioid misuse 1.5 months post-SA for those with a prior SA history; the PPRS video did not differ significantly from TAU [12]. This finding fits with broader literature suggesting that mindfulness interventions are effective at reducing opioid misuse and chronic pain [28-30]. Mindfulness interventions also have been used to effectively address PTSD and depression among SA survivors [31,32] and have been found acceptable among women with comorbid PTSD and substance use disorder more generally [33]. Therefore, to target PTSD, depression, cannabis use, cigarette use, alcohol use, suicidality, and opioid misuse, it is essential to incorporate mindfulness components into the PPRS video.

Despite the positive effects of PPRS, our finding that the video is more effective for survivors with a previous history of SA [12,15,25] suggests that it may be important to increase relevance to survivors broadly in future iterations of the video. Models of intervention development emphasize the importance of adapting interventions prior to rigorous effectiveness testing [34], and formative work like eliciting feedback from a community advisory board (CAB) can be an important initial step. CABs often consist of community members who have lived experience and expertise relevant to the topic in question. In this case, CAB members with SA and SAMFE experiences could provide feedback that increases both the relevance and acceptability of the video. Indeed, CABs have been shown to improve the quality and relevance of research for the communities the research impacts [35]. CABs have successfully informed interventions for specific populations in prior work [36,37] and thus could enhance the relevance of the intervention for SA survivors in this study.

Another challenge was that, understandably, some survivors were distressed at the time of the SAMFE [38], which happens in the acute postassault time frame, and may not have absorbed the information presented in PPRS as readily. Thus, providing the same core intervention content in the days and weeks following the SAMFE may help improve retention and subsequent outcomes for these individuals. An SMS text messaging program delivered after the SAMFE may be a viable strategy to push this content from the video over a longer period of time and via a different medium. A November 2024 report from the Pew Research Center found that 98% of Americans own a cellphone and 91% own a smartphone [39], so an SMS text messaging program could be viable. Indeed, SMS text messaging programs have been used to successfully improve mental health and substance use symptoms [40]. Indeed, a brief intervention for PTSD and alcohol misuse among SA survivors found that delivering 4 brief SMS text messages was highly acceptable to survivors and associated with reductions in PTSD symptoms, weekly drinks, and heavy episodic drinking at 8-week follow-up [41]. However, to our knowledge, an SMS text message program to address comorbid PTSD and other substance use has not been evaluated, particularly as a follow-on intervention amplification method after the acute timeframe-based intervention.

This study describes a formative collaborative process of working with a CAB of SA survivors to refine content for a new video called Skills Training in Active Recovery (STAR). STAR integrates the principles of the PPRS and mindfulness videos to better address PTSD, depression, and substance misuse including opioid misuse. To make STAR and TextSTAR more relevant and acceptable to women and nonbinary SA survivors, the CAB provided feedback on ways to make the interventions more relatable, accessible, understandable, and helpful. We analyzed feedback from meetings through thematic analysis

XSL•FO
RenderX

and discussions with the broader research team and made decisions collectively about how to refine the interventions to be responsive to this feedback. This was an important formative step before evaluating the acceptability and preliminary effectiveness of STAR and TextSTAR in a pilot RCT with recent SA survivors presenting for a SAMFE.

Methods

Recruitment of CAB Members

The target population and study protocol were considered when recruiting and selecting CAB members. Because SA disproportionately impacts women and those who identify outside the gender binary [1,2], those who present to SAMFE programs for care are disproportionately women [42], and our prior trials have only included women [16,17], we focused efforts on refining the content of this video and SMS text messaging program with women and nonbinary people specifically. Additionally, because people of color in the United States experience SA at high rates [1] and are often excluded from research, it was important to ensure that CAB members were racially diverse people drawn from across the United States. Given the focus on opioid and other substance misuse, we were interested in including those with substance misuse experiences. Additionally, because participants in the broader research study would be recruited at a SAMFE, we sought CAB members who had some experience with receiving post-SA medical care.

The study team developed a flyer that we distributed via email to 11 organizations across 10 states. The flyer stated that we were seeking SA survivors to provide feedback on an intervention for recent SA survivors. Those who were interested could scan the QR code or click the study link in the advertisement to complete a short screening survey that included questions about demographics, experiences of unwanted or nonconsensual sexual touching or oral, anal, or vaginal penetrations, medical care and services received after the unwanted or nonconsensual experience, substance use, and interest or willingness to be on a CAB focused on improving an intervention. We used these responses to select CAB members who were women or nonbinary, racially and ethnically diverse, and had SA, SAMFE, and varied substance misuse experiences, including opioid misuse.

CAB Participant Characteristics

CAB members were 5 women or nonbinary survivors of SA located in 4 states across the United States. Our budget for the project was limited, so we opted to recruit a smaller number of CAB members to participate regularly in a larger number of meetings. We wanted CAB members to have the time to share feedback and be heard, which the smaller group size accomplished. The smaller CAB allowed for more engagement with the material and more meaningful opportunities for CAB members to provide feedback and discuss suggestions with each other. CAB members were all 18 years or older and ranged in age from 25 to 54 years. Members had diverse racial identities, including White, South Asian, African American, and American Indian.

Ethical Considerations

The proposed CAB process to refine these interventions was considered “not human subjects research” (IRB 2023 - 1219).

CAB Process

CAB meetings were held in a HIPAA (Health Insurance Portability and Accountability Act)-compliant, passcode-enabled Zoom room. Meetings were 1 to 1.5 hours long and were audio-recorded and transcribed by an author (SP). Because we were not asking CAB members to share personal experiences with each other but rather to share feedback on the video and SMS text messaging program, we opted to audio-record meetings to ensure that we accurately recorded specific and detailed feedback. Recruitment emails and flyers clearly stated to potential CAB members that they would not be asked to share details of their nonconsensual sexual experiences with others and that we would be audio-recording meetings, so those who did not feel comfortable could make an informed decision not to participate. To ensure safety and comfort for all CAB members, the screening form confirmed that CAB members would be willing to have their cameras on and ensure that they were in a private space during meetings. We also developed a protocol that we reviewed with CAB members individually and in the first meeting about how to handle others unexpectedly entering their space or needing to leave during a CAB meeting. Although we set the expectation that CAB members did not need to share trauma details, the first author (KW), who led the meetings, is a licensed clinical psychologist and offered to speak with any CAB members after meetings if they found any of the content or discussions challenging or needed referrals for mental health resources.

Prior to each meeting, CAB members typically received slides or other content (eg, a video script, a document with SMS text messaging program messages) to familiarize themselves with the content we would be discussing. Two members of the research team (KW and SP) met with the CAB on 3 occasions to elicit feedback on the video script and on 3 occasions to elicit feedback on the texting program (Figure 1). CAB members were compensated US $50/hour for time spent in meetings, and they received an optional US $100 bonus if they agreed to receive the study text messages and provide feedback from the end-user perspective.

XSL•FO
RenderX

Figure 1. CAB meeting tasks and activities between meetings. CAB: community advisory board.

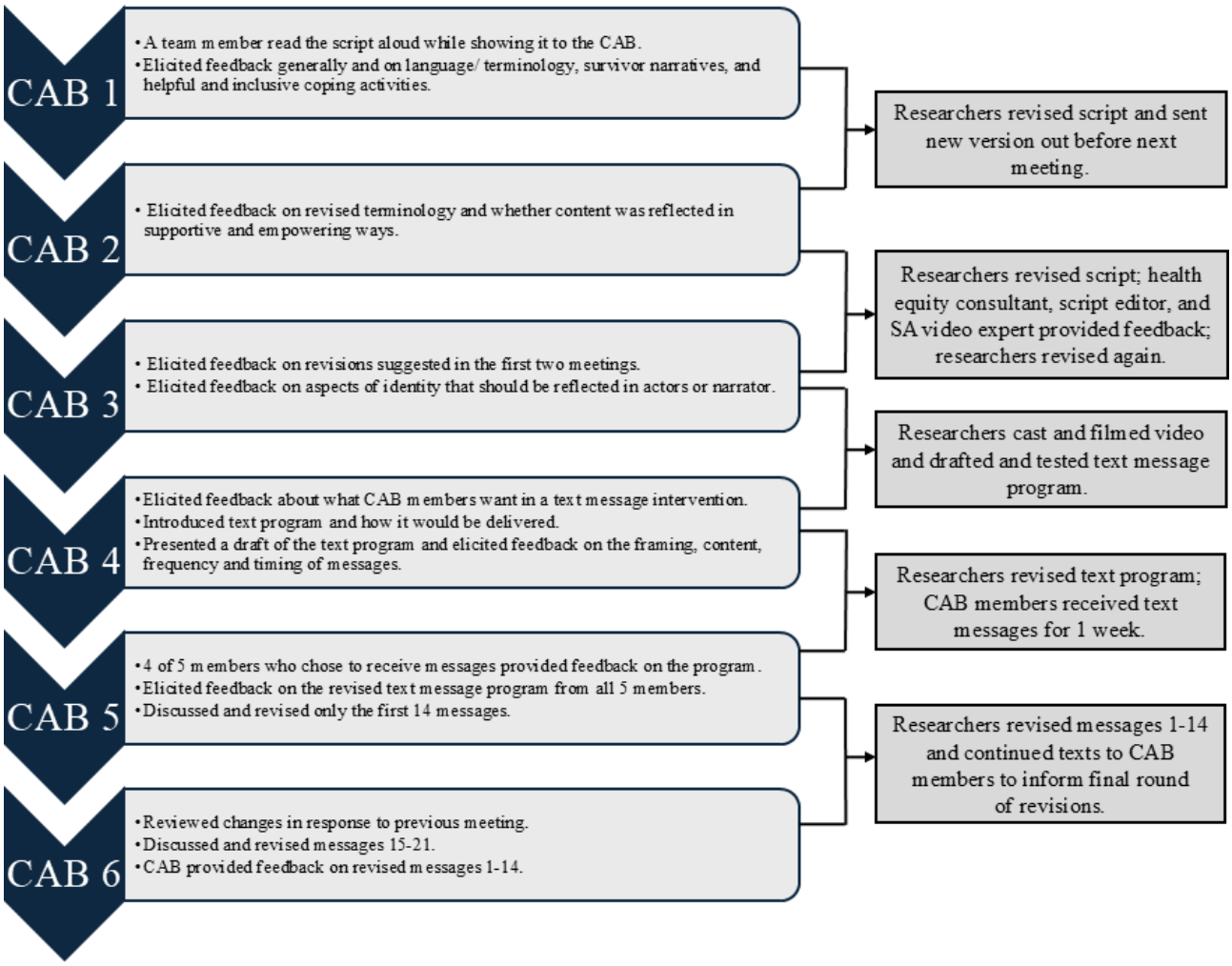

Iterative Video Intervention Refinement Process

The 5-member research team began refining the original video script to incorporate (1) SA survivor narratives to educate and normalize common experiences, (2) coping skills and tips told from survivor perspectives and modeled on screen in some cases (eg, how to do gradual exposure), (3) mindfulness skills (eg, urge surfing) to address opioid misuse specifically, and (4) distress tolerance skills (eg, temperature, intense exercise, and paced breathing) [43]. We also developed an empowering ending to the video that involved the narrator, actors, and many extras with diverse backgrounds and identities to highlight that anyone can experience SA and there is a large community of support.

In the first meeting, a team member read the script aloud while showing it to the CAB and asked for general feedback as well as feedback on language or terminology, examples of things that should be represented in survivor narratives, and activities survivors could suggest that would feel both helpful and inclusive. In between meetings, the research team revised the video script based on CAB feedback and sent these revised versions out to the CAB, so they could have a few days to review the changes before the next meeting. In the second meeting, we revisited the script to obtain additional feedback on some of the changes we made to ensure that revised terminology felt appropriate and that content was reflected in supportive and empowering ways. In the third meeting, we revisited themes from the first 2 meetings to ensure additional edits appropriately reflected earlier suggestions. We also asked the CAB for feedback to inform casting decisions prior to filming the video. Specifically, CAB members were asked to describe who they envisioned in each of the roles and what aspects of identity were important to have included or represented in the video.

Additional Consultation

Based on CAB feedback, we also had a health equity consultant review and edit the video script to ensure that it was inclusive for neurodiverse SA survivors. A scriptwriter then provided edits to make the video more conversational. Next, an SA expert consultant who developed the original video script reviewed and provided edits on the final version.

SMS Texting Program Refinement Process

The study team constructed approximately 3 weeks of SMS text messages using content and phrasing from the video. This material was organized into supportive, educational, and skills-focused messages, and we evaluated a version that had content types grouped by week and a version that had content types woven throughout. In the fourth meeting, we introduced the SMS text messaging intervention and how team members planned to implement it. We first elicited general feedback about what CAB members would like to see from an SMS text

XSL•FO
RenderX

messaging intervention. We then presented a draft of the SMS text messages the research team had created and asked for feedback on the framing of the program, content type, message ordering, and frequency and timing of messages.

In between meetings, the research team received SMS text content to gain clarity on how it felt to receive messages and refine the program with internal feedback. In the first CAB meeting, we described how the research team envisioned the texting program would be delivered and what alternatives had been considered to provide context. Specifically, although an interactive texting program staffed by those with SA advocacy expertise could be incredibly helpful and tailored to the specific concerns survivors experienced in a specific moment, we did not have the personnel or funding available to meet this need. Additionally, we had considered sending short video links (edited from the longer video) via texting. However, based on research showing the value of brief texting for SA survivors with PTSD and alcohol misuse [41], we opted to keep written texts limited largely to 160 characters or less (ie, the character limit for a single SMS text message). In the CAB meeting 4, which was the first that focused on feedback on the texting program, we asked CAB members for (1) general feedback on the structure of the proposed program, (2) feedback on the frequency and timing of messages, and (3) what would make the CAB likely to open or read the messages and try some of the skills. We also showed CAB members a few sample messages and elicited feedback. Prior to the fifth meeting, we invited CAB members to receive a week's worth of SMS text messages so we could get feedback about the end-user perspective of receiving the messages, including the order in which the first week of messages was delivered. Then, in the fifth meeting, 4 out of 5 members who chose to receive messages started by providing feedback on the program. One CAB member had a family emergency and opted not to receive SMS text messages during these weeks. A team member then presented the revised 21-day SMS text messaging program and elicited content-related feedback from all 5 members. In the allotted meeting time, we only discussed and revised the first 14 messages. Prior to the sixth meeting, the research team revised messages 1 - 14 based on CAB feedback and continued sending messages to CAB members to inform the final round of revisions. We asked for feedback on how it felt to receive the messages, the timing of when to receive messages during the day, whether any messages did not sit well with participants, and whether any messages should be edited. We also went through the first 2 weeks of SMS text messages and addressed wording, content, and ordering changes. For the third SMS text message meeting, CAB members received an additional week of SMS text messages and weighed in on the content, wording, and ordering of the third week of messages.

Reflexive Thematic Analysis

Using the 6 steps of reflexive thematic analysis [44,45], the first author (KW), who led the CAB meetings, read and reread transcripts to familiarize herself (step 1) with the data and took notes both during CAB meetings and while reading. Next, she generated initial codes (step 2) and searched for themes (step 3). With the research team, she reviewed themes (step 4) and defined, named, and organized themes according to intervention content (video vs text; step 5). These codes and themes were iteratively discussed with the broader research team, and an initial draft of the results was drafted, shared with CAB members, and iteratively refined (step 6). Additionally, because the purpose of the meetings was to improve the interventions, we have described the study team's response to the suggested changes.

Reflexivity Statement

Braun and Clarke [46] eschew a one-size-fits-all checklist for qualitative research and instead emphasize *reflexive openness* in their values-based guidelines for qualitative research. Predominantly, the women research team is composed of scholars who largely have doctoral-level training in psychology and public health. Several team members are licensed clinical psychologists who have provided trauma-focused and substance misuse–focused therapy for SA survivors and who strive to include the communities they serve in the design and production of research. Team members also have extensive experience collaborating with SAMFE programs to conduct research to better support SA survivors in the immediate aftermath of SA. Collaborating with the CAB illuminated biases in how the research team viewed content and messaging that could, in practice, be less acceptable or even off-putting for survivors. We practiced openness to receiving feedback and being responsive to concerns raised throughout the intervention refinement process by engaging in open discussion with the shared understanding that the research team and CAB both wanted to produce the most helpful possible content.

Results

Video Intervention Suggestions

Suggestions for video intervention improvements can best be described using a funnel or upside-down triangle where at the broadest level, CAB members made suggestions to improve the *relatability* of the video; at the middle level, they made suggestions to improve the *content* of the video, and at the bottom level, their suggestions focused on improving the *wording* of the video. Specific subthemes within each domain are identified in the text and Table 1.

XSL•FO
RenderX

Table . Summary of community advisory board recommendations and research modifications for the video and text messaging interventions.

Domain	Challenge	CAB[a] suggestion	Researcher modification
Video challenge			
Relatability	Technical language	• Do not use "mindfulness"	• Described techniques but did not label "mindfulness"
	Representation in actors	• Seek actors with visible physical disability	• Asked casting company for extra with visible disability
Content	Include messaging about social support	• Explicit suggestions to seek support from one's community	• Created more dialogue between the actors telling their stories in the video to model support
	Include substance-related assault	• Substance-related assault notably missing	• Had actor describe a substance-involved experience and feelings of self-blame
Wording	Normalize use of various SA[b] terms	• Use the word "violated" in addition to "rape/sexual assault/trauma" • Include discussion from the narrator and actors about different ways to talk about what happened • "Harm" too vague	• Narrator expressly says there are many different ways to talk about what happened and actors use different terms when describing their own experiences including "rape, sexual assault, and violated"
	Reduce insensitive language	• Use "challenge yourself to" instead of "do this" to recognize difficulty and reduce ableism	• Replaced recommendations to "do this" with "challenge yourself to"
Text program challenge			
Acceptability	Responding to messages	• Do not require responses to keep "low pressure"	• Removed response requirements
	Frequency	• No more than once/day	• Created a daily text program
	Timing	• Might be burdensome to anticipate when they'll need support • Transition from late afternoon to early evening	• Sent around 4 pm each day but ask for feedback about timing
	Formatting	• Prefer opening question followed by tip or suggestion preferred	• Use mostly opening questions followed by tip or suggestion preferred
Relatability	Introducing the program and texts	• Have avatar introduce program so there is a "person" sending texts • Start texts with "STAR[c] study:"	• Crafted introductory text from "Shari from the STAR study" with photo • Began each text with "STAR Study:"
Content	Messaging	• Change the content of the educational text	• Removed language about symptoms passing with time and included validating statement
	Ordering	• Place general text before specific	• Reordered so general text came before specific

XSL•FO
RenderX

Domain	Challenge	CAB[a] suggestion	Researcher modification
Wording	Clarity	• Clarify tempo of breathing • Expand on verb "use"	• Added the word "slowly" before counting instructions during breathing exercise • Added words "drugs or alcohol"
	Specificity	• "Approach" does not have teeth • "Danger" is not specific	• Changed to "confront" • Changed to "violation or traumatic event"

[a]CAB: community advisory board.

[b]SA: sexual assault.

[c]STAR: Skills Training in Active Recovery.

Relatability

Suggestions to improve the relatability of the video centered on ways to make it more appealing and survivor-centered. Examples included specific suggestions to destigmatize SA experiences and increase awareness of support, reduce technical language to make the video more appealing and easier for an average person to understand, and ensure diversity and representation in actors and in the kinds of experiences and narratives reflected in the video.

Destigmatize Assault and Increase Awareness of Support

A CAB member suggested adding a statement to the beginning of the video, destigmatizing the assault and highlighting the community of support:

> *I think in the beginning like a plain language statement of like, you're not alone. Like there is a community. What happened to you is real, and you should feel empowered to share your story because the more that we don't talk about this, the more it happens in silence to other people. Not to say it's our responsibility to stop abusers, but like, it's only a positive feedback loop. By us harboring the guilt and not talking about it, basically.*

To address this suggestion, we have the survivor-actors say, "What happened to you was not ok, it was not your fault, and you are not alone" at the beginning of the video.

Reduce Technical Language

The CAB noted that some of the content from the original video was technical and potentially difficult to follow for individuals in acute distress. For example, CAB members recommended not using the term "mindfulness" and instead focusing on the techniques:

> *I think there's a lot of mention of "mindfulness. I don't know, maybe it's just like a generational thing, where it's a lot of focus right now for a lot of things. Where right now if somebody tells to me, oh, am I having an anxiety attack? Why don't you try mindfulness? It just feels a bit less. But the techniques that was said was very specific. So it's just about rephrasing, you know, if somebody comes to me and say this is something bad that has happened to me. Why didn't you try this technique instead of mentioning mindfulness?*

The research team removed the term "mindfulness" from the script and instead described the actual techniques.

Increase Diversity and Representation in Actors

One goal for the new video was to move away from a single narrator describing intervention content to a format that allowed the narrator and multiple survivor-actors to personalize the information presented and show a breadth of experiences and coping strategies. The research team planned to show a narrator and actors who varied in terms of race and gender expression to make the content feel relatable to a larger swath of the population. However, the CAB expanded on this intention by recommending that we include actors of all shapes and sizes and at least one actor with a physical disability. These conversations allowed the research team to write casting calls to yield a diverse cast, seek an extra with a visible physical disability, and prioritize representation in the selection of actors.

Content

At the middle level, CAB members made suggestions to enhance or include new content that they viewed as highly relevant to the healing process. Content suggestions included promoting efforts to seek social support early and often, describing substance-facilitated SA, and providing feedback on suggested activities to cope, distract from substance use, or increase positive feelings.

Increase Social Support Content

CAB members noted that none of the skills the video promoted could be effectively implemented without social support. For example, one CAB member said:

> *Not too much of what was said or suggested is gonna more than likely happen without support first. And I only heard support in like the middle part of it as being a big, you know, or mentioned, it wasn't like support was like the number one because first of all, without support, with no support or...when you get a medical attention, if they don't believe you, you're more than likely to turn to those drugs and feelings. Yeah, but so the...the part that I felt was missing was the stressing support as number one. And, like, one of the first priorities.*

XSL•FO
RenderX

In response to suggestions to increase social support content, we had the narrator and actors emphasize the importance of social support early and throughout the video. Specifically, we created more dialogue between the actors telling their stories in the video and responding empathically to others who described what happened to them to model what support could and should look like. We also explicitly encourage seeking out supportive community members throughout the video and show clips of survivors calling supportive people and meeting up with them for a walk or other activity. We also end the video by encouraging survivors to consider other supports, including talking to a professional, if they find that their symptoms are not improving over the next few weeks.

Include Substance-Involved SA Content

CAB members raised that substance-involved assaults were especially difficult to identify as assault, and shame and guilt were common immediate reactions. They believed it was critical to include some acknowledgment of these kinds of experiences within the survivor narratives to be clear that this is a type of assault:

> *I was surprised to not see anything in it about or for survivors who were under the influence during their assault. I thought that might be applicable. I think it would be helpful, but I don't have like a clear thought of the messaging around it yet.*

Based on this suggestion, the study team revised the script to include a narrative from an actor-survivor that included feeling a lot of shame and self-blame for experiencing assault while using substances.

Activity Suggestions

One specific question the research team had for CAB members was which coping activities we should highlight and which may be less useful to emphasize. Organized religion (eg, going to church) was identified as an activity to exclude from the examples we provided:

> *It suggests like things you can do to help you when you're feeling like inundated or overwhelmed by feelings and it mentions prayer and going to church. While that is definitely helpful for some people, that's probably something that is better to stay away from considering this is a general audience. Like you don't know what the person is bringing, right, because a lot of people in that moment have given up on religion and so I think it's really insensitive to suggest that because, like, what, you pray for what?...those particular words really take away from the authenticity of what is trying to be accomplished here. So I guess trying to stay away so much from religion, sorry, structurally religious suggestions. Spiritually, I guess, reflection, meditation I think, like, that's totally different than prayer and going to church because I also think it's, you know, specific to Christianity.*

Wording

At the narrowest level, CAB members made wording suggestions about terminology to use to describe experiences of SA and to minimize potentially insensitive language.

Normalize Use of Various SA Terms

CAB members all used different terms to describe their SA experiences. They encouraged us to include acknowledgment of different terms to normalize the language survivors may use for SA experiences:

> *There were a lot of, um, like emotions and tactics that I know during my first assault I would not have like understood, like my emotions were so out of whack and I wouldn't have been able to make sense of a lot of it. Which I think might be why it resonated more with people who, um, had an earlier experience. I think kind of going back and focusing on the feelings of it rather than the term, I don't know, I'm not married to this term that I'm about to give, but like 'being violated.' I think that was a feeling that resonated with me and was clear even though I didn't understand it at the time. So maybe going back to the basics with some stuff like that.*

Another CAB member said:

> *I think telling survivors that they can talk about what happened without labeling it, without saying this is that or it felt like that, and just identifying that your feelings are valid and you can feel them. You don't have to do anything, you don't have to make any decisions. You can still talk to somebody about what happened without saying what it was, cause I think a lot of the time people don't wanna talk about it because they're not sure if it was assault, or if it was worth talking to anybody about and it's like, you can just talk about that without knowing. That's completely real and valid and shared by so many people.*

The research team addressed this concern by including the offered terms in narrative examples from the survivor-actors and using these terms interchangeably throughout the video.

Reduce Potentially Insensitive Language

CAB members noted some places where the video unintentionally came off as insensitive or ableist. For example, one CAB member said:

> *There is some language at the end about like make yourself get out of bed and like don't blame yourself that also rely on that like there's a standard way of doing things, that there is a standard level of capability. Like, saying, like, you know, make yourself do these things, that's a very neurotypical piece of advice and later, when I thought about like how to rephrase that I liked "challenge yourself to like go beyond your, your, your like bubble of comfort."*

Another CAB member noted:

XSL•FO
RenderX

> *When I was reading don't blame yourself at the end, umm, I was thinking one thing that to me feels more powerful is saying that this is not your fault. Umm, I think that that might feel a little bit more reassuring. And I think that that particularly applies to the substance-use, umm, scenarios.*

In response to these comments and suggestions, we changed our language to "challenge yourself" and "it's not your fault." Similarly, there were suggestions to change older, colloquial terms that have come to have a negative or stigmatizing connotation to less stigmatizing terminology. For example, a CAB member said:

> *I also want to note that there was a part early that said, like, we want to acknowledge that what you are feeling is real, you're not going crazy, and I also think that's a bit ableist because that's very loaded language too. I think that could be something more like, umm, you know, this is not all in your head.*

To address this language concern, we used the CAB member's language of "it's not all in your head."

SMS Texting Intervention Suggestions

Acceptability

Responding to Messages

The budget for the pilot would not allow for real-time monitoring and responding to SMS texts, which is clarified to participants in the informed consent as well as the introductory text of the SMS text messaging intervention. However, the team initially considered whether to assess message engagement by encouraging participants to respond to messages, including by typing back answers to questions or just rating the content of the message using "thumbs up" or "thumbs down" emojis. Although the CAB liked the "thumbs up" and "thumbs down" option, they preferred not requiring any sort of response because it was "low pressure." To ensure that no pressure was placed on participants in the SMS text messaging program, we ultimately opted to ask about the helpfulness of the SMS text messaging program in weekly surveys instead of after each message.

Timing

In an ideal world, survivors would be able to choose when to receive messages, but not all people know when they might need more support throughout the day, so we asked CAB members to weigh in on the preferred time of day for SMS text message delivery. While some CAB members had no preference for time of day, a few noted that late afternoon, when transitioning from the day into the evening, where plans might be less structured, could be a helpful time to receive messages encouraging them to seek support, although this assumes that people have more structured days and less structured evenings. We tested the late afternoon time-out with CAB members when they received SMS text messages; although one indicated that receiving messages in the morning might help her get out of bed on tough days, nearly everyone else reported liking the late afternoon time. Thus, we plan to initially pilot the SMS text messaging program being delivered in the late afternoon.

Frequency

Prior to receiving the SMS text messages, CAB members noted that they would not like to have messages more than once per day, and they were not sure that they would even want daily SMS text messages. After receiving daily SMS text messages for 2 weeks, we reassessed, and they all appreciated the daily SMS texts and thought once per day was the right frequency.

Format of Messages

CAB members viewed differently structured sample messages and strongly preferred messages that opened with a question and then provided a tip or suggestion (eg, Feeling down? Challenge yourself to get out of the house and spend time with supportive people even when you don't feel like it).

Relatability

Introducing Text Messages

The CAB recommended that we consider creating an avatar or identity that survivors would connect to the texting program to increase the acceptability and likelihood of engaging with the texting content. To address this suggestion, we had the narrator of the video film a brief clip introducing the texting program, and we sent an introductory message from "Shari from the STAR Project" with a photo of the narrator to accompany the SMS texts. CAB members preferred the introductory SMS text message from Shari as opposed to clicking a video link to view the introduction. CAB members were agnostic about including a photo, so we excluded it. Members of the research team proposed having participants save the STAR study phone number as a contact, and CAB members suggested having the messages open with "STAR:" to orient them to the content. We included all of these recommendations in the version of the text program we formally tested.

Content

Messaging

One CAB member experienced additional violence just before receiving the test SMS text messages and reported that the messages were particularly helpful and validating. Another said:

> *I especially liked the ones where they, where you gave specific advice, like steps to do. I thought that was a nice change of pace for a lot of automated messages. It didn't feel like in the void.*

However, CAB members also had suggestions for improvement in content. For example, one of the educational SMS texts was intended to provide information about the fight, flight, or freeze response but differed from other texts in that there was not a corresponding tip or action suggested. The research team tried out various "tips" that included just letting natural reactions happen, but the CAB suggested that a message that was simply about validating those experiences was sufficient. Based on this feedback, we decided not to include a tip and instead added validation.

Ordering

CAB members also recommended switching the ordering of SMS text messages so a general text about avoidance came before a more specific text about how to address avoidance.

XSL•FO
RenderX

The study team also received SMS text messages on multiple occasions and recommended including more supportive SMS text messages early in the program with more skill and activity suggestion text messages later in the program.

Wording

In the research team's efforts to make messages brief, we occasionally risked being unclear or lacked adequate specificity to be helpful.

Clarity

The CAB identified places where certain messages lacked clarity or could be misinterpreted. For example, one SMS text message read, "Trouble sleeping? Close your eyes and breathe in while you count to 4, hold your breath while you count to 4, then release your breath while you count to 4." A CAB member noted:

> *You haven't given any point of reference for how fast or slow you're supposed to be counting, so theoretically that could actually make it worse if you're counting way faster than you're supposed to be or in the alternative it could actually make someone hold their breath.*

We revised the message to include the word "slowly" before "breathe." Another SMS text message initially read, "Want to use? Some strategies people use to just wait out urges include taking long, slow deep breaths in and out, or picturing a peaceful scene." One CAB member felt that the verb "use" was specific to a certain type of drug use and recommended that we clarify the question to include "drugs or alcohol." We made this change to improve clarity and relevance.

Specificity

The CAB also identified messages that lacked adequate specificity. For example, one SMS text message read, "Approaching scary or difficult things can be hard," and a CAB member noted that the word approach "had no teeth" and a word like "confront" would be more aligned with what we were trying to convey.

Discussion

Summary

This paper describes a researcher-CAB collaboration to develop a 15-minute video and 21-day SMS texting program to prevent the onset and escalation of PTSD symptoms and opioid misuse for recent SA survivors who received a SAMFE. A prior 9-minute video was associated with reductions in a variety of mental health and substance use difficulties [17,25]. However, the revised video described here was designed to resonate more with recent survivors by including actors who described different kinds of experiences and different coping strategies in addition to a narrator, more relaxation and mindfulness skills to address opioid misuse (eg, urge surfing), and more empowering content. Consistent with other community-engaged approaches to developing interventions [36,37], CAB members provided feedback to improve the relatability, content, and wording of the video and SMS text messaging program to ensure greater acceptability and relevance to survivors with various backgrounds and experiences. The resulting interventions expand on a growing body of early interventions for recent SA survivors that typically target PTSD [13,47] and address comorbid substance use and PTSD symptoms [41]. Early interventions like the video and SMS text messaging program can be relatively easily administered and low cost, which has important implications for broad dissemination [47].

Some of the most critical suggestions the CAB made were around including more empowering statements and language to reduce the stigma of experiencing SA that often leads people to avoid talking about what happened [48], forego needed health care [49], and even experience revictimization [50]. CAB suggestions to prioritize seeking social support early and throughout the video and SMS text messaging program also fit with a large literature suggesting that social support is critical to recovery among sexual violence survivors [51-54]. However, some survivors may require more support than a video or SMS text messaging program can offer in developing a network of supportive people. Considering additional ways to enhance social support will be crucial for future work.

Limitations and Future Research

Although CAB members were diverse across a variety of dimensions, including gender, SA and substance misuse histories, and geographic locations, they comprised a small group of survivors at different stages of healing. Data on acceptability from a separate sample of survivors with more recent SA experiences will be critical to establish the utility of the interventions, and a fully powered RCT will be necessary to establish the efficacy of the interventions. Additionally, the CAB and the intervention refinement process only focused on women and nonbinary people, in part, because these groups are substantially more likely than men to seek SAMFE care [42]. Men also experience SA [1] but may experience shame and disempowerment that are entangled with feelings of emasculation due to rape myths that SA victimization does not happen to men [55]. Instead of trying to develop a one-size-fits-all intervention that risks not feeling personalized or relatable to anyone, a CAB of men that is led by a man could elicit feedback about how supportive content might be framed so that male survivors find it relatable and acceptable. Additional work will be needed to develop a video and SMS text messaging intervention appropriate for men. This study also did not include non–English-speaking individuals or those from non-US contexts. CABs should be developed for these populations to refine the messaging in ways that feel tailored to these groups to achieve maximum benefit.

Although SA tends to be concentrated among adolescents and young adults [1] who are likely to be more familiar with technology, the development of a texting program and reliance on Zoom to hold CAB meetings to connect with survivors from across the country may mean we are hearing from and creating interventions for those who are more technologically savvy. Two of our CAB members who stated they were not technologically savvy regularly joined Zoom from their smartphones and had highly positive feedback on the texting program. Additionally, in our pilot trial, we have budgeted for the small percentage of survivors who lack a phone to receive

XSL•FO
RenderX

a prepaid smartphone and service for the duration of the study. Nonetheless, future efforts should be made to ensure equitable access to both the intervention development or refinement process and the interventions.

Conclusions

This study highlights the importance of collaborating with people with lived experiences of violence to develop relevant and acceptable early interventions to prevent the onset and escalation of mental health and substance use difficulties.

Acknowledgments

Research reported in this publication was supported by the National Institute on Drug Abuse under grant number R61DA059897 (to KW). Content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflicts of Interest

None declared.

References

1. Basile KC, Smith SG, Kresnow MJ, Khatiwada S, Leemis RW. The national intimate partner and sexual violence survey: 2016/2017 report on sexual violence. : National Center for Injury Prevention and Control, US Centers for Disease Control and Prevention; 2022 Jun URL: https://www.cdc.gov/nisvs/documentation/nisvsReportonSexualViolence.pdf [accessed 2025-07-25]
2. Messinger AM, Guadalupe-Diaz XL, Kurdyla V. Transgender polyvictimization in the U.S. Transgender Survey. J Interpers Violence 2022 Oct;37(19-20):NP18810-NP18836. [doi: 10.1177/08862605211039250] [Medline: 34463599]
3. Basile KC, Smith SG, Chen J, Zwald M. Chronic diseases, health conditions, and other impacts associated with rape victimization of U.S. women. J Interpers Violence 2021 Dec;36(23-24):NP12504-NP12520. [doi: 10.1177/0886260519900335] [Medline: 31971055]
4. Dworkin ER, Menon SV, Bystrynski J, Allen NE. Sexual assault victimization and psychopathology: a review and meta-analysis. Clin Psychol Rev 2017 Aug;56:65-81. [doi: 10.1016/j.cpr.2017.06.002] [Medline: 28689071]
5. Peterson C, DeGue S, Florence C, Lokey CN. Lifetime economic burden of rape among U.S. adults. Am J Prev Med 2017 Jun;52(6):691-701. [doi: 10.1016/j.amepre.2016.11.014] [Medline: 28153649]
6. Berenz EC, McNett S, Paltell K. Development of comorbid PTSD and substance use disorders. In: Posttraumatic Stress and Substance Use Disorders: Routledge; 2019:11-27. [doi: 10.4324/9781315442648-2]
7. Brady KT, McCauley JL, Back SE. The comorbidity of post-traumatic stress disorder (PTSD) and substance use disorders. In: Textbook of Addiction Treatment: International Perspectives: Springer; 2021:1327-1339. [doi: 10.1007/978-3-030-36391-8_93]
8. Ullman SE, Relyea M, Peter-Hagene L, Vasquez AL. Trauma histories, substance use coping, PTSD, and problem substance use among sexual assault victims. Addict Behav 2013 Jun;38(6):2219-2223. [doi: 10.1016/j.addbeh.2013.01.027] [Medline: 23501138]
9. Dworkin ER, Jaffe AE, Bedard-Gilligan M, Fitzpatrick S. PTSD in the year following sexual assault: a meta-analysis of prospective studies. Trauma Violence Abuse 2023 Apr;24(2):497-514. [doi: 10.1177/15248380211032213] [Medline: 34275368]
10. Austin AE, Short NA. Sexual violence, mental health, and prescription opioid use and misuse. Am J Prev Med 2020 Dec;59(6):818-827. [doi: 10.1016/j.amepre.2020.06.019] [Medline: 33220753]
11. Ecker AH, Hundt N. Posttraumatic stress disorder in opioid agonist therapy: a review. Psychol Trauma 2018 Nov;10(6):636-642. [doi: 10.1037/tra0000312] [Medline: 28758767]
12. Gilmore AK, Walsh K, Frazier P, et al. Prescription opioid misuse after a recent sexual assault: a randomized clinical trial of a video intervention. Am J Addict 2019 Sep;28(5):376-381. [doi: 10.1111/ajad.12922] [Medline: 31242340]
13. Dworkin ER, Schumacher JA. Preventing posttraumatic stress related to sexual assault through early intervention: a systematic review. Trauma Violence Abuse 2018 Oct;19(4):459-472. [doi: 10.1177/1524838016669518] [Medline: 27681005]
14. Miller KE, Cranston CC, Davis JL, Newman E, Resnick H. Psychological outcomes after a sexual assault video intervention. J Forensic Nurs 2015;11(3):129-136. [doi: 10.1097/JFN.0000000000000080] [Medline: 26291847]
15. Resnick H, Acierno R, Waldrop AE, et al. Randomized controlled evaluation of an early intervention to prevent post-rape psychopathology. Behav Res Ther 2007 Oct;45(10):2432-2447. [doi: 10.1016/j.brat.2007.05.002] [Medline: 17585872]
16. Resnick HS, Acierno R, Amstadter AB, Self-Brown S, Kilpatrick DG. An acute post-sexual assault intervention to prevent drug abuse: updated findings. Addict Behav 2007 Oct;32(10):2032-2045. [doi: 10.1016/j.addbeh.2007.01.001] [Medline: 17275198]

XSL•FO
RenderX

17. Walsh K, Gilmore AK, Frazier P, et al. A randomized clinical trial examining the effect of video-based prevention of alcohol and marijuana use among recent sexual assault victims. Alcohol Clin Exp Res 2017 Dec;41(12):2163-2172. [doi: 10.1111/acer.13505] [Medline: 28940320]
18. Foa EB, Feske U, Murdock TB, Kozak MJ, McCarthy PR. Processing of threat-related information in rape victims. J Abnorm Psychol 1991;100(2):156-162. [doi: 10.1037/0021-843X.100.2.156]
19. Rawson RA, Obert JL, McCann MJ, Marinelli-Casey P. Relapse prevention models for substance abuse treatment. Psychotherapy: Theory, Research, Practice, Training 1993;30(2):284-298. [doi: 10.1037/0033-3204.30.2.284]
20. Lewinsohn PM. The behavioral study and treatment of depression. In: Progress in Behavior Modification: Elsevier; 1975:19-64. [doi: 10.1016/B978-0-12-535601-5.50009-3]
21. McLean CP, Levy HC, Miller ML, Tolin DF. Exposure therapy for PTSD: a meta-analysis. Clin Psychol Rev 2022 Feb;91:102115. [doi: 10.1016/j.cpr.2021.102115] [Medline: 34954460]
22. Irvin JE, Bowers CA, Dunn ME, Wang MC. Efficacy of relapse prevention: a meta-analytic review. J Consult Clin Psychol 1999 Aug;67(4):563-570. [doi: 10.1037//0022-006x.67.4.563] [Medline: 10450627]
23. Cuijpers P, van Straten A, Warmerdam L. Behavioral activation treatments of depression: a meta-analysis. Clin Psychol Rev 2007 Apr;27(3):318-326. [doi: 10.1016/j.cpr.2006.11.001] [Medline: 17184887]
24. Stein AT, Carl E, Cuijpers P, Karyotaki E, Smits JAJ. Looking beyond depression: a meta-analysis of the effect of behavioral activation on depression, anxiety, and activation. Psychol Med 2021 Jul;51(9):1491-1504. [doi: 10.1017/S0033291720000239]
25. Gilmore AK, Walsh K, Frazier P, et al. Post-sexual assault mental health: a randomized clinical trial of a video-based intervention. J Interpers Violence 2021 Nov;36(21-22):10614-10637. [doi: 10.1177/0886260519884674] [Medline: 31709903]
26. Walsh K, Gilmore AK, Schumacher JA, et al. Post-sexual assault cigarette smoking: findings from a randomized clinical trial of a video-based intervention. Addict Behav 2020 Jan;100:106121. [doi: 10.1016/j.addbeh.2019.106121] [Medline: 31622944]
27. Walsh K, Badour CL, Zuromski KL, et al. A secondary analysis of a brief video intervention on suicidal ideation among recent rape victims. Psychol Serv 2021 Nov;18(4):703-708. [doi: 10.1037/ser0000495] [Medline: 33661694]
28. Garland EL, Manusov EG, Froeliger B, Kelly A, Williams JM, Howard MO. Mindfulness-oriented recovery enhancement for chronic pain and prescription opioid misuse: results from an early-stage randomized controlled trial. J Consult Clin Psychol 2014 Jun;82(3):448-459. [doi: 10.1037/a0035798] [Medline: 24491075]
29. Garland EL. Mindfulness‐oriented recovery enhancement: implementing an evidence‐based intervention for chronic pain, opioid use, and opioid addiction in clinical settings. Br J Clin Pharmacol 2024 Dec;90(12):3028-3035. [doi: 10.1111/bcp.16147] [Medline: 39046159]
30. McClintock AS, McCarrick SM, Garland EL, Zeidan F, Zgierska AE. Brief mindfulness-based interventions for acute and chronic pain: a systematic review. J Altern Complement Med 2019 Mar;25(3):265-278. [doi: 10.1089/acm.2018.0351] [Medline: 30523705]
31. Goldsmith RE, Gerhart JI, Chesney SA, Burns JW, Kleinman B, Hood MM. Mindfulness-based stress reduction for posttraumatic stress symptoms: building acceptance and decreasing shame. J Evid Based Complementary Altern Med 2014 Oct;19(4):227-234. [doi: 10.1177/2156587214533703] [Medline: 24812075]
32. Scott Tilley D, Young CC, Richmond M, Humphrey J. Mindfulness-based interventions for adult survivors of sexual assault: a scoping review. J Sex Aggress 2024 Sep;30(3):275-291. [doi: 10.1080/13552600.2023.2187471]
33. Somohano VC, Kaplan J, Newman AG, O'Neil M, Lovejoy T. Formal mindfulness practice predicts reductions in PTSD symptom severity following a mindfulness-based intervention for women with co-occurring PTSD and substance use disorder. Addict Sci Clin Pract 2022 Sep 16;17(1):51. [doi: 10.1186/s13722-022-00333-2] [Medline: 36114577]
34. Wight D, Wimbush E, Jepson R, Doi L. Six steps in quality intervention development (6SQuID). J Epidemiol Community Health 2016 May;70(5):520-525. [doi: 10.1136/jech-2015-205952] [Medline: 26573236]
35. Matthews AK, Anderson EE, Willis M, Castillo A, Choure W. A community engagement advisory board as a strategy to improve research engagement and build institutional capacity for community-engaged research. J Clin Transl Sci 2018 Apr;2(2):66-72. [doi: 10.1017/cts.2018.14] [Medline: 31660220]
36. Hornbuckle LM, Rauer A. Engaging a community advisory board to inform an exercise intervention in older African-American couples. J Primary Prevent 2020 Jun;41(3):261-278. [doi: 10.1007/s10935-020-00589-x]
37. Maurer MA, Shiyanbola OO, Mott ML, Means J. Engaging patient advisory boards of African American community members with type 2 diabetes in implementing and refining a peer-led medication adherence intervention. Pharmacy (Basel) 2022 Mar 10;10(2):37. [doi: 10.3390/pharmacy10020037] [Medline: 35314618]
38. Resnick H, Acierno R, Holmes M, Kilpatrick DG, Jager N. Prevention of post-rape psychopathology: preliminary findings of a controlled acute rape treatment study. J Anxiety Disord 1999;13(4):359-370. [doi: 10.1016/s0887-6185(99)00010-9] [Medline: 10504107]
39. Mobile fact sheet: tech adoption trends. Pew Research Center. 2024 Nov 13. URL: https://www.pewresearch.org/internet/fact-sheet/mobile/ [accessed 2025-05-14]
40. Watson T, Simpson S, Hughes C. Text messaging interventions for individuals with mental health disorders including substance use: a systematic review. Psychiatry Res 2016 Sep 30;243:255-262. [doi: 10.1016/j.psychres.2016.06.050] [Medline: 27423123]

XSL•FO
RenderX

41. Bedard-Gilligan MA, Dworkin ER, Kaysen D, Ojalehto HJ, Stappenbeck CA, Lindgren KP. A pilot study on the feasibility, acceptability, and preliminary efficacy of a brief text message intervention for co-occurring alcohol misuse and PTSD symptoms in a community sample. J Anxiety Disord 2022 Oct;91:102615. [doi: 10.1016/j.janxdis.2022.102615] [Medline: 35988440]
42. Vogt EL, Jiang C, Jenkins Q, et al. Trends in US emergency department use after sexual assault, 2006-2019. JAMA Netw Open 2022 Oct 3;5(10):e2236273. [doi: 10.1001/jamanetworkopen.2022.36273] [Medline: 36264580]
43. Linehan MM. DBT Skills Training Manual: Guilford Publications; 2014. URL: https://tinyurl.com/ycy2zwr6 [accessed 2025-08-01]
44. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
45. Clarke V, Braun V. Thematic analysis. J Posit Psychol 2017 May 4;12(3):297-298. [doi: 10.1080/17439760.2016.1262613]
46. Braun V, Clarke V. Reporting guidelines for qualitative research: a values-based approach. Qual Res Psychol 2025 Apr 3;22(2):399-438. [doi: 10.1080/14780887.2024.2382244]
47. Geraets AFJ, van der Velden PG. Low-cost non-professional interventions for victims of sexual violence: a systematic review. Aggress Violent Behav 2020 Jul;53:101425. [doi: 10.1016/j.avb.2020.101425]
48. Deitz MF, Williams SL, Rife SC, Cantrell P. Examining cultural, social, and self-related aspects of stigma in relation to sexual assault and trauma symptoms. Violence Against Women 2015 May;21(5):598-615. [doi: 10.1177/1077801215573330] [Medline: 25732041]
49. Mellen EJ, Hatzenbuehler ML. Sexual violence-related stigma, mental health, and treatment-seeking: a multimodal assessment in a population-based study of young adults. J Interpers Violence 2023 Oct;38(19-20):11243-11271. [doi: 10.1177/08862605231179715] [Medline: 37491905]
50. Miller AK, Canales EJ, Amacker AM, Backstrom TL, Gidycz CA. Stigma-threat motivated nondisclosure of sexual assault and sexual revictimization. Psychol Women Q 2011 Mar;35(1):119-128. [doi: 10.1177/0361684310384104]
51. Bryant-Davis T, Ullman S, Tsong Y, et al. Healing pathways: longitudinal effects of religious coping and social support on PTSD symptoms in African American sexual assault survivors. J Trauma Dissociation 2015;16(1):114-128. [doi: 10.1080/15299732.2014.969468] [Medline: 25387044]
52. Howe ES, Dworkin ER. The day-to-day relationship between posttraumatic stress symptoms and social support after sexual assault. Eur J Psychotraumatol 2024 Dec 31;15(1):2311478. [doi: 10.1080/20008066.2024.2311478]
53. Lamade RV, Lopez EC, Calias S, Anderson EJ, Koss MP. Healing strategies of women who experienced sexual assault: a mixed method analysis. Psychol Trauma 2025 Feb;17(2):279-288. [doi: 10.1037/tra0001766] [Medline: 39172406]
54. Ullman SE, Relyea M. Social support, coping, and posttraumatic stress symptoms in female sexual assault survivors: a longitudinal analysis. J Trauma Stress 2016 Dec;29(6):500-506. [doi: 10.1002/jts.22143] [Medline: 27862347]
55. Hlavka HR. Speaking of stigma and the silence of shame: young men and sexual victimization. Men Masc 2017 Oct(4):482-505. [doi: 10.1177/1097184X16652656]

Abbreviations

CAB: community advisory board
HIPAA: Health Insurance Portability and Accountability Act
PPRS: Prevention of Post-Rape Stress
PTSD: posttraumatic stress disorder
RCT: randomized controlled trial
SA: sexual assault
SAMFE: sexual assault medical forensic exam
STAR: Skills Training in Active Recovery
TAU: treatment as usual

Edited by K Bond; submitted 03.02.25; peer-reviewed by SR Salim, S Holt; revised version received 16.05.25; accepted 22.06.25; published 18.08.25.

Please cite as:
Walsh K, Schipani-McLaughlin AM, Stappenbeck C, Khan S, Woodworth A, Panwalkar S, Acierno R, Gilmore AK
Refining a Video and Text Message Intervention (STAR, Skills Training in Active Recovery) to Prevent the Onset or Escalation of Posttraumatic Stress and Opioid Misuse Among Recent Sexual Assault Survivors: Community Engaged Study
JMIR Form Res 2025;9:e72095
URL: https://formative.jmir.org/2025/1/e72095
doi:10.2196/72095

XSL•FO
RenderX

© Kate Walsh, Anne Marie Schipani-McLaughlin, Cynthia Stappenbeck, Shahzarin Khan, Anna Woodworth, Sanika Panwalkar, Ronald Acierno, Amanda K Gilmore. Originally published in JMIR Formative Research (https://formative.jmir.org), 18.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

ADAPTed Cognitive Behavioral Therapy for Pediatric Functional Abdominal Pain in Community-Based Pediatric Care: Mixed Methods Study

Emma Ramsay Milford[1,2], DEdPsy; Sandra Buratti[1], PhD; Natoshia R Cunningham[3], PhD; Åsa Nilses[4], MSc; Ewa-Lena Bratt[5,6], PhD; Sandra Weineland[1,7,8], PhD

[1]Department of Psychology, University of Gothenburg, Haraldsgatan 1, Gothenburg, Sweden
[2]Region Västra Götaland, Pediatric Healthcare, Regionhälsan, Gothenburg, Sweden
[3]Department of Family Medicine, Michigan State University, College of Human Medicine, Grand Rapids, MI, United States
[4]Region Västra Götaland, Center for Progress in Children's Mental Health, Regionhälsan, Gothenburg, Sweden
[5]Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden
[6]Region Västra Götaland, Sahlgrenska University Hospital, Children's Heart Centre, Gothenburg, Sweden
[7]Region Västra Götaland, Primary Healthcare, Närhälsan, Södra Älvsborg, Sweden
[8]Sahlgrenska Academy, Institute of Medicine, General Practice/Family Medicine, School of Public Health/Community Medicine, Gothenburg, Sweden

Corresponding Author:
Emma Ramsay Milford, DEdPsy
Department of Psychology, University of Gothenburg, Haraldsgatan 1, Gothenburg, Sweden

Abstract

Background: The Aim to Decrease Anxiety and Pain Treatment (ADAPT) is a blended, digital, and live cognitive behavioral therapy program for children with functional abdominal pain disorder (FAPDs) and anxiety. Initially developed and evaluated in US pediatric gastroenterology settings, a culturally refined version was developed in Swedish, for potential use within community-based pediatric health care settings.

Objective: This study aimed to evaluate a modified version of ADAPT as an early intervention for FAPD within a community-based pediatric setting in Sweden, exploring both the potential treatment effect and participants' treatment experience.

Methods: Participants were aged 9 - 14 years, and all were diagnosed with FAPD. Using a mixed methods design, the study examined the preliminary effect through a single-arm pre-posttest and treatment experience through semistructured child interviews. Data were analyzed in three steps: nonparametric quantitative analysis of results on pre- and postintervention measures of self-rated pain-related functional disability, pain intensity, and anxiety; thematic qualitative analysis of the interviews; and conversion of qualitative data to enable both datasets to be analyzed and presented together.

Results: A total of 13 children (12 girls) participated in ADAPT, all completing the program. In total, 7 of the invited 13 participating children agreed to be interviewed following intervention completion. Quantitative results were analyzed using the Wilcoxon signed rank test, and Pearson *r* was used to calculate the effect size. Results showed a significant reduction in pain-related functional disability, with a median decrease from 14.00 (IQR 10 - 20) preintervention, to 5.00 (IQR 1 - 9) postintervention (P=.04), a large effect size, r=−0.58, and 46% (n=6) achieving a clinically meaningful change where a Functional Disability Inventory (FDI) score decrease of ≥7.8 points denoted a clinically meaningful treatment response. Pain intensity also significantly decreased from a median of 6.5 (IQR 5.25 - 8) to 4.00 (IQR 2.5 - 6.5; P=.02), a large effect size r=−0.70, with 33% (n=4) experiencing at least a 50% reduction. Clinically meaningful change was determined to be present if at least a 50% reduction in self-rated measures of pain intensity was observed. Qualitative thematic analysis identified three themes: "Starting from scratch," "Experiencing the treatment," and "Getting on with life." In terms of treatment experience, the blended live or digital format was perceived as a good fit for youth. Most children described finding some strategy that was effective and reported positive outcomes, such as increased participation in school.

Conclusions: As patient experiences were predominantly positive and quantitative results indicative of potential for increased function and reduced pain, our findings suggest that ADAPT may present as a possible early treatment option for FAPD in a Swedish community–based pediatric setting. To draw robust conclusions on effectiveness, further research is required using a larger sample size, as well as research aimed at following up treatment effects over time.

(JMIR Form Res 2025;9:e67106) doi:10.2196/67106

XSL•FO
RenderX

KEYWORDS

pediatric functional abdominal pain; digital intervention; cognitive behavioral therapy; primary care; cultural adaptation; psychological intervention; pediatrics; children; youth; mixed methods; patient experience; community-based; abdominal pain; anxiety; gastroenterology; child interview

Introduction

Functional abdominal pain disorders (FAPD) are gastrointestinal (GI) symptoms where medical evaluation does not reveal any underlying organic cause. FAPD is diagnosed following the Rome IV criteria, where symptom frequency and duration are taken into account when giving a diagnosis. The location of the abdominal pain, along with the type of gastro or intestinal symptoms, can enable the physician to assign a specific FAPD subtype diagnosis, such as irritable bowel syndrome (IBS), associated with changes in bowel patterns, such as constipation and diarrhea, or functional dyspepsia when upper epigastric pain is present. When criteria for a specific subtype are not met, a more general diagnosis, such as functional abdominal pain not otherwise specified, is given. Pediatric FAPD is a GI condition affecting up to 25% of children [1,2]. FAPD is associated with significant impairment, including academic impairment and poor quality of life, increased risk of developing other pain conditions, and long-term vulnerability to anxiety even if pain resolves [3-8]. Mental health comorbidities, such as anxiety, are common and associated with adverse pain-related outcomes [9-12]. FAPD also poses a considerable economic burden on patients, families, and health care systems [13,14]. Hence, developing and evaluating effective and early treatments is of great importance.

Although there is insufficient evidence supporting pharmacological treatments [15], including antidepressants [16] or probiotics [17], psychosocial interventions, such as cognitive behavioral therapy (CBT), improve pain levels [18], functional disability [19], and quality of life [20,21]. Internet-delivered exposure-based CBT has been found to be both an effective and cost-effective treatment for children with FAPD [21]. Anxiety impacts response to CBT for pain in youth [10]. Thus, Aim to Decrease Anxiety and Pain Treatment (ADAPT) was developed to address both pain and anxiety in youth with FAPD [22]. It was initially developed from evidence-based CBT protocols for the management of pediatric pain [23] and childhood anxiety disorders [24].

Psychological interventions such as CBT for pediatric patients presenting with pain [25] and FAPD [19] have been shown to be effective and of particular importance, as research has shown a connection between FAPD and anxiety [26]. Although comorbid anxiety is common among children with FAPD, it is more usual that gut symptoms occur first and that psychological distress is secondary [27]. Psychological interventions for children with FAPD can be effective in increasing the ability to participate and function in daily activities and alleviating symptoms of anxiety [19,21]. Pain-related functional disability refers to limitations in everyday life activities. Children with pain conditions such as FAPD are commonly limited in their ability to attend school and engage in physical and social activities [28]. Functional disability can co-occur with anxiety, which can adversely impact children's response to CBT [10].

Previous research conducted in Sweden, aimed at evaluating internet-delivered, exposure-based psychological treatment for IBS, a specific subtype of FAPD, has indicated both effective results and cost-effectiveness [20,21]. Exposure-based CBT typically focuses on avoidance behavior and, therefore, is suited to patients who, due to their GI symptoms, avoid pain- and anxiety-inducing stimuli. Exposure therapy aims to decrease fear of symptoms and reduce avoidant behaviors by gradual exposure to symptoms and feared situations and aversive stimuli. The mediating role of exposure has been evaluated in adolescents with IBS [29]. Change in GI-specific anxiety and GI-specific avoidance has been found to mediate changes in parent-reported abdominal symptoms for children with FAPD receiving exposure-based internet-CBT. Children with high baseline values of GI-anxiety and GI-avoidance are likely to benefit more from exposure-based internet CBT than children with low symptom-specific fear and avoidance [30]. Although common in patients with FAPD, not all experience symptom-related fear or display avoidance behavior, particularly those with less severe symptoms. Avoidance behaviors and ineffective pain-coping strategies may be learned from parents [31]. Hence, learning CBT strategies to manage pain in combination with anxiety management strategies at early onset FAPD could be a valuable contribution and treatment option within community-based care. Although the long-term effectiveness of broader multicomponent CBT, such as ADAPT, where several strategies are included as opposed to exposure-focused therapy, has been questioned [29]. However, it can be argued that there is a need for a range of CBT programs in order to provide treatment to children with varying FAPD symptoms. Furthermore, there is a need to develop and implement more treatment programs that can be scaled up and delivered using web-based approaches, both due to the large number of children and young people with FAPD and to keep in line with current developments to make health care more equitable and accessible.

The theoretical underpinnings of ADAPT are grounded in CBT and combine classic components such as cognitive restructuring and exposure with aspects of third-wave CBT [32]. This wave of CBT challenges the notion of behavioral and cognitive change as the only active components and places emphasis on the patient's perception and acceptance of their situation. The struggle to avoid unpleasant and undesirable events, thoughts, and feelings is central to CBT, which has its roots in learning theory. Third-wave CBT focuses on psychological flexibility, requiring acceptance of the concept that some things are not within our control. The ability to be mindful, in the moment, and focus inwards is an aspect of third-wave CBT that is incorporated into ADAPT and the strategies taught through the program. The ADAPT protocol is a CBT protocol and should not be confused with acceptance and commitment therapy. Acceptance and commitment therapy as a pain treatment aims to increase successful engagement in activities that are

XSL•FO
RenderX

meaningful and promote vitality for those who live with persistent pain. The classic CBT components of the ADAPT program are developed from the evidence-based CBT protocol "Cool Kids" for childhood anxiety disorders [24]. ADAPT is also grounded in the gate control theory of pain, which asserts that pain can be inhibited by activation of nerves that do not transmit pain signals. The pain management components of the ADAPT program are developed from an evidence-based protocol for pediatric pain [23]. ADAPT is associated with improvements in disability and anxiety levels as compared to standard medical care in youth with FAPD presenting to pediatric gastroenterology clinics [19].

ADAPT was redeveloped into a Swedish language version by two Swedish clinical psychologists (ÅN and ERM) and the original author of the American version (NRC). Portions of program content (live videos within web modules) were redesigned as animated films. Additionally, cultural adjustments to the original protocol were made (see Multimedia Appendix 1 for an overview of treatment content and session breakdown and Multimedia Appendix 2 for an overview of the original treatment and adapted Swedish language version). The cultural adaptation involved altering both content and language to align with a Swedish cultural setting. This involved changes such as altering terminology that is culturally specific to the United States, for example, spring break, and adjusting references to school grade terminology to align with the Swedish system. The module "Building Social Skills: How to be assertive" in the original version was left out of the Swedish version, as in Swedish school and society, children are not expected to handle bullying incidents. It is stipulated in Swedish educational law that children should not be subject to bullying in school; therefore, the responsibility is placed solely on adults to prevent bullying, and it would not have been appropriate to include this module in the treatment program. The adaptation involved going from videos with actors to creating animated characters, which also allowed for the opportunity to create an inclusive and culturally appropriate visual profile. The translation of the manual and manuscripts was done in several steps. Initially, an authorized translator was used. Following this, the texts were perused by clinical psychologists to ensure appropriate medical and psychological terminology was used. Finding appropriate Swedish names for ADAPT strategies was also part of the adaptation process. Some names were not possible to translate directly due to language differences—for example, calming statements, activity pacing, and pleasant activities, in which case, new terminology had to be applied. The cultural adaptation also involved finding appropriate names for the characters in the Swedish version.

An important reason to evaluate psychological treatment within a community-based pediatric setting is that it is in line with the ambition to decrease utilization of acute health care and offer earlier treatment. The importance of moving from reaction to prevention in relation to FAPD has been highlighted [33]. Multidisciplinary care for children with persistent FAPD has been associated with reduced utilization of acute health care [34]. It has been suggested that a multidisciplinary approach be used for optimal care of children with FAPD, but it is also implied that this may not always be feasible in primary care settings [35]. There is a need for accessible resources to enable FAPD to be managed within first-line health care, as well as practical strategies to assist primary care providers in the recognition and prompt diagnosis of FAPDs for early treatment interventions. It is therefore of importance to evaluate targeted treatments such as ADAPT that can be implemented as an early intervention.

A community-based pediatric setting that incorporates integrated medical assessment, diagnosis, and psychological treatment, such as ADAPT, could provide multidisciplinary care that may have a positive impact on reducing acute health care utilization for children with persistent symptoms. It could be that an integrated approach, including ADAPT as a psychological intervention for pain management within community-based pediatric care, could act as a preventative measure, thereby reducing referrals for repeated assessment and delaying treatment. Although excessive testing should be avoided, it is not uncommon that children with FAPD are subjected to numerous invasive and noninvasive tests. Extensive testing has not been found to have any impact on outcome or patient satisfaction [36]. The impact of pediatric FAPD on the child, parents, and society is substantial. The annual cost of health care for pediatric FAPD was explored in a Dutch study [13] where the cost per patient was estimated at approximately 2500 euros (US $2775). Inpatient and outpatient health care use were major cost drivers, followed by parental productivity loss. Differences in national health care structures prevent these figures from being broadly applied; however, the economic burden of pediatric FAPD can be concluded as being significant, which further implies the urgency of early access to treatment.

This formative study aims to evaluate a modified version of ADAPT as an early intervention for FAPD within a community-based pediatric care setting in Sweden. This is of importance as there is no prior research that has evaluated the potential of a CBT treatment program where all participants were recruited at this level of health care. Furthermore, this study is novel in that it is conducted within a clinical setting where recruitment was integrated into regular care. Due to the limited access to psychological treatment for pediatric FAPD, it is of importance to culturally adapt treatments available and evaluated in other countries, contexts, and languages. In addition to evaluating the effectiveness and accessibility of an adapted treatment program, it is equally important to evaluate children's experiences of the treatment. There is a shortage of research that has aimed to explore children's experiences of internet-delivered CBT for FAPD, and the need to access the opinions of children has been highlighted [37]. With formative research conducted in clinical settings, it is particularly important to obtain both quantitative and qualitative data. This mixed methodology contributes to an enhanced understanding of both the potential treatment effect and the participant's treatment experience. Both aspects are invaluable when making decisions regarding implementing a treatment into clinical practice. Previous research focused on participants' perceptions of ADAPT in relation to feasibility, tolerability, and patient outcomes has yielded valuable information in terms of understanding the potential of the treatment [22]. This study aims to explore experiential factors and outcomes further

XSL•FO
RenderX

through an analysis of the congruence between children's reported experiences of ADAPT and their results on pain-related outcome measures.

Methods

Design

A mixed methods approach [38] was used with a quantitative single-arm pre- and posttest design to explore treatment effect, a qualitative design with semistructured child interviews to explore experience of the treatment, and a conversion of qualitative data to enable both datasets to be analyzed and presented together. The study was approved by the Swedish Ethics Authority (Dnr 2022-03835-02).

Measures

The Functional Disability Inventory (FDI) was used in order to evaluate the level of self-rated pain-related functional disability[39]. The FDI is a 15-item measure of limitations in children's physical and psychosocial functioning due to their physical health. The psychometric properties of the FDI have been evaluated, confirming the reliability and validity of the measure for functional assessment of pediatric patients with chronic pain. Validity was found to be supported by significant correlations of child- and parent-report FDI scores with measures of school-related disability, pain, and somatic symptoms [39]. An example item is: "Being in school all day." The answers are given on a scale ranging from 0 to 4, where 0=No trouble, 1=A little trouble, 2=Some trouble, 3=A lot of trouble, and 4=Impossible. An FDI score decrease of ≥7.8 points was used as an indicator of clinically meaningful treatment response in a previous evaluation of ADAPT [19], based on previous research on pediatric pain [40]. Hence, the same indicator was applied in this study.

Self-rated level of anxiety was measured using the Screen for Child Anxiety Related Disorders (SCARED) [41]. The SCARED is a 41-item measure of anxiety over 3 months, regularly used in pediatric pain research. The psychometric properties of SCARED have been evaluated, indicating robust properties and that the measure is both clinically relevant and can be used in research to monitor intervention effectiveness [42]. Higher scores indicate greater anxiety. An example item is "I worry about being as good as other kids." The response items range from 0 to 2, where 0=Not true or Hardly ever true, 1=Somewhat true or Sometimes true, and 2=Very true or Often true. Clinical (≥25) and subclinical cut-offs (≥12) have been developed for children with chronic pain [43]. Given the adaptation of this program into pediatric community-based care, it was deemed that the subclinical range of anxiety would be optimal to identify patients who may benefit from an early intervention. Hence, patients with a self-rated score of ≥12 were eligible to participate.

Pain intensity was measured using a visual analogue scale (VAS) of 1 - 10 with both digits and faces. The psychometric properties of VAS for children with chronic abdominal pain have been evaluated, indicating strong correlations with other measures aimed at measuring pain [44].

The interview guide was semistructured, based on the semistructured interview used in the original study [22] and included questions about experience, and the impact of ADAPT (see Multimedia Appendix 3 for interview protocol).

Participants

Participating patients were aged 9 - 14 years and diagnosed with FAPD at a community-based pediatric health clinic. The criteria for inclusion were the presence of child-reported subclinical anxiety (total score ≥12), as measured using the SCARED [41,45] or pain-related functional disability as measured by the FDI [39,46], total score ≥7 for study participation. Participants had no organic painful medical conditions nor any identified learning disabilities. Participants were all Swedish language speakers. All participants were assessed by a licensed pediatric psychologist.

Procedure

Participants in this study were recruited from a Swedish community-based pediatric health clinic, which is a specialist clinic within regional care, organizationally located between primary care and hospital-based specialist services. While not traditional primary care, these clinics are part of the public health care system, do not require a hospital referral, and are often the first point of contact for pediatric patients with chronic and complex health care issues. Recruitment took place within the clinical setting. Participating children were referred to the clinic by their general practitioner. After FAPD was confirmed by a pediatrician at the clinic, children were referred, in-house, to a multidisciplinary team at the clinic. The children were then screened for pain-related functional disability with the FDI and presence of worry using a digital patient questionnaire as part of regular clinical practice (refer to Multimedia Appendix 4 for screening questionnaire). Children with an indication of functional disability and a yes response to a question regarding the presence of worry or stress were then assessed for inclusion by a pediatric psychologist at the clinic. Inclusion was based on participants' age, clinical judgment, and assessment, where children with scores below 12 on the SCARED or no presence of pain-related functional disability on the FDI were not deemed to need individual intervention and therefore excluded from the study. Excluded patients were allocated to another intervention where their parents were offered the opportunity to participate in a 2-hour group-based educational training held by a pediatric psychologist and a pediatric nurse. Patients with an FDI score of ≥7 or the presence of worry and avoidance behavior were approached to participate in the study. A psychologist informed the patients, children, and their carers, both verbally and in writing, about the study, whereby written consent was obtained from the parents of 13 participants, all of whom completed the intervention and quantitative assessment measures. All participating children consented verbally. Participants were approached by a psychologist during the final intervention session and asked whether they would consider participating in an optional qualitative interview about their experience of the program. In total, 7 participants (1 assigned male at birth and 6 assigned female at birth) were interviewed approximately 1 month postcompletion of ADAPT.

XSL•FO
RenderX

Data were collected between January 2022 and December 2023, and recruitment was consecutive during the initial 18 months of the period.

Intervention

ADAPT [22] is a blended, digital, or face-to-face CBT protocol targeted toward pediatric FAPD and anxiety symptoms. ADAPT is brief, consisting of two clinic-based sessions and four self-paced web sessions. Each self-paced web module was followed by a digital session (video call) with a psychologist. The treatment sessions were held at weekly intervals, allowing for a treatment period of approximately 6 weeks. ADAPT was delivered by three clinical psychologists affiliated with the participating clinic. Parents participated in both the clinic-based and digital sessions. In line with the treatment manual [22], the pain-coping strategies (deep-breathing, mindful breathing, guided imagery, progressive muscle relaxation, calming statements, and activity pacing) were taught to the child alone. The instructions were then repeated to the child and parent to facilitate practice between sessions. The duration of the two clinic sessions was approximately 1 hour, and the digital sessions with the clinical psychologist were approximately 20 minutes.

Analyses

The data were analyzed in three steps: (1) quantitative analysis of results on pre- and postintervention measures of self-rated pain-related functional disability, pain intensity, and anxiety; (2) qualitative analysis of the interviews; and (3) integration of results through assessment of the congruence of quantitative and qualitative results.

Step 1: Quantitative Analysis

Descriptive information and changes in outcomes at the group level, functional disability, pain intensity, and anxiety were explored via nonparametric Wilcoxon signed rank tests using SPSS (version 29; IBM Corp) software. Pearson *r* [47] was used to calculate effect size, where results are referred to as small (−0.1 to −0.3), medium (−0.3 to −0.5), or large (−0.5 or less). Reference points for clinically meaningful change in functional disability from previous evaluations of ADAPT were used. Hence, a total FDI score of 3.5 was considered indicative of healthy youth, and an FDI score decrease of ≥7.8 points was considered indicative of a clinically meaningful treatment response [19]. Clinically meaningful change was determined to be present if at least a 50% reduction in self-rated measures of pain intensity was observed [48]. A 50% reduction in scores on SCARED at posttreatment was also considered a clinically meaningful indicator of improvement or remission. Reduction of 50% on the SCARED has been found to optimally predict treatment response and can be used to systematically monitor treatment outcomes [49].

All participants completed the treatment and self-rated postmeasures. There was no missing data among postintervention measures; however, one participant's preintervention measure of pain intensity was missing due to an administrative error. Hence, n=12 for VAS preintervention.

Step 2: Qualitative Analysis

Reflexive thematic analysis was carried out by two researchers (ERM and ÅN) as part of the research project. Data were approached from a critical realist position as the participants' subjective descriptions of their experiences were obtained, and the researchers' roles and subjectivity were also recognized as informing the research. A critical realistic position in the context of reflexive thematic analysis takes the view that participants' perceptions are shaped by their own contexts and representations of reality, which are then interpreted and analyzed, and thereby shaped by the experiential and cultural context held by the researcher. The aim of thematic analysis from a critical realist standpoint is to establish a lucid and intelligible interpretation of the participants' experiences that resounds with situational realities [50].

To ensure comprehensive analysis, a team was assembled, consisting of three coauthors: ERM, a licensed psychologist and doctoral student; ÅN, a licensed psychologist; and SW, an associate professor, supervisor, and licensed psychologist. Drawing from our collective understanding and experience of internet-delivered CBT as an effective treatment for adults, we initiated the analysis process. Through iterative discussions, themes, subthemes, and categories were collaboratively refined and defined until consensus was achieved among the analysis team, initially comprising authors ERM and ÅN, and later including SW, SB, and ELB. Notably, ERM and ÅN's analysis was overseen by SW, progressing from themes and subthemes to categories, initial codes, and ultimately quotes, ensuring the reliability of the findings. Atlas.ti (version 23; ATLAS.ti Scientific Software Development GmbH) software was used as support in data processing.

Step 3: Convergent Analysis

A conversion of qualitative data was performed to enable both datasets to be analyzed together and presented in a joint display table [38]. During this process, themes were analyzed again, comparing them to individual responses. Responses were then reduced into a congruence measure to indicate to which extent the individual's responses aligned with the theme. Although the themes were drawn from the entire dataset and thereby guided by the individual responses, there was a range of responses where some were more congruent with or aligned to the theme. This analysis enabled any patterns between self-rated measures and the level of congruence with subthemes to be observed. A low congruence score was applied when the individual responses generally aligned with the subtheme, but some responses were conflicting. A medium congruence score was applied when the individual's responses generally aligned with the subtheme, but some responses were neutral. A high congruence score was applied when all the individual's responses aligned with the theme.

Ethical Considerations

The study adhered to the ethical principles outlined in the World Medical Association's Declaration of Helsinki. Participation was entirely voluntary. Participants and their guardians were provided with written and verbal information about the study. Guardians provided written informed consent, and participants

XSL•FO
RenderX

provided verbal assent. The study received ethical approval from the Swedish Ethics Review Authority (2022-03835-02). All participants and their legal guardians were informed about the study during an initial psychological assessment session. Guardians and patients subsequently attended an appointment for feedback regarding the assessment. If the participant met the inclusion criteria, participation in the study was offered. Upon agreement, the guardian signed a written consent form, and the participant provided verbal assent. This process also applied to secondary data used in this analysis; the original consent included approval for secondary analysis. To ensure participant confidentiality, all collected data were deidentified and assigned numerical codes. No personally identifiable information was retained in the analytic dataset. Audio recordings were securely stored. Access to the data was restricted to the research team only. Data handling and storage complied with the General Data Protection Regulation and the data management policies of Regionhälsan (Region Västra Götaland), including secure storage on encrypted devices and restricted physical access in accordance with institutional guidelines.

Results

Quantified Change

Results on measures of pain-related functional disability indicate a statistically significant decrease in disability from preintervention, median 14.00 (IQR 10‑20), indicative of moderate disability; range 10‑20, to postintervention, median 5.00 (IQR 1‑9), indicative of mild disability; n=13, and a large effect size, *z* score=−2.10; *P*=.04; *r*=−0.58. Given the 12-point change in disability, 46% (n=6) achieved a clinically significant [40] change (>7.8 points). Approximately 85% (11/13) of participants experienced decreased levels of pain-related functional disability (median change=8). The inclusion process and measurement points are provided in Figure 1.

Measures of pain intensity were also indicative of a statistically significant decrease from preintervention, median 6.5 (IQR 5.25‑8), n=12, to postintervention, median 4.00 (IQR 2.5‑6.5), n=13, and a large effect size was found, *z* score=−2.42; *P*=.02; *r*=−0.70. Clinically meaningful [48] pain reduction was observed in 4 of 12 (33%) participants who experienced at least a 50% reduction in pain intensity scores after ADAPT. Approximately 83% (10/12) of participants experienced decreased levels of pain intensity.

A total of 6 (46%) participants experienced at least some reduction across all three outcomes, with the 7 remaining participants experiencing a reduction in one (8%) or two (38%) categories. One participant did not improve in any category.

XSL•FO
RenderX

Figure 1. Allocation process and included participants. FAPD: functional abdominal pain disorders; FDI: Functional Disability Inventory; SCARED: Screen for Child Anxiety Related Disorders.

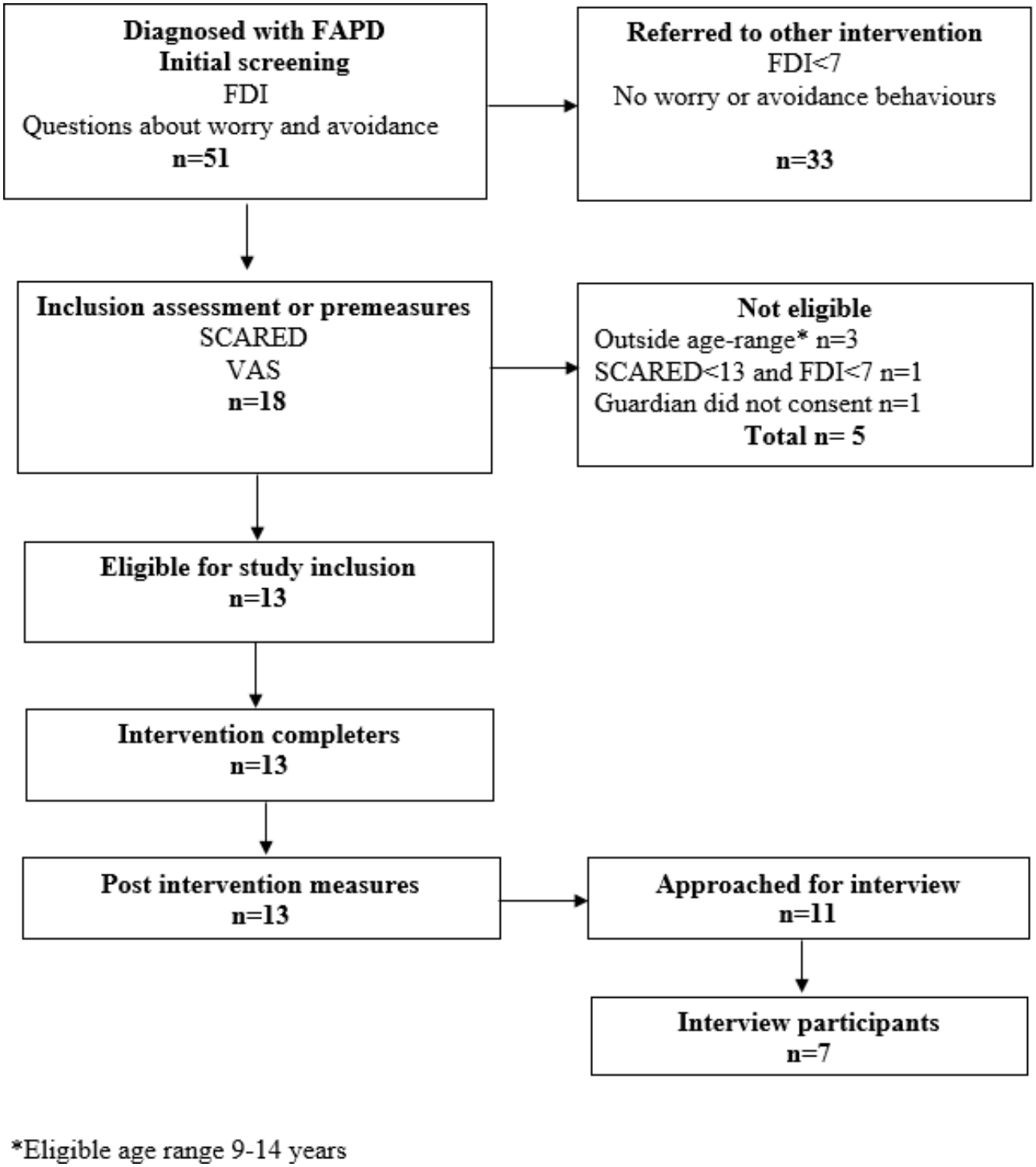

Qualitative Insights—Treatment Experience

Thematic analysis of the seven interviews resulted in three main themes and seven subthemes. The themes and subthemes are presented in Textbox 1 (see Multimedia Appendix 5 for themes, subthemes, and corresponding quotes from child interviews).

XSL•FO
RenderX

Textbox 1. Themes and subthemes

Starting from scratch

- No experience of psychological treatment
- Uncertain expectations

Experiencing the treatment

- Getting on with the content
- Engaging with a therapist
- Managing the digital format
- Getting help from a parent or carer

Getting on with life

- Using some of a range of strategies
- Learning about pain and worry
- Doing things despite symptoms

Starting From Scratch

This theme draws upon the children's expressions about thoughts, knowledge, and expectations of the treatment prior to commencing. Two subthemes were found, one relating to experience and the other to expectations. The overarching theme draws on the finding that children expressed both having little experience and being uncertain about what to expect from it.

Experiencing the Treatment

This theme relates to the children's attitude toward and experience of the blended format of the treatment, with the in-person sessions with a therapist, video calls, and digital modules including the animated films, as well as how they perceived the content. The digital content, such as films and assignments, was found to be acceptable by most participants. The content was perceived by some participants as somewhat tedious, and characters and examples were not found to be relatable by all participants. However, not being able to relate to the characters was not described as problematic. Generally, the blended format was perceived positively. Children made positive comments about the interaction with the therapist, but also mentioned the benefits of working from home. Saving time and choosing when to do the work was mentioned. Meeting the therapist was mostly appreciated and by some referred to as "real." Most of the participants expressed getting and needing help from a parent.

Getting on With Life

This theme captures the children's experiences of the strategies and perceived outcomes of the treatment. The children commented on the strategies, indicating a preference for some over others. Pain coping strategies were mentioned more than the strategies targeting anxiety. Most of the children gave examples of having applied some strategy. Some children reported increased understanding of symptoms of pain and worry. Increased participation in school was mentioned by most children as being the main impact of the program on their functioning.

Mixed Methods Integrated Results

Participant responses obtained through interviews varied in their level of congruence with established subthemes. The integrated results for the seven interviewed participants are displayed in Table 1. Most of the participants made comments that had a high to medium level of congruence with subthemes, but some responses were less or not congruent. The mixed methods joint display enables observation of patterns in responses and self-rated change. The convergent results in Table 1 suggest that patients with a low level of congruence on two or more subthemes are those whose ratings on postintervention measures increased. In general, a match between a high level of congruence and an overall decrease between pre- and postintervention ratings was observed.

XSL•FO
RenderX

Table . Pre- and postintervention results on self-rated measures of pain-related functional disability, anxiety, pain intensity, subthemes, and level of congruence between responses and subthemes among interviewed participants (n=7)[a].

	Participants[b] (age in years)						
	Ellen (10)	Alex (11)	Molly (9)	Julie (14)	Sofia (12)	Anne (12)	Lara (12)
Measures							
Functional Disability Inventory preintervention: median 16 (IQR 12.5-20; 14)	24	16	9	29	11	14	16
Functional Disability Inventory postintervention: median 6 (IQR 4.5-8.5; 5)	0	4	16	6	5	7	10
Change	−24	−12	+7	−23	−6	−7	−6
Pain intensity rating preintervention 0 - 10: median 7 (IQR 6.0-7.5; 6.5)	8	7	8	6	6	7	4
Pain intensity rating postintervention 0 - 10: median 5 (IQR 2.5-6.5; 4)	7	5	10	2	2	6	3
Change	−1	−2	+2	−4	−4	−1	−1
Screen for Child Anxiety Related Disorders preintervention: median 23 (IQR 16.0-33.5; 21)	12	13	39	37	30	23	19
Screen for Child Anxiety Related Disorders postintervention: median 23 (IQR 12.0-30.0; 23)	7	23	54	34	14	26	10
Change	−5	+10	+15	−3	16	+3	−9
Subthemes (congruence between responses and subthemes[c])							
No experience of psychological treatment	☒	☒	☒	☒	☒	☒	☒
Uncertain expectations	☒	☒	☒	☒	☒	☒	☒
Getting on with the content	☒	☒	☒	☒	☒	☒	☒
Getting help from a parent or carer	☒	☒	☒	☒	☒	☒	☒

XSL•FO
RenderX

	Participants[b] (age in years)						
	Ellen (10)	Alex (11)	Molly (9)	Julie (14)	Sofia (12)	Anne (12)	Lara (12)
Engaging with a therapist	☒	☒	☒	☒	☒	☒	☒
Managing the digital format	☒	☒	☒	☒	☒	☒	☒
Using some of a range of strategies	☒	☒	☒	☒	☒	☒	☒
Learning about pain and worry	☒	☒	☒	☒	☒	☒	☒
Doing things despite symptoms	☒	☒	☒	☒	☒	☒	☒

[a]Median scores for participants included in mixed methods analysis (n=7) are shown, as well as those for all intervention completers (n=13, shown in brackets) to enable comparison between scores for mixed methods participants and whole group.

[b]Not real names. Median age 12 (n=7).

[c]Interpretation key for level of congruence: low: ☒; medium: ☒; high: ☒.

Discussion

Principal Findings

Pediatric FAPD is common, associated with significant impairment, increased risk of developing other pain conditions, and long-term vulnerability to anxiety. Furthermore, the condition poses a considerable economic burden on patients, families, and health care systems. CBT as a treatment for FAPD has previously been found to be effective, and ADAPT has been evaluated for patients recruited from pediatric gastroenterology departments. This study evaluated a modified version of ADAPT as an early intervention for FAPD within a community-based pediatric care setting in Sweden.

Findings from this study indicate that ADAPT is a potentially effective and suitable treatment for pediatric FAPD within pediatric community-based clinics, particularly for pain-related outcomes. Qualitative results indicate that several interviewed patients reported positive experiences, and where the reported experience was mixed, positive quantitative outcomes were observed in at least one domain of outcome variables for most participants. The results are in line with previous studies [19,22], reinforcing the efficacy of ADAPT as a potentially viable therapeutic approach and broadening its usability to a community-based pediatric care context. Treating patients early when their symptoms are less compounded may prevent more long-term medical and associated mental health difficulties. Prior research into pediatric FAPD has generally focused on patients at secondary or tertiary level of care [19-22] and has not, as was done in this study, solely recruited patients seeking treatment within regular clinical care at pediatric community-based clinics. This may have represented a missed opportunity for earlier and more accessible care.

At the group level, our findings indicate both statistically and clinically significant changes in terms of pain-related functional disability and pain intensity. Due to the small sample size and single-arm design, these findings cannot be generalized and will need to be replicated in a controlled trial with more participants to draw conclusions about the treatment effect. Hence, next steps should include conducting larger, controlled studies to evaluate the treatment's effectiveness more robustly, as well as exploring long-term outcomes and mechanisms of change. One limitation of this study relates to the use of cut-off scores on the SCARED instrument. The use of categorical cut-offs in an uncontrolled, observational context may be problematic. Cut-offs can impose artificial dichotomies on what is inherently a dimensional construct and may lead to a loss of information about subclinical variations in symptom severity. In the absence of a control group, reliance on such thresholds may also overestimate or underestimate clinically meaningful change. Future studies should consider supplementing cut-off-based classifications with dimensional analyses to better capture nuanced symptom trajectories over time.

There was no significant change in measures of anxiety. Although individuals with subclinical anxiety were recruited, the sample also contained participants with higher levels of anxiety, either approaching or within the clinical range. It may be that patients with subclinical levels of anxiety could benefit from early treatment, particularly in terms of impact on pain-related outcomes. This would need to be explored further in future research, ideally with a larger sample size, before concluding whether early interventions delivered through community-based pediatric care have the potential to prevent the onset of a formal anxiety disorder diagnosis.

Future research may wish to address participants' quality of life and other mental health comorbidities to explore potential changes and issues that may impact the treatment effect. It would also be of interest to explore whether participants' anxiety was pain-related, specific, or more general, and if these aspects affect outcomes. Furthermore, repeated measurements at follow-up

XSL•FO
RenderX

intervals would enable a greater understanding of any changes in anxiety and treatment effect over time. As increased participation in school was mentioned by participants as an improvement to their functioning following the intervention, it would have been of interest to explore this in greater detail, particularly the presence of significant differences on specific questionnaire items relating to school attendance. This was not possible in this study due to the design and method of data collection. Future research may wish to explore ADAPT's potential specifically with regard to school attendance by measuring change on specific items within the FDI and SCARED relating to school.

The blended format appears to be a good fit as participants expressed both advantages of engaging with the therapist, as well as the digital format. Children's interview responses indicate that the web-based format allowed them to use their time more effectively; however, the personal connection obtained through clinic visits was also appreciated. There is evidence that any live intervention (whether in person or digital) is comparable [51]. Yet, future research may wish to explore whether face-to-face intervention in clinics differs from face-to-face intervention delivered via digital channels, such as video calls, specifically for this population. Such research could provide insight into whether components of blended treatments can be adapted further to increase treatment accessibility and patient satisfaction. All participants completed the whole treatment and all modules, and the reasons for the good compliance could be explored further.

Although psychological treatment programs are not expected to be equally effective for all patients, children's views can contribute important information in terms of improving outcomes. Accessing children's views is both relevant in terms of the United Nations conventions on the rights of the child and also in line with the current movement toward patient-centered care. Our findings indicate that children use at least one out of a range of strategies. These findings imply that there may be a demand for CBT for pediatric FAPD that includes a range of strategies. This would need to be explored further by obtaining children's views on a range of CBT components in further depth.

A great deal of effort was placed on constructing a visual profile that was representative of genders, ethnicity, and ability, and providing a range of examples of FAPD and worry. This did not appear to have impacted some children's views of how relatable the content was. However, as the findings indicate potential treatment effects on pain and functional disability, it may be that being able to relate to characters was considered more important by the treatment creators than by the participating children. This could be explored further in future studies seeking to evaluate factors contributing to the treatment effect. Furthermore, it is worth considering whether creating a treatment program with a team of adults and then asking children about their views of the content is the optimal design process. Perhaps cocreating the characters and using children's own examples as content would have been a better option. This should be considered in future projects aimed at designing digital treatments aimed at children.

Strengths and Limitations

A key strength of this study lies in its focus on the feasibility and acceptability of the ADAPT intervention within a real-world clinical context. By conducting the study in a community-based pediatric care setting, we were able to explore how the intervention functions under routine clinical conditions. The use of a mixed methods design further enhances the value of the findings, allowing for integration of both quantitative outcomes and qualitative data into participants' experiences. This dual perspective provides a richer understanding of how the intervention was received by children and which components may contribute to meaningful change.

However, several limitations should be acknowledged. The small sample size limits the generalizability of the findings and reduces the statistical power to detect more subtle effects. This constraint is partly due to the study being conducted within a single clinical setting and under limited funding conditions. Despite these limitations, the study offers an important foundation for future development and scale-up. Overall, this study contributes valuable insights into how a structured psychological intervention can be implemented and received in community-based care and serves as a foundation for continued adaptation and integration.

Transitioning psychological treatment for FAPD from hospital-based care to pediatric clinics offers several important benefits. These clinics often represent the first point of contact for children with recurrent abdominal pain, making them an ideal setting for early identification and intervention. Providing access to evidence-based interventions such as CBT could improve treatment accessibility, reduce referral delays, and decrease the burden on hospitals. Early intervention may also prevent the development of chronic symptoms, reduce unnecessary medical investigations, and improve overall outcomes for both children and families. Embedding interventions like ADAPT in community-based pediatric care could be an important step toward more integrated, potentially preventive, and cost-effective management of pediatric FAPD.

Conclusions

This study is the first of its kind to evaluate the use of ADAPT through regular clinical routine within a pediatric community-based care context. The findings may provide a way forward for early, potentially effective, and accessible treatment for FAPD, as well as shedding light on patients' treatment experience. For robust conclusions to be drawn, future research would need to expand upon these findings, further exploring ADAPT's potential as an accessible treatment for pediatric FAPD available within community-based care.

Acknowledgments

The authors of the study would like to thank the participants who participated in the study. The authors would also like to thank Tommy Skjulsvik and Robin Björk for their assistance in setting up the modules on the Swedish national support and treatment

XSL•FO
RenderX

platform. No generative artificial intelligence tools were used in the writing of the original manuscript. However, editorial assistance (language editing and formatting) during the revision phase was supported by a generative language model (ChatGPT, OpenAI), under the direct supervision of the authors. All content has been carefully reviewed and verified by the research team to ensure accuracy and integrity.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

ERM and SW conceived the idea presented. They jointly devised the study, including crafting the research plan and securing ethical approval. Initially, ERM, ÅN, and SW commenced data analysis, both qualitative and quantitative. Subsequently, regular meetings were held with SB and ELB to discuss findings and refine analytical approaches until consensus was reached on analysis and manuscript development. NRC was involved throughout all stages, contributing to the result discussions and the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Adapt content and session breakdown.
[DOCX File, 13 KB - formative_v9i1e67106_app1.docx]

Multimedia Appendix 2
Overview of the original treatment and Swedish language version.
[DOCX File, 14 KB - formative_v9i1e67106_app2.docx]

Multimedia Appendix 3
Interview protocol.
[DOCX File, 14 KB - formative_v9i1e67106_app3.docx]

Multimedia Appendix 4
Questions about abdominal pain and worry.
[DOCX File, 13 KB - formative_v9i1e67106_app4.docx]

Multimedia Appendix 5
Themes, subthemes, and related quotes.
[DOCX File, 17 KB - formative_v9i1e67106_app5.docx]

References

1. Thapar N, Benninga MA, Crowell MD, et al. Paediatric functional abdominal pain disorders. Nat Rev Dis Primers 2020 Nov 5;6(1):89. [doi: 10.1038/s41572-020-00222-5] [Medline: 33154368]
2. Robin SG, Keller C, Zwiener R, et al. Prevalence of pediatric functional gastrointestinal disorders utilizing the Rome IV criteria. J Pediatr 2018 Apr;195:134-139. [doi: 10.1016/j.jpeds.2017.12.012] [Medline: 29398057]
3. Mulvaney S, Lambert EW, Garber J, Walker LS. Trajectories of symptoms and impairment for pediatric patients with functional abdominal pain: a 5-year longitudinal study. J Am Acad of Child Adolesc Psychiatry 2006 Jun;45(6):737-744. [doi: 10.1097/10.chi.0000214192.57993.06]
4. Saps M, Seshadri R, Sztainberg M, Schaffer G, Marshall BM, Di Lorenzo C. A prospective school-based study of abdominal pain and other common somatic complaints in children. J Pediatr 2009 Mar;154(3):322-326. [doi: 10.1016/j.jpeds.2008.09.047]
5. Shelby GD, Shirkey KC, Sherman AL, et al. Functional abdominal pain in childhood and long-term vulnerability to anxiety disorders. Pediatrics 2013 Sep;132(3):475-482. [doi: 10.1542/peds.2012-2191] [Medline: 23940244]
6. Stanford EA, Chambers CT, Biesanz JC, Chen E. The frequency, trajectories and predictors of adolescent recurrent pain: a population-based approach. Pain 2008 Aug 15;138(1):11-21. [doi: 10.1016/j.pain.2007.10.032] [Medline: 18093737]
7. Walker LS, Dengler-Crish CM, Rippel S, Bruehl S. Functional abdominal pain in childhood and adolescence increases risk for chronic pain in adulthood. Pain 2010 Sep;150(3):568-572. [doi: 10.1016/j.pain.2010.06.018] [Medline: 20615615]

XSL•FO
RenderX

8. Wendland M, Jackson Y, Stokes LD. Functional disability in paediatric patients with recurrent abdominal pain. Child Care Health Dev 2010 Jul;36(4):516-523. [doi: 10.1111/j.1365-2214.2010.01087.x] [Medline: 20412144]
9. Cunningham NR, Jagpal A, Peugh J, et al. Risk categorization predicts disability in pain-associated functional gastrointestinal disorders after 6 months. J Pediatr Gastroenterol Nutr 2017 May;64(5):685-690. [doi: 10.1097/MPG.0000000000001342] [Medline: 27437930]
10. Cunningham NR, Jagpal A, Tran ST, et al. Anxiety adversely impacts response to cognitive behavioral therapy in children with chronic pain. J Pediatr 2016 Apr;171:227-233. [doi: 10.1016/j.jpeds.2016.01.018] [Medline: 26879812]
11. Cunningham NR, Lynch-Jordan A, Mezoff AG, Farrell MK, Cohen MB, Kashikar-Zuck S. Importance of addressing anxiety in youth with functional abdominal pain: suggested guidelines for physicians. J Pediatr Gastroenterol Nutr 2013 May;56(5):469-474. [doi: 10.1097/MPG.0b013e31828b3681] [Medline: 23412539]
12. Moorman EL, Farrell M, Santucci N, Denson L, Le C, Cunningham NR. Elevated fecal calprotectin is linked to psychosocial complexity in pediatric functional abdominal pain disorders. BMC Res Notes 2021 Sep 15;14(1):360. [doi: 10.1186/s13104-021-05776-5] [Medline: 34526124]
13. Hoekman DR, Rutten JMTM, Vlieger AM, Benninga MA, Dijkgraaf MGW. Annual costs of care for pediatric irritable bowel syndrome, functional abdominal pain, and functional abdominal pain syndrome. J Pediatr 2015 Nov;167(5):1103-1108. [doi: 10.1016/j.jpeds.2015.07.058] [Medline: 26329806]
14. Tack J, Stanghellini V, Mearin F, et al. Economic burden of moderate to severe irritable bowel syndrome with constipation in six European countries. BMC Gastroenterol 2019 May 7;19(1):69. [doi: 10.1186/s12876-019-0985-1] [Medline: 31064345]
15. Martin AE, Newlove-Delgado TV, Abbott RA, et al. Pharmacological interventions for recurrent abdominal pain in childhood. Cochrane Database Syst Rev 2017 Mar 6;3(3):CD010973. [doi: 10.1002/14651858.CD010973.pub2] [Medline: 28262913]
16. de Bruijn CMA, Rexwinkel R, Gordon M, Benninga M, Tabbers MM. Antidepressants for functional abdominal pain disorders in children and adolescents. Cochrane Database Syst Rev 2021 Feb 9;2(2):CD008013. [doi: 10.1002/14651858.CD008013.pub3] [Medline: 33560523]
17. Gordon M, Wallace C, Sinopoulou V, Akobeng AK. Probiotics for management of functional abdominal pain disorders in children. Cochrane Database Syst Rev ;2023(2). [doi: 10.1002/14651858.CD012849.pub2]
18. Gordon M, Sinopoulou V, Tabbers M, et al. Psychosocial interventions for the treatment of functional abdominal pain disorders in children: a systematic review and meta-analysis. JAMA Pediatr 2022 Jun 1;176(6):560-568. [doi: 10.1001/jamapediatrics.2022.0313] [Medline: 35404394]
19. Cunningham NR, Kalomiris A, Peugh J, et al. Cognitive behavior therapy tailored to anxiety symptoms improves pediatric functional abdominal pain outcomes: a randomized clinical trial. J Pediatr 2021 Mar;230:62-70. [doi: 10.1016/j.jpeds.2020.10.060] [Medline: 33130153]
20. Bonnert M, Olén O, Lalouni M, et al. Internet-delivered cognitive behavior therapy for adolescents with irritable bowel syndrome: a randomized controlled trial. Am J Gastroenterol 2017 Jan;112(1):152-162. [doi: 10.1038/ajg.2016.503] [Medline: 27845338]
21. Lalouni M, Ljótsson B, Bonnert M, et al. Clinical and cost effectiveness of online cognitive behavioral therapy in children with functional abdominal pain disorders. Clin Gastroenterol Hepatol 2019 Oct;17(11):2236-2244. [doi: 10.1016/j.cgh.2018.11.043] [Medline: 30502501]
22. Cunningham NR, Nelson S, Jagpal A, et al. Development of the aim to decrease anxiety and pain treatment for pediatric functional abdominal pain disorders. J Pediatr Gastroenterol Nutr 2018 Jan;66(1):16-20. [doi: 10.1097/MPG.0000000000001714]
23. Kashikar-Zuck S, Ting TV, Arnold LM, et al. Cognitive behavioral therapy for the treatment of juvenile fibromyalgia: a multisite, single-blind, randomized, controlled clinical trial. Arthritis Rheum 2012 Jan;64(1):297-305. [doi: 10.1002/art.30644] [Medline: 22108765]
24. Lyneham HJ, Abbott MJ, Wignall A, Rapee RA. The Cool Kids Anxiety Treatment Program: MUARU, Macquarie University; 2003.
25. Coakley R, Wihak T. Evidence-based psychological interventions for the management of pediatric chronic pain: new directions in research and clinical practice. Children (Basel) 2017 Feb 4;4(2):9. [doi: 10.3390/children4020009] [Medline: 28165415]
26. Cunningham NR, Cohen MB, Farrell MK, Mezoff AG, Lynch-Jordan A, Kashikar-Zuck S. Concordant parent-child reports of anxiety predict impairment in youth with functional abdominal pain. J Pediatr Gastroenterol Nutr 2015 Mar;60(3):312-317. [doi: 10.1097/MPG.0000000000000625] [Medline: 25714575]
27. Koloski NA, Jones M, Talley NJ. Evidence that independent gut-to-brain and brain-to-gut pathways operate in the irritable bowel syndrome and functional dyspepsia: a 1-year population-based prospective study. Aliment Pharmacol Ther 2016 Sep;44(6):592-600. [doi: 10.1111/apt.13738] [Medline: 27444264]
28. Stahlschmidt L, Friedrich Y, Zernikow B, Wager J. Assessment of pain-related disability in pediatric chronic pain: a comparison of the functional disability inventory and the pediatric pain disability index. Clin J Pain 2018 Dec;34(12):1173-1179. [doi: 10.1097/AJP.0000000000000646] [Medline: 30134282]

XSL•FO
RenderX

29. Bonnert M, Olén O, Bjureberg J, et al. The role of avoidance behavior in the treatment of adolescents with irritable bowel syndrome: a mediation analysis. Behav Res Ther 2018 Jun;105:27-35. [doi: 10.1016/j.brat.2018.03.006] [Medline: 29614378]
30. Lalouni M, Hesser H, Bonnert M, et al. Breaking the vicious circle of fear and avoidance in children with abdominal pain: a mediation analysis. J Psychosom Res 2021 Jan;140:110287. [doi: 10.1016/j.jpsychores.2020.110287] [Medline: 33227558]
31. Newton E, Schosheim A, Patel S, Chitkara DK, van Tilburg MAL. The role of psychological factors in pediatric functional abdominal pain disorders. Neurogastroenterol Motil 2019 Jun;31(6):e13538. [doi: 10.1111/nmo.13538] [Medline: 30729663]
32. Hofmann SG, Sawyer AT, Fang A. The empirical status of the "new wave" of cognitive behavioral therapy. Psychiatr Clin North Am 2010 Sep;33(3):701-710. [doi: 10.1016/j.psc.2010.04.006] [Medline: 20599141]
33. Lu PL, Saps M. Functional gastrointestinal disorders: all roads lead to prevention. Clin Gastroenterol Hepatol 2018 Jun;16(6):814-816. [doi: 10.1016/j.cgh.2018.01.033] [Medline: 29391269]
34. Chu AS, Torres L, Kao G, Gilbert C, Monico EC, Chumpitazi BP. Multidisciplinary care for refractory pediatric functional abdominal pain decreases emergency and inpatient utilization. J Pediatr Gastroenterol Nutr 2022 Feb 1;74(2):248-252. [doi: 10.1097/MPG.0000000000003305] [Medline: 34560729]
35. Vandenplas Y, Darma A, Indrio F, et al. Understanding functional abdominal pain disorders among children: a multidisciplinary expert consensus statement. Front Pediatr 2025;13:1576698. [doi: 10.3389/fped.2025.1576698] [Medline: 40433476]
36. Farrell P, Farrell L, Farrell MK. The approach to the management of a child with chronic abdominal pain. Curr Treat Options Peds 2024;10(2):64-78. [doi: 10.1007/s40746-023-00288-9]
37. Søgaard-Singh H, Roth Hulgaard D, Skovslund Nielsen E, et al. I'm a bit more free": lived symptom experiences in adolescents with functional abdominal pain disorders after internet-delivered psychological treatment. J Health Psychol 2025 Aug;30(9):2282-2296. [doi: 10.1177/13591053251320600] [Medline: 40069964]
38. Creswell J, Vicki LPC. Designing and Conducting Mixed Methods Research, 3rd edition: Sage Publications; 2017.
39. Claar RL, Walker LS. Functional assessment of pediatric pain patients: psychometric properties of the functional disability inventory. Pain 2006 Mar;121(1-2):77-84. [doi: 10.1016/j.pain.2005.12.002] [Medline: 16480823]
40. Sil S, Arnold LM, Lynch-Jordan A, et al. Identifying treatment responders and predictors of improvement after cognitive-behavioral therapy for juvenile fibromyalgia. Pain 2014 Jul;155(7):1206-1212. [doi: 10.1016/j.pain.2014.03.005] [Medline: 24650858]
41. Birmaher B, Khetarpal S, Brent D, et al. The Screen for Child Anxiety Related Emotional Disorders (SCARED): scale construction and psychometric characteristics. J Am Acad Child Adolesc Psychiatry 1997 Apr;36(4):545-553. [doi: 10.1097/00004583-199704000-00018] [Medline: 9100430]
42. Runyon K, Chesnut SR, Burley H. Screening for childhood anxiety: a meta-analysis of the screen for child anxiety related emotional disorders. J Affect Disord 2018 Nov;240:220-229. [doi: 10.1016/j.jad.2018.07.049]
43. Cunningham NR, Jagpal A, Nelson S, et al. Clinical reference points for the screen for child anxiety-related disorders in 2 investigations of youth with chronic pain. Clin J Pain 2019 Mar;35(3):238-246. [doi: 10.1097/AJP.0000000000000667] [Medline: 30362983]
44. Ginton L, Nurko S, Di Lorenzo C, Saps M. First study to assess the reliability of commonly used pain scales in children with disorders of gut-brain interaction. Neurogastroenterol Motil 2023 Sep;35(9):e14636. [doi: 10.1111/nmo.14636] [Medline: 37448182]
45. Ivarsson T, Skarphedinsson G, Andersson M, Jarbin H. The validity of the Screen for Child Anxiety Related Emotional Disorders Revised (SCARED-R) Scale and sub-scales in Swedish youth. Child Psychiatry Hum Dev 2018 Apr;49(2):234-243. [doi: 10.1007/s10578-017-0746-8]
46. Lind MC, Tornberg Å. Psychometric properties of the Swedish version of the Functional Disability Inventory. Eur J Physiother 2015 Jul 3;17(3):116-122. [doi: 10.3109/21679169.2015.1046486]
47. Lee Rodgers J, Nicewander WA. Thirteen ways to look at the correlation coefficient. Am Stat 1988 Feb;42(1):59-66. [doi: 10.1080/00031305.1988.10475524]
48. Dworkin RH, Turk DC, Wyrwich KW, et al. Interpreting the clinical importance of treatment outcomes in chronic pain clinical trials: IMMPACT recommendations. J Pain 2008 Feb;9(2):105-121. [doi: 10.1016/j.jpain.2007.09.005] [Medline: 18055266]
49. Caporino NE, Sakolsky D, Brodman DM, et al. Establishing clinical cutoffs for response and remission on the Screen for Child Anxiety Related Emotional Disorders (SCARED). J Am Acad Child Adolesc Psychiatry 2017 Aug;56(8):696-702. [doi: 10.1016/j.jaac.2017.05.018] [Medline: 28735699]
50. Braun V, Clarke V. Thematic Analysis: A Practical: Sage Publishing Ltd; 2021.
51. Cunningham NR, Ely SL, Barber Garcia BN, Bowden J. Addressing pediatric mental health using telehealth during coronavirus disease-2019 and beyond: a narrative review. Acad Pediatr 2021;21(7):1108-1117. [doi: 10.1016/j.acap.2021.06.002] [Medline: 34118499]

Abbreviations

ADAPT: Aim to Decrease Anxiety and Pain Treatment

XSL•FO
RenderX

CBT: cognitive behavioral therapy
FAPD: functional abdominal pain disorder
FDI: Functional Disability Inventory
GI: gastrointestinal
IBS: irritable bowel syndrome
SCARED: Screen for Child Anxiety Related Disorders
VAS: visual analogue scale

Edited by A Mavragani; submitted 08.10.24; peer-reviewed by C Brimeyer, M Bonnert; revised version received 01.07.25; accepted 02.07.25; published 20.08.25.

Please cite as:
Ramsay Milford E, Buratti S, Cunningham NR, Nilses Å, Bratt EL, Weineland S
ADAPTed Cognitive Behavioral Therapy for Pediatric Functional Abdominal Pain in Community-Based Pediatric Care: Mixed Methods Study
JMIR Form Res 2025;9:e67106
URL: https://formative.jmir.org/2025/1/e67106
doi:10.2196/67106

© Emma Ramsay Milford, Sandra Buratti, Natoshia R Cunningham, Åsa Nilses, Ewa-Lena Bratt, Sandra Weineland. Originally published in JMIR Formative Research (https://formative.jmir.org), 20.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Promoting Sustained Real-Life Benefits of Virtual Reality–Based Interventions in People With Mental Health and Substance Use Disorders: Qualitative Study

Jan Aasen[1], MMHC; Fredrik Nilsson[2]; Torgeir Sørensen[3,4], Prof Dr; Marja Leonhardt[1,5], PhD

[1]Research Centre for Substance Use Disorders and Mental Health Disorders (ROP-Forsk), Innlandet Hospital Trust, Erik Werenskiolds veg 3, Hamar, Norway

[2]RIO - Rusmisbrukerenes interesseorganisasjon, Oslo, Norway

[3]Centre for Diaconia and Professional Practice, VID Specialized University, Oslo, Norway

[4]Research Centre for Existential Health, Innlandet Hospital Trust, Ottestad, Norway

[5]Section for Clinical Addiction Research, Oslo University Hospital, Oslo, Norway

Corresponding Author:
Jan Aasen, MMHC
Research Centre for Substance Use Disorders and Mental Health Disorders (ROP-Forsk), Innlandet Hospital Trust, Erik Werenskiolds veg 3, Hamar, Norway

Abstract

Background: Concurrent mental health and substance use disorders (MHD/SUD) are one of the most prominent public health problems as of today, and the worldwide prevalence of MHD/SUD is currently increasing. Modern virtual reality technology may provide easy, unlimited, and safe access to social experiences and interactions that hold the potential to promote individuals' new learning for the benefit of their social participation and recovery. However, the clinical adoption of virtual reality–based interventions (VRIs) is still in its infancy. Human limitations in skills transfer from virtual to actual reality are a major challenge in designing efficient VRIs. Key working mechanisms of the interactive, digital social environments in virtual realities have yet to be identified. There is a lack of knowledge on how immersive learning experiences may be designed and structured to promote sustained real-life benefits for people with mental health and substance use disorders.

Objective: The main aims of this paper were to explain the factors affecting the outcomes of learning in multisensory virtual reality environments and to examine how they affect our particular target group. The overall purpose of this study was to understand how learning experiences in VRIs may be designed and orchestrated to promote sustained real-life benefits of VRIs in people with MHD/SUD.

Methods: Eight individual in-depth interviews with adults in recovery from mental health and substance use disorders were conducted in a medium-sized municipality in eastern Norway in fall 2022. The interviews were analyzed using template analysis, a form of codebook thematic analysis, in a process involving peer researcher collaboration.

Results: This study suggests that the human capacity to achieve sustained learning outcomes from multisensory immersive learning experiences was limited in general. This study also indicates that people with mental health and substance use disorders struggle with attention deficit, concentration, and memory to an extent that it affects their daily functioning.

Conclusions: Altogether, the theoretical framework and empirical findings provide added information on how we may develop learning experience designs in VRIs that accommodate human perceptual processes. VRI scenarios that may be repeated and structured according to individual learning prerequisites may enable the restructuring of maladaptive social schema. This may possibly promote the storage of new, repaired schemas in the user's long-term memory. It is therefore suggested that short, focused VRI scenarios, orchestrated in a sequenced and deliberately structured learning workflow, may promote sustained real-life benefits from VRIs in people with mental health and substance use disorders.

Trial Registration: ClinicalTrials.gov NCT05653167; https://clinicaltrials.gov/study/NCT05653167

(JMIR Form Res 2025;9:e57643) doi:10.2196/57643

KEYWORDS

mental health and substance use disorders; virtual reality–based interventions; immersive learning; cognitive load; social skills acquisition; template analysis

XSL•FO
RenderX

Introduction

Overview

Concurrent mental health and substance use disorder (MHD/SUD) is one of the most prominent public health problems as of today. The worldwide prevalence is currently increasing [1]. With the increasing prevalence of MHD/SUD globally, there is a need for developing more comprehensive, integrated, and evidence-based responses for harm reduction and support to recovery among people with MHD/SUD [1,2]. Simultaneously, the increasing demand for health and welfare services is a major concern for our future health and welfare services. Staff shortages are one of the most important factors threatening capacity going forward [3]. This leads to a need for understanding how the use of digital technologies, such as virtual reality–based interventions (VRIs), may be optimized in novel MHD/SUD treatment [1,2,4].

Focusing on social and functional recovery and supporting people in the process of becoming active and participating citizens are key parameters within the recovery process for people with MHD/SUD [5,6]. Modern virtual reality (VR) technology may provide easy and safe access to social experiences and interactions that hold the potential to promote individuals' learning for the benefit of their social participation and recovery [7-10]. Summarized research shows that VRIs have few challenges in recreating sufficient realism in the simulations to provide emotional responses and sufficient learning outcomes in experimental and clinical settings [9,11,12]. Current research shows that VRIs may be equally or more efficient than conventional treatment. The highest evidence and application maturity for VRIs are for treating PTSDs and anxiety disorders [13,14]. Once developed, VRIs can be distributed to an infinite number of hardware devices without any significant increase in costs [13]. However, the clinical adoption of VRIs is still in its infancy [13]. VRIs hold the potential to strengthen treatment capacity, treatment availability, and treatment quality in future mental health care services [7].

Lack of personalization of VR apps to patient needs is regarded as a major impediment to clinical adoption [15]. A limited amount of research has been conducted to date on the efficacy of VR-based cue exposure approaches for addiction, which aim to extinguish craving and prevent relapses. Small trials on subjective and physiological outcomes have emerged for nicotine and alcohol use disorder and pathological gambling [13,16]. Many of these studies, however, lack sufficient methodological rigor, and further large-scale randomized controlled trials are missing [13].

The literature reflects a wide variation of VR environments and software apps, in which all of them are regarded as VR, but they evidently provide different user experiences [17,18]. There is also a wide variation in VRI duration, dosage, and repetition, with little knowledge on optimizing learning workflow to promote efficient and sustained learning uptake [19]. There is also a lack of theoretical conceptualization and understanding of the characteristics of effective VRIs [7,11,13,20]. There is a need for a better understanding of factors that may affect positive VRI outcomes [16].

The key working mechanisms in virtual realities in terms of clinical outcomes and behavioral change have yet to be identified [8,21,22]. A main challenge when developing efficient VRIs is human limitations of transferring acquired skills from digital to real environments and their long-term maintenance [8,9,12,23-26]. A nearly 41% increase in frontal theta power was observed when testing cognitive load using an electroencephalogram with a fully immersive head-mounted VR display, compared with a 2D semi-immersive tablet display [27]. This points to the focus on understanding the impact of immersive VR experiences on the human perceptual processes [28]. Challenges associated with human factors in immersive learning in general include general health, physiology, adverse effects, cognitive load, and compatibility [29].

However, all learners are not equally capable of effective and efficient learning [30]. The characteristics of a population group targeted by VRIs are important determinants. These influence the impact that digital interventions may have on their functional trajectories [31]. It is also vital to understand the target group's perception of the use of digital interventions, as these are key uptake factors for all types of digital interventions as well [31].

Little research has been conducted on uptake factors for immersive learning or the feasibility of VRIs among people with MHD and SUD [20]. Studying key features, VRI uptake factors, and learning prerequisites among people with MHD/SUD is thus essential for creating VRIs that provide effective and persistent skills transfer from VR to real life.

Aim of the study

There is a lack of explanations of the factors affecting the outcomes of learning in multisensory VR environments. Explaining these factors and examining how they affect our particular target group was the main aim of this paper.

This main aim led to 2 subaims for this study. The first aim was to obtain a better understanding of human perceptual processes in fully immersive VR in general. The second aim was to examine how these perceptual processes were affected among people with MHD/SUD.

The overall purpose of this study was to understand how learning experiences in VRIs may be designed and orchestrated to promote sustained real-life benefits of VRIs in people with MHD/SUD.

Theoretical Framework

Learning experience design involves creating learning activities that induce an active process of constructing and retaining knowledge. This also includes orchestrating learning activities in a sequenced or otherwise carefully and deliberately structured learning workflow [30]. In immersive learning, the term "immersion" describes how technological affordances facilitate learning by inducing a sense of presence or copresence in digital environments. "Learning" refers to the response to immersive experiences in human perceptual and motor systems [28]. This indicates that a theoretical framework for the learning experience design in VRIs may be derived from basic principles of human learning. Within adult learning, this is referred to as didactics. The didactical aspect of social skills training in VRIs is

XSL•FO
RenderX

important to ensure that the participants achieve the learning goals related to the determined learning objectives (knowledge, skills, attitudes, behavior) from the VRI scenarios. Didactics is an umbrella term for the theoretical and practical application of learning in a wider sense than teaching. The didactical foundation of adult simulation training is most commonly based on Kolb's experiential learning theory [32,33]. Experiential learning is based on five crucial assumptions for adult learning: (1) adult learners are self-directed and self-regulated, (2) adult learners are intrinsically motivated to learn, (3) adult learners have previous knowledge and experience that form mental models which guide their behavior through this experience, (4) adults use analogical reasoning in learning and practice, and (5) one cannot force learning upon adults [33]. Adult learners must be explained the relevance the learning has to their work and guided in the understanding of how, when, and why to learn [33].

In VRIs, this learning process stems from constant multisensory inputs that depend on the individual's perceptual capacity to respond to multiple stimuli simultaneously. This requires executive shifts in focused attention according to the demands of the particular VRI scenario. This relies on human perceptual processes. In VRIs, learners make sense of information by organizing sensory input into understandable forms. They built meaningful mental pictures from the information given in the VR setting. This is known as constructivist learning [34]. According to the constructivist learning theory [34], learning occurs when learners are able to build referential connections between corresponding aspects of the visual and verbal representations of the learning object. Constructivist learning is fostered when the learner is able to hold a visual representation in visual working memory and a corresponding verbal representation in verbal working memory at the same time [34]. This also implies that the cognitive load on the working memory acts as a major impediment to social skills learning in multisensory learning environments, such as the VR scenarios in VRIs [34].

Attention and short-term memory processing are thus abilities for constructing sustainable learning outcomes from multisensory learning experiences in VRIs. This is a key distinction as the learning conveyance determines the learning uptake, whereas the retaining process determines the durability of the learning uptake.

This is important for the ecological validity of the VRI. The ecological validity of VRIs refers to the extent to which VRI users improve functioning, daily life skills, compensatory strategies, social participation, and psychological well-being in real-life settings and situations [8]. The process from sensory perception to using new behavioral repertoires in real-world situations in social learning is referred to as social skills acquisition [35]. The concept of acquisition refers to how persons acquire abilities and construct knowledge through mental information [35]. This refers to how cognitive structures develop and change, and how repertoires of new behaviors are acquired, used as practical intelligence, and used in capabilities to behavioral adaptation in real-life settings [34,35]. Learning deals with how people construct knowledge through mental information and how cognitive structures develop and change. Social performance is the process of retrieving social skills from the cognitive schemas in long-term memory and applying these skills to appropriate behaviors in relevant real-life settings [34,36]. The acquisition process is illustrated in Figure 1.

Figure 1. The process from sensory perception in virtual reality to practical application of new behavioral repertoires in real-world situations. Figure by authors, based on the description of social learning and constructivist learning in multimedia environments by Mayer et al [34] and Kopelowicz et al [37].

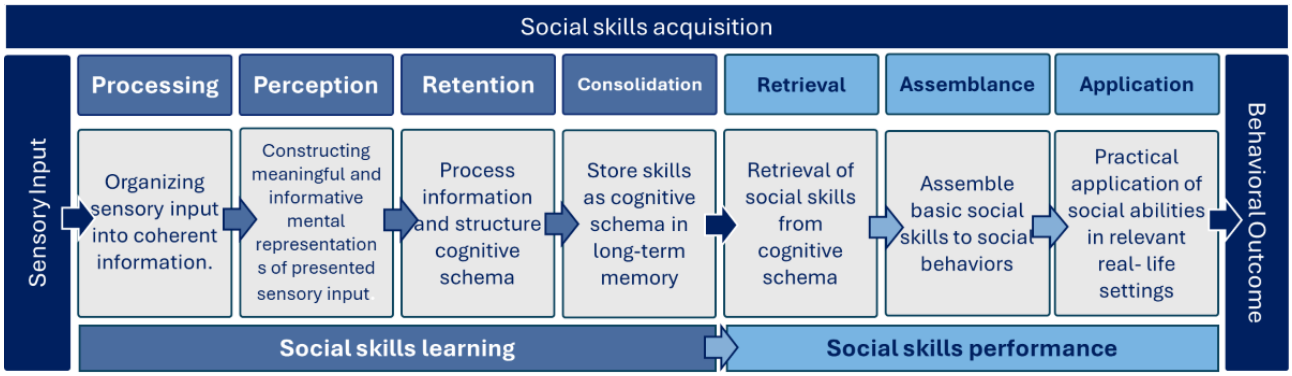

The link between high scenario authenticity, real-life fidelity, and ecological validity has been a dominating assumption in the field of simulation-based learning [38]. Learning tasks are mainly designed on the basis of real-life problem-solving with complex element interactivity, under the assumption that this leads to effective task performance in real-life settings [39]. According to cognitive load theory, this is contradictory to the functionality of human cognition [38]. Cognitive overload in learning leads to a collapse in memory processing and thus also in skills acquisition. Cognitive load theory is based on the assumption that the human cognitive system has a limited working memory [38]. This assumption is derived from Atkinson and Shiffrin's [40] classical work on multimodal memory. They explain human memory as a multimodal system consisting of sensory memory, a transient short-term memory buffer, often referred to as working memory, and a long-term memory store [40,41]. The working memory enables transient evaluation, manipulation, and synthesis of newly acquired information. The working memory operates within the short-term memory and interacts with attention and executive function [41]. According to Atkinson and Shiffrin's [40] theory on multimodal memory, human working memory is unable to hold more than 5 to 9 information elements and can actively process no more than 2 to 4 elements at the same time.

XSL•FO
RenderX

Additionally, all unprocessed information in the short-term memory is lost in about 20 seconds when dealing with novel memory obtained by information processing. Input that exceeds the sensory buffer capacity will be overwritten by new sensory inputs and forgotten before this sensory information is passed on for further processing [38]. In contrast to processing sensory input, short-term memory has no known limitations when dealing with information retrieved from long-term memory. Only skills stored as cognitive schemas are available and sustainable in long-term memory [38]. The cognitive limitations in the acquisition process are illustrated in Figure 2.

Figure 2. Cognitive limitations in the process from sensory perception in virtual reality to using new behavioral repertoires in real-world situations. Figure by authors, based on the description of social learning and constructivist learning in multimedia environments by Mayer et al [34], Kopelowicz et al [37], and Atkinson and Shiffrin [40].

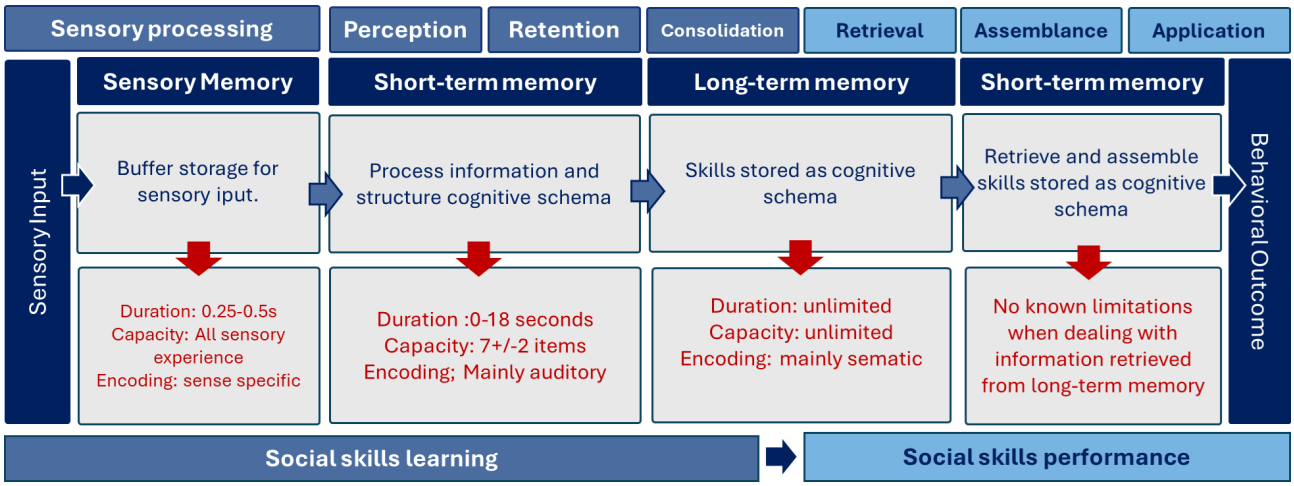

The long-term memory holds cognitive schemas that vary in their degree of complexity and automation. These schemas can then be treated as a single element in short-term memory. This greatly decreases the cognitive load associated with the performance of later tasks. Practicing skills already stored in long-term memory allows for a far higher task load than novel skills learning [38]. The cognitive load within the skills acquisition process is divided into 3 distinct cognitive load modalities: extraneous, intrinsic, and germane load [38]. Extraneous load is load generated by external factors that interfere with cognitive processing, negatively impacting the learner's capacity to buffer and process the relevant sensory inputs. Intrinsic load is load directly related to the material or task given in the learning situation and is often defined by the number of individual elements and interactivity between them. Germane load is the load on short-term memory capacity generated by accessing mental schemas and making connections necessary to move the material to long-term memory [38]. As the total cognitive load accumulates throughout the acquisition process, reducing the extraneous load will free up cognitive capacity for further perceptual information processing [38,39]. Unnecessary information in the VRI scenarios may thus hamper the end users' skills acquisition.

Methods

Context and Design

For this qualitative exploration, we used data from 8 in-depth interviews with people in recovery from MHD/SUD. The interviews were conducted in a medium-sized municipality in Eastern Norway during the fall of 2022.

Recruitment

Recruitment for this study was conducted at a nongovernmental support center for people in MHD/SUD recovery. The first author presented the research project to the whole group of service users present and invited them to participate in the study. Those who wished to participate then scheduled interviews with the staff of the center they attended. The inclusion criteria were as follows: (1) aged 18 years and older, (2) self-reported MHD/SUD with experience of social marginalization, and (3) the capacity to understand the study information and informed consent. This sampling strategy is a type of convenience sampling called volunteer sampling, where the decision to participate is strongly dependent on respondents themselves due to the nonindividualized nature of the invitations [42]. Convenience sampling is a sensitive recruiting strategy as we met all the participants in their own settings and caused little disturbance.

Participants

Eight persons with self-reported MHD/SUD, 3 females and 5 males, were recruited. The average age of the participants was 48 years, ranging from 42 to 61 years. One participant was an Eastern European immigrant, and one participant was not a native Norwegian but had grown up with Norwegian foster parents. The other participants were native Norwegians. Seven participants started substance use in early adolescence, with an average age of onset of 12 years. Six participants were in opiate agonist treatment at the time of the interviews. One participant reported polysubstance use without opiate dependency, and one participant reported opiate dependency only. All participants reported MHD experience, including anxiety, depression, attention deficit/hyperactivity disorder, posttraumatic stress disorder (PTSD), and personality disorders. All participants reported prior experienced social marginalization due to

XSL•FO
RenderX

MHD/SUD. The participant demographics are presented in Table 1.

Table . Participant demographics.

Participant	Sex	Age (years)	SUD[a] onset (years)	Self-reported MHD[b]
1	Male	51	44	PTSD[c], ADHD[d], and anxiety
2	Male	42	11	PTSD, ADHD, anxiety, and psychosis
3	Male	50	11	Anxiety
4	Male	44	12	Complex PTSD, ADHD, personality disorder, depression, and psychosis
5	Female	51	10	Anxiety
6	Female	43	13	PTSD and anxiety
7	Female	61	10	Anxiety
8	Male	45	14	PTSD/ ADHD/Anxiety

[a]SUD: substance use disorder.

[b]MHD: mental health disorder.

[c]PTSD: posttraumatic stress disorder.

[d]ADHD: attention deficit hyperactivity disorder.

Data Collection

The first author conducted the first 3 interviews independently, while the other 5 were comoderated by the first and second authors. The interviews lasted from 24 to 90 minutes and followed a semistructured interview guide. The interview guide consisted of open-ended questions on how the participants perceived their opportunities and prerequisites for participating in work, education, and cultural life in mainstream society. In order to ensure a comfortable and safe setting for the participants, we conducted our interviews at the support center where the participants were recruited. All the interviews were audiotaped and transcribed verbatim. The second author is a peer support worker who has lived experience of MHD and SUD. He participated in developing the interview guide, conducting the interviews, and collaborating in the analysis.

Analysis

Data were analyzed using template analysis with a deductive approach [43]. Template analysis is a variant of codebook thematic analysis [44,45]. Template analysis is theory-led and based on a set of a priori themes organized in a predetermined template. Key uptake factors for immersive learning were derived from the present theoretical construct and used as a priori themes. A priori themes are often based on theoretical ideas that have guided a particular study, as in this study [43].

The present theoretical framework implies that perceptual capacity, working memory, and executive functioning are key prerequisites for learning uptake from multisensory learning experiences in VRIs. Additionally, prior social modeling is a key factor for prior schema development. Hence, cognitive schemas act as key preconditions for social skills learning, as these schemas determine whether the users need to learn novel skills or practice skills already stored in the long-term memory. The a priori themes were developed from these theoretical principles. The thematic template is presented in Table 2.

In the first step of the template analysis, the first author familiarized himself with the data by exploring the verbatim transcriptions of the interviews and by reading the transcriptions afterwards. In the second step, the first and second authors generated initial codes from empirical statements within and across the transcripts.

The first author then presented the empirical statements and initial coding to the other authors in step 3. In step 4, the meaning and relevance of the empirical statements were discussed at both the semantic and latent levels with all authors. Step 5 consisted of the first author interpreting the data using the template. The results were then discussed with the second and fourth authors in step 6. Additionally, the empirical findings were presented, discussed with, and verified by the interview participants. The first author wrote up the analysis in step 7. The third and fourth authors edited the analytic description, and the first author drafted the final report of the results.

In this analysis process, the first author brought philosophical meta-theoretical assumptions, prior professional experience as a specialized psychiatric nurse, and theoretical knowledge to the analysis. The second author brought an additional layer of analytical perspective with lived experience, forming his preunderstandings in the shaping of the codes and revision of the themes. The themes are validated by the respondents, all themes are connected to quotes, and all participants are represented in the results. The themes, theme relations, theme hierarchy, and result descriptions are discussed and validated by the research team and the study participants.

XSL•FO
RenderX

Table . A priori themes and a priori subthemes.

	A priori themes	A priori subthemes
1	Digital uptake factors	• Digital literacy • Digital intervention receptiveness • VRI[a] expectations
2	Immersive learning capabilities	• Executive functioning • Memory processing • Cognitive capability
3	Social skills learning preconditions	• Prior social skills modeling

[a]VRI: virtual reality–based intervention.

Ethical Considerations

People with MHD/SUD are a highly vulnerable and particularly hard-to-reach population, which requires particularly considerate and gentle collaboration with our participants during recruitment and interviews. Within this target group, many are struggling with social anxiety and low trust in strangers. All recruitment and data collection in this study were conducted in line with the Norwegian guidelines for medical and health research. This study was approved, pursuant to the Norwegian Health Research Act, §10, by the South-Eastern Norway Regional Committee for Medical and Health Research Ethics (ref. 421376), and the Data Protection Officer of Innlandet Hospital Trust (ref. 18197741). All participants signed an informed consent that described the purpose of the project, personal data processing requirements, and data protection plan. The participants were allowed to withdraw from the study at any time. The researchers discussed potential vulnerability within the context of the research, potential consequences of their research participation, and potential positive impact of the research with each participant before signing the consent to participate. The participants were also informed about their right to withdraw their data from the research. The participants were not compensated for their study participation. The research data in their final version, after coding and deidentification, is stored in the form of text files (doc./docx.) and audio files (MP3). During the data collection, data processing, and analysis phases of the project, research data are stored in the researchers' personal, password-protected areas at the Innland Hospital Trusts' secured research server. At the end of the project period, the research data will be archived at Services for Sensitive Data (TSD) at the University of Oslo, in accordance with current agreements with Innlandet Hospital Trust and VID Specialized University. When archiving data, each dataset will have an identifier (personal identification).

Results

Overview

Three main themes were derived from the analysis based on the thematic template: (1) VRI receptiveness, (2) deficient immersive learning abilities, and (3) disadvantageous social learning prerequisites. The overall theme structure with a priori themes, final themes, and the associated subthemes is presented in the final thematic map in Table 3.

In the following, the main themes and the codes will be explained and related to empirical statements from the participants. These are presented in italics to show how the content appeared in the interviews. The statements have been translated from Norwegian with the original meaning preserved as far as possible.

Table . Final main themes and the associated subthemes.

	Final main themes	Final subthemes
1	Digital uptake factors	• Digital literacy • Digital intervention receptiveness • VRI[a] expectations
2	Immersive learning capabilities	• Executive functioning • Memory processing • Cognitive capability
3	Social skills learning preconditions	• Prior social skills modeling

[a]VRI: virtual reality–based intervention.

Theme 1: VRI Receptiveness

This theme refers to the participants' digital intervention uptake factors. This includes experience with digital technologies in general and VRT in particular. In general, the participants reported various levels of digital literacy. Some participants reported little experience with using computers and expressed a perceived lack of digital literacy in their daily lives.

XSL•FO
RenderX

These participants reported difficulty in using banking apps and web portals for public services. They also lacked access to digital technologies:

> *I haven't learned much computer stuff. I should though, otherwise you'll get stuck in today's society.* [P6]

Other participants used computers on a daily basis in work practice and reported average user competence. Others reported extensive experience with gaming and serious interest in digital technology:

> *I spend much of my days on Netflix and gaming […].* [P8]

With regard to VRT in particular, the participants reported having heard of VR and having understandings and opinions about VR and VRTs. They all had seen head-mounted VR displays, and some had experience of using recreational VR apps.

> *I have those cheap VR goggles at home, which I can put my phone inside, then I sit at the bottom of the sea watching sharks and so on.* [P3]

This theme also accommodates the participants' attitudes towards meeting technology-based interventions instead of those delivered by humans in MHD and SUD treatment. The participants with little digital experience reported interest in the possibility of new kinds of digital interventions to promote their recovery.

Some participants described how their social anxiety was triggered by human presence, regardless of the situation and their relationship with the person present. One participant explained how she felt that digital interventions could create a safe social learning space to benefit people who struggle with social anxiety:

> *Well, I think VR goggles would be really effective for me. So, I can sit at home and practice social skills so that I don't have to think about what I'm doing and what I'm saying and how I'm standing, because then I can focus on the things I want to focus on and not all the nonsense that I constantly get hung up about. So actually, I think VR would be very effective for me.* [P6]

Overall, the participants expressed positive attitudes towards VRI and were receptive to using VRIs as a tool for social skills learning.

Theme 2: Deficient Immersive Learning Abilities

This theme is related to the participants' abilities to acquire social skills through immersive learning experiences. The participants described a wide range of social and functional impairments in their daily lives. Their difficulties varied, but they all reported problems with concentration and short-term memory to an extent that impaired their everyday functioning.

Remembering appointments and messages was difficult for all the participants. They related many of their daily challenges to experiences they reported about difficulty in regulating attention, remembering things, or concentrating on tasks. They reported more profound impairments, such as the inability to recognize people and retain knowledge. The participants narrated their everyday struggles in different ways:

> *I cannot concentrate... I can't even read books, and I love to read.* [P1]

> *Well, I forget things, like if I'm just going to get something from the kitchen, I have to go back because I forget what I was supposed to be doing.* [P3]

These quotations illustrate the participants' attention deficits and short-term memory. Several of them reported better functioning prior to starting serious substance use and elaborated on how their mental capacity had deteriorated due to excessive polysubstance use and possible hypoxic brain damage from multiple overdoses:

> *[…] I reckon I've messed up quite a few things in my head because of my substance use, so I'm very forgetful.* [P2]

> *I've had 36 overdoses when I was given an antidote, so I should have been dead a long time ago.* [P8]

The participants claimed some improvement in their cognitive deficits after abstaining completely from substance use for an extended period. However, none of them had experienced a full recovery of their cognitive abilities. Several of them talked about how they did not understand the impact of sobriety until they became entirely sober. They described how abstaining totally from substance use had been important for their functioning and well-being.

They also reported that being clean was vital for benefiting from learning-based interventions. One of the participants expressed it in this way:

> *You have to give up powder and cut out tablets, the most important thing is you have to abstain from substance use enough to get to a point where you're able to think.* [P7]

The participants also talked about how opioid agonist therapy affected them in their everyday functioning. One participant explained his situation as follows:

> *I need to get off that methadone and all that stuff because it limits me so much in everything — I sweat and get tired and get totally worn out — just from washing up my dishes.* [P3]

Theme 3: Disadvantageous Learning Prerequisites

This theme describes the life experiences that have formed the participants' cognitive schemas for social skills and that determine their social learning preconditions. Some of the participants described adverse childhood experiences, such as neglect, domestic violence, substance use, or lack of empathic responses from their caregivers. Others described warm relations with their parents and siblings, but harsh social treatment with bullying and exclusion among their peers and in school. One of the participants explained:

> *We lived quite a sheltered life when I was little…until I started elementary school. There the whole school*

XSL•FO
RenderX

> *bullied me, and when I got home my father gave me a beating […] it was a lot like that.* [P3]

Additionally, the participants described many adverse experiences during their lives in substance use communities. Due to substance use onset early in their preadolescence, they were associated with their local substance use communities from a very young age. They described these scenes as harsh environments that exposed them to deviant behavior, such as violence, exploitation, and intimidation:

> *It's a kind of environment, on the streets, that comes with a lot of stuff…* [P8]

Several of the participants also stated that when they came off substances, they found themselves at the same socioemotional stage as when they started using them. They explained that they were still struggling with immature socioemotional functioning and described a perceived lack of adult behavioral repertoires. One participant put it this way:

> *[…] now I've just woken up from twenty-four years of substance use […] and I notice that I have the same mental problems I had when I was fourteen.* [P7]

The participants spoke about adverse school experiences. Several described learning difficulties and little help from teachers. They also talked of social exclusion and bullying in elementary school. They often skipped classes, and most of them were early dropouts. Overall, the participants expressed that they were missing much of the emotional and social skills they believed were required for appropriate adult functioning.

The participants also perceived their current social abilities as unfit for participation in mainstream society among people they considered normal. One of them summed it up in this way:

> *[….] you kind of have to learn everything all over again, learn to be human and relearn how to live.* [P6]

Discussion

Principal Findings

This study sought to provide a better understanding of the human perceptual processes in immersive VR and how these key VRI uptake factors were affected among people with MHD/SUD. This study suggests that the capabilities for social skills acquisition in VRIs are limited in people in general. The empirical findings in this study suggest that the immersive learning capabilities of people with MHD and SUD are inferior to those of the population in general. This study also showed disparities in digital literacy among the participants, but they were all positive towards using VRIs to promote their social participation and functional recovery.

The participants reported various levels of experience with digital technologies, but they all knew what VR was. They were receptive to meeting digital interventions instead of human-delivered measures in MHD and SUD treatment. The participants also suggested that digital interventions could alleviate social anxiety in social skills training, as such interventions enable social skills learning without the presence of humans, which triggers social anxiety. Overall, this study suggests that VRIs may be acceptable and feasible to use with people with MHD/SUD. This is in line with a systematic review on the feasibility of digital interventions among MHD and SUD populations, which shows that VRIs are feasible, but not consistently effective [20]. A study on the feasibility of VRIs for people who experience psychosis shows that they are feasible and are accepted by this patient group [9].

A key finding in this study is how the participants expressed difficulty with attention, concentration, and memory to an extent that it affected their daily functioning. Previous research shows that MHD and SUD often consist of various combinations of mental health disorders, neurodevelopmental disorders, PTSD, and long-term substance use complications. This can lead to substantial cognitive impairment [46-48]. Neuropsychological research from the last 2 decades has consistently demonstrated that impairments in cognitive functioning are frequent among people with substance use disorders [49]. It has also been found that a majority of people with MHD and SUD struggle with impaired learning prerequisites in various forms and severities [47].

Deficits in executive functioning, decision-making, and goal-directed behaviors are the most prominent problem areas observed in people with problematic substance use [49-51]. This includes difficulty with attention, planning, memory, problem-solving, and self-regulation. All these factors affect daily functioning, including the capacity to understand concepts and skills learned in an intervention and apply them to daily lives [50]. Tasks that measure working memory have revealed cognitive deficits in individuals with substance use disorders [52]. Working memory refers to the capacity to monitor and alter information held in mind temporarily, inhibition involves overriding an unwanted distraction to maintain task focus, and shifting pertains to flexibly switching attention between tasks or mental sets [53].

Deficits in short-term memory were also a prevalent topic among the participants. Short-term memory processing relies on attention and is crucial for constructing and reconstructing cognitive schemas that can be stored in long-term memory [38]. Attention is key to awareness and focusing on a stimulus, or a set of sensory stimuli. Attention is also key to the perception, selection, and filtering of sensory input [41]. Sustained attention is the capacity to maintain attention over an extended period, whereas selective, or focused, attention is the ability to preferentially attend to a subset of stimuli [41]. VRIs with head-mounted displays isolate the user from external distractions and are thus well suited for catching the VRI user's attention. However, the learning outcomes in VRIs also rely on divided attention.

The constructive learning process in VRIs stems from constant multisensory inputs that depend on the individual's capacity to respond to multiple stimuli simultaneously. This requires executive shifts in focused attention according to the demands of the particular VRI scenario [34,41]. Attention and short-term memory processing are thus paramount abilities for constructing sustainable learning outcomes from multisensory learning experiences in VRIs. Cognitive load theory indicates that in VRIs, sensory input that is not directly involved in the learning

XSL•FO
RenderX

conveyed may worsen rather than promote the desired learning outcome [39]. Hence, limiting cognitive tasks in VRI scenarios for social skills learning is likely to promote the transfer of skills from digital to real environments. VRI scenarios that only expose the user to sensory input necessary to convey particular learning goals are therefore likely to promote efficient learning uptake.

Another key finding in this study is the participants' descriptions of their childhood and adolescence. The participants reported multiple experiences of adverse life events from early childhood, as well as exposure as adolescents to deviant social behavior in substance use environments. This explains some of the problems the participants described with social participation in mainstream society. Adverse childhood experiences are one of the most robust predictors of substance use and are commonly found among people with MHD/SUD [54,55].

The participants' reports of violence and intimidation in substance use environments are also in line with previous research [56,57]. Adverse childhood experiences and other deviant social experiences heavily affect the development of individuals' cognitive schemas. This is commonly found in people with MHD and SUD [58]. Adverse childhood experiences include a set of highly correlated traumatic and negative events that include child maltreatment (eg, sexual, physical, and verbal abuse) and household dysfunction (eg, parental divorce, familial substance use, mental illness, and incarceration) [54]. The cognitive schemas for interpersonal relationships that specify what to expect from others and how to treat them are formed during preadolescence, in particular. Young people living in stressful or harmful environments can form atypical schemas that enable them to survive adolescence but lead to serious problems when they are applied to adult relationships later in life [58].

Preadolescents exposed to stressful or otherwise difficult interpersonal environments where their emotional needs are not met develop dysfunctional beliefs about themselves and the world. Over time, this can lead to cognitive and emotional impairments that directly affect learning abilities through neurocognitive pathways [54]. When the onset of substance use and mental illness occurs before adult social skills are learned through natural processes, such as those the participants described, individuals will often form dysfunctional and maladaptive cognitive schemas for the requisite skills for independent functioning and personal coping. Maladaptive schemas cannot be altered or adapted to new situations or new social contexts in the same way as functional schemas. With accumulated exposure to emotional stress and a lack of empathic responses throughout the adolescent's development, these schemas become more complex, dense, and rigid. Hence, they also become increasingly difficult to change [58,59]. This hampers the capability of developing social abilities and adapting them to new social circumstances, compared with people with functional and adaptive cognitive schemas.

Hence, people with MHD and SUD often have serious difficulty in adopting appropriate social behavior in society at large. This is a key finding in regard to VR-learning experience design, as people's maladaptive cognitive schemas require restructuring or overwriting in order to acquire and retain social skills [58]. Overwriting or restructuring old cognitive schemas or constructing new ones requires repeated practice and overlearning to ensure the assimilation and retention of social skills. Such learning is facilitated when errors are minimized and correct responses are strengthened with abundant positive reinforcement [37].

Strengths and Limitations

This study comes with some limitations. A major limitation of this study is that our limited sampling did not encompass the full range of disparities in our multifaceted target group. This study was aimed at a particularly vulnerable population, which is traditionally hard to reach. The sample size (n=8) is small, but it is within Braun and Clarke's [60] recommendation of 6 - 10 participants for small projects. The empirical findings in this study are in line with a number of previous studies and widely supported in the field of mental health and substance use care [49,61,62]. It is also found that cognitive impairment affects daily and social functioning across the entire range of psychiatric diagnoses [41,47]. Altogether, this suggests the sample size was adequate to meet the research purpose of this study. Another limitation of this study is the demographic context.

This study was conducted in a Nordic welfare state. Thus, generalization to different demographic contexts may be limited since social participation in society will always depend on civil rights and social opportunities, as well as the individual's social capabilities. The main strength of this study is the human-centered design process and the peer-researcher collaboration throughout the research process. The interviews with 2 interviewers, using a peer researcher as primary interviewer, also made a valuable contribution to the depth of our interviews and the richness of our data.

Conclusions

Altogether, the theoretical framework and empirical findings provide new information on how we may develop VRIs that accommodate limitations in human perceptual processes. Regarding uptake factors for VRIs in people with MHD/SUD, this study showed wide disparities in digital literacy. However, the participants expressed receptiveness to VRIs, regardless of their digital literacy. The present empirical data and prior research show that this group has deficient immersive learning abilities and disadvantageous social learning prerequisites. Minimizing cognitive load in the VRI scenarios and orchestrating the VRI scenarios in a deliberate learning workflow may promote sustained real-life benefits from VRIs in people with mental health and substance use disorders.

Practical Implications

The findings in this study suggest that people with MHD/SUD require precision learning, minimized cognitive load, and spaced repetition to achieve social skills that are constantly relevant for use in real-world situations. VRI learning experience designs may hence be aimed at errorless learning, structured in short scenarios that can be repeated until the desired learning outcome in each scenario is achieved by applying the principles of micro learning and spaced repetition to the learning experience design.

XSL•FO
RenderX

Microlearning involves a careful task analysis of the required VRI learning content to identify specific learning units. When the required main learning units have been identified, the identified learning goals are broken down into short, focused units of learning content, called chunks or micro units [63]. These micro units constitute the learning goal of each scenario for learning social skills in the VRI. Short, focused VRI scenarios tailored to convey one micro unit of learning in a low-fidelity environment with decreased element interactivity in order to minimize cognitive task load may promote learning uptake [39,63].

Chunking VRI learning goals into microscenarios that can be repeated and structured according to individuals' learning abilities and prerequisites may enable the restructuring of maladaptive social schemas and promote the storage of repaired and new schemas in the VRI users' long-term memory. Ebbinghaus [64] showed that rehearsals should be distributed over time to become more efficient. Instead of repeating a task many times in the same session, it is more efficient to reduce the number of repetitions and spread them over several sessions [64]. When the users have achieved their required social skills, they will be able to benefit from practicing these skills in more complex VRI scenarios, aimed at applying adequate social skills to relevant social situations simulated in high-fidelity scenarios with high-reality proximity. This may enhance the learning sustainability, which in turn is vital for VRI's learned skills to be available in real-life settings. This way of learning and practicing social skills may enable VRI users to further develop and adapt their new cognitive schemas based on real-world social experiences in the same manner as people with functional and adaptable cognitive schemas.

Acknowledgments

The National Norwegian Advisory Board on Concurrent Substance Use and Mental Health Disorders initiated this study. The study was conducted as a client-driven, collaborative project between Innlandet Hospital Trust, VID Specialized University, and RIO, a Norwegian service users' association in the field of alcohol and drugs. The study was funded by the Research Council of Norway (project number 326773). The Research Council of Norway is a national strategic and funding agency for research activities, and a chief source of advice on and input to research policy for the Norwegian Government. No artificial intelligence was used in manuscript writing.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

JA, ML, and TS were responsible for the conceptualization of the study. JA, FN, and ML contributed to the study design and data collection. JA and FN conducted the formal analysis. Interpretation of the results was carried out by JA, FN, ML, and TS. JA drafted the original manuscript. JA, ML, and TS reviewed and edited the manuscript. Project administration was handled by JA and ML. ML acquired the funding for the study.

Conflicts of Interest

None declared.

References

1. World mental health report: transforming mental health for all. : World Health Organization; 2022.
2. EMCDDA. The drug situation in Europe up to 2023 – an overview and assessment of emerging threats and new developments (European Drug Report 2023). : European Union; 2023.
3. Kunstig intelligens i klinikken seks trender for fremtidens helsetjeneste [Report in Norwegian]. : Teknologirådet; 2021.
4. Volkow ND, Blanco C. Substance use disorders: a comprehensive update of classification, epidemiology, neurobiology, clinical aspects, treatment and prevention. World Psychiatry 2023 Jun;22(2):203-229. [doi: 10.1002/wps.21073] [Medline: 37159360]
5. Norton MJ, Swords C. Social recovery: a new interpretation to recovery-orientated services – a critical literature review. JMHTEP 2021 Jan 13;16(1):7-20. [doi: 10.1108/JMHTEP-06-2020-0035]
6. Noordsy D, Torrey W, Mueser K, Mead S, O'Keefe C, Fox L. Recovery from severe mental illness: an intrapersonal and functional outcome definition. Int Rev Psychiatry 2002 Jan;14(4):318-326. [doi: 10.1080/0954026021000016969]
7. Freeman D, Reeve S, Robinson A, et al. Virtual reality in the assessment, understanding, and treatment of mental health disorders. Psychol Med 2017 Oct;47(14):2393-2400. [doi: 10.1017/S003329171700040X] [Medline: 28325167]
8. Riches S, Pisani S, Bird L, Rus-Calafell M, Garety P, Valmaggia L. Virtual reality-based assessment and treatment of social functioning impairments in psychosis: a systematic review. Int Rev Psychiatry 2021 May;33(3):337-362. [doi: 10.1080/09540261.2021.1918648] [Medline: 34121587]

XSL•FO
RenderX

9. Schroeder AH, Bogie BJM, Rahman TT, Thérond A, Matheson H, Guimond S. Feasibility and efficacy of virtual reality interventions to improve psychosocial functioning in psychosis: systematic review. JMIR Ment Health 2022 Feb 18;9(2):e28502. [doi: 10.2196/28502] [Medline: 35179501]
10. Cieślik B, Mazurek J, Rutkowski S, Kiper P, Turolla A, Szczepańska-Gieracha J. Virtual reality in psychiatric disorders: a systematic review of reviews. Complement Ther Med 2020 Aug;52:102480. [doi: 10.1016/j.ctim.2020.102480] [Medline: 32951730]
11. Bell IH, Nicholas J, Alvarez-Jimenez M, Thompson A, Valmaggia L. Virtual reality as a clinical tool in mental health research and practice. Dialogues Clin Neurosci 2020 Jun;22(2):169-177. [doi: 10.31887/DCNS.2020.22.2/lvalmaggia] [Medline: 32699517]
12. Piot MA, Dechartres A, Guerrier G, et al. Effectiveness of simulation in psychiatry for initial and continuing training of healthcare professionals: protocol for a systematic review. BMJ Open 2018 Jul 11;8(7):e021012. [doi: 10.1136/bmjopen-2017-021012] [Medline: 29997139]
13. Wiebe A, Kannen K, Selaskowski B, et al. Virtual reality in the diagnostic and therapy for mental disorders: a systematic review. Clin Psychol Rev 2022 Dec;98:102213. [doi: 10.1016/j.cpr.2022.102213] [Medline: 36356351]
14. Rudschies C, Schneider I. Ethical, legal, and social implications (ELSI) of virtual agents and virtual reality in healthcare. Soc Sci Med 2024 Jan;340:116483. [doi: 10.1016/j.socscimed.2023.116483] [Medline: 38064822]
15. Kouijzer M, Kip H, Bouman YHA, Kelders SM. Implementation of virtual reality in healthcare: a scoping review on the implementation process of virtual reality in various healthcare settings. Implement Sci Commun 2023 Jun 16;4(1):67. [doi: 10.1186/s43058-023-00442-2] [Medline: 37328858]
16. Dellazizzo L, Potvin S, Luigi M, Dumais A. Evidence on virtual reality-based therapies for psychiatric disorders: meta-review of meta-analyses. J Med Internet Res 2020 Aug 19;22(8):e20889. [doi: 10.2196/20889] [Medline: 32812889]
17. Rus-Calafell M, Garety P, Sason E, Craig TJK, Valmaggia LR. Virtual reality in the assessment and treatment of psychosis: a systematic review of its utility, acceptability and effectiveness. Psychol Med 2018 Feb;48(3):362-391. [doi: 10.1017/S0033291717001945] [Medline: 28735593]
18. Jahn FS, Skovbye M, Obenhausen K, Jespersen AE, Miskowiak KW. Cognitive training with fully immersive virtual reality in patients with neurological and psychiatric disorders: a systematic review of randomized controlled trials. Psychiatry Res 2021 Jun;300:113928. [doi: 10.1016/j.psychres.2021.113928] [Medline: 33857847]
19. Rowland DP, Casey LM, Ganapathy A, Cassimatis M, Clough BA. A decade in review: a systematic review of virtual reality interventions for emotional disorders. Psychosoc Interv 2022 Jan;31(1):1-20. [doi: 10.5093/pi2021a8] [Medline: 37362616]
20. Nesvåg S, McKay JR. Feasibility and effects of digital interventions to support people in recovery from substance use disorders: systematic review. J Med Internet Res 2018 Aug 23;20(8):e255. [doi: 10.2196/jmir.9873] [Medline: 30139724]
21. Bell I, Pot-Kolder RMCA, Wood SJ, et al. Digital technology for addressing cognitive impairment in recent-onset psychosis: a perspective. Schizophr Res Cogn 2022 Jun;28:100247. [doi: 10.1016/j.scog.2022.100247] [Medline: 35281550]
22. Ossenbrink L, Haase T, Timpel P, et al. Effectiveness of digital health interventions containing game components for the self-management of type 2 diabetes: systematic review. JMIR Serious Games 2023 Jun 1;11:e44132. [doi: 10.2196/44132] [Medline: 37261900]
23. Kaplan AD, Cruit J, Endsley M, Beers SM, Sawyer BD, Hancock PA. The effects of virtual reality, augmented reality, and mixed reality as training enhancement methods: a meta-analysis. Hum Factors 2021 Jun;63(4):706-726. [doi: 10.1177/0018720820904229] [Medline: 32091937]
24. Mann S, Furness T, Yuan Y, Iorio J, Wang Z. All reality: virtual, augmented, mixed (x), mediated (x,y), and multimediated reality. . Preprint posted online on Apr 20, 2018. [doi: 10.48550/arXiv.1804.08386]
25. Jacobs JW, Prince C, Hays RT, et al. A meta analysis of the flight simulator training research. : United States Navy; 1990.
26. Makransky G, Lilleholt L. A structural equation modeling investigation of the emotional value of immersive virtual reality in education. Educ Tech Res Dev 2018 Oct;66(5):1141-1164. [doi: 10.1007/s11423-018-9581-2]
27. Redlinger E, Shao C. Comparing brain activity in virtual and non-virtual environments: a VR & EEG study. Measurement: Sensors 2021 Dec;18:100062. [doi: 10.1016/j.measen.2021.100062]
28. Kuhail MA, ElSayary A, Farooq S, Alghamdi A. Exploring immersive learning experiences: a survey. Informatics (MDPI) 2022;9(4):75. [doi: 10.3390/informatics9040075]
29. Mystakidis S, Berki E, Valtanen JP. Deep and meaningful e-learning with social virtual reality environments in higher education: a systematic literature review. Appl Sci (Basel) 2021;11(5):2412. [doi: 10.3390/app11052412]
30. Britain S. A review of learning design: concept, specifications and tools. : Jisc; 2004.
31. van Kessel R, Roman-Urrestarazu A, Anderson M, et al. Mapping factors that affect the uptake of digital therapeutics within health systems: scoping review. J Med Internet Res 2023 Jul 25;25:e48000. [doi: 10.2196/48000] [Medline: 37490322]
32. Kolb DA. Experiential Learning: Experience as the Source of Learning and Development: Engelwood Cliffs, NJ: Prentice Hall; 2014.
33. Zigmont JJ, Kappus LJ, Sudikoff SN. Theoretical foundations of learning through simulation. Semin Perinatol 2011 Apr;35(2):47-51. [doi: 10.1053/j.semperi.2011.01.002] [Medline: 21440810]

XSL•FO
RenderX

34. Mayer RE, Moreno R, Boire M, Vagge S. Maximizing constructivist learning from multimedia communications by minimizing cognitive load. J Educ Psychol 1999;91(4):638-643. [doi: 10.1037/0022-0663.91.4.638]
35. Fenwick T, Tennant M. Understanding adult learners. In: Foley G, editor. Dimensions of Adult Learning: Adult Education and Training in a Global Era: Routlege; 2004:55-73.
36. Fenwick T, Tennant M. Understanding adult learners. In: Foley G, editor. Dimensions of Adult Learning: Adult Education and Training in a Global Era: Routledge; 2020:55-73.
37. Kopelowicz A, Liberman RP, Zarate R. Recent advances in social skills training for schizophrenia. Schizophr Bull 2006 Oct;32(Suppl 1):S12-S23. [doi: 10.1093/schbul/sbl023] [Medline: 16885207]
38. Bledsoe R, Richardson D. Maximizing learning using cognitive load theory. ETH Learn Teach J 2020;2(2):148-152.
39. van Merriënboer JJG, Sweller J. Cognitive load theory in health professional education: design principles and strategies. Med Educ 2010 Jan;44(1):85-93. [doi: 10.1111/j.1365-2923.2009.03498.x] [Medline: 20078759]
40. Atkinson RC, Shiffrin RM. Human memory: a proposed system and its control processes. In: Spence KW, Spence JT, editors. The psychology of learning and motivation: Academic Press; 1968:89-195. [doi: 10.1016/S0079-7421(08)60422-3]
41. Millan MJ, Agid Y, Brüne M, et al. Cognitive dysfunction in psychiatric disorders: characteristics, causes and the quest for improved therapy. Nat Rev Drug Discov 2012 Feb 1;11(2):141-168. [doi: 10.1038/nrd3628] [Medline: 22293568]
42. Vehovar V, Toepoel V, Steinmetz S. Non-probability sampling. In: Wolf C, editor. The Sage Handbook of Survey Methodology: Sage; 2016:329-345. [doi: 10.4135/9781473957893.n22]
43. King N, Brooks J, Tabari S. Template analysis in business and management research. In: Ciesielska M, Jemielniak D, editors. Qualitative Methodologies in Organization Studies: Palgrave Macmillian; 2018:179-206. [doi: 10.1007/978-3-319-65442-3_8]
44. Braun V, Clarke V. Can I use TA? should I use TA? should I not use TA? comparing reflexive thematic analysis and other pattern‐based qualitative analytic approaches. Couns and Psychother Res 2021 Mar;21(1):37-47. [doi: 10.1002/capr.12360]
45. Braun V, Clarke V. Conceptual and design thinking for thematic analysis. Qualitative Psychology 2022;9(1):3-26. [doi: 10.1037/qup0000196]
46. Braatveit K. Intellectual disability among in-patients with substance use disorders. : University of Bergen; 2018.
47. Chavez-Baldini U, Nieman DH, Keestra A, et al. The relationship between cognitive functioning and psychopathology in patients with psychiatric disorders: a transdiagnostic network analysis. Psychol Med 2023 Jan;53(2):476-485. [doi: 10.1017/S0033291721001781] [Medline: 34165065]
48. Hetland J, Braatveit KJ, Hagen E, Lundervold AJ, Erga AH. Prevalence and characteristics of borderline intellectual functioning in a cohort of patients with polysubstance use disorder. Front Psychiatry 2021;12:651028. [doi: 10.3389/fpsyt.2021.651028] [Medline: 34335320]
49. Manning V, Verdejo-Garcia A, Lubman DI. Neurocognitive impairment in addiction and opportunities for intervention. Curr Opin Behav Sci 2017 Feb;13:40-45. [doi: 10.1016/j.cobeha.2016.10.003]
50. Collings S, Allan J, Munro A. Improving treatment for people with cognitive impairment and substance misuse issues: lessons from an inclusive residential treatment program pilot in Australia. Disabil Health J 2022 Jun;15(2S):101295. [doi: 10.1016/j.dhjo.2022.101295] [Medline: 35365421]
51. Marceau EM, Berry J, Lunn J, Kelly PJ, Solowij N. Cognitive remediation improves executive functions, self-regulation and quality of life in residents of a substance use disorder therapeutic community. Drug Alcohol Depend 2017 Sep 1;178:150-158. [doi: 10.1016/j.drugalcdep.2017.04.023] [Medline: 28651150]
52. Ramey T, Regier PS. Cognitive impairment in substance use disorders. CNS Spectr 2019 Feb;24(1):102-113. [doi: 10.1017/S1092852918001426] [Medline: 30591083]
53. Hofmann W, Schmeichel BJ, Baddeley AD. Executive functions and self-regulation. Trends Cogn Sci (Regul Ed) 2012 Mar;16(3):174-180. [doi: 10.1016/j.tics.2012.01.006]
54. Rogers CJ, Pakdaman S, Forster M, et al. Effects of multiple adverse childhood experiences on substance use in young adults: a review of the literature. Drug Alcohol Depend 2022 May 1;234:109407. [doi: 10.1016/j.drugalcdep.2022.109407] [Medline: 35306395]
55. Farrugia PL, Mills KL, Barrett E, et al. Childhood trauma among individuals with co-morbid substance use and post-traumatic stress disorder. Ment Health Subst Use 2011 Nov;4(4):314-326. [doi: 10.1080/17523281.2011.598462]
56. Schifano F, et al. Substance-use disorders and violence. In: Carpinelli B, Vita A, Menacci C, editors. Violence and Mental Disorder: Springer; 2019:95-114. [doi: 10.1007/978-3-030-33188-7_6]
57. Chermack ST, Bonar EE, Ilgen MA, et al. Developing an integrated violence prevention for men and women in treatment for substance use disorders. J Interpers Violence 2017 Feb;32(4):581-603. [doi: 10.1177/0886260515586369] [Medline: 26002873]
58. André N, Baumeister RF. Dysfunctional schemas from preadolescence as one major avenue by which meaning has impact on mental health. Int J Environ Res Public Health 2023 Jun 25;20(13):6225. [doi: 10.3390/ijerph20136225] [Medline: 37444073]
59. Young JE, Klosko JS, Weishaar ME. Schema Therapy: A Practitioner's Guide: Guilford Press; 2006.
60. Braun V, Clarke V. Successful Qualitative Research: A Practical Guide for Beginners: Sage; 2013:1-400.

XSL•FO
RenderX

61. Aasen J, Galaaen K, Nilsson F, Sørensen T, Lien L, Leonhardt M. Promoting social participation and recovery using virtual reality-based interventions among people with mental health and substance use disorders: qualitative study. JMIR Form Res 2023 Apr 27;7:e46136. [doi: 10.2196/46136] [Medline: 37104000]
62. Aasen J. Møter med vold og suicidalitet [Book in Norwegian]. In: Problemer S, Sammenvede Tiltak LL, Lie TW, editors. Sammensatte Problemer, Sammenvede Tiltak: Fabokforlaget: Bergen, Norway; 2022:195-218.
63. Thillainadesan J, Le Couteur DG, Haq I, Wilkinson TJ. When I say … microlearning. Med Educ 2022 Aug;56(8):791-792. [doi: 10.1111/medu.14848] [Medline: 35654438]
64. Wollstein Y, Jabbour N. Spaced effect learning and blunting the forgetfulness curve. Ear Nose Throat J 2022 Nov;101(9_suppl):42S-46S. [doi: 10.1177/01455613231163726]

Abbreviations

MHD: mental health disorder
PTSD: posttraumatic stress disorder
SUD: substance use disorder
VR: virtual reality
VRI: virtual reality–based intervention

Edited by A Mavragani; submitted 22.02.24; peer-reviewed by A Adekoya, E Mitsea, R Pot-Kolder; revised version received 04.05.25; accepted 06.05.25; published 20.08.25.

Please cite as:
Aasen J, Nilsson F, Sørensen T, Leonhardt M
Promoting Sustained Real-Life Benefits of Virtual Reality–Based Interventions in People With Mental Health and Substance Use Disorders: Qualitative Study
JMIR Form Res 2025;9:e57643
URL: https://formative.jmir.org/2025/1/e57643
doi:10.2196/57643

© Jan Aasen, Fredrik Nilsson, Torgeir Sørensen, Marja Leonhardt. Originally published in JMIR Formative Research (https://formative.jmir.org), 20.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Evaluating the Prototype of a Clinical Decision Support System in Primary Care: Qualitative Study

Susanne M Köhler[1], Dr oec troph; Svea Holtz[1], Dr med; Michaela C Neff[2], MSc; Jannik Schaaf[2], Prof Dr; Michael von Wagner[3], Dr med; Beate S Müller[4], Prof Dr; Dania Schütze[1], Dr rer med

[1]Institute of General Practice, Goethe University Frankfurt, Theodor-Stern-Kai 7, Frankfurt am Main, Germany

[2]Institute of Medical Informatics, Goethe University Frankfurt, University Medicine Frankfurt, Frankfurt am Main, Germany

[3]Department of Medical Information Systems and Digitalization, University Medicine Frankfurt, Frankfurt am Main, Germany

[4]Institute of General Practice, University of Cologne, Faculty of Medicine and University Hospital Cologne, Cologne, Germany

Corresponding Author:
Susanne M Köhler, Dr oec troph
Institute of General Practice, Goethe University Frankfurt, Theodor-Stern-Kai 7, Frankfurt am Main, Germany

Abstract

Background: General practitioners are confronted with a wide variety of diseases and sometimes diagnostic uncertainty. Clinical decision support systems could be valuable to improve diagnosis, but existing tools are not adapted to the requirements and workflow in the primary setting. In project SATURN (Smart physician portal for patients with unclear disease), the prototype of a clinical decision support system based on artificial intelligence is being developed together with users specifically for primary care in Germany. It aims to reduce diagnostic uncertainty in cases of unclear and rare diseases and focuses on 3 medical fields. A user-centered design approach is applied for prototype development and evaluation.

Objective: This study evaluates the usability of a high-fidelity prototype and explores aspects of user experience like the subjective impression, satisfaction, and areas of improvement.

Methods: A total of 5 general practitioners participated in the evaluation, which consisted of (1) a remote think-aloud test, (2) a postsession interview, and (3) a survey with the System Usability Scale. During the think-aloud tests, the participants verbalized their thoughts and actions and solved several vignette-based tasks. Remarkable observations were logged, transcribed with quotes, and analyzed for usability problems and positive findings. All observations and interview responses were deductively assigned to the following categories: (1) content, (2) comprehensibility, (3) user-friendliness, (4) layout, (5) feedback, and (6) navigation. Usability problems were described in detail and solutions for improvement proposed. Median and total scores were calculated for all questionnaire items.

Results: The evaluation detected both strengths and areas for improvement. The participants particularly liked the clear and well-structured layout of the prototype. Key issues identified were content-related limitations, such as the inability to enter unlisted symptoms, medications, and examination findings. Also, participants found the terminology for laboratory not suitable to their needs. Another key issue was a lack of user-friendliness concerning the time required to input medication plans and lab values. Participants expressed a need for faster data entry, potentially through direct imports from practice management systems or laboratory files. Adding symptom duration, weighting symptoms, and incorporating hereditary factors were suggestions made for improvement. Overall, the SATURN prototype was deemed useful and promising for future clinical use, despite the need for further refinements, particularly in the areas of data entry, as this is a key obstacle to its use.

Conclusions: The usability evaluation methods combined proved to be location independent and easy to use. They provided important findings on usability issues and improvements that will be implemented in a second high-fidelity prototype, which will also be tested by users. Technically demanding user requirements, such as direct data transfer from the practice management system and entry options that require complex data models, were beyond the scope of this project, but should be considered in future development projects.

(JMIR Form Res 2025;9:e69875) doi:10.2196/69875

KEYWORDS

clinical decision support systems; primary care; user-centered design; usability testing; qualitative research

XSL•FO
RenderX

Introduction

General practitioners (GPs) are confronted with a wide variety of diseases and sometimes with diagnostic uncertainty [1]. As a large number of rare diseases are not commonplace in primary care, it is impossible for GPs to know all these symptoms and make differential diagnoses based purely on their experience and knowledge. Recognizing patterns in symptom complexes does not work in this case [2]. This can be very burdensome for patients, as the diagnosis of a rare disease can often take years. During this time, misdiagnosis and numerous consultations are the rule, and sometimes the diagnosis nevertheless remains unclear [3-5]. For GPs, clinical decision support systems (CDSS) could be especially valuable to improve diagnosis of frequent diseases with unusual manifestations of symptoms or rare diseases. However, "Google Search" or regular symptom checkers are not apt for the medical workflow, and specific tools for this setting are lacking. According to a scoping review of Harada et al [6], screening and diagnosis of chronic diseases are the main targets of CDSS in a primary care setting, but uncommon chronic diseases are not covered. A mixed-methods systematic review evaluating CDSS providing recommendations to primary care professionals by Meunier et al [7] including studies from 1998 to 2021 revealed that most systems were developed and used in the United States. Diagnosis was only presented in 7/45 of the investigated CDSS and only 1 system aiming at iatrogenic prevention was developed for the German market.

Rare diseases—defined in the European Union as diseases with a prevalence of not more than 5 affected persons per 10,000 [8]—are especially challenging in terms of diagnosis. Diagnosis support is supplied by specific databases for rare diseases (eg, "Orphanet") and 3 types of specific search engines: most of them are based on the analysis or comparison of genetic and phenotypic data (eg, "Phenomizer"), some on information retrieval (eg, "FindZebra"), and few on machine learning. Often, they lack direct data integration and do not provide information on when the system was last updated. Not all are fully developed systems or publicly available. Furthermore, these tools are not adapted to the requirements, needs, and workflow of GPs and are not designed together with users [9].

In general, when successfully designing eHealth apps, the iterative involvement of potential users at different stages of the development process is necessary, which is the basis of a user-centered design process [10-12]. To evaluate prototypes, repeated usability testing is crucial. Various behavioral and technology acceptance models describe that factors such as ease of use are an important factor for adoption and actual use [13,14]. Quantitative methods, especially standardized questionnaires like the System Usability Scale (SUS), which give an overall measure of usability, are most frequently used, often alone, and sometimes in combination with qualitative instruments like task completion, think-aloud (TA), interviews, heuristic testing, and focus groups [15]. While effectively depicting general user acceptance and facilitating comparisons with other systems, standardized questionnaires may fall short in identifying the specifics and extent of inherent usability problems, which can be captured by the application of qualitative methods. However, the exact process of qualitative usability testing, analyzing, and the reporting in literature is rare [15].

In the "Smart physician portal for patients with unclear disease" (SATURN) project, funded by the German Federal Ministry of Health, an artificial intelligence (AI)-based CDSS is being developed with and exclusively for GPs to aid in diagnosing unclear and rare diseases. The proof-of-concept study focuses on 3 medical fields (gastroenterology, pneumology, and endocrinology) and provides decision support based on medical guidelines and clinical cases from university hospitals. These fields were selected because the university hospitals involved in the project have clinical expertise in these areas and the quality of the available data was deemed suitable for the planned measures.

A total of 3 different AI methods are applied, to our knowledge for the first time in a CDSS for unclear diagnosis: diagnosis support based on (1) rule-based systems, (2) case-based reasoning, and (3) machine learning. In addition, when a rare disease is detected, further assistance is offered in terms of direct connections to specific databases delivering disease-specific knowledge and support concerning specialized care centers and patient organizations [16].

In a previous study [17], we described the requirements analysis as the first stage of the user-centered design process of the CDSS SATURN, which iteratively involves a group of 5 GPs. Initially, user requirements were identified through workshops and interviews, and a task model was developed in the follow-up. Based on the requirements analysis and subsequent prioritizing regarding impact and technical feasibility, first mockups and then a first high-fidelity, clickable prototype of the portal's user interface were designed [18]. The latter was initially tested through usability inspections and refined [19].

In this publication, we report the evaluation of the first high-fidelity SATURN prototype. The objective of the study was to investigate usability problems of the CDSS (eg, in terms of data entry, comprehensibility, layout, etc) at an early stage, as well as positive findings, advantageous system solutions that should not be changed. Also, we wanted to determine the GPs' satisfaction with the CDSS and improvement requests. Furthermore, we put special emphasis on our method mix applied for usability testing and on the process of data analysis and reporting which is not well documented in literature.

Methods

Design

We used a primarily qualitative design to evaluate the first high-fidelity, clickable prototype of the SATURN portal with a small group of GPs. The AI modules, which will later provide the results for decision support, were not yet connected in this version. For the usability test, sample data for diagnostic suggestions was, therefore, stored in the system.

The study consisted of three parts: (1) a remote TA test to identify usability issues (problems and positive findings), (2) a postsession interview, and (3) an SUS questionnaire to evaluate

XSL•FO
RenderX

user satisfaction and find out improvement requests. All 3 parts were carried out consecutively in individual remote sessions. During the test, the GPs had to perform tasks and were advised to "think aloud," that is, to utter all thoughts, for example, what they were looking for, what they were expecting, or what was irritating them. This approach was chosen to understand the requirements, objectives, reasons, impressions, and sensations behind the users' observed behavior [20,21]. The researchers who conducted and analyzed the tests (SMK, SH, and DS) were not part of the development team. The study was conducted and reported in accordance with the Consolidated Criteria for Reporting Qualitative Research (COREQ) [22].

Setting and Sampling

A group of 5 GPs, male and female, was invited. All group members took part in the usability testing between July and August 2023. The participants were recruited by email using purposeful sampling [23], all were interested in digitalization topics and had already been involved in the requirements analysis. They received written information on the study and provided written informed consent. All tests were conducted remotely in individual online sessions via Zoom (Zoom Communication). The SATURN prototype was accessible via a URL and an assigned user and password and could be used in the web browser. All sessions were audio- and videotaped.

Ethical Considerations

The Ethics Committee of the Goethe University Frankfurt am Main approved the SATURN study (case 2022 - 1088). The Ethics Committee waived the need for a separate ethics vote on the inquiry of GPs (case 2022 - 629). All participants received written information on the study and provided a signed informed consent. All data were pseudonymized before analysis. Participants received €150 (average conversion rate in August 2023: US $1 = €0.92) for taking part in the usability test.

Data Collection

SATURN Prototype

SATURN consists of a start page (see Figure 1) with the following options: (1) Patient and (2) More information.

Figure 1. Screenshot: Start page.

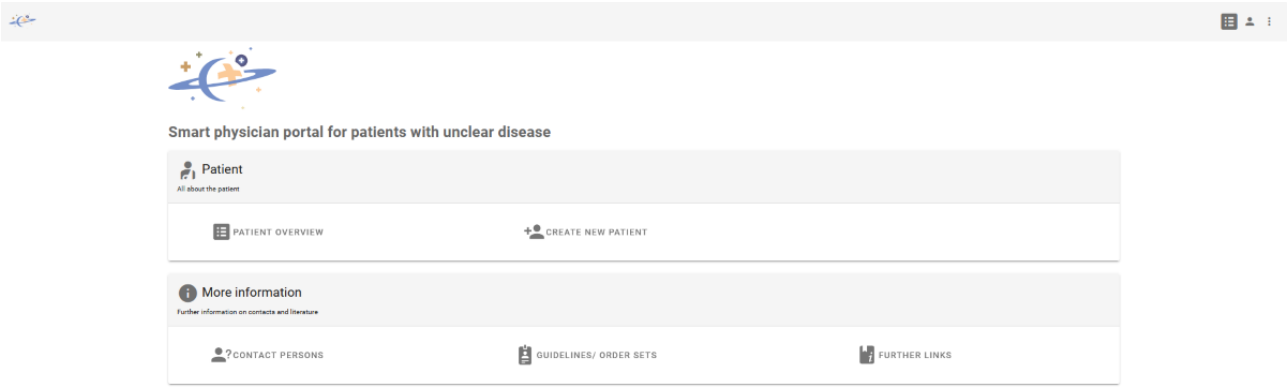

With the "Patient" option, logged-in GPs get an overview of all patients (Patient overview) and enter new patient data to receive diagnostic suggestions (Create new patient). There are seven steps for data entry under "Create new patient" for individual data categories (basic data, diagnoses, symptoms, medication, laboratory and vital values, and examinations) followed by an overview page of all entries (summary). From this data entry page, the GP can both access the page with possible AI diagnosis suggestions (results) and the case closure page where the GP can enter the final diagnosis once it has been confirmed.

With the "More information" option, GPs get more information on selected diseases and contact options (More information). In addition to contacts to specialists and support groups (Contacts), they will also find links to selected guidelines (Guidelines and Order sets) and websites (Further links).

A navigation bar on top leads through the data entry pages. Screenshots of all major SATURN pages are provided in Multimedia Appendix 1 (all screenshots are translated from German into English for this publication).

Remote Think-Aloud Test

The interdisciplinary study team consisted of researchers with backgrounds in sociology, medicine, medical sciences, and medical informatics, all of whom had experience with qualitative data collection. DS, SH, and SMK developed a test plan with test objectives, an interview guide, and a case vignette with 5 tasks to fulfill [10,20]. An internal pilot test was performed. All tests were moderated by SH.

The case vignette was based on a case report of Graves disease (an autoimmune thyroid disease) by Sharma [24]. The case report was selected as it depicted one of the SATURN disease patterns (endocrinology) and took place in an outpatient setting. It was shortened (parts about treatment and further course of the disease were removed) and information on medication dosages was added. It was narratively adapted to the primary care setting in Germany and checked for plausibility by an internist specializing in endocrinology. The following tasks, which were based on the task model developed in the requirements analysis [17], had to be performed: (1) entering patient data from the case vignette, (2) determining the most plausible diagnosis proposal, (3) returning to the start page, (4) concluding the case, and (5) finding information on patient organizations and care centers for rare diseases in the system.

Before the video conference, the participants received an email providing an access link and the case vignette with tasks (see Multimedia Appendix 2) but were advised not to study it before

XSL•FO
RenderX

the test. During the remote test, the GPs shared their screen while performing the tasks so that the process could be observed by the moderator and audio- and video-recorded (a written record was created later). The test persons verbalized their actions and what was going through their mind while using the system, for example, expectations, comprehension problems, or questions [11,20,25]. If necessary, the moderator supported the think-aloud process with questions like "What are you thinking?", "What are you looking at right now?", "What are you doing at the moment?", or "What would you be doing if I wasn't here"?

Postsession Interview

The postsession interview took place immediately after the usability test to evaluate the GPs' subjective impression of the CDSS and their desire for improvement (satisfaction and user experience). A total of three open-ended questions were asked: (1) What did you like about the system?, (2) What should we improve?, and (3) What is your overall impression: What do you think about the system?

System Usability Scale

The evaluation of the portal's usability was completed by applying the standardized SUS questionnaire, developed by Brooke [26]. The questionnaire was filled out online by the test persons right after the postsession interview. The SUS contains 10 items in the form of closed questions or statements on the assessment and learnability of software systems [27]. A total of 5 statements are formulated positively (eg, "I thought the system was easy to use"), 5 negatively (eg, "I found the system unnecessarily complex"). Agreement with the statements is measured using a 5-point Likert scale including verbal descriptions of the endpoints (1=strongly disagree and 5=strongly agree). We used the German version of the questionnaire [28].

Data Analysis

Remote Think-Aloud (TA) Test

Categories for Usability Issues

To evaluate the video recordings, a member of the study team (SH or SMK) created a protocol using Microsoft Excel 2016. Observations pointing out usability issues when carrying out the tasks were described, associated with quotes, and a time stamp was noted. The minute-takers checked each other's protocols for clarity and completeness. The observations were categorized into usability problems and positive findings. First, all detected usability problems were compiled in a table and sorted by location (On which system page did the problem occur?) and by similarity (Did several test persons experience the same problem?). Second, similar to the categories of Li et al [29] and Gorshid et al [30], who also evaluated user interfaces through TA-tests, we defined categories that depicted relevant aspects of the system's usability: (1) Content, (2) Comprehensibility, (3) User-friendliness, (4) Layout, (5) Feedback, and (6) Navigation (see Table 1). All findings were deductively assigned to these categories. In the course of the analysis, which was conducted by SH and SMK and supervised by DS, subcategories were inductively defined for each category to further classify the usability issues on a descriptive level. The subcategories were created in a discursive process between the researchers. Positive usability findings were summarized separately in an Excel file and sorted accordingly.

Table . Definition of categories for usability issues.

Category	Explanation of usability problem(s)
Content	The CDSS[a] texts are medically inaccurate or inappropriate or the terminology used is not understood by the test persons.
Comprehensibility	The meaning of texts, instructions, and the purpose of CDSS components cannot be quickly grasped by the test persons.
User-friendliness	The test persons are unable to handle the CDSS easily and use it with minimal effort (in terms of time, clicks, and mental energy).
Layout	The test persons are unable to recognize relevant information on the screen or they overlook it. They describe aesthetic problems and point out redundant information.
Feedback	The feedback the test persons receive from the user interface is not relevant or does not meet the users' expectations. A lack of feedback is criticized.
Navigation	The test persons express problems navigating the system.

[a]CDSS: clinical decision support system.

Description of Usability Problems

Finally, a detailed summary of each usability problem with the following information was created: (1) Title, (2) Severity, (3) Location, (4) Category, (5) Subcategory, (6) Description (partly with screenshot), (7) Number of test persons experiencing the problem, (8) Possible cause, and (9) Recommendation. A classification of 5 severity levels was used, based on the traditional Nielsen scale and its further development [31,32] (see Table 2). The severity level was determined in the team (SMK, SH, DS, and MCN), and the recommendations were developed by MCN.

XSL•FO
RenderX

Table . Definition of severity levels for usability problems.

Gradation	Severity level	Definition
Idea	0	The user makes a suggestion for an additional function.
Minor usability problem	1	The usability problem leads to temporary irritation of the user. The task can be completed.
Moderate usability problem	2	The usability problem leads to moderate irritation for the user. The task can be completed via detours or with a delay. Some users fail.
Critical usability problem	3	The usability problem causes considerable irritation or comprehension problems. The task cannot be completed.
Bug	4	Software error.

Postsession Interview

All responses were transcribed with quotes and time stamps, and analyzed according to structuring content analysis [33]. Quotes were assigned to the same categories and subcategories that were used for the usability issues. This made it possible to look at the results of the TA test and the interviews in conjunction.

System Usability Scale

With a large number of test persons, an overall score can be calculated from the SUS answers, representing the system's overall usability numerically [34]. Since the number of test subjects (n=5) in our study is not sufficient for a quantitative statement based on the overall score [35], we evaluated the median, total, and range per item. The aim was to identify possible outlier items and to triangulate the results with those from the usability tests and the postsession interviews.

Results

Participant and Interview Characteristics

The characteristics of the 5 participating GPs and the duration of the usability tests are presented in Table 3.

Table . Participant and interview characteristics (n=5).

Characteristic	Participant ID				
	1	2	3	4	5
Age (years)	44	45	34	50	36
Sex	Male	Female	Male	Male	Female
Additional qualifications or special focus	Emergency care, focus on geriatrics and palliative care.	Intensive and emergency care.	No additional qualifications.	Diabetologist, nephrologist, focus on hypertensiology and nutritional medicine.	Focus on rare diseases.
Time working as GP[a] (years)	10	7	2	9	3.5
Employment status	Self-employed	Employed	Employed	Self-employed	Employed
Duration of usability test and interviews (min:s)	55:50	56:00	47:47	37:15	42:50

[a]GP: general practitioner.

Remote Think-Aloud (TA Test) and Postsession Interview

Our results are presented below based on the defined categories (see Table 1). Individual findings are reported by way of example and supported by quotes. In addition to usability problems in the narrower sense (severity levels 1 - 3), the results include problems designated as "ideas," that is, missing functions that are not yet provided in the system (severity level 0). Also, positive usability findings are reported, that is, aspects that were praised by the participants and should be retained.

Table 4 provides an overview and brief description of all usability problems in the narrower sense (severity levels 1 - 3), including their frequency. Problem details can be found in Multimedia Appendix 3. Bugs (severity level 4) that occurred during the test were reported to the developers immediately after the sessions, fixed, and are not discussed further in this paper.

XSL•FO
RenderX

Table . Usability problems by category and subcategory.

Category	Subcategory	Usability problem Brief description	Persons (out of 5) who experienced the problem, n (%)
Content	Lab values input	Unusual naming of stored laboratory values.	5 (100)
		Inconsistent preset options for laboratory values.	1 (20)
		Contents of “Small blood count” and “Large blood count” not clear.	1 (20)
	Symptoms input	Input of symptoms not already stored in system not possible.	4 (80)
	Diagnoses input	Input of confirmed diagnosis only possible with *ICD*[a] code.	4 (80)
	Medication input	Input of medications not already stored in system not possible.	2 (40)
		Input of the dose for drug combinations unclear.	1 (20)
		“Night” medication administration time missing.	1 (20)
	Entering findings	Input of instrumental examination findings not already stored in system not possible.	1 (20)
	Entering vital signs	Imprecise labeling of blood pressure input field.	1 (20)
	ICD code display	Incomplete display of the *ICD* text.	1 (20)
Comprehensibility	Button label	Label of “Insert lab value set” button incomprehensible.	5 (100)
		“Positive/Negative” label not suitable for laboratory results.	2 (40)
		Label of “Back to overview” button imprecise.	1 (20)
		Label of “Go to case input” button incomprehensible.	1 (20)
		Label of “Contact persons” button not clear.	1 (20)
	Input field	Lack of instructions for input.	4 (80)
		Unclear function of the “Diagnostic history” input field.	1 (20)
	Icons	Unclear function of the Info icon.	1 (20)
		Inappropriate icons for diagnosis input.	1 (20)
	Explanatory text	Contradictory explanatory texts for diagnosis input.	1 (20)
User-friendliness	Time required	Excessive time required to input medication plan.	3 (60)
		Excessive time required to input lab values.	1 (20)
	Input support	Large number of stored units.	2 (40)
		Complicated input of confirmed diagnoses.	1 (20)
	Operation	Lack of export or print function.	1 (20)
		Lack of filter functions.	1 (20)

XSL•FO
RenderX

Category	Subcategory	Usability problem Brief description	Persons (out of 5) who experienced the problem, n (%)
Layout	Clarity	Poor localization of the laboratory value input field.	2 (40)
		Poor localization of the “Contact persons” button.	1 (20)
		Poor localization of the results list of patient organizations.	2 (40)
		Visibility of the findings text lacking.	1 (20)
	Input field	Poor recognition of the symptom input field.	1 (20)
	Structure	Awkward prioritization of medication preselection.	1 (20)
Feedback	Saving	Lack of confirmation when saving symptoms.	2 (40)
		Lack of confirmation when saving the confirmed diagnosis.	1 (20)
	Feedback lacking	Closed cases marking is lacking.	2 (40)
Navigation	Page navigation without any bar	Direct navigation back to home page not visible.	2 (40)
	Page navigation with bar	Nonadaptive labeling of the navigation bar.	1 (20)

[a]*ICD*: *International Classification of Diseases*.

Table 5 depicts a summary of all positive usability findings.

Table . Positive usability findings by category and subcategory.

Category	Subcategory	Brief description positive findings	Persons (out of 5) who mentioned the issue, n (%)
User-friendliness	Input support	Input support for diagnoses (*ICD*[a] code).	2 (40)
		Input support for symptoms (symptom list).	2 (40)
		Input support for medications (letters combination).	1 (20)
	Operation	Text can be transferred from other files.	1 (20)
Content	Lab values input	Input of capital letters possible.	1 (20)
	Probabilities	Diagnosis suggestions with indication of probability.	1 (20)
Layout	Clarity	Clear page design (start page and contact persons).	2 (40)
Navigation	Page navigation with bar	Quick movement through the pages possible.	1 (20)
	Further links	Links with contact details available.	1 (20)
Comprehensibility	Input field	Input field for notes available (Symptoms page).	1 (20)

[a]*ICD*: *International Classification of Diseases*.

Content

Many problems were of a content-related nature, including some that meant that a task could not be completed correctly. For example, it proved to be problematic that the participants did not have a suitable means of entering symptoms, medication, and examination findings that were not listed in the drop-down

XSL•FO
RenderX

menu. Some used the open text field "Notes" for this purpose, but entries in this field are not taken into account in the diagnosis search.

> *I can just write "daytime tiredness" and "nightmares," and "weight loss over weeks" in the comments [refers to the open text field "Notes"]. To give a time frame, I would add that too, as well as "gastrointestinal symptoms for about 4 weeks."* [GP 2, Usability problem 7]

When entering the laboratory values, some of the terminology used for the stored laboratory values was not appropriate. All test persons tried to search for laboratory values under common abbreviations (eg, TSH and GFR) but did not find them.

> *Now I'll try again with the TSH [thyroid stimulation hormone]. Maybe I have to scroll there - at least that's how it was with the others too. No, nothing comes up. No, it doesn't seem to exist.* [GP1, Usability problem 18]

The presentation of the results with probabilities, which corresponded to the expectations of the test subjects, was positively emphasized.

> *Yes, that's good with the probability indication here in percentage, I know it from Ada Health too, yes, great.* [GP 1, Positive finding]

During the usability tests, the GPs also made some suggestions as to how the content of the CDSS could possibly be improved (ideas). For example, it was suggested that the duration of symptoms should be requested and that the weighting of symptoms should be made possible. It was also suggested that the important information "rhythmic/arrhythmic" should be added to the heart rate and that physical examination findings should be entered (eg, enlarged liver or spleen). The interviews also emphasized the importance of family history in making a diagnosis.

> *What would not be bad either would be a sub-item on previous family illnesses…especially with regard to heredity, genetic predisposition - something like that would be very important [GP mentions diabetes, coronary heart disease] … that is definitely part of it.* [GP 2, Interview]

Comprehensibility

The usability test made it clear that some of the labeling and functions of buttons, input fields, and symbols were imprecise or incomprehensible. All test subjects had problems understanding the name and function of the "Insert lab value set" button, which was intended to support the input of the large and small blood count.

> *Can I click on here, somehow...it's not at this point...it's not quite clear to me why it's called something like a laboratory value set, but then there's only the option of large and small blood count. And if I want something else, I have to enter it at the bottom here. I don't quite understand that at this point, but it seems to be intentional, so I probably click on this here now.* [GP 1, Usability problem 23]

Some participants were unclear about the function of the "Diagnostic history" input field and the meaning of some symbols. A total of 4 people did not understand what should be entered on the "Case closure" page. Therefore, the corresponding task could not be completed.

> *Where am I now? Diagnostic accuracy. I'm not quite sure what I'm supposed to do at this point. Should I enter the final diagnosis? Or the presumptive diagnosis it churned out for me?* [GP 1, Usability problem 31]

User-Friendliness

The test persons rated the time required to enter the medication plan and laboratory values as not user-friendly. The large number of selectable laboratory value units also irritated users and delayed entry.

> *[GP 5 enters medication with dosage, form of administration, and unit.] And here too, now I'm already really annoyed that I somehow […]. Again, it's all too detailed for me. It would take far too long in the consultation for me.* [GP5, Usability problem 9]

During the interviews, 3 GPs again discussed the time-consuming entry of medication, laboratory values, and diagnoses. In this context, the desire for direct import from the practice management system was expressed several times.

> *I found it difficult to enter medication and laboratory values... and then having to enter the standard value myself, that kills it a bit'...or the best would be to import it directly from the system.* [GP 3, Interview]

To further improve user-friendliness, a test person suggested that it should be possible to mark diagnoses that have already been excluded in the results list.

> *Of course it would be nice for me if I could now exclude things directly… [gives an example]. if I could cross them out directly or something.* [GP 3, Idea 5]

The principle of input support through stored diagnoses, symptoms, and medications used on several pages appealed to the participants and was rated as user-friendly.

> *Ok, you also entered the ICD code. That's good.* [GP 5, Positive findings]

Layout

The GPs generally liked the clear, well-structured layout of the portal, with 4 people praising it in the interviews.

> *Relatively, "clean," ie, clear, not so many incomprehensible things, kept simple and easy to start, largely self-explanatory.* [GP 3, Interview]

However, a lack of clarity was observed regarding details on 2 pages. For example, 2 test subjects had to search for the field for entering new laboratory values. On the "Contact persons" page, only clinic contacts were directly recognizable on the screen. The list of patient organizations below was only accessible by scrolling and was not found by 2 test persons.

XSL•FO
RenderX

Feedback

The test persons missed provision of feedback from the system when saving the symptoms and the confirmed diagnosis. The fact that whether or not a case had already been closed was not recognizable in the patient overview led to considerable irritation.

> *And then I go to "Close case"...Okay. Then I go again here to "Close case," because I would have expected it to say that the case is closed… but… so I would have thought that now there would be a check mark, or it says "Case closed", but well, okay.* [GP 2, Usability problem 34]

Navigation

All GPs found the navigation bar at the top of the screen, which led them through the various pages of the SATURN portal, very easy to use. In the interviews, the navigation bar was praised several times.

> *The navigation in the portal is a great solution... that I can jump directly to these individual items with master data and symptoms in this bar at the top... I think that's very good.* [GP 5, Interview]

Overall Rating

Overall, the final personal assessment in the interviews was positive for all test persons, and the portal was rated as a very useful, helpful, and interesting diagnosis support tool and as a not unlikely option for the future. With few exceptions, the GPs succeeded in completing the tasks.

> *I think it's still very much in its infancy... I'm really excited to see what comes out of it in the end. I'm still convinced that it makes total sense and that it can still be really good.* [GP 2, Interview]

The subjectively perceived usability of the SATURN prototype was also positive in many respects (see Table 6). Using the SUS questionnaire, the median of the positively formulated items was 4 and 5 respectively (best rating: 5), and the median of the negatively formulated items was 2 and 1 respectively (best rating: 1). Overall, the highest approval was given to items concerning the ease of use and the speed with which the system can be learnt. Correspondingly, items stating the need for technical support and the necessity to learn a lot before getting started met with no approval. This corresponds to the findings of the usability test and interviews, especially the praise of the well-structured layout, the user-friendly input support, and the easy and intuitive navigation. No outlier items were identified, but some test persons found that there was too much inconsistency in the system.

Table . Usability rating by means of System Usability Scale (SUS) items (n=5).

Item	Median (minimum-maximum)	Summed rating of all participants
Positively formulated items		
I thought the system was easy to use.	5 (3-5)	23
I would imagine that most people would learn to use this system very quickly.	5 (3-5)	23
I found the various functions in this system were well integrated.	5 (3-5)	22
I think that I would like to use this system frequently.	4 (3-5)	20
I felt very confident using the system.	4 (3-5)	20
Negatively formulated items		
I thought there was too much inconsistency in this system.	2 (1-4)	11
I found the system unnecessarily complex.	1 (1-4)	8
I found the system very cumbersome to use.	1 (1-3)	8
I think that I would need the support of a technical person to be able to use this system.	1 (1-3)	7
I needed to learn a lot of things before I could get going with this system.	1 (1-1)	5

Discussion

Principal Findings

The evaluation highlighted strengths, usability problems, and areas for improvement. Key issues included content limitations, such as the inability to enter unlisted symptoms and medications, as well as unfamiliar terminology for laboratory values. In addition, data entry for medication and lab values was seen as cumbersome, with a need for faster input methods. Despite these challenges, participants appreciated the user-friendly navigation and clear layout, considering the prototype promising for future clinical use, though refinements in data entry are needed.

The remote setting allowed location-independent and natural environment testing for GPs, simulating a realistic use case of

XSL•FO
RenderX

the portal. The use of a case vignette with tasks also contributed to this. The TA method was quickly adapted and well-accepted by the GPs. It provided us with detailed feedback on already well-implemented user requirements, for example, the clear and concise structure and the simple, intuitive operation of the interface. It also helped us identify the extent and precise causes of the usability issues that were found.

The basic structure of the system was evaluated positively, and the problems encountered seem in line with the status of the prototype (eg, imprecise or incomprehensive labels, buttons or symbols, lack of feedback, etc). For further development, it was important to obtain feedback on the usability of the user interface at this stage. At the same time, dealing with some problems and suggestions is not easy. For example, aspects such as the desire for simple, straightforward entry on the one hand and the need for particularly detailed data collection on the other hand have to be reconciled. During the usability tests, the GPs made some suggestions as to how the content of the CDSS could possibly be improved (ideas). For example, they suggested that the duration of symptoms should be requested, that the weighting of symptoms should be enabled and that hereditary factors should be considered. Although these factors may be relevant for diagnosis, such data would require substantial amendments on the portal's database (university hospital data and elements), as well as complex algorithms and data models [36,37], which could not be met in our proof-of-concept project. However, such factors would be valuable and should be considered in the future.

It proved to be problematic that the participants did not have a suitable means of entering symptoms, medication, and examination findings that were not listed in the drop-down menu. Although the introduction and design of additional open-entry fields are easily performed, the potential input must also be mapped in the data model to contribute to diagnosis.

The test persons rated the time required to enter the medication plan and laboratory values as not user-friendly. The large number of selectable laboratory value units, which is due to different terminology standards being used (research vs practice and laboratory standard), also irritated users and delayed entry. However, the user requirement for quick data entry is, according to our test results, likely to be a crucial factor for the acceptance of CDSS in general practice settings. Increased workload also proved to be the greatest barrier to using CDSS in primary care in the systematic review by Meunier et al [7]. In this study, CDSS slowness was also addressed as an important technological barrier, next to missing user-friendliness and a lack of integration in the electronic health record (although 28/45 CDSS included in this study were already partially or fully integrated in the electronic health record). Another major organizational barrier identified in this study, which corresponds to our results, was the disruption of the usual workflow. The automated direct transfer of data from practice management systems or from the laboratory files remains a technical challenge for future development projects. This applies especially to countries like Germany where a wide variety of management systems and laboratory standards are being used [38,39]. Here, individual solutions for data transfer need to be developed for each system [40]. To address this issue, a workshop will be held later in the project with representatives from practice management software companies and other experts to discuss possibilities for interoperability standards for practice management systems in Germany. The results will be reported elsewhere.

Overall, the evaluation of the SATURN prototype shows that a system has been created that can be easily integrated into the workflow of GPs and meets the requirements of the participants to a high degree. Such a system offers good prerequisites for being used in everyday practice diagnosis searches and for contributing to faster and better diagnosis finding. At the current early stage of development, however, it is not yet possible to make any statements about the effectiveness and benefits in everyday practice.

The prototype of SATURN is now being further developed, and the usability issues and suggestions are being technically implemented. The next version of the prototype will be tested with 10 GPs, 5 of whom are not yet familiar with the CDSS. The respective evaluation will also include the presentation of AI-based diagnostic suggestions.

Strengths and Limitations

The combination of the TA usability test, the postsession interview, and the standardized questionnaire provided a comprehensive overall picture of the portal's usability. Classifying the findings into categories during the analysis enabled us to systematize and triangulate results. The detailed problem descriptions served as the basis for finding appropriate problem solutions and further system development.

A limitation of testing the usability of the SATURN prototype in a preliminary stage of development is that the AI modules were not yet working, so the test results had to be simulated. Therefore, it was not possible to fully evaluate the functionality and the perceived usefulness and acceptance of the system.

Another limitation might be due to the limited sample size in combination with the continuous involvement of the 5 GPs in the development process. For qualitative usability testing, this group size is common and recurring problems can be sufficiently recognized [10,20,41]. When selecting the participants, we made sure to create a diverse sample in terms of gender, age, professional experience, and specialization. During the development process, numerous ideas developed by the GPs were gradually implemented. This could have caused a bias in the subsequent portal evaluation. For this reason, additional GPs will be included in the next evaluation.

Conclusions

By combining usability tests, postsession interviews, and a standardized questionnaire, a high-fidelity CDSS prototype for GPs was efficiently evaluated with potential users in terms of usability and user satisfaction. The qualitative tools applied were location-independent and easy to use. Solutions for the usability problems detected could be developed and will be implemented in the next prototype. However, the implementation of some technically demanding user requirements, such as direct data transfer from the practice management system and entry options that require complex

XSL•FO
RenderX

data models, was beyond the scope of this project. They should be considered in future development projects. In the next iteration, the AI modules for generating results will be integrated, and the presentation of the results will be tested. In addition to usability, the second test will also focus on the system's acceptability and potential benefits for clinical practice.

Acknowledgments

We would like to thank the physicians who participated in our study. We are also grateful to Anna McSherry for the language review of the paper. SATURN is funded by the German Federal Ministry of Health as part of the research focus "Digital Innovation," Module 3: "Smart Algorithms and Expert Systems" (grant ZMI1-2520DAT02B). Open access funding was enabled and organized by DEAL-Konsortium.

Authors' Contributions

SMK, SH, DS, and MCN collected, analyzed, and interpreted the data. JS, DS, and BSM contributed to the conceptualization and methodology of the study. MVW, JS, and BSM applied for funding. SMK, DS, and SH drafted the manuscript, which was critically revised by all authors. All authors read and approved the final manuscript.

Multimedia Appendix 1
Screenshots of SATURN (smart physician portal for patients with unclear disease) pages.
[PDF File, 976 KB - formative_v9i1e69875_app1.pdf]

Multimedia Appendix 2
Case vignette and tasks.
[PDF File, 631 KB - formative_v9i1e69875_app2.pdf]

Multimedia Appendix 3
Details of usability problems.
[PDF File, 1084 KB - formative_v9i1e69875_app3.pdf]

References

1. Singh H, Schiff GD, Graber ML, Onakpoya I, Thompson MJ. The global burden of diagnostic errors in primary care. BMJ Qual Saf 2017 Jun;26(6):484-494. [doi: 10.1136/bmjqs-2016-005401] [Medline: 27530239]
2. Müller T, Jerrentrup A, Schäfer JR. Computerunterstützte Diagnosefindung bei Seltenen Erkrankungen [Article in German]. Internist 2018 Apr;59(4):391-400. [doi: 10.1007/s00108-017-0218-z]
3. Blöß S, Klemann C, Rother AK, et al. Diagnostic needs for rare diseases and shared prediagnostic phenomena: results of a German-wide expert Delphi survey. PLoS ONE 2017;12(2):e0172532. [doi: 10.1371/journal.pone.0172532] [Medline: 28234950]
4. Limb L, Nutt S, Sen A. Experiences of rare diseases: an insight from patients and families. 2010 URL: https://www.raredisease.org.uk/our-work/coordination-of-care/experiences-of-rare-diseases-an-insight-from-patients-and-families-2010/ [accessed 2024-08-08]
5. Gahl WA, Wise AL, Ashley EA. The undiagnosed diseases network of the National Institutes of Health: a national extension. JAMA 2015 Nov 3;314(17):1797-1798. [doi: 10.1001/jama.2015.12249] [Medline: 26375289]
6. Harada T, Miyagami T, Kunitomo K, Shimizu T. Clinical decision support systems for diagnosis in primary care: a scoping review. Int J Environ Res Public Health 2021 Aug 10;18(16):8435. [doi: 10.3390/ijerph18168435] [Medline: 34444182]
7. Meunier PY, Raynaud C, Guimaraes E, Gueyffier F, Letrilliart L. Barriers and facilitators to the use of clinical decision support systems in primary care: a mixed-methods systematic review. Ann Fam Med 2023;21(1):57-69. [doi: 10.1370/afm.2908] [Medline: 36690490]
8. Regulation - 141/2000 - EN - EUR-Lex: Regulation (EC) No 141/2000 of the European Parliament and of the Council of 16 December 1999 on orphan medicinal products. 2024. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32000R0141 [accessed 2024-10-14]
9. Schaaf J, Sedlmayr M, Schaefer J, Storf H. Diagnosis of rare diseases: a scoping review of clinical decision support systems. Orphanet J Rare Dis 2020 Sep 24;15(1):263. [doi: 10.1186/s13023-020-01536-z] [Medline: 32972444]
10. CPUX-f curriculum and glossary, version 3.16 EN. UXQB: International Usability and User Experience Qualification Board. 2020 Nov. URL: https://uxqb.org/public/documents/CPUX-F_EN_Curriculum-and-Glossary_v3.16.pdf [accessed 2025-08-20]
11. UXQB: international usability and user experience qualification board. CPUX-UR Curriculum Certified Professional for Usability and User Experience – User Requirements Engineering Version 322 EN. 2023 Mar. URL: https://isqi.org/media/6d/dd/e8/1740753386/CPUX-UR_EN_Curriculum.pdf [accessed 2025-08-20]

XSL•FO
RenderX

12. ISO. Ergonomics of Human-System Interaction—Part 210: Human Centred Design for Interactive Systems, ISO 9241–210. 2019. URL: https://www.iso.org/standard/77520.html [accessed 2025-08-20]
13. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989 Sep;13(3):319. [doi: 10.2307/249008]
14. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view. MIS Q 2003;27(3):425. [doi: 10.2307/30036540]
15. Maramba I, Chatterjee A, Newman C. Methods of usability testing in the development of eHealth applications: a scoping review. Int J Med Inform 2019 Jun;126:95-104. [doi: 10.1016/j.ijmedinf.2019.03.018] [Medline: 31029270]
16. SATURN Projekt [Website in German]. 2024. URL: https://www.saturn-projekt.de [accessed 2024-12-04]
17. Schütze D, Holtz S, Neff MC, et al. Requirements analysis for an AI-based clinical decision support system for general practitioners: a user-centered design process. BMC Med Inform Decis Mak 2023 Jul 31;23(1):144. [doi: 10.1186/s12911-023-02245-w] [Medline: 37525175]
18. Neff MC, Schaaf J, Noll R, et al. Initial user-centred design of an AI-based clinical decision support system for primary care. Stud Health Technol Inform 2024 Jan 25;310:1051-1055. [doi: 10.3233/SHTI231125] [Medline: 38269975]
19. Neff MC, Schütze D, Holtz S, et al. Development and expert inspections of the user interface for a primary care decision support system. Int J Med Inform 2024 Dec;192:105651. [doi: 10.1016/j.ijmedinf.2024.105651] [Medline: 39413613]
20. UXQB: international usability and user experience qualification board. CPUX-UT Curriculum Certified Professional for Usability and User Experience – Usability Testing and Evaluation. 2020. URL: https://www.uxqb.org/ux-training-and-certification/cpux-ut/ [accessed 2025-08-20]
21. Konrad K. Lautes denken. In: Mey G, Mruck K, editors. Handbuch Qualitative Forschung in Der Psychologie [Book in German], 1st edition: Wiesbaden: VS Verlag für Sozialwissenschaften; 2010. URL: https://download.e-bookshelf.de/download/0000/0175/96/L-G-0000017596-0002372728.pdf [accessed 2025-08-20]
22. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
23. Palinkas LA, Horwitz SM, Green CA, Wisdom JP, Duan N, Hoagwood K. Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. Adm Policy Ment Health 2015 Sep;42(5):533-544. [doi: 10.1007/s10488-013-0528-y] [Medline: 24193818]
24. Sharma P. The unlikely suspect: a case report of new-onset hyperthyroidism due to graves' disease in an 89-year-old gentleman and review of literature. Cureus 2022;14:e21546. [doi: 10.7759/cureus.21546]
25. Krug S. Don't Make Me Think: A Common Sense Approach to Web (and Mobile) Usability, 3rd edition: New Riders; 2014.
26. Brooke J. A 'quick and dirty' usability scale. In: Jordan P, Thomas B, Weerdmeester B, McClelland IL, editors. Usability Evaluation in Industry: Taylor & Francis; 1996.
27. Bangor A, Kortum PT, Miller JT. An empirical evaluation of the system usability scale. Int J Hum Comput Interact 2008 Jul 29;24(6):574-594. [doi: 10.1080/10447310802205776]
28. System usability scale – jetzt auch auf deutsch. SAP Blogs. 2023. URL: https://blogs.sap.com/2016/02/01/system-usability-scale-jetzt-auch-auf-deutsch [accessed 2023-01-26]
29. Li AC, Kannry JL, Kushniruk A, et al. Integrating usability testing and think-aloud protocol analysis with "near-live" clinical simulations in evaluating clinical decision support. Int J Med Inform 2012 Nov;81(11):761-772. [doi: 10.1016/j.ijmedinf.2012.02.009] [Medline: 22456088]
30. Gorshid GD, Mayer C, Rausch A, Seifried J. Das LUCA-dashboard im usability-test – eine gaze-cued retrospective think-aloud-studie. In: Schumann S, Seeber S, Abele S, editors. Digitale Transformation in Der Berufsbildung: Konzepte, Befunde Und Herausforderungen [Book in German]: Bielefeld: wbv Publikation; 2022.
31. Nielsen J. Severity ratings for usability problems. 2022 Mar 2. URL: https://www.nngroup.com/articles/how-to-rate-the-severity-of-usability-problems [accessed 2024-08-08]
32. Sauro J. MeasuringU. 2023. URL: https://measuringu.com/rating-severity [accessed 2023-11-27]
33. Schreier M. Ways of doing qualitative content analysis: disentangling terms and terminologies. Forum: Qualitative Social Research 2014;15(1). [doi: 10.17169/fqs-15.1.2043]
34. Bangor A, Kortum P, Miller J. Determining what individual SUS scores mean: adding an adjective rating scale. J Usability Stud 2009;4:114-123 [FREE Full text]
35. Lewis J, Sauro J. Sample sizes for a SUS score. MeasuringU. 2023. URL: https://measuringu.com/sample-sizes-for-sus-ci [accessed 2023-11-27]
36. Ahmadi N, Zoch M, Kelbert P, et al. Methods used in the development of common data models for health data: scoping review. JMIR Med Inform 2023 Aug 3;11:e45116. [doi: 10.2196/45116] [Medline: 37535410]
37. Ahmadi N, Zoch M, Guengoeze O, et al. How to customize common data models for rare diseases: an OMOP-based implementation and lessons learned. Orphanet J Rare Dis 2024 Aug 14;19(1):298. [doi: 10.1186/s13023-024-03312-9] [Medline: 39143600]

XSL•FO
RenderX

38. Söhl K, Wolf F, Kuniß N, et al. Nutzung von Daten aus dem Praxisverwaltungssystem eines großen hausärztlichen Verbundes medizinischer Versorgungszentren am Beispiel der Diagnosestellung einer ambulant erworbenen Pneumonie – eine qualitative Machbarkeitsstudie [Article in German]. Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen 2024 Apr;185:54-63. [doi: 10.1016/j.zefq.2023.12.005]
39. Kassenärztliche Bundesvereinigung (KBV), Praxisverwaltungssysteme (PVS). Die 20 führenden ADT-abrechnungssysteme in den einzelnen fachgruppen: TOP 20 systeme - allgemeinmediziner [Article in German]. Stand. 2024 Mar 31. URL: https://www.kbv.de/html/6989.php [accessed 2024-05-02]
40. Holmgren AJ, Esdar M, Hüsers J, Coutinho-Almeida J. Health information exchange: understanding the policy landscape and future of data interoperability. Yearb Med Inform 2023 Aug;32(1):184-194. [doi: 10.1055/s-0043-1768719] [Medline: 37414031]
41. Baxter K, Courage C, Caine K. Understanding Your Users: A Practical Guide to User Research Methods: Morgan Kaufmann; 2015. URL: https://shop.elsevier.com/books/understanding-your-users/baxter/978-0-12-800232-2 [accessed 2025-08-19]

Abbreviations

AI: artificial intelligence
CDSS: clinical decision support system
COREQ: Consolidated Criteria for Reporting Qualitative Research
GFR : glomerular filtration rate
GP: general practitioner
SATURN: smart physician portal for patients with unclear disease
SUS: System Usability Scale
TA: think-aloud

Edited by J Sarvestan; submitted 16.12.24; peer-reviewed by B Afolabi, J Pantel, K Schuler; revised version received 10.06.25; accepted 17.06.25; published 20.08.25.

Please cite as:
Köhler SM, Holtz S, Neff MC, Schaaf J, von Wagner M, Müller BS, Schütze D
Evaluating the Prototype of a Clinical Decision Support System in Primary Care: Qualitative Study
JMIR Form Res 2025;9:e69875
URL: https://formative.jmir.org/2025/1/e69875
doi:10.2196/69875

© Susanne M Köhler, Svea Holtz, Michaela C Neff, Jannik Schaaf, Michael von Wagner, Beate S Müller, Dania Schütze. Originally published in JMIR Formative Research (https://formative.jmir.org), 20.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Organizational Climate and Decision Aid Sustainability in Lupus Care: Mixed Methods Study

Aizhan Karabukayeva[1*], PhD; Larry R Hearld[2*], PhD; Nathan W Carroll[3*], PhD; Reena Kelly[4*], PhD; Jasvinder A Singh[5*], MD

[1]Hudson College of Public Health, University of Oklahoma Health Sciences Center, 801 NE 13th St., Oklahoma City, OK, United States
[2]Department of Health Services Administration, School of Health Professions, University of Alabama at Birmingham, Birmingham, AL, United States
[3]Department of Health Administration, College of Health Professions, Virginia Commonwealth University, Richmond, VA, United States
[4]Department of Population Health and Leadership, School of Health Sciences, University of New Haven, West Haven, CT, United States
[5]Section of Immunology, Allergy and Rheumatology, Baylor College of Medicine, Houston, TX, United States
[*]all authors contributed equally

Corresponding Author:
Aizhan Karabukayeva, PhD
Hudson College of Public Health, University of Oklahoma Health Sciences Center, 801 NE 13th St., Oklahoma City, OK, United States

Abstract

Background: Digital decision aids (DAs) are increasingly used in health care to support shared decision-making and promote patient engagement. In the context of systemic lupus erythematosus (SLE), a complex autoimmune disease characterized by diverse symptoms and uncertain prognoses, DAs offer guidance to patients in navigating treatment risks and benefits. Although numerous studies have examined the initial implementation of evidence-based tools, there is limited evidence on the organizational factors that influence their long-term sustainability in clinical practice. This gap is particularly salient for digital interventions, where integration into routine workflows and ongoing use require alignment with clinic readiness and culture. This study focuses on an evidence-based, electronic DA designed to support patients with lupus and investigates how dimensions of organizational climate, particularly learning climate and change readiness, are associated with the tool's sustained use across diverse practice settings.

Objective: This study aims to examine the relationship among the learning climate, change readiness climate, and perceived permanence of a DA for patients with lupus in 15 geographically diverse rheumatology clinics in the United States.

Methods: This study was conducted as part of a broader multisite implementation project. We used a concurrent mixed methods design, integrating longitudinal quantitative survey data with qualitative interviews. Quantitative data were collected via web-based surveys at 3 time points (6-, 12-, and 24-mo postimplementation) from physicians, nurses, medical assistants, and administrative personnel (n=204 responses across rounds). The primary outcome was perceived DA permanence, measured with a validated 5-item scale. Independent variables included internal and external learning climates, change commitment, and change efficacy. Data were aggregated to the clinic level and analyzed using generalized estimating equations with clustered SEs. Qualitative data were collected through 36 semistructured interviews with clinic staff to explore contextual factors affecting sustainment.

Results: Quantitative findings revealed that change efficacy climate was significantly associated with greater perceived permanence of the DA (β=4.00; P<.001) while internal (β=−1.39; P<.05) and external learning climates (β=−2.11; P<.01) were negatively associated. Change in commitment was not statistically significant. Qualitative data highlighted challenges to sustainment, including poor workflow integration, lack of physician buy-in, and limited applicability of the DA to certain patient populations.

Conclusions: Sustaining digital health tools like DAs requires not only technical integration but also a supportive organizational climate. This study demonstrates that perceptions of a clinic's collective ability to sustain change (change efficacy) are critical while learning climates may expose barriers that hinder long-term use. These findings underscore the importance of assessing organizational readiness and tailoring implementation strategies to foster DA sustainment in real-world settings.

(JMIR Form Res 2025;9:e69603) doi:10.2196/69603

KEYWORDS

organizational climate; organizational readiness for change; organizational learning; lupus; decision-aid; sustainment

XSL•FO
RenderX

Introduction

As health care increasingly embraces digital innovation, electronic decision aids (DAs) have emerged as vital tools for enhancing shared decision-making, improving patient engagement, and promoting patient-centered care [1,2]. These tools deliver evidence-based information about treatment options in accessible formats, empowering patients to make informed decisions aligned with their values and preferences [3,4]. In chronic and complex conditions, such as systemic lupus erythematosus (SLE), digital DAs can be especially impactful by helping patients understand long-term treatment implications and participate more meaningfully in their care [5], potentially improving adherence to complex medication regimens, a known challenge in SLE management [6].

Lupus is a chronic autoimmune condition characterized by a relapsing-remitting course and variable clinical presentations, ranging from mild symptoms to severe, life-threatening complications [7]. Characterized by disease unpredictability, complex medication regimens, and significant quality-of-life impacts [8], lupus often requires nuanced treatment decisions—particularly for women of childbearing age or patients with organ involvement [9,10]. However, time constraints in outpatient settings [11], variable health literacy [12], and limited consultation windows can inhibit the level of discussion needed for truly informed choices. Digital DAs offer a scalable, patient-centered solution to these challenges by preparing patients in advance of clinical encounters, encouraging questions, and fostering dialogue during visits [13,14], providing information at the patient's own pace, and facilitating structured discussions.

Despite their promise, many digital health tools, including DAs, are not sustained in routine care beyond their initial implementation phase [15]. Research indicates that adoption alone is not sufficient. Digital interventions must be actively maintained and integrated into clinical workflows to generate lasting impact [16]. Challenges to sustain often stem from a lack of organizational readiness, weak leadership support, poor workflow fit, and minimal clinician engagement [17-19]. While implementation science has extensively examined adoption and fidelity, the long-term sustainability of DAs in real-world clinical environments remains an understudied area—especially as it relates to digital health innovation [20].

Organizational climate—the shared perceptions of policies, practices, and procedures—plays a central role in shaping implementation success and long-term sustainment [17,21]. Two specific climate dimensions may be especially relevant for sustaining DAs: learning climate and readiness for change. A positive internal learning climate, characterized by team-level reflection and feedback, and an external learning climate, characterized by openness to outside knowledge and innovation, are theorized to foster the integration of digital tools into daily workflows (Edmondson [22]). Therefore, we hypothesize that a more positive internal learning climate (Hypothesis 1a; H1a) and external learning climate (Hypothesis 1b; H1b) will each be associated with greater perceived permanence of the DA.

In parallel, organizational readiness for change, defined as the collective willingness (change commitment) and confidence (change efficacy) to implement and sustain new practice, is also expected to support long-term use of innovations (Weiner [17]; Shea et al [23] 2014). Clinics with stronger change commitment may be more motivated to continue using DA, while higher change efficacy may promote creative problem-solving and proactive integration. Thus, we hypothesize that a more supportive climate of change commitment (H2) and change efficacy (H3) will be positively associated with the perceived permanence of DA in routine care. To test these hypotheses, we examined data from a multisite implementation of a digital lupus DA across 15 geographically diverse rheumatology clinics in the United States. Our goal was to identify key organizational factors associated with the long-term integration and sustained use of the DA, with particular attention to how learning climate and readiness for change contribute to its long-term integration.

Methods

Study Design and Setting

A concurrent mixed methods approach was used for the study. Quantitative data identified patterns across large numbers of study participants and sites while qualitative data were used to examine the underlying reasons for these patterns by understanding clinic personnel's perspectives on the perceived permanence of the DA [24]. Each data strand was collected and analyzed independently, followed by integration through triangulation to develop a comprehensive picture of DA sustainment. The study was part of a larger evaluation of the effectiveness of different implementation strategies for helping 15 clinics throughout the United States implement and sustain the use of a culturally competent, electronic DA designed to educate patients with lupus about their treatment options and help them engage in more shared decision-making with their clinical care team. The detailed description of the DA for patients with lupus is described elsewhere [25].

Data Sources

Quantitative Survey Data

The study relied primarily on 2 quantitative data sources. The first data source was a baseline web-based survey of members of the lupus clinics, including physicians, nurses, clinic administrators, medical assistants, front desk personnel, and research study coordinators. This "organizational assessment" survey included organizational climate measures (organizational readiness for change and learning), as well as clinic-level characteristics (eg, clinic size and ownership). A sampling frame was constructed for each site by asking the clinic study coordinator at each of the 15 clinics to provide the names and email addresses of all clinic personnel who would be working with or affected by the DA of patients with lupus. An email was then sent to each person in the sampling frame informing them about the purpose of the study and the survey, an estimate of the time required to complete the survey (10 min), and a hyperlink to the survey. Reminder emails were sent to nonrespondents 3 and 7 business days after the initial email.

XSL•FO
RenderX

The second data source was a web-based "implementation" survey administered at 3 different times over the study period: Round 1 was administered 6 months after the DA was first implemented in the clinic; Round 2 was administered 1 year after the DA was implemented in the clinic, and Round 3 was administered 2 years after the DA was implemented in each clinic. The main objective of this survey was to assess the clinic personnel's perception of the permanence of the DA. The sampling frame for the survey consisted of all clinic members who either directly interacted with the DA (eg, study coordinators) or were affected by its use (eg, physicians, nurses, medical assistants, and clinic administrators).

Qualitative Interview Data

The third data source was semistructured interviews with the clinic personnel to provide more in-depth understanding of the opportunities for sustaining the DA. Data collection took place during May-July 2022 via a cloud-based video conferencing service (Zoom; Zoom Communications). The average duration of the interviews was 20 (SD 10; range 10 - 30) minutes. Either the study coordinator, the principal investigator, or both at each site clinic selected key informants from a variety of positions within a clinic to assure that a wide variety of perspectives on potential challenges to sustaining the DA was captured. The eligibility criteria included being an employee of the clinic and being familiar with the clinic's research and patient care activities. Three research team members (LH, AK, and RK) piloted the initial interview protocol with 3 key informants to assess the clarity of questions, identify gaps in the protocol, and improve the flow of the interview. Subsequent interviews were conducted by these same 3 research team members.

The interview protocol for clinic personnel included questions related to the overall perceptions and experiences with the DA and views regarding its long-term sustainment (Multimedia Appendix 1). Participants were asked about their general involvement with the project, probing specifically for positive and negative experiences with the DA that might impact the sustainment. In addition, participants were asked to provide more targeted opinions on organizational factors that affect continued use of the DA, including the type of support and resources needed to effectively sustain the DA. The interview protocol for clinic personnel was constructed based on the patient experience with DAs [1]. Figure 1 provides an overview of the timing and duration of each data collection activity across the study period, including the preimplementation assessment, implementation-phase surveys, and postimplementation interviews.

Figure 1. Implementing the Decision-Aid for Lupus study timeline: Data collection phase.

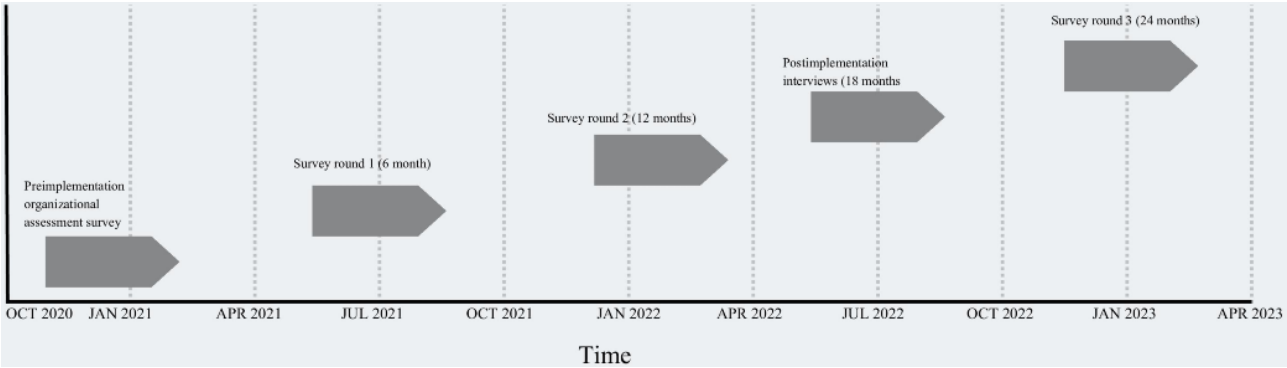

Measures

Outcome Variable

Perceived DA permanence of the DA, the primary outcome, was assessed via implementation survey at 6, 12, and 24 months postimplementation. Perceived DA permanence was measured using 5 validated items that were scored ranging from 1 ("Disagree") to 5 ("Agree") [26]. The 5 items were averaged to construct a composite score reflecting perceived DA permanence, where higher scores indicate perceptions of greater permanence of the DA in the clinic (ie, better outcome).

Predictor Variables

The primary predictor variables examined in this study were 4 summated scales based on multiple Likert-type items from the web-based survey (Multimedia Appendix 2).

There is psychometric evidence in support of a measure of organizational readiness for change, called Organizational Readiness for Implementing Change (ORIC) [23]. It is a brief, reliable, and valid measure that differentiates between change commitment and change efficacy. Commitment was operationalized as the average of 5 survey items, all of which were measured on a 5-point scale ranging from 1=Disagree to 5=Agree. Similarly, change efficacy was operationalized as the average of 7 survey items measured on the same 5-point scale.

The organizational learning variable was measured using the Team Learning Survey developed by Edmondson [27]. Organizational learning was divided into 2 dimensions, such as internal and external learning. Internal learning climate was an average of 8 survey items, and external learning climate was an average of 4 survey items, all items measured on a 7-point scale (1=Strongly disagree and 7=Strongly agree).

Covariates

Additional clinic-level variables included 5 clinic characteristics that reflected structural, cultural, and general staffing attributes of the clinic. Clinic specialty was a dichotomous variable that indicated whether a clinic included only a single medical specialty (ie, rheumatology; coded 0) or multiple medical specialties (coded 1). Ownership was also a dichotomous variable that indicated whether a clinic was owned by a university/part of an academic medical center (coded 0) or not

XSL•FO
RenderX

owned by a university (coded 1). Clinic size was operationalized as a series of dummy variables reflecting 3 categories: 1 - 10 members (referent); 11 - 30 members; and more than 30 members. The analysis included a count variable to account for the number of physical clinic locations for each participating site. Individual years of experience were averaged at the clinic level to create a continuous variable that reflected the work experience of clinic members. Finally, clinic culture was based on responses to 4 survey items that asked respondents to allocate percentage points (0% - 100%) to four different culture types [28,29]: (1) Clan/Team culture (friendly workplace where leaders act like mentors/facilitators and there is an emphasis on doing things together); (2) Rational/Hierarchical culture (structured and formalized workplace where leaders act like coordinators and value is placed on doing things right); (3) Adhocracy/Entrepreneurial culture (dynamic workplace with leaders that stimulate new ideas and value is placed on doing things first); and (4) Market culture (competitive workplace with leaders that are like producers and value placed on doing things fast). Individual responses were aggregated to the clinic level by averaging across clinic members. A clinic was then assigned a dominant culture type (eg, clan/team culture) if, on average, respondents assigned greater than 50% points to that culture type. For clinics where responses, on average, did not reveal a dominant culture type (ie, were more evenly distributed across the 4 culture types), they were labeled as having no dominant culture (referent group).

Quantitative Analytic Strategy

Organizational climate measures were aggregated to the clinic level. To justify aggregating individual responses to the clinic level, we calculated within-unit agreement and sufficient between-unit differences using 3 multi-item interrater agreement statistics: (1) Intraclass correlation coefficient; (2) Within group reliability (rwg(j)); and (3) Average Deviation (Adm(j)). Together, these analyses provided evidence that sufficient agreement existed among members' perception of climate measurements to support aggregation to the clinic level (Table 1).

Table . Characteristics of survey respondents by implementation round of the DA permanence survey.

Respondent characteristics	Round 1 (N=77)	Round 2 (N=60)	Round 3 (N=70)
Response rate (%)	56.8	46.5	54.8
Age, n (%)			
19-24	9 (12)	4 (6)	5 (7)
25-44	38 (49)	31 (50)	32 (47)
45-64	26 (33)	24 (39)	28 (42)
More than 65	4 (5)	3 (4)	2 (2)
Sex, n (%)			
Male	19 (25)	17 (28)	25 (38)
Female	58 (75)	44 (72)	41 (62)
Race, n (%)			
White	43 (56)	36 (60)	39 (61)
Black or African American	13 (17)	4 (7)	7 (11)
Asian	21 (27)	20 (33)	18 (28)
Ethnicity, n (%)			
Hispanic or Latino	5 (7)	6 (10)	6 (9)
Not Hispanic or Latino	70 (93)	54 (90)	58 (91)
Education, n (%)			
High school graduate	1 (1)	1 (2)	2 (3)
Undergraduate or less	35 (46)	22 (35)	23 (34)
Professional degree	17 (22)	16 (26)	18 (27)
Doctorate	24 (31)	23 (37)	24 (36)
Role in the clinic, n (%)			
Physician	28 (37)	29 (48)	26 (39)
Other clinical personnel	22 (29)	11 (18)	20 (30)
Administrative personnel	26 (34)	21 (34)	21 (31)

XSL•FO
RenderX

The multivariate analysis used generalized estimating equations and clustered SEs at the clinic level to account for potential correlations of individual respondents within the same clinic. All predictor variables were mean-centered to facilitate model interpretation [30]. All analyses were performed using Stata version 17 (StataCorp LLC) [31] and all relationships were considered statistically significant at $P<.05$.

Qualitative Analytic Strategy

We conducted a thematic analysis guided by Braun and Clarke's [32] 6-phase approach, which emphasizes reflexive and transparent practices in qualitative research. Each semistructured interview was recorded and transcribed verbatim. We used a hybrid inductive-deductive coding strategy where initial codes were drawn from the interview protocol (deductive) while additional themes were identified from the data (inductive) through iterative reading and comparison with literature on organizational change and sustainment. To ensure credibility and confirmability, 2 researchers (LH and AK) independently coded an initial subset of transcripts (n=3 - 4), achieving strong interrater reliability (k>0.81). Discrepancies were discussed and resolved through consensus, after which the remaining transcripts were coded by AK and RK. NVivo 12 Plus (QSR International) software was used to facilitate coding and organize themes. Reflexivity was maintained through memo writing and analytic journaling at the clinic level, supporting dependability and transferability by documenting analytic decisions and contextual factors. Themes were developed through familiarization, coding, theme generation, review, and refinement, with representative quotes used to support interpretations.

Triangulation of Qualitative and Quantitative Data Sources

Quantitative and qualitative data were integrated using a convergent mixed methods design, with findings triangulated to provide a comprehensive understanding of sustainment. Two strands were linked through a joint display table that aligned survey results and interview insights by thematic categories derived from the research questions. This allowed for direct comparison and contextualization of quantitative findings, for example, scores on organizational climate domains, with qualitative narratives that explained the underlying processes. Interviews added depth by uncovering contextual factors and "behind-the-scenes" mechanisms influencing continued use of the DA. Themes emerging from this integration informed the interpretation and organization of study implications.

Ethical Considerations

The study was approved by the Institutional Review Board (IRB) at the University of Alabama at Birmingham (UAB, 300002272; UAB Coordinating Center 200002554). Each study site's IRB (15 sites including UAB) also reviewed and approved the study. The IRB approved the use of verbal informed consent given that the interviews were conducted via Zoom. During the consent process, clinical personnel were informed that participation was voluntary, and they could withdraw from the study at any time. All investigations were conducted in conformity with ethical principles of research. All study procedures adhered to the ethical principles outlined in the Belmont Report and complied with the Health Insurance Portability and Accountability Act (HIPAA). All study data were deidentified and stored securely on IRB-approved services, accessible only to authorized personnel. Each study participant was paid US $50 for completing the baseline survey assessment and US $10 for follow-up assessments.

Results

Quantitative Results

Respondent Characteristics

The overall response rate for the DA permanence survey was as follows: 56.8% for the Round 1 survey; 46.5% for the Round 2 survey; and 48.5% for the Round 3 survey across the 15 clinics. The demographic characteristics of the respondents are described in Tables 2 and 3. In Round 1, 77 respondents completed the survey, and 58 (75%) respondents were female; 43 (56%) respondents were White, and 38 (49%) respondents were in the age group of 25 - 44 years. Most respondents had an undergraduate degree (35, 45%) and around 28 (37%) respondents were physicians while 26 (34%) respondents were administrative personnel. In Round 2, there were 60 respondents, and in Round 3, there were 70 respondents. The demographic profile of the later rounds was similar to the respondents of Round 1.

XSL•FO
RenderX

Table . Baseline organizational climate characteristics from the organizational assessment survey (N=135 respondents across 15 clinics).

Variables	Univariate statistics, mean (SD)	Aggregation statistics		
		ICC[a]	Rwg(j)[b]	Adm(j)[c]
Implementation climate				
Organizational Readiness for Implementing Change				
Commitment	3.97 (0.25)	0.093	0.84	0.71
Efficacy	3.93 (0.30)	0.044	0.87	0.71
Learning environment				
Internal learning environment	4.86 (0.52)	0.173	0.89	1.01
External learning environment	4.76 (0.63)	0.023	0.74	1.11
Psychological safety	3.75 (0.52)	0.27	0.83	1.02

[a]Intraclass correlation coefficient; higher values indicate that more variability in responses can be attributed to clinic membership, and thus, more support for aggregation.

[b]Within-group reliability; higher values indicate greater agreement and more support for aggregation.

[c]Average deviation index; higher values indicate more dispersion and less support for aggregation

Table . Baseline clinic level characteristics from the organizational assessment survey (N=135 respondents across 15 clinics).

Clinic Characteristics	Univariate Statistics
Specialty, n (%)	
Rheumatology only	6 (40)
Multispecialty	9 (60)
Ownership, n (%)	
University-owned	12 (80)
Not-university-owned	3 (20)
Size, n (%)	
0-10 clinic members	2 (13)
11-30 clinic members	6 (40)
31 or more clinic members	7 (47)
Number of locations (mean, SD)	2.27 (0.25)
Average clinic years of experience (mean, SD)	2.28 (0.25)
Culture, n (%)	
No dominant culture	11 (73)
Team culture	3 (20)
Rational/Hierarchical culture	1 (7)

Clinic Characteristics

The overall response rate for the baseline organizational assessment survey was 93% (n=135), ranging from 30% to 100% across the 15 clinics. Of the study clinics, 6 (40%) were rheumatology-only while the other 9 (60%) were multispecialty (Table 1); 12 (80%) of the study clinics were owned by a university while only 3 (20%) were not owned by a university. Overall, 2 (13%) study clinics had less than 10 members, while nearly half (n=7, 47%) had 31 or more clinic members. Notably, 11 (73%) study clinics had no dominant culture, while 3 (21%) study clinics had a team culture. The average number of years of experience in the clinic was 2.28 (SD 0.25).

On average, clinics in the study reported a relatively high level of organizational readiness to implement the DA, with a mean level of change commitment of 3.97 (SD 0.25) and a mean level of change efficacy of 3.93 (SD 0.30) across the 15 clinics. Similarly, clinics demonstrated high levels of learning climate, with a mean level of internal learning of 4.86 (SD 0.52) and a mean level of external learning of 4.76 (SD 0.63). Clinic personnel's perception of the permanence of the DA was

XSL•FO
RenderX

moderate in the first survey round (6 mo) with a mean of 3.22 (SD 0.74) across all 15 clinics. Univariate analysis also found that perceptions of permanence slightly decreased from the first survey period to the second and third periods (Table 4).

Table . Descriptive statistics for perceived permanence of the DA at 6-, 12-, and 24-month postimplementation.

Site	Round 1		Round 2		Round 3	
	Total (N)	Mean (SD)	Total (N)	Mean (SD)	Total (N)	Mean (SD)
1	9	3.55 (0.87)	5	3.40 (0.84)	7	3.20 (1.23)
2	7	3.62 (0.49)	4	3.64 (0.71)	6	3.13 (0.24)
3	3	2.33 (0.70)	7	3.24 (1.43)	4	2.65 (0.84)
4	3	2.93 (1.10)	3	2.26 (0.46)	4	1.65 (0.90)
5	7	3.62 (0.73)	3	2.98 (0.52)	5	2.84 (0.59)
6	3	2.86 (0.80)	4	2.90 (1.27)	3	2.20 (1.22)
7	6	3.36 (0.66)	8	3.70 (0.75)	7	3.46 (0.79)
8	4	3.49 (0.34)	6	2.76 (1.16)	2	3.59 (0.84)
9	6	2.70 (0.88)	3	2.73 (1.02)	1	1.60 (0.00)
10	5	3.12 (0.52)	2	3.30 (0.42)	1	1.40 (0.00)
11	11	3.30 90.67)	5	3.40 (0.89)	16	3.84 (0.89)
12	4	2.25 (0.44)	1	3.40 (0.00)	2	2.30 (1.27)
13	1	3.59 (0.00)	2	2.50 (0.70)	1	2.80 (0)
14	5	3.20 (0.20)	4	3.10 (0.20)	8	3.05 (0.65)
15	3	3.46 (0.90)	3	2.93 (0.80)	3	3.87 (0.83)
Total	77	3.22 (0.74)	60	3.13 (0.92)	70	3.12 (0.83)
Cronbach α	0.81		0.90		0.93	
ANOVA	F (14,62)=1.86; P=.04		F (14,45)=0.73; P=.72		F (14,55)=3.04; P=.001	

Associations Between Organizational Climate Measures and Perceived Permanence of the DA

The multivariate results are reported in Table 5. The internal learning climate was negatively associated with perceived permanence of the DA (β=−1.39; P<.05), as was external learning climate (β=−2.11; P<.01). Thus, we did not find support for hypotheses 1a and 1b.

XSL•FO
RenderX

Table . Generalized estimating equation regression results examining the association between organizational climate measures and perceived permanence of the DA, Adjusted for Clinic-level Clustering.

Variables	β (SE)
Independent variables	
Organizational Readiness for Implementing Change	
Commitment	−0.35 (0.73)
Efficacy	4.00 (1.02)[a]
Learning environment	
Internal learning environment	−1.39 (0.56)[b]
External learning environment	−2.11 (0.61)[c]
Psychological safety	2.68 (0.95)[c]
Clinic Characteristic	
Specialty	
Rheumatology only	Reference
Multispecialty	0.38 (0.36)
Ownership	
University-owned	Reference
Not-university-owned	1.04 (0.35)[c]
Size (number of personnel)	
1 - 10 clinic members	Reference
11 - 30 clinic members	−0.13 (0.30)
31 or more clinic members	0.88 (0.30)[c]
Culture, n, (%)	
No dominant culture	Reference
Team culture	0.73 (0.25)[c]
Respondent covariates	
Education	
High school	Reference
Undergraduate or less	0.04 (0.37)
Professional (eg, MSW and MBA)	−0.31 (0.25)
Doctorate (eg, MD and PharmD)	0.18 (0.26)
Average age (years)	−0.49 (0.53)
Average clinic years of experience	−0.28 (0.07)[a]
Time	
Baseline	Reference
Month	0.05 (0.69)
12-month	−0.19 (0.14)
Total (N)	204

[a]$P<.001$
[b]$P<.05$
[c]$P<.01$

The relationship between the change commitment climate and perceived DA permanence was negative and not statistically significant. However, findings indicated that the change efficacy climate, on average, was positively associated with perceived

XSL•FO
RenderX

permanence of the DA (β=4.00; *P*<.001), supporting hypothesis 3.

Clinic-Level Covariates Included in the Multivariate Models

The strongest and most robust results for control variables were observed for the years of clinic experience, ownership, size, and culture (Table 5). Specifically, the number of years of experience was negatively associated with the perceived permanence of the DA (β=−0.23; *P*<.01). This result suggests that those who have more clinical experience are less likely to continue using the DA. Respondents from nonuniversity clinics reported more positive perceptions of DA permanence than respondents from university-owned clinics (β=1.06; *P*<.01). Finally, relative to members of clinics with no dominant culture, members of clinics with a team culture reported more positive perceptions of DA permanence (β=0.65; *P*<.01).

Qualitative Results

Overview

In total, 36 interviews were conducted across all 15 clinics, with an average of 3.6 (SD 2.6; range 1-8) per clinic. The results below are organized around the major thematic areas of clinic-focused themes and intervention-focused themes.

Clinic-Focused Themes

Perceived Usefulness of the DA and Ease of Use

One of the key factors likely to support continued use of the DA was a positive attitude toward its usefulness. In general, there was consistent support for the DA from the interviews with clinic members from all sites. The DA was seen as a valuable tool that could simplify the office visit, as the patients potentially could come better informed. Appealing graphics, a user-friendly interface, and straightforward language were highly appreciated. Furthermore, several clinic members provided examples of patients who viewed the DA and became more involved in the decision-making process, asked more questions, or even changed their medication decision, as highlighted by one physician, “It’s a nice way to reinforce some of the most important aspects of how to take care of this disease long term instead of just thinking from visit to visit” (Site 11).

Buy-In from Physicians

Despite the perceived usefulness of the DA and alignment with the values of how treatment decisions should be made, a majority of physicians were less inclined to continue using it. One of the physicians noted that he was already adequately educating patients and involving them in decision-making; therefore, he did not have intentions to continue using the DA in the future. This is significant because, according to other nonphysician clinic members, the routine use of the DA would mainly depend on physicians’ interest and judgment. This was also important because patients seemed more open to using the DA when their physician introduced it to them, particularly if they had a good relationship and trusted their physician. Clinic members also noted that if the use of the DA is not promoted, patients may not take the initiative to click on the link and view the DA outside of the office visit. Therefore, the use of the DA in the clinic needed to be endorsed by the physicians if it was to be sustained.

Differential Utility Depending on the Disease Stage

Another consistent pattern was that, although the DA was regarded by clinic staff as a valuable tool for learning about lupus in general, many felt it would be particularly beneficial for patients who are newly diagnosed. According to several clinic members, patients who had been recently diagnosed with lupus reportedly expressed that they had learned something new from the DA. Given that clinic visits are typically brief, with an average of 15 (SD 5) minutes for return patients and 30 (SD 10) minutes for new patients, physicians feel compelled to be as efficient as possible when conveying information to patients. Therefore, the DA could serve as a valuable source of additional information or a reminder of topics that may have been missed or forgotten during appointments. However, several clinic members noted that, during the study period, the DA was primarily viewed by patients who have had lupus for a long time and did not require medication changes, making it difficult to gage its value. Moreover, because the participating clinics were outpatient, clinic staff believed most patients were not making an active treatment decision. As a result, clinic members observed that many patients, particularly those with longstanding lupus and stable treatment plans, did not perceive the DA as significantly enhancing their knowledge of lupus. However, clinic members suggested that the DA may be more beneficial for patients experiencing disease flares, those undergoing medication changes, or individuals receiving care in inpatient settings where active decision-making is more common.

Practice Pattern Differences

Clinic members from 5 sites noted that the medication choices presented in the DA did not reflect those used by their clinics for treating lupus. It was recommended to include more scenarios in the DA, as the current version covered either more intense immunosuppressive drugs or those not prescribed by the clinics. Given that the DA often was not applicable to their clinical situations, providers who used different medications inconsistent with the treatment options presented in the DA were unlikely to use it in the future.

Limited Integration into Routine Clinical Workflow

Almost all clinic members from all 15 sites noted that the DA was seen exclusively as a research tool, with the research coordinator responsible for recruiting and obtaining patient consent, as well as distributing the DA to patients. However, participating in the study imposed certain limitations, as clinics were required to adhere to institutional rules and the DA was not fully integrated into their daily workflow. Most of the time was devoted to research-related tasks, such as identifying eligible patients, obtaining consent, and ensuring that surveys were completed. As a result, the DA was viewed as an “add-on” rather than a clinical tool, making it challenging to garner buy-in from all members of the clinic. Some clinic members noted that if it had been introduced as a clinical tool, it could have been reinforced during weekly physician meetings and better integrated into the clinical workflow.

XSL•FO
RenderX

The DA being viewed only as a research tool had the consequence of limiting involvement from clinic personnel. In many clinics, nurses, medical assistants, and front desk staff were not heavily involved in the study. While they were aware of the study taking place, they did not play a substantial role in distributing the DA. This approach had both advantages and disadvantages. One advantage was that the research coordinator remained the same throughout the study in some clinics, which helped maintain consistency when clinic staff had to leave for various reasons. However, it also limited the potential sustainability of the DA as it was not perceived as a clinical tool that would continue to be used after the study was over.

Burden of Prescreening/Identifying Patients

Several clinics did not have a dedicated clinic for lupus, so a study coordinator would screen patients beforehand and identify those with lupus based on their *ICD* (*International Classification of Diseases*) code. The coordinator would then seek approval from the physician to include a patient in the study and hand the DA. In other clinics, the physician or principal investigator would identify eligible patients. Regardless of the approach, all clinic members acknowledged the need to identify patients who may be experiencing a flare-up or have recently been diagnosed with lupus in a proactive manner. Ideally, during an office visit, a physician would assess a patient's eligibility for the DA and inform them that a link would be sent. In addition, the physician should schedule a follow-up visit to answer any questions. However, participants noted that manually prescreening and identifying which patients should view the DA would be a barrier to long-term sustainability.

Intervention-Focused Themes

Customizing the Content and Ensuring Timeliness of the DA

During interviews, clinic members noted that the DA appeared to be more focused on kidney disease and women of childbearing potential. Therefore, it would be important to update and tailor the content to better meet the needs of a diverse group of patients. One frequently cited request was to update the content with new discoveries about medication side effects to increase the likelihood of continued use beyond the study period. Moreover, it would be important to have clearly defined roles/responsibilities so that the DA can be routinely maintained and updated to reflect the latest evidence.

Need for Follow-Up if Viewed Outside of Clinic Visit

Although many clinic members preferred the DA to be used independently outside of the consultation, preferably at home, it was hard to guarantee true shared decision-making with this format. The physicians noted that the DA was most effective when it was well-communicated with patients, that is, ensuring that patients are made aware of the availability of the DA, checking patients' understanding of the information provided, and tailoring the discussion to the patient's individual needs and preferences. Therefore, even if patients take the time to review the DA at home, it is necessary to have a follow-up call or visit to discuss any issues to really promote shared decision-making. This meaningful discussion would help increase the likelihood of continued use of the DA.

Automating/previsit Package

Participants noted that it would be beneficial if the scheduling system could identify patients with lupus and create a patient list to sustain the usage of the DA. Otherwise, turnover in staff makes it difficult to continue using it. Integrating the DA into the patient portal with an alarm for physicians when prescribing a new medication would also be helpful. Automating the process would allow patients to access the DA as a previsit package. Another idea was for physicians to give patients a link to view the DA independently. At the end of the visit, physicians could include the link as part of their instructions for patients to access at home.

Discussion

Principal Findings

This study examined the relationship between learning climate, change readiness climate, and the perceived permanence of a lupus DA for patients in 15 geographically diverse rheumatology clinics across the United States. We found that a positive change efficacy climate was significantly associated with the greater perceived permanence of the lupus DA, while change commitment was not significantly associated. Interestingly, both internal and external learning climates were negatively associated with DA permanence. Qualitative findings highlighted challenges to sustainment, including poor workflow integration, lack of physician buy-in, and perceptions of limited applicability of the DA to certain patient populations. Our findings underscore the importance of organizational context for the successful implementation and sustainment of evidence-based interventions, echoing prior theoretical and empirical work [33,34]. Respondents from clinics with a more positive change efficacy climate were more likely to have positive perceptions of DA permanence. This aligns with organizational readiness theory, which posits that efficacy, the belief in collective capability, is a critical driver of change implementation [17].

In contrast, change commitment was not significantly associated with DA permanence. This may be due to the setting of many participating clinics in academic environments with ongoing research and quality improvement projects. Interviews revealed that staff often viewed the DA as a temporary research tool, not a sustained clinical practice. This perception likely undermined broader organizational commitment, despite early engagement. However, the DA being viewed only as a research tool had the consequence of limiting clinic-wide commitment. Most of the clinics had only a handful of isolated research teams designated to implement the study. Thus, the DA did not have a pervasive presence and was rarely articulated as an ongoing explicit priority.

Unexpectedly, more favorable internal and external learning climates were negatively associated with perceived permanence. While this may seem counterintuitive, our qualitative data suggest that supportive learning environments facilitated critical reflection, leading clinic staff to identify limitations of the DA in their specific settings. For example, some physicians realized during implementation that the DA was less applicable to patients with stable disease or those not facing medication

XSL•FO
RenderX

changes. Others noted the absence of a reliable method to identify eligible patients, further complicating routine integration. This suggests that learning environments, while essential for innovation adoption, may also lead to early discontinuation if an intervention is perceived as misaligned with clinical needs.

Among the organizational factors, years of experience, culture, and ownership were associated with the perceived permanence of the DA. Interestingly, the results indicated that more experienced members were less optimistic about the likelihood of continuing to use the DA, which is in line with the literature on organizational tenure and its weak positive effect on employee innovative behavior [35]. The interviews shed light on why this might be the case. In most of the clinics, physicians have been with the clinic for 5 or more years, which meant strong institutional knowledge and possibly even a belief they were already adequately educating patients. In addition, as earlier stated, most of the physicians thought that the DA was not quite compatible with their current medication management; thus, they were less likely to continue using the DA. Experiential knowledge of what medications work best for their patients could have substantial influence on their prescription decisions and lead to variations seen in clinical practice [36]. Furthermore, high turnover among support staff meant that physicians were frequently the longest-tenured personnel in these clinics, making them critical for successful sustainment of the DA.

Practice Implications

The financial and human resources required to develop evidence-based interventions (EBIs) are substantial, making it crucial for practitioners and researchers to learn how to sustain them. This study, based on the implementation of the DA in 15 clinics, highlights the importance of organizational context in sustaining change. The focus should be on enhancing organizational climate, such as readiness for change and learning, before implementation. This study showed that change efficacy had a positive association with perceived DA permanence, while change commitment was not significant. Therefore, early planning that includes all clinic members and discussing the benefits of the EBI can improve buy-in. Likewise, leadership, particularly from physicians, plays a critical role in sustaining innovations in health care settings. A deliberate strategy should be in place to strengthen physician buy-in. Furthermore, if a continued use of the DA requires changes in organizational logistics, organizational management should focus on maintaining flexibility and compatibility of the DA across diverse settings. This includes tailoring the DA's format to meet the needs of patients and physicians, such as offering options to view it online, through an app, or in the clinic. Finally, a formal strategy for sustaining EBIs should include ongoing evaluation to identify improvement opportunities. Project tracking, reporting, and staff discussions can help address issues early in the implementation process. Regular dissemination of data on patient satisfaction and outcomes, as well as providing incentives for active DA use, can further strengthen the innovation's presence and sustainability within the organization.

Limitations

Several limitations should be considered when interpreting this study's findings. The analysis was based on just 15 clinics, mostly within academic medical centers, limiting the power to detect statistically significant relationships and generalizability. The outcome variable, perceived permanence, may have been biased as data collection occurred during implementation, possibly pressuring participants to respond positively. In addition, limited longitudinal analysis was possible due to staffing turnover, though the qualitative data helped clarify underlying mechanisms behind climate factors from multiple stakeholders' perspectives.

Conclusions

Our findings suggest that simply introducing a lupus DA does not guarantee its sustainment in clinical practice. Organizational climate can play a central role in shaping perceptions of an intervention's permanence. Future research should use longitudinal and multilevel designs to examine how organizational learning, readiness for change, and psychological safety influence the long-term adoption of decision aids. Tracking shifts in perception across time and context can help identify which organizational strategies best support sustained use.

In addition, clearer communication about the DA's relevance to different disease stages and care pathways is needed. Various versions of the DA were implemented during the COVID-19 pandemic, and confusion about their applicability likely reduced uptake. Future dissemination efforts should ensure that stakeholders understand how different versions of the DA align with specific patient populations. Doing so may improve fit, reduce resistance, and promote sustained integration into routine lupus care.

Acknowledgments

Research reported in this article was funded through a Patient-Centered Outcomes Research Institute (PCORI) Award (SDM-2017C2–8224). Dr. Singh is supported by the resources and use of facilities at the Birmingham VA Medical Center, Birmingham, Alabama, United States. The funders did not play any role in the design, collection, analysis, and interpretation of the data, in the writing of the manuscript, or in the decision to submit the manuscript for publication. The statements in this article are solely the responsibility of the authors and do not necessarily represent the views of the Patient-Centered Outcomes Research Institute (PCORI), its Board of Governors, or Methodology Committee.

XSL•FO
RenderX

Data Availability

The datasets generated or analyzed during the current study are not publicly available due to concerns regarding participant privacy, but may be made available by the principal investigator upon reasonable request by contacting Jasvinder Singh via email (jasvinder.md@gmail.com).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview Guide for Clinic Personnel.
[DOCX File, 23 KB - formative_v9i1e69603_app1.docx]

Multimedia Appendix 2
Survey instruments.
[DOCX File, 279 KB - formative_v9i1e69603_app2.docx]

References

1. Joseph-Williams N, Abhyankar P, Boland L, et al. What works in implementing patient decision aids in routine clinical settings? A rapid realist review and update from the International Patient Decision Aid Standards Collaboration. Med Decis Making 2021 Oct;41(7):907-937. [doi: 10.1177/0272989X20978208] [Medline: 33319621]
2. Staszewska A, Zaki P, Lee J. Computerized decision aids for shared decision making in serious illness: systematic review. JMIR Med Inform 2017 Oct 6;5(4):e36. [doi: 10.2196/medinform.6405] [Medline: 28986341]
3. O'Connor AM, Stacey D, Rovner D, et al. Decision aids for people facing health treatment or screening decisions. Cochrane Database Syst Rev 2001;2017(3):CD001431. [doi: 10.1002/14651858.CD001431] [Medline: 11686990]
4. Elwyn G, et al. Decision aids and beyond: the next decade of decision support technologies. In: Shared Decision-Making in Health Care Achieving Evidence-Based Patient Choice: Oxford University Press; 2009:389-394. [doi: 10.1093/oso/9780199546275.003.0054]
5. Wu CJ, Yeh TP, Chu G, Ho YF. Effectiveness of computerised decision aids for patients with chronic diseases in shared decision-making: a systematic review and meta-analysis. J Clin Nurs 2024 Jul;33(7):2732-2754. [doi: 10.1111/jocn.17095] [Medline: 38553843]
6. Petri M, Orbai AM, Alarcón GS, et al. Derivation and validation of the Systemic Lupus International Collaborating Clinics classification criteria for systemic lupus erythematosus. Arthritis Rheum 2012 Aug;64(8):2677-2686. [doi: 10.1002/art.34473] [Medline: 22553077]
7. Basta F, Fasola F, Triantafyllias K, Schwarting A. Systemic lupus erythematosus (SLE) therapy: the old and the new. Rheumatol Ther 2020 Sep;7(3):433-446. [doi: 10.1007/s40744-020-00212-9] [Medline: 32488652]
8. Olesińska M, Saletra A. Quality of life in systemic lupus erythematosus and its measurement. Reumatologia 2018;56(1):45-54. [doi: 10.5114/reum.2018.74750] [Medline: 29686443]
9. Dar S, Koirala S, Khan A, et al. A comprehensive literature review on managing systemic lupus erythematosus: addressing cardiovascular disease risk in females and its autoimmune disease associations. Cureus 2023 Aug;15(8):e43725. [doi: 10.7759/cureus.43725] [Medline: 37727166]
10. Wivel AE, Lapane K, Kleoudis C, Singer BH, Horwitz RI. Medicine based evidence for individualized decision making: case study of systemic lupus erythematosus. Am J Med 2017 Nov;130(11):1290-1297. [doi: 10.1016/j.amjmed.2017.06.013] [Medline: 28711556]
11. Khan R, Imami SK, Anwer Khan SE, Batool S, Naeem F, Zaffar MA. It's about time: a study of rheumatology patient consultation times. Cureus 2023 Oct;15(10):e48007. [doi: 10.7759/cureus.48007] [Medline: 38034181]
12. Katz P, Dall'Era M, Trupin L, et al. Impact of limited health literacy on patient‐reported outcomes in systemic lupus erythematosus. Arthritis Care Res (Hoboken) 2021 Jan;73(1):110-119. [doi: 10.1002/acr.24361] [Medline: 32741118]
13. Singh JA, Hearld LR, Hall AG, Beasley TM. Implementing the DEcision-Aid for Lupus (IDEAL): study protocol of a multi-site implementation trial with observational, case study design: implementing the DEcision-Aid for Lupus. Implement Sci Commun 2021 Mar 11;2(1):30. [doi: 10.1186/s43058-021-00118-9] [Medline: 33706813]
14. Joseph-Williams N, Elwyn G, Edwards A. Knowledge is not power for patients: a systematic review and thematic synthesis of patient-reported barriers and facilitators to shared decision making. Patient Educ Couns 2014 Mar;94(3):291-309. [doi: 10.1016/j.pec.2013.10.031] [Medline: 24305642]
15. Kaboré SS, Ngangue P, Soubeiga D, et al. Barriers and facilitators for the sustainability of digital health interventions in low and middle-income countries: a systematic review. Front Digit Health 2022;4:1014375. [doi: 10.3389/fdgth.2022.1014375] [Medline: 36518563]

XSL•FO
RenderX

16. Proctor E, Luke D, Calhoun A, et al. Sustainability of evidence-based healthcare: research agenda, methodological advances, and infrastructure support. Implement Sci 2015 Jun 11;10:88. [doi: 10.1186/s13012-015-0274-5] [Medline: 26062907]
17. Weiner BJ. A theory of organizational readiness for change. Implement Sci 2009 Oct 19;4(1):67. [doi: 10.1186/1748-5908-4-67] [Medline: 19840381]
18. Tsiknakis M, Kouroubali A. Organizational factors affecting successful adoption of innovative eHealth services: a case study employing the FITT framework. Int J Med Inform 2009 Jan;78(1):39-52. [doi: 10.1016/j.ijmedinf.2008.07.001] [Medline: 18723389]
19. Martin GP, Weaver S, Currie G, Finn R, McDonald R. Innovation sustainability in challenging health-care contexts: embedding clinically led change in routine practice. Health Serv Manage Res 2012 Nov;25(4):190-199. [doi: 10.1177/0951484812474246] [Medline: 23554445]
20. Borges do Nascimento IJ, Abdulazeem H, Vasanthan LT, et al. Barriers and facilitators to utilizing digital health technologies by healthcare professionals. NPJ Digit Med 2023;6(1):161. [doi: 10.1038/s41746-023-00899-4]
21. Powell BJ, Mettert KD, Dorsey CN, et al. Measures of organizational culture, organizational climate, and implementation climate in behavioral health: a systematic review. Implement Res Pract 2021;2:26334895211018862. [doi: 10.1177/26334895211018862] [Medline: 37090009]
22. Edmondson A. Psychological safety and learning behavior in work teams. Adm Sci Q 1999 Jun;44(2):350-383. [doi: 10.2307/2666999]
23. Shea CM, Jacobs SR, Esserman DA, Bruce K, Weiner BJ. Organizational readiness for implementing change: a psychometric assessment of a new measure. Implement Sci 2014 Jan 10;9(1):7. [doi: 10.1186/1748-5908-9-7] [Medline: 24410955]
24. Creswell JW, Creswell JD. Research Design: Qualitative, Quantitative, and Mixed Methods Approaches: Sage publications; 2017.
25. Karabukayeva A, Hearld LR, Choi S, Singh JA. Patient perspectives on a decision aid for systemic lupus erythematosus: insights and future considerations. J Rheumatol 2024 Aug 1;51(8):798-803. [doi: 10.3899/jrheum.2023-1178] [Medline: 38749565]
26. Goodman RM, McLeroy KR, Steckler AB, Hoyle RH. Development of level of institutionalization scales for health promotion programs. Health Educ Q 1993;20(2):161-178. [doi: 10.1177/109019819302000208] [Medline: 8491630]
27. Edmondson AC. Learning from mistakes is easier said than done: group and organizational influences on the detection and correction of human error. J Appl Behav Sci 1996 Mar;32(1):5-28. [doi: 10.1177/0021886396321001]
28. Cameron KS, Quinn RE. Diagnosing and Changing Organizational Culture: Based on the Competing Values Framework: John Wiley & Sons; 2011.
29. Denison DR, Spreitzer GM. Organizational culture and organizational development: a competing values approach. Res Organ Change Dev 1991;5(1):1-21.
30. Dalal DK, Zickar MJ. Some common myths about centering predictor variables in moderated multiple regression and polynomial regression. Organ Res Methods 2012 Jul;15(3):339-362. [doi: 10.1177/1094428111430540]
31. StataCorp. Stata: Release 17 Statistical Software. 2021. URL: https://www.stata.com/ [accessed 2025-07-28]
32. Clarke V, Braun V. Teaching thematic analysis: overcoming challenges and developing strategies for effective learning. Psychologist 2013;26(2):120-123 [FREE Full text]
33. Ehrhart MG, Aarons GA, Farahnak LR. Assessing the organizational context for EBP implementation: the development and validity testing of the Implementation Climate Scale (ICS). Implement Sci 2014 Oct 23;9:157. [doi: 10.1186/s13012-014-0157-1] [Medline: 25338781]
34. von Treuer K, Karantzas G, McCabe M, et al. Organizational factors associated with readiness for change in residential aged care settings. BMC Health Serv Res 2018 Dec;18(1):1-6. [doi: 10.1186/s12913-018-2832-4]
35. Woods SA, et al. Innovative work behavior and personality traits: examining the moderating effects of organizational tenure. J Manag Psychol 2018;33(1):29-42. [doi: 10.1108/JMP-01-2017-0016]
36. Kalkan A, Roback K, Hallert E, Carlsson P. Factors influencing rheumatologists' prescription of biological treatment in rheumatoid arthritis: an interview study. Implement Sci 2014 Oct 11;9:153. [doi: 10.1186/s13012-014-0153-5] [Medline: 25304517]

Abbreviations

ADM(J): Average deviation
DA: Decision aid
EBI: Evidence-based interventions
HIPAA: Health Insurance Portability and Accountability Act
***ICD*:** *International Classification of Diseases*
IRB: Institutional Review Board
ORIC: Organizational Readiness for Implementing Change
RWG(J): within group reliability

XSL•FO
RenderX

Edited by A Mavragani; submitted 03.12.24; peer-reviewed by EA Smitherman, M Gasmi, M Shoaran; revised version received 15.07.25; accepted 16.07.25; published 21.08.25.

Please cite as:
Karabukayeva A, Hearld LR, Carroll NW, Kelly R, Singh JA
Organizational Climate and Decision Aid Sustainability in Lupus Care: Mixed Methods Study
JMIR Form Res 2025;9:e69603
URL: https://formative.jmir.org/2025/1/e69603
doi:10.2196/69603

© Aizhan Karabukayeva, Larry R Hearld, Nathan W Carroll, Reena Kelly, Jasvinder A Singh. Originally published in JMIR Formative Research (https://formative.jmir.org), 21.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Comparing Approaches to Teaching Patients How to Use an App-Based Home Spirometer: Randomized Controlled Trial

Caitlin Morgan[1,2], MBChB; Daniel Higbee[2], MBBS, PhD; Catherine Dixon[2,3], MSc; Emma Buckroyd[2], MSc; Huzaifa Adamali[2], MB BCh BAO, MD; Shaney Barratt[1,2], MBBS, PhD; Rahul Shrimanker[1,2], MBBS, PhD; Catherine Hyams[1,2], MBBS, PhD; Paul White[3], MSc, PhD; James William Dodd[1,2], MBChB, PhD

[1]Academic Respiratory Unit, Southmead Hospital, Southmead Road, Bristol, United Kingdom

[2]Respiratory Department, Southmead Hospital, North Bristol NHS (National Health Service) Trust, Bristol, United Kingdom

[3]College of Arts, Technology and Environment, University of the West of England, Bristol, United Kingdom

Corresponding Author:
Caitlin Morgan, MBChB
Academic Respiratory Unit, Southmead Hospital, Southmead Road, Bristol, United Kingdom

Abstract

Background: Bluetooth-enabled, app-based home spirometry has been validated for use in the diagnosis and monitoring of respiratory disease. Remote teaching (virtual or self-directed) of patients on how to use an app-based home spirometry offers the opportunity to deliver diagnostics safely and at scale. However, the most appropriate method of teaching home spirometry to patients is unknown.

Objective: The aim of this pragmatic study undertaken during the COVID-19 pandemic was to determine whether virtual or self-directed teaching were valid methods for deploying home spirometry to patients referred for outpatient lung physiology testing.

Methods: REACH-SPIRO is a single-center, unblinded, randomized controlled trial of adults referred for spirometry. Participants were randomized (1:1:1) to be taught to use a Bluetooth, app-based spirometer either (A) face-to-face, (B) virtually (live video conferencing), or (C) self-directed. Forced vital capacity (FVC) and forced expiratory volume in 1 second (FEV_1) were recorded. Home spirometry (Spirobank Smart Spirometer) readings for each training modality were compared with gold standard hospital measurements (Vyaire Medical) using Bland-Altman Limits of Agreement (LoA). Patients' feedback questionnaires on acceptability and adherence were also collected, and common themes were described using content analysis.

Results: A total of 106 participants were randomized. Bland-Altman analysis between hospital and home FEV_1 measurements for group A showed a mean difference of 0.108L, LoA −0.331L to 0.548L, group B 0.152L, LoA −0.358L to 0.661L, and group C 0.153L, LoA −0.358L to 0.661L. FVC measurements in group A showed a mean difference of 0.123L, LoA −0.402L to 0.648L, group B 0.249L, LoA −0.297L to 0.795L, and group C 0.340L, LoA −0.556L to 1.235L. There was no significant difference between randomized arms for either mean FEV_1 (P=.41) or FVC (P=.84). Patient feedback was similar across all groups, with more positive feedback for face-to-face and virtual teaching methods.

Conclusions: There is no meaningful difference in spirometry measurement between patients taught to use a home spirometer remotely (virtual or self-directed) versus traditional face-to-face treatment. Patients' feedback was favorable to all three methods of teaching. This study supports the use of either virtual or self-directed teaching of home spirometry for the monitoring of respiratory disease. In doing so, we can improve access to spirometry in communities facing barriers to health care and clinically vulnerable populations.

Trial Registration: ISRCTN (International Standard Registered Clinical/Social Study Number) 18299685; https://www.isrctn.com/ISRCTN18299685

(JMIR Form Res 2025;9:e74125) doi:10.2196/74125

KEYWORDS

home spirometry; disease monitoring; medical education; mobile phone; access to healthcare; remote monitoring

Introduction

Bluetooth-enabled handheld spirometry devices have emerged as valuable tools for remote lung function monitoring, particularly for respiratory conditions such as interstitial lung disease, airway disease, and cystic fibrosis [1-3], but also other specialties, including cardiology and neurology. It requires participants to conduct forced breathing maneuvers, which

XSL•FO
RenderX

measure volume and flow of expired air. The gold standard, hospital spirometry, involves specialist equipment and training to perform and interpret.

Home spirometry devices have been used effectively in the prompt management of exacerbations (flares of disease), early detection of disease decline, and preventing cross-infection in vulnerable patient groups [1-3]. During the COVID-19 pandemic, spirometry was considered an aerosol-generating procedure posing an infection risk requiring considerable adaptations, including preclinic screening and full personal protective equipment, which ultimately reduced testing capacity and placed a significant burden on health systems. Home spirometry offers a solution in terms of increasing the capacity of physiology services, but the most appropriate method of teaching home spirometry to patients is unknown.

The validity of different home spirometry devices has been assessed in terms of their use diagnostically and in monitoring for disease progression, with some adopted into the clinical setting. It is widely recognized that home spirometry measurements, including forced expiratory volume in 1 second (FEV_1) and forced vital capacity (FVC), are often lower when compared to hospital spirometry, but the reliability of devices has been inferred by the repeatability shown with recurrent measurements [2-7].

Good quality home and hospital spirometry measurements are based on the Association for Respiratory Technology and Physiology (ARTP) guidelines and imply good supervision and understanding of the technique. Reduced supervision by a skilled physiologist, differences in teaching techniques, and differences in equipment are the key barriers to producing good quality home spirometry described in the literature [6,8,9]. Qualitative analysis of how patients view home spirometry has also highlighted the importance of bespoke use of devices according to patients' needs, so as not to risk psychological well-being [10].

This randomized controlled trial (RCT) evaluates 3 methods of teaching patients to use the MIR (Medical International Research) Spirobank Smart spirometer and assesses clinical and statistical differences in spirometry performance. The randomized groups included face-to-face teaching, virtual teaching, and self-directed learning. Additionally, patient feedback was collected to identify barriers to home spirometry adherence and assess the potential impact on the quality of measurements. This study aims to provide evidence as to whether remote teaching (self-directed or virtual) is an acceptable method by which to deploy home spirometry in the community. By doing so, this study hopes to enhance access to diagnostics, reduce health care burdens, and contribute to sustainable health care practices. Some of the results of this study have been reported previously in the form of an abstract [11].

Methods

Aim

We aim to determine if there is a difference between virtual, self-directed, and face-to-face teaching of home spirometry and gather patients' feedback on willingness to perform quality home spirometry with the MIR Spirobank Smart spirometer.

Design

This pragmatic, single-center, unblinded RCT compared virtual, self-directed, and face-to-face teaching of home spirometry on the Spirobank Smart spirometer in respiratory patients who required spirometry as part of their standard care for diagnosis or monitoring of disease at a central Bristol Hospital, UK (ISRCTN:18299685, Research Ethics Committee [REC] 21/NI/0063). It was designed, set up, and launched during the social distancing rules established in the SARS-CoV-2 pandemic with recruitment from July 2021 to May 2022. Participants were randomized to 1 of 3 groups: (A) face-to-face teaching (control), (B) virtual teaching, and (C) self-directed learning (Figure 1). Participants were asked to submit 1 or more home spirometry readings during this study's period. Statistical analysis was unblinded. The differences in FEV_1 and FVC between hospital (Vyaire Medical) and home spirometry were compared, and the differences in values between virtual, self-directed, and face-to-face teaching were also compared. Patient feedback was evaluated using feedback questionnaires. No changes were made to the methodology after trial commencement.

XSL•FO
RenderX

Figure 1. REACH-SPIRO randomized controlled trial design, Bristol, UK: recruitment from July 2021 to July 2022. Randomization groups shown 1:1:1. F2F: face-to-face; PIS: patient information sheet; SMH: Southmead Hospital.

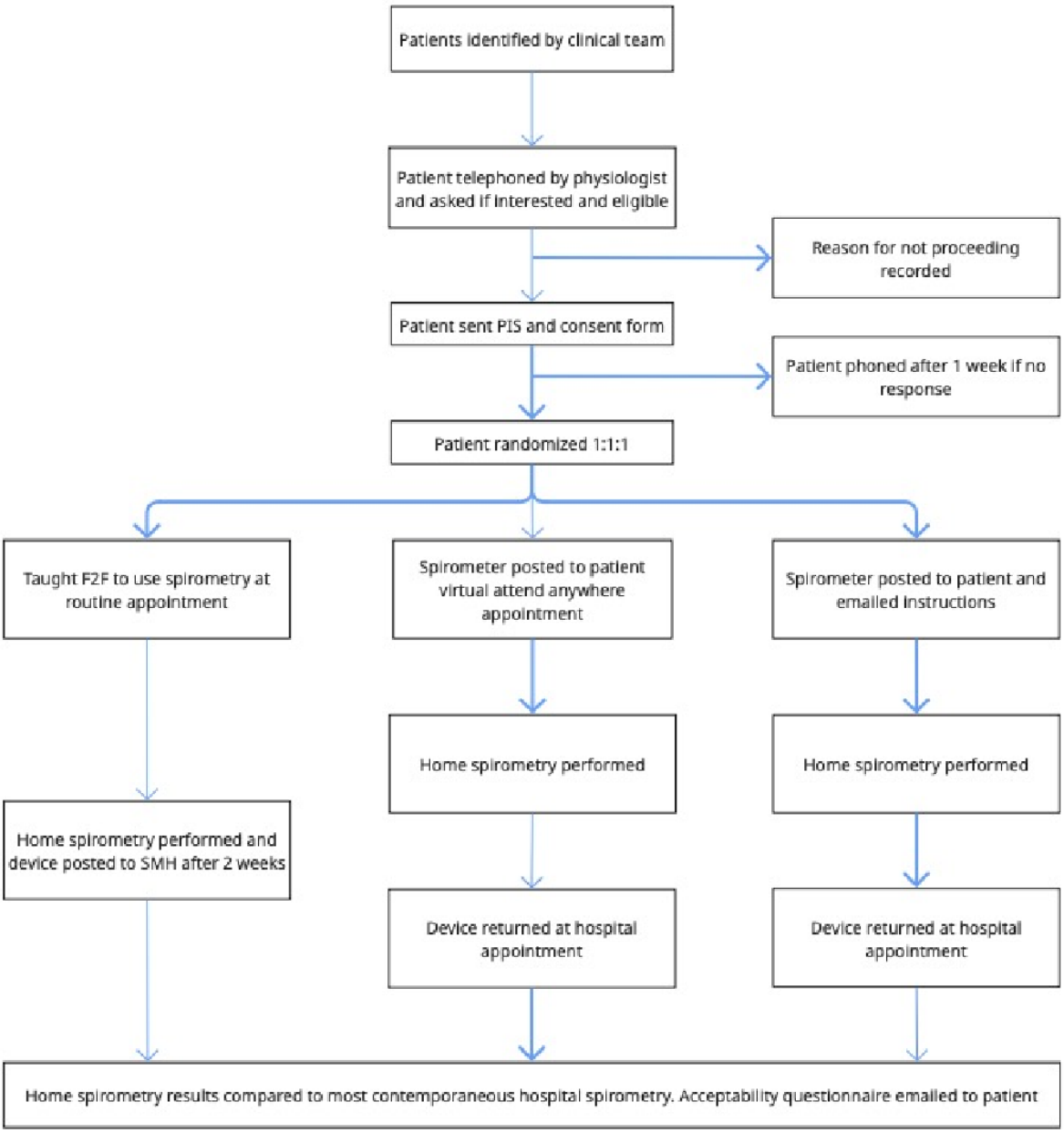

Participants

We included adult patients (aged ≥18 years) referred for hospital spirometry by primary or secondary care with respiratory symptoms. Participants were eligible if they were able to give written informed consent and they owned a smartphone or tablet with an internet connection that they were willing to use for this study. We excluded those with a contraindication to spirometry as per ARTP guidance [12].

Outcome Measures

Primary outcome: a comparison of FEV_1 and FVC measurements between face-to-face, virtual, and self-directed teaching of home spirometry using matched in-hospital spirometry measurements.

Secondary outcomes: understand and assess factors that affect patient willingness, adherence, and quality of home spirometry.

Recruitment and Randomization

All patients under the care of the North Bristol NHS (National Health Service) Trust (NBT) Respiratory service were eligible for recruitment. Patients had to be aged older than 18 years, require spirometry as part of their routine clinical care, have a smartphone and internet connection, and be able to give written informed consent. Patients were identified by a member of their clinical team as requiring spirometry for the diagnosis or monitoring of a respiratory disease. They were approached

XSL•FO
RenderX

within 3 weeks of their hospital referral via telephone by a member of the physiology team and screened for eligibility. If provided, reasons for declining to participate were recorded. Consented patients were randomized using a secure web-based system (REDCap [Research Electronic Data Capture; Vanderbilt University]). Participants were randomized in a 1:1:1 ratio to either face-to-face teaching (control arm), virtual teaching, or self-directed learning. It was not practical to use blind interventions in this study. Randomization was not stratified.

After Randomization

(A) Control Arm (Face-to-Face Teaching)

Patients allocated to the control arm attended the hospital for their routine laboratory-based hospital spirometry. They had a face-to-face teaching session with a physiologist who helped download the app to the patient's phone, set up the Spirobank Smart Spirometer, and perform spirometry using the equipment. The patient practiced the maneuvers with the physiologist. They then performed spirometry at home using the Spirobank Smart spirometer 1 or more times. This was supplemented by the NBT Home Spirometry Patient Information Leaflet and access to the Spirobank Smart spirometer patient information video produced by NBT and available on the NBT website.

(B) Intervention Arm 1 (Virtual Remote Teaching: National Health Service Attend Anywhere)

Participants allocated to the first intervention arm were sent the home spirometer and information leaflet by post. This was supplemented by the instructional video. Participants then had a virtual appointment with a member of the respiratory physiology team. At this virtual appointment, they were taught how to use the spirometer and app. They performed spirometry during the virtual appointment and then at home 1 or more times. Patients attended their routine in-hospital spirometry appointment after they had submitted measurements using the home device.

(C) Intervention Arm 2 (Self-Directed Learning)

Participants allocated to the second intervention arm were sent the spirometer and information leaflet by post. This was supplemented by the instructional video. They performed spirometry at home 1 or more times. Patients attended their routine in-hospital spirometry appointment after they had submitted measurements using the home device.

All Randomized Arms

Participants were asked to record their spirometry at least once, but as often as they liked or felt was clinically indicated. The most contemporaneous home and hospital measurements were used in the final analysis to assess diagnostic utility. All spirometers were calibrated before being issued to each participant, and spirometers were used according to the guidelines. All patients were invited by email to provide feedback questionnaires after returning the home equipment. A 1 - 10 Likert scale was used to answer questions on the ease of use of the spirometer, the app, and whether they would want to continue using the home spirometry device. Free text boxes were also provided for feedback on the "challenges to performing home spirometry" and "feedback on the teaching method used." A question on whether harm was caused as a result of home spirometry was also included in the questionnaire (see Multimedia Appendix 1). The questionnaire is not a validated tool for analysis in this setting and was created for the purposes of this study based on common barriers to home spirometry in the literature.

Sample Size

This is a pragmatic trial based on the number of patients referred for spirometry across this study's period. Given the time-sensitive nature of improving diagnostics and monitoring strategies for vulnerable patient groups during the COVID-19 pandemic and the long wait times of current spirometry services, a formal power calculation was not conducted.

Acceptability

Acceptable spirometry measurements are defined by European Respiratory Society and American Thoracic Society criteria as summarized by the most recent ARTP statement on pulmonary function testing [13,14]. Acceptability was assessed by trained physiologists.

Statistical Analysis

Primary Outcome

The differences in FEV_1 and FVC between hospital (Vyaire Medical) and home spirometry in the three arms were initially assessed using independent samples *t* tests. Within each arm, limits of agreement between hospital (gold standard) and home spirometry were assessed using Bland-Altman methodology (set to 2 SDs from the mean), with agreement bias between home and hospital measures quantified using 95% CIs. Per-protocol analysis, including only those who had produced acceptable hospital spirometry readings (per European Respiratory Society and American Thoracic Society criteria), was also performed at each stage to reduce bias introduced by unacceptable readings (available in Table S1 in Multimedia Appendix 1).

Secondary Objectives

Of those who submitted a questionnaire after completing this study, median scores and IQR were calculated for questions 1 - 3 of the feedback questionnaire, and content analysis was used to analyze the responses in the white boxes. The results of the qualitative analysis were combined with frequency statistics of missing data and "unacceptable results" (felt to reflect quality of home spirometry readings and therefore quality of teaching) to assess willingness, quality, and adherence to home spirometry.

Ethical Considerations

This study was fully approved by the Health Research Authority (HRA) Research Ethics Committee Northern Ireland (REC reference 21/NI/0063). The full study protocol has been registered with the ISRCTN registry (ISRCTN18299685). All participants were recruited with informed written consent by means of participant-dated signature by a member of the research team with GCP training. Informed written consent was obtained after the participant reviewed the patient information sheet detailing the exact nature of this study and any risks

XSL•FO
RenderX

involved in taking part. Data used in the analysis, both quantitative and qualitative, were anonymized. Participants were not compensated for their time as this study did not involve any extra visits outside of standard clinical appointments.

Results

Overview

Between July 13, 2021, and May 26, 2022, a total of 263 patients were screened and 106 were randomized. Principal reasons for exclusion from the study were patient choice not to be involved in this study (24/157, 19%), and an inability to use the technology or a home spirometer (23/157, 15%). The last patient completed this study on July 11, 2022. As per the study protocol, the study was open for 1 year. Baseline characteristics are described in Table 1.

Table . REACH-SPIRO randomized controlled trial participant flow diagram, Bristol, UK: recruitment from July 2021 to July 2022. Baseline demographic and clinical characteristics for each randomized group by age, gender, previous spirometry experience, and background general practitioner diagnosed respiratory condition.

Characteristics	Face-to-face teaching (n=34)	Virtual teaching (n=36)	Self-directed teaching (n=36)
Age (years)	57	53	58
Minimum-maximum (years)	30-74	29-73	34-82
Female sex, n (%)	19 (56)	21 (58)	24 (67)
Previous spirometry experience ever, n (%)	12 (35)	7 (19)	10 (28)
Hospital spirometry	8 (24)	6 (17)	6 (17)
GP[a] spirometry	4 (12)	1 (3)	4 (12)
Any respiratory diagnosis, n (%)	16 (47)	13 (36)	13 (36)
Asthma	9 (26)	8 (22)	9 (26)
COPD[b]	2 (6)	1 (3)	3 (8)
Mixed airways	2 (6)	0 (0)	0 (0)
ILD[c]	3 (9)	2 (6)	1 (3)
Other	0 (0)	2 (6)	0 (0)

[a]GP: general practitioner.

[b]COPD: chronic obstructive pulmonary disease.

[c]ILD: interstitial lung disease.

Data Completeness

Missing home and hospital measurements for the 2 metrics analyzed are outlined in Figure 2. A total of 100% (85/85) of those who provided both hospital and home spirometry measurements provided feedback.

XSL•FO
RenderX

Figure 2. REACH-SPIRO randomized controlled trial participant flow diagram, Bristol, UK: recruitment from July 2021 to July 2022. Randomization groups showed 1:1:1 with final intention-to-treat and per-protocol recruitment for analysis. Exclusion and drop-out reasons were documented. FEV_1: forced expiratory volume in 1 second; FVC: forced vital capacity; ITT: intention-to-treat.

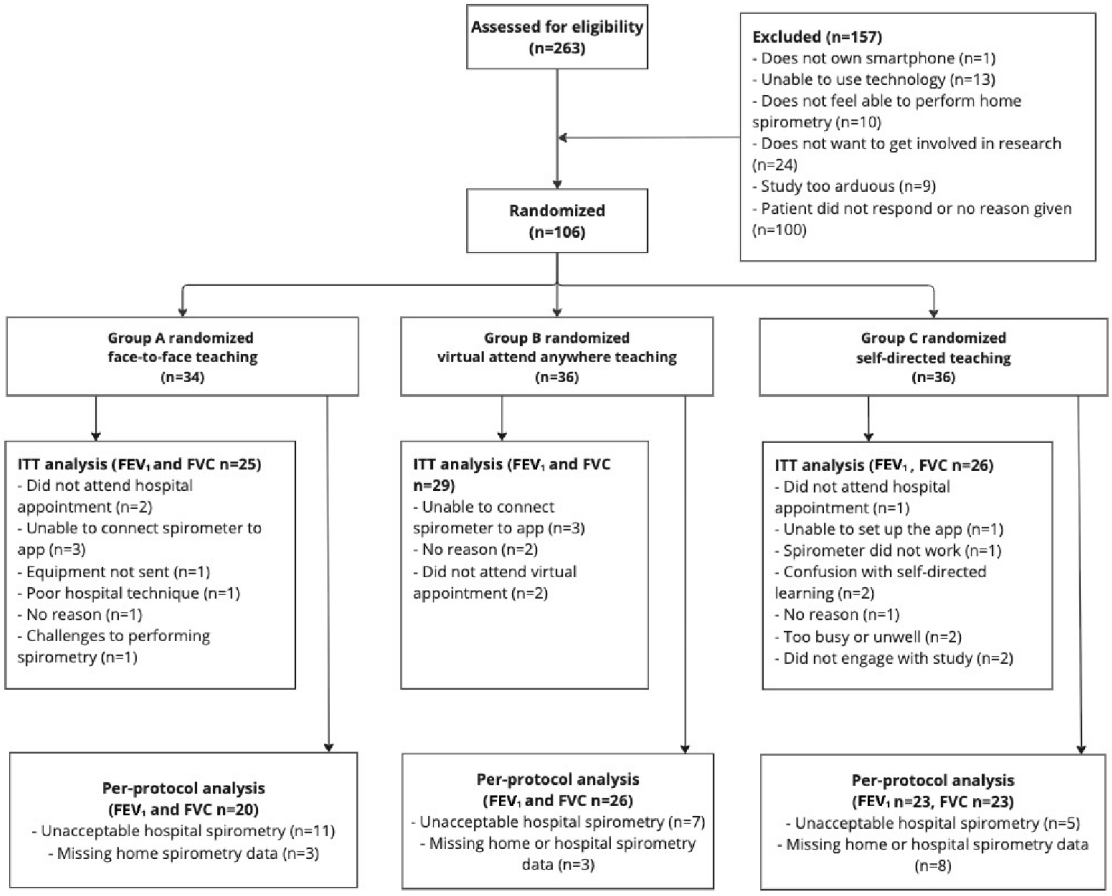

Clinical Outcomes

Baseline Demographics

As shown in Table 1, the mean participant age was 56 years, with 60% female and 95% of White ethnicity. Preexisting diagnoses were recorded based on general practitioner records. Asthma was the most common in all groups (n=9, 26% in groups A and C and n=8, 22% in group B). Previous spirometry experience was based on whether participants had ever performed hospital or general practitioner–based spirometry. This varied across groups, with the highest in group A (n=12, 35%).

Primary Outcome: Difference in FEV_1 and FVC Measurements Between Teaching Methods

The median number of days between hospital and home spirometry measurements was 3 (IQR 1‐9) in group A, 3 (IQR 1‐7) in group B, and 7 (IQR 1‐14) in group C. The difference in means between home and hospital measurements does not significantly vary between randomized arms in FEV_1 ($F_{2,76}$=2.262, mean squared error=2724.0, P=.11) or FVC ($F_{2,78}$=2.595, mean squared error=0.059, P=.08). Table 2 and Figure 3 compare FEV_1 and FVC measurements on the MIR Spirobank device at home versus hospital readings. The limits of agreement for the Bland-Altman plots were set to 2 SD from the mean.

XSL•FO
RenderX

Table . REACH-SPIRO home spirometry randomized controlled study, Bristol, UK: recruitment from July 2021 to July 2022. Intention-to-treat analysis in all randomized groups. Correlation between home and hospital measurements, mean difference, and limits of agreement with 95% CIs.

Variables	Pearson correlation between home and hospital (95% CI)	*P* value	Agreement bias or mean difference	Bland-Altman limits of agreement (95% CI)
FEV_1[a]				
Group A (n=25)	0.972 (0.936 to 0.988)	<.001	0.108	-0.218 to 0.435 (-0.331 to 0.548)
Group B (n=29)	0.972 (0.940 to 0.987)	<.001	0.152	-0.232 to 0.536 (-0.358 to 0.661)
Group C (n=26)	0.981 (0.957 to 0.991)	<.001	0.153	-0.217 to 0.523 (-0.342 to 0.649)
Overall (n=80)	0.975 (0.962 to 0.984)	<.001	0.138	-0.221 to 0.498 (-0.291 to 0.568)
FVC[b]				
Group A (n=26)	0.955 (0.902 to 0.980)	<.001	0.123	-0.402 to 0.648 (-0.581 to 0.827)
Group B (n=29)	0.958 (0.912 to 0.980)	<.001	0.249	-0.297 to 0.795 (-0.478 to 0.976)
Group C (n=26)	0.932 (0.853 to 0.969)	<.001	0.340	-0.556 to 1.235 (-0.860 to 1.539)
Overall (n=81)	0.942 (0.911 to 0.962)	<.001	0.238	-0.462 to 0.937 (-0.596 to 1.071)

[a]FEV_1: forced expiratory volume in 1 second.

[b]FVC: forced vital capacity.

XSL•FO
RenderX

Figure 3. REACH-SPIRO randomized controlled trial, Bristol, UK: recruitment from July 2021 to July 2022. Comparison of teaching methods by randomized arm for FEV_1 and FVC: A=face-to-face, B=virtual, and C=self-directed; P1 FEV_1 correlation coefficient scatterplot of all randomized groups. P2 FEV_1 Bland-Altman plot of all randomized groups. P3 FVC correlation coefficient scatterplot of all randomized groups. P4 FVC Bland-Altman plot of all randomized groups. FEV_1: forced expiratory volume in 1 second; FVC: forced vital capacity.

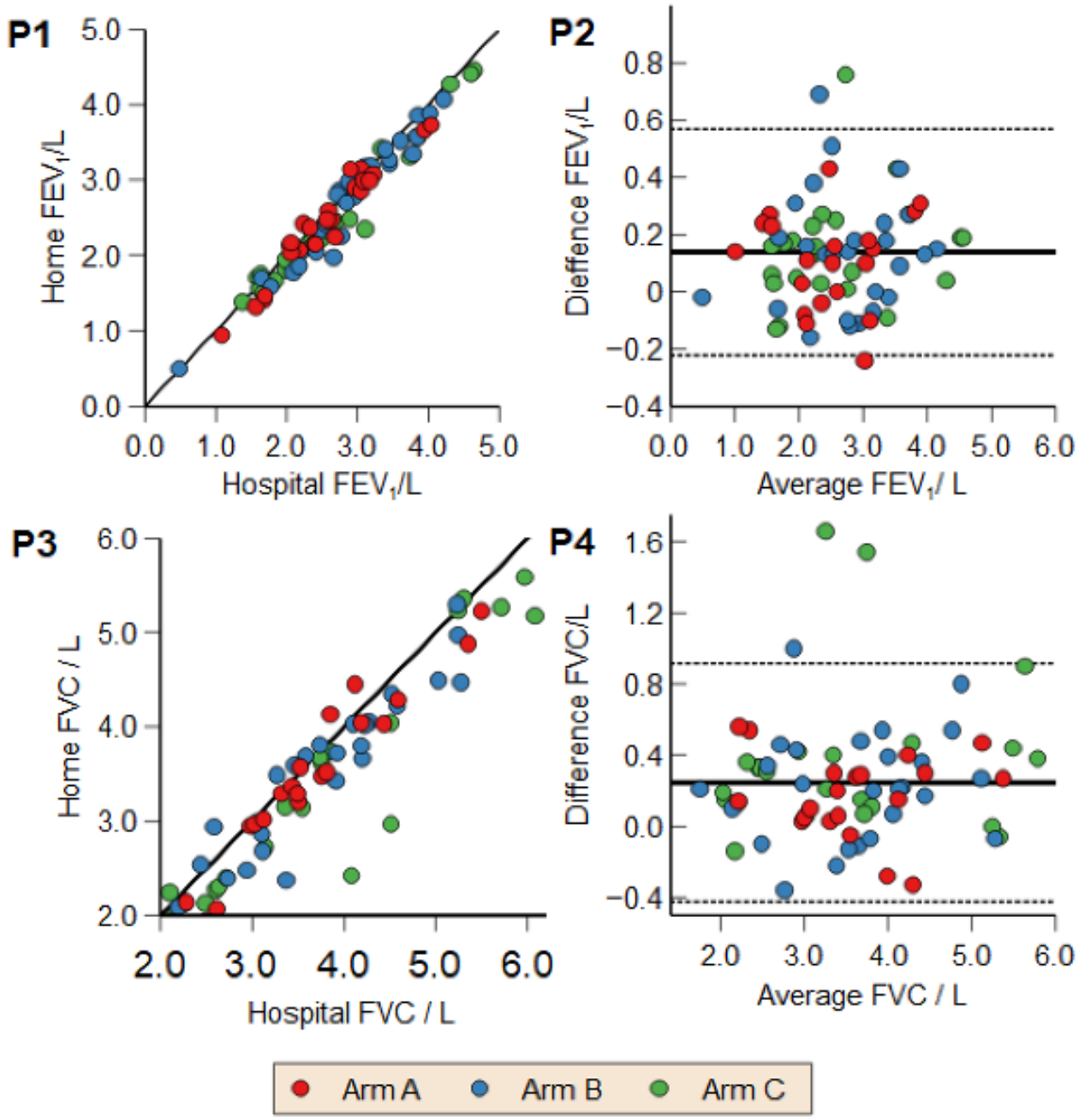

Difference Between Hospital and Home Spirometry Readings

Home and hospital spirometry were highly correlated, as shown in Table 2 (Pearson correlation) and in P1 and P3, Figure 3. In line with previous studies in the literature, FVC (mean difference 0.238L, $P<.001$, 95% CI 0.160 - 0.315) measurements were higher in the hospital compared to the most contemporary home spirometry measurement performed by each participant. This was the case in all teaching groups. FEV_1 mean difference (0.138L, $P<.001$, 95% CI 0.097 - 0.180) was significantly higher in hospital readings but still within ARTP reproducibility criteria of 150 ml. A subanalysis of only those participants who produced acceptable hospital spirometry readings was performed (Table S1 in Multimedia Appendix 1), but no difference was noted in any outcome from the findings of the intention-to-treat analysis.

Participant Feedback Questionnaires

Responses to acceptability questions were rated on a Likert scale of 1 - 10 (1=hard, 10=easy), and a median with IQR was calculated. In group A, 28 submitted feedback questionnaires. The response to question 1 “How easy did you find performing spirometry at home?” was 9 (IQR 7-9). The median response to question 2 “How easy did you find it to use the app and email your results?” was 8.5 (IQR 5.5-8.5). Question 3: “How happy

XSL•FO
RenderX

would you be to continue to use the device to monitor your lung function at home?" showed a median response of 9.5 (IQR 7.5-9.5). A total of 28 participants submitted feedback in group B. Question 1 showed a median response of 10 (IQR 8-10), question 2 a response of 9 (IQR 7-9), and question 3 a response of 10 (IQR 8.5-10). Finally, 29 participants in group C provided feedback. Question 1 had a response of 8 (IQR 3.5-8), question 2 a response of 9 (IQR 4.5-9), and question 3 a response of 9 (IQR 5.5-9).

Challenges to Home Spirometry and Feedback on Teaching Methods

Content analysis was conducted on written participant feedback. Common themes are outlined in Table 3.

Table . REACH-SPIRO home spirometry randomized controlled trial, Bristol, UK: recruitment from July 2021 to July 2022. Content analysis of participant feedback to "challenges to home spirometry" and "feedback on teaching method" with key themes.

Participant challenges and feedback	Group A (n=28), n (%)	Group B (n=28), n (%)	Group C (n=29), n (%)
Challenges to home spirometry			
Reported issues with the app	9 (32)	3 (11)	5 (18)
Connecting to the internet			
Downloading the app			
Emailing spirometry measurements			
Issues with the home spirometer device	0 (0)	1 (3.5)	0 (0)
Reports of not finding "the time" to conduct tests in an informal setting	3 (11)	2 (7)	1 (3.5)
Receiving an alert that the test was "unacceptable"	2 (7)	2 (7)	4 (14)
No problems reported	4 (14)	0 (0)	2 (7)
No feedback given	10 (36)	20 (71)	17 (59)
Feedback on teaching method			
More teaching requested	3 (11)	3 (11)	3 (10)
More info on downloading or navigating the app			
A demonstration video before each test			
Advice on app warnings			
Positive feedback on teaching	9 (32)	9 (32)	5 (17)
Convenient			
Easy to follow or clear			
Good leaflet			
No feedback given	16 (57)	15 (54)	21 (72)
Connectivity issues with virtual teaching	0 (0)	1 (3.5)	0 (0)

The face-to-face group had the most reported challenges with the home spirometer app on their device and struggled to find the time to perform home spirometry outside of the clinical setting. This group also provided the most feedback. Only 1 participant had issues with the spirometer device (group B). The face-to-face and virtual teaching groups had an equal (n=9, 32%) number of participants reporting positive feedback on the teaching methods used. The self-directed group had fewer positive comments on teaching methods (n=5, 17%). Participants reported alerts on the app advising them that their results were "unacceptable." Participants noted concern in feedback that this meant they had not performed the test correctly. More participants reported issues with "unacceptable" results in the self-directed arm (n=4, 14% vs n=2, 7% in groups A and B versus C).

As shown in Figure 1, missing data and "unacceptable" spirometry readings were recorded in all 3 groups. This was more apparent in the face-to-face and self-directed groups.

Harms

One participant in the virtual teaching group (group B) reported that trying to use the app was "stressful and frustrating." A total of 2 participants in the self-directed group (group C) expressed a desire for physician interpretation of results. No participants reported any side effects from performing home spirometry.

XSL•FO
RenderX

Discussion

This is the first RCT comparing methods of teaching home spirometry. It has been shown that there is no statistically significant difference between face-to-face, virtual, or self-directed teaching of home spirometry on the MIR Spirobank Smart spirometer. The qualitative analysis of patient feedback did not identify clear differences in usability or participant willingness and adherence to home spirometry depending on which teaching method was used. This trial supports the use of all 3 methods of teaching and therefore, the use of remote teaching (virtual or self-directed) in clinical practice to improve accessibility.

Participant feedback confirmed that technical difficulties with the app occurred across all 3 groups. Concerns were also raised about the app reporting that readings were "unacceptable" without giving a reason why. This led to concern that the test had not been performed correctly. This finding is supported by a recent study looking at virtual teaching of home spirometry in a cystic fibrosis population [3], which found that adherence after virtual teaching was high, but concerns were raised by participants regarding the accuracy of the results provided. Davis et al [3] also commented on the benefits of virtual teaching in allowing for "top-up" sessions to maintain the technique.

This trial has shown that in the outcomes measured (FEV_1 and FVC) in each randomized group, home spirometry readings were on average lower than those recorded using the in-hospital Vyntus equipment alongside a trained physiologist (138 and 238 ml, respectively). This is consistent with other studies, including a meta-analysis by Anand et al [7], which concluded that unsupervised spirometry is not a suitable replacement for supervised spirometry after assessing the results of 28 separate studies. A 2018 study, which assessed the Spirobank Smart spirometer in an interstitial lung disease cohort, also found an average FEV_1 mean difference between hospital and home spirometry of 220 ml even when participants were trained to produce acceptable readings before independent use at home [1]. Similarly, in 2012, the Spirobank Smart spirometer showed an underestimation of FEV_1 by 5% when compared to clinic readings [8]. Kerwin et al [4] also found home FEV_1 results were "moderately lower" than clinic readings, despite being an average of weekly measurements, but due to the high correlation between home and hospital measurements, this study concluded that home spirometry remained a reliable method of monitoring disease. Reasons for lower home versus hospital spirometry include the time of day that readings were taken, and also the impact of trained physiologists during in-hospital measurements [1,3,4,8]. Spirometry is a hard technique to implement, and acceptable measurements are challenging even when supported by a trained physiologist in a hospital setting. Therefore, it has been argued that only acceptable measurements should be relied upon clinically [3,9]. Based on this, we performed a subanalysis that included only those participants who performed "acceptable" hospital spirometry measurements, under the assumption that those who were unable to perform acceptable readings under optimum (hospital) conditions were less likely to produce acceptable readings without any supervision (Table S1 in Multimedia Appendix 1). However, this did not show a difference in results with FEV_1 and FVC, all underestimated in the home measurements by 133 and 259 ml, respectively. We also considered the potential impact of differences in the interval between hospital and home measurements. In our study, this was small (median 3 and 7, IQR 1-14 d) with the largest spread in group C. This short timeframe limits the influence of changes in the clinical state of the patient or the risk of patients forgetting the taught instruction on the spirometry result. However, in those with a large delay, these factors may have impacted the results. Together, our data and others suggest that home spirometry may be best suited to monitoring lung function rather than relying on an off-diagnostic threshold.

This trial has several limitations, including sample size. This was a pragmatic design during the COVID-19 pandemic and may be underpowered. However, it is one of the largest studies looking at home spirometry and one of the only studies comparing virtual or self-directed to face-to-face teaching. Researchers noted that just 9% (14/157) of those approached to take part declined due to a lack of a smartphone or a lack of confidence with the spirometer or app. Our population was young, with an average age of 53 - 58 years across the 3 groups, with a wide age range. Unlike other studies, we have not removed outlying results from our Bland-Altman analysis, and as such, we have noted large differences between home and hospital measurements. These limitations do not affect the generalizability of this study, and we were able to conclude that virtual or self-directed teaching of home spirometry was comparable to the gold standard (face-to-face teaching). A further limitation of this study was the lack of standardization for when home measurements were taken. Furthermore, participants were not asked to comment on clinical status during home or hospital measurements. However, in the context of this pragmatic RCT and given that all participants were presenting to clinicians with diagnosed or undiagnosed respiratory symptoms, some variation may be expected. Finally, the feedback questionnaire used in this study was designed for this study and, as such, has not been independently validated. There is no validated questionnaire for use in this setting. The questions were devised using common themes in the literature regarding barriers to home spirometry use.

In conclusion, virtual or self-directed instruction are effective methods for teaching home spirometry, yielding comparable results to face-to-face methods in both the quantitative and qualitative analysis. Like other studies, we have identified differences in home versus hospital readings, which suggest that home spirometry should not replace the gold standard diagnostic spirometry. Spirometry provision in the United Kingdom is unable to meet current demands and is extremely variable in different regions. The REACH-SPIRO trial has contributed evidence to support remote teaching of home spirometry (virtually or self-directed), which could help increase the provision of spirometry in clinical and research settings.

XSL•FO
RenderX

Acknowledgments

This study was funded by the Southmead Hospital Charitable Fund in the form of infrastructure funding and purchase of the home spirometry devices. This study was supported by the National Institute for Health and Care Research, Bristol Biomedical Research Centre. The views expressed are those of the authors and not necessarily those of the National Institute for Health and Care Research or the Department of Health and Social Care.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to patient confidentiality rules but are available from the corresponding author on reasonable request.

Authors' Contributions

DH, JWD, PW, CD, EB, HA, SB, and RS generated the research questions and analysis plans. CM, DH, CD, and EB were involved in data collection. CM, EB, PW, and JWD undertook data analysis. All authors (CM, DH, CD, EB, HA, SB, RS, CH, PW, and JWD) contributed to preparation of this paper and its revisions before publication. JWD provided oversight of the research.

Conflicts of Interest

CM, DH, CD, EB, HA, and PW have no conflicts of interest. JWD has received grants from AstraZeneca (AZ), Chiesi, and NIHR (National Institute for Health and Care Research) and honoraria from GSK (GlaxoSmithKline plc), AZ, Chiesi, and PulmonX. CH has received grants and honoraria from Pfizer. RS has received honoraria from AZ, GSK, and Chiesi. SB has received honoraria from Boehringer Ingelheim. All honoraria for presentations or meetings are unrelated to the submitted paper.

Multimedia Appendix 1
Per-protocol analysis and REACH-SPIRO RCT Questionnaire.
[DOCX File, 137 KB - formative_v9i1e74125_app1.docx]

Checklist 1
CONSORT checklist (V2). CONSORT: Consolidated Standards of Reporting Trials.
[PDF File, 219 KB - formative_v9i1e74125_app2.pdf]

References

1. Moor CC, Wapenaar M, Miedema JR, Geelhoed JJM, Chandoesing PP, Wijsenbeek MS. A home monitoring program including real-time wireless home spirometry in idiopathic pulmonary fibrosis: a pilot study on experiences and barriers. Respir Res 2018 May 29;19(1):105. [doi: 10.1186/s12931-018-0810-3] [Medline: 29843728]
2. Oppenheimer J, Hanania NA, Chaudhuri R, et al. Clinic vs home spirometry for monitoring lung function in patients with asthma. Chest 2023 Nov;164(5):1087-1096. [doi: 10.1016/j.chest.2023.06.029] [Medline: 37385337]
3. Davis J, Ryan M, Marchetti P, et al. Real-world feasibility of short-term, unsupervised home spirometry in CF. Pediatr Pulmonol 2022 Dec;57(12):3129-3135. [doi: 10.1002/ppul.26147] [Medline: 36124390]
4. Kerwin EM, Hickey L, Small CJ. Relationship between handheld and clinic-based spirometry measurements in asthma patients receiving beclomethasone. Respir Med 2019 May;151:35-42. [doi: 10.1016/j.rmed.2019.03.010] [Medline: 31047115]
5. van de Hei SJ, Flokstra-de Blok BMJ, Baretta HJ, et al. Quality of spirometry and related diagnosis in primary care with a focus on clinical use. NPJ Prim Care Respir Med 2020 May 15;30(1):22. [doi: 10.1038/s41533-020-0177-z] [Medline: 32415077]
6. Kołtowski Ł, Basza M, Bojanowicz W, et al. Remotely supervised spirometry versus laboratory-based spirometry during the COVID-19 pandemic: a retrospective analysis. Respir Res 2024 Jan 18;25(1):39. [doi: 10.1186/s12931-023-02586-0] [Medline: 38238745]
7. Anand R, McLeese R, Busby J, et al. Unsupervised home spirometry versus supervised clinic spirometry for respiratory disease: a systematic methodology review and meta-analysis. Eur Respir Rev 2023 Sep 30;32(169):169. [doi: 10.1183/16000617.0248-2022] [Medline: 37673426]
8. Degryse J, Buffels J, Van Dijck Y, Decramer M, Nemery B. Accuracy of office spirometry performed by trained primary-care physicians using the MIR Spirobank hand-held spirometer. Respiration 2012;83(6):543-552. [doi: 10.1159/000334907] [Medline: 22269344]
9. Rodriguez-Roisin R, Tetzlaff K, Watz H, et al. Daily home-based spirometry during withdrawal of inhaled corticosteroid in severe to very severe chronic obstructive pulmonary disease. Int J Chron Obstruct Pulmon Dis 2016;11:1973-1981. [doi: 10.2147/COPD.S106142] [Medline: 27578972]

XSL•FO
RenderX

10. Mandizha J, Lanario JW, Duckworth A, et al. Patient perspectives on home-spirometry in interstitial lung disease: a qualitative co-designed study. BMJ Open Respir Res 2023 Oct;10(1):e001837. [doi: 10.1136/bmjresp-2023-001837] [Medline: 37793682]
11. Morgan C, Higbee D, Dixon C, et al. A randomised controlled trial of self vs clinician-taught home spirometry. Eur Respir J 2023 Sep 9;62(suppl 67):PA3976. [doi: 10.1183/13993003.congress-2023.PA3976]
12. Cooper BG. An update on contraindications for lung function testing. Thorax 2011 Aug;66(8):714-723. [doi: 10.1136/thx.2010.139881] [Medline: 20671309]
13. Sylvester KP, Clayton N, Cliff I, et al. ARTP statement on pulmonary function testing 2020. BMJ Open Respir Res 2020 Jul;7(1):e000575. [doi: 10.1136/bmjresp-2020-000575] [Medline: 32631927]
14. Graham BL, Steenbruggen I, Miller MR, et al. Standardization of spirometry 2019 update. an official american thoracic society and european respiratory society technical statement. Am J Respir Crit Care Med 2019 Oct 15;200(8):e70-e88. [doi: 10.1164/rccm.201908-1590ST] [Medline: 31613151]

ABBREVIATIONS

ARTP: Association for Respiratory Technology and Physiology
FEV_1: forced expiratory volume in 1 second
FVC: forced vital capacity
LoA: limits of agreement
MIR: Medical International Research
NBT: North Bristol NHS (National Health Service) Trust
RCT: randomized controlled trial
REDCap: Research Electronic Data Capture

Edited by A Mavragani; submitted 18.03.25; peer-reviewed by M Doumit; revised version received 08.07.25; accepted 09.07.25; published 14.08.25.

Please cite as:
Morgan C, Higbee D, Dixon C, Buckroyd E, Adamali H, Barratt S, Shrimanker R, Hyams C, White P, Dodd JW
Comparing Approaches to Teaching Patients How to Use an App-Based Home Spirometer: Randomized Controlled Trial
JMIR Form Res 2025;9:e74125
URL: https://formative.jmir.org/2025/1/e74125
doi:10.2196/74125

© Caitlin Morgan, Daniel Higbee, Catherine Dixon, Emma Buckroyd, Huzaifa Adamali, Shaney Barratt, Rahul Shrimanker, Catherine Hyams, Paul White, James William Dodd. Originally published in JMIR Formative Research (https://formative.jmir.org), 14.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Value-Based Framework for Evaluating Pre-Commercial Procurement: Case Study of Value-Based Key Performance Indicators

Nils-Otto Ørjasæter[1], MSc, PhD; Kari Jorunn Kværner[1], MSc, MD, PhD; Linn Nathalie Støme[1,2], MSc, PhD

[1]Department of Early Psychosis, Oslo University Hospital, Forskningsveien 2B, 0373, Oslo, Norway

[2]Department of Addiction Research, Oslo University Hospital, Oslo, Norway

Corresponding Author:
Nils-Otto Ørjasæter, MSc, PhD
Department of Early Psychosis, Oslo University Hospital, Forskningsveien 2B, 0373, Oslo, Norway

Abstract

Background: The demographic shift toward older populations is placing increasing pressure on health care systems, and only 20% of patients with chronic issues in the industrial world's rural areas have guaranteed access to adequate health care services. This stresses the health care systems, emphasizing the need for innovative solutions. The Horizon 2020 Pre-Commercial Procurement (PCP) project, Crane, addresses these needs by facilitating the procurement of a digital self-management system for treating patients with chronic issues at home. Three rural European regions are participating in the project: Västerbotten (Sweden), Extremadura (Spain), and Agder (Norway).

Objective: This study aims to explore and identify key design criteria and value-based key performance indicators (VB-KPIs) to support the development and evaluation of digital health care solutions for patients with chronic issues in rural areas within the Crane PCP process.

Methods: A 3-iteration process was used to identify and prioritize the VB-KPIs in the Crane project. First, user needs were investigated based on stakeholder analyses in the participating rural regions. The early health technology assessment tool, Step Up, was used in 5 workshops (2 in Agder, 2 in Extremadura, and 1 in Västerbotten). Participants included patients and health care professionals. Second, post workshop, stakeholders were asked to comment on the summarized results, which were accordingly adjusted. Third, following the workshops, VB-KPIs were identified and prioritized, and discussions among representatives from the 3 buyer regions were conducted.

Results: Thirty-five VB-KPIs across 5 domains were identified. User-related (9 VB-KPIs), employee-related (9 key performance indicators), clinical (4 VB-KPIs), organizational (6 VB-KPIs), and economic (8 VB-KPIs) outcomes from the workshops and the subsequent discussions emphasized regional differences in terms of user needs and priorities. While Agder (Norway) and Västerbotten (Sweden) emphasized privacy, digital trust, and physical interaction as important, Extremadura (Spain) prioritized negotiation and shared decision-making. Despite differences, shared values were identified, including empowerment, flexibility, preventative care, and improved quality of life.

Conclusions: The identified and prioritized VB-KPIs are likely to provide a need-based foundation for the development and subsequent evaluation of the digital PCP, Crane, although regional socioeconomic and cultural differences may necessitate local adaptations.

(JMIR Form Res 2025;9:e71279) doi:10.2196/71279

KEYWORDS

internet-based self-management system for treating chronic diseases; treatment of chronic patients; pre-commercial procurement of internet and digital-based solutions in health care; evaluation of pre-commercial procurement solutions; value-based procurement; value-based key performance indicators in health care; early health technology assessment

Introduction

Background

Aging is one of the most important changes in population structure currently taking place in industrialized countries [1]. Increasing longevity and declining fertility rates are shifting the age distribution of populations in industrialized countries toward older age groups. As a result, the number of older adult people per 100 working-age people will nearly triple, from 20 in 1980 to 58 in 2060 [1]. The shift is more severe in rural areas than urban areas, mainly due to an increasing number of younger people moving to cities [1]. People living in rural areas face greater difficulties in accessing health and social care services.

XSL•FO
RenderX

Older people require these services more frequently and often have additional challenges due to loss of mobility and cognitive function. It is assumed that around 37% of Europeans older than 65 years have multiple chronic conditions, although a 3-fold variation between EU member states is reported [2]. Older than the age of 65 years, men and women have an additional life expectancy of approximately 18 and 21 years, respectively. For both genders, only 9 of these years are expected to be of high quality of life. The remaining time is expected to represent age-related morbidity due to one or more chronic diseases (multimorbidity) [3].

In rural areas, great geographical distances and a lack of transportation services commonly represent additional challenges [4]. Thus, only 20% of patients with chronic issues in these areas can be guaranteed adequate health care services. The remaining 70% - 80% need to be empowered to increase self-management of health care needs [5]. These demographic changes are expected to stress the health care systems and urge the need for innovations. To meet these challenges, moving more of the health care to the patient's home through digital solutions is a growing trend [6]. Crane [7] is a Horizon 2020 Pre-Commercial Procurement (PCP) project, compliant with the European Health Data Space [8] concept, addressing these challenges. PCPs enable the industry from the demand side to develop innovative solutions based on public sector needs. As such, the scheme provides a first customer reference that enables companies to create a competitive advantage in the marketplace. PCP allows public buyers to compare alternative potential solution approaches and filter the best possible solutions the market can offer to meet the public's needs [7].

The Basis for This Study

Crane is an internet and digital-based health care PCP project. The aim is to develop an integrated and flexible self-management system for treating patients with chronic issues at home, and thus improving the quality of life of the patients and optimizing the use of health resources.

Crane is run by partners from 3 rural regions: region Västerbotten in Sweden, Extremadura in Spain, and Agder in Norway. The project is divided into 3 phases. The first phase, "solution design," was a feasibility study where 5 bidders were selected. In the second phase, "prototype development," 3 proposals were selected from the first phase. This phase is still in progress, and the goal is to develop a prototype. Two suppliers will be selected in the third phase, "verification of prototype," for field testing their proposed solution in 3 rural regions. This process allows us to use the participant rural areas as testing grounds to build sustainable health care solutions that can be exported to the rest of the world. One of the main goals of the PCP process is to lead to the development of commercial volumes of end-products in the European market.

An important part of the evaluation process, selecting the best solutions and bidders throughout the PCP process, is a value-based approach. Value-based procurement is increasingly becoming part of the innovation procurement processes [9]. This procurement process must focus not only on the price of a particular product or service, but also on the overall value of the solution it can create, in terms of improved patient outcomes, reduced total cost of care, and benefits to stakeholders such as health care workers [9].

Porter and Teisberg [10] divide value-based procurement into 4 dimensions. These dimensions highlight different aspects of the procurement. At the center of the specification are the two dimensions that (1) relate to patient outcomes and (2) relate to the cost of producing them. (3) The next dimension consists of other important aspects, with more secondary relevance to patient outcome, typically benefits for the health care personnel, the hospital, and the complete health care system. (4) The last dimension relates to societal impact, and how the purchase will impact society in a larger context, an aspect that is emphasized in the European public procurement directive [10].

Aim of This Study

This study aims to explore and identify important design criteria for satisfying stakeholder needs in the PCP project, Crane. The aim of the study is 2-fold. First, to identify and prioritize value-based key performance indicators (VB-KPIs) to predict important positive and negative effects of the suggested PCP solutions. The VB-KPIs will be applied, prioritizing solutions from bidders throughout the PCP process. Second, the identification of regional differences in user needs. Other important key performance indicators (KPIs), such as evaluating technical compliance, capacity, and usability, were not part of this study.

Methods

Study Design

This study was reported according to the quality improvement reporting excellence (SQUIRE 2.0 [Standards for Quality Improvement Reporting Excellence]) publication guideline.

Context

The participating European regions in the Crane project are Västerbotten, a rural region in the northern parts of Sweden, Extremadura, a region in the central-western parts of Spain, and Agder in the south of Norway. A needs assessment in the user regions identified medical areas of importance for the Crane efforts. The selected areas were the following: diabetes, mental health, chronic obstructive pulmonary disease (COPD), and cardiac disease. In Europe, these conditions account for at least 25% of health spending [2]. Stakeholder needs for the identification of VB-KPIs were provided by patients with chronic conditions, their next of kin, and health personnel in each region. Each region recruited patient representatives and health care personnel, and participation was voluntary.

We used a framework for early health technology assessment (eHTA) to identify the VB-KPIs among the stakeholders. eHTA evaluates technologies still in development and can be defined as "a health technology assessment conducted to inform decisions about 20 subsequent developments, research and/or investment by explicitly evaluating the potential value of a conceptual or actual health technology" [11]. Thus, eHTA is a suitable prioritization methodology in the early development stages before a solution is final.

XSL•FO
RenderX

We used the decision support eHTA framework Step Up [12] to identify and structure stakeholder needs in 4 domains:

- The user benefits: positive and negative implications the innovation will have for the patient, next of kin, and employees.
- The clinical effects: clinical values and risks of implementing the innovation.
- The organizational effects: the positive and negative implications the innovation will have for the operations locally and nationally.
- The economic effects: positive and negative financial implications the innovation will have locally and nationally in the short term and long term.

A coassessment and visualization tool was used to discuss potential effects among the users and stakeholders [12]. Potential effects within the 4 domains were rated on a scale from low to medium and high for each patient group or health condition identified by the stakeholders. The larger the imprint, the higher the potential benefits for the users and stakeholders.

Intervention: Identifying VB-KPIs

Five workshops were conducted to identify VB-KPIs in the Crane project, 2 in Agder (Norway) and Extremadura (Spain) and 1 in Västerbotten (Sweden). Data collection of user needs was the focus of the workshops. The workshops were conducted digitally in Agder and Västerbotten and physically in Extremadura. Each workshop lasted approximately 2 hours and was organized with a facilitator and a secretary collecting and reporting the inputs. In the Agder workshop, with the 2 medical conditions, diabetes (10 participants) and mental health (9 participants), stakeholders were represented in 2 participating groups. The diabetes group consisted of 2 representatives of patients with diabetes, 1 diabetes coordinator, 1 physician, 1 nurse, and 5 other representatives from the health region. The mental health group consisted of 1 psychologist, 1 physician, 1 nurse, and 6 other representatives from the health region. In the Västerbotten workshop, COPD and cardiac diseases were represented. Present stakeholders were 2 nurses, 1 representative of patients with COPD, 2 representatives of patients with cardiac issues, and 2 other representatives from the health region. In Extremadura, 2 groups participated, 1 consisting of health care professionals and 1 with patient representatives. Both groups had 10 participants in total. The patient groups included 4 representatives of patients with diabetes and 4 representatives of patients with cardiac diabetes. The patient representatives were recruited by inviting various patient organizations within the selected chronic diseases in Crane. The participating patient representatives were the ones responding to the invitations. The participants in the workshops were recruited based on their connection to each chronic condition, not on the likelihood of engaging with the Crane solution in the future.

Study of the Intervention and Measures

Before and at the start of the workshop, the participants were informed of the Crane goals and vision, and informed consent was given for participation in this study. Further, a future scenario was presented to the participants in every workshop to invoke discussions on potential effects: “What if digital solutions are in place fulfilling the Crane goals and vision, and the users have sufficient digital competence using the solutions?”

Potential positive and negative effects were identified and prioritized within the 4 domains given by the eHTA framework Step Up. These prioritized effects gave inputs on user needs and a basis for defining the VB-KPIs for evaluating incoming solutions throughout the PCP process. The process is visualized in more detail in Figure 1.

XSL•FO
RenderX

Figure 1. Identifying VB-KPIs in the Crane project. eHTA: early health technology assessment; PCP: Pre-Commercial Procurement; VB-KPI: value-based key performance indicator.

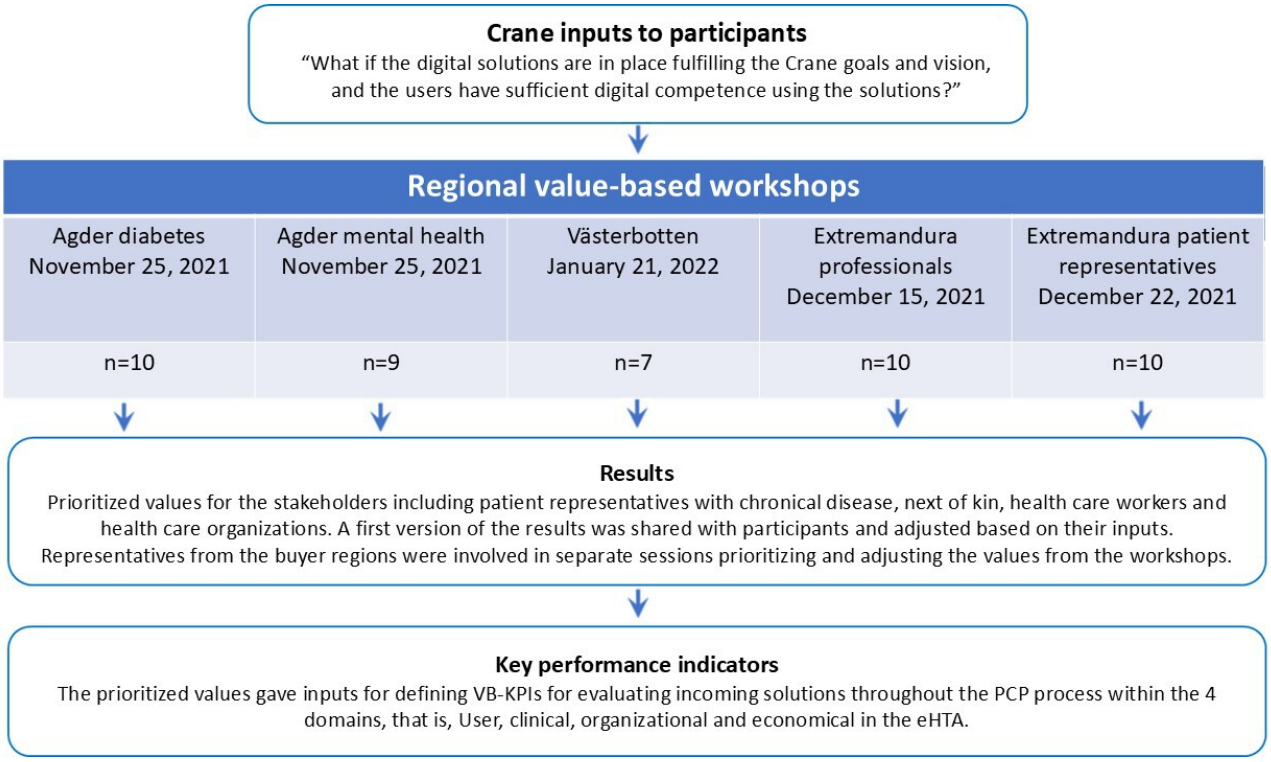

Analysis

A process of 3 iterations was used to collect, analyze, and finalize the VB-KPIs in the Crane project. (1) The user needs identified by the stakeholders in the workshops were prioritized based on the eHTA Step Up framework [12] and (2) adjusted based on comments postworkshop. (3) In separate sessions with the buyer regions, the VB-KPIs were then identified and prioritized based on stakeholder needs.

VB-KPIs were collected and structured within the 4 domains of the framework and simplified by merging patients and next of kin in the user domain. Local and national effects were merged in the economic domain. The organizational domain was simplified to positive and negative effects for the health care organization. The clinical domain was simplified to positive and negative health effects for the patient.

Ethical Considerations

This study is not subject to approval from the Regional Committees for Medical and Health Research Ethics (Norway), as it does not contain medical and health research on human beings, human biological material, or personal health data. This is not research on humans according to the definition in the Norwegian Health Research Act § 4 [13]. The study did not process special categories of personal data according to GDPR (General Data Protection Regulation), article 9, number 1 [14]. Participation in the 5 workshops was voluntary. Collected data from the workshops were based on informed consent and handled following GDPR. The legal basis for processing data is consent according to GDPR, article 6, number 1a. The informed consent form was given to the participants when invited to the workshop. The information was also given at the start of the workshops, and no reservations were given. Collected data from the 5 workshops were written down by referees in their native languages, anonymized, translated to English, and transferred in an encrypted key to the researchers at Oslo University Hospital. The recorded and transferred data were never linked to the informant, ensuring anonymity. No compensations were given to the participants.

The collected data are only personal data according to GDPR article 6. Processing of personal data followed the principles in GDPR article 5. The data were anonymized and transferred properly and securely. The specific situation of the data subject has been taken into consideration since there were representatives from patient organizations. They were invited as representatives from organizations, not as patients in a health care situation. The asymmetry is taken into consideration since the representatives are voluntary delegates in the organization, and the project is outside the health and medical care situation.

Results

Identified Needs

The workshops gave important insights into the potential positive and negative effects of implementing the Crane solution within the 4 domains.

Structuring the inputs from the workshops in all regions showed potential positive and negative effects within the following main areas: (1) increase in efficiency, (2) needs for ensuring safety, control, and trust in data sharing, (3) needs for knowledge, innovation, and research, (4) increased needs for interaction and networking, and (5) needs for considering the state of mind, attitude, and worries.

XSL•FO
RenderX

Mapping these main topics with the different domains gave overall results for all regions. When an iteration process with the stakeholders was applied to achieve consensus on the implications of Crane, the first iteration results from the workshops were shared with the stakeholders to provide inputs and comments. The result, including inputs and comments, is shown below (Tables 1 and 2). Complete tables of the results from the workshops are available in Multimedia Appendices 1 and 2.

XSL•FO
RenderX

Table . Stakeholders' needs in efficiency, safety, and knowledge versus the 4 domains (user, clinical, organizational, and economic).

Domains	Efficiency	Safety, control, and trust	Knowledge, innovation, and research
User			
Positive	• Handling of the disease • Increased flexibility • Less burden, traveling, use of medicine, hospitalizations, and consultations • Increased time for support and follow-up • Availability of information	• Better overview control, awareness, and feeling of safety • Fast response to changes and acute situations • Recognition from health care workers and relatives	• Empowerment of self-care • Increased knowledge and awareness • New type of patient contacts
Negative	• No input	• Privacy issues and loss of trust • Information overload and complexity	• Wrong use of the equipment • Technological literacy
Clinical			
Positive	• Fewer incidents, adverse events, and complications • Better follow-up and infection control • Right treatment at the right time	• Better control of the disease and the treatment • Less setbacks • Increased quality of life	• Learn from earlier actions • Potential preventive actions
Negative	• Increase in measuring parameters • Measurement failures • Cyber attacks • Technology-related problems	• Less physical contact • Technology and data focus • Too much standardization of treatment	• Lack of competencies
Organizational			
Positive	• Decrease in cancelations and work overload • Shorter treatment queues • Better allocation of resources • More consultations • Reduced demand for consultation per patient • Use of experts	• Increased quality of the health care service • Better documentation • Better routines and follow-up	• No inputs
Negative	• Increase in measurements • Implementation and scale-up challenges • Change management and transition • Interpretation of monitoring data • Increase in alarms and emergencies	• Increase in measurements • Operating the follow-up system; routines and processes • Data management and GDPR[a] compliance • Cybersecurity issues	• New training and educational activities
Economic			
Positive	• Decrease in cancelations • Fewer hospital admissions • More patients per health care worker • Easier and faster handling of patients • Less need for hospital space • Less travel costs	• Environment-friendly	• Research activity • Knowledge and training

XSL•FO
RenderX

Domains	Efficiency	Safety, control, and trust	Knowledge, innovation, and research
Negative	• Investments and funding • Cost of training patients and caregivers • Compensation challenges	• Increased need and cost of cybersecurity and GDPR issues • Increased system costs	• Increased training and education costs

[a]GDPR: General Data Protection Regulation.

Table . Stakeholders' needs in interaction and networking, and state of mind, attitude, and worries versus the 4 domains (user, clinical, organizational, and economic).

Domains	Interaction and networking	State of mind, attitude, and worries
User		
Positive	• Easier access to health care and health care workers • Maintain contact even in bad times • Less disturbance from health care workers entering at home • Create networks across country borders • Discuss how to improve the health parameters	• Increased self-esteem • More freedom • Information about the times and indicators in which you have been monitored • More exciting • New ways of working
Negative	• Still important with physical interactions between patients and health care workers • Empathic treatment must not be underestimated • There is a need for long-term relations • Less physical contact and the ability to see the complete picture • Interaction between systems	• Stress issues when looking at data • Information overload • Increased feeling of loneliness • "Burnout" • See that their values are becoming worse • Difficult to use for patients who are very sick • Negative influence from others
Clinical		
Positive	• No inputs	• Better control of the disease and the treatment • Less setbacks • Increased quality of life
Negative	• No inputs	• Less physical contact • Technology and data focus • Too much standardization of treatment
Organizational		
Positive	• Easier to work across disciplines • Better flow of data between health care actors • Closer interactions with patients	• More equal health care
Negative	• Many actors and interfaces to deal with • Third-party dependencies • Two systems: physical contact and digital interaction • Centralization: a system spanning over a larger geographical area may lose out on local and geographical context	• New ways of working • Will be confronted with ethical questions that are not seen today • Possible moral dilemmas among professionals
Economic		
Positive	• Opening for interaction between technology and people	• No inputs
Negative	• No inputs	• Possible abuse of evidence

XSL•FO
RenderX

While Agder and Västerbotten emphasized quite similar effects, Extremadura differed from the other 2 regions on the following topics:

1. Interaction and networking: none of the groups in Extremadura emphasized positive effects of interaction and networking; rather, the stakeholders in Extremadura weighted negotiations and shared decision-making between the patient and health care personnel. The groups from Agder and Västerbotten stressed the need for interaction between patients, experts, and health care organizations.
2. Privacy and GDPR compliance [14]: Agder and Västerbotten both emphasized the importance of safeguarding privacy and GDPR compliance, while this topic was not mentioned by Extremadura.
3. Hacking and cyber-attack issues were stressed by the groups in Agder and Västerbotten but not by Extremadura.
4. Physical presence for health care workers to see the overall picture and reduce the patient's feeling of loneliness. The need for still having physical meetings between patients and the health care workers was emphasized by the Agder and Västerbotten groups, but not by the Extremadura groups.

Some subtopics seemed to be important to all groups:

1. Quality of life: improved well-being and treatment quality
2. Empowerment of self-care and involvement of the patient and next of kin
3. Awareness: improved self-control
4. Preventive care: improvement in preventive care opportunities
5. Flexibility and freedom for patients, health care workers, and their organizations
6. Excitement: health care workers carrying out new health care pathways and achieving new knowledge; patients and next of kin getting better insights and control
7. Stress: health care workers receiving information overload, resulting in a new work reality; patients and next of kin feeling pressure to relate to the data output, getting alarms, and more responsibility

These are topics that should be implemented in the evaluation and selection of the solution throughout the PCP process.

Scoring of the Health Value Domains

The second iteration consisted of the rating within the 4 health value domains. The stakeholders from the workshops rated a potential value of the Crane solution, from 0 low value to 100 high value, using the eHTA framework, Step Up [12]. The results are shown in Figure 2 and Table 3. In Extremadura, only the professional group rated the potential of the future scenario.

The different regions' and groups' scoring of the domains. The scores from the health value spider are presented in the table (0, low-100, or high).

The user, economic, and organizational domains were assumed to be high in Extremadura. The clinical domain was predicted to be more moderate. In Agder, both the mental health and the diabetes groups scored the user domain as high and the economic domain as moderate. The clinical domain was assumed to be high for the patients with diabetes and low for the patients with mental health issues. In Västerbotten, all the domains were assumed to be moderate. The scoring was used as input in discussions of prioritizing domains and needs and used as input in defining VB-KPIs by the buyer regions in the third iteration.

Figure 2. The different regions and groups scored the domains using the Step Up coassessment tool.

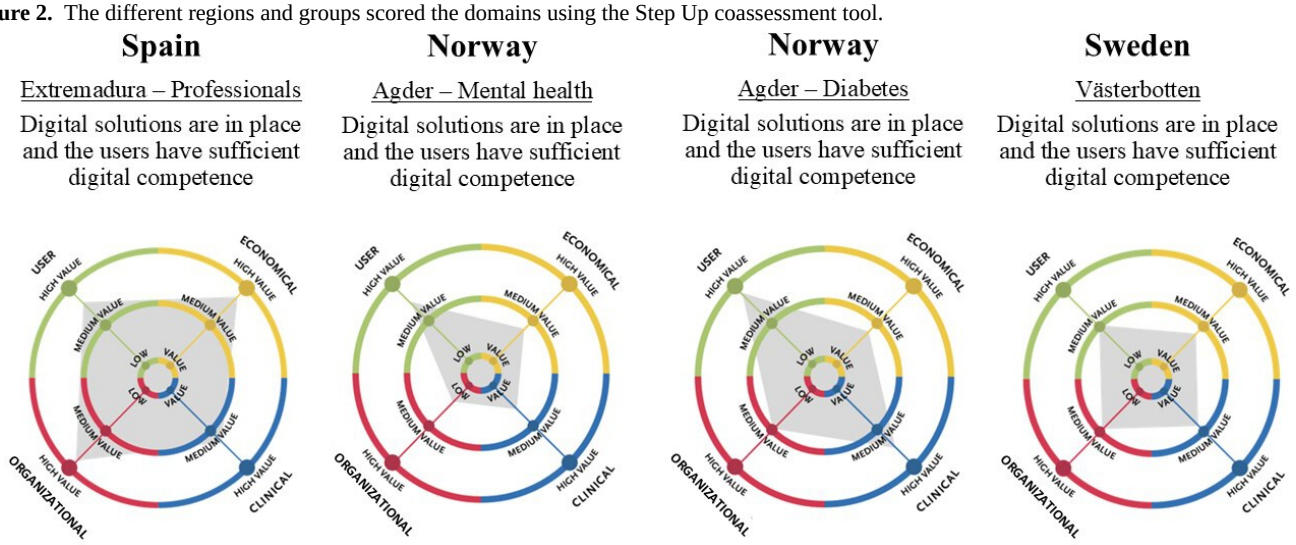

XSL•FO
RenderX

Table . The different regions and groups' scoring of the domains.

Regions	Groups	Domains			
		User	Economic	Organizational	Clinical
Extremadura	Professionals	80	88	90	66
Agder	Mental health	75	45	20	25
Agder	Diabetes	90	45	45	70
Västerbotten	Mixed	50	45	50	45

Values and KPIs

In the third and final iteration, the VB-KPIs were identified and prioritized. To actualize the identified needs in the VB-KPIs, analyses were performed by involving representatives from the buyer regions in separate sessions after the workshops. Discussions within the workshops and follow-up by professionals from the buyer regions resulted in a prioritized list of 35 potential needs (effects) to be considered in the evaluation process (Tables 4 and 5).

XSL•FO
RenderX

Table . Patient and employee–related VB-KPIs[a] were identified and prioritized by the buyer regions based on values identified and scored in the stakeholder workshops.

User domain		Values	KPIs[b] to be considered
Patients			
	1	Increased feeling of loneliness	Perceived degree of loneliness
	2	Improves the empowerment of self-care	Perceived level of self-control of the situation
	3	Relatives obtain insight into their loved one's health	Perceived relatives' insight into the patient
	4	Creates a higher level of trust in sharing data	Perceived level of trust in sharing data
	5	Creates an improved well-being	Perceived level of well-being
	6	Fast response to changes when it is an acute situation	Average response time in acute situations per patient per year
	7	Easier access to health care workers and the health care system in a hectic daily life	Perceived level of accessibility of health care workers
	8	Increased flexibility	Perceived level of flexibility
	9	Avoid stress	Perceived level of stress
Employee			
	10	Timesaving	Average (or perceived) use of time per patient per month
	11	More patients for follow-up	Average (or perceived) number of patients per month per health care worker
	12	Fast response	Average time from request to respond per month
	13	Good overview	Perceived a better overview of the patient's health condition
	14	Helps in decision-making by knowing the patient profile more accurately	Perceived ability to make better decisions for the patients
	15	Continued contact with patients	Average number of contacts per patient per month
	16	Increased flexibility	Perceived level of flexibility
	17	Increased stress	Perceived level of stress
	18	More exciting work conditions	Perceived as more exciting work conditions

[a]VB-KPI: value-based key performance indicator.

[b]KPI: key performance indicator.

XSL•FO
RenderX

Table . Economic, organizational, and clinical VB-KPIs[a] identified and prioritized by the buyer regions based on values identified and scored in the stakeholder workshops.

Domains		Values	KPIs[b] to be considered
Economic			
	19	Cost of damages and repair	Total cost of repair, that is, IT[c] and equipment per month
	20	Training and education costs, and time used for training and interpretation of results from the patients	Total cost of education and training, health care workers, and patients handling the solution
	21	Less hospitalization	Average hospitalization days per year per patient
	22	Less use of medicine	Average amount of medicine used per patient per month
	23	Less physical consultations	Average number of physical consultations with patients per month
	24	Lower travel costs for patients and health care workers	Average travel cost due to treatment and consultations for patients and health care workers
	25	Investments in the solution and support	Investment and cost calculations
Organizational			
	26	Decrease in cancelations	Average number of various types of cancelations per month
	27	Shorter treatment queues	Average time before treatment takes place per group of patients
	28	Increase in alarms and emergencies	Average number of alarms and emergencies per patient per month
	29	Better services and professional standards	Perceived quality of the service and professional standards
	30	Data management and GDPR[d] compliance	Evaluation of compliance and fulfillment of standards
	31	Better cooperation between health care services and actors	Perceived level of cooperation between the organization and other actors
Clinical			
	32	Better control of one's health and fewer setbacks	Average QALY[e], that is, EQ-5D
	33	Less focus on the complete health condition of the patient due to fewer physical contacts	Perceived change of focus, seeing the complete condition
	34	Better well-being and quality of life	Average QALY, that is, EQ-5D
	35	Fewer health risks and complications	Average number of alarms and emergencies per patient per month

[a]VB-KPI: value-based key performance indicator.

[b]KPI: key performance indicator.

[c]IT: information technology.

[d]GDPR: General Data Protection Regulation.

[e]QALY: quality-adjusted life year.

XSL•FO
RenderX

Discussion

Principal Findings

This study aimed to explore and identify key design criteria and VB-KPIs to support the development and evaluation of digital health care solutions for patients with chronic issues in rural areas within the Crane PCP process. The European rural regions of Västerbotten (Sweden), Extremadura (Spain), and Agder (Norway) are the project partners in Crane. A process of 3 iterations of stakeholder insight was used to identify and prioritize the VB-KPIs based on the eHTA framework Step Up. We identified 35 VB-KPIs across 5 domains. Despite differences in user needs, shared values were found, including empowerment, flexibility, preventative care, and improved quality of life. However, based on our findings, regional socioeconomic and cultural differences may necessitate local adaptations in the final solution in Crane.

Although we successfully identified a set of cross-national VB-KPIs to support the procurement process of Crane, these results should be applied with caution, considering the local context. One limitation to the transferring of the VB-KPIs between regions is the different prioritization of domains (Figure 2 and Table 3) and user needs identified by the participating regions in the project. Even though the 3 regions (Agder, Västerbotten, and Extremadura) had several similar characteristics and were defined as rural regions, there were geographical, cultural, and socioeconomic differences that became visible in the needs specification of the VB-KPIs. Needs that differed between Norway and Sweden compared with Spain were privacy and security, interaction between patient and health care personnel, and the need for physical interaction with patients. An analysis performed by the Cultural Group in 2024 [15] pointed to cultural differences between Spain and the northern countries. Spain scores lower on preferences for individualism, long-term orientation, and indulgence, and higher on dimensions such as power distance, motivation toward achievement and success, and uncertainty avoidance [15]. As for socioeconomic differences, the gross domestic product per capita in 2023 was US $ 90.50 in Norway, US $ 64.19 in Sweden, and US $ 46.35 in Spain [16]. Based on statistics from the European Commission from 2021, the 3 countries each spent between 10% - 11% of their gross domestic product on health care expenditures [17]. These differences will most likely influence the health care system and the patient's priorities, reactions, and willingness to adopt a service such as Crane. Thus, when planning for PCP projects crossing national borders, both cultural and socioeconomic differences should be addressed, and local adjustments to the final solution should be made accordingly.

Further Research

An important next step is the verification of the selected VB-KPIs by the stakeholders. In the next phase of the project, the tenders will be evaluated according to the identified VB-KPIs from this study. After the evaluation and prioritization of the tenders, the selected solutions will undergo field testing and piloting, and patients and health care workers will test the solutions in their local environments. The data collection from these feasibility tests will provide valid input in the final selection of the solution and a basis for regional adjustments. This makes room for continuous adjustments such that the final solution may comply with predefined and shifting user needs, that is, using an agile approach. As stated by Holbeche [18], the ability to continuously adjust and adapt (agility) is increasingly considered to be a vital success factor for organizations. This coincides with the formative evaluations of eHTA, where an innovation is evaluated while still under development.

Further research classifying KPIs for different types of PCP projects is a needed contribution to the selection of the most viable PCP solutions.

Broadening the scope by including other geographical factors, socioeconomic differences, and adjustments based on age groups in the VB-KPIs should be an area of further research.

Limitations

The outcomes from this study may not be generalized, as the VB-KPIs are retrieved from workshops and discussions with a limited number of participants and user groups. This affects the validity of this study, which may be strengthened by adjusting the needs and VB-KPIs based on continuous evaluations after each tender phase of the PCP process. Further, information on socioeconomic and geographical differences between the participating regions should be included. This may, for instance, entail variation in distance from the respective rural areas to more densely populated areas where satisfactory follow-up is available. Another limitation was that the stakeholder groups also differed in the health conditions they focused on in each region. Each workshop was also influenced by the personal interests and personalities of the stakeholders. Further, although 4 chronic disease conditions were represented, no information from other dominant chronic disorders in the regions was included. An important strength, however, is that the participants represented different European countries, thus increasing the relevance of the solution in a broader European context.

Conclusions

Although adjustments are needed to fit the specific aims of the given PCP project, the recommended VB-KPIs may represent a substantial contribution when evaluating digital PCP projects within health care. The identified and prioritized VB-KPIs provide a need-based foundation for the development and subsequent evaluation of Crane. The use of a structured framework based on eHTA provides an opportunity to verify the identified effects through economic modeling as the Crane solution develops, to ensure that user needs are met. One significant finding in the early assessment of stakeholder needs was that, although shared values were identified across nations in the rural regions, the need for local adjustments to the final solution should always be addressed.

XSL•FO
RenderX

Acknowledgments

This research has partly been funded by HORIZON 2020 (project ID 965277), and Oslo University Hospital (HORIZON participation ID 991104000). BI (Bedriftsøkonomisk Institutt) Norwegian Business School has been a collaborating partner, as well as the research and innovation center Forhelse. This study aims to explore and identify important design criteria for satisfying stakeholder needs in a digital Pre-Commercial Procurement project, Crane. We thank the patient representatives and health care workers for their insights. A special thanks to the regional Crane partners from Västerbotten in Sweden, Extremadura in Spain, and Agder in Norway for their insights and help in administering the process of data collection.

Data Availability

Data are not publicly available, but an anonymized table of data from the regions is available from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Complete output from the stakeholder workshops: efficiency, safety, control, and trust, and knowledge, innovation, and research.
[XLSX File, 18 KB - formative_v9i1e71279_app1.xlsx]

Multimedia Appendix 2
Complete output from the stakeholder workshops: interaction and networking, and state of mind, attitude, and worries.
[XLSX File, 18 KB - formative_v9i1e71279_app2.xlsx]

References

1. Jones K. The problem of an aging global population, shown by country. Visual Capitalist. 2020. URL: https://www.visualcapitalist.com/aging-global-population-problem/ [accessed 2025-07-08]
2. Introduction to healthcare systems. European Federation of Pharmaceutical Industries and Associations. URL: https://www.efpia.eu/about-medicines/use-of-medicines/healthcare-systems/introduction/ [accessed 2025-07-08]
3. Brennan P, Perola M, van Ommen GJ, Riboli E, European Cohort Consortium. Chronic disease research in Europe and the need for integrated population cohorts. Eur J Epidemiol 2017 Sep;32(9):741-749. [doi: 10.1007/s10654-017-0315-2] [Medline: 28986739]
4. Poverty and social exclusion in rural areas: final study report. : European Commission; 2008 URL: https://hdl.handle.net/11380/606205 [accessed 2025-07-08]
5. CRANE – Pre-Commercial Procurement Project. 2024. URL: https://crane-pcp.eu/ [accessed 2025-07-08]
6. Digital health. World Health Organization. URL: https://www.who.int/health-topics/digital-health#tab=tab_1 [accessed 2025-07-08]
7. Pre-Commercial Procurement. European Commission, Research and Innovation. 2024. URL: https://research-and-innovation.ec.europa.eu/strategy/support-policy-making/shaping-eu-research-and-innovation-policy/new-european-innovation-agenda/innovation-procurement/pre-commercial-procurement_en [accessed 2025-07-12]
8. European Health Data Space Regulation (EHDS). European Commission. 2024. URL: https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space_en [accessed 2025-07-08]
9. How procurement unlocks value-based health care. Boston Consulting Group (BCG). 2024. URL: https://www.bcg.com/publications/2020/procurement-unlocks-value-based-health-care [accessed 2025-07-08]
10. Value-Based Procurement (VBP): Knowledge, Guide, and Support All in the Value Chain of Medical Technology: Nordic Medtech Growth 2; 2017.
11. Grutters JPC, Bouttell J, Abrishami P, et al. Defining early health technology assessment: building consensus using Delphi technique. Int J Technol Assess Health Care 2025 Jun 2;41(1):e34. [doi: 10.1017/S0266462325100123] [Medline: 40452233]
12. Kværner KJ, Støme LN, Romm J, et al. Coassessment framework to identify person-centred unmet needs in stroke rehabilitation: a case report in Norway. BMJ Innov 2021 Jan;7(1):148-156. [doi: 10.1136/bmjinnov-2020-000432]
13. The Norwegian Health Research Act § 4. : Helse og omsorgsdepartementet (Norwegian Health Department); 2008 Jun 20 URL: https://app.uio.no/ub/ujur/oversatte-lover/data/lov-20080620-044-eng.pdf [accessed 2025-06-11]
14. General Data Protection Regulation (GDPR). GDPR.eu. 2024. URL: https://gdpr.eu/ [accessed 2025-07-08]
15. Country comparison tool. The Culture Factor Group. 2024. URL: https://www.theculturefactor.com/country-comparison-tool?countries=norway%2Cspain%2Csweden [accessed 2025-07-08]
16. GDP per capita, consumption per capita and price level indices. Eurostat. 2024. URL: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=GDP_per_capita,_consumption_per_capita_and_price_level_indices [accessed 2025-07-08]

XSL•FO
RenderX

17. Healthcare expenditure statistics – overview. Eurostat. URL: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Healthcare_expenditure_statistics_-_overview [accessed 2025-07-08]
18. Holbeche L. The Agile Organization: How to Build an Engaged, Innovative and Resilient Business: Kogan Page Limited; 2023.

Abbreviations

COPD: chronic obstructive pulmonary disease
eHTA: early health technology assessment
GDPR: General Data Protection Regulation
KPI: key performance indicator
PCP: Pre-Commercial Procurement
SQUIRE: Standards for Quality Improvement Reporting Excellence
VB-KPI: value-based key performance indicator

Edited by A Mavragani; submitted 21.01.25; peer-reviewed by H Stanislaw, P Valkonen; revised version received 13.06.25; accepted 14.06.25; published 14.08.25.

Please cite as:
Ørjasæter NO, Kværner KJ, Støme LN
Value-Based Framework for Evaluating Pre-Commercial Procurement: Case Study of Value-Based Key Performance Indicators
JMIR Form Res 2025;9:e71279
URL: https://formative.jmir.org/2025/1/e71279
doi:10.2196/71279

© Nils-Otto Ørjasæter, Kari Jorunn Kværner, Linn Nathalie Støme. Originally published in JMIR Formative Research (https://formative.jmir.org), 14.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Image-Based Dietary Assessment Using the Swedish Plate Model: Evaluation of Deep Learning–Based You Only Look Once (YOLO) Models

Gustav Chrintz-Gath[1*], MSocSc; Meena Daivadanam[2*], MBBS, MPH, PhD; Laran Matta[2*], MMedSci; Steve McKeever[1*], BSc(Eng), MSc, DPhil

[1]Department of Informatics and Media, Uppsala University, Box 513, Uppsala, Sweden

[2]Global Health and Migration Unit, Department of Women's and Children's Health, Uppsala University, Uppsala, Sweden

[*]all authors contributed equally

Corresponding Author:
Steve McKeever, BSc(Eng), MSc, DPhil
Department of Informatics and Media, Uppsala University, Box 513, Uppsala, Sweden

Abstract

Background: Recent advances in computer vision, particularly in deep learning, have significantly enhanced object recognition capabilities in images. Among these, real-time object detection frameworks such as You Only Look Once (YOLO) have shown promise across various domains. This study explores the application of YOLO-based object detection for food identification and portion estimation, with a focus on its alignment with the Swedish plate model recommended by the National Food Agency.

Objective: The primary aim of this study is to evaluate and compare the performance of 3 YOLO variants (YOLOv7, YOLOv8, and YOLOv9) in detecting individual food components and estimating their relative proportions within images, based on public health dietary guidelines.

Methods: A custom dataset comprising 3707 annotated food images spanning 42 food classes was developed for this study. A series of preprocessing and data augmentation techniques were applied to enhance dataset quality and improve model generalization. The models were evaluated using standard metrics, including precision, recall, mean average precision, and F_1-score.

Results: Among the evaluated models, YOLOv8 outperformed YOLOv7 and YOLOv9 in both peak precision and F_1-scores. It achieved a peak precision of 82.4%, compared with 73.34% for YOLOv7 and 80.11% for YOLOv9, indicating superior accuracy in both food classification and portion estimation tasks. YOLOv8 also demonstrated higher confidence in its predictions. However, all models faced challenges in distinguishing visually similar food items, underscoring the complexity of fine-grained food recognition.

Conclusions: While YOLO-based models, particularly YOLOv8, show strong potential for food and portion recognition aligned with dietary models, further refinement is needed. Improvements in model architecture and greater diversity in training data are essential before these systems can be reliably deployed in health and dietary monitoring applications.

(JMIR Form Res 2025;9:e70124) doi:10.2196/70124

KEYWORDS

food recognition; object detection models; dietary assessment; You Only Look Once; deep learning; machine learning

Introduction

The rise in type 2 diabetes presents a significant public health concern both globally and in Sweden, the context of this research [1]. Socioeconomically disadvantaged suburbs are disproportionately affected; these areas tend to be multicultural and have a high proportion of immigrants. Despite efforts to promote healthy eating behaviors, many individuals struggle to make informed dietary choices due to a range of individual socioeconomic barriers, as well as broader food environment factors. Additionally, dietary assessment approaches often rely on subjective self-reporting methods that carry a high respondent burden, leading to inaccuracies and inconsistencies in data collection and analysis [2]. This poses a challenge to accurately understanding and addressing the dietary needs and behaviors of diverse populations, thereby impeding the development of targeted interventions and policies to combat the escalating rates of type 2 diabetes and related health issues [3].

We aim to address the lack of effective tools and methodologies for accurately identifying food components within images to assess dietary patterns among multicultural populations. This gap highlights the need for innovative approaches that leverage advanced technologies, such as machine learning and computer vision, to develop robust solutions for dietary assessment and

XSL•FO
RenderX

intervention. Object detection in food images involves applying these advanced technologies to identify and localize different food items within an image. The purpose of our work was to evaluate the performance of You Only Look Once (YOLO) models in predicting the adherence of food images to the Swedish plate model. The plate is widely used as a model to convey dietary guidelines and serves as a practical proxy for assessing the healthiness of a meal. By leveraging machine learning techniques, we aim to provide insights into the effectiveness of object detection algorithms in dietary assessment and contribute to the development of innovative solutions that promote healthy eating habits.

Traditionally, food intake has been measured using food diaries, food frequency questionnaires, and other tools that, while informative, place a high burden on respondents and are prone to inaccurate reporting. We seek a less obtrusive approach to dietary self-reporting—one that is automated and relies solely on photos of food to determine whether one's diet is improving.

In recent years, there have been major breakthroughs in machine learning and computer vision, particularly in object detection and the development of deep learning techniques [4]. Object detection refers to the process of identifying and localizing objects within images or videos. This capability has diverse applications, ranging from surveillance and autonomous driving to medical imaging and dietary assessment.

Object detection has its roots in the early 1990s, when the first algorithms were developed. However, due to the limited quality of image representations at the time, no major models were constructed [5]. In 2001, the Viola-Jones Detector became the first real-time method for human face detection [6]. This was followed by the introduction of the histogram of oriented gradients detector [7], which was later superseded by the deformable part-based model [8], an extension of the histogram of oriented gradients detector. Since the 2010s, convolutional neural networks (CNNs) have formed the foundation for many successful object detection models. Among the most prominent are faster region-based CNN [9] and YOLO [10].

R-CNN is a 2-stage object detection system. In the first stage, it generates region proposals; in the second stage, it classifies these proposals and refines the bounding boxes. The regional proposal network is a component of the faster R-CNN architecture responsible for generating the region proposals. It takes the feature maps produced by the preceding convolutional layer of a CNN as input and outputs region proposals along with their objectness scores (ie, the probability that an object exists within a given region of the image). By contrast, YOLO is a real-time object detection system that is both fast and accurate. The main idea behind YOLO is to perform object detection on an image in a single forward pass through a CNN, rather than in multiple stages as done in faster R-CNN. While faster R-CNN is marginally more accurate than YOLO, it is also slower. More precisely, faster R-CNN performs better with small objects and instances involving significant overlap, owing to its more sophisticated region proposal and classification stages. Nonetheless, YOLO is highly efficient and capable of running at real-time speeds, even on low-end hardware. It is also easier to implement and has a relatively simple architecture compared with faster R-CNN. Over time, YOLO has evolved into a family of models that either supersede their predecessors or introduce new features such that the latest versions are competitive with faster R-CNN models [11]. Each member of the YOLO family is well supported, easy to install, and capable of detecting all the bounding boxes in a single pass. For these reasons, we chose to explore 3 recent YOLO models to evaluate which would be most suitable for food detection from images.

Existing work on dietary food intake assessment from images is extensive. Similar to our use case, NutriNet [12] and DeepFood [13] were deep learning–based models that offer practical solutions for automatically analyzing dietary intake from images captured using users' smartphones. Their focus is on promoting healthier eating habits and supporting informed decision-making about dietary intake. A systematic review of deep learning for food image recognition and nutrition analysis in the context of chronic disease monitoring is provided in [14]. The review covers 57 selected articles, evaluated based on their approach, employed models, datasets, experiment procedures, and results. The overarching goal was to achieve multirecognition of food items. The survey draws 2 key conclusions: first, that 5 different model types were predominantly used (mask R-CNN, R-CNN, fast R-CNN, faster R-CNN, and YOLO); second, that the field is progressing rapidly.

One framework commonly used in dietary assessment is the Swedish plate model, known locally as "Tallriksmodellen." It provides guidelines for constructing balanced meals based on the proportions of different food groups, with an emphasis on variety, portion sizes, and the relationship between food components [15]. The model divides the plate into different sections representing various food groups, promoting appropriate proportions for a well-rounded meal. Adhering to the Swedish plate model supports a balanced intake of essential nutrients, thereby promoting health and helping to prevent diet-related diseases. By leveraging advancements in computer vision, we aim to automate the analysis of food images to accurately identify food components within images and assess their adherence to the Swedish plate model.

This work was undertaken as part of the PREVENT project, which aims to cocreate, implement, and test an intervention targeting cardiometabolic diseases, such as type 2 diabetes, within the high-risk population of Region Uppsala. To evaluate changes in dietary behaviors in a "real-life" implementation trial, an innovative approach with low respondent burden was sought. The goal was to enable an assessment of broad dietary shifts over the intervention period by analyzing how the food components on participants' food plates adhered to the Swedish plate model. Our approach holds potential to support public health initiatives aimed at promoting healthier dietary choices and preventing cardiometabolic diseases such as type 2 diabetes.

The primary objective of this study is to assess the performance of 3 YOLO models in accurately identifying food components within images and their proportions in relation to the Swedish plate model. This process involves 2 main stages: data collection and the evaluation of object detection algorithms. Data collection is a crucial step in developing and training machine learning

XSL•FO
RenderX

models for object detection in food images, while evaluating the performance of object detection algorithms is essential for assessing their performance and comparing their effectiveness across different datasets and tasks.

The remainder of this paper is structured as follows. The "Methods" section introduces the YOLO framework for single-pass image classification and outlines the 3 versions used in our experiments. It also describes the data collection and analysis procedures. The "Results" section presents the data preprocessing steps, model configurations, and experimental results. Finally, the "Discussion" section discusses the broader implications of our findings in the context of the PREVENT project.

Methods

Overview

This section describes the methods used to evaluate real-time object detection models. We outline the 3 YOLO versions used, the creation of the custom dataset, and the approach taken to assess model performance.

YOLO Model for Object Detection

At the core of the YOLO architecture lies a CNN that processes the entire image in a single pass, enabling simultaneous localization and classification of objects. This unified approach eliminates the need for complex region proposal networks and postprocessing steps, thereby streamlining the object detection pipeline and supporting real-time inference.

The YOLO architecture comprises several key components, each playing a crucial role in the detection process [16]. First, the convolutional layers are responsible for feature extraction, capturing hierarchical representations of the input image. These layers leverage convolutional filters to detect meaningful patterns and features, such as edges, textures, and shapes, which are essential for object recognition. By hierarchically stacking multiple convolutional layers, the network learns increasingly abstract and discriminative features, enabling it to distinguish between different object classes and background clutter effectively. Second, the downsampling layers play a pivotal role in reducing the spatial resolution of feature maps while increasing their receptive field size. This downsampling process enables the model to capture more spatial information at multiple scales, which enhances its ability to detect objects of varying sizes and aspect ratios. Additionally, down-sampling facilitates the integration of contextual information from surrounding regions, enabling more robust and accurate object localization. Finally, the detection layers are responsible for generating the bounding box coordinates and class probabilities for detected objects. These layers typically consist of a combination of convolutional and fully connected layers, followed by specialized activation functions such as sigmoid or softmax (a mathematical function used in machine learning algorithms to convert a vector of numbers into a probability distribution). The output from the detection layers comprises a set of bounding boxes, each associated with a confidence score and corresponding class probabilities.

An essential aspect of optimizing YOLO models is the tuning of hyperparameters, which are external settings that influence both the training process and the model architecture [17]. Key hyperparameters include learning rate, batch size, number of epochs, and YOLO-specific parameters such as backbone architecture and object confidence thresholds. Proper tuning of these parameters is crucial, as it directly impacts the model's learning efficiency and overall performance. For example, an inappropriate learning rate may lead to suboptimal convergence, while unsuitable batch sizes can impact training stability. Hyperparameter optimization is therefore vital for achieving high performance in YOLO-based object detection tasks, ensuring the model functions both robustly and accurately.

Since its initial publication in 2016, successive versions of the YOLO architecture have been developed to further enhance performance. Each version features distinct architectural and network design elements, with their own characteristics and capabilities. Hence, when researchers have evaluated these models on specific datasets, newer versions have not always demonstrated superior performance [18].

YOLOv7 introduces modifications to the efficient layer aggregation network (ELAN) architecture [19,20], resulting in the extended efficient layer aggregation network (E-ELAN). Unlike ELAN, the E-ELAN architecture incorporates the principles of expand, shuffle, and merge cardinality to improve model learning while preserving gradient flow paths. In its computational blocks, E-ELAN uses group convolution to expand both channel dimensions and cardinality, applying a consistent channel multiplier and group parameter across all blocks in a layer. This strategy ensures scalability and enhances the model's adaptability across diverse datasets.

The YOLOv8 model was developed by Ultralytics, the same team behind YOLOv5. YOLOv8 introduces modifications to the cross-stage partial (CSP) modules [21]. Its feature map module combines a simpler algorithm than that used in YOLOv5 with one adapted from YOLOv7. This strategic integration enhances model efficiency while maintaining strong performance. By leveraging the strengths of both architectures, YOLOv8 achieves a refined balance between computational complexity and feature representation. These adaptations improve training dynamics, scalability, object detection accuracy, and processing speed.

YOLOv9 builds on the foundation of YOLOv5 by introducing 2 key innovations: programmable gradient information (PGI) and the generalized efficient layer aggregation network (GELAN) [22]. These design choices allow users to select appropriate computational blocks depending on the target inference device [23]. Together, PGI and GELAN address the problem of information loss in deep neural networks, ensuring high efficiency, accuracy, and adaptability. PGI helps preserve essential data across network layers, while GELAN optimizes parameter utilization and computational efficiency.

Data Collection

The selection, preparation, and preprocessing of the dataset play a pivotal role in ensuring the models' effectiveness and generalizability [24]. The goal is to create a comprehensive and

XSL•FO
RenderX

representative dataset that captures a diverse range of food items and dishes, thereby facilitating robust model training and evaluation. It was also essential to address inherent imbalances and variability within food classes.

The custom-created dataset used in this study was mainly sourced from publicly available repositories, including Roboflow Universe [25] and Kaggle [26]. It comprised 3707 images spanning 42 different food classes. These images were collected from a variety of settings and contexts—some depicting home-cooked meals, though the majority featured simpler food items. The dataset was then divided into training, validation, and test sets. The training set mainly comprised images from the public repositories.

Care was taken to include images that reflected a range of real-world conditions. These included variations in lighting, partial occlusions (eg, partially visible food items), diverse backgrounds, and inconsistent plate arrangements. Such real-world variations can impact object detection accuracy and the reliability of nutritional estimations. Nonetheless, this diversity helps simulate typical user-captured meal images and promotes better generalization by the detection models. However, we acknowledge that these variations were not formally quantified. Future work should thus incorporate a more systematic curation and analysis of environmental diversity and its effect on model performance.

The validation set consisted of anonymized images of complete meal plates, including a diverse range of multicultural foods, to better reflect real-world model usage. These images were collected through private networks by 2 bachelor students conducting thesis work within the PREVENT project. The annotation process involved manually labeling each food item in the images to create ground truth data for training and evaluation. This included drawing bounding boxes around individual items and assigning them the appropriate class labels. Annotation was performed using the Roboflow tool [27], which enabled annotation on the entire dataset and export into COCO JSON format. The resulting annotations captured essential information about the type and quantity of food items present in the images, allowing the models to learn and recognize food components.

The dataset was split into 3570 images for training, 131 for validation, and 6 for testing, covering a broad spectrum of dietary items. The test set comprised just a few images from the PREVENT group, used primarily for inference testing, allowing the model to be evaluated on individual images. Additionally, the dataset included annotation labels corresponding to the location, name, and size of food items within the images, facilitating supervised learning.

Preprocessing techniques have a noticeable impact on dataset quality, especially given the presence of noise in images and the influence of uncontrolled/unlearned environmental factors [28]. The preprocessing steps applied were *Auto-Orient*, *Resizing*, and *Auto-Adjust Contrast*. *Auto-Orient* removes EXIF-based rotations and standardizes pixel ordering. This is important, as each image contains metadata that dictate the orientation in which it is displayed relative to how the pixels are arranged on disk [29]. *Auto-Orient* eliminates any misalignment between the image and its annotations. Next, *Resizing* downsizes images to smaller, standardized dimensions, resulting in faster training. Finally, *Auto-Adjust Contrast* boosts contrast based on the image's histogram to improve normalization and line detection in varying lighting conditions. This is mainly used to enhance edge detection.

Data augmentation was used to create new training examples for the models to learn from by generating augmented versions of each image in the training set. The techniques used were *Rotation*, *Grayscale*, and *Noise*. *Rotation* introduced image variations rotated by 15°, helping the model detect objects even when images are not perfectly aligned [30]. *Grayscale* converted the image color to grayscale in 10% of the cases, adding variance to the training set. Lastly, *Noise* added random distortions to the images, helping the model become more resilient to imperfections that are typically ignored by the human eye.

Analysis

The evaluation and analysis of the YOLO models' performance in identifying food components and predicting their proportions based on the Swedish plate model is vital for assessing our methodology. The Pascal Visual Object Classes Challenge [31] has played a pivotal role in defining standard evaluation protocols and metrics for object detection tasks. It aims to provide a benchmark dataset for assessing object detection algorithms through standardized evaluation metrics. The selected evaluation metrics are precision, recall, mean average precision (mAP), and F_1-score. Together, these offered a comprehensive insight into the models' ability to detect and classify food components, an essential step for estimating proportions according to the Swedish plate model.

Precision is a metric that measures the accuracy of the model's positive predictions. It is crucial for ensuring that the model accurately identifies and classifies food items without generating an excessive number of false positives. Precision is calculated as the ratio of true-positive predictions to the total number of positive predictions, indicating the proportion of correctly identified instances among all instances predicted as positive [32]. A high precision value indicates that the model makes fewer incorrect predictions, resulting in more reliable and trustworthy outcomes [33]. By evaluating precision, we can assess the model's ability to make accurate and precise predictions regarding the presence of food items in images.

Precision=TruePostiveTruePostive + FalsePostive

Recall measures the model's ability to correctly identify all relevant instances of a particular class [32]. It is calculated as the ratio of true-positive predictions to the total number of actual positive instances in the dataset, indicating the completeness of the model's predictions:

Recall=TruePostiveTruePostive + FalsePostive

In the context of food image analysis, recall is essential to ensure that the model captures all food items present in the images, minimizing the risk of overlooking important dietary components. A high recall value indicates that the model effectively detects and identifies most food items, thereby improving its usefulness for dietary assessment. By evaluating

XSL•FO
RenderX

recall, we can assess the model's ability to comprehensively detect and classify food items, contributing to a more holistic understanding of its predictive performance.

The mAP metric is widely used for evaluating the performance of object detection models, particularly in scenarios involving multiple object classes and varying levels of difficulty [33]. It calculates the average precision for each class and then computes the mean of these values, providing a comprehensive measure of the models' overall performance.

Average Precision (AP)=1N∑k=1NP(k)

mean Average Precision (mAP)=1Q∑q=1QAP(q)

where N is the number of relevant items; Q is the number of classes; P is the precision; and K is the relevant item.

There exist multiple variants of the *mAP* metric depending on the intersection over union (IoU) thresholds. The two used in this study are *mAP@*[0.5] and *mAP@*[0.5:0.95]. *mAP@*[0.5] is calculated at a fixed IoU threshold of 0.5, and measures the model's ability to accurately localize objects. It evaluates the overlap between predicted bounding boxes and ground truth boxes, requiring at least 50% overlap for a correct detection. By contrast, *mAP@*[0.5:0.95] is computed over varying IoU thresholds, ranging from 50% to 95%. This provides insights into the model's performance under varying levels of localization strictness in bounding box overlap criteria, and offers a more comprehensive view of detection performance across different difficulty levels [32]. The mean average precision was chosen because of its ability to consider the average precision across all classes while providing a holistic measure of the models' effectiveness and robustness in food image analysis, facilitating informed decision-making and dietary recommendations.

The F_1-score is a commonly used metric in binary classification tasks, including object detection. It provides a balance between precision and recall by considering both false positives and false negatives in the model's predictions. The F_1-score is particularly valuable when class distribution is imbalanced or when there is a trade-off between precision and recall [32]. In this study, the F_1-score was used to evaluate the overall accuracy and effectiveness of the YOLO models in detecting and classifying food items within images. By combining both precision and recall into a single metric, the F_1-score offers a more comprehensive assessment of the models' predictive performance, enabling a more nuanced evaluation of their predictive capabilities.

F1 score=2×Precision×RecallPrecision+Recall

Ethical Considerations

This study involved the development and evaluation of deep learning models for dietary assessment using a custom image dataset. The dataset comprises images of food items that do not contain any identifiable personal information, human participants, or metadata linked to individuals. All images were either sourced from publicly available repositories or captured in controlled settings without the inclusion of human participants. Therefore, ethical approval and informed consent were not required, in accordance with the guidelines of the Swedish Ethical Review Authority (Etikprovningsmyndigheten), available online [34].

Results

Overview

This section presents the results of our study. We begin by outlining the model configurations, including architectural choices and hyperparameter settings. This is followed by a summary of the experiments and a detailed analysis of the performance metrics for each model.

Model Configuration

In addition to extensive data augmentation techniques, such as rotation, noise injection, and grayscale conversion, we applied transfer learning by initializing the YOLO models with weights pretrained on the COCO dataset. This large-scale image recognition dataset was designed for various tasks, including object detection, and enhances generalization for the food detection task. Training was performed using an NVIDIA 3070 GPU with 16 GB VRAM, which offers significant computational advantages by speeding up model convergence and reducing training time. Hyperparameters define key aspects of the learning process. The backbone, as described in the "Methods" section, is a critical architectural component of each YOLO model. It directly influences the model's ability to capture intricate features and patterns in the input data, thereby affecting overall detection accuracy and robustness [18]. Training epochs, which denote the number of complete passes through the dataset during training, were set to 50 for all 3 models. The batch size, that is, the number of samples processed before the internal model's parameters are updated, was set to 8 for both YOLOv7 and YOLOv9, and 16 for YOLOv8. The higher batch size for YOLOv8 was due to its more efficient memory utilization during training. Unlike YOLOv7's E-ELAN and YOLOv9's GELAN/PGI modules, YOLOv8 employs streamlined CSP modules, making it less computationally demanding per batch. This architectural efficiency allowed for a larger batch size without exceeding GPU memory limits. A larger batch size also helps stabilize gradient updates and can accelerate convergence, as observed during early training trials. Both epochs and batch size play a crucial role in determining the convergence behavior of the model during the training phase. An optimal configuration ensures sufficient iterations for learning while avoiding issues such as underfitting or excessive computational overhead. Similarly, adjusting the learning rate, weight decay, and momentum was essential for fine-tuning the model's optimization process; these parameters were configured identically across all 3 models. Finally, the choice of image size had significant implications for both computational efficiency and detection performance, and was set to 640 for all 3 models.

Performance Results

The 3 YOLO models were trained using a custom dataset consisting of images annotated with bounding boxes corresponding to various food items. The training process involved optimizing the model's parameters to minimize a composite loss function, which included components for

XSL•FO
RenderX

bounding box regression and object classification. Table 1 presents the metrics on a sample of the data before training, while Table 2 shows the same metrics after training was completed, highlighting the superior performance of YOLOv8.

Table . Lowest results at the start of training.

Model	Precision	Recall	mAP[a]@[0.5]	mAP@[0.5:95]
YOLO[b]v7	0.02079	0.4463	0.04819	0.02756
YOLOv8	0.35552	0.40618	0.3243	0.22303
YOLOv9	0.40253	0.35802	0.35684	0.22947

[a]mAP: mean average precision.
[b]YOLO: You Only Look Once.

Table . Highest results at the end of training.

Model	Precision	Recall	mAP[a]@[0.5]	mAP@[0.5:95]	F_1-score
YOLO[b]v7	0.7334	0.6863	0.6941	0.5446	0.65@0.419
YOLOv8	0.82382	0.67553	0.70447	0.53606	0.63@0.723
YOLOv9	0.80114	0.64477	0.6892	0.55459	0.64@0.480

[a]mAP: mean average precision.
[b]YOLO: You Only Look Once.

Detailed Performance Metrics

Performance Metrics Across Model Variants

For each of the 3 models, the detailed progression of performance metrics is shown in Figure 1A-C. The validation set consists of real-world food dishes, making the resulting metrics meaningful, as they reflect the models' ability to perform in practical, everyday scenarios.

Figure 1. Performance metrics overview. The x-axis represents the number of epochs, and the y-axis represents the metric value. mAP: mean average precision; YOLO: You Only Look Once.

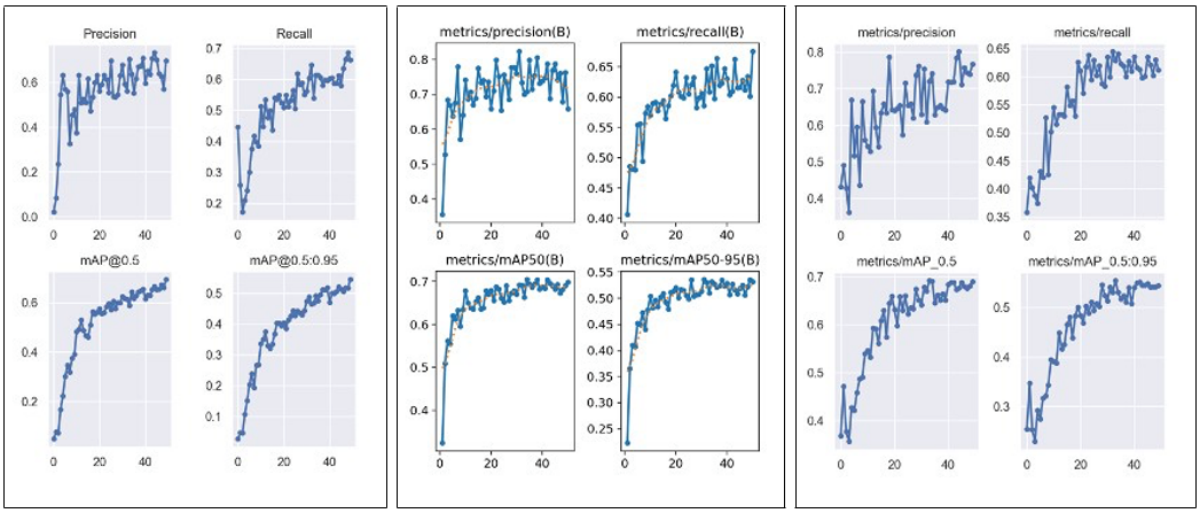

(a)YOLOv7 (b)YOLOv8 (c)YOLOv9

Precision

Precision measures the accuracy of the model's positive class predictions. For YOLOv7, it started very low at 0.02079, as expected during the early stages of training, but gradually improved over the epochs (iterations), peaking at 0.7334 by epoch 44. This indicates that 73.34% of the model's positive predictions were correct, reflecting its ability to accurately identify relevant objects in the dataset.

For YOLOv8, precision started low at 0.35552 and gradually improved over the epochs. It peaked around the midpoint of training, reaching 0.82382 at epoch 31, while the final epoch showed a precision of 0.6584. This indicates that at its best, 82.382% of the model's positive predictions were correct.

YOLOv9 started with a higher precision of 0.43127. Interestingly, this was not its lowest value—that occurred at epoch 3, where it dropped to 0.40253. Precision then steadily

XSL•FO
RenderX

increased, peaking at 0.80114 in epoch 45, and ending at 0.76679 in the final epoch. This indicates that at its best, 80.114% of the model's positive predictions were correct.

Recall

Recall evaluates the model's ability to capture all relevant instances. For YOLOv7, it started at 0.4463 and experienced fluctuations before stabilizing and reaching its highest value of 0.6863 by epoch 48. This indicates that the model correctly identified around 68.63% of all positive instances in the dataset.

YOLOv8 started with a recall of 0.40618 and steadily improved, reaching its highest value of 0.67553 at epoch 50 (the final epoch). This indicates that the model correctly identified around 67.553% of all positive instances in the dataset.

Meanwhile, YOLOv9 began with its lowest recall value of 0.35802 and then steadily increased, peaking at 0.64477 at epoch 33. By the final epoch, it reached 0.61206, indicating that the model correctly identified around 61.206% of all positive instances.

Mean Average Precision

The mAP provides insights into the model's overall detection performance. At an IoU threshold of 50%, YOLOv7's mAP steadily increased from 0.04819 to 0.6941 by epoch 50, indicating consistent improvement in object detection accuracy. Similarly, the mAP for IoU thresholds ranging from 0.5 to 0.95 rose from 0.02756 to 0.5446 over the training epochs, reflecting the model's ability to maintain precision across a broader range of IoU criteria.

At an IoU threshold of 50%, YOLOv8's mAP steadily increased from 0.3243 in the first epoch, reaching its peak of 0.70447 by epoch 38. Similar to the precision value, it peaked slightly past the midpoint of training and ended at 0.69816 by epoch 50. Similarly, the mAP at IoU thresholds ranging from 0.5 to 0.95 rose from 0.22303 in the first epoch, peaked at 0.53606 in epoch 31, and settled at 0.53174 in the final epoch, demonstrating the model's ability to maintain strong detection performance across varying IoU levels.

For YOLOv9, the mAP at an IoU threshold of 50% increased steadily throughout training, reaching its highest value of 0.6892 in the final epoch. For the mAP at IoU thresholds ranging from 0.5 to 0.95, performance improved consistently until peaking at 0.55459 in epoch 34, before ending at 0.54468 by epoch 50. These results reflect the model's ability to maintain precision across varying levels of detection difficulty.

F_1-Score

The F_1-scores, shown in Figure 2A-C, reflect the harmonic mean of precision and recall. For YOLOv7, the F_1-score peaked at 0.65 at a confidence level of 0.419. This suggests a balanced performance between precision and recall, demonstrating the YOLOv7 model's effectiveness in both identifying relevant objects and minimizing false positives. For YOLOv8, the F_1-score reached its peak of 0.63 at a confidence level of 0.723, indicating a slightly better performance between precision and recall at higher confidence thresholds. Finally, the F_1-score for YOLOv9 reached its peak of 0.64 at a confidence level of 0.480, similar to YOLOv7, although the curves appear quite different. The tailing-off pattern observed in Figure 2A for YOLOv7's F_1-score may be indicative of overfitting during the final epochs of training. Unlike YOLOv8 and YOLOv9, which maintain more stable F_1-scores, YOLOv7 exhibits a decline after its peak, suggesting that the model may have begun to memorize training patterns rather than generalize effectively. This behavior could also be attributed to its architectural configuration, where the extended E-ELAN modules may have increased model complexity relative to the training data size, making it more prone to such fluctuations.

Figure 2. F_1-scores. The x-axis represents the confidence score, and the y-axis represents the F_1-score. YOLO: You Only Look Once.

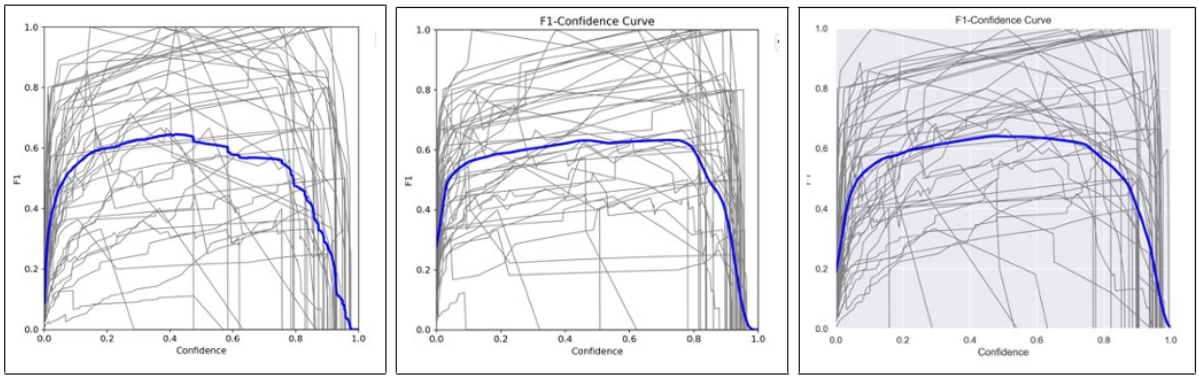

(a)YOLOv7 (b)YOLOv8 (c)YOLOv9

Confusion Matrix

A confusion matrix visually represents the model's performance across different classes using the validation set. Each box represents the model's prediction for a specific class, with predicted classes displayed along the y-axis and actual classes along the x-axis. A strong diagonal trend reflects accurate predictions, where the predicted class matches the true class. The darker the color, the greater the number of times that class was predicted, relative to the actual frequency of that class in the dataset. This visualization offers insights into the model's strengths and weaknesses in classifying different objects. For YOLOv7, as shown in Figure 3, most classes are accurately predicted according to their true labels. However, there are

XSL•FO
RenderX

classes—such as mushroom and radish—where the model misclassified them as background. These misclassifications may stem from a combination of factors. One plausible reason is the one-object-per-grid-cell constraint inherent in YOLO architectures, which may lead to the suppression of smaller or overlapping objects, particularly when they are clustered with more dominant items on the plate. Additionally, the underrepresentation of these classes in the training dataset likely contributed to reduced recognition accuracy. When certain food categories appear infrequently, the model struggles to learn their distinct visual features, increasing the likelihood of ignoring them entirely during prediction.

For YOLOv8, shown in Figure 4, most classes are accurately predicted based on their true labels. However, the predictions show slightly more dispersion compared with YOLOv7's confusion matrix. Notably, classes such as radish and shrimp were occasionally predicted as background. Such misclassifications typically occur when the model fails to detect a component that is actually present in the image.

Figure 5 shows the matrix for YOLOv9. Most classes are accurately predicted based on their labels, although there are more incorrect predictions than for either YOLOv7 or YOLOv8, where the model missed components in the image and therefore categorized them as background. There are no distinct classes where the model missed all of the instances, such as in the previous models, where both had 2 classes each that were wrong.

Figure 3. YOLOv7 confusion matrix. Darker colors indicate a higher number of correct predictions by the food plate algorithm. Squares off the diagonal represent misclassifications. Background false negative (FN) indicates that the model fails to detect a true object that is present in the image, leaving it as part of the background. Background false positive (FP) indicates that the model incorrectly predicts an object when there is none (ie, it finds an object in the background). YOLO: You Only Look Once.

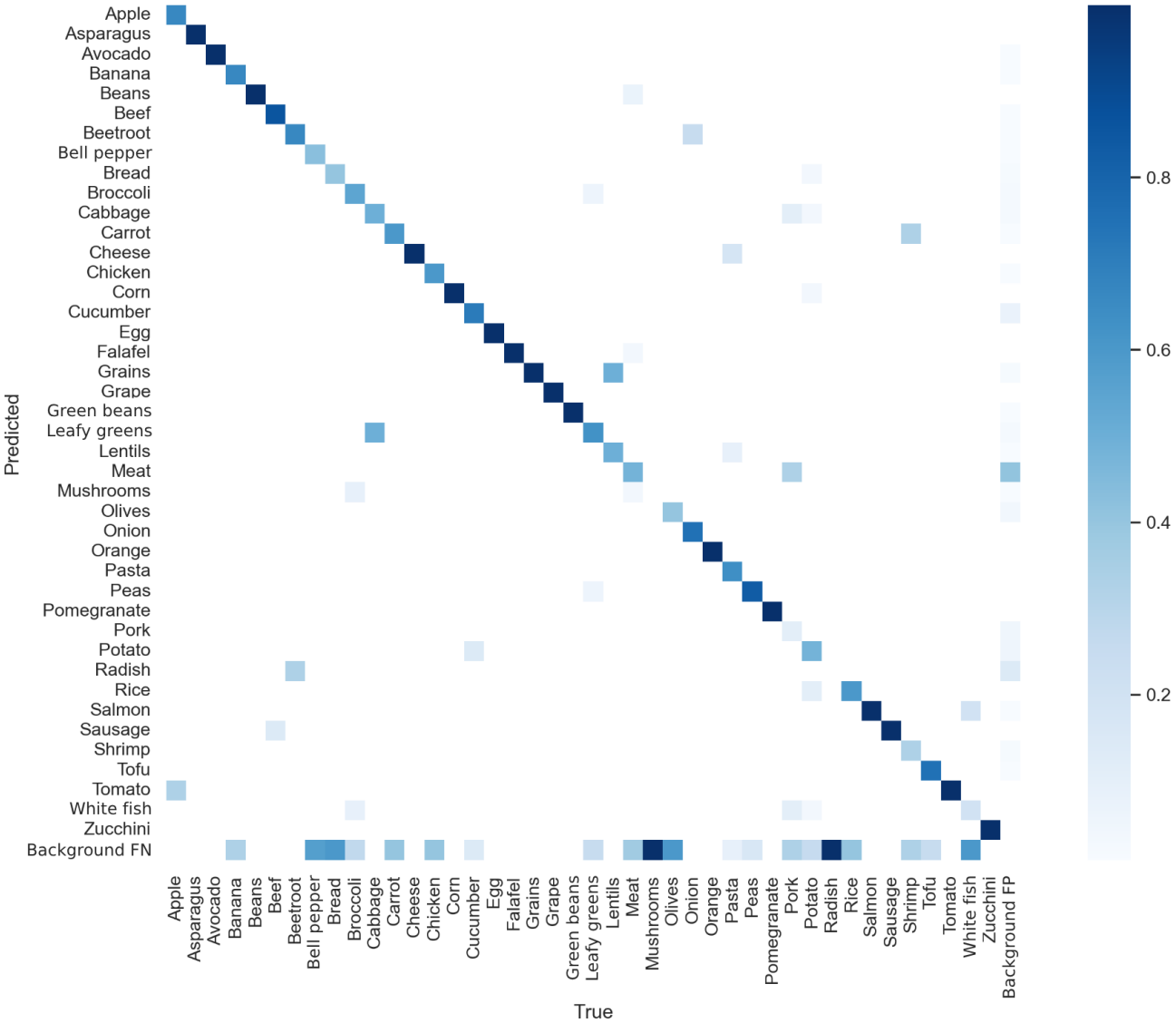

XSL•FO
RenderX

Figure 4. YOLOv8 confusion matrix showing that most classes are accurately predicted. YOLO: You Only Look Once.

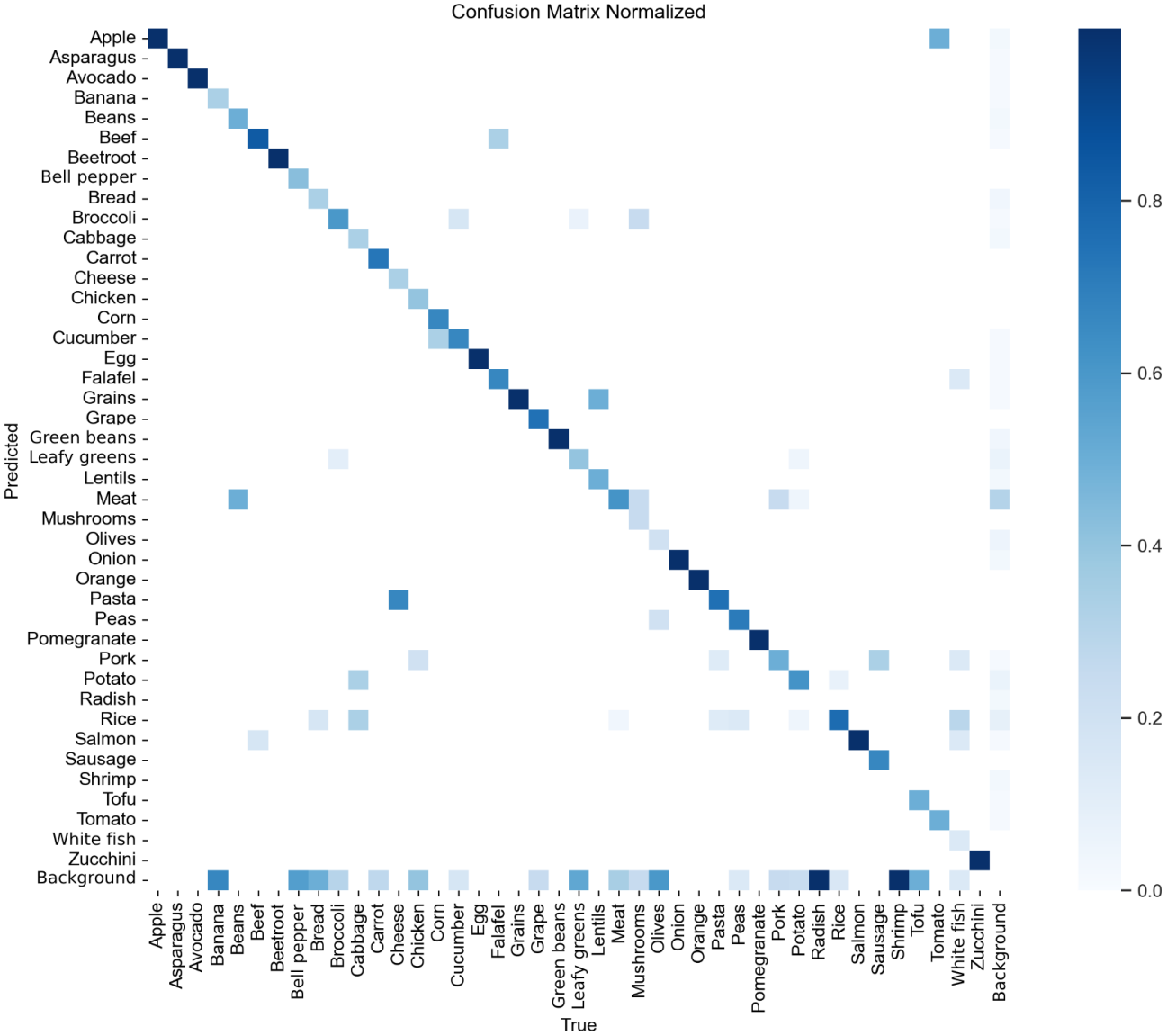

XSL•FO
RenderX

Figure 5. YOLOv9 confusion matrix demonstrating good predictability, but less than YOLOv7 and YOLOv8. YOLO: You Only Look Once.

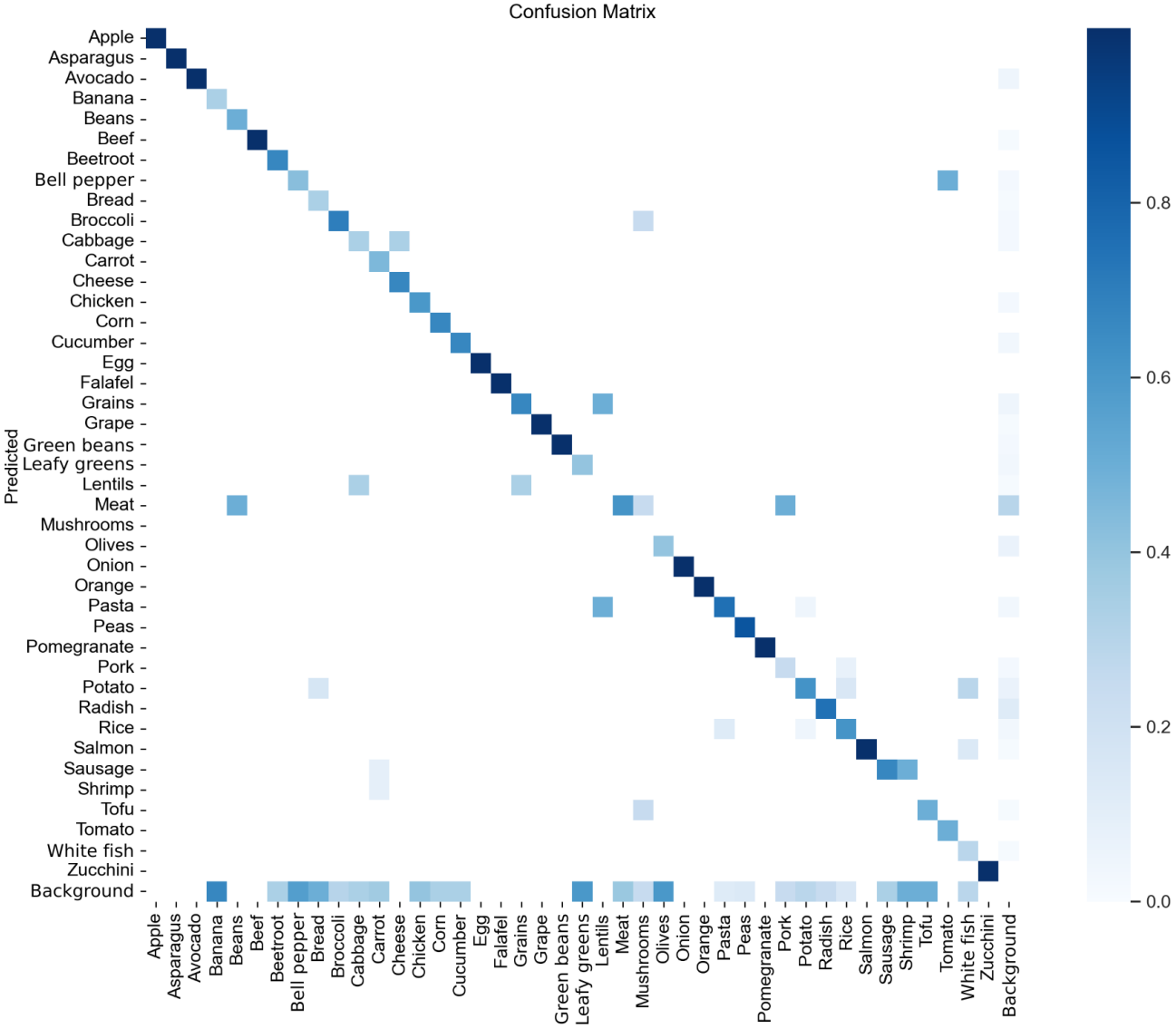

Summary of YOLO Models

Based on the results from the performance metrics—precision, recall, mAP, and F_1-score—all 3 YOLO models (YOLOv7, YOLOv8, and YOLOv9) demonstrated competitive performance in identifying and classifying food components. However, YOLOv8 emerges as the most suitable model overall. With a peak precision of 82.382%, YOLOv8 exhibits high accuracy in identifying positive instances, which is essential for classifying food components within images and estimating their proportions according to the Swedish plate model. While YOLOv7 exhibits the highest recall rate, YOLOv8 closely follows, indicating its effectiveness in capturing relevant objects. Additionally, YOLOv8 surpasses YOLOv7 in mAP, achieving a peak value of 0.70447 at the 50% IoU threshold, highlighting its capability to localize objects with high confidence. Although its F_1-score is slightly lower compared with those of YOLOv7 and YOLOv9, YOLOv8's higher optimal confidence level (0.723) suggests stronger certainty in its predictions, an advantage in practical applications that require reliable food classification.

As the models were trained using similar hyperparameters and the same dataset, the observed performance differences can be attributed primarily to architectural innovations. YOLOv8 employs a modified CSPDarknet backbone, which enhances feature representation while maintaining computational efficiency. This architecture supports improved gradient flow and better preservation of spatial information, both critical for distinguishing visually similar food items, such as lentils and grains. By contrast, YOLOv7's E-ELAN and YOLOv9's GELAN/PGI, although powerful, may introduce added complexity or optimization challenges that slightly hinder detection precision in this domain. YOLOv8's combination of a lightweight yet expressive architecture and more stable learning dynamics likely contributes to its superior performance in food detection tasks, which demand both fine-grained object recognition and real-time speed. Future ablation studies could help isolate the contributions of specific architectural elements, such as CSP modules and feature pyramid designs, to better understand their roles in this performance gain.

XSL•FO
RenderX

Discussion

Principal Findings

This study evaluated the application of 3 YOLO-based object detection models for automated food recognition and portion estimation aligned with the Swedish plate model. Among the models, YOLOv8 demonstrated superior performance across several key evaluation metrics, achieving the highest precision and exhibiting greater confidence in its predictions compared with YOLOv7 and YOLOv9. These findings underscore the potential of modern deep learning architectures to accurately identify food items and estimate their proportions in plated meals, highlighting their promise for automated dietary assessment.

Dietary assessment plays a crucial role in public health, particularly in addressing the growing prevalence of type 2 diabetes [35]. In Sweden, as in many other countries, the incidence of type 2 diabetes and other cardiometabolic diseases has risen markedly, prompting efforts to develop effective strategies for prevention and management. Supporting individuals in adopting healthier behaviors and promoting active lifestyles is essential to curbing this trend. Enabling people at risk to easily monitor their food intake, such as through photo-based analysis, is an appealing option. This study evaluated the effectiveness of YOLO models in identifying food components within images and estimating their proportions according to the Swedish plate model, which offers a simple yet practical framework for encouraging balanced and healthy eating habits. The model divides the plate into sections representing various food groups, emphasizing appropriate portion sizes and proportions to support a well-balanced meal. By following this model, individuals can achieve a nutrient-rich diet that promotes overall health and helps prevent diet-related conditions. The performance metrics used in this study offered valuable insights into each model's strengths and limitations, enabling a clear comparison of their ability to identify food components. Among the 3, the YOLOv8 model demonstrated the best overall performance and is therefore recommended for implementation in PREVENT's pilot program.

Comparisons With Prior Work

Our research aimed to investigate the potential of using object detection models to identify food components within images and estimate their proportions according to the Swedish plate model. This approach distinguishes our work from earlier contributions [12,13], as we trained our models specifically on this healthy meal planning template rather than a general food recognition dataset.

Limitations

One limitation of our study was the restricted size of the food plate library folder, which resulted in many food items being underrepresented. This gap is reflected in the confusion matrices, which reveal uncertainties in food identification, indicating insufficient training data. Some foods, such as oranges, onions, and grapes, were identified with high accuracy. By contrast, lentils were often misclassified as pasta or grains, possibly because these categories were better represented in the training data or because their visual features are more distinct, leading the food plate algorithm to favor them. Another challenge was the identification of foods with similar appearance. For instance, red paprika was frequently misidentified as tomato, suggesting the absence of a dedicated class for (red) paprika. These findings highlight the need for expanding and refining the dataset to ensure the models are robust enough for accurate dietary data collection [36].

Another important limitation concerns the model's performance under varying real-world conditions, such as different lighting scenarios, partial occlusions, and camera angles. While the validation set includes some images taken in uncontrolled environments, such as home kitchens and multicultural meal settings, these conditions were not systematically varied, tested, or curated.

It is also crucial to clarify that an object detection model does not directly predict the healthiness of food based on the Swedish plate model. Rather, by identifying food components and their bounding boxes, the model can estimate the relative area each food occupies. This information can then be used to analyze the proportions of different food groups on the plate. Based on this analysis, a healthiness score can be generated—for example, using a Likert scale from 1 to 10—allowing users to compare meals and track their dietary progress. However, assessing a plate is more nuanced than simply identifying its components and their respective proportions. Factors such as the content of salt or sugar, the type and amount of fat, and the cooking method also play important roles in determining a meal's healthiness—yet these are difficult to evaluate through photographic analysis alone. Nevertheless, identifying the components of a meal from a photo can serve as a useful proxy for dietary assessment from a health perspective.

Conclusions

The broader societal relevance of this study lies in its potential to enhance dietary assessment and health promotion initiatives. Applying machine learning techniques such as YOLO could enable more accurate and efficient methods for monitoring and improving dietary habits, thereby contributing to the prevention of diet-related diseases and the promotion of healthier lifestyles. By leveraging the principles of the Swedish plate model, this study extends the application of object detection toward the development of actionable and practical tools that support informed dietary choices. Future work should explore alternative object detection models beyond the YOLO series, such as faster R-CNN and cascade R-CNN [37]. Additionally, creating more comprehensive and diverse datasets is essential to improve the accuracy and robustness of these models, thereby expanding their potential utility.

XSL•FO
RenderX

Acknowledgments

We acknowledge funding support from Diabetesfonden (grant DIA2022-693). We thank food, nutrition, and dietetics course students Zebastian Lindblom Gaurwitsch, and Roz Daham for annotating the food images using the Roboflow tool and for assessing the YOLOv9 model. The authors are grateful to the anonymous reviewers for their insightful, constructive, and valuable suggestions.

Data Availability

The dataset used in this study is experimental and currently under internal development. Owing to its preliminary nature and ongoing validation, it is not yet publicly available. We plan to release the dataset once quality assurance processes are completed and appropriate documentation is in place. Researchers interested in early access are encouraged to contact the corresponding author (SM) to explore potential collaborations or data-sharing agreements.

Authors' Contributions

All authors contributed to model development, dataset evaluation, literature research, drafting and discussing the manuscript, and critically revising the work.

Conflicts of Interest

None declared.

References

1. Bennet L. Hög diabetesrisk bland icke-västerländska invandrare. Läkartidningen. 2018. URL: https://lakartidningen.se/klinik-och-vetenskap-1/artiklar-1/temaartikel/2018/02/hog-diabetesrisk-bland-icke-vasterlandska-invandrare/ [accessed 2025-07-28]
2. Bailey RL. Overview of dietary assessment methods for measuring intakes of foods, beverages, and dietary supplements in research studies. Curr Opin Biotechnol 2021 Aug;70:91-96. [doi: 10.1016/j.copbio.2021.02.007]
3. Bountziouka V, Panagiotakos DB. Statistical methods used for the evaluation of reliability and validity of nutrition assessment tools used in medical research. Curr Pharm Des 2010;16(34):3770-3675. [doi: 10.2174/138161210794455102] [Medline: 21128895]
4. Murthy CB, Hashmi MF, Bokde ND, Geem ZW. Investigations of object detection in images/videos using various deep learning techniques and embedded platforms—a comprehensive review. Appl Sci (Basel) 2020;10(9):3280. [doi: 10.3390/app10093280]
5. Zou Z, Chen K, Shi Z, Guo Y, Ye J. Object detection in 20 years: a survey. Proc IEEE 2023;111(3):257-276. [doi: 10.1109/JPROC.2023.3238524]
6. Viola P, Jones M. Rapid object detection using a boosted cascade of simple features. 2001 Presented at: 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR; Dec 8-14, 2001; Kauai, HI. [doi: 10.1109/CVPR.2001.990517]
7. Dalal N, Triggs B. Histograms of oriented gradients for human detection. 2005 Presented at: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05); Jun 20-25, 2005; San Diego, CA p. 886-893. [doi: 10.1109/CVPR.2005.177]
8. Felzenszwalb P, McAllester D, Ramanan D. A discriminatively trained, multiscale, deformable part model. 2008 Presented at: IEEE Conference on Computer Vision and Pattern Recognition; Jun 23-28, 2008; Anchorage, AK p. 1-8.
9. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 2017;39(6):1137-1149. [doi: 10.1109/TPAMI.2016.2577031]
10. Redmon J, Divvala S, Girshick R, Farhadi A. You Only Look Once: unified, real-time object detection. 2016 Presented at: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Jun 27-30, 2016; Las Vegas, NV p. 779-788. [doi: 10.1109/CVPR.2016.91]
11. Terven J, Córdova-Esparza DM, Romero-González JA. A comprehensive review of YOLO architectures in computer vision: from YOLOv1 to YOLOv8 and YOLO-NAS. Mach Learn Knowl Extr 2023;5(4):1680-1716. [doi: 10.3390/make5040083]
12. Mezgec S, Koroušić Seljak B. NutriNet: a deep learning food and drink image recognition system for dietary assessment. Nutrients 2017 Jun 27;9(7):657. [doi: 10.3390/nu9070657] [Medline: 28653995]
13. Liu C, Cao Y, Luo Y, Chen G, Vokkarane V, Ma Y. Deepfood: deep learning based food image recognition for computer-aided dietary assessment. In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Springer; 2016, Vol. 9677:37-48. [doi: 10.1007/978-3-319-39601-9_4]
14. Mansouri M, Benabdellah Chaouni S, Jai Andaloussi S, Ouchetto O. Deep learning for food image recognition and nutrition analysis towards chronic diseases monitoring: a systematic review. SN Comput Sci 2023;4(5). [doi: 10.1007/s42979-023-01972-1]

XSL•FO
RenderX

15. Tallriksmodellen. Livsmedelsverket. 2024. URL: https://www.livsmedelsverket.se/matvanor-halsa--miljo/kostrad/tallriksmodellen [accessed 2025-07-28]
16. Wong A, Famuori M, Shafiee MJ, Li F, Chwyl B, Chung J. YOLO nano: a highly compact you only look once convolutional neural network for object detection. 2019 Presented at: 2019 Fifth Workshop on Energy Efficient Machine Learning and Cognitive Computing - NeurIPS Edition (EMC2-NIPS); Dec 13, 2019; Vancouver, BC, Canada p. 22-25. [doi: 10.1109/EMC2-NIPS53020.2019.00013]
17. Wahyudi D, Soesanti I, Nugroho HA. Optimizing hyperparameters of yolo to improve performance of brain tumor detection in MRI images. 2023 Presented at: 2023 6th International Conference on Information and Communications Technology (ICOIACT); Nov 10-11, 2023; Yogyakarta, Indonesia p. 413-418. [doi: 10.1109/ICOIACT59844.2023.10455813]
18. Gašparović B, Mauša G, Rukavina J, Lerga J. Evaluating YOLOv5, YOLOv6, YOLOv7, and YOLOv8 in underwater environment: is there real improvement? 2023 Presented at: 8th International Conference on Smart and Sustainable Technologies (SpliTech); Jun 20-23, 2023; Split and Bol (Island of Brac), Croatia p. 1-4. [doi: 10.23919/SpliTech58164.2023.10193505]
19. Liu K, Sun Q, Sun D, Peng L, Yang M, Wang N. Underwater target detection based on improved YOLOv7. JMSE 2023;11(3):677. [doi: 10.3390/jmse11030677]
20. Samma H, Al-Azani S, Luqman H, Alfarraj M. Contrastive-based YOLOv7 for personal protective equipment detection. Neural Comput Applic 2024 Feb;36(5):2445-2457. [doi: 10.1007/s00521-023-09212-6]
21. Ang GJN, Goil AK, Chan H, et al. A novel application for real-time arrhythmia detection using YOLOv8. arXiv. Preprint posted online on January 8, 2023. [doi: 10.48550/arXiv.2305.16727]
22. Acharya A. YOLOv9: sota object detection model explained. Encord. 2024. URL: https://encord.com/blog/yolov9-sota-machine-learning-object-dection-model/ [accessed 2024-04-25]
23. Wang CY, Yeh IH, Liao HM. YOLOv9: learning what you want to learn using programmable gradient information. In: Computer Vision (LNCS ECCV) 2024, Vol. 15089:1-21. [doi: 10.1007/978-3-031-72751-1_1]
24. Zhao W, Bhushan A, Santamaria AD, Simon MG, Davis CE. Machine learning: a crucial tool for sensor design. Algorithms 2008 Dec 1;1(2):130-152. [doi: 10.3390/a1020130] [Medline: 20191110]
25. Roboflow, Inc. 2024. URL: https://universe.roboflow.com [accessed 2024-10-29]
26. Kaggle. 2024. URL: https://www.kaggle.com [accessed 2024-10-29]
27. Roboflow Docs. Use roboflow annotate. 2024. URL: https://docs.roboflow.com/annotate/use-roboflow-annotate [accessed 2024-05-02]
28. Wang TS, Kim GT, Kim M, Jang J. Contrast enhancement-based preprocessing process to improve deep learning object task performance and results. Appl Sci (Basel) 2023;13(19):10760. [doi: 10.3390/app131910760]
29. Roboflow Docs. Preprocess images. 2023. URL: https://docs.roboflow.com/datasets/image-preprocessing [accessed 2024-05-10]
30. Roboflow docs. Generate augmented images. 2019. URL: https://docs.roboflow.com/datasets/image-augmentation [accessed 2024-05-10]
31. Everingham M, Van Gool L, Williams CKI, Winn J, Zisserman A. The Pascal Visual Object Classes (VOC) Challenge. Int J Comput Vis 2010 Jun;88(2):303-338. [doi: 10.1007/s11263-009-0275-4]
32. Jocher G, Munawar R, Vina A. YOLO performance metrics. Ultralytics YOLO Docs. 2023. URL: https://docs.ultralytics.com/guides/yolo-performance-metrics [accessed 2024-03-02]
33. Padilla R, Netto SL, da Silva EAB. A survey on performance metrics for object-detection algorithms. 2020 Presented at: 2020 International Conference on Systems, Signals and Image Processing (IWSSIP); Jul 1-3, 2020; Niterói, Brazil p. 237-242. [doi: 10.1109/IWSSIP48289.2020.9145130]
34. Värnar människani forskning [protecting the individual in research]. Etikprövningsmyndigheten. URL: https://etikprovningsmyndigheten.se [accessed 2025-04-22]
35. Lazarou C, Panagiotakos D, Matalas AL. The role of diet in prevention and management of type 2 diabetes: implications for public health. Crit Rev Food Sci Nutr 2012;52(5):382-389. [doi: 10.1080/10408398.2010.500258] [Medline: 22369258]
36. Gaurwitsch ZL, Daham R. Identifiering av måltidskomponenter med hjälp av ett ai-verktyg: en valideringsstudie (bachelor's thesis, uppsala university, department of food studies, nutrition and dietetics). 2024 URL: https://www.uppsatser.se/uppsats/e15011d6d7/ [accessed 2025-07-28]
37. Shah B, Bhavsar H. Depth-restricted convolutional neural network—a model for Gujarati food image classification. Vis Comput 2024 Mar;40(3):1931-1946. [doi: 10.1007/s00371-023-02893-z]

Abbreviations

CNN: convolutional neural network
CSP: cross-stage partial
E-ELAN: extended efficient layer aggregation network
ELAN: efficient layer aggregation network
GELAN: generalized efficient layer aggregation network

XSL•FO
RenderX

IoU: intersection over union
mAP: mean average precision
PGI: programmable gradient information
R-CNN: regions with CNN
YOLO: You Only Look Once

Edited by A Mavragani; submitted 16.12.24; peer-reviewed by AA Khan, FM Ruhad, MI Idris; revised version received 17.06.25; accepted 18.06.25; published 14.08.25.

Please cite as:
Chrintz-Gath G, Daivadanam M, Matta L, McKeever S
Image-Based Dietary Assessment Using the Swedish Plate Model: Evaluation of Deep Learning–Based You Only Look Once (YOLO) Models
JMIR Form Res 2025;9:e70124
URL: https://formative.jmir.org/2025/1/e70124
doi:10.2196/70124

© Gustav Chrintz-Gath, Meena Daivadanam, Laran Matta, Steve McKeever. Originally published in JMIR Formative Research (https://formative.jmir.org), 14.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Using Artificial Intelligence ChatGPT to Access Medical Information About Chemical Eye Injuries: Comparative Study

Layan Yousef Alharbi[1*], MBBS; Rema Rashed Alrashoud[1*], MBBS; Bader Shabib Alotaibi[1], MBBS; Abdulaziz Meshal Al Dera[1], MBBS; Raghad Saleh Alajlan[1], MBBS; Reem Rashed AlHuthail[2], MD; Dalal Ibrahim Alessa[2], MD

[1]College of Medicine, Imam Mohammad Ibn Saud Islamic University (IMSIU), Prince Mohammed Ibn Salman Ibn Abdulaziz Road, Riyadh, Saudi Arabia

[2]Department of Ophthalmology, College Of Medicine, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, Saudi Arabia

[*]these authors contributed equally

Corresponding Author:
Layan Yousef Alharbi, MBBS
College of Medicine, Imam Mohammad Ibn Saud Islamic University (IMSIU), Prince Mohammed Ibn Salman Ibn Abdulaziz Road, Riyadh, Saudi Arabia

Abstract

Background: Chemical ocular injuries are a major public health issue. They cause eye damage from harmful chemicals and can lead to severe vision loss or blindness if not treated promptly and effectively. Although medical knowledge has advanced, accessing reliable and understandable information on these injuries remains a challenge. This is due to unverified web-based content and complex terminology. Artificial intelligence tools like ChatGPT provide a promising solution by simplifying medical information and making it more accessible to the general public.

Objective: This study aims to assess the use of ChatGPT in providing reliable, accurate, and accessible medical information on chemical ocular injuries. It evaluates the correctness, thematic accuracy, and coherence of ChatGPT's responses compared with established medical guidelines and explores its potential for patient education.

Methods: A total of 9 questions were entered into ChatGPT regarding various aspects of chemical ocular injuries. These included the definition, prevalence, etiology, prevention, symptoms, diagnosis, treatment, follow-up, and complications. The responses provided by ChatGPT were compared with the *International Classification of Diseases-9* and *International Classification of Diseases-10* guidelines for chemical (alkali and acid) injuries of the conjunctiva and cornea. The evaluation focused on criteria such as correctness, thematic accuracy, and coherence to assess the accuracy of ChatGPT's responses. The inputs were categorized into 3 distinct groups, and statistical analyses, including Flesch–Kincaid readability tests, ANOVA, and trend analysis, were conducted to assess their readability, complexity, and trends.

Results: The results showed that ChatGPT provided accurate and coherent responses for most questions about chemical ocular injuries, demonstrating thematic relevance. However, the responses sometimes overlooked critical clinical details or guideline-specific elements, such as emphasizing the urgency of care, using precise classification systems, and addressing detailed diagnostic or management protocols. While the answers were generally valid, they occasionally included less relevant or overly generalized information. This reduced their consistency with established medical guidelines. The average Flesch Reading Ease Score was 33.84 (SD 2.97), indicating a fairly challenging reading level, while the Flesch–Kincaid Grade Level averaged 14.21 (SD 0.97), suitable for readers with college-level proficiency. The passive voice was used in 7.22% (SD 5.60%) of sentences, indicating moderate reliance. Statistical analysis showed no significant differences in the Flesch Reading Ease Score (P=.38), Flesch–Kincaid Grade Level (P=.55), or passive sentence use (P=.60) across categories, as determined by one-way ANOVA. Readability remained relatively constant across the 3 categories, as determined by trend analysis.

Conclusions: ChatGPT shows strong potential in providing accurate and relevant information about chemical ocular injuries. However, its language complexity may prevent accessibility for individuals with lower health literacy and sometimes miss critical aspects. Future improvements should focus on enhancing readability, increasing context-specific accuracy, and tailoring responses to a person's needs and literacy levels.

(JMIR Form Res 2025;9:e73642) doi:10.2196/73642

KEYWORDS

artificial intelligence; ChatGPT; chemical eye injuries; ophthalmology; medical information; patient education; ICD-9; ICD-10; readability

XSL•FO
RenderX

Introduction

The internet has become a significant source of health-related information. In the early 2000s, 4.5% of global internet searches were health-related. However, much of the web-based medical information is often inappropriate or even harmful due to unverified content and insufficient internet regulations. Additionally, even when the information is accurate, it may be presented in language too complex for the general public, making it inaccessible [1-3]. Artificial intelligence (AI), a field of computer science, focuses on creating systems capable of tasks traditionally requiring human intelligence, such as reasoning, learning, problem-solving, perception, and language comprehension [4]. In November 2022, OpenAI introduced "ChatGPT," a large language model trained on extensive multilingual text datasets, enabling it to generate human-like responses to textual inputs. Chatbots, as AI-powered apps, are designed to simulate human conversations effectively through text or voice interactions [5].

With its extensive knowledge base and complex parameters, it allows users to ask questions and receive answers that demonstrate advanced, human-like reasoning [6]. Since its release, ChatGPT has gained significant popularity and is rapidly becoming a common tool for individuals to seek various types of information via the web [7]. Currently, many studies are assessing the capabilities and potential applications of this chatbot, but the reliability of the information provided by ChatGPT still requires validation.

ChatGPT has been recognized for its potential to revolutionize the entire medical field across several areas, such as patient care, health care systems and professionals, research, and education and training [8]. A 2020 survey indicated that 55% of Europeans aged 16 - 74 years searched for health-related information on the internet, reflecting a 21% increase since 2010. In the United States, the percentage rose from 62.8% in 2008% to 74.7% in 2017, showing an 11.9% increase [9,10]. Because ChatGPT is still new, little is known about how it will affect society in the future, especially in the medical field. In the future, both patients and health care workers might move from using the internet in a basic way to getting more interactive, AI-supported information. Instead of looking up facts about a disease or treatment, they might ask the AI directly, "How should this disease be treated?" or "What are the complications of this treatment?" Learning how AI works and checking how good its answers are will be a big but necessary job for future health care workers. Mistakes or wrong medical information from AI could have serious and even harmful effects in the worst cases [11].

AI has made significant progress in health care, especially in several areas of ophthalmology [12,13]. Since the complexity of ophthalmology makes it highly challenging for ChatGPT to provide accurate answers to questions in this field, the use of ChatGPT to access medical information in ophthalmology remains unexplored. For this reason, further validation studies are crucial, particularly to evaluate the effectiveness of ChatGPT in addressing a broad spectrum of eye disease–related questions within the clinical practice setting in Saudi Arabia.

Considering the growing trend and increasing popularity of ChatGPT, it is likely that people will rely on it to ask questions related to ophthalmology, particularly in cases of emergency eye injuries such as those caused by chemicals. As a result, we aimed to assess the accuracy of the information provided by ChatGPT in ophthalmology, particularly on chemical injury of the conjunctiva and cornea. Additionally, we sought to evaluate the potential for incorrect or even harmful information that could pose risks to patients using the chatbot independently. This study aims to assess the use of AI ChatGPT for accessing medical information on chemical corneal and conjunctival injuries.

Methods

Generating Medical Information

The following data were generated on December 21, 2024, using the official ChatGPT-4-turbo app by OpenAI. To evaluate the ability of ChatGPT to provide accurate medical information, we presented the AI ChatGPT with a series of questions related to chemical ocular injury. These questions covered topics such as the definition, prevalence, etiology, prevention, symptoms, diagnosis, treatment, follow-up, and complications of chemical ocular injury. A total of 9 distinct ChatGPT responses were analyzed, corresponding to the 9 queries submitted.

ChatGPT responses were then compared with the *International Classification of Diseases-9* (ICD-9) and *ICD-10* guidelines for chemical (alkali and acid) injuries of the conjunctiva and cornea. These responses were evaluated based on criteria such as correctness, thematic accuracy, and coherence to assess their accuracy. ChatGPT-4-turbo uses a probabilistic algorithm, meaning it uses random sampling to generate a variety of responses, which may include different answers to the same query. This investigation focused solely on ChatGPT's initial response to each query, without regenerating answers. Additional clarifications or explanations were not allowed. All queries were entered into a ChatGPT account owned by the author in a single day, ensuring accuracy in grammar and syntax. Each query was placed into a separate dialogue window to eliminate confounding factors and ensure the accuracy and precision of the responses, as ChatGPT-4-turbo can adapt based on the details of each interaction. The study did not require institutional review board approval, as it did not involve human participants, and no patient-identifying information was used.

Evaluation of Answers

In evaluating the responses to each question, we compared them with established guidelines and assessed them based on their correctness, thematic accuracy, and coherence. The criteria are organized as shown in Textbox 1.

XSL•FO
RenderX

Textbox 1. Criteria for evaluation of answers.

Correctness

Correct: The entire answer is correct.

Partially correct: The core statement of the answer is correct, but the rest of the answer contains one or more errors.

Incorrect: The core statement of the answer is incorrect.

Thematic accuracy

Applicable: The answer identifies the central concept and is thematically specific.

Partially applicable: The answer identifies the central concept, but also partially addresses an unrelated topic.

Not applicable: The answer does not identify the central concept or target an unrelated topic.

Coherence

Coherent: The core message of the answer is fully supported by the rest of the answer.

Partially coherent: The core statement of the answer is essentially confirmed by the rest of the answer, but there are deviating statements/contradictions in the rest of the answer.

Incoherent: The core statement of the answer contradicts the rest of the answer [14].

Data Analysis

We performed an in-depth linguistic analysis to evaluate the readability and complexity of the AI-generated responses. To achieve this, 2 main readability scores, the Flesch Reading Ease Score (FRES) and the Flesch–Kincaid Grade Level (FKGL), were computed using standard formulas. FRES evaluates text complexity on a scale from 0 to 100, where lower scores suggest more difficult reading.

$$FRES=206.835-1.015\times(WordsSentences)-84.6\times(SyllablesWords)$$

FKGL approximates the US education grade level required to comprehend the text.

$$FKGL=0.39\times(WordsSentences)+11.8\times(SyllablesWords)-15.59$$

Sentence Identification and Data Extraction

All textual analyses were performed on the final output for each of the 9 sections (definition, prevalence, causes, preventive measures, symptoms, diagnosis, treatment, follow-up, and complications). Each section's text was analyzed separately, without alteration, ensuring that the analysis reflected the original language. Sentences were identified by their terminal punctuation (periods, exclamation points, or question marks).

Passive Voice Identification, Word and Syllable Counting

Words were counted by recognizing sequences of characters separated by whitespace or punctuation. Contractions (eg, "don't") were treated as single words. Hyphenated words (eg, "well-labeled") were similarly counted as single words, ensuring uniform treatment of such constructions. A standard English syllable-counting heuristic was applied to each word, examining vowel groupings and common exceptions (eg, silent "e" or diphthongs). This automated count was then manually reviewed for potential errors, especially in acronyms or chemical abbreviations, to ensure accurate syllable totals. Each sentence was manually inspected to detect passive constructions, characterized by a form of "be" followed by a past participle (eg, "...was performed"). Such sentences were marked as passive. The total count of passive sentences in each section was divided by the total sentence count to calculate the percentage of passive use.

In our study, we first calculated the average (mean) and variability (SD) for both the FRES and FKGL indices. We also calculated the percentage of passive sentences to assess the overall readability performance of ChatGPT. We included the analysis of the percentage of passive sentences because it can impact the readability and comprehension of the text. Passive structures are typically more difficult to read and understand, especially for individuals with lower reading proficiency.

Subsequently, we categorized the 9 questions about chemical ocular injury into three distinct groups: (1) definition, prevalence, and etiology; (2) prevention, symptoms, and diagnosis; and (3) treatment, follow-up, and complications. For each category, we calculated the mean and SD for the FRES, FKGL, and the percentage of passive sentences in ChatGPT responses. These initial steps set the stage for more detailed statistical analyses, including ANOVA and trend analysis, which allowed us to investigate whether the readability performance of ChatGPT varied across the different question categories. ANOVA was used to compare the means of 3 or more groups to determine if there were any statistically significant differences among them. In this study, ANOVA was applied to compare the means of the readability metrics across the 3 categories of questions. Trend analysis was performed to identify patterns and shifts in the readability metrics across the categories. The mean and SD for each readability metric were plotted to visualize trends in the data. This analysis helped uncover any systematic changes in readability and complexity of the responses based on the type of information provided. Hypothesis testing was performed at a 5% significance level. This comprehensive analytical approach enabled us to gain a deeper understanding of the textual characteristics of the AI-generated content and thoroughly assess its suitability for patient education.

Ethical Considerations

This study did not involve human participants, clinical data, or personal identifiers and was therefore exempt from institutional

XSL•FO
RenderX

ethics review. As no human participants were involved, informed consent was not applicable. Additionally, no personal data were collected or analyzed, as all data were generated using the ChatGPT-4-turbo app developed by OpenAI. Since no participants were recruited or compensated, compensation considerations were not applicable. Furthermore, no identifiable images of individuals were included in the manuscript or any supplemental materials.

Results

Qualitative Analysis

Figures 1,2,3,4,5,6,7,8,9,10,11,12,13-15 illustrate ChatGPT responses to a series of questions addressing various aspects of chemical ocular injuries, including their definition, prevalence, etiology, prevention, symptoms, diagnosis, treatment, follow-up, and potential complications. The initial objective of this analysis is to evaluate the accuracy of ChatGPT's answers to each of these questions, ensuring that the information provided aligns with established medical knowledge.

Figure 1. ChatGPT-4-turbo's response to the question "What is the definition of chemical eye injury?" generated on December 21, 2024, as part of a comparative study assessing artificial intelligence–generated medical information on chemical ocular injuries. The study was conducted in Riyadh, Saudi Arabia.

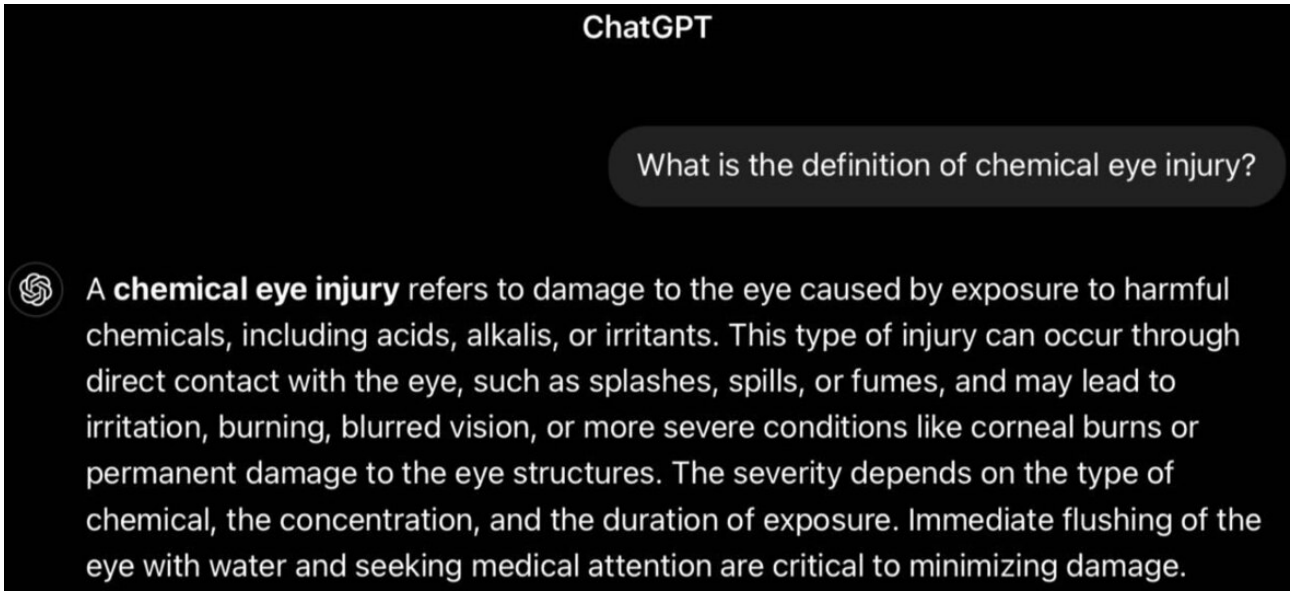

XSL•FO
RenderX

Figure 2. ChatGPT-4-turbo's response to the question "Can you tell me the prevalence of chemical eye injury?" generated on December 21, 2024, as part of a comparative study assessing artificial intelligence–generated medical information on chemical ocular injuries. The study was conducted in Riyadh, Saudi Arabia.

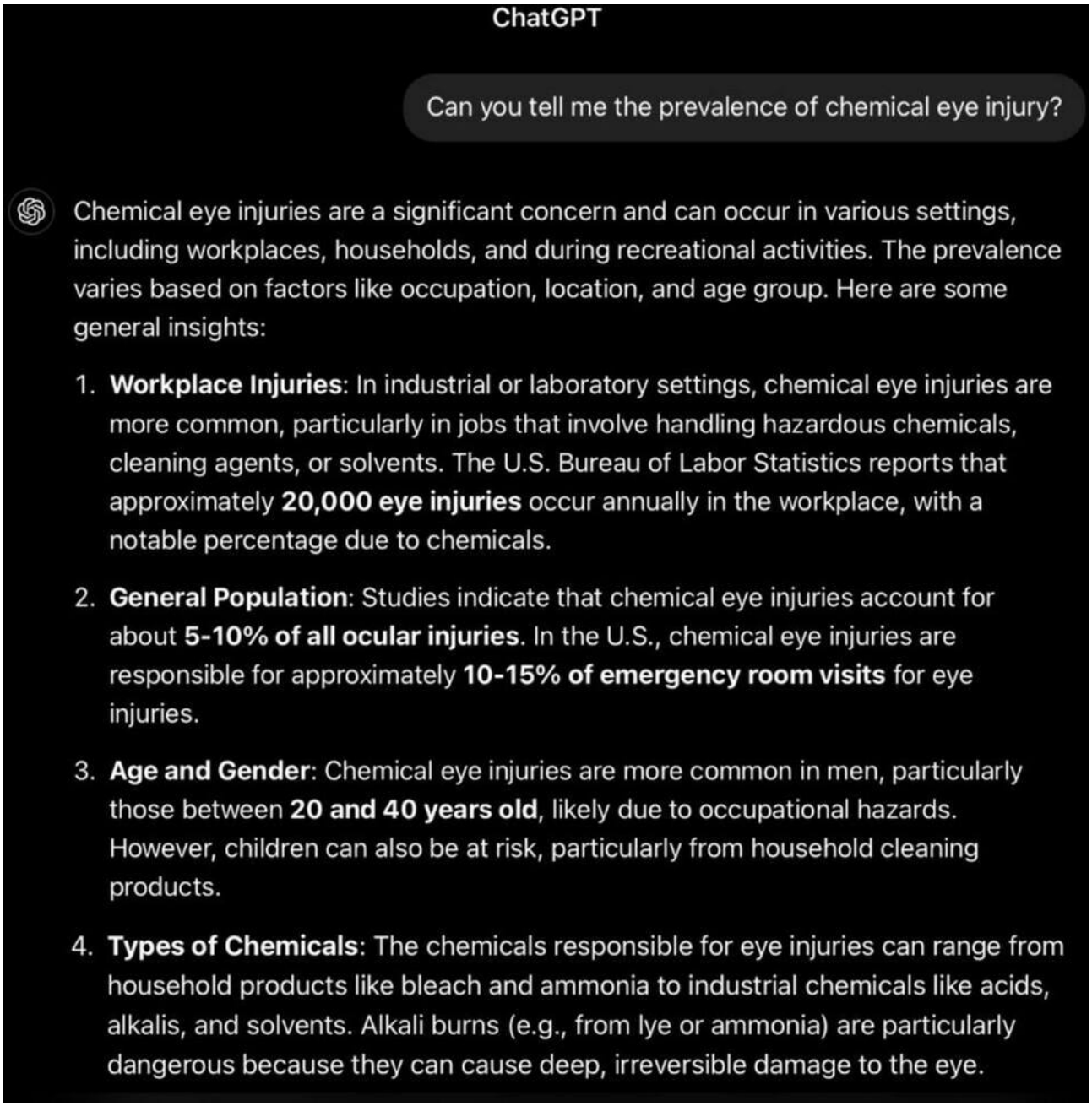

Figure 3. The rest of ChatGPT-4-turbo's response to the question "can you tell me the prevalence of chemical eye injury?" generated on December 21, 2024, as part of a comparative study assessing artificial intelligence–generated medical information on chemical ocular injuries. The study was conducted in Riyadh, Saudi Arabia.

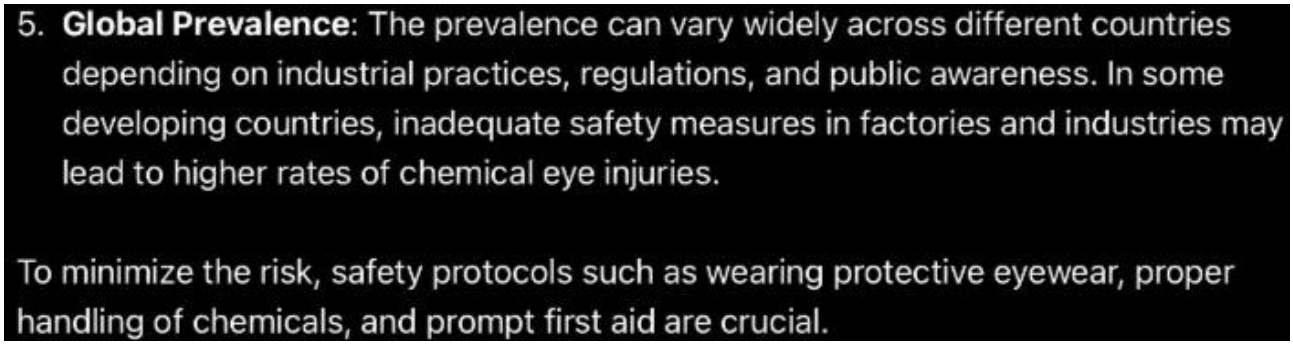

XSL•FO
RenderX

Figure 4. ChatGPT-4-turbo's response to the question "can you tell me the causes of chemical eye injury" generated on December 21, 2024, as part of a comparative study assessing artificial intelligence–generated medical information on chemical ocular injuries. The study was conducted in Riyadh, Saudi Arabia.

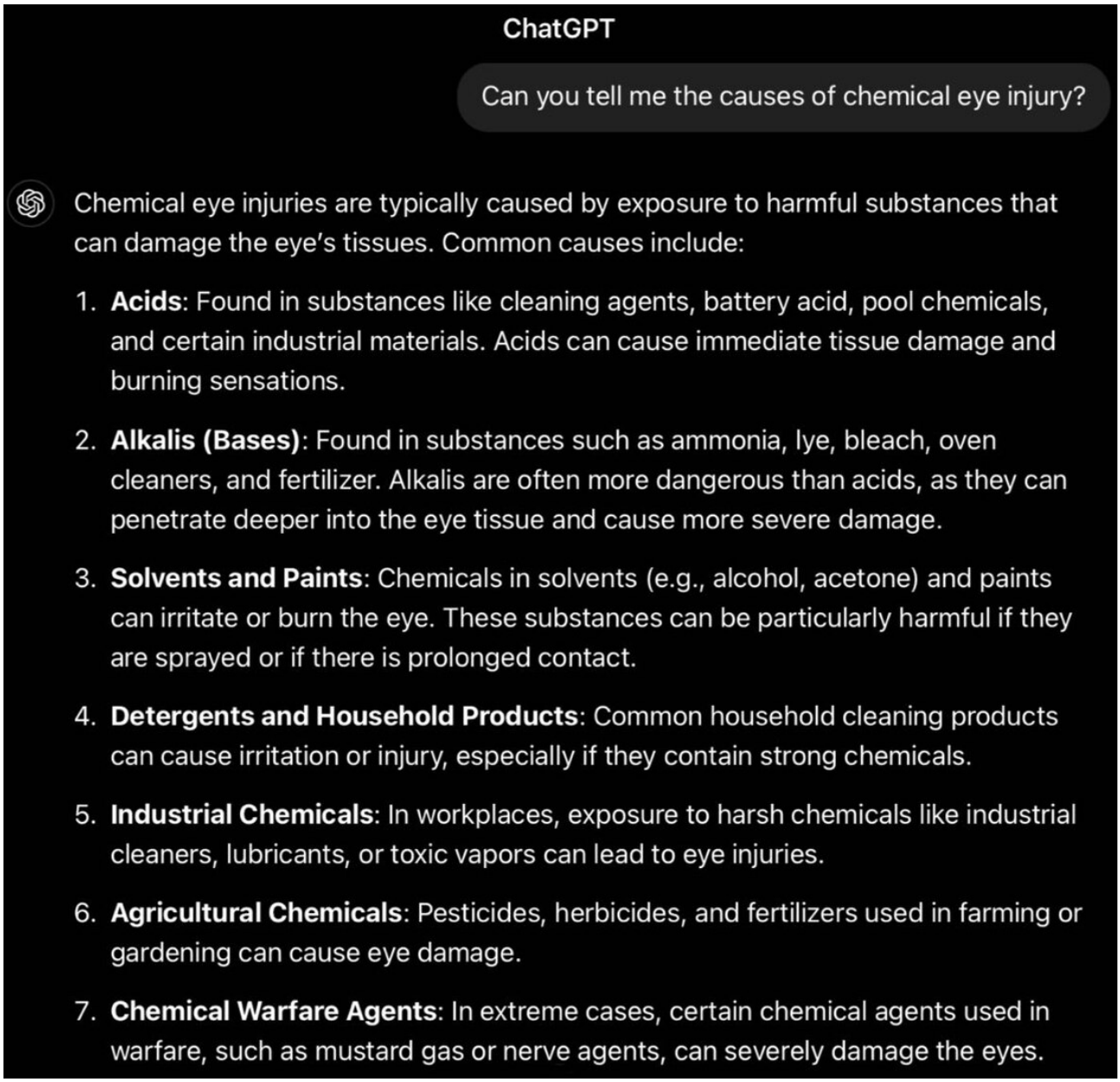

XSL•FO
RenderX

Figure 5. ChatGPT-4-turbo's response to the question "What are the preventive measures of chemical eye injury?" generated on December 21, 2024, as part of a comparative study assessing artificial intelligence–generated medical information on chemical ocular injuries. The study was conducted in Riyadh, Saudi Arabia.

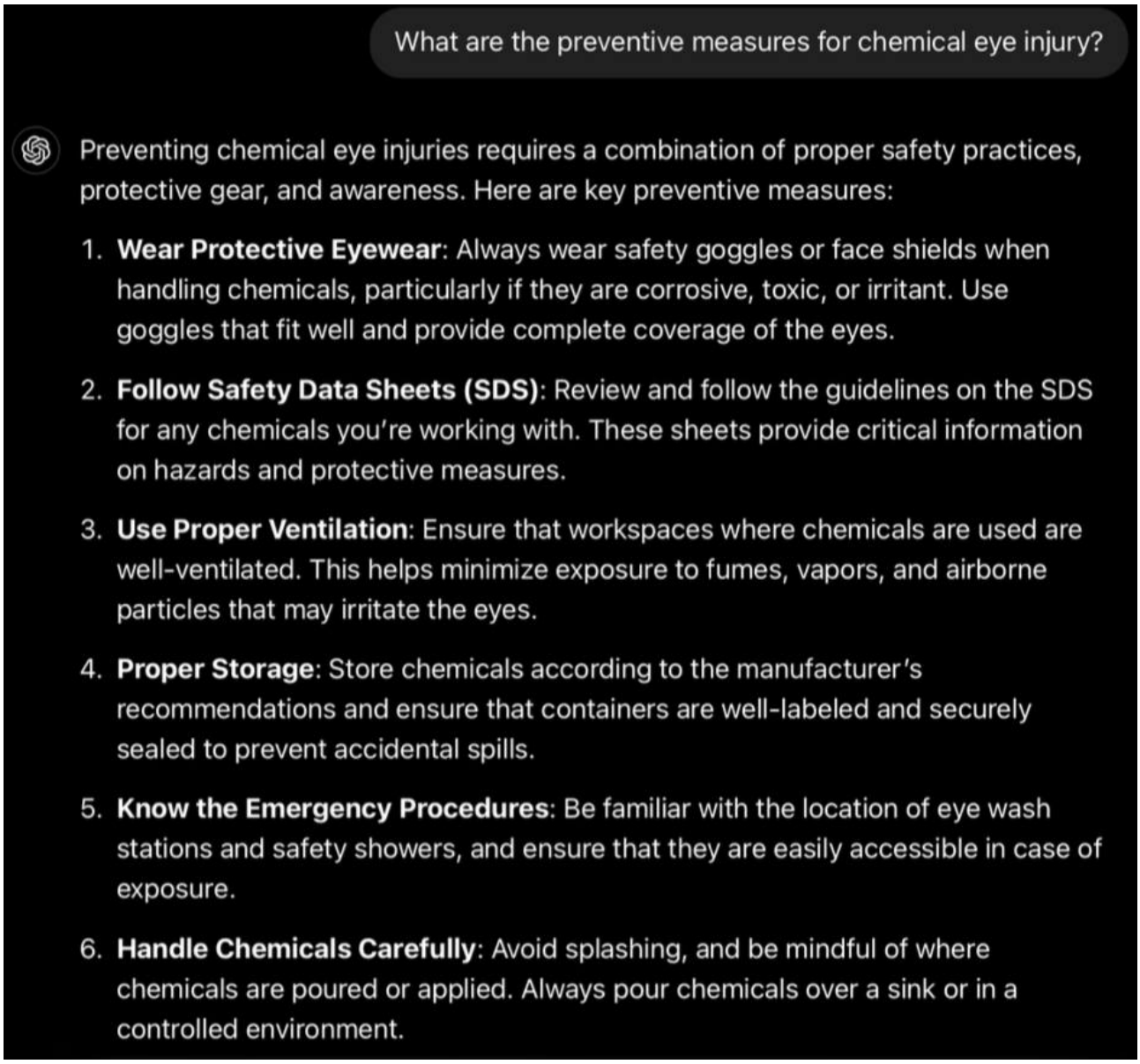

Figure 6. Rest of ChatGPT-4-turbo's response to the question "What are the preventive measures of chemical eye injury?" generated on December 21, 2024, as part of a comparative study assessing artificial intelligence–generated medical information on chemical ocular injuries. The study was conducted in Riyadh, Saudi Arabia.

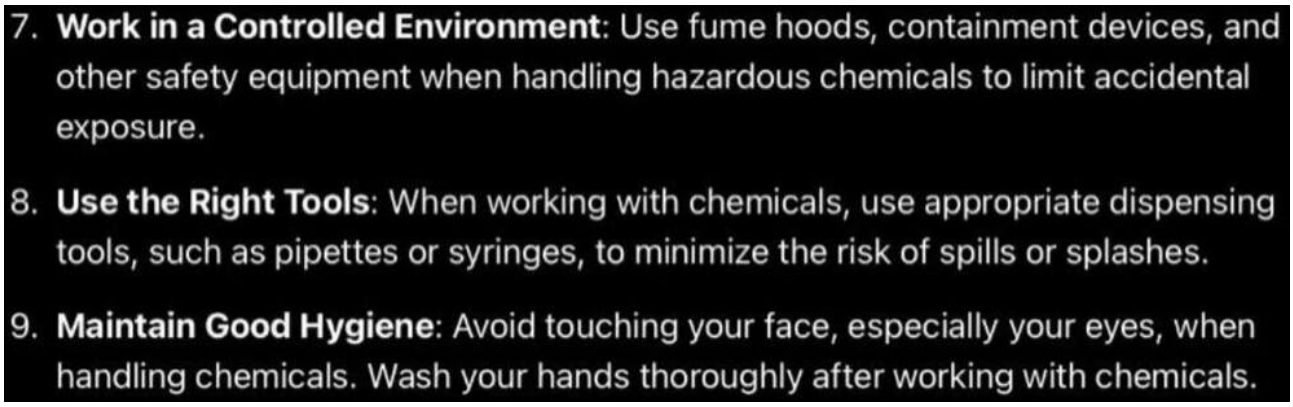

XSL•FO
RenderX

Figure 7. ChatGPT-4-turbo's response to the question "What are the symptoms of chemical eye injury?" generated on December 21, 2024, as part of a comparative study assessing artificial intelligence–generated medical information on chemical ocular injuries. The study was conducted in Riyadh, Saudi Arabia.

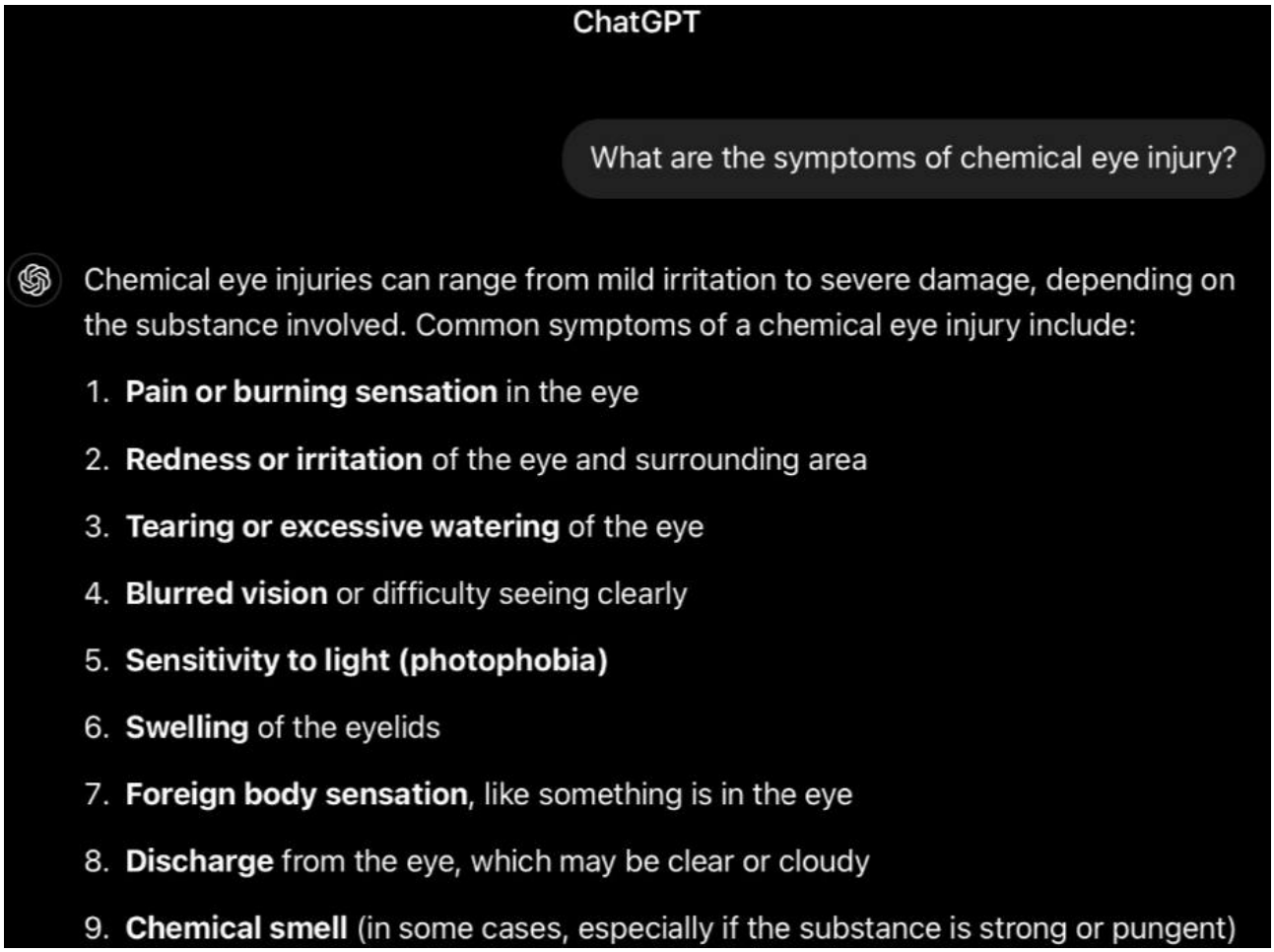

XSL•FO
RenderX

Figure 8. ChatGPT-4-turbo's response to the question "Can you tell me how to diagnose chemical eye injury?" generated on December 21, 2024, as part of a comparative study assessing artificial intelligence–generated medical information on chemical ocular injuries. The study was conducted in Riyadh, Saudi Arabia.

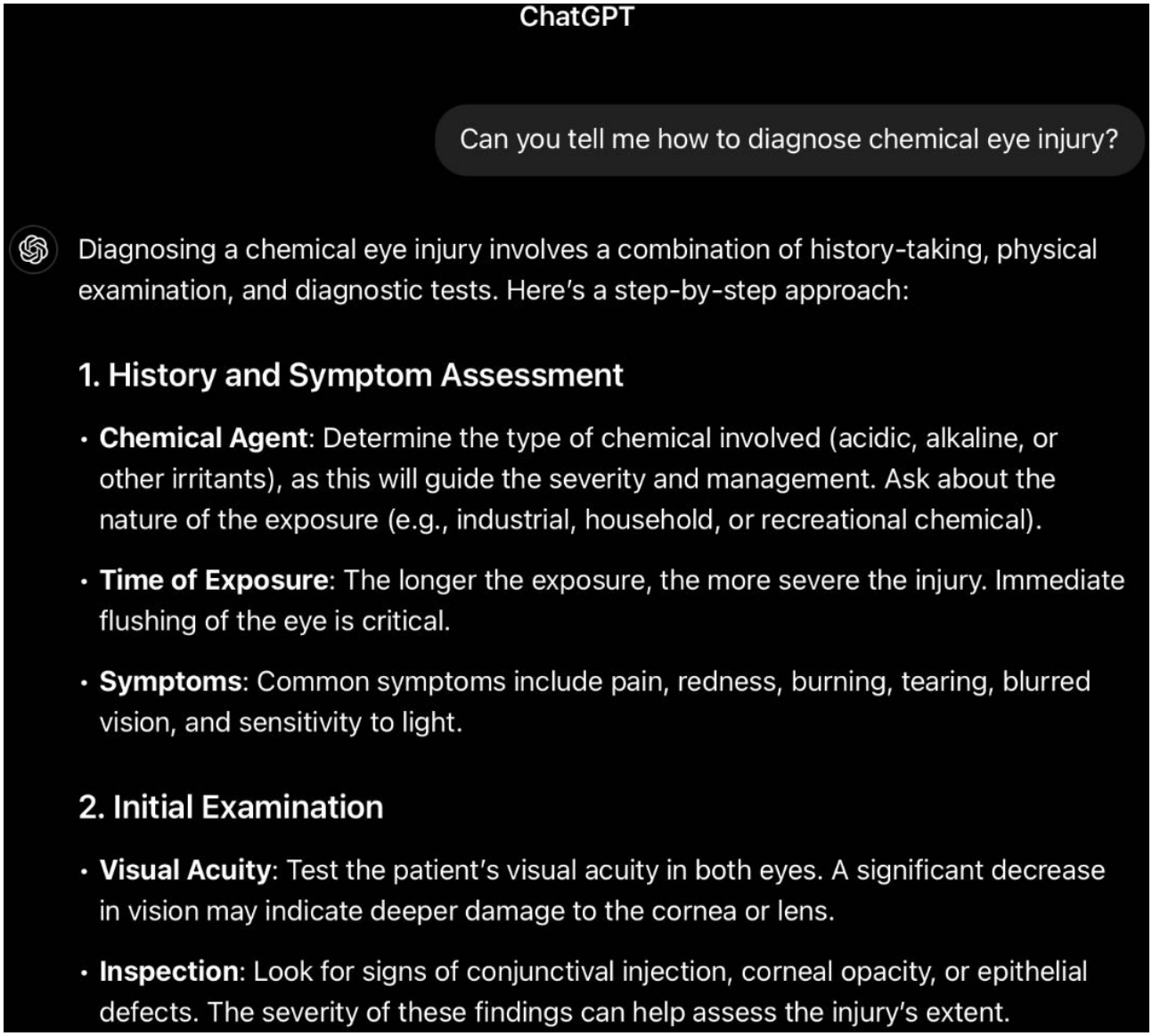

XSL•FO
RenderX

Figure 9. The rest of ChatGPT-4-turbo's response to the question "can you tell me how to diagnose chemical eye injury?" generated on December 21, 2024, as part of a comparative study assessing artificial intelligence–generated medical information on chemical ocular injuries. The study was conducted in Riyadh, Saudi Arabia.

3. Fluorescein Staining

- Apply fluorescein dye to the cornea and examine under a blue light to detect any epithelial defects, which may indicate a corneal injury.

4. Assessing pH Level

- **pH Test**: Use pH paper to test the eye's pH level. Normal pH is around 7.0. For chemical injuries:
  - Alkaline substances (e.g., lye, ammonia) can cause deeper tissue damage.
  - Acidic substances (e.g., sulfuric acid) typically cause more superficial injuries.
- A pH of less than 7.0 or greater than 7.5 may indicate ongoing chemical exposure and necessitate further irrigation.

5. Slit Lamp Examination

- A slit lamp is crucial to assess for corneal and conjunctival damage. Look for:
  - Corneal abrasions or ulcerations.
  - Chemical burns or opacification.
  - Iris or lens injury.
- Assess the anterior chamber for any signs of inflammation or hyphema (blood in the anterior chamber), which may indicate more severe injury.

XSL•FO
RenderX

Figure 10. The rest of ChatGPT-4-turbo's response to the question "Can you tell me how to diagnose chemical eye injury?" generated on December 21, 2024, as part of a comparative study assessing artificial intelligence–generated medical information on chemical ocular injuries. The study was conducted in Riyadh, Saudi Arabia.

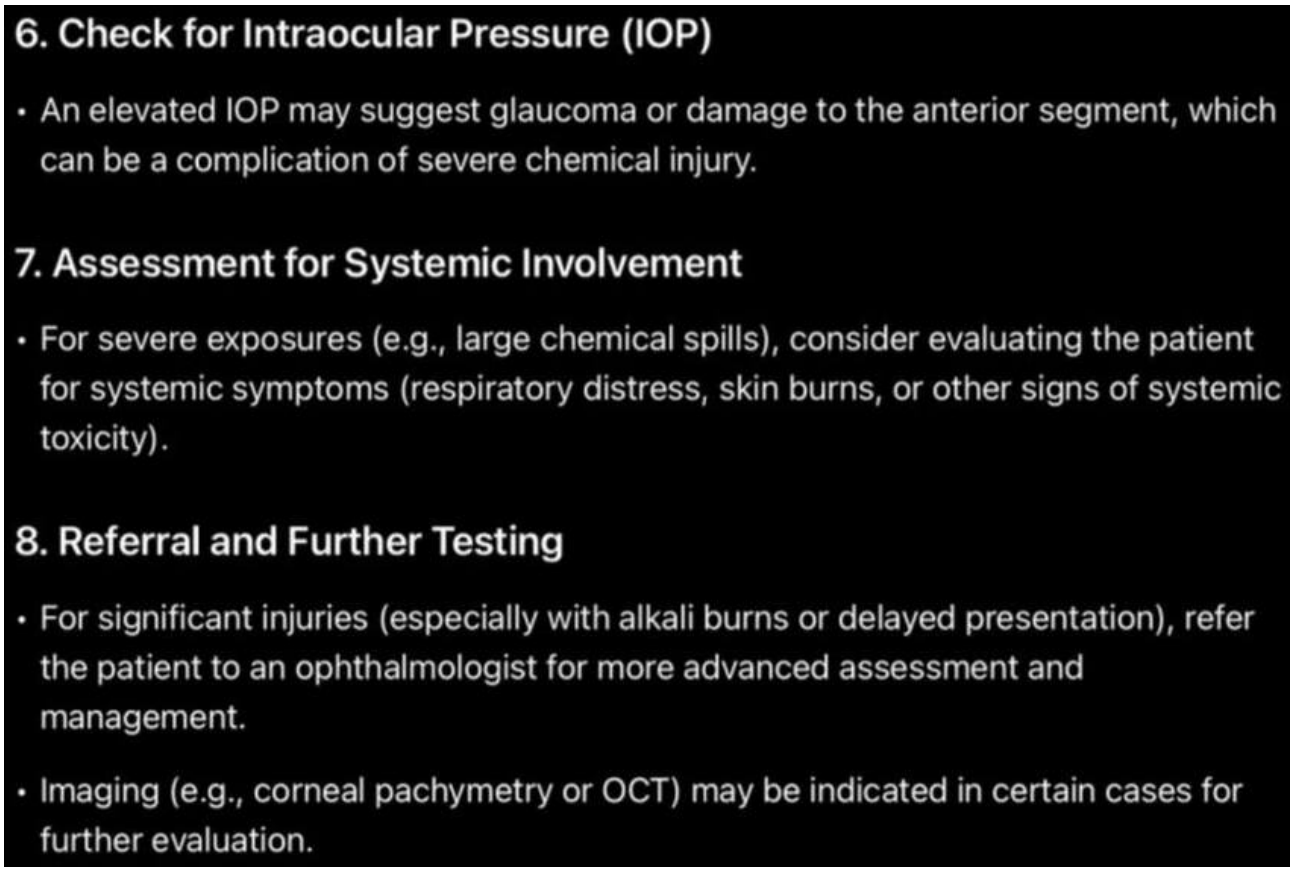

XSL•FO
RenderX

Figure 11. ChatGPT-4-turbo's response to the question "can you tell me what the treatment of chemical eye injury is?" generated on December 21, 2024, as part of a comparative study assessing artificial intelligence–generated medical information on chemical ocular injuries. The study was conducted in Riyadh, Saudi Arabia.

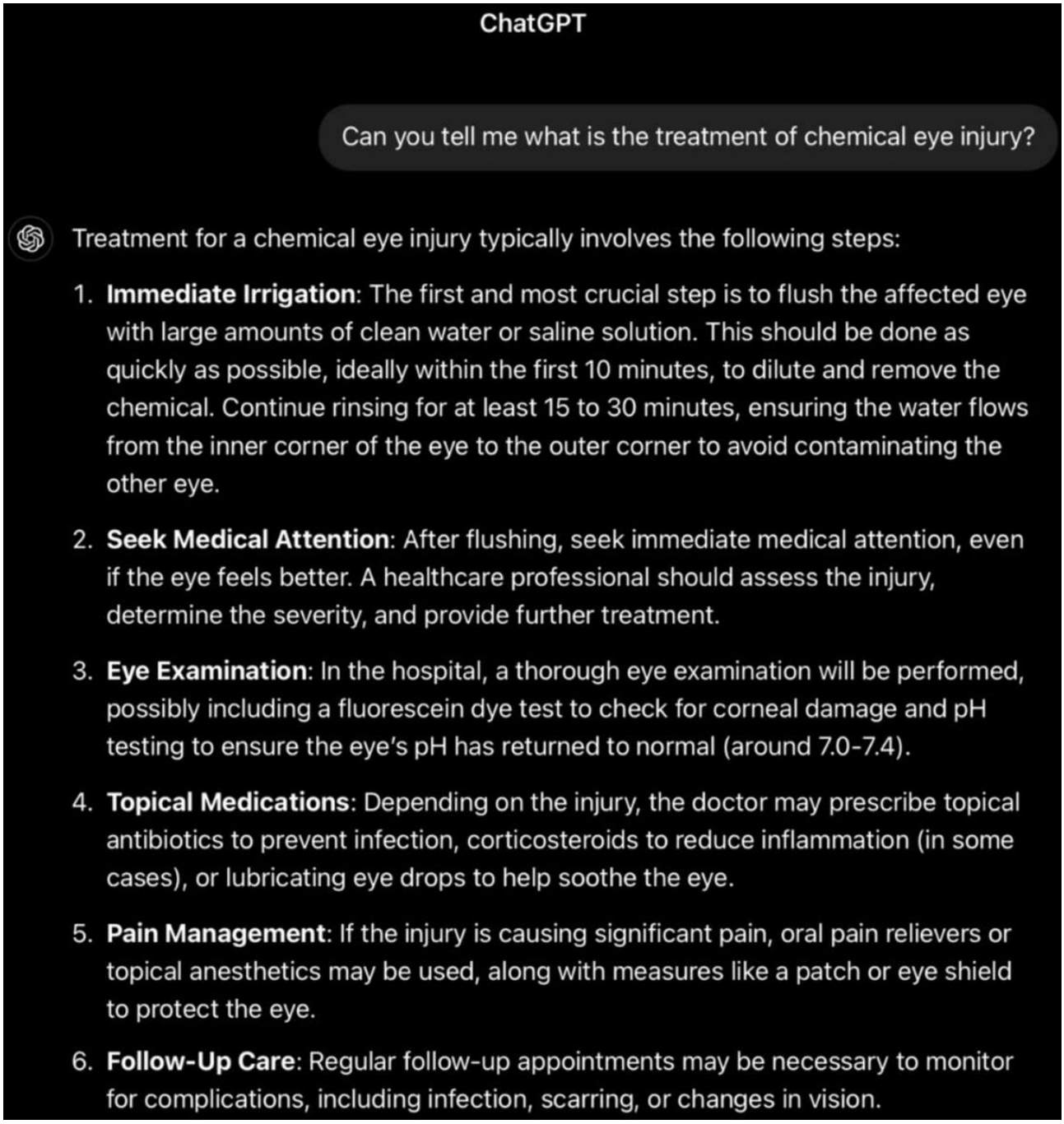

XSL•FO
RenderX

Figure 12. ChatGPT-4-turbo's response to the question "What is the follow-up care for chemical eye injury" generated on December 21, 2024, as part of a comparative study assessing artificial intelligence–generated medical information on chemical ocular injuries. The study was conducted in Riyadh, Saudi Arabia.

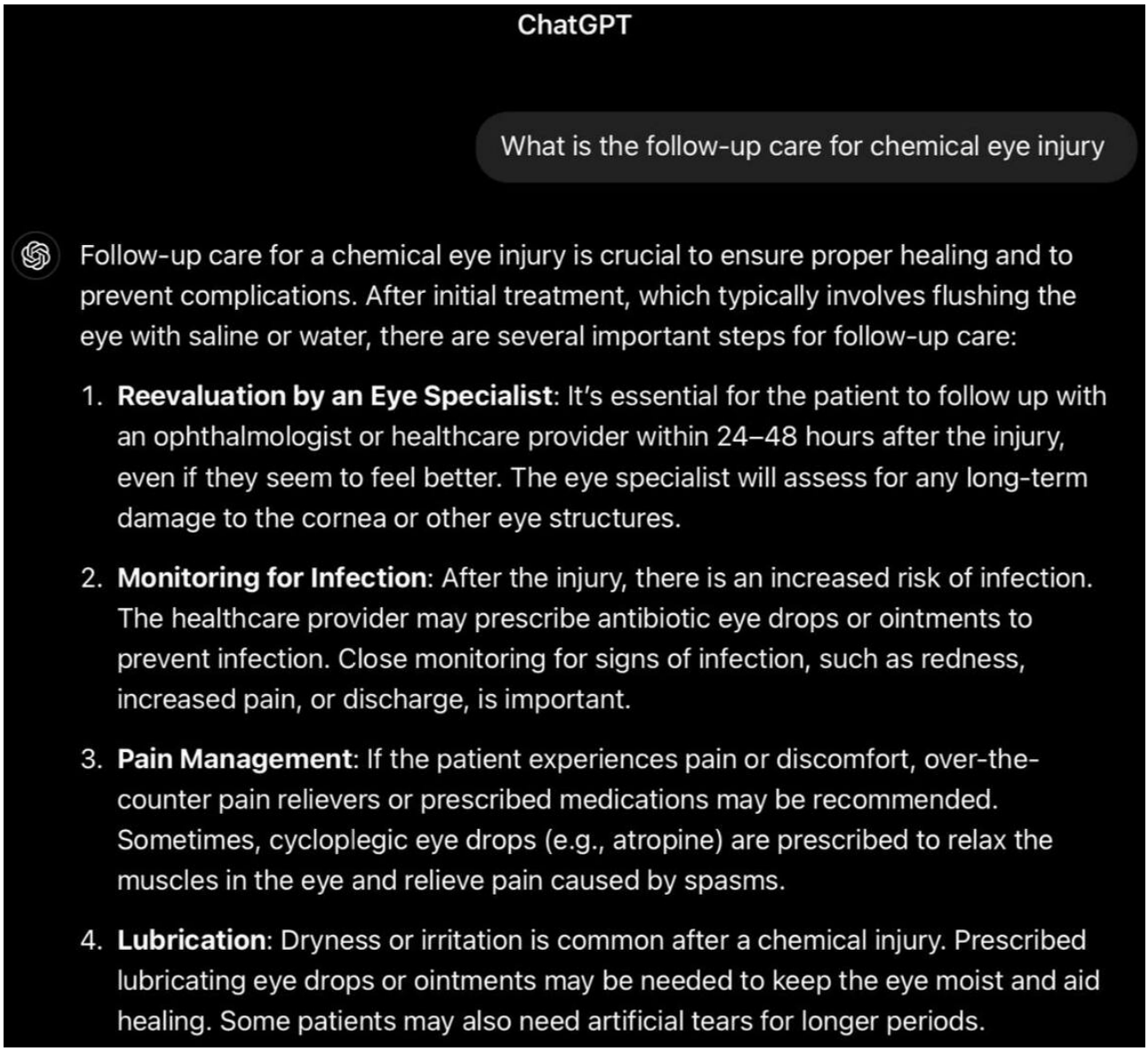

XSL•FO
RenderX

Figure 13. The rest of ChatGPT-4-turbo's response to the question "What is the follow-up care for chemical eye injury" generated on December 21, 2024, as part of a comparative study assessing artificial intelligence–generated medical information on chemical ocular injuries. The study was conducted in Riyadh, Saudi Arabia.

5. **Avoiding Strain or Further Injury**: The patient should avoid rubbing their eyes, exposure to bright lights, and activities that may lead to further trauma (e.g., sports or heavy physical activity). The use of protective eyewear may be recommended during the healing period.
6. **Corneal Healing and Scarring**: Chemical injuries can cause scarring on the cornea, which may impair vision. The ophthalmologist will monitor corneal healing over time and may suggest interventions such as a corneal transplant if significant scarring develops.
7. **Follow-Up Appointments**: Depending on the severity of the injury, follow-up visits may be required at intervals (e.g., 1 week, 1 month, 3 months) to ensure proper healing, assess vision, and adjust the treatment plan if necessary.
8. **Long-Term Care**: In severe cases, patients may require long-term care to manage complications like chronic dryness, vision changes, or scarring. In some cases, vision rehabilitation or corrective surgeries may be necessary.

It's important to adhere to the healthcare provider's instructions for follow-up visits and treatment to ensure the best possible outcome.

XSL•FO
RenderX

Figure 14. ChatGPT-4-turbo's response to the question "what is the complication of chemical eye injury" generated on December 21, 2024, as part of a comparative study assessing artificial intelligence–generated medical information on chemical ocular injuries. The study was conducted in Riyadh, Saudi Arabia.

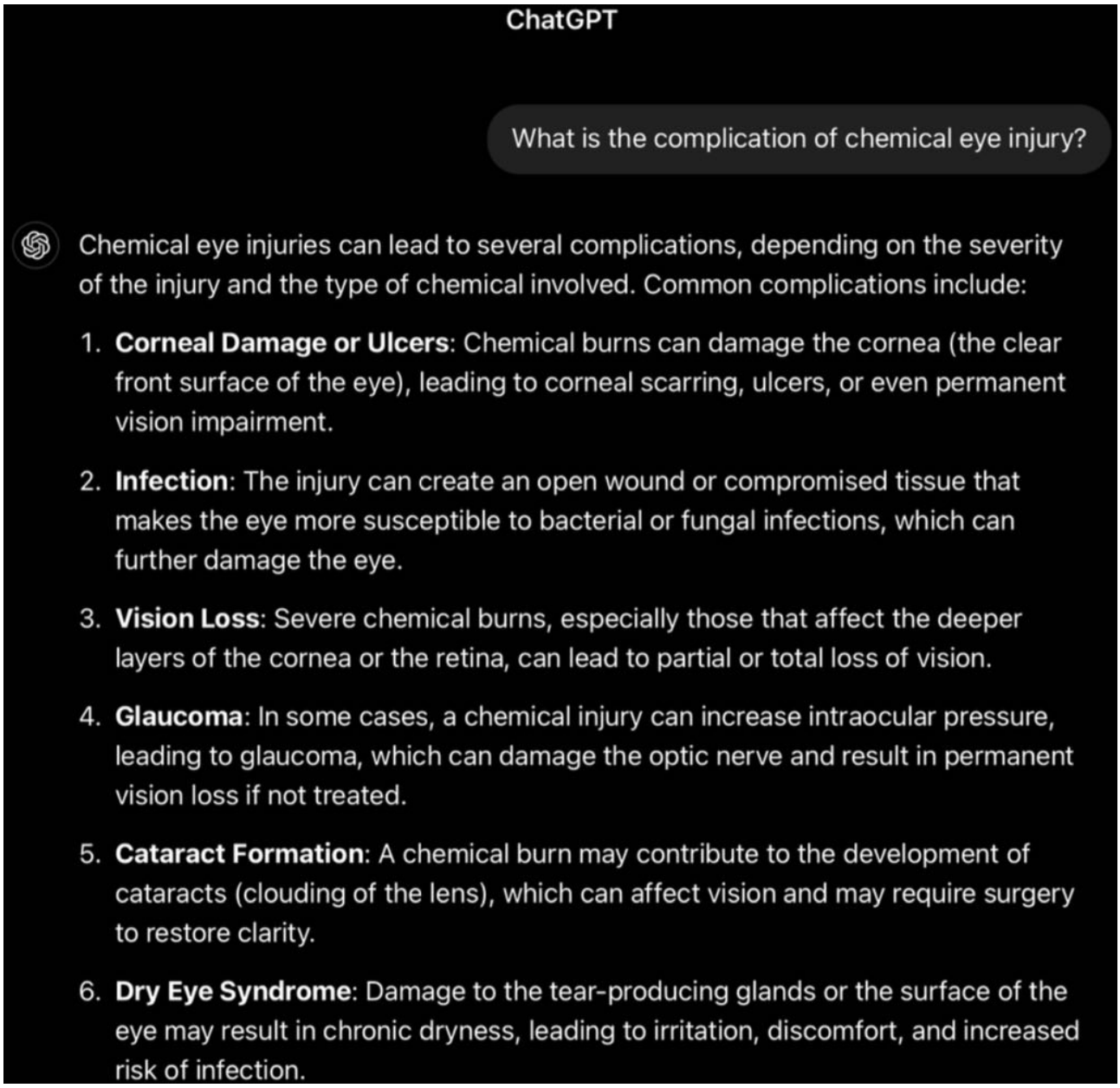

XSL•FO
RenderX

Figure 15. The rest of ChatGPT-4-turbo's response to the question "what is the complication of chemical eye injury" generated on December 21, 2024, as part of a comparative study assessing artificial intelligence–generated medical information on chemical ocular injuries. The study was conducted in Riyadh, Saudi Arabia.

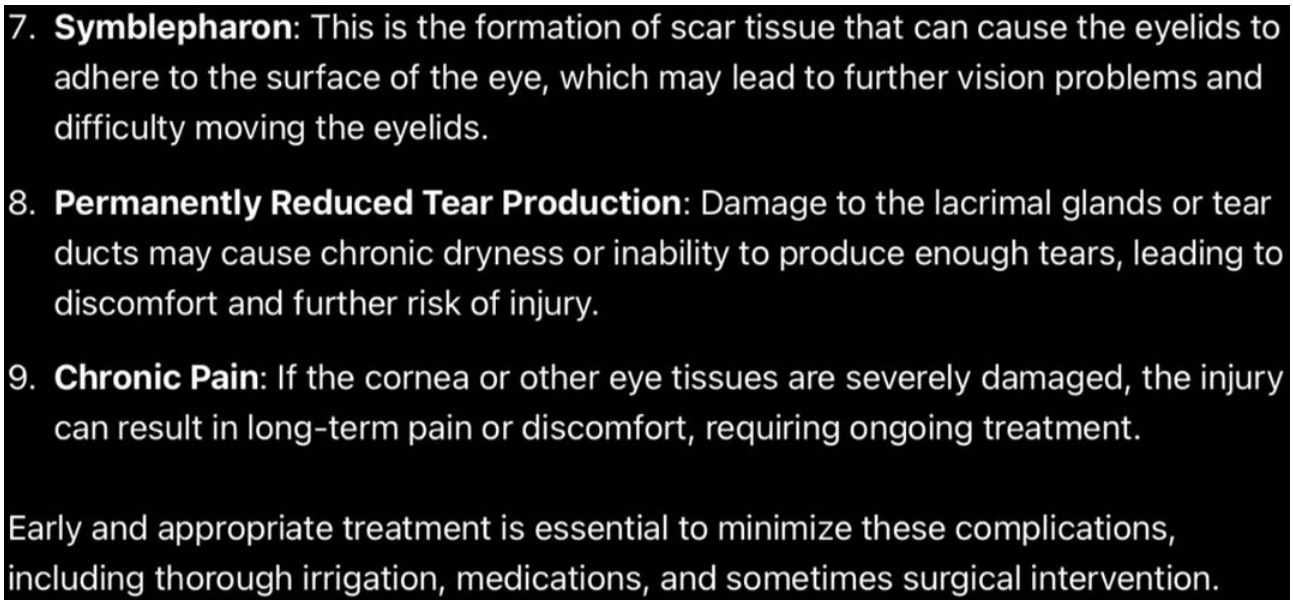

For the definition question (Figure 1), ChatGPT defines chemical eye injury as it is commonly defined in the medical literature. The analysis of ChatGPT's response shows that it is correct, providing an accurate description of chemical eye injuries, including their causes, symptoms, and some complications. It also briefly notes the importance of immediate treatment and the factors that determine the severity. It is thematically accurate, as it identifies the central concept and remains fully focused on the topic. The response is also coherent, with its core statement supported by the rest of the answer without contradictions. In comparison, ChatGPT's response offers a broader explanation, detailing causes (acids, alkalis, and irritants), symptoms (irritation, burning, and blurred vision), and a brief mention of treatment (flushing and medical attention). Meanwhile, the *ICD-9* and *ICD-10* guidelines are more concise, emphasizing the urgency and severity of chemical eye injuries and targeting medical professionals or individuals needing a quick understanding of the condition.

ChatGPT's response, while accurate and comprehensive, fails to sufficiently emphasize that chemical eye injuries are true emergencies that demand immediate attention and prompt intervention. It primarily focuses on describing general information, including causes, symptoms, and general treatment. This lack of emphasis on the emergency nature of such injuries may diminish the critical need for quick action.

For the prevalence question (Figures 2 and 3), the analysis of ChatGPT's response reveals that it is partially correct. It provides accurate and detailed information on the prevalence of chemical eye injuries, including key statistics and contributing factors such as workplace exposure, gender, and age. However, in the general population, ChatGPT's response states that chemical eye injuries account for about 5% - 10% of all ocular injuries, which is incorrect according to *ICD-9* and *ICD-10* guidelines. These guidelines indicate that chemical injuries to the eye represent between 11.5% and 22.1% of ocular traumas. The response is also partially accurate in terms of thematic accuracy. Although it stays focused on the central topic of the prevalence of chemical eye injuries and addresses relevant subtopics, such as industrial versus domestic settings and global variations, it also discusses the types of chemicals, which was not the main question. Furthermore, it is coherent, with the core statements consistently supported by the detailed explanations provided throughout the response.

In comparison, ChatGPT's response is more detailed and comprehensive, covering multiple aspects, including workplace statistics (eg, 20,000 injuries annually in the United States), demographic factors (eg, men aged 20 - 40 years), and the influence of chemical types. In contrast, the guidelines provide a more concise overview, primarily focusing on general prevalence rates (11.5% - 22.1%), demographic risks (young men and children), and the primary setting of workplace injuries. While both answers are accurate and coherent, ChatGPT's response offers a broader and more informative perspective by incorporating additional data and global context, whereas the guidelines focus on essential statistics and trends, making it more concise. ChatGPT's response failed to mention that alkali injuries are more common than acid injuries and are often caused by building materials and cleaning agents, as this exclusion is significant for understanding and managing chemical eye injuries effectively.

For the causes question (Figure 4), the analysis of ChatGPT's response to the etiology of chemical eye injuries reveals that it is partially correct, as it identifies that chemical injuries can be caused by acids and alkalis. However, some of the listed etiologies are not mentioned in the guidelines, including solvents, detergents, industrial chemicals, agricultural chemicals, and chemical warfare agents. Moreover, the response lacks specificity in its categorization and examples compared with the *ICD-9* and *ICD-10* guidelines. For instance, ChatGPT's response mentions "cleaning agents, battery acid, and pool chemicals" for acids but does not specify compounds like sulfuric acid (H_2SO_4) or hydrofluoric acid, as outlined in the

XSL•FO
RenderX

guidelines. ChatGPT's response lacks the specific and systematic presentation of substances that the guidelines offer, which is essential for clinical and educational purposes. Thematic accuracy is applicable. The response identifies the central concept (causes of chemical eye injuries) and stays fully focused on the topic. The response is coherent, with its main statement consistently supported by the rest of the answer and free from contradictions.

For the preventive question (Figures 5 and 6), the analysis of ChatGPT's response shows that it is correct, providing accurate and comprehensive preventive measures that align with standard practices. It emphasizes wearing protective eyewear, following safety protocols, proper chemical handling, and the use of emergency equipment. Thematic accuracy is applicable. The response identifies the central concept (preventing chemical eye injuries) and stays fully focused on the topic. All measures are practical and thematically relevant. The response is coherent, and the answer's core statement (preventing chemical eye injuries through safety practices) is well-supported throughout. It does not contradict itself and maintains logical progression.

While the response provides detailed preventive strategies, it could more effectively highlight the Occupational Safety and Health Administration regulations mentioned in the guidelines to emphasize compliance with mandatory workplace safety standards. The guidelines specifically mandate the use of protective eyewear in workplace settings, and although the response addresses this and its proper use, it does not sufficiently stress its "mandatory" nature under regulatory requirements. Additionally, the response includes broader prevention measures, such as hygiene practices, proper ventilation, and emergency procedures, which are practical for general audiences but not clearly covered in the guidelines. To improve, the response should emphasize the mandatory nature of protective eyewear per Occupational Safety and Health Administration regulations.

For the symptoms question (Figure 7), the analysis of ChatGPT's response shows that it is partially correct. The answer includes several accurate and commonly reported symptoms, such as pain and tearing. However, some of the listed symptoms deviate from the most clinically significant findings emphasized in the guidelines. The inclusion of "chemical smell" is not supported by standard guidelines and is not a defining symptom of chemical eye injury. Similarly, "foreign body sensation" may occur, but is not a primary or hallmark symptom. Moreover, the answer neglects "blepharospasm" and "reduced visual acuity", which are key symptoms in the guidelines. Thematic accuracy is partially applicable. While the answer broadly addresses the concept of chemical eye injury symptoms, it does not focus on the critical symptoms highlighted in the guidelines (severe pain, epiphora, blepharospasm, and reduced visual acuity), and by including less relevant symptoms (eg, "chemical smell"), it partially diverts attention from the more urgent diagnostic features. The response is also partially coherent. The core statement that chemical eye injuries can cause a range of symptoms is supported by the list. However, the inclusion of peripheral symptoms ("chemical smell" and "foreign body sensation") introduces inconsistencies, as these are not central to the condition and can detract from the urgency of hallmark symptoms.

For the diagnosis question (Figures 8-10), the analysis of ChatGPT's response shows that it is correct. It accurately describes the diagnostic process for chemical eye injuries. It includes key steps such as history-taking to determine the type of chemical, exposure time, and symptoms, followed by an initial examination of visual acuity and inspection for physical damage, including the use of fluorescein staining, pH testing, slit lamp examination, and intraocular pressure measurement. It is thematically accurate, as it identifies the central concept and remains fully focused on the topic. The response is also coherent, with its core statement supported by the rest of the answer without contradictions.

In comparison, ChatGPT's response appropriately begins with the importance of obtaining a patient's history in diagnosing chemical eye injuries. It suggests asking about the chemical agent to identify its type, the nature of exposure (household or industrial), the timing of exposure, and common symptoms such as pain, redness, burning, tearing, blurred vision, and light sensitivity. However, it omits several critical aspects that have been mentioned in *ICD-9* and *ICD-10* guidelines, including assessing the chemical's toxicity, understanding the depth of ocular penetration, determining the area of involvement, and asking whether the eyes were rinsed afterward and for how long. Additionally, it fails to mention the need to ask about the use of eye protection at the time of injury and the importance of examining the chemical's packaging for product information, including chemical composition. If such information is unavailable, contacting the local poison control center, such as the American Association of Poison Control Centers at 1-800-222-1222, is advisable.

Following the history, the physical examination is the next critical step. ChatGPT's response highlights essential components, such as conducting an initial examination to assess visual acuity and inspect for signs of conjunctival injection, corneal opacity, and epithelial defects. It also emphasizes the importance of fluorescein staining to detect epithelial damage, assessing pH levels to ensure they are within the physiologic range (not less than 7 or greater than 7.5, which may indicate ongoing chemical exposure requiring further irrigation), and performing a slit lamp examination. Additionally, intraocular pressure measurement and systemic evaluation for potential involvement are mentioned. For severe injuries, referral for further testing is recommended. However, the response overlooks the use of established classification schemes, such as the Roper-Hall classification and the Dua classification, which have been mentioned in *ICD-9* and *ICD-10* guidelines to evaluate the extent and depth of the injury. Specifically, these schemes document corneal, conjunctival, and limbal involvement, which are critical for predicting the ultimate visual outcome.

For the treatment question (Figure 11), the analysis of ChatGPT's response shows that it is partially correct. ChatGPT provides an outline of the appropriate treatment of chemical eye injury in 6 steps: immediate irrigation, seeking medical attention, eye examination, topical medication, pain

XSL•FO
RenderX

management, and finally, follow-up. It does not clarify the recommended grading approach depending on the severity of injury, since mild burns respond well to medical treatments and lubrication, while more severe burns necessitate more intensive medical therapy and surgery. Moreover, ChatGPT's answer missed mentioning essential components of the management, such as the preirrigation instructions, the performing technique, and the correct way to remove the foreign body.

It is thematically accurate, as it identifies the central concept and remains fully focused on the topic. The response is also coherent, with its core statement supported by the rest of the answer without contradictions. In comparison, the *ICD-9* and *ICD-10* guidelines provide detailed clarification of each step along with detailed explanations, such as providing the types of irrigating fluids, each with different benefits. Whereas ChatGPT's response seems to be short and brief. In addition, the *ICD-9* and *ICD-10* guidelines demonstrate that there are standard treatments and other treatments that are prescribed in special circumstances, unlike ChatGPT's response, which categorizes them into topical medication and pain management. The probability of the need for surgery, such as debridement of necrotic epithelium, amniotic membrane disease transplantation, limbal stem cell transplant, and cultivated oral mucosal epithelial transplantation, is not mentioned in ChatGPT's response.

For the follow-up question (Figures 12 and 13), ChatGPT provides a detailed and structured explanation of follow-up care for chemical eye injuries. The analysis of ChatGPT's response reveals that it is partially correct, as it accurately describes many key elements, such as reevaluation, infection monitoring, pain management, and long-term care. However, it misses critical points emphasized in the *ICD-10* and *ICD-9* guidelines, such as the necessity of daily follow-up for severe burns and inpatient admission for compliance concerns or pediatric patients. Furthermore, while the response mentions long-term care, it fails to address specific complications like glaucoma monitoring or interventions for eyelid and conjunctival damage. The response is partially applicable. While it addresses the central concept of follow-up care and includes relevant details, it does not fully align with the thematic focus of the guidelines. Key guideline-specific elements, such as the progression of follow-up intervals, the importance of early aggressive management of complications, and the long-term need for glaucoma and dry eye monitoring, are underemphasized or omitted. The response is coherent, with its core message regarding follow-up care being consistently supported by subsequent details. The organization and flow of information are logical, providing a stepwise approach from reevaluation to long-term care. However, coherence does not compensate for the absence of critical guideline-specific details.

For the complication question (Figures 14 and 15), ChatGPT identifies the complications of chemical eye injuries as commonly described in medical literature. The analysis of ChatGPT's response shows that it is partially correct, providing a general overview of common complications such as corneal damage, glaucoma, and dry eye syndrome. However, certain points, such as "permanently reduced tear production" and "chronic pain," are unnecessary or overly broad.

The response is applicable, as it focuses on the central concept of complications associated with chemical eye injuries. It lists key complications and provides brief explanations for most of them. However, the response does not fully align with the *ICD-10*/*ICD-9* guidelines, which offer more specific details about the pathophysiology and mechanisms underlying these complications. For example, the guidelines emphasize the multifactorial causes of glaucoma (eg, inflammatory debris and damage to the trabecular meshwork) and the detailed impact of dry eye syndrome, including mucus deficiency and its effect on the tear film, which ChatGPT does not sufficiently address. The response is partially coherent. While the listed complications are thematically relevant and logically connected, the explanation lacks depth and fails to detail the progression or severity of complications.

Quantitative Analysis

Overall Readability

Across all 9 sections, we found (Table 1) that the text presented an average FRES of 33.84 (SD 0.28), indicating a fairly challenging reading level. The FKGL averaged at 14.21 (SD 0.97), suggesting that the content is suitable for readers with college-level proficiency. Regarding sentence construction, the passive voice was used in 7.22% (SD 5.60%) of the sentences, implying a moderate reliance on the passive form.

XSL•FO
RenderX

Table . Overall and question-specific readability metrics (Flesch Reading Ease Score [FRES], Flesch–Kincaid Grade Level [FKGL], and passive sentence percentages) from a comparative study assessing ChatGPT-4-turbo's responses to 9 questions on chemical ocular injuries. The artificial intelligence responses were collected on December 21, 2024, in Riyadh, Saudi Arabia.

	Sentences	Words	Syllables	FRES	FKGL	Passive sentences	Percentage of passive sentences
Section							
Definition	4	92	161	35.44	14.03	0	0.00
Prevalence	14	297	500	32.15	14.86	2	14.30
Causes	10	210	367	33.12	14.47	1	10.00
Preventive measures	11	241	416	31.97	14.72	1	9.10
Symptoms	3	78	129	40.24	12.16	0	0.00
Diagnosis	19	414	716	30.12	15.24	2	10.50
Treatment	9	212	354	34.59	13.98	1	11.10
Follow-up	10	235	406	31.84	14.66	1	10.00
Complications	4	99	166	35.06	13.78	0	0.00
Mean (SD)	9.33 (5.20)	208.67 (108.66)	357.22 (199.59)	33.84 (2.97)	14.21 (0.97)	0.89 (0.78)	7.22 (5.60)

Category-Specific Readability Metrics

Table 2 displays the metrics across the category groupings. The mean FRES, FKGL, and passive-voice percentages for the definition, prevalence, and causes categories were 33.57 (SD 1.69), 14.45 (SD 0.42), and 8.10% (SD 7.34%), respectively. The average scores were 34.11 (SD 5.39), 14.04 (SD 1.65), and 6.53% (SD 5.70%), respectively, for the Preventive Measures, Symptoms, and Diagnosis category. In contrast, the averages were 33.83 (SD 1.74), 14.14 (SD 0.46), and 7.03% (SD 6.12%) for the treatment, follow-up, and complications categories, respectively. These findings reveal relatively low Flesch scores, indicating text complexity, with the highest average Grade Level observed in the first category (14.45). Although the percentage of passive sentences varies across categories, each remains under 10% on average.

Table . Readability metrics by category, comparison of readability metrics by grouped question categories based on ChatGPT-4-turbo responses generated during a comparative study on December 21, 2024, in Riyadh, Saudi Arabia.

Category	Sentences	Words	Syllables	FRES[a]	FKGL[b]	Passive sentences	Percentage of passive sentences
Definition, prevalence, and causes, mean (SD)	9.33 (5.03)	199.67 (102.91)	342.67 (170.76)	33.57 (1.69)	14.45 (0.42)	1.00 (1.00)	8.10 (7.34)
Preventive measures, symptoms, and diagnosis, mean (SD)	11.00 (8.00)	244.33 (168.09)	420.33 (293.58)	34.11 (5.39)	14.04 (1.65)	1.00 (1.00)	6.53 (5.70)
Treatment, follow-up, and complications, mean (SD)	7.67 (3.22)	182.00 (72.80)	308.67 (126.32)	33.83 (1.74)	14.14 (0.46)	0.67 (0.58)	7.03 (6.12)

[a]FRES: Flesch Reading Ease Score.

[b]FKGL: Flesch–Kincaid Grade Level.

Statistical Significance

There was no statistically significant difference in FRES (P=.38), FKGL (P=.55), or the percentage of passive sentences (P=.60) across the categories, as determined by one-way ANOVA. This indicates that any observed variations in readability measures were likely attributable to random fluctuations rather than systematic differences.

Trend Analysis

A trend analysis showed that (Figure 16) the FRES was consistently within the low-to-mid-30s range for all categories, suggesting challenging reading material. The FKGL similarly remains around 14, underscoring the advanced reading level required to analyze the text. Minor variations in passive-voice use (ranging from 6.53% to 8.10%) indicate some stylistic

XSL•FO
RenderX

differences. Overall, the readability was relatively constant across the 3 categories.

Figure 16. Average score for FKGL, FRES, and percentage of passive sentences across 3 question categories, based on ChatGPT-4-turbo responses about chemical eye injuries. A comparative study was conducted on December 21, 2024, in Riyadh, Saudi Arabia. FKGL: Flesch–Kincaid Grade Level;

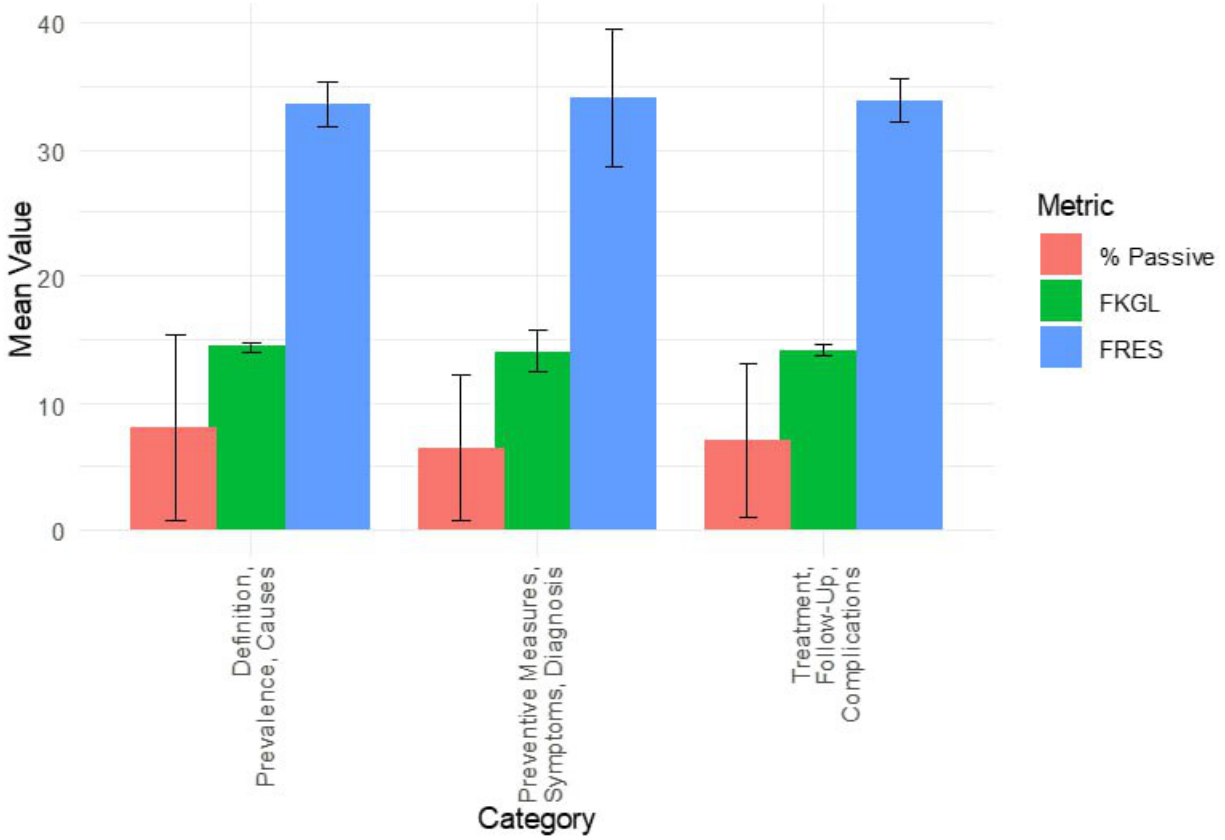

Discussion

Principal Findings

This study assessed ChatGPT's performance in providing information about chemical ocular injuries. ChatGPT provided accurate responses to most queries but missed critical details such as the urgency of care and specific management protocols. The average FRES was 33.84 (± SD 0.28), and the FKGL was 14.21 (SD 0.22), indicating a high complexity level suitable for college-level readers. Statistical analysis using ANOVA showed no significant differences in readability scores across content categories (*P*=.38 for FRES, *P*=.55 for FKGL).

Limitations

Despite its strengths, some limitations are also associated with ChatGPT. Most notably, it sometimes misses the inclusion of very important information in its responses, which is crucial for appropriate management. These issues support concerns that ChatGPT may sometimes give incomplete or biased answers, which could affect patient care in a harmful way [15]. In addition, ChatGPT cannot adjust its answers based on the user's situation, like a patient's history, how serious the condition is, or personal needs. This was seen in patient education, where it could not change its response depending on how simple or complex the medical case was [16].

Comparison With Prior Work

ChatGPT responded to all queries about chemical ocular injuries. These included questions related to definition, prevalence, causes, preventive measures, symptoms, diagnosis, treatment, follow-up, and complications. The responses were mostly accurate and thematically relevant. They adhered closely to the sentence structure and grammar. Nonetheless, certain details were overlooked. They include the urgency of care in managing chemical ocular injuries and the need for exact classification systems and management protocols based on *ICD-9* and *ICD-10* guidelines. These omissions align with findings from similar studies. They highlight ChatGPT's tendency to provide generalized but incomplete responses, reducing its clinical utility. ChatGPT accurately described the basic etiology and treatment steps but failed to include important information, such as the recommended preirrigation instructions or specific surgical interventions, which are needed in case of severe injuries. These results are consistent with those reported in another study, which evaluated the performance of ChatGPT in ophthalmology-related fields. The results showed that ChatGPT performed correctly on 46% to 58% of questions across categories. The best performance was reported in general medicine, while the poorest was reported in areas related to retina and vitreous topics. The results further support those from this study and indicate that ChatGPT provided accurate yet generalized information and often missed essential details [17].

XSL•FO
RenderX

In general, the responses were focused on the central concept of each question. There was minimal deviation from the topics. The responses related to prevalence provided statistics and global variations that were thematically relevant, although they were partially incorrect when compared with *ICD-9*/*ICD-10* standards. Thematic relevance was maintained in other topics, such as the description of diagnostic tools and complications. Nonetheless, ChatGPT introduced tangential or unrelated information, which reduced the overall coherence and thematic focus. Such a finding was also reported in prior studies that assessed the performance of AI models in medical domains [18].

The FRES and FKGL were used to assess the readability of the responses regarding chemical ocular injuries. Results showed a moderately low readability score, which suggests that the output was aligned with academic or professional standards and can limit accessibility for individuals with low levels of health literacy. The previous findings are consistent with findings from studies that evaluated the performance of ChatGPT on medical licensing examinations. The study showed that the responses were clear but overly complex for lay audiences [19]. Interestingly, there was no statistically significant difference in readability scores between the different content categories, which indicates that ChatGPT maintains a consistent language complexity regardless of the inherent difficulty. Basic information about the definition or prevalence of chemical ocular injuries was presented at the same level of complexity as more complex topics, such as treatment and surgical procedures. The same findings were reported in a meta-analysis of studies that evaluated the performance of ChatGPT in board examinations. The meta-analysis concluded that AI provided the same level of complexity across question types, irrespective of whether the questions were simple or advanced [20]. In a study assessing the performance of ChatGPT on radiology board-style examination questions, the AI demonstrated minimal variability in the complexity of responses, which indicates a uniformity in language complexity regardless of the level of difficulty of the questions [21].

While this may be beneficial to achieve uniformity, it underlines the deficiency of ChatGPT in presenting variability in speech, depending on the audience it addresses. It is desirable that simpler concepts, such as causes and symptoms, should be presented in more accessible terms, while advanced topics like surgical interventions may justifiably require technical language. This rigidity in adaptability is a finding similarly drawn from studies in other medical contexts where the responses of ChatGPT lacked subtlety in adjusting to the diversity of user groups [22]. However, readability metrics do not account for health literacy, which includes knowledge of medical terms. Even simple text can confuse a user if it includes unexplained medical jargon or lacks explanations. Besides, readability assessments do not address the accuracy of the information. This is a critical factor for patient safety and effective decision-making [19].

In this regard, ChatGPT has proven to be an exceptionally good tool for dissemination regarding chemical ocular injuries. The strong points assure that it provides structured, comprehensive answers that are logically organized, even for complex topics, in an intelligible and easy-to-understand way. Moreover, responses from ChatGPT come rather fast, allowing one to access information quickly—in some cases, this speed can be crucial in emergency situations where quick reference material is required. Recent studies, such as the work by Kianian et al [23], highlight ChatGPT's potential to improve the accessibility of complex medical information. In their study, ChatGPT reduced the FKGL of 12 peer-reviewed uveitis papers from an average of 15th grade to 7th grade, making the content more accessible to the general public. This simplification was achieved without compromising the scientific accuracy of the information, as assessed by expert ophthalmologists [23]. Another study by Kianian et al [24] explored how ChatGPT can assist in educating patients about surgical management in glaucoma. Despite prompts aimed at generating content at a 6th-grade reading level, the resulting materials maintained a higher reading level (average 11th grade) and were deemed of "fair" quality by the DISCERN instrument. This suggests that while ChatGPT can assist in developing educational content, it may require further refinement to meet the readability needs of the average patient [24].

This is also in line with observations that ChatGPT can make information more accessible. It also leads to tailored, opportune responses, especially in medical and educational domains [25]. This also makes ChatGPT a resource for users looking for initial guidance on chemical ocular injuries, given its capability for fact retrieval and the integration of information from various topics. For instance, in medical education, ChatGPT has been noted to enhance personalized learning, improve clinical reasoning, and facilitate a better understanding of complex concepts. This is consistent with its ability to explain complex medical concepts related to ocular injuries [26].

While the complexity of language used by ChatGPT may be useful for professionals, it can be difficult for patients with lower literacy levels to understand. This mismatch in the levels of language complexity with users' literacy is still far from optimization, according to the studies reviewed, because such an AI application could unconsciously exclude an audience that does not have a high educational background. ChatGPT's dual nature presents both challenges and opportunities. This requires further development and control to make it more useful and reduce risks. The concept of health literacy has to be taken into consideration while educating the masses about chemical ocular injuries. Studies show that only a small percentage of adults possess proficient health literacy. This makes communicating complex medical information a significant challenge [25]. Most users struggle to read and comprehend complex medical terminology or instructions for treatment or procedures. Our findings indicate the need for ChatGPT's content to be as readable as possible.

Future versions of AI-based platforms may adjust language complexity based on user comprehension. This could be achieved by incorporating features such as preliminary assessments or interactive dialogue to gauge the user's understanding and tailor responses accordingly [26]. Additionally, integrating visual aids, such as step-by-step infographics or flowcharts, could enhance comprehension, particularly for inherently complex topics like chemical burn

XSL•FO
RenderX

management and surgical interventions [16]. However, this study has certain limitations. Data collection was performed at one point in time within this dynamic field of AI. Furthermore, as in any qualitative research study, investigator bias might occur. In addition, the wording of the input questions can introduce some influence on the results. Future studies can overcome some of these limitations by comparing ChatGPT with other models to place it within a broader perspective in terms of its strengths and weaknesses. Other studies may wish to examine the extent to which ChatGPT is consistent and reliable across successive responses and even when asked the same question in different ways [15,27]. This might allow some insights into aspects of the capability of AI platforms to act as a reliable tool in both medical and patient education.

Conclusions

ChatGPT shows significant potential in providing information about chemical ocular injuries, offering factually accurate and relevant responses that are generally clear and well-constructed. However, its language complexity may sometimes make information less accessible to individuals with lower health literacy. Additionally, the responses sometimes lack key information relevant to treatment, such as the urgency when dealing with chemical burns. Although ChatGPT can serve as a valuable supplementary tool for patients and health care professionals, it should not replace professional medical advice, as some responses may not align with clinical practice or address the needs of patients with different levels of education. Future iterations should aim to enhance readability, improve context-specific accuracy, and better tailor responses to user needs and literacy levels. This highlights the importance of ensuring AI tools like ChatGPT provide safe and accurate responses to support patient care and reduce the risk of misinformation.

Acknowledgments

This work was supported and funded by the Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University (IMSIU), (grant number IMSIU-DDRSP2501). Generative artificial intelligence tools were not used to write any part of this manuscript independently. All content was reviewed and validated by the authors to ensure accuracy and integrity.

Data Availability

All data generated or analyzed during this study are included in this published article. The data consist of responses generated by ChatGPT-4 Turbo to 9 predefined questions related to chemical ocular injury. No patient data were used in this study.

Conflicts of Interest

None declared.

References

1. Morahan-Martin JM. How internet users find, evaluate, and use online health information: a cross-cultural review. Cyberpsychol Behav 2004 Oct;7(5):497-510. [doi: 10.1089/cpb.2004.7.497]
2. AlGhamdi KM, Moussa NA. Internet use by the public to search for health-related information. Int J Med Inform 2012 Jun;81(6):363-373. [doi: 10.1016/j.ijmedinf.2011.12.004] [Medline: 22217800]
3. de Boer MJ, Versteegen GJ, van Wijhe M. Patients' use of the internet for pain-related medical information. Patient Educ Couns 2007 Sep;68(1):86-97. [doi: 10.1016/j.pec.2007.05.012] [Medline: 17590563]
4. Briganti G. Artificial intelligence: an introduction for clinicians. Rev Mal Respir 2023 Apr;40(4):308-313. [doi: 10.1016/j.rmr.2023.02.005] [Medline: 36894376]
5. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel). 2023. URL: https://doi.org/10.3390/healthcare11060887 [accessed 2025-08-07]
6. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023 Aug;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
7. Vallance BC. ChatGPT: new AI chatbot has everyone talking to it. BBC. 2022 Dec 7. URL: https://www.bbc.com/news/technology-63861322 [accessed 2025-08-07]
8. Ting DSJ, Tan TF, Ting DSW. ChatGPT in ophthalmology: the dawn of a new era? Eye (Lond) 2024 Jan;38(1):4-7. [doi: 10.1038/s41433-023-02619-4] [Medline: 37369764]
9. One in two EU citizens look for health information online. Eurostat. 2021 Apr 6. URL: https://ec.europa.eu/eurostat/web/products-eurostat-news/-/edn-20210406-1 [accessed 2025-08-07]
10. Finney Rutten LJ, Blake KD, Greenberg-Worisek AJ, Allen SV, Moser RP, Hesse BW. Online health information seeking among US adults: measuring progress toward a healthy people 2020 objective. Public Health Rep 2019;134(6):617-625. [doi: 10.1177/0033354919874074] [Medline: 31513756]
11. Walker HL, Ghani S, Kuemmerli C, et al. Reliability of medical information provided by ChatGPT: assessment against clinical guidelines and patient information quality instrument. J Med Internet Res 2023 Jun 30;25:e47479. [doi: 10.2196/47479] [Medline: 37389908]

XSL•FO
RenderX

12. Ting DSW, Pasquale LR, Peng L, et al. Artificial intelligence and deep learning in ophthalmology. Br J Ophthalmol 2019 Feb;103(2):167-175. [doi: 10.1136/bjophthalmol-2018-313173] [Medline: 30361278]
13. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng 2018 Oct;2(10):719-731. [doi: 10.1038/s41551-018-0305-z] [Medline: 31015651]
14. Strzalkowski P, Strzalkowska A, Chhablani J, et al. Evaluation of the accuracy and readability of ChatGPT-4 and Google Gemini in providing information on retinal detachment: a multicenter expert comparative study. Int J Retina Vitreous 2024 Sep 2;10(1):61. [doi: 10.1186/s40942-024-00579-9] [Medline: 39223678]
15. Wu Y, Zheng Y, Feng B, Yang Y, Kang K, Zhao A. Embracing ChatGPT for medical education: exploring its impact on doctors and medical students. JMIR Med Educ 2024 Apr 10;10:e52483. [doi: 10.2196/52483] [Medline: 38598263]
16. Halaseh FF, Yang JS, Danza CN, Halaseh R, Spiegelman L. ChatGPT's role in improving education among patients seeking emergency medical treatment. West J Emerg Med 2024 Sep;25(5):845-855. [doi: 10.5811/westjem.18650] [Medline: 39319818]
17. Mihalache A, Popovic MM, Muni RH. Performance of an artificial intelligence chatbot in ophthalmic knowledge assessment. JAMA Ophthalmol 2023 Jun 1;141(6):589-597. [doi: 10.1001/jamaophthalmol.2023.1144] [Medline: 37103928]
18. Cappellani F, Card KR, Shields CL, Pulido JS, Haller JA. Reliability and accuracy of artificial intelligence ChatGPT in providing information on ophthalmic diseases and management to patients. Eye (Lond) 2024 May;38(7):1368-1373. [doi: 10.1038/s41433-023-02906-0] [Medline: 38245622]
19. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? the implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 8;9:e45312. [doi: 10.2196/45312] [Medline: 36753318]
20. Liu M, Okuhara T, Chang X, et al. Performance of ChatGPT across different versions in medical licensing examinations worldwide: systematic review and meta-analysis. J Med Internet Res 2024 Jul 25;26:e60807. [doi: 10.2196/60807] [Medline: 39052324]
21. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations. Radiology 2023 Jun;307(5):e230582. [doi: 10.1148/radiol.230582] [Medline: 37191485]
22. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
23. Kianian R, Sun D, Rojas-Carabali W, Agrawal R, Tsui E. Large language models may help patients understand peer-reviewed scientific articles about ophthalmology: development and usability study. J Med Internet Res 2024 Dec 24;26:e59843. [doi: 10.2196/59843] [Medline: 39719077]
24. Kianian R, Sun D, Giaconi J. Can ChatGPT aid clinicians in educating patients on the surgical management of glaucoma? J Glaucoma 2024 Feb 1;33(2):94-100. [doi: 10.1097/IJG.0000000000002338] [Medline: 38031276]
25. Farrokhnia M, Banihashem SK, Noroozi O, Wals A. A SWOT analysis of ChatGPT: implications for educational practice and research. Innovations in Education and Teaching International 2024 May 3;61(3):460-474. [doi: 10.1080/14703297.2023.2195846]
26. Sallam M, Salim NA, Barakat M, Al-Tammemi AB. ChatGPT applications in medical, dental, pharmacy, and public health education: a descriptive study highlighting the advantages and limitations. Narra J 2023 Apr;3(1):e103. [doi: 10.52225/narra.v3i1.103] [Medline: 38450035]
27. Gödde D, Nöhl S, Wolf C, et al. A SWOT (strengths, weaknesses, opportunities, and threats) analysis of ChatGPT in the medical literature: concise review. J Med Internet Res 2023 Nov 16;25:e49368. [doi: 10.2196/49368] [Medline: 37865883]

Abbreviations

AI: artificial intelligence
FKGL: Flesch–Kincaid Grade Level
FRES: Flesch Reading Ease Score
ICD: International Classification of Diseases

Edited by A Mavragani; submitted 08.03.25; peer-reviewed by A Rouhi, S Mohanadas; revised version received 29.06.25; accepted 30.06.25; published 13.08.25.

Please cite as:
Alharbi LY, Alrashoud RR, Alotaibi BS, Al Dera AM, Alajlan RS, AlHuthail RR, Alessa DI
Using Artificial Intelligence ChatGPT to Access Medical Information About Chemical Eye Injuries: Comparative Study
JMIR Form Res 2025;9:e73642
URL: https://formative.jmir.org/2025/1/e73642
doi:10.2196/73642

XSL•FO
RenderX

© Layan Yousef Alharbi, Rema Rashed Alrashoud, Bader Shabib Alotaibi, Abdulaziz Meshal Al Dera, Raghad Saleh Alajlan, Reem Rashed AlHuthail, Dalal Ibrahim Alessa. Originally published in JMIR Formative Research (https://formative.jmir.org), 13.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Promoting Dairy Consumption Among Families: Development and User Experience Study of a Web-Based Nutrition Intervention

Juliette Lemay[1,2,3,4], MSc; Jacynthe Roberge[5], PhD; Véronique Provencher[2,3,6], PhD; Angelo Tremblay[1,2,7], PhD; Shirin Panahi[1,2,7], PhD; Raphaëlle Jacob[1,2,3], PhD; Lucie Brunelle[7], MBA, MSc; Gabrielle Saintonge[5], MA; Vicky Drapeau[1,2,6,7], PhD

[1]Quebec Heart and Lung Institute Research Center, Laval University, Quebec, QC, Canada

[2]Institute of Nutrition and Functional Foods, Université Laval, Quebec, QC, Canada

[3]School of Nutrition, Faculty of Agricultural and Food Sciences, Laval University, Quebec, QC, Canada

[4]Institut universitaire de cardiologie et de pneumologie de Québec, Laval University, Quebec, QC, Canada

[5]School of Desing, Laval University, Quebec, QC, Canada

[6]Centre Nutrition, santé et société, Laval University, Quebec, QC, Canada

[7]Department of Kinesiology, Faculty of Medecine, Laval University, Pavillon de l'Éducation Physique et des Sports, 2300, rue de la Terrasse, Quebec, QC, Canada

Corresponding Author:
Vicky Drapeau, PhD
Quebec Heart and Lung Institute Research Center, Laval University, Quebec, QC, Canada

Abstract

Background: Insufficient adherence to dietary guidelines underscores the need for effective interventions promoting healthy eating, including dairy consumption, among Canadian families. Research suggests that web-based interventions grounded in user research and behavior change theories can effectively support dietary improvements. However, few theory-driven digital interventions specifically target dairy consumption in families.

Objective: This study aims to describe the development of a web-based nutrition intervention, Dairyathlon, designed to promote dairy consumption among families using the IDEAS (Ideate, Design, Assess, and Share) framework. In addition, it evaluates user experience (UX) with the web-based platform.

Methods: Following the IDEAS framework, family perspectives and beliefs regarding dairy consumption were explored through ethnographic research and interviews. Behavior change techniques, based on the theory of planned behavior, were integrated to enhance attitudes and perceived behavioral control toward dairy intake. These techniques underwent iterative design, prototype testing, and refinement. UX was assessed with the AttrakDiff questionnaire, comparing families using Dairyathlon to those using the Canadian Food Guide (CFG). Children and parents completed the questionnaire after the presentation of the platform (PRE) and following 8 weeks of use (POST). AttrakDiff evaluates pragmatic quality (PQ), hedonic stimulation (HSQ), hedonic identity (HIQ), and attractiveness dimension (ATT) on a scale from –3 to +3, with >1 considered optimal, 0–1 acceptable, and < 0 suboptimal.

Results: Between April 2019 and August 2020, Dairyathlon was developed to enhance families' attitudes and perceived control over dairy consumption, adhering to the IDEAS framework. Users' experience assessments were conducted among 29 families and showed significantly higher scores for Dairyathlon compared to the reference platform (CFG) at both pre- and postassessments (P<.001). Although both platforms were initially rated as optimal, UX ratings decreased after use: PRE (1.7, SD 0.6) to POST (1.4, SD 0.8) in the Dairyathlon group (mean difference of 0.4, 95% CI 0.2-0.7; P=.002), and (1.4, SD 0.6) to (0.9, SD 0.6) in the CFG group (mean difference = 0.6, 95% CI 0.5-0.6; P<.001). After using Dairyathlon, children (n=45) rated all UX dimensions as optimal, with scores of PQ (1.4, SD 1.0), HSQ (1.6, SD 1.0), HIQ (1.4, SD 1.1), and ATT (1.7, SD 0.9). Parents (n=50) also rated most dimensions as optimal, with scores of 1.2 (SD 1.0) for PQ, 1.4 (SD 0.8) for HIQ, and 1.6 (SD 0.8) for ATT. However, the HSQ dimension received a slightly lower rating of 0.9 (SD 0.8), indicating a need for improvement in adult stimulation.

Conclusions: This study highlights the effectiveness of the IDEAS framework in developing a web-based intervention to promote dairy consumption. The Dairyathlon platform's UX was rated as optimal, especially for visual attractiveness, though the stimulation dimension requires improvement for adults. Future research will evaluate its impact on dairy consumption, diet quality, and family health status.

Trial Registration: ClinicalTrials.gov NCT05417347; https://clinicaltrials.gov/study/NCT05417347

(JMIR Form Res 2025;9:e66582) doi:10.2196/66582

XSL•FO
RenderX

KEYWORDS

IDEAS framework; eHealth; dairy product; family; children; user appreciation; user experience; nutrition; consumption; RCT; randomized; controlled trial; dietary; Canada; behavior change; IDEAS; dairy products; Dairyathlon; ethnographic; interview; planned behavior; questionnaire; diet quality

Introduction

Dairy products are known for their specific nutrient content, including protein, calcium, vitamin D, and magnesium, which have well-documented benefits for bone health and growth [1,2]. In addition to high nutrient content, diets rich in dairy products, notably milk, yogurt, and cheese, are known to play a role in the management of body weight and fat composition [3,4]. In one of our studies, an increase in milk consumption (and fruit) was associated with less body weight and fat gain over time [5]. Furthermore, dairy products have a specific food matrix, particularly yogurt, which has been shown to enhance satiety and reduce energy intake in short-term clinical trials [6]. Epidemiological evidence also links regular dairy consumption to reduced risk of chronic diseases [7,8]. According to a study on the Canadian Community Health Survey 2015 cohort, many Canadians do not meet vitamin D, magnesium, and calcium recommendations, all of which are dairy-related nutrients [9]. This portrait reflects the situation before the Canada's Food Guide (CFG) change in 2019. Since the release of the new CFG, where dairies are no longer presented as a specific category [10], an analysis of nutrient intake provided by a dietary pattern according to the 2019 CFG guidelines suggests a high risk of inadequate intake of calcium and vitamin D [11]. Another study found that the more Canadians adhere to the 2019 CFG, the more they are at risk of calcium and vitamin D deficiencies [12]. Given their significance in growth, nutrient intake, disease prevention, and possibly obesity management, promoting dairy in nutrition programs is relevant, especially since the 2019 CFG no longer distinguishes dairy as a separate group.

Eating habits are acquired early in life. Parents play a pivotal role in teaching healthy eating habits to their children [13]. The family environment, including food availability and parental eating behaviors, influences children's eating habits [14]. Family-based behavioral interventions have been found effective in fostering the adoption and sustained practice of healthier eating habits [15]. In the context where dairy consumption and healthy eating behaviors are important in children and adults, family intervention should be considered.

Using technology in health promotion is a cost-effective solution that reaches diverse populations [16]. Once implemented, it requires minimal upkeep and can be widely distributed. Grounded in behavior and design theories, web-based interventions offer a promising avenue for promoting healthy family habits [17]. In 2009, our research team developed Team Nutriathlon, a web-based nutrition challenge for schools [18,19], later adapted in 2016 as Family Nutriathlon to encourage fruit, vegetable, and dairy consumption in families [20]. This adaptation successfully increased dairy consumption in the short term, demonstrating the potential of web-based approaches [20]. Building on these findings, this study refines and expands the Family Nutriathlon approach through Dairyathlon, a web-based intervention specifically targeting dairy consumption. Key engagement strategies and family-oriented content from Family Nutriathlon were integrated to enhance Dairyathlon's effectiveness and sustainability.

To optimize Dairyathlon's design, we used the Integrate, Design, Assess, and Share (IDEAS) framework, a validated model for developing and adapting web-based health interventions [21]. This structured approach integrates behavior change theory and has been applied successfully in physical activity promotion [22] and health care programs [23]. Given our goal to build on Family Nutriathlon while improving efficacy and usability, IDEAS provided a systematic and efficient framework to guide development. This ensured that Dairyathlon was tailored to family needs while maximizing engagement and impact.

This study aims to describe the development process of a web-based nutrition intervention to promote dairy products consumption among families using the IDEAS framework and to evaluate users' appreciation of the web-based platform compared to a reference nutrition platform. The hypotheses are: (1) designing the platform based on the IDEAS theoretical framework will lead to a web-based platform that users highly appreciate and (2) both children and parents are expected to find the platform more enjoyable and useful than the reference nutrition platform, as it is specifically designed to address the needs of Quebec families.

Methods

Overview

This study adapted the Family Nutriathlon web-based intervention [20], developed to promote healthy eating among families. It was selected for adaptation because of its beneficial effect on dairy intakes in the short term, its similar targeted users, the inclusion of health behavior change techniques (ie, self-determination theory), and its focus on specific food groups such as dairy consumption. The adaptation process included 3 phases inspired by the IDEAS framework. In phase 1, the focus was on integrating the target users and specifying the target behavior. Phase 2 involves the design process in an iterative process (ie, designing the user interface and experience and creating features of the platform). Phase 3 assessed the user experience (UX) of the platform's design (see Figure 1).

XSL•FO
RenderX

Figure 1. Overview of the Dairyathlon development process: timeline, phases, and human resources. UX: user experience.

Phase 1: Integration of User Needs and Behavioral Theories

Phase 1 of the IDEAS framework sought to gather insights from the target users, specify the target behavior, and ground in behavioral theory [21]. As specified by Curtis et al [24], the target population should be included in all steps of the development process. In this study, our target population included families with children aged between 8 and 16 years. The age of 8 years marks the minimum age at which children can reliably complete questionnaires and navigate on the web platform, while 16 years is the typical upper limit for parental influence on food choices. In addition, this age range aligns with calcium needs during growth and was effective in our previous Nutriathlon study [20]. To better understand the targeted population's thoughts, opinions, and experiences, an ethnographic process was undertaken [25]. First, a 27-item questionnaire was created to explore families' activities, dairy consumption, grocery shopping habits, cooking, meal planning, time availability, perception of clinical studies, and use of electronic devices. This questionnaire was shared on Facebook (Meta Platforms) groups targeting parents living in Quebec, Canada. The questionnaire was designed by the UX design team in collaboration with the nutritionists from the research team. The questions were not validated before being sent to the target population. The survey was voluntary and open to all site visitors, but only participants with at least one child between 8 and 16 years old were retained for analysis. No incentives were offered. In May 2019, for 1 week, a total of 327 families completed the internet-based survey. Second, to complement the quantitative results, we conducted 7 interviews among the families who agreed to be contacted, to explore their needs, eating habits, and motivations toward dairy consumption. Based on the results from the questionnaire and the interviews, two family profiles were established: (1) A young family with children (8 - 12 y): they cook on weekends, plan meals, maintain a well-stocked kitchen, and aim to expand their nutrition knowledge. They consume a lot of milk and moderately consume yogurt and cheese. (2) A family with teenagers (13 - 16 y): teens participate in meal preparation and grocery, meal choices vary among family members, and their motivation is driven by competition and rewards. They prefer cheese and consume less yogurt and milk. These profiles were considered during the next steps (ie, the design process) to ensure that the intervention is specifically created for the target user and that it is based on the family's needs.

Second, we specified the target behavior. Since we aimed to develop a web-based nutrition intervention to increase dairy consumption, we defined the target behavior as consuming 4 dairy servings per day (ie, 250 ml of milk, 50 g of cheese, and 2 servings of 175 ml of yogurt). This quantity provides 1200 mg of calcium, which has been shown to improve appetite and body weight control in adults and children [26]. All flavored types, including sweetened and plain, were permitted to enhance yogurt acceptability. Since the family interviews revealed that they were less structured and usually overloaded on weekends, only weekdays (ie, Monday through Friday) were targeted for the intervention to accommodate families.

Finally, the next step was to ground in behavior theories. The theory of planned behavior (TPB), which has emerged as particularly effective in web intervention on healthy lifestyle habits aimed at changing behavior [27], was used to target 2 determinants of behavioral change: attitude and perceived behavioral control toward dairy consumption. From each determinant, salient attitude beliefs (ie, advantages and disadvantages of consuming dairies), and salient control beliefs (ie, facilitators and barriers to dairy consumption) were assessed [28]. First, parent's salient beliefs related to dairy consumption were identified based on a qualitative study from Lacroix et al [29], where beliefs about dairy products were gathered during focus group sessions among 161 men and women aged 16 to 50 years in the Quebec City, Montreal, and Toronto regions. Second, in the absence of existing literature to identify salient beliefs in children, our research team developed a questionnaire based on the key elements from the study by Lacroix et al [29]. In December 2019, the 18-item questionnaire was sent by email to a family mailing list from Université Laval. Parents who received the questionnaire were informed of its objective: to identify the key beliefs related to dairy consumption among children aged 8 to 16 years. The first question asked about the child's age, and only responses from children within this age range were accepted. No financial compensation was provided. A total of 87 children in the Quebec City area completed the questionnaire. Attitude and perceived control about dairy consumption were then targeted with a specific behavior change technique (BCT). These BCTs were selected from the taxonomy by Michie et al [30], such as pros and cons, health consequences,

XSL•FO
RenderX

instruction on how to perform a behavior, and social comparison, to favor a positive attitude and increase perceived behavioral control toward dairy consumption. In addition, motivational aspects, inspired by Robinson [31], were incorporated, such as gamification [32], goals and visual representations of the goals [33], a leaderboard, and digital trophies (eg, badges displayed on the intervention platform) [25]. A challenge between families was added in the web-based intervention, where families could earn points and symbolic prizes by meeting dairy consumption targets or trying some of the BCTs available on the web-based platform, such as true or false quizzes, tips and advice section, and recipes. The leaderboard, one important gamification feature of the intervention, recognized high-achieving families anonymously, promoting healthy competition and motivation [25]. The salient beliefs linked to the selected BCTs are outlined in Multimedia Appendix 1.

Phase 2: Iterative Design of the Web Platform

After identifying the BCTs for the features of the Dairyathlon platform, a collaborative effort involving nutrition researchers, the UX design team, full-stack developer, and the registered dietitian team was undertaken in an iterative design process. Phase 2 involved a human-centered design process focusing on nutrition challenge and education content, the conceptualization of the platform, and usability tests.

Initially, the education information content, based on the salient beliefs identified in the first phase, was developed and validated by a team of dietitians, ensuring that the literacy level was appropriate for children aged 8 to 16 years. An informal review of the content was conducted with a group of 3 children and 3 parents, who provided qualitative feedback on clarity of the content. Based on this feedback, adjustments were made to improve understanding.

The UX design team prototyped the web-based intervention, starting with wireframes to visualize the platform. After refining these with the research team, flowcharts were created to map out user navigation. With the help of a full-stack developer, the first versions of the intervention were developed for both mobile and computer formats.

A total of 2 rounds of usability testing were conducted to refine the platform. The A/B test compared 2 versions of the intervention, asking users to indicate their preferences (eg, whether a data table with numerical values or color-coding was more effective). The A/B test involved observing users performing tasks on the platform, guided by an audio scenario, and evaluating success, partial success, or failure for each task. Feedback from the tests, including task completion time and user satisfaction, helped identify areas for improvement.

Only 1 round of testing was made, involving 5 participants (3 adults and 2 children), and adjustments were made based on feedback, which was key in improving usability. The final version of the Dairyathlon platform was developed in collaboration with a full-stack developer, and project management tools (eg, DevOps and Microsoft Azure) ensured efficient tracking of the project's timeline and features. This platform was designed to be highly flexible and easily adaptable. For example, it allows for the addition of new food groups or modifications to better accommodate different target audiences and populations. The platform is hosted on Université Laval's secure servers, allowing for the simultaneous registration of multiple families. It is accessible across computers, tablets, and mobile phones. Security measures include password authentication, SSL encryption, and real-time access monitoring. The technical architecture and security of the platform are summarized in Multimedia Appendix 2.

Phase 3: Users Experience Assessment

The third phase of the study focused on assessing the UX of the Dairyathlon platform among families. This assessment was carried out using the AttrakDiff questionnaire before and after an 8-week intervention period. Following sections present the UX assessment methodology.

Ethical Considerations

This study was approved by the Research Ethics Board of Université Laval (approval NCT05417347). All parents and children provided written informed consent and assent, respectively, which explicitly included approval for analyses of the collected data (AttrakDiff). As part of the informed consent process, participants were also informed about data privacy and confidentiality measures. All research documents will be stored in a coded (deidentified) form for 5 years and then anonymized and retained for 25 years by the principal investigators. Participants in the intervention group received free yogurt coupons as compensation for the cost of the dairy products. No monetary compensation was offered to participating families.

Participants

Families were recruited between June 2021 and September 2023 in the Quebec City area through various channels, including Facebook family groups, email lists from Université Laval and primary schools, and public spaces like community centers. This study was single-blind, meaning that participants were unaware of whether they had been randomized to the intervention or control group. Recruitment materials described the study's objective as evaluating a family-based nutrition intervention, without specific reference to dairy products. Interested families could contact the research team via email or phone for further information and to complete an eligibility screening. The inclusion criteria were parents with children aged 8 to 16 years, with at least one parent presenting overweight (BMI ≥27 kg/m² and a waist circumference ≥88 cm for females and ≥102 cm for males). Families with food allergies or dairy restrictions were excluded.

Intervention and Control Group

Families were randomly assigned to intervention or control group for the 8-week intervention. The intervention group received a 1-hour face-to-face presentation of the Dairyathlon challenge, which involved the consumption of 4 dairy servings/day and was introduced to the Dairyathlon web-based platform. They were encouraged to use the Dairyathlon web-based platform during the 8 weeks to help them reach their dairy consumption target and participate in the challenge as a family. The control group received a 1-hour presentation on the 2019 CFG website, which provides dietary recommendations,

XSL•FO
RenderX

including the healthy plate model (half fruits and vegetables, one-quarter protein, and one-quarter whole grains) [10]. They were invited to consult the website and follow its recommendations for the 8-week period. During the 8-week intervention, all families had 4 nutritional follow-up sessions with a dietitian every 2 weeks. The dietitian checked progress toward goals and monitored the use of the Dairyathlon platform. The families' attendance at the follow-up meetings allowed us to measure the fidelity of platform use.

Measurement

UX was measured before (PRE) and after (POST) the 8-week intervention using the validated French version of the AttrakDiff questionnaire for parents and a validated short version for children [34-36]. The AttrakDiff questionnaire, developed by Hassenzahl et al [36], is designed to assess the UX of interactive products. It consists of 28 pairs of opposite items (eg, "good – bad") on a 7-point Likert scale for adults and 10 pairs for children. The questionnaire evaluates four dimensions: (1) Pragmatic quality (PQ): practicality and functionality of the platform, (2) Hedonic-stimulation quality (HSQ): joy of use and the sense of challenge, (3) Hedonic-identity quality (HIQ): Personal expression and control of one's image in a social context, and (4) Attractiveness dimension (ATT): visual appeal of the interface. Each dimension is scored from −3 (very negative) to +3 (very positive). Scores below 0 are suboptimal, scores between 0 and 1 are acceptable, and scores above 1 are optimal [36]. A global score is calculated, representing the overall UX appreciation [37]. This evaluation allows a comparison of the Dairyathlon platform to a reference platform (in this case, the CFG website). The AttrakDiff questionnaire provides insight into the strengths and weaknesses of the Dairyathlon platform, with a specific focus on differences in UX between children and parents.

Statistical Analysis

Participants were randomly assigned to the intervention or control group using a computer-generated randomization list. The randomization was performed before the start of the intervention to ensure unbiased assignment. As the statistical power required for evaluating user appreciation with the AttrakDiff questionnaire was achieved (ie, more than 20 participants), these analyses were conducted [38]. Statistical analyses were performed only in families who completed the full 8-week intervention. For the global appreciation assessment of both Dairyathlon and CFG platforms, paired *t* tests and Welch *t* tests (when assumptions were not met) were used to assess changes over time within groups (PRE vs POST in each group). An independent *t* test was conducted to compare the Dairyathlon group with the control group (CFG) at both PRE and POST. Only families who completed the full 8-week intervention were included in the analysis.

The AttrakDiff questionnaire (UX) used paired *t* tests to compare pre- and post-ratings for each dimension, assessing significant changes over time in children and parents. This analysis focused exclusively on children and parents in the Dairyathlon intervention group, as the primary goal was to assess the UX design of the platform. Data are presented as mean (SD), with statistical significance set at $P<.05$. Statistical analyses were performed using IBM SPSS 29.

Results

Visual of the Dairyathlon Platform

In May 2021, the final version of the Dairyathlon platform was launched via the Université Laval server, ready to be used by the families. Figure 2 illustrates the visual of the web-based platform, and Figure 3 shows the mobile representation.

XSL•FO
RenderX

Figure 2. Dairyathlon homepage design (Dashboard), computer format. (A) Tips and advice section; (B) True or false quizzes; (C) Compilation tool; (D) Daily value in %; (E) Recipes; (F) Leaderboard; (G) Challenge progression view with family member avatar.

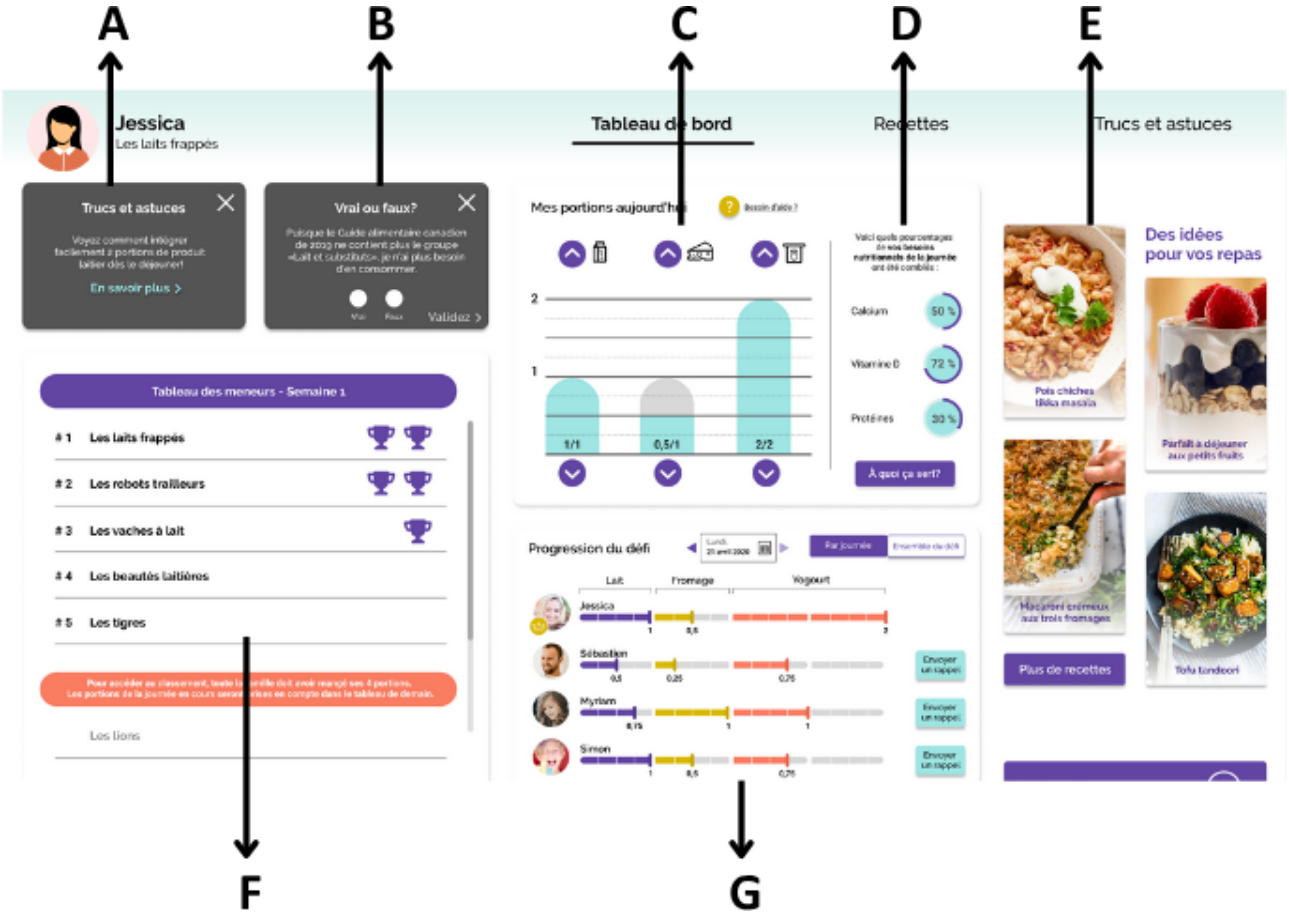

XSL•FO
RenderX

Figure 3. Mobile format of the Dairyathlon web-based intervention with true or false quizzes and the compilation tool.

XSL•FO
RenderX

User Experience

Global Appreciation

As of January 2024, a total of 44 families completed the full 8-week intervention. In the intervention group, 29 families were randomized, and 15 families were randomized in the control group. The study is expected to conclude in February 2025, with a target of 60 families. In the intervention group, 42 children and 46 parents completed the AttrakDiff questionnaire before and after the intervention, while 26 children and 28 parents in the control group. Platform usage compliance was assessed based on attendance at follow-up sessions with the dietician. Families in the intervention group attended an average of 3.5 out of 4 sessions (88%), while families in the control group attended an average of 3.8 out of 4 sessions (95%).

First, results show a significant decrease in scores from PRE to POST in both groups. In the Dairyathlon group, the mean score decreased from 1.7 (SD 0.6) at PRE to 1.4 (SD 0.8) at POST, with a mean difference of 0.4 (95% CI 0.2-0.7), significant at P=.002 (paired t test, assuming equal variances). In the control group (CFG), the mean score dropped from 1.4 (SD 0.6) to 0.9 (SD 0.6), with a mean difference of 0.6 (95% CI 0.5-0.6), significant at P<.001 (Welch-Satterthwaite t test, due to unequal variances). An independent t test revealed a significant difference between the two groups at both time points, with the Dairyathlon group scoring higher than the control group at both PRE (P<.001) and POST (P<.001). Figure 4 illustrates these results.

XSL•FO
RenderX

Figure 4. Family global appreciation score, before (PRE) and after (POST) an 8-week use of the Dairyathlon platform and the nutrition reference platform (Canada's Food Guide). Significant differences were observed before (PRE) *P*<.001, and after (POST) *P*<.001, and the green line indicates the minimal score for optimal. CFG: Canada's Food Guide

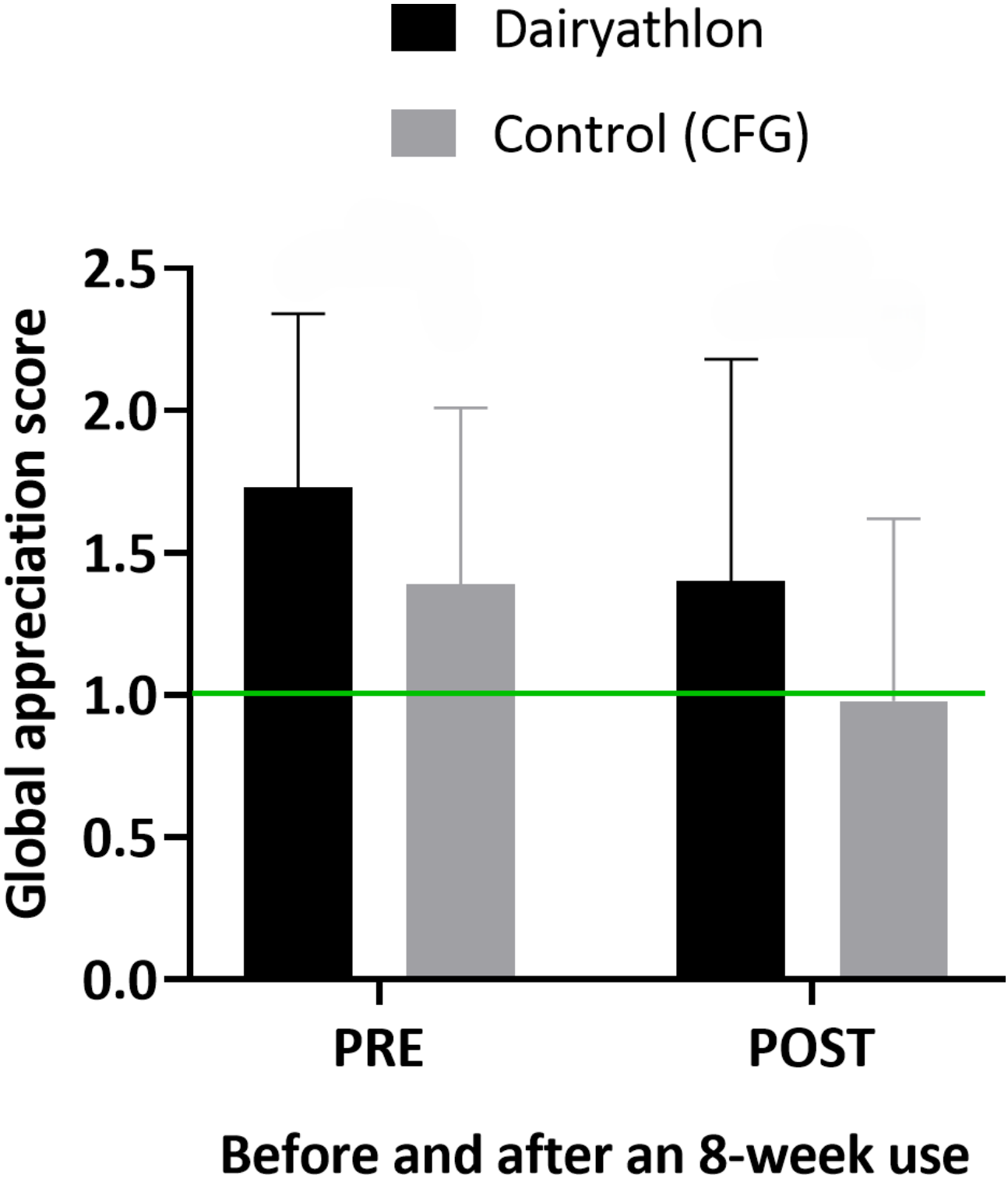

Dimensions Appreciation

In the 29 intervention families, 42 children and 46 parents completed the AttrakDiff questionnaire before and after the 8-week intervention. Among children, all 4 dimensions were rated above 1 at both time points (see Figure 5) , while parents only rated all the dimensions above 1 before the intervention (see Figure 6). After the intervention, parents rated the stimulation dimension as acceptable. For both children and parents, the highest dimension before and after the use of the Dairyathlon platform is related to its aesthetics (attractiveness).

Paired *t* tests were conducted to compare PRE and POST ratings for each dimension to determine if scores significantly decreased over time. For children, the HIQ (1.7, SD 0.9 vs 1.5, SD 0.8; *P*=.04) and ATT (2.0, SD 0.9 vs 1.7, SD 0.9; *P*=.01) dimensions showed significant decreases (see Figure 5).

XSL•FO
RenderX

For parents, significant decreases were observed in the PQ (1.6, SD 0.9 vs 1.2, SD 1.0; *P*=.004), HIQ (1.8, SD 0.9 vs 1.4, SD 1.1; *P*=.001), and ATT (2.0, SD 0.9 vs 1.6, SD 0.9; *P*<.001) dimensions (see Figure 6).

After its use, items with the highest score for children were: "Clearly structured," "Creative," and "Good," while the lowest scores were "Unpredictable" and "Dull" (see Figure 7).

For parents, the lowest items were: "Cautious" and "Undemanding," suggesting that the gamification part (hedonic-stimulation dimension) was not optimal to stimulate them (see Figure 8).

Figure 5. Children's user experience of the Dairyathlon platform, before (PRE) and after (POST) an 8-week use for pragmatic quality (PQ), hedonic-stimulation quality (HSQ), hedonic-identification quality (HIQ), and attractiveness dimension (ATT). *P*<.05 for HIQ and ATT. The green line indicates the minimal score for optimal.

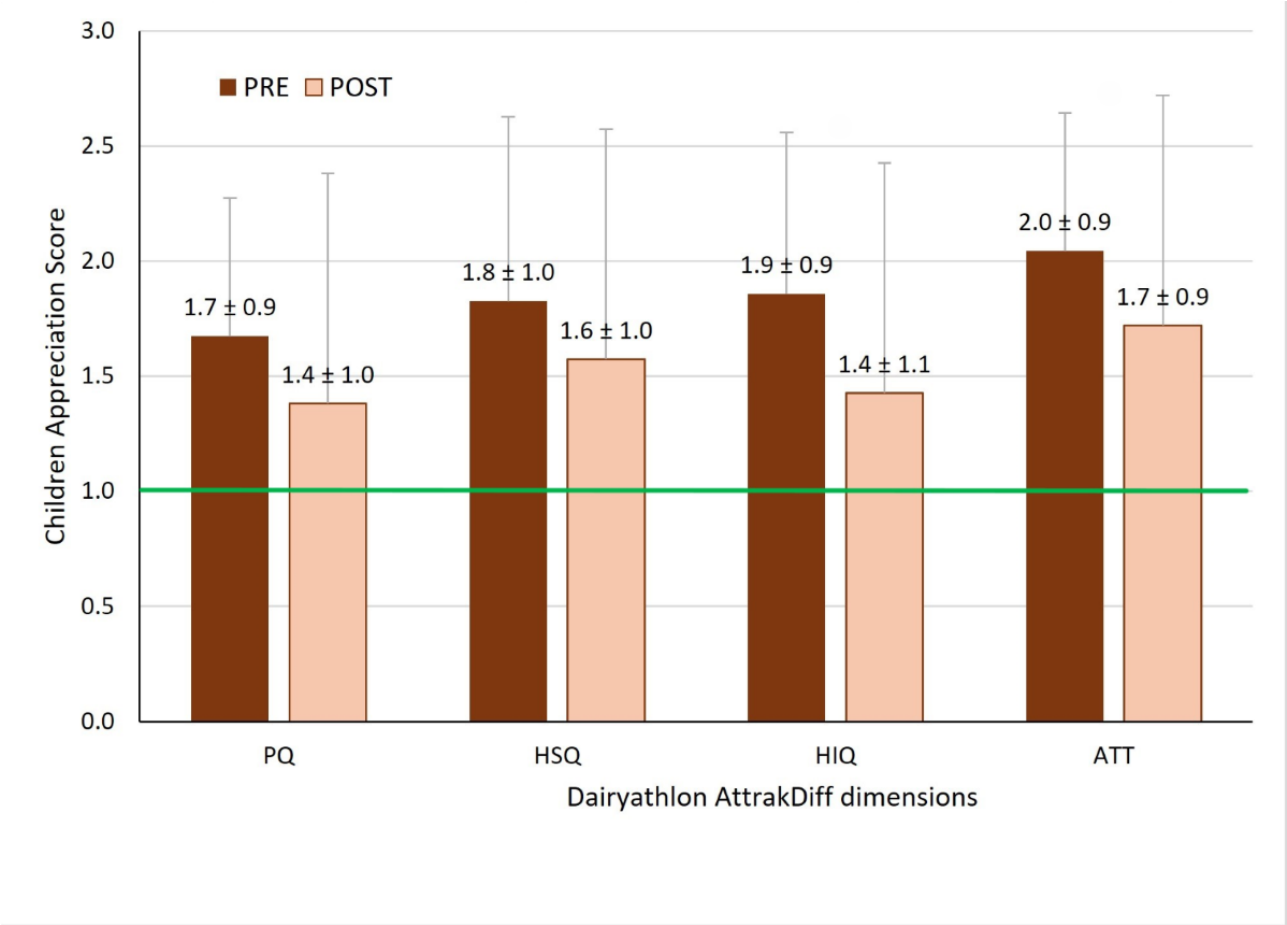

XSL•FO
RenderX

Figure 6. Parents' user experience of the Dairyathlon platform, before (PRE) and after (POST) an 8-week use for pragmatic quality (PQ), hedonic-stimulation quality (HSQ), hedonic-identification (HIQ) qualities, and attractiveness dimension (ATT). *P*<.05 for PQ, HIQ, and ATT. The green line indicates the minimal score for optimal.

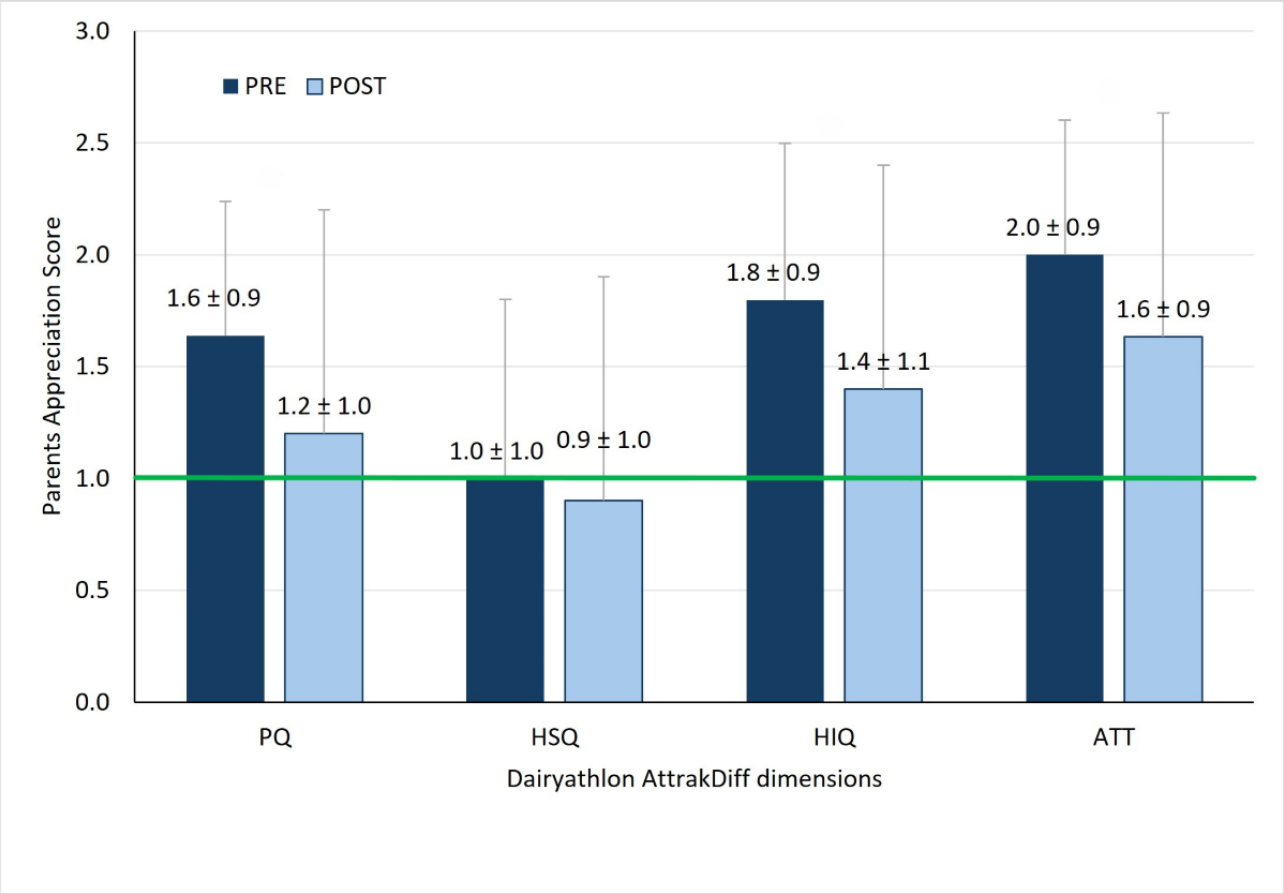

Figure 7. Children's user experience score for each item (10), after its use. Dimensions: pragmatic quality (PQ), hedonic-stimulation quality (HSQ), hedonic-identity quality (HIQ), and attractiveness dimension (ATT).

Dimension		0.0	0.5	1.0	1.5	2.0	2.5	3.0		Score
PQ	Complicated								Simple	1.4
	Impractical								Practical	1.7
	Unpredictable								Predictable	1.0
	Confusing								Clearly structured	1.9
HSQ	Unimaginative								Creative	2.0
	Dull								Captivating	1.2
HIQ	Tacky								Stylish	1.6
	Cheap								Premium	1.3
ATT	Ugly								Attractive	1.6
	Bad								Good	1.8

XSL•FO
RenderX

Figure 8. Parents' user experience score for each item (28), after its use. Dimensions: pragmatic quality (PQ), hedonic-stimulation quality (HSQ), hedonic-identity quality (HIQ), and attractiveness dimension (ATT).

Dimension		0.0	0.5	1.0	1.5	2.0	2.5	3.0		Score
PQ	Technical								Human	0.6
	Complicated								Simple	1.3
	Impractical								Practical	1.3
	Cumbersome								Straightforward	1.0
	Unpredictable								Predictable	1.3
	Confusing								Clearly structured	1.7
	Unruly								Manageable	1.3
HSQ	Conventional								Inventive	1.1
	Unimaginative								Creative	1.6
	Cautious								Bold	0.2
	Conservative								Innovative	0.9
	Dull								Captivating	1.0
	Undemanding								Challenging	-0.1
	Ordinary								Novel	1.1
HIQ	Isolating								Connective	0.9
	Unprofessional								Professional	1.5
	Tacky								Stylish	2.1
	Cheap								Premium	1.3
	Alienating								Integrating	1.2
	Separates me								Brings me closer	0.7
	Unpresentable								Presentable	2.0
ATT	Unpleasant								Pleasant	1.4
	Ugly								Attractive	2.0
	Disagreeable								Likeable	1.5
	Rejecting								Inviting	1.7
	Bad								Good	1.8
	Repelling								Appealing	1.8
	Discouraging								Motivating	1.0

Discussion

Principal Findings

This study examined the development process and UX of Dairyathlon, a web-based nutrition intervention designed to promote dairy consumption among families. Guided by the IDEAS framework, the development process began with assessing families' needs and beliefs about dairy consumption. An iterative, multidisciplinary approach integrated gamification and educational elements to enhance engagement. The Dairyathlon platform provided an optimal UX, outperforming the reference nutrition website both before and after use. While children responded positively to the platform, adults showed lower appreciation, particularly in the stimulation dimension, underscoring the challenge of designing universally appealing family-based interventions. Despite differences in engagement levels among users, the Dairyathlon shows promising potential for improving dietary habits in families. Beyond its immediate impact, investing in the platform could contribute to long-term benefits, including better bone health, a reduced risk of chronic diseases, and enhanced nutrition literacy. This study is part of

XSL•FO
RenderX

an ongoing randomized controlled trial, and future studies will assess the short-term (after an 8-week use) and long-term (4 months after) effects of the Dairyathlon platform on dairy consumption, diet quality, and weight management among families.

Grounding the intervention content in behavioral theory, in the present case the TPB, was a complex and long process due to the lack of studies in the literature on dairy consumption beliefs. In this study, a qualitative analysis was conducted on adults' beliefs about dairy, but research on children was still required. The identification of dairy consumption salient beliefs was important for targeting BCTs aimed at fostering positive attitude and decreasing barriers toward the consumption of dairy products [39]. Recognizing the salient belief in adults and children was key to identifying specific BCTs and the platform features. Supporting this approach, a meta-analysis by Webb et al [27] revealed a correlation between the use of BCTs and the effectiveness of web-based interventions. This further underscores the importance of a well-founded theoretical basis in designing impactful health promotion strategies.

The IDEAS framework ensures high-quality scientific communication and supports the development of interactive user-centered content [25]. The iterative design process within a multidisciplinary team helps create evidence-based and targeted content for users, though it may extend development timelines [40]. Despite initial and ongoing costs for software, security, and content maintenance, web-based platforms like Dairyathlon offer economic advantages over traditional interventions by reducing individual counseling time, travel, and printed materials. While few studies have applied the IDEAS framework to web-based behavior change interventions, examples in physical activity and nutrition suggest its relevance [22,41]. In a randomized controlled trial, a mobile app designed with this framework led to increased vegetable consumption among overweight adults [42]. These applications highlight the framework's adaptability for health-related behavioral interventions.

The global UX assessment using the AttrakDiff questionnaire showed that families were generally satisfied with the Dairyathlon platform. When compared to the reference nutrition platform, the CFG, the global appreciation for the Dairyathlon platform was higher before and after its use, highlighting the importance of a UX design approach in the development of web-based nutrition interventions. However, both the Dairyathlon and the reference platforms showed a decrease for user appreciation over time, which can suggest a familiarity effect [43,44]. A decline in user engagement with interactive web-based platforms appears to be common, particularly when gamified elements are involved, as they can lead to boredom if the content becomes too simplistic for sustained interest [45,46]. While long-term studies on web-based nutrition platforms are limited, research on other eHealth interventions has shown a decrease in user interest after 3 to 6 months of use [47]. To encourage longer-term engagement, future iterations of the platform could include weekly new interactive quizzes, complemented by email notifications to prompt ongoing participation, as outlined in a recent scoping review [48]. Furthermore, features such as push notifications for new recipes or challenges and a chat system with a registered dietitian could promote more regular engagement. However, time and funding constraints limited the development of these features in this study.

The assessment of each dimension before and after using the Dairyathlon platform helped identify both the strengths and weaknesses of the platform. Although evaluating UX with AttrakDiff is suggested before and after use [36], most studies assess it only at baseline [49].

First, for children, the ratings were optimal across all dimensions, which is promising. A study assessing the relationship between children's UX satisfaction and learning in a nutrition-focused video game found that UX satisfaction was positively correlated with their learning outcomes among children aged 8 to 10 years [50]. However, these results should be interpreted with caution, as validation studies of the AttrakDiff short version among children are still limited. To our knowledge, only a single study has assessed the internal consistency of the AttrakDiff dimensions in children, finding no significant associations [35], and another study has evaluated the AttrakDiff short version (10 items) among adults, reporting internal consistency [51]. More studies are needed to validate the use of the AttrakDiff short version in children further. While future research will assess the impact of the Dairyathlon intervention on dairy consumption, this study provides promising results, particularly in terms of children's satisfaction and potential learning outcomes. For parents, among the 4 UX dimensions, only hedonic stimulation was suboptimal, indicating that the gamification elements did not provide enough challenge or engagement. This suggests that the tasks and rewards may have been too simplistic or not well aligned with the parents' needs and motivations. In our study, we observed that parents tended to complete the quizzes early on, often within the first few weeks. To increase adult engagement, future interventions could introduce more structure and interactive tasks with additional educational content. Weekly quizzes increasing in difficulty levels could be introduced, with email notifications, which have been shown to increase dairy-derived calcium intake and could inform parents about new quizzes and content to boost participation [52]. Tailored features, for example, educational videos with nutrition tips or interactive quizzes for parents and children, could help maintain attention, engagement, and support learning [48,53].

Limitations

Our study had several limitations. The sample was limited to participants from the Quebec region, Canada, which may affect the generalizability of the results. The platform could be adapted for other cultural contexts using the IDEAS framework. This would first require an assessment of the target population's needs, cultural beliefs, and attitudes toward dairy consumption. Following this, both the content and the format would need to be adjusted to align as closely as possible with the target culture. As highlighted in a recent narrative review, ensuring a culturally relevant and accessible UX, through appropriate design and translation, is essential for effective adaptation [54]. Another limitation of this study is that it relied solely on quantitative evaluations of UX, which may limit the depth of result

XSL•FO
RenderX

interpretation. A qualitative study involving interviews with families who used the Dairyathlon platform would have provided valuable insights beyond the AttrakDiff score, highlighting specific features that were particularly appreciated. If the platform is adapted for broader use in the future, conducting qualitative studies with user families and organizing focus groups would be important for refining and enhancing the platform. Finally, this study did not cover all steps of the IDEAS framework, including evaluating our web-based intervention's effectiveness and impact on dairy consumption among families [21]. These last steps are part of another ongoing study by this study's present team.

Conclusion

This study outlines the development process of a web-based nutrition intervention for families using the IDEAS framework. It includes a concrete example and discusses timelines, issues, and reflections. The UX assessment of the Dairyathlon platform, using AttrakDiff, showed optimal appreciation before and after its use, surpassed the reference nutrition platform, and stimulation dimension for parents could be improved.

Acknowledgments

The authors thank Marie-Claude Roy, Lola Bourgoin, and Charlotte Brochu (research assistants) for their collaboration during the project; Rachel Bouchard and Mélissa Gaboury for their contribution in the design process; Guillaume Brassard (full-stack developer) for the web version of the platform; and Isabelle Frappier, Chloé Caredda, and Marie Cyrenne-Dussault for their help in researching and creating content.

This work was supported by the Dairy Farmers of Canada (no grant number). RJ was supported by a PhD scholarship from the Canadian Institutes of Health Research (CIHR, MFE-171302) at the time of the intervention development and is now supported by a postdoctoral fellowship from CIHR (MFE-176530).

Data Availability

The datasets generated or analyzed during this study are not publicly available due to confidentiality but are available from the corresponding author on reasonable request.

Conflicts of Interest

VD and AT have received grants from Dairy Farmers of Canada. None declared by the other authors.

Multimedia Appendix 1
Behavior change techniques (BCTs) targeting adults' and children's salient beliefs and motivation toward dairy consumption.
[DOCX File, 27 KB - formative_v9i1e66582_app1.docx]

Multimedia Appendix 2
Overview of platform architecture and security features.
[PDF File, 215 KB - formative_v9i1e66582_app2.pdf]

References

1. Huth PJ, DiRienzo DB, Miller GD. Major scientific advances with dairy foods in nutrition and health. J Dairy Sci 2006 Apr;89(4):1207-1221. [doi: 10.3168/jds.S0022-0302(06)72190-7] [Medline: 16537954]
2. Moore LL, Bradlee ML, Gao D, Singer MR. Effects of average childhood dairy intake on adolescent bone health. J Pediatr 2008 Nov;153(5):667-673. [doi: 10.1016/j.jpeds.2008.05.016] [Medline: 18701115]
3. Dougkas A, Reynolds CK, Givens ID, Elwood PC, Minihane AM. Associations between dairy consumption and body weight: a review of the evidence and underlying mechanisms. Nutr Res Rev 2011 Jun;24(1):72-95. [doi: 10.1017/S095442241000034X] [Medline: 21320381]
4. Mena-Sánchez G, Becerra-Tomás N, Babio N, Salas-Salvadó J. Dairy product consumption in the prevention of metabolic syndrome: a systematic review and meta-analysis of prospective cohort studies. Adv Nutr 2019 May 1;10(suppl_2):S144-S153. [doi: 10.1093/advances/nmy083] [Medline: 31089736]
5. Drapeau V, Després JP, Bouchard C, et al. Modifications in food-group consumption are related to long-term body-weight changes. Am J Clin Nutr 2004 Jul;80(1):29-37. [doi: 10.1093/ajcn/80.1.29] [Medline: 15213024]
6. Douglas SM, Ortinau LC, Hoertel HA, Leidy HJ. Low, moderate, or high protein yogurt snacks on appetite control and subsequent eating in healthy women. Appetite 2013 Jan;60(1):117-122. [doi: 10.1016/j.appet.2012.09.012] [Medline: 23022602]
7. Alvarez-Bueno C, Cavero-Redondo I, Martinez-Vizcaino V, Sotos-Prieto M, Ruiz JR, Gil A. Effects of milk and dairy product consumption on type 2 diabetes: overview of systematic reviews and meta-analyses. Adv Nutr 2019 May 1;10(suppl_2):S154-S163. [doi: 10.1093/advances/nmy107] [Medline: 31089734]

XSL•FO
RenderX

8. Feng Y, Zhao Y, Liu J, et al. Consumption of dairy products and the risk of overweight or obesity, hypertension, and type 2 diabetes mellitus: a dose-response meta-analysis and systematic review of cohort studies. Adv Nutr 2022 Dec 22;13(6):2165-2179. [doi: 10.1093/advances/nmac096] [Medline: 36047956]
9. Ahmed M, Praneet Ng A, L'Abbe MR. Nutrient intakes of Canadian adults: results from the Canadian Community Health Survey (CCHS)-2015 Public Use Microdata File. Am J Clin Nutr 2021 Sep 1;114(3):1131-1140. [doi: 10.1093/ajcn/nqab143] [Medline: 34020449]
10. Canada's Food Guide [Article in French]. Government of Canada. 2024. URL: https://guide-alimentaire.canada.ca/fr/ [accessed 2024-02-11]
11. Barr SI. Is the 2019 Canada's Food Guide Snapshot nutritionally adequate? Appl Physiol Nutr Metab 2019 Dec;44(12):1387-1390. [doi: 10.1139/apnm-2019-0432]
12. Brassard D, Chevalier S. Relationship between adherence to the 2019 Canada's Food Guide recommendations on healthy food choices and nutrient intakes in older adults. J Nutr 2023 Sep;153(9):2699-2708. [doi: 10.1016/j.tjnut.2023.07.005] [Medline: 37460086]
13. Magarey A, Watson J, Golley RK, et al. Assessing dietary intake in children and adolescents: considerations and recommendations for obesity research. Int J Pediatr Obes 2011 Feb;6(1):2-11. [doi: 10.3109/17477161003728469] [Medline: 20874449]
14. Patrick H, Nicklas TA. A review of family and social determinants of children's eating patterns and diet quality. J Am Coll Nutr 2005 Apr;24(2):83-92. [doi: 10.1080/07315724.2005.10719448] [Medline: 15798074]
15. Chai LK, Farletti R, Fathi L, Littlewood R. A rapid review of the impact of family-based digital interventions for obesity prevention and treatment on obesity-related outcomes in primary school-aged children. Nutrients 2022 Nov 15;14(22):4837. [doi: 10.3390/nu14224837] [Medline: 36432522]
16. Kracht CL, Hutchesson M, Ahmed M, et al. E-&mHealth interventions targeting nutrition, physical activity, sedentary behavior, and/or obesity among children: A scoping review of systematic reviews and meta-analyses. Obes Rev 2021 Dec;22(12):e13331. [doi: 10.1111/obr.13331] [Medline: 34476890]
17. Antwi FA, Fazylova N, Garcon MC, Lopez L, Rubiano R, Slyer JT. Effectiveness of web-based programs on the reduction of childhood obesity in school-aged children: a systematic review. JBI Database System Rev Implement Rep 2013 Jun;11(6):1-44. [doi: 10.11124/jbisrir-2013-459]
18. Drapeau V, Savard M, Gallant A, Nadeau L, Gagnon J. The effectiveness of a school-based nutrition intervention on children's fruit, vegetables, and dairy product intake. J Sch Health 2016 May;86(5):353-362. [doi: 10.1111/josh.12385] [Medline: 27040473]
19. Chamberland K, Sanchez M, Panahi S, Provencher V, Gagnon J, Drapeau V. The impact of an innovative web-based school nutrition intervention to increase fruits and vegetables and milk and alternatives in adolescents: a clustered randomized trial. Int J Behav Nutr Phys Act 2017 Oct 16;14(1):140. [doi: 10.1186/s12966-017-0595-7] [Medline: 29037203]
20. Drapeau V, Harvey AA, Jacob R, Provencher V, Panahi S. The impact of a family web-based nutrition intervention to increase fruit, vegetable, and dairy intakes: a single-blinded randomized family clustered intervention. Nutr J 2022 Dec 20;21(1):75. [doi: 10.1186/s12937-022-00825-6] [Medline: 36539753]
21. Mummah SA, Robinson TN, King AC, Gardner CD, Sutton S. IDEAS (Integrate, Design, Assess, and Share): a framework and toolkit of strategies for the development of more effective digital interventions to change health behavior. J Med Internet Res 2016 Dec 16;18(12):e317. [doi: 10.2196/jmir.5927] [Medline: 27986647]
22. Hartson KR, Della LJ, King KM, Liu S, Newquist PN, Rhodes RE. Application of the IDEAS framework in adapting a web-based physical activity intervention for young adult college students. Healthcare (Basel) 2022 Apr 9;10(4):700. [doi: 10.3390/healthcare10040700] [Medline: 35455877]
23. Fedele DA, McConville A, Moon J, Thomas JG. Topical review: design considerations when creating pediatric mobile health interventions: applying the IDEAS framework. J Pediatr Psychol 2019 Apr 1;44(3):343-348. [doi: 10.1093/jpepsy/jsy084] [Medline: 30346561]
24. Curtis KE, Lahiri S, Brown KE. Targeting parents for childhood weight management: development of a theory-driven and user-centered healthy eating app. JMIR Mhealth Uhealth 2015 Jun 18;3(2):e69. [doi: 10.2196/mhealth.3857] [Medline: 26088692]
25. Yardley L, Morrison L, Bradbury K, Muller I. The person-based approach to intervention development: application to digital health-related behavior change interventions. J Med Internet Res 2015 Jan 30;17(1):e30. [doi: 10.2196/jmir.4055] [Medline: 25639757]
26. Tremblay A, Panahi S. Yogurt consumption as a signature of a healthy diet and lifestyle. J Nutr 2017 Jul;147(7):1476S-1480S. [doi: 10.3945/jn.116.245522] [Medline: 28615373]
27. Webb TL, Joseph J, Yardley L, Michie S. Using the internet to promote health behavior change: a systematic review and meta-analysis of the impact of theoretical basis, use of behavior change techniques, and mode of delivery on efficacy. J Med Internet Res 2010 Feb 17;12(1):e4. [doi: 10.2196/jmir.1376] [Medline: 20164043]
28. Sutton S, French DP, Hennings SJ, et al. Eliciting salient beliefs in research on the theory of planned behaviour: the effect of question wording. Curr Psychol 2003 Sep;22(3):234-251. [doi: 10.1007/s12144-003-1019-1]

XSL•FO
RenderX

29. Lacroix MJ, Desroches S, Turcotte M, et al. Salient beliefs among Canadian adults regarding milk and cheese consumption: a qualitative study based on the theory of planned behaviour. BMC Nutr 2016 Dec;2(1):48. [doi: 10.1186/s40795-016-0087-1]
30. Michie S, Richardson M, Johnston M, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013 Aug;46(1):81-95. [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
31. Robinson TN. Chapter 25 - stealth interventions for obesity prevention and control: motivating behavior change. In: Dubé L, Bechara A, Dagher A, editors. Obesity Prevention: The Role of Brain and Society on Individual Behavior: Academic Press; 2010. [doi: 10.1016/B978-0-12-374387-9.00025-8]
32. Bitrián P, Buil I, Catalán S. Enhancing user engagement: the role of gamification in mobile apps. J Bus Res 2021 Aug;132:170-185. [doi: 10.1016/j.jbusres.2021.04.028]
33. Cheema A, Bagchi R. The effect of goal visualization on goal pursuit: implications for consumers and managers. J Mark 2011 Mar;75(2):109-123. [doi: 10.1509/jm.75.2.109]
34. Lallemand C, Koenig V, Gronier G, Martin R. Création et validation d'une version française du questionnaire AttrakDiff pour l'évaluation de l'expérience utilisateur des systèmes interactifs [Article in French]. European Review of Applied Psychology 2015 Sep;65(5):239-252. [doi: 10.1016/j.erap.2015.08.002]
35. Giardi A. Evaluate children's User eXperience with AttrakDiff method: USiena experience. Open Journal Per La Formazione in Rete Dec 2019;19(3):155-174. [doi: 10.13128/form-7464]
36. Hassenzahl M, Burmester M, Koller F. AttrakDiff: ein fragebogen zur messung wahrgenommener hedonischer und pragmatischer qualität. In: Szwillus G, Ziegler J, editors. Mensch & Computer 2003: Interaktion in Bewegung [Book in German]: Vieweg+Teubner Verlag; 2003:187-196. [doi: 10.1007/978-3-322-80058-9_19]
37. Lallemand C, Gronier G, Koenig V. User experience: a concept without consensus? Exploring practitioners' perspectives through an international survey. Comput Human Behav 2015 Feb;43:35-48. [doi: 10.1016/j.chb.2014.10.048]
38. Díaz-Oreiro I, López G, Quesada L, Guerrero LA. UX evaluation with standardized questionnaires in ubiquitous computing and ambient intelligence: a systematic literature review. Adv Hum Comput Interact 2021 May 4;2021(1):1-22. [doi: 10.1155/2021/5518722]
39. Van Voorhees BW, Gladstone T, Cordel S, et al. Development of a technology-based behavioral vaccine to prevent adolescent depression: a health system integration model. Internet Interv 2015 Sep;2(3):303-313. [doi: 10.1016/j.invent.2015.07.004] [Medline: 30473992]
40. Kelders SM, Pots WTM, Oskam MJ, Bohlmeijer ET, van Gemert-Pijnen JEWC. Development of a web-based intervention for the indicated prevention of depression. BMC Med Inform Decis Mak 2013 Feb 20;13:26. [doi: 10.1186/1472-6947-13-26] [Medline: 23425322]
41. Mummah SA, King AC, Gardner CD, Sutton S. Iterative development of Vegethon: a theory-based mobile app intervention to increase vegetable consumption. Int J Behav Nutr Phys Act 2016 Aug 8;13:90. [doi: 10.1186/s12966-016-0400-z] [Medline: 27501724]
42. Mummah S, Robinson TN, Mathur M, Farzinkhou S, Sutton S, Gardner CD. Effect of a mobile app intervention on vegetable consumption in overweight adults: a randomized controlled trial. Int J Behav Nutr Phys Act 2017 Sep 15;14(1):125. [doi: 10.1186/s12966-017-0563-2] [Medline: 28915825]
43. Cox D, Cox AD. Beyond first impressions: the effects of repeated exposure on consumer liking of visually complex and simple product designs. J Acad Mark Sci 2002 Apr 1;30(2):119-130. [doi: 10.1177/03079459994371]
44. Cechetti NP, Bellei EA, Biduski D, Rodriguez JPM, Roman MK, De Marchi ACB. Developing and implementing a gamification method to improve user engagement: a case study with an m-Health application for hypertension monitoring. Telemat Inform 2019 Aug;41:126-138. [doi: 10.1016/j.tele.2019.04.007]
45. Sardi L, Idri A, Fernández-Alemán JL. A systematic review of gamification in e-Health. J Biomed Inform 2017 Jul;71:31-48. [doi: 10.1016/j.jbi.2017.05.011] [Medline: 28536062]
46. Seah ML, Cairns P. From immersion to addiction in videogames. In: People and Computers XXII Culture, Creativity, Interaction: The British Computer Society; 2008. [doi: 10.14236/ewic/HCI2008.6]
47. Bente BE, Wentzel J, Schepers C, et al. Implementation and user evaluation of an eHealth technology platform supporting patients with cardiovascular disease in managing their health after a cardiac event: mixed methods study. JMIR Cardio 2023 Mar 24;7:e43781. [doi: 10.2196/43781] [Medline: 36961491]
48. Oakley-Girvan I, Yunis R, Longmire M, Ouillon JS. What works best to engage participants in mobile app interventions and e-health: a scoping review. Telemed J E Health 2022 Jun;28(6):768-780. [doi: 10.1089/tmj.2021.0176] [Medline: 34637651]
49. Oliveira ER, Branco AC, Carvalho D, et al. An iterative process for the evaluation of a mobile application prototype. SN COMPUT SCI 2022 Jul;3(4):262. [doi: 10.1007/s42979-022-01153-6]
50. Espinosa-Curiel IE, Pozas-Bogarin EE, Martínez-Miranda J, Pérez-Espinosa H. Relationship between children's enjoyment, user experience satisfaction, and learning in a serious video game for nutrition education: empirical pilot study. JMIR Serious Games 2020 Sep 17;8(3):e21813. [doi: 10.2196/21813] [Medline: 32940609]
51. Vieira D, Providência B, Carvalho H. Design of a smart garment for fencing: measuring attractiveness using the AttrakDiff Mini method. Hum-Intell Syst Integr 2023 Nov;5(1-2):1-9. [doi: 10.1007/s42454-023-00047-z]

XSL•FO
RenderX

52. Jung ME, Latimer-Cheung AE, Bourne JE, Martin Ginis KA. Targeted messages increase dairy consumption in adults: a randomized controlled trial. Ann Behav Med 2017 Feb;51(1):57-66. [doi: 10.1007/s12160-016-9828-2] [Medline: 27797014]
53. Domagk S, Schwartz RN, Plass JL. Interactivity in multimedia learning: an integrated model. Comput Human Behav 2010 Sep;26(5):1024-1033. [doi: 10.1016/j.chb.2010.03.003]
54. Naderbagi A, Loblay V, Zahed IUM, et al. Cultural and contextual adaptation of digital health interventions: narrative review. J Med Internet Res 2024 Jul 9;26:e55130. [doi: 10.2196/55130] [Medline: 38980719]

Abbreviation:

ATT: attractiveness dimension
BCT: behavior change technique
CFG: Canada's Food Guide
HIQ: hedonic-identity quality
HSQ: hedonic-stimulation quality
IDEAS: Ideate, Design, Assess, and Share
PQ: pragmatic quality
TPB: theory of planned behavior
UX: user experience

Edited by A Mavragani; submitted 17.09.24; peer-reviewed by L Ou, W Zhou; revised version received 12.03.25; accepted 12.03.25; published 13.08.25.

Please cite as:
Lemay J, Roberge J, Provencher V, Tremblay A, Panahi S, Jacob R, Brunelle L, Saintonge G, Drapeau V
Promoting Dairy Consumption Among Families: Development and User Experience Study of a Web-Based Nutrition Intervention
JMIR Form Res 2025;9:e66582
URL: https://formative.jmir.org/2025/1/e66582
doi:10.2196/66582

© Juliette Lemay, Jacynthe Roberge, Véronique Provencher, Angelo Tremblay, Shirin Panahi, Raphaëlle Jacob, Lucie Brunelle, Gabrielle Saintonge, Vicky Drapeau. Originally published in JMIR Formative Research (https://formative.jmir.org), 13.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Technology Integration in Syrian Medical Education From the Perspective of Students and Faculty: A Cross-Sectional Evaluation

Subhiya Hassoun[1], DDS; Mayssoon Dashash[1,2], DDS, MSc, PhD; Adnan Baddour[1,3], MSc, PhD

[1]Medical Education Department, Syrian Virtual University, P.O. Box 35329, Damascus, Syrian Arab Republic

[2]Department of Pediatric Dentistry, Faculty of Dental Medicine, Damascus University, Damascus, Syrian Arab Republic

[3]Faculty of Pharmacy, University of Kalamoon, Deir-Attia, Syrian Arab Republic

Corresponding Author:
Subhiya Hassoun, DDS
Medical Education Department, Syrian Virtual University, P.O. Box 35329, Damascus, Syrian Arab Republic

Abstract

Background: Technology-enhanced learning (TEL) has become increasingly vital in global medical education, offering significant advantages in knowledge acquisition, communication, motivation, and student engagement. In Syria, a country facing prolonged crises, there is an urgent need to evaluate the integration of technology within medical education to address institutional limitations and support student learning.

Objective: The aim of this study is to evaluate the awareness, perceived challenges, and needs regarding the integration of technology in medical education from the perspectives of students and faculty at Syrian medical colleges.

Methods: A cross-sectional survey was conducted during the 2023 - 2024 academic year across Syrian universities. Stratified random sampling was used to recruit 500 medical students and 200 faculty members. Two tailored, self-administered questionnaires were used, covering motivation, perceived benefits, challenges, and suggestions for technology integration. Validity was assessed through expert review and pilot testing (n=30), and internal consistency was confirmed (Cronbach α=0.6 - 0.7). Quantitative data were analyzed using descriptive statistics, *t* tests, ANOVA, and Kruskal-Wallis tests.

Results: Among medical students, 94% (470\500) agreed that integrating technology into medical education is essential, with similar agreement from 93.5% (187\200) of faculty. No significant differences were found in student responses based on specialization (P=.32) or university type (P=.11). Likewise, faculty perspectives did not significantly differ by academic qualification or years of experience (P>.05). There were several perceived benefits; for instance, 93.2% (n=466) of students reported that technology kept them up to date with new developments, 88% stated it enhanced research skills, and 86.8% found TEL more enjoyable than traditional learning methods. Most respondents (95% n=475) said TEL created a flexible, interactive environment. Among faculty, 77% (n=154) agreed TEL improves clinical skill development. Respondents noted there were some challenges; specifically, 57% (n=285) of students cited poor internet service, 33% (n=165) noted the financial burden, and 82.2% (n=411) called for behavioral guidelines. Among faculty, 85.5% (n=171) cited lack of institutional support and 90% (n=180) emphasized the need for training. Both groups supported the development of communication platforms, curriculum revisions, and faculty development programs.

Conclusions: There is a strong consensus among Syrian medical students and faculty on the value and necessity of integrating technology in medical education. Despite infrastructure and administrative challenges, both groups recognize TEL as a powerful tool for improving clinical competencies, student motivation, and academic engagement. Institutional commitment, curricular reform, and tailored training are essential to achieving sustainable, effective technology integration.

(JMIR Form Res 2025;9:e76958) doi:10.2196/76958

KEYWORDS

technology integration; technology-enhanced learning environments; blended learning; medical education; faculty members

Introduction

The landscape of medical education is rapidly evolving in response to the exponential growth of medical knowledge, the demand for lifelong learning, and changing societal expectations regarding professionalism, patient safety, and evidence-based care [1,2]. Technology-enhanced learning (TEL) has emerged as a key strategy to meet these demands by offering flexible, interactive, and student-centered educational environments. TEL methods—including web-based modules, virtual simulation, and mobile learning—are shown to improve knowledge retention, clinical reasoning, motivation, and engagement across various health disciplines [3-5].

XSL•FO
RenderX

Globally, blended learning models that combine face-to-face instruction with web-based components have gained widespread acceptance, particularly after the COVID-19 pandemic highlighted the need for resilient educational systems [6,7]. In the context of medical education, TEL fosters not only cognitive and technical skill development but also collaboration, self-directed learning, and emotional engagement [8]. These advantages make TEL an essential component of modern medical curricula, particularly in low-resource or crisis-affected settings.

In Syria, however, technology integration into medical education remains underexplored. The country's prolonged conflict and resulting humanitarian crisis have severely impacted higher education institutions, contributing to gaps in infrastructure, access, and pedagogical innovation [9,10]. Despite these challenges, there is growing interest in leveraging TEL to improve teaching and learning outcomes in Syrian universities. Initiatives such as web-based modules, digital assessments, and virtual classrooms have gained traction, though systematic evaluations remain limited [11].

Recognizing the gap in technology integration within Syrian medical education, this study aims to provide a comprehensive formative evaluation of the extent to which TEL has been incorporated in this context. By examining and analyzing the perspectives of both medical students and faculty members, the study seeks to assess current levels of awareness about TEL, identify institutional challenges and requirements, and generate practical, actionable insights to enhance its implementation. Understanding stakeholder perceptions is essential for developing culturally appropriate and context-sensitive strategies that support innovation and resilience in Syrian medical education. This formative approach goes beyond purely descriptive assessments; this study presents a formative evaluation of technology integration in Syrian medical education. By examining the perspectives of both medical students and faculty members, this research aims to inform future educational planning and policy reform in Syrian higher education. The study's findings can guide the development of support systems, capacity-building programs, and infrastructure investments aligned with both local realities and global standards in medical education.

Methods

Study Design

A cross-sectional study was conducted across multiple public and private medical colleges in the Syrian Arab Republic during the winter semester of the 2023‑2024 academic year. The participating institutions included faculties of medicine, dentistry, pharmacy, and health sciences affiliated with universities located in major Syrian cities such as Damascus, Aleppo, Latakia, and Homs.

Study Population

The study population consisted of key stakeholders involved in the integration of technology into medical education at Syrian universities. This included faculty members and educational personnel with PhD degrees, master's degrees, and other degrees in various fields including human medicine, pharmacy, dental medicine, and health sciences. The sample accounted for years of experience and included both academic and administrative staff, regardless of their decision-making roles. Additionally, medical college students were included, taking into consideration their respective colleges, year of study, and whether they attended public or private universities.

Eligibility criteria for students included enrollment in one of the targeted colleges during the academic year, while faculty participants were required to be actively employed at the time of data collection.

Given the diverse nature of the study community, stratified random sampling was used. The population was divided into two homogeneous samples from which random samples were drawn. The number of participants (n) in each sample was proportional to the size of the stratum (N).

Survey Design

The study used a self-administered anonymous questionnaire, and a systematic approach was used to ensure equal representation within the targeted sample, including only those participants who completed the questionnaires.

Two questionnaires were designed, each tailored to a specific group from the study population. The questionnaires were created in Arabic on the basis of prior studies related to the topic, field observations, and expert consultations. An English version of the questionnaire is available in Multimedia Appendices 1 and 2. A 5-point Likert scale, ranging from 1 (strongly disagree) to 5 (strongly agree), was used to assess attitudes related to study variables.

Questionnaire A was designed for sample A to evaluate medical college students' awareness of the technology integration process at Syrian universities. Demographic data such as gender, specialization, academic year, and university type were collected. This questionnaire consisted of 23 items divided into 4 sections. The first section consisted of 7 items focused on evaluating motivation to learn in TEL environments. The second section included 7 items exploring skills and experiences gained in TEL environments. The third and fourth sections discussed the challenges and needs associated with the technology integration process in medical education, comprising 9 items. Additionally, an open-ended section was included for participants to provide further recommendations or suggestions regarding technology integration in medical education (Multimedia Appendix 1).

Questionnaire B was directed toward sample B to assess the awareness of educational members in medical colleges at Syrian universities. It gathered information on academic specialization, qualifications, and years of experience. This questionnaire included 24 items divided into 3 sections. The first section consisted of 11 items aimed at analyzing the respondent's personal point of view on the goal of technology integration in medical education. The second and third sections consisted of 13 items focused on the requirements and challenges faced during the integration process, alongside an open-ended field for additional recommendations or proposals related to

XSL•FO
RenderX

technology integration in medical education (Multimedia Appendix 2).

Data Collection

Participant recruitment was conducted through a combination of direct email invitations, university announcements, and social media channels used by academic institutions (eg, WhatsApp groups, Telegram channels). Interested individuals were provided with an explanation of the study purpose, ethical approval details, and a secure Google Forms link to the relevant questionnaire.

Participation was entirely voluntary, and responses were collected anonymously to ensure privacy and minimize response bias. Only fully completed questionnaires were included in the final analysis. A total of 500 students and 200 faculty members completed the surveys, yielding a 100% response rate based on the targeted sample.

Ethical Considerations

This study received ethical approval from the ethical research committee of the Syrian Virtual University (number 2258/0, dated August 8, 2024). The research was conducted in compliance with the Declaration of Helsinki and relevant national ethical standards.

All participants were informed about the objectives of the study and their rights as participants. Participation was voluntary, and informed consent was obtained electronically before the questionnaire was administered.

Participants' confidentiality and anonymity were strictly maintained. All data were anonymized and securely stored in password-protected files accessible only to authorized research team members. The data are available upon request from the corresponding author.

This study did not involve any clinical interventions or procedures. It consisted solely of a self-administered questionnaire assessing perceptions and experiences of faculty and students regarding technology integration in medical education.

No images or supplementary materials in this study contain identifiable information or visual representations of individual participants. No identifying personal data were collected, and all responses remained anonymous. No compensation was provided to participants in this study.

Instrument Development and Validation

Two separate, structured, self-administered questionnaires were developed to capture the perspectives of medical students and faculty members regarding the integration of technology into Syrian medical education. The instruments were created in Arabic, drawing on existing literature, institutional observations, and expert consultations in medical education and educational technology. English translations are available in Multimedia Appendices 1 and 2.

The questionnaire for students (Sample A) consisted of 23 items across four domains: (1) motivation to learn in technology-enhanced environments, (2) skills and experiences gained, (3) perceived challenges, and (4) institutional needs. The faculty questionnaire (Sample B) contained 24 items across three domains: (1) attitudes toward technology in medical education, (2) perceived requirements, and (3) challenges. Both instruments included open-ended items for qualitative input. Face and content validity were established through a panel of 5 medical education experts who assessed each item's relevance, clarity, and alignment with study objectives. Based on their feedback, items were revised for clarity and cultural appropriateness.

A pilot test was conducted with 30 participants (15 students, 15 faculty members) from Syrian universities who were not part of the final sample. Respondents provided structured feedback on questionnaire clarity, response burden, and item comprehension. Minor adjustments were made accordingly.

Internal consistency reliability was assessed using Cronbach α. For both instruments, α values ranged from 0.6 to 0.7 across all domains, indicating acceptable reliability for exploratory research in social and educational contexts [4]. All items used a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree).

Sample Validity and Reliability Test

Reliability is a critical characteristic of a study tool. The importance of measuring the reliability lies in obtaining consistent results. The reliability of the overall questionnaire was verified via the Cronbach α coefficient, which ranges from 0 to 1, with an acceptable threshold of 60% in social studies. The reliability coefficient values for the study sections ranged from 0.6 to 0.7, indicating acceptable reliability and suggesting that the questionnaire is valid and reliable.

Data Analysis

To meet the study objectives, the data were transferred from Google Forms to Microsoft Excel, and the Statistical Package for the Social Sciences software (version 25; IBM Corp) was used to analyze the final data. Descriptive statistics, including the mean (SD), were calculated. Statistical methods were used to analyze the data, test the study hypotheses, and describe the participants' responses. Quantitative variables are reported as mean (SD), whereas qualitative variables are reported as frequencies and percentages. A 5-point Likert scale was adopted, with the mean calculated for each statement (out of 5), each section of the evaluation, and the total questionnaire. Mean values and ranges were used to provide a verbal interpretation of the mean responses: strongly disagree=1.00 - 1.8, disagree=1.81 - 2.60, not sure=2.61 - 3.40, agree=3.41 - 4.20, strongly agree=4.21 - 5.

For Sample A, an independent sample *t* test was used to determine significant differences in skills and experiences gained in TEL environments according to the type of university. The Kruskal-Wallis test was used to determine significant differences in motivation to learn in TEL environments on the basis of the specialization variable. A significance level ($P<.05$) was established to determine statistical significance.

For Sample B, an independent sample *t* test was used to determine significant differences in the personal opinions of

XSL•FO
RenderX

faculty members in medical colleges, regardless of whether they held administrative positions (decision-makers). A 1-way ANOVA test was used to determine significant differences in faculty opinions regarding technology integration in medical education on the basis of years of experience, whereas the Welch test was conducted to determine significant differences in faculty opinions about technology integration on the basis of the academic degree. A significance level of $P<.05$ was applied to determine statistical significance.

Qualitative Data Analysis

Responses to open-ended questions were analyzed using a basic thematic content analysis approach. After initial reading, the responses were coded and categorized into emerging themes. The analysis was conducted manually by two researchers independently, who then discussed and resolved any discrepancies to ensure consistency. Thematic saturation was assessed informally, and illustrative quotes were selected to represent common viewpoints and highlight participant insights.

Results

Sample Analysis

Sample A consisted of 500 medical college students from Syrian universities who met the criteria and voluntarily agreed to participate in the study (Table 1). Among the participants, females represented 61.2%, whereas males accounted for 38.8%. In terms of university affiliation, 37.2% attended governmental universities, whereas the majority, 62.8%, attended private universities. The highest percentage of students by specialization was among dental students (35.6%), followed by pharmacy students (32.4%) and medical students (32%). The largest group by academic year was fifth-year students at 25.6%, followed by second-year students at 20.6%, fourth-year students at 19.6%, third-year students at 18%, and first-year students at 16.2%.

XSL•FO
RenderX

Table . Demographic characteristics of Syrian medical students (N=500) and faculty members (N=200) who participated in a cross-sectional study on technology-enhanced learning conducted during the 2023 - 2024 academic year.

Characteristics		Values, n (%)
Sample A (students)		
Gender		
	Male	194 (38.8)
	Female	306 (61.2)
University		
	Governmental	186 (37.2)
	Private	314 (62.8)
College		
	Medicine	160 (32)
	Pharmacy	162 (32.4)
	Dentistry	178 (35.6)
Year of study		
	First	81 (16.2)
	Second	103 (20.6)
	Third	90 (18)
	Fourth	98 (19.6)
	Fifth	128 (25.6)
Sample B (faculty)		
Specialization		
	Human medicine	57 (28.5)
	Dental medicine	65 (32.5)
	Pharmacy	49 (24.5)
	Health sciences	29 (14.5)
Academic degree		
	Bachelor's degree	74 (37)
	Master's degree	85 (42.5)
	PhD	41 (20.5)
Administrative position		
	Yes	20 (10)
	No	180 (90)
Years of experience		
	Less than 5 years	80 (40)
	5-10 years	79 (39.5)
	More than 10 years	41 (20.5)

Conversely, Sample B included 200 faculty members from medical colleges at Syrian universities (Table 1). The largest percentage came from the faculty of dentistry (32.5%). This was followed by specialists in medicine (28.5%), pharmacy (24.5%), and health sciences (14.5%). With respect to academic qualifications, 42.5% held a master's degree, 37% held a bachelor's degree, and 20.5% held a PhD. In terms of professional experience, 40% had less than 5 years of experience, 39.5% had 5 - 10 years, and 20.5% had more than 10 years. Additionally, 10% of the faculty members were decision-makers and held administrative positions.

Statistical Analysis

All participants completed the questionnaire, resulting in a 100% response rate. The overall evaluation of the questionnaires was positive as indicated in Table 2 (the English versions of the questionnaires are provided in Multimedia Appendices 1 and 2). Most participants agreed that integrating technology into

XSL•FO
RenderX

medical education is necessary today, with a mean score of 4.5 for sample A and 4.4 for sample B. Furthermore, participants from both samples agreed that TEL environments enhance communication skills among students, with mean scores of 3.6 for this statement. The participants reported challenges and requirements associated with the integration of technology, with a mean score of 3.25 for sample A and 3.79 for sample B for the entire section (Table 2). We also obtained the mean Likert-scale responses from medical students and faculty members regarding motivation (students: 4.02), skills (students: 4.08; faculty: 3.04), challenges (students: 3.25; faculty: 3.62), and needs related to TEL (students: 3.50; faculty:3.79) in Syrian medical education.

XSL•FO
RenderX

Table . Descriptive statistics for items evaluating technology-enhanced learning Results are based on responses from students (Questionnaire A) and faculty (Questionnaire B) at Syrian medical universities.

Statement evaluation and item or section number	Mean	Standard deviation	Expressions
Questionnaire A			
Section 1			
1.1	4.5	0.6	Strongly agree
1.2	4.4	0.5	Strongly agree
1.3	4.3	0.7	Strongly agree
1.4	3.3	1	Not sure
1.5	4.1	0.8	Agree
1.6	4.3	0.7	Strongly agree
1.7	3.3	1	Not sure
Section 2			
2.1	4.3	0.6	Strongly agree
2.2	3.9	0.9	Agree
2.3	3.8	0.9	Agree
2.4	3.6	1	Agree
2.5	4.22	0.7	Strongly agree
2.6	4.4	0.6	Strongly agree
2.7	4.4	0.6	Strongly agree
Section 3			
3.1	3.06	1.1	Not sure
3.2	3.5	1.1	Agree
3.3	3.01	1	Not sure
3.4	3.45	1	Agree
Section 4			
4.1	4	0.7	Strongly agree
4.2	3.4	1.1	Agree
4.3	3.3	1.1	Not sure
4.4	3.6	1.1	Agree
4.5	3.2	1.2	Not sure
Section evaluation			
S.1	4.02	0.7	Agree
S.2	4.08	0.7	Agree
S.3	3.25	1.05	Not sure
S.4	3.5	1.04	Agree
Questionnaire B			
Section 1			
1.1	4.4	0.6	Strongly agree
1.2	3.14	1.2	Not sure
1.3	3.9	0.9	Agree
1.4	2.9	1.1	Not sure
1.5	3.7	0.9	Agree
1.6	3.9	0.8	Agree

XSL•FO
RenderX

Statement evaluation and item or section number	Mean	Standard deviation	Expressions
1.7	3.6	1.08	Agree
1.8	2.7	1.05	Not sure
1.9	4.2	0.7	Strongly Agree
1.10	3.8	1.02	Agree
1.11	3.8	0.8	Agree
Section 2			
2.1	3.09	1.2	Not sure
2.2	4.28	0.8	Strongly agree
2.3	4.29	0.7	Strongly agree
2.4	4.39	0.7	Strongly agree
2.5	3.6	1.07	Agree
2.6	3.09	0.9	Not sure
Section 3			
3.1	4.19	0.8	Agree
3.2	4.18	0.7	Agree
3.3	3.9	0.9	Agree
3.4	3.48	1.19	Agree
3.5	3.22	1.1	Not sure
3.6	2.9	1.14	Not sure
3.7	3.53	0.9	Agree
Section evaluation			
S.1	3.64	0.9	Agree
S.2	3.79	0.8	Agree
S.3	3.62	0.9	Agree

Independent sample *t* test showed no significant difference between the dependent variable and the independent variable regarding the skills and experience gained in TEL environments among students from governmental (mean 4.0845, SD 0.47542) and private universities (mean 4.1306, SD 0.53298; P=.11; P<.05). When comparing motivation to learn in TEL environments on the basis of students' specialization, the Kruskal-Wallis Test was applied after confirming the test hypotheses and conditions, ensuring the normal distribution of the sample, and validating it through the QQ plot. The results revealed no significant differences, as the average responses for motivation to learn in TEL environments were similar. Medicine students had an average score of 264.23, dental students had an average score of 247.13, and pharmacy students had an average score of 241.22, with a P value of .32. These findings are presented in Table 3. Similarly, no significant differences were observed when comparing faculty members' opinions about TEL environments on the basis of years of experience and academic degree, according to the results of a 1-way ANOVA and Welch test (Table 4).

Table . Kruskal-Wallis test comparing student motivation to learn in technology-enhanced learning environments across academic specializations (medicine, dentistry, pharmacy) among students in Syrian medical colleges.

College	Mean	SD	*P* value
Medicine	4.136	.5307	.32
Dentistry	4.068	.5446	
Pharmacy	4.028	.5753	

XSL•FO
RenderX

Table . One-way ANOVA and Welch test comparing faculty members' perceptions of technology-enhanced learning environments based on academic degree and years of teaching experience in Syrian medical institutions.

Variable	Mean	SD	*P* value
Academic degree			.22
Bachelor	2.6462	.92423	
Master	2.6298	.62768	
PhD	2.5543	.69019	
Years of teaching experience			.92
Less than 5 years	2.5352	.77935	
5-10 years	2.6974	.70759	
More than 10 years	2.6364	.81616	

Student Preferences for Learning Environments

Among the 500 surveyed students, preferences for learning environments varied notably. We found that 47.2% (n=236) of students preferred traditional, in-person instruction, while 18.6% (n=93) favored TEL environments. The remaining 34.2% (n=171) expressed a neutral stance, indicating no strong preference for either mode.

These findings suggest that while TEL is recognized as beneficial, a substantial portion of students still favor traditional approaches, possibly due to familiarity, infrastructure limitations, or insufficient exposure to well-implemented digital tools.

Perceived Challenges of TEL

Figure 1 shows the most commonly reported challenges faced by students when engaging in TEL environments. The majority 57% (n=285) cited poor internet access as a major barrier, followed by the need for behavioral or usage guidelines 82.2% (n=411), and difficulty navigating complex platforms 54.2% (n=271). Financial burden was also a concern, with 33% (n=165) of students reporting cost-related challenges. These findings highlight persistent infrastructure and support issues that may limit the effective adoption of TEL in Syrian medical institutions.

XSL•FO
RenderX

Figure 1. Percentage of students facing challenges in technology-enhanced learning enviroments in syrian medical educationn.

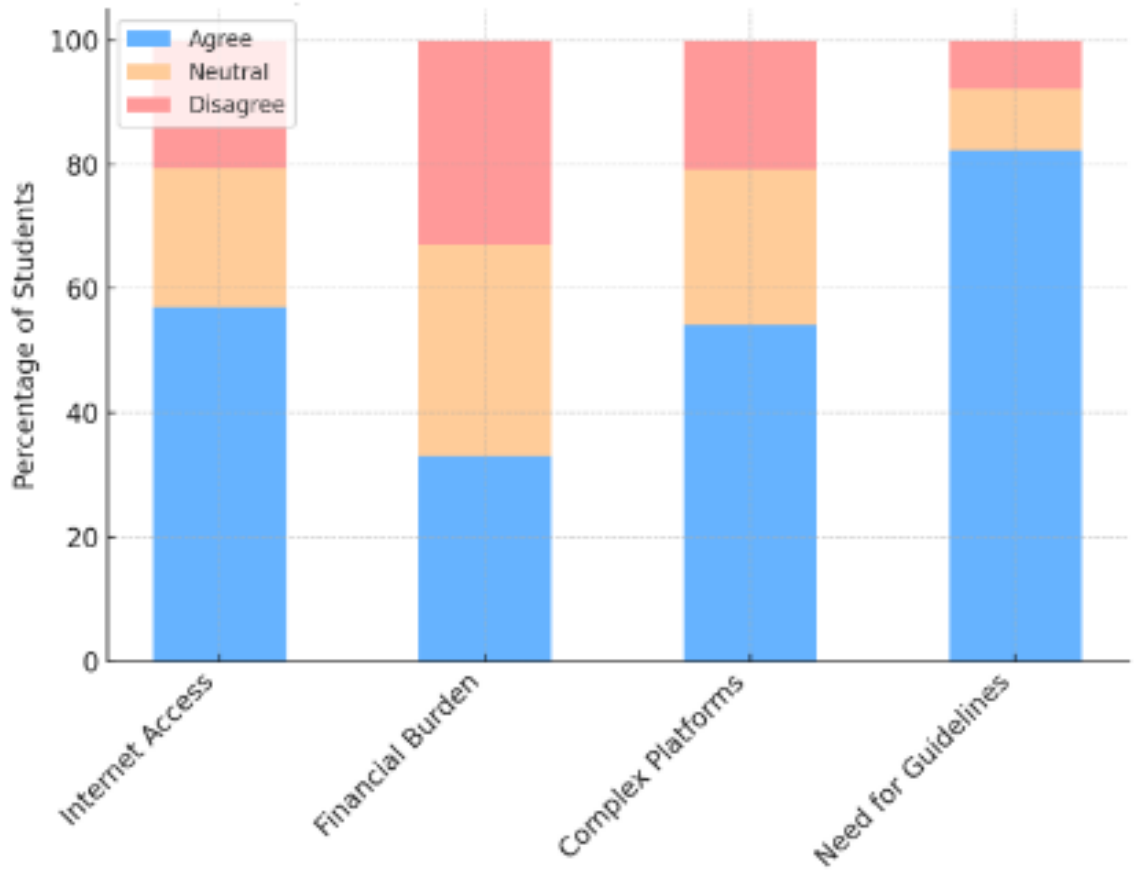

Qualitative Analysis

Open-ended responses provided further insight into participants' perspectives on technology integration. Thematic analysis revealed several recurring themes among both students and faculty members. These themes reflected practical needs, perceived barriers, and suggestions for improvement in TEL environments. Table 5 summarizes the most frequently mentioned themes along with representative anonymized quotations.

Table . Summary of qualitative themes derived from open-ended responses regarding technology-enhanced learning in Syrian medical education, including frequency and representative quotations from students and faculty.

Theme	Frequency, n	Sample quotation
Need for digital training	62	"We need proper training on how to use new platforms."
Lack of institutional support	48	"Our college has no clear policy for using technology."
Internet connectivity issues	85	"The connection is so bad I often miss online lectures."
Motivation from interactive tools	51	"Videos and interactive quizzes make learning enjoyable."

Discussion

Principal Findings

This study examined the perceptions of medical students and faculty members regarding the integration of technology into medical education across Syrian universities. The findings revealed high levels of agreement from both groups on the importance and value of TEL. Students and faculty identified similar benefits—such as improved clinical skills, communication, and flexibility—as well as common challenges, including poor internet access, institutional limitations, infrastructure and administrative challenges, and lack of training. No significant differences in perceptions were found based on demographics such as university type, specialization, or faculty experience. However, there was a strong consensus among

XSL•FO
RenderX

Syrian medical students and faculty on the value and necessity of integrating technology in medical education, and both groups recognized TEL as a powerful tool for improving clinical competencies, student motivation, and academic engagement. Institutional commitment, curricular reform, and tailored training are essential to achieving sustainable, effective technology integration.

Interpretation and Comparison With Prior Studies

Perceptions of Technology Integration and Its Impact on Skills and Learning

This study was designed to evaluate the current situation of technology integration into medical education at universities in Syria via a self-administrated questionnaire. A total of 500 medical students from medical colleges (medicine, dentistry, and pharmacy) and 200 faculty members completed the questionnaire regarding their knowledge and awareness of the technology integration process and its applications in medical education.

The findings indicated a high level of awareness and attention among participants regarding the importance and necessity of TEL in medical education. Specifically, 94% of medical students and 93.5% of faculty members agreed on the necessity of integrating technology in medical education, despite the challenges posed by the current situation in the country. However, no significant relationship was found between students' specialization and their perception of motivation in TEL. Additionally, there was no significant relationship between faculty members' academic qualifications or years of teaching experience and their perceptions of TEL environments. A positive trend was observed in agreement across all sections of the questionnaire, likely reflecting a growing awareness among faculty members regarding the need to shift away from traditional and outdated teaching methods in favor of modern methods.

In examining the significance of TEL environments, it is crucial to analyze their impact on the acquisition of cognitive and clinical skills. The results revealed that 77% of faculty members and 66.6% of medical students agreed that TEL environments facilitate the development of better clinical skills. Furthermore, 92.6% of the students held a positive attitude toward the role of technology in enhancing their understanding of course materials, whereas only 39.5% of the faculty members shared the same opinion, with 24.5% remaining neutral and 36% disagreeing. Additionally, 86.2% of medical students stated that technology aids them in their understanding and review of lessons, and 93.4% agreed that it provides access to a wide range of educational resources. Furthermore, 93.2% indicated that technology enables them to stay up to date with new trends in their field of study. These results align with a study on the effectiveness of the gastroenterology e-module for undergraduate medical students at Syrian universities, which concluded that e-learning significantly enhances knowledge levels, with no significant difference between synchronous and asynchronous methods, and good satisfaction and self-confidence among all participants [11]. Another study that evaluated the impact of a web-based restorative dentistry course on learning outcomes reported similar findings, indicating high satisfaction rates and improved understanding of restorative dentistry [5]. A third similar study of a web-based course regarding dental pain management and local anesthesia found that it effectively enhanced students' knowledge in these areas [12].

Additionally, 86.8% of the students agreed that the educational process is more enjoyable when technology is present, and 95% affirmed that integrating technology into education creates a flexible learning environment that facilitates interaction and skills acquisition. However, 47.2% preferred traditional learning environments, perceiving them as clearer and easier to navigate than TEL environments, which they still find complex and difficult to access. Furthermore, 34.2% of students remained neutral on this issue. Moreover, 54.2% acknowledged the challenges of engaging in TEL environments, attributing this to various difficulties, barriers, and requirements.

Communication, Engagement, and Educational Value

Although 33% of faculty members expressed concerns that TEL environments may diminish the role of the teacher, 75.5% agreed that these environments enhance student-centered learning, and 46% believed that they promote individual differences among students. Moreover, 58.5% agreed that these environments enhance communication skills. Additionally, 59.4% of medical students felt that technology enhances their communication skills, and 64% believed that it allows them to express their opinions freely and without hesitation. Furthermore, 88% reported that technology has helped them develop their research skills, whereas 84.2% stated that it encourages their continuous medical education. Faculty members also recognized the positive impact of technology across these dimensions, with agreement rates ranging from 75% to 86%.

A recent publication evaluating the effectiveness and feasibility of an asynchronous web-based course aimed at improving the skills of Syrian health professionals in evidence-based medicine reported that the course was highly effective, fostering positive attitudes toward continuous medical education training programs [7].

Requirements and Institutional Challenges

The requirements and challenges encountered by the sample members during the integration of technology were highlighted to identify appropriate solutions for overcoming these issues. For example, 33% of medical students agreed that TEL environments impose a financial burden on them, while 57% noted the inadequacy of sufficient and favorable internet services for using these methods, with 22.4% remaining neutral. Furthermore, 82.2% emphasized the need for behavior guidelines to regulate these environments, and 87% highlighted the urgent need for educational communication platforms tailored to each university. Such platforms would facilitate the exchange and accessibility of scientific materials at any time and from any location, as well as enhance communication with professors and peers for sharing information and experiences.

Concerns about requirements and challenges expressed by faculty members showed a high level of agreement; the study indicated that 85.5% of faculty members recognized the urgent

XSL•FO
RenderX

need for institutional support at all levels to optimally implement blended learning in universities. Additionally, 85% emphasized the importance of support from department heads, deans, and decision-makers. Moreover, 90% agreed on the necessity of training and developing faculty members' skills. Notably, 75% of the faculty members identified a lack of teaching experiences as a challenge. A recent study exploring the lived experience of Syrian dentists enrolled in a web-based course on traumatic dental injuries revealed that the course was effective overall, with participants highly motivated to benefit from it despite personal and technical problems and challenges related to the crisis [13].

Curriculum Reform and Stakeholder Role

Our study revealed strong agreement among participants regarding the urgent need to revise curricula to align with the rapid growth of medical knowledge and changing health care environments, as well as with the societal expectations of graduates. These findings are consistent with a study conducted among undergraduate medical students at Ain Shams University in spring 2023, which assessed knowledge and attitudes toward artificial intelligence (AI) and its applications in medical education and practice. The study revealed that medical curricula must evolve to adequately prepare students by providing a comprehensive understanding of AI, including its applications, potential, limitations, and ethical considerations, ensuring students are well-prepared for their future careers [14].

Furthermore, 87% of the participants acknowledged the inadequate infrastructure and services within medical colleges. Additionally, 56.5% noted the insufficient time allocated for lectures to use technology and modern techniques. These findings align with a study aimed at identifying the difficulties faced by the Faculty of Basic Education of Mustansiriya University regarding modern education and the use of technology in teaching. This revealed obstacles faced by faculty members, such as inadequate infrastructure, insufficient training on the use of modern educational technology, and a lack of material and moral incentives and suitable lecture halls [15]. Conversely, no significant differences were found in personal opinions about integrating technology in medical education among faculty members, regardless of their administrative roles or years of experience. Similar findings were reported in a study of faculty members at Al-Ahliyya Amman University, which revealed no significant differences in term of gender, college, academic rank, or years of experience [16]. On the other hand, medical curricula evolve alongside advancements in medical practice, and steps should be taken to integrate technology into the medical education curricula. Both medical faculty and students must understand the concept and implantation of technology integration to maximize its utility.

Responses to the open-ended questions concerning students' opinions about the requirements and challenges of integrating technology in medical education reflected the collective findings from the faculty opinions. One student said that "most teachers present the subject in a narrative, non-interactive manner, without considering presentation methods that engage students." Another student added, "it is very important to provide appropriate technologies for learning purposes in universities." A third student emphasized that "medical education is a very precise field and using technology (when applied correctly) will facilitate both theoretical and possibly clinical studies." A fourth student expressed concerns about the prevalence of technology in medical education, stating "there is an issue with the appropriate use of technology in the learning process; it should be used in a way that avoids monotony and boredom." Another student expressed that "integrating technology into medical education should be a requirement, not an option, for universities and relevant authorities." Regarding barriers, one student said, "the difficulty lies in the often poor internet connectivity, and there is insufficient infrastructure for using technology in education." Another student called for "updating our curricula to keep pace with modern medicine," while another stressed "it is important to have email access for students and faculty to inquire about academic subjects, communicate with professors, and ask about curriculum, study dates, and quizzes."

Policy Context and National Strategy

Ministers of higher education are striving to create educational environments that meet the needs of modern societies by implementing strategies that combine technology-enhanced learning with traditional methods. Over time, traditional learning is expected to diminish as students progress through their academic years [17]. Additionally, there is an emphasis on developing models that assist educators in their teaching programs and provide frameworks for obtaining accreditation for technology-enhanced learning programs, along with standards to maintain high quality [18].

Despite various challenges such as inadequate infrastructure, limited technical capabilities, problematic virtual learning environments, and insufficient technical skills [9], there is a long-term commitment in higher education to support students and enhance their engagement with faculty members, who are continuously seeking to create effective learning activities that produce competent graduates. However, meaningful technology integration cannot be initiated or supported if the institution lacks knowledge, self-efficacy, pedagogical beliefs, and culture conductive to technology integration [19]. When stakeholders' perspectives shift to acknowledge that effective learning requires the appropriate use of assistive technology resources, critical milestones will be achieved in ensuring quality as a means of assurance, accountability, monitoring, and enhancement [18]. Moreover, various aspects related to TEL, such as the quality and reliability of learning management systems and the availability of resources, rely heavily on feedback and verification of student learning and activities [18].

Solutions and Future Work

Our findings are consistent with global trends in the adoption of educational technologies across medical curricula. A range of modalities can complement traditional instruction and address the specific needs and barriers reported by students and faculty in this study. Technology integration in education leverages digital tools to enhance instruction, collaboration, and professional skill development. Mobile learning, powered by smartphones and apps, helps students build digital fluency through video, audio, and storytelling-based platforms [20], supporting the development of communication, clinical, and

XSL•FO
RenderX

problem-solving skills [21]. Social media platforms such as Twitter, Facebook, and YouTube foster interactive learning environments, enabling students to create content, receive feedback, and engage with educational communities [22,23]. Messaging and collaboration tools like WhatsApp, Telegram, and Wikipedia contribute to real-time communication and team-based performance tracking [21]. Flipped classrooms, which involve reviewing video lectures before class [24], transform traditional instruction into active, discussion-based sessions. These in-person sessions often use case-based or team-based learning strategies [25-27], turning the classroom into a dynamic environment for applied learning. Simulation-based learning further complements this approach by offering a safe, controlled space to develop clinical, communication, and decision-making skills [28-31]. Simulations also allow learners to practice rare or complex scenarios, building confidence through experiential methods.

Emerging technologies such as wearable devices and immersive simulations provide hands-free access to instructional content, reduce errors, and support procedural training [32,33]. Virtual and augmented reality enhance engagement and empathy by allowing students to experience clinical and ethical situations in realistic ways, especially in fields like geriatrics and end-of-life care [6,34]. To ensure the safe adoption of these technologies, verification of users' professional identities is essential [35]. AI is also reshaping medical education, offering personalized content delivery and real-time feedback through AI-based tutoring systems [36,37]. However, integrating AI into curricula requires both educator readiness and the thoughtful design of content to ensure meaningful and ethical engagement with the technology [8].

Financial support and infrastructure improvement are essential to facilitating technology integration. Institutions should consider establishing dedicated classrooms, well-equipped laboratories, display devices, and simulation facilities to enhance learning outcomes [38]. These improvements align with findings that robust infrastructure, funding, and technical support are crucial for successful educational technology implementation in higher education [39]. To establish the goals of medical education, providing sufficient financial support and improving infrastructure and services to facilitate the use of technology in medical education are crucial. This includes the creation of dedicated classrooms and well-equipped laboratories, as well as display devices and medical simulation equipment. Additionally, having specialized technicians available to address issues and specialized programmers to support university e-learning platforms is essential to ensuring the easy accessibility of educational materials. Enhancing communication skills through the exchange of opinions and ideas, discussing cases via the chat feature, and simplifying communication with teachers are also vital. Moreover, deploying specialized technicians and programmers to manage e-learning platforms and troubleshoot technical issues ensures easy access to educational materials, consistent with best practices recommended in systematic reviews [39]; dedicated units and offices are urgently needed at each university to monitor, assess, and evaluate the process of integrating technology into education. These units should provide detailed reports and evaluations of all completed operations for inclusion in the university's annual report, which may be submitted to relevant authorities as needed. Training courses for faculty, seminars for stakeholders, and documentation of staff tutorials on the integration process should also be organized. Collaborating with companies that specialize in educational technologies is important, and research should be conducted to create tools, standards, and frameworks for performance assessment, which can be adopted to ensure the quality of higher education. Additionally, further studies are needed to evaluate the impact of modern technologies on the quality and development of higher education in key areas including students, curricula, teacher performance, and college/university administration [40].

Effective technology integration requires a deep understanding of its potential advantages and challenges, as well as adequate training and substantial support and incentives for both educators and learners [8]. Importantly, these technologies should not be viewed as replacements for "real-world" medical educators; instead, they should be considered additional tools for learning [35]. However, it is evident that integrating technology in medical education will take time and require adaptability to rapidly changing technologies [4]. Despite the current situation in Syria, the optimal effectiveness of technology integration in medical education presents a significant opportunity for medical students to overcome educational circumstances and enhance their learning skills, even amid economic, social, and health difficulties, ultimately saving time and effort for both students and educators.

Limitations

This study has several limitations that should be acknowledged. First, the data were collected through self-reported questionnaires, which may be subject to response bias or social desirability effects. Second, although stratified sampling was used to ensure representation from different universities and disciplines, the findings may not be generalizable beyond the Syrian medical education context. Third, while the inclusion of open-ended questions provided some qualitative insight, the study did not use in-depth qualitative methods such as interviews or focus groups, which could have offered a deeper understanding of the participants' experiences. Additionally, because data collection relied on digital platforms, individuals with limited internet access or lower digital literacy may have been underrepresented. These factors should be considered when interpreting the results and applying them to broader contexts.

Conclusion and Broader Implications

As technology integration becomes increasingly common in both medical education and health care, equipping future doctors with digital competencies is critical to enhancing clinical outcomes, improving system efficiency, and fostering lifelong learning. This study provides evidence that TEL not only motivates medical students but also significantly improves their engagement, communication skills, clinical competencies, and access to current scientific knowledge. TEL fosters collaboration, supports research, and helps students remain up to date in a rapidly evolving field, making it essential for the training of future-ready physicians. Interestingly, the study revealed no significant correlations between demographic

XSL•FO
RenderX

variables (such as student specialization or university type or faculty academic rank or teaching experience) and attitudes toward TEL. This suggests that literacy and acceptance of TEL may be more strongly influenced by exposure and hands-on experience than by traditional academic hierarchies or roles, indicating a broad, cross-demographic readiness for technology adoption in medical education. To translate this readiness into meaningful reform, institutions must commit to developing well-supported e-learning platforms, revising outdated curricula, and shifting pedagogical strategies toward student-centered approaches. Building sustainable TEL environments will require institutional investment in infrastructure, ongoing faculty development, and monitoring systems that assess both learning outcomes and community needs.

Importantly, TEL should not be seen as a replacement for traditional education, but rather as a complementary and empowering tool for both learners and educators. Future efforts must include rigorous follow-up studies to evaluate the long-term impact of TEL implementation and inform adaptive strategies that ensure continuous quality improvement. Despite the economic and logistical challenges faced in Syria and similar contexts, the findings of this study offer a hopeful foundation for the effective and scalable integration of technology into medical education.

Acknowledgments

The authors would like to thank Syrian Virtual University for its support, as well as all participating students and faculty who provided comments and suggestions that led to this study's improvement. The authors also acknowledge JMIR Publications for providing a publication fee waiver/discount for this paper.

Data Availability

The data sets generated and analyzed during this study are not publicly available due to participant confidentiality and institutional restrictions. However, deidentified data may be made available from the corresponding author upon reasonable request and with appropriate ethical clearance.

Authors' Contributions

SH is the principal researcher, collected participants' data, and analyzed the results. Both MD and AB supervised the study; participated in the design and implementation of the study; and contributed to the study design, execution, analysis, and interpretation. The authors reviewed and agreed on all versions of the article before submission, agreed on the journal to which the article was submitted, and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Evaluating technology integration in Syrian medical education (student questionnaire).
[PDF File, 137 KB - formative_v9i1e76958_app1.pdf]

Multimedia Appendix 2
Evaluating technology integration in Syrian medical education (faculty questionnaire).
[PDF File, 202 KB - formative_v9i1e76958_app2.pdf]

References

1. Naeem NIK, Yusoff MSB, Hadie SNH, Ismail IM, Iqbal H. Understanding the functional components of technology-enhanced learning environment in medical education: a scoping review. MedSciEduc 2023;33(2):595-609. [doi: 10.1007/s40670-023-01747-6]
2. Jha S. Role of technology in medical education: SWOC analysis. SBV Journal of Basic, Clinical and Applied Health Science 2022 Jun 4;5(1):19-21. [doi: 10.5005/jp-journals-10082-03131]
3. Cabero-Almenara J, De-La-Portilla-De-Juan F, Barroso-Osuna J, Palacios-Rodríguez A. Technology-enhanced learning in health sciences: improving the motivation and performance of medical students with immersive reality. Appl Sci (Basel) 2023 Jul 21;13(14):8420. [doi: 10.3390/app13148420]
4. Michalik B, Sęk M, Szypuła A, Hajduk-Maślak K, Skóra A, Galasińska I. New technological developments in medical education. J Educ Health Sport 2024;60:204-220. [doi: 10.12775/JEHS.2024.60.014]
5. Al-Sbei R, Ataya J, Jamous I, Dashash M. The impact of a web-based restorative dentistry course on the learning outcomes of dental graduates: pre-experimental study. JMIR Form Res 2024 Mar 5;8(1):e51141. [doi: 10.2196/51141] [Medline: 38441921]

XSL•FO
RenderX

6. Dyer E, Swartzlander BJ, Gugliucci MR. Using virtual reality in medical education to teach empathy. J Med Libr Assoc 2018 Oct;106(4):498-500. [doi: 10.5195/jmla.2018.518] [Medline: 30271295]
7. Kenjrawi Y, Dashash M. The first asynchronous online evidence-based medicine course for Syrian health workforce: effectiveness and feasibility pilot study. JMIR Form Res 2022 Oct 25;6(10):e36782. [doi: 10.2196/36782] [Medline: 36282556]
8. Wood EA, Ange BL, Miller DD. Are we ready to integrate artificial intelligence literacy into medical school curriculum: students and faculty survey. J Med Educ Curric Dev 2021;8:23821205211024078. [doi: 10.1177/23821205211024078] [Medline: 34250242]
9. Dashash M. The Implementation of Online Medical Education in the Arab World 2023:271-294. [doi: 10.1007/978-3-031-33568-6_14]
10. Latifeh Y, Dashash M. A critical analysis and a suggested reform of psychiatric curricula in medical faculties during Syrian crisis. American Journal of Health Research 2016;4(6) [FREE Full text]
11. Almasri B, Dashash M. Evaluating the effectiveness of gastroenterology e-module for undergraduate medical students, a pilot study from Syria. ResearchSquare. Preprint posted online on Aug, 2021. [doi: 10.21203/rs.3.rs-808231/v1]
12. Shamsy E, Dashash M. The effectiveness of online course about pain management and local anesthesia methods for undergraduate dental students. Azerbaijan Medical Association Journal 2022;62(10) [FREE Full text]
13. Dashash M, Alkhadragy R, Scanlan GM. A phenomenological exploration of experience of Syrian dentists with online course “Traumatic dental injuries”. Heliyon 2024 Jul 15;10(13):e34045. [doi: 10.1016/j.heliyon.2024.e34045] [Medline: 39071621]
14. Khater AS, Zaaqoq AA, Wahdan MM, Ashry S. Knowledge and attitude of Ain Shams University medical students towards artificial intelligence and its application in medical education and practice. Educational Research and Innovation Journal 2023 Jul 1;3(10):29-42. [doi: 10.21608/erji.2023.306718]
15. Issa RI, Saleh AJ. The difficult application of modern education technology from the point view of the members of a teaching staff. JUBPAS 2019;27(1):206-227. [doi: 10.29196/jubpas.v27i1.2114]
16. AlNaimi S, Mahmoud H, Ghoneim F. The degree to which the use of modern educational technologies on the quality of education and its development in Al-Ahliyya Amman University from the viewpoint of faculty members. AJRS 2020;23(2):65-76. [doi: 10.35875/1105-023-002-005]
17. Al-Momani MO, Rababa EM. Mixed education and quality standard in the university teaching: a theoretical study. Indonesian J Educ Res Technol 2022;2(3):155-174. [doi: 10.17509/ijert.v2i3.43655]
18. Casanova D, Moreira A. A model for discussing the quality of technology-enhanced learning in blended learning programmes. International Journal of Mobile and Blended Learning 2017 Oct;9(4):1-20. [doi: 10.4018/IJMBL.2017100101]
19. Ertmer PA, Ottenbreit-Leftwich AT. Teacher technology change: how knowledge, confidence, beliefs, and culture intersect. Journal of Research on Technology in Education 2010 Mar;42(3):255-284. [doi: 10.1080/15391523.2010.10782551]
20. Briz-Ponce L, Juanes-Méndez JA, García-Peñalvo FJ, Pereira A. Effects of mobile learning in medical education: a counterfactual evaluation. J Med Syst 2016 Jun;40(6):136. [doi: 10.1007/s10916-016-0487-4] [Medline: 27098779]
21. Popoiu MC, Grosseck G, Holotescu C. What do we know about the use of social media in medical education? Procedia - Social and Behavioral Sciences 2012;46:2262-2266. [doi: 10.1016/j.sbspro.2012.05.466]
22. Sharma S, Tannor EK, Duarsa R, et al. Twitter as educational tool: a global web-based cross-sectional study on social media behavior of nephrologists. Kidney Int Rep 2023 Nov;8(11):2469-2473. [doi: 10.1016/j.ekir.2023.08.009] [Medline: 38025224]
23. Davis WM, Ho K, Last J. Advancing social media in medical education. CMAJ 2015 May 19;187(8):549-550. [doi: 10.1503/cmaj.141417] [Medline: 25852033]
24. Bridges J, Stefaniak J, Baaki J. A formative design examining the effects of elaboration and question strategy with video instruction in medical education. J Form Des Learn 2018 Dec;2(2):129-143. [doi: 10.1007/s41686-018-0025-5]
25. Tune JD, Sturek M, Basile DP. Flipped classroom model improves graduate student performance in cardiovascular, respiratory, and renal physiology. Adv Physiol Educ 2013 Dec;37(4):316-320. [doi: 10.1152/advan.00091.2013] [Medline: 24292907]
26. Hodgson P, Hui D. Enriching Literacy Learning with Mobile Devices in a Flipped Classroom 2017:827-832.
27. Phillips J, Wiesbauer F. The flipped classroom in medical education: a new standard in teaching. Trends Anaesth Crit Care 2022 Feb;42:4-8. [doi: 10.1016/j.tacc.2022.01.001] [Medline: 38620968]
28. Bradley P, Postlethwaite K. Simulation in clinical learning. Med Educ 2003 Nov;37 Suppl 1:1-5. [doi: 10.1046/j.1365-2923.37.s1.1.x] [Medline: 14641632]
29. Bradley P. The history of simulation in medical education and possible future directions. Med Educ 2006 Mar;40(3):254-262. [doi: 10.1111/j.1365-2929.2006.02394.x] [Medline: 16483328]
30. Issenberg SB, McGaghie WC, Petrusa ER, Lee Gordon D, Scalese RJ. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. Med Teach 2005 Jan;27(1):10-28. [doi: 10.1080/01421590500046924] [Medline: 16147767]
31. Studnicka K, Zarzycka D, Dziura M, Kijowska A. Basic guidelines for high-fidelity medical simulation. Zenodo 2018 Dec 17. [doi: 10.5281/ZENODO.2365275]

XSL•FO
RenderX

32. Khosravi S, Bailey SG, Parvizi H, Ghannam R. Learning enhancement in higher education with wearable technology. Sensors (Basel) 2022 Oct 8;22(19):7633. [doi: 10.3390/s22197633]
33. Sultan N. Reflective thoughts on the potential and challenges of wearable technology for healthcare provision and medical education. Int J Inf Manage 2015 Oct;35(5):521-526. [doi: 10.1016/j.ijinfomgt.2015.04.010]
34. López-Ojeda W, Hurley RA. Extended-reality technologies: an overview of emerging applications in medical education and clinical care. JNP 2021 Jul;33(3):A4-177. [doi: 10.1176/appi.neuropsych.21030067]
35. Ahuja AS, Polascik BW, Doddapaneni D, Byrnes ES, Sridhar J. The digital metaverse: applications in artificial intelligence, medical education, and integrative health. Integr Med Res 2023 Mar;12(1):100917. [doi: 10.1016/j.imr.2022.100917] [Medline: 36691642]
36. Chan KS, Zary N. Applications and challenges of implementing artificial intelligence in medical education: integrative review. JMIR Med Educ 2019 Jun 15;5(1):e13930. [doi: 10.2196/13930] [Medline: 31199295]
37. Wartman SA, Combs CD. Medical education must move from the information age to the age of artificial intelligence. Acad Med 2018 Aug;93(8):1107-1109. [doi: 10.1097/ACM.0000000000002044] [Medline: 29095704]
38. Rizana AF, Hediyanto U, Ramadhan F, Kurniawati A. E-learning success determinants in higher education: a systematic literature review from users' perspective. IOP Conf Ser: Mater Sci Eng 2020 Apr 1;830(3):032012. [doi: 10.1088/1757-899X/830/3/032012]
39. Chugh R, Turnbull D, Cowling MA, Vanderburg R, Vanderburg MA. Implementing educational technology in Higher Education Institutions: a review of technologies, stakeholder perceptions, frameworks and metrics. Educ Inf Technol 2023 Dec;28(12):16403-16429. [doi: 10.1007/s10639-023-11846-x]
40. Cabaleiro-Cerviño G, Vera C. The impact of educational technologies in higher education. GIST 2020 Jun 11;20:155-169. [doi: 10.26817/16925777.711]

Abbreviations

AI: artificial intelligence
TEL: technology-enhanced learning

Edited by A Mavragani; submitted 05.05.25; peer-reviewed by C Burgos, M Lepore, Y Nenko; revised version received 27.06.25; accepted 05.07.25; published 08.08.25.

Please cite as:
Hassoun S, Dashash M, Baddour A
Technology Integration in Syrian Medical Education From the Perspective of Students and Faculty: A Cross-Sectional Evaluation
JMIR Form Res 2025;9:e76958
URL: https://formative.jmir.org/2025/1/e76958
doi:10.2196/76958

© Subhiya Hassoun, Mayssoon Dashash, Adnan Baddour. Originally published in JMIR Formative Research (https://formative.jmir.org), 8.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Development of a Sham Smartphone App for Opioid Use Disorder: Acceptability and Suitability Study

Kierstyn S Gallegos[1*], MPH; Jennifer S Potter[1,2*], MPH, PhD; Van L King[1,2*], MD; Gregg Siegel[3*], MS; Leslie H Siegel[3*], MFA; Elise N Marino[1,2*], PhD

[1]Be Well Institute on Substance Use and Related Disorders, University of Texas Health Science Center at San Antonio, 5109 Medical Drive, San Antonio, TX, United States

[2]Department of Psychiatry and Behavioral Sciences, University of Texas Health Science Center at San Antonio, San Antonio, TX, United States

[3]Biomedical Development Corporation, San Antonio, TX, United States

[*]all authors contributed equally

Corresponding Author:
Elise N Marino, PhD
Be Well Institute on Substance Use and Related Disorders, University of Texas Health Science Center at San Antonio, 5109 Medical Drive, San Antonio, TX, United States

Abstract

Background: Despite having evidence-based medication for opioid use disorder (OUD), dropout is one of the most common issues noted with this treatment. Prescription digital therapeutics, which are app-based interventions prescribed by a health care professional, have the potential to increase adherence to medication for OUD and retention while overcoming treatment barriers, including provider capacity and patient access. Using a sham app as a control condition for a randomized clinical trial is an innovative method to establish the true efficacy of these apps.

Objective: This study included the development and testing of a sham smartphone app for OUD.

Methods: After the sham app was developed, participants were enrolled in a 4-week trial examining the use and suitability of the sham app as a control condition. Criteria for determining suitability included (1) participants believing the sham app is an active intervention and (2) participants experiencing no clinical improvements in depression severity or quality of life after using the sham app. Self-reported depression severity and quality of life were captured before and after using the sham app. A user satisfaction survey and semistructured interviews were conducted at the end of the study. Quantitative analyses included paired 2-tailed *t* tests. The semistructured interviews were conducted with 20 of the 21 participants, and these interviews were analyzed using rapid qualitative analysis.

Results: Overall, 21 participants ($mean_{age}$ 42.0, SD 6.4 years; female: n=9, 43% and male: n=12, 57%) were enrolled. The average number of app log-ins was 17.8 (SD 10.6; range 1-41). There were 2 participants who only logged in 1 time, and 15 (71%) participants completed the goal of logging in an average of 3 times per week. No significant differences were found in depression severity (P=.50) or quality of life (quality of life: P=.42, physical health: P=.58, psychological health: P=.07, environmental health: P=.44, and social relationships: P=.86) after using the sham app. Of the 20 participants who completed the semistructured interview, 19 (95%) believed that they were using an active intervention. The user satisfaction survey revealed high overall satisfaction with the sham app with a score of 91%. Qualitative analyses revealed several recurring themes, including perceived value and impact, potential for behavior change, use patterns and engagement, perspective and usability, and perceptions of authenticity.

Conclusions: Our sham app met our a priori criteria for suitability as a sham app. No clinical improvements from baseline were observed at the end of the study period, and all but 1 participant believed that they were using an active intervention. Demonstrating that this sham app is suitable as a control condition elevates the rigor of randomized clinical trials and ensures the efficacy of prescription digital therapeutics.

(JMIR Form Res 2025;9:e71105) doi:10.2196/71105

KEYWORDS

prescription digital therapeutic; sham software; believable; placebo; digital health

XSL•FO
RenderX

Introduction

In 2017, the estimated total economic burden of opioid use disorder (OUD) and opioid-related overdose deaths in the United States was US $1.02 trillion [1,2]. The high prevalence of OUD [3] coupled with the significant health care cost of treating it has resulted in the need for an innovative and cost-effective approach to care that will complement existing evidence-based treatment.

Medication for opioid use disorder (MOUD), including methadone, buprenorphine, and naltrexone, is standard of care with the latter 2 available in office-based settings [4]. Despite robust evidence demonstrating the effectiveness of MOUD [5], there are limitations to receiving this in-office treatment [6]. Wait times are often extended due to the availability of providers working within standard clinic operating hours [7]. Such delays in treatment are detrimental and potentially fatal [8]. Being unable to receive care in times of crisis and a strictly regulated time allotted for patient sessions [9] are significant barriers to entering and maintaining engagement in treatment, creating potential gaps in service [10]. Reliable transportation, childcare, and clinic proximity also serve as barriers.

Overcoming these barriers to receiving in-office treatment necessitates enhancing MOUD treatment with evidence-based interventions that are not limited by clinician availability or patient access. App-based interventions, which are easily accessible and available at any time, have demonstrated efficacy for those with OUD [11-14]. Prescription digital therapeutics are app-based interventions that are used to treat or manage physical or mental health conditions, such as OUD. They can be used as monotherapy or as an adjunctive treatment. Unique from other app-based interventions, prescription digital therapeutics must be prescribed by a health care professional and are eligible for reimbursement from some health insurance companies. To receive this designation, a prescription digital therapeutic must receive approval from the US Food and Drug Administration (FDA) after demonstrating its safety and efficacy in a clinical trial. These apps are empirically validated interventions, a classification not attributed to most freely available apps, which increases the likelihood that patients will benefit from them. Although prescription digital therapeutics require a prescription and thus at least 1 clinician visit, once given access to the app, patients are not limited in their engagement with the intervention.

In 2018, the FDA approved reSET-O, an 84-day prescription digital therapeutic, to improve retention in outpatient treatment for OUD [15]. reSET-O received FDA approval after the completion of a randomized trial that used treatment as usual for the control condition. In doing so, the efficacy demonstrated by reSET-O may be attributable to having access to an app rather than the intervention provided within reSET-O itself. This uncertainty is inherent in an unblinded study design, highlighting the essential role of using a sham app.

A sham condition should be indistinguishable from the experimental condition in appearance and functionality. It must also provide no therapeutic benefit to serve as an acceptable control [16]. Distinct from drug trials that use an inactive substance, simply engaging with a digital platform poses risk for measurable improvement [17], making the development of a sham app much more challenging. The magnitude of naming an app-based intervention an FDA-approved prescription digital therapeutic nevertheless requires a blinded study design.

Ensuring the efficacy of prescription digital therapeutics is paramount to the complexities of designing a sham app. Recognizing this, the field is evolving. One of the first to use a sham condition in pursuit of receiving FDA approval as a prescription digital therapeutic is KIOS, an innovative app that was developed using contextual design methodology to augment MOUD treatment [16]. As opposed to the module-based intervention provided by reSET-O, the KIOS app delivers customized, patient-driven behavioral interventions in real time [9].

The study described here was designed to test the suitability of the KIOS sham app as a control condition for a randomized trial. A priori criteria for determining suitability included (1) participants believing the sham app is an active intervention and (2) participants experiencing no clinical improvements in depression severity or quality of life after using the sham app. Being among the first to develop a sham app, we are transforming the standard of clinical trials and consequently prescription digital therapeutics.

Methods

Overview

This study used a mixed methods design to evaluate whether our sham app could serve as a believable control condition in a randomized controlled trial as well as to evaluate participant perspectives of usability and satisfaction with the app.

Ethical Considerations

This fully web-based study was approved by the University of Texas Health Science Center at San Antonio Institutional Review Board (protocol 20230232HU). Participants provided written informed consent prior to participation. During the consent process that was conducted remotely, participants were informed that the purpose of the study was to evaluate engagement, usability, and user satisfaction with a mobile app that offered information about OUD and other substance use disorders. Participants received US $200 for their participation in the form of a ClinCard MasterCard. Compensation was automatically credited after completion of the study. All data were deidentified prior to being analyzed.

Participants

Participants were all diagnosed with OUD, currently receiving MOUD, and in stable outpatient treatment for at least 4 weeks at the time of enrollment. They were recruited from outpatient MOUD treatment facilities in Texas. Recruitment strategies included placing flyers in clinic waiting rooms, announcements during web-based trainings sponsored by our staff, sending emails about the study to our network of MOUD providers, and outreach personnel speaking with people in line at methadone clinics. Individuals interested in participating in the study were directed to contact study staff via phone. Inclusion criteria for

XSL•FO
RenderX

the study included (1) male or female outpatients 18 years of age or older; (2) diagnosis of OUD; (3) currently stable in OUD outpatient treatment, defined as medication adherence and attending outpatient appointments for 4 weeks or longer; and (4) ability to access the KIOS sham app via a smartphone or tablet. Participants were excluded if they (1) were unwilling or unable to comply with study requirements, (2) had a psychiatric or medical disorder interfering with the ability to use the app, (3) were incarcerated, or (4) were pregnant.

Measures

The Patient Health Questionnaire-9 (PHQ-9) is a 9-item validated self-report assessment that was used to measure depression severity over the past 2 weeks [18]. The World Health Organization Quality of Life-BREF (WHOQOL-BREF) assessment was used to assess self-reported quality of life across 4 domains: physical health, psychological health, social relationships, and environment over the past 2 weeks [19]. A User Satisfaction Survey was administered to assess the degree of satisfaction with the sham app [20]. The WHOQOL-BREF and PHQ-9 were chosen as outcomes because they assess complementary facets of recovery (eg, depression has a high comorbidity with alcohol and drug use) [21-23]. Improvements in depression and quality of life are some potential benefits that may arise when using an app-based intervention; thus, treatment targets for the app are not limited to opioid use but rather to all aspects of recovery that may impact opioid use and retention in MOUD treatment.

Procedure

As this was a web-based study, communication between research assistants and participants occurred via telephone and Zoom (Zoom Video Communications). No changes were made to the app after the trial began.

This was a 1-sample study with no control group using a pre- and posttest design. Prior to engaging with the sham app, participants completed the PHQ-9 and WHOQOL-BREF delivered via an electronic data capture system. Surveys were completed while on the phone with research assistants in the event that participants had any questions. They were then given access to the app for 4 weeks, during which time they were instructed to engage with the app at a minimum of 3 times per week and as often as once daily. Participants were encouraged to explore all the functions within the app. At the end of the 4 weeks, they completed the PHQ-9 and WHOQOL-BREF as well as the User Satisfaction Survey.

Upon completion of the study, semistructured interviews were conducted to gather detailed information on believability, usability, and satisfaction with the sham app. Six questions were asked assessing app believability and satisfaction (eg, “Tell us what you thought of KIOS?” and “If you weren’t in the study anymore, would you use the app? Why or why not?”). Following these questions, a debriefing statement was delivered informing participants of the sham nature of the app. Once informed, participants were asked to reshare their perception of the sham app and indicate whether they suspected it was anything other than what it was presented as.

Sham App Development

The sham app was developed for use in a randomized controlled trial testing the KIOS app, and it was created at the request of the FDA. The proposed design of the sham app was presented to the FDA for review prior to proceeding with development.

The sham app was meticulously designed to convince the user that it was an active intervention despite being designed to have no therapeutic value. To create a believable sham app, the active KIOS design was replicated to create a visually similar version that lacked key therapeutic components. The nonspecific factor of the sham app mirrored the active condition by replicating the layout and functionality of account log-in, color palette, button shapes, emergency contact functions, and landing pages. Features that incorporated elements of an active intervention were removed from the sham and were replaced with general drug and alcohol educational content taken from the National Institute on Drug Abuse (NIDA).

We identified, selected, and analyzed information about drugs using specific criteria. Drug-related facts were explored through key questions addressing the drug’s definition, use, short- and long-term effects, physical and chemical properties, and health impacts. Questions using a true or false format were included in the app, and they were designed to generate generic feedback that was presented to participants. These questions were developed by selecting topics from the NIDA website, identifying relevant passages, and rephrasing or editing them for clarity and conciseness. Correct “true” answers were based on accurate passages, while “false” answers included explanations derived from the content available from NIDA. The process also involved assessing the grade level of each question to ensure readability at or below a sixth-grade reading level, excluding drug names, and documenting both the grade level and source for each question. Participants were presented with the true or false questions 3 times a week followed by questions assessing their level of interest in the educational content being provided. Because education alone is unlikely to change long-term behavior [24], the final product was a fully functional, education-based learning tool that had no active components capable of creating meaningful improvements in clinical outcomes.

The second fundamental difference between the active and sham versions of KIOS is the reminder push notification. The push notification frequency was customizable to the user’s preference, ranging from daily, every other day, to not at all. The active KIOS app that was developed is a tailored, patient-centered app, thus necessitating a similar approach to the reminder notifications. Participants using the active version receive personalized reminders (eg, “Savor this one moment with a slow, deep breath. Then, connect with yourself further in KIOS”). In contrast, the sham app was designed to be standardized across users, including the reminders that used generic language (eg, “This is the KIOS reminder you set for yourself”). This difference is essential because the personalized reminders are designed to simulate the therapeutic intent of an active intervention while generic reminders do not.

XSL•FO
RenderX

Statistical Analysis

Quantitative analyses included paired 2-tailed *t* tests to assess differences in PHQ-9 and WHOQOL-BREF scores after using the sham app. To determine overall satisfaction with the app, a total percentage of the aggregate responses of the User Satisfaction Survey was calculated.

Rapid qualitative analysis was used to systematically review the semistructured interviews that were conducted following completion of the study. Transcripts were thoroughly reviewed to identify recurring themes. Each question was separated with participant responses to the question being condensed into bullet points. After grouping the questions into overarching domains (eg, use patterns and engagement and perceptions of authenticity), recurrent themes were cataloged within the domains to capture larger research focus areas. The final summary template that was used to aggregate participant responses was standardized across participants to ensure consistency.

A matrix was then developed to display the condensed summaries. The columns were organized by domain with subsections for each question in the domain. Each row in the matrix was dedicated to a participant interview. The matrix was filled using the data from the summaries and analyzed across rows and down columns to identify central themes, allowing for a comprehensive understanding of key findings from the interviews. The qualitative analysis was completed manually, and the quantitative analyses were completed in Microsoft Excel.

Results

Consistent with our recruitment goal, 21 participants provided informed consent, used the app during the 4-week trial, and completed all pre- and postassessments. The average age was 42.0 (SD 6.4) years, 9 (43%) were female, and 12 (57%) were male.

Among the full sample, the average number of app log-ins was 17.8 (SD 10.6; range 1-41). There were 2 participants who only logged in 1 time, and the rest of the sample logged in a minimum of 6 times. A total of 15 (71%) participants completed the goal of logging in an average of 3 times per week. Figure 1 displays the app log-in landing page. Quantitative analyses indicated that there were no statistically significant differences in depression severity or quality of life from pre- to postapp use (Table 1). The User Satisfaction Survey indicated high overall satisfaction with the sham app, with an average score of 91%.

Of the full sample, 20 (95%) completed the end of study semistructured interview. Qualitative analysis of these 20 participants revealed several recurring themes, including perceived value and impact, potential for behavior change, use patterns and engagement, perspective and usability, and perceptions of authenticity (Table 2). Participants characterized the sham app as simple, user-friendly, and educational. Participants also reported high satisfaction with the software and increased app use, as they became more familiar with it. Participants valued the educational content and increased awareness of drug-related health risks. There was a notable desire among users to retain access to the software beyond the study. Despite positive feedback, most participants did not believe that the sham app would change their substance use.

In total, 19 of 20 (95%) participants reported during the semistructured interview that they believed they were enrolled in an active intervention.

XSL•FO
RenderX

Figure 1. App log-in landing page.

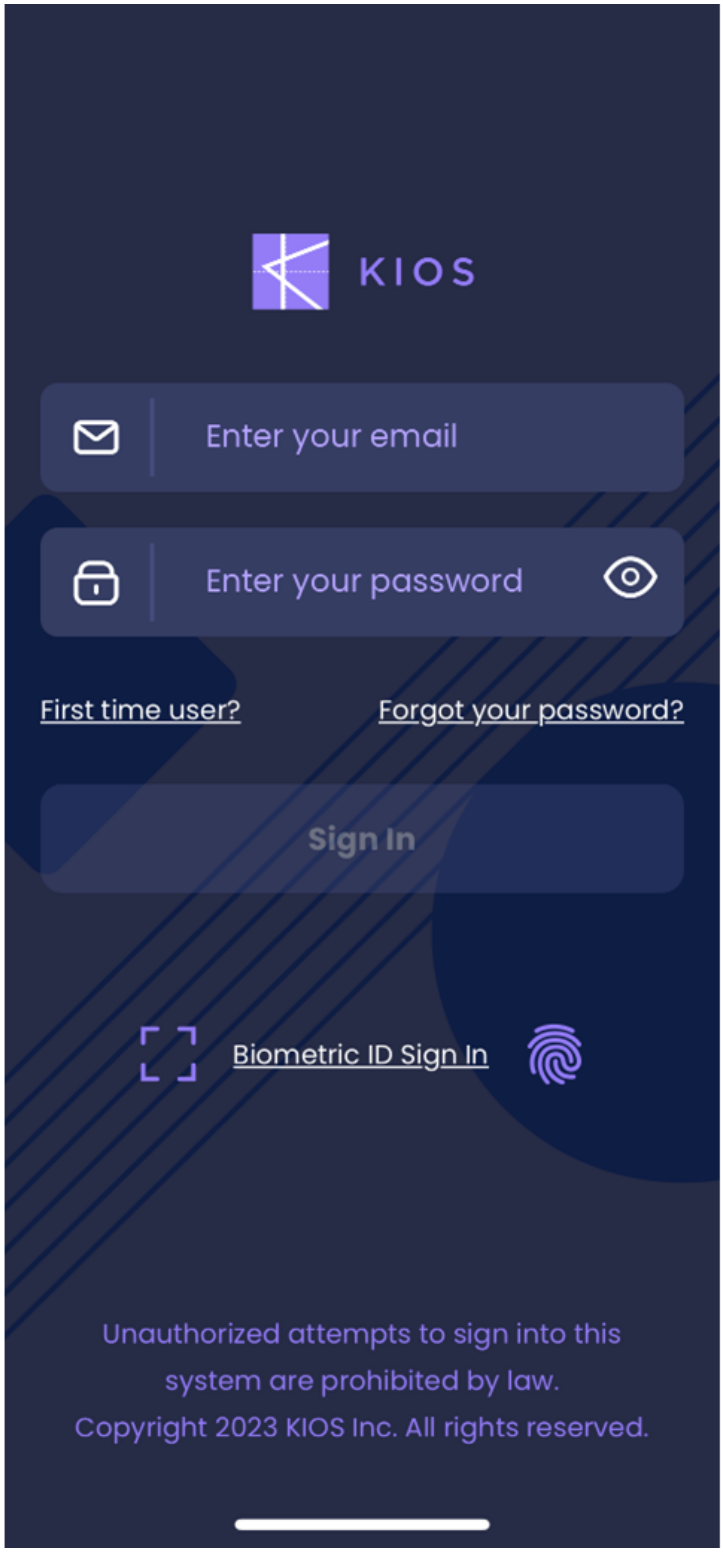

XSL•FO
RenderX

Table . Bivariate analyses examining depression severity and quality of life from pre- to postuse of the sham app.

Measure	Mean (SD)	Coefficient	*P* value
PHQ-9[a]		2.1	.50
Pre	10.7 (37.6)		
Post	11.8 (33.2)		
WHOQOL-BREF[b]			
Quality of Life		2.1	.42
Pre	66.7 (411.5)		
Post	69.6 (415.2)		
Physical health		2.1	.58
Pre	56.6 (255.5)		
Post	55.1 (263.5)		
Psychological health		2.1	.07
Pre	64.9 (209.3)		
Post	58.9 (446.7)		
Environmental health		2.1	.44
Pre	59.7 (338.8)		
Post	56.7 (481.7)		
Social relationships		2.1	.86
Pre	57.1 (606.2)		
Post	56.3 (693.8)		

[a]PHQ-9: Patient Health Questionnaire-9.

[b]WHOQOL-BREF: World Health Organization Quality of Life-BREF.

Table . Qualitative themes and quotes from participants about their perception of the sham app following software engagement.

Theme	Representative quotations
Perceived value and impact	• "I took it as a learning process… like the long-term effects like people catching the hepatitis, HIV, blood clots, high blood pressure. And that stuck in my head." • "Like learning, learning certain things that I've never thought I would."
Potential for behavioral change	• "The version that I used to if I was a full drug blown back into drugs again would that help me in trying to recover? No." • "I don't think it changed the way I was or the way I am."
Use patterns and engagement	• "I found myself wanting to know more ... I found myself using it more and more often." • "I would say I would still use it because of the information."
Perspective and usability	• "... it was super user friendly and some of the questions that were I was I was not expecting." • "It was like no difficulty with it at all. Like literally."
Perceptions of authenticity	• "I didn't know that there was more to it, but I mean, I'm glad." • "I thought that was it, you know, I mean. I didn't think it was gonna be anything different."

XSL•FO
RenderX

Discussion

Principal Findings

Our findings indicated no significant change in self-reported depression severity or quality of life after using the sham app, and all but 1 participant believed that the app was an active intervention. Together, these findings meet our a priori criteria that the sham app appears to be a suitable control condition for a randomized trial.

Prescription digital therapeutics should be held to the highest standard of research. This makes blinded research designs crucial for determining true efficacy before prescribing these apps to patients. Because a randomized controlled trial is the optimal method for testing the evidence base of a treatment, a sham app should be expected. The creation of a sham app for testing a prescription digital therapeutic is a relatively new and modernizing advancement in research designs [25]. Recognized by the FDA as a uniquely difficult undertaking [17], this novel approach addresses the need for a more scientifically robust methodology in the study of prescription digital therapeutics.

With an appealing user interface, the design and execution of the sham software were pivotal in presenting the app to be nearly indistinguishable from an active intervention. The sham app was reported to be engaging, yet nontherapeutic. The latter signifies the possibility to surmount the potential benefits inherent in engaging with an interactive digital platform. The identical user interface, testing prior to using it in a randomized controlled trial, and positive engagement results we demonstrated here overcome previously cited limitations in the study and use of sham apps [25]. This sham app holds significant potential for OUD treatment and recovery, paving the way for creating studies to identify effective prescription digital therapeutics. Without it, the gap between OUD and a good prognosis will remain.

The success of the sham app extends beyond the immediate study, and it represents an opportunity to elevate the standards of study designs for all digital health research. Using the framework we have demonstrated here can further the development of future sham apps. We have shown some components of a sham app that make it believable without producing clinical improvements (eg, generic push reminders and using education in place of the active intervention). These components can be modified and adapted for future studies testing app interventions, making our framework appropriate for creating a sham app that will be suitable as a control condition. This innovation holds significant implications not only for KIOS but also for the broader research field, advancing the science, testing, and deployment of prescription digital therapeutics.

Limitations

There are limitations that should be mentioned. This study included a small sample size, and the study duration was short. Because our initial pilot study demonstrated the feasibility and acceptability of the KIOS app and this study was designed to create a comparison, sham app for future testing of the KIOS app, we used a pre- and posttest design with no control group [9]. Participants were not required to see a medical provider to be included, and we do not have information regarding their opioid use or adherence to their medication regimen. There may have been variations in interview techniques among researchers. It is possible that a sham app may have some therapeutic benefits among those early in recovery. Finally, all participants were in Texas, making additional research in other geographic regions necessary. Despite these caveats, the insight gained here is invaluable.

Conclusions

With the rapid integration of technology in health care, there has been a call for more methodologically rigorous studies. Developing this sham app marks a historic accomplishment in the emerging field of prescription digital therapeutics. This advancement will drive high-quality research, providing the foundation for a new placebo that will revolutionize the future of clinical trials.

Acknowledgments

Funding was provided by the National Institute on Drug Abuse (award R42DA054881). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute on Drug Abuse.

Data Availability

The dataset generated and analyzed during this study is not publicly available.

Conflicts of Interest

ENM and JSP have served as consultants to Biomedical Development Corporation. GS and LHS work for or own stocks or stock options of the company that developed or owned the app that was analyzed in this paper.

References

1. Florence C, Luo F, Rice K. The economic burden of opioid use disorder and fatal opioid overdose in the United States, 2017. Drug Alcohol Depend 2021 Jan 1;218:108350. [doi: 10.1016/j.drugalcdep.2020.108350] [Medline: 33121867]

XSL•FO
RenderX

2. Hsu DJ, McCarthy EP, Stevens JP, Mukamal KJ. Hospitalizations, costs and outcomes associated with heroin and prescription opioid overdoses in the United States 2001-12. Addiction 2017 Sep;112(9):1558-1564. [doi: 10.1111/add.13795] [Medline: 28191702]
3. Opioid data analysis and resources. Centers for Disease Control and Prevention. 2023 Aug. URL: https://www.cdc.gov/overdose-prevention/?CDC_AAref_Val=https://www.cdc.gov/opioids/data/analysis-resources.html [accessed 2024-04-28]
4. Information about medication-assisted treatment (MAT). US Food and Drug Administration. 2023 May. URL: https://www.fda.gov/drugs/information-drug-class/information-about-medications-opioid-use-disorder-moud [accessed 2024-04-28]
5. Becker SJ, Bowen CA, Reed EN, et al. Sustainment of contingency management within opioid treatment programs: COVID-related barriers and innovative workflow adaptations. Drug Alcohol Depend Rep 2021 Dec;1:100003. [doi: 10.1016/j.dadr.2021.100003] [Medline: 36277301]
6. Wolfe MK, McDonald NC, Holmes GM. Transportation barriers to health care in the United States: findings from the National Health Interview Survey, 1997-2017. Am J Public Health 2020 Jun;110(6):815-822. [doi: 10.2105/AJPH.2020.305579] [Medline: 32298170]
7. Austin EJ, Chen J, Briggs ES, et al. Integrating opioid use disorder treatment into primary care settings. JAMA Netw Open 2023 Aug 1;6(8):e2328627. [doi: 10.1001/jamanetworkopen.2023.28627] [Medline: 37566414]
8. Issue brief: lowering the barriers to medication treatment for people with opioid use disorder. Penn LDI. 2022. URL: https://ldi.upenn.edu/our-work/research-updates/lowering-the-barriers-to-medication-treatment-for-people-with-opioid-use-disorder/#:~:text=This%20brief%20summarizes%20the%20barriers,high%2Dbarrier%2C%20standard%20care [accessed 2024-08-08]
9. King VL Jr, Siegel G, Priesmeyer HR, Siegel LH, Potter JS. Development and evaluation of a digital app for patient self-management of opioid use disorder: usability, acceptability, and utility study. JMIR Form Res 2024 Apr 1;8:e48068. [doi: 10.2196/48068] [Medline: 38557501]
10. Browne T, Priester MA, Clone S, Iachini A, DeHart D, Hock R. Barriers and facilitators to substance use treatment in the rural south: a qualitative study. J Rural Health 2016;32(1):92-101. [doi: 10.1111/jrh.12129] [Medline: 26184098]
11. DeFulio A, Rzeszutek MJ, Furgeson J, Ryan S, Rezania S. A smartphone-smartcard platform for contingency management in an inner-city substance use disorder outpatient program. J Subst Abuse Treat 2021 Jan;120:108188. [doi: 10.1016/j.jsat.2020.108188] [Medline: 33298295]
12. DeFulio A, Brown HD, Davidson RM, Regnier SD, Kang N, Ehart M. Feasibility, acceptability, and preliminary efficacy of a smartphone-based contingency management intervention for buprenorphine adherence. Behav Anal Pract 2023 Jun;16(2):450-458. [doi: 10.1007/s40617-022-00730-8] [Medline: 37187840]
13. Velez FF, Anastassopoulos KP, Colman S, et al. Reduced healthcare resource utilization in patients with opioid use disorder in the 12 months after initiation of a prescription digital therapeutic. Adv Ther 2022 Sep;39(9):4131-4145. [doi: 10.1007/s12325-022-02217-y] [Medline: 35799080]
14. Marino EN, Karns-Wright T, Perez MC, Potter JS. Smartphone app-based contingency management and opioid use disorder treatment outcomes. JAMA Netw Open 2024 Dec 2;7(12):e2448405. [doi: 10.1001/jamanetworkopen.2024.48405] [Medline: 39621343]
15. FDA clears mobile medical app to help those with opioid use disorder stay in recovery programs. US Food and Drug Administration. 2020 Mar. URL: https://www.prnewswire.com/news-releases/fda-clears-mobile-medical-app-to-help-those-with-opioid-use-disorder-stay-in-recovery-programs-300762887.html [accessed 2024-05-29]
16. Bowden CL, Priesmeyer R, Tohen M, et al. Development of a patient-centered software system to facilitate effective management of bipolar disorder. Psychopharmacol Bull 2021 Mar 16;51(2):8-19. [Medline: 34092819]
17. Lutz J, Offidani E, Taraboanta L, Lakhan SE, Campellone TR. Appropriate controls for digital therapeutic clinical trials: a narrative review of control conditions in clinical trials of digital therapeutics (DTx) deploying psychosocial, cognitive, or behavioral content. Front Digit Health 2022;4:823977. [doi: 10.3389/fdgth.2022.823977] [Medline: 36060538]
18. Kroenke K, Spitzer RL, Williams JBW. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613. [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
19. Skevington SM, Lotfy M, O'Connell KA, WHOQOL Group. The World Health Organization's WHOQOL-BREF quality of life assessment: psychometric properties and results of the international field trial. A report from the WHOQOL group. Qual Life Res 2004 Mar;13(2):299-310. [doi: 10.1023/B:QURE.0000018486.91360.00] [Medline: 15085902]
20. Brooke J. SUS: a "quick and dirty" usability scale. In: Usability Evaluation in Industry: CRC Press; 1996.
21. Delgadillo J, Payne S, Gilbody S, et al. How reliable is depression screening in alcohol and drug users? A validation of brief and ultra-brief questionnaires. J Affect Disord 2011 Nov;134(1-3):266-271. [doi: 10.1016/j.jad.2011.06.017] [Medline: 21723619]
22. McGovern MP, Dunn J, Bonnell LN, et al. The association between depression and substance use among primary care patients with comorbid medical and behavioral health conditions. J Prim Care Community Health 2023;14:21501319231200302. [doi: 10.1177/21501319231200302] [Medline: 37728047]
23. Hunt GE, Malhi GS, Lai HMX, Cleary M. Prevalence of comorbid substance use in major depressive disorder in community and clinical settings, 1990-2019: systematic review and meta-analysis. J Affect Disord 2020 Apr 1;266:288-304. [doi: 10.1016/j.jad.2020.01.141] [Medline: 32056890]

XSL•FO
RenderX

24. Arlinghaus KR, Johnston CA. Advocating for behavior change with education. Am J Lifestyle Med 2018;12(2):113-116. [doi: 10.1177/1559827617745479] [Medline: 30283247]
25. Bos E, Preller KH, Kaur G, Malhotra P, Kharawala S, Motti D. Challenges with the use of digital sham: systematic review and recommendations. J Med Internet Res 2023 Oct 24;25:e44764. [doi: 10.2196/44764] [Medline: 37874638]

Abbreviations

FDA: US Food and Drug Administration
MOUD: medication for opioid use disorder
NIDA: National Institute on Drug Abuse
OUD: opioid use disorder
PHQ-9: Patient Health Questionnaire-9
WHOQOL-BREF : World Health Organization Quality of Life-BREF

Edited by A Mavragani; submitted 09.01.25; peer-reviewed by R MacLean, R Bhandari, VR Votaw; revised version received 30.05.25; accepted 02.06.25; published 08.08.25.

Please cite as:
Gallegos KS, Potter JS, King VL, Siegel G, Siegel LH, Marino EN
Development of a Sham Smartphone App for Opioid Use Disorder: Acceptability and Suitability Study
JMIR Form Res 2025;9:e71105
URL: https://formative.jmir.org/2025/1/e71105
doi:10.2196/71105

© Kierstyn S Gallegos, Jennifer S Potter, Van L King, Gregg Siegel, Leslie H Siegel, Elise N Marino. Originally published in JMIR Formative Research (https://formative.jmir.org), 8.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Feasibility and Acceptability of Remotely Accessed Compensatory Cognitive Training for Japanese People With Schizophrenia: Pilot Study

Ritsuko Aijo[1,2], MHS; Mie Matsui[3], PhD

[1]Department of Neuroscience, Graduate School of Medical Sciences, Kanazawa University, Kanazawa, Japan

[2]Department of Nursing, Faculty of Health Sciences, Komatsu University, He 14-1 Mukaimotoorimachi Suehiro Campus, Komatsu, Ishikawa 923-0961, Komatsu, Japan

[3]Laboratory of Clinical Cognitive Neuroscience, Institute of Liberal Arts and Science, Kanazawa University, Kanazawa, Japan

Corresponding Author:
Ritsuko Aijo, MHS
Department of Neuroscience, Graduate School of Medical Sciences, Kanazawa University, Kanazawa, Japan

Abstract

Background: Compensatory cognitive training (CCT) is an evidence-based treatment for improving cognitive function in patients with schizophrenia. However, the need for patients to commute to treatment sites hinders its widespread use. Using a remote device to conduct CCT could improve its accessibility, making it easier for participants to adjust their schedules and reducing their burden.

Objective: The objective of this study was to (1) investigate the creation and participant acceptability of CCT using a remote compensatory cognitive training (r-CCT) device, (2) determine the feasibility of implementing the developed intervention, and (3) collect preliminary data for future studies of the effectiveness of r-CCT in Japan.

Methods: To reduce participant movement during training, CCT was conducted remotely in real time, using borrowed iPads. The training was conducted in a group format through video conferencing once a week for 2 h, for a total of 12 sessions. In total, 4 patients with schizophrenia who underwent r-CCT were recruited to determine participation or dropout rates across 12 training sessions. In addition, their diagnostic assessment (the Scale of Positive Symptoms and the Scale of Negative Symptoms), cognitive function (eg, the Japanese version of the Trail Making Test Part A [TMT-A] and Trail Making Test Part B [TMT-B], digit span, and digit symbol), social functioning (Social Functioning Scale Japanese version [SFS-J]), and quality of life (Japanese Schizophrenia Quality of Life Scale [JSQLS]) were assessed before, immediately after, and 3 months after implementation.

Results: The average participation rate of the 3 participants (a male in his 30s was excluded) was high at 92%. Immediately after the r-CCT, positive trends were observed in cognitive function—excluding prospective memory. For example, the TMT-A scores improved for all 3 participants: Participant A (from 58 s to 56 s), Participant B (from 52 s to 49 s), and Participant C (from 65 s to 49 s). The Japanese Verbal Learning Test (JVLT) immediate scores also increased: Participant A (from 16 to 19), Participant B (from 13 to 14), and Participant C (from 14 to 21). Functional outcomes, assessed using the SFS-J, showed limited improvement immediately postintervention but tended to return to or fall below preintervention levels at the 3-month follow-up. Quality of life (QOL) scores, measured using the JSQLS, remained relatively stable or improved immediately following the r-CCT and at the 3-month follow-up.

Conclusions: Despite this study's small number of participants and lack of randomization, it suggests that the accessibility and implementation potential of r-CCT may be high. The ability to participate in training from any location could be expected to increase participation rates or reduce dropout rates. In the future, the authors will develop the implementation method further and increase the sample size to demonstrate the training's effectiveness.

(JMIR Form Res 2025;9:e70916) doi:10.2196/70916

KEYWORDS

schizophrenia; remote; CCT; compensatory cognitive training; case study

Introduction

Schizophrenia is a chronic mental disorder that often negatively affects the daily life of individuals living with the condition, including their community and social life, due to positive and negative symptoms as well as cognitive dysfunction [1]. The main treatment for schizophrenia is pharmacotherapy, with antipsychotics primarily improving positive symptoms but affecting negative symptoms and cognitive dysfunction less [2]. Because negative symptoms and cognitive dysfunction

XSL•FO
RenderX

contribute to a lower quality of life (QOL) and lifelong disability in patients with schizophrenia [3,4], emphasis is placed on improving their cognitive function [5].

Cognitive remediation therapy (CRT) is an evidence-based therapy used to improve cognition in patients with schizophrenia. The therapeutic goal is to positively impact their ability to function and live in the community through improved cognitive function [6,7]. At the 2010 International Schizophrenia Research Conference in Florence, Italy, CRT for schizophrenia was defined as "a behavioral training-based intervention aimed at improving cognitive processes (attention, memory, executive functioning, social cognition, or metacognition) with persistence and generalization [7]." The most recent meta-analysis of cognitive remediation found small-to-moderate effects in cognitive tests as well as in psychosocial function and psychiatric symptom severity [7,8]. Current treatment guidelines suggest long-term treatment with antipsychotic medication in conjunction with psychological interventions for people with psychosis [1].

Patients with schizophrenia may be hesitant to approach traditional mental health treatment settings due to stigma, which may prevent them from engaging in help-seeking behaviors [9]. Typically, clinicians run CRT sessions at least twice a week to ensure that participants have adequate practice with the cognitive exercises, which are usually provided over the internet [6]. However, multiple in-person sessions each week can pose transportation and scheduling burdens on participants, which could limit their ability to participate [10]. A study reported that reducing the need to travel to once a week may encourage more people to participate in such training [6]. In sum, while CRT exhibits efficacy, it imposes burdens on participants. These factors lead to lower CRT participation rates and higher dropout rates, which in turn significantly affect CRT's penetration rate [11]. Therefore, it is first necessary to develop training that patients with schizophrenia can receive without the need to move to the greatest extent possible, as well as to establish the acceptability of the training for participants and the feasibility of implementing it.

Accordingly, this study aimed to evaluate the acceptability of CRT training by examining participation and dropout rates. Furthermore, it aimed to examine whether compensatory cognitive training (CCT), a type of CRT, improves patients' cognitive function, daily life functioning, and QOL as a result of its implementation at a remote site using quantitative methods. Furthermore, the study aimed to examine the feasibility of remote compensatory cognitive training (r-CCT) implementation in Japan. However, if patients with schizophrenia lack experience with computers, they could experience frustration in learning if a clinician is not present for support and guidance [10]. Thus, to support participants' acceptance and continuation of the r-CCT, only the therapist participated remotely while the participants attended sessions together at a location where an assistant was available to provide technical and other support. The meeting locations for the r-CCT sessions were places that the participants commuted to on a daily basis and did not involve longer travel. To collect preliminary data for a future study of the effectiveness of r-CCT in Japan, r-CCT was administered to 4 patients with schizophrenia in this study. The remote approach to CCT piloted in this study is, to our knowledge, the first to be evaluated in Japan. Furthermore, a quantitative approach was used to examine the training's acceptability to the patients as well as the feasibility of its implementation.

Methods

Recruitment

The 4 participants were recruited from different social welfare facilities in Japan, specifically from 4 facilities in proximity to each other. To avoid selection bias in participant recruitment, facility administrators—rather than researchers—approached the users themselves. While the patients belonged to the same social welfare facility, 3 of them worked at the same location and 1 worked somewhere else. The r-CCT was conducted at the location where the 3 participants normally worked; therefore, only 1 participant was required to move to participate in the study. The inclusion criteria were as follows: (1) aged 18 years or older, a diagnosis of schizophrenia according to the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* or the *International Classification of Diseases 10th Revision*, fluency in Japanese, and stable enough to participate in group treatment. The exclusion criteria were as follows: head trauma, cerebrovascular and neurodegenerative disorders, diagnosis of alcohol and drug dependence, and a premorbid estimated IQ of 70 on the Japanese version of the Adult Reading Test [12].

Data were collected between February and September 2022. The 4 participants (3 men and 1 woman) had a mean age of 55.8 years, mean age of onset of illness of 30 years, mean duration of illness of 27 years, mean education of 12.8 years, estimated IQ before illness onset of 101.3, and mean daily dose of antipsychotic medication (chlorpromazine equivalent) of 1163.8 mg. All of them worked at a job of support B-type; however, their job descriptions and working hours differed. Of the 4 participants, 1 used a cellphone and 1 used a smartphone, but none had ever used an iPad or a computer. Table 1 presents detailed information on each participant. Initially, we planned to recruit a larger number of participants; however, due to a low number of applicants at the target facilities and practical constraints imposed by the COVID-19 pandemic, only 4 participants ultimately participated. Among them, Participant D exhibited pronounced positive symptoms of schizophrenia at the start of the intervention and, at their own request, was excluded from most of the program, completing only 2 training sessions. Consequently, they were considered to have withdrawn from the study and were excluded from the analyses of cognitive function, functional outcomes, and QOL.

XSL•FO
RenderX

Table . Sample characteristics before compensatory cognitive training using a remote compensatory cognitive training (r-CCT) device intervention for individuals diagnosed with schizophrenia who participated in this pilot study. Participants were recruited from social welfare facilities in Japan between February and September 2022.

Participant		Age (years)	Age at onset (years)	Illness duration (years)	Education (years)	Occupation	Use of internet-based equipment	Estimate of premorbid IQ	Antipsychotic dose (mg/day)[a]
A	Male	68	28	40	12	During the day, they go to a job support B-type, in which they work for approximately 2 h a day, doing coffee shop work and cooking.	No previous experience with cell phones, smartphones, iPads, and personal computers.	104	1600
B	Male	54	40	15	12	During the day, they attend a job support type B, in which they mainly work for 1 - 2 h a day bagging electronic components.	No previous experience with cell phones, smartphones, iPads, and personal computers.	104	600
C	Female	64	28	36	15	During the day, they attend a job support B-type and do work commissioned by the facility for approximately 2 h a day. Sometimes, they exhibit their crocheted works.	Have used and still use cell phones; no previous experience with smart phones, iPads, and computers.	96	1200
D	Male	37	22	15	12	During the day, they go to a job support B-type, in which they inspect goods, process food, and pack bags for approximately 3 h a day. However, even within those 3 h, frequent breaks are required because of the illness.	Have used and still use cell phones and smart phones; no previous experience with iPad and personal computers	101	1255

[a]Antipsychotic medication dose is chlorpromazine equivalent (mg/day).

XSL•FO
RenderX

Procedures

The r-CCT's acceptability was assessed by examining the 4 patients' participation and dropout rates across the 12 training sessions. In addition, the feasibility of implementing r-CCT in Japan was evaluated by assessing the participants' cognitive function, social functioning, and QOL at baseline, 3 months (ie, postintervention), and 6 months. The goal here was to determine whether they exhibited trends similar to those of patients who receive face-to-face CCT.

CCT Intervention and Remote Methods

CCT is a type of cognitive remediation training developed that targets the following 4 cognitive domains: prospective memory, attention, learning and memory, and executive functioning [13]. The r-CCT used in this study is not a newly developed method of intervention but is based on an existing manualized CCT training developed by Twamley et al [13]. The Japanese version of CCT had already been translated and culturally adapted by researchers independent of the research team. In this study, the basic treatment structure and session content were maintained; however, the format was modified to accommodate remote delivery. The changes include sharing Microsoft PowerPoint files through the internet-based meeting platform (Microsoft Teams), writing session content on the slides in real time, and using audio and video equipment. The compensatory strategies trained are both internal (eg, learning and recollection of information using acronyms or visual imagery) and external (eg, learning and recollection of information by writing it down somewhere it will be seen). These strategies lead to a wide range of improvements, ranging from those in cognitive function to social functioning to QOL [13]. The CCT intervention is a face-to-face, manual training program that comprises 4 to 6 participants and several therapists, who meet once a week for 2 hours for a total of 12 sessions [13]. Each session uses a workbook and an interactive format that allows for interactivity between participants and therapists as well as among participants [13]. In this study, a workbook originally translated into Japanese was used, and the approach of manual group therapy for 2 hours per week for 12 weeks was followed [14]. During the training, participants learn many strategies, practice their use in CCT, and habitually use them in the real world—even after the CCT is over [13]. A simple homework assignment (eg, carry a calendar with you every day) is given at the end of each session to encourage participants to continue using the strategies [13].

This paper is the first report concerning CCT performed remotely. To emphasize technical assistance, the 4 participants were asked to undergo r-CCT at the same location, assisted by 1 or 2 assistants. The remote approach involved placing 1 Apple iPad Pro in a room with the participants and 1 assistant so that the therapist could communicate with them internet-based from another room. A Microsoft Teams was used for internet-based interaction. For video, iPad images were projected onto a TV screen so that all participants could view them under the same conditions. For audio, 2 speakerphones (eMeet Office Core) were used to easily pick up the voices of multiple participants. At the first training session, workbooks and remote equipment were given to the participants and assistants (see Figure 1).

Figure 1. Training scene of remote compensatory cognitive training (r-CCT) for people with schizophrenia conducted at a social welfare facility in Japan. The intervention was conducted from February to September 2022.

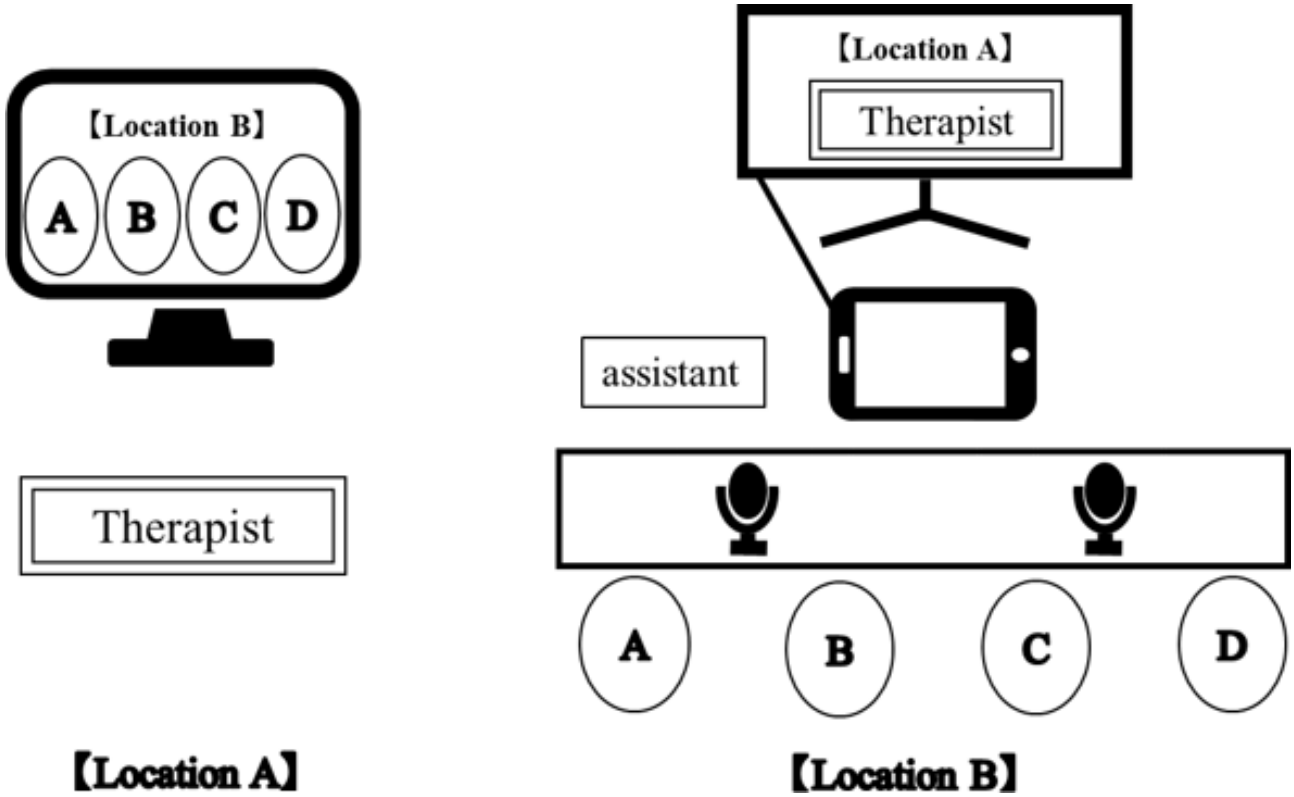

The assistant was asked to connect 1 iPad and 2 eMeet Office Core devices, project the iPad video, and set up participation in a Microsoft Teams meeting. In this study design, participants were not expected to operate iPads; therefore, all remote device

XSL•FO
RenderX

operations were performed by assistants. However, rather than increasing the number of staff relative to that required for face-to-face training, we devised a method to conduct remote training using the same number of staff. The therapist and assistant tested the connection once at a time and date that differed from the r-CCT. The assistants were instructed to speak only to participants who experienced problems during the r-CCT, with the exception of the provision of technical support. To examine the remote method of delivery, a preliminary survey was conducted with 4 healthy people in advance. Through these preliminary interviews, the implementation of the existing CCT was confirmed to be possible through adjusting the volume and screen of the remote devices; then, the main study could begin.

Quantitative Data Collection

The r-CCT's acceptability was assessed by recording how many times participants were able or unable to participate during the 12 training sessions as a percentage. The feasibility of its implementation was assessed by performing the following inspections on participants before, immediately after, and 3 months after the implementation of the r-CCT. The inspections were performed by personnel trained in standardized procedures.

With regard to symptom assessment, the severity of positive symptoms was measured using the Scale of Positive Symptoms (SAPS [15]), while the severity of negative symptoms was measured using the Scale of Negative Symptoms (SANS [16]). Both the SAPS and the SANS are rated on a 6-point scale (0 - 5 for each item), with higher scores indicating a greater severity of symptoms. Both scales score the severity of psychiatric symptoms based on observations and interviews with patients with schizophrenia.

In addition, a combination of existing tests was used that is thought to correspond to the 4 cognitive domains covered by CCT (ie, prospective memory, attention, learning and memory, and executive functioning). Specifically, the belongings subtests of the Rivermead Behavioral Memory Test (RBMT; Japanese version), the Trail Making Test Part A (TMT-A; Japanese version), the digit symbol and digit span subtests of the Wechsler Adult Intelligence Scale–Fourth Edition (WAIS-IV), the Japanese Verbal Learning Test (JVLT), the story recall subtest of the RBMT, and the Trail Making Test Part B (TMT-B; Japanese version). The tests were administered in a specific order each time.

First, prospective memory was assessed with the belongings subtests of the RBMT [17,18]. The belongings task involved the examiner hiding the participant's belongings (eg, their keys and watch) under a desk or another location that was not visible to the participant at the beginning of the cognitive test; then, the participant was asked to recall the hidden belongings at the end of all cognitive tests. Thus, their memory of the need to request the belongings as well as of the hidden location was assessed. The participants received 2 points each for recalling the location and belongings without hints, and 1 point each for recalling the location and belongings with hints, for a maximum of 4 points.

Second, attention was assessed using the TMT-A [19], the digit symbol subtest of the WAIS-IV [20,21], and the digit span subtest of the WAIS-IV [20,21]. First, in the TMT-A, the participant must tie 25 numbers in order as quickly as possible without making mistakes, and the time taken to complete the task is measured. The shorter the completion time, the higher their cognitive ability. Second, in the digit symbol subtest, participants have 2 minutes to write down symbols that correspond to digits as quickly and as accurately as possible. The study used a maximum score of 19 points, with the number of correctly described signs adjusted for age. Finally, in the digit span subtest, participants must reproduce digits read aloud in the correct order. The number of digits is gradually increased and assessed by the maximum number of digits they can correctly reproduce. The maximum number of digits in this study was 9.

Third, learning and memory were assessed with the JVLT [22] and the story recall subtest of the RBMT [18]. First, the JVLT consisted of a list of 16 words divided into 4 categories. The words were presented orally, and participants were asked to recall the words for 3 trials each time the 16 words were presented. In this study, the total number of correct responses to recall over the 3 trials (ie, highest score=48) was used to evaluate the participants. Second, the story recall subtest is a task scored up to 25 points in which participants remember and replay a short story that was told orally.

Finally, executive function was assessed using the TMT-B [19]. The task involves connecting 13 numbers and 12 hiragana (1 of the 3 writing systems in Japanese), alternately in sequence (eg, 1a - 2 - i - 3u.), as quickly as possible and without making mistakes. The time taken to complete the task was measured, and the shorter the completion time, the higher the participant's cognitive ability.

Furthermore, functional outcomes were assessed using the Japanese version of the Scale of Social Functioning (SFS-J). The SFS-J is a self-administered rating scale with 7 subscales [23]. It consists of a total of 75 items that relate to basic skills and social behaviors required for daily living, and they are assessed using subscale scores and total scores. For the question items in the subscales of withdrawal, interpersonal relationships, social participation, recreation, independence and competence, independence and performance, and employment, the participants indicated how often they perform the stated behaviors (response options: "never," "rarely," "sometimes," and "frequently") or how well they perform them (response options: "do not know how to do them," "cannot do them," "can do them with help," and "can do them well enough"). Each item was scored on a 4-point scale from 0 to 3, with higher scores indicating better functioning. The total score was used in this study. QOL was assessed using the Japanese version of the Schizophrenia Quality of Life Scale (JSQLS), which is a self-administered scale that measures the QOL of patients with schizophrenia in three domains (psychosocial relationships, motivation and vitality, and symptoms and side effects). It consists of 30 items. Responses are rated on a 5-point scale from 0 to 4, with lower scores indicating a higher QOL [24].

Statistical Analyses

This was a preliminary study using a small sample size to examine the feasibility and acceptability of remote compensatory

XSL•FO
RenderX

cognitive function improvement therapy. Due to the very small sample size (n=3), statistical tests (such as calculations of significance or effect size) were not performed. Instead, we provided a descriptive report on r-CCT participation rates and changes in each participant's measurements, ensuring confidentiality. In addition, we investigated the feasibility and potential effects of the intervention, showing trends at three points in time: before r-CCT (T1), immediately after r-CCT (T2), and 3 months post r-CCT (T3).

Ethical Considerations

This study was approved by the Medical Ethics Committee of Kanazawa University (approval 2020‐327 [089]) and was conducted in accordance with the principles of the Declaration of Helsinki. All participants were informed of the study's purpose and procedures in writing and orally, and written informed consent was obtained from each participant before participation. All data collected were anonymized, and personally identifiable information was removed before analysis. This study and its supplementary materials do not contain any images or information that could identify individuals. Participants did not receive any monetary compensation; however, the intervention sessions were provided for free, and technical support was offered to ensure the participants could engage in the training sessions with confidence and ease. This study is not a secondary analysis of existing data but a new study in which participants were recruited and consent was obtained. Informed consent waivers were not applied.

Results

Acceptability of r-CCT to Participants

Of the 12 r-CCT sessions, Participant A attended 75% (9/12) of them, Participant B and Participant C attended 100% (12/12), and Participant D attended 17% (2/12). The average participation rate of the 4 participants was 73%, while that of the 3 participants excluding Participant D was 92%. Participant D had the lowest participation rate because he, unlike the other participants, had to drive for 15 minutes to attend the training sessions. Furthermore, they exhibited strong positive symptoms from the first session and were only able to participate in the fourth and fifth sessions from start to finish. A staff member who knew Participant D stated the following: "Mr. D sometimes said it was hard for them to leave the house when their auditory hallucinations got worse." Participant A was absent from 3 of the 12 r-CCT sessions but never forgot their training; one of their absences was due to ill health while 2 were deliberate. The reason for their deliberate absences was an unpleasant experience during the r-CCT, which was resolved through discussion with the therapist. It was caused by emotional discomfort rather than technical difficulties or physical risks. Specifically, the therapist's comments (eg, praising other participants) may have induced feelings of jealousy and dissatisfaction, as they did not align with the participants' expectations. This concern was raised by participants during discussions with the research team, and through these discussions, the participants were able to gain a clearer understanding of the situation, enabling them to continue with the intervention without dropping out. Participant B was able to participate in all r-CCT sessions; however, they exhibited little voluntary participation and were often asked to participate by the staff as they forgot about the training. Participant C was able to voluntarily attend all r-CCT sessions. Regarding staff support, all of the participants indicated that they did not feel inconvenienced because the staff essentially operated the internet-based equipment. Regarding their willingness to continue the training, all participants stated that it did not matter whether staff were present or not. Furthermore, Participant C, who was absent without tardiness, said the following: "At first, I was worried that I wouldn't understand the training, but once I did it, I had a goal and then I joined in with a challenging spirit." Few negative comments were made about the volume and pictures on the screen. Furthermore, participants stated that they "didn't mind that it was online," found the content "easy to understand even online," and described the experience as "online training that was interesting and fun, not too hard."

Feasibility of Implementing R-CCT

Table 2 presents the results for cognitive performance, functional outcomes, and QOL. Participant D was excluded from the analysis because they fully participated in only 2 training sessions. Therefore, this study presents the results based on data from the remaining 3 participants, excluding Participant D. Furthermore, the study design was not suitable for group comparisons or statistical significance testing. Therefore, the results are presented with a focus on changes in individual data, and descriptive statistics (such as means and SDs) are not used. This approach aims to accurately reflect the specific responses and clinical changes observed in each participant. The following sections detail these results.

XSL•FO
RenderX

Table . Changes in scores for each indicator among individuals diagnosed with schizophrenia who participated in this pilot study. Data collected from social welfare facilities in Japan between February and September 2022.

	Participant A					Participant B					Participant C					Participant D				
	T1[a]	T2[b]	T3[c]	T2-T1	T3-T1	T1	T2	T3	T2-T1	T3-T1	T1	T2	T3	T2-T1	T3-T1	T1	T2	T3	T2-T1	T3-T1
Symptom assessment																				
SAPS[d] positive symptoms	14	7	2	−7[f]	−12[f]	31	30	19	−1[f]	−12[f]	6	1	2	−5[f]	−4[f]	36	48	36	+12[g]	0[h]
SANS[e] negative symptoms	35	40	18	+5[g]	−17[f]	18	14	37	−4[f]	+19[g]	48	39	41	−9[f]	−7[f]	24	16	13	−8[f]	−11[f]
Prospective memory																				
Belongings (total 4 points)	4	2	4	−2[g]	0[h]	2	1	2	−1[g]	0[h]	4	2	4	−2[g]	0[h]	4	4	4	0[h]	0[h]
Attention																				
TMT-A[j,e] (s)	58	56	60	−2[f]	+2[g]	52	49	43	−4[f]	−9[f]	65	49	38	−16[f]	−27[f]	49	52	28	+3[g]	−21[f]
Digit symbol (total 19 points)	10	11	11	+1[f]	+1[f]	6	6	8	0[h]	+2[f]	9	8	10	−1[g]	1[f]	6	4	6	−2[g]	0[h]
Digit span (total 9 digits)	5	5	6	0[h]	+1[f]	7	7	6	0[h]	−1	6	5	6	−1[g]	0[h]	6	8	7	+2[f]	+1[f]
Learning and memory																				
JVLT[k] immediate (total 48 pieces)	16	19	15	+3[f]	−1	13	14	7	+1[f]	−5[g]	14	21	23	+7[f]	+2[f]	29	24	20	−5[g]	−9[g]

XSL•FO
RenderX

	Participant A					Participant B					Participant C					Participant D				
	T1[a]	T2[b]	T3[c]	T2-T1	T3-T1	T1	T2	T3	T2-T1	T3-T1	T1	T2	T3	T2-T1	T3-T1	T1	T2	T3	T2-T1	T3-T1
Story recall (total 25 pieces)	7.0	7.5	9.5	+0.5[f]	+2.5[f]	3.0	7.5	1.0	+4.5[f]	−2.0[g]	7.5	6.5	2.5	−1.0[g]	−5.0[g]	4.0	7.0	1.0	+3.0[f]	−3.0[g]
Executive function																				
TMT-B[l,e] (s)	127	106	116	−21[f]	−11[f]	91	103	76	+12[g]	−15[f]	192	159	116	−33[f]	−43[f]	71	54	37	−17[f]	−34[f]
Functional outcome																				
SFS-J[m] total score	157	147	130	−10[g]	−27[g]	112	122	113	+10[f]	+1[f]	109	114	105	+5[f]	−4[g]	106	114	121	+8[f]	+15[f]
QOL[n]																				
JSQLS[o]	57	61	56	+4[g]	−1[f]	60	45	53	−15[f]	−7[f]	55	37	43	+18[f]	+12[f]	41	41	36	0[h]	+5[f]

[a]T1 is before remote compensatory cognitive training (r-CCT).

[b]T2 is immediately after r-CCT.

[c]T3 is 3 months after r-CCT.

[d]SAPS: Scale of Positive Symptoms.

[e] SAPS, SANS, TMT-A, TMT-B, and SQOL indicate that the lower the score, the better the condition.

[f]Red indicates an increase in score.

[g]Blue indicates a decline in score.

[h]Green indicates stability of scores.

[i]SANS: Scale of Negative Symptoms.

[j]TMT-A: Trail Making Test Part A.

[k]JVLT: Japanese Verbal Learning Test.

[l]TMT-B: Trail Making Test Part B.

[m]SFS-J: Social Functioning Scale Japanese version.

[n]QOL: quality of life.

[o]JSQLS: Japanese Schizophrenia Quality of Life Scale.

First, the specific changes in cognitive function for Participant A, Participant B, and Participant C, as shown in Table 2, are as follows. The scores from 3 time points—before the intervention (T1), after (T2), and 3 months postintervention (T3)—are compared. Prospective memory was measured using the "belongings subtests," with Participant A scoring 4, 2, and 4; Participant B scoring 2, 1, and 2; and Participant C scoring 4, 2, and 4. All 3 participants exhibited a decline in scores immediately after the intervention; however, their scores returned to preintervention levels at the 3-month follow-up. Attention function was measured using the TMT-A, digit symbol, and digit span subtests. In the TMT-A, shorter completion times indicated higher cognitive function. Participant A's recorded times of 58 s, 56 s, and 60 s; Participant B recorded 52 s, 49 s, and 43 s; and Participant C recorded 65 s, 49 s, and 38 s at T1, T2, and T3, respectively. Immediately after the intervention, the performance of all 3 participants improved; however, by the 3-month follow-up, Participant A's scores had declined, whereas Participants B and C maintained their scores. The digit symbol test, which uses a 19-point scale that takes age into account, yielded the following scores. Participant A scored 10, 11, and 11; Participant B scored 6, 6, and 8; and Participant C scored 9, 8, and 10. Immediately after the intervention, Participants A and B showed improvement; however, Participant C's performance declined. After 3 months, all 3 participants maintained their performance levels. The digit spans were 5, 5, and 6 for Participant A; 7, 7, and 6 for Participant B; and 6, 5, and 6 for Participant C. Subsequently, the performance of Participants A and B improved, while that of Participant C declined. After 3 months, the scores of Participants A and C remained stable, whereas those of Participant B declined. Learning and memory were measured using the JVLT immediate test and the story recall test. The JVLT immediate scores were 16, 19, and 15 for Participant A;

XSL•FO
RenderX

13, 14, and 7 for Participant B; and 14, 21, and 23 for Participant C, with all 3 showing improvement in their immediate task scores. After 3 months, Participant A's and Participant B's scores declined, while Participant C maintained their performance levels. The story recall scores were 7.0, 7.5, and 9.5 for Participant A; 3.0, 7.5, and 1.0 for Participant B; and 7.5, 6.5, and 2.5 for Participant C. In the immediate task, Participant A's performance improved, whereas Participant B's and Participant C's performance levels declined. After 3 months, Participant A maintained their earlier performance levels, whereas those of Participant B and Participant C declined. Executive function was measured using the TMT-B. In the TMT-B, shorter task completion times indicate higher cognitive function. Participant A scored 127 s, 106 s, and 116 s; Participant B scored 91 s, 103 s, and 76 s; and Participant C scored 192 s, 159 s, and 116 s. Immediately after the intervention, Participants A and C showed improvement in task performance, whereas Participant B's performance levels declined. After 3 months, all 3 maintained their task performance levels.

To summarize the trends, with the exception of prospective memory, the attention, learning and memory, and executive functions targeted by r-CCT tended to improve immediately after r-CCT. After 3 months, the trends tended to remain stable, except for learning and memory.

Next, Table 2 shows the changes in functional outcome and QOL for Participants A, B, and C before (T1), after (T2), and 3 months after (T3) the r-CCT intervention. Functional outcome measured using the SFS-J for Participant A was 157, 147, and 130; for Participant B, it was 112, 122, and 113; and for Participant C, it was 109, 114, and 105. Immediately after the intervention, Participant A showed a decrease, whereas Participants B and C showed improvements. Furthermore, 3 months after the intervention, Participants A and B showed a decline, whereas Participant C returned to their pretraining level. QOL was measured using the JSQOL, a scale for measuring QOL in schizophrenia, with lower scores indicating higher QOL. Participant A scored 57, 61, and 56; Participant B scored 60, 45, and 53; and Participant C scored 55, 37, and 43. Immediately after the intervention, Participant A's task performance levels decreased, while those of Participant B and Participant C showed improvement. In addition, 3 months after the intervention, all 3 maintained their QOL scores. The participants' scores on the SFS-J, which measures functional outcomes, tended to remain stable immediately after the r-CCT. However, 3 months later, the effects had returned to pretraining levels or decreased slightly. Their scores on the JSQLS, which measures QOL, were generally maintained or improved immediately after the r-CCT, and these effects continued.

Discussion

Principal Findings

CRT is an evidence-based practice that is increasingly used in clinical practice to treat the widespread and significant cognitive deficits that contribute to poor functional outcomes in patients with schizophrenia [10]. Recommendations for practicing CRT include ensuring that patients receive sufficient exposure to cognitive exercises to promote cognitive change [25], usually at least twice a week [10]. Nevertheless, the challenge of attending in-person sessions several times a week may limit the treatment's scalability and patients' access to care. Therefore, it is crucial to consider other ways to implement CRT, which include providing opportunities for participants to access the training remotely [10].

This study was a preliminary exploratory study to examine the acceptability and feasibility of CCT using remote devices in Japanese patients with schizophrenia. As this study was the first attempt to conduct CCT entirely remotely in Japan, it was intentionally conducted with a small number of participants to confirm implementation procedures, assess participants' responses, and identify potential challenges during the introduction phase. Due to the small sample size (n=4, with 3 completing the study) and the fact that participants were recruited from a single facility, statistical analyses were not conducted and conclusions regarding efficacy or the generalizability of results could not be drawn. However, this study provided foundational insights to inform future large-scale studies. The results obtained using this method suggest that participants with fewer movement requirements may demonstrate higher participation rates and lower dropout rates.

The 3 participants (excluding Participant D) did not have to travel to attend the r-CCT sessions, which may have facilitated their participation and resulted in fewer absences. To reduce movement, the training was delivered remotely; however, the assistants were asked to be present in the same location as the participants who were determined to require technical support and assistance with the training content. However, the assistants mostly assisted with the operation of the internet-based equipment and provided little assistance with the training content. The participants could ask the internet-based therapist any questions they had without assistance. Although the telecommunication conditions were poor and communication was sometimes choppy, no negative comments were provided by the participants. In this study, remote training was implemented as an intervention method; however, it was not a completely independent format in which participants conducted training from their homes or other locations. Instead, they visited familiar facilities and received technical support from local staff while therapists conducted the intervention remotely. This partially remote format was selected to accommodate individuals who may be unfamiliar with operating digital devices. Therefore, the results of this study do not directly indicate the feasibility of a fully remote intervention. Conclusions regarding the acceptability and scalability of r-CCT based on this study are provisional, and caution is warranted when generalizing findings to other environments.

Keef et al [26] reported that increasing the enjoyment of digital training interventions may lead to greater adherence and cognitive benefits. Furthermore, Best et al [27] reported that programs designed to increase motivation and provide opportunities to practice daily living skills improve treatment retention rates. Originally, CCT was more likely to be motivational because the sessions mainly involved group work, included fun activities such as games, and were relevant to the participants' everyday life [28]. This study was designed to

XSL•FO
RenderX

enable all of the participants to receive training in the same location, as they may have been resistant to the use of remote equipment. However, if the training itself is designed to be enjoyable and motivational, training with remote devices may not be as burdensome for those who are not used to using them. This study's findings suggest that even remote methods may be acceptable to patients with schizophrenia if they are easy to operate. Currently, based on this study, we are conducting research on the acceptability and effectiveness of a fully home-based r-CCT to investigate its applicability in a wider range of settings.

In this study, standardized scales or structured interviews were not used to evaluate participants' acceptability of the intervention. Instead, acceptability was primarily evaluated based on participation rates, number of absences, and anecdotal feedback. This represents one of the study's limitations. However, as this was the first study to implement CCT remotely, its primary aim was to verify whether remote CCT could be conducted safely and consistently. Future studies should address this limitation by refining implementation methods, setting a larger sample size, and including structured qualitative interviews to gain a deeper understanding of participants' experiences and perceptions.

Due to differences in measurement scales and participant characteristics, direct comparisons are difficult. However, the effectiveness of r-CCT was evaluated by assessing participants' cognitive function, functional outcomes, and QOL before intervention, after intervention, and 3 months post r-CCT implementation to determine whether the training effects showed a trend similar to those of face-to-face CCT. The 4 cognitive domains targeted by r-CCT (ie, prospective memory, attention, learning and memory, and executive function) exhibited the following specific trends:

Prospective memory was measured using a belongings task; however, all 3 participants showed a decline in performance levels immediately after the intervention. By the 3-month follow-up, their performance had returned to preintervention levels, with no stable scores or improvement observed using r-CCT. Attention function was measured using the TMT-A, digit symbol, and digit span subtests, whereas executive function was measured using the TMT-B. In addition, 2 or 3 participants demonstrated an improvement in performance immediately after training, which was sustained at the 3-month follow-up. Learning and memory were measured using the JVLT and the story recall subtest. Furthermore, 2 to 3 participants demonstrated improved performance immediately after training; however, 2 participants exhibited a decline in performance at the 3-month follow-up. These results suggest that r-CCT may have beneficial effects on attention, learning and memory, and executive function, with prospective memory representing a potential exception. Furthermore, the 3-month follow-up results indicate that performance levels tend to remain stable, with the exception of learning and memory.

Previous studies on face-to-face CCT have found that, immediately after intervention, participants in the intervention group showed significant improvements across 4 cognitive-functional domains—prospective memory, attention, learning and memory, and executive function—relative to those receiving regular treatment [8,13]. This trend was observed in the 3 cognitive functions, excluding prospective memory, showing improvement trends similar to previous studies. However, the lack of improvement in prospective memory may be attributable to inconsistencies between the assessment tasks employed and the intervention content. The r-CCT uses a strategy called "the usual place," which involves designating a specific location within the home—referred to as "home"—where essential items, such as keys and wallets, are consistently placed to prevent them from being forgotten [13]. This "usual place" strategy was particularly emphasized during training and was the only strategy to be included in all homework assignments. By contrast, in the belongings task, the examiner hides the participant's belongings at the start of the cognitive function test and then asks them to recall the hidden items and their locations at the end of the test. A possible explanation for the lack of improvement in cognitive function following remote CCT is that the "usual place" strategy may have become integrated into participants' daily routines, making it challenging for them to recall items hidden in locations outside of their homes during the belongings task. Future research on prospective memory should consider the careful selection of alternative assessment methods and inclusion of larger sample sizes.

Furthermore, previous studies on face-to-face CCT have reported that, at the 3-month follow-up, the intervention group showed significant improvements across 4 cognitive domains (prospective memory, attention, learning and memory, and executive function) compared with those receiving regular treatment [8,13]. This trend was observed in 3 cognitive functions excluding learning and memory, yielding results similar to those reported in previous studies. However, the lack of improvement in learning and memory in this study may be attributed to the nature of the homework assignments. Habit learning in individuals with schizophrenia is generally preserved, and declarative memory acquired through habitual practice is less likely to be forgotten [29]. CCT is designed to facilitate habit formation by assigning homework after each training session, enabling participants to apply strategies in their daily lives. However, since the training was conducted remotely this time, the implementation of homework was only confirmed with participants, and no visual confirmation was made. As a result, the extent to which participants actually implemented the homework remains unclear, which may explain why improvements in language memory were not sustained at the 3-month follow-up. Future studies should address these limitations by incorporating methods more suitable for remote settings. For example, measures such as having participants display the contents of their homework on the screen during homework confirmation or extending the time allocated for homework confirmation could be considered.

In this study on social functioning, SFS-J was used. However, when looking at the specific results, the scores of 2 participants improved immediately after the intervention; however, this improvement was not sustained at the 3-month follow-up. Scores on the JSQLS remained relatively stable in 2 out of 3 participants immediately after the intervention and in all 3

XSL•FO
RenderX

participants at the 3-month follow-up. In previous studies on face-to-face CCT targeting Japanese participants, effect sizes for functional indicators such as functional capacity and subjective QOL were reported to range from 0.47 to 1.56, suggesting that these effects were greater than those observed for cognitive function [14]. As this is a case study, effect sizes were not calculated, making direct comparisons with previous research difficult. However, the observed improvement trend in JSQLS scores immediately after the intervention and the sustained trend at the 3-month follow-up may reflect similar improvement trends observed in previous studies. However, the lack of stable SFS-J scores at the 3-month follow-up differs from the findings of previous studies. These findings may have been influenced by the participants' living environment. Social life functioning is influenced by factors like motivation and social support [29]. Participants A, B, and C worked at the same social welfare facility, and the results may have been influenced by events at the facility and the nature of their work. Future studies may need to recruit participants from a wider population.

Overall, the results indicate that training without moving is relatively less burdensome for participants with schizophrenia as well as feasible to implement. This study has several limitations. First, the participants gathered at the same facility and received technical support during training; however, this format may not substantially reduce the burden on patients compared with traditional face-to-face interventions. Furthermore, it should be noted that devices such as iPads, television screens, and speakerphones may not be easily available or accessible to all participants and facilities. However, if the use of remote devices allows participants to receive training without having to visit the therapist's location, then it should facilitate their participation and make it possible for more patients with schizophrenia to receive CCT. The authors suggest that if training could be developed in the future to allow therapists and participants to participate remotely with ease, then higher participation rates could be achieved. Although this was a small-scale pilot study, it identified specific challenges in implementing cognitive training using remote devices and provided initial insights into the practical feasibility of r-CCT. While several instances of unstable communication occurred during training sessions, no serious issues related to privacy or connectivity were reported, and no complaints were received from participants or assistants. In this study, the internet-based environment was set up smoothly by providing assistants with clear instructions about the entire process. If the operations and settings themselves were simplified slightly, participants may be able to operate the remote devices independently.

Without this study, it would be difficult to proceed with new research projects in this field. Currently, based on the results of this study, we are conducting a large-scale study with all participants who are able to participate from their homes. This new research is positioned as a separate project from the current research and is planned to be reported as an independent study in the future. Therefore, the data being collected currently are not included in this study.

Limitations and Future Challenges

This study had some limitations. First, the small sample and the fact that participants were recruited from the same welfare facility made it difficult to generalize the results, which may have been strongly influenced by the facility that recruited the participants. Second, a target group was not established and a randomized controlled trial was not conducted; therefore, the effects of the intervention were not rigorously proven. However, this study was a preliminary study to explore the feasibility and acceptability of remote compensatory cognitive function improvement therapy. It was difficult to conduct a randomized controlled trial due to resource and practical constraints. While acknowledging that the absence of randomization and a control group may reduce internal validity and introduce selection bias, efforts were made to avoid arbitrary selection by researchers in the recruitment process. Participants were recruited from four social welfare facilities, with the cooperation and assistance of facility managers. In the initial approach, researchers did not directly approach participants but instead requested managers to approach them. By delegating the selection process to facility staff rather than researchers directly determining participants, the influence of researcher bias was minimized. Therefore, although this was not an intervention conducted under strict control, it reflected a natural implementation method tailored to the local situation and provided valuable insights for future implementation and research design. Third, as the subjects were Japanese, their schooling and life backgrounds differ slightly from subjects in previous studies; therefore, the results of this study should be interpreted with caution and replicated in future studies. Fourth, the intervention in this study was designed to be remote; however, it was not entirely home-based. Participants attended familiar welfare facilities and received technical support from assistants during. The assistants primarily provided support with device operation, while therapists conducted the intervention remotely. Although this partially remote intervention was an appropriate option for participants who were unfamiliar with digital devices, its ecological validity was limited due to the discrepancy between the participants' actual living environments and the intervention setting. Such contextual factors are critically important when evaluating the feasibility of the intervention and should be systematically assessed alongside changes in cognitive function in future expansion studies.

This study also has strengths. First, although it was a case study, it found that cognitive function and QOL immediately after the r-CTT training were generally similar to those after face-to-face CCT, except for prospective memory. Second, the training's effects on QOL were found to possibly persist after 3 months.

Conclusion

This study examined the acceptability and feasibility of distance training for patients with schizophrenia participating in r-CCT. The study was conducted remotely to avoid participant movement, although internet-based problems were simultaneously considered. The 3 participants attended more than two-thirds of the 12 training sessions and expressed few negative comments about the remote training. In addition, cognitive functions targeted by r-CCT, except for prospective

XSL•FO
RenderX

memory, could usually be maintained or improved immediately after training, even with r-CCT. However, whether this effect persists three months later should be thoroughly researched in the future. Meanwhile, it was suggested that patients could benefit in terms of QOL from the training, even when the remote method is used. These findings demonstrate that the remote method, depending on how it is used, may lead to improved training participation and participant acquisition. These are crucial findings regarding the acceptability and feasibility of r-CCT for patients with schizophrenia, and they indicate that the remote approach is likely to be suitable in Japan.

Acknowledgments

The authors would like to thank all patients with schizophrenia who cooperated with them in conducting this study. They also wish to thank the stakeholders who referred participants. Finally, the authors acknowledge the role of the Enago website for the English language review. This work was supported by the Japan Society for the Promotion of Science KAKENHI (Grant JP 23K09961) and a Research Development and Improvement Fund grant through Komatsu University.

Data Availability

Due to privacy concerns and ethical constraints, the data obtained and analyzed in this study are not publicly available. The datasets generated or analyzed during this study are available from the corresponding author upon reasonable request and with the approval of the Ethical Review Board of our institution.

Authors' Contributions

RA contributed to conceptualization, data curation, funding acquisition, investigation, methodology, project administration, supervision, validation, visualization, writing the original draft, and writing - review and editing. MM contributed to investigation and supervision.

Conflicts of Interest

None declared.

References

1. Krogmann A, Peters L, von Hardenberg L, Bödeker K, Nöhles VB, Correll CU. Keeping up with the therapeutic advances in schizophrenia: a review of novel and emerging pharmacological entities. CNS Spectr 2019 Aug;24(S1):38-69. [doi: 10.1017/S109285291900124X] [Medline: 31482779]
2. Stępnicki P, Kondej M, Kaczor AA. Current concepts and treatments of schizophrenia. Molecules 2018 Aug 20;23(8):2087. [doi: 10.3390/molecules23082087] [Medline: 30127324]
3. Correll CU, Schooler NR. Negative symptoms in schizophrenia: a review and clinical guide for recognition, assessment, and treatment. Neuropsychiatr Dis Treat 2020;16:519-534. [doi: 10.2147/NDT.S225643] [Medline: 32110026]
4. Millan MJ, Fone K, Steckler T, Horan WP. Negative symptoms of schizophrenia: clinical characteristics, pathophysiological substrates, experimental models and prospects for improved treatment. Eur Neuropsychopharmacol 2014 May;24(5):645-692. [doi: 10.1016/j.euroneuro.2014.03.008] [Medline: 24820238]
5. Carbon M, Correll CU. Thinking and acting beyond the positive: the role of the cognitive and negative symptoms in schizophrenia. CNS Spectr 2014 Dec;19 Suppl 1(suppl 1):38-52. [doi: 10.1017/S1092852914000601] [Medline: 25403863]
6. Medalia A, Saperstein AM, Erlich MD, Sederer LI. Cognitive remediation in large systems of psychiatric care. CNS Spectr 2019 Feb;24(1):163-173. [doi: 10.1017/S1092852918000822] [Medline: 29716665]
7. Wykes T, Huddy V, Cellard C, McGurk SR, Czobor P. A meta-analysis of cognitive remediation for schizophrenia: methodology and effect sizes. AJP 2011 May;168(5):472-485. [doi: 10.1176/appi.ajp.2010.10060855]
8. Twamley EW, Vella L, Burton CZ, Heaton RK, Jeste DV. Compensatory cognitive training for psychosis: effects in a randomized controlled trial. J Clin Psychiatry 2012 Sep;73(9):1212-1219. [doi: 10.4088/JCP.12m07686] [Medline: 22939029]
9. Angermeyer MC, Matschinger H, Schomerus G. Attitudes towards psychiatric treatment and people with mental illness: changes over two decades. Br J Psychiatry 2013 Aug;203(2):146-151. [doi: 10.1192/bjp.bp.112.122978] [Medline: 23787060]
10. Medalia A, Saperstein AM, Stefancic A, et al. Feasibility and acceptability of remotely accessed cognitive remediation for schizophrenia in public health settings. Psychiatry Res 2021 Jul;301:113956. [doi: 10.1016/j.psychres.2021.113956] [Medline: 33962354]
11. Biagianti B, Fisher M, Neilands TB, Loewy R, Vinogradov S. Engagement with the auditory processing system during targeted auditory cognitive training mediates changes in cognitive outcomes in individuals with schizophrenia. Neuropsychology 2016 Nov;30(8):998-1008. [doi: 10.1037/neu0000311] [Medline: 27617637]
12. Matsuoka K, Kin Y. Japanese Adult Reading Test: JART: Shinkou Igaku Shuppan; 2006.

XSL•FO
RenderX

13. Twamley EW, Savla GN, Zurhellen CH, Heaton RK, Jeste DV. Development and pilot testing of a novel compensatory cognitive training intervention for people with psychosis. Am J Psychiatr Rehabil 2008 Apr;11(2):144-163. [doi: 10.1080/15487760801963678] [Medline: 19198664]
14. Otsuka S, Matsui M, Hoshino T, Miura K, Higuchi Y, Suzuki M. The effectiveness and applicability of compensatory cognitive training for Japanese patients with schizophrenia: a pilot study. Advances in Psychiatry 2015 Jul 7;2015:1-12. [doi: 10.1155/2015/314804]
15. Andreasen NC. The scale for the assessment of positive symptoms (SAPS). : The University of Iowa; 1984 URL: https://eprovide.mapi-trust.org/instruments/scale-for-the-assessment-of-positive-symptoms [accessed 2025-07-26]
16. Andreasen NC. Scale for assessment of negative symptoms (SANS). : University of Iowa Press; 1983 URL: https://eprovide.mapi-trust.org/instruments/scale-for-assessing-negative-symptoms [accessed 2025-07-26]
17. Wilson B, Cockburn J, Baddeley A, Hiorns R. The development and validation of a test battery for detecting and monitoring everyday memory problems. J Clin Exp Neuropsychol 1989 Dec;11(6):855-870. [doi: 10.1080/01688638908400940] [Medline: 2592527]
18. Watamori S, Hara H, Miyamori T, Eto F. The Rivermead Behavioral Memory Test (RBMT) Japanese Version: Tokyo: Chiba Test Center; 2015, Vol. Revised version. URL: https://www.chibatc.co.jp/cgi/web/index.cgi?c=catalogue-zoom&pk=165 [accessed 2025-07-26]
19. Brain Function Test Committee, Japan Society for Higher Brain Dysfunction. Trail Making Test, Japanese Edition (TMT-J)" Edited by the Brain Function Test Committee of the Japan Society for Higher Brain Dysfunction: Shinkoh Igaku Shuppan Co. Ltd; 2019, Vol. 1. URL: https://shinkoh-igaku.jp/cgi-bin/order_inspection/tmt-j/ordermail.cgi [accessed 2025-07-26]
20. Wechsler D. Wechsler Adult Intelligence Scale-IV: Pearsom; 2008. [doi: 10.1037/t15169-000]
21. Wechsler D, Coalson D. WAIS-IV Technical and Interpretive Manual: Pearsom; 2008. [doi: 10.1037/t15169-000]
22. Matsui M, Yuuki H, Kato K, Kurachi M. Impairment of memory organization in patients with schizophrenia or schizotypal disorder. J Int Neuropsychol Soc 2006 Sep;12(5):750-754. [doi: 10.1017/S1355617706060905] [Medline: 16961956]
23. Nemoto T, Fujii C, Miura U, Chino B, Kobayashi H, Yamazawa R, et al. Reliability and validity of the Social Functioning Scale Japanese version (SFS-J). JPN Bull Soc Psychiat 2008;17:188-195 [FREE Full text]
24. Kaneda Y, Imakura A, Fujii A, Ohmori T. Schizophrenia quality of life scale: validation of the Japanese version. Psychiatry Res 2002 Dec 15;113(1-2):107-113. [doi: 10.1016/s0165-1781(02)00240-8] [Medline: 12467950]
25. Bowie CR, Bell MD, Fiszdon JM, et al. Cognitive remediation for schizophrenia: an expert working group white paper on core techniques. Schizophr Res 2020 Jan;215:49-53. [doi: 10.1016/j.schres.2019.10.047] [Medline: 31699627]
26. Keefe RSE, Cañadas E, Farlow D, Etkin A. Digital intervention for cognitive deficits in major depression: a randomized controlled trial to assess efficacy and safety in adults. Am J Psychiatry 2022 Jul;179(7):482-489. [doi: 10.1176/appi.ajp.21020125] [Medline: 35410496]
27. Best MW, Milanovic M, Tran T, et al. Motivation and engagement during cognitive training for schizophrenia spectrum disorders. Schizophr Res Cogn 2020 Mar;19:100151. [doi: 10.1016/j.scog.2019.100151] [Medline: 31828022]
28. Green MF, Kern RS, Heaton RK. Longitudinal studies of cognition and functional outcome in schizophrenia: implications for MATRICS. Schizophr Res 2004 Dec 15;72(1):41-51. [doi: 10.1016/j.schres.2004.09.009] [Medline: 15531406]
29. Bayley PJ, Frascino JC, Squire LR. Robust habit learning in the absence of awareness and independent of the medial temporal lobe. Nature New Biol 2005 Jul 28;436(7050):550-553. [doi: 10.1038/nature03857]

Abbreviations

CCT: compensatory cognitive training
CRT: cognitive remediation therapy
JSQLS: Japanese Schizophrenia Quality of Life Scale
JVLT: Japanese Verbal Learning Test
QOL: quality of life
r-CCT: remote compensatory cognitive training
RBMT: Rivermead Behavioral Memory Test
SANS: Scale of Negative Symptoms
SAPS: Scale of Positive Symptoms
SFS-J: Social Functioning Scale Japanese version
TMT-A: Trail Making Test Part A
TMT-B: Trail Making Test Part B
WAIS-IV: Wechsler Adult Intelligence Scale–Fourth Edition

XSL•FO
RenderX

Edited by A Mavragani; submitted 07.01.25; peer-reviewed by A Lynham, A Vignapiano, M Han; revised version received 16.06.25; accepted 05.07.25; published 12.08.25.

Please cite as:
Aijo R, Matsui M
Feasibility and Acceptability of Remotely Accessed Compensatory Cognitive Training for Japanese People With Schizophrenia: Pilot Study
JMIR Form Res 2025;9:e70916
URL: https://formative.jmir.org/2025/1/e70916
doi:10.2196/70916

© Ritsuko Aijo, Mie Matsui. Originally published in JMIR Formative Research (https://formative.jmir.org), 12.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Assessing the Ability to Use eHealth Resources Among Older Adults: Cross-Sectional Survey Study

Bernard Aoun*, MD; Jon O Ebbert*, MD, MSc; Priya Ramar*, MPH; Daniel L Roellinger*, BS; Lindsey M Philpot*, MPH, PhD

Division of Community Internal Medicine Geriatrics and Palliative Care, Department of Quantitative Health Sciences, Mayo Clinic, 200 First Street SW, Rochester, MN, United States

*all authors contributed equally

Corresponding Author:
Lindsey M Philpot, MPH, PhD
Division of Community Internal Medicine Geriatrics and Palliative Care, Department of Quantitative Health Sciences, Mayo Clinic, 200 First Street SW, Rochester, MN, United States

Abstract

Background: Increasing reliance on digital health resources can create disparities among older patients. Understanding health-related, mobility, and socioeconomic factors associated with the use of eHealth technologies is important for addressing inequitable access to health care.

Objective: We sought to assess digital health literacy among patients aged ≥65 years and identify factors associated with their ability to access, understand, and use digital health resources.

Methods: We developed a survey instrument grounded in the Technology Acceptance Model and conducted a cross-sectional, mixed-mode survey of patients aged ≥65 years from an integrated, multispecialty medical center. Digital health literacy was measured using the eHeals health literacy scale, and responses were analyzed across self-rated health, self-reported mobility, and socioeconomic deprivation assessed with the Area Deprivation Index (ADI). Counts (n) and frequencies (%) are reported across response groups, and analyses for differences are performed using the χ^2 test for independence or the Fisher exact test.

Results: Analyses included 878 responses (response rate=878/2847; 30.8%). There was a significant difference in the distribution of race between responders and nonresponders (P<.001) but no significant differences were observed by age (P=.053) or gender (P=.73). Respondents with lower self-rated health had lower levels of digital health literacy; only 54.2% (n=13/25) participants with poor self-rated health were able to send a message to their doctor compared to 89.5% (n=68/77) of patients with excellent self-rated health. All comparisons across the digital health literacy domains revealed significant differences across self-rated health groups (P<.05). Respondents with mobility restrictions had lower levels of digital health literacy, including lower frequencies of reporting knowledge of what health resources are available on the internet (mobility restricted, n=92/182; 52.0% vs no mobility restriction, n=433/688; 64.7%), knowledge of how to find health resources on the internet (mobility restricted, n=120/182; 67.4% vs no mobility restriction, n=513/688; 76.8%), and ability to use a camera or video with a doctor easily (mobility restricted, n=58/182; 32.6% vs no mobility restriction, n=321/688; 48.0%). Older adults experiencing increased socioeconomic deprivation, as measured by the ADI, reported lower rates of digital health literacy across most categories, including knowledge of how to find health resources on the internet (high ADI, n=28/49; 59.6% vs low ADI, n=551/751; 75.5%) and the ability to send an electronic message to their doctor easily (high ADI, n=27/49; 57.4% vs low ADI, n=584/751; 80.2%).

Conclusions: Our findings highlight the need for targeted interventions to improve engagement with eHealth among patients aged ≥65 years, who are impacted by poor health, limited mobility, and socioeconomic deprivation. Enhancing digital health literacy can help bridge the gap in access to digital health resources and improve overall health outcomes for this population.

(JMIR Form Res 2025;9:e70672) doi:10.2196/70672

KEYWORDS

patient portal; digital health; digital literacy; older adults; socioeconomic deprivation

Introduction

Digital health care (eHealth) has become a fundamental feature of modern clinical practice. eHealth encompasses a range of digital tools that support health care delivery and patient-provider communication. This includes patient portals, telehealth services, mobile apps, and online health resources that are increasingly integrated into everyday care. Its adoption has accelerated during the COVID-19 pandemic [1] with the development of dependable platforms, health care workforce shortages, and increases in demand for health care services in

XSL•FO
RenderX

growing urban and rural areas. Digital health literacy is essential for patients to use online health tools effectively. It includes the ability to find, interpret, and apply electronic health information, as well as skills for participating in telehealth visits and communicating with providers.

Older adults are expected to constitute a larger share of the population [2], and they face barriers to digital engagement due to changes in health, mobility, and finances placing them at a disadvantage with the growing introduction of digital health modalities. Internet usage decreases with advancing age, especially among those lacking a college education, which is further constrained by chronic health conditions and health care access [3]. In addition, patient health status changes, such as the onset of cognitive impairment, placement in a long-term care facility, or declining physical capabilities can compromise their ability to use technological solutions [4]. Higher education levels, autonomous internet use, possession of internet-capable devices, and earlier engagement with patient portal messaging increase the probability of technology adoption [5].

Investigations of factors influencing the adoption of eHealth by seniors have highlighted the impact of age, educational level, socioeconomic status, health literacy, and ability to engage with digital resources [6,7]. Higher levels of socioeconomic deprivation, measured by the Area Deprivation Index (ADI), are consistently associated with reduced eHealth technology adoption, particularly telemedicine (eg, reliance on audio-only over video) [8-11], but extant research is limited in the examination of US-based older adult residents, ADI-specific subgroups, and contemporary data since the end of the COVID-19 pandemic. Patients with mobility limitations have been observed to use the internet for health-related tasks [12], which may have increased during the COVID-19 pandemic [13], but gaps may still remain. Direct links between self-reported health status and eHealth disparities are also underexamined, with most studies addressing health status via proxy measures (ie, comorbidities) rather than subjective health ratings [14-16].

In the current study, we sought to advance our understanding of the relationship between self-reported health, mobility, and socioeconomic deprivation and digital health literacy, specifically the confidence, skills, and use of eHealth resources. We analyzed data from a survey including 871 individuals aged ≥65 years responding to a questionnaire about digital health literacy. Moreover, we aimed to identify potential barriers to eHealth engagement in older adults that could be addressed through targeted interventions by policymakers and health care systems.

Methods

Study Overview

As the current investigation is a subanalysis of a larger study, full details of the study approach can be found in the original article [17]. In brief, a cross-sectional survey was developed and deployed using 2 modalities based on patient-indicated preference for contact: electronic deployment via Qualtrics Survey Software (Provo) to patient-provided e-mail address or paper survey delivered via US mail to patient-provided permanent residence. Stamped return envelopes were included with all paper-based, mailed surveys. Responses to both survey modalities were appended into a single dataset for analysis purposes. The Mayo Clinic Survey Research Center performed the build, deployment, and management of both survey modalities. The survey was deployed from August to October of 2023. This study investigated our research questions among the respondents aged ≥65 years at the time of survey completion.

Study Population

As part of the routine process for clinical care delivery at any Mayo Clinic practice site, patients are offered enrollment within an electronic medical record-based patient portal. Within this portal, patients can access their medical chart information, view upcoming medical appointments, and send and receive messages from their medical team, among other features. At our institution, most patients register to use the patient portal (90% in 2022), while approximately 10% of patient accounts are inactive due to infrequent recent use, pending activation by the end user, or the patient declined enrollment. We sampled patients in a 2:1 ratio of nonregistered patients to registered and active patients to gain representation across both populations. Patients were included within the original sample if they were 18 years and older at the time of survey sampling, were still living according to the US Census and our internal medical records, and had selected English as their primary language within our medical record system. We sampled a total of 11,424 individuals for our primary analysis, of which 1850 (16.2%) completed our survey instrument and 871 were aged ≥65 years and included within the present secondary analysis.

Survey Instrument

Based on the technology acceptance model, a widely used framework that explains how users come to accept and use a technology, an interdisciplinary team developed a survey instrument containing a mix of validated survey instruments and with novel survey items focusing on assessment of patient perceptions on the usefulness of an electronic medical record-based patient portal, patient perceptions on the ease of use of the patient portal, patient intent to use the patient portal, and actual use behavior of the portal. In addition, we incorporated a validated measure of digital health literacy. Face validation of the survey instrument was performed through 3 rounds of iterative testing and rewording. A copy of the complete deployed survey instrument is provided in Multimedia Appendix 1.

For this secondary analysis, a variety of demographic and technological ability assessments were used, including the eHEALS health literacy scale (an 8-item measure that assesses an individual's knowledge, comfort, and perceived skills at finding, evaluating, and applying electronic health information to health problems), developed by Norman and Skinner [18]. Demographic information included age at the time of survey deployment (years and extracted from the medical record), gender, and race. Respondents were also asked to rate their level of comfort reading and speaking English, their ability to connect to the internet for general purposes, their ability to access connection devices (smartphone, laptop, personal computer,

XSL•FO
RenderX

tablet, other), their ability to access the internet from home, and satisfaction with their internet connection from home.

To address our study aims, we asked respondents to indicate their current level of self-rated health on a 5-point Likert scale using the single item deployed within the National Health and Nutrition Examination Survey. This single-item report of self-rated health has demonstrated a correlation with the multidimensions of health and well-being, including social, economic, physical, environmental, and mental aspects of one's life [19]. Reliability measures of this single-item self-rated health have been moderate to strong within large, population-based samples in the United States [20]. Answers reflect an individual's comprehensive or holistic belief of their own well-being, and unlike objective clinical measures which require medical record access or clinical assessment, self-rated health can be easily and reliably collected. Furthermore, how people perceive their health may influence whether they see value in digital health resources, potentially being a more direct factor in technology adoption than objective disease measurements. To assess respondents' self-reported mobility, we used the mobility question from the EuroQual-5 dimension (EQ-5D) scale [21] ("I have no problems in walking about" versus "I have some problems in walking about" and "I am confined to bed" combined). The overall performance of the EQ-5D has been well documented among older adults, including validation and reliability [22,23]. To assess the relationship between socioeconomic deprivation (SED) and digital disparities, we deployed the standardized ADI [24]. ADI is a neighborhood-level approximation for SED based on a census block of 17 individual indicators, including median family income, percent of single-family households, and population with less than a high school education, among others. High scores of SED measured by the ADI correlate with poorer health outcomes, including higher rates of chronic disease associated with limited access to health care services, healthy food options, and safe spaces for recreation, shorter life expectancy due to higher exposure to environmental hazards and increased stress levels, and higher rates of mental health issues [25]. The reliability and validity of the ADI have been previously reported [26]. We created a dichotomized SED flag at the 80% standardized ADI rank value, as has been published elsewhere [27].

Survey Fielding

Our cross-sectional survey was deployed via two modes to address the study aims, an electronic survey and a paper survey instrument. Contact information for the sampled population was derived from the institution's integrated electronic medical record system. Deployment mode was selected based on patient-provided preferences for communications documented in the electronic medical record system. Electronic surveys were designed, managed, and deployed using Qualtrics survey software (Provo; Qualtrics) and included 3 total invitations to complete the survey instrument. Paper surveys were created using InDesign software (Microsoft Corporation) and deployed in a scannable booklet format via the US Postal Service in a single distribution wave. Stamped return envelopes were included with all mailed surveys. Those who did not respond to their electronic survey received a paper survey within 90 days of nonresponse. Responses to electronic and paper surveys were appended into a single dataset for analysis purposes. The build, deployment, and management of survey instruments were performed by the Mayo Clinic Survey Research Center.

Statistical Analysis

Simple count (n) and proportion (%) are provided as descriptive statistics for all reported data. Comparisons across groups are performed either using the χ^2 test for independence or the Fisher exact test when the assumptions of the χ^2 test were not met. All data management and statistical analyses were performed using Statistical Analysis Software (SAS, version 9.2). Results were considered statistically significant if $P<.05$. Assessment for bias due to nonresponse is included in Multimedia Appendix 2.

Ethical Considerations

This study (IRB#22‑008356; principal investigator: LM Philpot) underwent expedited review by the Mayo Clinic Institutional Review Board (Rochester, MN, USA) and was approved as exempt (45 CFR 46.104d, Category 2). As protected health information was not being requested from survey respondents, HIPAA (Health Insurance Portability and Accountability Act) authorization was not required in accordance with 45 CFR 160.103. Study data were anonymized, deidentified, and only presented in aggregate form. No compensation was provided to survey respondents.

Results

An overview of the study inclusion is shown in Figure 1. A total of 2847 surveys were sent to patients aged ≥65 years. Of those, we received 878 complete responses (response rate 30.8%); 49.7% (436/878) of respondents completed the paper-based survey and 50.3% (442/878) of respondents completed the electronic survey. Survey respondents had a reported average age of 74.5 (SD 6.3) years; 51.3% (n=448) were women and 82.1% (n=715) were White. Across the entire response population, more than 81% (n= 711) of respondents had access to a computer or a smartphone and more than 93% (n=816) of respondents could access and connect to the internet to surf and use email (Table 1).

XSL•FO
RenderX

Figure 1. Flowchart of patients included in the survey sample.

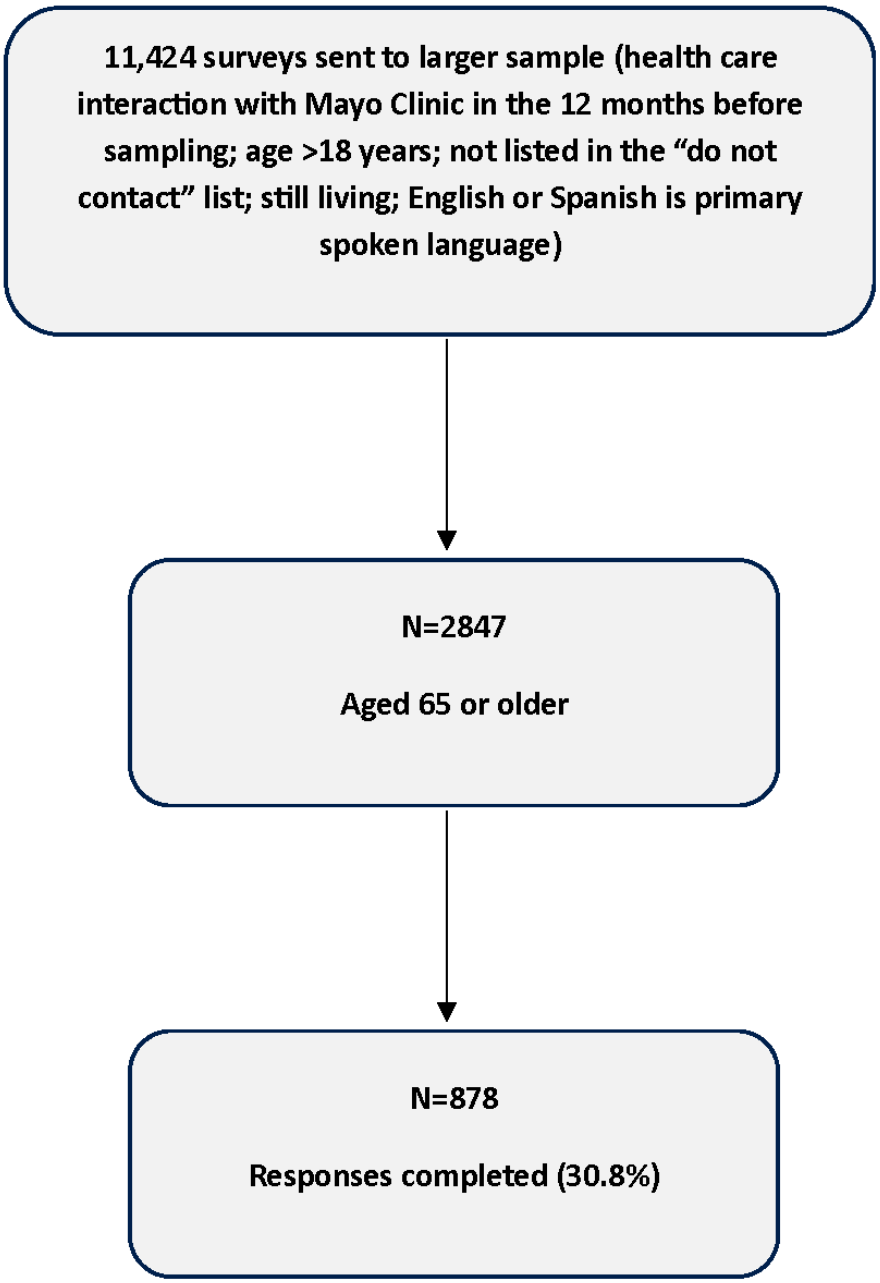

XSL•FO
RenderX

Table . Demographic characteristics of patients aged ≥65 years completing a survey on digital health literacy.

*Demographic characteristics	Participants (N=878), n (%)
Gender	
Missing	5 (0.6)
Women	448 (51.3)
Men	417 (47.8)
Transgender women	1 (0.1)
Prefer not to answer	5 (0.6)
Other	2 (0.2)
Race	
Missing	7 (0.8)
Asian/Pacific Islander	37 (4.2)
Black/African/African American	34 (3.9)
Mexican	9 (1.0)
Middle Eastern	2 (0.2)
Mixed	27 (3.1)
Native American/Native Hawaiian	16 (1.8)
No Answer	15 (1.7)
None/Other	16 (1.8)
White	715 (82.1)
Self-rated health	
Missing	7 (0.8)
Excellent	77 (8.8)
Very Good	287 (33)
Good	325 (37.3)
Fair	157 (18)
Poor	25 (2.9)
Comfort level: reading English	
Missing	9 (1.0)
Comfortable	848 (97.6)
Neutral	19 (2.2)
Uncomfortable	2 (0.2)
Comfort level: spoken English	
Missing	5 (0.6)
Comfortable	856 (98.1)
Neutral	16 (1.8)
Uncomfortable	1 (0.1)
Socioeconomic deprivation (Area Deprivation Index >=80%)	
Missing	78 (8.9)
Yes	49 (6.1)
No	751 (93.9)
Is patient's mobility restricted?	
Missing	8 (0.9)
Yes	182 (20.9)

XSL•FO
RenderX

*Demographic characteristics	Participants (N=878), n (%)
No	688 (79.1)
Connect to internet to surf web/email?	
Missing	20 (2.3)
Yes	802 (93.5)
No	52 (6.1)
I don't know	4 (0.5)
Access internet from home?	
Missing	23 (2.6)
Yes	824 (96.4)
No	29 (3.4)
I don't know	2 (0.2)
Satisfaction in ability to access internet?	
Missing	53 (6.0)
Satisfied	694 (84.1)
Neither	100 (12.1)
Dissatisfied	31 (3.8)
Device usage	
Desktop/Laptop Computer	
Yes	714 (81.3)
No	164 (18.7)
Tablet	
Yes	481 (54.8)
No	397 (45.2)
Smartphone	
Yes	747 (85.1)
No	131 (14.9)
None of the above	
Yes	12 (1.4)
No	866 (98.6)

Self-Rated Health and Digital Health Literacy

Overall, most respondents felt confident in knowing what health resources were available, where, and how to find them, how to use the resources, and how to evaluate and use the information found in those resources (Table 2). However, less than half of the respondents felt confident in using and being able to determine the quality of health information found on the internet. While nearly 80% (n=673) of respondents could send a message to a doctor easily on the day, less than 45% (n=379) felt they could easily use a camera or video with a doctor on the day. Most respondents in each self-rated health group felt that they could send a message to a doctor easily on the day, though there were significant differences in distributions becoming more pronounced with fewer respondents reporting being able to do this among groups with lower self-rated health. In contrast, most (n=41, 54.7%) of the respondents in the excellent self-rated health group reported being able to use a camera or video with a doctor on the day, while this was the case for less than half of respondents with very good (n=132, 47.3%) and good (n=147, 46.4%) self-rated health, and lower for those in fair (n=54, 35.5%) and poor (n=5, 20.8%) self-rated health.

XSL•FO
RenderX

Table . Self-rated health by digital literacy among patients aged ≥65 years completing a survey on digital health literacy.

Self-rated digital health literacy variables	Excellent (n=77), n (%)	Very good (n=287), n (%)	Good (n=325), n (%)	Fair (n=157), n (%)	Poor (n=25), n (%)	Total (N=871), n (%)	*P* value
Health resources - What is available							.01
Yes	53 (71.6)	190 (68.1)	190 (60.1)	81 (52.9)	11 (45.8)	525 (62.1)	
No	7 (9.5)	45 (16.1)	63 (19.9)	39 (25.5)	5 (20.8)	159 (18.8)	
I don't know	14 (18.9)	44 (15.8)	63 (19.9)	33 (21.6)	8 (33.3)	162 (19.1)	
Health resources - Where to find							.02
Yes	63 (84)	208 (74)	239 (74.7)	93 (61.2)	15 (62.5)	618 (72.5)	
No	4 (5.3)	36 (12.8)	37 (11.6)	28 (18.4)	5 (20.8)	110 (12.9)	
I don't know	8 (10.7)	37 (13.2)	44 (13.8)	31 (20.4)	4 (16.7)	124 (14.6)	
Health resources - How to find them							.005
Yes	64 (86.5)	215 (77.1)	245 (77.3)	95 (62.5)	15 (62.5)	634 (74.9)	
No	4 (5.4)	34 (12.2)	36 (11.4)	28 (18.4)	5 (20.8)	107 (12.6)	
I don't know	6 (8.1)	30 (10.8)	36 (11.4)	29 (19.1)	4 (16.7)	105 (12.4)	
Health resources - How to use							.02
Yes	62 (83.8)	221 (79.5)	238 (74.8)	97 (63.8)	15 (62.5)	633 (74.8)	
No	6 (8.1)	27 (9.7)	43 (13.5)	29 (19.1)	5 (20.8)	110 (13)	
I don't know	6 (8.1)	30 (10.8)	37 (11.6)	26 (17.1)	4 (16.7)	103 (12.2)	
Health information - How to use							.02
Yes	54 (73)	214 (76.4)	223 (70.1)	95 (62.1)	15 (62.5)	601 (70.8)	
No	7 (9.5)	28 (10)	39 (12.3)	28 (18.3)	7 (29.2)	109 (12.8)	
I don't know	13 (17.6)	38 (13.6)	56 (17.6)	30 (19.6)	2 (8.3)	139 (16.4)	
Health information - skills to evaluate							.003
Yes	58 (78.4)	166 (59.5)	186 (59.2)	78 (51)	10 (43.5)	498 (59.1)	
No	6 (8.1)	55 (19.7)	54 (17.2)	41 (26.8)	8 (34.8)	164 (19.5)	
I don't know	10 (13.5)	58 (20.8)	74 (23.6)	34 (22.2)	5 (21.7)	181 (21.5)	
Health information - differentiate quality							.03
Yes	49 (66.2)	136 (48.9)	146 (46.2)	63 (41.2)	12 (50)	406 (48)	
No	9 (12.2)	70 (25.2)	72 (22.8)	41 (26.8)	8 (33.3)	200 (23.7)	
I don't know	16 (21.6)	72 (25.9)	98 (31)	49 (32)	4 (16.7)	239 (28.3)	
Health information - confidence in using							<.001
Yes	49 (66.2)	122 (43.9)	135 (42.7)	46 (30.3)	10 (41.7)	362 (42.9)	
No	15 (20.3)	85 (30.6)	105 (33.2)	63 (41.4)	10 (41.7)	278 (32.9)	
I don't know	10 (13.5)	71 (25.5)	76 (24.1)	43 (28.3)	4 (16.7)	204 (24.2)	

XSL•FO
RenderX

Self-rated digital health literacy variables	Excellent (n=77), n (%)	Very good (n=287), n (%)	Good (n=325), n (%)	Fair (n=157), n (%)	Poor (n=25), n (%)	Total (N=871), n (%)	*P* value
Able to send a message to doctor easily today?							<.001
Yes	68 (89.5)	233 (84.1)	260 (82.8)	99 (65.1)	13 (54.2)	673 (79.8)	
No	7 (9.2)	30 (10.8)	33 (10.5)	39 (25.7)	10 (41.7)	119 (14.1)	
I don't know	1 (1.3)	14 (5.1)	21 (6.7)	14 (9.2)	1 (4.2)	51 (6)	
Able to use camera/video with doctor easily today?							.007
Yes	41 (54.7)	132 (47.3)	147 (46.4)	54 (35.5)	5 (20.8)	379 (44.7)	
No	29 (38.7)	121 (43.4)	148 (46.7)	77 (50.7)	18 (75)	393 (46.4)	
I don't know	5 (6.7)	26 (9.3)	22 (6.9)	21 (13.8)	1 (4.2)	75 (8.9)	

Mobility Restrictions and Digital Health Literacy

Respondents who reported restricted mobility were significantly different from those who reported no mobility restrictions in their ability to use digital health resources (Figure 2). They more often reported feeling unsure or not knowing how to use health resources online; however, they were more comfortable using health information found online. While most respondents felt they could send a message to a doctor, those reporting mobility restrictions differed significantly in their ability compared to those with no restricted mobility (n=121, 68.0% vs n=553, 83.2%). Most respondents were unsure or not able to use a camera or video with a doctor (n=468, 55.2%), and this was more often the case for those with mobility restrictions compared to those with no restricted mobility (n=120, 67.4% vs n=348, 52.0%). A significant difference in satisfaction with access to internet or a computer was observed with almost a quarter of respondents with mobility restriction being unsatisfied with their ability to access the internet (n=43, 23.5% vs n=94, 13.9%) or a computer (n=46, 25.3% vs n=114, 16.6%), compared to those with no restricted mobility.

Figure 2. Mobility restrictions versus digital health literacy among older adults.

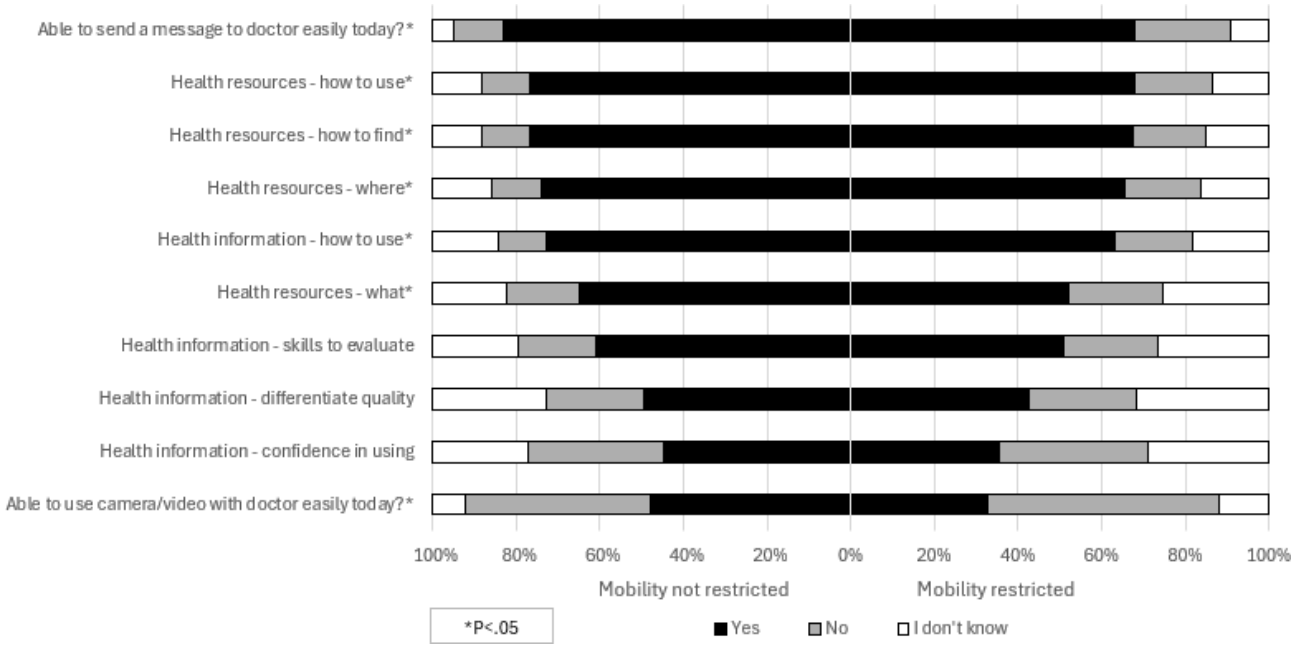

ADI and Digital Health Literacy

When comparing respondents who experience socioeconomic deprivation as measured by ADI, a significantly higher proportion of those at or above 80% ADI reported not knowing how to find or use health resources and health information (Figure 3). Overall, 78.8% (n=611) of respondents were able to send a message to a doctor on the day; however, there was a significant difference between ADI groups with fewer of those at or above 80% ADI being able to do this compared with 45.3% (n=341) among those below 80% ADI. Less than half of the respondents overall were able to use a camera or video with a doctor; there was no significant difference between the ADI groups. There was a significant difference in access to a

XSL•FO
RenderX

computer with 36.7% (n=18) of respondents at or above 80% ADI reporting no access to a computer versus 17.8% (n=134) among those below 80% ADI.

Figure 3. Social deprivation (measured by the Area Deprivation Index) by digital literacy among older adults.

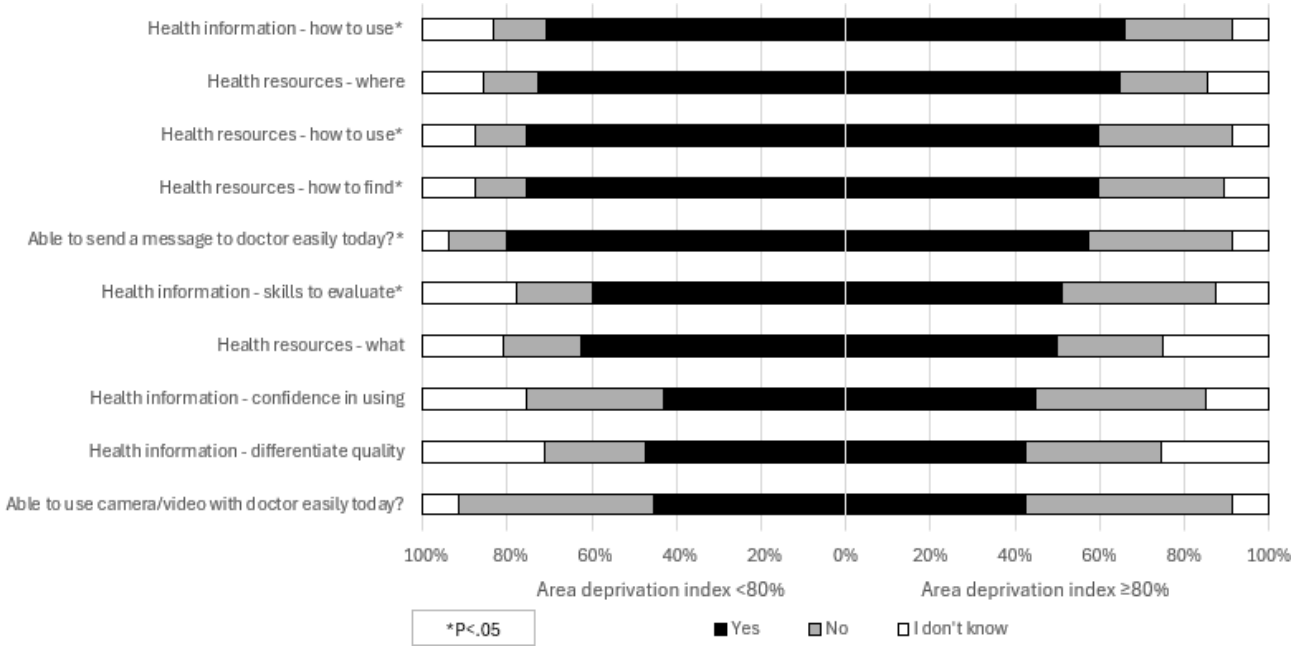

Discussion

Principal Results

In our survey among adults aged ≥65 years after the end of the COVID-19 pandemic, we successfully identified factors associated with older adults' ability to access and use digital health resources. Our findings demonstrate that low self-rated health, mobility impairments, and socioeconomic disadvantages were associated with lower digital health literacy. Specifically, respondents with poor self-rated health showed substantially reduced ability to send messages to doctors (54.2% vs 89.5% with excellent health) and use video communication. Similarly, those with mobility restrictions showed lower confidence in finding and using health resources online and were less able to send messages to providers (68.0% vs 83.2% without restrictions). Respondents living in areas with high socioeconomic deprivation also showed significant disparities, mainly in messaging providers (57.4% vs 80.2% in low-deprivation areas) and access to computers (36.7% reporting no access vs 17.8% in low-deprivation areas).

Interpretations, Implications, and Comparison With Previous Work

We observed significant relationships between low self-rated health and lower confidence and ability to use eHealth resources. This finding aligns with previous reporting that poor health is associated with lower digital health literacy and competence in the use of technology [4]. Our study also supports earlier research that seniors face more challenges in adopting innovative technologies compared with the general population [6,28]. Older adults who are suffering from health issues may have other care requirements taking precedence over the knowledge and use of digital resources, particularly when those resources are complicated, do not function as expected, or troubleshooting resources are not available.

A significant need exists to emphasize the importance of usability and accessibility in the design and development of digital health platforms [28,29]. Anxiety about the use of information and communication technologies, an issue disproportionately affecting seniors [30], can be reduced through encouragement and support of professionals, such as point-of-care teaching or advice by telephone. A patient portal eLearning program with older adults has been proven to increase self-efficacy and portal use among older adults [31] and could be tailored to specific subpopulations. Patients with poor self-rated health or suffering from cognitive decline may need shorter, more frequent sessions. Implementing a standardized assessment of patients' digital readiness before scheduling telehealth visits would help identify those needing more support.

Our finding that less than 45% of responders felt confident using video with the provider suggests that previsit technical preparation is essential. Our data show that only 54.2% of individuals with poor self-rated health could message their provider, highlighting the need to establish a formal proxy user such as a family member or caregiver to help with digital health tasks while supporting privacy protections. Our finding that mobility limitations are associated with lower confidence and ability to use eHealth technologies aligns with earlier research. In a survey of 7609 community-dwelling Medicare beneficiaries, technology use decreased significantly with greater limitations in physical capacity and greater disability after adjustment for sociodemographic and health characteristics [12]. In a survey of 8019 individuals between the ages of 75 and 99, individuals with impaired functional ability or vision perceived digital health to be less beneficial than their counterparts with less impairment [29].

XSL•FO
RenderX

Older adults who are homebound and disabled due to mobility issues may have difficulties using nonadaptable devices. User interfaces requiring physical coordination (eg, camera installations or adjustments), dexterity (eg, holding a handheld device steadily), or sitting at a computer make these platforms less useful to individuals with physical limitations. Overcoming these obstacles requires addressing both the technical aspect of access with adaptable access options to people with age-related physical decline [32] and improving the functional aspect of access. Technology developers should prioritize user-centered designs such as larger text options, simplified navigation, sharp contrast displays, and voice command capabilities that can accommodate those with visual, cognitive, or dexterity limitations. Our findings that only 32.6% of mobility-restricted older adults felt comfortable using video with their doctor suggest that video interfaces require significant redesign for this population. Enhancing remote training capabilities to enable telemedicine visit engagement has been evaluated [33] and may provide opportunities to bridge the digital divide in the older adult subpopulation with mobility limitations.

Our finding that older adults living with socioeconomic deprivation report lower rates of knowledge of how to find or use eHealth resources and the ability to message their provider adds to extant literature. Previously published literature has proven that higher ADI correlates with lower telemedicine usage, particularly video-based consultations [8-11]. Compared with patients with a lower ADI in our survey, patients with high ADI were not more likely to report difficulty with engaging with video but did report more difficulty with sending messages to their doctor.

A digital divide has also been shown among patients with socioeconomic deprivation which varies by technology. Among patients living in economically disadvantaged zip codes, 32% - 40% of patients own a computer or an iPad compared to 59% from more affluent neighborhoods [34]. However, among individuals reporting a pretax annual income of < $35,000 per year, 96% report owning a smartphone, although only 15% report using a medical app [35]. Furthermore, consumer technology is being embraced by older adults (aged ≥65 y) [36] and the number of older adults who own a smartphone has risen from 18% in 2013 to 61% in 2021 among individuals aged ≥65 years [37]. In our survey, 85.1% of patients reported having a smartphone. Patients with socioeconomic deprivation based upon census data are more likely to access their medical records from a mobile device [38]. Facilitating care through mobile devices, templated instructions, and anticipatory guidance has been proposed as a practical solution to bridge the digital divide among patients with socioeconomic deprivation [39]. Given that 36.7% of respondents in high-deprivation areas reported no access to a computer, health care organizations should consider setting up technology lending programs that provide devices (tablets and laptops) and Internet connectivity to socioeconomically disadvantaged older patients. These programs could be integrated with existing social services and community outreach initiatives.

Limitations and Strengths

Our study has several limitations. First, the cross-sectional design does not allow for assessing changes over time. Longitudinal studies are needed to understand how digital literacy evolves as health status, mobility, and socioeconomic circumstances change. Second, the use of self-reported data is limited by recall bias and may not accurately reflect participants' actual digital capabilities. Third, we did not identify the specific comorbidities, functional limitations, or deprivation domains associated with lower use of eHealth resources. Fourth, surveys were sent to patients receiving care at a single health system, with most respondents being White and within a limited geographical area, which reduces generalizability to more diverse populations, those receiving care in different health care settings, or those living in different geographic regions with varying internet infrastructure. Analysis of nonresponse showed a significant difference in the distribution of race between responders and nonresponders, with higher proportions of non-White race groups among nonresponders. Finally, although our study was conducted after the acute phase of the COVID-19 pandemic, the extraordinary circumstances of the pandemic may have altered digital health usage in ways that will not persist in the future.

Our study has several strengths. First, we oversampled patients who were not registered with our patient portal to ensure representation from this group. Second, we assessed perceptions, intent, and use grounded in a well-known theoretical model (ie, technology acceptance model) which expands our understanding of actual uptake and use of eHealth solutions. Third, our use of validated assessment tools enhances the accuracy and applicability of the findings.

Conclusions

Our findings highlight digital disparities of older adults associated with lower self-rated health, mobility limitations, and socioeconomic deprivation. These results have important implications for health care policy and practice and may allow health care systems to proactively identify older patients who would benefit from targeted interventions.

Rather than adopting a one-size-fits-all approach to digital health implementation, our data suggests the need for user-centered platforms with adaptable accessibility options that can accommodate the diverse needs of aging populations. This includes developing mobile-optimized solutions for socioeconomically disadvantaged patients who may lack computer access, implementing proxy user systems for those with poor health status, and redesigning video interfaces for mobility-restricted individuals.

Digital health tools present both an opportunity and a challenge: they can help close longstanding gaps in care or further marginalize those already facing barriers, especially older adults. Our findings underscore the importance of designing digital health systems with equity in mind.

XSL•FO
RenderX

Acknowledgments

We thank the Mayo Clinic Survey Research Center for their help in planning, testing, and implementing the survey.

This publication was made possible by CTSA Grant Number UL1 TR002377 from the National Center for Advancing Translational Sciences (NCATS), a component of the National Institutes of Health (NIH). Contents of this manuscript are solely the responsibility of the authors and do not necessarily represent the official view of NIH.

The Center for Digital Health at Mayo Clinic supported this work through funding from the Noaber Foundation (TN006478). The authors bear full responsibility for the content and any errors or omissions, and the views expressed do not necessarily reflect those of the funding agency.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

BA, JOE, PR, and LMP performed conceptualization. PR, DLR, and LMP performed data curation. PR and DLR conducted formal analysis. LMP acquired funding. BA, JOE, and LMP carried out the investigation. PR and LMP developed the methodology. LMP provided resources. PR, DLP, and LMP developed the software. LMP and JOE provided supervision. PR and LMP conducted validation. BA, JOE, PR, DLR, and LMP wrote the original draft. BA, JOE, PR, DLR, and LMP reviewed and edited the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Digital health inclusion survey.
[PDF File, 141 KB - formative_v9i1e70672_app1.pdf]

Multimedia Appendix 2
Assessment for nonresponse bias.
[DOCX File, 19 KB - formative_v9i1e70672_app2.docx]

References

1. Koonin LM, Hoots B, Tsang CA, et al. Trends in the use of telehealth during the emergence of the COVID-19 pandemic - United States, January-March 2020. MMWR Morb Mortal Wkly Rep 2020 Oct 30;69(43):1595-1599. [doi: 10.15585/mmwr.mm6943a3] [Medline: 33119561]
2. Caplan Z, Rabe M. The older population: 2020. United States Census Bureau. 2023. URL: https://www2.census.gov/library/publications/decennial/2020/census-briefs/c2020br-07.pdf [accessed 2025-07-15]
3. Fealy S, McLaren S, Seaman CE, et al. Exploring the sociodemographic and health-related determinants of telehealth use among a cohort of older Australians during the COVID-19 pandemic: repeated cross-sectional study. JMIR Aging 2024 Aug 23;7:e58594. [doi: 10.2196/58594] [Medline: 39178035]
4. Levine DM, Lipsitz SR, Linder JA. Changes in everyday and digital health technology use among seniors in declining health. J Gerontol A Biol Sci Med Sci 2018 Mar 14;73(4):552-559. [doi: 10.1093/gerona/glx116] [Medline: 28605446]
5. Gordon NP, Crouch E. Digital information technology use and patient preferences for internet-based health education modalities: cross-sectional survey study of middle-aged and older adults with chronic health conditions. JMIR Aging 2019 Apr 4;2(1):e12243. [doi: 10.2196/12243] [Medline: 31518291]
6. Ausserhofer D, Piccoliori G, Engl A, et al. Community-dwelling older adults' readiness for adopting digital health technologies: cross-sectional survey study. JMIR Form Res 2024 Apr 30;8:e54120. [doi: 10.2196/54120] [Medline: 38687989]
7. Crouch E, Gordon NP. Prevalence and factors influencing use of internet and electronic health resources by middle-aged and older adults in a US health plan population: cross-sectional survey study. JMIR Aging 2019 Mar 26;2(1):e11451. [doi: 10.2196/11451] [Medline: 31518256]
8. Ostovari M, Zhang Z, Patel V, et al. Telemedicine and health disparities: association between the area deprivation index and primary care telemedicine utilization during the COVID-19 pandemic. J Clin Transl Sci 2023;7(1):e168. [doi: 10.1017/cts.2023.580] [Medline: 37588680]

XSL•FO
RenderX

9. Brown SH, Griffith ML, Kripalani S, et al. Association of health literacy and area deprivation with initiation and completion of telehealth visits in adult medicine clinics across a large health care system. JAMA Netw Open 2022 Jul 1;5(7):e2223571. [doi: 10.1001/jamanetworkopen.2022.23571] [Medline: 35877125]
10. Hassan AM, Chu CK, Liu J, et al. Determinants of telemedicine adoption among financially distressed patients with cancer during the COVID-19 pandemic: insights from a nationwide study. Support Care Cancer 2022 Sep;30(9):7665-7678. [doi: 10.1007/s00520-022-07204-1] [Medline: 35689108]
11. Wang F, Zeng Y, Liu L, et al. Disparities in spatial accessibility of primary care in Louisiana: from physical to virtual accessibility. Front Public Health 2023;11(1154574):1154574. [doi: 10.3389/fpubh.2023.1154574] [Medline: 37143988]
12. Gell NM, Rosenberg DE, Demiris G, et al. Patterns of technology use among older adults with and without disabilities. Gerontologist 2015 Jun;55(3):412-421. [doi: 10.1093/geront/gnt166] [Medline: 24379019]
13. Friedman C, VanPuymbrouck L. Telehealth use by persons with disabilities during the COVID-19 pandemic. Int J Telerehabil 2021;13(2):e6402. [doi: 10.5195/ijt.2021.6402] [Medline: 35646237]
14. Ng BP, Park C, Silverman CL, et al. Accessibility and utilisation of telehealth services among older adults during COVID-19 pandemic in the United States. Health Soc Care Community 2022 Sep;30(5):e2657-e2669. [doi: 10.1111/hsc.13709] [Medline: 34994028]
15. Rivera V, Aldridge MD, Ornstein K, et al. RESEARCHRacial and socioeconomic disparities in access to telehealth. J Am Geriatr Soc 2021 Jan;69(1):44-45. [doi: 10.1111/jgs.16904] [Medline: 33075143]
16. Ekwegh T, Cobb S, Adinkrah EK, et al. Factors associated with telehealth utilization among older African Americans in South Los Angeles during the COVID-19 pandemic. Int J Environ Res Public Health 2023 Feb 2;20(3):2675. [doi: 10.3390/ijerph20032675] [Medline: 36768041]
17. Philpot LM, Ramar P, Roellinger DL, et al. Individual-level digital determinants of health and technology acceptance of patient portals: cross-sectional assessment. JMIR Form Res 2024 Jun 10;8:e56493. [doi: 10.2196/56493] [Medline: 38695754]
18. Norman CD, Skinner HA. eHEALS: The eHealth Literacy Scale. J Med Internet Res 2006 Nov 14;8(4):e27. [doi: 10.2196/jmir.8.4.e27] [Medline: 17213046]
19. Gallagher JE, Wilkie AA, Cordner A, et al. Factors associated with self-reported health: implications for screening level community-based health and environmental studies. BMC Public Health 2016 Jul 26;16(640):640. [doi: 10.1186/s12889-016-3321-5] [Medline: 27460934]
20. Zajacova A, Dowd JB. Reliability of self-rated health in US adults. Am J Epidemiol 2011 Oct 15;174(8):977-983. [doi: 10.1093/aje/kwr204] [Medline: 21890836]
21. Brooks R. EuroQol: the current state of play. Health Policy 1996 Jul;37(1):53-72. [doi: 10.1016/0168-8510(96)00822-6] [Medline: 10158943]
22. Buchholz I, Janssen MF, Kohlmann T, et al. A systematic review of studies comparing the measurement properties of the Three-Level and Five-Level Versions of the EQ-5D. Pharmacoeconomics 2018 Jun;36(6):645-661. [doi: 10.1007/s40273-018-0642-5] [Medline: 29572719]
23. Feng YS, Kohlmann T, Janssen MF, et al. Psychometric properties of the EQ-5D-5L: a systematic review of the literature. Qual Life Res 2021 Mar;30(3):647-673. [doi: 10.1007/s11136-020-02688-y] [Medline: 33284428]
24. Kind AJH, Buckingham WR. Making neighborhood-disadvantage metrics accessible - the neighborhood Atlas. N Engl J Med 2018 Jun 28;378(26):2456-2458. [doi: 10.1056/NEJMp1802313] [Medline: 29949490]
25. Fleitas Alfonzo L, King T, You E, et al. Theoretical explanations for socioeconomic inequalities in multimorbidity: a scoping review. BMJ Open 2022 Feb 23;12(2):e055264. [doi: 10.1136/bmjopen-2021-055264] [Medline: 35197348]
26. Maroko AR, Doan TM, Arno PS, et al. Integrating social determinants of health with treatment and prevention: a new tool to assess local area deprivation. Prev Chronic Dis 2016 Sep 15;13:E128. [doi: 10.5888/pcd13.160221] [Medline: 27634778]
27. Montano AR, Ge A, Halladay CW, et al. Association of home-based primary care enrollment with social determinants of health for older veterans. R I Med J (2013) 2021 May 3;104(4):32-38. [Medline: 33926156]
28. Gordon NP, Hornbrook MC. Older adults' readiness to engage with eHealth patient education and self-care resources: a cross-sectional survey. BMC Health Serv Res 2018 Mar 27;18(1):220. [doi: 10.1186/s12913-018-2986-0] [Medline: 29587721]
29. Kainiemi E, Saukkonen P, Virtanen L, et al. Perceived benefits of digital health and social services among older adults: A population-based cross-sectional survey. Digit Health 2023;9(20552076231173559):20552076231173559. [doi: 10.1177/20552076231173559] [Medline: 37312955]
30. Reid M, Aleti T, Figueiredo B, et al. Factors influencing seniors' anxiety in using ICT. Soc Sci (Basel) 2024;13(9):496. [doi: 10.3390/socsci13090496]
31. Nahm E, Zhu S, Bellantoni M, et al. The effects of a theory-based patient portal elearning program for older adults with chronic illnesses. Innovation in Aging 2018 Nov 1;2(suppl_1):731-731. [doi: 10.1093/geroni/igy023.2700]
32. Lukkahatai N, Junxin L. Challenges and solutions for mobile health intervention research in older adults with multiple chronic conditions. Innov Aging 2022 Dec 20;6(Supplement_1):208-208. [doi: 10.1093/geroni/igac059.829]
33. Robinson KS, Morin PC, Shupe JAC, et al. Use of three computer training methods in elderly underserved rural patients enrolled in a diabetes telemedicine program. Comput Inform Nurs 2010;28(3):172-177. [doi: 10.1097/NCN.0b013e3181d785d5] [Medline: 20431360]

XSL•FO
RenderX

34. Livingood WC, Bautista MAB, Smotherman C, et al. Comparative study of different SES neighborhood clinics for health literacy and internet access. Digit Health 2022;8(20552076221123715):20552076221123715. [doi: 10.1177/20552076221123715] [Medline: 36081750]
35. Adepoju O, Dang P, Nguyen H, et al. Equity in digital health: assessing access and utilization of remote patient monitoring, medical apps, and wearables in underserved communities. Inquiry 2024;61(469580241271137):469580241271137. [doi: 10.1177/00469580241271137] [Medline: 39323052]
36. Mace RA, Mattos MK, Vranceanu AM. Older adults can use technology: why healthcare professionals must overcome ageism in digital health. Transl Behav Med 2022 Dec 30;12(12):1102-1105. [doi: 10.1093/tbm/ibac070] [Medline: 36073770]
37. Faverio M. Share of those 65 and older who are tech users has grown in the past decade. : Pew Research Center; 2022 URL: https://www.pewresearch.org/short-reads/2022/01/13/share-of-those-65-and-older-who-are-tech-users-has-grown-in-the-past-decade/ [accessed 2025-07-15]
38. Graetz I, Huang J, Brand RJ, et al. Bridging the digital divide: mobile access to personal health records among patients with diabetes. Am J Manag Care 2018 Jan;24(1):43-48. [Medline: 29350505]
39. Eyrich NW, Andino JJ, Fessell DP. Bridging the digital divide to avoid leaving the most vulnerable behind. JAMA Surg 2021 Aug 1;156(8):703-704. [doi: 10.1001/jamasurg.2021.1143] [Medline: 33769441]

Abbreviations

ADI: Area Deprivation Index
EQ-5D: EuroQual-5 dimension scale
SED: socioeconomic deprivation

Edited by A Mavragani; submitted 29.12.24; peer-reviewed by B Mudashiru, S Olisakwe; revised version received 23.05.25; accepted 25.05.25; published 06.08.25.

Please cite as:
Aoun B, Ebbert JO, Ramar P, Roellinger DL, Philpot LM
Assessing the Ability to Use eHealth Resources Among Older Adults: Cross-Sectional Survey Study
JMIR Form Res 2025;9:e70672
URL: https://formative.jmir.org/2025/1/e70672
doi:10.2196/70672

© Bernard Aoun, Jon O Ebbert, Priya Ramar, Daniel L Roellinger, Lindsey M Philpot. Originally published in JMIR Formative Research (https://formative.jmir.org), 6.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Development of a Customizable Web-Based Dashboard for Remote Blood Pressure Monitoring: Feasibility and Usability Implementation Study

Joshua Wollen[1*], PharmD; Thuy Nguyen[2*], PharmD; Charles Coton[3], MSBI; Gina DeWildt[3], BA; Gabretta Cooksey[4], AA; Audrey Cohen[5], MSBI; Nicholas Hoang[6], MSBI; Munachi Okpala[4], MSN, MBA, DNP; Michael Gonzalez[3], MSCS; Mengxi Wang[3], PhD; Chigozirim Izeogu[4], MD, MPH; Elmer Bernstam[5], MSE, MD; Charles Green[4,7], PhD; Sean Savitz[4,8], MD; Jose-Miguel Yamal[3,9], PhD; Anjail Sharrief[4], MPH, MD

[1]Department of Pharmacy Practice and Translational Research, College of Pharmacy, University of Houston, Houston, TX, United States

[2]Department of Pharmacy, Texas Medical Center, Memorial Hermann–Texas Medical Center, Houston, TX, United States

[3]Coordinating Center for Clinical Trials, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX, United States

[4]Department of Neurology, McGovern Medical School, The University of Texas Health Science Center at Houston, 6431 Fannin St, MSB 7.110, Houston, TX, United States

[5]Department of Clinical and Health Informatics, McWilliams School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, United States

[6]Health Services Research, College of Medicine, Baylor University, Houston, TX, United States

[7]Department of Pediatrics, McGovern Medical School, The University of Texas Health Science Center at Houston, Houston, TX, United States

[8]Institute for Stroke and Cerebrovascular Disease, UT Health Houston, The University of Texas Health Science Center at Houston, Houston, TX, United States

[9]Department of Biostatistics and Data Science, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX, United States

[*]these authors contributed equally

Corresponding Author:
Anjail Sharrief, MPH, MD
Department of Neurology, McGovern Medical School, The University of Texas Health Science Center at Houston, 6431 Fannin St, MSB 7.110, Houston, TX, United States

Abstract

Background: Effective blood pressure (BP) management is essential for several disease states, including secondary stroke prevention. Remote BP monitoring used in combination with an electronic data capture (EDC) system offers potential for improved BP control for patients. While such systems currently exist, there are many patient-, provider-, and system-specific challenges these systems currently do not address. In addition, such systems are rarely described transparently and comprehensively in the literature with enough detail to be reproduced by other clinicians.

Objective: This study aims to describe the development and design of a web-based EDC system tailored for remote BP monitoring and management to overcome common patient-, provider-, and system-specific challenges.

Methods: The EDC system was developed for use in the Video-based Intervention to Reduce Treatment and Outcome Disparities in Adults living with Stroke (VIRTUAL) and transient ischemic attack clinical trial, a randomized study involving adult stroke survivors receiving care for secondary stroke prevention. The EDC system supports remote monitoring using a BP monitor, which was specifically and methodically chosen for its suitability to the population. The EDC system integrates the data from the remote monitors into a web-based application, facilitating real-time monitoring for health care team intervention. The system includes a BP dashboard displaying time-series data, visual alerts, and email notifications for out-of-range readings. Proxy measures assessed system effectiveness in addressing patient-, provider-, and system-specific needs.

Results: Seventeen BP monitors were evaluated, and the BlipcareBlip BP800 was selected based on its cellular connectivity, ergonomic design, and application programming interface compatibility. The EDC system successfully linked patient data to the dashboard, enabling real-time visualization of BP trends and provider alerts. A total of 97% (184/190) of participants in the intervention arm engaged in BP monitoring, averaging 1.2 readings per day. Clinicians actively used the system, with mean monthly logins ranging from 19 to 44 per provider. Alerts were triggered for 69% (131/190) of participants, primarily for systolic BP abnormalities. Data transmission delays were minimal, with a median time of 1.4 minutes from data collection to dashboard entry. Outlier alerts were efficiently processed, with an average notification time of 17.9 hours. The system demonstrated high

XSL•FO
RenderX

functionality in addressing patient-, provider-, and system-specific needs, supporting effective BP monitoring and intervention delivery.

Conclusions: The VIRTUAL EDC system facilitates remote monitoring of BP in stroke survivors, offering a model for similar systems in clinical trials. Clinicians can analyze, track, and interpret monitored data in near real-time, remotely, which may enhance BP management approaches, particularly in communities with limited access to care. This approach has the potential to improve hypertension management for patients and clinicians when standard data transparency is insufficient, standard monitor features are incompatible with the user, or when integrating a commercial dashboard may threaten electronic health record security.

Trial Registration: ClinicalTrials.gov NCT05264298; https://clinicaltrials.gov/study/NCT05264298

(JMIR Form Res 2025;9:e62700) doi:10.2196/62700

KEYWORDS

electronic data capture; remote monitoring; telehealth; customization; web based; blood pressure; stroke; heart; development; video based; patient management

Introduction

Elevated systolic and diastolic blood pressure (BP) are associated with an increased risk for stroke [1]. An American Heart Association (AHA) Presidential Advisory Committee projected that hypertension prevalence will increase from 51.2% in 2020 to 61.0% in 2050, leading to higher rates of cardiovascular disease and stroke. The committee also projected an increase in total cardiovascular disease by 3.7% [2]. A separate report from AHA projected health care costs related to cardiovascular risk factors will increase from US $400 billion in 2020 to US $1344 billion in 2050 while accounting for inflation [3]. The AHA supports managing BP as one of 8 strategies for cardiovascular disease prevention [4]. The 2024 Guideline for the Primary Prevention of Stroke recommends screening for hypertension in all adults to identify those at increased risk for stroke [5].

Improved BP control depends on the health care team's ability to collect, interpret, and act on these findings collectively rather than making the data available to a single clinician. The clinical practice guidelines for cardiovascular disease prevention, management of coronary artery disease, primary and secondary prevention of stroke, management of stroke, and prevention and management of hypertension all explicitly recommend a health care team or team-based care approach [5-10]. In addition, the American College of Cardiology (ACC) Foundation has endorsed ambulatory (clinic-based) BP monitoring by various outpatient health care professionals to improve screening [11]. Pharmacists can play an important role in detecting medication-related causes of suboptimal BP control [12]. Compared with well-resourced patients, underresourced patients may have additional socioeconomic barriers to attending clinic visits, such as transportation, wait times, unknown reasons for visits, adverse clinic experiences, health status, and competing responsibilities [13,14]. Use of remote BP monitoring (RBPM), or monitoring patients using telecommunications (internet, telephone, etc) while patients are outside the health system, may be one method for improving collaboration among health care team members while minimizing the clinical burden on patients. Evidence supporting the use of RPBM for BP control is available, but there are still unresolved challenges with this approach [15].

BP varies in response to genetics and social stressors as well as environmental factors, such as monitoring location and environmental stress [16-18]. Studies that have shown benefits of clinical BP monitoring often do not describe in detail how the BP data were collected, received, cleaned, reported, and acted upon [19-22]. In addition, patients may have difficulty using home BP monitors. The monitors' properties and features aren't always chosen based on patient population or the individual patient's ability to operate the device but instead are chosen based on cost or familiarity. Integration of BP data into the electronic health record (EHR) faces challenges, such as workflow, data compatibility, data autonomy, and EHR security [15].

Dashboards have been developed to monitor BP in patients with dementia [23], renal transplant [24], hypertension [25-29], myocardial infarction [30], heart failure [31-35], diabetes mellitus [36-38], and COVID-19 [39,40], and in patients exposed to pollution [41]. While these examples of other dashboards monitor, track, and trend BP readings, none have been explicitly developed or described to address patient-, provider-, and system-specific challenges. There are multiple RBPM systems available on the market; however, opportunities for end-users to customize these according to patient and provider needs are limited. In addition, many of these systems are embedded in the institution's EHR rather than as a free-standing web-based application, which makes collaboration across disciplines and institutions difficult. Furthermore, detailed descriptions of the process by which the dashboards were developed and modified are not provided by most of the applications, thereby limiting opportunities for easy adaptation in new systems. The description and design of a dashboard programmed to monitor BP using a BP monitor specifically chosen for stroke survivors has not yet been described in the literature.

Modern technology can address integration challenges related to patient-, provider-, and health system-specific factors in BP management. Real-time remote BP monitoring systems should be selected with consideration of potential barriers. These solutions are accessible to health systems, researchers, and organizations with the financial resources to contract IT solution companies; however, in-house systems can also be developed to meet specific needs.

XSL•FO
RenderX

Patient-specific challenges to remote BP monitoring relate to self-efficacy, motivation, health literacy, digital literacy, functional limitations, wireless access, social support, and data protection [15-18]. Provider-specific factors involve ease of access, data summarization (date and time), data visualization, and identification of outlier values [15,19-22]. System-specific considerations include data aggregation, EHR integration, privacy considerations, data quality and maintenance, system updates, and usage tracking [15].

This paper aims to describe the development and design of a web-based dashboard for clinicians to monitor and manage BP while addressing patient-specific barriers, process transparency, provider challenges, and system constraints. This framework can serve as a blueprint for health systems, researchers, and clinicians seeking to develop in-house remote BP monitoring systems tailored to patient-, provider-, and system-specific needs.

Methods

Overview

The dashboard described in this text was developed for use in a clinical trial that incorporates remote BP monitoring and telehealth into the care of stroke survivors (NCT05264298). Briefly, the Video-based Intervention to Address Treatment and Outcome Disparities in Adults Living with transient ischemic attack (TIA) or Stroke (VIRTUAL) trial is a randomized comparative effectiveness study of a multidisciplinary telehealth intervention aimed at improving BP control following stroke. Participants are enrolled during their stroke or TIA hospitalization and randomized to the VIRTUAL intervention or standard care. Participants in the intervention arm receive an RBPM prior to hospital discharge and receive follow-up via telehealth with a stroke provider, pharmacist, and social worker. BP is monitored by a pharmacist who adjusts medications according to a standard protocol. Participants in the standard care arm receive a standard BP monitor with remote monitoring capabilities and receive follow-up care with a stroke provider. Study pharmacists relay patient-reported BP to the treating provider for patients in the standard arm but do not treat elevated BP directly, and thus the dashboard's primary use is for the intervention group. The intervention lasts for 6 months, and the primary outcome is 6-month 24-hour ambulatory BP.

The study team took the following recommended steps to customize the remote BP monitoring system to meet patient, provider, and system needs

Patient-Specific Needs

Many patients who have experienced stroke bear lasting physical and cognitive impairment [42,43]. For the VIRTUAL study, the selection of at-home BP monitors surveyed potential options and capabilities to address the needs of stroke survivors who often have physical and cognitive impairment after stroke [44]. Investigators considered application programming interface (API) push ability, transmission modality, ergonomics, reusability, cost, and customer service. The final BP monitor was selected based on these criteria. Patient use of the selected BP device was assessed by the RBPM usage rate (percentage) and the mean number of BP measurements per patient per day over the 180-day intervention period of the VIRTUAL study.

System-Specific Needs

The principal investigator (AS) contracted with Carematix, the digital health software solutions company (SSC) that distributes the selected monitor for back-end data acquisition. This device is connected to a cellular network and addresses barriers to remote monitoring associated with limited broadband access [44]. Using a cellular network avoids a potentially confusing and complicated internet setup and allows patients without wireless internet access to use the device [44]. Participant BP and heart rate data measured by these devices is transmitted via cellular connection to the SSC's servers. Using the SSC's API, the study's Data Coordinating Center (DCC) developed a web application to accept participant data from the SSC's servers via an internet connection. The SSC regularly transmits or "pushes" participant data to the DCC's servers via the API, where it is then stored in the DCC's database for use in a dashboard displaying patients' longitudinal BP and heart rate readings. Participant values typically appear in the DCC database and dashboard 2 - 5 minutes after readings are taken using the device.

The DCC modified an existing, locally developed clinical trial management system (CTMS) called the Clinical Electronic Data Access and Management System (CEDAMS) in close collaboration with clinicians to make the VIRTUAL system. Figure 1 describes the data flow. CEDAMS includes CTMS functionality, electronic Case Report Forms (eCRF), and a dashboard.

XSL•FO
RenderX

Figure 1. Data map for VIRTUAL Electronic Data Capture System. BP: blood pressure; CEDAMS: Clinical Electronic Data Access and Management System.

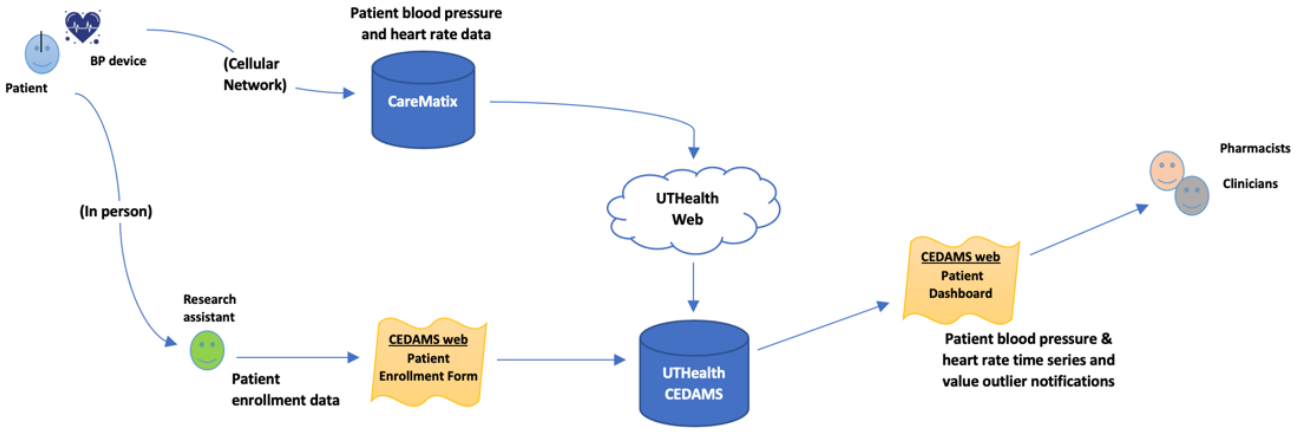

In CEDAMS, the participant ID and the device serial number are used to link the data from the device to the patient's profile in the database. The CTMS manages user accounts and study site information, as well as control over user access permissions, eCRF configuration, and usage monitoring. CEDAMS retrieves patient data from the institution's database and displays each patient's BP readings as a time series along with other variables on the dashboard screen. The implementation of CEDAMS was customized for the VIRTUAL trial. System-specific needs were addressed in the design. Alert frequencies, time from alert to notification, and time from data collection to dashboard entry (BP data transmission times) were analyzed as measures of system efficiency.

Provider-Specific Needs

The development and testing phases followed the DCC's established protocols for system development. This includes the DCC's standard operating procedures regarding source code control, nightly system builds, and testing. Separate web servers and databases were used for system development, testing, and production. Testing was performed first by the development staff, then the data coordinating staff, and then the users. The clinician users provided casual feedback regarding the API during this phase, and the feedback was incorporated as appropriate.

The CEDAMS electronic data capture (EDC) system was developed over approximately 3 months, including front-end development, front-end testing, and API development. The data gathered before the API's completion were then integrated into the completed EDC system at the end of May 2022. Informal user training was provided at clinical research meetings, and a detailed user manual was issued at the launch of the project for clinical users. Provider usage was measured using the number of monthly logins per clinician.

Ethical Considerations

The VIRTUAL trial was approved by the University of Texas Health Science Center at Houston Committee for the Protection of Human Subjects (study HSC-MS-21 - 0549) on August 5, 2021. Patients were provided informed consent, and confidentiality was achieved through the use of generated patient identifiers. Patients were provided compensation in cash per the protocol and were loaned the devices necessary to take and upload readings to the investigators.

Results

Patient-Specific Needs

Seventeen BP monitors were considered from 6 different manufacturers and were evaluated based on API push ability, transmission modality, ergonomics, reusability, cost, and customer service. These monitors were further screened for barriers to integrating their data into the EHR data, such as (1) compatibility and synchronicity of data with the EHR; (2) API programming challenges; and (3) the burden on patients to execute the data transfer. Of the 17 monitors, 3 were excluded for lack of API-push ability. Five additional monitors were excluded due to poor ergonomic design. Integration barriers removed another 3 monitors from consideration. The remaining 5 monitors were considered for use in the study, and the Blipcare Blip BP800 distributed by Carematix was determined to be the most suitable for VIRTUAL's needs based on the previously mentioned criteria. Among participants in the intervention arm, there was 97% engagement with a mean number of 1.3 BP measures per day.

System-Specific Needs

Figure 1 depicts the VIRTUAL data capture system. The system was designed to operate independently of the EHR to address data security concerns while allowing for potential EHR integration in the future. The system was designed to capture delays in data transmission, identify outliers, and track BP monitoring adherence, enabling effective data quality monitoring. Delays in BP data transmission occurred when the device was turned off before the data could be sent, with transmission resuming the next time the device was powered on. Time from data collection (BP measurement) to transmission to the dashboard database (via the BP device online server) was a median of 1.4 (IQR 1.0-5.8) minutes. Based on input from clinical providers, BP and heart rate outlier alerts were generated each morning, capturing all outlier readings from the previous day. The average time from a unique alert to provider email notification was 17.9 (range 6.4-30.3) hours. Additional details on alert frequency are provided in Table 1.

XSL•FO
RenderX

Table . Summary statistics of number of alerts in patients in the intervention group who have been followed up for 6 months (N=190).

Alert type	Number of patients, n (%)	Minimum number of alerts	Maximum number of alerts	Mean (SD)	Median (IQR)
Overall	131 (68.9)	0	109	5.9 (14.4)	2 (0-5.8)
Abnormal systolic BP[a] (<91 or >179)	82 (43.2)	0	45	1.7 (4.6)	0 (0-2)
High systolic BP (>179)	32 (16.8)	0	44	0.8 (3.6)	0 (0-0)
Low systolic BP (<91)	60 (31.6)	0	24	1 (2.9)	0 (0-1)
Abnormal diastolic BP (<41 or >119)	53 (27.9)	0	80	1.4 (6.6)	0 (0-1)
High diastolic BP (>119)	47 (24.7)	0	80	1.4 (6.6)	0 (0-0)
Low diastolic BP (<41)	7 (3.7)	0	3	0 (0.3)	0 (0-0)
Abnormal heart rate (<50)	59 (31.1)	0	106	2.8 (11.4)	0 (0-1)

[a]BP: blood pressure.

Provider-Specific Needs

The VIRTUAL system was designed with a login page, a home page, a face sheet, and a BP dashboard screen. The login page is the initial page visible to the user with username and password textboxes and is shown in Figure 2. There is a security feature required upon first login to create 3 security questions from a preselected dropdown and a subsequent screen to change the password from the temporary password provided by the DCC. Once logged in, the user is directed to the Home Page which has 2 icons as seen in Multimedia Appendix 1—one for the face sheet and one for the dashboard screen.

Clinical providers provided feedback regarding the need to easily toggle between patient demographic data and BP data and the ability to modify the time window for BP data visualization. The face sheet participant screen displays a list of participants already enrolled and allows the user to add a new participant, select a participant, search participants, and view or sort participants based on participant ID, acrostic participant code, enrollment site, date entered, completion status, assigned clinician, and eCRF reference number Multimedia Appendix 2. The dashboard screen has a dropdown button to select a patient, or the patient's dashboard screen panel can be accessed from their face sheet. The dashboard screen, as seen in Multimedia Appendix 2, displays systolic BP, diastolic BP, and heart rate during enrollment as well as a 14-day moving average. Each parameter can be selected separately, which removes the other parameters from the chart. The dashboard screen also shows averaged day (08:00-19:59 h), night (<08:00 or >19:59 h), and combined systolic BP, diastolic BP, and heart rates. This component of the dashboard screen can also select day, night, or combined parameters from the chart which will remove the others.

Other CEDAMS VIRTUAL design features include real-time visual cues and email notifications for out-of-range systolic BP, diastolic BP, and heart rate values. Email notifications are sent to appropriate study staff for missing readings when a patient's device hasn't sent any data for over 5 days. Also, the system sends reminders to study staff to follow up with patients at specified times after enrollment. The version of the CEDAMS that supports the VIRTUAL trial also supports outbound API function calling to interface with other clinical study support systems such as REDCap® (Research Electronic Data Capture; Vanderbilt University). Having separate systems allowed for the storing of protected health information in CEDAMS and deidentified data in the trial database.

The EDC system successfully performed the intended RBPM functions. The patient-, provider-, and system-specific challenges from traditional monitoring procedures were addressed by the EDC system.

The system was designed to aggregate and summarize data for clinical use. It provides daytime, nighttime, and average readings in both graphical and tabular formats. Pharmacists can adjust the data capture window based on specific needs. Providers receive alerts for out-of-range data, which may prompt changes in medications. Patient and caregiver contact information is securely linked to the capture system to facilitate care. Clinicians effectively used the EDC system to monitor BP. Mean monthly CTMS logins for 4 clinicians were as follows: PharmD1=19, PharmD2=22, PharmD3=29, and PharmD4=44. Alerts for out-of-range BP or heart rate were triggered for 69% (n/N) of patients (Table 1), with most alerts due to abnormal systolic BP (43%), particularly low systolic BP (32%). The distribution of alerts was right-skewed, with an average of 6 alerts per enrolled person and a median of 2 (IQR 2-5.8) alerts per person.

XSL•FO
RenderX

Figure 2. Login page and homepage screens.

Discussion

Principal Findings

Despite the known long-term detrimental effects of uncontrolled BP on cardiovascular and all-cause mortality, more than 50% of US adults, including survivors of stroke, have uncontrolled BP. RBPM has been proven to be effective for BP reduction and is established as a practice for reducing cardiovascular risk and improving team-based BP control in the 2017 ACC/AHA Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults [10]. However, RBPM has not been routinely implemented into the care of patients, including those at high risk for cardiovascular disease. This is partly due to the lack of published methods for integrating RBPM into clinical practice while addressing patient, provider, and system-level needs and barriers. Herein, we describe steps to developing an RBPM EDC while considering our population's specific needs.

The monitor was chosen for ease of use in a population of patients with stroke. Among enrolled participants, rates of participation are high, with 97% engagement. The system was also built to address provider needs while overcoming system challenges. The EDC system was successfully developed and implemented within three months to function alongside the EHR without requiring full integration. Clinician feedback shaped features, such as alert timing, customizable visualization windows, and streamlined access to patient data. Clinicians consistently logged in without issues and accessed patients' readings for monitoring. Patient data were securely protected while remaining accessible to providers. The system supports data monitoring for quality assurance purposes.

Limitations

This study has several limitations related to its design and measures. We described steps used within a single, large health care system with access to expertise in IT, bioinformatics, health systems research, and clinical trial design. The framework and steps we provide may be more challenging to implement in settings with limited resources. We were limited in some of the data we reported because of the ongoing nature of the clinical trial. The data we can report serve as proxies for system fitness and do not comprehensively address all potential usage barriers faced by patients and providers or integration challenges encountered by the system. Nonetheless, combining these parameters with the context in which they were gathered may offer valuable insights for extrapolation and future exploration.

XSL•FO
RenderX

There were also some data collection limitations of the VIRTUAL EDC system. When taking family members' BP or when clinic staff demonstrate how to use the BP monitor, the reading is logged as the patient's BP in the interface. This challenge was mitigated with clinician and patient retraining; however, this can have implications for the duration of clinic visits. We observed delays in transmission from the BP device to the dashboard due to the device being powered off by the user too soon after a reading. While the overall median time for entry into the EDC was minimal, there were some readings that were delayed as a result. For subsequent enrolled patients, we trained them to leave the monitor on for a few minutes after the completion of the reading to complete data transmission.

Future Directions

Future work in this area could include more human-centered design features to more directly refine the system based on patient, caregiver, and clinician experiences, and test outcomes could include those related to ergonomics and other factors impacting usability. BP monitoring systems can be more specifically tailored to populations and end-users based on different health conditions, environmental factors, and accessibility concerns. Finally, it can be further refined to incorporate clinical decision-making tools, including artificial intelligence to assist with pharmacotherapy recommendations and identify other clinical and social factors related to monitoring adherence and BP control.

Conclusions

The VIRTUAL system was developed for remote BP monitoring in a population at high risk for cardiovascular disease. Clinicians can analyze, track, and interpret monitored data in near real-time, remotely, which may enhance BP management approaches, particularly in communities with limited access to care. The successful development of this system and its capabilities could serve as a blueprint for creating similar systems for clinical BP monitoring when standard data transparency is insufficient, standard monitor features are incompatible with the user, or when integrating a commercial dashboard may threaten EHR security.

Acknowledgments

The authors would like to acknowledge the patients enrolled in the VIRTUAL trial for their participation. In addition, the institution's information technology team and the clinician users who developed and gave feedback throughout the process were invaluable to this work. This work was supported by the National Institutes of Health [R01 MD016465-04 and R01 MD016465-04S1].

Data Availability

The datasets generated or analyzed during this study are not publicly available due to ongoing study for the associated clinical trial but are available from the corresponding author on reasonable request upon completion of the trial.

Authors' Contributions

JW contributed to writing the original draft. CC, TN, JW, and JMY were responsible for visualization. AS, JMY, EB, and SS were involved in supervision, conceptualization, project administration, and funding acquisition. CC, GD, MG, MW, and JMY developed and maintained the software. AS, SS, JMY, EB, CG, TN, MO, CI, NH, AC, and GC carried out the investigation.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Face sheet screen and patient information screen.
[PNG File, 581 KB - formative_v9i1e62700_app1.png]

Multimedia Appendix 2
Dashboard screen with data selection function.
[PNG File, 1195 KB - formative_v9i1e62700_app2.png]

References

1. Rost NS, Kasner SE. Overview of secondary prevention of ischemic stroke. UpToDate. 2023 Jan 27. URL: https://www.uptodate.com/contents/overview-of-secondary-prevention-of-ischemic-stroke [accessed 2025-07-16]
2. Joynt Maddox KE, Elkind MSV, Aparicio HJ, et al. Forecasting the burden of cardiovascular disease and stroke in the United States through 2050-prevalence of risk factors and disease: a presidential advisory from the American Heart Association. Circulation 2024 Jul 23;150(4):e65-e88. [doi: 10.1161/CIR.0000000000001256] [Medline: 38832505]

XSL•FO
RenderX

3. Kazi DS, Elkind MSV, Deutsch A, et al. Forecasting the economic burden of cardiovascular disease and stroke in the United States through 2050: a presidential advisory from the American Heart Association. Circulation 2024 Jul 23;150(4):e89-e101. [doi: 10.1161/CIR.0000000000001258] [Medline: 38832515]
4. Lloyd-Jones DM, Allen NB, Anderson CAM, et al. Life's Essential 8: updating and enhancing the American Heart Association's construct of cardiovascular health: a presidential advisory from the American Heart Association. Circulation 2022 Aug 2;146(5):e18-e43. [doi: 10.1161/CIR.0000000000001078] [Medline: 35766027]
5. Bushnell C, Kernan WN, Sharrief AZ, et al. 2024 Guideline for the primary prevention of stroke: a guideline from the American Heart Association/American Stroke Association. Stroke 2024 Dec;55(12):e344-e424. [doi: 10.1161/STR.0000000000000475] [Medline: 39429201]
6. Arnett DK, Blumenthal RS, Albert MA, et al. 2019 ACC/AHA guideline on the primary prevention of cardiovascular disease: a report of the American College of Cardiology/American Heart Association task force on clinical practice guidelines. Circulation 2019 Sep 10;140(11):e596-e646. [doi: 10.1161/CIR.0000000000000678] [Medline: 30879355]
7. Virani SS, Newby LK, Arnold SV, et al. 2023 AHA/ACC/ACCP/ASPC/NLA/PCNA guideline for the management of patients with chronic coronary disease: a report of the American Heart Association/American College of Cardiology joint committee on clinical practice guidelines. Circulation 2023 Aug 29;148(9):e9-e119. [doi: 10.1161/CIR.0000000000001168] [Medline: 37471501]
8. Kleindorfer DO, Towfighi A, Chaturvedi S, et al. 2021 Guideline for the prevention of stroke in patients with stroke and transient ischemic attack: a guideline from the American Heart Association/American Stroke Association. Stroke 2021 Jul;52(7):e364-e467. [doi: 10.1161/STR.0000000000000375] [Medline: 34024117]
9. Powers WJ, Rabinstein AA, Ackerson T, et al. Guidelines for the early management of patients with acute ischemic stroke: 2019 update to the 2018 guidelines for the early management of acute ischemic stroke: a guideline for healthcare professionals from the American Heart Association/American Stroke Association. Stroke 2019 Dec;50(12):e344-e418. [doi: 10.1161/STR.0000000000000211] [Medline: 31662037]
10. Whelton PK, Carey RM, Aronow WS, et al. 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA guideline for the prevention, detection, evaluation, and management of high blood pressure in adults: a report of the American College of Cardiology/American Heart Association task force on clinical practice guidelines. Hypertension 2018 Jun;71(6). [doi: 10.1161/HYP.0000000000000065]
11. Dunn SP, Birtcher KK, Beavers CJ, et al. The role of the clinical pharmacist in the care of patients with cardiovascular disease. J Am Coll Cardiol 2015 Nov 10;66(19):2129-2139. [doi: 10.1016/j.jacc.2015.09.025] [Medline: 26541925]
12. Khazan E, Anastasia E, Hough A, Parra D. Pharmacist-managed ambulatory blood pressure monitoring service. Am J Health Syst Pharm 2017 Feb 15;74(4):190-195. [doi: 10.2146/ajhp160113] [Medline: 28179243]
13. Crosby LE, Modi AC, Lemanek KL, Guilfoyle SM, Kalinyak KA, Mitchell MJ. Perceived barriers to clinic appointments for adolescents with sickle cell disease. J Pediatr Hematol Oncol 2009 Aug;31(8):571-576. [doi: 10.1097/MPH.0b013e3181acd889] [Medline: 19636266]
14. Pesata V, Pallija G, Webb AA. A descriptive study of missed appointments: families' perceptions of barriers to care. J Pediatr Health Care 1999;13(4):178-182. [doi: 10.1016/S0891-5245(99)90037-8] [Medline: 10690082]
15. Lu Z. Clinical data management: current status, challenges, and future directions from industry perspectives. Open Access J Clin Trials 2010:93-105. [doi: 10.2147/OAJCT.S8172]
16. Arora P, Newton-Cheh C. Blood pressure and human genetic variation in the general population. Curr Opin Cardiol 2010 May;25(3):229-237. [doi: 10.1097/HCO.0b013e3283383e2c] [Medline: 20224392]
17. Spruill TM. Chronic psychosocial stress and hypertension. Curr Hypertens Rep 2010 Feb;12(1):10-16. [doi: 10.1007/s11906-009-0084-8] [Medline: 20425153]
18. Franklin SS, Thijs L, Hansen TW, O'Brien E, Staessen JA. White-coat hypertension: new insights from recent studies. Hypertension 2013 Dec;62(6):982-987. [doi: 10.1161/HYPERTENSIONAHA.113.01275] [Medline: 24041952]
19. Kitagawa K, Yamamoto Y, Arima H, et al. Effect of standard vs intensive blood pressure control on the risk of recurrent stroke. a randomized clinical trial and meta-analysis. JAMA Neurol 2019 Nov 1;76(11):1309-1318. [doi: 10.1001/jamaneurol.2019.2167] [Medline: 31355878]
20. Mant J, McManus RJ, Roalfe A, et al. Different systolic blood pressure targets for people with history of stroke or transient ischaemic attack: PAST-BP (Prevention After Stroke--Blood Pressure) randomised controlled trial. BMJ 2016 Feb 24;352:i708. [doi: 10.1136/bmj.i708] [Medline: 26919870]
21. Benavente OR, Coffey CS, Conwit R. Blood-pressure targets in patients with recent lacunar stroke: the SPS3 randomised trial. The Lancet 2013 Aug;382(9891):507-515. [doi: 10.1016/S0140-6736(13)60852-1]
22. Bath PM, Scutt P, Blackburn DJ, et al. Intensive versus guideline blood pressure and lipid lowering in patients with previous stroke: main results from the pilot "prevention of decline in cognition after stroke trial" (PODCAST) randomised controlled trial. PLoS ONE 2017;12(1):e0164608. [doi: 10.1371/journal.pone.0164608] [Medline: 28095412]
23. Cassarino N, Bergstrom B, Johannes C, Gualtieri L. Monitoring older adult blood pressure trends at home as a proxy for brain health. Online J Public Health Inform 2021;13(3):e16. [doi: 10.5210/ojphi.v13i3.11842] [Medline: 35079322]

XSL•FO
RenderX

24. Aberger EW, Migliozzi D, Follick MJ, Malick T, Ahern DK. Enhancing patient engagement and blood pressure management for renal transplant recipients via home electronic monitoring and web-enabled collaborative care. Telemed J E Health 2014 Sep;20(9):850-854. [doi: 10.1089/tmj.2013.0317] [Medline: 25046403]
25. Al-Khatib A, Andreski M, Pudlo A, Doucette WR. An evaluation of community pharmacies' actions under value-based payment. J Am Pharm Assoc (2003) 2020;60(6):899-905. [doi: 10.1016/j.japh.2020.06.014] [Medline: 32819876]
26. Hartkopf KJ, Heimerl KM, McGowan KM, Arndt BG. Expansion and evaluation of pharmacy services in primary care. Pharmacy (Basel) 2020 Jul 22;8(3):124. [doi: 10.3390/pharmacy8030124] [Medline: 32707794]
27. De Marchi ACB, Alves A, Gonçalves CBC, et al. An electronic health platform for monitoring health conditions of patients with hypertension in the Brazilian public health system: protocol for a nonrandomized controlled trial. JMIR Res Protoc 2020 Jan 20;9(1):e15299. [doi: 10.2196/15299] [Medline: 31958068]
28. Koopman RJ, Canfield SM, Belden JL, et al. Home blood pressure data visualization for the management of hypertension: designing for patient and physician information needs. BMC Med Inform Decis Mak 2020 Aug 18;20(1):195. [doi: 10.1186/s12911-020-01194-y] [Medline: 32811489]
29. Tseng TW, Su CF, Lai F. Fast Healthcare Interoperability Resources for inpatient deterioration detection with time-Series vital signs: design and implementation study. JMIR Med Inform 2022 Oct 13;10(10):e42429. [doi: 10.2196/42429] [Medline: 36227636]
30. Jameie S, Haybar H, Aslani A, Saadat M. Development and usability evaluation of web-based telerehabilitation platform for patients after myocardial infarction. Stud Health Technol Inform 2019;261:68-74. [Medline: 31156093]
31. Park C, Otobo E, Ullman J, et al. Impact on readmission reduction among heart failure patients using digital health monitoring: feasibility and adoptability study. JMIR Med Inform 2019 Nov 15;7(4):e13353. [doi: 10.2196/13353] [Medline: 31730039]
32. Alnosayan N, Chatterjee S, Alluhaidan A, Lee E, Houston Feenstra L. Design and usability of a heart failure mHealth system: a pilot study. JMIR Hum Factors 2017 Mar 24;4(1):e9. [doi: 10.2196/humanfactors.6481] [Medline: 28341615]
33. Holthe H, Serrano JA. ePoint.telemed–an open web-based platform for home monitoring of patients with chronic heart failure. Stud Health Technol Inform 2015;216:74-78. [Medline: 26262013]
34. Serrano JA, Holthe H. Development and trial of ePoint.telemed - An open web-based platform for home monitoring of chronic heart failure patients. Stud Health Technol Inform 2015;210:311-315. [Medline: 25991156]
35. Safdari R, Jafarpour M, Mokhtaran M, Naderi N. Designing and implementation of a heart failure telemonitoring system. Acta Inform Med 2017 Sep;25(3):156-162. [doi: 10.5455/aim.2017.25.156-162] [Medline: 29114106]
36. Ayub A, Ng C, Portales-Casamar E, Metzger D, Amed S. Toward building a provincial diabetes registry of children and youth living with diabetes in British Columbia, Canada. Can J Diabetes 2022 Jun;46(4):346-352. [doi: 10.1016/j.jcjd.2021.11.006] [Medline: 35527205]
37. Hull S, Chowdhury TA, Mathur R, Robson J. Improving outcomes for patients with type 2 diabetes using general practice networks: a quality improvement project in east London. BMJ Qual Saf 2014 Feb;23(2):171-176. [doi: 10.1136/bmjqs-2013-002008] [Medline: 24003237]
38. Dixon BE, Alzeer AH, Phillips EO, Marrero DG. Integration of provider, pharmacy, and patient-reported data to improve medication adherence for type 2 diabetes: a controlled before-after pilot study. JMIR Med Inform 2016 Feb 8;4(1):e4. [doi: 10.2196/medinform.4739] [Medline: 26858218]
39. Wong CK, Ho DTY, Tam AR, et al. Artificial intelligence mobile health platform for early detection of COVID-19 in quarantine subjects using a wearable biosensor: protocol for a randomised controlled trial. BMJ Open 2020 Jul 22;10(7):e038555. [doi: 10.1136/bmjopen-2020-038555] [Medline: 32699167]
40. Fisher L, Curtis HJ, Croker R, et al. Eleven key measures for monitoring general practice clinical activity during COVID-19: A retrospective cohort study using 48 million adults' primary care records in England through OpenSAFELY. Elife 2023 Jul 27;12:e84673. [doi: 10.7554/eLife.84673] [Medline: 37498081]
41. Zhao JH, Hu LX, Xiao S, et al. Screening and prioritization of organic chemicals in a large river basin by suspect and non-target analysis. Environ Pollut 2023 Sep 15;333:122098. [doi: 10.1016/j.envpol.2023.122098] [Medline: 37352960]
42. Gittins M, Lugo-Palacios D, Vail A, et al. Stroke impairment categories: a new way to classify the effects of stroke based on stroke-related impairments. Clin Rehabil 2021 Mar;35(3):446-458. [doi: 10.1177/0269215520966473] [Medline: 33131321]
43. El Husseini N, Katzan IL, Rost NS, et al. Cognitive impairment after ischemic and hemorrhagic stroke: a scientific statement from the American Heart Association/American Stroke Association. Stroke 2023 Jun;54(6):e272-e291. [doi: 10.1161/STR.0000000000000430] [Medline: 37125534]
44. Sharrief AZ, Guzik AK, Jones E, et al. Telehealth trials to address health equity in stroke survivors. Stroke 2023 Feb;54(2):396-406. [doi: 10.1161/STROKEAHA.122.039566] [Medline: 36689591]

Abbreviations

ACC: American College of Cardiology
AHA: American Heart Association
API: application programming interface

XSL•FO
RenderX

BP: blood pressure
CEDAMS: Clinical Electronic Data Access and Management System
CTMS: clinical trial management system
DCC: data coordinating center
eCRF: electronic case report form
EDC: electronic data capture
EHR: electronic health record
RBPM: remote blood pressure monitoring
REDCap: Research Electronic Data Capture
SSC: software solutions company
TIA: transient ischemic attack
VIRTUAL: Video-based Intervention to Reduce Treatment and Outcome Disparities in Adults Living with Stroke or Transient Ischemic Attack

Edited by A Coristine; submitted 03.06.24; peer-reviewed by IJ McGurgan, K Fakes; revised version received 28.04.25; accepted 05.05.25; published 06.08.25.

Please cite as:
Wollen J, Nguyen T, Coton C, DeWildt G, Cooksey G, Cohen A, Hoang N, Okpala M, Gonzalez M, Wang M, Izeogu C, Bernstam E, Green C, Savitz S, Yamal JM, Sharrief A
Development of a Customizable Web-Based Dashboard for Remote Blood Pressure Monitoring: Feasibility and Usability Implementation Study
JMIR Form Res 2025;9:e62700
URL: https://formative.jmir.org/2025/1/e62700
doi:10.2196/62700

© Joshua Wollen, Thuy Nguyen, Charles Coton, Gina DeWildt, Gabretta Cooksey, Audrey Cohen, Nicholas Hoang, Munachi Okpala, Michael Gonzalez, Mengxi Wang, Chigozirim Izeogu, Elmer Bernstam, Charles Green, Sean Savitz, Jose-Miguel Yamal, Anjail Sharrief. Originally published in JMIR Formative Research (https://formative.jmir.org), 6.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Factors Associated With Portal and Telehealth Uptake and Use in a Minoritized, Low-Income Community: Mixed Methods Study

Robin T Higashi[1,2], PhD; Emily C Repasky[1], MA; Antara Gupta[3], BSA; MinJae Lee[2,4], PhD; Catherine M DesRoches[5], DrPH; Aimee Israel[2], MSW; Sandi L Pruitt[1,2], PhD

[1]Department of Social and Behavioral Sciences, Peter O'Donnell Jr School of Public Health, The University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd, Dallas, TX, United States

[2]Harold C Simmons Comprehensive Cancer Center, Dallas, TX, United States

[3]The University of Texas Southwestern Medical School, Dallas, TX, United States

[4]Department of Health Data Science & Biostatistics, O'Donnell School of Public Health, The University of Texas Southwestern Medical Center, Dallas, TX, United States

[5]Department of Medicine, Harvard Medical School, Harvard University, Boston, MA, United States

Corresponding Author:
Robin T Higashi, PhD
Department of Social and Behavioral Sciences, Peter O'Donnell Jr School of Public Health, The University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd, Dallas, TX, United States

Abstract

Background: Despite evidence that use of patient portals and telehealth is associated with many health benefits, disparities exist in awareness, adoption, and use. Understanding factors and strategies specific to underserved populations is key to achieving digital equity and better health.

Objective: This study assesses portal and telehealth experiences among residents of a minoritized and lower-resource area of Dallas, Texas.

Methods: Using an explanatory sequential design, we conducted surveys and semistructured interviews with English- and Spanish-speaking adults in 15 ZIP Codes surrounding a community-based clinic. We recruited participants via a patient portal, flyers, emails distributed by clinic and community partners, and in person. Surveys were offered online and on paper. We used Fisher exact tests to identify factors associated with telehealth and/or portal use. We also recruited a subsample of survey participants to expound on survey findings in semistructured interviews. Our thematic analysis assessed convergence in survey and interview findings.

Results: Among 182 survey respondents, most were older (n=109, 66%; age ≥60 years), African American or Black (n=120, 65%), and female (n=142, 79%); a little more than half (n=97, 54%) had completed ≥1 telehealth appointment, and a majority (n=131, 72%) had used a patient portal at least once. Compared with those who used the portal and/or telehealth, those reporting no use of portal or telehealth were more likely to have a high school education or less ($P<.001$) or be Spanish speakers ($P<.011$). A majority, regardless of portal or telehealth use, agreed with health promotion activity survey statements like “Using the Internet for health-related activities makes me feel actively involved with my health care” (n=103, 59%) and “I find the Internet useful for monitoring my health” (n=100, 58%). In interviews with 20 individuals, most of whom were older, Black, female, and had digital technology experience, seven factors were key to increased engagement in portals and telehealth: (1) improving patient autonomy, (2) integrating digital health technology into daily life, (3) receiving recommendations from trusted individuals, (4) appreciating the value of digital health technologies, (5) enlisting the support of care partners or peers, (6) managing severe or chronic illness, and (7) accessing test results rapidly.

Conclusions: This study builds on previous work by confirming and contributing insights about factors key to technology uptake and use among underserved populations. Interventions using digital health technologies should focus on these factors to promote digital and health equity and achieve better health outcomes. Future research should explore which clinical settings and contexts are most conducive to increasing digital technology uptake and use, and implementation should leverage partnerships with community groups.

(JMIR Form Res 2025;9:e70146) doi:10.2196/70146

KEYWORDS

digital health technology; digital divide; digital inclusion; social determinants of health; health disparities; technology access; technology uptake; 21st Century Cures Act

XSL•FO
RenderX

Introduction

The use of patient portals and synchronous video visits, or telehealth, has become an integral part of health care, particularly since the beginning of the COVID-19 pandemic [1-4]. Portals and telehealth have remained valuable tools for care delivery due to their many documented benefits, including convenience (reduced time, money, and burden, especially for patients living in rural areas), reduced transportation barriers, and inclusion of family members from remote areas [5,6]. Furthermore, telehealth and patient portal use are associated with increased health knowledge, patient satisfaction, and preventive services use [7-10].

Portal and telehealth use are especially valuable for people with chronic or acute illnesses such as cancer, given their high need for medical appointments, services, medications, and numerous communications with the health care system. When individuals are juggling information from multiple sources over a prolonged period, portals and telehealth serve as vital tools to organize, track, and make information and communication accessible from home or anywhere through mobile technologies. Telehealth is also especially beneficial for people with cancer, HIV, or other conditions that should avoid exposure to infection due to their compromised immune system, and because these individuals may be too physically frail or ill to attend clinic appointments [11,12].

At the same time, digital health disparities and inequities in using and accessing virtual health information remain a significant problem [13]. The digital divide has been mapped by Gallardo [14], who demonstrated using a Digital Divide Index that disparities in technological advancements parallel geographic distributions of poorer health outcomes. When populations with lower digital literacy access and skills are also the populations experiencing poorer health outcomes, these problems can be mutually exacerbating [15-17].

Low-income, minoritized populations face unique barriers to the adoption and use of telehealth and patient portals. For example, low income and low levels of education are negatively associated with uptake of digital health interventions due to low level of English literacy [18]. Additionally, language and cultural barriers present challenges to uptake of digital health interventions among minoritized and non–English-speaking populations because resources are often not tailored to fit their specific needs [19]. These challenges increase the importance of engaging local community partners to enhance the design, development, and implementation of digital health interventions [19-21]. Several studies have demonstrated the importance of user-centered design as a key strategy to addressing digital inclusion [19,22]. Partnerships with local community groups can also facilitate outreach and recruitment to specific user groups by promoting trust in the community [23]. O'Connor et al [24] in their 2016 meta-analysis of 19 qualitative studies demonstrated that telehealth utilization in underserved communities is most strongly affected by 4 key factors: improving patient autonomy, seamless integration into daily life, personal recommendations from trusted individuals, and perceived quality of the intervention.

The goal of this study was to explore attitudes toward and experiences with telehealth and the portal among a sample of residents from an underserved community in Dallas, Texas, and to explore factors associated with the use of these health technologies. We compare our findings with the key factors of O'Connor et al [24] and offer 3 additional factors that may be relevant to the development of strategies to address digital inclusion in future interventions with low-resource populations.

Methods

Study Design

This study used an explanatory sequential mixed methods design [25] consisting of surveys and semistructured interviews with a geographically defined sample recruited from both community and clinic settings. This study adhered to the SRQR (Standards for Reporting Qualitative Research) guidelines (the SRQR checklist is provided in Checklist 1).

Setting

This study was conducted with English- and Spanish-speaking individuals (age ≥18 years) and residence within the RedBird neighborhood of Dallas. We defined the "RedBird area" as comprising 15 ZIP Codes surrounding a community-based clinic affiliated with UT Southwestern Medical Center and Simmons Comprehensive Cancer Center, hereafter referred to as "RedBird Clinic." The RedBird area in the Southern Sector of Dallas is known to be adversely affected by the digital divide, illustrated in Figure 1 [14]. Gallardo [14] calculated the Digital Divide Index by adding a composite socioeconomic score (age, poverty, lower than high school education, disability status, and internet-to-income ratio) with an infrastructure and adoption score (broadband access, computer devices, and download and/or upload speed).

XSL•FO
RenderX

Figure 1. Map of Dallas County including the study recruitment area surrounding RedBird Clinic and Digital Divide Index in tertiles based on the distribution across the United States. RedBird Clinic is the community-based clinic affiliated with UT Southwestern Medical Center and the geographic center around which we based recruitment for this study. Digital Divide Index refers to a multicomponent scoring system developed by Gallardo [14] that combines socioeconomic and infrastructure elements, and "low access" reflects lowest tertile of digital access within this score.

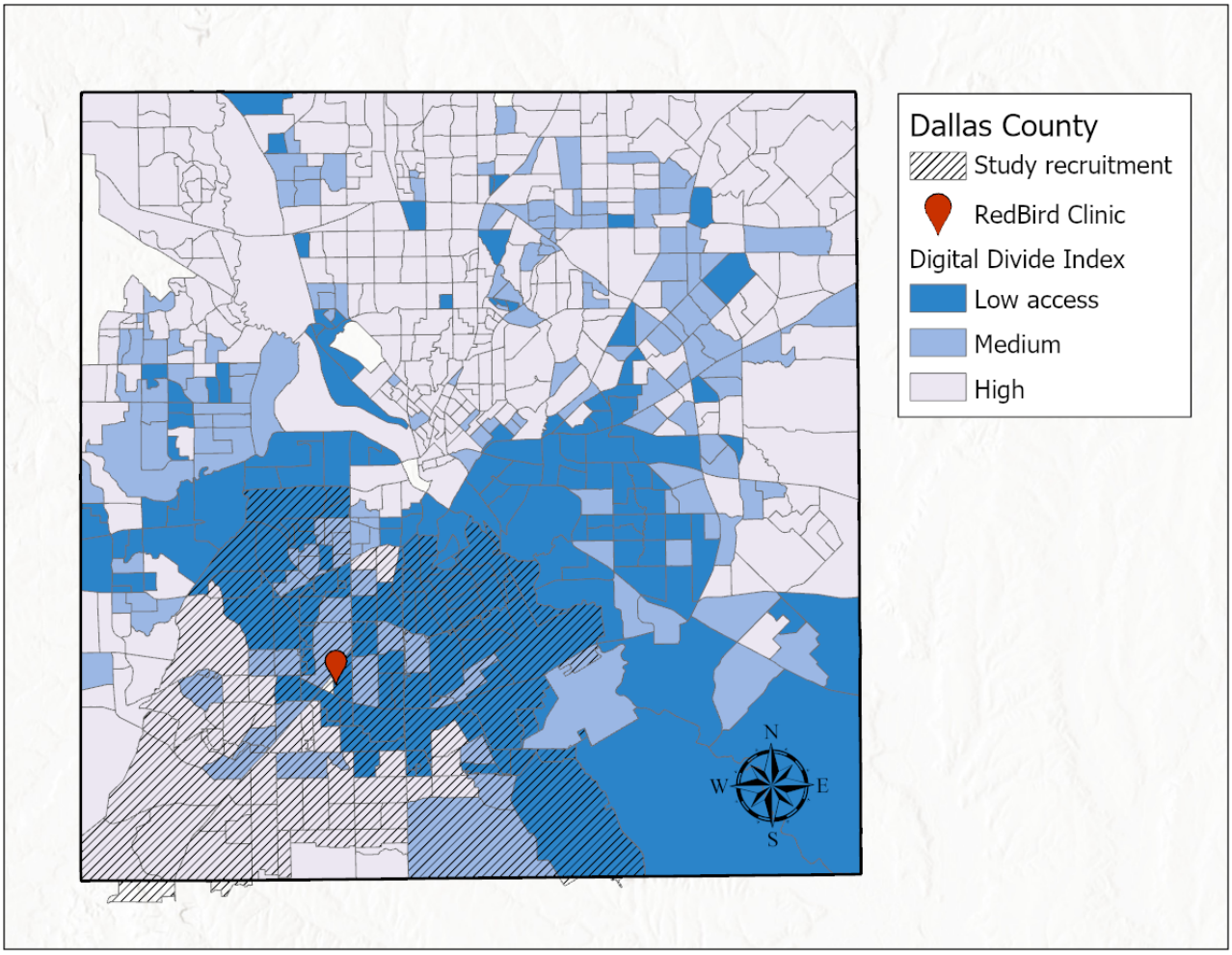

Survey Sampling and Recruitment

We recruited survey participants through the RedBird Clinic, the Simmons Comprehensive Cancer Center (hereafter "Cancer Center"), and through local community partners including faith-based organizations, nonprofit organizations, and support groups. All organization partners were located in the RedBird neighborhood from which we recruited study participants, and all were dedicated to advancing the social, economic, and/or health needs of RedBird area residents. Cancer Center participants were sampled through the patient registry, limited by those with residential addresses within the 15 Zip Code area and having completed an appointment within the past 5 years.

To capture data from participants of varying levels of digital literacy, we deployed multiple survey recruitment and data collection methods in accordance with each community partner's preference. Ultimately, most surveys were completed either on paper or through REDCap (Research Electronic Data Capture) [26], a web-based data collection and management tool. Recruitment strategies included mailing paper surveys, sending a link for distribution in the organization's newsletter, distributing flyers with a QR code and a phone number, and in-person recruitment. Patients of the Cancer Center who had portal accounts received a survey invitation link through the portal, while nonportal account holders received a postcard with multiple options for survey administration. In all formats, the first question of the survey explained the voluntary nature of participation, and that submission of survey responses indicated consent to participate.

Survey Data Collection and Analysis

The survey included questions about technology access, health literacy, information seeking and literacy, and portal and telehealth exposure. Descriptive statistics were used to summarize participant characteristics and survey responses. Next, to identify factors associated with portal and telehealth use, we compared their use by demographic characteristics and by health promotion behaviors using bivariate Fisher exact tests.

The survey also contained 6 open-ended questions focusing on portal and telehealth use, preferences, and experiences. We included questions related to both the portal and telehealth, going beyond the study by O'Connor et al [24] that focused on telehealth alone. Portal use is critical to study because portals are used by a greater share of patients than telehealth, and increasingly, health care systems use the portal as the gateway to their telehealth application [27,28]. Open-ended survey text

XSL•FO
RenderX

data were analyzed using axial coding techniques, and thematic findings were used to refine the semistructured interview guide [29].

Interview Sampling and Recruitment

Based on survey findings, we created a semistructured guide to probe interview participants' perceptions and experiences with the portal and telehealth, including factors that motivated their enrollment and use. We also added questions about the role of care partner support and experiences related to accessing test results, given emergent findings from the open-ended survey questions that these factors were associated with technology uptake and use (Multimedia Appendix 1).

We recruited interview participants from a subsample of survey completers who expressed a willingness to be contacted for follow-up interviews. We stratified interview sampling to recruit half of the participants from the survey cohort recruited through the Cancer Center and half through various RedBird area community organizations. Within each group, we prioritized recruitment of participants whose responses to the open-ended survey questions were more robust, presuming that this might correlate with greater engagement in the interviews. As we accrued participants, we monitored the demographics of our sample to ensure that we achieved diversity by gender and race or ethnicity that reflected the RedBird area population. We made up to 3 attempts to reach each participant via their preferred mode of contact. During recruitment, we also verified whether the individual had "consistent" or "mostly reliable" access to the web and/or access to a smartphone to support our focus on individuals with potential access to the portal and telehealth. That is, we chose to focus on those without device or Wi-Fi access barriers because the goal of our inquiry was to identify potential opportunities for educational interventions, such as barriers due to misperceptions, attitudes, or limited skills. Furthermore, device affordability and Wi-Fi infrastructure were improving under the Affordable Connectivity Act and broadband expansion by leading telecommunications companies.

Interview Data Collection and Analysis

A trained qualitative interviewer (ER) conducted interviews by phone using the semistructured interview guide. Interviews lasted an average of 30 minutes (range 16‑61 minutes). Interviewers conducted rapid data analysis [30,31] during interview data collection to allow for immediate exploration of emerging patterns and insights. Rapid data analysis also serves as a tool to enhance communication among members of the interview team, facilitate monitoring effectiveness of the interview guide and determination of thematic saturation, and promote consistency of data capture.

Concurrently, we also conducted an in-depth qualitative analysis of interview transcripts. All transcript data were thematically coded by trained qualitative researchers using NVivo (Lumivero). The lead qualitative investigator (RTH) drafted an initial codebook deductively driven by concepts in the interview guide and inclusive of emergent findings extracted from the rapid analysis. Next, 2 qualitatively trained staff (ER and AG) engaged in open coding of 15% of the dataset (n=3 transcripts), iteratively refining the codes and definitions to match themes and patterns. They tested the codebook by independently coding an additional 10% of the dataset (n=2 transcripts) and then met to compare codes and calculate an interrater reliability score using Cohen κ. Because the rate of concordance exceeded 0.75, indicating substantial agreement, the remaining transcripts (n=15) were divided between the coders and analyzed independently. The research team met weekly to discuss and interpret findings. None of the qualitatively trained researchers engaged in data collection or analysis were physicians or had clinical relationships with participants.

Ethical Considerations

All study-affiliated personnel have current certification in the ethical conduct of human subjects research on file with the University of Texas Southwestern Medical Center (UTSW)'s Human Research Protection Program. We obtained informed verbal consent for participation and audio recording in accordance with the approved study protocol from the institutional review board at UTSW (STU-2023‑0469). All audio recordings were transcribed verbatim and deidentified by a professional transcription vendor contracted with UTSW under a Health Insurance Portability and Accountability Act waiver. Participants were mailed a $30 gift card in appreciation for their participation.

Results

Participant Characteristics

We distributed 1262 total surveys via (1) digital methods: 646 were distributed and 64 completed, and (2) paper: 616 were distributed and 118 completed. Characteristics of the 182 total participants (182/1262, 14.4% overall response rate) did not differ by survey modality (data not shown, chi-square *P* value >.05) (Table 1). Of 182 survey participants, most were older (n=109, 66%; age ≥60 years), African American or Black (n=120, 66%), and women (n=142, 79%). Survey participants reported experience with health technology: a little more than half (97/182, 54%) had completed a telehealth appointment, and a majority (131/182, 72%) had used a patient portal.

For interviews, we contacted 59 individuals, 3 declined to participate (reason cited was "not interested"), 7 were ineligible due to lack of a Wi-Fi connection or a smartphone, and 29 did not respond to ≤3 recruitment attempts. Of the 20 semistructured interview participants, most were older (n=14, 70%; age ≥60 years) and 40% (n=8) were men. We had attempted to diversify our interview sample with individuals of varying levels of portal and telehealth experience; however, due to lack of recruitment response among portal and/or telehealth nonusers, our interview sample included a greater proportion of those with greater technology experience (Multimedia Appendix 2). Nearly all interview participants had used the portal at least once (18/20, 90%), and a majority completed at least 1 telehealth appointment (12/20, 60%).

XSL•FO
RenderX

Table . Survey (N=182) and interview (N=20) participant characteristics.

Characteristics	Survey participants	Interview participants
Age (years), n (%)		
<40	22 (13.3)	2 (10)
40 - 59	35 (21.1)	4 (35)
60 - 69	40 (24.1)	7 (35)
≥70	69 (41.6)	7 (35)
Unknown	16	0
Gender, n (%)		
Female	142 (78.5)	12 (60)
Male	39 (21.5)	8 (40)
Unknown	1	0
Race/ethnicity (check all that apply), n (%)		
Non-Hispanic African American or Black	120 (65.9)	11 (55)
Non-Hispanic White	28 (15.4)	6 (30)
Hispanic	23 (12.6)	2 (10)
Other	11 (6.0)	1 (5)
Education, n (%)		
High school graduate or less	51 (28.5)	6 (30)
Some college/college graduate	91 (50.8)	12 (60)
Postgraduate	37 (20.7)	2 (10)
Unknown	3	0
Language, n (%)		
English	174 (95.6)	20 (100)
Spanish	8 (4.4)	0
“Have you ever used the portal?”, n (%)		
Yes	131 (72.0)	18 (90)
No	51 (28.0)	2 (10)
“Have you ever completed a telehealth appointment?”, n (%)		
Yes	97 (53.3)	12 (60)
No	83 (46.4)	8 (40)
Unknown	2	0
Portal and telehealth use combined, n (%)		
Used both the portal and telehealth	91 (50.0)	12 (60)
Used the portal only	40 (22.0)	3 (15)
Used telehealth only	6 (3.3)	3 (15)
Used neither the portal nor telehealth	43 (23.6)	2 (10)
Used portal and unknown telehealth use	2	0
Recruitment method, n (%)		
Cancer Center	98 (53.8)	10 (50)
Community outreach	84 (46.2)	10 (50)

XSL•FO
RenderX

Survey Findings

As shown in Table 2, more than half of our sample used both the portal and telehealth (91/180, 51%), and nearly a quarter (43/180, 24%) reported not using either technology; the remainder used 1 technology or the other. We combined those using only 1 technology or the other due to a very small number of patients reporting use of telehealth but no portal use (n=6). Compared with technology users, those reporting no use of portal or telehealth had a lower level of education (*P*<.001) or were more likely to be Spanish speaking (*P*<.011).

Table 3 illustrates frequency of endorsement of each statement about using the web for health promotion activities. The most commonly endorsed statement was "I can use my computer or mobile phone to communicate about my health needs" (81.5%). The percent agreeing to the statements varied statistically significantly by portal and/or telehealth (*P*<.001), with greater agreement among those using portal and telehealth use.

Table . Demographic characteristics of survey participants by ever use of the portal and/or telehealth[a].

Characteristics	Categories of participants by ever use of the portal and/or telehealth[a]				
	Both, 91 (51%)[b]	One technology, 46 (26%)[b]	Neither, 43 (24%)[b]	Overall (N=180)	*P* value[c]
Age (years), n (%)					.71
<40	13 (15.3)	6 (14.3)	3 (8.1)	22 (13.4)	
40 - 59	15 (17.6)	12 (28.6)	7 (18.9)	34 (20.7)	
60 - 69	21 (24.7)	10 (23.8)	9 (24.3)	40 (24.4)	
≥70	36 (42.4)	14 (33.3)	18 (48.6)	68 (41.5)	
Unknown	6	4	6	16	
Gender, n (%)					.22
Female	72 (79.1)	39 (84.8)	29 (69.0)	140 (78.2)	
Male	19 (20.9)	7 (15.2)	13 (31.0)	39 (21.8)	
Unknown	0	0	1	1	
Race/ethnicity, n (%)					.095
Non-Hispanic African American or Black	56 (61.5)	32 (69.6)	31 (72.1)	119 (66.1)	
Non-Hispanic White	21 (23.1)	5 (10.9)	2 (4.7)	28 (15.6)	
Hispanic	9 (9.9)	5 (10.9)	8 (18.6)	22 (12.2)	
Other	5 (5.5)	4 (8.7)	2 (4.7)	11 (6.1)	
Education, n (%)					<.001
High school graduate or less	9 (10.0)	11 (24.4)	29 (69.0)	49 (27.7)	
Some college/college graduate	55 (61.1)	25 (55.6)	11 (26.2)	91 (51.4)	
Postgraduate	26 (28.9)	9 (20.0)	2 (4.8)	37 (20.9)	
Unknown	1	1	1	3	
Survey language, n (%)					.011
English	89 (97.8)	46 (100.0)	38 (88.4)	173 (96.1)	
Spanish	2 (2.2)	0 (0.0)	5 (11.6)	7 (3.9)	
Recruitment location, n (%)					.43
Community	39 (42.9)	25 (54.3)	19 (44.2)	83 (46.1)	
Simmons	52 (57.1)	21 (45.7)	24 (55.8)	97 (53.9)	

[a]Two individuals who had unknown telehealth use were not included in this analysis.

[b]n (%).

[c]Fisher exact test.

XSL•FO
RenderX

Table . Survey responses regarding health promotion activity by ever use of the portal and/or telehealth.

Variable	Categories of participants by ever use of the portal and/or telehealth				
	Both, 91 (51%)[a]	One technology, 46 (26%)[a]	Neither, 43 (24%)[a]	Overall (N=180)	*P* value[b]
Overall technology access: I can use my computer or mobile phone to communicate about my health needs, n (%)					<.001
Agree	85 (93.4)	36 (78.3)	24 (58.5)	145 (81.5)	
Disagree	4 (4.4)	6 (13.0)	13 (31.7)	23 (12.9)	
Neither agree nor disagree	2 (2.2)	4 (8.7)	4 (9.8)	10 (5.6)	
Unknown	0	0	2	2	
Health literacy: I can use the internet to learn about health topics that are relevant to me, n (%)					<.001
Agree	83 (91.2)	37 (82.2)	24 (58.5)	144 (81.4)	
Disagree	5 (5.5)	3 (6.7)	13 (31.7)	21 (11.9)	
Neither agree nor disagree	3 (3.3)	5 (11.1)	4 (9.8)	12 (6.8)	
Unknown	0	1	2	3	
Improving health autonomy: I know how to find helpful health resources on the internet, n (%)					<.001
Agree	86 (94.5)	36 (80.0)	19 (46.3)	141 (79.7)	
Disagree	4 (4.4)	4 (8.9)	17 (41.5)	25 (14.1)	
Neither agree nor disagree	1 (1.1)	5 (11.1)	5 (12.2)	11 (6.2)	
Unknown	0	1	2	3	
Appreciating the value of digital health technology: I can use the internet to learn how to manage my health in a positive way, n (%)					<.001
Agree	71 (78.0)	31 (70.5)	19 (46.3)	121 (68.8)	
Disagree	8 (8.8)	7 (15.9)	17 (41.5)	32 (18.2)	
Neither agree nor disagree	12 (13.2)	6 (13.6)	5 (12.2)	23 (13.1)	
Unknown	0	2	2	4	
Health literacy: I can summarize basic health information from the internet in my own words, n (%)					<.001
Agree	74 (81.3)	25 (54.3)	13 (33.3)	112 (63.6)	
Disagree	7 (7.7)	11 (23.9)	15 (38.5)	33 (18.8)	
Neither agree nor disagree	10 (11.0)	10 (21.7)	11 (28.2)	31 (17.6)	
Unknown	0	0	4	4	
Confidence using technology: The internet improves my communication with health professionals, n (%)					<.001
Agree	68 (74.7)	22 (48.9)	16 (39.0)	106 (59.9)	
Disagree	8 (8.8)	7 (15.6)	19 (46.3)	34 (19.2)	
Neither agree nor disagree	15 (16.5)	16 (35.6)	6 (14.6)	37 (20.9)	
Unknown	0	1	2	3	
Integrating use of digital health technology into one's daily life: I can use information on the internet to make informed decisions about my health, n (%)					<.001
Agree	64 (70.3)	23 (51.1)	18 (43.9)	105 (59.3)	
Disagree	9 (9.9)	10 (22.2)	17 (41.5)	36 (20.3)	

XSL•FO
RenderX

Variable	Categories of participants by ever use of the portal and/or telehealth				
	Both, 91 (51%)[a]	One technology, 46 (26%)[a]	Neither, 43 (24%)[a]	Overall (N=180)	*P* value[b]
Neither agree nor disagree	18 (19.8)	12 (26.7)	6 (14.6)	36 (20.3)	
Unknown	0	1	2	3	
Confidence using technology: Using the internet for health-related activities makes me feel actively involved with my health care, n (%)					<.001
Agree	63 (69.2)	25 (55.6)	15 (37.5)	103 (58.5)	
Disagree	9 (9.9)	8 (17.8)	17 (42.5)	34 (19.3)	
Neither agree nor disagree	19 (20.9)	12 (26.7)	8 (20.0)	39 (22.2)	
Unknown	0	1	3	4	
Improving health autonomy: I find the internet useful for monitoring my health, n (%)					<.001
Agree	66 (72.5)	25 (56.8)	9 (22.5)	100 (57.1)	
Disagree	7 (7.7)	8 (18.2)	22 (55.0)	37 (21.1)	
Neither agree nor disagree	18 (19.8)	11 (25.0)	9 (22.5)	38 (21.7)	
Unknown	0	2	3	5	

[a]n (%).
[b]Fisher exact test.

Interview Findings

We found evidence supporting 4 key factors by O'Connor et al [24]: improving individuals' health autonomy, integrating use of digital health technology into individuals' daily lives, obtaining recommendations from trusted individuals, and appreciating the value of digital health technology.

Improving Individuals' Health Autonomy

Many participants reported that using technology, such as the patient portal, empowered them with easy access to information about their health and in a way that was understandable to them.

> *Well, I'd probably describe it as a good way to stay up on your personal health without being worried about wasting time or having to wait for pathology. If you have questions, you can always message the clinic or there is some place within MyChart, which explains certain texts and what the results mean. It's really pretty easy for you to stay up on what's going on with you personally.* [P04, experience with both the portal and telehealth]

Using the portal also facilitated participants' self-efficacy and the belief in their ability to improve their health.

> *Well, I like MyChart because it makes you focus on your health more better. [It lets] you know what's wrong with you or [it lets] you know that you need to work on fixing this. It's not too late to fix it.* [P10, no telehealth or direct portal experience but a care partner helps her use the portal]

The portal also helped patients take a more active role in making decisions about their health and in leading conversations with their health care providers.

> *I would say using MyChart is about being pro-help-for-yourself and knowing how to say, hey, this is what's going on with me. I can make my own healthy choices and also confirm with my doctor, my nurse practitioner, of what these choices are and what they think or suggest about it. So MyChart is a tool for a person to make healthy choices for themselves and stay informed.* [P17, experience with both the portal and telehealth]

Integrating Use of Digital Health Technology Into Individuals' Daily Lives

Many participants found the portal to be a helpful tool to access all their health information, especially given the fact that they have multiple medications.

> *There have been many times where I need to go back and look...like say, for example, I call Medco and say, 'I need to go over my medications.' I don't have them written down anywhere, but I can go to my last visit and look at all my medications that are listed.... Or if I've been given special instructions, I'll go back to the visit summary and look at whatever the details were for that particular visit.* [P15, experience with both the portal and telehealth]

Participants liked the ease of access through their cell phone, presumably a technology with which they were already familiar.

XSL•FO
RenderX

> *I think it's actually pretty cool that you can get into medical charts through your phone and be able to set appointments.* [P13, experience with the portal only]

Obtaining Recommendations From Trusted Individuals

Many participants who enrolled in the portal reported first learning about it from a medical provider or clinical staff.

> *[At the] doctor's office... They explained how it worked and that I could keep up with my appointments and my drugs and everything–pertinent information.* [P05, experience with both the portal and telehealth]

Establishing a relationship with the provider first was the key to a successful portal interaction for 1 patient.

> *Once you visit your doctor, you know who he or she is and you're comfortable with them, I don't think that's much of a problem, 'cause you have a relationship.* [P11, experience with the portal only]

Appreciating the Value of Digital Health Technology

For 1 participant, the portal offered a more rapid and effective solution to a billing problem than troubleshooting by telephone.

> *I was getting a bill for some treatment... I called the billing department, and they gave me a voicemail and they just put me on hold. It don't work [sic]. But then I went into my MyChart and went to the billing link and left a message stating that I shouldn't have been billed for it... They answered it and removed that and sent me an email back in MyChart saying that they had been verified that I did have an insurance coverage and I shouldn't have been billed.* [P09, experience with the portal only]

Some participants also valued the portal's lesser-known features, such as appointment reminders and the option to view appointment locations on a map, which streamlined navigation across an expansive medical campus.

> *Well, for one thing, it keeps me up to date on when I have appointments and also with all the different buildings as to which building it's in and where is it on the map... That's very helpful to me... Yeah, you can see the schematic of where all the buildings are that are running around the [location].* [P02, experience with both the portal and telehealth]

Multiple participants noted that the portal offers an efficient and dependable channel for communicating with their care team. They valued the rapid response via the portal, especially in comparison with phone interactions.

> *It's an easy way to communicate with the doctor because sometimes [when] you call on your phone, you'll be on that phone, nobody will answer. Or they'll answer, they keep you on hold [so long] that you get frustrated, and you hang up the phone. With MyChart, when you send the email to them, they will respond. It may take about five or six hours or two hours of work, but they will still reply the same day to let you know what is going on and that will give you hope that you'll get it.* [P01, experience with both the portal and telehealth]

In addition, participants in this study uniquely reported 3 additional key factors that encouraged uptake and use of the portal and telehealth: enlisting support from care partners or peers, managing severe or chronic illness, and accessing test results rapidly.

Enlisting Support From Care Partners or Peers

Several participants relied on family members to use the portal and/or telehealth.

> *My daughter–she works at a breast cancer clinic so she pretty much know how to do all that [portal and telehealth] at home...she will tell me my blood pressure was very high. I would tell her, 'yeah, the doctor told me I need to work on my blood pressure.' And then she looked at it again and she said, 'Oh, your blood pressure has went down.' It just lets you know. It show you what's going on.* [P10, no direct telehealth or portal experience but uses portal through a care partner]

Other participants spoke of being a care partner for family members or friends and highlighted how technology facilitated management of the patient's medical needs.

> *Well...it helps me, with him being in a wheelchair, it saves me time to be able to help him and to get all the things done that he needs as far as his physical therapy and all of his appointments in line and all of his blood work that lets me know that he's doing okay from day to day.* [P13, personal experience with the portal and care partner for her husband to use the portal]

Managing Severe or Chronic Illness

Several participants who used the portal emphasized its value when coordinating care across multiple providers.

> *I have become more active in [the portal] as the number of doctors have increased. I rely on it more and more.* [P15, experience with both the portal and telehealth]

Others described how they used the portal to manage their chronic illness, such as the following participant who had kidney disease.

> *They checked me and they made an appointment and it's all written on [the portal]. I look into my test results. And if I forget something I'll just go back and look. I double check on my A1C or what's my potassium because they say it's kind of high, and now it went down. So it's good. I can always check back.* [P08, experience with both the portal and telehealth]

Four individuals recruited from the Cancer Center specifically mentioned that they joined the portal when they were diagnosed with cancer and started cancer treatment. One said,

> *I'm not sure exactly when I started, but I had anti-cancer surgery, and I was recuperating after the stay at the hospital and [the portal] was a good way*

XSL•FO
RenderX

> *to keep up with what was going on.* [P02, experience with both the portal and telehealth]

Another participant recognized that telehealth offered a convenient alternative to in-person visits, especially during times of acute illness or frailty.

> *If you're feeling really, really sick and you just really can't get out, that's very helpful that if you need to stay in bed and talk to the doctor, you can do that.* [P15, experience with both the portal and telehealth]

Accessing Test Results Rapidly

Participants from both the Cancer Center and the community cohorts generally liked the portal's facilitation of faster access to test results compared with formerly having to rely on waiting for results to be communicated via the telephone.

> *I like that it don't take a week to hear back from the doctor. Within the day or so, the doctors get back with you... it is not a hassle of having to wait on the phone or wait for the doctor to call me or get an email.* [P13, experience with the portal only]

> *I wouldn't have known the results until probably my next appointment. But I went on MyChart and I found out what it was, so I didn't have to wait until my next appointment with the doctor to get the results.* [P18, experience with both the portal and telehealth]

Many people felt that waiting for the clinician to contact them with the results was more anxiety-provoking than receiving the results, even if they might be negative.

> *I get this anxiety like, 'Damn, what's the result?' I got thyroid [issues] so I'm like, 'Is it good?' I like to get mine fast…. If you want to know, then at the end of the day, that's the whole point of having MyChart, because you want to know.* [P16, experience with both the portal and telehealth]

At the same time, 1 participant acknowledged that, in some circumstances, receiving one's results before a doctor had a chance to speak with them could cause more anxiety or a potentially negative experience if the test result was life-threatening.

> *I'm in the last [cancer] stage. I wouldn't want to find that out through MyChart. I would rather through a phone call… because I can't talk to MyChart.* [P18, experience with both the portal and telehealth]

Discussion

Principal Findings

In this iterative, mixed methods study, we describe attitudes toward and experiences with telehealth and the portal among a minoritized population in Dallas, Texas. We identify 7 factors that are key to the uptake and use of digital health technologies. Four of these factors had previously been identified by the meta-analysis of qualitative studies by O'Connor et al [24]: improving health autonomy, integrating digital health technology into individuals' daily lives, receiving technology recommendations from trusted individuals, and appreciating the value of digital health technologies. The 3 additional factors we identified in this study—enlisting support from care partners, managing severe or chronic illness, and accessing test results rapidly—are valuable contributions because they provide insight into potential intervention opportunities among specific populations and clinical settings. Survey findings also align with the factors by O'Connor et al [24] that are key to telehealth utilization in underserved populations (improving health autonomy, integrating use of digital health technology into one's daily life, and appreciating the value of digital health technology), as well as other factors we know from the literature to be relevant to successful engagement with the portal or telehealth (eg, health literacy, confidence using technology, and overall technology access). Our finding that language and level of education are associated with nonusers of the portal and telehealth underscores the importance of making the portal and telehealth accessible to populations that are Spanish-speaking and, among English speakers, with lower literacy levels. Notably, although a Spanish version of the patient portal is available in some health care systems, multiple challenges have dampened enthusiasm, uptake, and use of it [32].

Support from care partners has been described as a key facilitator for technology access among other populations, especially older individuals and those with chronic illness or disabilities [33,34]. To our knowledge, this is the first study to qualitatively describe care partner support as a key facilitator of digital health technology engagement among minoritized populations. Prior work suggests that care partners value having access to patient information and report benefits related to greater understanding of medications and follow-up steps and increased feelings of control [35]. Care partner assistance with patient portals is more common when patients are older, have limited English proficiency or lower overall literacy, and greater illness severity [36]. While our sample was older, our findings around care partner support as a facilitator suggest that care partners may also play a more prominent role among lower-resource, minoritized populations.

Second, our finding about the value placed on rapid access to test results has far-ranging implications across clinical settings and populations. Participants ranked rapid access to this information as a key to motivating portal use. For example, individuals recruited from the Cancer Center expressed mostly positive and some negative opinions of rapid access to test results. Specifically, a few participants felt that patients should not be given a cancer diagnosis via the portal, but that access to laboratory results and appointments was very important. At the same time, individuals recruited from the community (who presumably were less likely to be in active cancer treatment) also felt strongly that rapid access to appointments, medications, and the opportunity to communicate with their providers was a strong motivator for portal use. While we did not ask about community participants' medical conditions, this finding may reflect higher rates of chronic illness in lower-resource populations, such as diabetes and hypertension, which also involve regular appointments and coordination across multiple providers.

The 21st Century Cures Act [37], which functionally required that institutions immediately make available to patients, on

XSL•FO
RenderX

request, information in the electronic health record, has spurred debate since it was passed in 2021 [38,39]. Most health care organizations choose to meet this regulatory requirement by making results automatically viewable on the patient portal as soon as they are available in electronic health records, meaning patients are often able to see results before speaking with a clinician. Consequently, patients and families living with cancer will likely receive "nonnormal" results with some frequency as they use the portal to review data, laboratory test results, and imaging. There is controversy about providing patients with immediate access to test results in part because this information can be conveyed prior to the patient speaking with their clinician [38,40]. The potential positive and negative impacts of immediate test result access are likely to vary immensely across individuals in both patient and clinician groups [41,42]. Health care organizations may wish to provide additional education and support to patients and care partners, particularly around how to configure portal settings to either receive or mute new test result notifications. Our findings showing some complex preferences about information provision among our participants suggest that a "one-size-fits-all" approach will not be sufficient.

Digital inclusion, or "[ensuring] that all individuals and communities, including the most disadvantaged, have access to and use of digital tools," [43] must be emphasized in order to overcome the existing digital divide [13]. With tools such as Gallego's Digital Divide Index and the Centers for Disease Control and Prevention's social vulnerability index [44], among others, social epidemiologists have provided ample evidence of where behavioral interventionists should implement more intensive interventions to maximize use of resources to improve health outcomes [45]. Additionally, intervention strategies shown to be effective in addressing digital health equity have emphasized user design (eg, compatibility with existing devices and culturally appropriate content), supportive infrastructure (eg, assistance with devices and connectivity), and educational support (ie, from professionals, family, or friends) [46]. Our findings of the importance of care partners, rapid access to test results, and coordinated care for individuals with more complex disease contribute to this shared understanding of what, where, and to whom digital health equity interventions should be applied. And as a next step, using an implementation science framework such as Glasgow's RE-AIM (1999) may help us optimize how the intervention is applied [47].

Finally, it is important to recognize that providers and health care systems contribute to the persistence of digital health disparities. Multiple studies using data from the Health Information National Trends Survey have documented bias in provider recommendations for portal and telehealth use, with greater encouragement directed toward patients who are female, White, and college-educated [48-50]. Other Health Information National Trends Survey–based studies have found that provider recommendation is a key driver of telehealth use [51] and that adults with limited English proficiency are more likely than those proficient in English to seek health information from professionals [52]. Notably, 1 study found that Spanish-preferring participants placed greater trust in information from government agencies and often turned to web-based videos for health-related information [52]. These findings illustrate the importance of educating providers and health care institutions about how provider bias can exacerbate the digital divide in health care access and disseminating reliable health information through multiple channels.

Although significant progress has been achieved in strengthening digital health access and use among underserved communities [53], several barriers remain. These include lower educational attainment [54], lack of awareness of the benefits of digital health technologies [55-57], mistrust in the health care system evidenced as concern for data privacy [58,59], and ambivalence toward technology [60]. These barriers often exacerbate existing health disparities that affect the same minoritized populations with multiple chronic conditions, who stand to benefit greatly from portal and telehealth access. Future implementation strategies to address the persistent digital divide will need to provide more personalized options, be supported by robust financial and human resources, and include trust-building with cultural minority communities [60].

Strengths and Limitations

With only 20 interviews, in a mostly older female Non-Hispanic Black or African American population, it is not possible to generalize these findings broadly. We also chose to focus our interviews on individuals who did not have barriers in access to Wi-Fi or smartphone technology in order to delve more deeply into potential intervention opportunities that could be addressed through education versus systemic infrastructure barriers over which we had limited control. While the infrastructure for Wi-Fi is increasing, and cellular phone ownership exceeds 90%, even among low-income populations [61], this sampling strategy resulted in the exclusion of the perspectives of a more diverse range of participants, which limits generalizability and precludes a more comprehensive understanding of the issues. Additionally, while we identified care partners as an important facilitator and explored this topic in depth in interviews, we missed an opportunity to obtain quantifiable data on care partner demographics and technology-related descriptors of individuals acting as care partners by not including such questions in the survey. This would have enabled us to contribute to the literature on the demographics, technology-related skills, and attitudes of people who use technology as care partners versus those who need care partner assistance. Understanding the relationship between care providers and receivers may facilitate conceptualization of new hypotheses and strategies for interventions.

However, our study has several strengths. Our finding that the uptake and use of technology may be related to chronicity and severity of illness merits attention. While perhaps those with a more life-threatening condition may be more acutely motivated to engage in their health care, findings from this study suggest that those with chronic conditions such as diabetes, certain cancers, and hypertension may also be persuaded to use the portal to monitor their test results. Also, our study contributes uniquely to the literature by focusing both quantitatively and qualitatively on the perspectives of a multiracial, minoritized population toward digital health technologies. Our finding that language and level of education are associated with lower portal and telehealth use is useful for informing future intervention

XSL•FO
RenderX

efforts. Finally, future research may be enhanced by examining behavioral outcomes, which may reveal insights regarding the possible associations between digital health interventions, health care utilization, and patient outcomes.

Conclusions

We contribute evidence that seven key factors are relevant to the development of strategies to promote digital inclusion in future interventions for underserved populations: (1) improving patient autonomy, (2) integrating digital health technology into daily life, (3) receiving recommendations from trusted individuals, (4) appreciating the value of the technology [24], (5) enlisting support from care partners, (6) managing severe or chronic illness, and (7) accessing test results rapidly. Future studies should focus on making the portal and telehealth more accessible to English-speaking populations of lower educational attainment and Spanish speakers and should address how different clinical settings and health conditions affect the uptake of the patient portal and telehealth.

Acknowledgments

The authors thank their community partners: Shelia Patterson at Oak Cliff Bible Fellowship, Renae Perry at The Senior Source, Sherri Nixon at the Hoover Community Development Corporation, Dr Alisha Orton and Victoria Doby at the UT Southwestern Clinic at RedBird, and Alexandra Huffman and members of the Patient and Family Advisory Council at Simmons Comprehensive Cancer Center. They are grateful for assistance from Dr Meera Muthukrishnan for statistical analysis; Kathryn Shahan for Figure 1 and Cynthia Ortiz for Table 1; and Zoe Osei-Appiah, Tiffany Hang, and Christelle King. This work was funded by The Cary Council Foundation (PI: RTH); the Harold C Simmons Comprehensive Cancer Center, Office of Community Outreach, Equity & Engagement (PI: SLP); and an award from the National Institutes of Health (NIH; NIH P30CA142543; PI: Arteaga). Research reported in this publication was also supported by the National Center for Advancing Translational Sciences of the NIH (award number UL1 TR003163). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

Unrelated to this work, SLP reports consulting fees from Pfizer and Gilead; the remaining authors declare that they have no conflicts of interest to disclose.

Multimedia Appendix 1
Interview guide.
[PDF File, 163 KB - formative_v9i1e70146_app1.pdf]

Multimedia Appendix 2
Interview recruitment CONSORT (Consolidated Standards of Reporting Trials) diagram.
[PDF File, 92 KB - formative_v9i1e70146_app2.pdf]

Checklist 1
SRQR (Standards for Reporting Qualitative Research) checklist.
[PDF File, 51 KB - formative_v9i1e70146_app3.pdf]

References

1. Chimonas S, Lipitz-Snyderman A, Spiegelhoff Z, et al. Persistence of telemedicine usage for breast and prostate cancer after the peak of the COVID-19 pandemic. Cancers (Basel) 2023 Oct 12;15(20):37894328. [doi: 10.3390/cancers15204961] [Medline: 37894328]
2. Cancer Center Cessation Initiative Telehealth Working Group. Telehealth delivery of tobacco cessation treatment in cancer care: an ongoing innovation accelerated by the COVID-19 pandemic. J Natl Compr Canc Netw 2021 Nov;19(Suppl 1):S21-S24. [doi: 10.6004/jnccn.2021.7092] [Medline: 34872049]
3. El-Tallawy SN, Pergolizzi JV, Vasiliu-Feltes I, et al. Innovative applications of telemedicine and other digital health solutions in pain management: a literature review. Pain Ther 2024 Aug;13(4):791-812. [doi: 10.1007/s40122-024-00620-7] [Medline: 38869690]
4. Tolu-Akinnawo O, Ezekwueme F, Awoyemi T. Telemedicine in cardiology: enhancing access to care and improving patient outcomes. Cureus 2024 Jun;16(6):e62852. [doi: 10.7759/cureus.62852] [Medline: 38912070]

XSL•FO
RenderX

5. Cox A, Lucas G, Marcu A, et al. Cancer survivors' experience with telehealth: a systematic review and thematic synthesis. J Med Internet Res 2017 Jan 9;19(1):e11. [doi: 10.2196/jmir.6575] [Medline: 28069561]
6. Leader AE, Capparella LM, Waldman LB, et al. Digital literacy at an urban cancer center: implications for technology use and vulnerable patients. JCO Clin Cancer Inform 2021 Aug;5(872-80):872-880. [doi: 10.1200/CCI.21.00039] [Medline: 34428075]
7. Han HR, Gleason KT, Sun CA, et al. Using patient portals to improve patient outcomes: systematic review. JMIR Hum Factors 2019 Dec 19;6(4):e15038. [doi: 10.2196/15038] [Medline: 31855187]
8. Sarkar U, Lyles CR, Parker MM, et al. Use of the refill function through an online patient portal is associated with improved adherence to statins in an integrated health system. Med Care 2014 Mar;52(3):194-201. [doi: 10.1097/MLR.0000000000000069] [Medline: 24374412]
9. Harris LT, Haneuse SJ, Martin DP, Ralston JD. Diabetes quality of care and outpatient utilization associated with electronic patient-provider messaging: a cross-sectional analysis. Diabetes Care 2009 Jul;32(7):1182-1187. [doi: 10.2337/dc08-1771] [Medline: 19366959]
10. Reed M, Huang J, Graetz I, et al. Outpatient electronic health records and the clinical care and outcomes of patients with diabetes mellitus. Ann Intern Med 2012 Oct 2;157(7):482-489. [doi: 10.7326/0003-4819-157-7-201210020-00004] [Medline: 23027319]
11. McBride SM, Hughes HK, MacDonald SM. Future of telemedicine in radiation oncology. Semin Radiat Oncol 2024 Oct;34(4):463-467. [doi: 10.1016/j.semradonc.2024.07.010] [Medline: 39271281]
12. Saragih ID, Tonapa SI, Osingada CP, Porta CM, Lee BO. Effects of telehealth-assisted interventions among people living with HIV/AIDS: a systematic review and meta-analysis of randomized controlled studies. J Telemed Telecare 2024 Apr;30(3):438-450. [doi: 10.1177/1357633X211070726] [Medline: 34967240]
13. van Kessel R, Wong BLH, Clemens T, Brand H. Digital health literacy as a super determinant of health: more than simply the sum of its parts. Internet Interv 2022 Mar;27(100500):100500. [doi: 10.1016/j.invent.2022.100500] [Medline: 35242586]
14. Gallardo R. Digital Divide Index.: Purdue Center for Regional Development; 2024. URL: http://pcrd.purdue.edu/ddi [accessed 2025-06-26]
15. Cousins MM, Van Til M, Steppe E, et al. Age, race, insurance type, and Digital Divide Index are associated with video visit completion for patients seen for oncologic care in a large hospital system during the COVID-19 pandemic. PLoS One 2022;17(11):e0277617. [doi: 10.1371/journal.pone.0277617] [Medline: 36395112]
16. Paro A, Rice DR, Hyer JM, et al. Telehealth utilization among surgical oncology patients at a large academic cancer center. Ann Surg Oncol 2022 Nov;29(12):7267-7276. [doi: 10.1245/s10434-022-12259-9] [Medline: 35896926]
17. Saeed SA, Masters RM. Disparities in health care and the digital divide. Curr Psychiatry Rep 2021 Jul 23;23(9):61. [doi: 10.1007/s11920-021-01274-4] [Medline: 34297202]
18. Chesser A, Burke A, Reyes J, Rohrberg T. Navigating the digital divide: a systematic review of eHealth literacy in underserved populations in the United States. Inform Health Soc Care 2016;41(1):1-19. [doi: 10.3109/17538157.2014.948171] [Medline: 25710808]
19. Bailey JE, Gurgol C, Pan E, et al. Early patient-centered outcomes research experience with the use of telehealth to address disparities: scoping review. J Med Internet Res 2021 Dec 7;23(12):e28503. [doi: 10.2196/28503] [Medline: 34878986]
20. Williams K, Markwardt S, Kearney SM, et al. Addressing implementation challenges to digital care delivery for adults with multiple chronic conditions: stakeholder feedback in a randomized controlled trial. JMIR Mhealth Uhealth 2021 Feb 1;9(2):e23498. [doi: 10.2196/23498] [Medline: 33522981]
21. Tebb KP, Rodriguez F, Pollack LM, et al. Assessing the effectiveness of a patient-centred computer-based clinic intervention, Health-E You/Salud iTu, to reduce health disparities in unintended pregnancies among Hispanic adolescents: study protocol for a cluster randomised control trial. BMJ Open 2018 Jan 10;8(1):e018201. [doi: 10.1136/bmjopen-2017-018201] [Medline: 29326184]
22. Lama Y, Davidoff AJ, Vanderpool RC, Jensen RE. Telehealth availability and use of related technologies among Medicare-enrolled cancer survivors: cross-sectional findings from the onset of the COVID-19 pandemic. J Med Internet Res 2022 Jan 25;24(1):e34616. [doi: 10.2196/34616] [Medline: 34978531]
23. Horowitz CR, Brenner BL, Lachapelle S, Amara DA, Arniella G. Effective recruitment of minority populations through community-led strategies. Am J Prev Med 2009 Dec;37(6 Suppl 1):S195-S200. [doi: 10.1016/j.amepre.2009.08.006] [Medline: 19896019]
24. O'Connor S, Hanlon P, O'Donnell CA, Garcia S, Glanville J, Mair FS. Understanding factors affecting patient and public engagement and recruitment to digital health interventions: a systematic review of qualitative studies. BMC Med Inform Decis Mak 2016 Sep 15;16(1):120. [doi: 10.1186/s12911-016-0359-3] [Medline: 27630020]
25. Fetters MD, Curry LA, Creswell JW. Achieving integration in mixed methods designs-principles and practices. Health Serv Res 2013 Dec;48(6 Pt 2):2134-2156. [doi: 10.1111/1475-6773.12117] [Medline: 24279835]
26. Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: Building an international community of software platform partners. J Biomed Inform 2019 Jul;95(103208):103208. [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]
27. Johnson C, Richwine C, Patel V. Individuals' access and use of patient portals and smartphone health apps, 2020. ASTP Health IT Data Brief.: Office of the Assistant Secretary for Technology Policy; 2012. [Medline: 39571059]

XSL•FO
RenderX

28. Lucas JW, Wang X. Declines in telemedicine use among adults: United States, 2021 and 2022. Natl Health Stat Report 2024 Jun;2024(205):1-11 [FREE Full text] [Medline: 38912919]
29. Miles MB, Huberman AM, Saldaña J. Qualitative Data Analysis: A Methods Sourcebook, 4th edition: Sage Publications, Inc; 2020.
30. Vindrola-Padros C, Johnson GA. Rapid techniques in qualitative research: a critical review of the literature. Qual Health Res 2020 Aug;30(10):1596-1604. [doi: 10.1177/1049732320921835] [Medline: 32667277]
31. Gale RC, Wu J, Erhardt T, et al. Comparison of rapid vs in-depth qualitative analytic methods from a process evaluation of academic detailing in the Veterans Health Administration. Implement Sci 2019 Feb 1;14(1):11. [doi: 10.1186/s13012-019-0853-y] [Medline: 30709368]
32. Higashi RT, Thakur B, Repasky EC, et al. Digital health technology use among Spanish speakers in the US: a scoping review. JAMA Netw Open 2025 May 1;8(5):e2510386. [doi: 10.1001/jamanetworkopen.2025.10386] [Medline: 40372754]
33. Riffin C, Cassidy J, Smith JM, et al. Care partner perspectives on the use of a patient portal intervention to promote care partner identification in dementia care. J Appl Gerontol 2025 Jan;44(1):136-146. [doi: 10.1177/07334648241262649] [Medline: 38901834]
34. Benmessaoud C, Pfisterer KJ, De Leon A, et al. Design of a dyadic digital health module for chronic disease shared care: development study. JMIR Hum Factors 2023 Dec 25;10:e45035. [doi: 10.2196/45035] [Medline: 38145480]
35. Jackson SL, Mejilla R, Darer JD, et al. Patients who share transparent visit notes with others: characteristics, risks, and benefits. J Med Internet Res 2014 Nov 12;16(11):e247. [doi: 10.2196/jmir.3363] [Medline: 25405911]
36. Gleason KT, Peereboom D, Wec A, Wolff JL. Patient portals to support care partner engagement in adolescent and adult populations: a scoping review. JAMA Netw Open 2022 Dec 1;5(12):e2248696. [doi: 10.1001/jamanetworkopen.2022.48696] [Medline: 36576738]
37. Office of the National Coordinator for Health Information Technology (ONC). Health IT legislation: 21st Century Cures Act. HealthIT.gov. URL: https://www.healthit.gov/topic/laws-regulation-and-policy/health-it-legislation [accessed 2023-04-14]
38. Gerber DE. 21st Century Cures Act: implementation without understanding implication? JCO Oncol Pract 2022 Feb;18(2):85-87. [doi: 10.1200/OP.21.00436] [Medline: 34436944]
39. Blease C, Salmi L, Rexhepi H, Hägglund M, DesRoches CM. Patients, clinicians and open notes: information blocking as a case of epistemic injustice. J Med Ethics 2021 May 14;48(10):785-793. [doi: 10.1136/medethics-2021-107275] [Medline: 33990427]
40. Durieux BN, DeCamp M, Lindvall C. 21st Century Cures Act: ethical recommendations for new patient-facing products. J Am Med Inform Assoc 2022 Sep 12;29(10):1818-1822. [doi: 10.1093/jamia/ocac112] [Medline: 35876830]
41. Brooks JV, Zegers C, Sinclair CT, et al. Understanding the Cures Act Information Blocking Rule in cancer care: a mixed methods exploration of patient and clinician perspectives and recommendations for policy makers. BMC Health Serv Res 2023 Mar 6;23(1):216. [doi: 10.1186/s12913-023-09230-z] [Medline: 36879318]
42. Petrovskaya O, Karpman A, Schilling J, et al. Patient and health care provider perspectives on patient access to test results via web portals: scoping review. J Med Internet Res 2023 Oct 19;25:e43765. [doi: 10.2196/43765] [Medline: 37856174]
43. Rodriguez JA, Shachar C, Bates DW. Digital inclusion as health care—supporting health care equity with digital-infrastructure initiatives. N Engl J Med 2022 Mar 24;386(12):1101-1103. [doi: 10.1056/NEJMp2115646] [Medline: 35302722]
44. Centers for Disease Control and Prevention. CDC Health Disparities and Inequalities Report—United States, 2011. MMWR Suppl 2011;60:1-116.
45. Mah JC, Penwarden JL, Pott H, Theou O, Andrew MK. Social vulnerability indices: a scoping review. BMC Public Health 2023 Jun 28;23(1):1253. [doi: 10.1186/s12889-023-16097-6] [Medline: 37380956]
46. Wilson S, Tolley C, Mc Ardle R, et al. Recommendations to advance digital health equity: a systematic review of qualitative studies. NPJ Digit Med 2024 Jun 29;7(1):173. [doi: 10.1038/s41746-024-01177-7] [Medline: 38951666]
47. Glasgow RE, Vogt TM, Boles SM. Evaluating the public health impact of health promotion interventions: the RE-AIM framework. Am J Public Health 1999 Sep;89(9):1322-1327. [doi: 10.2105/ajph.89.9.1322] [Medline: 10474547]
48. Shimoga SV, Lu YZ. Role of provider encouragement on patient engagement via online portals. J Am Med Inform Assoc 2019 Oct 1;26(10):968-976. [doi: 10.1093/jamia/ocz026] [Medline: 30925585]
49. Richwine C. Progress and persistent disparities in patient access to electronic health information. JAMA Health Forum 2023 Nov 3;4(11):e233883. [doi: 10.1001/jamahealthforum.2023.3883] [Medline: 37948063]
50. Richwine C, Johnson C, Patel V. Disparities in patient portal access and the role of providers in encouraging access and use. J Am Med Inform Assoc 2023 Jan 18;30(2):308-317. [doi: 10.1093/jamia/ocac227] [Medline: 36451262]
51. Hong YR, Xie Z, Nguyen OT, Turner K, Walker AF. Telehealth service use and quality of care among US adults with diabetes: a cross-sectional study of the 2022 Health Information National Trends Survey. BMJ Open 2024 Nov 3;14(10):e086418. [doi: 10.1136/bmjopen-2024-086418] [Medline: 39489536]
52. Swoboda C, Stamos A, Fareed N. Traditional and online health information seeking among individuals with limited English proficiency in the United States: cross-sectional study. Am J Health Promot 2025 Mar;39(3):469-478. [doi: 10.1177/08901171241302011] [Medline: 39568317]

XSL•FO
RenderX

53. Budhwani S, Fujioka J, Thomas-Jacques T, et al. Challenges and strategies for promoting health equity in virtual care: findings and policy directions from a scoping review of reviews. J Am Med Inform Assoc 2022 Apr 13;29(5):990-999. [doi: 10.1093/jamia/ocac022] [Medline: 35187571]
54. Estrela M, Semedo G, Roque F, Ferreira PL, Herdeiro MT. Sociodemographic determinants of digital health literacy: a systematic review and meta-analysis. Int J Med Inform 2023 Sep;177(105124):105124. [doi: 10.1016/j.ijmedinf.2023.105124] [Medline: 37329766]
55. Wanless B, Hassan N, McClellan C, et al. How do we better serve excluded populations when delivering digital health technology? Inclusion evaluation of a digital musculoskeletal self-management solution. Musculoskeletal Care 2024 Sep;22(3):e1925. [doi: 10.1002/msc.1925] [Medline: 39180193]
56. Jongebloed H, Anderson K, Winter N, et al. The digital divide in rural and regional communities: a survey on the use of digital health technology and implications for supporting technology use. BMC Res Notes 2024 Mar 28;17(1):90. [doi: 10.1186/s13104-024-06687-x] [Medline: 38549176]
57. Poncette AS, Spies C, Mosch L, et al. Clinical requirements of future patient monitoring in the intensive care unit: qualitative study. JMIR Med Inform 2019 Apr 30;7(2):e13064. [doi: 10.2196/13064] [Medline: 31038467]
58. Harry C, Goodday S, Chapman C, et al. Using social media to engage and enroll underrepresented populations: longitudinal digital health research. JMIR Form Res 2025 Apr 15;9:e68093. [doi: 10.2196/68093] [Medline: 40233355]
59. Livieri G, Mangina E, Protopapadakis ED, Panayiotou AG. The gaps and challenges in digital health technology use as perceived by patients: a scoping review and narrative meta-synthesis. Front Digit Health 2025;7(1474956):1474956. [doi: 10.3389/fdgth.2025.1474956] [Medline: 40212901]
60. Mumtaz H, Riaz MH, Wajid H, et al. Current challenges and potential solutions to the use of digital health technologies in evidence generation: a narrative review. Front Digit Health 2023;5:1203945. [doi: 10.3389/fdgth.2023.1203945] [Medline: 37840685]
61. Pew Research Center. Mobile fact sheet.: The Pew Charitable Trusts; 2024. URL: https://www.pewresearch.org/internet/fact-sheet/mobile/ [accessed 2023-04-15]

Abbreviations

REDCap: Research Electronic Data Capture
SRQR: Standards for Reporting Qualitative Research
UTSW: University of Texas Southwestern Medical Center

Edited by A Mavragani; submitted 16.12.24; peer-reviewed by A Kapoor, A Sheon; revised version received 29.05.25; accepted 02.06.25; published 31.07.25.

Please cite as:
Higashi RT, Repasky EC, Gupta A, Lee M, DesRoches CM, Israel A, Pruitt SL
Factors Associated With Portal and Telehealth Uptake and Use in a Minoritized, Low-Income Community: Mixed Methods Study
JMIR Form Res 2025;9:e70146
URL: https://formative.jmir.org/2025/1/e70146
doi:10.2196/70146

© Robin T Higashi, Emily C Repasky, Antara Gupta, MinJae Lee, Catherine M DesRoches, Aimee Israel, Sandi L Pruitt. Originally published in JMIR Formative Research (https://formative.jmir.org), 31.7.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Students' Perceptions of Learning Analytics for Mental Health Support: Qualitative Study

Aglaia Freccero[1], MSc, MPH; Miriam Onwunle[2], MPH; Jordan Elliott[1*], BSc; Nathalie Podder[3,4*], MSc; Julia Purrinos De Oliveira[5*], BSc; Lindsay H Dewa[2,6], PhD

[1]Division of Psychiatry, Department of Brain Sciences, Imperial College London, 2nd Floor Commonwealth Building, Du Cane Road, London, United Kingdom

[2]School of Public Health, Faculty of Medicine, Imperial College London, London, United Kingdom

[3]Institute of Security Science & Technology, Imperial College London, London, United Kingdom

[4]Department of Physics, Imperial College London, London, United Kingdom

[5]Department of Life Sciences, Faculty of Natural Sciences, Imperial College London, London, United Kingdom

[6]NIHR Biomedical Research Centre, Imperial College London, London, United Kingdom

[*]these authors contributed equally

Corresponding Author:
Aglaia Freccero, MSc, MPH
Division of Psychiatry, Department of Brain Sciences, Imperial College London, 2nd Floor Commonwealth Building, Du Cane Road, London, United Kingdom

Abstract

Background: Poor mental health among higher education students is a global public health concern. Learning analytics, which involves collecting and analyzing big data to support learning, could detect changes in behavior, learning patterns, as well as mental health and well-being. This could help inform mental health interventions in university settings. However, research has yet to explore students' perspectives on using learning analytics for mental health and well-being purposes.

Objective: This study aimed to explore students' perspectives on using learning analytics to support students' mental health and well-being at university.

Methods: Semistructured interviews were conducted online using Microsoft Teams between June and July 2023. Participants were identified through university student unions, social media, and snowball sampling. In total, 3 university students aged 20‐26 years joined our team and formed our student advisory group (SAG). They informed the design, analysis, and dissemination stages of the research cycle. Braun and Clarke's approach guided our thematic analysis. Data were triangulated by comparing codes from 2 transcripts across 2 independent researchers over a 2-hour online meeting. A coding framework was cocreated with the SAG to code the remaining transcripts and ensure data saturation. Themes were finalized and presented in a thematic map during a 2-hour meeting with the SAG and 2 researchers.

Results: In total, 15 participants were interviewed. We identified three main themes: (1) potential of learning analytics for mental health and well-being innovation, (2) student involvement in decision-making regarding learning analytics, and (3) integration of learning analytics with existing support. Despite being initially unaware, students recognized the potential of using learning analytics as a monitoring and early intervention tool to support university students' mental health. However, students raised concerns regarding data reliability and identified several ethical issues, such as privacy and lack of transparency. They also expressed the need to be involved in decision-making regarding learning analytics design, practices, and policies. Overall, students welcomed the possible integration of learning analytics with the existing university support.

Conclusions: This is the first qualitative study to explore students' perceptions of using learning analytics to support student mental health and well-being. Students' generally positive attitudes toward learning analytics suggest that this tool could be effectively integrated into the existing university support systems. Considering the ethical concerns raised by students, our findings suggest the need to bring the student voice into learning analytics development and implementation.

(JMIR Form Res 2025;9:e70327) doi:10.2196/70327

KEYWORDS

learning analytics; big data; e-learning; higher education; data mining; mental health; well-being

XSL•FO
RenderX

Introduction

Higher education students' mental health is worsening in the United Kingdom and is an increasing public health and policy concern. Students are particularly vulnerable to poor mental health due to academic, financial, and psychosocial pressures [1]. In a recent survey by the charity Student Minds, 57% of students reported experiencing a current mental health problem, and 27% had a diagnosed mental health condition, the most common being depression and anxiety [2]. Over the past 10 years, the number of students who disclosed a mental health condition to their institution has increased almost fivefold [3], reflecting rising need, improved awareness, reduced stigma, and clearer reporting pathways. Poor mental health can result in poor academic performance, study discontinuation, and even death by suicide [3]. However, universities struggle to meet the growing demand for mental health support, highlighting the need to adopt mental health as a priority [4]. Innovative solutions are required to effectively address this crisis while meeting students' mental health needs.

The integration of digital technology in higher education offers the opportunity to use students' data to detect signs of deteriorating mental health and provide targeted interventions [5,6]. Learning analytics refers to the collection, analysis, and reporting of large datasets (often referred to as "big data") of student data to understand and support student learning [7]. Learning analytics systems draw on data from various university systems, including student information systems, virtual learning environments, library usage, and attendance records [8,9]. While there is compelling evidence that learning analytics interventions can improve student outcomes (eg, retention, academic performance, and engagement) [10], there is limited research on their impact on mental health and well-being [11-13]. Changes in learning behavior may indicate changes in student mental health, although the current evidence remains scarce [14-16].

Few qualitative studies have explored students' perceptions of learning analytics to support learning [17-25]. There is also a lack of research across different academic disciplines and education levels [17,19,20]. Only 3 studies focused on more than 1 institution [17-19], and the only study conducted in the United Kingdom used a small sample of final-year students from a single discipline [21]. Overall, these studies report that students generally welcome the use of learning analytics to enhance their educational experience [17-21], but also raise concerns about possible privacy breaches, lack of consent, equity issues, and poor data reliability [22-26].

Learning analytics has the potential not only to improve academic outcomes but also to identify and address students' mental health and well-being needs in order to provide better support services [11-13]. Nonetheless, students are often not actively involved in decision-making regarding their implementation [4]. Understanding students' views and engaging them in consultation is essential to inform learning analytics practices and the acceptable adoption of learning analytics for mental health purposes in higher education [15,16,27-30]. However, no studies have explored university students' perceptions of using learning analytics for mental health and well-being. This study aims to fill this gap by exploring students' perspectives on using learning analytics to support students' mental health and well-being in university settings.

Methods

Study Design and Methodological Orientation

The study followed the Consolidated Criteria for Reporting Qualitative Research (COREQ) checklist (Checklist 1).

This study used a qualitative methodology to explore a "how" research question, focusing on the depth and complexity of an underresearched topic. Guided by a relativist ontology, subjectivist epistemology, and an interpretivist paradigm [31], our approach is well-suited to research in psychology, mental health, and education, where lived experience and sociocultural context are central [32,33]. A grounded theory methodology supported the inductive development of insights from participants' perspectives [34]. Semistructured interviews were used to gather rich, nuanced data, aligned with the study's exploratory aims [35].

Participant Selection

Participants were identified through student networks, social media platforms (Instagram [Meta], X, and LinkedIn), and snowball sampling, an approach well-suited for exploring sensitive topics and effective in reaching diverse student populations, including international students [32,33]. Interested participants were encouraged to email the researcher to express their interest. Participants were eligible to take part if they were students (aged 18‑25 y) enrolled in an undergraduate or postgraduate course at a university in London, fluent English speakers, and willing to provide informed written consent. The age range represented students enrolled in undergraduate or postgraduate taught programs, which typically involve more structured online learning. Efforts were made to include diversity in gender, ethnicity, sexual orientation, degree, and year of study. All individuals who expressed interest and were contacted by the researcher met the eligibility criteria; none declined to participate or dropped out of the study. Participation was voluntary, and as this study was part of a master's thesis, no compensation was provided.

Data Generation

Participants' demographic data were collected at recruitment (Table 1). Semistructured interviews were conducted by the lead researcher (AF) via Microsoft Teams between June 6 and July 27, 2023. Open-ended questions explored participants' knowledge, attitudes, and concerns regarding the use of learning analytics to support university students' mental health and well-being. The semistructured interview questions were informed by a review of literature on student mental health, learning analytics, and help-seeking in higher education. They were also guided by the technology acceptance model [36] to explore perceptions of learning analytics and self-determination theory [37] to consider factors influencing student motivation and engagement. Furthermore, 3 university students informed the topic guide (refer to Patient and Public Involvement section). It consisted of five sections: (1) introductory questions to help

XSL•FO
RenderX

participants feel comfortable and at ease, (2) existing mental health support at university, (3) use of learning analytics to support university students' mental health and well-being, (4) facilitators and barriers to learning analytics adoption for mental health and well-being purposes, and (5) integration of learning analytics within existing university mental health support services (Appendix S1 in Multimedia Appendix 1). All interviews were audio- and video-recorded, then transcribed verbatim. No one else was present during the interview except for the participants and the researcher. Transcripts were accessible to participants through Microsoft Teams. Field notes were taken during and after the interview and were considered when coding the transcripts.

Table . Participants' demographics.

Characteristics	Value
Age (y), mean (SD)	21.47 (1.60)
Gender, n (%)	
Female	10 (67)
Male	5 (33)
Other	0 (0)
Ethnicity, n (%)	
White[a]	5 (33)
Asian or Asian British[b]	5 (33)
Mixed[c]	3 (20)
Black or Black British[d]	1 (7)
Other[e]	1 (7)
Qualification, n (%)	
Undergraduate	11 (73)
Postgraduate	4 (27)
Fee status, n (%)	
Home[f]	8 (53)
International[g]	7 (47)

[a]White: English, Welsh, Scottish, Northern Irish, British, Irish, Gypsy or Irish Traveller, Roma, or any other White background.

[b]Asian or Asian British: Indian, Pakistani, Bangladeshi, Chinese, or any other Asian background.

[c]Mixed or multiple ethnic groups: White and Black Caribbean, White and Black African, White and Asian, or any other mixed or multiple ethnic group.

[d]Black or Black British: Caribbean, African, or any other Black, Black British, or Caribbean background.

[e]Other ethnic group: Arab or any other ethnic group.

[f]Home fee status: students ordinarily resident or settled in the United Kingdom.

[g]International fee status: overseas or international students not subject to home fee status.

Patient and Public Involvement

We advertised for individuals with lived experience (ie, university students whose data were being used by the university) through the Imperial College Union Mental Health Network Instagram account to join our team as advisors and inform the project. As a result, 3 students, aged 21 - 26 years (2 female and 1 male), from the Faculty of Natural Sciences and the Faculty of Medicine at Imperial College London expressed interest and joined the team to form a student advisory group (SAG). The SAG included 1 postgraduate and 2 undergraduate students and informed 3 key stages of the research cycle [38]: designing and managing research, carrying out the research study, and dissemination.

We held a 2-hour online meeting (Microsoft Teams) on April 7, 2023, where the SAG informed the research objectives and accessibility and appropriateness of project documentation (interview topic guide, participant information sheet, consent form, and protocol). The following changes were made to the topic guide: language was simplified (ie, a lay definition of learning analytics was agreed upon), and a question was added regarding the impact of learning analytics on academic outcomes. The final topic guide was piloted with the SAG in 3 mock interviews. After providing all positive feedback, no subsequent changes were made to the topic guide. Following the generation of initial codes, a coding framework was cocreated with the SAG during an additional 2-hour online meeting (Microsoft Teams) on July 14, 2023. We held a final meeting on August 2, 2023, where the SAG and the 2 researchers (AF and MO) refined and consolidated themes. Several themes were revised, consolidated, and more effectively linked through student advisor involvement (Figure S1 in

XSL•FO
RenderX

Multimedia Appendix 1). The SAG was not compensated for their time, as no budget was associated with this study, being a master's thesis. Patient and public involvement (PPI) was reported using the Guidance for Reporting Involvement of Patients and the Public, Version 2-short version [39] (Table S1 in Multimedia Appendix 1).

Data Analysis

An inductive and deductive method to Braun and Clarke's 6-stage thematic analysis guided our analysis [40]. We used a data-driven approach to theme development while also drawing from the researcher's previous knowledge of learning analytics. Transcripts were imported into Microsoft Word. After transcription, the lead researcher (AF) familiarized herself with the data by reading each transcript multiple times. All transcripts were manually coded in Microsoft Word, and relevant codes were compiled. Manual coding was chosen as it supported reflexivity, the development of the lead researcher's skills, and enabled deep immersion in the data [41-43]. Furthermore, 2 independent researchers (AF and MO) compared codes from 2 transcripts during a 2-hour online meeting on Microsoft Teams. Data triangulation helped ensure the reliability and validity of the findings by identifying potential individual researcher biases and confirming the consistency of information. The coding framework was applied to code the remaining transcripts. Interviews were conducted until data saturation was reached, assessed through concurrent data analysis, so that no new codes emerged, and responses across participants consistently reinforced existing patterns. AF then organized codes into potential themes and subthemes and gathered all data relevant to each of these. Some themes were combined to avoid redundancy, while others were broken down into separate themes and judged for internal homogeneity and external heterogeneity. The thematic map was finalized with the SAG.

Ethical Considerations

All procedures involving human participants were approved by the Head of Department of the School of Public Health and the Research Governance and Integrity Team at Imperial College London (Imperial College Research Ethics Committee reference 6604462). The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008.

Participants provided informed consent by signing and returning a consent form via email. A copy co-signed by both the participant and the researcher was emailed back to the participant and stored securely on Imperial College London encrypted drives accessible only to the study team. Participants could withdraw at any time before the start of data analysis. Audio and video recordings were deleted after transcription, and study-specific ID numbers were used to maintain participants' anonymity. The principal investigator ensured confidentiality and General Data Protection Regulation compliance for health and care research. Participation was voluntary and uncompensated, as the study formed part of a master's thesis.

Results

Participants' Characteristics

In total, 15 participants took part in the study (10 female, 5 male), with a mean age of 21.47 (SD 1.60, range 19 - 24) years (Table 1). Students from 4 London universities participated in the interview (Table S2 in Multimedia Appendix 1). Most participants were undergraduate students (11/15, 73%). All students were enrolled in a science, technology, engineering, and mathematics (STEM) degree. Interviews lasted between 19 and 34 minutes.

Themes and Subthemes

There were three themes: (1) potential of learning analytics for mental health and well-being innovation, (2) student involvement in decision-making regarding learning analytics, and (3) integration of learning analytics with existing support (Figure 1).

XSL•FO
RenderX

Figure 1. Thematic map displaying themes (dark blue), subthemes (light blue), and their relationships. Red arrows represent relationships between themes. Blue arrows represent the relationships between themes and their corresponding subthemes. LA: learning analytics.

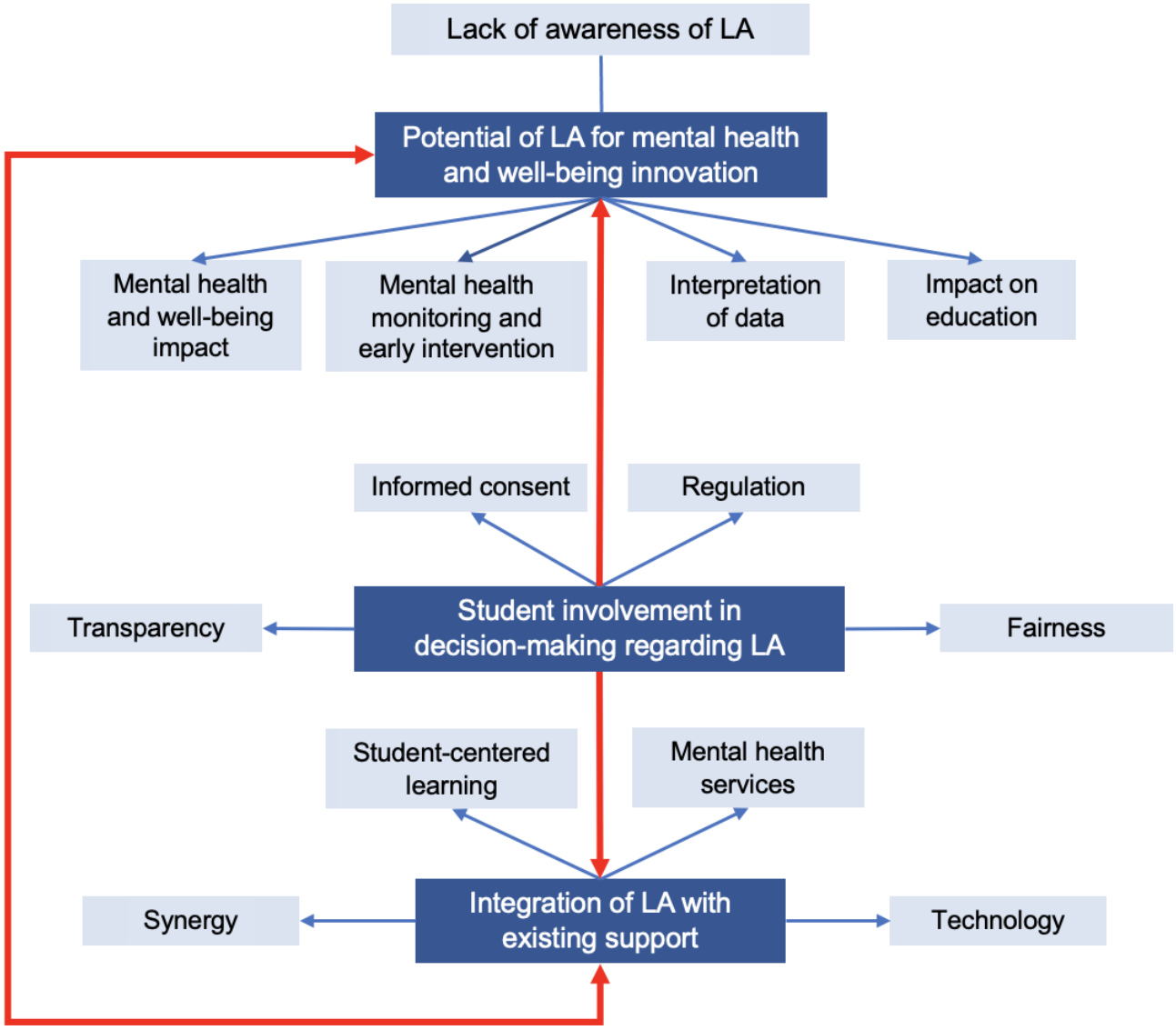

Potential of Learning Analytics for Mental Health and Well-Being Innovation

At the beginning of the interview, all participants were generally unaware of learning analytics. While they were confident that their university had collected their data, such as their grades and feedback, most participants were unsure how these data were used. After providing them with an accurate definition of learning analytics (Appendix B in Multimedia Appendix 1), all participants felt more confident in their understanding of learning analytics and then recognized the potential impact of learning analytics to go beyond academic purposes.

> *I think it has potential […] being used as an early indicator to intervene with mental health issues. I think when used correctly, you could potentially help students get the support they need.* [ID10, female]

Specifically, they indicated learning analytics might serve as a tool for monitoring students' mental health and well-being and informing appropriate intervention. Some students said that learning data collected by learning analytics could flag disengagement with studies, which might indicate a worsening in mental health.

> *So let's say you've never turned up to a lecture, that's a red flag and that can be used by the mental health services… to intervene.* [ID13, female]

They also highlighted the potential of learning analytics to serve as an early warning system for mental health issues and inform early intervention. Suggested interventions included automated messages via email or virtual learning environments, as well as direct contact from staff, such as mental health support teams and academic tutors, to check in with students in case of disengagement from studies and to signpost them to appropriate support.

> *If someone is not attending because they are not doing well, having an email from a welfare perspective…may be a way to intervene a little bit earlier.* [ID15, female]

Most students agreed that higher engagement with studies generally suggests better academic performance and that mental health and academic success are closely interconnected. They

XSL•FO
RenderX

recognized that mental health difficulties could negatively impact academic engagement and performance, and that this underscored the need for learning analytics to consider these links when supporting students.

> *It (learning analytics) would improve academic outcomes like through the well-being route […] because you know there's no good academic outcome without good mental health.* [ID4, female]

However, some students were skeptical about whether changes in learning behavior truly reflect changes in mental health. They reported concerns about the difficulty of interpreting data and how students have varied learning behaviors and mental health symptoms manifest differently from person to person.

> *I can see why from that sort of behavioural data you can make conclusions about the learning…I'm just not sure how representative that is of one's mental health.* [ID14, female]

For instance, 1 participant suggested that some students might disengage from their studies when struggling, while others may immerse themselves in work as a coping mechanism. A few participants emphasized the importance of contextualizing behavioral changes and personalizing learning analytics to account for individual differences.

> *With mental health…you can't be definite. It's so broad between individuals and not every mental health issue presents the same.* [ID10, female]

Student Involvement in Decision-Making Regarding Learning Analytics

Most participants emphasized the need for clear, accessible, and transparent communication with students regarding the use of learning analytics for mental health and well-being. Many students reported that they should be made fully aware of what data will be collected, who will have access, where it will be stored, and how it will be used.

> *Making it clear…why and what (data) is being collected, and that's for your benefit.* [ID8, male]

Some participants said better transparency would allow students to build trust in learning analytics, understand its benefits, and increase engagement with learning analytics systems.

> *There needs to be transparency between the university and the students, otherwise…you can get to a point where students don't necessarily trust the university.* [ID13, female]

Importantly, many students stressed the importance of involving students as stakeholders in decision-making regarding the future development and implementation of learning analytics to ensure that learning analytics systems accurately reflect the realities of students' mental health experiences.

> *The thing I'm going to stress is involving the relevant stakeholders in deciding who's going to have this data. How is it going to be used? What intervention, so what are you going to do when you identify these trends? Is that something that would be considered acceptable for students and staff and would it work? Would it also be helpful for those who find it acceptable?* [ID15, female]

One student mentioned this involvement could mean coproducing learning analytics systems between students and university staff. Finally, 1 participant believed implementing learning analytics would require continuous review and improvement based on student feedback to maximize its benefits for mental health and better meet the needs of students.

However, some students highlighted barriers to learning analytics adoption, including the potential hesitancy among students to give informed consent due to privacy concerns.

> *Probably the main issue would be students not wanting to consent based on privacy.* [ID6, female]

A few students mentioned that consent is usually given without the students' full understanding and approval. Consent is often buried in long documents that students rarely read, leaving them uneasy about how their data are being used. A few suggested offering opt-in and opt-out options to maintain freedom of choice regarding data sharing. Ethical and regulatory issues were also reported. Some students voiced worries about learning analytics being used as a tool for inappropriate monitoring, which could lead to potential privacy breaches and invasive interventions.

> *If some people have issues with being monitored, … the data processing and being spied on…they might find it actually uncomfortable… and maybe even a little bit threatened.* [ID1, female]

In addition, a few students reported that learning analytics interventions could come across as accusatory and might unfairly impact students' grades, for example, if low engagement were to be detected by learning analytics.

> *…as long is not held against me like "ohh this will impact you…on your grades..."* [ID8, male]

Therefore, they expressed the desire for learning analytics to be implemented equitably for all students, despite their level of engagement and academic performance, ensuring that the support learning analytics might offer is uniform and nondiscriminatory.

Integration of Learning Analytics With Existing Support

A common issue expressed by several participants was the lack of motivation for students to engage with learning analytics, such as attendance tracking or feedback. However, a few participants suggested monetary incentives to help engage students with learning analytics. Engagement was regarded to be dependent on the perceived relevance and ease of use of learning analytics systems. As such, several participants noted that learning analytics must be effortlessly integrated with existing academic practices.

> *I think take the amount of effort away from the students as much as possible, or else they won't want to participate.* [ID5, male]

For example, many students reported that engagement with learning analytics could be useful in assessing students' learning journeys and responses to workload or assessments.

XSL•FO
RenderX

> *…they can figure out which way is best for students' learning, and they could then be able to personalise that a bit more.* [ID13, female]

Therefore, some believed it would help inform how to adapt teaching practices to improve students' academic outcomes and experiences, which could motivate students to engage with learning analytics.

> *As college knows at which period students feel most stressed, they can change their teaching objectives and make students feel less stressed.* [ID12, female]

Many students believed that learning analytics could be integrated not only within existing academic practices but also with existing mental health support. Therefore, learning analytics could be an innovative tool to link academic and welfare services, providing more holistic support for students.

> *Academics and welfare are really closely linked and actually…you asked me….what would I suggest as improvements, and one more thing to add there is integrating these two services. So why is welfare completely separate?* [ID15, female]

However, some highlighted concerns about implementing this integration. For example, some felt that implementing learning analytics to link services would require time, resources, and service restructuring.

> *I feel like in theory it should help improve it [mental health]. But I guess it just depends how they use them.* [ID11, male]

Many students identified artificial intelligence (AI) as a potential solution for tracking students' mental health and implementing timely interventions with minimal effort from academic staff. For example, they suggested that algorithms adjusted to individual differences could offer personalized recommendations to maximize mental health. However, some participants expressed skepticism about the accuracy of these automated systems and that they would be overly intrusive in their lives.

> *Maybe just an AI system that keeps updating what their baseline should be and so any deviation from that baseline of normal someone could kind of alert a member of staff.* [ID4, female]

Discussion

Key Findings and Comparison With Previous Work

To the authors' knowledge, this is the first study to explore students' perceptions of learning analytics to support students' mental health and well-being. There were three main themes: (1) potential of learning analytics for mental health and well-being innovation, (2) student involvement in decision-making regarding learning analytics, and (3) integration of learning analytics with existing support.

All participants supported the use of learning analytics as a tool to support students' mental health and well-being. The majority of participants felt that learning analytics could enhance access to and delivery of support services, advocating for an evidence-based and student-centered approach, in line with existing qualitative literature and policy recommendations on improving student mental health support [35-38]. However, some students also recognized potential challenges in using learning analytics to detect and intervene in case of student mental health deterioration. Overall, our findings show that students' positive attitudes and acceptability of learning analytics hold promise for implementation within current university support systems. Learning analytics may help enhance coordination between academic and mental health services, supporting UK policy recommendations to adopt a whole-university approach to mental health [4].

Despite an initial lack of awareness of learning analytics, all students recognized its potential beyond the learning context to support mental health and well-being. This aligns with qualitative literature that reports students' generally positive attitudes toward learning analytics [17,25] and digital tools for mental health monitoring and early intervention in young people [44-46]. Most participants believed that learning analytics could provide objective and reliable evidence for monitoring students' mental health. They also welcomed its potential to inform early interventions and integrate with existing university support systems. However, they also recognized the potential challenges of using learning analytics for mental health purposes and stressed the need to be involved in decision-making about learning analytics development and implementation.

Several barriers and facilitators were identified for adopting learning analytics to support students' mental health and well-being. Some students raised concerns regarding the reliability of learning data and identified several ethical issues, in line with other qualitative studies [14,17,18,25]. Students worried that data inaccuracy could lead to delivering inappropriate interventions, highlighting the need for a context-specific understanding of students' behavior. Students also reported the importance of balancing the need to share data for effective interventions with concerns about privacy breaches and ensuring informed consent, reflecting findings from other studies [20,26,47-49] and aligning with proposed recommendations for ethics and privacy in learning analytics [50,51]. Participants agreed these risks might subside if universities were more transparent about using learning analytics and abided by good practice guidelines [14,15,28,30].

Overall, students' ethical concerns underscore the need to adopt a coproduced approach in the development, implementation, and governance of learning analytics that strategically incorporates the student voice and aligns with the principles of PPI [15,16]. Meaningfully engaging students as active stakeholders, beyond their role as data subjects, is essential to fostering trust and ensuring that learning analytics systems are transparent, ethical, and responsive to student needs [27-30]. Regarding facilitators for learning analytics adoption, students stressed the need for incentives to engage with learning analytics, as well as staff inclusion and training, in line with other studies [19,21,23,52]. Our results support previous research [17,26], showing that learning analytics can be used to assess students' learning methods and adjust teaching approaches and curricula to better support their mental health and academic needs. AI-driven computational approaches, including machine learning, natural language processing, and rule-based systems, were viewed positively. These methods

XSL•FO
RenderX

were considered valuable additions to learning analytics systems, offering predictive capabilities and data-informed recommendations to enhance student learning and support students' mental health when implemented with appropriate student involvement, as highlighted in a recent systematic review [53]. This aligns with qualitative research reporting young people's acceptance of AI [54]. Therefore, our study suggests that students' positive views on learning analytics support the potential for integrating learning analytics with existing academic and mental health support, as per UK policy recommendations [4].

Reflexivity

The researcher's perspective was informed by her lived experience as a student with mental health issues (Appendix S2 in Multimedia Appendix 1). This insider positionality facilitated access, built rapport, and enabled the development of informed, contextually relevant questions. However, it also introduced potential for unconscious bias, including overidentification with participants and possible influence from pre-existing views on university mental health provision. Furthermore, the researcher's involvement in mental health advocacy within her institution may have led participants to have some awareness of her motivations for conducting the research.

To address these biases, a reflexive approach was used throughout the study to critically examine how personal assumptions, values, and previous knowledge might influence design, data collection, and interpretation [55,56]. Attention to the insider-outsider dynamic, reflexive journaling, PPI, and data triangulation with other researchers was used to ensure transparency and analytical rigor, grounding the findings in participants' perspectives rather than the researcher's personal views.

Strengths and Limitations

There are several strengths to this study. A total of 3 people with lived experience and interest in mental health research informed the design, analysis, and interpretation research stages [57]. Students had personal knowledge and experience of mental health support at university, which made the research more relevant, focused, and accessible.

There are also study limitations. Despite efforts to maximize variation, participants' backgrounds were not diverse in terms of gender and sexual orientation. Our sample was predominantly female, and we did not include individuals who did not identify with the gender assigned at birth. We also did not explore differences in sexual orientation, and our sample was predominantly heterosexual. LGBTQ+ students are particularly vulnerable to mental health problems [58], and their views would enrich the field. While the sample included both home and international students from a range of cultural and ethnic backgrounds, the study did not explicitly explore how cultural factors shape mental health experiences or help-seeking behaviors [59]. Furthermore, although participants came from multiple London universities, they were all STEM students, so our results may not be generalizable to other universities in the United Kingdom or students from other educational backgrounds. Future research could incorporate focus groups to explore the topic in a more naturalistic setting, thereby validating findings and minimizing researcher bias. Larger studies are also needed to enhance generalizability. Quantitative approaches, such as exploratory factor analysis, could be used to examine broader patterns based on the themes identified in this study.

Implications and Policy Recommendations

This research focused on students' perceptions of learning analytics to support students' mental health and well-being. To ensure the successful adoption of learning analytics systems to support students' mental health and well-being in higher education in the future, there is now a need to explore the perceptions of all other stakeholders, including academic staff, counseling and disability services, and IT services [60]. Our qualitative findings could also be complemented with a national survey across all UK universities and internationally to capture broader perspectives on the interview themes and explore relationships between the data. As learning analytics systems are implemented, future research should investigate the feasibility, acceptability, and effectiveness of these systems in improving students' mental health and well-being.

There are also several practical and policy implications to this study. First, students' general positive perceptions around the use and integration of learning analytics for mental health reflect the need to align learning analytics practices with mental health strategic priorities as part of a whole-university and data-driven approach to mental health [4]. Second, students' ethical concerns regarding learning analytics adoption, particularly around transparency, consent, and privacy, highlight the importance of involving students in their development and implementation, echoing previous policy calls [15,16,27-30]. Universities should, therefore, develop clear institutional policies and protocols coproduced with students regarding the development and implementation of learning analytics. Third, institutional learning analytics policies should align with national higher-education–wide guidelines to ensure sector-wide consistency. A coordinated government strategy on student mental health is needed to align institutional practices with broader policy goals. Ultimately, while learning analytics offer strong potential to enhance student support and service delivery, their success relies on meaningful student involvement, institutional accountability, and coherent policy at both institutional and national levels.

Conclusion

This was the first study to explore students' perceptions of using learning analytics to support students' mental health and well-being. Our findings suggest that students perceive value in utilizing learning analytics to enhance mental health and well-being in higher education, as well as their integration with existing university services, which holds significant importance for UK policy. However, considering the ethical concerns raised by students, this study also highlights the need to inform students about learning analytics–based systems being developed or implemented at their institution, echoing previous calls for student engagement in the decision-making process regarding learning analytics. Ensuring learning analytics systems meet students' needs and expectations is essential to avoid resistance

XSL•FO
RenderX

to learning analytics adoption and empower students to make decisions regarding their welfare and educational experience. Overall, learning analytics holds promises for its integration within university support, but not without the involvement of students.

Acknowledgments

We are grateful for support from the National Institute for Health Research under the Applied Research Collaboration program for North West London (NIHR ARC NWL) and the Imperial National Institute for Health Research Biomedical Research Centre (NIHR BRC). The views expressed are those of the authors and not necessarily those of Imperial NIHR BRC or NHIR ARC NWL.

Data Availability

The interview data generated and analyzed during this study are not publicly available due to the maintenance of confidentiality and the presence of identifiable participant information. Deidentified data are available from the corresponding author upon reasonable request.

Authors' Contributions

AF and LHD designed the study. AF was responsible for the overall organization and coordination of the study. AF conducted data collection and processing. AF and MO were responsible for data analysis. LHD was the chief investigator and oversaw all study aspects. JE, JPDO, and NP contributed equally to informing the research design, data analysis, interpretation of results, and future dissemination as members of a student advisory group (SAG). All authors contributed to the writing of the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary materials include the interview topic guide, figures of thematic map development, documentation of patient and public involvement activities, a table of full participant demographics, and a reflexivity statement.
[DOC File, 642 KB - formative_v9i1e70327_app1.doc]

Checklist 1
Consolidated Criteria for Reporting Qualitative Research (COREQ) checklist.
[DOCX File, 19 KB - formative_v9i1e70327_app2.docx]

References

1. Campbell F, Blank L, Cantrell A, et al. Factors that influence mental health of university and college students in the UK: a systematic review. BMC Public Health 2022 Sep 20;22(1):1778. [doi: 10.1186/s12889-022-13943-x] [Medline: 36123714]
2. Student Minds research briefing – February ‘23. Student Minds. 2023. URL: https://www.studentminds.org.uk/uploads/3/7/8/4/3784584/student_minds_insight_briefing_feb23.pdf [accessed 2025-06-04]
3. Thorley C. Not by degrees: improving student mental health in the UK’s universities. Institute for Public Policy Research. 2017. URL: https://www.ippr.org/articles/not-by-degrees [accessed 2025-06-04]
4. Stepchange: mentally healthy universities. Universities UK. 2017. URL: https://www.universitiesuk.ac.uk/what-we-do/policy-and-research/publications/stepchange-mentally-healthy-universities [accessed 2025-06-04]
5. Baig MI, Shuib L, Yadegaridehkordi E. Big data in education: a state of the art, limitations, and future research directions. Int J Educ Technol High Educ 2020 Dec;17(1). [doi: 10.1186/s41239-020-00223-0]
6. Chatti MA, Muslim A, Schroeder U. Toward an Open Learning Analytics Ecosystem Big Data and Learning Analytics in Higher Education: Springer; 2017:195-219. [doi: 10.1007/978-3-319-06520-5_12]
7. Siemens G, Gasevic D. Guest editorial - learning and knowledge analytics. Educational Technology & Society 2012;15(3):1-2 [FREE Full text]
8. Samuelsen J, Chen W, Wasson B. Integrating multiple data sources for learning analytics—review of literature. RPTEL 2019 Dec;14(1). [doi: 10.1186/s41039-019-0105-4]
9. Wong BT, Li KC. A review of learning analytics intervention in higher education (2011–2018). J Comput Educ 2020 Mar;7(1):7-28. [doi: 10.1007/s40692-019-00143-7]
10. Scatler N. Learning Analytics Explained: London: Routledge; 2017.

XSL•FO
RenderX

11. Bodily R, Verbert K. Trends and issues in student-facing learning analytics reporting systems research. Presented at: LAK '17: Proceedings of the Seventh International Learning Analytics & Knowledge Conference; Mar 13-17, 2017; Vancouver British Columbia Canada. [doi: 10.1145/3027385.3027403]
12. Foster C, Francis P. A systematic review on the deployment and effectiveness of data analytics in higher education to improve student outcomes. Assessment & Evaluation in Higher Education 2020 Aug 17;45(6):822-841. [doi: 10.1080/02602938.2019.1696945]
13. Viberg O, Hatakka M, Bälter O, Mavroudi A. The current landscape of learning analytics in higher education. Comput Human Behav 2018 Dec;89:98-110. [doi: 10.1016/j.chb.2018.07.027]
14. Tsai YS, Whitelock-Wainwright A, Gašević D. More than figures on your laptop: (dis)trustful implementation of learning analytics. JLA 2021;8(3):81-100. [doi: 10.18608/jla.2021.7379]
15. SHEILA framework. SHEILA Project. 2018. URL: http://sheilaproject.eu/wp-content/uploads/2018/08/SHEILA-framework_Version-2.pdf [accessed 2025-06-04]
16. Piper R, Emmanuel T. Co-producing mental health strategies with students: a guide for the higher education sector. Student Minds. URL: https://www.studentminds.org.uk/uploads/3/7/8/4/3784584/cpdn_document_artwork.pdf [accessed 2025-06-04]
17. Gray G, Schalk AE, Cooke G, Murnion P, Rooney P, O'Rourke KC. Stakeholders' insights on learning analytics: perspectives of students and staff. Comput Educ 2022 Oct;187:104550. [doi: 10.1016/j.compedu.2022.104550]
18. Heiser R, Dello Stritto ME, Brown A, Croft B. Amplifying student and administrator perspectives on equity and bias in learning analytics. Learning Analytics 2023;10(1):8-23. [doi: 10.18608/jla.2023.7775]
19. Jones KML, Asher A, Goben A, et al. "We're being tracked at all times": Student perspectives of their privacy in relation to learning analytics in higher education. Asso for Info Science & Tech 2020 Sep;71(9):1044-1059. [doi: 10.1002/asi.24358]
20. Roberts LD, Howell JA, Seaman K, Gibson DC. Student attitudes toward learning analytics in higher education: "The Fitbit Version of the Learning World". Front Psychol 2016;7:1959. [doi: 10.3389/fpsyg.2016.01959] [Medline: 28066285]
21. Bennett L, Folley S. Four design principles for learner dashboards that support student agency and empowerment. JARHE 2019 May 13;12(1):15-26. [doi: 10.1108/JARHE-11-2018-0251]
22. Lim LA, Dawson S, Gašević D, et al. Students' perceptions of, and emotional responses to, personalised learning analytics-based feedback: an exploratory study of four courses. Assessment & Evaluation in Higher Education 2021 Apr 3;46(3):339-359. [doi: 10.1080/02602938.2020.1782831]
23. Roberts LD, Howell JA, Seaman K. Give me a customizable dashboard: personalized learning analytics dashboards in higher education. Tech Know Learn 2017 Oct;22(3):317-333. [doi: 10.1007/s10758-017-9316-1]
24. Schumacher C, Ifenthaler D, Azevedo R, Cromley JG, Seibert D, Moos DC. Features students really expect from learning analytics. Comput Human Behav 2018 Jan;78:397-407. [doi: 10.1016/j.chb.2017.06.030]
25. Rets I, Herodotou C, Bayer V, Hlosta M, Rienties B. Exploring critical factors of the perceived usefulness of a learning analytics dashboard for distance university students. Int J Educ Technol High Educ 2021;18(1):46. [doi: 10.1186/s41239-021-00284-9] [Medline: 34778534]
26. Knight DB, Brozina C, Novoselich B. An investigation of first-year engineering student and instructor perspectives of learning analytics approaches. Learning Analytics 2016;3(3):215-238. [doi: 10.18608/jla.2016.33.11]
27. Ahern SJ. Making a #Stepchange? Investigating the alignment of learning analytics and student wellbeing in United Kingdom higher education institutions. Front Educ 2020 Nov 4;5. [doi: 10.3389/feduc.2020.531424]
28. Cormack AN, Reeve D. Developing a code of practice for using data in wellbeing support. Learning Analytics 2022;9(2):253-264. [doi: 10.18608/jla.2022.7533]
29. Code of practice for learning analytics. Jisc. URL: https://www.jisc.ac.uk/guides/code-of-practice-for-learning-analytics [accessed 2025-06-04]
30. Sclater N. Developing a code of practice for learning analytics. Learning Analytics ;3(1). [doi: 10.18608/jla.2016.31.3]
31. Denzin NK, Lincoln YS. The Sage Handbook of Qualitative Research, 4th edition: London: Sage; 2011.
32. Badu E, O'Brien AP, Mitchell R. An integrative review on methodological considerations in mental health research - design, sampling, data collection procedure and quality assurance. Arch Public Health 2019;77(1):37. [doi: 10.1186/s13690-019-0363-z] [Medline: 31624592]
33. Cohen L, Manion L, Morrison K. Research Methods in Education, 8th edition: Routledge; 2018. [doi: 10.4324/9781315456539]
34. Urquhart C. Grounded Theory for Qualitative Research: A Practical Guide: SAGE Publication; 2013.
35. Rubin H, Rubin I. Qualitative Interviewing: The Art of Hearing Data: SAGE Publishers; 2012.
36. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989 Sep;13(3):319. [doi: 10.2307/249008]
37. Ryan RM, Deci EL. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. Am Psychol 2000 Jan;55(1):68-78. [doi: 10.1037//0003-066x.55.1.68] [Medline: 11392867]
38. NIHR. Standards for public involvement in research—better public involvement for better health and social care research. Invo. 2019. URL: https://www.invo.org.uk/wp-content/uploads/2019/11/UK-standards-for-public-involvement-v6.pdf [accessed 2025-06-04]

XSL•FO
RenderX

39. Staniszewska S, Brett J, Simera I, et al. GRIPP2 reporting checklists: tools to improve reporting of patient and public involvement in research. BMJ 2017:j3453. [doi: 10.1136/bmj.j3453]
40. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
41. Saldana J. The Coding Manual for Qualitative Researchers: SAGE Publication; 2021.
42. Berger R. Now I see it, now I don't: researcher's position and reflexivity in qualitative research. Qual Res 2015 Apr;15(2):219-234. [doi: 10.1177/1468794112468475]
43. Nowell LS, Norris JM, White DE, Moules NJ. Thematic analysis: striving to meet the trustworthiness criteria. Int J Qual Methods 2017;16(1). [doi: 10.1177/160940691773384]
44. Armstrong CC, Odukoya EJ, Sundaramurthy K, Darrow SM. Youth and provider perspectives on behavior-tracking mobile apps: qualitative analysis. JMIR Ment Health 2021 Apr 22;8(4):e24482. [doi: 10.2196/24482] [Medline: 33885364]
45. Wies B, Landers C, Ienca M. Digital mental health for young people: a scoping review of ethical promises and challenges. Front Digit Health 2021;3:697072. [doi: 10.3389/fdgth.2021.697072] [Medline: 34713173]
46. Dewa LH, Lavelle M, Pickles K, et al. Young adults' perceptions of using wearables, social media and other technologies to detect worsening mental health: a qualitative study. PLoS ONE 2019;14(9):e0222655. [doi: 10.1371/journal.pone.0222655] [Medline: 31532786]
47. Ifenthaler D, Schumacher C. Student perceptions of privacy principles for learning analytics. Education Tech Research Dev 2016 Oct;64(5):923-938. [doi: 10.1007/s11423-016-9477-y]
48. Prinsloo P, Slade S. Student privacy self-management: implications for learning analytics. Presented at: LAK '15: Proceedings of the Fifth International Conference on Learning Analytics And Knowledge; Mar 16-20, 2015; Poughkeepsie New York. [doi: 10.1145/2723576.2723585]
49. Slade S, Prinsloo P. Student perspectives on the use of their data: between intrusion, surveillance and care. Presented at: Proceedings of the European Distance and E-Learning Network 2014 Research Workshop; Oct 27-28, 2014; Oxford, England.
50. Cormack AN. A data protection framework for learning analytics. Learning Analytics 2016;3(1):91-106. [doi: 10.18608/jla.2016.31.6]
51. Ferguson R, Hoel T, Scheffel M, Drachsler H. Guest editorial: ethics and privacy in learning analytics. Learning Analytics 2016;3(1):5-15. [doi: 10.18608/jla.2016.31.2]
52. Priestley M, Broglia E, Hughes G, Spanner L. Student perspectives on improving mental health support services at university. Couns and Psychother Res 2022 Mar;22(1). [doi: 10.1002/capr.12391]
53. Alfredo R, Echeverria V, Jin Y, et al. Human-centred learning analytics and AI in education: a systematic literature review. Computers and Education: Artificial Intelligence 2024 Jun;6(5):100215. [doi: 10.1016/j.caeai.2024.100215]
54. Götzl C, Hiller S, Rauschenberg C, et al. Artificial intelligence-informed mobile mental health apps for young people: a mixed-methods approach on users' and stakeholders' perspectives. Child Adolesc Psychiatry Ment Health 2022 Nov 17;16(1):86. [doi: 10.1186/s13034-022-00522-6] [Medline: 36397097]
55. Darwin Holmes AG. Researcher positionality - a consideration of its influence and place in qualitative research - a new researcher guide. education 2020;8(4):1-10. [doi: 10.34293/education.v8i4.3232]
56. Mercer J. The challenges of insider research in educational institutions: wielding a double‐edged sword and resolving delicate dilemmas. Oxf Rev Educ 2007 Feb;33(1):1-17. [doi: 10.1080/03054980601094651]
57. Ting L. Diagram outlining the research process and how patient and public involvement can be embedded in it. NIHR Applied Research Collaboration (ARC) West. URL: https://arc-w.nihr.ac.uk/patient-and-public-involvement/resources/working-effectively-with-public-contributors/research-process-and-ppi-diagram [accessed 2025-06-04]
58. Russell ST, Fish JN. Mental health in lesbian, gay, bisexual, and transgender (LGBT) youth. Annu Rev Clin Psychol 2016;12:465-487. [doi: 10.1146/annurev-clinpsy-021815-093153] [Medline: 26772206]
59. Zeng F, John WCM, Qiao D, Sun X. Association between psychological distress and mental help-seeking intentions in international students of national university of Singapore: a mediation analysis of mental health literacy. BMC Public Health 2023 Nov 28;23(1):2358. [doi: 10.1186/s12889-023-17346-4] [Medline: 38017406]
60. Kollom K, Tammets K, Scheffel M, et al. A four-country cross-case analysis of academic staff expectations about learning analytics in higher education. Internet High Educ 2021 Apr;49:100788. [doi: 10.1016/j.iheduc.2020.100788]

Abbreviations

AI: artificial intelligence
COREQ: Consolidated Criteria for Reporting Qualitative Research
PPI: patient and public involvement
SAG: student advisory group
STEM: science, technology, engineering, and mathematics

XSL•FO
RenderX

Edited by A Mavragani; submitted 20.12.24; peer-reviewed by L Andrew, S Ahern; revised version received 05.06.25; accepted 23.06.25; published 01.08.25.

Please cite as:
Freccero A, Onwunle M, Elliott J, Podder N, Purrinos De Oliveira J, Dewa LH
Students' Perceptions of Learning Analytics for Mental Health Support: Qualitative Study
JMIR Form Res 2025;9:e70327
URL: https://formative.jmir.org/2025/1/e70327
doi:10.2196/70327

© Aglaia Freccero, Miriam Onwunle, Jordan Elliott, Nathalie Podder, Julia Purrinos De Oliveira, Lindsay H Dewa. Originally published in JMIR Formative Research (https://formative.jmir.org), 1.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Evaluating Fitbits for Assessment of Physical Activity and Sleep in Pediatric Pain: Feasibility and Acceptability Pilot Study

Bridget A Nestor[1,2,3], PhD; Andreas M Baumer[2,4], MSc; Justin Chimoff[2], BS; Benoit Delecourt[2], MSc; Camila Koike[2], MD; Nicole Tacugue[2], BS; Roland Brusseau[2,3], MD; Nathalie Roy[3,5], MD; Israel A Gaytan-Fuentes[6], MPH, MD; Navil Sethna[2,3], MBChB; Danielle Wallace[7], MPH, PhD; Joe Kossowsky[2,3], MMSc, PhD

[1]Department of Psychology, Endicott College, Beverly, MA, United States

[2]Department of Anesthesiology, Critical Care, and Pain Medicine, Boston Children's Hospital, 333 Longwood Ave, Boston, MA, United States

[3]Harvard Medical School, Boston, MA, United States

[4]Institute for Implementation Science in Healthcare, University of Zürich, Zurich, Switzerland

[5]Department of Cardiac Surgery, Boston Children's Hospital, Boston, MA, United States

[6]OBEDI Obesity Clinic, Mexico City, Mexico

[7]Rollins School of Public Health, Emory University, Atlanta, United States

Corresponding Author:
Joe Kossowsky, MMSc, PhD
Department of Anesthesiology, Critical Care, and Pain Medicine, Boston Children's Hospital, 333 Longwood Ave, Boston, MA, United States

Abstract

Background: Consumer-grade wearables, such as Fitbits, are a promising, cost-effective methodology for objectively assessing sleep and physical activity in youth with pain.

Objective: This study investigated the acceptability and feasibility of implementing Fitbits for youth with acute and chronic pain in and out of hospital settings while maintaining data security and patient confidentiality.

Methods: We investigated participant experience of Fitbit use over 3 to 4 weeks for a sample of youth with acute pain undergoing either orthopedic or cardiac surgical procedures (N=34, mean age 14.46, SD 3.70 years, 47.06% [n=36] female) and a sample of youth with chronic pain enrolled in an intensive interdisciplinary pain treatment program (N=28, mean age 15.00, SD 2.33 years, 82.14% [n=23] female). We assessed the acceptability of Fitbit use through survey items probing comfort (0=extremely uncomfortable to 10=extremely comfortable), perceived burdensomeness (0=not burdensome at all to 10=extremely burdensome), and open-ended issues or concerns. Feasibility was assessed by tracking the daily compliant wear of the Fitbit device, which was operationalized as more than 600 minutes of daily wear time. We tested for group differences in acceptability and feasibility between orthopedic and cardiac patients within the acute pain sample and between the acute pain and chronic pain samples. We created an automated data pipeline to ensure data security, patient confidentiality, and quality.

Results: Acceptability findings revealed high levels of reported comfort (acute pain: mean 8.56, SD 1.43; chronic pain: mean 8.27, SD 1.69) and low levels of perceived burdensomeness (acute: mean 0.68, SD 1.17; chronic: mean 1.15, SD 1.38) related to Fitbit wearing in both samples. No significant differences in these acceptability outcomes emerged between orthopedic and cardiac patients or between the acute pain and chronic pain groups (P values>.10). Transient concerns of mild wrist irritation and sleep discomfort were occasionally reported across both samples (15.79% [n=9] of participants). Feasibility findings indicated high feasibility (acute: median compliance rate of 86.67%; chronic: median compliance rate of 96.65%) for the study duration in both samples. Mann-Whitney U tests indicated significantly higher median compliance rates per participant among orthopedic as compared with cardiac patients (U=146.5, P=.04) and significantly higher median compliance rates per participant among the chronic pain group as compared with the acute pain group (U=186, P<.001), including significantly higher median compliant days (U=162, P<.001) and study days (U=234.5, P<.001) per participant.

Conclusions: These findings support the use of Fitbits as an acceptable and feasible method for collecting objective data on sleep and physical activity for youth experiencing pain. Findings also highlight the logistics of implementing consumer-grade wearable devices throughout all stages of the clinical research process.

(JMIR Form Res 2025;9:e59074) doi:10.2196/59074

KEYWORDS

Fitbit; pediatric pain; sleep; pediatric; acceptability; feasibility; physical activity; youth; wearable devices

XSL•FO
RenderX

Introduction

Pediatric pain affects nearly 40% of youth globally [1-3] and is associated with significant functional impairment across multiple life domains [4]. Youth who develop a chronic pain disorder are also at increased risk for such disorders persisting into adulthood [5], often resulting in costly health care utilization [6] and positioning pediatric pain as a serious public health concern. The gold-standard intervention for pediatric chronic pain assumes an interdisciplinary approach [7], guided and informed by the biopsychosocial pain model [8], where treatment aims to address multifactorial influences on pain [9].

Physical activity and sleep are 2 biopsychosocial factors critical to the experience of pediatric pain and are identified as core outcome domains for its treatment and study [6]. Prior research has identified bidirectional relations between decreased physical activity and increased pain [10] and between decreased sleep duration and quality and increased pain [11,12]. A reliable and valid assessment of physical activity and sleep is necessary for intervention efforts in pediatric pain. Such assessment must also accurately and precisely monitor and measure changes in these constructs over time (ie, improvements and deteriorations in physical activity and sleep).

New multisensory, consumer-facing technology has emerged as a promising means to continuously assess physical activity and sleep in research [13]. Of these consumer-grade devices, Fitbits have garnered the most research [14] and are significantly more affordable, widely available, and easy to use, highlighting their potential for large-scale studies and long-term monitoring. Fitbits capture physical activity and sleep metrics using accelerometer data collected in 1-minute and 30-second epochs, respectively [15,16]. Technologically, newer Fitbits differ from actigraphy devices in that they leverage multisensory capabilities, differentiating them from actigraphy devices, which primarily use accelerometer data. For example, Fitbits incorporate an optical heart rate sensor in addition to accelerometry, allowing for more refined sleep staging capabilities by distinguishing between light, deep, and REM sleep phases. With this new multisensory technology, newer versions of Fitbit devices can also accurately capture sleep stages and heart rate and demonstrate a specificity of 61% - 69% [17], which is significantly higher than that of actigraphy [16]. In contrast, traditional actigraphy devices estimate sleep solely based on movement patterns, often resulting in lower specificity.

Fitbits have high user acceptability over long periods of time [18], and have demonstrated high levels of accuracy and sensitivity for correctly identifying sleep episodes compared with polysomnography [19], a laboratory-based sleep methodology that leverages simultaneous recordings of multiple physiological indicators, including brain activity, eye movements, heart rate, and breathing patterns. Findings show that Fitbits provide measures of total sleep time that are statistically indistinguishable from polysomnography - and actigraphy-measured total sleep time [20], including in pediatric samples [21-23]. Fitbits also boast cost-effective prices, and as wrist-worn devices, are noninvasive methodologies. Accordingly, these methodological strengths position Fitbits as a scalable method. Indeed, large-scale, longitudinal, multi-site studies funded by the National Institutes of Health, such as the Adolescent Brain Cognitive Development study, have incorporated Fitbits into their data collection for youth [21]. Other research has begun to examine the use of Fitbits to assess physical activity or sleep in youth [24,25], both in healthy samples [26], as well as those with chronic health conditions [27,28] and physical and developmental disabilities [29,30]. Investigations in pediatric pain, however, have been minimal. With an exception, a recently published pilot study examined the feasibility of using a much older iteration of Fitbit for assessing physical activity and sleep in adolescents at an intensive interdisciplinary pain rehabilitation program [31]. However, the data from this study were somewhat incomplete due to challenges with compliance, and the device used did not incorporate multisensory technology.

Further investigation is needed to use Fitbits in pediatric pain populations for several reasons. First, from a clinical perspective, reliable and valid physical activity and sleep assessment are integral to treatment progress for acute and chronic pediatric pain populations [32]. Second, from a research perspective, a more detailed description of the practical issues and challenges relevant to the research implementation of Fitbit is critical for developing best practices for using consumer-grade methodologies, particularly those with proprietary, “black-box” algorithms [14]. Third, more research using Fitbits is needed to augment and contrast other large-scale adolescent studies already underway that have adopted Fitbit data collection (eg, Adolescent Brain Cognitive Development).

Therefore, the current article describes our implementation of Fitbits for assessing physical activity and sleep in samples of youth with acute and chronic pain – those undergoing orthopedic or cardiac surgery (acute pain) and those in an interdisciplinary pain rehabilitation program (chronic pain). We chose to investigate 2 samples of surgical patients with acute pain to determine the extent to which Fitbits could be acceptably and feasibly implemented in varied surgical populations. Similarly, we chose to investigate an acute and a chronic sample to assess the extent to which Fitbit’s acceptability and feasibility may vary by pain acuity and patient care environment. We hypothesized that Fitbits would be acceptable and feasible devices for the assessment of physical activity and sleep in youth with acute and chronic pain. We further hypothesized that greater acceptability and feasibility would be observed in youth with chronic pain, as opposed to acute pain, due to the inherent challenges associated with inpatient and postoperative care. These challenges may range from the clinical workflow of the surgical process to difficulty wearing a Fitbit with surgical wounds or drains to potential reliance on medical personnel for refitting the Fitbit after surgery.

Methodologically, we explain our quantitative methods for determining the acceptability and feasibility of Fitbit assessment in this population, and we also explain narratively our process of Fitbit research implementation, including data collection and data management. We then present acceptability and feasibility data collected from our samples. Finally, we discuss our viewpoints on the research implementation of Fitbits in pediatric populations with pain and provide our recommendations for

XSL•FO
RenderX

best practices of this assessment method for future pediatric research.

Methods

Ethical Considerations

Institutional Review Board (IRB) of Boston Children's Hospital approved this study of the 3 patient groups (cardiac surgery: IRB-P0003441, April 12, 2021; orthopedic surgery: IRB-P00038993, July 27, 2021; chronic pain: IRB-P00036475, October 07, 2021). Informed consent was obtained. Privacy and confidentiality of research participants' data and identity were maintained. Compensation varied by study protocol: acute pain participants received US $25, and chronic pain participants received US $90 as part of a comprehensive parent study.

Participants

Two samples of participants seeking pain treatment services for acute or chronic pain were recruited from a tertiary pediatric hospital in the Northeast United States. The first sample consisted of youth with acute pain who were recruited before undergoing either cardiac or periacetabular osteotomy orthopedic surgery (acute pain group). The second sample consisted of youth with disabling chronic musculoskeletal pain who were enrolled in an outpatient intensive interdisciplinary pain treatment (IIPT) program (chronic pain group), which operates on a day-hospital model of rehabilitation.

Participants were included in the study if they were (1) between 8 and 18 years of age, (2) willing to wear a Fitbit watch, and (3) requiring pain treatment services immediately following cardiac surgery, periacetabular osteotomy orthopedic surgery, or disabling chronic pain. Participants were excluded for intellectual disability. All participants understood and responded to questions in English. Data collection occurred from February 2022 to February 2024.

Eligible patients were identified by research staff through weekly screening of clinic schedules. A research staff contacted the caregiver or legal guardians of identified patients via mail, email, or secure patient portal message with information about the study. Several days after the information was sent, patients were contacted via phone or in the clinic to explain the study requirements and answer any questions. Patients who expressed interest were then consented to the study.

The total acute pain group consisted of 34 participants (mean age 14.46, SD 3.70 years, 47.06% [n=16] female) who provided acceptability or feasibility data. Of these 34 patients, 24 were cardiac patients (mean age 12.92, SD 2.24 years, 33.33% [n=8] female), and 10 were orthopedic patients (mean age 18.30, SD 3.92 years, 80.00% [n=8] female). Some participants in the acute pain group withdrew from the study due to changes in the scheduling of surgery and location of family prohibiting sending/returning the device. The total chronic pain group consisted of 28 participants (mean age 15.00, SD 2.33 years, 82.14% [n=23] female) who provided acceptability and feasibility data. Participants in the chronic pain group who withdrew from study participation cited illness (ie, contracting COVID-19), lack of interest in the study, schedule changes, and withdrawal from IIPT as reasons for discontinuing their participation. Notably, one participant withdrew from participation due to getting a rash from the device. Missing Fitbit data was due only to participants not wearing the device.

Participant Onboarding Procedures

Acute Pain Group

For patients with adequate time before their planned scheduled preoperative visit or date of surgery (DOS), a research staff member obtained written informed consent from those who provided remote consenting (eConsenting) on REDCap (Research Electronic Data Capture) [33] via phone call or Zoom. Patient assent and parental consent were obtained via electronic signature on REDCap. When patients were scheduled close to a preoperative visit or DOS, a research staff member obtained written assent from patients and consent from their parent/guardian on the preoperative visit day, typically the day before or a few days before the DOS.

Surgical patients received a Fitbit device and account information at the end of their consent meeting on their preoperative visit day. Those who consented via eConsenting received a mailed Fitbit device and account information. After receiving their device in person or via mail, a research staff member contacted the families via Zoom to explain the Fitbit setup and answer any questions. Participants were instructed to wear their Fitbit continuously throughout the study. Specifically, they were asked to wear their Fitbit into the hospital on their DOS; a research staff member removed the device before the surgery and replaced it after. Participants were allowed to keep their Fitbit on during bathing and showering, if desired, and they could take small breaks as needed. Participants were also instructed to sync their devices every 2 to 3 days and charge them every 7 days.

During the study, participants indicated their understanding of device setup procedures, and they were given a copy of the standard operating procedures (SOPs) when they received their Fitbit device (either in person or via mail). Multimedia Appendix 1 contains the SOP of Fitbit configuration process. Participants set up their devices independently with support from the SOPs. Contact information for the research staff was provided if families requested additional support, either via phone call or Zoom meeting.

Chronic Pain Group

Patients who expressed interest during the recruitment phone call were informed that they would be consented on the admission day to the outpatient IIPT program. On the first admission day, a research staff member obtained written informed assent from patients and consent from their parents/guardians. On the second day, these patients were provided with Fitbit devices and account information. This onboarding was not completed during the patients' first day of admission due to the busy admission process by different disciplines at the IIPT. A research staff member explained the Fitbit setup, answered any questions, and provided the same standard instructions to sync their device every 2 to 3 days and charge it every 7 days. Multimedia Appendix 1 contains the SOP of Fitbit configuration process.

XSL•FO
RenderX

Fitbit Account Creation and User Agreements

The Fitbit setup process included creating a participant account. Before onboarding, the study team created a unique deidentified email account for each new participant. A unique email account was provided to each newly enrolled participant when registering/joining Fitbit during the study setup.

Instruments and Measures

Fitbit Charge 3

The Fitbit Charge 3 (Fitbit Inc; 2018)was implemented for the cardiac surgical participants in the acute pain group. The Fitbit Charge 3 is a wrist-worn consumer-grade device that uses multisensory technology to provide relevant physical activity and sleep metrics. For physical activity, the Fitbit Charge 3 records steps taken and mileage with its 3D accelerometer, allowing motion and acceleration tracking in all directions. For sleep assessment, it leverages accelerometer technology and a PPG sensor to calculate heart rate, monitor sleep patterns, and identify sleep stages (eg, rapid eye movement and nonrapid eye movement).

Fitbit Inspire 2

The Fitbit Inspire 2 (Fitbit Inc; 2020) was implemented for orthopedic surgical and chronic pain group participants. Like the Fitbit Charge 3, the Inspire 2 also assesses heart rate and provides the same activity metrics of interest (ie, step count) and sleep-tracking capabilities. The newer Fitbit Inspire 2 model offers a longer battery life and a slightly thinner and lighter build than the Fitbit Charge 3. Both devices use the same sleep assessment algorithm.

Acceptability

Acceptability was assessed through self-report scales and open-ended questions developed by the research team. Specifically, on a 10-point scale, participants rated their comfort (“How comfortable was it to wear the watch?” 0=extremely uncomfortable to 10=extremely comfortable) and burden (“How much of a burden/chore was it to wear the watch?” 0=not burdensome at all/not a chore at all to 10=extremely burdensome/extreme chore). Participants in the chronic pain group were additionally asked on the extent to which wearing the Fitbit disturbed their sleep on a 10-point scale (“How much did wearing the Fitbit disturb your sleep?” 0=no disturbance at all to 10=extremely disturbed). All participants were also asked open-ended questions about potential adverse reactions from wearing the Fitbit (“Did you have any adverse reactions or experience any other distressing issues while wearing the Fitbit?”) and about how much they enjoyed participating in the study.

Feasibility

This study determined feasibility by participant compliance with wearing the Fitbit, as indicated by device use. Consistent with other thresholds defined for wrist-worn consumer wearable assessment in biobehavioral research, a 600-minute cut-off was implemented to determine daily compliant use for activity tracking [34]. Participants were considered compliant if they wore their device for 600 minutes or more during the day, excluding sleep periods, and for 180 continuous minutes or more during the night while asleep, without off-wrist indicators. Compliance assessment was facilitated by minute-level heart rate data from Fitbit, based on the assumption that recorded heart rate data signified the device was worn. If the heart rate sensor was deactivated or malfunctioned, the Fitbit would still register activity data. Thus, minutes without both heart rate and activity data were assumed to be times the device was not worn, signifying noncompliance.

Compliance ratings for the total number of compliant days out of the total number of study days were generated for the acute pain group and the chronic pain group as described above. Due to variability in the time of day that participants began and stopped wearing their devices on the first and last day of the study, actigraphy data were trimmed to exclude the first and last partial day of measurement. Similarly, sleep data were trimmed to exclude the first night of measurement. Compliance rates throughout the study were computed for each participant and each sample.

Study Timeline

Acute Pain Group

After receiving their Fitbit device at the preoperative appointment, participants in the acute pain group were instructed to continuously wear their Fitbit before their surgery, immediately after their surgery, while inpatient, and for 2 weeks after discharge. Devices were not worn during the surgical procedures. Participants answered acceptability and feasibility survey questions as part of a semistructured interview with a research staff member at the end of their study participation.

Chronic Pain Group

After receiving their Fitbit device on day 2 of IIPT, participants in the chronic pain group were instructed to continuously wear their Fitbit devices for the duration of their admission to the IIPT program. Treatment duration in this program is typically 4 weeks (8 h/day, 5 d/week), and participants were also instructed to wear the devices on the weekends. Participants completed the acceptability and feasibility survey items at the end of their study participation at their discharge.

Data Collection and Processing

Because no prior automated process was available for data collection using Fitbits for research purposes in our hospital, we created an automated pipeline using Python (version 3.11; Python Software Foundation) to securely collect and manage participant device data (Figure 1). This pipeline enacted 3 steps: data extraction, transformation, and report creation.

Our pipeline was initiated by connecting to the participants’ accounts using the provided anonymous emails and passwords. Data was then requested via the Fitbit web application programming interface, storing the resulting JSON data files on our secure hospital servers. These JSON files were then reformatted into structured, tabular CSV files, encapsulating relevant sleep, heart rate, and activity metrics at either 30-second or 1-minute intervals, from which daily summary statistics were derived. Subsequently, the pipeline generated reports featuring data visualizations to monitor compliance and provide participants with an intuitive summary of their health data

XSL•FO
RenderX

(detailed in the Results section). This process was automated to iterate for all active participant Fitbit accounts. To ensure ongoing participant compliance, a research staff member performed daily reviews of the generated reports. This allowed for the identification and subsequent contact of participants who either failed to sync their devices for 3 consecutive days or exhibited 2 consecutive days of compliance issues, to provide reminders and troubleshooting support.

Figure 1. Visual depiction of the automated Fitbit data pipeline. First, raw Fitbit data from each participant is extracted via API calls and stored locally in an unstructured JSON format. Second, these raw JSON files are processed and transformed into structured tabular CSV files for analysis. Third, the pipeline generates automated reports with data visualizations, enabling both researchers and participants to monitor compliance and review summarized activity and sleep trends. API: application programming interface; CSV: Comma-Separated Values; JSON: JavaScript Object Notation.

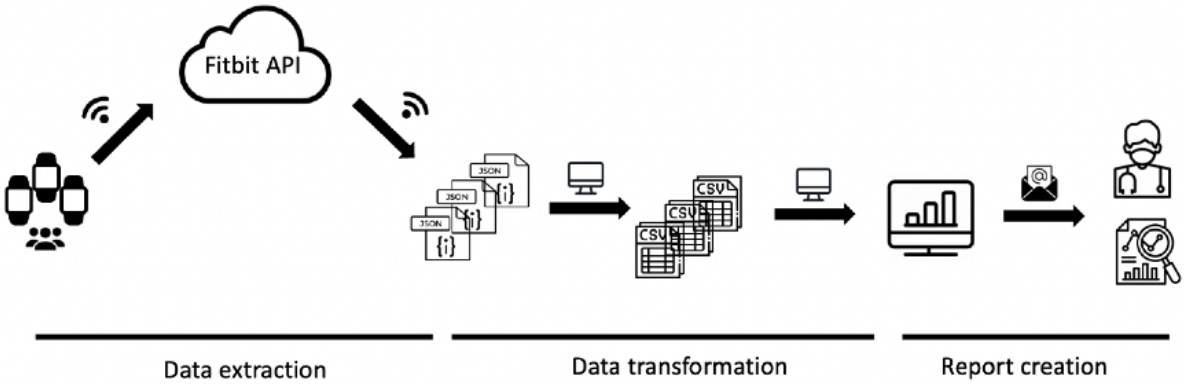

Statistical Analysis

Analyses conducted in this study were largely descriptive. For acceptability outcomes, participants' self-reported device comfort, perceived burdensomeness, and sleep disturbance were summarized for means and standard deviations. Adverse reactions and responses to open-ended inquiries were reported descriptively. For feasibility outcomes, compliance rates were summarized using medians and ranges. Tests of differences using independent samples *t* tests were conducted to investigate mean differences in acceptability between acute pain and chronic pain groups, as well as between orthopedic and cardiac samples. Similarly, Mann-Whitney U tests were conducted to examine median differences in compliance between acute pain and chronic pain groups and between orthopedic and cardiac samples.

Results

Acceptability

Overview

A total of 31 participants in the acute pain group and 26 participants in the chronic pain group completed items related to acceptability. Findings indicated high levels of comfort and low levels of burdensomeness reported by participants in both samples and very low levels of sleep disturbance in the chronic pain group (Table 1). Figure 2 depicts the CONSORT diagram.

Results of independent samples *t* tests (Table 1) indicated no significant differences in participant comfort or perceived burdensomeness between orthopedic and cardiac samples or between the acute pain and chronic pain groups (P values>.10).

XSL•FO
RenderX

Table . Participant reported comfort with and the burdensomeness of wearing Fitbit devices.

	Acute pain					Chronic pain		
	Orthopedic (N=10), mean (SD)	Cardiac (N=21), mean (SD)	Test of mean differences[a]		Total (N=31), mean (SD)	IIPT[b] (N=26), mean (SD)	Test of mean differences[c]	
			t test (*df*=29)	*P* value			*t* test (*df*=55)	*P* value
Comfort[d]	8.80 (1.14)	8.48 (1.60)	−0.57	.58	8.56 (1.43)	8.27 (1.69)	−0.71	.49
Burden[e] /Chore	0.30 (0.48)	0.86 (1.35)	1.27	.22	0.68 (1.17)	1.15 (1.38)	1.39	.17
Sleep disturbance[f]	NA[g]	NA	NA	NA	NA	0.69 (1.12)	NA	NA

[a]Independent samples *t* test comparing orthopedic and cardiac samples.

[b]IIPT: intensive interdisciplinary pain treatment.

[c]Independent samples *t* test comparing Acute Pain and Chronic Pain [35].

[d]Comfort was assessed on a 0 - 10 scale (0=extremely uncomfortable to 10=extremely comfortable).

[e]Burdensomeness was assessed on a 0 - 10 scale (0=not burdensome at all/not a chore at all to 10=extremely burdensome/extreme chore).

[f]The degree to which the Fitbit disturbed their sleep was assessed on a 0 - 10 scale (0=no disturbance at all to 10=extremely disturbed).

[g]NA: not assessed.

Figure 2. CONSORT diagram of participants included in the study. IIPT: intensive interdisciplinary pain treatment.

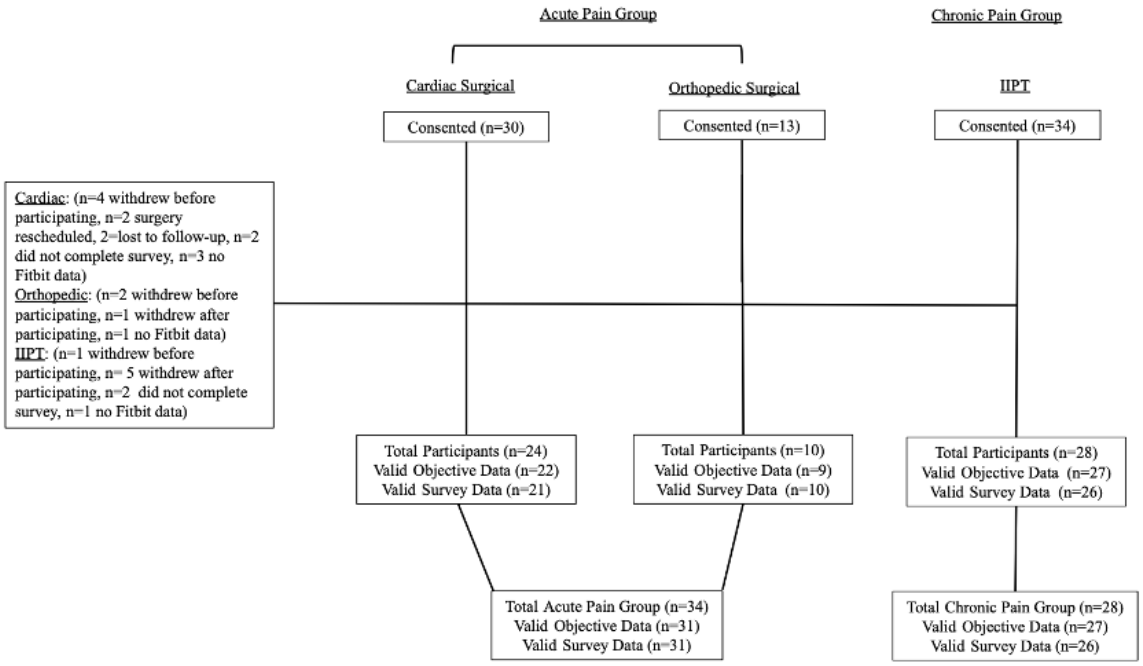

Adverse Reactions

In the acute pain group, 90.3% (28/31) reported no adverse reactions or distressing issues related to their Fitbit wear. In the chronic pain group, 76.9% (20/26) reported no adverse reactions or distressing issues while wearing their Fitbit. For participants in both groups who did report adverse reactions or other distressing issues, responses are presented in Textbox 1.

XSL•FO
RenderX

Textbox 1. Participants reported adverse reactions and other distressing issues while using Fitbits.

Acute pain (N=3)

- "Sometimes put on too tight; user error" [Cardiac sample]
- "Hot during sleep, may be due to tightness on wrist" [Cardiac sample]
- "Once after shower it was red" [Orthopedic sample]

Chronic pain (N=6)

- "Got a rash after taking it off"
- "Brightness of watch when falling asleep"
- "Yes, a mild rash on left side. Removed watch and put it on right side"
- "My wrist got slightly irritated"
- "Sometimes it left a mark on my skin but it was fine"
- "Slight red irritation, dry skin, not itchy/painful"

Open-Ended Responses

Responding to open-ended inquiries, some participants reported that wearing the Fitbit device was a positive experience. For example, participants in the acute pain group reported they "liked the Fitbit," that "the Fitbit was cool," that "the Fitbit was very fun," and that it was "fun to have [the] watch." Similarly, other participants in the chronic pain group reported that it was "kind of fun, nice seeing how well I slept and heart rate," that they "liked wearing fitbit, helpful to use it," that they "liked to use the Fitbit and the phone," and that they "liked wearing [the] fitbit and seeing data."

Other participants provided responses that indicated issues with wearing the Fitbit device. For example, participants in the acute pain group reported that "sleeping in the Fitbit and wearing the Fitbit started to be uncomfortable for a little bit," that they "sometimes forget to put on a watch, like a secondary Apple watch," and that "Fitbit hurt during sleep because I move around a lot and put the watch tight." Participants in the chronic pain group reported that it was "annoying sometimes that there weren't any breaks from wearing it" and that they "kept losing [their] watch."

Feasibility

In the acute pain group, compliance data were obtained from 31 participants (22 cardiac surgical participants and 9 orthopedic surgical participants). All participants in the acute pain group had 604 total days of study duration. Of those, 485 days were classified as compliant per the 600-minute threshold. The median compliance rate for the acute pain group was 86.67%.

In the chronic pain group, compliance data were obtained from 27 participants. There were 624 total days of study duration across all participants in the chronic pain group. Of those, 583 days were classified as compliant per the 600-minute threshold. The median compliance rate was 95.65% for the chronic pain group. Table 2 contains additional details about compliance rates for both samples.

Results of Mann-Whitney U tests (Table 2) indicated significantly higher median compliance rates per participant among orthopedic as compared with cardiac samples (U=146.5, P=.04). Results also indicated significantly higher median compliance rates per participant among the chronic pain group as compared with the Acute Group (U=186, P<.001), including significantly higher median compliant days (U=162, P<.001) and study days (U=234.5, P<.001) per participant.

In the acute pain group, sleep compliance data were obtained from 31 participants (22 cardiac surgical participants and 9 orthopedic surgical participants). All participants in the acute pain group had 524 total days of study duration. Of those, 478 days were classified as compliant per the 180-minute threshold. The median compliance rate for the acute pain group was 90.91%.

In the chronic pain group, compliance data were obtained from 27 participants. There were 631 total days of study duration across all participants in the chronic pain group. Of those, 596 days were classified as compliant per the 180-minute threshold. The median compliance rate was 95.65% for the chronic pain group. Table 3 contains additional details about compliance rates for both samples.

Results of Mann-Whitney U tests (Table 3) did not indicate significant differences in median compliance rates per participant among orthopedic as compared with cardiac samples (U=338, P=.20).

As noted, compliance monitoring was enacted through daily checks of generated reports to identify participants who had not synced their devices in 3 days or demonstrated compliance rates beneath the threshold of 600 minutes of daily wear time. These reports also provided visualizations of both sleep and physical activity (Multimedia Appendix 1).

XSL•FO
RenderX

Table . Day compliance data during the study period. For 3 participants in the acute pain group, device-wearing did not begin until postdischarge due to medical complications during hospitalization.

	Acute pain					Chronic pain		
	Orthopedic (N=9)	Cardiac (N=22)	Test of median differences[a]		Total (N=31)	IITP[b] (N=27)	Test of median differences[c]	
			U test	*P* value			U test	*P* value
Number of compliant days per participant, median (range)	18 (11-29)	15.5 (6-29)	128.5	.20	16 (6-29)	22 (9-27)	162.5	<.001
Number of study days per participant, median (range)	19 (14-29)	21 (11-30)	96	.91	20 (11-30)	24 (10-18)	234.5	<.001
Total number of compliant days	160	325	__[d]	—	485	583	—	—
Total number of study days	175	429	—	—	604	624	—	—
Compliance rate per participant (%), median (range)	90 (78.57 - 100)	80 (36.36 - 96.67)	146.5	.04	86.67 (36.36 - 100)	95.65 (60 - 100)	186	<.001

[a]Mann-Whitney U test comparing orthopedic and cardiac samples.

[b]IIPT: intensive interdisciplinary pain treatment.

[c]Mann-Whitney U test comparing Acute Pain vs Chronic Pain [36].

[d]Not applicable.

XSL•FO
RenderX

Table . Sleep compliance data during the study period. For 3 participants in the acute pain group, device-wearing did not begin until postdischarge due to medical complications during hospitalization.

	Acute pain					Chronic pain		
	Orthopedic (N=9)	Cardiac (N=22)	Test of median differences[a]		Total (N=31)	IITP[b] (N=27)	Test of median differences[c]	
			U test	*P* value			U test	*P* value
Number of compliant nights per participant, median (range)	17 (8-30)	12.5 (4-29)	84	.53	13 (4-30)	23 (5-29)	200	<.001
Number of study nights per participant, median (range)	20 (9-30)	14.5 (6-31)	87.5	.63	15 (6-31)	23 (8-29)	191.5	<.001
Total number of compliant nights	156	322	__[d]	—	478	596	—	—
Total number of study nights	166	358	—	—	524	631	—	—
Compliance rate per participant (%), median (range)	90 (80.96 - 100)	92.23 (55.56 - 100)	85	.55	90.91 (55.5 - 100)	95.65 (62.50 - 100)	338	.20

[a]Mann-Whitney U test comparing orthopedic and cardiac samples.

[b]IIPT: intensive interdisciplinary pain treatment.

[c]Mann-Whitney U test comparing Acute Pain vs Chronic Pain [36].

[d]Not applicable.

Discussion

This study investigated the implementation of Fitbits for assessing physical activity and sleep in youth with acute and chronic pain. Consistent with our hypotheses, our findings position Fitbits as acceptable and feasible methods of assessment in these populations. To our knowledge, this is the first study investigating the acceptability and feasibility of using Fitbits in youth with pain after surgery and in outpatient rehabilitation settings. Below, we discuss our findings, highlight the logistical challenges of this work, and suggest avenues for future research.

In both acute and chronic pain groups of this study, Fitbits garnered similar and high levels of acceptability from youth. Adolescents in both samples and across both Fitbit devices reported a high degree of comfort and a very low degree of burdensomeness with wearing their devices, echoing prior research that has also found high levels of acceptability for wearable trackers in other pediatric samples [25,37-39]. Open-ended responses also highlighted adolescents' positive experiences with Fitbit-wearing. For example, youth reported that they enjoyed wearing their Fitbit, that it was fun to have their device, and that they enjoyed raising awareness of their sleep pattern and physical activity behaviors. This highlights the potential for using this method for intervention to help youth self-monitor and improve their own sleep and activity levels, which in turn could improve their ability to manage their pain.

Adolescents' positive perspectives on using Fitbits could have been related to the youth's general enthusiasm for technology [40]. For pediatric samples with pain, our study's positive responses to Fitbits may have reinforced adaptive health-focused behaviors to engage in clinical treatment and monitor functional progress, such as sleep hygiene, activity pacing, behavioral activation, etc. On the other hand, the few negative responses may indicate low tolerance or concerns about discomfort at night, minor skin irritations, and the device's light emission, which have been reported in other pediatric populations [41]. Future studies using Fitbits in youth may benefit from addressing these concerns and specifically emphasizing appropriate tightness of the device, adequate cleanliness of the band, and brief removal of the device after periods of extended wear. Additionally, it would be important to guide patients on how to adjust relevant Fitbit settings, for example, dimming the display brightness to not interfere with sleep and circadian rhythms, and ensuring the heart rate sensor is activated.

In addition to acceptability, our results also highlight Fitbit's feasibility as an assessment method for youth with pain. Despite some variability, particularly in our cardiac patients, these results showed high median compliance rates of over 85% across youth with acute and chronic pain. These findings indicate higher compliance rates than were observed in a recently published feasibility study using a much older iteration of Fitbit for assessing physical activity and sleep in adolescents at a similar intensive interdisciplinary pain rehabilitation program [31]. However, it is also worth noting that compliance rates were

XSL•FO
RenderX

significantly higher for participants enrolled in IIPT than for the surgical samples, which was consistent with our hypotheses. In this study, the lower compliance rate for surgical patients, as compared with chronic patients, might be related to the clinical workflow of the surgical process, postoperative inpatient setting, patients' distress with pain, anxiety, caring for the surgical wound/drains, and poor sleep, particularly after cardiac surgery, and reliance on medical personnel for refitting the Fitbit after surgery. Indeed, our findings also revealed significantly lower compliance rates for cardiac patients as compared with orthopedic patients, likely due to the extra challenges and stressors inherent to cardiac versus orthopedic surgery. In general, compliance rates should be interpreted cautiously, particularly following surgical procedures, as medical necessities may override device compliance during this critical period, such as intravenous therapy, invasive monitoring lines, extended ICU stays, etc. It should also be noted, in this study, that orthopedic (and chronic pain) patients wore the Fitbit Inspire 2, whereas cardiac patients wore the Fitbit Charge. Although both devices use similar assessment technology, the Inspire 2 has a thinner and sleeker design, which may be relevant to these differential compliance rates.

Several practical challenges and limitations emerged that are relevant to the feasibility of using Fitbits for research purposes. First, the Fitbit device requires participants to charge it approximately every 7 days. In this study, the period of device charging was a particularly risky window for low compliance, suggesting that a reminder strategy may help youth replace their device after charging. Fortunately, compared with other consumer-grade wearable devices, Fitbits have a much longer battery life and thus require less frequent charging (eg, compared with Apple Watch, which has a shorter 18 h battery life). Failure to sync the device promptly posed a further challenge for data collection. Because of limited storage space on Fitbit devices, it can hold high-resolution (ie, minute or second) data for about 7 days, with the exact duration varying across different models. If the device is not synced in this timeframe, it will discard the oldest high-detail information and keep only the daily summary values. Participants were instructed to sync their devices at least every 3 days to ensure we had access to the highest possible data resolution. Our research team reached out to the participants with low compliance data and low battery status, as indicated by daily checks of the data pipeline. Second, we found that specific Fitbit settings must be activated for proper data collection. For example, Fitbits can be worn on the wrist or a clip. However, when the device is switched to "clip mode," heart rate and sleep are not recorded due to the deactivation of the optical PPG sensor (skin sensor). For one participant in the Chronic Pain Sample, the Fitbit was inadvertently switched into "clip mode," resulting in the loss of 2 weeks' worth of data. This highlights the importance of providing clear instructions to participants and periodic reminders on the proper use of the device for data fidelity. To address these challenges, we suggest that future work develop comprehensive SOPs to avoid the recurrent compromising of data. Third, we acknowledge that the Fitbit Charge 3 was released in 2018, making it not the most current device. However, in clinical practice, many institutions and families still use these devices due to cost and accessibility. Additionally, prior to launching the acute pain study, we conducted a pilot trial to assess the initial acceptability and feasibility of the Fitbit Charge 3 within this population. Based on the positive findings from that pilot, as well as the prior institutional purchase of these devices, we chose to maintain consistency by continuing to use the Charge 3. This approach minimized variability in data collection while also leveraging existing resources to enhance study feasibility. Finally, given the nature of the IIPT as a day program with close contact with study personnel over the course of 4 weeks, we, unfortunately, could not control for differences in the amount of contact with study staff and its potential influence on our findings.

Ensuring patient confidentiality and the highest possible level of transparency during patient consent was also central to the feasible implementation of Fitbits in this study. We took several measures to ensure this throughout the study duration in close communication with our IRB. First, we created deidentified Fitbit accounts and linked them to anonymous email addresses only accessible by the research team. The account information was stored in private, protected folders on central computers at Boston Children's Hospital, as the Privacy Rule of the Health Insurance Portability and Accountability Act (HIPAA) recommended. Second, each new participant was required to agree to the end user license agreement of Fitbit during participant onboarding. In addition, we used devices that could not record GPS data to mitigate further confidentiality concerns that may stem from location tracking. Developing a secure, automated data collection pipeline was also integral to a feasible and scalable Fitbit implementation. Without automated data logging or an application that accesses the Fitbit application programming interfaces, data can only be visualized on the Fitbit app or manually downloaded, neither of which represents a feasible approach to large-scale data management.

In summary, findings from this study support the utility of Fitbits as an acceptable and feasible method of objective data collection from youth experiencing acute and chronic pain. Future directions should explore how this wearable technology, coupled with passive sensing and ecological momentary assessment, may be leveraged toward continuous and valid assessment of pain-impacted functional measures and biopsychosocial correlates in pediatric pain populations [42,43]. The sleep, activity, and heart rate can be correlated to medications and clinical pain interventions in the future, as well as perioperative adverse events. Youth self-monitoring of physical activity and sleep patterns may also better promote patient-centered collaborative care between providers and youth with pediatric pain. Our findings also underscore the importance of patient confidentiality and data privacy when implementing wearable devices. We hope our preliminary research methodology of collecting and analyzing data from a wearable device in youth with acute and chronic pain treated in outpatient and inpatient settings can help future pediatric pain studies answer specific questions and test hypotheses.

XSL•FO
RenderX

Acknowledgments

JK discloses support for the research of this work from NIH grant K01DA057374 and the Boston Children's Hospital Anesthesia Research Distinguished Trailblazer and Ignition Awards. DW discloses grant support from NIH-NHLBI K99HL166700. Generative artificial intelligence was not used in any portion of the manuscript writing.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

BAN, BD, AMB, and NT assisted with data analyses. BAN drafted the initial manuscript and revised and reviewed all drafts of the manuscript. NT, JC, and IAGF assisted in creating the materials, carrying out the data acquisition, and reviewing and revising the manuscript. BD and AMB assisted with data acquisition and reviewed and revised the manuscript. CK, DW, and NS reviewed and revised the manuscript. RB and NR assisted in conceptualizing the study, assisted in data acquisition, and reviewed and revised the manuscript. JK conceptualized and designed the study, coordinated and supervised data collection and analyses, and reviewed and revised the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary materials.
[PDF File, 1468 KB - formative_v9i1e59074_app1.pdf]

References

1. Huguet A, Miró J. The severity of chronic pediatric pain: an epidemiological study. J Pain 2008 Mar;9(3):226-236. [doi: 10.1016/j.jpain.2007.10.015] [Medline: 18088558]
2. King S, Chambers CT, Huguet A, et al. The epidemiology of chronic pain in children and adolescents revisited: a systematic review. Pain 2011 Dec;152(12):2729-2738. [doi: 10.1016/j.pain.2011.07.016] [Medline: 22078064]
3. Swain MS, Henschke N, Kamper SJ, Gobina I, Ottová-Jordan V, Maher CG. An international survey of pain in adolescents. BMC Public Health 2014 May 13;14:447. [doi: 10.1186/1471-2458-14-447] [Medline: 24885027]
4. Rabbitts JA, Holley AL, Groenewald CB, Palermo TM. Association between widespread pain scores and functional impairment and health-related quality of life in clinical samples of children. J Pain 2016 Jun;17(6):678-684. [doi: 10.1016/j.jpain.2016.02.005] [Medline: 26924379]
5. Walker LS, Dengler-Crish CM, Rippel S, Bruehl S. Functional abdominal pain in childhood and adolescence increases risk for chronic pain in adulthood. Pain 2010 Sep;150(3):568-572. [doi: 10.1016/j.pain.2010.06.018] [Medline: 20615615]
6. Groenewald CB, Essner BS, Wright D, Fesinmeyer MD, Palermo TM. The economic costs of chronic pain among a cohort of treatment-seeking adolescents in the United States. J Pain 2014 Sep;15(9):925-933. [doi: 10.1016/j.jpain.2014.06.002] [Medline: 24953887]
7. Hechler T, Kanstrup M, Holley AL, et al. Systematic review on intensive interdisciplinary pain treatment of children with chronic pain. Pediatrics 2015 Jul;136(1):115-127. [doi: 10.1542/peds.2014-3319] [Medline: 26101358]
8. Gatchel RJ, Peng YB, Peters ML, Fuchs PN, Turk DC. The biopsychosocial approach to chronic pain: scientific advances and future directions. Psychol Bull 2007 Jul;133(4):581-624. [doi: 10.1037/0033-2909.133.4.581] [Medline: 17592957]
9. Kozlowska K, Rose D, Khan R, Kram S, Lane L, Collins J. A conceptual model and practice framework for managing chronic pain in children and adolescents. Harv Rev Psychiatry 2008;16(2):136-150. [doi: 10.1080/10673220802069723] [Medline: 18415885]
10. Rabbitts JA, Holley AL, Karlson CW, Palermo TM. Bidirectional associations between pain and physical activity in adolescents. Clin J Pain 2014 Mar;30(3):251-258. [doi: 10.1097/AJP.0b013e31829550c6] [Medline: 23669450]
11. Liu X, Yang Y, Liu ZZ, Jia CX. A longitudinal study of bidirectional associations between frequent pain and insomnia symptoms in adolescents. Sleep Health 2022 Oct;8(5):467-474. [doi: 10.1016/j.sleh.2022.05.009] [Medline: 35872148]
12. Bromberg MH, Gil KM, Schanberg LE. Daily sleep quality and mood as predictors of pain in children with juvenile polyarticular arthritis. Health Psychol 2012 Mar;31(2):202-209. [doi: 10.1037/a0025075] [Medline: 21842997]
13. Baron KG, Duffecy J, Berendsen MA, Cheung Mason I, Lattie EG, Manalo NC. Feeling validated yet? a scoping review of the use of consumer-targeted wearable and mobile technology to measure and improve sleep. Sleep Med Rev 2018 Aug;40:151-159. [doi: 10.1016/j.smrv.2017.12.002] [Medline: 29395985]
14. de Zambotti M, Rosas L, Colrain IM, Baker FC. The sleep of the ring: comparison of the ŌURA sleep tracker against polysomnography. Behav Sleep Med 2019;17(2):124-136. [doi: 10.1080/15402002.2017.1300587] [Medline: 28323455]

XSL•FO
RenderX

15. Lopez X, Afrin K, Nepal B. Examining the design, manufacturing and analytics of smart wearables. Med Devices Sens 2020 Jun;3(3). [doi: 10.1002/mds3.10087]
16. Lujan MR, Perez-Pozuelo I, Grandner MA. Past, present, and future of multisensory wearable technology to monitor sleep and circadian rhythms. Front Digit Health 2021;3:721919. [doi: 10.3389/fdgth.2021.721919] [Medline: 34713186]
17. Eylon G, Tikotzky L, Dinstein I. Performance evaluation of Fitbit Charge 3 and actigraphy vs. polysomnography: sensitivity, specificity, and reliability across participants and nights. Sleep Health 2023 Aug;9(4):407-416. [doi: 10.1016/j.sleh.2023.04.001] [Medline: 37270397]
18. Maher C, Ryan J, Ambrosi C, Edney S. Users' experiences of wearable activity trackers: a cross-sectional study. BMC Public Health 2017 Nov 15;17(1):880. [doi: 10.1186/s12889-017-4888-1] [Medline: 29141607]
19. Haghayegh S, Khoshnevis S, Smolensky MH, Diller KR, Castriotta RJ. Accuracy of wristband Fitbit models in assessing sleep: systematic review and meta-analysis. J Med Internet Res 2019 Nov 28;21(11):e16273. [doi: 10.2196/16273] [Medline: 31778122]
20. Lee XK, Chee N, Ong JL, et al. Validation of a consumer sleep wearable device with actigraphy and polysomnography in adolescents across sleep opportunity manipulations. J Clin Sleep Med 2019 Sep 15;15(9):1337-1346. [doi: 10.5664/jcsm.7932] [Medline: 31538605]
21. Bagot KS, Matthews SA, Mason M, et al. Current, future and potential use of mobile and wearable technologies and social media data in the ABCD study to increase understanding of contributors to child health. Dev Cogn Neurosci 2018 Aug;32:121-129. [doi: 10.1016/j.dcn.2018.03.008] [Medline: 29636283]
22. Burkart S, Beets MW, Armstrong B, et al. Comparison of multichannel and single-channel wrist-based devices with polysomnography to measure sleep in children and adolescents. J Clin Sleep Med 2021 Apr 1;17(4):645-652. [doi: 10.5664/jcsm.8980] [Medline: 33174529]
23. Menghini L, Yuksel D, Goldstone A, Baker FC, de Zambotti M. Performance of Fitbit Charge 3 against polysomnography in measuring sleep in adolescent boys and girls. Chronobiol Int 2021 Jul;38(7):1010-1022. [doi: 10.1080/07420528.2021.1903481] [Medline: 33792456]
24. Creaser AV, Clemes SA, Costa S, et al. The acceptability, feasibility, and effectiveness of wearable activity trackers for increasing physical activity in children and adolescents: a systematic review. Int J Environ Res Public Health 2021 Jun 8;18(12):6211. [doi: 10.3390/ijerph18126211] [Medline: 34201248]
25. Ridgers ND, McNarry MA, Mackintosh KA. Feasibility and effectiveness of using wearable activity trackers in youth: a systematic review. JMIR Mhealth Uhealth 2016 Nov 23;4(4):e129. [doi: 10.2196/mhealth.6540] [Medline: 27881359]
26. Lammers AE, Romanowski AL, Baumgartner H, Diller GP, Uebing A. Reference values for wristband accelerometry data in children aged 6-11 years of age. Front Pediatr 2022;10:808372. [doi: 10.3389/fped.2022.808372] [Medline: 35498803]
27. Cummings C, Crochiere R, Lansing AH, Patel R, Stanger C. A digital health program targeting physical activity among adolescents with overweight or obesity: open trial. JMIR Pediatr Parent 2022 Mar 28;5(1):e32420. [doi: 10.2196/32420] [Medline: 35343903]
28. Kasparian AM, Badawy SM. Utility of Fitbit devices among children and adolescents with chronic health conditions: a scoping review. Mhealth 2022;8:26. [doi: 10.21037/mhealth-21-28] [Medline: 35928511]
29. Bremer E, Arbour-Nicitopoulos KP, Tsui B, et al. Feasibility and utility of a Fitbit tracker among ambulatory children and youth with disabilities. Pediatr Exerc Sci 2023 Nov 1;35(4):249-257. [doi: 10.1123/pes.2022-0121] [Medline: 37236617]
30. Garcia JM, Leahy N, Brazendale K, Quelly S, Lawrence S. Implementation of a school-based Fitbit program for youth with autism spectrum disorder: a feasibility study. Disabil Health J 2021 Apr;14(2):100990. [doi: 10.1016/j.dhjo.2020.100990] [Medline: 33011113]
31. Junghans-Rutelonis A, Sim L, Harbeck-Weber C, Dresher E, Timm W, Weiss KE. Feasibility of wearable activity tracking devices to measure physical activity and sleep change among adolescents with chronic pain-a pilot nonrandomized treatment study. Front Pain Res (Lausanne) 2023;4:1325270. [doi: 10.3389/fpain.2023.1325270] [Medline: 38333189]
32. Palermo TM, Li R, Birnie KA, et al. Updated recommendations on measures for clinical trials in pediatric chronic pain: a multiphase approach from the core outcomes in pediatric persistent pain (Core-OPPP) workgroup. Pain 2024 May 1;165(5):1086-1100. [doi: 10.1097/j.pain.0000000000003105] [Medline: 38112633]
33. Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: building an international community of software platform partners. J Biomed Inform 2019 Jul;95:103208. [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]
34. Nelson BW, Low CA, Jacobson N, Areán P, Torous J, Allen NB. Guidelines for wrist-worn consumer wearable assessment of heart rate in biobehavioral research. NPJ Digit Med 2020;3:90. [doi: 10.1038/s41746-020-0297-4] [Medline: 32613085]
35. Comparison of means calculator. MedCalc. URL: https://www.medcalc.org/calc/comparison_of_means.php [accessed 2024-02-18]
36. Mann Whitney U test calculator (wilcoxon rank-sum). Statistics Kingdom. URL: https://www.statskingdom.com/170median_mann_whitney.html [accessed 2025-02-18]
37. Mackintosh KA, Chappel SE, Salmon J, et al. Parental perspectives of a wearable activity tracker for children younger than 13 years: acceptability and usability study. JMIR Mhealth Uhealth 2019 Nov 4;7(11):e13858. [doi: 10.2196/13858] [Medline: 31682585]

XSL•FO
RenderX

38. Wing D, Godino JG, Baker FC, et al. Recommendations for identifying valid wear for consumer-level wrist-worn activity trackers and acceptability of extended device deployment in children. Sensors (Basel) ;22(23):9189. [doi: 10.3390/s22239189]
39. Yurkiewicz IR, Simon P, Liedtke M, Dahl G, Dunn T. Effect of Fitbit and iPad wearable technology in health-related quality of life in adolescent and young adult cancer patients. J Adolesc Young Adult Oncol 2018 Oct;7(5):579-583. [doi: 10.1089/jayao.2018.0022] [Medline: 29924668]
40. Valor J, Sieber S. Uses and attitudes of young people toward technology and mobile telephony. SSRN J 2003:01. [doi: 10.2139/ssrn.459222]
41. Kuan MTY, Voss C, Lopez J, Hemphill NM, Harris KC. Children with congenital heart disease exhibit seasonal variation in physical activity. PLoS One 2020;15(11):e0241187. [doi: 10.1371/journal.pone.0241187] [Medline: 33151936]
42. Nestor BA, Chimoff J, Koike C, et al. Adolescent and parent perspectives on digital phenotyping in youths with chronic pain: cross-sectional mixed methods survey study. J Med Internet Res 2024 Jan 11;26:e47781. [doi: 10.2196/47781] [Medline: 38206665]
43. Henry LM, Hansen E, Chimoff J, et al. Selecting an ecological momentary assessment platform: tutorial for researchers. J Med Internet Res 2024 Jan 4;26:e51125. [doi: 10.2196/51125] [Medline: 38175682]

Abbreviations

DOS: date of surgery
HIPAA: Health Insurance Portability and Accountability Act
IIPT: intensive interdisciplinary pain treatment
IRB: Institutional Review Board
REDCap: Research Electronic Data Capture
SOP: standard operating procedure

Edited by A Mavragani; submitted 01.04.24; peer-reviewed by G Badicu, J Hooker, L Henry; revised version received 06.05.25; accepted 25.05.25; published 30.07.25.

Please cite as:
Nestor BA, Baumer AM, Chimoff J, Delecourt B, Koike C, Tacugue N, Brusseau R, Roy N, Gaytan-Fuentes IA, Sethna N, Wallace D, Kossowsky J
Evaluating Fitbits for Assessment of Physical Activity and Sleep in Pediatric Pain: Feasibility and Acceptability Pilot Study
JMIR Form Res 2025;9:e59074
URL: https://formative.jmir.org/2025/1/e59074
doi:10.2196/59074

© Bridget A Nestor, Andreas M Baumer, Justin Chimoff, Benoit Delecourt, Camila Koike, Nicole Tacugue, Roland Brusseau, Nathalie Roy, Israel A Gaytan-Fuentes, Navil Sethna, Danielle Wallace, Joe Kossowsky. Originally published in JMIR Formative Research (https://formative.jmir.org), 30.7.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Agreement Between TDK Silmee W22 and ActiGraph wGT3X-BT for Estimating Daily Step Counts and Moderate to Vigorous Physical Activity in Free-Living Adults: Comparative Study

Hiroko Shimura[1,2], PhD; Shinpei Okada[3,4], MEd; Kazushi Maruo[1,5], PhD; Kaori Daimaru[1], MPH; Naoki Deguchi[1], PhD; Yoshinori Fujiwara[6], MD, PhD; Hiroyuki Sasai[1], PhD

[1]Research Team for Promoting Independence and Mental Health, Tokyo Metropolitan Institute for Geriatrics and Gerontology, Tokyo, Japan
[2]Department of Liberal Arts, Tokyo University of Technology, 1404-1 Katakura, Hachioji, Tokyo, Japan
[3]Physical Education and Medicine Research Foundation, Nagano, Japan
[4]Department of Preventive Medicine and Public Health, Tokyo Medical University, Tokyo, Japan
[5]Department of Biostatistics, Institute of Medicine, University of Tsukuba, Ibaraki, Japan
[6]Tokyo Metropolitan Institute for Geriatrics and Gerontology, Tokyo, Japan

Corresponding Author:
Hiroko Shimura, PhD
Research Team for Promoting Independence and Mental Health, Tokyo Metropolitan Institute for Geriatrics and Gerontology, Tokyo, Japan

Abstract

Background: Wearable Internet of Things (IoT) devices are powerful tools for remotely collecting intensive longitudinal data. The TDK Silmee W22, a wristband-type wearable IoT device with a built-in 3-axis acceleration sensor, provides minute-by-minute physical activity data such as estimated metabolic equivalents (METs) and step counts. These measurements can be aggregated to daily estimates; however, their accuracies have not been fully explored in adults under free-living settings.

Objective: This study aims to assess the agreement between the TDK Silmee W22 and the research-grade activity monitor, ActiGraph wGT3X-BT, in estimating daily steps and time spent in moderate to vigorous physical activity (MVPA ≥3 METs) in adults under free-living settings.

Methods: A convenience sample of young to older adults was recruited from communities across several prefectures in Japan. Participants concurrently wore a TDK Silmee W22 on their nondominant wrist and an ActiGraph wGT3X-BT on the left side of the waist during waking hours for 7 consecutive days. Data were aggregated to daily steps and time spent in MVPA (≥1952 vertical axis counts/minute for ActiGraph) for each participant. A valid day was defined as having ≥ 10 hours of accumulated ActiGraph wear time with ≥100 and <50,000 accumulated steps from both devices. Each valid day was classified as either an active day (≥10,000 steps/day or ≥21.4 minutes MVPA per day) or an inactive day. Bland-Altman plots combined with multilevel analysis and κ statistics were used to assess the agreement between physical activity estimates from the devices.

Results: A total of 129 participants (n=66, 51.2% women) aged 23 - 89 years provided the final dataset of 884 observations (5 to 7 daily observations/participant). The TDK Silmee W22 estimated an overall mean of 6369 (SE 242) steps/day and 40.3 (SE 1.9) minutes/day spent in MVPA. Although Bland-Altman plots suggested no obvious proportional bias, fixed biases were observed; the TDK Silmee W22 estimated −1203 steps/day (95% limits of agreement [LoA] −4202 to 1796) and +5 minutes/day (LoA−23 to 34) spent in MVPA compared with those estimated by the ActiGraph wGT3X-BT. TDK Silmee W22 and ActiGraph wGT3X-BT, respectively, classified 14% and 23.1% as active by daily step counts (κ=0.65, 95% CI 0.59 - 0.72), and 70.4% and 60.9% as active by daily time spent in MVPA (κ=0.64, 95% CI 0.59 - 0.69), both indicating moderate agreement.

Conclusions: TDK Silmee W22 underestimated step counts and overestimated time spent in MVPA compared with the research-grade ActiGraph wGT3X-BT, which may lead to misclassification of active and inactive days. Caution is warranted when using TDK Silmee W22 data over relatively short periods, as discrepancies—particularly when compared with research-grade monitors—may affect feedback or goal setting.

(JMIR Form Res 2025;9:e64602) doi:10.2196/64602

KEYWORDS

wearable; internet of things; activity tracker; mHealth; intensive longitudinal data; physical activity; accelerometry; ambulatory monitoring; mobile phone

XSL•FO
RenderX

Introduction

Wearable, accelerometer-based activity monitors have been used to assess physical behavior such as physical activity, sedentary behavior, and sleep in free-living conditions [1]. The embedded sensors can monitor an individual's behavior objectively in noninvasive and nonintrusive ways without interrupting their daily activities. Such monitors typically provide time-stamped data, which allow researchers to investigate within-individual fluctuations (eg, day-to-day changes) over time.

Over the past decade, wristband-type wearable activity trackers (eg, Fitbit, Garmin, and Xiaomi) have become increasingly popular among consumers, and there has been growing interest in using these devices for research [2,3]. These devices were originally designed to motivate individuals to increase their physical activity, which is of public health interest due to the well-documented benefits of physical activity on health [4,5]. Based on the measured triaxial acceleration signals, these activity trackers use the manufacturers' proprietary algorithms to generate behavioral data such as step count, energy expenditure, and sleep time, allowing users to track their daily behaviors [3]. In addition, embedded rechargeable batteries and charging by the wearers themselves enable long-term monitoring, while Internet of Things (IoT) technologies facilitate automatic data upload to cloud storage through a gateway [6,7]. These features make it possible for researchers to remotely collect intensive longitudinal data [8], which consist of repeated observations over time to capture dynamic processes and individual trajectories.

The TDK Silmee W22 (TDKsw; TDK Corporation; see Figure 1) is a wristband-type wearable IoT device intended for use mainly in medical and health care services [9]. It contains multiple sensors (eg, acceleration, photoplethysmography, and skin temperature sensors), a microcontroller for analog-to-digital conversion, flash memory for storage, and a Bluetooth chip for communication, and provides behavioral data such as metabolic equivalents (METs), step count, sleep state, pulse rate, and skin temperature [9]. Data can be transferred from TDKsw to cloud storage via a gateway (smartphone or dedicated hardware), and the device is expected to be used in large-scale epidemiological studies or for delivering just-in-time adaptive interventions [10] to promote physical activity in a manner similar to other wearable IoT devices.

Figure 1. TDK Silmee W22.

However, to date, only 1 study has used the TDKsw to assess physical activity [11], possibly due to insufficient examination of the accuracy of the output measures. Regarding step count, although a previous study reported a significant correlation between video observation and the TDKsw, the data were collected under a short-duration protocol (9 minutes of walking) with a limited sample of older adults [11]. Also, to our knowledge, the accuracy of the estimated METs values remains unclear. Furthermore, whereas the evaluation of the device under daily life conditions is particularly important because the results are more likely to reflect day-to-day activity, no published study has investigated TDKsw accuracy under free-living conditions.

Given these considerations, this study examined the accuracy of the TDKsw in estimating measures of physical activity using day-to-day data obtained from young to older adults in their daily lives. Daily accumulation of step counts and time spent in moderate to vigorous physical activity (MVPA ≥3 METs) were the measures examined. As a reference monitor, we chose the ActiGraph wGT3X-BT (AG; ActiGraph LLC). ActiGraph monitors have been the most frequently used to assess physical activity, accounting for approximately 50% of published studies [12], and a previous model in the same product line, the GT3X+, has been used in large-scale surveillance studies such as the US National Health and Nutrition Examination Survey, supporting the credibility and wide acceptance of this product line. This study aimed to assess the agreement between the TDKsw and AG in estimating the daily accumulation of step counts and time spent in MVPA in adults under free-living conditions.

XSL•FO
RenderX

Methods

Participants

A convenience sample was recruited by word of mouth through friends and colleagues of the authors (H Shimura and SO) from communities across several prefectures in Japan, including both the National Capital Region (Tokyo, Kanagawa, Saitama, and Yamanashi) and other regions (Nagano and Gifu). Eligibility criteria included age between 20 and 89 years, no gait abnormalities, having one's own smartphone, and willingness to wear activity monitors during the monitoring period. Participants aged 70 years and older were recruited from the National Capital Region, while those aged 70 years and younger were recruited from both the National Capital Region and other regions. Efforts were made to achieve a balanced distribution of participants across 10-year age groups (20s to 80s), and to recruit men and women in roughly similar numbers within each age group wherever possible.

Ethical Considerations

Participants who attended the introductory session were given a full explanation of the purpose and potential risks of the study. They signed an institutionally approved informed consent form before data collection. This study was conducted after obtaining approval from the institutional review board of the Tokyo Metropolitan Institute for Geriatrics and Gerontology (R22-013). It was part of the Smart Watch Innovation for Next Geriatrics and Gerontology (SWING) study, an intensive longitudinal study designed to investigate the use of behavioral data, such as physical activity and sleep monitored by wearable devices, to promote health among older adults.

Data Collection

Schedule

Data were collected for approximately 1 week between January and May 2023; the specific data-collection period differed depending on the participant. At the introductory session (day 0), participants completed questionnaires regarding their demographic and anthropometric characteristics. In addition, they started to wear the TDKsw (firmware version 1.F.18) and the AG (firmware version 1.9.2). The Silmee Connect app (version 1.1.2 for iOS and version 1.1.3 for Android; TDK Corporation) was installed on each participant's smartphone, after which their smartphone was paired with their TDKsw. They were instructed to wear both TDKsw and AG simultaneously during waking hours until day 7 and return the devices to the research team via postal mail using a prepaid return envelope on day 8.

After returning the devices, participants received a ¥5000 (approximately US $36) gift card as compensation for their time and effort in the study. This compensation was provided regardless of the amount of data collected to ensure voluntary participation.

Demographic and Anthropometric Characteristics

Age, gender, living arrangement ("living alone" or not), work status ("work full-time," "work part-time," or "not working"), perceived health (5-point Likert scale ranging from "poor" to "excellent"), height, and weight were self-reported. BMI was calculated as kg/m^2. These data were collected on day 0, before wearing the devices.

Physical Activity in Free-Living Conditions

Physical activity data were collected for 7 consecutive days (from day 1 to day 7). Participants were asked to simultaneously wear a TDKsw on their nondominant wrist (using an original nylon band) and an AG on the left side of the waist (using an elastic belt) during waking hours, removing only for water-based activities (eg, showering, bathing, or swimming) or contact sports.

The TDKsw uses signals from an embedded triaxial acceleration sensor to provide proprietary processed minute-by-minute physical activity data in CSV format. In this study, we focused on 2 outputs: METs and step count. The TDKsw is a small (device body size=52 mm×24.5 mm×13.5 mm), light (weight=26 g, including band), and rechargeable wristband-type device embedded with multiple sensors for health monitoring [9], and it can store minute-by-minute data for up to approximately 30 days. In this study, a dummy account was created for each participant, and the participant's smartphone was used as a gateway to send the data to the cloud server, where the TDKsw data were downloaded. To handle missing data on the server, such as that caused by a communication error, the SilmeePro Wx software (version 2.1.0.0; TDK Corporation) was used to download the data from the device when needed.

The waist-worn AG was used as a reference to assess physical activity in free-living conditions because it is a valid and reliable monitor for estimating METs [13] and step count [14,15]. The AG is a research-grade triaxial accelerometer-based activity monitor (size=46 mm×33 mm×15 mm and weight=19 g); both raw and proprietary processed data can be downloaded using dedicated software. In this study, 1-minute epoch data on activity count and step count were downloaded in CSV format, using ActiLife software (version 6.13.4; ActiGraph LLC), in order to match data output from the TDKsw. Because the ActiLife algorithm uses only the vertical axis data in determining steps taken [16,17], the waist, which is near the center of mass, was considered more appropriate than the wrist to prevent false-positive step counts due to nonambulatory wrist movement under free-living conditions [16,18].

Activity Monitor Data Processing

The data from the TDKsw and the AG were merged using the participant ID and the timestamp as keys. In doing this, records (minute-by-minute observations) were excluded if there were any missing values (eg, due to TDKsw device malfunction) or if the AG data were classified as nonwear [19,20]. There was no validated algorithm for TDKsw to identify nonwear; therefore, it was deemed that the participants followed the instruction to wear or remove both AG and TDKsw simultaneously. A valid day was defined as having ≥10 hours of accumulated AG wear time (convention for compliant wear time) [21] with an accumulation of ≥100 and <50,000 steps (for exclusion of outliers) [22] from both AG and TDKsw.

Data from each device were aggregated to daily step counts and time spent in MVPA (≥3 METs for the TDKsw and ≥1952

XSL•FO
RenderX

vertical axis activity counts/minute for the AG [13]) for each participant. Also, each valid day was classified as either an active day or an inactive day, using the data of step counts and time spent in MVPA from each device; valid days with ≥10,000 steps/day [23,24] or ≥21.4 minutes MVPA per day (corresponds to ≥150 minutes MVPA per week [5]) were classified as active days.

Statistical Analysis

Baseline characteristic data were summarized as mean (SD) for continuous variables and as frequency and percentage for categorical variables.

The dataset in this study had a hierarchical structure, with daily aggregated data (maximum 7 valid daily observations per participant) nested within individuals. Therefore, multilevel models were used to appropriately account for this clustering. An intercept-only multilevel model was employed to estimate the overall mean (SE) of step counts and time spent in MVPA. Bland-Altman plots [25] were used to visualize the agreement between the TDKsw and the AG for each measure across each day where valid data were available. An intercept-only multilevel model was also applied to estimate the overall mean of the paired differences between the devices (ie, fixed bias) and the 95% limits of agreement (LoA) [26-28]. Proportional bias was evaluated based on the slope parameter of a multilevel model that included the average value between the devices for each outcome as a fixed effect.

To address potential model misspecification, robust inference was applied to all multilevel model analyses [29,30]. In addition, residual normality and homoscedasticity were assessed using diagnostic analyses for the Bland-Altman analysis models.

The proportion of agreement in classification as active day or inactive day was assessed using κ statistics. For interpretation of the κ statistics, we followed the ratings suggested by McHugh [31]: none (0.00 - 0.20), minimal (0.21 - 0.39), weak (0.40 - 0.59), moderate (0.60 - 0.79), strong (0.80 - 0.90), and almost perfect (above 0.90).

Statistical analyses were performed using SPSS 29.0.1.0 (IBM Corporation) and R (version 4.5.0; R Foundation for Statistical Computing). The *lme4* and *clubSandwich* packages in R were used for multilevel modeling [32,33].

Results

All 130 recruited participants agreed to be enrolled in the study and completed data collection. Of these, 1 participant was excluded from the analysis due to unexpected TDKsw malfunction attributable to the device or the user (only 371 out of 10,080 minutes of data were downloaded). Given this substantial data loss, we compared the data downloaded from the server with those extracted directly from each device using the SilmeePro Wx software. As a result, missing data in 27 participants due to communication errors were fully recovered using the software, while 35 participants had unrecoverable missing data caused by device malfunction or human error. Nevertheless, these 35 participants retained at least 92.6% of data during the measurement period. The final sample included 129 participants aged 23-89 years, of whom 51.2% (n=66) were women. Table 1 shows the full details of the participants' characteristics.

XSL•FO
RenderX

Table . Descriptive characteristics of the participants.

Characteristics	Overall (N=129)	Men (n=63)	Women (n=66)
Age (years), mean (SD)	53.5 (18.9)	52.3 (18.5)	54.6 (19.3)
Age group (years), n (%)			
20 - 29	16 (12.4)	6 (9.5)	10 (15.2)
30 - 39	21 (16.3)	14 (22.2)	7 (10.6)
40 - 49	20 (15.5)	10 (15.9)	10 (15.2)
50 - 59	23 (17.8)	13 (20.6)	10 (15.2)
60 - 69	14 (10.9)	4 (6.3)	10 (15.2)
70 - 79	22 (17.1)	9 (14.3)	13 (19.7)
80 - 89	13 (10.1)	7 (11.1)	6 (9.1)
Living area, n (%)			
National Capital Region	59 (45.7)	27 (42.9)	32 (48.5)
Other	70 (54.3)	36 (57.1)	34 (51.5)
Living arrangement, n (%)			
Living alone	36 (27.9)	11 (17.5)	25 (37.9)
Not living alone	93 (72.1)	52 (82.5)	41 (62.1)
Work status, n (%)			
Work full-time	80 (62)	45 (71.4)	35 (53)
Work part-time	20 (15.5)	7 (11.1)	13 (19.7)
Not working	29 (22.5)	11 (17.5)	18 (27.3)
Perceived health, n (%)			
Excellent	23 (17.8)	12 (19)	11 (16.7)
Very good	54 (41.9)	26 (41.3)	28 (42.4)
Good	52 (40.3)	25 (39.7)	27 (40.9)
Height (cm), mean (SD)	163.3 (9.4)	171.3 (5.1)	155.6 (5.4)
Weight (kg), mean (SD)	59.7 (11.4)	67.6 (9.1)	52.2 (7.7)
BMI (kg/m^2), mean (SD)	22.3 (3.2)	23.0 (3.0)	21.6 (3.3)
BMI ≥25, n (%)	24 (18.6)	13 (20.6)	11 (16.7)

A total of 884 daily observations were used for the analysis (5 to 7 daily observations per participant). Estimated overall means of step counts and time spent in MVPA were 6369 (SE 242) steps/day and 40.3 (SE 1.9) minutes/day for the TDKsw, and 7572 (SE 237) steps/day and 34.8 (SE 1.9) min/day for the AG.

Figures 2 and 3 show Bland-Altman plots comparing TDKsw and AG for each metric. The Bland-Altman plots suggested no obvious proportional bias, as the regression slopes were not significant for step counts (mean 0.010, SE 0.017; P=.53) and MVPA (mean 0.014, SE 0.034; P=.676). Meanwhile, fixed biases were observed; the TDKsw estimated 1203 fewer steps per day (LoA −4202 to 1796; see Figure 2) and 5 more minutes per day (LoA −23 to 34) spent in MVPA (see Figure 3) compared with those estimated by the AG.

XSL•FO
RenderX

Figure 2. Bland-Altman plot comparing the TDK Silmee W22 to the ActiGraph wGT3X-BT for daily step counts (884 observations of 129 participants). The solid line represents the mean difference, the dashed lines represent the 95% limits of agreement, and the dot-dash line represents the linear regression lines by multilevel model. AG: ActiGraph wGT3X-BT; TDKsw: TDK Silmee W 22.

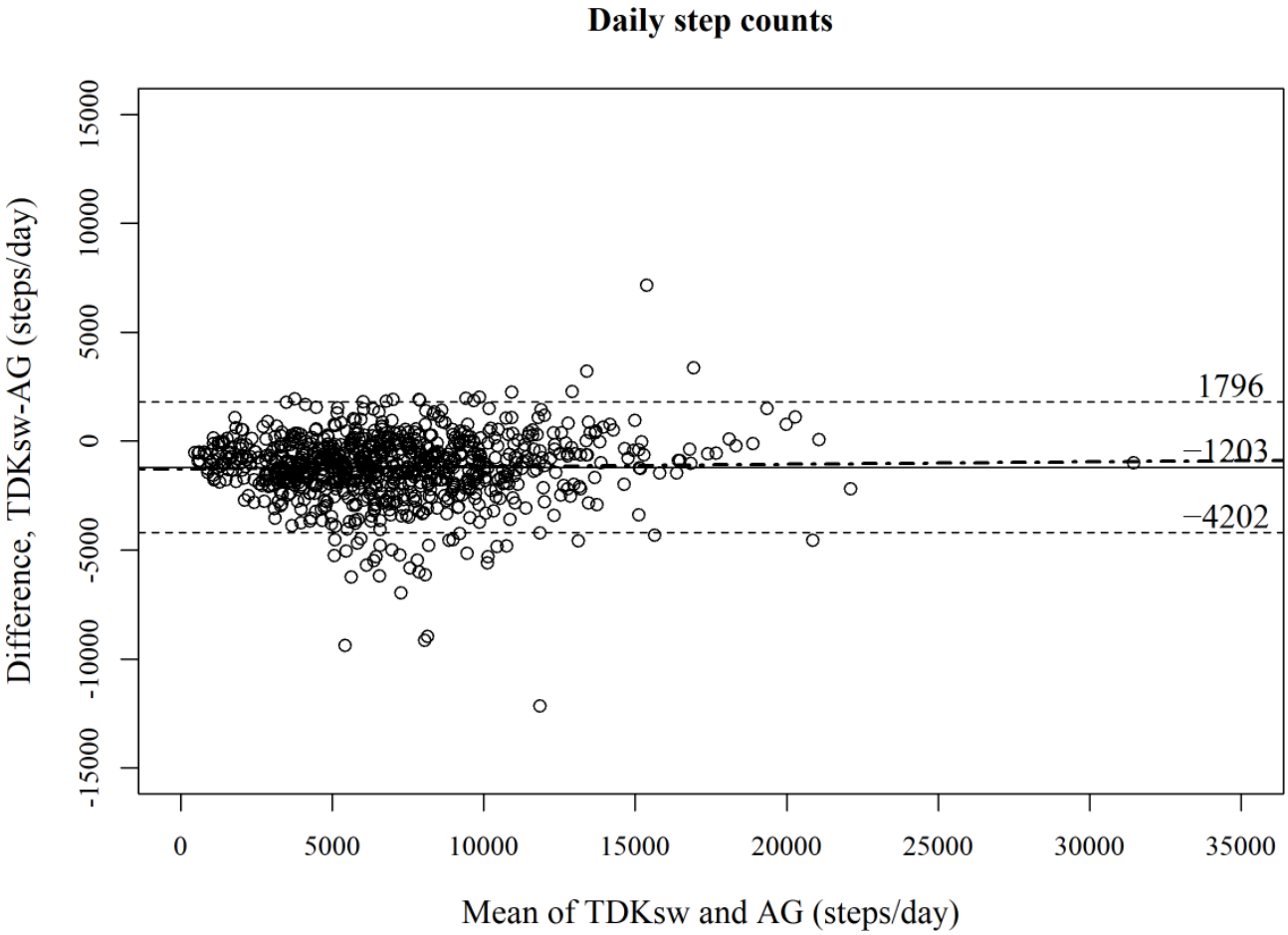

XSL•FO
RenderX

Figure 3. Bland-Altman plot comparing the TDK Silmee W22 to the ActiGraph wGT3X-BT for daily time spent in MVPA (884 observations of 129 participants). The solid line represents the mean difference, the dashed lines represent the 95% limits of agreement, and the dot-dash line represents the linear regression lines by multilevel model. AG: ActiGraph wGT3X-BT; TDKsw: TDK Silmee W 22; MVPA: moderate to vigorous physical activity.

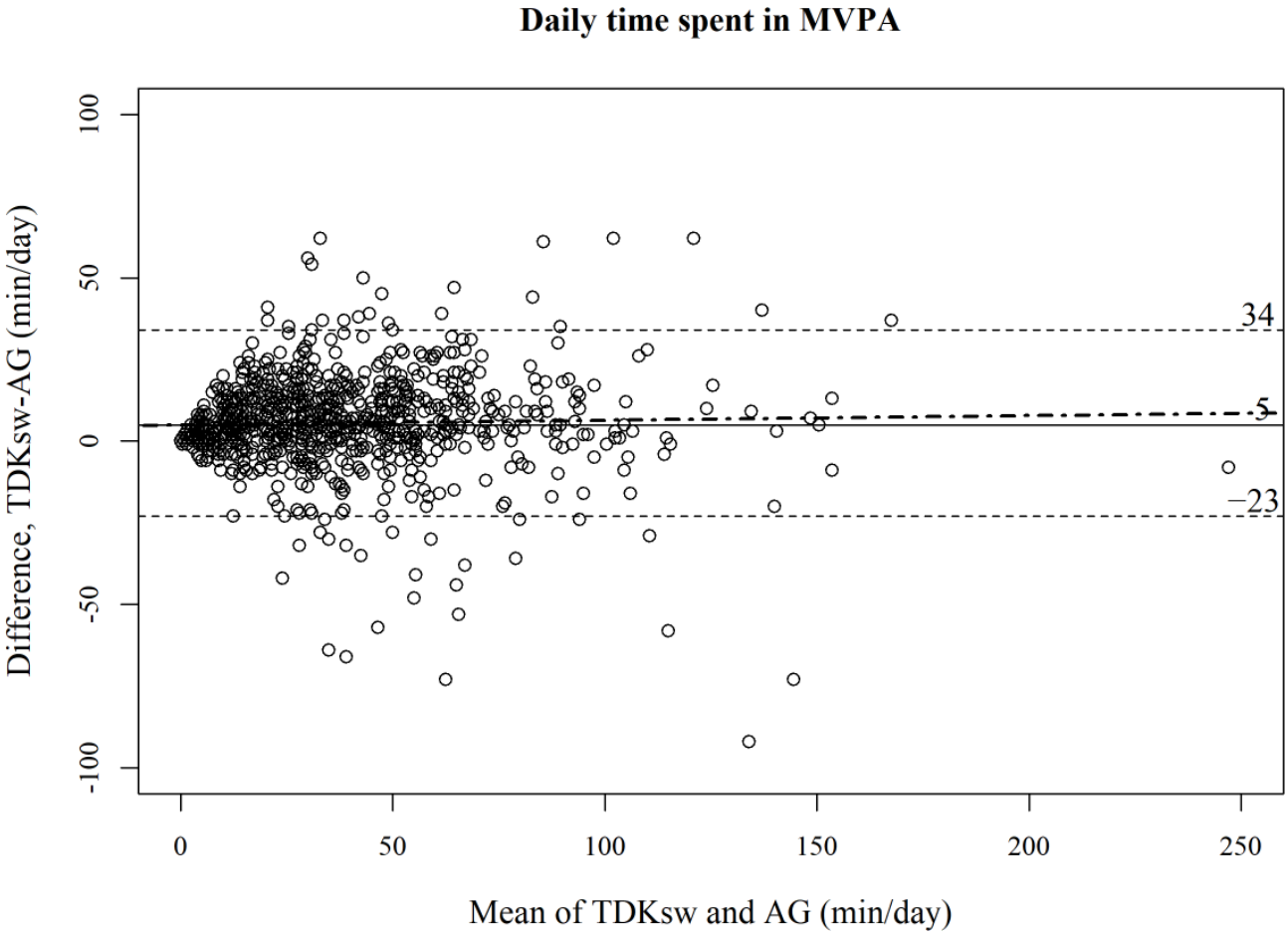

Multimedia Appendix 1 presents the results of residual normality and homoscedasticity diagnostics. Although the distribution of residuals appeared slightly heavy-tailed at both ends, no substantial skewness was observed in either direction. In addition, the residual versus fitted plots did not show any clear pattern of heteroscedasticity, with the spread of residuals remaining approximately constant across the range of fitted values. Based on these results, we considered it acceptable to interpret the Bland-Altman metrics under the assumptions of approximate normality and homoscedasticity.

Tables 2 and 3 show the number of days classified as active or inactive by each device. The TDKsw and AG respectively classified 14% (124/884) and 23.1% (204/884) as active by daily step counts (see Table 2), and 70.4% (622/884) and 60.9% (538/884) by daily time spent in MVPA (see Table 3). The κ statistic was 0.65 (95% CI 0.59 - 0.72) when classified by daily step counts and was 0.64 (95% CI 0.59 - 0.69) when classified by daily time spent in MVPA, both indicating moderate agreement.

Table . Number of days classified as active or inactive by the TDK Silmee W22 and ActiGraph wGT3X-BT, based on daily step counts (884 observations of 129 participants).

	Days classified by AG[b], n		
Days classified by TDKsw[a], n	Inactive	Active[c]	Total
Inactive	673	87	760
Active[c]	7	117	124
Total	680	204	884

[a]TDKsw: TDK Silmee W22.

[b]AG: ActiGraph wGT3X-BT.

[c]Valid days with ≥10,000 steps/day were classified as active.

XSL•FO
RenderX

Table . Number of days classified as active or inactive by the TDK Silmee W22 and ActiGraph wGT3X-BT, based on daily time spent in moderate to vigorous physical activity (MVPA) (884 observations of 129 participants).

	Days classified by AG[b], n		
Days classified by TDKsw[a], n	Inactive	Active[c]	Total
Inactive	231	31	262
Active[c]	115	507	622
Total	346	538	884

[a]TDKsw: TDK Silmee W22.

[b]AG: ActiGraph wGT3X-BT.

[c]Valid days with ≥21.4 minutes MVPA per day were classified as active.

Discussion

Principal Findings

This study assessed the agreement between a wristband-type wearable IoT device, the TDKsw, and a research-grade activity monitor, the AG, worn on the waist, in estimating the daily accumulation of step counts and time spent in MVPA. Agreement was investigated at the day level under free-living conditions in young to older adults. Our results showed some discrepancies between the devices. Compared with the AG, the TDKsw underestimated step counts by approximately 1200 steps per day, whereas it overestimated the time spent in MVPA by 5 minutes per day. These differences seemed to be mostly constant over the measurement range. In addition, the results indicated moderate agreement in classifying whether a given person was active or not on that day; the levels of agreement were similar whether we used the data of daily step count or time spent in MVPA.

Our findings seemed to be counterintuitive, given that the daily step counts are expected to increase as people spend more time in MVPA [34]. A possible explanation for why step counts were underestimated, whereas MVPA was overestimated in this study, is that the MVPA measured by the TDKsw was not explicitly identified as walking. However, the exact reason remains unclear because this study used data that were preprocessed by the manufacturers' proprietary algorithms.

The differences between the two devices are not trivial, equivalent to approximately 12 minutes of walking per day (assuming 100 steps per minute [35]) and one-fifth of the recommended level of physical activity (150 minutes of MVPA per week [5]). In addition, our results may have important implications for promoting physical activity. That is, some TDKsw users may be identified as active on that day even if in fact they are not, and vice versa. Such misclassification, which stems from the fixed bias observed in step counts and MVPA estimates, may be considered unacceptable for research or clinical purposes when assessing the amount or intensity of physical activity at a given point in time. However, when examining within-individual fluctuations of physical activity behavior over time, such differences would not substantially affect the interpretation of the data. The key advantage of the TDKsw is that it enables long-term monitoring and remote collection of large amounts of data by using IoT technology.

Although the κ statistics indicated moderate agreement, whether this level is acceptable depends on the intended use. As noted above, it may be insufficient for point-in-time assessments in clinical or research settings, but potentially acceptable for tracking temporal trends within individuals.

Comparison With Previous Work

Comparing the present results with other free-living studies was challenging because there were differences in terms of methodology and the devices used. The only study to date that has investigated the accuracy of the TDKsw is Kimura et al [11], which reported a significant correlation (r=0.987; P<.001) between TDKsw step count and 9 minutes of video observation in older adults (n=20). Other related studies include research on the agreement between wristband-type activity trackers and waist-worn research-grade devices on a day-to-day basis. Chu et al [36] (n=35 males and n=69 females; median 31.0, IQR 26.0 - 42.8 years; 7 days) reported that the Fitbit Flex overestimated steps/day relative to the waist-worn AG (median difference 1300 steps/day). Mikkelsen et al [37] (n=10 males and n=31 females; mean 47.6, SD 10.4 years; 7 days) reported that compared with the waist-worn ActiGraph GT3X, the Fitbit Charge 2 measured 1492 more steps per day (LoA −2250 to 5234), and 31 fewer MVPA minutes per day (LoA −132 to 71) with increasing difference as time in MVPA increased, suggesting that counterintuitive results might be obtained, as observed in our study. It should be noted that there was variability among the studies in the choice of ActiGraph count data (vertical axis or vector magnitude) and data-processing algorithms, which might have contributed to the mixed findings.

Strength and Limitations

This study has several strengths. First, it included a relatively large sample of participants with a wide range of ages, equal gender balance, and various sociodemographic backgrounds. Second, the study was conducted under free-living conditions, which improved the ecological validity and the generalizability of the findings to real-life settings [38]. Third, we kept the observations (ie, daily aggregated data) as repeated measures for each participant and used multilevel analysis to calculate the LoA. This provides a more accurate representation of the actual agreement by lowering the risk of producing LoA that is too narrow [26]. Earlier agreement studies have rarely used multilevel analysis, despite the nested structure of the repeated measurement [27,37].

XSL•FO
RenderX

The major limitation of this study is the use of non–gold standard methods for comparing steps and MVPA. Having said that, there is no feasible gold standard method for assessing these measures in free-living conditions. In this study, we regarded the waist-worn AG as the best reference because the algorithms it uses to estimate METs and step count have been validated [13-15] and frequently used in earlier studies. Furthermore, although we instructed the participants to wear and remove both TDKsw and AG at exactly the same time, there is no way to ensure they followed this instruction. In addition, because the algorithms used in the TDKsw are proprietary and not publicly available, we were unable to provide a direct comparison of the algorithms.

Conclusions

Differences were found in estimating the daily accumulation of step counts and time spent in MVPA when comparing outputs between the wrist-worn TDKsw and the research-grade waist-worn AG. It was noted that the TDKsw underestimated step counts and overestimated the time spent in MVPA, which might lead to misclassification of active days and inactive days. Wearable IoT devices such as the TDKsw are expected to be used for large-scale epidemiological studies and for promoting physical activity through long-term monitoring. When using TDKsw data over relatively short periods, caution is warranted, as misclassification—particularly when compared with data from research-grade monitors such as the ActiGraph—may affect feedback or goal-setting. Adjustments to classification thresholds may be necessary depending on the intended application.

Acknowledgments

The authors would like to thank all the study participants for their time and commitment, the research assistants for their help with data collection, and Dr. Shuichi Awata for his helpful discussions and encouragement during this work. This study was supported by Smart Watch Innovation for Next Geriatrics and Gerontology (SWING) project of the Tokyo Metropolitan Government and by a Grant-in-Aid for Early-Career Scientists 19K20177 from the Japan Society for the Promotion of Science (Shimura).

Data Availability

The data sets generated and analyzed during this study are not publicly available, as participants were informed that their data would not be shared with other research institutions or made publicly available, and consent for public data sharing was therefore not obtained.

Authors' Contributions

Shimura H was responsible for conceptualization, investigation, data collection, data curation, formal analysis, and methodology, and drafted the original manuscript. SO and KD contributed to investigation, data collection, and critically reviewed and edited the manuscript. KM supervised formal analysis and critically reviewed and edited the manuscript. ND and YF contributed to the conceptualization of the study and critically reviewed and edited the manuscript. Sasai H served as the project leader, overseeing conceptualization, methodology, supervision of analysis and interpretation of data, and critical review and editing the manuscript. All authors reviewed and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Results of residual normality and homoscedasticity diagnostics.
[DOC File, 237 KB - formative_v9i1e64602_app1.doc]

References

1. Burchartz A, Anedda B, Auerswald T, et al. Assessing physical behavior through accelerometry – State of the science, best practices and future directions. Psychol Sport Exerc 2020 Jul;49:101703. [doi: 10.1016/j.psychsport.2020.101703]
2. Feehan LM, Geldman J, Sayre EC, et al. Accuracy of Fitbit devices: systematic review and narrative syntheses of quantitative data. JMIR Mhealth Uhealth 2018 Aug 9;6(8):e10527. [doi: 10.2196/10527] [Medline: 30093371]
3. Shin G, Jarrahi MH, Fei Y, et al. Wearable activity trackers, accuracy, adoption, acceptance and health impact: a systematic literature review. J Biomed Inform 2019 May;93:103153. [doi: 10.1016/j.jbi.2019.103153] [Medline: 30910623]
4. Haskell WL, Blair SN, Hill JO. Physical activity: health outcomes and importance for public health policy. Prev Med 2009 Oct;49(4):280-282. [doi: 10.1016/j.ypmed.2009.05.002] [Medline: 19463850]
5. Global recommendations on physical activity for health. World Health Organization. 2010. URL: https://www.who.int/publications/i/item/9789241599979 [accessed 2024-07-22]

XSL•FO
RenderX

6. Sim I. Mobile devices and health. N Engl J Med 2019 Sep 5;381(10):956-968. [doi: 10.1056/NEJMra1806949] [Medline: 31483966]
7. Majumder S, Mondal T, Deen MJ. Wearable sensors for remote health monitoring. Sensors (Basel) 2017 Jan 12;17(1):130. [doi: 10.3390/s17010130] [Medline: 28085085]
8. Walls TA, Jung H, Schwartz JE. Models for intensive longitudinal data. In: Walls TA, Schafer JL, editors. Models for Intensive Longitudinal Data: Oxford University Press; 2006:3-37. [doi: 10.1093/acprof:oso/9780195173444.003.0001]
9. Miyamoto K, Hashimoto K, Kasaoka M, Kakumu M. Wearable sensors corresponding to various applications in medical/healthcare field. Presented at: 2020 27th International Workshop on Active-Matrix Flatpanel Displays and Devices (AM-FPD); Sep 1-4, 2020; Kyoto, Japan. [doi: 10.23919/AM-FPD49417.2020.9224474]
10. Nahum-Shani I, Smith SN, Spring BJ, et al. Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. Ann Behav Med 2018 May 18;52(6):446-462. [doi: 10.1007/s12160-016-9830-8] [Medline: 27663578]
11. Kimura N, Aso Y, Yabuuchi K, et al. Modifiable lifestyle factors and cognitive function in older people: a cross-sectional observational study. Front Neurol 2019;10:401. [doi: 10.3389/fneur.2019.00401] [Medline: 31068892]
12. Wijndaele K, Westgate K, Stephens SK, et al. Utilization and harmonization of adult accelerometry data. Med Sci Sports Exerc 2015;47(10):2129-2139. [doi: 10.1249/MSS.0000000000000661]
13. Freedson PS, Melanson E, Sirard J. Calibration of the Computer Science and Applications, Inc. accelerometer. Med Sci Sports Exerc 1998 May;30(5):777-781. [doi: 10.1097/00005768-199805000-00021] [Medline: 9588623]
14. Lee JA, Williams SM, Brown DD, Laurson KR. Concurrent validation of the Actigraph gt3x+, Polar Active accelerometer, Omron HJ-720 and Yamax Digiwalker SW-701 pedometer step counts in lab-based and free-living settings. J Sports Sci 2015;33(10):991-1000. [doi: 10.1080/02640414.2014.981848] [Medline: 25517396]
15. Kaminsky LA, Ozemek C. A comparison of the Actigraph GT1M and GT3X accelerometers under standardized and free-living conditions. Physiol Meas 2012 Nov;33(11):1869-1876. [doi: 10.1088/0967-3334/33/11/1869] [Medline: 23111061]
16. Tudor-Locke C, Barreira TV, Schuna JM Jr. Comparison of step outputs for waist and wrist accelerometer attachment sites. Med Sci Sports Exerc 2015 Apr;47(4):839-842. [doi: 10.1249/MSS.0000000000000476] [Medline: 25121517]
17. Toth L, Paluch AE, Bassett DR Jr, et al. Comparative analysis of ActiGraph step counting methods in adults: a systematic literature review and meta-analysis. Med Sci Sports Exerc 2024 Jan 1;56(1):53-62. [doi: 10.1249/MSS.0000000000003282] [Medline: 37703308]
18. John D, Morton A, Arguello D, Lyden K, Bassett D. “What Is a Step?” differences in how a step is detected among three popular activity monitors that have impacted physical activity research. Sensors (Basel) 2018 Apr 15;18(4):1206. [doi: 10.3390/s18041206] [Medline: 29662048]
19. Choi L, Liu Z, Matthews CE, Buchowski MS. Validation of accelerometer wear and nonwear time classification algorithm. Med Sci Sports Exerc 2011 Feb;43(2):357-364. [doi: 10.1249/MSS.0b013e3181ed61a3] [Medline: 20581716]
20. Choi L, Beck C, Liu Z, Moore R, Matthews CE, Buchowski MS. PhysicalActivity: process accelerometer data for physical activity measurement. r package version 0.2-4. R Project. 2021. URL: https://CRAN.R-project.org/package=PhysicalActivity [accessed 2025-07-14]
21. Lee IM, Shiroma EJ, Evenson KR, Kamada M, LaCroix AZ, Buring JE. Accelerometer-measured physical activity and sedentary behavior in relation to all-cause mortality: the women’s health study. Circulation 2018 Jan 9;137(2):203-205. [doi: 10.1161/CIRCULATIONAHA.117.031300] [Medline: 29109088]
22. The national health and nutrition survey in japan [Article in Japanese]. Office of Nutrition, Health Service Division, Health Service Bureau, Ministry of Health, Labour and Welfare, Japan. 2019. URL: https://www.mhlw.go.jp/content/001066903.pdf [accessed 2024-07-22]
23. Tudor-Locke C, Hatano Y, Pangrazi RP, Kang M. Revisiting “how many steps are enough?”. Med Sci Sports Exerc 2008 Jul;40(7 Suppl):S537-S543. [doi: 10.1249/MSS.0b013e31817c7133] [Medline: 18562971]
24. Tudor-Locke C, Craig CL, Brown WJ, et al. How many steps/day are enough? For adults. Int J Behav Nutr Phys Act 2011 Jul 28;8(1):79. [doi: 10.1186/1479-5868-8-79] [Medline: 21798015]
25. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986 Feb 8;1(8476):307-310. [Medline: 2868172]
26. Bland JM, Altman DG. Agreement between methods of measurement with multiple observations per individual. J Biopharm Stat 2007;17(4):571-582. [doi: 10.1080/10543400701329422] [Medline: 17613642]
27. Parker RA, Weir CJ, Rubio N, et al. Application of mixed effects limits of agreement in the presence of multiple sources of variability: exemplar from the comparison of several devices to measure respiratory rate in COPD patients. PLoS ONE 2016;11(12):e0168321. [doi: 10.1371/journal.pone.0168321] [Medline: 27973556]
28. Parker RA, Scott C, Inácio V, Stevens NT. Using multiple agreement methods for continuous repeated measures data: a tutorial for practitioners. BMC Med Res Methodol 2020 Jun 12;20(1):154. [doi: 10.1186/s12874-020-01022-x] [Medline: 32532218]
29. Bell RM, McCaffrey DF. Bias reduction in standard errors for linear regression with multi-stage samples. Surv Methodol 2002;28(2):169-181 [FREE Full text]

XSL•FO
RenderX

30. Pustejovsky JE, Tipton E. Small-sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. J Bus Econ Stat 2018 Oct 2;36(4):672-683. [doi: 10.1080/07350015.2016.1247004]
31. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22(3):276-282. [Medline: 23092060]
32. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. J Stat Softw 2015:1-48. [doi: 10.18637/jss.v067.i01]
33. ClubSandwich: cluster-robust (sandwich) variance estimators with small-sample corrections. R package version 0.6.0. : R Project; 2025 URL: https://cran.r-project.org/web/packages/clubSandwich/clubSandwich.pdf [accessed 2025-07-14]
34. Tudor-Locke C, Leonardi C, Johnson WD, Katzmarzyk PT, Church TS. Accelerometer steps/day translation of moderate-to-vigorous activity. Prev Med 2011;53(1-2):31-33. [doi: 10.1016/j.ypmed.2011.01.014] [Medline: 21295063]
35. Marshall SJ, Levy SS, Tudor-Locke CE, et al. Translating physical activity recommendations into a pedometer-based step goal: 3000 steps in 30 minutes. Am J Prev Med 2009 May;36(5):410-415. [doi: 10.1016/j.amepre.2009.01.021] [Medline: 19362695]
36. Chu AHY, Ng SHX, Paknezhad M, et al. Comparison of wrist-worn Fitbit Flex and waist-worn ActiGraph for measuring steps in free-living adults. PLoS One 2017;12(2):e0172535. [doi: 10.1371/journal.pone.0172535] [Medline: 28234953]
37. Mikkelsen MLK, Berg-Beckhoff G, Frederiksen P, et al. Estimating physical activity and sedentary behaviour in a free-living environment: a comparative study between Fitbit Charge 2 and Actigraph GT3X. PLoS ONE 2020;15(6):e0234426. [doi: 10.1371/journal.pone.0234426] [Medline: 32525912]
38. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol 2008;4(1):1-32. [doi: 10.1146/annurev.clinpsy.3.022806.091415] [Medline: 18509902]

Abbreviations

AG: ActiGraph wGT3X-BT
IoT: Internet of Things
LoA: limits of agreement
METs: metabolic equivalents
MVPA: moderate to vigorous physical activity
TDKsw: TDK Silmee W22

Edited by J Sarvestan; submitted 24.07.24; peer-reviewed by F Clemente, J Tripette; revised version received 17.06.25; accepted 18.06.25; published 30.07.25.

Please cite as:
Shimura H, Okada S, Maruo K, Daimaru K, Deguchi N, Fujiwara Y, Sasai H
Agreement Between TDK Silmee W22 and ActiGraph wGT3X-BT for Estimating Daily Step Counts and Moderate to Vigorous Physical Activity in Free-Living Adults: Comparative Study
JMIR Form Res 2025;9:e64602
URL: https://formative.jmir.org/2025/1/e64602
doi:10.2196/64602

© Hiroko Shimura, Shinpei Okada, Kazushi Maruo, Kaori Daimaru, Naoki Deguchi, Yoshinori Fujiwara, Hiroyuki Sasai. Originally published in JMIR Formative Research (https://formative.jmir.org), 30.7.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Short-Form Psychoeducation Videos: Process Development Study

Louise Turtle[1], MSc; Helen Alexandra Wesson[1], MSc; Simon Williamson[1,2], MRCPsych; Nathan Hodson[1], MRCPsych

[1]Warwick Medical School, University of Warwick, Gibbett Hill Road, Coventry, United Kingdom

[2]Department of Brain Sciences, Imperial College London, London, United Kingdom

Corresponding Author:
Louise Turtle, MSc
Warwick Medical School, University of Warwick, Gibbett Hill Road, Coventry, United Kingdom

Abstract

Background: Every year, around 1.8 million people in the United Kingdom are referred to NHS Talking Therapies, predominantly for cognitive behavioral therapy (CBT), which is the first-line treatment for common affective and anxiety disorders. However, more than a million of these do not complete their course. Supporting this "missing million" to attend and complete CBT is a policy priority.

Objective: We aimed to coproduce a series of video resources to help patients better prepare for and complete their CBT sessions.

Methods: We structured this project around a development cycle and documented outcomes against the Template for Intervention Description and Replication (TIDieR) checklist to ensure transparent intervention reporting. We assembled an interdisciplinary team to undertake an iterative video development process, composed of 3 subteams. An expert contributor subteam of 21 therapists shared their priorities and preferences for video content and style. A creative subteam of 4 members was responsible for scripting, filming, and editing video content. A project management subteam comprising 4 members (2 project managers, 1 designer, and 1 psychiatrist) distilled insights from the expert contributors and shared them with the creative team; they also presented video content to expert contributors and collected feedback. The process was terminated when expert contributors were satisfied that the videos developed could be shared with their patients.

Results: We conducted 2 development cycles over 7 months between February and August 2024. In total, we produced 12 short-form videos, each 1 minute 14 seconds to 4 minutes 46 seconds long, across 4 distinct presentation styles (animation, patient narrative, therapist vignette, and expert interview). Videos covered topics such as the format of CBT (eg, why there is work to do between therapy sessions) and the psychological content (the value of developing healthy habits). Between 4 and 11 expert contributors reviewed any given batch of videos. Based on early feedback, we removed checklist formats in favor of positive storytelling, slowed pacing, and added subtitles to ensure readability and reduce cognitive load. The termination condition was achieved; expert contributors agreed to share videos with their patients.

Conclusions: We successfully collaborated to produce a series of psychoeducation videos. A major strength of this process was the large number of people from different professional backgrounds involved; this diversity boosted both the validity of the content and the creativeness of the videos. This approach was well-suited to the setting of psychotherapy, where therapists have a detailed understanding of the anxieties and uncertainties of their patients, but we would advise caution in fields where professionals are less attuned to their patients' needs. Support to engage the "missing million" is urgently needed, and psychoeducational videos provide one suitable approach.

(JMIR Form Res 2025;9:e66884) doi:10.2196/66884

KEYWORDS

psychoeducation; co-production; videos; interdisciplinary development; Cognitive Behavioral Therapy; anxiety disorders; depression; psychotherapy; online therapy; mental health

Introduction

Background

In 2023, 1.8 million people were referred for NHS Talking Therapies [1]. Only 1.2 million attend their first session (the "initial assessment"), and less than 700,000 attend their second session (the first session with their allocated therapist) [1]. Improving attendance and completion of talking therapy is a key strategic goal for the National Health Service (NHS) [2]. Uncertainty and misconceptions about therapy are common barriers to attending [3], and likely contribute to the dissatisfaction that frequently underlies dropout [4].

The United Kingdom's National Institute for Health and Social Care Research (NICE) recommends cognitive behavioural therapy (CBT) as first line for mild and moderate depression, with or without accompanying antidepressant

XSL•FO
RenderX

psychopharmacological therapy [5]. NICE also recommends CBT and related therapies as first-line treatment for anxiety disorders, such as generalized anxiety disorder and social phobia, again, with or without antidepressant psychopharmacological therapy [6]. In the United Kingdom, primary care doctors (known as general practitioners) commonly commence antidepressant medication, but historically referred less than 5% of patients with anxiety and depression to an evidence-based psychological treatment [7]. In response, the UK government created the initiative *Improving Access to Psychological Therapies*, now known as "NHS Talking Therapies," in 2008. This initiative increased therapy uptake by offering people in the United Kingdom with depression or anxiety disorders the opportunity to self-refer for a course of talking therapy, typically lasting 8 sessions, while also increasing the therapy workforce by training more than 10,000 new therapists over 10 years [8]. People who self-referred would receive an initial evaluation to appraise their suitability for talking therapy and to establish what form of talking therapy would be most suitable. NHS Talking Therapies has proven an effective approach. Meta-analytic evidence indicates large reductions in depression and anxiety and a medium-sized improvement in work and social adjustment [2].

While demand for NHS Talking Therapies support has increased over the intervening period, engagement has remained inconsistent; so much so that the program has been described as "hemorrhaging patients" [9]. In 2013 - 2014, there were 1.3 million referrals into *Improving Access to Psychological Therapies*; 42% (468,881/1,267,193) completed sufficient sessions [10]. In 2023 - 2024, 37% (670,000) of the 1.8 million people referred actually completed therapy [1]. Unfortunately, there is also no evidence that the "missing million" are either fully recovered or unsuitable for therapy. Therefore, there is a clear need to improve engagement among this group.

High rates of attrition from psychotherapy are widely acknowledged and present a significant challenge for evidence-based psychotherapy services [1,2,11]. In 2010, Self et al [3] drew on data from the United Kingdom NHS to argue that the only demographic factor consistently correlated with poor attrition in the literature is socioeconomic status, leaving extensive scope for individual differences based on beliefs about mental illness and therapy. Since then, there has also been a clear divergence of Pakistani and Bangladeshi British people whose engagement in NHS Talking Therapies has been lower than the median, and this may also be partially attributable to differences in beliefs about mental illness and therapy among other factors [1,12]. Globally, the picture is different, and international meta-analyses drawing heavily on other health care systems have found lower rates of discontinuation (ranging from 20% to 47%) [13,14]. Minority racial status has been associated with dropout in a predominantly North American meta-analysis, although socioeconomic status could not be included due to lack of data [13]. In another North American meta-analysis, race was not a significant predictor of engagement [14]. Diagnosis of a personality disorder or treatment in a university predicted lower engagement, but it is not relevant to NHS Talking Therapies, which focuses on uncomplicated anxiety and depression treated in local NHS clinics. These findings demonstrate, first, the value of considering pathways in isolation, rather than pooling data across health systems [3]. Second, the lack of a clear picture of demographic drivers of attrition may indicate that individual differences in health beliefs explain why some people attend and others do not [3].

One approach to long CBT waiting lists is fully remote, asynchronous CBT. The potential advantages include the convenience of therapy in your own home at your own pace [15]. These approaches, with or without light-touch human support, have had some success, but engagement with fully remote asynchronous programs remains as low as 3% to 25%, partially due to the burdensome and unexciting nature of the content, and patients' wariness of digital therapy [15-17]. Engagement challenges in remote asynchronous CBT have become so significant that Batterham et al [18] developed a digital intervention to increase engagement with their digital intervention. While the long-term solution may be fully remote asynchronous therapy, over the medium term, other means of increasing engagement with human therapists are needed.

In 1992, Hunt and Andrews [19] argued that "the finding that dropouts are ubiquitous in psychotherapy is very damaging, for if patients do not stay for treatment, then there is little point in developing effective treatment" and in 1998, Harris [20] made this call to action: "it is necessary for investigators to move beyond research on correlations of attrition to propose and test theoretical models with clearer implications for preventing attrition." Over the subsequent 27 years, the knotty problem of attrition from psychotherapy remains inadequately addressed, but meta-analysis data shows that increasing the credibility of CBT leads to improved treatment outcomes [21]. In this paper, we worked with NHS Talking Therapies professionals to gather their understanding about why people drop out before or during therapy and with a creative team to develop a psychoeducation intervention to address these concerns.

Aims

This study aimed to develop a series of video resources to reduce nonadherence by providing accurate and trustworthy information about NHS Talking Therapy in an engaging manner.

Methods

We assembled an interdisciplinary team and undertook an iterative process of video development [22-24]. All development work took place over a 6-month period between February and August 2024.

Content Team Composition

Building on the methodology reported by Rosaasen et al [23], we assembled a team with the appropriate set of skills to address the development challenges. A project management subteam facilitated collaboration between a creative subteam (1 psychiatrist, 2 actors, and 1 videography expert) and a subteam of NHS Talking Therapies expert contributors. The project management team liaised with the expert contributors to establish their content ideas and proposed changes, and to liaise with the creative team to present feedback and discuss content development, as illustrated in Figure 1. Although patients were

XSL•FO
RenderX

not included in any of the 3 teams, preparatory work was conducted during early 2024 with a range of stakeholders, including 3 patient groups and 2 consultant experts-by-experience, as this creative process was being developed.

Figure 1. The 3 subteams.

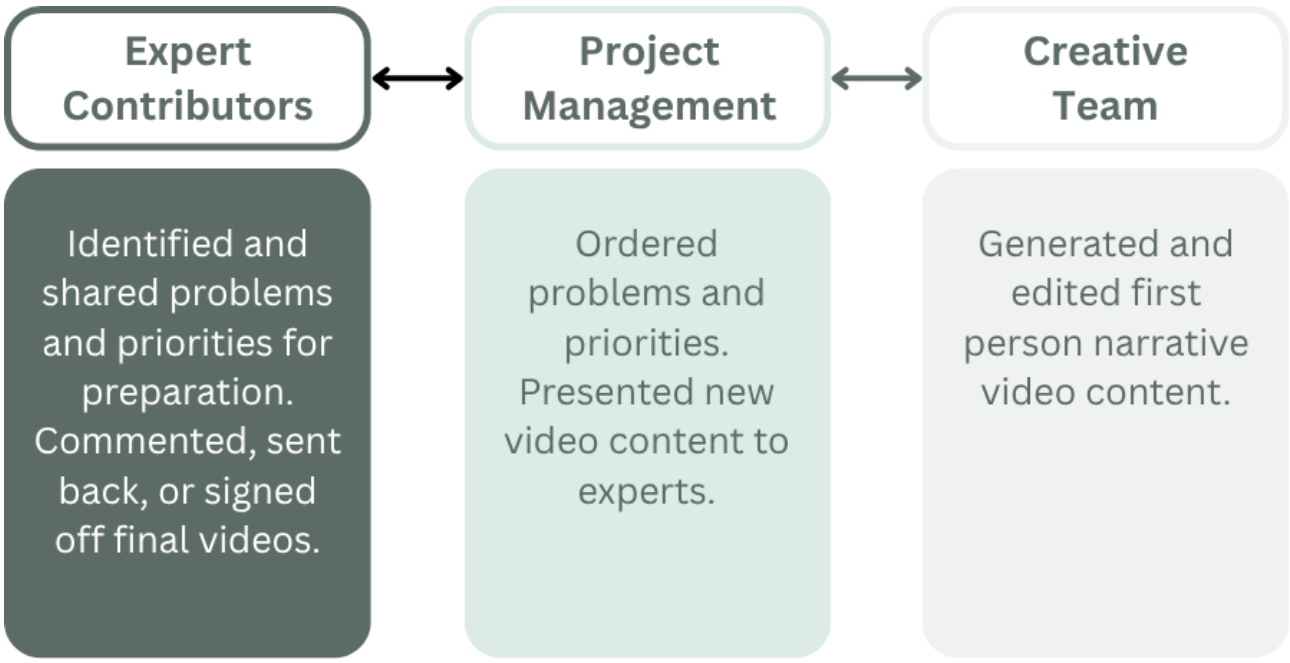

The expert contributors and project management team conducted frequent meetings over the course of the 6-month project. The project management team collated the insights of the expert contributors and presented these to the creative team. The creative team generated multimedia insights and shared them with the project management team, who in turn sought feedback from the expert contributors.

The project management team, recruited through the University of Warwick, had previous expertise in digital project development. It comprised 2 project managers, 1 designer, and 1 psychiatrist. The expert contributors were recruited directly and through snowballing on the basis that they were involved in the provision of CBT services; they included 16 clinicians from NHS mental health trust A and 5 from NHS mental health trust B. A total of 11 of these subject matter experts made repeated contributions throughout the feedback process. The creative team was identified through participation in medical school or community creative projects and included 5 people (3 actors, 1 animator, and 1 psychiatrist).

Content Development Cycle

Video presentations of the psychotherapy experience have previously been identified as an effective way of communicating relevant information, as they have in other contexts [25]. In stage 1a, the project management team established the priorities and preferences of the expert contributors. This was done through a series of meetings between the 2 subteams. Minutes were taken by HAW at each meeting and integrated with information gathered in previous meetings.

In stage 1b, these minutes were presented to the creative team. Within the creative team, the core problem or priority from the expert contributors was distilled, and a narrative was developed. Then, 1 or 2 members of the creative team were allocated the task of writing a short script. This script was edited by other members of the creative team and the project management team. Once the script was agreed, 1 member of the creative team was allocated to make the video through live action filming or animation. The final product was edited using VEED.IO, video production platform (VEED Ltd), followed by final touches (including subtitling).

In stage 2a, the videos were presented to the project management team with any necessary context. The project management team presented the videos back to the expert contributors for their comments. Any further comments from the expert contributors were recorded and delivered back to the creative team with the next set of priorities and preferences. In stage 2b, any further changes required were made either in VEED.IO or through rerecording, and the video was exported and returned to the expert contributors.

This cycle of development and feedback continued for each video until the expert contributors reported that they were happy to share the video with potential users. Given the busy schedules of the expert contributors, most videos were seen by more than 1 expert contributor; any 1 expert contributor could “sign off” on a video or send it back. This process resulted in a series of videos that could be offered to potential service users by expert contributors [26].

Describing the Intervention

The resulting psychoeducation intervention was produced and described using the Template for Intervention Description and Replication (TIDieR) checklist [27]. The TIDieR checklist provides a consistent framework for researchers to describe interventions.

XSL•FO
RenderX

Ethical Considerations

The Biomedical Sciences Research Ethics Committee at the University of Warwick confirmed that this project was exempt from human subject research ethics review as there were no human research subjects. In the absence of human subjects, informed consent was not relevant, nor was confidentiality or compensation.

Results

Timeline of Events

Table 1 illustrates a timeline of the development events.

Table . Timeline of events.

Month	Stage	Subject matter expert progress		Video progress
		Subject matter experts	Meetings	
June	Stage 1	11	2	__[a]
July	Stage 2 - 3	9	5	Initial content production
August	Stage 4 - 5	10	8	Updates from creative team
September	Stage 5	4	2	Termination criteria reached and intervention shared

[a]Not applicable.

Stage 1: Preferences and Priorities Identified by Expert Contributors

The contributing experts reported a wide range of priorities to the project management team. Their recommendations included topics relating to explaining the format of CBT, describing some psychological theory, and using an accessible design style. With respect to the CBT format, the expert contributors suggested describing some features that are uniform to CBT and others that are optional, as well as considering practical considerations for sessions. With respect to the more psychological content, they advised describing the value of building healthy habits and relaxation skills before therapy, the importance of safety from deliberate self-harm and suicide, and the importance of participation in recovery. Finally, they suggested that the style should reflect modern video provision on social media, such as short-form videos, and should draw on color schemes that are suitable for people with neurodevelopmental differences.

Stage 2: Initial Content Production

The creative team scripted, filmed, and edited a pilot series of 8 videos illustrating these messages and including “checklists” of steps to take before beginning CBT. Topics covered were as follows: (1) What is CBT? (2) Where did CBT come from? (3) What can CBT help with? (4) How I told my boss about CBT? (5) How I told my children about CBT? (6) How I told my partner about CBT? (7) A day in the life of a CBT therapist. (8) There is homework in CBT? Furthermore, 4 styles were developed: a corporate information animation (videos 1-3), a friendly former patient (videos 4-6), a therapist (video 7), and an expert interview (video 8). Videos ranged from 1 minute 14 seconds to 4 minutes 46 seconds.

Stage 3: Improvements Suggested by Expert Contributors

The expert contributors rereviewed the initial content developed. They were generally content that the information given about the content of CBT was accurate. However, they were concerned that a proposed checklist of things to do before CBT could be overwhelming for service users. Expert contributors identified several technical points on which the videos could be improved. With respect to the breadth of conditions treated with CBT, they suggested that mentioning dementia could be confusing for service users, but thought that the potential for CBT to help with chronic health conditions should be discussed. They also advised against using the concept of misconceptions and instead focused on facts and positive stories of CBT’s success. Many additional comments related to style were made; they felt initial videos presented too much information too fast, and that key content should be in written form too, not just audio.

Stage 4: Updates Made by Creative Team

The creative team responded to all of the above comments. The concept of checklists was removed completely to remove the binary implication that some people are “ready” and others are “not ready” for CBT. Named conditions were changed, and the focus on anxiety and mood disorders was emphasized, while chronic physical health conditions were included. Likewise, the idea of misconceptions was removed and replaced with more positive framing. Stories of CBT were made more prominent by developing further scripts from the perspective of the friendly former patient. Videos were slowed down, scripts were shortened, and subtitles were added.

Table 2 describes the development process from stage 1 to stage 4.

XSL•FO
RenderX

Table . Summarizing stage 1 to stage 4.

	Stage 1: Examples of key preferences and priorities	Stage 2: Which content covered it?	Stage 3: Suggested improvements	Stage 4: Changes made by the creative team
Information on CBT[a] format				
	There is homework in CBT. Therapists are trained professionals.	7, 8	Checklists to describe practical steps are overwhelming	Checklists were removed.
	Tell the therapist about any personal preferences.	7, 8	—[b]	—
	Consider practicalities such as where CBT happens, transport to CBT during the working day, and balancing CBT with other commitments.	4, 5, 6	—	—
Information about psychological theory				
	Develop healthy habits in advance such as relaxation activities and sleep.	1, 3, 7	Remove dementia, mention chronic conditions.	Main conditions changed.
	Psychological concepts such as thoughts-body-actions-physical feelings.	1, 2, 3, 8	Add stories of CBT benefits and consider mentioning safety on the waiting list	—
	Recovery is not immediate or not guaranteed. Recovery is more likely with CBT.	3, 5, 6	Avoid the concept of "misconceptions"	The concept of misconceptions is replaced with positive framing and positive stories.
Stylistic suggestions				
	Vary audio-visual style and include colors to suit neurodevelopmental differences.	1, 2, 3, 4, 5, 6, 7, 8	Too fast, keep information manageable, make it readable for people who prefer to read	Scripts shortened, subtitles added.

[a]CBT: cognitive behavioral therapy.

[b]Not applicable.

Stage 5: Resulting Intervention Presented to Expert Contributors

After the stage 4 changes were made, the project management team presented the resulting intervention to the expert contributors. The expert contributors were happy to share the videos with their users, and the cycle was terminated. Table 3 describes the resulting intervention, a series of short-form videos, using the TIDieR format [27].

XSL•FO
RenderX

Table . Describing the digital intervention.

TIDieR[a] criteria	Component
Brief name	Get ready for therapy video series
Why	After referral, there are low rates of initiation and completion of CBT[b] through NHS[c] Talking Therapies. Setting expectations and informing prospective patients about CBT is an appropriate first step in addressing this.
What	A short-form video series was developed and no additional interactions were required. Exemplar materials can be accessed online [28].
Who provided	Requires an intervention assistant to share sequential videos via WhatsApp (Meta) or email.
How	Internet, asynchronous, individual.
Where	Internet connection required, access to WhatsApp or email either via smartphone, tablet, or computer.
When and how much	12× short form videos each lasting 1 - 5 minutes. Order given: Watched all at once, or one by one, with the option to rewatch videos later.
Tailoring	Tailoring not required, just targeted for people awaiting NHS Talking Therapies.
Modifications	—[d]
How well (planned)	—
How well (actual)	—

[a]TIDier: Template for Intervention Description and Replication.

[b]CBT: cognitive behavioral therapy.

[c]NHS: National Health Services.

[d]Not applicable.

Discussion

Summary of Findings

The aim of this study was to develop a series of videos to help people prepare for CBT within the NHS Talking Therapies service. Between February and August 2024, we conducted an iterative process of development, combining 21 expert contributors, 4 members of the project management team, and 4 members of the creative team. We produced 12 short-form videos lasting 1-5 minutes. The videos used 4 different styles and covered information about the format of CBT as well as psychological theory. By the end of the second cycle of development, partners at NHS trusts were satisfied by the quality and content and shared videos with patients, achieving the termination condition. We concluded that an iterative interdisciplinary approach is a feasible approach to coproducing psychoeducational videos to the standard required by NHS psychotherapists.

Comparison With the Literature

Approaches to Increasing Engagement With Therapy

Web-based interventions similar to these videos have been designed to increase engagement with eating disorders services. Muir et al [29] developed the MotivATE website to improve engagement with NHS eating disorders services and drew on patient stories. McLean et al [30] developed the Reach Out and Recover website to improve engagement with Australian Eating Disorders services. One advantage in the case of NHS Talking Therapies waiting lists is that patients have already taken the step to seek help, although the long waiting list creates a risk of disengagement. In keeping with those interventions, we anticipate that the use of patient stories of success will reduce levels of disengagement.

Generalizability of the Content Produced

This approach has produced several videos which are highly specific and targeted to the context of NHS Talking Therapies service. While NHS Talking Therapies is by far the largest route to psychotherapy in the United Kingdom, other routes include private providers, university well-being services, and specialist mental health teams [31]. The videos produced in this project are of mixed relevance to these other routes into therapy. Likewise, some would be partially appropriate for those accessing health care systems in other countries, but not all. Videos addressing the core mechanisms of CBT would be applicable widely. However, several “confessional” style videos specifically mentioned a referral from a general practitioner (primary care physician) which could be distracting for those accessing psychotherapy through other routes and could be confusing in many countries where primary care is arranged differently.

Strengths and Limitations

This is the first study to describe the process of developing psychoeducation videos. A major strength is the large number of professionals involved, which boosted both the validity of the content and the creativity of the videos. The use of subteams

XSL•FO
RenderX

(expert contributors, project management, and creative) ensured that we benefited from individuals' areas of expertise, without expecting people to work outside their areas of expertise (eg, asking a psychotherapist to script a video or asking an actor to research the details of therapy). However, the team composition also included a psychiatrist and an actor who is a CBT trainee in the creative team; their presence enabled rapid factual clarification rather than relying entirely on the development cycle.

Another strength arose from the commitment of the expert contributors. The novel concept and high-quality production of the initial videos built enthusiasm among expert contributors who were invested in reviewing the products of the second cycle when they saw the initial videos addressing their preferences and priorities.

The primary limitation of the study is that we did not include anybody currently on the waiting list in the development process. Some would argue that psychoeducation videos should be developed with people who are on the waiting list or were on the waiting list in the past. Yet, many other interventions designed to improve engagement with therapy have also been developed by expert panels without direct patient input [32,33]. In our case, each of the expert contributors had seen tens or hundreds of patients, so drawing on the expert opinion of therapists acted as an impact multiplier. There may be some features of the therapy experience which patients hide from therapists, yet demographically it is highly likely that several expert contributors have lived experience as psychotherapy patients (as colleagues, it would have been inappropriate to ask about this directly). Furthermore, 2 of the paper authors (HAW and NH) are disclosed former psychotherapy patients, further addressing this limitation. Finally, to be sure, former patients turned mental health professionals are likely a more engaged subset of patients than those who do not pursue careers in related fields, but any subgroup of patients willing to spend time in coproduction would also be self-selecting and more enthusiastic than the average participant. By contrast, the professional opinion of our expert contributors was derived from experience treating patients across the entire spectrum of people attending CBT. Nevertheless, we acknowledge that some patients fail to attend even an initial assessment appointment, and therefore, the therapists included do not necessarily have an insight into their mindset beyond what they read in the initial referral. This population is inherently difficult to engage, but in further research, we aim to include them as participants.

Implications for Research and Policy

In general, we cautiously commend this iterative interdisciplinary method of video development to other researchers. This project succeeded in producing videos that therapists considered valuable, illustrating the value of employing a nodal project management team to coordinate a creative team and an expert contributor team. However, our approach relied on the deep understanding of expert contributors regarding the preferences and feelings of their patients; we cannot guarantee that professionals who are not psychotherapists will have the same depth of insight into patients' thoughts, so this approach should be used with caution in other fields. In addition, it is as yet unclear whether these videos are in fact beneficial for people on the waiting list; further study is required to investigate this.

Further research should formally evaluate whether these videos are acceptable and engaging for patients. Then, the objective effects of the intervention on patient engagement and outcomes of psychotherapy should be assessed. Standardized measures of engagement and outcomes are used by the national NHS Talking Therapies program, but qualitative interviewing is more likely to be suitable for evaluating the effect of the videos of subsequent experience of psychotherapy [11]. Randomized controlled trial evidence is important in this field because behavior change interventions in this field have previously led to counterintuitive results [18,34]. If further evidence supports the usefulness of these videos, they could be disseminated through a website or app to reach a wide number of patients in the NHS and other health care systems in a highly scalable way.

Finally, this study illustrates the feasibility of rapidly and cheaply developing psychoeducation materials in a format readily accessible to contemporary tastes. The continued use of paper leaflets and text-dense websites is increasingly difficult to justify; indeed, the validity of *informed* consent could be called into question if patient information materials fail to keep up with the progressing expectations and preferences of patients [35]. It is incumbent upon health care professionals to produce information materials in a convenient and engaging format for patients, and this project has demonstrated that it is possible to do so efficiently and effectively.

Conclusions

We developed a video series to help people prepare for NHS Talking Therapies CBT through an iterative interdisciplinary process. Our experience highlights the value in collaboration between mental health professionals and creatives in helping the "missing million" make the most of their opportunity to benefit from CBT through the NHS.

Acknowledgments

The authors would like to acknowledge the expert contributors and the creative team who supported this project.

Conflicts of Interest

None declared.

XSL•FO
RenderX

References

1. Psychological Therapies, Annual Report on the Use of IAPT Services, 2023-24: NHS Digital: London, UK; 2024.
2. Wakefield S, Kellett S, Simmonds-Buckley M, Stockton D, Bradbury A, Delgadillo J. Improving access to psychological therapies (IAPT) in the United Kingdom: a systematic review and meta-analysis of 10-years of practice-based evidence. Br J Clin Psychol 2021 Mar;60(1):1-37. [doi: 10.1111/bjc.12259] [Medline: 32578231]
3. Self R, Oates P, Pinnock-Hamilton T, Leach C. The relationship between social deprivation and unilateral termination (attrition) from psychotherapy at various stages of the health care pathway. Psychol Psychother 2005 Mar;78(Pt 1):95-111. [doi: 10.1348/147608305X39491] [Medline: 15826408]
4. O'Keeffe S, Martin P, Target M, Midgley N. "I Just Stopped Going": a mixed methods investigation into types of therapy dropout in adolescents with depression. Front Psychol 2019;10:75. [doi: 10.3389/fpsyg.2019.00075] [Medline: 30804827]
5. Depression in Adults: Treatment and Management: National Institute for Health and Care Excellence: London, UK; 2022.
6. Generalised Anxiety Disorder and Panic Disorder in Adults: Management: National Institute for Health and Care Excellence: London, UK; 2011.
7. Clark DM. Realizing the mass public benefit of evidence-based psychological therapies: the IAPT program. Annu Rev Clin Psychol 2018 May 7;14(14):159-183. [doi: 10.1146/annurev-clinpsy-050817-084833] [Medline: 29350997]
8. Stanton E, Wheelan B, Surendranathan T. Have we improved access to mental health services. Health Services Journal. 2013. URL: https://www.hsj.co.uk/have-we-improved-access-to-mental-health-services/5053577.article [accessed 2025-07-28]
9. Richards DA, Borglin G. Implementation of psychological therapies for anxiety and depression in routine practice: two year prospective cohort study. J Affect Disord 2011 Sep;133(1-2):51-60. [doi: 10.1016/j.jad.2011.03.024] [Medline: 21501876]
10. Psychological Therapies: Annual Report on the Use of IAPT Services England, 2014/15: Health and Social Care Information Centre: London, UK; 2015.
11. Final including a report on the IAPT employment advisers pilot, england, may 2023 in psychological therapies: reports on the use of IAPT services. : NHS Digital: London, UK; 2023.
12. Hassan H, Lack S, Salkovskis PM, Thew GR. Acknowledging religion in cognitive behavioural therapy: the effect on alliance, treatment expectations and credibility in a video-vignette study. Br J Clin Psychol 2024 Sep;63(3):347-361. [doi: 10.1111/bjc.12464] [Medline: 38528321]
13. Cooper AA, Conklin LR. Dropout from individual psychotherapy for major depression: a meta-analysis of randomized clinical trials. Clin Psychol Rev 2015 Aug;40:57-65. [doi: 10.1016/j.cpr.2015.05.001] [Medline: 26067572]
14. Swift JK, Greenberg RP. Premature discontinuation in adult psychotherapy: a meta-analysis. J Consult Clin Psychol 2012 Aug;80(4):547-559. [doi: 10.1037/a0028226] [Medline: 22506792]
15. Hodson N, Williamson S. Can large language models replace therapists? Evaluating performance at simple cognitive behavioral therapy tasks. JMIR AI 2024 Jul 30;3:e52500. [doi: 10.2196/52500] [Medline: 39078696]
16. Ebert DD, Berking M, Cuijpers P, Lehr D, Pörtner M, Baumeister H. Increasing the acceptance of internet-based mental health interventions in primary care patients with depressive symptoms. A randomized controlled trial. J Affect Disord 2015 May 1;176:9-17. [doi: 10.1016/j.jad.2015.01.056] [Medline: 25682378]
17. Batterham PJ, Calear AL, Sunderland M, Carragher N, Brewer JL. Online screening and feedback to increase help-seeking for mental health problems: population-based randomised controlled trial. BJPsych Open 2016 Jan;2(1):67-73. [doi: 10.1192/bjpo.bp.115.001552] [Medline: 27703756]
18. Batterham PJ, Calear AL, Sunderland M, Kay-Lambkin F, Farrer LM, Gulliver A. A brief intervention to increase uptake and adherence of an online program for depression and anxiety: protocol for the enhancing engagement with psychosocial interventions (EEPI) randomized controlled trial. Contemp Clin Trials 2019 Mar;78:107-115. [doi: 10.1016/j.cct.2019.01.015] [Medline: 30711664]
19. Hunt C, Andrews G. Drop-out rate as a performance indicator in psychotherapy. Acta Psychiatr Scand 1992 Apr;85(4):275-278. [doi: 10.1111/j.1600-0447.1992.tb01469.x] [Medline: 1595361]
20. Harris PM. Attrition revisited. Am J Eval 1998;19(3):293-305. [doi: 10.1016/S1098-2140(99)80213-6]
21. Kumpasoğlu GB, Campbell C, Saunders R, Fonagy P. Therapist and treatment credibility in treatment outcomes: a systematic review and meta-analysis of clients' perceptions in individual and face-to-face psychotherapies. Psychother Res 2025 Jan;35(1):139-154. [doi: 10.1080/10503307.2023.2298000] [Medline: 38176020]
22. Rola J, Gibbs H, Robertson H, et al. The development of a video-based nutrition education curriculum for patients undergoing radical cystectomy. Bladder Cancer 2023;9(2):151-157. [doi: 10.3233/BLC-230021] [Medline: 38993291]
23. Rosaasen N, Mainra R, Kukha-Bryson A, et al. Development of a patient-centered video series to improve education before kidney transplantation. Patient Educ Couns 2018 Sep;101(9):1624-1629. [doi: 10.1016/j.pec.2018.04.014] [Medline: 29729857]
24. Li EW, Lee A, Vaseghi-Shanjani M, et al. Development and evaluation of a whiteboard video series to support the education and recruitment of committed unrelated donors for hematopoietic stem cell transplantation. Biol Blood Marrow Transplant 2020 Nov;26(11):2155-2164. [doi: 10.1016/j.bbmt.2020.07.008] [Medline: 32673743]
25. Geller JD, Farber BA. Ready when you are: answering your questions about psychotherapy. J Clin Psychol 2020 Aug;76(8):1438-1446. [doi: 10.1002/jclp.22996] [Medline: 32558940]

XSL•FO
RenderX

26. Hodson N, Woods P, Sobolev M, Giacco D. A digital microintervention supporting evidence-based parenting skills: development study using the agile scrum methodology. JMIR Form Res 2024 Jun 28;8:e54892. [doi: 10.2196/54892] [Medline: 38941594]
27. Hoffmann TC, Glasziou PP, Boutron I, et al. Better reporting of interventions: Template for Intervention Description and Replication (TIDieR) checklist and guide. BMJ 2014 Mar 7;348:g1687. [doi: 10.1136/bmj.g1687] [Medline: 24609605]
28. What is CBT? One minute explainer. YouTube. 2024. URL: https://www.youtube.com/watch?v=Ptg0S70Cb5s [accessed 2025-07-01]
29. Muir S, Newell C, Griffiths J, et al. MotivATE: a pretreatment web-based program to improve attendance at UK outpatient services among adults with eating disorders. JMIR Res Protoc 2017 Jul 26;6(7):e146. [doi: 10.2196/resprot.7440] [Medline: 28747295]
30. McLean SA, Caldwell B, Roberton M. Reach out and recover: intentions to seek treatment in individuals using online support for eating disorders. Int J Eat Disord 2019 Oct;52(10):1137-1149. [doi: 10.1002/eat.23133] [Medline: 31298791]
31. Worsley JD, Pennington A, Corcoran R. Supporting mental health and wellbeing of university and college students: a systematic review of review-level evidence of interventions. PLoS ONE 2022;17(7):e0266725. [doi: 10.1371/journal.pone.0266725] [Medline: 35905058]
32. Reis BF, Brown LG. Preventing therapy dropout in the real world: the clinical utility of videotape preparation and client estimate of treatment duration. Professional Psychology: Research and Practice 2006;37(3):311-316. [doi: 10.1037/0735-7028.37.3.311]
33. Olson AL, Gaffney CA, Hedberg VA, Gladstone GR. Use of inexpensive technology to enhance adolescent health screening and counseling. Arch Pediatr Adolesc Med 2009 Feb;163(2):172-177. [doi: 10.1001/archpediatrics.2008.533] [Medline: 19188650]
34. Krampe H, Salz AL, Kerper LF, et al. Readiness to change and therapy outcomes of an innovative psychotherapy program for surgical patients: results from a randomized controlled trial. BMC Psychiatry 2017 Dec 29;17(1):417. [doi: 10.1186/s12888-017-1579-5] [Medline: 29284443]
35. Gigerenzer G, Gaissmaier W, Kurz-Milcke E, Schwartz LM, Woloshin S. Helping doctors and patients make sense of health statistics. Psychol Sci Public Interest 2007 Nov;8(2):53-96. [doi: 10.1111/j.1539-6053.2008.00033.x] [Medline: 26161749]

Abbreviations

CBT: cognitive behavioral therapy
NHS: National Health Service
NICE: National Institute for Health and Social Care Research
TIDieR: Template for Intervention Description and Replication

Edited by A Mavragani; submitted 01.10.24; peer-reviewed by AAL Sawafi, M Zimmermann; revised version received 25.04.25; accepted 02.05.25; published 30.07.25.

Please cite as:
Turtle L, Wesson HA, Williamson S, Hodson N
Short-Form Psychoeducation Videos: Process Development Study
JMIR Form Res 2025;9:e66884
URL: https://formative.jmir.org/2025/1/e66884
doi:10.2196/66884

© Louise Turtle, Helen Alexandra Wesson, Simon Williamson, Nathan Hodson. Originally published in JMIR Formative Research (https://formative.jmir.org), 30.7.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Acceptability and Utility of a Web-Based Patient-Completed Clinical Decision Aid for the Differential Diagnosis of Transient Loss of Consciousness: Qualitative Interview Study

Alistair Wardrope[1,2], MBChB, MSc; Lindsay Blank[3], BSc, PhD; Melloney Ferrar[2], BSc, MSc; Steve Goodacre[2,3], MBChB, PhD; Daniel Habershon[2], BSc, MSc; Markus Reuber[1,2], MD, PhD

[1]Division of Neuroscience, University of Sheffield, Royal Hallamshire Hospital, Glossop Road, Sheffield, United Kingdom

[2]Sheffield Teaching Hospitals NHS Foundation Trust, Sheffield, United Kingdom

[3]Sheffield Centre for Health and Related Research (ScHARR), University of Sheffield, Sheffield, United Kingdom

Corresponding Author:
Alistair Wardrope, MBChB, MSc
Division of Neuroscience, University of Sheffield, Royal Hallamshire Hospital, Glossop Road, Sheffield, United Kingdom

Abstract

Background: Web-based patient-completed clinical decision aids (CDAs) have the potential to reduce inefficient resource use and patient risk in acute and emergency settings while minimizing additional clinician time burdens. However, such interventions must be acceptable for use by their target audience—patients.

Objective: The objective of this study is to assess acceptability and utility to patients of a novel online patient-completed CDA for the differential diagnosis of transient loss of consciousness (TLoC).

Methods: Within a larger validation study of a patient-completed CDA, we conducted nested qualitative semistructured interviews with a purposive sample of 20 patients who used the CDA in the study and performed thematic analysis of interview transcripts.

Results: We identified 11 themes within the data: 3 addressing the content of the CDA, 3 addressing the online implementation, and 4 addressing usability and acceptability of the CDA. Respondents generally felt an online CDA was easy to complete and acceptable, though they felt that increased options to personalize descriptions of their experience would be helpful and offered guidance on how to make it a more useful resource for patients as well as clinicians. We present good practice points for the design of patient-completed online CDAs on the basis of our thematic analysis.

Conclusions: Findings suggest that patient-completed CDAs may be accessible and feasible in acute and emergency settings, though further research is needed to explore their real-world usability. In designing such tools, clinicians should endeavor to maintain their accessibility for all relevant patient groups and to use them to provide direct patient benefit, as well as to support clinical decision-making, for example, through simultaneous patient-directed outputs.

Trial Registration: ClinicalTrials.gov NCT05367999; https://clinicaltrials.gov/study/NCT05367999

(JMIR Form Res 2025;9:e67608) doi:10.2196/67608

KEYWORDS

user experience; virtual medicine; triage; acceptability; qualitative research; qualitative interview; qualitative interview study; online patient-completed clinical decision aids; online; clinical decision; patient risk; burden; consciousness; interview; thematic analysis; usability; clinical decision-making; virtual; medicine; digital; technology; digital health

Introduction

With rising demand for health care services and increasing resource pressures, new tools are needed to ensure more efficient and effective care delivery. Clinical decision aids (CDAs) are sets of instructions that use accessible clinical data to provide clear guidance on appropriate next steps in patient management. Appropriately applied, they reduce inefficient resource use and patient risk in acute and emergency settings [1].

One barrier to the application of CDAs is clinician time burden. This could be addressed through patient self-completion prior to or following clinician assessment. Existing work demonstrates the feasibility of inviting patients to participate in the application of CDAs for their own care [2,3]. Increasingly mobile health technologies such as tablet or smartphone applications are allowing for patient-led data collection and CDA deployment [4]. These have largely been used in outpatient and elective settings for the care of patients with chronic health problems [4-6], but there is evidence that such tools are feasible in the Emergency Department (ED) setting [7-9] and no less acceptable

XSL•FO
RenderX

than clinician-completed use of the same tool [10]. Given that around 9 in 10 patients attending EDs in England in 2018‑19 spent over an hour in the department [11], ED attendances provide ample opportunity for self-administration. Such CDAs could be completed in the waiting room prior to assessment, using departmental tablets (as in this study), or patients' own smartphones.

However, in order for such complex interventions to become applicable to general clinical practice, it is important to understand acceptability to their target audience—patients. Guidelines on developing such interventions highlight the importance of acceptability assessment as part of the process [12,13].

In this paper, we report the acceptability of a novel patient-completed self-assessment CDA to support differential diagnosis of transient loss of consciousness (TLoC) at first presentation—the Paroxysmal Event Symptoms Questionnaire (PESQ). TLoC is a common emergency presentation, but 20%‑30% are not accurately diagnosed or treated at presentation [14-16]. This has significant resource implications—in 2002, the annual direct medical costs of epilepsy misdiagnosis in England and Wales were estimated to be £29,000,000 (US $40-46 million in 2002) [17].

However, the use of CDAs in the diagnosis and management of TLoC presentations has thus far been limited. There are no well-supported criteria for the distinction between all common causes of TLoC. While there are some candidate CDAs designed to discriminate between syncope and bilateral tonic-clonic seizures [18,19] or epilepsy and functional/dissociative seizures [20,21], none covers all relevant presentations, nor has been prospectively validated in mixed TLoC populations. This lack of robust criteria is acknowledged in the UK National Institute of Clinical Excellence guidance on TLoC assessment, which provides very limited suggestions of features that should lead doctors to "suspect epileptic seizures" based on low- and very low-quality evidence and expert opinion alone [22].

In this paper, we use patient experiences of interacting with the PESQ to elucidate understanding of patient requirements of an online, patient-completed CDA. We performed a concurrent nested qualitative study within a larger quantitative research project developing and validating the PESQ in a first-presentation TLoC population. We aimed to identify whether users found the tool acceptable for use, easy to use, and relevant to their clinical presentation.

Methods

The data reported in this paper come from interviews conducted as the qualitative arm of a mixed methods study developing and validating the PESQ for use in a first-presentation TLoC population [23]. The design, setting, participants, interviews, and analysis are described fully elsewhere [24].

Setting and Participants

We conducted the study within a single large hospital Trust in the United Kingdom, screening all patients presenting to the ED with TLoC or referred to first seizure or syncope clinics within the window of February 10, 2022, to January 9, 2023. All participants recruited to the quantitative study were also invited to interview; we approached those who consented to interview sequentially.

We had a provisional target of 30 participants, using empirical data showing 24 participants reliably achieves saturation, the narrow specification of subject matter, and the study team's expertise in the subject matter [25].

Instruments

All participants in the quantitative study completed the PESQ at or shortly after presentation to the ED or primary care with TLoC. The majority completed it using a simple online platform. The PESQ comprises a brief (52-item) questionnaire comprising 4 demographic questions, 13 regarding patient medical history, and 35 peri-ictal symptoms. The online platform delivered these with sequential presentation of each question, with the option for binary yes or no responses (except for the demographic questions of age and years of education, which allowed for any integer answer). Participants could skip questions or navigate back and forth between them. Multimedia Appendix 1 lists PESQ items. For this study, all PESQ data were stored in anonymized fashion on a university secure database.

Patients could complete the questionnaire at the time of assessment and presentation using a tablet computer provided by the study team or else at home after their initial presentation through a simple browser-based application.

Interviews

Two researchers (AW and DH) conducted remote (video, via Microsoft Teams, or telephone) semistructured interviews following a predefined interview schedule (Multimedia Appendix 2). Interviews lasted between 30 minutes and 1 hour. Interviewers noted initial reflections in contemporaneous logs, to support reflexive engagement with later analysis. An independent, nonclinical, professional transcription service transcribed all interviews for subsequent analysis.

Analysis

We undertook thematic analysis [26] of transcribed interviews [27]. One researcher (LB; an experienced qualitative researcher with no clinical background or experience of TLoC) imported transcribed interviews into NVivo (Lumivero), and 2 researchers (LB and AW; a specialty registrar [senior resident] in neurology, with prior experience as a core trainee [junior resident] in cardiology and emergency medicine) developed and refined themes through iterative coding.

We assessed for saturation through interim analyses, stopping recruitment when saturation was reached.

Ethical Considerations

We preregistered the study protocol on ClinicalTrials.gov (NCT05367999). Ethical approval came from the NHS Health Research Authority Edgbaston Research Ethics Committee (IRAS: 304114). We report results in line with the SRQR (Standards for Reporting Qualitative Research; checklist provided in Checklist 1) [28].

XSL•FO
RenderX

Results

Participants and Demographics

Of 2811 potential participants screened for recruitment, 1181 were eligible. Of these, 186 responded to the invitation to participate, and 133 also consented to approach for an interview. We approached 40 participants for interviews, aiming for diversity of age and gender. After 20 interviews, we achieved data saturation.

Of interview participants, 14 (70%) received final diagnoses of syncope, and 6 (30%) epilepsy. In total, 12 participants (60%) were female. The median age was 69 (range 17-90) years. Interviews were held a median of 69 (range 35 - 283) days from initial presentation. Table 1 displays full participant demographics.

Table . Demographics and diagnoses of interview participants. Participants were recruited from all adult referrals to the Emergency Department, Acute Medical Unit, and syncope or seizure clinics with a first transient loss of consciousness (TLoC). They participated in semistructured interviews regarding their experiences of interacting with a prototype patient-completed clinical decision aid for the differential diagnosis of TLoC.

ID	Age	Gender	Diagnosis
TL095	27	F	Syncope
TL099	69	F	Syncope
TL101	70	M	Syncope
TL106	75	M	Syncope
TL109	90	F	Syncope
TL122	17	F	Epilepsy
TL133	73	M	Syncope
TL146	69	F	Syncope
TL150	78	F	Syncope
TL152	40	M	Syncope
TL157	46	M	Epilepsy
TL169	27	F	Epilepsy
TL173	74	F	Syncope
TL174	24	F	Epilepsy
TL176	36	F	Epilepsy
TL178	82	M	Epilepsy
TL180	79	F	Syncope
TL181	70	F	Syncope
TL184	50	M	Syncope
TL188	44	M	Syncope

Themes

The semistructured interview protocol included assessment of both participants' impressions of using the PESQ and of their assessments of its utility and acceptability in clinical practice. We identified 6 themes addressing the former—3 concerning the content of the questions and 3 the design of the tool—and 4 themes addressing the latter. These are summarized in Multimedia Appendix 3.

Content of the Questionnaire

Appropriate Questions

Respondents were largely happy with the content and clarity of the questions asked and did not suggest significant changes. Several respondents (eg, TL109; TL169) reported that the very act of being asked questions—and given the words with which to frame their experience—helped to remind them of the symptoms they had experienced and find the means with which to express them.

Limiting Structure

Other respondents reported that the question structure—in particular, comprising binary yes or no questions—was limiting; they would value the opportunity to elaborate on some of their answers or provide more than yes or no responses. One felt that the questions were too simple and that the opportunity to gain better understanding or "go deeply" into their condition had been missed. More commonly, respondents found they were not always able to give a clear "yes" or "no" answer to some questions; "I appreciate sometimes you wanna force people to make a yes or a no but I didn't […] find that particularly helpful cos it wasn't really yes or no answers" (TL101).

XSL•FO
RenderX

Questions Not Matching Experience

Some participants queried the content of the questions. These issues came in 3 broad forms: an assumption of multiple blackouts when a person had only experienced one, including too many questions that bore little apparent relevance to their experience; or not including questions that were relevant to their experience.

The PESQ used in this study was derived from a questionnaire originally issued to people with long-standing disorders causing TLoC; as a consequence, many questions referred to "my attacks." Respondents who had only experienced a single episode found this phrasing confusing.

Others (eg, TL184) noted that many of the questions referenced experiences or historical characteristics that did not appear relevant to them and found the process of going through lots of questions to give them negative responses difficult; it was "just a little bit frustrating to go through and go no I've not had that and, no I've not seen that, no this hasn't happened to me." The face validity of the questionnaire to respondents appeared to fall when this was the case.

Lastly, some felt that the questions did not sufficiently cover the aspects of their experience—or their general background—that they felt most important. This might be individual symptoms ("I'd lost control of my bladder [but] there was nothing like that to say yes or no" [TL173]), comorbidities, eg, coeliac disease (TL146), or the social milieu of their TLoC ("the first one was actually different to the second one […] if my wife hadn't […] kinda like pushed me and shouted at me, I would have probably gone out" [TL101]).

Design of the Tool

Ease of Use

In nearly all cases, respondents reported that the tool was easy to use and the interface easy to operate. A number of respondents (eg, TL150) noted that the interface was accessible for people who were "not a hundred percent au fait with the […] computer" (ie, low self-evaluated digital literacy). There were no technical issues raised with the exception of 1 respondent (TL173) who had trouble with an outdated internet browser rather than with the tool itself. The simple format with few options and little text per page was highlighted as making navigation straightforward.

Language Used

Respondents overwhelmingly felt that the questions they were answering made sense to them. Some struggled with individual items. This arose in some cases due to language use—either specific terms (eg, "febrile seizures" for TL101) or, more generally, language of a higher complexity that was not "plain English" (TL178).

Going Beyond the Questionnaire

Several participants made suggestions regarding how the tool could be extended beyond the original questionnaire and its intended applications. For some, this concerned other means in which users could narrate their experience than just the questionnaire, for example, supplementing the yes or no questions with free-text descriptions of ictal symptoms. Others felt that the tool could be adapted to support users' informational needs. This could be as simple as providing users with a record of their answers for a reference description of ictal experience they could turn to later (TL099); others thought that after questionnaire completion would be a helpful time for users to be provided with further information about TLoC—either the model's predicted likely diagnosis for the user, with tailored information depending on cause (TL106), or general information regarding the TLoC assessment pathway; users wanted support in "explaining […] the process […] what the next steps would be […] cos I don't want to have to keep going to my GP or the A+E every time I experience these episodes" (TL169).

Clinical Utility and Acceptability

Expectation of Benefit

Many respondents did think that the tool would be helpful in an emergency setting. Those still in situations of diagnostic uncertainty (eg, TL173) felt that it might streamline the assessment pathway or give them answers sooner. Some (eg, TL095) saw the tool as licensing access to definitive investigations. Others (eg, TL169) thought the existence of the questionnaire itself—independent of any diagnostic predictions it could generate—would improve the emergency assessment, by overcoming challenges related to the description of episodes—"it's hard to put things into words yourself…there can be…a lot of confusion at appointments, trying to explain the…episodes."

Unable to Gauge Benefit

Others felt they were unable to assess the potential for benefit in the emergency setting. For some (eg, TL099) this related to the user-generated nature of the questionnaire outputs, which they saw as "my interpretation," contrasted with the "objective method" of medical assessment (TL099). For others, the difficulty rather came from the confusion or disorientation they experienced at the time of assessment, making this a difficult issue for them to adjudicate.

Many patients reported feeling confused or disoriented when asked to complete the questionnaire, even though this was done subsequently in their own home. In at least one case the respondent reported that they had difficulty remembering actually doing it. This in itself raises questions about whether the participants would have been able to complete the questionnaire in an ED shortly after their blackout occurred and whether they would have required help from someone who had witnessed their blackout in order to answer the questions accurately.

Unlikely to Benefit

Two respondents did not think there was any real prospect for the tool improving patient care; however, in both cases, this was because they were entirely satisfied with the standard of care and felt nothing needed to be added to their assessment.

XSL•FO
RenderX

A Tool to Benefit Clinicians and Researchers, Not Patients

Two respondents challenged the question of benefit by drawing a distinction between clinician (or researcher) benefit, on the one hand, and patient benefit, on the other. For these respondents, the questionnaire treated users as "sources of information" [29]—passive objects to be "read" for the clinician's or researcher's epistemic gain—rather than "informants"—peers in the creation and exchange of knowledge. These respondents (eg, TL106) felt that they would want to be able to use the outputs themselves—rather than them simply being passed to clinicians—to experience direct benefit. Information on the assessment pathway was again suggested as one means of realizing patient benefit: "to actually be able to say 'Look, you know, we're gonna investigate this and it's looking like you might have something like this going on so here's some reading'" (TL184).

Discussion

Overview

The experiences of our respondents provide general guidance for the development of patient-completed CDAs in the emergency setting, as well as for subsequent refinement of our specific tool. In general, respondents found a simple online questionnaire tool acceptable and understandable, though they were divided on their assessment of likely clinical benefit. They highlighted some specific features of such tools that they found of particular value; other uses to which such a tool could be put to support the patient navigation of their emergency assessment pathway; and identified important considerations in their design to maximize acceptability and potential benefit.

Value and Limitations of Structured History-Taking

Some respondents found that having a questionnaire as a prompt for ictal recall and description was itself of value. It could serve as an aide-memoire for particular ictal experiences ("it…highlighted…some of the things that applied to me"; TL109) or even could provide the conceptual resources for articulating experiences that are "hard…to put into words yourself, not understanding it" (TL169). Previous work comparing structured and open interviews to extract seizure histories has demonstrated the value of such systematic closed questioning to enhance the yield of ictal history-taking [30,31] and support patients' self-understanding and interpretation of experience [32]; our respondents suggest similar benefits can be found when answers are elicited in a patient-administered, self-report format (more practical in resource-limited settings, with online capture easily facilitating computer-aided interpretation of results). This is consonant with previous work demonstrating that, in general, electronic data collection is more effective than paper-based administration of the same questionnaires [6,33].

However, our participants also highlighted the limitations of this format. They described various ways in which a finite symptom list, with binary yes or no answers, limited their ability to articulate their experience—whether from a surfeit of irrelevant questions, a lack of relevant ones, or feeling constrained by the insistence on the dichotomous presence or absence of a given experience, regarding which they might feel ambivalent or uncertain. Part of this issue may stem from a failure of the study team to articulate fully the purpose of the questionnaire—not to capture a complete description of the person's TLoC experience, but rather to use certain previously identified highly discriminating features to predict likely etiology of the TLoC—but in part their concerns reflect those found more widely in the experiences of those completing symptom questionnaires. Exploring conditions as diverse as breathlessness and depression, patients completing symptom questionnaires describe ambivalent reactions to the depictions of their experience thus captured [34,35]. They may seek to reformulate questions, recontextualize them (as with TL101's avoiding of the present or absent dichotomy, or TL106's looking for follow-up exploration of symptoms), or reject their framing (as with TL184 concluding that the research was not directed toward their presentation) [35].

Beyond just highlighting this issue, our data suggest means to address it. Respondents were clear on the need for such a CDA to complement, not replace, in-person clinical assessment; and they suggested that a more nuanced description of their experience could be captured by complementing the structured questions with free-text symptom reporting. The combination of questionnaires and AI-supported analysis of free speech seizure descriptions may address this potential deficiency of an approach based on closed questions alone [36,37].

Emergency Department Implementation of Patient-Completed CDAs

While much of the research on patient-completed CDAs has focused on their role in the management of chronic conditions [4,38], there is increasing recognition of the opportunities they offer in acute and emergency care. In particular, several studies have explored the role of patient-completed symptom checkers in supporting triage at the ED front door [8,9,39]. While self-triage CDAs for unselected presentations have been found to perform inadequately [39], quantitative assessments of acceptability and utility in this setting have generally shown patients are willing and able to interact with such tools [8,9]. Our study enriches this finding with qualitative data exploring patient preferences for interacting with such tools. As well as having a simple, navigable interface, they reported a preference for patient-facing outcomes to demonstrate benefit to them (explored more below).

In order to develop CDAs fit for supporting emergency and primary care practice, it is important to understand not just how the CDA performs in ideal conditions, but also how it will perform in the "real world" and how clinicians will use it in practice [40,41]; in particular, what are the barriers to and facilitators of physician uptake of a new CDA. Not all factors affecting uptake of CDAs depend on the design of the CDA itself; from a thematic analysis of semistructured interviews with emergency clinicians in the UK, Hayes identified that, along with the design of the CDA itself, factors concerning the care providers, the environment of their practice, and the institutions within which they worked might enable or impede widespread use of CDAs [42]. The Ottawa Acceptability of

XSL•FO
RenderX

Decision Rules Instrument [43], a validated scale for measuring the likely clinical acceptability of new candidate CDAs for emergency care use, similarly considers factors both related to the CDA itself and the context in which it is to be used. The tool-related factors that affect uptake of CDAs can broadly be divided into 3 main areas of concern: how useful the CDA is; how accessible it is to apply in the working context; and how justifiable its decisions are—whether to patients, other clinicians, or in legal contexts. The 5 core principles of CDA design dictate that they must provide the right information to the right person in the right format through the right channel at the right time [44,45]. The majority of CDAs are designed to be completed and used by clinicians, but attending to the core principles of CDA design shows that the right information, right person, right format, and right time for information provision could also come through patients. Patient completion of CDAs prior to clinician assessment could empower patients to participate in their care and support shared decision-making [44], while reducing the time burden on clinicians and so improving the utility and accessibility of the CDA. Computerized CDAs already exist for the management of chronic conditions such as hypertension and multiple sclerosis [46,47]; patient-facing checklists have also been used successfully in ED settings to support patient-centered care [2]. Before real-world implementation, the PESQ would require further research addressing these implementation challenges, as well as multicenter validation of its performance.

It is important that such tools be able to be implemented in the ED setting, not only to allow swift and accurate triage but also because it maximizes capture of reliable information. The richness of subjective symptom descriptions decreases with time from the event for many TLoC presentations, reducing the utility of information captured [48]. Our respondents demonstrated this, with several unable to recall elements of their TLoC they reported in the CDA or even using the CDA itself. Allowing patients to return to their answers to support recall for other clinical purposes would be one means of enhancing the patient utility of such tools. Allowing delayed access to review answers through an online platform would also provide patients with the opportunity to reflect on their responses, an option valued in other work on patient-completed CDAs [38].

Implementation in an ED setting would therefore benefit both from an online application to which users could return after their initial attendance and from hardware available in the ED so that patients could interact with the CDA as early as possible in their pathway. In this study we achieved this through making available to participants in the acute hospital setting a tablet PC, through which they could complete the CDA in its browser-based form; but making it online in this fashion also allowed home completion at a later date. However, further implementation research would be necessary to see how successfully this approach could be used for all patients, including, for example, those with language barriers or who were transiently disoriented at the time of presentation.

Maximizing Patient Benefit From Patient-Completed CDAs

Respondents drew a distinction between clinical benefit—that which achieves the assessing clinician's aims (in particular, diagnosis)—and patient benefit—addressing the patient's immediate needs (prominently, information and management of uncertainty) [24]. They proposed ways in which a CDA could be used for patient benefit. Access to answers after the time of completion (eg, through a print-out, or option to log in and return to questionnaire answers) could help them as a reference point in later assessment and care. They also indicated that the time of questionnaire completion would be an ideal opportunity to address patients' informational needs. An online interface could easily provide patients as well as clinicians information about the results of the CDA and its implications for ongoing management or links to explanations of assessment pathways and interim management while awaiting definitive diagnosis. As reported elsewhere [24], these were predominant concerns of respondents after their first assessment for TLoC, and this represents a significant opportunity to add value to patient-completed CDAs.

While it was not a theme explored extensively by our participants, perhaps because of the study literature clarifying the protocol for research data storage that would differ from widescale health care implementation, data governance will always be important for online CDAs. As with any digital health tool, future iterations of PESQ should consider patient data privacy and security, ensuring compliance with standards such as General Data Protection Regulations (GDPR) or the US Health Insurance Portability and Accountability Act (HIPAA).

Existing tools for supporting CDA development and implementation (such as the Theoretical Domains Framework [38,49] or GUIDES checklist [13]) provide high-level overviews of barriers and enablers of implementation; our data allow for more granular practical guidance. On the basis of our data, we offer a set of good practice points for the development of patient-completed online CDAs, summarized in Table 2.

XSL•FO
RenderX

Table . Considerations to maximize utility and acceptability of a patient-completed online clinical decision aid (CDA), based on recommendations drawn from thematic analysis of semistructured interviews with adult patients using a prototype CDA at first presentation of transient loss of consciousness.

Theme	Considerations
Content	• Language—consider reading age and educational background of your patient group. Pilot questions for intelligibility. • Relevance—consider explaining relevance of different items so that patients do not disengage from questions not matching their experience.
Interface	• Simplicity—keeping the CDA as simple as possible aids engagement for those not otherwise used to interacting with online or computer interfaces. • Flexibility—patients may value the opportunity to adapt their input, eg, through free-text supplements. • Compatibility—If patients will complete the CDA on their own device, ensure cross-compatibility with different browsers and both smartphone and desktop access.
Outputs	• Data storage—patients may want to be able to retrieve responses and outputs for their own use later. Consider how this can be achieved without increasing time demands or complexity of the interface (eg, by accounts and log-ins). Such storage must be done securely, in line with relevant data protection regulations (eg, GDPR[a], HIPAA)[b]. • Direct patient benefit—patients may have different priorities from clinicians in the use of a CDA. Consider what else might be important to patients using the CDA (eg, information about the clinical problem and self-management or lifestyle advice) and how it can be addressed (eg, with CDA outputs providing links to patient support groups or information and self-management guidance).

[a]GDPR: General Data Protection Regulations.

[b]HIPAA: Health Insurance Portability and Accountability Act.

Limitations

We note some important limitations. The inclusion criteria for the quantitative study necessarily left some groups underrepresented, particularly those with insufficient English language proficiency to complete the questionnaire or learning disabilities that would similarly make self-completion difficult. To an extent this is by design—since cross-language and cross-cultural validation would be necessary anyway after development of the English-language version of such a CDA prior to its employment in different languages. We did include within our sample a diverse age range, including older populations who may be less familiar with online tools—it is a strength of this study to note that even those who self-described as “not a hundred percent au fait” with computer use were able to use the tool without difficulty. Previous research has suggested that older adults are less likely to engage with tablet-based mobile health interventions [4], but that when they do are no less likely to complete them than younger patients, suggesting that perceptions of digital literacy may be lower amongst older patients than actual inability to engage with such tools. To remove such barriers, CDA design can emphasize aspects more likely to be acceptable in this demographic [50], such as simple user interfaces and minimizing numbers of clicks and scrolls, without intrusive features like pop-ups.

We also recruited only a small proportion of those eligible to participate. Anecdotally, the study team noted that recruitment rates appeared higher when potential participants were approached directly and contemporaneously with their assessment, rather than retrospectively by letter. Future research could assess differential recruitment rates and effects on sample representativeness of different recruitment pathways.

Additionally, as a measure of acceptability of such a CDA, we only describe qualitative findings; further work could complement this with quantitative data, both survey responses identifying users’ evaluations of acceptability (along dimensions guided by the data presented in this study) and objective measures, eg, noncompletion of the questionnaire.

Further Work

In reporting the main quantitative outcomes of this study, we discuss the further validation and implementation research necessary to determine the practical value of the use of such a CDA in routine practice [23]. The results from this study highlight the need to combine this with further research refining the acceptability and utility of the CDA to patients and clinicians. Furthermore, before application in different cultural contexts and different languages, cross-language and cross-cultural validation would be required. Our present research provides guidance—summarized in Table 2—for how such a program of research could proceed.

Conclusions

In this study we find that an online, patient-completed questionnaire tool to support a CDA for the differential diagnosis of TLoC is generally acceptable to users, though opinions diverge on the likely utility of such a tool. Respondents highlighted the distinction between clinician and patient

XSL•FO
RenderX

benefit—noting that while interests overlap, they are not coextensive—and offered means to improve the tool such that, in addition to supporting a CDA for clinical benefit, it could address the informational needs of patients at the time of their first assessment for TLoC.

Acknowledgments

This study is funded by the NIHR Research for Patient Benefit (RfPB) program (grant NIHR201992). The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

Data Availability

Data are available from the corresponding author upon reasonable request. Data in the form of transcripts are stored on a secure institutional repository. Excerpts will be made available on request to the extent they do not compromise participant anonymity.

Conflicts of Interest

AW received institutional grant funding from the UK National Institute for Health Research to the University of Sheffield to undertake this study. LB received institutional grant funding from the UK National Institute for Health Research to the University of Sheffield to undertake this study. MF received institutional grant funding from the UK National Institute for Health Research to Sheffield Teaching Hospitals NHS Foundation Trust to undertake this study. SG received institutional grant funding from the UK National Institute for Health Research to the University of Sheffield to undertake this study. DH has no relevant conflicts of interest to declare in accordance with the journal's guidelines. MR received institutional grant funding from the UK National Institute for Health Research to the University of Sheffield to undertake this study.

Multimedia Appendix 1
Paroxysmal Event Symptoms Questionnaire.
[DOCX File, 17 KB - formative_v9i1e67608_app1.docx]

Multimedia Appendix 2
Interview schedule.
[DOC File, 34 KB - formative_v9i1e67608_app2.doc]

Multimedia Appendix 3
Summary table of themes and sub-themes with illustrative quotations.
[DOCX File, 19 KB - formative_v9i1e67608_app3.docx]

Checklist 1
SRQR (Standards for Reporting Qualitative Research) checklist.
[DOCX File, 17 KB - formative_v9i1e67608_app4.docx]

References

1. Stiell IG, Bennett C. Implementation of clinical decision rules in the emergency department. Acad Emerg Med 2007 Nov;14(11):955-959. [doi: 10.1111/j.1553-2712.2007.tb02372.x]
2. Clancy M, Health Found. A patient-centred checklist to promote safe, high-quality practice and improved outcomes. URL: https://www.health.org.uk/improvement-projects/a-patient-centred-checklist-to-promote-safe-high-quality-practice-and-improved [accessed 2020-07-20]
3. Blackham JEJ, Claridge T, Benger JR. Can patients apply the Ottawa ankle rules to themselves? Emerg Med J 2008 Nov 1;25(11):750-751. [doi: 10.1136/emj.2008.057877]
4. Kouri A, Yamada J, Sale JEM, Straus SE, Gupta S. Primary Care Pre-Visit Electronic Patient Questionnaire for Asthma: uptake analysis and predictor modeling. J Med Internet Res 2020 Sep 18;22(9):e19358. [doi: 10.2196/19358] [Medline: 32945779]
5. Ferré F, Laurent R, Furelau P, et al. Perioperative risk assessment of patients using the MyRISK digital score completed before the preanesthetic consultation: prospective observational study. JMIR Perioper Med 2023 Jan 16;6(1):e39044. [doi: 10.2196/39044] [Medline: 36645704]
6. Cho J, Han JY, Cho A, Yoo S, Lee HY, Kim H. Enhancing clinical history taking through the implementation of a streamlined electronic questionnaire system at a pediatric headache clinic: development and evaluation study. JMIR Med Inform 2024 Nov 8;12(1):e54415. [doi: 10.2196/54415] [Medline: 39622694]

XSL•FO
RenderX

7. Larkin C, Djamasbi S, Boudreaux ED, et al. ReachCare mobile apps for patients experiencing suicidality in the emergency department: development and usability testing using mixed methods. JMIR Form Res 2023 Jan 27;7(1):e41422. [doi: 10.2196/41422] [Medline: 36705961]
8. Fraser HSF, Cohan G, Koehler C, et al. Evaluation of diagnostic and triage accuracy and usability of a symptom checker in an emergency department: observational study. JMIR Mhealth Uhealth 2022 Sep 19;10(9):e38364. [doi: 10.2196/38364] [Medline: 36121688]
9. Knitza J, Hasanaj R, Beyer J, et al. Comparison of two symptom checkers (Ada and Symptoma) in the emergency department: randomized, crossover, head-to-head, double-blinded study. J Med Internet Res 2024 Aug 20;26(1):e56514. [doi: 10.2196/56514] [Medline: 39163594]
10. Larkin C, Tulu B, Djamasbi S, et al. Comparing the acceptability and quality of intervention modalities for suicidality in the emergency department: randomized feasibility trial. JMIR Ment Health 2023 Oct 24;10(1):e49783. [doi: 10.2196/49783] [Medline: 37874619]
11. NHS Digital, NHS England, NHS Improvement. Hospital accident and emergency activity 2018-19. 2019 Sep. URL: https://files.digital.nhs.uk/F5/ACF07A/AE1819_Annual_Summary.pdf [accessed 2020-07-20]
12. Skivington K, Matthews L, Simpson SA, et al. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. Br Med J 2021 Sep 30;374:n2061. [doi: 10.1136/bmj.n2061] [Medline: 34593508]
13. Van de Velde S, Kunnamo I, Roshanov P, et al. The GUIDES checklist: development of a tool to improve the successful use of guideline-based computerised clinical decision support. Implement Sci 2018 Jun 25;13(1):86. [doi: 10.1186/s13012-018-0772-3] [Medline: 29941007]
14. Smith D, Defalla BA, Chadwick DW. The misdiagnosis of epilepsy and the management of refractory epilepsy in a specialist clinic. QJM 1999 Jan;92(1):15-23. [doi: 10.1093/qjmed/92.1.15] [Medline: 10209668]
15. Jungilligens J, Michaelis R, Popkirov S. Misdiagnosis of prolonged psychogenic non-epileptic seizures as status epilepticus: epidemiology and associated risks. J Neurol Neurosurg Psychiatry 2021 Dec;92(12):1341-1345. [doi: 10.1136/jnnp-2021-326443] [Medline: 34362852]
16. Xu Y, Nguyen D, Mohamed A, et al. Frequency of a false positive diagnosis of epilepsy: a systematic review of observational studies. Seizure 2016 Oct;41:167-174. [doi: 10.1016/j.seizure.2016.08.005] [Medline: 27592470]
17. Juarez-Garcia A, Stokes T, Shaw B, Camosso-Stefinovic J, Baker R. The costs of epilepsy misdiagnosis in England and Wales. Seizure 2006 Dec;15(8):598-605. [doi: 10.1016/j.seizure.2006.08.005] [Medline: 17011217]
18. Sheldon R, Rose S, Ritchie D, et al. Historical criteria that distinguish syncope from seizures. J Am Coll Cardiol 2002 Jul 3;40(1):142-148. [doi: 10.1016/s0735-1097(02)01940-x] [Medline: 12103268]
19. Hoefnagels WA, Padberg GW, Overweg J, van der Velde EA, Roos RA. Transient loss of consciousness: the value of the history for distinguishing seizure from syncope. J Neurol 1991 Feb;238(1):39-43. [doi: 10.1007/BF00319709] [Medline: 2030371]
20. Kerr WT, Janio EA, Braesch CT, et al. Identifying psychogenic seizures through comorbidities and medication history. Epilepsia 2017 Nov;58(11):1852-1860. [doi: 10.1111/epi.13888] [Medline: 28895657]
21. Syed TU, Arozullah AM, Loparo KL, et al. A self-administered screening instrument for psychogenic nonepileptic seizures. Neurology (ECronicon) 2009 May 12;72(19):1646-1652. [doi: 10.1212/WNL.0b013e3181a55ef7] [Medline: 19433737]
22. Westby M, Bullock I, Cooper PN, Davis S, Guideline Development Group. Transient loss of consciousness--initial assessment, diagnosis, and specialist referral: summary of NICE guidance. Br Med J 2010 Sep 2;341:c4457. [doi: 10.1136/bmj.c4457] [Medline: 20813824]
23. Wardrope A, Ferrar M, Goodacre S, et al. Validation of a machine-learning clinical decision aid for the differential diagnosis of transient loss of consciousness. Neurol Clin Pract 2025 Apr;15(2):e200448. [doi: 10.1212/CPJ.0000000000200448] [Medline: 40196464]
24. Wardrope A, Blank L, Ferrar M, Goodacre S, Habershon D, Reuber M. “It is just a big question mark”: a qualitative interview study of patient experiences of the initial assessment of transient loss of consciousness. BMJ Open 2025 Mar 4;15(3):e098045. [doi: 10.1136/bmjopen-2024-098045] [Medline: 40037677]
25. Guest G, Bunce A, Johnson L. How many interviews are enough? An experiment with data saturation and variability. Field methods 2006 Feb;18(1):59-82. [doi: 10.1177/1525822X05279903]
26. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
27. Mason J. Qualitative Researching: Sage Publications Ltd; 2002.
28. O’Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014 Sep;89(9):1245-1251. [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]
29. Fricker M. Epistemic Injustice: Power and the Ethics of Knowing: Oxford University Press; 2009.
30. Devinsky O, Feldmann E, Bromfield E, Emoto S, Raubertas R. Structured interview for partial seizures: clinical phenomenology and diagnosis. J Epilepsy 1991 Jan;4(2):107-116. [doi: 10.1016/S0896-6974(05)80069-6]

XSL•FO
RenderX

31. Whitfield A, Wardrope A, Ardern K, Garlovsky J, Oto M, Reuber M. Subjective seizure symptom reporting in functional/dissociative seizures and epilepsy: effects of sampling technique and patient characteristics. Epilepsy Behav 2023 Aug;145:109331. [doi: 10.1016/j.yebeh.2023.109331] [Medline: 37406557]
32. Wardrope A, Reuber M. The hermeneutics of symptoms. Med Health Care Philos 2022 Sep;25(3):395-412. [doi: 10.1007/s11019-022-10086-z] [Medline: 35503189]
33. Meirte J, Hellemans N, Anthonissen M, et al. Benefits and disadvantages of electronic patient-reported outcome measures: systematic review. JMIR Perioper Med 2020 Apr 3;3(1):e15588. [doi: 10.2196/15588] [Medline: 33393920]
34. Malpass A, Mcguire C, Macnaughton J. “The body says it”: the difficulty of measuring and communicating sensations of breathlessness. Med Humanit 2022 Mar;48(1):63-75. [doi: 10.1136/medhum-2019-011816] [Medline: 33509802]
35. Galasinski D. Constructions of the self in interaction with the Beck Depression Inventory. Health (London) 2008 Oct;12(4):515-533. [doi: 10.1177/1363459308094423] [Medline: 18818278]
36. Pevy N, Christensen H, Walker T, Reuber M. Feasibility of using an automated analysis of formulation effort in patients’ spoken seizure descriptions in the differential diagnosis of epileptic and nonepileptic seizures. Seizure 2021 Oct;91:141-145. [doi: 10.1016/j.seizure.2021.06.009] [Medline: 34157636]
37. Pevy N, Christensen H, Walker T, Reuber M. Predicting the cause of seizures using features extracted from interactions with a virtual agent. Seizure 2024 Jan;114:84-89. [doi: 10.1016/j.seizure.2023.11.022] [Medline: 38091849]
38. Yamada J, Kouri A, Simard SN, Segovia SA, Gupta S. Barriers and enablers to using a patient-facing electronic questionnaire: a qualitative theoretical domains framework analysis. J Med Internet Res 2020 Oct 8;22(10):e19474. [doi: 10.2196/19474] [Medline: 33030437]
39. Trivedi SV, Batta R, Henao-Romero N, Mondal P, Wilson T, Stempien J. A comparison of self-triage tools to nurse driven triage in the emergency department. PLoS One 2024;19(8):e0297321. [doi: 10.1371/journal.pone.0297321] [Medline: 39196994]
40. Brehaut JC, Stiell IG, Visentin L, Graham ID. Clinical decision rules “in the real world”: how a widely disseminated rule is used in everyday practice. Acad Emerg Med 2005 Oct 1;12(10):948-956. [doi: 10.1197/j.aem.2005.04.024]
41. Chan TM, Mercuri M, Turcotte M, Gardiner E, Sherbino J, de Wit K. Making decisions in the era of the clinical decision rule: how emergency physicians use clinical decision rules. Acad Med 2020 Aug;95(8):1230-1237. [doi: 10.1097/ACM.0000000000003098] [Medline: 31789846]
42. Hayes TF. A qualitative exploratory study of emergency medicine clinician perspectives on clinical decision support systems (CSSS) rooted in machine learning in england. : Harvard University; 2020 URL: https://www.proquest.com/docview/2467636298?%20Theses&fromopenview=true&pq-origsite=gscholar&sourcetype=Dissertations%20
43. Brehaut JC, Graham ID, Wood TJ, et al. Measuring acceptability of clinical decision rules: validation of the Ottawa acceptability of decision rules instrument (OADRI) in four countries. Med Decis Making 2010;30(3):398-408. [doi: 10.1177/0272989X09344747] [Medline: 20042533]
44. Sittig DF, Boxwala A, Wright A, et al. A lifecycle framework illustrates eight stages necessary for realizing the benefits of patient-centered clinical decision support. J Am Med Inform Assoc 2023 Aug 18;30(9):1583-1589. [doi: 10.1093/jamia/ocad122] [Medline: 37414544]
45. NHS England. Supporting clinical decisions with health information technology: An implementation guide for clinical decision support systems. 2023. URL: https://www.england.nhs.uk/long-read/supporting-clinical-decisions-with-health-information-technology/ [accessed 2025-01-06]
46. Dorr D, D’Autremont C, Richardson JE, et al. Patient-facing clinical decision support for high blood pressure control: patient survey. JMIR Cardio 2023 Jan 23;7(1):e39490. [doi: 10.2196/39490] [Medline: 36689260]
47. Ziemssen T, Vandercappellen J, Jordan Mondragon V, Giovannoni G. Clinical decision support tool for identifying multiple sclerosis progression. J Clin Med 2022 Jul 28;11(15):4401. [doi: 10.3390/jcm11154401] [Medline: 35956018]
48. Mielke H, Meissner S, Wagner K, Joos A, Schulze-Bonhage A. Which seizure elements do patients memorize? A comparison of history and seizure documentation. Epilepsia 2020 Jul;61(7):1365-1375. [doi: 10.1111/epi.16550] [Medline: 32515852]
49. Atkins L, Francis J, Islam R, et al. A guide to using the Theoretical Domains Framework of behaviour change to investigate implementation problems. Implement Sci 2017 Jun 21;12(1):77. [doi: 10.1186/s13012-017-0605-9] [Medline: 28637486]
50. Hettiarachchi Senarath GM, Delir Haghighi P, Bai L, et al. Barriers and facilitators to the uptake of electronic collection and use of patient-reported measures in routine care of older adults: a systematic review with qualitative evidence synthesis. JAMIA Open 2024 Oct;7(3):ooae068. [doi: 10.1093/jamiaopen/ooae068] [Medline: 39100988]

Abbreviations

CDA: clinical decision aid
ED: emergency department
GDPR: General Data Protection Regulations
HIPAA: Health Insurance Portability and Accountability Act
PESQ: Paroxysmal Event Symptoms Questionnaire
SRQR: Standards for Reporting Qualitative Research

XSL•FO
RenderX

TOLC: transient loss of consciousness

Edited by A Mavragani; submitted 16.10.24; peer-reviewed by K Savolainen, S Joshi; revised version received 16.04.25; accepted 08.05.25; published 24.07.25.

Please cite as:
Wardrope A, Blank L, Ferrar M, Goodacre S, Habershon D, Reuber M
Acceptability and Utility of a Web-Based Patient-Completed Clinical Decision Aid for the Differential Diagnosis of Transient Loss of Consciousness: Qualitative Interview Study
JMIR Form Res 2025;9:e67608
URL: https://formative.jmir.org/2025/1/e67608
doi:10.2196/67608

© Alistair Wardrope, Lindsay Blank, Melloney Ferrar, Steve Goodacre, Daniel Habershon, Markus Reuber. Originally published in JMIR Formative Research (https://formative.jmir.org), 24.7.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

At-Home Bioelectrical Impedance Analysis (BIA) Monitoring of Adult Females at Risk of Breast Cancer-Related Lymphedema: Nonrandomized One-Year Longitudinal Feasibility Study

Cheri Teranishi-Hashimoto[1*], DPT; Monica Padilla[2], PhD; Yoomi Heo[3], MS; Lori Bravi[4*], MS, OTR/L, BCPR, CLT-LANA

[1]REHAB Hospital of the Pacific, Honolulu, HI, United States
[2]InBody BWA, 2550 Eisenhower Avenue, Suite C209, Trooper, PA, United States
[3]InBody Co, Ltd, Seoul, Republic of Korea
[4]Shirley Ryan Ability Lab, Chicago, IL, United States
[*]these authors contributed equally

Corresponding Author:
Monica Padilla, PhD
InBody BWA, 2550 Eisenhower Avenue, Suite C209, Trooper, PA, United States

Abstract

Background: Breast cancer, the most common type of diagnosed cancer in women worldwide, is often associated with the development of lymphedema as a treatment-related effect. Patients undergoing surgery, radiation, or chemotherapy present a higher risk for this side effect. Historically, patients are often not referred to rehabilitation for lymphedema management until the swelling is visible and has progressed, which reduces the chance of reversing the disease progression. Surveillance is key to identifying the earliest signs of breast cancer–related lymphedema, initiating treatment, reversing the disease process, and reducing the impact on function.

Objective: The primary goal of this study was to assess the feasibility of monitoring fluid levels in patients with cancer using a bioimpedance analysis home device and to detect any relevant changes that may correlate with an increased risk of developing lymphedema. Remote monitoring by a clinician has not previously been possible, and a comparable bioelectrical water analyzer device has never been available to patients within the comfort of their homes.

Methods: The study included 8 adult patients diagnosed with unilateral breast cancer who underwent unilateral lumpectomy or mastectomy, bilateral mastectomy, or reconstruction and were followed for 12 months. Patients also underwent radiation or chemotherapy as part of their treatment before the study and, in some cases, during participation in the study. Clinic visits were required every 3 months, with standard care treatment administered by the clinician, as well as daily fluid monitoring, using the extracellular to total body water (ECW/TBW) ratio obtained with the home device.

Results: Preliminary findings from the 8 cases showed that daily monitoring with the home device is possible, and may aid in the detection of fluid changes due to interventions like radiation or chemotherapy; these changes typically subside after treatment, compared to a permanent fluid increase that may indicate lymphedema. While one participant developed lymphedema, there is insufficient data to generalize the feasibility of early detection using the home device. A significant difference between ECW/TBW ratio measurements taken in the morning and evening (P values<.016) was observed for 6 participants, with morning values being higher than evening ones. Additionally, 7 participants showed a higher ECW/TBW ratio in the affected arm compared to the unaffected arm. On average, the ratio between the two values was higher than 1, approximately 75% of the time. The daily monitoring empowered patients to take charge of their health, with more than half expressing a desire to continue using the home device beyond the end of the study period.

Conclusions: This case report shows the feasibility of daily remote monitoring for patients at risk of developing BCRL using a home bioimpedance analysis device.

(JMIR Form Res 2025;9:e73978) doi:10.2196/73978

KEYWORDS

breast cancer-related lymphedema; home-based bioimpedance analyzer (BIA); extracellular water; remote monitoring; case report

XSL•FO
RenderX

Introduction

BCRL (Breast cancer-related lymphedema) is the most common complication arising after breast cancer surgery and treatment, characterized by abnormal swelling postsurgery or even years after treatment [1,2]. While lymphedema can be managed once detected, it remains a lifetime risk, for which patients should remain vigilant [3]. Early surveillance allows risk reduction through therapy interventions and reversal of the early-stage BCRL, particularly stages 0-1 as defined by the International Society of Lymphology. Data show that one in every five women with a history of breast cancer develop BCRL [4], most likely caused by surgical removal of lymph nodes or fluid circulation or radiotherapy-induced scarring of lymph nodes and surrounding circulation, leading to an imbalance in lymphatic fluid. Early diagnosis and precise evaluation are necessary to minimize edema progression and to prevent lymphatic drainage failure. Lymphedema is a chronic condition characterized by swelling from lymphatic fluid buildup and may be primary (hereditary or congenital) or secondary (caused by trauma such as surgery, radiation therapy, or infection). Recent studies have found that other factors, including predisposition to disease may increase the incidence of secondary lymphedema [5]. Early diagnosis and management through therapies (eg, manual drainage and compression bandages), compression garments, exercise, and limb elevation help control symptoms and prevent complications [6,7]. Standard clinical practice involves circumferential or volumetric measurements to monitor patients with cancer after surgery, which occur as often as monthly to every 3 or 6 months, depending on the treatment protocol at the clinic, and whether the patient is referred to therapy. However, subclinical (ie, not visible) lymphedema may go unnoticed as using tape measurement is not sensitive enough to detect such a slight increase in volume. Organizations such as the National Lymphedema Network, the International Society of Lymphology, and the National Accreditation Program for Breast Cancer Centers recommend that patients should be assessed before surgery and then monitored at regular intervals, although no gold standard measurement has been defined [2]. Furthermore, despite the growing awareness and prevalence of BCRL, a significant deficit remains in patient education and breast cancer treatment paradigm for early surveillance of lymphedema [8]. Often, patients are not proactively referred to therapy, but reactively once symptoms are already apparent, and have progressed from Stage 0 to Stage 1 [6,7,9]. Although invasive procedures such as new microsurgery and surgery interventions are being used, complete decongestive therapy remains the most common treatment for BCRL. In advanced stages of lymphedema, patients can develop fatty fibrosis, which may require liposuction procedures to remove the affected tissue [8].

Professional bioelectrical impedance analysis (BIA) devices have been previously used in clinical settings to aid in the detection of lymphedema [10] and to evaluate the effectiveness of provided treatment [11]. These devices send low-amplitude alternating currents through the body at different frequencies to measure the resistance (ie, impedance) to the flow of that current in the body water. These data allow estimation of the complete amount of water in the body (TBW, total body water), the water outside the cells (ECW, extracellular water), and the water inside the cells (ICW, intracellular water) can be estimated. The ECW/TBW ratio serves as an indication of inflammation or edema (ie, EI – Edema Index). The InBody BWA ON (InBody Co. Ltd., Seoul, Korea, via InBody BWA) device uses BIA to assess body composition, including hydration levels, fat, and muscle mass, segmentally, which may aid in lymphedema diagnosis. Like other BIA devices, it is easy to use, fast, reproducible, and in the case of the BWA ON, also portable. Regular measurements with home-use devices such as the BWA ON had not been available before; these tools may enable early lymphedema detection and management, reducing health care burden and quality-of-life costs associated with BCRL diagnosed in later stages. Real-time data measurements are shared with health care providers for customized consultations and prescriptions, evaluation of treatment effectiveness, as well as patient compliance, while also monitoring weight, muscle mass, and body fat percentage—further supporting prevention and management of lymphedema.

Methods

Participants

Eight adult female patients diagnosed with unilateral breast cancer who underwent surgery with lymph node removal 1 - 12 months before study enrollment participated in the study. Exclusion criteria included bilateral cancer, as the aim was to detect fluid differences between the affected and unaffected arms; previous history of breast cancer; existing lymphedema diagnosis; and cellulitis or other active infection at the time of enrollment. Additionally, other exclusion criteria were pregnancy (as it could affect body composition and body fluids), being unable to provide informed consent, prisoners, and other vulnerable populations. As indicated by BIA testing, individuals with pacemakers were excluded. Inclusion criteria were: 18 years old or older, unilateral breast cancer diagnosis, and having undergone a procedure that would increase their risk of developing lymphedema. Patients undergoing bilateral mastectomy (prophylactic) and breast reconstruction could be included. Participants were required to have a phone so they could install the app and test themselves at home. There were no exclusions based on age, BMI, or comorbidities, unless these conditions (eg, kidney disease or heart failure) could affect fluid levels. Participants included in this initial report were tested at two recognized institutions: REHAB Hospital of the Pacific in Honolulu (Site 1) and Shirley Ryan AbilityLab in Chicago (Site 2).

Ethical Considerations

The local institutional review board (IRB) at each site reviewed the study protocol, consent form, and all required details for human research. The IRB used at the REHAB Hospital of the Pacific was the University of Hawaii IRB (UH IRB, IRB2022-00055) and for the SRALAB in Chicago, the Northwestern University IRB was used (NU IRB,

XSL•FO
RenderX

STRU00216604). No study-related activities were initiated before IRB approval was granted.

Informed consent approved by the IRB at each site was discussed with possible participants, and consent was obtained before any research activity was conducted. Possible participants were either referred to or contacted by the researcher or clinician if they met the inclusion criteria, or they reached out directly. Patients who might qualify for the study and required follow-up after unilateral breast cancer diagnosis and treatment were referred by their surgeons or oncologists. They could contact the site directly or be contacted by the researcher or therapist at the corresponding site, based on their preference. After receiving information about the study and what participation would entail, patients could schedule a clinic visit to complete enrollment. Patients referred to the sites for standard treatment or who had already been seen and had expressed willingness to participate in research, were approached by the therapist to determine if they qualified for the study, and details of the study were shared with them.

All questions regarding the informed consent and study activities were answered before consent forms were signed. As with any other research study, participants were allowed to withdraw from the study without any repercussions or without affecting treatment, if they were receiving any type of therapy related to their condition or risks.

Privacy and confidentiality were guaranteed in accordance with IRB-approved protocols and were clearly explained to participants. All data was deidentified, and unique IDs were assigned to each participant; only the researcher or clinician at each site kept that information locked on their site).

All participants were compensated for their participation. They received US $100 for each of the 5 clinic visits (totaling US $500). At Site 1, participants received an additional US $100 for study completion. Site 2 decided not to make this final payment to stay within the US $500 IRS limits. As stated before, clinic visits with standard-of-care were provided at no cost, as well as the ability to monitor themselves with the home device. Given the amount of time involved in daily measurements, compensation for their time was recommended by the researchers.

Procedure

A total of five clinic visits were scheduled at months 0 (the first visit upon enrollment), 3, 6, 9, and 12 months. The following measurements were recorded at each clinic visit: limb volume using tape measurement, BIA, using an InBody professional device (InBody 770; InBody Co. Ltd., Seoul, Korea, via InBody BWA) that has been used for research and clinic [12], range of motion, and quality of life assessments using DASH (The Disabilities of the Arm, Shoulder and Hand) and LLIS (Lymphedema Life Impact Scale), following the standard of care procedure. The clinician also examined the patient for any visible inflammation and asked if they were experiencing any discomfort. According to all the information, they determined the patient's lymphedema stage from 0 to 2 and recommended follow-up treatment, if needed. Figure 1 shows a timeline of the study for reference, with the time before enrollment, starting from cancer diagnosis, surgery and enrollment and continuing with study visits and testing with the BIA home device specified.

Figure 1. Timeline before study initiation and during the study.

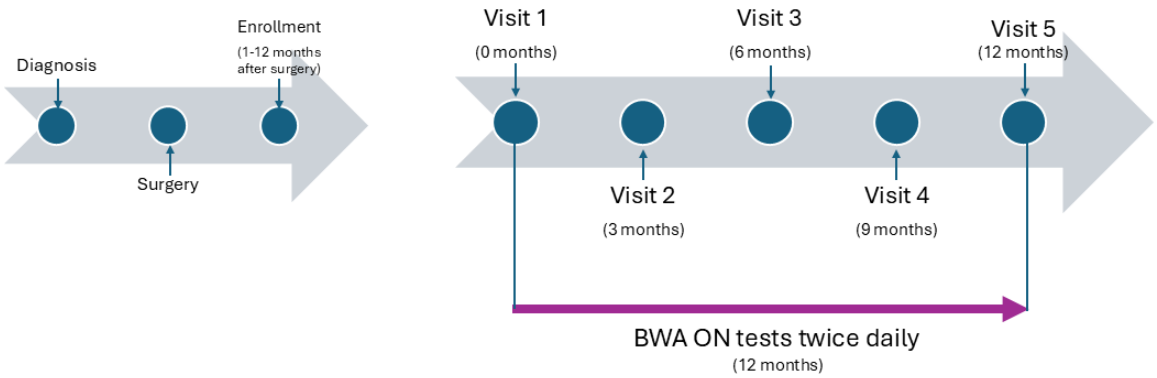

To allow each participant to test themselves at home with a home-use BIA device, the BWA ON (InBody Co. Ltd., Seoul, Korea, via InBody BWA) was provided to them during their first clinic visit. This home BIA device uses 8-touch electrodes to apply small alternating electrical currents to the body at 3 different frequencies (5, 50, and 250 kHz). The BWA ON uses the same technology as professional InBody devices, validated against different gold standards. Devices were validated to DEXA (Dual-Energy X-ray Absorptiometry) by the Mayo Clinic, showing a concordance correlation coefficient (CCC) of 0.98 for fat-free mass (FFM) (5% limits of agreement: −3.5 to +5.2) and 0.97 for percent body fat (PBF) (5% limits of agreement: 6.0 to +4.2). The total body water (TBW) was validated in multiethnic populations against D2O (Deuterium Oxide Solution), showing that it produces valid estimates. Although some home BIA monitoring devices have been used previously [13], to the best of our knowledge, this is the first time that a home device has been available to patients for assessing segmental fluid levels.

XSL•FO
RenderX

The clinicians helped participants with setting up the device, assisted with the first measurement, and ensured the appropriate phone app (Apple or Google) was installed and working on the patient's phone. The participant was registered by the clinician and data sharing with the clinic monitoring software was established.. Participants measured themselves daily: 30 minutes to 1 hour after waking and after 5 PM, before going to bed. Through the Apple or Google app, participants had access to their body composition (eg, weight, skeletal muscle, fat, and visceral fat level) and their extracellular-to-total body water (ECW/TBW) ratio levels. Clinicians were also able to monitor their progress remotely.

Figure 2 shows the home device used by participants at home. The black screen indicated the devices' connection to the phone app via Bluetooth, and the patient could initiate the test. The phone app checked the electrode connections and started the test, showing its progress. Once the test was completed, patients were able to add details about exercise or any issues that may help explain detected changes in fluid levels or share if there was nothing new to report.

Figure 2. Bioelectrical body water analyzer for home use, BWA ON (InBody Co. Ltd., Seoul, Korea, via InBody BWA).

Figure 3 shows screenshots of the information available to participants in the Apple (or Google) app, as well as what the clinicians may see in the computer application to monitor their patients' day-to-day fluid levels.

At the end of the study, participants were asked a few questions to assess the ease of use of the home device. They answered questions regarding daily use and the measurements provided.

XSL•FO
RenderX

Figure 3. (A) Information displayed by the clinicians software and the patient's Apple (or Google) phone app (B) ECW/TBW ratio difference between the two arms is shown in graph mode, (C) segmental ECW/TBW ratio is provided as level for limbs and trunk, and other body composition values (and history) are shown.

Results

Table 1 shows detailed demographic characteristics, surgery dates, and additional treatments required before or during the study, as these factors could affect water levels without lymphedema onset.

XSL•FO
RenderX

Table . Participants' detailed demographics.

Case number	Age (yrs)	Dominant arm (A/L/R)[a]	Diagnosis date	Site	Surgery Date	Type	SLNB[b] or ALND[c]	Radiation (Y/N)[d] (B/D)[e]	Chemotherapy (Y/N)[d] (B/D)[e]
100	66	A	01/24/23	L	01/30/23	Lumpectomy	SLNB	Y-B	N
101	66	R	02/23/23	L	03/21/23	Lumpectomy	SLNB	Y-D	N
102	64	R	02/07/23	L	03/30/23	Lumpectomy	SLNB	Y-D	N
105	52	R	02/20/23	L	03/20/23	Lumpectomy	SLNB	Y-D	N
110	54	R	02/14/23	L	03/10/23	Lumpectomy	SLNB	Y-D	N
112	54	R	04/21/23	L	06/14/23	Lumpectomy	SLNB	Y-D	Y-D
113	52	R	11/19/22	R	08/02/23	Lumpectomy	SLNB	Y-D	Y-B
200[f]	60	R	10/01/22	1. L 2. R	1. 03/10/23 2. 06/15/23	1. Lumpectomy 2. Mastectomy	1. ALND 2. n/a[g]	1. Y-D 2. n/a	1. Y-D 2. n/a

[a](A/L/R): Ambidextrous/Left/Right.

[b]SLNB: Sentinel Lymph Node Biopsy.

[c]ALND: Axillary Lymph Node Dissection.

[d](Y/N): Yes/No.

[e](B/D): Before the study/During the study.

[f]200 was the only participant to undergo prophylactic mastectomy.

[g]n/a: Not applicable.

Previous studies have shown that the prevalence of lymphedema diagnosis among patients undergoing breast cancer surgery is expected to be approximately 21.4% [14]. With eight cases presented in this initial report, only one case of lymphedema—or none—may have developed.

A nonpermanent change in water levels was detected in one participant undergoing chemotherapy treatment. Water levels returned to baseline once the treatment was completed, which coincides with previous reports of chemotherapy causing temporary swelling [15].

Fluid movement during the day could affect fluid distribution. This study evaluated the ECW/TBW ratio difference between the two arms (affected minus unaffected) in the morning and evening measurements and the interlimb (ILB) ECW/TBW ratio (affected arm/unaffected arm).

The daily ECW/TBW ratio measurements for the morning (AM) and evening (PM) are presented for the affected and unaffected arm in Figure 4. A Mann-Whitney U-test was performed on the data to determine if there was a significant difference between AM and PM measurements for each arm for each participant. AM and PM measurements were shown to be significantly different, except for 110's and 113's affected arms. A Rank-biserial correlation to determine effect size was used. Although values were small, most of them are positive, showing that AM values (red) of ECW/TBW ratio are higher than PM values (green), except for both arms for participant 102 and for 113's unaffected arm, which showed that ECW/TBW ratio AM (red) values were lower than PM (green) values (case 100 and 102 Affected, have an opposite direction to what would be expected from the value shown in the table, due to rounding). This information is presented in Table 2, which also shows the number of tests performed by each participant in the AM and PM. Some participants are missing measurements in both the AM and PM, but in most cases, PM measurements were less frequent than AM measurements.

XSL•FO
RenderX

Table . ECW/TBW[a] ratio values and number of measurements for each participant in the morning (AM) and evening (PM).

Case number and arm	ECW/TBW measurements (morning, AM), mean(SD)	Morning (AM) tests, n	ECW/TBW measurements (evening, PM), mean (SD)	Evening (PM) tests, n	*P* value Mann-Whitney *U*-test	Rank biserial correlation, *r*
100		341		314		
Affected	0.386 (0.0013)		0.386 (0.0016)		.001	0.132
Unaffected	0.387 (0.0012)		0.385 (0.0016)		<.001	0.507
101		356		317		
Affected	0.384 (0.0019)		0.382 (0.0023)		<.001	0.467
Unaffected	0.382 (0.0023)		0.380 (0.0026)		<.001	0.457
102		335		319		
Affected	0.383 (0.0020)		0.383 (0.0018)		.02	−0.093
Unaffected	0.382 (0.0021)		0.383 (0.0018)		<.001	−0.142
105		202		124		
Affected	0.381 (0.0046)		0.380 (0.0020)		<.001	0.344
Unaffected	0.381 (0.0049)		0.379 (0.0024)		<.001	0.432
110		313		191		
Affected	0.381 (0.0036)		0.381 (0.0020)		.15	0.064
Unaffected	0.380 (0.0030)		0.380 (0.0020)		<.001	0.222
112		365		350		
Affected	0.383 (0.0048)		0.381 (0.0028)		<.001	0.273
Unaffected	0.380 (0.0025)		0.378 (0.0054)		<.001	0.337
113		315		120		
Affected	0.385 (0.0024)		0.385 (0.0028)		.77	0.014
Unaffected	0.387 (0.0028)		0.389 (0.0024)		<.001	−0.229
200		200		83		
Affected	0.388 (0.0044)		0.386 (0.0039)		.005	0.166
Unaffected	0.381 (0.0028)		0.379 (0.0025)		<.001	0.227
All patients						
Affected	0.384 (0.0039)	2427	0.383 (0.0030)	1818	<.001	0.144

[a]ECW/TBW: extracellular to total body water.

XSL•FO
RenderX

Figure 4. Morning, AM (red) and Evening, PM (green) ECW/TBW ratio measurements of the affected and unaffected arm. The bar plot shows the distribution of all values, and the *P* value of the comparison between the two measurements is shown above to determine if the difference between morning (AM) and evening (PM) measurements is significant. ECW/TBW: extracellular to total body water.

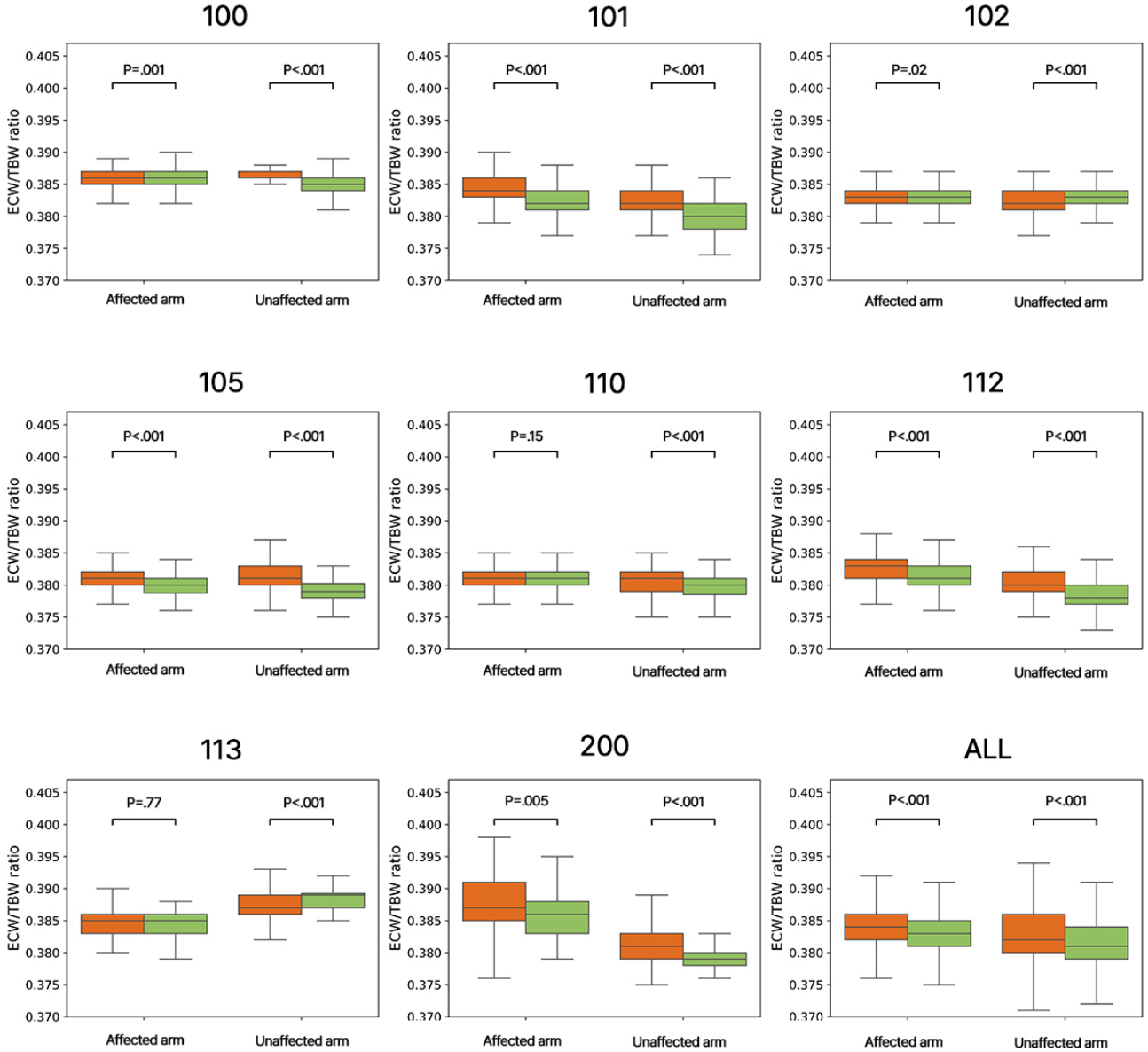

All participants, except for 113, demonstrated at least slightly higher ECW/TBW ratios in the affected arm compared to the unaffected arm, both AM and PM. One potential difference between this participant and the other participants is that participant 113's diagnosis and surgery occurred in the dominant arm. Lymphedema literature has found different cutoffs for dominant and nondominant arms' volume, and BIA impedance ratio [6], so this could help explain this result. Participant 200 shows the largest interarm difference.

As seen in Figure 5, all participants have an interlimb (ILB) value ([$ECW/TBW_{affected\ arm}$]/[$ECW/TBW_{unaffected\ arm}$]) larger than '1' at some point, meaning that the ECW/TBW ratio in the affected arm is larger than the ratio in the unaffected arm. Ideally, the arms are balanced, and the value is '1'. Others have shown that a ratio close to 1 is related to lymphedema onset [16-18]. Considering all participants, occurrences of ILB larger than '1' took place in 75.95%. Participants 101, 110, 112, and 200 show a higher percentage (>80%), with 112 being the highest. Treatments during the study (Table 1) could affect fluid levels, so this average may not confirm lymphedema but may help determine if those changes are permanent or not. Occurrences of values larger than 1 are below average for 100, 102, and 113. Further analysis is needed to determine if they underwent less radiation/chemotherapy treatment.

Although not significantly different, all participants showed a decrease in weight and ECW/TBW ratio in both arms by the end of the study.

At the end of the study, 7/8 participants agreed that the device was easy to use daily, and the measurements were easy to understand. A few participants had some issues that were solved either by replacing the device or by troubleshooting, but they all agreed that they felt there was a benefit to having the home device to use.

XSL•FO
RenderX

Figure 5. Distribution of the interlimb ratio (ILB) of the ECW/TBW ratio of the affected and unaffected limb larger than '1' (orange). The average value and standard deviation of the interlimb ratio are shown along with the percentage of times the ratio is larger than 1 (below). ECW: extracellular water; TBW: total body water; ILB:interlimb; ECW ratio: ECW/TBW ratio

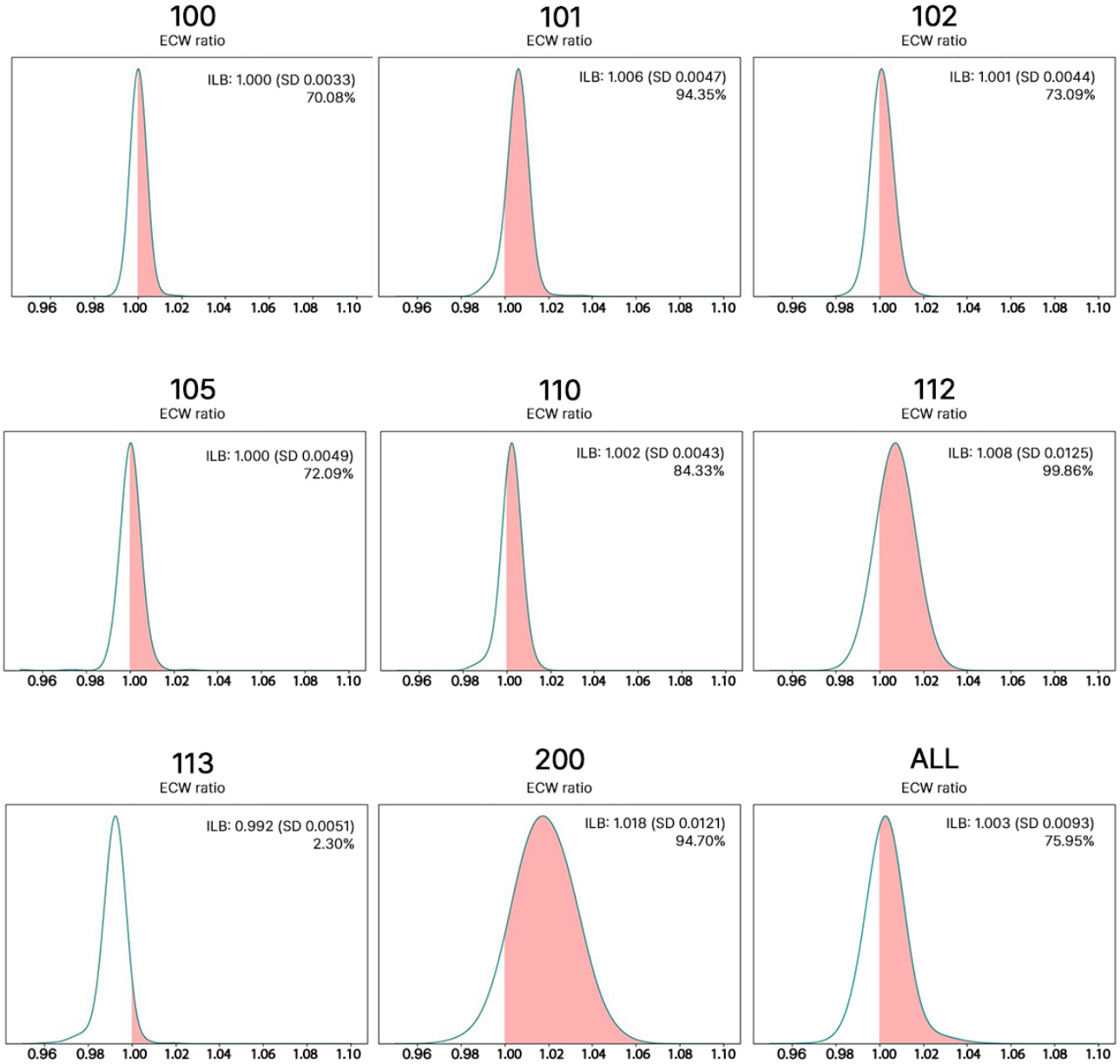

Discussion

Principal Findings

Significant differences between AM and PM fluid levels measurements were found, which indicates the need to establish a proper protocol to follow up patients, either in the morning or afternoon to reduce their burden. More data would be needed to determine if the ECW/TBW ratio difference, as well as the ratio between the affected and unaffected arms, is also affected by the time of day. This would help confirm if measurements need always to be conducted in AM and PM, or if only one measurement, one time during the day, may suffice when monitoring bilateral differences (unilateral lymphedema risk). In this study, we only considered unilateral lymphedema risk when a risk of bilateral lymphedema exists, tests may also have to be conducted at the same time of the day, always, as there would not be a contralateral reference limb (unaffected) to compare to, but a follow-up study would be needed to determine this.

Although one participant (113) whose affected arm was the dominant arm showed a different behavior compared to those whose affected arm was the nondominant arm, additional data will be needed to confirm that the BWA ON data is affected the same way as has been shown by previous literature [12], with lymphedema cutoff values being different.

Cases where treatment, such as radiotherapy or chemotherapy, resulted in significant changes in ECW/TBW or substantial differences between the two arms were reported. Possible BCRL was reported using the home BIA device for participant 200, as the ECW/TBW ratio between the affected and unaffected arms reached cutoff values (around 1.03, with small differences in each paper, which also depends on the population being used in the study - different stages) defined by previous literature [16,17,19].

XSL•FO
RenderX

Comparison to Previous Work

This study aimed to verify whether it is feasible to use a home BIA device that monitors fluid levels (BWA ON) with patients at risk of developing breast cancer-related lymphedema. Traditionally, patients visit the hospital every 3 to 6 months for lymphedema assessment. However, daily home monitoring has not been available, and patients have not been able to be tested both in the AM and PM to determine if there may be significant differences in measurements, which may affect lymphedema assessment. Home-based BIA monitoring is meaningful as it could quickly alert patients and clinicians to fluid changes.

Strengths and Limitations

This is the first study in which patients could objectively monitor fluid levels at home and relate those measurements to their subjective symptoms of discomfort or tightness. This allowed them to reach out to the therapist and discuss any changes they may have seen, being more proactive with their care. Additionally, participants expressed that they felt they benefited from using the device.

This study proved that it is feasible for patients to use the home device while allowing clinicians to monitor them without having to visit the clinic.

The first limitation of this study was the small number of participants, which did not allow for further analysis, such as comparing participants who developed lymphedema to those who did not, as only one participant appeared to have developed the disease. This is a case report to determine the feasibility of following participants at home, so no additional participants were included.

Second, the morning and evening home measurements provided an additional burden to participants, besides their necessary treatment, which caused at least one participant to withdraw from the study earlier. AM and PM measurements were necessary to determine possible differences between times of day. In practice, the goal is to reduce patient burden by encouraging testing themselves once at the same time of day.

Future Work

Increasing the number of participants monitoring themselves daily and with the professional BIA device in the clinic would help determine the correlation with other traditional methods (eg, tape measurements, range of motion, therapists' assessments) of determining lymphedema to aid therapists in making treatment decisions.

Currently, lymphedema assessment considers different measurements, including BIA in many cases. Staging lymphedema is an even more complex process that also depends on the therapist's concept and experience. In the future, a large study that considers staging may be helpful to improve patient outcomes. It may also help guide patients better when using a home monitoring device, to empower them more to take care of their health.

Conclusion

Lymphedema is a well-known disease that is difficult to improve after irreversible symptoms affect quality of life. The ECW/TBW ratio of the upper extremities, measured through BIA, could be an indicator of the development and severity of BCRL [18]. Preliminary data suggest that daily at-home BIA monitoring is feasible, which in the future may aid early detection and prevention by empowering patients as well as clinicians.

Acknowledgments

This research was supported by a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea. (grant number: HI22C106100). We appreciate the participants'' dedication to testing with the home device.

A basic initial proposal was provided by the sponsor company and approved by the Korean Government during the grant process. This was discussed with the researchers/clinicians at each study site, who created the study protocols according to best practices, standard of care, and appropriate inclusion/exclusion criteria.

Data Availability

The datasets generated or analyzed during the study are not publicly available due to the home device not having been launched in the market yet, and future work based on this dataset may be possible. Datasets are available from the corresponding author on reasonable requests.

Authors' Contributions

YH helped consolidate data from the two sites for the statistical analysis, which was discussed and verified by the other authors. MP assisted in reviewing the manuscript and adjusting it to the required journal format, including all the relevant figures. Both employees also provided technical information on the InBody devices used.

XSL•FO
RenderX

Conflicts of Interest

Authors acknowledge that Monica Padilla, PhD, is an employee of a US company (InBody BWA) associated with the parent company InBody Co., Ltd., where Yoomi Heo, MS, is an employee. The parent company manufactures the BIA device used in the study.

References

1. Paskett ED, Naughton MJ, McCoy TP, Case LD, Abbott JM. The epidemiology of arm and hand swelling in premenopausal breast cancer survivors. Cancer Epidemiol Biomarkers Prev 2007 Apr;16(4):775-782. [doi: 10.1158/1055-9965.EPI-06-0168] [Medline: 17416770]
2. Sharifi N, Ahmad S. Breast cancer-related lymphedema: a critical review on recent progress. Surg Oncol 2024 Oct;56:102124. [doi: 10.1016/j.suronc.2024.102124] [Medline: 39208532]
3. Kalra R, Anand B, Sharma H. A case report on breast cancer-related lymphedema in adulthood. Cureus 2022 May;14(5):e25042. [doi: 10.7759/cureus.25042] [Medline: 35719770]
4. Fourgeaud C, Vignes S. New insights in breast cancer-related lymphedema. J Med Vasc 2024 Sep;49(3-4):135-140. [doi: 10.1016/j.jdmv.2024.06.001] [Medline: 39278693]
5. Rockson SG. Breast cancer-related lymphedema: the primary/secondary conundrum. Lymphat Res Biol 2023 Apr;21(2):99-100. [doi: 10.1089/lrb.2023.29141.editorial] [Medline: 37093173]
6. Executive Committee of the International Society of Lymphology. The diagnosis and treatment of peripheral lymphedema: 2020 Consensus Document of the International Society of Lymphology. Lymphology 2020;53(1):3-19. [Medline: 32521126]
7. Soran A, D'Angelo G, Begovic M, et al. Breast cancer-related lymphedema--what are the significant predictors and how they affect the severity of lymphedema? Breast J 2006;12(6):536-543. [doi: 10.1111/j.1524-4741.2006.00342.x] [Medline: 17238983]
8. Donahue PMC, MacKenzie A, Filipovic A, Koelmeyer L. Advances in the prevention and treatment of breast cancer-related lymphedema. Breast Cancer Res Treat 2023 Jul;200(1):1-14. [doi: 10.1007/s10549-023-06947-7] [Medline: 37103598]
9. McLaughlin SA, Staley AC, Vicini F, et al. Considerations for clinicians in the diagnosis, prevention, and treatment of breast cancer-related lymphedema: recommendations from a multidisciplinary expert ASBrS panel : Part 1: definitions, assessments, education, and future directions. Ann Surg Oncol 2017 Oct;24(10):2818-2826. [doi: 10.1245/s10434-017-5982-4] [Medline: 28766232]
10. Zhuang L, Chen H, Zheng X, et al. Bioelectrical impedance analysis for early screening of upper limb subclinical lymphedema: A case-control study. PLoS ONE 2022;17(9):e0274570. [doi: 10.1371/journal.pone.0274570] [Medline: 36121834]
11. Yasunaga Y, Yanagisawa D, Nakajima Y, et al. Water reductive effect of lymphaticovenular anastomosis on upper-limb lymphedema: bioelectrical impedance analysis and comparison with lower-limb lymphedema. J Reconstr Microsurg 2020 Nov;36(9):660-666. [doi: 10.1055/s-0040-1713670] [Medline: 32659799]
12. Jung M, Jeon JY, Yun GJ, Yang S, Kwon S, Seo YJ. Reference values of bioelectrical impedance analysis for detecting breast cancer-related lymphedema. Medicine (Baltimore) 2018 Nov;97(44):e12945. [doi: 10.1097/MD.0000000000012945] [Medline: 30383644]
13. Lee YJ, Hong JH, Hur MH, Seo EY. Effects of virtual reality exercise program on blood glucose, body composition, and exercise immersion in patients with type 2 diabetes. Int J Environ Res Public Health 2023 Feb 26;20(5):4178. [doi: 10.3390/ijerph20054178] [Medline: 36901191]
14. DiSipio T, Rye S, Newman B, Hayes S. Incidence of unilateral arm lymphoedema after breast cancer: a systematic review and meta-analysis. Lancet Oncol 2013 May;14(6):500-515. [doi: 10.1016/S1470-2045(13)70076-7] [Medline: 23540561]
15. Zhu W, Li D, Li X, et al. Association between adjuvant docetaxel-based chemotherapy and breast cancer-related lymphedema. Anticancer Drugs 2017 Mar;28(3):350-355. [doi: 10.1097/CAD.0000000000000468] [Medline: 27997437]
16. Kim L, Jeon JY, Sung IY, Jeong SY, Do JH, Kim HJ. Prediction of treatment outcome with bioimpedance measurements in breast cancer related lymphedema patients. Ann Rehabil Med 2011 Oct;35(5):687-693. [doi: 10.5535/arm.2011.35.5.687] [Medline: 22506192]
17. Liu S, Zhao Q, Ren X, et al. Determination of bioelectrical impedance thresholds for early detection of breast cancer-related lymphedema. Int J Med Sci 2021;18(13):2990-2996. [doi: 10.7150/ijms.53812] [Medline: 34220327]
18. Kim WJ, Jo GY, Park JH, Do HK. Feasibility of segmental bioelectrical impedance analysis for mild- to moderate-degree breast cancer-related lymphedema. Medicine (Baltimore) 2021;100(4):e23722. [doi: 10.1097/MD.0000000000023722]
19. Yasunaga Y, Kinjo Y, Kondoh S, Yanagisawa D, Yuzuriha S, Kondoh S. Screening for breast cancer-related lymphedema development using extracellular water ratio. Lymphat Res Biol 2023 Oct;21(5):447-455. [doi: 10.1089/lrb.2022.0060] [Medline: 36951667]

Abbreviations

BCRL: breast cancer related lymphedema
BIA: bioelectrical impedance analysis

XSL•FO
RenderX

ECW: extracellular water
TBW: total body water
ECW/TBW: extracellular to total body water
ECW ratio: ECW/TBW ratio
ILB: interlimb

Edited by A Mavragani; submitted 15.03.25; peer-reviewed by SANDRA OLISAKWE, S Yoshida; revised version received 26.06.25; accepted 27.06.25; published 24.07.25.

Please cite as:
Teranishi-Hashimoto C, Padilla M, Heo Y, Bravi L
At-Home Bioelectrical Impedance Analysis (BIA) Monitoring of Adult Females at Risk of Breast Cancer-Related Lymphedema: Nonrandomized One-Year Longitudinal Feasibility Study
JMIR Form Res 2025;9:e73978
URL: https://formative.jmir.org/2025/1/e73978
doi:10.2196/73978

© Cheri Teranishi-Hashimoto, Monica Padilla, Yoomi Heo, Lori Bravi. Originally published in JMIR Formative Research (https://formative.jmir.org), 24.7.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Translation, Cultural Adaptation, and Mixed-Methods User Feedback of the SelfBack App to the Arabic Language for Patients With Low Back Pain: Pilot Mixed-Methods Study

Mai Aldera[1], PT, PhD; Hana Alsobayel[1], PT, PhD; Jonathan Hill[2], PT, PhD; Christian Lodberg Jensen[3], BA

[1]Department of Rehabilitation Science, College of Applied Medical Science, King Saud University, P. O. Box 800, Riyadh, Saudi Arabia

[2]School of Medicine, Keele University, Keele, United Kingdom

[3]SelfBack, Odense, Denmark

Corresponding Author:
Mai Aldera, PT, PhD
Department of Rehabilitation Science, College of Applied Medical Science, King Saud University, P. O. Box 800, Riyadh, Saudi Arabia

Abstract

Background: Low back pain (LBP) is a pervasive global health concern, significantly impacting the quality of life and burdening health care systems. Effective self-management strategies are essential for mitigating the effects of LBP and empowering individuals in their recovery. Digital health interventions, particularly smartphone apps, offer a promising avenue for delivering accessible and personalized self-management support, potentially improving patient adherence to treatment plans. The SelfBack app exemplifies such digital innovations, demonstrating clinical effectiveness in LBP management and prioritizing personalized user experiences.

Objective: This study describes the comprehensive process of translating and culturally adapting the SelfBack app from English to Arabic for the Saudi context. It further evaluates the perceived usability of the adapted Arabic SelfBack app within a cohort of Saudi individuals experiencing LBP, aiming to ensure accessibility and cultural appropriateness for Arabic-speaking users.

Methods: A rigorous, 5-stage process was implemented. Stage 1, exchange, focused on adapting the app's core content. English self-assessment scales, used for tailored treatment plans, were replaced with validated, culturally appropriate Arabic versions. Stage 2, translation and cultural adaptation, which translated the content of the patient self-management plan and adapted it for cultural relevance. Stage 3, audio conversion, addressed educational resources. English audio content was professionally translated and rerecorded in Arabic. Stage 4, laboratory usability testing, integrated the Arabic content, verifying interface functionality and right-to-left script compatibility. Stage 5, field usability testing, evaluated the app with 11 Saudi participants experiencing nonspecific LBP using the Arabic System Usability Scale (A-SUS) and semistructured interviews.

Results: The translation and adaptation processes are detailed, highlighting the work of an expert panel of linguists, health care professionals, and cultural consultants. The panel identified minimal discrepancies and no significant misunderstandings, demonstrating the accuracy and cultural appropriateness of the adaptation. The mean System Usability Scale (SUS) score was 70%, indicating good usability. Interviews corroborated these results, with participants generally reporting the app as clear, intuitive, and easy to use. However, feedback highlighted areas for improvement, including the perceived number of mandatory questions, a perceived lack of interactivity, repetitive content, and unmet expectations regarding certain functionalities.

Conclusions: The Arabic SelfBack app has been successfully developed, demonstrating high user satisfaction, ease of use, and interface efficiency within the target cultural context. The identified criticisms provide actionable insights for future updates. These results suggest that the app is ready for clinical research with Arabic-speaking participants, potentially improving LBP management in Saudi Arabia. Further research will be essential to confirm these initial findings and establish the app's long-term effectiveness.

(JMIR Form Res 2025;9:e59777) doi:10.2196/59777

KEYWORDS

low back pain; self-management; treatment; digital intervention; quality of life; language; smartphone; mobile app; Arabic-speaking; Arabic translation; cultural adaptation; cohort; mixed methods; user feedback; mobile health; usability; interview; Saudi Arabia

XSL•FO
RenderX

Introduction

Low back pain (LBP) is a significant global health concern, adversely impacting a significant portion of the population throughout their lives [1]. It ranks among the leading causes of disability in adults and generates a substantial public health burden due to its impact on affected individuals' well-being and quality of life [2]. The prevalence of LBP is particularly concerning, with estimates suggesting that up to 80% of adults will experience LBP at some point in their lives [1]. This underscores the need for effective and accessible interventions.

Clinical guidelines advocate that patients with LBP be provided with supported self-management strategies to empower them to manage their pain. These strategies include exercise, education, pain relief, lifestyle advice around sleep, diet, and stress, as well as advice regarding work and maintaining normal activities [3]. However, people with LBP often find it hard to implement these strategies due to lack of knowledge, fears about doing the wrong thing, lack of confidence, competing routines, feeling overwhelmed by information, limited access to physical activity resources, lack of trust in what to do, and lack of feedback about the benefits of adhering to their treatment regimens [4]. These barriers highlight the gap between recommended care and actual practice.

Digital interventions, including smartphone apps, have emerged as promising solutions to overcome these challenges [5]. They offer evidence-based and cost-effective methods for delivering self-management solutions for people with LBP [4]. Digital health interventions can increase access to care, particularly in remote areas or for individuals with mobility limitations. They also offer the potential for personalized interventions tailored to individual needs and preferences [6].

While a systematic review found that numerous LBP apps are available, their overall quality was often poor, characterized by a lack of engaging features, unattractive interfaces, and unreliable information [7]. This highlights the importance of rigorous evaluation and quality assurance in the development and deployment of LBP apps. The recent National Institute for Health and Care Excellence (NICE) Early Value Assessment of digital self-management apps for LBP [8] identified 10 apps, 5 of which were recommended for use in the National Health Service (NHS), including SelfBack, and 5 that can only be used in research. The NICE evaluation underscores the growing potential of digital tools in chronic pain management but emphasizes the need for rigorous assessment before mainstream integration. This evaluation process provides a valuable framework for assessing the clinical and cost-effectiveness of digital health interventions.

Various smartphone apps, including Spine Zone [9], Pain Coach [10], and getUBetter [11], address LBP self-management. The SelfBack app stands out for its unique evidence-based approach and dedication to personalized user experiences. Its efficacy has been shown in a large randomized controlled trial [4], where participants using SelfBack demonstrated significant reductions in pain and disability compared to a control group after 3 months. This robust evidence base distinguishes SelfBack from many other LBP apps. Furthermore, its personalized approach, based on individual assessments and feedback, may contribute to improved engagement and adherence.

At present, there are no digital tools available in the Arabic language for supporting self-management for people with LBP. Other self-management apps for different health conditions, such as the Sehhaty app (the Saudi governmental health platform) [12] and the Sehhaty Wa Daghty app, have demonstrated effectiveness in the Saudi cultural context [13]. These successful examples of digital health interventions in Saudi Arabia highlight the potential for culturally adapted LBP apps. Therefore, the development of an app supporting self-management for Arabic-speaking people with LBP is important. However, translating software app content involves much more than just basic word-for-word translation. It requires a nuanced approach that considers cultural context, user experience, and technical limitations. The objectives of this study were, therefore, to translate and culturally adapt the English language version of the SelfBack app to the Arabic language and evaluate user satisfaction.

Methods

Design

A prospective study was conducted to assess user satisfaction with the SelfBack app after its translation and cultural adaptation from English to Arabic.

The SelfBack App

The SelfBack app is a smartphone app that uses artificial intelligence to create personalized self-management plans for people with nonspecific LBP based on individual person data within a case-based reasoning system, as shown in Figure 1 [4]. It is designed for self-management and uses an evidence-based decision support system. The app works by first asking the person to fill out a range of scales and questionnaires about their LBP condition. For example, it asks about their pain severity, functional ability, fear avoidance behaviors, workability, perceived barriers to self-management, pain self-efficacy, sleep, perceived stress, and mood. This information is then used to create a personalized treatment plan, which consists of a combination of physical exercises and educational materials.

The treatment plan is updated weekly based on physical activity monitoring and self-reported adherence to the self-management plan. The app uses the person's smartphone to track their physical activity levels or can link to a wearable step counter if the person has one. This information is used to ensure that the treatment plan is challenging, but not too difficult, and to identify any areas where the person may need additional support.

To use the SelfBack app, a person first needs to download the app from the App Store or Google Play. For the purpose of this research, patients were initially screened by their treating clinician as appropriate, before being signposted to the app and provided with a token to enable the app to be downloaded for free. Once the app was installed on their phone, they created an account and filled out a range of questions about their condition to initiate their individual treatment plan. The treatment plan remains dynamic and depends on the person's progress, which is adjusted weekly as needed.

XSL•FO
RenderX

Figure 1. SelfBack app workflow. (adapted from SelfBack website [14].

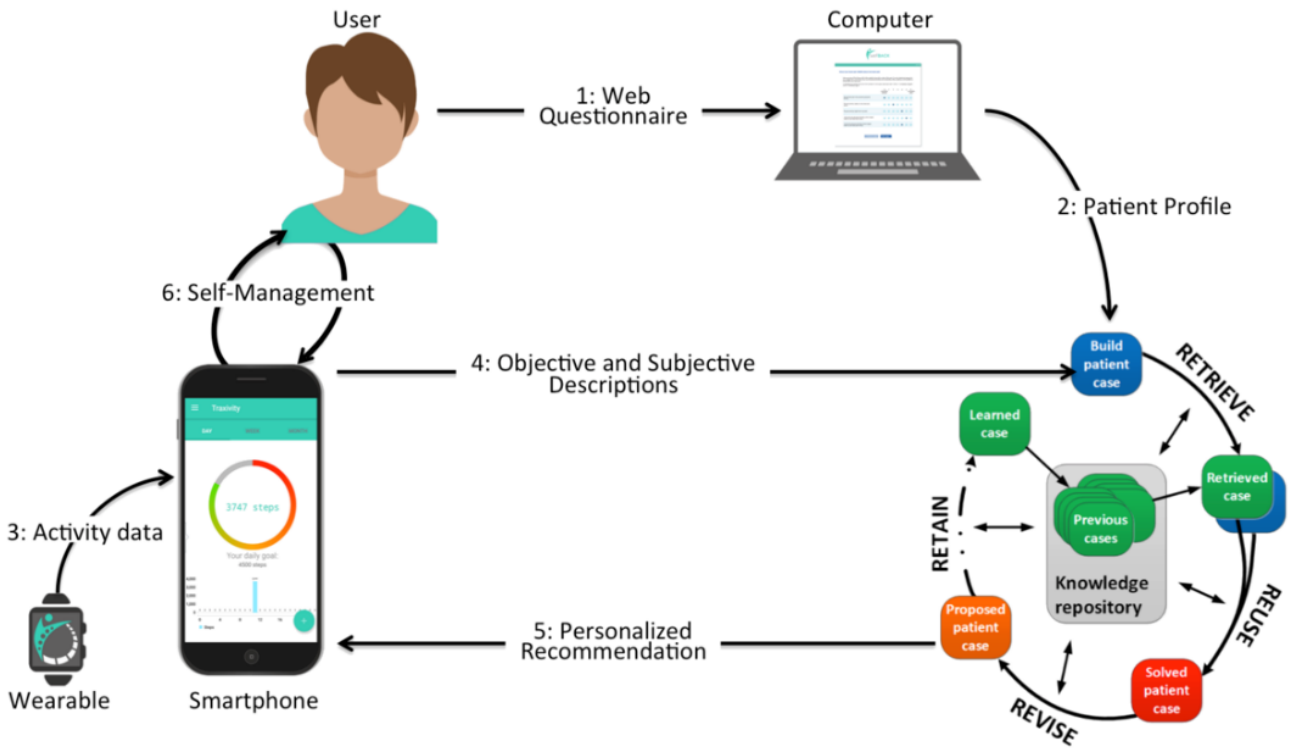

Translation and Cultural Adaptation

The translation and cultural adaptation of the SelfBack app contents from the English language to the Arabic language required a multistep design approach to ensure it was of high quality [15]. The method comprised 5 phases:

Exchange Phase

This phase involved exchanging the SelfBack self-assessment scales and questionnaires used at the baseline and weekly question and answer sessions from the English language version to translated and validated Arabic versions. The investigators (HS and MA) identified ready-to-use Arabic-translated scales and, if necessary, obtained permission for their use.

Translation and Cultural Adaptation Phase

This phase focused on the text content of the SelfBack patient self-management plan, translating and culturally adapting it into the Arabic language. This was based on the principles described by Wild et al [16] using a multistep approach: preparation, forward translation to the Arabic language, cultural adaptation, quality assurance, comparison of the 2 versions by an expert panel, proofreading, and adding the final Arabic version to the SelfBack app. To ensure the quality of translation, various steps involving translation were undertaken by native Arabic and English speakers at an authorized language translation service center. The cultural adaptation step focused on replacing common English sayings with Arabic sayings (eg, the statement “trouble sleeping” was replaced by “difficulty sleeping”); This was done by expert bilingual physiotherapists (HA and MA). The quality assurance step was carried out by the experts (HS and MA). Comparison of the 2 versions (English and Arabic) was undertaken by an expert panel comprising 2 investigators (HS and MA) and a consultant psychologist who compared the original English version with the translated Arabic version to assess linguistic equivalency and correct spelling, and that the 2 versions were conceptually equivalent. Final proofreading was conducted by an investigator (HS). Adding the final Arabic version to the SelfBack app was undertaken by the software developers of the SelfBack app.

Audio Conversion Phase

This phase involved converting the English language audio educational materials to the Arabic language, using the principles described by Wild et al [16], and requiring multiple steps as follows: preparation, forward translation to the Arabic language, cultural adaptation, quality assurance, comparison of the 2 versions by an expert panel, audio recording, proof listening, and adding the final Arabic version to the SelfBack app. These steps were, mostly, undertaken as described for the translation and cultural adaptation phase. However, in this phase, the expert panel selected 1 person out of 3 candidates to record the audio recordings based on their ability to speak clearly, understandably, and in a professional tone.

Laboratory Usability Test Phase

This phase addressed the software interface text of all features by allocating the Arabic materials of the SelfBack app to the appropriate place in the app. This phase involved carefully considering the differences between the English and Arabic languages, in particular in terms of the direction of text writing being right to left in Arabic and that it is not possible to divide Arabic words when starting new lines. Thus, in this phase, the main focus of the investigators (HS and MD) was to address the software interface text of all features by allocating the Arabic materials to the appropriate place in the SelfBack app.

XSL•FO
RenderX

Clinical Usability Testing Phase

This phase involved assessing the usability of the Arabic SelfBack app for both the Android and iOS operating systems using the validated and translated Arabic version of the System Usability Scale (A-SUS) [17]. The SUS is a questionnaire used to evaluate the usability of a wide variety of new software systems [18]. It comprises 10 questions which are answered using a 5-point Likert scale ranging from "I strongly disagree" to "I strongly agree." The SUS generates a score (range 0-100) and offers information about overall customer satisfaction and usability [19]. In addition, to assess user experience with the app, one-to-one phone interviews were undertaken using one open-ended question, "What feedback do you have after using the Arabic version of the SelfBack software app?" List 3 important features."

This study was conducted as part of a pilot randomized controlled trial aimed at translating, culturally adapting, and evaluating the usability of the SelfBack app for Arabic-speaking users in Saudi Arabia (IRB E-22 - 7106). Participants for this phase of the study were randomly recruited from the Primary Care Clinic at King Khalid University Hospital in Riyadh. The participants (N=11) represent a random sample of 50% of the control group who used the app and adhered to it for more than 3 months. Inclusion criteria were adult patients (≥18 y) diagnosed with LBP who had used the SelfBack app for more than 3 months. Exclusion criteria encompassed previous back surgery, specific spine pathologies (eg, spinal stenosis and herniated disc), pregnancy-related back pain, and any other clinical contraindication to self-management via the SelfBack app as determined by the treating clinician.

Data Analysis

The SUS questionnaire data (10 items) were analyzed using IBM SPSS Statistics for Windows (version 25.0; IBM Corp). The SUS yields a global score reflecting perceived usability, with higher scores indicating better usability and lower scores suggesting potential usability problems. A benchmark of 70% is often used to differentiate acceptable usability from areas needing improvement, while scores below 50% are typically indicative of low usability and necessitate further investigation [19].

The SUS comprises 10 items, 5 of which are positively worded (Q1, Q3, Q5, Q7, and Q9) and 5 negatively worded (Q2, Q4, Q6, Q8, and Q10). Scoring involves a 2-step process. First, responses to each item are converted to numerical values: strongly disagree=1, disagree=2, neutral=3, agree=4, and strongly agree=5. Second, intermediate scores are calculated for the odd- and even-numbered items. For the odd-numbered items, the sum of the item scores is subtracted from 5 (x=Σ(Q1, Q3, Q5, Q7, Q9) - 5). For the even-numbered items, the sum of the item scores is subtracted from 25 (y=25 - Σ(Q2, Q4, Q6, Q8, Q10)). The final SUS score is then calculated using the following formula: SUS score = (x+y) * 2.5. This calculation method reflects the 100-point scale of the SUS, where each item contributes a maximum of 10 points [19].

Data from the open question regarding user experience were analyzed using qualitative content analysis by counting the frequency of the created codes in the transcribed text from the phone interviews. A thematic analysis of interview data was conducted manually, and the participants' answers were translated from Arabic to English by 2 experts (HS and MD).

Ethical Considerations

This study involving the translation and cultural adaptation of the SelfBack app, along with usability testing, was reviewed and approved by the research committees of the College of Medicine, King Saud University with the approval number E-22 - 7106, and the procedures followed were in accordance with the ethical standards of the responsible committee in human experimentation and with the WMA Declaration of Helsinki. While the usability testing involved human participants, the study focused on evaluating the app's interface and functionality, not on collecting sensitive personal data.

All participants in the usability testing provided informed consent after receiving a detailed explanation of the study's purpose, procedures, and rights. The Arabic language consent form explained that the data collected would be used solely for evaluating the app's usability and would not be used for any other purpose. Participants were also informed of their right to withdraw from the study at any time without penalty.

To ensure privacy and confidentiality, all data collected during the usability testing were anonymized. Participant names and other identifying information were removed from all data files. The interview transcripts were coded, and any potentially identifying information was redacted. The anonymized data will be stored securely and will only be accessible to the research team.

Participants in the usability testing did not receive any compensation for their time and contribution to the study. No images of individual participants or users are included in this paper. Therefore, no consent for image use was required.

Results

Translation and Cultural Adaptation Phase

The expert panel participants during the translation and cultural adaptation phases did not report any misunderstandings and only a minor few discrepancies with the translated and culturally adapted instructions and educational materials.

Clinical Usability Testing

Eleven people with nonspecific LBP participated in the field usability testing. As shown in Table 1, there were 10 (91%) females with ages ranging from 31 - to 58 years.

XSL•FO
RenderX

Table . Demographic data of the participants (N=11) who completed 3 months using the SelfBack app, included in the field usability test.

Participant	Sex	Age (y)	MSK-HQ[a]	StartBack tool[b]	Physical activity level (d/wk)
1	Female	37	29	High	4
2	Female	40	46	Low	2
3	Female	36	41	Low	0
4	Female	58	43	Medium	2
5	Female	38	35	High	0
6	Female	52	32	Medium	5
7	Female	38	30	High	2
8	Female	38	38	Medium	0
9	Female	31	44	Low	5
10	Female	38	35	Medium	2
11	Male	31	30	Low	2

[a]MSK-HQ: Musculoskeletal Health Questionnaire.

[b]Keel StartBack Tool (the grade level indicates high, medium, and low levels of risk of chronicity and disability).

Arabic System Usability Scale

The mean (SD) A-SUS score was 70% (SD 5), reflecting a “good” level of usability [18]. This has been driven by 2 steps: first, convert the scale into numerical values for each of the 10 questions, as follows: strongly disagree: 1 point, disagree: 2 points, neutral: 3 points, agree: 4 points, and strongly agree: 5 points. Then, calculate the scores as the following formula: SUS score=(x+y) * 2.5, where (x) is the sum of the item scores and is subtracted from 5 for all odd-numbered questions, and (y) is the sum of the item scores and is subtracted from 25 for even-numbered questions. In addition, the answers to the open-ended question were consistent with the SUS outcomes.

Qualitative Analysis

The thematic analysis data were organized hierarchically using a code frame to develop sentiment analysis by identifying patterns and themes. Based on the feedback received, 2 main themes emerged: positive and negative. The positive feedback theme comprised 3 subthemes: ease of use, clear language, and good audiovisual features. The negative feedback theme consisted of 4 subthemes: answering mandatory questions, lack of interaction, repetitive content, and below-average app features.

Positive and negative feedback are listed in Textbox 1.

XSL•FO
RenderX

Textbox 1. Positive and negative feedback.

Positive Feedback (A)

- Theme (A1) easy to use (number of mentions=7): feedback highlighted the user-friendly nature of the app, with participants (n=7) expressing appreciation for its simplicity and ease of use.
  - "The app was clear and easy to use." [P1]
  - "It is my first time using such these kinds of applications. Once I understood the idea, then the uses were easy for me." [P10]
- Theme (A2) clear language: (number of mentions=2): this theme reflected the clarity of the Arabic language and its integration into the app. [P3]
  - "Easy to use and the language is understandable and direct."
- Theme (A3) good audiovisual features: number of mentions (n=2): positive feedback was received for the SelfBack app's useful audio and visual features (n=2).
  - "The exercise pictures, sound, and instructions are clear." [P7]

Negative Feedback (B)

- Theme (B1) answering the mandatory questions (number of mentions, n=4): participants (n=4) expressed their opinions on the mandatory and complicated process of having to ask consistent questions every time they use the app. They provided different feedback, including that it consumed too much of their time and led to frustration or quitting using the app altogether.
  - "Everything in the app is easy, but every time I log in, I have to answer the same question over and over." [P3]
- Theme (B2) lack of interaction (number of mentions n=3): this theme pertained to the uncertainty around whether they are using the app correctly or whether they should stop using it. In addition, this feedback suggested that there was a lack of real interaction with health care practitioners (n=3).
  - "No interaction." [P6]
  - "It is my first time using such these kinds of applications. Once I understood the idea, then the uses were easy for me. But I want someone to follow up with me." [P10].
- Theme (B4) repetitive content (number of mentions, n=2): this theme concerned negative feedback regarding repeating instructions and exercises over time, regardless of LBP severity (n=2).
  - "It was beneficial, my comment is for how long should I have to use it? Because the app keeps repeating the same exercise." [P9]
- Theme (B5) below expectations (number of mentions, n=1): this subtheme pertained to a user's expectations of the app in comparison to other apps (n=1).
  - "The app was below my expectations and there is nothing special about it. There is another app with superior features that gives you more exercise and interaction." [P11]

Discussion

Primary Findings

This study successfully translated and culturally adapted the English SelfBack app for the Saudi Arabian context, subsequently evaluating user satisfaction through a mixed-methods approach. The translation and adaptation process was efficiently executed, with an expert panel review identifying minimal discrepancies and confirming the absence of significant misunderstandings attributable to language. Usability testing with 11 participants yielded a positive overall usability rating. Qualitative data derived from semistructured interviews revealed a range of user perceptions. Positive feedback themes centered on the app's intuitive interface, ease of navigation, clarity of the Arabic language implementation, and effective integration of audiovisual components. Conversely, negative feedback addressed concerns regarding the length of mandatory questionnaires, limited opportunities for user interaction, perceived redundancy in content delivery, and unmet expectations concerning the breadth and functionality of available features.

The usability and acceptability findings of this study resonate with prior SelfBack app assessments conducted in Aberdeen, Scotland and Trondheim, Norway [20], which also reported user feedback concerning desired improvements in app functionality and mixed reactions to motivational notifications. While the majority of Scottish participants appreciated the notifications they received, user reception in Norway was less favorable, with only half finding them useful. This variability in user response across different cultural contexts underscores the need for further refinement of the notification system to ensure its resonance with a broader range of users. This suggests that cultural adaptation of digital health interventions extends

XSL•FO
RenderX

beyond language translation and requires careful consideration of user preferences and contextual factors related to how and when information is delivered. Future iterations of the notification system should consider incorporating user customization options, allowing individuals to select notification frequency, content, and delivery methods.

The requirement for mandatory sign-up questions during initial app interaction, while designed to collect data for personalized treatment plan updates, elicited negative feedback from some participants in this study, potentially contributing to user burden. Given that initial onboarding experiences can significantly influence app adherence [21], this step warrants streamlining and optimization to prioritize user comfort and engagement. This finding aligns with the broader literature on user experience in digital health, which emphasizes the importance of minimizing barriers to entry and maximizing user-friendliness, particularly during initial interactions. Future versions of the Arabic SelfBack app could explore simplifying the sign-up process, perhaps by reducing the number of mandatory fields or by clearly communicating the rationale behind the information requested to manage user expectations and prevent negative first impressions. Alternatively, a staged approach to data collection could be implemented, gathering essential information initially and allowing users to provide additional details later.

This study's findings regarding the desire for personal interaction with health care professionals are consistent with a qualitative study by Alzahrani et al (2022) [21], which explored Saudi patients' perspectives on mobile health apps. Their research revealed a preference for personal interaction with health care providers to enhance adherence and suggested incorporating social media groups for peer support within apps. This convergence of findings across studies highlights a potentially crucial cultural factor in the adoption and effective use of digital health interventions in Saudi Arabia. The desire for human connection and support, even within a digital context, may reflect cultural values emphasizing interpersonal relationships and community. These insights can inform future app development and design, suggesting that incorporating features that facilitate interaction with health care practitioners, such as secure messaging or telehealth integration, and fostering peer support communities within the app could significantly enhance its acceptability and effectiveness. Furthermore, addressing concerns regarding repetitive content and expanding the variety of exercises offered could further improve user engagement and satisfaction, aligning the app's functionality more closely with user expectations and potentially leading to better outcomes.

While this study used a rigorous and evidence-based methodology for the translation and cultural adaptation of the SelfBack app, including a comprehensive expert review process, it is important to acknowledge certain limitations. The primary limitation lies in the relatively small sample size (N=11) used in the field usability testing. This limited sample size may restrict the generalizability of the findings and may not fully capture the diversity of experiences and perspectives within the target population. While the qualitative data gathered through interviews provided valuable insights into user perceptions, a larger and more representative sample would have strengthened the study's conclusions regarding usability and acceptability. Furthermore, this study focused primarily on usability and user satisfaction; it did not assess the clinical effectiveness of the Arabic SelfBack app in managing LBP. Therefore, future research using a larger sample size and incorporating a robust evaluation of clinical outcomes, such as pain intensity, functional capacity, and quality of life, is warranted to determine the app's efficacy in the Saudi Arabian context. Such research would provide a more comprehensive understanding of the app's potential impact on LBP management and inform its wider implementation.

Conclusions

This study successfully translated, culturally adapted, and evaluated the usability of the SelfBack app for Arabic-speaking users in Saudi Arabia, addressing a critical need for culturally sensitive digital health resources. Positive usability findings suggest the app's potential for effective adoption and use. These results contribute to the growing evidence supporting mobile health for chronic pain management, particularly in underserved populations, and highlight the importance of cultural adaptation in digital health intervention development. The Arabic SelfBack app has the potential to improve access to evidence-based LBP self-management in Saudi Arabia. Future research should evaluate its clinical effectiveness to inform wider implementation and contribute to improved LBP care. This study also serves as a model for developing and adapting digital health interventions for other conditions and cultural contexts.

Acknowledgments

The authors acknowledge support and funding from the British Council.

The authors would like to thank Ongoing Research Funding Program, (ORFFT-2025-010-1), King Saud University, Riyadh, Saudi Arabia for financial support.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

XSL•FO
RenderX

References

1. Freburger JK, Holmes GM, Agans RP, et al. The rising prevalence of chronic low back pain. Arch Intern Med 2009 Feb 9;169(3):251-258. [doi: 10.1001/archinternmed.2008.543] [Medline: 19204216]
2. Wu A, March L, Zheng X, et al. Global low back pain prevalence and years lived with disability from 1990 to 2017: estimates from the Global Burden of Disease Study 2017. Ann Transl Med 2020 Mar;8(6):299. [doi: 10.21037/atm.2020.02.175] [Medline: 32355743]
3. Bernstein IA, Malik Q, Carville S, Ward S. Low back pain and sciatica: summary of NICE guidance. BMJ 2017;356:i6748. [doi: 10.1136/bmj.i6748]
4. Sandal LF, Bach K, Øverås CK, et al. Effectiveness of app-delivered, tailored self-management support for adults with lower back pain-related disability: a selfBACK randomized clinical trial. JAMA Intern Med 2021 Oct 1;181(10):1288-1296. [doi: 10.1001/jamainternmed.2021.4097] [Medline: 34338710]
5. Michie S, Yardley L, West R, Patrick K, Greaves F. Developing and evaluating digital interventions to promote behavior change in health and health care: recommendations resulting from an international workshop. J Med Internet Res 2017 Jun 29;19(6):e232. [doi: 10.2196/jmir.7126] [Medline: 28663162]
6. Olawade DB, Bolarinwa OA, Adebisi YA, Shongwe S. The role of artificial intelligence in enhancing healthcare for people with disabilities. Soc Sci Med 2025 Jan;364:117560. [doi: 10.1016/j.socscimed.2024.117560]
7. Machado GC, Pinheiro MB, Lee H, et al. Smartphone apps for the self-management of low back pain: a systematic review. Best Pract Res Clin Rheumatol 2016 Dec;30(6):1098-1109. [doi: 10.1016/j.berh.2017.04.002] [Medline: 29103552]
8. Digital technologies for managing non-specific low back pain: early value assessment. National Institute for Health and Care Excellence. 2024. URL: https://www.nice.org.uk/guidance/indevelopment/gid-hte10021 [accessed 2025-06-23]
9. SpineZone is your trusted physical therapy partner for spine and joint pain. SpineZone. 2024. URL: https://www.spinezone.com [accessed 2025-06-23]
10. Pre and post-operative pain management that's there every step of the way. Pain Coach. 2024. URL: https://www.paincoachapp.com [accessed 2025-06-23]
11. Support your local population to recover, live and work well. getUBetter Platform. 2024. URL: https://www.getubetter.com [accessed 2025-06-23]
12. Sehhaty platform. Ministry of Health. 2023. URL: https://www.moh.gov.sa/en/eServices/Sehhaty/Pages/default.aspx [accessed 2025-06-23]
13. Alzahrani A, Gay V, Alturki R. Enabled artificial intelligence (AI) to develop Sehhaty Wa Daghty app of self-management for Saudi patients with hypertension: a qualitative study. Information 2023;14(6):334. [doi: 10.3390/info14060334]
14. Mork PJ, Bach K, selfBACK Consortium. A decision support system to enhance self-management of low back pain: protocol for the selfBACK project. JMIR Res Protoc 2018 Jul 20;7(7):e167. [doi: 10.2196/resprot.9379] [Medline: 30030208]
15. Grasaas E, Fegran L, Helseth S, et al. iCanCope with pain: cultural adaptation and usability testing of a self-management app for adolescents with persistent pain in Norway. JMIR Res Protoc 2019 Jun 3;8(6):e12940. [doi: 10.2196/12940] [Medline: 31162132]
16. Wild D, Grove A, Martin M, et al. Principles of good practice for the translation and cultural adaptation process for patient-reported outcomes (PRO) measures: report of the ISPOR Task Force for Translation and Cultural Adaptation. Value Health 2005;8(2):94-104. [doi: 10.1111/j.1524-4733.2005.04054.x] [Medline: 15804318]
17. AlGhannam BA, Albustan SA, Al-Hassan AA, Albustan LA. Towards a Standard Arabic System Usability Scale: psychometric evaluation using Communication Disorder App. Int J Hum-Comput Int 2018 Sep 2;34(9):799-804. [doi: 10.1080/10447318.2017.1388099]
18. Lewis JR. The System Usability Scale: past, present, and future. Int J Hum-Comput Int 2018 Jul 3;34(7):577-590. [doi: 10.1080/10447318.2018.1455307]
19. Jordan PW, Thomas B, McClelland IL, Weerdmeester B. Usability Evaluation In Industry: CRC Press; 1996. URL: http://books.google.ie/books?id=ujFRDwAAQBAJ&printsec=frontcover&dq=Brooke+J.+Usability+Evaluation+in+Industry&hl=&cd=1&source=gbs_api [accessed 2025-07-08]
20. Nordstoga AL, Bach K, Sani S, et al. Usability and acceptability of an app (SELFBACK) to support self-management of low back pain: mixed methods study. JMIR Rehabil Assist Technol 2020 Sep 9;7(2):e18729. [doi: 10.2196/18729] [Medline: 32902393]
21. Alzahrani A, Gay V, Alturki R. Exploring Saudi individuals' perspectives and needs to design a hypertension management mobile technology solution: qualitative study. Int J Environ Res Public Health 2022 Oct 10;19(19):12956. [doi: 10.3390/ijerph191912956] [Medline: 36232254]

Abbreviations

A-SUS: Arabic System Usability Scale
LBP: low back pain
NICE: National Institute for Health and Care Excellence
SUS: System Usability Scale

XSL•FO
RenderX

Edited by A Mavragani; submitted 22.04.24; peer-reviewed by P Nohl-Deryk; revised version received 11.02.25; accepted 02.03.25; published 25.07.25.

Please cite as:
Aldera M, Alsobayel H, Hill J, Jensen CL
Translation, Cultural Adaptation, and Mixed-Methods User Feedback of the SelfBack App to the Arabic Language for Patients With Low Back Pain: Pilot Mixed-Methods Study
JMIR Form Res 2025;9:e59777
URL: https://formative.jmir.org/2025/1/e59777
doi:10.2196/59777

© Mai Aldera, Hana Alsobayel, Jonathan Hill, Christian Lodberg Jensen. Originally published in JMIR Formative Research (https://formative.jmir.org), 25.7.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Generative AI–Powered Mental Wellness Chatbot for College Student Mental Wellness: Open Trial

Jazmin A Reyes-Portillo[1*], PhD; Amy So[1*], BS; Kelsey McAlister[2*], MS, PhD; Christine Nicodemus[3], BS; Ashleigh Golden[3*], PsyD; Colleen Jacobson[4*], PhD; Jennifer Huberty[2*], PhD

[1]Montclair State University, 1 Normal Ave, Montclair, NJ, United States
[2]Fit Minded, Phoenix, AZ, United States
[3]Wayhaven, Charlotte, NC, United States
[4]Continuum Psychiatric Services, LLP, New York, NY, United States
[*]these authors contributed equally

Corresponding Author:
Jazmin A Reyes-Portillo, PhD
Montclair State University, 1 Normal Ave, Montclair, NJ, United States

Abstract

Background: Colleges have turned to digital mental health interventions to meet the increasing mental health treatment needs of their students. Among these, chatbots stand out as artificial intelligence–driven tools capable of engaging in human-like conversations that have demonstrated some effectiveness in reducing depression and anxiety symptoms.

Objective: This study aimed to assess the feasibility and acceptability of using Wayhaven, an artificial intelligence chatbot, among college students with elevated depression or anxiety symptoms. We also aimed to examine the preliminary effectiveness of Wayhaven in improving symptoms of anxiety and depression, hopelessness, agency, and self-efficacy among students.

Methods: Participants were 50 racially and ethnically diverse college students with elevated depression or anxiety symptoms (n=45, 80% female; mean age 22.12, SD 4.42 years). Students were asked to use Wayhaven over the course of 1 week and completed assessments at preintervention, after 1 session, and 1 week.

Results: Wayhaven use was associated with a significant decrease in depression (β=−1.62; P<.001), anxiety (β=−2.15; P<.001), and hopelessness (β=−.64; P<.001) and a significant increase in agency (β=.64; P=.32), self-efficacy (β=.53; P=.02), and well-being (t_{40}=2.90; P=.006; d=0.45) across the study period. Most students also reported being satisfied with Wayhaven and it being a tool they would recommend to their peers.

Conclusions: Findings suggest that Wayhaven may be a viable mental wellness resource for diverse students with elevated depression or anxiety symptoms.

(JMIR Form Res 2025;9:e71923) doi:10.2196/71923

KEYWORDS

artificial intelligence; chatbots; anxiety; depression; college students

Introduction

Rates of depression, anxiety, and suicidal ideation have surged among college students since the COVID-19 pandemic [1,2]. From 2013 to 2021, the past-year prevalence rate of depression, anxiety, and suicidal ideation has increased by 134.6%, 109.5%, and 64%, respectively, among college students [3]. More than ever before, college counseling centers are unable to meet the increasing mental health needs of their students [4,5]. Students of color, in particular, have higher unmet treatment needs than their White peers [3]. At the same time, many who would benefit from services never seek treatment. Among college students, lack of perceived need for treatment, limited time, preference for self-help, and fear of stigma impede help-seeking [6,7].

To help address these issues, colleges have turned to digital mental health interventions (DMHIs). DMHIs refer to a range of health information technologies (eg, websites, mobile apps, and chatbots) designed to intervene in health conditions by changing behaviors, cognitions, and emotional states [8]. Among these, chatbots stand out as artificial intelligence (AI)–driven tools capable of engaging in human-like conversations. Chatbots can help monitor a patient's progress, track symptoms and behaviors, facilitate the transfer of therapeutic content into their daily lives, and provide personalized support by delivering additional mental health resources [9]. Beyond offering convenient access to resources, AI-powered DMHIs are well-liked by college-aged individuals [10]. Features like active listening for rapport building and problem exploration through AI have been found to be not only acceptable but, in some cases,

XSL•FO
RenderX

even preferred over human-to-human interactions [11]. This user-centered focus of DMHIs appears to be a key factor in their appeal to college-aged individuals [12].

Chatbots, such as Woebot and Tess, have been used to deliver cognitive behavioral therapy (CBT)–based self-help content in a text-based conversational format to college students experiencing depression and anxiety, demonstrating acceptability and some effectiveness in reducing depression and anxiety symptoms compared to psychoeducational control conditions [13,14]. While these tools use user-centered approaches, such as mood monitoring, psychoeducation, and validation, their interactions rely heavily on preprogrammed scripts and structured conversation pathways. This limits their ability to adapt dynamically to the nuanced, real-time needs of users or incorporate personalized context, such as past conversations or external user-specific data (eg, demographics).

To address these gaps, Wayhaven, a generative AI-powered mental wellness chatbot, delivers dynamic, context-aware mental well-being support tailored to each user's unique needs. Its AI mental wellness coaches deliver personalized, context-aware support by incorporating user demographics, university-specific resources, and a tailored AI coach personality. Unlike scripted chatbots, Wayhaven dynamically adapts to users' real-time needs while ensuring safety through advanced crisis detection and rerouting. This innovative approach positions Wayhaven as a unique tool for college students with elevated anxiety and depression symptoms, offering dynamic, evidence-based mental wellness support tailored to diverse needs. However, there is a need to evaluate its feasibility, acceptability, and preliminary effectiveness in a real-world setting.

The purpose of this study was to explore how college students interact with Wayhaven's AI-powered mental wellness chatbot and its feasibility for improving mental wellness among college students with elevated depression or anxiety symptoms. This study aimed to: (1) determine the acceptability of Wayhaven; (2) determine the preliminary effectiveness of Wayhaven in improving mental health symptoms among students, including anxiety, depression, hopelessness, agency, and self-efficacy; and (3) explore student experiences and perceptions of Wayhaven. In this study, we use engagement (ie, actual use of the app) and participant satisfaction (ie, postuse evaluation) as proxies for acceptability.

Methods

Recruitment

Participants were recruited from a public university in New Jersey. Students learned about the study through an email newsletter that lists on-campus research opportunities that were sent to all students. The email provided information about the study and a link to a brief web-based survey to assess eligibility. After providing informed consent, participants were directed to take the screening survey to determine eligibility for the study. To be included, participants need to: (1) be aged 18 years or older; (2) be a student enrolled in the participating university; (3) live in the United States; (4) read and understand English; and (5) have elevated depression (Patient Health Questionnaire-2 [PHQ-2] [15] score ≥3) or anxiety symptoms (Generalized Anxiety Disorder 2 [GAD-2; [16] score ≥3).

Procedures

Eligible participants were asked to complete a preintervention survey and to spend at least 5 minutes during their first Wayhaven conversation. At the end of the first session, they were directed to complete a postintervention survey. Following this initial conversation, participants continued to have access to Wayhaven for 1 week and were encouraged to engage with the AI mental wellness chatbot as much as needed. Participants received a final follow-up survey 1 week after enrollment. Each survey took 30 - 45 minutes to complete. Participants were recruited from September 20 to September 25, 2024. Recruitment was closed once 40 students completed the study (ie, all assessments and at least one conversation). After completing the study, students were granted unlimited access to Wayhaven.

Ethical Considerations

The Montclair State University Institutional Review Board approved this study (IRB-FY24-25-3875). All participants provided informed consent prior to particpating and were informed that they could to opt out of the study at any time. Participants received a US $20 gift card for completing each survey. All participant data was de-identified.

Wayhaven AI Mental Wellness Coach

Participants were given access to Wayhaven upon study enrollment. Wayhaven is a generative AI-powered mental wellness chatbot, designed to deliver brief, evidence-based text conversations that support college students' mental wellness needs. These conversations are rooted in CBT, dialectical behavior therapy, acceptance and commitment therapy, and mindfulness-based practices, developed in collaboration with mental health experts, including psychologists and mental health counselors. Student feedback has also been incorporated into Wayhaven's working model. Specifically, before initiating this study, 74 students across 8 college campuses tested Wayhaven and responded to a brief open-ended survey eliciting feedback on the app. Students were also asked to participate in 45-minute focus groups (3 - 6 students per group) where they completed design thinking sessions related to a particular feature or conversation flow change in the app. Six focus groups were conducted prior to the study.

Wayhaven aims to provide comprehensive mental wellness support to college students by offering tailored psychoeducational content and emotional support resources. The AI is designed to replicate empathic responses and offer specific techniques to address emotions or concerns shared by users. For example, a user expressing stress might first be validated and supported by the coach's comments and then guided through a relaxation technique, ensuring a responsive and meaningful experience. Before starting a conversation, users select a coach persona that aligns with their preferences, such as a first-generation Latinx student or a university professor. This user-centered approach enhances engagement and flexibility, allowing users to switch between coaches while maintaining comprehensive support. Wayhaven also integrates

XSL•FO
RenderX

university-specific context, such as details about campus resources, student clubs, and upcoming events, to ensure users have quick access to services available on their campus. By combining evidence-based practices with personalized AI interactions, Wayhaven not only addresses mental wellness concerns but also amplifies the utility of a school's existing support infrastructure.

The user journey begins with account creation and onboarding, where participants provide demographic information and choose an AI mental wellness coach. Participants are able to choose from four options: Marcela (Latinx college student), Michelle (Black life coach), Professor Wilson (White college professor), and Coach Thompson (Black sports coach). Conversations follow a structured single-session intervention format, starting with the identification of a concern and a primary goal. The AI mental wellness chatbot then offers evidence-based tools, helps the user practice selected techniques, and collaborates to create an actionable plan. Each session concludes with a summary, and users can return for on-demand support at any time. To promote continued engagement, Wayhaven also sends push notifications to check in on users, ensuring ongoing and adaptive support. Figure 1 shows example conversations with the AI mental wellness coaches.

Figure 1. Example student conversations with the Wayhaven AI mental wellness coach. AI: artificial intelligence.

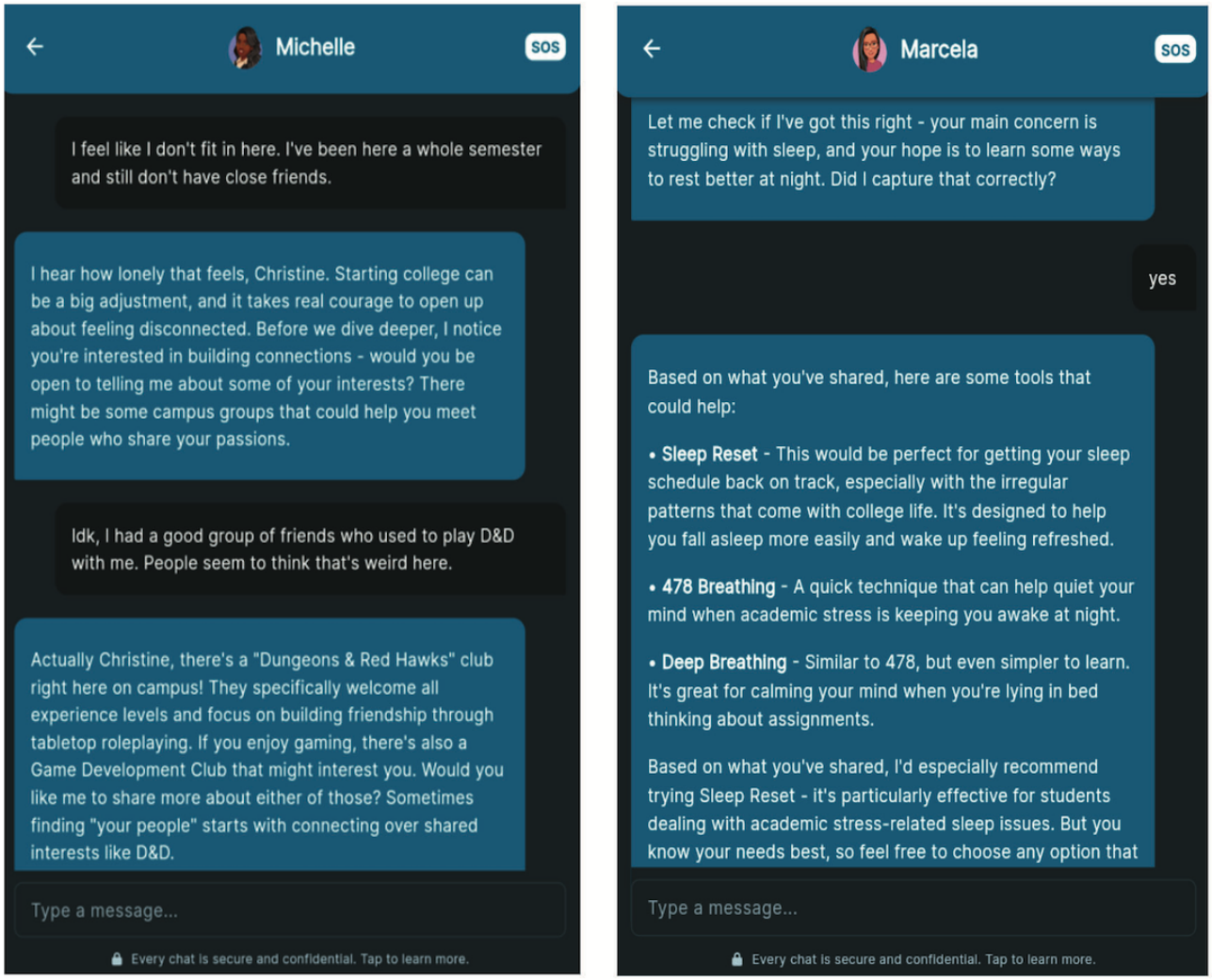

When creating an account, users are also informed that whenever they use Wayhaven, they will be chatting with an AI and not a real person, and that Wayhaven is not designed to address crises or substitute professional mental health care. Users seeking mental health services or reporting elevated distress are directed to campus-specific resources, such as the college counseling center. The app also features an "SOS" button, which allows users to connect to the Suicide and Crisis Lifeline (988) or the Crisis Text Line directly from the app. This feature appears at the top of each screen a user sees.

Measures

Overview

In addition to participant demographics, the preintervention survey included measures of anxiety, depression, hopelessness, agency, self-efficacy, and well-being. The postintervention survey included measures of anxiety, depression, hopelessness, agency, self-efficacy, and satisfaction. The final follow-up survey included all of the aforementioned mental health outcomes and satisfaction measures.

XSL•FO
RenderX

Screening Survey

To determine eligibility for the study, students were asked to indicate whether they were currently enrolled in the participating university, aged 18 years or older, currently living in the United States, and could read and understand English. Participants also completed the PHQ-2 [15] and the GAD-2 [16] to assess depression and anxiety symptoms, respectively.

Demographics

In the preintervention survey, participants were asked to report their age, sex assigned at birth, gender identity, sexual orientation, race or ethnicity, year in school, enrollment status, where they currently live, financial aid status, if they were a first-generation college student, if they were working outside of attending school, and whether they were currently receiving mental health treatment.

Anxiety

The Generalized Anxiety Disorder-7 (GAD-7) [17] was used to assess generalized anxiety symptoms at baseline, postintervention, and 1-week follow-up. Time periods for this measure were adjusted such that there was no overlap prior to using the AI chatbot, during the 1-week use period, and after the 1-week use period. Participants reported how often during the past 2 weeks they were bothered by seven anxiety symptoms (eg, feeling nervous, anxious, or "on edge") from 0=not at all to 3=every day. Scale scores were computed as the sum of all items. Cronbach α for this measure in this study was 0.74.

Depression

Depression symptoms were assessed using the Patient Health Questionnaire (PHQ-8) [18], an 8-item self-report measure that evaluates depressive symptoms within the previous 2 weeks, at baseline, postintervention, and 1-week follow-up. Time periods for this measure were adjusted such that there was no overlap prior to using the AI chatbot, during the 1-week use period, and after the 1-week use period. Participant responses ranged from 0=not at all to 3=nearly every day. Scale scores were computed as the sum of all items. Cronbach α for this measure in this study was 0.79.

Hopelessness

A modified version of the Beck Hopelessness Scale [19,20] was used to assess participants' feelings of hopelessness. This scale consists of 4 items, including 3 items from the original measure (eg, "My future seems dark to me") and 1 item from the Beck Depression Inventory [21] (eg, "I feel that the future is hopeless and that things cannot improve"). These 4 items have been found to correlate highly with the original Beck Hopelessness Scale and demonstrate excellent internal consistency [22-24]. Participant responses ranged from 0=absolutely disagree to 3=absolutely agree. Scale scores were computed as the sum of all items. Cronbach α for this measure in this study was 0.87.

Agency

The agency subscale of the State of Hope Scale [25] was used to assess agency or participants' beliefs in their ability to initiate and sustain goals or actions. This subscale consists of 3 items (eg, "At the present time, I am energetically pursuing my goals") measuring agency for goals. Participants rated how true each item was for them using an 8-point Likert scale, going from 1=definitely false to 8=definitely true. This measure has been found to be valid in previous research examining single-session interventions among adults [22]. Cronbach α for this measure in this study was 0.79.

Self-Efficacy

A modified version of the General Self-Efficacy Scale [26] was used to examine participants' self-efficacy or the belief that one is able to control challenging environmental demands by taking adaptive action. This scale consists of 6 items from the original measure (eg, "It is easy for me to stick to my aims and accomplish my goals"). Participant responses ranged from 1=not at all to 4=exactly true. This scale has been found to be valid and reliable [26]. Cronbach α for this measure in this study was 0.81.

Well-Being

The Short Warwick-Edinburgh Mental Wellbeing Scale [27] was used to assess both subjective and psychological well-being. Participants completed this measure at postintervention and follow-up. Participants were asked to indicate how frequently they had experienced the feelings described in each of the 7 items on the Short Warwick-Edinburgh Mental Wellbeing Scale over the past 2 weeks (eg, "I've been feeling optimistic about the future"). Responses ranged from 1=none of the time to 5=all of the time. This measure has been found to be valid and reliable among college student samples [28]. Cronbach α for this measure in this study was 0.68.

Participants were also asked to rate their mental wellness following their use of Wayhaven via one item developed by the study investigators. Participants were first provided with the following definition of mental wellness: "Mental wellness refers to our ability to thrive emotionally and psychologically (eg, feeling in control of life, able to manage stress or emotions, positive view of self, and knowing how to find support)." They were then asked to rate the extent to which their mental wellness changed compared to before their session with the AI mental wellness coach. Response options ranged from 1=much worse to 5=much better. Participants completed the single item at postintervention.

Engagement

Engagement with Wayhaven was assessed using data acquired through the app on use and engagement with the intervention, including unique sessions, number of messages sent, and time spent per session.

Satisfaction

Participant satisfaction with Wayhaven was assessed via 14 questions developed by the study investigators that were informed by the Bowen et al [29] framework. Specifically, 5 items examine approval with Wayhaven (eg, "I was satisfied with my experience using Wayhaven"), 4 items assessed demand (eg, "I would choose to use Wayhaven over other mental wellness resources available to me, such as the college counseling center"), 3 items measured practicality (eg, "Using Wayhaven fit easily into my daily routine"), and 1 item was

XSL•FO
RenderX

used to assess implementation (eg, "Wayhaven worked as intended without technical issues"). Participants were also asked to rate how engaging they found interactions with the Wayhaven AI mental wellness coach. Responses to this item ranged from 1=not at all to 5=extremely engaging. The list of items is provided in the Results section.

Statistical Analysis

Analyses were conducted using SPSS (version 28; IBM Corp). Descriptive statistics were used to summarize participant characteristics, primary outcomes, app use and engagement, and satisfaction with the AI coach. Univariate outliers were identified as having *z* scores greater than 3. When outliers were present, univariate analyses were run without outliers. Multilevel modeling (MLM) was used to examine the main study hypotheses. MLM was used because (1) it allows the inclusion of all participants, regardless of missing data, while efficiently handling missing data; and (2) it is considered the preferred method to analyze longitudinal psychiatric data [30]. The linear mixed-effects models procedure of SPSS 28 was used for these analyses. For depression on the PHQ-8, a two-level MLM model using maximum likelihood estimation was computed. At Level 1, depression symptoms varied within participants over time as a function of a person-specific growth curve. This level tested for time effects, estimating the change in depression across the pre, postintervention, and follow-up assessment points. Time was coded such that the intercept reflected the initial level of depression at preintervention. Similar MLM models were computed for anxiety, hopelessness, self-efficacy, and agency. For each MLM model, we also computed the proportional reduction explained at the individual level, conceptualized as a change in pseudo R^2 when the time was added to the empty model (ie, random intercept-only model without any predictors). We used the following Snijders and Bosker [31] formula: $R^2=1-(\sigma^2_F + \tau^2_F / \sigma^2_E + \tau^2_E)$, where σ^2_F represents the level-one random error variance for the full model with time; τ^2_F represents the level-two random error variance for the full model; σ^2_E represents the level-one random error variance for the empty model; and τ^2_E represents the level-two random error variance for the empty model. The effect size measure related to the variance explained for the overall model is $^2=R^2/1-R^2$ [32,33]. Guidelines for interpretation of ^{2} indicate that 0.02 is a small effect, 0.15 is a medium effect, and 0.35 is a large effect [34]. Finally, for outcome measures assessed at only two time points, such as well-being, 2-tailed paired samples *t* tests were used to examine within-participant change. Cohen *d* was calculated as a measure of effect size, with values of 0.2, 0.5, and 0.8 interpreted as small, medium, and large effects, respectively [32].

Results

Preliminary Analyses

Of the 113 students who were eligible to participate, 50 students completed the pre- and postintervention surveys and at least 1 session with the app, and 41 students completed the 1-week follow-up survey. No significant differences were found between those who completed the follow-up survey and those who did not in terms of gender or baseline severity of depression or anxiety symptoms. A significant difference was found in terms of age (t_{48}= 2.42; P=.02), such that participants who did not complete the follow-up survey were significantly older (mean age 25, SD 6.41 y) than participants who completed the follow-up survey (mean age 21.4, SD 3.51 y). However, this difference is relatively modest and may not represent a clinically meaningful distinction among this age group. Accuracy checks assessing the range of summed scores (minimum and maximum scores) for each scale indicated that participants accurately completed the scales.

Of the 50 students who completed at least one session, 90% (n=45) were female. The mean age of the sample was 22.12 (SD 4.42) years. About 46% (n=23) self-identified as Latinx, 34% (n=17) non-Latinx White, 8% (n=4) non-Latinx Black, 8% (n=4) non-Latinx Asian, and 4% (n=2) multiracial. Most students were seniors (n=14, 28%). Half the students self-identified as heterosexual or straight (n=25). Most students were enrolled full-time (n=46, 92%), and 58% (n=29) were commuters or lived off-campus. Most students (n=36, 72%) also reported receiving federal or state financial assistance for their college education and being first-generation college students (n=29, 58%; see Table 1 for demographic characteristics of the sample).

The mean baseline PHQ-8 and GAD-7 scores were 13.50 (SD 5.06; range 4 - 21) and 14 (SD 4.03; range 3 - 23), respectively, with 80% (n=40) of students scoring above the clinical cutoff on the PHQ-8 (score ≥10) and 88% (n=44) on GAD-7 (score ≥10). The average hopelessness score was 5.38 (SD 2.67; range 0 - 12). The mean agency score was 14.14 (SD 5.29; range 3 - 23), self-efficacy was 15.98 (SD 3.98; range 4 - 23), and wellness was 19.50 (SD 4.52; range 4 - 29). Table 2 presents the mean score for each outcome of interest across the three study time points.

XSL•FO
RenderX

Table . Demographic characteristics.

Variable	Value, n (%)
Sex[a]	
Female	45 (90)
Male	5 (10)
Gender[a]	
Female	40 (80)
Male	4 (8)
Transgender man or transman	1 (2)
Genderqueer	4 (8)
Other	1 (2)
Sexual orientation	
Straight or heterosexual	25 (50)
Bisexual	11 (22)
Gay or lesbian	4 (8)
Pansexual	3 (6)
Queer	3 (6)
Asexual	2 (4)
Other	2 (4)
Race or ethnicity	
Latinx	23 (46)
Non-Latinx White	17 (34)
Non-Latinx Black	4 (8)
Non-Latinx Asian	4 (8)
Multiracial	2 (4)
Year in school	
Freshman	8 (16)
Sophomore	10 (20)
Junior	11 (22)
Senior	14 (28)
Graduate student	7 (14)
Enrollment status	
Full-time	46 (92)
Part-time	4 (8)
Current living	
On-campus (ie, university housing)	21 (42)
Off-campus (ie, nonuniversity housing near campus)	6 (12)
Commuter (ie, permanent home or residence a significant distance from campus)	23 (46)
Receiving financial aid	
Yes	36 (72)
No	14 (28)
First-generation college student	
Yes	29 (58)

XSL•FO
RenderX

Variable	Value, n (%)
No	21 (42)
Working outside of attending school	
Yes (employed part-time)	24 (48)
Yes (employed full-time)	5 (10)
No	19 (38)
Prefer not to disclose	2 (4)
Current mental health treatment	
Yes	18 (36)
No	27 (54)
Prefer not to disclose	5 (10)

[a]Both sex and gender are reported to reflect biological and self-identified dimensions, respectively. While the distributions are similar in this sample, we report both for transparency and inclusivity.

Table . Means and SDs for outcomes across assessments.

Variable	Pre (n=50), mean (SD)	Post[a] (n=50), mean (SD)	1-week follow-up (n=41), mean (SD)
Depression (PHQ-8)[b]	13.50 (5.06)	13.46 (2.92)	10.17 (5.31)
Anxiety (GAD-7)[c]	14.00 (4.03)	11.14 (4.08)	9.76 (5.64)
Hopelessness (BHS)[d]	5.38 (2.67)	4.18 (2.65)	4.27 (2.71)
Agency (SHS)[e]	14.14 (5.29)	15.52 (4.98)	15.24 (5.34)
Self-Efficacy (GSE)[f]	15.98 (3.98)	16.84 (4.07)	16.95 (3.82)
Well-Being (SWEMWBS)[g]	10.50 (4.52)	N/A[h]	21.66 (5.05)

[a]Postassessment occurred after the first session.
[b]PHQ-8: Patient Health Questionnaire-8.
[c]GAD-7: Generalized Anxiety Disorder-7.
[d]BHS: Beck Hopelessness Scale.
[e]SHS: State of Hope Scale.
[f]GSE: General Self-Efficacy Scale.
[g]SWEMWBS: Short Warwick-Edinburgh Mental Well-being Scale.
[h]Not applicable.

The mean number of sessions attended was 2.02 (SD 1.86; range 1 - 10), suggesting that, on average, students completed around two sessions over the 1-week period with the AI mental health wellness coach. Of the four coach characters offered, the mean number of coaches used was 1.14 (SD 0.45; range 1 - 3), which may suggest that students often opted to use the same coach character across sessions. Specifically, Michelle was selected a total of 19 times, Marcela was selected a total of 20 times, Professor Wilson was selected a total of 16 times, and Coach Thompson was selected a total of 2 times. On average, students used 1.1 (SD 1.13; range 0 - 5) CBT-based skills during their conversations (ie, grounding techniques, mindfulness exercises, and exploring and challenging cognitive distortions). One univariate outlier was identified with regard to minutes spent per session (ie, *z* score above 3). After removing this participant from the analysis, the average minutes students spent in each session was 15.43 (SD 10.11; range 3.18 - 53.48). Two outliers were also present with regard to message count. The average message count was 29.08 (SD 22.43; range 5 - 115) after these two students were removed from the analysis. Finally, two outliers were identified for students' response time. The students' mean response time to each AI-generated response was 7.33 (SD 24.55; range 0.13 - 143.37) minutes once these cases were excluded.

Improvement in Outcome Scores

A linear model with random slopes and intercepts revealed that students' GAD-7 scores decreased significantly from preintervention to 1-week follow-up (β=−2.15; P<.001). About 14% of the variance in anxiety was explained by time (R^2=0.13; ^{2}=0.14). Similarly, there was a significant reduction in students' PHQ-8 (β=−1.62; P<.001) and hopelessness scores (β=−.64; P<.001), with time accounting for 8% and 3% of the variance, respectively (R^2=0.07, ^{2}=0.08; R^2=0.02, ^{2}=0.03, respectively). With regard to agency and efficacy, students' scores significantly increased from preintervention to follow-up (β=.64,

XSL•FO
RenderX

P=.32; β=.53, *P*=.02, respectively). However, time accounted for less than 1% of the variance in agency (R^2=0.007; ^{2}=0.007) and efficacy (R^2=0.009 ^{2}=0.009). See Table 3 and Multimedia Appendix 1 for results of MLM analyses. To examine changes in well-being, a 2-tailed paired samples *t* test was computed. Students' well-being scores significantly increased from preintervention (mean 18.98, SD 4.35) to follow-up (mean 21.66, SD 0.79; t_{40}=2.90; *P*=.006, *d*=0.45). Relatedly, at postintervention, 62% (n=31) of students reported that their mental wellness was better or much better after one session with Wayhaven. As a sensitivity analysis, the current mental health treatment status was added as a covariate to the linear mixed models. This variable did not meaningfully affect the pattern or significance of the results. Thus, it was not included in the final models reported here.

Table . Parameters for time predicting change in outcomes over 1 week post intervention.

Outcome and predictor	Estimate	SE	T ratio	95% CI	2[a]
Anxiety					0.14
Intercept	13.76	0.59	23.34[b]	12.60 to 14.93	
Time	−2.15	0.47	−4.60[b]	−3.08 to −1.22	
Depression					0.08
Intercept	14.03	0.60	23.50[b]	12.84 to 15.21	
Time	−1.62	0.43	−3.74[b]	−2.47 to −0.76	
Hopelessness					0.03
Intercept	5.19	0.37	13.99[b]	4.45 to 5.95	
Time	−.64	0.16	−4.09[b]	−0.94 to −0.33	
Agency					0.007
Intercept	14.39	0.72	19.95[b]	12.95 to 15.83	
Time	.64	0.30	2.17[c]	0.06 to 1.23	
Self-efficacy					0.009
Intercept	16.09	0.55	29.37[b]	14.99 to 17.18	
Time	.54	0.23	2.35[c]	0.08 to 0.99	

[a] 2: effect size.

[b]*P*<.001.

[c]*P*<.05.

Satisfaction With the App

Immediately after using Wayhaven (ie, at postintervention), participants generally endorsed that they found Wayhaven to be acceptable. Specifically, 90% (n=45) of participants agreed that Wayhaven was easy to use, and 74% (n=37) of participants indicated that they were satisfied with their experience with Wayhaven. Most participants (n=36, 72%) also agreed that they would use Wayhaven again in the future. The majority of students (n=42, 84%) participants also agreed that there is a need for a resource like Wayhaven among college students. Many students also reported finding the app quite engaging (n=26, 53%).

At 1-week follow-up, there was a slight decrease in students' satisfaction, with 65% (27/41) of participants indicating that they were satisfied with their experience using Wayhaven. Additionally, 59% (24/41) of participants indicated that they would use Wayhaven again in the future. Despite these shifts, 68% (27/41) of participants noted that they would recommend Wayhaven to a friend who may be experiencing similar mental wellness concerns. Participants also endorsed the strong feasibility of Wayhaven, such that 80% (33/41) of participants felt that Wayhaven fit easily into their daily routine (see Table 4 for detailed frequency of student responses to satisfaction questions at postintervention and follow-up). More than half of the students also continued to find the app quite engaging (22/41, 56%).

XSL•FO
RenderX

Table . Frequency of student responses to items on satisfaction survey (items 1 - 13).

Item and time point	Strongly disagree, n (%)	Disagree, n (%)	Neither agree nor disagree, n (%)	Agree, n (%)	Strongly agree, n (%)	Prefer not to say, n (%)
I found Wayhaven easy to use						
Postintervention	1 (2)	0 (0)	3 (6)	17 (34)	28 (56)	1 (2)
Follow-up	1 (2)	1 (2)	1 (2)	20 (49)	18 (44)	0 (0)
I was satisfied with my experience using Wayhaven						
Postintervention	1 (2)	4 (8)	8 (16)	22 (44)	15 (30)	0 (0)
Follow-up	1 (2)	5 (12)	8 (20)	13 (32)	14 (34)	0 (0)
I would use Wayhaven again in the future						
Postintervention	2 (4)	3 (6)	9 (18)	20 (40)	16 (32)	0 (0)
Follow-up	3 (7)	7 (17)	7 (17)	7 (17)	17 (42)	0 (0)
Wayhaven fits well with how I prefer to address my mental wellness concerns						
Postintervention	4 (8)	4 (8)	18 (36)	14 (28)	10 (20)	0 (0)
Follow-up	6 (15)	6 (15)	13 (32)	7 (17)	9 (22)	0 (0)
People close to me agree or would agree with me using Wayhaven						
Postintervention	6 (12)	3 (6)	14 (28)	18 (36)	9 (18)	0 (0)
Follow-up	3 (7)	4 (10)	7 (17)	18 (44)	9 (22)	0 (0)
I would recommend Wayhaven to a friend experiencing similar mental wellness concerns						
Postintervention	4 (8)	4 (8)	6 (12)	24 (48)	11 (22)	1 (2)
Follow-up	4 (10)	3 (7)	6 (15)	16 (39)	11 (27)	1 (2)
I would choose to use Wayhaven over other mental wellness resources available to me, such as the college counseling center						
Postintervention	8 (16)	10 (20)	13 (26)	11 (22)	7 (14)	1 (2)
Follow-up	9 (22)	12 (29)	5 (12)	4 (10)	10 (24)	1 (2)
I intend to continue using Wayhaven						
Postintervention	3 (6)	6 (12)	7 (14)	20 (40)	14 (28)	0 (0)
Follow-up	5 (12)	3 (7)	7 (17)	14 (34)	11 (27)	1 (2)
I believe there is a need for a resource like Wayhaven among college students						
Postintervention	2 (4)	1 (2)	4 (8)	17 (34)	25 (50)	1 (2)
Follow-up	2 (5)	1 (2)	2 (5)	14 (34)	22 (54)	0 (0)
I was able to use Wayhaven without requiring any additional resources or support						
Postintervention	2 (4)	2 (4)	5 (10)	11 (22)	30 (60)	0 (0)
Follow-up	2 (5)	1 (2)	3 (7)	15 (37)	20 (49)	0 (0)
Using Wayhaven fits easily into my daily routine						
Postintervention	1 (2)	1 (2)	6 (12)	24 (48)	17 (34)	1 (2)
Follow-up	1 (2)	0 (0)	7 (17)	14 (34)	19 (46)	0 (0)
The advice that Wayhaven provided was practical and easy for me to carry out in my daily life						
Postintervention	2 (4)	3 (6)	6 (12)	23 (46)	15 (30)	1 (2)
Follow-up	1 (2)	3 (7)	10 (24)	14 (34)	13 (32)	0 (0)
Wayhaven worked as intended without technical issues						
Postintervention	0 (0)	3 (6)	4 (8)	16 (32)	25 (50)	2 (4)
Follow-up	1 (2)	4 (10)	3 (7)	13 (32)	19 (46)	1 (2)

XSL•FO
RenderX

Discussion

Principal Results

The purpose of this study was to explore how college students interact with Wayhaven's AI-powered mental wellness chatbot and its feasibility for improving mental wellness among college students with elevated depression or anxiety symptoms. Specifically, this study sought to assess the feasibility and acceptability (ie, engagement and satisfaction) of using Wayhaven with students with elevated depression or anxiety symptoms, as well as whether app use was linked with improvement across several mental health outcomes over a 1-week period. Most participants self-identified as racially and ethnically minoritized and were experiencing clinically elevated depression or anxiety symptoms. Overall, Wayhaven use was associated with a significant decrease in depression, anxiety, and hopelessness from preintervention to follow-up. Wayhaven use was also related to a significant increase in agency, efficacy, and well-being across the study period. Additionally, students reported being satisfied with Wayhaven and seeing the app as a new important resource for college students.

More than half of the participants scored above the clinical cutoff on the PHQ-8 and GAD-7, which is in line with previous research documenting elevated rates of internalizing symptoms among college students [2]. It is clear that there is a need for increased access to efficacious and equitable mental health resources for college students, particularly those self-identifying as racially and ethnically minoritized [35,36]. Of note, 66% (n=33) of students in our sample were racially and ethnically minoritized. DMHIs may be particularly appealing to racially and ethnically minoritized students who are more likely than White students to indicate a preference for web-based treatment versus in-person therapy [37,38]. Our findings suggest that Wayhaven may be a viable mental health resource to such populations, given the reduction in depression, anxiety, and hopelessness scores in a relatively short period of time.

Results also suggest that the majority of students were satisfied with their experience using Wayhaven. Most students felt that there was a need for a resource like Wayhaven among college students and that they would recommend it to their peers. With regard to the coach personalities offered, most students opted to use the same coach persona across sessions and selected Marcela, the Latinx college student, most frequently. This preference among students aligns with previous research suggesting that college students tend to favor seeking help from informal sources, like peers [39]. In addition to rating the app as engaging, students completed two sessions on the app on average across the 1-week study period, with the average message count across sessions being around 29 (SD 22.43). The number of messages students exchanged is comparable to those reported by He et al [40], who tested a cognitive behavioral therapy−based mental health chatbot, XiaoE, for young adults with depressive symptoms. They found that the average number of messages exchanged with XiaoE was 25.54 (SD 26.25) daily across a 1-week period. It should be noted that participants in this study were instructed and prompted to engage with XiaoE daily, whereas students in this study were asked to use Wayhaven as needed. It is unclear whether user engagement (ie, sessions and messages) would be higher if students were asked to use the tool daily. Taken together, these data suggest that Wayhaven appears to be engaging, acceptable, and feasible for college students with elevated depression or anxiety symptoms.

Limitations

While these preliminary findings support the applicability of Wayhaven in college populations, this study is not without its limitations; caution should be used when considering the generalizability of the findings. This was a small, uncontrolled pilot study, and we did not include a control group. Therefore, we cannot infer that the observed symptom improvements were a result of the intervention. However, our goal was not to conduct a randomized controlled trial (RCT); instead, our aim was to test the feasibility and acceptability of a novel AI mental wellness chatbot in a real-world setting. Testing an intervention's feasibility and acceptability via an open trial is an important first step prior to an RCT [41]. As noted above, this research should be followed up with a larger RCT comparing Wayhaven to an active control condition matched by age and sex and a longer follow-up period. The use of self-report assessments is also limited in that participants may have under- or overreported the severity of their symptoms. Importantly, self-report questionnaires of depression and anxiety symptoms have been found to be moderately to strongly correlated with clinician-rated scales, and previous work using these measures has proven to be clinically impactful [42,43]. The majority of participants were also female, possibly limiting generalizability to male participants. Additionally, our sample was racially and ethnically diverse, participants were obtained from a single public university in New Jersey, and results may not be generalizable to other areas of the United States.

Although Wayhaven use was associated with a statistically significant improvement in outcome scores across time, the magnitude of these effects was small. The observed effect sizes were strongest for anxiety, followed by depression, which were our main clinical outcomes. These findings are consistent with prior research on brief interventions that suggests such programs have the strongest effects on anxiety symptoms [44,45]. Researchers have noted that categorizing effect sizes as "small," "medium," or "large" is not inherently meaningful without a contextual frame of reference [46]. In public health, even small effects can yield a substantial impact when scaled across populations. Small shifts in mean scores can disproportionately affect the extreme end of a population distribution and drive higher rates of clinical referrals, as illustrated by the surge in mental health service use among youth during the COVID-19 pandemic despite only a modest increase in mental health outcome scores [47]. Moreover, brief interventions with small average effects may still yield substantial population-level benefits by offering a low-cost means to expand the reach and impact of mental health services, particularly for individuals in need, like college students who may be reluctant to seek higher-intensity support [45,48]. Nevertheless, future research should explore ways to strengthen and sustain the impact of Wayhaven, such as incorporating booster sessions or embedding the intervention within a broader stepped-care framework.

XSL•FO
RenderX

Comparison With Prior Work

Our findings mirror previous research on the effectiveness of AI chatbots in reducing depression and anxiety symptoms among college students [13,14,41]. However, in contrast to previous studies testing scripted chatbots, Wayhaven's generative AI-powered mental wellness chatbot offered students personalized, context-sensitive support by integrating demographic information, university-specific resources, and a customized AI coach persona. Given our preliminary findings, future research should focus on comparing the effectiveness of Wayhaven to an active control condition, such as a mental health app or internet-based intervention, and among a larger sample of students. If proven effective, Wayhaven can serve as a lower-intensity intervention that college counseling centers can offer to students presenting with subclinical or acute depression or anxiety symptoms, which can help free in-person services to students experiencing more severe symptoms or distress.

Conclusions

Despite limitations, results revealed that Wayhaven led to an improvement in student mental well-being outcomes (ie, depression, anxiety, hopelessness, agency, efficacy, and well-being) and students were satisfied with the tool. Wayhaven may be a promising resource to deliver brief, evidence-based text conversations to support college students' mental wellness needs. Future research should examine the effectiveness of Wayhaven by comparing it to an active control condition and including a larger student sample. With the already overwhelming and unmet mental health needs of this population, a DMHI like Wayhaven may offer a pathway to free, accessible, and equitable mental wellness care.

Acknowledgments

The authors give special thanks to all the students who participated in the study as well as to Erik Murphy for his assistance with the app use data.

Data Availability

The data presented in this study are not publicly available due to privacy concerns but are available from the corresponding author upon reasonable request.

Authors' Contributions

JAR-P and AS contributed substantially to the conception and design of the study, the collection and analysis of data, and the writing of the manuscript. KM and JH contributed significantly to the conception and design of the study and writing of the manuscript. CN and AG provided a critical review of the manuscript and significantly contributed to the development of the Wayhaven artificial intelligence mental wellness chatbot and the acquisition of the app use data. CJ provided a critical review and significant contributions to the manuscript. All authors contributed to the article and approved the submitted version.

Conflicts of Interest

CN and AG are employees of Wayhaven. JAR-P, JH, and CJ serve on the Wayhaven advisory board. CJ owns shares of Wayhaven stock. None of the authors was paid for the results of the study.

Multimedia Appendix 1
Multilevel models for time predicting change in outcomes over 1 week post intervention.
[DOCX File, 19 KB - formative_v9i1e71923_app1.docx]

References

1. McLafferty M, Brown N, McHugh R, et al. Depression, anxiety and suicidal behaviour among college students: comparisons pre-COVID-19 and during the pandemic. Psychiatry Res Commun 2021 Dec;1(2):100012. [doi: 10.1016/j.psycom.2021.100012] [Medline: 34977911]
2. Wang X, Hegde S, Son C, Keller B, Smith A, Sasangohar F. Investigating mental health of US college students during the COVID-19 pandemic: cross-sectional survey study. J Med Internet Res 2020 Sep 17;22(9):e22817. [doi: 10.2196/22817] [Medline: 32897868]
3. Lipson SK, Diaz Y, Davis J, Eisenberg D. Mental health among first-generation college students: findings from the National Healthy Minds Study, 2018-2021. Cogent Ment Health 2023;2(1):2220358. [doi: 10.1080/28324765.2023.2220358] [Medline: 38145239]
4. Gorman B, Hargraves L, Huynh S, Parker S, Wiafe B. Early findings from Transitions I: mental health, stress and health-related behaviors. : Transitions; 2020 URL: https://transitions.web.unc.edu/wp-content/uploads/sites/22616/2020/10/Transitions-Wave-One-Report.pdf [accessed 2025-01-01]
5. Walls KL. Navigating a new model for therapy at a university counseling center: a case example. Psychol Serv 2023 Nov;20(4):952-961. [doi: 10.1037/ser0000784] [Medline: 37307319]

XSL•FO
RenderX

6. Gulliver A, Griffiths KM, Christensen H. Perceived barriers and facilitators to mental health help-seeking in young people: a systematic review. BMC Psychiatry 2010 Dec 30;10:113. [doi: 10.1186/1471-244X-10-113] [Medline: 21192795]
7. Vidourek RA, King KA, Nabors LA, Merianos AL. Students' benefits and barriers to mental health help-seeking. Health Psychol Behav Med 2014 Jan 1;2(1):1009-1022. [doi: 10.1080/21642850.2014.963586] [Medline: 25750831]
8. Lipschitz JM, Pike CK, Hogan TP, Murphy SA, Burdick KE. The engagement problem: a review of engagement with digital mental health interventions and recommendations for a path forward. Curr Treat Options Psychiatry 2023 Sep;10(3):119-135. [doi: 10.1007/s40501-023-00297-3] [Medline: 38390026]
9. Boucher EM, Harake NR, Ward HE, et al. Artificially intelligent chatbots in digital mental health interventions: a review. Expert Rev Med Devices 2021 Dec;18(sup1):37-49. [doi: 10.1080/17434440.2021.2013200] [Medline: 34872429]
10. Koulouri T, Macredie RD, Olakitan D. Chatbots to support young adults' mental health: an exploratory study of acceptability. ACM Trans Interact Intell Syst 2022 Jun 30;12(2):1-39. [doi: 10.1145/3485874]
11. Kuhail MA, Alturki N, Thomas J, Alkhalifa AK. Human vs. AI counseling: college students' perspectives. Comput Hum Behav Rep 2024 Dec;16:100534. [doi: 10.1016/j.chbr.2024.100534]
12. Rasouli S, Ghafurian M, Nilsen ES, Dautenhahn K. University students' opinions on using intelligent agents to cope with stress and anxiety in social situations. Comput Human Behav 2024 Apr;153:108072. [doi: 10.1016/j.chb.2023.108072]
13. Fitzpatrick KK, Darcy A, Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. JMIR Ment Health 2017 Jun 6;4(2):e19. [doi: 10.2196/mental.7785] [Medline: 28588005]
14. Klos MC, Escoredo M, Joerin A, Lemos VN, Rauws M, Bunge EL. Artificial intelligence-based chatbot for anxiety and depression in university students: pilot randomized controlled trial. JMIR Form Res 2021 Aug 12;5(8):e20678. [doi: 10.2196/20678] [Medline: 34092548]
15. Kroenke K, Spitzer RL, Williams JBW. The Patient Health Questionnaire-2: validity of a two-item depression screener. Med Care 2003 Nov;41(11):1284-1292. [doi: 10.1097/01.MLR.0000093487.78664.3C] [Medline: 14583691]
16. Kroenke K, Spitzer RL, Williams JBW, Monahan PO, Löwe B. Anxiety disorders in primary care: prevalence, impairment, comorbidity, and detection. Ann Intern Med 2007 Mar 6;146(5):317-325. [doi: 10.7326/0003-4819-146-5-200703060-00004] [Medline: 17339617]
17. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
18. Kroenke K, Strine TW, Spitzer RL, Williams JBW, Berry JT, Mokdad AH. The PHQ-8 as a measure of current depression in the general population. J Affect Disord 2009 Apr;114(1-3):163-173. [doi: 10.1016/j.jad.2008.06.026] [Medline: 18752852]
19. Beck AT, Weissman A, Lester D, Trexler L. The measurement of pessimism: the hopelessness scale. J Consult Clin Psychol 1974 Dec;42(6):861-865. [doi: 10.1037/h0037562] [Medline: 4436473]
20. Forintos DP, Rózsa S, Pilling J, Kopp M. Proposal for a short version of the Beck Hopelessness Scale based on a national representative survey in Hungary. Community Ment Health J 2013 Dec;49(6):822-830. [doi: 10.1007/s10597-013-9619-1] [Medline: 23756722]
21. Beck AT, Ward CH, Mendelson M, Mock J, Erbaugh J. An inventory for measuring depression. Arch Gen Psychiatry 1961 Jun;4:561-571. [doi: 10.1001/archpsyc.1961.01710120031004] [Medline: 13688369]
22. Schleider JL, Mullarkey MC, Fox KR, et al. A randomized trial of online single-session interventions for adolescent depression during COVID-19. Nat Hum Behav 2022 Feb;6(2):258-268. [doi: 10.1038/s41562-021-01235-0] [Medline: 34887544]
23. Brott KH, Veilleux JC. Examining state self-criticism and self-efficacy as factors underlying hopelessness and suicidal ideation. Suicide Life Threat Behav 2024 Apr;54(2):207-220. [doi: 10.1111/sltb.13034] [Medline: 38112324]
24. Sung JY, Bugatti M, Vivian D, Schleider JL. Evaluating a telehealth single-session consultation service for clients on psychotherapy wait-lists. Pract Innov (Wash D C) 2023;8(2):141-161. [doi: 10.1037/pri0000207]
25. Snyder CR, Sympson SC, Ybasco FC, Borders TF, Babyak MA, Higgins RL. Development and validation of the State Hope Scale. J Pers Soc Psychol 1996 Feb;70(2):321-335. [doi: 10.1037//0022-3514.70.2.321] [Medline: 8636885]
26. Romppel M, Herrmann-Lingen C, Wachter R, et al. A short form of the General Self-Efficacy Scale (GSE-6): development, psychometric properties and validity in an intercultural non-clinical sample and a sample of patients at risk for heart failure. Psychosoc Med 2013;10:Doc01. [doi: 10.3205/psm000091] [Medline: 23429426]
27. Stewart-Brown S, Tennant A, Tennant R, Platt S, Parkinson J, Weich S. Internal construct validity of the Warwick-Edinburgh Mental Well-being Scale (WEMWBS): a Rasch analysis using data from the Scottish Health Education Population Survey. Health Qual Life Outcomes 2009 Feb 19;7:15. [doi: 10.1186/1477-7525-7-15] [Medline: 19228398]
28. Barbayannis G, Bandari M, Zheng X, Baquerizo H, Pecor KW, Ming X. Academic stress and mental well-being in college students: correlations, affected groups, and COVID-19. Front Psychol 2022;13:886344. [doi: 10.3389/fpsyg.2022.886344] [Medline: 35677139]
29. Bowen DJ, Kreuter M, Spring B, et al. How we design feasibility studies. Am J Prev Med 2009 May;36(5):452-457. [doi: 10.1016/j.amepre.2009.02.002] [Medline: 19362699]
30. Seidel A, Presnell K, Rosenfield D. Mediators in the dissonance eating disorder prevention program. Behav Res Ther 2009 Aug;47(8):645-653. [doi: 10.1016/j.brat.2009.04.007] [Medline: 19457474]

XSL•FO
RenderX

31. Snijders TAB, Bosker RJ. Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling, 2nd edition: Sage Publishers; 2012.
32. Cohen J. Statistical Power Analysis for the Behavioral Sciences, 2nd edition: Lawrence Erlbaum Associates Publishers; 1988.
33. Lorah J. Effect size measures for multilevel models: definition, interpretation, and TIMSS example. Large-scale Assess Educ 2018 Dec;6(1). [doi: 10.1186/s40536-018-0061-2]
34. Cohen J. A power primer. Psychol Bull 1992 Jul;112(1):155-159. [doi: 10.1037//0033-2909.112.1.155] [Medline: 19565683]
35. Lipson SK, Lattie EG, Eisenberg D. Increased rates of mental health service utilization by U.S. college students: 10-year population-level trends (2007-2017). Psychiatr Serv 2019 Jan 1;70(1):60-63. [doi: 10.1176/appi.ps.201800332] [Medline: 30394183]
36. Miranda R, Soffer A, Polanco-Roman L, Wheeler A, Moore A. Mental health treatment barriers among racial/ethnic minority versus white young adults 6 months after intake at a college counseling center. J Am Coll Health 2015;63(5):291-298. [doi: 10.1080/07448481.2015.1015024] [Medline: 25692661]
37. Dunbar MS, Sontag-Padilla L, Kase CA, Seelam R, Stein BD. Unmet mental health treatment need and attitudes toward online mental health services among community college students. Psychiatr Serv 2018 May 1;69(5):597-600. [doi: 10.1176/appi.ps.201700402] [Medline: 29540117]
38. Lungu A, Sun M. Time for a change: college students' preference for technology-mediated versus face-to-face help for emotional distress. Telemed e-Health 2016 Dec;22(12):991-1000. [doi: 10.1089/tmj.2015.0214] [Medline: 27223266]
39. Eisenberg D, Hunt J, Speer N. Help seeking for mental health on college campuses: review of evidence and next steps for research and practice. Harv Rev Psychiatry 2012;20(4):222-232. [doi: 10.3109/10673229.2012.712839] [Medline: 22894731]
40. He Y, Yang L, Zhu X, et al. Mental health chatbot for young adults with depressive symptoms during the COVID-19 pandemic: single-blind, three-arm randomized controlled trial. J Med Internet Res 2022 Nov 21;24(11):e40719. [doi: 10.2196/40719] [Medline: 36355633]
41. Rounsaville BJ, Carroll KM, Onken LS. A stage model of behavioral therapies research: getting started and moving on from stage I. Clin Psychol Sci Pract 2001;8(2):133-142. [doi: 10.1093/clipsy.8.2.133]
42. Rush AJ, Carmody TJ, Ibrahim HM, et al. Comparison of self-report and clinician ratings on two inventories of depressive symptomatology. Psychiatr Serv 2006 Jun;57(6):829-837. [doi: 10.1176/ps.2006.57.6.829] [Medline: 16754760]
43. Uher R, Perlis RH, Placentino A, et al. Self-report and clinician-rated measures of depression severity: can one replace the other?.. Depress Anxiety 2012 Dec;29(12):1043-1049. [doi: 10.1002/da.21993] [Medline: 22933451]
44. Cape J, Whittington C, Buszewicz M, Wallace P, Underwood L. Brief psychological therapies for anxiety and depression in primary care: meta-analysis and meta-regression. BMC Med 2010 Jun 25;8:38. [doi: 10.1186/1741-7015-8-38] [Medline: 20579335]
45. Schleider JL, Zapata JP, Rapoport A, et al. Single-session interventions for mental health problems and service engagement: umbrella review of systematic reviews and meta-analyses. Annu Rev Clin Psychol 2025 May;21(1):279-303. [doi: 10.1146/annurev-clinpsy-081423-025033] [Medline: 39874601]
46. Funder DC, Ozer DJ. Evaluating effect size in psychological research: sense and nonsense. Adv Methods Pract Psychol Sci 2019 Jun;2(2):156-168. [doi: 10.1177/2515245919847202]
47. Carey EG, Ridler I, Ford TJ, Stringaris A. Editorial perspective: when is a "small effect" actually large and impactful? J Child Psychol Psychiatry 2023 Nov;64(11):1643-1647. [doi: 10.1111/jcpp.13817] [Medline: 37226639]
48. Greenberg MT, Abenavoli R. Universal interventions: fully exploring their impacts and potential to produce population-level impacts. J Res Educ Eff 2017 Jan 2;10(1):40-67. [doi: 10.1080/19345747.2016.1246632] [Medline: 22180789]

Abbreviations

AI: artificial intelligence
CBT: cognitive behavioral therapy
DMHI: digital mental health intervention
GAD: Generalized Anxiety Disorder
MLM: multilevel modeling
PHQ: Patient Health Questionnaire
RCT: randomized controlled trial

XSL•FO
RenderX

Edited by A Bucher; submitted 30.01.25; peer-reviewed by L Hecht, P Watson; revised version received 26.05.25; accepted 26.05.25; published 28.07.25.

Please cite as:
Reyes-Portillo JA, So A, McAlister K, Nicodemus C, Golden A, Jacobson C, Huberty J
Generative AI–Powered Mental Wellness Chatbot for College Student Mental Wellness: Open Trial
JMIR Form Res 2025;9:e71923
URL: https://formative.jmir.org/2025/1/e71923
doi:10.2196/71923

© Jazmin Reyes-Portillo, Amy So, Kelsey McAlister, Christine Nicodemus, Ashleigh Golden, Colleen Jacobson, Jennifer Huberty. Originally published in JMIR Formative Research (https://formative.jmir.org), 28.7.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Feasibility of a Mental Health App Intervention for Emergency Service Workers and Volunteers: Single-Arm Pilot Study

Denise Meuldijk[1], PhD; Mark Deady[1], PhD; Daniel A J Collins[1], MSc; Douglas O Williams[1], MSc; Richard A Bryant[1,2], PhD; Samuel B Harvey[1], PhD

[1]Black Dog Institute, Faculty of Medicine and Health, University of New South Wales, High St, Kensington, Sydney, Australia
[2]School of Psychology, University of New South Wales, Kensington, Australia

Corresponding Author:
Denise Meuldijk, PhD
Black Dog Institute, Faculty of Medicine and Health, University of New South Wales, High St, Kensington, Sydney, Australia

Abstract

Background: Emergency service workers are at an elevated risk for stressor-related mental health (MH) issues, such as anxiety, depression, and posttraumatic stress disorder. Barriers to help-seeking are widespread across this sector, necessitating interventions tailored to the unique needs of this population. Build Back Better is a smartphone-based intervention designed to provide evidence-based strategies for the prevention of anxiety, depression, and posttraumatic stress disorder among emergency service workers.

Objective: This study aimed to evaluate the usability, acceptability, feasibility, and preliminary effectiveness of the Build Back Better app among emergency service workers.

Methods: A single-group (N=67), 1-month pilot study assessing the impact of the Build Back Better app on MH outcomes, including general distress, anxiety, depression, and traumatic stress coping, was undertaken with emergency service workers. Participants completed baseline and 1-month follow-up assessments using the Kessler Psychological Distress Scale, 9-item Patient Health Questionnaire, 7-item Generalized Anxiety Disorder, World Health Organization Well-Being Index, and the Trauma Coping Self-Efficacy Scale. The app's usability and acceptability were also evaluated through participant feedback and usage data.

Results: Of the 71 participants enrolled, 67 completed the baseline assessment and downloaded the app, with 33 participants providing follow-up data. The mean age of participants was 44.73 (SD 11.4) years, and 64% (n=43) were male. The majority of respondents rated the app quality as very high (n=27, 79%), felt that the app was easy to use (n=20, 61%), easily understood (n=18, 55%), improved their mental fitness (n=27, 80%), and would recommend the app to others (n=20, 61%). Encouraging trends toward improvement were found across symptom and well-being outcomes; however, these trends were not significant: general distress (t_{32}=0.65, P=.52), depression (t_{32}=0.75, P=.46), anxiety (t_{32}=1.08, P=.29), or traumatic stress coping (t_{32}=−0.27, P=.79), with effect sizes ≤0.2, likely due to the small sample size.

Conclusions: The Build Back Better app demonstrated satisfactory levels of usability and acceptability. While the pilot study showed encouraging trends toward improved MH, further research with a larger sample size is needed to determine its efficacy. Participants furthermore suggested improvements in app navigation and content clarity, emphasizing the need for a more intuitive user experience. Given the positive feedback and improvement in MH outcomes, a larger-scale efficacy trial is warranted to further assess the app's potential for MH support in this high-risk population.

Trial Registration: New Zealand Clinical Trials Registry ACTRN12621001006831; https://anzctr.org.au/Trial/Registration/TrialReview.aspx?id=382290

(JMIR Form Res 2025;9:e50995) doi:10.2196/50995

KEYWORDS

digital mental health; emergency service workers; high risk professions; prevention; psychological distress; smartphone applications; user-engagement

XSL•FO
RenderX

Introduction

Background

Emergency service workers and volunteers are those workers (both paid and volunteer) who attend scenes of emergencies and disasters to minimize risk to community safety and security. These include (but are not limited to) police, paramedics, firefighters, rescue, and emergency service personnel. They perform an indispensable function in the community; however, the nature of this work can come at a cost to their mental health (MH) and well-being [1]. These workers are exposed to elevated rates of stress, trauma, and adversity on a regular basis, putting them at increased risk for a range of MH conditions, including, but not limited to, depression, anxiety, drug and alcohol-related disorders, and posttraumatic stress disorder (PTSD) [2-5].

The question of what measures may be able to enhance or protect MH in emergency service workers is garnering growing interest internationally. A recent survey by Kyron et al [6] among 14,868 Australian ambulance, fire and rescue, police, and state emergency services employees highlights the range of MH and well-being conditions associated with work in the emergency services sector. While 10% of employees screened positive for probable PTSD, 30% had low well-being, and 30% had high or very high psychological distress, suggesting that depression and anxiety, as well as PTSD, are likely to be common MH issues for emergency service workers. Similar results were found in a systematic review of 27 international studies [7] reporting on 30,878 ambulance personnel, which found estimated prevalence rates of 11% for PTSD, 15% for depression, 15% for anxiety, and 27% for general psychological distress. In a study of 5813 Canadian safety personnel (correctional workers, dispatchers, firefighters, paramedics, and police officers), Carleton et al [8] found that 44.5% screened positive for clinically significant symptoms of one or more diagnosable mental disorders, approximately 4 times higher than rates of diagnosis for the general population. Of concern, these rates were noted to be higher than those reported in previously conducted studies in the area, suggesting that mental disorders may be increasing among emergency service workers [8-10].

Despite the high prevalence of MH symptoms, significant barriers continue to hinder and delay help-seeking among emergency service workers, including stigma, perceived unavailability of resources, the need for self-reliance in solving problems, and fear about the intervention or treatment [11]. Emergency service workers' work constraints are also different from many other professions, which can influence seeking help for MH problems even more. For example, many emergency service workers are often "on duty" for 24 to 72 hours, thereby having to respond to unpredictable emergencies, which supports the importance of tailored interventions, accessible at any time [12,13]. Their culture is different as well; when emergency service workers return home, they are expected to "function" in daily living, such as work and family life, which increases the likelihood that any MH symptoms may be ignored. Finally, emergency service workers are reluctant to approach mainstream MH services because of fear that awareness by their organization of MH problems may negatively impact certain privileges, such as promotion, deployment to select units, or the right to carry a firearm, and so on. Taken together, these factors point to the importance of considering emergency service workers' perspectives and needs when devising MH interventions and the necessity of developing novel MH approaches with widespread reach.

Mobile phone–based health initiatives have become an increasingly popular way to provide MH care and support to individuals [14]. The increased use of technology, alongside the widespread treatment gap, has led researchers to explore the use of digital platforms for MH engagement. Smartphone apps are considered a suitable prevention approach, since they have the potential to deliver effective psychological interventions via frequent brief exchanges, offering a unique opportunity for disorder prevention and symptom management [15,16]. The multiple features of mobile phones furthermore offer great opportunities for the delivery of health promotion interventions in different formats, thereby having the potential to overcome traditional service barriers [17], especially in difficult-to-access populations, such as emergency service workers. Apps enable flexible user engagement, allowing individuals to access interventions at their convenience, in private, and at a time and location of choice, thereby facilitating real-time monitoring of mood and behavior [18]. Meta-analyses of randomized controlled trials have shown promising results for evidence-based MH apps in reducing depressive symptoms, stress, anxiety, and substance use [19,20]. Moreover, MH applications can successfully use persuasive strategies, supporting people with MH issues to adopt healthy lifestyles and improving user-oriented behavior changes [17,21].

Prior to efficacy testing, a critical aspect in the development of mobile phone apps is exploring the acceptability and feasibility of the interventions [22]. Engaging the target audience in the process of development is essential for developing optimum treatments in specific settings and is likely to contribute to the wider acceptability and utility of digital MH interventions [22,23].

Objectives

This pilot study aims to explore the use, acceptability, and preferences of emergency service workers regarding potential uptake of a smartphone-based MH intervention designed to provide evidence-based MH prevention strategies, the Build Back Better app. As well as assessing acceptability and feasibility, this study also aims to assess the impact of the Build Back Better app on depression, anxiety, well-being, and other symptoms of common mental disorders.

Methods

Study Design and Setting

The study was conducted with emergency service workers from 5 different emergency and rescue services organizations across Australia. Participants were recruited via organizational advertisements, which included QR codes or links to the study website. Participants were recruited to test usability, acceptability, feasibility, and preliminary MH effects of a

XSL•FO
RenderX

smartphone-based MH intervention: the Build Back Better app. Data were collected via pre- and postintervention questionnaires and in-app data collection, with participants using the app remotely. Recruitment started on August 16, 2021, and was finalized on December 18, 2021. Data collection of the trial was completed by January 2022.

Ethical Considerations

The study was granted ethical approval by the University of New South Wales Human Research Ethics Committee (HC210399). Informed consent was collected electronically before participation. This approval covers secondary analysis without additional consent. Participants were compensated with an Aus $40 (US $26) gift voucher for completing the post-assessment and qualitative interview. Alternatively, participants could opt for an Aus $40 (US $26) donation to a charity if their organization had gift restrictions. The pilot study was registered in the Australian New Zealand Clinical Trials Registry (ACTRN12621001006831p), and the app was notified with the Therapeutic Goods Administration through the Clinical Trial Notification scheme (CT-2021-CTN-02958 - 1). Study data were deidentified to ensure confidentiality and participant privacy.

Participants

Eligibility

Eligible participants were Australian residents, aged 18 years and older, currently working, or previously (ie, in the last 2 years), as an emergency service worker (volunteer or paid). Retired emergency service workers (ie, emergency service workers who have worked in the last 2 years) were also deemed eligible to participate due to both the delayed onset of many trauma-related symptoms and the fact that such symptoms may have been implicated in a decision to retire [1]. Other inclusion criteria included the ability to speak and understand English, willingness and ability to provide informed consent, and access to an Android or iOS smartphone. Participants were also required to have a current email address or be willing to obtain one.

Recruitment

Participants were recruited via organizational advertisements calling for emergency service workers to test a new MH app. QR codes or links on the advertisements directed interested individuals to the study website, which included more detailed information about the study procedures and inclusion criteria.

All interested participants were required to provide informed consent electronically. Baseline assessment occurred immediately after participant consent. Upon completion of the baseline assessment, participants were prompted to download Build Back Better from either the App Store (for iPhone users) or the Google Play Store (for Android users) and use the app for approximately 30 days. Postassessment occurred 1 month after the completion of the baseline assessment. After the completion of the postassessment surveys, participants were invited to participate in a qualitative interview to ascertain in-depth information on app use and user preferences (also see Qualitative Data Collection section for details). See Table 1 for an overview of the different measures administered at baseline and 1-month follow-up. All assessments were administered online using Black Dog Institute's Online Research Platform.

Table . Overview of study measures and data collection time points.

Instrument	Domain	Time points		
		Baseline	Postassessment[a]	In-app data collection[b]
Demographic Inventory	Demographics	✓		
Kessler Psychological Distress Scale	Psychological distress	✓	✓	
9-item Patient Health Questionnaire	Depression	✓	✓	
7-item General Anxiety Disorder	Anxiety	✓	✓	
5-item World Health Organization Well-Being Index	Well-being	✓	✓	
Trauma Coping Self-Efficacy	Trauma coping	✓	✓	
App engagement (survey and qualitative interviews)	App feasibility and acceptability		✓	
Objective app engagement outcomes	App feasibility and acceptability			✓

[a]Postassessment administered 1 month after baseline assessment.

[b]In-app data collection throughout the entire duration of study (ie, from baseline to 1-month follow-up).

XSL•FO
RenderX

Intervention

Build Back Better was developed by a multidisciplinary team with content adapted from a previously developed app, HeadGear, designed to prevent depression and anxiety and improve well-being among workers in male-dominated industries [18,22]. A randomized controlled trial of HeadGear with a large sample of workers showed that the app was effective in the prevention of depression and anxiety, as well as improving a range of MH and work-related outcomes [24,25]. While many other MH apps offer generic advice and often fail to offer personalized care, the Build Back Better app was carefully developed through extensive collaboration with clinical psychologists, psychiatrists, IT professionals, user experience and app design experts, and, most importantly, emergency service workers themselves. This approach ensures that the app content and design are tailored to the specific needs of emergency service workers and their families. By integrating trauma-focused content alongside strategies for addressing depression and anxiety, the app tailors its support based on the preferences and needs expressed by emergency service workers, leading to better, more effective outcomes [26]. Personalized changes to the app visual presentation and design included a more gender-neutral look and feel (achieved through color scheme, imagery, language, and theme) and streamlined features to promote use in rural populations with unreliable internet access or phone service (ie, videos or activities that can be accessed without an internet connection) and those with limited mobile phone digital literacy. In addition, changes ensuring that the app's structure would be suitable for those performing shift work were included, by allowing day-to-day use at all times, and whenever desired. For example, whereas previous apps developed by the research team [18,22] present app content in a linear form to be completed on consecutive days, the Build Back Better app content is tailored to the individual needs, via an upfront needs analysis, and through an in-app search function, maximizing user control and autonomy [27]. As it is suggested that this target population typically uses apps in their own time without clinical oversight, they must be intrinsically motivated to engage with the app [26,28]. Therefore, use flexibility was maximized by allowing the user to complete tasks freely within the app, according to their priorities and preferences. To further target the needs of emergency service workers, trauma-informed content was developed specifically for the app through consultation with clinical psychologists working primarily with emergency service workers. This clinical content incorporates thought monitoring or challenging, behavioral activation, problem solving, mindfulness, elements of positive psychology (eg, gratitude), and techniques to enhance healthy coping behaviors. Other components of the Build Back Better app include a tracker for monitoring mood, physical activity, sleep, eating habits, and work and life balance, and a variety of breathing and grounding exercises, based on the expressed preferences of emergency service workers. See Figure 1 for screenshots of the Build Back Better app.

Figure 1. Screenshots of key features of the Build Back Better app.

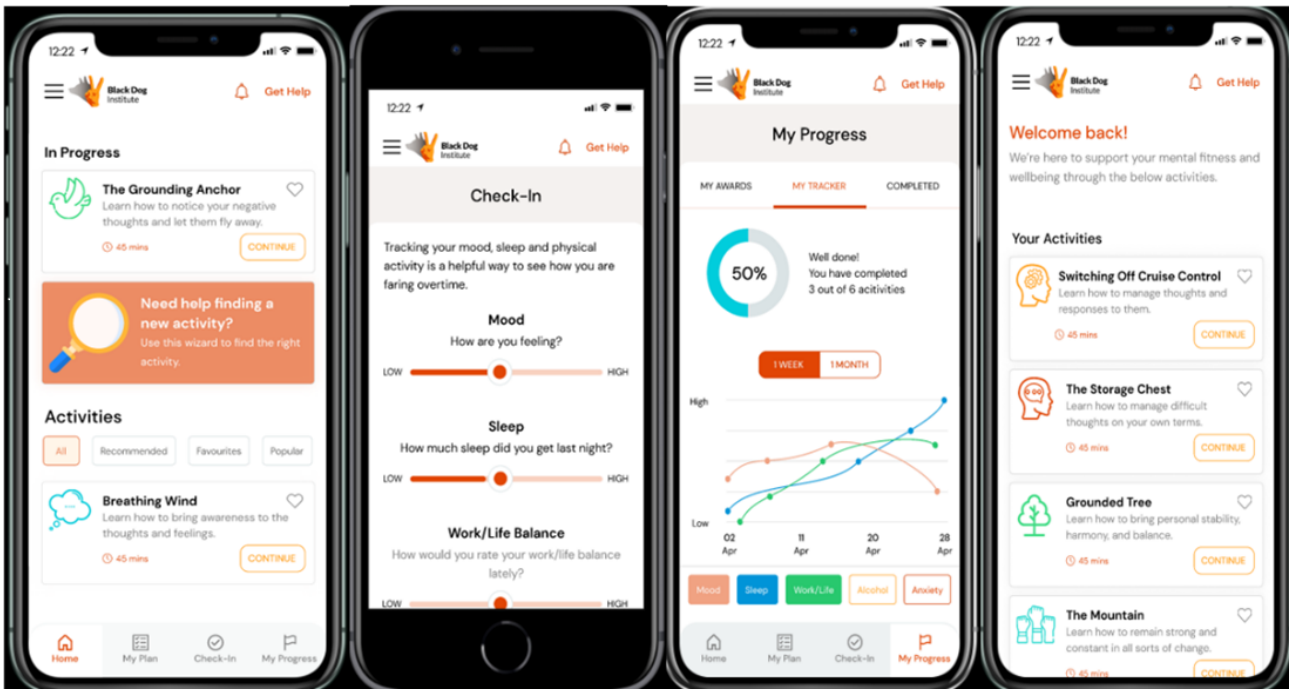

Data Collection

At baseline, participants were asked questions regarding demographics, MH illness in the previous 2 years, help-seeking behaviors for an MH problem in the previous month (from any source including health professionals, telephone or online support services, family, partner, or friends), and whether they were currently taking medication for an MH issue.

App usability, feasibility, and acceptability were measured using questions adapted from the Mobile Application Rating Scale (MARS; Stoyanov et al [29]) at postassessment, including questions pertaining to current smartphone use, potential future use, and preferences for design and content. The MARS is a simple, objective, and reliable tool for classifying and assessing the quality of mobile health apps [29]. The app quality criteria were clustered within the engagement, functionality, aesthetics,

XSL•FO
RenderX

information quality, and subjective quality categories, from which 7 individual MARS items were included. Each MARS item used a 5-point scale ("not at all" to "completely" or "definitely"). In addition, 4 open-ended questions were administered, allowing the user to express their opinion about the app, in their own words. In addition, those agreeing to take part in the qualitative semistructured interview were asked a series of open-ended questions pertaining broadly to functionality, engagement, aesthetics, information exchange, and the user's overall experience with the app to further explore app usability and acceptance. To monitor app usage, in-app data was automatically collected for the duration of the study (ie, from baseline to 1-month follow-up), such as number of times app opened, number of activities accessed, number of activities completed, and total time spent in app.

Psychological distress was measured using the Kessler Psychological Distress Scale (K10) [30]. The K10 is a 10-item measure of psychological distress over the past 4 weeks and is intended to yield a global measure of distress based on questions about anxiety and depressive symptoms. The K10 uses a 5-value response option for each question, which can be scored from 5 ("all of the time") through 1 ("none of the time"). The total K10 score is based on the sum of the answers given to the 10 questions and was used to indicate psychological distress across all study time points. The maximum score is 50, indicating severe distress; the minimum score is 10, indicating no distress.

Depressive symptoms were measured using the 9-item Patient Health Questionnaire (PHQ-9) ([31]. The PHQ-9 is a reliable and valid 9-item self-report tool designed to assess depression severity. The items duplicate the 9 diagnostic criteria for major depressive disorder covered in the *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition* (DSM-IV). The PHQ-9 asks how often participants have been bothered by problems in the past 2 weeks. Each item is scored on a 4-point Likert scale, ranging from 0 ("not at all") to 3 ("nearly every day"). Items are summed to provide a total score (ranging from 0 to 27), which was used as a marker of depressive symptom severity across all study time points. Scores ≤4 suggest minimal depression, which may not require treatment, whereas scores of ≥10 are indicative of depression.

Anxiety symptoms were measured using the 7-item Generalized Anxiety Disorder (GAD-7) [32], a reliable and valid 7-item measure of generalized anxiety symptoms [31]. GAD-7 scores can range from 0 to 27, with 5, 10, and 15 representing cutoffs for mild, moderate, and severe levels of anxiety, respectively. A score of 10 or greater on the GAD-7 represents a reasonable cut point for identifying cases of GAD.

Participant well-being was measured using the 5-item World Health Organization Well-Being Index (WHO-5) [33]. The WHO-5 is a short, self-administered measure of well-being over the last 2 weeks. It consists of 5 positively worded items that are rated on a 6-point Likert scale, ranging from 0 ("at no time") to 5 ("all of the time"). The raw scores were transformed to a score from 0 to 100, with lower scores indicating worse well-being. A score of ≤50 indicates poor well-being and suggests further investigation into possible symptoms of depression. A score of 28 or below is indicative of depression.

Perceived self-efficacy for coping with challenges and threats was measured using the 9-item Trauma Coping Self-Efficacy Scale (CSE-T) [34]. Participants were asked to rate their capability to handle a series of posttraumatic situations (eg, "Deal with my emotions [anger, sadness, depression, anxiety] since the traumatic event," "Manage distressing dreams or images about the traumatic experience"), on a 7-point scale ranging from 1 ("not at all capable") to 7 ("totally capable"). Total CSE-T scores (ranging between 1 and 63) were created by summing the item ratings across all time points to analyze whether CSE-T increased as participants continued to use the application (ie, reflecting improvement in self-efficacy perceptions).

Qualitative Data Collection

After the completion of the postassessment surveys, participants were invited to participate in a qualitative interview to ascertain in-depth information on app use and user preferences. Interviews (approximately 1 hour) occurred within 2 weeks of postassessment and were conducted virtually (ie, using telephone or videoconference).

Participation in postassessment surveys and the qualitative interviews was entirely voluntary and did not affect the use of the app. Users could engage with the app as they wish, regardless of their involvement in the postassessment surveys and qualitative interviews. Participants were, however, offered an Aus $40 (US $26) gift voucher as time reimbursement upon completion of the postassessments and the qualitative interview; if this was against the gift policy of their organization, they could select an Aus $40 (US $26) donation to a charity of their choice.

Sample Size

Based on earlier comparable (pilot) studies using eHealth interventions [22,35,36], it was estimated that a sample size of 63 was required to detect a small within-group effect difference (Cohen d=0.3) with 80% power and a 0.05 significance level. With the expectation that 30% (n=20) of participants will be ineligible, lost to follow-up, drop-outs, or noncompliant, at least 90 participants would need to be enrolled.

Data Analysis

Analyses were both quantitative and qualitative. Study investigators examined descriptive characteristics of the study sample and overall feasibility and acceptability of the smartphone app within the pilot study. Potential efficacy of the app in reducing psychological distress was assessed by evaluating changes in clinical outcomes after 1 month of use in terms of both statistically significant changes as well as effect size estimates. Changes in participants' well-being and MH levels from baseline to posttest were examined using paired sample t tests. Qualitative data collected through semistructured interviews were examined to understand user needs, experience, satisfaction, and expectations of the app. Study investigators extracted themes based on participants' responses to understand participants' acceptability of the study design and overall experience with the smartphone app.

XSL•FO
RenderX

Results

Participants Characteristics

A total of 71 participants enrolled in the study, of which 67 both completed the baseline assessment and downloaded the app (see Figure 2 for the CONSORT flow diagram of this single-arm study). The baseline characteristics of the sample (N=67) are presented in Table 2. The mean age of the participants was 44.73 (SD 11.40; range 29 - 66) years, with the majority being male (n=43, 64%). Participants included firefighters, disaster response, surf lifesaving, police, and correctional services across 4 Australian states: New South Wales (n=12), South Australia (n=5), Tasmania (n=17), and Western Australia (n=33). The sample was largely drawn from major cities (>60%), and approximately half were volunteer emergency service workers (48%). More than 80% reported an episode of poor MH lasting at least 1 month within the last 4 weeks, with 44% having sought some form of MH support in the previous month (eg, from a general practitioner, MH professional, telephone or online support services, family, partner, or friends), and 16% (n=11) currently using medication for an MH issue.

Although the majority of the sample (n=41, 62%) reported experiencing no current distress (ie, a score under 20) on the K10, more than a quarter (n=18, 27%) reported either moderate or severe distress (ie, a score of 25 or higher) on the K10 at baseline (refer to Table 2). Around half of the participants (n=35, 52.2%) reported minimal anxiety symptoms (ie, a score of 4 or under), whereas 45% (n=30) scored between 15 and 24 on the GAD-7, indicative of mild to moderate anxiety symptoms. A total of 3% (n=2) of the participants reported severe anxiety symptoms (ie, a score >15 on the GAD-7). Prevalence of depression symptoms (ie, a score of ≥10 on the PHQ-9) was reported in 24% (n=16), and minimal to mild depression severity (ie, a score of <10 on the PHQ-9) was reported in 76% (n=51) of the participants.

Of the 67 app users who completed the baseline assessment, 51% (n=34) failed to complete the follow-up questionnaire, resulting in complete follow-up data for 33 users. The only significant difference between postcompleters and noncompleters was sex, with males being less likely to complete the postassessment. Noncompleters reported worse MH on all measures at baseline, although none of these differences reached statistical significance.

XSL•FO
RenderX

Figure 2. CONSORT (Consolidated Standards of Reporting Trials) flow diagram illustrating the study procedures. GP: general practitioner; I/E: inclusion/exclusion.

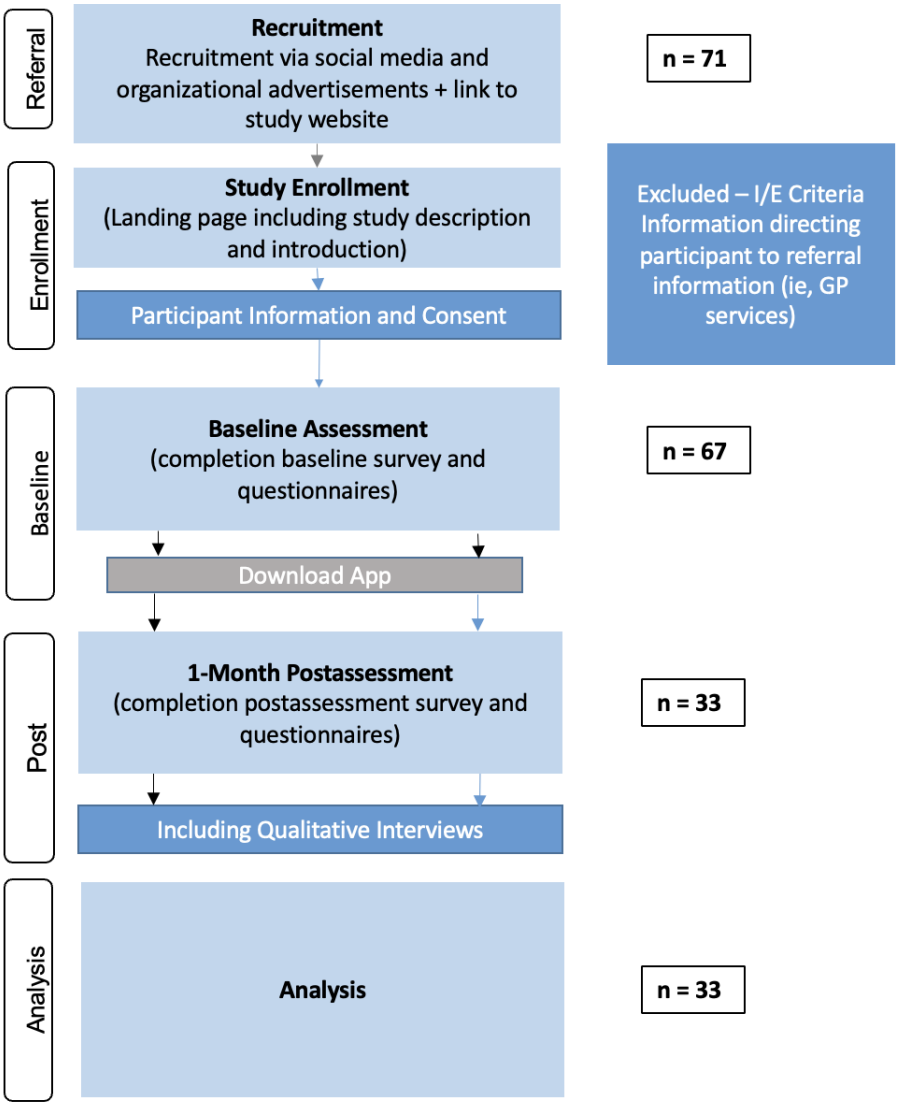

XSL•FO
RenderX

Table . Sociodemographic and clinical characteristics of study participants.

Demographics	Values
Age (years)	
Mean (SD)	44.73 (11.4)
Range	29-66
Male, n (%)	43 (64.2)
Highest level of education, n (%)	
University	13 (19.4)
Postgraduate	9 (13.4)
Non-university diploma/college	14 (20.9)
Trade or other certificate	17 (25.4)
Year 12 certificate	11 (16.4)
Year 10 certificate	2 (3)
Other	1 (1.5)
State, n (%)	
New South Wales	12 (17.9)
South Australia	5 (7.5)
Tasmania	17 (25.4)
Western Australia	33 (49.3)
Agency, n (%)	
Airservices Australia[a]	2 (3)
DBCA[b]	3 (4.5)
DFES[c]	30 (44.8)
DPFEM[d]	16 (23.9)
RFS[e]	4 (6)
SAFECOM[f]	5 (7.5)
SLSA[g]	7 (10.4)
Emergency service worker - volunteer, n (%)	32 (47.8)
Location, n (%)	
Metropolitan	44 (65.7)
Rural	11 (16.4)
Remote	12 (17.9)
General distress (K10)[h], n (%)	
No distress	41 (61.2)
Mild distress	8 (11.9)
Moderate distress	9 (13.4)
Severe distress	9 (13.4)

[a]Airservices: Aviation Rescue Fire Fighting Service - Airservices Australia (New South Wales).
[b]DBCA: Department of Biodiversity, Conservation and Attractions (Western Australia).
[c]DFES: Department of Fire and Emergency Services (Western Australia).
[d]DPFEM: Department of Police, Fire & Emergency Management (Tasmania).
[e]RFS: New South Wales Rural Fire Service (New South Wales).
[f]SAFECOM: South Australian Fire and Emergency Services Commission (South Australia).

XSL•FO
RenderX

[g]SLSA: Surf Life Saving Australia (New South Wales).

[h]K10: Kessler Psychological Distress Scale.

Intervention Delivery

The frequency of app usage ranged from once to 22 days (mean 5.5, SD 5.62), with usage averaging 25 minutes per session. In-app activities related to trauma, recognizing poor MH, identifying thought patterns, breathing, and mindfulness exercises were the most accessed by users. The daily check-in feature (allowing users to rate their mood or behaviors and displaying a weekly and monthly graph of their past mood check-ins) was the most popular app feature, being accessed by 43% (n=15) of participants during the trial period. On average, after opening an activity within the app, 67% (n=22) of participants completed the full activity content.

Figure 3 presents the results for the different acceptability variables. At 1-month follow-up, most of the participants (n=29, 88%) felt that the Build Back Better app had at least "somewhat" improved their mental fitness. The majority of participants (n=27, 79%) rated the overall quality of the app as "very high" or "high," with a mean score of 4 out of 5. Although more than half of the participants (n=18, 55%) claimed that they "completely" understood the app content, 30% (n=10) indicated the content of the app was "somewhat well" understood. The majority of participants (n=20, 61%) reported that using the app was very easy (with a mean rating of 4 out of 5), and that they would "definitely" be willing to recommend the app to others. Around half (n=17, 52%) indicated that the app content (visual information, language, and design) was appropriate, whereas one-third (n=11, 33%) indicated the app content was "somewhat" appropriate for them. In terms of app interest and engagement, the participant's response was somewhat mixed. A total of 36% (n=12) of the participants indicated the app was "completely or highly" engaging, whereas 42% (n=14) of participants rated the app as "somewhat interesting."

Figure 3. Participant ratings of acceptability variables using a 5-star rating scale. The specific meanings of the 5-star rating scale labels may vary depending on the item, but in general, 5 points represent excellent or outstanding performance, while 1 point represents a poor experience or performance.

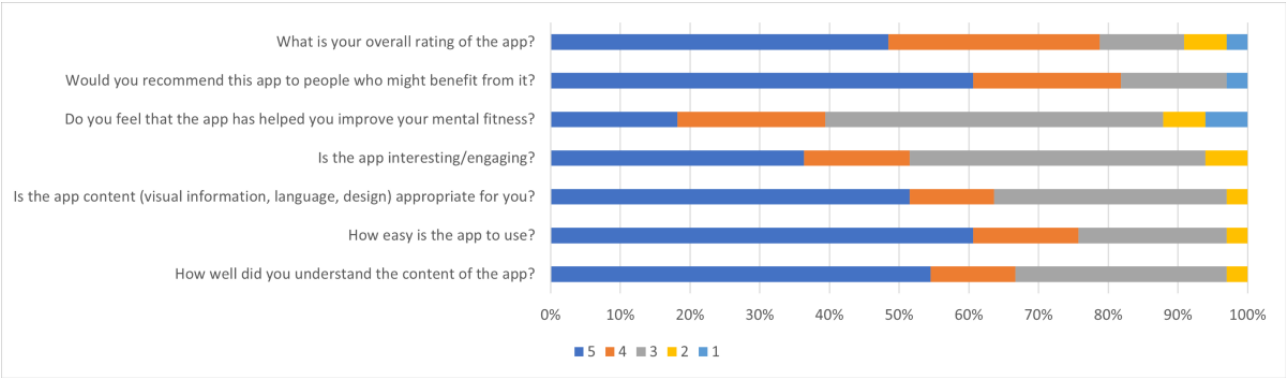

Qualitative Data

Open feedback on the app from 33 participants who completed their 1-month follow-up assessment was generally positive and encouraging:

> *It is a great app, a very useful tool.*
>
> *Really good work, thank you.*
>
> *Well done and I hope it helps many people along the way.*

Users commented on the presentation of app content and features. Overall, they appreciated the mixture of audio, visual, or text within the different in-app activities and liked having a variety of activities and interactive components:

> *For me it was especially helpful the components with video and audio. I do not like the components with reading as I am trying to minimise my screen time.*
>
> *It's very interactive which is a great feature to get focus back on track.*

In terms of the usefulness of specific therapeutic elements within the app, the relaxation practices, such as breathing or meditation, were particularly well received. Users indicated that these activities assisted them with stress reduction and anxiety relief, and experienced these activities to be beneficial for their overall well-being and joy:

> *This a great tool for anyone who is suffering any kind of anxiety and doesn't feel they can approach anyone with their problems.*

With respect to the thematic analysis, we further classified feedback into three categories: (1) qualities or features users liked, (2) features requested or missed by users, and (3) weaknesses of the app identified by users. These 3 categories helped us further understand how each component of the app helped (or could be improved) to meet user needs and promote adherence and acceptability.

Several app users (7/33, 21%) indicated they liked the daily check-in feature; it was suggested that this feature was very helpful for self-management of symptoms:

> *I liked the daily check-in and to see how my two selected categories were tracked; mood and sleep. It was interesting to see the correlation between the two.*
>
> *I like how the app asked how you are currently feeling, and then displayed content relevant to your feelings that you could do within the app.*

XSL•FO
RenderX

Around a quarter of users (8/33, 24%) provided suggestions to further improve the app, referring to specific app features; participants emphasized the importance of using "real-life" characters (ie, characters with lived experience) throughout the app, and more audio and visual elements to stimulate engagement with the app. Accurate presentation of time estimates for each of the activities was also recommended. The main weakness of the app, identified by users (8/33, 24%), related to difficulties navigating the app:

> *Too many steps to find results.*
>
> *Not very much direction in how to go through the app.*
>
> *Was a little scattered to navigate.*

Others (5/33, 15%) highlighted problems with user engagement:

> *Sometimes getting back into a repeat activity was tedious.*
>
> *Maybe a little too simple, I tended to forget about using it.*
>
> *Could be more interactive.*
>
> *The repetitive nature of all the activities. It got a bit boring and started feeling less impactful.*

Qualitative Interviews

In total, 8 out of the 33 participants (24%) who completed both the baseline and follow-up assessment agreed to participate in a semistructured interview. During the qualitative interviews, participants were asked to reflect on their experiences with the app and provide more detailed feedback on certain features of the app and its overall design and functionality. Participants felt that the app was very easy to use and provided relevant information for emergency service workers who are feeling overwhelmed and need some help managing stress or anxiety:

> *I thought the app content was well balanced and covered a whole range of different issues and problems that people like me might helpful*
>
> *For everyone who has a stressful job or stressful home life or school life or just life in general I think the app is really relevant, it's really suitable for pretty much everybody across the board in our communities.*
>
> *Of great use for first responders and community members who don't have access to the mental health resources of the agencies.*

Participants also commented on the quality and credibility of the app, referring to the content as being of high standard, easy to use, and very educational. Participants furthermore felt that the app increased their awareness of MH issues not only within themselves, but also regarding work colleagues or employees:

> *I highly recommend using the app for its insight into mood tracking. It showed me the relationship between anxiety and sleeping and the effect that it has a mental health when you don't know unless you track it.*
>
> *It makes it easier to start that conversation when thinking about suicide, gives you some ideas how to start that conversation:*
>
> *I'm not feeling so great, Tom, can I chat with you.*

Participants suggested some minor design and content changes:

> *There were too many layers. There were too many pages, sometimes that confused me.*

Participants recommended including additional reminders and adjunct components to help users stay engaged with the app:

> *Maybe come up with a reminder, on your phone saying don't forget to finish this. You know that you started yesterday or the day before just something like that.*

Further suggestions for improvement by emergency service workers app users were therefore largely focused on elements of user interface and user experience, such as a more intuitive home screen and changes to the background color within the app to reduce confusion around navigation and increase clarity in the presentation of the different activities.

Symptom Change

Table 3 presents the means of the outcomes at both baseline and 1-month follow-up. Overall, at 1-month follow-up, there was a nonsignificant trend for improvement in general distress (t_{32}=0.65; P=.52), depressive symptoms (t_{32}=0.75; P=.46), and anxiety symptoms (t_{32}=1.08; P=.29). This same nonsignificant trend was present in participants' ability to manage traumatic stress (t_{32}=−0.27; P=.79). Effect size estimates from baseline to 1-month follow-up were ≤0.2 for all outcomes.

XSL•FO
RenderX

Table . Changes in general distress, depression, anxiety, well-being, and traumatic stress coping outcome scores: mean difference, effect sizes, and standardized mean differences from baseline to 1-month follow-up.

	Baseline, mean (SD)	Postassessment[a], mean (SD)	*t* test (*df*)	*P* value	Effect size (95% CI)	Standardized mean difference
K10[b] (general distress)	18.67 (6.63)	18.18 (5.78)	0.65 (32)	.52	0.08 (−0.41 to 0.56)	0.48
PHQ-9[c] (depression)	5.52 (4.14)	5.12 (3.98)	0.75 (32)	.46	0.10 (−0.39 to 0.58)	0.39
GAD-7[d] (anxiety)	5.21 (4.38)	4.49 (4.20)	1.08 (32)	.29	0.20 (−0.25 to 0.58)	0.87
WHO-5[e] (well-being)	56.97 (21.89)	56.12 (24.67)	0.28 (32)	.78	−0.06 (−0.55 to 0.42)	−1.55
CSE-T[f] (traumatic stress coping)	47.82 (11.07)	48.12 (11.60)	−0.27 (32)	.79	−0.12 (−0.60 to 0.37)	−1.33

[a]Postassessment administered 1 month after baseline assessment.

[b]K10: Kessler Psychological Distress Scale.

[c]PHQ-9: 9-item Patient Health Questionnaire.

[d]GAD-7: 7-item Generalized Anxiety Disorder.

[e]WHO-5: 5-item World Health Organization Well-Being Index.

[f]CSE-T: Trauma Coping Self-Efficacy Scale.

Discussion

Principal Findings

The study evaluated the usability, feasibility, acceptability, and preliminary effectiveness of Build Back Better, an app-based MH intervention aimed at enhancing emotional well-being and positive MH among emergency service workers and volunteers.

The results from the postassessment surveys, in-app feedback, and qualitative interviews indicated that the Build Back Better app was found to have satisfactory levels of usability and acceptability. Participants indicated that the app content was relevant for emergency service workers and felt it was beneficial in increasing their capacity to manage stress and anxiety and in improving their general well-being. Users also appreciated the variety of features, functionality, and content within the app. The overall rating of the app was high to very high, with the majority of respondents indicating the app was easy to use, appropriate, and recommendable to others. These findings suggest that the app and its components are likely to meet emergency service workers' expectations and requirements. However, qualitative analyses provided a more nuanced understanding of the perceived acceptability and usability of the app. Despite positive feedback on the user experience of the app, navigation concerns were raised that require design changes to enhance user experience and flow within the app. In addition, although most respondents understood the app content mostly or completely, a sizable proportion reported they only "somewhat" understood the content. Coupled with these navigation issues, both self-report feedback and objective app data highlighted issues pertaining to engagement and functionality. Not only was the mean total number of activities accessed within the app flow, but also questionnaire data suggested that half of the respondents found the app at best "somewhat" engaging. The 51% dropout rate for postsurvey completion may also be indicative of a lack of engagement with the app. Even though it is well-established that interventions delivered via MH apps are susceptible to low levels of engagement and poor adherence by their users [37-39], this can have implications for the effectiveness of the intervention delivered [40-42].

A secondary goal of this study was to determine the preliminary effectiveness of the Build Back Better app in reducing general distress and supporting (mental) well-being of emergency service workers. Although encouraging trends toward improvement in most symptom and well-being outcomes were found, these failed to meet significance. This raises vital questions that need to be addressed before larger testing. Considering the overall low symptom levels at baseline, floor effects are a possible implication of this lack of significance. However, as mentioned above, high rates of trial attrition left the trial underpowered, which would have also contributed to the risk of type 2 error. Furthermore, a related factor that requires attention is low program adherence. Despite the short length of the trial and wide variety of available exercises within the app, engagement was low, with the app being used, on average, 5 to 6 times for the duration of the 1-month trial. Surprisingly, adherence was considerably lower than that observed in previous trials of similar apps [18,24,35,36]. This raises questions about both the app content and the free-form structure of the app. Compared to previous trials in which the app content was delivered in a linear fashion (introducing one new activity at a time), users of this app had greater choice and were able to access recommended content on demand. Generally, such an individualized approach using personally tailored content is viewed as a positive means to enhance engagement in digital interventions [43,44], but in this study, it may have left users feeling overwhelmed and undirected, as from the qualitative data findings, some participants reported feeling overwhelmed or confused by the lack of direction in

XSL•FO
RenderX

how to navigate the app content. Previous research has shown that providing users with more choices and actions leads to increased cognitive strain, and this may negatively affect acceptability [41,45,46]. Further research should investigate which activities users prefer to do without those activities becoming burdensome or overwhelming, while also creating an app that allows independence and control.

Limitations

There are several limitations of this study, which should also be acknowledged. Emergency service workers are an at-risk population due to high rates of trauma exposure and related illness [47]. Although the self-reported ability of participants to manage traumatic stress was captured for the duration of the trial, trauma exposure or levels of traumatic distress were not examined in the current sample, and there were no restrictions on eligibility related to these factors. Overall, the sample reported very few symptoms of MH illness. Interestingly, despite low baseline levels of anxiety, depression, and psychological distress, most participants also scored in the lower percentile of well-being scores. Regardless, not only did this potentially impact capacity for symptom change, but it also limited comparison of outcomes based on severity. Such a program is likely to have a differing effect on individuals experiencing differing levels of psychological well-being at baseline, possibly because there are considerable between-individual differences in the response to these interventions [48,49]. The Build Back Better app itself was designed as an MH promotion and selective prevention app, directed toward at-risk groups or individuals with physical, psychological, or social risk factors associated with mental illness development, but this study lacked the power to adequately test this.

Another limitation of this study was the high attrition rate, with nearly half of the participants failing to complete their follow-up assessments after 30 days. Although high attrition is common in unguided trials [50-52], nonrespondents in this study may have been affected by digital access issues. Of the study sample, 34.3% lived in regional or remote areas, and these rural residents are still less likely than those living in suburban areas to have home broadband and could have faced challenges to accessing the internet and completing their follow-up surveys. While the Build Back Better app was designed to allow the user to run the app in different conditions and locations, a lack of internet connectivity could have prevented users from completing the online follow-up surveys via Black Dog Institute's Online Research Platform.

Future Directions

Further investigation is needed to determine the potential protective benefits of the Build Back Better app under more controlled conditions, in light of current knowledge on between-individual differences in how individuals respond to interventions. In addition to a larger, appropriately powered controlled trial, future evaluation should consider baseline cut-offs for symptom severity, which may influence the effectiveness of the app. This study used a heterogeneous sample of emergency service workers, which may have impacted findings. While this diversity was seen as a strength of the study, and the app was designed to cater to a range of emergency service workers, there is the potential that the content may have been received differently across professions. This suggests that a more tailored version of the app, designed specifically for different professions or roles, could improve engagement and content relevance.

We recognize that the lower participation rate in the follow-up questionnaires and qualitative interviews is a limitation of this study. Although the interviews were structured with thoughtful, targeted questions designed to elicit rich, in-depth responses, the low response rate limits the representativeness and diversity of the perspectives gathered. Similarly, the high dropout rate from the follow-up assessments reduces the robustness of our conclusions and the generalizability of our findings. Further analysis comparing postcompleters and noncompleters revealed that the only significant difference between these groups was gender, with males being less likely to complete the postassessment. Dropout was not strongly associated with baseline symptom severity or well-being; however, additional research with a larger sample size would help clarify these patterns and improve the reliability of the findings.

The high attrition rate, furthermore, highlights the need for future studies to address potential barriers to engagement, such as digital access issues. Identifying and mitigating these barriers through improved app design or more effective follow-up communication strategies could provide more reliable data on the effectiveness of mobile phone–based health initiative interventions in high-risk populations. For example, testing different engagement strategies, such as varying the frequency of reminders or using multiple forms of communication (eg, SMS text messages and phone calls), could help improve follow-up rates. The protocol for this trial involved 2 email reminders for follow-up assessment. In future evaluations of MH apps, consideration is needed regarding other reminder options. SMS text message reminders and phone calls might be a more reliable form of contact for this group, considering their location and the nature of many emergency service workers' work. Moreover, varying frequency of email reminders and varying content could help engage participants [53]. Nevertheless, once recruited, participants seemed motivated to start using the app, and the 94% (67/71) baseline completion rate was strong.

Finally, feedback from users in this pilot study revealed several areas for improvement in the app to increase usability and engagement. Participants in this study identified issues particularly around app navigation and user interface, that is, clearer design, short interactions with the app, easy access, but the app users also provided suggestions for improving user engagement, for example, via incorporating reminders to engage with the app, and involving those with lived experience in the activities within the app. This feedback is consistent with observations from reviews of efficacy studies on smartphone interventions, in which apps incorporating more elements aimed at promoting user engagement have shown larger effect sizes on several MH outcomes [42]. Proposed design changes therefore include (1) improved navigation via refined home screen emphasizing the presentation of activities in modules, (2) a more visually engaging home screen, and (3) increased accessibility and understanding of the in-app search function

XSL•FO
RenderX

via clearer instructions and use of different design elements. In addition, future marketing and promotion efforts should emphasize involving those with lived experience to strengthen the app's relevance and appeal. Finally, this study furthermore highlighted the need for more nuanced app usage data to further explore use patterns.

Conclusions

The findings of this pilot study support the usability and acceptability of the Build Back Better app among emergency service workers, complemented by valuable user input that will inform modification of the app design to further enhance the in-app user experience. The app was found to have satisfactory levels of engagement, usability, feasibility, and acceptability. However, evidence of change in MH outcomes could not be gained from this small pilot study. Comparison of the Build Back Better app with a control condition in a larger trial will help to address the limitations of this study and further elucidate (clinical) MH outcome findings. To further advance the field in digital MH options, it is furthermore important for future research to explore which engagement metrics (log-ins, activities completed, time spent in app, etc) could have an impact on improving MH outcomes.

Acknowledgments

The authors would like to acknowledge and thank the Black Dog Institute IT Team, the Curve Tomorrow Team, as well as the User Experience Team at Black Dog Institute for their integral role in the development of the Build Back Better app and the technical implementation of this study. The development of the Build Back Better app and its pilot study was funded by the National Bushfire Recovery Agency as part of the Department of Home Affairs.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author upon reasonable request for noncommercial purposes.

Conflicts of Interest

Authors DM, DOW, and MD were involved in the development of the Build Back Better app. SBH works for the Black Dog Institute, a not for profit research institute that runs the Australian National Emergency Workers Supports (NEWS) Service. There is no other conflict of interest to declare. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Harvey SB, Milligan-Saville JS, Paterson HM, et al. The mental health of fire-fighters: an examination of the impact of repeated trauma exposure. Aust N Z J Psychiatry 2016 Jul;50(7):649-658. [doi: 10.1177/0004867415615217] [Medline: 26607303]
2. Benedek DM, Fullerton C, Ursano RJ. First responders: mental health consequences of natural and human-made disasters for public health and public safety workers. Annu Rev Public Health 2007;28:55-68. [doi: 10.1146/annurev.publhealth.28.021406.144037] [Medline: 17367284]
3. Berger W, Coutinho ESF, Figueira I, et al. Rescuers at risk: a systematic review and meta-regression analysis of the worldwide current prevalence and correlates of PTSD in rescue workers. Soc Psychiatry Psychiatr Epidemiol 2012 Jun;47(6):1001-1011. [doi: 10.1007/s00127-011-0408-2] [Medline: 21681455]
4. Fraess-Phillips A, Wagner S, Harris RL. Firefighters and traumatic stress: a review. Int J Emerg Serv 2017 May 2;6(1):67-80. [doi: 10.1108/IJES-10-2016-0020]
5. Skogstad M, Skorstad M, Lie A, Conradi HS, Heir T, Weisæth L. Work-related post-traumatic stress disorder. Occup Med (Lond) 2013 Apr;63(3):175-182. [doi: 10.1093/occmed/kqt003] [Medline: 23564090]
6. Kyron MJ, Rikkers W, Bartlett J, et al. Mental health and wellbeing of Australian police and emergency services employees. Arch Environ Occup Health 2022;77(4):282-292. [doi: 10.1080/19338244.2021.1893631] [Medline: 33653231]
7. Petrie K, Milligan-Saville J, Gayed A, et al. Prevalence of PTSD and common mental disorders amongst ambulance personnel: a systematic review and meta-analysis. Soc Psychiatry Psychiatr Epidemiol 2018 Sep;53(9):897-909. [doi: 10.1007/s00127-018-1539-5] [Medline: 29869691]
8. Carleton RN, Afifi TO, Turner S, et al. Mental disorder symptoms among public safety personnel in Canada. Can J Psychiatry 2018 Jan;63(1):54-64. [doi: 10.1177/0706743717723825] [Medline: 28845686]
9. Lawn S, Roberts L, Willis E, Couzner L, Mohammadi L, Goble E. The effects of emergency medical service work on the psychological, physical, and social well-being of ambulance personnel: a systematic review of qualitative research. BMC Psychiatry 2020 Jul 3;20(1):348. [doi: 10.1186/s12888-020-02752-4] [Medline: 32620092]
10. Kyron MJ, Rikkers W, Page AC, et al. Prevalence and predictors of suicidal thoughts and behaviours among Australian police and emergency services employees. Aust N Z J Psychiatry 2021 Feb;55(2):180-195. [doi: 10.1177/0004867420937774] [Medline: 32615800]

XSL•FO
RenderX

11. Harvey S, Forbes D, Glozier N, et al. Expert guidelines: diagnosis and treatment of post-traumatic stress disorder in emergency service workers. Black Dog Institute. 2024. URL: https://www.blackdoginstitute.org.au/wp-content/uploads/2024/11/BDI_PTSD_Guidelines_A4_DIGITAL_V2.pdf
12. Jones S, Agud K, McSweeney J. Barriers and facilitators to seeking mental health care among first responders: "removing the darkness". J Am Psychiatr Nurses Assoc 2020;26(1):43-54. [doi: 10.1177/1078390319871997] [Medline: 31509058]
13. Beck C, McSweeney JC, Richards KC, Roberson PK, Tsai PF, Souder E. Challenges in tailored intervention research. Nurs Outlook 2010;58(2):104-110. [doi: 10.1016/j.outlook.2009.10.004] [Medline: 20362779]
14. Steinhubl SR, Muse ED, Topol EJ. The emerging field of mobile health. Sci Transl Med 2015 Apr 15;7(283):283rv3. [doi: 10.1126/scitranslmed.aaa3487] [Medline: 25877894]
15. Linardon J, Cuijpers P, Carlbring P, Messer M, Fuller-Tyszkiewicz M. The efficacy of app-supported smartphone interventions for mental health problems: a meta-analysis of randomized controlled trials. World Psychiatry 2019 Oct;18(3):325-336. [doi: 10.1002/wps.20673] [Medline: 31496095]
16. Price M, Yuen EK, Goetter EM, et al. mHealth: a mechanism to deliver more accessible, more effective mental health care. Clin Psychol Psychother 2014;21(5):427-436. [doi: 10.1002/cpp.1855] [Medline: 23918764]
17. Donker T, Petrie K, Proudfoot J, Clarke J, Birch MR, Christensen H. Smartphones for smarter delivery of mental health programs: a systematic review. J Med Internet Res 2013 Nov 15;15(11):e247. [doi: 10.2196/jmir.2791] [Medline: 24240579]
18. Collins DAJ, Harvey SB, Lavender I, Glozier N, Christensen H, Deady M. A pilot evaluation of a smartphone application for workplace depression. Int J Environ Res Public Health 2020 Sep 16;17(18):6753. [doi: 10.3390/ijerph17186753] [Medline: 32947994]
19. Firth J, Torous J, Nicholas J, Carney R, Rosenbaum S, Sarris J. Can smartphone mental health interventions reduce symptoms of anxiety? A meta-analysis of randomized controlled trials. J Affect Disord 2017 Aug 15;218:15-22. [doi: 10.1016/j.jad.2017.04.046] [Medline: 28456072]
20. Firth J, Torous J, Nicholas J, et al. The efficacy of smartphone-based mental health interventions for depressive symptoms: a meta-analysis of randomized controlled trials. World Psychiatry 2017 Oct;16(3):287-298. [doi: 10.1002/wps.20472] [Medline: 28941113]
21. Luxton DD, McCann RA, Bush NE, Mishkind MC, Reger GM. mHealth for mental health: Integrating smartphone technology in behavioral healthcare. Prof Psychol: Res Pract 2011;42(6):505-512. [doi: 10.1037/a0024485]
22. Deady M, Choi I, Calvo RA, Glozier N, Christensen H, Harvey SB. eHealth interventions for the prevention of depression and anxiety in the general population: a systematic review and meta-analysis. BMC Psychiatry 2017 Aug 29;17(1):310. [doi: 10.1186/s12888-017-1473-1] [Medline: 28851342]
23. Campbell M, Fitzpatrick R, Haines A, et al. Framework for design and evaluation of complex interventions to improve health. BMJ 2000 Sep 16;321(7262):694-696. [doi: 10.1136/bmj.321.7262.694] [Medline: 10987780]
24. Deady M, Glozier N, Calvo R, et al. Preventing depression using a smartphone app: a randomized controlled trial. Psychol Med 2022 Feb;52(3):457-466. [doi: 10.1017/S0033291720002081] [Medline: 32624013]
25. Deady M, Glozier N, Collins D, et al. The utility of a mental health app in apprentice workers: a pilot study. Front Public Health 2020;8:389. [doi: 10.3389/fpubh.2020.00389] [Medline: 33014953]
26. Bakker D, Kazantzis N, Rickwood D, Rickard N. Mental health smartphone apps: review and evidence-based recommendations for future developments. JMIR Ment Health 2016 Mar 1;3(1):e7. [doi: 10.2196/mental.4984] [Medline: 26932350]
27. Alqahtani F, Orji R. Insights from user reviews to improve mental health apps. Health Informatics J 2020 Sep;26(3):2042-2066. [doi: 10.1177/1460458219896492]
28. Chandrashekar P. Do mental health mobile apps work: evidence and recommendations for designing high-efficacy mental health mobile apps. Mhealth 2018;4(6):6. [doi: 10.21037/mhealth.2018.03.02] [Medline: 29682510]
29. Stoyanov SR, Hides L, Kavanagh DJ, Zelenko O, Tjondronegoro D, Mani M. Mobile app rating scale: a new tool for assessing the quality of health mobile apps. JMIR Mhealth Uhealth 2015 Mar 11;3(1):e27. [doi: 10.2196/mhealth.3422] [Medline: 25760773]
30. Slade T, Grove R, Burgess P. Kessler Psychological Distress Scale: normative data from the 2007 Australian National Survey of Mental Health and Wellbeing. Aust N Z J Psychiatry 2011 Apr;45(4):308-316. [doi: 10.3109/00048674.2010.543653] [Medline: 21332432]
31. Kroenke K, Spitzer RL. The PHQ-9: a new depression diagnostic and severity measure. Psychiatr Ann 2002 Sep;32(9):509-515. [doi: 10.3928/0048-5713-20020901-06]
32. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
33. Topp CW, Østergaard SD, Søndergaard S, Bech P. The WHO-5 Well-Being Index: a systematic review of the literature. Psychother Psychosom 2015;84(3):167-176. [doi: 10.1159/000376585] [Medline: 25831962]
34. Benight CC, Shoji K, James LE, Waldrep EE, Delahanty DL, Cieslak R. Trauma coping self-efficacy: a context-specific self-efficacy measure for traumatic stress. Psychol Trauma 2015 Nov;7(6):591-599. [doi: 10.1037/tra0000045] [Medline: 26524542]

XSL•FO
RenderX

35. Deady M, Johnston D, Milne D, et al. Preliminary effectiveness of a smartphone app to reduce depressive symptoms in the workplace: feasibility and acceptability study. JMIR Mhealth Uhealth 2018 Dec 4;6(12):e11661. [doi: 10.2196/11661] [Medline: 30514694]
36. Deady M, Johnston DA, Glozier N, et al. Smartphone application for preventing depression: study protocol for a workplace randomised controlled trial. BMJ Open 2018 Jul 13;8(7):e020510. [doi: 10.1136/bmjopen-2017-020510] [Medline: 30007927]
37. Torous J, Nicholas J, Larsen ME, Firth J, Christensen H. Clinical review of user engagement with mental health smartphone apps: evidence, theory and improvements. Evid Based Ment Health 2018 Aug;21(3):116-119. [doi: 10.1136/eb-2018-102891] [Medline: 29871870]
38. Ng MM, Firth J, Minen M, Torous J. User engagement in mental health apps: a review of measurement, reporting, and validity. Psychiatr Serv 2019 Jul 1;70(7):538-544. [doi: 10.1176/appi.ps.201800519] [Medline: 30914003]
39. Baumel A, Muench F, Edan S, Kane JM. Objective user engagement with mental health apps: systematic search and panel-based usage analysis. J Med Internet Res 2019 Sep 25;21(9):e14567. [doi: 10.2196/14567] [Medline: 31573916]
40. Donkin L, Christensen H, Naismith SL, Neal B, Hickie IB, Glozier N. A systematic review of the impact of adherence on the effectiveness of e-therapies. J Med Internet Res 2011 Aug 5;13(3):e52. [doi: 10.2196/jmir.1772] [Medline: 21821503]
41. Yardley L, Spring BJ, Riper H, et al. Understanding and promoting effective engagement with digital behavior change interventions. Am J Prev Med 2016 Nov;51(5):833-842. [doi: 10.1016/j.amepre.2016.06.015] [Medline: 27745683]
42. Gan DZQ, McGillivray L, Han J, Christensen H, Torok M. Effect of engagement with digital interventions on mental health outcomes: a systematic review and meta-analysis. Front Digit Health 2021;3:764079. [doi: 10.3389/fdgth.2021.764079] [Medline: 34806079]
43. Yardley L, Morrison L, Bradbury K, Muller I. The person-based approach to intervention development: application to digital health-related behavior change interventions. J Med Internet Res 2015 Jan 30;17(1):e30. [doi: 10.2196/jmir.4055] [Medline: 25639757]
44. Schubart JR, Stuckey HL, Ganeshamoorthy A, Sciamanna CN. Chronic health conditions and internet behavioral interventions: a review of factors to enhance user engagement. Comput Inform Nurs 2011 Feb;29(2):81-92. [doi: 10.1097/NCN.0b013e3182065eed] [Medline: 21164337]
45. Engdahl P, Svedberg P, Bejerholm U. Acceptability of a digital return-to-work intervention for common mental disorders: a qualitative study on service user perspectives. BMC Psychiatry 2021 Aug 3;21(1):384. [doi: 10.1186/s12888-021-03386-w] [Medline: 34344327]
46. Yardley L, Choudhury T, Patrick K, Michie S. Current issues and future directions for research Into digital behavior change interventions. Am J Prev Med 2016 Nov;51(5):814-815. [doi: 10.1016/j.amepre.2016.07.019] [Medline: 27745680]
47. Jones S. Describing the mental health profile of first responders: a systematic review [formula: see text]. J Am Psychiatr Nurses Assoc 2017 May;23(3):200-214. [doi: 10.1177/1078390317695266] [Medline: 28445653]
48. Pavani JB, Berna G, Andreotti E, et al. Between-individual differences in baseline well-being and emotion regulation strategy use moderate the effect of a self-help cognitive-behavioral intervention for typical adults. Appl Psychol Health Well Being 2020 Jul;12(2):411-431. [doi: 10.1111/aphw.12189] [Medline: 31869005]
49. Snow RE. Aptitude-treatment interaction as a framework for research on individual differences in psychotherapy. J Consult Clin Psychol 1991 Apr;59(2):205-216. [doi: 10.1037//0022-006x.59.2.205] [Medline: 2030178]
50. Farvolden P, Denisoff E, Selby P, Bagby RM, Rudy L. Usage and longitudinal effectiveness of a Web-based self-help cognitive behavioral therapy program for panic disorder. J Med Internet Res 2005 Mar 26;7(1):e7. [doi: 10.2196/jmir.7.1.e7] [Medline: 15829479]
51. Klein DN, Kotov R, Bufferd SJ. Personality and depression: explanatory models and review of the evidence. Annu Rev Clin Psychol 2011;7:269-295. [doi: 10.1146/annurev-clinpsy-032210-104540] [Medline: 21166535]
52. Morgan C, Mason E, Newby JM, et al. The effectiveness of unguided internet cognitive behavioural therapy for mixed anxiety and depression. Internet Interv 2017 Dec;10:47-53. [doi: 10.1016/j.invent.2017.10.003] [Medline: 30135752]
53. Fry JP, Neff RA. Periodic prompts and reminders in health promotion and health behavior interventions: systematic review. J Med Internet Res 2009 May 14;11(2):e16. [doi: 10.2196/jmir.1138] [Medline: 19632970]

Abbreviations

CSE-T: Trauma Coping Self-Efficacy Scale
***DSM-IV*:** *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition*
GAD-7: 7-item Generalized Anxiety Disorder
K10: Kessler Psychological Distress Scale
MARS: Mobile Application Rating Scale
MH: mental health
PHQ-9: 9-item Patient Health Questionnaire
PTSD: posttraumatic stress disorder
RCT: randomized controlled trial
WHO-5: 5-item World Health Organization Well-Being Index

XSL•FO
RenderX

Edited by A Mavragani; submitted 19.07.23; peer-reviewed by C Cha, S Hatcher, T Benham; revised version received 03.04.25; accepted 08.04.25; published 28.07.25.

Please cite as:
Meuldijk D, Deady M, Collins DAJ, Williams DO, Bryant RA, Harvey SB
Feasibility of a Mental Health App Intervention for Emergency Service Workers and Volunteers: Single-Arm Pilot Study
JMIR Form Res 2025;9:e50995
URL: https://formative.jmir.org/2025/1/e50995
doi:10.2196/50995

© Denise Meuldijk, Mark Deady, Daniel A J Collins, Douglas O Williams, Richard A Bryant, Samuel B Harvey. Originally published in JMIR Formative Research (https://formative.jmir.org), 28.7.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Gait Disturbances in Older Adults With Cerebral Small Vessel Disease: Mixed Methods Study Using Smartphone Sensors and Video Analysis

Xiaojun Lai[1]; Li-Yan Qiao[2]; Pei-Luen Patrick Rau[1]; Yankuan Liu[1]

[1]Department of Industrial Engineering, Tsinghua University, Beijing, China

[2]Department of Neurology, Yuquan Hospital of Tsinghua University, No. 5 Shijingshan Road, Shijingshan District, Beijing, China

Corresponding Author:
Li-Yan Qiao
Department of Neurology, Yuquan Hospital of Tsinghua University, No. 5 Shijingshan Road, Shijingshan District, Beijing, China

Abstract

Background: Cerebral small vessel disease (CSVD) significantly impacts motor functions, particularly gait dynamics. However, its analysis often lacks the integration of comprehensive tools that capture the multifaceted nature of gait disturbances. Traditional methods may not fully address the complexity of CSVD's impact on gait, underscoring the need for a detailed exploration of gait characteristics through advanced technological means.

Objective: This study aims to identify the distinct gait patterns and postural adaptations present in patients with CSVD compared to a healthy older population, using an integrative analysis combining sensor and video data to provide a holistic understanding of gait dynamics in CSVD.

Methods: This study involved 90 participants older than 50 years (mean age 68.85, SD 9.74 years; 47 males and 43 females), with 24 categorized as normal controls (mean age 66.42, SD 7.51 years) and 66 diagnosed with CSVD (mean age 69.74, SD 10.37 years). Participants performed three walking tasks: normal walking, dual-task walking (with concurrent mental arithmetic), and fast walking. Gait parameters were collected through video data for image posture parameters using the OpenPose BODY_25 key point model, and the "Pocket Gait Test" smartphone app for sensor-based parameters sampled at approximately 40 Hz. Data analysis included 5 sensor-based parameters (step frequency, root mean square (RMS), step variability, step regularity, and step symmetry) and 6 key video-based parameters (including knee angle, ankle angle, elbow angle, body trunk angles, and head posture).

Results: Among the 29 participants with complete sensor and video data (10 normal controls and 19 patients with CSVD), significant differences were observed in step regularity (normal walking: mean 0.76, SD 0.09 vs mean 0.61, SD 0.25; $P<.003$ and dual-task: mean 0.74, SD 0.13 vs mean 0.57 SD 0.24; $P<.005$), RMS (normal walking: mean 1.64, SD 0.45 vs mean 1.43, SD 0.42; $P<.006$), and forward head posture angles (head-to-body angle during normal walking: mean 132.96, SD 7.78 vs mean 128.07, SD 7.99; $P<.02$ and head-to-ground angle: mean 134.11, SD 8.28 vs mean 128.40, SD 9.75; $P<.008$) between the CSVD and control groups. The CSVD group exhibited a more pronounced forward head posture across all walking tasks, with the greatest difference observed during dual-task walking (head-to-ground angle: mean 134.43, SD 8.29 vs mean 125.02, SD 8.42; $P<.02$).

Conclusions: The study provides compelling evidence of distinct gait disturbances in patients with CSVD, characterized by reduced step regularity (15% - 23% lower than controls), altered acceleration patterns, and significant postural adaptations, particularly forward head positioning (4° - 7° more pronounced than controls). These quantifiable differences, detectable through accessible smartphone and video technology, offer potential biomarkers for early CSVD detection and monitoring. The integration of sensor and video analysis provides a more comprehensive assessment approach that could be implemented in both clinical and home settings for longitudinal monitoring of disease progression and rehabilitation outcomes.

(JMIR Form Res 2025;9:e58864) doi:10.2196/58864

KEYWORDS

cerebral small vessel disease; gait analysis; motor function assessment; cerebral; gait; motor function; mixed methods; mixed methods study; sensors; video analysis; video; tools; accelerometer; video data

XSL•FO
RenderX

Introduction

Background and Motivation

Gait analysis is commonly used to detect abnormalities in walking behavior, analyze physical function, and assess disease rehabilitation, indicating the occurrence of poor health problems or the progression of neurological diseases. Cerebral small vessel disease (CSVD) is characterized by brain white matter lesions, lacunar infarcts, and microbleeds on cranial imaging (especially head magnetic resonance imaging [MRI]) [1,2], and clinically can present as stroke, cognitive impairment, gait and balance dysfunction, emotional disorders, urinary disorders, etc [3,4]. CSVD has an estimated prevalence of 5% - 10% in the general population older than 50 years of age, with numbers increasing significantly with advancing age [5,6]. Recent evidence indicates that CSVD is the most common pathology underlying vascular cognitive impairment and is increasingly recognized as a primary contributor to gait disturbances in older adults [7,8]. Gait disturbance is a major and important clinical manifestation of patients with CSVD [9], but it is severely underestimated in clinical practice. There is an urgent need in clinical practice to intelligently collect, analyze, and report data on the gait of patients with CSVD and to diagnose and follow up on their condition.

Current assessment methods for CSVD-related gait disturbances are primarily limited to traditional clinical evaluations or laboratory-based systems that lack real-world applicability [3]. These existing approaches often fail to capture the subtle, multidimensional aspects of gait abnormalities specific to CSVD, particularly in everyday settings [10]. Recent systematic reviews and meta-analyses have highlighted that patients with CSVD commonly exhibit reduced walking speed, shorter stride length, increased stance time, and greater gait variability compared to healthy controls [3,8]. These gait characteristics are detectable even in early-stage CSVD and may precede cognitive symptoms, underscoring their potential value as early biomarkers. However, in the medical diagnostic process, many motor function impairments and gait disturbances associated with CSVD cannot be conveniently assessed using traditional cumbersome clinical tools. These motor manifestations, which often precede cognitive symptoms in patients with CSVD, require more accessible and efficient measurement approaches for timely diagnosis and monitoring.

Due to its convenience, remote application, and quantifiability, an intelligent wearable gait analysis measurement system offers a promising alternative direction in gait analysis. This approach directly addresses the limitations of conventional methods by enabling continuous, ecological monitoring of gait parameters in patients' natural environments. Smartphone-based gait analysis has emerged as a valid and reliable method for capturing gait parameters, with recent validation studies demonstrating high agreement between smartphone accelerometry data and gold-standard systems in measuring key spatiotemporal gait parameters [10,11]. These smartphone-based systems can detect subtle gait abnormalities and have shown particular promise in neurological conditions where traditional assessment methods may be insensitive to early changes. On the one hand, medical costs are too high, and on the other hand, patients may delay diagnosis or treatment due to high medical expenses, requiring simpler and easier-to-use tools to assist in diagnosing diseases. In addition, the physical walking posture characteristics of patients at various stages, such as during illness and recovery, are significantly different, requiring tools that can effectively assist in assessing daily rehabilitation conditions over a long period. This study proposes combining image and sensor data to establish an intelligent wearable gait measurement system. This system allows for analyzing gait differences between patients with CSVD and healthy older groups. It also enables tracking of patient data during illness and recovery periods. Through this approach, key gait disturbances specifically related to CSVD can be identified. These findings can provide valuable reference points for the diagnosis of patients with CSVD and assist in their health monitoring and rehabilitation assessment.

This study aimed to conduct a detailed analysis of gait characteristics in patients with CSVD, using both sensor and video data. The objective was to identify and characterize distinct gait patterns and postural adaptations in patients with CSVD and to compare these with those of a healthy older population. Specifically, this study's contribution includes the collection and analysis of real-world data from patients with CSVD. This effort sought to contribute to a deeper understanding of the impact of CSVD on motor functions, potentially guiding clinical assessments and rehabilitation strategies.

Related Studies

Gait is the behavioral characteristic of human walking. Normal gait is the gait of a healthy adult walking in a natural state and feeling most comfortable, characterized by stable body posture, appropriate stride length, and minimal energy consumption [12]. Quantitative gait analysis is an important clinical indicator and, as a routine analysis in patient management, can be used together with medical history, physical examination, and other special investigations. Clinical gait analysis typically includes five elements: videotape examination, measurement of general spatiotemporal gait parameters, kinematic analysis, kinetic measurements, and electromyography [13]. Spatiotemporal gait data includes stride length, stride width, support time, rhythm, walking speed, swing time, and double-limb support time [14,15]. Gait parameters, including gait intensity, smoothness or regularity [16], variability, and complexity [17], can also be described by collecting acceleration signals during walking using specialized devices. Using kinematic (such as joint angles and angular velocities) and kinetic (such as ground reaction forces and joint torques) data, joint torques and forces can be calculated more precisely in 3D, and electromyography from specific muscles provides a detailed description of the gait mechanism [13,18].

Attributing gait abnormalities to a specific disease is difficult, as many diseases present similar gait abnormalities [19]. Special and precise quantification of gait abnormalities for specific diseases is needed. Diseases related to the cerebellum usually accompany specific gait differences, such as ataxic gait in spinocerebellar ataxia and cerebellar stroke. These conditions

XSL•FO
RenderX

typically present with a wide-based stance, irregular stepping patterns, and poor coordination [20]. Similarly, CSVD has its own characteristic gait disturbances that require specific identification and quantification. Research on CSVD found that patients have reduced gait capabilities such as walking speed and stride length. De Laat et al [21] conducted MRI, voxel-based morphometry analysis, and walking ability assessments on 429 patients aged 50-80 years with CSVD but without dementia or Parkinson, finding a significant relationship between white matter lesions in CSVD and reduced gait ability, especially in terms of walking speed, stride length, and broad-based gait. In addition, they analyzed the relationship between the characteristic microbleeds of CSVD and gait ability using MRI to calculate the location and proportion of microbleeds, and using the GAITRite system, Tinetti gait scale, and Timed Up and Go measurements, finding that "microbleeds" resulted in shorter walking stride lengths and significantly lower Tinetti and Timed Up and Go scores [22]. Zong et al [23] used scales such as the ataxia rating scale, Tinetti gait score, Tinetti balance score, etc, for behavioral ability assessment and conducted gait and balance disorder studies on 57 patients with CSVD, finding that more than half of the participants showed gait disturbances such as reduced speed, dragging, wide base, and unequal bilateral stride lengths.

Traditional gait analysis methods are limited to fixed laboratory measurement environments, often using complex instruments such as single-camera image processing [24] and walking leap sensors [25] to record human gait motion, but they struggle to reflect real-world gait conditions. Moreover, nonwearable system equipment and testing are too expensive, making it difficult for people to measure related data on their own [26]. In recent years, portable, wireless real-time recording of daily gait information has become the main method of data collection. Zhong and Rau [27] used smartphones for gait assessment based on embedded accelerometers, finding that the reliability and effectiveness of smartphones in measuring gait parameters (stride length variability, autocorrelation, and root mean square [RMS] of acceleration) were comparable to external accelerometers. More recently, Tao et al [28] validated smartphone-based gait analysis across different attachment positions and walking speeds, including applications in patients with CSVD. Their study confirmed measurement validity in controlled environments, establishing that smartphones can provide reliable gait measurements comparable to traditional inertial measurement units. In addition, using OpenPose (developed by Cao et al [29] at Carnegie Mellon University) image recognition technology to collect key body metrics during walking and to calculate body posture parameters during walking can also serve as important gait feature data. At present, the key gait disturbances in patients with CSVD have not been fully identified, and there is ongoing research to establish a complete intelligent wearable gait measurement system based on smartphone-collected gait parameters, for use in assisting clinical diagnosis and rehabilitation monitoring. Other influencing factors, such as demographic information (age, sex, BMI, etc), disease severity, and other concurrent diseases, may also bring about changes or differences in gait. For example, gait abnormalities increase with age; men are more likely to have neurological gait abnormalities (such as stroke), while women are more likely to have non-neurological gait abnormalities (such as arthritis, cardiac, or respiratory diseases) [30]. Therefore, in addition to gait data, tracking and combining medical diagnostic results for in-depth research are needed to better assist in the diagnosis, assessment, and rehabilitation monitoring of diseases.

Methods

Overview

This study was designed to assess gait characteristics in older adults with and without CSVD using a combination of sensor-based and video-based analysis methods. Participants completed three walking tasks—normal walking, dual-task walking, and fast walking—while wearing a smartphone equipped with accelerometers and being recorded by a camera. Data were collected on spatiotemporal gait parameters and postural angles to identify specific gait disturbances. The complete setup and data collection process are presented in Figures 1A and 1B.

XSL•FO
RenderX

Figure 1. Field experimental environment setup and data collection method for gait assessment in patients with cerebral small vessel disease and healthy controls. (A) A participant wearing a waist bag containing a smartphone positioned at the L3 vertebra region for collecting accelerometer data and (B) the positioning of the GoPro HERO8 camera mounted on a tripod to capture a 5-meter walking segment.

Recruitment

The study was conducted at Yuquan Hospital of Tsinghua University, a major medical center in Beijing, China, specializing in neurological disorders and geriatric care. Participants were not recruited through conventional research recruitment processes. Instead, a collaborative clinical screening approach was used. Physicians at Yuquan Hospital identified potentially eligible patients from their clinical practice who met the study criteria.

The recruitment process began with hospital neurologists reviewing medical records to identify individuals older than 50 years who either had a confirmed CSVD diagnosis or were free of neurological disease (for the control group). The physicians then conducted preliminary assessments of these individuals during their regular clinical visits to determine their suitability for the study based on mobility status and cognitive ability to follow instructions. After this clinical screening, suitable candidates were approached about potential participation in the study during their hospital visits. The physicians explained the study's purpose, procedures, and potential benefits and risks. Those who expressed interest were provided with detailed written and verbal information about the study. All participation was strictly voluntary, and patients were assured that their decision regarding participation would not affect their clinical care.

For the CSVD group, inclusion was based on the presence of characteristic neuroimaging findings (white matter lesions, lacunar infarcts, or microbleeds on MRI) and clinical symptoms consistent with CSVD, as determined by the attending neurologists. Control group participants were recruited from the same hospital setting but were confirmed to be free of CSVD and other neurological conditions through clinical and radiological assessment.

This study involved 90 participants older than 50 years, of whom 24 participants were categorized as normal and 66 participants with CSVD. The groups were demographically comparable: the average age was 66.42 years in the normal group and 69.74 in the disease group (P=.14); sex distribution was 11 males to 13 females in the normal group and 36 men to 30 women in the disease group (P=.47). Heights and weights were also similar: normal group averaged 163.62 cm and 65.58 kg, while the disease group averaged 164.30 cm and 64.78 kg (P=.67 and .63, respectively). BMI showed no significant difference either, with 24.45 in the normal group and 23.81 in the disease group (P=.40). The demographic statistics for the participants are shown in Table 1.

XSL•FO
RenderX

Table . Demographic characteristics of participants in a cross-sectional observational study comparing gait parameters between patients with cerebral small vessel disease and healthy controls, conducted at Yuquan Hospital.

Variable	Normal, mean (SD)	CSVD[a], mean (SD)	*P* value
Age (years)	66.42 (7.51)	69.74 (10.37)	.14
Height (cm)	163.62 (8.03)	164.30 (8.62)	.67
Weight (kg)	65.58 (9.47)	64.78 (12.91)	.63
BMI	24.45 (2.55)	23.81 (3.50)	.40

[a]CSVD: cerebral small vessel disease.

Variables and Measurements

This study involved collecting gait parameters through accelerometer data for sensor-based parameters and video data for image posture parameters. Initially, 30-second walking data were gathered using the "Pocket Gait Test," a gait measurement app developed by the authors' research team (Zhong and Rau [27]). While other smartphone validation approaches, such as those by Tao et al [28], focus primarily on validating measurement accuracy, the "Pocket Gait Test" was specifically designed for analyzing subtle gait parameters in neurological conditions and older adults. This app uses the accelerometer embedded in smartphones and provides voice prompts to start and end the gait assessment, with a sampling rate of approximately 40 Hz. Custom algorithms developed in MATLAB (MathWorks) processed these data, resampling the discontinuous raw data to 100 Hz using interpolation.

Sensor gait parameters included step frequency, acceleration RMS, step variability, step regularity, and step symmetry. Step frequency refers to the rate of walking; RMS represents the magnitude of force generated during walking; step variability indicates the potential unsteadiness in gait, where lower values are preferable; step regularity refers to the consistency in walking pattern throughout; and step symmetry measures the difference between left and right leg movements, with higher symmetry indicating better gait balance.

Furthermore, walking videos covering approximately 5 meters were recorded using a GoPro HERO8 camera mounted on a tripod, set at a resolution of 1080p and a frame rate of 60 fps. The BODY_25 key point model was used with the OpenPose algorithm [29] to identify key body points and output JSON files for each video frame. This approach has been validated in previous human motion analysis studies [31-33]. Python was used to connect key points on the left (or right) side of the body into vectors and to process the videos in batches, following a methodology similar to that described by Viswakumar et al [31].

For signals with anomalies, especially when participants entered or exited the screen and key points were incomplete, parameters were adjusted to reselect video segments for analysis. MATLAB was then used for signal processing, using Db4 for three-level wavelet transform decomposition and reconstruction before data extraction. The output image posture parameters included the knee angle, defined as the angle between the thigh and calf segments; the ankle angle, representing the angle between the calf and foot segments; the elbow angle, indicating the angle at the elbow joint; the angle between the upper arm and body trunk, reflecting upper body posture; the angle between the body trunk and thigh, which reflects hip flexion or extension; the angle between the neck and body trunk, indicating forward head position; body lean, measured as the angle of the trunk relative to the vertical; sight line, representing the orientation angle of the head; heel height, defined as the vertical distance of the heel from the ground; step length, which is the distance between consecutive heel strikes; and step speed, which is derived from step length and time.

Particularly for KA, ankle angle, EA, the angle between the body trunk and thigh, and the angle between the upper arm and body trunk, which dynamically change with walking, the discussion focused on their mean and variance, while the analysis of the other parameters primarily concentrated on their mean values. The measurement of these angles follows standardized biomechanical conventions [34].

Procedure

The experiment was conducted in the corridors of Yuquan Hospital. Before its commencement, participants' age, sex, height, and weight were recorded. Their balance abilities and confidence were assessed using the Tinetti scale [35] and the Activities-Specific Balance Confidence (ABC) scale [36]. During the experiment, participants wore a waist bag containing a smartphone placed in the pocket at the L3 area of the waist to collect walking data for calculating sensor-based gait parameters.

Participants were instructed to perform three different walking tasks in order: normal walking, dual-task walking, and fast walking. These three tasks did not require specific speed requirements. For the first two tasks, participants were asked to walk at a natural and comfortable pace. The third task required participants to walk as quickly as possible. In addition, the second task involved a dual-task load, wherein participants were required to continuously perform mental arithmetic (addition and subtraction of two-digit numbers) while walking, to assess the impact of cognitive load on gait. Tasks were assigned randomly to eliminate the influence of the order of the experiment. Each condition was repeated twice. During the walking tasks, a GoPro HERO8 camera mounted on a tripod was positioned on the side of the corridor. This setup captured a 5-meter midsection of the participant's walking path, including videos of both the left and right sides of the body.

The experimental setup controlled for variables independent of the video by managing the distance between the tripod and the participant, ensuring the participant occupied a consistent proportion of the video frame. This control allowed for the

XSL•FO
RenderX

collection of consistent and reliable data across all participants and conditions. The field experimental environment is shown in Figure 1.

Illustration of the experimental setup at Yuquan Hospital corridors, where participants performed walking tasks while being recorded. The image shows (A) a participant wearing a waist bag containing a smartphone positioned at the L3 vertebra region for collecting accelerometer data, and (B) the positioning of the GoPro HERO8 camera mounted on a tripod to capture a 5-meter walking segment. Three walking conditions were tested: normal walking, dual-task walking (with mental arithmetic), and fast walking. Data were collected between January and September 2023 from 90 participants (24 healthy controls and 66 patients with CSVD), with complete data available for 29 participants (10 controls and 19 patients with CSVD).

Data Exclusion

This study implemented a rigorous data exclusion process to maintain the quality and reliability of its results. After exclusions, the final analysis included 29 participants (10 normal controls and 19 patients with CSVD). The demographic characteristics of these participants are shown in Table 2. Similar to the initial recruitment cohort, these groups remained demographically comparable. This comparability ensures the validity of subsequent gait parameter comparisons between groups in the final analyzed cohort.

Table . Demographic characteristics of participants with complete data included in the final analysis of a cross-sectional study on gait parameters in cerebral small vessel disease.

Variables	Normal, mean (SD)	CSVD[a], mean (SD)	*P* value
Age (years)	67.40 (6.79)	71.79 (7.56)	.12
Height (cm)	165.10 (9.65)	168.11 (7.65)	.29
Weight (kg)	66.10 (11.30)	65.13 (12.34)	.82
BMI	24.12 (2.23)	22.92 (3.22)	.37

[a]CSVD: cerebral small vessel disease.

Factors contributing to incomplete data included technical issues with sensor equipment (9.8% of exclusions), errors in video recording (37.7% of exclusions), and participants' inability to complete all tasks due to physical limitations (29.5% of exclusions) or scheduling conflicts (26.2% of exclusions). In addition, data segments in which participants had just entered or exited the screen, resulting in incomplete key point detection by the OpenPose algorithm (approximately 10% of video recordings), were excluded to maintain consistency in the posture and gait analysis.

This stringent exclusion criterion was essential to ensure that the analysis was based on complete and accurate representations of each participant's gait characteristics, thereby enhancing the study's overall validity and reliability. While this reduced our effective sample size, it ensured that all analyzed data met the highest quality standards necessary for drawing meaningful conclusions about gait disturbances in CSVD.

Ethical Considerations

This study follows the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines for cross-sectional studies and was conducted in accordance with ethical standards. Ethical approval was obtained from the Ethics Committee of Yuquan Hospital, Tsinghua University, and the Science and Technology Ethics Committee (Humanities, Social Sciences, and Engineering) of Tsinghua University. The formal IRB approval number is THU-04-2025-1023. All participants provided written informed consent before they participated in the study, which included consent for the collection, analysis, and publication of anonymized data and results. Participants did not receive financial compensation. The experiment was conducted on-site at Yuquan Hospital, where participants were already present for clinical care. Participation was entirely voluntary, with participants signing up to take part in the study of their own accord. All experimental procedures were strictly followed by the guidance of attending physicians to ensure participant safety and comfort. All assessments and data collection procedures were scheduled to accommodate participants' daily routines and were conducted with careful attention to minimize fatigue, discomfort, or disruption to their normal activities. Their privacy and confidentiality were ensured throughout the study. All collected data were anonymized by assigning unique identification codes to each participant, and personal identifiers were stored separately from research data. During data analysis and reporting, no personally identifiable information was used, and results were presented as aggregated data. The video recordings and images used for gait analysis were processed using the OpenPose algorithm, which converts human figures into skeletal representations and removes identifiable facial features and body characteristics. No identifiable images of participants are included in the paper or supplementary materials. For the field experimental environment setup image (Figure 1), consent was obtained from the individual shown, and the image was processed to obscure any identifiable features.

Results

Sensor Data

Mann-Whitney *U* tests were used to compare gait characteristics between the normal and CSVD groups across various walking tasks, revealing several patterns. During a normal walking task, significant differences were noted in RMS (P=.006) and step regularity (P=.003), with the normal group having higher values, while step frequency, step variability, and step symmetry showed no significant differences between groups.

XSL•FO
RenderX

During dual walking conditions, RMS (*P*=.006) and step regularity (*P*=.005) again differed significantly, with the normal group showing higher values. In addition, step symmetry was significantly higher in the normal group (*P*=.001), while step frequency (*P*=.83) and step variability (*P*=.48) remained not significantly different.

In the fast walking condition, RMS (*P*<.001) and step regularity (*P*=.01) continued to show significant differences, with the normal group exhibiting higher values. Step frequency, step variability, and step symmetry did not differ significantly between groups.

These findings suggest that while certain aspects of gait, such as step frequency and variability, are not significantly affected by CSVD, other parameters, such as RMS and step regularity, consistently show noticeable differences across various walking conditions, with the CSVD group demonstrating significantly lower values than the normal group. Detailed sensor results are presented in Table 3.

Table . Comparison of sensor-based gait parameters between normal controls and patients with cerebral small vessel disease during three walking tasks.

Characteristic	Normal, mean (SD)	CSVD[a], mean (SD)	*P* value
Normal walking			
Step frequency	1.75 (0.19)	2.00 (0.60)	.56
RMS[b]	1.64 (0.45)	1.43 (0.42)	.006
Step variability	0.10 (0.06)	0.12 (0.08)	.59
Step regularity	0.76 (0.09)	0.61 (0.25)	.003
Step symmetry	0.90 (0.06)	0.85 (0.15)	.11
Dual-task walking			
Step frequency	1.70 (0.17)	1.92 (0.43)	.82
RMS	1.57 (0.45)	1.28 (0.38)	.006
Step variability	0.10 (0.06)	0.13 (0.08)	.48
Step regularity	0.74 (0.13)	0.57 (0.24)	.005
Step symmetry	0.90 (0.09)	0.83 (0.15)	.001
Fast walking			
Step frequency	1.90 (0.18)	2.02 (0.42)	.58
RMS	2.17 (0.68)	1.86 (0.69)	<.001
Step variability	0.12 (0.06)	0.14 (0.06)	.72
Step regularity	0.80 (0.07)	0.63 (0.27)	.01
Step symmetry	0.94 (0.05)	0.81 (0.22)	.26

[a]CSVD: cerebral small vessel disease.

[b]RMS: root mean square.

Video Data

Mann-Whitney *U* tests were used. Analysis was conducted on the mean angles of key body joints on both the left and right sides of 29 participants, focusing on the ankle, knee, elbow, thigh, head, and back. This analysis examined the changes in angles and angular accelerations of these joints across various tasks. The results indicated significant differences between the normal group and the CSVD group in three specific measurements: the mean angle between the calf and the vertical line from the ground, the angle between the head and the body, and the angle between the head and the ground.

Under normal walking conditions, significant differences were observed in the angles of the head with respect to the body (*P*=.023) and the ground (*P*=.008). The normal group had a mean head-to-body angle of 132.96 (SD 7.78) degrees and a head-to-ground angle of 134.11 (SD 8.28) degrees, compared to 128.07 (SD 7.99) degrees and 128.40 degrees (SD 9.75) degrees, respectively, for the CSVD group.

During dual walking, the CSVD group exhibited a significantly greater calf angle of 163.20 (SD 3.07) degrees compared to 159.97 (SD 2.40) degrees in the normal group (*P*=.004). The head-to-body (*P*=.021) and head-to-ground angles (*P*=.015) also showed significant differences, with the disease group demonstrating a more pronounced forward head posture.

In fast walking, although the calf angle did not differ significantly between the groups, the angles involving the head continued to show significant differences, with the CSVD group maintaining a more downward head position (head-to-body: *P*=.024; head-to-ground: *P*=.003).

These results suggest that patients with CSVD tend to lower their heads more when walking. This alteration in head posture could be indicative of balance or visual scanning changes in

XSL•FO
RenderX

this population, potentially affecting their gait dynamics and overall mobility. Detailed video results are presented in Table 4.

Table . Comparison of video-based postural parameters between normal controls and patients with cerebral small vessel disease across three walking tasks.

Characteristic	Normal, mean (SD)	CSVD[a], mean (SD)	*P* value
Normal walking			
Calf	159.81 (2.06)	161.54 (2.94)	.26
Head and body	132.96 (7.78)	128.07 (7.99)	.02
Head and ground	134.11 (8.28)	128.40 (9.75)	.008
Dual walking			
Calf	159.97 (2.40)	163.20 (3.07)	.004
Head and body	133.64 (8.83)	126.38 (7.08)	.02
Head and ground	134.43 (8.29)	125.02 (8.42)	.02
Fast walking			
Calf	159.26 (2.63)	160.91 (3.31)	.40
Head and body	133.66 (6.21)	128.39 (6.40)	.02
Head and ground	134.25 (6.20)	126.98 (7.75)	.003

[a]CSVD: cerebral small vessel disease.

Discussion

Principal Findings

This study has identified three key findings that significantly advance the understanding of gait disturbances in CSVD: (1) patients with CSVD exhibit a distinct forward head posture during walking tasks, with significantly different head-to-body and head-to-ground angles compared to healthy controls (2) step regularity and RMS values are significantly reduced in patients with CSVD across all walking conditions, suggesting altered gait dynamics and reduced stability; and (3) the integration of sensor and video data provides complementary insights that neither method alone could capture, demonstrating the value of a multimodal approach to gait analysis in CSVD. These findings collectively address the initial objective to identify and characterize distinct gait patterns and postural adaptations in patients with CSVD compared to healthy older adults, using the integrative analysis approach combining sensor and video data.

The forward head posture observed in patients with CSVD through video data might reflect adaptive mechanisms to maintain balance or compensate for proprioceptive deficits. This raises new questions about the nature of postural adaptations in CSVD. Notably, the sensor data provided precise measurements of these postural changes, while the video data offered a more qualitative understanding of the overall gait pattern. In contrast, the lack of significant differences in step frequency and variability between the normal and CSVD groups suggests that basic temporal aspects of gait might be less affected by CSVD. This finding is intriguing as it contrasts with some previous research [21], possibly due to methodological differences or the variability in disease severity among participants. It is important to consider that traditional gait metrics, while valuable, might not be sensitive enough to capture all aspects of gait disturbances in CSVD.

Furthermore, this study highlights the importance of considering both sensor and video data in gait analysis. The sensor data's high precision in measuring joint angles and movements provides a detailed view of the mechanical aspects of gait, while video data offers a more holistic picture of gait, including body posture and coordination. This dual approach is crucial for a comprehensive analysis, as it captures both the quantitative and qualitative aspects of gait. The head angle, identified in this study as one of the differing factors in gait between patients with CSVD and healthy individuals, is challenging to capture through body sensors. The integration of video image processing enables the analysis of the overall gait in patients with CSVD, thereby providing more information that can be used to identify gait abnormalities in these individuals.

Integrating the clinical diagnosis of CSVD with the observed gait dynamics in patients with CSVD, this study underscores a critical linkage between early-stage CSVD manifestations and their impact on motor functions, as demonstrated by gait disturbances. The clinical findings from CSVD diagnosis reveal a pattern of gait disturbances that precede cognitive impairments, highlighting the significance of gait assessment in early CSVD identification. These disturbances, notably in dual-task walking conditions where patients exhibited lower average amplitude, gait regularity, and symmetry, align with observations of pronounced forward head posture in patients with CSVD through sensor and video analysis. This posture suggests an adaptation mechanism to maintain balance, potentially compensating for proprioceptive deficits intrinsic to CSVD. The convergence of clinical diagnosis with gait analysis in this study presents a compelling case for incorporating comprehensive gait assessment tools in the early detection and management of CSVD, fostering a holistic approach to patient

XSL•FO
RenderX

care. The correlation between clinical indicators of CSVD and specific gait disturbances further supports the integration of sensor and video gait analysis into CSVD's diagnostic and monitoring frameworks (refer to Table 5).

Table . Correlation between cerebral small vessel disease clinical findings and gait parameters in the observational study.

Parameter	Study finding	Clinical finding	*P* value (if applicable)
Forward head posture	Increase in patients with CSVD[a]	Confirmed	.008
Step regularity	Lower in patients with CSVD	Confirmed	.003
Step symmetry	Decrease in patients with CSVD	Confirmed	—
Head-to-body angle	Altered in patients with CSVD	Confirmed	.02
Gait disturbances	Identified	Confirmed	—

[a]CSVD: cerebral small vessel disease.

Limitations

This study has several limitations that should be acknowledged. Although information was collected from a total of 90 participants, due to constraints such as experimental duration, scheduling, and task difficulty, complete data were obtained from only 29 participants (including sensor and video data across 3 tasks). This smaller number of participants could potentially impact the results to a certain extent and limit the ability to conduct more detailed subgroup analyses or to control for potential confounding factors such as age, sex, or comorbidities.

Technical challenges also affected the data collection. Approximately 9.8% of the exclusions were due to issues with sensor equipment, while 37.7% resulted from errors in video recording. In addition, 29.5% of participants were unable to complete all tasks due to physical limitations, and 26.2% had scheduling conflicts. These high exclusion rates highlight the practical challenges of conducting comprehensive gait analysis in older populations, particularly those with neurological conditions such as CSVD.

The OpenPose algorithm used for video analysis occasionally resulted in incomplete key point detection (approximately 10% of video recordings), particularly when participants were entering or leaving the screen. This necessitated careful selection of video segments for analysis, which may have introduced some selection bias. Furthermore, while our approach of combining sensor and video data provides complementary insights, it also introduces complexity in data integration and interpretation that may not be readily transferable to clinical settings without specialized expertise.

Conclusions

This study offers valuable insights into the intricate gait disturbances experienced by patients with CSVD. By adopting a comprehensive approach that integrates sensor-based quantitative assessments with video-based qualitative observations, this study has enriched the understanding of CSVD's impact on motor functions. The identification of specific gait parameters that differ between patients with CSVD and controls—particularly forward head posture and reduced step regularity—provides potential biomarkers for early detection and monitoring of CSVD progression.

These findings have several important implications for clinical practice and future research. First, they suggest that smartphone-based gait assessment combined with video analysis could serve as a cost-effective screening tool for early CSVD detection, potentially enabling earlier intervention before significant cognitive decline occurs. Second, the quantifiable nature of these gait parameters enables objective monitoring of disease progression over time, which could serve as outcome measures in clinical trials of new therapeutic interventions. Third, understanding the specific gait adaptations in CSVD suggests targeted rehabilitation approaches that might focus on improving proprioception and balance to address the forward head posture observed in these patients.

Furthermore, the smartphone-based assessment method demonstrated in this study has potential for remote monitoring applications, allowing clinicians to track patients' gait parameters outside of clinical settings and providing a more ecological assessment of functional mobility. This approach could be particularly valuable in resource-limited settings or for patients with mobility restrictions that make frequent clinic visits challenging.

Future research should aim to validate these findings in larger, more diverse cohorts and explore the potential of longitudinal gait assessment to predict disease progression and response to interventions. In addition, the development of more user-friendly, automated analysis tools could facilitate the integration of comprehensive gait assessment into routine clinical care for patients with CSVD, potentially improving early detection, monitoring, and personalized treatment planning for this common but often underdiagnosed condition.

Acknowledgments

No generative AI tools were used in the conceptualization, experimental design, data analysis, or generation of original content for this article. A large language model was used to improve the readability and language. The authors affirm that they take full responsibility for all content and have verified all information presented in this paper.

XSL•FO
RenderX

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

References

1. Song MR, Qiao-shu W. A development history of cerebral small vessel disease. Chinese Journal of Contemporary Neurology & Neurosurgery 2016;16(11):736. [doi: 10.3969/j.issn.1672-6731.2016.11.003]
2. Mancuso M, Arnold M, Bersano A, et al. Monogenic cerebral small-vessel diseases: diagnosis and therapy. Consensus recommendations of the European Academy of Neurology. Eur J Neurol 2020 Jun;27(6):909-927. [doi: 10.1111/ene.14183] [Medline: 32196841]
3. Sharma B, Wang M, McCreary CR, Camicioli R, Smith EE. Gait and falls in cerebral small vessel disease: a systematic review and meta-analysis. Age Ageing 2023 Mar 1;52(3):afad011. [doi: 10.1093/ageing/afad011] [Medline: 37000039]
4. Kancheva AK, Wardlaw JM, Lyall DM, Quinn TJ. Clinical phenotypes associated with cerebral small vessel disease: an overview of systematic reviews. Neurology 2024 Apr;102(8):e209267. [doi: 10.1212/WNL.0000000000209267] [Medline: 38552192]
5. Hilal S, Mok V, Youn YC, Wong A, Ikram MK, Chen CLH. Prevalence, risk factors and consequences of cerebral small vessel diseases: data from three Asian countries. J Neurol Neurosurg Psychiatry 2017 Aug;88(8):669-674. [doi: 10.1136/jnnp-2016-315324] [Medline: 28600443]
6. Chojdak-Łukasiewicz J, Dziadkowiak E, Zimny A, Paradowski B. Cerebral small vessel disease: a review. Adv Clin Exp Med 2021 Mar;30(3):349-356. [doi: 10.17219/acem/131216] [Medline: 33768739]
7. Markus HS, de Leeuw FE. Cerebral small vessel disease: recent advances and future directions. Int J Stroke 2023 Jan;18(1):4-14. [doi: 10.1177/17474930221144911] [Medline: 36575578]
8. Blumen HM, Jayakody O, Verghese J. Gait in cerebral small vessel disease, pre-dementia, and dementia: a systematic review. Int J Stroke 2023 Jan;18(1):53-61. [doi: 10.1177/17474930221114562] [Medline: 35797006]
9. Pantoni L. Cerebral small vessel disease: from pathogenesis and clinical characteristics to therapeutic challenges. Lancet Neurol 2010 Jul;9(7):689-701. [doi: 10.1016/S1474-4422(10)70104-6] [Medline: 20610345]
10. Xu K, Wang Y, Jiang Y, et al. Analysis of gait pattern related to high cerebral small vessel disease burden using quantitative gait data from wearable sensors. Comput Methods Programs Biomed 2024 Jun;250:108162. [doi: 10.1016/j.cmpb.2024.108162] [Medline: 38631129]
11. Strongman C, Cavallerio F, Timmis MA, Morrison A. A scoping review of the validity and reliability of smartphone accelerometers when collecting kinematic gait data. Sensors (Basel) 2023 Oct 20;23(20):8615. [doi: 10.3390/s23208615] [Medline: 37896708]
12. Espy DD, Yang F, Bhatt T, Pai YC. Independent influence of gait speed and step length on stability and fall risk. Gait Posture 2010 Jul;32(3):378-382. [doi: 10.1016/j.gaitpost.2010.06.013] [Medline: 20655750]
13. Whittle MW. Clinical gait analysis: a review. Hum Mov Sci 1996 Jun;15(3):369-387. [doi: 10.1016/0167-9457(96)00006-1]
14. Brach JS, Perera S, Studenski S, Katz M, Hall C, Verghese J. Meaningful change in measures of gait variability in older adults. Gait Posture 2010 Feb;31(2):175-179. [doi: 10.1016/j.gaitpost.2009.10.002] [Medline: 19889543]
15. Verghese J, Holtzer R, Lipton RB, Wang C. Quantitative gait markers and incident fall risk in older adults. J Gerontol A Biol Sci Med Sci 2009 Aug;64(8):896-901. [doi: 10.1093/gerona/glp033] [Medline: 19349593]
16. Moe-Nilssen R, Helbostad JL. Estimation of gait cycle characteristics by trunk accelerometry. J Biomech 2004 Jan;37(1):121-126. [doi: 10.1016/s0021-9290(03)00233-1] [Medline: 14672575]
17. van Schooten KS, Pijnappels M, Rispens SM, et al. Daily-life gait quality as predictor of falls in older people: a 1-year prospective cohort study. PLoS ONE 2016;11(7):e0158623. [doi: 10.1371/journal.pone.0158623] [Medline: 27389583]
18. Tao W, Liu T, Zheng R, Feng H. Gait analysis using wearable sensors. Sensors (Basel) 2012;12(2):2255-2283. [doi: 10.3390/s120202255] [Medline: 22438763]
19. Alexander NB. Gait disorders in older adults. J Am Geriatr Soc 1996 Apr;44(4):434-451. [doi: 10.1111/j.1532-5415.1996.tb06417.x] [Medline: 8636592]
20. Morton SM, Bastian AJ. Mechanisms of cerebellar gait ataxia. Cerebellum 2007;6(1):79-86. [doi: 10.1080/14734220601187741] [Medline: 17366269]
21. de Laat KF, Tuladhar AM, van Norden AGW, Norris DG, Zwiers MP, de Leeuw FE. Loss of white matter integrity is associated with gait disorders in cerebral small vessel disease. Brain (Bacau) 2011 Jan;134(Pt 1):73-83. [doi: 10.1093/brain/awq343] [Medline: 21156660]
22. de Laat KF, van den Berg HAC, van Norden AGW, Gons RAR, Olde Rikkert MGM, de Leeuw FE. Microbleeds are independently related to gait disturbances in elderly individuals with cerebral small vessel disease. Stroke 2011 Feb;42(2):494-497. [doi: 10.1161/STROKEAHA.110.596122] [Medline: 21164137]

XSL•FO
RenderX

23. Zong LX, et al. Gait and balance disorders in 57 patients with cerebral small vessel disease. Chin J Stroke 2015;10(12):1000-1005 [FREE Full text]
24. Pratheepan Y, Condell JV, Prasad G. The use of dynamic and static characteristics of gait for individual identification. : IEEE Presented at: 2009 13th International Machine Vision and Image Processing Conference. Dublin, Ireland.
25. Derawi MO, Ali H, Cheikh FA. Gait recognition using time-of-flight sensor. Presented at: BIOSIG 2011–Proceedings of the Biometrics Special Interest Group; Sep 8-9, 2011; Darmstadt, Germany URL: https://dl.gi.de/items/09c926d6-7242-4a1d-a112-7b4ccf1774f4 [accessed 2025-07-18]
26. Muro-de-la-Herran A, Garcia-Zapirain B, Mendez-Zorrilla A. Gait analysis methods: an overview of wearable and non-wearable systems, highlighting clinical applications. Sensors 2014 Feb 19;14(2):3362-3394. [doi: 10.3390/s140203362] [Medline: 24556672]
27. Zhong R, Rau PLP. A mobile phone-based gait assessment app for the elderly: development and evaluation. JMIR Mhealth Uhealth 2020 Feb 29;8(2):e14453. [doi: 10.2196/14453] [Medline: 32452821]
28. Tao S, Zhang H, Kong L, Sun Y, Zhao J. Validation of gait analysis using smartphones: reliability and validity. Digit Health 2024;10:20552076241257054. [doi: 10.1177/20552076241257054] [Medline: 38817844]
29. Cao Z, Hidalgo G, Simon T, Wei SE, Sheikh Y. OpenPose: realtime multi-person 2D pose estimation using part affinity fields. IEEE Trans Pattern Anal Mach Intell 2019;43(1):172-186. [doi: 10.1109/TPAMI.2019.2929257]
30. Verghese J, LeValley A, Hall CB, Katz MJ, Ambrose AF, Lipton RB. Epidemiology of gait disorders in community-residing older adults. J Am Geriatr Soc 2006 Feb;54(2):255-261. [doi: 10.1111/j.1532-5415.2005.00580.x] [Medline: 16460376]
31. Viswakumar A, Rajagopalan V, Ray T, Parimi C. Human gait analysis using openpose. : IEEE Presented at: 2019 Fifth International Conference on Image Information Processing (ICIIP). Shimla, India. [doi: 10.1109/ICIIP47207.2019.8985781]
32. Ota M, Tateuchi H, Hashiguchi T, Ichihashi N. Verification of validity of gait analysis systems during treadmill walking and running using human pose tracking algorithm. Gait Posture 2021 Mar;85:290-297. [doi: 10.1016/j.gaitpost.2021.02.006]
33. D'Antonio E, Taborri J, Mileti I, Rossi S, Patane F. Validation of a 3D markerless system for gait analysis based on OpenPose and two RGB webcams. IEEE Sensors J 2021;21(15):17064-17075. [doi: 10.1109/JSEN.2021.3081188]
34. Baker R. Measuring Walking: A Handbook of Clinical Gait Analysis: Mac Keith Press; 2013. URL: https://journals.lww.com/pedpt/citation/2014/26040/measuring_walking__a_handbook_of_clinical_gait.21.aspx [accessed 2025-07-18]
35. Raîche M, Hébert R, Prince F, Corriveau H. Screening older adults at risk of falling with the Tinetti balance scale. The Lancet 2000 Sep;356(9234):1001-1002. [doi: 10.1016/S0140-6736(00)02695-7]
36. Newell AM, VanSwearingen JM, Hile E, Brach JS. The modified Gait Efficacy Scale: establishing the psychometric properties in older adults. Phys Ther 2012 Feb;92(2):318-328. [doi: 10.2522/ptj.20110053] [Medline: 22074940]

Abbreviations

ABC: Activities-Specific Balance Confidence
CSVD: cerebral small vessel disease
MRI: magnetic resonance imaging
RMS: root mean square
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by A Mavragani; submitted 27.03.24; peer-reviewed by N Yu, X Lei; revised version received 14.04.25; accepted 23.04.25; published 28.07.25.

Please cite as:
Lai X, Qiao LY, Rau PLP, Liu Y
Gait Disturbances in Older Adults With Cerebral Small Vessel Disease: Mixed Methods Study Using Smartphone Sensors and Video Analysis
JMIR Form Res 2025;9:e58864
URL: https://formative.jmir.org/2025/1/e58864
doi:10.2196/58864

© Xiaojun Lai, Li-Yan Qiao, Pei-Luen Patrick Rau, Yankuan Liu. Originally published in JMIR Formative Research (https://formative.jmir.org), 28.7.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Barriers to Patient Portal Adoption Among a Bilingual Patient Population by Analysis of Survey Findings from English- and Spanish-Speaking Patients: Information Needs Study

Jiahua Yang[1*], PhD; Michael Mackert[1,2,3*], PhD; Daniela De Luca[1*], PhD; Sophia Annette Dove[1*], MA

[1]Stan Richards School of Advertising and Public Relations, University of Texas at Austin, 300 W. Dean Keeton St., Austin, TX, United States

[2]Dell Medical School, University of Texas at Austin, Austin, TX, United States

[3]Center for Health Communication, University of Texas at Austin, Austin, TX, United States

[*]all authors contributed equally

Corresponding Author:
Sophia Annette Dove, MA
Stan Richards School of Advertising and Public Relations, University of Texas at Austin, 300 W. Dean Keeton St., Austin, TX, United States

Abstract

Background: Despite legislative action, pre-existing barriers continue to prevent patients from using patient portals. Patients, especially older people, people of color, and people with limited English proficiency continue to experience difficulty in adopting patient portals.

Objective: The aim of this study was to advance understanding, explore willingness to adopt an electronic portal, and examine differences between language preferences.

Methods: English- and Spanish-speaking patients (N=106) were surveyed from a community clinic regarding access to electronic devices and the internet, barriers to using a patient portal, willingness to adopt such a portal, preference mode of communication with health care providers, and preferred features in the current clinic's portal. Linear and logistic regressions were performed to predict the probability that patients would adopt the patient portal.

Results: Only 65% (n=69)of participants said they envisioned themselves using a patient portal. English-speaking patients were more willing to exchange electronic information with their health care providers. Spanish-speaking patients reported language as a significant barrier to portal use. A logistic regression revealed that patients with more positive attitudes and higher perceived behavioral control are more likely to sign up and use the patient portal (Nagelkerke R^2=.51, classification=90.8%, efficacy B=2.38, Wald-1=5.93, P=.02 and Exp[B]=12.44, attitude B=1.87, Wald=6.45, P=.01, Exp[B]=7.49).

Conclusions: Understanding language preference differences while predicting portal use based on attitudes and perceptions empowers patients to have a more meaningful experience with their physician, potentially overcoming low health literacy–related barriers.

(JMIR Form Res 2025;9:e66717) doi:10.2196/66717

KEYWORDS

patient portal adoption; patient portals; health communication; patient access to information; patient information; electronic health records; EHRs; electronic medical records; EMRs; patient record; health records; personal health record; PHR

Introduction

Beginning in 2000, the Office of Civil Rights of the Department of Health and Human Services required physicians to comply with requests for copies of medical records [1]. By 2014, physicians participating in the Electronic Health Record Incentive Program were expected to transition to electronic medical recordkeeping, and finally, the 2021 Cures Act granted patients near immediate access to their medical records and prohibited practices that interfere with the access, exchange, or use of electronic health information [1-6]. Despite legislative action to protect the rights of patients, there remain pre-existing barriers that prevent patients from understanding, signing up for, and using patient portals.

Despite patients reporting generally high access and perceived usefulness attitudes toward patient portals, especially for administrative tasks such as reviewing lab results, patients continue to experience barriers in adopting patient portals [7,8]. Those who use patient portals less frequently tend to be older adults, people of color, adolescents, and people with limited English proficiency (LEP) [9-13]. Previous research has shown that race and ethnicity are the strongest predictors when

XSL•FO
RenderX

investigating portal use, especially for Latinx individuals, as reports revealed that people of color and LEP were offered access to patient portals at much lower rates yet are more likely to download and transmit health information, highlighting the need for providers to promote portal use [4,10-12].

In order to provide the best patient portal experience, a multifaceted, multilingual portal platform is necessary. Understanding the limitations of existing clinical electronic systems is critical, as a complete redesign of portals is impractical. Education on current platforms and consideration of language in messages can help alleviate confusion and increase portal uptake [9,12]. Previous studies suggest that patients of color prefer face-to-face interaction with their providers; therefore, it may be beneficial to leverage existing relationships to provide training on the proper use and promotion of existing patient portals [4,11,13,14]. Provider endorsement increases portal use and patient empowerment and self-efficacy [5,8,10,13,15].

Patient portals can help patients interact with their health care team, manage appointments, access medical notes, and have been shown to improve medication adherence and health care quality [4,6,9,16,17]. A 2017 study revealed patient portal use was found to decrease emergency visits and hospital admissions [17]. On some clinical notes sharing platforms, such as Our Notes, patients have the ability to work alongside physicians by contributing their own notes to their digital medical records, resulting in an increase of medication and treatment plan adherence [5,6]. This shared work increases the patients' perceived ability to control their health, motivating further engagement with patient portals [6,10,15]. In addition, physicians can overcome distrust toward the health care system, mainly felt in patients of color, by increasing communication quality [4,7,14]. Engaging in meaningful conversation also combats the barriers of low health and technology literacy that is largely observed in the older population [6,9,11].

The purpose of this paper was to advance understanding of potential barriers to patient portal use by exploring patients' willingness to adopt an electronic portal. This understanding is achieved by measuring attitudes and perceived expectations leading to intention and examining differences between distinct linguistic backgrounds, specifically for patients who might experience language and technological barriers. Based on previous research, we hypothesize a presence of distinct differences in barriers to patient portal adoption between clinic patients based on preferred language (English and Spanish). The rest of this paper provides an overview of study methods, results, and a discussion of implications for future research and practice to increase adoption of a potentially important health technology by understanding barriers to use.

Methods

Data Collection

This study relied on convenience sampling of patients from a nonprofit Federally Qualified Health Center (FQHC) serving marginalized community members from June 2022 to July 2022. There was no existing user of the patient portal as the study was conducted prior to the full launch to improve the design of the patient portal. The access to the digital survey was published on the clinic's website and the recruitment material was available at the clinic's waiting room with a QR code directing to the digital survey. The survey took about 7 minutes to finish. Those who agreed had the chance to opt into a drawing for 20 gift cards (US $25 each). All questions were translated into Spanish by certified translators and double-checked by a native Spanish speaker graduate research assistant.

Design

The survey asked questions regarding participants' access to electronic devices and the internet, barriers to using a patient portal, willingness to adopt such a portal, preference mode of communication with health care providers, and preferred features in the current clinic's portal. To determine the participants' preferred language, participants were asked what their primary language was, and more specifically, what language they speak most of the time. Potential barriers were identified from previous literature as well as input from the clinical staff's experience [4,6,11,18,19]. For the full survey instrument, please see Multimedia Appendix 1.

Grounded in the Theory of Planned Behavior, a psychological framework focused on behavior prediction, which states intention as the strongest predictor to behavior measured by the attitude toward the behavior, perceived behavioral norms, and perceived behavioral control shown in [20] (Textbox 1). Measures were adopted from previous studies [20-22] and modified to fit the research context of interest and participants' literacy level per clinic staff's feedback.

XSL•FO
RenderX

Textbox 1. Theory of Planned Behavior measures

Attitude

- Using the Patient Portal will support critical aspects of my health care.
- Using the Patient Portal will enhance my effectiveness in managing my health care.
- Overall, the Patient Portal will be useful in managing my health care.

Perceived behavioral norms

- People who are important to me think that I should use the Patient Portal.
- People who influence my behavior think that I should use the Patient Portal.
- People whose opinions that I value prefer that I use the Patient Portal.

Perceived behavioral control

- Learning how to use the Patient Portal is easy for me.
- My interaction with the Patient Portal is clear and understandable.
- I find the Patient Portal easy to use.It is easy for me to become skillful at using the Patient Portal.

Data Analysis

A series of independent-sample *t* tests and χ^2 analyses were performed to identify differences among English- and Spanish-speaking participants. A logistic regression analysis was performed to investigate the relationship between attitudes, perceived behavioral control, and perceived behavioral norms with the intention to adopt the patient portal. Age and gender were controlled in the logistic regression.

Ethical Considerations

The University of Texas at Austin Institutional Review Board determined this study as exempt and did not require further review (STUDY00002291). Participants reviewed and acknowledged informed consent materials prior to completing the electronic survey. To compensate participants, each person was entered into a drawing to win one of the 20 gift cards (US $25 each). For privacy and confidentiality, all study data were deidentified.

Results

The demographic information of the participants is shown in Table 1. The majority of respondents (96/106, 93%) reported having good internet access, with 91/106 (88%) having access at home and 64/106 (60%) respondents reported owning an internet-enabled device. Specifically, 80 (75%) participants reported accessing the internet at home, 66 (62%) through mobile phones, 24 (22%) at work, 21/106 (20%) via a tablet, 7 out of 106 (7%) at school, 6/106 (6%) public spaces such as a cafe, 6 (6%) from someone else's home, 4 out of 106 (4%) at a public library, and 1 participant (0.9%) at a community center.

Regarding potential barriers to using an electronic patient portal, 29 (27%) participants reported preferring in-person visits as their main concern, followed by 25 (23%) worry about privacy, 14 (13%) lack of access to the internet, 10 (9%) language difficulties, 7 (7%) lack of comfort with computers and internet-enabled devices, and 3 (3%) participants reported poorly designed or implemented portals.

Specific to this clinic's patient portal, only 69 (65%) of participants said they saw themselves signing up and using it, and 8 (8%) reported not being sure.

A series of analyses revealed differences among English- and Spanish-speaking participants. An independent-sample *t* test showed that those who speak English were more willing to get electronic information about symptoms than Spanish speakers (P=.006). Further, χ^2 analyses showed significant association between language and internet access, communication with doctor preferences, barriers to using a portal, and features that participants preferred about the clinic's patient portal (Table 2).

Based on the Theory of Planned Behavior, Figure 1 shows the results of correlations and the logistic regression. Overall, a logistic regression revealed that patients with more positive attitudes and higher perceived behavioral control are more likely to sign up and use a patient portal (Nagelkerke R^2=.51, classification=90.8%, control B=2.38, odds ratio[OR] 10.81, P=.02; attitude B=1.87, OR 6.52, P=.02); however, the relationship between perceived norm and sign-up was not significant (norm B=−1.04, OR .35, P=.27).

XSL•FO
RenderX

Table . Baseline demographic details of study population.

Baseline demographics	Participants (N=107)[a]
Sex, n (%)	
Female	90/106 (85)
Male	15/106 (14)
Chose not to disclose	1/106 (1)
Race or ethnicity[b], n (%)	
African American or Black	16/109 (15)
Asian, Asian Indian, or Asian American	1/109 (1)
American Indian or Alaska Native	3/109 (3)
White	19/109 (17)
Hispanic or Latino or Latina	70/109 (65)
Age (years), mean (SD)	40 (15.35)
Language, n (%)	
English	67/107 (63)
Spanish	40/107 (37)

[aa]107 participants completed at least 60% of the survey instrument; however, only 106 completed the entire survey instrument.

[b]Participants were able to select multiple race or ethnicity options.

XSL•FO
RenderX

Table . Chi-square analyses to examine language associations (N=106).

Variables	Total (N=106), n (%)	English-speaking (n=67), n (%)	Spanish-speaking (n=40), n (%)	χ^2 (df)	*P* value
Internet access and device ownership					
Internet access	96 (93)	64 (100)	32 (82)	12.32[a]	<.001
Internet access at home	91 (88)	62 (95)	29 (74)	9.85[a]	.002
Computer access at home	64 (62)	47 (73)	17 (43)	9.95[a]	.002
Internet use, home	27 (75)	56 (84)	24 (60)	7.38[a]	.007
Internet use, work	24 (22)	20 (30)	4 (10)	5.67[a]	.02
Owning personal computers	48 (45)	40 (60)	8 (20)	15.96[a]	<.001
Communication with doctor preferences					
In-person	63 (59)	34 (51)	29 (73)	4.90[b]	.03
Telephone	52 (49)	41 (61)	11 (28)	11.38[a]	<.001
Email	33 (31)	26 (39)	7 (18)	5.33[a]	.02
Barriers to using an electronic patient portal					
Privacy concerns	25 (23)	22 (33)	3 (8)	8.98[a]	.003
Language	10 (9)	3 (5)	7 (18)	5.01[b]	.03
Most liked portal features					
Paperwork and records	53 (50)	39 (58)	14 (35)	5.40[a]	.02
Asking for medical refills	46 (43)	35 (52)	11 (28)	6.25[a]	.01
Send questions to provider online	44 (41)	36 (54)	8 (20)	11.77[a]	.001

[a]English-speaking participants were overrepresented.

[b]Spanish-speaking participants were overrepresented.

Figure 1. Theory of Planned Behavior correlations.

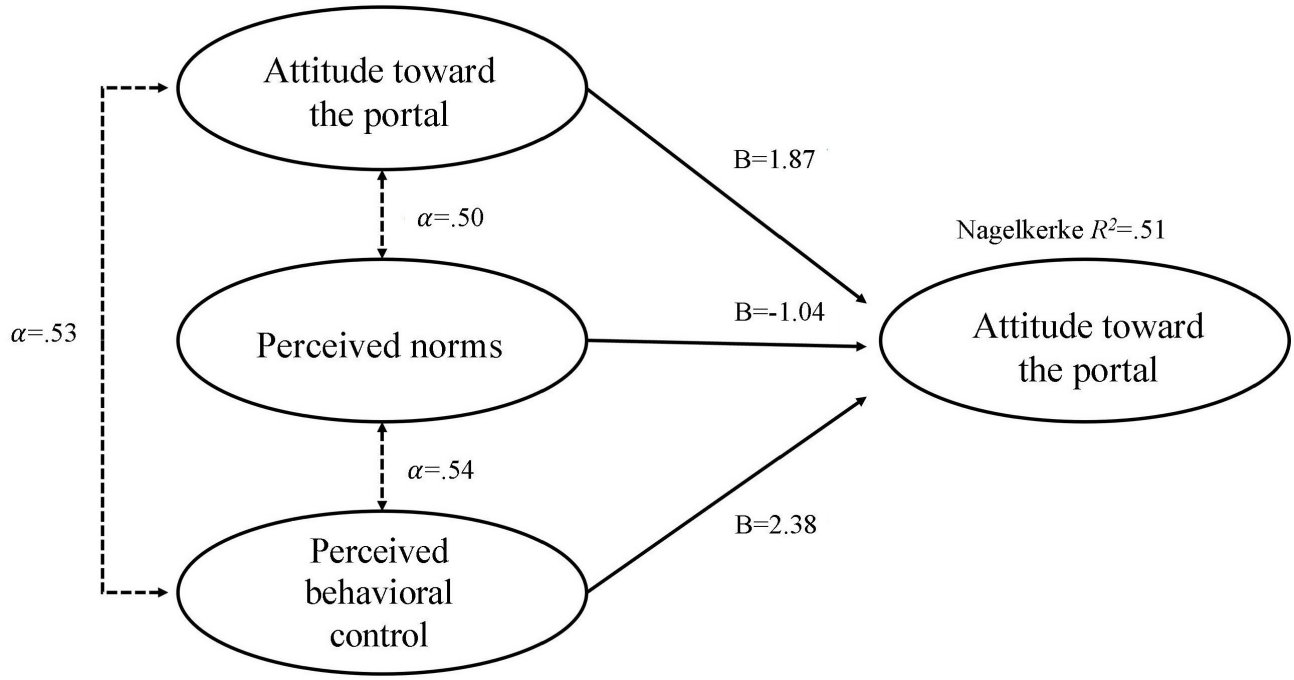

XSL•FO
RenderX

Discussion

Primary Barriers to Patient Portal Adoption

The differences between English-speaking and Spanish-speaking participants revealed further insights such as Spanish-speaking participants' lower willingness to exchange electronic information than English-speaking participants. Even so, compared to English-speaking participants, Spanish-speaking participants prefer in-person communication with their physicians. This finding, in conjunction with the language barrier, supports previous research where more emphasis on portal training for patients with LEP is needed to increase comfort and combat distrust [4].

Historically, patient portal uptake has been challenging due to a lack of access to technology [6]. However, the findings of this study revealed many patients have good internet access and device availability. Despite adequate access to the technology, only 65% of participants said they envisioned themselves signing up for and using the patient portal. While this is a fairly strong intention response, there are still barriers worth acknowledging, such as a stronger preference for in-person communication with their physician, concerns about privacy, language difficulties, lack of comfort with portals, and poor design. The most prominent barrier reported was the preference for in-person communication with their doctor, highlighting an opportunity for physicians and staff to emphasize the potential benefit of completing administrative tasks through portal use such as scheduling appointments, sending secure messages, and reviewing lab results.

Comparison to Prior Work

Many studies have investigated the uptake of patient portals in Hispanic populations; however, often in specialized areas of care such as asthma or with small sample sizes [8,12,19]. Few have investigated patients' willingness to adopt an electronic portal and assess potential barriers while specifically exploring the impact of language preferences and measuring behavioral intention. This study successfully fills such a gap, as it surveyed English- and Spanish-speaking participants from a community clinic in a Southwestern city to identify patient portal uptake barriers based on language preferences and measured differences in their intention to sign up and use the portal. Understanding the existing barriers is the first step in strategizing ways to overcome obstacles patients experience when considering a patient portal.

Limitations and Considerations

The Theory of Planned Behavior is criticized for simply measuring the intentions to predict actual behavior; however, this study uniquely addresses this by simultaneously identifying barriers and measuring patients' attitudes and perceptions (perceived norms and perceived behavioral control) toward portal adoption. The first notable limitation is that while this study did not measure actual adoption of the clinic's patient portal, the findings still provide insight into the challenges faced by a multilingual patient population. Specifically, by understanding attitudes and perceptions along with language preference differences, we can tailor approaches and predict how patients with higher perceived behavioral control and more positive attitudes toward the patient portal were most likely to adopt and use the portal. The second limitation is the reliance on convenience sampling limiting the ability to make broad generalizations from the findings; however, the strength of this study is that it helps understand the need for further assessment on ways to increase perceived behavioral control and positive attitudes toward patient portals through in-person relationships, particularly among patients with LEP.

Future Directions

The study's findings identify areas for health care organizations to focus on when seeking to increase patient portal uptake among a linguistically diverse population. Future research can investigate in-person preferences as well as communication surrounding privacy concerns, which were the top 2 barriers expressed by participants in this study. Empowering patients to have more autonomy over their health and decision-making leads to a more meaningful experience with their physicians, which can tackle low health literacy–related barriers. Yet, understanding these differences based on language preference confirms that a more targeted approach is needed. These findings add to the current body of literature by revealing that barriers such as lack of access require more investigation, as even with adequate access to internet and technology, there is still a need for patients to feel secure with the level of privacy and ease of use regarding portal design.

Conclusions

To understand the willingness to adopt an electronic portal and examine differences between language preferences, this study presented survey findings from a FQHC regarding access to electronic devices and the internet, and identified barriers to using a patient portal. We conducted a series of independent-sample t tests and χ^2 analyses to identify differences among English- and Spanish-speaking participants. A logistic regression analysis was performed to investigate the relationship between attitudes, perceived behavioral control, and perceived behavioral norms with the intention to adopt the patient portal. Despite adequate access to technology, only 65% of participants said they envisioned themselves using a patient portal. English-speaking patients were more willing to exchange electronic information with their health care providers. Spanish-speaking patients reported language as a significant barrier to portal use. Finally, patients with more positive attitudes and higher perceived behavioral control are more likely to sign up and use the patient portal. These findings can inform strategies to achieve a more meaningful experience between physician and patients with potential to tailor approaches to patients based on preferred language.

XSL•FO
RenderX

Acknowledgments

A special thanks to Sahithi Kasa for the assistance and support in this research.

Data Availability

The datasets generated and analyzed during this study are available upon reasonable request from the corresponding author.

Authors' Contributions

JY and MM took the lead on conceptualization. JY developed the methodology and conducted the formal analysis. JY, MM, DDL, and SAD contributed to portions of the original draft. SAD conducted the review and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Survey instrument.
[DOCX File, 17 KB - formative_v9i1e66717_app1.docx]

References

1. ONC's cures act final rule. Assistant Secretary for Technology Policy/Office of the National Coordinator for Health IT. 2020. URL: https://www.healthit.gov/topic/oncs-cures-act-final-rule [accessed 2025-03-09]
2. EHR incentive programs 2014 CEHRT rule: quick guide. CMS. 2014. URL: https://www.cms.gov/regulations-and-guidance/legislation/ehrincentiveprograms/downloads/cehrt2014_finalrule_quickguide.pdf [accessed 2025-03-09]
3. Dworkowitz A. Provider obligations for patient portals under the 21st century cures act. HealthAffairs. 2022. URL: https://www.healthaffairs.org/do/10.1377/forefront.20220513.923426 [accessed 2025-06-25]
4. Richwine C, Johnson C, Patel V. Disparities in patient portal access and the role of providers in encouraging access and use. J Am Med Inform Assoc 2023 Jan 18;30(2):308-317. [doi: 10.1093/jamia/ocac227] [Medline: 36451262]
5. Griffin AC, Troszak LK, Van Campen J, Midboe AM, Zulman DM. Tablet distribution to veterans: an opportunity to increase patient portal adoption and use. J Am Med Inform Assoc 2022 Dec 13;30(1):73-82. [doi: 10.1093/jamia/ocac195] [Medline: 36269168]
6. Fant C, Adelman D, Zak C. Post-cures act: implications for nurse practitioners. J Nurse Pract 2022 Nov;18(10):1102-1106. [doi: 10.1016/j.nurpra.2022.04.001]
7. Sakumoto M, Ye J, Kalu R, et al. Patient portal perceptions in an urban community health center setting: insights for telehealth. Telehealth Med Today 2022;7(5):1-29. [doi: 10.30953/thmt.v7.373]
8. Ochoa A 3rd, Kitayama K, Uijtdehaage S, et al. Patient and provider perspectives on the potential value and use of a bilingual online patient portal in a Spanish-speaking safety-net population. J Am Med Inform Assoc 2017 Nov 1;24(6):1160-1164. [doi: 10.1093/jamia/ocx040] [Medline: 28460130]
9. Perzynski AT, Roach MJ, Shick S, et al. Patient portals and broadband internet inequality. J Am Med Inform Assoc 2017 Sep 1;24(5):927-932. [doi: 10.1093/jamia/ocx020] [Medline: 28371853]
10. Casillas A, Perez-Aguilar G, Abhat A, et al. Su salud a la mano (your health at hand): patient perceptions about a bilingual patient portal in the Los Angeles safety net. J Am Med Inform Assoc 2019 Dec 1;26(12):1525-1535. [doi: 10.1093/jamia/ocz115] [Medline: 31373362]
11. Saif N, Blackstone S, Sebring A, Compton R. Improving electronic portal use in groups with low utilization at a family medicine practice. Online J Nurs Inform 2022;26(2):1-14 [FREE Full text]
12. Chu JN, Sarkar U, Rivadeneira NA, Hiatt RA, Khoong EC. Impact of language preference and health literacy on health information-seeking experiences among a low-income, multilingual cohort. Patient Educ Couns 2022 May;105(5):1268-1275. [doi: 10.1016/j.pec.2021.08.028] [Medline: 34474924]
13. Goel MS, Brown TL, Williams A, Hasnain-Wynia R, Thompson JA, Baker DW. Disparities in enrollment and use of an electronic patient portal. J Gen Intern Med 2011 Oct;26(10):1112-1116. [doi: 10.1007/s11606-011-1728-3] [Medline: 21538166]
14. Lyles CR, Sarkar U, Ralston JD, et al. Patient-provider communication and trust in relation to use of an online patient portal among diabetes patients: the Diabetes and Aging Study. J Am Med Inform Assoc 2013;20(6):1128-1131. [doi: 10.1136/amiajnl-2012-001567] [Medline: 23676243]
15. Bajracharya AS, Crotty BH, Kowoloff HB, Safran C, Slack WV. Patient experience with family history tool: analysis of patients' experience sharing their family health history through patient-computer dialogue in a patient portal. J Am Med Inform Assoc 2019 Jul 1;26(7):603-609. [doi: 10.1093/jamia/ocz008] [Medline: 30946464]

XSL•FO
RenderX

16. Avdagovska M, Menon D, Stafinski T. Capturing the impact of patient portals based on the quadruple aim and benefits evaluation frameworks: scoping review. J Med Internet Res 2020 Dec 8;22(12):e24568. [doi: 10.2196/24568] [Medline: 33289677]
17. Sorondo B, Allen A, Fathima S, Bayleran J, Sabbagh I. Patient portal as a tool for enhancing patient experience and improving quality of care in primary care practices. EGEMS (Wash DC) 2016;4(1):1262. [doi: 10.13063/2327-9214.1262] [Medline: 28203611]
18. Patel PD, Cobb J, Wright D, et al. Rapid development of telehealth capabilities within pediatric patient portal infrastructure for COVID-19 care: barriers, solutions, results. J Am Med Inform Assoc 2020 Jul 1;27(7):1116-1120. [doi: 10.1093/jamia/ocaa065] [Medline: 32302395]
19. Localio AM, Klusaritz H, Morales KH, Ruggieri DG, Han X, Apter AJ. Primary language and the electronic health record patient portal: barriers to use among Spanish-speaking adults with asthma. J Asthma 2022 Oct;59(10):2081-2090. [doi: 10.1080/02770903.2021.1989462] [Medline: 34634975]
20. Ajzen I. Constructing a theory of planned behavior questionnaire. : UMass Amherst; 2006 URL: http://www.people.umass.edu/aizen/pdf/tpb.measurement.pdf [accessed 2025-07-07]
21. Ajzen I. Perceived behavioral control, self‐efficacy, locus of control, and the theory of planned behavior. J Applied Social Pyschol 2002 Apr;32(4):665-683. [doi: 10.1111/j.1559-1816.2002.tb00236.x]
22. Chu PY, Chiu JF. Factors influencing household waste recycling behavior: test of an integrated model. J Applied Social Pyschol 2003 Mar;33(3):604-626. [doi: 10.1111/j.1559-1816.2003.tb01915.x]

Abbreviations

FQHC: Federally Qualified Health Center
LEP: limited English proficiency
OR: odds ratio

Edited by A Mavragani; submitted 20.09.24; peer-reviewed by C Richwine, M Sakumoto; revised version received 20.05.25; accepted 21.05.25; published 28.07.25.

Please cite as:
Yang J, Mackert M, De Luca D, Dove SA
Barriers to Patient Portal Adoption Among a Bilingual Patient Population by Analysis of Survey Findings from English- and Spanish-Speaking Patients: Information Needs Study
JMIR Form Res 2025;9:e66717
URL: https://formative.jmir.org/2025/1/e66717
doi:10.2196/66717

© Jiahua Yang, Michael Mackert, Daniela M De Luca, Sophia Annette Dove. Originally published in JMIR Formative Research (https://formative.jmir.org), 28.7.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Toward a Smartphone-Based and Conversational Agent–Delivered Just-in-Time Adaptive Holistic Lifestyle Intervention for Older Adults Affected by Cognitive Decline: Two-Week Proof-of-Concept Study

Esther Brill[1*], PhD; Rasita Vinay[2,3,4*], PhD; Marcia Nißen[3,4,5*], PhD; Priyam Joshi[6], MPH; Stefan Klöppel[1*], MD; Tobias Kowatsch[3,4,5*], PhD

[1]University Hospital of Old Age Psychiatry and Psychotherapy, University of Bern, Bern, Switzerland

[2]Institute of Biomedical Ethics and History of Medicine, University of Zurich, Zurich, Switzerland

[3]School of Medicine, University of St. Gallen, St. Gallen, Switzerland

[4]Centre for Digital Health Interventions, Department of Management, Technology, and Economics, ETH Zurich, Weinbergstrasse 56/58, Zurich, Switzerland

[5]Institute for Implementation Science in Health Care, University of Zurich, Zurich, Switzerland

[6]Department of Psychology, Georgia State University, Atlanta, GA, United States

[*]these authors contributed equally

Corresponding Author:
Rasita Vinay, PhD
Institute of Biomedical Ethics and History of Medicine, University of Zurich, Zurich, Switzerland

Abstract

Background: Dementia is projected to impact 152 million people by 2050, making it one of the most pressing global health challenges. The neurodegenerative process initiates well before clinical symptoms manifest, advancing from subjective cognitive decline (SCD) to mild cognitive impairment (MCI) and ultimately to dementia. Despite the growing prevalence, awareness of dementia prevention is limited, and many individuals express a desire to cease living upon diagnosis. Lifestyle interventions can mitigate cognitive decline, but there is a need for effective, scalable approaches to deliver these interventions to older adults. Digital health interventions, such as app-based just-in-time adaptive interventions, offer a promising solution, but their application in cognitively impaired older populations remains underexplored.

Objective: This formative study evaluated the plausibility, acceptability, and adherence to a smartphone-based just-in-time adaptive digital lifestyle intervention delivered by a rule-based conversational agent (CA) among older adults with SCD or MCI. The primary focus was on adherence to the CA-initiated conversational turns (measured objectively via interaction logs), and secondary objectives included perceptions of technology acceptance, working alliance with the CA, self-reported adherence to the suggested health-promoting activity, and feedback for future improvements (through a questionnaire and short interview).

Methods: This monocentric study investigated 15 participants (mean age 70.3, SD 5.01; 10 female and 5 male participants) with SCD (n=12) or MCI (n=3). Participants used the study app that delivered daily health-promoting activities through a CA over 2 weeks. Participants received notifications to engage in 7 health-related activities, and adherence to the activities was self-reported. Post intervention, participants rated their experience with the app and assessed their working alliance with the CA through the 6-item session alliance inventory. Data on smartphone use, demographic information, and cognitive performance (via Montreal Cognitive Assessment) were collected during a preintervention visit.

Results: Participants rated the study app positively, especially regarding ease of use and a subset of the working alliance. Adherence to the CA-initiated conversational turn was measured at an average of 81% across 14 days. In total, 27% (mean 4.07, SD 2.27) of participants indicated being vulnerable, and 100% then responded with their state of receptivity, of which 83% (mean 3.14, SD 1.61) were receptive to completing the activity, and 69% (mean 2.86, SD 1.70) self-reported adherence to the activity. There was no significant decline in adherence across the study period. Qualitative results support these findings and present two emerging themes: app enjoyment and enhancing engagement.

Conclusions: This study demonstrates that smartphone-based just-in-time adaptive interventions are feasible and generally well-accepted by older adults with SCD or MCI. However, the findings underscore the need for robust technological infrastructure and potential personal assistance to optimize adherence. Future interventions could benefit from integrating wearables to improve real-time engagement and accurately monitor adherence, ultimately supporting healthy aging and cognitive health in older populations.

XSL•FO
RenderX

(JMIR Form Res 2025;9:e66885) doi:10.2196/66885

KEYWORDS

conversational agents; older adults; just-in-time adaptive intervention; health intervention; smartphone-based; mHealth; mobile health; digital health; smartphone; digital; technology; single-arm feasibility study; feasibility; dementia; global health; cognitive impairment; awareness; dementia prevention; digital health interventions; older person; aging; mobile phone

Introduction

Dementia is expected to become one of the most critical global health challenges, with an estimated prevalence of 152 million patients worldwide by 2050 [1]. The neurodegenerative cascade in dementia commences well before clinically significant symptoms appear and can be conceptualized as a continuum extending from subjective cognitive decline (SCD) in the preclinical phase to mild cognitive impairment (MCI) to more advanced stages of the disease [2]. Recent international data suggest that approximately one in four individuals aged 60 years and older experience SCD [3], representing a significant risk factor with a subsequent conversion rate of 27% from SCD to MCI and a 14% progression from SCD to dementia within a decade [4,5]. Crucially, in Switzerland, 25% of individuals surveyed by the Swiss "Demenzbarometer" expressed a preference to cease living upon a dementia diagnosis, reflecting profound personal and societal implications [6].

Despite growing evidence from high-quality randomized controlled trials that indicates that lifestyle interventions can improve or maintain quality of life (QoL), slow cognitive decline, and even mitigate epigenetic age acceleration [7-9], awareness of dementia prevention remains limited: one-third of individuals living in Switzerland are unaware that lifestyle choices can impact dementia risk [6]—a gap underscoring the need for accessible, scalable approaches that promote health literacy and support everyday health behavior change.

Indeed, in recent years, a growing number of mobile and digital health interventions [10,11] have been developed to address modifiable risk factors associated with cognitive decline. Notable examples include the Maintain Your Brain platform [12-14] or the GrayMatters app [15-17], which combines psychoeducational content with self-tracking options across multiple domains such as nutrition, cognition, physical activity, and sleep. While these interventions show promising results (eg, with regard to cognitive decline, diet quality, and physical activity), they are often static, requiring high user initiative and offering limited real-time personalization or behavioral guidance adapted to an individual's changing needs or contexts [18].

Smartphone-based, just-in-time adaptive interventions (JITAIs) offer a promising alternative for delivering personalized behavioral support when individuals need it [19]. JITAIs are "interventions that adapt over time to an individual's changing status and circumstances with the goal to address the individual's need for support, whenever this need arises" [20]. They are, by definition, designed to provide tailored support to individuals during states of vulnerability or opportunity (eg, moments of adverse health behavior), receptivity (eg, when they can receive and process the support provided), and adherence (eg, using or adopting the support) [19,21]. They are also considered underdeveloped and underresearched, with a recent review identifying them as "a significant missed opportunity" for tailoring digital interventions [18,22].

While early research on JITAIs has demonstrated some efficacy in targeting lifestyle behaviors [23,24], they have so far primarily been designed and tested in young, healthy, and cognitively intact populations [25,26] in diverse settings or among different demographic groups [27]. Yet, their implementation in older populations affected by cognitive decline, especially given the added barriers of reduced digital literacy and potential accessibility constraints [28,29], presents unique challenges. Individuals with SCD or MCI may experience difficulties with abstract app navigation or complex content, which can compromise engagement [30].

In this context, relational agents—software that is designed to emulate conversations through voice or SMS text messages, also known as chatbots or digital coaches [31]—offer a compelling interface for delivering JITAIs. Relational agents have been shown to foster therapeutic rapport with users [32-34], support long-term engagement [35], and have been deployed across various clinical [35-37] and nonclinical populations [38-40], including (cognitively healthy) older adults [41-43]. However, most relational agent studies have been domain-specific (eg, targeting physical activity [41] or adherence to home blood pressure monitoring [44]), and have not been integrated into a broader JITAI framework to deliver holistic, multidomain sets of lifestyle interventions tailored to older adults with cognitive decline.

Taken together, the potential for combining (1) JITAI logic and (2) relational agents for the delivery of a (3) holistic, multidomain lifestyle intervention in (4) particularly underserved populations, such as older adults with SCD or MCI—remains completely underexplored.

To address this gap, this study explores the plausibility of a smartphone-based and relational agent–delivered just-in-time adaptive holistic lifestyle intervention in older adults with SCD or MCI, guided by the following research questions (RQs): (RQ1) To what extent do older adults with SCD or MCI adhere to a smartphone-based, CA-delivered JITAI over a 2-week period? (RQ2) How do participants perceive the usability and acceptability of the intervention?

By answering these questions, this study will contribute to the emerging research space at the intersection of digital lifestyle interventions, JITAIs, and relational agents—and thus to the integration of critical but often siloed research areas—by providing insights into how such tools can be adapted for cognitively vulnerable, aging populations.

XSL•FO
RenderX

Methods

Study Design

This study was designed as a 2-week, single-arm proof-of-concept study to evaluate the acceptability, adherence, and technical implementation of a smartphone-based, holistic just-in-time adaptive digital lifestyle intervention delivered by a rule-based CA. The interventions focused on promoting daily health behaviors across 7 domains relevant to cognitive and overall health, including physical activity, nutrition, sleep, cognitive exercises, hydration, and social engagement. The target population consisted of older adults with either SCD or MCI.

As a proof-of-concept study, the primary objectives were to assess plausibility, acceptability, and adherence patterns to the CA-initiated dialog sequences, and participants' subjective evaluation of the intervention's usability and acceptability. Secondary objectives included exploring the perceived working alliance with the CA and collecting qualitative feedback to inform future iterations of the intervention.

Intervention Design

We adopted the JITAI framework to design a low-burden, chatbot-delivered digital lifestyle intervention tailored to the behavioral needs and capacities of older adults with SCD or MCI—individuals in the early stages of cognitive decline who generally retain the ability to provide reliable, momentary self-reports when prompted with low-burden, structured questions and limited recall windows [45].

Following the conceptual framework by Nahum-Shani et al [20], which defines key elements such as decision points, tailoring variables, intervention options, and decision rules, and proximal or distal outcomes (Table 1), the intervention was designed to assess a participant's momentary states and only deliver a lifestyle suggestion when two criteria were met: (1) the participant had not yet engaged in the suggested lifestyle behavior (ie, determining the state of vulnerability or opportunity), and (2) the participant indicated availability and willingness to engage in the suggested lifestyle behavior (ie, state of receptivity).

XSL•FO
RenderX

Table . Key terms of a JITAI[a] and its operationalization[b] to the Elsa/Erik[c] study among older adults with SCD[d] or MCI[e].

Key term	Operationalization
Distal outcome	For example, quality of life and cognitive health.
Proximal outcomes	Implementation (yes/no) of a specific daily lifestyle behavior, such as a walk in the park or an activity nutrient-rich diet.
State of vulnerability or opportunity	An individual is classified as vulnerable by the CA[f] Elsa/Erik (yes/no), if a specific health-promoting lifestyle behavior (see Table 2 for all behaviors) has not been implemented until the early afternoon (see Decision points below);
State of receptivity	An individual is classified as receptive and open to implementing a health-promoting lifestyle suggestion (yes/no) by the CA Elsa/Erik, if she or he has time in the afternoon or early evening to implement such a behavior.
Decision points	Every day in the early afternoon, either 1 PM, 2 PM, or 3 PM; the specific time when Elsa/Erik should ask the participant about the state of vulnerability and receptivity was defined by the study participant during the first conversational turns with Elsa/Erik.
Intervention options	Low-burden, easy-to-implement lifestyle interventions focusing on diet and nutrition (eg, fruits and vegetables), physical activity (eg, a walk in the park), sleep (eg, 8 h of uninterrupted sleep), stress management, etc (see Table 2 for a full list of the options) recommended by Elsa/Erik if someone was classified as vulnerable and receptive.
Tailoring variables	Participants' response (yes/no) to her or his state of vulnerability and receptivity across a predefined lifestyle behavior (eg, if the target person has not yet implemented a health-promoting behavior such as physical activity, a variety of physical activity options would have been suggested by Elsa/Erik).
Decision rules	If someone is vulnerable (see tailoring variables above) and if that person also indicates that she or he is receptive to implementing a suggested behavior, then the intervention option that matches the vulnerable state is delivered by Elsa/Erik; also refer to the visualization of the JITAI logic in Figure 1.

[a]JITAI: just-in-time adaptive intervention.

[b]Table outline adapted from Nahum-Shani et al [19].

[c]Elsa/Erik was a 2-week proof-of-concept study, and we were primarily interested in the response rates to the conversational agent and the adherence to the recommended lifestyle behaviors as proximal outcomes. We did not expect effects on potential distal outcomes.

[d]SCD: subjective cognitive decline.

[e]MCI: mild cognitive impairment.

[f]CA: conversational agent.

Table . An overview of the 7 health-promoting activities prompted through the CAs[a], Elsa/Erik, during the 2-week single-arm proof-of-concept study among older adults with SCD[b] or MCI[c].

Aspect of health	Example
Days 1 and 8: Exercise	Moderate exercise such as 30 minutes of walking, swimming, and cycling
Days 2 and 9: Balanced diet	Eating 4 different types of fruit and vegetables throughout the day
Days 3 and 10: Sleep	Sleep for 8 hours the past night, if not, nap for 25 minutes
Days 4 and 11: Cognitive exercise	Crossword puzzle, sudoku, scrabble
Days 5 and 12: Movement	Bike ride, work in the garden, dancing, yoga
Days 6 and 13: Social engagement	Meet or call a close friend or relative
Days 7 and 14: Hydration	Drinking 5 glasses (0.2 L) of water

[a]CA: conversational agent.

[b]SCD: subjective cognitive decline.

[c]MCI: mild cognitive impairment.

XSL•FO
RenderX

Figure 1. A decision tree indicating the JITAI logic for this 2-week single-arm proof-of-concept study among older adults with SCD or MCI, delivered by a rule-based CA through a smartphone app. CA: conversational agent; JITAI: just-in-time adaptive intervention; MCI: mild cognitive impairment; SCD: subjective cognitive decline.

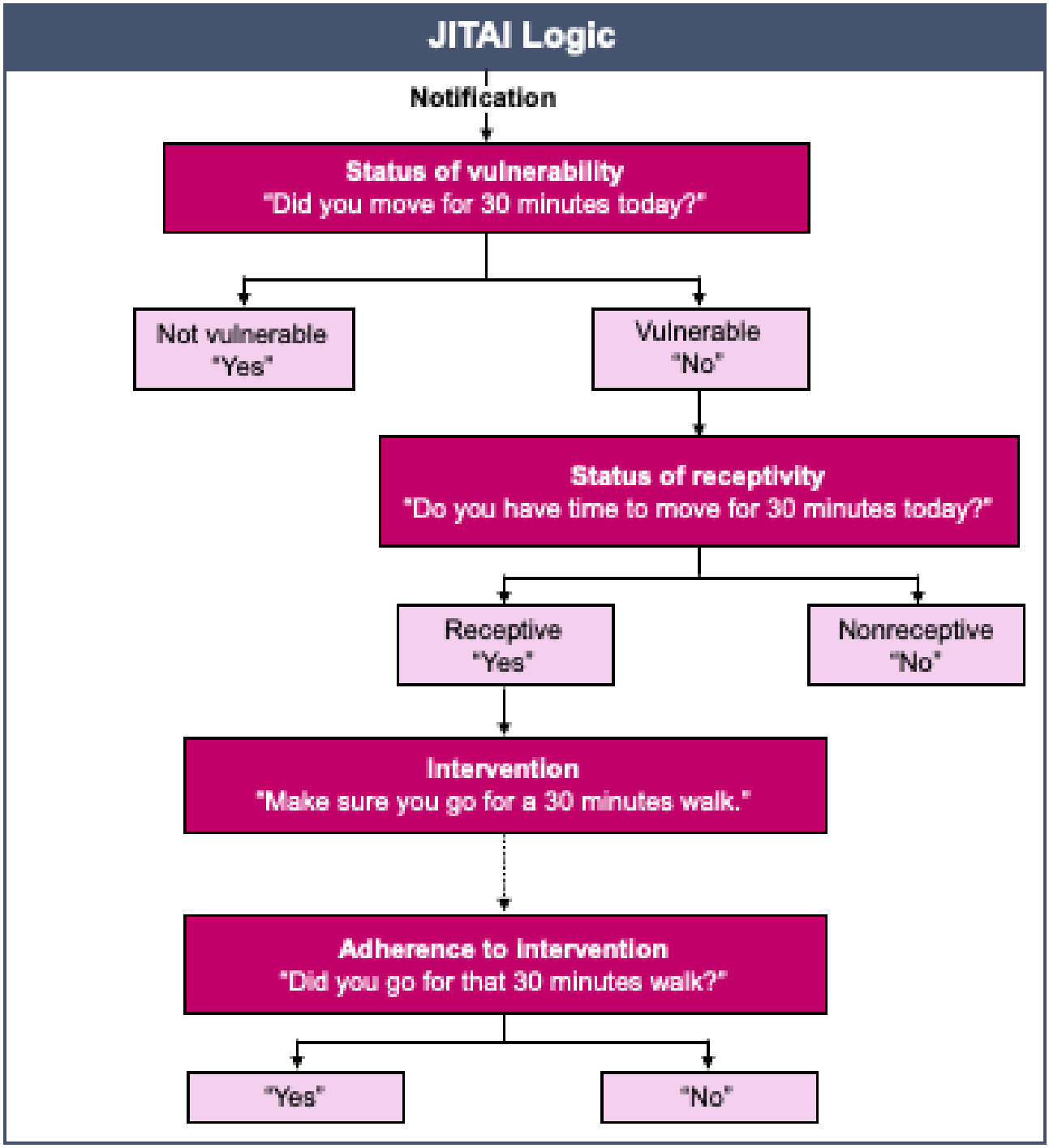

Each day's JITAI was structured around a decision point, occurring at a time predefined by each participant (1 PM, 2 PM, or 3 PM). At this time, the CA first assessed the participant's state of vulnerability—a key tailoring variable—by asking whether they had already completed the targeted activity (eg, physical activity, hydration, or social contact).

If vulnerability (=an opportunity) was detected, because the response was "no," a second tailoring variable—the participant's state of receptivity was assessed to determine a second decision point. Therefore, the CA asked whether the individuals felt able and willing to complete the suggested activity later that day. Only when both vulnerability and receptivity were affirmed, the intervention option was triggered: a low-burden, tailored suggestion encouraging engagement in a target behavior (eg, "Why don't you try and go for a 30-min walk. I'm sure you'll feel great after").

A third decision point occurred in the evening, where the CA inquired whether the participant had completed the suggested behavior. This response was used to capture the proximal outcome of adherence—that is, whether the support had led to real-world action that day.

XSL•FO
RenderX

Given the short duration and early-phase nature of this proof-of-concept, no distal outcome (eg, change in the QoL or long-term cognitive functioning) was assessed at this stage.

The complete mapping of each JITAI component to its operationalization in this study is also summarized in Table 1 and visualized in Figure 1. This table further clarifies how each core construct—decision points, tailoring variables, intervention options, decision rules, and outcomes—was instantiated within the JITAI logic of the Elsa/Erik study.

Technical Implementation

The smartphone-based JITAI was developed with the latest version of the MobileCoach software platform (April 2024) [46,47], which has already been used for various health interventions [31,37,39,40,48-51]. MobileCoach offers data collection capabilities (eg, via self-reports and interaction logs) and the delivery of JITAIs with the help of a rule-based CA. The user interface follows the principles of messaging apps, such as Apple's iMessage or WhatsApp. However, answer options are predominantly predefined with a few exceptions (eg, when the CA asks for the nickname of the study participant). MobileCoach was hosted by Pathmate Technologies. Participants could choose from two CAs, Elsa or Erik, to be their digital coach over the 2-week study period. Figure 2 demonstrates the three conversational turns identifying the state of vulnerability, state of receptivity, and adherence to the CA-suggested health-promoting activity.

Figure 2. The three conversational turns with the CA Elsa providing the JITAI in the 2-week single-arm proof-of-concept study among older adults with SCD or MCI. Translated from German. From left to right; first chat panel: determining the state of vulnerability (ie, "Have you completed an activity today?"); second chat panel: determining the state of receptivity (ie, if no: "Would you be able to complete the activity later?"); and third chat panel: determining adherence to the CA-suggested health-promoting activity (ie, "Did you complete this activity today?"). CA: conversational agent; JITAI: just-in-time adaptive intervention; MCI: mild cognitive impairment; SCD: subjective cognitive decline.

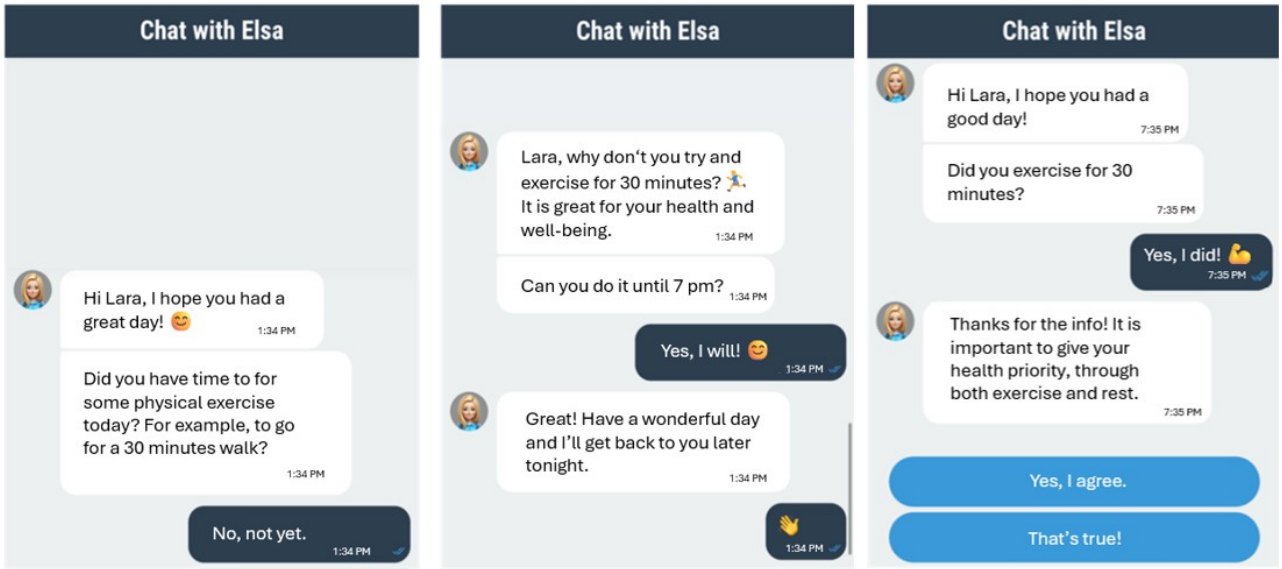

Recruitment

Participants were recruited in the Bern Memory Clinic and local cognitive training groups [52] in July 2024. First, a telephone screening for a first check of inclusion and exclusion criteria and a detailed explanation of the study was performed. An investigator explained to each participant the nature of the study, its purpose, the procedures involved, the expected duration, the potential risks and benefits, and any discomfort it may entail. As inclusion criteria, participants were required to be aged between 60 and 85 years, exhibit cognitive decline classified as SCD or MCI, possess a smartphone (either iOS or Android operated), demonstrate fluency in German, and have the capacity to provide written informed consent. Participants were excluded if they had severe psychiatric disorders or a history of substance abuse. Cognitive performance was assessed using the Montreal Cognitive Assessment (MoCA) score, with >25 indicating SCD, and scores equal to or below 25 indicating MCI [53,54] by expert investigators in a standardized clinical setting. Detailed study information and consent forms were sent to potential participants several days prior to the on-site visit. Participants gave written informed consent and were informed of their right to withdraw from the study at any time. Participation was voluntary, and no incentive for participation was provided.

Procedure

During the on-site preintervention visit, demographic data, general cognitive performance (assessed using the MoCA Score), and technology readiness were evaluated. Questions on technological readiness included the duration, modality, and frequency of smartphone use (Multimedia Appendix 1).

Participants then downloaded the study app, MobileCoach, and received 2 weeks of JITAI promoting healthy lifestyle behaviors through their CA. The CA started a maximum of two daily dialogs with the participant, one in the afternoon (to assess the state of vulnerability and receptivity) and one in the evening (to assess whether the health-promoting activity was implemented) if the participant indicated that they were vulnerable in the afternoon. Participants preselected their preferred times for these dialogs with the CA: 1 PM, 2 PM, or 3 PM for the afternoon dialog and 7 PM, 8 PM, or 9 PM for the evening dialog. The notifications prompted health-promoting

XSL•FO
RenderX

activities and assessed the participant's vulnerability and receptivity to these activities. In total, the CA suggested health-promoting activities from 7 aspects of health, twice each during the 2-week study period (Table 2). The health-promoting activities were decided among three authors (EB, RV, and PJ), based on evidence-based domains affecting positive health outcomes in the literature [55].

After the 2-week study period, a follow-up telephone call was conducted to confirm study completion and gather feedback on app use. Participants were sent (digitally or by post) a postintervention questionnaire (Multimedia Appendix 2).

Data Collection

The primary end point is adherence to the CA-initiated conversational turns, measured objectively via app interaction logs. Data from the app were collected using the MobileCoach software platform in anonymized form using participant IDs. The platform allowed the data to be exported and further used for analyzing the measured metrics reported in the results section. States of vulnerability, receptivity, and adherence to both the CA-initiated conversational turns and CA-suggested health-promoting activities, as well as the daily response rates to vulnerability questions, were measured.

Secondary end points include perceptions of technology acceptance, working alliance with the CA, self-reported adherence to the CA-suggested health-promoting activity, and suggestions for improving the JITAI (conducted through a questionnaire and short telephone interview). As a clarification, adherence to the CA-initiated conversational turns refers to whether the participants responded to the first notification assessing their state of vulnerability. Adherence to the CA-suggested health-promoting activity refers to whether the participant was able to complete the suggested activity.

A postintervention questionnaire was used to assess technology acceptance constructs, based on prior work on CA [31]. Constructs measured in this study included perceived ease of use, enjoyment, usefulness (as a reminder of the health-promoting behavior and as a motivator to complete the activity), control over the CA, and also the intention to interact again with the CA [56-58]. Working alliance, an important relationship quality, has previously been shown to relate strongly to treatment outcomes [59-62] and was measured with the 6-item session alliance inventory [63]. A 7-point Likert scale was used for each item and anchored from 1=strongly disagree to 7=strongly agree.

During the follow-up telephone call, a short semistructured interview was conducted to ask participants: (1) What did you particularly like about the interaction with Elsa/Erik? (2) What needs to be improved about the interaction with Elsa/Erik? (3) Do you have any further comments or suggestions for improvement? The postintervention questionnaire also included these three open-ended questions, which were then updated and added to their interview responses. Additionally, participants were also provided with space to provide open feedback and details regarding their smartphone and mobile operating system.

Data Analysis

Based on the mobile app data collected by MobileCoach, metrics for vulnerability, receptivity, and both adherences (ie, CA-initiated conversational turns and CA-suggested health-promoting activity) could be analyzed. Adherence to the CA-initiated conversational turns was calculated from the response to notifications on the status of vulnerability—the first afternoon message. Participants were categorized as vulnerable (who had not yet completed the activity), not vulnerable (who had already completed the activity), or nonresponsive. Receptivity was assessed only among vulnerable participants, who were further categorized based on their intent to complete the activity. Participants were either classified as receptive (intended to complete the activity), not receptive (did not intend to complete the activity), or nonresponsive. Proportions were calculated by dividing the number of participants in each category by the total number of vulnerable participants. Adherence to the CA-suggested health-promoting activity was measured by tracking whether participants who were vulnerable in the afternoon completed the activity by the evening notification time. Among these participants, those who completed the activity were considered to have adhered, while those who did not were categorized as not adhered. The proportion of adherence was calculated by dividing the number of participants who completed the activity by the total number of vulnerable participants who responded to the evening message.

For the self-report items of the postintervention questionnaires, we conducted descriptive statistics to report mean and SD calculations. Shapiro-Wilk tests indicated that the self-reported data were not normally distributed, and therefore, Wilcoxon signed rank tests were conducted to assess whether the median of the evaluations significantly differed from the test value of 4, representing a neutral response on the 7-point Likert scale. An overall score for the 6-item session alliance inventory was computed by averaging the six individual items, as Cronbach α of 0.90 demonstrated high internal consistency [9].

Inductive thematic analysis was conducted with the interview responses using MAXQDA (VERBI) [64]. Codes were determined from the raw questionnaire and interview data (translated into English) by authors EB and RV and further classified into overarching themes based on their trends. Two rounds of thematic analysis were conducted, and the results are presented as code and thematic maps, also using MAXQDA.

Ethical Considerations

The study was reviewed by the local Ethics Committee in Bern (identifier 2024 - 01029) and categorized as not falling under the Human Research Act [65]. Written informed consent was obtained from all participants in the study language, German. Participants were sent a study information sheet that included the inclusion and exclusion criteria, detailed information about the study, and their rights as participants, as well as a consent form for participation in the study and a consent form for further use of data. All data collected in this study were anonymized, and no identification of individual participants in the data, study, or supplementary files is possible. Collection, processing, and storage procedures for this study strictly adhered to Swiss legal

XSL•FO
RenderX

and ethical frameworks, including the Swiss Federal Act on Research involving Human Beings (Human Research Act) and the new Federal Act on Data Protection. Participants were not provided compensation for the study.

Results

Characteristics of the Study Population and Technology Readiness

Overall, 17 participants were recruited for the study. One participant was excluded due to outdated hardware, and another participant due to technical issues (notifications could not be delivered), resulting in a total analyzable sample of 15 participants. All participants were assigned to the CA-delivered app intervention in this single-arm study, and reporting followed the CONSORT (Consolidated Standards of Reporting Trials) checklist (Checklist 1). Participant sample characteristics indicated 67% (n=10) identified as female, and 33% (n=5) as male, where the mean age was 70.3 (SD 5.01) years, and the mean MoCA Score was 27.73 (SD 2.01). Based on the suggested cut-off [53,54], 12 participants were classified as having SCD and 3 as MCI. There was no significant difference between SCD and MCI participants in any of the technology readiness outcomes described in Table 3.

Table . Data collected from the 2-week, single-arm proof-of-concept study among older adults with SCD[a] or MCI[b], showing information on smartphone use and frequency, including the technology readiness of participants.

	Value
Smartphone and mobile operating system, n (%)	
Android smartphone	8 (53)
Apple iPhone	7 (47)
Frequency of carrying the smartphone, n (%)	
Always (including at home)	5 (33)
When leaving the house	8 (53)
Only when needed	2 (13)
Using the smartphone (on a 5-point Likert scale, where 1 is “never” and 6 is “multiple times per day”), mean (SD)	
Daily use for	
Chatting	4.26 (0.88)
Researching information on the internet	4 (1.3)
Checking the weather	3.66 (1.11)
Multiple times per week for	
Phone calls	3.46 (1.06)
Reading emails	3.46 (1.68)
Taking photos	3.13 (0.99)
Consulting their calendar	2.87 (1.77)
Planning routes	2.8 (1.42)
Writing emails	2.67 (1.59)
Once per week for	
Tracking health-related data	2.46 (1.72)
Listening to music	2.2 (0.94)
Playing games	1.8 (1.2)

[a]SCD: subjective cognitive decline.

[b]MCI: mild cognitive impairment.

Response Rates to the JITAI

On average, 81% (mean 12.14) of participants responded to the afternoon notification each day over the 14-day study period (SD 1.51 responses per day to the state of vulnerability dialogue), demonstrating adherence to the CA-initiated conversational turns. Conversely, 19% (mean 2.86, SD 1.51) of participants, on average, did not respond each day. A detailed breakdown of the daily response rates according to the response type is presented in Figure 3.

XSL•FO
RenderX

Figure 3. Daily response rates for afternoon vulnerability notifications to determine whether the participant has already completed the suggested activity or not, or if the participant does not respond. Graphed over the 2-week single-arm proof-of-concept study among older adults with subjective cognitive decline or mild cognitive impairment.

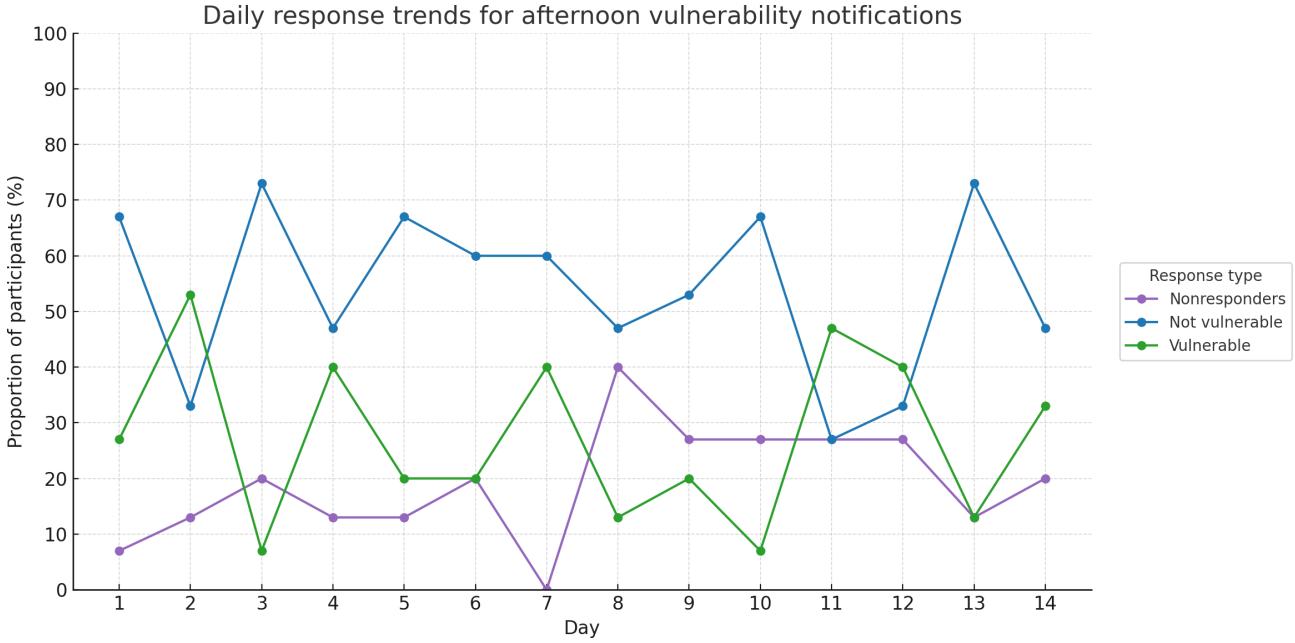

For each aspect of health, participant engagement was analyzed across three key stages: determining vulnerability, identifying receptivity, and measuring adherence to the CA-suggested health-promoting activity. Table 4 presents these metrics as percentages, with results grouped by specific activities. On average, 27% (mean 4.07, SD 2.27) of participants reported being vulnerable, and of these, 100% responded to the state of receptivity notification. An average of 83% (mean 3.14, SD 1.61) of participants were receptive to completing the health-promoting activity, with 69% (mean 2.86, SD 1.70) of participants self-reporting adherence to the activity. An average of 95% (mean 3.93, SD 2.27) of participants responded to the adherence evening notification.

XSL•FO
RenderX

Table . Data collected from the 2-week, single-arm proof-of-concept study among older adults with SCD[a] or MCI[b], showing results of participants' vulnerability, receptivity, and adherence to the CA[c]-suggested health-promoting activity, as proportion (%) and absolute values (n).

	State of vulnerability			State of receptivity			Adherence		
Health-promoting activity	No response[d]	Not vulnerable[e]	Was vulnerable[f]	No response[g]	Was not receptive[h]	Was receptive[i]	No response[j]	Did not adhere[k]	Did adhere[l]
Exercise (%)	23.30	56.70	20	0	0	100	0	37.50	62.50
Day 1 (n)	1	10	4	0	0	4	0	1	3
Day 8 (n)	6	7	2	0	0	2	0	1	1
Balanced diet (%)	20	43.30	36.70	0	25	75	0	25	75
Day 2 (n)	2	5	8	0	4	4	0	4	4
Day 9 (n)	4	8	3	0	0	3	0	0	3
Sleep (%)	23.30	70	6.70	0	0	100	0	50	50
Day 3 (n)	3	11	1	0	0	1	0	1	0
Day 10 (n)	4	10	1	0	0	1	0	0	1
Cognitive exercise (%)	20	36.70	43.30	0	29.80	70.20	7.10	38.10	54.80
Day 4 (n)	2	7	6	0	1	5	0	2	4
Day 11 (n)	4	4	7	0	3	4	1	3	3
Movement (%)	20	50	30	0	25	75	0	16.70	83.30
Day 5 (n)	2	10	3	0	0	3	0	0	3
Day 12 (n)	4	5	6	0	3	3	0	2	4
Social engagement (%)	16.70	66.70	16.70	0	41.70	58.30	25	16.70	58.30
Day 6 (n)	3	9	3	0	1	2	0	1	2
Day 13 (n)	2	11	2	0	1	1	1	0	1
Hydration (%)	10	53.30	36.70	0	0	100	0	0	100
Day 7 (n)	0	9	6	0	0	6	0	0	6
Day 14 (n)	3	7	5	0	0	5	0	0	5

[a]SCD: subjective cognitive decline.

[b]MCI: mild cognitive impairment.

[c]CA: conversational agent.

[d]This is the proportion of participants who did not respond to the vulnerability message for each intervention type, relative to the total participants for that interventio.n

[e]This is the proportion of participants who indicated they were "not vulnerable" and had already completed the activity, as a percentage of the total participants for the intervention type.

[f]This proportion represents participants who indicated they were "vulnerable" for the activity type, again relative to the total participants for that intervention.

[g]This shows the percentage of participants who did not respond to the receptivity message after indicating vulnerability.

[h]The proportion of participants who were "not receptive" to completing the activity, among those who were "vulnerable."

[i]This is the percentage of vulnerable participants who were "receptive" to completing the activity.

[j]This represents the percentage of participants who did not respond to the evening check-in for adherence, calculated only for participants who were marked as "vulnerable."

[k]The proportion of participants who did "not adhere" to the activity, again calculated only for those marked as "vulnerable."

[l]This is the percentage of participants who "adhered" to the activity, calculated based on those who were "vulnerable."

XSL•FO
RenderX

However, engagement patterns varied significantly across health domains. For example, the cognitive exercise domain showed the highest combined engagement: 43% (n=6.5) of participants reported being vulnerable, 70% (n=4.5) were receptive, and 55% (n=3.5) adhered to the suggested activity. In contrast, sleep was the domain with the lowest proportion of vulnerability: only 1 (7% of all participants) person indicated being vulnerable at all who was receptive on both days and consistently reported that they had already engaged in the recommended sleep behavior by the time the just-in-time prompt was delivered. This suggests that many participants already had stable evening routines or did not perceive a need for behavioral adjustment in this area.

Technology Acceptance and Working Alliance Assessment

Participants were asked to rate their experience interacting with the CAs Elsa/Erik on a 7-point scale (where 1=strongly disagree and 7 =strongly agree). Further, participants rated their perceived relationship with Elsa/Erik on a 7-point scale (where 1=never and 7=always). The results are displayed in Table 5.

Table . Data collected from the 2-week, single-arm proof-of-concept study among older adults with SCD[a] or MCI[b], showing the postinteraction assessment of 15 participants on PEU[c], PU[d], PEN[e], PC[f], ITI[g], and PSA[h].

#Item	Mean (SD)[i]	V[j]
PEU: I thought the response from Erik/Elsa was easy to understand.	6.60 (0.83)	105.0[k]
PU1: I found it useful to work on the intervention with Erik/Elsa.	5.13 (1.77)	76.0[l]
PU2: I found Erik/Elsa motivated me to perform the exercises.	4.67 (2.02)	55.0
PEN: I had fun using Erik/Elsa.	5.20 (2.08)	85.5[l]
PC: I was able to control the interaction with Erik/Elsa.	5.33 (2.38)	90.0
ITI: I would continue interacting with Erik/Elsa.	4.40 (2.53)	61.0
PSA		
I felt that Erik/Elsa and I respected each other.	6.20 (1.42)	89.5[m]
I felt that Erik/Elsa appreciated me.	6.07 (1.44)	115.0[m]
I felt that Erik/Elsa cared about me even when I did things it did not approve of.	5.67 (1.50)	98.0[m]
I felt that Erik/Elsa and I are working toward mutually agreed upon goals.	4.93 (1.67)	54.0
I felt that Erik/Elsa and I agree on what is important for me to work on.	5.13 (1.68)	76.5[l]
I believe the way Erik/Elsa and I are working with my problem is correct.	5.53 (1.36)	75.5[m]
Overall	5.59 (1.23)	102.0[m]

[a]SCD: subjective cognitive decline.
[b]MCI: mild cognitive impairment.
[c]PEU: perceived ease of use.
[d]PU: perceived usefulness.
[e]PEN: perceived enjoyment.
[f]PC: perceived control.
[g]ITI: intention to interact.
[h]PSA: perceived session alliance.
[i]7-point Likert scales were used from 1=strongly disagree to 7=strongly agree.
[j]V is the test statistic of the Wilcoxon signed rank with test value 4.
[k]$P<.001$.
[l]$P<.05$.
[m]$P<.01$.

Results indicate that all mean values lie above the neutral scale value of 4. Moreover, Wilcoxon signed rank tests indicate that the median of perceived ease of use and two items of the session alliance inventory (items 1 and 2) were significantly higher compared to the neutral test value of 4. Additionally, the median values of perceived usefulness as a reminder (PU1), perceived

XSL•FO
RenderX

enjoyment, perceived control, and the overall session alliance inventory items were also rated highly. The remaining construct items (perceived usefulness as a motivator [PU2] and intention to interact) still had medians above the value of 4. Therefore, it can be concluded that participants evaluated the CAs Elsa/Erik favorably.

Qualitative Interview Feedback

To complement the quantitative findings, all participants were contacted postintervention for a short semistructured interview (approximately 10 min). Results focusing on (1) positive feedback, (2) potential for improvement, and (3) open remarks (cf questions mentioned in the Methods) were thematically analyzed and coded. Figure 4, therefore, presents a thematic map of two overarching themes—app enjoyment and enhancing engagement—along with related subthemes and their frequency counts. Multimedia Appendix 3further provides direct participant quotes to illustrate each subtheme and clarify sentiment.

Figure 4. Qualitative results of the 2-week proof-of-concept study conducted with a smartphone-based CA among older adults with SCD or MCI. This presents a thematic map of the main emerging themes, app enjoyment and enhancing engagement, from the qualitative interviews. Note: Numbers in circles represent the frequency count, or how many times the participant theme and its concept were mentioned by participants. Some participants mentioned multiple examples of a single theme during their interview, which then led to multiple frequency counts being identified during thematic analysis. CA: conversational agent; MCI: mild cognitive impairment; SCD: subjective cognitive decline.

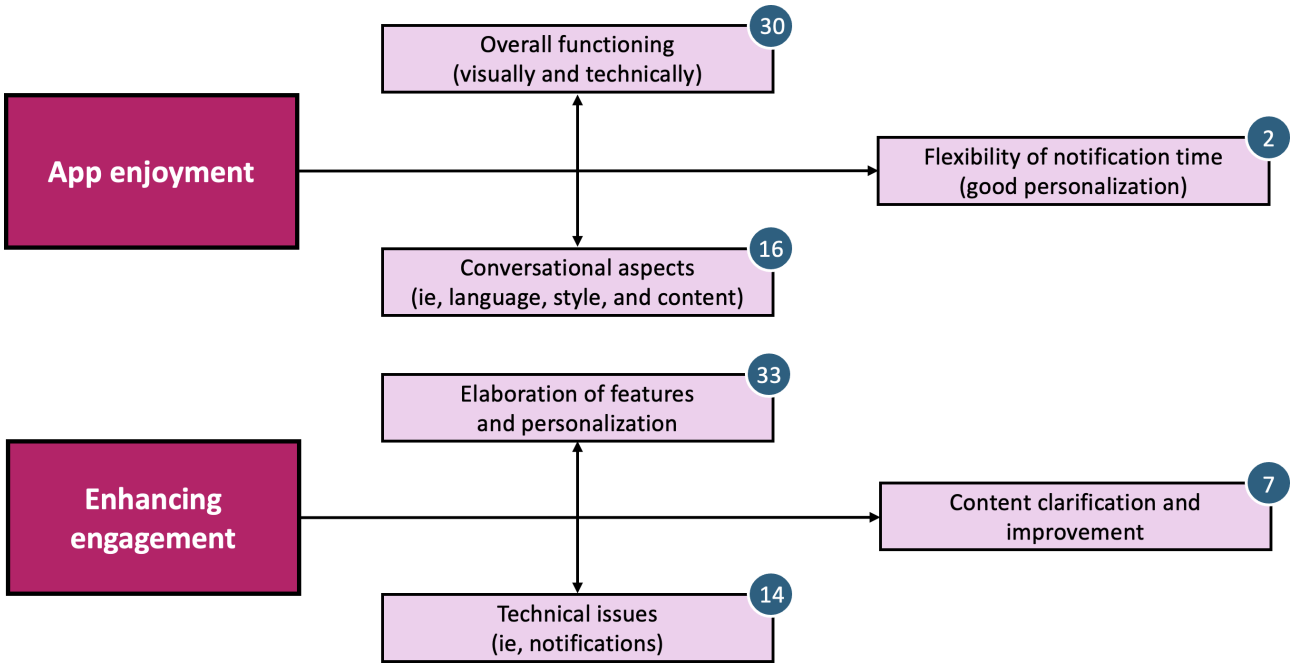

Participants generally appreciated the overall functioning of the app (both visually and technically), with several describing it as “easy to use” (P06, P07, P12, and P17), “appealing, nothing to criticize” (P14), or “the design was ok” (P10, P11, and P12). Participants also responded positively to the tone and conversational style of the agent: “There were no accusations if something wasn’t done, which was very pleasant” (P07), but also highlighted the need for adaptations for older adults: “The language was quite youthful, it should be adjusted since the app is for older people” (P09). The timing and delivery of notifications also emerged as an area for improvement, with several users noting they had missed prompts or found the timing suboptimal: “It might be better if the notifications came in the morning […] especially if you are still working” (P03, P04, and P05).

Discussion

Principal Results

Overview

This proof-of-concept study investigated the plausibility and acceptability of a smartphone-based JITAI delivered via a rule-based CA (Elsa/Erik) among older adults with SCD or MCI. While our intervention draws on the well-established tradition of relational agents [20,66], it introduces several novel elements: integrating (1) a JITAI logic delivered via a relational agent; (2) embedding a multicomponent, holistic set of lifestyle interventions; and (3) tailoring the JITAI and relational agent to minimize burden for a vulnerable population such as cognitively impaired older adults.

Delivering JITAI Logic via a CA

Overall, participants showed high engagement with the JITAI decision-making flow that included daily conversational check-ins to assess states of vulnerability, receptivity, and adherence. Across the 14-day intervention, the overall response rate to the initial question assessing the state of vulnerability amounted to 81%, suggesting strong adherence to the CA-initiated conversational turns (RQ1). Vulnerability was then reported in 27% (n=4.07) of interactions, and 100% of vulnerable participants responded to the follow-up receptivity notification. Notably, 69% (n=2.86) of those classified as receptive also reported completing a suggested activity. Moreover, participants generally reported a positive experience interacting with the CA, as reflected in their ratings of the

XSL•FO
RenderX

working alliance. Taken together, these findings support the plausibility of delivering a JITAI via a CA.

Toward Personalized Multicomponent, Holistic Lifestyle Intervention

The JITAI intervention delivered lifestyle recommendations across 7 distinct domains of modifiable risk factors relevant to dementia prevention and healthy aging: exercise, balanced diet, sleep, cognitive exercises, movement, social engagement, and hydration [9]. Analysis across domains revealed considerable variation in momentary vulnerability: for example, cognitive exercise emerged as the area with the highest vulnerability, whereas sleep had the lowest reported vulnerability, with only one individual indicating a need for behavioral support on relevant days. On the one hand, this supports the value of a multidomain approach, as it allows tailoring a JITAI to varying individual needs. On the other hand, it suggests that future versions of the JITAI should be more flexible to allow users to select focal domains or automatically personalize content based on historical preferences. Indeed, participants also mentioned that providing supporting information regarding suggested activities or alternatives for certain activities in case they could not complete a specific task on a given day, but still wished to engage, would have been appreciated.

Tailoring the Intervention to Prioritize Minimal Burden for Vulnerable Intervention

Results show that participants generally responded positively to the rule-based CA-delivered JITAI. Quantitative and qualitative feedback confirmed perceived ease of use, perceived usefulness (ie, reminders of the health-promoting behavior), enjoyment, and an overall positive working alliance (ie, participant-CA relationship). Taken together, these findings support the usability and acceptability of a structured, CA-delivered JITAI designed for low-burden interactions (RQ2).

However, the findings of this study also highlight several critical aspects related to the deployment of JITAIs in older adults with varying degrees of cognitive impairment, such as MCI and most likely mild Alzheimer disease. One of the most significant considerations is the adequacy of the hardware used to deliver the intervention. As observed, two participants were excluded due to outdated hardware or technical issues, underscoring the importance of future studies to ensure that the technology used is sufficiently robust and up to date [67]. The interview study also revealed a high frequency of mentions regarding issues with notifications. Overall, this emphasizes the need for addressing potential hardware challenges in future studies, as these issues were not related to participant behavior but to technical malfunctions, which is also a key takeaway from our proof-of-concept study.

Limitations and Future Research Directions

This proof-of-concept study was deliberately designed as a formative, discovery-oriented investigation rather than a hypothesis-driven trial. This said, no a priori benchmarks for adherence or acceptability were defined, given the novelty of the target population and the limited existing evidence on smartphone-based and CA-driven JITAIs in individuals with SCD or MCI. Instead, this design allowed for open-ended observations of use patterns and participant responses to interventions. While this approach supported initial insights into the plausibility and user engagement of the JITAI, future studies will benefit from the use of literature- or theory-informed benchmarks to support clearer interpretability and “go/no-go” decision-making as suggested, for instance, by emerging frameworks such as the Digital Therapeutics Real-World Evidence framework [68].

Another key limitation of the current JITAI design lies in its reliance on self-reported data only to assess both states of vulnerability, receptivity, and adherence. While the target population—individuals with SCD and MCI—is generally capable of providing momentary self-reports over short timeframes, this approach may not scale to individuals with more advanced cognitive decline [45]. To minimize cognitive burden, we limited the recall window to same-day behaviors and used simple, structured prompts. Nevertheless, future iterations of the intervention could benefit from the integration of passive sensing modalities (eg, step count, sleep patterns, location-based context, or even social interaction cues from Bluetooth proximity) to reduce reliance on self-reports and enable more accurate, inclusive data collection. Wearable devices, such as smartwatches or fitness bands, hold particular promise for older adults with cognitive impairments [69]. In addition to enabling continuous, real-time tracking of health behaviors like physical activity, sleep wearables offer a more reliable notification interface: unlike smartphones, which may be overlooked or out of reach, wearables provide discreet, wrist-based notifications through visual cues or haptic feedback, increasing the likelihood that prompts are noticed and acted upon [70].

Further, this early-stage proof-of-concept study was designed to investigate plausibility-related outcomes such as acceptability, usability, and engagement, rather than to demonstrate broad effectiveness. The outcomes observed here reflect the experiences of a predominantly female and technology-aware sample, reflecting common patterns in research study participation, but limit generalizability across genders.

Additionally, embedding the JITAI into a stepped-care program or study with in-person contacts, as a form of blended treatment [71], would allow for additional collection of proxy (or “other”) reports from caregivers or clinicians, which could serve as alternative data source to strengthen the reliability of the decision logic. A blended approach could also offer the necessary support to enhance the overall effectiveness and user experience with the hardware [72,73].

The use of a rather simple, low-burden rule-based CA further constrained the depth and flexibility of user interactions. Participants could only select from predefined response options, and the CA delivered one daily activity suggestion based solely on binary assessments of vulnerability and receptivity. This approach did not accommodate individual preferences or provide alternative options if the suggested behavior was, for instance, irrelevant (cf the low vulnerability reported for days with a sleep intervention). Indeed, several participants expressed a desire for more contextual flexibility and choice. While this JITAI was intentionally designed to reduce cognitive load and mitigate

XSL•FO
RenderX

confusion or misinformation, its limited adaptability also constrained the intervention's alignment with the defining characteristics of a "true" JITAI, which requires nuanced, context-sensitive tailoring over time. Future research may address this limitation by exploring hybrid architectures that retain the safety and clarity of rule-based logic while incorporating the adaptive potential of generative artificial intelligence (eg, large language models). Such models may allow for richer personalization, deeper therapeutic engagement, and more responsive intervention delivery—provided that appropriate safeguards and oversight mechanisms are in place to preserve safety and therapeutic integrity for cognitively vulnerable users [74,75].

Conclusions

The multicomponent approach used in this study, which integrates various aspects of health such as sleep, hydration, physical fitness, and cognitive exercises, is particularly promising for enhancing the QoL in older adults. By addressing multiple dimensions of well-being simultaneously, this approach has the potential to produce more comprehensive and sustained benefits compared to interventions that focus on a single health aspect. For older adults, maintaining a balance across these domains is crucial, as deficits in one area, like poor sleep or inadequate hydration, can negatively impact other aspects of health, such as cognitive function or physical mobility. By promoting holistic health behaviors, the JITAI not only supports cognitive health but also contributes to overall well-being, potentially delaying the onset of more severe impairments and enhancing daily functioning. This integrated strategy aligns with current geriatric care principles, which emphasize the importance of a well-rounded approach to aging and hold great promise for improving the QoL in this population.

Acknowledgments

We would like to sincerely thank all participants for generously contributing their time and effort to this study. We also wish to express our special thanks to our master's student, Marina Gisler, for her dedicated assistance with data collection.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Authors' Contributions

EB, RV and MN contributed equally and share first authorship. SK and TK also contributed equally and share last authorship.

Conflicts of Interest

RV, MN, and TK are affiliated with the Center for Digital Health Interventions (CDHI), a joint initiative of the Institute for Implementation Science in Health Care, University of Zurich; the Department of Management, Technology, and Economics at ETH Zurich; and the Institute of Technology Management and School of Medicine at the University of St. Gallen. CDHI is funded in part by the Swiss health insurer CSS, the Austrian health care provider (and corporate start-up of UNIQA) Mavie Next, and the Swiss investor MTIP. TK is also a cofounder of Pathmate Technologies, a university spin-off company that creates and delivers digital clinical pathways. However, neither CSS, Mavie Next (UNIQA), nor MTIP were involved in this study. Furthermore, TK has neither shares of Pathmate Technologies nor any formal role in the company.

Multimedia Appendix 1
Preintervention questionnaire (German).
[DOCX File, 58 KB - formative_v9i1e66885_app1.docx]

Multimedia Appendix 2
Postintervention questionnaire (German).
[DOCX File, 63 KB - formative_v9i1e66885_app2.docx]

Multimedia Appendix 3
Code groups and participant quotes (translated in English).
[PDF File, 75 KB - formative_v9i1e66885_app3.pdf]

Checklist 1
CONSORT (Consolidated Standards of Reporting Trials) checklist (English).
[PDF File, 153 KB - formative_v9i1e66885_app4.pdf]

References

XSL•FO
RenderX

1. GBD Dementia Forecasting Collaborators. Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the Global Burden of Disease Study 2019. Lancet Public Health 2022;7. [doi: 10.1016/S2468-2667(21)00249-8]
2. Alzheimer's Association. Alzheimer's disease: facts and figures special report race, ethnicity and Alzheimer's in America. Alzheimers Dement 2021;17:327-406.
3. Röhr S, Pabst A, Riedel-Heller SG, et al. Estimating prevalence of subjective cognitive decline in and across international cohort studies of aging: a COSMIC study. Alzheimers Res Ther 2020 Dec 18;12(1):167. [doi: 10.1186/s13195-020-00734-y] [Medline: 33339532]
4. Jessen F, Wiese B, Bachmann C, et al. Prediction of dementia by subjective memory impairment. Arch Gen Psychiatry 2016;67:414-422. [doi: 10.1001/archgenpsychiatry.2010.30]
5. Aj M, Beaumont H, Ferguson D, Yadegarfar M, Risk SB. Risk of dementia and mild cognitive impairment in older people with subjective memory complaints: meta-analysis. Acta Psychiatr Scand 2014:1-13. [doi: 10.1111/acps.12336]
6. Seifert A, Schelling HR. Dementia barometer 2018—knowledge, attitudes and experiences of the swiss population [Article in German]. : Universität Zürich ZORA; 2019. URL: https://www.zora.uzh.ch/id/eprint/170355/ [accessed 2025-07-11]
7. Xu M, Zhu J, Liu XD, Luo MY, Xu NJ. Roles of physical exercise in neurodegeneration: reversal of epigenetic clock. Transl Neurodegener 2021 Aug 13;10(1):30. [doi: 10.1186/s40035-021-00254-1] [Medline: 34389067]
8. Fitzgerald KN, Hodges R, Hanes D, et al. Potential reversal of epigenetic age using a diet and lifestyle intervention: a pilot randomized clinical trial. Aging (Albany NY) 2021 Apr 12;13(7):9419-9432. [doi: 10.18632/aging.202913] [Medline: 33844651]
9. Livingston G, Huntley J, Sommerlad A, et al. Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. Lancet 2020 Aug 8;396(10248):413-446. [doi: 10.1016/S0140-6736(20)30367-6] [Medline: 32738937]
10. Jacobson N, Kowatsch T, Marsch L. Digital Therapeutics for Mental Health and Addiction: Elsevier; 2023. [doi: 10.1016/C2020-0-02801-X]
11. Kowatsch T, Fleisch E. Digital health interventions. In: Gassmann O, Ferrandina F, editors. Connected Business: Springer International Publishing; 2021:71-95. [doi: 10.1007/978-3-030-76897-3_4]
12. Brodaty H, Chau T, Heffernan M, et al. An online multidomain lifestyle intervention to prevent cognitive decline in at-risk older adults: a randomized controlled trial. Nat Med 2025 Feb;31(2):565-573. [doi: 10.1038/s41591-024-03351-6] [Medline: 39875685]
13. Heffernan M, Andrews G, Fiatarone Singh MA, et al. Maintain Your Brain: protocol of a 3-year randomized controlled trial of a personalized multi-modal digital health intervention to prevent cognitive decline among community dwelling 55 to 77 year olds. J Alzheimers Dis 2019;70(s1):S221-S237. [doi: 10.3233/JAD-180572] [Medline: 30475762]
14. Anupama GJ, Christos B, Megan H, Carlos SJJ, Igor C, Tiffany C. Fully-online, interoperable clinical trial management system for multi-interventional RCT: Maintain Your Brain digital platform. Stud Health Technol Inform 2020;268:97-112. [doi: 10.3233/SHTI200009]
15. Schiwal AT, Fauth EB, Wengreen H, Norton M. The Gray Matters app targeting health behaviors associated with Alzheimer's risk: improvements in intrinsic motivation and impact on diet quality and physical activity. J Nutr Health Aging 2020;24(8):893-899. [doi: 10.1007/s12603-020-1421-5] [Medline: 33009542]
16. Hartin PJ, Nugent CD, McClean SI, et al. The empowering role of mobile apps in behavior change interventions: The Gray Matters randomized controlled trial. JMIR Mhealth Uhealth 2016 Aug 2;4(3):e93. [doi: 10.2196/mhealth.4878] [Medline: 27485822]
17. Norton MC, Clark CJ, Tschanz JT, et al. The design and progress of a multidomain lifestyle intervention to improve brain health in middle-aged persons to reduce later Alzheimer's disease risk: The Gray Matters randomized trial. Alzheimers Dement (N Y) 2015 Jun;1(1):53-62. [doi: 10.1016/j.trci.2015.05.001] [Medline: 29854925]
18. Oikonomidi T, Ravaud P, LeBeau J, Tran VT. A systematic scoping review of just-in-time, adaptive interventions finds limited automation and incomplete reporting. J Clin Epidemiol 2023 Feb;154:108-116. [doi: 10.1016/j.jclinepi.2022.12.006] [Medline: 36521653]
19. Nahum-Shani I, Smith SN, Spring BJ, et al. Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. Ann Behav Med 2018 May 18;52(6):446-462. [doi: 10.1007/s12160-016-9830-8] [Medline: 27663578]
20. Nahum-Shani I, Hekler EB, Spruijt-Metz D. Building health behavior models to guide the development of just-in-time adaptive interventions: a pragmatic framework. Health Psychol 2015 Dec;34S:1209-1219. [doi: 10.1037/hea0000306] [Medline: 26651462]
21. Naughton F. Delivering "Just-In-Time" smoking cessation support via mobile phones: current knowledge and future directions. Nicotine Tob Res 2016;19(3):379-383. [doi: 10.1093/ntr/ntw143]
22. Zheng S, Edney SM, Goh CH, et al. Effectiveness of holistic mobile health interventions on diet, and physical, and mental health outcomes: a systematic review and meta-analysis. EClinicalMedicine 2023 Dec;66:102309. [doi: 10.1016/j.eclinm.2023.102309] [Medline: 38053536]
23. Jiang X, Ming WK, You JH. The cost-effectiveness of digital health interventions on the management of cardiovascular diseases: systematic review. J Med Internet Res 2019;21(6):e13166. [doi: 10.2196/13166]

XSL•FO
RenderX

24. Kouvari M, Karipidou M, Tsiampalis T, et al. Digital health interventions for weight management in children and adolescents: systematic review and meta-analysis. J Med Internet Res 2022 Feb 14;24(2):e30675. [doi: 10.2196/30675] [Medline: 35156934]
25. Brewer LC, Fortuna KL, Jones C, et al. Back to the future: achieving health equity through health informatics and digital health. JMIR Mhealth Uhealth 2020 Jan 14;8(1):e14512. [doi: 10.2196/14512] [Medline: 31934874]
26. Huh J, Koola J, Contreras A, et al. Consumer health informatics adoption among underserved populations: thinking beyond the digital divide. Yearb Med Inform 2018 Aug;27(1):146-155. [doi: 10.1055/s-0038-1641217] [Medline: 30157518]
27. Skivington K, Matthews L, Simpson SA, et al. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. BMJ 2021 Sep 30;374:n2061. [doi: 10.1136/bmj.n2061] [Medline: 34593508]
28. Makri A. Bridging the digital divide in health care. Lancet Digital Health 2019 Sep;1(5):e204-e205. [doi: 10.1016/S2589-7500(19)30111-6]
29. Mitchell UA, Chebli PG, Ruggiero L, Muramatsu N. The digital divide in health-related technology use: the significance of race/ethnicity. Gerontologist 2019 Jan 9;59(1):6-14. [doi: 10.1093/geront/gny138] [Medline: 30452660]
30. Vaportzis E, Clausen MG, Gow AJ. Older adults perceptions of technology and barriers to interacting with tablet computers: a focus group study. Front Psychol 2017 Oct 4;8:1687. [doi: 10.3389/fpsyg.2017.01687] [Medline: 29071004]
31. Kowatsch T, Schachner T, Harperink S, et al. Conversational agents as mediating social actors in chronic disease management involving health care professionals, patients, and family members: multisite single-arm feasibility study. J Med Internet Res 2021 Feb 17;23(2):e25060. [doi: 10.2196/25060] [Medline: 33484114]
32. Beatty C, Malik T, Meheli S, Sinha C. Evaluating the therapeutic alliance with a free-text CBT conversational agent (Wysa): a mixed-methods study. Front Digital Health 2022;4:847991. [doi: 10.3389/fdgth.2022.847991] [Medline: 35480848]
33. Malouin-Lachance A, Capolupo J, Laplante C, Hudon A. Does the digital therapeutic alliance exist? Integrative review. JMIR Ment Health 2025 Feb 7;12:e69294. [doi: 10.2196/69294] [Medline: 39924298]
34. Jabir AI, Martinengo L, Lin X, Torous J, Subramaniam M, Tudor Car L. Evaluating conversational agents for mental health: scoping review of outcomes and outcome measurement instruments. J Med Internet Res 2023 Apr 19;25:e44548. [doi: 10.2196/44548] [Medline: 37074762]
35. Schläpfer S, Schneider F, Santhanam P, et al. Engagement with a relaxation and mindfulness mobile app among people with cancer: exploratory analysis of use data and self-reports from a randomized controlled trial. JMIR Cancer 2024 May 31;10:e52386. [doi: 10.2196/52386] [Medline: 38819907]
36. Griffin AC, Khairat S, Bailey SC, Chung AE. A chatbot for hypertension self-management support: user-centered design, development, and usability testing. JAMIA Open 2023 Oct;6(3):ooad073. [doi: 10.1093/jamiaopen/ooad073] [Medline: 37693367]
37. Stasinaki A, Büchter D, Shih CHI, et al. Effects of a novel mobile health intervention compared to a multi-component behaviour changing program on body mass index, physical capacities and stress parameters in adolescents with obesity: a randomized controlled trial. BMC Pediatr 2021 Jul 9;21(1):308. [doi: 10.1186/s12887-021-02781-2] [Medline: 34243738]
38. Castro RP, Haug S, Filler A, Kowatsch T, Schaub MP. Engagement within a mobile phone-based smoking cessation intervention for adolescents and its association with participant characteristics and outcomes. J Med Internet Res 2017;19(11):e356. [doi: 10.2196/jmir.7928]
39. Kramer JN, Künzler F, Mishra V, et al. Which components of a smartphone walking app help users to reach personalized step goals? Results from an optimization trial. Ann Behav Med 2020 Jun 12;54(7):518-528. [doi: 10.1093/abm/kaaa002] [Medline: 32182353]
40. Stieger M, Flückiger C, Rüegger D, Kowatsch T, Roberts BW, Allemand M. Changing personality traits with the help of a digital personality change intervention. Proc Natl Acad Sci U S A 2021 Feb 23;118(8):e2017548118. [doi: 10.1073/pnas.2017548118] [Medline: 33558417]
41. Bickmore TW, Silliman RA, Nelson K, et al. A randomized controlled trial of an automated exercise coach for older adults. J Am Geriatr Soc 2013 Oct;61(10):1676-1683. [doi: 10.1111/jgs.12449] [Medline: 24001030]
42. Vardoulakis LP, Ring L, Barry B, Sidner CL, Bickmore T. Designing relational agents as long term social companions for older adults. In: Intelligent Virtual Agents: 12th International Conference, IVA 2012, Santa Cruz, CA, USA, September, 12-14, 2012 Proceedings: Springer; 2012:289-302. [doi: 10.1007/978-3-642-33197-8_30]
43. Bickmore TW, Caruso L, Clough-Gorr K, Heeren T. 'It's just like you talk to a friend' relational agents for older adults. Interact Comput 2005 Dec;17(6):711-735. [doi: 10.1016/j.intcom.2005.09.002]
44. Beger C, Rüegger D, Lenz A, et al. Effect of a digital blood pressure coach on hypertension management in primary care practices-a pragmatic, randomised controlled trial. Front Digital Health 2025;7:1516600. [doi: 10.3389/fdgth.2025.1516600] [Medline: 40124106]
45. Ibnidris A, Robinson JN, Stubbs M, Piumatti G, Govia I, Albanese E. Evaluating measurement properties of subjective cognitive decline self-reported outcome measures: a systematic review. Syst Rev 2022 Jul 18;11(1):144. [doi: 10.1186/s13643-022-02018-y] [Medline: 35850915]
46. Kowatsch T, Volland D, Shih I, et al. Design and Evaluation of a Mobile Chat App for the Open Source Behavioral Health Intervention Platform MobileCoach: Springer; 2017:485-489. [doi: 10.1007/978-3-319-59144-5_36]

XSL•FO
RenderX

47. Filler A, Kowatsch T, Haug S, Wahle F, Staake T, Fleisch E. MobileCoach: a novel open source platform for the design of evidence-based, scalable and low-cost behavioral health interventions: overview and preliminary evaluation in the public health context. Presented at: 2015 Wireless Telecommunications Symposium (WTS); Apr 15-17, 2015; New York City, NY. [doi: 10.1109/WTS.2015.7117255]
48. Chan WS, Cheng WY, Lok SHC, et al. Assessing the short-term efficacy of digital cognitive behavioral therapy for insomnia with different types of coaching: randomized controlled comparative trial. JMIR Ment Health 2024 Aug 7;11:e51716. [doi: 10.2196/51716] [Medline: 39110971]
49. Castro O, Mair JL, Salamanca-Sanabria A, et al. Development of “LvL UP 1.0”: a smartphone-based, conversational agent-delivered holistic lifestyle intervention for the prevention of non-communicable diseases and common mental disorders. Front Digital Health 2023;5:1039171. [doi: 10.3389/fdgth.2023.1039171] [Medline: 37234382]
50. Ulrich S, Gantenbein AR, Zuber V, Von Wyl A, Kowatsch T, Künzli H. Development and evaluation of a smartphone-based chatbot coach to facilitate a balanced lifestyle in individuals with headaches (BalanceUP App): randomized controlled trial. J Med Internet Res 2024 Jan 24;26:e50132. [doi: 10.2196/50132] [Medline: 38265863]
51. Ulrich S, Lienhard N, Künzli H, Kowatsch T. A chatbot-delivered stress management coaching for students (MISHA App): pilot randomized controlled trial. JMIR Mhealth Uhealth 2024 Jun 26;12:e54945. [doi: 10.2196/54945] [Medline: 38922677]
52. Memory Clinic Bern. URL: https://www.memory-clinic-bern.ch/ [accessed 2025-06-20]
53. Nasreddine ZS, Phillips NA, Bédirian V, et al. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. J Am Geriatr Soc 2005 Apr;53(4):695-699. [doi: 10.1111/j.1532-5415.2005.53221.x] [Medline: 15817019]
54. Davis DH, Creavin ST, Yip JL, Noel-Storr AH, Brayne C, Cullum S. Montreal Cognitive Assessment for the detection of dementia. Cochrane Database Syst Rev 2021 Jul 13;7(7):CD010775. [doi: 10.1002/14651858.CD010775.pub3] [Medline: 34255351]
55. Chia F, Huang WY, Huang H, Wu CE. Promoting healthy behaviors in older adults to optimize health-promoting lifestyle: an intervention study. Int J Environ Res Public Health 2023 Jan 16;20(2):1628. [doi: 10.3390/ijerph20021628] [Medline: 36674395]
56. Davis FD, Bagozzi RP, Warshaw PR. User acceptance of computer technology: a comparison of two theoretical models. Manage Sci 1989 Aug;35(8):982-1003. [doi: 10.1287/mnsc.35.8.982]
57. van der Heijden H. User acceptance of hedonic information systems. MIS Q 2004;28(4):695. [doi: 10.2307/25148660]
58. Venkatesh V, Thong JYL, Xu X. Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology. MIS Q 2012;36(1):157. [doi: 10.2307/41410412]
59. Del Re AC, Flückiger C, Horvath AO, Wampold BE. Examining therapist effects in the alliance-outcome relationship: a multilevel meta-analysis. J Consult Clin Psychol 2021 May;89(5):371-378. [doi: 10.1037/ccp0000637] [Medline: 33829817]
60. Flückiger C, Del Re AC, Wampold BE, Symonds D, Horvath AO. How central is the alliance in psychotherapy? A multilevel longitudinal meta-analysis. J Couns Psychol 2012;59(1):10-17. [doi: 10.1037/a0025749]
61. Flückiger C, Del Re AC, Wampold BE, Horvath AO. The alliance in adult psychotherapy: a meta-analytic synthesis. Psychotherapy (Chic) 2018 Dec;55(4):316-340. [doi: 10.1037/pst0000172] [Medline: 29792475]
62. Horvath AO, Greenberg LS. Development and validation of the Working Alliance Inventory. J Couns Psychol 1989;36(2):223-233. [doi: 10.1037/0022-0167.36.2.223]
63. Falkenström F, Hatcher RL, Skjulsvik T, Larsson MH, Holmqvist R. Development and validation of a 6-item working alliance questionnaire for repeated administrations during psychotherapy. Psychol Assess 2015 Mar;27(1):169-183. [doi: 10.1037/pas0000038] [Medline: 25346997]
64. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
65. Federal act on research involving human beings (Human Research Act, HRA). Swiss Federal Authorities. 2023. URL: https://www.fedlex.admin.ch/eli/cc/2013/617/en [accessed 2025-11-07]
66. Nahum-Shani S, Smith SN, Tewari A, et al. Just-in-time adaptive interventions (JITAIs): an organizing framework for ongoing health behavior support. IHMC Public Cmaps.: The Methodology Center, Penn State; 2014. URL: https://cmapspublic2.ihmc.us/rid=1P0PJNP7V-12LPKBQ-273F/Nahum-Shanietal.-2014-Justintimeadaptiveinterventions(jitais)Ano.pdf [accessed 2025-07-11]
67. Wilson SA, Byrne P, Rodgers SE, Maden M. A systematic review of smartphone and tablet use by older adults with and without cognitive impairment. Innov Aging 2022;6(2):igac002. [doi: 10.1093/geroni/igac002] [Medline: 35243008]
68. Kim M, Patrick K, Nebeker C, et al. The digital therapeutics real-world evidence framework: an approach for guiding evidence-based digital therapeutics design, development, testing, and monitoring. J Med Internet Res 2024 Mar 5;26:e49208. [doi: 10.2196/49208] [Medline: 38441954]
69. Moore K, O’Shea E, Kenny L, et al. Older adults’ experiences with using wearable devices: qualitative systematic review and meta-synthesis. JMIR Mhealth Uhealth 2021 Jun 3;9(6):e23832. [doi: 10.2196/23832] [Medline: 34081020]
70. Huang Y, Upadhyay U, Dhar E, Kuo LJ, Syed-Abdul S. A scoping review to assess adherence to and clinical outcomes of wearable devices in the cancer population. Cancers (Basel) 2022 Sep 13;14(18):4437. [doi: 10.3390/cancers14184437] [Medline: 36139602]

XSL•FO
RenderX

71. Keller OC, Budney AJ, Struble CA, Teepe GW. Blending digital therapeutics within the healthcare system. In: Digital Therapeutics for Mental Health and Addiction: The State of the Science and Vision for the Future: Academic Press; 2023:45-64. [doi: 10.1016/B978-0-323-90045-4.00016-2]
72. Czaja SJ, Charness N, Fisk AD, et al. Factors predicting the use of technology: findings from the Center for Research and Education on Aging and Technology Enhancement (CREATE). Psychol Aging 2006 Jun;21(2):333-352. [doi: 10.1037/0882-7974.21.2.333] [Medline: 16768579]
73. Czaja SJ, Boot WR, Charness N, Rogers WA, Sharit J. Improving social support for older adults through technology: findings from the PRISM randomized controlled trial. Gerontologist 2018 May 8;58(3):467-477. [doi: 10.1093/geront/gnw249] [Medline: 28201730]
74. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 2023 Mar 30;388(13):1233-1239. [doi: 10.1056/NEJMsr2214184] [Medline: 36988602]
75. Hastings J. Preventing harm from non-conscious bias in medical generative AI. Lancet Digital Health 2024 Jan;6(1):e2-e3. [doi: 10.1016/S2589-7500(23)00246-7] [Medline: 38123253]

Abbreviations

CA: conversational agent
CONSORT: Consolidated Standards of Reporting Trials
JITAI: just-in-time adaptive intervention
MCI: mild cognitive impairment
MoCA: Montreal Cognitive Assessment
QoL: quality of life
RQ: research question
SCD: subjective cognitive decline

Edited by A Mavragani; submitted 25.09.24; peer-reviewed by E Hekler, F Temidayo; revised version received 07.05.25; accepted 12.05.25; published 28.07.25.

Please cite as:
Brill E, Vinay R, Nißen M, Joshi P, Klöppel S, Kowatsch T
Toward a Smartphone-Based and Conversational Agent–Delivered Just-in-Time Adaptive Holistic Lifestyle Intervention for Older Adults Affected by Cognitive Decline: Two-Week Proof-of-Concept Study
JMIR Form Res 2025;9:e66885
URL: https://formative.jmir.org/2025/1/e66885
doi:10.2196/66885

© Esther Brill, Rasita Vinay, Marcia Nißen, Priyam Joshi, Stefan Klöppel, Tobias Kowatsch. Originally published in JMIR Formative Research (https://formative.jmir.org), 28.7.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Virtual Reality Interventions of Daily Versus Weekly Data Collection in Patient-Reported Outcomes Among Adults With Cancer: Pilot Survey Study

Matthew Browning[1], PhD; Olivia McAnirlin[1*], PhD; Fu Li[1*], MS; Jeffrey Bertrand[2], PhD; Denise Davis[3], PhD; Kapil Madathil[2], PhD; George Fredric Mau[4], PhD; Teny Henry Gomez[5], MD

[1]Virtual Reality and Nature Lab, Department of Parks, Recreation and Tourism Management, Clemson University, 515 Calhoun Drive, Clemson, SC, United States

[2]Department of Industrial Engineering, Clemson University, Clemson, SC, United States

[3]Department of Integrated Information Technology, Molinaroli College of Engineering and Computing, University of South Carolina, Columbia, SC, United States

[4]Watermark Counseling, Columbia, SC, United States

[5]Division of Hospice and Palliative Medicine, Department of Internal Medicine, Prisma Health Upstate, University of South Carolina School of Medicine Greenville (USCOM-Greenville), Columbia, SC, United States

[*]these authors contributed equally

Corresponding Author:
Matthew Browning, PhD
Virtual Reality and Nature Lab, Department of Parks, Recreation and Tourism Management, Clemson University, 515 Calhoun Drive, Clemson, SC, United States

Abstract

Background: Virtual reality (VR) interventions are increasingly used in health care settings to improve patient-reported outcomes (PROs). PROs are commonly evaluated at weekly intervals with data collected via digital surveys. While weekly assessments have benefits, VR devices enable more frequent in-device data collection. It remains unclear whether PROs collected more frequently provide more information on these interventions than PROs collected more infrequently.

Objective: This pilot study explored differences between daily and weekly PRO data collection in a VR intervention with nature imagery, with and without guided imagery, among patients with cancer.

Methods: Patients with cancer (n=8) were randomly assigned to one of four intervention groups: (1) virtual reality–assisted guided imagery (VRAGI), (2) VR without guided imagery, (3) desktop VR with guided imagery, or (4) desktop VR without guided imagery. Devices were mailed to participants' homes for 15 - 20 minutes of daily use over 3 weeks. Weekly outcomes (pain, anxiety, depression, and well-being) were assessed using items from the Edmonton Symptom Assessment Scale. Daily outcomes were captured via in-device pre-post surveys. Data were analyzed descriptively, using visual trend comparisons to explore patterns.

Results: Of 41 patients who consented, 8 provided complete and usable data. Weekly outcomes showed no consistent trends. In contrast, daily data revealed more nuanced patterns, such as early symptom relief, plateaus, and "double-bottom" effects. The addition of guided imagery did not consistently enhance outcomes beyond VR alone, although the VRAGI condition showed the greatest improvement in well-being. Given the small sample size, these findings should be considered exploratory.

Conclusions: This pilot study suggests that daily PRO data might offer richer insight into intervention effects than weekly assessments. Further research with larger samples is needed to confirm these patterns.

Trial Registration: ClinicalTrials.gov NCT05348174; https://clinicaltrials.gov/study/NCT05348174

International Registered Report Identifier (IRRID): RR2-10.1136/bmjopen-2022-064363

(JMIR Form Res 2025;9:e73506) doi:10.2196/73506

KEYWORDS

virtual reality; guided imagery; survey administration; pain management; research design

XSL•FO
RenderX

Introduction

Health care interventions using virtual reality (VR) are often evaluated by testing their efficacy at weekly intervals [1,2]. Data can be collected via digital surveys, paper questionnaires, SMS text messages, or physician notes on medication use and psychological outcomes (ie, pain, anxiety, and depressive symptoms) and compared with these same measures weeks later [3,4]. For example, Pallavicini et al [5] compared stress, anxiety, and depression levels before and after 1 week of VR at-home intervention, and Hernandez et al [6] compared depression, well-being, and quality of life immediately after VR, as well as 3 and 6 months after VR.

Collecting data relatively infrequently (ie, weekly) has benefits. Using weekly push notifications from the REDCap (Research Electronic Data Capture) system, patients can complete HIPAA (Health Insurance Portability and Accountability Act)-compliant surveys with minimal burden [7]. Additionally, patients may be more likely to complete surveys if assessments are spaced further apart [8]. However, while less frequent data collection can provide insight into the outcomes of VR interventions, they may overlook shorter-term symptom changes between assessments.

What is relatively unknown is whether frequent data collection is more informative for measuring patient-reported outcomes (PROs) than infrequent data collection. In a broader review, Peasgood et al [8] found that daily assessments tend to yield lower symptom ratings than 7-day recall, suggesting that frequency may impact PRO sensitivity. Conveniently, VR devices already allow for frequent data collection. In-device surveys can use approaches such as visual analog scales [9], asking participants how they feel in the moment to collect data as close to the VR intervention as possible with minimal recall bias [10]. Such data can be collected and stored on VR devices for retrieval when devices are returned to the research team. Alternatively, health care providers' daily notes can be used to describe patients' symptoms of their patients experiencing VR [11].

To the best of our knowledge, no studies have directly compared daily and weekly PRO data with a VR intervention, especially one administered at home, with head-mounted display (HMD) and desktop options, and with or without guided imagery—a therapeutic technique shown to enhance VR's effects in some contexts [12]. This knowledge gap is especially relevant in cancer care, where patients experience fluctuating symptoms and may benefit from more tailored, flexible interventions and data collection methods [13,14].

Therefore, in this study, we examined weekly and daily data from a VR intervention on PROs among patients with varying stages of cancer diagnoses. Our intervention consisted of a virtual computer-generated natural environment similar to other virtual environments that distract and induce relaxation [15,16]. We additionally examined how adding guided imagery to our intervention influenced PROs. Guided imagery is audio narration meant to invoke one or more of the senses and guide an individual to mentally visualize experiences to reduce pain, anxiety, and depression and to improve well-being [12]. To generalize to more types of digital health interventions, we provided the intervention at 2 immersion levels—delivery in an HMD and in "desktop VR"—where patients interact with a 3D multimedia environment on a computer screen [17,18]. This investigation builds off past research showing VR and guided imagery interventions in isolation can improve PROs among patients with cancer [19-28]. Our research question was: Is collecting data at weekly intervals sufficient to predict trends in PROs, or is daily data collection more informative?

Methods

Study Context

This study was part of a larger project examining the impact of virtual reality–assisted guided imagery (VRAGI) dimensions on cancer pain in the home setting [29]. The original intent was to assess the efficacy and safety of VRAGI on opioid use, anxiety, depression, and fatigue and compare those effects with a VR program absent of guided imagery and laptop versions of these two interventions (with and without guided imagery). The results reported here are limited to outcomes available from data collected at both daily and weekly intervals. Only these specific data and outcomes could answer the research question for this study. The completed CONSORT (Consolidated Standards of Reporting Trials) checklist for pilot and feasibility studies [30] is provided in Checklist 1.

Ethical Considerations

The study was approved by the Prisma Health institutional review board (Pro00114598) and published on ClinicalTrials.gov (NCT05348174). Two clinics at Prisma Health Cancer Institute in Greenville, SC (outpatient oncology and palliative care) were the sites of recruitment. This institute is accredited by the American College of Surgeons Commission and serves over 10,000 cases per year by more than 30 adult oncology specialists. Recruitment took place through flyers as well as word-of-mouth by the clinical and research staff during inpatient visits. All patients provided verbal or written informed consent, which occurred 2 - 3 weeks before the intervention start date. Patients earned a US $25 e-gift card at the end of weeks 1 and 2 if they completed ≥75% of the surveys during each of those weeks, a third US $25 e-gift card if they completed ≥75% of the surveys and initiated the device return after week 3, and a fourth US $25 e-gift card if they completed ≥75% of the survey at week 6. Respondents were assigned a unique, nonidentifiable participant number, and collected data were associated with this number rather than identifiable information to maintain their confidentiality.

Eligibility and Preliminary Screening

Verification of eligibility relied on electronic health records (EHRs) and follow-up phone conversations with the study coordinator. Inclusion criteria included patients >18 years of age, baseline pain score on the Edmonton Symptom Assessment Scale (ESAS [31]) ≥4 (mean score at 7-day screening), ability to provide consent and willingness to comply with all study procedures, and ability to comprehend spoken and written English. We initially set the inclusion criteria to also include an advanced cancer diagnosis—defined as incurable cancer,

XSL•FO
RenderX

including locally advanced and metastatic cancers with no plan for resection during the study period—but broadened this criterion to a cancer diagnosis to expand the eligible sample pool. Exclusion criteria are as follows: patients >65 years of age; having a condition that interfered with VR such as a history of seizures, facial injury precluding safe placement of an HMD, or other visual or hearing impairment; previously participated in a VR clinical trial; underwent a surgical procedure within 8 weeks; diagnosed with serious mental illness; have brain metastases; have a neurocognitive disorder according to past medical history; have a prognosis of <3 months from the time of enrollment per treating oncologist; experience current substance abuse; or have experienced complex childhood trauma.

All eligible patients completed a 4-day screening, which evaluated their adherence to complete surveys. The screening survey used the same instruments as the weekly surveys in the actual study. Patients who completed fewer than 50% of the surveys provided during the screening were dismissed before the randomization process and intervention.

All but one member of the research team were blinded to the intervention group assigned to each participant. The one exception was a single researcher who conducted the randomization and device shipping and receiving. Patients were partially blinded to the treatment group. They knew the device type (laptop vs HMD) but were unaware of assignments to AGI (with guided imagery) versus NG (no guided imagery) groups.

Intervention

Patients were randomly assigned, while considering gender, to one of four conditions: VRAGI, VR-NG (virtual reality with no guided imagery), Laptop-AGI, and Laptop-NG. VR meant the intervention was delivered via an HMD, specifically a Meta Quest 2, while the laptop meant it was delivered on a Windows laptop (13″ or 15″ screen with 1080p resolution). AGI meant the intervention included guided imagery. NG meant the content included no guided imagery and presented simply the nature-based imagery and soundscape.

All 4 conditions were experienced at the patients' home while seated. During the study, patients were asked to engage in the intervention daily (15‑20 minutes/day) for 3 consecutive weeks. During the intervention, patients were allowed to undergo any cancer treatments outside the exclusion criteria, as prescribed by their health care provider.

All devices and accessories were packaged, shipped to patients' homes, and ready to be used once unpacked. Virtual content was preloaded as an offline app, so no internet was needed, and patients navigated to the app with the provided instructions. Patients in the VR conditions were trained on using the HMD through a printed color manual. Patients in the laptop conditions were provided with a similar manual. To align with the audio experience of the HMD's near-ear speaker, we provided headphones for patients in the laptop conditions to wear during the intervention. Patients could receive help from researchers by email and a telephone hotline for technical support.

Virtual Content

The virtual content consisted of a computer-generated immersive virtual landscape with nature-based imagery (ie, trees, birds, mountains, water, and birds) aligned with the landscape preferences and perceived restorative outcomes of past virtual nature scenery [16,32,33]. All conditions also had a nature-based soundscape (ie, water flowing and birds chirping). Additionally, the AGI conditions included an omnipresent vocal narration that overlapped with the imagery in the virtual scene and steered users through mental escape, redirection of attention, anxiety alleviation, and a nonjudgmental acceptance psychological process [29]. Patients were asked to complete the intervention in calendar order (ie, week 1 was the spring season, week 2 was the summer, and week 3 was the fall) (Figure 1).

The virtual content was developed across 14 months (December 2020 to February 2022) by an interdisciplinary collaboration of physicians, mental health counselors, human systems engineers, and VR content experts. Patients in all 4 conditions saw the same virtual content and could look around from all angles by moving their heads in the HMD or laptop cursor. No intervention allowed for "walking" within the virtual world.

Figure 1. Seasons available in the virtual environments.

Measures

Patient Characteristics

Demographic data were provided from the patient's EHR. These data included date of birth, age, sex, marital status, education, household income, employment status, insurance status, and place of birth. An EHR review was also used to collect patients' medical record numbers, type of cancer, and other health status measures.

Weekly Data Collection of Patient-Reported Outcomes

Weekly outcomes of pain, anxiety, depression, and well-being were measured with the ESAS [31]. The ESAS is a 9-item patient-rated symptom instrument developed for use with patients receiving palliative care. In this analysis, we used the individual items measuring patient-reported anxiety, depression, pain, and well-being on an 11-point visual analog scale (ie, 0=none to 10=worst possible). The ESAS has been extensively validated for assessing symptom intensity, including among

XSL•FO
RenderX

patients with cancer [34]. A change of 1 point in the ESAS has been suggested as clinically significant [35].

All patients were prompted via an email through REDCap and asked to fill out the ESAS at the start and end of each intervention week (days 1, 7, 8, 14, 15, and 21). They also received an email reminder through REDCap if they did not complete the ESAS before the end of that day. The weekly response was recorded and stored in REDCap alongside other instruments collected weekly [29].

Daily Data Collection of Patient-Reported Outcomes

Daily outcomes of pain, anxiety, depression, and well-being were measured with an in-device survey instrument designed for this study that was administered before and after each VR session on an 11-point visual analog scale (Figure 2). Patients were required to answer all the survey items before starting the VR intervention and again before quitting the VR application. The order of the survey items was randomized each time they were shown. The daily responses were recorded and stored on the study device (HMD or laptop) and retrieved by the researchers once the patients returned the device.

Figure 2. In-device survey for daily data collection of patient-reported outcomes.

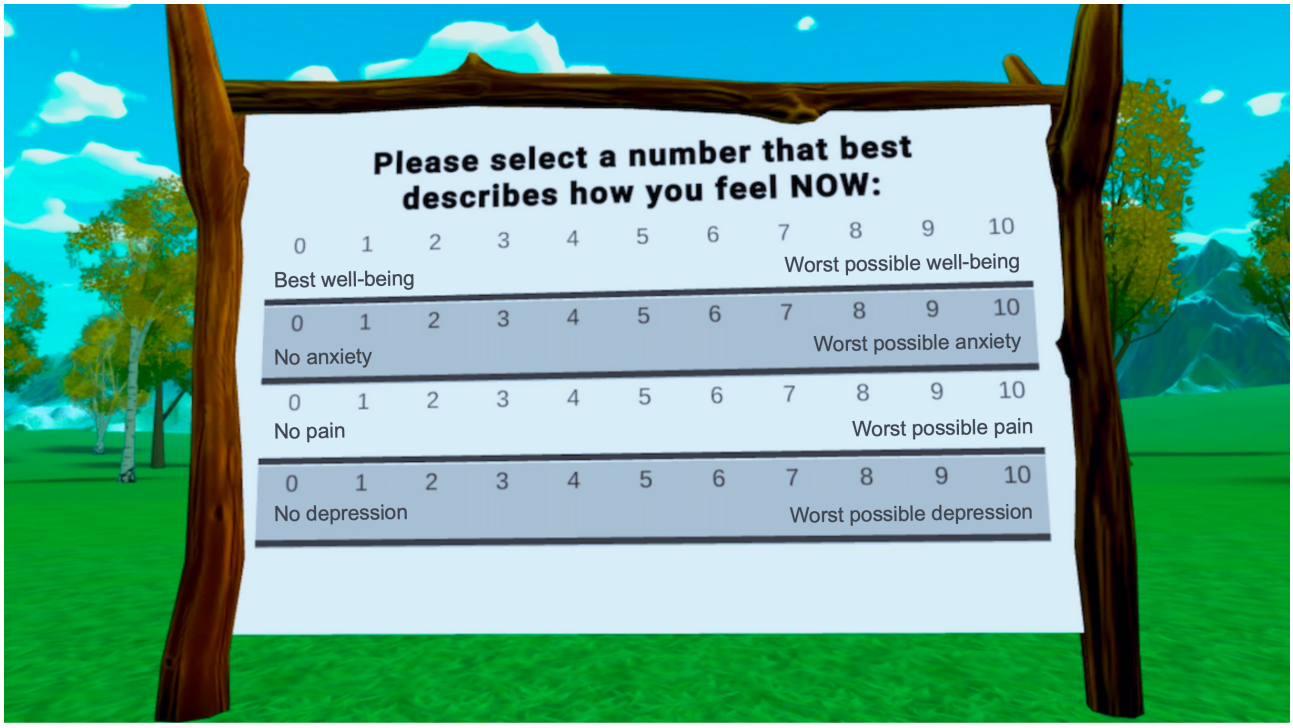

Data Availability and Analysis

We faced considerable challenges in recruiting, consenting, and retaining our sample of patients with cancer. Across 17 months of recruitment (February 2023 to January 2024), we collected usable daily and weekly data from 8 patients. All members of the research team were in consensus to stop the trial due to these challenges and initiate data analysis. Due to the small sample size, this pilot study was not powered for inferential statistical analyses. Accordingly, we employed a descriptive approach using visual inspection of trends to explore preliminary patterns in PROs. For daily data, we calculated pre-post differences for each session and applied LOESS (locally estimated scatterplot smoothing) to visualize time trends by intervention group. For weekly data, raw group means were plotted without smoothing, given the limited number of time points. No imputation was performed; only complete observations were used for each data stream. Ultimately, this visual strategy of analysis was aimed at addressing our research question and generating hypotheses that could inform future research.

Results

Patient Recruitment and Characteristics

We consented 41 eligible patients, but only 11 passed the screening week criteria (Figure 3). Nine patients were then randomized to VRAGI (n=3), VR-NG (n=1), Laptop-AGI (n=3), and Laptop-NG (n=2), but one patient in the Laptop-AGI condition only completed the intervention once, providing 1 day of in-device data. Therefore, our final sample size for analysis was 8.

Table 1 shows the characteristics of the patients. One-quarter were male, and the average age was 49.1 (range 30-60). All patients were non-Hispanic White. In contrast, dropout participants (n=33) had a mean age of 55 years, with 55% (n=18) female, 67% (n=22) White, 33% (n=11) African American, and 91% (n=30) non-Hispanic.

XSL•FO
RenderX

Figure 3. Study flowchart of patient recruitment and available data. AGI: with guided imagery; NG: no guided imagery; VRAGI: virtual reality–assisted guided imagery; VR-NG: virtual reality with no guided imagery.

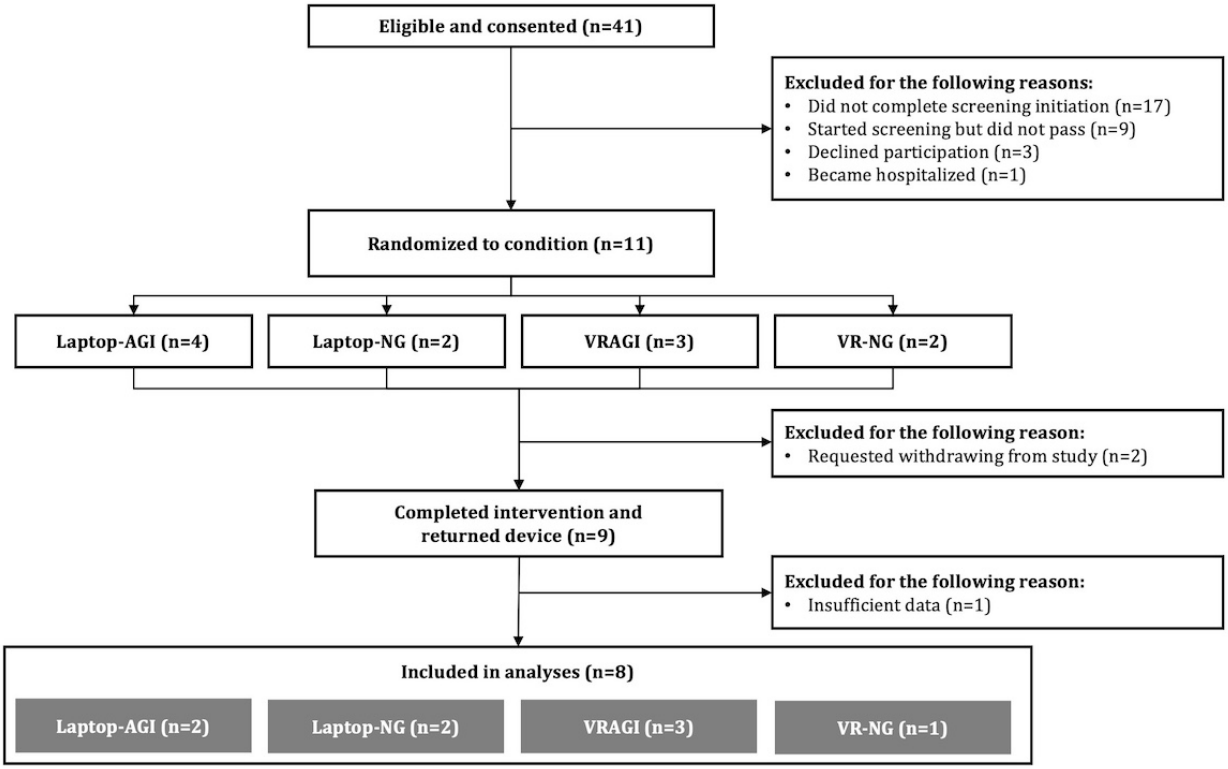

Table . Characteristics of patients.

Characteristic	Total (n=8)	Laptop-AGI[a] (n=2)	Laptop-NG[b] (n=2)	VRAGI[c] (n=3)	VR-NG[d] (n=1)
Age (years), mean (SD)	49 (10)	59 (2)	51 (8)	44 (13)	42 (0)
Sex, n (%)					
Male	2 (25)	0 (0)	1 (50)	1 (33)	0 (0)
Female	6 (75)	2 (100)	1 (50)	2 (67)	1 (100)
Race, n (%)					
White	8 (100)	2 (100)	2 (100)	3 (100)	1 (100)
Ethnicity, n (%)					
Non-Hispanic	8 (100)	2 (100)	2 (100)	3 (100)	1 (100)
Marital status, n (%)					
Single	1 (13)	0 (0)	0 (0)	0 (0)	1 (100)
Married	6 (75)	2 (100)	2 (100)	2 (67)	0 (0)
Not reported	1 (13)	0 (0)	0 (0)	1 (33)	0 (0)

[a]AGI: with guided imagery.
[b]NG: no guided imagery.
[c]VRAGI: virtual reality–assisted guided imagery.
[d]VR-NG: virtual reality with no guided imagery.

Weekly Versus Daily Data on Pain

Weekly pain score data showed no consistent patterns (Figure 4). In contrast, daily data from the Laptop-NG and VR-NG conditions revealed double-bottom patterns (ie, initial decline, recovery, and secondary decline). Patients reported improvements in pain within the first few days of the intervention and again 1-2 weeks afterward. While these results showed differences in PROs between conditions with and without guided imagery, there was insufficient evidence to suggest that the addition of guided imagery improved PROs.

XSL•FO
RenderX

Figure 4. Changes in pain scores based on data collected at weekly versus daily intervals. (**A**) Weekly ESAS pain scores. (**B**) Daily in-device pain score differences. AGI: with guided imagery; ESAS: Edmonton Symptom Assessment Scale; NG: no guided imagery; VRAGI: virtual reality–assisted guided imagery; VR-NG: virtual reality with no guided imagery.

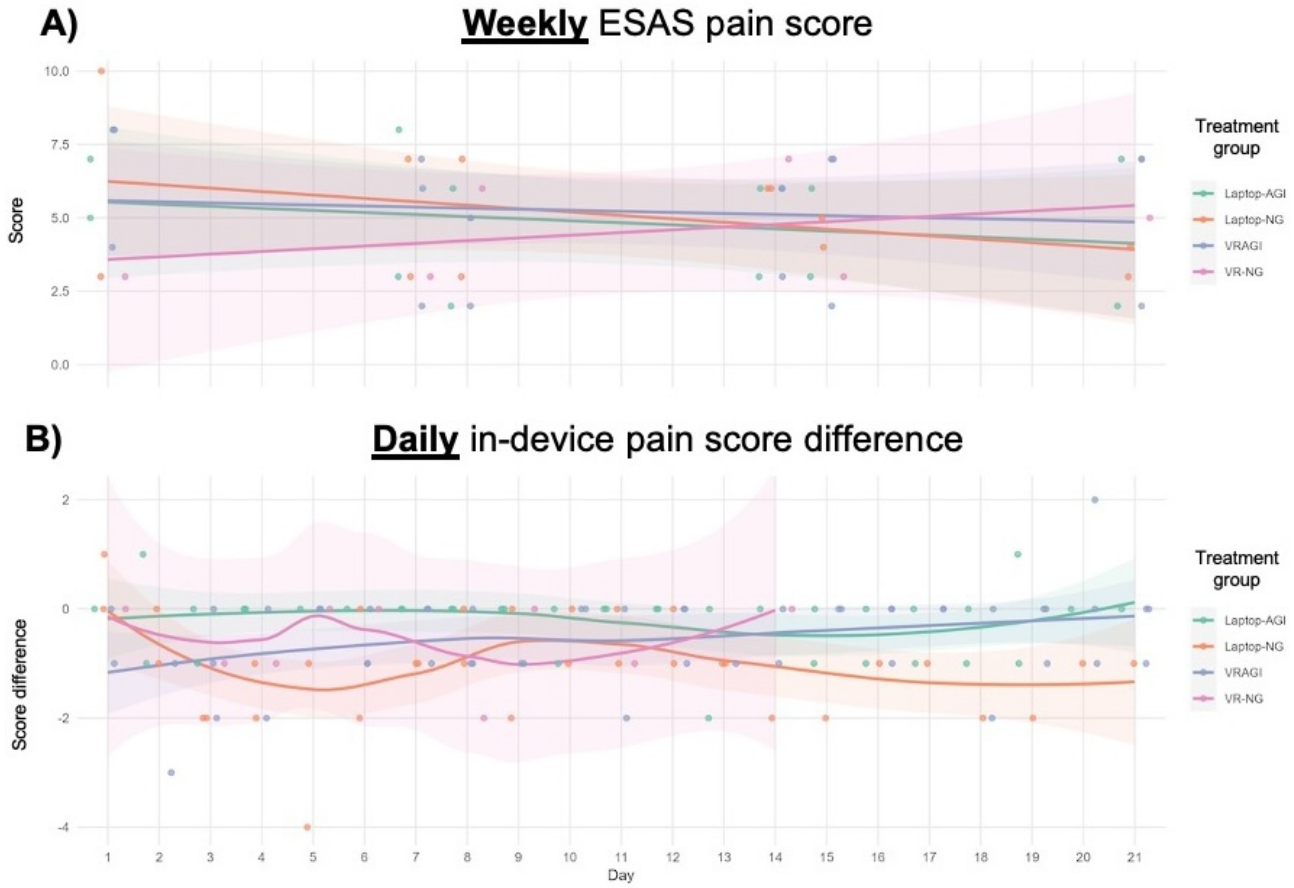

Weekly Versus Daily Data on Anxiety

Weekly data from the intervention showed no strong effects on anxiety (Figure 5). The Laptop-NG condition exhibited the greatest decrease in anxiety over time, while the VR-NG condition showed the greatest increase, though both effects were modest. In contrast, daily data revealed clearer trends for the VR-NG and VRAGI conditions, with anxiety levels plateauing over time. Initial benefits were observed within the first 2 days of the VR-NG intervention and within the first week of VRAGI. However, these effects did not continue to increase with prolonged use of the intervention. These results also provided insufficient evidence that the addition of guided imagery assisted with PROs.

XSL•FO
RenderX

Figure 5. Changes in anxiety scores based on data collected at weekly versus daily intervals. (**A**) Weekly ESAS pain scores. (**B**) Daily in-device pain score differences. AGI: with guided imagery; ESAS: Edmonton Symptom Assessment Scale; NG: no guided imagery; VRAGI: virtual reality–assisted guided imagery; VR-NG: virtual reality with no guided imagery.

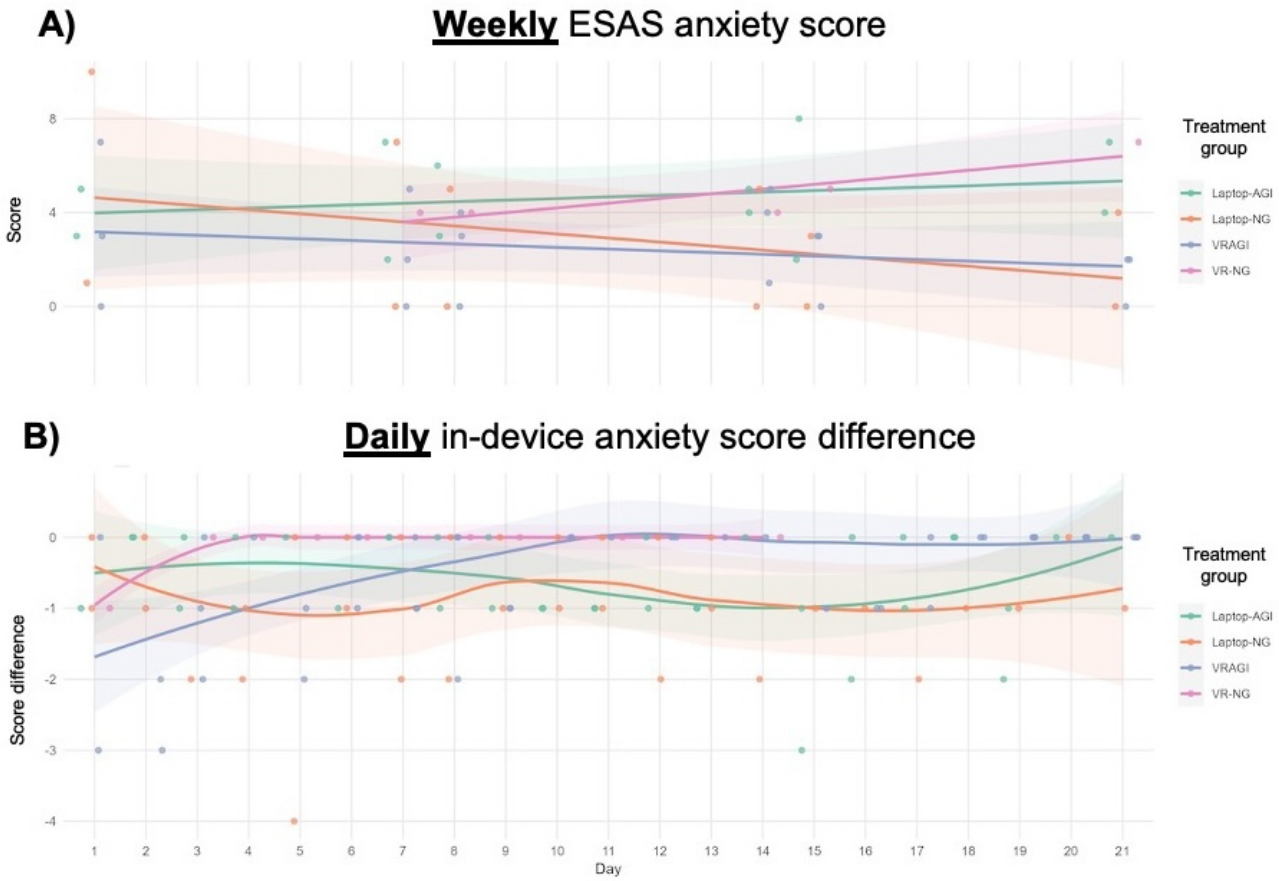

Weekly Versus Daily Data on Depression

Weekly data from the intervention showed no strong effects on depression (Figure 6). However, daily data from VR-NG showed a plateauing pattern where the intervention's benefits were observed in the first 2 days of device use but not afterward. These results once again provided no evidence that the addition of guided imagery assisted with PROs.

XSL•FO
RenderX

Figure 6. Changes in depression scores based on data collected at weekly versus daily intervals. (**A**) Weekly ESAS pain scores. (**B**) Daily in-device pain score differences. AGI: with guided imagery; ESAS: Edmonton Symptom Assessment Scale; NG: no guided imagery; VRAGI: virtual reality–assisted guided imagery; VR-NG: virtual reality with no guided imagery.

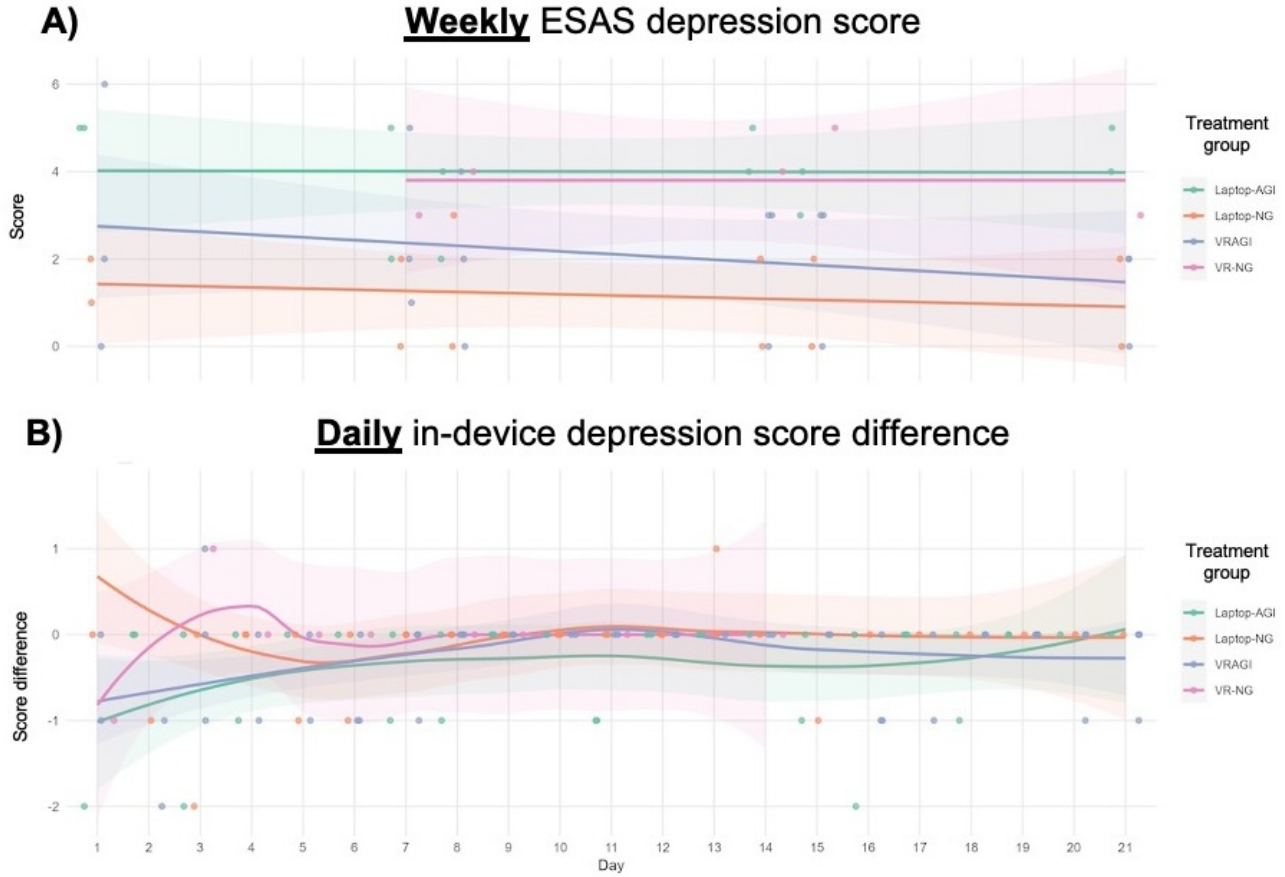

Weekly Versus Daily Data on Well-Being

Weekly data from well-being scores suggested downward trends for VRAGI and VR-NG (Figure 7). Daily data from VR-NG showed a plateauing pattern similar to the depression data pattern, where the intervention's benefits were observed in the first couple of days but not afterward.

These results provided evidence that the addition of guided imagery may have assisted with PROs measured weekly but not daily—results which may inform future trials. No adverse events or harms were reported.

XSL•FO
RenderX

Figure 7. Changes in well-being scores based on data collected at weekly versus daily intervals. (**A**) Weekly ESAS pain scores. (**B**) Daily in-device pain score differences. AGI: with guided imagery; ESAS: Edmonton Symptom Assessment Scale; NG: no guided imagery; VRAGI: virtual reality–assisted guided imagery; VR-NG: virtual reality with no guided imagery.

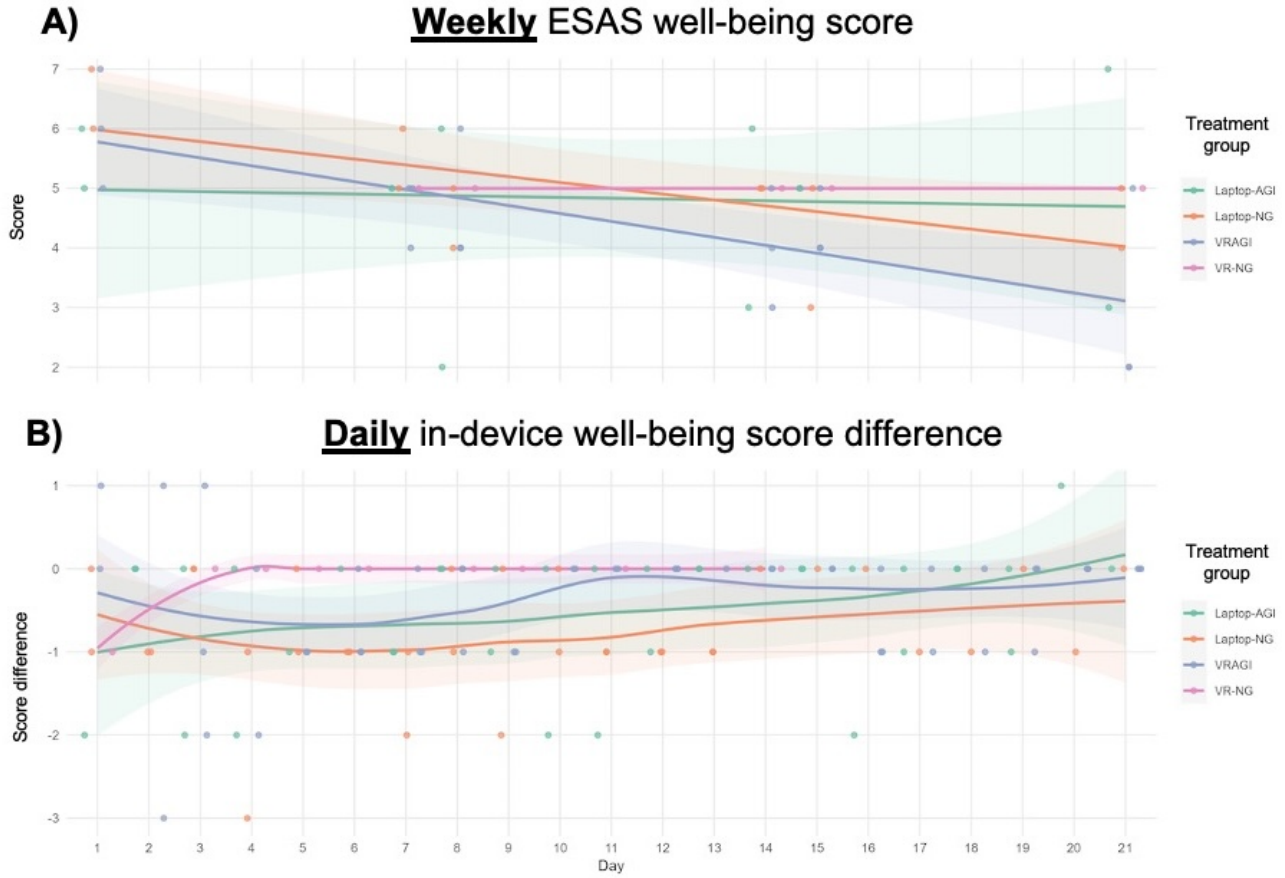

Discussion

Summary of Main Findings

In summary, 8 patients analyzed within this technology-based intervention provided PRO data on daily and weekly schedules. Weekly data showed inconsistent patterns, while daily data revealed more informative trends such as double-bottom patterns and plateauing effects for some outcomes and conditions.

Weekly data for pain, anxiety, and depression PROs had no consistent patterns over time. However, weekly data suggested that the laptop intervention with virtual nature imagery (and no guided imagery) had the greatest decrease in anxiety. The VR HMD intervention with virtual nature imagery (and again, no guided imagery) showed the greatest increase in anxiety over time.

Daily data showed different findings. The greatest benefit for pain was observed in the laptop and HMD interventions (without guided imagery) but only within the first few days and then 1-2 weeks later. Daily data also showed benefits for anxiety among the HMD conditions but only within the first few days to 1 week. Likewise, depression and well-being plateaued after a few days in the VR condition without guided imagery.

Comparing weekly and daily data illustrates the importance of PROs at varying time points during a digital intervention. Without daily data, for instance, there may be fewer insights into the trends and patterns over the course of the intervention.

The rationale behind combining VR and guided imagery lies in complementary mechanisms for modulating attention and emotional processing. Nature-based VR can benefit patients through distraction and relaxation since natural landscapes can capture people's soft fascination, align with our evolutionary needs of survival, and recall positive memories [15,16,33]. Meanwhile, guided imagery can redirect the user's attention away from the negative emotional processing of pain [12]. While our results did not confirm these additive effects of nature-based VR and guided imagery, further work with fully powered samples is warranted to disentangle whether these 2 modalities complement each other and what mechanisms drive therapeutic benefits.

Implications for Researchers and Clinicians

This study has implications for researchers and clinicians interested in using VR for improving PROs. It is important to consider when designing a VR clinical trial how frequently to measure PROs. There are benefits and drawbacks to daily versus weekly survey administrations, such as patient burden or biases resulting from repeated survey instruments. Still, daily surveys may allow for in-depth tracking of an intervention's efficacy. Daily approaches may thus allow for a more personalized approach to these interventions. Research coordinators could communicate with patients after PROs severely decrease, increase, or plateau and offer intervention support or follow-up data collection on the severe changes in PROs.

XSL•FO
RenderX

Daily PRO collection may also allow researchers and clinicians to track when a patient group or individual may discontinue an intervention. For example, a patient group may be on a standardized intervention schedule (ie, 8 weeks), but their daily PROs may plateau in week 4. This indicates that patients may not need to continue the intervention, saving resources and time and ultimately helping patients reap the benefits of PROs quicker and through a less demanding or intensive intervention process. As a hypothetical example, an adaptive clinical trial with cancer patients may begin with daily VR sessions and complete short in-device PROs after each session. If a participant reports sustained symptom relief (eg, ≥1-point ESAS improvement maintained for 3 consecutive days), the intervention could be reduced in frequency to every other day. Alternatively, if a patient reports worsening symptoms or stagnation, the device could trigger the researchers to provide the patient with additional support (eg, switching VR modalities or initiating a follow-up call) or terminate the intervention. Still, it should be acknowledged that plateauing of daily PRO surveys could also be due to familiarity of survey items, as patients may have defaulted to providing the same values each time they saw the in-device surveys, even when items are randomized.

Patient samples provide opportunities for applied research where VR and VR interventions benefit those in greatest need. However, we experienced challenges when using patients, specifically those with cancer. Previous literature emphasizes several barriers to participation, including fluctuating symptom burdens, fatigue, logistical issues, and psychological stress associated with illness progression, which collectively complicate sustained involvement in clinical research [36]. In our study, among the 41 patients who provided informed consent, we observed substantial attrition before the intervention initiation, with 80% (33/41) not choosing to follow up, including all 11 patients who were Black—a disproportionate dropout raising concerns about equity and suggesting that our recruitment and screening may have inadvertently failed to support the inclusion of underrepresented populations. This high attrition among Black participants aligns with a broader historical pattern, rooted in longstanding systemic mistrust stemming from unethical research practices, which have understandably led many Black Americans to be cautious about participating in clinical research [37]. Integrating trusted health care providers early in the recruitment and retention process, as well as culturally tailored communication strategies, may have helped mitigate these issues and improve retention rates [38,39].

More broadly, this relatively high dropout rate during screening may be attributed to multiple factors. For instance, the extended timeline between consent and screening (1 week) and the involvement of non–health care personnel (the screening-intervention process was supervised by nonclinical research team members) may have introduced communication inefficiencies and spurred dropout. Streamlining recruitment-to-screening processes and integrating health care team members in screening processes might improve early retention rates. Embedding trusted health care providers as contacts for VR interventions may also ensure patients are invested throughout the study process. Trusted health care providers are likely positioned to build trust and address concerns throughout the study process. This trust-based approach may be crucial in studies requiring sustained patient engagement, such as daily interventions.

Despite the high preintervention attrition, the retention rate among patients who commenced the intervention was encouraging, with 73% (8/11) of patients completing the study protocol. Patients who demonstrated higher engagement and motivation during the screening process tended to show better adherence during the intervention. This suggests that VR interventions measuring PROs may benefit from screening processes that mimic the surveys and patient responsibilities required during the intervention.

For researchers and clinicians using at-home VR interventions, there are challenges to balancing device functionality and monitoring. We ran into minor incidents, including unauthorized device use by a minor, concerns about illegal data entry, devices malfunctioning and needing to be replaced, and packages missing before arrival. Still, we expect our challenges to have been greater if devices were connected to the internet. Future protocols can be designed to assist VR researchers and clinicians to balance competing priorities, including stable device performance, patient safety, sensitive information, and unauthorized use, while still enabling data collection of PROs and adherence.

Strengths and Limitations

This study had both strengths and limitations. One strength was the combination of immersive nature imagery in VR with guided imagery that could be self-administered at home without an internet connection. The completion rate among our intervention initiators was high (>70%), suggesting this intervention was generally feasible once patients completed the consent and screening criteria. However, our findings were further limited by the small sample size given significant recruitment constraints and our inability to conduct live monitoring or biofeedback to track physiological indicators and validate PROs.

Further limiting our study were the methodological differences between the daily and weekly survey measures, given that the original study was not designed for this purpose. Patients were required to complete daily surveys to access the intervention, whereas weekly surveys were optional and incentivized. These differences in requirements, incentives, and reminder frequencies between daily and weekly data collection methods could have influenced participant response patterns and introduced biases, complicating direct comparisons of the datasets. Still, our exploratory findings may be informative for several reasons. First, patterns emerged in the daily data—such as plateau effects—which weekly data failed to detect, highlighting the inherent sensitivity of more frequent assessments. Second, our findings align with other studies suggesting that daily assessments may more accurately capture transient symptom changes [8]. Finally, recognizing such nuances from our exploratory data can inform future study designs, emphasizing the need to consider data collection frequency carefully to balance the sensitivity and respondent burden of PROs.

XSL•FO
RenderX

Conclusions

When designing VR interventions, it may be important to consider how often to collect PROs. This pilot study suggests that daily data collection may be more informative for researchers and clinical teams than less frequent data collection. Future research is needed to substantiate these results—which were not only mixed but also based on a small sample size—and to evaluate how data collection frequency and VR interventions such as guided imagery impact PROs.

Acknowledgments

The authors thank Mondira Bardhan for assisting with data collection and recruitment, Andrew Todd for developing an early version of the software and for assisting with developing video tutorials on how to use the headset, Kailan Sindelar for assisting with developing video tutorials on how to use the headset and contributing to the study design, George Hartshorn for assisting with contributing to the study design, Rosalba Hernandez for assisting with contributing to the study design, and Shyam Ranganathan for assisting with contributing to the study design. This research was funded by the Clemson University Research Foundation (CU-Prisma Innovation Fund PPN 2022001197). The virtual reality intervention used in this study, formerly referred to as VRAGI (virtual reality–assisted guided imagery), has since been branded as GuidedNature XR and is intended for broader dissemination through Carpe Reality LLC, a company founded by 2 coauthors (OM and MB) following study completion. Finally, we used the generative AI tool ChatGPT by OpenAI [40] to draft small sections and improve the readability of the main text.

Data Availability

Deidentified data on patient-reported outcomes are available from the corresponding author on reasonable request.

Authors' Contributions

MB contributed to conceptualization, funding acquisition, software, formal analysis, methodology, visualization, original draft writing and editing, and supervision. OM contributed to conceptualization, project administration, methodology, original draft writing and editing, and supervision. FL contributed to project conceptualization, administration, investigation, software, formal analysis, methodology, visualization, original draft writing and editing, data curation, and supervision. JB contributed to conceptualization, project administration, software, methodology, visualization, original draft reviewing and editing, data curation, and supervision. DD contributed to investigation, software, formal analysis, methodology, original draft editing, and data curation. GFM contributed to conceptualization, software, methodology, narration, and original draft reviewing and editing. KM contributed to conceptualization, funding acquisition, project administration, software, methodology, visualization, original draft reviewing and editing, and supervision. THG contributed to conceptualization, funding acquisition, project administration, methodology, original draft reviewing and editing, data curation, and supervision. All authors approved the final version of the manuscript.

Conflicts of Interest

The virtual reality software will be made publicly available by Carpe Reality LLC, a company cofounded by 2 of the authors (OM and MB). While the intervention was initially developed with academic and grant support, the authors disclose this potential commercial interest in the platform.

Checklist 1
CONSORT-eHEALTH checklist (v1.6.1).
[PDF File, 148 KB - formative_v9i1e73506_app1.pdf]

References

1. Birckhead B, Eberlein S, Alvarez G, et al. Home-based virtual reality for chronic pain: protocol for an NIH-supported randomised-controlled trial. BMJ Open 2021 Jun 15;11(6):e050545. [doi: 10.1136/bmjopen-2021-050545] [Medline: 34130965]
2. Spiegel B, Fuller G, Lopez M, et al. Virtual reality for management of pain in hospitalized patients: a randomized comparative effectiveness trial. PLoS One 2019;14(8):e0219115. [doi: 10.1371/journal.pone.0219115] [Medline: 31412029]
3. Rocque GB, Patel MI. Asking the next generation of electronic patient-reported outcome intervention questions-what components are needed? JAMA Netw Open 2024 Sep 3;7(9):e2433164. [doi: 10.1001/jamanetworkopen.2024.33164] [Medline: 39269709]
4. Aiyegbusi OL, Roydhouse J, Rivera SC, et al. Key considerations to reduce or address respondent burden in patient-reported outcome (PRO) data collection. Nat Commun 2022 Oct 12;13(1):6026. [doi: 10.1038/s41467-022-33826-4] [Medline: 36224187]
5. Pallavicini F, Orena E, Arnoldi L, et al. Effects and acceptability of a 1-week home-based virtual reality training for supporting the management of stress and anxiety: randomized pilot trial. JMIR Serious Games 2025 Mar 6;13:e50326. [doi: 10.2196/50326] [Medline: 40053782]

XSL•FO
RenderX

6. Hernandez R, Wilund K, Solai K, et al. Positive psychological intervention delivered using virtual reality in patients on hemodialysis with comorbid depression: protocol and design for the joviality randomized controlled trial. JMIR Res Protoc 2023 Jun 16;12:e45100. [doi: 10.2196/45100] [Medline: 37327026]
7. Soni H, Ivanova J, Wilczewski H, et al. User preferences and needs for health data collection using Research Electronic Data Capture: survey study. JMIR Med Inform 2024;12:e49785. [doi: 10.2196/49785] [Medline: 38917448]
8. Peasgood T, Caruana JM, Mukuria C. Systematic review of the effect of a one-day versus seven-day recall duration on patient reported outcome measures (PROMs). Patient 2023 May;16(3):201-221. [doi: 10.1007/s40271-022-00611-w] [Medline: 36786931]
9. McCormack HM, de L. Horne DJ, Sheather S. Clinical applications of visual analogue scales: a critical review. Psychol Med 1988 Nov;18(4):1007-1019. [doi: 10.1017/S0033291700009934]
10. Tsigarides J, Grove V, Chipping J, et al. Feasibility of attention-based virtual reality interventions in fibromyalgia syndrome: comparing systems, virtual environments and activities. Br J Pain 2025 Jan 4;19(2):20494637241310696. [doi: 10.1177/20494637241310696] [Medline: 39763629]
11. Appel L, Appel E, Kisonas E, et al. Evaluating the impact of virtual reality on the behavioral and psychological symptoms of dementia and quality of life of inpatients with dementia in acute care: randomized controlled trial (VRCT). J Med Internet Res 2024 Jan 30;26:e51758. [doi: 10.2196/51758] [Medline: 38289666]
12. Roffe L, Schmidt K, Ernst E. A systematic review of guided imagery as an adjuvant cancer therapy. Psychooncology 2005 Aug;14(8):607-617. [doi: 10.1002/pon.889] [Medline: 15651053]
13. Basch E, Deal AM, Dueck AC, et al. Overall survival results of a trial assessing patient-reported outcomes for symptom monitoring during routine cancer treatment. JAMA 2017 Jul 11;318(2):197-198. [doi: 10.1001/jama.2017.7156] [Medline: 28586821]
14. Mooney KH, Beck SL, Friedman RH, Farzanfar R. Telephone-linked care for cancer symptom monitoring: a pilot study. Cancer Pract 2002;10(3):147-154. [doi: 10.1046/j.1523-5394.2002.103006.x] [Medline: 11972569]
15. White MP, Yeo NL, Vassiljev P, et al. A prescription for “nature” – the potential of using virtual nature in therapeutics. Neuropsychiatr Dis Treat 2018;14:3001-3013. [doi: 10.2147/NDT.S179038] [Medline: 30510422]
16. Browning M, Saeidi-Rizi F, McAnirlin O, Yoon H, Pei Y. The role of methodological choices in the effects of experimental exposure to simulated natural landscapes on human health and cognitive performance: a systematic review. Environ Behav 2021 Aug;53(7):687-731. [doi: 10.1177/0013916520906481]
17. Upadhyay B, Madathil KC, Bertrand J, et al. Virtual reality enabled asynchronous learning modules for fiber optic preform manufacturing education. Appl Opt 2023 Nov 1;62(31):H1-H8. [doi: 10.1364/AO.491404] [Medline: 38037917]
18. Ausburn LJ, Ausburn FB. Desktop virtual reality: a powerful new technology for teaching and research in industrial teacher education. J STEM Teach Educ 2004;41(4) [FREE Full text]
19. Niki K, Okamoto Y, Maeda I, et al. A novel palliative care approach using virtual reality for improving various symptoms of terminal cancer patients: a preliminary prospective, multicenter study. J Palliat Med 2019 Jun;22(6):702-707. [doi: 10.1089/jpm.2018.0527] [Medline: 30676847]
20. Tsuda K, Sudo K, Goto G, et al. A feasibility study of virtual reality exercise in elderly patients with hematologic malignancies receiving chemotherapy. Intern Med 2016;55(4):347-352. [doi: 10.2169/internalmedicine.55.5275] [Medline: 26875958]
21. House G, Burdea G, Grampurohit N, et al. A feasibility study to determine the benefits of upper extremity virtual rehabilitation therapy for coping with chronic pain post-cancer surgery. Br J Pain 2016 Nov;10(4):186-197. [doi: 10.1177/2049463716664370] [Medline: 27867508]
22. Bani Mohammad E, Ahmad M. Virtual reality as a distraction technique for pain and anxiety among patients with breast cancer: a randomized control trial. Palliat Support Care 2019 Feb;17(1):29-34. [doi: 10.1017/S1478951518000639] [Medline: 30198451]
23. Glennon C, McElroy SF, Connelly LM, et al. Use of virtual reality to distract from pain and anxiety. Oncol Nurs Forum 2018 Jul 2;45(4):545-552. [doi: 10.1188/18.ONF.545-552] [Medline: 29947355]
24. Coelho A, Parola V, Sandgren A, Fernandes O, Kolcaba K, Apóstolo J. The effects of guided imagery on comfort in palliative care. J Hosp Palliat Nurs 2018 Aug;20(4):392-399. [doi: 10.1097/NJH.0000000000000460] [Medline: 30063633]
25. Mahdizadeh MJ, Tirgari B, Abadi O, Bahaadinbeigy K. Guided imagery: reducing anxiety, depression, and selected side effects associated with chemotherapy. Clin J Oncol Nurs 2019 Oct 1;23(5):E87-E92. [doi: 10.1188/19.CJON.E87-E92] [Medline: 31538976]
26. Tusek DL, Church JM, Strong SA, Grass JA, Fazio VW. Guided imagery: a significant advance in the care of patients undergoing elective colorectal surgery. Dis Colon Rectum 1997 Feb;40(2):172-178. [doi: 10.1007/BF02054983]
27. Forward JB, Greuter NE, Crisall SJ, Lester HF. Effect of structured touch and guided imagery for pain and anxiety in elective joint replacement patients—a randomized controlled trial: M-TIJRP. Perm J 2015 Dec;19(4):18-28. [doi: 10.7812/TPP/14-236]
28. Kwekkeboom KL, Wanta B, Bumpus M. Individual difference variables and the effects of progressive muscle relaxation and analgesic imagery interventions on cancer pain. J Pain Symptom Manage 2008 Dec;36(6):604-615. [doi: 10.1016/j.jpainsymman.2007.12.011] [Medline: 18504089]

XSL•FO
RenderX

29. Hartshorn G, Browning M, Chalil Madathil K, et al. Efficacy of virtual reality assisted guided imagery (VRAGI) in a home setting for pain management in patients with advanced cancer: protocol for a randomised controlled trial. BMJ Open 2022 Dec 5;12(12):e064363. [doi: 10.1136/bmjopen-2022-064363] [Medline: 36576188]
30. Eldridge SM, Chan CL, Campbell MJ, et al. CONSORT 2010 statement: extension to randomised pilot and feasibility trials. BMJ 2016 Oct 24;355:i5239. [doi: 10.1136/bmj.i5239] [Medline: 27777223]
31. Bruera E, Kuehn N, Miller MJ, Selmser P, Macmillan K. The Edmonton Symptom Assessment System (ESAS): a simple method for the assessment of palliative care patients. J Palliat Care 1991;7(2):6-9. [doi: 10.1177/082585979100700202] [Medline: 1714502]
32. Bolouki A, McAnirlin O, H. E. M. Browning M, Maynard A. Optimizing virtual nature for psychological and physiological well-being: a systematic review of the moderating effects of duration, nature type, sample characteristics, and immersiveness and potential risks of bias. Int J Hum-Comput Interact 2025 Jan 2;41(1):775-791. [doi: 10.1080/10447318.2024.2338327]
33. McAnirlin O, Browning M, Fasolino T, et al. Co-creating and delivering personalized, nature-based VR experiences: proof-of-concept study with four U.S. adults living with severe COPD. Wellb Space Soc 2024 Dec;7:100212. [doi: 10.1016/j.wss.2024.100212]
34. Hui D, Bruera E. The Edmonton Symptom Assessment System 25 years later: past, present, and future developments. J Pain Symptom Manage 2017 Mar;53(3):630-643. [doi: 10.1016/j.jpainsymman.2016.10.370]
35. Hui D, Shamieh O, Paiva CE, et al. Minimal clinically important differences in the Edmonton Symptom Assessment Scale in cancer patients: a prospective, multicenter study. Cancer 2015 Sep;121(17):3027-3035. [doi: 10.1002/cncr.29437]
36. Mills EJ, Seely D, Rachlis B, et al. Barriers to participation in clinical trials of cancer: a meta-analysis and systematic review of patient-reported factors. Lancet Oncol 2006 Feb;7(2):141-148. [doi: 10.1016/S1470-2045(06)70576-9] [Medline: 16455478]
37. Scharff DP, Mathews KJ, Jackson P, Hoffsuemmer J, Martin E, Edwards D. More than Tuskegee: understanding mistrust about research participation. J Health Care Poor Underserved 2010 Aug;21(3):879-897. [doi: 10.1353/hpu.0.0323] [Medline: 20693733]
38. Fisher JA, Kalbaugh CA. Challenging assumptions about minority participation in US clinical research. Am J Public Health 2011 Dec;101(12):2217-2222. [doi: 10.2105/AJPH.2011.300279] [Medline: 22021285]
39. Langford AT, Resnicow K, Dimond EP, et al. Racial/ethnic differences in clinical trial enrollment, refusal rates, ineligibility, and reasons for decline among patients at sites in the National Cancer Institute's Community Cancer Centers Program. Cancer 2014 Mar 15;120(6):877-884. [doi: 10.1002/cncr.28483] [Medline: 24327389]
40. ChatGPT [large language model]. OpenAI. URL: https://chat.openai.com/chat [accessed 2025-06-19]

Abbreviations

AGI: with guided imagery
CONSORT: Consolidated Standards of Reporting Trials
EHR: electronic health record
ESAS: Edmonton Symptom Assessment Scale
HIPAA: Health Insurance Portability and Accountability Act
HMD: head-mounted display
LOESS: locally estimated scatterplot smoothing
NG: no guided imagery
PRO: patient-reported outcome
REDCap: Research Electronic Data Capture
VR: virtual reality
VR-NG: virtual reality with no guided imagery
VRAGI: virtual reality–assisted guided imagery

Edited by A Mavragani; submitted 05.03.25; peer-reviewed by E Maslowski, F Weinstein; revised version received 16.05.25; accepted 20.05.25; published 29.07.25.

Please cite as:
Browning M, McAnirlin O, Li F, Bertrand J, Davis D, Madathil K, Mau GF, Henry Gomez T
Virtual Reality Interventions of Daily Versus Weekly Data Collection in Patient-Reported Outcomes Among Adults With Cancer: Pilot Survey Study
JMIR Form Res 2025;9:e73506
URL: https://formative.jmir.org/2025/1/e73506
doi:10.2196/73506

XSL•FO
RenderX

© Matthew Browning, Olivia McAnirlin, Fu Li, Jeffrey Bertrand, Denise Davis, Kapil Madathil, Fredric Mau, Teny Henry Gomez. Originally published in JMIR Formative Research (https://formative.jmir.org), 29.7.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

K-Means Clustering and Classification of Breast Cancer Images Using Histogram of Oriented Gradients Features and Convolutional Neural Network Models: Diagnostic Image Analysis Study

Said Salloum, PhD

School of Computing, Skyline University College, Al Taawun, Sharjah, United Arab Emirates

Corresponding Author:
Said Salloum, PhD
School of Computing, Skyline University College, Al Taawun, Sharjah, United Arab Emirates

Abstract

Background: Breast cancer has proven to be the most common type of cancer among females around the world. However, mortality rates can be reduced if it is diagnosed at the initial stages. Interpretation made by an expert is required by conventional diagnostic tools such as biopsies and mammograms, and this interpretation can be erroneous. Artificial intelligence has increasingly been used to detect and classify breast cancer. Yet, the acquisition of impressive reliability and accuracy continues to be challenging with these automated systems.

Objective: This study aimed to develop an innovative hybrid technique for the classification of breast cancer images involving unsupervised analysis by K-means clustering, feature extraction using Histogram of Oriented Gradients (HOG), and classification of images through a convolutional neural network (CNN).

Methods: This study used a dataset of 2788 breast cancer images categorized into benign (n=1480) and malignant (n=1308) classes. The proposed hybrid method included three stages: (1) unsupervised clustering using K-means to group visually similar features; (2) feature extraction using Histogram of Oriented Gradients (HOG) to capture texture and shape patterns; and (3) classification using a CNN trained on the extracted features. The model's performance was evaluated using standard metrics such as accuracy, precision, recall, and F_1-score.

Results: The CNN achieved a classification accuracy of 98%, with precision, recall, and F_1-score values of 0.98 for both benign and malignant cases. K-means clustering revealed distinct groups corresponding to benign and malignant tumors, indicating natural separability in the image data.

Conclusions: The combination of HOG feature extraction and CNN-based classification demonstrated high performance in breast cancer detection. The model demonstrates potential utility for automated diagnosis, with possible clinical applications to assist radiologists in identifying malignant tumors more efficiently. Future research will explore additional imaging modalities and real-world clinical testing.

(JMIR Form Res 2025;9:e71974) doi:10.2196/71974

KEYWORDS

breast cancer; K-means clustering; image classification; machine learning; Histogram of Oriented Gradients; convolutional neural network

Introduction

The most common cancer among females all across the world is breast cancer. Every year, around 2.3 million women are diagnosed with breast cancer. In 2020, more than 685,000 women died because of this disease [1]. It can be treated effectively if it is diagnosed in the early stages. Mortality rates can therefore be reduced through screening tests and early detection [2]. Conventional modalities used for the diagnosis of breast cancer, including mammograms [3], magnetic resonance imaging, and biopsies [4], call for manual interpretation of images by expert radiologists—this task can be tiring and erroneous. Even though contemporary medical imaging modalities are quite efficient, different radiologists may hold different opinions about them [5]. This diagnostic variability can be addressed by using machine learning (ML) and artificial intelligence (AI), as they have been attracting the attention of many clinicians as an efficient means of detecting breast cancer automatically.

The modalities of AI-based image classification have proven to be capable of facilitating radiologists in making diagnoses, as they can extensively analyze medical images and identify cancerous regions. This not only decreases the workload of professionals but also allows diagnosis with better accuracy [6]. In this context, the popularity of convolutional neural networks

XSL•FO
RenderX

(CNNs) has increased in recent years as a deep learning modality for the classification of images. The assessment of histopathological slides, ultrasound scans, mammograms, and other breast cancer images is made using CNNs [7,8]. It has been established that CNNs can learn characteristics from images at different levels, thereby taking into consideration high-level and low-level representations, which have proven to be crucial for distinguishing between malignant and benign lesions [9].

Even though CNNs have demonstrated high efficiency in classifying images, the implementation of CNN-based modalities for the diagnosis of breast cancer is challenged by numerous factors. First is the requirement of huge datasets to acquire impressive accuracy. Acquisition of these huge datasets becomes a hard task in health care settings because of the complexity of medical imaging and privacy concerns [10]. Moreover, numerous reports are based only on CNN-based image classification and do not take unsupervised learning methods such as K-means clustering under consideration, though they can offer a detailed assessment of the inherent data structure. The efficiency of classification models can be improved through an understanding of the grouping of breast cancer images into clusters. In this way, extensive visual clues can be provided to radiologists relating to the nature of lesions [11].

These challenges have been addressed by our study that uses a combinatory approach involving supervised as well as unsupervised learning techniques for classifying breast cancer cases. In particular, the natural categorization of images of breast cancer has been evaluated through the application of K-means clustering. Next, feature extraction was carried out through the Histogram of Oriented Gradients (HOG), and finally, a CNN model was used for the classification. Grouping of data points based on similarities in features using the K-means clustering offers added evaluation which proves to be informative for the process of classification [12]. Likewise, an extensively used feature extraction method, HOG, evolves the local structure of objects through the computation of edge directions or gradients that play a vital role in the differentiation of malignant and benign tissues [13].

The efficiency of CNNs for breast cancer detection through the evaluation of medical images has been established. CNN architectures have been used by Hameed et al [14] who were successful in classifying breast cancer images accurately by applying deep learning methodology to histopathology images of breast cancer. Another example from the literature is the classification of skin cancer images using a deep learning model by Esteva et al [15], thereby validating the efficiency of a dermatologist—it validates the capability of CNNs to deal with complicated datasets of medical images. Yet, a small number of researchers have used the combination of CNNs, feature extraction, and unsupervised clustering for the classification of breast cancer. This study is aimed at integrating K-means clustering with the classification procedure so that an unsupervised approach of analysis could support the procedure and ensure improved interpretation.

This paper makes several key contributions to the field of AI in medical image analysis. First, we introduce an innovative approach that combines K-means clustering with HOG feature extraction and CNNs for breast cancer image classification. This approach allows us to gain both unsupervised and supervised insights into the data, which enhances the interpretability of the results. In addition, we demonstrate that K-means clustering can reveal natural groupings in breast cancer images, providing a useful method for obtaining an initial understanding of the dataset's structure before applying more complex deep learning models. Our experiments show that this method achieves a high classification accuracy of 98%, with strong precision and recall, demonstrating the effectiveness of CNNs when combined with robust feature extraction techniques such as HOG. Finally, the proposed framework integrates unsupervised clustering, feature extraction, and deep learning, offering a comprehensive approach to breast cancer diagnosis that contributes to the development of more interpretable and accurate AI models in health care.

The rest of this paper is organized as follows: Methods covers data preprocessing, K-means clustering, HOG feature extraction, and CNN architecture. Results include clustering analysis, classification performance, and HOG feature visualization. Discussion focuses on interpreting findings and comparing them with related work; it also summarizes contributions and proposes future directions.

Methods

Overview

Different procedures performed for the preprocessing, feature extraction, K-means clustering, and classification for analysis of breast cancer images are described in this section. A set of breast cancer images grouped into malignant and benign categories was used in this study. The combinatory approach used in this study included preprocessing of data, K-means clustering, followed by feature extraction through HOG and finally classification of images with CNNs. All of these procedures are explained in the ensuing sections.

Data Collection

The data were collected for this study from the Breast Cancer Diagnosis dataset on Kaggle (Google LLC), which is accessible to the public and is supplied by Faysal Miah [16]. Labeled images of breast cancer have been grouped into malignant and benign categories and are presented in this data collection. Histopathological findings obtained from the tissue biopsies are presented in these images which are extensively used by researchers for AI-based modalities, especially those for the classification of breast cancer. The collected data were properly labeled and curated to facilitate binary classification and AI-based breast cancer diagnosis. However, it is worth mentioning that the dataset may not reflect the diversity of clinical scenarios in the real medical environment although it is helpful for the algorithm designs and evaluation. It is limited by the paucity of diversity in scanning protocols, staining styles, scanner modalities, and patients' demographics. This is a clinically common variability that can drastically impact the performance and generalizability of ML models. By leveraging this dataset, the model's exposure to these broader diagnostic challenges is thus restricted. This study is expected to be a

XSL•FO
RenderX

control for an experimental proof rather than have a clinical use. In the future, we will validate and further optimize the proposed model on multi-institutional larger datasets with different clinical situations to examine its robustness and generalization to clinical application.

The dataset contains a total of 2788 breast cancer images. These images are divided into 2 distinct categories (Figure 1): benign (1480 images) and malignant (1308 images).

The 2 basic diagnostic groups for breast cancer include malignant and benign images. So every image in the collected data was labeled as either malignant or benign. In general, benign tumors demonstrate lower aggression as they are noncancerous. Conversely, malignant tumors need to be attended to instantly as they are cancerous.

Figure 1. Sample images from the dataset showing representative benign and malignant breast cancer histopathology slides.

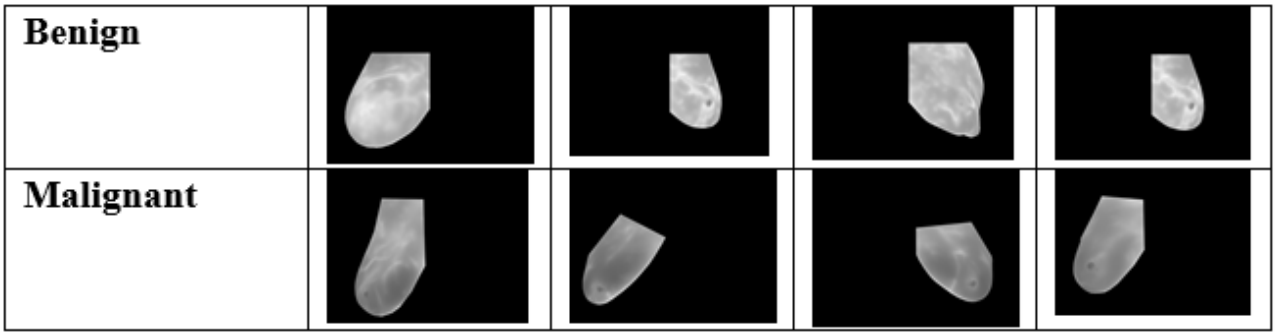

Image Characteristics

All images collected for the study were in RGB format demonstrating different resolutions. However, images were subjected to resizing at 128x128 pixels for model standardization. Even though the images were colored (blue, green, and red channels), their conversion to grayscale images was performed during preprocessing. This was done to lower memory usage and computational complexity while the important information relating to the structure and patterns of the tumor remains unaffected.

Data Preprocessing

An important step in the classification of breast cancer images is preprocessing. Preprocessing methods used during this study included resizing images to acquire uniform resolution, conversion to grayscale images; usage of Gaussian filtering for noise removal, and normalizing data to obtain pixel values in the range of 0 - 1. Such a preprocessing allowed the acquisition of consistent data and greater efficiency for feature extraction. Preprocessing and model development procedures were carried out with the help of Python 3.9 (Python Software Foundation [PSF]). The main libraries used for this purpose included OpenCV, scikit-learn, and TensorFlow. Image handling was done with OpenCV, for K-means clustering as well as feature extraction. Finally, the primary deep learning framework used during this study for the development and training of the CNN model was TensorFlow with Keras.

- Resizing: the collected breast cancer images were subjected to resizing to obtain images of 128x128 pixels so that all input dimensions are standardized for feature extraction as well as the development of a classification model.
- Normalization: normalization of values of pixel intensity of images was carried out by dividing the values by 255 so that they get normalized between 0 - and 1. During the training process, convergence becomes speedy due to this normalization.
- Data augmentation: the researchers also carried out data augmentation methods for better generalization of the CNN model and to minimize the risk of overfitting. To produce augmented images, random transformations were applied such as zooming, horizontal flips, shifts in height and width, and rotations. These methods resulted in the artificial expansion of the training data size and enabled the model to acquire detailed features from the image.

K-Means Clustering

To study the natural structure of the breast cancer image dataset, the researchers conducted K-means clustering before the application of deep learning. K-means clustering refers to an unsupervised ML algorithm that is used to categorize any dataset into definite clusters based on image similarity.

Feature Extraction for Clustering

Flattening of every image into a 1-dimensional array (128x128=16,384 values per image) was done to allow the extraction of features from the images for clustering. Also, every image was treated as a feature vector. The K-means algorithm was provided with the pixel intensity values of the flattened images. In this way, the clustering algorithm becomes capable of grouping images based on their pixel intensity distributions.

Dimensionality Reduction Using Principal Component Analysis

Taking the increased dimensionality of flattened images (16,384 dimensions) into consideration, the researchers used principal component analysis (PCA) in an attempt to decrease the dimensionality before clustering. PCA has proven to be a popular method for decreasing the number of dimensions and sustaining the data variance at the same time. The clustering results could be visualized in a 2-dimensional space as the dataset has been reduced to the first 2 principal components. The computational costs were also decreased due to the application of PCA as it increased the K-means algorithm efficiency.

XSL•FO
RenderX

K-Means Clustering Application

Following the reduction of features through the application of PCA, the images were clustered into 2 groups (K=2) through the application of the K-means algorithm. These groups corresponded to malignant and benign images. The clusters were plotted in a 2-dimensional plane by making use of the first 2 principal components to visualize the outcomes of the clustering. Such a visualization enabled the researchers to study the natural grouping of images based on pixel intensity features. Results obtained following the K-means clustering delivered a preliminary unsupervised analysis of the structure of the dataset.

HOG Feature Extraction

The researchers used the HOG for feature extraction. HOG has proven to be a popular technique for acquiring gradient and edge structures in images. This technique illustrates the gradient distribution (directional variations in pixel intensity) in specific areas of an image. For that reason, it is an efficient method for the identification of contours and shapes in clinical images.

HOG Feature Extraction Process

Each image was divided into small, connected regions termed as cells to compute the HOG descriptor for every image. This was followed by the computation of the histogram of edge orientations and gradient directions for every cell. The following steps were included in this task:

- Computation of gradient: the difference in intensities of pixels was used to compute the gradient of each pixel. Moreover, the computation of gradients was carried out in vertical (y) as well as horizontal (x) directions.
- Orientation binning: HOG feature extraction also involved the division of gradients into orientation bins (such as 8 bins for 8 directions). The gradient's magnitude was considered to be the weight for the relevant orientation bin.
- Block normalization: normalization of histograms was carried out over overlapping blocks of cells so that the HOG features were not affected by the variations in contrast or illumination. This made the features more robust and stable.

After resizing the images to acquire images with 128x128 pixels, extraction of HOG features was carried out which led to the acquisition of feature vectors which enabled the capturing of local shape information.

Visualizing HOG Features

As mentioned in the Results section, a visual picture of the main components within malignant and benign images was offered by the HOG features. The features extracted in this way were used as input for the classification task performed by CNN.

CNN Architecture

After carrying out the extraction of features, the next step was training the CNN using HOG features for binary classification. It has been established that CNNs are effective tools for image classification owing to their potential to acquire hierarchical patterns from pixels of raw data.

CNN Model Design

The CNN model used during this study consisted of layers mentioned below.

1. Input layer: the grayscale image with size 128x128x1 serves to be the input for the CNN.
2. Convolutional layers: three convolutional layers are present in the CNN. Their filter sizes are 64, 128, and 256. Each of these layers uses a 3x3 kernel for learning feature maps from the image serving as input. Nonlinearity is then introduced by ReLU activation functions.
3. Pooling layers: to decrease the spatial dimensions of the feature maps, max pooling layers (2x2) are introduced following every convolutional layer. Important information is retained during this process.
4. Completely connected layers: a pair of completely connected layers with 512 and 256 units is used following the flattening of feature maps. The learned features are combined by these layers, which also execute the final classification.
5. Output layer: a single sigmoid unit is present in the output layer. It is responsible for predicting the probability of the image being either malignant or benign.

Training the CNN

The Adam optimizer was used for the training of the CNN model. For this purpose, binary cross-entropy loss was used and the learning rate was 0.001. The model was initially trained for 20 epochs, and early stopping was executed based on validation loss. Overfitting was avoided through the usage of dropout layers having a dropout rate of 0.5. During training, units were dropped randomly, thereby promoting generalization.

Model Evaluation Metrics

The efficiency of the model can be assessed through the below-mentioned dimensions.

- Accuracy: it referred to the percentage of malignant and benign images that were classified accurately.
- Precision: it referred to the true positive results out of the total number of positive results, that is, malignant classification.
- Recall: it referred to the percentage of true positive results obtained from the model out of the results which were positive.
- F_1-score: it referred to the harmonic mean of recall and precision, thereby offering a balanced dimension for assessment of the efficiency of the model.
- Confusion matrix: the confusion matrix was used to visualize the number of true positives, true negatives, false positives, and false negatives predicted by the model.

Ethical Considerations

This study did not involve secondary data analysis. Instead, it was conducted using a publicly available dataset titled Breast Cancer Diagnosis, provided by Faysal Miah on Kaggle [16]. The dataset is openly accessible, anonymized, and intended for research purposes.

XSL•FO
RenderX

Since no human participants were directly involved and no identifiable or private data were collected or processed by the authors, formal ethical approval from an institutional review board was not required. The study complies with World Medical Association Declaration of Helsinki ethical standards for research involving publicly available, deidentified datasets [17].

Results

Overview

This study involved the usage of K-means clustering, feature extraction through HOG, and classification of breast cancer images through CNNs. Results obtained through this study are presented in this section. The efficiency of the model was evaluated with the help of several parameters such as precision, accuracy, F_1-score, and recall. The training process, validation curves, and other visual assessments are also illustrated in this section.

K-Means Clustering Results

For unsupervised analysis, the application of K-means clustering to the dataset was performed before the supervised classification. This task was intended to find out if the collected images of breast cancer are capable of forming distinct groups based on pixel intensities only. To ensure better visualization, the dimensionality of the dataset was decreased through the application of PCA. Two discrete clusters were obtained through the K-means clustering of the collected dataset as shown in Figure 2. Benign images and malignant images are represented by yellow circles and purple triangles, respectively. PCA was applied to reduce dimensionality while preserving essential variance, enabling clear cluster visualization. The clustering outcome suggests that the image features differ sufficiently to support effective classification into benign and malignant categories.

Figure 2. Visualization of K-means clustering applied to the dataset using 2 principal components after principal component analysis (PCA) dimensionality reduction. Yellow dots and purple triangles represent benign and malignant clusters, respectively, indicating clear visual separability.

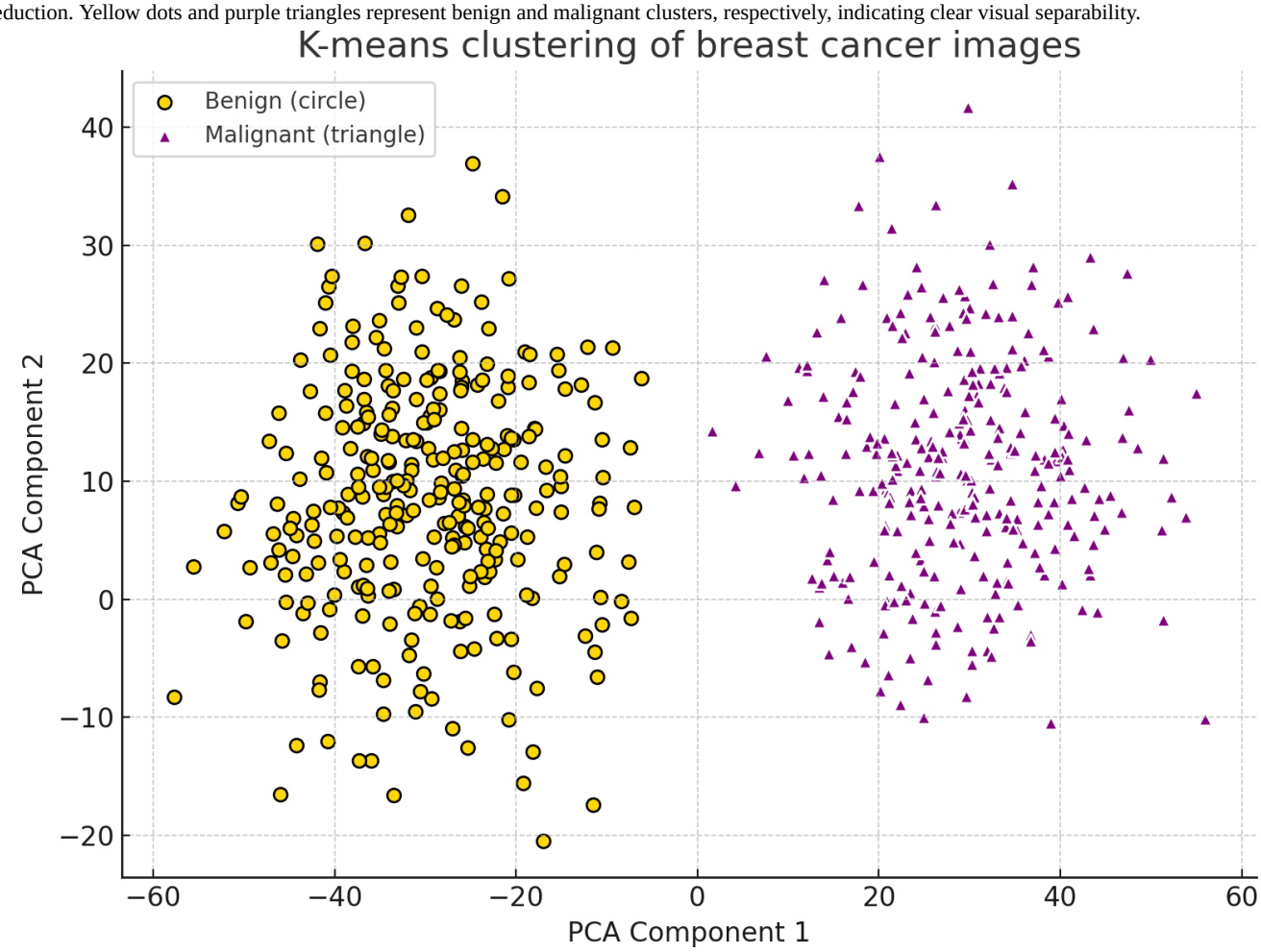

Dataset Distribution

The data were uniformly divided into 640 malignant and 640 benign images to make sure that distinct groups were represented with a balance during the tasks of training and testing. Such a balance is unavoidable to escape any bias during the training of the model. The distribution of images among the 2 groups is shown in . When the class distribution is balanced, the dimensions of the classification offer a genuine assessment of the efficiency of the model for malignant and benign images. The dataset was balanced, comprising 640 benign and 640 malignant breast cancer images, ensuring equal class representation during model training and testing to avoid classification bias.

XSL•FO
RenderX

Pixel Intensity Distribution and HOG Features

When pixel intensity distributions for malignant and benign images were assessed, it was found that discrete patterns are present between the 2 groups. In particular, pixel intensities of benign images were concentrated in lower ranges as shown in Figure 3. On the other hand, pixel intensities of malignant images were evenly distributed. This shows substantial variation in structure and texture. Such variations in pixel intensities are suggested to positively affect the potential of the model to execute the correct classification of the images. Extraction of HOG features was performed to acquire image information about the gradient and edge. The HOG features for sample malignant and benign images are presented in Figure 4. These HOG features indicate the structural distinctions between malignant and benign tumors, specifically in terms of shape and gradient orientation. The CNN model was then added with the HOG features as input, thereby improving its potential for differentiation between the 2 groups based on shape and texture. In Figure 4, benign images showed pixel intensities concentrated within a narrow grayscale range with low variability, whereas malignant images displayed a broader intensity distribution, indicating higher variability across grayscale values. The pixel intensity histograms reveal distinct grayscale characteristics between benign and malignant images. Benign images show a strong peak near zero, with most pixel values concentrated in a narrow intensity band around 120–150, indicating low variability. In contrast, malignant images also have a dominant peak near zero but display a broader and more dispersed pixel intensity range across 100–200. This indicates greater grayscale variation, reflecting structural complexity in malignant tissue. These differences support the model's ability to distinguish between the two classes based on intensity patterns.

Figure 3. Pixel intensity histograms for benign and malignant images, showing differences in grayscale intensity distributions. Benign images exhibit lower and narrower intensity ranges, while malignant images show higher variability.

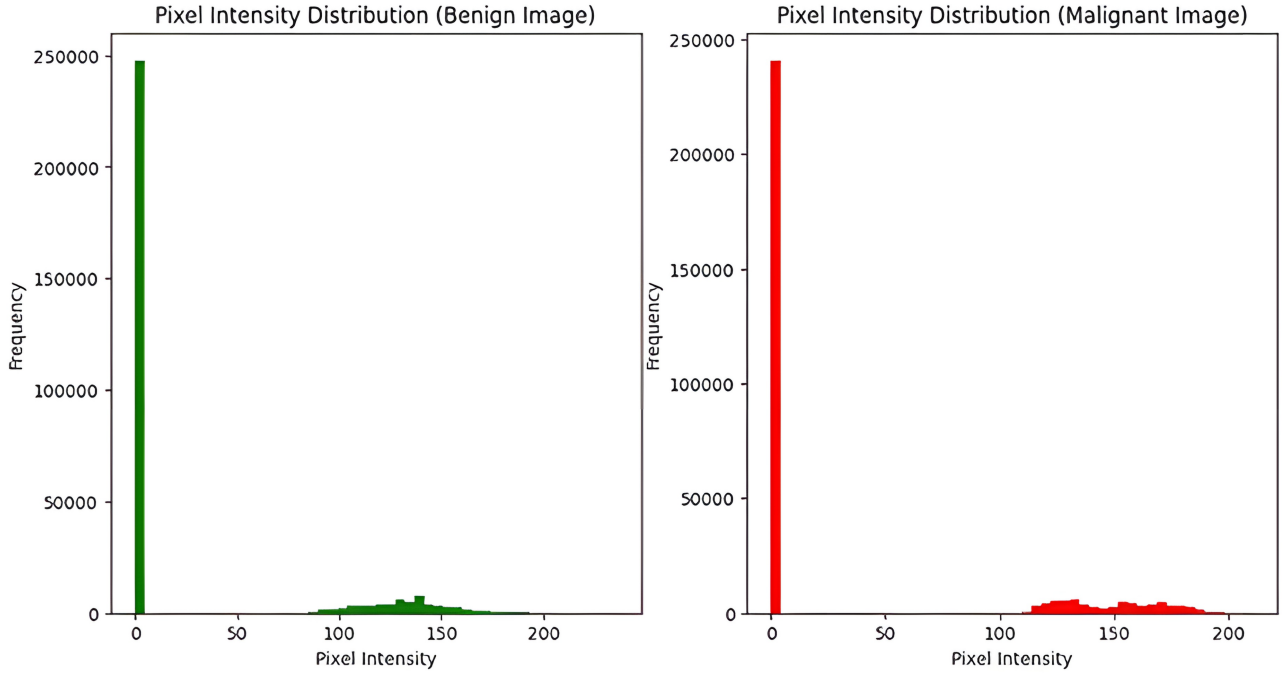

XSL•FO
RenderX

Figure 4. The Histogram of Oriented Gradients (HOG) features for representative benign and malignant images. The edge and gradient patterns highlight texture variations important for convolutional neural network–based classification.

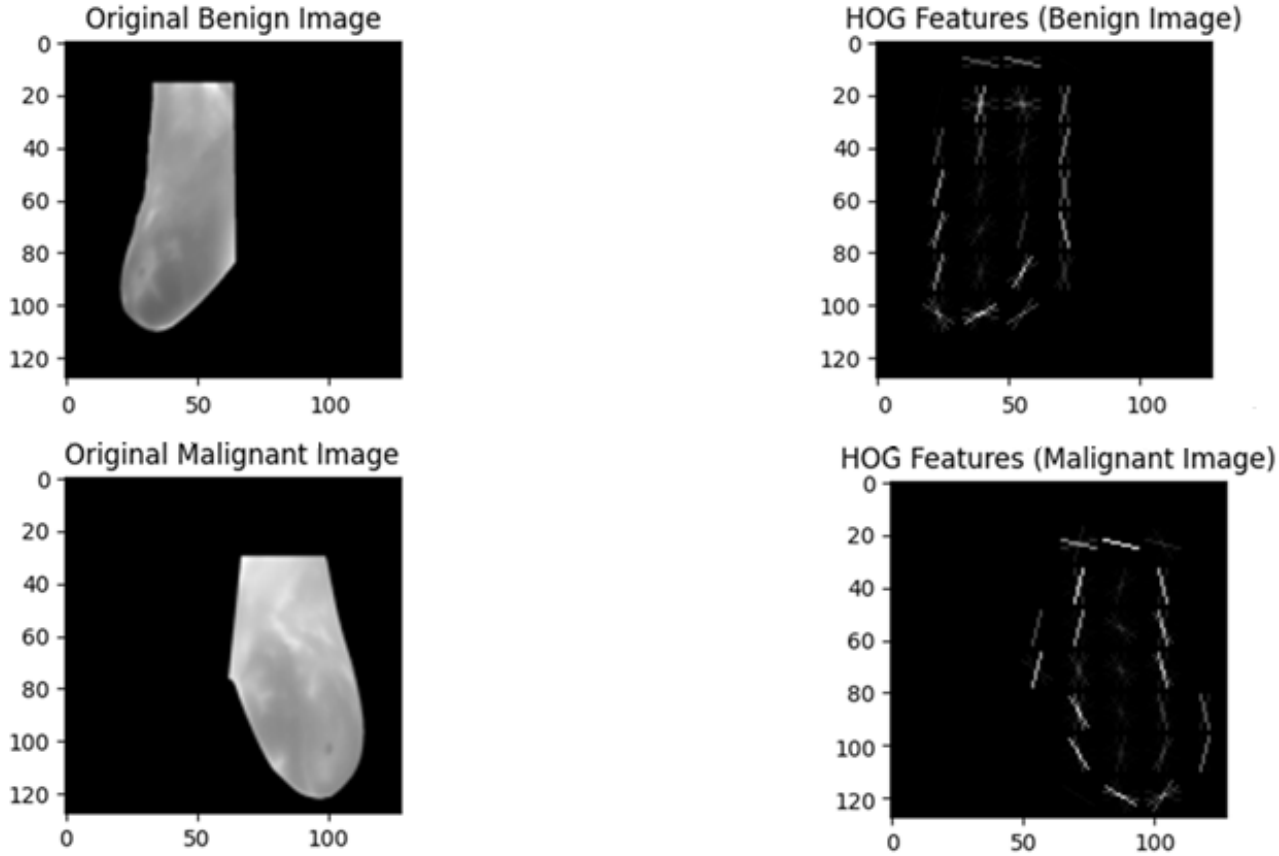

CNN Classification Results

Training of the CNN model was executed using the HOG features acquired from the images of breast cancer. The efficiency of the CNN model was analyzed with the help of dimensions such as F_1-score, recall, precision, and accuracy. The classification report is summarized in Table 1 and as per these results, the model has exhibited an accuracy of 98% with the dataset comprising 1280 images in total (640 malignant and 640 benign). Besides accuracy, the model demonstrated impressive precision, F_1-scores, and recall for the 2 groups validating the fact that it can efficiently differentiate between malignant and benign tumors.

Table . Classification performance of the convolutional neural network model using Histogram of Oriented Gradients features (N=1280).

Class	Precision	Recall	F_1-score	Support, n (%)
Benign	0.98	0.97	0.98	840 (65.6)
Malignant	0.98	0.95	0.99	840 (65.6)
Accuracy[a]			0.98	1280 (100)
Macro avg	0.98	0.98	0.98	1280 (100)
Weighted avg	0.98	0.98	0.98	1280 (100)

[a]Overall accuracy (0.98) and total support (1280, 100%) emphasize the model's performance and dataset size.

Training and Validation Performance

The efficiency of the CNN model or the training and validation process has been evaluated over 20 epochs. The validation accuracy and training accuracy are shown in Figure 5A. The validation loss and training loss are shown in Figure 5B. After 20 epochs, the training accuracy became as high as 100%. However, the validation accuracy continued to be near 98%. It shows an impressive generalization of the model toward unseen data. It is evident from the training and validation loss curves that there was a steady reduction in the model's loss during training. It can therefore be stated that the CNN model was capable of learning meaningful patterns from the given data. Results obtained during this study showed that breast cancer images can be classified efficiently through a combination of deep learning and HOG feature extraction. Impressive F_1-scores and 98% accuracy of the model indicate that the model is efficient in differentiating between malignant and benign tumors.

XSL•FO
RenderX

Findings of the K-means clustering and analysis of pixel intensity distribution also point toward the fact that malignant and benign tumors demonstrate discrete visual patterns that can be identified using supervised as well as unsupervised learning techniques. Besides these, the classification report and confusion matrix clearly show that the model is very effective in breast cancer classification with misclassification in rare cases. Usage of a balanced dataset resulted in increased efficiency of the model. Efficiency was also increased due to the usage of the data augmentation approach for improving variability in the data used for training.

Figure 5. Training and validation accuracy across 20 epochs, indicating stable learning and convergence (left). Training and validation loss curves, showing consistent decrease and minimal overfitting (right).

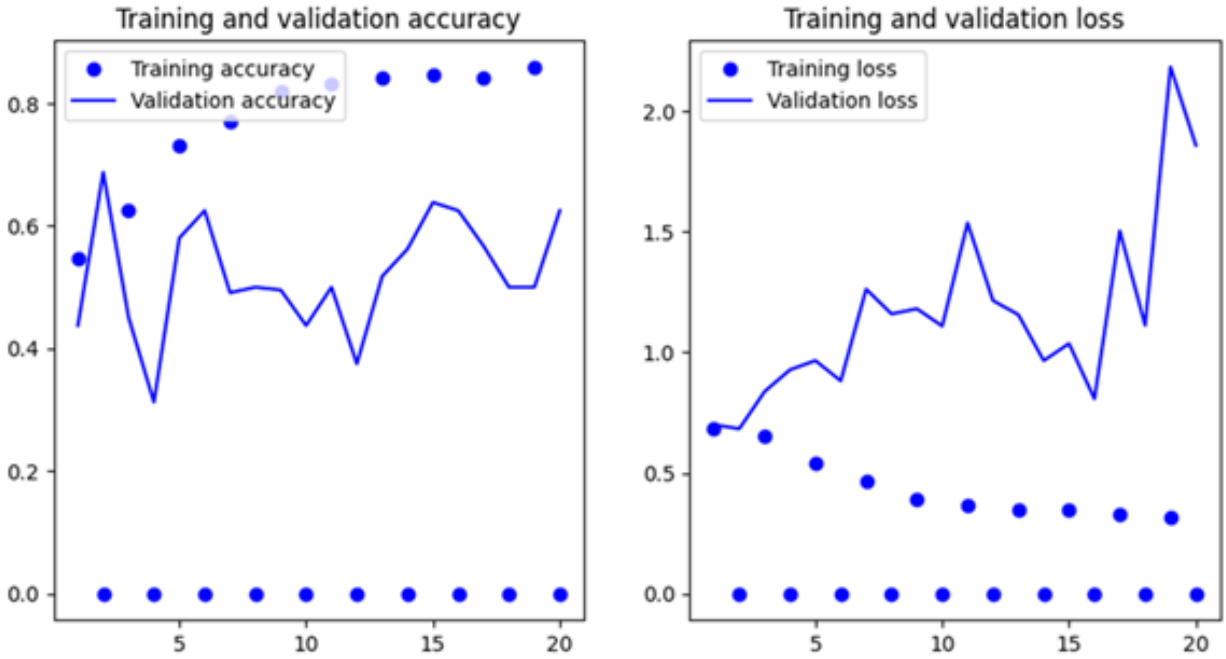

Discussion

Principal Findings

The findings of this study revealed that when K-means clustering is performed together with feature extraction through HOG and image classification through CNNs, an efficient combinatory approach is achieved for the classification of images of breast cancer into malignant and benign tumors. Having an accuracy of 98%, this modality is far better than other contemporary modalities used for the same purpose. This section discusses different implications of this study besides discussing its strengths and weaknesses. Finally, certain areas with opportunities for future research are also mentioned.

Comparison with Related Works

ML and deep learning techniques have been used by numerous researchers for the interpretation of histopathological images to diagnose breast cancer. A method based on deep learning has been used by Hameed et al [14] for the classification of histopathological images of breast cancer. The researchers used CNNs but did not conduct feature extraction through any procedure. Their method gave 95.29% accurate results. This study revealed the efficiency of CNNs and showed that classification efficiency could be increased by using any preprocessing method. This discrepancy has been considered during this study, and hence feature extraction was performed through HOG. This led to improved accuracy of the CNN model.

CNN has also been used by Bardou et al [18] for the classification of histopathological images of breast cancer with 96.15% accuracy demonstrated by the classification model. Although CNN proved to be efficient in detecting breast cancer images, feature extraction through any procedure such as HOG or unsupervised learning was not taken into consideration. This study shows that usage of HOG can lead to considerable improvement in the efficiency of the model in capturing information relating to edge and texture. This is the reason behind the improved accuracy demonstrated by the combinatory approach used in this study for breast cancer image classification.

Usage of the BreakHis dataset has been reported by some researchers such as Spanhol et al [19] for the classification of breast cancer images. With the help of a CNN-based model, these researchers have acquired 72% accuracy for the classification. Reasons behind such accuracy may include the usage of limited data. Moreover, the researchers did not use any method for feature extraction as well.

Han et al [20] have used a combinatory approach for detecting breast cancer. Their method involved CNNs and transfer learning, and this method demonstrated 93.2% accuracy. Transfer learning has proven to be a strong tool for increasing efficiency, specifically when small datasets are at hand. Yet, our study shows that the efficiency of CNNs can be increased with the help of HOG feature extraction and without using transfer learning.

XSL•FO
RenderX

Success of the K-Means Clustering and HOG Features

Usage of K-means clustering to integrate unsupervised learning proved to be an important aspect of the strategy proposed during this study. The majority of researchers have relied only on supervised learning, such as the usage of CNNs [21,22]; however, this study has attempted to integrate the unsupervised technique to decipher the inherent structure of the data at hand. It is evident from Figure 1 that malignant and benign images can be grouped into distinct clusters using K-means clustering with PCA. This implies that the images can be differentiated based on certain low-level characteristics, such as pixel intensities, even when they do not contain labeled data.

The current model proved to be quite efficient because of the usage of HOG for feature extraction. Identification of distinguishing features between malignant and benign tumors essentially requires the collection of edge information, and HOG has proven to be good at it. Usage of HOG enabled the CNN to concentrate on extensive abstract characteristics which were revealed by low-level gradient patterns. This increased the accuracy of classification. Because diagnosis is substantially affected by tumor morphology, the efficiency of HOG in collecting information about shape and texture is of great value in medical imaging [13].

Model Generalization and Overfitting Concerns

Overfitting has proven to be a considerable challenge faced frequently during the analysis of medical images, particularly in cases with small datasets. The researchers made use of numerous approaches to deal with this issue. These approaches included data augmentation and dropout layers in the architecture of CNN. Augmentation of data involved the expansion of the dataset artificially through the application of transformations such as zooms, flips, and rotations. This led to improved generalization of unseen data. Since 98% validation accuracy was achieved, it shows that the training data did not overfit because of the current model. Figure 5 shows a comparison between training and validation loss curves, which reveals that the current approach was capable of acquiring meaningful patterns from the dataset. Besides this, the model was successful in sustaining steady efficiency across varying epochs. Such stability shows the significance of the usage of the dropout layers. This approach deactivates units at random during the training so that the model does not become increasingly dependent on specific characteristics. Such approaches made the model highly generalizable, enabling it to demonstrate impressive efficiency on unseen test data. However, it should be noted that internal validation measures may indicate good generalization, but the true overfitting potential can only be estimated by external validation on a data series from another source. Small, even augmented, datasets may not completely capture real-world complexity. Accordingly, future studies applying the model on external multicenter data that have diverse imaging characteristics to validate its robustness under different clinical environments are needed.

Implications and Conclusion

Breast cancer has long been established as an important cause of cancer-associated mortalities all across the globe. However, patient outcomes can be significantly improved if the disease is diagnosed earlier. Conventional approaches to the diagnosis of breast cancer are substantially dependent on the interpretation made by an expert professional, though this interpretation can be erroneous and time-consuming. Consequently, the identification and classification of breast cancer through analysis of medical images has now increasingly been done with AI and ML. Accordingly, a hybrid approach has been proposed during this study, which is based on a combination of HOG and K-means clustering to execute feature extraction together with CNN to classify images. This combinatory approach is used in an attempt to improve diagnostic accuracy and to enable supervised as well as unsupervised analysis of data obtained for breast cancer images. As per the findings of this study, the combinatory approach involving HOG and K-means clustering-based feature extraction and CNN-based classification offers an efficient model for classifying images of breast cancer into malignant and benign groups. Classification accuracy of 98% has been demonstrated by the CNN model trained on HOG-extracted features. Moreover, the recall and precision values for this framework were found to be over 0.95 in most cases for malignant as well as benign groups. Explicit distinctions between malignant and benign image clusters were acquired through the usage of K-means clustering and PCA. This confirms the efficiency of the current model for obtaining the inherent structure of data. When HOG features and distributions of pixel intensity were taken from the images, they indicated distinguishing visual patterns in malignant and benign tumors. The potential of CNN to perform accurate classification of these images with impressive F_1-scores indicates that the diagnosis of breast cancer can be substantially improved through integrating feature extraction modalities with deep learning frameworks. The efficiency exhibited by the model in terms of training and validating proved its potential to generalize unseen data effectively without any overfitting. It can therefore be stated that the results of this study are quite valuable for AI-based medical imaging. The framework suggested during this study can be considered a dependable support tool for diagnosis, which can decrease the effort of radiologists as well as the chances of errors in diagnosis. The current model takes advantage of both supervised learning (CNNs) and unsupervised learning (K-means clustering) to extensively evaluate breast cancer images and achieve impressive interpretability and classification accuracy. Improvement in the efficiency of the classification model through the usage of HOG feature extraction demonstrates the significance of using gradient and edge-based characteristics for evaluating clinical images. The model has proven to be capable of identifying small distinctions in the texture and structure of tumors. This distinction plays a crucial role in differentiating malignant tumors from benign tumors. Findings of this study show that this combinatory approach can be implemented in the current diagnostic system to facilitate radiologists, especially in settings with limited resources. However, there are certain limitations to the study which must be considered before implementing the model in the system. A small dataset has been used in this study as compared to the huge datasets of the real world. It is important to note that the efficiency of this model has not been evaluated in any clinical setting having a wide variety of patients being attended and

XSL•FO
RenderX

several imaging modalities being used. The generalizability of this model requires validation using huge datasets involving varying clinical settings. Besides these, grayscale images had been used in this study, which made the task easy for the model, enabling it to focus on shape and texture. However, it could have led to the dropping of any data relating to color. The conversion of data into grayscale images indeed lowers computational complexity; colored images can be used in future. Similarly, future studies can consider the usage of other imaging techniques such as mammography and magnetic resonance imaging in an attempt to improve the efficiency of the model. Future research in the same area can consider working in different directions, such as investigating huge datasets with the same approach. The generalizability and efficiency of the model can be better evaluated with other datasets that include other kinds of cancers. The usage of transfer learning can improve this approach since this technique involves fine-tuning models that have been trained on large-scale datasets such as ImageNet. The usage of multiple imaging modalities can also be considered by future studies, as this study involved usage of only histopathological images. Future studies may use other kinds of clinical information in combination with medical images. The classification model may consider the clinical histories of patients or their genetic data. In this way, a holistic approach to the disease can be actualized with the aim of improved diagnosis of breast cancer. Interpretability and accuracy of the classification models can also be improved by considering highly efficient deep learning architectures such as Capsule Networks, EfficientNet, or ResNet as they deliver impressive results in the classification of images. Application of these architectures to breast cancer diagnosis is expected to allow for improved capturing of complicated information present in the data. A novel technique for the classification of breast cancer images has been proposed in this study which involves K-means clustering, HOG feature extraction, and CNNs. 98% accurate results were obtained through this approach. It can therefore be stated that this approach can serve as a useful diagnostic tool to assist radiologists in health care settings. Even though there are considerable limitations of this study, including the small size of the dataset and usage of grayscale images, it has formed a basis for AI-based diagnostic models for breast cancer. More advanced and efficient automated breast cancer diagnostic systems with greater accuracy and reliability can be developed in the future through the usage of datasets larger in size, several different imaging procedures, and contemporary deep learning methods.

Conflicts of Interest

None declared.

References

1. Breast cancer fact sheet. World Health Organization. 2021. URL: https://www.who.int/news-room/fact-sheets/detail/breast-cancer [accessed 2025-07-08]
2. Xiong X, Zheng LW, Ding Y, et al. Breast cancer: pathogenesis and treatments. Signal Transduct Target Ther 2025 Feb 19;10(1):49. [doi: 10.1038/s41392-024-02108-4] [Medline: 39966355]
3. Ozcan BB, Dogan BE, Xi Y, Knippa EE. Patient perception of artificial intelligence use in interpretation of screening mammograms: a survey study. Radiol Imaging Cancer 2025 May;7(3):e240290. [doi: 10.1148/rycan.240290] [Medline: 40249272]
4. Comparetto C, Borruto F. Cutting-edge imaging breakthroughs for early breast cancer detection. Cancer Screen Prev 2025 Mar 30;4(1):21-39. [doi: 10.14218/CSP.2024.00032]
5. Lehman CD, Wellman RD, Buist DSM, et al. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. JAMA Intern Med 2015 Nov;175(11):1828-1837. [doi: 10.1001/jamainternmed.2015.5231] [Medline: 26414882]
6. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017 Dec;42:60-88. [doi: 10.1016/j.media.2017.07.005] [Medline: 28778026]
7. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM 2017 May 24;60(6):84-90. [doi: 10.1145/3065386]
8. Giger ML. Machine learning in medical imaging. J Am Coll Radiol 2018 Mar;15(3 Pt B):512-520. [doi: 10.1016/j.jacr.2017.12.028] [Medline: 29398494]
9. Litjens G, Sánchez CI, Timofeeva N, et al. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Sci Rep 2016 May 23;6(1):26286. [doi: 10.1038/srep26286] [Medline: 27212078]
10. Greenspan H, van Ginneken B, Summers RM. Guest editorial deep learning in medical imaging: overview and future promise of an exciting new technique. IEEE Trans Med Imaging 2016;35(5):1153-1159. [doi: 10.1109/TMI.2016.2553401]
11. Islam MM, Poly TN. Machine learning models of breast cancer risk prediction. bioRxiv. Preprint posted online on Aug 2, 2019. [doi: 10.1101/723304]
12. Jain AK. Data clustering: 50 years beyond K-means. Pattern Recognit Lett 2010 Jun;31(8):651-666. [doi: 10.1016/j.patrec.2009.09.011]
13. Dalal N, Triggs B. Histograms of oriented gradients for human detection. Presented at: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05); Jun 20-25, 2005; San Diego, CA, USA. [doi: 10.1109/CVPR.2005.177]

XSL•FO
RenderX

14. Hameed Z, Zahia S, Garcia-Zapirain B, Javier Aguirre J, María Vanegas A. Breast cancer histopathology image classification using an ensemble of deep learning models. Sensors (Basel) 2020 Aug 5;20(16):4373. [doi: 10.3390/s20164373] [Medline: 32764398]
15. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature New Biol 2017 Feb 2;542(7639):115-118. [doi: 10.1038/nature21056] [Medline: 28117445]
16. Breast cancer diagnosis. Kaggle. 2023. URL: https://www.kaggle.com/datasets/faysalmiah1721758/breast-cancer-diagnosis [accessed 2025-07-08]
17. World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA 2013 Nov 27;310(20):2191-2194. [doi: 10.1001/jama.2013.281053] [Medline: 24141714]
18. Bardou D, Zhang K, Ahmad SM. Classification of breast cancer based on histology images using convolutional neural networks. IEEE Access 2018;6:24680-24693. [doi: 10.1109/ACCESS.2018.2831280]
19. Spanhol FA, Oliveira LS, Petitjean C, Heutte L. Breast cancer histopathological image classification using convolutional neural networks. Presented at: 2016 International Joint Conference on Neural Networks (IJCNN); Jul 24-29, 2016; Vancouver, BC, Canada. [doi: 10.1109/IJCNN.2016.7727519]
20. Han Z, Wei B, Zheng Y, Yin Y, Li K, Li S. Breast cancer multi-classification from histopathological images with structured deep learning model. Sci Rep 2017 Jun 23;7(1):4172. [doi: 10.1038/s41598-017-04075-z] [Medline: 28646155]
21. Almuqren AA. Cybersecurity threats, countermeasures and mitigation techniques on the IoT: Future research directions. J Cyber Secur Risk Audit 2025 Jan 22;1(1):1-11. [doi: 10.63180/jcsra.thestap.2025.1.1]
22. Aljumaiah O, Jiang W, Reddy Addula S, et al. Analyzing cybersecurity risks and threats in IT infrastructure based on NIST framework. J Cyber Secur Risk Audit 2025 Apr 4;2025(2):12-26. [doi: 10.63180/jcsra.thestap.2025.2.2]

Abbreviations

AI: artificial intelligence
CNN: convolutional neural network
HOG: Histogram of Oriented Gradients
ML: machine learning

Edited by J Sarvestan; submitted 31.01.25; peer-reviewed by D Patel, K Alomari, MA Almaiah; revised version received 23.06.25; accepted 24.06.25; published 29.07.25.

Please cite as:
Salloum S
K-Means Clustering and Classification of Breast Cancer Images Using Histogram of Oriented Gradients Features and Convolutional Neural Network Models: Diagnostic Image Analysis Study
JMIR Form Res 2025;9:e71974
URL: https://formative.jmir.org/2025/1/e71974
doi:10.2196/71974

© Said Salloum. Originally published in JMIR Formative Research (https://formative.jmir.org), 29.7.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Usefulness of Interventions Using a Smartphone Cognitive Behavior Therapy Application for Children With Mental Health Disorders: Prospective, Single-Arm, Uncontrolled Clinical Trial

Shinichiro Nagamitsu[1], MD; Ayumi Okada[2], MD; Ryoichi Sakuta[3], MD; Ryuta Ishii[4], MD; Kenshi Koyanagi[5], MD; Chizu Habukawa[6], MD; Takashi Katayama[7], BEng; Masaya Ito[8], PhD; Ayako Kanie[8], MD; Ryoko Otani[3], MD; Takeshi Inoue[3], MD; Tasuku Kitajima[3], MD; Naoki Matsubara[3], PhD; Chie Tanaka[2], MD; Chikako Fujii[2], MD; Yoshie Shigeyasu[2], MD; Michiko Matsuoka[9], MD; Tatsuyuki Kakuma[10], PhD; Masaru Horikoshi[8], PhD

[1]Department of Pediatrics, Faculty of Medicine, Fukuoka University, 7-45-1 Nanakuma, Jonan-ku, Fukuoka, Japan
[10]Biostatistics Center, Kurume University, Kurume, Japan
[2]Department of Pediatrics, Okayama University Graduate School of Medicine, Dentistry and Pharmaceutical Sciences, Okayama, Japan
[3]Child Development and Psychosomatic Medicine Center, Dokkyo Medical University Saitama Medical Center, Shimotsugagun, Japan
[4]Department of Pediatrics & Child Health, Kurume University, School of Medicine, Kurume, Japan
[5]Nagasaki Prefectural Center of Medicine and Welfare for Children, Isahaya, Japan
[6]Department of Pediatric Allergy, Minami Wakayama Medical Center, Tanabe, Japan
[7]L2B Inc, Shibuya, Japan
[8]National Center for Cognitive Behavior Therapy and Research, National Center of Neurology and Psychiatry, Kodaira, Japan
[9]Department of Neuropsychiatry, Kurume University School of Medicine, Kurume, Japan

Corresponding Author:
Shinichiro Nagamitsu, MD
Department of Pediatrics, Faculty of Medicine, Fukuoka University, 7-45-1 Nanakuma, Jonan-ku, Fukuoka, Japan

Abstract

Background: The prevalence of mental health disorders among children in Japan has increased rapidly, and these children often show depressive symptoms and reduced quality of life (QOL). We previously developed a smartphone-based self-monitoring app to deliver cognitive behavioral therapy (CBT), implemented it in healthy children, and reported its effectiveness for health promotion.

Objective: This study aims to examine the usefulness of the CBT app for improvement in depressive symptoms and QOL in children with mental health disorders.

Methods: The participants were 115 children with mental health disorders (eg, school refusal, orthostatic hypotension, eating disorders, developmental disorders, among others) and aged 12 - 18 years. The CBT app–based program comprised 1 week of psychoeducation followed by 1 week of self-monitoring. After reading story-like scenarios, participants created a self-monitoring sheet with 5 panels: events, thoughts, feelings, body responses, and actions. All participants received regular mental health care from physicians in addition to the app-based program. To evaluate the participants' depressive symptoms and QOL, Patient Health Questionnaire for Adolescents (PHQ-9A), Depression Self-Rating Scale for Children (DSRS-C), and Pediatric Quality of Life Inventory (PedsQL) were measured at the beginning of the intervention, and at 2 and 6 months thereafter. Questionnaire for Triage and Assessment with 30 items (QTA30), and Rosenberg Self-Esteem Scale (RSES) were also used to measure their health and self-esteem. Participants were divided into 4 groups on the basis of the PHQ-9A score (above or below the cutoff; PHQ-9A≥5 or PHQ-9A<5) and completion or noncompletion of the CBT app–based program (app [+] or app [-]). The primary outcome was improvement in the DSRS-C score, and secondary outcomes were improvement in other psychometric scales including PedsQL, QTA30, and RSE. A paired-samples *t* test was used for statistical analysis. The Medical Ethics Committee of Fukuoka University Faculty of Medicine (approval U22-05-002) approved the study design.

Results: There were 48, 18, 18, and 7 participants in the PHQ-9A≥5 app (+), PHQ-9A≥5 app (-), PHQ-9A<5 app (+), and PHQ-9A<5 app (-) groups, respectively. A total of 24 participants dropped out. No improvement in the DSRS-C score was observed in all groups. However, PedsQL scores improved significantly at 2 and 6 months in the PHQ-9A<5 app (+) group (t_{17}=6.62; P<.001 and t_{17}=6.11; P<.001, respectively). There was a significant positive correlation between the PHQ-9A scores and the number of self-monitoring sheets completed.

Conclusions: The CBT app was useful for improving PedsQL scores of children with mental health disorders. However, a higher-intensity CBT program is necessary for more severely depressed children.

XSL•FO
RenderX

Trial Registration: University Hospital Medical Information Network Clinical Trials Registry UMIN000046775; center6.umin.ac.jp/cgi-open-bin/ctr_e/ctr_view.cgi?recptno=R000053360

(JMIR Form Res 2025;9:e60943) doi:10.2196/60943

KEYWORDS

smartphone; cognitive behavioral therapy; application; adolescent; youth; teen; pediatric; mental health; psychoeducation; self-monitoring; questionnaire; depressive symptoms; effectiveness; Japan; statistical analysis; single-arm uncontrolled study; mobile phone

Introduction

The prevalence of mental health disorders among children has increased rapidly, and mental health care for children is an increasingly important part of pediatric health care delivery [1,2]. The number of children with school refusal is rapidly rising, and the "hikikomori" (severe social withdrawal) phenomenon may become more common in the postpandemic era [3]. In addition, it has been reported that COVID-19 has led to an increase in the number of children worldwide with depression, eating disorders, and sleep disorders [4-7]. Disability-adjusted life years of Japanese adolescents indicated that mental health disorders accounted for approximately 20% of the burden of disease [8]. Children with mental health disorders are more prone to experiencing depressive symptoms, which in turn can lead to a decline in their quality of life (QOL) [9-11] and potentially increase their risk of suicide. Suicide has become the leading cause of death among adolescents in many countries [12]. Therefore, it is important to establish a system for the treatment of these children's mental health disorders.

Cognitive behavioral therapy (CBT) is one of the most effective means of treating mental health disorders in children. CBT, which focuses on improving the emotional state by changing thinking and behavior patterns, has been applied to a range of mental health disorders including depression, anxiety disorders, insomnia, and eating disorders [13-15]. It has also been reported that CBT improves QOL in patients with various physical conditions and mental health disorders [16]. In a study applying CBT in children with school refusal, improvements in school attendance, depression, and anxiety were reported [17]. However, the use of CBT treatments for mental health disorders in children and adolescents is often limited, where adolescents frequently refuse face-to-face psychotherapy because of embarrassment or time constraints. In recent years, mental health research has increased on CBT treatment provided via the internet and smartphones (ICBT) as an alternative to face-to-face CBT treatment [16,18-21], and the effectiveness of ICBT is reportedly no different from that of face-to-face CBT [22]. According to Aemissegger et al [23], the baseline characteristics of patients in internet-based and face-to-face intervention trials revealed no significant differences. However, internet-based trials showed a longer duration of depression and a lower proportion of patients with previous treatment experience, suggesting that such trials may attract patients who tend to delay seeking treatment [23].

ICBT includes therapist-guided ICBT and unguided ICBT, which does not involve a therapist. Guided ICBT provides individualized support and is considered highly effective, but the time of use is restricted. Unguided ICBT, on the other hand, offers convenience but no individualized guidance and is strongly dependent on the motivation of the user. In both forms of ICBT, while their effectiveness has been recognized, there are instances where the quality is lacking or randomized controlled trials (RCTs) meeting adequate standards have not been carried out [24-26]. However, time and effort tend to be spent on children's mental illness in primary care [27], and it is hoped that ICBT will play a supplementary role to general practice.

We previously developed a smartphone-based self-monitoring application to deliver CBT (CBT app), implemented it in 217 healthy children, and reported its effectiveness for health promotion [28]. The CBT app was highly effective in terms of providing users with self-monitoring skills and reducing depressive symptoms. This study sought to examine the usefulness of the CBT app for improvement of depressive symptoms and QOL in children with mental health disorders including school refusal, eating disorders, and developmental disorders, among others. The primary outcome of this intervention is the improvement in the Depression Self-Rating Scale for Children (DSRS-C) scores after implementation of the CBT app.

Methods

Study Design

We conducted a prospective, single-arm clinical trial involving 115 children with mental health disorders aged 11 - 18 years (mean age 14.9 SD 1.6 y). The trial was registered in the University Hospital Medical Information Network Clinical Trials Registry (registration UMIN000046775). Participants were enrolled from 6 collaborating institutions, and there were 49 and 66 male and female participants, respectively. The primary diagnoses were school refusal (n=54), orthostatic dysregulation (n=15), eating disorders (n=13), developmental disorders (n=10), irritable bowel syndrome (n=5), sleep disorders (n=5), somatoform disorders (n=4), obsessive-compulsive disorder (n=3), and other mental health disorders (n=6). The diagnoses were carried out entirely by pediatricians and conducted in accordance with the Japanese version of the Pediatric Psychosomatic Medicine Guidelines and the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition*. Following the diagnosis, the primary physician introduced the study to the patient and their family, explained the use of the app as part of routine care, and obtained their consent. The app itself did not provide any information about the primary illness. Postdiagnosis interventions focused on disease education, lifestyle management guidance, dietary and

XSL•FO
RenderX

sleep counseling, as well as directive counseling. Participants who were receiving medication did not change their medications during the study period. Participants were divided into 4 groups on the basis of the PHQ-9A score (above or below the cutoff; PHQ-9A≥5 or PHQ-9A<5) and completion or noncompletion of the CBT app-based program (app [+] or app [-]. PHQ-9A score of 5 or higher (PHQ-9A≥5) is considered indicative of a tendency toward depression. The program completion was verified on the basis of server records.

Ethical Considerations

The design of this study and the procedures for obtaining informed consent were approved by the Medical Ethics Committee of Fukuoka University Faculty of Medicine (approval: U22-05-002). Informed consent was obtained from the children and their guardians, and the electronic consent forms were stored on a secure server. The research institution managed personal information by creating a correspondence table that linked participant names to research identification numbers, with the file protected by a password. Information provided by the collaborating research institutions to the primary research institution was anonymized, with participants identified only by their research identification numbers. As compensation, each participant received a US $15 QUO card per visit, provided directly to the individual. Payments were made three times: at the initial visit, 2 months later, and 6 months later.

Procedure

The study inclusion criteria were as follows: (1) aged 11 - 18 years, (2) first visit to a collaborating institution within 6 months, (3) approval for concurrent psychiatric treatment by the attending physician, and (4) access to a smartphone or Wi-Fi network. After the participants and parents signed the informed consent form, the CBT app was installed on their smartphones. While using the CBT app at home, participants visited the hospital at 1 - 2-month intervals to receive psychological counseling for their primary illnesses from their physicians. The observation period was 6 months, and psychometric scales were assessed at the beginning of the intervention and 2 and 6 months thereafter.

CBT App

A smartphone-based CBT app, named Mugimaru (L2B Corporation). The program based on the app was described in detail previously [28]. Briefly, the program comprised 1 week of psychoeducation followed by 1 week of self-monitoring. Mugimaru presents story-like scenarios to provide psychoeducation, so that adolescents can easily understand the rationale of the CBT and are motivated to continue using the app. The story featured an adolescent boy, an adolescent girl, and a cat (Mugimaru). In the story, the boy and girl have troubles in their relationships with friends and concerns about their future. Mugimaru teaches them how feelings, thoughts, and actions mutually affect each other. Similarly, they learn that their feelings are associated with their thoughts and actions. The story consisted of 10 scenarios, and participants could browse 1 - 2 scenarios each day. After reading a scenario, a new scenario could be read on the following day. The ending of the story was available 1 week after the participants had read the rest of the story. During the intervention period, participants completed several self-monitoring sheets comprising 5 panels: events, thoughts, feelings, body responses, and actions. The participants inputted their thoughts, feelings, body responses, and actions in association with daily events. In another window, the adolescents could input comments or advice for a friend who had experienced the same event. This input was used by adolescents to practice cognitive reappraisal and problem solving. Figure 1 shows screenshots of the smartphone CBT app. By repeatedly creating self-monitoring sheets, the adolescents could monitor their own experiences, develop solutions, and make necessary changes. The shortest time in which Mugimaru can be completed is 2 weeks. All of the data were stored in the main server, and the participants were informed in advance that only the principal investigator could view the data.

XSL•FO
RenderX

Figure 1. Screenshots of the smartphone cognitive behavior therapy app.

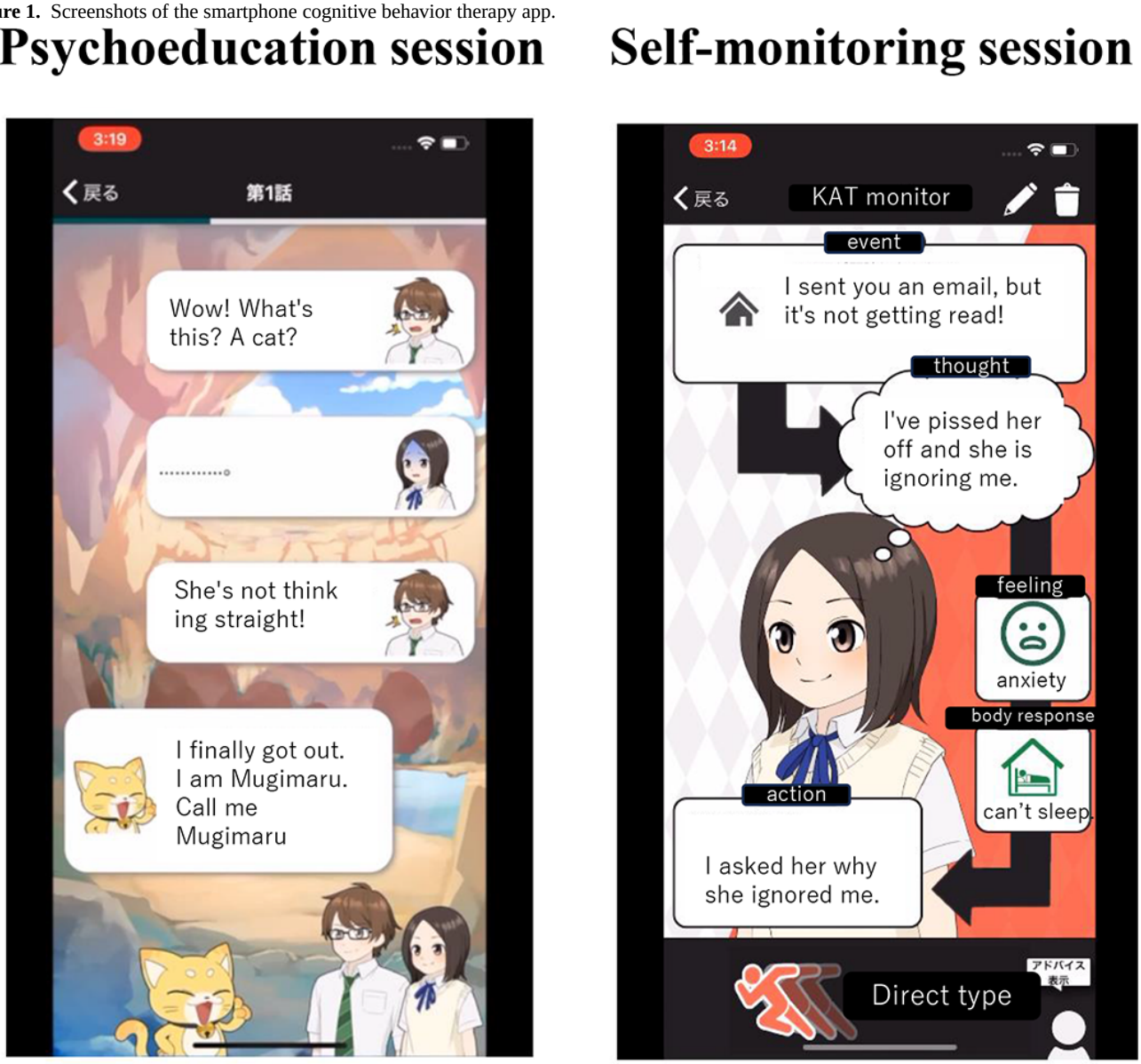

Psychometric Scales

The following 4 psychometric scales were used for outcome assessment. DSRC-S and PedsQL were used for measurement of the extent of depressive symptoms and QOL, respectively. QTA30 and RSES were used for measurement of health scores and self-esteem, respectively.

Depression Self-Rating Scale for Children (DSRS-C)

The DSRS-C is an 18-item self-report questionnaire that measures depressive symptoms [29]. Participants are asked to select 1 of 3 response options for each item: “most of the time” (score=2), “sometimes” (score=1), or “never” (score=0). The maximum score is 36, and higher scores indicate stronger depressive tendencies. The Japanese version of the DSRS-C has good reliability and validity [30]. The cutoff score for possible depression on the Japanese version is 16 points.

Pediatric Quality of Life Inventory (PedsQL)

The PedsQL is a brief measure of adolescents’ health-related quality of life [31]. The 23 items are distributed among 4 generic core scales: physical functioning, emotional functioning, social functioning, and school functioning. Items are scored as 0 (“never”), 1 (“almost never”), 2 (“sometimes”), 3 (“often”), or 4 (“almost always”). The total scale score is calculated as the mean score of all the items transformed to a 0 - 100 scale. Higher scores indicate better health-related quality of life. The Japanese version of the PedsQL has good reliability and validity [32].

QTA30

The QTA30, a psychosomatic evaluation scale for pediatric populations created by the Japanese Society of Psychosomatic Pediatrics [33], is a standardized tool for triage and assessment of pediatric patients with mental health disorders that takes Japanese cultural and social factors into account. The QTA30 comprises 4 scales: physical symptoms, depression symptoms, sense of self-efficacy, and anxiety symptoms. Higher scores on all scales correlate with poorer mental health. A total score of ≥37 points is considered to indicate poor mental health.

XSL•FO
RenderX

Rosenberg Self-Esteem Scale (RSES)

The RSES is the most widely recognized and used measure of global positive and negative attitudes toward the self [34]. It comprises 10 items, and responses are provided via a 4-point Likert scale (4, “strongly agree”; 3, “agree”; 2, “disagree”; and 1, “strongly disagree”). Negatively worded items are reverse scored, and total scores range from 10 to 40. Higher scores reflect greater levels of self-esteem. The Japanese version of the RSES has good reliability and validity [35].

Data Analysis

As stated previously, the primary outcome of this study is the improvement in DSRS-C scores after completion of the CBT app-based program. The secondary outcomes were improvements in other psychometric scales including PedsQL, QTA30, and RSE. A paired-samples *t* test was used for statistical analysis. The mean baseline PHQ-9A scores between the app completion and noncompletion were compared. In addition, we explored the relationship between participants’ suicidal ideation and their completion or noncompletion of the app. To assess this, we calculated the number of individuals who selected “More than half the days” or “Nearly every day” for the PHQ-9A item, “Thoughts that you would be better off dead or of hurting yourself in some way,” and analyzed it in relation to their completion of the app. In addition, to determine the effect of app implementation for PHQ-9A score, multiple linear regression models were tested with the change in depression score (DSRS-C) or the change in QOL score (PedsQL) as the dependent variable and the PHQ-9A score and app implementation as independent variables. We also evaluated the correlations between the number of self-monitoring sheets completed by participants and their psychometric scale scores.

Results

In total, 115 children with mental health disorders participated in this study. According to the PHQ-9A data, 79 of the children showed PHQ-9A scores above the cutoff (PHQ-9A≥5), and 36 were PHQ-9A scores below the cutoff (PHQ-9A<5; see (Figure 2). Although 91 participants completed the 6-month outcome measurements, 24 dropped out (13 participants with PHQ-9A≥5 and 11 participants with PHQ-9A<5) before the 6-month measurements. Of the 91 participants who completed the 6-month outcome measurements, 66 also completed the CBT app-based program (48 participants with PHQ-9A≥5 and 18 participants with PHQ-9A<5); of the remaining 25 participants, 18 participants were PHQ-9A≥5 and 7 participants were PHQ-9A<5. The 91 participants who completed the 6-month outcome measurements were divided into 4 groups on the basis of PHQ-9A score (above or below the cutoff; PHQ-9A≥5 or PHQ-9A<5) and completion or noncompletion of the CBT app-based program (PHQ-9A≥5 app [+], PHQ-9A≥5 app [-], PHQ-9A<5 app [+], and PHQ-9A<5 app [-] groups). The primary diagnosis was not related to whether they completed the app. Table 1 shows the preintervention scores on each psychological scale.

Figure 2. Participant flow chart. CBT: cognitive behavioral therapy; PHQ-9A: Patient Health Questionnaire for Adolescents.

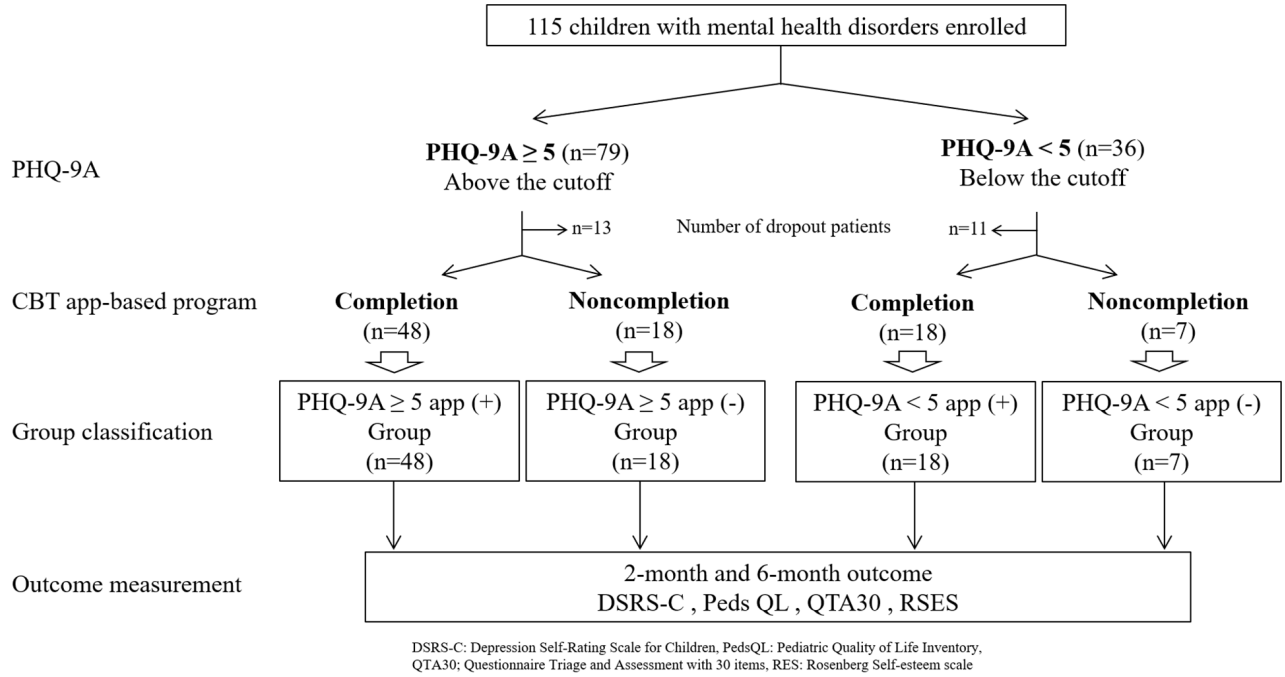

XSL•FO
RenderX

Table . Comparison of the mean scores for each psychometric scale between baseline and follow-up measurements in 4 groups, categorized by the Patient Health Questionnaire for Adolescents (PHQ-9A) cutoff value and completion or noncompletion of the app program.

Follow-up period	PHQ-9[a]A≥5						PHQ-9A<5					
	App completion (n=48), mean (SD)	*t* test (*df*)	*P* value	App non-completion (n=18), mean (SD)	*t* test (*df*)	*P* value	App completion (n=18), mean (SD)	*t* test (*df*)	*P* value	App non-completion (n=7), mean (SD)	*t* test (*df*)	*P* value
DSRS-C[b]												
Baseline	18.77 (5.33)	—	—	17.83 (5.53)	—	—	10.17 (3.24)	—	—	11.57 (6.32)	—	—
2-month	18.75 (5.51)	0.03 (47)	.98	16.22 (5.81)	1.68 (17)	.11	8.83 (4.63)	1.90 (17)	.07	10.86 (3.29)	0.47 (6)	.65
6-month	19.04 (5.06)	0.39 (47)	.70	15.72 (5.68)	1.61 (17)	.13	8.56 (5.07)	1.87 (17)	.08	9.14 (4.45)	0.98 (6)	.37
PedsQL[c]												
Baseline	64.22 (17.53)	—	—	67.81 (13.01)	—	—	84.96 (8.42)	—	—	78.57 (12.13)	—	—
2-month	67.44 (17.03)	1.85 (47)	.07	71.26 (11.57)	1.59 (17)	.13	92.57 (7.44)	6.62 (17)	<.001	83.39 (11.72)	1.25 (6)	.26
6-month	71.22 (14.55)	3.36 (47)	.002	75.00 (10.67)	2.20 (17)	.04	94.75 (5.96)	6.11 (17)	<.001	81.21 (10.70)	0.54 (6)	.61
QTA30[d]												
Baseline	54.33 (14.19)	—	—	55.61 (12.35)	—	—	27.94 (12.13)	—	—	25.43 (6.02)	—	—
2-month	52.06 (18.01)	1.19 (47)	.24	49.78 (10.82)	2.03 (17)	.06	21.78 (12.76)	2.32 (17)	.03	26.43 (8.20)	0.34 (6)	.75
6-month	47.56 (16.12)	3.08 (47)	.003	47.11 (16.85)	2.44 (17)	.03	26.56 (22.43)	0.23 (17)	.82	23.43 (11.83)	0.42 (6)	.69
RSES[e]												
Baseline	20.85 (4.31)	—	—	21.50 (4.83)	—	—	27.94 (3.90)	—	—	26.57 (5.97)	—	—
2-month	20.40 (4.46)	0.96 (47)	.34	21.22 (4.18)	0.32 (17)	.76	28.11 (4.71)	0.25 (17)	.81	25.14 (3.76)	0.81 (6)	.45
6-month	20.38 (4.57)	0.76 (47)	.45	20.78 (3.59)	0.64 (17)	.53	27.94 (5.22)	0.00 (17)	≥.99	26.57 (3.87)	0.00 (6)	≥.99

[a]PHQ-9A; Patient Health Questionnaire-9.

[b]DSRS-C: Depression Self-Rating Scale for Children.

[c]PedsQL: Pediatric Quality of Life Inventory

[d]QTA30; Questionnaire Triage and Assessment with 30 items.

[e]RSE: Rosenberg Self-esteem scale,

Outcomes at 2 and 6 Months

DSRS-C: Regardless of the completion or noncompletion of the CBT app-based program or whether the PHQ-9A value is above or below the cutoff, the DSRS-C scores did not improve significantly at 2 or 6 months (see Table 1 and Figure 3).

PedsQL: In the participants with PHQ-9A ≥5, the PedsQL scores at 6 months improved significantly regardless of the completion or noncompletion of the CBT app-based program. In the participants with PHQ-9A <5, the PedsQL scores at 2 and 6 months improved significantly only among those who completed the program.

QTA30: In the participants with PHQ-9A <5, the QTA30 score at 2 months improved significantly only among those who completed the program. In the participants with PHQ-9A ≥5, the QTA30 scores at 6 months improved significantly regardless of the completion or noncompletion of the CBT app-based program.

XSL•FO
RenderX

RSES: Regardless of the completion or noncompletion of the CBT app-based program or whether the PHQ-9A value is above or below the cutoff, the RSES scores did not improve significantly at 2 or 6 months.

Figure 3. Changes in psychometric scores in the 4 groups during the follow-up period. DSRS-C: Depression Self-Rating Scale for Children; PHQ-9A: Patient Health Questionnaire for Adolescents; PedsQL: Pediatric Quality of Life Inventory; QTA30: Questionnaire for Triage and Assessment with 30 items; RSES: Rosenberg Self-Esteem Scale.

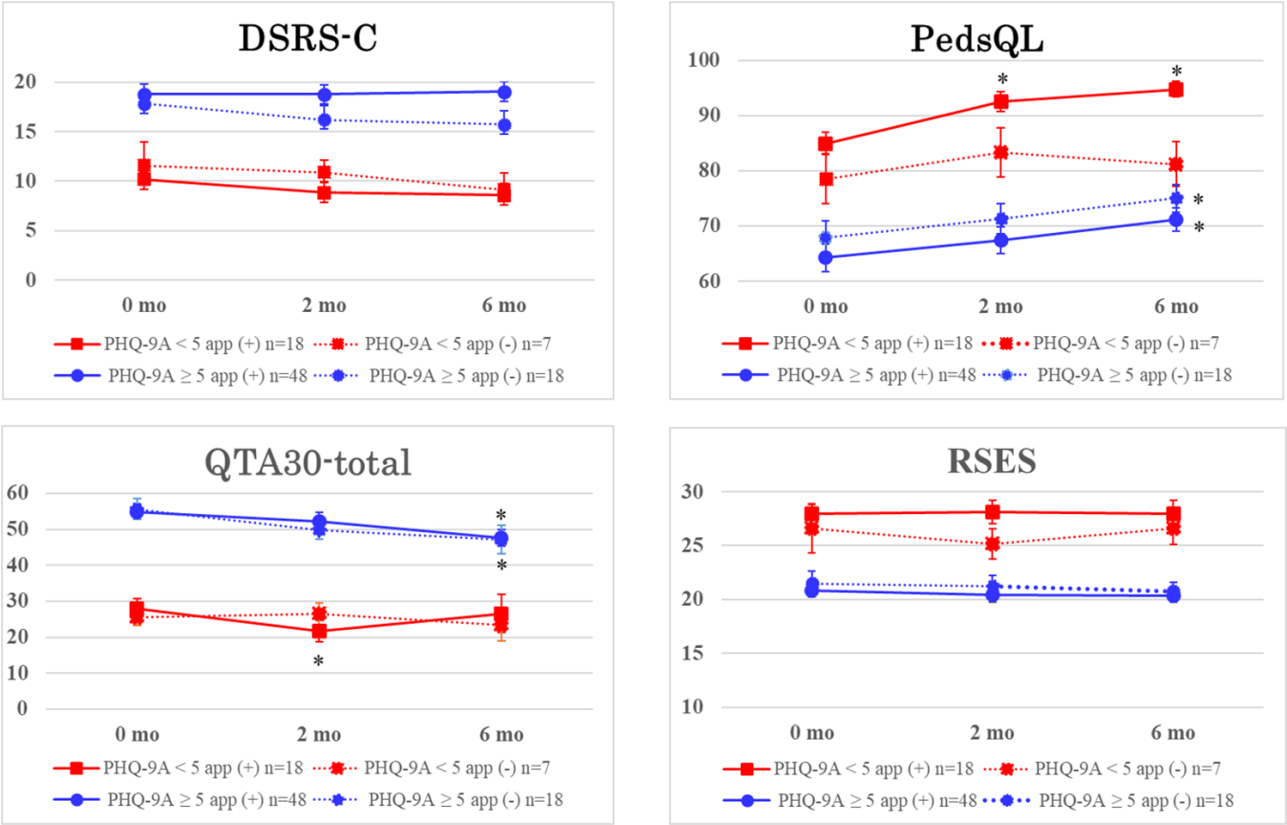

Comparison of the Mean Baseline PHQ-9A Scores Between the App Completion and Noncompletion Groups

In the group with PHQ-9A values above the cutoff (PHQ-9A≥5), the PHQ-9A score showed a tendency to be higher in the app noncompletion group (mean 13.33, SD 4.51), compared to the app completion group (mean 10.94, SD 5.02); however, the difference was not statistically significant ($t_{33,79}$=1.86; P=.07). Conversely, in the group with PHQ-9A values below the cutoff (PHQ-9A<5), the PHQ-9A score (mean 3.29, SD 0.95) was significantly higher in the app noncompletion group (mean 1.72, SD 1.56) ($t_{18,17}$=1.86; P<.01). For the PHQ-9A question, "Thoughts that you would be better off dead or of hurting yourself in some way," 17 individuals responded with "More than half the days" or "Nearly every day." Among them, 14 successfully completed the app.

Multiple Linear Regression Analysis

Multiple linear regression models were used to test if PHQ-9A score and completion or noncompletion of the CBT app-based program significantly predicted the change in DSRS-C or the change in PedsQL score. These regression models were not statistically significant (see Tables 2 and 3).

Table . Results of the regression model for changes in Depression Self-Rating Scale (DSRS) at 6 months of follow-up.

Terms	B[a]	SE	*t* test (*df*)	95% CI	*P* value
Intercept	1.210	1.659	0.729 (90)	−2.088 to 4.507	.47
PHQ-9A[b]	0.080	0.087	0.916 (90)	−0.094 to 0.253	.36
Implementation of the app	−2.125	1.173	1.182 (90)	−4.456 to 0.206	.07

[a]B: unstandardized coefficient.

[b]PHQ-9A: Patient Health Questionnaire for Adolescents.

XSL•FO
RenderX

Table . Results of the regression model for changes in Pediatric Quality of Life Inventory (PedQL) at 6 months of follow-up.

Terms	B[a]	SE	*t* test (*df*)	95% CI	*P* value
Intercept	7.887	4.365	1.807 (90)	–0.787 to 16.561	.07
PHQ-9A[b]	0.271	0.230	1.180 (90)	–0.185 to 0.727	.24
Implementation of the app	2.412	3.085	0.782 (90)	–0.8542 to 3.719	.44

[a]B:unstandardized coefficient.

[b]PHQ-9A: Patient Health Questionnaire for Adolescents.

Correlation Between the Number of Monitoring Sheets Completed and Baseline Psychometric Scale Scores

There was a significant positive correlation between the PHQ-9A score at baseline and the number of monitoring sheets completed (see Figure 4).

Figure 4. Correlation between the number of self-monitoring sheet completed by the participants and the PHQ-9A score. PHQ-9A: Patient Health Questionnaire for Adolescents.

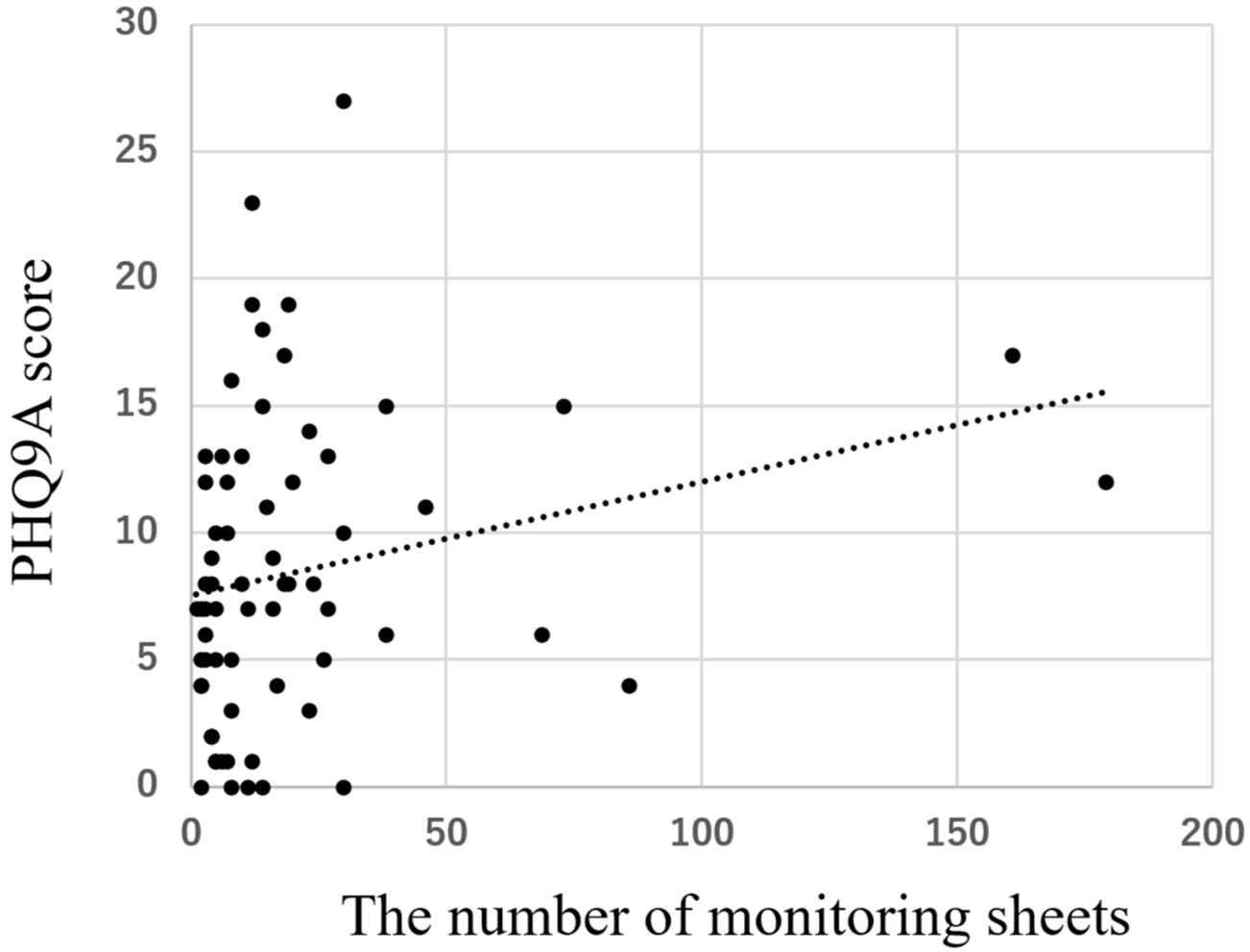

Discussion

This single-arm clinical trial demonstrated the usefulness of our CBT app-based program in terms of improving QOL (PedsQL) and health (QTA30) scores among children with mental health disorders who exhibited PHQ-9A scores below cutoff (PHQ-9A<5). However, children with PHQ-9A scores above the cutoff (PHQ-9A≥5) also showed improvements in QOL (PedsQL) and health (QTA30) scores at 6 months, even in the group that did not use the app. The app did not demonstrate effectiveness in improving depressive symptoms (DSRS-C) or self-esteem (RSE) at any time point. While a higher-intensity CBT program is necessary for more severely depressed children, this unguided app may play a supplementary role in the treatment of children with mental health disorders.

XSL•FO
RenderX

The improvement in depressive symptoms through the use of the app in children with mental health disorders was the primary outcome of this study; however, the app failed to show improvement in DSRS-C scores at any point, regardless of the initial PHQ-9A value. Although the CBT app has potential as an easily accessible, low-intensity intervention, the app might not be sufficient to reduce depressive symptoms in children with mental health disorders. Skar et al [36] analyzed the factors associated with nonresponse among children receiving CBT and identified higher baseline stress and older age as significant predictors; they also identified high intensity of the CBT intervention as key to prevent nonresponse. Nevertheless, the usefulness of app-based low-intensity CBT has also been noted. A meta-analysis of 66 randomized controlled trials of app-supported smartphone interventions for mental health problems found no significant difference in effectiveness versus active face-to-face interventions, showing the potential of apps to serve as cost-effective, easily accessible, low-intensity interventions for those who cannot receive standard high-intensity psychological treatment [37]. Adding standard CBT after low-intensity CBT was also reportedly effective in improving anxiety disorders in children [38,39]. The present CBT app consisted of 2 modules (psychoeducation and self-monitoring); additional modules including cognitive restructuring and behavioral activation are necessary to increase the effectiveness of the CBT app for children with more severe depression.

A notable finding of this study was improvements in PedsQL scores at 2- and 6-month follow-ups among children with PHQ-9A scores below the cutoff (PHQ-9A<5), achieved through the use of the app. The mechanism by which CBT improves QOL is believed to occur alongside improvements in depressive symptoms and self-esteem facilitated by CBT [40]. In this study, however, neither improvements in depressive symptoms nor self-esteem were observed, suggesting that the cartoon-style psychoeducational stories may have activated behavioral activation, a component of CBT, in the participants. On the other hand, improvements in PedsQL associated with the app were also observed at the 6-month follow-up among children with PHQ-9A scores above the cutoff (PHQ-9A≥5); however, similar improvements were noted in children who did not complete the CBT app-based program. As a medical examination was also performed and psychosocial counseling was provided in conjunction with the CBT app, it was not possible to evaluate the effects of the CBT app when analyzing the changes in the PedsQL scores in children with PHQ-9A ≥5. In general, smartphone and internet-based CBT apps are of low intensity, and their effects do not persist for a long time [41]; this likely explains why many intervention studies have 3-month outcome assessment periods [42]. In practice, it is difficult to provide treatment for mental health disorders using an app alone, that is, without also providing psychosocial counseling, so intervention studies comparing groups receiving psychosocial counseling with and without an app-based intervention are necessary to precisely determine the effects of apps.

One of the goals of this study was to learn about the differences in clinical characteristics between the CBT app-based program completion and noncompletion groups. In general, the noncompletion rate for smartphone- and web-based CBT is between 10% and 30% [43-45]. Predictors of discontinuation of CBT app use include worse physical health, lower education level, cognitive deficits, and less positive expectations of the outcome [43,46,47]. More severe pretreatment symptoms, comorbidities, lack of social support, and parental conflict are also associated with nonresponse to CBT [43,48-51]. In our study, 24 participants (20.9%) dropped out and another 25 (21.7%) discontinued the CBT app-based program. Among the latter cases, the PHQ-9A scores showed a trend toward being higher in those who did not complete the program compared with the scores among those who completed it. On the other hand, participants with strong suicidal ideation were more likely to complete the app. Furthermore, given that the participants with higher PHQ-9A scores completed more monitoring sheets, it appears that, although individuals with more severe depressive symptoms may be more likely to use a CBT app initially, they may also be more likely to give up. A stepped-care approach, which begins with lower-intensity internet-delivered CBT and then proceeds to more intensive treatments with face-to-face therapist involvement, could be effective for those who might otherwise show a poor response [38,39].

A few limitations of this study need to be addressed. First, this was a single-arm clinical trial involving children with mental health disorders who presented with varying PHQ-9A scores; a control group not using the app derived from the same population would be needed to precisely determine the effects of our CBT app. In the future, studies using randomized controlled trials (RCTs) with control groups are anticipated to provide more comprehensive and robust evidence. Second, the inclusion of participants with many disorders, such as school refusal, eating disorders, and developmental disorders, among others, may have led to variability in the results; conducting an intervention study including a population with a single disorder may have increased the reproducibility of the results. Third, 45 of the 115 participants dropped out of the study or discontinued their app use. There was a greater tendency toward dropout or discontinuation of the app among participants without depressive symptoms, and it is necessary to devise ways to encourage them to continue using the program.

Our CBT app-based program was useful for improving the QOL of children with mental health disorders with PHQ-9A<5. The program also seemed to be useful for children with PHQ-9A≥5, although the importance of regular psychosocial counseling is undeniable. A higher-intensity CBT program is necessary for more severely depressed children. By integrating this app into routine clinical practice, it serves as a supplementary tool to therapy, based on the principles of CBT, allowing individuals to observe their own mental states from a broader perspective.

XSL•FO
RenderX

Acknowledgments

This work was supported by grants from the Japan Agency for Medical Research and Development (Grants 21gk0110055h0001, 22gk0110055h0002, and 23gn0110055h0003), JSPS KAKENHI (Grant 23K02324), and Fukuoka University (Grant GR2416). We would like to thank all participants. We also thank Michael Irvine, PhD, from Edanz for editing a draft of this manuscript.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

SN, AO, RS, RI, KK, and CH were the lead researchers in the 6 collaborating institutions and participated in the design of the study. SN compiled the manuscript. RO, TI, Kitajima T, NM, CT, CF, YS, and MM recruited and examined participants and provided psychological counseling. MH, AK, MI, and Katayama T designed the CBT app. Kakuma T conducted the statistical analyses. All authors read and approved the manuscript.

Conflicts of Interest

None declared.

References

1. Madigan S, Racine N, Vaillancourt T, et al. Changes in depression and anxiety among children and adolescents from before to during the COVID-19 pandemic: a systematic review and meta-analysis. JAMA Pediatr 2023 Jun 1;177(6):567-581. [doi: 10.1001/jamapediatrics.2023.0846] [Medline: 37126337]
2. Araújo LD, Veloso CF, Souza MDC, Azevedo JD, Tarro G. The potential impact of the COVID-19 pandemic on child growth and development: a systematic review. J Pediatr (Rio J) 2021;97(4):369-377. [doi: 10.1016/j.jped.2020.08.008] [Medline: 32980318]
3. Fong TC, Yip PS. Prevalence of hikikomori and associations with suicidal ideation, suicide stigma, and help-seeking among 2,022 young adults in Hong Kong. Int J Soc Psychiatry 2023 Nov;69(7):1768-1780. [doi: 10.1177/00207640231174376] [Medline: 37191282]
4. Bignardi G, Dalmaijer ES, Anwyl-Irvine AL, et al. Longitudinal increases in childhood depression symptoms during the COVID-19 lockdown. Arch Dis Child 2021 Aug;106(8):791-797. [doi: 10.1136/archdischild-2020-320372]
5. Chen F, Zheng D, Liu J, Gong Y, Guan Z, Lou D. Depression and anxiety among adolescents during COVID-19: a cross-sectional study. Brain Behav Immun 2020 Aug;88:36-38. [doi: 10.1016/j.bbi.2020.05.061] [Medline: 32464156]
6. Ünver H, Rodopman Arman A, Erdoğdu AB, İlbasmış Ç. COVID-19 pandemic-onset anorexia nervosa: three adolescent cases. Psychiatry Clin Neurosci 2020 Dec;74(12):663-664. [doi: 10.1111/pcn.13160] [Medline: 33000521]
7. Sharma M, Aggarwal S, Madaan P, Saini L, Bhutani M. Impact of COVID-19 pandemic on sleep in children and adolescents: a systematic review and meta-analysis. Sleep Med 2021;84:259-267. [doi: 10.1016/j.sleep.2021.06.002.27]
8. Igarashi T. Research report on the biopsychosocial health issues of children, kodomo/kosodate shiensuishin chousakenkyujigyou [article in japanese]. 2018 URL: https://www.mhlw.go.jp/content/11900000/000520474.pdf
9. Shigeyasu Y, Okada A, Fujii C, et al. Quality of life and physical/psychosocial factors in children and adolescents with orthostatic intolerance. Biopsychosoc Med 2023 Jun 12;17(1):23. [doi: 10.1186/s13030-023-00278-1] [Medline: 37308984]
10. Mierau JO, Kann-Weedage D, Hoekstra PJ, et al. Assessing quality of life in psychosocial and mental health disorders in children: a comprehensive overview and appraisal of generic health related quality of life measures. BMC Pediatr 2020 Jul 3;20(1):329. [doi: 10.1186/s12887-020-02220-8] [Medline: 32620157]
11. Dey M, Landolt MA, Mohler-Kuo M. Health-related quality of life among children with mental disorders: a systematic review. Qual Life Res 2012 Dec;21(10):1797-1814. [doi: 10.1007/s11136-012-0109-7] [Medline: 22298200]
12. Saito T, Reines EH, Florea I, Dalsgaard MK. Management of depression in adolescents in Japan. J Child Adolesc Psychopharmacol 2019 Dec;29(10):753-763. [doi: 10.1089/cap.2019.0023] [Medline: 31268356]
13. Coull G, Morris PG. The clinical effectiveness of CBT-based guided self-help interventions for anxiety and depressive disorders: a systematic review. Psychol Med 2011 Nov;41(11):2239-2252. [doi: 10.1017/S0033291711000900] [Medline: 21672297]
14. Turner H, Marshall E, Wood F, Stopa L, Waller G. CBT for eating disorders: the impact of early changes in eating pathology on later changes in personality pathology, anxiety and depression. Behav Res Ther 2016 Feb;77:1-6. [doi: 10.1016/j.brat.2015.11.011] [Medline: 26690743]
15. Alimoradi Z, Jafari E, Broström A, et al. Effects of cognitive behavioral therapy for insomnia (CBT-I) on quality of life: a systematic review and meta-analysis. Sleep Med Rev 2022 Aug;64:101646. [doi: 10.1016/j.smrv.2022.101646] [Medline: 35653951]

XSL•FO
RenderX

16. O'Connor KA, Bagnell A, Rosychuk RJ, et al. A randomized controlled trial evaluating the effect of an internet-based cognitive-behavioral program on anxiety symptoms in a community-based sample of adolescents. J Anxiety Disord 2022 Dec;92:102637. [doi: 10.1016/j.janxdis.2022.102637] [Medline: 36179438]
17. Strömbeck J, Palmér R, Sundberg Lax I, Fäldt J, Karlberg M, Bergström M. Outcome of a multi-modal CBT-based treatment program for chronic school refusal. Glob Pediatr Health 2021;8:2333794X211002952. [doi: 10.1177/2333794X211002952] [Medline: 33855128]
18. Sakata M, Toyomoto R, Yoshida K, et al. Components of smartphone cognitive-behavioural therapy for subthreshold depression among 1093 university students: a factorial trial. Evid Based Ment Health 2022 Dec;25(e1):e18-e25. [doi: 10.1136/ebmental-2022-300455] [Medline: 35577537]
19. McCloud T, Jones R, Lewis G, Bell V, Tsakanikos E. Effectiveness of a mobile app intervention for anxiety and depression symptoms in university students: randomized controlled trial. JMIR Mhealth Uhealth 2020 Jul 31;8(7):e15418. [doi: 10.2196/15418] [Medline: 32735221]
20. Donker T, Cornelisz I, van Klaveren C, et al. Effectiveness of self-guided app-based virtual reality cognitive behavior therapy for acrophobia: a randomized clinical trial. JAMA Psychiatry 2019 Jul 1;76(7):682-690. [doi: 10.1001/jamapsychiatry.2019.0219] [Medline: 30892564]
21. Chan CS, Wong CYF, Yu BYM, Hui VKY, Ho FYY, Cuijpers P. Treating depression with a smartphone-delivered self-help cognitive behavioral therapy for insomnia: a parallel-group randomized controlled trial. Psychol Med 2023 Apr;53(5):1799-1813. [doi: 10.1017/S0033291721003421] [Medline: 37310329]
22. Wagner B, Horn AB, Maercker A. Internet-based versus face-to-face cognitive-behavioral intervention for depression: a randomized controlled non-inferiority trial. J Affect Disord 2014 Jan;152-154:113-121. [doi: 10.1016/j.jad.2013.06.032] [Medline: 23886401]
23. Aemissegger V, Lopez-Alcalde J, Witt CM, Barth J. Comparability of patients in trials of ehealth and face-to-face psychotherapeutic interventions for depression: meta-synthesis. J Med Internet Res 2022 Sep 14;24(9):e36978. [doi: 10.2196/36978] [Medline: 36103217]
24. Moberg C, Niles A, Beermann D. Guided self-help works: randomized waitlist controlled trial of Pacifica, a mobile app integrating cognitive behavioral therapy and mindfulness for stress, anxiety, and depression. J Med Internet Res 2019 Jun 8;21(6):e12556. [doi: 10.2196/12556] [Medline: 31199319]
25. Olthuis JV, Watt MC, Bailey K, Hayden JA, Stewart SH. Therapist-supported internet cognitive behavioural therapy for anxiety disorders in adults. Cochrane Database Syst Rev 2015 Mar 5(3):CD011565. [doi: 10.1002/14651858.CD011565] [Medline: 25742186]
26. Pelucio L, Quagliato LA, Nardi AE. Therapist-guided versus self-guided cognitive-behavioral therapy: a systematic review. Prim Care Companion CNS Disord 2024 Mar 21;26(2):23r03566. [doi: 10.4088/PCC.23r03566] [Medline: 38512281]
27. O'Brien D, Harvey K, Howse J, Reardon T, Creswell C. Barriers to managing child and adolescent mental health problems: a systematic review of primary care practitioners' perceptions. Br J Gen Pract 2016 Oct;66(651):e693-e707. [doi: 10.3399/bjgp16X687061] [Medline: 27621291]
28. Nagamitsu S, Kanie A, Sakashita K, et al. Adolescent health promotion interventions using well-care visits and a smartphone cognitive behavioral therapy app: randomized controlled trial. JMIR Mhealth Uhealth 2022 May 23;10(5):e34154. [doi: 10.2196/34154] [Medline: 35604760]
29. Poznanski EO, Grossman JA, Buchsbaum Y, Banegas M, Freeman L, Gibbons R. Preliminary studies of the reliability and validity of the children's depression rating scale. J Am Acad Child Psychiatry 1984 Mar;23(2):191-197. [doi: 10.1097/00004583-198403000-00011] [Medline: 6715741]
30. Murata T, Shimizu A, Mori Y, Oushima S. Childhood depressive state in the school situation-consideration from the Birleson's Scale [Article in Japanese]. Jpn J Psychiatr 1996;1:131-138 [FREE Full text]
31. Varni JW, Seid M, Rode CA. The PedsQL: measurement model for the pediatric quality of life inventory. Med Care 1999 Feb;37(2):126-139. [doi: 10.1097/00005650-199902000-00003] [Medline: 10024117]
32. Kobayashi K, Kamibeppu K. Measuring quality of life in Japanese children: development of the Japanese version of PedsQL. Pediatr Int 2010 Feb;52(1):80-88. [doi: 10.1111/j.1442-200X.2009.02889.x] [Medline: 19496977]
33. Ishii R, Nagamitsu S, Sakurai R, Koyanagi K, Kanbara Y, Furusho J, et al. The development and standardization of the questionnaire for triage and assessment: a self-report health-related outcome scale for children with somataization [Article in Japanese]. J Jpn Pediatr Soc ;121:1000-1008 [FREE Full text]
34. Rosenberg M. Society and the Adolescent Self-Image: Princeton, NJ: Princeton University Press; 1965. [doi: 10.1515/9781400876136]
35. Mimura C, Griffiths P. A Japanese version of the Rosenberg Self-Esteem Scale: translation and equivalence assessment. J Psychosom Res 2007 May;62(5):589-594. [doi: 10.1016/j.jpsychores.2006.11.004] [Medline: 17467414]
36. Skar AMS, Braathu N, Jensen TK, Ormhaug SM. Predictors of nonresponse and drop-out among children and adolescents receiving TF-CBT: investigation of client-, therapist-, and implementation factors. BMC Health Serv Res 2022 Sep 29;22(1):1212. [doi: 10.1186/s12913-022-08497-y] [Medline: 36175864]

XSL•FO
RenderX

37. Linardon J, Cuijpers P, Carlbring P, Messer M, Fuller-Tyszkiewicz M. The efficacy of app-supported smartphone interventions for mental health problems: a meta-analysis of randomized controlled trials. World Psychiatry 2019 Oct;18(3):325-336. [doi: 10.1002/wps.20673] [Medline: 31496095]
38. Ollendick TH, Öst LG, Farrell LJ. Innovations in the psychosocial treatment of youth with anxiety disorders: implications for a stepped care approach. Evid Based Ment Health 2018 Aug;21(3):112-115. [doi: 10.1136/eb-2018-102892] [Medline: 29976564]
39. Pettit JW, Rey Y, Bechor M, et al. Can less be more? Open trial of a stepped care approach for child and adolescent anxiety disorders. J Anxiety Disord 2017 Oct;51:7-13. [doi: 10.1016/j.janxdis.2017.08.004] [Medline: 28843575]
40. Watson HJ, Allen K, Fursland A, Byrne SM, Nathan PR. Does enhanced cognitive behaviour therapy for eating disorders improve quality of life? Eur Eat Disord Rev 2012 Sep;20(5):393-399. [doi: 10.1002/erv.2186] [Medline: 22730260]
41. Grist R, Porter J, Stallard P. Mental health mobile apps for preadolescents and adolescents: a systematic review. J Med Internet Res 2017 May 25;19(5):e176. [doi: 10.2196/jmir.7332] [Medline: 28546138]
42. Pennant ME, Loucas CE, Whittington C, et al. Computerised therapies for anxiety and depression in children and young people: a systematic review and meta-analysis. Behav Res Ther 2015 Apr;67:1-18. [doi: 10.1016/j.brat.2015.01.009] [Medline: 25727678]
43. Bennemann B, Schwartz B, Giesemann J, Lutz W. Predicting patients who will drop out of out-patient psychotherapy using machine learning algorithms. Br J Psychiatry 2022 Apr;220(4):192-201. [doi: 10.1192/bjp.2022.17] [Medline: 35177132]
44. Burke D, Lennon O, Blake C, et al. An internet-delivered cognitive behavioural therapy pain management programme for spinal cord injury pain: a randomized controlled trial. Eur J Pain 2019 Aug;23(7):1264-1282. [doi: 10.1002/ejp.1402] [Medline: 31002442]
45. Axelsson E, Hedman-Lagerlöf E. Unwanted outcomes in cognitive behavior therapy for pathological health anxiety: a systematic review and a secondary original study of two randomized controlled trials. Expert Rev Pharmacoecon Outcomes Res 2023;23(9):1001-1015. [doi: 10.1080/14737167.2023.2250915] [Medline: 37614181]
46. Schindler A, Hiller W, Witthöft M. What predicts outcome, response, and drop-out in CBT of depressive adults? a naturalistic study. Behav Cogn Psychother 2013 May;41(3):365-370. [doi: 10.1017/S1352465812001063] [Medline: 23211066]
47. van Emmerik-van Oortmerssen K, Blankers M, Vedel E, et al. Prediction of drop-out and outcome in integrated cognitive behavioral therapy for ADHD and SUD: results from a randomized clinical trial. Addict Behav 2020 Apr;103:106228. [doi: 10.1016/j.addbeh.2019.106228] [Medline: 31838443]
48. Rotondi AJ, Belnap BH, Rothenberger S, Feldman R, Hanusa B, Rollman BL. Predictors of use and drop out from a web-based cognitive behavioral therapy program and health community for depression and anxiety in primary care patients: secondary analysis of a randomized controlled trial. JMIR Ment Health 2024 Jan 17;11:e52197. [doi: 10.2196/52197] [Medline: 38231552]
49. Lundkvist-Houndoumadi I, Thastum M. Anxious children and adolescents non-responding to CBT: clinical predictors and families' experiences of therapy. Clin Psychol Psychother 2017 Jan;24(1):82-93. [doi: 10.1002/cpp.1982] [Medline: 26514088]
50. Lundkvist-Houndoumadi I, Hougaard E, Thastum M. Pre-treatment child and family characteristics as predictors of outcome in cognitive behavioural therapy for youth anxiety disorders. Nord J Psychiatry 2014 Nov;68(8):524-535. [doi: 10.3109/08039488.2014.903295] [Medline: 24754469]
51. Argiros A, Venanzi L, Dao A, et al. Social support and parental conflict as predictors of outcomes of group cognitive behavioral therapy for adolescent depression. Int J Cogn Ther 2023;16(2):202-221. [doi: 10.1007/s41811-023-00159-z] [Medline: 37228300]

Abbreviations

CBT: cognitive behavioral therapy
DSRS-C: Depression Self-Rating Scale for Children
PedsQL: Pediatric Quality of Life Inventory
PHQ-9A: Patient Health Questionnaire for Adolescents
QOL: quality of life
QTA30: Questionnaire for Triage and Assessment with 30 items
RSES: Rosenberg Self-Esteem Scale

XSL•FO
RenderX

Edited by A Mavragani; submitted 01.06.24; peer-reviewed by F Yang, P Croarkin, V Huhn; revised version received 09.05.25; accepted 12.05.25; published 29.07.25.

Please cite as:
Nagamitsu S, Okada A, Sakuta R, Ishii R, Koyanagi K, Habukawa C, Katayama T, Ito M, Kanie A, Otani R, Inoue T, Kitajima T, Matsubara N, Tanaka C, Fujii C, Shigeyasu Y, Matsuoka M, Kakuma T, Horikoshi M
Usefulness of Interventions Using a Smartphone Cognitive Behavior Therapy Application for Children With Mental Health Disorders: Prospective, Single-Arm, Uncontrolled Clinical Trial
JMIR Form Res 2025;9:e60943
URL: https://formative.jmir.org/2025/1/e60943
doi:10.2196/60943

© Shinichiro Nagamitsu, Ayumi Okada, Ryoichi Sakuta, Ryuta Ishii, Kenshi Koyanagi, Chizu Habukawa, Takashi Katayama, Masaya Ito, Ayako Kanie, Ryoko Otani, Takeshi Inoue, Tasuku Kitajima, Naoki Matsubara, Chie Tanaka, Chikako Fujii, Yoshie Shigeyasu, Michiko Matsuoka, Tatsuyuki Kakuma, Masaru Horikoshi. Originally published in JMIR Formative Research (https://formative.jmir.org), 29.7.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Game-Based Social-Emotional Learning for Youth: School-Based Qualitative Analysis of Brain Agents

Elizabeth Liverman[1], BS; David Antognoli[2,3], MS; Cordelia Elaiho[1], MPH; Madison McGuire[1], BS; Abbey Stoltenburg[1], MA; Angel Navarrete[4], BS; Garrett Bates[5], MPH; Thomas Chelius[1], MS; Constance Gundacker[6], MPH, MD; Paula Lumelsky[7], EdD; Brandon Currie[7], PhD; John Meurer Sr[1], MBA, MD

[1]School of Medicine, Medical College of Wisconsin, 8701 Watertown Plank Road, Milwaukee, WI, United States

[2]School of Design, Columbia College Chicago, Chicago, IL, United States

[3]Night City, Chicago, IL, United States

[4]School of Medicine, Southern Illinois University, Springfield, IL, United States

[5]School of Graduate Studies, Institute for Health & Humanity, Medical College of Wisconsin, Milwaukee, WI, United States

[6]Department of Pediatrics, Medical College of Wisconsin, Milwaukee, WI, United States

[7]STRYV365, Milwaukee, WI, United States

Corresponding Author:
John Meurer Sr, MBA, MD
School of Medicine, Medical College of Wisconsin, 8701 Watertown Plank Road, Milwaukee, WI, United States

Abstract

Background: Adverse childhood experiences such as violence, substance use, and family disruption disproportionately affect youth in urban communities, increasing the risk of emotional and behavioral challenges. Social-emotional learning (SEL) and trauma-informed programming are effective strategies for mitigating these effects, fostering resilience, and promoting mental well-being. Game-based learning is a promising, engaging method for delivering SEL content. STRYV365 developed *Brain Agents*, a trauma-informed, game-based SEL intervention aimed at improving emotional regulation, coping strategies, and interpersonal skills among students in grades 5 through 9.

Objective: This study explored students' experiences with and perceptions of *Brain Agents*, evaluating its effectiveness in fostering SEL skills and resilience across 4 diverse urban schools in Milwaukee, Wisconsin.

Methods: A cluster-randomized, incomplete block factorial crossover design was implemented from 2022-2024. Of 1626 eligible students, 329 (20%) had caregiver consent and student assent. Among these, 180 students in grades 5‑9 played *Brain Agents* at school over 4‑5 weeks, for an average of 10 sessions and 23 minutes per session. SEL-related outcomes were assessed using surveys, focus groups, and interviews. Qualitative data were analyzed using Dedoose software, with thematic coding conducted by multiple coders to ensure reliability.

Results: Student demographics included 189/321 (58.9%) Black, 112/321 (34.9%) White, and 221/321 (68.8%) from economically disadvantaged backgrounds. Baseline surveys of 277 children revealed that 202 (72.9%) of students had experienced the death of someone close, 147 (53.1%) had a close contact incarcerated, and 39 (14.1%) reported feeling nervous or anxious daily. Strengths included 230 (83.0%) students reporting life satisfaction and 183 (66.1%) able to calm down when upset. Game performance data from 328 students indicated varying levels of achievement, with a median of 3 (IQR 1.5-4) missions completed, 4 (IQR 2-6) stars earned, 8 positive energies collected, and 2 (IQR 1-2.5) crew members rescued. Grades 7‑8 had the highest engagement, while grade 9 students had the lowest participation. Qualitative analysis from 62 participants identified 8 core themes: qualities of most pride, neighborhood relationships, challenges in life, emotions associated with loss of control, coping strategies, future goals, experiences with *Brain Agents*, and suggestions to improve the game. Students most frequently cited anger as a cause of emotional dysregulation and named coping strategies such as self-calming, asking for help, and perseverance. Feedback on *Brain Agents* highlighted improved focus, emotional control, and critical thinking, with younger students more positively engaged. Suggested improvements included better graphics, more customization, and cooperative play.

Conclusions: *Brain Agents* was positively received by students, particularly those in earlier grades, and demonstrated potential as an effective trauma-informed SEL tool. The findings support the role of game-based interventions in enhancing resilience and emotional intelligence among youth exposed to adversity. Broader implementation may extend benefits to diverse student populations and settings.

(JMIR Form Res 2025;9:e67550) doi:10.2196/67550

XSL•FO
RenderX

KEYWORDS

trauma-informed programming; social-emotional learning; game-based learning; adolescent mental health; qualitative research; educational video games

Introduction

Adverse Childhood Experiences

Child and adolescent health and well-being are influenced by various factors, including early life experiences and educational opportunities. One way to measure trauma in early life is through adverse childhood experiences (ACEs), which include familial incarceration, drug use in the home, and witnessing or experiencing violence [1]. Youth exposure to trauma is linked to poor health, school absenteeism, and behavioral issues [2]. The COVID-19 pandemic is a recent ACE that disrupted typical childhood interactions, causing significant hardships for many children [3]. The COVID-19 pandemic saw a rise in gaming as a coping mechanism, reflecting the role of games in promoting relaxation, emotional regulation, and social connection during periods of stress [4]. Despite indicators of positive mental health in most children, such as curiosity, persistence, and self-control, depression and anxiety have increased over the past 2 decades. For adolescents, depression, substance use, and suicide are significant concerns [1]. Given the prevalence of ACEs and rising concerns for student mental health, programs to support student resilience and behavioral and emotional regulation have become essential.

Trauma-Informed Programming

Trauma-informed programming is one method used to address these issues. This approach sensitizes educators to trauma while improving school engagement, reducing behavioral problems, and enhancing students' social and emotional skills [5]. The goal of trauma-informed programming is to reduce stress, increase coping abilities, and decrease impulsive behavior in youth [6]. Another complementary approach is social-emotional learning (SEL), which focuses on self-awareness, self-management, and social mindfulness to help youth make responsible decisions, achieve their goals, and build relationships [7]. SEL is closely tied to emotional intelligence, the ability to monitor and manage one's own and others' emotions and use this information to guide thinking and actions [8]. The ability model of emotional intelligence focuses on the capacity to accurately perceive, understand, and manage emotions—skills that are essential for building resilience and social-emotional competence in youth [9]. School-based SEL programs can lead to positive attitudes toward self and others, positive behavior, academic success, and mental wellness [10]. Research has shown that video games can support emotion regulation, a core component of SEL [11]. This evidence reinforces the use of a game-based tool to help students identify and manage emotions.

Game-Based Learning

Game-based learning is an innovative way to integrate trauma-informed programming and SEL into school curricula. This strategy applies game elements such as scoring, challenge, narrative, role-playing, and progression to motivate and engage players in educational content where they can learn by doing [12]. While excessive use of video games can have negative health and behavior implications, video games also offer many benefits. These include more accurate attention, higher spatial resolution in visual processing, enhanced mental rotation abilities, improved mood, increased relaxation, and lower levels of anxiety [13-15]. When used in moderation, video games can be a useful SEL tool for fostering positive behaviors and emotions. For instance, the Spanish video game *Aislados* was designed to improve youth psychological well-being and help them avoid risk behaviors related to addiction, violence, or emotional disorders. Those who played the game experienced increased health-related quality of life, positive affect, and mental health [13]. Additionally, research on the video game *Emociomente* showed that video games could promote emotional intelligence in adolescents when integrated with a guided and assisted framework [16]. Similarly, a systematic review highlighted the therapeutic and preventive benefits of video games in child and adolescent psychiatry [17], further supporting the potential of *Brain Agents* to address the needs of youth facing adversity.

Game-based learning has also been beneficial in clinical settings for mental health outcomes, provided there is a successful integration of health care and entertainment elements [18]. Research has demonstrated that games for mental health are generally accessible, feasible, and effective, supporting their potential to promote emotional and cognitive well-being in children and adolescents [19]. This aligns with the purpose of *Brain Agents* as a tool to foster resilience and SEL. *Brain Agents* was developed through an iterative process, incorporating student feedback to refine gameplay and enhance learning outcomes [20].

Video games, often viewed skeptically regarding their impact on child health, are favored leisure activities among children [21]. Engaging with games might help to develop emotional self-regulation skills, promoting mental health and well-being among children [11]. Besides emotional regulation, video games can also improve social skills [22]. This benefit supports the idea that game-based programs provide valuable opportunities to promote social-emotional competencies and adjustment in adolescence [23]. Research on prosocial video games highlights their ability to improve social outcomes in adolescence and emerging adulthood [24], consistent with *Brain Agents'* focus on fostering emotional and interpersonal skills. By leveraging the engaging nature of digital tools, such programs can spark children and adolescents' interest in emotions and influence various personal, emotional, and social factors.

Moreover, the challenges in video games, such as facing failure and persisting through it, offer valuable life skills [25]. Persistence, analytical thinking, and self-efficacy are among the traits nurtured by engaging with challenging games, contributing to personal development. Overall, video games, when designed and used effectively, can positively impact child health, social development, and emotional regulation, presenting valuable opportunities for intervention and education.

XSL•FO
RenderX

STRYV365 Interventions

Following principles from these studies, STRYV365, a nonprofit organization based in Milwaukee, Wisconsin, worked to advance and support the idea that educational video games could help children cope with ACEs. The goal was to be one of the first US studies to show that video games can improve SEL, resiliency, and positive life experiences in school children. STRYV365 partnered with the development studio Night City to design and develop the video game *Brain Agents*, with SEL and trauma-informed content as part of a multifaceted approach to foster resilience, promote mindfulness, and improve problem-solving and critical thinking in fifth to ninth-grade students. As others have found, we expect that when players participate and engage with *Brain Agents*, they can demonstrate openness, self-reflection, empathy, and promote tolerance of others' perspectives [26].

For this block randomization study, in addition to providing *Brain Agents* to students in grades 5‑9 at 4 schools, the students also experienced the STRYV365 program *peak team*. The *peak team* coach mentor program strengthens students' social-emotional skills and resilience through a trauma-informed approach integrating sports and mentorship, complemented by the *Brain Agents* video game. Grounded in principles of trauma-informed care, the program promotes emotional regulation, resilience, and decision-making by engaging students in physical activities such as teamwork-based games, reflective writing, and discussions. Delivered over 4 to 10 weeks in gym or physical education classes, the program combines fun and educational activities to empower students to overcome adversity and thrive. Qualitative research about *peak team* is reported in another paper.

The primary aim of this research is to describe the experiences of students participating in the study and their perceptions of the *Brain Agents* video game. The qualitative study uses the game to help students identify and describe their emotions, increasing awareness of their actions. This study also included an analysis of whether *Brain Agents* can be integrated into school settings to equip youth affected by trauma and adverse experiences with the skills and resilient mindset needed for a successful future. In surveys, focus groups, and interviews, the research team asked students to provide general feedback on the game, describe their personal experiences with *Brain Agents*, and offer suggestions for improving future versions of the game.

Methods

Brain Agents Game Design and Objectives

Brain Agents, created in collaboration between STRYV365, development studio Night City, and behavioral health professionals, was designed to foster resilience and support SEL by translating STRYV365's established curriculum to an engaging digital game format. Targeting ages 10-15 years, the game uses narrative context and minigames to deliver SEL content, including trauma resilience, coping strategies, cognitive reframing, communication, relationship building, and emotional awareness. The game also incorporates STRYV365's CRAFT process: Catch Yourself, Relax, Assess, Focus, Think, Decide, and Act.

Development Process

Development of the game initiated in the spring of 2020 and has involved Columbia College Chicago game design and programming students, hired as interns through the US Federal Work-Study program. *Brain Agents* was developed using the Unity game engine and initially designed for mobile devices. However, due to Chromebook use in participating schools, the game was adapted for this platform. The game's low polygon 3D art style ensures compatibility with lower-powered hardware while maintaining immersion and enjoyment from contemporary 3D game design.

Early Access Model

Brain Agents uses an "early access" model, allowing students to play while the game is still in development. This approach can attract students interested in game development but may disappoint those expecting a finished product. Continuous, iterative improvements and content additions mean players' experiences can vary, with later participants likely having a more refined experience. The development team has balanced student desires for features from popular games with the need to incorporate a learning agenda and the scope limitations of a nonprofit indie game project.

Access and Future Availability

As of publication, *Brain Agents* is not publicly available. Study participants access the game via a designated website using a special code name. Future plans include making the game available in schools, youth organization programs, and potentially at home to support parents.

Narrative Context and Gameplay Mechanics

In *Brain Agents*, players are explorers on a space vessel, responding to a distress signal. They encounter an alien computer program named MAL that infects both computer systems and human minds, causing cognitive distortions in the crew. Players, guided by the ship's captain, must gather positive energy to rescue the crew by reframing these cognitive distortions (Figures 1 and 2).

XSL•FO
RenderX

Figure 1. *Brain Agents* gameplay showing narrative context. The ship captain explains that the cryo-sleeping crew's positive thoughts have been displaced by an alien computer virus that feeds off chaos and negativity. The displaced positive thoughts have been scattered around the ship in the form of "positive energy collectibles" that the player must find and collect.

Figure 2. *Brain Agents* gameplay showing energy collection via positive thought construction. Upon gathering enough positive energy, the player attempts to save a crew member from the alien computer virus.

Minigames and Learning Objectives

During this study, *Brain Agents* included 5 minigames that help players practice coping mechanisms, connect with emotions, and build social relationships with nonplayer characters. After this study, additional minigames have been developed.

Of the reframing minigame, players deposit positive energy into a reframing machine to transform cognitive distortions into constructive thoughts, rescuing crew members from MAL's influence (Figures 3 and 4).

XSL•FO
RenderX

Figure 3. *Brain Agents* reframing minigame. To rescue members, the player deposits collected positive energy into a "reframing machine" that shows a cognitive distortion within a slumbering crew member's mind. Each thought is divided into 2 sections, and the player must identify which is an unhelpful cognitive distortion and eject it.

Figure 4. *Brain Agents* reframed positive phrases from the reframing minigame. Upon identifying the unhelpful thought to reframe, the player must adjust a dial to the "sweet spot" to reframe it into a constructive, positive idea.

Of the breathing minigame, players practice controlled breathing to calm themselves after in-game challenges, reinforcing mindfulness and relaxation techniques.

Of the body chart minigame, players identify physical manifestations of emotions on a body chart and associate colors with feelings, helping them become more attuned to their emotions (Figures 5 and 6).

XSL•FO
RenderX

Figure 5. *Brain Agents* character HEART from the body chart minigame for identifying the physical manifestation of emotions is accessed through a robot named HEART that needs the player's help in learning about human emotions. HEART: Human Emotion Acquisition Robot.

Figure 6. *Brain Agents* body chart minigame HEART asks the player to identify how their physical body feels when they experience various emotions. HEART: Human Emotion Acquisition Robot.

Of balance minigame, players balance a tower of blocks to build a firewall, emphasizing calmness under pressure and teamwork (Figure 7).

Of the hacking minigame, players solve logic puzzles, reinforcing problem-solving skills and agency through the theme that systems can be changed and improved (Figure 8).

XSL•FO
RenderX

Figure 7. *Brain Agents* balance minigame. The player must catch blocks and lean a tower without tipping it over. The minigame uses gyroscope motion input to capture physical movement on devices equipped with one.

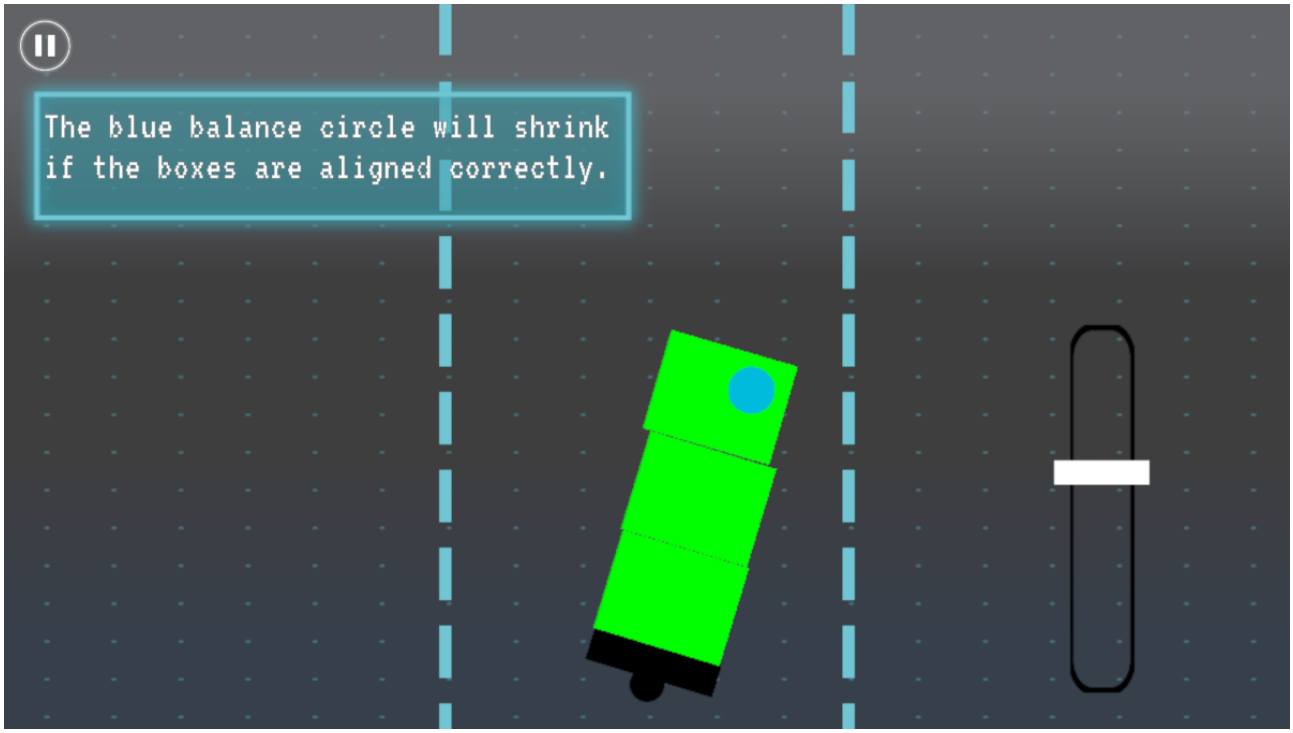

Figure 8. *Brain Agents* hacking minigame. Players must rearrange directional command blocks to create a path from the beginning to the goal. This image shows the tutorial where the simplest puzzle is presented to teach how to play.

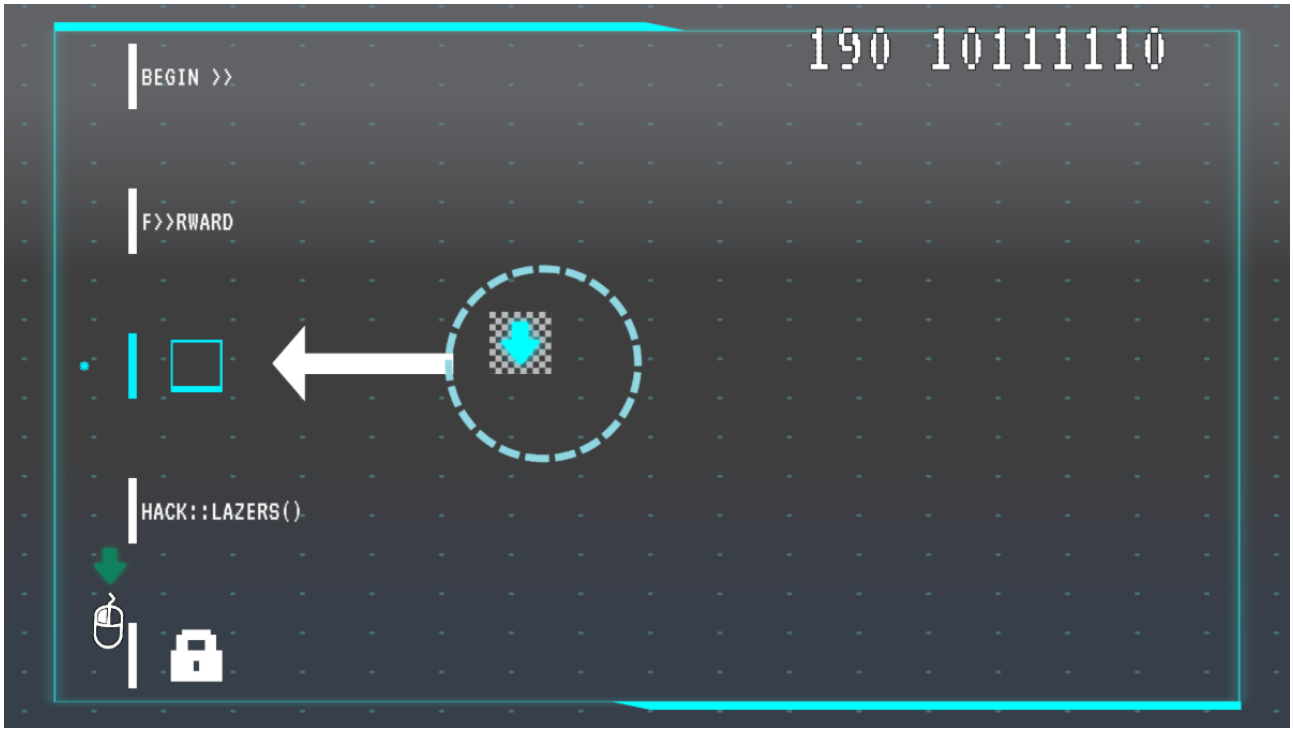

Ethical Considerations

The Medical College of Wisconsin Human Research and Review Board (PRO037500) approved the research protocol for this evaluation. Parents and caregivers consented for their eligible children to participate in the research, students assented to their participation, and teachers and school administrators consented to joining focus groups. Privacy and confidentiality were protected throughout this study. Students provided their identity in surveys to link the data longitudinally over 2 years, then the data were deidentified for analyses. Students and teachers were deidentified in qualitative data analysis. Participants were not compensated.

XSL•FO
RenderX

School Study Setting

STRYV365 partnered with 4 Milwaukee schools (2 public and 2 private choice schools), involving 1626 students in grades 5 - 9. The demographic breakdown from school data was 991 (60.9%) Black or African American, 309 (19.0%) White, 179 (11.0%) Hispanic or Latinx, 98 (6.0%) Asian or Hmong, and 49 (3.0%) mixed racial or ethnic backgrounds, with 1122 (69.0%) experiencing economic disadvantage. In state standardized examinations, 943 (58.0%) were below basic, 439 (27.0%) basic, and 244 (15.0%) proficient in English, language arts, and mathematics. A total of 455 (28.0%) of students were chronically absent. Despite these challenges, 1431 (88.0%) of students at these schools earned high school diplomas when the study began, indicating a strong commitment to education and resilience among the student population.

Researcher Qualifications and Relationships (Reflexivity)

This study was conducted from a positivist paradigm, using sociocultural theory to inform the conceptual framework. This approach allowed for the coconstruction of data and analysis through participants' and researchers' voices. The research team included members with similar racial or ethnic backgrounds and traumatic experiences as the participants, enhancing reflexivity. The coauthors are professionals in the health and education fields.

Sampling, Recruitment, Consent, and Assent

CITI (Collaborative Institutional Training Initiative)–trained STRYV365 and school staff recruited participants from a convenience sample of 1626 students' parents or caregivers, who provided informed consent digitally. A video at the STRYV365 website [27] summarized the consent form. Parents or caregivers were recruited at back-to-school events, open houses, parent-teacher conferences, and through emails in education tools PowerSchool, Google Classroom, ClassDojo, and school dashboards. Students provided written assent in REDCap (Research Electronic Data Capture; Vanderbilt University) before completing surveys and participating in focus groups or interviews.

Children were selected for focus groups and interviews based on their participation in the *peak team* and *Brain Agents* programs during the preceding semester. The randomized selection aimed to ensure a diverse representation of students in terms of grade level, gender, and race or ethnicity. Focus groups and interviews of students were conducted in a comfortable and confidential setting. Focus groups with teachers and school administrators were done in person or via videoconferencing. The findings have been summarized and discussed with parents or caregivers and students at the end of this study.

Conceptual Framework and Study Design

This study hypothesizes that *Brain Agents* will positively impact student feelings, attitudes, behaviors, and school performance. A pilot cross-sectional series of surveys and focus groups was conducted in the 2021-2022 academic year [28], followed by a cluster randomized, incomplete block factorial, crossover design in the 2022-2024 academic years (Multimedia Appendix 1). Surveys were conducted at 6 time points, with key outcomes to be reported in 2025 (Multimedia Appendix 2).

Data Collection

Baseline surveys were completed in October-November 2022, with additional surveys and focus groups or interviews in January-February and April-May 2023. Research staff facilitated focus groups with younger students and individual interviews with older students, ensuring comfort and openness. Audio recordings were transcribed for analysis.

Qualitative Methods

Qualitative data were collected through focus groups and interviews conducted at the end of the fall and spring semesters to evaluate both *Brain Agents* and *peak team* programs' impact on students' SEL outcomes. Research staff facilitated focus groups with fifth to seventh graders, which encouraged peer interaction and group discussion dynamics, while individual interviews with eighth to ninth graders allowed for more in-depth exploration of personal experiences and perspectives. This approach ensured both comfort and openness for participants across age groups.

The qualitative evaluation was guided by a semistructured interview protocol (Multimedia Appendix 3), designed to elicit both open-ended responses and targeted insights aligned with the study objectives. Audio recordings of all focus groups and interviews were professionally transcribed and securely stored to maintain confidentiality.

Data processing and analysis were conducted using Dedoose (Sociocultural Research Consultants, LLC) software. A thematic analysis approach was used, involving an iterative process of generating and refining codes and themes. Coding began inductively, allowing themes to emerge naturally from the data, with additional subthemes identified during subsequent analysis.

A team of 7 coders participated in the analysis, with 2 coders independently reviewing each transcript. Discrepancies in coding were resolved through discussion to optimize intercoder reliability and enhance the rigor of the analysis. This systematic approach ensured a comprehensive understanding of the qualitative data, with a focus on identifying patterns and themes related to the interventions' impact on students' SEL outcomes. Qualitative findings focused on teachers' perspectives [29] and students' views of *peak team* are published elsewhere.

Results

Participants

A total of 329 parents or caregivers consented to the participation of their children in this study and children assented to join this study. The demographics of 321 children included 132 (41.1%) girls, 170 (53.0%) boys, 10 (3.1%) nonbinary, 6 (1.9%) who preferred not to answer, and 3 (0.9%) other. Their reported race or ethnicity was 189 (58.9%) Black or African American, 112 (34.9%) White, 32 (10.0%) Hispanic or Latino, 22 (6.9%) Asian, 16 (5.0%) American Indian, and 6 (1.9%) Pacific Islander; the total exceeds 321 due to mixed race/ethnicity children. As of the winter of the second year of this study (2023), nearly 32/329 (9.7%) of students had dropped

XSL•FO
RenderX

out, primarily due to leaving the school or parental withdrawal of consent.

Qualitative Data Collection

Focus groups and interviews were conducted in private rooms at the schools. One-third of participants, equally split between boys and girls, who were exposed to *Brain Agents* during the 2-year study, participated in a focus group or interview about their experiences. Twenty-two students from grades 5 - 7, aged 10 - 12 years, participated in 1 of 6 focus groups, each lasting an average of 26 minutes. Forty students from grades 8 - 9, aged 13 - 15 years, participated in individual interviews, each averaging 22 minutes.

Baseline Survey Results

Using SAS (IBM Corp) to calculate percentages, the baseline survey of 277 students revealed that 202 (72.9%) of students had someone close to them die, 147 (53.1%) had someone close in jail, and 80 (28.9%) had someone close with drinking or drug problems. Additionally, 39 (14.1%) felt nervous or anxious daily, and 33 (11.9%) felt down or depressed daily; school psychologists or counselors were notified to follow up with their families. Strengths and assets reported included 230 (83.0%) being satisfied with life and aware of their feelings, 183 (66.1%) able to calm down when upset, 175 (63.2%) having no physical fights in the past 3 months, and 199 (71.8%) talking with friends about problems. Additional survey findings are reported in another paper.

Game Use

During year 1 (2022-2023), a total of 328 children enrolled in the research played *Brain Agents* at school only. These students were distributed as 98 (29.9%) in grades 5-6, 145 (44.2%) in grades 7-8, and 85 (25.9%) in grade 9. Students played the game 2 to 3 days weekly for 4 to 5 weeks, averaging 10 sessions per semester and 23 minutes per session, totaling approximately 3.8 hours of playtime. Most students did not complete the entire game.

In *Brain Agents*, players aim to rescue all 15 crew members and defeat the alien malware by collecting up to 5 positive energies in each of the 30 playable missions or levels. Players start with 3 missions unlocked. New missions can be unlocked by earning up to 3 stars per mission: 1 for completing the mission, 1 for finding all 5 positive energies, and 1 for completing the mission under a target time. Thus, the amount of positive energy collected and the number of stars earned reflect player success and progress in the game.

Game performance data, analyzed with statistical software SAS, showed that grades 5 - 6 collected the most positive energy at the 75th percentile, grades 7 - 8 had the highest median positive energy of 8 (IQR 3-16), and grade 9 had the lowest performance in all measures (Table 1). The association of game performance with individual survey, focus group, and interview responses is pending further analysis.

Table . *Brain Agents'* game performance by grade after 4 semesters of play.

	Participants, n	Minimum	Maximum	Median (IQR)
Overall				
Missions completed	328	0	13	3 (1.5-4)
Stars awarded	328	0	38	4 (2-6)
Positive energy	328	0	67	7 (3-14)
Crew rescued	328	1	9	2 (1-2.5)
Grades 5 - 6				
Missions completed	98	0	13	2 (2-4)
Stars awarded	98	0	25	4 (2-6)
Positive energy	98	0	63	7 (4-12)
Crew rescued	98	1	9	2 (1-4)
Grades 7 - 8				
Missions completed	145	0	13	3 (2-5)
Stars awarded	145	0	38	4 (2-7)
Positive energy	145	0	67	8 (3-16)
Crew rescued	145	1	9	2 (1-3)
Grade 9				
Missions completed	57	0	8	2 (1-4)
Stars awarded	57	0	17	3 (1-4)
Positive energy	57	0	42	6 (2-11)
Crew rescued	57	1	7	2 (1-2)

XSL•FO
RenderX

Synthesis of Themes

Eight core themes emerged from student focus groups and interviews: experience with *Brain Agents*, awareness, and ability to control emotions, goals, current challenges, and perceptions of the neighborhood (Table 2). Themes were derived from specific research questions about life experiences and goals, SEL, resilience, and game evaluation.

Table . Themes and subthemes used to code transcripts of focus groups and interviews of participating students in year 1.

Theme	Subtheme
Experience with *Brain Agents*	• Positive experience with *Brain Agents* • Negative experience with *Brain Agents* • Neutral *Brain Agents* experience • How *Brain Agents* was not helpful • Feedback on improving *Brain Agents*
Experience with *peak team*	• Positive experience with *peak team* • Negative experience with *peak team* • Neutral *peak team* experience • How to improve *peak team*
Qualities of most pride	• Extrinsic qualities (good at sport, etc) • Intrinsic qualities (perseverance, etc) • Unable to name quality (negative self-image)
Emotion associated with loss of control	• Anger • Annoyance • Stress or anxiety • Positive reactions • Suppress emotions • No emotion • Other
Coping strategies	• Ask for help • Perseverance or grit • Self-calm • Escape, distraction, or sleep • Self-improvement, self-efficacy, or choices • Time management • Retaliation • No coping strategies
Future goal in 5 years	• Educational goal • Career goal • Stability goal • Extracurricular activities or hobbies • Other goal • No goal
Challenges in life	• Family challenges • School challenges • School focus • Sports or extracurricular challenges • Anxiety • Depression • New responsibilities
Neighborhood or neighbors	• Peaceful or calm neighborhood • Loud or noisy neighborhood • Rough neighborhood • Busy neighborhood • Can rely on neighbors • Cannot rely on neighbors • Feel safe in neighborhood • Feel unsafe in neighborhood

XSL•FO
RenderX

Experiences With Brain Agents

Brain Agents was generally described as fun and helpful, particularly in improving self-awareness, focus, and emotional regulation. Positive feedback included comments on its calming effects and its role in enhancing critical thinking and problem-solving skills. Some students found it boring or had trouble with gameplay, but overall, the sentiment was positive.

Suggestions to Improve Brain Agents

Students suggested improvements such as better graphics, customizable avatars, cooperative play, clearer instructions, more problem-solving opportunities, expanded maps, additional missions, and difficulty adjustments. They also recommended incorporating self-reflection before gameplay.

Neighborhood Relationships

Students generally felt safe in their neighborhoods and trusted their neighbors, although some reported neighborhood problems. Descriptions ranged from “quiet” and “safe” to “loud” with frequent emergency responses and occasional gunshots. A few students mentioned living in “rough” neighborhoods with violence, shootings, and break-ins. Most students, however, felt they could rely on their neighbors for safety and support.

Qualities of Most Pride

Students expressed pride in various traits, both intrinsic (eg, work ethic, determination, intelligence, and creativity) and extrinsic (eg, athleticism, artistic skills, and relationships). However, some students had negative views of themselves, with 1 student explicitly stating they had nothing to be proud of. Despite these few negative outlooks, most students could identify positive qualities.

Challenges in Life

Students discussed family challenges, academic focus, and new responsibilities as significant challenges. Some described dealing with family losses, while others focused on academic pressures like procrastination and anxiety about school projects. Older students mentioned balancing home responsibilities with school demands.

Emotions Associated With Loss of Control

Anger was the most frequently reported emotion leading to loss of control, often triggered by video game frustration or sibling annoyances. Some students reported trying to suppress their emotions to avoid negative outcomes, while others described actively trying to calm themselves.

Coping Strategies

Students reported various coping strategies, including self-calming, self-improvement, avoidance, asking for help, and perseverance. *Brain Agents* encouraged students to remain calm, consider options, and breathe when stressed, which some students mentioned as helpful in managing their emotions.

Future Goals

Students’ goals for the next 5 years varied but included themes of education, career aspirations, extracurricular activities, and personal ambitions. Educational goals included attending college, pursuing specific degrees, and finishing high school. Career aspirations ranged from engineering and medicine to real estate and entrepreneurship. Extracurricular goals involved sports, the arts, and club participation. Other personal goals included caring for family, obtaining a driver’s license, and living independently.

Discussion

Summary of Main Findings

This study used a qualitative design with triangulation of focus groups, interviews, and surveys to conduct an in-depth exploration of students' experiences with the *Brain Agents* game and its impact on their learning, resilience, and life experiences. After 1 year of STRYV365 *Brain Agents* gameplay, qualitative insights highlighted both challenges and strengths for participating youth. These preliminary results will guide future coaching and game improvements.

The research aimed to gather qualitative insights about student experiences with the *Brain Agents* video game and its ability to promote SEL in schools. Overall, students had a generally positive experience with *Brain Agents*, reporting improvements in self-awareness, focus, and critical thinking skills. Many found *Brain Agents* entertaining, although some reported negative experiences, including glitches that caused frustration or boredom. Younger students (grades 5 - 6) found the game more challenging compared to older students (grades 8 - 9), who had lower engagement and scores, indicating the game may be more effective for younger audiences. Suggestions for future improvements included better graphics, customizable avatars, clearer instructions, and appropriate difficulty levels.

The game-based SEL of *Brain Agents* had a positive impact on students, demonstrating its potential as a tool to promote SEL in schools. This study highlighted the prevalence of ACEs among students and showed that *Brain Agents* could be used to build coping skills and resilience in a fun and competitive environment. Teachers, parents, and students can use the game’s lessons both in school and at home to foster resilience and teach SEL skills. For researchers and game designers, this study highlighted both successes and areas for improvement in the game, emphasizing the potential of game-based learning in education.

Comparison With Other Educational Games

The *Brain Agents* game shares similarities with other educational games such as the Spanish *Aislados* and Italian *Emociomente*. The *Aislados* study reported increases in health-related quality of life, positive affect, and mental health but no improvement in emotional intelligence [13]. While this study did not directly measure quality of life or mental health, many students reported positive experiences and joyful emotions after playing *Brain Agents*, suggesting potential improvements in these areas. Positive life experiences are protective against ACEs. Future survey data may provide more concrete evidence of improved mental health.

The *Emociomente* study linked their game to improved emotional intelligence [16]. Similarly, students reported increased awareness of emotions, use of coping strategies, and improved behavioral responses after playing *Brain Agents*.

XSL•FO
RenderX

Techniques such as breathing exercises and self-calming taught in the game were applied by students in real-life situations, demonstrating the game's practical impact on emotional regulation.

The potential of video games as therapy extends beyond educational settings, with a research agenda proposed by Carras et al [30], highlighting the value of commercial video games in therapeutic contexts. This underscores the broader applicability of *Brain Agents* as a tool for fostering resilience and emotional intelligence.

Transferability of *Brain Agents*

Brain Agents, accessible online through personal or school computers and mobile devices, has high potential for dissemination to other schools and communities. Studies examining serious games for children with specific conditions, such as ADHD, have identified subgroup variations in effectiveness [31]. Insights from such research could inform future adaptations of *Brain Agents* to maximize its impact across diverse student populations. Currently offered only in English, the game could be developed further and translated to make a global impact, promoting youth coping and resilience worldwide.

Limitations

A limitation of this preliminary study is the lack of complete objective survey measures as the primary outcome to assess students' SEL, resilience, and positive life experiences. The survey results will be reported after the 2-year study is completed. The nonspecific qualitative insights from students who participated in both *Brain Agents* and other STRYV365 programs may not solely reflect their experiences with the game. Additionally, the diverse student population in 4 Milwaukee schools may not be generalizable to children in other settings. For example, systematic reviews of cognitive training interventions for autism spectrum disorder [32] suggest that similar game-based approaches may be tailored to support other neurodiverse populations.

In some cases, individual student game performance tracking was unreliable due to issues such as students sharing accounts and leaving the game running while not actively playing. These issues affected the measurement of average playtime and other game metrics. In year 2, the research team enforced the use of unique accounts for each student and encouraged proper game usage to ensure accurate data collection.

The qualitative nature of focus group and interview analyses also introduces potential bias. Although each transcript was coded by at least 2 individuals to minimize bias, the qualitative analysis process has inherent limitations due to human factors.

Conclusions

This study suggests that a digital video game intervention can positively impact SEL skills, resilience, and behaviors such as coping skills. While similar studies have been conducted in Europe, this is one of the first of its kind in the United States. This study's inclusion of students from grades 5 to 9 provides unique insights into the responsiveness of different age groups to SEL video games, with younger students appearing more engaged with *Brain Agents*. This study also captured experiences unique to marginalized children in urban schools, demonstrating that video games can be a useful tool in SEL education and building resilience and positive life experiences.

By the end of the 2023‑2024 academic year, students completed 2 years of surveys. The STRYV365 research team expects that these surveys, combined with school record analysis, will demonstrate the program's impact on students' SEL, resilience, and positive life experiences. In the next year, *Brain Agents* will undergo several updates, including a new logo, more varied and engaging mission content, refined target times for missions, additional minigames, deeper dialogue choices with crew members, and in-game customization options.

To enhance teacher experience, STRYV365 will work to better connect the game's learning scenarios to classroom curricula, allowing teachers to reinforce concepts more effectively. From a research perspective, the game will implement improved performance tracking. Ongoing studies of *Brain Agents* will continue to use student feedback to refine and enhance the game, aiming to improve student SEL, resilience, and positive life experiences further.

Acknowledgments

We gratefully acknowledge the work of Cody Hallowell, Melanie Murphy, Souki Aouita, Kristin Bolan, Jen Stock, Danielle Winston, and the STRYV365 coaches. We also deeply appreciate the partnership and support of leaders, teachers, staff, parents or caregivers, and students or scholars at Brown Deer School District, Dr Howard Fuller Collegiate Academy, Milwaukee Academy of Science, and St Francis School District.

Data Availability

The datasets generated or analyzed during this study are not publicly available because the participants did not consent or assent to sharing them with researchers not directly involved in this study.

Conflicts of Interest

PL and BC are leaders in nonprofit STRYV365, Inc, which delivered the intervention programs. Schools were not charged for the programs.

XSL•FO
RenderX

Multimedia Appendix 1
Incomplete block factorial design for academic years 1 and 2.
[DOCX File, 37 KB - formative_v9i1e67550_app1.docx]

Multimedia Appendix 2
STRYV365 student survey instrument.
[PDF File, 58 KB - formative_v9i1e67550_app2.pdf]

Multimedia Appendix 3
STRYV365 guide for conducting focus groups and interviews.
[DOCX File, 31 KB - formative_v9i1e67550_app3.docx]

References

1. Data and statistics on children's mental health. Centers for Disease Control and Prevention. 2024 Aug 19. URL: https://www.cdc.gov/children-mental-health/data-research/index.html [accessed 2025-06-28]
2. Christian-Brandt AS, Santacrose DE, Barnett ML. In the trauma-informed care trenches: teacher compassion satisfaction, secondary traumatic stress, burnout, and intent to leave education within underserved elementary schools. Child Abuse Negl 2020 Dec;110:104437. [doi: 10.1016/j.chiabu.2020.104437]
3. McManus MA, Ball E. COVID-19 should be considered an adverse childhood experience (ACE). JCSWB 2020;5(4):164-167. [doi: 10.35502/jcswb.166]
4. Pallavicini F, Pepe A, Minissi ME. Gaming in the time of COVID-19: what do people play and why? Int J Environ Res Public Health 2021;18(11):5722. [doi: 10.3390/ijerph18115722]
5. Menu of trauma-informed programs for schools REL appalachia cross-state collaborative to support schools in the opioid crisis. Institute of Education Sciences. 2017. URL: https://ies.ed.gov/ncee/rel/regions/appalachia/events/materials/04-8-20-Handout2_menu-trauma-informed-programs-for-schools.pdf [accessed 2025-06-28]
6. Giordano F, Cipolla A, Ungar M. Tutor of resilience: a model for psychosocial care following experiences of adversity. Front Psychiatry 2021;12:559154. [doi: 10.3389/fpsyt.2021.559154] [Medline: 33833694]
7. What is social emotional learning (SEL): why it matters. National University. 2025. URL: https://tinyurl.com/3cvb23wh [accessed 2025-06-28]
8. Toh W, Kirschner D. Developing social-emotional concepts for learning with video games. Comput Educ 2023 Mar;194:104708. [doi: 10.1016/j.compedu.2022.104708]
9. Mayer JD, Caruso DR, Salovey P. The ability model of emotional intelligence: principles and updates. Emot Rev 2016 Oct;8(4):290-300. [doi: 10.1177/1754073916639667]
10. Mahoney JL, Durlak JA, Weissberg RP. An update on social and emotional learning outcome research. Phi Delta Kappan 2018 Dec;100(4):18-23. [doi: 10.1177/0031721718815668]
11. Villani D, Carissoli C, Triberti S, Marchetti A, Gilli G, Riva G. Videogames for emotion regulation: a systematic review. Games Health J 2018 Apr;7(2):85-99. [doi: 10.1089/g4h.2017.0108] [Medline: 29424555]
12. Aldrich C. The Complete Guide to Simulations and Serious Games: How the Most Valuable Content Will Be Created in the Age Beyond Gutenberg to Google, 1st edition: Pfeiffer; 2009.
13. Cejudo J, Losada L, Feltrero R. Promoting social and emotional learning and subjective well-being: impact of the "Aislados" intervention program in adolescents. IJERPH 2020;17(2):609. [doi: 10.3390/ijerph17020609]
14. Green CS, Bavelier D. Learning, attentional control, and action video games. Curr Biol 2012;22(6):R197-R206. [doi: 10.1016/j.cub.2012.02.01215]
15. Ho RC, Zhang MWB, Tsang TY, et al. The association between internet addiction and psychiatric co-morbidity: a meta-analysis. BMC Psychiatry 2014 Jun 20;14:183. [doi: 10.1186/1471-244X-14-183] [Medline: 24947851]
16. Carissoli C, Villani D. Can videogames be used to promote emotional intelligence in teenagers? results from EmotivaMente, a school program. Games Health J 2019 Dec;8(6):407-413. [doi: 10.1089/g4h.2018.0148] [Medline: 31241356]
17. Zayeni D, Raynaud JP, Revet A. Therapeutic and preventive use of video games in child and adolescent psychiatry: a systematic review. Front Psychiatry 2020;11:36. [doi: 10.3389/fpsyt.2020.00036] [Medline: 32116851]
18. Lukka L, Palva JM. The development of game-based digital mental health interventions: bridging the paradigms of health care and entertainment. JMIR Serious Games 2023 Sep 4;11:e42173. [doi: 10.2196/42173] [Medline: 37665624]
19. Lau HM, Smit JH, Fleming TM, Riper H. Serious games for mental health: are they accessible, feasible, and effective? A systematic review and meta-analysis. Front Psychiatry 2020;11:508. [doi: 10.3389/fpsyt.2020.00508] [Medline: 28149281]
20. Scholten H, Granic I. Use of the principles of design thinking to address limitations of digital mental health interventions for youth: viewpoint. J Med Internet Res 2019 Jan 14;21(1):e11528. [doi: 10.2196/11528] [Medline: 31344671]
21. Kovess-Masfety V, Keyes K, Hamilton A, et al. Is time spent playing video games associated with mental health, cognitive and social skills in young children? Soc Psychiatry Psychiatr Epidemiol 2016 Mar;51(3):349-357. [doi: 10.1007/s00127-016-1179-6] [Medline: 26846228]

XSL•FO
RenderX

22. Craig AB, Brown ER, Upright J, DeRosier ME. Enhancing children's social emotional functioning through virtual game-based delivery of social skills training. J Child Fam Stud 2016 Mar;25(3):959-968. [doi: 10.1007/s10826-015-0274-8]
23. de la Barrera U, Mónaco E, Postigo-Zegarra S, Gil-Gómez JA, Montoya-Castilla I. EmoTIC: impact of a game-based social-emotional programme on adolescents. PLoS ONE 2021;16(4):e0250384. [doi: 10.1371/journal.pone.0250384]
24. Holl E, Bernard S, Melzer A. Examining the social outcomes of playing prosocial video games in late adolescence and emerging adulthood. J Appl Dev Psychol 2020;71:101212. [doi: 10.1016/j.appdev.2020.101212]
25. Hefkaluk N, Linehan C, Trace A. Fail, fail again, fail better: how players who enjoy challenging games persist after failure in "Celeste". Int J Hum Comput Stud 2024 Mar;183:103199. [doi: 10.1016/j.ijhcs.2023.103199]
26. Whitby M, Kowert R. Designing games to challenge the stigma around mental health. : Take This; 2022 URL: https://tinyurl.com/2u7s9tpe [accessed 2025-06-28]
27. STRYV365. URL: https://www.stryv365.org/ [accessed 2025-07-01]
28. Meurer JR, Schevers-Lumelsky P, Chelius TH, Murphy M, Hallowell CC, Currie B. peak team: survey of student strengths and needs and teacher's evaluation of an innovative trauma-informed school program. J Child Adolesc Trauma 2025 Jun;18(2):257-264. [doi: 10.1007/s40653-024-00664-z] [Medline: 40469782]
29. Stoltenburg A, McGuire M, Liverman E, et al. STRYV365 peak team and Brain Agents: teacher perspectives on school impact of a trauma-informed, social-emotional learning approach for students facing adverse childhood experiences. Front Psychol 2024;15:1388499. [doi: 10.3389/fpsyg.2024.1388499] [Medline: 39439751]
30. Carras MC, Van Rooij AJ, Spruijt-Metz D, et al. Commercial video games as therapy: a new research agenda to unlock the potential of a global pastime. Front Psychiatry 2017;8:300. [doi: 10.3389/fpsyt.2017.00300] [Medline: 29403398]
31. Bul KCM, Doove LL, Franken IHA, Oord SD, Kato PM, Maras A. A serious game for children with attention deficit hyperactivity disorder: who benefits the most? PLoS ONE 2018;13(3):e0193681. [doi: 10.1371/journal.pone.0193681] [Medline: 29543891]
32. Espinoza J, Ruiz-Parra E, Uribe-Restrepo JM, et al. Computer-based cognitive training interventions for autism spectrum disorder: a systematic review. Front Psychiatry 2020;11:1310. [doi: 10.3389/fpsyt.2020.00722]

Abbreviations

ACE: adverse childhood experience
CITI: Collaborative Institutional Training Initiative
CRAFT: Catch Yourself, Relax, Assess, Focus, Think, Decide, and Act
REDCap: Research Electronic Data Capture
SEL: social-emotional learning

Edited by A Mavragani; submitted 16.10.24; peer-reviewed by JC Diperna, JM Mestre; revised version received 09.06.25; accepted 10.06.25; published 24.07.25.

Please cite as:
Liverman E, Antognoli D, Elaiho C, McGuire M, Stoltenburg A, Navarrete A, Bates G, Chelius T, Gundacker C, Lumelsky P, Currie B, Meurer Sr J
Game-Based Social-Emotional Learning for Youth: School-Based Qualitative Analysis of Brain Agents
JMIR Form Res 2025;9:e67550
URL: https://formative.jmir.org/2025/1/e67550
doi:10.2196/67550

© Elizabeth Liverman, David Antognoli, Cordelia Elaiho, Madison McGuire, Abbey Stoltenburg, Angel Navarrete, Garrett Bates, Thomas Chelius, Constance Gundacker, Paula Lumelsky, Brandon Currie, John Meurer Sr. Originally published in JMIR Formative Research (https://formative.jmir.org), 24.7.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Digital Infrared Thermographic Imaging for Limb Salvage in Patients at Risk of Amputation: Prospective Observational Study

Víctor Manuel Loza-González[1], MSc; Eleazar Samuel Kolosovas-Machuca[2], PhD; Patricia Aurea Cervantes-Báez[3], MD; José Luis Ramírez-GarcíaLuna[4,5], PhD; Edgar Guevara[2,5], PhD; Mario Aurelio Martínez-Jiménez[3], PhD

[1]Doctorado Institucional en Ingeniería y Ciencia de Materiales (DICIM-UASLP), Universidad Autónoma de San Luis Potosí, San Luis Potosí, Mexico

[2]Coordinación para la Innovación y Aplicación de la Ciencia y Tecnología, Universidad Autónoma de San Luis Potosí, Sierra Leona No. 550 Col. Lomas 2da. Sección, Planta Baja, San Luis Potosí, Mexico

[3]Burn Care Unit and Wound Care Center, Hospital Central Dr. Ignacio Morones Prieto, San Luis Potosí, Mexico

[4]Division of Experimental Surgery, Faculty of Medicine, McGill University, Montreal, QC, Canada

[5]Faculty of Science, Universidad Autónoma de San Luis Potosí, San Luis Potosí, Mexico

Corresponding Author:
Eleazar Samuel Kolosovas-Machuca, PhD
Coordinación para la Innovación y Aplicación de la Ciencia y Tecnología, Universidad Autónoma de San Luis Potosí, Sierra Leona No. 550 Col. Lomas 2da. Sección, Planta Baja, San Luis Potosí, Mexico

Abstract

Background: Scores and prediction models such as the mangled extremity severity score (MESS) for trauma and the Wound, Ischemia, and foot Infection (WIfI) classification for diabetic foot ulcers help in the decision-making process for amputation. However, these tools can be subjective as they depend on the experience of the medical staff applying them.

Objective: This study aimed to assess the impact of temperature measurement using infrared thermal imaging in extremity salvage in patients at risk of limb amputation.

Methods: We included 29 patients who sought a second opinion after an amputation recommendation. Infrared thermographic images were acquired to measure the temperature difference (ΔT) between the injured and uninjured limbs. For patients with salvaged limbs, we provided clinical follow-up for up to 12 weeks.

Results: Of the 29 patients enrolled in the study, 27 limbs were salvaged. Thermographic imaging allowed the distinction into two groups: the first group of 18 patients with mean ΔT value of −3.6 °C (SD 1.99 °C) , and the second group of 9 patients with mean ΔT of 3.36 °C (SD 2.71 °C). None of the patients in either group showed progression in ΔT values within the first 5 days; at the twelfth week, ΔT values approached 0 °C at wound closure. Of the two patients who required amputation, one showed an initial ΔT of −4.3 °C, which worsened to −5 °C by the fifth day, and the other patient showed an initial ΔT of −4.5 °C, which worsened to −5.8 °C by the fifth day.

Conclusions: Digital infrared thermography is a tool that may help guide limb salvage in patients with uncertain clinical diagnoses. This imaging modality allows visualization of thermal differences and patterns derived from thermal changes in patients at risk of limb amputation.

(JMIR Form Res 2025;9:e69072) doi:10.2196/69072

KEYWORDS

infrared thermography; amputation; diabetic foot; traumatic injuries; limb salvage

Introduction

Amputations are performed due to both traumatic and nontraumatic causes, including peripheral vascular disease, diabetes, and cancer. Diabetes and peripheral disease are the leading causes of nontraumatic amputations of the lower extremities [1,2]. Along with its high prevalence, the number of patients with diabetes mellitus is expected to increase to 700 million by 2045. It is estimated that 53% of diabetic patients are at risk of developing diabetic foot, with an augmented risk of amputation resulting from neuropathy, peripheral arterial disease, biomechanical deformity, limited articular mobility, and ulceration [1,3-6]. Following major amputation, the patients are at a higher mortality risk than those with some malignant diseases—up to 70% after 5 years—often dying due to cardiovascular disease [1,7]. While many classification systems are available, the WIfI classification was designed to assist with risk stratification of patients with diabetes. Despite its acceptance, its correlation with stage and clinical outcomes remains questionable, as it does not consider hemodynamic measures of limb ischemia [8].

XSL•FO
RenderX

Amputations related to accidents have decreased in frequency, but the decision to perform an amputation or reconstruct a lower extremity is a complex one for surgical teams. This is supported by the mangled extremity severity score (MESS), which evaluates acute limb ischemia, soft tissue damage, and physiological conditions. This score was developed over 32 years ago; however, medical advancements have raised questions about its current accuracy [9,10]. Hand injuries account for 12% of all traumatic injuries and 90% of amputations in the United States, with fingertip injuries being the most common cause. Despite this fact, the evaluation of patients with finger injuries is done on a case-by-case basis, and no formal clinical practice guidelines with treatment strategies are available for surgeons [11-15].

Body temperature has long been used as a natural indicator of diseases; it can be affected by conditions such as peripheral arterial disease and infections [4]. Medical infrared thermography can objectively measure body temperature changes related to the blood by detecting the long-wave infrared radiation emitted from the body. It creates temperature maps with abnormal regions that are detected and analyzed by the software, comparing regions of interest with the unaffected contralateral regions [16]. Based on the principle that any object with a temperature greater than 5 K emits infrared radiation, infrared thermography can detect heat derived from microcirculation flow 1 to 2 mm below the surface of the skin using a dense grid of sensory pixels [17,18]. It is noninvasive, fast, reliable, cost-effective, contactless, avoids exposure to ionizing radiation, and can produce multiple images within short time intervals without the discomfort of the patients [19,20]. The use of infrared thermography over the years has allowed scanning foot temperature in diabetic patients and evaluating the state of an injured limb to help decide if a patient's limb must be amputated [17]. It has also been used for screening fever, breast cancer, and diagnosis of vascular disorders by observing blood flow patterns, showing that the skin can reflect the status of the cardiovascular system. In cases of compromised small vessels and unaffected large vessels, thermography could be more effective than ultrasound-based techniques [21-23]. Recent advances have enabled thermography tools to be coupled with smartphones, facilitating the extended use of this technology [22].

Goal of This Study

Based on the thermal changes caused by blood flow and our previous study which predicted the treatment modalities in burn patients using thermal imaging, we aimed to use infrared thermography in patients at risk of amputation [21]. As wounds have heterogeneous thermographic patterns, measurements were taken from the fingers of the affected limbs and compared with the contralateral side, as finger evaluations have been done previously with devices that evaluate tissue oxygenation and limb blood flow [24]. By analyzing thermal changes between healthy and affected extremities, we proposed a linear regression model to help in the decision-making process by physicians treating such patients. The objective of this study was to determine the impact of digital infrared thermographic imaging for limb salvage in patients at risk of limb amputation by detecting temperature differences between the affected and healthy limbs.

Methods

Patients

This prospective cohort study included patients referred to the surgery department at Hospital Central "Dr. Ignacio Morones Prieto," a tertiary-level hospital in San Luis Potosí, Mexico, from another hospital or wound clinic for a second opinion regarding limb salvage between November 2023 to November 2024. Recruitment was performed during routine clinical consultations by the wound care team. Eligible patients were invited to participate, and informed consent was obtained before enrollment. All included patients had received recommendations for amputation, based on the clinical assessment of traumatic limb injuries or diabetic foot ulcers. Inclusion criteria were patients of any sex, older than 18 years, with a MESS of <7 or WIfI classification of ≤4. Exclusion criteria included patients with frank tissue nonviability, hemodynamic instability, severe wound infection, sepsis, or septic shock. Demographic variables such as the patients' gender and age were registered and information about the etiology of the lesion, patient comorbidities, anatomical site of the wound, and wound area were recorded.

Infrared Thermography

Infrared thermography scanning was performed using Flir T600 infrared camera (97070, FLIR System) in a room with a controlled temperature of 22 °C, atmosphere humidity of 40%, and under controlled light conditions. Scanning was performed from a distance of 0.5m at the patient's bedside. Before each scan, the dressings were removed, and after cleaning, the wounds were dried with sterile gauze. The affected extremities were left uncovered for 20 minutes before scanning The camera was powered on 10 minutes prior to image acquisition to allow for sensor stabilization. Thermographic images were taken following the protocol as proposed in the thermographic imaging in sports and exercise medicine (TISEM) checklist, with skin emissivity set at 0.98. A single expert in thermographic imaging interpreted the images using the FLIR Tools software (version 5.13.18031.2002; Teledyne FLIR) [17,25].

Scanning was conducted on the day 1 of medical evaluation, day 5 after hospital admission, and at the 12 weeks in outpatient follow-up, after complete wound healing. The regions of interest included toes or fingers of both the affected and healthy limbs. The difference in average temperatures (ΔT) of the compared areas between the affected limb and the contralateral healthy limb was calculated and used for the statistical analysis.

Statistical Analysis

Demographic data are expressed as the mean (SD) or proportions, as appropriate. Statistical analysis was performed using the R software (version 4.1.2; R Foundation for Statistical Computing, including the *R Commander* package (version 2.9 - 5). The Shapiro-Wilk test was used to verify the normality of the data distribution, and the nonparametric Friedman test to assess differences across the time points within each group. When the Friedman test indicated significant results, we

XSL•FO
RenderX

conducted post-hoc multiple comparisons using the Bonferroni correction to identify specific pairs of time points with statistically significant differences. Linear regression was used to compare the obtained ΔT and identify confounding factors such as sex, age, wound area, evolution time, day of wound closure, diagnosis, comorbidities, and clinical outcome. Wound healing in both groups was evaluated with Kaplan-Meier survival analysis. A *P* value <.05 was considered statistically significant.

Ethical Considerations

The study was conducted according to the guidelines of the Declaration of Helsinki and was approved by the Research Ethics Committee of the Hospital Central "Dr. Ignacio Morones Prieto" with Cofepris 17 CI 24 028 093 (protocol code 08 - 23) at the Department of Surgery and Wound Care. Informed consent was obtained from all subjects involved in the study. The consent form included use of clinical and thermographic imaging data. All data were anonymized prior to analysis no personal identifying information was included in the study database or in any publication materials. Participants did not receive financial or material compensation for their participation in the study. No identifying images of individual participants were included in the manuscript or supplementary material. All thermographic and clinical images were captured in a way that preserved patient anonymity.

Results

Demographic findings

Over a 12-month period, we included 29 patients with wounds referred for a second opinion regarding limb salvage, with a mean age of 42 (SD 18) years (Figure 1). Seventeen patients were male and 12 were female participants. Twelve patients were evaluated due to trauma, 14 had diabetic foot, 2 had Raynaud disease, and 1 had osteomyelitis. The average wound duration at the time of evaluation was 30 (SD 3) days.

Figure 1. Flowchart of patient inclusion and outcomes in a prospective cohort study of 29 patients at risk of limb amputation evaluated with infrared thermography at Hospital Central "Dr. Ignacio Morones Prieto," Mexico (Nov 2023-Nov 2024). Among them, 20 patients presented with a negative ΔT, with limb salvage achieved in 18 patients and amputation required in 2 patients.

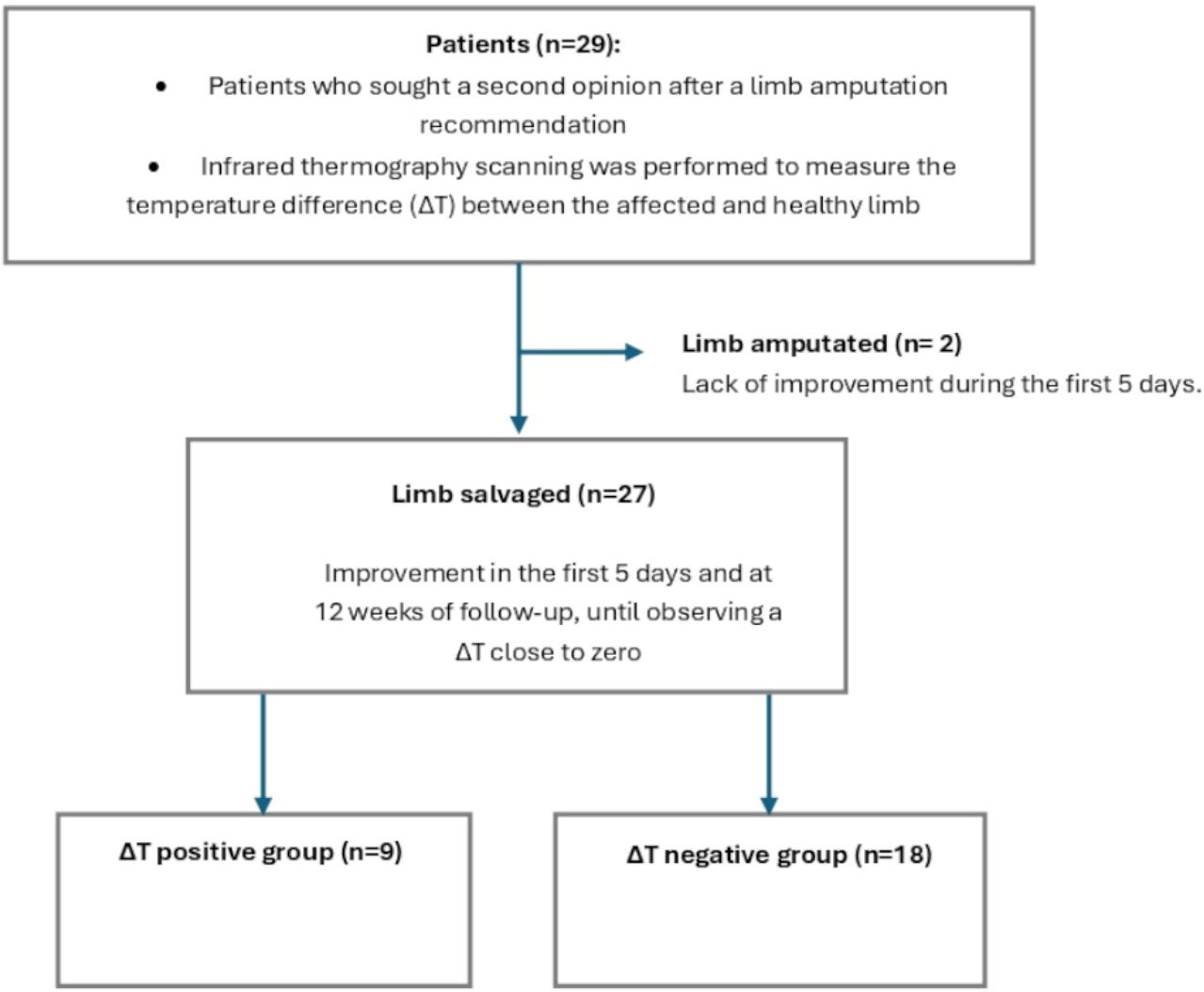

Of the 29 patients, 27 limbs were salvaged, and 2 underwent amputation (Table 1). A further subgroup analysis of the salvaged limbs revealed two distinct groups: one with a positive ΔT between the affected limb and healthy limb, and the other with negative ΔT. The Friedman test showed a significant difference across time points for both the positive ΔT group

XSL•FO
RenderX

(n=9; χ^2=10.89, *P*=.004) and the negative ΔT group (n=20; χ^2=32.4, *P*<.001) groups. Furthermore, the Mann-Whitney *U* test revealed significant differences between the median ΔT values of both groups on day 1 and day 84 of measurement (*P*<.01). The negative ΔT group presented statistically significant differences between day 1 and day 5, as well as between day 5 and day 84 (Table 2). The temperature difference between the affected limb and the healthy one over time tended toward zero in both groups. (Figure 2).

Table . Baseline demographic and clinical characteristics of patients at risk of limb amputation, stratified by ΔT[a] group (positive vs negative).

Variables	Patients (N=27), n (%)	ΔT[a] positive group (n=9), median (IQR)/n (%)	ΔT negative group (n=18), median (IQR)/n (%)	*P* value[b]
Age (years), n	42.3	42 (18 - 67)[c]	42.44 (18-68)[c]	.95[d]
Sex				.45[e]
Male	15 (55.5)	4 (14.8)[f]	11 (40.7)[f]	
Female	12 (44.4)	5 (18.5)[f]	7 (25.9)[f]	
Etiology				.39[g]
Diabetes	12 (44.4)	4 (14.8)[f]	8 (29.6)[f]	
Raynaud-disease	2 (7.4)	0	2 (7.4)[f]	
Osteomyelitis	1 (3.7)	1 (3.7)[f]	0	
Trauma	12 (44.4)	4 (14.8)[f]	8 (29.6)[f]	
Time since onset (days)				.03[e]
≤30	24 (88.8)	25 (21-30)[c]	9 (1-30)[c]	
>30	3 (11.1)	220 (60-320)[c]	–[h]	
Comorbidities	19 (70.3)	2 (7.4)[f]	17 (62.9)[f]	.98[g]
Affected region				.24[g]
Toe	7 (27)	2 (7.4)[f]	5 (18.5)[f]	
Finger	5 (18.5)	1 (3.7)[f]	4 (14.8)[f]	
Foot	9 (33.3)	2 (7.4)[f]	7 (25.9)[f]	
Hand	2 (7.4)	2 (7.4)[f]	0[f]	
Leg	4 (14)	2 (7.4)[f]	2 (7.4)[f]	
WIfI[i] classification, n (%)				.004[e]
2	1 (3.7)	0[f]	1 (3.7)[f]	
3	7 (25.9)	1 (3.7)[f]	6 (22.2)[f]	
4	4 (14.8)	0[f]	4 (14.8)[f]	

[a]ΔT: temperature difference.
[b]*P* values exhibit the comparison analysis between the group of preserved limbs and amputated limbs.
[c]median (IQR)
[d]Mann-Whitney *U* Test,
[e]Fisher Exact Test,
[f]n (%)
[g]χ^2 test.
[h]Not applicable.
[i]WIfI: Wound, ischemia, and foot infection.

XSL•FO
RenderX

Table . ΔT[a] values in patient ΔT positive and ΔT negative groups.

Time point 1 and Time point 2	Difference in temperature (ΔT), (°C), median (IQR)	*P* value[bc]
ΔT positive group		
Day 1 and Day 5	0.78 (0.35 to 1.91)	.30
Day 1 and Day 84	1.56 (0.43 to 2.68)	.003
Day 5 and Day 84	–0.35 (0.78 to 1.91)	.30
ΔT negative group		
Day 1 and Day 5	–1.66 (–0.90 to –0.14)	.01
Day 1 and Day 84	–2.56 (–1.80 to –1.04)	<.001
Day 5 and Day 84	–1.66 (–0.90 to –0.14)	.01

[a]ΔT: temperature difference.

[b]*P* values show the temperature differences between the ΔT positive and ΔT negative groups on day 1, day 5, and day 84 of evaluation.

[c] Post-hoc pairwise test: Wilcoxon-Mann-Whitney test and Bonferroni correction were used for multiple comparisons.

XSL•FO
RenderX

Figure 2. ΔT values of patients at risk of limb amputation, stratified by thermal asymmetry pattern: (A) ΔT-positive group and (B) ΔT-negative group. Each plot shows ΔT values across three time points (Day 1, Day 5, and Day 84). Asterisks indicate levels of statistical significance (*P < .05, **P < .01, ***P < .001). 20 patients presented with a negative ΔT, with limb salvage achieved in 18 cases and amputation required in 2.

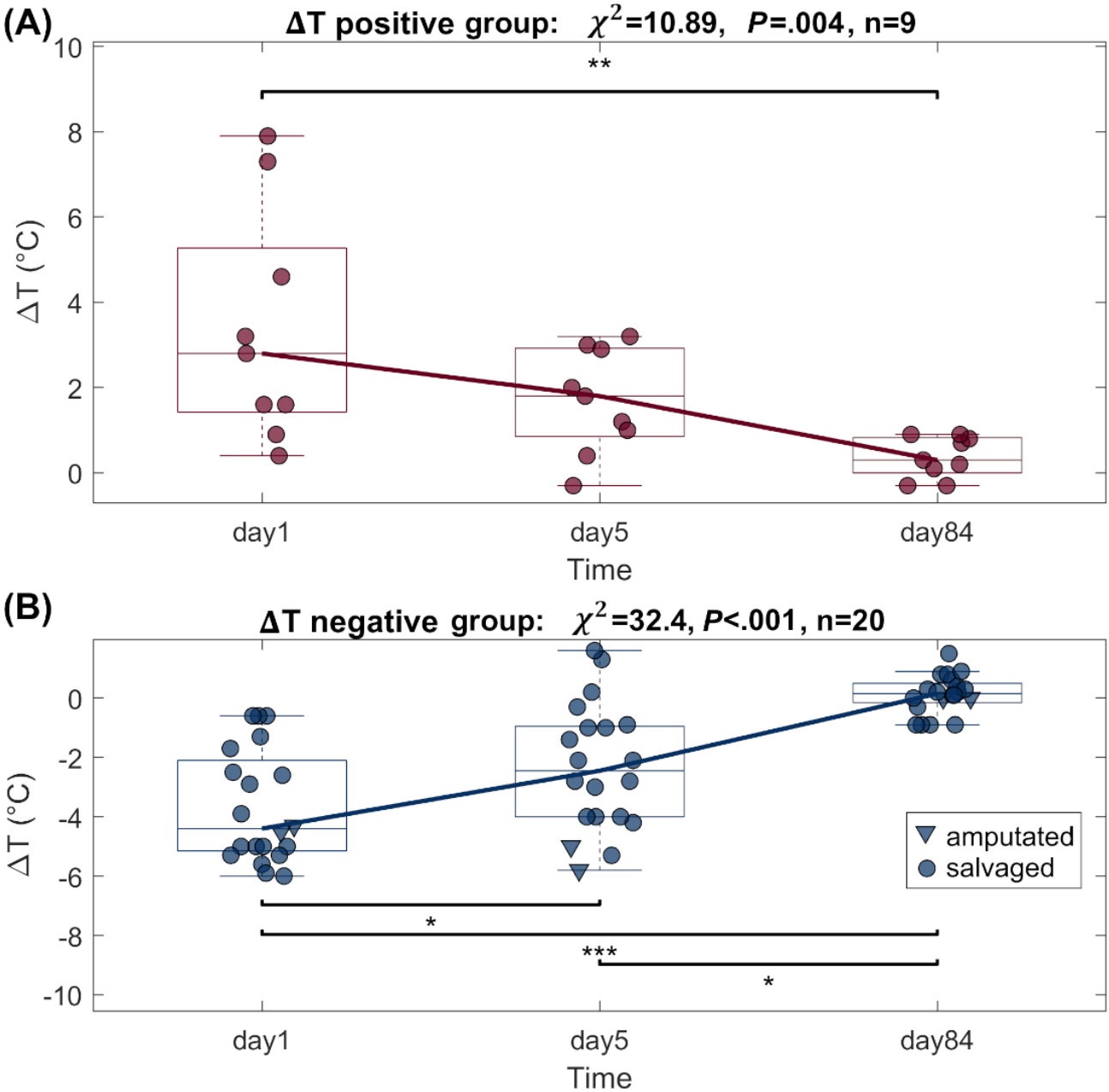

To evaluate the relationship between ΔT and time, linear regression was performed for each group (Figure 3). The regression graph for the positive ΔT group shows a decreasing trend in ΔT values over time, with the following equation:

$$y=-0.0274x+2.6293$$

In contrast, the negative ΔT group shows an increasing trend in ΔT values over time:

$$(y=0.0391x-3.1369)$$

The two patients who required amputation showed a greater decrease in negative ΔT between days 1 and days 5, in contrast to the other patients in the negative ΔT group, who showed an increased ΔT during the same period.

Kaplan-Meier analysis using the long-rank test showed no significant difference in survival distributions between the ΔT positive and ΔT negative groups ($P<.95$). Therefore, the survival distributions between both groups were similar (Figure 4).

XSL•FO
RenderX

Figure 3. Linear regression of temperature change (ΔT, °C) over time (days) in (A) the ΔT-positive group and (B) the ΔT-negative group. The data points are represented as circles; solid lines indicate the fitted regression lines, and shaded bands represent the 95% CI.

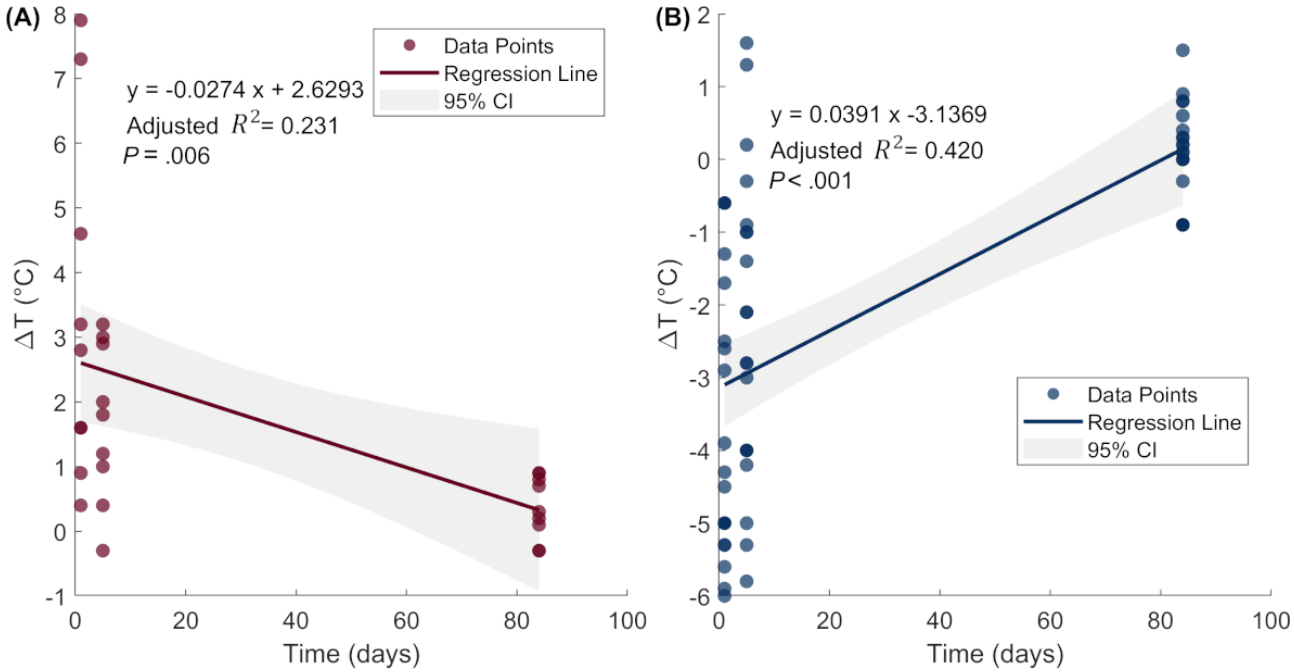

XSL•FO
RenderX

Figure 4. Difference in healing time between ΔT positive and ΔT negative groups using the Kaplan-Meier curve.

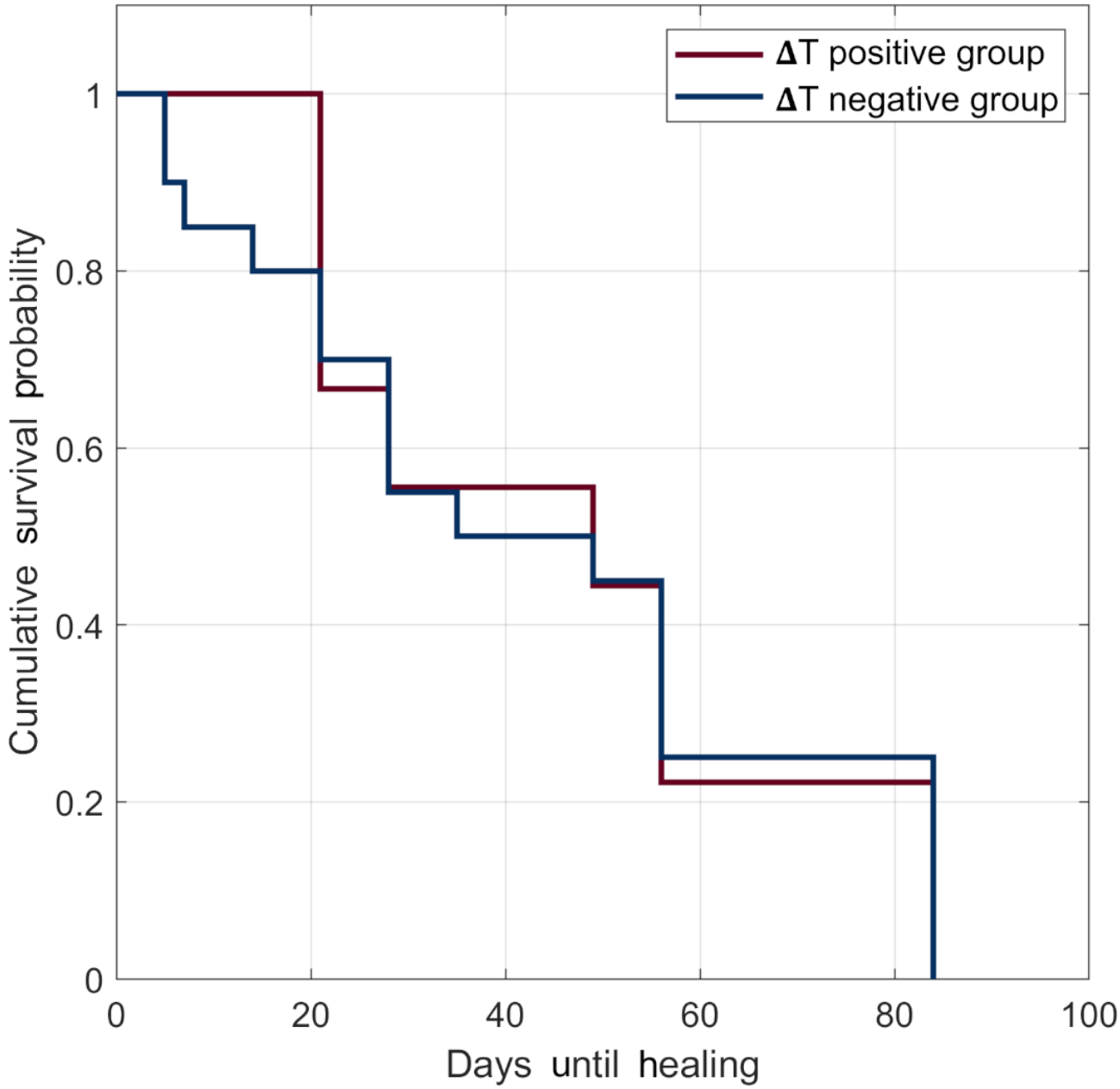

Discussion

Principal Results

This study shows the thermographic progression of patients at risk of amputation. It reveals two patterns of temperature, one with an increase and another with decrease in temperature in the regions of interest. The tendency of both groups was toward equalizing the temperature of the healthy limb. The two patients who required amputation showed an initial negative ΔT >5 °C, which worsened by the day 5. Two clinical vignettes, each showing a clinical case from one of the groups, are depicted in Figures 5 and 6.

XSL•FO
RenderX

Figure 5. Thermographic assessment of limb salvage. A 67-year-old female patient with a previous history of long-term diabetes and three previous episodes of diabetic foot presented to the Burn Care Unit and Wound Center. (A) She arrived with a 7-day-old wound on the first toe with local necrotic areas. (B) After evaluation, she received treatment with antibiotics, local debridement, and silver wound coverage. (C) Two weeks later the patient was discharged and periodically evaluated until wound closure during follow-up consultations. Areas with the lowest temperatures are displayed in black (D-F).

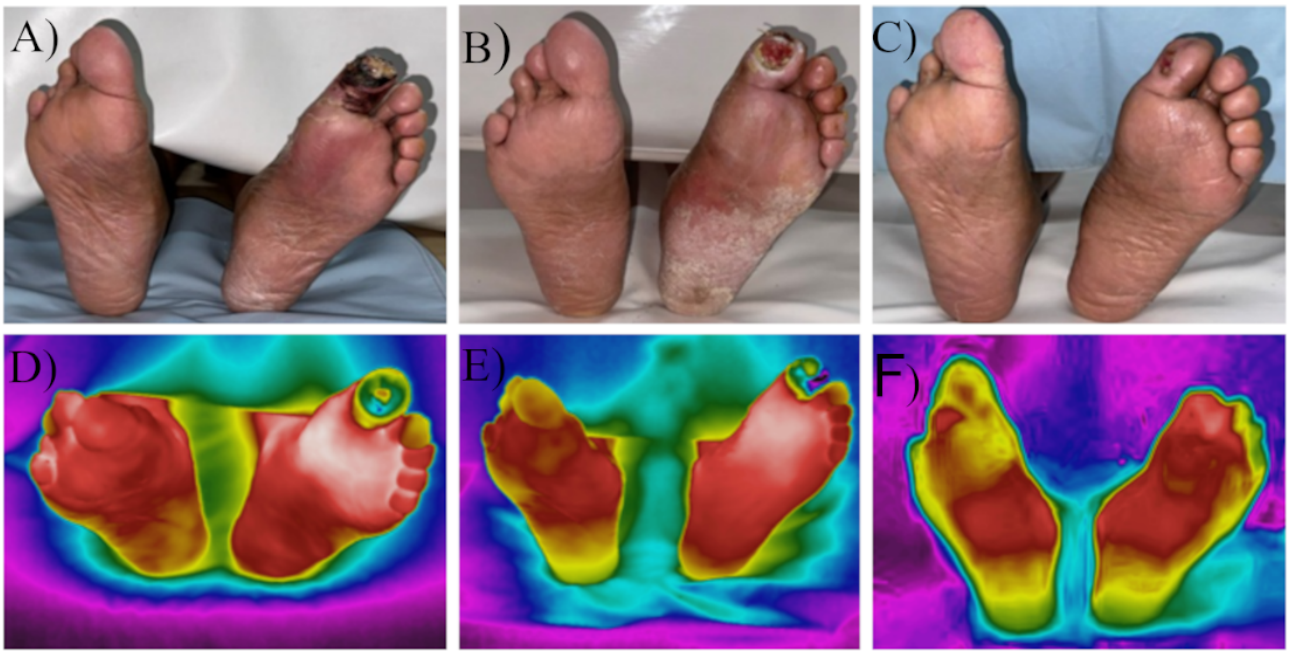

XSL•FO
RenderX

Figure 6. Thermographic assessment of limb salvage. (A) A 31-year-old male patient with a traffic injury on the right lower limb was evaluated after experiencing flap necrosis and wound infection. (B) The patient received treatment with antibiotics, progressive eschar debridement, and silver wound coverage. (C) Two months later, the patient was discharged and periodically evaluated until wound closure during follow-up consultations. Areas with the highest temperatures are displayed in white (D-F).

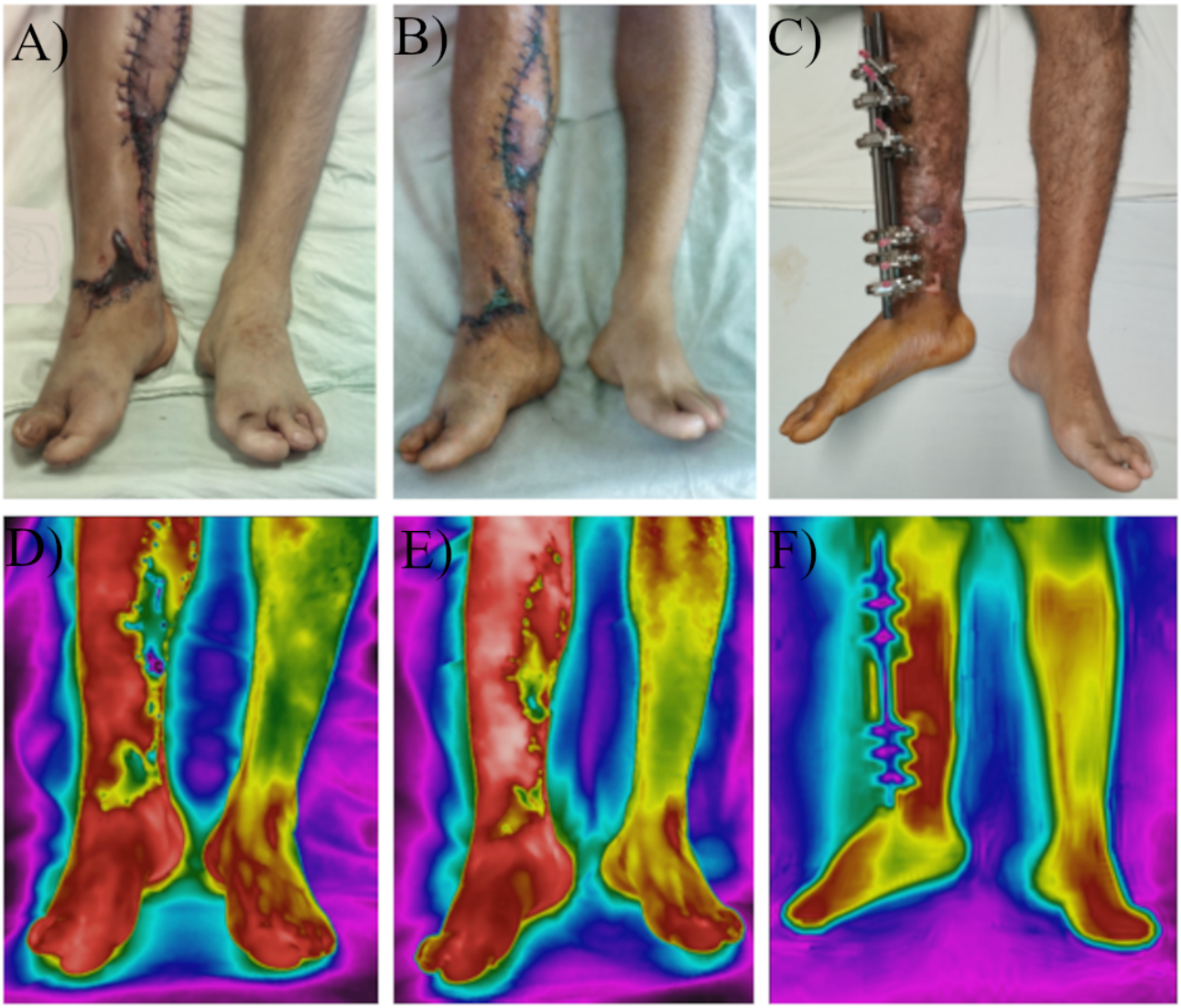

Previous studies have developed classification systems to support decision-making for amputations, as well as developed methods to stratify the risk of amputations in patients with ulcers. For example, the International Working Group on the Diabetic Foot criteria establishes the risk of ulceration in diabetic patients, while Hüsers et al [6] developed a tool to evaluate amputation risk based on data from the PEDIS (perfusion, extent, depth, infection and sensation) classification for diabetic foot. Prognostic systems including prediction models, offer better assessments than scoring systems by indicating the probability of amputation to physicians. Therefore, adding additional thermographic data to a model could help to improve the actual models. Due to improvements in modern cameras, the production of high-resolution thermal images has allowed the use of thermographic imaging in the detection of diseases related to abnormalities in blood flow such as arterial occlusive disease [1,6].

Limitations

The sample size and its single-center design are among the main limitations of this work. glycosylated hemoglobin levels, prior interventions, infection severity, and vascular status were not included as variables in the statistical analysis, which may have limited the interpretation of data. The decision to perform an amputation depends on factors beyond temperature measurements. Further studies are required to collect data from a larger cohort of nonamputated patients throughout their clinical evolution, including functional assessment of the affected limb in terms of range of motion, strength, and sensitivity. Additionally, more data is required regarding the observed decrease in temperature values among patients whose limbs were amputated. A longer follow-up study of patients with longer healing times is necessary to determine whether prolonged healing is associated with a reduced life expectancy and poorer prognosis caused by impairment in other organ systems [12,14,15].

Comparison With Prior Work

Studies on skin microcirculation have established that infrared thermography is equivalent to other methods such as laser Doppler flowmetry, transcutaneous oxygen measurement, and capillaroscopy [4,19,26]. Arteriography is the gold standard

XSL•FO
RenderX

method to evaluate blood vessels but is not routinely used due to its invasive nature and the risks associated with radiation and contrast, limiting its application in surgical planning for patients with ischemia [20]. In contrast, the current trend in medicine is to offer personalized and evidence-based care, performed using intravital noninvasive imaging technologies such as thermography. These tools provide valuable data to interdisciplinary teams to enhance the quality of life for patients and providing offloading systems for public health [27]. In future studies, it will be necessary to add more variables to the model such as critical limb ischemia, due to its association with diabetes mellitus and renal insufficiency [28]. Establishing the correlation between temperature, vascular flow, and tissular oxygen impairment requires the incorporation of other techniques such as finger plethysmography, micromyography, noninvasive tissular oxygen measurement, and laser Doppler flowmetry [29]. In the future, thermal data could be used to evaluate the limbs at risk by complementing the existing scores, which is a necessary addition considering that the rates of diabetes mellitus will continue to increase in the future, leading to increased costs and associated complications such as ulcerations, amputations, and deaths; notably 85% of the patients with a lower limb amputation had an ulcer and 47.9% of amputated patients are die within a year [1,30]. Early identification of the need for amputation can reduce these complications, reduce hospital stay and follow-up visits[2]. The global use of infrared thermography as a complementary tool is a feasible approach; it can be implemented not only in cities and large hospitals but also in remote areas and small clinics, as the use of thermal cameras by coupling with smartphones is easy, cheap and requires short-time capacitation sessions for health professionals and students. These results can be interpreted remotely using telemedicine, and can improve the use of resources, referral to a multicenter hospitals, and the medical teams' response times for limb salvage [23,31].

Conclusions

Limb salvage was performed using infrared thermography in patients at risk of limb amputation. This method allows comparisons of the thermal difference between the affected and nonaffected contralateral limbs, showing the thermal changes occurring within the treated limb over time. Further studies with more patients should be performed to provide insights into the clinical and progression-related differences between these group of patients.

Acknowledgments

The authors acknowledge the Universidad Autónoma de San Luis Potosí for providing the thermographic camera used in this research, as well as the chief of the Surgery Department, Dr. Fidel Francisco Bear Morales, for allowing its use. This research received no external funding.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Authors' Contributions

Conceptualization: MAM-J
Data curation: VML-G
Formal analysis: EG (lead), VML-G (supporting)
Investigation: VML-G (lead), PAC-B (supporting)
Methodology: VML-G (lead), JLR-G (equal)
Supervision: ESK-M (lead), MAM-J (equal)
Validation: JLR-G (lead), EG (equal), MAM-J (equal)
Writing-original draft: VML-G
Writing-review and editing: ESK-M (lead), PAC-B (supporting)

Conflicts of Interest

None declared.

Checklist 1
STROBE (Strengthening the Reporting of Observational studies in Epidemiology) checklist
[PDF File, 90 KB - formative_v9i1e69072_app1.pdf]

References

1. Maldonado-Valer T, Pareja-Mujica LF, Corcuera-Ciudad R, et al. Prevalence of diabetic foot at risk of ulcer development and its components stratification according to the International Working Group on the Diabetic Foot (IWGDF): a systematic review with metanalysis. PLoS ONE 2023;18(11):e0284054. [doi: 10.1371/journal.pone.0284054] [Medline: 38015974]
2. Yao PF, Diao YD, McMullen EP, Manka M, Murphy J, Lin C. Predicting amputation using machine learning: A systematic review. PLOS ONE 2023;18(11):e0293684. [doi: 10.1371/journal.pone.0293684] [Medline: 37934767]

XSL•FO
RenderX

3. Falanga V, Isseroff RR, Soulika AM, et al. Chronic wounds. Nat Rev Dis Primers 2022 Jul 21;8(1):50. [doi: 10.1038/s41572-022-00377-3] [Medline: 35864102]
4. Gauci J, Falzon O, Formosa C, et al. Automated region extraction from thermal images for peripheral vascular disease monitoring. J Healthc Eng 2018;2018:5092064. [doi: 10.1155/2018/5092064] [Medline: 30651943]
5. González-Pérez S, Perea Ström D, Arteaga-Marrero N, et al. Assessment of registration methods for thermal infrared and visible images for diabetic foot monitoring. Sensors (Basel) 2021 Mar 24;21(7):2264. [doi: 10.3390/s21072264] [Medline: 33804926]
6. Hüsers J, Hafer G, Heggemann J, Wiemeyer S, John SM, Hübner U. Predicting the amputation risk for patients with diabetic foot ulceration - a Bayesian decision support tool. BMC Med Inform Decis Mak 2020 Aug 24;20(1):200. [doi: 10.1186/s12911-020-01195-x] [Medline: 32838777]
7. Brix ATH, Rubin KH, Nymark T, Schmal H, Lindberg-Larsen M. Mortality after major lower extremity amputation and association with index level: a cohort study based on 11,205 first-time amputations from nationwide Danish databases. Acta Orthop 2024 Jun 19;95:358-363. [doi: 10.2340/17453674.2024.40996] [Medline: 38895969]
8. Mathioudakis N, Hicks CW, Canner JK, et al. The Society for Vascular Surgery Wound, Ischemia, and foot Infection (WIfI) classification system predicts wound healing but not major amputation in patients with diabetic foot ulcers treated in a multidisciplinary setting. J Vasc Surg 2017 Jun;65(6):1698-1705. [doi: 10.1016/j.jvs.2016.12.123] [Medline: 28274750]
9. Stern JR, Wong CK, Yerovinkina M, et al. A meta-analysis of long-term mortality and associated risk factors following lower extremity amputation. Ann Vasc Surg 2017 Jul;42:322-327. [doi: 10.1016/j.avsg.2016.12.015] [Medline: 28389295]
10. Khan S, Hawkins BM. Acute limb ischemia interventions. Interv Cardiol Clin 2020 Apr;9(2):221-228. [doi: 10.1016/j.iccl.2019.12.002] [Medline: 32147122]
11. Loja MN, Sammann A, DuBose J, et al. The mangled extremity score and amputation: Time for a revision. J Trauma Acute Care Surg 2017 Mar;82(3):518-523. [doi: 10.1097/TA.0000000000001339] [Medline: 28030489]
12. Billig JI, Nasser JS, Cho HE, Chou CH, Chung KC. Association of interfacility transfer and patient and hospital characteristics with thumb replantation after traumatic amputation. JAMA Netw Open 2021 Feb 1;4(2):e2036297. [doi: 10.1001/jamanetworkopen.2020.36297] [Medline: 33533928]
13. Pastor T, Hermann P, Haug L, Gueorguiev B, Pastor T, Vögelin E. Semi-occlusive dressing therapy versus surgical treatment in fingertip amputation injuries: a clinical study. Eur J Trauma Emerg Surg 2023 Jun;49(3):1441-1447. [doi: 10.1007/s00068-022-02193-6]
14. Speth KA, Yoon AP, Wang L, Chung KC, FRANCHISE Group. Assessment of tree-based statistical learning to estimate optimal personalized treatment decision rules for traumatic finger amputations. JAMA Netw Open 2020 Feb 5;3(2):e1921626. [doi: 10.1001/jamanetworkopen.2019.21626] [Medline: 32083690]
15. Yoon AP, Mahajani T, Hutton DW, Chung KC, Finger Replantation and Amputation Challenges in Assessing Impairment, Satisfaction, and Effectiveness (FRANCHISE) Group. Cost-effectiveness of finger replantation compared with revision amputation. JAMA Netw Open 2019 Dec 2;2(12):e1916509. [doi: 10.1001/jamanetworkopen.2019.16509] [Medline: 31790567]
16. Martin-Playa P, Foo A. Approach to fingertip injuries. Clin Plast Surg 2019 Jul;46(3):275-283. [doi: 10.1016/j.cps.2019.02.001] [Medline: 31103072]
17. Moreira DG, Costello JT, Brito CJ, et al. Thermographic imaging in sports and exercise medicine: A Delphi study and consensus statement on the measurement of human skin temperature. J Therm Biol 2017 Oct;69:155-162. [doi: 10.1016/j.jtherbio.2017.07.006] [Medline: 29037377]
18. Cooke WH, Moralez G, Barrera CR, Cox P. Digital infrared thermographic imaging for remote assessment of traumatic injury. J Appl Physiol (1985) 2011 Dec;111(6):1813-1818. [doi: 10.1152/japplphysiol.00726.2011] [Medline: 21940851]
19. Grech J, Mizzi S, Falzon O. A technical review of foot temperature measurement systems. J Am Podiatr Med Assoc 2022 Apr 27;112(2):21-065. [doi: 10.7547/21-065] [Medline: 35239503]
20. Antônio de Carvalho Abreu J, de Oliveira RA, Martin AA. Correlation between ankle-brachial index and thermography measurements in patients with peripheral arterial disease. Vascular 2022 Feb;30(1):88-96. [doi: 10.1177/1708538121996573] [Medline: 33752532]
21. Martínez-Jiménez MA, Ramirez-GarciaLuna JL, Kolosovas-Machuca ES, Drager J, González FJ. Development and validation of an algorithm to predict the treatment modality of burn wounds using thermographic scans: prospective cohort study. PLoS One 2018;13(11):e0206477. [doi: 10.1371/journal.pone.0206477] [Medline: 30427892]
22. Passos MDD, Cordeiro AH, Tavares SDC, Alves LM, Godoy-Gomes I, da Rocha AF. Acute limb ischemia after self-injection of crushed morphine tablets into the radial artery: role of infrared thermography in the assessment of tissue perfusion. Am J Case Rep 2022 Mar 5;23:e935336. [doi: 10.12659/AJCR.935336] [Medline: 35246501]
23. Fraiwan L, AlKhodari M, Ninan J, Mustafa B, Saleh A, Ghazal M. Diabetic foot ulcer mobile detection system using smart phone thermal camera: a feasibility study. Biomed Eng Online 2017 Oct 3;16(1):117. [doi: 10.1186/s12938-017-0408-x] [Medline: 28974212]
24. Harford M, Areia C, Villarroel M, et al. Study protocol for an exploratory interventional study investigating the feasibility of video-based non-contact physiological monitoring in healthy volunteers by Mapping Of Lower Limb skIn pErfusion (MOLLIE). BMJ Open 2020 Jun 11;10(6):e036235. [doi: 10.1136/bmjopen-2019-036235] [Medline: 32532774]

XSL•FO
RenderX

25. Ammer K. The Glamorgan Protocol for recording and evaluation of thermal images of the human body. Thermol Int 2008 Oct 1;18:125-129 [FREE Full text]
26. Kesztyüs D, Brucher S, Wilson C, Kesztyüs T. Use of infrared thermography in medical diagnosis, screening, and disease monitoring: a scoping review. Medicina (Kaunas) 2023 Dec 9;59(12):2139. [doi: 10.3390/medicina59122139] [Medline: 38138242]
27. Kayama T, Sano M, Inuzuka K, et al. A pilot study investigating the use of regional oxygen saturation as a predictor of ischemic wound healing outcome after endovascular treatment in patients with chronic limb-threatening ischemia. Ann Vasc Dis 2021 Mar 25;14(1):23-30. [doi: 10.3400/avd.oa.20-00132] [Medline: 33786096]
28. Masaki S, Kawamoto T. Fingertip amputation injury of Allen Type III managed conservatively with moist wound dressings. Am J Case Rep 2021 Feb 23;22:e928950. [doi: 10.12659/AJCR.928950] [Medline: 33621217]
29. Chalise U, Becirovic-Agic M, Lindsey ML. The cardiac wound healing response to myocardial infarction. WIREs Mech Dis 2023 Jan;15(1):e1584. [doi: 10.1002/wsbm.1584] [Medline: 36634913]
30. Ramirez-GarciaLuna JL, Bartlett R, Arriaga-Caballero JE, Fraser RDJ, Saiko G. Infrared thermography in wound care, surgery, and sports medicine: a review. Front Physiol 2022;13:838528. [doi: 10.3389/fphys.2022.838528] [Medline: 35309080]
31. Piva G, Crepaldi A, Zenunaj G, et al. The value of infrared thermography to assess foot and limb perfusion in relation to medical, surgical, exercise or pharmacological interventions in peripheral artery disease: a systematic review. Diagnostics (Basel) 2022 Dec 1;12(12):3007. [doi: 10.3390/diagnostics12123007] [Medline: 36553014]

Abbreviations

MESS: Mangled Extremity Severity Score
WIfI: Wound, ischemia, and foot infection
ΔT: Difference of temperature between two areas of interest

Edited by A Mavragani; submitted 11.12.24; peer-reviewed by YG Ang; revised version received 11.05.25; accepted 20.05.25; published 21.07.25.

Please cite as:
Loza-González VM, Kolosovas-Machuca ES, Cervantes-Báez PA, Ramírez-GarcíaLuna JL, Guevara E, Martínez-Jiménez MA Digital Infrared Thermographic Imaging for Limb Salvage in Patients at Risk of Amputation: Prospective Observational Study JMIR Form Res 2025;9:e69072
URL: https://formative.jmir.org/2025/1/e69072
doi:10.2196/69072

© Víctor Manuel Loza-González, Eleazar Samuel Kolosovas-Machuca, Patricia Aurea Cervantes-Báez, José Luis Ramírez-GarcíaLuna, Edgar Guevara, Mario Aurelio Martínez-Jiménez. Originally published in JMIR Formative Research (https://formative.jmir.org), 21.7.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

GamePlan4Care, a Web-Based Adaptation of the Resources for Enhancing Alzheimer's Caregiver Health II Intervention for Family Caregivers of Persons Living With Dementia: Formative, Qualitative Usability Testing Study

Jinmyoung Cho[1,2], PhD; Thomas Birchfield[1], MS; Jennifer L Thorud[1], MPH; Marcia G Ory[3], PhD; Alan B Stevens[1,3], PhD

[1]Center for Applied Health Research, Baylor Scott & White Research Institute, 2401 S 31st St, Temple, TX, United States

[2]Saint Louis University School of Medicine, St. Louis, MO, United States

[3]Texas A&M School of Public Health, College Station, TX, United States

Corresponding Author:
Alan B Stevens, PhD
Center for Applied Health Research, Baylor Scott & White Research Institute, 2401 S 31st St, Temple, TX, United States

Abstract

Background: The negative consequences of caregiving can be mitigated by providing caregivers with support programs that increase their dementia care skills and provide emotional and tangible support. Web-based technology can increase the availability of evidence-based caregiver interventions. GamePlan4Care (GP4C) is a web-based adaptation of the Resources for Enhancing Alzheimer's Caregiver Health II (REACH II) intervention, redesigned and reformatted for web-based delivery.

Objective: The goal of GP4C is to create a web-based family caregiver support platform that facilitates self-directed exposure to evidence-based skills training and support for caregivers of persons living with dementia. This multidimensional approach of using technology enhanced with live support has the potential for improved scalability and sustainability. In preparation for a randomized clinical trial of the new intervention, the GP4C platform underwent user interface/user experience (UI/UX) testing with caregivers as part of an iterative design process.

Methods: UI/UX testing of caregivers' reactions to technical and content-related aspects of the platform was conducted with 31 caregivers recruited through partnerships with community-based organizations in central Texas. Usability testing consisted of performing system tasks, answering open-ended questions on the tasks, and providing feedback on their experience with the platform. Two researchers used an inductive thematic approach to data analysis using transcripts of individual audio and screen-recorded sessions with each participant. The analysis consisted of 3 phases: data familiarization, coding, and theme formulation.

Results: In total, 18 participants tested technical-related aspects of the GP4C platform, and 13 participants tested content-related aspects. The average age of participants was 62 (SD 12.2, range 31‐86). A majority of participants were female (27/31, 87.1%) and White or Caucasian (26/31, 83.1%) while almost one-third were Hispanic (10/31, 32.3%). The thematic analysis revealed 3 themes: supportive resources as a common theme, active engagement for technical aspects of the platform, and a comprehensive approach for content aspects of the platform. Participants also suggested changes in navigation and content.

Conclusions: Findings from the usability testing sessions indicate that the platform provided engaging, useful content that the caregiver would continue to use, resonated with their caregiving experience, helped the caregivers think through their choices and emotions, and could be used to help communicate with the person living with dementia. Caregivers appreciated the personalization based on what they had already completed and the concept of having a Dementia Care Navigator when they needed additional help. Caregivers also provided multiple suggestions on how to improve the system, including changes for easier navigation and inclusiveness. This positive feedback indicates that with a few changes, the platform would be beneficial to meet the needs and provide resources for caregivers of persons living with dementia. The process of involving end users in usability testing during the development stage ensures that the finished tool will better meet users' expectations and current needs.

(JMIR Form Res 2025;9:e60143) doi:10.2196/60143

KEYWORDS

dementia caregivers; user interface/user experience (UI/UX); technology; web-based caregiving program; user experience; training; online resource; peer support; Alzheimer disease; dementia; caregiving; caregiver; carer; informal care; family care;

XSL•FO
RenderX

spousal care; usability; clinical trial; aging; elderly; geriatrics; gerontology; mhealth; mobile health; digital health; online platforms; web health; online support; neurodegeneration; brain degeneration; memory loss

Introduction

Background

Communities across the United States are challenged to meet the physical, emotional, and social needs of persons living with dementia. Currently, nearly 7 million Americans are living with dementia, and that number is expected to almost double in the next 25 years [1]. While efforts to increase the availability of formal health and social care services are essential, support of family caregivers is paramount if they are expected to continue to provide the vast majority of daily care and support of persons living with dementia. As reported in the National Academies' report, Families Caring for an Aging America, nearly 70% of caregivers in a nationally representative survey had provided 2 to 10 years of care, and 15% had already provided care for more than 10 years at the time of the survey [2]. In other words, an average person in their fourth decade of life is expected to eventually spend 5.0 years (or 10% of their remaining life) caring for an older adult, a situation known to have negative consequences [2]. This is significant because caregivers are more likely than noncaregivers to experience negative consequences on their well-being such as increased depression, anxiety, burden, social isolation, and family conflict [3-5].

The negative consequences of caregiving, however, can be mitigated by engaging caregivers in programs that increase their dementia care skills while providing emotional and tangible support. A recent systematic review shows that caregivers are more likely to engage in and benefit from multicomponent interventions that assess the caregiver's challenges as well as their emotional response to providing care [6]. These considerations respect the unique circumstances of each family caregiving situation and lead to a tailored program of education, skills training, and support [7-9]. While caregiver interventions that use a tailored approach to education, skills training, and support exist, services derived from these evidence-based interventions are not routinely provided by health and social support organizations. Two major barriers to the provision of such services include the cost of developing a workforce trained in evidence-informed dementia care strategies and the lack of formal payment systems (eg, commercial insurance and Centers for Medicare and Medicaid Services) [10-12].

Web-based technology provides an alternative approach for moving evidence-based interventions into widely available caregiving education, skills training, and support programs. Web-based caregiver support programs have the potential to increase the availability of caregiver interventions that have previously been delivered through an in-person or telephone format [13,14]. Moreover, web-based technology can be designed to meet the expressed desire of caregivers for an on-demand, self-paced, learning and skills training experience to support their role as a dementia caregiver [15]. Use of web-based technology in the provision of existing evidence-based interventions is supported by research on caregiver interventions designed for web-based or mobile delivery [16-19].

Multiple meta-analyses of caregiver interventions designed for internet delivery have shown promise, but findings have been mixed [13,20-22]. Some reviews found that technology-based interventions were just as effective as face-to-face interventions; however, mixed delivery methods showed greater improvements than web-based, telephone, or DVD-based interventions alone [21]. Leng et al [13] found that personalized internet-based interventions had a greater effect size than nonpersonalized interventions on depressive symptoms and perceived stress. These findings suggest the need for more attention to the engagement of family caregivers in the design and user testing of web-based platforms for dementia caregivers.

To address this need, the research team created a web-based family caregiver support platform, GamePlan4Care (GP4C), by applying technology to an existing evidence-based intervention (ie, Resources for Enhancing Alzheimer's Caregiver Health [REACH II]). This study explored the usability of GP4C among family caregivers of persons living with dementia.

GP4C: A Web-Based Adaptation of REACH II

GP4C incorporates a risk-based approach and aligns the self-reported needs of a caregiver with therapeutic content consistent with the REACH II domains of safety, stress, health, emotions, care services, support, and behaviors. Similar to REACH II, the GP4C therapeutic strategies include educational materials, tools (eg, worksheets), skills training exercises, and access to peer support. Importantly, the web-based format of GP4C allows strategies to also be presented in video format, which was not done in REACH II. The therapeutic process is facilitated by a Dementia Care Navigator, a trained interventionist who assists the caregiver in the navigation of the web-based content and personalized goal setting via telephone and email support.

After an initial build of the platform to include the functionality of an initial assessment of the caregiving situation followed by access to the safety and emotions domains of the GP4C content, we conducted caregiver usability testing on these domains to identify additional design needs in preparation for a randomized trial. This manuscript describes the caregivers' experience in using the initial design of the GP4C platform and key aspects of how they experienced the functionality of the platform regarding the initial assessment and tailored education and skills training on the 2 domains of GP4C: Safety and Emotions.

Methods

Process

The GP4C research team worked with a large integrated health care system's Digital Health Department to develop a task hierarchy cataloging individual tasks to accomplish within the platform and salient design questions appropriate for experimenter prompts. Using phenomenological methods, the research team conducted usability testing concurrently with content enhancement via a method of iterative evaluation using the think-aloud technique [23]. This method characterizes the

XSL•FO
RenderX

ease with which a user can complete a task, by what means a user attains mastery of system features, and problems a user encounters while using the system. Such tasks included registering, logging in, answering user assessments, and navigating to education and skill-building content (Textbox 1). Participants tested either the technical-related aspects (user interface and design) or content-related aspects (wording or appropriateness of questions and feedback, satisfaction with education and skill-building content) of the GP4C platform. User testing sessions for each participant were audio and screen recorded. Participants performed system tasks while vocalizing their thoughts, feelings, and satisfaction with the platform. Research staff prompted the participant with a question to elicit specific feedback regarding that feature. Upon completion, participants also provided their overall impression or opinions regarding their experience with the platform. Participants were not given access to a Dementia Care Navigator as part of this user interface/user experience (UI/UX) testing.

Textbox 1. Summary of tasks that were tested by the participant, a caregiver of a person living with dementia, during GamePlan4Care usability testing.

Technical-related aspects

- Logging into the platform
- Responding to questions regarding the context of caregiving
- Managing messaging feature
- Reviewing automated feedback
- Assessing video content
- Managing goal progress

Content-related aspects

- Logging into the platform
- Responding to questions that assessed caregiving risk associated with the domain
- Reviewing two core skill videos
- Reviewing three related exercise videos
- Reviewing materials, including worksheets

Recruitment

Participants for the usability testing were recruited through referrals from local partnerships with community-based organizations (such as Area Agencies on Aging and Alzheimer's focused not-for-profit organizations) from October 2019 to February 2020. Program champions at our partner organization recruitment sites identified interested participants and sent referrals through a secure Research Electronic Data Capture (REDCap; Vanderbilt University) website after receiving verbal consent by the interested participant. Upon receipt of a referral, research staff reached the interested participant by telephone to complete eligibility screening. If eligible, interested participants scheduled a usability testing session at a recruitment site. Informed consent was conducted electronically at the time of the usability testing session immediately before the intake assessment and usability testing occurred. Interested participants received information about the study from the program champion prior to verbal consent to be referred and from research staff during the eligibility screen telephone conversation.

Participants

Eligibility criteria for this study was for caregivers to be 18 years of age or older, reside in our recruitment area, be able to read and speak English, provide care or supervision to a friend or family member for an average of at least 8 hours per week for the last 6 months, and have access to a home computer or tablet with internet access and use it for an average of at least 3 times per week. The care recipient, or person receiving care, must also have a diagnosis of Alzheimer disease or a related dementia (self-reported diagnosis by the caregiver was accepted) and have signs or symptoms of dementia (defined as a score of 2 or greater on the Ascertain Dementia 8, [24]). Criteria that excluded caregivers from participating included being currently enrolled in another evidence-based caregiver education and support intervention.

Program champions referred 113 individuals to the program, of which 102 were distinct, unduplicated referrals. Of these, 61 individuals were screened for eligibility by the research team and 41 were not screened (1 care recipient passed away, 29 were unable to be reached, and 10 were not interested, had time constraints, or were unable to travel for the testing session). Forty individuals were eligible for the intervention and 21 were ineligible. Reasons for being ineligible included the care recipient not having a dementia diagnosis (n=9), participating in another caregiving intervention (n=5), living outside of the recruitment area (n=2), not having access to or using a home computer (n=2), and not providing care for at least 8 hours per week (n=3). Reasons for not enrolling include time constraints (n=2) and other reasons (n=7). A total of 31 caregiver participants completed the sessions, 18 were recruited to complete user testing of the technical-related aspects of the platform and then an additional 13 were recruited for testing of the content-related aspects.

Tables 1 and 2 summarize the demographic characteristics and caregiving experience of participants. A majority of participants

XSL•FO
RenderX

were female (27/31, 87.1%) and White or Caucasian (26/31, 83.9%). The average age of participants was 62 years old (range 31 - 86). Nearly a third of participants were Hispanic (10/31, 32.3%). Caregivers were providing care to a variety of family relations with a majority being parents (-in-law) or spouses (26/31 participants, 83.9%).

Table . Demographic characteristics of GP4C[a] usability testing participants (n=31, caregivers of persons living with dementia) recruited from October 2019-February 2020.

Participant characteristics	Values
Age, mean (SD)	62.0 (12.2)
Sex, n (%)	
Female	27 (87.1)
Male	4 (12.9)
Race, n (%)	
White/Caucasian	26 (83.9)
Black or African American	3 (9.7)
Other or more than 1 race	2 (6.5)
Hispanic ethnicity	10 (32.3)
Marital status, n (%)	
Married or living as married	20 (64.5)
Widowed	2 (6.5)
Divorced or separated or never married	9 (29.1)
Employment, n (%)	
Full-time	8 (25.8)
Part-time	2 (6.5)
Homemaker	2 (6.5)
Retired	13 (41.9)
Other	6 (19.3)
Relationship to care recipients, n (%)	
Parents (-in-law)	15 (48.4)
Spouses	11 (35.5)
Brother, grandfather, or friend	5 (16.1)

[a]GP4C: GamePlan4Care.

XSL•FO
RenderX

Table . Caregiving experience of GP4C[a] usability testing participants (n=31, caregivers of persons living with dementia) recruited from October 2019–February 2020.

Caregiving characteristics	Values, n (%)
CR[b] diagnosis	
Alzheimer disease	13 (41.9)
Dementia	15 (48.4)
Vascular dementia or TIA[c]	4 (12.9)
Parkinson's disease	2 (6.5)
Lewy-body disease/Fronto-temporal disease	5 (16.1)
MCI[d]	2 (6.5)
Distance to CR	
Living with CR	4 (12.9)
Less than 15 min	22 (71.0)
Within 15 - 30 min	1 (3.2)
Within 1 hour	4 (12.9)
Indirect care hours per week	
Up to 8 hours	1 (3.2)
9 - 19 hours	2 (6.5)
20 - 39 hours	3 (9.7)
40+ hours	23 (74.2)
Direct care hours per week	
Up to 8 hours	1 (3.2)
9 - 19 hours	5 (16.1)
20-39 hours	10 (32.3)
40+ hours	15 (48.4)
Caregiving duration	
6 months - 2 years	9 (29.0)
2 - 5 years	13 (41.9)
5+ years	9 (29.0)

[a]GP4C: GamePlan4Care.
[b]CR: care recipient.
[c]TIA: transient ischemic attacks.
[d]MCI: mild cognitive impairment.

Participants reported that the majority of care recipients were diagnosed with either Alzheimer disease (13/31, 41.9%) or dementia (15/31, 48.4%). Over 80% of caregivers were living with the care recipient (4/31, 12.9%) or living nearby (less than 15 minutes; 22/31, 71.0%). Most participants also reported providing over 20 hours of either indirect (26/31 participants, 83.9%) or direct (25/31 participants, 80.6%) care weekly for at least 2 years (22/31 participants, 71.0%).

Data Collection

Usability testing sessions were conducted by 2 members of our research team (one male senior project manager, author TB, and one female research project coordinator). Both were masters-trained researchers with psychology and public health backgrounds with training in counseling and research ethics. Sessions lasted approximately 2 hours with each participant. Only the participant and the member of the research team were in the room for the usability testing. Research staff read each task individually to the participant with prompts about each task. Throughout the testing session, research staff prompted the participant with generic prompts to explain what was happening during the testing and the participant's thoughts. At the end of the testing session, research staff asked participants questions about their overall impression of the platform, what could be improved, and the positive and negative aspects of the platform. Research staff had blank paper available during the usability testing sessions for any notes that could be used in future analysis of the participant's performance while

XSL•FO
RenderX

completing the tasks, including an overall sense of the participant's computer literacy.

Analysis

Analysis of the data was performed using an inductive thematic analysis approach [25] with 31 transcripts using no statistical software. The research team focused on participants' progress on the tasks and their valuable feedback on their experience as they completed the tasks. Participants were observed, vocalized what they were thinking and feeling, and then provided final feedback after completion of the tasks. All of this information was triangulated to determine the users' experience with the platform [26].

Computerized transcripts were reviewed initially for quality assurance by research staff. Two researchers completed the thematic analysis. Both researchers independently reviewed the transcripts and revisited the recorded audio of participants completing the tasks. Initial ideas were documented and discussed during the initial phase of data familiarization. The coding phase of the data was then initiated. Key terms (words and short or long phrases) were identified to illustrate the concepts of the caregivers' experience and feedback found in the transcripts and were saved in an excel spreadsheet. Key terms were discussed over multiple meetings. Once agreement was reached on the key terms, they were brought to the full team for further discussion. These key terms represented underlying ideas from the data that the research team considered pertinent to the aim of the study. The third stage consisted of formulating themes by grouping together key terms that reflected similar overarching ideas. Common themes from technical and content-related aspects and a major theme specifying each aspect emerged. In addition, participant-suggested themes also emerged. Under each major theme, subthemes related to the main theme, but separate and distinct enough to be differentiated, were also revealed. Major themes and subthemes applied to the data were defined and labeled. Finally, examples were chosen from the transcripts to illustrate theme elements. IBM SPSS software was used for descriptive statistics of the participants [27].

Ethical Considerations

This study was approved by the institutional review board at the participating institution (Baylor Scott & White Research Institute, IRB Number 018‑622) and participants gave informed consent. Informed consent was given electronically before data collection and usability testing, and participants were given a paper copy to follow along with during the electronic consent process. Participants chose to receive either an electronic or mailed copy of the signed consent form. All data presented are deidentified. Participants received US $100 gift cards after completion of the usability testing.

Results

Overview

Overall, many participants seemed satisfied with the quality of the content and flow of the platform. One participant said, "Oh, it is very clear and understandable…" Another participant commented that "it [videos] weren't exhausting or boring." Another participant felt that the videos "…are very clear, very simple, they are not overwhelming, they are not over my head, so it is good." Several other participants also commented that the videos were "very soothing." A participant also mentioned that they thought the videos would be used by caregivers to refer back to when brainstorming different strategies to use with the care recipient. An overview of the themes and subthemes is presented in Figure 1.

XSL•FO
RenderX

Figure 1. Summary of major themes and subthemes of the GP4C usability testing data from caregivers of persons living with dementia recruited from October 2019–February 2020. Note: Main themes are underlined. GP4C: GamePlan4Care, CR: care recipient.

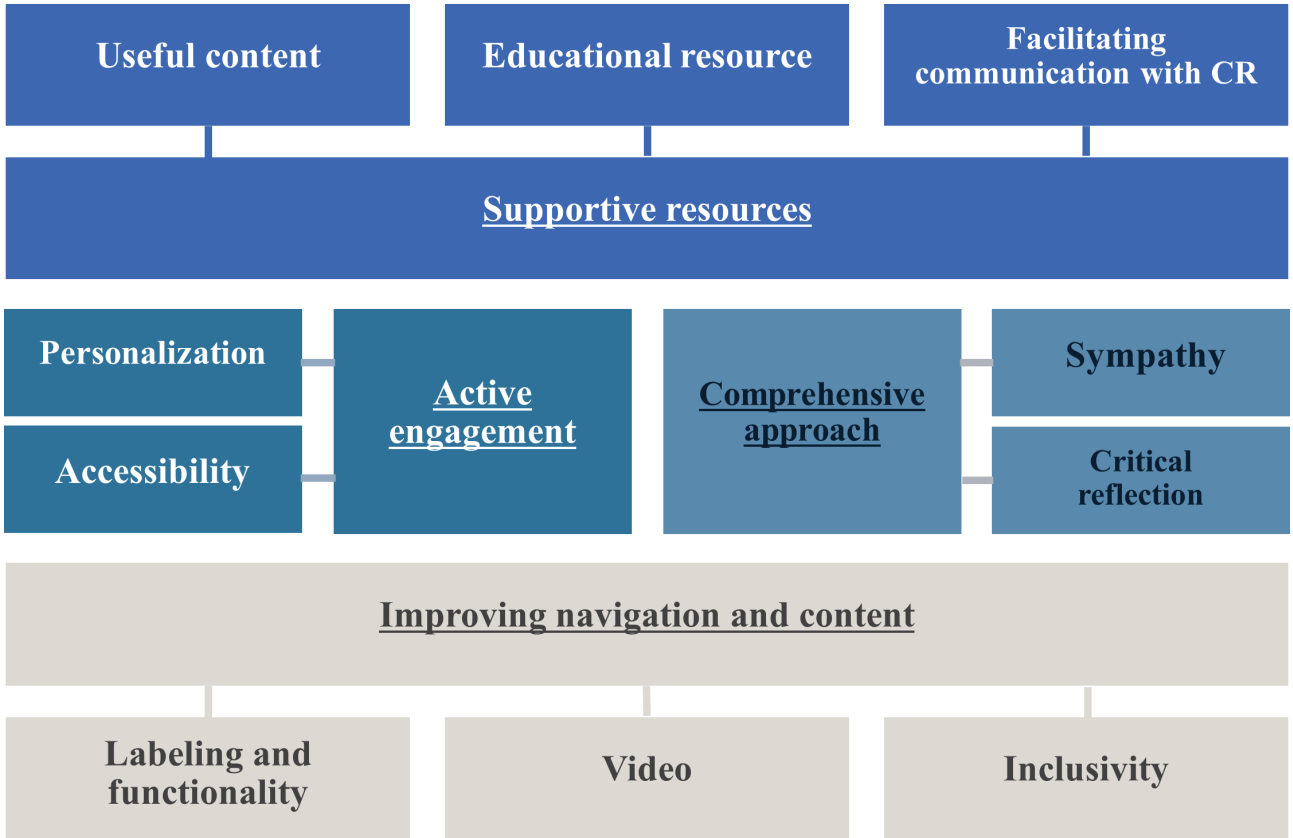

Common Theme: Supportive Resources

In addition to the general positive responses, all the participants agreed that the platform provided very supportive resources for their caregiving responsibilities. Caregivers described the content as useful and educational, suggesting that the platform would be beneficial to facilitate their caregiving tasks and communication with their care recipients.

Useful Content

Participants felt that the content was appropriate. When they were asked to review the safety topic in particular, some caregivers seemed to agree that the safety tips on driving are an important strategy that they should address with their care recipients. For example, one caregiver said, “It is very clear that there is going to be some strategies and…hints to help the caregiver.” Another participant noted, “I am liking what it says. This is a reminder of what I need to be looking for.” Similarly, another participant explained, “It is not too long, yet it is very informative.” Another participant echoed a similar perspective by saying, “I think this is very good as far as different topics and stuff.” Another participant commented that the content was “conversational” and that they liked the “vocabulary” used throughout the platform. One participant expressed that they felt the videos were useful, they said, “…reading something is one thing, but to have the video show, it just really keeps your mind, you know, wanting to continue.”

Educational Resource

Several participants commented that the program would help them persist with caregiver education. For example, one participant said, “It is not something I would just watch and put it away. I could have these printed out in my little book.” Another participant stated, “I like how they are almost making it [the program] like a continuing education.” Similarly, another participant said, “It’s not too long, it gives information, it makes me want to come back and see what else is there.”

Facilitating Communication With Care Recipients

As many participants recognized the materials as helpful and educational, a participant commented that the worksheets could help facilitate difficult conversations with the care recipient and help them decide on changes together.

> *So, to discuss those worksheets…that’s another good reason for worksheets… I'm saying I feel like I'm always the one to having to say no you can't, no you can't, and so the person receiving care, this is a way… to say we're going to go through this and we're going to help you identify what things are safe for you, what you feel comfortable with and what you may not. So, it’s a platform that opens up the communication with that individual to try to keep them safe as it relates to driving.*

Technical Theme: Active Engagement

Participants noted that they appreciated the active engagement required for dementia caregivers. One participant said that they liked “That you are actually trying to do something to help the caregivers out there rather than just saying oh you got a local support group down the street.” A caregiver also mentioned that she appreciated she could access the information on her

XSL•FO
RenderX

own time. She said, "…it is giving you support where there is nothing expected of you, you know you don't have to come back and do 20 visits, it is all online when you have time." Another participant felt similarly and said, "I like the thought of tutorials that I can go through and work on myself at my pace and all that." The feature that the information was gathered and organized in one central location was appreciated. For example,

> *I see this [is] very unique and different than I've accessed so far, and I've been doing a whole lot of research…I've been online for weeks and week and weeks now, so this is definitely something that I think will be used a lot…*

Personalization

Participants liked the level of personalization. For example, after entering information about the care recipient, caregivers would go on to receive questions that included that relationship (eg, "your mother"). One participant said, "the most positive are definitely the suggestions that strictly apply to me from what they heard from my answers." Similarly, another participant said, "It is nice to know that they have recognized my answers."

Accessibility

Participants also seemed to appreciate that they could connect with a professional (ie, dementia care navigator) for additional help, especially in times of crisis. A participant said,

> *It would be good to have [a dementia care navigator], as a caregiver you get flustered with everything and bogged down with everything and you cannot think correctly sometimes because you are overwhelmed with trying to figure out how to fix the situation.*

Another participant said,

> *I think that that's important, but there are some times where you just feel alone because there is nobody out there who gets what you are going through…and you have a specialist out there who is willing to listen without judgement…*

One participant also stressed the importance of a care specialist (or other professional) providing thoughtful and specific feedback when reviewing information that had been completed on the platform. They said,

> *I would want to know that somebody was reviewing and understanding what my situation is. I wouldn't want somebody to be, you know, just looking at it and doing a cursory review.*

Content Theme: Comprehensive Approach

The themes from content interviews or sessions can be divided into 2 subthemes: sympathy and critical reflection. Caregivers expressed that the features of the platform resonate with their experience.

Sympathy

A participant appreciated that the topic they visited (ie, caregiver guilt) was featured in the videos, which resonated with their experience. The participant explained,

> *I'm very happy to hear that guilt is something that they are definitely touching on…guilt is a big part, and as a caregiver, you can find yourself going down that road…very quickly.*

Another participant also appreciated that addressing caregiver guilt was included in the platform. The participant said,

> *I like that [it's included] because…sometimes you feel guilty and it's just a little bit, but sometimes you feel guilty and it's extreme.*

Another participant agreed and said,

> *When I feel guilty I feel like God how can I be tired or I will think to myself…you know she took care of you and lot of thoughts…but maybe [I] felt ashamed because I felt that and then of course I felt sorry.*

Similarly, another participant commented that

> *Another thing…that needs to be said here is that you're going to have negative feelings and you're going to have the negative responses, and you have to learn to forgive yourself. I think because in the sense telling me I shouldn't have the negative thoughts increases my guilt.*

Critical Reflection

Caregivers also felt that the videos did a nice job of helping caregivers think critically about their choices. A participant said, "…it makes you think where can they drive, what do I cut out first." Another participant commented that the video made them think differently about their emotions. The participant said,

> *sometimes I think 'mom, are you just playing with me?'…they are not, you just got to be patient and then the second part of it about the guilt in that, that really makes you think and makes you analyze yourself.*

Several caregivers also noted that they appreciated that the videos reminded caregivers that they could not change the care recipient. A participant said,

> *So, I also very much appreciate the way that they are subtly reminding the us as caregivers that the person with dementia cannot change and that is a very big part of dealing as a caregiver for with someone with dementia and Alzheimer's is that they don't know how they are and have to be…to remind yourself this is not who the person was. This is who they are now, and they cannot change it. So, I like seeing that in here.*

Participant Suggested Themes: Improving Navigation and Content

As a main goal of the tasks, participants provided feedback on how to improve the platform features and contents. Three areas of improvement were identified: labeling and functionality, video, and inclusivity.

XSL•FO
RenderX

Labeling and Functionality

Clearer labeling or using different colors on the platform was suggested by several participants. For example, incorporating a "box that says next" rather than just an arrow. Another participant suggested "the button area and the fonts" could be bigger. Another participant suggested making completed sections "a different color than the rest" and several participants mentioned that they were color blind and had a difficult time seeing text show up. The participant said, "…I'm color blind, so text itself, I can't tell what that looks like until I go to the box." However, other participants commented that they liked the colors that were used. One participant said,

> *Oh, I like all the blue and gold and stuff like that. I think they are very soothing to the eye. The whites good because you need to make everything else pop up on the screen better. I mean, I think this is good because at least the dark blue highlights all the stuff over here.*

Video

One of the few criticisms of the videos was given by a participant who felt that the videos were "a little slow in some parts." However, another participant disagreed and felt that at times the narrators in the video were "speaking a little fast." Another participant suggested that the video reflections have more parameters. For example, they suggested that the video narrator say, "tell me what you're thinking in two sentences" rather than "what are you thinking?" One participant felt that the video narrator should introduce themselves each time and provide a brief overview of the topic. The participant felt this would be particularly helpful when caregivers were skipping around the site. They said, "…maybe you need to … tell what you're going to talk about."

Inclusivity

Participants addressed the importance of making sure the materials were inclusive for a wide range of individuals. For example, a caregiver also noted that cost efficiency is important when creating any materials that caregivers are expected to print. The caregiver said,

> *I have to control cost in the household, because I'm looking at hiring people to come in, so if I can print this, even though this looks pretty, if it would print in black only without me having to do anything, then that would be great because this is extra ink and cost.*

Another participant mentioned concerns about how a caregiver might access materials, particularly if "some of the [computer] software is really, really old." The caregiver went on to explain that if it was too complicated, caregivers would likely not download the worksheets and that "technical support" may be useful for some caregivers. One participant also mentioned the materials were needed for individuals who didn't speak English. The participant said, "…within the United States there are many Spanish-speaking individuals, so at some point certainly should be translated…"

Discussion

Principal Findings

This study presents participants' experiences from user testing of GP4C, a web-based adaptation of the REACH II caregiver intervention. Similar to other usability studies [28,29], both content- and technology-related aspects of our platform were explored. We focused on the way participants experienced and reacted to the GP4C platform while completing their tasks. Results showed that GP4C was appreciated by participants because it offered a unique resource for those who provide care to a person living with dementia. In general, participants recognized the platform's comprehensive approach reflecting their experience as caregivers. Active engagement, including personalization for their own needs and accessibility to resources such as the Dementia Care Navigator intervention staff, was emphasized as the best features of the platform. After completing tasks with the platform, participants provided valuable suggestions to improve navigation and platform content. Thus, the overall positive responses from participants were promising and encouraged the research team to move forward in finalizing the platform.

Implications of Platform Revision Findings

The qualitative findings of the study were used by the research team to revise the platform before testing in a randomized clinical trial. The following design changes were implemented in the final iteration of the GP4C platform.

Navigation and basic layout

Based on user feedback, the research team worked with the development team to change the navigation and basic layout. For example, the platform included explicit buttons of "Next" and "Back," which were resized and recolored to increase visibility, callouts (eg, "new message") were added to explicitly capture the caregiver's attention, and clues (eg, images, spinners, and prompts) were added to make workflow and progression smoother.

Assessment items and response choices

Not surprising, the participants preferred a conversational tone in assessment items and real-life reflective responses. For instance, one assessment question was changed from "How severe are the dementia symptoms?" to "How would you describe his or her thinking and memory changes?" and one of the choices to the question, "Does he or she drive?" was changed from "Never" to "Never, or not anymore." This feedback from participants was considered in the development of the risk assessments being tested and for the additional 5 content domains that were created after UI/UX testing.

Tools to orient users to the site and its use

Some participants, especially older caregivers, were frustrated with the newly developed interface because they were not familiar with it. To assure simplicity and logical progression of the interface, the platform added more tools to assist caregivers (eg, welcome video presented upon first visit to user dashboard, opening and closing scenes to remind users how to start and complete goals, and demonstrations of live use of worksheets).

XSL•FO
RenderX

Each of the 7 content domains included in the final version of GP4C shared a similar format in presentation of information, support in goal setting, and functionality.

Comparison of Findings to Existing Platforms Designed for Caregivers

Usability testing of other web-based caregiver support interventions found findings similar to our study. Feedback from two websites, one web-based training and support program for dementia caregivers [30] and one informational website for people with dementia and their carers [31], indicated that changes needed to be made to the labeling and functionality of "next" buttons [30,31], indicating a need to ensure easy readability and navigation. A systematic review of web-based interventions for dementia caregivers found similar findings regarding the importance of the use of text for labeling, font size, and color choice for buttons [32]. Similar to our platform personalizing content for users, the program by Teles et al [30] allowed users to self-personalize the platform by completing a user profile, which was found to be appraised positively. A usability testing of a web-based intervention for informal caregivers [33] and an acceptability and feasibility pilot study of a self-guided, web-based intervention for dementia caregivers [17] found, similar to our results, caregivers appreciated the use of video content and thought the videos were engaging and helped communicate key takeaways in a timely manner. The systematic review by Ottaviani et al [32] also found that the use of videos and practical examples with real care situations was valuable to users. One area of recommendation from the usability testing of a web-based program for caregivers of persons living with Alzheimer disease that our platform included in its design and implementation was the ability for users to interact with professionals [34]. Accessing a professional was a planned component of the program by Cristancho-Lacroix et al [34], but was not implemented due to a lack of resources. A feasibility study of a web-based intervention for dementia caregivers included motivational coaching sessions incorporated into their platform [35]. Feedback from participants indicated that they greatly appreciated the availability of a coach through multiple means (video call, phone call, and email) and thought the personal feedback provided by the coaches increased their active engagement with the intervention [35]. Another qualitative study of a web-based intervention for caregivers of older adults with dementia and multiple chronic conditions found that participants recommended the availability of a person to help answer questions and direct the participant to resources that met their specific needs [36]. These findings support the benefit of the inclusion of a Dementia Care Navigator in our platform. Although web-based interventions are designed to be self-directed, users appreciate the ability to have access to a professional when they need additional support.

Limitations

This exploratory usability testing is not intended for generalization but rather to provide user feedback to a specific digital health tool. Thus, a limitation of this study may be its lack of representativeness in the sample and generalizability to a broader population. The participants of the study were recruited from one geographic area of the United States. Other caregiver populations in different regions might present different opinions due to their accessibility to local resources and support. Next, the themes identified and presented in this study were identified in the midst of the GP4C platform development. Participants were only allowed to visit portions of the domains (ie, safety and emotions) rather than the full platform. This was because the iterative design resulted in updates being made to the platform as the usability testing was being completed. We acknowledge that participants may express different opinions on the final version of the platform if they were able to visit all of the domains. Moreover, given the limitations of the sample size, demographic variation in participants was not included when analyzing findings. Furthermore, the digital divide may lead to voluntary bias among caregivers from diverse backgrounds. The participants who provided feedback for this study had internet access and were already inclined to use it for web-based information searches. Additional recruitment efforts should be made to minimize this bias. Despite these limitations, this study provides a model for conducting usability testing of web-based caregiver support programs with the end user.

Conclusions

This study presented the first stage of the development of GP4C, a web-based family caregiver support platform. This study demonstrates the value of having end users participate in the development and early evaluation of digital tools. The UI/UX approach provided feedback on both the technical and content aspects of the web-based platform developed by the research team. Caregiver feedback led to changes in the functionality of the system and guided the research team in the development of additional content to be added to the system. Usability testing and participants' in-depth and positive feedback showed its functionality and utility as a potential caregiver support program for dementia caregivers. The research team finalized the platform based on these results and is currently conducting a randomized clinical trial to show the effectiveness of the platform.

Acknowledgments

This study was supported by the National Institute on Aging (5R01AG061973 – 05, 3R01AG061973-04S1).

Data Availability

Findings from this study represent the initial development and evaluation of the GamePlan4Care web-based platform for a Phase I clinical trial testing of the GamePlan4Care web-based platform. Due to its exploratory nature, funding was not provided to make the datasets generated or analyzed for this study publicly available but the dataset is available upon reasonable request from the corresponding author.

XSL•FO
RenderX

Conflicts of Interest

None declared.

References

1. The Alzheimer's Association. 2024 Alzheimer's disease facts and figures. Alzheimers Dement 2024 May;20(5):3708-3821. [doi: 10.1002/alz.13809] [Medline: 38689398]
2. Committee on Family Caregiving for Older Adults, Board on Health Care Services, Health and Medicine Division, National Academies of Sciences, Engineering, and Medicine. Families Caring for an Aging America: National Academies Press (US); 2016. URL: https://www.ncbi.nlm.nih.gov/pubmed/27905704
3. Gwyther LP. Social issues of the Alzheimer's patient and family. Am J Med 1998 Apr 27;104(4A):17S-21S. [doi: 10.1016/s0002-9343(98)00024-2] [Medline: 9617848]
4. Ory MG, Hoffman RR 3rd, Yee JL, Tennstedt S, Schulz R. Prevalence and impact of caregiving: a detailed comparison between dementia and nondementia caregivers. Gerontologist 1999 Apr;39(2):177-185. [doi: 10.1093/geront/39.2.177] [Medline: 10224714]
5. Pinquart M, Sörensen S. Differences between caregivers and noncaregivers in psychological health and physical health: a meta-analysis. Psychol Aging 2003 Jun;18(2):250-267. [doi: 10.1037/0882-7974.18.2.250] [Medline: 12825775]
6. He J, Wang J, Zhong H, Guan C. The effectiveness of multi-component interventions on the positive and negative aspects of well-being among informal caregivers of people with dementia: A systematic review and meta-analysis. Int J Environ Res Public Health 2022 Jun 7;19(12):6973. [doi: 10.3390/ijerph19126973]
7. Belle SH, Czaja SJ, Schulz R, et al. Using a new taxonomy to combine the uncombinable: integrating results across diverse interventions. Psychol Aging 2003 Sep;18(3):396-405. [doi: 10.1037/0882-7974.18.3.396] [Medline: 14518803]
8. Schulz R, Belle SH, Czaja SJ, Gitlin LN, Wisniewski SR, Ory MG. Introduction to the special section on Resources for Enhancing Alzheimer's Caregiver Health (REACH). Psychol Aging 2003 Sep;18(3):357-360. [doi: 10.1037/0882-7974.18.3.357]
9. Richards KC, Enderlin CA, Beck C, McSweeney JC, Jones TC, Roberson PK. Tailored biobehavioral interventions: a literature review and synthesis. Res Theory Nurs Pract 2007;21(4):271-285. [doi: 10.1891/088971807782428029] [Medline: 18236771]
10. Gitlin LN, Marx K, Stanley IH, Hodgson N. Translating evidence-based dementia caregiving interventions into practice: State-of-the-science and next steps. Gerontologist 2015 Apr;55(2):210-226. [doi: 10.1093/geront/gnu123] [Medline: 26035597]
11. Maslow K. Translating Innovation to Impact: Evidence-based interventions to support people with Alzheimer's disease and their caregivers at home and in the community. 2012. URL: https://www.agingresearch.org/wp-content/uploads/2018/03/508-Compliant-AoA-White-Paper-for-Release.pdf [accessed 2025-07-03]
12. Weiss J, Tumosa N, Perweiler E, et al. Critical workforce gaps in dementia education and training. J Am Geriatr Soc 2020 Mar;68(3):625-629. [doi: 10.1111/jgs.16341] [Medline: 31967320]
13. Leng M, Zhao Y, Xiao H, Li C, Wang Z. Internet-based supportive interventions for family caregivers of people with dementia: Systematic review and meta-analysis. J Med Internet Res 2020 Sep 9;22(9):e19468. [doi: 10.2196/19468] [Medline: 32902388]
14. Waller A, Dilworth S, Mansfield E, Sanson-Fisher R. Computer and telephone delivered interventions to support caregivers of people with dementia: a systematic review of research output and quality. BMC Geriatr 2017 Nov 16;17(1):265. [doi: 10.1186/s12877-017-0654-6] [Medline: 29145806]
15. Ramirez M, Duran MC, Pabiniak CJ, et al. Family caregiver needs and preferences for virtual training to manage behavioral and psychological symptoms of dementia: Interview study. JMIR Aging 2021 Feb 10;4(1):e24965. [doi: 10.2196/24965] [Medline: 33565984]
16. Hong YA, Shen K, Han HR, et al. A WeChat-based Intervention, Wellness Enhancement for Caregivers (WECARE), for Chinese American dementia caregivers: Pilot assessment of feasibility, acceptability, and preliminary efficacy. JMIR Aging 2023 Apr 5;6:e42972. [doi: 10.2196/42972] [Medline: 37018042]
17. Kwok I, Lattie EG, Yang D, et al. Acceptability and feasibility of a socially enhanced, self-guided, positive emotion regulation intervention for caregivers of individuals with dementia: Pilot intervention study. JMIR Aging 2023 Sep 6;6:e46269. [doi: 10.2196/46269] [Medline: 37672311]
18. Premanandan S, Ahmad A, Cajander Å, Ågerfalk P, Dolezel M, van Gemert-Pijnen L. Designing a mobile e-coaching app for immigrant informal caregivers: Qualitative study using the persuasive system design model. JMIR Mhealth Uhealth 2023 Nov 9;11:e50038. [doi: 10.2196/50038] [Medline: 37943598]
19. Scott AF, Ayers S, Pluye P, et al. Impact and perceived value of iGeriCare e-learning among dementia care partners and others: Pilot evaluation using the IAM4all questionnaire. JMIR Aging 2022 Dec 22;5(4):e40357. [doi: 10.2196/40357] [Medline: 36150051]

XSL•FO
RenderX

20. Deeken F, Rezo A, Hinz M, Discher R, Rapp MA. Evaluation of technology-based interventions for informal caregivers of patients with dementia-a meta-analysis of randomized controlled trials. Am J Geriatr Psychiatry 2019 Apr;27(4):426-445. [doi: 10.1016/j.jagp.2018.12.003] [Medline: 30642650]
21. Naunton Morgan B, Windle G, Sharp R, Lamers C. eHealth and web-based interventions for informal carers of people with dementia in the community: Umbrella review. J Med Internet Res 2022 Jul 22;24(7):e36727. [doi: 10.2196/36727] [Medline: 35867388]
22. Zhao Y, Feng H, Hu M, et al. Web-based interventions to improve mental health in home caregivers of people with dementia: Meta-analysis. J Med Internet Res 2019 May 6;21(5):e13415. [doi: 10.2196/13415] [Medline: 31066680]
23. Kushniruk AW, Patel VL. Cognitive and usability engineering methods for the evaluation of clinical information systems. J Biomed Inform 2004 Feb;37(1):56-76. [doi: 10.1016/j.jbi.2004.01.003] [Medline: 15016386]
24. Galvin JE, Roe CM, Powlishta KK, et al. The AD8: a brief informant interview to detect dementia. Neurology (ECronicon) 2005 Aug 23;65(4):559-564. [doi: 10.1212/01.wnl.0000172958.95282.2a] [Medline: 16116116]
25. Braun V, Clarke V. One size fits all? What counts as quality practice in (reflexive) thematic analysis? Qual Res Psychol 2021 Jul 3;18(3):328-352. [doi: 10.1080/14780887.2020.1769238]
26. Carter N, Bryant-Lukosius D, DiCenso A, Blythe J, Neville AJ. The use of triangulation in qualitative research. Oncol Nurs Forum 2014 Sep;41(5):545-547. [doi: 10.1188/14.ONF.545-547] [Medline: 25158659]
27. IBM Corp. IBM SPSS statistics, version 29.0. 2024.
28. Haggerty T, Brabson L, Grogg KA, et al. Usability testing of an electronic health application for patient activation on weight management. Mhealth 2021;7:45. [doi: 10.21037/mhealth-20-119] [Medline: 34345622]
29. Steele Gray C, Khan AI, Kuluski K, et al. Improving patient experience and primary care quality for patients with complex chronic disease using the electronic patient-reported outcomes tool: Adopting qualitative methods into a user-centered design approach. JMIR Res Protoc 2016;5(1):e28. [doi: 10.2196/resprot.5204]
30. Teles S, Paúl C, Lima P, Chilro R, Ferreira A. User feedback and usability testing of an online training and support program for dementia carers. Internet Interv 2021 Sep;25:100412. [doi: 10.1016/j.invent.2021.100412] [Medline: 34401371]
31. Zheng J, Gresham M, Phillipson L, et al. Exploring the usability, user experience and usefulness of a supportive website for people with dementia and carers. Disabil Rehabil Assist Technol 2024 May;19(4):1369-1381. [doi: 10.1080/17483107.2023.2180546] [Medline: 37086036]
32. Ottaviani AC, Monteiro DQ, Oliveira D, et al. Usability and acceptability of internet-based interventions for family carers of people living with dementia: systematic review. Aging Ment Health 2022 Oct;26(10):1922-1932. [doi: 10.1080/13607863.2021.1975095] [Medline: 34511028]
33. Semonella M, Marchesi G, Andersson G, Dekel R, Pietrabissa G, Vilchinsky N. Usability study of SOSteniamoci: An internet-based intervention platform to support informal caregivers in Italy. Digit Health 2024;10:20552076231225082. [doi: 10.1177/20552076231225082] [Medline: 38235418]
34. Cristancho-Lacroix V, Moulin F, Wrobel J, et al. A web-based program for informal caregivers of persons with Alzheimer's disease: an iterative user-centered design. JMIR Res Protoc 2014 Sep 15;3(3):e46. [doi: 10.2196/resprot.3607] [Medline: 25263541]
35. Atefi GL, van Knippenberg RJM, Bartels SL, et al. A web-based intervention based on acceptance and commitment therapy for family caregivers of people with dementia: Mixed methods feasibility study. JMIR Aging 2024 Apr 4;7:e53489. [doi: 10.2196/53489] [Medline: 38574360]
36. Ploeg J, McAiney C, Duggleby W, et al. A web-based intervention to help caregivers of older adults with dementia and multiple chronic conditions: Qualitative study. JMIR Aging 2018 Apr 23;1(1):e2. [doi: 10.2196/aging.8475] [Medline: 31518230]

Abbreviations:

GP4C: GamePlan4Care
REACH II: Resources for Enhancing Alzheimer's Caregiver Heath II
REDCap: Research Electronic Data Capture
UI/UX: User interface/user experience

XSL•FO
RenderX

Edited by A Mavragani; submitted 03.05.24; peer-reviewed by H Feng, J Martindale-Adams; revised version received 21.04.25; accepted 23.04.25; published 11.07.25.

Please cite as:
Cho J, Birchfield T, Thorud JL, Ory MG, Stevens AB
GamePlan4Care, a Web-Based Adaptation of the Resources for Enhancing Alzheimer's Caregiver Health II Intervention for Family Caregivers of Persons Living With Dementia: Formative, Qualitative Usability Testing Study
JMIR Form Res 2025;9:e60143
URL: https://formative.jmir.org/2025/1/e60143
doi:10.2196/60143

© Jinmyoung Cho, Thomas Birchfield, Jennifer L Thorud, Marcia G Ory, Alan B Stevens. Originally published in JMIR Formative Research (https://formative.jmir.org), 11.7.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Bringing Executive Function Testing Online: Assessment Validation Study

Mouna Attarha[1], PhD; Ana Carolina de Figueiredo Pelegrino[2], MSc; Lydia Ouellet[2], MSc; Sarah-Jane Grant[1], MA; Etienne de Villers-Sidani[2], MDCM; Thomas Van Vleet[1], PhD

[1]Posit Science Corporation, 160 Pine St Suite 200, San Francisco, CA, United States

[2]Montreal Neurological Institute and Hospital, McGill University, Montreal, QC, Canada

Corresponding Author:
Mouna Attarha, PhD
Posit Science Corporation, 160 Pine St Suite 200, San Francisco, CA, United States

Abstract

Background: Executive function encompasses a set of higher-order cognitive processes, including planning, cognitive flexibility, and inhibitory control, that are essential for goal-directed behavior. These abilities are adversely affected by age, with executive dysfunction ultimately impairing the performance of activities of daily living.

Objective: This study aimed to assess the validity of a computerized cognitive assessment in predicting executive function performance in healthy older adults.

Methods: This retrospective analysis utilized baseline data from the Improving Neurological Health in Aging via Neuroplasticity-Based Computerized Exercise (INHANCE) trial. The study provides normative data for cognitively healthy older adults (aged 65 years and above) and evaluates the usability and validity of Freeze Frame, a cognitive assessment available on the BrainHQ platform. Performance on Freeze Frame was analyzed in relation to self-reported demographic variables and neuropsychological function, using a standardized measure of executive function, the National Institutes of Health Executive Abilities: Measures and Instruments for Neurobehavioral Evaluation and Research (NIH EXAMINER).

Results: The intent-to-treat analysis included 92 cognitively healthy older adults (mean age 71.9, SD 4.86, range: 65 - 83 years), of whom 66% (61/92) were female, with a mean education level of 16.45 (SD 3.40, range: 9 - 27) years. Performance on Freeze Frame was modestly associated with executive function scores on NIH EXAMINER (P=.02), accounting for 6.8% of the variance. The assessment showed a small but statistically significant relationship to age (ρ=−0.22, P=.046) and gender, with no significant influence of education. Psychometric evaluation supported its usability, with an average completion time of 4 (SD 0.16) minutes.

Conclusions: Freeze Frame is a brief, scalable, and accessible computerized cognitive assessment with demonstrated concurrent validity for executive function. Its efficiency and ease of administration across internet-connected devices suggests potential applications for cognitive screening. Future research should explore its utility in detecting executive dysfunction in clinical populations and its potential role in predicting functional performance across activities of daily living.

Trial Registration: ClinicalTrials.gov NCT04149457; https://clinicaltrials.gov/study/NCT04149457

International Registered Report Identifier (IRRID): RR2-10.2196/59705

(JMIR Form Res 2025;9:e75687) doi:10.2196/75687

KEYWORDS

executive function; web-based cognitive assessment; brain health; cognitive status; National Institutes of Health Executive Abilities: Measures and Instruments for Neurobehavioral Evaluation and Research; NIH EXAMINER; assessment accessibility

Introduction

Executive function is a critical domain of cognition that influences functional independence [1-5]. In healthy older adults, executive function abilities are closely associated with essential activities such as self-care, financial management, transportation use, and meal preparation. These skills support the capacity to make decisions, control action, and adapt to novel situations, which are necessary for daily living. However, executive function is known to decline with age [6,7]. Age-related changes can contribute to difficulties in maintaining autonomy and increase reliance on caregivers and support systems. Assessing executive function decline in aging populations facilitates the early identification of individuals at risk for functional impairments.

The National Institutes of Health Executive Abilities: Measures and Instruments for Neurobehavioral Evaluation and Research (NIH EXAMINER) was developed to provide a comprehensive, standardized assessment of executive function [8]. Traditional

XSL•FO
RenderX

in-person neuropsychological tests are often limited in their ability to capture the complex, multifaceted nature of executive function. Computerized assessment batteries such as Cogstate, Cognifit, CNS Vital Signs, CANTAB, and NIH Toolbox are broader-purpose neurocognitive batteries that largely aim to evaluate general cognition by including domains such as speed, attention, memory, language, and emotion processing, with executive functioning being relatively underrepresented within the overall battery. NIH EXAMINER was designed to address these gaps by incorporating multiple tasks to reliably and validly evaluate various aspects of executive function [9]. Research utilizing NIH EXAMINER has shown strong associations between executive function and age, due to changes in prefrontal cortex function [10]. Additionally, studies have widely used NIH EXAMINER in clinical research, showing sensitivity to executive dysfunction across neurodevelopmental conditions such as attention deficit disorder, acquired conditions such as traumatic brain injury, and neurodegenerative diseases such as Alzheimer disease and related dementias [11]. Performance deficits on NIH EXAMINER have been associated with impairments in activities of daily living [12]. As a result, NIH EXAMINER serves as a valuable tool for both research and clinical applications, aiding in the detection of executive dysfunction and informing targeted interventions to support cognition and functional independence.

There remains a growing need for scalable, computerized cognitive assessments due to significant barriers of traditional methods, such as prolonged administration times, clinician training, and the burden of patient travel. Addressing these limitations, this study examines the usability and preliminary validity of the Freeze Frame assessment to evaluate inhibitory control in a cohort of healthy older adults (aged 65 years and above) from the Improving Neurological Health in Aging via Neuroplasticity-Based Computerized Exercise (INHANCE) trial [13]. Inhibitory control represents a critical subprocess of executive functioning, involving the rapid suppression of prepotent or automatic responses in favor of task-relevant actions to regulate behavior in dynamic environments and support cognitive flexibility, working memory, and goal maintenance [14,15]. Empirical evidence from neuroimaging and meta-analyses have demonstrated that impairments to inhibitory control are observed across neurodevelopmental, psychiatric, and neurodegenerative diseases [16-19]. Inhibitory control therefore serves as a sensitive and mechanistically informative index of executive function integrity, making it a valuable target for cognitive assessment. Virtually all existing batteries of general cognition (eg, CANTAB, Cognifit, CNS Vital Signs) include at least 1 measure of inhibitory control.

Freeze Frame is similar to other measures of inhibitory control, particularly the Sustained Attention to Response Task (SART) [20]. Response contingencies in Freeze Frame were inspired by SART and work completed in hemispatial neglect and traumatic brain injury [20]. However, the timing varies considerably between the 2 tasks in that Freeze Frame utilizes a variable intertrial interval whereas SART has a consistent intertrial interval. The advantage of a variable time interval is that unlike consistent intertrial intervals, which increases attentional lapses and response anticipations [21], the unpredictable timing of the appearance of each image in Freeze Frame requires a more alert state and enhanced response control [22]. The most unique aspect of Freeze Frame apart from the SART or Conners Continuous Performance Task [23] is the real-time adaptation of the core features of the assessment, including target frequency, the timing of the intertrial interval, and the variable and often longer duration of the task that allows time for participants to wane in their sustained attentional and inhibitory control capacities. These novel task parameters distinguish Freeze Frame from other measures of inhibitory control.

Although computerized batteries of global cognition are preferred when a comprehensive evaluation of neurocognitive profiles is needed to assist with diagnostic decisions, a single brief inhibitory control task offers several advantages in terms of speed, scale, and frequent monitoring. A targeted measure can provide interpretable performance metrics without confounding from broader executive or global cognition domains. Short completion times make brief assessments well-suited for clinical use where patient burden must be minimized, longitudinal studies that require daily or weekly monitoring over time, or high-throughput research that requires remote testing of hundreds, thousands, or millions of participants. Finally, assessments that do not require clinical training or specialized hardware can be deployed by patients using mobile devices at home and monitored by clinicians or researchers remotely.

This study evaluates the psychometric properties of Freeze Frame, its associations with demographic variables (ie, age, gender, and education), and its concurrent validity relative to NIH EXAMINER. The findings provide preliminary data that contribute to the growing need for scalable, digital tools that assess executive function, offering a promising avenue for identifying individual differences in cognition and advancing research in aging.

Methods

Setting, Design, and Sample

Participants were recruited near McGill University, Canada. Participants were community-dwelling, healthy older adults aged 65 years and above. The data reported reflect the intent-to-treat population (N=92) from the INHANCE randomized clinical trial (see published protocol for details [13]). The study enrolled participants from July 2021 to December 2023, with the final follow-up in June 2024.

Participants and Inclusion and Exclusion Criteria

Participants were community-dwelling, healthy older adults. Inclusion criteria included individuals aged 65 years or older, proficient in English or French, capable of fulfilling study requirements, and cognitively intact with a Montreal Cognitive Assessment total score of ≥23, a cut-off that optimizes diagnostic sensitivity and specificity [24]. Exclusion criteria included neurocognitive disorders, suicidal ideation, major depression indicated by a total score of >10 on the Geriatric Depression Scale–Short Form, prior experience using BrainHQ within the past 5 years, concurrent clinical trial participation, pregnancy,

XSL•FO
RenderX

substance abuse, or medical conditions hindering study engagement. Full details are provided in the protocol [13].

Freeze Frame Assessment

The Freeze Frame assessment is a cognitive task designed to measure speeded inhibitory control using a reverse go/no-go paradigm [25-27]. Participants are shown a target image at the beginning of each block, followed by a rapid sequence of images consisting of both targets and foils, presented with unequal probabilities under constrained viewing durations. The interstimulus interval varied randomly between 500 and 1500 milliseconds. Participants were instructed to withhold their response when a target appeared while executing a speeded motor response to all foil stimuli. See Figure 1 for a visual representation of the task structure.

Figure 1. The Freeze Frame computerized cognitive assessment. Participants remember a target image presented at the start of the block after which a rapid stream of targets and foils are interleaved with unequal probability. Participants withhold a motor response to all targets and enter a speeded response to all foils. (**A**) Guided tutorial presenting the target image. (**B**) Example of sequentially presented trials of the target and sample foils. Responses are correct if the participant selects "no" to all foil images within a limited response window.

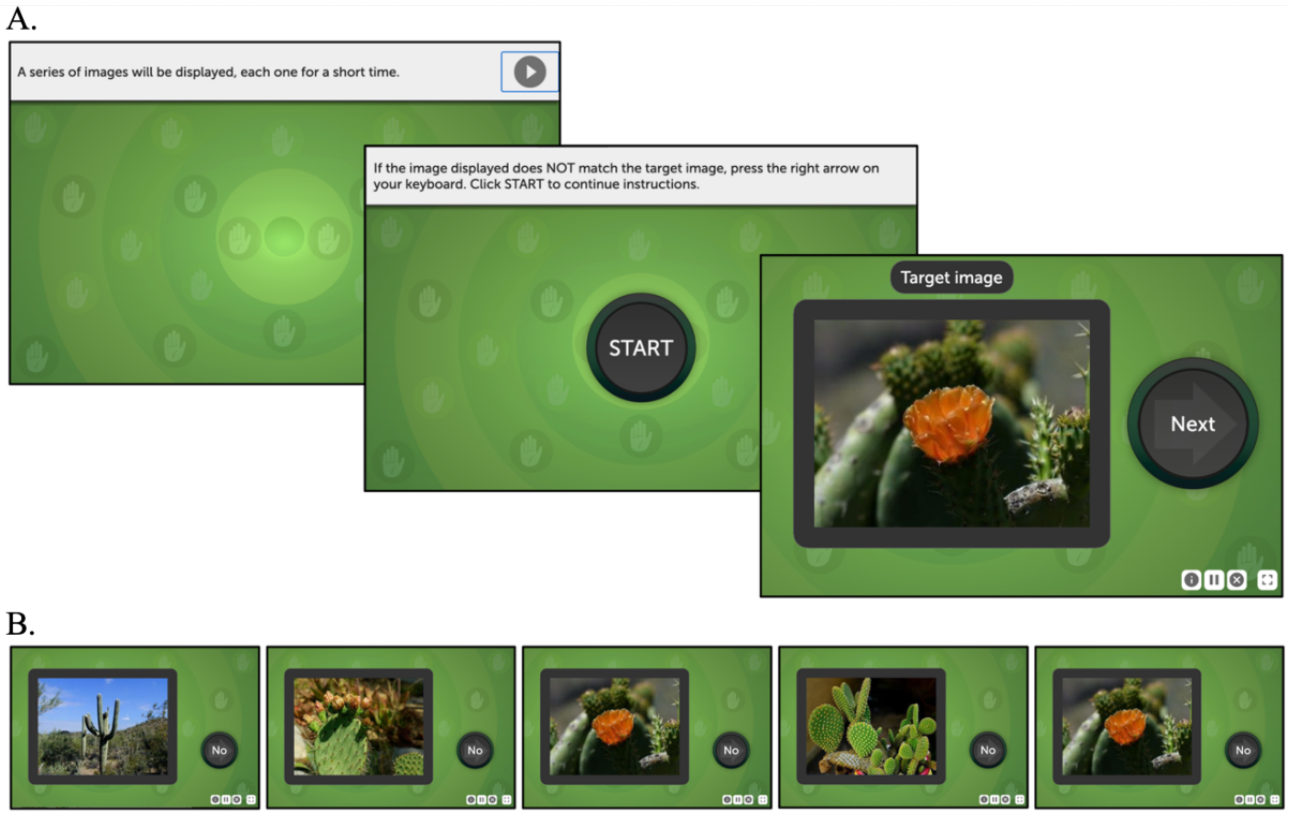

The adaptive dimension of the assessment was target frequency, with lower target frequencies increasing task difficulty. There are 7 possible target frequency levels: 40%, 35%, 30%, 25%, 20%, 15%, and 10%. All participants began the assessment at a target frequency of 30%. Adaptation occurred over 5 epochs, with each epoch consisting of 30 trials. Performance-based adjustments to target frequency were made based on the abilities of each participant. If the participant correctly withheld responses to at least 80% of targets and correctly responded to at least 80% of foils, the task progressed to a more difficult level (ie, a lower target frequency). If these accuracy thresholds were not met, the task became easier by increasing the target frequency on the subsequent epoch.

The raw threshold score was recorded as an integer ranging from 1 to 7, where 1 corresponded to the easiest condition (40% target frequency) and 7 represented the most challenging condition (10% target frequency). While threshold scores were used in the raw data output, these thresholds were converted to mean accuracy, as discussed in the protocol [13].

Ethical Considerations

The study was developed in accordance with the Declaration of Helsinki guidelines and was approved by the Western Institutional Review Board (IRB00000533) and Research Ethics Board of McGill University Health Centre (2020 - 6474). All participants provided written informed consent.

Study data were recorded into a secure, web-based electronic case report form at the study site through the Longitudinal Online Research and Imaging System (LORIS). This system meets relevant privacy and security standards for electronic trial data entry and storage, as well as the Health Insurance Portability and Accountability Act and Personal Information Protection and Electronic Documents Act standards for confidentiality and privacy [28]. Following consent, each participant was assigned a standardized Participant Identification Number composed of digits to identify the study and digits to identify the participant. All electronic case report form data entries were deidentified, using the Participant Identification Number and not the participant's name.

XSL•FO
RenderX

Participants received CAD $40 (approximately US $30) for completing the baseline visit, which included the measures described below. Payment followed the completion of the visit. Additional details are provided in the protocol [13].

The study was registered on November 4, 2019, before participant enrollment under the name "Improving Neurological Health in Aging Via Neuroplasticity-based Computerized Exercise (INHANCE)" with identification number NCT04149457.

Measures

Blinded staff conducted assessment administration and scoring. Following informed consent, participants completed a structured clinical interview, the Freeze Frame assessment, and a validated neuropsychological assessment of executive function (NIH EXAMINER). The structured clinical interview (20 minutes, in person) collected key demographic (eg, age, gender, education). The Freeze Frame assessment (8 minutes, in person) evaluated inhibitory control. NIH EXAMINER (20 minutes, in person) assessed executive function [8], and performance was measured using the executive composite score [10], which is automatically generated by the assessment program and converted to a *z* score per the registered protocol [13].

Analyses

Overview

Recruitment spanned from July 2021 to December 2023. The final follow-up visit was on June 7, 2024, and the study team was unblinded on June 14, 2024, after database lock. For all analyses, a $P<.05$ determined statistical significance [13].

Demographic Characteristics

Participants self-reported their age, gender, and educational level and completed the NIH EXAMINER assessment. Descriptive statistics including arithmetic means, SDs, and ranges were provided for continuous variables (age, education, NIH EXAMINER) to summarize central tendency and variability. For categorical variables such as gender, we calculated the frequency and percentage of participants who selected each response option.

Assessment Characteristics

To assess the usability of the Freeze Frame assessment, we used descriptive statistics (arithmetic mean and SD), characteristics of the distribution (skew and kurtosis), and psychometric properties including the performance histogram (the proportion of participants achieving each possible assessment score), as well as the time spent completing the assessment. To assess the presence and extent of ceiling and floor effects, we determined the proportion of participants who obtained the lowest and highest possible scores. A high frequency of extreme scores may suggest limitations in the assessment's ability to discriminate performance across different levels of ability.

Associations Between Freeze Frame, Demographics, and NIH EXAMINER

To examine the relationship between performance on the Freeze Frame assessment and key demographic variables, we used Spearman ρ for age and education. For gender-based comparisons, we applied the Wilcoxon rank-sum test to evaluate differences in assessment performance between women and men. We used Pearson *r* to investigate the relationship between assessment performance and NIH EXAMINER. We further used simple linear regression to evaluate the relationship between Freeze Frame and NIH EXAMINER to assess the extent to which performance on the assessment could account for variance in executive function scores, thereby providing concurrent validity with an established neuropsychological measure.

Results

Demographic Characteristics

Participants (N=92) had a mean age of 71.90 (SD 4.86; range: 65 - 83) years and a mean education level of 16.45 (SD 3.40, range: 9 - 27) years. Of the 92 participants, 61 (66%) identified as female and 31 (34%) identified as male. The baseline NIH EXAMINER executive composite mean was 0.43 (SD 0.57; range: −1.50 to 1.52; see Table 1). A complete characterization of the intent-to-treat population has been published previously [29].

XSL•FO
RenderX

Table . Baseline characteristics of the intent-to-treat population (N=92).

Characteristic	Participants
Age (years), mean (SD)	71.9 (4.9)
Education (years), mean (SD)	16 (3.4)
Gender, n (%)	
Female	61 (66)
Male	31 (34)
Race, n (%)	
White	88 (96)
Asian	2 (2)
Black or African American	1 (1)
American Indian or Alaska Native	0 (0)
Native Hawaiian or other Pacific Islander	0 (0)
More than one race	0 (0)
Unknown or not reported	1 (1)
Ethnicity, n (%)	
Not Hispanic or Latino	91 (99)
Hispanic or Latino	0 (0)
Unknown or not reported	1 (1)
Freeze Frame (accuracy), raw score (SD)	92.01 (4.70)
NIH EXAMINER[a], executive composite score (SD)	0.43 (0.57)

[a]NIH EXAMINER: National Institutes of Health Executive Abilities: Measures and Instruments for Neurobehavioral Evaluation and Research.

Assessment Characteristics

The distribution of Freeze Frame assessment scores is displayed in Figure 2A. Mean accuracy was 92.01% (SD 4.70%), with a skew of −0.99 and a kurtosis of 0.37. With respect to extreme scores, the highest recorded mean accuracy was 98%, achieved by 3 (3.75%) participants with no participant obtaining a perfect score. The lowest recorded mean accuracy was 77.33%, observed in 1 (1.25%) participant. Participants required an average of 4.27 (SD 0.16) minutes to complete the assessment.

XSL•FO
RenderX

Figure 2. Outcome measures. (**A**) Performance histogram of Freeze Frame scores (mean accuracy) across the intent-to-treat population. Association of Freeze Frame with demographic characteristics including (**B**) age, (**C**) years of education, and (**D**) gender. (**E**) Association of Freeze Frame with the NIH EXAMINER executive composite, a standard validated measure of executive function. NIH EXAMINER: National Institutes of Health Executive Abilities: Measures and Instruments for Neurobehavioral Evaluation and Research.

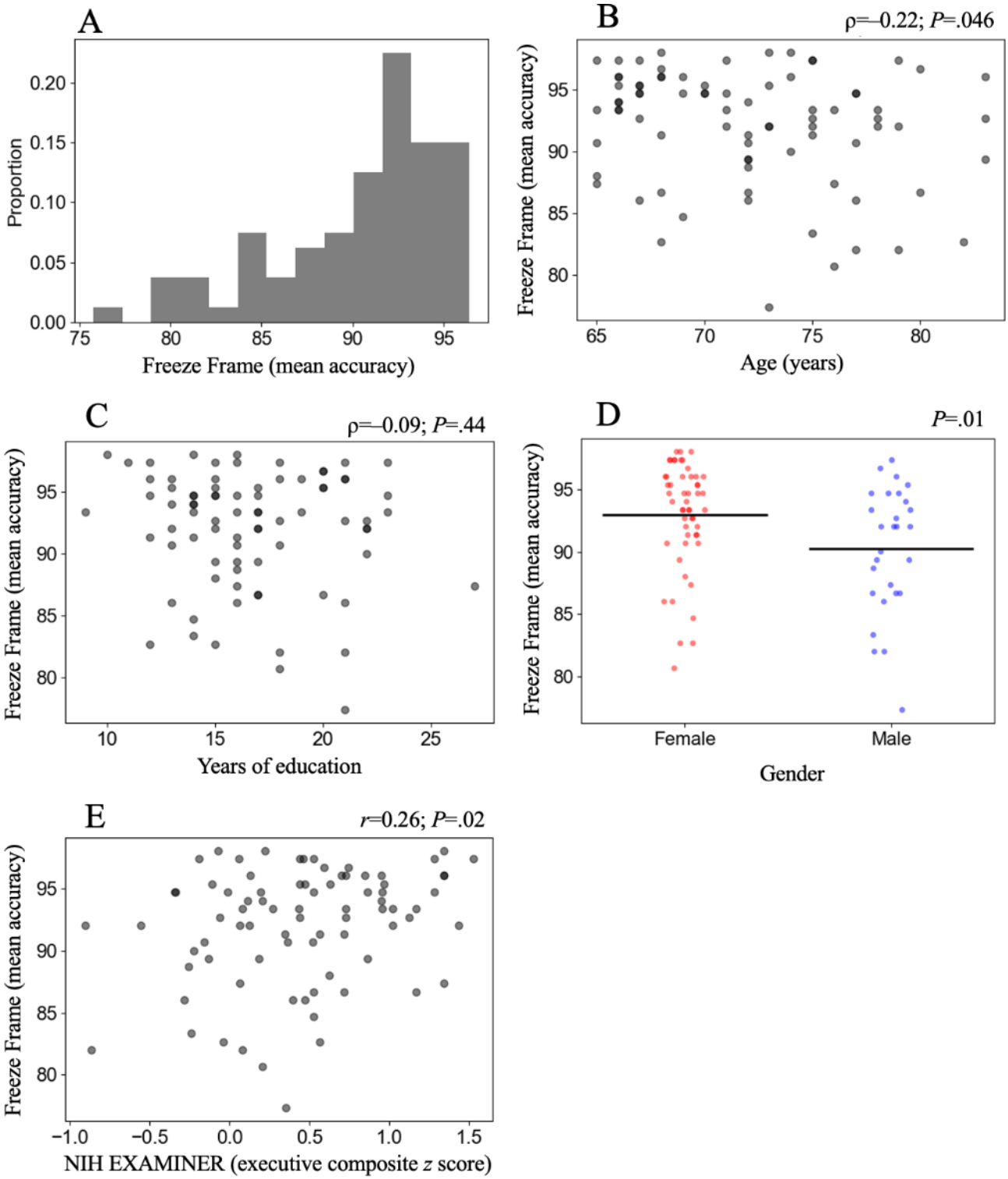

Associations Between Freeze Frame, Demographics and NIH EXAMINER

Performance on the Freeze Frame assessment showed a small but statistically significant negative association with age (ρ=−0.22, *P*=.046), and a significant effect of gender with women performing better than men (*P*=.01). In contrast, the association between Freeze Frame performance and years of education was nonsignificant (ρ=−0.09, *P*=.44; see Figure 2B–D).

With respect to executive function, Freeze Frame performance was positively correlated with the NIH EXAMINER executive composite score (*r*=0.26; *P*=.02; see Figure 2E). Linear regression showed that Freeze Frame scores were significantly associated with the NIH EXAMINER executive composite

XSL•FO
RenderX

score ($F_{1,78}$=5.693; P=.02) and modestly accounted for 6.8% of the variance (R^2=0.068).

Missing Data

A total of 12 participants (13% of the sample) did not have mean accuracy scores that could be calculated from threshold scores due to administrative errors. Missing data were not imputed for baseline data of predefined exploratory outcomes data, and Freeze Frame associations reflect participants from the intent-to-treat population with complete data at baseline. There were no statistical differences between those with versus without Freeze Frame mean accuracy scores across demographic variables or NIH EXAMINER scores.

Discussion

Principal Findings

Freeze Frame assesses inhibitory control and, in this study, captured modest variance in overall executive function, with better Freeze Frame performance associated with better NIH EXAMINER performance. Older participants exhibited lower accuracy on the task compared to participants who were relatively younger. Educational background did not influence task performance. Extreme scores were minimal, suggesting that the assessment did not have notable ceiling or floor effects. The duration of the assessment was brief, at 4 minutes, and the SD suggests that completion times were stable and did not introduce significant variability due to differences in participant pacing. These results provide preliminary evidence that Freeze Frame may serve as a useful indicator of executive function and that healthy older adults are generally able to complete the assessment within a feasible administration time.

Performance on most cognitive assessments declines with age, largely due to the complex interplay of structural, functional, and neurochemical changes that occur in the aging brain [30]. While considerable individual differences exist, age-related reductions in cortical thickness, functional connectivity, and neuromodulatory control generally result in measurable decreases in cognitive function over one's lifespan, such as slower information processing and diminished executive functioning [7]. In contrast, assessments associated with crystallized intelligence, such as general content knowledge, vocabulary, and spelling, often remain stable or improve with age [31]. In this study, performance on the Freeze Frame assessment was found to decrease modestly with age, with relatively younger participants outperforming older participants. This pattern is consistent with the intended design of Freeze Frame to evaluate cognitive networks that are vulnerable to age-related decline [25-27,32,33].

In this study, women demonstrated significantly better performance than men on the Freeze Frame assessment, which is consistent with previous evaluations of Freeze Frame [34] as well as other validated assessments of executive function [35]. This finding aligns with prior research suggesting that women often exhibit better response inhibition and cognitive control, particularly in tasks requiring rapid decision-making using go/no-go paradigms [36]. These differences may be due to strategy differences [37] or may be influenced by sociocultural factors, including differences in early-life experiences and cognitive engagement. Neurobiological factors, including structural and functional variations in prefrontal networks, which play a critical role in executive function abilities, may also contribute to these observed performance disparities [38]. The findings underscore the importance of considering sex-based differences in cognitive assessments and highlight the need for further research to explore underlying mechanisms driving such differences and their implications for cognitive aging and disease vulnerability.

The relatively small proportion of variance in NIH EXAMINER scores explained by Freeze Frame (6.8%) suggests that the assessment captures limited aspects of the broader executive function construct. The modest shared variance highlights that most individual differences in executive function remain unexplained by Freeze Frame alone. This is not unexpected given that the NIH EXAMINER executive composite score integrates performance across multiple domains of executive function, whereas Freeze Frame is narrowly focused on a speeded version of inhibitory control. Freeze Frame may nevertheless serve as a useful, fully remote, and self-administered proxy within research or clinical screening contexts that rely on rapid evaluation. The findings reported further highlight the value of incorporating additional tasks using the same remote assessment methodology (such as those capturing other aspects of executive function like working memory) to develop a comprehensive evaluation of overall functioning.

Limitations

The generalizability of the study findings is constrained by the limited demographic diversity of the participant sample, which may not fully represent racial and ethnic minority groups [29]. As a result, the applicability of these findings to more diverse and lower-education populations remains uncertain, beyond preliminary evidence from a prior study involving a sample with greater racial and educational diversity [34], and warrants further investigation. Additionally, the age range of participants (65-83 years) was restricted due to the study's inclusion criteria, potentially limiting the ability to detect associations across a broader age spectrum. Another methodological limitation pertains to the Freeze Frame assessment specifically, where 13% of threshold scores could not be converted to mean accuracy, which reduced the analytic sample to 80 and likely diminished the statistical power of the analyses and may have attenuated the magnitude of observed effects.

Strengths

The INHANCE study is distinguished by several methodological strengths, including a rigorous study design, the implementation of stringent inclusion criteria, and adherence to a prespecified statistical analysis plan. A key advantage of this study is the use of a computerized assessment platform, which enhances standardization, reliability, and efficiency in data collection. Furthermore, the digital nature of the assessment allows for broad scalability, making it accessible to individuals across diverse geographic and socioeconomic backgrounds who have access to an internet-enabled device. This broad accessibility

XSL•FO
RenderX

supports the potential for large-scale implementation in both research and clinical settings pending further validation studies.

Conclusions

Findings from this study highlight the preliminary sensitivity of Freeze Frame to evaluate higher-order cognitive abilities. Freeze Frame may serve as a scalable behavioral proxy for detecting executive function deficits in aging populations. Future research should investigate whether lower performance on this task is an early indicator of heightened risk for cognitive decline. Additionally, further studies are needed to determine the utility of Freeze Frame in clinical populations known to exhibit executive dysfunction, such as individuals with attention deficit disorder, traumatic brain injury, or mild cognitive impairment. Examining the predictive validity of Freeze Frame in relation to real-world functional decline will be critical in establishing its clinical applicability as a screening tool.

Acknowledgments

Research reported in this publication was supported by the National Institute on Aging of the National Institutes of Health under awards R44AG039965 and 3R44AG039965-06S1, awarded to MA and TVV. We would like to thank the members of McGill's Clinical Research Unit and the software engineers and quality assurance team at Posit Science. This content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Data Availability

On May 31, 2026, the deidentified participant data from INHANCE (Improving Neurological Health in Aging via Neuroplasticity-Based Computerized Exercise) will be made available to verified academic and industry researchers. Investigators can conduct both confirmatory and exploratory analyses using LORIS (Longitudinal Online Research and Imaging System), a freely accessible, open-source data archive with provenance-sharing features. The dataset will be accessible at the INHANCE Study site on LORIS [39]

Authors' Contributions

MA and TVV are principal investigators. EdV-S is the site principal investigator. MA, EdV-S, and TVV were responsible for the concept and the design of the study. MA conducted the analyses and wrote the first version of the manuscript. SJG was the clinical trial manager at Posit Science and LO administered the assessments at McGill. MA, EdV-S, and ACdFP defined the data analysis plan. All authors read, edited and approved the final manuscript.

Conflicts of Interest

MA, TVV, and SJG are employees at Posit Science, the company that develops BrainHQ's brain training and assessment programs. They are shareholders of Posit Science stock. EdV-S, ACdFP, and LO declare no competing interests.

References

1. Royall DR, Lauterbach EC, Kaufer D, et al. The cognitive correlates of functional status: a review from the Committee on Research of the American Neuropsychiatric Association. J Neuropsychiatry Clin Neurosci 2007;19(3):249-265. [doi: 10.1176/jnp.2007.19.3.249] [Medline: 17827410]
2. Cahn-Weiner DA, Malloy PF, Boyle PA, Marran M, Salloway S. Prediction of functional status from neuropsychological tests in community-dwelling elderly individuals. Clin Neuropsychol 2000 May;14(2):187-195. [doi: 10.1076/1385-4046(200005)14:2;1-Z;FT187] [Medline: 10916193]
3. Jefferson AL, Paul RH, Ozonoff A, Cohen RA. Evaluating elements of executive functioning as predictors of instrumental activities of daily living (IADLs). Arch Clin Neuropsychol 2006 May;21(4):311-320. [doi: 10.1016/j.acn.2006.03.007] [Medline: 16814980]
4. Johnson JK, Lui LY, Yaffe K. Executive function, more than global cognition, predicts functional decline and mortality in elderly women. J Gerontol A Biol Sci Med Sci 2007 Oct;62(10):1134-1141. [doi: 10.1093/gerona/62.10.1134] [Medline: 17921427]
5. Tucker-Drob EM. Neurocognitive functions and everyday functions change together in old age. Neuropsychology 2011 May;25(3):368-377. [doi: 10.1037/a0022348] [Medline: 21417532]
6. Christ SE, White DA, Mandernach T, Keys BA. Inhibitory control across the life span. Dev Neuropsychol 2001;20(3):653-669. [doi: 10.1207/S15326942DN2003_7] [Medline: 12002099]
7. Ferguson HJ, Brunsdon VEA, Bradford EEF. The developmental trajectories of executive function from adolescence to old age. Sci Rep 2021 Jan 14;11(1):1382. [doi: 10.1038/s41598-020-80866-1] [Medline: 33446798]
8. Kramer JH, Mungas D, Possin KL, et al. NIH EXAMINER: conceptualization and development of an executive function battery. J Int Neuropsychol Soc 2014 Jan;20(1):11-19. [doi: 10.1017/S1355617713001094] [Medline: 24103232]
9. Kramer JH. Special series introduction: NIH EXAMINER and the assessment of executive functioning. J Int Neuropsychol Soc 2014 Jan;20(1):8-10. [doi: 10.1017/S1355617713001185] [Medline: 24161072]

XSL•FO
RenderX

10. Possin KL, LaMarre AK, Wood KA, Mungas DM, Kramer JH. Ecological validity and neuroanatomical correlates of the NIH EXAMINER executive composite score. J Int Neuropsychol Soc 2014 Jan;20(1):20-28. [doi: 10.1017/S1355617713000611] [Medline: 23764015]
11. Stiver J, Kramer JH. Executive abilities: methods and instruments for neurobehavioral evaluation and research. In: Kreutzer J, DeLuca J, Caplan B, editors. Encyclopedia of Clinical Neuropsychology: Springer International Publishing; 2017:1-3. [doi: 10.1007/978-3-319-56782-2_1890-2]
12. Staffaroni AM, Bajorek L, Casaletto KB, et al. Assessment of executive function declines in presymptomatic and mildly symptomatic familial frontotemporal dementia: NIH-EXAMINER as a potential clinical trial endpoint. Alzheimers Dement 2020 Jan;16(1):11-21. [doi: 10.1016/j.jalz.2019.01.012] [Medline: 31914230]
13. Attarha M, de Figueiredo Pelegrino AC, Toussaint PJ, Grant SJ, Van Vleet T, de Villers-Sidani E. Improving neurological health in aging via neuroplasticity-based computerized exercise: protocol for a randomized controlled trial. JMIR Res Protoc 2024 Aug 8;13(1):e59705. [doi: 10.2196/59705] [Medline: 39116435]
14. Munakata Y, Herd SA, Chatham CH, Depue BE, Banich MT, O'Reilly RC. A unified framework for inhibitory control. Trends Cogn Sci 2011 Oct;15(10):453-459. [doi: 10.1016/j.tics.2011.07.011] [Medline: 21889391]
15. Schall JD, Palmeri TJ, Logan GD. Models of inhibitory control. Philos Trans R Soc Lond B Biol Sci 2017 Apr 19;372(1718):20160193. [doi: 10.1098/rstb.2016.0193] [Medline: 28242727]
16. Rabi R, Vasquez BP, Alain C, Hasher L, Belleville S, Anderson ND. Inhibitory control deficits in individuals with amnestic mild cognitive impairment: a meta-analysis. Neuropsychol Rev 2020 Mar;30(1):97-125. [doi: 10.1007/s11065-020-09428-6] [Medline: 32166707]
17. Senkowski D, Ziegler T, Singh M, et al. Assessing inhibitory control deficits in adult ADHD: a systematic review and meta-analysis of the stop-signal task. Neuropsychol Rev 2024 Jun;34(2):548-567. [doi: 10.1007/s11065-023-09592-5] [Medline: 37300725]
18. Lipszyc J, Schachar R. Inhibitory control and psychopathology: a meta-analysis of studies using the stop signal task. J Int Neuropsychol Soc 2010 Nov;16(6):1064-1076. [doi: 10.1017/S1355617710000895] [Medline: 20719043]
19. Jacobson McEwen SC, Connolly CG, Kelly AMC, et al. Resting-state connectivity deficits associated with impaired inhibitory control in non-treatment-seeking adolescents with psychotic symptoms. Acta Psychiatr Scand 2014 Feb;129(2):134-142. [doi: 10.1111/acps.12141] [Medline: 23621452]
20. Robertson IH, Manly T, Andrade J, Baddeley BT, Yiend J. "Oops!": performance correlates of everyday attentional failures in traumatic brain injured and normal subjects. Neuropsychologia 1997 Jun;35(6):747-758. [doi: 10.1016/s0028-3932(97)00015-8] [Medline: 9204482]
21. Smallwood J, Davies JB, Heim D, et al. Subjective experience and the attentional lapse: task engagement and disengagement during sustained attention. Conscious Cogn 2004 Dec;13(4):657-690. [doi: 10.1016/j.concog.2004.06.003] [Medline: 15522626]
22. Ryan M, Martin R, Denckla MB, Mostofsky SH, Mahone EM. Interstimulus jitter facilitates response control in children with ADHD. J Int Neuropsychol Soc 2010 Mar;16(2):388-393. [doi: 10.1017/S1355617709991305] [Medline: 20003583]
23. Shaked D, Faulkner LMD, Tolle K, Wendell CR, Waldstein SR, Spencer RJ. Reliability and validity of the Conners' Continuous Performance Test. Appl Neuropsychol Adult 2020;27(5):478-487. [doi: 10.1080/23279095.2019.1570199] [Medline: 30793982]
24. Carson N, Leach L, Murphy KJ. A re-examination of Montreal Cognitive Assessment (MoCA) cutoff scores. Int J Geriatr Psychiatry 2018 Feb;33(2):379-388. [doi: 10.1002/gps.4756] [Medline: 28731508]
25. Van Vleet TM, DeGutis JM, Merzenich MM, Simpson GV, Zomet A, Dabit S. Targeting alertness to improve cognition in older adults: a preliminary report of benefits in executive function and skill acquisition. Cortex 2016 Sep;82:100-118. [doi: 10.1016/j.cortex.2016.05.015] [Medline: 27372902]
26. Van Vleet TM, Chen A, Vernon A, Novakovic-Agopian T, D'Esposito MT. Tonic and phasic alertness training: a novel treatment for executive control dysfunction following mild traumatic brain injury. Neurocase 2015;21(4):489-498. [doi: 10.1080/13554794.2014.928329] [Medline: 24984231]
27. Degutis JM, Van Vleet TM. Tonic and phasic alertness training: a novel behavioral therapy to improve spatial and non-spatial attention in patients with hemispatial neglect. Front Hum Neurosci 2010;4(60):60. [doi: 10.3389/fnhum.2010.00060] [Medline: 20838474]
28. FAQ. LORIS. URL: https://loris.ca/faq.html [accessed 2024-06-06]
29. Pelegrino ADF, Attarha M, Toussaint PJ, et al. Cholinergic neurotransmission in the anterior cingulate cortex is associated with cognitive performance in healthy older adults: baseline characteristics of the Improving Neurological Health in Aging via Neuroplasticity-based Computerized Exercise (INHANCE) trial. Neuroimage Rep 2025 Mar;5(1):100234. [doi: 10.1016/j.ynirp.2025.100234] [Medline: 40191405]
30. Craik FIM, Salthouse TA, editors. The Handbook of Aging and Cognition, 3rd edition: Psychology Press; 2007. [doi: 10.4324/9780203837665]
31. Salthouse TA. What and when of cognitive aging. Curr Dir Psychol Sci 2004 Aug;13(4):140-144. [doi: 10.1111/j.0963-7214.2004.00293.x]

XSL•FO
RenderX

32. Van Vleet TM, DeGutis JM. The nonspatial side of spatial neglect and related approaches to treatment. Prog Brain Res 2013;207:327-349. [doi: 10.1016/B978-0-444-63327-9.00012-6] [Medline: 24309261]
33. DeGutis J, Grosso M, VanVleet T, Esterman M, Pistorino L, Cronin-Golomb A. Sustained attention training reduces spatial bias in Parkinson's disease: a pilot case series. Neurocase 2016;22(2):179-186. [doi: 10.1080/13554794.2015.1088035] [Medline: 26360648]
34. Attarha M, Mahncke H, Merzenich M. The real-world usability, feasibility, and performance distributions of deploying a digital toolbox of computerized assessments to remotely evaluate brain health: development and usability study. JMIR Form Res 2024 May 13;8(1):e53623. [doi: 10.2196/53623] [Medline: 38739916]
35. Gaillard A, Fehring DJ, Rossell SL. A systematic review and meta-analysis of behavioural sex differences in executive control. Eur J Neurosci 2021 Jan;53(2):519-542. [doi: 10.1111/ejn.14946] [Medline: 32844505]
36. Weafer J. Sex differences in neural correlates of inhibitory control. In: de Wit H, Jentsch JD, editors. Recent Advances in Research on Impulsivity and Impulsive Behaviors: Springer International Publishing; 2020:73-89. [doi: 10.1007/7854_2020_146]
37. Gaillard A, Fehring DJ, Rossell SL. Sex differences in executive control: a systematic review of functional neuroimaging studies. Eur J Neurosci 2021 Apr;53(8):2592-2611. [doi: 10.1111/ejn.15107] [Medline: 33423339]
38. Garavan H, Hester R, Murphy K, Fassbender C, Kelly C. Individual differences in the functional neuroanatomy of inhibitory control. Brain Res 2006 Aug 11;1105(1):130-142. [doi: 10.1016/j.brainres.2006.03.029] [Medline: 16650836]
39. Inhance Study. URL: https://inhance.loris.ca/ [accessed 2025-06-18]

Abbreviations

INHANCE: Improving Neurological Health in Aging via Neuroplasticity-Based Computerized Exercise
LORIS: Longitudinal Online Research and Imaging System
NIH EXAMINER: National Institutes of Health Executive Abilities: Measures and Instruments for Neurobehavioral Evaluation and Research
SART: Sustained Attention to Response Task

Edited by A Mavragani; submitted 08.04.25; peer-reviewed by A Kairys, B Tosti; revised version received 09.06.25; accepted 11.06.25; published 14.07.25.

Please cite as:
Attarha M, Carolina de Figueiredo Pelegrino A, Ouellet L, Grant SJ, de Villers-Sidani E, Van Vleet T
Bringing Executive Function Testing Online: Assessment Validation Study
JMIR Form Res 2025;9:e75687
URL: https://formative.jmir.org/2025/1/e75687
doi:10.2196/75687

© Mouna Attarha, Ana Carolina De Figueiredo Pelegrino, Lydia Ouellet, Sarah-Jane Grant, Etienne de Villers-Sidani, Thomas Van Vleet. Originally published in JMIR Formative Research (https://formative.jmir.org), 14.7.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Acceptability of a Web-Based Financial Education Intervention for Latino Caregivers: Mixed Methods Evaluation

Susanna Mage[1], MA, PhD; Donna Benton[1], PhD; Kathleen Wilber[1], LCSW, PhD; Rocio Aguila Rodriguez[2]; David Silva[2], BS; Alexander Gonzalez[1], BS; Frank Puga[3], PhD; Kylie Meyer[2], MSc, PhD

[1]Leonard Davis School of Gerontology, University of Southern California, 3715 McClintock Avenue, Los Angeles, CA, United States

[2]Frances Payne Bolton School of Nursing, Case Western Reserve University, Cleveland, OH, United States

[3]School of Nursing, University of Alabama at Birmingham, Birmingham, AL, United States

Corresponding Author:
Susanna Mage, MA, PhD
Leonard Davis School of Gerontology, University of Southern California, 3715 McClintock Avenue, Los Angeles, CA, United States

Abstract

Background: Latino caregivers of persons living with dementia face distinct financial challenges, spending nearly half of their annual household income on caregiving expenses, almost twice as much as non-Latino caregivers. Cultural norms, limited access to financial resources, and underrepresentation in intervention research compound this economic burden. There remains a critical gap in evidence-based, culturally tailored programs designed to reduce financial stress among Latino caregivers.

Objective: This study examines the acceptability and perceived value of the Confidently Navigating Financial Decisions and Enhancing Financial Well-Being in Dementia Caregiving (CONFIDENCE) program, a culturally tailored, web-based psychoeducational intervention designed to reduce financial stress and improve financial well-being among Latino caregivers of persons living with dementia.

Methods: We applied a multimethod approach to evaluate the acceptability of the CONFIDENCE pilot. Following completion, 14 caregivers participated in in-depth, semistructured interviews, and 27 (39% of program participants) completed a 14-item satisfaction survey. The study was guided by a theoretical framework of acceptability composed of 7 domains: affective attitude, burden, ethicality, intervention coherence, opportunity costs, perceived effectiveness, and self-efficacy. Interview transcripts were analyzed independently by 4 coders (SM, RAR, DS, and KM) using thematic analysis, and survey responses were summarized using descriptive statistics.

Results: Participants overall reported satisfaction with the CONFIDENCE program. Qualitative analysis of interview transcripts revealed four themes: (1) the perceived need for financial intervention; (2) perceived intervention effectiveness, particularly in improving financial knowledge, communication skills, and mental health; (3) positive emotional responses to participation, with high praise for the program's group setting and cultural relevance; and (4) recommendations for intervention improvement, including shorter sessions and technology support. Survey results confirmed high acceptability, with 96% (n=26) of respondents reporting they enjoyed the program. All participants (n=27, 100%) said they would recommend CONFIDENCE to other Latino caregivers. Participants reported improved awareness of available resources, increased confidence in financial decision-making, and reduced financial stress. Caregivers appreciated the group setting, which allowed for mutual interaction and learning; discussions fostered emotional connection and peer learning. Participants praised the content's trustworthiness and relevance. Although most agreed that participation required minimal effort, barriers such as time constraints and technology challenges were noted.

Conclusions: Findings suggest that CONFIDENCE is acceptable among Latino caregivers of persons living with dementia. Culturally tailored content, group learning, and an emphasis on practical financial strategies were central to the program's success. Findings will inform program refinements, including limiting burden by addressing barriers while maintaining strengths such as group learning to optimize engagement. Considering the unique economic burdens Latino caregivers face, it is vital to develop and support interventions tailored to their unique needs. This study adds to the limited but growing literature on financial well-being interventions for Latino caregivers and highlights the value of culturally informed, community-driven approaches to supporting caregiver financial resilience.

(JMIR Form Res 2025;9:e70471) doi:10.2196/70471

KEYWORDS

informal caregiving; caregiver stress; dementia; disparities; financial strain; ethnicity; acceptability; intervention; Latino; Latina

XSL•FO
RenderX

Introduction

Background

Family caregivers play a crucial role in supporting older adults living with chronic illness and disability; yet, many face significant economic challenges. Caregivers of persons living with dementia experience substantial out-of-pocket expenses due to medical copayments, necessary home modifications, specialized caregiving supplies, personal care services, and home care. These costs precipitate far-reaching economic and psychological consequences, including financial strain and anxiety [1-3]. Households of individuals diagnosed with dementia have been shown to experience more than a 2-fold increase in out-of-pocket health care spending and a decline in net worth exceeding 60% within 8 years of diagnosis [2]. This kind of financial insecurity can have both immediate and long-term effects on caregivers' ability to meet basic needs, career development, generational wealth, and overall economic well-being [4,5].

The financial burden of caregiving is acutely evident among resource-limited populations disproportionately impacted by dementia [6], with recent findings showing that out-of-pocket costs and related financial strain disproportionately affect historically marginalized groups, including racial and ethnic minority populations [4]. Latino caregivers are particularly vulnerable to these financial pressures [7,8]. This is due in part to delayed access to services, more unpaid care responsibilities, and limited availability of culturally tailored support. Despite depending heavily on informal care and choosing more affordable health care options (ie, adult day care over nursing facilities) [8], Latino families of persons living with dementia still incur higher costs than non-Latino families [7,9]: on average, caregivers spend 26% of their annual household income on caregiving, whereas Latino caregivers spend nearly twice as much (47%) [3]. One study found that Mexican American caregivers contend with caregiving burdens that are intensified by socioeconomic constraints and disrupted life opportunities, contributing to greater financial strain and worse physical and mental health outcomes [8].

Understanding values and norms central to Latino culture is crucial for addressing needs and developing financial support programs. Cultural values that influence approaches to caregiving and spending patterns include *Famalismo* and *Marianismo* [7,10-12]. *Famalismo*—prioritizing family welfare over individual desires—can contribute to career compromises that reduce potential earnings when a caregiver provides direct care over formal services. *Marianismo*—the characterization of women as family-centered and selfless—may reduce household income by Latinas' prioritizing unpaid care roles over paid labor force participation. Within Latino families, caregiving expectations often fall disproportionately on women, which can limit their participation in the paid workforce and contribute to unequal caregiving burdens [7,8]. Caregivers under strain are more likely to consider skilled nursing facilities for their loved ones [13]. However, despite providing more intensive care, Latino caregivers historically prefer providing care in the community, possibly due to familial obligations and concerns about the quality of nursing facilities [8,11,14].

Strengthening caregivers' financial security is a national priority in the United States, reflected in the first-ever National Strategy on Caregiving [15]. Financial education and training interventions have been found to improve money management and the overall well-being of participants [16], and several studies have shown the positive impact of psychosocial interventions in easing caregiver burden and mitigating distress [7,16-18]. Interventions tailored to meet participant needs, preferences, and values increase satisfaction and engagement [8,19,20]. Despite increasing recognition of the unequal caregiving burden felt among Latino families, most evidence-based interventions have not been adapted for their cultural needs, and few address financial strain directly [21].

This Study

The Confidently Navigating Financial Decisions and Enhancing Financial Well-Being in Dementia Caregiving (CONFIDENCE) intervention was developed to reduce financial stress and lower Latino caregivers' out-of-pocket spending on care costs by displacing such spending with community and informal resources and facilitating help-seeking behaviors (personal communication by Dr Kylie Meyer, 2024). This study aims to test the acceptability of its pilot program using Sekhon's theoretical framework of acceptability (TFA) as a guide [20]. Evaluating acceptability can identify factors affecting adherence, execution, and outcomes, facilitating necessary adjustments to the intervention's design to boost its practicality and acceptance. The TFA breaks acceptability into 7 validated components: affective attitude, burden, perceived effectiveness, ethicality, intervention coherence, opportunity costs, and self-efficacy, defined further in Multimedia Appendix 1. We selected this framework to guide our study because it provides a comprehensive structure for assessing emotional, cognitive, and practical dimensions of intervention acceptability. Given the strong influence of cultural values, caregiving expectations, and financial vulnerabilities within Latino communities, it was important to use a framework that could not only capture general satisfaction with the program but also be in alignment with participants' values and perceived burdens. The TFA's multidimensional focus makes it well-suited to evaluate culturally tailored interventions like CONFIDENCE.

Methods

Study Design

This study used qualitative and quantitative methods to examine the perspectives of Latino communities caring for persons living with dementia who participated in CONFIDENCE. Data were collected through in-depth, semistructured qualitative interviews (n=14) and a 14-item quantitative satisfaction survey (n=27). The TFA guided the study design, including the interview guide, coding, and interpretation processes [20].

XSL•FO
RenderX

Intervention

Overview

The CONFIDENCE intervention is a culturally tailored 5-session psychoeducational program delivered via teleconference to up to 12 caregivers at a time. Sessions include the delivery of didactic content interspersed with practice exercises and discussions. The program was facilitated by interventionists who self-identified as Latino and, when delivered in Spanish, were native Spanish speakers. After each session, caregivers were encouraged to practice what they learned. Topics included navigating and seeking help from formal and informal resources, balancing employment and caregiving, and basic money management skills applied to caregiving (eg, using a budget to make financial decisions related to caregiving). To ensure that content was culturally tailored, the investigator team drew upon findings from qualitative interviews with Latino family caregivers about their experiences managing caregiving-related financial strain [22] alongside ongoing guidance from a community advisory council. The council provided input on key elements, including the length of discussion sessions and culturally appropriate phrasing to describe practice activities. A more detailed description of the cultural tailoring process can be found in Meyer et al [23]. The CONFIDENCE intervention was delivered at a community-based organization 11 times, including twice in Spanish, from 2022 to 2023.

Steps to Support Intervention Fidelity

Intervention fidelity was supported through the use of a structured facilitator guide, standardized slide decks to guide group discussion sessions, and template reminder emails encouraging participants to complete recommended between-session activities. All facilitators had at least a bachelor's degree and prior experience providing support services to family caregivers within the service setting where the intervention was delivered. The lead facilitator, who helped develop the intervention and had prior personal caregiving experience, led the training for other facilitators. This training included role-playing of the intervention delivery, supervised delivery of at least 2 sessions prior to independently delivering sessions, and fidelity procedures. To monitor fidelity, an independent observer used a standardized checklist to review recordings from 5 randomly selected sessions. All observed sessions were delivered in the intended sequence, with at least 80% of intervention steps completed according to the facilitator guide. Facilitators also submitted attendance reports to the evaluation team, which were reviewed to help monitor participant engagement. Weekly evaluation team meetings, including the lead facilitator, and ongoing review of attendance and fidelity reports by the principal investigator (KM) helped address delivery issues, such as low participation during sessions.

Data Collection

Eligibility and Recruitment

The CONFIDENCE program was delivered as a part of routine services at a community-based organization that regularly provides caregiver education and support. Approximately 60% of caregiving clients at this organization identify as Latino, and nearly 40% care for persons living with dementia. Caregivers were primarily informed about CONFIDENCE through email "blasts" and staff referrals. In total, 20 caregivers participated in an optional pre- and posttest study, which required registration through the organization's web-based event platform. To support study recruitment, the program was also promoted more broadly, including flyers and social media posts. For more information on the recruitment methods used in the feasibility study, see Meyer et al [23].

A satisfaction survey with 14 items was sent by email to caregivers who registered for CONFIDENCE on the web-based event platform (ie, convenience sample) and met the following eligibility criteria: self-identified as Latino, provided care to persons living with dementia, 18 years or older, and attended at least 1 of the 5 CONFIDENCE sessions. Those who completed the satisfaction survey were also asked if they would be willing to complete a qualitative interview. A purposeful sample of 14 caregivers was recruited via email for qualitative interviews, including those who completed the satisfaction survey and indicated that they would be willing to be contacted, as well as feasibility study participants. Participants were selected to include caregivers of various genders, ages, geographic locations, and kin relationships with the care recipient.

Interview Guide

The TFA and recent literature on Latino caregivers informed the qualitative interview's open-ended questions [20,22]. The research team reviewed the interview guide (Multimedia Appendix 2), which discussed the program's purpose, the skills caregivers believed they acquired, how participants perceived it helped protect financial well-being, which parts were most helpful, and barriers and challenges to participation.

Interviews

Participants provided verbal consent to participate in individual interviews, which were conducted in Spanish or English, based on the participant's language preference, via Zoom (Zoom Video Communications; 2022 - 2023). All interviews were recorded and then transcribed by a professional transcription service, which also translated Spanish transcriptions. The interview method followed the recommendations of Baker and Edwards [24] to include at least 12 interviews until no new themes or codes emerged (ie, thematic saturation) [25], which was achieved after the 10th interview.

Satisfaction Survey

A 14-item satisfaction survey was sent to participants after the last session of the program (Table 1). The survey assessed the acceptability, overall perception of the program, and participants' evaluation of the information presented. Satisfaction levels were rated using a 5-point Likert scale (1=strongly disagree, 2=disagree, 3=neutral, 4=agree, and 5=strongly agree). For each item, responses were categorized as "satisfied" if the participant selected "agree" or "strongly agree." This survey was a newly developed tool grounded in Sekhon's TFA [20].

XSL•FO
RenderX

Table . Survey items used to assess the acceptability of the CONFIDENCE[a] intervention, aligned with seven domains of Sekhon's theoretical framework of acceptability.

Construct of the theoretical framework of acceptability	Questions asked
Affective attitude: how an individual feels about the intervention	• I felt comfortable and included in the CONFIDENCE program. • I felt comfortable sharing about my caregiving experience and personal finances during the CONFIDENCE program. • I enjoyed participating in the CONFIDENCE program. • I trusted the information I received in the CONFIDENCE program.
Burden: the perceived amount of effort that is required to participate in the intervention	• It was easy to understand the information presented during the CONFIDENCE program. • The pace of the program was just right (ie, not too fast and not too slow). • I found it easy to access the CONFIDENCE program on the web, including using Zoom.
Ethicality: the extent to which the intervention has a good fit with the individual's value system.	• The information presented in the CONFIDENCE program applied to my caregiving situation.
Intervention coherence: the extent to which the participant understands the intervention and how it works	• I knew who to contact if I had any questions about the CONFIDENCE program. • Overall, I found the program to be interactive.
Opportunity costs: the extent to which benefits, profits, or values must be given up to engage in the intervention	• Overall, I found the program to be convenient to attend.
Perceived effectiveness: the extent to which the intervention is perceived as likely to achieve its purpose	• I learned from the other caregivers whom I attended the program with. • I would recommend for specifically other Latino or Latinx or Hispanic caregivers to participate in the CONFIDENCE program.
Self-efficacy: the participant's confidence that he or she can perform the behaviors required to participate in the intervention	• I found that take-home assignments added to my learning.

[a]CONFIDENCE: Confidently Navigating Financial Decisions and Enhancing Financial Well-Being in Dementia Caregiving.

Data Analysis

We applied thematic analysis to examine interview transcripts [26]. Codes were devised to address each component of acceptability from Sekhon's TFA, facilitating triangulation between qualitative findings and survey responses (Table 2). Four research team members (SM, RAR, DS, and KM) independently coded transcripts using NVivo (version 1.7.1; Lumivero). After initial coding, the team met with the lead coder (SM) for peer debriefing sessions. Discrepancies identified were resolved through discussion after each round, leading to iterative refinements of the codebook. As new codes were identified, like those related to culture, they were incorporated into the codebook and retrospectively applied to earlier transcripts. New codes and unresolved discrepancies were discussed with KM, who was available to adjudicate coding disagreements; however, no such adjudication was ultimately required. To assess participant satisfaction, we summarized survey responses using means and percentages.

XSL•FO
RenderX

Table . Codebook used for thematic analysis of qualitative interviews with Latino caregivers of persons living with dementia, grounded primarily in Sekhon's theoretical framework of acceptability [20], with a supplemental inductive code to capture culturally specific relevance.

Codebook	Definition
Affective attitude	Caregiver expresses how they feel about the intervention.
Burden	Caregiver describes the perceived amount of effort that was required to participate in the intervention.
Ethicality	Caregiver indicates that the intervention is a good fit with their value system.
Intervention coherence	Caregiver shows an understanding of the intervention and how it works.
Opportunity costs	Caregiver reports that they had to give up benefits, profits, or values to engage in the intervention.
Perceived effectiveness	Caregiver perceives the intervention as likely to achieve its purpose.
Self-efficacy	Caregiver's confidence that they can perform the behaviors required to participate in the intervention.
Relevance of Latino cultural and traditional values	Caregiver describes caregiving in relation to a sense of duty to their family (ie, Familialism) or traditional gender roles.

Ethical Considerations

The University of Southern California institutional review board (IRB) reviewed the study protocol and issued an exempt determination (UP-23‑00020) for the secondary analysis of deidentified data. Data collected from the primary study were approved under 2 separate protocols by the Case Western Reserve University IRB: an expedited protocol (STUDY20221010) for caregivers enrolled in the optional pre- and posttest study and an exempt protocol (STUDY20221415) for caregivers completing the CONFIDENCE program outside the pre- and posttest study. Caregivers who were eligible for the pre- and posttest study provided written consent to participate prior to enrollment. For participants not enrolled in the study who were invited to complete the satisfaction survey, consent was implied upon survey completion following a review of an information sheet. Caregivers who participated in qualitative interviews gave verbal consent after reviewing a study information sheet. All consent forms and information sheets stated that deidentified data may be retained for future use. The University of Southern California IRB did not require additional consent for the use of deidentified data shared from Case Western Reserve University. Participants in the optional pre- and posttest study received up to $25 USD in Amazon gift card value, which included compensation for completing the satisfaction survey. Caregivers who participated in a qualitative interview received a $25 USD gift card, and those not enrolled in the study but who completed a satisfaction survey were given the option to provide their email address to receive a $25 USD gift card.

Results

Sample Characteristics

In total, 14 Latino caregivers of persons living with dementia participating in CONFIDENCE between May 2022 and September 2023 consented to an interview following completion of the program (Table 3); 11 (79%) in English and 3 (21%) in Spanish. The mean age of participants was 60 (SD 7) years. Most were women (95%). Nearly two-thirds (64%) had attended at least 3 CONFIDENCE sessions; roughly half (44%) completed all sessions.

XSL•FO
RenderX

Table . Demographic characteristics of Latino caregivers of persons living with dementia (n=14) who participated in qualitative interviews following the CONFIDENCE[a] intervention.

Age (years)	Gender	Below 250% FPL[b]	Language	Sessions attended
52	Woman	Yes	English	4
66	Woman	No	English	1
62	Woman	No	English	5
56	Woman	Yes	English	4
56	Woman	No	English	5
54	Woman	No	English	4
57	Woman	Yes	Spanish	5
46	Woman	__[c]	Spanish	5
63	Woman	—	English	2
73	Man	No	English	3
59	Woman	Yes	Spanish	2
63	Woman	—	English	4
63	Woman	—	English	4
68	Woman	Yes	English	5

[a]CONFIDENCE: Confidently Navigating Financial Decisions and Enhancing Financial Well-Being in Dementia Caregiving.

[b]FPL: federal poverty level.

[c]Missing data.

Thematic Insights

Interview findings revealed four key themes: (1) perceived need for financial intervention, (2) perceived effectiveness of the intervention, (3) positive responses to participation, and (4) recommendations to improve the intervention.

Theme 1: Perceived Need for Financial Intervention

The first theme characterized the perceived need for a program like CONFIDENCE to address financial well-being and lower out-of-pocket care costs. Applying the TFA, this theme primarily reflects "intervention coherence" [20].

Caregivers Found the Content Applicable

Caregivers appreciated building financial management skills through CONFIDENCE, recognizing its importance: "It's to help caregivers, especially those of us who have not managed our finances on our own, teach us and guide us on how to begin the process of taking over some of the skills of managing our finances" (Participant 3). They expressed eagerness for such a program. "It's the class I was hoping for a long time," shared one participant, "it was my New Year's resolution to be more financially confident" (Participant 4). The unpredictable progression of dementia highlighted the need to save for future care costs:

> *Dementia can go on for many years. It doesn't matter your financial situation. People have no clue how long it runs, what home care costs, or the financial issues if you need placement.* [Participant 14]

CONFIDENCE Is Relevant to the Latino Community

During interviews, caregivers acknowledged the program's benefit for Latino families due to a perceived gap in financial knowledge, partially related to the division of household tasks between genders. They emphasized its value to women, who are often expected to be caregivers:

> *It is on a very personal level for us Mexican-Americans. Typically, the man takes over the finances ... so this program is very important ... whether it's the wife or daughter, so she can learn how to handle finances.* [Participant 3]

One caregiver felt the program could "help primarily Hispanic families to deal with strategies for why our out-of-pocket expenses are so much higher than other groups" (Participant 11). Caregivers valued the program's approach, which respects multiple ways of providing care rather than questioning care choices.

> *There's a reason why we pay more out-of-pocket than other people. It's a different mindset, and [CONFIDENCE] doesn't say, "Stop doing that." "Why are you doing what you're doing?" Instead, it gives strategies and ideas.* [Participant 11]

Theme 2: Perceived Effectiveness of the Intervention

Theme 2 reflected the TFA's categories of "perceived effectiveness" and "self-efficacy" [20]. Analyses of interview transcripts indicated that participants believed they acquired new knowledge and skills to help them manage their financial situation.

XSL•FO
RenderX

Caregivers Felt CONFIDENCE Improved Their Knowledge of Finances and Resources

Participants, especially those new to managing household finances, shared increased awareness of spending and praised CONFIDENCE for making financial well-being information accessible: "This information is amazing. It's here for us to dive in and have the confidence to explore it. This class did that; it gave me the confidence" (Participant 3). The program empowered caregivers to think of innovative solutions to financial challenges, as they became more aware of the wide range of available resources: "I was very surprised finding out how many resources are available" (Participant 6).

> *We didn't know that [resources] were available, and, unfortunately, we didn't know this was part of the stress that all caregivers experience when the sick person depends on you.* [Participant 8]

Several caregivers provided examples of successfully connecting with resources they learned about, including In-Home Supportive Services (Participant 5), adult daycare centers (Participant 1), and Medicaid workshops (Participant 12). Knowledge of community resources helped some caregivers offset the costs of caregiving:

> *I got so much help with providing [equipment] for my mom. I got a chair for the bath and grab-hold bars. I wasn't aware that I qualified for $250 to buy stuff for my mom.* [Participant 5]

Caregivers Felt the Program Improved Their Financial Well-Being

Strategies covered in the program to help manage care-related costs included tracking expenses, using a budget to make care-related decisions, asking for help from external resources, and setting financial goals.

> *It gave different strategies that I can modify for what I need. Everybody is gonna be different, but the great thing that it gave was strategies and tips, and you use what you can. I came from a pretty high place, but this put me over the top.* [Participant 11]

Caregivers noted that the skills gained improved their financial habits: "It has helped me a lot to manage my expenses more reasonably" (Participant 7). Several caregivers credited goal-setting activities for improving finances; clear, achievable targets fostered savings and debt reduction:

> *I got my tax return. Paid off a credit card. Paid down two credit cards. Put money in my emergency thing. I knocked off three goals ... It was awesome to break down goals in a way where I knew for sure what I was gonna do and how it was gonna make me feel as a result.* [Participant 5]

Learning Communication Skills Helped Caregivers Get Their Needs Met

Learning communication skills and engagement strategies, such as how to seek help from service providers, helped caregivers access support to offset care-related expenses.

> *This class gave me the ability to retool how I speak to someone who would be able to help me with different resources to help my mom and me and language better to outside resources to help us as I move forward with her.* [Participant 4]

CONFIDENCE taught caregivers techniques for effectively discussing support needs with their social circle. Many successfully sought family assistance after learning to ask for help through the program: "It helped me to plan that I had to get help from my family ... And so, I did speak to my family, and now I got a little bit more help from them" (Participant 6). Skills related to help-seeking are particularly notable, given the perceived cultural reticence around money talk within Latino families, a "stigma," as one caregiver described (Participant 4).

Caregivers Felt the Program Improved Mental Health

Interviews showed that participation in CONFIDENCE was believed to alleviate financial stress and improve mental and emotional well-being.

> *I was sobbing so much. Then, once I started applying, doing the little exercises, and reading the material, something shifted in me. Something had gone to sleep for a while ... And what it did was it–it recharged it. It woke that problem-solving part of my brain again.* [Participant 5]

CONFIDENCE reinforced self-confidence and provided assurance: "It reassured me, yeah, you are going in the right direction" (Participant 10). Another caregiver echoed, "It helped reassure me, hey, you have these skills. You can do that" (Participant 12). Caregivers felt validated in their decision-making and equipped to handle surprises and assert their needs.

Theme 3: Positive Responses to Participation

In alignment with the "ethicality" and "affective attitude" TFA domains [20], caregivers expressed a favorable opinion toward the intervention, indicating that it aligned well with their values and could benefit other caregivers.

Caregivers Enjoyed Participating in CONFIDENCE

Caregivers expressed satisfaction with the program and that they would recommend other caregivers to enroll: "This class is incredible. It's an amazing gift because of this class" (Participant 4). "It was very successful for those of us in the class," said one caregiver. "I think it would be great for any caretaker" (Participant 1). Others conveyed a desire to re-enroll and to continue with training.

> *Honestly, there's nothing negative to say. I'd like to take the class again after a year 'cause there's always new things going on. I try to take classes more than once because I always learn something.* [Participant 10]

Caregivers Valued the Group Discussions and Appreciated the Flexibility of the Program Design

Most caregivers felt solidarity with other participants and that this collective experience added to the program's perceived

XSL•FO
RenderX

impact. Participants reported learning from hearing about others' financial problem-solving strategies.

> *Somebody shared a skill that was really good. Every time they get a $10 bill, they put it away for a whole year ... I started doing something like that with larger denominations, and I've saved a few hundred dollars.* [Participant 10]

Group sessions helped caregivers recognize similar challenges faced by others; they were not alone: "I was feeling stressed and sad because [I thought] I was the only one, but hearing other stories and learning that there are resources to help, besides family, it helped me emotionally" (Participant 6). This awareness fostered emotional support within the group, making participants more receptive to sharing and accepting financial guidance.

The program's flexibility, which allows caregivers to access materials at their convenience both now and in the future, was well-received: "My mother's dementia goes in stages. I don't need this right now, but I am going to need it, and now I have a place to look when that moment comes" (Participant 7).

Theme 4: Recommendations to Improve the Intervention

Caregivers were asked to reflect on their experiences and provide feedback, offering insights into potential modifications and improvements to the intervention. This feedback elucidated potential "burden" and "opportunity costs" [20] associated with participation.

Help Caregivers Manage the Demands on Time

Caregivers reported difficulty attending weekly discussion sessions: "Sometimes it's difficult to come to class because our loved ones need our care," shared one caregiver. Another caregiver shared that the length of sessions posed a barrier.

> *Two hours is hard for me. I can't get that kind of time away from Mom. It would be better, I think, like an hour and a half or even an hour, but an hour might not give you enough time to get that energy going, you know?* [Participant 11]

Considering their already busy schedules, some caregivers found that the time required for take-home assignments was burdensome. They felt that the inability to complete tasks outside the designated course time could deter caregivers from joining future sessions.

> *Sometimes people are just burnt out. That might be what causes people to fall out of coming, because they feel like, "I was supposed to do an assignment, and I didn't do it. I'm embarrassed."* [Participant 5]

Facilitate Engagement During Group Sessions

Although participants were enthusiastic about interacting with fellow caregivers, some felt that the facilitators could have been more engaging: "I had the experience of being more talked at than engaged. It was a lot of reading off the presentation ... You have to be engaged to be engaging" (Participant 2). Some participants observed a reluctance among peers to participate in discussions. "Trying to get most of the people engaged was rough," said one caregiver, "but it's always gonna be rough. Some people are gonna sit there and listen, and pick things up as opposed to getting involved" (Participant 9).

Address Technology-Related Issues

Caregivers with limited technological proficiency encountered challenges that hindered participation in the web-based program: "Since I haven't learned to use a computer very well, I missed several sessions" (Participant 13). Other technological barriers affecting program delivery, like poor audio quality during video-sharing, also impeded caregivers' engagement. "We want to get every piece that they're saying and discussing. If the audio is poor, it loses it" (Participant 4).

Satisfaction Survey

Following the intervention, 27 of 69 (39%) caregivers who participated in the CONFIDENCE program completed satisfaction surveys. Most were women (78%) and lived with the care recipient (70%), 74% were children or children-in-law, and 26% were spouses or partners. In total, 25 (93%) respondents identified that they had served as the primary caregiver for over 2 years. Near half (41%) provided 15 or more care hours per week; 48% were employed. Additional respondent characteristics can be found in Table 4.

XSL•FO
RenderX

Table . Demographic characteristics of Latino caregivers of persons living with dementia (n=27) who completed a satisfaction survey following participation in the CONFIDENCE[a] program.

Gender	State	Marital status	Education	Employment status	Relationship to CR[b]	Live with CR	Hours of care per week	Years of care
Woman	CA	Partner	Some college	Working (self-employed)	Parent or PIL[c]	No	4	2 - 5
Man	AZ	Married	College or more	Retired	Spouse or partner	Yes	8	1 - 2
Woman	CA	Never married	College or more	Working (paid employee)	Parent or PIL	Yes	6	5+
Woman	CA	Divorced	Some college	Not working—looking for work	Parent or PIL	Yes	24	2 - 5
Man	CA	Married	College or more	Retired	Parent or PIL	No	13	1 - 2
Woman	CA	Married	College or more	Working as a paid employee	Parent or PIL	No	2	5+
Man	TX	Never married	College or more	Not working—other	Parent or PIL	Yes	24	__[d]
Woman	TX	Married	College or more	Retired	Spouse or partner	Yes	18	5+
Woman	CA	Divorced	College or more	Working, self-employed	Parent or PIL	Yes	24	2 - 5
Woman	CA	Married	Some college	Full-time student or trainee	Parent or PIL	No	12	5+
Woman	CA	Married	Some college	Working as a paid employee	Parent or PIL	Yes	6	2 - 5
Woman	NJ	Never married	College or more	Not working—looking for work	Parent or PIL	Yes	1	2 - 5
Woman	CA	Separated	Some college	Working, self-employed	Parent or PIL	Yes	17	5+
Woman	—	Never married	Some college	Working as a paid employee	Parent or PIL	No	12	2 - 5
Man	CA	Married	HS[e] or less	Retired	Spouse or partner	Yes	24	2 - 5
Woman	CA	Widowed	Some college	Working as a paid employee	Parent or PIL	Yes	16	1 - 2
Woman	—	Never married	Some college	Disabled	Parent or PIL	Yes	15	5+
Woman	CA	Married	College or more	Working as a paid employee	Parent or PIL	No	2	5+
Woman	CA	Married	College or more	Retired	Spouse or partner	Yes	16	5+
Man	CA	Married	Some college	Working as a paid employee	Parent or PIL	No	4	2 - 5
Woman	—	Married	Some college	Retired	Spouse or partner	Yes	5	2 - 5
Woman	CA	Never married	HS or less	Working as a paid employee	Parent or PIL	Yes	—	2 - 5
Woman	AZ	Never married	Some college	Retired	Parent or PIL	Yes	12	5+

XSL•FO
RenderX

Gender	State	Marital status	Education	Employment status	Relationship to CR[b]	Live with CR	Hours of care per week	Years of care
Woman	CA	Married	Some college	Retired	Spouse or partner	Yes	24	5+
Woman	AZ	Married	Some college	Retired	Parent or PIL	No	4	1 - 2
Woman	—	Never married	HS or less	Working as a paid employee	Parent or PIL	Yes	—	2 - 5
Man	CA	Married	Some college	—	Spouse or partner	Yes	24	2 - 5

[a]CONFIDENCE: Confidently Navigating Financial Decisions and Enhancing Financial Well-Being in Dementia Caregiving.

[b]CR: care recipient.

[c]PIL: parent-in-law.

[d]Missing data.

[e]HS: high school.

Caregivers expressed high satisfaction with CONFIDENCE's content, setting, and delivery. All respondents (27/27, 100%) agreed (ie, responding "agreed" or "strongly agreed") that they would recommend the program to all caregivers and specifically to Latino caregivers. All respondents (27/27, 100%) agreed that the information presented was trustworthy and applied to their caregiving situation and that they felt comfortable sharing personal experiences and financial details. Nearly all caregivers (26/27, 96%) enjoyed participating in the program and found the content easy to understand. Most respondents (25/27, 93%) found CONFIDENCE easy to access via Zoom, with nearly 90% (24/27, 89%) responding that it was convenient to attend the group discussion sessions. Additionally, most respondents (25/27, 93%) agreed that take-home assignments added to their learning. A full summary of the survey results is presented in Figure 1.

Figure 1. Ratings from 27 Latino caregivers of persons living with dementia on key program acceptability domains, based on a 14-item postintervention satisfaction survey following participation in the CONFIDENCE intervention. The "strongly disagree" response option was not selected by any participant and is therefore excluded. CONFIDENCE: Confidently Navigating Financial Decisions and Enhancing Financial Well-Being in Dementia Caregiving.

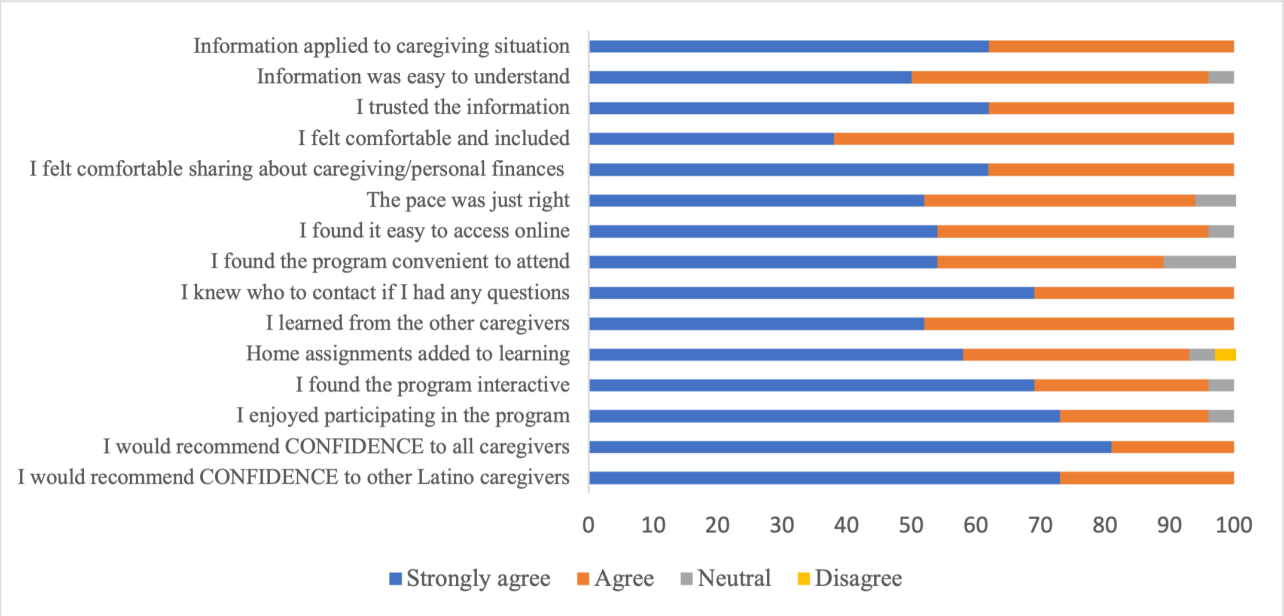

Discussion

Principal Findings

This study examined the acceptability of a financial psychoeducation program, CONFIDENCE, for Latino caregivers of persons living with dementia by assessing its perceived value, effectiveness, and participants' overall satisfaction. Feedback was gathered on the program's structure, relevance, most beneficial aspects, and suggestions for improvement. Caregivers endorsed the program, reporting increased financial awareness and high satisfaction with its content and delivery. Quantitative survey data corroborated qualitative findings, indicating that caregivers found the intervention engaging, understandable, relevant, and trustworthy (Figure 1). The use of an underlying framework (Sekhon's TFA) facilitated the evaluation of the overlap between qualitative and quantitative findings. Survey responses and interviews reflected comfort and willingness to discuss caregiving experiences and financial situations. In

XSL•FO
RenderX

addition, participants shed light on areas of the intervention that might need modification or improvement.

Financial Interventions for Caregivers

Caregiver interventions seldom address financial well-being [27], despite evidence that most caregivers are unprepared for health care expenses not covered by insurance. The CONFIDENCE program targeted caregivers from the Latino community who generally experience even higher levels of financial strain than the average caregiver population [3]. Content for the CONFIDENCE financial psychoeducational intervention was developed based on interviews with Latino caregivers and iterative feedback from a community advisory board. These strategies followed best practices in culturally tailoring interventions and likely enhanced the program's acceptability by ensuring that content was relevant to caregivers [28-30]. CONFIDENCE also incorporated proven strategies such as service systems and resource navigation, financial education, and asset-building to alleviate financial strain and enhance financial well-being [27].

Acceptability of Psychoeducation Interventions for Caregivers

Despite different content and aims, caregiver satisfaction reports align with findings from the literature, highlighting the benefits of group-based programs for interactive problem-solving regarding preparing for the future, identifying and reorganizing the support team, and self-care [31]. Findings of high satisfaction suggest the perceived value of interactive problem-solving among caregivers, resonating with results from similar interventions for caregivers [31-33]. Communication skills training is considered essential for caregivers, as it helps strengthen emotional connections with other caregivers, family, and friends [34]. Including this type of training in CONFIDENCE fostered relationships and understanding between the caregivers during group sessions—and with their network of friends and family outside of the intervention to promote help-seeking behaviors. Active participation and skill-building increase financial and self-confidence among caregivers, supporting the TFA [20].

Previous studies have established the appeal of web-based interventions for family caregivers [32,33,35]. Although previously reported challenges of web-based interventions, like time constraints and scheduling conflicts, were also found in this study, participants were satisfied with the overall design. Like other digital group-based interventions, such as TeleSavvy, CONFIDENCE prioritized mental health, which helps create a safe, positive learning environment [36]. Mirroring the advantages of such programs, the hybrid model blended web-based group interactions with the flexibility of asynchronous delivery. Our research supports findings that peer support is beneficial for managing the complexities of financial well-being among caregivers, as it encourages open discussions on financial issues, enhancing the group's collective problem-solving capacity [31,32]. CONFIDENCE's flexible format, which allows caregivers to engage with materials at their own pace, was well-received and well-suited to varying schedules. The design facilitated independent learning, underscoring the effectiveness and practicality of self-paced learning models.

Culturally Tailored Interventions for Latino Caregivers

Findings indicate that CONFIDENCE is an acceptable culturally tailored financial education program for Latino caregivers, a demographic with higher-than-average out-of-pocket care costs and a greater likelihood of limited formal financial engagement [3,37]. Latino caregivers may also face unique financial strains, often shouldering out-of-pocket expenses not covered by Medicare or Medicaid [7,11]. The US Latino population is both increasing and aging, adding to increases in dementia. Combined with lower use of formal care services and barriers to accessing care, overall expenses for Latino persons living with dementia and their caregivers are expected to increase at a faster rate than for non-Latino persons living with dementia [7,9]. There is a need for culturally targeted financial programs to address their higher levels of out-of-pocket costs [21,22,38].

Research acknowledges the importance of culturally informed support for caregivers belonging to communities where caregiving is heavily influenced by ethnic identity and cultural values [10,11,38,39]. However, many existing programs fail to meet the unique needs of diverse ethnic groups [21,40]. Preparing families to navigate available resources and financial decisions could mitigate financial risks while aligning with cultural norms about caregiving within the Latino community [7,11]. CONFIDENCE addresses the unique challenges of juggling caregiving with financial duties while acknowledging the preference for in-home care. Caregivers in this study reported that the program addressed the needs of Latino family caregivers by providing strategies to manage high costs, such as communicating with other family members and acknowledging cultural norms around caregiving financial decisions.

Improving Future Intervention Delivery

Despite enthusiasm for the program, lower-than-expected attendance and completion rates suggest that caregivers may experience time barriers to participation [33]. Previous research suggests that live-in caregivers with intense caregiving duties lack time for web-based interventions [31]. Shorter or fewer sessions may increase participation and commitment. Further, paralleling other web-based studies, technical difficulties, audio issues, and schedule clashes impacted user experience and potential program adoption [33]. Caregivers also encountered logistical challenges that hampered their ability to complete assignments. Helping caregivers identify strategies to complete homework could foster greater program acceptance and treatment adherence. Examining participants' expectations before delivering the program may benefit future implementations of CONFIDENCE [41].

Limitations

Despite high acceptance levels among respondents, only 39% of caregivers who completed the CONFIDENCE intervention answered satisfaction survey questions. Moreover, the investigators did not record data on the proportion of participants who accepted qualitative interview invitations. A high

XSL•FO
RenderX

nonresponse rate and lack of information about the qualitative response rate hinder a comprehensive assessment of caregiver experiences, as it is unknown whether survey completers had similar experiences to noncompleters. Future studies should prioritize strategies to enhance survey completion rates for a more representative analysis of the intervention's impact, including providing a payment incentive for satisfaction surveys. The demographic composition of our sample also presents limitations. Most participants were women, such that male caregiver perspectives may not be adequately represented.

Additionally, most sessions were delivered in English, and most caregivers chose to complete interviews in English. A deeper analysis of language preferences and cultural differences could provide valuable insights. Coding in Spanish could uncover subtleties that go unnoticed, thus capturing nuances lost in translation and enriching the depth of our findings. Finally, while web-based delivery offers the benefit of reaching caregivers across geographic regions, it imposes a constraint by requiring a certain level of technological proficiency, restricting who can participate. Our study may not capture the acceptability experiences of caregivers without internet access, whose perspectives on financial circumstances may differ.

Finally, variation in session attendance and the voluntary nature of participation in surveys and interviews may have introduced self-selection bias. Participants who completed the evaluation components may have been more motivated or held more favorable views of the intervention. Missing demographic data in some survey responses further limits the ability to fully characterize the participant population. These factors should be considered when evaluating the generalizability and transferability of findings to the broader Latino caregiver population.

Conclusions

Findings from this study suggest that the CONFIDENCE intervention is acceptable to Latino caregivers of persons living with dementia who seek to enhance their financial well-being and reduce out-of-pocket caregiving expenses. They also identified areas of improvement to make the program more accessible and more effective. This study contributes to the small but growing body of research on financial interventions designed to support caregivers, particularly within the Latino community. These caregivers have a higher risk of poor financial health while providing care, especially considering the rising rates of dementia among the Latino community [42]. Using culturally relevant psychoeducational interventions can enhance financial resilience and address the disproportionate costs borne by Latino caregivers.

Acknowledgments

The authors wish to acknowledge the support and guidance from the Confidently Navigating Financial Decisions and Enhancing Financial Well-Being in Dementia Caregiving Community Advisory Council members for their input in developing this program. The authors also want to thank community members who helped to develop intervention materials, including Mabel Brown, Sonia Rodriguez, Carol Bertche, JD, and Aleida Villareal, JD. This work was supported by the AARP Foundation (INC-2020-07-002 to DB and KM), the Administration for Community Living (90CGPS0005-01-00 to KW), and the National Institute on Aging (R01AG072569 to FP and 5P30AG024978, Kaye [pilot awardee: KM]). The content does not necessarily represent the official views of the AARP Foundation, the Administration on Aging, nor the National Institute on Aging.

Data Availability

The qualitative data generated and analyzed during this study are not publicly available due to the small sample size and the sensitive nature of the information shared by participants but are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The 7 component constructs of Sekhon's theoretical framework of acceptability.
[DOCX File, 18 KB - formative_v9i1e70471_app1.docx]

Multimedia Appendix 2
Interview guide.
[DOCX File, 18 KB - formative_v9i1e70471_app2.docx]

References

1. Brodaty H, Donkin M. Family caregivers of people with dementia. Dialogues Clin Neurosci 2009;11(2):217-228. [doi: 10.31887/DCNS.2009.11.2/hbrodaty] [Medline: 19585957]

XSL•FO
RenderX

2. Choi H, Langa KM, Norton EC, Cho TC, Connell CM. Changes in care use and financial status associated with dementia in older adults. JAMA Intern Med 2023 Dec 1;183(12):1315-1323. [doi: 10.1001/jamainternmed.2023.5482] [Medline: 37843869]
3. Skufca L, Rainville C. Caregiving out-of-pocket costs: 2021 report. : AARP; 2021.
4. Schmidt T, Juday C, Patel P, Karmarkar T, Smith-Howell ER, Fendrick AM. Expanding the catalog of patient and caregiver out-of-pocket costs: a systematic literature review. Popul Health Manag 2024 Feb;27(1):70-83. [doi: 10.1089/pop.2023.0238] [Medline: 38099925]
5. Templeman ME, Badana ANS, Haley WE. The relationship of caregiving to work conflict and supervisor disclosure with emotional, physical, and financial strain in employed family caregivers. J Aging Health 2020;32(7-8):698-707. [doi: 10.1177/0898264319848579] [Medline: 31096825]
6. Kalipeni J, Kashen J. Building our care infrastructure for equity, economic recovery and beyond. : The Care for All with Respect and Equity (CARE) Fund; 2022 URL: https://carefund.org/wp-content/uploads/2021/07/Building-Our-Care-infrastructure_FINAL.pdf [accessed 2025-06-27]
7. Garcia RI, Khemka S, Roghani AK, et al. Caring for individuals with Alzheimer's disease: a spotlight on Hispanic caregivers. J Alzheimers Dis Rep 2024;8(1):877-902. [doi: 10.3233/ADR-240035] [Medline: 38910940]
8. Mehdipanah R, Briceño EM, Malvitz M, et al. Exploring pathways to caregiver health: the roles of caregiver burden, familism, and ethnicity. J Aging Health 2025 Mar;37(3-4):148-155. [doi: 10.1177/08982643241235970] [Medline: 38388008]
9. Wu S, Vega WA, Jin H. Latinos and Alzheimer's disease: new numbers behind the crisis. : USC Edward R Roybal Institute on Aging and the Latinos Against Alzheimer's Network; 2016.
10. Adames HY, Chavez-Dueñas NY, Fuentes MA, Salas SP, Perez-Chavez JG. Integration of Latino/a cultural values into palliative health care: a culture centered model. Palliat Support Care 2014 Apr;12(2):149-157. [doi: 10.1017/S147895151300028X] [Medline: 23768861]
11. Jaldin MA, Balbim GM, Colin SJ, et al. The influence of Latino cultural values on the perceived caregiver role of family members with Alzheimer's disease and related dementias. Ethn Health 2023 May 19;28(4):619-633. [doi: 10.1080/13557858.2022.2115018]
12. Mendez-Luck CA, Anthony KP. Marianismo and caregiving role beliefs among U.S.-born and immigrant Mexican Women. J Gerontol B Psychol Sci Soc Sci 2016 Sep;71(5):926-935. [doi: 10.1093/geronb/gbv083] [Medline: 26362602]
13. Spillman BC, Long SK. Does high caregiver stress predict nursing home entry? Inquiry 2009;46(2):140-161. [doi: 10.5034/inquiryjrnl_46.02.140] [Medline: 19694388]
14. Rote SM, Moon H. Racial/ethnic differences in caregiving frequency: does immigrant status matter? J Gerontol B Psychol Sci Soc Sci 2018 Aug 14;73(6):1088-1098. [doi: 10.1093/geronb/gbw106] [Medline: 27573991]
15. National strategy to support family caregivers. : Administration for Community Living (ACL); 2022.
16. Tsai J, Kinney RL, Elbogen EB, Gluff J. Systematic review of financial interventions for adults experiencing behavioral health conditions. Psychiatr Serv 2024 Jun 1;75(6):570-579. [doi: 10.1176/appi.ps.20230271] [Medline: 38321921]
17. National Academies of Sciences, Engineering, and Medicine (NASEM). Considerations for the design of a systematic review of care interventions for individuals with dementia and their caregivers: letter report. : The National Academies Press; 2018. [doi: 10.17226/25326] [Medline: 30516946]
18. Leszko M. The effectiveness of psychoeducational and financial intervention to support caregivers of individuals with Alzheimer's disease in Poland. Innov Aging 2019 Jul;3(3):igz026. [doi: 10.1093/geroni/igz026] [Medline: 31392287]
19. Bowen DJ, Kreuter M, Spring B, et al. How we design feasibility studies. Am J Prev Med 2009 May;36(5):452-457. [doi: 10.1016/j.amepre.2009.02.002] [Medline: 19362699]
20. Sekhon M, Cartwright M, Francis JJ. Acceptability of healthcare interventions: an overview of reviews and development of a theoretical framework. BMC Health Serv Res 2017 Jan 26;17(1):88. [doi: 10.1186/s12913-017-2031-8] [Medline: 28126032]
21. Meinertz NR, Lee J, Kim Y. Are family caregiving programs appropriate for minority caregivers? A systematic review of 37 caregiving programs. J Appl Gerontol 2024;64(1):8-23. [doi: 10.1177/07334648241298109]
22. Mage S, Benton D, Gonzalez A, et al. "I Lay Awake at Night": Latino family caregivers' experiences covering out-of-pocket costs when caring for someone living with dementia. Gerontologist 2024 Jan 1;64(1):gnad011. [doi: 10.1093/geront/gnad011] [Medline: 36786288]
23. Meyer K, Mage SM, Gonzalez A, et al. Lessons from a pilot study of a culturally tailored financial well-being intervention among Latino family caregivers. J Appl Gerontol 2025 Jun;44(6):938-948. [doi: 10.1177/07334648241293524] [Medline: 39420566]
24. Baker SE, Edwards R. How many qualitative interviews are enough? Expert voices and early career reflections on sampling and cases in qualitative research. : NCRM Methods Review Papers; 2012.
25. Guest G, Bunce A, Johnson L. How many interviews are enough? Field Methods 2006 Feb;18(1):59-82. [doi: 10.1177/1525822X05279903]
26. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]

XSL•FO
RenderX

27. Glenn NM, Allen Scott L, Hokanson T, et al. Community intervention strategies to reduce the impact of financial strain and promote financial well-being: a comprehensive rapid review. Glob Health Promot 2021 Mar;28(1):42-50. [doi: 10.1177/1757975920984182] [Medline: 33601965]
28. Elk R. Partnering with minoritized communities to reduce health disparities: a focus on advance care planning. J Am Geriatr Soc 2023 Aug;71(8):2369-2372. [doi: 10.1111/jgs.18491] [Medline: 37432098]
29. Neidre DB, Delgado RE, Peacock KS, Luy LP, White CL. A scoping review to contribute to knowledge about culturally adapting interventions for Latino family caregivers of persons living with dementia. J Transcult Nurs 2025 Jan;36(1):43-56. [doi: 10.1177/10436596241256328] [Medline: 38828565]
30. Parker LJ, Marx KA, Nkimbeng M, et al. It's more than language: cultural adaptation of a proven dementia care intervention for Hispanic/Latino caregivers. Gerontologist 2023 Mar 21;63(3):558-567. [doi: 10.1093/geront/gnac120] [Medline: 35951488]
31. Samia LW, Hepburn K, Nichols L. "Flying by the seat of our pants": what dementia family caregivers want in an advanced caregiver training program. Res Nurs Health 2012 Dec;35(6):598-609. [doi: 10.1002/nur.21504]
32. White CL, Barrera A, Turner S, et al. Family caregivers' perceptions and experiences of participating in the learning skills together intervention to build self-efficacy for providing complex care. Geriatr Nurs 2022;45:198-204. [doi: 10.1016/j.gerinurse.2022.04.012] [Medline: 35533583]
33. Meyer K, Gonzalez A, Benton D. Qualitative evaluation of family caregivers' experiences participating in Knowledge and Interpersonal Skills to Develop Exemplary Relationships (KINDER): web-based intervention to improve relationship quality. JMIR Form Res 2023 Aug 22;7:e42561. [doi: 10.2196/42561] [Medline: 37606980]
34. Van't Leven N, Prick AEJC, Groenewoud JG, Roelofs PDDM, de Lange J, Pot AM. Dyadic interventions for community-dwelling people with dementia and their family caregivers: a systematic review. Int Psychogeriatr 2013 Oct;25(10):1581-1603. [doi: 10.1017/S1041610213000860] [Medline: 23883489]
35. Hopwood J, Walker N, McDonagh L, et al. Internet-based interventions aimed at supporting family caregivers of people with dementia: systematic review. J Med Internet Res 2018 Jun 12;20(6):e216. [doi: 10.2196/jmir.9548] [Medline: 29895512]
36. Hepburn K, Nocera J, Higgins M, et al. Results of a randomized trial testing the efficacy of tele-savvy, an online synchronous/asynchronous psychoeducation program for family caregivers of persons living with dementia. Gerontologist 2022 Apr 20;62(4):616-628. [doi: 10.1093/geront/gnab029]
37. Blanco LR, Ponce M, Gongora A, Duru OK. A qualitative analysis of the use of financial services and saving behavior among older African Americans and Latinos in the Los Angeles area. SAGE Open 2015;5(1):1-12. [doi: 10.1177/2158244014562388] [Medline: 26064788]
38. Rote SM, Angel JL, Moon H, Markides K. Caregiving across diverse populations: new evidence from the national study of caregiving and Hispanic EPESE. Innov Aging 2019 May;3(2):igz033. [doi: 10.1093/geroni/igz033] [Medline: 31517066]
39. Napoles AM, Chadiha L, Eversley R, Moreno-John G. Reviews: developing culturally sensitive dementia caregiver interventions: are we there yet? Am J Alzheimers Dis Other Demen 2010 Aug;25(5):389-406. [doi: 10.1177/1533317510370957] [Medline: 20508244]
40. Wennberg A, Dye C, Streetman-Loy B, Pham H. Alzheimer's patient familial caregivers: a review of burden and interventions. Health Social Work 2015 Nov;40(4):e162-e169. [doi: 10.1093/hsw/hlv062]
41. Peck BM, Asch DA, Goold SD, et al. Measuring patient expectations: does the instrument affect satisfaction or expectations? Med Care 2001 Jan;39(1):100-108. [doi: 10.1097/00005650-200101000-00011] [Medline: 11176547]
42. Quiroz YT, Solis M, Aranda MP, et al. Addressing the disparities in dementia risk, early detection and care in Latino populations: highlights from the second Latinos & Alzheimer's Symposium. Alzheimers Dement 2022 Sep;18(9):1677-1686. [doi: 10.1002/alz.12589] [Medline: 35199931]

Abbreviations

CONFIDENCE: Confidently Navigating Financial Decisions and Enhancing Financial Well-Being in Dementia Caregiving
IRB: institutional review board
TFA: theoretical framework of acceptability

Edited by A Mavragani; submitted 22.12.24; peer-reviewed by L Ou; revised version received 14.05.25; accepted 16.05.25; published 15.07.25.

Please cite as:
Mage S, Benton D, Wilber K, Aguila Rodriguez R, Silva D, Gonzalez A, Puga F, Meyer K
Acceptability of a Web-Based Financial Education Intervention for Latino Caregivers: Mixed Methods Evaluation
JMIR Form Res 2025;9:e70471
URL: https://formative.jmir.org/2025/1/e70471
doi:10.2196/70471

XSL•FO
RenderX

© Susanna Mage, Donna Benton, Kathleen Wilber, Rocio Aguila Rodriguez, David Silva, Alexander Gonzalez, Frank Puga, Kylie Meyer. Originally published in JMIR Formative Research (https://formative.jmir.org), 15.7.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Reducing Methamphetamine Use in Aboriginal and Torres Strait Islander Communities With the "We Can Do This" Web App: Qualitative Evaluation of Acceptability and Feasibility

Leda Sivak[1], BA, BSci, MA, PhD; Rachel Reilly[1,2], BA, MClinSci, DPsych; Shani Crumpen[3,4], BA; Carla Treloar[5], BSci, PhD; Rebecca McKetin[6], BSci, PhD; Julia Butt[7], BSci, PhD; Yvette Roe[8], PhD; Nadine Ezard[9], BA, MPH, MBBS, PhD; Brendan Quinn[10], BA, PhD; Jack Nagle[11], DipComServ; Wade Longbottom[12]; Clifford Warrior[13]; James Ward[14], PhD

[1]Warliparingga Aboriginal Health Equity, South Australian Health and Medical Research Institute, North Terrace, PO Box 11060, Adelaide, Australia
[10]Burnet Institute, Melbourne, Australia
[11]Real Drug Talk, Melbourne, Australia
[12]South Coast Aboriginal Medical Service Corporation, Nowra, Australia
[13]Port Lincoln Aboriginal Health Service, Port Lincoln, Australia
[14]Poche Centre for Indigenous Health, University of Queensland, Brisbane, Australia
[2]School of Psychology, University of Adelaide, Adelaide, Australia
[3]Kaiela Institute, Shepparton, Australia
[4]Poche Centre for Indigenous Health, Department of Rural Health, University of Melbourne, Shepparton, Australia
[5]Centre for Social Research in Health, University of New South Wales, Sydney, Australia
[6]National Drug and Alcohol Research Centre, University of New South Wales, Sydney, Australia
[7]School of Arts and Humanities, Edith Cowan University, Perth, Australia
[8]Molly Wardaguga Research Centre, Charles Darwin University, Brisbane, Australia
[9]National Centre for Clinical Research on Emerging Drugs, National Drug and Alcohol Research Centre, University of New South Wales, Sydney, Australia

Corresponding Author:
Rachel Reilly, BA, MClinSci, DPsych
Warliparingga Aboriginal Health Equity, South Australian Health and Medical Research Institute, North Terrace, PO Box 11060, Adelaide, Australia

Abstract

Background: Preventing and treating methamphetamine-related harm in Aboriginal and Torres Strait Islander populations is a significant challenge for health care services. Digital health care may offer opportunities to support individuals and families in ways that complement existing methamphetamine treatment options. This study responds to a community-identified priority as Aboriginal Community Controlled Health Services identified methamphetamine use as a key concern and sought support to respond to the needs of people who use methamphetamine and their families.

Objective: This paper reports on a process evaluation of the web application's acceptability and feasibility when used by clients and clinicians in residential rehabilitation services and primary care. This study is part of a larger project entitled "Novel Interventions to address Methamphetamine use in Aboriginal and Torres Strait Islander Communities" (NIMAC), which seeks to develop culturally appropriate and strengths-based prevention and treatment interventions to reduce methamphetamine related harm. "We Can Do This" was a web application developed for Aboriginal and Torres Strait Islander people who are seeking to reduce or stop methamphetamine use.

Methods: Clinicians and clients who had used the web application were recruited through Aboriginal Community Controlled Health Services and Aboriginal residential rehabilitation services in urban and regional Victoria and South Australia. Unidentified usage data was collected from all participants. After using the web application, those who indicated a willingness to be interviewed were contacted and interviewed by phone or in person and asked about the feasibility and acceptability of the web application. The framework method of analysis was used to structure and summarise the resulting qualitative data.

Results: Interviews with 24 clients and 11 clinicians explored the acceptability and feasibility of the web application. Acceptability incorporated the following domains: affective attitude, burden, ethicality, cultural appropriateness, coherence, opportunity cost, perceived effectiveness, and self-efficacy. The evaluation of feasibility assessed barriers and facilitators to the implementation of the program, with a focus on demand, practicality, fidelity, and integration. Results indicated that both clients and clinicians

XSL•FO
RenderX

found the web application content coherent, relatable, empowering, and culturally safe. Barriers to using the web application for clients included a lack of internet connectivity and personal issues such as scheduling.

Conclusions: Process evaluation is often under-valued. However, as "We Can Do This" was new, innovative and targeted a hard-to-reach population, understanding its feasibility and acceptability as a clinical tool was essential to understanding its potential. "We Can Do This" is unique as the only evidence-based, culturally appropriate internet-based therapeutic program specifically designed for Aboriginal and Torres Strait Islander people who use methamphetamine. Findings suggest it was both acceptable and feasible as a low-cost adjunct to usual care in residential rehabilitation and primary care settings.

(JMIR Form Res 2025;9:e58369) doi:10.2196/58369

KEYWORDS

Aboriginal and Torres Strait Islander; framework analysis; methamphetamine; process evaluation; qualitative; web-app; app; application; acceptability; feasibility; Australia; primary care; rehabilitation

Introduction

Background

Preventing and treating methamphetamine-related harm in Aboriginal and Torres Strait Islander populations is a significant challenge for health care services. Regular methamphetamine use contributes to adverse mental health outcomes, including substance use disorder, depression, anxiety, and psychosis [1]. In addition to increased hospital admissions [2], the use of methamphetamine is associated with more complex health and behavioral challenges for young people in corrective custody [3] and damages family and social relationships more broadly [4]. Available data indicates that Aboriginal and Torres Strait Islander people are approximately 2.2 times more likely to use methamphetamine than non-Indigenous Australians [5]. This study responds to a community-identified priority as Aboriginal Community Controlled Health Services (ACCHSs) identified methamphetamine use as a key concern [6] and sought support to respond to the needs of people who use methamphetamine and their families.

Treatment for Methamphetamine Use Disorder

Both residential treatment facilities and outpatient counseling therapies are used for the treatment of methamphetamine use disorder in Australia [7,8]. Psychological interventions such as cognitive behavior therapy (CBT) and motivational interviewing (MI) have shown modest effectiveness [9,10], while strengths-based, collaborative, culturally safe, and family-inclusive approaches to treatment are recommended for Aboriginal and Torres Strait Islander clients [11]. ACCHSs, which provide comprehensive primary health care, are well-placed to provide services to people who use methamphetamines and their families, but demonstrably effective treatment programs for Aboriginal and Torres Strait Islander people are rare. Digital health care may offer opportunities to support individuals and families in ways that complement existing treatment options.

Web-Based Therapeutic Interventions

Web-based therapeutic interventions (WBTI) have the potential to extend the reach of conventional treatment for methamphetamine [12]. Our previous work has shown that culturally appropriate, evidence-based WBTI have the potential to improve outcomes on a range of health issues and reduce inequalities in access to health care services [12]. The popularity of WBTI stems in part from their social and geographic reach. Chapman et al [13] have further noted that advantages of smartphone apps and web-based interventions for accessing health-related information include anonymity, increased accessibility, low cost, portability of information, and the ability to provide tailored feedback and support [13]. While digital applications have been developed to support Aboriginal and Torres Strait Islander people to address alcohol and other drug use [14], there currently are no WBTIs designed specifically for Aboriginal and Torres Strait Islander people who use methamphetamine.

NIMAC Study and We Can Do This

This study is part of a larger project entitled "Novel Interventions to address Methamphetamine use in Aboriginal and Torres Strait Islander Communities" (NIMAC). The first 2 phases of the project focused on gathering quantitative and qualitative data on patterns of methamphetamine use and risk and protective factors for use [15], while the latter 2 phases focused on using that information to inform the development of culturally appropriate and strengths-based prevention and treatment interventions to reduce methamphetamine related harm [16].

"We Can Do This" is a web application developed for Aboriginal and Torres Strait Islander people who are seeking to reduce or stop methamphetamine use [17]. It comprises 7 modules incorporating evidence-based therapeutic approaches, reflecting elements of CBT, MI, acceptance and commitment therapy (ACT), and narrative approaches, presented in a culturally relevant format. The web application was developed in partnership with 10 ACCHSs nationally, as has been described elsewhere [17].

"We Can Do This" was the subject of a randomized waitlist control trial in which individuals who self-identified as using methamphetamine could access the web application as a stand-alone intervention in the community [17]. The results of the trial are being reported separately. This acceptability and feasibility study arose from a strong interest from clinicians working with Aboriginal and Torres Strait Islander people who use methamphetamine and their desire to have access to the web application as an adjunct to client consultations. Process evaluations are an essential part of testing complex interventions [18] as they identify why an intervention might work and how it can be improved [19]. Successful implementation requires

XSL•FO
RenderX

that an intervention be acceptable to both those delivering and those receiving the intervention. This study sought to understand the acceptability and feasibility of the web application when used in the context of primary health care and residential rehabilitation services, as opposed to when used as a self-directed intervention in the community.

Methods

Overview

This study is framed within an Indigenous research paradigm [20]. The project was led by an Aboriginal academic and governed by an Aboriginal research governance group to ensure Aboriginal and Torres Strait Islander oversight, guidance and input into the study design, process and outputs. In line with ethical practice supporting data sovereignty for Aboriginal and Torres Strait Islander communities, all data are owned by the participating ACCHSs. This paper was drafted with reference to the CONSIDER Statement [21], which provides consolidated criteria for reporting of health research involving Indigenous peoples, and COREQ (Consolidated Criteria for Reporting Qualitative Research) [22], which provides consolidated criteria for reporting qualitative research.

Ethical Considerations

The study was approved by the Aboriginal Health Ethics Committee in South Australia (#04-19-810) and St Vincent's Hospital Human Research Ethics Committee in Victoria (#241 - 19). Informed consent was provided by all interview participants before data collection and included permission for secondary analysis of data. All study data were deidentified before analysis and reported anonymously. While clinician participants were not financially compensated, client participants were reimbursed Aus $40 (US $25) in appreciation of their time.

Interview Guide Development

The semistructured interview guide incorporated questions addressing each of the domains within the overarching constructs of acceptability and feasibility. Acceptability, a multifaceted construct, incorporates the following domains: affective attitude, how the participant feels about the program; burden, the perceived effort required to use the program; ethicality and cultural appropriateness, the extent to which the intervention is a good fit with the participant's value system; coherence, the extent to which a participant understands the program and how it works; opportunity cost, the extent to which a participant feels they have to compromise values or benefits to engage in the program; perceived effectiveness, the extent to which the program is perceived to achieve its purpose; and self-efficacy, a participant's confidence that they can use the program [23].

The evaluation of feasibility assessed barriers and facilitators to the implementation of the program, with a focus on: demand, degree of use and interaction with the program; practicality, the extent to which the program can be delivered within current resource and time restraints; fidelity, whether the program was delivered as designed or whether adaptations were made to fit the service context; and integration, the level of system change required to incorporate the program into the service.

Sampling, Recruitment, and Data Collection

Building on earlier phases of the NIMAC study [17], perspectives regarding acceptability and feasibility were sought from 2 groups of stakeholders who had been involved in earlier phases of NIMAC: Aboriginal and Torres Strait Islander clients of participating health services, aged 16 years and older, who had used the internet-based program at least once; and clinicians from the same services, who had used the program with clients. Clinicians could include nurses, doctors, counselors, Aboriginal health workers or practitioners, psychologists, drug and alcohol workers, youth workers or social and emotional well-being workers.

Clinicians and clients were recruited by the study team through Aboriginal Community Controlled Health Services and Aboriginal residential rehabilitation services in urban and regional Victoria and South Australia that had participated in earlier phases of the NIMAC study. These existing relationships supported the staged approach of the study. After using the web application, those who indicated a willingness to be interviewed were contacted and offered an interview by telephone, in-person or in writing. All interviews were conducted by telephone and took between 40 and 75 minutes.

Analysis

Interviews were audio recorded, transcribed, and deidentified prior to analysis. Participants were offered the opportunity to review the deidentified transcript of their interview and provide feedback. Two participants requested transcripts, but none provided feedback. Deidentified transcripts were uploaded into NVivo 12 software (Lumivero) for thematic analysis. The interview guide provided a framework for deductive coding, as acceptability and feasibility factors were organized under these categories. Themes that emerged inductively outside of this framework were coded separately. The framework method of analysis was used to structure and summarize the data [24]. Framework analysis comprises five steps: (1) familiarization with the data, (2) identification of a thematic framework, (3) indexing of study data against the framework, (4) charting the indexed data, and (5) mapping and interpretation [25].

Analysis was carried out by 3 researchers, with all data being coded thematically by at least 2 researchers in an iterative process of coding, discussion, and cross-checking. As 2 of these researchers were non-Indigenous, community-based researchers and Aboriginal and Torres Strait Islander members of the investigator team also reviewed the analysis to ensure that findings were expressed in culturally appropriate ways.

Results

Participants

Participants in this study included 24 Aboriginal or Torres Strait Islander participants and 11 clinicians who had used the web application. A total of 14 Aboriginal or Torres Strait Islander participants female and 10 were male and ranged in age from 17 to 58 years. Clinicians included a youth worker, 2 psychologists, 3 peer workers, and 5 alcohol and other drug counselors. The following codes are used to identify: P=person

XSL•FO
RenderX

with experience, also referred to as "clients"; C=clinician; and W=peer worker.

Acceptability

Affective Attitude

Affective attitude refers to how a participant feels about the program being evaluated. There were several examples of positive affect. As one clinician stated:

> *I had a good feeling that there was this kind of app out there available for clients. So going through it, I was like "Yeah, this is looking pretty good."* [C2]

Not all participants had clear expectations of the web application, and expressed wariness:

> *I was a bit wary of it at first, to be honest, and when I went through it, I thought, "You know what, this is pretty good" and yeah, I actually thought it was better than what I thought it was going to be.* [C4]

Another participant stated:

> *At first I was like, "This is not going to work." And then after a while, after a couple of goes at it, I was like "Oh yeah, this is actually helping." And then I got used to it and I enjoyed it.* [P38]

> *At first I was like, "This is not going to work." And then after a while, after a couple of goes at it, I was like "Oh yeah, this is actually helping." And then I got used to it and I enjoyed it. [P38]*

Another said, "I wasn't sure what to expect at first, but yeah, it was really good" (P153). Some participants appeared satisfied:

> *I was hoping that the app would help me on the way and it definitely support me in that direction. I went into the app knowing that I needed to try a lot of things to cut back and get off it [methamphetamine].* [W3]

Some participants felt challenged by the content of the web application. As one client commented, "Sometimes it hit a little too close to home, what these people are saying" (P184). Others, who initially felt scared about the web application, found the video stories relatable and true to life, "It's good with the videos and stuff – they do hit home with me. I could relate" (P190).

Qualities of the web application associated with positive affect included that it focuses on Aboriginal and Torres Strait Islander people, is easy to understand and straightforward to use, is available at home, and is nonintrusive. Overwhelmingly, participants found the graphics, including the themes and colours, inviting and relatable. One clinician reported, "I thought there was a good Aboriginal theme going through, which is important" (C4). In relation to the films, another participant suggested:

> *It's good to have other people's comments and their addiction included in it, so you can have a little look and see what we relate to and what we don't, and how it confronts us.* [P158]

Burden

Participants were equally divided between those who experienced time pressures as a barrier to engaging with the web application, and those who either found it easy to prioritise using it or had experienced no time barriers. More significantly, several participants reported that a lack of internet connection presented a barrier to using the web application at times, and 2 reported having trouble signing in.

A facilitator of use identified by 2 participants was the support of nonjudgemental peer-workers or clinicians who encouraged them to use the web application or helped them to log in. Several participants described features of the web application that they found helpful, such as the flexibility of being able to use it privately, anywhere, any time; being able to access the content in any order; and receiving notifications and reminders from the web application itself.

For similar reasons, most participants reported that it was not an effort to use the web application, although 1 participant did report that it was "emotionally challenging sometimes" (P190). Another suggested that "If we are putting ourselves in difficult situations to use the meth, it shouldn't be difficult to use the web app" (P200). Although 2 participants found the web application "too wordy" (P142, W1), the accessibility of the language used in the web application was highlighted as a strength of the web application by most participants who responded to this question (n=12).

Coherence

All 12 participants who commented on coherence of the web application agreed that the content made sense and was clear in its purpose, with 1 client commenting, "The content of information provided was helpful and clear in how it was meant to help people" (P165). Another participant commented:

> *What I took from it, it was self-motivated. So being given and being aware of certain tools, creating that support network and prompting us to think of some supports that we could put down on them that we could call up. It really made sense.* [W3]

Only 2 clinicians commented on the therapeutic alignment of the web application with their existing practice. Both could see the value of the web application, with 1 clinician noting that the web application will complement a range of therapeutic approaches, given that "There's no one-fix therapy or approach" (C1). The other clinician, who used a combination of motivational interviewing and harm reduction, commented that the web application "aligned pretty well" (C2).

Perceived Effectiveness

A total of 7 participants (including 2 peer workers) commented on the effectiveness of the web application for reducing their methamphetamine use, with almost all finding the web application effective. For example, 1 client observed that the web application "helped when I was craving; it helped at that time when I didn't have a support person to turn to – I just turned to the app" (P38). Likewise, a peer worker commented:

> *The app was really helpful for me, because once I tuned into the app, and I was able to go through and*

XSL•FO
RenderX

> *read everything that app had to offer in terms of the modules and the videos and it really helped me. …It really made me think about what stage of change I am at. … It really helped me therapeutically, it helped me.* [W3]

A total of 5 participants commented on the effectiveness of the web application for avoiding situations where they might feel stigma. In contrast, 1 participant shared that the web application was less helpful than expected, especially as the cravings module increased their experience of cravings.

Furthermore, 5 participants observed that the web application was also helpful for making changes in other areas of their life, with 1 peer worker commenting:

> *It really helped me to just really reach out to people when anything was going wrong. It didn't have to be about ice or anything. I've just known that I've got some people in my corner and how much they really mean to me.* [W3]

A client, who noted that the web application helped with their breathing, anxiety and fitness commented that, "It's mentally empowering, if that's the word. It helps your mental side" (P54).

Ethicality

Most participants were content with the ethicality of the web application. As a client explained, "I felt really safe to just be open and honest on there" (P190). However, one clinician was concerned about participant mental health due to the self-led modality of the application. They suggested incorporating a question function to further support the social and emotional well-being of individuals using the application. Other participants raised concerns around the collection and use of personal data and data security. The majority of participants felt comfortable to raise concerns about the application with a clinician or the research team.

Cultural Appropriateness

The web application was found to be culturally appropriate with the inclusion of Aboriginal and Torres Strait Islander people, language, artwork, images, and fonts as relatable for many participants. One client stated:

> *It didn't seem institutionalised. It did not have that government smell about it… You could tell that it is real people trying to help other people, and it's – it's real.* [P190]

Whilst 4 out of 10 participants were unsure of the extent to which the web application aligned with their value system, others felt it was a good fit. For example, one peer worker said:

> *It really did fit with my values. And not just my values, my community's values as well, about mob trying to break this cycle. It's such an important app. And it really made me feel like I was part of something. It sat with my values really well.* [W3]

Opportunity Costs

A total of 10 participants commented on the opportunity costs associated with the web application, with all observing that there was no need to compromise their other commitments to engage in the program. The flexibility of the web application was especially helpful in this regard, as one client commented:

> *It was good in a sense that you could just do it whenever and, if you did have another commitment, it didn't have to interfere with that because you could just do it afterwards or beforehand.* [P158]

Self-Efficacy

In total, 9 participants commented on their confidence using the web application, with almost all feeling immediately confident. In contrast, 1 participant needed to engage with the web application more than once to build confidence with it:

> *At the start, I felt sort of like a little bit just in unfamiliar territory. Once I had the option to go back into stuff, it made me a lot more confident to go back in and just give the questions a second crack and read them again and really make sense of it. Just going over them the first time it didn't really sink in. But when I had an option to go back in and check them out, and go back into ones that sort of sparked my interest more, that really made a lot more an impact, sunk in more.* [W3]

Furthermore, 8 participants observed that the web application did not require any new skills, with 1 client commenting:

> *I think for most people that are familiar with using either phones or any sort of device, it was fairly straightforward how to access the different parts of it.* [P153]

Feasibility

Demand

Participant use and interaction with the web application varied. Duration and frequency with which an individual recounted using the web application were between 10 and 60 minutes each time, 1 - 5 days per week or 2 - 4 times in total. Furthermore, 6 out of 10 clients and 1 clinician reported using the application outside of counseling sessions. However, clinicians also described using the web application with clients before giving "them the opportunity to do it themselves" (W1). Participants accessed the web application by desktop computer, tablet, or mobile phone.

Fidelity

Fidelity refers to whether participants found they used the web application as designed and suggested by the researchers, or whether they made adjustment to it to suit their own personal or organizational (in the case of clinicians) context. Very few participants responded to this question. The 2 who did respond suggested that they "self-explored" (P165) the web application, or "did it in my own time" (P200), reflecting the flexibility that was intentionally built into the web application design. Given this flexibility, the lack of responses to this question could indicate that the question of fidelity was less relevant to participants' experiences of using the web application.

XSL•FO
RenderX

Integration

Similarly, only 2 clinicians commented on integration, or the systems changes required to incorporate the web application into the service. One commented that due to time constraints within existing service delivery, using the web application with clients had been undertaken opportunistically thus far and that dedicating time for using the web application with clients would make their approach more strategic.

Practicality

Most participants reported that they had the equipment they needed to use the web application. The remainder either did not respond directly to this question (n=5) or reported that a lack of reliable internet was a problem for them (n=2). One participant described having intermittent internet but not finding this a problem as receiving text messages from the web application meant that they "could still be reminded of those things" (P153). One clinician reported that using the web application on a tablet was preferable to using it on a desktop computer as it facilitated a more relaxed, conversational setting and they saw their role as supportive rather than directive:

> *We'll share it together and then I can give it to them and they're in total control of what's happening with it. I'll support them with utilising it but they're in control whether they use it and what they do with it.* [C1]

Out of the 21 participants who responded to questions about navigation within the web application, 1 participant had trouble going back one step once they were in the module, and 1 peer-worker suggested there were "too many options in some modules" (W1). Another peer-worker had trouble logging people in initially. However, the majority of respondents to this question had no trouble with navigation, and one clinician stated:

> *What I liked the most, the interface and the way it controls was very user friendly, especially to those ones who aren't necessarily the most tech savvy.* [C5]

Participants varied in terms of whether they went through each module sequentially or in random order; however, most seemed to go through all the modules at least once, then return to those they found most useful. For example, one participant said:

> *I went from the first one to the last one. And then, like I said, the option of being able to go back into them. Then I picked the ones that I really was interested in.* [W3]

A total of 11 participants gave feedback on the practicality of finding time within their routines to use the web application. Of these, one said they could not find the time. Again, flexibility was highlighted as a facilitator for people finding time to use the web application. A participant stated:

> *They just said, 'You can just sit on it anytime you want. You can use it on the bus.' Because I told them I use public transport. And they just more or less helped me identify all this time I do have to use it.* [W3]

Another stated:

> *Just that it was something that could be there all the time and I could do it when I had time in between other commitments that I had.* [P153]

Another suggested that finding time to use the web application was harder at first, "But once I started using it, it only took 10 - 15 minutes each time" (C1).

Discussion

Principal Findings

The results of this study demonstrate that clinicians and clients generally found the "We Can Do This" web application to be both acceptable and feasible when used in the context of primary health care and residential rehabilitation services. "We Can Do This" was generally found by both clients and clinicians to be coherent, relatable, empowering, and culturally safe. In particular, study participants valued the flexibility of the web application and noted how it reduced the stigma they sometimes felt in seeking help in other ways. They also liked the video vignettes with Aboriginal and Torres Strait Islander actors and the ability to share or print the module summaries. Challenges with the web application in its tested form included difficulty signing in and barriers with internet connections.

Methamphetamine use is associated with higher use of acute health services and reduced use of community-based or primary health care [2]. McKetin et al [2] note that this may also reflect a nonacute service system that is ill-prepared to meet the needs of people who use methamphetamine. In general practice, drug use can be a difficult topic to broach with patients and to tackle effectively. For example, stigma has been identified as a barrier to treatment-seeking [26,27], with 1 in 3 Australians who use methamphetamine reporting experiences of discrimination and 87% expressing self-stigma [28]. Digital interventions can provide a useful entry point for clinicians to start conversations, break down stigma, normalize some of the challenges experienced by people who use methamphetamine, and guide people toward making positive change.

In contexts where access to trained alcohol and other drug counselors may be limited, web applications have the potential to increase the capacity of health services to respond to the needs of people seeking help for methamphetamine use. For example, having shown that the eHealth platform, Cracks in the Ice, improved knowledge and decreased stigma about methamphetamine, Kershaw et al [29] demonstrate that digital information and support platforms play an important role in improving knowledge and reducing stigma. As digital health care becomes more widely used [13], ensuring that Aboriginal and Torres Strait Islander people can access internet-based support that is culturally appropriate is essential to achieving equity in health service access [30]. "We Can Do This" may increase the capacity of mainstream services to better meet the needs of Aboriginal and Torres Strait Islander people who use methamphetamine in culturally relevant and safe ways.

Limitations

Interview participants in our sample self-nominated to take part, likely favouring those with strong views and opinions. While findings may not be representative of the experiences and

XSL•FO
RenderX

concerns of all Aboriginal and Torres Strait Islander communities in Australia or Indigenous peoples in other countries, the congruence between both client and clinician participants suggests the broad acceptability and feasibility of this web app in primary care and residential rehabilitation settings.

Process evaluation is an essential and often under-valued aspect of program evaluation. Given that the "We Can Do This" web application is a new and innovative program, targeting a unique and often hard-to-reach population, understanding how it is implemented, its feasibility, acceptability, and overall attractiveness as a clinical tool are essential to understanding its value and potential.

Conclusion

The "We Can Do This" web application is unique and innovative, being the only culturally appropriate internet-based therapeutic program developed specifically for Aboriginal and Torres Strait Islander people who use methamphetamine. This web application makes a contribution to the growing suite of digital resources developed for Aboriginal and Torres Strait Islander health and well-being. The findings of this study provide tentative support for the usability of the program in supported health service settings. While future research should explore its effectiveness at supporting a reduction in methamphetamine use in this context, this study suggests that "We Can Do This" may provide a valuable adjunct to more traditional counseling methods, at low cost, without significant burden on health services, making culturally appropriate care more readily available to those who need it.

Conflicts of Interest

None declared.

Checklist 1
COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist.
[PDF File, 426 KB - formative_v9i1e58369_app1.pdf]

References

1. Degenhardt L, Sara G, McKetin R, et al. Crystalline methamphetamine use and methamphetamine-related harms in Australia. Drug Alcohol Rev 2017 Mar;36(2):160-170. [doi: 10.1111/dar.12426] [Medline: 27286742]
2. McKetin R, Degenhardt L, Shanahan M, Baker AL, Lee NK, Lubman DI. Health service utilisation attributable to methamphetamine use in Australia: Patterns, predictors and national impact. Drug Alcohol Rev 2018 Feb;37(2):196-204. [doi: 10.1111/dar.12518] [Medline: 28294443]
3. Kaye S, Lewandowski A, Bowman J, Doyle MF. Crystal methamphetamine use among young people entering custody: prevalence, correlates and comorbidity. Drug Alcohol Rev 2021 Nov;40(7):1266-1274. [doi: 10.1111/dar.13210] [Medline: 33155354]
4. Watanabe-Galloway S, Ryan S, Hansen K, Hullsiek B, Muli V, Malone AC. Effects of methamphetamine abuse beyond individual users. J Psychoactive Drugs 2009 Sep;41(3):241-248. [doi: 10.1080/02791072.2009.10400534] [Medline: 19999677]
5. National drug strategy household survey 2019: drug statistics series no 32. : Australian Institute of Health and Welfare; 2020 URL: https://www.aihw.gov.au/reports/illicit-use-of-drugs/national-drug-strategy-household-survey-2019/contents/summary [accessed 2025-02-25]
6. Aboriginal and torres strait islander health organisations online services report—key results 2015-16. : Australian Institute of Health and Welfare; 2017 URL: https://www.aihw.gov.au/reports/indigenous-australians/health-organisations-osr-key-results-2015/report-editions [accessed 2025-02-25]
7. McKetin R, Najman JM, Baker AL, et al. Evaluating the impact of community-based treatment options on methamphetamine use: findings from the Methamphetamine Treatment Evaluation Study (MATES). Addiction 2012 Nov;107(11):1998-2008. [doi: 10.1111/j.1360-0443.2012.03933.x] [Medline: 22564065]
8. McKetin R, Kothe A, Baker AL, Lee NK, Ross J, Lubman DI. Predicting abstinence from methamphetamine use after residential rehabilitation: Findings from the Methamphetamine Treatment Evaluation Study. Drug Alcohol Rev 2018 Jan;37(1):70-78. [doi: 10.1111/dar.12528] [Medline: 28421682]
9. Alammehrjerdi Z, Briggs NE, Biglarian A, Mokri A, Dolan K. A randomized controlled trial of brief cognitive behavioral therapy for regular methamphetamine use in methadone treatment. J Psychoactive Drugs 2019;51(3):280-289. [doi: 10.1080/02791072.2019.1578445] [Medline: 30835643]
10. Paulus MP, Stewart JL. Neurobiology, clinical presentation, and treatment of methamphetamine use disorder: a review. JAMA Psychiatry 2020 Sep 1;77(9):959-966. [doi: 10.1001/jamapsychiatry.2020.0246] [Medline: 32267484]
11. MacLean S, Harney A, Arabena K. Primary health-care responses to methamphetamine use in Australian Indigenous communities. Aust J Prim Health 2015;21(4):384-390. [doi: 10.1071/PY14126] [Medline: 25704260]

XSL•FO
RenderX

12. Reilly R, Stephens J, Micklem J, et al. Use and uptake of web-based therapeutic interventions amongst Indigenous populations in Australia, New Zealand, the United States of America and Canada: a scoping review. Syst Rev 2020 May 31;9(1):123. [doi: 10.1186/s13643-020-01374-x] [Medline: 32475342]
13. Chapman C, Champion KE, Birrell L, et al. Smartphone apps about crystal methamphetamine ("Ice"): systematic search in app stores and assessment of composition and quality. JMIR Mhealth Uhealth 2018 Nov 21;6(11):e10442. [doi: 10.2196/10442] [Medline: 30463834]
14. Lee KSK, Conigrave JH, Wilson S, et al. Patterns of drinking in Aboriginal and Torres Strait Islander peoples as self-reported on the Grog Survey App: a stratified sample. BMC Med Inform Decis Mak 2019 Sep 5;19(1):180. [doi: 10.1186/s12911-019-0879-8] [Medline: 31488135]
15. Reilly R, Wand H, McKetin R, et al. Survey methods and characteristics of a sample of Aboriginal and Torres Strait Islander and non-Indigenous people who have recently used methamphetamine: the NIMAC survey. Drug Alcohol Rev 2020 Sep;39(6):646-655. [doi: 10.1111/dar.13085] [Medline: 32573069]
16. Reilly R, Gendera S, Treloar C, et al. Identifying risk and protective factors, including culture and identity, for methamphetamine use in Aboriginal and Torres Strait Islander communities: Relevance of the "communities that care" model. Soc Sci Med 2020 Dec;266:113451. [doi: 10.1016/j.socscimed.2020.113451] [Medline: 33126097]
17. Reilly R, McKetin R, Wand H, et al. A web-based therapeutic program (We Can Do This) for reducing methamphetamine use and increasing help-seeking among Aboriginal and Torres Strait Islander people: protocol for A randomized wait-list controlled trial. JMIR Res Protoc 2019 Jul 26;8(7):e14084. [doi: 10.2196/14084] [Medline: 33932278]
18. Moore GF, Audrey S, Barker M, et al. Process evaluation of complex interventions: Medical Research Council guidance. BMJ 2015 Mar 19;350:h1258. [doi: 10.1136/bmj.h1258] [Medline: 25791983]
19. Pals RAS, Drejer S, Laursen RH, et al. Implementing a collaborative model in health education practice: a process evaluation of a health education programme targeting users with mental health problems. BMC Health Serv Res 2020 Jan 14;20(1):38. [doi: 10.1186/s12913-019-4819-1] [Medline: 31937316]
20. Rix EF, Wilson S, Sheehan N, Tujague N. Indigenist and decolonizing research methodology. In: Handbook of Research Methods in Health Social Sciences: Springer; 2019:253-267.
21. Huria T, Palmer SC, Pitama S, et al. Consolidated criteria for strengthening reporting of health research involving indigenous peoples: the CONSIDER statement. BMC Med Res Methodol 2019 Aug 9;19(1):173. [doi: 10.1186/s12874-019-0815-8] [Medline: 31399058]
22. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
23. Sekhon M, Cartwright M, Francis JJ. Acceptability of healthcare interventions: an overview of reviews and development of a theoretical framework. BMC Health Serv Res 2017 Jan 26;17(1):88. [doi: 10.1186/s12913-017-2031-8] [Medline: 28126032]
24. Gale NK, Heath G, Cameron E, Rashid S, Redwood S. Using the framework method for the analysis of qualitative data in multi-disciplinary health research. BMC Med Res Methodol 2013 Sep 18;13:117. [doi: 10.1186/1471-2288-13-117] [Medline: 24047204]
25. Goldsmith L. Using framework analysis in applied qualitative research. TQR 2021. [doi: 10.46743/2160-3715/2021.5011]
26. Cumming C, Troeung L, Young JT, Kelty E, Preen DB. Barriers to accessing methamphetamine treatment: A systematic review and meta-analysis. Drug Alcohol Depend 2016 Nov 1;168:263-273. [doi: 10.1016/j.drugalcdep.2016.10.001] [Medline: 27736680]
27. Axelsson M, Snell TL. Breaking the ice: narratives of recovery from crystal methamphetamine. Aust Psychol 2021 Jan 2;56(1):81-92. [doi: 10.1080/00050067.2021.1893600]
28. Deen H, Kershaw S, Newton N, et al. Stigma, discrimination and crystal methamphetamine ('ice'): Current attitudes in Australia. Int J Drug Policy 2021 Jan;87:102982. [doi: 10.1016/j.drugpo.2020.102982] [Medline: 33126164]
29. Kershaw S, Birrell L, Deen H, et al. Evaluation of a digital health initiative in illicit substance use: cross-sectional survey study. J Med Internet Res 2021 Aug 10;23(8):e29026. [doi: 10.2196/29026] [Medline: 34383690]
30. Brusse C, Gardner K, McAullay D, Dowden M. Social media and mobile apps for health promotion in Australian Indigenous populations: scoping review. J Med Internet Res 2014 Dec 10;16(12):e280. [doi: 10.2196/jmir.3614] [Medline: 25498835]

Abbreviations

ACCHS: Aboriginal Community Controlled Health Service
ACT: acceptance and commitment therapy
CBT: cognitive behavioral therapy
COREQ: Consolidated Criteria for Reporting Qualitative Research
MI: motivational interviewing
NIMAC: Novel Interventions to address Methamphetamine use in Aboriginal and Torres Strait Islander Communities

XSL•FO
RenderX

WBTI: web-based therapeutic interventions

Edited by A Mavragani; submitted 13.03.24; peer-reviewed by S Maclean; revised version received 29.09.24; accepted 30.09.24; published 15.07.25.

Please cite as:
Sivak L, Reilly R, Crumpen S, Treloar C, McKetin R, Butt J, Roe Y, Ezard N, Quinn B, Nagle J, Longbottom W, Warrior C, Ward J Reducing Methamphetamine Use in Aboriginal and Torres Strait Islander Communities With the "We Can Do This" Web App: Qualitative Evaluation of Acceptability and Feasibility
JMIR Form Res 2025;9:e58369
URL: https://formative.jmir.org/2025/1/e58369
doi:10.2196/58369

© Leda Sivak, Rachel Reilly, Shani Crumpen, Carla Treloar, Rebecca McKetin, Julia Butt, Yvette Roe, Nadine Ezard, Brendan Quinn, Jack Nagle, Wade Longbottom, Clifford Warrior, James Ward. Originally published in JMIR Formative Research (https://formative.jmir.org), 15.7.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Usability and Usefulness of SMS-Based Artificial Intelligence Intervention (Mwana) on Breastfeeding Outcomes in Lagos, Nigeria: Pilot App Development Study

Anisha Musti

Independent Researcher, 10685-B Hazelhurst Dr. #40992, Houston, TX, United States

Corresponding Author:
Anisha Musti
Independent Researcher, 10685-B Hazelhurst Dr. #40992, Houston, TX, United States

Abstract

Background: Nigeria has one of the highest child mortality rates globally, with 111 deaths per 1000 live births. Exclusive breastfeeding (EBF) improves infant survival by providing essential nutrients and antibodies that protect against infections and diseases. Despite its benefits, EBF rates in Nigeria remain low at 29%, largely due to limited health care support and breastfeeding guidance. With the proliferation of mobile phones in Nigeria, mobile health (mHealth) interventions are being explored as scalable solutions. SMS text messaging interventions have demonstrated success in delivering behavioral interventions; yet, few use artificial intelligence (AI) for personalized breastfeeding support.

Objective: This study evaluates the effectiveness of Mwana, an AI-powered SMS-based app, in improving breastfeeding outcomes for postpartum mothers in Lagos, Nigeria.

Methods: Mwana was developed using TextIt for SMS integration and Meta's Wit.ai for natural language processing (NLP). The chatbot provides breastfeeding support via SMS, offering personalized tips, addressing common challenges, and connecting users to human agents when necessary. The intervention was piloted with 216 postpartum mothers recruited through local health care networks, focusing on usability, usefulness, and engagement. The study used a mixed methods approach, using structured surveys and observation to assess participant experiences at multiple intervals over a 6-month period. Primary outcomes measured were app usability, usefulness, and breastfeeding adherence.

Results: The intervention was well-received, with high scores for both usefulness (mean 4.01, SD 1.41) and usability (mean 3.92, SD 1.35) on a 5-point scale. The majority of respondents, 57% (118/206), rated the chatbot's usefulness at the highest score of 5. Qualitative feedback statements identified areas for improvement, including enhancing AI comprehension, response times, and human-like interaction.

Conclusions: The study highlights the potential of Mwana to improve breastfeeding outcomes in resource-limited settings, contributing to the growing body of evidence supporting mHealth interventions in maternal and child health. By leveraging personalized messaging, SMS-based delivery, and language localization, Mwana offers a scalable, accessible solution. However, challenges remain regarding AI comprehension, and further research is necessary to evaluate Mwana's effectiveness among populations not actively engaged with health care services. Future iterations will expand AI training datasets, refine NLP capabilities, and scale to broader populations.

(JMIR Form Res 2025;9:e65157) doi:10.2196/65157

KEYWORDS

breastfeed; mother; maternal; mHealth; mobile health; mobile intervention; femtech; child mortality; maternal healthcare; pediatric; infant; newborn; artificial intelligence; SMS; text message; Nigeria; Africa

Introduction

Background

In Nigeria, child mortality remains a critical public health concern. As of 2020, the under-five mortality rate was estimated at 111 deaths per 1000 live births, considering both perinatal and preterm births [1]. This rate places Nigeria as the country with the second-highest under-five mortality rate globally and the highest number of annual under-five deaths worldwide, with 844,321 deaths recorded in 2020 [2].

Breastfeeding is widely recognized as a critical intervention to improve child survival outcomes for both perinatal and preterm babies, a leading cause of child mortality [2]. Secretory immunoglobulin, a key component of breast milk, provides essential immune protection to infants, particularly against respiratory and gastrointestinal infections—two leading causes of child morbidity and mortality [3,4]. Exclusive breastfeeding

XSL•FO
RenderX

(EBF) during the first 6 months of life is strongly recommended by the World Health Organization (WHO) and United Nations International Children's Emergency Fund (UNICEF) to optimize infant health outcomes [5]. Despite its benefits, Nigeria's EBF rate remains low at approximately 29%, far below the global target of 70% by 2030 [6]. Addressing this gap could prevent over 95,000 child deaths annually in Nigeria [7].

One of the key barriers to improving EBF rates in Nigeria is the lack of adequate support within the health care system. Many mothers struggle to access accurate and timely information on breastfeeding techniques, proper positioning, and other challenges. Furthermore, limited counseling from health care providers often results in suboptimal breastfeeding practices [8]. While breastfeeding is widely practiced in Nigeria, the persistently low EBF rate of 29% highlights the gap between awareness and sustained EBF [9].

Prior Work

The rapid growth of mobile phone use has significantly impacted health care delivery, particularly in resource-limited settings [10]. With increasing accessibility and affordability, mobile technology—especially SMS—offers a powerful tool for delivering targeted health interventions to mothers [11]. Mobile health (mHealth) initiatives have already demonstrated success in improving maternal and child health outcomes [11].

In Nigeria, where mobile phone penetration continues to rise, mHealth tools could help address barriers to EBF by providing mothers with timely, accessible information [12]. With approximately 88% of women owning basic cell phones [13], mobile technology presents a promising avenue for intervention. However, research on integrating artificial intelligence (AI) into SMS-based breastfeeding interventions remains limited.

One example of mobile technology leveraging AI is the iCow app. Developed by Su Kahumbu, iCow was originally created to be a mobile-based calendar for farmers in Tanzania and Ethiopia. With nearly 60,000 registered users, iCow illustrates the potential of mobile apps to effectively engage large populations, a relevant insight for Nigeria's expanding smartphone user base [14].

Su Kahumbu

Su Kahumbu originally developed iCow as a mobile-based calendar for farmers in Tanzania and Ethiopia. With nearly 60,000 registered users, iCow illustrates the potential of mobile apps to effectively engage large populations—a relevant insight for Nigeria's expanding smartphone user base [14].

Similarly, the Moommae mobile app, developed by P Chaovalit and S Pongnumkul, was studied as a breastfeeding intervention in Thailand. It offered tracking tools for feeding activities and support for breastfeeding outside the home. While assessments confirmed its potential as a self-management tool, usability challenges underscored the need for refinements to enhance the user experience [15].

These findings suggest that mobile-based interventions can positively influence breastfeeding behaviors. However, further research is needed to explore AI-driven SMS apps for improving breastfeeding support in Nigeria.

Goals of This Study

This research study aims to evaluate the effectiveness of an SMS-based AI app in improving breastfeeding outcomes for mothers in Lagos, Nigeria. The intervention leverages culturally accepted breastfeeding practices in Nigeria while addressing gaps in medical support through an AI-powered virtual assistant. Specifically, we implemented Mwana using TextIt and Meta AI, piloted it with a focus group of postpartum mothers in Lagos, Nigeria, and evaluated our implementation using usability and usefulness metrics to assess its potential for long-term impact on breastfeeding practices.

By focusing on EBF as a key strategy for reducing child mortality, this research contributes valuable insights into the role of AI-driven mHealth interventions in resource-limited settings.

Methods

Study Design

This formative pilot study combined extensive opinion surveys and observational data collection to assess the feasibility and effectiveness of the SMS-based AI intervention, Mwana, in supporting breastfeeding mothers.

Data collection involved externally administered surveys that captured both quantitative engagement metrics and qualitative participant experiences. The surveys included structured response options to track breastfeeding adherence, response rates, and interaction patterns with the intervention, alongside open-ended questions that allowed mothers to share insights on their challenges, preferences, and overall experience using Mwana. Multiple data collection points were scheduled at specific intervals, including at the commencement of the study, at 1 month, 3 months, and 6 months later. No formal interim analyses were conducted due to the small sample size and exploratory nature of the pilot study.

In addition, an observational pilot study was conducted to evaluate how Mwana was integrated into postpartum care without direct intervention or variable manipulation. Platform use was remotely monitored and recorded, focusing on its role in breastfeeding routines among postpartum mothers. The study spanned 6 months (January 1, 2023, to July 31, 2023) in Lagos, Nigeria, with all participant interactions occurring through the platform and surveys, without any in-person engagement.

No major methodological changes were made after the trial commenced. However, based on preliminary user feedback in the first month, additional predefined responses were introduced to improve the chatbot's comprehension of common breastfeeding concerns, and message queuing was optimized to reduce response delays.

Participants and Recruitment

Participants were recruited using convenience sampling. Postpartum mothers were recruited from local health care networks, including Breastfeeding Support Lagos, the Family Youth and Health Initiative, and the Ibadan North East Primary Health Care Center in Oyo. Local partners identified and contacted eligible participants during routine antenatal and

XSL•FO
RenderX

postnatal visits, provided study details, and obtained informed consent prior to enrollment (Textbox 1).

Textbox 1. Eligibility criteria for the Mwana pilot.

Inclusion criteria

- Within 9 months post partum.
- Owns a mobile phone with SMS capability.
- Provides consent to engage with Mwana for 6 months.
- Access to the internet is required to complete the survey questionnaires.

Exclusion criteria

- Mothers without access to an SMS-capable mobile phone, as the intervention relied on text-based engagement.
- Unwilling to participate for the full 6-month period.
- Mothers with medical conditions significantly affecting breastfeeding were excluded to ensure data consistency.

The sample size of 216 mothers was determined based on feasibility and the need for a sufficiently large cohort to capture diverse breastfeeding behaviors and engagement patterns. This number ensured adequate representation across various maternal health care settings while maintaining logistical feasibility for data collection and follow-ups. Efforts were made to ensure a diverse and representative sample, including mothers from both urban and periurban areas. Participants were randomly selected based on breastfeeding status to maintain a balanced representation of exclusive and non-EBF mothers.

Among the enrolled mothers, 85.6% had been pregnant within 6 months of participation, with all within 9 months. Most had prior access to maternal health care, with 88 attending specialist hospitals, 46 visiting general hospitals, and 72 using other health care facilities.

Software and Technology

Operating Mode of the Mwana Platform

The Mwana platform was designed as an SMS-based intervention to provide breastfeeding support, eliminating the need for a mobile app. SMS was chosen due to its broad accessibility and lower technological barriers compared to smartphone apps [16]. While smartphone penetration is increasing, many individuals—particularly in low-resource settings—still rely on basic feature phones with limited internet connectivity [17]. By leveraging SMS, Mwana ensured that the intervention was available to a wider audience, making it a cost-effective and scalable solution.

In addition, SMS is a widely used and familiar communication method in Nigeria, increasing the likelihood of participant engagement [18]. Unlike mobile apps that require regular updates and compatibility across operating systems, SMS-based solutions require minimal infrastructure while maintaining high usability and accessibility. As shown in Table 1, the majority of mothers preferred SMS as their communication channel (n=206), followed by WhatsApp (n=161).

Table . Distribution of mothers' access to different communication platforms.

Category	Mother
Smart apps	74
WhatsApp	161
SMS text messaging	206

Language and Localization

To enhance accessibility, the Mwana platform was translated into Hausa and Swahili, two of the most widely spoken languages in the region. Hausa is commonly spoken across Northern Nigeria and West Africa, while Swahili is prevalent in East Africa, including among communities within Nigeria [19].

To ensure linguistic accuracy and cultural appropriateness, the translation process was conducted in collaboration with local translators who had in-depth knowledge of both language nuances and cultural contexts. This approach ensured that all messages, prompts, and interactions were not only grammatically correct but also culturally sensitive and relevant, ultimately improving user engagement and comprehension.

Mwana System Architecture

The Mwana system architecture is designed to provide structured, AI-powered maternal health support through an interactive chatbot interface. It integrates TextIt for building conversation flows and Wit.ai for backend natural language processing (NLP), ensuring smooth and dynamic interactions with users.

The system begins processing when a user submits a query, which the AI interprets as an utterance—a user-provided phrase such as "I have nipple pain." These utterances are analyzed

XSL•FO
RenderX

using a training dataset of 7000 prior utterances to identify the intent behind the query. Wit.ai, a machine learning–based NLP engine, performs this classification, mapping the user's statement to a predefined intent such as "breastfeeding pain" or "latching issues."

Once the AI successfully determines the intent, the system routes the user to the appropriate predesigned flow within TextIt. TextIt structures interactions using a decision-tree format, where users navigate a sequence of questions and responses. Each flow consists of a series of conditional prompts—typically yes or no questions—that guide users toward a resolution. For example, if a user reports nipple pain, the system may ask follow-up questions such as "Is the pain sharp or dull?" and "Does the pain persist after feeding?" to refine the diagnosis and provide relevant guidance.

By leveraging Wit.ai for intent recognition and TextIt for structured decision-tree flows, Mwana delivers a responsive and scalable maternal support system, ensuring that users receive context-aware assistance tailored to their specific concerns.

Mwana User Experience Design and Functionality

Mwana's user design is structured to provide an intuitive and supportive experience for new mothers, ensuring seamless interaction with the AI-powered system. The interface is designed with a conversational approach, mimicking human interaction to foster engagement and trust. The chatbot's structured flow categorizes user interactions into 3 key components: onboarding, daily check-ins, and problem management, allowing for a progressive and context-aware experience (Figure 1).

Figure 1. Selection of Mwana chatbot interface screenshots, illustrating different user interactions. (**A**) Greeting and onboarding screen: the chatbot introduces itself, asks for the user's preferred language and name, and reassures the user about privacy and support for breastfeeding issues. (**B**) Daily check-in screen: the chatbot engages in a routine check-in, asking about the baby's weight and the user's confidence in their breastfeeding technique. (**C**) Problem management and questioning screen: the chatbot assesses a concern regarding a crying baby by asking follow-up diagnostic questions to determine if the symptoms align with colic, ultimately providing reassurance and guidance.

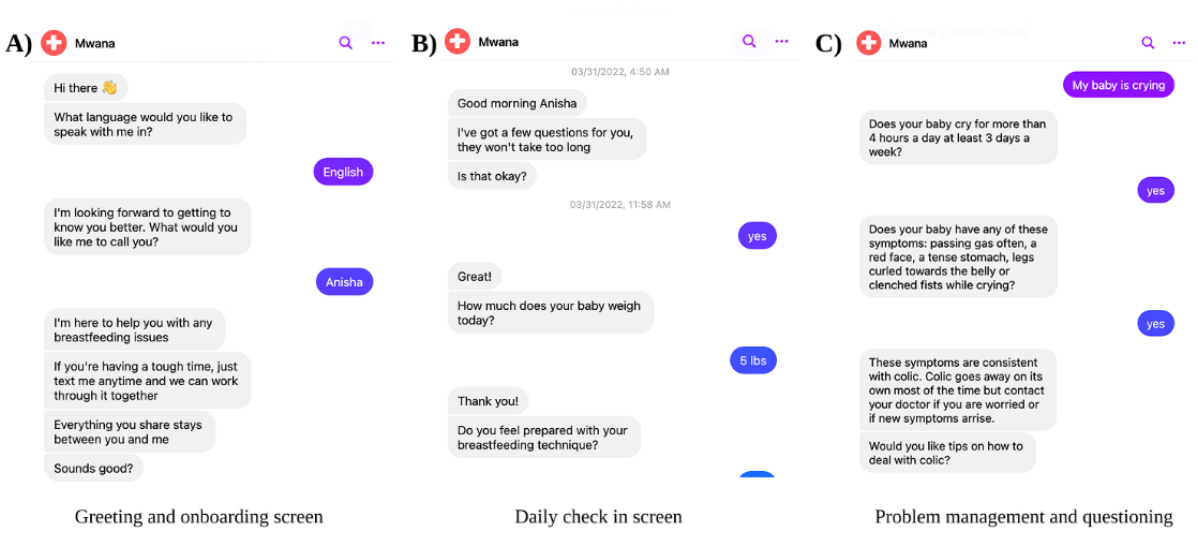

The onboarding process introduces the user to the platform, allowing personalization through language selection and name input. This personalization helps create a sense of familiarity and user-centric engagement. Daily check-ins serve as proactive touchpoints where the system collects essential health data, such as infant weight and breastfeeding confidence, reinforcing habitual engagement and real-time support. If an infant's reported weight deviates from expected growth patterns or if a mother expresses low confidence, frustration, or distress, the system flags these as potential concerns and will initiate a predefined response system for the identified issues.

The problem management system leveraged predefined flows structured as decision trees to enable the chatbot to ask targeted questions based on the user's concern. When an issue was identified, the platform guided users through automated support flows and, when necessary, escalated cases to human agents for further assistance. By integrating automated classification and escalation mechanisms, Mwana ensured timely and effective support while minimizing the need for continuous human intervention. This approach improved user experience, engagement, and accessibility, making breastfeeding support more efficient and responsive for postpartum mothers.

Ethical Considerations

This study adheres to ethical guidelines for human participants research and complies with relevant regulatory standards. It is deemed exempt from Institutional Review Board review under 45 CFR 46.104(d)(2), as it involved anonymous survey data with minimal risk to participants. In addition, the research was conducted in accordance with the Declaration of Helsinki, ensuring ethical integrity and participant protection. To safeguard privacy and confidentiality, all survey data were collected anonymously and focused solely on opinions, preventing the identification of individual participants. No compensation was provided to participants, and no identifiable individual data were included in the study. Because the survey did not collect identifiable personal information and was limited

XSL•FO
RenderX

to nonsensitive opinion-based questions, a formal IRB assessment was not pursued.

Results

Summary of User Feedback to the Mwana Platform

The feedback collected from participants in the Mwana pilot study provided valuable insights into the app's usability, usefulness, and areas for improvement. A total of 52 feedback statements were analyzed, revealing several recurring themes that shaped user experience and satisfaction. Feedback was categorized as positive, negative (eg, delayed responses, chatbot rigidity, or comprehension issues), or mixed (Table 2).

Table . Qualitative feedback categorization from postpartum mothers using the Mwana chatbot (n=52).

Feedback categorization	Percentage of sample (%)
Positive feedback	60
Negative feedback	
Delayed, no doctor response	6
Chat rigidity, comprehension issues	12
Other	12
Mixed feedback	10

Among participants who provided written qualitative feedback, the majority (31 participants) shared positive feedback, reflecting overall satisfaction with the app. Negative feedback (16 participants) primarily centered on concerns such as delays or no doctor response (3 participants), chatbot rigidity and comprehension issues (6 participants), and other unspecified issues (6 participants). A small subset (5 participants) provided mixed feedback, indicating neither strong positive nor negative feedback. These insights underscore the platform's strengths while also identifying key areas for improvement, particularly in chatbot responsiveness and AI comprehension.

Statistical Analysis of Usability and Usefulness

To assess the impact of the Mwana intervention, a statistical analysis was conducted comparing baseline and postintervention scores for usability and usefulness. The results indicate a statistically significant improvement in both categories, demonstrating the intervention's effectiveness.

Over a 6-month period, 206 of the 216 recruited postpartum mothers completed the pilot study, resulting in a 95% completion rate. Participant engagement with Mwana was measured using the average daily check-in response rate. Initially, engagement was strong, with a 90% response rate in the early weeks of the study. However, this rate gradually declined and stabilized at 60% by the end of the study period.

Analysis of engagement trends revealed that the most frequently accessed chatbot flows included "Too Much Milk," "Breastfeeding Technique," and "Nipple Pain," highlighting the key areas where mothers sought support. This suggests that users primarily relied on Mwana for immediate breastfeeding guidance.

In the final evaluation, participants rated Mwana's usability and usefulness on a 5-point scale. The majority assigned the highest rating of 5 in both categories, with 124 respondents for usefulness and 103 respondents for usability. The average usefulness score was 4.01 (SD 1.41), while usability received a slightly lower mean score of 3.92 (SD 1.35; Table 3). Notably, no participants reported injuries or safety concerns, and qualitative feedback was overwhelmingly positive, with many expressing enthusiasm for continued engagement and additional support.

Table . Distribution of usability and usefulness scores among postpartum mothers in Lagos, Nigeria (N=206).

Score	Usability response, n (%)	Usefulness response, n (%)
1	19 (9.2)	20 (9.7)
2	20 (9.7)	21 (10)
3	23 (11)	20 (9.7)
4	41 (20)	27 (13.1)
5	103 (50)	118 (57.2)

The mean usefulness score was 4.01, with a 95% CI of 3.82 to 4.20, and the mean usability score was 3.92, with a 95% CI of 3.74 to 4.10. These CIs indicate that, if the study were repeated, the true population mean would likely fall within these ranges, reinforcing the high level of user satisfaction with the intervention.

The paired *t* tests showed significant improvements in both usability and usefulness ($P<.001$), indicating that the likelihood of these results occurring by chance is extremely low. This supports the conclusion that the chatbot intervention had a real, measurable positive effect on how postpartum mothers perceived its usefulness and ease of use.

XSL•FO
RenderX

The effect size, measured using Cohen *d*, further supports the impact of the intervention. The calculated effect sizes were 0.48 for usefulness and 0.51 for usability, both indicating a moderate effect. In practical terms, this suggests that the Mwana chatbot made a meaningful difference in users' perceptions, improving their experience in a way that was both statistically and practically significant.

Discussion

Principal Findings

This study assessed the feasibility and effectiveness of Mwana, an AI-driven SMS-based intervention designed to support EBF among postpartum mothers in Lagos, Nigeria. On a scale of 5, the findings indicate that Mwana was well-received, with high usefulness (mean 4.01, SD 1.41) and usability (mean 3.92, SD 1.35) scores. More than half of the participants (57%, 118/206) rated the chatbot's usefulness at the highest level, suggesting strong demand for mobile-based breastfeeding support.

The consistently high usability and usefulness scores across the 6-month study period demonstrate the potential of SMS-based AI interventions to serve as accessible tools for maternal health support. Statistically significant improvements in both metrics ($P<.001$), along with moderate effect sizes (Cohen $d=0.48$ for usefulness and 0.51 for usability), provide compelling evidence of impact. These findings align with previous research showing that mHealth interventions can effectively support health care delivery in areas with limited resources [10,11].

A notable strength of the Mwana platform was its language localization capabilities, which enabled interactions in Hausa and Swahili alongside English. This multilingual approach expanded accessibility to diverse linguistic communities within Nigeria, addressing a critical barrier to health information delivery in multilingual settings [19]. In addition, the selection of SMS as the delivery method proved advantageous, as evidenced by the high study completion rate (95%), confirming its appropriateness in contexts with limited smartphone penetration.

Despite these positive reception indicators, the study also revealed critical challenges. The AI system struggled to accurately interpret complex user queries, leading to frustration and delayed responses. Many users reported difficulty in receiving immediate answers to their breastfeeding concerns, with some noting that they had to repeat queries multiple times before receiving a relevant response. In addition, the translation feature improved accessibility for Hausa- and Swahili-speaking mothers, but some cultural nuances in breastfeeding practices were not fully accounted for in chatbot responses.

Another key finding was engagement trends, as measured through response rates and retention over the 6-month period. While initial engagement was high, a gradual decline in response rates was observed in later months (50% reduction), suggesting that long-term user retention remains a challenge for SMS-based interventions. This could be due to the static nature of SMS interactions or the absence of multimedia content, which is known to enhance engagement in digital health interventions [20]. The most frequently accessed chatbot flows—"Too Much Milk," "Breastfeeding Technique," and "Nipple Pain"—provide valuable insights into the primary concerns of new mothers and can inform future content development for breastfeeding support interventions.

Comparison With Prior Work

This study contributes to the expanding field of AI-driven mHealth interventions by introducing an SMS-based chatbot designed to support breastfeeding. Although previous mHealth solutions have shown promise in improving maternal health outcomes, few have leveraged AI to provide real-time breastfeeding guidance in low-resource environments.

Several breastfeeding support interventions have focused on mobile apps rather than SMS. For example, the Moommae app in Thailand successfully enhanced self-management behaviors among breastfeeding mothers, but its reliance on smartphone technology limited its accessibility to low-income populations [15]. Similarly, the LactApp in Spain, developed by Maria Berruezo, Alba Padró, and Enric Pallarés, offered AI-driven breastfeeding support, but its requirement for internet access restricted its use in areas with poor connectivity. In contrast, our intervention operates via SMS, ensuring that mothers with basic mobile phones and no internet access can also benefit [21].

Our findings also compare favorably with traditional SMS-based health interventions that lack AI capabilities. While conventional SMS programs typically deliver static, scheduled messages, Mwana's interactive approach enables more personalized support based on individual user queries. This represents an important evolution in mHealth design, moving beyond one-way communication toward conversational engagement that can adapt to diverse user needs.

Other mHealth initiatives, such as iCow in East Africa, have effectively used AI-driven mobile engagement for agricultural education. While iCow demonstrated that AI can drive behavioral change through mobile messaging, its model has not been widely applied to maternal health [14]. Our study builds upon this concept by using conversational AI for real-time maternal support, marking a new era of AI-powered SMS in public health.

The observed engagement patterns in our study—high initial participation followed by gradual decline—mirror trends seen in other digital health interventions [20]. However, our retention rates remained relatively strong compared to similar studies, possibly due to the immediate practical value of the breastfeeding support provided. This suggests that when mHealth interventions address concrete, immediate needs rather than abstract health goals, they may sustain user interest more effectively.

Limitations

This study has several limitations that should be considered when interpreting the results. First, the intervention relied solely on SMS, which limited the ability to deliver rich educational content, such as images or videos, that might enhance user understanding. This could potentially impact the overall educational impact of the intervention. In addition, while

XSL•FO
RenderX

SMS-based interventions are highly accessible, participants may become desensitized to frequent messages, leading to decreased responsiveness [20].

Second, the study population was drawn from health care–engaged mothers who were already accessing maternal health services. This may limit the generalizability of findings to more vulnerable populations, such as mothers in rural areas who may not regularly visit health care facilities. Future research should explore how Mwana performs in non–health care engaged populations to assess its broader impact.

Another limitation was the chatbot's dependency on a predefined NLP dataset. Despite training the AI model on 7000 breastfeeding-related queries, gaps in language processing remained, particularly for complex or nonstandard user inputs. Expanding the chatbot's training dataset and refining NLP algorithms through machine learning could improve response accuracy and overall user experience.

Finally, the study faced occasional challenges with translator availability for Hausa and Swahili. Although the AI system provided automatic responses in these languages, user feedback indicated that some responses lacked cultural nuance, affecting trust in the chatbot's guidance. Addressing this issue will require additional cultural adaptation and expert verification of translated content.

Conclusions

This study demonstrates the feasibility of an AI-powered SMS intervention for breastfeeding support in Nigeria. Our findings indicate high usability and usefulness scores (mean 4.01, SD 1.41), aligning with our hypothesis that an AI-driven SMS intervention could serve as a feasible maternal health tool.

The intervention's strengths—including its accessibility through basic mobile phones, multilingual capabilities, and ability to provide personalized guidance—position it as a potentially valuable complement to traditional health care services in resource-limited settings. The integration of AI with SMS technology represents an innovative approach that balances technological sophistication with accessibility. By leveraging NLP capabilities within a widely available communication channel, Mwana demonstrates a pragmatic solution that doesn't require smartphone ownership or internet connectivity—key barriers that have limited the reach of previous mobile app interventions [15,21]. This approach has particular relevance for Nigeria, where basic mobile phone penetration is high (88% among women), but smartphone and internet access remain limited [13].

Future work should focus on expanding the chatbot's training dataset to improve response accuracy, integrating real-time expert consultation to supplement AI-driven guidance, and testing engagement strategies to enhance long-term retention. In addition, further research should assess the impact of Mwana among nonhealth care–engaged populations to evaluate its scalability beyond the current study setting.

From a public health perspective, these findings have significant implications for addressing Nigeria's low EBF rates. With child mortality rates remaining alarmingly high (111 deaths per 1000 live births) [1] and EBF rates at only 29% [9], scalable interventions that can effectively support breastfeeding mothers are urgently needed. Longitudinal studies measuring the impact of AI-powered SMS interventions on actual breastfeeding rates and duration would provide more definitive evidence of their public health value.

The statistically significant improvements observed in usability and usefulness, coupled with positive qualitative feedback, suggest that Mwana effectively addresses a critical gap in breastfeeding support services. Mwana represents a scalable, cost-effective solution for addressing breastfeeding support gaps in resource-limited settings. With continued development, AI-powered mHealth interventions such as Mwana have the potential to significantly improve maternal and child health outcomes in Nigeria and beyond.

Acknowledgments

Special thanks to the team at FAYOHI for providing continuous support, resources, and time that were vital in making this project a reality. To our on-the-ground partners, thank you for building with us every step of the way. Finally, thanks to all the mothers, stakeholders, and community leaders whose contributions enriched the project discussions.

Data Availability

The datasets generated during this study are not publicly available due to agreements with third-party partners, hospitals, or companies that limit distribution but are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

References

1. Nigeria. UNICEF. 2023 Dec 7. URL: https://data.unicef.org/country/nga/ [accessed 2025-06-09]
2. Pérez-Escamilla R, Tomori C, Hernández-Cordero S, et al. Breastfeeding: crucially important, but increasingly challenged in a market-driven world. Lancet 2023 Feb 11;401(10375):472-485. [doi: 10.1016/S0140-6736(22)01932-8] [Medline: 36764313]

XSL•FO
RenderX

3. Justiz AA, Wilson AM. Transient hypogammaglobulinemia of infancy. In: StatPearls: StatPearls Publishing; 2023. URL: https://www.ncbi.nlm.nih.gov/books/NBK544356 [accessed 2025-07-08]
4. Richard RM, Maziashvili G, Tran M, et al. Breast milk conferred immunity to infants against COVID-19. Cureus 2023 Jul;15(7):e42075. [doi: 10.7759/cureus.42075] [Medline: 37602015]
5. Breastfeeding. World Health Organization. 2019 Nov 11. URL: https://www.who.int/health-topics/breastfeeding#tab=tab_1 [accessed 2025-06-09]
6. Abdullahi M. Report: proper breastfeeding could prevent over 95k child deaths in Nigeria. : TheCable; 2023 Dec 7 URL: https://www.thecable.ng/report-proper-breastfeeding-could-prevent-over-95k-child-deaths-in-nigeria/ [accessed 2025-07-08]
7. Rio-Aige K, Azagra-Boronat I, Castell M, et al. The breast milk immunoglobulinome. Nutrients 2021 May 26;13(6):1810. [doi: 10.3390/nu13061810] [Medline: 34073540]
8. Agunbiade OM, Ogunleye OV. Constraints to exclusive breastfeeding practice among breastfeeding mothers in Southwest Nigeria: implications for scaling up. Int Breastfeed J 2012 Apr 23;7(1):5. [doi: 10.1186/1746-4358-7-5] [Medline: 22524566]
9. Nigeria's exclusive breastfeeding rate unacceptably low at 29%, says CS-SUNN. THISDAY. 2024 Sep. URL: https://www.thisdaylive.com/index.php/2024/09/25/nigerias-exclusive-breastfeeding-rate-unacceptably-low-at-29-says-cs-sunn/ [accessed 2025-06-09]
10. Haleem A, Javaid M, Singh RP, et al. Telemedicine for healthcare: capabilities, features, barriers, and applications. Sens Int 2021;2:100117. [doi: 10.1016/j.sintl.2021.100117] [Medline: 34806053]
11. Mishra M, Parida D, Murmu J, et al. Effectiveness of mHealth interventions for monitoring antenatal care among pregnant women in low- and middle-income countries: a systematic review and meta-analysis. Healthcare 2023 Sep 27;11(19):2635. [doi: 10.3390/healthcare11192635] [Medline: 37830672]
12. Soyombo DA. Digital health tools for breastfeeding mothers: a comparative review between Nigeria and the US. Int J appl res soc sci 2024;6(4):549-566. [doi: 10.51594/ijarss.v6i4.1009]
13. 2 in 3 nigerian women missing out on mobile internet boom. BusinessDay. URL: https://businessday.ng/technology/article/2-in-3-nigerian-women-missing-out-on-mobile-internet-boom/ [accessed 2025-03-17]
14. Marwa ME, Mburu J, Oburu REJ, et al. Impact of ICT based extension services on dairy production and household welfare: the case of iCow service in Kenya. J Agric Sci 2020;12(3):141. [doi: 10.5539/jas.v12n3p141]
15. Wang CJ, Chaovalit P, Pongnumkul S. A breastfeed-promoting mobile app intervention: usability and usefulness study. JMIR Mhealth Uhealth 2018 Jan 26;6(1):e27. [doi: 10.2196/mhealth.8337] [Medline: 29374000]
16. The differences between text and chat messaging. Sendbird. 2024 Jul 9. URL: https://sendbird.com/blog/in-app-chat-vs-sms-which-option-is-right-for-your-app [accessed 2025-06-09]
17. Feature phone and smartphone battle it out in africa. DW Akademie. 2015 May 27. URL: https://akademie.dw.com/en/feature-phone-and-smartphone-battle-it-out-in-africa/a-18465485 [accessed 2025-06-09]
18. Nigerians exchanged 2593 billion text messages in 2022, says NCC. Punch. URL: https://punchng.com/nigerians-exchanged-25-93-billion-text-messages-in-2022-says-ncc/ [accessed 2025-03-17]
19. Idiatov D, Velde MVD. Nigeria as a Linguistic Terra Incognita: The Two Languages of Lau 2020:322-328.
20. Suffoletto B. Deceptively simple yet profoundly impactful: text messaging interventions to support health. J Med Internet Res 2024 Aug 27;26:e58726. [doi: 10.2196/58726] [Medline: 39190427]
21. LactApp women health. MWC Barcelona. URL: https://www.mwcbarcelona.com/exhibitors/32456-lactapp-women-health [accessed 2025-03-17]

Abbreviations

AI: artificial intelligence
EBF: exclusive breastfeeding
mHealth: mobile health
NLP: natural language processing
UNICEF: United Nations International Children's Emergency Fund
WHO: World Health Organization

Edited by A Mavragani; submitted 06.08.24; peer-reviewed by NH Mohamad Zainal; revised version received 13.04.25; accepted 30.04.25; published 16.07.25.

Please cite as:
Musti A
Usability and Usefulness of SMS-Based Artificial Intelligence Intervention (Mwana) on Breastfeeding Outcomes in Lagos, Nigeria: Pilot App Development Study
JMIR Form Res 2025;9:e65157
URL: https://formative.jmir.org/2025/1/e65157
doi:10.2196/65157

XSL•FO
RenderX

© Anisha Musti. Originally published in JMIR Formative Research (https://formative.jmir.org), 16.7.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Acceptability of Guided Symptom Entry and Asynchronous Clinical Communication Software Among Primary Care Staff: Qualitative Study

Riina Raudne, PhD; Taavi Tillmann, PhD

Institute of Family Medicine and Public Health, University of Tartu, Ravila 19, Tartu, Estonia

Corresponding Author:
Taavi Tillmann, PhD
Institute of Family Medicine and Public Health, University of Tartu, Ravila 19, Tartu, Estonia

Abstract

Background: Patients often communicate with primary care centers remotely (eg, by telephone or email) before seeking in-person care. A comparatively novel addition might be patient-facing symptom entry websites, where subsequent questions are automatically guided by previous responses. However, the acceptability of such systems to health care staff remains unclear, particularly in terms of what features staff perceive as useful.

Objective: This study aimed to investigate a patient-facing algorithm-guided symptom-entry software (developed by Certific OÜ, Estonia), which also supports subsequent asynchronous communication, for its acceptability and perceived utility to primary health care providers.

Methods: In-depth and open-ended interviews were conducted in 8 primary care centers in Estonia, including 8 nurses and 6 doctors, 3 - 6 months after the implementation of a novel patient-facing website. Transcripts were coded inductively, using grounded theory and phenomenological approaches to uncover themes most salient to providers. Two family doctors provided feedback on the final analysis.

Results: Staff perceived unstructured communication (via email and phone calls) as a burden that increased their cognitive load. Sometimes, this arises out of the perceived mismatch between needing to identify and document critical symptom information and being unable to standardize the supply of such information, due to a heterogeneous and unpredictable communication processes whose duration, quality, and risk of miscommunication are hard to predict and control. All interviewees expressed the desire that more patients initiate their remote query via the algorithm-guided symptom-entry software. The software was reported to satisfy perceived feature needs for patient verification, privacy and data security, editable plain-language symptom summaries of symptoms, and integration with prewritten response templates (particularly for staff who were nonnative speakers). Safety of the new software was perceived as high, on account of integration alongside traditional telephone requests. Staff reported the challenge that great effort was needed to persuade patients to use the website. Among perceived challenges, some providers reported difficulty in onboarding patients, digital literacy gaps, and limited time savings. While previous research has criticized poorly designed multiple-choice systems, our findings suggest that an appropriately designed and personalized multiple-choice system can be preferable to health care staff, as they may lower cognitive demands and enhance well-being.

Conclusions: Interviewed primary health care staff felt that this symptom entry software was acceptable and desirable. They valued a perceived reduction in cognitive demands. This holds promise for increasing staff well-being and increasing efficiency, which needs to be quantified in future studies.

(JMIR Form Res 2025;9:e59620) doi:10.2196/59620

KEYWORDS

health information exchange; telemedicine; symptom entry; health portal; online consultations; cognitive load; user experience; E-Visits; asynchronous care; primary care.

Introduction

Literature Review

Historically, most primary care patients discussed most of their symptoms with a health care provider face-to-face. Usually, this was preceded by registering new patients using paper documentation, and appointments were scheduled by short phone calls with receptionists. Various information and communication technologies have disrupted this model of care, particularly after the pandemic. Initially, these focused on replacing synchronous in-person care with synchronous online care (eg by switching to a video, phone, or instant messaging communication route) where the traditional concept of a

XSL•FO
RenderX

time-limited "medical appointment" was maintained [1,2]. More recently, asynchronous technologies are emerging to complement, and in some instances, replace aspects of synchronous care. Although its terminology continues to evolve, typically these technologies span one or many domains of the following 5 chronological processes: (1) patient intake technology allows new patients to register and identify themselves online without an in-person visit; (2) symptom entry websites typically have a one-size-fits all structure and do not adapt in the light of previous responses; (3) symptom data analytics include self-triage and the provision of automated medical advice; (4) symptom data transmission software allows patient-entered data to be read and actioned by clinicians at an asynchronous timepoint; and (5) patient-clinician communication platforms enable bidirectional message exchange in either a synchronous or asynchronous manner [3].

These 5 domains could be seen as modular, whereby health care providers may use multiple IT providers whose functions may overlap with one another, and some of the 5 domains may not be provided. Sometimes, the integration of many of these domains may allow simpler and lower-risk queries to be handled and closed without requiring an in-person visit, and this is sometimes referred to as electronic visits or (e-visits) [4].

Academic study of principally asynchronous clinical care technologies may help inform applied questions of if, where, when and how to adopt certain technologies and how these technologies may be improved [1,5]. Patient satisfaction with asynchronous care appears high [6-8]; however, clinician acceptability has been less studied. The few quantitative studies to date have found mixed findings, with some suggesting that clinicians may prefer asynchronous to synchronous models of communication [9], while others reporting low levels of acceptability [10-13], potentially due to increased cognitive exhaustion from taking in information in new formats [14-16]. One theoretical model suggested that online consultation technologies could ease physician burdens if they do not spur higher demand from patients and can resolve half of patients' queries without face-to-face follow-up [17]. However, most platforms have not met these criteria in evaluations, thereby risking exacerbating physician workloads. What else clinicians want from such software, and how the technology can deliver this, remains understudied and unclear [6]. Previous research has identified some features that have facilitated greater acceptability by clinicians, such as direct integration with electronic health record (EHRs), accurate patient identification, trust in privacy and security, and quasi-automated triage [7,8,11], more complex workflow such as initial nurse review before escalating to physicians, and free text options [1,17].

Certific's Guided Symptom-Entry and Asynchronous Communication Software

Patients and clinicians access a secure website using any internet-enabled device (incl. smartphones and PCs), without needing new software or hardware. A multifactor authentication process is designed to be tailored for each country. In Estonia (the only country where Certific currently operates), all residents are identified by a 11-digit personal identification code, which users operationalize for digital identification by one of 3 hardware options: Smart-ID is a phone app for smartphones that is initially activated in-person at a police station or bank branch; Mobile-ID is an SMS text messaging-based verification service activated by telecommunications operators, making it suitable for older mobile phones; and the combination of a PC or mac with a USB-card reader, plus a credit-card sized national ID-card, allows for digital identification for those without mobiles (including at public PCs).

Patients select their presenting complaint from a list of available options (eg, headache). A short adaptive questionnaire covers related symptoms (eg neck stiffness), onset, and a text box. The pathways are tailored to dynamically adapt in light of the severity of the data entered, with intent of reducing the need for follow-up queries and supporting faster decision-making. As optional features, patients can submit a health complaint on behalf of a relative or a next of kin, and upload images.

The software does not provide triage or urgency categorization. It lists the presenting complaint in the query title, thereby facilitating clinicians when performing triage as they scan lists of query titles. After opening a query, clinicians view a structured, plain-language summary of the presenting complaint. In case of a mismatch between the languages used by patients and clinicians, clinicians may opt to view a machine translation version (provided by DeepL). They can edit this summary before manually copying and pasting it into an external EHR system. Depending on the complexity of the query, clinicians can close it outright, send an instant (or delayed) message (or SMS text message) to the patient, or invite them for an appointment. Many of these can be streamlined by using predefined or user-generated templates.

Once a clinician has responded, patients receive a short email notification directing them to view the detailed message on the secure platform. Patients may also respond to the clinician's message (if permitted by the clinician).

The software has been available in Estonia since the beginning of 2023, with 69 primary care centers (out of the country's 400, giving about one-fifth market share) subscribing to its services at the end of 2023, when this study was done. For further background about the types of patients who were using the system during this time, the IT service provider (Certific OÜ) has described their demographics and presenting complaints in Multimedia Appendix 1. This appendix also lists further technical details and screenshots that illustrate the above text.

For additional context, many Estonian primary care practices have permitted unverified individuals to email their practice with medical queries since the late 2000s. In 2014, the Estonian National Health Insurance Fund (which is similar to the National Health Service in the United Kingdom) formally recognized e-consultations (when general practitioners ask advice from other specialists using remote asynchronous communication like email) as an official and billable health care activity. Thus the culture of asking for medical advice over email has been long widespread in the Estonian health care system.

This study investigates the perceived experiences, utility, and acceptability reported by primary care clinicians shortly after the adoption Certific's software, which enables algorithm-guided

XSL•FO
RenderX

symptom-entry for patients, and its asynchronous medical review and communication by clinicians.

Methods

The Research Team

The core research team consisted of the 2 coauthors of this article. We had quarterly meetings with 2 staff working at the IT provider (Certific), Andres Lasn and Helena Volmer, whose input helped inform the study design, interview guide, and sampling approach. All interviews were conducted, and the qualitative analysis was led by one academic experienced in qualitative research (first author RR) with a public health background, either in person or via video conferencing. The researcher did not have previous expert knowledge on web-based consultation software, but was experienced with health services, and generic software development, so she conducted interviews as someone professionally familiar with the health system. The researcher was employed exclusively at the affiliated university to work on this project; there was no prior or subsequent professional or personal relationship with Certific. The second researcher (TT) is an academic clinician experienced in public health interventions, including the design, implementation, and evaluation of digital public health tools. He has previously not worked with web-based consultation software for symptomatic patients. He provided the theoretical context for the study and helped interpret findings in light of existing digital health literature.

Certific assisted with drawing up the purposive sample and sending out introductory recruitment emails. The interviewer had no prior relationship with any of the interviewees. The participants were told by Certific that a researcher from the University of Tartu is seeking feedback on provider experiences with Certific, as part of a qualitative study. Both researchers were motivated to work on this project out of interest in innovation in health services and an opportunity to learn and contribute to health technology development. They are otherwise not involved with, and have no competing interests in, this or similar IT solutions that address the health care needs of symptomatic patients in primary care.

Study Design

This qualitative study assessed the perceived feasibility, utility and acceptability among primary health care staff in Estonia during the first months after implementing Certific's communication technology. Its open-ended design sought primary care providers' rich descriptions and own explanations, perceptions, mental models and framings. Rather than using and potentially departing from established theories, frameworks and metrics of what might potentially makes one software acceptable to primary care staff, we asked participants to tell us in their own words what they perceive remote consulting to mean, what are its aims, channels, and which tools allow them to best achieve what they are trying to do.

Grounded theory and phenomenological approaches were employed in all phases of study design to prioritize providers' own descriptions of challenges in their practice and experiences (phenomenological aim) with remote consulting, including Certific. Emergent themes were also systematically analyzed (grounded theory aim). This analytic approach was preferred to more predefined evaluation frameworks, because one of the aims in our funding proposal (which we had to adhere to in delivering the research) was to use this independent formative evaluation to quickly contribute to software development, by feeding back important user needs to the software developer. Accordingly, open-ended interviews with relatively few questions and only one researcher conducting the interviews were chosen as a method of data collection. As compared to focus groups or interviews with several researchers, conversations with solo researchers have limitations (see Discussion); however, they can also increase privacy and trust, potentially leading to more disclosive conversations.

Recruitment

Participants were purposively sampled from among Certific's recently onboarded clients. In a relatively small and homogeneous country like Estonia (with just 1.3 million inhabitants), we sought maximum variety in practice location, size, and demographics. Recruitment and interviews took place from May to September 2023, involving both native-Estonian and native Russian-speaking professionals proficient in Estonian. Invitations were sent via email to primary care physicians and nurses, with 6 physicians and 8 nurses consenting to participate. One invited recipient declined, due to increased workload owing to the summer vacations. Altogether, 14 face-to-face interviews were conducted either in person or via video conferencing. There was one repeat interview with a doctor at the end of the initial analysis to check our interpretation of core findings. One interview involved two nurses who wished to be interviewed together.

Data Collection

The interviews, which were audio-recorded and transcribed, explored participants' views on the current realities and changing nature of their roles and their practices, especially in the context of the increasing role of remote consultations. The openly framed questions of the interview guide (Textbox 1) sought provider perceptions of strengths and weaknesses of using different remote communication tools such as the telephone, email, and other online consultation platforms, if applicable, in the context of what providers sought to achieve. Finally, we enquired about the specific experiences with Certific.

XSL•FO
RenderX

Textbox 1. Interview guide used in the interviews

Introduction:

- Please tell us a little bit about your role and your practice. How is it changing?
- What are the current main challenges?

On the rise of remote consulting:

- How is remote care organized in your practice? What channels do you use for patient communication?
- What are the benefits and drawbacks of each communication channel?

- Phone
- Email
- Online consultation software

On Certific:

- Can you try to summarize, with very short and simple words what Certific's software is and does?
- How and why did your practice decide to adopt this software?
- What has been your experience with it so far?
- What are its results and impacts, for example, has it changed workflow?
- What are the upsides and downsides of Certific's software when compared to other options?
- Can you walk me through exactly what happens when a patient reaches out to you via Certific?
- What other tools do you have to complete one patient inquiry?
- What is missing at the moment from this software?

Data Analysis

Interviews were transcribed verbatim and coded by the same researcher (RR) who conducted the interviews using Atlas.ti software. Analysis involved a systematic reflection on the collected data during the interview and coding phases. Some themes were identified beforehand and resulted from the interview guide (such as the benefits and drawbacks of phone, email, and Certific software in solving patient's queries). Other themes and surprising subthemes emerged during the analysis, such as perceived increases in workload arising from greater patients' (remote) access to care; overuse of medical services; desire for greater gatekeeping to prevent undue use; variable digital literacy among patients; unpredictable quality of email queries; and descriptions of the mental load while staff try to identify and categorize key diagnostic features.

Emergent themes and findings were discussed throughout, during meetings with the study team and with Certific to provide feedback. A version of the analysis was also presented to 2 family doctors, one of whom had also been interviewed, for sense-checking and feedback. When drafting this article, we followed the "Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups" which is provided in Checklist 1.

Ethical Considerations

Formal ethics approval was not required according to Estonian regulations and the Oviedo Convention [18], as this study involved voluntary interviews with health care professionals about their professional roles, posed minimal risk, and collected no sensitive personal or patient data. All participants gave informed consent for recording, transcription, and pseudonymized publication of quotations. Participants were not compensated for their time (as this took place during usual work hours). This approach aligns with the Estonian Code of Conduct for Research Integrity [19], which stipulates that researchers must obtain ethics approval when required by law or institutional rules, but allows for researcher discretion in low-risk studies involving professionals and nonsensitive data. In addition, the study complies with the principles outlined in the Oviedo Convention (Convention on Human Rights and Biomedicine), which Estonia ratified in 2002. The Convention emphasizes the protection of human rights in the biomedical field and allows for national discretion in ethical oversight of research not involving medical interventions or sensitive personal data [20].

Role of the Funder

The study was funded by the Estonian government's Applied Research Program. Specifically, this qualitative study was one part of a larger grant made available to the software developer, to provide an external and independent academic assessment.

Results

Overview

We interviewed 14 primary care providers (8 nurses and 6 doctors) between June and September 2023. Certific's software had been in use for 3 - 6 months before our interview. The number of patients per center ranged from 1500 to 10,000, with staff numbers ranging from 2 to 16. All centers used telephone, website, Gmail-based email, and Certific's software for remote

XSL•FO
RenderX

communication; 3 clinicians had also used other remote consultation software. During the initial open-ended interviews, some new themes (with example types that turned into codes) emerged such as provider cognitive load and the communication patterns of patients. Time and follow-up attention was afforded for these themes to be fully elaborated in subsequent interviews. Saturation was noted in all reported themes.

The following is a summary of selected themes that emerged in data analysis. Examples of coding trees are presented in Multimedia Appendix 2.

How Remote Consultations Have Changed Doctor-Patient Communication and Demand

Interviewees described significant changes in patient communication during the recent years Multimedia Appendix 3. While patient uptake of remote consultations varied, COVID-19 had made remote consulting the new default. Several practices had stopped and never reintroduced walk-ins without previous appointments.

Remote Consulting Using Email

Face-to-face contact enables providers to quickly carry out multisensory examinations of patients' signs and symptoms. Their omission over email revealed a significant variability in patients' ability to communicate their symptoms in writing. This creates additional work for staff, who have to identify, categorize, identify uncertainty, and sometimes narrow this by explicitly asking about critical diagnostic features from lengthy verbose texts or vague, unsigned and sometimes impolite one-line questions that echoe chat culture. Often, patients' emails were incomplete, missing critical administrative details or symptoms necessary to complete a task.

> *Well, this patient just wants to tell me all their concerns in one big sitting, but I have no use for all this excess information… this kind of massive letter I can't even copy and I have to spend considerable time in distilling it…* [GP 4]

> *The aim of good software might be that such patients' can't write all those endless lengthy emails* [GP 3]

Staff are required to record a summary of each patient interaction in an EHR system. Staff said that it had become a significant challenge during the remote era to interpret patients' written queries and asynchronously follow up for isolated critical details across multiple channels (such as a long email, followed by a short telephone call) to quickly compile a sufficient summary of the patient's symptoms.

In general, interviewees were not concerned about remote access for older and less digitally literate patient groups, having witnessed their adaptation to email over the past decade. However, providers were more concerned about the ways in which email in particular appeared almost too convenient of a method for reaching out to doctors and they perceived an increased demand they considered sometimes as frivolous and unnecessary:

> *There are some people who are completely helpless. One week they call and say their child has a fever. I ask if they've taken their temperature and given paracetamol. They haven't. So I tell them to give some paracetamol, use steam, etc. Our nurse explains everything in detail, which takes about 10 minutes. They then want a sick note. The child recovers. Three weeks later, the same child develops the same symptoms, and we go through exactly the same steps once again. The nurse repeats the exact same advice, asks if they took paracetamol. They have not. I understand if you call once a year, but calling twice a month to be told to take paracetamol is ridiculous. This could easily be handled by a chatbot.* [GP 4]

> *The downside of email is that someone can have a thought at 1 AM, and they pour it all into an email. Then in the morning, I read it all and think, 'Wow, someone was in a bad mood last night.' But with remote consultation where you have to first log in and jump through hoops, people don't bother as much. Emails are just too accessible; everything comes through. It's too easy to send messages, and I already know some people here who vent a lot in their emails. I know I don't have to respond to everything, and sometimes I think it's nice that people can vent, but it's simply overwhelming.* [Nurse 8]

Remote Consulting Using Phone

Providers described synchronous phone-based communication to have both advantages and disadvantages. First, hearing vocal cues brought back additional multisensory signs that staff have been trained for, and increased confidence in their diagnostic abilities.

Second, synchronous interactions allow perceived information gaps to be completed, follow-up appointments to be arranged, and understanding to be checked quickly.

The option for patients to phone the practice was perceived as the most accessible alternative to in-person appointments. As practices were contemplating options such as restricting email access and walk-ins, knowing that their practice offers telephone access as a back-up to those in need provided a relief for concerns about potential unequal access.

As drawbacks, telephone calls required simultaneously focusing on listening and typing for documentation, the potential for information to get lost in this transfer from speech to text, and disruptions into other ongoing tasks that call-backs caused. Participants noted the stress of handling emotional or upset patients over the phone, where they had to simultaneously listen, document, diagnose, and search for helpful information and communication routes while maintaining adequate awareness and control over their own feelings.

The combination of email and phone created a time-pressed, stressful, and fragmented flow of communication, particularly for nurses who typically were frontline responders.

Especially during the flu season, several providers reported coming to work on Monday mornings and finding an overflow of email queries, around 50 to handle in one day. Larger practices had established clearer workloads, such as "never letting one person respond to remote queries for more than 4

XSL•FO
RenderX

hours at a time." Smaller practices did not have such luxury of choice, and nurses expressed a sentiment of despondent acceptance.

First and Subsequent Impressions

> *At the beginning I thought 'Gosh, not another login, not another window to look at. But in hindsight I was mistaken. Now I really like this program.* [Nurse 6]

Data Security and Identity Verification

In most cases, Certific was first adopted for its data security and privacy capabilities, as Gmail was deemed insufficient for meeting legal requirements.

> *Frankly, Gmail is still a little bit more convenient and familiar, but we know that it's not secure. That's why we adopted Certific. We had to make a choice as something had to replace Gmail.* [GP 1]

The use of national ID for login streamlined patient identification, reducing the burden of confirming identities from ambiguous emails.

> *Sometimes I get an e-mail from, say 'abcde999@hot.ee' (a made-up email address) and then I have to begin from writing back to them and asking what their name is and whether they are really our patient in the first place.* [GP 4]

On the other hand, the identification requirement on the patient side was also deemed by the providers to constitute a suitable barrier to appropriately help reduce excess demand.

> *The ease of writing an email leads people to send messages impulsively—an idea pops into their head, and they immediately write to us. Certific requires a bit more effort from them. Now, sometimes we get people writing to us with just one sentence, without even a greeting or closing. Communicating via Gmail is too convenient; its liberties are frankly sometimes abused. A single thought comes to the patient's mind, and they write to the doctor straight away.* [GP 2]

Structure and Clarity in Communication

Participants reported that Certific's software brought more structure and clarity to patient communications. The software's guided questions on symptoms was perceived to provide in most cases a comprehensive summary for subsequent triage, and to also educate patients about the information that clinicians might need:

> *What I like is that patients approach us through guided questions. This forces the person to give a little thought to their request, those guiding questions are very good. On the phone or via email they will just say the first things that comes to their mind.* [Nurse 2]

In addition, participants noted that the requirement to log in provided a barrier to excessive convenience.

> *It kind of disciplines the patient. Otherwise, we have to ask over three or four emails where to direct their attention. But now they can directly just talk to the machine…* [Nurse 4]

Documentation Ease and Standard Response Templates

Interviewees appreciated Certific's structured and plain language summaries, which filtered out less relevant details. This reduces the effort needed to synthesize a summary and it was perceived to free up time to focus on speaking with patients. The editable templates and links to guidelines reduced workload, particularly during high-volume periods.

> *I think nurses have to answer a lot of questions from people with colds and coughs. Now that they have standard template answers, I think things have improved for them.* [GP 6]

Reduction of Emotional Burden

Patient queries through Certific's software provided relatively complete case facts, minimizing the need for follow-up questions and allowing pauses before responding to patients. This reportedly mitigated stress from continuous communication with unwell patients and occasional encounters with rudeness.

Time-Saving Perceptions

When asked whether the use of Certific OC platform saves time, most providers were not sure. In their perception, Certific's software had nonetheless improved the ease, structure, and clarity of case management, contributing to a more efficient workflow.

> *I cannot say that I am spending less (time solving a patient's case) than on the phone or on e-mail, I think it'll end up around the same amount of time. It's just that… now it's very convenient that when I respond, I click-click-click on prefilled fields… I cannot entirely automate it, I still have to focus and think about it, and add little notes.* [Nurse 7]

Challenges With Implementation and Patient Reception

Despite the improvements, remote consultation on a platform such as Certific's could not, according to participants, replace the immediacy of face-to-face consultations and telephone, which participants considered best for lowering diagnostic risks by getting extra sensory input about a patient's condition through the voice, body language, and general vitality.

Concerns were raised by some staff about the patient digital literacy required for using the software (eg logging in), especially among older patients in smaller towns. However, other staff elaborated how phone and email were still sufficient options for these patients, and they did not express accessibility concerns. Some patients reportedly opposed the mandatory requirement of accepting "cookies" on Certific's website, leading them to opt-out of the service.

Patient Onboarding

Informing patients about adopting Certific's software was described as slow and complex, with centers lacking the resources for marketing and onboarding. During data collection,

XSL•FO
RenderX

patients mainly learned about this communication route reactively, when approaching their primary physician often in person, as few proactive and systematic communication attempts were made by centres due to incomplete email lists.

Furthermore, a temporary technical inability to transfer all patient contacts to Certific at the time of conducting the interviews, requiring manual contact entry, had temporarily slowed workflows and reduced motivation to invest time in the platform. This problem has since been solved, alongside a new onboarding process that allows SMS text messaging notification campaigns to all registered patients; email footers, prerecorded phone messages, and website links.

Discussion

Principal Findings

In this qualitative study, we found that primary care doctors and nurses in Estonia supported novel technological solutions to the historical problem of overwhelming unstructured written information when patients emailed practices with free text that often led to long-winded need for clarifications. The most useful feature that persuaded clinicians to consider trying the software in the first place was its ability to deliver a high degree of security and privacy against the risk of sensitive medical information being leaked via insecure means, such as Gmail servers residing around the world. Though participants described greater ease of managing patient cases using Certific, this did not translate into provider perceptions of saved time. Rather, the main perceived benefits after having adopting the platform after a few months came from the number of cognitively challenging tasks that had been automated (eg summarizing patient symptom descriptions, providing pre-written response templates and links to treatment guidelines, reducing the need to formulate replies for frequent problems from scratch).

Comparison With Previous Work

Previous research has identified technological features that facilitate greater acceptability by clinicians. Some of these features also emerged in our study, adding consistency to their transferability, such as direct data transfer into the electronic health record and other integrations; initial review by nurse before escalating to physician [21]; accurate identification of patients, privacy, security and confidentiality [20]; and quasi-automated triage including self-care and in-person options [1]. However, while previous studies suggested that free-text entry is superior for capturing patient queries [1,22], our findings contradict this. Clinicians in our study felt overwhelmed by free text and preferred the plain language summaries generated from multiple-choice symptom-entry systems, which minimized the need for follow-up.

The design of Certific's multiple choice system, which guides patients through thematic and adaptive symptom-gathering pathways and ensures all guideline-relevant questions are asked, seemingly improves the signal-to-noise ratio of the output summary [10,14,23]. A related consideration is how appropriately the remote system is embedded into other modalities of care. Excessively strict gatekeeping of in-person care can lead to patients manipulating their symptoms to get appointments [7,10], but this was not an issue in our study as in-person appointments were usually available within a week.

Suggestions for Future Research

First, our findings suggest that a well-designed, guided and highly personalized multiple choice symptom entry system may for staff lower improve cognitive demands, and increase workplace satisfaction and well-being. Future studies could evaluate this quantitatively, using questionnaires or absence data, ideally using a randomized design. Second, there remains considerable uncertainty whether this may increase or decrease total work time. As it is difficult to technologically measure how many seconds a staff member uses one or 2 IT systems (especially if these are being used in parallel) while concurrently responding to email, telephone calls, or in-person conversations, a traditional design using an observer with a stopwatch may help shed light on this question. Third, consistent with previous studies, we observed patients who used the system had more younger and female patients. This risks marginalizing older and male patients. Service developers could consider explicitly involving them, to design services around this demographic. Other dimensions of equity (including patients' education, income, and ethnicity, etc) should also be monitored and services developed in an inequality-reducing manner, where possible.

Limitations

This study has several constraints, which may have lowered the trustworthiness and credibility of our findings. First, the limited sample size, stemming from a small number of potentially "early adopter" primary care practices who had employed this software at this early phase, in one country, could limit broader transferability. More specifically, we conducted comparatively few interviews with those health care staff such as those who were working in rural areas, male staff members, and doctors. Second, the study did not investigate how acceptable the software was to patients. Third, centres participating in the study had invested effort and finances into Certific, potentially introducing plan continuation bias, the psychological need to defend a previous decision. Fourth, a single researcher both collected the data and analyzed it. Despite regular team discussions, the single-researcher approach (while offering certain advantages) may have increased the risk of personal perceptions and biases influencing data collection and analysis. For example, this interviewer's background and interest in patient journeys and user journey may have influenced the choice of themes that she found most relevant (both in terms of deciding which clues to further elaborate during the interview, and which themes were pursued during the analysis (refer Multimedia Appendix 2 - coding tree). Furthermore, empathy towards health care providers and their communication struggles grew over the course of the interviews, possibly impacting the analysis by siding with the health care staff members' perspective. However, these limitations were partly offset by all interviewees having no prior contact with the interviewer, and the interviewer being broadly trained and experienced in understanding and managing such biases as best as possible.

XSL•FO
RenderX

Conclusions

Primary health care staff in Estonia felt that Certific's guided symptom entry software, which facilitates subsequent asynchronous clinical communication, was acceptable and desirable. When compared to the conventional alternative of emails and telephone calls, staff initially felt an interest in trialing the software due to its superior security and privacy features. After a few months of use, they reported a perceived reduction in cognitive demands when compared to email and telephone routes. This holds promise for increasing staff well-being, which could be assessed in future studies. Staff did not explicitly perceive the software to lead to time savings, which also be assessed in future studies.

Acknowledgments

The study was funded by the Estonian government's business development grant scheme, specifically Small Grants of the Applied Research Program. TT is additionally supported by the Estonian Research Council's Starting grant (no. PSG809). The funding sources had no role in the design, conduct, or analysis of the study, or in the decision to submit the manuscript for publication.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request, provided that participant confidentiality (ie, non-identification) can be maintained.

Authors' Contributions

TT conceived the presented idea. RR conducted the interviews, analyzed the data, and led writing, with cowriting from TT.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Technical details of Certific's system (Design, Ux, UI and workflow), plus presenting complaints made by typical patients.
[DOCX File, 4455 KB - formative_v9i1e59620_app1.docx]

Multimedia Appendix 2
Coding tree used in the analysis.
[DOCX File, 10 KB - formative_v9i1e59620_app2.docx]

Multimedia Appendix 3
Description of the sample used in recruitment of interviewees.
[DOCX File, 8 KB - formative_v9i1e59620_app3.docx]

Checklist 1
Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups - completed by the study authors.
[XLSX File, 18 KB - formative_v9i1e59620_app4.xlsx]

References

1. Darley S, Coulson T, Peek N, et al. Understanding how the design and implementation of online consultations affect primary care quality: systematic review of evidence with recommendations for designers, providers, and researchers. J Med Internet Res 2022 Oct 24;24(10):e37436. [doi: 10.2196/37436] [Medline: 36279172]
2. Campbell K, Greenfield G, Li E, et al. The impact of virtual consultations on the quality of primary care: systematic review. J Med Internet Res 2023 Aug 30;25(25):e48920. [doi: 10.2196/48920] [Medline: 37647117]
3. Fuster-Casanovas A, Vidal-Alaball J. Asynchronous remote communication as a tool for care management in primary care: a rapid review of the literature. Int J Integr Care 2022;22(3):7. [doi: 10.5334/ijic.6489] [Medline: 36043028]
4. Nguyen OT, Alishahi Tabriz A, Huo J, Hanna K, Shea CM, Turner K. Impact of asynchronous electronic communication-based visits on clinical outcomes and health care delivery: systematic review. J Med Internet Res 2021 May 5;23(5):e27531. [doi: 10.2196/27531] [Medline: 33843592]
5. Salisbury C, Murphy M, Duncan P. The impact of digital-first consultations on workload in general practice: modeling study. J Med Internet Res 2020 Jun 16;22(6):e18203. [doi: 10.2196/18203] [Medline: 32543441]
6. Mold F, Hendy J, Lai YL, de Lusignan S. Electronic consultation in primary care between providers and patients: systematic review. JMIR Med Inform 2019 Dec 3;7(4):e13042. [doi: 10.2196/13042] [Medline: 31793888]

XSL•FO
RenderX

7. Carter M, Fletcher E, Sansom A, Warren FC, Campbell JL. Feasibility, acceptability and effectiveness of an online alternative to face-to-face consultation in general practice: a mixed-methods study of webGP in six Devon practices. BMJ Open 2018 Feb 15;8(2):e018688. [doi: 10.1136/bmjopen-2017-018688] [Medline: 29449293]
8. Eccles A, Hopper M, Turk A, Atherton H. Patient use of an online triage platform: a mixed-methods retrospective exploration in UK primary care. Br J Gen Pract 2019 May;69(682):e336-e344. [doi: 10.3399/bjgp19X702197]
9. Waschkau A, Traulsen P, Steinhäuser J. Evaluation of synchronous and asynchronous telemedical applications in primary care in rural regions of Northern Germany-results and lessons learned from a pilot study. Int J Environ Res Public Health 2022 Nov 11;19(22):14860. [doi: 10.3390/ijerph192214860] [Medline: 36429577]
10. Banks J, Farr M, Salisbury C, et al. Use of an electronic consultation system in primary care: a qualitative interview study. Br J Gen Pract 2018 Jan;68(666):e1-e8. [doi: 10.3399/bjgp17X693509] [Medline: 29109115]
11. Baines R, Tredinnick-Rowe J, Jones R, Chatterjee A. Barriers and enablers in implementing electronic consultations in primary care: scoping review. J Med Internet Res 2020 Nov 12;22(11):e19375. [doi: 10.2196/19375] [Medline: 33035177]
12. Edwards HB, Marques E, Hollingworth W, et al. Use of a primary care online consultation system, by whom, when and why: evaluation of a pilot observational study in 36 general practices in South West England. BMJ Open 2017 Nov 22;7(11):e016901. [doi: 10.1136/bmjopen-2017-016901] [Medline: 29167106]
13. Adler-Milstein J, Zhao W, Willard-Grace R, Knox M, Grumbach K. Electronic health records and burnout: Time spent on the electronic health record after hours and message volume associated with exhaustion but not with cynicism among primary care clinicians. J Am Med Inform Assoc 2020 Apr 1;27(4):531-538. [doi: 10.1093/jamia/ocz220] [Medline: 32016375]
14. Ambrose L. Remote consulting: recognising the cognitive load. Br J Gen Pract 2020 Jun;70(695):295. [doi: 10.3399/bjgp20X710213] [Medline: 32467208]
15. Tai-Seale M, Dillon EC, Yang Y, et al. Physicians' well-being linked to in-basket messages generated by algorithms in electronic health records. Health Aff (Millwood) 2019 Jul;38(7):1073-1078. [doi: 10.1377/hlthaff.2018.05509] [Medline: 31260371]
16. Kroth PJ, Morioka-Douglas N, Veres S, et al. The electronic elephant in the room: physicians and the electronic health record. JAMIA Open 2018 Jul;1(1):49-56. [doi: 10.1093/jamiaopen/ooy016] [Medline: 31093606]
17. Moschogianis S, Darley S, Coulson T, Peek N, Cheraghi-Sohi S, Brown BC. Patient experiences of an online consultation system: a qualitative study in English primary care post-COVID-19. Br J Gen Pract 2024 Jul;74(744):e489-e497. [doi: 10.3399/BJGP.2023.0076] [Medline: 38164550]
18. Council of Europe. Convention for the Protection of Human Rights and Dignity of the Human Being with regard to the Application of Biology and Medicine: Convention on Human Rights and Biomedicine. 1997 Apr 4. URL: https://www.coe.int/en/web/human-rights-and-biomedicine/oviedo-convention [accessed 2025-04-16]
19. Estonian research council. Estonian code of conduct for research integrity. 2023. URL: https://eetika.ee/en/content/estonian-code-conduct-research-integrity [accessed 2025-04-16]
20. Hanna L, May C, Fairhurst K. The place of information and communication technology-mediated consultations in primary care: GPs' perspectives. Fam Pract 2012 Jun;29(3):361-366. [doi: 10.1093/fampra/cmr087] [Medline: 22006040]
21. Davis K, Schoenbaum SC, Audet AM. A 2020 vision of patient-centered primary care. J Gen Intern Med 2005 Oct;20(10):953-957. [doi: 10.1111/j.1525-1497.2005.0178.x] [Medline: 16191145]
22. Gardner RL, Cooper E, Haskell J, et al. Physician stress and burnout: the impact of health information technology. J Am Med Inform Assoc 2019 Feb 1;26(2):106-114. [doi: 10.1093/jamia/ocy145] [Medline: 30517663]
23. Farr M, Banks J, Edwards HB, et al. Implementing online consultations in primary care: a mixed-method evaluation extending normalisation process theory through service co-production. BMJ Open 2018 Mar 19;8(3):e019966. [doi: 10.1136/bmjopen-2017-019966] [Medline: 29555817]

Abbreviations

COREQ: Consolidated criteria for reporting qualitative research
EHR: electronic health record

Edited by A Mavragani; submitted 18.04.24; peer-reviewed by S Thawillarp, VM Nymberg; revised version received 16.04.25; accepted 23.04.25; published 16.07.25.

Please cite as:
Raudne R, Tillmann T
Acceptability of Guided Symptom Entry and Asynchronous Clinical Communication Software Among Primary Care Staff: Qualitative Study
JMIR Form Res 2025;9:e59620
URL: https://formative.jmir.org/2025/1/e59620
doi:10.2196/59620

XSL•FO
RenderX

© Riina Raudne, Taavi Tillmann. Originally published in JMIR Formative Research (https://formative.jmir.org), 16.7.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Breath-Focused Mindfulness and Compassion Training in Parent-Child Dyads: Pilot Intervention Study

Satish Jaiswal[1], PhD; Jason Nan[1], PhD; Seth Dizon[1], BS; Jessica O Young[1], BS; Suzanna R Purpura[1], MS; James K Manchanda[1], BS; Dhakshin Ramanathan[2], MD, PhD; Dennis J Kuo[1], MS, MD; Jyoti Mishra[1], MBA, PhD

[1]Department of Psychiatry, University of California, San Diego, 9500 Gilman Drive, Guava Building, Room 130, La Jolla, CA, United States
[2]VA San Diego Medical Center, San Diego, CA, United States

Corresponding Author:
Satish Jaiswal, PhD
Department of Psychiatry, University of California, San Diego, 9500 Gilman Drive, Guava Building, Room 130, La Jolla, CA, United States

Related Article:

This is a corrected version. See correction statement: https://formative.jmir.org/2025/1/e81122

Abstract

Background: Depression in children is a concerning societal issue and can be associated with poor academic performance, school dropout, and poor overall quality of life. Additionally, child depression is often associated with parallel stress and depression in parents.

Objective: This scenario highlights the urgent need for the development and implementation of accessible and scalable solutions that may cobenefit child and parent mental well-being.

Methods: This pilot study introduced "Cooperative Compassion" (CoCo), a parent-child cotraining digital application aimed at promoting mindfulness and compassion through brief, performance-adaptive sessions. A community sample of 24 parent-child dyads (children's mean age 9.5, SD 3.27 years; female: n=14, male: n=10; Asian: n=5, White: n=11, mixed race: n=7, other race: n=1; and parents' mean age 44.5, SD 6.5 years; 20 female: n=20, male: n=4; Asian: n=8, White: n=14, mixed race: n=2) of high average affluence socioeconomic scores participated in the study. These parent-child dyads completed 30 sessions of CoCo training over 3 months with baseline and postintervention assessments occurring within 2 weeks of training initiation or completion, respectively.

Results:: The program was feasible, with 80% (n=19) of families completing over 90% (n=22) of sessions and providing positive feedback. Mental health assessments showed a nonsignificant effect in the expected direction in children's depression scores (Cohen d=−0.19; 95% CI −8.89 to 1.74; P=.07) and significant reductions in parental stress (d=−0.41; 95% CI −2.63 to −0.16; P=.02), anxiety (d=−0.47; 95% CI −2.67 to −0.20; P=.02), and depression (d=−0.50; 95% CI −3.25 to −0.08; P=.03), with sustained benefits at the 3-month follow-up. Parental mindfulness improvements were correlated with stress reduction (ρ=−0.45; P=.03). On an emotion bias task used as an objective assessment of cognition, children demonstrated improved processing speed after the intervention (d=0.54; 95% CI 0.012-0.083; P=.005), and a marginal improvement was also observed in parents (d=0.19; 95% CI −0.004 to 0.030; P=.05). Cortical source imaging of electroencephalographic recordings was acquired simultaneous to an attention-to-breathing assessment that showed significant reduction in task-related default mode network activity (d=−0.62; 95% CI −0.0096 to −0.0002; P=.01).

Conclusions: Post-CoCo intervention decrease in default mode network activity on the attention-to-breath task in parent-child dyads may be indicative of cortical plasticity reflecting reduced mind-wandering and thereby, enhanced focus after training. The current promising behavioral and cognitive results suggest the need for a larger sample size and a randomized controlled study design. Overall, these findings highlight the potential for brief, digital mindfulness and compassion cotraining to improve family mental health and well-being.

Trial Registration: International Standard Randomized Controlled Trial Number Registry ISRCTN89594822; https://www.isrctn.com/ISRCTN89594822

(JMIR Form Res 2025;9:e69607) doi:10.2196/69607

KEYWORDS

child; parent; depression; compassion; mindfulness; default mode network; EEG; electroencephalograph

XSL•FO
RenderX

Introduction

The Centers for Disease Control and Prevention estimates a 5% prevalence of depression in US children [1]. Depression in children is comorbid with anxiety and attention deficit [2-5] and may be associated with poor academic performance and school dropout, foretelling poor overall quality of life for these children [6,7]. In serious cases, children with depression can be susceptible to the grave risk of suicide [8]. Experts note that this scenario has been further exacerbated by the COVID-19 pandemic [9-11]. Additionally, depression in children is often associated with parental stress and concomitant parental depression [12-14]. Overall, this scenario highlights the urgent need for the development and implementation of accessible and scalable solutions that cobenefit child and parent mental well-being.

Parents play significant roles as either supportive influences, protective buffers, or potential sources of risk depending on the nature of their caregiving relationship with their children [15,16]. Several studies have underscored the detrimental mental health effects on children of poor parental mental health and psychological distress [17-19]. Additionally, developmental studies on parent-child dyads have consistently shown that parental involvement not only affects child behaviors but also influences emotion regulation as well as cognitive and neural responses in children [20-24]. With regard to interventions, a recent meta-analytic study [25] of parent-child mental health interventions recommended that interventions should recognize the family as an integrated unit by taking into account the collective needs of parents and children. By offering external guidance and emotional scaffolding, caregivers help shape children's ability to regulate their emotions—a process referred to as coregulation [26,27]. Thus, we hypothesize that parent-child cointervention programs that promote joint participation may leverage coregulation to foster parent-child co–well-being. Of note, research also shows that in this context, families prefer nonpharmacological treatments for their children whenever effective behavioral solutions are available [28].

To address this intervention gap, we developed and implemented a digital parent-child cotraining program that focuses on imparting mindfulness and compassion skills to both parent and child. Our focus on designing a digital program was driven by the need for greater accessibility and scalability of such interventions. Studies have shown that mindfulness and compassion training in children and parents can benefit interpersonal behaviors and even improve stress physiology [29-32]. For instance, a study of mindful parenting training found significantly reduced posttraining psychopathology in parents as well as in their children [33]. Another study demonstrated that Cognitively-Based Compassion Training in parents significantly reduced stress cortisol levels in their infants and young children [31]. Furthermore, many school-based mindfulness interventions have been introduced to enhance well-being in children, and such programs have grown rapidly since the pandemic [30,34,35]. Yet, all of these studies have examined the efficacy of mindfulness or compassion-based intervention separately applied to parents or children but not as cotraining. A recent randomized controlled trial compared parallel (but not simultaneous) parent and child mindfulness training to pharmacological intervention in children with attention-deficit/hyperactivity disorder (ADHD) and concluded that mindfulness training can be considered a useful nonpharmacological alternative or add-on to pharmacological treatment [36]. Other studies have also shown the utility of parent and child parallel mindfulness trainings in ADHD [37,38]. In addition, in a previous study, we showed that digitally delivered meditative trainings can benefit attentive function in adolescents who report childhood neglect and show signs of ADHD [39]. Overall, this literature suggests that a parent-child mindfulness and compassion cotraining application is an innovative approach.

In the interventional program that we designed and implemented, parent-child dyads jointly participated in working through 10-minute attention-to-breath mindfulness exercises in each session alongside prompts for compassion cultivation. Each session was kept brief to facilitate adherence in busy families, and up to 30 training sessions were available for families to practice over 3 months. We referred to the program as Cooperative Compassion (CoCo) training, and both parent and child underwent the practices together. To evaluate the outcomes of this intervention, we used self-report mental health measures at baseline, after training, and 3-month follow-up and also measured objective cognition in the presence of emotions at post- versus pretraining [40-43]. Additionally, we acquired electroencephalographic (EEG) recordings on an attention-to-breath assessment at pre- and posttraining as a neural outcome measure in all participants [44,45].

Since prior intervention research in this field has mostly relied on self-report measures and there have been limited comprehensive attempts at examining the effects of longitudinal parent-child cointervention programs [25,46-48], in this pilot study, we used multimodal assessments, that is, a combination of self-report mental health measures, an objective emotion-based cognitive processing measure, and a neural EEG-based measure in both parents and children. We hypothesized that after CoCo training, we would observe an improvement in mental health behaviors as the primary outcome and changes in cognitive processing and underlying neuroplasticity in both children and parents as secondary outcomes. Additionally, in the context of emotion regulation, prior interventions have focused on changing and evaluating parental behaviors but with limited assessment of children's emotion processing skills [22]. Therefore, in this study, we explored the dynamics of emotion-related cognitive processing changes in both children and parents. Finally, we used an EEG-based neural recording on an attention-to-breath assessment, using variability in breath-monitoring responses as a direct objective marker of variability in performance on an interoceptive attention task [45]. We adopted this task, as we have found that neurophysiology on this task correlates with trait mindfulness in a large cross-sectional study in individuals across the lifespan and have further demonstrated that this neurophysiology undergoes plasticity in the context of a mindfulness and compassion intervention in adults [44]. Specifically, we have observed intervention-driven reduction in default mode network (DMN) activity on the

XSL•FO
RenderX

attention-to-breathing task demonstrating posttraining plasticity of the DMN. Using neuroimaging, several other studies have shown that DMN activity is modulated by contemplative interventions [49-52]. The DMN is a functional network that has been consistently associated with autobiographical memory, self-referencing, but also with on-task behavioral variability, mind-wandering, and rumination [53-58]. Given its role in mind-wandering, we specifically hypothesized that the CoCo training would drive a reduction in DMN activity at postintervention relative to baseline. In summary, we hypothesized improvement in mental health behaviors as the primary outcome and improvement in emotion-related cognitive processing and reduction in EEG-derived DMN activity, reflective of reduction in mind-wandering and enhancement of present moment focus, as secondary outcomes.

Methods

Participants

A total of 24 parent-child dyads participated in this study. Dyads were recruited from the local community, that is, schools and university-affiliated pediatric clinics in the San Diego area through flyer advertisements and clinic referrals, respectively. The recruitment of parent-child dyads began on February 1, 2021, and ended on August 13, 2023. The study was registered on October 9, 2024, in the International Standard Randomized Controlled Trial Number Registry (ISRCTN89594822).

Ethical Considerations

Signed informed consent was obtained from the parents, and signed assent was obtained from the children for study participation following the guidelines outlined in the Declaration of Helsinki. The study protocol was approved by the institutional review board of the University of California San Diego (protocol #180140). The study data were deidentified as per Health Insurance Portability and Accountability Act (HIPAA) compliance. All the parent-child dyads were paid US $250 for completing all assessments and digital trainings.

Demographics

Parent and child demographics including age, sex, race, ethnicity, and socioeconomic status (SES) scores are shown in Table 1. SES was measured on the Family Affluence Scale [40,59-64]; this scale measures individual wealth based on ownership of objects of value (eg, car or computer) and number of vacations in the past year. The sum of items produces a composite score ranging from 0 - 2=low affluence, 3 - 5=middle affluence, and >5=high affluence. The average family SES score in our study sample was of high affluence. All children and parents were right-handed and had normal or corrected-to-normal vision.

XSL•FO
RenderX

Table . Summary of demographics and baseline mental health for parent-child dyad study participants[a].

Demographics and baseline mental health	Parents (n=24)	Children (n=24)
Age (years)		
Mean (SD)	44.5 (6.5)	9.5 (3.27)
Range	28-54	5 - 12
Sex, n (%)		
Male	4 (16.7)	10 (41.7)
Female	20 (83.3)	14 (58.3)
Race, n (%)		
Asian	8 (33.3)	5 (20.8)
Black or African American	0 (0)	0 (0)
Native American	0 (0)	0 (0)
Native Hawaiian or Other Pacific	0 (0)	0 (0)
White	14 (58.3)	11 (45.8)
More than 1 ethnicity	2 (8.3)	7 (29.2)
Other	0 (0)	1 (4.2)
Ethnicity, n (%)		
Hispanic or Latino	5 (20.8)	5 (20.8)
Not Hispanic or Latino	18 (75)	19 (79.2)
Unknown	1 (4.2)	0 (0)
Socioeconomic status		
Mean (SD)	6.5 (1.4)	—[b]
Range	4-8	—
Child Depression Index *T* scores		
Mean (SD)	54.08 (9.06)	57.96 (13.75)
Range	40-90	37-74
Parental stress (DASS-21)		
Mean (SD)	4.91 (3.61)	—
Range	0 - 13	—
Parental anxiety (GAD-7)		
Mean (SD)	4.39 (3.28)	—
Range	0-14	—
Parental depression (PHQ-9)		
Mean (SD)	5.13 (3.52)	—
Range	0 - 13	—

[a]Parental stress was measured using the 7 stress items on the 21-item Depression Anxiety Stress Scale (DASS-21), anxiety was measured on the 7-item General Anxiety Disorder (GAD-7) scale, and depression was measured on the 9-item Patient Health Questionnaire scale (PHQ-9).

[b]Not available.

Child and parent baseline mental health scores are also shown in Table 1. In our community-recruited sample, children were assessed on the Child Depression Index (CDI) [40,59-65] for study inclusion, that is, must have average or above-average CDI scores (*T* score>40). Parents in the study did not report any diagnosed illness. Exclusion was based on any self-reported severe illness for parent or child that would not allow time for study participation.

Sample Size and Power

This single-arm study was powered to detect a medium effect size (Cohen d >0.5) comparing pre- versus postintervention differences at β power of .8 and α level of .05 for each

XSL•FO
RenderX

assessment measure (behavioral, cognitive, and neural). Effect sizes were calculated a priori using the G*Power software (Heinrich Heine Universität) [66].

Feasibility

We assessed intervention feasibility by monitoring the total number of assigned intervention sessions completed. Additionally, parents completed a feasibility survey at the end of the study that we have standardized in previous digital training studies [67]. The survey queried 16 questions about the training as elaborated in the Results section (Table 2), and each question required a response on a 7-point Likert scale. In total, 11 of 16 questions solicited positive feedback regarding the intervention, while 5 questions solicited negative feedback and are shown in italics format in Table 2; for all questions, anchors were 1=strongly disagree to 7=strongly agree, and negative feedback questions were reverse coded and then averaged with positive questions to obtain overall training feasibility. The Cronbach α measure of reliability for the training survey was high (α=0.92).

Table . Results of the training feasibility survey completed by all participating parents (N=24) at the end of the study.

Training feasibility survey	Response on 1 - 7 scale, mean (SD)
We enjoyed the training.	4.96 (1.76)
We felt frustrated after the training.[a]	3.33 (2.18)
We felt satisfied with the training.	4.88 (1.85)
We felt tired after the training.	3.17 (1.74)
The training was easy to understand.	5.67 (1.71)
The training was difficult to use.	2.63 (1.69)
The training was easy to navigate.	5.75 (1.07)
We were worried about our data security.	1.58 (0.88)
The training was easy to initiate each day.	5.29 (1.71)
The training fit in our daily schedule.	4.79 (1.86)
The training time passed by quickly.	4.13 (1.83)
The training felt therapeutic.	4.21 (1.93)
The training felt useless.	3.21 (1.93)
We would recommend this training outside of this study.	3.96 (2.26)
We would recommend this training to others.	4.08 (2.28)
This training positively affected our family's life.	4.25 (1.85)
Average positive feedback	4.72 (1.83)
Average negative feedback	2.78 (1.68)
Average program feedback	4.97 (1.76)

[a]Questions soliciting negative feedback are marked in italics format.

Assessments

Overview

Each parent-child dyad made 2 visits (baseline [preintervention] and 3 months later [postintervention]) to the Neural Engineering and Translational Labs and participated in behavioral and neurocognitive assessments (Figure 1A); these assessments were conducted within 1 - 2 weeks of intervention initiation or completion, respectively. Behavioral assessments were also completed on the web at a 3-month follow-up after intervention completion.

XSL•FO
RenderX

Figure 1. Study design and breath-focused compassion training. (**A**) The study design incorporated behavioral, cognitive, and EEG assessments at pre- and posttraining conducted in parent-child dyads. Behavioral assessments were additionally repeated at a 3-month follow-up after intervention completion. (**B**) The digital *Cooperative Compassion* training delivered attention to breath-focused, performance-adaptive training in which the parent and child engaged simultaneously. {Level} refers to the number of breaths monitored by both parent and child ranging from 1 to 10. Higher levels were accessed when performance at lower levels was consistent for at least two-thirds of the level time. Performance was monitored, as users tapped the mobile screen after instructed {level} number of breaths while keeping their eyes closed; the screen was digitally split to simultaneously track both parent and child performance. At every sixth session of training, compassion instructions were relayed by text and audio before the start of the breath-focused exercises so that users could discuss and keep these instructions in mind during their practice. There were a total of 5 levels of standard compassion training instructions provided focusing on settling the mind, compassion for a loved one, compassion for self, loving kindness for self, and embracing common humanity. A distinct, calming nature scene unveiled at the end of each session. (**C**) Stimuli for the emotion bias cognitive assessment are shown and presented neutral, happy, sad, or angry faces superimposed on an arrow, whose direction was discriminated by participants. (**D**) In the attention-to-breath task, participants were instructed to close their eyes, breathe naturally, and respond after every 2 breaths by tapping on the spacebar, while simultaneously EEG was being recorded. EEG: electroencephalography.

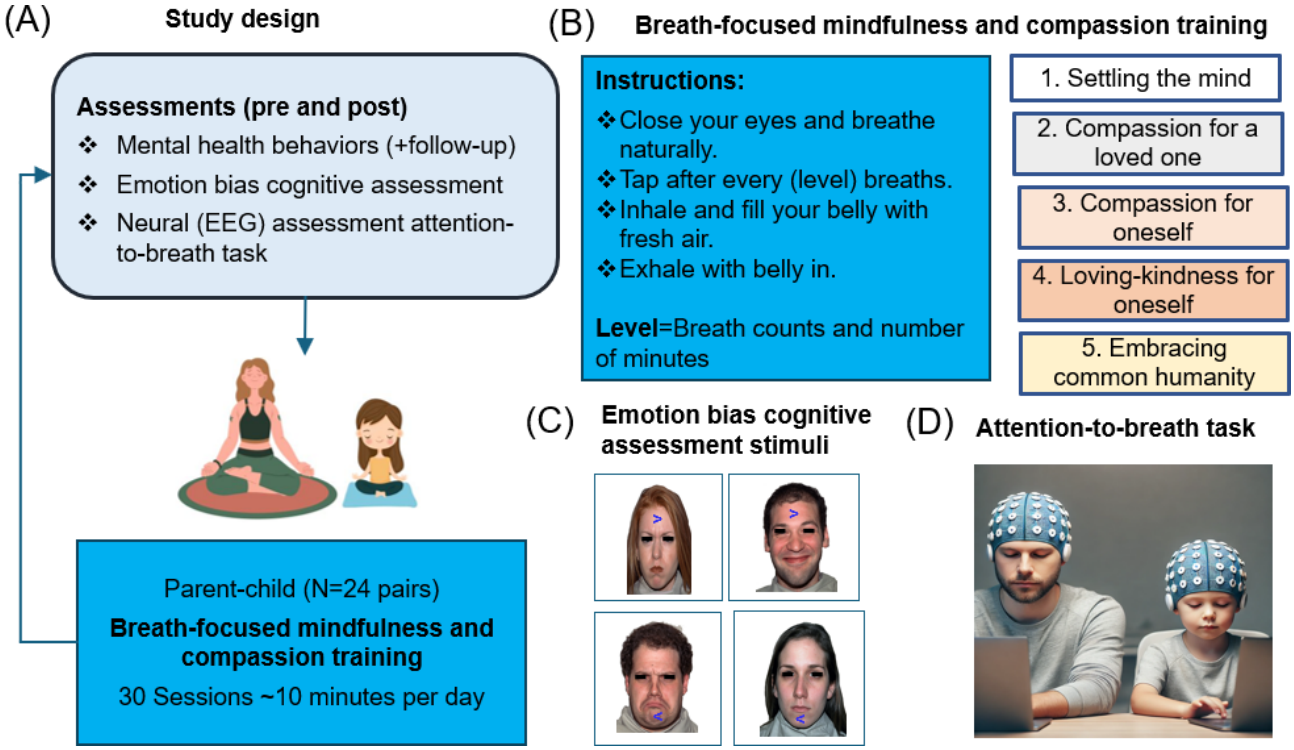

Behavioral Assessments

Each child self-reported on the CDI, and parents also provided CDI reports for their child. The CDI self-report and parent report are 12-item and 17-item scales, respectively, with higher scores reflecting greater child depression. Each CDI child report question requests the child to pick 1 of 3 sentences based on which one best described their emotion or behavior over the past 2 weeks rated 0=least to 2=most negative emotion or behavior. The CDI parent report enquires about child emotions or behaviors on a 4-point scale with anchors 0=not at all to 3=much or most of the time. Each parent additionally completed standard surveys assessing their own mental health including stress, anxiety, and depression, evaluated using the 21-item Depression Anxiety Stress Scale (DASS-21) with 4-point anchors from 0=did not apply to me at all to 3=applied to me very much or most of the time [68], 7-item General Anxiety Disorder (GAD-7) scale with 4-point anchors from 0=not at all to 3=nearly every day [69], and the 9-item Patient Health Questionnaire (PHQ-9) for depression scale with same anchors as for GAD-7; higher scores in each of these scales denoted greater symptom severity [70]. Parents also completed a 14-item mindfulness measure, Mindful Attention Awareness Scale (MAAS), that inquires about being unaware in the present moment in the context of different daily life activities with 6-point anchors from 1=almost always to 6=almost never, and higher scores reflecting higher trait mindfulness [71]. The Cronbach α measure of reliability calculated at baseline was high for each of these measures (CDI child report: α=0.83, CDI parent report: α=0.83, DASS-21: α=0.89, GAD-7: α=0.84, PHQ-9: α=0.74, MAAS: α=0.94).

Cognitive Assessment

Participants accessed a game-like emotion bias assessment adapted from studies of attentional bias in emotional contexts [40-42,71]. The task was delivered on the BrainE Unity-coded platform developed and deployed by Neural Engineering and Translational Labs [60,72] with stimuli displayed at a comfortable viewing distance. The task integrated a standardized set of culturally diverse faces from the NimStim database [73], examples shown in Figure 1C. We used an equivalent number of male and female faces, each face with 4 sets of emotions: neutral, positive (happy), negative (sad), or threatening (angry), presented on an equivalent number of trials in each task block. Postfixation cue on each trial, participants viewed an emotional face with a superimposed arrow of 300-millisecond duration. The arrow occurred in either the upper or lower central visual field on an equal number of trials. Participants responded to the

XSL•FO
RenderX

direction of the arrow (left or right) within an ensuing 1-second response window. For neutral emotion trials, this response window adapted in a 3-up-1-down scheme (−33 ms after correct trials and +100 ms after incorrect trials) that maintains accuracy at ~80% and engages the user by avoiding ceiling performance [67,74]. An adaptive scheme also reduces practice effects that affect repeat assessment sessions. All other emotion trials followed the same response window as their previous neutral emotion trial. Participants completed 144 trials presented over 3 equipartitioned blocks. Processing speed across trials was monitored as the main outcome of this assessment.

Neural (EEG) Assessment

We assessed EEG neurophysiology on an attention-to-breathing task assessment that we have recently shown to relate to mindfulness in a large cross-sectional study [45] and further show modulation in the context of a mindfulness and compassion intervention in adults [44]. As described in these studies, participants were instructed to close their eyes, breathe naturally, and respond every 2 breaths by tapping on the spacebar. The lab streaming layer protocol was used to time-stamp all user response events. The task was of 5-minute duration, implemented in two 2.5-minute blocks.

The median response time (RT) on the breath monitoring task was monitored for all participants, so that we could identify and contrast neurophysiological activity on high consistency breath monitoring trials (trials with RT ≤1 median absolute deviation of median RT) versus low consistency trials (trials with RT >1 median absolute deviation of median RT). The underlying premise of this analysis is that when participants are in a state of higher interoceptive attention, they will show more consistent awareness of their respiratory cycle (ie, responding more consistently after every 2 breath cycle), while states of lower interoceptive attention may be associated with inconsistent breath monitoring. Thus, by comparing brain activity on more versus less consistent trials within each participant, we can examine brain activity underlying performance variability on this interoceptive attention task while controlling for interindividual differences in baseline characteristics of respiration.

EEG data were collected simultaneous to this assessment using a 24-channel cap with saline-soaked electrodes following the 10‑20 system and a wireless SMARTING amplifier. The signals were digitized with a sampling rate of 500 Hz and 24-bit resolution and stored as .xdf files.

Intervention

Akin to a similar intervention that we deployed in adults [44], the digital CoCo training for parent-child dyads was delivered on the HIPAA-compliant BrainE app with a deidentified and password-protected login provided to each family. Participants accessed the iOS- or Android-compatible smartphone app in their own free time and engaged in ~10 minutes of training per session for up to 30 sessions (Figure 1B). The training was delivered in a game-like format and was performance-adaptive. The single-user interface design was already vetted in our prior study in adults [44]. For this study, the dual user interface for parent-child dyads was informally tested for usability by the study team's family members.

Specifically to practice breath-focused mindfulness, individuals were requested to close their eyes, pay attention to their breathing, and tap the mobile screen after a specific number of breaths from 1 breath up to 10 breaths at a time. Consistency of performance was monitored, as users tapped the mobile screen after instructed number of breaths while keeping their eyes closed; the screen was digitally split to simultaneously track both parent and child performance, that is, one-half of the screen kept track of child finger taps, while the other half kept track of parent taps. If both parent and child consistently monitored breathing for at least two-thirds of the duration at any given level, where level refers to the number of breaths monitored, then they would together progress to monitoring the next level or count of breaths. At every sixth session of training, compassion instructions were relayed by text and audio before the start of the breath-focused exercises so that users could discuss and keep these instructions in mind during their practice. Over 30 sessions, there were a total of five levels of standard compassion training instructions provided focusing on (1) settling the mind, (2) compassion for a loved one, (3) compassion for self, (4) loving kindness for self, and (5) embracing common humanity; these instructions followed guidance from the Compassion Cultivation Training program [75]. These compassion training instructions appeared as text and audio on the app's daily introductory screen; the text details at each level have been provided in Multimedia Appendix 1. Finally, a distinct, calming nature scene unveiled at the end of each session as a form of training reward. Parents received app notifications once a day reminding them to complete their training.

Data Analyses

Behavioral outcomes were analyzed using paired tests comparing pre- and posttraining assessments as well as pre- and follow-up assessments. Outliers >3 median absolute deviation from the median were removed, and all metrics were inspected using the Anderson-Darling normality test in MATLAB prior to statistical analyses. All measures, that is, CDI reports from children and parents and parental stress (per DASS-21), anxiety (GAD-7), and depression (PHQ-9) measures, were not normally distributed; hence, they were analyzed for pre- to posttraining changes using signed rank tests. Parent mindfulness data were normally distributed and hence, analyzed using 1-tailed *t* tests. One-sided paired tests were used in all cases, as score improvement in a single direction was meaningful for all measures. While 2-sided tests are more conservative, we have ensured that only strong effects are reported by applying (fdr) corrections for multiple statistical comparisons.

Cognitive outcomes corresponded to the pre- to postintervention difference in processing speed on the emotion bias task. These data were not normally distributed, and pre- to posttraining differences in processing speed across participants were analyzed using the signed rank test, with group as a factor analyzed using the Wilcoxon rank sum test.

XSL•FO
RenderX

Neural data on the eyes-closed attention-to-breathing task were analyzed according to methodologies used in our recently published studies [44,45]. Adult parent data were available for 22 of 24 parents, and child data were available for 19 of 24 children; thus, overall source analysis was conducted on 41 samples with children and adults data combined with individual age-specific head models applied (Multimedia Appendix 1) [76-78]. EEG data from the remaining adults or children were too noisy for analysis and were excluded; no imputations were made for these missing data. EEG analysis included (1) EEG channel data processing and (2) cortical source localization of the EEG data to estimate source-level neural activity. Details of this analysis are provided in Multimedia Appendix 1. Briefly, cleaned EEG data were epoched to the response taps made by participants after every 2 breaths with trials separated by either high consistency (trials with RT ≤1 median absolute deviation of median RT in each participant) or low consistency (trials with RT >1 median absolute deviation of median RT in each participant). As α band (8 - 12 Hz) oscillations are dominant in the eyes-closed state, source localized activity was analyzed in the α band in 3 canonical brain networks, the frontoparietal network (FPN), the cingulo-opercular network (CON), and the DMN. Specifically, we quantified the average network source activity for the low versus high consistency trial differential in the 0 - to 4-second period prior to breath response, as this provided a neural correlate for low versus high consistency task performance. Neural data were not normally distributed as verified by the Anderson-Darling test; hence, the pre- to posttraining difference in α source activity across participants was analyzed using the signed rank test in each of the 3 networks (FPN, DMN, and CON), with group as a factor analyzed using the Wilcoxon rank sum test.

Results

Intervention Feasibility

On average, families completed a mean of 27.58 (SD 4.82) of the total 30 assigned sessions, with 19 of 24 families completing more than 90% of the assigned sessions.

All parents also completed a feasibility survey (Table 2) responding to feasibility questions on a 7-point Likert scale. The average positive feedback across questions such as “We enjoyed the training,” “We felt satisfied with the training,” “The training was easy to understand,” and “This training positively affected our family’s life” was 4.72 (SD 1.83). In contrast, the average negative feedback across questions such as “We felt frustrated after the training,” “We felt tired after the training,” “The training was difficult to use,” and “The training felt useless” was 2.78 (SD 1.68). Overall, training program feedback was positive, that is, a mean response score of 4.97 (SD 1.76) of 7.

Behavioral Outcomes

Pre- versus postintervention change in CDI scores for children by self-report showed a nonsignificant effect in the expected direction (signed rank test z=1.50; Cohen d=−0.19; 95% CI −8.89 to 1.74; P=.07). Change in parent-reported CDI scores was nonsignificant (z=1.05; d=−0.20; 95% CI −4.93 to 1.27; P=.15). Parental stress (DASS-21, z=2.03; d=−0.41; 95% CI −2.63 to −0.16; P=.02), anxiety (GAD-7, z=2.15; d=−0.47; 95% CI −2.67 to −0.20; P=.02), and depression (PHQ-9, z=1.96; d=−0.50; 95% CI −3.25 to −0.08; P=.03) were all significantly improved at the posttraining time point.

Notably, these patterns of behavioral change were mostly sustained at the 3-month follow-up after training completion (Figure 2). At follow-up, self-reported CDI scores were not significant relative to baseline (z=1.21; d=−0.24; 95% CI −11.43 to 3.54; P=.11); yet, parent-reported CDI scores showed a nonsignificant improvement in the expected direction (z=1.46; d=−0.25; 95% CI −5.57 to 0.90; P=.07). Parental self-reports at follow-up showed sustained stress reduction (z=1.72; d=−0.43; 95% CI −2.31 to 0.21; P=.05) as well as improvement in anxiety (z=1.87; d=−0.39; 95% CI −2.53 to 0.23; P=.03) and depression (PHQ-9, z=2.47; d=−0.81; 95% CI −4.46 to −0.68; P=.007) symptoms. Parents did not show a significant change in mindfulness (MAAS) at either postintervention or follow-up (P>.1); yet, notably, pre- versus posttraining reduction in parental stress correlated with improvement in mindfulness (ρ=−0.45; P=.03). We applied fdr corrections to account for multiple self-reported behavior comparisons in parents, and significant results did not change either for pre versus post or pre versus follow-up testing for parental stress, anxiety, and depression (P<.05 fdr-corrected).

XSL•FO
RenderX

Figure 2. Swarm box charts of behavioral outcomes are shown for children (left column) and parents (middle and right columns). Box plots show the median with the lower and upper quartiles as the bottom and top edges of the boxes, respectively. The whiskers denote the data range, and the scatter points show individual values. The correlation between post- versus prestress reduction and improved mindfulness in parents is shown at the bottom right. Significant effects in parents survived multiple comparison false discovery rate (fdr) corrections (fdr corrected *P*<.05) for stress, anxiety, and depression. CDI: Child Depression Index. * represents that the *P* value is significant.

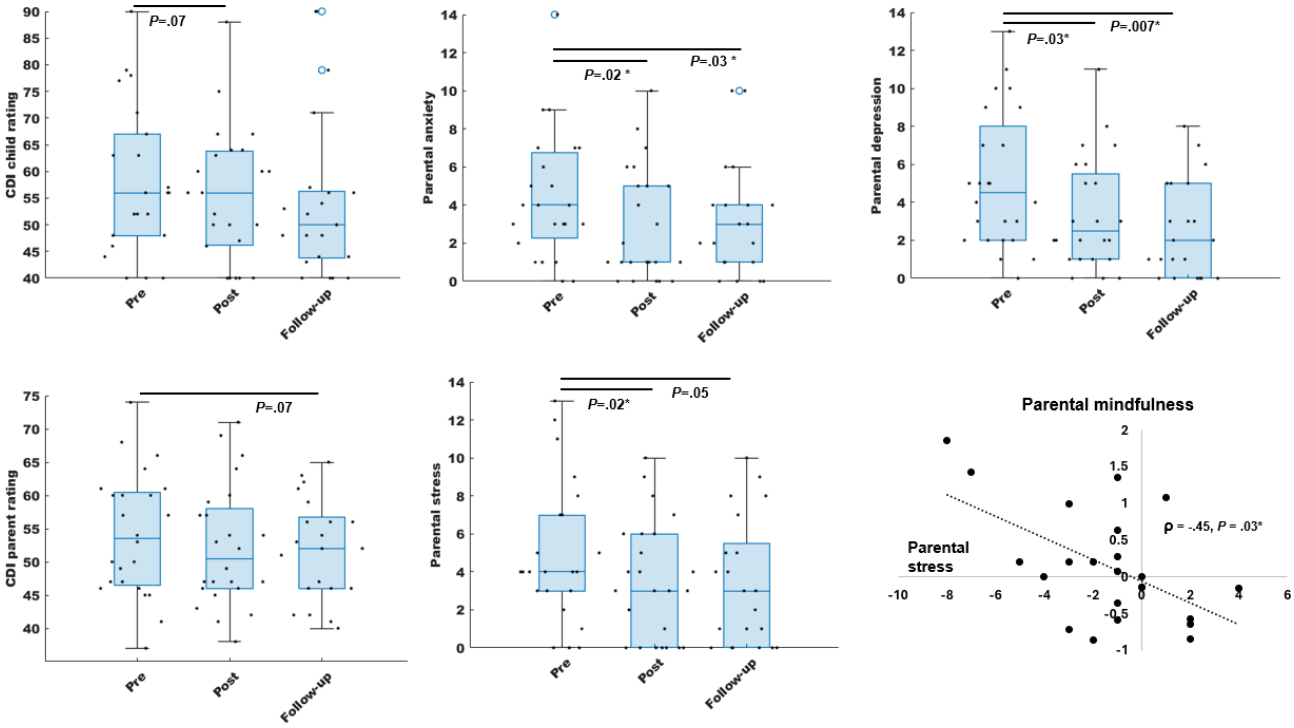

Cognitive Outcomes

Cognition was assessed as processing speed on the emotion bias task. Across all participants (n=48), there was a significant improvement in processing speed from pre- to postintervention (signed rank test *z*=3.42; *P*=.0006; *d*=0.40); yet, there was a significant group difference in this speed improvement (rank sum test *z*=2.01; *P*=.04). Post-hoc within-group pre- to postintervention signed rank tests revealed that only children showed a significant processing speed improvement across all emotion stimuli (*d*=0.54; 95% CI 0.012 to 0.083; *P*=.005), while adults showed a nonsignificant improvement in the expected direction (*d*=0.19; 95% CI −0.004 to 0.030; *P*=.05; Figure 3). In addition, children showed a consistent processing speed improvement for every emotion: neutral, happy, angry, or sad; that is, there was no bias toward a specific emotion (Table 3). The significant *P* values in children across emotion conditions survived fdr correction (*P*<.05).

Figure 3. Swarm box charts of emotion bias task outcomes are shown for (A) children and (B) parents. Box plots show the median with the lower and upper quartiles as the bottom and top edges of the boxes, respectively. The whiskers denote the data range, and the scatter points show individual values. Processing speed units are log(1/sec), where sec is the seconds unit of average trial response time on the task. Greater processing speed corresponds to speedier (ie, smaller) response times or speedier cognitive task processing. * represents that the *P* value is significant.

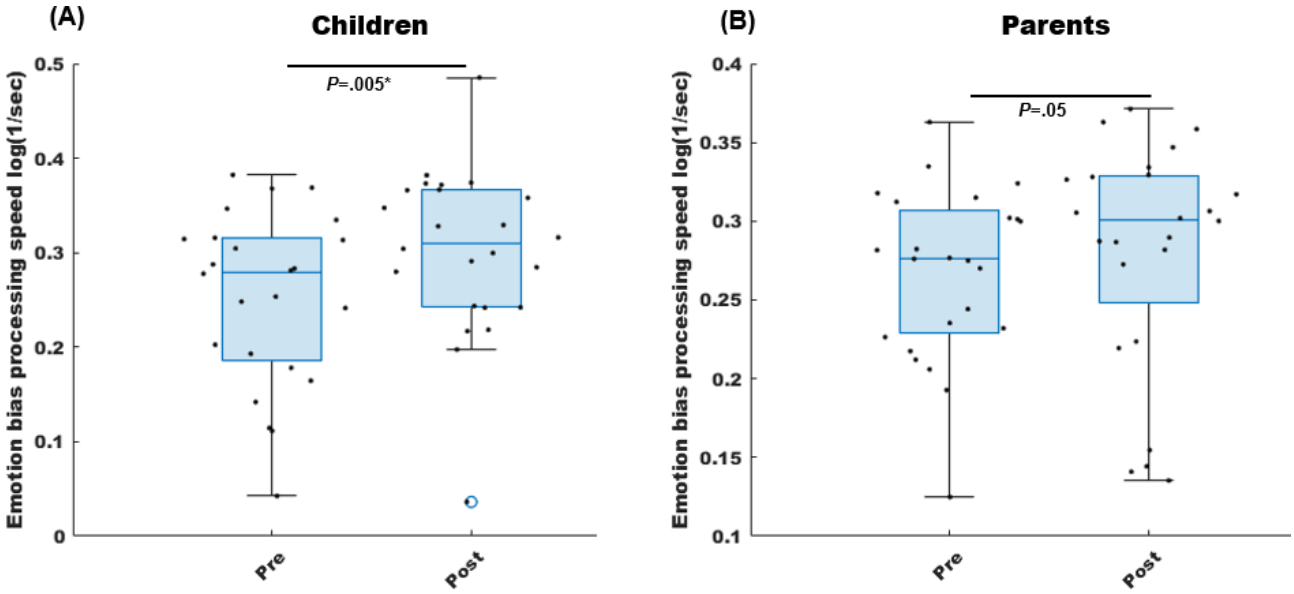

XSL•FO
RenderX

Table . Processing speeds on the emotion bias task in children and adults.

Processing speed	Children			Parents		
	Pre, mean (SD)[a]	Post, mean (SD)	*P* value	Pre, mean (SD)	Post, mean (SD)	*P* value
All faces	0.253 (0.09)	0.302 (0.09)	.005	0.268 (0.05)	0.280 (0.07)	.05
Neutral faces	0.255 (0.10)	0.307 (0.08)	.008	0.272 (0.06)	0.280 (0.07)	.21
Happy faces	0.249 (0.09)	0.306 (0.08)	.003	0.267 (0.05)	0.278 (0.07)	.09
Angry faces	0.258 (0.09)	0.301 (0.09)	.01	0.269 (0.05)	0.282 (0.08)	.14
Sad faces	0.250 (0.09)	0.296 (0.09)	.006	0.263 (0.06)	0.282 (0.07)	.05

[a]Mean and SD values are in log(1/sec), where sec is the seconds unit of average trial response time on the task. Post- versus precomparison *P* values are per signed rank tests.

Neural Outcomes

Neural changes were evaluated in the 8 - 12 Hz α frequency band of the EEG signal, as all participants showed peak processing in this band on the eyes-closed attention-to-breathing task. Across all participants, there was no significant change in FPN or CON source localized α activity ($P>.4$), but DMN activity was significantly reduced at postintervention relative to preintervention (signed rank test $z=-2.48$; $d=-0.62$; 95% CI −0.0096 to −0.0002; $P=.01$). In addition, there was no significant group difference in post- versus pre-DMN activity for children versus parents (rank sum test, $P=.09$; Figure 4). Group-specific post- versus prechanges in DMN activity showed significant reduction in DMN activity in children (signed rank test $z=-2.56$; $d=-1.09$; 95% CI −0.0015 to −0.0003; $P=.01$), but no change in parent ($P>.5$), suggesting that this neural outcome was exclusively driven by post- versus prechange in children.

Figure 4. Cooperative Compassion training–related neurophysiological changes evaluated on the attention-to-breath monitoring assessment. (**A**) Schematic of attention-to-breath task instructions. (**B**) Power frequency plot of scalp channel data across all participants and sessions showed peak processing in the α frequency band (8 - 12 Hz); dashed lines are 95% CIs. (**C**) Source-reconstructed electroencephalographic data were analyzed for 3 cognitive control networks: frontoparietal network (FPN), the cingulo-opercular network (CON), and the default mode network (DMN); regions of interest (ROIs) averaged within each network are highlighted. The identity of the ROIs within the 3 cognitive control networks is detailed in Multimedia Appendix 1. (**D**) Comparisons of the pre- to postintervention network changes across all participants are shown as swarm box plots. Box plots show median values with lower and upper quartiles as the bottom and top edges of the boxes, respectively. The whiskers denote the data range, and the scatter points show individual α source activity values.

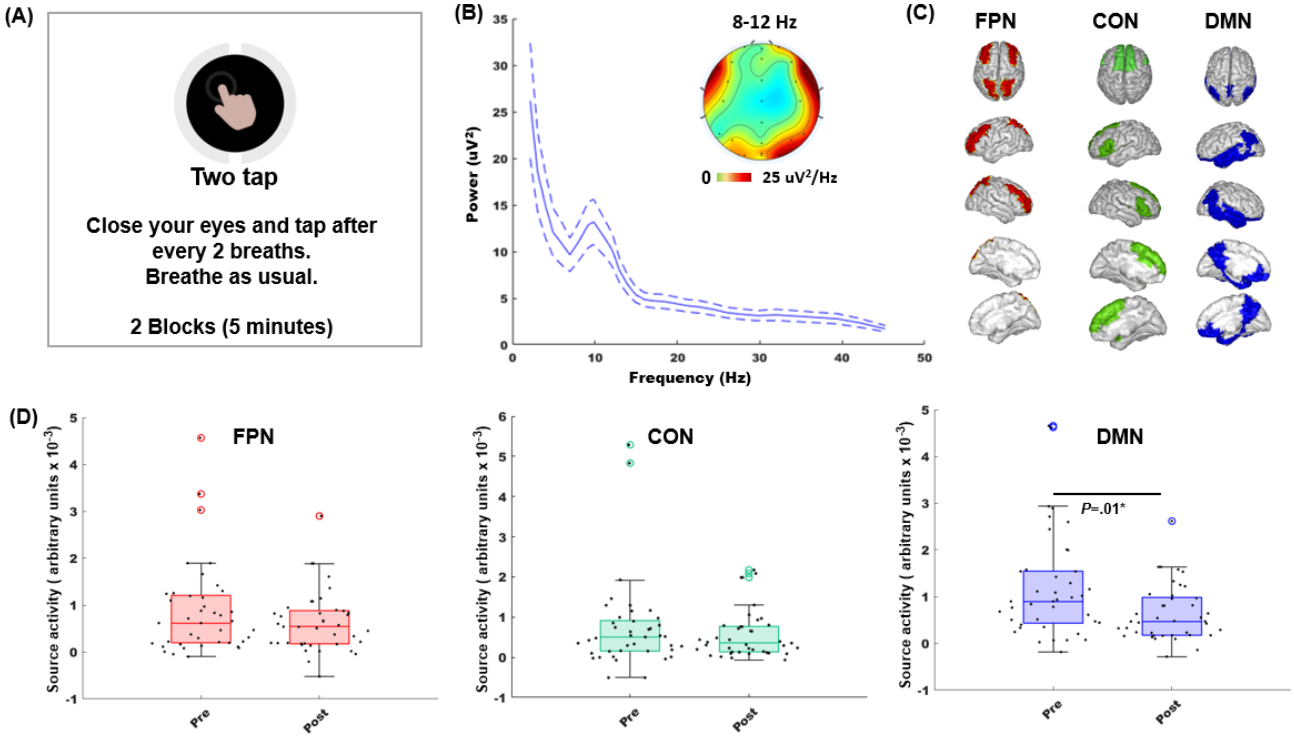

The summary of post- versus prechanges in behavioral, cognitive, and neural outcome measures in parents and children is provided in Table 4.

XSL•FO
RenderX

Table . Summary of post- versus prechanges in behavioral, cognitive, and neural outcome measures in parents and children[a].

Post- versus preintervention outcomes	Cohen *d* effect size	Mean difference (95% CI)	fdr[b] corrected *P* value
Child depression (self-reported)	−0.19	−3.57 (−8.89 to 1.74)	ns[c]
Child depression (parent-reported)	−0.20	−1.83 (−4.93 to 1.27)	ns
Parental stress	−0.41	−1.39 (−2.63 to −0.16)	.03
Parental anxiety	−0.47	−1.44 (−2.6 to −0.20)	.03
Parental depression	−0.50	−1.67 (−3.25 to −0.08)	.03
Children emotion bias cognitive processing (all faces)	0.54	0.047 (0.012 to 0.083)	.01
Parent emotion bias cognitive processing (all faces)	0.19	0.013 (−0.004 to 0.030)	ns
DMN[d] neural activity on attention-to-breath task (children and parents)	−0.62	-0.5631×10^{-3} (−0.0096 to −0.0002)	.04

[a]Children did not show any significant improvement in depressive symptoms, while parents showed significant improvement in stress, anxiety, and depression symptoms. Cognitively only children showed better emotion bias processing. Reduction in DMN neural activity was observed across all participants; there was no group interaction on this measure, but post-hoc tests showed this result to be driven by DMN neuroplasticity in children (see text).

[b]fdr: false discovery rate.

[c]ns: nonsignificant.

[d]DMN: default mode network.

Discussion

The primary objective of our study was to develop a digital cotraining program that would facilitate parent and child copractice of breath-focused mindfulness and compassion. We found the CoCo digital trainings to be feasible with 80% (n=19) of families completing >90% (n=22) of sessions, and all families reporting positive feedback on an exit survey. We evaluated the pilot efficacy of this intervention program using behavioral surveys assessing depression in children by both self and parental reports as well as parental self-reported stress, anxiety, and depression symptoms. Overall, children showed a statistical trend toward, that is, nonsignificant improvement in the expected direction in self-reported depression at posttraining, while the parent-reported child depression symptoms showed a similar effect at 3-month follow-up. The parents showed significant reductions in their stress, anxiety, and depression symptoms at postintervention with sustained effects at the 3-month follow-up. Furthermore, the posttraining reduction in parental stress was significantly correlated with improvement in dispositional mindfulness. We further measured pre- versus posttraining cognition in the presence of emotional faces and found that emotion bias processing speed was significantly improved in children, and a nonsignificant effect in the expected direction was observed in parents. Finally, on a neural assessment of breath-focused attention, we found a significant reduction in DMN activity at postintervention relative to preintervention that was especially robust in children.

Our observation of training-related reduction in parental stress, anxiety, and depression symptoms resonates with previous meta-analyses of behavioral interventions showing positive mental health benefits in parents [25,48]. Notably, we found significant results with a brief digital intervention of only ~10 minutes per session practiced over 30 sessions in 3 months, which suggests that even brief interventions that families can practically engage with in their own time can be beneficial for parents. Statistically, we were powered for medium effect size outcomes (d>0.5) that are also clinically meaningful. These medium to large effect size outcomes were observed for parental depression (d=0.5 at post and d=0.8 at follow-up) but not for parental anxiety or stress. This suggests that the CoCo training may most robustly ameliorate parental depression. In terms of neurocognitive outcomes, the emotion bias processing speed improvements in children were robust with medium effect size results (d=0.54). This neurocognitive assessment is in line with recommendations that studies of parent-child interventions must include valid assessments of children's emotion regulation abilities [22].

The neuroplasticity signal, specifically, reduction in DMN activity at postintervention relative to preintervention replicates our recent findings of digital mindfulness and compassion training–related neural changes observed in adults [44]. Specifically, this DMN activity modulation was observed on the attention-to-breath task, as the differential neural activity on high versus low consistency performance trials compared post- versus preintervention. As stated in the Methods section, the high versus low consistency trial contrast in neural activity controls for interindividual differences in baseline characteristics of respiration and reduction in DMN activity at postintervention relative to baseline may reflect reduced mind-wandering. This observation of intervention-related reduction in DMN activity also resonates with broader contemplative empirical research showing lesser DMN activity with greater mindfulness or meditation experience [49,50,52]. The nonsignificant effects within FPN and CON during the attention-to-breathing task may suggest the task-specific recruitment of DMN, as observed

XSL•FO
RenderX

in our recent studies [44,45]. In this recent work, in a large sample of participants (n=324) across the lifespan, we showed that lesser DMN activity on the breath attention task relates to greater trait mindfulness [45]. Thus, the significant postintervention decrease in DMN activity, particularly in children, in our study may represent neuroplasticity reflective of reduced mind-wandering and greater present moment awareness and focus after the training [55-58]. It is also notable that we obtained these results using EEG, a more scalable approach for measuring neural markers compared to other more costly neuroimaging modalities like functional magnetic resonance imaging.

Major limitations of this pilot intervention study include a small sample size and no control intervention comparator. Given the small sample size, we can only be confident of the medium to large effect size results in this pilot study obtained for parental depression, cognitive processing speed on the emotion bias task in children, and DMN neuroplasticity analyzed across all participants. Another limitation of the study was the lack of inclusion of Black and multiracial parent-child dyads as well as families from low affluence socioeconomics. We also acknowledge that parent-reported CDI measures may have potential bias; hence, we considered both parent and self-reports on this measure. Additionally, the CoCo training included both mindfulness and compassion instructions to maximize benefit, and disentangling their respective contribution to the outcomes is not possible in this study. Future studies should certainly include an active digital training control (eg, a relaxation app) with a larger sample size and ensure that the sample is more representative of racial and SES diversity. Further, since the training program was designed to target compassion, a future trial should include a compassion outcome measure as well as a measure of caregiver-child relational interaction.

Overall, this study examined a promising, accessible, and scalable digital program for cotraining parents and children on aspects of mindfulness and compassion. Our results showed significant medium effect size improvements particularly for parental depression and smaller effect size improvements for parental stress and anxiety, as well as a nonsignificant reduction in the expected direction in depressive symptoms in children. Notably, these outcomes in parents were sustained at follow-up. Additionally, children showed significant medium effect size improvement in emotion bias–related cognitive processing and large effect size reduction of DMN activity on the attention-to-breathing task as a neural indicator of less mind-wandering or greater focus at posttraining. Compared to prior interventions, this study is novel in its use of dyadic parent-child cotraining design as well as multimodal outcome measures that combine EEG, cognitive, and behavioral data. Prior studies conducted separately in children and parents have particularly found improvements in mindfulness and stress-related behavioral measures of medium effect size, and one study showed improvements in stress-related cortisol levels in parents [79-82]. Mindfulness training in parents has also been shown to reduce externalizing psychopathologic behaviors in children [83], and such trainings in parents and children have been shown to reduce child ADHD symptoms [84-86]. No other studies to date in this literature have measured emotion-related cognitive processing or intervention-related DMN neuroplasticity. Future studies should certainly include an active digital training control (eg, a relaxation app that may also reduce stress) with a larger sample size and ensure that the sample is more representative of racial and SES diversity. Training program implementation in schools, clinical populations such as ADHD children, and underserved or trauma-exposed populations is also relevant. In conclusion, our encouraging findings support future scale-up of the CoCo program to nurture family well-being.

Acknowledgments

This work was supported by grants from the Sanford Institute for Empathy and Compassion at the University of California San Diego (JM and SJ) and the NIH/NCCIH 5U24AT011289 Plasticity of Wellbeing Network subaward (JM and SJ). The authors thank Alankar Misra for the software development of the BrainE software. The BrainE software is copyrighted for commercial use (Regents of the University of California Copyright #SD2018-816) and is free for research and educational purposes. There was no use of generative artificial intelligence in the preparation of any portion of this manuscript.

Data Availability

The datasets generated or analyzed during this study are available in the Data Dryad repository [87]. The Cooperative Compassion digital intervention implemented on the BrainE app [72] is available for scientific research and is copyrighted for commercial use (Regents of the University of California Copyright SD2025-331).

Authors' Contributions

DR, DJK, and JM contributed to the conception, design, and implementation of the study. SJ conducted data management and data analytics and wrote the first draft of the manuscript. SRP and JKM collected and managed the data, and JN and JOY managed and analyzed the data. SD assisted in data management and manuscript writing. JM supervised all data collection, data management, analytics, and manuscript writing. All authors contributed to manuscript edits and revisions and approved the submitted version.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
Additional material.
[DOCX File, 60 KB - formative_v9i1e69607_app1.docx]

References

1. Data and statistics on children's mental health. Centers for Disease Control and Prevention. 2023. URL: https://www.cdc.gov/children-mental-health/data-research/?CDC_AAref_Val=https://www.cdc.gov/childrensmentalhealth/data.html [accessed 2023-05-13]
2. Biederman J, Newcorn J, Sprich S. Comorbidity of attention deficit hyperactivity disorder with conduct, depressive, anxiety, and other disorders. Am J Psychiatry 1991 May;148(5):564-577. [doi: 10.1176/ajp.148.5.564] [Medline: 2018156]
3. Michielsen M, Comijs HC, Semeijn EJ, Beekman ATF, Deeg DJH, Sandra Kooij JJ. The comorbidity of anxiety and depressive symptoms in older adults with attention-deficit/hyperactivity disorder: a longitudinal study. J Affect Disord 2013 Jun;148(2-3):220-227. [doi: 10.1016/j.jad.2012.11.063] [Medline: 23267726]
4. Schein J, Childress A, Gagnon-Sanschagrin P, et al. Treatment patterns among patients with attention-deficit/hyperactivity disorder and comorbid anxiety and/or depression in the United States: a retrospective claims analysis. Adv Ther 2023 May;40(5):2265-2281. [doi: 10.1007/s12325-023-02458-5] [Medline: 36913128]
5. Ware OD, Zerden LD, Duron JF, et al. Prevalence of co-occurring conditions among youths receiving treatment with primary anxiety, ADHD, or depressive disorder diagnoses. Front Child Adolesc Psychiatry 2024;3:1340480. [doi: 10.3389/frcha.2024.1340480] [Medline: 39816615]
6. DeBerard M, Spielmans GI, Julka DL. Predictors of academic achievement and retention among college freshmen: a longitudinal study. Coll Stud J 2004;38(1):66-80.
7. Esch P, Bocquet V, Pull C, et al. The downward spiral of mental disorders and educational attainment: a systematic review on early school leaving. BMC Psychiatry 2014 Aug 27;14(1):237. [doi: 10.1186/s12888-014-0237-4] [Medline: 25159271]
8. Janiri D, Doucet GE, Pompili M, et al. Risk and protective factors for childhood suicidality: a US population-based study. Lancet Psychiatry 2020 Apr;7(4):317-326. [doi: 10.1016/S2215-0366(20)30049-3] [Medline: 32171431]
9. Deng J, Zhou F, Hou W, et al. Prevalence of mental health symptoms in children and adolescents during the COVID-19 pandemic: a meta-analysis. Ann N Y Acad Sci 2023 Feb;1520(1):53-73. [doi: 10.1111/nyas.14947] [Medline: 36537131]
10. Kauhanen L, Wan Mohd Yunus WMA, Lempinen L, et al. A systematic review of the mental health changes of children and young people before and during the COVID-19 pandemic. Eur Child Adolesc Psychiatry 2023 Jun;32(6):995-1013. [doi: 10.1007/s00787-022-02060-0] [Medline: 35962147]
11. Madigan S, Racine N, Vaillancourt T, et al. Changes in depression and anxiety among children and adolescents from before to during the COVID-19 pandemic: a systematic review and meta-analysis. JAMA Pediatr 2023 Jun 1;177(6):567-581. [doi: 10.1001/jamapediatrics.2023.0846] [Medline: 37126337]
12. Fang Y, Luo J, Boele M, Windhorst D, van Grieken A, Raat H. Parent, child, and situational factors associated with parenting stress: a systematic review. Eur Child Adolesc Psychiatry 2024 Jun;33(6):1687-1705. [doi: 10.1007/s00787-022-02027-1] [Medline: 35876894]
13. Farmer AY, Lee SK. The effects of parenting stress, perceived mastery, and maternal depression on parent–child interaction. J Soc Serv Res 2011 Oct;37(5):516-525. [doi: 10.1080/01488376.2011.607367]
14. Tan S, Rey J. Depression in the young, parental depression and parenting stress. Australas Psychiatry 2005 Mar;13(1):76-79. [doi: 10.1080/j.1440-1665.2004.02155.x] [Medline: 15777418]
15. Feldman R. What is resilience: an affiliative neuroscience approach. World Psychiatry 2020 Jun;19(2):132-150. [doi: 10.1002/wps.20729] [Medline: 32394561]
16. Masten AS, Barnes AJ. Resilience in children: developmental perspectives. Children (Basel) 2018 Jul 17;5(7):98. [doi: 10.3390/children5070098] [Medline: 30018217]
17. Rosa EM, Putri R, Ismail D, Haksari EL. The influence of parenting style and mother's mental health from mental health in children who have siblings with autism. Bali Med J 2021;10(3):1181-1185. [doi: 10.15562/bmj.v10i3.2879]
18. Sander JB, McCarty CA. Youth depression in the family context: familial risk factors and models of treatment. Clin Child Fam Psychol Rev 2005 Sep;8(3):203-219. [doi: 10.1007/s10567-005-6666-3] [Medline: 16151618]
19. Tran AGTT. Family contexts: parental experiences of discrimination and child mental health. Am J Community Psychol 2014 Mar;53(1-2):37-46. [doi: 10.1007/s10464-013-9607-1] [Medline: 24146093]
20. Alonso A, McDorman SA, Romeo RR. How parent-child brain-to-brain synchrony can inform the study of child development. Child Dev Perspect 2024 Mar;18(1):26-35. [doi: 10.1111/cdep.12494] [Medline: 39421441]
21. Deng X, Chen K, Chen X, et al. Parental involvement affects parent-adolescents brain-to-brain synchrony when experiencing different emotions together: an EEG-based hyperscanning study. Behav Brain Res 2024 Feb 26;458:114734. [doi: 10.1016/j.bbr.2023.114734] [Medline: 37926335]
22. England-Mason G, Gonzalez A. Intervening to shape children's emotion regulation: a review of emotion socialization parenting programs for young children. Emotion 2020 Feb;20(1):98-104. [doi: 10.1037/emo0000638] [Medline: 31961186]

XSL•FO
RenderX

23. Feldman R. The relational basis of adolescent adjustment: trajectories of mother-child interactive behaviors from infancy to adolescence shape adolescents' adaptation. Attach Hum Dev 2010 Jan;12(1-2):173-192. [doi: 10.1080/14616730903282472] [Medline: 20390528]
24. Santamaria L, Noreika V, Georgieva S, Clackson K, Wass S, Leong V. Emotional valence modulates the topology of the parent-infant inter-brain network. Neuroimage 2020 Feb 15;207:116341. [doi: 10.1016/j.neuroimage.2019.116341] [Medline: 31712166]
25. Newham JJ, McLean K, Ginja S, et al. Brief evidence-based interventions for universal child health services: a restricted evidence assessment of the literature. BMC Public Health 2020 Jun 24;20(1):993. [doi: 10.1186/s12889-020-09104-7] [Medline: 32580720]
26. Bornstein MH, Esposito G. Coregulation: a multilevel approach via biology and behavior. Children (Basel) 2023 Jul 31;10(8):1323. [doi: 10.3390/children10081323] [Medline: 37628322]
27. Paley B, Hajal NJ. Conceptualizing emotion regulation and coregulation as family-level phenomena. Clin Child Fam Psychol Rev 2022 Mar;25(1):19-43. [doi: 10.1007/s10567-022-00378-4] [Medline: 35098427]
28. National Survey of Children's Health. Data Resource Center for Child & Adolescent Health. URL: https://www.childhealthdata.org/learn-about-the-nsch/NSCH [accessed 2024-07-31]
29. Curl EL, Hampton LH. Virtual mindfulness workshops for parents of children on the autism spectrum. J Early Interv 2024 Sep;46(3):323-337. [doi: 10.1177/10538151231169950]
30. Garro A, Janal M, Kondroski K, Stillo G, Vega V. Mindfulness initiatives for students, teachers, and parents: a review of literature and implications for practice during COVID-19 and beyond. Contemp Sch Psychol 2023;27(1):152-169. [doi: 10.1007/s40688-022-00446-0] [Medline: 36741424]
31. Poehlmann-Tynan J, Engbretson A, Vigna AB, et al. Cognitively-Based Compassion Training for parents reduces cortisol in infants and young children. Infant Ment Health J 2020 Jan;41(1):126-144. [doi: 10.1002/imhj.21831] [Medline: 31583748]
32. Roberts LR, Boostrom GG, Dehom SO, Neece CL. Self-reported parenting stress and cortisol awakening response following mindfulness-based stress reduction intervention for parents of children with developmental delays: a pilot study. Biol Res Nurs 2020 Apr;22(2):217-225. [doi: 10.1177/1099800419890125] [Medline: 31755301]
33. Meppelink R, de Bruin EI, Wanders-Mulder FH, Vennik CJ, Bögels SM. Mindful parenting training in child psychiatric settings: heightened parental mindfulness reduces parents' and children's psychopathology. Mindfulness (N Y) 2016;7(3):680-689. [doi: 10.1007/s12671-016-0504-1] [Medline: 27217845]
34. Styck KM, Malecki CK, Ogg J, Demaray MK. Measuring COVID-19-related stress among 4th through 12th grade students. School Psych Rev 2021 Oct 2;50(4):530-545. [doi: 10.1080/2372966X.2020.1857658]
35. van de Weijer-Bergsma E, Langenberg G, Brandsma R, Oort FJ, Bögels SM. The effectiveness of a school-based mindfulness training as a program to prevent stress in elementary school children. Mindfulness (N Y) 2014;5(3):238-248. [doi: 10.1007/s12671-012-0171-9]
36. Meppelink R, de Bruin EI, Zoller BK, Oort FJ, Bögels SM. Child and parent mindfulness-based training versus medication for childhood ADHD: a randomised clinical trial. Mindfulness (N Y) 2024 Feb;15(2):295-309. [doi: 10.1007/s12671-024-02305-w]
37. Singh NN, Singh AN, Lancioni GE, Singh J, Winton ASW, Adkins AD. Mindfulness training for parents and their children with ADHD increases the children's compliance. J Child Fam Stud 2010 Apr;19(2):157-166. [doi: 10.1007/s10826-009-9272-z]
38. van der Oord S, Bögels SM, Peijnenburg D. The effectiveness of mindfulness training for children with ADHD and mindful parenting for their parents. J Child Fam Stud 2012 Feb;21(1):139-147. [doi: 10.1007/s10826-011-9457-0] [Medline: 22347788]
39. Mishra J, Sagar R, Parveen S, et al. Closed-loop digital meditation for neurocognitive and behavioral development in adolescents with childhood neglect. Transl Psychiatry 2020 May 18;10(1):153. [doi: 10.1038/s41398-020-0820-z] [Medline: 32424253]
40. Grennan G, Balasubramani PP, Alim F, et al. Cognitive and neural correlates of loneliness and wisdom during emotional bias. Cereb Cortex 2021 Jun 10;31(7):3311-3322. [doi: 10.1093/cercor/bhab012] [Medline: 33687437]
41. López-Martín S, Albert J, Fernández-Jaén A, Carretié L. Emotional distraction in boys with ADHD: neural and behavioral correlates. Brain Cogn 2013 Oct;83(1):10-20. [doi: 10.1016/j.bandc.2013.06.004] [Medline: 23867737]
42. López-Martín S, Albert J, Fernández-Jaén A, Carretié L. Emotional response inhibition in children with attention-deficit/hyperactivity disorder: neural and behavioural data. Psychol Med 2015 Jul;45(10):2057-2071. [doi: 10.1017/S0033291714003195] [Medline: 25708692]
43. Thai N, Taber-Thomas BC, Pérez-Edgar KE. Neural correlates of attention biases, behavioral inhibition, and social anxiety in children: an ERP study. Dev Cogn Neurosci 2016 Jun;19:200-210. [doi: 10.1016/j.dcn.2016.03.008] [Medline: 27061248]
44. Jaiswal S, Purpura SR, Manchanda JK, et al. Design and implementation of a brief digital mindfulness and compassion training app for health care professionals: cluster randomized controlled trial. JMIR Ment Health 2024;11:e49467. [doi: 10.2196/49467]

XSL•FO
RenderX

45. Ramanathan D, Nan J, Grennan G, et al. Modulation of posterior default mode network activity during interoceptive attention and relation to mindfulness. Biol Psychiatry Glob Open Sci 2024 Nov;4(6):100384. [doi: 10.1016/j.bpsgos.2024.100384] [Medline: 39416659]
46. Ballal D, Navaneetham J, Chandra PS. Children of parents with mental illness: the need for family focussed interventions in India. Indian J Psychol Med 2019;41(3):228-234. [doi: 10.4103/IJPSYM.IJPSYM_430_18] [Medline: 31142923]
47. Christopoulos C, Hamoudi A, Murray DW, Rosanbalm KD. Self-regulation and toxic stress: foundations for understanding self-regulation from an applied developmental perspective. : Duke University; 2015 URL: https://dukespace.lib.duke.edu/items/a684cb28-7275-464e-a33a-d1ce28d5188e [accessed 2025-07-03]
48. Kulasinghe K, Whittingham K, Mitchell AE, Boyd RN. Psychological interventions targeting mental health and the mother-child relationship in autism: systematic review and meta-analysis. Dev Med Child Neurol 2023 Mar;65(3):329-345. [doi: 10.1111/dmcn.15432] [Medline: 36208472]
49. Brewer JA, Worhunsky PD, Gray JR, Tang YY, Weber J, Kober H. Meditation experience is associated with differences in default mode network activity and connectivity. Proc Natl Acad Sci USA 2011 Dec 13;108(50):20254-20259. [doi: 10.1073/pnas.1112029108]
50. Garrison KA, Zeffiro TA, Scheinost D, Constable RT, Brewer JA. Meditation leads to reduced default mode network activity beyond an active task. Cogn Affect Behav Neurosci 2015 Sep;15(3):712-720. [doi: 10.3758/s13415-015-0358-3] [Medline: 25904238]
51. Scheibner HJ, Bogler C, Gleich T, Haynes JD, Bermpohl F. Internal and external attention and the default mode network. Neuroimage 2017 Mar 1;148:381-389. [doi: 10.1016/j.neuroimage.2017.01.044] [Medline: 28110087]
52. Simon R, Engström M. The default mode network as a biomarker for monitoring the therapeutic effects of meditation. Front Psychol 2015;6:776. [doi: 10.3389/fpsyg.2015.00776] [Medline: 26106351]
53. Harrison BJ, Pujol J, López-Solà M, et al. Consistency and functional specialization in the default mode brain network. Proc Natl Acad Sci U S A 2008 Jul 15;105(28):9781-9786. [doi: 10.1073/pnas.0711791105] [Medline: 18621692]
54. Koban L, Gianaros PJ, Kober H, Wager TD. The self in context: brain systems linking mental and physical health. Nat Rev Neurosci 2021 May;22(5):309-322. [doi: 10.1038/s41583-021-00446-8] [Medline: 33790441]
55. Kucyi A, Esterman M, Riley CS, Valera EM. Spontaneous default network activity reflects behavioral variability independent of mind-wandering. Proc Natl Acad Sci U S A 2016 Nov 29;113(48):13899-13904. [doi: 10.1073/pnas.1611743113] [Medline: 27856733]
56. Malinowski P. Neural mechanisms of attentional control in mindfulness meditation. Front Neurosci 2013;7:8. [doi: 10.3389/fnins.2013.00008] [Medline: 23382709]
57. Poerio GL, Sormaz M, Wang HT, Margulies D, Jefferies E, Smallwood J. The role of the default mode network in component processes underlying the wandering mind. Soc Cogn Affect Neurosci 2017 Jul 1;12(7):1047-1062. [doi: 10.1093/scan/nsx041] [Medline: 28402561]
58. Zhou HX, Chen X, Shen YQ, et al. Rumination and the default mode network: meta-analysis of brain imaging studies and implications for depression. Neuroimage 2020 Feb 1;206:116287. [doi: 10.1016/j.neuroimage.2019.116287] [Medline: 31655111]
59. Boudreau B, Poulin C. An examination of the validity of the Family Affluence Scale II (FAS II) in a general adolescent population of Canada. Soc Indic Res 2009 Oct;94(1):29-42. [doi: 10.1007/s11205-008-9334-4]
60. Balasubramani PP, Ojeda A, Grennan G, et al. Mapping cognitive brain functions at scale. Neuroimage 2021 May 1;231:117641. [doi: 10.1016/j.neuroimage.2020.117641] [Medline: 33338609]
61. Balasubramani PP, Diaz-Delgado J, Grennan G, et al. Distinct neural activations correlate with maximization of reward magnitude versus frequency. Cereb Cortex 2023 May 9;33(10):6038-6050. [doi: 10.1093/cercor/bhac482] [Medline: 36573422]
62. Grennan G, Balasubramani PP, Vahidi N, Ramanathan D, Jeste DV, Mishra J. Dissociable neural mechanisms of cognition and well-being in youth versus healthy aging. Psychol Aging 2022 Nov;37(7):827-842. [doi: 10.1037/pag0000710] [Medline: 36107693]
63. Grennan GK, Withers MC, Ramanathan DS, Mishra J. Differences in interference processing and frontal brain function with climate trauma from California's deadliest wildfire. In: Jia F, editor. PLOS Clim 2023;2(1):e0000125. [doi: 10.1371/journal.pclm.0000125]
64. Mo Z, Grennan G, Kulkarni A, Ramanathan D, Balasubramani PP, Mishra J. Parietal alpha underlies slower cognitive responses during interference processing in adolescents. Behav Brain Res 2023 Apr 12;443:114356. [doi: 10.1016/j.bbr.2023.114356] [Medline: 36801472]
65. Kovacs M. Childen's Depression Inventory 2nd EditionTM: APA PsycTests; 2012.
66. Faul F, Erdfelder E, Buchner A, Lang AG. Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. Behav Res Methods 2009 Nov;41(4):1149-1160. [doi: 10.3758/BRM.41.4.1149] [Medline: 19897823]
67. Mishra J, Anguera JA, Gazzaley A. Video games for neuro-cognitive optimization. Neuron 2016 Apr 20;90(2):214-218. [doi: 10.1016/j.neuron.2016.04.010] [Medline: 27100194]

XSL•FO
RenderX

68. Henry JD, Crawford JR. The short-form version of the Depression Anxiety Stress Scales (DASS-21): construct validity and normative data in a large non-clinical sample. Br J Clin Psychol 2005 Jun;44(Pt 2):227-239. [doi: 10.1348/014466505X29657] [Medline: 16004657]
69. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
70. Kroenke K, Spitzer RL, Williams JBW, Löwe B. The Patient Health Questionnaire Somatic, Anxiety, and Depressive Symptom Scales: a systematic review. Gen Hosp Psychiatry 2010;32(4):345-359. [doi: 10.1016/j.genhosppsych.2010.03.006] [Medline: 20633738]
71. Brown KW, Ryan RM. The benefits of being present: mindfulness and its role in psychological well-being. J Pers Soc Psychol 2003 Apr;84(4):822-848. [doi: 10.1037/0022-3514.84.4.822] [Medline: 12703651]
72. Misra A, Ojeda A, Mishra J. BrainE: a digital platform for evaluatng, engaging and enhancing brain function. : Regents of the University of California Copyright SD2018-816; 2018.
73. Tottenham N, Tanaka JW, Leon AC, et al. The NimStim set of facial expressions: judgments from untrained research participants. Psychiatry Res 2009 Aug 15;168(3):242-249. [doi: 10.1016/j.psychres.2008.05.006] [Medline: 19564050]
74. García-Pérez MA. Forced-choice staircases with fixed step sizes: asymptotic and small-sample properties. Vision Res 1998 Jun;38(12):1861-1881. [doi: 10.1016/s0042-6989(97)00340-4] [Medline: 9797963]
75. Jinpa T. Compassion Cultivation Training© (CCTTM). : Compassion Institute; 2009. URL: https://www.compassioninstitute.com/cct/ [accessed 2023-04-20]
76. Neurodevelopmental MRI database. John E Richards Lab. URL: https://jerlab.sc.edu/projects/neurodevelopmental-mri-database/ [accessed 2025-09-25]
77. Richards JE, Sanchez C, Phillips-Meek M, Xie W. A database of age-appropriate average MRI templates. Neuroimage 2016 Jan 1;124(Pt B):1254-1259. [doi: 10.1016/j.neuroimage.2015.04.055] [Medline: 25941089]
78. Richards JE, Xie W. Brains for all the ages: structural neurodevelopment in infants and children from a life-span perspective. Adv Child Dev Behav 2015;48:1-52. [doi: 10.1016/bs.acdb.2014.11.001] [Medline: 25735940]
79. Chaplin TM, Turpyn CC, Fischer S, et al. Parenting-focused mindfulness intervention reduces stress and improves parenting in highly stressed mothers of adolescents. Mindfulness (N Y) 2021 Feb;12(2):450-462. [doi: 10.1007/s12671-018-1026-9]
80. Ho RTH, Lo HHM, Fong TCT, Choi CW. Effects of a mindfulness-based intervention on diurnal cortisol pattern in disadvantaged families: a randomized controlled trial. Psychoneuroendocrinology 2020 Jul;117:104696. [doi: 10.1016/j.psyneuen.2020.104696] [Medline: 32353816]
81. Shorey S, Ng ED. The efficacy of mindful parenting interventions: a systematic review and meta-analysis. Int J Nurs Stud 2021 Sep;121:103996. [doi: 10.1016/j.ijnurstu.2021.103996] [Medline: 34175531]
82. Burgdorf V, Szabó M, Abbott MJ. The effect of mindfulness interventions for parents on parenting stress and youth psychological outcomes: a systematic review and meta-analysis. Front Psychol 2019;10:1336. [doi: 10.3389/fpsyg.2019.01336] [Medline: 31244732]
83. van de Weijer-Bergsma E, Formsma AR, de Bruin EI, Bögels SM. The effectiveness of mindfulness training on behavioral problems and attentional functioning in adolescents with ADHD. J Child Fam Stud 2012 Oct;21(5):775-787. [doi: 10.1007/s10826-011-9531-7] [Medline: 22993482]
84. Valero M, Cebolla A, Colomer C. Mindfulness training for children with ADHD and their parents: a randomized control trial. J Atten Disord 2022 Mar;26(5):755-766. [doi: 10.1177/10870547211027636] [Medline: 34189992]
85. Sultan MA, Nawaz FA, Alattar B, et al. Assessing the impact of mindfulness programs on attention-deficit/ hyperactivity disorder in children and adolescents: a systematic review. BMC Pediatr 2025 Jan 14;25(1):32. [doi: 10.1186/s12887-024-05310-z] [Medline: 39810118]
86. Haydicky J, Wiener J, Badali P, Milligan K, Ducharme JM. Evaluation of a mindfulness-based intervention for adolescents with learning disabilities and co-occurring ADHD and anxiety. Mindfulness (N Y) 2012 Jun;3(2):151-164. [doi: 10.1007/s12671-012-0089-2]
87. Jaiswal S, Nan J, Dizon S, et al. Breath-focused mindfulness and compassion training in parent-child dyads: a pilot intervention study. Dryad 2025 May 12. [doi: 10.5061/DRYAD.9GHX3FFSS]

Abbreviations

ADHD: attention-deficit/hyperactivity disorder
CDI: Child Depression Index
CoCo: Cooperative Compassion
CON: cingulo-opercular network
DASS-21: 21-item Depression Anxiety Stress Scale
DMN: default mode network
EEG: electroencephalographic
fdr: false discovery rate
FPN: frontoparietal network

XSL•FO
RenderX

GAD-7: 7-item General Anxiety Disorder
HIPAA: Health Insurance Portability and Accountability Act
MAAS: Mindful Attention Awareness Scale
PHQ-9: 9-item Patient Health Questionnaire
RT: response time
SES: socioeconomic status

Edited by A Mavragani; submitted 03.12.24; peer-reviewed by J Traxler, L Bacalhau, TRA Kral; revised version received 14.05.25; accepted 15.05.25; published 17.07.25.

Please cite as:
Jaiswal S, Nan J, Dizon S, Young JO, Purpura SR, Manchanda JK, Ramanathan D, Kuo DJ, Mishra J
Breath-Focused Mindfulness and Compassion Training in Parent-Child Dyads: Pilot Intervention Study
JMIR Form Res 2025;9:e69607
URL: https://formative.jmir.org/2025/1/e69607
doi:10.2196/69607

© Satish Jaiswal, Jason Nan, Seth Dizon, Jessica O Young, Suzanna R Purpura, James K Manchanda, Dhakshin Ramanathan, Dennis J Kuo, Jyoti Mishra. Originally published in JMIR Formative Research (https://formative.jmir.org), 17.7.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Performance of a Retinal Imaging Camera With On-Device Intelligence for Primary Care: Retrospective Study

Matthew Silvestrini*, PhD; Clarissa Lui*, PhD; Anil Patwardhan, PhD; Ying Chen; Tayyeba Ali, MD; Elie Glik; Honglei Wu; Brian Levinstein; Adrianna Wenz; Nathan Shemonski, PhD; Lin Yang; Ian Atkinson, PhD; Sam Kavusi, PhD

Verily Life Sciences, 999 Bayhill Dr, San Bruno, CA, United States

*these authors contributed equally

Corresponding Author:
Matthew Silvestrini, PhD
Verily Life Sciences, 999 Bayhill Dr, San Bruno, CA, United States

Abstract

Background: Access to screening continues to be a barrier for the early detection of diabetic retinopathy (DR). Primary care–based diabetic retinopathy screening could improve access, but operational challenges, such as cost and workflow management, hamper the widespread adoption of retinal camera systems in primary care clinics in the United States.

Objective: This study aimed to develop and evaluate a retinal screening system suitable for integration into a primary care workflow.

Methods: We developed a nonmydriatic, 45° field imaging retinal camera system, the Verily Numetric Retinal Camera (VNRC; Verily Life Sciences LLC), able to generate high-fidelity retinal images enabled by on-device intelligent features. The VNRC output flows into cloud-based software that accepts and routes digitized images for grading. We evaluated the performance and usability of the VNRC in 2 studies. A retrospective performance study compared the performance of VNRC against a reference camera (Crystalvue NFC-700 [Crystalvue Medical]) as well as the correlation between VNRC capture status and gradability (as determined by ophthalmologist graders). The usability study simulated a primary care setting for a combined cohort of trained and untrained users (corresponding to patients in the simulation) and operators (corresponding to health care personnel in the simulation), where respondents completed a questionnaire about their user experience after attempting to capture images with the VNRC.

Results: In the comparative performance study (N=108, K=206 images), a total of 98.5% (203/206) of images captured by the VNRC were graded as sufficient for clinical interpretation compared to 97.1% (200/206) of Crystalvue NFC-700 images (difference in proportion was 0.015, 95% CI –0.007 to 0.033). In the quality control algorithm evaluation (N=172, K=343 images), we found a positive association (ϕ=0.58, 95% CI 0.45 - 0.69) between gradability status (gradable or nongradable) as determined by ophthalmologists and the capture status (recapture not-needed or needed) as determined by the VNRC quality control algorithm. In the usability study (n=15 users and n=30 operators), all participating users (15/15) indicated that they were able to have both eyes screened easily. Most users and operators indicated agreement (from somewhat agree to strongly agree) with statements describing the imaging process as intuitive (15/15, 100% and 29/30, 97%), comfortable (15/15, 100% and 30/30, 100%), and allowing for a positive experience (15/15, 100% and 30/30, 100%), of users and operators, respectively.

Conclusions: Our findings about the performance and usability of this retinal camera system support its deployment as an integrated end-to-end retinal service for primary care. These results warrant additional studies to fully characterize real-world usability across a wider and diverse set of primary care clinics.

(JMIR Form Res 2025;9:e70331) doi:10.2196/70331

KEYWORDS

diabetic retinopathy screening; retinal camera; retinal imaging; fundus imaging; diabetic retinopathy

Introduction

Diabetes mellitus (DM) is a growing epidemic that impacts more than 38 million Americans [1], and accounts for a substantial burden to the health care system (approximately US $412 billion) in both direct costs and lost productivity [2]. One of the most common complications associated with DM is diabetic retinopathy (DR), a progressive condition that impacts the microvasculature of the eye and results in irreversible damage to the retina [3,4]. DR, which occurs in approximately 25% of all patients with DM [5], is the leading cause of blindness in the United States and represents a major quality of

XSL•FO
RenderX

life burden. Even though studies have found that early detection of DR can mitigate vision loss by over 90% [6], diligent screening of patients with diabetes for DR in the United States remains encumbered by the need for a referral into an ophthalmology clinic. The logistical and financial barriers associated with specialty care are a major cause of noncompliance with the American Diabetes Association screening guidelines [7-9]. These barriers disproportionately impact rural and resource-poor areas, where limitations in the availability of specialists and ophthalmologic equipment may be associated with screening rates lower than 20% [10].

Implementation of DR screening in primary care clinics could improve access [11], and there are retinal camera systems suitable for primary care deployment, such as tabletop Crystalvue (Crystalvue Medical) or Topcon (Topcon Healthcare, Inc) models, and handheld Retinavue (Hillrom), Phelcom (Phelcom Technologies), Remidio (Remidio Innovative Solutions Pvt Ltd), or Volk (Volk Optical) models. Efficient DR screening in primary care, however, faces hurdles. Common issues that impact the consistency of image quality, such as handling small pupil sizes, cataracts, lid and lash occlusions, eye movements, or user positioning, may represent technical challenges for primary care settings. Mydriasis can be another technical challenge; while nonmydriatic imaging is operationally more approachable than mydriatic imaging, it yields some degradation in image quality [12,13]. In addition, the adoption of retinal camera systems in primary care clinics requires resource investments and efficient integration into the clinical workflow. It is therefore unsurprising that DR screening programs tend to succeed in large and well-resourced health organizations that can adopt mydriatic-based systems [14-16], while resource-constrained or small primary-care clinics face difficulties acquiring and operating DR programs [17]. Apart from the direct equipment purchasing costs, training and maintaining dedicated staff (in settings prone to high staff turnover) may become problematic [18]. Furthermore, these environments may often lack optimal infrastructures for electronic health record integration, information technology and Picture Archiving And Communication System systems, and tele-retinal grading services.

These issues are top concerns when tailoring DR screening systems toward primary care. Desirable features in this regard would include relatively affordable cost, as well as the feasibility of integration into existing workflows and of use by nondedicated staff. Yet, it remains imperative to achieve these attributes while maintaining diagnostic accuracy and integrity. DR screening systems need to produce images of sufficient quality to enable ophthalmologists to assess the presence of retinal disease appropriately. Retinal image quality is therefore a necessary condition that correlates with downstream gradability and clinical interpretability.

We report the development and performance characterization of a retinal screening camera system for use in a primary care practice. This system, developed as “Verily Retinal Camera” during the initial investigational period, now termed Verily Numetric Retinal Camera (VNRC; Verily Life Sciences LLC) and provides nonmydriatic, color fundus photos of the eye. Our objectives were threefold: (1) to evaluate the comparative performance of the VNRC against a standard state-of-the-art reference device, the Crystalvue NFC-700 (Crystalvue Medical), for providing high-quality images sufficient for clinical interpretation in a retina image analysis study; (2) to characterize the utility of the VNRC quality control algorithm, investigating the association between the algorithm outputs (scoring the image quality) and human graders’ assessments of whether images were gradable or ungradable; and (3) to assess the usability of the VNRC among potential operators and users in a simulated primary care setting.

Methods

Devices

The VNRC and the Crystalvue NFC-700 were used in this study. The VNRC is a nonmydriatic, macula-centered, 45° field imaging retinal camera system that uses a set of proprietary machine learning algorithms to generate high-fidelity composite color posterior chamber images of the eye by capturing and compiling a series of retinal images from each perceived single flash (<200 ms). The camera system is incorporated into a cloud-based software platform that accepts digitized images transferred from the VNRC and has the capacity to store, convert formats, display, and transfer data or imaging data between cameras (Figure S1 in Multimedia Appendix 1).

The Crystalvue NFC-700 camera [19,20] uses a conventional technique [21] to automatically adjust focus and capture the best quality image using a single annulus type of illumination. The image is captured by the built-in color complementary metal-oxide-semiconductor camera module.

Retrospective Retina Image Analysis Study

This was an analysis of previously collected paired-image acquisitions (discussed further in this study). The objective of this study was to characterize the overall performance of the VNRC, articulated in 2 aims, that is, first, characterize the comparative performance of the VNRC (as an investigational device) against a reference instrument to generate clinically interpretable images, and second, evaluate the utility of the image quality control feature (IQCF) embedded in the camera system (discussed in section “Development and Description of a Retinal Imaging Camera for Primary Care”), by investigating the association between the outputs of this algorithm scoring the quality (ie, the capture status) of the images and image gradeability as determined by ophthalmologists.

Study Participants and Image Eligibility

The original trial enrolled 212 participants of at least 22 years of age and who provided informed consent to the protocol; it was conducted in 4 separate phases at 2 sites (Verily Life Sciences South San Francisco headquarters and Diablo Clinical Research). This retrospective study was based on a trial subcohort consisting of non-Verily participants with self-reported history of diabetes (N=172; Figure S2 in Multimedia Appendix 1); this exclusion was to remove sponsor employees as participants in the evaluation of the device.

Analyses were conducted on retinal fundus images collected without mydriasis. For the comparative performance analysis

XSL•FO
RenderX

(first aim), eligible images were those from participants who achieved successful image acquisitions without the aid of mydriasis by both cameras (investigational and reference) from both eyes, and for which annotations about clinical interpretability from 3 ophthalmologists were available (n=108 participants). These excluded images captured with only one of the cameras, and also those lacking proper annotations.

For the IQCF utility analysis (second aim), eligible images were those from participants with at least 1 evaluable eye image obtained with the VNRC with a corresponding output from the reference human graders, namely, an annotation as "ungradable" or "gradable" after human inspection (n=172 participants). This excluded those images without proper grader annotation.

Study Procedures

Nonmydriatic retinal fundus images were acquired on 2 separate cameras from both right and left eyes of each participant, during a single session. Operators took up to 3 images per eye at their discretion, first on the Crystalvue NFC-700 followed by the VNRC.

Graders received images according to an established protocol. For the Crystalvue NFC-700 camera, graders received the last image acquired for each eye, and for the VNRC, graders received either the first image that passed the quality control feature for each eye, or the last image acquired per eye if no image had passed the quality control feature. This sequencing approach aimed to achieve parity using the highest quality image captured from each camera.

Images from both cameras were presented in random order to 3 board-certified ophthalmologist graders one by one using a grading platform (Figure S3 in Multimedia Appendix 1). Graders were external (non–sponsor-affiliated) professionals who assessed gradability and clinical significance as well as image quality characteristics according to a predefined rubric. The final outcomes were based on the majority vote across 3 graders.

Statistical Analyses

Aim 1

We compared the performance of VNRC to the reference camera (Crystalvue NFC-700) based on the proportion of acquired images that were determined to be "sufficient quality for clinical interpretation." This derived binary outcome was based on the majority vote of 3 board-certified ophthalmologists who each reviewed and annotated the same image as "Yes" or "No." Each participant could contribute up to 1 image from each eye for annotation and subsequent end point evaluation. Duplicate acquisitions of the same eye were not included in any analysis.

The estimated parameter for the end point was the difference in the proportion of images considered sufficient quality for clinical interpretation (ie, Yes), between those from the VNRC camera and those from the Crystalvue NFC-700 camera. Point estimates and 95% CIs for the reported proportions and difference in proportions were based on bootstrap (ie, "cluster" or "block" bootstrap) that accounted for the paired-image design (ie, correlated proportions) and within-participant clustering, as participants provided images from both the right and left eye. This cluster bootstrap was applied as follows:

Define: J=cluster unit = participant, where there may be multiple observation units (ie, eyes) within a participant. The sampling is based on the total number of J clusters.

The first step is to randomly select J number of clusters with replacement. For each cluster selected (with some clusters selected more than once and others not selected at all), all observations (ie, eyes) within that cluster are selected. Original cluster sizes are maintained.

In addition, due to the paired design, whenever an eye image is selected based on the output from 1 camera, the corresponding eye image from the comparative camera is also selected.

A difference in proportion is calculated based on the bootstrapped sample, and the process is repeated B number of times. Our analysis used B=10,000.

The point estimate for the difference in proportion is based on the 50th percentile of the resulting bootstrap distribution. Nonparametric 95% CIs for the difference in proportion were derived based on the 2.5% and 97.5% quantiles of the resulting bootstrap distribution.

Aim 2

We investigated the utility of the IQCF based on the association of the scores generated by this algorithm with the classification as "ungradable" or "gradable" by human graders. The algorithm scores image quality, generating an output of "recapture not needed" (eg, sufficient quality) or "recapture needed" (eg, insufficient quality) based on a score threshold. We reported the results from this portion of the study using summaries appropriate for categorical or ordinal data (counts and percentages). The correlation of gradeability status with outputs from the IQCF algorithm was summarized using both a contingency table and the Phi-coefficient.

Usability Study

Study Participants

This study was conducted across two groups (N=45) in a simulated primary care setting: (1) an operator cohort consisting of individuals with health care degrees or licenses (registered nurse, nurse practitioner, Licensed Practical Nurse, and physician assistant; n=15), or some health care training (master's degree, pharmacy technician, and phlebotomist; n=15), and (2) a user cohort consisting of participants with diabetes without health care training, who were asked to complete the VNRC retinal screening workflow unaided by a technician (n=15). This study was determined to be exempt research that did not require IRB approval.

Study Procedures

Operators (group 1) were trained in 2 subgroups. Individuals with health care degrees or licenses and those with some health care training. Training consisted of a visual or auditory slide presentation, demonstration, hands-on activities, and time for the sponsor to answer questions. Approximately half of each subgroup received in-person training, while the other half received one-on-one training from a remote sponsor representative via videoconference. All operators had access to the camera system during training sessions. Training lasted

XSL•FO
RenderX

approximately 1 hour. Trained participants experienced a decay period of at least 1 hour and up to 7 hours between their training and test sessions.

The simulated environment (ie, clinic) was equipped with an adjustable height table and all accessories required to use the retinal camera. Participants were asked to perform tasks within representative, naturalistic use scenarios; participants in the operator cohort interacted with pretrained actors (as stand-ins for hypothetical patients) who behaved consistently across simulations to elicit specific responses from operator participants. Moderators used a series of questionnaires around camera ease of use to collect participant feedback.

Analysis

Operator and user questionnaires followed Likert scales (from 1=strongly disagree to 7=strongly agree). We analyzed responses using descriptive statistics.

Development and Description of a Retinal Imaging Camera for Primary Care

The Verily Retinal Service was developed for use in primary care clinics. The VNRC is a lightweight (approximately 6 kg), 45° field imaging camera system consisting of custom electronics, optics, LEDs, and a retinal camera (Figures 1A and 1B). The VNRC has a range of pivot=0° to 45° to adjust for user height, posture, and comfort. The black face rest is light blocking, which enables retinal imaging in bright ambient lighting conditions, and the built-in handle allows for flexibility for camera placement within a clinic space.

Figure 1. Overview of the Verily Numetric Retinal Camera. (**A**) Illustration of the Verily Numetric Retinal Camera and the (**B**) architecture. (**C**) Operator's and user's view during image acquisition. Cam: camera.

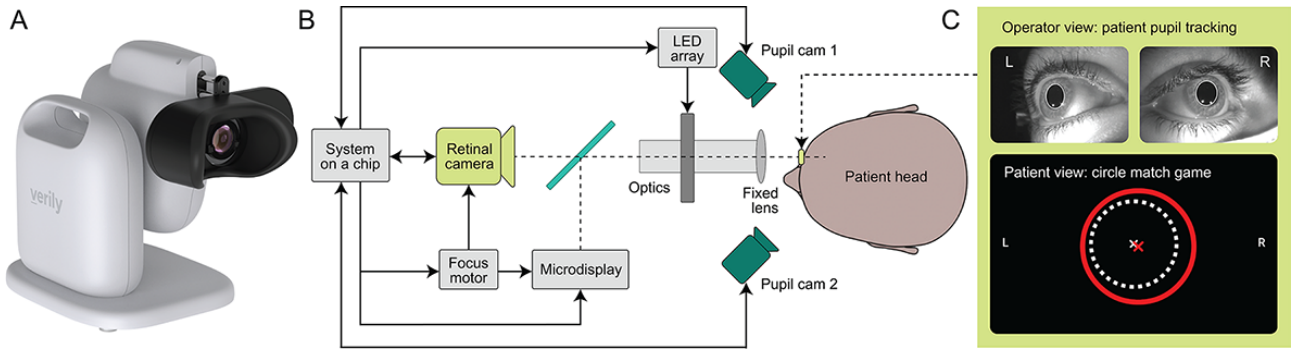

The VNRC uses a stereo infrared camera system (pupil camera 1 and 2 in Figure 1B) to achieve proper pupil alignment and facilitate fast image acquisition (120 frames per second [fps]). Users can adjust an image to their best perceived focus with a focus knob, and an interactive game-like interface allows users to optimize proper pupil alignment using small eye or head movements within the face rest, while operators can simultaneously oversee this process to ensure proper eye location (Figure 1C). The focus motor controls both the retinal camera and microdisplay, allowing the user to view an interactive image and video inside the camera. This is similar to an experience in a virtual reality headset, but monocular.

After the initial focus step, users position their head into the VNRC face rest in order to align their eye with the camera lens. Once proper focal length is established by dialing the focus knob, users can confirm by pressing the top of the focus knob. Alternatively, for users unable to use the focus knob, the VNRC can also be brought into focus using an operator-driven setting controlled via the software graphical user interface.

The VNRC then uses a proprietary redundant illumination system to automatically collect a series of retinal images (up to 30). This is in contrast to conventional techniques that only capture 1 image frame per flash [21] and is made possible by the use of high-speed high-sensitivity image sensors (120 fps) deeply integrated with the system-on-chip. The flash duration is limited to ensure sufficient quality, minimizing dazzle. Here, the system-on-chip uses an LED array with the camera system to capture a series of full-field (>45°) images of the retina with various illuminations focused on different areas of the crystalline lens of the eye (Figure 2). The proprietary redundant illumination system dynamically adapts to specific eye characteristics, such as pupil size and position, and optimizes a series of retinal illumination configurations, regardless of corneal clarity.

XSL•FO
RenderX

Figure 2. Verily Numetric Retinal Camera's pupil tracking adaptive illumination. (**A**) Diagram of the eye indicating where the retinal camera projects the illumination dot location onto the pupil area; (**B**) comparison between conventional flash and exposure cameras, where only 1 image is captured and has a higher likelihood of artifacts, and the Verily Numetric Retinal Camera method of oversampling, where a series of images are captured and merged into a final image and hence mitigating the artifacts from any one single image.

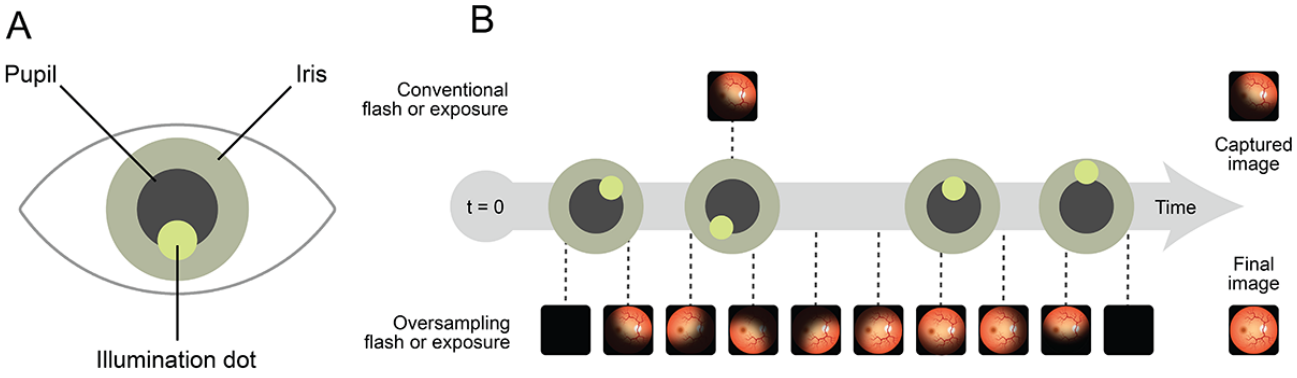

This sequence of flashes, or "burst imaging," illuminates the entire field of view, occurs within 200 milliseconds, and is perceived as a single flash. The acquisition time falls below the typical latency of the pupil's response time [22], so that the imaging process is not expected to interfere with natural pupil constriction. The typical burst image set contains many fully illuminated images, with varying levels of artifacts on the images.

The VNRC uses a proprietary "Burst Reduce Algorithm" to generate a single high-fidelity retinal image (Figures 3A and 3B), by merging a variable number of frames. The total number of frames is typically 20 images, but is highly dependent on the presence of artifacts and the amount of eye motion during the camera flash. The algorithm draws similarities from pixel-level high dynamic range imaging [23] but incorporates an assignment of a score layer to separate photons due to retinal reflection from scattered photons due to cataract or cornea reflection, or lid or lash occlusion, after compensating for intra-acquisition gaze shift. It is notable that no fine spatial filtering was applied to enhance the contrast or segment-specific features or pathologies [24,25].

Figure 3. Creation of a single composite image with Verily Numetric Retinal Camera. (**A**) Burst imaging captures up to 30 images from a single flash via oversampling (<200 ms). (**B**) Using a burst reduction algorithm, the Verily Numetric Retinal Camera creates a single high-fidelity composite image. (**C**) A quality control algorithm (the image quality control feature) is then used to predict the signal quality from the retina, and a score is assigned. A passing image quality control feature score indicates that an image is of sufficient quality and does not require recapture. IQCF: image quality control feature; QC: quality control.

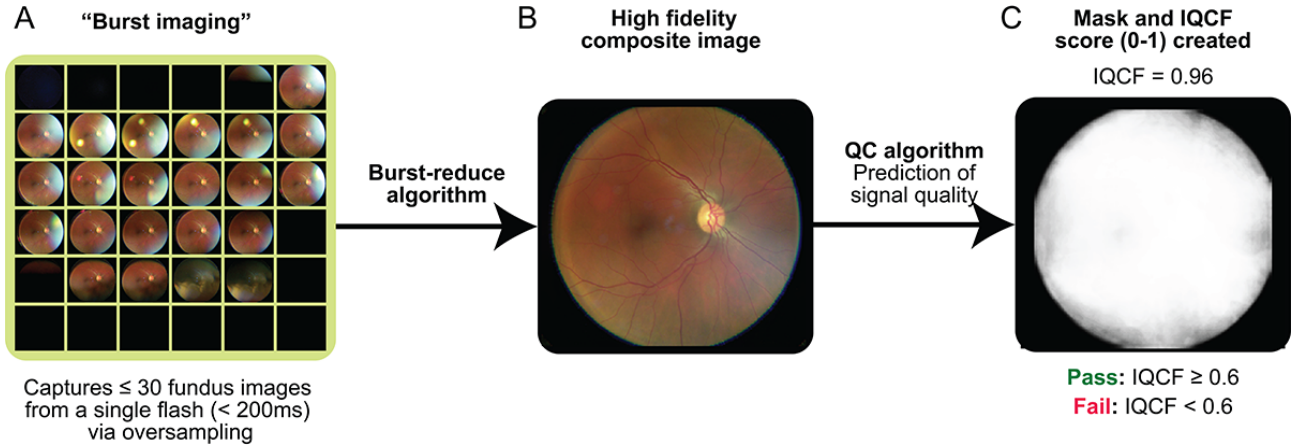

A subsequent algorithm, the IQCF algorithm, then performs a quality control assessment of the resulting composite image to determine if a recapture is needed (Figure 3C). The model was trained to compute areas where the retina is clearly visible, without impaired visibility due to (1) darkness, (2) saturation, (3) blur, or (4) haze. The IQCF algorithm produces a single score per image by calculating the probability that each pixel contributes to that of the retina signal, as represented by an image mask. The algorithm incorporates a fixed cutoff to determine if the image needs to be recaptured. If the score does not exceed this operating point, then the camera will output instructions to the operator that another image acquisition is required. Thus, the VNRC provides real-time image quality characterization information to help reduce the chances that a substandard image is used for clinical interpretation. Importantly, the IQCF does not score image gradeability.

The final output of the VNRC can integrate into a primary care workflow. It feeds into a cloud-based software platform that accepts digitized images transferred from the camera and has the capacity to transfer, store, convert formats, display, and transfer medical device data or medical imaging data between medical devices (Figure S1 in Multimedia Appendix 1). The VNRC also continuously collects metadata such as details on camera use, uptime, and operation in order to allow for real-time error handling. This passively collected metadata is uploaded along with the retinal images.

XSL•FO
RenderX

Ethical Considerations

This was a retrospective study conducted in a subset of the participants in a prospective technical feasibility trial (Verily protocol 103535; approved by institutional review board [IRB] Western IRB, before initiation, IRB Protocol #20214693) [26]. All participants signed informed consent (Multimedia Appendix 2) approved by the IRB and received nominal compensation for their time ($25 for the screening procedure, an additional $75 for each completed study visit). Personal study-related data were managed in accordance with local data protection law.

Results

Comparative Performance of the VNRC Against the Crystalvue NFC-700 Camera

Eye images (K=206) were captured from 108 participants (Table 1) with both the VNRC and the Crystalvue NFC-700 Camera.

Table . Demographic characteristics of participants in the comparative performance analysis within the retrospective retina image study (as reference, refer to the characteristics of all participants in the original technical feasibility trial in the 2 right columns).

Characteristics	Retrospective study comparative performance (N=108)	Initial technical feasibility trial (N=212)	
		Phase 1 - 3 (n=192)	Phase 4 (n=20)
Age (y)			
Mean (SD)	57.8 (13.2)	54.8 (16.6)	49.0 (10.3)
Median (range)	59.5 (23.0 - 84.0)	__[a]	—
Sex, n (%)			
Female	60 (55.6)	94 (49)	11 (55)
Male	48 (44.4)	98 (51)	9 (45)
Race, n (%)			
Asian	9 (8.3)	44 (22.9)	<5 (≤5)
Black or African American	18 (16.7)	18 (9.4)	<5 (≤5)
White	76 (70.4)	115 (59.9)	18 (90)
Other	5 (4.6)	15 (7.9)	—
Diagnosed diabetes[b], n (%)	108 (100)	123 (64.1)	20 (100)
For 0 - 15 y	—	35 (18.3)	5 (25)
For 15+ y	—	88 (45.8)	15 (75)

[a]Not available.

[b]A post hoc analysis showed 39 (36.1%) participants with mild diabetic retinopathy, 40 (37%) with moderate diabetic retinopathy, and none with severe diabetic retinopathy; 31 (28.7%) had proliferative diabetic retinopathy (based on image evaluation obtained with the reference camera).

The proportion of images of sufficient quality for clinical interpretation was 0.985 (203/206) and 0.971 (200/206) for the VNRC and Crystalvue NFC-700 cameras, respectively. The difference in proportion was 0.015 (95% CI –0.007 to 0.033; Table 2, examples in Figure 4).

Table . Comparative performance results.

N=108 (K=206 images)	Investigational VNRC[a] and reference Crystalvue
Images with sufficient quality for clinical interpretation[b], k (%)	203/206 (98.5) and 200/206 (97.1)
Difference in proportion (95% CI)	0.015 (–0.007 to 0.033)
Uncaptured images, k (n)	3 (2) and 42 (25)

[a]VNRC: Verily Numetric Retinal Camera.

[b]Some of the reasons that the images were of insufficient quality: eyelash artifacts, eyelash artifacts, low contrast, or extended areas where detail was lost.

XSL•FO
RenderX

Figure 4. Examples of images with or without sufficient quality for clinical interpretation from each of the cameras in the comparison. VNRC: Verily Numetric Retinal Camera.

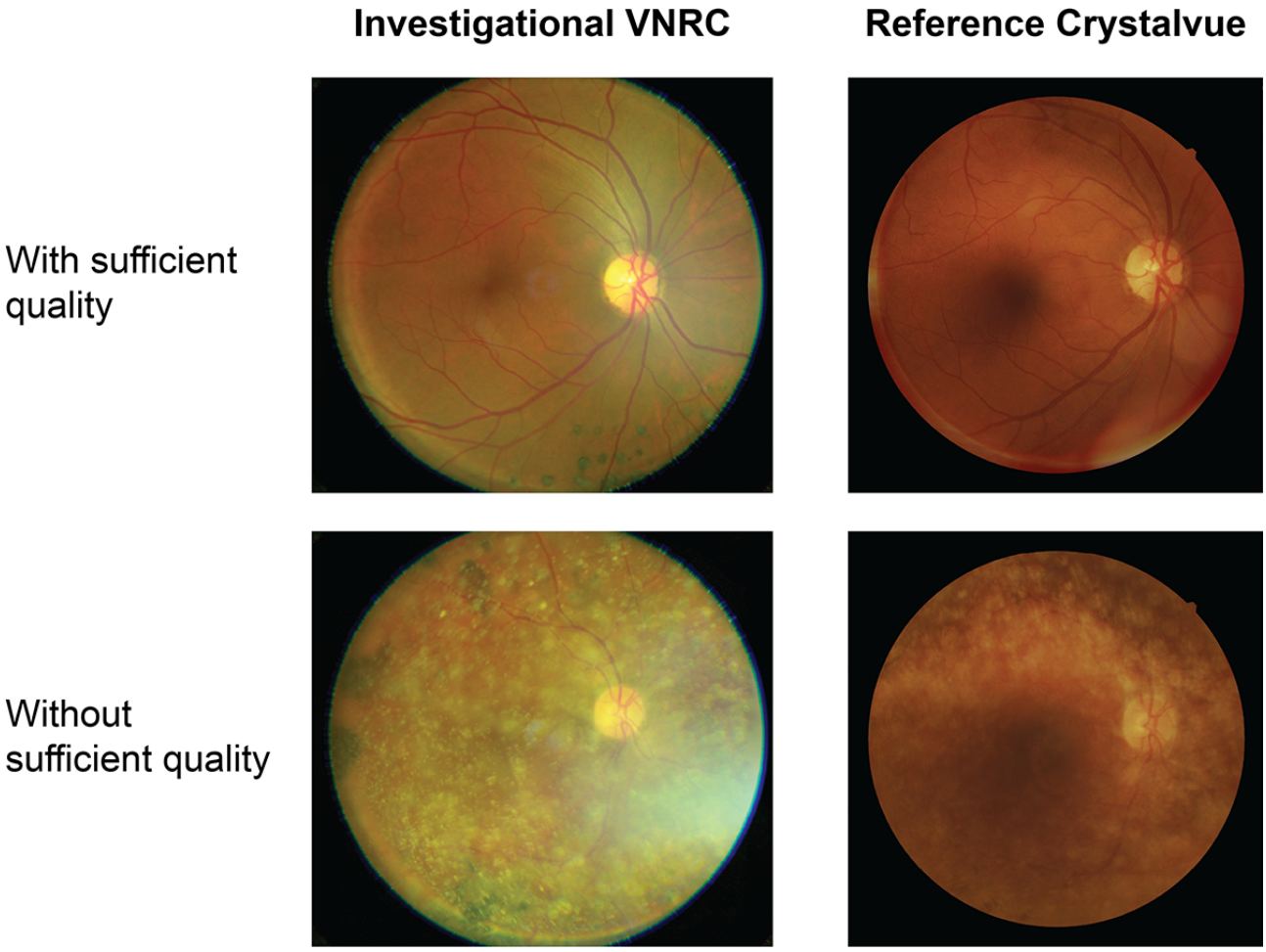

Graders also computed the quality of the images across several specific criteria, for both the VNRC and the reference camera (Tables S1 and S2 in Multimedia Appendix 1), indicating whether the quality of specific image aspects was sufficient for clinical interpretation (note that these were not pooled for a single majority classification; there were 3 separate adjudications for each factor in each image). Graders evaluated the visualization of the optic disc and determined that it was adequate in a majority of images for both devices (in at least 90%, 225/248 of images by the VNRC and in at least 93%, 195/209 of the reference). There were similar results in the classification of macula visualization (deemed appropriate in at least 92%, 229/248 of VNRC images and 83%, 175/209 of reference images), and the visualization of the retinal vessels (appropriate in at least 94%, 233/248 of the VNRC images and 92%, 194/209 of reference images). Graders also found a majority of images to be of sufficient quality regarding key imaging features, such as adequate focus (in at least 91%, 228/248 of VNRC images and 88%, 185/209 of reference images), appropriate brightness (in at least 94%, 213/248 and 198/209 of images, for both), adequate field of view (in at least 91%, 228/248 and 191/209 of images, for both), no significant image defects (in at least 91%, 227/248 and 92%, 194/209), no small pupil interference (at least 89%, 222/248 and 87%, 183/209), and no ocular media opacity (in at least 73%, 183/248 and 85%, 178/209 of VNRC and reference images, respectively).

Performance of the IQCF

We found a moderate association (ϕ=0.58) between ophthalmologists' assessments of a retinal image's gradeability and the IQCF algorithm's scoring of capture status (Table 3).

XSL•FO
RenderX

Table . Contingency table of image quality control factor scores of image quality and gradeability assessments determined by ophthalmology graders.

IQCF[a] score (N=172; K=343 images)	Graders' rating, n (%)		Total
	Nongradable	Gradable	
Recapture needed (IQCF=not pass)	20 (40[b])	30 (60)	50
Recapture not needed (IQCF=pass)	1 (0.3)	292 (99.7)	293

[a]IQCF: image quality control factor.

[b]Percentages reflect the total number of image quality control factor classifications that were determined to be nongradable or gradable during human assessment (ie, graders').

Overall, the IQCF scored 50 images as needing recapture, but the human assessment was "gradable" for 60% (30/50) of these. Conversely, the vast majority of the images scored by the IQCF as not needing recapture, 292 out of 293 (99.7%), were found "gradable" by human assessment. Refer to Figure 5 for examples of images with concordant and discordant IQCF scoring and human assessment.

XSL•FO
RenderX

Figure 5. Example images of different image quality control feature scores and human assessments. (**A**) Image quality control feature "pass" with acceptable image quality, but ungradable due to optic disc shift. (**B**) Image quality control feature "not pass" with poor image quality, but gradable due to visible lesions. (**C**) Image quality control feature "pass" and gradable. (**D**) Image quality control feature "not pass" and ungradable, due to eyelash artifacts.

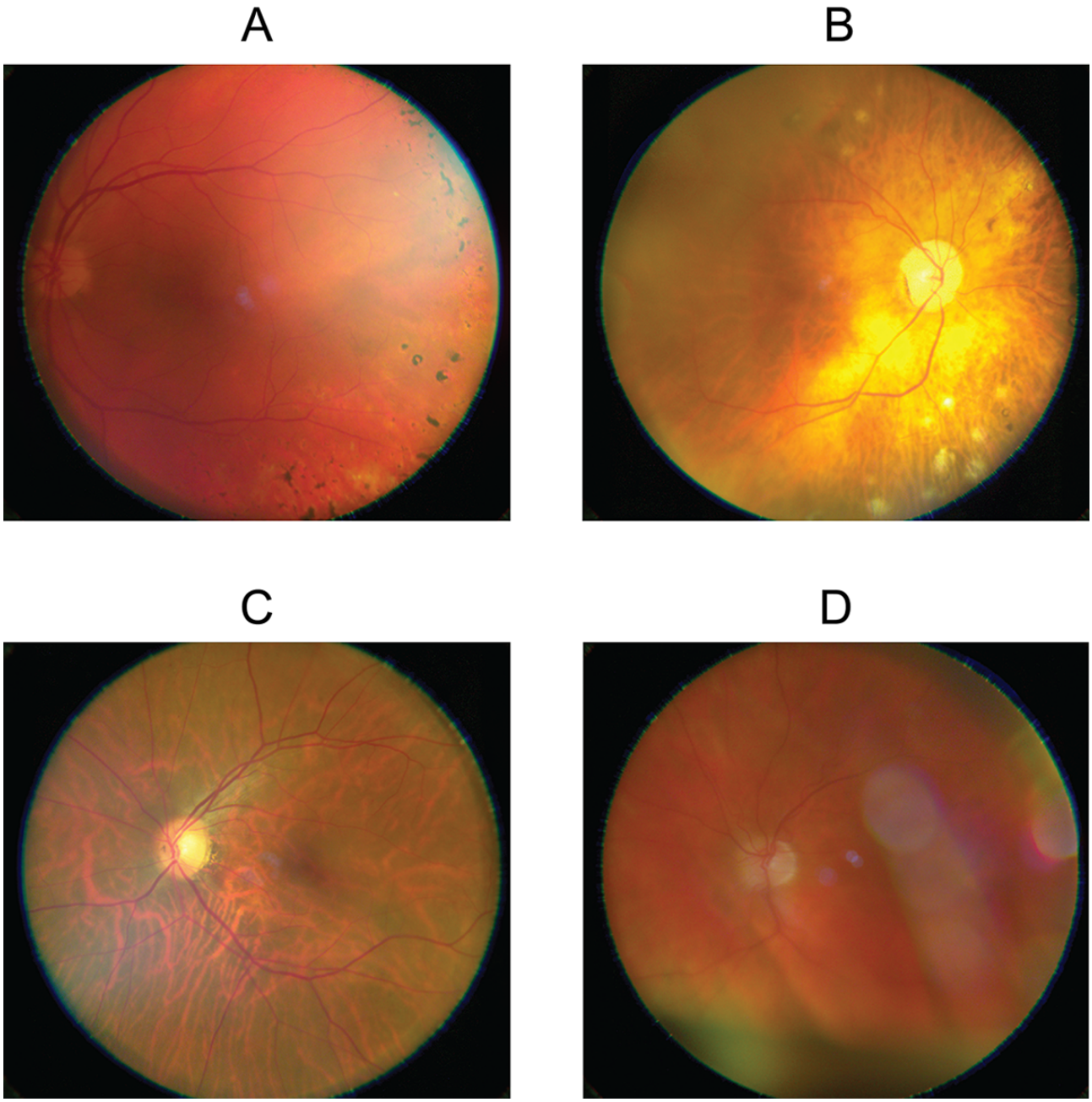

User Research Study

There was agreement among 100% (15/15) of the participating users in the simulated clinic environment that they were able to have both eyes screened easily. In addition, they somewhat agreed or strongly agreed (rating of 5 to 7 on a 1 - 7 Likert scale) with statements indicating that they were confident in knowing how to complete the screening after watching the video, that they found it intuitive to set themselves up with the camera properly, felt comfortable while completing the screening, and had a positive experience with the camera (Table 4, Table S3 in Multimedia Appendix 1).

XSL•FO
RenderX

Table . Summary questionnaire results for simulated users (no health care training).

Scale and question	Users (n=15), n (%)
Likert scales (range 1-7)[a], responses>4	
Q1: I felt confident that I knew how to do the screening after watching the video	15 (100)
Q2: I found it intuitive to get myself set-up with the camera properly	15 (100)
Q3: I felt comfortable when doing the screening	15 (100)
Q4: I had a positive experience using the camera	15 (100)
Binary (yes or no), response "yes"	
Q5: I had both my eyes screened easily	15 (100)

[a]Likert scale, where 1=strongly disagree and 7=strongly agree.

All participating operators under simulated conditions (30/30, 100%), who had either health care degrees, licenses, or some training somewhat agreed or strongly agreed (ratings of 5 to 7 on a 1 - 7 Likert scale) with statements indicating that they felt comfortable completing the screening, had a positive experience while using the camera, found training easy and useful to understand, found it easy to capture a retinal image with the camera, and found the camera easy to clean (Table 5, and Table S4 and S5 in Multimedia Appendix 1). A majority (23/30) felt that hands-on help for users was probably not needed, and approximately half (14/30) felt that they would not have to apply their relevant clinical training to complete specific tasks.

Table . Summary questionnaire results for simulated operators (with health care licenses or training).

Questions on Likert scale (range 1-7)[a]	Participants with responses >4, n (%)		
	Health care degree or license (n=15)	Health care training (n=15)	Pooled (n=30)
Q1: I felt comfortable when doing the screening	15 (100)	15 (100)	30 (100)
Q2: I feel like I needed to provide hands on help for the patient	1 (7)	6 (40)	7 (23)
Q3: I had a positive experience using the camera	15 (100)	15 (100)	30 (100)
Q4: I found the training easy to understand and useful	15 (100)	15 (100)	30 (100)
Q5: I found it easy to capture a retinal image with the camera	15 (100)	15 (100)	30 (100)
Q6: I had to apply my relevant clinical training to complete specific tasks	8 (53)	8 (53)	16 (53)
Q7: I found the camera user interface intuitive and easy to understand	15 (100)	14 (93)	29 (97)
Q8: I found the camera easy to clean	15 (100)	15 (100)	30 (100)

[a]Likert scale, where 1=strongly disagree and 7=strongly agree.

Most of the issues reported by users related to involuntary clicks back and forth in the user interface, particularly during the "focus image" steps. These issues, at most, caused delays (not failures) in the image capture process.

Discussion

Principal Findings

We report here the results of a series of analyses to characterize the performance and usability of a new retinal camera system aimed for implementation in primary care settings. The performance was satisfactory in 2 main aspects. First, the VNRC performed at comparable levels to a reference camera for the generation of quality retinal images; the numeric difference in the percentage proportion of images with quality to be clinically interpretable was low (0.015), and the CI for that difference straddled 0, indicating a likelihood that there was no difference between the 2 devices. Further reinforcing the main results, gradability across a variety of image quality metrics appeared numerically similar across VNRC and the reference camera. Second, the quality-scoring outputs of the quality control algorithm embedded in the system showed a moderate association (ϕ=0.58) with the classification of images as "gradable" or "ungradable" by human graders. Furthermore, operators and users (ie, individuals with and without previous health care training) found the system to be generally intuitive and approachable to use, allowing them to feel comfortable

XSL•FO
RenderX

performing image captures on their own (at least 95%, 44/45 of survey participants agreed to the corresponding statements).

Our findings indicate that VNRC can perform at the level of a standard tabletop retinal camera system. Thus, the VNRC may provide a balance of features that could mitigate some of the reservations from primary care providers to the adoption of DR screening programs. Across both primary care and specialty clinics, tabletop equipment may be the highest quality option for retinal imaging [27]; however, high direct costs and the resource and space demands of optimal installation (ie, dedicated darkened room) act as deterrents, particularly in underresourced environments. While handheld and smartphone-based cameras overcome these aspects, reports are mixed regarding the quality of their image capture and their ease of successful use for DR screening (for instance, the skill and finesse required to obtain proper eye alignment [13,28], particularly when imaging is performed without pharmacologic pupil dilation [13,18,29-31]. Most of the images excluded from the comparative analysis (for lack of bilateral counterparts) were in the reference camera group (probably due to the inability of that camera to capture images for pupil sizes below 3.5 mm). The rates of ungradable images we observed are encouragingly low, considering that other studies with mydriatic devices (where they would be expected to be lower) have reported approximately 6% [32], or ranges from 0% to approximately 28% with mydriatic cameras (when grading was done via algorithm) [33]. A possible downside of producing a high rate of ungradable images (which can be due to a variety of causes, such as small pupil size, eye pathologies) may be an inflation of referrals that can overwhelm downstream eye care services and reduce the overall cost-effectiveness of teleophthalmology programs [18,30]. The VNRC has advantages closely associated with handheld systems in cost, maneuverability, relatively low weight (10‑20 kg lighter than other tabletop systems), and relatively low space demands. Our results suggest that this system can produce images of quality comparable with expensive and more sophisticated tabletop equipment, overall and across specific image metrics, possibly reaching a balance between operational requirements and quality performance, which is particularly well-suited to primary care clinics.

Our results also indicate that the quality control feature within the VNRC system, the IQCF, functions as intended and could effectively filter out the intake of undesirable, poor-quality images into practical clinical workflows. Capturing retinal images of quality is a necessary requisite for the subsequent clinical utility of those images during diagnosis; entering a high proportion of low-quality images could render a clinical workflow inefficient and impair effective care. The approach presented in this work was to embed an algorithm to score image quality in real time, prompting users to discard low-quality images (those below a score threshold) and recapture before entering an image into the clinical workflow. Thus, it was important to establish the correlation between the proportion of retinal images passing the VNRC's IQCF threshold and downstream assessments of gradability according to ophthalmic graders. We found a moderate correlation, as nearly all the images cleared by the IQCF as "not needing recapture" were indeed found "gradable" by human assessment. The IQCF scored noticeably more images as "needing recapture" than human graders deemed "ungradable" (ie, 60%, 30/50 of images that IQCF labeled as needing recapture were actually deemed gradable during human assessment). This is probably due to the fact that the IQCF is a visibility measure, with an output that is pathology-independent; in contrast, human graders may grade images if a lesion is clearly identifiable, even in a context of low overall quality or visibility that the IQCF probably would "not pass." This excess of images with "recapture needed" scores, however, may not represent a major practical problem, since image recapture in real time may not be a burdensome procedure. Ultimately, this ensures that the IQCF facilitates an inflow of quality images, without erring in a direction that would create backflows or inefficiencies, impairing practical clinic workflows after the fact.

Our third major result showed that participants found that handling the VNRC system was easy and intuitive and felt comfortable with it. This addresses another relevant concern for primary care practice managers considering retinal screening systems, namely, the actual or perceived need for dedicated trained staff. Furthermore, systems with which users and operators experience repeated lack of success in producing images of sufficient quality may undermine confidence. In turn, this can nudge personnel toward lower usage [18] and depress the cost-effectiveness of a screening program. Our survey results indicate that the VNRC could be an approachable system that mitigates usability barriers.

While our analyses yielded promising results, the characterization of this new DR screening system had some limitations, largely related to the generalizability of our findings. First, the images for the comparative performance analysis were collected first with the VNRC, followed by the reference camera, in order to maintain internal consistency with the larger study from which this retrospective analysis was undertaken [26]. Therefore, we cannot discount the possibility of experimental bias related to the order of measurement (it could be a learning effect or a fatigue effect) that could have influenced the results. Second, our analyzable dataset for the comparative analysis images excluded participants without complete bilateral, nonmydriatic image sets and annotations, which could have exerted some selection bias toward better-quality cases; while that effect would be expected to some extent for both groups in the comparison, it may have had a differential effect that we cannot discount. Third, we cannot interpret the potential impact of not-captured images in our comparative performance analysis. Some of those noncaptures may correspond to smaller pupil sizes, since the instructions for use for Crystalvue NFC-700 require a minimum pupil size of 4 mm. Therefore, different study conditions (eg, dimming ambient light in order to dilate pupil for an attempted recapture) may yield different comparative results from the ones in this report. Fourth, the VNRC IQCF would encourage image recaptures for low-quality scores in an actual clinic setting, until the operator has a total of 3 eye images or one of the images scores above the threshold. In this study, graders received image captures according to a prespecified sequence that may not have consistently maximized quality. Therefore, future studies are warranted to investigate performance across a more diverse set of conditions. For

XSL•FO
RenderX

instance, future postmarket analyses could evaluate performance relying on the images with the top IQCF score across all eye captures, or in different environmental setups (such as lighting), to better reflect performance in real-clinic conditions. Another area for future study is the characterization of variability in larger cohorts, more widely representative of primary-care screening populations. Importantly, disease characteristics (burden and type of disease) may impact the relative performance in terms of gradeability and the association of image quality with gradeability. Finally, we did not collect information on visual acuity from the participants in the user study; while results were overall favorable, visual acuity (or lack thereof) may affect user-device interactions; therefore, it will be worthwhile to gather appropriate information in future studies to better understand it.

Conclusions

In summary, barriers to primary case-based DR screening overall are multifactorial [18,34,35]. Providers tend to perceive that rigorous and cost-effective implementation of teleophthalmology in primary care settings is expensive and difficult, demanding restructuring of work processes and increasing the burden on clinic staff [18,36]. Our results indicate that the VNRC system could mitigate some of these issues, particularly in underresourced environments. It has the positive operational characteristics akin to handheld systems, produces images of quality comparable with standard tabletop retinal cameras, and is able to optimize the inflow of quality images into clinical workflows. In addition, users note that they are able to handle the system and produce usable images with ease. Furthermore, the transfer and flow of the digital output are adaptable to typical primary care workflows. These results support considering this system as an integrated end-to-end retinal service suitable for primary care and warrant additional studies across a wider and diverse set of primary care clinics. Novel DR screening systems that address primary care adoption barriers may represent an advance toward more widespread access, with the potential to curtail rates of severe disease progression at the population level and ultimately contribute to better patient outcomes.

Acknowledgments

The Retrospective Retina Image Analysis Study and this analysis were funded by Verily Life Sciences, San Bruno, California.

Verily Life Sciences is the funding source for this study and is responsible for data collection. Authors were fully responsible for the data analysis, interpretation presented herein, and the writing of this article. AP, CL, and MS had access to the raw data. Authors had access to the full dataset for the study, reviewed, and approved the final manuscript for submission.

Data Availability

The raw data in this analysis are not available for sharing.

Authors' Contributions

Study concept and design were contributed by CL, SK, and AP. Data collection was handled by Verily Life Sciences. Data analysis and interpretation were contributed by AP, CL, and SK. All authors contributed to draft writing and review and draft approval for submission. All authors wish to acknowledge writing and editing support from Julia Saiz from Verily Life Sciences.

Conflicts of Interest

All authors report employment in Verily Life Sciences. MS, CL, AP, YC, EG, HW, BL, AW, NS, LY, IA, and SK report equity ownership in Verily Life Sciences.
The Verily Retinal Camera or Verily Numetric Retinal Camera described in this report is a product for research use only, not for use in diagnostic procedures.

Multimedia Appendix 1
Additional methods and results.
[DOCX File, 2492 KB - formative_v9i1e70331_app1.docx]

Multimedia Appendix 2
Consent form.
[PDF File, 185 KB - formative_v9i1e70331_app2.pdf]

References

1. National diabetes statistics report. : Centers for Disease Control and Prevention; 2024 May 15 URL: https://www.cdc.gov/diabetes/php/data-research/?CDC_AAref_Val=https://www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetes-statistics-report.pdf [accessed 2024-11-18]
2. Parker ED, Lin J, Mahoney T, et al. Economic costs of diabetes in the U.S. in 2022. Diabetes Care 2024 Jan 1;47(1):26-43. [doi: 10.2337/dci23-0085] [Medline: 37909353]

XSL•FO
RenderX

3. Duh EJ, Sun JK, Stitt AW. Diabetic retinopathy: current understanding, mechanisms, and treatment strategies. JCI Insight 2017 Jul 20;2(14):e93751. [doi: 10.1172/jci.insight.93751] [Medline: 28724805]
4. Markle J, Shaia JK, Araich H, Sharma N, Talcott KE, Singh RP. Longitudinal trends and disparities in diabetic retinopathy within an aggregate health care network. JAMA Ophthalmol 2024 Jul 1;142(7):599-606. [doi: 10.1001/jamaophthalmol.2024.0046] [Medline: 38869883]
5. Lundeen EA, Burke-Conte Z, Rein DB, et al. Prevalence of diabetic retinopathy in the US in 2021. JAMA Ophthalmol 2023 Aug 1;141(8):747-754. [doi: 10.1001/jamaophthalmol.2023.2289] [Medline: 37318810]
6. American Diabetes Association Professional Practice Committee. 12. Retinopathy, neuropathy, and foot care: standards of care in diabetes-2024. Diabetes Care 2024 Jan 1;47(Suppl 1):S231-S243. [doi: 10.2337/dc24-S012] [Medline: 38078577]
7. Chou CF, Sherrod CE, Zhang X, et al. Barriers to eye care among people aged 40 years and older with diagnosed diabetes, 2006-2010. Diabetes Care 2014;37(1):180-188. [doi: 10.2337/dc13-1507] [Medline: 24009300]
8. Hartnett ME, Key IJ, Loyacano NM, Horswell RL, Desalvo KB. Perceived barriers to diabetic eye care: qualitative study of patients and physicians. Arch Ophthalmol 2005 Mar;123(3):387-391. [doi: 10.1001/archopht.123.3.387] [Medline: 15767483]
9. Chan AX, McDermott Iv JJ, Lee TC, et al. Associations between healthcare utilization and access and diabetic retinopathy complications using All of Us nationwide survey data. PLoS ONE 2022;17(6):e0269231. [doi: 10.1371/journal.pone.0269231] [Medline: 35704625]
10. Daskivich LP, Vasquez C, Martinez C, Tseng CH, Mangione CM. Implementation and evaluation of a large-scale teleretinal diabetic retinopathy screening program in the Los Angeles County Department of Health Services. JAMA Intern Med 2017 May 1;177(5):642-649. [doi: 10.1001/jamainternmed.2017.0204] [Medline: 28346590]
11. Scanlon PH. The English National Screening Programme for diabetic retinopathy 2003-2016. Acta Diabetol 2017 Jun;54(6):515-525. [doi: 10.1007/s00592-017-0974-1] [Medline: 28224275]
12. Malerbi FK, Morales PH, Farah ME, et al. Comparison between binocular indirect ophthalmoscopy and digital retinography for diabetic retinopathy screening: the multicenter Brazilian type 1 diabetes study. Diabetol Metab Syndr 2015;7(1):116. [doi: 10.1186/s13098-015-0110-8] [Medline: 26697120]
13. Salongcay RP, Aquino LAC, Salva CMG, et al. Comparison of handheld retinal imaging with ETDRS 7-standard field photography for diabetic retinopathy and diabetic macular edema. Ophthalmol Retina 2022 Jul;6(7):548-556. [doi: 10.1016/j.oret.2022.03.002] [Medline: 35278726]
14. Jalkiewicz JF. A screening success. NYC Health+Hospitals takes diabetic retinopathy screening to new heights. Ophthalmol Management 2022:46-47 [FREE Full text]
15. Fong DS, Gottlieb J, Ferris FL, Klein R. Understanding the value of diabetic retinopathy screening. Arch Ophthalmol 2001 May;119(5):758-760. [doi: 10.1001/archopht.119.5.758] [Medline: 11346406]
16. Hudson SM, Modjtahedi BS, Altman D, Jimenez JJ, Luong TQ, Fong DS. Factors affecting compliance with diabetic retinopathy screening: a qualitative study comparing English and Spanish speakers. Clin Ophthalmol 2022;16(16):1009-1018. [doi: 10.2147/OPTH.S342965] [Medline: 35400992]
17. Liu Y, Zupan NJ, Shiyanbola OO, et al. Factors influencing patient adherence with diabetic eye screening in rural communities: a qualitative study. PLoS ONE 2018;13(11):e0206742. [doi: 10.1371/journal.pone.0206742] [Medline: 30388172]
18. Liu Y, Torres Diaz A, Benkert R. Scaling up teleophthalmology for diabetic eye screening: opportunities for widespread implementation in the USA. Curr Diab Rep 2019 Aug 2;19(9):74. [doi: 10.1007/s11892-019-1187-5] [Medline: 31375932]
19. NFC-700 fully automated non-mydriatic retinal camera. Crystalvue. URL: https://www.crystalvue.com.tw/data/files/1623122722411841591.pdf [accessed 2024-11-18]
20. 510(k) summary. : US Food and Drug Administration; 2019 Jan 2 URL: https://www.accessdata.fda.gov/cdrh_docs/pdf18/K182199.pdf [accessed 2024-11-18]
21. DeHoog E, Schwiegerling J. Fundus camera systems: a comparative analysis. Appl Opt 2009 Jan 10;48(2):221-228. [doi: 10.1364/ao.48.000221] [Medline: 19137032]
22. Bista Karki S, Coppell KJ, Mitchell LV, Ogbuehi KC. Dynamic pupillometry in type 2 diabetes: pupillary autonomic dysfunction and the severity of diabetic retinopathy. Clin Ophthalmol 2020;14:3923-3930. [doi: 10.2147/OPTH.S279872] [Medline: 33244218]
23. Kavusi S, El Gamal A. Quantitative study of high-dynamic-range image sensor architectures. : SPIE; 2004 Jun 7 Presented at: Sensors and Camera Systems for Scientific, Industrial, and Digital Photography Applications V; San Jose, CA p. 264-275. [doi: 10.1117/12.544517]
24. Sahu S, Singh AK, Ghrera SP, Elhoseny M. An approach for de-noising and contrast enhancement of retinal fundus image using CLAHE. Optics & Laser Technology 2019 Feb;110:87-98. [doi: 10.1016/j.optlastec.2018.06.061]
25. Youssif AA, Ghalwash AZ, Ghoneim AS. Comparative study of contrast enhancement and illumination equalization methods for retinal vasculature segmentation. 2006 Presented at: Cairo International Biomedical Engineering Conference (CIBEC'06) p. 1-5 URL: https://www.researchgate.net/publication/224014522_Comparative_Study_of_Contrast_Enhancement_and_Illumination_Equalization_Methods_for_Retinal_Vasculature_Segmentation [accessed 2025-07-03]

XSL•FO
RenderX

26. Makedonsky K, Patwardhan A, Kim A, Silvestrini M, Lui C, Kavusi S. Technical feasibility study for a retinal camera system. Endocrinology (including Diabetes Mellitus and Metabolic Disease). Preprint posted online on 2024. [doi: 10.1101/2024.12.20.24319469]
27. Bascaran C. Key considerations when choosing a retinal camera for diabetic retinopathy screening. Community Eye Health 2023;36(119):4-5. [Medline: 37600682]
28. Cuadros J, Bresnick G. Can commercially available handheld retinal cameras effectively screen diabetic retinopathy? J Diabetes Sci Technol 2017 Jan;11(1):135-137. [doi: 10.1177/1932296816682033] [Medline: 28264174]
29. Leonard C. Retinal imaging on the go. Rev Ophthalmol 2020 [FREE Full text]
30. Widner K, Virmani S, Krause J, et al. Lessons learned from translating AI from development to deployment in healthcare. Nat Med 2023 Jun;29(6):1304-1306. [doi: 10.1038/s41591-023-02293-9] [Medline: 37248297]
31. Dow ER, Khan NC, Chen KM, et al. AI-human hybrid workflow enhances teleophthalmology for the detection of diabetic retinopathy. Ophthalmol Sci 2023 Dec;3(4):100330. [doi: 10.1016/j.xops.2023.100330] [Medline: 37449051]
32. Aquino LAC, Salongcay RP, Alog GP, et al. Comparison of 2-field and 5-field mydriatic handheld retinal imaging in a community-based diabetic retinopathy screening program. Ophthalmologica 2023;246(3-4):203-208. [doi: 10.1159/000530903] [Medline: 37231995]
33. Kubin AM, Huhtinen P, Ohtonen P, Keskitalo A, Wirkkala J, Hautala N. Comparison of 21 artificial intelligence algorithms in automated diabetic retinopathy screening using handheld fundus camera. Ann Med 2024 Dec;56(1):2352018. [doi: 10.1080/07853890.2024.2352018] [Medline: 38738798]
34. Piyasena M, Murthy GVS, Yip JLY, et al. Systematic review on barriers and enablers for access to diabetic retinopathy screening services in different income settings. PLoS One 2019;14(4):e0198979. [doi: 10.1371/journal.pone.0198979] [Medline: 31013274]
35. Lu Y, Serpas L, Genter P, Anderson B, Campa D, Ipp E. Divergent perceptions of barriers to diabetic retinopathy screening among patients and care providers, Los Angeles, California, 2014–2015. Prev Chronic Dis 2014;13(13). [doi: 10.5888/pcd13.160193]
36. Panwar N, Huang P, Lee J, et al. Fundus photography in the 21st century--a review of recent technological advances and their implications for worldwide healthcare. Telemed J E Health 2016 Mar;22(3):198-208. [doi: 10.1089/tmj.2015.0068] [Medline: 26308281]

Abbreviations

DM: diabetes mellitus
DR: diabetic retinopathy
fps: frames per second
IQCF: image quality control feature
IRB: institutional review board
VNRC: Verily Numetric Retinal Camera

Edited by J Sarvestan; submitted 23.12.24; peer-reviewed by FK Malerbi, M Shen, V Villabona-Martinez; revised version received 01.05.25; accepted 02.05.25; published 17.07.25.

Please cite as:
Silvestrini M, Lui C, Patwardhan A, Chen Y, Ali T, Glik E, Wu H, Levinstein B, Wenz A, Shemonski N, Yang L, Atkinson I, Kavusi S
Performance of a Retinal Imaging Camera With On-Device Intelligence for Primary Care: Retrospective Study
JMIR Form Res 2025;9:e70331
URL: https://formative.jmir.org/2025/1/e70331
doi:10.2196/70331

© Matthew Silvestrini, Clarissa Lui, Anil Patwardhan, Ying Chen, Tayyeba Ali, Elie Glik, Honglei Wu, Brian Levinstein, Adrianna Wenz, Nathan Shemonski, Lin Yang, Ian Atkinson, Sam Kavusi. Originally published in JMIR Formative Research (https://formative.jmir.org), 17.7.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Remote Patient Monitoring for Global Emergencies: Case Study in Patients With COVID-19

Ramin Ramezani[1], PhD; Wenhao Zhang[1], PhD; Minh Cao[2], MSc; Alex Bui[3], PhD; Antonia Petruse[3], MBA; Amelia Weldon[3], MA; Arash Naeim[3], MD, PhD

[1]Department of Computer Science, University of California, Los Angeles, 550, 404 Westwood Plaza, Engineering 6, Los Angeles, CA, United States
[2]Department of Bioengineering, University of California, Los Angeles, Los Angeles, CA, United States
[3]Clinical and Translational Science Institute (CTSI), University of California, Los Angeles, Los Angeles, CA, United States

Corresponding Author:
Ramin Ramezani, PhD
Department of Computer Science, University of California, Los Angeles, 550, 404 Westwood Plaza, Engineering 6, Los Angeles, CA, United States

Abstract

Background: The COVID-19 pandemic has highlighted the critical need for telehealth and remote patient monitoring in health care delivery. Despite the growing use of on-body wearable sensors for continuous monitoring and predicting adverse events, their widespread adoption remains a significant challenge. While the pandemic has accelerated the acceptance of these technologies, achieving widespread integration requires their sustained incorporation into routine health care practices beyond emergencies. In this study, we extend the application of our previously developed remote patient monitoring system to patients with COVID-19.

Objective: Our objective is to assess whether the metrics obtained from our previously developed system can provide additional insights into the recovery trajectory of individuals affected by COVID-19. This case study aims to demonstrate that remote patient monitoring systems can be adapted to diverse patient cohorts during emergencies. We aim to illustrate the ease of deployment, particularly when these systems are already integrated into the existing health care ecosystem.

Methods: From November 2020 to July 2021, a total of 73 patients were recruited through the University of California, Los Angeles, Center for Smart Health, after having consented to participate in this study for 2 weeks. The research concentrated on an exploratory analysis, focusing on the detailed examination of characteristics and behaviors of patients with COVID-19 as captured by the remote patient monitoring system. We collected day-to-day changes in the following sensor measurements: daily activity, daily energy expenditure, indoor localization, SpO_2, respiratory rate, heart rate, and temperature.

Results: Out of the 73 patients satisfying the inclusion criteria, 41 successfully adhered to using the monitoring technology, with only 22 providing substantial watch data (>4 h). Among the participants, 39 used the pulse oximeter, 37 used the thermometer, and 36 used respiratory monitoring at night. This study demonstrated an overall increase in patients' activity levels toward the end of this study, with many beginning to leave their homes after 2 weeks. Additionally, respiratory rates shifted toward healthier lower levels, and oxygen saturation improved. Fatigue and headache were identified as the most prevalent symptoms, followed by cough and loss of smell.

Conclusions: The conclusion highlights the critical importance of monitoring patients outside of hospital settings, especially during pandemics, when patients travel to hospitals or receive home visits by health care professionals, which could increase the risk of disease transmission. Studies demonstrating the benefits and efficacy of remote monitoring in home settings can better prepare health care professionals for future pandemic events. Continuous monitoring of a wide range of patient metrics, from activities to vital signs, and integration of these data into electronic health records would not only improve accuracy and reduce the burden of data collection but also pave the way for enhanced home care, offering higher quality care at a lower cost.

(JMIR Form Res 2025;9:e66773) doi:10.2196/66773

KEYWORDS

COVID-19; physical activity; health care delivery models; wearable sensors; indoor localization; Bluetooth Low Energy beacons; smartwatches; mobile phone; remote sensing technology

Introduction

In late 2019, an outbreak of COVID-19 emerged and spread rapidly worldwide. As of May 5, 2020, when we began this study, there were more than 3.6 million confirmed cases and >250,000 deaths attributed to COVID-19 worldwide, with over 69,000 deaths alone in the United States. These numbers have reached 7 million deaths worldwide, including 1.2 million in

XSL•FO
RenderX

the United States alone [1,2]. Despite the development of multiple vaccines, the virus and its new variants continue to wreak havoc in the public health sector. While the mortality rate for COVID-19 was reported to be around 2%, the number was significantly increased among older patients with underlying coexisting conditions. Notably, in hospitalized patients, the death rate approached 15%. Individuals who were at mild to moderate risk, such as those with cancer, could demonstrate significant clinical deterioration within 24 - 48 hours.

In recent years, during the COVID-19 pandemic, we observed the implementation of policies such as "shelter in place" aimed at preventing the unnecessary strain on the health care system, measures that may become necessary in the face of another pandemic. Some high-risk patients were instructed to "weather the storm" at home without adequate monitoring, despite the potential for rapid deterioration in their health [3]. Over the past decade, the emergence of commercially available, affordable, and lightweight sensors has significantly accelerated the adoption of remote patient monitoring systems within the health care system for continuous and comprehensive patient tracking. Even before the COVID-19 pandemic, substantial evidence indicated that continuous monitoring of vital signs, such as pulse oximetry and heart rate, was associated with reduced mortality [4,5]. This aligns with numerous studies conducted during the pandemic, which explored the use of pulse oximetry and other wearable devices for continuous vital sign monitoring [4,6,7]. During the COVID-19 pandemic, numerous studies evaluated the use of technology for patient management and assessed its effectiveness in reducing hospitalizations [3,6,8-11]. While some initiatives were labeled as remote patient monitoring, they often focused more on telehealth, e-visits, and the use of patients' existing technologies, such as smartphones to connect to health care professionals. These studies demonstrated encouraging results, indicating that technology can significantly enhance patient management [6,8,9].

In a series of studies, we had introduced and documented the development and implementation of our remote patient monitoring system and its clinical validation within older patient populations at risk of various health conditions [12-15]. This platform, known as Sensing At-Risk Population (SARP), encompasses activity monitoring through smartwatches, indoor localization using stationary beacons, and the collection of additional physiological data via wireless sensors. Data are securely and automatically transmitted to a Health Insurance Portability and Accountability Act (HIPAA)–compliant cloud infrastructure. For this study, we expanded the capabilities of our system to serve as a monitoring hub and to support additional devices tailored to the needs of COVID-19 positive patients. We incorporated commercially available Bluetooth-enabled thermometers, pulse oximeters, and respiratory distress monitors. These additions enabled us to capture relevant information for monitoring and assessing COVID-19 positive cases. The primary objective of this study was to enhance existing outpatient COVID-19 positive clinical trials, which incorporate the variables that our system captures as secondary or exploratory endpoints. By leveraging our platform, we aimed to minimize COVID-19 positive patients' exposure while collecting crucial vital information. We intended to gather this data within a natural environment, shedding light on the diverse array of symptoms presented by COVID-19 positive patients, thus contributing to a better understanding of this disease and the ongoing pandemic using the emerging remote patient monitoring systems.

Methods

Design, Setting, and Participants

Of study design and participants, patient recruitment occurred from November 2020 to March 2021 at the University of California, Los Angeles (UCLA) Health Center. We emphasized that participation in the study would not affect their care at UCLA and was entirely separate from their medical treatment. Participation would conclude either after 2 weeks or if the patient was admitted to the hospital during that timeframe. This study's protocol was obtained from the Institutional Review Board of UCLA.

Eligible participants were individuals aged 18 years or older with a COVID-19 diagnosis who were not hospitalized or had known exposure to a COVID-19 positive case. They required the ability to manage their condition at home, access to home Wi-Fi, proficiency in either English or Spanish, and a willingness to provide informed consent by signing the approved form (IRB #20 - 001565) from the UCLA, titled "Early Detection of Health Improvement and Decline Through Remote Health Monitoring in COVID-19 Positive Patients."

Inclusion and Exclusion Criteria

Inclusion criteria were wearable device compatibility (ability to wear a watch), willingness to host the remote monitoring system for 2 weeks, aged 18 years or older, current UCLA Health patient, confirmed positive COVID-19 laboratory result, or known exposure to COVID-19.

Exclusion criteria were clinical diagnosis of movement disorders (eg, Parkinson Disease) and failure to meet inclusion criteria.

About SARP

SARP is a remote patient monitoring system developed by UCLA's Center for SMART Health, designed to cater to patients beyond the confines of health care institutions. Its primary purpose is to monitor vulnerable, at-risk populations by simulating the measurement of activities of daily living and instrumental activities of daily living [16] through cost-effective sensor technology. SARP has been used to generate prognostic data and predictive models for mortality and functional decline [12-15].

The core components of SARP encompass hardware, including an Android Smartwatch and readily available proximity Bluetooth Low Energy (BLE) beacons, and clinically validated software featuring activity recognition and indoor localization algorithms. Additionally, SARP incorporates a remotely triggered adaptive smart questionnaire mechanism, data visualization tools, and algorithms for assessing frailty, all within a HIPAA-compliant infrastructure [17].

Activity features were derived from three groups of parameters using smartwatches and BLE beacons: (1) activity recognition

XSL•FO
RenderX

(eg, sitting time or standing time), (2) indoor localization (eg, time in bed or time in the bathroom), and (3) raw acceleration quantification (mean absolute deviation [MAD] in accelerometer signal). By combining these attributes, we created features such as sitting time in bed and energy expenditure during walking or while in bed [12,13].

To ensure fair comparisons among patients with different watch wearing times, we normalized features by dividing the time spent on activities or in locations by the total wear time (uptime). Energy-related features were also normalized by uptime to yield energy intensity and by the total daily value to calculate energy percentage.

BLE beacons are used in the SARP system to estimate indoor patient locations by measuring the received signal strength indicator values via smartwatches [18-20]. Patients were instructed to place a BLE beacon in each designated indoor location: the kitchen, bathroom, bedroom, dining room, and television or sitting room. If the system does not detect any beacons, it infers that the patient is outside, indicating they are not at home.

Ethical Considerations

This study was approved by the UCLA Ethics Review Board (IRB #20 - 001565), ensuring adherence to ethical standards and the protection of human participants' rights and welfare. Participants were recruited through oral consent, which included a clear explanation of the study's purpose, duration, procedures, potential risks (such as discomfort from wearing a smartwatch or placing sensors at home), and confidentiality measures. Patients were also informed that the data collected could be used for future research studies, may be shared with other investigators, or used in collaboration with the private sector to develop smart algorithms, without requiring additional consent. Data collected were anonymized and deidentified before analysis, following strict confidentiality protocols. No financial compensation was offered to participants. Additionally, no identifiable images or data of participants are presented in this paper or supplementary materials.

Data Collection

For this study, SARP was modified to integrate a series of US Food and Drug Administration (FDA)–cleared, Bluetooth-enabled devices, including pulse oximeters, thermometers, and ultra-wideband radar technology for respiratory monitoring. The utility of these metrics for patient assessment has been demonstrated in numerous studies, with further research validating the accuracy of Bluetooth-enabled devices for home and remote monitoring [5,7,9,10,21-24]. While evaluating the efficacy of these devices is beyond the scope of this paper, we assume their reliability, given that pulse oximeters, thermometers, and respiratory monitoring devices incorporated into the SARP system are FDA-cleared. However, it is worth noting that some studies indicate ultra-wideband radar may overestimate respiratory rates at low levels and underestimate them at high levels [25,26]. While FDA clearance validates accuracy in controlled settings, real-world performance may vary due to user compliance, environmental factors, and device positioning. Our study did not independently validate these devices but relied on their FDA clearance as a baseline for integration into remote monitoring workflows.

Among the integrated devices, the smartwatch was designed for continuous monitoring, inferring metrics such as activity tracking, energy expenditure, and time spent in various indoor locations. In contrast, other devices were used on a spot-check basis, voluntarily recorded by patients. Patients were advised to use the pulse oximeter and thermometer at least twice daily, while the respiratory monitoring device was mounted next to the bed for continuous tracking whenever the patient was present.

All data were securely uploaded to a HIPAA-compliant Azure internet of things cloud infrastructure, expanding the SARP system's remote patient monitoring capabilities. The integration of Bluetooth-enabled devices allowed for the remote collection of additional vital data, complementing existing infrastructure. Specifically, for this project, pulse oximeters and thermometers were used to detect potential COVID-19 symptoms, such as shortness of breath and fever, aiding in testing recommendations. Additionally, SARP incorporated an FDA-cleared Bluetooth device equipped with radar, a microphone, and a light sensor to monitor respiratory function, enabling the detection of hypoventilation, hyperventilation, and impaired lung function. This device was recommended to be placed adjacent to the patient's bed, primarily for nighttime monitoring or when the patient was in the room.

The devices used included the AndesFit Non-Contact Forehead and Surface Thermometer with Bluetooth 4.0 (ADF-B38A), AndesFit Pulse Oximeter with Bluetooth 4.0, Circadia ultra-wideband Respiration Monitoring Device, TicWatch Pro 3 Smartwatch with Bluetooth module, Proximity Beacons (MCU ARM Cortex-M4 32-bit processor with floating-point unit), and Amazon Fire Tablet 7.

We recorded the day-to-day changes in the following sensor measurements and compiled the assessments shown in Textbox 1:

1. Position: lying down, sitting, standing, and walking inferred from continuous accelerometer data collection.
2. Active: active (walking), active (not walking), and nonactive inferred from continuous accelerometer data collection.
3. Steps: total daily steps is the total distance traveled inferred from continuous accelerometer data collection.
4. Indoor localization: bedroom, bathroom, kitchen, dining room, family room, and office inferred from accelerometer data and indoor location beacon received signal strength indicator scans during and immediately after each movement detected by the smartwatch.
5. Sleep quality: duration and toss and turn, inferred from continuous accelerometer.
6. Heart rate: spot check by patient using pulse oximeter.
7. Temperature with infrared technology spot check by the patient using a Bluetooth thermometer.
8. SpO_2 oxygen saturation spot check by the patient using pulse oximetry.

XSL•FO
RenderX

9. Breaths per minute measured using ultra-wideband radar technology, captured during sleep by a device positioned next to the patient's bed.
10. Respiration waveform (inhalation or exhalation patterns) with ultra-wideband radar technology.

Textbox 1. Patient assessments from sensor measurements.

Summary sensor assessments

- Percent of time out of bedroom
- Percent of time out of house
- Percent of time walking
- Total daily steps
- Walking speed
- Counts of activities and locations
- Total active time
- Total energy
- Weighted total activity score
- Spot check of temperature
- Spot check of SpO_2
- Spot check of average breaths per minute
- Respiratory rate and chest displacement
- Early warning score

As described in the study by Ramezani et al [12], active or nonactive is determined by an empirical threshold of 0.02 m/s^2 (2 cm/s^2) imposed on the MAD value of an accelerometer signal every 10 seconds. The threshold translates to 1 meter of displacement of the hand in 10 seconds (hand wearing the smartwatch with accelerometer). Furthermore, the energy expenditure is estimated to be proportional to the MAD of the accelerometer magnitude signal.

Exploratory Analysis

The primary objective of this study was to conduct an observational analysis aimed at developing predictive models for forecasting adverse COVID-19 outcomes, including hospitalization within 24 or 48 hours. Given the absence of adverse events among the participants, the research emphasis shifted toward an exploratory analysis, concentrating on the delineation of characteristics and behaviors of patients with COVID-19 as captured by the remote patient monitoring system.

All the analyses were performed using the Python Programming Language (version 3.11.3; Python Software Foundation) libraries, Pandas (version 2.2.1), NumPy (version 1.25.2), SciPy (version 1.11.1), Scikit-learn (version 1.3), and Seaborn library (version 0.12.2) were used for statistical data visualization.

Analytics Inclusion Criteria

To ensure the meaningfulness of the activity data, we imposed an additional inclusion criterion requiring a minimum daily wear time of the smartwatch of at least 4 hours. However, the use of the remote patient monitoring system kit was not limited solely to the smartwatch; it also included a thermometer and pulse oximeter for spot checks, along with a respiratory monitor capable of continuous data capture in the bedroom. Data collection was conducted either continuously or at specific intervals from these devices. For analysis, we included individuals with measurements taken on more than 7 distinct days and calculated the mean value of the metric for each day.

Activity and Energy

Energy intensity, as detailed in the SARP section, was analyzed for all individuals on their first day of the study (baseline) and their last day using a box plot. The plot displays the median line, representing the central tendency, and the overall distribution of energy intensity values. It is important to note that these observations were made after applying the inclusion criteria, which only considered days where patients had more than 4 hours of watch wearing time.

Using Bluetooth beacons, the time each patient spent in specified locations within their home was calculated to better form a daily storyline of their activity patterns. The goal was to assess whether patients spent less time in the bedroom as this study progressed and whether they spent more time outside their residence. The percentage of time spent in each location was calculated by dividing the duration in that location by the total uptime (watch wearing time) for that day. A heatmap was generated to visualize the average percentage of time spent in each location longitudinally over 2 weeks.

It is important to note that data were analyzed after applying inclusion criteria, and some participants were in this study for fewer than 2 weeks. For each day of this study, the time spent in each location was averaged across the number of patients observed on that particular day. Given that the start and end dates varied across patients, the analysis was conducted by calculating each patient's individual day. For example, "day 1" represents the average time percentage spent in each location

XSL•FO
RenderX

for all patients on their respective first day, which may differ chronologically between patients.

Vital Signs

The distributions of SpO_2, pulse rate, respiratory rate, and temperature were visualized using kernel density estimate (KDE) plots. The KDE plots depict patients' baseline (the first day of data collection) and their final day (the last available day of this study). For each patient, the mean value of observations was used for both the baseline and the final day. A minimum interval of 7 days between the baseline and final day was enforced to exclude patients with only a few days of data, as their differences could not accurately represent longitudinal trends. It is important to note that the KDE visualization may include out-of-bounds values due to the edge effect, where Gaussian distribution-based estimation extends beyond the actual data range. When the observations are close to the edges of the data range, the KDE, which is formed by centering around each data point, may extend beyond the boundary. This can be addressed by truncating the graph and clipping the x-axis out-of-bound values, although this is merely a visualization adjustment. Alternatively, the edge effect can be mitigated by smoothing the KDE curve through bandwidth adjustment. In the Seaborn library, the default bandwidth value is determined by Scott rule (set to 1) [27] and automatically adapts to the data characteristics [28]. While increasing the bandwidth reduces sensitivity to individual observations and produces a smoother curve, the authors opted not to adjust the bandwidth due to the limited number of observations and the need to maintain the graph's sensitivity to changes.

Self-Reported Survey

Daily self-reported surveys, collected using the SARP app provided to patients on tablets, were integrated with sensor data. The purpose of this integration was to later align the data with potential complications, including cardiovascular issues or hospitalizations, to retrospectively investigate whether any early indications could have been inferred from the observational data. This analysis aimed to identify potential early warning signs and develop timely interventions.

Results

Demographic Characteristics

Of 176 interested patients who were approached by our team, 73 met the eligibility criteria and were enrolled to use our remote patient monitoring kit, SARP, as detailed in Table 1. Of these, 41 patients successfully connected to and complied with the system.

XSL•FO
RenderX

Table . Sociodemographic characteristics of the patient cohort (N=73).

Characteristics	Patients, n (%)
Race	
Asian	3 (4.1)
Black or African American	6 (8.2)
Hawaiian or Other Pacific Islander	1 (1.4)
White	42 (57.5)
Other	21 (28.8)
Ethnicity	
Not Hispanic or Latino	44 (60.3)
Hispanic or Latino	26 (35.6)
Unknown or not reported	3 (4.1)
Sex	
Female	38 (52)
Male	31 (42.5)
Prefer not to say	4 (5.5)
Highest education level	
High school diploma	6 (8.2)
Some college	19 (26)
Bachelor's degree	28 (38.3)
Graduate degree	18 (24.7)
No response	2 (2.73)
Employment status	
Employed	44 (69.8)
Unemployed	10 (13.7)
Self-employed	8 (11)
Retired	4 (5.5)
Marital status	
Married	34 (46.6)
Divorced	2 (2.7)
Widowed	2 (2.7)
Single	29 (45.3)
Separated	2 (2.7)
Age (years)	
20 - 30	14 (19.2)
30 - 40	15 (20.5)
40 - 50	17 (23.3)
50 - 60	14 (19.2)
60 - 70	9 (12.3)
70 - 80	3 (4.1)
80 - 90	1 (1.4)
Weight (lbs)	
110 - 150	20 (27.3)
150 - 190	17 (23.3)

XSL•FO
RenderX

Characteristics	Patients, n (%)
190 - 230	24 (32.9)
230 - 270	9 (12.3)
270 - 310	2 (2.8)
310 - 360	1 (1.4)
Not reported	1 (1.4)
Height (feet)	
3 - 4	1 (1.4)
4 - 5	15 (20.5)
5 - 6	56 (76.7)
Not reported	1 (1.4)

To reiterate, the SARP remote patient monitoring system included a smartwatch, thermometer, pulse oximeter, and a respiratory rate monitoring device. We established an additional inclusion criterion that mandated a minimum daily wear time for the smartwatch of at least 4 hours.

There were 22 patients with substantial watch data (>4 h per day) with an average usage of 10.1 (SD 4.3) days of watch wearing time. A total of 39 patients used a pulse oximeter to check the SpO_2 and heart rate, with an average usage of 10.1 (SD 4.7) days of using the device. The number of patients who used the thermometer decreased to 37 with an average usage period of 9.7 (SD 4.8) days. The number of patients with respiratory data is 36, with an average of 11.4 (SD 4.7) nights of continuous monitoring of breathing rate per minute.

Activity and Energy

The box plot in Figure 1 shows that the overall energy levels of patients increased on the last day of this study compared to the first day. Both the mean intensity and the variability of energy levels showed an upward trend. The figure indicates an overall improvement in the physical activity of patients with COVID-19 over time, though with varying degrees of recovery across the patients.

Figure 1. Average energy intensity: first day versus last day.

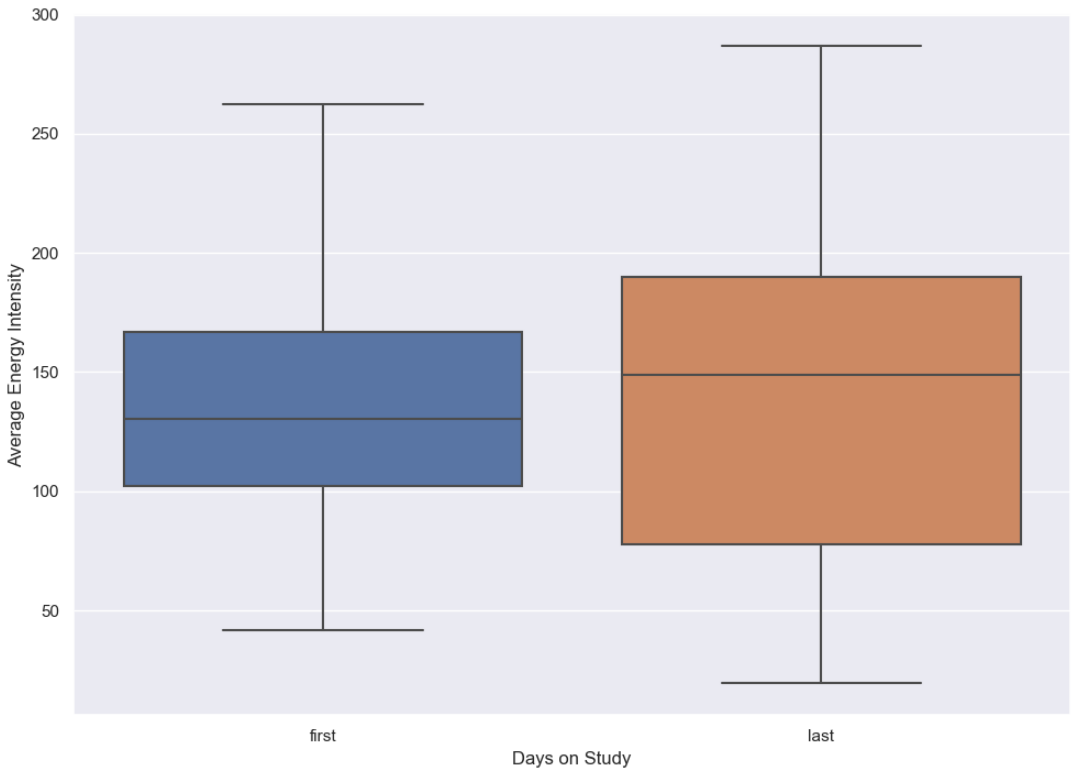

XSL•FO
RenderX

By analyzing the heatmap of patient activities shown in Figure 2, it is evident that toward the end of this study, patients were expending less time in the bedroom and more in other areas outside the home. The fading color in the bedroom area toward the end of this study suggests a decrease in time spent in that location. Similarly, there is a noticeable decline in activity within other home settings, such as the television or sitting area, as captured by indoor localization beacons. This trend likely reflects patients' gradual recovery and increased mobility, allowing them to engage more in activities outside their homes.

Figure 2. Heatmap of patients' activities during the 2-week study.

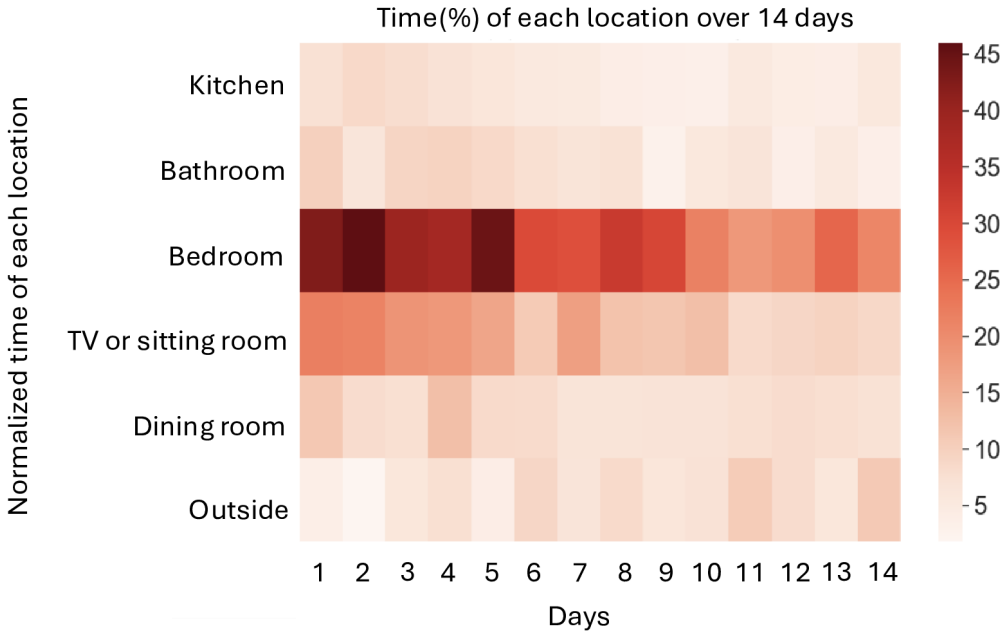

Vital Signs

In Figure 3, KDE graphs for 4 vital signs show that the distribution for SpO_2 has shifted slightly to the right from baseline to the last day, indicating an overall improvement in oxygen saturation levels by the end of this study. The value of the last day's SpO_2 distribution is higher than the baseline, suggesting more patients reached higher SpO_2 levels on the last day.

XSL•FO
RenderX

Figure 3. Biomarkers: first day versus last day.

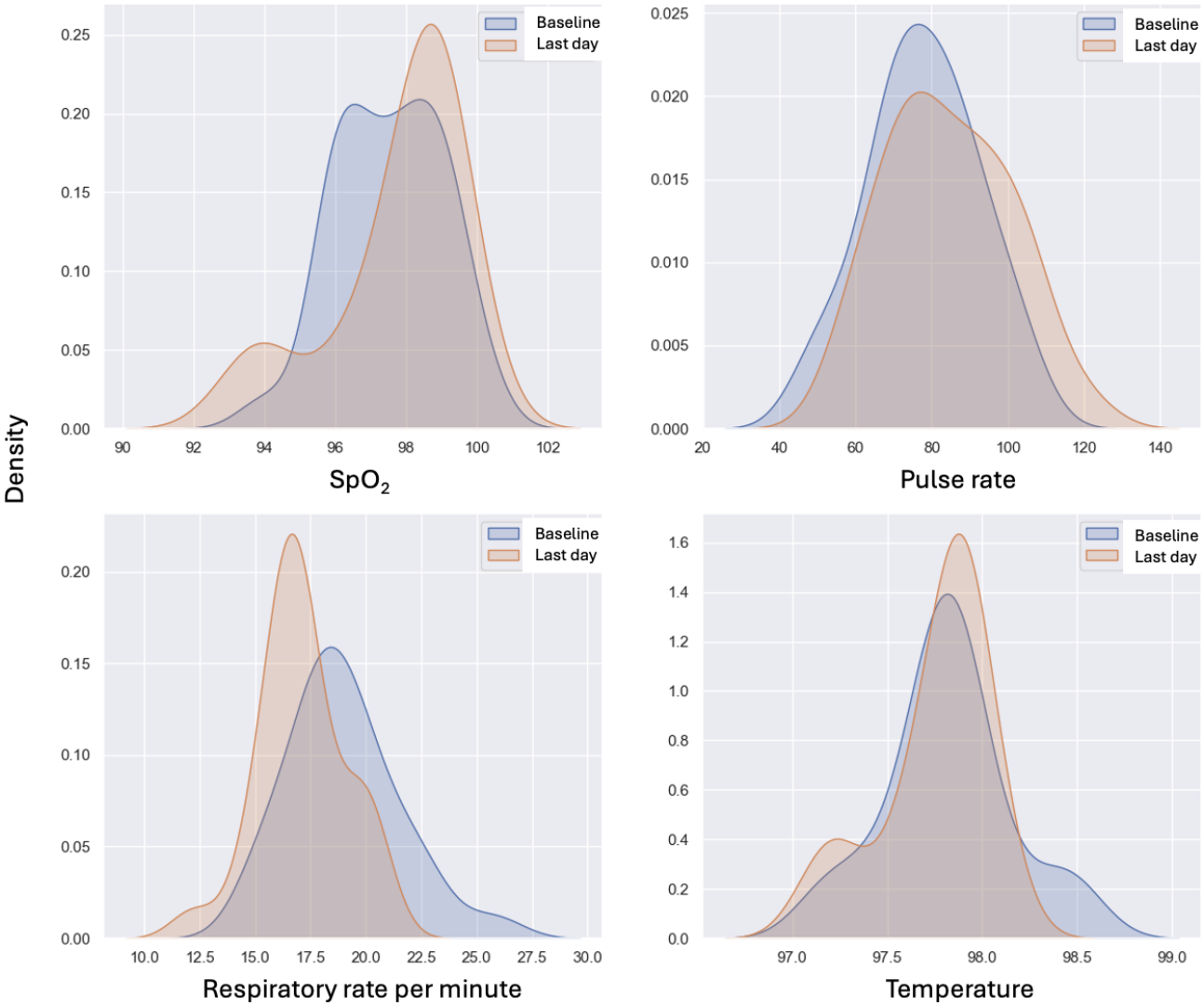

The pulse rate distribution on the last day shows a slight shift to the right compared to the baseline. This suggests that patients' pulse rates tended to increase slightly by the end of this study. The distribution also appears to broaden slightly, indicating greater variability in pulse rates among patients on the last day.

On the respiratory rate subgraph, there is a noticeable shift in distribution toward lower rates on the last day compared to the baseline. This may indicate that, on average, patients had a lower respiratory rate by the end of this study, potentially reflecting improved respiratory function. The last day's distribution is narrower, suggesting reduced variability in respiratory rates, with most patients converging around healthier respiratory measures.

The temperature distributions, however, remain relatively similar in both baseline and the last day, with only slight differences.

Self-Reported Survey

The tablet provided to each patient, serving as a device hub, was also available for completing protocol-mandated surveys. Participants could voluntarily report their symptoms and overall wellness at least once daily. The survey responses are included in Multimedia Appendix 1, which illustrates the prevalence of symptoms reported by patients at least once during the 2-week observation period.

Multimedia Appendix 1 indicates that fatigue (22/75, 30%) and headache (22/75, 30%) were the most prevalent symptoms among patients with COVID-19, followed by cough (19/75, 25%) and loss of smell or taste (20/75, 26%). Despite these symptoms, only 6% (5/75) of patients contacted a health care provider, suggesting either mild severity or potential difficulties in seeking medical advice. As there was a direct contact number for the patients in this study and the ease of contacting them, it would be safe to assume that mild severity was the main reason. This underscores the need for careful monitoring of common COVID-19 symptoms. Other symptoms were chest pain (12/75, 16%), shortness of breath (10/75, 13%), fever (6/75, 8%), runny nose (17/75, 22%), diarrhea (10/75, 13%), and body aches (15/75, 20%).

Discussion

Principal Findings

Our analysis of activity and indoor location suggests that a positive trend in data from wearable and remote monitoring devices is aligned with the recovery of patients with COVID-19

XSL•FO
RenderX

over this study's period. It is important to emphasize that the lower numbers in each category (smartwatch, pulse oximeter, thermometer, and respiratory monitoring) compared to the 73 participants result from not meeting the analytics inclusion criteria. This includes insufficient compliance with smartwatch wear time, voluntary use of Bluetooth-enabled devices for spot checking, and willingness to use the respiratory monitoring device next to their bed.

Although we did not directly ask patients about their compliance or the reasons for not using the system more frequently in this study, insights from our previous studies using the same infrastructure [12]—albeit with a different cohort (at-risk older population)—suggest that the bulkiness of smartwatches was a primary issue. Most commercially available smartwatches, despite recent advancements in sensor quality approaching medical-grade standards, remain relatively bulky due to their large battery packs. This design choice enables them to perform various functions, including connectivity with phones, but makes them less suitable for medical use, where comfort and continuous wearability are crucial. In short, most smartwatches are not specifically designed for patient use, and the current battery technology often results in devices that are bulky and somewhat inconvenient for continuous wear. In contrast, other sensors, such as temperature monitors and pulse oximeters, were used more frequently, as they were used for spot checks rather than requiring continuous wear, making them more acceptable to patients.

Figure 1 shows an increase in overall energy levels, with both the mean energy intensity and variability rising from the first to the last day, indicating an overall improvement in physical activity. Figure 2 offers additional insights, revealing that patients progressively expended less energy in the bedroom and other indoor locations, such as the television or sitting area, as this study progressed. This decline in indoor activity suggests that patients were recovering and becoming more mobile, with greater energy expenditure in locations outside their homes. The shift in activity distribution likely reflects improved physical health, as indicated by the higher intensity color in the heatmap corresponding to outdoor locations toward the end of this study.

Our study results are consistent with a systemic scoping review of studies published between 2019 and 2022, in which smartwatches or fitness trackers were reported as the preferred type of wearable technologies for early and presymptomatic detection [29-32]. Much of the published literature is focused on using wearables for early detection [33,34]. These studies do show that wearables can track heart rate and heart rate variability with improvement over 7 days [35]. Moreover, SpO_2 and activity tracking from wearable devices might be able to help identify patients with COVID-19 at risk for sudden death [36] based on a systematic review. However, the best results are looking at physiological features in a multimodal approach where one can achieve sensitivity as high as 90% and specificity as high as 80% [37]. Our study asked about self-reported symptoms and the importance of combining symptoms with sensor data that has been shown to allow for better predictive models [38]. The use of wearables and remote monitoring has also been explored in managed long COVID-19 [39].

The prevalence of fever among patients with COVID-19 varies across studies and cohorts, with a high percentage observed in hospitalized patients [40]. However, in nonhospitalized patients, fever is less frequently observed or may be entirely absent. In studies such as the one by Shi et al [41], fever was recorded for fewer than 2 days throughout the illness, primarily in frail patients, even after lowering the temperature threshold from 100.4°F to 100°F. Body temperature in our cohort never exceeded 98.6°F. Figure 3 indicates that patients' body temperatures were relatively stable throughout this study, with a marginal increase by the last day. Both distributions of body temperatures in baselines and the last day are tightly clustered, indicating consistent body temperatures across patients. This suggests that either the cohort did not exhibit fever, or the noncontact infrared thermometers used were not sufficiently sensitive, despite being FDA-cleared. This aligns with an FDA study [42] highlighting misleading labeling of FDA-cleared devices during COVID-19, showing that noncontact infrared thermometers may fail to reliably detect fevers when used on adults and may not meet the accuracy specifications advertised in manufacturers' instructions for use. Notably, 8% of patients in this study reported fever in the self-reported surveys (Multimedia Appendix 1), indicating that the subjective experience of fever was prevalent in at least 8% of the cohort.

Figure 3 of vital signs overall suggests that by the end of this study, patients generally showed signs of physiological improvement, with higher SpO_2 levels, slightly elevated pulse rates, and lower respiratory rates. Such trends indicate a recovery or improvement in the monitored patients' vital signs.

Study Limitations

Patient compliance with wearing a smartwatch and using the Bluetooth-enabled thermometer and pulse oximeters was one of the main challenges of this study, and we anticipate this to be a common obstacle in similar studies that aim to use wearable technology for patient monitoring. In contrast, the ultra-wideband respiration monitoring device required no patient interaction and could monitor respiration passively. However, the accuracy of this device could be compromised if more than 1 person were in the bedroom or if the patient left the room. To mitigate this, we ensured that only the results from patients who lived alone in the monitored bedroom were included in the analysis. While setting up the devices was designed to be straightforward—requiring only a tablet and a simple app we provided to connect all devices to the patients' home internet in 1 step—it is reasonable to assume that certain patients, particularly those less familiar with technology, may have found the setup process challenging.

As the cohort was randomly selected, the predominance of younger participants indicated in Table 1 may reflect self-selection bias, accessibility to technology, or willingness to participate in remote monitoring studies rather than a deliberate sampling choice. However, this demographic skew could impact the relevance and generalizability of the findings, particularly if younger individuals have different health profiles, technology engagement levels, or adherence behaviors compared to older populations. To address this limitation, we acknowledge that further studies with more diverse age representation are

XSL•FO
RenderX

needed to validate the findings across broader demographics, particularly in populations more likely to benefit from remote health monitoring, such as older adults or individuals with chronic conditions.

Conclusions

The COVID-19 pandemic demonstrated the importance of telehealth and remote monitoring of at-risk patients. A recent cost-utility analysis estimates that daily pulse oximetry use with a follow-up after 3 weeks could reduce the mortality rate to 6 per 1000 patients, compared to 26 per 1000 without at-home monitoring. Various studies suggest that remote patient monitoring in patients with COVID-19 could potentially reduce hospitalizations and deaths by as much as 80% and yield cost savings of around $12,000 per patient [10,43]. This underscores the growing importance of using technology for continuous at-home monitoring, which can enhance the quality of care while significantly reducing costs.

In this study, we demonstrated that affordable remote monitoring devices can effectively track individuals over time, revealing trends in their health status. However, a significant challenge in remote and home monitoring systems is ensuring participant adherence. Our study found that only 41 of 73 individuals consistently used all the provided devices. This finding aligns with previous research, which suggests that patients are more likely to engage in remote monitoring or mHealth when they perceive an immediate personal benefit. Based on anecdotal feedback, patients appeared to prefer passive sensing over more demanding active monitoring. However, no formal assessment of patient satisfaction was conducted.

This study also generated novel data specific to COVID-19, comparing self-reported symptoms with objective monitoring data. Additionally, the findings highlight the potential benefits of implementing remote patient monitoring systems in future pandemics, as they provide continuous, scalable, and contactless health tracking, which could aid in early detection, triage, and disease management while minimizing the burden on health care facilities.

Acknowledgments

The authors would like to express their gratitude to the researchers cited here and apologize to those whose work, because of page restrictions, could not be mentioned. This research was mainly funded by the W M Keck Foundation COVID-19 research fund at the University of California, Los Angeles (UCLA). No generative artificial intelligence tool was used in the ideation or writing of this text.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to University of California, Los Angeles (UCLA) policies governing the transfer or sharing of health or research data from the electronic medical record with third parties outside the university, which require prior approval from the UCLA Health Data Oversight Committee (HDOC), but are available from the corresponding author on reasonable request.

Conflicts of Interest

The Sensing At-Risk Populations (SARP) system is protected by a patent (US Patent 10937547) [15] in which RR and AN are listed as coinventors. RR and AN are cofounders of InvistaHealth LLC. Other authors have declared no potential conflict of interest regarding the publication of this paper.

Multimedia Appendix 1
Prevalence of symptoms and health care contact among patients.
[PNG File, 146 KB - formative_v9i1e66773_app1.png]

References

1. The true death toll of COVID-19 estimating global excess mortality. World Health Organization. 2024. URL: https://www.who.int/data/stories/the-true-death-toll-of-covid-19-estimating-global-excess-mortality [accessed 2025-06-19]
2. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis 2020 May;20(5):533-534. [doi: 10.1016/S1473-3099(20)30120-1] [Medline: 32087114]
3. Rapid COVID-19 testing underway for all patient admittance to peacehealth hospitals. KLCC. 2020. URL: https://www.klcc.org/health-medicine/2020-04-21/rapid-covid-19-testing-underway-for-all-patient-admittance-to-peacehealth-hospitals [accessed 2025-06-19]
4. Greenhalgh T, Knight M, Inda-Kim M, Fulop NJ, Leach J, Vindrola-Padros C. Remote management of COVID-19 using home pulse oximetry and virtual ward support. BMJ 2021 Mar 25;372:n677. [doi: 10.1136/bmj.n677] [Medline: 33766809]
5. McGrath SP, McGovern KM, Perreard IM, Huang V, Moss LB, Blike GT. Inpatient respiratory arrest associated with sedative and analgesic medications: impact of continuous monitoring on patient mortality and severe morbidity. J Patient Saf 2021 Dec 1;17(8):557-561. [doi: 10.1097/PTS.0000000000000696] [Medline: 32175965]

XSL•FO
RenderX

6. Seshadri DR, Davies EV, Harlow ER, et al. Wearable sensors for COVID-19: a call to action to harness our digital infrastructure for remote patient monitoring and virtual assessments. Front Digit Health 2020;2:8 [FREE Full text] [doi: 10.3389/fdgth.2020.00008] [Medline: 34713021]
7. Shah S, Majmudar K, Stein A, et al. Novel use of home pulse oximetry monitoring in COVID-19 patients discharged from the emergency department identifies need for hospitalization. Acad Emerg Med 2020 Aug;27(8):681-692. [doi: 10.1111/acem.14053] [Medline: 32779828]
8. Crotty BH, Dong Y, Laud P, et al. Hospitalization outcomes among patients with COVID-19 undergoing remote monitoring. JAMA Netw Open 2022 Jul 1;5(7):e2221050. [doi: 10.1001/jamanetworkopen.2022.21050] [Medline: 35797044]
9. Wurzer D, Spielhagen P, Siegmann A, et al. Remote monitoring of COVID-19 positive high-risk patients in domestic isolation: a feasibility study. PLoS One 2021;16(9):e0257095. [doi: 10.1371/journal.pone.0257095] [Medline: 34559832]
10. Pronovost PJ, Cole MD, Hughes RM. Remote patient monitoring during COVID-19: an unexpected patient safety benefit. JAMA 2022 Mar 22;327(12):1125-1126. [doi: 10.1001/jama.2022.2040] [Medline: 35212725]
11. Annis T, Pleasants S, Hultman G, et al. Rapid implementation of a COVID-19 remote patient monitoring program. J Am Med Inform Assoc 2020 Aug 1;27(8):1326-1330. [doi: 10.1093/jamia/ocaa097] [Medline: 32392280]
12. Ramezani R, Zhang W, Xie Z, et al. A combination of indoor localization and wearable sensor-based physical activity recognition to assess older patients undergoing subacute rehabilitation: baseline study results. JMIR mHealth uHealth 2019 Jul 10;7(7):e14090. [doi: 10.2196/14090] [Medline: 31293244]
13. Moatamed B, Shahmohammadi F, Ramezani R, Naeim A, Sarrafzadeh M. Low-cost indoor health monitoring system. Presented at: 2016 IEEE 13th International Conference on Wearable and Implantable Body Sensor Networks (BSN); Jun 14-17, 2016; San Francisco, CA, United States URL: https://ieeexplore.ieee.org/document/7516252 [accessed 2025-06-19] [doi: 10.1109/BSN.2016.7516252]
14. Wong K, Shen J, Ramezani R, et al. A pilot study of a wearable monitoring system as an adjunct to geriatric assessment in older adults with cancer. JCO 2020 May 20;38(15_suppl):2062-2062 [FREE Full text] [doi: 10.1200/JCO.2020.38.15_suppl.2062] [Medline: 32874355]
15. Ramezani R, Moatamed B, Naeim A, Sarrafzadeh M. US10937547B2 - subject assessment using localization, activity recognition and a smart questionnaire. Google Patents. 2021 Mar 2. URL: https://patents.google.com/patent/US10937547B2/en [accessed 2025-06-19]
16. Mlinac ME, Feng MC. Assessment of activities of daily living, self-care, and independence. Arch Clin Neuropsychol 2016 Sep;31(6):506-516 [FREE Full text] [doi: 10.1093/arclin/acw049] [Medline: 27475282]
17. Ramezani R, Cao M, Earthperson A, Naeim A. Developing a smartwatch-based healthcare application: notes to consider. Sensors (Basel) 2023 Jul 25;23(15):6652 [FREE Full text] [doi: 10.3390/s23156652] [Medline: 37571436]
18. Bouchard K, Ramezani R, Naeim A. Features based proximity localization with bluetooth emitters. Presented at: 2016 IEEE 7th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON); Oct 20-22, 2016; New York City, NY, United States URL: https://ieeexplore.ieee.org/document/7777845 [accessed 2025-06-19] [doi: 10.1109/UEMCON.2016.7777845]
19. Bouchard K, Ramezani R, Arjun A, Naeim A. Evaluation of Bluetooth beacons behavior. Presented at: 2016 IEEE 7th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON); Oct 20-22, 2016; New York, NY, United States URL: https://ieeexplore.ieee.org/document/7777846 [accessed 2025-06-19] [doi: 10.1109/UEMCON.2016.7777846]
20. Bouchard K, Eusufzai MR, Ramezani R, Naeim A. Generalizable spatial feature for human positioning based on Bluetooth beacons. Presented at: 2016 IEEE 7th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON); Oct 20-22, 2016; New York City, NY, United States URL: https://ieeexplore.ieee.org/document/7777884 [accessed 2025-06-19] [doi: 10.1109/UEMCON.2016.7777884]
21. Ishikawa M, Sakamoto A. Patient SafetyNet for the evaluation of postoperative respiratory status by nurses: a presurvey and postsurvey study. J Perianesth Nurs 2021 Feb;36(1):14-17 [FREE Full text] [doi: 10.1016/j.jopan.2020.03.005] [Medline: 32978050]
22. Nicolò A, Massaroni C, Schena E, Sacchetti M. The importance of respiratory rate monitoring: from healthcare to sport and exercise. Sensors (Basel) 2020 Nov 9;20(21):6396. [doi: 10.3390/s20216396] [Medline: 33182463]
23. McFadden JP, Price RC, Eastwood HD, Briggs RS. Raised respiratory rate in elderly patients: a valuable physical sign. Br Med J (Clin Res Ed) 1982 Feb 27;284(6316):626-627. [doi: 10.1136/bmj.284.6316.626] [Medline: 6802262]
24. Goldfine CE, Oshim MFT, Chapman BP, Ganesan D, Rahman T, Carreiro SP. Contactless monitoring system versus gold standard for respiratory rate monitoring in emergency department patients: pilot comparison study. JMIR Form Res 2024 Feb 16;8:e44717 [FREE Full text] [doi: 10.2196/44717] [Medline: 38363588]
25. Lazaro A, Girbau D, Villarino R. Analysis of vital signs monitoring using an IR-UWB radar. PIER 2010;100:265-284 [FREE Full text] [doi: 10.2528/PIER09120302]
26. Immoreev I, Tao TH. UWB radar for patient monitoring. IEEE Aerosp Electron Syst Mag 2008;23(11):11-18 [FREE Full text] [doi: 10.1109/MAES.2008.4693985]

XSL•FO
RenderX

27. Scott DW. Multivariate density estimation and visualization. In: Gentle JE, Härdle WK, Mori Y, editors. Handbook of Computational Statistics: Springer; 2012:549-569 URL: http://link.springer.com/10.1007/978-3-642-21551-3_19 [accessed 2025-06-19]
28. Waskom M. seaborn: statistical data visualization. JOSS 2021;6(60):3021 [FREE Full text] [doi: 10.21105/joss.03021]
29. Cheong SHR, Ng YJX, Lau Y, Lau ST. Wearable technology for early detection of COVID-19: a systematic scoping review. Prev Med 2022 Sep;162:107170 [FREE Full text] [doi: 10.1016/j.ypmed.2022.107170] [Medline: 35878707]
30. Mishra T, Wang M, Metwally AA, et al. Pre-symptomatic detection of COVID-19 from smartwatch data. Nat Biomed Eng 2020 Dec;4(12):1208-1220. [doi: 10.1038/s41551-020-00640-6] [Medline: 33208926]
31. Mirjalali S, Peng S, Fang Z, Wang CH, Wu S. Wearable sensors for remote health monitoring: potential applications for early diagnosis of COVID-19. Adv Mater Technol 2022 Jan;7(1):2100545. [doi: 10.1002/admt.202100545] [Medline: 34901382]
32. Channa A, Popescu N, Skibinska J, Burget R. The rise of wearable devices during the COVID-19 pandemic: a systematic review. Sensors (Basel) 2021 Aug 28;21(17):5787 [FREE Full text] [doi: 10.3390/s21175787] [Medline: 34502679]
33. Cleary JL, Fang Y, Sen S, Wu Z. A caveat to using wearable sensor data for COVID-19 detection: the role of behavioral change after receipt of test results. PLOS ONE 2022;17(12):e0277350 [FREE Full text] [doi: 10.1371/journal.pone.0277350] [Medline: 36584148]
34. Sanches CA, Silva GA, Librantz AFH, Sampaio LMM, Belan PA. Wearable devices to diagnose and monitor the progression of COVID-19 through heart rate variability measurement: systematic review and meta-analysis. J Med Internet Res 2023 Nov 14;25:e47112 [FREE Full text] [doi: 10.2196/47112] [Medline: 37820372]
35. Erratum to “COVID-19 surveillance based on consumer wearable devices”. Digit Health 2024;10:20552076241261219 [FREE Full text] [doi: 10.1177/20552076241261219] [Medline: 39099679]
36. Takahashi S, Nakazawa E, Ichinohe S, Akabayashi A, Akabayashi A. Wearable technology for monitoring respiratory rate and SpO2 of COVID-19 patients: a systematic review. Diagnostics (Basel) 2022 Oct 21;12(10):2563 [FREE Full text] [doi: 10.3390/diagnostics12102563] [Medline: 36292252]
37. Mason AE, Hecht FM, Davis SK, et al. Author correction: detection of COVID-19 using multimodal data from a wearable device: results from the first TemPredict study. Sci Rep 2022 Mar 16;12(1):4568 [FREE Full text] [doi: 10.1038/s41598-022-08723-x] [Medline: 35296773]
38. Quer G, Radin JM, Gadaleta M, et al. Wearable sensor data and self-reported symptoms for COVID-19 detection. Nat Med 2021 Jan;27(1):73-77 [FREE Full text] [doi: 10.1038/s41591-020-1123-x] [Medline: 33122860]
39. Khondakar KR, Kaushik A. Role of wearable sensing technology to manage long COVID. Biosensors (Basel) 2022 Dec 31;13(1):62. [doi: 10.3390/bios13010062] [Medline: 36671900]
40. Islam MA, Kundu S, Alam SS, Hossan T, Kamal MA, Hassan R. Prevalence and characteristics of fever in adult and paediatric patients with coronavirus disease 2019 (COVID-19): a systematic review and meta-analysis of 17515 patients. PLOS ONE 2021;16(4):e0249788 [FREE Full text] [doi: 10.1371/journal.pone.0249788] [Medline: 33822812]
41. Shi SM, Bakaev I, Chen H, Travison TG, Berry SD. Risk factors, presentation, and course of coronavirus disease 2019 in a large, academic long-term care facility. J Am Med Dir Assoc 2020 Oct;21(10):1378-1383 [FREE Full text] [doi: 10.1016/j.jamda.2020.08.027] [Medline: 32981664]
42. Sullivan SJL, Rinaldi JE, Hariharan P, et al. Clinical evaluation of non-contact infrared thermometers. Sci Rep 2021 Nov 11;11(1):22079 [FREE Full text] [doi: 10.1038/s41598-021-99300-1] [Medline: 34764438]
43. Iuliano AD, Chang HH, Patel NN, et al. Estimating under-recognized COVID-19 deaths, United States, March 2020-May 2021 using an excess mortality modelling approach. Lancet Reg Health Am 2021 Sep;1:100019. [doi: 10.1016/j.lana.2021.100019] [Medline: 34386789]

Abbreviations

BLE: Bluetooth Low Energy
FDA: US Food and Drug Administration
HIPAA: Health Insurance Portability and Accountability Act
KDE: kernel density estimation
MAD: mean absolute deviation
SARP: Sensing At-Risk Populations
UCLA: University of California, Los Angeles

XSL•FO
RenderX

Edited by A Mavragani; submitted 23.09.24; peer-reviewed by M Donati, M Wilkes; revised version received 11.04.25; accepted 13.04.25; published 18.07.25.

Please cite as:
Ramezani R, Zhang W, Cao M, Bui A, Petruse A, Weldon A, Naeim A
Remote Patient Monitoring for Global Emergencies: Case Study in Patients With COVID-19
JMIR Form Res 2025;9:e66773
URL: https://formative.jmir.org/2025/1/e66773
doi:10.2196/66773

© Ramin Ramezani, Wenhao Zhang, Minh Cao, Alex Bui, Antonia Petruse, Amelia Weldon, Arash Naeim. Originally published in JMIR Formative Research (https://formative.jmir.org), 18.7.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

A Digital Lifestyle App for Hypertension During Pregnancy: Mixed Methods Intervention Development Study Using the Person-Based Approach

Lucy Goddard[1], DPhil; Katherine Tucker[1], PhD; Nerys M Astbury[1], PhD; Cristian Roman[2], PhD; Yuan Chi[2], PhD; Katherine Morgan[3]; Patricia Devitt[3]; Richard J McManus[1,4], PhD

[1]Nuffield Department of Primary Care Health Sciences, University of Oxford, Radcliffe Observatory Quarter, Woodstock Road, Oxford, United Kingdom

[2]NIHR Oxford Biomedical Research Centre, University of Oxford, Oxford, United Kingdom

[3]Patient and Public Contributor, United Kingdom

[4]Brighton and Sussex Medical School, University of Brighton and University of Sussex, Brighton, United Kingdom

Corresponding Author:
Lucy Goddard, DPhil
Nuffield Department of Primary Care Health Sciences, University of Oxford, Radcliffe Observatory Quarter, Woodstock Road, Oxford, United Kingdom

Abstract

Background: Chronic hypertension affects 1% - 5% of pregnancies, increasing women's risks of adverse pregnancy outcomes and life-long cardiovascular disease risk. Therefore, care management during pregnancy includes close monitoring of blood pressure and medication. Healthy dietary and physical activity behaviors have proven beneficial effects on blood pressure outside and during pregnancy. However, little is known about the best way to support women with chronic hypertension during pregnancy to adopt such behaviors, which could improve pregnancy outcomes, as well as future cardiovascular health.

Objective: This study aims to develop and optimize a digital lifestyle intervention—the DAPHNY (Diet and Activity for Pregnancy Hypertension) app—with those who have experienced chronic hypertension during pregnancy.

Methods: Guided by the person-based approach to intervention development, a review of literature and continuous expert input, including from patient and public representatives, informed the planning stage. This was followed by focus groups with maternity health professionals (n=23) and think-aloud interviews with women who had experience of chronic hypertension during pregnancy (n=11). A content analysis, underpinned by theoretical modeling using the capability opportunity motivation-behavior model, informed 3 logic models to visualize modifications for meaningful engagement with an intervention and sustained behavior change. The intervention was modified iteratively, leading to a first version of the digital intervention that was tested by women (n=10) to further optimize acceptability and engagement. App use data and user engagement patterns were captured.

Results: An evidence-based, theoretically informed lifestyle app, named DAPHNY, was developed. Key features included in logic models and implemented into a first version of the app comprised supportive messaging to acknowledge challenges of hypertensive pregnancy, goal setting and progress reports for feedback on behaviors, information about health consequences to shape knowledge, credible source endorsement, and a reward or recognition system to acknowledge effort had been made. Engagement with the DAPHNY app during user testing demonstrated variability across users, with a mean of 13 (SD 6.84) sessions per participant. Session duration was variable, with a median of 36 seconds (range: 5 seconds to 5 minutes, 20 seconds). Action-based pages, including recording blood pressure (40 sessions) and step count (39 sessions), were accessed more frequently than informational pages, which required a deeper level of app engagement.

Conclusions: Development of the DAPHNY app, underpinned by an established behavioral framework for developing digital interventions, provided new data insights about how to support women with chronic hypertension to engage in healthy behaviors, a currently overlooked aspect of blood pressure management. Future iterations should focus on increasing engagement and supporting implementation through streamlined content and integration with existing health systems and self-monitoring data. Rigorous, larger-scale studies including comprehensive process evaluation would determine potential clinical effectiveness, implementation strategies, and impact for women and health care professionals.

(JMIR Form Res 2025;9:e68927) doi:10.2196/68927

KEYWORDS

pregnancy; hypertension; lifestyle; behavior change; person-based approach; intervention development; mobile phone

XSL•FO
RenderX

Introduction

Supporting women with pregnancies that are complicated by hypertension to adopt healthy behaviors could play an important role in blood pressure control during pregnancy and beyond [1]. However, an understanding of how best to support women with existing, chronic hypertension to manage their blood pressure and improve future cardiovascular health through targeted lifestyle modification during pregnancy is unknown [2,3].

Pregnant women with chronic hypertension are at increased risk of pregnancy and birth complications, including pre-eclampsia, which occurs in a quarter of women with chronic hypertension [4]. They have approximately double the risk of major adverse cardiovascular events in later life compared to women who do not have any hypertension during pregnancy [5]. The prevalence of chronic hypertension in pregnancy is increasing, likely due to rising maternal age and obesity rates in the population, which further increase pre-eclampsia risk [6-9]. This indicates a public health concern requiring attention [10].

Lifestyle behaviors, including diet and physical activity, play a vital role in preventing and managing cardiovascular disease and are well-evidenced to provide sustained, clinically relevant reductions in blood pressure, outside of pregnancy [1]. Weight loss is a key recommendation for hypertension management outside of pregnancy, but as weight loss is not recommended during pregnancy, attenuation of excessive gestational weight gain is the focus of lifestyle interventions to improve pregnancy and birth outcomes [11]. Syntheses of study-level data have produced mixed results, but there appears to be some, albeit low-quality, evidence that interventions may have a protective effect on the incidence of pre-eclampsia, hypertension, gestational diabetes, and preterm delivery [12-15]. A recent systematic review identified no studies that specifically developed and tested lifestyle interventions to support healthy lifestyle behaviors for women who have chronic hypertension, who may benefit the most from lifestyle modification [3]. Since this is an important population at risk of poor pregnancy experiences and outcomes, a clear gap in the literature exists, which this study aims to start addressing.

Delivering health promotion messaging in a way that has a meaningful impact on patient outcomes is a challenge within current National Health Service (NHS) models of care and system constraints. However, a shift toward prevention and proactive care, including the prevention of cardiovascular events, is an NHS priority [16]. Alongside this, improving digital systems to improve population health and support health professionals to deliver unbiased information is at the forefront of maternity service delivery, with its potential to increase women's choice, women's engagement with services, and facilitate existing care pathways [17].

App use in pregnancy is prolific, but many freely available apps lack robust scientific input, lack regulation of content, screening for comorbidities, health care professional involvement, and theoretical underpinning [18,19]. This is especially relevant for lifestyle information, which can be sought from alternative, potentially less reputable sources [20,21]. Encouragingly, remote monitoring apps to manage hypertensive pregnancy have been developed and robustly trialed on large populations and rolled out widely and found to be acceptable and feasible for women [22-25]. An intervention to manage chronic hypertension during pregnancy with embedded lifestyle support to meaningfully change behavior, that is collaboratively developed with women who have long-term hypertension, has not been scientifically explored and could be an important avenue to improving pregnancy, birth, and future cardiovascular health.

This intervention development study aimed to design, develop, and test a digital lifestyle intervention with women who have long-term hypertension and have been pregnant, and health care professionals.

This paper is reported following the GUIDED (Guidance for the Reporting of Intervention Development) checklist (Checklist 1) [26].

Methods

Design

This intervention development study incorporated two phases of the person-based approach (PBA): (1) planning phase and (2) optimization phase. The PBA is frequently used in the development of digital interventions and incorporates user design methods alongside evidence-based behavior change methods, aiming to ensure that interventions address the needs of the target population, increasing their likelihood of successful roll-out [27,28]. There was fluidity between the two phases. A visual demonstration of this dynamic approach is displayed in Figure 1.

XSL•FO
RenderX

Figure 1. Planning and optimization phases of the person-based approach to develop the DAPHNY (Diet and Activity for Pregnancy Hypertension) intervention. PPI: patient and public involvement.

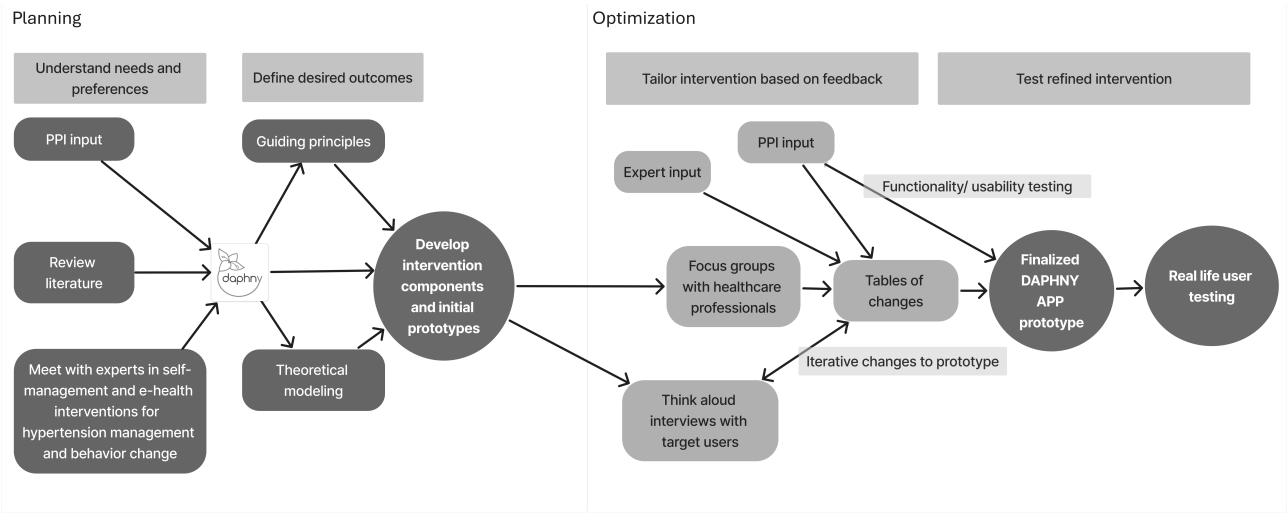

Ethical Considerations

Ethical approval was gained from Health and Care Research Wales (REC ref: 22/WA/0130), and written informed consent was obtained from all participants. Participant data were deidentified as soon as it was practical to do so through assigning unique participant IDs. All women who took part in the study received a £10 (US $13.77) shopping voucher to thank them for their time and participation.

Planning Phase

Patient and Public Involvement

Patient and public involvement (PPI) is vital to promote the interests of patients throughout the research cycle to shape more acceptable and sustainable study methodologies and interventions [29-31]. A group of 12 women with experience of hypertension during pregnancy met in groups or individually, providing an informal space to discuss current management of hypertension and ideas around what would support healthy behaviors during pregnancy, if this was something they were seeking to do. Recurring topics of discussion were that lifestyle advice was rarely discussed in relation to blood pressure, with some unaware that changes in diet and physical activity could reduce blood pressure. The potential to reduce medication if blood pressure improved was a motivating factor to engage in healthy behaviors for some, due to side effects acting as a barrier to being active, and some shared concerns around the impact on their baby.

Two PPI representatives joined the study team and had continuous involvement throughout to contribute ideas, support the study design and procedures, review study materials, and share perspectives on the developing intervention. This valuable insight contributed to optimizing the conduct of the study and the developed intervention.

Scoping the Literature

Exploring the literature was a necessarily rapid process. An overview of the foundation of evidence on which the study was built is summarized in Multimedia Appendix 1. This also serves to provide a rationale for the app prototypes that were further refined and iteratively developed.

Intervention Guiding Principles

Guiding principles summarize the intervention objectives [28]. Three key design objectives and the ways the intervention could address these were referred to throughout when making development decisions (Textbox 1).

XSL•FO
RenderX

Textbox 1. Planning phase of the person-based approach—DAPHNY (Diet and Activity for Pregnancy Hypertension) intervention guiding principles.

Design objective 1: To increase awareness of the association between lifestyle and high blood pressure

- Provide evidence for the association between diet and activity patterns and blood pressure
- Provide education on nutrients that support good blood pressure management

Design objective 2: To motivate women and increase confidence to undertake positive and sustainable lifestyle changes during pregnancy

- Emphasis on how to make small changes to build over time as a way to protect their own health in the short- and long-term health and a way to support their baby's growth and development
- Create a platform that provides feedback to increase confidence and motivation
- Motivational messages to prompt movement
- Notifications to prompt uptake of healthy dietary habits
- Repetition of information to enable understanding
- Opportunity to self-monitor
- Acknowledgment of the barriers to change and ways to overcome these

Design objective 3: Intervention content is simple, clear, and appropriate for lower health literacy

- Short sentences, avoiding complex language, terminology, abbreviations, and jargon
- Create a platform that is visual, attractive, and easy to navigate
- Wording and content of the platform supportive and encouraging
- Content and messages piloted with a diverse group of women, including those whose first language is not English

Theoretical Modeling

The capability opportunity motivation-behavior model (COM-B model) underpinned the intervention development, given its incorporation of other theories and practical application through identifying key barriers and facilitators that disable or enable positive behavior change [32,33]. It posits that behavior change is the result of direct or indirect interactions between the components: capability (skills and knowledge), opportunity (available resources and environmental factors), and motivation (brain processes such as intentions, emotional responses, and beliefs) to enact a target behavior. An introductory logic model (Figure 2) hypothesized how the intervention may work. Logic models provide a visual and testable representation to hypothesize how an intervention and its features may work. These provide fine-grain detail on the barriers, facilitators, and implemented features and behavior change techniques (BCTs), the smallest components of an intervention that influence behavior, to address these [34].

XSL•FO
RenderX

Figure 2. Planning phase of the person-based approach—introductory logic model of the digital lifestyle intervention. DASH: Dietary Approaches to Stop Hypertension; GWG: gestational weight gain.

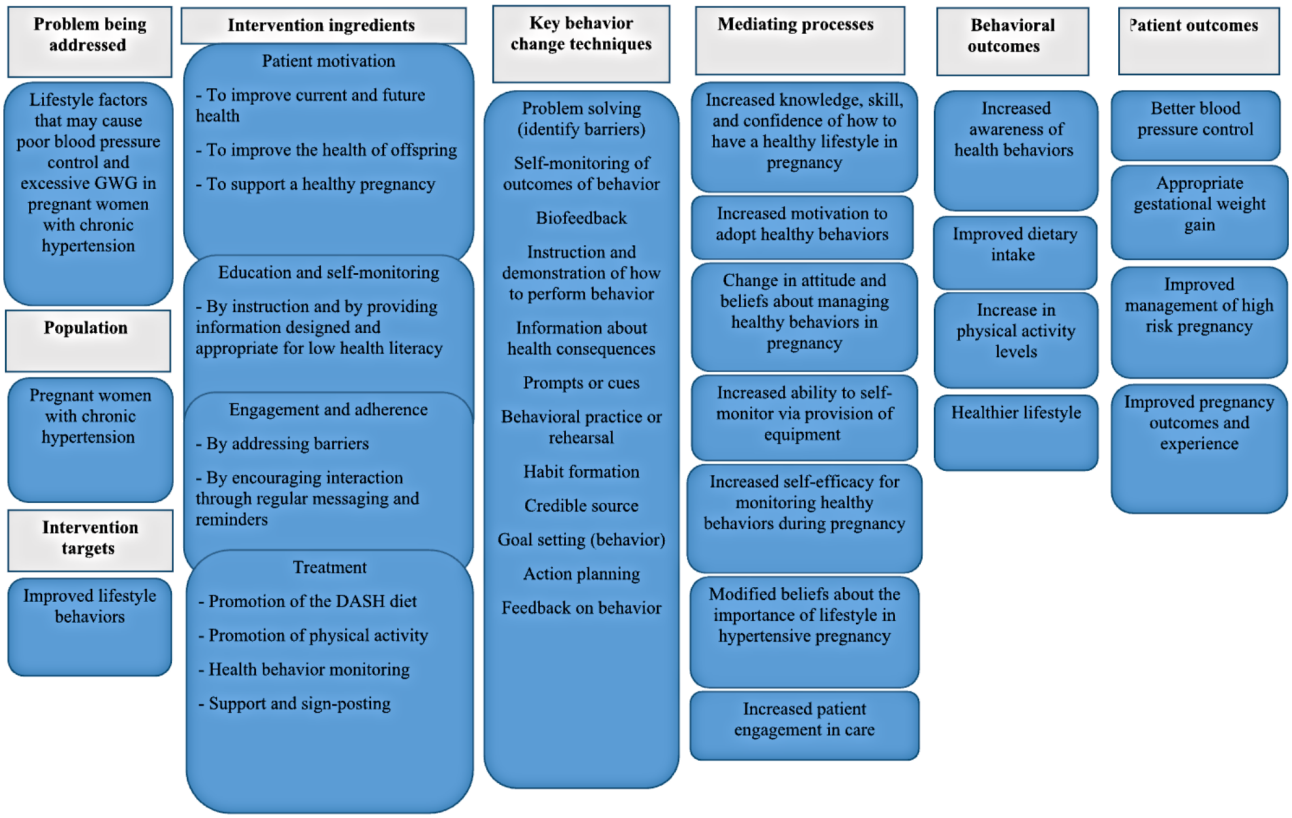

Optimization Phase

Overview

Focus groups and early feasibility testing involving think-aloud interviews and user testing took place at the maternity units within Oxford University Hospital Trust and Manchester University Foundation Trust. All focus groups and interviews were conducted by the lead researcher.

Health Care Professional Focus Groups

Focus groups with maternity health care professionals were carried out to understand views of the developing intervention, as well as important insight into experiences of hypertension management and lifestyle advice for this population.

Recruitment

Participants were any health care professional providing care for hypertensive pregnancies with varying experience working in different clinical settings. Participant information was shared, and time was given to consider participation prior to obtaining informed consent and commencing study activities.

Data Collection

Semistructured interviews or focus groups involved discussions about usual care pathways for pregnant women with chronic hypertension and the experience of providing lifestyle advice. Prototypes of the app were presented, using FIGMA (a web app for user interface design), to stimulate discussion and questions oriented around these. Discussions were fluid between directly relating to the intervention, to wider challenges experienced, and views on providing care for these women.

Early Feasibility Testing With Women

Overview

Early feasibility testing involved first, and in parallel to health care professional focus groups, think-aloud interviews to design and refine intervention features. Subsequently, user testing to understand the use of the developed intervention in real-life contexts was conducted.

Recruitment

Participants were pregnant, or recently pregnant (for think-aloud interviews), women with chronic hypertension. For the user testing, participants were required to be currently pregnant and have a compatible smartphone (Android or iOS) for the developed DAPHNY (Diet and Activity for Pregnancy Hypertension) app.

Purposive sampling was used to recruit eligible participants from specialist antenatal clinics, antenatal wards, or via email. Information about the study was provided, and following time to consider participation, consent was obtained prior to study enrollment.

Data Collection

Think-aloud interviews are an evidence-based qualitative interview technique commonly adopted in the PBA to gather important insights from target users about how the intervention can be adapted to meet their needs and overcome barriers [35,36].

Interviews began with general questions about the experience of lifestyle advice in the context of hypertension, as well as current behaviors (lifestyle or other) to manage their blood

XSL•FO
RenderX

pressure. Participants then freely explored the prototype, navigating into different intervention components as they wished, sharing immediate reactions and thoughts throughout to establish user acceptability and inform changes. Prompts used were informed by the COM-B model, and questions asked responded to women's interaction with the prototype to elicit the most useful information for intervention development and optimization.

Following think-aloud interviews, user testing with a different group of participants was carried out. Participants used the developed DAPHNY app at home within their daily lives for at least 2 weeks, and app use data were collected. At enrollment, participants were given access to the DAPHNY app. An initial orientation was given, and brief written instructions were provided. Qualitative interviews following this period were carried out and will be published elsewhere.

Qualitative Data Management

Data collection continued until no new major information was appearing with additional interviews, that is, comments related more to personal preferences than key intervention functions that influenced user engagement or behavior change, as recommended in the PBA.

All focus groups and interviews were audio-recorded, deidentified, and transcribed by the lead researcher or an independent transcriber. All quotes directly relating to the intervention design, content, component, and features were extracted and tabulated into an Excel (Microsoft Corp) workbook, which organized and managed the dataset for a content analysis.

Analysis

Health Care Professional Focus Groups and Early Feasibility Testing, Think-Aloud Interviews

Content analysis was conducted throughout data collection so that modifications were made and then shown to subsequent participants. There were 3 main iterations throughout data collection with minor changes in between. Each comment was considered against a coding framework, commonly used in PBA, including important for behavior change, easy and uncontroversial, said repeatedly, supported by experience, does not contradict, and not changed, to support decision-making about potential intervention modification (Multimedia Appendix 2). The MoSCOW (Must have, Should have, Could have, Would like) prioritization model, commonly used for developing digital interventions, prioritized changes to make [37]. A "table of changes" documented the decisions and actions throughout the process, providing an audit trail of how the intervention developed over the course of the study.

The "table of changes" was summarized into logic models and data mapped against the COM-B model to define what participants felt the barriers and facilitators were to its use and engagement. Modifications to the intervention with the associated BCT, which were derived from the behavior change taxonomy, were detailed in the logic model [34], successfully linking the data to core elements of the behavior change wheel (COM-B) that underpins the intervention.

Early Feasibility Testing, User Testing

App use logbooks with raw data for each app page were exported in CSV format and analyzed in Excel. The following insights were obtained: the total duration of app use (including number of days and minutes), the duration and number of individual sessions (presented as mean, SD, median, and range), and the pages that were most frequently accessed. This data is presented in summary graphs and tables to provide a view of how the app was used. Data entered by the participants, including step count and blood pressure, are reported descriptively.

Data extracted included sessions occurring 20 minutes after initial installation of the app (to exclude the initial orientation session) to 5 PM on the date of the participant interview. A session was defined from the moment a participant opened the app until they exited the app, or if the app was not exited but it was running in the background, but the participant did not return to it again within half an hour. Sessions that were less than 5 seconds long were removed from the analysis to exclude any unintentional interactions that were too short to be meaningful [38]. Though this removed interactions such as looking at notifications, it meant that the higher-level pages did not appear as the most popular.

Results

Participant Characteristics

A multidisciplinary mix of 23 health care professionals took part in a focus group (n=7, consisting of between 2 and 4 participants) or an individual interview (n=4; Multimedia Appendix 3).

A total of 11 pregnant or recently pregnant women took part in think-aloud interviews, and 10 pregnant women tested the DAPHNY app at home. Participant characteristics are displayed in Multimedia Appendix 3.

Findings

Health Care Professional Focus Groups and Early Feasibility Testing, Think-Aloud Interviews

The table of changes informed the production of three logic models. Logic model 1 (Figure 3) acts as a mediating process that is necessary to lead to logic models 2 (Figure 4) and 3 (Figure 5); that is, accessing and engaging with the app is required to increase knowledge and capability to increase healthy eating and physical activity levels during pregnancy. Subsequent changes in blood pressure are outcomes associated with logic models 2 and 3, but were not a testable part of the intervention or within the scope of this study.

Table 1 displays an abridged excerpt from the table of changes to demonstrate an example of participant views and changes implemented.

XSL•FO
RenderX

Table . Abridged excerpt from the table of changes that informed changes to the intervention and logic models.

App page	Blood pressure–focused foods	Blood pressure portal	Blood pressure and your lifestyle	Movement portal	Weight portal	Reflections or free writing page
Participant comments	A. "really break things down into achievable steps ... they can seem overwhelming altogether ... like 1 little thing to think about" B. "it's quite information intense, isn't it?..."	"...include some information that what will be the effects if blood pressure is not controlled during pregnancy"	A. "I feel like if you have loads of information some people might read it all, but you might need it punchier with the option to click on it and read more. Need easy and quick to absorb information." B. "you don't want to be accusatory do you, I think that can be, people shut off immediately" "it would have probably ended up in another layer of guilt over stuff I wasn't doing that I knew I should have been doing"	A. "Even if it was a slow paced swim and then it was just kind of, told me Great, you've done something today, well done. I'd quite like that." B. "does the app have a pedometer built into it because that would be good"	"I would do it but I think ... I would be probably a bit upset to see it on a graph not showing where I want it to be. This would demotivate me."	"I like tracking my blood pressure. That data I put into an app because I get something from it, which is an important monitoring of an important metric, but kind of moods and even symptoms when it's not looking for a particular pattern, though, I just I just wouldn't bother."
Decision for change[a]	A. Said repeatedly, Important for behavior change. B. Important for behavior change	Experience, Important for behavior change	A. Important for behavior change, easy and uncontroversial. B. Supported by experience, said repeatedly, important for behavior change	A. Experience, important for behavior change B. Not changed	Experience	Said repeatedly, experience, does not contradict
MoScoW[b] criteria[c]	A. Should have B. Must have	Must have	A. Should have B. Must have	A. Should have B. Would like	Should have	Could have
Action taken	A. App presents a list of top tips which participants can choose to receive as a notification. Emphasis on how to break goals down into small steps. B. Simplified language and reduced text. Links to further information if wanted by some.	Short- and long-term effects of high blood pressure for both mother and baby made clear.	A. Links to more information for those who wish to read more. B. Tone and terms used throughout app supportive. Section to acknowledge difficulties to avoid adding additional burden or feelings of guilt.	A. "Well done" message prompted when an activity is entered. B. Not feasible but important consideration for future iterations.	Information about weight not included on main homepage. Made an optional section, women can choose to engage with this or not.	Page removed as unlikely to impact target behaviour therefore its removal does not contradict guiding principles.

[a]See Multimedia Appendix 2 for coding framework definitions.

[b]MoSCOW: Must have, Should have, Could have, Would like.

[c]Bradbury et al [37].

XSL•FO
RenderX

Figure 3. Logic model 1: target behavior—accessing and engaging with a lifestyle app. BCT: behavior change technique; COM-B: capability opportunity motivation-behavior model.

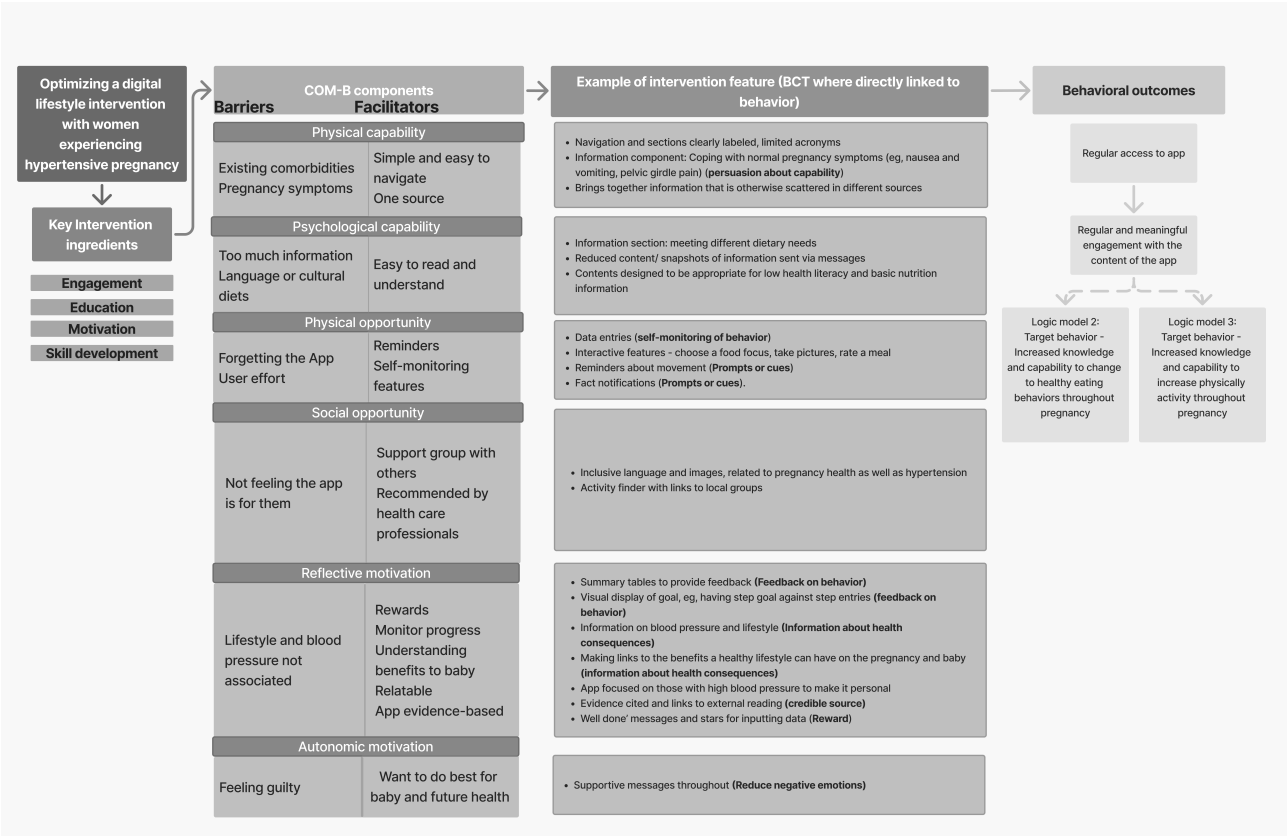

Figure 4. Logic model 2: target behavior—increased knowledge and capability to change to healthy eating behaviors throughout pregnancy. BCT: behavior change technique; COM-B: capability opportunity motivation-behavior model.

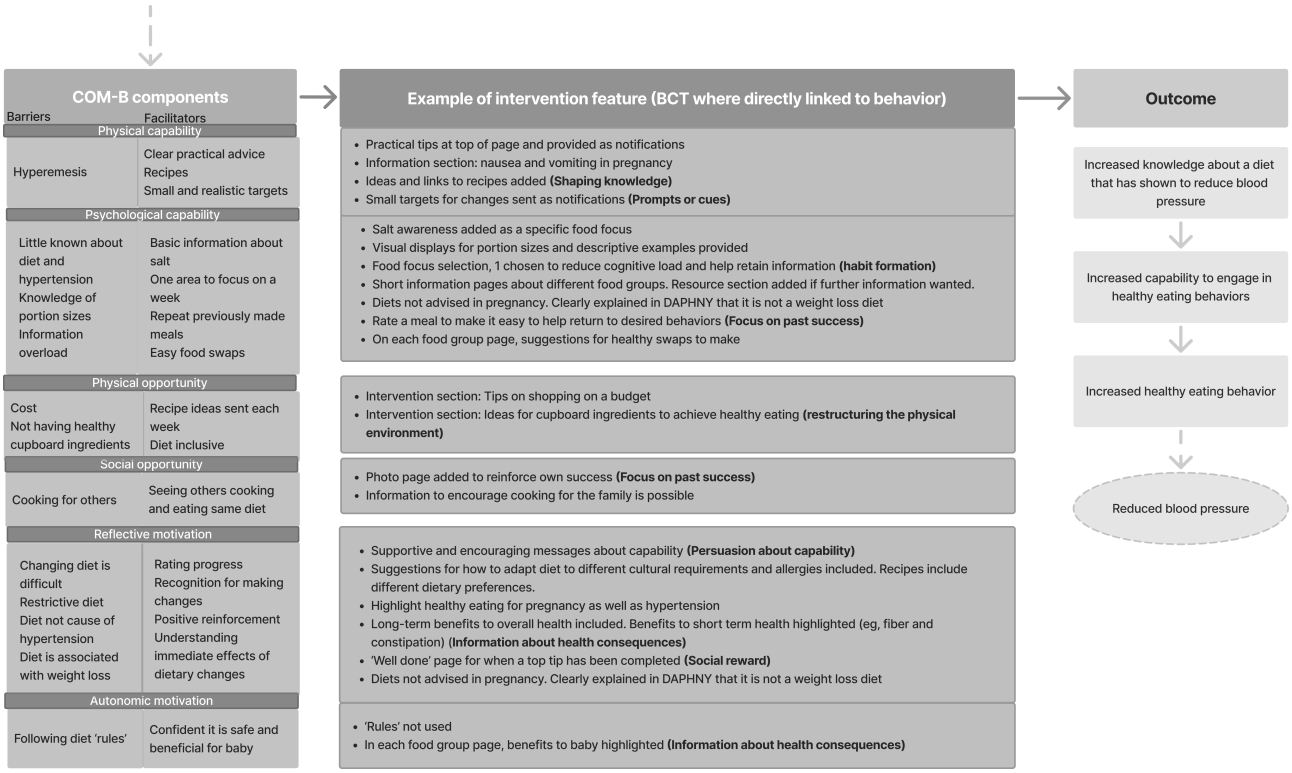

XSL•FO
RenderX

Figure 5. Logic model 3: target behavior: Increased knowledge and capability to increase physical activity throughout pregnancy. BCT: behavior change technique; COM-B: capability opportunity motivation-behavior model.

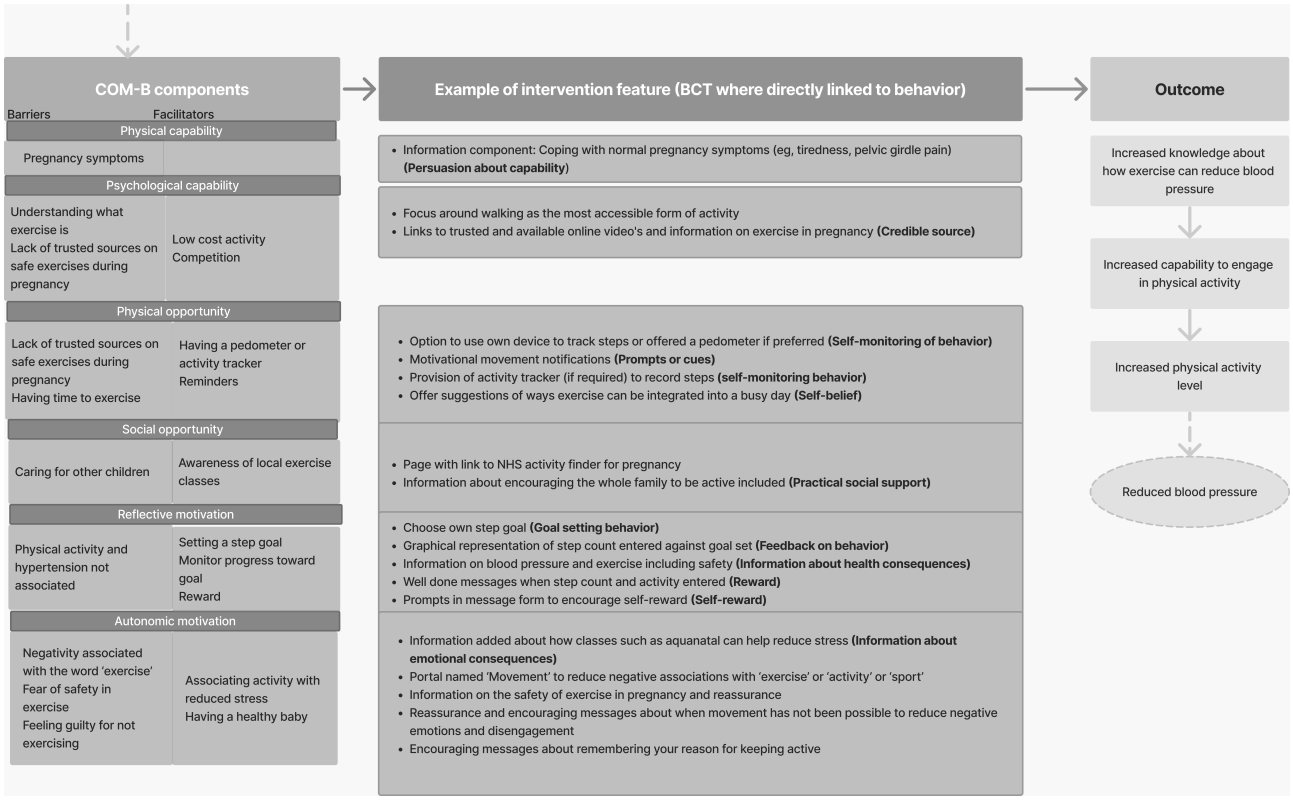

Key features included supportive messaging, acknowledging the challenges of being pregnant, goal setting, rewards and recognition that effort has been made, feedback on behavior including self-monitoring data, prompts via push notifications, shaping knowledge around how to have healthy behaviors, information about health consequences such as pregnancy outcomes associated with poor blood pressure control, and the information on the app being from a credible source and therefore sources used clearly referenced.

Figure 6 displays example screenshots with associated BCTs. Further screenshots can be viewed in Multimedia Appendix 4.

Some desirable features were not feasible within this intervention development study but are important findings to inform future iterations, supporting access, engagement, and implementation of the intervention (Table 2).

XSL•FO
RenderX

Figure 6. Screenshots of the DAPHNY (Diet and Activity for Pregnancy Hypertension) app prototype with associated behavior change techniques following iterative feedback from participant interviews and focus groups.

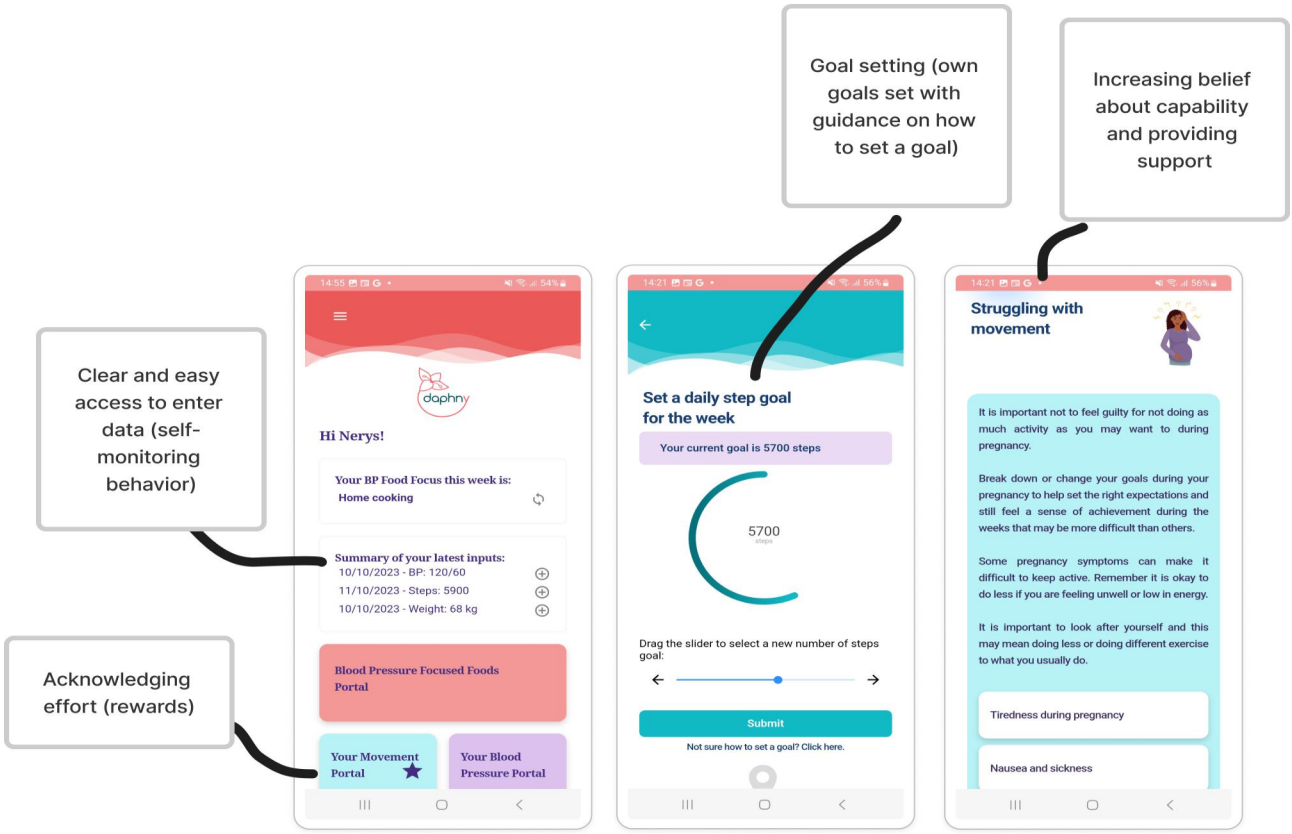

Table . Findings indicating important considerations for future iterations.

COM-B[a] component	Findings	Future features to consider
Physical capability	• Not all women have smartphones or use apps • Downloadable on different devices • Clear meal plans	• Printable or web version • Weekly meal plans with shopping list
Psychological capability	• Only available in English	• Available in different languages
Physical opportunity	• Health care professionals do not have training on lifestyle factors and hypertension • Tool to help facilitate health care professionals talk about the app • Linked devices • Entering food intake to learn about intake • Being able to measure portion sizes	• Specific training for health care professionals about lifestyle factors and hypertension • Implement a prompt for consultations to remind professionals to introduce the app • Link to other devices and apps (relevant to self-monitoring features) • Optional food tracking feature • Specific lifestyle-related training • Provision of food scales
Social opportunity	• Recommended by a health care professional • Support community • Seeing other people's stories • Competition with others (steps)	• Chat forum with other users or similar • Success stories about making lifestyle changes • Gamification with other users
Reflective motivation	• Using the app in the postnatal or preconception period	• Talk to GPs[b] and other health care professionals about the app • Consider application across the life course
Autonomic motivation	• None identified	• None identified

[a]COM-B: capability opportunity motivation-behavior.

[b]GP: general practitioner.

XSL•FO
RenderX

Following changes derived from focus groups and think-aloud interviews, Version 0.0.1 of the DAPHNY app was built. The app was tested for usability and functionality within the research team, including PPI contributors. Technical issues were resolved prior to user testing.

Early Feasibility Testing, User Testing

Summary of Total App Use

Participants had access to the app for between 16 and 27 days. The proportion of days participants accessed the app within this time varied from 14% to 90%. The total number of minutes a participant accessed the app over the course of their participation ranged from 5 minutes 57 seconds (PPT19) to 51 minutes 8 seconds (PPT18; Figure 7).

Figure 7. Total app use by each participant (ordered by no, sessions, high to low) during user testing.

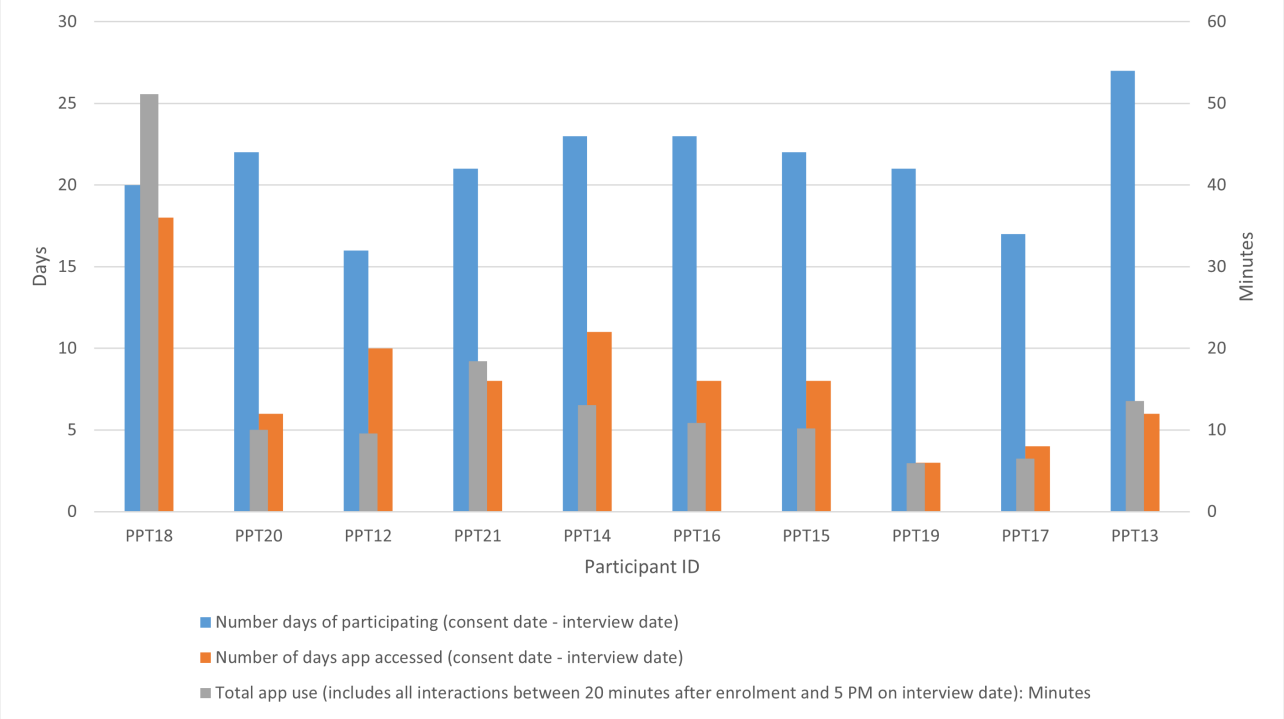

PPT18 spent the most time, had the greatest number of sessions (23) on the app, and accessed it on the majority of days during participation (90%).

Spending fewer days on the app did not necessarily equate to less time on the app. For example, PPT20 and PPT12 spent a similar amount of time on the app (10 min 1 s and 9 min 34 s, respectively), yet for PPT12, this was spread over 4 more days than PPT20, indicating more frequent access to the app but shorter sessions.

There were 116 sessions in total (Table 3). The majority of the sessions had durations that were more akin to the minimum duration, with a median length of 36 seconds.

Table . Summary of app sessions during the user-testing period.

	Mean (SD)	Median	Range (min-max)
Number of sessions per participant	13 (6.84)	13	3 - 23
Duration of sessions (mm:ss)[a]	01:19 (02:27)	00:36	00:05-05:20

[a]A 21:40 minute session was removed as only one page (step summary page) was accessed during this time, suggesting the app was left open but with no active interaction from the participant. Mean and median are calculated from the inclusion of 21:40.

Pages Accessed

The number of individual pages that the participants accessed ranged from 7 to 23 pages. The pages requiring the user to take an action, such as blood pressure readings and choosing a food focus, were accessed by more participants than the pages with written information, which required a deeper level of engagement through multiple layers of the app. Data entry pages “Enter a clinic blood pressure reading” (40 sessions) and “Record today’s step count” (39 sessions) had the greatest number of sessions, indicating that participants’ use of the app was predominantly action-based as opposed to reading information.

The least explored area of the app was the food portal. Despite this finding, participants spoke about how it was useful having this information in one place where they could find out more if they needed it. Furthermore, information about food was received in notifications, which may have led to reduced interaction with this section.

XSL•FO
RenderX

The blood pressure portal was the most frequently accessed portal (24 sessions). The “Explore more” portal had 3 sessions. This portal included information on weight and healthy living after giving birth, as well as stress, sleep, alcohol, and smoking information, indicating that few participants accessed these additional subsections, contradicting a desire for these in think-aloud interviews.

Entering Data

Although participants were instructed to only enter blood pressure readings from the clinic, four participants entered blood pressure readings between 4 and 19 submissions, suggesting some were using it to record their self-monitoring readings. There was a keenness to have some of the information available on the DAPHNY app in the same place as they were entering their self-monitoring blood pressure reading.

Five participants set a step goal, 3 of whom did this during the enrollment session. In total, 6 participants entered a step count between 2 and 12 times, of whom 2 of these 6 did not set a step goal. Some participants did not enter a step count because they recorded this elsewhere on apps linked to existing activity trackers or their phone, indicating a need for interconnectivity between devices.

Few participants used the weight tracking feature. The weight section required navigation through multiple pages of the app, and therefore, some participants were not aware it existed. In some cases, participants did not feel this was needed as there was enough to explore in the movement, diet, and blood pressure sections. The focus on nutrients, as opposed to weight, was more acceptable to some participants.

Discussion

Summary of Findings

This study led to the development of an evidence-based, theoretically informed mobile lifestyle app (the DAPHNY app) that, through methods of the PBA, prioritized the needs and perspectives of those with chronic hypertension during pregnancy. Participant feedback informed three logic models that identified key intervention features to optimize the intervention. Multiple iterations of the app implementing these changes led to a first version that was user tested by a group of women. Action-based features were accessed more frequently than educational pages. Strategies to increase engagement, improve acceptability, and support implementation were offered by participants.

Comparison to the Wider Literature

Women in the study were accustomed to self-monitoring their blood pressure, and some of the information on the app was considered an acceptable addition to existing practices. However, engagement in the lifestyle information was relatively limited, considering the extent of available content. A study that developed an intervention for nonpregnant patients with hypertension found that the lifestyle component had little traction with participants, especially if medication sufficiently controlled blood pressure, suggesting a lack of overall engagement in this part of blood pressure management [27]. This has also been demonstrated in a systematic review, which found that promoting positive lifestyle behaviors within hypertensive populations was met with a lack of willingness, even where there was knowledge of the benefits [39]. Embedding lifestyle-promoting messages effectively that lead to meaningful change within hypertensive populations remains a challenge.

Barriers to engaging with healthy behaviors that were identified by women with chronic hypertension in the study were similar to those identified by the general pregnant population. These included common pregnancy symptoms such as general discomfort, sickness or tiredness, and work or family commitments, as well as safety concerns [40-42]. Furthermore, cost and time have been found by others to be major barriers in consuming heart-healthy diets among women with risk factors for hypertension [43]. The DAPHNY app addressed some of these barriers through codeveloping supportive, practical, and realistic messaging to manage cognitive load and reduce feelings of guilt. Sharing goals, progress, and efforts with a health care professional, who women view as a credible source of information and support, alongside the app, was important to mitigate some concerns shared by participants in the study. This aligns with research that developed an activity app for women with gestational diabetes [44]. Health care professionals’ health promotion role and how they deliver this information must be examined to understand if the app could be used as a tool to facilitate this partnership with women.

Women were keen to be acknowledged for their efforts through a reward and recognition system via the app. This may reflect women’s desire to be recognized for their work as expectant mothers whilst navigating a system that is oriented around monitoring and medication regimes. Bringing women back into the conversation about their own bodies, including ways they may manage blood pressure in their personal domain, beyond the maternity care setting, could contribute toward opening up the dialogue where women can believe in the capabilities of their own bodies. The DAPHNY app gives women a tool by providing content that is woman-centered, the underlying philosophy of the app, to acknowledge the experiences through hypertensive pregnancy, its challenges, and provide support to manage or overcome these.

The notorious challenge of multiple, already available systems and their integration was brought to light in developing this digital intervention. For example, some women did not use the DAPHNY app to monitor their steps, because this was recorded elsewhere and automatically. Similar to others, health care professional participants expressed enthusiasm for the app to be linked to existing apps and electronic health records [45]. This supports the NHS England Better Births report and the NHS Long Term plan that value the role of digital tools in maternity care [17,46]. Though there is a significant preference across all participants for embedding and linking existing technologies, the complex landscape of both the digital field and large health care systems continues to be a barrier to the seamless implementation of apps for the benefit of the end user [47]. Synchronizing a digital intervention across different trusts and professionals would require funding, national-level stakeholder involvement, and a host organization to manage,

XSL•FO
RenderX

update, and provide governance to oversee the safety and ongoing usability of an intervention. Collaboration with those driving the digital agenda within maternity care is essential to bolster the development of a lifestyle intervention beyond its own silo. As a small first step toward interconnected systems, adjoining the findings from this development study with apps that are used within current care would be appropriate to investigate. This would benefit from embedding implementation theory to address barriers, prior to proceeding to a larger-scale feasibility trial to validate and more comprehensively understand if and how a digital lifestyle intervention for hypertensive pregnancy can be embedded within the NHS existing system and women's lives.

Limitations

The DAPHNY app was designed with experts and relevant users of the intervention, with continuous input from PPI representatives. It builds on a strong body of evidence that women can and are willing to self-manage their hypertension at home. Therefore, this study appropriately addresses further self-management strategies that may influence blood pressure management through pregnancy that have not been explored before [24,48,49].

The perspectives of those with long-term hypertension in relation to their lifestyle behaviors were elicited, which, to our knowledge, has not been done before. Pregnant women whose first language was not English were included, as well as those who followed their own cultural diets, which gave important insight when discussing the food-related sections of the app. There was limited diversity in the sample of participants in the think-aloud interviews in relation to ethnicity. However, 6 participants enrolled in user testing were from different non-White British ethnic backgrounds, thereby increasing the diversity of views heard compared to the first interviews. Furthermore, there was a range of educational attainment, which is a strength of the study, as those with lower education levels are underrepresented in research [50]. An important next step is to speak to women attending different NHS Trusts with diverse backgrounds to further learn how the provision of lifestyle information within the context of hypertension management may be adapted and optimized. Digital interventions may not address needs across populations, for example, those with low digital health literacy, limited internet access, or infrequent device use. However, insights from this study can inform the design of different models of information delivery to support health promotion that better meet the needs of women with chronic hypertension. The UK context of the study means findings may not translate across different economic and health care landscapes.

Conclusions

The PBA led to the development of a novel, theoretically underpinned, evidence-based lifestyle app designed to address the needs and perspectives of women with chronic hypertension during pregnancy. Valuable insight from this important population provides new data that adjoins lifestyle behavior and hypertension during pregnancy that has previously been overlooked but has the potential to improve outcomes, as evidenced outside of pregnancy.

Future work must address the recommended changes and the learnings from this work with a view to moving toward the implementation and evaluation stage of the PBA so it can be formally piloted and tested for effectiveness on behavior change and subsequent blood pressure control. Good lifestyle support alongside regular monitoring and management of hypertension has the potential to improve cardiovascular health during pregnancy and beyond. This study is a step toward putting this on the public health and maternity agenda, which should be noted by professionals and researchers.

Acknowledgments

This research was funded by the National Institute of Health Research (NIHR) Applied Research Collaboration (ARC) Oxford and Thames Valley. LG, KT, and RJM were funded by the NIHR ARC Oxford and Thames Valley during the time of this study. NMA is funded by Diabetes UK, National Institutes of Health Research School of Primary Care Research, and the Diabetes Research & Wellness Foundation. CR is supported by the Oxford NIHR Biomedical Research Center. The views expressed are those of the authors and not necessarily those of the National Health Service (NHS), the NIHR, or the Department of Health and Social Care. This study was part of a body of work submitted for a PhD thesis. Thank you to all of the participants who kindly gave up their time to take part in this study. A particular thanks to Katherine and Patricia, our patient and public contributors, who gave such important insight throughout the studies. Thank you to Prof Jenny Myers and Dr Lucy MacKillop, who were the site principal investigators, and the research midwives and clinical staff who provided support with recruitment and also took part in focus groups.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

LG, RJM, KT, and NMA conceived the study. LG developed the protocol with support from all coauthors. LG led the study with supervision from RJM, KT, and NMA. LG conducted all focus groups and interviews and conducted the analysis with support from all coauthors. LG created the prototypes, and the app versions were subsequently built by CR and YC, who also provided

XSL•FO
RenderX

technical support throughout. LG wrote the first draft of the manuscript, which was reviewed, and a final version was agreed upon by all coauthors.

Conflicts of Interest

RJM has developed a telemonitoring intervention with Omron for which the University of Oxford received licensing and consultancy fees.

Multimedia Appendix 1
Planning phase of the person-based approach—overview of scoping the relevant literature for intervention development to inform prototypes.
[DOCX File, 18 KB - formative_v9i1e68927_app1.docx]

Multimedia Appendix 2
Coding framework used to facilitate decisions about intervention modifications.
[DOCX File, 13 KB - formative_v9i1e68927_app2.docx]

Multimedia Appendix 3
Participant characteristics for the health care professional focus groups (n=23), think-aloud interviews (n=11), and user testing (n=10).
[DOCX File, 15 KB - formative_v9i1e68927_app3.docx]

Multimedia Appendix 4
Screenshots of key behavior change techniques within the intervention.
[DOCX File, 674 KB - formative_v9i1e68927_app4.docx]

Checklist 1
GUIDED (Guidance for the Reporting of Intervention Development) checklist.
[PDF File, 219 KB - formative_v9i1e68927_app5.pdf]

References

1. Valenzuela PL, Carrera-Bastos P, Gálvez BG, et al. Lifestyle interventions for the prevention and treatment of hypertension. Nat Rev Cardiol 2021 Apr;18(4):251-275. [doi: 10.1038/s41569-020-00437-9] [Medline: 33037326]
2. Poon LC, Nguyen‐Hoang L, Smith GN, et al. Hypertensive disorders of pregnancy and long‐term cardiovascular health: FIGO best practice advice. Intl J Gynecology Obste 2023 Jan;160(S1):22-34. [doi: 10.1002/ijgo.14540]
3. Goddard L, Patel R, Astbury NM, Tucker K, McManus RJ. Evidence of lifestyle interventions in a pregnant population with chronic hypertension and/or pre-existing diabetes: a systematic review and narrative synthesis. Pregnancy Hypertens 2023 Mar;31:60-72. [doi: 10.1016/j.preghy.2022.12.004] [Medline: 36608587]
4. Yoder SR, Thornburg LL, Bisognano JD. Hypertension in pregnancy and women of childbearing age. Am J Med 2009 Oct;122(10):890-895. [doi: 10.1016/j.amjmed.2009.03.036] [Medline: 19786154]
5. Hypertension in pregnancy: diagnosis and management. National Institute for Health and Care Excellence (NICE). 2019. URL: https://www.nice.org.uk/guidance/NG133 [accessed 2025-07-09]
6. Lamminpää R, Vehviläinen-Julkunen K, Gissler M, Heinonen S. Preeclampsia complicated by advanced maternal age: a registry-based study on primiparous women in Finland 1997-2008. BMC Pregnancy Childbirth 2012 Jun 11;12(1):47. [doi: 10.1186/1471-2393-12-47] [Medline: 22687260]
7. Bodnar LM, Ness RB, Markovic N, Roberts JM. The risk of preeclampsia rises with increasing prepregnancy body mass index. Ann Epidemiol 2005 Aug;15(7):475-482. [doi: 10.1016/j.annepidem.2004.12.008] [Medline: 16029839]
8. Kent L, McGirr M, Eastwood KA. Global trends in prevalence of maternal overweight and obesity: a systematic review and meta-analysis of routinely collected data retrospective cohorts. Int J Popul Data Sci 2024;9(2):2401. [doi: 10.23889/ijpds.v9i2.2401] [Medline: 40313349]
9. Health survey for England 2021: part 4—trends. NHS Digital. 2023. URL: https://digital.nhs.uk/data-and-information/publications/statistical/health-survey-for-england/2021/part-4-trends [accessed 2025-07-09]
10. Global report on hypertension: the race against a silent killer. World Health Organization. 2023. URL: https://www.who.int/teams/noncommunicable-diseases/hypertension-report [accessed 2025-07-09]
11. Maternal and child nutrition: nutrition and weight management in pregnancy, and nutrition in children up to 5 years. National Institute for Health and Care Excellence (NICE). 2025. URL: https://www.nice.org.uk/guidance/ng247 [accessed 2025-07-09]

XSL•FO
RenderX

12. International Weight Management in Pregnancy (i-WIP) Collaborative Group. Effect of diet and physical activity based interventions in pregnancy on gestational weight gain and pregnancy outcomes: meta-analysis of individual participant data from randomised trials. BMJ 2017 Jul 19;358:j3119. [doi: 10.1136/bmj.j3119] [Medline: 28724518]
13. Fair F, Soltani H. A meta-review of systematic reviews of lifestyle interventions for reducing gestational weight gain in women with overweight or obesity. Obes Rev 2021 May;22(5):e13199. [doi: 10.1111/obr.13199] [Medline: 33459493]
14. Thangaratinam S, Rogozinska E, Jolly K, et al. Effects of interventions in pregnancy on maternal weight and obstetric outcomes: meta-analysis of randomised evidence. BMJ 2012 May 16;344:e2088. [doi: 10.1136/bmj.e2088] [Medline: 22596383]
15. Muktabhant B, Lawrie TA, Lumbiganon P, Laopaiboon M. Diet or exercise, or both, for preventing excessive weight gain in pregnancy. Cochrane Database Syst Rev 2015 Jun 15;2015(6):CD007145. [doi: 10.1002/14651858.CD007145.pub3] [Medline: 26068707]
16. 2025/26 priorities and operational planning guidance. NHS England. 2025. URL: https://www.england.nhs.uk/long-read/2025-26-priorities-and-operational-planning-guidance/#our-national-priorities-for-2025-26 [accessed 2025-07-09]
17. Maternity transformation programme. NHS England. 2024. URL: https://www.england.nhs.uk/mat-transformation/ [accessed 2025-07-09]
18. Brammall BR, Garad RM, Boyle JA, et al. Assessing the content and quality of digital tools for managing gestational weight gain: systematic search and evaluation. J Med Internet Res 2022 Nov 25;24(11):e37552. [doi: 10.2196/37552] [Medline: 36427237]
19. Carrandi A, Hayman M, Harrison CL. Safety considerations for assessing the quality of apps used during pregnancy: a scoping review. Digital Health 2023;9:20552076231198683. [doi: 10.1177/20552076231198683] [Medline: 37675058]
20. van den Heuvel JF, Groenhof TK, Veerbeek JH, et al. eHealth as the next-generation perinatal care: an overview of the literature. J Med Internet Res 2018 Jun 5;20(6):e202. [doi: 10.2196/jmir.9262] [Medline: 29871855]
21. Vogels-Broeke M, Daemers D, Budé L, de Vries R, Nieuwenhuijze M. Sources of information used by women during pregnancy and the perceived quality. BMC Pregnancy Childbirth 2022 Feb 8;22(1):109. [doi: 10.1186/s12884-022-04422-7] [Medline: 35135487]
22. Band R, Hinton L, Tucker KL, et al. Intervention planning and modification of the BUMP intervention: a digital intervention for the early detection of raised blood pressure in pregnancy. Pilot Feasibility Stud 2019;5(1):153. [doi: 10.1186/s40814-019-0537-z] [Medline: 31890265]
23. Chappell LC, Tucker KL, Galal U, et al. Effect of self-monitoring of blood pressure on blood pressure control in pregnant individuals with chronic or gestational hypertension: The BUMP 2 randomized clinical trial. JAMA 2022 May 3;327(17):1666-1678. [doi: 10.1001/jama.2022.4726] [Medline: 35503345]
24. Hinton L, Tucker KL, Greenfield SM, et al. Blood pressure self-monitoring in pregnancy (BuMP) feasibility study; a qualitative analysis of women's experiences of self-monitoring. BMC Pregnancy Childbirth 2017 Dec 19;17(1):427. [doi: 10.1186/s12884-017-1592-1] [Medline: 29258469]
25. Tucker KL, Mort S, Yu LM, et al. Effect of self-monitoring of blood pressure on diagnosis of hypertension during higher-risk pregnancy: the BUMP 1 randomized clinical trial. JAMA 2022 May 3;327(17):1656-1665. [doi: 10.1001/jama.2022.4712] [Medline: 35503346]
26. Duncan E, O'Cathain A, Rousseau N, et al. Guidance for reporting intervention development studies in health research (GUIDED): an evidence-based consensus study. BMJ Open 2020 Apr 8;10(4):e033516. [doi: 10.1136/bmjopen-2019-033516] [Medline: 32273313]
27. Bradbury K, Morton K, Band R, et al. Using the person-based approach to optimise a digital intervention for the management of hypertension. PLoS One 2018;13(5):e0196868. [doi: 10.1371/journal.pone.0196868] [Medline: 29723262]
28. Yardley L, Ainsworth B, Arden-Close E, Muller I. The person-based approach to enhancing the acceptability and feasibility of interventions. Pilot Feasibility Stud 2015;1(1):37. [doi: 10.1186/s40814-015-0033-z] [Medline: 27965815]
29. Baines R, Bradwell H, Edwards K, et al. Meaningful patient and public involvement in digital health innovation, implementation and evaluation: a systematic review. Health Expect 2022 Aug;25(4):1232-1245. [doi: 10.1111/hex.13506] [Medline: 35526274]
30. Public involvement. Health Research Authorty. 2024. URL: https://www.hra.nhs.uk/planning-and-improving-research/best-practice/public-involvement/ [accessed 2025-07-09]
31. Mockford C, Staniszewska S, Griffiths F, Herron-Marx S. The impact of patient and public involvement on UK NHS health care: a systematic review. Int J Qual Health Care 2012 Feb;24(1):28-38. [doi: 10.1093/intqhc/mzr066] [Medline: 22109631]
32. Michie S, van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. Implement Sci 2011 Apr 23;6(42):42. [doi: 10.1186/1748-5908-6-42] [Medline: 21513547]
33. Michie S, Atkins L, West R. The behaviour change wheel. In: A Guide to Designing Interventions, 1st edition: Silverback Publishing; 2014:1003-1010.
34. Michie S, Richardson M, Johnston M, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013 Aug;46(1):81-95. [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]

XSL•FO
RenderX

35. Yardley L, Morrison LG, Andreou P, Joseph J, Little P. Understanding reactions to an internet-delivered health-care intervention: accommodating user preferences for information provision. BMC Med Inform Decis Mak 2010 Sep 17;10(1):52. [doi: 10.1186/1472-6947-10-52] [Medline: 20849599]
36. van den Haak MJ, de Jong MDT, Schellens PJ. Evaluation of an informational web site: three variants of the think-aloud method compared. Tech Commun 2007;54(1):58-71 [FREE Full text]
37. Bradbury K, Watts S, Arden-Close E, Yardley L, Lewith G. Developing digital interventions: a methodological guide. Evid Based Complement Alternat Med 2014;2014:561320. [doi: 10.1155/2014/561320] [Medline: 24648848]
38. Henriksson P, Migueles JH, Söderström E, Sandborg J, Maddison R, Löf M. User engagement in relation to effectiveness of a digital lifestyle intervention (the HealthyMoms app) in pregnancy. Sci Rep 2022 Aug 13;12(1):13793. [doi: 10.1038/s41598-022-17554-9] [Medline: 35963935]
39. Khatib R, Schwalm JD, Yusuf S, et al. Patient and healthcare provider barriers to hypertension awareness, treatment and follow up: a systematic review and meta-analysis of qualitative and quantitative studies. PLOS One 2014;9(1):e84238. [doi: 10.1371/journal.pone.0084238] [Medline: 24454721]
40. McKeough R, Blanchard C, Piccinini‐Vallis H. Pregnant and postpartum women's perceptions of barriers to and enablers of physical activity during pregnancy: a qualitative systematic review. J Midwife Womens Health 2022 Jul;67(4):448-462. [doi: 10.1111/jmwh.13375]
41. Grenier LN, Atkinson SA, Mottola MF, et al. Be healthy in pregnancy: exploring factors that impact pregnant women's nutrition and exercise behaviours. Matern Child Nutr 2021 Jan;17(1):e13068. [doi: 10.1111/mcn.13068] [Medline: 32705811]
42. Sui Z, Turnbull D, Dodd J. Enablers of and barriers to making healthy change during pregnancy in overweight and obese women. Australas Med J 2013;6(11):565-577. [doi: 10.4066/AMJ.2013.1881] [Medline: 24348873]
43. Kovell LC, Sibai D, Wilkie GL, et al. Identifying barriers, facilitators, and interventions to support healthy eating in pregnant women with or at risk for hypertensive disorders of pregnancy. Cardiovasc Digital Health J 2022 Dec;3(6 Suppl):S1-S8. [doi: 10.1016/j.cvdhj.2022.10.001] [Medline: 36589759]
44. Smith R, Michalopoulou M, Reid H, et al. Applying the behaviour change wheel to develop a smartphone application "stay-active" to increase physical activity in women with gestational diabetes. BMC Pregnancy Childbirth 2022 Mar 26;22(1):253. [doi: 10.1186/s12884-022-04539-9] [Medline: 35346075]
45. Tucker K, Rose F, Lavallee L, Roman C, Goddard L, McManus RJ. Intervention development and optimisation of a multi-component digital intervention for the monitoring and management of hypertensive pregnancy: the My Pregnancy Care Intervention. Pilot Feasibility Stud 2024 Nov 12;10(1):139. [doi: 10.1186/s40814-024-01562-9] [Medline: 39533431]
46. The NHS long term plan. NHS England. 2019. URL: https://www.longtermplan.nhs.uk/publication/nhs-long-term-plan/ [accessed 2025-07-09]
47. Marwaha JS, Landman AB, Brat GA, Dunn T, Gordon WJ. Deploying digital health tools within large, complex health systems: key considerations for adoption and implementation. NPJ Digital Med 2022 Jan 27;5(1):13. [doi: 10.1038/s41746-022-00557-1] [Medline: 35087160]
48. Tucker KL, Hinton L, Green M, Chappell LC, McManus RJ. Using self-monitoring to detect and manage raised blood pressure and pre-eclampsia during pregnancy: the BUMP research programme and its impact. Hypertens Res 2024 Mar;47(3):714-720. [doi: 10.1038/s41440-023-01474-w] [Medline: 38062200]
49. Pealing L, Tucker KL, Fletcher B, et al. Perceptions and experiences of blood pressure self-monitoring during hypertensive pregnancy: a qualitative analysis of women's and clinicians' experiences in the OPTIMUM-BP trial. Pregnancy Hypertens 2022 Dec;30:113-123. [doi: 10.1016/j.preghy.2022.09.006] [Medline: 36174484]
50. Witham MD, Anderson E, Carroll C, et al. Developing a roadmap to improve trial delivery for under-served groups: results from a UK multi-stakeholder process. Trials 2020 Aug 1;21(1):694. [doi: 10.1186/s13063-020-04613-7] [Medline: 32738919]

Abbreviations

BCT: behavior change technique
COM-B: capability opportunity motivation-behavior model
DAPHNY: Diet and Activity for Pregnancy Hypertension
GUIDED: Guidance for the Reporting of Intervention Development
MoSCOW: Must have, Should have, Could have, Would like
NHS: National Health Service
PBA: person-based approach
PPI: patient and public involvement

XSL•FO
RenderX

Edited by A Mavragani; submitted 22.11.24; peer-reviewed by G Katageri, SF Albasri; revised version received 12.06.25; accepted 14.06.25; published 18.07.25.

Please cite as:
Goddard L, Tucker K, Astbury NM, Roman C, Chi Y, Morgan K, Devitt P, McManus RJ
A Digital Lifestyle App for Hypertension During Pregnancy: Mixed Methods Intervention Development Study Using the Person-Based Approach
JMIR Form Res 2025;9:e68927
URL: https://formative.jmir.org/2025/1/e68927
doi:10.2196/68927

© Lucy Goddard, Katherine Tucker, Nerys M Astbury, Cristian Roman, Yuan Chi, Katherine Morgan, Patricia Devitt, Richard J McManus. Originally published in JMIR Formative Research (https://formative.jmir.org), 18.7.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Future Me, a Prospection-Based Chatbot to Promote Mental Well-Being in Youth: Two Exploratory User Experience Studies

Martin Dechant[1], PhD; Eva Lash[2], MSc; Sarah Shokr[1], MSc; Ciarán O'Driscoll[1], PhD

[1]University College London, 1-19 Torrington Place, London, United Kingdom
[2]University of Oxford, Oxford, United Kingdom

Corresponding Author:
Martin Dechant, PhD
University College London, 1-19 Torrington Place, London, United Kingdom

Abstract

Background: Digital interventions have been proposed as a solution to meet the growing demand for mental health support. Large language models (LLMs) have emerged as a promising technology for creating more personalized and adaptive mental health chatbots. While LLMs generate responses based on statistical patterns in training data rather than through conscious reasoning, they can be designed to support important psychological processes. Prospection—the ability to envision and plan for future outcomes—represents a transdiagnostic process altered across various mental health conditions that could be effectively targeted through such interventions. We designed "Future Me," an LLM-powered chatbot designed to facilitate future-oriented thinking and promote goal pursuit using evidence-based interventions including visualization, implementation intentions, and values clarification.

Objective: This study aims to understand how users engage with Future Me, evaluate its effectiveness in supporting future-oriented thinking, and assess its acceptability across different populations, with particular attention to postgraduate students' stress management needs. We also seek to identify design improvements that could enhance the chatbot's ability to support users' mental well-being.

Methods: In total, 2 complementary studies were conducted. Study 1 (n=20) examined how postgraduate students used Future Me during a single guided session, followed by semistructured interviews. Study 2 (n=14) investigated how postgraduate students interacted with Future Me over a 1-week period, with interviews before and after usage. Both studies analyzed conversation transcripts and interview data using thematic analysis to understand usage patterns, perceived benefits, and limitations.

Results: Across both studies, participants primarily engaged with Future Me to discuss career or education goals, personal obstacles, and relationship concerns. Users valued Future Me's ability to provide clarity around goal-setting (85% of participants), its nonjudgmental nature, and its 24/7 accessibility (58%). Future Me effectively facilitated self-reflection (80%) and offered new perspectives (70%), particularly for broader future-oriented concerns. However, both studies revealed limitations in the chatbot's ability to provide personalized emotional support during high-stress situations, with participants noting that responses sometimes felt formulaic (50%) or lacked emotional depth. Postgraduate students specifically emphasized the need for greater context awareness during periods of academic stress (58%). Overall, 57% of requests occurred outside office hours, dropping from 40 on day 1 to 12 by day 7.

Conclusions: Future Me demonstrates promise as an accessible tool for promoting prospection skills and supporting mental well-being through future-oriented thinking. However, effectiveness appears context-dependent, with prospection techniques more suitable for broader life decisions than acute stress situations. Future development should focus on creating more adaptive systems that can adjust their approach based on the user's emotional state and immediate needs. Rather than attempting to replicate human therapy entirely, chatbots like Future Me may be most effective when designed as complementary tools within broader support ecosystems, offering immediate guidance while facilitating connections to human support when needed.

(JMIR Form Res 2025;9:e74411) doi:10.2196/74411

KEYWORDS

chatbot; artificial intelligence; prospection; mental health; student well-being; stress; anxiety

Introduction

The rising demand for mental health support has led to a significant gap between those seeking help and available resources. Approximately 11% of people in the United States seek psychotherapy annually [1], with demand consistently outstripping supply. Additionally, postgraduate students face unique mental health challenges, including academic workload, financial insecurity, and social isolation, which can negatively

XSL•FO
RenderX

impact their well-being [2,3]. Digital interventions have emerged as accessible and effective early intervention or prevention techniques to address mental health problems, potentially reducing the burden on traditional services and providing support when human alternatives are unavailable.

Digital mental health interventions face significant challenges, including low adherence rates [4], financial and logistical barriers to accessibility [5], and a lack of personalization. Many users disengage from digital tools due to generic, repetitive responses that fail to adapt to individual emotional states and stress levels [6,7]. Addressing these challenges is crucial for developing effective, scalable mental health tools that align with users' needs, particularly for younger populations like students who may be more receptive to digital solutions.

Recent advancements in large language models (LLMs) have transformed the landscape of digital mental health interventions. Unlike earlier rule-based chatbots, LLM-based conversational agents can gather situation-specific information through natural conversation to provide personalized and context-specific interventions. These chatbots could improve equity and accessibility [8] by personalizing interventions, reminding users, tracking progress, and offering multilingual and adaptable communication options for users of diverse backgrounds and technology literacy.

Therapeutic chatbots have been used across a broad spectrum of mental health conditions, focusing on specific issues such as anxiety, depression, and autism [1]. Most of these interventions position the chatbot as a different entity, like a digital therapist. However, LLM-powered chatbots may have the capability to enact more convincing alternative roles, offering novel therapeutic approaches beyond traditional digital therapy paradigms.

One promising application of LLM chatbots in mental health involves enabling goal pursuit, an aspect of functioning compromised in various mental health conditions, including depression, addiction, and schizophrenia [9]. The extent to which we can imagine, plan for, and enjoy thinking about the future affects the decisions we make now. This cognitive process (prospection) involves the ability to envision and plan for possible future outcomes, allowing individuals to adaptively integrate information from the past and present to simulate potential futures [10]. When used appropriately, prospection enables goal pursuit, facilitates flexible decision-making, and contributes to maintaining a coherent sense of personal identity [11].

Research suggests that individuals with stronger identification with their future self tend to make better decisions in daily life, such as exercising more frequently [12], staying on the right side of the law [13], achieving better academic grades [14], experiencing greater well-being [15], and attaining superior financial outcomes [16]. These benefits are understood as a consequence of the reduced difference between the future and present self, which results in improved inter-temporal decision-making—with individuals more likely to select "should" behaviors (eg, exercising, saving money, going to bed early) over "want" behaviors (eg, staying on the sofa, spending money, staying up late) [17]. Within the context of mental health, future-oriented therapies aim to promote goal-directed behavior [18]. Therapeutic approaches have long recognized the significance of targeting prospection to improve well-being. Cognitive behavioral therapy uses techniques that aim to readjust clients' overly pessimistic predictions by addressing cognitive distortions like catastrophizing and fortune-telling [19] and incorporates Socratic questioning to help clients develop more accurate predictions about their future [20]. Other prospective therapeutic practices with promising results include future-directed therapy [21], mental contrasting and implementation intentions [22], and goal-setting and planning techniques [23].

Using an LLM chatbot may further facilitate these future-oriented approaches by enabling users to engage with their future self, developing the capacity for intertemporal decision-making. These conversational agents may offer users support and guidance as and when they need it. However, due to the inherent opacity of LLMs, it is crucial to evaluate how they provide such support and, more significantly, how individuals interact with them.

While chatbots have been shown to provide valuable support through their 24/7 availability and nonjudgmental nature [24], most research on mental health interventions has relied on retrospective self-reports, which may overlook moment-to-moment variations in stress levels and emotional states [25]. Few studies have explored on-demand, chatbot-based interventions that incorporate prospection-based techniques, particularly for student populations who may benefit significantly from improved future-oriented thinking. Prior research points to several well-documented risk factors faced by postgraduate students, including academic pressures [26], financial concerns [27,28], future uncertainty [29], and social isolation [30], where prospective thinking may help to overcome these factors and improve the overall mental well-being.

To address these gaps, we developed "Future Me," a chatbot prototype that uses GPT-3.5 to facilitate future-oriented thinking by simulating a conversation with the user's future self. The interaction with Future Me aims to stimulate a future-oriented mindset by strengthening people's ability and motivation to consider their future and fostering identification with their future self.

This paper presents 2 complementary studies examining the usability, acceptability, and potential efficacy of Future Me:

- Study 1 investigates how postgraduate participants use Future Me during a single guided session, followed by interviews to assess their experience.
- Study 2 examines how postgraduate students interact with Future Me over a 1-week period, with interviews before and after usage to understand expectations and experiences.

Together, these studies aim to answer the following research questions: (1) How do people use Future Me, and what types of support do they seek from it? (2) In what ways does Future Me enable or support the process of future-oriented thinking? (3) How are the acceptability and usability of Future Me perceived by different user groups? (4) What are the specific perceptions and needs of postgraduate students regarding

XSL•FO
RenderX

chatbot-based support for stress management? (5) How can chatbot design be improved to better address users' mental well-being needs?

By exploring these questions, we aim to provide comprehensive insights into the potential of LLM-based chatbots for promoting prospection skills and supporting mental well-being and identify key design considerations for developing more effective digital mental health interventions.

Methods

Design Overview

Both studies used qualitative approaches to evaluate the usability and effectiveness of Future Me, a chatbot designed to facilitate future-oriented thinking. Study 1 focused on a single guided session, while Study 2 examined sustained usage over 1 week. Both studies used semistructured interviews and analyzed conversation transcripts between users and Future Me. We used the SRQR (Standards for Reporting Qualitative Research) checklist [31] for reporting our study (see Checklist 1).

Participants

Study 1 comprised a convenience sample of 20 postgraduate students (12 female, 6 male, and 2 nonbinary), aged 21 - 40 (mean 25.1, SD 4.39) years. Study 2 consisted of 14 postgraduate students (3 men and 11 women) aged 22 to 33 years (mean 23.71, SD 2.73) years.

Eligibility criteria included being 18 years or older, the ability to read and understand English, being based in the United Kingdom, having a personal mobile phone, and not currently experiencing mental health problems.

Participants were recruited for both studies through social media postings and advertisements on campus, outside of any clinical context.

The Future Me Prototype

Both studies used the same core Future Me chatbot prototype, which used GPT-3.5 to facilitate future-oriented thinking based on principles of Mental Contrasting with Implementation Intentions. Mental contrasting involves imagining a future desired outcome and reflecting on present obstacles that impede its attainment. By making the connection between current reality and the desired future more explicit, individuals gain a deeper understanding of the challenges they may face and the feasibility of their goals.

Implementation intentions involve "if... then..." thinking, which promotes consideration of anticipated obstacles, planning to overcome these obstacles, and, consequently, more successful goal pursuit. Together, these therapeutic techniques have been shown to consistently aid goal attainment and promote a wide range of beneficial outcomes.

Future Me was also designed to use Socratic questioning, a key feature of cognitive behavioral therapy, involving a series of graded questions to guide thought toward a therapeutic goal strategically. The aim of bringing these therapeutic techniques together was to enable users to link their current behavior and future consequences, facilitating skills in decision-making, goal-setting, and behavior change.

In both studies, participants interacted with Future Me by sending SMS messages to a registered phone number. Each message was stored on a local server, processed to create prompts for the GPT-3.5 model, and responses were sent back to the user's phone. The system used an adaptive prompt creation mechanism, which refined the initial manual prompt as conversations progressed using each user's message history to 'train' subsequent responses. This approach merges previous responses from the system and questions from the user into an expanded prompt sent back to the GPT-3.5 model to ensure that the responses can include previously discussed topics.

Participants were instructed that the prototype is not intended for crisis support, and external support resources (eg, contact information for crisis support services) were handed out during the onboarding for the experiments in case participants needed support. Due to the low-level nature of the prototype, Future Me was not able to sense or detect signs of distress (eg, measuring the user's heart rate).

Procedures

Study 1

The intervention involved a 15-minute interaction with Future Me via SMS text messaging on their phone in a quiet location chosen by the participants. The duration was considered long enough to have a meaningful interaction and assess the user experience in line with similar research [32]. Before involvement, participants were instructed that Future Me was designed to be open and self-directed, with the primary aim of facilitating thinking about the future and assisting in the exploration of prospective decisions, goals, and problems. To have relevant material for the conversation, participants were asked to come up with approximately 3 obstacles or decisions they were currently facing. Following the interaction, semistructured interviews were conducted with participants via video conference to assess dimensions of usability, acceptability, and signs of preliminary efficacy.

Study 2

The study was conducted in 3 phases.

Preinteraction Interviews

Semistructured interviews lasting approximately 45 minutes were conducted web-based via videoconference. These explored students' stress experiences, coping strategies, and expectations from digital mental health tools. Participants discussed their stress management routines, daily stressors, responses to the Perceived Stress Scale (PSS)-10, and attitudes toward digital mental health tools and chatbots.

Interaction With Future Me

After the first interviews, participants received instructions for using the Future Me chatbot. They were informed that the chatbot was not intended for crisis intervention and were provided with external support resources. Participants were asked to interact with Future Me over 1 week. After 7 days, they completed a second survey that included the PSS-10 and

XSL•FO
RenderX

questions about their experiences with Future Me. Participants were required to use the system at least once in the beginning but were otherwise free to use it as often as they wished.

Follow-Up Interviews

The final phase involved 25- to 30-minute follow-up interviews to assess the usability and effectiveness of the Future Me chatbot and explore changes in perceptions of digital mental health tools.

Measurements

Common Measures

Both studies used semistructured interviews to capture individuals' reactions to and experiences with Future Me. Interview questions were designed to assess dimensions of usability, acceptability, and preliminary efficacy. Additionally, both studies recorded and analyzed the conversation transcripts between participants and Future Me.

Study 2: Additional Measures

PSS-10

Perceived stress was measured using the PSS-10 [33], a widely used tool with high reliability. This 10-item scale assesses how often individuals have recently felt their lives were overwhelming, uncontrollable, and overloaded using a 5-point Likert scale. For this study, the timeframe was modified to measure stress over the past week rather than the past month, to correspond with the period participants interacted with Future Me.

Interaction Measurements

For each SMS message participants sent to the chatbot, the timestamp and message content were stored, along with the chatbot's response, allowing for analysis of usage patterns throughout the week.

Data Analysis

Conversation Analysis

In Study 1, conversations between participants and Future Me were analyzed using framework analysis with NVivo [34]. The thematic framework used to code participant inputs was based on the COM-B model for behavior change [35], which identifies 3 key requirements needed for behavior change: capabilities (C), opportunities (O), and motivations (M). The transtheoretical model of behavior change was used to develop the framework for coding Future Me's responses, which posits that behavior change occurs through stages: pre-contemplation, contemplation, preparation, action, maintenance, and termination [36].

In Study 2, a reflexive bottom-up thematic analysis was adopted to analyze conversation transcripts. The data were codified into key themes, which were then reviewed and consolidated into overarching themes. To ensure reliability, a second researcher independently reviewed the transcripts using a top-down analysis based on the initial codebook created by the first researcher.

Interview Analysis

In Study 1, interview transcripts were thematically analyzed using the method outlined by Braun and Clarke [37]. Coding was inductive, based on the frequency of sentiments across interviews.

In Study 2, a similar thematic analysis approach was used to assist in data coding and theme development. The interaction logs with Future Me were also analyzed to examine the average number and timing of interactions, assessing whether patterns changed throughout the week.

Reflexivity Statement

The researchers' backgrounds in psychology, human-computer interaction, and computer science could influence both chatbot design and analytical approach, potentially favoring technological solutions for mental health. Personal experiences with academic stress and mental health apps could also shape our assumptions about user needs. We implemented reflexive practices to mitigate these potential biases.

Ethical Considerations

Both studies were approved by the research ethics committee of University College London (ethics approval IDs: CEHP/2024/597 (Study 1) and UCLIC_2023_003_Dechant_Manning (Study 2) and followed a similar procedure for providing informed consent. All participants provided informed consent electronically before the beginning of each study. The consent form detailed the study purpose, procedures, potential risks and benefits, data privacy protection, and voluntary participation. Participants in both studies received a £20 (US $26.86) gift card after completion of the whole study.

All participant data were deidentified before analysis. Identifiable information, including phone numbers and email addresses, was removed before the analysis. Data were stored on encrypted servers compliant with institutional and GDPR (General Data Protection Regulation) guidelines.

Results

Study 1

User Inputs

Our analysis revealed that goals fell into 3 core categories: "Career/education," "Personal," and "Relationships." The "Career/education" theme occupied more than half of user inputs, aligning well with the COM-B model, with most inputs falling into the capability component. This was divided into 2 separate codes to recognize that participants inquired about both their professional capability (eg, skills, education, and work experience) and their emotional capacity to reach these professional capacities.

The second largest theme was "Personal," in which users expressed current obstacles in their lives and asked for advice to overcome these issues. Most inputs focused on internal factors, specifically psychological and behavioral aspects, falling under the "Capability" category, followed by "Motivation" and "Combination," with only a few falling into "Opportunity."

Within "Relationships," most inputs fell into the "Motivation" category, in which individuals described relationship dynamics and nuance that informed their desire to commit to a behavior

XSL•FO
RenderX

change or decision. The remaining inputs were coded under "Feasibility," which consisted of a combination of their capability and opportunity to action a behavior related to a personal relationship.

The final theme, "Anthropomorphizing," fell outside the COM-B framework and consisted of comments indicating the degree to which users were humanizing Future Me. Most frequently, participants used conversational markers and social cues similar to a typical human-human conversation (eg, introducing themselves, saying thank you). The other codes in "Anthropomorphizing" suggested that users were either keenly aware of Future Me being an AI and therefore did not use typical conversational cues ("expressionless tone") or directly questioned Future Me's capacity and limits ("curiosity").

Multimedia Appendix 1 summarizes the themes participants discussed during the usage of Future Me.

Future Me Outputs

The transtheoretical model was used to categorize Future Me's responses according to the presumed stage of change of the user. Responses were suggestive of users being in 2 out of the 6 stages—"Contemplation" and "Preparation."

Under "Contemplation," users identified a problem or goal and were evaluating whether and how to pursue it. Future Me primarily responded by "Inquiring" about the context of their decision, "Suggesting" information or techniques they could use, and "Evaluating" by highlighting relevant factors and nuances affecting a decision.

The second stage seen was "Preparation," in which users had committed to a decision and were given action-based steps. There were significantly fewer outputs in this stage compared to "Contemplation" (42 vs 176).

'Recognition' was present in both stages and formed its own theme. In this theme, Future Me reflected information received back to the user through "Acknowledgment" (neutral recognition or statement of understanding) and "Positive reinforcement" (recognition with a hopeful and encouraging tone).

Multimedia Appendix 1 outlines the themes of the answers generated by Future Me collected during the conversations with the chatbot.

Participant Interviews

Thematic analysis of the interviews revealed several key themes.

Initial Perceptions of Future Me

Participants had 3 main expectations before interacting with Future Me: "Novelty of AI as a therapeutic tool," "Interest" in testing AI capabilities or exploring personal issues, and "Skepticism" that Future Me would provide generic or vague responses.

Topics of Conversation

Participants engaged in multidimensional conversations, combining practical and professional elements with personal preferences and motivations. Many discussed only professional goals, believing Future Me was best suited for objective and unemotional advice and expressed skepticism about using it for personal matters due to perceived limitations of an AI not being human.

Positive Perceptions of Future Me

The main positive aspect mentioned was that Future Me provided "Clarity and practicality" (85%), helping with idea mapping, exploring options, and evaluating pros and cons. Participants also valued how Future Me facilitated personal reflection by asking questions rather than providing answers (80%), finding it "unbiased" and "nonjudgmental."

Additionally, 70% reported that Future Me provided new information and perspectives, particularly about career paths and techniques for personal issues. Participants found Future Me "reaffirming and consolidating" (65%) prior knowledge regarding personal obstacles, making them feel "reassured" and more "confident."

In "Comparison to human conversation" (65%), participants described the interaction as "easy" and "natural," noting that Future Me exhibited features traditionally associated with being human, such as "emotional intelligence" and "empathy." The "Text message format" (50%) was appreciated for its accessibility and comfort, making the conversation feel less formal and more human.

Negative Perceptions of Future Me

The main limitation mentioned was "User-dependence" (70%), with users feeling they had to take responsibility for guiding the conversation. The main criticism (60%) was that Future Me provided "formulaic" or overly general responses, though some acknowledged that advice from human sources can also be vague.

Half the participants thought Future Me was "lacking something human" (50%), finding answers less meaningful due to a lack of "compassion," "intention," or "empathy." Some participants (45%) felt uneasy about the "Conversation flow," particularly the fluctuating response time that disrupted normal conversation patterns, leading to feelings described as "disconcerting" and "frustrated."

Multimedia Appendix 1 outlines the positive and negative themes found in the user feedback collected after the interaction with Future Me.

Study 2: Results

General Perceptions and Needs for Stress Management From Chatbots

Multimedia Appendix 1 outlines the themes from the preintervention interview transcripts.

Trust and Emotional Openness

While students found chatbots useful for immediate advice and practical support, they highlighted a lack of emotional depth as a significant limitation, particularly in high-stress situations. Participants expressed that, during moments of heightened stress, genuine compassion and nuanced emotional support were critical qualities they felt chatbots could not provide, leading to a reluctance among some to rely on chatbots for deep emotional support.

XSL•FO
RenderX

Concerns over data privacy also influenced students' trust in chatbots, particularly regarding integration within university systems. Several participants expressed hesitation about the possibility of their conversations being accessed by university staff, fearing potential academic or professional implications.

Role and Effectiveness

The potential for chatbots to detect early signs of stress was viewed positively, with students noting that early warnings could help prevent stress from escalating. However, there was a widespread perception that chatbots lacked the emotional depth necessary for high-stress situations, leading to a reluctance among some students to use chatbots for emotional support.

Personalization and Contextual Awareness

Students expressed a need for chatbots to provide personalized support tailored to their specific academic stressors, envisioning chatbots having detailed background knowledge of their courses for more targeted advice. They highlighted the need for support appropriately tailored to their current stress levels, emphasizing that different situations require different responses.

Support Preferences

Students consistently valued the 24/7 accessibility of chatbots, highlighting the convenience of having support available at any time. Anonymity was also a significant factor, with many noting that it encouraged more openness in sharing experiences. The non-judgmental nature of chatbot interactions was cited as reducing the pressure to filter thoughts, making it easier to discuss sensitive topics.

Expressions of Stress and Support Sought From a Chatbot in Real Time

Multimedia Appendix 1 outlines the themes of participants' inputs during conversations with Future Me.

Academic Stress and Time Management Advice

Postgraduate students, particularly those in their final term, experienced significant stress related to academic demands, with a strong focus on dissertation work. They sought practical, immediate relief through quick advice and activities, appreciating Future Me's swift SMS responses for short-term stress management.

Support to Tackle Negative Emotions and Low Motivation

Students reported emotional distress linked to academic workload and personal challenges. They found self-reflective questions useful for future uncertainties but frustrating when applied to immediate academic stress, as they felt these questions were time-consuming and added to their emotional burden.

Future Uncertainty and Advice on How to Handle It

International students expressed heightened stress about future uncertainties, particularly concerning finances, student loans, and job prospects before visa expiration. This theme was more pronounced among international students compared to their domestic peers.

Advice About Social Relationships

Although not as prominently discussed, social relationships were indirectly mentioned in the context of overall stress. The focus was primarily on academic and future-related concerns rather than detailed issues with social relationships.

Future Me Engagement Metrics

Participants' interactions with Future Me were spread throughout the day, with 49 out of 115 chatbot interactions occurring between 9 AM and 5 PM, and the remaining 66 occurring outside the typical work hours. Figure 1 shows the distribution of requests over time. Participants sent most of the requests at the beginning of the experiment (N=40), with a linear decrease in the interaction over the week (see Figure 2).

A paired-samples *t* test of PSS scores before (mean 20.25, SD 5.96) and after (mean 17.67, SD 6.51) using Future Me for 1 week showed no significant difference (t_{11}=1.765, *P*=.105).

XSL•FO
RenderX

Figure 1. Frequency distribution of requests per hour over the full experiment.

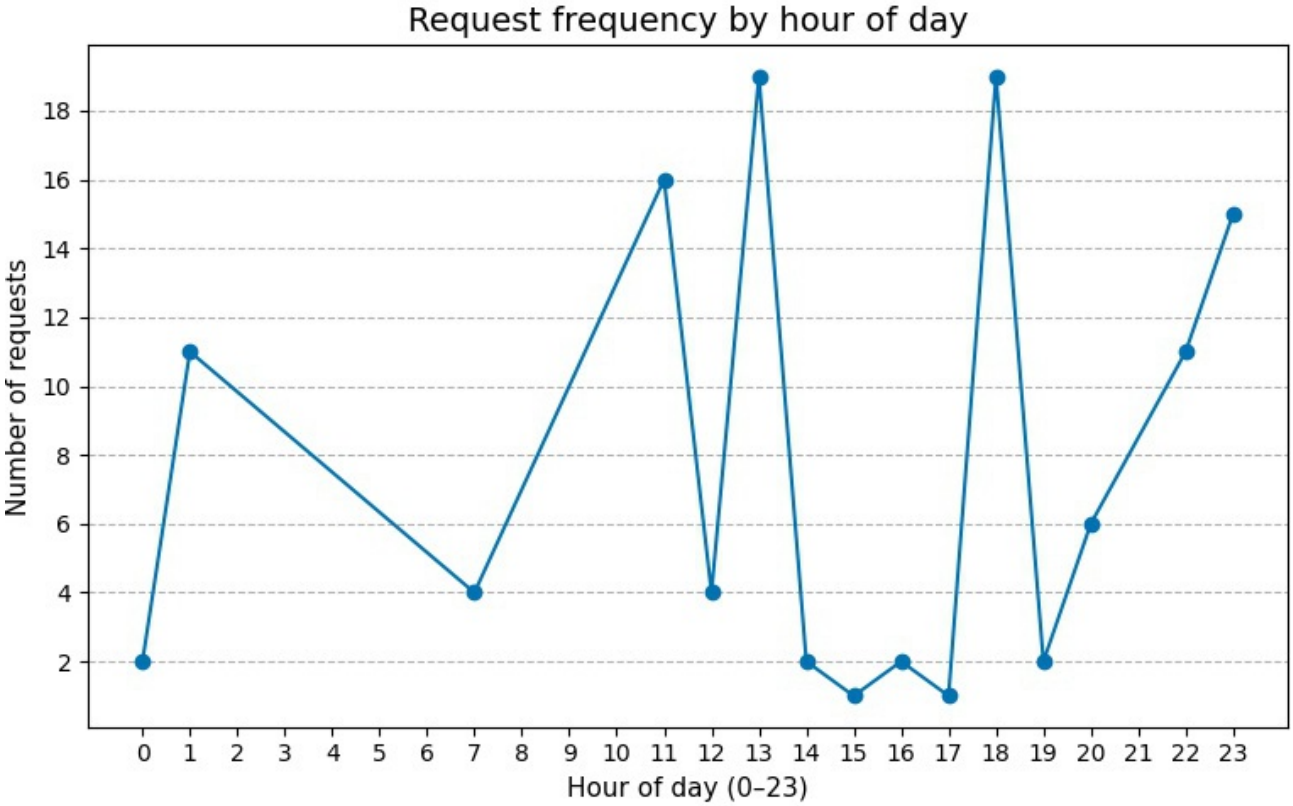

Figure 2. Distribution of requests sent to the chatbot via SMS per day.

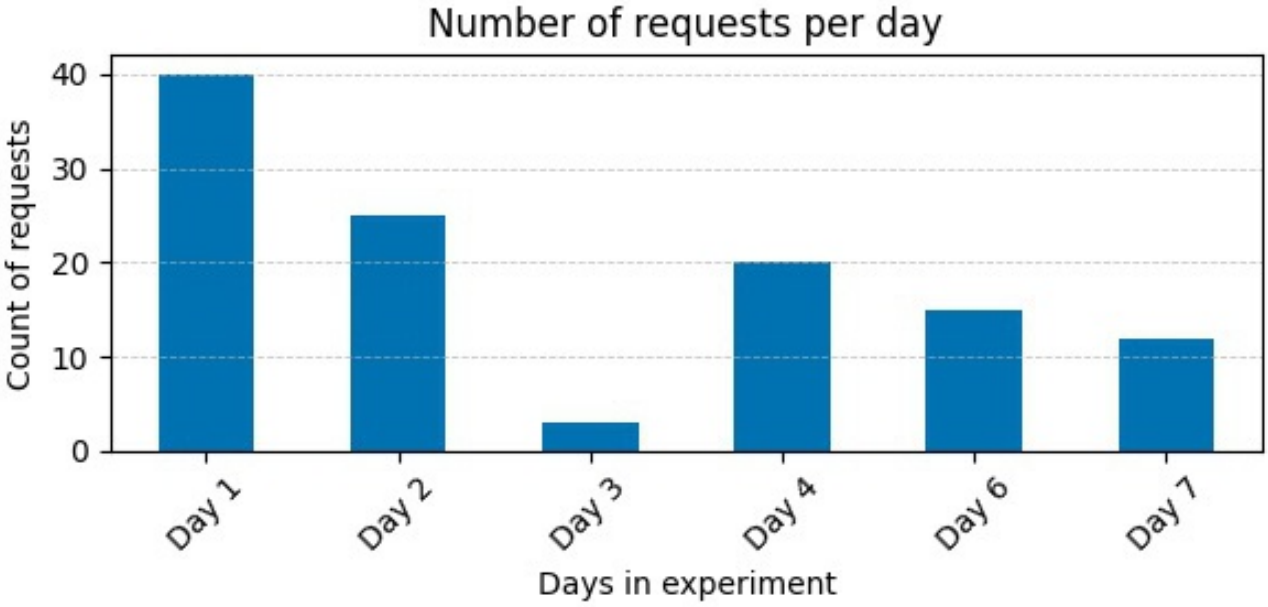

User Engagement, Specific Perceptions, and Experiences of Using Chatbots for Stress Management

Multimedia Appendix 1 outlines the themes revealed from the follow-up interview transcripts.

Accessibility and Convenience

Postgraduate students found Future Me highly accessible and convenient, particularly outside working hours. They appreciated its personalized interactions and conversational tone, comparing the experience to speaking with a friend. In total, 3 students noted that simply expressing their current situations to Future Me sometimes helped relieve stress, similar to venting to friends.

Personalized and Relatable

Students valued the practical advice Future Me offered but felt there was a gap in its ability to provide empathetic support during high-stress times. They appreciated Future Me's reflective questions for broader issues like future uncertainty but preferred these questions to focus more on understanding emotions during immediate concerns. When reflective questions

XSL•FO
RenderX

were used during acute stress, students found the responses scripted and impersonal, leading to frustration.

Need for Better Balance Between Advice and Emotional Support

Several students found it difficult to express themselves clearly to Future Me during high-stress periods, noting that humans can often fill in blanks using body language and facial expressions. Students emphasized that they would be more open to using chatbots during high-stress periods if there were other ways to interpret their emotions, suggesting voice systems to make expression easier.

Increased Demand for Context Awareness

Students called for Future Me to be more context-aware and responsive to their immediate emotional states. They felt that integrating Future Me with professional mental health services could provide a more comprehensive support system, allowing the chatbot to assist with less serious issues while referring students to appropriate professionals for specialized support.

Connection With Human Support Systems

Students viewed Future Me as a valuable complement to human support rather than a replacement. They saw a need for integrating the chatbot with professional mental health services and believed it could help navigate the complexity of university services by referring them to appropriate professionals. While generally willing to share their chatbot data to streamline this process, they preferred those referrals be handled sensitively to avoid feeling dismissed.

Discussion

Summary of Findings

Both studies provide valuable insights into how users interact with Future Me, a chatbot designed to facilitate future-oriented thinking through simulated conversations with one's future self. While Study 1 focused on a single guided session with non-clinical adults, Study 2 examined sustained usage over 1 week among postgraduate students. These preliminary investigations point toward several promising directions for further development and research.

First, both studies demonstrate that users primarily engaged with Future Me to discuss career and education goals, personal obstacles, and, to a lesser extent, social relationship concerns. The prevalence of career-related topics suggests that users find the chatbot particularly suitable for objective, goal-oriented discussions that benefit from structured future thinking. This aligns with findings from both studies and prior work [38] that participants viewed Future Me as less suitable for emotionally complex or high-stress situations that might require more nuanced human support.

Second, both studies revealed that Future Me effectively facilitated 2 key processes: guided introspective questioning and collaborative solution generation. Analysis of conversation transcripts in Study 1 showed that Future Me's responses primarily fell under the "contemplation" stage of the transtheoretical model, helping users gather information, evaluate options, and move toward preparation for action. Similarly, Study 2 found that Future Me encouraged reflection and offered new perspectives, particularly valuable for addressing broader concerns like future uncertainty.

Third, users across both studies appreciated similar qualities in Future Me. Nonclinical participants in Study 1 and students in Study 2 both valued the chatbot's ability to provide clarity around goal setting [39], its nonjudgmental nature, and its accessibility through SMS text messaging [40,41]. Both groups also found value in Future Me's ability to facilitate personal reflection without the influence of others' perspectives or judgment [42,43].

However, both studies also identified similar limitations. Participants in both samples noted that Future Me sometimes provided formulaic or generic responses that lacked personalization. Users in both studies felt that the chatbot lacked something inherently human, despite its ability to generate conversational responses. This was particularly evident during high-stress moments in Study 2, where students found reflective questions, although important for managing stress, frustrating when seeking immediate emotional support.

Implications for Design

Balancing Reflective and Directive Approaches

A key direction emerging from both studies is the need for chatbots to balance reflective questioning with more directive guidance based on the user's emotional state and immediate needs. While the Socratic questioning and prospective thinking techniques used by chatbots, like Future Me, were generally well-received in prior work [44], our studies indicate that these approaches may be less effective during periods of acute stress or when users seek immediate practical advice.

Our findings strongly advocate for mental health chatbot designers to adopt a more adaptive and nuanced approach that considers both the user's emotional state and their surrounding context, including cultural factors [45]. This perspective aligns with prior research demonstrating that the chatbot's communication style significantly influences user acceptance [46]. Whenever a system detects signs of high stress or anxiety using various techniques [47], the chatbot could temporarily shift from reflection-promoting questions to more direct, solution-focused responses that address immediate concerns. Once the acute stress has been acknowledged and addressed, the chatbot could then gradually reintroduce reflective elements to promote longer-term thinking and behavior change.

Enhancing Personalization and Context Awareness

Both studies highlighted the importance of personalization in maintaining user engagement. Study 1 found that 70% of participants appreciated when Future Me provided new information specific to their situation, while Study 2 revealed that students wanted the chatbot to understand their academic context and tailor support accordingly.

Future iterations of mental health chatbots could substantially enhance personalization by implementing a layered approach to user understanding. By developing progressive profiling capabilities, these systems could build increasingly nuanced

XSL•FO
RenderX

user models over time, remembering important details from previous conversations and creating a more continuous experience. This historical awareness would allow the chatbot to reference past concerns or achievements, creating a sense of ongoing relationships rather than disconnected interactions. Alongside this individual history, context integration would enable chatbots to recognize the broader circumstances affecting users' well-being. For instance, a system aware of university examination periods or job application deadlines could proactively adjust its support approach during these high-stress times, offering relevant coping strategies without users needing to explain their situation repeatedly.

Integration With Human Support Systems

Study 2 particularly emphasized the potential value of integrating chatbots within broader support systems. Participating students viewed Future Me as a complement to human support rather than a replacement, suggesting that mental health chatbots could serve as an initial point of contact that guides users toward appropriate professional services when needed.

This finding has important implications for institutional implementation of mental health chatbots, particularly within university settings. Rather than positioning chatbots as stand-alone solutions, designers should consider how these tools can form part of a comprehensive support ecosystem. This might include features that facilitate warm handoffs to human professionals, provide information about available resources, or offer continuous support between professional sessions.

Theoretical Implications

Prospection and Digital Mental Health Support

Both studies contribute to our understanding of how digital tools can facilitate prospection skills. The findings suggest that LLM-based chatbots can effectively support aspects of future-oriented thinking by helping users clarify goals, identify obstacles, and develop implementation intentions. This aligns with existing literature on the benefits of prospection for mental well-being and extends it by demonstrating the potential of AI-mediated conversations to strengthen future self-identification.

However, the studies also reveal nuances in how prospection techniques should be applied in digital interventions. While these approaches appear beneficial for addressing broader life decisions and goals, they may be less effective during moments of acute stress when immediate relief is the priority. This suggests that prospection-based digital interventions should be designed with sensitivity to timing and emotional context.

Anthropomorphism and Therapeutic Alliance

Both studies provide insights into how users anthropomorphize and form relationships with AI chatbots. Study 1 found that 65% of participants compared their interaction with Future Me to human conversation, noting qualities like "emotional intelligence" and "empathy." Similarly, Study 2 revealed that students compared talking to Future Me to speaking with a friend.

These findings suggest that users can form a type of therapeutic alliance with chatbots, even while remaining aware of their nonhuman nature. The nonjudgmental quality of chatbot interactions appears particularly valuable, potentially reducing barriers to self-disclosure that might exist in human relationships. However, both studies also found that users perceived limitations in the emotional depth and authenticity of the chatbot, highlighting the complex nature of human-AI therapeutic relationships.

Ethical Concerns

The use of artificial intelligence in mental health is challenging traditional practices and offering novel approaches to address existing barriers. However, these innovations also raise significant ethical concerns that clinicians and researchers must carefully consider. One major issue is bias in AI models, which are often trained on datasets that underrepresent diverse populations. This can lead to neglect of users' cultural and personal needs, potentially compromising the quality and relevance of support provided. For example, in the context of Future Me, overlooking cultural nuances may harm users' sense of identity and well-being [48]. Privacy is another critical concern. AI-based mental health tools inherently handle sensitive personal data, raising risks around data security and user confidentiality [49]. Moreover, safeguarding is essential in AI systems used for mental health. Studies have shown that systems like ChatGPT can, under certain prompt conditions, generate harmful content [50,51]. Despite ongoing efforts to enhance safety mechanisms, AI may still fail to detect subtle cues related to self-harm or distress, posing risks to vulnerable users.

Limitations and Future Directions

Several limitations should be considered when interpreting the findings from these studies. First, both studies used relatively small, non-representative samples, limiting the generalizability of findings. Study 1 used convenience sampling with non-clinical adults in the United Kingdom, while Study 2 focused specifically on postgraduate students, with international students overrepresented.

Second, the controlled research context may have influenced how participants engaged with Future Me. In both studies, participants were aware they were part of a research trial, which may have affected their privacy concerns and usage patterns. Future research should examine more naturalistic usage in nonresearch contexts to better understand how users would engage with the chatbot in everyday life.

Third, the timeframes in both studies were relatively short—a single session in Study 1 and 1 week in Study 2. While Study 2 provided valuable insights into sustained engagement, longer-term studies are needed to understand how usage patterns evolve over time and whether benefits are maintained. Previous research has indicated that online mental health interventions often suffer from low retention rates, making longitudinal studies particularly important for evaluating practical utility.

Fourth, Future Me was only tested with English as the input language. Future work may explore other languages and

XSL•FO
RenderX

multimodal ways, such as voice and visual inputs, to increase the accessibility of the approach.

Fifth, *Future Me* relied on GPT-3.5, a model known to exhibit various biases—gender, racial, cultural, linguistic, and ideological [52]. Users could not choose a more suitable or nuanced model tailored to their needs, leaving them exposed to these underlying biases of GPT 3.5. Additionally, *Future Me* could only gather information from user input during conversations, lacking any mechanism to collect essential contextual data such as gender, age, or personal preferences. This limitation further increased the risk of biased or inappropriate responses.

Future research should address these limitations while pursuing several promising avenues. Longer-term studies tracking users over 8 - 12 weeks would yield more comprehensive insights into engagement patterns, attrition rates, and the sustainability of benefits beyond initial interactions. Additionally, expanding testing to diverse populations, including those with clinical conditions and from varied demographic backgrounds, would enhance our understanding of how Future Me's benefits and usage patterns might differ across different contexts and user needs. The limitations of text-only communication, particularly during high-stress moments, could be addressed through multimodal interfaces, making it not only easier to communicate with the device but also giving more data to detect signs of mental needs such as high stress [53] or anxiety [54]. Such approaches might better capture emotional nuances and reduce the burden on users when articulating complex feelings. Integration studies represent another crucial direction, examining how Future Me could function within existing mental health ecosystems to complement professional services and therapies, potentially serving as a bridge between self-help and clinical intervention. Finally, comparative effectiveness research directly measuring Future Me against other digital interventions and traditional therapeutic approaches would provide valuable insights into its unique contributions and limitations. Such comparisons would help situate Future Me within the broader landscape of mental health support tools and clarify its most appropriate applications and user groups. Together, these research directions would substantially advance our understanding of how LLM-based chatbots can effectively support mental well-being through prospection-focused interventions.

Conclusions

The 2 studies presented in this paper provide complementary insights into the potential of LLM-powered chatbots for promoting prospection skills and supporting mental well-being. Future Me demonstrates promise as an accessible, non-judgmental tool that can facilitate reflection, provide practical guidance, and help users connect their present actions to future outcomes. The findings suggest that such chatbots can fulfill a valuable supportive role, particularly when human alternatives are unavailable or when users prefer anonymous, pressure-free interaction.

However, the studies also highlight important limitations and design considerations. Future Me's effectiveness appears context-dependent, with prospection techniques more suitable for broader life decisions than acute stress situations. Users appreciated the chatbot's accessibility and reflective capabilities but noted limitations in emotional depth and personalization that affected sustained engagement.

These findings suggest that future development of mental health chatbots should focus on creating more adaptive, context-aware systems that can adjust their approach based on the user's emotional state and immediate needs. Rather than attempting to replicate human therapy entirely, chatbots like Future Me may be most effective when designed as complementary tools within broader support ecosystems, offering immediate, accessible reflection and guidance while facilitating connections to human support when needed.

As LLM technology continues to advance, the potential for chatbots to provide increasingly personalized and adaptive mental health support will likely grow. Future Me represents an important early exploration of how these technologies can be harnessed to promote prospection skills and support well-being, providing a foundation for continued research and development in this promising field.

Acknowledgments

We thank our participants for their great insights, time, and effort they put into participating in our study.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Summary of all quotes in tables.
[DOCX File, 183 KB - formative_v9i1e74411_app1.docx]

Checklist 1

XSL•FO
RenderX

SRQR (Standards for Reporting Qualitative Research) checklist.
[DOCX File, 158 KB - formative_v9i1e74411_app2.docx]

References

1. Abd-Alrazaq AA, Rababeh A, Alajlani M, Bewick BM, Househ M. Effectiveness and safety of using chatbots to improve mental health: systematic review and meta-analysis. J Med Internet Res 2020 Jul 13;22(7):e16021. [doi: 10.2196/16021] [Medline: 32673216]
2. Christiansen S, Oettingen G, Dahme B, Klinger R. A short goal-pursuit intervention to improve physical capacity: a randomized clinical trial in chronic back pain patients. Pain 2010 Jun;149(3):444-452. [doi: 10.1016/j.pain.2009.12.015] [Medline: 20199846]
3. Chung NC, Dyer G, Brocki L. Challenges of large language models for mental health counseling. arXiv. Preprint posted online on Nov 23, 2023. [doi: 10.48550/arXiv.2311.13857]
4. Huckvale K, Nicholas J, Torous J, Larsen ME. Smartphone apps for the treatment of mental health conditions: status and considerations. Curr Opin Psychol 2020 Dec;36:65-70. [doi: 10.1016/j.copsyc.2020.04.008] [Medline: 32553848]
5. Ramsey A, Lord S, Torrey J, Marsch L, Lardiere M. Paving the way to successful implementation: identifying key barriers to use of technology-based therapeutic tools for behavioral health care. J Behav Health Serv Res 2016 Jan;43(1):54-70. [doi: 10.1007/s11414-014-9436-5] [Medline: 25192755]
6. Andrews B, Wilding JM. The relation of depression and anxiety to life-stress and achievement in students. Br J Psychol 2004 Nov;95(Pt 4):509-521. [doi: 10.1348/0007126042369802] [Medline: 15527535]
7. Bewick B, Koutsopoulou G, Miles J, Slaa E, Barkham M. Changes in undergraduate students' psychological well‐being as they progress through university. Studies in Higher Education 2010 Sep;35(6):633-645. [doi: 10.1080/03075070903216643]
8. Tzelios C, Contreras C, Istenes B, et al. Using digital chatbots to close gaps in healthcare access during the COVID-19 pandemic. Public Health Action 2022 Dec 21;12(4):180-185. [doi: 10.5588/pha.22.0046] [Medline: 36561900]
9. Hallford DJ, Austin DW, Takano K, Raes F. Psychopathology and episodic future thinking: a systematic review and meta-analysis of specificity and episodic detail. Behav Res Ther 2018 Mar;102(November 2017):42-51. [doi: 10.1016/j.brat.2018.01.003] [Medline: 29328948]
10. Gilbert DT, Wilson TD. Prospection: experiencing the future. Science 2007 Sep 7;317(5843):1351-1354. [doi: 10.1126/science.1144161] [Medline: 17823345]
11. O'Donnell S, Oluyomi Daniel T, Epstein LH. Does goal relevant episodic future thinking amplify the effect on delay discounting? Conscious Cogn 2017 May;51:10-16. [doi: 10.1016/j.concog.2017.02.014]
12. Rutchick AM, Slepian ML, Reyes MO, Pleskus LN, Hershfield HE. Future self-continuity is associated with improved health and increases exercise behavior. J Exp Psychol Appl 2018 Mar;24(1):72-80. [doi: 10.1037/xap0000153] [Medline: 29595304]
13. Hershfield HE. Future self-continuity: how conceptions of the future self transform intertemporal choice. Ann N Y Acad Sci 2011 Oct;1235(1):30-43. [doi: 10.1111/j.1749-6632.2011.06201.x] [Medline: 22023566]
14. Bixter MT, McMichael SL, Bunker CJ, et al. A test of a triadic conceptualization of future self-identification. PLoS ONE 2020;15(11):e0242504. [doi: 10.1371/journal.pone.0242504] [Medline: 33232353]
15. Reiff JS, Hershfield HE, Quoidbach J. Identity over time: perceived similarity between selves predicts well-being 10 years later. Soc Psychol Personal Sci 2020 Mar;11(2):160-167. [doi: 10.1177/1948550619843931]
16. Ersner-Hershfield H, Garton MT, Ballard K, Samanez-Larkin GR, Knutson B. Don't stop thinking about tomorrow: individual differences in future self-continuity account for saving. Judgm Decis Mak 2009 Jun 1;4(4):280-286. [doi: 10.1017/S1930297500003855] [Medline: 19774230]
17. Duckworth AL, Milkman KL, Laibson D. Beyond willpower: strategies for reducing failures of self-control. Psychol Sci Public Interest 2018 Dec;19(3):102-129. [doi: 10.1177/1529100618821893] [Medline: 30760176]
18. Mertens ECA, van der Schalk J, Siezenga AM, van Gelder JL. Stimulating a future-oriented mindset and goal attainment through a smartphone-based intervention: study protocol for a randomized controlled trial. Internet Interv 2022 Mar;27(October 2021):100509. [doi: 10.1016/j.invent.2022.100509] [Medline: 35242590]
19. Gautam M, Tripathi A, Deshmukh D, Gaur M. Cognitive behavioral therapy for depression. Indian J Psychiatry 2020 Jan;62(Suppl 2):S223-S229. [doi: 10.4103/psychiatry.IndianJPsychiatry_772_19] [Medline: 32055065]
20. Carey TA, Mullan RJ. What is Socratic questioning? PTRPT 2004;41(3):217-226. [doi: 10.1037/0033-3204.41.3.217]
21. Vilhauer J, Cortes JJ, Moali NN, Chung SS, Mirocha JJ, Ishak WW. Improving quality of life for patients with major depressive disorder by increasing hope and positive expectations with future directed therapy (FDT). Innov Clin Neurosci 2013 Mar;10(3):12-22.
22. Oettingen G. Future thought and behaviour change. Eur Rev Soc Psychol 2012 Mar;23(1):1-63. [doi: 10.1080/10463283.2011.643698]
23. West R, Michie S. A brief introduction to the COM-b model of behaviour and the PRIME theory of motivation. Qeios Preprint posted online on 2020. [doi: 10.32388/WW04E6]

XSL•FO
RenderX

24. Ejim OC, Livanou M, Khan H, Lindenmeyer A, Uwom C, Manaseki-Holland S. Depression, anxiety and stress among international postgraduate students in a UK university: a cross-sectional study. In Review. Preprint posted online on 2021. [doi: 10.21203/rs.3.rs-287018/v1]
25. Manning A, Cronin P, Monaghan A, Rawlings-Anderson K. Supporting students in practice: an exploration of reflective groups as a means of support. Nurse Educ Pract 2009 May;9(3):176-183. [doi: 10.1016/j.nepr.2008.07.001] [Medline: 18786861]
26. Woolston C. PhDs: the tortuous truth. Nature New Biol 2019 Nov 14;575(7782):403-406. [doi: 10.1038/d41586-019-03459-7]
27. Webster P, North SE. Health professions educational debt: personal, professional, and psychological impacts 5 years post-graduation. Front Med (Lausanne) 2022;9:746463. [doi: 10.3389/fmed.2022.746463] [Medline: 35223888]
28. Jamilah A, Haque M, Muhammad F, et al. Depression and associated factors among international students in a private university of Bangladesh. Global Psychiatry 2020. [doi: 10.2478/gp-2020-0021] [Medline: 33294191]
29. Chi T, Cheng L, Zhang Z. Global prevalence and trend of anxiety among graduate students: a systematic review and meta-analysis. Brain Behav 2023 Apr;13(4):e2909. [doi: 10.1002/brb3.2909] [Medline: 36852520]
30. Vasileiou K, Barnett J, Barreto M, et al. Coping with loneliness at University: a qualitative interview study with students in the UK. Ment Health Prev 2019 Mar;13:21-30. [doi: 10.1016/j.mhp.2018.11.002]
31. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014 Sep;89(9):1245-1251. [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]
32. Entenberg GA, Dosovitsky G, Aghakhani S, et al. User experience with a parenting chatbot micro intervention. Front Digit Health 2022;4:989022. [doi: 10.3389/fdgth.2022.989022] [Medline: 36714612]
33. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. J Health Soc Behav 1983 Dec;24(4):385-396. [Medline: 6668417]
34. NVivo 15 - the most trusted qualitative analysis software (QDA) is even better. Lumivero. 2023. URL: www.lumivero.com [accessed 2025-07-14]
35. West R, Michie S. A brief introduction to the COM-b model of behaviour and the PRIME theory of motivation. Qeios 2020. [doi: 10.32388/WW04E6.2]
36. Smith KJ, Subich LM, Kalodner C. The transtheoretical model's stages and processes of change and their relation to premature termination. J Couns Psychol 1995;42(1):34-39. [doi: 10.1037//0022-0167.42.1.34]
37. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
38. Yokotani K, Takagi G, Wakashima K. Advantages of virtual agents over clinical psychologists during comprehensive mental health interviews using a mixed methods design. Comput Human Behav 2018 Aug;85:135-145. [doi: 10.1016/j.chb.2018.03.045]
39. Vowels LM, Francois-Walcott RRR, Darwiche J. AI in relationship counselling: evaluating ChatGPT's therapeutic capabilities in providing relationship advice. Comput Hum Behav Artif Hum 2024 Aug;2(2):100078. [doi: 10.1016/j.chbah.2024.100078]
40. de Gennaro M, Krumhuber EG, Lucas G. Effectiveness of an empathic chatbot in combating adverse effects of social exclusion on mood. Front Psychol 2019;10:3061. [doi: 10.3389/fpsyg.2019.03061] [Medline: 32038415]
41. Derksen F, Bensing J, Lagro-Janssen A. Effectiveness of empathy in general practice: a systematic review. Br J Gen Pract 2013 Jan;63(606):e76-e84. [doi: 10.3399/bjgp13X660814] [Medline: 23336477]
42. Kettle L, Lee YC. User experiences of well-being chatbots. Hum Factors 2024 Jun;66(6):1703-1723. [doi: 10.1177/00187208231162453]
43. Nadarzynski T, Miles O, Cowie A, Ridge D. Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: a mixed-methods study. Digit Health 2019;5(5):2055207619871808. [doi: 10.1177/2055207619871808] [Medline: 31467682]
44. Pataranutaporn P, Winson K, Yin P, et al. Future you: a conversation with an AI-generated future self reduces anxiety, negative emotions, and increases future self-continuity. : IEEE Presented at: 2024 IEEE Frontiers in Education Conference (FIE); May 11, 2024; Washington, DC, USA p. 1-10. [doi: 10.1109/FIE61694.2024.10893443]
45. Ng SH, Soon LK, Su TT, Association for Computing Machinery. Emotion-aware chatbot with cultural adaptation for mitigating work-related stress. 2023 Apr 28 Presented at: Asian CHI '23: Proceedings of the Asian HCI Symposium 2023; Oct 3, 2023; Online Indonesia p. 41-50. [doi: 10.1145/3604571.3604578]
46. Xiao R, Yazan M, Situmeang FBI. Rethinking conversation styles of chatbots from the customer perspective: relationships between conversation styles of chatbots, chatbot acceptance, and perceived tie strength and perceived risk. Int J Hum Comput Interact 2025 Jan 17;41(2):1343-1363. [doi: 10.1080/10447318.2024.2314348]
47. Dobson R, Li LL, Garner K, Tane T, McCool J, Whittaker R. The use of sensors to detect anxiety for in-the-moment intervention: scoping review. JMIR Ment Health 2023 Feb 2;10:e42611. [doi: 10.2196/42611] [Medline: 36729590]
48. James S, Prilleltensky I. Cultural diversity and mental health. Towards integrative practice. Clin Psychol Rev 2002 Nov;22(8):1133-1154. [doi: 10.1016/s0272-7358(02)00102-2] [Medline: 12436808]
49. Wu X, Duan R, Ni J. Unveiling security, privacy, and ethical concerns of ChatGPT. J Inf Intell 2024 Mar;2(2):102-115. [doi: 10.1016/j.jiixd.2023.10.007]
50. Blease C, Torous J. ChatGPT and mental healthcare: balancing benefits with risks of harms. BMJ Ment Health 2023 Nov;26(1):e300884. [doi: 10.1136/bmjment-2023-300884] [Medline: 37949485]

XSL•FO
RenderX

51. Kalam KT, Rahman JM, Islam MR, Dewan SMR. ChatGPT and mental health: friends or foes? Health Sci Rep 2024 Feb;7(2):e1912. [doi: 10.1002/hsr2.1912] [Medline: 38361805]
52. Ray PP. ChatGPT: a comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. IoT CPS 2023;3:121-154. [doi: 10.1016/j.iotcps.2023.04.003]
53. Pisanski K, Sorokowski P. Human stress detection: cortisol levels in stressed speakers predict voice-based judgments of stress. Perception 2021 Jan;50(1):80-87. [doi: 10.1177/0301006620978378] [Medline: 33302780]
54. Riad R, Denais M, de Gennes M, et al. Automated speech analysis for risk detection of depression, anxiety, insomnia, and fatigue: algorithm development and validation study. J Med Internet Res 2024 Oct 31;26:e58572. [doi: 10.2196/58572] [Medline: 39324329]

Abbreviations

GDPR: General Data Protection Regulation
LLM: large language model
PSS-10: Perceived Stress Scale
SRQR: Standards for Reporting Qualitative Research

Edited by A Mavragani; submitted 24.03.25; peer-reviewed by J Abua, M Adnan; revised version received 14.05.25; accepted 21.05.25; published 18.07.25.

Please cite as:
Dechant M, Lash E, Shokr S, O'Driscoll C
Future Me, a Prospection-Based Chatbot to Promote Mental Well-Being in Youth: Two Exploratory User Experience Studies
JMIR Form Res 2025;9:e74411
URL: https://formative.jmir.org/2025/1/e74411
doi:10.2196/74411

© Martin Dechant, Eva Lash, Sarah Shokr, Ciarán O'Driscoll. Originally published in JMIR Formative Research (https://formative.jmir.org), 18.7.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Use of a Wearable Self-Tracking Instrument by Refugees With Complex Posttraumatic Stress Disorder: A Qualitative Study of Psychotherapeutic Mediation and Engagement

Lisa Groenberg Riisager[1,2], PhD; Stine Bjerrum Moeller[1,2], PhD; Jakob Eg Larsen[3], PhD; Thomas Blomseth Christiansen[4], MSc; Jesper Aagaard[5], PhD; Lotte Huniche[1], PhD

[1]Department of Psychology, Faculty of Health Sciences, University of Southern Denmark, Campusvej 55, Odense, Denmark
[2]Department of Multidisciplinary Trauma Treatment, Mental Health Services in the Region of Southern Denmark, Middelfart, Denmark
[3]Department of Applied Mathematics and Computer Science Cognitive Systems, Technical University of Denmark, Kongens Lyngby, Denmark
[4]Konsulent Blomseth, Hjoerring, Denmark
[5]Department of Psychology and Behavioral Sciences, Aarhus University, Aarhus, Denmark

Corresponding Author:
Lisa Groenberg Riisager, PhD
Department of Psychology, Faculty of Health Sciences, University of Southern Denmark, Campusvej 55, Odense, Denmark

Abstract

Background: Wearable self-tracking technologies are increasingly used in mental health care to enhance engagement and personalize treatment. However, most existing instruments focus on passive data collection or predefined symptom monitoring. Less attention has been given to tools that enable patients to actively track personally meaningful, self-defined mental health experiences as part of psychotherapy, particularly in vulnerable populations such as refugees with complex posttraumatic stress disorder (CPTSD).

Objective: This study aimed to explore how the One Button Tracker (OBT), a novel single-purpose wearable self-tracking instrument that enables in-the-moment active registration of self-defined, personally relevant mental health phenomena, supports therapeutic engagement among refugees receiving psychotherapeutic treatment for CPTSD.

Methods: This qualitative study was part of a larger participatory action research project conducted from 2022 to 2024 at a specialized clinic for trauma-affected refugees in Denmark. A total of 9 adult refugees diagnosed with CPTSD used the OBT during psychotherapy to actively track a personally relevant and collaboratively defined target phenomenon through a button press. The OBT provided vibrotactile feedback and stored timestamped data for therapeutic use. A total of 25 semistructured interviews were conducted across 3 time points: before, during, and after treatment. Reflexive thematic analysis guided by a postphenomenological framework was used to explore how the technologies mediate experience.

Results: Participants (6 women and 3 men, median age 47 years, IQR 31–57 years) had lived in Denmark for an average of 21.4 years. The duration of OBT use ranged from 22 to 366 days. Participants tracked between 2 and 14 different phenomena and registered between 37 and 4733 events in total during their courses of treatment. Thematic analysis revealed five central themes that captured the multistable character of the OBT: (1) from external instrument to extension of the self, (2) mental switch, (3) faithful companion, (4) scarlet letter, and (5) emergency lifeline. Patients described the OBT as a meaningful anchor in distressing moments, enhancing emotional regulation, self-awareness, and continuity between sessions. The OBT's vibrotactile feedback was experienced as affirming and responsive, reinforcing a sense of being acknowledged and connected, even in the absence of direct therapist contact. However, the visibility of the OBT also introduced challenges, including stigma and altered social dynamics.

Conclusions: The OBT functioned as an active mediator in therapy. It supported in-the-moment tracking of personal experiences, encouraged agency and emotional regulation, and helped patients feel connected to their therapist outside of sessions. The vibrotactile feedback played a key role in how the OBT was embodied and interpreted. These findings highlight the value of designing digital mental health technologies incorporating active self-tracking that emphasize relational use, experiential meaning, and personalization. A focus on simplistic design, adaptability, and patient-defined use may enhance therapeutic relevance, especially in settings where stigma or complexity limits engagement.

(JMIR Form Res 2025;9:e70511) doi:10.2196/70511

XSL•FO
RenderX

KEYWORDS

wearable technology; self-tracking; One Button Tracker; psychotherapy; engagement; postphenomenology; complex PTSD; personalized treatment; refugees; complex posttraumatic stress disorder

Introduction

Background

Wearable self-tracking technologies are increasingly recognized as valuable tools in mental health care, offering innovative approaches to enhance patient engagement, promote empowerment, and personalize treatment [1,2]. The rapid advancement of digital health technologies, particularly smartphones and wearable devices, has eased the integration of these tools into daily life, providing users with accessible means to monitor their health and well-being [3].

These wearable technologies are grounded in the principles of ubiquitous computing [4] and often rely on passive data collection, such as heart rate, sleep patterns, body temperature, and physical activity [5]. This passive monitoring approach minimizes the burden on patients while generating clinically relevant data that can be used for diagnostic and clinical purposes [1,6,7].

Within the field of digital mental health, 2 distinct types of data collection are recognized, namely passive and active. Passive data collection involves the automatic gathering of information without requiring user intervention, thus facilitating continuous monitoring of health metrics. In contrast, active data collection necessitates intentional engagement from users, allowing them to log specific experiences and emotions in real time, known as Ecological Momentary Assessment. This can be achieved through techniques such as electronic diaries [8] or the use of smartphone apps that prompt users to engage in assessments and interventions, known as Ecological Momentary Assessment and Intervention [9].

However, despite the advantages of both passive and active data collection methods, there is a tendency in current research to prioritize data collection at the expense of exploring how wearable self-tracking technologies can actively mediate the therapeutic process (eg, [10-12]). While there is an increasing focus on delivering treatment in patients' daily lives and natural settings [13], interventions often consist of predefined assessment points with limited personalization [9]. This narrow approach may lead to missed opportunities to personalize treatment and use digital health technologies for in-the-moment engagement during the treatment process.

Engaging patients in tracking their subjective experiences as they occur can redirect focus to the present, fostering self-awareness, emotional regulation, and agency while producing personalized, fine-grained data [14]. By facilitating engagement in therapeutic interventions, wearable self-tracking technologies can empower patients to access support precisely when they need it, thereby enhancing their ability to navigate challenges in their daily lives. This dual role, acting as both data collectors and active mediators in therapeutic interventions, remains underexplored in mental health care research. A better understanding of this could reveal how wearable self-tracking technologies, when thoughtfully integrated, can extend therapeutic support beyond the clinical setting into patients' everyday lives, ensuring they can engage in personalized interventions at the moments that matter most.

The One Button Tracker: A Novel Approach to Engagement

This study introduces the One Button Tracker (OBT), a research prototype of a single-purpose wearable self-tracking instrument specifically for in-the-moment registration of a self-chosen subjectively experienced mental health phenomenon [14]. In contrast to multifunctional devices, such as smartphones or smartwatches, the OBT has one specific purpose: tracking self-defined mental health phenomena, while reducing cognitive and technical barriers.

The OBT is compact, measuring 41×31×12.5 mm, and operates with a user-friendly single-button interface (Figure 1). This design enables patients to record occurrences of their self-defined mental health phenomenon with a button press, promoting immediacy and precision in data collection. Each button press captures a timestamp and duration of the event [15] in the OBT's internal storage [14]. These data can be accessed through a web-based data visualization tool collaboratively analyzed during therapy sessions.

XSL•FO
RenderX

Figure 1. The One Button Tracker, a novel single-purpose wearable self-tracking instrument used in this study, shown configured with a neck strap and wristband.

Figure 1 displays the OBT that was used by refugee patients diagnosed with Complex Posttraumatic Stress Disorder during a clinical qualitative study conducted at the Clinic for Trauma and Torture Survivors in Denmark between 2022 and 2024. Emphasizing user adaptability, the OBT can be integrated into daily life without visual engagement, ensuring minimal disruption to routines. The instrument can be carried in a pocket, worn around the neck, or strapped to the wrist, catering to individual preferences (Figure 1). Furthermore, strong vibrotactile feedback accompanies each button press, providing tangible interaction with the instrument. This unique feature enhances the tracking experience, promoting user engagement without the need for visual attention.

Addressing an Epistemic Gap in Mental Health Technology

Despite a growing interest in wearable technologies in mental health care [1], a significant epistemic gap remains: how do these tools mediate therapeutic experiences and shape patients' perceptions of themselves and their mental health? Traditionally, wearable technology research has focused on measuring behavioral data, neglecting the dynamic, lived interactions between patients and these tools. In line with Torous et al [16] recommendations to define engagement in digital health in terms of the process rather than through behavioral measures, this study focuses on the OBT as an active mediator in the therapeutic process. By exploring the dynamic interactions between patients and the OBT, we aim to uncover how the instrument supports engagement, self-awareness, and becomes part of the therapeutic process. To guide this exploration, we turn to postphenomenology, a philosophy of technology developed by Ihde [17] and expanded by Verbeek [18] to provide us with a theoretical framework for understanding how technologies mediate human experiences.

Research Question

The research question "How does the OBT mediate the psychotherapeutic process in everyday life?" was asked. Specifically, we explore how patients engage with the OBT during their daily lives, what the OBT means to them, and how it enables or hinders recovery in their treatment process.

Methods

Participants

This study was part of a larger participatory action research project aimed at co-developing a self-tracking assisted psychotherapy treatment concept for refugees diagnosed with complex posttraumatic stress disorder (CPTSD) [14]. The study was conducted at the Clinic for Trauma and Torture Survivors (CTTS) in the Region of Southern Denmark, a specialized outpatient facility within the Danish public health care system that provides interdisciplinary mental health care to approximately 550 trauma-affected refugees annually. The clinic offers evidence-based treatment, including psychotherapy and psychiatric support, for refugees with trauma-related mental health conditions.

Recruitment was conducted through routine clinical pathways. Psychologists at the clinic introduced the study to eligible patients either during the intake process or in the context of ongoing treatment. Some patients were already engaged in therapy for CPTSD when they were invited to participate, while others were introduced to the project as part of the initial assessment and planning of their treatment. Patients who expressed interest received additional information about the study from their treating psychologist, who also assessed eligibility and obtained informed consent. To be eligible, participants had to be aged 18 years or older, have a refugee background, receive a diagnosis of CPTSD based on the International Trauma Interview [19], be accepted for psychiatric

XSL•FO
RenderX

treatment at CTTS, and be fluent in Danish to ensure meaningful participation in therapy and interviews without an interpreter.

Intervention: Self-Tracking Assisted Psychotherapy

The OBT was introduced to the patient by their treating therapist as a digital personal diary designed to track subjectively defined and relevant mental health phenomena during daily life [14,20]. This specific phenomenon is referred to as the target phenomenon, which is pertinent to the patient's unique mental health challenges (Figure 2 for an outline of the treatment process). Together in the therapy session, the patient and therapist identified and defined the target phenomenon (eg, a symptom, impulse, or behavior) to be tracked. The target phenomenon should present itself clearly to the patient when it occurs; therefore, patients were encouraged to identify somatic or behavioral markers associated with its occurrence [14]. This target phenomenon could be adjusted throughout the treatment period based on the patient's experiences with tracking and the insights gained from therapy sessions. Patients were encouraged to carry the OBT with them during their daily routines, using it as a wristband, around their neck, or in a pocket or bag.

Figure 2. Flowchart outlining the self-tracking assisted psychotherapy process using the One Button Tracker (OBT).

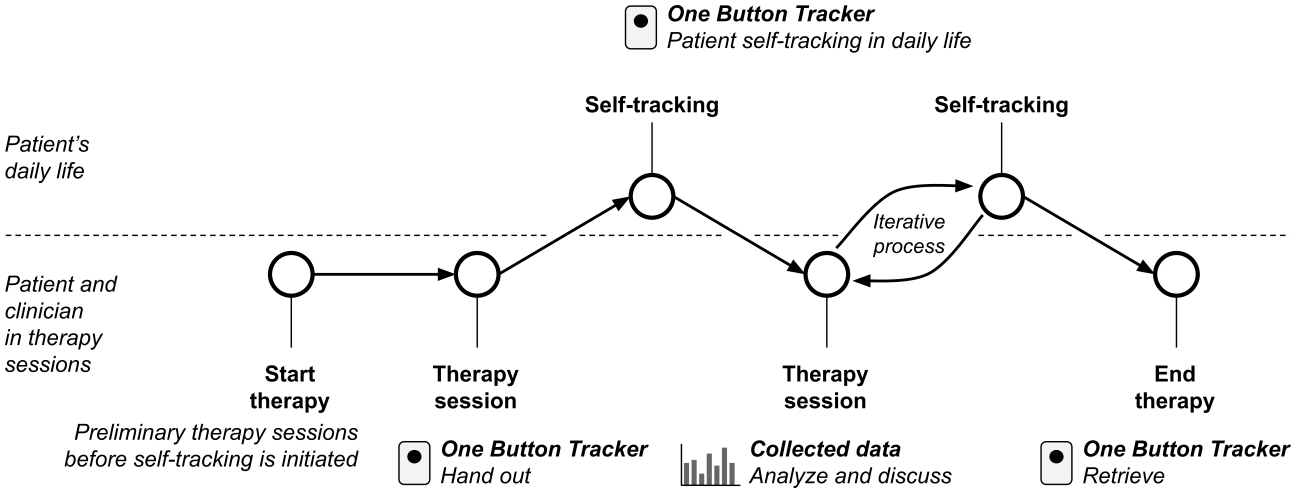

Figure 2 displays the flowchart, which illustrates the collaborative definition of target phenomena, data collection in daily life, and therapeutic reflection using a web-based visualization tool. This process was developed and evaluated between 2022 and 2024 in a Danish clinical study with refugee patients diagnosed with complex posttraumatic stress disorder.

Based on the patients' expectations of the occurrences of the target phenomenon during their daily life, the therapist and patient developed a hypothesis list to predict patterns and dynamics in the target phenomenon. An observation protocol was then created to specify tracking criteria, including characteristics of the target phenomenon and how the OBT would be carried. Initially in therapy, patients could choose 1 phenomenon to track during their daily life. If deemed relevant, patients were introduced to the option of tracking 2 phenomena, assigning single or double button presses accordingly [20]. Frequently, a single press represented a distressing phenomenon, while a double press corresponded to having completed a therapeutic intervention at home as agreed upon in therapy. Upon returning to the therapy session, the collected data were analyzed collaboratively using a web-based data visualization tool as illustrated in Figure 3. This figure comprises blue dots representing single button presses (typically associated with distressing events), and orange dots representing double presses (often indicating completed interventions). Dot size reflects press duration. Data were collected using the One Button Tracker as part of an intervention conducted at a Danish trauma clinic between 2022 and 2024. These discussions focused on developing a shared understanding of the occurrences of the target phenomena and personalizing interventions based on the insights gained from the data.

XSL•FO
RenderX

Figure 3. Calendar visualization of self-tracking data from participant P2 in a clinical qualitative study with refugees diagnosed with complex posttraumatic stress disorder.

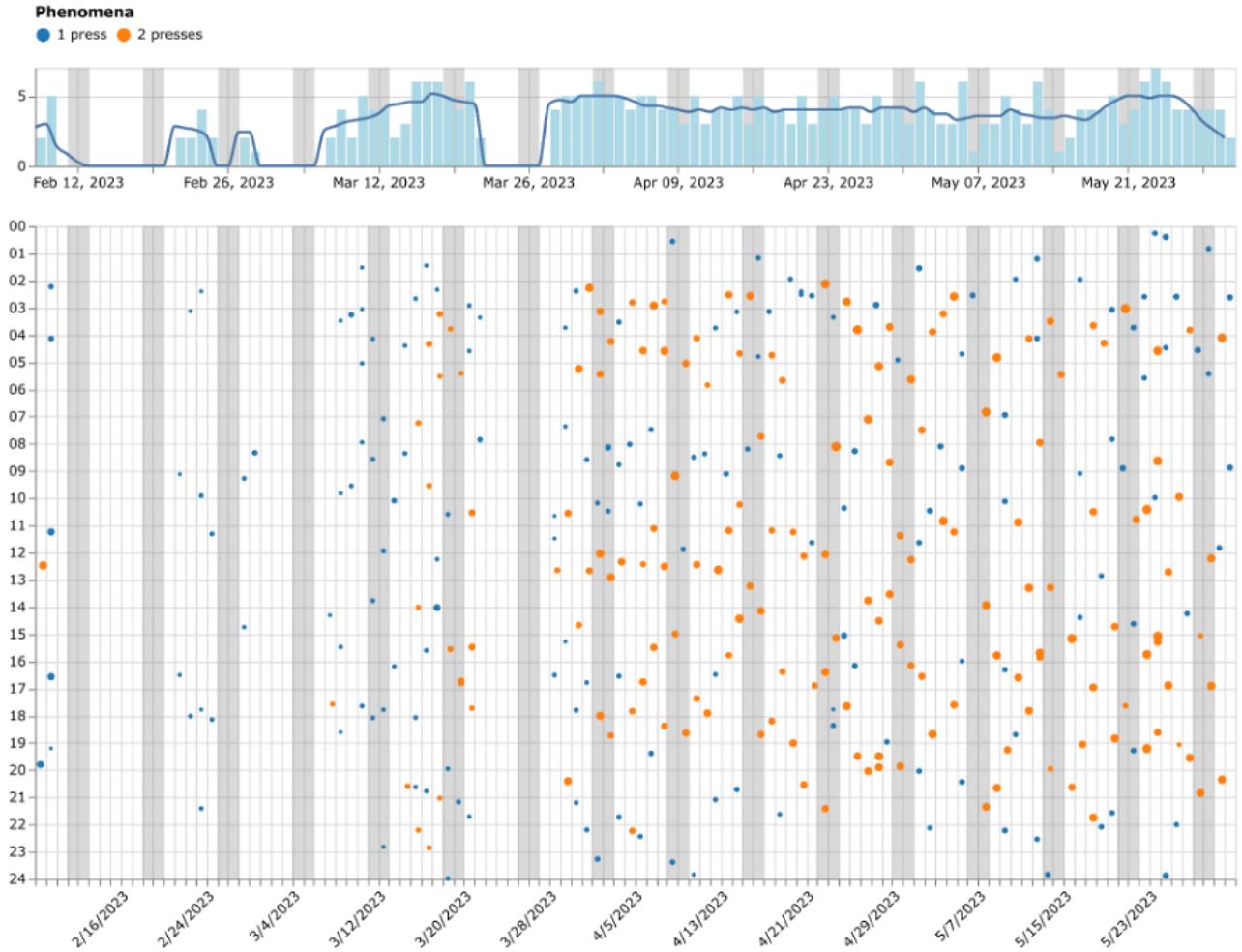

Interviews

The interviews were semistructured, meaning they followed an interview guide but remained flexible to allow exploration of new and unexpected insights into patients' lived experiences [21]. The interview guide was based on both a clinical and a postphenomenological framework, focusing on patients' interactions with the OBT across 3 key areas: its role in their treatment (eg, "What have you and your therapist decided your target phenomenon should be?"), their experiences of using the OBT in daily life (eg, "How has having the OBT with you in your daily life been?"), and its impact on their social relationships (eg, "Have you shown your family and friends the OBT?"). A total of 25 interviews were conducted with 9 patients at 3 stages: before, during, and after treatment. Two patients participated in only 2 interviews, due to the short duration of their treatment.

Participants were given the option to participate in the interview either at the clinic or in their own homes. Interviews were conducted face-to-face in Danish by the first author (LGR) except for 1 interview where LGR served as the therapist; in this case, author (SBM) conducted the interview. The duration of the interviews ranged from 19 to 93 minutes, with a median of 53 minutes. At the clinic, interviews were held in consultation rooms. In total, 22 (88%) interviews took place at the clinic, while the remaining 3 interviews were conducted in patients' homes (2 patients). All interviews were audio-recorded and subsequently transcribed verbatim by a research assistant. To ensure accuracy, LGR cross-checked all transcripts against the audio recordings. Quotations presented in this study have been translated from Danish into English by LGR. Each quote is accompanied by a patient ID to ensure anonymity.

Data Analysis

The analysis was conducted by LGR under the supervision of co-authors [JEL, TBC, JA, and LH] using reflexive thematic analysis as described by Braun and Clarke [22,23]. An inductive approach was used, grounded in a postphenomenological theoretical framework [17,18,24], which provided the lens for understanding how the OBT mediated patients' experiences in their everyday life. In this analysis, we drew on Ihde's [17] 4 human-technology relations to interpret patients' engagement with the OBT: technologies can become embodied as extensions of the body (eg, glasses), function as hermeneutic relations representing reality (eg, a thermometer), be experienced as quasi-others in alterity relations (eg, chatbots), or recede into the background as part of the environment (eg, noise from a refrigerator). Importantly, these relations are not mutually exclusive. Instead, as Ihde [17] argues, technologies exhibit multistability, the ability to take on different roles depending on the user's context and intentions [24,25]. According to Ihde

XSL•FO
RenderX

[17], technologies are inherently ambiguous and transform the user's experience as per the cultural context in which it is used.

The analysis followed Braun and Clarke's [23] 6-phase reflexive thematic analysis, which emphasizes flexibility, researcher reflexivity, and the interpretive nature of thematic development. In the first phase, LGR familiarized herself with the dataset by listening to audio recordings and reviewing the interview transcripts multiple times. This immersion allowed for an in-depth understanding of both explicit content and latent meanings embedded in the patients' narratives. In the second phase, initial codes were generated to identify significant features of the data relevant to the research question. Codes reflected variations in patients' experiences with the OBT and were iteratively refined as patterns were identified.

The third phase involved clustering the codes into initial themes, reflecting patterns related to the OBT's role in the lives of the patients. These themes were developed and reviewed iteratively throughout phases 3 and 4 through repeated engagement with the data and ongoing discussions with coauthors to ensure coherence and alignment with the study's research question. The postphenomenological framework served as a theoretical lens during this phase, allowing us to interpret the data in terms of how the OBT mediated patients' experiences through its multistability. In phases 5 and 6, themes were refined, clearly defined, and named to ensure that they captured the nuances of patients' perspectives while using the theoretical framework as an analytical resource and contextualizing findings within existing literature.

Ethical Considerations

This study was reviewed and approved by the Regional Committees on Health Research Ethics for Southern Denmark (project ID: S-20210019 CSF) and conducted following the Declaration of Helsinki. All participants received written and oral information about the study before participation and provided informed consent. This included consent for the use of anonymized quotations, interviews, and self-tracking data. No financial compensation was provided to participants. To protect privacy, all data were pseudonymized and securely stored following Danish data protection regulations. Interview transcripts and tracking data were deidentified, and access was restricted to members of the research team. No personally identifying information appears in the manuscript or supplementary materials, and all figures presenting self-tracking data have been fully anonymized to ensure confidentiality.

Results

Overview

In total, 9 patients participated in this study (n=6 females, n=3 males), aged 22-63 years, with a median age of 47 years (IQR 31–57 years). Patients had been living in Denmark for an average of 21.4 years. Three patients were originally from Bosnia, while the remaining patients were from Caucasus, Lebanon, Iran, Afghanistan, Turkey, and Syria (refer to Table 1 for additional demographic data).

Table . Demographic characteristics of 9 refugee patients diagnosed with complex posttraumatic stress disorder, enrolled in a clinical qualitative study on the use of a novel wearable self-tracking instrument, the One Button Tracker, during psychotherapy.

Patient ID	Age (years)	Sex	Employment status	Years in Denmark
P1	31	Female	On sick leave	20
P2	63	Male	On sick leave	30
P3	59	Female	On sick leave	25
P4	22	Female	Student	22
P5	32	Female	Student	20
P6	57	Male	On sick leave	30
P7	52	Female	Temporary employed	15
P8	47	Male	Employed	22
P9	22	Female	Student	9

Table 1 presents the data collected at a specialized trauma clinic in Denmark between 2022 and 2024.

Patients used the OBT for durations ranging from 22 to 366 days and tracked between 2 and 14 different phenomena each throughout their treatment. Over this period, the number of recorded observations varied widely, from 37 to 4733, reflecting diverse patterns of engagement with the OBT (Table 2). The table includes the number of tracking sessions, duration of use, number of phenomena tracked, and button press frequency. Data were collected between 2022 and 2024 as part of a qualitative intervention study at a Danish trauma clinic.

XSL•FO
RenderX

Table . Individual-level summary of self-tracking activity using the One Button Tracker for 9 refugee patients with complex posttraumatic stress disorder.

Patient ID	Total sessions with tracking	Total days with tracking	Days with data	Days without data	Number of observations	Daily average observations	Protocols used	Target phenomena tracked
P1	12	92	62	30	1276	20.6	4	3
P2	17	112	87	25	347	4	4	4
P3	24	324	289	35	2038	7.1	17	14
P4	22	262	185	77	130	0.7	10	8
P5	13	366	278	88	980	3.5	8	8
P6	12	246	128	118	4733	37	5	5
P7	20	252	212	40	708	3.3	5	4
P8	8	175	58	117	188	3.2	1	1
P9	4	22	22	0	37	1.7	3	3

Themes

The analysis resulted in five key themes: (1) from external instrument to extension of the self, (2) mental switch, (3) faithful companion, (4) scarlet letter, and (5) emergency lifeline. In the results that follow, we see how the OBT is multistable, functioning as an embodied extension of therapeutic interventions, a friend, a link to their therapist and treatment, and a scarlet letter.

From External Instrument to Extension of the Self

Patients varied in how they carried the OBT, reflecting individual preferences and comfort levels. In total, 5 opted for discrete placement, such as in a pocket, a bag, or a designated place in their home, while 4 patients wore it more openly on their wrist or around their neck. While these choices often remained consistent throughout treatment, a few patients adjusted their approach based on practicality or personal needs in specific contexts.

As the OBT became embedded in the daily routine of the patients, it evolved from being an external instrument to an extension of themselves. Within approximately 2 weeks, many patients described using the OBT as a natural and almost automatic part of their lives. One patient articulated this experience vividly:

> *Yes. It's like my body. Sometimes if I forget it, like if I'm driving, for example, to go shopping, then I go back, grab it and go again.* [P7]

This embodiment was particularly evident in patients' nighttime routines, where the OBT remained within reach. Some patients placed the instrument under their pillow, enabling them to record observations easily during the night. One patient shared how this practice became a habitual and embodied part of her routine:

> *Well, at first I tried sleeping while wearing it, but in the middle of the night I noticed that I was very sore around my chest. So in the end, I just thought I'd put it under the pillow, so it's easier for me.* [P1]
>
> *That makes sense.* [Interviewer]
>
> *Yes, it's better. But sometimes when I wake up, without noticing, because I'm holding it.* [P1]
>
> *You hold it in the morning?* [Interviewer]
>
> *Without knowing why I've done it. So I hold it so much that I think I have it.* [P1]

Why did the OBT become so indispensable to patients? For many, its significance extended beyond mere habitual use; it took on emotional and even physiological importance, integrating into their daily lives as an extension of their own bodies. This emotional attachment was especially pronounced when patients experienced the absence of the OBT, whether due to technical errors or returning the OBT at the end of treatment. The void left by the OBT was deeply felt, as 1 patient described a sense of loss when asked to stop using it:

> *There was always something missing… I look at the spot where it used to be, you know? And then I say, "It's not here." Also, when I go out, you know? I miss it. I miss it a bit.* [P6]

The physiological bond with the OBT was further reinforced by the tactile sensation of pressing the button, which patients came to associate closely with their emotional and physiological states. One patient vividly elaborated on the vibrotactile feedback's profound effect, saying:

> *It goes to my heart when I press. A lot. I feel it. It goes through the blood. It brings oxygen.* [P3]

This enduring sense of presence, even when the OBT was no longer physically available, illustrates the OBT's role as a stable extension of self-regulation that, once embodied, remains psychologically significant.

Mental Switch

For many patients, the OBT acted as more than a self-tracking instrument; it became a tangible instrument for self-regulation strategies, functioning as a mental switch that fostered a sense of agency during moments of distress. This role as a mental switch was characterized by multistability, adapting to the diverse intentions and emotional states of the patient. One patient illustrated how the act of pressing the button held

XSL•FO
RenderX

different meanings depending on her intention of using it, while consistently carrying personal meaning:

> *Or maybe it's because, when I press the button, it starts to mean something.* [P3]
>
> *It means something?* [Interviewer]
>
> *Yes. It means maybe I'm anxious, maybe I'm scared, maybe I'm depressed, maybe it's about my medication, or maybe I'm calling too often. It's different each time. It's starting to feel meaningful to me every time I press.* [P3]

For some, this act of pressing the button was not merely functional but became a moment of reflective self-awareness, helping patients recognize and manage specific emotions or behaviors. The dynamic adaptability of the OBT allowed it to align with the unique needs of each patient, making it a personalized and integrated aspect of their therapeutic process. One patient described how the OBT became instrumental in resisting emotional eating, transforming an overwhelming impulse into a manageable act of self-regulation:

> *It has helped a lot, really. Before, I knew I was eating away my feelings. But with pressing the button, you realize, "Okay, I have some feelings that I'm trying to eat away." So, in a way, it stops it. Yes, it does something to the brain. Suddenly the brain is like, "Okay, I have some feelings, now I feel hungry, I should press, but I've pressed many times today. So I've eaten enough." It gives something in the mind that suddenly makes you stop.* [P9]

In these cases, pressing the button became a hermeneutic act, enabling patients to interpret their patterns and reflect on their impulses. This reflective process not only helped them resist impulsive behaviors but also fostered a sense of agency. At other times, the OBT receded into a background relation in the patient's daily life, offering a constant and supportive presence.

For some patients, the act of pressing the button carried symbolic meaning, fostering hope and reinforcing their commitment to recovery. One patient described how the availability of the OBT as a mental switch gave her reassurance in moments of anxiety:

> *Yes, if it isn't there… then it's like I'm empty, like I'm missing someone. When anxiety comes, what should I do? Yes, of course, we have many strategies. But it's also… I don't know. But if I press the button, I believe I will get better.* [P7]

For others, pressing the button offered emotional regulation during moments of distress, helping them regain a sense of agency. One patient shared how the OBT helped him manage intrusive and distressing thoughts:

> *When I think about the war. When I think about my village. When I think about everything. Then I press too. It's like... How should I explain it? It's like it says to me: "Relax, it will be okay."* [P6]

Here, the act of pressing the button was shaped by the therapeutic context in which the OBT was introduced in, its macroperception, which informed how the sensory experience of pressing the button, or microperception [18], was interpreted. This context imbued the act with a sense of agency, enabling patients to interrupt intrusive and distressing thoughts. For 1 patient, this process symbolized a mental switch, allowing her to shift distressing thoughts and reinforcing her connection to therapy:

> *When I press the button, it shifts my bad thoughts so quickly. But it comes back again. But I have to do it because it helps me. That's because, for example, when you don't use the light here, right? When you go outside, you turn off the light, right? It helps me so much.* [P5]

Over time, the OBT became an anchor that paused impulsive behavior and allowed time for thoughtful decision-making. Patients noted how pressing the button encouraged them to slow down and reconsider their actions. One patient articulated this shift:

> *But I would say that the button itself has, in a way… I don't know, somehow made me think twice about things and not just do things impulsively. It's actually given me a chance to be a bit more sensible in some situations and really think things through.* [P4]

For many patients, pressing the button became a meaningful act, encouraging emotional regulation and self-reflection during moments of distress. Reflecting on her use of the OBT, 1 patient described how it helped her reconnect with her emotions:

> *Well, it has somehow forced me to feel something. Or at least to check in with myself and feel something. I've been pretty numb in relation to my emotions. But over time, it's gotten me to feel more. And it's strange, since it's just a button, but it's gotten me to feel more. Because I've been forced to do it. I've also forced myself to do it. And I don't even know, I don't think I ever really understood its purpose. I knew it was supposed to track a lot of things, but I didn't know exactly for what purpose. But for me, the purpose has been completely personal, that it could get me to feel again.* [P4]

This theme reflects the depth and complexity of the act of pressing the button of the OBT. While appearing as a simple and insignificant act, it served as a mental switch, a way for patients to regulate their emotions, provide hope for recovery, and intentional self-awareness in the moments they needed it the most.

Faithful Companion

For 7 patients, the OBT transcended its role as a self-tracking instrument, assuming the role of a "quasi-other" that provided support. Patients consistently anthropomorphized the OBT, attributing it with a sense of agency and giving it a "voice" that offered them encouragement and companionship. For example, 1 patient, described feeling like the OBT was an advisor, guiding her decisions and reinforcing behavior change:

> *It's like it says to me: "[Her name], you have a lot of anxiety. [Her name], why don't you go outside? [Her name], why aren't you sitting with others and talking like everyone else?"* [P3]

XSL•FO
RenderX

This sense of connection and guidance was closely tied to the OBT's vibrotactile feedback, which, for some patients, became an affirmation of support that symbolized companionship in moments of distress. The vibrotactile feedback, in particular, was perceived as a sign of life, enhancing the felt connection between the patient and the instrument. One patient articulated this as a feeling of not being alone:

> *It was like there was someone with you the whole way. Yes. You get this feeling that you're not alone in this process. There's someone who is also keeping an eye on what you're doing. Or on your challenges. Feeling the vibration was just wonderful. It was just great. Because it's like it responds to you.* [P9]

Here, the tactile feedback served not just as a technical acknowledgment of a recorded observation but mediated an emotional relation, making the OBT feel like a quasi-other that provided comfort during challenging moments and in periods of solitude. One patient described how he held the OBT in his hands at night as if it were a supportive friend, substituting for his wife's presence:

> *How should I say it? Because the others are sleeping. And I mostly wake up at night. I wake up many times. So I take it in my hand and hold it. Instead of holding my wife, for example. She's sleeping; she has to work in the morning, so I hold it. It's like it helps me. "I'm here," for example. In that way.* [P6]

For this patient, the alterity relation was disrupted during technical outages, which revealed the extent of their emotional reliance on the OBT. When the vibrotactile feedback stopped functioning, the patient expressed it as a loss akin to losing a friend:

> *It's not the same. It's like my friend has died. Because he doesn't respond. So if it vibrates, then it's like, "There's life in you." Then it's fine.* [P6]

In these moments, the OBT's quasi-other role in patients' lives became evident, embodying a form of companionship that fostered a therapeutic extension beyond the clinical setting.

Scarlet Letter

The OBT served as a visible, purpose-specific object in patients' daily lives, shaping their self-awareness and social interactions. While its physical presence bridged the gap between therapy and everyday life, it also introduced challenges and opportunities, amplifying or reducing engagement depending on the context.

For 3 patients, the visibility of the OBT took on the meaning of a scarlet letter of shame leading to discomfort and self-consciousness, particularly in public or familial settings. One patient explained how its presence conflicted with her preference for keeping her mental health challenges private:

> *Well, I don't need people to see it. Like, starting to ask questions, and I don't need people to know that side of me. My colleagues know that I see a psychologist, but they don't need to know more than that. So, I'm fine with not wearing it. Because then it becomes obvious that something's going on.* [P4]

Another patient described how the visibility of the OBT influenced family dynamics, amplifying family members' negative views toward mental health care treatment. One patient recounted how his teenage daughter disliked seeing him wearing the OBT, as she expressed how she feared it might lead to him being taken away and institutionalized. This patient also shared how his wife discouraged him from wearing the instrument around extended family, reflecting concerns about social stigma:

> *She [his wife] doesn't want me to bring it with me. She says we should stay together, and you don't need to have it with you. Maybe she... Because we have family down there, and maybe they would ask, "What is that?" All sorts of things. So maybe she's ashamed or something. That could be.* [P6]

Although this patient was comfortable wearing the OBT among family members, his wife's discomfort led him to stop using it in certain settings. The visibility of the OBT amplified existing stigmas surrounding mental health treatment, which ultimately influenced his engagement with the instrument.

Conversely, for some patients, the OBT's visibility transformed from a scarlet letter of shame into a symbol of strength and proactive management of mental health. One patient shared how her daughter supported her use of the instrument:

> *I sometimes forget it [the OBT] when I'm lying down somewhere. My daughter says, "Mom, I brought your instrument", and then she places it under the pillow for me. She reminds me of it.* [P1]

In this instance, the OBT encouraged familial support and established a sense of shared responsibility in the therapeutic process. This dynamic not only enhanced the patient's accountability but also strengthened family bonds, contrasting the negative associations of the scarlet letter with a sense of empowerment. While some patients found the OBT supportive, others struggled with its aesthetics, which created mixed feelings about wearing it openly. One patient remarked on its appearance:

While some patients found the OBT supportive, others struggled with its aesthetics, which created mixed feelings about wearing it openly. One patient remarked on its appearance:

> *I wasn't happy with it. Like, because it was ugly and looked like an ankle monitor. It was kind of a bit of a love-hate relationship we had. Because it was so ugly, but it still made me think about what I was tracking.* [P4]

For her, the OBT's appearance compromised her willingness to wear it, evoking associations with social judgment similar to the stigma of a scarlet letter. Similarly, another patient found that the OBT's appearance triggered memories of wartime imprisonment:

> *And I haven't received much information. What does it mean for me? At first, I thought: "Oh, for hell's sake, it's like I'm back in prison." Like I have to report to the police that I'm here.* [P2]

These varied reactions illustrate the OBT's multistability, highlighting its ability to take on different roles and meanings depending on the specific contexts and situations in which it is

XSL•FO
RenderX

used. For some contexts, the OBT served as a source of stigma and discomfort, while in others, it fostered engagement and support, underlining how the meaning of the instrument is not fixed, but rather shaped by the circumstances in which it is integrated.

Emergency Lifeline

The OBT served as an emergency lifeline for many patients, facilitating nonverbal communication that bridged their daily lives and therapy sessions. The OBT mediated patients' inner dialogues, reinforcing the therapeutic relationship and acting as an embodied extension of emotional communication. One patient described how pressing the button created a sense of being "seen" by her therapist, even outside of sessions:

> *It was kind of like: "I'm eating my feelings away. Please help me."* [P9]
>
> *And what happens when you press, and you sort of say, please help me? What happens then?* [Interviewer]
>
> *Yes, well, then I can feel that, okay, I'm not alone in this. My therapist will see it, maybe she'll do something about it or say something to me that helps. Figure out why I've eaten so much this past week, for example.* [P9]

This interaction established a feedback loop, strengthening the therapeutic relationship and fostering a sense of accountability. Another patient highlighted how pressing the button encouraged reflection on therapy sessions:

> *I feel like I'm sitting right in front of my therapist, and we're talking. What did we talk about last time? What helps me?* [P5]

In such moments, pressing the button allowed patients to "call on" their therapist's advice, creating a silent connection to their treatment. Similarly, another patient described how pressing the button created a sense of support in moments of need:

> *It's like someone is saying to you: "I'm here."* [P9]
>
> *Yes. So, when you press it and it vibrates?* [Interviewer]
>
> *Yes. It's like my whole body felt that someone is here. Right here beside me. But it also made me feel like something in my mind, like my therapist is with me all the time because of the tracker.* [P9]

In such moments, the OBT mediated an immediate and nonverbal imagined dialogue with the therapist, fostering a sense of support during the patient's daily life. In both cases, the OBT mediates a nonverbal dialogue with the therapist, allowing patients to feel supported between sessions. As such, the OBT embodied the therapeutic relationship, reinforcing patients' sense of accountability and engagement. One patient described how the OBT served as a daily reminder of her therapeutic homework:

> *As I told my therapist, it's kind of like a task for me, actually. That's how I see it. It's a task for me, so it's supposed to help me. So I don't know when it will help me, but that's how I think about it.* [P2]

The OBT thus channels patients toward therapeutic engagement, embodying both the presence and expectations of their therapist in their everyday lives. For some, even after returning the OBT, its influence persisted, though not as strongly.

Discussion

Principal Findings

This study is the first to explore how refugees with CPTSD engage with a single-purpose, wearable (the OBT) to track a personally relevant and collaboratively defined mental health phenomenon in daily life. The findings showed that the OBT-mediated therapy promoted agency, helping patients regulate emotions and stay connected to their therapist between sessions. As an active mediator, it shaped how patients related to their symptoms, treatment, and themselves. These findings reveal both the potential and limitations of the OBT, offering a more nuanced understanding of how self-tracking can support psychotherapy.

Design Simplicity and Integration

The OBT's minimalistic design made it easy to incorporate into daily life. Most patients described it as intuitive within the first 2 weeks, aligning with previous research showing that low cognitive demand improves engagement [2,26]. Its simplicity allowed patients to register experiences in the moment without distraction, creating a sense of immediacy that supported embodied engagement. Some even described the instrument as becoming part of their body.

These experiences can be understood through Ihde's [17] concept of multistability and Verbeek's [18] theory of technological mediation, which describe how technologies shape human intentionality and experience. As the OBT faded into the background, patients could focus less on the instrument itself and more on the intention of pressing the button. In this way, the OBT became an active part of the therapeutic process, not just an instrument for collecting data, but something that supported agency and intentionality in the moment.

At the same time, the simplicity of the design of the OBT placed responsibility on the patient. With no structured prompts or categories, patients had to remember their target phenomena and interpret what their data meant over time. Several adapted or expanded their target phenomena (from one to two target phenomena) as therapy progressed. While this flexibility enriched the therapeutic process, it also required therapists to stay attuned to how patients were using the OBT and what their patterns meant. Supporting this evolving engagement calls for an open and adaptive therapeutic approach.

Taken together, these findings suggest that simplicity and personalization are central to making self-tracking meaningful in a therapeutic context. When patients define what is relevant to track, the OBT can support a stronger sense of agency and bring therapy into the flow of everyday life. This reframes the role of wearable technologies in mental health, not just as passive data collectors, but as instruments that actively mediate change. At the same time, striking the right balance between openness and support is key. Simplicity must not come at the

XSL•FO
RenderX

cost of clarity or overwhelm patients with too much interpretive responsibility.

Personalization Through Multistability

The OBT's multistability [17] was central to how patients engaged with it. Although pressing the button was a uniform act, the meaning and function varied widely. Some patients described it as a "mental switch" that helped them pause or reflect in difficult moments. Others experienced it as a "faithful companion" that offered emotional support between sessions. A few used it as a silent link to their therapist, an "emergency lifeline" that created a sense of connection and accountability.

One patient described how the vibration moved through her body "like blood," illustrating how the OBT, in some cases, became more than a tool. It was experienced as part of the self. This points to what Verbeek [25] describes as a cyborg relation, where the boundary between person and technology dissolves. Here, the OBT no longer simply mediates therapeutic engagement, but shapes how patients experience themselves and their treatment.

This deep integration may support agency, but it also raises concerns about dependency. Two patients asked to keep or buy the OBT after treatment, and 1 delayed their final session to continue using it. While strong engagement can be beneficial during therapy, it may complicate the end of treatment. Therapists using similar tools should be prepared to help patients gradually internalize what the instrument supported, so that progress continues without it.

The vibrotactile itself played an important role. For many, it was not just feedback, but a form of affirmation. The immediate tactile response gave a sense of being acknowledged, reinforcing their actions and sometimes symbolizing the presence of the therapist. Unlike tools that provide delayed or abstract feedback [27], the OBT offered something more grounded in both the body and the therapeutic relationship.

These findings highlight the value of personalization, especially when contrasted with existing literature that has primarily focused on collecting physiological, contextual, or environmental data to tailor ecological momentary interventions [9]. While such approaches often emphasize objective metrics, the OBT allowed patients to define for themselves what was relevant to track in daily life. This made the act of self-tracking more personally meaningful and emotionally resonant. It also provided detailed, in-the-moment collected data that therapists could draw on to guide conversations and adapt treatment in ways that felt more aligned with each patient's lived experience.

Still, the multistability of the OBT sometimes introduced challenges. Not all of the patients' interpretations were helpful. One patient believed the vibration would "absorb" his anger and felt disappointed when that did not happen. These examples show why clear communication is essential when introducing wearable instruments. Patients need guidance and support to make sense of what the OBT can and cannot do, and how to use it in ways that strengthen rather than disrupt the therapeutic process [28].

Visibility and Its Dual Impact

The visibility of the OBT mediated how patients managed their self-presentation in everyday life. While the instrument often blended into daily routines [17], it also acted as both a supportive and disruptive mediator. For some, it encouraged openness and relational support; for others, it amplified feelings of exposure and stigma. Drawing on Feenberg's [29] expansion of Ihde's [30] concept of the extended body, the OBT served both as a medium of self-awareness and as a visual representation of the patient. This dual role reflects the complexity of visible self-tracking technologies. The OBT's physical presence influences how patients navigated social and cultural expectations around mental health. While some welcomed the visibility as a sign of taking action, others found it uncomfortable, comparing it with an ankle monitor.

These findings raise broader questions about the design of wearable health technologies. As Nunes et al [27] argue, devices that look overly clinical may alienate users, particularly in contexts where mental health remains stigmatized. While mobile technologies are often preferred for their neutral appearance [9,31], they also face high dropout rates and low engagement [32]. The OBT's single-purpose design appears to support stronger therapeutic involvement, but its visibility may also create barriers in settings where discretion is essential.

This suggests that therapists should actively explore the social contexts in which patients will use such wearable technologies. Helping patients prepare for how the OBT may be interpreted by others can reduce discomfort and support sustained use. Future research should also consider how these dynamics unfold across different cultural settings, where visibility and stigma may carry distinct meanings.

Limitations

This study has some limitations. First, none of the participants were native Danish speakers, which may have affected the depth and nuance of their interview responses. Language barriers likely limited their ability to fully articulate their experiences with the OBT. The decision to exclude interpreters and avoid translation was a deliberate trade-off, made to prioritize the authenticity of the patients' voices. However, this approach also meant that patients had to express themselves in Danish, which could lead to some loss of depth and nuance in their responses. In addition, limiting the sample to Danish-speaking patients may have excluded newly arrived refugees who had not yet learned the language, thereby limiting the cultural diversity of the sample and reducing visibility of culturally specific experiences with the OBT. Future studies could address this limitation by incorporating trained interpreters who are familiar with both the language and the study context, ensuring accurate and reliable data collection from nonnative speakers.

Second, the dual role of LGR as both researcher and PhD candidate may have influenced participants' responses, as evidenced when 2 participants inquired about LGR's academic progress. This interaction suggests participants' awareness of the researcher's personal stake in the study, potentially biasing their accounts toward more favorable descriptions of the OBT. To overcome this, interview questions were carefully designed

XSL•FO
RenderX

to explore both positive experiences and challenges related to OBT use, ensuring a balanced and comprehensive perspective.

Conclusion

This study explored how the OBT mediates patients' psychotherapeutic process in daily life. Addressing the research question, our results show that the OBT actively shaped patients' engagement with therapy by fostering agency, emotional regulation, and a sense of connection to their therapeutic process beyond clinical sessions.

The OBT was integrated into patients' daily routines, where it took on multiple roles based on their needs and contexts. These results challenge the traditional emphasis on passive data collection in mental health technologies, showing that self-tracking instruments can act as active mediators of therapeutic processes. Through its vibrotactile feedback and the act of tracking, the OBT became a tangible tool for self-awareness, nonverbal communication, and agency in moments of need. This study emphasizes the need for digital mental health technologies to focus not only on data but on their relational and experiential dimensions. Future research should build on these findings, examining how wearable technologies mediate therapeutic processes across diverse contexts and exploring design improvements to address issues such as stigma and visibility.

Taken together, the OBT illustrates how a simple, single-purpose, wearable self-tracking instrument can mediate patients' treatment during their daily lives, thereby bridging the gap between therapy sessions and real-world challenges. These insights contribute to a deeper understanding of how design principles for personalized wearable self-tracking instruments can support mental health care beyond passive data collection.

Acknowledgments

These materials have received financial support from the Danish Victims Fund (grant number: 20-610-00094). The execution, content, and results of the materials are the sole responsibility of the authors. The analysis and viewpoints that have been made evident from the materials belong to the authors and do not necessarily reflect the views of the Council of the Danish Victims Fund. We are grateful for the patients participating in this study, whose willingness to share their experiences was invaluable. Generative artificial intelligence tools were used to support grammar correction, language, and clarity. No content was generated, interpreted, or rewritten by artificial intelligence. All scientific content, analysis, and interpretations were developed entirely by the authors, who reviewed and approved all language adjustments.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to the sensitive nature of the data and the small participant group, which could compromise participant anonymity. The participants were refugees receiving psychotherapy for complex posttraumatic stress disorder, and consent was granted for the use of anonymized data within the scope of this study only. Deidentified excerpts or materials may be available from the corresponding author on reasonable request, subject to approval by the relevant ethics committee and in accordance with Danish data protection regulations.

Authors' Contributions

LGR, LH, SBM, TBC, and JEL designed the study. LGR conducted the analysis and the supervision of TBC, JEL, JA, and LH. JEL and TBC conducted the quantitative analysis of the data from the One Button Tracker and produced data visualizations. LGR wrote the manuscript which was reviewed by co-authors TBC, JEL, SBM, JA, and LH.

Conflicts of Interest

The One Button Tracker was invented and developed by authors JEL and TBC.

References

1. Borghare PT, Methwani DA, Pathade AG. A comprehensive review on harnessing wearable technology for enhanced depression treatment. Cureus 2024 Aug;16(8):e66173. [doi: 10.7759/cureus.66173] [Medline: 39233951]
2. Kang HS, Exworthy M. Wearing the future-wearables to empower users to take greater responsibility for their health and care: scoping review. JMIR Mhealth Uhealth 2022 Jul 13;10(7):e35684. [doi: 10.2196/35684] [Medline: 35830222]
3. Matthews M, Murnane E, Snyder J. Quantifying the changeable self: the role of self-tracking in coming to terms with and managing bipolar disorder. Human–Computer Interaction 2017 Nov 2;32(5-6):413-446. [doi: 10.1080/07370024.2017.1294983]
4. Weiser M. The computer for the 21st century. Sci Am 1991 Sep;265(3):94-104. [doi: 10.1038/scientificamerican0991-94]
5. Lupton D, Maslen S. The more-than-human sensorium: sensory engagements with digital self-tracking technologies. The Senses and Society 2018 May 4;13(2):190-202. [doi: 10.1080/17458927.2018.1480177]
6. Fedor S, Lewis R, Pedrelli P, Mischoulon D, Curtiss J, Picard RW. Wearable technology in clinical practice for depressive disorder. In: Friend SH, Ginsburg GS, Picard RW, Drazen JM, editors. N Engl J Med 2023 Dec 28;389(26):2457-2466. [doi: 10.1056/NEJMra2215898] [Medline: 38157501]

XSL•FO
RenderX

7. Rottstädt F, Becker E, Wilz G, Croy I, Baumeister H, Terhorst Y. Enhancing the acceptance of smart sensing in psychotherapy patients: findings from a randomized controlled trial. Front Digit Health 2024;6:1335776. [doi: 10.3389/fdgth.2024.1335776] [Medline: 38698889]
8. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol 2008;4:1-32. [doi: 10.1146/annurev.clinpsy.3.022806.091415] [Medline: 18509902]
9. Balaskas A, Schueller SM, Cox AL, Doherty G. Ecological momentary interventions for mental health: a scoping review. In: Myers B, editor. PLOS ONE 2021;16(3):e0248152. [doi: 10.1371/journal.pone.0248152] [Medline: 33705457]
10. Faurholt-Jepsen M, Vinberg M, Frost M, et al. Behavioral activities collected through smartphones and the association with illness activity in bipolar disorder. Int J Methods Psychiatr Res 2016 Dec;25(4):309-323. [doi: 10.1002/mpr.1502] [Medline: 27038019]
11. Pahwa M, McElroy SL, Priesmeyer R, et al. KIOS: a smartphone app for self-monitoring for patients with bipolar disorder. Bipolar Disord 2024 Feb;26(1):84-92. [doi: 10.1111/bdi.13362] [Medline: 37340215]
12. Pedrelli P, Fedor S, Ghandeharioun A, et al. Monitoring changes in depression severity using wearable and mobile sensors. Front Psychiatry 2020;11:584711. [doi: 10.3389/fpsyt.2020.584711] [Medline: 33391050]
13. Heron KE, Smyth JM. Ecological momentary interventions: incorporating mobile technology into psychosocial and health behaviour treatments. Br J Health Psychol 2010 Feb;15(Pt 1):1-39. [doi: 10.1348/135910709X466063] [Medline: 19646331]
14. Riisager LG, Larsen JE, Huniche L, Christiansen TB, Moeller SB. Collaborative development of a self-tracking assisted psychotherapy treatment concept for refugees with complex PTSD: participatory action research. JMIR Preprints. Preprint posted online on Sep 19, 2024. [doi: 10.2196/preprints.66663]
15. van de Belt TH, de Croon A, Freriks F, Blomseth Christiansen T, Eg Larsen J, de Groot M. Barriers to and facilitators of using a one button tracker and web-based data analytics tool for personal science: exploratory study. JMIR Form Res 2022 Mar 1;6(3):e32704. [doi: 10.2196/32704] [Medline: 35230247]
16. Torous J, Michalak EE, O'Brien HL. Digital health and engagement-looking behind the measures and methods. JAMA Netw Open 2020 Jul 1;3(7):e2010918. [doi: 10.1001/jamanetworkopen.2020.10918] [Medline: 32678446]
17. Ihde D. Technology and the Lifeworld: From Garden to Earth: Bloomington: Indiana University Press; 1990.
18. Verbeek PP. What Things Do: Philosophical Reflections on Technology, Agency, and Design 2 Printing: Pennsylvania State University Press; 2005.
19. Roberts N, Cloitre M, Brewin CR. International trauma interview (ITI) for ICD-11 PTSD and complex PTSD (test version 3.2). 2019.
20. Riisager LG, Christiansen TB, Moeller SB, Huniche L, Larsen JE. Snapping out of it: how a wearable for self-tracking assisted psychotherapy bridges the gap between thoughts and the world. Presented at: CHI '24: CHI Conference on Human Factors in Computing Systems; May 11-16, 2024; Honolulu HI USA. [doi: 10.1145/3613905.3637112]
21. Kvale S, Brinkmann S. Interviews: Learning the Craft of Qualitative Research Interviewing, 3rd edition: Sage Publications; 2014.
22. Braun V, Clarke V. One size fits all? What counts as quality practice in (reflexive) thematic analysis? Qual Res Psychol 2021 Jul 3;18(3):328-352. [doi: 10.1080/14780887.2020.1769238]
23. Braun V, Clarke V. Thematic analysis: a practical guide. bpsqmip 2022;1(33):46-50. [doi: 10.53841/bpsqmip.2022.1.33.46]
24. Rosenberger R, Verbeek P. Postphenomenological Investigations: Essays on Human-Technology Relations: Lexington Books; 2015.
25. Verbeek PP. Cyborg intentionality: rethinking the phenomenology of human–technology relations. Phenom Cogn Sci 2008 Sep;7(3):387-395. [doi: 10.1007/s11097-008-9099-x]
26. Feng S, Mäntymäki M, Dhir A, Salmela H. How self-tracking and the quantified self promote health and well-being: systematic review. J Med Internet Res 2021 Sep 21;23(9):e25171. [doi: 10.2196/25171] [Medline: 34546176]
27. Nunes F, Verdezoto N, Fitzpatrick G, Kyng M, Grönvall E, Storni C. Self-care technologies in HCI: trends, tensions, and opportunities. ACM Trans Comput-Hum Interact 2015 Dec 14;22(6):1-45. [doi: 10.1145/2803173]
28. Schroeder J, Karkar R, Murinova N, Fogarty J, Munson SA. Examining opportunities for goal-directed self-tracking to support chronic condition management. Proc ACM Interact Mob Wearable Ubiquitous Technol 2019 Dec;3(4):1-26. [doi: 10.1145/3369809] [Medline: 32656490]
29. Feenberg A. Active and passive bodies: comments on Don Ihde's bodies in technology. Techné Res Philos Technol 2003;7(2):125-130. [doi: 10.5840/techne2003725]
30. Ihde D. Bodies in Technology: University of Minnesota Press; 2002.
31. Faurholt-Jepsen M, Frost M, Christensen EM, Bardram JE, Vinberg M, Kessing LV. The effect of smartphone-based monitoring on illness activity in bipolar disorder: the MONARCA II randomized controlled single-blinded trial. Psychol Med 2020 Apr;50(5):838-848. [doi: 10.1017/S0033291719000710] [Medline: 30944054]
32. Pratap A, Neto EC, Snyder P, et al. Indicators of retention in remote digital health studies: a cross-study evaluation of 100,000 participants. NPJ Digit Med 2020;3(1):21. [doi: 10.1038/s41746-020-0224-8] [Medline: 32128451]

XSL•FO
RenderX

Abbreviations

CPTSD: complex posttraumatic stress disorder
CTTS: Clinic for Trauma and Torture Survivors
OBT: One Button Tracker

Edited by A Mavragani; submitted 23.12.24; peer-reviewed by E Nelson; revised version received 20.05.25; accepted 21.05.25; published 09.07.25.

Please cite as:
Riisager LG, Moeller SB, Larsen JE, Christiansen TB, Aagaard J, Huniche L
Use of a Wearable Self-Tracking Instrument by Refugees With Complex Posttraumatic Stress Disorder: A Qualitative Study of Psychotherapeutic Mediation and Engagement
JMIR Form Res 2025;9:e70511
URL: https://formative.jmir.org/2025/1/e70511
doi:10.2196/70511

© Lisa Groenberg Riisager, Stine Bjerrum Moeller, Jakob Eg Larsen, Thomas Blomseth Christiansen, Jesper Aagaard, Lotte Huniche. Originally published in JMIR Formative Research (https://formative.jmir.org), 9.7.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Correlation of Biomechanical Variables of Lower Extremity Movement During Functional Tests and Tasks in Youth League Football Players: Cross-Sectional Correlation Study

Anna Davidoviča[1], MHS; Sergejs Davidovičs[1], MHS; Guna Semjonova[1], PhD; Alexei Katashev[2], Prof Dr; Alexander Oks[3], Prof Dr; Linda Lancere[4], PhD; Signe Tomsone[1], Prof Dr Med; Maksims Zolovs[5,6], PhD

[1]Department of Rehabilitation, Riga Stradiņš University, Anniņmuižas bulvāris 26A, Riga, Latvia
[2]Institute of Mechanical and Biomedical Engineering, Riga Technical University, Riga, Latvia
[3]Institute of Architecture and Design, Riga Technical University, Riga, Latvia
[4]Department of Sociotechnical Systems Modelling, Vidzeme University of Applied Sciences, Valmiera, Latvia
[5]Statistics Unit, Riga Stradiņš University, Riga, Latvia
[6]Institute of Life Sciences and Technology, Daugavpils University, Daugavpils, Latvia

Corresponding Author:
Anna Davidoviča, MHS
Department of Rehabilitation, Riga Stradiņš University, Anniņmuižas bulvāris 26A, Riga, Latvia

Abstract

Background: Football is the most widely played sport globally but carries a high risk of lower limb injuries, particularly among youth athletes. Real-time biomechanical monitoring can play a critical role in injury prevention. However, traditional lab-based systems are often complex and impractical for field use. Recent advances in wearable technology, such as inertial sensors and smart socks, provide more accessible solutions for movement analysis. The DAid smart sock system is a promising tool, but further evidence is needed to support its use in lower extremity functional assessments.

Objective: This correlational study aimed to investigate the correlation between lower limb joint angles, muscle activity, and plantar pressure distribution during the “Single Leg Squat” (SLS) and its variations in youth football players, using wireless wearable sensors in a field-based setting.

Methods: In total, 32 youth football players (16 male and 16 female individuals; mean age 14.6, SD 0.5 years) performed SLS movements while wearing the NOTCH inertial motion sensors, DAid smart socks (plantar pressure), and PLUX muscleBAN EMG system. Spearman correlation was used to explore relationships between hip, knee, and ankle joint kinematics; muscle activity (gluteus medius, gluteus maximus [GMx], vastus lateralis, and biceps femoris); and changes in center of pressure (COP) on the plantar surface.

Results: A strong positive correlation was found between hip adduction and medial foot pressure (COP1X: $\rho=0.785$, $P<.001$). Knee flexion was strongly correlated with gluteus medius ($\rho=0.809$) and GMx ($\rho=0.841$) muscle activity. Hip internal rotation also showed moderate to strong correlations with both COP variables (COP1Y: $\rho=0.585$) and GMx activation ($\rho=0.477$). Significant gender-specific differences were identified: male individuals showed stronger correlations between joint angles and muscle activation, while female individuals demonstrated a strong correlation between knee flexion and overall plantar pressure (COP2W: $\rho=0.818$). Several moderate correlations ($0.35<\rho<0.47$) further confirmed interactions between joint movement, muscle activity, and plantar pressure.

Conclusions: The findings support the feasibility and utility of wireless wearable sensors—including inertial measurement units and smart socks—for in-field biomechanical analysis in youth football players. The study confirmed key relationships between joint mechanics and plantar pressure distribution, suggesting their relevance in injury risk screening. The DAid smart sock system, in particular, demonstrated reliable performance for assessing medial-lateral loading patterns associated with hip and knee movement. These insights may help guide neuromuscular training and individualized injury prevention strategies in young athletes.

(JMIR Form Res 2025;9:e69046) doi:10.2196/69046

KEYWORDS

biomechanics; youth athletes; wearable technology; postural control; muscle activity; center of pressure; sports injury prevention; wireless sensors; functional movement

XSL•FO
RenderX

Introduction

Football is the most widely played sport globally, with around 400 million participants across 208 countries [1]. Among the 38 million registered players, over half are under 18 years old [2]. Due to its high-intensity nature, football ranks among the top 5 sports with the highest injury risk [2,3]. Injuries commonly involve the lower extremities, particularly the thighs, ankles, knees, and hips, with frequent issues including muscle damage, strains, and contusions [4]. Youth football players are especially vulnerable to injuries, particularly during growth spurts, with one in 3 players experiencing injuries each season, many of which are noncontact [3,5]. Altered movement patterns such as dynamic knee valgus are a significant biomechanical risk factor for injuries, such as anterior cruciate ligament (ACL) tears [6]. Understanding and managing this risk factor is crucial for effective injury risk reduction. Preventive exercise programs like Fédération Internationale de Football Association (FIFA) 11+ [7] have proven effective in reducing injury risk by focusing on muscle strength, balance, and dynamic stability, with studies demonstrating a 30% reduction in injury rates [1]. A key component of this program, the "Single Leg Squat" exercise, is designed to improve lower limb movement patterns and balance. However, despite the success of FIFA 11+, there remains a gap in providing personalized, real-time biofeedback on lower limb biomechanics during training, particularly concerning the functional tasks that expose youth players to injury risks [8]. Assessing an entire team simultaneously during exercises and tests limits the ability to provide personalized feedback, which is essential for progressing athletes to more advanced exercises and more intensive training sessions [8]. Coaches often provide individualized feedback during training, but this feedback can be subjective and varies based on the coach's experience [9]. An objective approach to providing feedback can be achieved through motion-capture biofeedback systems [10-12]. While optical motion capture, Microsoft Kinect cameras [13,14], 3D kinematic analysis [15], PhaseSpace systems [16], and force plates [17] are commonly used for biomechanical assessments, they present challenges that hinder their practical application in football training. Optical motion capture systems require a fixed setup with a restricted field of view, making them unsuitable for on-field use [18,19]. Force plates, which measure ground reaction forces and foot plantar pressure, are typically confined to laboratory settings, limiting their practicality for daily football practice [20,21]. Given these limitations, wireless smart sensor systems present a promising alternative for providing real-time biofeedback on lower limb motion and biomechanics. Inertial measurement unit systems can capture data on specific body forces, angular velocity, and magnetic fields [22,23]. Wireless sensors that monitor muscle activity in real-time, using electromyographic signals, can provide insights into muscle performance and condition [24,25]. In addition, smart insole systems and smart socks, such as DAid Pressure Sock systems, measure biomechanical parameters of the foot, including the center of pressure (COP) and pressure distribution across the plantar surface [26-30]. These sensor systems offer significant benefits in sports medicine due to their cost-effectiveness, ease of use, and lightweight design, making advanced monitoring technology accessible to a broader range of sports teams and medical professionals [31,32]. However, challenges include synchronizing data from multiple sensors, signal interference, and potential restrictions on natural movement [30,33]. Previous studies have demonstrated a relationship between lower limb biomechanical variables during functional tasks. Increased knee muscle strength is associated with reduced knee valgus motion [34-38]. In addition, higher levels of lower limb muscle preactivity—particularly in the quadriceps and hamstrings—are linked to a greater peak knee valgus angle [35,39]. Regarding foot biomechanics, a positive correlation exists between greater foot pronation and increased pressure on the front to medial plantar surface during functional squat activities [21]. Considering the difficulties involved in using multiple wireless sensors simultaneously and the importance of accurately tracking foot biomechanics and limb movement, this study aims to investigate correlations between plantar pressure measurements, electrical activity of muscles, and lower extremity movement angles during functional tests like the "Single Leg Squat" in youth league football players. The hypothesis states that there will be strong correlations among key biomechanical variables—joint angles, muscular activation, and foot plantar pressure—during the "Single Leg Squat" test, using measurements from wireless sensor systems in field tests. Specifically, the position of the COP on the plantar surface is expected to be significantly associated with lower limb movement patterns and muscle activity. This study will demonstrate the relevance of wireless sensor systems for assessing lower limb biomechanics in football players, with the DAid smart sock system serving as one of the tools to facilitate this assessment.

Methods

Study Design

To obtain the aim of the study, the correlation research design was used, where associations between variables can be positive (variables change in the same direction), negative (variables change in opposite directions), or null (no relationship). The sample size was estimated based on Fraenkel and Wallen [40], who recommend a minimum of 30 participants for correlational studies to ensure reliable results. Given the target population and study design, a total of 32 participants was considered sufficient for meaningful statistical analysis.

Participants

The study involved 32 youth football players (both male and female) from the Latvian Youth Football League (U-14 and U-15), recruited between September 2023 and December 2023. Participants were required to be between 14 and 15 years old, actively playing in a youth football league, and have at least 5 years of experience in football. They had to be free of pain in the hip, knee, or ankle joints during movement and voluntarily agree to participate in the study. Participants were excluded if they had any lower limb disease, deformity, injury, or surgery within the past 12 months. Those with vestibular disorders that prevented them from performing the "Single Leg Squat" test and its variations were also excluded. Additional exclusion criteria included the presence of metallic implants in areas where sensors were applied and the use of knee or ankle braces or

XSL•FO
RenderX

kinesiology tape. Both limbs were evaluated in this study; however, differences between the dominant and nondominant legs were not specifically analyzed. Limb dominance was determined based on the leg preferred for kicking a ball, a commonly used method in sports biomechanics research [41]. All participants completed the "Single Leg Squat" functional test, wearing tight-fitting sportswear that allowed accurate measurements and sports shoes with appropriate insoles.

Ethical Considerations

The study was conducted in compliance with the Declaration of Helsinki and Latvian laws, adhered to the European Union's General Data Protection Regulation 2016/679. Before participation, the study's procedures were explained to participants and their guardians, and informed consent was obtained. Ethical approval was granted by the Riga Stradins University Research Ethics Committee (approval number 2-PĒK-4/294/2023). Participation was voluntary, with the right to withdraw at any time, and participant confidentiality was maintained throughout the study. The participants did not receive any compensation for their participation in the study.

Instrumentation and Data Collection

Data collection involved three main instruments: the DAid Pressure Sock System (Riga, Latvia), the NOTCH Inertial Sensor System (Notch Interfaces Inc, New York, United States), and the PLUX Wireless Biosignals system (PLUX Wireless Biosignals, Lisbon, Portugal).

DAid Pressure Sock System

This system, equipped with 6 sensors in the sole, measured plantar pressure at up to 200 Hz per channel. The sensors were positioned under the heel, arch, and metatarsal heads to monitor gait and assess foot pressure distribution. Data were transmitted via Bluetooth to a remote device for storage. The system was calibrated using participants' weight shifts. Activated before testing, the socks were connected via Bluetooth to "Fastreader" software in the LabView environment. The DAid records relative pressure values under each sensor to calculate the COP coordinates along the mediolateral (COPx) and anteroposterior (COPy) axes, as well as the overall COP position (COPw). More details about this system have been described in previous studies [29,30].

NOTCH Inertial Sensor System

The NOTCH inertial measurement unit system, comprising 9-axis inertial sensors, was used to measure lower limb angles during movement. Using 6 sensors placed on the lower body, data were wirelessly transmitted to a mobile device and processed to map a 3D human skeleton. The system recorded at 40 Hz for 10-second intervals during the "Single Leg Squat" test. Sensors were calibrated and placed according to the user manual. The 6 inertial sensors were positioned at specific anatomical landmarks for accurate measurement: one on the sternum, one on the waist, and one on each thigh and calf (placed laterally at the midthigh and midcalf regions). Sensors were color-coded and placed according to the app's guidelines. The "Lower body + hip" configuration was used, and participants followed specific positioning instructions to ensure accuracy.

PLUX Wireless Biosignals (MuscleBAN Kit)

The PLUX Wireless Biosignals system with an electromyographic add-on was used to collect muscle electrical activity data. The system includes an electromyographic sensor, triaxial accelerometer, and magnetometer, capturing data at up to 1000 Hz with 16-bit resolution. The device, powered by an internal battery and featuring dual Bluetooth modules, provided real-time muscle activity and movement data for analysis. Four electrodes were placed to measure muscle activity in this study. Electrode placement was identified according to Surface Electromyography for the Non-Invasive Assessment of Muscles (SENIAM) guidelines [42]. Electrodes were positioned on the vastus lateralis (VL), gluteus maximus (GMx), gluteus medius (GM), and biceps femoris (BF). After placing the electrodes, participants performed a maximum voluntary contraction (MVC) to evaluate muscle activity, achieved by performing maximal contractions against a stable resistance [42]. Muscle activity was recorded throughout the execution of each "Single Leg Squat" repetition, ensuring continuous data collection during movement. The recording duration per trial was 10 seconds, covering the entire squat cycle. Electromyographic signals were normalized [43], using electromyographic data recorded during MVC as the reference. The signals were processed by calculating the root mean square from the rectified signal with a 0.2-second window. The highest root mean square value from all test repetitions was used to normalize the electromyographic signals, assessing muscle activity relative to its maximum neural activation. During the "Single Leg Squat" test and its variations, the PLUX system continuously measured the electrical activity of the VL, GM, GMx, and BF muscles, providing a dynamic assessment of neuromuscular activation patterns.

Study Procedure

Participants' demographic and physical characteristics, such as age, height, weight, shoe size, playing position, and dominant leg, were recorded. Each participant performed the "Single Leg Squat" (SLS) and its 3 variations, completing 3 repetitions of each type (Textbox 1):

Textbox 1. Description of Single Leg Squat (SLS) Variations: Front (SLS_F), Middle (SLS_M), and Back (SLS_B) based on non-supporting leg position while maintaining a 60° knee flexion in the supporting leg.

- Single Leg Squat – Front: Hands placed on the hips, the non-supporting leg extended forward, and the supporting leg performing a squat to a 60-degree knee flexion angle.
- Single Leg Squat – Middle: Hands placed on the hips, the non-supporting leg positioned in the middle and bent at a 90-degree knee angle, with the supporting leg performing a squat to a 60-degree knee flexion angle.
- Single Leg Squat – Back: Hands placed on the hips, the non-supporting leg extended backward and bent at a 90-degree knee angle, with the supporting leg performing a squat to a 60-degree knee flexion angle.

XSL•FO
RenderX

During the execution of the "Single Leg Squat" and its variations, participants ensured continuous foot contact with the ground. Valid trials were considered those in which participants maintained balance throughout all 3 repetitions. If an incorrect execution occurred, participants were asked to repeat the attempt. A 3-minute rest period was observed between trials to prevent muscle fatigue.

Throughout the execution of all "Single Leg Squat" tests and their variations, participants wore the DAid smart sock system, the PLUX Wireless Biosignals (muscleBAN kit) for electromyography, and the NOTCH inertial sensor system, as shown in Figure 1.

Figure 1. Study participants interacting with the DAid smart sock system, the PLUX Wireless Biosignals (muscleBAN kit) wireless electromyography device, and the NOTCH inertial sensor system.

Biomechanical Variables for Determining the Movement of the Lower Extremities

Each youth league football player provided data to assess biomechanical movement variables during the "Single Leg Squat" functional test and its variations.

Angles from the NOTCH inertial sensor system: The dataset for the "Lower body+ hip" configuration included dynamic knee valgus determining angles: (1) Angles_Thigh: Flexion and extension, abduction and adduction, and internal and external rotation angles of the hip joint. (2) Angles_LowerLeg: Flexion and extension angles of the knee joint.

COP from the DAid smart sock system: This dataset included COP1X, COP1Y, COP1W, COP2X, COP2Y, and COP2W values, which represent the center of pressure on the plantar surface of the foot. Positive COPX values indicate the medial position, negative values indicate the lateral position, positive COPY values indicate the anterior position, and negative values indicate the posterior position. COPW values represent the total center of pressure. Two mathematical methods (COP1 and COP2) were used to calculate the center of pressure.

Electromyography data from the wireless PLUX system: Electrical activity of VL, GM, GMx, and BF and MVC values.

Data Statistical Analysis

To synchronize datasets from the NOTCH, DAid, and PLUX for biomechanical movement analysis of the lower extremities, data consolidation was performed using the LabView environment. Each participant provided 64 data files, resulting in 2048 files from 32 participants, which were compiled into a unified Database.txt file for correlation analysis. Data analysis was conducted with Microsoft Excel v.16.77.1 and jamovi v.2.3.28.0 (an open-source graphical user interface for the R programming language) using descriptive statistics to describe participant characteristics and results.

XSL•FO
RenderX

Spearman's correlation analysis, a nonparametric method, was used to determine correlations, considering the Spearman correlation coefficient (ρ). Positive correlations indicate that as one variable increases, so does the other, while negative correlations mean that as one increases, the other decreases. The strength of these relationships is measured by the (ρ up to 0.2 indicates a weak correlation, 0.2 to 0.5 indicates a moderate correlation, and 0.5 to 1 indicates a strong correlation. These correlation strength levels are based on Cohen (1988) [44,45]. Statistical significance was set at $P<.05$, with analyses conducted using jamovi v.2.3.28.0.

Results

Participant Description

A total of 32 participants, including 16 female and 16 male individuals, all of whom were youth league football players, were involved in this study. Participants were youth football league athletes who met specific inclusion and exclusion criteria (Table 1).

The positions of the youth league football players (study participants) on the field varied—12 players were defenders, 9 were midfielders, and 11 were forwards.

Table . Demographic and anthropometric characteristics of youth football players participating in the study (n=32; 16 male and 16 female individuals). Variables include age, body mass index (BMI), height, weight, and European shoe size. These characteristics support contextual interpretation of the biomechanical and sensor data collected during Single Leg Squat testing.

Variable	Mean (SD)	Median	Minimum	Maximum	Unit
BMI	20.8 (2.065)	20.8	17.3	27.2	kg/m²
Age	14.6 (0.495)	15.0	14	15	years
Weight	59.8 (6.413)	58.0	50.0	78.6	kg
Height	169.4 (4.662)	169.0	159	181	cm
EU[a] Size	40.6 (1.961)	40.0	37.0	44.0	EU shoe size

[a]EU size: European shoe size measurement.

Correlation Analysis

Strong Positive Correlation

A statistically significant relationship was found among the study participants between hip joint adduction movement and changes in the center of pressure on the plantar surface of the foot COP1X (ρ=0.785; $P<.001$), representing the position of the pressure center in the medial part of the plantar surface of the foot (Figure 2).

Figure 2. Strong positive correlation (ρ=0.785, $P<.001$) between hip joint adduction and medial plantar pressure (COP1X) during the SLS in youth football players (n=32, aged 14 - 15 years). Joint kinematic data were recorded using the NOTCH inertial motion capture system, and plantar pressure was measured using the DAid smart sock system.

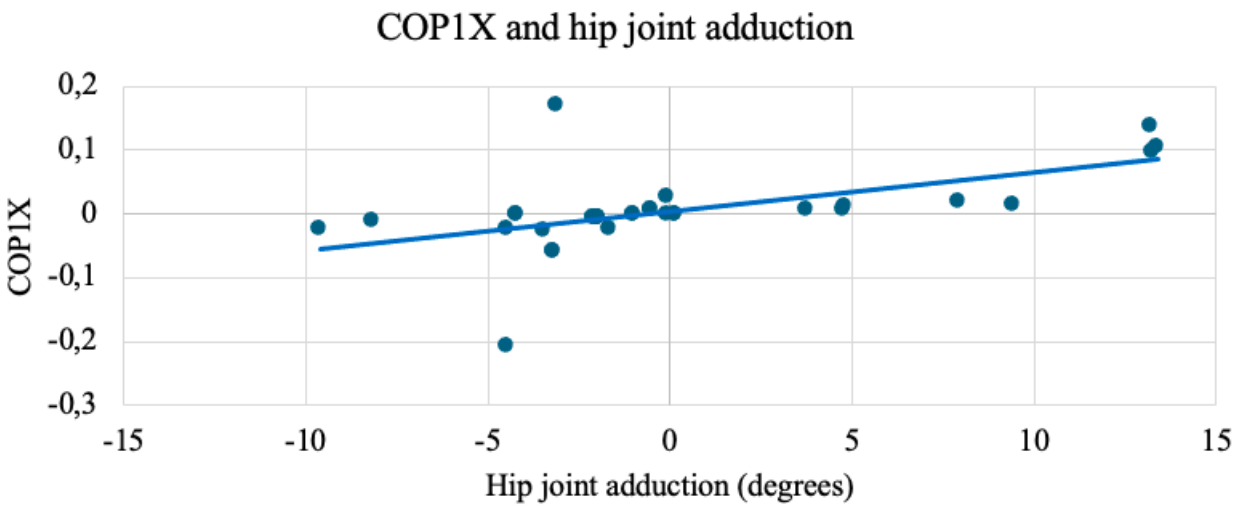

A statistically significant relationship was also found between knee joint flexion and the electrical activity of GM muscle (ρ=0.66; $P<.001$; Figure 3) and between hip joint internal rotation with changes in the COP on the plantar surface of the foot COPY1, representing the position of the pressure center in the anterior part of the plantar surface of the foot (ρ=0.585; $P<.01$; Figure 4), and between changes in the center of pressure on the plantar surface of the foot COP2X, representing the position of the pressure center in the medial part of the plantar surface of the foot, and the electrical activity of GM muscle (ρ=0.568; $P<.001$; Figure 5), and between the electrical activities of the GM and VL muscles (ρ=0.696; $P<.001$; Figure 6).

XSL•FO
RenderX

Figure 3. Strong positive correlation (ρ=0.66, *P*<.001) between knee joint flexion and GM muscle activity during the Single Leg Squat test in youth football players. Knee flexion was measured using the NOTCH® inertial sensor system, and muscle activity was recorded via the PLUX Wireless Biosignals surface electromyography system. GM: gluteus medius.

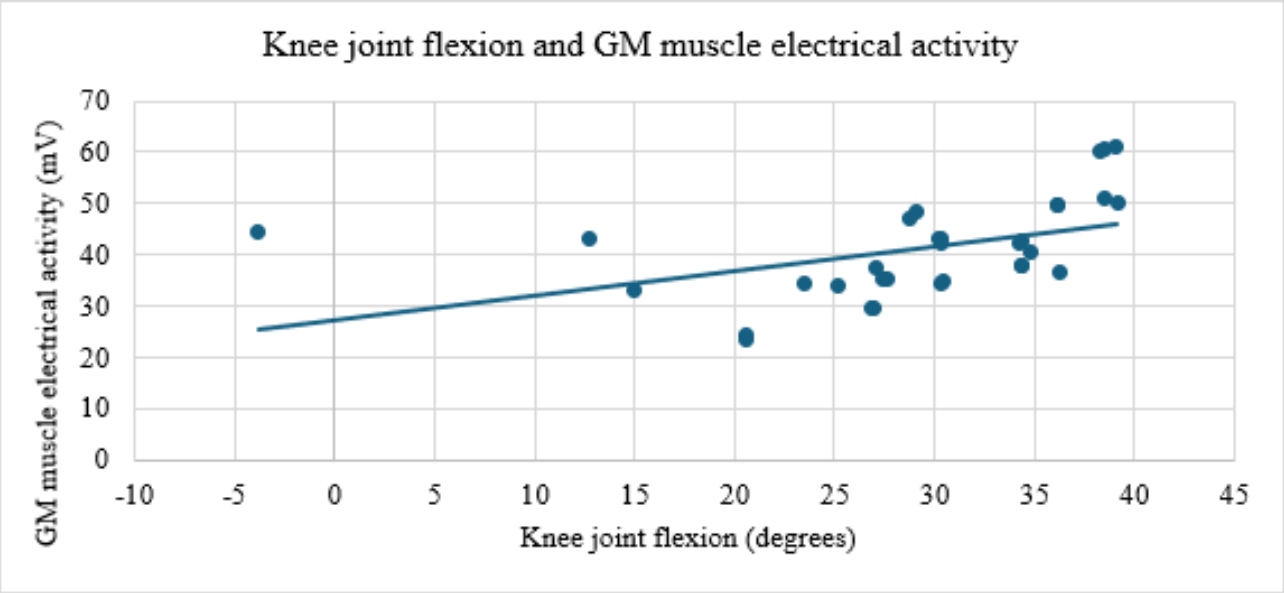

Figure 4. Strong positive correlation (ρ=0.585, *P*<.01) between hip internal rotation and anterior plantar pressure (COPY1) during the SLS. Hip rotation was captured using NOTCH motion sensors, and plantar pressure shifts were measured with the DAid smart sock system.

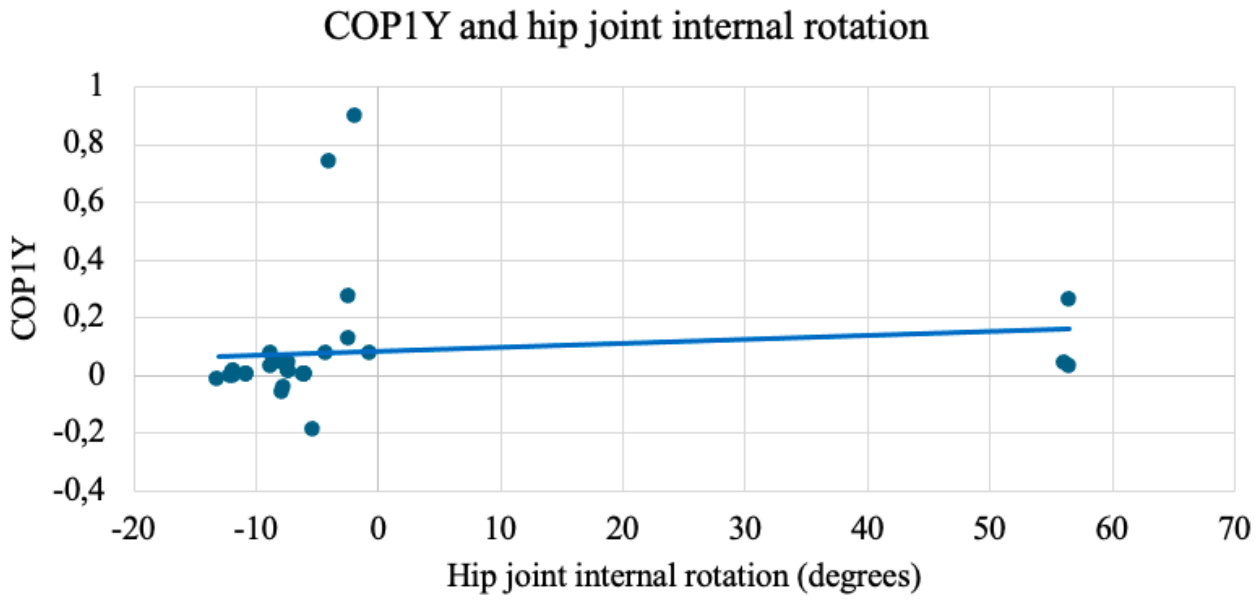

XSL•FO
RenderX

Figure 5. Strong positive correlation (ρ=0.568, *P*<.001) between changes in the center of pressure on the plantar surface of the foot (COP2X) and gluteus medius muscle activity during functional movement assessment. Data were collected using DAid smart socks for pressure and PLUX surface electromyography sensors for muscle activity. GM: gluteus medius.

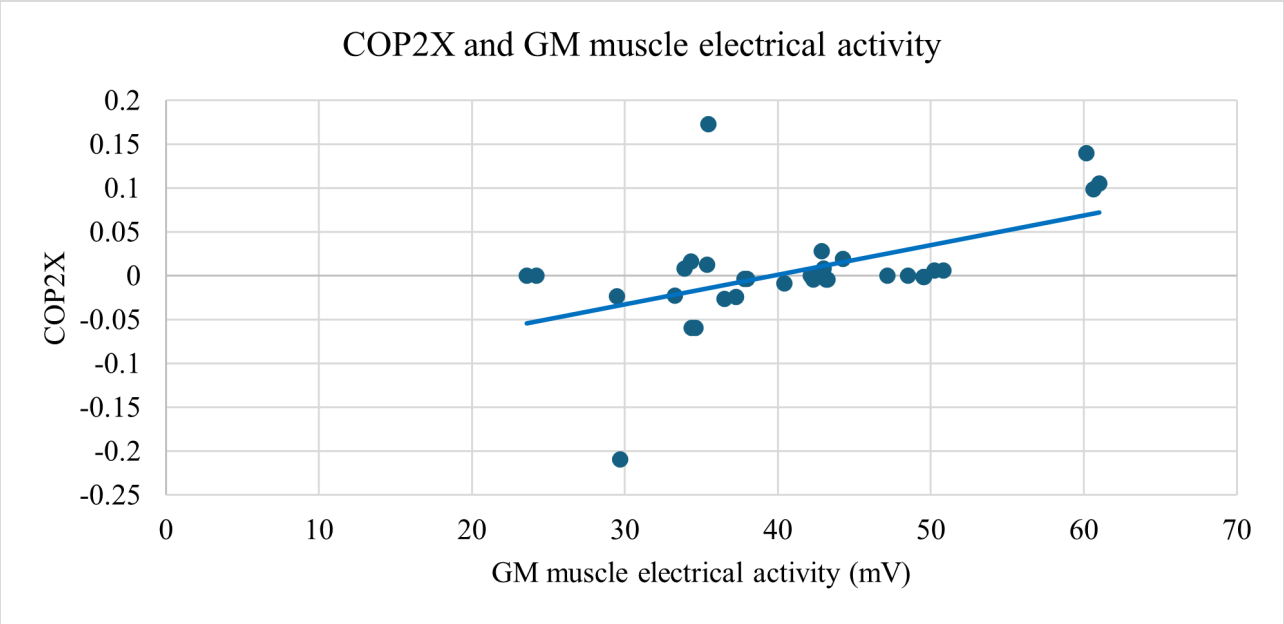

Figure 6. Strong positive correlation (ρ=0.696, *P*<.001) between GM and VL muscle activity during the SLS test. Both muscle signals were recorded using PLUX surface electromyography. GM: gluteus medius; VL: vastus lateralis.

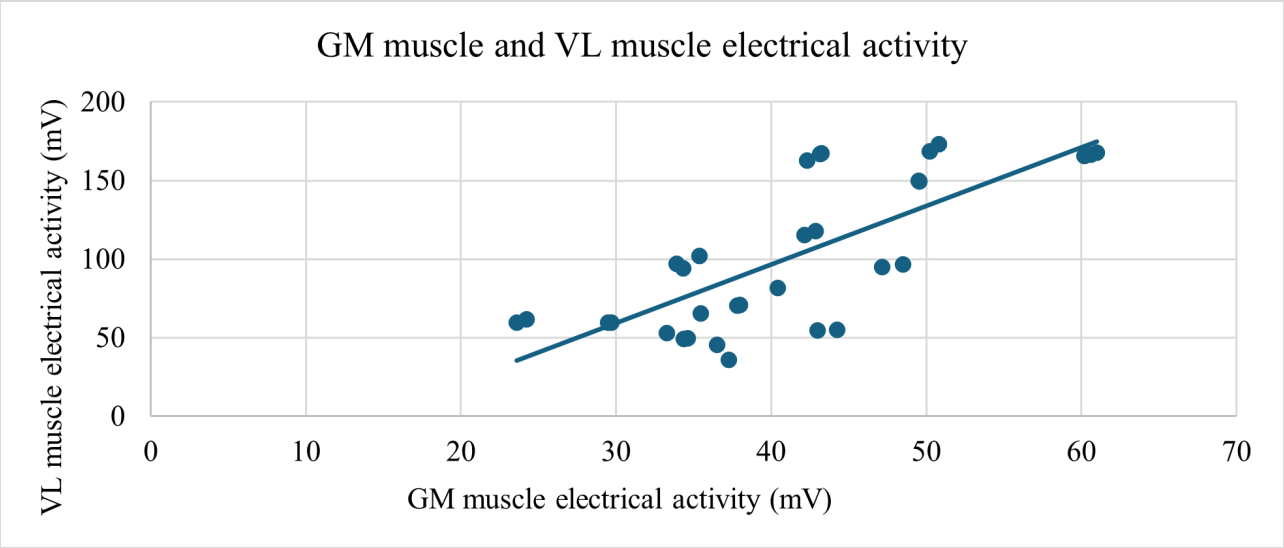

Strong Negative Correlation

A statistically significant relationship was found among the study participants between changes in the COP on the plantar surface of the foot COP2W, representing the overall position of the pressure center on the plantar surface of the foot, and the electrical activity of GMx muscle (ρ=−0.592; *P*<.001) (see Figure 7) and between the electrical activities of BF and VL muscles (ρ=−0.539; *P*<.01; Figure 8).

XSL•FO
RenderX

Figure 7. Strong negative correlation (ρ=-0.592, *P*<.001) between overall plantar pressure (COP2W) and GMx muscle activity. Plantar pressure was measured using the DAid smart sock system; muscle activity was measured with PLUX surface electromyography. GMx: gluteus maximus.

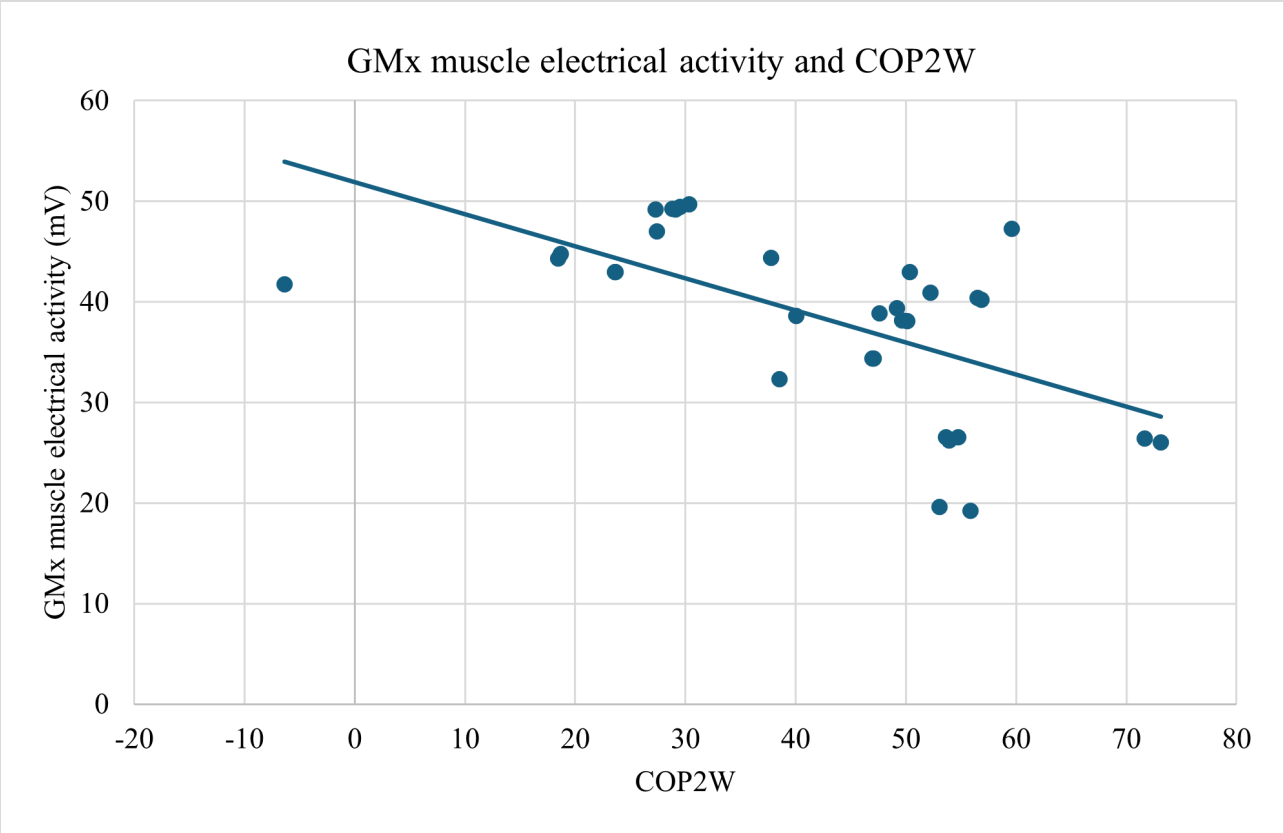

Figure 8. Strong negative correlation (ρ=-0.539, *P*<.01) between BF and VL muscle activation during the SLS. Both muscle activities were assessed using the PLUX Wireless Biosignals surface electromyography system. BF: biceps femoris; GM: gluteus medius; VL: vastus lateralis.

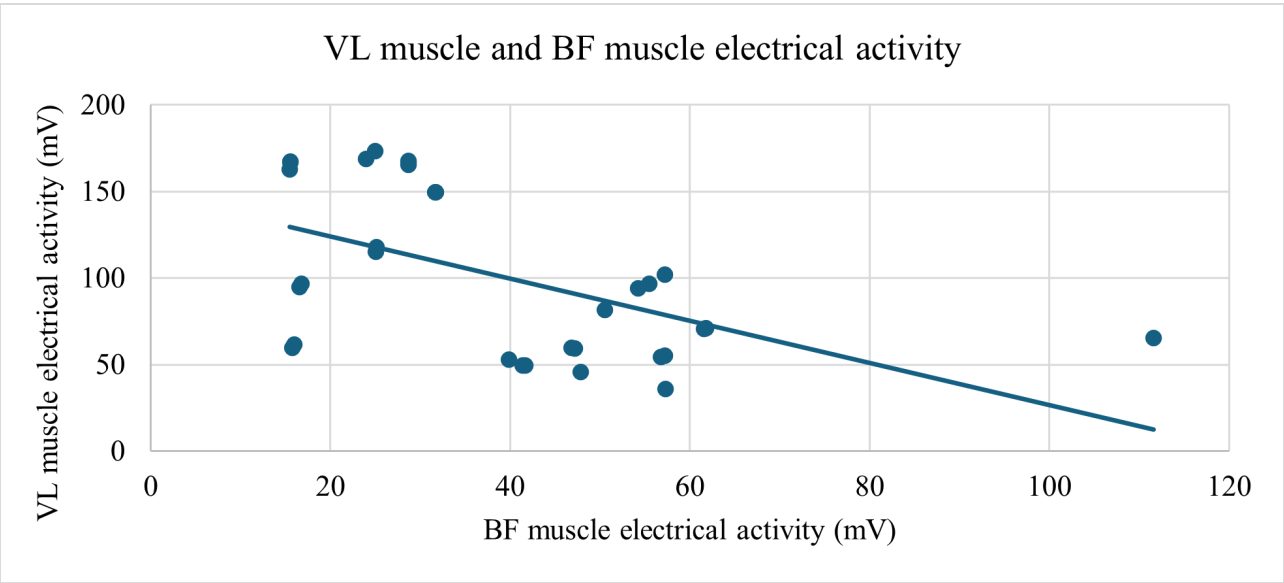

Moderate Positive and Negative Correlation

Several statistically significant moderate correlations were found among the study participants (Table 2), indicating biomechanical relationships of varying strengths. A moderate positive correlation was observed between hip joint adduction and the electrical activity of the VL muscle (ρ=0.356, *P*<.05), suggesting that increased adduction movement corresponded with greater activation of the vastus lateralis. Similarly, hip joint internal rotation exhibited a moderate positive correlation with the electrical activity of the BF muscle (ρ=0.36, *P*<.05), indicating a link between internal rotation and posterior thigh muscle activation. Changes in the center of pressure on the plantar surface of the foot (COP1Y) were moderately correlated with the electrical activity of the GMx muscle (ρ=0.367, *P*<.05), while the mediolateral shifts of the center of pressure (COP1X) were moderately correlated with anterior-posterior changes in COP1Y (ρ=0.398, *P*<.05), reinforcing the interaction between

XSL•FO
RenderX

foot pressure distribution and lower limb movement. Further significant moderate positive correlations were found between COP2X and the electrical activity of the VL muscle (ρ=0.415, *P*<.05), hip joint flexion and COPY2 (ρ=0.426, *P*<.05), and COP2Y and the electrical activity of the BF muscle (ρ=0.426, *P*<.05). The COP1X variable was also moderately correlated with the electrical activity of the GM muscle (ρ=0.445, *P*<.05) and VL muscle (ρ=0.47, *P*<.01), demonstrating a relationship between pressure shifts and muscle activation. Additionally, hip joint internal rotation showed a moderate correlation with the electrical activity of the GMx muscle (ρ=0.477, *P*<.01).

Table . Moderate positive and moderate negative correlations between joint kinematics, muscle activity, and plantar pressure distribution during the SLS in youth football players. Data were derived using the NOTCH inertial sensor system, DAid smart sock system, and PLUX Wireless Biosignals surface electromyographic. All correlations are statistically significant (*P*<.05 or *P*<.01*)*.

Correlation type	First parameter	Second parameter	Coefficient (ρ)	Significance (*P* values)
Moderate positive				
	Hip joint adduction (°)[a]	VL[b] muscle electrical activity (mV)[c]	0.356	<.05
	Hip joint internal rotation (°)	BF[d] muscle electrical activity (mV)	0.36	<.05
	COP1Y (a.u.)[e]	GMx[f] muscle electrical activity (mV)	0.367	<.05
	COP1X (a.u.)	COP1Y (a.u.)	0.398	<.05
	COP2X (a.u.)	VL muscle electrical activity (mV)	*r*=0.415	<.05
	Hip joint flexion (°)	COPY2 (a.u.)	0.426	<.05
	COP2Y (a.u.)	BF muscle electrical activity (mV)	0.426	<.05
	COP1X (a.u.)	GM[g] muscle electrical activity (mV)	0.445	<.05
	COP1X (a.u.)	VL muscle electrical activity (mV)	0.47	<.01
	Hip joint internal rotation (°)	GMx muscle electrical activity (mV)	0.477	<.01
Moderate negative				
	Hip joint internal rotation (°)	COP2W (a.u.)	−0.368	<.05
	COP1W (a.u.)	Hip joint internal rotation (°)	-0.39	<.05
	COP2Y (a.u.)	GMx muscle electrical activity (mV)	−0.394	<.05
	COP2X (a.u.)	COP2W (a.u.)	−0.415	<.05
	COP1Y (a.u.)	COP2W (a.u.)	−0.437	<.05

[a](°): degrees - Joint angles.
[b]VL: vastus lateralis.
[c](mV): millivolts.
[d]BF: biceps femoris.
[e](a.u.): arbitrary units.
[f]GMx: gluteus maximus.
[g]GM: gluteus medius.

Moderate negative correlations were also identified in the study. Hip joint internal rotation was negatively correlated with COP2W (ρ=-0.368, *P*<.05), indicating that increased internal rotation was associated with a posterior shift in the pressure center. A negative correlation was also observed between the overall pressure center COP1W and hip joint internal rotation (ρ=-0.39, *P*<.05). In addition, COP2Y exhibited a moderate negative correlation with the electrical activity of the GMx muscle (ρ=-0.394, *P*<.05), while COP2X and COP2W were inversely related (ρ=-0.415, *P*<.05). Another significant moderate negative correlation was found between COP1Y and COP2W (ρ=-0.437, *P*<.05), indicating an interaction between anterior-posterior shifts in plantar pressure distribution.

XSL•FO
RenderX

Correlation Matrix Analysis by Gender

The study also examined correlations based on gender distribution (see Table 3).

Table . Correlation matrix for male youth football players (n=16), showing strong positive and negative associations between hip and knee joint angles, plantar pressure distribution, and muscle activity during the Single Leg Squat test. Measures were collected using a multi-sensor setup: NOTCH® for motion, DAid® for pressure, and PLUX for surface electromyography.

Correlation type	First parameter	Second parameter	Coefficient (ρ)	Significance (*P* values)
Strong positive				
	Hip joint adduction (°)[a]	Hip joint internal rotation (°)	0.806	<.001
	Hip joint adduction (°)	COP1X (a.u.)[b]	0.785	<.001
	Knee joint flexion (°)	GM[c] muscle electrical activity (mV)	0.809	<.001
	Knee joint flexion (°)	GMx[d] muscle electrical activity (mV)	0.841	<.001
	Knee joint flexion (°)	VL[e] muscle electrical activity (mV)	0.776	<.001
	COP1X (a.u.)	COP2X (a.u.)	0.832	<.001
	COP2X (a.u.)	GM muscle electrical activity (mV)	0.791	<.001
Strong negative	COP1Y (a.u.)	COP1W (a.u.)	-0.832	<.001

[a](°): degrees - Joint angles.

[b](a.u.): arbitrary units.

[c]GM: gluteus medius.

[d]GMx: gluteus maximus.

[e]VL: vastus lateralis.

Among male participants, a strong positive correlation was observed between hip joint adduction and hip joint internal rotation (ρ=0.806, *P*<.001), indicating that increased adduction movement was associated with greater internal rotation of the hip joint. In addition, hip joint adduction showed a strong positive correlation with changes in the COP on the plantar surface of the foot (COP1X) (ρ=0.785, *P*<.001), suggesting that medial shifts in the pressure center were linked to increased hip adduction. Knee joint flexion demonstrated a strong positive relationship with the electrical activity of the GM muscle (ρ=0.809, *P*<.001), GMx muscle (ρ=0.841, *P*<.001), and VL muscle (ρ=0.776, *P*<.001), confirming that increased knee flexion corresponded with heightened muscle activation in these regions. Moreover, a strong correlation was found between the changes in the COP variables COP1X and COP2X (ρ=0.832, *P*<.001), indicating a strong interaction between these mediolateral pressure shifts.

Similarly, COP2X showed a strong positive correlation with the electrical activity of the GM muscle (ρ=0.791, *P*<.001), reinforcing the link between plantar pressure adjustments and muscle activation patterns. A strong negative correlation was found between COP1Y and COP1W (ρ=-0.832, *P*<.001), suggesting that anterior-posterior shifts in the COP were inversely related to overall pressure distribution.

For female participants, a strong positive correlation was identified between knee joint flexion and the overall center of pressure on the plantar surface of the foot (COP2W; ρ=0.818, *P*<.001).

Discussion

Principal Results

The key findings of this study reveal strong positive and negative correlations between lower limb biomechanics during the “Single Leg Squat” in youth football players, using wireless sensor systems. These results provide important biomechanical insights that may help in identifying movement patterns associated with increased injury risk. Understanding the interaction between foot pressure distribution, joint angles, and muscle activity can guide targeted interventions to improve stability and prevent common football-related injuries, such as ACL injuries and patellofemoral pain syndrome [46,47]. A study by Ahn et al [21] demonstrated that altered foot biomechanics, particularly excessive pronation, may lead to compensatory changes in hip mechanics, such as increased internal hip rotation and hip adduction. This suggests that foot pronation may lead to altered load distribution and neuromuscular adaptations at the hip joint. As a result, foot mechanics play a crucial role in shaping lower limb kinematics and should be carefully considered when designing training programs for young athletes [21,48]. Similarly, Kim et al [10] found that foot pronation is associated with dynamic knee valgus, which involves a combination of hip adduction and internal rotation, contributing to medial knee displacement and increased risk of lower limb injuries. The findings of this study align with those of Kim et al [10], demonstrating that a medial shift in the COP was associated with hip adduction and internal rotation, supporting the link between altered foot mechanics

XSL•FO
RenderX

and knee valgus in dynamic movements. Our study further supports these associations, reinforcing the idea that foot pressure analysis can be a valuable screening tool for injury risk assessment. Findings from the DAid smart sock system reinforced the link between medial plantar pressure distribution and hip adduction, which aligns with studies by Sanchis et al [49] and Tran et al [50]. Similarly, hip internal rotation was associated with increased muscle activity in the gluteus maximus and biceps femoris, supporting the role of proximal muscle engagement in dynamic lower limb stability [51]. These findings emphasize the importance of strengthening the hip and hamstring muscles in football players, as improved proximal control may help mitigate excessive knee valgus and reduce injury risk [52]. Further examination of muscle activity patterns reveals the role of neuromuscular control in maintaining lower limb stability. For instance, gluteus maximus and biceps femoris activation correlated with hip internal rotation, which is consistent with previous findings by Wilczynski et al [51] suggesting that proximal muscle control influences knee joint alignment during functional tasks [53]. Increased dynamic knee valgus was associated with higher gluteus maximus activity, reinforcing earlier studies on the importance of hip muscle engagement in knee biomechanics and valgus control [54,55]. This aligns with findings from a systematic review, which highlighted the association between gluteus maximus activation and hip and knee control during single-leg tasks [56], as well as research indicating that higher gluteal activation levels are linked to improved functional performance in tasks requiring lower limb stability and power, such as the 6-meter timed hop test [57]. From a practical perspective, these results indicate that strength and neuromuscular training focusing on the gluteus maximus and hamstrings could be beneficial for reducing valgus-related injuries in youth athletes [58]. While the current study focused on lower limb kinematics and pressure distribution, future research should further explore the dynamic interactions between foot posture, knee mechanics, and proximal muscle function in movement stability. Incorporating real-time feedback using wireless sensor systems during training may enhance movement efficiency and reduce injury risk, an area that warrants further investigation [59,60].

Limitations

This study highlights key limitations that should be addressed in future research. One important aspect is the distinction between muscle activation and muscle force generation. While surface electromyography provides valuable insights into muscle activity, it does not directly quantify neural drive or muscle force [61]. The relationship between these variables is influenced by multiple factors, including electrode positioning, motor unit recruitment, and tissue conductivity, making it difficult to infer strength solely from electromyographic amplitude. Future studies should further investigate neuromuscular control strategies and their implications for lower limb biomechanics, ensuring that clinical applications account for these complexities and uncertainties [61,62]. In addition, this study focused on youth football players aged 14 - 15 years old, a group at higher risk of neuromuscular adaptations and injury [4]. While these findings provide valuable insights into adolescent biomechanics, broader age groups and different sports, such as basketball, should be included in future research to determine the generalizability of the results. Future research could include individuals with lower limb pathologies, such as patellofemoral pain syndrome or postanterior cruciate ligament reconstruction, to provide data on biomechanical parameters and their relationships during functional tasks, aiming to identify risk factors and improve rehabilitation strategies to reduce the risk of progressive lower limb injuries. The research design and methodology effectively captured key biomechanical associations; however, additional experimental conditions—such as fatigue testing or perturbation trials—could provide deeper insights into neuromuscular responses under dynamic conditions.

Comparison With Previous Work

The use of wireless sensor systems in this study allowed for detailed in-field analysis of lower limb biomechanics, but alternative technologies could further refine measurement accuracy and expand analytical capabilities. For instance, the PLUX Wireless Biosignals (muscleBAN kit) was used to assess muscle activation, but the Ultium Wireless Surface Electromyographic system could offer even greater resolution, with its high sampling rates, low noise levels, and SmartLead options, making it a strong alternative for future research [63]. The ability to capture real-time movement data outside laboratory settings provides a significant advantage for applied sports science [19]. These results demonstrate that wearable sensors can provide meaningful biomechanical insights without the constraints of traditional motion capture systems, making them more feasible for training environments [64]. Similarly, the NOTCH inertial sensor system effectively captured lower limb kinematics, but alternative motion capture systems such as OptiTrack could provide more precise 3D movement analysis [65,66]. However, considering the practical challenges of optical motion capture in sports settings [65,66], the use of wearable inertial sensors presents a valuable alternative for continuous monitoring and individualized movement assessment [67].

Gender-Based Biomechanical Differences

Among male participants, hip adduction and hip internal rotation showed a strong correlation, which suggests that males may be more reliant on hip-dominant movement strategies for single-leg stability [56]. This could have implications for training programs, as focusing on hip abductor and external rotator strength may help correct excessive hip adduction patterns that could lead to knee valgus [51,68]. For female participants, the most significant correlation was between knee joint flexion and the overall COP (COP2W). This suggests that female athletes may demonstrate different weight distribution strategies during squatting movements, potentially increasing their vulnerability to ACL injuries [57]. Given that previous research has highlighted greater valgus angles in female athletes, our findings suggest that targeted neuromuscular training, including balance and landing mechanics, may help mitigate these risks [69]. In addition, studies have demonstrated that eccentric hip abductor and lateral rotator strength are significantly associated with lower limb kinematics, particularly in women, reinforcing the importance of proximal control in mitigating excessive knee valgus and femur adduction [69]. Future research should further investigate these gender-based variations in lower limb

XSL•FO
RenderX

biomechanics to better understand potential injury risk factors and inform targeted training interventions for male and female athletes [70]. Coaches and physiotherapists could use this information to create individualized training plans to optimize performance while minimizing injury risk.

Conclusions

This study identified significant correlations between lower limb biomechanics during the "Single Leg Squat" functional test in youth football players. Strong associations were observed between foot COP, lower limb movement patterns, and muscle activity, demonstrating the feasibility of using wireless sensor systems for biomechanical assessment in field conditions. The findings confirm that wireless sensor technology, including the DAid smart sock system, provides reliable data for evaluating lower limb biomechanics in football players.

Acknowledgments

This research is funded by the Latvian Council of Science, project Smart Textile Solutions as Biofeedback Method for Injury Prevention for Latvian Football Youth League Players, project No. lzp-2023/1 - 0027.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author upon reasonable request. The statistical analysis code used in this study is also available from the corresponding author on request. Additionally, the data management plan is registered and can be accessed via Zenodo [71].

Conflicts of Interest

None declared.

References

1. Sadigursky D, Braid JA, De Lira DNL, Machado BAB, Carneiro RJF, Colavolpe PO. The FIFA 11+ injury prevention program for soccer players: a systematic review. BMC Sports Sci Med Rehabil 2017 Dec;9(1):1-8. [doi: 10.1186/s13102-017-0083-z]
2. Mandorino M, J. Figueiredo A, Gjaka M, Tessitore A. Injury incidence and risk factors in youth soccer players: a systematic literature review. Part I: epidemiological analysis. bs 2023;40(1):3-25. [doi: 10.5114/biolsport.2023.109961] [Medline: 36636179]
3. Robles-Palazón FJ, López-Valenciano A, De Ste Croix M, et al. Epidemiology of injuries in male and female youth football players: a systematic review and meta-analysis. J Sport Health Sci 2022 Nov;11(6):681-695. [doi: 10.1016/j.jshs.2021.10.002]
4. Wik EH. Growth, maturation and injuries in high-level youth football (soccer): a mini review. Front Sports Act Living 2022;4:975900. [doi: 10.3389/fspor.2022.975900] [Medline: 36385783]
5. Jones S, Almousa S, Gibb A, et al. Injury incidence, prevalence and severity in high-level male youth football: a systematic review. Sports Med 2019 Dec;49(12):1879-1899. [doi: 10.1007/s40279-019-01169-8]
6. Hewett TE, Myer GD, Ford KR, et al. Biomechanical measures of neuromuscular control and valgus loading of the knee predict anterior cruciate ligament injury risk in female athletes: a prospective study. Am J Sports Med 2005 Apr;33(4):492-501. [doi: 10.1177/0363546504269591]
7. Barengo NC, Meneses-Echávez JF, Ramírez-Vélez R, Cohen DD, Tovar G, Bautista JEC. The impact of the FIFA 11+ training program on injury prevention in football players: a systematic review. Int J Environ Res Public Health 2014 Nov 19;11(11):11986-12000. [doi: 10.3390/ijerph111111986] [Medline: 25415209]
8. Bizzini M, Dvorak J. FIFA 11+: an effective programme to prevent football injuries in various player groups worldwide—a narrative review. Br J Sports Med 2015 May;49(9):577-579. [doi: 10.1136/bjsports-2015-094765]
9. Di Paolo S, Zaffagnini S, Pizza N, Grassi A, Bragonzoni L. Poor motor coordination elicits altered lower limb biomechanics in young football (soccer) players: implications for injury prevention through wearable sensors. Sensors (Basel) 2021;21(13):4371. [doi: 10.3390/s21134371]
10. Kim HS, Yoo HI, Hwang UJ, Kwon OY. Comparison of dynamic knee valgus during single-leg step down between people with and without pronated foot using two-dimensional video analysis. Physical Therapy Korea 2021 Nov 20;28(4):266-272. [doi: 10.12674/ptk.2021.28.4.266]
11. Hribernik M, Umek A, Tomažič S, Kos A. Review of real-time biomechanical feedback systems in sport and rehabilitation. Sensors (Basel) 2022 Apr 14;22(8):3006. [doi: 10.3390/s22083006] [Medline: 35458991]
12. Di Paolo S, Nijmeijer EM, Bragonzoni L, Gokeler A, Benjaminse A. Definition of high-risk motion patterns for female ACL injury based on football-specific field data: a wearable sensors plus data mining approach. Sensors (Basel) 2023 Feb 15;23(4):2176. [doi: 10.3390/s23042176] [Medline: 36850776]
13. Tipton CC, Telfer S, Cherones A, Gee AO, Kweon CY. The use of microsoft kinect for assessing readiness of return to sport and injury risk exercises: a validation study. Int J Sports Phys Ther 2019 Sep;14(5):724-730. [Medline: 31598410]

XSL•FO
RenderX

14. Haimovich Y, Hershkovich O, Portnoy S, Schwartz I, Lotan R. Evaluating lower limb kinematics using microsoft's kinect: a simple, novel method. Physiother Can 2021 Nov 1;73(4):391-400. [doi: 10.3138/ptc-2020-0051] [Medline: 34880546]
15. Longo UG, De Salvatore S, Carnevale A, et al. Optical motion capture systems for 3D kinematic analysis in patients with shoulder disorders. Int J Environ Res Public Health 2022 Sep 23;19(19):12033. [doi: 10.3390/ijerph191912033] [Medline: 36231336]
16. Schmitz RJ, Cone JC, Tritsch AJ, et al. Changes in drop-jump landing biomechanics during prolonged intermittent exercise. Sports Health 2014 Mar;6(2):128-135. [doi: 10.1177/1941738113503286] [Medline: 24587862]
17. Chen B, Liu P, Xiao F, Liu Z, Wang Y. Review of the upright balance assessment based on the force plate. Int J Environ Res Public Health 2021 Mar 8;18(5):2696. [doi: 10.3390/ijerph18052696] [Medline: 33800119]
18. Jackson BE, Evangelista DJ, Ray DD, Hedrick TL. 3D for the people: multi-camera motion capture in the field with consumer-grade cameras and open source software. Biol Open 2016 Sep 15;5(9):1334-1342. [doi: 10.1242/bio.018713] [Medline: 27444791]
19. Suo X, Tang W, Li Z. Motion capture technology in sports scenarios: a survey. Sensors (Basel) 2024 May 6;24(9):2947. [doi: 10.3390/s24092947] [Medline: 38733052]
20. Zamankhanpour M, Sheikhhoseini R, Letafatkar A, Piri H, Asadi Melerdi S, Abdollahi S. The effect of dual-task on jump landing kinematics and kinetics in female athletes with or without dynamic knee valgus. Sci Rep 2023 Aug 31;13(1):14305. [doi: 10.1038/s41598-023-41648-7] [Medline: 37652971]
21. Ahn IK, Gwak GT, Hwang UJ, Yoo HI, Kwon OY. Comparison of foot pressure distribution during single-leg squat in individuals with and without pronated foot. Physical Therapy Korea 2024 Apr 20;31(1):40-47. [doi: 10.12674/ptk.2024.31.1.40]
22. Rodríguez-Martín D, Pérez-López C, Samà A, Cabestany J, Català A. A wearable inertial measurement unit for long-term monitoring in the dependency care area. Sensors (Basel) 2013 Oct 18;13(10):14079-14104. [doi: 10.3390/s131014079] [Medline: 24145917]
23. Khan D, Al Mudawi N, Abdelhaq M, et al. A wearable inertial sensor approach for locomotion and localization recognition on physical activity. Sensors (Basel) 2024;24(3):735. [doi: 10.3390/s24030735]
24. Rodriguez KM, Palmieri-Smith RM, Krishnan C. Quadriceps motor evoked torque is a reliable measure of corticospinal excitability in individuals with anterior cruciate ligament reconstruction. J Electromyogr Kinesiol 2022 Dec;67:102700. [doi: 10.1016/j.jelekin.2022.102700] [Medline: 36063566]
25. Tanaka A, Visi F, Donato BD, Klang M, Zbyszyński M. An end-to-end musical instrument system that translates electromyogram biosignals to synthesized sound. Computer Music Journal 2023 Mar 1;47(1):64-84. [doi: 10.1162/comj_a_00672]
26. Brindle RA, Bleakley CM, Taylor JB, Queen RM, Ford KR. Validity of estimating center of pressure during walking and running with plantar load from a three-sensor wireless insole. Wearable Technol 2022;3:e8. [doi: 10.1017/wtc.2022.5] [Medline: 38486902]
27. Ziagkas E, Loukovitis A, Zekakos DX, Chau TDP, Petrelis A, Grouios G. A novel tool for gait analysis: validation study of the smart insole PODOSmart®. Sensors (Basel) 2021 Sep 6;21(17):5972. [doi: 10.3390/s21175972] [Medline: 34502861]
28. Oks A, Katashev A, Eizentals P, Rozenstoka S, Suna D. Smart socks: new effective method of gait monitoring for systems with limited number of plantar sensors. Health Technol 2020 Jul;10(4):853-860. [doi: 10.1007/s12553-020-00421-w]
29. Januskevica A, Semjonova G, Oks A, Katashev A, Eizentals P. Evaluation of the foot performance in "single leg squat" test of female athletes using smart socks. Presented at: 8th International Conference on Sport Sciences Research and Technology Support; Budapest, Hungary p. 161-168. [doi: 10.5220/0010146701610168]
30. Semjonova G, Davidovica A, Kozlovskis N, Okss A, Katashevs A. Smart textile sock system for athletes' self-correction during functional tasks: formative usability evaluation. Sensors (Basel) 2022 Jun 24;22(13):4779. [doi: 10.3390/s22134779] [Medline: 35808274]
31. Li RT, Kling SR, Salata MJ, Cupp SA, Sheehan J, Voos JE. Wearable performance devices in sports medicine. Sports Health: A Multidisciplinary Approach 2016 Jan;8(1):74-78. [doi: 10.1177/1941738115616917]
32. Liu L, Zhang X. A focused review on the flexible wearable sensors for sports: from kinematics to physiologies. Micromachines (Basel) 2022;13(8):1356. [doi: 10.3390/mi13081356]
33. Masalskyi V, Čičiurėnas D, Dzedzickis A, Prentice U, Braziulis G, Bučinskas V. Synchronization of separate sensors' data transferred through a local Wi-Fi network: a use case of human-gait monitoring. Future Internet 2024;16(2):36. [doi: 10.3390/fi16020036]
34. Claiborne TL, Armstrong CW, Gandhi V, Pincivero DM. Relationship between hip and knee strength and knee valgus during a single leg squat. J Appl Biomech 2006 Feb;22(1):41-50. [doi: 10.1123/jab.22.1.41] [Medline: 16760566]
35. Palmieri-Smith RM, Wojtys EM, Ashton-Miller JA. Association between preparatory muscle activation and peak valgus knee angle. J Electromyogr Kinesiol 2008 Dec;18(6):973-979. [doi: 10.1016/j.jelekin.2007.03.007] [Medline: 17498972]
36. Norcross MF, Almog R, Huang YL, Chang E, Hannigan KS, Johnson ST. Greater explosive quadriceps strength is associated with greater knee flexion at initial contact during landing in females. Sports Biomech 2024 Aug;23(8):1028-1042. [doi: 10.1080/14763141.2021.1908413] [Medline: 33821760]

XSL•FO
RenderX

37. Llurda-Almuzara L, Perez-Bellmunt A, Labata-Lezaun N, et al. Relationship between lower limb EMG activity and knee frontal plane projection angle during a single-legged drop jump. Phys Ther Sport 2021 Nov;52:13-20. [doi: 10.1016/j.ptsp.2021.07.007] [Medline: 34365085]
38. Hu Z, Kim Y, Zhang Y, et al. Correlation of lower limb muscle activity with knee joint kinematics and kinetics during badminton landing tasks. IJERPH 2022;19(24):16587. [doi: 10.3390/ijerph192416587]
39. Bencke J, Aagaard P, Zebis MK. Muscle activation during ACL injury risk movements in young female athletes: a narrative review. Front Physiol 2018;9:445. [doi: 10.3389/fphys.2018.00445] [Medline: 29867521]
40. Fraenkel RJ, Wallen EN. How to Design and Evaluate Research in Education, 7th edition: McGraw-Hill; 2009. URL: https://tinyurl.com/yh3rzymn [accessed 2025-07-03]
41. Dos'Santos T, Bishop C, Thomas C, Comfort P, Jones PA. The effect of limb dominance on change of direction biomechanics: a systematic review of its importance for injury risk. Phys Ther Sport 2019 May;37:179-189. [doi: 10.1016/j.ptsp.2019.04.005] [Medline: 30986764]
42. Hermens HJ, Freriks B, Disselhorst-Klug C, Rau G. Development of recommendations for SEMG sensors and sensor placement procedures. J Electromyogr Kinesiol 2000 Oct;10(5):361-374. [doi: 10.1016/s1050-6411(00)00027-4] [Medline: 11018445]
43. Halaki M, Gi K. Normalization of EMG signals: to normalize or not to normalize and what to normalize to? InTech 2012. [doi: 10.5772/49957]
44. Gignac GE, Szodorai ET. Effect size guidelines for individual differences researchers. Pers Individ Dif 2016 Nov;102:74-78. [doi: 10.1016/j.paid.2016.06.069]
45. Cohen J. Statistical Power Analysis for the Behavioral Sciences: Routledge; 2013. [doi: 10.4324/9780203771587]
46. Weiss K, Whatman C. Biomechanics associated with patellofemoral pain and ACL injuries in sports. Sports Med 2015 Sep;45(9):1325-1337. [doi: 10.1007/s40279-015-0353-4] [Medline: 26130304]
47. Belkhelladi M, Cierson T, Martineau PA. Biomechanical risk factors for increased anterior cruciate ligament loading and injury: a systematic review. Orthop J Sports Med 2025 Feb;13(2):23259671241312681. [doi: 10.1177/23259671241312681] [Medline: 39958696]
48. Mei Q, Gu Y, Xiang L, Baker JS, Fernandez J. Foot pronation contributes to altered lower extremity loading after long distance running. Front Physiol 2019;10:573. [doi: 10.3389/fphys.2019.00573] [Medline: 31191329]
49. Sanchis GJB, Lopes JM, Assis SJCD, Catão RLA, Santos TRT, Roncalli AGDCO. Dynamic knee valgus prevalence in children and its association with pain intensity, foot mobility, and sex- a cross-sectional study. Heliyon 2022 Oct;8(10):e10984. [doi: 10.1016/j.heliyon.2022.e10984] [Medline: 36276716]
50. Tran AA, Gatewood C, Harris AHS, Thompson JA, Dragoo JL. The effect of foot landing position on biomechanical risk factors associated with anterior cruciate ligament injury. J Exp Orthop 2016 Dec;3(1):13. [doi: 10.1186/s40634-016-0049-1] [Medline: 27315816]
51. Wilczyński B, Zorena K, Ślęzak D. Dynamic knee valgus in single-leg movement tasks. potentially modifiable factors and exercise training options. a literature review. Int J Environ Res Public Health 2020 Nov 6;17(21):8208. [doi: 10.3390/ijerph17218208] [Medline: 33172101]
52. Vincent HK, Brownstein M, Vincent KR. Injury prevention, safe training techniques, rehabilitation, and return to sport in trail runners. Arthrosc Sports Med Rehabil 2022 Jan;4(1):e151-e162. [doi: 10.1016/j.asmr.2021.09.032] [Medline: 35141547]
53. Rabello R, Brunetti C, Bertozzi F, Rodrigues R, Sforza C. Different neuromuscular parameters are associated with knee abduction and hip adduction angles during functional tasks. J Electromyogr Kinesiol 2023 Oct 14;73:102833. [doi: 10.1016/j.jelekin.2023.102833] [Medline: 39492375]
54. Rinaldi VG, Prill R, Jahnke S, Zaffagnini S, Becker R. The influence of gluteal muscle strength deficits on dynamic knee valgus: a scoping review. J Exp Orthop 2022 Aug 17;9(1):81. [doi: 10.1186/s40634-022-00513-8] [Medline: 35976534]
55. Alzahrani AM, Alzhrani M, Alshahrani SN, Alghamdi W, Alqahtani M, Alzahrani H. Is hip muscle strength associated with dynamic knee valgus in a healthy adult population? a systematic review. Int J Environ Res Public Health 2021 Jul 19;18(14):7669. [doi: 10.3390/ijerph18147669] [Medline: 34300118]
56. Nascimento MB, Vilarinho LG, Lobato DFM, Dionisio VC. Role of gluteus maximus and medius activation in the lower limb biomechanical control during functional single-leg Tasks: A systematic review. Knee 2023 Aug;43:163-175. [doi: 10.1016/j.knee.2023.05.005] [Medline: 37422984]
57. Souza MSA, Valadao VMA, Teruyu AI, et al. Can gluteal muscle activation discriminate functional performance in moderately trained women? - A cross-sectional study. Sports Biomech 2024 Dec;23(12):3602-3614. [doi: 10.1080/14763141.2023.2228281] [Medline: 37370023]
58. Ramezani F, Saki F, Tahayori B. Neuromuscular training improves muscle co‐activation and knee kinematics in female athletes with high risk of anterior cruciate ligament injury. Eur J Sport Sci 2024 Jan;24(1):56-65. [doi: 10.1002/ejsc.12046]
59. Van Hooren B, Plasqui G, Meijer K. The effect of wearable-based real-time feedback on running injuries and running performance: a randomized controlled trial. Am J Sports Med 2024 Mar;52(3):750-765. [doi: 10.1177/03635465231222464] [Medline: 38287728]

XSL•FO
RenderX

60. Davidovica A, Semjonova G, Kamynina L, et al. Concept protocol for developing a DAID® smart socks-based biofeedback system: enhancing injury prevention in football through real-time biomechanical monitoring and mixed reality feedback. Appl Sci (Basel) 2025;15(3):1584. [doi: 10.3390/app15031584]
61. Farina D. Interpretation of the surface electromyogram in dynamic contractions. Exerc Sport Sci Rev 2006 Jul;34(3):121-127. [doi: 10.1249/00003677-200607000-00006] [Medline: 16829739]
62. Martinez-Valdes E, Negro F, Falla D, De Nunzio AM, Farina D. Surface electromyographic amplitude does not identify differences in neural drive to synergistic muscles. J Appl Physiol (1985) 2018 Apr 1;124(4):1071-1079. [doi: 10.1152/japplphysiol.01115.2017] [Medline: 29420155]
63. Wu YD, Ruan SJ, Lee YH. An ultra-low power surface EMG sensor for wearable biometric and medical applications. Biosensors (Basel) 2021 Oct 21;11(11):411. [doi: 10.3390/bios11110411] [Medline: 34821627]
64. Hafer JF, Vitali R, Gurchiek R, Curtze C, Shull P, Cain SM. Challenges and advances in the use of wearable sensors for lower extremity biomechanics. J Biomech 2023 Aug;157:111714. [doi: 10.1016/j.jbiomech.2023.111714] [Medline: 37423120]
65. Leardini A, Sawacha Z, Paolini G, Ingrosso S, Nativo R, Benedetti MG. A new anatomically based protocol for gait analysis in children. Gait Posture 2007 Oct;26(4):560-571. [doi: 10.1016/j.gaitpost.2006.12.018] [Medline: 17291764]
66. Aurand AM, Dufour JS, Marras WS. Accuracy map of an optical motion capture system with 42 or 21 cameras in a large measurement volume. J Biomech 2017 Jun 14;58:237-240. [doi: 10.1016/j.jbiomech.2017.05.006] [Medline: 28549599]
67. Prisco G, Pirozzi MA, Santone A, et al. Validity of wearable inertial sensors for GAIT analysis: a systematic review. Diagnostics (Basel) 2024 Dec 27;15(1):36. [doi: 10.3390/diagnostics15010036] [Medline: 39795564]
68. Mozafaripour E, Seidi F, Minoonejad H, Bayattork M, Khoshroo F. The effectiveness of the comprehensive corrective exercise program on kinematics and strength of lower extremities in males with dynamic knee valgus: a parallel-group randomized wait-list controlled trial. BMC Musculoskelet Disord 2022 Jul 22;23(1):700. [doi: 10.1186/s12891-022-05652-8] [Medline: 35869467]
69. Baldon RDM, Lobato DFM, Carvalho LP, Santiago PRP, Benze BG, Serrão FV. Relationship between eccentric hip torque and lower-limb kinematics: gender differences. J Appl Biomech 2011;27(3):223-232. [doi: 10.1123/jab.27.3.223]
70. Lin CY, Casey E, Herman DC, Katz N, Tenforde AS. Sex differences in common sports injuries. PM R 2018 Oct;10(10):1073-1082. [doi: 10.1016/j.pmrj.2018.03.008] [Medline: 29550413]
71. Kreišmanis I. Project: "Smart textile solutions as biofeedback method for injury prevention for Latvian football youth league players (Nr. lzp-2023/1-0027)". Zenodo. 2024 Jul 24. URL: https://doi.org/10.5281/zenodo.12804841 [accessed 2025-07-07]

Abbreviations

ACL: anterior cruciate ligament
BF: biceps femoris
COP: center of pressure
FIFA: Fédération Internationale de Football Association
GM: gluteus medius
GMx: gluteus maximus
MVC: maximum voluntary contraction
SENIAM: Surface Electromyography for the Non-Invasive Assessment of Muscles
SLS: Single Leg Squat
VL: vastus lateralis

Edited by A Mavragani; submitted 20.11.24; peer-reviewed by DFM Lobato, M Hatefi; revised version received 11.04.25; accepted 13.04.25; published 10.07.25.

Please cite as:
Davidoviča A, Davidovičs S, Semjonova G, Katashev A, Oks A, Lancere L, Tomsone S, Zolovs M
Correlation of Biomechanical Variables of Lower Extremity Movement During Functional Tests and Tasks in Youth League Football Players: Cross-Sectional Correlation Study
JMIR Form Res 2025;9:e69046
URL: https://formative.jmir.org/2025/1/e69046
doi:10.2196/69046

© Anna Davidoviča, Sergejs Davidovičs, Guna Semjonova, Alexei Katashev, Alexander Oks, Linda Lancere, Signe Tomsone, Maksims Zolovs. Originally published in JMIR Formative Research (https://formative.jmir.org), 10.7.2025. This is an open-access

XSL•FO
RenderX

article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

A Web-Based Well-Being and Resilience Intervention for Family Members and Friends Supporting a Loved One Using Alcohol and Other Drugs: Mixed Methods Pilot Study

Steph Kershaw[1], PhD; Jessica Deng[1], MClinPsych; Madeleine Keaveny[1], MClinPsych; Bronte Speirs[2,3], MClinPsych; Anna Grager[1], BHlthSc; Dara Sampson[2,3], PhD; Kate Ross[1], MA; Nicola Newton[1], PhD; Maree Teeson[1], PhD; Frances Kay-Lambkin[2,3], PhD; Cath Chapman[1], PhD

[1]The Matilda Centre for Research in Mental Health and Substance Use, University of Sydney, 160 City Road, Level 6 G02 Jane Foss Russell Building, Sydney, Australia

[2]The University of Newcastle, Newcastle, New South Wales, Australia

[3]Hunter Medical Research Institute, Newcastle, New South Wales, Australia

Corresponding Author:
Steph Kershaw, PhD
The Matilda Centre for Research in Mental Health and Substance Use, University of Sydney, 160 City Road, Level 6 G02 Jane Foss Russell Building, Sydney, Australia

Abstract

Background: Despite the known psychosocial challenges associated with supporting a loved one using alcohol and other drugs (AOD), there is a scarcity of mental health and well-being interventions for affected friends and family members (AFFMs). Stigma has also been shown to discourage help-seeking among AFFMs. Web-based interventions may facilitate help-seeking by ensuring privacy and anonymity.

Objective: This pilot study examines the usability, acceptability, and feasibility of the Family and Friend Support Program (FFSP), a world-first, evidence-based web-based resilience and well-being program designed with, and for, people caring for someone using AOD. This study also examined AFFM's experiences of caring for a loved one using AOD and their help-seeking behaviors and barriers.

Methods: In 2021 (November-December), participants across Australia completed a baseline web-based cross-sectional survey that assessed the impact of caring for a loved one using AOD (adapted Short Questionnaire for Family Members-Affected by Addiction), and distress levels (Kessler Psychological Distress Scale [K-10]). Following baseline, participants were invited to interact with the FFSP over 10 weeks. Postprogram and follow-up surveys (10 and 14 wk postbaseline, respectively) and semistructured interviews assessed the usability and acceptability of the program, as well as help-seeking experiences and barriers.

Results: Baseline surveys were completed by 131 AFFMs, with 37% (n=49) completing the postprogram survey and 24% (n=32) completing the follow-up survey. A total of 5 participants took part in individual semistructured interviews at postprogram. On average, K-10 scores fell in the moderate to severe range at baseline (mean 28.4, SD 8.6). At postprogram, the majority of participants (n=27, 55.1%) reported that they did not seek help to cope with or manage their role supporting their loved one and the most common endorsed barrier was cost (n=11, 28.6%). Overall, participants found the FFSP easy to use and provided them with relevant, helpful, and validating information. The majority (n=35, 71.5%) of participants said they would be likely to recommend the FFSP to a person supporting a loved one using AOD. Qualitative responses highlighted the need for free, accessible support for AFFMs such as the FFSP. Limitations included low program engagement and high attrition.

Conclusions: Overall, the FFSP appears to be a promising mental health intervention for AFFMs. This study builds on existing research finding high levels of distress among AFFMs, while highlighting the ongoing barriers to help-seeking. Limitations and future directions for refinements and future efficacy evaluation of the FFSP are discussed including ways to address attrition and increase engagement.

(JMIR Form Res 2025;9:e72425) doi:10.2196/72425

KEYWORDS

affected friends and family members; substance use; help-seeking; web-based intervention; pilot trial; mobile phone

XSL•FO
RenderX

Introduction

The use of alcohol and other drugs (AOD) not only affects the person using the substance but also impacts their family, friends, and community, often in adverse ways [1,2]. The 2022-2023 National Drug Strategy Household Survey found that around 1 in 3 Australians (31% or 6.6 million) reported drinking alcohol above guideline recommendations [3], and around 1 in 5 (18% or 3.9 million) people had used an illicit drug in the previous 12 months [4]. Additionally, the Australian Burden of Disease Study reported that AOD use was estimated to be responsible for 7.5% of the total burden of disease and injury [5]. Considering the high rates of AOD use and associated burden in Australia, the number of family members and friends who are supporting someone using AOD, or are affected by someone's use, is likely to be high [6].

The term "affected friends and family members" (AFFMs), used in this paper, represents a heterogeneous group that includes parents, children, significant others, relatives, friends, and caregivers of someone using AOD. AFFMs often have the responsibility of advocating and caring for their loved one, while also managing their loved one's changes in behavior, increased relationship strain and conflict, and financial insecurity [1,7-9]. Taking on this support role in these circumstances can be difficult and unpredictable and it has been found that AFFMs experience a substantially poorer quality of life compared to the general population, including poorer physical and mental health [10-13]. AFFMs may be at higher risk of developing AOD use, depression, and anxiety disorders themselves [6,14]. As AOD use and dependence can be a lifelong issue, there is a need to support AFFMs to develop effective long-term coping strategies and access support for their own needs and well-being [15].

AFFMs also face significant stigma and social isolation, which can further exacerbate stress and their ability to cope and seek support [16,17]. As well as public and self-stigma associated with substance use, AFFMs may be subject to stigma by association or "courtesy stigma" which involves public disapproval and negative social interactions invoked by narratives of blame, shame, and contamination around those supporting a loved one using AOD [18,19]. Both public stigma and stigma by association have been found to be a significant barrier to help-seeking. AFFMs may be deterred from seeking help due to shame, fear of judgment, and concerns around privacy for both themselves and their loved ones [20]. However, help-seeking, both informally and formally, has been identified as an important strategy for helping AFFMs manage and cope adaptively with stressors and reduce levels of impact and burden [7,21,22]. Thus, addressing barriers of stigma and shame associated with seeking help is critical when developing and promoting effective forms of support and interventions for AFFMs.

Previous research suggests that the experiences and impact of supporting a loved one may depend on characteristics such as the AFFM's gender and relationship with their loved one, as well as on sociocultural differences [23-25]. For example, female partners, parents, and those who live in the same household as their loved one are more likely to experience a greater cumulative burden compared to other family members and friends [26,27]. Differences have also been found based on country and geographical location, suggesting cultural variations in impact and coping among AFFMs [23,24,28,29]. Therefore, it is important to consider potential differences in impact and coping based on demographic factors to appropriately capture the different needs of this diverse population.

Research now conceptualizes AFFMs as an independent population who are uniquely impacted by the significant uncertainty, stress, and strain that may arise around a loved one's AOD use. Historically, AFFMs have been seen as "part of the problem" or only as an adjunct to treatment of the individual affected by AOD use rather than as help-seekers in and of themselves [16,30,31]. In response to this, Orford et al [32] developed the Stress-Strain-Coping-Support Model which is a nonpathological model that recognizes the chronic stress of living with or supporting an individual experiencing drug dependence and the way this stress leads to strain which is mediated by their coping style and quality of social support. The 5-step method puts this model into practice with an emphasis on empowering family members, reducing distress, providing information, and enhancing coping and support [33]. This model has been shown to be effective in reducing the strain experienced by family members with some positive knock-on effects for the person they are supporting [15]. This finding is consistent with previous research suggesting that family members and family function play a key role in preventing and minimizing the risks and harms associated with a loved one's AOD use, as well as promoting resilience and more positive and longer-lasting treatment outcomes [30,34-36]. Specifically, family-focused approaches to intervention and recovery improve family functioning, reduce relapse, and help both affected family and their loved ones improve their quality of life [35]. This signifies the importance of fostering help-seeking among AFFMs, and the need for effective, targeted interventions for this population.

To date, there are few interventions available to AFFMs that do not require the involvement of the person using AOD, and the interventions that do exist are rarely evaluated [7]. Two recent web-based group programs developed specifically for AFFMs have shown promising results in terms of self-reported outcomes and feasibility [37,38]. The peer-led web-based support program by Peart et al [38] found significant improvements in self-efficacy and overall satisfaction with the program, and Rushton et al [37] found that the SMART Family and Friends, a mutual-support group targeting families delivered via videoconferencing, was associated with significant improvements in psychological distress, family impact, strain symptoms, and total family burden. These programs demonstrate the need for support and interventions specifically developed to address the unique challenges faced by AFFMs, such as web-based groups and peer-based support.

While web-based peer support groups offer a valuable avenue for accessing help, it is important to consider the ongoing barriers to help-seeking including stigma, gaps in help-seeking knowledge, concern for their own and their loved one's privacy, and practical concerns (eg, cost, location, and time) which can

XSL•FO
RenderX

deter AFFMs from accessing support even when it exists [17,39,40]. Additionally, AFFMs often face multiple competing stressors on top of and related to caring for a loved one using AOD and this can lead to carer burnout, introducing an additional barrier to participating in structured, group-based interventions [41,42]. These barriers and the isolating impacts of stigma highlight the need for interventions to be privately accessible anywhere at any time for AFFMs.

To address this gap, the Family and Friend Support Program (FFSP) [43] was developed in 2019 [44]. The FFSP is a world-first, evidence-based, web-based intervention designed with, and for, family members and friends supporting loved ones using AOD. The development of the FFSP was based on the Stress-Strain-Coping-Support Model and the 5-step method [15] and included consultation and collaboration with AFFMs, capturing their concerns and needs through web-based surveys and interviews. The program provides tailored support for AFFMs that is free, confidential, and accessible digitally anywhere at any time across Australia. This allows vulnerable and burdened AFFMs to access support without anxieties around compromising their own or their loved one's privacy. It also ensures that at-risk groups and marginalized communities have access to the program including women (who often bear the greater caregiver burden), low socioeconomic communities (who experience greater financial stress), and those living in rural, regional, or remote areas with limited resources [1,39,45]. In this way, the FFSP is positioned as a low-barrier early-intervention option for AFFMs learning to manage their role as a carer and advocate for their loved ones using AOD. The FFSP features a package of modules informed by principles of cognitive behavioral therapy including psychoeducation, personalized activities, and information on other web-based resources and services [44]. With permission, the FFSP also included excerpts from initial AFFM consultations and interviews conducted during the development of FSSP as quotes and vignettes throughout the program to highlight lived experience perspectives (see Multimedia Appendix 1 for program screenshots).

This pilot study aimed to assess the usability, acceptability, and feasibility of FFSP. A secondary aim was to capture a cross-sectional snapshot of a diverse sample of AFFM's experiences of caring for a loved one using AOD.

Methods

Study Design

The study involved web-based surveys completed at three timepoints: (1) baseline, (2) postprogram (10 wk postbaseline), and (3) one-month follow-up (14 wk postbaseline; see Multimedia Appendix 2 for study flow). The baseline survey assessed demographic details about participants and their loved ones, psychological distress, and family burden including impact, stress, coping, and social support. Following baseline surveys, participants were invited to interact with the FFSP [43] and complete 11 evidence-informed modules, including 4 core modules and 7 mini modules, over a 10-week period. The postprogram survey was offered at 10 weeks postbaseline to all AFFMs who completed baseline, regardless of whether they had accessed all, part, or none of the FFSP. The same sample was offered a follow-up survey 4 weeks later. The postprogram and follow-up surveys assessed previous help-seeking experiences and barriers, program usability and feedback, and repeated baseline measures of psychological distress and family burden. Participants were offered the opportunity to volunteer for an in-depth phone interview at both postprogram and follow-up timepoints.

Participants and Recruitment

Between November and December 2021, Australian residents from the general community aged 18 years or older were recruited for the web-based survey via social media (Facebook, Twitter) along with e-newsletters. Recruitment advertisements targeted AFFMS (eg, "Have you been affected by someone else's use of drugs or alcohol? Researchers at the University of Newcastle and the University of Sydney have developed a web-based program and information website to help families or friends"). All 3 surveys were administered via REDCap (Research Electronic Data Capture; Vanderbilt University) [46,47], a secure, web-based data capture tool. Semistructured phone interviews were conducted via Zoom (Zoom Video Communications Inc) or telephone depending on the participants' preference with participants confirming at the start of the interview that they were in a safe, private space (eg, a room in their own house). Inclusion criteria were being an Australian resident, older than 18 years of age, and having a close family member or friend whose AOD use was causing them concern. Since this was a pilot study, a sample size calculation was not performed. The researchers aimed for 120 participants as it was believed this would be a large enough sample to gain insight into the usability, acceptability, and feasibility of the web-based program accounting for a moderate level of attrition.

Measures

Demographics

Demographic information included age (in years), gender ("How do you identify?"), residential postcode, geographic region (metropolitan, regional, or rural or remote), and cultural heritage [48]. Participants were also asked how they typically accessed the internet (smartphone, PC, or tablet), and questions about their family member or friend who uses AOD including demographics, the substance of concern, length of use, and the participant's relationship to that person (eg, partner, friend).

Psychological Distress

Psychological distress was measured using the Kessler Psychological Distress Scale (K-10) [49]. The K-10 is a well-validated and widely recommended simple screening measure of psychological distress [50]. Participants responded to 10 items about their emotional states in the past 4 weeks on a 5-point Likert scale (where 1=none of the time and 5=all of the time). Scores were then summed with the maximum score of 50 indicating severe distress, and the minimum score of 10 indicating no distress.

XSL•FO
RenderX

Impact, Stress, Coping, and Social Support

Stress, coping, and social support were measured via the adapted Short Questionnaire for Family Members-Affected by Addiction (SQFM-AA) [51]. The SQFM-AA has shown satisfactory to good internal reliability and validity [27]. The SQFM-AA consists of 33 questions about a loved one's AOD-related behaviors in the preceding 3 months and participants rated how often they have engaged in certain responses using a 4-point Likert scale (1=never, 2=once or twice, 3=sometimes, and 4=often). The SQFM-AA includes subscales measuring impact (worry and disturbance), symptoms (physical and psychological), coping (engaged emotional, engaged assertive, tolerant, and withdrawal), social support (helpful formal, helpful informal, and unhelpful informal), and total family burden. Total family burden scores include the cumulative score of impact, symptoms, and coping with scores ranging from 0 to 48 and higher scores indicating a higher degree of burden experienced by the family member.

Help-Seeking Experiences

At postprogram and follow-up, participants were also asked about their previous help-seeking experiences: "Over the past month, did you seek help to help cope with or manage your role supporting your loved one (eg, mental health support)? Yes or No" and if they answered "Yes," they were prompted to select all that applied to them from a list of services including "Online," "GP," "Counsellor or psychologist," and "Friend or family member." Participants were then asked, "Did you receive the help that you needed?" (Yes or No), followed by "Please comment on the help you did or did not receive" (free textbox). All participants were asked about barriers to help-seeking. Barriers to help-seeking were assessed using the barriers to help-seeking scale, adapted from the widely used, reliable, and valid Perceived Need for Care Questionnaire administered in the Australian National Survey of Mental Health and Well-being [52,53]. Participants were asked "Were there any barriers that have stopped you from seeking help (Select all that apply)?" and then asked to select from 11 predefined statements. Additionally, there was an option to select "Other" followed by the prompt "Please specify other barriers that have stopped you from seeking help?" and a free textbox.

Usability, Acceptability, and Perceived Usefulness of the Program

Program usability was assessed via the System Usability Scale (SUS) [54]. The SUS reliably measures usability and consists of 10 items including several facets such as ease of website use and website complexity. These 10 items were scored on a 5-point Likert scale ranging from 0=strongly disagree to 5=strongly agree. Higher scores indicate better usability and range from 0 to 100. Scores of 80 or above indicate a strong performance, with the average SUS score being 68 as based on more than 500 studies [54].

Additionally, participants answered the Likert scale (eg, "How easy was it to find the information you wanted? (Select an option: Very easy, moderately easy, somewhat easy, not very easy, not easy at all, unsure)") and open-ended questions (eg, "Was there anything missing from FFSP that you expected or wanted to be included?") about the acceptability and usefulness of the FFSP in terms of structure and content. Participants also answered open-ended questions about perceived barriers to accessing the program ("What barriers stopped you from accessing FFSP (specify)?"), likelihood of recommending to someone else ("On a scale of 1 (would not recommend at all) to 10 (extremely likely to recommend), how likely would you be to recommend FFSP to a person supporting a loved one using alcohol or other drugs?"), and feedback on ways to improve the program ("Was there anything missing from FFSP that you expected or wanted to be included? If yes, then, what was missing (specify)").

In-Depth Phone Interviews

Further, optional in-depth semistructured phone interviews were conducted by a clinical psychologist or social worker experienced in interviewing. Confidentiality was explained to the interviewee before obtaining consent to audio record the interview which was later transcribed by a member of the research team.

Data Analysis

Data were analyzed using the SPSS Statistics (version 25; IBM Corp). Independent samples 2-tailed *t* tests and one-way ANOVAs were conducted to assess differences in baseline K-10 and SQFM-AA Total Family Burden scores based on demographic factors (age, gender, geographical location, and relationship to loved one using AOD). To assess attrition, independent samples 2-tailed *t* tests were conducted to compare participants who completed only baseline versus both baseline and postprogram, as well as participants who accessed versus did not access the program at postprogram. Although analysis of changes in outcomes across timepoints was intended, it was not conducted due to the small sample size (n=11) of those who reported accessing the program and completed all 3 timepoints. This paper therefore reports findings from the quantitative data collected at baseline, as well as the qualitative data from postprogram and follow up.

Ethical Considerations

Ethics approval was obtained from the University of Newcastle (H-2017‑0040) Human Research Ethics Committee. All participants provided informed consent. All data collected from participants were anonymous and nonidentifiable via the secure REDCap survey platform. Participants who completed the postprogram and follow-up surveys were reimbursed with AUD $50 (US $32) grocery digital gift cards for each survey. Participants who completed the interview were reimbursed with an AUD $25 (US $16) grocery digital gift card.

Results

Demographics and Sample Characteristics

A total of 131 participants completed baseline measures (refer to Table 1 for characteristics). Mean age was 50.9 (SD 12.7) years and 80.2% (n=105) identified as female. Just over half of the participants were from metropolitan areas (n=68, 51.9%) with the majority born in Australia (n=107, 81.7%). The most common relationship of the loved one was a child (n=50,

XSL•FO
RenderX

38.2%), and the most reported substance of concern was ice or crystal methamphetamine (n=64, 48.9%).

A total of 49 (37%) participants went on to complete the postprogram survey and 32 (24%) participants completed the follow-up survey. At the postprogram timepoint, 17 (13% of the baseline sample) participants reported accessing the FFSP, while at the follow-up survey, 14 (11% of the baseline sample) participants had accessed the FFSP since postprogram. A total of 5 (4%) participants took part in individual semistructured interviews. Figure 1 shows the study flowchart.

Table . Demographic characteristics of the affected friends and family members sample at baseline survey.

Characteristics	Value (n=131)
Age (years), mean (SD)	50.9 (12.7)
Gender, n (%)	
Woman	105 (80.15)
Man	25 (19.1)
Nonbinary	1 (0.8)
Country of birth, n (%)	
Australia	107 (81.7)
Other[a]	24 (18.3)
Geographical area, n (%)	
Metropolitan	68 (51.9)
Regional	42 (32.1)
Rural or remote	21 (16)
Loved one using AOD[b], n (%)	
Child[c]	50 (38.2)
Partner[d]	29 (22.1)
Sibling	11 (8.4)
Friend	10 (7.6)
Parent	10 (7.6)
Other[e]	21 (16)
AOD of concern, n (%)	
Crystal methamphetamine	64 (48.9)
Alcohol	40 (30.5)
Cannabis	12 (9.2)
Heroin	4 (3.1)
Prescription opiates	4 (3.1)
Other	7 (5.4)

[a]Other includes Afghanistan, Canada, Indonesia, Iran, Malaysia, Netherlands, New Zealand, Scotland, Singapore, South Africa, United Kingdom, and the United States.

[b]AOD: alcohol and other drugs.

[c]Child includes son, daughter, step-son, and step-daughter.

[d]Partner includes husband, wife, spouse, and partner.

[e]Other includes ex-partner, sibling-in-law, nephew, roommate, and neighbor.

XSL•FO
RenderX

Figure 1. Flowchart of the study design for affected friends and family members (n=131 at baseline), which includes the sample sizes at each survey timepoint (baseline, postprogram, and follow-up), whether they accessed the FFSP, and participation in optional in-depth interviews. FFSP: Family and Friend Support Program.

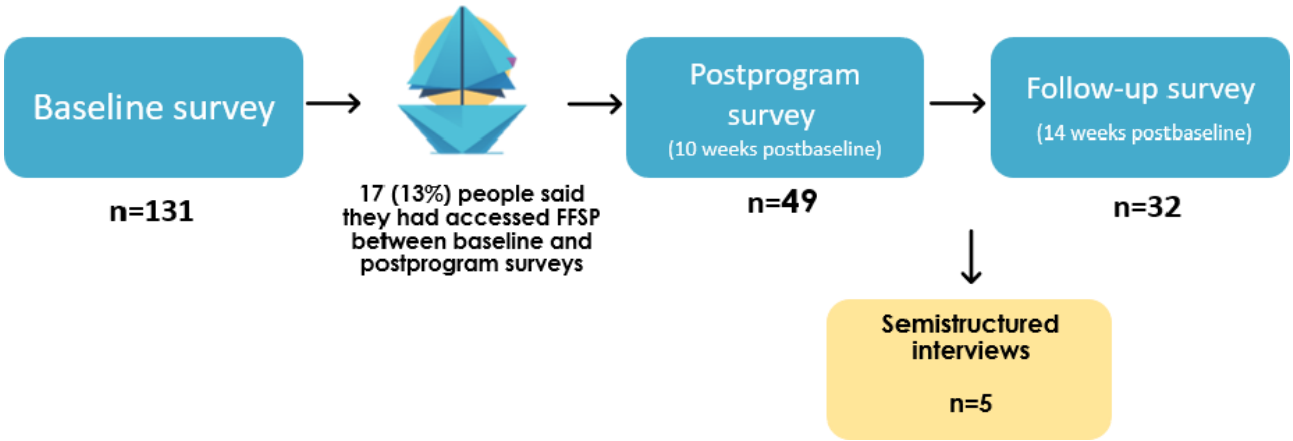

Analysis of Dropout and Attrition

As attrition (62.6% dropout) between baseline and postprogram timepoint was high, analysis was undertaken to explore if there were any demographic differences between those who completed only the baseline survey and those who completed both the baseline and postprogram survey. Age, gender, geographical location, and relationship to the affected loved one were not significantly associated with dropout from baseline to postprogram. There were also no significant differences in basic demographic factors between those who did and did not access the program.

Psychological Distress at Baseline

The mean K-10 score at baseline among the total sample was 28.4 (SD 8.6), which falls in the moderate to severe range of psychological distress (Multimedia Appendix 3). There were no significant differences in K-10 scores based on age, gender, relationship to affected loved one, or geographical location. However, participants who accessed the FFSP had significantly higher baseline K-10 scores (mean 31.6, SD 10.3) compared to those who did not access the FFSP (mean 24.2, SD 7.6; t_{47}=2.86; P=.006).

Stress, Coping, and Social Support

Stress, Strain, and Total Family Burden

The mean total family burden score at baseline was 44.96 (SD 9.46), indicating a high degree of burden experienced by participants associated with their loved one's AOD use. Total family burden did not differ significantly based on demographics such as age, gender, relationship to loved one, or geographical location.

In terms of the impact of their loved one's AOD use over the previous 3 months, the majority of participants reported that their loved one's AOD use at least sometimes affected the family's finances (n=89, 68%), got in the way of their social life (n=99, 75.6%), led to worry about their loved one's neglect of self-care (n=104, 79.4%), led to fights or arguments (n=96, 73.3%), or upset family or social occasions (n=81, 61.9%). More than half reported that their loved one has threatened them at least once or twice (n=71, 54.2%). The majority (n=93, 71%) of participants reported often worrying about their loved one's AOD use. Almost half the participants reported sometimes having difficulty concentrating (n=77, 48.9%) and feeling weak in parts of their body (n=64, 48.9%).

Coping Styles

One-third (n=39, 29.8%) of participants endorsed often watching their loved one's every move, 15% (n=19) had often started an argument with their loved ones about their AOD use, and 30% (n=37) often got moody and emotional with their loved ones. One-third of participants endorsed often talking frankly about their AOD use and making clear to their loved ones their expectations of how they should contribute to the family. Over a third (n=45, 35%) endorsed often putting themselves out for their loved ones, for example, getting them to bed or clearing up the mess after they had been drinking or taking drugs.

Social Support

A majority of participants (n=88, 64.1%) endorsed that friends or relations sometimes or often listened to them when they talked about their feelings, that their friends or relations sometimes or often have been there for them (n=83, 63.3%), and that friends or relations have sometimes or often talked to them about their loved one and listened to what they had to say (n=82, 62.6%). Over half the participants endorsed that they never had health or social care workers give them helpful information about problem drinking or drug taking (n=76, 58%), had health or social care workers make themselves available to them (n=77, 58.8%), or confided in their health or social care worker about their situation (n=70, 53.4%; see Multimedia Appendix 4 for a full breakdown of SQFM-AA scores).

Help-Seeking Experiences

Among the participants who completed the postprogram survey (n=49), the majority (n=27, 55.1%) reported that they did not seek help to cope with or manage their role in supporting their loved ones. Of the 22 (44.9%) participants who reported that they sought help, the most common form of support accessed was a counselor or psychologist (n=13, 26.5%). Participants provided a mix of positive and negative comments on their previous help-seeking experiences. Examples of positive comments include, "My psychologist has been a great listener and has been very understanding and supportive" and "NA

XSL•FO
RenderX

(Narcotics Anonymous) has been useful in that we now have a small group of people who understand our situation and are supportive. Otherwise, NA and 12 steps in particular don't feel relevant or useful to me." Examples of negative comments include, "(health workers') Judgment or jumping to conclusions" and "too long of waitlists." These sentiments were also reflected in interviews, for example, one interviewee highlighted that stigma is an ongoing barrier to asking for help.

> *One barrier that I think I could identify there is that it's very common, in my experience, that people, family and friends of people who've got a substance abuse problem ... feel almost embarrassed to admit that this is something that's completely out of their control, and participating in programs can often be a very challenging thing for them to do, because they fear that they're going to be held accountable.*

Barriers to help-seeking are summarized in Table 2.

Table . Barriers to help-seeking endorsed by affected friends and family members at postprogram survey completed 10 weeks postbaseline[a].

Barrier to help-seeking	Value (n=49), n (%)
I couldn't afford the money	14 (28.6)
I didn't need help	11 (22.4)
I preferred to manage myself	11 (22.4)
I didn't think anything could help	9 (18.4)
I was afraid of legal implications	9 (18.4)
I previously asked but didn't get help	9 (18.4)
I couldn't afford the time off work	8 (16.3)
I previously asked for help but had a negative or bad experience	6 (12.2)
I was afraid of what others would think of me	4 (8.2)
I didn't know where to seek help	4 (8.2)
I was afraid to seek help	2 (4.1)
Other[b] (service unhelpful, not needing support, general stigma, other support)	6 (12.2)

[a]Note that participants were able to select multiple responses.

[b]Free textbox.

Usability, Acceptability, and Perceived Usefulness of the Program

The average SUS score for the FFSP was 70 (SD 18; n=17), which is slightly above the average score of 68 from 500 studies [54]. Among the participants who completed the postprogram survey, the majority (n=36, 65.3%) of participants reported that it was easy to find the information that they wanted, and most (n=35, 71.4%) reported that the information was easy to understand. Further, over half (n=26, 53.1%) felt that the activities were somewhat, moderately, or very helpful. After interacting with the FFSP, 1 in 3 (n=17, 34.7%) felt very or moderately confident when faced with AOD-related issues. The majority (n=35, 71.5%) of participants rated that they would be likely to recommend the FFSP to a person supporting a loved one using alcohol or other drugs.

All interviewees in the semistructured interviews (n=5) identified the need for programs such as the FFSP to address the lack of accessible support for AFFMs, to validate and normalize AFFMs experiences, and to combat misinformation and stigma surrounding AOD use. For example, one interviewee found the vignettes of lived experience stories validating.

> *It's extremely comforting... to find that there is somebody else who is just like you.*

Another interviewee spoke about the important role of web-based support in overcoming shame.

> *There is a big role for online support to play because it helps people not feel so bad because I think ... half the problem is, people don't tell anyone because it's so shameful, particularly once they're an adult.*

Another interviewee highlighted the need for web-based, 24/7 access particularly when living in a small town.

> *Being in a smaller town where everyone seems to know everyone's business ... makes it a bit harder to front up to those sort of things [ask for help]. So when you can do stuff online it can be obviously done in your own time.*

Some interviewees described the length and amount of content as a barrier to completion. For example, one participant found that "FFSP felt quite complex and needed considerable time commitment" and another participant noted that "there was just so much information, it required considerable time to work through." In terms of ways to improve the content of the FFSP, participants suggested including more practical strategies and skills for supporting their loved ones, more content focused on lived experiences, and follow-up with in-person community groups and resources. For example, when asked what they would like to include, one interviewee said, "more stories of what

XSL•FO
RenderX

being a friend or family support person might involve." Another interviewee wanted more content addressing stigma and shame in the carer role.

> *I think it would be very helpful if you could give to people going through the program, tips to enable them (a) to understand this is not your fault, (b) to come to the realization that you don't have to be angry at the person who's in substance abuse.*

Finally, another interviewee said that they would like the web-based format to be complemented by in-person services and resources.

> *It needs direct references into live communities of belonging, not just web resources.*

Reasons for not accessing the program between the survey timepoints included time restrictions, other stressors (eg, their own sickness, work stress, and carer burnout), and perceived lack of need (eg, receiving other formal or informal support and loved ones doing well or in recovery). Several participants also noted technical difficulties (eg, links not working and forgetting passwords).

Discussion

Principal Findings

This pilot study investigated the usability, acceptability, and feasibility of the FFSP, a world-first, evidence-based web-based intervention developed to support the well-being of family members and friends with a loved one using AOD.

Overall, the FFSP demonstrated good usability and acceptability among AFFMs and was found to be useful in validating users' experiences and addressing stigma around seeking help. Most participants reported that the FFSP was easy to use, and it was easy to find and understand the content. Participants reported that the lived experience vignettes on the FFSP were validating and normalized their own experiences. This feedback is reflective of broader research highlighting the value of sharing lived experiences to reduce stigma and discrimination around mental health and substance use for both individuals and the general public [55-57]. Similar to peer support programs, the sharing of lived experience stories plays an important role in empowering AFFMs by validating their unique difficulties supporting a loved one with AOD use while also normalizing seeking and accessing support for themselves [58]. Participants highlighted that they would like to see even more emphasis on stories and voices of lived experience in the FFSP.

The findings indicate that the FFSP was feasible to administer as a web-based, self-paced program that users are able to access at their discretion. The private and confidential format helps to address participant's concerns about stigma and fear of judgment. Further, the program being free, self-paced, and available 24/7 helps overcome other help-seeking barriers that participants endorsed such as financial strain, time restrictions, and other life stressors.

This study also examined AFFM's experiences of caring for a loved one using AOD in relation to their levels of distress, as well as their experiences of stress and strain, coping, and social support. On average, AFFMs in this study reported moderate to severe levels of psychological distress. This is in line with previous research highlighting AFFMs as a population may be vulnerable to increased risks of psychological distress and mental health difficulties [2,13,14]. Distress levels were similar across demographic variables including age, gender, geographical location (metropolitan, regional, or remote areas), and relationship to their loved ones. Similarly, there were no significant differences in SQFM-AA scores across these participant characteristics. For example, scores measuring the degree of burden falling on an AFFM due to the effects of their loved one's AOD use did not differ based on their relationship with the loved one. This finding is interesting as previous studies have found significant differences in symptoms of stress and level of harm based on gender, cultural factors, and relationship to the family member [26]. For example, a previous study in Brazil found that wives and mothers reported higher levels of burden compared to other family members including fathers [26]. A smaller sample size in this study, as well as uneven distribution across types of relationships and cultural backgrounds, may partially explain this different finding. Although this study did not focus on this, it would be important for future studies to explore the differences between social and cultural factors in order to inform the development of services that are culturally sensitive and responsive [59,60]. Finally, the majority of participants did not access or receive helpful support from formal services, consistent with previous findings of low rates of help-seeking among AFFMs which may also reflect a lack of accessible or appropriate formal support available [17].

Finally, higher levels of psychological distress were significantly associated with accessing the program. This is in line with this study's finding that the most common reason for participants not accessing the program was a perceived lack of need. Almost one-third of the reasons given for not accessing the program were related to the participant's loved one being in recovery or generally doing well. This may reflect the fact that recovery is a lifelong dynamic process dependent on individual, social, and contextual factors [61-64]. Knowing that AFFMs are greatly impacted by their loved one's AOD use, AFFMs' own mental health and support needs are also likely to change over time depending on both their loved one and their own life events. This points to the importance of ensuring programs such as the FFSP are a known option for AFFMs to access when they need it. This might include liaising with clinicians and health workers who are likely to be in contact with people during periods of difficulty who may then be able to refer individuals to the FFSP, positioning the program as a complementary tool in their broader recovery journey.

Strengths, Limitations, and Future Directions

A key limitation of this study was the low rate of program uptake and high attrition, which hindered the study's aim to assess the program's effectiveness, as well as limited the generalizability of the findings. As feedback on usability and acceptability was drawn from a small sample, these should be interpreted as preliminary, and future studies with larger, more representative samples are needed to build upon these findings. Low uptake of the program may be due to the recruitment strategy being focused on completing the surveys rather than

XSL•FO
RenderX

as an opportunity to use and test the program itself. Future research recruitment could benefit from clearly communicating the commitment required in signing up for the study, reimbursing participants at baseline to incentivize return, and increasing email reminders and contact with the research team between surveys to troubleshoot technical barriers. Further, the reliance on web-based recruitment platforms may have also biased our sample toward more educated and digitally literate populations. Future research recruitment could benefit from partnering with key AOD organizations (eg, Family Drug Support) or posting in AOD community forums to help reach a more diverse group of potential participants. High attrition may also be partly due to the impact of the pandemic and natural disasters occurring over the course of recruitment, however, poor adherence to web-based mental health and substance use interventions is a universal problem [65,66]. Further, AFFMs are a particularly vulnerable and hard-to-reach population facing ongoing barriers of stigma and carer burnout [17]. It is possible that some participants simply were not at a place in their journey where they felt able and ready to participate in the program or that they needed any additional support.

Previous research has found that integrating therapist or facilitator elements into web-based interventions tends to be associated with higher rates of engagement compared to self-guided interventions [67,68]. Thus, future strategies to promote engagement and retention of the FFSP might include nonclinician facilitation such as periodic email or text reminders, technical support, “gamification” (more structured or personalized progress tracking), peer-support elements (eg, moderated discussions), or clinician guidance such as complementing brief psychosocial support from a trained therapist. Further, enhancing usability and accessibility through optimizing user experience, particularly for mobile may increase engagement and adherence as most people access the internet on their smartphone [69]. To address technical difficulties encountered by participants (eg, broken links or forgetting passwords), it would be valuable to incorporate and evaluate user-friendly enhancements such as a digital skills guide or tutorial embedded into the program to help AFFMs navigate and benefit from the program. In addition, regular reviews of the program and provision of digital technical support (eg, through the chat function) would help to quickly identify and resolve issues. Finally, future content improvements could include more lived experience stories, prioritizing clarity and conciseness of information per page, and additional resources and links to in-person services to provide follow-up and encourage people to build a network of support options. These learnings will guide a future large-scale trial evaluating the efficacy and outcomes of the FFSP.

This study also had a number of strengths. First, an intention-to-treat approach was followed which aimed to minimize the risk of bias related to adherence and attrition in regard to assessing the acceptability of the FFSP. This allowed us to capture more diverse perspectives and account for potential differences between those who do and do not access the program and those who do and do not complete the postprogram and follow-up surveys.

Second, as reflected in the broader literature, AFFMs are a particularly hard-to-reach population in research and health care considering the unpredictability of their carer roles and the concurrent burdens they face, as well as the ongoing impacts of stigma and social isolation. This study was successful in recruiting a relatively representative sample of AFFMs which included a range of genders, ages, geographical locations, and relationships to the loved one using AOD. The flexibility of the FFSP allows family members and friends to access the program 24/7 when they have the time, privacy, and space to access support. This flexibility is key to reaching at-risk groups including women (who often bear the greater caregiver burden), low socioeconomic communities (who experience greater financial stress), and those living in rural, or regional, or remote areas with limited resources [1,39,45]. Further, a number of participants reported a perceived lack of need for treatment at the time of completing the study which reflects the nonlinear nature of AOD recovery and fluctuating need for support. This points to the importance of continuing to make the FFSP readily available so that AFFMs can access the program when and wherever they may need it. As reflected in the quantitative and qualitative feedback on the program, participants provided positive feedback on the ease, convenience, and need for web-based programs.

Conclusions

This study adds to the growing body of literature demonstrating high levels of distress and strain among people caring for a loved one using AOD, while highlighting the ongoing barriers to help-seeking among this population, as well as the opportunities for early intervention and support. This pilot study found that the FFSP is feasible to administer, acceptable, and has the potential to fill an important gap in services specific to family and friends supporting a loved one using AOD. This study also identified areas for improvement of the program and highlighted important learnings prior to conducting an evaluation of program efficacy. While the FFSP is a promising intervention for AFFMs, there is an ongoing need for increased research and investment into reducing stigma and other barriers to help-seeking, as well as developing and evaluating early intervention and support for both people who use AOD and the family members and friends who care for them.

Acknowledgments

The development and implementation of the Family and Friends Support program was funded by the Australian Government Department of Health. The Alcohol and Drug Foundation provided funding to the University of Newcastle and the Matilda Centre for Research in Mental Health and Substance Use to conduct the evaluation of the program. Thank you to Miss Hannah Wells, Miss Jessica Wilson, and Miss Caitlin Bray who also assisted in the development and data collection of this study. We would

XSL•FO
RenderX

also like to acknowledge and thank the members of the Australian community who generously shared their experiences and contributed valuable perspectives and insights to this study.

Data Availability

The anonymous datasets generated and analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

SK, DS, FK-L, and CC conceptualized the research design and paper. SK, MK, BS, and AG undertook data collection and cleaning. JD led the analysis of data. SK and JD led the manuscript drafting and submission. DS, KR, NN, FK-L, MT, and CC provided expert guidance throughout the research design and analysis. All authors reviewed the manuscript and contributed to the intellectual content of the article.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Screenshots from the program.
[DOCX File, 2246 KB - formative_v9i1e72425_app1.docx]

Multimedia Appendix 2
Study procedure.
[DOCX File, 32 KB - formative_v9i1e72425_app2.docx]

Multimedia Appendix 3
SQFM-AA (Short Questionnaire for Family Members-Affected by Addiction) individual item endorsement.
[DOCX File, 59 KB - formative_v9i1e72425_app3.docx]

Multimedia Appendix 4
K-10 (Kessler Psychological Distress Scale) and SQFM-AA (Short Questionnaire for Family Members-Affected by Addiction) descriptive statistics.
[DOCX File, 17 KB - formative_v9i1e72425_app4.docx]

References

1. Kourgiantakis T, Ashcroft R, Mohamud F, Fearing G, Sanders J. Family-focused practices in addictions: a scoping review. J Soc Work Pract Addict 2021 Jan 2;21(1):18-53. [doi: 10.1080/1533256X.2020.1870287]
2. Orford J, Velleman R, Copello A, Templeton L, Ibanga A. The experiences of affected family members: a summary of two decades of qualitative research. Drugs Educ Prev Policy 2010 Dec;17(sup1):44-62. [doi: 10.3109/09687637.2010.514192]
3. Australian Guidelines to Reduce Health Risks from Drinking Alcohol: NHMRC; 2020.
4. National Drug Strategy Household Survey 2022–2023: Australian Institute of Health and Welfare; 2024.
5. Australian Burden of Disease Study 2018: Interactive Data on Risk Factor Burden 2021: Australian Institute of Health and Welfare; 2021.
6. Orford J, Velleman R, Natera G, Templeton L, Copello A. Addiction in the family is a major but neglected contributor to the global burden of adult ill-health. Soc Sci Med 2013 Feb;78:70-77. [doi: 10.1016/j.socscimed.2012.11.036]
7. Merkouris SS, Rodda SN, Dowling NA. Affected other interventions: a systematic review and meta-analysis across addictions. Addiction 2022 Sep;117(9):2393-2414. [doi: 10.1111/add.15825] [Medline: 35129234]
8. McCann TV, Lubman DI, Boardman G, Flood M. Affected family members' experience of, and coping with, aggression and violence within the context of problematic substance use: a qualitative study. BMC Psychiatry 2017 Jun 2;17(1):209. [doi: 10.1186/s12888-017-1374-3] [Medline: 28578666]
9. Sampson D, Heinsch M, Geddes J, et al. "I no longer know that person": experiences of families living with someone using crystal methamphetamine. PLoS One 2023;18(4):e0284156. [doi: 10.1371/journal.pone.0284156] [Medline: 37043501]
10. Ray GT, Mertens JR, Weisner C. Family members of people with alcohol or drug dependence: health problems and medical cost compared to family members of people with diabetes and asthma. Addiction 2009 Feb;104(2):203-214. [doi: 10.1111/j.1360-0443.2008.02447.x] [Medline: 19149814]

XSL•FO
RenderX

11. Birkeland B, Foster K, Selbekk AS, Høie MM, Ruud T, Weimand B. The quality of life when a partner has substance use problems: a scoping review. Health Qual Life Outcomes 2018 Nov 20;16(1):219. [doi: 10.1186/s12955-018-1042-4] [Medline: 30453992]
12. Vederhus JK, Kristensen Ø, Timko C. How do psychological characteristics of family members affected by substance use influence quality of life? Qual Life Res 2019 Aug;28(8):2161-2170. [doi: 10.1007/s11136-019-02169-x] [Medline: 30895489]
13. Ólafsdóttir J, Orjasniemi T, Hrafnsdóttir S. Psychosocial distress, physical illness, and social behaviour of close relatives to people with substance use disorders. J Soc Work Pract Addict 2020 Apr 2;20(2):136-154. [doi: 10.1080/1533256X.2020.1749363]
14. Di Sarno M, De Candia V, Rancati F, Madeddu F, Calati R, Di Pierro R. Mental and physical health in family members of substance users: a scoping review. Drug Alcohol Depend 2021 Feb 1;219:108439. [doi: 10.1016/j.drugalcdep.2020.108439] [Medline: 33333362]
15. Copello A, Templeton L, Orford J, Velleman R. The 5-step method: evidence of gains for affected family members. Drugs Educ Prev Policy 2010 Dec;17(sup1):100-112. [doi: 10.3109/09687637.2010.514234]
16. McDonagh D, Connolly N, Devaney C. “Bury don’t discuss”: the help-seeking behaviour of family members affected by substance-use disorders. Child Care in Practice 2019 Apr 3;25(2):175-188. [doi: 10.1080/13575279.2018.1448258]
17. McCann TV, Lubman DI. Stigma experience of families supporting an adult member with substance misuse. Int J Ment Health Nurs 2018 Apr;27(2):693-701. [doi: 10.1111/inm.12355] [Medline: 28593690]
18. Goffman E. Stigma: Notes on the Management of Spoiled Identity: Simon and Schuster; 2009.
19. Corrigan PW, Watson AC, Miller FE. Blame, shame, and contamination: the impact of mental illness and drug dependence stigma on family members. J Fam Psychol 2006 Jun;20(2):239-246. [doi: 10.1037/0893-3200.20.2.239] [Medline: 16756399]
20. McCann TV, Lubman DI. Help-seeking barriers and facilitators for affected family members of a relative with alcohol and other drug misuse: a qualitative study. J Subst Abuse Treat 2018 Oct;93:7-14. [doi: 10.1016/j.jsat.2018.07.005] [Medline: 30126543]
21. McCann TV, Lubman DI. Adaptive coping strategies of affected family members of a relative with substance misuse: a qualitative study. J Adv Nurs 2018 Jan;74(1):100-109. [doi: 10.1111/jan.13405] [Medline: 28771795]
22. Duggan C. The experiences of families seeking support in coping with heroin use. : Lenus; 2007 URL: http://hdl.handle.net/10147/45167 [accessed 2025-06-12]
23. Arcidiacono C, Velleman R, Procentese F, et al. Italian families living with relatives with alcohol or drugs problems. Drugs Educ Prev Policy 2010 Dec;17(6):659-680. [doi: 10.3109/09687630902824262]
24. Dussaillant F, Fernandez M. Alcohol’s harm to others’ well-being and health: a comparison between Chile and Australia. Alcohol Alcohol 2015 May;50(3):346-351. [doi: 10.1093/alcalc/agv002] [Medline: 25694495]
25. Velleman R, Orford J, Templeton L, et al. 12-month follow-up after brief interventions in primary care for family members affected by the substance misuse problem of a close relative. Addict Res Theory 2011 Aug;19(4):362-374. [doi: 10.3109/16066359.2011.564691]
26. Orford J. How does the common core to the harm experienced by affected family members vary by relationship, social and cultural factors? Drugs Educ Prev Policy 2017 Jan 2;24(1):9-16. [doi: 10.1080/09687637.2016.1189876]
27. Orford J, Cousins J, Smith N, Bowden-Jones H. Stress, strain, coping and social support for affected family members attending the National Problem Gambling Clinic, London. Int Gambl Stud 2017 May 4;17(2):259-275. [doi: 10.1080/14459795.2017.1331251]
28. Orford J, Padin MFR, Canfield M, Sakiyama HMT, Laranjeira R, Mitsuhiro SS. The burden experienced by Brazilian family members affected by their relatives’ alcohol or drug misuse. Drugs Educ Prev Policy 2019 Mar 4;26(2):157-165. [doi: 10.1080/09687637.2017.1393500]
29. Ahuja A, Orford J, Copello A. Understanding how families cope with alcohol problems in the UK West Midlands Sikh community. Contemp Drug Probl 2003 Dec;30(4):839-873. [doi: 10.1177/009145090303000406]
30. Velleman RDB, Templeton LJ, Copello AG. The role of the family in preventing and intervening with substance use and misuse: a comprehensive review of family interventions, with a focus on young people. Drug Alcohol Rev 2005 Mar;24(2):93-109. [doi: 10.1080/09595230500167478] [Medline: 16076580]
31. Orr LC, Barbour RS, Elliott L. Involving families and carers in drug services: are families ‘part of the problem’? Fam Relatsh Soc 2014;3(3):405-424. [doi: 10.1332/204674313X669900]
32. Orford J, Copello A, Velleman R, Templeton L. Family members affected by a close relative’s addiction: the stress-strain-coping-support model. Drugs Educ Prev Policy 2010 Dec;17(sup1):36-43. [doi: 10.3109/09687637.2010.514801]
33. Copello A, Templeton L, Orford J, Velleman R. The 5-step method: principles and practice. Drugs Educ Prev Policy 2010 Dec;17(sup1):86-99. [doi: 10.3109/09687637.2010.515186]
34. Xia Y, Gong Y, Wang H, Li S, Mao F. Family function impacts relapse tendency in substance use disorder: mediated through self-esteem and resilience. Front Psychiatry 2022;13:815118. [doi: 10.3389/fpsyt.2022.815118] [Medline: 35237192]
35. Copello AG, Templeton L, Velleman R. Family interventions for drug and alcohol misuse: is there a best practice? Curr Opin Psychiatry 2006 May;19(3):271-276. [doi: 10.1097/01.yco.0000218597.31184.41] [Medline: 16612212]

XSL•FO
RenderX

36. Roozen HG, de Waart R, van der Kroft P. Community reinforcement and family training: an effective option to engage treatment-resistant substance-abusing individuals in treatment. Addiction 2010 Oct;105(10):1729-1738. [doi: 10.1111/j.1360-0443.2010.03016.x] [Medline: 20626372]
37. Rushton CM, Kelly PJ, Thomas T, et al. SMART family and friends: feasibility and outcomes of a video-conference delivered intervention for families impacted by another's methamphetamine use. J Subst Use Addict Treat 2024 Jun;161:209355. [doi: 10.1016/j.josat.2024.209355] [Medline: 38548059]
38. Peart A, et al. The experiences of family members attending an online addiction education program: a qualitative study. Drugs Educ Prev Policy 2024;32(1):1-8. [doi: 10.1080/09687637.2023.2184248]
39. Degenhardt L, Wolfe D, Hall W, et al. Strategies to reduce drug-related harm: responding to the evidence base. Lancet 2019 Oct;394(10208):1490-1493. [doi: 10.1016/S0140-6736(19)32232-9]
40. Help-seeking by family and friends of people who use alcohol and drugs: survey report. : Alcohol and Drug Foundation; 2024 URL: https://cdn.adf.org.au/media/documents/IS_Survey_Report_2024_Help-Seeking_by_Family_and_Friends.pdf [accessed 2025-06-15]
41. McCann TV, Stephenson J, Lubman DI. Family stress, well-being and hope for the future when supporting a relative with AOD misuse: a cross-sectional survey. Issues Ment Health Nurs 2021 May;42(5):430-436. [doi: 10.1080/01612840.2020.1817207] [Medline: 33886421]
42. McCann TV, Polacsek M, Lubman DI. Experiences of family members supporting a relative with substance use problems: a qualitative study. Scand J Caring Sci 2019 Dec;33(4):902-911. [doi: 10.1111/scs.12688] [Medline: 31033023]
43. Family and friend support program. FFSP. 2025. URL: https://ffsp.com.au [accessed 2024-10-21]
44. Kay-Lambkin F. Content development for Family and Friends Support Program. : Australian Government Department of Health; 2020.
45. Laslett AM, Room R, Kuntsche S, et al. Alcohol's harm to others in 2021: who bears the burden? Addiction 2023 Sep;118(9):1726-1738. [doi: 10.1111/add.16205] [Medline: 37052510]
46. Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: building an international community of software platform partners. J Biomed Inform 2019 Jul;95:103208. [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]
47. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research Electronic Data Capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381. [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
48. Location: census. Australian Bureau of Statistics. 2021. URL: https://www.abs.gov.au/statistics/people/people-and-communities/location-census/latest-release [accessed 2024-07-11]
49. Kessler RC, Andrews G, Colpe LJ, et al. Short screening scales to monitor population prevalences and trends in non-specific psychological distress. Psychol Med 2002 Aug;32(6):959-976. [doi: 10.1017/s0033291702006074] [Medline: 12214795]
50. Wojujutari AK, Idemudia ES. Consistency as the currency in psychological measures: a reliability generalization meta-analysis of Kessler Psychological Distress Scale (K-10 and K-6). Depress Anxiety 2024;2024(1):3801950. [doi: 10.1155/2024/3801950] [Medline: 40226709]
51. Orford J, Templeton L, Velleman R, Copello A. Family members of relatives with alcohol, drug and gambling problems: a set of standardized questionnaires for assessing stress, coping and strain. Addiction 2005 Nov;100(11):1611-1624. [doi: 10.1111/j.1360-0443.2005.01178.x] [Medline: 16277623]
52. Meadows G, Harvey C, Fossey E, Burgess P. Assessing perceived need for mental health care in a community survey: development of the Perceived Need for Care Questionnaire (PNCQ). Soc Psychiatry Psychiatr Epidemiol 2000 Sep;35(9):427-435. [doi: 10.1007/s001270050260] [Medline: 11089671]
53. Kershaw S, Sunderland M, Grager A, et al. Perceived barriers to help-seeking for people who use crystal methamphetamine: perspectives of people with lived experience, family members and health workers. Drug Alcohol Rev 2024 Nov;43(7):1929-1939. [doi: 10.1111/dar.13897] [Medline: 38965840]
54. Sauro J. Measuring usability with the System Usability Scale (SUS). Measuring U. 2011. URL: https://measuringu.com/sus/ [accessed 2024-07-11]
55. Treloar C, Pienaar K, Dilkes-Frayne E, Fraser S. Lives of substance: a mixed-method evaluation of a public information website on addiction experiences. Drugs Educ Prev Policy 2019 Mar 4;26(2):140-147. [doi: 10.1080/09687637.2017.1397602]
56. Jorm AF. Effect of contact-based interventions on stigma and discrimination: a critical examination of the evidence. Psychiatr Serv 2020 Jul 1;71(7):735-737. [doi: 10.1176/appi.ps.201900587] [Medline: 32188364]
57. Morgan AJ, Wright J, Reavley NJ. Review of Australian initiatives to reduce stigma towards people with complex mental illness: what exists and what works? Int J Ment Health Syst 2021 Jan 18;15(1):10. [doi: 10.1186/s13033-020-00423-1] [Medline: 33461567]
58. Tari-Keresztes N, Armstrong N, Smith JA, Gupta H, Goding S, Endemann SA. "You don't get that from professionals": a consumer-led peer recovery program for families and friends of individuals with alcohol and other drugs use issues in Darwin. Int J Environ Res Public Health 2023 Apr 14;20(8):5514. [doi: 10.3390/ijerph20085514] [Medline: 37107795]
59. Marel C, et al. People from culturally and linguistically diverse backgrounds, in guidelines on the management of co-occuring alcohol and other drug and mental health conditions in alcohol and other drug treatment settings. In: Matilda Centre for Research in Mental Health and Substance Use: The University of Sydney; 2022:311-315.

XSL•FO
RenderX

60. Rickwood D. Pathways of Recovery Preventing Further Episodes of Mental Illness (Monograph): Commonwealth of Australia; 2006.
61. Laudet AB, Savage R, Mahmood D. Pathways to long-term recovery: a preliminary investigation. J Psychoactive Drugs 2002;34(3):305-311. [doi: 10.1080/02791072.2002.10399968] [Medline: 12422942]
62. Vaillant GE. A 60-year follow-up of alcoholic men. Addiction 2003 Aug;98(8):1043-1051. [doi: 10.1046/j.1360-0443.2003.00422.x] [Medline: 12873238]
63. Scott CK, Foss MA, Dennis ML. Pathways in the relapse—treatment—recovery cycle over 3 years. J Subst Abuse Treat 2005;28 Suppl 1(2):S63-S72. [doi: 10.1016/j.jsat.2004.09.006] [Medline: 15797640]
64. Laudet AB. The road to recovery: where are we going and how do we get there? Empirically driven conclusions and future directions for service development and research. Subst Use Misuse 2008;43(12-13):2001-2020. [doi: 10.1080/10826080802293459] [Medline: 19016176]
65. Forbes A, Keleher MR, Venditto M, DiBiasi F. Assessing patient adherence to and engagement with digital interventions for depression in clinical trials: systematic literature review. J Med Internet Res 2023 Aug 11;25:e43727. [doi: 10.2196/43727] [Medline: 37566447]
66. Shams F, Tai AMY, Kim J, et al. Adherence to e-health interventions for substance use and the factors influencing it: systematic review, meta-analysis, and meta-regression. Digital Health 2023;9:20552076231203876. [doi: 10.1177/20552076231203876] [Medline: 37780062]
67. Borghouts J, Eikey E, Mark G, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. J Med Internet Res 2021 Mar 24;23(3):e24387. [doi: 10.2196/24387] [Medline: 33759801]
68. Baumeister H, Reichler L, Munzinger M, Lin J. The impact of guidance on internet-based mental health interventions—a systematic review. Internet Interventions 2014 Oct;1(4):205-215. [doi: 10.1016/j.invent.2014.08.003]
69. Jakob R, Harperink S, Rudolf AM, et al. Factors influencing adherence to mHealth apps for prevention or management of noncommunicable diseases: systematic review. J Med Internet Res 2022 May 25;24(5):e35371. [doi: 10.2196/35371] [Medline: 35612886]

Abbreviations

AFFM: affected friends and family members
AOD: alcohol and other drugs
FFSP: Family and Friend Support Program
K-10: Kessler Psychological Distress Scale
REDCap: Research Electronic Data Capture
SQFM-AA: Short Questionnaire for Family Members-Affected by Addiction
SUS: System Usability Scale

Edited by A Mavragani; submitted 11.02.25; peer-reviewed by M Adnan, T Adebambo; revised version received 13.05.25; accepted 21.05.25; published 09.07.25.

Please cite as:
Kershaw S, Deng J, Keaveny M, Speirs B, Grager A, Sampson D, Ross K, Newton N, Teeson M, Kay-Lambkin F, Chapman C
A Web-Based Well-Being and Resilience Intervention for Family Members and Friends Supporting a Loved One Using Alcohol and Other Drugs: Mixed Methods Pilot Study
JMIR Form Res 2025;9:e72425
URL: https://formative.jmir.org/2025/1/e72425
doi:10.2196/72425

© Stephanie Kershaw, Jessica Deng, Madeleine Keaveny, Bronte Speirs, Anna Grager, Dara Sampson, Kate Ross, Nicola Newton, Maree Teeson, Frances Kay-Lambkin, Cath Chapman. Originally published in JMIR Formative Research (https://formative.jmir.org), 9.7.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Vibrotactile Feedback Strategies for Trunk-Stabilizing Exercises in a Home-Based Scenario: Qualitative Interview Study Among Physiotherapists

Kaya Holzmeyer[1], MSc; Lisa-Marie Lüneburg[2,3], Dipl.-Ing.; Luca Oppici[4], PhD; Philipp Flößel[5], MSc; Doris Lachmann[5], Dr med; Jens Krzywinski[2,3], Prof Dr; Susanne Narciss[1,3], Prof Dr

[1]Psychology of Learning and Instruction, Faculty of Psychology, Technische Universität Dresden, Dresden, Germany

[2]Industrial Design Engineering, Faculty of Mechanical Engineering, Technische Universität Dresden, Helmholtzstr. 10, Dresden, Germany

[3]Cluster of Excellence CeTI: "Centre for Tactile Internet with Human-in-the-Loop", Technische Universität Dresden, Dresden, Germany

[4]The Department of Teacher Education and Outdoor Studies, Norwegian School of Sport Sciences, Oslo, Norway

[5]Center of Orthopedic, Trauma and Plastic Surgery, University Hospital Carl Gustav Carus, Medical Faculty Carl Gustav Carus, Technische Universität Dresden, Dresden, Germany

Corresponding Author:
Lisa-Marie Lüneburg, Dipl.-Ing.
Industrial Design Engineering, Faculty of Mechanical Engineering, Technische Universität Dresden, Helmholtzstr. 10, Dresden, Germany

Abstract

Background: Physiotherapy treatments frequently incorporate a combination of hands-on, therapist-led exercises and hands-off, home-based exercises, whereby patients perform exercises independently. A key challenge in home-based rehabilitation is providing patients with effective feedback to guide their movements. A range of strategies is being reviewed to address this issue and a starting point for developing effective and implementable strategies to gather feedback requirements from the end users is also being considered.

Objective: This study aims to derive requirements for designing a vibrotactile feedback strategy for a wearable device for physiotherapy home exercises. To achieve this, it is essential to consider the principles of feedback design as set forth by the field of instructional psychology and to involve physiotherapists in the process.

Methods: The feedback behavior of 9 physical therapists was observed during a staged training scenario. Guided interviews were conducted to explore their recommendations for a vibrotactile feedback system. Observational data and interviews were analyzed using a combined deductive and inductive category system. For each exercise-specific motion pattern, it was recorded whether feedback was provided. Instances of feedback were systematically coded according to feedback modality, timing, and content; haptic feedback was additionally categorized by localization. Interview statements referring to the use of vibrotactile feedback were categorized by modality, task requirements, localization, time, content, frequency, function, and individualization. Quantitative data from observations were evaluated using frequency distributions, and qualitative interview data were analyzed using content structuring content analysis.

Results: On average, 50.2% (SD 13.4%) of the observed feedback content provided by each therapist consisted of the presentation of correct exercise execution. A smaller proportion was exclusively confirmatory feedback or the additional provision of elaborated information. Therapists provided feedback in 57.3% (SD 16.3%) of the movement repetitions, on average, while no feedback was given in 42.7% (SD 16.3%) of the repetitions. The interview data supported a reduction in feedback. Regarding feedback timing, 70.8% (SD 13.1%) of the feedback observations were given concurrently with the task across therapists; the proportion of feedback given after the exercise execution was smaller. On average, across therapists, 51.9% (SD 12.1%) of the feedback was auditory, and a further proportion was multimodal with an auditory component. Regarding feedback localization, haptic feedback was often given proximal to the trunk, in the knee and foot region. Frequent combinations of the categories were auditory, simple confirmatory feedback, often given when terminal, or the presentation of the correct exercise execution given auditorily or multimodally concurrent to exercise execution.

Conclusions: This study offers insights into the design of a vibrotactile feedback strategy. We identified key feedback characteristics recommended by physical therapists: reduced frequency, differentiated content, multimodal feedback, and trunk-proximal vibrotactile cues. These approaches should be tested in hypotheses-driven research and further patient studies should be conducted.

(JMIR Form Res 2025;9:e62903) doi:10.2196/62903

XSL•FO
RenderX

KEYWORDS

vibrotactile feedback; physiotherapy; home exercises; motor learning; Internet of Things; haptic; wearable; qualitative research; user-centered design; digital health; back pain

Introduction

Background

Feedback is a key component in physiotherapy, playing a crucial role in achieving therapy goals [1,2]. Physiotherapists provide hands-on feedback during treatment, a form that is lacking for patients following an at-home exercise regime. The use of home exercises has seen a notable rise in recent years. This growth can be attributed to 2 key factors: the effectiveness of active methods over passive treatment approaches and the shortage of health care professionals [3]. Standard technologies include apps and devices that provide visual or auditory feedback [4-11], with some using gamification approaches [12,13]. Recent advances in tactile internet technologies (tactile internet with human in the loop) have opened up new possibilities for manipulating and controlling virtual or real objects with almost no latency, ensuring minimal delay in data transmission. These technologies allow for real-time tactile feedback, which has been shown to positively impact motor skill performance and learning [14-19]. Such feedback is increasingly used in rehabilitation, offering innovative approaches to physical therapy [20-23]. Specifically, wearables, which are devices that embed technology into garments or accessories directly worn close to the skin, such as shirts, bracelets, watches, anklets, or waistcoats, are suitable for those purposes. These devices use vibration actuators to deliver tactile feedback to the wearer while using sensors to monitor and classify movement data during exercises. In the context of home-based physiotherapy, external tactile feedback offers a promising approach to address the absence of hands-on guidance. This tactile feedback can be delivered through wearable systems. This study is part of a broader effort to develop a haptic physiotherapy assistant, including a wearable device that provides vibrotactile feedback. Among the wearable formats, the shirt stands out as particularly promising due to its ability to cover larger areas of the body, allowing for precise and distributed feedback during rehabilitation exercises.

Recent reviews highlight both the potential of these wearable technologies for home-based rehabilitation and the need for empirically driven strategies to optimize feedback mechanisms [24,25]. By addressing these gaps, this study aims to establish a robust foundation for designing and implementing effective haptic feedback systems for physiotherapy practice.

Therefore, challenges from the perspective of instructional psychology and interaction design are considered when developing a vibrotactile feedback strategy. The interactive tutoring feedback (ITF) model, which originated in instructional psychology, posits that feedback information can be sourced from 2 distinct domains: the internal domain, which encompasses information generated internally by the feedback recipient, such as proprioception, and the external domain, which encompasses feedback originating from external sources, such as wearables [26,27]. A haptic physiotherapy assistant, which is targeted by this research project, is an external feedback source. It is crucial to design the external feedback in a manner that does not override the internal feedback of the feedback recipients. As internal feedback is available to the feedback recipient in the long term, external feedback should be designed in such a way that the recipient is supported in generating internal feedback. This approach minimizes dependence on the external system and facilitates long-term learning [26-28]. Therefore, this study follows a strategic approach to design effective external feedback, that is, a coordinated plan that implies decisions about specific feedback properties [26,27]. A feedback strategy integrates decisions about at least the following properties:

Which content should be provided by the feedback? Feedback always provides a certain type of evaluative information (eg, knowledge of the result in terms of a single confirmation or rejection or knowledge of the correct result). Furthermore, it can provide additional, elaborated information components. Elaborated components can offer information about task requirements, conceptual knowledge, error type, learning process, or strategies, that is, information supporting the feedback recipient in further knowledge-gaining or learning process, or a combination of these components [26].

How is the feedback content presented using the formal and technical design properties that are available in a task context? Designing the presentation of the feedback contents includes making decisions on feedback timing, scheduling, frequency, and the feedback modality. Concerning the feedback timing, literature differentiates between concurrent feedback, that is, information delivered during task execution, and terminal feedback, that is, information that is either delivered immediately or with delay after finishing a task. Feedback can be presented continuously after each event or intermittently with a reduced frequency, such as fading-in or fading-out strategies. A technical feedback system can provide feedback in a unimodal vibrotactile way or in a multimodal way by supplementing, for example, vibrotactile feedback with other modalities.

A challenge in the design process is the early integration of potential users. Human-centered design aims to increase the usability of end products. An important goal is to promote and improve the understanding of tasks, contexts, and users, as well as the resulting requirements [29-31]. This work is settled in the early phase of the development of a wearable feedback system, the unstructured and uncertain fuzzy front-end. So, human-centered design could be an effective method of improving the design process [31].

To the best of our knowledge, no extant literature addresses the characteristics of tactile feedback in the context of physiotherapy home exercises, highlighting the need for further investigation.

Objectives

The objective of this study is to derive requirements for effective external vibrotactile feedback for home exercises in

XSL•FO
RenderX

physiotherapy. Specifically, the investigation will encompass the characteristics that, following the principles of instructional psychology, must be considered during the design process to ensure the effectiveness of external feedback (modality, content, and timing) [26-28]. Concurrently, we are integrating feedback behavior (modality, content, timing, and localization) from physiotherapists to understand the context of use.

Methods

Overview

We conducted a qualitative study to gain a fundamental understanding of physiotherapists' feedback behavior when providing home-based exercise instructions. The reporting and writing of this study were guided by the standards for reporting qualitative research and the COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist [32,33] (Checklist 1). The data collection was structured in 2 parts. In the first part, we observed how physical therapists delivered feedback to a standardized patient in a typical hands-on scenario. In the second part, we asked the physical therapist to recommend the most appropriate use of a wearable device to provide feedback to a standardized patient in a hands-off home-based scenario. The data was collected using video-based participant observation and guided interviews. We staged a training scenario with home exercises for low back pain. The study was conducted over the period from February to March 2022.

Participants

Physical therapists working in physiotherapy facilities in Dresden, Germany, were recruited as study participants. A purposive sample of 30 public physiotherapy facilities was initially contacted via email and telephone to assess their interest in study participation. All facilities offered training therapy or active physiotherapy, as indicated on their websites. At the initial point of contact, which was conducted for a specific purpose and by the protocol for engaging with public institutions, physiotherapists were informed about this study's aim, procedures, and data collection. Physiotherapists were invited to participate after confirming their interest. This recruitment procedure was authorized by the Data Protection Officer of TU Dresden, as required by § 12 (1) of the Saxon Data Protection Implementation Act, with no alternative data collection methods identified. The planned inclusion criteria of participants were at least 3 years of professional experience as a physiotherapist and expertise in the treatment of home exercises. The investigator and participants were not previously acquainted, except for the initial recruitment contact.

Ethical Considerations

This study was approved by the Ethics Committee of Technische Universität Dresden (SR-EK-529112021) and was conducted in accordance with the principles of the Declaration of Helsinki.

The research involved human participants and therefore underwent full ethical review. This study's design, including procedures for data collection via video recording, was assessed and approved under the stated institutional review board protocol. No exemptions or waivers were sought. The approval documentation is available upon request.

Before participation, all individuals were provided with detailed written information about this study's aims, procedures, data handling practices, and their rights, including the right to withdraw at any time without penalty. Written informed consent was obtained from all participants. The consent process explicitly addressed the use of video data and the scope of data processing involved.

All data collected during this study were either anonymized or deidentified to protect participant privacy. In cases where full anonymization was not feasible due to the nature of the video recordings, strict access controls and data encryption were implemented to safeguard the information. Data access was restricted to authorized research personnel only.

Participants were compensated €15 (US $17.15) for their time and any inconvenience incurred due to their involvement in this study, including potential work time lost. Compensation was not contingent on completion of this study.

Staged Training and Interview

The staged training session consisted of 5 home exercises for the treatment of nonspecific low back pain. We preselected home exercises based on a multicenter controlled trial [34]. Following a systematic selection process, 3 physiotherapists evaluated and selected effective and commonly used exercises through interviews. The physiotherapists added additional exercises for daily use to ensure practical applicability. We extracted requirements and 20 typical errors for the 5 exercises. To ensure comparability of the exercises and the physiotherapists' feedback behavior, we trained 1 healthy individual to perform the exercises in both correct and error conditions. This approach allowed for a reproducible staged training session with a standardized patient (Figure 1). A detailed description of the staged training is provided in Multimedia Appendix 1. We obtained an explicit consent from the individual in Figure 1 as well as in Multimedia Appendix 1 to use their image for publication.

XSL•FO
RenderX

Figure 1. Overview of home exercises and predefined error conditions used during the staged training session.

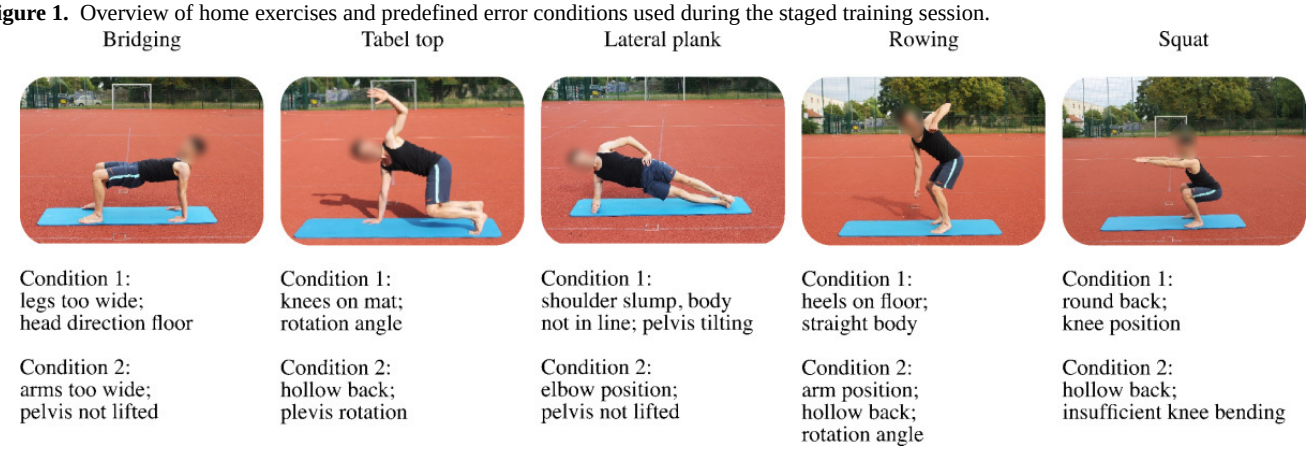

Interview guide development was grounded in the ITF model, which describes the influencing factors, interdependencies and effects of feedback in competence acquisition and can thus form the basis for the development of effective feedback strategies [26,27]. The semistructured interview mainly contained open-ended questions and was supplemented by individual closed-ended questions (Table 1). The order and phrasing of the questions were flexible and could be adapted to the course of the interview. To be able to go into more detail about unforeseen, thematically relevant statements, additional ad hoc questions were asked. The interview lasted approximately 15 minutes and was audio-recorded. Notes and photos were taken to supplement the audio information, for example, when a participant visualized a statement.

Table . Semistructured interview schedule.

Topic	Questions and prompts
Task requirements	Please describe a possible behavior that could initiate vibrotactile feedback.
Localization	Besides the body localization you already showed, are there other body parts where vibrotactile feedback could be delivered?
Timing	To schedule the feedback to movement repetitions: when should vibrotactile feedback be presented?
Frequency	How often should vibrotactile feedback be delivered? Within a single training session? Within the training period? Which technique of reducing frequency is effective?
Content	Which information should be given by vibrotactile feedback?
Individualization	Which feedback properties should be adaptable to individuals?

A preliminary pilot study was conducted in an unpublished format before the main study. The piloting was performed with a trained physiotherapist with 10 years of professional experience; he fulfilled the inclusion criteria of the selected sample. The piloting aimed to check the comprehensibility and accuracy of the instructions, questions, interview, and home exercises and to develop them further. Following the piloting phase, we refined the semistructured interview questions and modified the execution of 2 exercises (rowing and squatting). The piloting also allowed us to validate the research design and test its implementation. This process also enhanced the study's feasibility and confidence for both the standardized patient and the researcher.

Study Procedure

This study took place at the participants' facilities and involved individual observation and interviews (Figure 2).

XSL•FO
RenderX

Figure 2. Study procedure with 3 main phases of data collection. TaHiL: tactile internet with human in the loop.

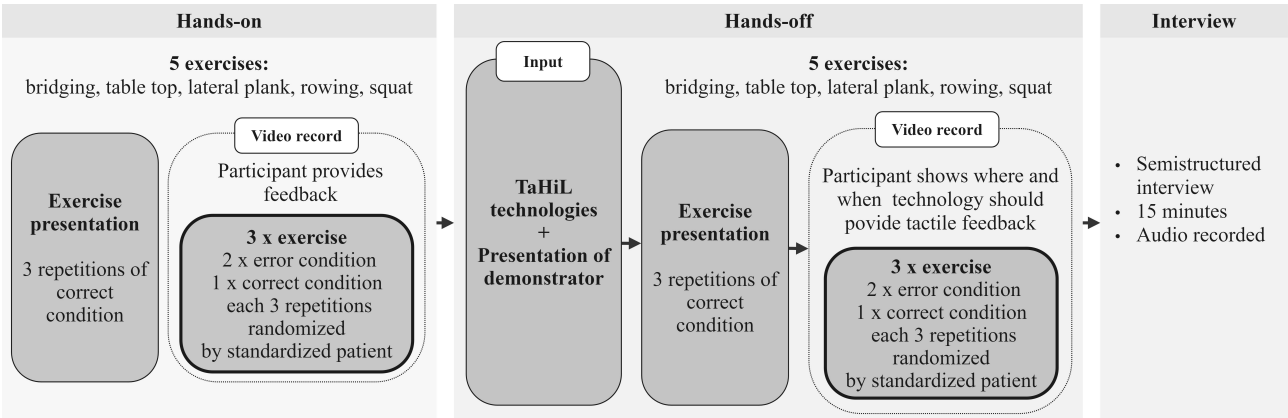

After completing demographic inquiries, we started the staged training which consists of a hands-on and hands-off part followed by an interview as mentioned above. Participants were instructed to provide feedback to the standardized patient during the staged training session, deciding independently on the type and timing of feedback. Each exercise began with a correct demonstration to illustrate the target execution, during which no feedback was given. Participants then completed 3 runs of the same exercise: 1 without errors and 2 with errors. Each run consisted of 3 repetitions of a movement pattern or 5 seconds of stabilization. After completing 1 exercise, participants moved on to the next, performing it across 3 feedback runs. When participants provided feedback, the standardized patient responded by performing additional repetitions or maintaining stabilization in the corrected form, explicitly addressing the feedback without correcting unmentioned errors.

After finishing the hands-on part, participants were informed about tactile internet with human in the loop interventions and a simple demonstrator of a possible wearable feedback device was presented (Figure 3). This assistant is a wearable device—a body-worn cyber-physical system designed to enable advanced multimodal human-machine interactions through vibrotactile feedback. The envisioned system provides real-time haptic feedback through vibrations on specific body parts where a posture needs to be corrected during the exercise. Furthermore, the concept entails an AI-supported expert system with embedded sensors to classify movement data during exercises, triggering the vibrotactile feedback when deviations from the intended movement pattern are detected. This technology can provide intuitive, hands-free guidance and support by incorporating vibrotactile stimulation into clothing.

XSL•FO
RenderX

Figure 3. Demonstrator of wearable haptic physiotherapy assistant.

The authors instructed them to imagine a wearable device, similar to the one shown in Figure 3, capable of generating vibrations anywhere on the body (also on arms and legs) in response to specific situations that they felt required feedback. They should imagine a hands-free scenario in which the standardized patient uses such a device independently at home while performing their exercises. Subsequently, they were asked to indicate where and when a tactile feedback system should provide feedback while the standardized patient repeated the entire staged training session. Two video cameras recorded the staged training sessions for further analysis.

Data Analysis

We transcribed the audio recordings (interviews) with MAXQDA Plus 2022 (release 22.2.0; MAXQDA - Distribution by VERBI GmbH) using a content semantic transcription [35]. The aim of this data analysis was a systematic comparison as well as an exploration of feedback properties. To systematically compare and explore feedback properties, we conducted a content-structuring content analysis, the core of which is a coding system [36]. The main and subcategories of the coding system were developed deductively based on the ITF model [26,27] and prior empirical work [2]. For each category, we created names, content descriptions, examples, and decision rules. Using data from 2 participants (videos and audio transcripts), we performed a trial coding, adding subcategories inductively through subsumption [37]. The same coder repeated the trial coding after 4 weeks, and the agreement between these 2 codings was 87.85% (κ=0.88). A κ value of κ=0.88 is interpreted as a strong agreement in the literature [38]. Additionally, the category system was reviewed and discussed with a research team member. After this trial's coding, definitions of main and subcategories were modified, for example, if segments could not be clearly assigned to a category or if segments were assigned to different categories when coded again. The main analysis was conducted using a complete and modified category system. The entire category system, including the definitions and examples, is provided in Multimedia Appendix 2. The videos (hands-on and hands-off observations) and audio transcripts (interviews) were segmented and coded.

In the analysis of the hands-on observations, the exercise-specific motion pattern served as the first observation segment for feedback delivery and was categorized as either "feedback given" or "no feedback given." Each identified feedback was systematically coded into the following subcategories: feedback modality, timing, and content. Additionally, each haptic feedback was assigned to the feedback localization category (Figure 4). In the analysis of the hands-off observations, the behavioral sequences that were relevant were those associated with the use of a vibrotactile feedback system. Touches that presented information for using vibrotactile

XSL•FO
RenderX

feedback systems and statements that referred to using a vibrotactile feedback systems formed relevant coding segments. If a coding segment was detected, it was classified into main and subcategories (Figures 4 and 5).

Figure 4. Category system for observed feedback in hands-on context (segmented motion patterns annotated with corresponding feedback events, categorized by modality, timing, content, and localization [body areas 1 - 32]) and in hands-off context (feedback instances categorized by localization and task requirements).

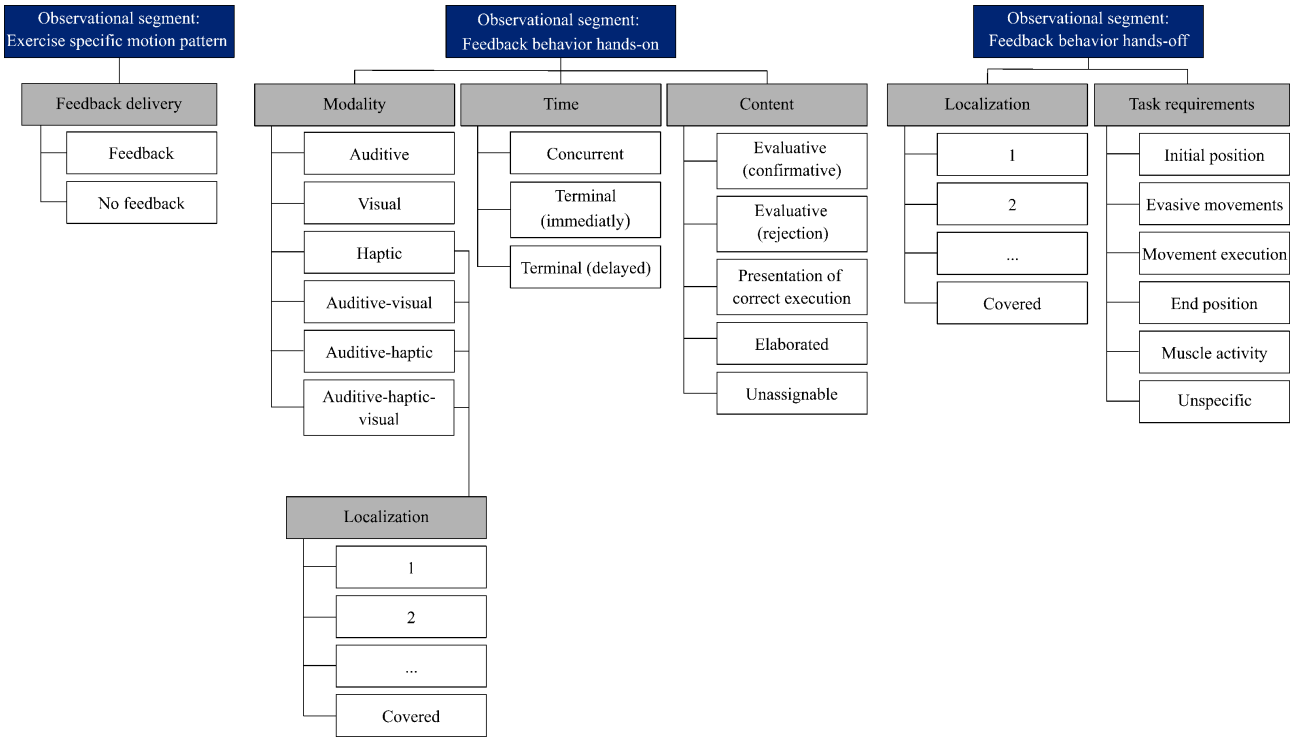

XSL•FO
RenderX

Figure 5. Category system for interview statements on hands-off feedback during staged training: categorized by modality, task requirements, localization, timing, content, frequency, function, and individual parameters. fb: feedback.

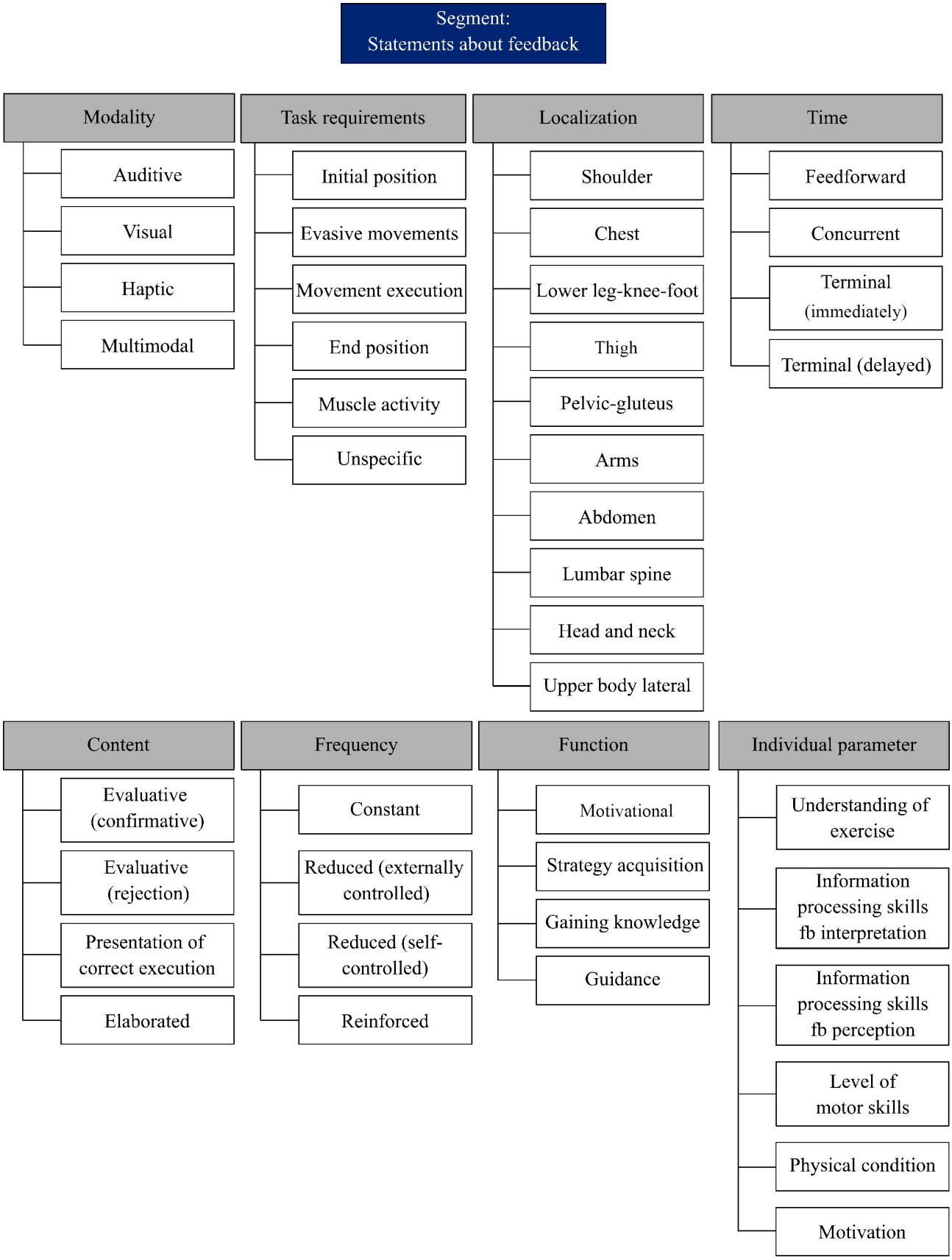

We quantified the observational data and calculated the relative frequencies of each subcategory for each participant. For the entire sample, we calculated the means and SDs of each subcategory from the participant-based relative frequencies. To provide a more comprehensive account of feedback behavior, the combination of observed properties of feedback presentation

XSL•FO
RenderX

was analyzed. We calculated the relative frequencies of subcategory combinations for each participant. We then calculated the means and SDs of these combinations for the entire sample. We thematically summarized the interview data for each subcategory on a case-by-case basis and analyzed it qualitatively. Additionally, interrelations between the subcategories were analyzed. We analyzed the data using MAXQDA Plus 2022 (release 22.2.0) and R (version 4.2.1; R Foundation).

Results

Participant Statistics

Table 2 shows the demographic and occupational characteristics of participants. This study was conducted with a total of 9 participants. Seven were trained physiotherapists, while the others were sport and rehabilitation trainers. The sports therapist and the rehabilitation trainer stated that almost every patient was treated with home exercises. Therefore, we considered them as experienced.

Table . Participant demographics, occupations, and working experience.

Participant	Education	Working experience (years), n	Treatment with home exercises[a]	Age (years), n	Gender
P1[b]	Physiotherapist	4	Almost every patient	23	Male
P2	Physiotherapist	24	For many patients	52	Male
P3	Physiotherapist	20	Almost every patient	43	Male
P4	Physiotherapist	16	Almost every patient	39	Male
P5	Sport therapist	5	Almost every patient	28	Female
P6	Physiotherapist	1	Almost every patient	26	Female
P7	Rehabilitation trainer	8	Almost every patient	46	Female
P8	Physiotherapist	2.5	Almost every patient	26	Male
P9	Physiotherapist	8	Almost every patient	32	Male

[a]Possible answers: “almost every patient,” “many patients,” “some patients,” “very few patients.”

[b]P: participant.

Feedback Content

As shown in Table 3, on average across therapists 50.2% (SD 13.4%) of observed feedback content was coded as a presentation of correct execution. Elaborated and evaluative feedback was provided in smaller but similar amounts.

Types of evaluative feedback mentioned by 8 of 9 physiotherapists in hands-off context, including nonspecific confirmatory feedback (eg, “Yes, well done”), error signaling (eg, “That’s not fully correct”), and specific confirmatory feedback linked to movement execution. The presentation of correct execution feedback addressed various task requirements, such as muscle activation, spatial movement patterns, initial positioning, movement velocity, and training persistence. The specific task request associated with a particular (haptic) signal would depend on the individual interpretation of the patients.

Four participants made statements about elaborated feedback content. A detailed description of feedback properties named by physiotherapists in hands-off interviews, grouped by theme and linked to task requirements and individual parameters is provided in Multimedia Appendix 3.

Feedback Frequency

On average, physical therapists provided feedback in 57.3% (SD 16.3%) of repetitions, while no feedback was given in 42.7% (SD 16.3%).

Four participants made statements about elaborated feedback content. A detailed description of feedback properties named by physiotherapists in hands-off interviews, grouped by theme and linked to task requirements and individual parameters is provided in Multimedia Appendix 3.

Interview data revealed unanimous support among participants for reducing feedback frequency (). Reductions were recommended both within individual training sessions and over more extended training periods. Suggested techniques included starting with less feedback and gradually increasing it to refine proprioception, prevent sensory overload, and foster acclimatization in the early stages. Feedback fading aimed to enhance intrinsic body awareness and reduce dependency on external feedback, though no consensus emerged on the optimal fading rate. Suggested initiation periods ranged from 2‑6 weeks.

While most participants supported feedback reduction, 2 highlighted concerns about declining endurance and muscle activity during training. Explicit feedback was recommended for exercise repetitions and signs of muscle fatigue. Feedback frequency was also noted to depend on physical conditions, such as illness or age. In cases of significant error patterns (eg, severe joint overload), immediate corrective feedback was deemed essential. Finally, self-controlled feedback was perceived as less effective for patients lacking intrinsic motivation to improve.

XSL•FO
RenderX

Table . Observed relative frequencies (%) of feedback properties by category (content, modality, and timing) during staged training in hands-on context (means and SDs).

Feedback propertiemean percentages (SD)s	Mean percentages (SD)
Content	
Evaluative (confirmation)	28.3 (8.4)
Evaluative (rejection)	0 (0)
Presentation correct execution	50.2 (13.4)
Elaborated	21.5 (11.4)
Unassignable	5.6 (1.1)
Modality	
Visual	0 (0)
Auditive	51.9 (12.1)
Haptic	0.8 (1.2)
Auditive or visual	16.4 (16.0)
Auditive or haptic	29.2 (24.4)
Auditive, haptic, or visual	1.9 (2.6)
Timing	
Concurrent	70.8 (13.1)
Terminal immediately	27.1 (13.6)
Terminal delayed	2.3 (4.6)

Feedback Timing

On average, across therapists, mainly concurrent feedback could be observed (70.8%, SD 13.1%; Table 3). Additionally, the interview data showed that physiotherapists would give concurrent or feedforward information.

Feedback Modality

Across therapists, unimodal auditory feedback (51.9%, SD 12.1%) or multimodal feedback with an auditory component was observed very frequently. The presentation of unimodal haptic feedback was an exception, with an average of 0.8% (SD 1.2%). The proportion of auditory or haptic feedback shows large interindividual differences (29.2%, SD 24.4%), that is, some participants used this modality very often, some did not use it at all.

Six participants emphasized that visual or auditory feedback should pertain to the initial starting position. In parallel, 4 participants highlighted the appropriateness of haptic feedback for guiding movement execution and indicating muscle activity.

Integration of Content, Modality, and Timing

On average, terminal feedback was often coded in combination with the auditory subcategory for 18.2% (SD 5.1%). Simple confirming feedback was exclusively coded with the auditory modality for 28% across therapists, and the SD was relatively low, 8.5%. Conversely, the presentation of correct execution and elaborated feedback was coded in combination with auditory, auditory or visual, and auditory or haptic modality, and relatively large SDs were detected. Evaluative confirming feedback was almost equally coded with concurrent and immediate terminal feedback subcategories. The feedback that was coded as presentation of correct execution was often coded with the subcategory concurrent for, on average, 42.9% (SD 15.4%), and the same for elaborated feedback with 14.2% (SD 9.5%).

The interview data showed that there were single statements about when and how simple evaluative feedback should be presented. Unspecific confirmation was suggested as concurrent or terminal-delayed vibrotactile cues, while error signaling could be conveyed concurrently via vibrotactile, auditory, or visual channels. Correct execution feedback was exclusively described as concurrent and delivered through haptic, auditory, visual, or multimodal formats. Regarding elaborated feedback, comments were sparse. One participant recommended terminal-delayed feedback for addressing common errors, such as: “Yes, you could also show at the end that you had 90% of 100% correct posture. Or pay more attention to your abdominal tension in the next session because it needs a lot of stimuli now.” Additionally, multimodal concurrent or visual concurrent relevant hints could be presented about task-specific muscle activity (illustrated, for instance, by color-coded body regions).

Feedback Localization

Figure 6 shows the means of the event-based relative frequencies (%) of individual body regions, both for the hands-on and the hands-off observation data.

XSL•FO
RenderX

Figure 6. Mean relative frequencies (%) for body locations in hands-on and hands-off contexts during staged training, based on observed therapist-patient interactions; error bars indicate SD; values above the median (dotted line) are shown as colored areas on body diagrams. rel.: relative.

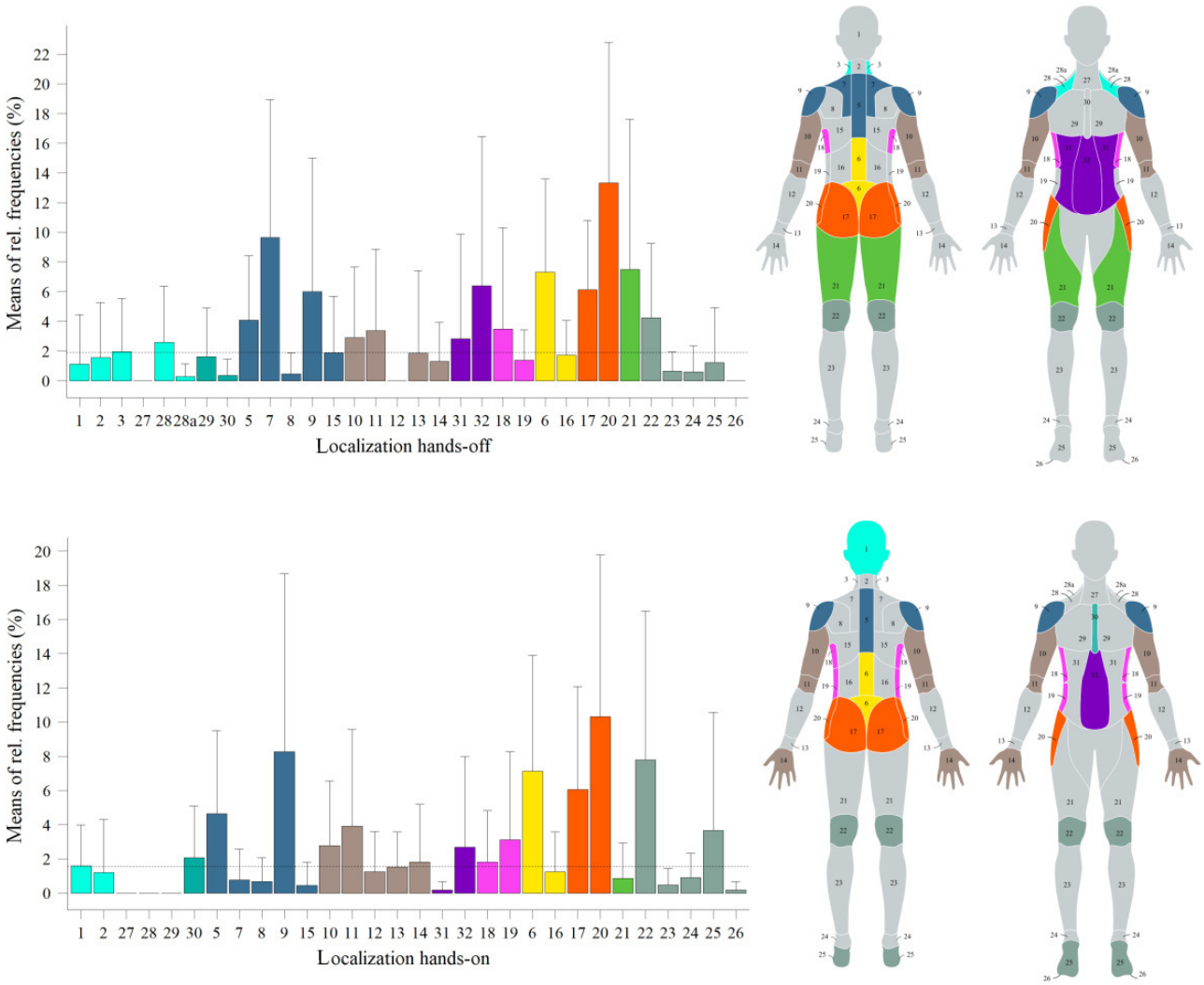

Feedback on muscle activity was frequently associated with the abdomen and upper body, while movement execution feedback was linked to the shoulder, pelvis, gluteus, and arms. Starting position feedback was often related to the head, shoulders, thighs, lower legs, knees, feet, pelvis, gluteus, and lumbar spine. Evasive movements were predominantly associated with the pelvis, gluteus, and shoulders.

The interview data showed that half of the participants advocated providing vibrotactile feedback in specific anatomical regions, including the upper back or shoulder and lower back areas. An insightful suggestion emerged from 1 participant who underscored the necessity for individually tailored actuator settings in these regions, considering physical conditions such as kyphosis or lordosis tendencies and individual sensitivity to tactile stimuli in the scapula region. Additionally, 4 participants endorsed the application of vibrotactile cues to the thigh and gluteus or pelvis region. Opinions diverged among participants concerning feedback localization for the knee, leg, or foot, upper extremities, and neck or head regions. Notably, 2 participants opposed tactile feedback in the neck or head area.

Four of seven participants recommended the adaptability of localization to individual needs, enabling flexible actuator configurations based on physical attributes or personalized training progress. Three participants highlighted conceptual differentiation between static and mobile muscles during exercise execution, prompting the suggestion for diverse feedback forms (eg, variation in frequency) to accommodate these distinctions.

Feedback on starting position was associated with the upper extremities by 4 participants and with the lower extremities by 2. Muscle activity feedback was linked to the abdomen (3 participants), gluteus or pelvis (2 participants), lower back, and thighs.

Discussion

Foundations for Vibrotactile Feedback in Home Exercise Contexts

This paper aimed to derive requirements for a vibrotactile feedback strategy for home exercises. We gained initial insights into the provision of feedback and recommendations for a feedback strategy. Hypothesis-driven empirical studies based on initial approaches should follow. A strength of this work is the integration of a design-driven methodology with insights from both empirical and theoretical domains of instructional psychology.

XSL•FO
RenderX

Feedback Content

The presentation of correct execution was frequently observed, and somewhat less regularly simple confirmatory feedback or elaborated feedback. Few participants recommended the presentation of elaborated information, and some recommended the use of presentation of correct results concerning different task requirements. For the design of the feedback strategy, deciding which task requirements should be covered in terms of content (eg, muscle activity, spatial movement accuracy, and velocity) seems relevant. The literature presents mixed results regarding the effectiveness of informational feedback, confirmatory feedback, or a combination of both in facilitating motor learning [39]. What content is effective depends on moderating factors such as task difficulty and skill level. The informational property of feedback becomes more important as task difficulty increases [40]. Especially for people with low skill levels, the presentation of information about solutions and ways of solving problems, as well as elaborated feedback, seems to be effective [41,42].

To design a feedback strategy for home exercising, it seems worthwhile to analyze the effects of feedback content on different learning states. Whether pain patients could benefit from elaborated feedback should be determined in further studies. Therefore, interaction effects should be considered. If, for example, elaborated feedback is combined directly with the presentation of the target performance, elaborated information has no additional effect [43].

The qualitative data can also identify reasons against the use of elaborated feedback. The participants mainly mentioned structural aspects such as the limited time frame of a treatment or the patients' goal and motivation to become pain-free quickly, without wanting to achieve long-term learning effects. Further studies should examine the user acceptance and usability of the feedback strategy.

Feedback Frequency

Participants recommended and showed reduced feedback frequency. They did not provide feedback for every movement repetition and error. This aligns with previous research linking improved motor performance with decreased feedback frequency during training and retention [44-47]. Conflicting results show no effects of reduced feedback as well as effects of dependency on task complexity, learning phase, and individual learner characteristics [2,45,46,48-50]. Physical therapists aimed to reduce system dependency and train proprioception. This also matches previous literature, which postulates that by reducing the frequency of feedback, the dependence on extrinsic feedback is reduced, and thus, intrinsic sources of feedback can be used [2,49]. Participants discussed reducing feedback frequency within a training session and over a more extended training period, suggesting the adaptability of the feedback strategy to both time frames. Externally controlled feedback was recommended, while self-controlled feedback was neither used nor suggested. The empirical literature does not clearly indicate a preferred technique for reducing feedback frequency [48,51].

Individual factors, such as susceptibility to errors, patient age, and physical condition, were identified as potential influences on feedback reduction. These recommendations are partly in line with previous research literature. For example, the optimal reduction in feedback frequency is related to the feedback recipient's motor skill level and is influenced by perceived task difficulty [52]. For the design of a feedback strategy, adapting the feedback frequency to these parameters seems relevant.

Feedback Timing

Results show that physical therapists provided frequently and unanimously recommended concurrent feedback. Delivering concurrent feedback can hinder learning, but it can support the performance of motor skills [17,48]. To apply a feedback strategy, feedback timing should be connected to its goal: performance enhancement or motor learning.

Additionally, the level of motor skills should be considered. Presenting concurrent feedback is initially beneficial for people with low motor skill levels [53]. So, there is a need to analyze the skill levels of patients.

Furthermore, physical therapists recommended the presentation of feedforward information. They suggested presenting information related to muscle activity or the starting position.

Feedback Modality

Physical therapists frequently use auditory or multimodal feedback with an auditory component. This creates the need to translate or integrate verbal information into vibrotactile feedback. To translate verbal information, unimodal vibrotactile feedback can be presented with solid links to verbal information, for example, to movement directions in different planes. Literature describes this as a push or pull mechanism, moving the body away from the vibration or following the vibration [18,48]. Previous research shows individual preferences for intelligibility and the use of these mechanisms [18,54,55]. By activating multiple actuators, spatiotemporally bound vibration patterns can instruct movements [56].

Multimodal feedback can be used to integrate verbal information. This can cover the entire usage behavior that we found in our results (eg, some participants used certain modalities very often or very little). Literature also shows the benefits of providing multimodal feedback on motor performance and learning [15,57,58]. Our results show that visual or auditory information was connected to various task requirements. Six participants connected them to the correction of the initial position. Four participants connected tactile cues with spatial movement and muscle activity. To design a feedback strategy, it seems useful to conduct studies that investigate the advantages of unimodal auditory versus haptic versus multimodal feedback about initial position, spatial movements, and muscle activity.

Integration of Content, Modality, and Timing

It was found that simple confirmatory feedback was mainly presented auditorily, concurrently, or terminally to a movement execution. The presentation of correct execution, as well as elaborated feedback, was mainly presented concurrently, in some cases terminally. To further develop the feedback strategy, we can define a type of feedback: confirmatory auditory feedback. It is then considered and needs to be studied whether

XSL•FO
RenderX

this type of feedback differs from the presentation of correct execution and elaborated feedback.

Feedback Localization

Feedback in the abdominal area was associated with muscle activity. Spatial motion accuracy was often coded with the pelvis, shoulder, or arm area. The starting position was frequently coded with head, shoulder, thigh, lumbar spine, pelvis, and knee. Thus, we see a tendency for feedback on muscle activity and spatial motion accuracy relating more to specific body parts than feedback on the starting position. A request for the feedback strategy could be to make vibrotactile cues distinguishable between muscle activity and spatial movements regarding the mentioned body areas.

Both in the hands-off context and from the interview data, haptic stimuli appear less set or intended on the forearm, hands, lower legs, and feet. Further studies should investigate whether auditory or visual information is sufficient to correct the positions of extremities.

Limitations

This study has limitations that must be considered. First, the use of a healthy individual to demonstrate home exercises may limit the generalizability of our findings to patients with back pain, as patients may exhibit different movement patterns and errors [59,60]. We selected this approach to ensure exercise comparability and to guarantee the occurrence of comparable errors during the performance of the exercises. Having gained an overview of potential feedback behaviors, engaging in hypothesis-driven research on patient samples is a necessity. Additionally, the feedback behavior observed in this study may have been influenced by the presence of the experimenter and the use of cameras, which could have led to a socially desired behavior. Participants may have altered their actions to appear as correct as possible, potentially distorting the natural use of feedback. In some cases, the participants expressed that they were under pressure to act. Future studies should focus on real-world practice settings to minimize such biases.

Our study primarily focused on spatial accuracy, specifically whether participants' movement paths aligned with the prescribed exercises. However, temporal factors [61], such as the number of repetitions, sets, and movement speed, were not considered in this study design. As a result, feedback related to these factors was rarely provided in the hands-on context. The interviews conducted in the hands-off context may not fully reflect physical therapists' consideration of temporal factors, suggesting the need for future research that includes both spatial and temporal aspects in feedback strategies. Numerous factors influence whether home exercises are performed correctly, for example, memory capacity, motivation, attitudes, volition, or training conviction of the patients [61]. For designing an effective feedback strategy, the function of feedback should be differentiated. Further research is needed to determine how feedback can support the prescribed execution of home exercises.

Practical Implementation

First, it should be noted that a feedback strategy should be flexible and adaptable to the patient. The present work suggests that a feedback system should consider the baseline of motor skill expression (eg, in the form of error frequency). The error thresholds should be set according to this baseline; thus, the feedback frequency can be influenced. Ideally, a feedback system can measure learning progress and adjust feedback frequency. In terms of feedback content, a language needs to be developed that patients can understand, both in terms of tactile presentation of the correct execution of exercise and in terms of confirmation of goal-directed action. Whether the presentation of elaborated information can increase the effectiveness of the feedback strategy needs to be investigated in subsequent studies. Although the focus of this paper is on a vibrotactile feedback strategy, the addition of auditory cues or visual information seems advisable. Suitable interfaces for the presentation of auditory and visual signals should be developed. Concurrent presentation of feedback is recommended. This feedback is presented vibrotactilely on muscle activity or spatial motion accuracy. For this purpose, actuators should be positioned primarily in the regions of the gluteus, pelvis, spine, deltoid, trapezius, abdominals, lateral to the trunk, thigh, and knee. Another implication for practice is the possibility of providing feedforward information on task-relevant musculature and the starting position.

Conclusions

This work highlights that reduced feedback frequency, differentiated feedback content, multimodal feedback, and trunk-proximal vibrotactile feedback were recommended by physical therapists for a technical feedback system for the correct performance of home exercises. The results of this study provide an overview of relevant feedback characteristics and can serve as a starting point for further empirical research. Hypothesis-driven studies in sufficiently large patient samples should be followed up for the further development of a feedback strategy. Research-relevant topics include the perception, interpretation, and reaction of patients to vibrotactile feedback, as well as the integration of a technical-economic perspective regarding cost factors and hardware and software solutions.

Acknowledgments

This study was funded by the German Research Foundation (DFG, Deutsche Forschungsgemeinschaft) as part of Germany's Excellence Strategy—EXC 2050/1—Project ID 390696704—Cluster of Excellence "Centre for Tactile Internet with Human-in-the-Loop" (CeTI) of Technische Universität Dresden.

XSL•FO
RenderX

Data Availability

The datasets generated or analyzed during this study are not publicly available due to data privacy, as the audio and video recordings cannot be fully anonymized, but they are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Description of exercise execution.
[PDF File, 1172 KB - formative_v9i1e62903_app1.pdf]

Multimedia Appendix 2
Descriptions of category systems (hands-on and hands-off).
[DOC File, 825 KB - formative_v9i1e62903_app2.doc]

Multimedia Appendix 3
Detailed results of qualitative interview data hands-off.
[DOCX File, 22 KB - formative_v9i1e62903_app3.docx]

Checklist 1
A 32-item COREQ checklist for qualitative interviews. COREQ: Consolidated Criteria for Reporting Qualitative Research.
[DOCX File, 179 KB - formative_v9i1e62903_app4.docx]

References

1. Hartveld A, Hegarty JR. Augmented feedback and physiotherapy practice. Physiotherapy 1996 Aug;82(8):480-490. [doi: 10.1016/S0031-9406(05)66414-0]
2. Ribeiro DC, Sole G, Abbott JH, Milosavljevic S. Extrinsic feedback and management of low back pain: a critical review of the literature. Man Ther 2011 Jun;16(3):231-239. [doi: 10.1016/j.math.2010.12.001] [Medline: 21269869]
3. Budde K. KI für Gesundheitsfachkräfte. Chancen und Herausforderungen von Medizinischen und Pflegerischen Ki-Anwendungen [Report in German]. : Whitepaper München: Plattform Lernende Systeme; 2023 URL: https://www.acatech.de/publikation/ki-fuer-gesundheitsfachkraefte/ [accessed 2025-05-29] [doi: 10.48669/pls_2023-2]
4. Orlowski K, Eckardt F, Nusser JE, Witte K. Feedback system for physiotherapy and popular athletes. Int J Physiother 2018;5(6):194-201. [doi: 10.15621/ijphy/2018/v5i6/178057]
5. Burghoorn AW, Dhaeze ER, Faber JS, Wories JWH, Feijs LMG, Timmermans AAA. Mirrorcle: enhanced visual feedback to support motor learning in low back pain. Presented at: REHAB '15: Proceedings of the 3rd 2015 Workshop on ICTs for improving Patients Rehabilitation Research Techniques; Oct 1-2, 2015; Lisbon, Portugal. [doi: 10.1145/2838944.2838951]
6. Ranasinghe I, Dantu R, Watts S, Ocana R. Cyber-physiotherapy: rehabilitation to training. Presented at: Proceedings of the IM 2021 - 2021 IFIP/IEEE International Symposium on Integrated Network Management; May 17-21, 2021; Bordeaux, France.
7. Tang R, Yang XD, Bateman S, Jorge J, Tang A. Physio@Home: exploring visual guidance and feedback techniques for physiotherapy exercises. Presented at: CHI '15: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems; Apr 18-23, 2015; Seoul, Republic of Korea. [doi: 10.1145/2702123.2702401]
8. Hamilton C, Lovarini M, van den Berg M, McCluskey A, Hassett L. Usability of affordable feedback-based technologies to improve mobility and physical activity in rehabilitation: a mixed methods study. Disabil Rehabil 2022 Jul;44(15):4029-4038. [doi: 10.1080/09638288.2021.1884904] [Medline: 33645384]
9. Lambert TE, Harvey LA, Avdalis C, et al. An app with remote support achieves better adherence to home exercise programs than paper handouts in people with musculoskeletal conditions: a randomised trial. J Physiother 2017 Jul;63(3):161-167. [doi: 10.1016/j.jphys.2017.05.015] [Medline: 28662834]
10. Toelle TR, Utpadel-Fischler DA, Haas KK, Priebe JA. App-based multidisciplinary back pain treatment versus combined physiotherapy plus online education: a randomized controlled trial. NPJ Digit Med 2019;2(1):34. [doi: 10.1038/s41746-019-0109-x] [Medline: 31304380]
11. Ananthanarayan S, Sheh M, Chien A, Profita H, Siek K. PT viz: towards a wearable device for visualizing knee rehabilitation exercises. Presented at: Conference on Human Factors in Computing Systems – Proceedings; Apr 27 to May 2, 2013; Paris, France. [doi: 10.1145/2470654.2466161]
12. Hügli AS, Ernst MJ, Kool J, et al. Adherence to home exercises in non-specific low back pain. A randomised controlled pilot trial. J Bodyw Mov Ther 2015 Jan;19(1):177-185. [doi: 10.1016/j.jbmt.2014.11.017] [Medline: 25603757]

XSL•FO
RenderX

13. Meinke A, Peters R, Knols R, Karlen W, Swanenburg J. Exergaming using postural feedback from wearable sensors and exercise therapy to improve postural balance in people with nonspecific low back pain: protocol for a factorial pilot randomized controlled trial. JMIR Res Protoc 2021 Aug 26;10(8):e26982. [doi: 10.2196/26982] [Medline: 34435954]
14. Lee BC, Chen S, Sienko KH. A wearable device for real-time motion error detection and vibrotactile instructional cuing. IEEE Trans Neural Syst Rehabil Eng 2011 Aug;19(4):374-381. [doi: 10.1109/TNSRE.2011.2140331] [Medline: 21511568]
15. Lieberman J, Breazeal C. Development of a wearable vibrotactile feedback suit for accelerated human motor learning. Presented at: 2007 IEEE International Conference on Robotics and Automation; Apr 10-14, 2007; Rome, Italy. [doi: 10.1109/ROBOT.2007.364093]
16. Cuppone AV, Squeri V, Semprini M, Masia L, Konczak J. Robot-assisted proprioceptive training with added vibro-tactile feedback enhances somatosensory and motor performance. PLoS ONE 2016;11(10):e0164511. [doi: 10.1371/journal.pone.0164511] [Medline: 27727321]
17. Sigrist R, Rauter G, Riener R, Wolf P. Terminal feedback outperforms concurrent visual, auditory, and haptic feedback in learning a complex rowing-type task. J Mot Behav 2013;45(6):455-472. [doi: 10.1080/00222895.2013.826169] [Medline: 24006910]
18. Spelmezan D, Jacobs M, Hilgers A, Borchers J. Tactile motion instructions for physical activities. Presented at: Conference on Human Factors in Computing Systems; Apr 4-9, 2009; Boston, MA, United States. [doi: 10.1145/1518701.1519044]
19. van Erp JBF. Vibrotactile spatial acuity on the torso: effects of location and timing parameters. Presented at: First Joint Eurohaptics Conference and Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems; Mar 18-20, 2005; Pisa, Italy. [doi: 10.1109/WHC.2005.144]
20. Moesgen T, Salovaara A, Pouta E, Pyykko R, Xiao Y. Vibrotactile motion guidance for stroke rehabilitation: a comparative study. Presented at: ICMHI 2022; May 13-15, 2022. [doi: 10.1145/3545729.3545759]
21. Bao T, Klatt BN, Carender WJ, et al. Effects of long-term vestibular rehabilitation therapy with vibrotactile sensory augmentation for people with unilateral vestibular disorders – a randomized preliminary study. J Vestib Res 2020 Jan 2;29(6):323-334. [doi: 10.3233/VES-190683]
22. Bao T, Carender WJ, Kinnaird C, et al. Effects of long-term balance training with vibrotactile sensory augmentation among community-dwelling healthy older adults: a randomized preliminary study. J Neuroeng Rehabil 2018 Jan 18;15(1):5. [doi: 10.1186/s12984-017-0339-6] [Medline: 29347946]
23. Du Q, Luo J, Cheng Q, Wang Y, Guo S. Vibrotactile enhancement in hand rehabilitation has a reinforcing effect on sensorimotor brain activities. Front Neurosci 2022;16:935827. [doi: 10.3389/fnins.2022.935827] [Medline: 36267238]
24. Brennan L, Zubiete ED, Caulfield B. Feedback design in targeted exercise digital biofeedback systems for home rehabilitation: a scoping review. Sensors (Basel) 2019 Dec 28;20(1):181. [doi: 10.3390/s20010181] [Medline: 31905653]
25. Handelzalts S, Ballardini G, Avraham C, Pagano M, Casadio M, Nisky I. Integrating tactile feedback technologies into home-based telerehabilitation: opportunities and challenges in light of COVID-19 pandemic. Front Neurorobot 2021;15:617636. [doi: 10.3389/fnbot.2021.617636] [Medline: 33679364]
26. Narciss S. Designing and evaluating tutoring feedback strategies for digital learning environments on the basis of the interactive tutoring feedback model. Digital Education Review 2013;23(1):7-26. [doi: 10.1344/der.2013.23.7-26]
27. Narciss S. Feedback strategies. In: Seel NM, editor. Encyclopedia of the Sciences of Learning: Springer; 2012:1289-1293.
28. Nicol D. The power of internal feedback: exploiting natural comparison processes. Assess Eval Higher Educ 2021 Jul 4;46(5):756-778. [doi: 10.1080/02602938.2020.1823314]
29. Bobbe T, Oppici L, Lüneburg LM, et al. What early user involvement could look like—developing technology applications for piano teaching and learning. MTI 2021;5(7):38. [doi: 10.3390/mti5070038]
30. van der Linden J, Waights V, Rogers Y, Taylor C. A blended design approach for pervasive healthcare: bringing together users, experts and technology. Health Informatics J 2012 Sep;18(3):212-218. [doi: 10.1177/1460458212442934] [Medline: 23011816]
31. Ergonomics of human-system interaction – human-centred design for interactive systems (ISO standard no 9241- 210:2019). International Organization for Standardization.: ISO; 2019. URL: https://www.iso.org/standard/77520.html [accessed 2025-05-29]
32. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014 Sep;89(9):1245-1251. [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]
33. Tong A, Sainsbury P, Craig J. Consolidated Criteria for Reporting Qualitative Research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007;19(6):349-357. [doi: 10.1093/intqhc/mzm042]
34. Niederer D, Vogt L, Wippert PM, et al. Medicine in spine exercise (MiSpEx) for nonspecific low back pain patients: study protocol for a multicentre, single-blind randomized controlled trial. Trials 2016 Oct 20;17(1):507. [doi: 10.1186/s13063-016-1645-1] [Medline: 27765058]
35. Dresing T, Pehl T. Praxisbuch Interview, Transkription & Analyse: Anleitungen Und Regelsysteme Für Qualitativ Forschende [Book in German], 8th edition: Eigenverlag; 2018.
36. Kuckartz U. Computergestützte analyse qualitativer daten [Book in German]. In: Buber R, Holzmüller HH, editors. Qualitative Marktforschung, 2nd edition: Gabler; 2009:713-730. [doi: 10.1007/978-3-8349-9441-7_45]

XSL•FO
RenderX

37. Schreier M. Qualitative content analysis. In: Flick U, editor. The SAGE Handbbok of Qualitative Data Analysis: SAGE Publications; 2014:170-183. [doi: 10.4135/9781446282243.n12]
38. Rössler P. Inhaltsanalyse [Book in German], 3rd edition: UVK Verlagsgesellschaft; 2017.
39. Zhou Y, Shao WD, Wang L. Effects of feedback on students' motor skill learning in physical education: a systematic review. IJERPH 2021;18(12):6281. [doi: 10.3390/ijerph18126281]
40. Bugnon M, Wälchli M, Taube W. How to benefit from augmented feedback? The influence of motivational and informational content of augmented feedback and the influence of task complexity. Eur J Sport Sci 2023 Jul;23(7):1435-1445. [doi: 10.1080/17461391.2023.2178975] [Medline: 36799077]
41. Oppici L, Dix A, Narciss S. When is knowledge of performance (KP) superior to knowledge of results (KR) in promoting motor skill learning? A systematic review. Int Rev Sport Exerc Psychol 2024 Dec 31;17(1):182-207. [doi: 10.1080/1750984X.2021.1986849]
42. Heckler AF, Mikula BD. Factors affecting learning of vector math from computer-based practice: feedback complexity and prior knowledge. Phys Rev Phys Educ Res 2016 Jun;12(1):010134. [doi: 10.1103/PhysRevPhysEducRes.12.010134]
43. Narciss S. Informatives Tutorielles Feedback: Entwicklungs- und Evaluationsprinzipien auf der Basis instruktionspsychologischer Erkenntnisse [Book in German]: Waxmann; 2006.
44. Herbert WJ, Heiss DG, Basso DM. Influence of feedback schedule in motor performance and learning of a lumbar multifidus muscle task using rehabilitative ultrasound imaging: a randomized clinical trial. Phys Ther 2008 Feb;88(2):261-269. [doi: 10.2522/ptj.20060308] [Medline: 18042655]
45. Winstein CJ, Schmidt RA. Reduced frequency of knowledge of results enhances motor skill learning. J Exp Psychol Learn Mem Cogn 1990;16(4):677-691. [doi: 10.1037//0278-7393.16.4.677]
46. Wulf G, Schmidt RA, Deubel H. Reduced feedback frequency enhances generalized motor program learning but not parameterization learning. J Exp Psychol Learn Mem Cogn 1993 Sep;19(5):1134-1150. [doi: 10.1037//0278-7393.19.5.1134] [Medline: 8409852]
47. Hebert EP, Coker C. Optimizing feedback frequency in motor learning: self-controlled and moderate frequency KR enhance skill acquisition. Percept Mot Skills 2021 Oct;128(5):2381-2397. [doi: 10.1177/00315125211036413] [Medline: 34338053]
48. Sigrist R, Rauter G, Riener R, Wolf P. Augmented visual, auditory, haptic, and multimodal feedback in motor learning: a review. Psychon Bull Rev 2013 Feb;20(1):21-53. [doi: 10.3758/s13423-012-0333-8] [Medline: 23132605]
49. Park JH, Shea CH, Wright DL. Reduced-frequency concurrent and terminal feedback: a test of the guidance hypothesis. J Mot Behav 2000 Sep;32(3):287-296. [doi: 10.1080/00222890009601379] [Medline: 10975276]
50. McKay B, Hussien J, Vinh MA, Mir-Orefice A, Brooks H, Ste-Marie DM. Meta-analysis of the reduced relative feedback frequency effect on motor learning and performance. Psychol Sport Exerc 2022 Jul;61:102165. [doi: 10.1016/j.psychsport.2022.102165]
51. Goodwin JE. Scheduling concurrent visual feedback in learning a continuous balance task. J Mot Learn Dev 2019;7(2):261-272. [doi: 10.1123/jmld.2017-0067]
52. Guadagnoli MA, Lee TD. Challenge point: a framework for conceptualizing the effects of various practice conditions in motor learning. J Mot Behav 2004 Jul;36(2):212-224. [doi: 10.3200/JMBR.36.2.212-224]
53. Marchal-Crespo L, van Raai M, Rauter G, Wolf P, Riener R. The effect of haptic guidance and visual feedback on learning a complex tennis task. Exp Brain Res 2013 Nov;231(3):277-291. [doi: 10.1007/s00221-013-3690-2] [Medline: 24013789]
54. Tannert I, Schulleri KH, Michel Y, et al. Immediate effects of vibrotactile biofeedback instructions on human postural control. Presented at: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); Nov 1-5, 2021; Mexico. [doi: 10.1109/EMBC46164.2021.9630992]
55. Jansen C, Oving A, Veen HJ. Vibrotactile movement initiation. Presented at: Proceedings of the 4th International Conference EuroHaptics; Jun 5-7, 2004; Munich, Germany.
56. Rosenthal J, Edwards N, Villanueva D, Krishna S, McDaniel T, Panchanathan S. Design, implementation, and case study of a pragmatic vibrotactile belt. IEEE Trans Instrum Meas 2011;60(1):114-125. [doi: 10.1109/TIM.2010.2065830]
57. Bark K, Hyman E, Tan F, et al. Effects of vibrotactile feedback on human learning of arm motions. IEEE Trans Neural Syst Rehabil Eng 2015 Jan;23(1):51-63. [doi: 10.1109/TNSRE.2014.2327229] [Medline: 25486644]
58. Frikha M, Chaâri N, Elghoul Y, Mohamed-Ali HH, Zinkovsky AV. Effects of combined versus singular verbal or haptic feedback on acquisition, retention, difficulty, and competence perceptions in motor learning. Percept Mot Skills 2019 Aug;126(4):713-732. [doi: 10.1177/0031512519842759] [Medline: 31033405]
59. Luomajoki H, Kool J, de Bruin ED, Airaksinen O. Movement control tests of the low back; evaluation of the difference between patients with low back pain and healthy controls. BMC Musculoskelet Disord 2008 Dec 24;9:1-12. [doi: 10.1186/1471-2474-9-170] [Medline: 19108735]
60. Descarreaux M, Blouin JS, Teasdale N. Repositioning accuracy and movement parameters in low back pain subjects and healthy control subjects. Eur Spine J 2005 Mar;14(2):185-191. [doi: 10.1007/s00586-004-0833-y] [Medline: 15759173]
61. Kowatsch T, Lohse KM, Erb V, et al. Hybrid ubiquitous coaching with a novel combination of mobile and holographic conversational agents targeting adherence to home exercises: four design and evaluation studies. J Med Internet Res 2021 Feb 22;23(2):e23612. [doi: 10.2196/23612] [Medline: 33461957]

XSL•FO
RenderX

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Research
ITF: interactive tutoring feedback

Edited by A Mavragani; submitted 07.06.24; peer-reviewed by B Luciani, B Carpallo, P Cordova; revised version received 09.05.25; accepted 12.05.25; published 08.07.25.

Please cite as:
Holzmeyer K, Lüneburg LM, Oppici L, Flößel P, Lachmann D, Krzywinski J, Narciss S
Vibrotactile Feedback Strategies for Trunk-Stabilizing Exercises in a Home-Based Scenario: Qualitative Interview Study Among Physiotherapists
JMIR Form Res 2025;9:e62903
URL: https://formative.jmir.org/2025/1/e62903
doi:10.2196/62903

© Kaya Holzmeyer, Lisa-Marie Lüneburg, Luca Oppici, Philipp Flößel, Doris Lachmann, Jens Krzywinski, Susanne Narciss. Originally published in JMIR Formative Research (https://formative.jmir.org), 8.7.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Assessment of Recommendations Provided to Athletes Regarding Sleep Education by GPT-4o and Google Gemini: Comparative Evaluation Study

Lukas Masur[1], MSci; Matthew Driller[2], PhD; Haresh Suppiah[2], PhD; Manuel Matzka[3], PhD; Billy Sperlich[3], PhD; Peter Düking[1], PhD

[1]Department of Sports Science and Movement Pedagogy, Technische Universität Braunschweig, Pockelsstraße 11, Braunschweig, Germany
[2]School of Allied Health, Human Services, and Sport, La Trobe University, Melbourne, Australia
[3]Department of Sports Science, Integrative and Experimental Exercise Science & Training, University of Würzburg, Würzburg, Germany

Corresponding Author:
Peter Düking, PhD
Department of Sports Science and Movement Pedagogy, Technische Universität Braunschweig, Pockelsstraße 11, Braunschweig, Germany

Abstract

Background: Inadequate sleep is prevalent among athletes, affecting adaptation to training and performance. While education on factors influencing sleep can improve sleep behaviors, large language models (LLMs) may offer a scalable approach to provide sleep education to athletes.

Objective: This study aims (1) to investigate the quality of sleep recommendations generated by publicly available LLMs, as evaluated by experienced raters, and (2) to determine whether evaluation results vary with information input granularity.

Methods: Two prompts with differing information input granularity (low and high) were created for 2 use cases and inserted into ChatGPT-4o (GPT-4o) and Google Gemini, resulting in 8 different recommendations. Experienced raters (n=13) evaluated the recommendations on a 1 - 5 Likert scale, based on 10 sleep criteria derived from recent literature. A Friedman test with Bonferroni correction was performed to test for significant differences in all rated items between the training plans. Significance level was set to $P<.05$. Fleiss κ was calculated to assess interrater reliability.

Results: The overall interrater reliability using Fleiss κ indicated a fair agreement of 0.280 (range between 0.183 and 0.296). The highest summary rating was achieved by GPT-4o using high input information granularity, with 8 ratings >3 (tendency toward good), 3 ratings equal to 3 (neutral), and 2 ratings <3 (tendency toward bad). GPT-4o outperformed Google Gemini in 9 of 10 criteria ($P<.001$ to $P=.04$). Recommendations generated with high input granularity received significantly higher ratings than those with low granularity across both LLMs and use cases ($P<.001$ to $P=.049$). High input granularity leads to significantly higher ratings in items pertaining to the used scientific sources ($P<.001$), irrespective of the analyzed LLM.

Conclusions: Both LLMs exhibit limitations, neglecting vital criteria of sleep education. Sleep recommendations by GPT-4o and Google Gemini were evaluated as suboptimal, with GPT-4o achieving higher overall ratings. However, both LLMs demonstrated improved recommendations with higher information input granularity, emphasizing the need for specificity and a thorough review of outputs to securely implement artificial intelligence technologies into sleep education.

(JMIR Form Res 2025;9:e71358) doi:10.2196/71358

KEYWORDS

individualization; personalization; artificial intelligence; education; recovery; monitoring

Introduction

Sleep is essential for the health and well-being of individuals across all age groups, as it supports cognitive function, mood, mental health, as well as cardiovascular, cerebrovascular, and metabolic health [1,2]. Short-term sleep deprivation, chronic sleep restriction, circadian misalignment, and untreated sleep disorders can significantly harm physical health, mental health, and mood [1,3,4].

One population frequently experiencing inadequate sleep or poor sleep quality is athletes [5-7]. Reasons for poor sleep may stem from multiple sport and nonsport factors including high training loads, long-haul travel, early morning training, family commitments, lifestyle choices including diet, or work or study commitments [6]. While it is beyond the scope of this paper to dive into the effects of inadequate sleep in detail, we refer the reader to existing papers on this topic [6,8]. Briefly, athletes' sleep is considered a primary mechanism facilitating both

XSL•FO
RenderX

psychological and physiological recovery [9,10]. Poor sleep may be detrimental for athletes, with negative impacts on mental well-being, cognition, learning and memory consolidation, growth and repair of cells, glucose metabolism, and immune responses (eg, the resistance to respiratory infection) [6,11,12].

To improve aspects of sleep, a first step is to educate athletes and staff on the negative effects of poor sleep and factors affecting sleep [6], and there is evidence that sleep education improves the sleep behavior of team sport athletes [13]. Such sleep education might include (1) education of athletes on potential factors negatively impacting sleep [14] and (2) education on reducing the impact of factors negatively impacting sleep.

It was reported that even in elite athlete cohorts, there is a lack of sleep knowledge [15], and consequently, there is a need for sleep education. For this, artificial intelligence and, more specifically, publicly available large language models (LLMs) might offer a scalable solution. By simulating human-like conversations, LLMs leverage deep learning to process information and generate nuanced responses. LLMs such as ChatGPT (OpenAI) are rapidly gaining popularity among the general population [16] as well as in various scientific domains such as medical research and education [17-20], health care [21-23], or nutrition [24,25]. Thereby, it is likely that individuals turn to LLMs to receive responses to questions they face, for example, regarding sleep. However, it is currently unknown if recommendations regarding sleep generated by publicly available LLMs are appropriate and in line with recent scientific evidence and thus suitable for a specific athlete. Here, we aim (1) to investigate sleep recommendations provided to athletes generated by different publicly available LLMs as evaluated by experienced raters and (2) to investigate if sleep recommendations differ depending on the provided information by the user.

Methods

General Design

To evaluate sleep recommendations provided by LLMs, we followed methodologies of similar papers in the medical field [26-29] or in the exercise science literature [30-33] and adjusted these methodologies to the aim of our research. Figure 1 depicts the experimental workflow of the study.

For this, we (1) define a specific use case, (2) define criteria of relevance for sleep education in this use case, (3) define information input into publicly available LLMs, and (4) involve experienced raters in the topic of sleep within the athletic population to rate outcomes of LLM-generated responses.

Figure 1. Experimental workflow of the study.

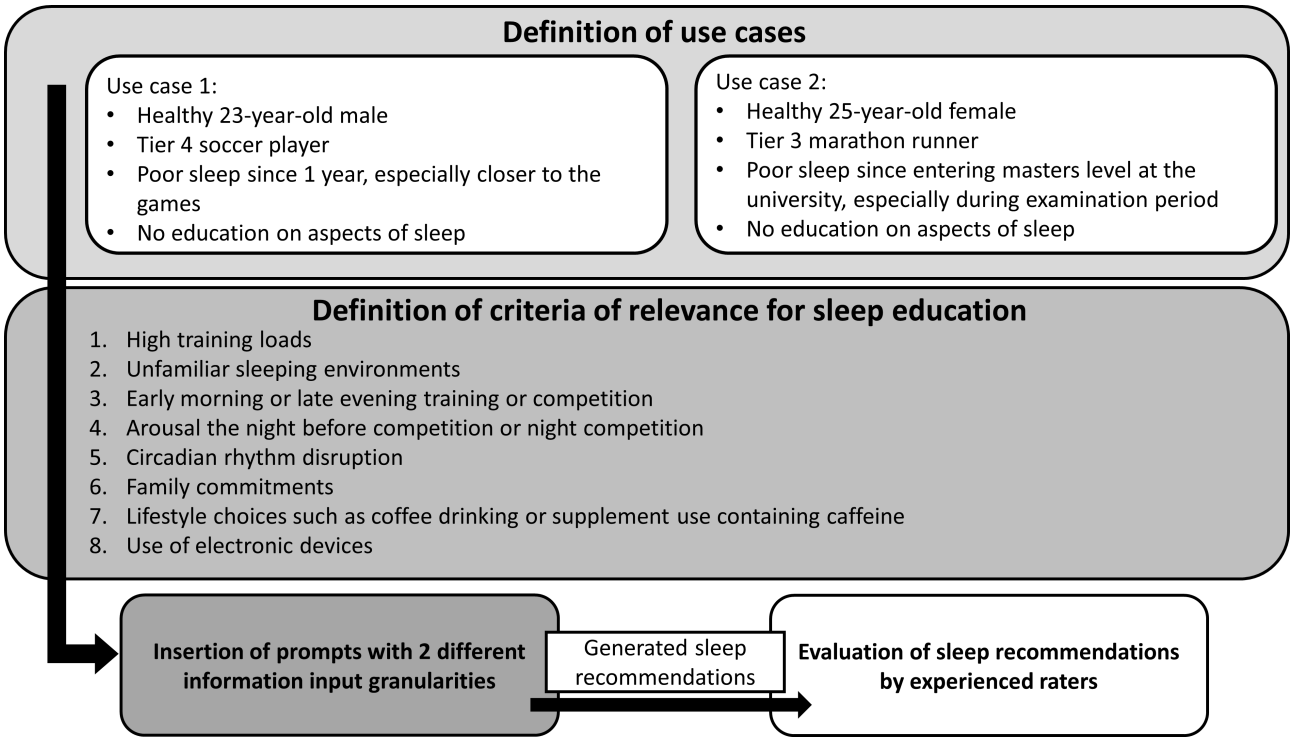

Ethical Considerations

The ethics committee of the Faculty of Exercise Science and Training at the University of Würzburg approved the study (reference: EV2025/5-0606). Raters were informed about procedures and gave their consent to participate in the study. No compensation was given to the raters. After receiving the ratings, these were deidentified.

Definition of 2 Use Cases

Use Case 1: Male Tier 4 Soccer Player

For use case 1, we define a healthy, 23-year-old male tier 4, elite soccer player [34] who trains 5 times a week and has 1 - 2 competitive games per week at a national level. The soccer player experiences poor sleep for approximately a year, and sleep disturbance is closer to important games. The individual in our use case has no formal education on, for example, factors

XSL•FO
RenderX

affecting sleep or on the effectiveness of countermeasures to improve sleep and no access to experienced and educated personnel who could educate him on sleep. We define that the major reasons for impaired sleep are high training loads, arousal the night before competition, and unfamiliar sleeping environments in the case of away games.

Use Case 2: Female Tier 3 Marathon Runner

For use case 2, we define a healthy, 25-year-old female tier 3, highly trained or national-level marathon runner [34] who runs around 100 km per week following a pyramidal training intensity distribution [35]. The female runner competes at the national level. In addition to her running training, the runner is enrolled at a master level at university.

The female runner experiences poor sleep since entering master level at university (approximately 6 months ago), and sleep is compromised especially during examination periods, which also affects running training and performance. The individual in our use case has no formal education on, for example, factors affecting sleep or on the effectiveness of countermeasures to improve sleep and no access to experienced and educated personnel who could educate her on sleep. We define that the major reasons for impaired sleep are training loads and disturbances stemming from university obligations.

In line with the aims of this research, we challenge the publicly available LLMs (1) to identify the reasons for sleep disturbances and (2) to give evidence-based guidance on how to reduce the impact of factors negatively impacting sleep.

Criteria of Relevance for Sleep Education

There are many sport and nonsport factors that impact the sleep of athletes. However, factors influencing sleep that are most commonly mentioned in the literature include high training loads [6,14], unfamiliar sleeping environments [6], early morning or late evening training or competition [6,14], arousal the night before competition or the night after competition [6,14], circadian rhythm disruption (eg, due to long-haul travel) [6,14], family commitments [6], lifestyle choices such as coffee drinking or supplement use containing caffeine [6,14], and use of electronic devices (eg, smartphone use) [14].

Prompts Inserted Into the LLMs

Given the chatbot nature of LLMs, we assume that the input provided by individuals seeking sleep education will vary, like any other conversation. Depending on factors such as previous knowledge about sleep or personal experiences, we assume that some individuals may provide minimal information, while others may be more detailed. To accommodate this diversity in the input information, we developed 2 distinct input information scenarios for both our use cases. Scenario 1 resembles an individual who inserts little to no information and only asks superficial check-backs once a response is given by LLMs. Scenario 2 resembles an individual who inserts more information and asks more detailed check-backs once a response is given by LLMs. The complete conversations with LLMs are available in Multimedia Appendices 1 and 2. Prompts of the scenarios were designed by the authors who are frequently in conversations with athletes on aspects of sleep.

For use case 1, the initial prompts were as follows:

Scenario 1:

> *I feel tired in the morning and after waking up. I do not know why. This is especially worse close to soccer games. Can you give me advice on how to improve my sleep?*

Scenario 2:

> *I am a 23 year old male highly trained/national level soccer player. I train 5 times a week and plus a game per week. I feel tired due to poor sleep at night since approximately a year. Especially the night before a game I sleep poorly. Can you give me advice on factors which might affect my sleep and how I can reduce the impact of these factors? Use only scientific literature and specifically, for each advice, state this literature and provide a reference list at the end.*

For use case 2, the initial prompts were as follows:

Scenario 1:

> *I feel tired in the morning and after waking up. I do not know why. Maybe it is due to my running training or due to entering master level at university. Can you give me advice on how to improve my sleep?*

Scenario 2:

> *I am a 25 year old female tier 3, highly trained/national level marathon runner player and I run approx. 100km per week with a pyramidal intensity distribution. I feel tired in the morning since approximately entering my master course at university. Can you give me advice on factors which might affect my sleep and how I can reduce the impact of these factors? Use only scientific literature and specifically, for each advice, state this literature and provide a reference list at the end.*

Here, we used GPT-4o and Google Gemini Advanced without any use of plug-ins, as both are publicly available and thereby can be used by individuals who seek sleep education. Prompts were inserted on May 15, 2024.

Raters

We reached out to well-educated and experienced raters on sleep and athletes to assess the provided recommendations on the outlined aspects relevant for sleep education on a 1 to 5 Likert scale. We included a total of 13 experienced raters (age span: 28 - 42 years; n=6 with a PhD, n=7 with a master degree in human physiology or sleep science) working for an average of 9 (SD 7) years with athletes on matters of sleep and indicating a mean sleep research experience of 4 (SD 6) years participated in this study.

Statistical Analysis

We calculated descriptive statistics for the Likert scores on all rated items for each question. To test for significant differences in all rated items between the training plans, a Friedman test with Bonferroni correction was performed. Significance level was set to $P<.05$. Fleiss κ was calculated to assess interrater

XSL•FO
RenderX

reliability [36]. Interpretation of Fleiss κ results was conducted according to the classification by Landis and Koch [37]. Fleiss κ values were interpreted as follows: a value of 0.00-0.20 as "slight," 0.21-0.40 as "fair," 0.41-0.60 as "moderate," 0.61-0.80 as "substantial," and >0.80 as "almost perfect" [37]. All statistical analyses were performed in SPSS (version 28; IBM Corp).

Results

Overview

Table 1 represents Fleiss κ values for the different LLMs and use cases. The analysis of Fleiss κ indicated a fair agreement of the overall interrater reliability (0.280), while results ranged between 0.183 and 0.296 (Table 1).

Descriptive statistics of the evaluated sleep recommendations are presented in Table 2.

Table . Fleiss κ results.

Large language model, use case, and scenario		Fleiss κ
Google Gemini		
	Use case 1—scenario 1 (Gem_C1-S1)	0.270
	Use case 1—scenario 2 (Gem_C1-S2)	0.198
	Use case 2—scenario 1 (Gem_C2-S1)	0.296
	Use case 2—scenario 2 (Gem_C2-S2)	0.183
GPT-4o		
	Use case 1—scenario 1 (GPT_C1-S1)	0.258
	Use case 1—scenario 2 (GPT_C1-S2)	0.264
	Use case 2—scenario 1 (GPT_C2-S1)	0.256
	Use case 2—scenario 2 (GPT_C2-S2)	0.229

XSL•FO
RenderX

Table . Descriptive analysis of Likert-scale ratings[a].

Relevant aspects when deploying sleep recommendations	Gem_C1-S1[b]	GPT_C1-S1[c]	Gem_C1-S2[d]	GPT_C1-S2[e]	Gem_C2-S1[f]	GPT_C2-S1[g]	Gem_C2-S2[h]	GPT_C2-S2[i]
General aspects, median (IQR)								
Overall training plan	3 (3-3)	3 (3-4)	3 (3-4)	4 (4-4)	3 (3-4)	3 (3-4)	4 (3-4)	4 (3.75-4)
Training load	0 (0-3)	0 (0-0)	3 (0-4)	4 (3-4)	3 (2-4)	4 (3-4)	4 (3-4)	4 (4-5)
Unfamiliar sleeping environments	0 (0-3)	2 (0-3)	4 (3-4)	3 (0-4)	3 (0-3)	3 (0-4)	3 (0-4)	3 (0-4)
Early morning or late evening training or competition	3 (0-4)	3 (3-3)	4 (4-4)	4 (4-4)	4 (3-4)	3 (0-4)	0 (0-4)	4 (3-4)
Arousal the night before competition or training	0 (0-3)	3 (3-4)	4 (3-4)	3 (3-4)	0 (0-0)	0 (0-0)	0 (0-0)	0 (0-4)
Circadian rhythm disruptions or consistency of sleep schedule	3 (3-4)	4 (3-4)	4 (4-4)	4 (3-5)	4 (3-4)	3 (3-4)	4 (3-4)	3 (3-4)
Family commitments	0 (0-0)	0 (0-0)	0 (0-0)	0 (0-0)	0 (0-0)	0 (0-0)	0 (0-0)	0 (0-0)
Lifestyle choices	3 (0-3)	3 (0-4)	2 (0-3)	0 (0-3)	3 (2-4)	3.5 (2.25-4)	3 (0-3)	3 (3-3)
Use of electronic devices	3 (3-3.25)	4 (3-5)	0 (0-0)	4 (4-5)	3 (0-4)	4 (3-4)	4 (3-4)	4 (3-5)
Nutrition	3 (3-3)	4 (3-4)	4 (3-4)	3 (3-5)	4 (3-4)	3 (3-4)	4 (3-4)	4 (3-5)
Summary rating, n (%)								
<3	4 (40)	3 (30)	3 (30)	2 (20)	2 (20)	2 (20)	3 (30)	2 (20)
3	6 (60)	4 (40)	2 (20)	3 (30)	5 (50)	5 (50)	2(20)	3 (30)
>3	0 (0)	3 (30)	5 (50)	5 (50)	3 (30)	3 (30)	5 (50)	5 (50)
Scientific sources, median (IQR)								
Is the real and existing literature stated (no “fake citations”)?	0 (0-0)	0 (0-0)	4 (2-4)	5 (5-5)	0 (0-0)	0 (0-0)	4 (3-4)	5 (5-5)
Appropriateness of provided scientific evidence	0 (0-0)	0 (0-0)	4 (3-4)	4 (4-4)	0 (0-0)	0 (0-0)	4 (4-5)	4 (4-5)
Quality of provided scientific evidence in this specific context	0 (0-0)	0 (0-0)	4 (3-4)	4 (4-4)	0 (0-0)	0 (0-0)	4 (3-4)	4 (4-4)

XSL•FO
RenderX

Relevant aspects when deploying sleep recommendations	Gem_C1-S1[b]	GPT_C1-S1[c]	Gem_C1-S2[d]	GPT_C1-S2[e]	Gem_C2-S1[f]	GPT_C2-S1[g]	Gem_C2-S2[h]	GPT_C2-S2[i]
Summary rating, n (%)								
<3	3 (100)	3 (100)	0 (0)	0 (0)	3 (100)	3 (100)	0 (0)	0 (0)
3	0 (0)	0 (0)	0 (0)	1 (33.3)	0 (0)	0 (0)	0 (0)	0 (0)
>3	0 (0)	0 (0)	3 (100)	2 (66.6)	0 (0)	0 (0)	3 (0)	3 (100)

[a]Results range from 1=bad to 5=good with 0=not applicable.

[b]Gem_C1-S1: Google Gemini, use case 1—scenario 1.

[c]GPT_C1-S1: ChatGPT, use case 1—scenario 1.

[d]Gem_C1-S2: Google Gemini, use case 1—scenario 2.

[e]GPT_C1-S2: ChatGPT, use case 1—scenario 2.

[f]Gem_C2-S1: Google Gemini, use case 2—scenario 1.

[g]GPT_C2-S1: ChatGPT, use case 2—scenario 1.

[h]Gem_C2-S2: Google Gemini, use case 2—scenario 2.

[i]GPT_C2-S2: ChatGPT, use case 2—scenario 2.

Differences Regarding Input Information Granularity and Between Google Gemini and GPT-4o

Results for significance testing regarding different input information granularities and differences between Google Gemini and GPT-4o are presented in Table 3.

Significance testing of the comparison between identical prompts across different LLMs shows that GPT-4o attained significantly higher Likert-scale scores in 9 of 10 criteria of relevance for sleep education ($P<.001$ to $P=.045$). Higher input information granularity, independent of the LLM used, exhibits significantly higher Likert-scale ratings in 28 of 52 criteria items of relevance for sleep education ($P<.001$ to $P=.049$).

In this paper, we compare the output of different LLMs when the same information was inserted. We do not show comparisons of different LLMs and different information input (eg, Gem_C1-S1 vs GPT_C1-S2) but provide these to the interested reader in Multimedia Appendix 3.

XSL•FO
RenderX

Table . Results of the significance testing comparing training plans: between Google Gemini and GPT-4o for the same input information granularity and within Google Gemini or GPT-4o for different input information granularity.

Relevant aspects when deploying sleep recommendations	Significance testing (*P* value) (Google Gemini versus GPT-4o; same prompt, different LLM)[a]				Significance testing (*P* value) (input information granularity; different prompt, same LLM)			
	Gem_C1-S1[b] versus GPT_ C1-S1[c]	Gem_C1-S2[d] versus GPT_ C1-S2[e]	Gem_C2-S1[f] versus GPT_ C2-S1[g]	Gem_C2-S2[h] versus GPT_ C2-S2[i]	Gem_C1-S1 versus Gem_C1-S2	GPT_C1-S1 versus GPT_C1-S2	Gem_C2-S1 versus Gem_C2-S2	GPT_C2-S1 versus GPT_C2-S2
General aspects								
Overall training plan	.92	*.02*	.16	.07	.47	*.003*	.38	*.005*
Training load	.19	.21	.06	.21	*.003*	*<.001*	*.005*	*.03*
Unfamiliar sleeping environments	.63	.21	.76	.72	*.006*	.28	.15	.37
Early morning or late evening training or competition	.88	.88	.08	*.004*	*.01*	*.01*	*.009*	*.04*
Arousal the night before competition or training	*.005*	.85	.11	*.02*	*.001*	.54	.14	.35
Circadian rhythm disruptions or consistency of sleep schedule	.16	.96	.43	.96	*.01*	.25	.71	.71
Family commitments	>.99	*.049*	>.99	>.99	*.049*	>.99	>.99	>.99
Lifestyle choices	.74	.96	.74	.55	.37	.21	.28	.41
Use of electronic devices	.07	*<.001*	.27	.38	*<.001*	.22	.48	.64
Nutrition	*.03*	.70	.96	.28	*.046*	.83	.74	.14
Scientific sources								
Is real and existing literature stated (no “fake citations”)?	.62	*<.001*	.22	*<.001*	*<.001*	*<.001*	*<.001*	*<.001*
Appropriateness of provided scientific evidence	.92	.52	.17	.40	*<.001*	*<.001*	*<.001*	*<.001*
Quality of provided scientific evidence in this specific context	*.04*	.72	.25	.08	*<.001*	*<.001*	*<.001*	*<.001*

[a]LLM: large language model.

XSL•FO
RenderX

[b]Gem_C1-S1: Google Gemini, use case 1—scenario 1.
[c]GPT_C1-S1: ChatGPT, use case 1—scenario 1.
[d]Gem_C1-S2: Google Gemini, use case 1—scenario 2.
[e]GPT_C1-S2: ChatGPT, use case 1—scenario 2.
[f]Gem_C2-S1: Google Gemini, use case 2—scenario 1.
[g]GPT_C2-S1: ChatGPT, use case 2—scenario 1.
[h]Gem_C2-S2: Google Gemini, use case 2—scenario 2.
[i]GPT_C2-S2: ChatGPT, use case 2—scenario 2.

Discussion

Principal Findings

We aimed (1) to investigate the quality of sleep recommendations provided to athletes generated by different publicly available LLMs as evaluated by experienced raters and (2) to investigate if the quality of sleep recommendations differs depending on the provided information by the user.

Our main results are as follows:

- The highest Likert-scale rating was achieved by GPT-4o using the prompt with high input granularity (use case 2, scenario 2) with 8 ratings >3 (tendency toward good), 3 ratings equal 3 (neutral), and 2 ratings <3 (tendency toward bad) on a 1 - 5 Likert Scale. This indicates that even the highest-ranked recommendations provided by the herein investigated LLMs are not optimal.
- Sleep recommendations by GPT-4o received higher Likert-scale ratings compared to those by Google Gemini (9 of 10 significant differences, with $P<.001$ to $P=.04$). This suggests a tendency that GPT-4o outperforms Google Gemini in the investigated sleep deficiency scenarios.
- Quality of sleep recommendations enhances with higher input information granularity (significantly higher Likert-scale ratings in 28 of 52 criteria items of relevance for sleep education; $P<.001$ to $P=.049$); however, some criteria of relevance for sleep education were partly or completely omitted, irrespective of the input information granularity.

Ratings of Generated Recommendations Regarding Sleep

Our results indicate that sleep recommendations of publicly available LLMs are not rated optimally, even when inserting a prompt with a high input information granularity. Although prompting GPT-4o with high input information granularity (user case 2, scenario 2) gained the highest Likert-scale ratings by the experienced raters (n=8 >3 Likert-scale rating), the summarized ratings demonstrate that the sleep recommendations exhibit deficiencies (n=3 ratings equal 3 [neutral]; n=2 ratings <3 [tendency toward bad]).

The insufficiency in providing optimal recommendations aligns with previous studies assessing LLMs in other research fields. For example, a systematic review and meta-analysis on ChatGPT's performance in answering medical questions showed that ChatGPT has an overall accuracy of 56%, suggesting potential but inadequacy for independent clinical decision-making [38]. Research in the field of nutrition reported inappropriate recommendations generated by ChatGPT, indicating that personalized dietary recommendations by ChatGPT involve unpredictable errors [32,39]. Therefore, LLMs should not be relied on to provide current nutritional advice without nutrition professionals [32,39].

Our results revealed further limitations pertaining to the negligence of criteria that are relevant for sleep education by the LLMs. In particular, the criterion "family commitments" did not exceed a Likert-scale median of 0 (IQR 0-0) across all LLMs and levels of input information granularities. Regarding the evaluation of all LLMs and input information granularities, "arousal the night before competition" attained a median of 0 (IQR 0-0 to 0-4) in 5 of 8 Likert-scale ratings, while "lifestyle" gained a maximum median of 3.5 (IQR 2.25-4), reflecting a neutral rating. These criteria were highlighted by recent research to influence both sleep quality and quantity [6,14]. Nastasi et al [40] reported similar results in assessing ChatGPT's ability to provide appropriate responses to medical questions within care contexts. While the authors noted appropriate responses corresponding to clinical guidelines, they indicated insufficient recommendations in regard to personalized medical advice, especially neglecting social factors [40].

Collectively, our results reveal that tailored sleep recommendations generated by GPT-4o and Google Gemini exhibit deficiencies according to received Likert-scale ratings by experienced raters, particularly in neglecting relevant criteria for sleep education. Therefore, sleep recommendations by LLMs should be carefully reviewed by a qualified coach or sleep professional before being applied in sleep educational settings.

Differences in Ratings of GPT-4o and Google Gemini

Our results demonstrate that sleep recommendations by GPT-4o generally were rated higher compared to those by Google Gemini. Of the 10 significant differences between the 2 LLMs in Likert-scale ratings, 9 favored GPT-4o ($P=.001$ to $P=.04$). Although the remaining scenarios (n=42) did not display significant results, our findings suggest a better quality of recommendations for sleep education by GPT-4o compared to Google Gemini.

Similar to our work, different authors compared different publicly available LLMs in different scenarios. For example, Günay et al [41] assessed GPT-4o and Google Gemini on 40 electrocardiogram cases and their responses to the most likely diagnosis. The results revealed that GPT-4o achieved higher accuracy compared to Gemini in electrocardiogram diagnostics. Carlà et al [42] evaluated ChatGPT and Google Gemini on 4 retinal detachment cases related to planned surgeries and revealed that ChatGPT received higher ratings for accuracy and precision compared to Gemini. Hieronimus et al [43] analyzed the completeness and accuracy of dietary reference intake in

XSL•FO
RenderX

meal plans for different dietary patterns created by ChatGPT and Google Bard (subsequently rebranded Gemini) and revealed higher quality in the response of ChatGPT. In hand surgery, it was shown that Google Gemini outperformed ChatGPT in classifying injuries, while ChatGPT provided more sensitive recommendations regarding surgical interventions [44].

Collectively, it appears that different LLMs differ in output quality, irrespective of the use case, and there seems to be a tendency that versions of ChatGPT outperform Google Gemini (which is in line with the results of our study), even though such statements need further investigation.

Differences in Quality Regarding Prompt Information Granularity

Our results indicate that higher input granularity leads to better-rated sleep recommendations, independent of the LLM. In the combined results of both use cases and LLMs, scenario 2 (higher input information granularity) received significantly higher Likert-scale ratings in 28 of 52 criteria items of relevance for sleep education (P=.001 to P=.049) compared to scenario 1 (low input information granularity), while scenario 1 attained higher Likert-scale ratings in 2 of 52 items (P=.001 to P=.009) compared to scenario 2.

Our finding is in line with research examining LLMs in other contexts [30,31,45-48]. For example, Kunze et al [48] inserted 20 knee complaints necessitating triage into ChatGPT-4 and investigated the accuracy and suitability rated by orthopedic sports medicine physicians. The authors stated that when providing additional input information, the accuracy of information output by ChatGPT-4 improved, particularly with enhancements in conservative management, surgical approaches, and related treatments [48].

In the context of endurance sports, Düking et al [30] provided 3 prompts with different levels of input information granularities to ChatGPT, resulting in 3 training plans aimed to improve running performance. Following an evaluation of these training plans by coaching experts, the authors indicated an increased quality with more input information provided [30].

Collectively, when using LLMs, it seems important for users to input a sufficient amount of information into LLMs to improve output quality, at least for the herein investigated scenario.

An additional point of interest is the quality and appropriateness of the scientific sources. Our results demonstrate that more input information leads to significantly higher ratings in items pertaining to the used scientific sources (P<.001), irrespective of the analyzed LLM. While higher input information granularity leads to improved sleep recommendations, the higher quality and appropriateness of the scientific sources may also contribute to these improvements. Since our analysis did not involve a differentiation between the effect of input information granularity and the use of scientific resources, future research should investigate both factors separately.

Conclusively, higher input information granularity leads to improved sleep recommendations and might be influenced by higher quality and appropriateness of the scientific sources. When supplying LLMs for sleep education, coaches and athletes should be highly aware of the appropriate input granularity and conduct a thorough review of the received output, potentially in consultation with an individual with strong experience in the field of sleep and athletic populations.

Strengths, Limitations, and Future Research

Strengths of our study include the assessment of different publicly available LLMs (ie, GPT-4o and Google Gemini) in different use cases and with different input information granularity, allowing a detailed analysis of provided sleep recommendations to the athletic population. Another strength of our study is the involvement of experienced raters evaluating the recommendations of LLMs.

Our results are limited to GPT-4o and Google Gemini on the versions available on May 15, 2024. As LLMs show fast developments (eg, by being able to search the internet), transferring our results to newer versions should be performed with caution. Due to such fast developments, it appears necessary to develop assessment methods or frameworks to evaluate the quality of LLMs in different scenarios to inform practitioners of the quality of currently available LLMs. Additionally, our results are, strictly speaking, only valid for the herein tested prompts, and other prompts might yield different results.

Despite the fact that all raters were experienced or well-educated in the field of sleep, the interrater reliability of our study ranges between slight and fair agreements (0.183 to 0.296). The tendency toward low interrater reliability is in line with previous research [30]. It may indicate that no single approach is universally optimal or perfect regarding athlete sleep recommendations. This suggests that different experts might rate the recommendations provided by LLMs in the respective use case differently, for example, based on personal preferences or personal experiences. While our results hold practical insights for athletes seeking sleep recommendations from LLMs (eg, to insert detailed information), individual sleep coaches might disagree with the recommendations provided by LLMs.

It seems important to note that even though it was shown that sleep education approaches can result in enhanced sleep behavior [13,49-51], to the best of our knowledge, there is no scientific evidence available that sleep recommendations generated by LLMs improve aspects of sleep in the athletic population. Future studies should evaluate the effectiveness of sleep education provided by publicly available LLMs to improve aspects of sleep.

Conclusions

Our study indicates that sleep recommendations generated by GPT-4o or Google Gemini are not rated optimally, independently of the level of input information granularity. However, our results demonstrate that GPT-4o provides better sleep recommendations to athletes compared to Google Gemini and that sleep recommendations improved with more detailed input information for both herein investigated LLMs. Collectively, for LLMs to be used in practice, it seems essential to insert detailed information into LLMs and to thoroughly

XSL•FO
RenderX

review the provided sleep recommendations for athletic populations.

Acknowledgments

The authors acknowledge the use of large language models for assisting with grammar and language checks during the final preparation of this manuscript. The authors checked the final manuscript and approved it for publication. JMIR Publications has provided article processing fee (APF) support for the publication of this paper. The authors also acknowledge funding for the open-access publication fees by TU Braunschweig.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Conversation with ChatGPT4o.
[DOCX File, 23 KB - formative_v9i1e71358_app1.docx]

Multimedia Appendix 2
Conversation with Google Gemini.
[DOCX File, 28 KB - formative_v9i1e71358_app2.docx]

Multimedia Appendix 3
Results of the significance testing comparing training plans generated by Google Gemini and GPT-4o in response to different input information granularity.
[DOCX File, 15 KB - formative_v9i1e71358_app3.docx]

References

1. Ramar K, Malhotra RK, Carden KA, et al. Sleep is essential to health: an American Academy of Sleep Medicine position statement. J Clin Sleep Med 2021 Oct 1;17(10):2115-2119. [doi: 10.5664/jcsm.9476] [Medline: 34170250]
2. Agarwal AK, Gonzales R, Scott K, Merchant R. Investigating the feasibility of using a wearable device to measure physiologic health data in emergency nurses and residents: observational cohort study. JMIR Form Res 2024 Feb 22;8:e51569. [doi: 10.2196/51569] [Medline: 38386373]
3. Xu Q, Kim Y, Chung K, Schulz P, Gottlieb A. Prediction of mild cognitive impairment status: pilot study of machine learning models based on longitudinal data from fitness trackers. JMIR Form Res 2024 Jul 18;8:e55575. [doi: 10.2196/55575] [Medline: 39024003]
4. Knowlden AP, Winchester LJ, MacDonald HV, Geyer JD, Higginbotham JC. Associations among cardiometabolic risk factors, sleep duration, and obstructive sleep apnea in a southeastern US rural community: cross-sectional analysis from the SLUMBRx-PONS Study. JMIR Form Res 2024 Nov 8;8:e54792. [doi: 10.2196/54792] [Medline: 39514856]
5. Gupta L, Morgan K, Gilchrist S. Does elite sport degrade sleep quality? A systematic review. Sports Med 2017 Jul;47(7):1317-1333. [doi: 10.1007/s40279-016-0650-6] [Medline: 27900583]
6. Walsh NP, Halson SL, Sargent C, et al. Sleep and the athlete: narrative review and 2021 expert consensus recommendations. Br J Sports Med 2021 Apr;55(7):356-368. [doi: 10.1136/bjsports-2020-102025]
7. Driller MW, Suppiah H, Rogerson D, Ruddock A, James L, Virgile A. Investigating the sleep habits in individual and team-sport athletes using the Athlete Sleep Behavior Questionnaire and the Pittsburgh Sleep Quality Index. Sleep Sci 2022;15(1):112-117. [doi: 10.5935/1984-0063.20210031] [Medline: 35662975]
8. Halson SL. Sleep in elite athletes and nutritional interventions to enhance sleep. Sports Med 2014 May;44(Suppl 1):S13-S23. [doi: 10.1007/s40279-014-0147-0] [Medline: 24791913]
9. Cunha LA, Costa JA, Marques EA, Brito J, Lastella M, Figueiredo P. The impact of sleep interventions on athletic performance: a systematic review. Sports Med Open 2023 Jul 18;9(1):58. [doi: 10.1186/s40798-023-00599-z] [Medline: 37462808]
10. Venter RE. Perceptions of team athletes on the importance of recovery modalities. Eur J Sport Sci 2014;14 Suppl 1:S69-S76. [doi: 10.1080/17461391.2011.643924] [Medline: 24444246]
11. Czeisler CA, Klerman EB. Circadian and sleep-dependent regulation of hormone release in humans. Recent Prog Horm Res 1999;54:97-130. [Medline: 10548874]

XSL•FO
RenderX

12. Fullagar HHK, Skorski S, Duffield R, Hammes D, Coutts AJ, Meyer T. Sleep and athletic performance: the effects of sleep loss on exercise performance, and physiological and cognitive responses to exercise. Sports Med 2015 Feb;45(2):161-186. [doi: 10.1007/s40279-014-0260-0] [Medline: 25315456]
13. Caia J, Scott TJ, Halson SL, Kelly VG. The influence of sleep hygiene education on sleep in professional rugby league athletes. Sleep Health 2018 Aug;4(4):364-368. [doi: 10.1016/j.sleh.2018.05.002] [Medline: 30031530]
14. Cook JD, Charest J. Sleep and performance in professional athletes. Curr Sleep Med Rep 2023;9(1):56-81. [doi: 10.1007/s40675-022-00243-4] [Medline: 36683842]
15. Miles KH, Clark B, Fowler PM, Miller J, Pumpa KL. Sleep practices implemented by team sport coaches and sports science support staff: a potential avenue to improve athlete sleep? J Sci Med Sport 2019 Jul;22(7):748-752. [doi: 10.1016/j.jsams.2019.01.008] [Medline: 30685228]
16. Salah M, Alhalbusi H, Ismail MM, Abdelfattah F. Chatting with ChatGPT: decoding the mind of Chatbot users and unveiling the intricate connections between user perception, trust and stereotype perception on self-esteem and psychological well-being. Curr Psychol 2024 Mar;43(9):7843-7858. [doi: 10.1007/s12144-023-04989-0]
17. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023 Mar 19;11(6):887. [doi: 10.3390/healthcare11060887] [Medline: 36981544]
18. Roos J, Kasapovic A, Jansen T, Kaczmarczyk R. Artificial intelligence in medical education: comparative analysis of ChatGPT, Bing, and medical students in Germany. JMIR Med Educ 2023 Sep 4;9:e46482. [doi: 10.2196/46482] [Medline: 37665620]
19. Dagli MM, Oettl FC, Gujral J, et al. Clinical accuracy, relevance, clarity, and emotional sensitivity of large language models to surgical patient questions: cross-sectional study. JMIR Form Res 2024 Jun 7;8:e56165. [doi: 10.2196/56165] [Medline: 38848553]
20. Skryd A, Lawrence K. ChatGPT as a tool for medical education and clinical decision-making on the wards: case study. JMIR Form Res 2024 May 8;8:e51346. [doi: 10.2196/51346] [Medline: 38717811]
21. Wang Y, Chen Y, Sheng J. Assessing ChatGPT as a medical consultation assistant for chronic hepatitis B: cross-language study of English and Chinese. JMIR Med Inform 2024 Aug 8;12:e56426. [doi: 10.2196/56426] [Medline: 39115930]
22. Kim J, Vajravelu BN. Assessing the current limitations of large language models in advancing health care education. JMIR Form Res 2025 Jan 16;9:e51319. [doi: 10.2196/51319] [Medline: 39819585]
23. Yang X, Li G. Psychological and behavioral insights from social media users: natural language processing-based quantitative study on mental well-being. JMIR Form Res 2025 Jan 20;9:e60286. [doi: 10.2196/60286] [Medline: 39832365]
24. Zheng J, Wang J, Shen J, An R. Artificial intelligence applications to measure food and nutrient intakes: scoping review. J Med Internet Res 2024 Nov 28;26:e54557. [doi: 10.2196/54557] [Medline: 39608003]
25. Tagi M, Hamada Y, Shan X, et al. A food intake estimation system using an artificial intelligence-based model for estimating leftover hospital liquid food in clinical environments: development and validation study. JMIR Form Res 2024 Nov 5;8:e55218. [doi: 10.2196/55218] [Medline: 39500491]
26. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023 Jun 1;183(6):589. [doi: 10.1001/jamainternmed.2023.1838]
27. Lukac S, Dayan D, Fink V, et al. Evaluating ChatGPT as an adjunct for the multidisciplinary tumor board decision-making in primary breast cancer cases. Arch Gynecol Obstet 2023 Dec;308(6):1831-1844. [doi: 10.1007/s00404-023-07130-5] [Medline: 37458761]
28. Seth I, Cox A, Xie Y, et al. Evaluating chatbot efficacy for answering frequently asked questions in plastic surgery: a ChatGPT case study focused on breast augmentation. Aesthet Surg J 2023 Sep 14;43(10):1126-1135. [doi: 10.1093/asj/sjad140] [Medline: 37158147]
29. Samaan JS, Yeo YH, Rajeev N, et al. Assessing the accuracy of responses by the language model ChatGPT to questions regarding bariatric surgery. Obes Surg 2023 Jun;33(6):1790-1796. [doi: 10.1007/s11695-023-06603-5] [Medline: 37106269]
30. Düking P, Sperlich B, Voigt L, Van Hooren B, Zanini M, Zinner C. ChatGPT generated training plans for runners are not rated optimal by coaching experts, but increase in quality with additional input information. J Sports Sci Med 2024 Mar;23(1):56-72. [doi: 10.52082/jssm.2024.56] [Medline: 38455449]
31. Dergaa I, Saad HB, El Omri A, et al. Using artificial intelligence for exercise prescription in personalised health promotion: a critical evaluation of OpenAI's GPT-4 model. Biol Sport 2024 Mar;41(2):221-241. [doi: 10.5114/biolsport.2024.133661] [Medline: 38524814]
32. Dergaa I, Ben Saad H, Ghouili H, et al. Evaluating the applicability and appropriateness of ChatGPT as a source for tailored nutrition advice: a multi-scenario study. NAJM 2024;2(1):1-16. [doi: 10.61838/kman.najm.2.1.1]
33. Havers T, Masur L, Isenmann E, et al. Reproducibility and quality of hypertrophy-related training plans generated by GPT-4 and Google Gemini as evaluated by coaching experts. Biol Sport 2025 Apr;42(2):289-329. [doi: 10.5114/biolsport.2025.145911] [Medline: 40182716]
34. McKay AKA, Stellingwerff T, Smith ES, et al. Defining training and performance caliber: a participant classification framework. Int J Sports Physiol Perform 2022 Feb 1;17(2):317-331. [doi: 10.1123/ijspp.2021-0451] [Medline: 34965513]

XSL•FO
RenderX

35. Casado A, González-Mohíno F, González-Ravé JM, Foster C. Training periodization, methods, intensity distribution, and volume in highly trained and elite distance runners: a systematic review. Int J Sports Physiol Perform 2022;17(6):820-833. [doi: 10.1123/ijspp.2021-0435]
36. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull 1971;76(5):378-382. [doi: 10.1037/h0031619]
37. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159. [doi: 10.2307/2529310]
38. Wei Q, Yao Z, Cui Y, Wei B, Jin Z, Xu X. Evaluation of ChatGPT-generated medical responses: a systematic review and meta-analysis. J Biomed Inform 2024 Mar;151:104620. [doi: 10.1016/j.jbi.2024.104620] [Medline: 38462064]
39. Agne I, Gedrich K. Personalized dietary recommendations for obese individuals—a comparison of ChatGPT and the Food4Me algorithm. Clin Nutr Open Sci 2024 Aug;56:192-201. [doi: 10.1016/j.nutos.2024.06.001]
40. Nastasi AJ, Courtright KR, Halpern SD, Weissman GE. A vignette-based evaluation of ChatGPT's ability to provide appropriate and equitable medical advice across care contexts. Sci Rep 2023 Oct 19;13(1):17885. [doi: 10.1038/s41598-023-45223-y] [Medline: 37857839]
41. Günay S, Öztürk A, Yiğit Y. The accuracy of Gemini, GPT-4, and GPT-4o in ECG analysis: a comparison with cardiologists and emergency medicine specialists. Am J Emerg Med 2024 Oct;84:68-73. [doi: 10.1016/j.ajem.2024.07.043] [Medline: 39096711]
42. Carlà MM, Gambini G, Baldascino A, et al. Exploring AI-chatbots' capability to suggest surgical planning in ophthalmology: ChatGPT versus Google Gemini analysis of retinal detachment cases. Br J Ophthalmol 2024 Sep 20;108(10):1457-1469. [doi: 10.1136/bjo-2023-325143] [Medline: 38448201]
43. Hieronimus B, Hammann S, Podszun MC. Can the AI tools ChatGPT and Bard generate energy, macro- and micro-nutrient sufficient meal plans for different dietary patterns? Nutr Res 2024 Aug;128:105-114. [doi: 10.1016/j.nutres.2024.07.002] [Medline: 39102765]
44. Pressman SM, Borna S, Gomez-Cabello CA, Haider SA, Forte AJ. AI in hand surgery: assessing large language models in the classification and management of hand injuries. J Clin Med 2024 May 11;13(10). [doi: 10.3390/jcm13102832] [Medline: 38792374]
45. Dergaa I, Fekih-Romdhane F, Hallit S, et al. ChatGPT is not ready yet for use in providing mental health assessment and interventions. Front Psychiatry 2023;14:1277756. [doi: 10.3389/fpsyt.2023.1277756] [Medline: 38239905]
46. Washif JA, Pagaduan J, James C, Dergaa I, Beaven CM. Artificial intelligence in sport: exploring the potential of using ChatGPT in resistance training prescription. Biol Sport 2024 Mar;41(2):209-220. [doi: 10.5114/biolsport.2024.132987] [Medline: 38524820]
47. Zaleski AL, Berkowsky R, Craig KJT, Pescatello LS. Comprehensiveness, accuracy, and readability of exercise recommendations provided by an AI-based chatbot: mixed methods study. JMIR Med Educ 2024 Jan 11;10:e51308. [doi: 10.2196/51308] [Medline: 38206661]
48. Kunze KN, Varady NH, Mazzucco M, et al. The large language model ChatGPT-4 exhibits excellent triage capabilities and diagnostic performance for patients presenting with various causes of knee pain. Arthroscopy 2025 May;41(5):1438-1447. [doi: 10.1016/j.arthro.2024.06.021]
49. O'Donnell S, Driller MW. Sleep-hygiene education improves sleep indices in elite female athletes. Int J Exerc Sci 2017;10(4):522-530. [doi: 10.70252/DNOL2901] [Medline: 28674597]
50. Fullagar HHK, Skorski S, Duffield R, Julian R, Bartlett J, Meyer T. Impaired sleep and recovery after night matches in elite football players. J Sports Sci 2016 Jul;34(14):1333-1339. [doi: 10.1080/02640414.2015.1135249] [Medline: 26750446]
51. Driller MW, Lastella M, Sharp AP. Individualized sleep education improves subjective and objective sleep indices in elite cricket athletes: a pilot study. J Sports Sci 2019 Sep;37(17):2121-2125. [doi: 10.1080/02640414.2019.1616900] [Medline: 31076021]

Abbreviations

GPT-4o: ChatGPT-4o
LLM: large language model

Edited by A Mavragani; submitted 16.01.25; peer-reviewed by M Schmidt, S Haghayegh; revised version received 06.03.25; accepted 11.03.25; published 08.07.25.

Please cite as:
Masur L, Driller M, Suppiah H, Matzka M, Sperlich B, Düking P
Assessment of Recommendations Provided to Athletes Regarding Sleep Education by GPT-4o and Google Gemini: Comparative Evaluation Study
JMIR Form Res 2025;9:e71358
URL: https://formative.jmir.org/2025/1/e71358
doi:10.2196/71358

XSL•FO
RenderX

© Lukas Masur, Matthew Driller, Haresh Suppiah, Manuel Matzka, Billy Sperlich, Peter Düking. Originally published in JMIR Formative Research (https://formative.jmir.org), 8.7.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

AI in Qualitative Health Research Appraisal: Comparative Study

August Landerholm

Physiotherapy Department, Healthscience Faculty, Mälardalen University, Avdelningen för Fysioterapi Akademin för Hälsa, Vård Och Välfärd Mälardalens Universitet, Västerås, Sweden

Corresponding Author:
August Landerholm
Physiotherapy Department, Healthscience Faculty, Mälardalen University, Avdelningen för Fysioterapi Akademin för Hälsa, Vård Och Välfärd Mälardalens Universitet, Västerås, Sweden

Abstract

Background: Qualitative research appraisal is crucial for ensuring credible findings but faces challenges due to human variability. Artificial intelligence (AI) models have the potential to enhance the efficiency and consistency of qualitative research assessments.

Objective: This study aims to evaluate the performance of 5 AI models (GPT-3.5, Claude 3.5, Sonar Huge, GPT-4, and Claude 3 Opus) in assessing the quality of qualitative research using 3 standardized tools: Critical Appraisal Skills Programme (CASP), Joanna Briggs Institute (JBI) checklist, and Evaluative Tools for Qualitative Studies (ETQS).

Methods: AI-generated assessments of 3 peer-reviewed qualitative papers in health and physical activity–related research were analyzed. The study examined systematic affirmation bias, interrater reliability, and tool-dependent disagreements across the AI models. Sensitivity analysis was conducted to evaluate the impact of excluding specific models on agreement levels.

Results: Results revealed a systematic affirmation bias across all AI models, with "Yes" rates ranging from 75.9% (145/191; Claude 3 Opus) to 85.4% (164/192; Claude 3.5). GPT-4 diverged significantly, showing lower agreement ("Yes": 115/192, 59.9%) and higher uncertainty ("Cannot tell": 69/192, 35.9%). Proprietary models (GPT-3.5 and Claude 3.5) demonstrated near-perfect alignment (Cramer V=0.891; P<.001), while open-source models showed greater variability. Interrater reliability varied by assessment tool, with CASP achieving the highest baseline consensus (Krippendorff α=0.653), followed by JBI (α=0.477), and ETQS scoring lowest (α=0.376). Sensitivity analysis revealed that excluding GPT-4 increased CASP agreement by 20% (α=0.784), while removing Sonar Huge improved JBI agreement by 18% (α=0.561). ETQS showed marginal improvements when excluding GPT-4 or Claude 3 Opus (+9%, α=0.409). Tool-dependent disagreements were evident, particularly in ETQS criteria, highlighting AI's current limitations in contextual interpretation.

Conclusions: The findings demonstrate that AI models exhibit both promise and limitations as evaluators of qualitative research quality. While they enhance efficiency, AI models struggle with reaching consensus in areas requiring nuanced interpretation, particularly for contextual criteria. The study underscores the importance of hybrid frameworks that integrate AI scalability with human oversight, especially for contextual judgment. Future research should prioritize developing AI training protocols that emphasize qualitative epistemology, benchmarking AI performance against expert panels to validate accuracy thresholds, and establishing ethical guidelines for disclosing AI's role in systematic reviews. As qualitative methodologies evolve alongside AI capabilities, the path forward lies in collaborative human-AI workflows that leverage AI's efficiency while preserving human expertise for interpretive tasks.

(JMIR Form Res 2025;9:e72815) doi:10.2196/72815

KEYWORDS

artificial intelligence; qualitative research appraisal; systematic reviews; interrater agreement; CASP checklist; Critical Appraisal Skills Programme; JBI checklist; Joanna Briggs Institute; ETQS; Evaluative Tools for Qualitative Studies; large language models; affirmation bias; human-AI collaboration

Introduction

Importance of Quality Assessment in Qualitative Research

Systematic quality assessment is foundational for establishing the credibility, dependability, and transferability of qualitative research findings [1]. Rigorous appraisal enables readers to evaluate the trustworthiness of study conclusions and their applicability to real-world contexts. Unlike quantitative methodologies, qualitative research prioritizes contextual richness and interpretive depth, necessitating frameworks that account for methodological diversity across paradigms (eg, phenomenology and ethnography).

Cross-study comparison is crucial for building a robust evidence base in any given field; yet, qualitative studies have seen little attention in systematic reviews. Facilitating and synthesizing

XSL•FO
RenderX

research in qualitative methodologies is challenging because these approaches are based on diverse philosophical foundations, such as phenomenology, ethnography, and grounded theory, as well as a wide variety of analytical methods [2,3]. The chosen analysis method of each research group may have varying degrees of subjective data interpretation, leading to diverse findings and conclusions [4]. Further complicating is the contextual nature of each study, where the findings may be highly dependent on its unique context [3]. The richness of data in qualitative datasets also encourages attempts at cohesive summaries, which can be challenging without losing important details or context [1]. Qualitative studies typically use smaller sample sizes than their quantitative methodological counterparts and use purposively selected samples [5]. Depending on the degree of rigor in the selection criteria, it becomes difficult to determine transferability or comparability [6]. Facilitating comparison and research synthesis becomes easier through standardized quality assessment criteria.

Systematic Assessment for Qualitative Research

Various assessment tools exist for systematically addressing qualitative study quality and have all been developed to address the unique challenges of evaluating qualitative research [7]. If a research group wishes to systematically address a field that has been given qualitative attention, their choice of assessment tool will provide different weight to studies in the final synthesis, potentially affecting the review's conclusions. There are also potential limitations of an overly systematic approach to addressing qualitative study quality, as overly rigid quality criteria may not capture the diversity of qualitative research approaches more broadly [2].

There are methodological issues with systematic quality assessment that endanger the potential credibility of a systematic review. Quality assessment of qualitative research is often time-consuming and labor-intensive, limiting the number of studies that systematic reviews tend to include. Human reviewers may have varying interpretations of quality criteria, leading to inconsistencies [8]. Different research groups might dissolve this in different ways, with bias present [9]. Humans' ability to recognize patterns is ever so present in systematic reviews, as a reviewer can detect recurring themes or quality indicators across supposedly independent quality assessments. These aspects may be augmented using another, nonhuman, reviewer [10].

In the context of health research, the systematic appraisal of qualitative studies is particularly critical, as it directly shapes the evidence base used to inform health care practice, policy, and patient care [11]. Qualitative research in health not only helps to understand patient experiences, barriers, and preferences but also guides the development of interventions and health technologies that are responsive to real-world needs. Therefore, the rigor and consistency of quality assessment tools are deemed trustworthy and ultimately influence clinical decision-making, health policy recommendations, and the quality of care delivered to diverse patient populations.

The Role of Artificial Intelligence in Research Quality Assessment

Artificial intelligence (AI) has received attention in the qualitative research field, offering new possibilities for many steps of the research process [12,13]. AI's role in systematic reviews has been theorized and tested in the quantitative space, but its potential in the qualitative field is given less attention [14,15]. AI's ability to work efficiently and at scale holds the potential for systematic appraisal of study quality, and the system's consistency could reduce human bias and variability [16]. Assessments done by AI could also enable the inclusion or exclusion of pattern recognition, as an AI can be instructed to complete a task independently or dependent on previous tasks. However, while AI can process complex data, it may struggle with the nuanced, context-dependent nature of qualitative research.

Large language models are among the most popular AI tools for researchers today [17]. These advanced AI systems offer capabilities that could potentially revolutionize the process of qualitative research quality assessment, as they are pretrained to analyze vast amounts of textual data to enable nuanced analysis [18]. Furthermore, these systems are multilingual, potentiating the inclusion of a paper written in a language unknown to the human researcher. Finally, these specific systems can be fine-tuned or prompted (post training) to follow specific assessment criteria, suitable for the systematic quality assessment of qualitative studies [13]. The reliability and validity testing of the tools are problematic, as new versions or updates of the tools push their capabilities faster than the scientific community can assess their usefulness. The need to test these tools still exists, as the prevalence of its use is growing [19].

Aims of the Study

Given the complex landscape of qualitative research quality assessment and the emerging potential of augmenting AI in research processes, this study aims (1) to evaluate and compare the performance of different AI models in assessing the quality of qualitative research studies using various assessment tools; (2) to compare the ratings given by 5 AI models (GPT-3.5, Claude 3.5, Sonar Huge, GPT-4, and Claude 3 Opus) when assessing qualitative studies; (3) to evaluate the interrater agreement among these AI models using 3 different assessment tools: Critical Appraisal Skills Programme (CASP), Joanna Briggs Institute (JBI), and Evaluative Tool for Qualitative Studies (ETQS); (4) to analyze how the exclusion of individual AI models affects the overall interrater agreement for each assessment tool; and (5) to identify specific items or criteria within these assessment tools that lead to prominent disagreements among the AI raters.

Methods

Overview

These models are chosen based on their diverse architectures and capabilities, which are crucial for a comprehensive analysis of AI augmentation in qualitative research. The selected models are included in Table 1.

XSL•FO
RenderX

Table . Overview of artificial intelligence (AI) models used for qualitative research appraisal.

Model	Developer	Release date	Size (B parameters)
GPT-3.5	OpenAI	2022	175
Claude 3.5	Anthropic	2024	Not disclosed
Sonar Huge	Perplexity AI, based on Llama 3.1	2024	405
GPT-4	OpenAI	2023	Not disclosed
Claude 3 Opus	Anthropic	2024	Not disclosed

This diverse selection of models detailed in Table 1 aims to identify which AI performance metrics are most beneficial for qualitative research quality assessment and to highlight areas where AI may complement or challenge each other. To evaluate the performance of these models, specific criteria will be used, including accuracy in coding, contextual understanding, and bias detection. The evaluation will use a standardized dataset of qualitative research papers to ensure a robust comparison across models.

Quality Assessment Tools for Qualitative Research

Overview

The AI models will be instructed to use 3 widely recognized quality assessment tools for qualitative research. These tools have been comparatively analyzed in previous studies [20]. The CASP checklist was chosen for its widespread use and accessibility in various research fields and has been previously used in qualitative assessments and syntheses [21-23]. The JBI Critical Appraisal Checklist for Qualitative Research was chosen for its focus on the alignment between research objectives and methodological choices [22,24,25]. Finally, the ETQS was selected for its comprehensive approach to evaluating qualitative research integrity, offering a more nuanced assessment of methodological rigor. Each tool offers perspectives on qualitative research quality [20], which will facilitate a multifaceted assessment of the AI models' ability to understand and evaluate different aspects of qualitative studies. The combination of these tools will provide a robust framework for comparing AI performance across various dimensions of qualitative research quality.

Assessment Process

Three peer-reviewed qualitative research papers have been selected as the source material for this study. These papers represent diverse topics within health and physical activity research, providing a robust basis for evaluating the AI models' performance across varied contexts (Table 2).

Table . Summary of qualitative health studies evaluated by artificial intelligence models.

Study title	Year	Focus area	Methodology
Paper A: A qualitative study examining the validity and comprehensibility of physical activity items: developed and tested in children with juvenile idiopathic arthritis [26]	2019	Physical activity in children with juvenile idiopathic arthritis	Qualitative interviews
Paper B: "If only balls could talk...": barriers and opportunities to participation for students with blindness and visual impairment in specialized PE [27]	2023	Participation barriers for students with visual impairments in PE[a]	Focus groups
Paper C: A qualitative study of exercise and physical activity in adolescents with pediatric-onset multiple sclerosis [28]	2019	Exercise and physical activity in adolescents with MS[b]	Semistructured interviews

[a]PE: physical education.

[b]MS: multiple sclerosis.

Application of AI Models to Each Assessment Tool

Each AI model will be tasked with applying all 3 quality assessment tools (CASP, ETQS, and JBI) to the selected studies. All papers are free text, and the AI models will be provided with the full text of each study as well as the assessment criteria for each tool. All AI-generated assessments will be collected and stored for analysis. AI-generated assessments will be formalized in a standardized format.

Repeated Assessments for Consistency

The repeated assessments for consistency are as follows:

1. Input preparation: The full text of each study will be provided to the AI models along with the complete assessment criteria for each tool. To ensure consistency, the input format will be standardized across all models.
2. Assessment protocol: AI models will be instructed to conduct a comprehensive quality assessment of each study using all 3 tools independently. Clear instructions will be

XSL•FO
RenderX

provided to ensure the models understand the task requirements.

3. Structured output: To facilitate comparative analysis, AI models will be required to provide their assessments in a standardized format for each tool. This may include numerical scores, categorical ratings, and textual explanations.
4. Reasoning transparency: The AI models will be prompted to explain their reasoning for each assessment criterion, providing insights into their decision-making process and allowing for evaluation of their understanding of qualitative research principles.
5. Consistency evaluation: Each AI model will perform the assessment task multiple times to evaluate the consistency of their outputs and identify any variability in their assessments.
6. Data collection and storage: All AI-generated assessments, including explanations and any variations in repeated assessments, will be systematically collected and stored in a secure database for subsequent analysis. This will ensure data integrity and facilitate comprehensive evaluation.
7. Bias mitigation: To minimize potential biases, the order of presenting studies and assessment tools to the AI models will be randomized for each evaluation session.

Ethical Considerations

All analyzed studies were previously published and had undergone their own ethical review processes. No new data were collected from individuals. The role of AI in the research process was disclosed, and all AI-assisted assessments were documented and stored securely. No personal or sensitive data were collected or processed. The research did not involve any intervention or interaction with human participants. The study posed no risk to individuals or groups, as it relied solely on secondary analysis of published material.

Results

All AI models showed high "Yes" rates (75.9% - 85.4%), with Claude 3.5 achieving the highest affirmation (164/192, 85.4%), as detailed in Table 3. GPT-4 diverged significantly, showing lower agreement ("Yes": 115/192, 59.9%) and elevated uncertainty ("Cannot tell": 69/192, 35.9%). GPT-3.5 and Claude 3.5 exhibited near-perfect alignment (Cramer V=0.891; P<.001). Sonar Huge ("Yes": 148/188, 78.7%) and Claude 3 Opus (145/191, 75.9%) demonstrated moderate consistency. GPT-4's exclusion boosted CASP agreement by 20% (α=0.784 vs 0.653 baseline), highlighting its role as a variability driver. Statistical associations weakened with open-source models (Cramer V=0.496 - 0.545), suggesting that architectural differences influence assessment patterns.

Table . Frequency of ratings by rater.

Model	Yes, n/N (%)	Cannot tell, n/N (%)	No, n/N (%)
GPT-3.5	158/192 (82.3)	34/192 (17.7)	0/192 (0)
Claude 3.5	164/192 (85.4)	28/192 (14.6)	0/192 (0)
Sonar Huge	148/188 (78.7)	33/188 (17.6)	7/188 (3.7)
GPT-4	115/192 (59.9)	69/192 (35.9)	8/192 (4.2)
Claude 3 Opus	145/191 (75.9)	38/191 (19.9)	8/191 (4.2)

Table 4 demonstrates GPT-3.5's significant agreement across all models (χ^2=47.0 - 152.3; P<.001), with effect sizes revealing distinct patterns such as perfect concordance with Claude 3.5 (Cramer V=0.891; χ^2_1=152.3), moderate agreement with Sonar Huge (V=0.539; χ^2_2=54.6) and Claude 3 Opus (V=0.496; χ^2_2=47.0), and finally, a weaker association with GPT-4 (V=0.545; χ^2_2=57.0) despite shared commercial development.

Table . GPT-3.5 associations.

	Chi-square[a] (*df*)	*P* value	Cramer *V*
Versus Claude 3.5	152.3 (1)	<.001	0.891
Versus Sonar Huge	54.6 (2)	<.001	0.539
Versus GPT-4	57.0 (2)	<.001	0.545
Versus Claude 3 Opus	47.0 (2)	<.001	0.496

[a]Chi-square test results showing associations between GPT-3.5 assessments and those of other artificial intelligence models.

The sensitivity analysis (summarized in Table 5) revealed tool-specific impacts of model exclusion on interrater agreement. For the CASP tool, excluding GPT-4 increased agreement by 20% (α=0.784), while Sonar Huge exclusion raised it by 18% (α=0.773), suggesting that these models introduce divergent interpretations of methodological rigor criteria. Conversely, JBI agreement improved most when excluding Sonar Huge (+18%; α=0.561) but dropped sharply without GPT-3.5 (−17%; α=0.398), indicating its stabilizing role for JBI appraisals. ETQS maintained the lowest baseline agreement (α=.376), with marginal improvements when excluding GPT-4 or Claude 3 Opus (+9%; α=0.409). This aligns with findings from Table 6, where ETQS criteria like policy implications (item 35) showed full-spectrum disagreements across models. Notably, proprietary

XSL•FO
RenderX

models (GPT-3.5 or Claude 3.5) consistently supported consensus-building, as their exclusion reduced CASP or JBI agreement by 12% - 17%. This pattern mirrors architectural similarities observed in GPT-3.5 and Claude 3.5's coding behaviors.

Table . Sensitivity analysis of interrater agreement (Krippendorff α)[a] across model exclusion scenarios.

Model exclusion scenario	JBI[b] (Δ%[c])	CASP[d] (Δ%)	ETQS[e] (Δ%)
All 5 models	0.477	0.653	0.376
Exclude GPT-3.5	0.398 (−17)	0.572 (−12)	0.346 (−8)
Exclude Claude 3.5	0.468 (−2)	0.572 (−12)	0.356 (−5)
Exclude Sonar Huge	*0.561 (+18)* [f]	*0.773 (+18)*	0.359 (−5)
Exclude GPT-4	0.494 (+3)	*0.784 (+20)*	0.409 (+9)
Exclude Claude 3 Opus	0.468 (−2)	0.572 (−12)	0.409 (+9)

[a]α values represent Krippendorff interrater reliability coefficient.
[b]JBI: Joanna Briggs Institute.
[c]Δ%=percentage change from full model agreement.
[d]CASP: Critical Appraisal Skills Programme.
[e]ETQS: Evaluative Tools for Qualitative Studies.
[f]Values in italics format highlight agreement improvements ≥10% across all tools.

Table . High-discrepancy Evaluative Tools for Qualitative Studies (ETQS) criteria across artificial intelligence models.

ETQS item	Criteria description	GPT-3.5	Claude 3.5	Sonar Huge	GPT-4	Claude 3 Opus	Disagreement score[a]
35	Generalizability to settings	Yes	Yes	Cannot tell	No	No	3
36	Generalizability to populations	Yes	Yes	Cannot tell	No	No	3
38	Policy implications	Cannot tell	Cannot tell	Yes	Cannot tell	No	2
43	Reviewer identification	Cannot tell	Cannot tell	Cannot tell	No	Yes	2
44	Review date verification	Cannot tell	Cannot tell	Cannot tell	No	Yes	2
8	Methodological framework alignment	Yes	Yes	No	Yes	Yes	2

[a]Number of distinct response categories (yes or cannot tell or no) per criterion.

Figure 1 illustrates how excluding specific AI models affects interrater agreement across 3 qualitative research assessment tools: JBI, CASP, and ETQS. The CASP tool demonstrated the highest baseline agreement (α=0.653), with notable improvements observed when GPT-4 or Sonar Huge was excluded, increasing agreement to 0.784 and 0.773, respectively. These findings suggest that GPT-4 and Sonar Huge may introduce variability in CASP assessments. In contrast, the exclusion of GPT-3.5, Claude 3.5, or Claude 3 Opus reduced agreement to 0.572, highlighting their role in fostering consensus.

XSL•FO
RenderX

Figure 1. Radar-chart visualization upon model exclusion. AI: artificial intelligence; CASP: Critical Appraisal Skills Programme; ETQS: Evaluative Tools for Qualitative Studies; JBI: Joanna Briggs Institute.

For the JBI tool, excluding Sonar Huge resulted in the largest improvement in agreement (α=0.561), while removing GPT-3.5 led to a significant drop to 0.398, indicating that GPT-3.5 is a key contributor to maintaining consistency in JBI assessments. The ETQS tool exhibited the lowest baseline agreement (α=0.376), with marginal gains observed when GPT-4 or Claude 3 Opus were excluded, both increasing agreement to 0.409. This suggests that ETQS assessments are generally consistent across models, with GPT-4 and Claude 3 Opus introducing slight variability.

These results underscore the importance of model selection in AI-assisted qualitative research assessment, as certain models contribute more significantly to consensus, while others may introduce variability depending on the assessment tool used.

Table 6 highlights the ETQS criteria where AI models demonstrated the most significant disagreements in their assessments. Items such as generalizability to settings and populations (items 35 and 36) exhibited the full spectrum of possible responses ("Yes," "Cannot tell," and "No"), indicating substantial variability in model interpretation. Other items, including policy implications (item 38) and methodological framework alignment (item 8), also showed notable disagreement, albeit with fewer distinct response categories. These findings underscore the challenges AI models face in achieving consensus on nuanced qualitative criteria, particularly those requiring contextual or interpretive judgment.

Discussion

Principal Findings

This study reveals critical insights about AI's role in qualitative research appraisal, particularly in health science contexts where methodological rigor directly impacts evidence-based practice [29,30]. All AI models demonstrated systematic affirmation bias, with "Yes" rates ranging from 75.9% to 85.4%, suggesting an inherent tendency toward favorable assessments regardless of the assessment tool (Table 3). Model-specific variability emerged as a key factor, particularly with GPT-4 diverging significantly ("Yes": 115/192, 59.9%) compared to proprietary models like GPT-3.5 and Claude 3.5, which showed near-perfect alignment (Cramer V=0.891; P<.001) as detailed in Table 4.

Tool-dependent disagreements were evident, particularly with ETQS criteria like policy implications (item 35) and generalizability (item 36), which elicited the full spectrum of responses across models. This highlights current limitations of AI in contextual interpretation. In health research, such biases could distort evidence syntheses informing clinical guidelines or public health policies, especially for studies like Paper C (multiple sclerosis), where AI's inability to contextualize structural barriers (eg, health care access disparities) risks undermining person-centered care models [31].

Comparison to Prior Work

The findings of this study are consistent with emerging research on AI-augmented qualitative analysis. The consensus-building role of proprietary models mirrors previous findings regarding ChatGPT's utility in thematic analysis [18]. AI's challenges with nuanced criteria such as policy implications corroborate [13] known limitations in interpretive tasks critical for health policy design, such as balancing clinical efficacy with ethical or logistical constraints (eg, insurance coverage gaps in Paper A). The ongoing need for human validation supports the framework proposed by Hitch [12], which positions AI as a "team member" rather than a standalone evaluator. This approach is reinforced by the importance of patient-centered

XSL•FO
RenderX

transparency in health care AI [32], where oversight mechanisms and impact on care experience directly influence trust.

Strengths and Limitations

The strengths of the study are as follows:

- Standardized protocols: The use of standardized protocols and independent verification of AI outputs helped mitigate potential bias, especially given the lead author's (AL) dual role as investigator and participant in Paper A.
- Diverse model selection: The inclusion of multiple AI models with varied architectures and capabilities facilitated a comprehensive analysis of AI's potential and limitations in qualitative research appraisal.
- Tool variety: The application of 3 widely recognized assessment tools (CASP, JBI, and ETQS) provided a robust framework for evaluating AI performance across different dimensions of qualitative research quality.

The limitations of the study are as follows:

- Proprietary model opacity: The proprietary nature of commercial models (GPT-3.5 and Claude 3.5) obscures the architectural factors driving their consensus patterns, potentially masking biases that disproportionately affect vulnerable populations (eg, Paper B's findings on physical education participation barriers).
- Dataset scope: The focused dataset of 3 health science papers limits generalizability, although the inclusion of pediatric and chronic disease contexts underscores current challenges for large language models in appraising life span–specific health narratives (Table 6).
- Author dual role: The lead author's (AL) involvement as both investigator and participant in Paper A introduced potential interpretation bias, mitigated but not eliminated by standardized protocols.
- Absence of human expert ratings: The lack of human expert ratings prevents definitive conclusions about whether AI's "favorable bias" reflects accuracy or systemic overestimation.

Conclusions

This study demonstrates that AI models exhibit both promise and limitations as evaluators of qualitative research quality. This comprehensive analysis revealed 3 critical insights: first, affirmation bias was evident, with "Yes" ratings ranging from 75.9% to 85.4% across models, highlighting AI's tendency to favor positive assessments, a pattern that could overstate the feasibility of interventions in health research. Second, model-specific variability emerged, as seen in GPT-4's divergent ratings, which lowered CASP agreement by 20% and underscored the influence of model architecture on appraisal consistency. Third, disagreements were often tool-dependent, particularly for ETQS criteria like policy implications and generalizability, exposing current limitations in AI's contextual interpretation.

The findings emphasize that AI cannot yet replace human judgment in nuanced qualitative appraisal but could enhance efficiency when strategically implemented. In health research, strong alignment of proprietary models (Cramer V=0.891) may expedite systematic reviews of patient experience studies, but their affirmation bias risks inflating confidence in underpowered qualitative evidence used for clinical guidelines. Open-source variability, while requiring oversight, could help counterbalance systemic optimism in AI-driven health syntheses.

Key limitations, including proprietary model opacity, which obscures biases affecting marginalized health populations, dataset scope constraints, and the author's dual role in Paper A warrant cautious interpretation. The absence of human expert ratings is particularly consequential for health research, where patient narratives and clinician insights require a nuanced ethical appraisal that AI's binary frameworks may oversimplify.

Future research should prioritize three areas: (1) health-specific AI training protocols emphasizing qualitative epistemology to better capture patient-centered care priorities, (2) benchmarking against expert panels to validate accuracy thresholds, and (3) establishing ethical frameworks for disclosing AI's role in health evidence synthesis, ensuring transparency in policy recommendations. As qualitative methodologies evolve alongside AI capabilities, the path forward lies not in human-machine competition but in hybrid workflows that leverage AI's scalability while preserving human expertise for contextual and interpretive tasks.

Acknowledgments

The author declared that they had insufficient funding to support open access publication of this manuscript, including from affiliated organizations or institutions, funding agencies, or other organizations. JMIR Publications provided article processing fee (APF) support for the publication of this paper. Artificial intelligence tools are cited for transparency; their inclusion does not imply endorsement.

Authors' Contributions

AL served both investigator and author in Paper A, a qualitative study included in the analysis. Standardized protocols and independent verification of artificial intelligence outputs mitigated potential bias.

Conflicts of Interest

This study used ChatGPT-3.5, ChatGPT-4 (OpenAI), Claude 3.5, Claude 3 Opus (Anthropic), and Sonar Huge (Perplexity AI) for structured quality assessments. The author is not employed by, holds stock in, or has received financial compensation from

XSL•FO
RenderX

these companies. AL holds no affiliation with JMIR journals, holds no patents related to this work, or has financial or personal relationships with individuals or organizations that could influence this research.

References

1. McKenna L. Translation of research interviews: do we have a problem with qualitative rigor? Nurse Author Ed 2022 Mar;32(1):1-3. [doi: 10.1111/nae2.31]
2. Pilbeam C, Anthierens S, Vanderslott S, Tonkin-Crine S, Wanat M. Methodological and ethical considerations when conducting qualitative interview research with healthcare professionals: reflections and recommendations as a result of a pandemic. Int J Qual Methods 2022 Apr;21:160940692210777. [doi: 10.1177/16094069221077763]
3. Pyo J, Lee W, Choi EY, Jang SG, Ock M. Qualitative research in healthcare: necessity and characteristics. J Prev Med Public Health 2023 Jan;56(1):12-20. [doi: 10.3961/jpmph.22.451] [Medline: 36746418]
4. Duden GS. Challenges to qualitative evidence synthesis—aiming for diversity and abstracting without losing meaning. Methods Psychol 2021 Dec;5:100070. [doi: 10.1016/j.metip.2021.100070]
5. Sebele-Mpofu FY. Saturation controversy in qualitative research: complexities and underlying assumptions. A literature review. In: Serpa S, editor. Cogent Soc Sci 2020 Jan;6(1):1838706. [doi: 10.1080/23311886.2020.1838706]
6. Noyes J, Booth A, Moore G, Flemming K, Tunçalp Ö, Shakibazadeh E. Synthesising quantitative and qualitative evidence to inform guidelines on complex interventions: clarifying the purposes, designs and outlining some methods. BMJ Glob Health 2019;4(Suppl 1):e000893. [doi: 10.1136/bmjgh-2018-000893] [Medline: 30775016]
7. Santiago-Delefosse M, Gavin A, Bruchez C, Roux P, Stephen SL. Quality of qualitative research in the health sciences: analysis of the common criteria present in 58 assessment guidelines by expert users. Soc Sci Med 2016 Jan;148:142-151. [doi: 10.1016/j.socscimed.2015.11.007]
8. Sanjari M, Bahramnezhad F, Fomani FK, Shoghi M, Cheraghi MA. Ethical challenges of researchers in qualitative studies: the necessity to develop a specific guideline. J Med Ethics Hist Med 2014;7:14. [Medline: 25512833]
9. Buetow S. Apophenia, unconscious bias and reflexivity in nursing qualitative research. Int J Nurs Stud 2019 Jan;89:8-13. [doi: 10.1016/j.ijnurstu.2018.09.013] [Medline: 30316055]
10. Christou P. How to use artificial intelligence (AI) as a resource, methodological and analysis tool in qualitative research? TQR 2023;28(7):1968-1980. [doi: 10.46743/2160-3715/2023.6406]
11. Al-Busaidi ZQ. Qualitative research and its uses in health care. Sultan Qaboos Univ Med J 2008 Mar;8(1):11-19. [Medline: 21654952]
12. Hitch D. Artificial intelligence augmented qualitative analysis: the way of the future? Qual Health Res 2024 Jun;34(7):595-606. [doi: 10.1177/10497323231217392] [Medline: 38064244]
13. Tai RH, Bentley LR, Xia X, et al. An examination of the use of large language models to aid analysis of textual data. Int J Qual Methods 2024 Jan;23:16094069241231168. [doi: 10.1177/16094069241231168]
14. Datt M, Sharma H, Aggarwal N, Sharma S. Role of ChatGPT-4 for medical researchers. Ann Biomed Eng 2024 Jun;52(6):1534-1536. [doi: 10.1007/s10439-023-03336-5] [Medline: 37526801]
15. van Dijk SHB, Brusse-Keizer MGJ, Bucsán CC, van der Palen J, Doggen CJM, Lenferink A. Artificial intelligence in systematic reviews: promising when appropriately used. BMJ Open 2023 Jul 7;13(7):e072254. [doi: 10.1136/bmjopen-2023-072254] [Medline: 37419641]
16. Alanazi A. Clinicians' views on using artificial intelligence in healthcare: opportunities, challenges, and beyond. Cureus ;15(9):e45255. [doi: 10.7759/cureus.45255]
17. De Angelis L, Baglivo F, Arzilli G, et al. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. Front Public Health 2023;11:1166120. [doi: 10.3389/fpubh.2023.1166120] [Medline: 37181697]
18. Lee VV, van der Lubbe SCC, Goh LH, Valderas JM. Harnessing ChatGPT for thematic analysis: are we ready? J Med Internet Res 2024;26:e54974. [doi: 10.2196/54974]
19. Senthil R, Anand T, Somala CS, Saravanan KM. Bibliometric analysis of artificial intelligence in healthcare research: trends and future directions. Future Healthc J 2024 Sep;11(3):100182. [doi: 10.1016/j.fhj.2024.100182] [Medline: 39310219]
20. Hannes K, Lockwood C, Pearson A. A comparative analysis of three online appraisal instruments' ability to assess validity in qualitative research. Qual Health Res 2010 Dec;20(12):1736-1743. [doi: 10.1177/1049732310378656] [Medline: 20671302]
21. Aydogdu ALF. Exploring different aspects of nursing leadership: an integrative review of qualitative studies. Mod Care J 2023;20(4). [doi: 10.5812/modernc-130402]
22. Gottvall M, Brunell C, Eldebo A, Johansson Metso F, Jirwe M, Carlsson T. Post-migration psychosocial experiences and challenges amongst LGBTQ+ forced migrants: a meta-synthesis of qualitative reports. J Adv Nurs 2023 Jan;79(1):358-371. [doi: 10.1111/jan.15480] [Medline: 36320151]
23. Jayachandran S, Hill K, Walmsley AD. A critical review of qualitative research publications in dental implants from 2006 to 2020. Clin Oral Implants Res 2021 Jun;32(6):659-671. [doi: 10.1111/clr.13743] [Medline: 33715249]
24. Moreno Gómez A, Guo P, de la Llave Rincón AI, Efstathiou N. Women's experiences of primary dysmenorrhea symptoms: a systematic review of qualitative evidence and meta-aggregation. Women Health 2023 Sep 14;63(8):658-668. [doi: 10.1080/03630242.2023.2255289]

XSL•FO
RenderX

25. Au A, Murad-Kassam S, Mukanoheli V, et al. Immigrant older adults' experiences of aging in place and their neighborhoods: a qualitative systematic review. Int J Environ Res Public Health 2024 Jul 10;21(7):904. [doi: 10.3390/ijerph21070904] [Medline: 39063481]
26. Flodén A, Broström EW, von Heideken J, et al. A qualitative study examining the validity and comprehensibility of physical activity items: developed and tested in children with juvenile idiopathic arthritis. Pediatr Rheumatol Online J 2019 Apr 25;17(1):16. [doi: 10.1186/s12969-019-0317-6] [Medline: 31023371]
27. Meier S, Höger B, Giese M. "If only balls could talk…": barriers and opportunities to participation for students with blindness and visual impairment in specialized PE. Front Sports Act Living 2023;5:1286909. [doi: 10.3389/fspor.2023.1286909] [Medline: 38162696]
28. Sikes EM, Richardson EV, Motl RW. A qualitative study of exercise and physical activity in adolescents with pediatric-onset multiple sclerosis. Int J MS Care 2019;21(2):81-91. [doi: 10.7224/1537-2073.2018-033] [Medline: 31049039]
29. Stickley T, O'Caithain A, Homer C. The value of qualitative methods to public health research, policy and practice. Perspect Public Health 2022 Jul;142(4):237-240. [doi: 10.1177/17579139221083814] [Medline: 35362352]
30. Bajwa J, Munir U, Nori A, Williams B. Artificial intelligence in healthcare: transforming the practice of medicine. Future Healthc J 2021 Jul;8(2):e188-e194. [doi: 10.7861/fhj.2021-0095] [Medline: 34286183]
31. Cresswell K, de Keizer N, Magrabi F, et al. Evaluating artificial intelligence in clinical settings—let us not reinvent the wheel. J Med Internet Res 2024 Aug 7;26(1):e46407. [doi: 10.2196/46407] [Medline: 39110494]
32. Stroud AM, Minteer SA, Zhu X, Ridgeway JL, Miller JE, Barry BA. Patient information needs for transparent and trustworthy cardiovascular artificial intelligence: a qualitative study. PLOS Digit Health 2025 Apr;4(4):e0000826. [doi: 10.1371/journal.pdig.0000826] [Medline: 40258073]

Abbreviations

AI: artificial intelligence
CASP: Critical Appraisal Skills Programme
ETQS: Evaluative Tools for Qualitative Studies
JBI: Joanna Briggs Institute

Edited by A Mavragani; submitted 18.02.25; peer-reviewed by S Mohanadas, VK nomula; revised version received 08.05.25; accepted 15.05.25; published 08.07.25.

Please cite as:
Landerholm A
AI in Qualitative Health Research Appraisal: Comparative Study
JMIR Form Res 2025;9:e72815
URL: https://formative.jmir.org/2025/1/e72815
doi:10.2196/72815

© August Landerholm. Originally published in JMIR Formative Research (https://formative.jmir.org), 8.7.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Evaluating a Large Language Model's Ability to Synthesize a Health Science Master's Thesis: Case Study

Pål Joranger, PhD; Sara Rivenes Lafontan, PhD; Asgeir Brevik, PhD

Department of Nursing and Health Promotion, Faculty of Health Sciences, OsloMet – Oslo Metropolitan University, P.O. Box 4 St. Olavs plass, Oslo, Norway

Corresponding Author:
Asgeir Brevik, PhD
Department of Nursing and Health Promotion, Faculty of Health Sciences, OsloMet – Oslo Metropolitan University, P.O. Box 4 St. Olavs plass, Oslo, Norway

Abstract

Background: Large language models (LLMs) can aid students in mastering a new topic fast, but for the educational institutions responsible for assessing and grading the academic level of students, it can be difficult to discern whether a text has originated from a student's own cognition or has been synthesized by an LLM. Universities have traditionally relied on a submitted written thesis as proof of higher-level learning, on which to grant grades and diplomas. But what happens when LLMs are able to mimic the academic writing of subject matter experts? This is now a real dilemma. The ubiquitous availability of LLMs challenges trust in the master's thesis as evidence of subject matter comprehension and academic competencies.

Objective: In this study, we aimed to assess the quality of rapid machine-generated papers against the standards of the health science master's program we are currently affiliated with.

Methods: In an exploratory case study, we used ChatGPT (OpenAI) to generate 2 research papers as conceivable student submissions for master's thesis graduation from a health science master's program. One paper simulated a qualitative health science research project and another simulated a quantitative health science research project.

Results: Using a stepwise approach, we prompted ChatGPT to (1) synthesize 2 credible datasets, and (2) generate 2 papers, that—in our judgment—would have been able to pass as credible medium-quality graduation research papers at the health science master's program the authors are currently affiliated with. It took 2.5 hours of iterative dialogue with ChatGPT to develop the qualitative paper and 3.5 hours to develop the quantitative paper. Making the synthetic datasets that served as a starting point for our ChatGPT-driven paper development took 1.5 and 16 hours for the qualitative and quantitative datasets, respectively. This included learning and prompt optimization, and for the quantitative dataset, it included the time it took to create tables, estimate relevant bivariate correlation coefficients, and prepare these coefficients to be read by ChatGPT.

Conclusions: Our demonstration highlights the ease with which an LLM can synthesize research data, conduct scientific analyses, and produce credible research papers required for graduation from a master's program. A clear and well-written master's thesis, citing subject matter authorities and true to the expectations for academic writing, can no longer be regarded as solid proof of either extensive study or subject matter mastery. To uphold the integrity of academic standards and the value of university diplomas, we recommend that master's programs prioritize oral examinations and school exams. This shift is now crucial to ensure a fair and rigorous assessment of higher-order learning and abilities at the master's level.

(JMIR Form Res 2025;9:e73248) doi:10.2196/73248

KEYWORDS

master's thesis; large language model; LLM; ChatGPT; health science; qualitative; quantitative

Introduction

Students represent a group that benefits from the use of large language models (LLMs). By using LLMs, students can gain knowledge on virtually any topic as well as edit or create scholarly text instantaneously [1-3]. In a survey from 2023, it was documented that 89% of college students in America used ChatGPT to complete their assignments, and 53% used the tool for writing papers [4]. Another report documents that LLMs perform on par with or better than university students across 32 university courses [5]. Although potentially beneficial for learning, the ubiquitous availability and use of LLMs challenge traditional methods of measuring the quality of students' academic performance. To obtain a master's degree, a written thesis is traditionally required. If theses, independent student research projects, or written home examinations can now be easily completed using LLMs, their role as a robust demonstration of higher-level abilities and learning might not be justified going forward.

XSL•FO
RenderX

Despite their already powerful capabilities, LLMs are still relatively new tools with major improvements yet to emerge, making them even more powerful in the future, autonomous executing AI agents being an example [3,6]. It may be argued that many corners of academia are already struggling to keep up with the pace of AI development [7], and more powerful tools can give students an even greater advantage against plagiarism detection attempts. While plagiarism detection tools are widely used at universities, it has been observed that traditional plagiarism detection tools may struggle to detect AI-generated text [8]. This raises concerns that faculty staff could be left with the unappealing task of grading a large number of almost similar student projects written largely by LLMs, falsely believing that they are assessing the student's academic ability. The value and status of university diplomas could be threatened by the widespread adoption of LLMs.

As health science lecturers and master's thesis censors at a Norwegian university, we believe there is an urgent need to acknowledge how much of a typical master's thesis can now be written by LLMs and what this means for assessing students' abilities in higher education. We were curious to test the ability of one of the most accessible LLMs to assist budding master's theses' authors in completing their graduation theses. Hence, we set out to briefly test the power of ChatGPT-4o to assist in the professional analysis and presentation of health science data, both in the realm of qualitative and quantitative data analysis, as well as the writing of an academic paper based on the results. Our objective was to assess the quality of machine-generated papers against the standards of the health science master's program we are currently affiliated with.

Methods

Study Design

In an exploratory case study, we asked ChatGPT (ChatGPT-4o and ChatGPT-o1) to synthesize two datasets: one qualitative and one quantitative, inspired by existing real-world datasets the authors had previously analyzed. The generation of synthetic datasets and subsequent data analysis took place in October and November of 2024.

Two prerequisites were kept during the development of the papers: (1) the process should be feasible for an average student who had completed the methodology courses in our local master's program, and (2) the process should be possible to complete fast, that is, during a long work session, guided entirely by chatbot dialogue.

Generation of Qualitative Data

We started by outlining the study aim and instructed ChatGPT-4o mini to prepare interview transcripts of 7 students with different levels of education, age, and gender. We also prompted ChatGPT to develop an interview guide (Multimedia Appendix 1). The first transcripts generated by ChatGTP4o mini were quite short, hence, in a follow-up prompt to ensure that each interview reflected 30 - 45 minutes of conversation, we requested interview transcripts that included longer and more detailed responses detailing specific events relevant to the study aim, including respondents' experiences and feelings. We also prompted for the artificial interviewer to ask follow-up questions, emulating a natural conversation. We prompted 1 interview transcript at a time to ensure that these aspects were included in each transcript. This resulted in 7 transcripts that were considered suitable for further analysis. The qualitative dataset is shown in Multimedia Appendix 2.

Generation of Quantitative Dataset

For the quantitative data, we used ChatGPT-o1 to develop a script that was run in ChatGPT-4o to produce the data matrix. To work with meaningful associations that resembled what could be found in international literature, we had to provide ChatGPT-o1 with information about how our synthetic dataset's 10 variables correlated with each other in pairs. To avoid inconsistencies in how the variables correlated as an overall system, we relied on a correlation matrix from real data for similar variables. All the correlation coefficients given to ChatGPT-o1 were discretionarily slightly changed (usually increased slightly) before admission, but without changing the direction (positive or negative) of the correlation. The quantitative dataset is shown in Multimedia Appendix 3. The fabrication of the quantitative dataset took 16 hours. This included the learning process, the time it took to create tables for reporting variables, estimate relevant bivariate correlation coefficients, and prepare these coefficients to be read by ChatGPT-4o.

Overall Approach to ChatGPT-4o–Assisted Generation of Papers

Conceptually our workflow advanced in the following logical sequence (with continuous prompt optimization based on the output from the LLM): (1) upload dataset, prompt the LLM to analyze and present results using specified method, (2) suggest methods chapter based on user predetermined context, (3) suggest introduction chapter, with emphasis on providing relevant scientific citations, (4) develop a discussion, with emphasis on integrating relevant scientific citations, (5) suggest abstract and title, and (6) combine elements to a final paper and update list of references. We opted for a "bite-sized" chapter-by-chapter approach to our ChatGPT-4o writing, because in our experience, ChatGPT-4o tends to respond to single prompts with a relatively limited amount of text output.

Preparation of Qualitative Paper

To conduct the data analysis and generate the research article, we prompted ChatGTP-4o mini to develop a qualitative analysis of the 7 transcripts, repeating the study aim and areas of interest. We also specified the method of qualitative data analysis that we wanted to use, thematic analysis by Braun and Clarke [9], and that 3 - 4 themes should be developed with 2 illustrative statements for each theme. Following ChatGTP-4o's qualitative data analysis output, we asked it to develop the results chapter for a paper for a scientific publication. The initial text output lacked sufficient detail, and an additional prompt was needed to request a more elaborate analysis. ChatGPT was then prompted to write an introduction, methods chapter, discussion, methodological discussion, and conclusion, as well as the abstract. ChatGPT was also asked to suggest suitable titles for the paper and state which one it recommended. After prompting

XSL•FO
RenderX

all the citations to the bottom of the text, we made a Microsoft Word (Microsoft Corp) text file of the complete paper. The file was then uploaded to ChatGPT with the prompt to suggest areas of improvement. A final set of prompts was necessary for the incorporation of the suggested improvements for each chapter of the paper (see Multimedia Appendix 4 for a full list of prompts for the development of a qualitative paper).

Preparation of Quantitative Paper

We started by informing ChatGPT-4o about the overarching research question. Having specified which variables were dependent variables, explanatory variables, and control variables, a multiple linear regression analysis was requested. We explained which parameters we wanted to have estimated and included in a table. Next, we prompted ChatGPT-4o to do a Spearman rank-correlation analysis where all the variables used above were correlated against each other 2 by 2 with corresponding r and *P* values. Finally, a descriptive analysis was requested. For all three analyses, we prompted ChatGPT-4o to set up the results in tables suitable for publication in scientific journals.

ChatGPT-4o was prompted to prepare the results chapter that reports the results of the 3 analyses with an emphasis on the multiple regression analysis. For the preparation of this chapter and for the subsequent ones, we instructed ChatGPT-4o to use the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) checklist [10] for preparing cross-sectional analyses. The checklist was uploaded as an attachment to the prompt. In separate dialogs, we asked ChatGPT-4o to prepare the methods chapter and the introductory chapter, respectively. We promoted ChatGPT-4o to discuss the methods and the results independently of each other. Conclusions were requested at the end.

We entered all the subbibliographies from individual chapters into a document and prompted ChatGPT-4o to remove duplicates, to check for authenticity, and to print the list in American Psychological Association style. In separate prompts, we requested ChatGPT-4o to prepare a summary and to suggest a title for the paper. The various parts were combined manually into 1 paper. Lastly, ChatGPT-4o was prompted to suggest improvements to its own paper. Based on ChatGPT-4o's suggestions for improvements, we instructed ChatGPT-4o to refine the paper, chapter by chapter, and at the end, the summary. The various optimized parts were edited together manually into a complete paper. All results from the bi- and multivariate analyses performed by ChatGPT-4o were tested against analyses in SPSS (version 28; IBM Corp) and StataMP (version 18; StataCorp LLC). The full list of prompts used for the quantitative paper is shown in Multimedia Appendix 5.

Ethical Considerations

No human research was included in this study; hence, approval from an institutional review board or regional ethics committee was not required. The research datasets, on which the evaluation was based, were fabricated using ChatGPT to avoid potential conflicts of interest with real datasets. No students were involved in this project. The evaluation was carried out for illustrative purposes, to show the emerging post-LLM challenge of measuring acquired competence among university students.

Results

Overview

Our results consist of 2 full-length scientific papers (see Multimedia Appendices 6 and 7 for the qualitative and quantitative papers, respectively). In the following we present our observations of the strengths and weaknesses of the papers developed for our particular use, that is, to emulate a health science master's student's research paper submitted for graduation, followed by a short presentation of our judgement of the final full-size papers, as well as estimates of total time spent on dataset and paper development.

Examination and Assessment of the Qualitative Paper

The qualitative paper produced contained the key elements expected of a scientific article (Multimedia Appendix 6). It was concise, with 4529 words. The thematic analysis reflected the study aim, with a clear identification of themes and relevant illustrative quotes included for each theme. The generated paper provided a logical basis for discussion, and the abstract followed standard conventions, offering an adequate summary of the main findings and included reference to Braun and Clarke 6-phase analytical method. The paper presented 4 main themes—motivation to mentor, personal and professional growth, challenges faced, and the need for enhanced training. However, the paper had some weaknesses. The thematic analysis lacked depth in exploring the various experiences of the "participants," and many interpretations were superficial. For example, the theme "motivation to become a mentor" highlights the participants' desire to support others and develop leadership skills while largely reiterating generic motivations without exploring individual nuances or cultural influences on these motivations. There was also repetition and an overlap in certain quotes. While there were some references to existing literature, the use of references was limited, and as such, the paper failed to connect findings to broader evidence-based scientific discussion. Additionally, of 11 references, 7 were suspected to have been fabricated by ChatGPT-4o. Despite the weaknesses, the paper was assessed to have received a pass grade according to the university's censor guidelines. That said, it is doubtful it would have passed if the fabricated references were discovered by the censors.

The generation of the paper itself was both fast and efficient. While the total time spent generating both the synthesized data and the paper was approximately 2.5 hours, of which approximately 1.5 hours were used to generate the synthesized data as it required several prompts to ensure detailed reflections on experiences and emotional states.

Examination and Assessment of the Quantitative Paper

The resulting paper (Multimedia Appendix 7) includes the main elements one would expect and generally aligns with the STROBE guidelines. It is relatively concise, with 2930 words excluding the abstract, tables, and references. ChatGPT-4o performed correlation analyses and multiple linear regression analyses, but the multiple linear regression did not include robust

XSL•FO
RenderX

SE adjustments. Chapter coherence and focus on the research question are considered to be satisfactory. Relevant discussions on methodological challenges and comparisons with existing literature are present and seem genuine, while the existence of 1 publication could not be confirmed. The abstract followed the standard thematic structure and provided a satisfactory summary.

One area where ChatGPT-4o failed was in conducting multiple regression analyses based on robust SEs. After the initial multiple regression analyses were performed, we asked ChatGPT-4o to check if the assumptions for these analyses were met. According to ChatGPT-4o, they were met except for potential issues with the assumption of homoscedasticity. When we asked ChatGPT-4o how this issue could be resolved, it provided 6 different methods to address the problem. We chose the first method suggested, which was to conduct multiple regression analyses based on robust SEs. We then prompted ChatGPT-4o to perform these analyses, and the transformer produced new estimates for the relevant parameters, explicitly stating, "here are the results from the multiple regression analysis with robust standard errors summarized in a table." Surprisingly, the numbers reported in this table were identical to the numbers in the original table without robust SEs.

Further miscalculations by ChatGPT-4o were exposed by the fact that 9 of the 28 bivariate correlation analyses (Spearman) were incorrect. This can be observed by comparing Table 3 with the Spearman correlation matrix in Multimedia Appendix 8, estimated using SPSS. All bivariate analyses between life satisfaction and the other variables appear to have been correctly calculated by ChatGPT-4o. We checked whether the 9 discrepancies aligned with the corresponding bivariate correlation analyses performed using Pearson rather than Spearman correlation, but this was not the case.

Assuming the examiners do not conduct their own analyses and therefore do not detect the errors in the "robust" analyses and the miscalculations of Spearman correlations, we assess that this paper would pass as part of a master's thesis. At the university we are affiliated with, it is not common practice for examiners to perform their own control analyses based on students' data. However, if the examiners did indeed perform their own calculations, and discovered the error, we believe the fictional student could still pass, albeit with a lower grade. Readers can make their own assessments of the quality of the paper, which is attached as Multimedia Appendix 7.

A total of approximately 3.5 hours was spent preparing the paper, including verifying the authenticity of the references. This excludes the time spent creating the synthetic data and the time required to prepare the table of variables used (16 h). The simulated situation assumed that the student was using secondary data, with an accompanying description of included variables. We excluded the time spent verifying calculations done in SPSS and Stata, as we assumed that the imaginary students relied solely on the analyses performed by ChatGPT-4o.

For both the quantitative and qualitative analyses, we have emphasized being as transparent as possible by including in the appendices the data used, the prompts, and the paper produced by the LLM. This allows readers to both create similar articles themselves and to make their own assessment of the papers' quality against their university's requirements for a master's thesis.

Discussion

Principal Findings

This case study evaluates how easy it is to use ChatGPT-4o to rapidly develop credible scientific papers that comply with current standards and expectations for a master's thesis. The machine-generated example papers that were developed in a single work session did not represent elite students' highest conceivable potential, but we rate them sufficiently to score a pass grade at the health science master's program the authors are currently affiliated with. We also demonstrated that it is possible to use ChatGPT (ChatGPT-o1 and ChatGPT-4o) to synthesize credible research data. It was surprisingly easy to generate research data through dialogue with ChatGPT, in particular, qualitative data. With a couple of additional steps and prompts, we were also able to generate a quantitative dataset. This synthetic quantitative dataset appeared to possess sufficient complexity and internal consistency to be useful as an open dataset for illustrating our experiments with ChatGPT-4o. However, readers who conduct a closer examination of the data may observe that several variables exhibit distributions that are not entirely realistic. The fact that LLMs can be used to falsify or fabricate research data, and the associated potential for academic misconduct, has been recognized by other authors [11,12].

It could be argued that collectively the authors possess more research experience than a typical master's student, and as such are better equipped to compose effective prompts for the fabrication of research data and master's thesis elements. Clearly, an average student starting from scratch, with limited experience, would likely take longer to arrive at useful results. At the same time, one would expect that students representing a generation well-versed in online information retrieval would be able to quickly acquire the know-how of generating a master's thesis using LLMs. One can easily envision a predefined, standardized set of LLM prompts designed for the efficient generation of master's theses. In an evolving "arms race," emerging AI agents could further enhance students' ability to circumvent university antiplagiarism systems.

ChatGPT-4o produced shorter text outputs than what is required for the specific sections of a research paper. This tendency of ChatGPT-4o to produce well-written superficial answers that fail to capture the level of factual detail, terminology, and nuance that is often expected in academia has been observed by other authors [13]. By issuing more specific prompts, with clear expectations for context and output, it was usually possible to extract a lot more detail than what was generated from the initial prompts. Prompt engineering for academic writers has been covered by Giray [14]. ChatGPT-4o sometimes generated false references when instructed to provide references supporting its claims. In our case study, the tendency to create false references was more pronounced in the generation of the qualitative research paper than in the quantitative paper. We do not know what caused this difference. The tendency to, despite prompts, omit scientific citations and generate nonexistent

XSL•FO
RenderX

references has been reported by other authors [12,13,15]. A recent report looking specifically at hallucination rates documented high hallucination rates and reference inaccuracies in ChatGPT-3.5 and ChatGPT-4 and Bard (now Google Gemini; Google LLC) and recommended against using LLMs as a primary or exclusive tool for systematic reviews [16]. High hallucination rates are corroborated by the HALoGEN project, which looked at hallucination rates spanning 9 domains across 14 LLMs and suggested a new classification system for the study of LLM hallucinations, which often can be difficult to explain [17]. ChatGPT-4o's suggested references could be compromised in two ways: (1) they could be nonexistent, and (2) they could be real but not relevant or at least not the ones that would have been chosen by the human expert in the specific field. While it is easy to check for the existence of cited references checking for the quality and relevance of the cited references requires a modicum of expertise. There is no question that in-depth knowledge of the subject matter at hand, including the methods used, would represent an advantage for confirming the relevance of the cited references. Unfortunately, due to budgetary constraints and increased student volume, master's thesis censors in our context are not necessarily subject matter experts, often unable to verify each citation of the many master's theses they are required to evaluate. Hence, suboptimal references could in many instances be expected to slip through unnoticed and fail to draw the critical remarks they deserved in an examination situation. In the ChatGPT-4o–assisted quantitative paper, we manually checked the accuracy of the statistical calculations performed by ChatGPT-4o. Although the mistakes did not stand out as obvious and easily detectable, we did, on closer scrutiny, discover some errors and irregularities. The errors were not grave enough to set off an immediate red flag and would likely go undetected should the censors not perform their own control calculations. Nevertheless, our experience illustrates that research integrity may endure overreliance on LLMs in their current state.

One argument that is made against the prospects of an LLM-assisted streamlined master's thesis expressway is that the student's supervisor would surely detect any attempts at cheating. This belief may be based on how things used to be in a bygone era. In a modern world where technological tools increasingly allow for economically efficient online-based master's programs, daily contact between faculty staff and graduate students is no longer guaranteed. With increasingly limited oversight, it might be tempting for the graduate student to rely heavily on AI assistance to complete their academic tasks, especially when the universities' antiplagiarism tools are unable to detect AI generated text [18]. Furthermore, should heavy reliance on LLMs be considered cheating at all, or is it just a new power tool for all students to enjoy unrestrictedly, as they would be able to do in a real-life work situation? While the challenges related to student's use of LLMs by now are well-known in higher education [5,19,20], we have been unable to identify clear solutions from the current literature to what academic institutions can put up as a defense against LLM-assisted cheating or plagiarism. Master's programs still rely heavily on thesis-based assessment; some exclusively rely on thesis-based assessment. Based on our case study, demonstrating how easy it is to cheat by using heavy AI assistance in the development of a master's thesis, we believe that the assessment of academic achievement should no longer be based purely on the student's written work. An oral examination is often conducted after the submitted master's thesis to determine an overall grade for the thesis. Traditionally, at least at our local university, the written thesis has been more heavily weighed than the student's performance at a subsequent oral defense. The term "adjusting oral defense" has been used in-house by faculty staff and the university administration, implying that only minor adjustments, such as one grade up or down, are to be expected following oral examination. As a perhaps unfortunate but necessary consequence of the LLM revolution, we suggest flipping the script on this practice. When it comes to grading the student's academic achievement and maturity, placing more emphasis on the results of the oral examination of the student than on the student's submitted thesis now makes sense. After all, if universities are not able to enforce correctness, and what they consider to be the most ethical student use of LLMs, it might be argued that the reasonable thing to do is to lift most if not all restrictions, and instead rely primarily on oral examinations where students are allowed to demonstrate deep knowledge of their subject, scientific reasoning, the relevant analytical steps of their analysis, as well as the inherent strengths and weaknesses of the methods used in their thesis. This switch does not have to incur extra expenses on an often-challenged university economy. Evaluator resources spent on repeated reading of students' written texts could be better spent on broad oral examinations. Oral examinations could include questions from the student's written master's thesis, but also a set of thesis-independent questions on scientific methods that required less preparation from the examiner.

One limitation of our ChatGPT-4o–generated master's paper assessment is that the finished products were neither put through a proper peer assessment nor submitted to the university examination system as sham student projects. As master's program lecturers, supervisors, course leaders, and master's thesis examiners, we believe we are capable of assessing whether the fake papers would have been able to obtain a pass grade. A master's thesis in our system contains more than just the scientific paper that we have used as an example in our case study. A short introduction usually precedes the paper. However, we believe that the process of developing scientific papers is relevant to illustrate how easy it is to produce scientific work at a high level using LLMs. We only made 2 example papers. More papers using a diverse set of analyses could perhaps have revealed relevant nuances between analytical methods. Given the highly competitive field of AI tools, we could have tested more than 1 LLM. Further prompt optimization could have reduced the number of false references and prompted the use of the COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist [21] or a similar tool, which could have further enhanced the quality of the qualitative paper.

Conclusions

While the universally available LLMs constitute a family of extraordinarily capable new software tools that promise to revolutionize scientific productivity, they will also be of great help for students at all levels. An emerging realization is that the ubiquitous use of LLMs will pose a challenge for academic

XSL•FO
RenderX

institutions entrusted with the task of ensuring a certain level of acquired knowledge by graduating students and issuing diplomas based on demonstrated academic achievements. In this study, we have demonstrated that LLMs can produce credible student master's thesis papers at a fraction of the time normally allocated to the completion of a thesis submitted for graduation. We also demonstrate that LLMs can create entirely new artificial datasets with relative ease when modeled upon previously published research.

Traditionally, universities have relied heavily on a written academic thesis to evaluate academic achievement and issue grades. With the introduction of LLMs, capable of generating credible scientific texts, academic institutions now risk issuing grades based on machine-generated output. Based on our role-playing experience as hypothetical busy student actors looking for a heavily AI-assisted fast track to a master's thesis, we recommend against pure text-based ascertainment of academic achievement for graduate students. Increased emphasis on oral examinations for grade-defining graduation work is a logical consequence of a looming LLM caused disruption of the ingrained methods of assessing students' performance in higher education.

Data Availability

All data generated or analyzed during this study are included in this published article [and its supplementary information files].

Authors' Contributions

PJ and AB conceptualized this study. PJ, SRL, and AB performed the methodology, investigation, data curation, writing of the original draft, and review and editing of the writing. All authors have read and agreed to the published version of this paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview guide from the qualitative study.
[DOCX File, 27 KB - formative_v9i1e73248_app1.docx]

Multimedia Appendix 2
Qualitative dataset.
[DOCX File, 85 KB - formative_v9i1e73248_app2.docx]

Multimedia Appendix 3
Quantitative dataset.
[XLSX File, 59 KB - formative_v9i1e73248_app3.xlsx]

Multimedia Appendix 4
Prompts used for the qualitative study paper.
[DOCX File, 27 KB - formative_v9i1e73248_app4.docx]

Multimedia Appendix 5
Prompts used for the quantitative study paper.
[DOCX File, 30 KB - formative_v9i1e73248_app5.docx]

Multimedia Appendix 6
Qualitative paper.
[DOCX File, 38 KB - formative_v9i1e73248_app6.docx]

Multimedia Appendix 7
Quantitative paper.
[DOCX File, 54 KB - formative_v9i1e73248_app7.docx]

Multimedia Appendix 8
Spearman correlations.
[DOCX File, 93 KB - formative_v9i1e73248_app8.docx]

XSL•FO
RenderX

References

1. Meyer JG, Urbanowicz RJ, Martin PCN, et al. ChatGPT and large language models in academia: opportunities and challenges. BioData Min 2023 Jul 13;16(1):20. [doi: 10.1186/s13040-023-00339-9] [Medline: 37443040]
2. Patil R, Gudivada V. A review of current trends, techniques, and challenges in large language models (LLMs). Appl Sci (Basel) 2024 Mar 1;14(5):2074. [doi: 10.3390/app14052074]
3. Wang S, Xu T, Li H, et al. Large language models for education: a survey and outlook. arXiv. Preprint posted online on Mar 26, 2024. [doi: 10.48550/arXiv.2403.18105]
4. Yu H. Reflection on whether Chat GPT should be banned by academia from the perspective of education and teaching. Front Psychol 2023;14:1181712. [doi: 10.3389/fpsyg.2023.1181712] [Medline: 37325766]
5. Ibrahim H, Liu F, Asim R, et al. Perception, performance, and detectability of conversational artificial intelligence across 32 university courses. Sci Rep 2023 Aug 24;13(1):12187. [doi: 10.1038/s41598-023-38964-3] [Medline: 37620342]
6. Pachegowda C. The global impact of AI-artificial intelligence: recent advances and future directions. a review. arXiv. Preprint posted online on Dec 22, 2023. [doi: 10.48550/arXiv.2401.12223]
7. Strzelecki A. 'As of my last knowledge update': how is content generated by ChatGPT infiltrating scientific papers published in premier journals? Learn Publ 2025 Jan;38(1):e1650. [doi: 10.1002/leap.1650]
8. Santra PP, Majhi D. Scholarly communication and machine-generated text: is it finally AI vs AI in plagiarism detection? SRELS 2023 Jun 30:175-183. [doi: 10.17821/srels/2023/v60i3/171028]
9. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
10. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol 2008 Apr;61(4):344-349. [doi: 10.1016/j.jclinepi.2007.11.008] [Medline: 18313558]
11. Taloni A, Scorcia V, Giannaccare G. Large language model advanced data analysis abuse to create a fake data set in medical research. JAMA Ophthalmol 2023 Dec 1;141(12):1174-1175. [doi: 10.1001/jamaophthalmol.2023.5162] [Medline: 37943569]
12. Hosseini M, Rasmussen LM, Resnik DB. Using AI to write scholarly publications. Account Res 2024 Dec 6;31(7):715-723. [doi: 10.1080/08989621.2023.2168535] [Medline: 36697395]
13. Reis F, Lenz C, Gossen M, Volk HD, Drzeniek NM. Practical applications of large language models for health care professionals and scientists. JMIR Med Inform 2024 Sep 5;12:e58478. [doi: 10.2196/58478] [Medline: 39235317]
14. Giray L. Prompt engineering with ChatGPT: a guide for academic writers. Ann Biomed Eng 2023 Dec;51(12):2629-2633. [doi: 10.1007/s10439-023-03272-4] [Medline: 37284994]
15. Salvagno M, Taccone FS, Gerli AG. Artificial intelligence hallucinations. Crit Care 2023 May 10;27(1):180. [doi: 10.1186/s13054-023-04473-y] [Medline: 37165401]
16. Chelli M, Descamps J, Lavoué V, et al. Hallucination rates and reference accuracy of ChatGPT and Bard for systematic reviews: comparative analysis. J Med Internet Res 2024 May 22;26:e53164. [doi: 10.2196/53164] [Medline: 38776130]
17. Ravichander A, Ghela S, Wadden D, Choi Y. HALoGEN: fantastic LLM hallucinations and where to find them. arXiv. Preprint posted online on Jan 14, 2025. [doi: 10.48550/arXiv.2501.08292]
18. Pudasaini S, Miralles-Pechuán L, Lillis D, Llorens Salvador M. Survey on AI-generated plagiarism detection: the impact of large language models on academic integrity. J Acad Ethics 2024. [doi: 10.1007/s10805-024-09576-x]
19. Yan L, Sha L, Zhao L, et al. Practical and ethical challenges of large language models in education: a systematic scoping review. Brit J Educational Tech 2024 Jan;55(1):90-112. [doi: 10.1111/bjet.13370]
20. Jiao J, Afroogh S, Chen K, Atkinson D, Dhurandhar A. The global landscape of academic guidelines for generative AI and large language models. arXiv. Preprint posted online on May 26, 2024. [doi: 10.48550/arXiv.2406.18842]
21. Tong A, Sainsbury P, Craig J. Consolidated Criteria for Reporting Qualitative Research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]

Abbreviations

COREQ: Consolidated Criteria For Reporting Qualitative Research
LLM: large language model
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

XSL•FO
RenderX

Edited by A Mavragani; submitted 28.02.25; peer-reviewed by R Vergaz, S Karekar; revised version received 13.05.25; accepted 15.05.25; published 03.07.25.

Please cite as:
Joranger P, Rivenes Lafontan S, Brevik A
Evaluating a Large Language Model's Ability to Synthesize a Health Science Master's Thesis: Case Study
JMIR Form Res 2025;9:e73248
URL: https://formative.jmir.org/2025/1/e73248
doi:10.2196/73248

© Pål Joranger, Sara Rivenes Lafontan, Asgeir Brevik. Originally published in JMIR Formative Research (https://formative.jmir.org), 3.7.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Co-Design of a Health Screening Program Fact Sheet by People Experiencing Homelessness and ChatGPT: Focus Group Study

Nóra Radó[1,2*], PhD; Orsolya Németh[3,4*], PhD; Sándor Békási[1,2*], MD

[1]DocRoom Health Research Program, Hungarian Charity Service of the Order of Malta, 28 Bem Rakpart, Budapest, Hungary
[2]Institute of Behavioural Sciences, Faculty of Medicine, Semmelweis University, Budapest, Hungary
[3]Department of Community Dentistry, Faculty of Dentistry, Semmelweis University, Budapest, Hungary
[4]Centre for Translational Medicine, Semmelweis University, Budapest, Hungary
[*]all authors contributed equally

Corresponding Author:
Nóra Radó, PhD
DocRoom Health Research Program, Hungarian Charity Service of the Order of Malta, 28 Bem Rakpart, Budapest, Hungary

Abstract

Background: People experiencing homelessness have worse oral health outcomes and a notable health informational asymmetry compared to the general population. Screening programs present a viable option for this population; however, barriers to access, such as lower levels of health literacy, lack of information, and mistrust, narrow their chances to participate in such programs.

Objective: The aim of this study is to investigate the applicability of generative artificial intelligence (AI) in designing a homeless health screening program fact sheet with experts by experience using co-design principles.

Methods: Six fact sheet text variants were created by the open-access version of ChatGPT 3.5 for an oral cancer screening program targeting people experiencing homelessness in Budapest, Hungary. Clients of homeless social services (N=23) were invited to a short questionnaire survey and 3 semistructured focus group discussions between May and July 2024. General opinions regarding generative AI technology and direct feedback on the text variants were obtained. Additionally, a standardized readability assessment of the text variants was completed via the Sydney Health Literacy Lab Editor.

Results: Almost two-thirds of participants (17/23) stated that they had previously heard about AI; however, their self-assessment regarding the extent of their knowledge resulted in an average of 2.38 (n=16) on a 5-point Likert scale. During the first focus group discussion, all 6 variants received a high score (between 4.63 and 4.92 on a 5-point Likert scale). In the next sessions, participants remained positive when the pool was narrowed to 4 versions, although they scored the texts lower. During open discussions, text variants were considered understandable, while difficulties with medical expressions, lengthiness of sentences, and references to a stereotypical homeless subgroup (rough sleepers) were also reported. The health literacy editor showed that most AI-generated text variants were difficult to read and too complex for the target group.

Conclusions: The co-design process revealed that focus group participants actively wanted to shape the fact sheet drafts. They shared their insights on how to make the text variants more appealing for the target audience. Moreover, the involvement of generative AI technology revealed that the participants have heard about the concept of AI and text generation as a potential function, and they have not rejected its use in health care settings.

(JMIR Form Res 2025;9:e68316) doi:10.2196/68316

KEYWORDS

digital health; co-design; oral health; focus group; expert by experience; ChatGPT; health information; leaflet; screening; mixed method study; screening programs; oral cancer; vulnerable population; generative AI; co-design; Hungary; homeless; qualitative method; quantitative method; artificial intelligence

Introduction

Homelessness and Oral Health

Homelessness is a complex social phenomenon that leaves individuals for a shorter or longer period in an extremely vulnerable life situation. According to previous research, homelessness is associated with significantly higher disease burden [1-3] and higher mortality rates for both women and men than the average population [4]. In Western, high-income countries, studies have also shown that homelessness is an independent risk factor for mortality, and life expectancy varies between 50 - 65 years on average [5].

Previous research on the oral health of people experiencing homelessness found that this population has poor outcomes; they are in great need of restorative, oral hygiene, and periodontal treatment. They also have inadequate access to

XSL•FO
RenderX

dental services, mostly relying on emergency treatment, in parallel with unmet treatment needs [6-8]. In the United Kingdom, dental health was identified as this group's largest unmet health need [9]. In the United States, a national study of homeless adults using Health Care for the Homeless services found that approximately half of homeless adults had an unmet need for dental care [10]. Higher rates of substance use (alcohol, tobacco, drugs) further put the oral and general health of people experiencing homelessness at risk [11]. Freitas et al [10] found strong associations between having lost half or more of their teeth and evidence of problem drinking, cocaine use, or having ever smoked. In 1997, in Hungary, precancerous lesions were found in or around the oral cavity in 14% of people experiencing homelessness or participating in alcohol withdrawal treatment, benign tumors in 2.33%, and malignancies in 2.66% [12].

Access to oral care also comes with serious barriers for this population; the cost of care for private service providers, lengthy waiting lists for publicly funded institutions, competing priorities (which might lead them to secure food and accommodation before health care), a lack of information, mistrust of health care systems, and experiences of discrimination in care settings all drive people experiencing homelessness away from dental care services, resulting in them needing to rely on emergency treatment in cases of acute problems [11,13-15]. Moreover, psychosocial factors play a significant role; higher levels of dental anxiety and dental phobia were found in the homeless adult population [16].

Screening Programs, Health Literacy, and Information Asymmetry

As literature shows, the potential implications of a health screening program in dental practice are reductions in morbidity, mortality, and onward cost to health care systems by avoiding acute presentations of late-stage chronic diseases [17]. Moreover, Nunez et al [18] found that in the United States, veterans who received dental care were found to stay in homeless intervention programs significantly longer than veterans who did not. Their findings also indicated that the impact of the provision of dental care on outcomes among homeless veterans is equivalent to the impact of psychological treatments for depression.

To overcome the barriers to dental care for people experiencing homelessness in Hungary, the Charity Service of the Order of Malta, in collaboration with Semmelweis University and Óbuda University, launched an oral cancer screening program with digital capabilities in Budapest in 2024. The initiative fits into the wider digital health research agenda of the Charity Service, which previously completed numerous digital health projects [14,19-21]. Using advanced asynchronous telecare solutions in this vulnerable community, the new digital platform Lesionwizard was designed to deliver an oral cancer screening program for people experiencing homelessness using teledentistry [22].

As an additional barrier, a lack of information seriously burdens vulnerable populations. One of the main problems is information asymmetry between providers and people experiencing homelessness, coupled with lower levels of (oral) health literacy. In our previous study in collaboration with the Digital Health Working Group at Semmelweis University, Budapest, Hungary, we found that difficulties in gaining reliable information from service providers might result in the phenomenon that people experiencing homelessness look up medical information online or turn to alternative sources [14]. Csikar et al [23] also identified the level of (oral) health literacy as a barrier for people experiencing homelessness who had difficulties understanding letters sent to them. The authors concluded that it impacted their prioritization of oral health, as individuals may have yet to understand the importance of oral care or their options for accessing it.

The Application of Co-Design and Generative Artificial Intelligence

To facilitate participation in our oral cancer screening program, the research team decided to aid the initiative with an A5-format, awareness-raising, short health information fact sheet that presents the initiative as acceptable, available, and effective for this vulnerable population [13]. Co-design principles and the technological assistance of the generative artificial intelligence (AI) tool ChatGPT (OpenAI) were applied.

Co-design has previously been defined as a participatory approach that brings individuals together to collaborate and combine their knowledge, skills, and resources to accomplish a design task [24], also in the area of digital health for tool, educational, and health information material development [24-28]. It involves the meaningful engagement of end users recognized as experts by experience [29]. Previous research found that co-design, co-creation, or co-production can be empowering for socially marginalized or excluded groups, such as people experiencing homelessness, while it is also a pivotal approach to tackling stigmatization and promoting inclusivity. Co-design techniques resulted in increased applicability and acceptance of research questions, outputs, participant engagement, and knowledge of different contexts, as well as an improved community network for the researchers [30].

Generative AI software, such as ChatGPT, is a large language model (LLM) combined with a user-friendly interface that uses deep learning algorithms trained on vast amounts of data to generate multimodal humanlike responses to user prompts [31]. Its applicability in medicine is currently under scrutiny, but it has great promise in aiding doctor-patient communication and providing patient information. It has performed satisfactorily in answering physician-generated medical queries across 12 distinct specialties [32]. It has also been shown to simplify online health information [33], to generate dermatologic patient education materials according to specific reading levels [34], and to translate patient education materials from English into other languages [35].

In this research project, we aimed to co-design an awareness-raising fact sheet for an oral cancer screening program with people experiencing homelessness as experts by experience and ChatGPT. The latter was used to present textual alternatives for this health information piece, so we could also test the usability of ChatGPT in designing adequate information materials serving the needs of people experiencing homelessness.

XSL•FO
RenderX

Methods

Participants and Recruitment Procedure

The study followed the Consolidated Criteria for Reporting Qualitative Research (COREQ) checklist, adapted to focus groups [34] (see Checklist 1). Three focus group discussions were organized to provide feedback regarding patient information materials for an oral cancer screening program. One of them was an already existing group of experts by experience; in addition, two ad hoc groups were formed from clients of 3 shelters in Budapest, Hungary (Miklós Street Integrated Homeless Care Center, Homeless Care Center at Bem rakpart, and Galvani Street Homeless Care Center), operated by the Hungarian Charity Service of the Order of Malta. The sample constituted a convenience sample; the researchers advertised ad hoc focus groups in the shelters, and clients over 18 years without mental health problems and dementia who expressed their interest participated voluntarily, without any compensation.

The experts by experience group was established in 2023 to assist in co-designing initiatives targeting relevant health issues of people experiencing homelessness. Expert group meetings were organized on a monthly schedule with the attendance of 6 - 9 experts. The option to participate in the experts by experience group was open to adult clients (>18 years) of homeless shelters operated by the Hungarian Charity Service of the Order of Malta, without mental health problems or dementia.

From the recruited sample (N=26), three people decided not to participate (2 people due to scheduling problems and 1 person due to the difficulty of the topic). The number of participants in the 3 focus groups was 6, 10, and 7, and the demographic characteristics are shown in Table 1. The focus group discussions took place on May 16, June 4, and July 4, 2024; their length varied between 40 and 55 minutes.

Table . Demographic characteristics of the focus groups discussing ChatGPT-generated text variants. Participants were clients of 3 homeless shelters operated by the Hungarian Charity Service of the Order of Malta.

Group	Age (years), mean (SD)	Gender, female (male)
Experts by experience (n=6)	55.83 (14.97)	1 (5)
Focus group 2 (n=10)	61.50 (7.11)	2 (8)
Focus group 3 (n=7)	53.57 (5.19)	0 (7)

Text Generation

Six text variants of basic client information materials were generated on May 13, 2024, by the open-access version of ChatGPT 3.5 developed by OpenAI. The researchers chose OpenAI's most advanced freely available product because, according to statistics, it is the most widely available [36]. Prompts were applied in English, while the results were given in Hungarian. Each text version was limited to a word count of 150 due to the limitations of an A5-size one-sided fact sheet. All prompts emphasized the target population (people experiencing homelessness), the main aim of the text (to raise the level of participation), and a reasoning or style/tonal requirement. These requirements were the following: (1) scientific evidence regarding oral cancer, (2) statistical evidence regarding oral cancer, (3) as motivating as possible, (4) based on an informal, familiar tone, using slang expressions, (5) formatted as a clickbait news article, and (6) structured in a bullet-point format. Otherwise, the prompts were formulated as plain texts produced by people without relevant expertise in prompt design, as the researchers had the intention to involve ChatGPT as a tool that would be used by nonexpert social sector users. The prompts used in this study and the resulting Hungarian text variants, as well as the texts translated into English, are provided in Multimedia Appendix 1.

Feedback Questionnaires and Semistructured Group Discussions

A 2-part short feedback questionnaire developed by the research team was used to quantify different aspects of AI in general and AI-generated text variants, and it was also used to catalyze an open group discussion. The first part consisted of three items: (1) whether the participants have heard about AI (in a Yes or No scheme), (2) self-assessment of knowledge regarding AI technology (on a 5-point Likert scale), (3) and trust in its use in health care settings (on a 5-point Likert scale). The second part included each text variant with 7 items. Assessment of understandability and clarity, the quality of information content, the tone and style of the texts, and the convincing factor were conducted on a 5-point Likert scale. Lastly, 3 open questions inquired about the strengths and weaknesses of the texts, any changes suggested, and the applicability of the texts in the screening program. The quantified values were obtained in paper and pencil form, while answers to open questions were discussed by the group members, and notes were taken by the research team.

Text Evaluation via the Sydney Health Literacy Lab Health Literacy Editor

After the focus groups analyzed the text variants, we also assessed the texts according to standardized readability measurement tools. There are several methods to calculate the readability scores of texts, such as the Flesch-Kincaid method [37], the Gunning fog index [38], or the SMOG readability formula [39]. The third method is frequently used in health research [40]. In this study, we used the framework by Ayre et al [33] entitled the Sydney Health Literacy Lab (SheLL) Health Literacy Editor as it is a web-based tool designed to objectively assess the extent to which health information is written in plain language, while all the other methods serve as general tools for readability measurement. The SHeLL Editor, available as a web-based tool [41], assesses the number of words, readability

XSL•FO
RenderX

as grade reading score, language complexity, passive voice usage, and the use of bullet points for lists [33]. Based on this framework, we made the first 4 assessments and left out bullet points for lists as they only appeared in 1 text variant. The text assessments were then compared with the focus group assessment.

Ethical Considerations

Participation in the focus group discussion was voluntary and without any compensation. Data collection from the questionnaires was anonymous, and notes from the focus group discussions were deidentified. After a verbal summary of the study tasks and setting the ground rules of the focus groups, consent was obtained from all members of the group, and questionnaires were collected anonymously. During the focus group discussions, no dropout occurred. As an observational, noninterventional, and nonbiomedical investigation of the study subjects' sociological behavior, it was exempt from ethical review, as it is out of the scope of the Hungarian Act CLIV of 1997 on Health Care, the Decree 23/2002 (9 May) of the Ministry of Health on Medical Research on Human Subjects, and the Decree 35/2005 (26 VIII) of the Ministry of Health on the Clinical Investigation of Investigational Medicinal Products for Human Use and the Application of Good Clinical Practice [42]. For the same reason, the Semmelweis University Committee for Regional Institutional Scientific and Research Ethics could not issue an institutional review board exemption.

Results

General Acceptance of AI

During the focus group discussions, participants were asked about AI technology as a starting point. Of the 23 participants, 17 (74%) stated they had heard about AI in a Yes or No scheme. On a 5-point Likert scale asking about the extent of their knowledge of AI, they were more hesitant, resulting in an average of 2.38 (n=16), where 1 was not familiar at all and 5 was totally familiar. As examples of the possible functions of AI, text or picture generation was mentioned the most (8 times), and in 3 cases, AI-generated content was attributed as "fake" or "not real." One participant said:

> *I know it can also generate fake photos.*

After a general impression of AI, its application in health care was also discussed. For the question "Would you trust in AI-generated medical texts, documents, or tools?" the answers averaged 3.06 (n=16) on a 5-point Likert scale (where 1 was no trust at all and 5 was complete trust). When participants were asked about the reasoning behind their answers, the need for human involvement was emphasized concerning decision-making regarding health issues. Two participants said the following:

> *Even if it was created by humans, machines can have errors, so I would have less confidence in it if my health were at stake.*
>
> *I have no opposition regarding artificial intelligence if they use it as a helping tool, but it would be frightening for me if it were to make decisions without human oversight.*

Applicability of Text Variants

In the focus group discussions, the AI-generated text variants were presented. As the first step before using these texts, 2 independent researchers reviewed the AI-generated draft text variants. Modifications were applied in only 2 cases due to severe grammatical errors in the Hungarian language that limited the integrity of these texts. Otherwise, all variants were intact and brought to the focus groups in their original form. The source of each text was clarified for members of the groups only during the closure of group sessions.

First, participants were asked to provide general feedback on the applicability of each text variant in the context of a future oral cancer screening program. Scores measured on a 5-point Likert scale were detected in 4 dimensions (understandability and clarity, the quality of information content, the tone and style of the texts, and their convincing factor), and the results are shown as the average of these 4 items. During the first focus group discussion with experts by experience, all 6 variants were presented to the group.

Although the expert group members were highly positive regarding all variants, there were slight differences in the scoring of the text versions. The ranking turned out to be the following: (1) scientific reasoning (4.92; n=6), (2) informal, familiar tone (4.83; n=6), (3) focusing on motivation (4.75; n=6), (4) clickbait news article style (4.71; n=6), (5) statistical reasoning (4.67; n=6), and (6) bullet-point format (4.63; n=6). Participants were also asked to agree on the two most promising text variants that represented the highest opportunity to raise the attendance rate according to their experience. A consensus was reached after a short discussion, resulting in the variant based on scientific reasoning being selected as the top choice, and the informal, familiar version as the second choice, without knowing the quantitative results. Participants were convinced that different text variants could address different subgroups of people experiencing homelessness. One participant remarked the following:

> *The familiar one will motivate the youth more. It sounds not so official.*

After the first focus group discussion, 2 text variants (number 2 with statistical reasoning and number 6 with a bullet-point format) were removed from the pool as these were highly redundant according to the previous participants, and going through 6 texts challenged their attention, limiting the effectiveness of group discussions. The remaining 4 variants were presented to both remaining focus groups in the same form.

Participants of the latter two group discussions (n=17) were more critical in all aspects of the quantitative survey. The results of the 5-point Likert scale scoring were the following: (1) informal, familiar tone (3.77; n=13), (2) focusing on motivation (3.69; n=15), (3) scientific reasoning (3.69; n=16), and (4) clickbait news article style (3.50; n=12).

XSL•FO
RenderX

Evaluation of AI-Generated Content by Research Participants

After scoring all text versions, an open discussion took place. All group discussions concluded that the texts are almost fully understandable. Two participants remarked the following:

> *I can totally get what they are speaking about.*
>
> *The main point is clear, even if there are difficult words.*

However, there were suggestions for certain changes related to wording for ease of reading. The replacement of medical jargon—from "oral cancer" to "mouth cavity tumor," as the latter is a more commonly used term by the general population in the Hungarian language—was mentioned 7 times and affected all variants, while words with Latin roots, for example, "informing" and "early staging," were advised to be changed to a more widely used expression one time each.

In addition, the length of sentences as a factor causing gaps in readability was mentioned twice in the context of the versions based on scientific and statistical reasoning. Furthermore, participants accommodated in night shelters and other temporary housing solutions mentioned that the phrasing in two-thirds (4/6) of the text variants was not inclusive enough, as the term "rough sleepers" was used as a synonym for the homeless population, and this might result in the alienation of other subgroups. As one participant said:

> *They say people living on the streets only. That's not very motivating for me, who is living in a shelter.*

Based on the focus group discussions, the research group summarized the main strengths and weaknesses of the text variants created by ChatGPT 3.5 in Textbox 1.

Textbox 1. Evaluation of strengths and weaknesses of the ChatGPT-generated health information content of 6 text variants by people experiencing homelessness.

Strengths

- No significant opposition was detected against AI-created content from people experiencing homelessness.
- It is easy to generate many text outputs with open-access tools quickly.
- The results are almost ready to use, with minimal modification needed from the textual coherence point of view (in the Hungarian language).
- In most cases, participants were positive about whether the texts could fulfill the goal of motivating the target population to attend the program.
- Text variants in various tones and styles can attract different age groups.

Weaknesses

- There was a level of disapproval, mostly regarding AI-based decision-making processes concerning health issues.
- Text variants repeated the same problems (eg, medical jargon is difficult to understand for vulnerable populations).
- The motivational elements of text variants were stereotypical to a subgroup of people experiencing homelessness (rough sleepers) and lacking other prominent subgroups (eg, people accommodated in community shelters or temporary hostels).

Assessment of Text Variants With the SHeLL Editor

The research participants mentioned during the focus group discussions that text variants presented words that were difficult for them to understand, so to have a more comprehensive understanding of the ChatGPT-generated text variants' readability level, we evaluated the text with the help of the SHeLL Editor [33,38,43]. As this tool is only available in English, we translated the Hungarian text variants into English. The assessment of the text variants based on their word count, grade reading score, language complexity in percentages, and passive voice usage is summarized in Table 2.

Table . Evaluation of the readability of the 6 ChatGPT-generated text variants with the Sydney Health Literacy Lab Health Literacy Editor.

Text versions	Word count (sentence count)	Grade reading score	Text complexity, %	Passive voice, word count
Scientific evidence	87 (6)	11.5	26.6	0
Statistical evidence	81 (5)	10.4	21.4	0
Motivational	90 (7)	9.2	16.2	0
Informal	94 (11)	7.4	7.3	0
Clickbait news article	100 (12)	10.3	22.3	0
Bullet point format	65 (6)	8.3	21.2	0

Grade reading score refers to how difficult a text is to read and roughly corresponds to the expected reading ability for US school students in different grades [33]. Text complexity means the proportion of the text (%) that contains acronyms, uncommon words (as defined by an existing English-language corpus), or terms listed as public health or medical jargon [43].

XSL•FO
RenderX

Discussion

Main Findings

Our aim to co-design an awareness-raising fact sheet for an oral cancer screening program with people experiencing homelessness and ChatGPT was realized. We were also able to test the usability of ChatGPT in designing adequate information materials serving the needs of people experiencing homelessness by having focus group participants evaluate the ChatGPT-generated text variants. Moreover, focus group participants expressed prior knowledge of the concept of AI. Of potential functions of AI, they mentioned text or image generation the most. It also turned out that they did not reject the medical use of AI, although they indicated hesitancy in trusting it, especially without human oversight.

The text evaluation included cohesiveness, wording, tone, and style, and the results showed that, overall, the texts were able to fulfill their purpose of motivating the target group to participate in the screening activities, although participants suggested that the wording could be less stereotypical and less difficult to read. They also mentioned that text variants with different tones and styles could attract different age groups from the diverse population of people experiencing homelessness. The readability assessment of the texts underpinned their findings as the readability level of the majority of the text variants was above the readability level recommended for health-related texts by the literature [44].

Applicability of Generative Software in Health Care

Many fields of possible applications have been raised in using generative AI in clinical settings, such as writing discharge summaries [44], medical notes based on transcripts of physician-patient encounters, summaries of laboratory test results [45], medical education [46], medical research [47], providing a communication platform for patients, and facilitating health information dissemination [47]. One of the most obvious applications is generating tailored patient information on a predetermined topic, as collecting massive amounts of available evidence on different topics and human-like reasoning are easily achievable with open-access versions of generative software.

However, vulnerable populations might have different contexts, motivations, challenges, and medical needs than the general population and often require tailored medical treatment approaches to ensure the safety and efficacy of the treatment alongside potentially optimal health outcomes [48]. Moreover, concerns have arisen that the quality of AI-generated results depends on the user's ability to develop effective prompts, input accurate text for inquiries, and access advanced features through subscriptions; as a result, individuals with limited health literacy, insufficient prompt development skills, or an inability to afford premium subscriptions may miss out on these technological benefits, potentially exacerbating health disparities [49].

Vulnerable Groups and Their Knowledge and Trust Around AI

In health care, underserved subgroups are known to have limited access to care pathways and possess altered demands in addition to an existing systematic information asymmetry, as our previous study also revealed [14]. As the results showed, anxiety, misunderstanding, discrimination, and negative experiences related to this information deficit could be compensated for by using co-design principles. Better usability of such services might play an important role in the more equitable management of health issues. Moreover, the usage of ChatGPT as a co-design element might unburden health care and social care personnel tasked with the formulation of client information, as creating relevant materials with appropriate prompts takes significantly less time than building them from scratch. On the other hand, editing the draft or iterating the prompt sequence may require a level of expertise and take additional time, making the time-saving element of the use of ChatGPT unclear. Further on, as another potential downside of relying on the technology, it is questionable whether and for how long the subscription-based model of OpenAI or any other generative software development company will allow vulnerable populations to benefit from the advantages of generative AI in the future.

Our study recruited people experiencing homelessness, one of the most underserved populations. The randomly invited study participants had a nonnegligible prior knowledge of AI technology's existence, although they self-evaluated their knowledge as slightly below average. Previous research shows that people with lower socioeconomic status are slower to adopt new technology, and the rates of smartphone and internet use among people experiencing homelessness were lower than for those with similarly low socioeconomic status but more stable housing [14,50]. A 2023 international, multicenter, cross-sectional study assessing the attitudes of hospital patients toward AI in health care across 43 countries, including Hungary, found that patients have a predominantly favorable general view of AI in health care [51]. In Hungary, a representative survey published in September 2024 found that 79% of the population believed they knew what AI was, and 31% of respondents used chatbots and virtual customer service assistants [52].

In our study, participants' attitudes toward the medical use of AI were slightly above average, meaning that they might be hesitant or neutral when it comes to trusting such services. This is in line with other Hungarian general populational findings [53]. In this survey, researchers asked respondents how they would feel if their family doctor or medical specialists would partly rely on AI during their care; overall, 41.2% of the respondents were neutral, 27.5% said they would feel rather bad or very bad, while 31.3% reported they would feel rather well or very well about it [53].

Text Quality Evaluation

The ChatGPT-generated draft text variants had to be modified by the researchers as these versions contained a few severe grammatical errors in the Hungarian language; however, after such modifications were made, the texts were presentable and positively accepted by the focus groups. This might be due to generative AI software being most predominantly trained on Standard English texts, which means that in the case of small languages such as Hungarian, there is limited data available online for model training; therefore, large language models perform worse in such a "low resource" language compared to

XSL•FO
RenderX

English or other "high resource" languages such as Spanish, Chinese, or Arabic [54].

This could partly explain why the assessment of the text variants by the SHeLL Editor showed such strong differences by readability, although the target group was defined in the prompts as people experiencing homelessness, implying generally lower health literacy levels. As Ayre et al [33] found when they were experimenting with prompt design, prompts that described specific health literacy principles (eg, simple language, active voice, minimal jargon) worked better with ChatGPT than prompts that described the target audience. This could suggest that social sector employees would greatly benefit from prompting skill enhancement concerning AI health-related text generation. The use of the official ChatGPT prompt engineering guide [55] or specific prompt design elements, such as in-context learning, could also aid the process [56].

Members of the focus groups generally stated that the various styles and tones might attract various subgroups and generations of people experiencing homelessness; however, they also noted that the motivational elements of text variants were stereotypical of a subgroup of people experiencing homelessness (rough sleepers) and lacking other prominent subgroups (eg, people accommodated in community shelters or temporary hostels). This could partly stem from the generalization bias challenge of LLM models such as the one behind ChatGPT. These models are trained on large datasets that may contain biases, stereotypes, and prejudiced language [49,57]. As a result, the model may unintentionally learn these biases and produce responses that are offensive or perpetuate harmful stereotypes, such as the one about people experiencing homelessness being represented as rough sleepers.

Co-Design With Experts by Experience and Technology

In recent years, co-design, co-creation, co-production, or different forms of citizen engagement and collaboration of stakeholders have gained popularity in various fields, including social services for people experiencing homelessness [58]. The involvement of individuals with lived experience has also been shown to increase recruitment and follow-up rates in research projects, add to the validation of research findings, and generate more useful outputs [59,60]. This research project highlights these previous findings, as the involvement of the experts by experience group, as well as two focus groups, generated useful insights.

This experimental focus group study offered the opportunity to bring generative AI technology into the co-design process as a potential new element for the consideration of personnel working in the social or health sectors dealing with vulnerable subgroups, although the final benefits of this approach require further research and analysis. Our results showed that ChatGPT could produce usable material as a solid base for a health information material draft, which was acceptable for the target group, while the co-design process revealed additional benefits.

Limitations

Our study had some limitations. As a qualitative study relying on focus groups and feedback questionnaires, the methods themselves posed certain drawbacks. Although focus groups encourage participation from vulnerable populations and do not rely on participant literacy, they offer a space where those individual perspectives that differ from the majority opinion might remain hidden due to overriding behavioral or cultural norms or a desire to be seen as conforming [61,62].

The study participants were selected from the urban homeless population in Budapest, Hungary, where socioeconomic conditions might differ from those in the countryside. In addition, participants represented people experiencing homelessness who had a connection to the social infrastructure; therefore, others not in touch with the Hungarian social service architecture were not represented in the study sample. For a qualitative study using focus groups and feedback questionnaires, the sample size was small, and this should be taken into account when drawing conclusions.

Regarding technology, the researchers used OpenAI's most advanced freely accessible technology, ChatGPT 3.5, at the time of the research, while other generative AI software, such as Google's Gemini (previously Bard), Claude, or Synthesia, were not used. The use of ChatGPT 3.5, or any other generative software for that matter, also raises the question of replicability; with the constant and rapid development of LLMs, it might become uncertain whether this research could be replicated with the same technological conditions. Regarding the text evaluation aspect, we did not use a baseline text variant produced by human hands, as we had the intention to involve only ChatGPT in the co-creation process, as well as to assess the quality of the text variants that emerged during the process.

Conclusions

Our study revealed that health information materials generated by AI can be used by people experiencing homelessness in an oral cancer screening program. The co-design process revealed that the participants in the focus groups wanted to actively shape the drafts for the screening program and shared their ideas and insights on how to finalize the texts to avoid prevailing stereotypes about people experiencing homelessness and include more subgroups, as well as how to frame the text for various target audiences.

The group discussion also revealed some challenges of current LLM technology when using it without prior prompting experience. Based on our results, using the most up-to-date LLM technology, considering the health literacy and general language skills of vulnerable populations and avoiding generalization bias for this underrepresented group, and extensive prompt design upskilling of social workers and other groups of people aiming to produce health information material would be beneficial for future applications.

Moreover, via co-creation with members of the target audience, the final product might be more appealing to the target group of a health screening program. As a recommendation for its efficient use, offering prompt design training to personnel working in the social or health sectors may help maximize the impact of AI in client care.

XSL•FO
RenderX

Acknowledgments

The publication of this research project was supported by the European Social Fund Plus (ESF+) under the Project Code SOLACE-CEE 101172625. Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them.

Conflicts of Interest

None declared.

Multimedia Appendix 1
ChatGPT prompts and responses.
[DOCX File, 26 KB - formative_v9i1e68316_app1.docx]

Checklist 1
Consolidated Criteria for Reporting Qualitative Research checklist.
[DOCX File, 19 KB - formative_v9i1e68316_app2.docx]

References

1. Schreiter S, Bermpohl F, Krausz M, et al. The prevalence of mental illness in homeless people in Germany. Dtsch Arztebl Int 2017 Oct 6;114(40):665-672. [doi: 10.3238/arztebl.2017.0665] [Medline: 29070426]
2. Zhang L, Norena M, Gadermann A, et al. Concurrent disorders and health care utilization among homeless and vulnerably housed persons in Canada. J Dual Diagn 2018;14(1):21-31. [doi: 10.1080/15504263.2017.1392055] [Medline: 29494795]
3. Graffy P, McKinnon S, Lee G, Remington P. Life outside: a narrative ethnographic inquiry into the determinants of homelessness. J Poverty 2019 Apr 16;23(3):202-228. [doi: 10.1080/10875549.2018.1550133]
4. Aldridge RW, Story A, Hwang SW, et al. Morbidity and mortality in homeless individuals, prisoners, sex workers, and individuals with substance use disorders in high-income countries: a systematic review and meta-analysis. Lancet 2018 Jan 20;391(10117):241-250. [doi: 10.1016/S0140-6736(17)31869-X] [Medline: 29137869]
5. van Dongen SI, van Straaten B, Wolf J, et al. Self-reported health, healthcare service use and health-related needs: a comparison of older and younger homeless people. Health Soc Care Community 2019 Jul;27(4):e379-e388. [doi: 10.1111/hsc.12739] [Medline: 31020738]
6. Beaton L, Coles E, Freeman R. Homeless in Scotland: an oral health and psychosocial needs assessment. Dent J (Basel) 2018 Dec 1;6(4):67. [doi: 10.3390/dj6040067] [Medline: 30513713]
7. Daly B, Newton T, Batchelor P, Jones K. Oral health care needs and oral health-related quality of life (OHIP-14) in homeless people. Community Dent Oral Epidemiol 2010 Apr;38(2):136-144. [doi: 10.1111/j.1600-0528.2009.00516.x] [Medline: 20074294]
8. Figueiredo RLF, Hwang SW, Quiñonez C. Dental health of homeless adults in Toronto, Canada. J Public Health Dent 2013;73(1):74-78. [doi: 10.1111/j.1752-7325.2012.00355.x] [Medline: 22881462]
9. Simons D, Pearson N, Movasaghi Z. Developing dental services for homeless people in East London. Br Dent J 2012 Oct;213(7):E11. [doi: 10.1038/sj.bdj.2012.891] [Medline: 23059693]
10. Freitas DJ, Kaplan LM, Tieu L, Ponath C, Guzman D, Kushel M. Oral health and access to dental care among older homeless adults: results from the HOPE HOME study. J Public Health Dent 2019 Dec;79(1):3-9. [doi: 10.1111/jphd.12288] [Medline: 30295922]
11. Bedmar MA, Bennasar-Veny M, Artigas-Lelong B, et al. Health and access to healthcare in homeless people: protocol for a mixed-methods study. Medicine (Baltimore) 2022 Feb 18;101(7):e28816. [doi: 10.1097/MD.0000000000028816] [Medline: 35363172]
12. Szabó G, Klenk G, Veér A. Correlation between the combination of alcohol consumption and smoking in oral cancer (screening of the population at risk). Orv Hetil 1997 Dec 28;138(52):3297-3299. [Medline: 9463184]
13. Stormon N, Pradhan A, McAuliffe A, Ford PJ. Does a facilitated pathway improve access to dental services for homeless and disadvantaged adults? Eval Program Plann 2018 Dec;71:46-50. [doi: 10.1016/j.evalprogplan.2018.08.002] [Medline: 30114615]
14. Radó N, Békási S, Győrffy Z. Health technology access and peer support among digitally engaged people experiencing homelessness: qualitative study. JMIR Hum Factors 2024 May 14;11:e55415. [doi: 10.2196/55415] [Medline: 38743937]
15. Liu M, Hwang SW. Health care for homeless people. Nat Rev Dis Primers 2021 Jan 14;7(1):5. [doi: 10.1038/s41572-020-00241-2] [Medline: 33446661]

XSL•FO
RenderX

16. Goode J, Hoang H, Crocombe L. Homeless adults' access to dental services and strategies to improve their oral health: a systematic literature review. Aust J Prim Health 2018 Jul 9. [doi: 10.1071/PY17178] [Medline: 29983138]
17. Doughty J, M Gallier S, Paisi M, Witton R, J Daley A. Opportunistic health screening for cardiovascular and diabetes risk factors in primary care dental practices: experiences from a service evaluation and a call to action. Br Dent J 2023 Nov;235(9):727-733. [doi: 10.1038/s41415-023-6449-6] [Medline: 37945870]
18. Nunez E, Gibson G, Jones JA, Schinka JA. Evaluating the impact of dental care on housing intervention program outcomes among homeless veterans. Am J Public Health 2013 Dec;103 Suppl 2(Suppl 2):S368-S373. [doi: 10.2105/AJPH.2012.301064] [Medline: 23678921]
19. Radó N, Girasek E, Békási S, Győrffy Z. Digital technology access and health-related internet use among people experiencing homelessness in Hungary: quantitative survey. J Med Internet Res 2022 Oct 19;24(10):e38729. [doi: 10.2196/38729] [Medline: 36260379]
20. Győrffy Z, Békási S, Döbrössy B, et al. Exploratory attitude survey of homeless persons regarding telecare services in shelters providing mid- and long-term accommodation: the importance of trust. PLoS One 2022;17(1):e0261145. [doi: 10.1371/journal.pone.0261145] [Medline: 34990458]
21. Békási S, Girasek E, Győrffy Z. Telemedicine in community shelters: possibilities to improve chronic care among people experiencing homelessness in Hungary. Int J Equity Health 2022 Dec 17;21(1):181. [doi: 10.1186/s12939-022-01803-4] [Medline: 36528777]
22. Sanders E. From user-centered to participatory design approaches. In: Des Soc Sci Mak Connect 2002:1-7. [doi: 10.1201/9780203301302.ch1]
23. Csikar J, Vinall-Collier K, Richemond JM, Talbot J, Serban ST, Douglas GVA. Identifying the barriers and facilitators for homeless people to achieve good oral health. Community Dent Health 2019 May 30;36(2):137-142. [doi: 10.1922/CDH_4488Csikar06] [Medline: 31070874]
24. Sayani A, Hussain A, Freedman H, et al. EP.04A.10 Creating safe connections: usability testing of an intervention co-designed to increase equitable access to lung cancer screening. J Thorac Oncol 2024 Oct;19(10):S463-S464. [doi: 10.1016/j.jtho.2024.09.849]
25. Fox S, Brown LJE, Antrobus S, et al. Co-design of a smartphone app for people living with dementia by applying agile, Iierative co-design principles: development and usability study. JMIR mHealth uHealth 2022 Jan 14;10(1):e24483. [doi: 10.2196/24483] [Medline: 35029539]
26. Latulippe K, Hamel C, Giroux D. Co-design to support the development of inclusive eHealth tools for caregivers of functionally dependent older persons: social justice design. J Med Internet Res 2020 Nov 9;22(11):e18399. [doi: 10.2196/18399] [Medline: 33164905]
27. Noorbergen TJ, Adam MTP, Teubner T, Collins CE. Using co-design in mobile health system development: a qualitative study with experts in co-design and mobile health system development. JMIR mHealth uHealth 2021 Nov 10;9(11):e27896. [doi: 10.2196/27896] [Medline: 34757323]
28. Lewis CC, Taba M, Allen TB, et al. Developing an educational resource aimed at improving adolescent digital health literacy: using co-design as research methodology. J Med Internet Res 2024 Aug 7;26(1):e49453. [doi: 10.2196/49453] [Medline: 39110967]
29. Visser FS, Stappers PJ, van der Lugt R, Sanders EBN. Contextmapping: experiences from practice. CoDesign 2005 Apr;1(2):119-149. [doi: 10.1080/15710880500135987]
30. Rodriguez A, Shambhunath S, Wijesiri TID, Biazus-Dalcin C, Mc Goldrick N. Co-design of health educational materials with people experiencing homelessness and support workers: a scoping review. Front Oral Health 2024;5:1355349. [doi: 10.3389/froh.2024.1355349] [Medline: 38919732]
31. Lambert R, Choo ZY, Gradwohl K, Schroedl L, Ruiz De Luzuriaga A. Assessing the application of large language models in generating dermatologic patient education materials according to reading level: qualitative study. JMIR Dermatol 2024 May 16;7:e55898. [doi: 10.2196/55898] [Medline: 38754096]
32. Jia JL, Nguyen B, Sarin KY. Assessment of readability and content of patient-initiated Google search results for epidermolysis bullosa. Pediatr Dermatol 2019 Nov;36(6):1004-1006. [doi: 10.1111/pde.13975] [Medline: 31468562]
33. Ayre J, Mac O, McCaffery K, et al. New frontiers in health literacy: using ChatGPT to simplify health information for people in the community. J Gen Intern Med 2024 Mar;39(4):573-577. [doi: 10.1007/s11606-023-08469-w] [Medline: 37940756]
34. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
35. Dzuali F, Seiger K, Novoa R, et al. ChatGPT may improve access to language-concordant care for patients with non-English language preferences. JMIR Med Educ 2024 Dec 10;10:e51435. [doi: 10.2196/51435] [Medline: 39657144]
36. Liu Y, Wang H. Who on Earth Is Using Generative AI: World Bank; 2024. [doi: 10.1596/1813-9450-10870]
37. Williamson JML, Martin AG. Analysis of patient information leaflets provided by a district general hospital by the Flesch and Flesch-Kincaid method. Int J Clin Pract 2010 Dec;64(13):1824-1831. [doi: 10.1111/j.1742-1241.2010.02408.x] [Medline: 21070533]

XSL•FO
RenderX

38. Świeczkowski D, Kułacz S. The use of the Gunning Fog Index to evaluate the readability of Polish and English drug leaflets in the context of Health Literacy challenges in Medical Linguistics: an exploratory study. Cardiol J 2021;28(4):627-631. [doi: 10.5603/CJ.a2020.0142] [Medline: 33140389]
39. McLaughlin GH. SMOG grading — a new readability formula. J Read 1969;22:639-646 [FREE Full text]
40. Mac O, Ayre J, Bell K, McCaffery K, Muscat DM. Comparison of readability scores for written health information across formulas using automated vs manual measures. JAMA Netw Open 2022 Dec 1;5(12):e2246051. [doi: 10.1001/jamanetworkopen.2022.46051] [Medline: 36508219]
41. SHeLL Editor. URL: https://shell.techlab.works/ [accessed 2025-06-30]
42. 1997. évi CLIV. törvény az egészségügyről (1997th CLIV. Law on Health Care). WHO MiNDbank. URL: https://extranet.who.int/mindbank/item/3817 [accessed 2025-05-29]
43. Ayre J, Bonner C, Muscat DM, et al. Multiple automated health literacy assessments of written health information: development of the SHeLL (Sydney Health Literacy Lab) Health Literacy Editor v1. JMIR Form Res 2023 Feb 14;7:e40645. [doi: 10.2196/40645] [Medline: 36787164]
44. Patel SB, Lam K. ChatGPT: the future of discharge summaries? Lancet Digit Health 2023 Mar;5(3):e107-e108. [doi: 10.1016/S2589-7500(23)00021-3] [Medline: 36754724]
45. Lee P, Bubeck S, Petro J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. N Engl J Med 2023 Mar 30;388(13):1233-1239 [FREE Full text] [doi: 10.1056/NEJMsr2214184] [Medline: 36988602]
46. Preiksaitis C, Rose C. Opportunities, challenges, and future directions of generative artificial intelligence in medical education: scoping review. JMIR Med Educ 2023 Oct 20;9(1):e48785. [doi: 10.2196/48785] [Medline: 37862079]
47. Fatima A, Shafique MA, Alam K, Fadlalla Ahmed TK, Mustafa MS. ChatGPT in medicine: a cross-disciplinary systematic review of ChatGPT's (artificial intelligence) role in research, clinical practice, education, and patient interaction. Medicine (Baltimore) 2024 Aug 9;103(32):e39250. [doi: 10.1097/MD.0000000000039250] [Medline: 39121303]
48. Tao H, Liu L, Cui J, Wang K, Peng L, Nahata MC. Potential use of ChatGPT for the treatment of infectious diseases in vulnerable populations. Ann Biomed Eng 2024 Dec;52(12):3141-3144. [doi: 10.1007/s10439-024-03600-2] [Medline: 39133388]
49. Uddin J, Feng C, Xu J. Health communication on the internet: promoting public health and exploring disparities in the generative AI era. J Med Internet Res 2025 Mar 6;27(1):e66032. [doi: 10.2196/66032] [Medline: 40053755]
50. Raven MC, Kaplan LM, Rosenberg M, Tieu L, Guzman D, Kushel M. Mobile phone, computer, and internet use among older homeless aadults: results from the HOPE HOME cohort study. JMIR mHealth uHealth 2018 Dec 10;6(12):e10049. [doi: 10.2196/10049] [Medline: 30530464]
51. Busch F, Hoffmann L, Xu L, et al. Multinational attitudes towards AI in healthcare and diagnostics among hospital patients. medRxiv. Preprint posted online on Sep 2, 2024. [doi: 10.1101/2024.09.01.24312016]
52. Broadest ever research on AI readiness in Hungary published. IT Business. 2024. URL: https://itbusiness.hu/english/research-on-ai-awareness-in-hungary/ [accessed 2024-10-21]
53. Győrffy Z, Radó N, editors. E-Patients and E-Physicians in Hungary 2024. URL: https://www.semmelweiskiado.hu/termek/2027/e-patients-and-e-physicians-in-hungary [accessed 2025-06-30]
54. Nicholas G, Bhatia A. Lost in translation: large language models in non-English content analysis. arXiv. Preprint posted online on Jun 12, 2023. [doi: 10.48550/arXiv.2306.07377]
55. Prompt engineering best practices for ChatGPT. OpenAI. URL: https://help.openai.com/en/articles/10032626-prompt-engineering-best-practices-for-chatgpt [accessed 2025-05-08]
56. Ferber D, Wölflein G, Wiest IC, et al. In-context learning enables multimodal large language models to classify cancer pathology images. Nat Commun 2024 Nov 21;15(1):10104. [doi: 10.1038/s41467-024-51465-9] [Medline: 39572531]
57. Ray PP. ChatGPT: a comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet of Things and Cyber-Physical Systems 2023;3:121-154. [doi: 10.1016/j.iotcps.2023.04.003]
58. Meriluoto T. Between expertise and authenticity: co-creation in Finnish Housing First Initiatives from the perspective of experts-by-experience. Eur J Homelessness 2018;12(1):61-83 [FREE Full text]
59. Crooks J, Flemming K, Shulman C, Casey E, Hudson B. Involving people with lived experience of homelessness in palliative and end of life care research: key considerations from experts in the field. Res Involv Engagem 2024 Jan 30;10(1):16. [doi: 10.1186/s40900-024-00549-3] [Medline: 38291483]
60. Staley K. 'Is it worth doing?' Measuring the impact of patient and public involvement in research. Res Involv Engagem 2015;1(1):6. [doi: 10.1186/s40900-015-0008-5] [Medline: 29062495]
61. Kitzinger J. Qualitative research. Introducing focus groups. BMJ 1995 Jul 29;311(7000):299-302. [doi: 10.1136/bmj.311.7000.299] [Medline: 7633241]
62. Rice PL, Ezzy D. Qualitative Research Methods: A Health Focus: Oxford University Press.

Abbreviations

AI: artificial intelligence
COREQ: Consolidated Criteria for Reporting Qualitative Research

XSL•FO
RenderX

LLM: large language model
SHeLL: Sydney Health Literacy Lab

Edited by A Mavragani; submitted 02.11.24; peer-reviewed by F Chen, V Fu; revised version received 02.06.25; accepted 03.06.25; published 04.07.25.

Please cite as:
Radó N, Németh O, Békási S
Co-Design of a Health Screening Program Fact Sheet by People Experiencing Homelessness and ChatGPT: Focus Group Study
JMIR Form Res 2025;9:e68316
URL: https://formative.jmir.org/2025/1/e68316
doi:10.2196/68316

© Nóra Radó, Orsolya Németh, Sándor Békási. Originally published in JMIR Formative Research (https://formative.jmir.org), 4.7.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Feasibility and Acceptability of a Digital Health Portal to Improve HIV Care Engagement Among Kenyan Youth: Mixed Methods Study

Eric Nturibi*, MD, MSc, MMed; Jared Mecha*, MD, MSc, MMed, MPH; Florence Kaara*, BSc, MPH; Faith Musau*, BSc, MA; Christine Mwangi*, BA, MPH; Elizabeth Kubo*, MD, MPH; Albert Orwa*, MSc

Centre for Health Quality and Innovation, Department of Clinical Medicine and Therapeutics, University of Nairobi, P.O. Box 30197, Nairobi, Kenya
*all authors contributed equally

Corresponding Author:
Eric Nturibi, MD, MSc, MMed
Centre for Health Quality and Innovation, Department of Clinical Medicine and Therapeutics, University of Nairobi, P.O. Box 30197, Nairobi, Kenya

Abstract

Background: Adolescents and young people aged 15‑24 years in Kenya bear a disproportionate burden of HIV, necessitating innovative, youth-friendly approaches to improve care engagement. Digital solutions such as patient health portals (PHPs) offer scalable ways to enhance access, understanding, and support.

Objective: This study aims to evaluate the feasibility and acceptability of a customized PHP tailored for Kenyan adolescents and young people living with HIV, and to explore participant preferences for features, content delivery modes, and privacy concerns.

Methods: A cross-sectional, mixed methods feasibility study was conducted among 163 adolescents and young people recruited during psychosocial support group meetings across seven high-volume clinics in Kiambu and Kirinyaga counties in August 2021. Participants interacted with a prototype PHP through guided demonstrations and completed structured questionnaires assessing health literacy, digital readiness, feature preferences, emotional needs, and data security perceptions. Quantitative data were analyzed using descriptive and bivariate statistics. While qualitative data were also collected to inform portal development, this paper focuses exclusively on quantitative findings.

Results: The median age was 18 (IQR 15‑24) years, with 60% (98/163) female participants. Most participants were students (73/163, 45%) and had secondary-level education (83/163, 51%). Approximately 17% (27/163) participants reported difficulty understanding written medical information. A large majority (146/160, 91%) expressed interest in using a web-based portal, with 78% (124/159) participants rating it as easy to use. Weekly use was anticipated by 56% (90/161) of participants. Top-rated features included appointment scheduling (129/163, 79%), access to test results (129/163, 79%), and communication with the doctor (117/163, 72%). Emotional health tools like a mood tracker and PHQ-9 (Patient Health Questionnaire-9) screening were highly valued, with 59% (92/156) reporting difficulties coping emotionally with HIV. Concerns about data privacy were minimal, and 62% (99/159) were willing to share access with a trusted family member. No statistically significant associations were observed between portal preference and age, sex, education level, or employment status.

Conclusions: The study demonstrates the strong feasibility and acceptability of a digital health portal among Kenyan adolescents and young people. High readiness for digital health, combined with clearly expressed content and feature preferences, underscores the potential for such tools to improve engagement in HIV care. These findings support the need for co-designed, youth-centered digital interventions that address both medical and psychosocial needs in resource-limited settings.

(JMIR Form Res 2025;9:e59661) doi:10.2196/59661

KEYWORDS

HIV; online portal; antiretroviral; STI; STD; sexual; portal; young adult; ART; preference; engagement; experience; attitude; opinion; perception; perspective; survey; questionnaire; Africa

Introduction

The patient health portal (PHP) study, a stepped wedge cluster randomized controlled trial conducted between May 2022 and December 2023, evaluated the effectiveness of a tailored web-based health portal in fostering patient engagement among adolescents and young people aged 15 to 24 years. Detailed in our study protocol, the study reflects an innovative approach to

XSL•FO
RenderX

address the unique needs of this demographic, who often face barriers to consistent health care access and adherence [1].

Central to the design of the PHP was the Capability, Opportunity, Motivation, and Behavior model, a framework developed by Michie et al [2] to conceptualize behavior as a dynamic interplay between capability (the possession of knowledge and skills required to perform an action), opportunity (the external conditions that enable or prompt behavior), and motivation (reflexive and automatic processes that drive actions). The Capability, Opportunity, Motivation, and Behavior model's versatility has made it a cornerstone in designing effective health interventions, particularly for identifying and addressing barriers to behavior change [2].

The PHP portal featured an array of user-focused tools, including an e-learning module, a peer support forum, access to patient health information, a symptom diary, questionnaires, appointment reminders, and provider-specific functionalities (Figure 1).

Figure 1. Schematic overview of the proposed digital health portal for adolescents and young people living with HIV.

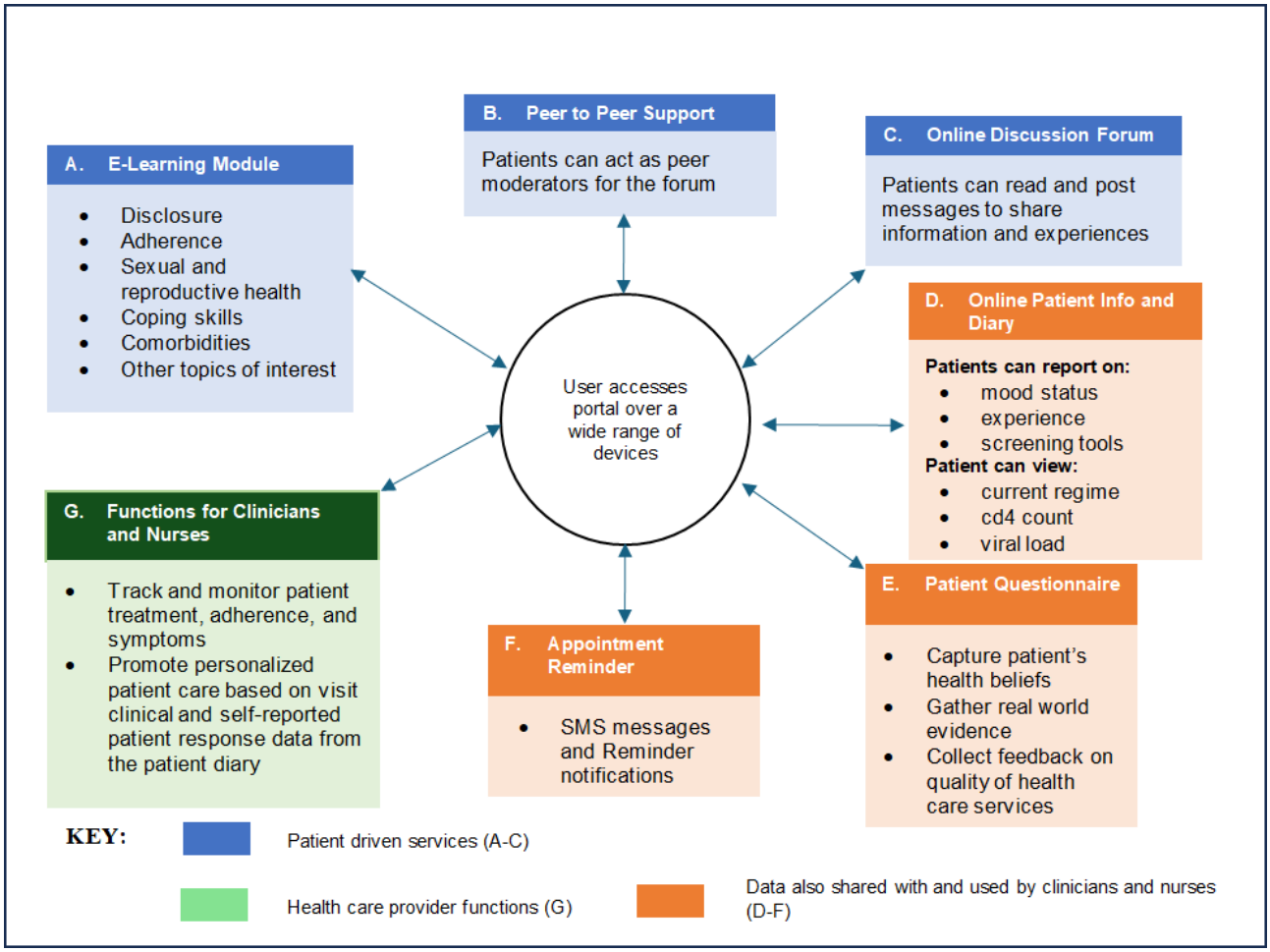

Delivered via internet-enabled smartphones, the portal seamlessly integrates with key elements of the patient's electronic medical record, ensuring a personalized and cohesive health care experience.

The development of the web-based health portal was motivated by the pressing need to address the disproportionate burden of HIV among young people, who account for a significant share of new infections and HIV-related deaths globally. Leveraging the readiness of adolescents and young people to adopt internet-based digital health solutions, this initiative aimed to bridge critical gaps in health care delivery [3,4]. The portal was integrated into a large-scale antiretroviral therapy program in Kenya, which, as of September 30, 2021, had provided antiretroviral therapy to 38,079 individuals living with HIV, including 3459 adolescents and young people, across 78 public health facilities (program data).

The PHP study was conducted in Kiambu and Kirinyaga counties, which represent distinct socioeconomic and epidemiological contexts for addressing HIV. Kiambu, a predominantly urban and peri-urban county, had an HIV prevalence of 4.0% among adults and 1.8% among youth aged 15‑24 years by the end of 2019 and had recorded approximately 510 new infections and 230 AIDS-related deaths. While Kiambu benefits from a thriving economy driven by agriculture, industrialization, and services, significant health care challenges persist in its informal settlements.

In contrast, Kirinyaga, a largely rural county, reported an HIV prevalence of 3.2% among adults and 1.4% among youth aged 15‑24 years, with 300 new infections and 160 AIDS-related deaths over the same period. The county's economy relies heavily on tea, coffee, and subsistence farming, yet health care access remains constrained by economic vulnerabilities. Nationally, as of 2019, Kenya reported an HIV prevalence of

XSL•FO
RenderX

4.5% among adults, with 41,416 new infections and 21,000 AIDS-related deaths [5].

To determine the feasibility of the PHP study, we investigated smartphone ownership and internet use patterns (separate manuscript), and gathered end user feedback on the portal's proposed features. The study objectives were designed to comprehensively assess the portal's potential impact. Specifically, the questionnaire sought to (1) characterize the sociodemographic profiles of adolescents and young people aged 15‑24 years attending HIV care clinics; (2) evaluate readiness to adopt a web-based health portal; (3) gather and prioritize feature preferences to tailor the portal to user needs; and (4) identify specific life-stage domains critical for inclusion in the portal's content.

Methods

Study Setting and Recruitment Procedures

Study Context

The study was implemented across seven high-volume public HIV care facilities located in Kiambu and Kirinyaga counties. These sites were intentionally selected for their large youthful populations, well-established psychosocial support programs, and participation in national adolescent-centered initiatives such as Operation Triple Zero (OTZ). The OTZ initiative targets adolescents and young people aged 10‑24 years and promotes HIV care adherence through a focus on achieving "zero missed visits, zero missed drugs, and zero viral load".

Participant recruitment was embedded within triannual psychosocial support group (PSSG) meetings held during the August 2021 school holiday. Eligible individuals, identified through program rosters and clinic records, were aged 15‑24 years, enrolled in HIV care at one of the study clinics, and attended the PSSG session. Attendees were informed about the study through brief plenary announcements during the morning session. Interested individuals were then invited to designated study booths for further explanation.

Study staff provided oral and written information in English, Swahili, or Sheng (local slang), and obtained verbal and written consent. For minors aged 15‑17 years, assent was obtained alongside parental/guardian consent, consistent with ethical guidelines. Participation was voluntary, and no material or financial incentives were offered beyond routine PSSG refreshments and transport reimbursements. Individuals who chose not to participate in the focused group discussions (FGDs) were encouraged to engage in alternative PSSG activities.

Study Procedures

Overview

The PSSGs provided a strategic setting for introducing the proposed health portal. A designated project representative delivered step-by-step slide presentations, using screenshots to demonstrate the portal's login process, core functionalities, navigation tools, and operational features. A total of seven FGDs were conducted—one at each participating facility—comprising 20 to 25 participants each. Groups were stratified based on key attributes such as age, viral suppression status, and treatment duration to allow targeted feedback aligned with programmatic priorities.

Verbal consent for audio recording and note-taking was obtained, and participants formalized their agreement by signing a combined "attendance and consent form".

The FGDs followed a structured format, guided by a pretested questionnaire distributed to all participants. Facilitators introduced each question in English and, when necessary, paraphrased the content in Swahili or local slang to ensure comprehension and inclusivity. Participants were encouraged to engage in open discussions, and adequate time was allocated for them to document their responses. Secondary facilitators provided assistance to those requiring additional support, ensuring comprehensive engagement before progressing to subsequent sections. Key discussion points were recorded by a designated note-taker, and all completed questionnaires were collected at the conclusion of each session.

The questionnaire was thoughtfully developed by the project's psychosocial support team, drawing on over a decade of experience conducting PSSG sessions with adolescents and young people living with HIV. Its design incorporated insights from Kenya's National HIV Guidelines (2022) [6] OTZ program tools, and validated instruments including PHQ-9 (Patient Health Questionnaire-9) and CRAFFT (Car, Relax, Alone, Forget, Friends, Trouble) [7,8].

The final version of the questionnaire was organized into four distinct sections to ensure clarity and comprehensiveness.

Demographics

This section captured detailed sociodemographic characteristics of participants to understand their backgrounds and how these might influence engagement with the proposed digital health portal. Data collected included age, recorded in years and calculated from date of birth, to represent younger adolescents aged 15 to 19 years and older youth aged 20 to 24 years; sex, documented to analyze gender distribution and differences in preferences and readiness for digital solutions; educational status, categorized as primary, secondary, postsecondary such as diploma or undergraduate, or no formal education to assess the impact of literacy on portal use; and employment status, classified as student, unemployed, casual worker, self-employed, or formally employed to examine socioeconomic factors affecting access to digital tools.

Health Literacy and Portal Readiness

This part of the questionnaire evaluated the participants' ability to comprehend and use medical information, as well as their willingness to adopt the proposed digital portal. Key aspects included health literacy, where participants rated their ability to understand written medical information and instructions, highlighting gaps in traditional approaches; satisfaction with existing methods, measuring preferences and perceived limitations in current health education methods such as face-to-face sessions and printed materials; and portal readiness, assessing familiarity with smartphones and internet use, impressions of the portal prototype presented during sessions, and confidence in navigating its features independently. The

XSL•FO
RenderX

findings highlighted the participants' current health literacy levels and openness to adopting innovative digital health solutions, forming a baseline for designing user-centered features.

Portal Features and Privacy Concerns

This section of the questionnaire examined participants' preferences for portal functionalities, concerns about data privacy and security, and anticipated use patterns. Participants ranked features such as appointment management, access to medical records, test results, symptom diaries, educational content, and communication tools with health care providers. They also shared views on data security, emphasizing confidentiality and their willingness to share portal access with family members or caregivers. Additionally, expectations regarding the frequency of portal use, whether daily, weekly, or as needed, and the specific circumstances under which the portal would be used were recorded. These insights informed the prioritization of features that balance functional needs with the sensitive nature of managing personal health information.

Content Relevance

This section identified the life-stage domains most critical to participants and assessed their preferred methods for receiving health-related information. Participants prioritized topics such as disease progression, symptom management, psychosocial challenges, nutrition, physical activity, financial literacy, and social support resources. Preferred delivery methods included videos, SMS updates, pictorials, question-and-answer forums, storytelling, and animations, evaluated for their effectiveness in presenting information in an engaging and accessible manner.

The questionnaire was pretested at nonstudy sites over 2 months, incorporating feedback from peer mentors and small youth groups during OTZ clinic days to refine clarity, validity, and reliability. Completion time ranged from 30 to 45 minutes.

This study used a mixed methods design, integrating quantitative survey data with qualitative insights from FGDs to enhance understanding of participant preferences. Quantitative reporting in this manuscript follows the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines. Qualitative procedures were guided by the SRQR (Standards for Reporting Qualitative Research) framework. This paper primarily reports the quantitative findings.

Ethical Considerations

This study was conducted as part of a quality improvement initiative under the University of Nairobi CRISSP+ (Central Kenya Response: Integration, Strengthening, and Sustainability - Plus) Project. The project received ethics approval from the AMREF Health Africa Ethics and Scientific Review Committee (ESRC), with final approval granted in April 2022 (Ref: ESRC P1164/2022). All procedures were carried out in accordance with institutional and national ethical standards, and conformed to the principles outlined in the Declaration of Helsinki. As this was a quality improvement initiative involving minimal risk, and consistent with ESRC guidance, a waiver of written informed consent was granted. Nonetheless, participants were informed about the purpose of the initiative, assured of confidentiality, and given the option to decline participation without penalty.

Sample Size Estimation

This feasibility study uses a mixed methods approach, incorporating both quantitative and qualitative components. The larger trial for which this study provides preliminary data has a computed sample size of 480 participants, with follow-up planned for 12 months [1].

For the quantitative component of this study, Cochran formula was used to determine the appropriate sample size [9]:

$$n=Z2.p.(1-p)e2$$

where:

- Z=1.96 (z score for 95% confidence)
- p=0.5 (estimated proportion for maximum variability)
- e=0.1 (margin of error).

Given the finite population of 2647 young people in the study sites, the sample size was adjusted using the finite population correction:

$$nadjusted=n1+n-1N$$

This yielded a minimum sample size of approximately 93 participants. Our survey included 163 participants.

Analysis

The dataset underwent rigorous cleaning and coding processes prior to analysis to ensure accuracy and integrity. Univariate analyses were conducted to summarize participant characteristics and key variables of interest. Continuous variables, such as age, were described using medians and ranges, while categorical variables, such as sex, education level, and clinic type, were summarized using frequencies and percentages. The results of these univariate analyses were presented in tables and figures for clarity and ease of interpretation.

Bivariate analyses were then conducted to explore associations between participant demographic characteristics and their preferences or readiness to use the web-based health portal. The chi-square test was used to examine relationships between categorical variables, such as sex, education level, and employment status, with outcomes such as portal preference and perceived ease of use. For continuous variables like age, the mean and variance were calculated and compared across subgroups where relevant.

All statistical analyses were performed using Stata (version 13; StataCorp LP) [10].

Results

Background Characteristics

The study included 163 participants aged 15 - 24 years from diverse settings within Kiambu and Kirinyaga counties, including urban, peri-urban, rural, and informal settlements (Table 1).

XSL•FO
RenderX

Table . Geographic and settlement profile of study participants aged 15 - 24 years in Kiambu and Kirinyaga counties, Kenya (August 2021).

County and population	Participants (n=163), n (%)
Kiambu	
Urban	47 (29)
Informal settlement	16 (10)
Peri-urban	47 (29)
Kirinyaga	
Rural	53 (32)

A total of 60% (98/163) of the participants were female, 15% (24/163) of the participants attended prevention of mother to child transmission clinics, while 3% (5/163) of the participants had high viral loads. The median age was 18 years (Table 2). Educational backgrounds varied, with most participants (81/163, 51%) having completed secondary school. Most participants were students (72/163, 45%).

Table . Demographic summary of adolescents and young people living with HIV enrolled in a feasibility assessment of a proposed health portal in Kenya (August 2021).

Characteristic	Participants (n=163)
Sex (female), n (%)	98 (60)
Age (years), median (range)	18 (15-24)
Clinic type, n (%)	
Comprehensive Care Center	139 (85)
Prevention of mother to child transmission (PMTCT)	24 (15)
Education type, n (%)	
Primary	58 (36)
Secondary	83 (51)
College diploma	7 (4)
Undergraduate	15 (9)
Occupation, n (%)	
Student	73 (45)
Jobless	38 (23)
Casual	15 (9)
Self employed	29 (18)
Employed (formal)	8 (5)

Health Literacy and Readiness to Use a Web-Based Portal

In this second section, participants were asked to assess their ability to understand medical information, readiness to adopt a web-based health portal, and the ease of use of the portal prototype. Approximately 17% (27 of 163) of respondents reported frequent difficulties in understanding written medical information (Table 3).

Table . Proportion of participants reporting challenges in understanding medical information during a portal usability study in Kenya (August 2021).

Rating	Values (n=163), n (%)
Always	18 (11)
Often	10 (6)
Sometimes	72 (44)
Occasionally	18 (11)
Never	45 (28)

XSL•FO
RenderX

A majority (146/160, 91%) expressed a preference for using a web-based health portal to manage their health (Table 4), and 78% (124 of 159) found the prototype demonstrated during the PSSG meetings to be user-friendly (Table 5).

Table . Agreement with the statement: "I would use a portal to manage my health" among young people living with HIV in Kenya (August 2021).

Rating	Values (n=160), n (%)
Strongly agree	93 (58)
Agree	53 (33)
Neutral	6 (4)
Disagree	6 (4)
Strongly disagree	2 (1)

Table . Ease of use rating for the portal prototype demonstrated during support group meetings with young people living with HIV in Kenya (August 2021).

Rating	Values (n=159), n (%)
Very easy	41 (26)
Easy	83 (52)
Somewhat difficult	13 (8)
Difficult	13 (8)
Very difficult	9 (6)

More than half of the participants (90/161, 56%) anticipated weekly use of the portal once implemented (Table 6).

Bivariate analysis using chi-square tests revealed no statistically significant associations between portal preference and demographic factors such as age (P=.82), sex (P=.62), education level (P=.58), or employment status (P=.25).

Table . Participant responses on how frequently they anticipate using the digital health portal if implemented (Kenya, August 2021).

Frequency	Values (n=161), n (%)
Daily	43 (27)
Weekly	47 (29)
Monthly	37 (23)
3 to 6 monthly	26 (16)
Rarely or not at all	8 (5)

Portal Features and Privacy Concerns

Section 3 explored the specific features respondents desired in the portal, their concerns about privacy and security, and how they envisioned using the platform, including use frequency and access rights. Features such as appointment management, access to medical records and test results, and communication with health care providers were rated the highest (Table 7).

Additionally, the majority of respondents (105/159, 66%) indicated a willingness to grant access rights to a family member (Table 8).

XSL•FO
RenderX

Table . Feature preferences rated by participants for inclusion in a digital health portal tailored to young people living with HIV (Kenya, August 2021).

Feature	Values (n=163), n (%)
Allergy information	64 (39)
Caregiver access	64 (39)
Record treatment supporter	91 (56)
Lifestyle choices	98 (60)
Service requests	103 (63)
Sign up for reminders & updates	106 (65)
List of providers	109 (67)
Receive health information	109 (67)
Medication information	109 (67)
Communicate with peers	111 (68)
View medical records	116 (71)
Communicate with doctor	117 (72)
Test results	129 (79)
Appointments	129 (79)

Table . Participant views on sharing portal access with a trusted family member in the context of HIV care (Kenya, August 2021).

Access rights	Values (n=159), n (%)
Family member	99 (62)
No one	51 (32)
Other	9 (6)

Content Relevance

Section 4 aimed to identify key life-stage concerns of young individuals to determine the most valuable content areas for inclusion as “health tips” and to assess preferred methods for content delivery. Participants evaluated the extent to which the portal prototype might enhance various aspects of their health care experience, including the quality of patient-provider communication, their understanding of health, sense of control, reduction of worries, satisfaction with care, and the safeguarding of data privacy and confidentiality (Table 9).

Participants then rated the importance of content related to their disease, overall health, symptomatology, progression of their condition over time, and aspects of psychosocial, physical, and financial well-being (Table 10).

Table . Perceived influence of portal use on different aspects of health care experience among Kenyan adolescents and young people (August 2021).

Domain	Rating by participants, n (%)				
	Greatly improved	Slightly improved	No effect	Worsen slightly	Worsen greatly
Communication with my doctor (n=153)	113 (74)	26 (16)	12 (8)	0 (0)	2 (1)
Understanding of my health (n=152)	119 (78)	18 (12)	14 (9)	0 (0)	2 (2)
My sense of control (n=148)	104 (70)	24 (16)	16 (11)	0 (0)	4 (3)
My worries (n=149)	98 (66)	18 (12)	22 (15)	7 (5)	3 (3)
Safety of my care (n=146)	101 (69)	20 (14)	19 (13)	4 (3)	1 (1)
Satisfaction with care (n=148)	108 (73)	24 (16)	12 (8)	4 (3)	0 (0)
Overall quality of my health care (n=147)	107 (73)	28 (19)	12 (7)	0 (0)	0 (0)
Data privacy and security (n=154)	108 (70)	18 (12)	18 (12)	6 (4)	3 (2)

XSL•FO
RenderX

Table . Perceived influence of portal use on different aspects of health care experience among Kenyan adolescents and young people (August 2021).

Domain	Rating by participants, n (%)				
	Very important	Important	Not really important	Unimportant	Very unimportant
Being aware of all details about my health (n=163)	130 (80)	29 (18)	2 (1)	2 (1)	0 (0)
Understand the cause of my symptoms (n=162)	100 (62)	55 (33)	6 (4)	0 (0)	0 (0)
Knowing whether I am healthy (n=161)	126 (78)	32 (20)	0 (0)	0 (0)	3 (2)
Knowing how my condition will evolve (n=161)	118 (73)	39 (24)	2 (1)	3 (2)	0 (0)
Knowing that people in my community do not discriminate me (n=161)	76 (47)	47 (29)	31 (19)	5 (3)	3 (2)
Find information that is relevant to my personal health situation (n=160)	109 (68)	42 (26)	5 (4)	3 (2)	2 (1)
Trust the reliability of the health information I find myself (n=160)	106 (66)	50 (31)	3 (3)	0 (0)	2 (1)
To be informed of factors that influence your health without realizing them (n=157)	108 (69)	36 (22)	8 (5)	5 (3)	0 (0)
To be educated about the future of my social and financial security (n=161)	106 (66)	45 (28)	5 (3)	2 (1)	3 (2)
How important is it for you to understand exactly what impact your diet has on your health? (n=161)	121 (75)	34 (21)	3 (2)	2 (1)	2 (1)
How important is it for you to understand exactly what impact your physical activity has on your health? (n=161)	98 (61)	58 (36)	2 (1)	0 (0)	3 (2)

To evaluate the utility of proposed digital tools—such as the PHQ-9 for depression screening and the CRAFFT for identifying alcohol use disorders—participants were asked to rate their emotional well-being. Of the 156 adolescents and young people living with HIV who participated, 59% reported some level of difficulty coping emotionally with their condition. Specifically, 23% found it somewhat difficult, 21% found it difficult, and 15% found it very difficult. Conversely, 30% found it easy, and 11% very easy to manage the emotional impact of their diagnosis.

Since the portal was designed to include a symptom diary and a Tuberculosis screening tool, participants were asked to rate the importance of disease symptom awareness (Table 11).

To determine the most effective format for delivering content through the portal, participants were asked to rate common delivery methods. Video, SMS, and question-and-answer formats emerged as the most preferred options, while role-playing, animations, and storytelling were the least favored (Table 12).

XSL•FO
RenderX

Table . Importance attributed to symptom awareness and monitoring in the context of HIV care among Kenyan adolescents and young people (August 2021).

Domain	Rating by participants, n (%)				
	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
Sometimes I wonder how common the symptoms I have are (n=161)	56 (35)	82 (51)	16 (10)	5 (3)	2 (1)
Sometimes, I wonder whether my symptoms are severe enough to take action (n=161)	82 (51)	60 (37)	13 (8)	5 (3)	2 (1)
Sometimes, I wonder whether there are symptoms I should watch out for, so I can take actions in time (n=161)	90 (56)	61 (38)	3 (2)	5 (3)	2 (1)

Table . Preferred digital formats for health content delivery as selected by adolescents and young people participants during the feasibility study (Kenya, August 2021).

Mode	Rating by participants, n (%)				
	Strongly preferred	Preferred	Neutral	Disliked	Strongly disliked
SMS (n=162)	86 (53)	52 (32)	10 (6)	10 (6)	5 (3)
Pictorials (n=155)	54 (35)	51 (33)	22 (14)	11 (7)	17 (11)
Animations (n=157)	50 (32)	44 (28)	25 (16)	19 (12)	19 (12)
Role plays (n=154)	57 (37)	32 (21)	25 (16)	22 (14)	18 (12)
Storytelling (n=155)	48 (31)	48 (31)	23 (15)	12 (8)	23 (15)
Question and answer (n=156)	84 (54)	47 (30)	11 (7)	12 (8)	2 (1)
Audio (n=157)	66 (42)	55 (35)	19 (11)	9 (6)	8 (5)
Video (n=155)	98 (63)	40 (26)	8 (5)	5 (3)	5 (3)

Discussion

Principal Findings

This preliminary investigation explores the perspectives of young people on a digital health portal designed to enhance sustained engagement in HIV care and improve treatment outcomes. Our findings suggest that the portal is both timely and feasible, addressing a critical gap identified in a systematic review, which noted that many mobile health interventions fall short of aligning with user preferences [11].

Participants were recruited from a range of settings—urban, peri-urban, rural, and informal settlements—capturing the diversity of clinic populations typical of sub-Saharan Africa. The gender distribution reflects the elevated HIV burden among adolescent girls and young women aged 15 - 24, who remain disproportionately affected relative to their male peers [12]. Perspectives from subgroups such as antenatal clinic attendees, individuals with high viral loads, and those from varied occupational backgrounds further underscore the portal's broad relevance and adaptability.

Approximately 60% of participants reported difficulty navigating traditional medical information formats. This finding echoes research from a high HIV-burden county in Kenya, where low functional health literacy hindered tasks like interpreting appointment cards and understanding medical forms [13,14]. These results highlight the importance of integrating audio-visual formats tailored to young users' learning preferences, an approach reinforced by a Nairobi-based study that advocates for participatory and engaging educational methods to enhance comprehension and user involvement [15].

Participants showed high readiness to adopt the portal, with overwhelmingly positive feedback on its usability—an encouraging sign for scalability. This reception aligns with the e-commerce acceptance model, where perceived ease of use and perceived usefulness are critical determinants of technology adoption [16]. Nevertheless, nearly 20% of respondents anticipated challenges in using the portal, emphasizing the need for ongoing user feedback and iterative refinement to uphold a human-centered design approach [17].

Encouragingly, 80% of participants indicated they would engage with the portal at least once a month, offering the content

XSL•FO
RenderX

management team a basis for update scheduling. Essential features—such as appointment reminders, access to medical records, and test results—were purposefully integrated to align with existing clinic electronic medical records. Notably, 62% of participants expressed willingness to share access with family members, suggesting the need for built-in guidance on disclosure practices. In similar Kenyan settings, structured disclosure services have been associated with improved viral suppression among adolescents [18].

Emotional wellness features also resonated strongly with participants. Tools such as the mood tracker, PHQ-9 questionnaire, and coping resources were particularly well-received. Their inclusion was guided by national guidelines and reinforced by previous studies demonstrating the centrality of emotional support in adolescent HIV care [6-8,19]. Likewise, the expressed interest in a "symptom diary" underscores the value of early symptom detection—especially for comorbid conditions such as tuberculosis—where mobile applications have proven effective in enhancing screening and accelerating treatment initiation [20,21].

Beyond clinical features, participants prioritized content on nutrition and physical fitness, mirroring global trends in youth health information needs [11,22]. Moreover, strong demand for content addressing social and financial security points to broader age-specific concerns, including the economic burdens of chronic disease management. These findings are aligned with initiatives like DREAMS (determined, resilient, empowered, AIDS-free, mentored, and safe), which seek to mitigate structural drivers of HIV risk—such as poverty and gender inequality—among adolescent girls and young women [23-25]. Emerging evidence continues to support the effectiveness of this integrated, multisectoral approach [26].

Limitations

This survey, as a preliminary step to inform a controlled trial, relied on a convenience sample of 163 participants, representing approximately 6% of the target population. The use of a convenience sampling approach may limit the generalizability of the findings. As such, the perspectives of some young individuals living with HIV in Kenya or other contexts may not have been fully captured. Additionally, potential selection bias may have influenced the results, as participants attending PSSG meetings might differ systematically from those not engaged in these support structures. Future studies should consider random sampling strategies and broader recruitment frameworks to enhance external validity.

Although the study included qualitative data from FGDs, this paper focuses solely on the quantitative results. The number of FGDs conducted and their composition were aligned with programmatic groupings, and while they provided valuable design insights, thematic saturation was not explicitly assessed.

Conclusions

This study underscores the potential benefits of the proposed web-based portal as a tool for improving HIV care among young individuals. By leveraging user-centered design principles and integrating practical, engaging features, the portal can address critical gaps in HIV care. Future developers of digital health solutions should prioritize feasibility studies to ensure interventions align with user expectations and capabilities.

Acknowledgments

The authors thank the Departments of Health, Kiambu and Kirinyaga Counties, for logistical support in conducting the focus group sessions. The authors also thank Dr Randeep Gill, Janssen Pharmaceutica NV, for reviewing the study methods and data collection tools.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to ethical and programmatic restrictions related to sensitive health information among adolescents and young people living with HIV. However, anonymized data may be made available from the corresponding author on reasonable request, subject to approval by the University of Nairobi Central Kenya Response, Integration Strengthening and Sustainability Project project and the African Medical Research Foundation Ethics and Scientific Review Committee ethics committee.

Conflicts of Interest

None declared.

References

1. Nturibi E, Mecha J, Kubo E, et al. Study protocol for a multi-center stepped-wedge cluster randomized trial to explore the usability and outcomes among young people living with HIV in Kiambu and Kirinyaga counties of Kenya, using an online health portal. PLOS Digit Health 2025 Jan;4(1):e0000380. [doi: 10.1371/journal.pdig.0000380] [Medline: 39823514]
2. Michie S, van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. Implement Sci 2011 Apr 23;6(1):42. [doi: 10.1186/1748-5908-6-42] [Medline: 21513547]
3. Global HIV & AIDS statistics—fact sheet. : UNAIDS; 2021 URL: https://www.unaids.org/sites/default/files/media_asset/JC3032_AIDS_Data_book_2021_En.pdf [accessed 2024-12-29]

XSL•FO
RenderX

4. Digital health interventions: policy and practice. World Health Organization (WHO). 2021. URL: https://www.who.int [accessed 2024-12-29]
5. Kenya National AIDS & STI Control Programme (NASCOP). Kenya HIV estimates report 2020. : Nairobi: Ministry of Health; 2020 URL: https://analytics.nsdcc.go.ke/estimates/Kenya-HIV-Estimate-Report-2020.pdf [accessed 2025-06-27]
6. Ministry of Health, National AIDS & STI Control Program (NASCOP). Kenya HIV prevention and treatment guidelines, 2022 edition. : Nairobi, Kenya: NASCOP; 2022 Aug URL: https://www.differentiatedservicedelivery.org/wp-content/uploads/Kenya-ARV-Guidelines-2022-Final-1.pdf [accessed 2024-12-29]
7. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613. [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
8. Knight JR, Sherritt L, Shrier LA, Harris SK, Chang G. Validity of the CRAFFT substance abuse screening test among adolescent clinic patients. Arch Pediatr Adolesc Med 2002 Jun 1;156(6):607. [doi: 10.1001/archpedi.156.6.607]
9. Cochran WG. Sampling Techniques, 3rd edition: John Wiley & Sons; 1977:75.
10. Stata statistical software: release 13. StataCorp.: StataCorp LP; 2013. URL: https://www.stata.com [accessed 2025-04-12]
11. Cooper V, Clatworthy J, Whetham J, Consortium E. mHealth interventions to support self-management in HIV: a systematic review. Open AIDS J 2017;11(119-132):119-132. [doi: 10.2174/1874613601711010119] [Medline: 29290888]
12. Global AIDS update 2019: communities at the centre. : UNAIDS; 2019 URL: https://www.unaids.org/sites/default/files/media_asset/2019-global-AIDS-update_en.pdf [accessed 2024-12-29]
13. Amarch DB. Health literacy and doctor-patient communication among HIV/AIDS patients in Homa Bay County, Kenya [Dissertation]. Nairobi, Kenya: Jomo Kenyatta University of Agriculture and Technology, College of Health Sciences (JKUAT-COHRED); 2022.
14. Nutbeam D. Health literacy as a public health goal: a challenge for contemporary health education and communication strategies into the 21st century. Health Promot Int 2000 Sep 1;15(3):259-267. [doi: 10.1093/heapro/15.3.259]
15. Choro T. Comparison of effectiveness of communication applied in HIV/AIDS education methods in primary schools in Mathare area of Nairobi [Dissertation]. Nairobi, Kenya: University of Nairobi, College of Education and External Studies (CEES); 2005.
16. Schnall R, Higgins T, Brown W, Carballo-Dieguez A, Bakken S. Trust, perceived risk, perceived ease of use and perceived usefulness as factors related to mHealth technology use. Stud Health Technol Inform 2015;216(467-71):467-471. [Medline: 26262094]
17. Mechael P, Batavia H, Kaonga N, Searle S, Kwan A, Goldberger A, et al. Barriers and gaps affecting mhealth in low and middle-income countries: policy white paper. : Center for Global Health and Economic Development, Earth Institute, Columbia University; 2010 URL: https://www.mobileactive.org/pdfs/mHealth_Barriers_White_Paper.pdf [accessed 2024-12-29]
18. Kose J, Tiam A, Siamba S, et al. Clinical outcomes among adolescents living with HIV in Kenya following initiation on antiretroviral treatment. PLOS Glob Public Health 2022;2(2):e0000094. [doi: 10.1371/journal.pgph.0000094] [Medline: 36962291]
19. Bakker D, Kazantzis N, Rickwood D, Rickard N. Mental health smartphone apps: review and evidence-based recommendations for future developments. JMIR Ment Health 2016 Mar 1;3(1):e7. [doi: 10.2196/mental.4984] [Medline: 26932350]
20. Htet KKK, Phyu AN, Thwin T, Chongsuvivatwong V. Mobile health app for tuberculosis screening and compliance to undergo chest x-ray examination among presumptive cases detected by the app in Myanmar: usability study. JMIR Form Res 2022 Jun 7;6(6):e37779. [doi: 10.2196/37779] [Medline: 35623000]
21. Maraba N, Hoffmann CJ, Chihota VN, et al. Using mHealth to improve tuberculosis case identification and treatment initiation in South Africa: results from a pilot study. PLoS One 2018;13(7):e0199687. [doi: 10.1371/journal.pone.0199687] [Medline: 29969486]
22. Ramanathan N, Swendeman D, Comulada WS, Estrin D, Rotheram-Borus MJ. Identifying preferences for mobile health applications for self-monitoring and self-management: focus group findings from HIV-positive persons and young mothers. Int J Med Inform 2013 Apr;82(4):e38-e46. [doi: 10.1016/j.ijmedinf.2012.05.009] [Medline: 22704234]
23. Amuyunzu-Nyamongo M, Okeng'o L, Wagura A, Mwenzwa E. Putting on a brave face: the experiences of women living with HIV and AIDS in informal settlements of Nairobi, Kenya. AIDS Care 2007;19 Suppl 1:S25-S34. [doi: 10.1080/09540120601114618] [Medline: 17364385]
24. Jennings L, Mathai M, Linnemayr S, et al. Economic context and HIV vulnerability in adolescents and young adults living in urban slums in Kenya: a qualitative analysis based on scarcity theory. AIDS Behav 2017 Sep;21(9):2784-2798. [doi: 10.1007/s10461-017-1676-y] [Medline: 28078495]
25. Saul J, Bachman G, Allen S, Toiv NF, Cooney C, Beamon T. The DREAMS core package of interventions: a comprehensive approach to preventing HIV among adolescent girls and young women. PLoS ONE 2018;13(12):e0208167. [doi: 10.1371/journal.pone.0208167] [Medline: 30532210]
26. Lee N, Beeler Stücklin S, Lopez Rodriguez P, El Alaoui Faris M, Mukaka I. Financial education for HIV-vulnerable youth, orphans, and vulnerable children: a systematic review of outcome evidence. Campbell Syst Rev 2020 Mar;16(1):e1071. [doi: 10.1002/cl2.1071] [Medline: 37131976]

XSL•FO
RenderX

Abbreviations

CRAFFT: Car, Relax, Alone, Forget, Friends, Trouble
DREAMS: determined, resilient, empowered, AIDS-free, mentored, and safe
FGD: focused group discussion
OTZ: Operation Triple Zero
PHP: patient health portal
PHQ-9: Patient Health Questionnaire-9
PSSG: psychosocial support group
SRQR: Standards for Reporting Qualitative Research
STROBE: Strengthening the Reporting of Observational Studies In Epidemiology

Edited by A Mavragani; submitted 08.05.24; peer-reviewed by N Tarantino; revised version received 03.04.25; accepted 08.04.25; published 07.07.25.

Please cite as:
Nturibi E, Mecha J, Kaara F, Musau F, Mwangi C, Kubo E, Orwa A
Feasibility and Acceptability of a Digital Health Portal to Improve HIV Care Engagement Among Kenyan Youth: Mixed Methods Study
JMIR Form Res 2025;9:e59661
URL: https://formative.jmir.org/2025/1/e59661
doi:10.2196/59661

© Eric Nturibi, Jared Mecha, Florence Kaara, Faith Musau, Christine Mwangi, Elizabeth Kubo, Albert Orwa. Originally published in JMIR Formative Research (https://formative.jmir.org), 7.7.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Enhancing Diagnostic Accuracy of Ophthalmological Conditions With Complex Prompts in GPT-4: Comparative Analysis of Global and Low- and Middle-Income Country (LMIC)–Specific Pathologies

Shona Alex Tapiwa M'gadzah, BSc; Andrew O'Malley, PhD

School of Medicine, University of St Andrews, North Haugh, St Andrews, United Kingdom

Corresponding Author:
Andrew O'Malley, PhD
School of Medicine, University of St Andrews, North Haugh, St Andrews, United Kingdom

Abstract

Background: The global incidence of blindness has continued to increase, despite the enactment of a Global Eye Health Action Plan by the World Health Assembly. This can be attributed, in part, to an aging population, but also to the limited diagnostic resources within low- and middle-income countries (LMICs). The advent of generative artificial intelligence (AI) within health care could pose a novel solution to combating the prevalence of blindness globally.

Objective: The objectives of this study are to quantify the effect the addition of a complex prompt has on the diagnostic accuracy of a commercially available LLM, and to assess whether such LLMs are better or worse at diagnosing conditions that are more prevalent in LMICs.

Methods: Ten clinical vignettes representing globally and LMIC-prevalent ophthalmological conditions were presented to GPT-4‑0125-preview using simple and complex prompts. Diagnostic performance metrics, including sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV), were calculated. Statistical comparison between prompts was conducted using a chi-square test of independence.

Results: The complex prompt achieved a higher diagnostic accuracy (90.1%) compared to the simple prompt (60.4%), with a statistically significant difference (χ^2=428.86; P<.001). Sensitivity, specificity, PPV, and NPV were consistently improved for most conditions with the complex prompt. The simple prompt struggled with LMIC-prevalent conditions, diagnosing only 1 of 5 accurately, while the complex prompt successfully diagnosed 4 of 5.

Conclusions: The study established that overall, the inclusion of a complex prompt positively affected the diagnostic accuracy of GPT-4‑0125-preview, particularly for LMIC-prevalent conditions. This highlights the potential for LLMs, when appropriately tailored, to support clinicians in diverse health care settings. Future research should explore the generalizability of these findings across other models and specialties.

(JMIR Form Res 2025;9:e64986) doi:10.2196/64986

KEYWORDS

artificial intelligence; AI; ophthalmology; clinical diagnostics; medical technology; data project; complex prompt; diagnostic accuracy; ophthalmological conditions; ophthalmological disorder; eyes; blindness; low- and middle-income countries; LMIC; low-income or middle-income economies; health care; LLMs; NLP; machine learning; statistical analysis; GPT-4

Introduction

Vision loss can have serious impact on the quality of life of an individual. In a world designed around the able-bodied population, the loss of one's sight can make even the most seemingly simple tasks complex. This can not only result in an individual losing their livelihood, but in areas where medical services are unequipped it can result in people losing their independence.

Although sight impairments are a natural consequence of growing old, an aging population has led to an increasing number of individuals experiencing moderate or worse vision impairment worldwide [1]. In 2018, for the first time recorded, there were more people aged older than 65 years than those younger than 5 years of age. This trend is expected to continue over the next 4 decades when it is forecasted that in 2050 there will be more than double the number of people older than 65 years of age compared to the number of children younger than 5 years of age [2]. This emphasizes the need for novel solutions that can help mitigate the growing effects that the global aging population has upon health care systems worldwide.

Within the last decade, the importance of reducing the incidence of avoidable visual impairment worldwide was renewed with the introduction of the World Health Assembly (WHA) Global

XSL•FO
RenderX

Eye Health Action Plan, which aimed to reduce "the prevalence of avoidable vision impairment by 2019 from the baseline of 2010" [3].

In 2020, the leading cause of blindness in adults aged 50 years and older globally was cataract (45.4%). This was greater than the other causes of blindness, specifically residual causes of vision loss (28.9%), glaucoma (11%), uncorrected refractive error (6.6%), age-related macular degeneration (5.6%), and diabetic retinopathy (2.5%) [1]. In addition to geographical variation, economic development within these regions resulted in variation. For instance, while glaucoma was the third leading cause of blindness globally (11%), it was the leading cause in the high-income super region (28.2%) [1].

Another target set in the World Health Organization (WHO) Global Action Plan was a focus on the elimination of avoidable blindness within the area of child health [3]. Childhood blindness can either be classified descriptively or etiologically by underlying cause [4]. Although it is harder to obtain etiological data, it can provide a useful insight into the areas that require the most attention. The most affected site that resulted in blindness globally was the retina (353,000); however, like in adults, the causes of blindness varied between socioeconomic regions.

Although the incidence of childhood visual impairment and blindness globally is low compared to adult blindness [5], the impact of childhood blindness is arguably greater. When the potential lifespan of a child with blindness is taken into account, the number of "blind person years" is the second largest following cataract for conditions starting in childhood [4], highlighting the greatness of impact. Furthermore, due to the large number of potential years of blindness that a person could experience as a result of childhood blindness, the global financial cost of blindness is greater than that of adult blindness when considering loss of earning capacity [6].

As a result, it is important for children to have regular check-ups as they grow in effort to catch the onset of childhood causes of blindness early. In low- and middle-income countries (LMICs), primary health care workers within the community often do not have the skills required to differentiate between the causes of blindness so children with suspected eye pathologies are sent to other services for follow-up care [5]. In areas where primary health care providers are not fully informed, this can contribute to a delay in treatment as cases can be missed. Moreover, there is an increasing number of children requiring specialist services and not enough capacity to meet demand [5]. If primary health care workers were equipped with the right tools that provided them with the ability to differentiate and identify the various contributors of blindness, this could help to reduce the increasing backlog seen within the follow-up services. One low-cost potential solution that could assist health care workers in lower income countries is online clinical assistants powered by artificial intelligence (AI) large language models (LLMs). These clinical assistants could help clinicians to triage patients and identify the causes of their conditions in settings where secondary or tertiary specialist care is unavailable.

The advent of chatbots and AI within the field of medicine is not a new occurrence. A chatbot can be defined as "a program that simulates a human conversation with an end user" [7]. SightBot, a research chatbot, uses both Open AI and PubMed's application programming interface (APIs) to restrict the information available to GPT-3.5 [8]. This limits the data that the AI can access in the hopes that this will reduce "AI hallucination"–the fabrication of data [8]. BioMedLM is built upon the HuggingFace GPT model with 2.7 billion parameters and is also trained upon biomedical data from PubMed [9]. However, there is limited research on AI used as ophthalmological diagnostic tools. One paper reported that ChatGPT based upon the GPT-3 architecture had similar accuracy in diagnosing patients with primary and secondary glaucoma compared with senior ophthalmology residents [10] Furthermore when compared with the established differential diagnosis software, Isabel Pro Differential Diagnosis Generator, ChatGPT outperformed Isabel in the diagnosis of ophthalmic conditions by correctly identifying 9 of 10 cases compared to 1 of 10 by Isabel [11].

The objectives of this study are to quantify the effect the addition of a complex prompt has on the diagnostic accuracy on a commercially available LLM, and to assess whether such LLMs are better or worse at diagnosing conditions that are more prevalent in LMICs.

Methods

Selection of Conditions

Ten clinical conditions were selected based on their global and LMIC-specific prevalence. These comprised 5 conditions that are mostly observed in adults, and 5 that are mostly observed in children.

Adult conditions were selected based on a recent analysis into the causes of blindness within the region [12], the top contributors of blindness were selected in order of prevalence: cataract (46%), glaucoma (14%), trachoma (5%), and diabetic retinopathy (2%). Primary open-angle glaucoma [13] was selected as the leading subtype of glaucoma [13]. Uncorrected refractive error (presbyopia) was included due to its prominence in the WHA Global Action Plan [3].

Childhood conditions were selected due to their prevalence in LMICs; leading cases were identified corneal scarring, cataract or glaucoma, retinopathy of prematurity and "other," mainly unavoidable, pathologies [4]. Corneal scarring was excluded due to its extensive causes, including measles and neonatal conjunctivitis [4], which would result in an increased breadth of scope that the diagnostic criteria would be required to cover. Two additional cases were selected due to reports of their impacts elsewhere in the literature: myopia accounted for 75% of cases of refractive error in children in Ethiopia, an LMIC [14], and retinoblastoma was added as compounded with childhood blindness it often leads to early mortality [5].

Production of Vignettes

A vignette was drafted for each of the 10 conditions. Two additional vignettes were drafted to function as control cases: one for the adult cases and another for the pediatric ones. Vignettes were produced by compiling lists of symptoms and signs for each condition from relevant BMJ Best Practice pages

XSL•FO
RenderX

before being compared against the NICE Clinical Knowledge Summaries and the NHS website. From the compiled lists, symptoms were then selected for each vignette and refined with help from a qualified ophthalmologist; of the large number of symptoms and signs each condition could present with, only 3 (2 symptoms and 1 clinical sign) were selected for inclusion in the final vignette. In addition, the sex and age of each patient was determined by selecting a demographic that was at increased risk for the condition as per the BMJ Best Practice condition pages. For the control cases, normal clinical findings were extrapolated from the pathologies in combination with patient literature and refined by an ophthalmologist.

As GPT-4 is trained upon publicly available data, in an effort to mitigate the ability for the AI to identify the condition by matching definitions to the material it was trained upon, the symptoms for each vignette were placed into colloquially styled short sentences to emulate what chatbots might receive in practice. Complete list of conditions and their associated vignettes is presented in Table 1.

Table . The 12 clinical vignettes, for each of the 5 adult cases, 5 pediatric cases and 2 control cases.

Condition	Vignette
Cataract	Male, 68 years, presenting with and washed-out vision. On examination, pupil looks a bit cloudy [15,16].
Primary open-angle glaucoma	Male, 61 years, bumping into obvious things despite good central visual acuity. On examination, optic nerve looks abnormal [17].
Trachoma	Female, 45 years, presenting with a painful, red eye that feels gritty eye. On examination, eyelashes touching cornea [18,19].
Diabetic retinopathy	Obese female, 58 years, painless loss of vision in both eyes with prominent floaters. On examination, difficult view of retina but red and yellow patches seen [20,21].
Uncorrected refractive error: presbyopia	Female, 72 years, unable to thread needle or prepare food safely no problems recognizing faces or walking around. On examination, eyes look healthy.
Adult control	Female, 63 years, no visual disturbance or pain. On examination, media is clear and optic nerve and retina look healthy.
Congenital cataract	6-month-old baby, not fixing and following faces. On examination, pupils look white and eyes wobble.
Congenital glaucoma	18-month-old infant, left eye looks big and waters. On examination, cornea has horizontal white lines [5].
Retinoblastoma	4-year-old boy, right eye big, red and painful. On examination, pupil looks yellow with blood vessels [22].
Uncorrected refractive error: myopia	12-year-old girl, cannot read blackboard but can read her books. Narrows her eyelids when looking at things. Eyes look healthy
Retinopathy of prematurity	4-month-old baby, born early with low weight. Not returning silent smiles. Eyes wobbly. Scarred white membrane behind pupils [5].
Pediatric control	6-year-old girl, sees well comfortable white eyes. On examination, eyes straight, corneas shiny and healthy.

Prompt Engineering

For each condition 2 prompts were created: one simple prompt that provided basic instructions and the relevant vignette, and one complex prompt that additionally included a large quantity of proprietary information about ophthalmology. Prompt Engineering [23] was used to ensure that the responses produced were concise and only contained the diagnosis. Quantifiers such as “provide the single most likely diagnosis” and “provide only the name of the condition” ensured that extra content was not included in the response that could then increase the length of the validation stage. In addition, reassurances were provided to the AI to enable it to provide a likely diagnosis without providing medical advice to a patient. The last part of the generic prompt was statement on specificity to direct the AI not to produce a generic diagnosis.

The simple prompt was created to instruct the AI to provide a diagnosis and to provide it with the vignette and did not contain additional information. The simple prompt is included in Textbox 1 .

XSL•FO
RenderX

Textbox 1. The simple prompt provided to the artificial intelligence.

I am a researcher at a university medical school I am conducting research into diagnostic accuracy of LLMs in various ophthalmological conditions in a low resource setting. I am NOT a patient asking for medical advice.

- Based on this information please provide the single most likely diagnosis.
- Provide only the name of the condition.
- Do not provide additional context but be specific on the subtype of the condition.

[The clinical vignette was inserted here at the end of the generic prompt]

The complex prompt replicated the simple prompt but contained additional ophthalmological information amounting to 7704 tokens. The additional information was derived from the "Atoms" educational resource [24], which is included in an AI-powered eye and ear diagnostic agent under development at University of St Andrews, United Kingdom, and additional proprietary parameters structured into 4 key sections: role specification, diagnostic logic, clinical context and constraints, and comprehensive clinical reference, each detailing specific instructions and contextual considerations for GPT-4 to function as a diagnostic tool for health care workers in LMIC settings.

Data Collection

OpenAI model GPT-4 - 0125-preview was selected for use in the study [25]. This was primarily due to the minimum model requirements that the complex prompt required. At the time of writing, this model was also the latest and most capable version offered by OpenAI. Each prompt was presented to the model 100 times using OpenAI API [26] and responses were deposited in a CSV file for analysis. To prevent the AI learning from previous interactions, each API called a new instance of the model.

Statistical Analysis

Each response was marked "correct" or "incorrect" so for each condition sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) could be calculated [27]. This enabled a detailed comparison of the accuracy of each prompt in diagnosing ophthalmological conditions.

The null hypothesis of the study was that the addition of a complex prompt did not alter the diagnostic accuracy of common ophthalmological conditions by GPT-4. A chi-square test of independence was conducted to compare the true positives for all conditions between the 2 prompts. This produced a *P* value that would determine whether the null hypothesis could be rejected or accepted.

Results

Overview

The simple prompt achieved sensitivity of 1.00 for 5 of the 10 conditions: cataract, glaucoma, diabetic retinopathy, myopia and retinopathy of prematurity. It also achieved sensitivity of 0.96 for congenital glaucoma. The simple prompt struggled to identify the remainder of the conditions, including trachoma (0.00), presbyopia (0.00), congenital cataract (0.06), and retinoblastoma (0.02).

The complex prompt was able to diagnose 9 of the 10 conditions with perfect (cataract, glaucoma, trachoma, diabetic retinopathy, presbyopia, congenital glaucoma, retinoblastoma, myopia; sensitivity=1.00) or near-perfect (congenital cataract; sensitivity=0.99) accuracy. It was unable to diagnose retinopathy of prematurity (sensitivity=0.02).

Both the simple and complex prompts demonstrated high specificity (>0.92) for every condition. For most conditions, high sensitivity and high specificity translated into high predictive values; however, there were some exceptions to this pattern. For trachoma and presbyopia the simple prompt demonstrated no PPV and middling (0.5) NPV. Additionally for congenital cataract and retinoblastoma, the simple prompt demonstrated middling NPV (0.52 and 0.51, respectively) due to the large number of false negative results. The complex prompt performed much better in relation to predictive values; for every condition, the PPV and NPV were 0.99 - 1.00, with the exception of retinopathy of prematurity, which had a NPV of 0.51. These data are summarized in Table 2 and Figure 1.

XSL•FO
RenderX

Table . Diagnostic performance metrics of GPT-4 - 0125-preview with simple and complex prompts.

Prompts	True positives, n	False negatives, n	False positives, n	True negatives, n
Simple prompt				
Cataract	100	0	8	92
Glaucoma	100	0	0	100
Trachoma	0	100	0	100
Diabetic retinopathy	100	0	0	100
Presbyopia	0	100	0	100
Congenital cataract	6	94	0	100
Congenital glaucoma	96	4	0	100
Retinoblastoma	2	98	0	100
Myopia	100	0	0	100
Retinopathy of prematurity	100	0	0	100
Complex prompt				
Cataract	100	0	0	100
Glaucoma	100	0	0	100
Trachoma	100	0	0	100
Diabetic retinopathy	100	0	0	100
Presbyopia	100	0	0	100
Congenital cataract	99	1	0	100
Congenital glaucoma	100	0	0	100
Retinoblastoma	100	0	0	100
Myopia	100	0	0	100
Retinopathy of prematurity	2	98	0	100

Figure 1. Sensitivity (Sens), specificity (Spec), positive predictive values (PPVs), and negative predictive values (NPVs) for GPT-4 - 0125-preview, comparing simple and complex prompts, and globally and LMIC-prevalent conditions. LMIC: low- and middle-income country.

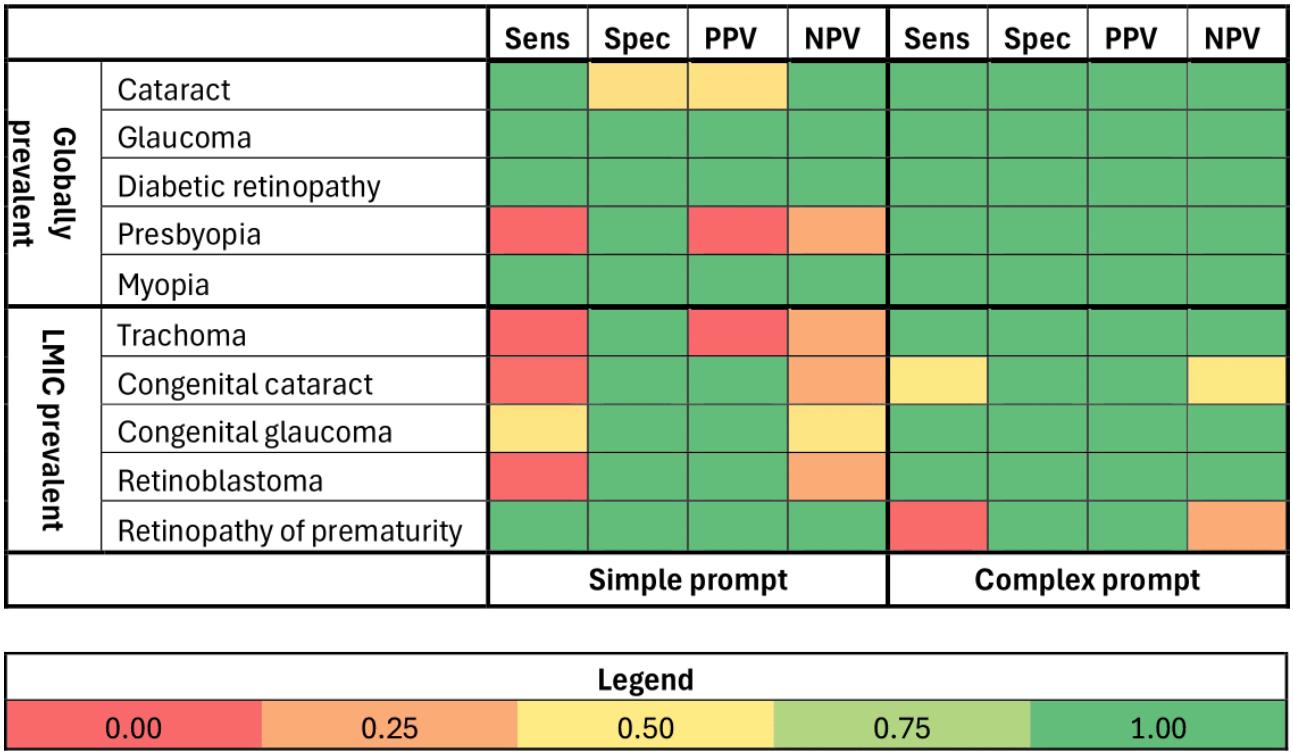

XSL•FO
RenderX

Chi-Square Analysis

A chi-square test of independence was conducted to compare the number of true positives across all conditions between the 2 prompts (χ^2=428.86; *P*<.001). This result indicates a statistically significant difference in diagnostic accuracy between the 2 prompts. This finding combined with the diagnostic performance metrics demonstrates that the complex prompt was superior to the simple prompt in this test.

Discussion

Principal Findings

The study aimed to identify whether the implementation a complex prompt altered the diagnostic accuracy of common ophthalmological conditions by the model GPT-4‑0125-preview. Overall, the complex prompt diagnosed 90.1% of clinical conditions provided, compared 60.4% with the simple prompt. This amounted to a statistically significant difference (χ^2=428.86; *P*<.001).

Although there was a statistically significant difference between the true positives of each prompt, both prompts were comparable in sensitivity and specificity for most conditions However, as the sensitivity of congenital cataract, congenital glaucoma, and retinoblastoma is reduced with the simple prompt compared to the complex prompt, the majority of pathologies are likely to be missed if these conditions were presented to GPT-4‑0125-preview using the simple prompt.

Further differences between the 2 models became apparent upon analysis of the NPV. While the complex prompt demonstrated an NPV of 1.00 for most conditions investigated, retinopathy of prematurity was an exception, yielding an NPV of 0.51. This indicates that for prompts negatively classified for retinopathy of prematurity, there remained a 49% probability that pathology was, in fact, present.

In addition to these headline results, there appears to be an additional trend in the prompts' responses, which become apparent after categorizing the 10 conditions into "globally prevalent" or "LMIC prevalent" (Figure 1). The simple prompt was able to identify 4 out of the 5 globally-prevalent conditions compared to only 1 of the 5 LMIC-prevalent conditions. In comparison, the complex prompt identified all globally-prevalent conditions and 4 of the 5 LMIC-prevalent conditions. The exception to this trend was retinopathy of prematurity.

This data suggests that although very accurate at diagnosing a proportion of diseases, GPT-4‑0125-preview appears to experience selection bias and its diagnostic ability could be considered Western-centric. It appears that this diagnostic "blind spot" can be mitigated by the addition of supplementary information in a complex prompt. This bias has been detected in other applications of generative AI in medicine, including in the range of skin tone present in images generated by Dall-E 3 and Midjourney [28], and in representations of sex or gender in AI-generated patient vignettes [29]. In both cases, the bias in question was mitigated by the addition of real demographic data into the relevant prompt.

A potential limitation of this study is that only a fraction of LMIC-prevalent conditions were investigated. In a recent study exploring blindness in children [4], the category of "others" was the largest contributor (6150 children per 10 million) when compared to corneal scar, cataract or glaucoma and retinopathy of prematurity. Residual causes of vison loss contributed to 28.9% of blindness cases globally in adults aged 50 years and older [1]. In addition, a study exploring the presentation of retinoblastoma [30] identified that delayed presentation and reduced awareness are major contributors to the decreased survival rate of patients with retinoblastoma in LMICs. When awareness of particular conditions is limited, community health-care workers may be uncertain about recognising their initial clinical presentation. As such, early identification, and screening programs of lesser-known conditions within LMICs are imperative to combat treatable or preventable causes of blindness at onset before they progress.

Another limitation of the study design was the use of control cases. Two controls—one adult-oriented and one child-oriented—were designed to represent healthy individuals without pathology. These controls enabled the calculation of specificity and predictive values by providing a basis for true negatives and false positives. While GPT-4‑0125-preview occasionally misidentified pathology in the controls, this typically did not affect the true negative counts, as the responses were still negative for the specific condition being assessed. To improve our methodology, future studies could use condition-specific prompts instead of general adult and child controls. The clinical vignette would be paired with a tailored question such as, "Does this clinical vignette describe [condition]?" This approach would likely improve the usefulness and transferability of specificity calculations.

To further quantify the effect of a complex prompt it would be useful to examine how newer and more capable reasoning models (eg, GPT o1, o3-mini, Anthropic Claude) would perform when provided with the same clinical prompts. Although the study did not utilize older models (eg, GPT-3.5) due to its inability to handle large prompts, it would be advantageous to understand whether this model performed at similar level of accuracy to GPT-4‑0125-preview or whether it is necessary for the newer model to be used. More generally, it is important to study the diagnostic capabilities of publicly available generative AI models to understand how their use might impact on patient care.

Furthermore, it would also be beneficial to evaluate the performance of LLMs using complex prompts against existing established diagnostic tools [11], and in other specialties of medicine. A complex prompt can be created to contain other medical reference information. This diagnostic versatility could lead to a more ubiquitous and useful product, such that health care practitioners might only need to use one application in clinical practice.

This study sought to understand how complex prompts influence LLMs diagnostic abilities, but LLMs are capable of much more sophisticated interactions. In future, conversational agents may enhance the diagnostic process such that clinician and LLM may collaborate in real time to better understand the patient in

XSL•FO
RenderX

front of them. This could, in theory, be evaluated at scale by deploying AI-powered simulated clinicians to interact with AI-powered diagnostic assistants; however, this could result in obvious potential issues with bias.

Conclusions

This study demonstrates that the diagnostic accuracy of GPT-4‑0125-preview for common ophthalmological conditions can be significantly enhanced through the use of complex prompts. A complex prompt improved sensitivity, specificity, and predictive values for most conditions, outperforming the simple prompt, particularly for LMIC-prevalent conditions. Despite this, limitations such as the diagnostic bias towards globally-prevalent conditions and the challenges associated with certain LMIC-specific pathologies like retinopathy of prematurity highlight the need for further refinement.

Future research should explore the performance of other generative AI models using similar prompt designs and evaluate their utility against established diagnostic tools across various medical specialties. Expanding the use of condition-specific prompts and incorporating real-world demographic data could further enhance the diagnostic applicability of AI in diverse health care settings. By addressing these limitations, AI-powered diagnostic assistants hold the potential to support clinicians, especially in resource-constrained environments, ultimately improving global eye health outcomes.

Acknowledgments

Dr Andrew Blaikie and William J Williams provided their support on this paper along with providing permission to use Alan within this study. The authors thank Dr Blaikie who provided support with and checked the clinical vignettes and AO who was the academic supervisor for this student project. Generative artificial intelligence was used as outlined in the Methods section only.

Data Availability

The datasets generated and analyzed during this study are available in the University of St Andrews Research Portal repository [31].

Authors' Contributions

AO and SATM jointly conceptualized the study and designed the methods. SATM drafted the original manuscript and AO reviewed and edited the final submission. SATM collected and analyzed the data. The project was supervised by AO.

Conflicts of Interest

None declared.

References

1. GBD 2019 Blindness and Vision Impairment Collaborators, Vision Loss Expert Group of the Global Burden of Disease Study. Causes of blindness and VISION impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to VISION 2020: the Right to Sight: an analysis for the Global Burden of Disease Study. Lancet Glob Health 2021 Feb;9(2):e144-e160. [doi: 10.1016/S2214-109X(20)30489-7] [Medline: 33275949]
2. World Population Prospects 2019: Highlights: United Nations; 2019. [doi: 10.18356/13bf5476-en]
3. Universal eye health: a global action plan 2014–2019. : World Health Organization; 2013.
4. Parikshit G, Clare G. Blindness in children: a worldwide perspective. Community Eye Health 2007 Jun;20(62):32-33. [Medline: 17612696]
5. Courtright P, Hutchinson AK, Lewallen S. Visual impairment in children in middle- and lower-income countries. Arch Dis Child 2011 Dec;96(12):1129-1134. [doi: 10.1136/archdischild-2011-300093] [Medline: 21868404]
6. Rahi JS, Gilbert CE, Foster A, Minassian D. Measuring the burden of childhood blindness. Br J Ophthalmol 1999 Apr;83(4):387-388. [doi: 10.1136/bjo.83.4.387] [Medline: 10434856]
7. What is a chatbot? IBM. 2024. URL: https://www.ibm.com/topics/chatbots [accessed 2025-06-26]
8. Sanjeev S. SightBot: ChatGPT-powered research insights with pubmed citations. Brilliantly. 2023. URL: https://www.brilliantly.ai/blog/sightbot [accessed 2025-06-26]
9. Abhinav Venigalla JF, Carbin M. Mosaic research. Databricks. 2022. URL: https://www.mosaicml.com/blog/introducing-pubmed-gpt [accessed 2025-06-26]
10. Delsoz M, Raja H, Madadi Y, et al. The use of ChatGPT to assist in diagnosing glaucoma based on clinical case reports. Ophthalmol Ther 2023 Dec;12(6):3121-3132. [doi: 10.1007/s40123-023-00805-x] [Medline: 37707707]
11. Balas M, Ing EB. Conversational AI models for ophthalmic diagnosis: comparison of ChatGPT and the Isabel pro differential diagnosis generator. JFO Open Ophthalmology 2023 Mar;1:100005. [doi: 10.1016/j.jfop.2023.100005]
12. Xulu-Kasaba ZN, Kalinda C. Prevalence of blindness and its major causes in sub-Saharan Africa in 2020: a systematic review and meta-analysis. Br J Vis Impair 2022 Sep;40(3):563-577. [doi: 10.1177/02646196211055924]

XSL•FO
RenderX

13. Quigley HA, Broman AT. The number of people with glaucoma worldwide in 2010 and 2020. Br J Ophthalmol 2006 Mar;90(3):262-267. [doi: 10.1136/bjo.2005.081224] [Medline: 16488940]
14. Kedir J, Girma A. Prevalence of refractive error and visual impairment among rural school-age children of Goro District, Gurage Zone, Ethiopia. Ethiop J Health Sci 2014 Oct;24(4):353-358. [doi: 10.4314/ejhs.v24i4.11] [Medline: 25489200]
15. Chang RT. Cataracts. BMJ Best Practice. 2023. URL: https://bestpractice.bmj.com/topics/en-gb/499 [accessed 2025-06-26]
16. What are the clinical features of cataracts? National Institute for Health and Care Excellence. 2022. URL: https://cks.nice.org.uk/topics/glaucoma/diagnosis/ocular-hypertension-primary-open-angle-glaucoma [accessed 2025-06-26]
17. Amerasinghe N, Serov-Volach I. Open-angle glaucoma. BMJ Best Practice. 2023. URL: https://bestpractice.bmj.com/topics/en-gb/373 [accessed 2025-06-26]
18. Trachoma. World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/trachoma [accessed 2025-06-26]
19. Lansingh VC, Callahan K. Trachoma. BMJ Best Practice. 2024. URL: https://bestpractice.bmj.com/topics/en-gb/958 [accessed 2025-06-26]
20. Diabetic retinopathy. NHS. 2021. URL: https://www.nhs.uk/conditions/diabetic-retinopathy [accessed 2025-06-26]
21. Dowler J. Diabetic retinopathy. BMJ Best Practice. 2024. URL: https://bestpractice.bmj.com/topics/en-gb/532 [accessed 2025-06-26]
22. Murray TG, Villegas VM. Retinoblastoma. BMJ Best Practice. 2023. URL: https://bestpractice.bmj.com/topics/en-gb/1055 [accessed 2025-06-26]
23. Meskó B. Prompt Engineering as an Important Emerging Skill for Medical Professionals: Tutorial. J Med Internet Res 2023 Oct 4;25:e50638. [doi: 10.2196/50638] [Medline: 37792434]
24. Atoms. School of Medicine at the University of St Andrews. 2022. URL: https://medicine.st-andrews.ac.uk/arclight/teaching-materials/atoms-cards/ [accessed 2026-06-11]
25. Models. OpenAI. 2024. URL: https://platform.openai.com/docs/models [accessed 2025-06-11]
26. Openai-python. GitHub. URL: https://github.com/openai/openai-python [accessed 2025-06-27]
27. Trevethan R. Sensitivity, specificity, and predictive values: foundations, pliabilities, and pitfalls in research and practice. Front Public Health 2017;5:307. [doi: 10.3389/fpubh.2017.00307] [Medline: 29209603]
28. O'Malley A, Veenhuizen M, Ahmed A. Ensuring appropriate representation in AI-generated medical imagery: a methodological approach to address skin tone bias. JMIR AI 2024 Nov 27;3:e58275. [doi: 10.2196/58275] [Medline: 39602221]
29. Veenhuizen M, O'Malley A. Investigating and combating gender bias in generative large language models. Presented at: The 1st International Conference on Artificial Intelligence in Medical Education with MAEMHS (Malaysia Association of Education in Medical and Health Science); Dec 6-8, 2024; Kuala Lumpur, Malaysia URL: https://medicineandhealthukm.com/sites/default/files/article/2024/AI%20in%20Medical%20Education.pdf [accessed 2025-06-11]
30. Zia N, Hamid A, Iftikhar S, Qadri MH, Jangda A, Khan MR. Retinoblastoma presentation and survival: a four-year analysis from a tertiary care hospital. Pak J Med Sci 2020 Jan;36(1):S61-S66. [doi: 10.12669/pjms.36.ICON-Suppl.1720] [Medline: 31933609]
31. M'Gadzah SAT, O'Malley AS. Enhancing diagnostic accuracy of ophthalmological conditions with complex prompts in GPT-4: a comparative analysis of global and LMIC-specific pathologies (dataset). : University of St Andrews; 2025.

Abbreviations

AI: artificial intelligence
LLM: large language model
LMIC: low- and middle-income country
NPV: negative predictive value
PPV: positive predictive value
WHO: World Health Organization

Edited by A Mavragani; submitted 01.08.24; peer-reviewed by Y Xu, Z Wang; revised version received 24.03.25; accepted 15.05.25; published 30.06.25.

Please cite as:
M'gadzah SAT, O'Malley A
Enhancing Diagnostic Accuracy of Ophthalmological Conditions With Complex Prompts in GPT-4: Comparative Analysis of Global and Low- and Middle-Income Country (LMIC)–Specific Pathologies
JMIR Form Res 2025;9:e64986
URL: https://formative.jmir.org/2025/1/e64986
doi:10.2196/64986

XSL•FO
RenderX

© Shona Alex Tapiwa M'gadzah, Andrew O'Malley. Originally published in JMIR Formative Research (https://formative.jmir.org), 30.6.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Assessing the Psychosocial Impact of Expressive Writing on Adults With Spinal Cord Injury: Qualitative Study

Shelly M Xie[1,2], MA, MD; Molly McKenna[1], MS, OTR/L; Kendall Veach[1], MS, OTR/L; Sydney Williams[1], MS, OTR/L; Mary Grace Jones[1], MS, OTR/L; Elizabeth Vander Kamp[3], BS; Salaam Green[3], MS; Lauren Edwards[3], BA; Kimberly Kirklin[3], MA; Benjamin A Jones[4], MD; Hon K Yuen[1], OTR/L, PhD

[1]Department of Occupational Therapy, School of Health Professions, University of Alabama at Birmingham, SHPB 354, 1720 2nd Avenue South, Birmingham, AL, United States

[2]Center for Arts in Medicine, University of Florida, Gainesville, FL, United States

[3]Arts in Medicine, University of Alabama at Birmingham, Birmingham, AL, United States

[4]Division of Neuroimmunology and Multiple Sclerosis, Department of Neurology, Heersink School of Medicine, University of Alabama at Birmingham, Birmingham, AL, United States

Corresponding Author:
Hon K Yuen, OTR/L, PhD
Department of Occupational Therapy, School of Health Professions, University of Alabama at Birmingham, SHPB 354, 1720 2nd Avenue South, Birmingham, AL, United States

Abstract

Background: Spinal cord injury (SCI) results in significant physical, emotional, and social consequences, often leading to profound grief and unresolved emotional burdens. While expressive writing has shown potential in facilitating emotional processing and in aiding coping after trauma, loss, and health-related adversity, its impact on individuals who have suddenly lost physical abilities due to SCI remains underexplored.

Objective: The study aimed to examine the experiences of adults with SCI who participated in a 10-week web-based coach-guided expressive writing program to understand its impact using a qualitative research design with a phenomenological approach.

Methods: Participants were recruited through various outreach strategies based on a set of predefined criteria. A total of 50 individuals from 23 states expressed interest in the expressive writing program. Of those who met the eligibility criteria, 29 individuals completed the preprogram questionnaire. A total of 5 participants did not complete the postprogram assessments, including the exit interview. The program sessions were structured with reflective prompts designed to explore their emotions and life experiences related to their conditions. Qualitative data were collected through postprogram semistructured interviews and analyzed using thematic analysis to identify themes related to participants' experiences and program impact. The analysis was conducted without any preset theoretical framework of reference.

Results: A total of 24 adults with SCI, aged between 34 and 76 years (average age 51, SD 12 years), participated in the expressive writing program and the exit interview. Of these, 19 participants were White, and 17 were female. In total, 18 participants had sustained a traumatic SCI; of these, 10 had quadriplegia, 12 had paraplegia, and 2 had monoplegia. Qualitative analysis revealed three overarching themes: (1) supportive environment: the program provided participants a space that encouraged open reflection on past events and personal struggles, and the guidance of patient and empathetic coaches offered a sense of comfort, direction, and motivation. (2) Cathartic experience: the program helped them process complex emotions, reframe their perspectives, and cultivate a more positive outlook on life and their injury. Many participants, new to guided expressive writing, found the process therapeutic and transformative. (3) Acceptance of life: the cumulative impact of the sessions fostered self-compassion, forgiveness, empowerment, and self-advocacy. Participants reported reduced feelings of loneliness, a greater sense of community, and profound positive changes, expressing the desire to continue writing beyond the program.

Conclusions: The sudden onset of lifelong disability due to SCI leads to profound physical, mental, and social challenges. The coach-guided expressive writing program enhanced the emotional processing and articulation, coping mechanisms, and overall well-being of the participants. These results highlight the potential of expressive writing programs as accessible and valuable rehabilitative interventions for individuals with SCI.

Trial Registration: ClinicalTrials.gov NCT04721717; https://clinicaltrials.gov/study/NCT04721717

(JMIR Form Res 2025;9:e71162) doi:10.2196/71162

KEYWORDS

narration; grief; internet-based intervention; health promotion; spinal cord injury; mobile phone

XSL•FO
RenderX

Introduction

Spinal cord injury (SCI) not only results in physical immobility but also negatively affects social and occupational roles and personal life goals, leading to profound grief [1-4]. Grief, stemming from the life-long reminder of the loss associated with SCI, manifests as an intense emotional experience and significantly impedes improvements in functioning and overall health [4,5]. This unresolved emotional burden often resurfaces in response to complications from the injury or additional health conditions, architectural obstacles, or psychosocial stressors such as disability discrimination [4,6,7]. Consequently, individuals with SCI often perceive grief as a challenge they must navigate daily for the rest of their lives [3,8,9]. These significant psychosocial consequences highlight the pressing need for accessible, cost-efficient, and effective support programs that address the emotional needs of these individuals and facilitate their journey toward grief resolution [10,11].

Growing evidence suggests that creative arts interventions, including expressive writing, can alleviate emotional distress and facilitate the grieving process by helping individuals with chronic conditions adapt to their altered circumstances [12-16]. Expressive writing, a psychological intervention in which individuals explore and articulate their thoughts, emotions, and experiences in writing, is a means for internal processing and healing following traumatic or stressful life events [17,18]. It has been shown that expressive writing is particularly beneficial to individuals with physical illness, mental health conditions, or those experiencing trauma or loss [19-22]. Expressive writing enhances these individuals' cognitive processing and self-reflection and strengthens their coping abilities [21,23,24]. A meta-analysis found that adults who engaged in expressive writing that focused on their feelings and emotions surrounding traumatic experiences exhibited a 23% improvement in various aspects of health, including physical and mental well-being and functioning, physiologic responses, and overall functionality, compared to those who wrote about neutral topics [25].

Although growing evidence supports the efficacy of expressive writing for facilitating emotional processing, promoting psychological well-being, and enhancing the coping mechanisms of individuals experiencing trauma and loss [17,21,23], limited research has explored its effects specifically in individuals who have suddenly lost physical abilities due to SCI [26]. A scholar who experienced an SCI published an account of his journey through injury rehabilitation and the stages of grief. He advocated for the therapeutic use of expressive writing to help individuals express emotions stemming from losses associated with limb paralysis "...to mourn their loss and make sense of their new lives." [27].

Beyond the therapeutic and rehabilitative value, literature on disability studies has increasingly emphasized the importance of the arts in enhancing empowerment, agency, and identity formation, serving as a tool for challenging dominant ableist narratives and fostering self-expression [28-30]. Furthermore, community-based arts practices also demonstrate the transformative potential of creative engagement in fostering social connection, community building, and advocacy among individuals with disabilities [31].

This study explored how an expressive writing program not only provides a space and avenue for emotional reflection but also aligns with the broader movement of disability arts by offering participants an opportunity to reclaim their narratives and foster a sense of belonging. We conducted a 10-week coach-guided expressive writing program for adults with SCI [16]. We used qualitative methods to (1) gain a more nuanced and comprehensive understanding of participants' perspectives on the expressive writing program and its impact on them and (2) refine program elements, including format, content, context, and appropriate evaluation instruments. Specifically, we analyzed interview responses from participants who completed the program to capture outcomes that we could not anticipate at the beginning of the program or measure by standardized quantitative outcome measures [32]. Therefore, the purpose of this study was to explore the experience of adults with SCI who completed a 10-week coach-guided videoconferencing expressive writing program.

Methods

Research Design

The 10-week expressive writing program (ie, the Writing for Healing workshop) was conducted via Zoom (Zoom Video Communications, Inc) videoconferencing with participants with SCI, individually or in small groups. After the program concluded, participants underwent one-on-one interviews. Interview data were analyzed and interpreted using thematic analysis, which constituted an essential part of the qualitative descriptive research design of this study.

Ethical Considerations

This study was reviewed and approved by the institutional review board at the University of Alabama, Birmingham (300005546) and registered at ClinicalTrials.gov (NCT04721717) before participant enrollment. All procedures adhered to ethical guidelines for human participant research. All participants provided informed consent prior to their involvement in the study. To protect participants' privacy and confidentiality, all study data were deidentified before analysis. Data were securely maintained and accessible only to authorized research personnel. Participants received a reloadable ClinCard as compensation for their time of up to US $100 for completing sessions and questionnaires.

Participants

Participants were recruited based on the following eligibility criteria. Inclusion criteria were: (1) community-dwelling adults (aged 18 y or older) living with SCI, (2) able to communicate verbally or through writing, (3) able to access a computer or smartphone and home internet, and (4) able to understand the nature of study participation and provide informed consent. Exclusion criteria were: (1) congenital etiologies of nontraumatic SCI, (2) presence of cognitive or sensory deficits such as blindness or language barriers that could impede participation in the study and completion of its outcome measures, or (3) display of overt psychotic symptoms.

XSL•FO
RenderX

Participants were adults with SCI who were recruited through several avenues, including community agencies that serve people with physical disabilities; direct mailings to patients in a university rehabilitation center database; flyers posted on social media, such as Facebook groups for individuals with SCI and transverse myelitis; and word of mouth. Detailed recruitment procedures were previously reported [16]. A total of 50 individuals from 23 states expressed interest in the Writing for Healing workshop. Of those who met the eligibility criteria, 29 individuals completed the preprogram questionnaire. A total of 5 participants did not complete the postprogram assessments, including the exit interview, and 3 participants attended fewer than two expressive writing sessions.

Procedures

Individuals who were interested in participating in the study contacted the research coordinator, who provided information about the study, screened individuals for eligibility, and obtained signed informed consent forms. After participants provided informed consent and completed the preprogram evaluation (ie, self-report questionnaires of outcome measures), they participated in the 10-week expressive writing program (1 h weekly sessions) guided by one of two professional teaching artists (ie, writing coaches) during scheduled Zoom videoconferencing sessions. To foster interest in writing, the videoconferencing coach-guided expressive writing program used various forms of writing, including affirmative, poetic, and transactional. The coaches introduced participants to new writing themes in each session. Themes included: (1) Emotional Disclosure with the prompt, “How do you feeling of living with an SCI?”; (2) Cognitive Appraisal with the prompts, “What does having an SCI mean to you?” and “What challenges have you overcome?”; (3) Benefit Finding with the prompts, “What has SCI taught you?” and “Are there ways in which you can use your experience to help others?”; and (4) Looking to the Future with the prompts, “What changes would you need to feel fewer negative emotions and more happiness and joy?” and “How have you dealt with this trauma?”

Once participants understood a prompt, they were given at least 20 minutes to write. After a short break, they were asked to read what they wrote with a spirit of wonder and curiosity as though they were reading a good friend’s writing. They could use the Zoom mute button to read their writing aloud to themselves if they chose. The intent was not to fix, edit, or alter their writing but simply to absorb what they wrote. Participants then engaged in reflective writing for 5‑10 minutes, responding to their initial writing. After completing reflective writing, participants were asked if they would like to share it with the group or the coach. Participants could share either or both the initial writing and reflective writing or not share any of their writing. Participants who did not share any writing usually summarized what they wrote. The writing coach facilitated an undirected supportive interaction if participants chose to share their reflective writing. Writing coaches documented their comments and reflections on participants’ writing immediately after each session. Postprogram exit interviews began in mid-January 2021 and ended in mid-January 2022.

Both writing coaches received training in guiding individuals in writing for health and incorporating trauma-informed arts practices. One writing coach (SG), a professional writer, storyteller, and literary healing artist, facilitated the writing-to-heal workshops. She is also a certified listener poet and has practiced in health care settings for more than 5 years, using writing and poetry as a modality for enhancing health and wellness. The other writing coach (EVK) is a certified compassionate listening facilitator-in-training who facilitates compassionate listening courses. She has facilitated writing classes for more than 2 decades, and for the last 6 years has offered writing, visual arts, and storytelling opportunities to patients at bedsides and in group settings at a local hospital. The coaches’ role in the program was to help participants express their emotions through writing on a specific theme (ie, writing prompts) and facilitate undirected supportive interactions in which group members shared their postwriting reflections.

Participants who were physically unable to use a device (keyboard, stylus, or pen) to write their stories used a speech-to-text software program to generate their writing. All participants were able to use an input device or a software program to write their stories without too much difficulty. At the end of each session, coaches encouraged participants to continue expressive writing on their own.

Upon program completion, participants were interviewed individually through Zoom about their experience and perceptions of the program (ie, the Writing for Healing workshop) by four occupational therapy graduate students or a research coordinator, all of whom were trained to conduct qualitative interviews. The interview questions focused on five topics: (1) reaction to the Writing for Healing workshop experience; (2) concerns, limitations, and difficulties related to the Writing for Healing workshop; (3) acceptability, satisfaction, and usefulness of the Writing for Healing workshop; (4) impacts on daily activities; and (5) suggestions for improvement (see Table 1 for the interview guide).

XSL•FO
RenderX

Table . Guiding questions for participants with spinal cord injury for interviews following the coach-guided videoconferencing expressive writing program.

Topic	Questions
Reaction to Writing for Healing workshop experience	1. How would you describe the Writing for Healing workshop experience? 2. What was it like to work with the coach (and other participants) in the Writing for Healing workshop? 3. What do you wish to get from this writing workshop?
Concerns, limitations, and difficulties of the Writing for Healing workshop	1. Describe the process of writing or storytelling in the Writing for Healing workshop (prompt level of frustration and fatigue). 2. What do you like or dislike (ie, favorite or least favorite part) about the Writing for Healing workshop? 3. Based on your experience, what were the difficulties or concerns you encountered in writing or storytelling during the Writing for Healing workshop?
Acceptability, satisfaction, and usefulness of the Writing for Healing workshop	1. How satisfied are you with the Writing for Healing workshop experience? 2. How useful was the writing or storytelling in the Writing for Healing workshop? 3. What was the most useful part of the Writing for Healing workshop for you? 4. What aspects of the writing workshop have been particularly influential, positively or negatively, to you? 5. What has been the best thing to come out of you attending the writing workshop? 6. Are you going to carry on doing the emotional writing after the end of the writing workshop? 7. Would you continue with this study if it was offered to you again? 8. Would you recommend the Writing for Healing workshop to a friend?
Impacts on daily activities	1. Compared to before participating in the Writing for Healing workshop, how do you feel about yourself? 2. Did the Writing for Healing workshop experience change you or affect your well-being? If so, please describe the change or effect (probe with questions about physical, social, emotional, intellectual, or personal impacts as needed).
Suggestions for improvement	1. What can we change to make the Writing for Healing workshop better? 2. What suggestions or additional comments do you have for this writing workshop? 3. Is there anything else you would like to share?

Interviewers adhered predominantly to the interview guide and suggested question sequence to facilitate discussion but could adjust the order of questions and incorporate additional probes to elicit responses based on each interview's unique flow. Member checking was used actively during interviews through a process of reflecting and probing to verify that interviewers correctly understood what each participant said and meant [33]. Interviews were audio-recorded with participant consent and transcribed verbatim for thematic analysis. On average, the core of each interview lasted 15.6 (SD 7.4) minutes (ranging from 4.4 to 36.4 min). Participants were assigned numerical pseudonyms.

After completing the web-based expressive writing program, many participants expressed their desire to continue with the group writing program. Consequently, we organized a web-based evening gathering 3 months after the project's completion. More than 75% (n=19) of participants who completed the writing program attended the 2-hour web-based gathering to share their experiences.

Data Analysis

A phenomenological approach using thematic analysis was used to identify emerging themes from the interview transcripts related to participants' experiences with the expressive writing program (ie, the Writing for Healing workshop) and the impact of these experiences on their emotions, moods, thoughts, and daily lives.

Two unique steps in the phenomenological approach, bracketing and intuiting, were implemented before data analysis [34]. Bracketing is a powerful strategy to minimize investigator biases [35] in which interviewers and coders bracket out their presuppositions and biases related to the study topic and content and approach the data with an open mind. To ensure data were approached neutrally, we selected interviewers and coders for this study who had no prior clinical knowledge of people with

XSL•FO
RenderX

SCI or their experiences with expressive writing and had not conducted any literature searches on this topic before data collection (ie, interviews) and coding. By fully immersing themselves in participants' perspectives, coders drew on their insights and understanding to grasp the essence of participants' lived experience with the program (ie, intuiting). This attempt by coders to access the "meaning" behind participants' words and descriptions of their program experience allowed a deeper interpretation of the data. The coders spent a full semester analyzing the data.

After verbatim responses from all participants were imported into a Word (Microsoft Corp) document, the qualitative data analysis was conducted using the following steps [36]. To gain an overall impression of participants' responses and to formulate tentative ideas, the same 4 occupational therapy graduate students who interviewed participants read all transcripts independently and multiple times to fully immerse themselves in the data and participants' perspectives. To generate initial codes (ie, open coding), they closely read each transcript, focusing on the content, context, and language used by participants. The coders ensured codes aligned with the original texts and preserved the participants' meaning throughout the process. As analysis progressed, codes expressing related concepts were grouped into categories (ie, axial coding) that captured participants' experiences. The coders then clustered categories into themes and subthemes by checking whether preliminary themes and subthemes were appropriate relative to the coded extracts and modifying them as needed (an iterative process) to ensure themes reflected the original data accurately. They integrated participants' background information and writing coaches' comments and reflections into the analysis as appropriate. Finally, coders defined and named the themes and subthemes. In addition, we presented our preliminary findings to participants who attended the web-based gathering held 3 months after the project's completion. During the gathering, participants verbally shared their experiences with the expressive writing program. Information collected from the gathering served as member checking to verify findings and interpretations (ie, themes and subthemes). The main themes and subthemes that emerged from the data are presented in the Results section of this paper along with examples of participant quotations to provide additional context.

To reduce bias in the analysis and enhance credibility, the coders reviewed each other's theme categorization, comparing and contrasting their findings. They met several times to check the plausibility of the data categorization and interpretation and discussed the meaning of themes and subthemes until they reached a consensus. To increase the rigor of the analytic process, the interpretations and derived meanings were constantly checked against the interview transcripts to verify that themes and subthemes reflected the participants' actual words. When disagreements occurred, the team reviewed participants' transcripts, and then discussed and resolved disagreements. To enhance the trustworthiness of transcript interpretations, an independent arbiter (HKY) with substantial experience in qualitative research and familiarity with the study served as an external auditor to verify the coding and emerged themes and subthemes.

Results

Overview

A total of 24 adults with SCI completed the 10-week expressive writing program and the exit interview; their characteristics are reported in Table 2. A total of 19 (79%) participants were White, and 17 (71%) participants were female. In total, 18 (75%) participants had sustained a traumatic SCI; of these, 10 (42%) had quadriplegia, 12 (50%) had paraplegia, and 2 (8%) had monoplegia. The mean age of the participants was 51 (SD 12) years (range: 34 - 76 y). The mean year of postinjury was 17 (SD 13) years (range: 1.5 - 52 y), and 6 (25%) participants had sustained their injury within 5 years of study enrollment. In total, 5 (21%) participants lived alone, and 10 (42%) participants received disability assistance.

Qualitative analysis of participant experiences with the program revealed three overarching themes with nine subthemes: supportive environment (overarching theme 1, with 3 subthemes); cathartic experience (overarching theme 2, with 2 subthemes); and acceptance of life (overarching theme 3, with 3 subthemes).

XSL•FO
RenderX

Table . Characteristics of participants with spinal cord injury who completed the coach-guided videoconferencing expressive writing program (n=24).

Name	Age (years)	Gender[a]	Race	Education[b]	Living situation[c]	Employment[d]	Diagnosis[e]	Injury level[f]	Postinjury duration (years)	Session type
P1	62	F	White	SC	WS	D	NS	P	4.4	Group
P2	60	F	White	PG	A	D	SCI	OLL	20.9	Individual
P3	44	F	White	SC	WS	HM	TM	OLL	3.1	Group
P4	73	M	White	PG	WS	D	SCI	Q	3.3	Individual
P5	34	F	White	PG	WS	FT	SCI	Q	11.8	Group
P6	50	M	Black	BA/BS	WS	D	SCI	P	6.5	Group
P7	59	M	White	V	WS	NW	SCI	P	15.8	Individual
P8	60	F	White	PG	A	D	SCI	Q	17.8	Group
P9	76	M	White	BA/BS	WS	D	SCI	P	33.2	Group
P10	43	F	White	PG	WS	FT	SCI	P	22.5	Individual
P11	49	F	White	PG	WS	PT	E	P	17.7	Individual
P12	53	M	White	BA/BS	WS	FT	SCI	Q	36	Individual
P13	45	F	White	BA/BS	WS	D	SCI	Q	29.3	Individual
P14	54	F	White	SC	A	PT	SCI	P	22.5	Group
P15	68	F	White	PG	WS	PT	TM	P	52	Group
P16	35	F	White	BA/BS	WS	D	SCI	Q	17.1	Group
P17	52	M	White	BA/BS	WS	FT	SCI	Q	34.6	Group
P18	39	F	Asian	BA/BS	WS	FT	SCI	Q	18.3	Group
P19	67	F	White	PG	A	R	TM	P	Missing	Group
P20	42	F	Black	SC	WS	D	SCI	P	16.3	Individual
P21	40	F	White	PG	A	PT	SCI	Q	6.4	Individual
P22	36	F	Mixed	BA/BS	WS	NW	SCI	P	1.4	Group
P23	49	F	White	BA/BS	WS	FT	TM	P	3.5	Group
P24	36	M	Black	<HS	WS	D	SCI	Q	3.3	Group

[a]Gender: F=female; M=male

[b]Education: PG=postgraduate; BA/BS =bachelor's degree; SC=some college; V=vocation; <HS=less than high school.

[c]Living situation: A=alone; WS=with someone.

[d]Employment: FT=fulltime; PT=parttime; NW=not working; R=retired; D=disability; HM=homemaker.

[e]Diagnosis: E=ependymoma; NS=neurosarcoidosis; SCI=spinalcord injury; TM=transverse myelitis.

[f]Injury level: P=paraplegia; Q=quadriplegia; OLL=one lower limb.

Overarching Theme 1: Supportive Environment

The Writing for Healing workshop created a supportive environment for participants by establishing an accepting environment and safe space (subtheme 1a), providing a supportive coach (subtheme 1b), and offering structured sessions (subtheme 1c).

Subtheme 1a: Establishing an Acceptance Environment and Safe Space

Participants reported that the sessions were a safe space for discussing uncomfortable topics and disclosing personal struggles associated with their SCI. Some sessions were held in a small group setting in which participants were unfamiliar with each other. This was not seen as a drawback, however, as indicated in this comment.

> *I actually like being with someone else, hearing someone else talk allows you to hear someone else's side. So it kind of like made it easier to be a little more open with your own issues or your own path when someone else is being open with theirs.* [P16]

Given the safe environment, participants felt the sessions allowed them to look at past life events that they typically brushed aside. In addition, the writing coaches played a crucial role in creating an accepting environment where participants felt comfortable sharing their emotions and experiences. One participant stated:

XSL•FO
RenderX

> *I sometimes came into the sessions, and I was really sad or down from the week I was having. I could write about that and share, and I always felt accepted when I didn't feel like opening up. The coach did a really good job of making it okay for when I felt like that in a session. I felt like all my emotions were safe here and I could write about anything I wanted, and it was never dismissed.* [P3]

Subtheme 1b: Providing a Supportive Coach

The coaches selected for the Writing to Healing workshop were highly praised by participants for their support and patience. For example, a participant noted that:

> *Elizabeth, who was the leader of my group [ie, writing coach] did an exceptional job of explaining the rules, guidelines, and making us feel comfortable in an environment where you are willing to express things [emotions or injury specifics] that you may not be able to or willing to talk about in every type of environment.* [P17]

Another participant noted:

> *I just think she's [Elizabeth, the writing coach] gifted in what I don't know what how to name her skill set exactly but she just the way she interacted with us, just gave us liberty to say, or share anything, and put it out on the table. She didn't hinder us or make us feel like that's not important, or that's not even what I was looking for. She just gave us a large space in which to pour things out that just bubbled up to the surface, depending on what writing prompt she would give us.* [P15]

Subtheme 1c: Offering Structured Sessions

Participants also commented on the workshop structure and session duration. Each web-based session was led by a coach and lasted 1 hour. This format provided participants with clear guidance and motivation. The discussions led by the writing coaches allowed the participants to explore and discuss various emotions related to their SCI, such as feelings of depression, guilt, forgiveness, acceptance, pain, and success. Many participants said they approved of the duration of each session and the overall program. For example, one noted:

> *I found that structure beneficial and it gave structure to what would have been an unstructured day, and I appreciated that.* [P8]

The participants also noted that "there was always a set plan for the direction of the sessions," (P10) and that "the structure was motivating" (P18). For example, a participant stated, "I appreciated the length [1 h] of the sessions because it didn't feel like a burden" (P21). Another stated, "I enjoyed the length of time in each session, I thought it was beneficial to my week" (P4).

Regarding the appropriate timing for people with SCI to engage in a Writing for Healing workshop, one participant suggested:

> *This [the workshop] should be a standard of care for anyone who suffered from a major injury or accident. Maybe not immediately because you are still processing [your injury] but definitely within a year of your accident.* [P11]

Many participants also expressed their desire to continue writing as a means of healing after the workshop concluded.

Overarching Theme 2: Cathartic Experience

The Writing for Healing workshop provided a cathartic outlet for participants by allowing them to experience therapeutic and insightful emotions (subtheme 2a) and gain a positive perspective and attitude on life (subtheme 2b).

Subtheme 2a: Experiencing Therapeutic and Insightful Emotions

The workshop was most participants' first experience in a coach-led expressive writing program. Participants described the sessions as therapeutic. For example, one participant stated:

> *I saw it as like a form of therapy, to be able to write your thoughts down, good, bad [and] mixed in or whatever, you just get it out and it's sort of the same way, like when I'd be really stressed, and I'd write lists. And if I could put my list on paper quite often. I don't come back to them, but it's just the process of getting it out. That's freeing.* [P2]

Another described the writing process as "just therapeutic to be able to talk to someone and to write down your feelings and learn how to express thoughts in a different way" (P6). One of the writing coaches, EVK, further validated this by noting that, for this participant (P6), "[the sessions] pulled emotions out of him that he did not even know he had."

Participants also discussed how the workshop helped them clarify some emotions about their injury or changed the way they viewed their emotions. For example, one participant stated:

> *It [The Writing for Healing workshop] helped me to address feeling that I probably suppressed about having a spinal cord injury, along with other emotions and feelings that I was having. And by week 8 or 9, I felt like a burden was lifted off of me.* [P20]

Subtheme 2b: Gained a Positive Perspective and Attitude on Life

The Writing for Healing workshop provided participants with opportunities to discuss past adversities related to their SCI in one-on-one web-based sessions with a writing coach or a small group setting. A participant stated that the sessions allowed her "… to reflect on the past in a way that was kind of guided in a naturally positive and accepting way. But that also gave me a chance to kind of reflect on my growth over the years" (P13). Participants appreciated the supportive environment the sessions created, which encouraged open discussion and sharing of strategies for coping with challenges. They found the workshop helped them cultivate a positive attitude and served as a tool for processing emotions, as illustrated in this comment.

> *I sort of felt like, there was an art of the workshop. In the beginning, I feel like there were some tougher really emotional things. And by the end, they weren't talking about those emotions as much, but it felt like*

XSL•FO
RenderX

> *we moved to a space of greater positivity. Like more forward-thinking and so that was something that I felt like that was the art of the workshop working through some emotions and then getting to a place of sort of thinking forward in life. I thought that was a really powerful way of running it and I really appreciated that. So, the prompts at the end, I felt like they were more imaginative towards the future as opposed to processing tough things in the past.* [P21]

Participants discussed how the expressive writing sessions allowed them to reflect on life and provided unexpected benefits. Through the reflective writing prompts, participants talked about how the workshop helped them gain a new perspective on their injuries and the challenges in their lives. Some participants viewed the experience as an opportunity to gain clarity and reflect on the impact of their injuries. As one participant stated, the sessions forced them "…to look at the ins and outs of not just daily living, but how it's [my transverse myelitis] impacted me how it has impacted my daily living, but also how it's impacted me looking at things. How it's impacted me and my relationships, how it's impacted me in my relationship with myself and my viewing of myself" (P23).

Overarching Theme 3: Acceptance of Life

Overview

The Writing for Healing sessions allowed participants to approach and accept different feelings within their lives, specifically, they felt less lonely and gained a greater sense of community (subtheme 3a), felt less stress (subtheme 3b), and were more forgiving of themselves and others (subtheme 3c), which enhanced their sense of self-advocacy and empowerment.

Subtheme 3a: Reduced Loneliness and a Greater Sense of Community

The cumulative effects of the Writing for Healing sessions were particularly beneficial for participants. They reported that, toward the end of the program, they felt less lonely and isolated as they connected with others who shared similar experiences. For example, one participant stated, "At one point I felt lonely, but now I don't feel so lonely" (P24), and "I think I still have a lot to process, but I don't feel as alone" (P22).

Participants also placed a high value on the sense of community fostered by this relatively brief 10-week workshop. For example, one stated:

> *I think probably the most beneficial aspect of it to me was that it was with two other ladies who had the same diagnosis as I do so they have gone through some of the same struggles and hearing how it's impacted them and how they've dealt with it and knowing you're not alone and being able to support each other when sometimes those conversations and our writing got very emotional. And so having the other ladies there who have gone through similar stuff, I felt like that was a big important part of this healing process. In addition to the writing, but the fact that our writing got feedback in a positive supportive way.* [P23]

Another noted:

> *There were times when the other person (participant), would write something and then we would just start talking, and it wasn't just Salaam [a writing coach] reflecting, but also, I would reflect and the other participant would reflect, and sometimes those conversations were just very meaningful for both of us.* [P3]

This exchange with P3 was followed by the interviewer, the research coordinator, asking, "That's great. And so, is it correct if, tell me if I'm wrong, but is it correct in saying that you felt like there was a sense of community?" P3 responded, "Absolutely." The sense of community and shared experience also contributed to participants' sense of empowerment. The participant continued, "I view life as more capable now. I am more open to sharing my truth [transverse myelitis] and feel more free to express myself" [P3].

Subtheme 3b: Feeling Less Stressed

Throughout the workshop, participants viewed the Writing for Healing sessions as opportunities to talk freely and enjoy a sense of community. Many participants, depending on how recently their SCI had occurred, were still dealing with daily physical and emotional stress. A participant noted, "…it [the sessions] allowed me to escape from life's adversity" (P11). Participants also noted that they experienced less stress and an increased ability to advocate for themselves after completing the program. One participant said:

> *People would ask me things which would increase my stress about my injury, but now when I am asked about myself, I am able to answer them, and I feel more comfortable with being able to be vocal about certain things that I wouldn't have before.* [P16]

Subtheme 3c: Feeling Forgiving of Themselves and Others

When participating in the Writing for Healing workshop, some participants felt that they could forgive people who had hurt them in the past or did not understand their diagnosis and the complications that arose from it. A participant noted:

> *I think I've identified some, some things and it surprised me, especially with the forgiveness prompt. I was like, oh, wow, you know, I have a lot of people that I am really angry about. It was it was kind of a revelation like, well, man, I'm holding a lot of negativities. And I think, you know, I feel lighter, I guess after writing for that prompt and doing a second one around the same kind of topics.* [P18]

Participants also expressed that they were better able to open their minds to new thoughts and forgiveness of others and themselves at the end versus the start of the workshop. For example, one participant noted:

> *Well, sometimes my mind kind of just looks at the past and picks very specific things to focus on to tell a narrative and it's not that great sometimes. You know, especially when it revolves around trauma, and I think that this helped me look at it with kind of a bird's eye*

XSL•FO
RenderX

view or allowed me to give myself some grace and forgiveness and not be so hard on myself. [P13]

Discussion

Principal Findings

The sudden onset of lifelong disability caused by SCI often results in profound physical, mental, and social challenges. This study demonstrates that the coach-guided videoconferencing expressive writing program is a promising approach that can help adults with SCI gain a positive perspective and attitude toward life and experience therapeutic and insightful emotions, facilitating their journey toward grief resolution. Overall, participants experienced profound, positive changes resulting from participating in the expressive writing program, highlighting its transformative impact. A high percentage of participants completed the 10-week writing program and attended the postproject web-based gathering, which indicates they felt that the coach-guided expressive writing program added value to their lives.

The coach-guided expressive writing program provided a supportive environment in which participants with similar experiences felt less lonely and isolated. Participants in small group settings encouraged and supported one another, highlighting the importance of acceptance and establishing a safe space to discuss personal challenges related to SCI and share coping strategies. The two writing coaches strictly adhered to the expressive writing program protocol, which included weekly themes or activities and writing prompts regardless of whether sessions with participants were conducted individually or in small groups. However, participants in small groups had the opportunity to share postwriting reflections with others, which reduced feelings of loneliness and increased their sense of community (subtheme 3a). This unintended positive consequence could not have been achieved had we conducted all sessions individually.

The emphasis on creating a supportive environment and the role of the coaches in guiding participants through the writing process resonates with the importance of interpersonal dynamics in therapeutic interventions [24]. This study's supportive environment emerged as a pivotal aspect of the intervention, as engaging with peers facilitated a shift toward a more positive outlook on life. The participants reported that, by confronting their past, making sense of their losses, and embracing their injury-related limitations, they experienced decreased distress and an enhanced sense of meaning and purpose.

Findings from this study also echo the broader discussion in disability studies on how the arts play an integral role in individuals' cultural and social lives, shifting from a solely clinical perspective. The participants' descriptions of increased confidence, empowerment, and sense of agency align with findings that artistic expression serves as a platform for self-advocacy and representation [37]. Additionally, the expressive writing program fostered a shared space where participants felt validated and connected to others with similar experiences, reinforcing the significance of community-oriented artistic engagements [31,38].

Central to the program's success was the creation of a safe space by the coaches, who demonstrated patience and attentive listening. Their guidance enabled participants to overcome their inhibitions gradually and engage in the writing process wholeheartedly. The participant's acceptance of the coaches was crucial to the program's effectiveness. Participants in previous coach-led expressive writing programs have also emphasized the value and impact of writing coaches in helping them explore their emotional states and improve their ability to express themselves and their emotions regarding important personal health issues [39-41]. Moreover, many participants expressed a desire to continue expressive writing beyond the program's duration, underscoring its value as a long-term resource for healing and self-discovery.

Qualitative analysis of participants' exit interview responses revealed that the expressive writing sessions were a cathartic experience, enabling them to articulate and release their thoughts and emotions and engage in introspection and self-expression. Furthermore, participants noted a shift toward a positive perspective on life, indicating the program's potential to foster resilience and optimism in the face of adversity. The findings of this study align with existing literature on the benefits of expressive writing interventions for individuals facing adversity (ie, trauma and loss), including physical illness and mental health conditions [17,19-22,40]. The expressive writing program may have encouraged participants to rethink their behaviors, leading to changes in cognitive processing and coping mechanisms, as demonstrated in other studies on expressive writing [24,42].

The theme of acceptance of life further elucidated the program's impact on participants' emotional well-being. By fostering a sense of community and connection, the program helped alleviate the feelings of loneliness and isolation commonly experienced by individuals with SCI. Additionally, participants reported reduced stress levels and an increased capacity for forgiveness toward themselves and others. These findings highlight the transformative potential of expressive writing in promoting emotional resilience and self-compassion in individuals with SCI. Notably, participants reported enhanced self-advocacy and empowerment as a result of completing the program. Many expressed newfound confidence in discussing their diagnoses and sharing their stories, demonstrating the program's broader impact on their sense of agency and self-expression.

As this was a pilot study, we did not exclude people with SCI based on the duration of their injury. Although psychological support may be more beneficial to those with a recent injury compared to those with an injury of longer duration [43,44], all our participants had sustained their injuries more than a year before entering the study. Most individuals with SCI in the subacute stage would likely not have been mentally prepared to participate in this program, which was not part of clinical rehabilitation and was conducted in a home setting. This pilot program demonstrated a high degree of success for individuals with SCI and can offer them a complementary or alternative avenue of psychological intervention. Therefore, our next step is to refine the program for scalability to support individuals

XSL•FO
RenderX

with other conditions that involve a sudden loss of physical abilities, such as stroke, brain injury, and amputation.

Although participants reported an overall positive experience with the program, several initially expressed concerns about its 10-week duration. At the outset, some felt that 10 weeks was too long, however, by the program's end, many felt it was not long enough. A proposed strategy to encourage enrollment of people with SCI is to advertise the program as a 5-week program while offering the option of a 10-week extension. In addition, scheduling group meetings with members who resided in different time zones was challenging. As a result, some participants met with the coach individually. Finally, some participants had unstable internet connections that led them to reschedule sessions. To accommodate group members who missed sessions due to scheduling conflicts, the coach provided individual make-up sessions within the same week to ensure they were prepared to rejoin the following group session.

Limitations

This study's sample size was limited and predominantly included highly educated, non-Hispanic White women, which may affect the transferability of the findings. Further research is needed to explore the implementation of the expressive writing program across diverse demographics and settings and advance the understanding of effective interventions for individuals with SCI. Since the interviews were conducted immediately after the program ended, this study's findings reflect only the immediate perceived impact of the program. Future studies may consider conducting interviews twice, immediately after program completion and several months later, to explore both immediate and sustained benefits of the expressive writing program perceived by participants.

As we did not have permission to video-record writing sessions or access participants' writing, we could not use transcripts and field notes to examine group dynamics. Instead, we relied on the writing coach's reflections on the nature of interactions (between coaches and participants and among participants) to understand the group communication process and its contributions to participants' insights and feedback about their experiences. The purpose of this investigation was to understand participants' perspectives on the expressive writing program and its impact, rather than their emotional reaction to their SCI. Therefore, analyzing participants' actual writing (initial and reflective) completed during the program would not add significant value to the research findings. Additionally, if participants had expected that their actual writing, which included personal health information, would be shared with and read by unknown individuals, they would have been less likely to express their emotions freely, defeating the program's purpose.

Conclusions

In conclusion, the coach-guided expressive writing program has emerged as a valuable tool and an achievable avenue for individuals with SCI to process and alleviate grief and stress across various dimensions. The results from this study suggest that the program promoted psychosocial adjustment and well-being in individuals living with SCI, which could ultimately strengthen coping abilities and resilience in the face of adversity. This study demonstrates the feasibility and potential efficacy of a coach-guided expressive writing program specifically tailored for individuals with SCI, addressing a critical gap in the literature on rehabilitative interventions for this population. The study provides an important lens for understanding the role of nonclinical programs and services in the lives of people with SCI. While the study's limitations should be acknowledged, the findings hold promise for further exploration to extend these results to diverse populations with limb paralysis.

Acknowledgments

We thank all the study participants for their precious time participating in this study. This research was supported by the Christopher & Dana Reeve Foundation (2022631, 2027598), and Craig H Neilsen Foundation (967751).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author upon request.

Conflicts of Interest

None declared.

References

1. deRoon-Cassini TA, de St. Aubin E, Valvano A, Hastings J, Horn P. Psychological well-being after spinal cord injury: perception of loss and meaning making. Rehabil Psychol 2009;54(3):306-314. [doi: 10.1037/a0016545]
2. Mc Ateer MF. Some aspects of grief in physiotherapy. Physiotherapy 1989 Jan;75(1):55-58. [doi: 10.1016/S0031-9406(10)62950-1]
3. Middleton J, Craig A. Psychological challenges in treating persons with spinal cord injury. In: Psychological Dynamics Associated with Spinal Cord Injury Rehabilitation: New Directions and Best Evidence: Nova Science Publishing; 2008:3-53.
4. Burkhart L, Kale IO, LaVela SL. Grief and loss among veterans with spinal cord injury: a qualitative study. Rehabil Nurs 2021;46(5):270-278. [doi: 10.1097/RNJ.0000000000000303] [Medline: 33264175]

XSL•FO
RenderX

5. Niemeier JP, Burnett DM. No such thing as “uncomplicated bereavement” for patients in rehabilitation. Disabil Rehabil 2001 Oct 15;23(15):645-653. [doi: 10.1080/09638280110049324] [Medline: 11720114]
6. Angel S, Kirkevold M, Pedersen BD. Getting on with life following a spinal cord injury: regaining meaning through six phases. Int J Qual Stud Health Well-being 2009 Jan;4(1):39-50. [doi: 10.1080/17482620802393492]
7. Jackson SA, Hough S. Adjustment to the process of grief following a spinal cord injury/dysfunction. SCI Nurs 2005;22(1):15-19. [Medline: 15794430]
8. Chuang CH, Yang YO, Kuo LT. Finding a way to cope: a qualitative study of the experiences of persons with spinal cord injury. J Neurosci Nurs 2015 Dec;47(6):313-319. [doi: 10.1097/JNN.0000000000000169] [Medline: 26528948]
9. Pollard C, Kennedy P. A longitudinal analysis of emotional impact, coping strategies and post-traumatic psychological growth following spinal cord injury: a 10-year review. Br J Health Psychol 2007 Sep;12(Pt 3):347-362. [doi: 10.1348/135910707X197046] [Medline: 17640451]
10. O’Donnell ML, Bryant RA, Creamer M, Carty J. Mental health following traumatic injury: toward a health system model of early psychological intervention. Clin Psychol Rev 2008 Mar;28(3):387-406. [doi: 10.1016/j.cpr.2007.07.008] [Medline: 17707563]
11. Parrish MM, O’Malley K, Adams RI, Adams SR, Coleman EA. Implementation of the care transitions intervention: sustainability and lessons learned. Prof Case Manag 2009;14(6):282-293. [doi: 10.1097/NCM.0b013e3181c3d380] [Medline: 19935345]
12. Mankad A, Gordon S. Psycholinguistic changes in athletes’ grief response to injury after written emotional disclosure. J Sport Rehabil 2010 Aug;19(3):328-342. [doi: 10.1123/jsr.19.3.328] [Medline: 20811081]
13. Baker FA, Rickard N, Tamplin J, Roddy C. Flow and meaningfulness as mechanisms of change in self-concept and well-being following a songwriting intervention for people in the early phase of neurorehabilitation. Front Hum Neurosci 2015;9(May):299. [doi: 10.3389/fnhum.2015.00299] [Medline: 26082702]
14. Glover E, Rice G, Phillips V, Williamson C. Shadow into Light: a Bristol-based arts project for bereavement. Bereave Care 2016 Jan 2;35(1):7-12. [doi: 10.1080/02682621.2016.1160613]
15. Kállay É. Physical and psychological benefits of written emotional expression. Eur Psychol 2015 Oct;20(4):242-251. [doi: 10.1027/1016-9040/a000231]
16. Yuen HK, Kamp EV, Green S, et al. Effects of a coach-guided video-conferencing expressive writing program on facilitating grief resolution in adults with SCI. J Spinal Cord Med 2024 Nov;47(6):1016-1025. [doi: 10.1080/10790268.2023.2253390] [Medline: 37682297]
17. Garisch D, Giddy J, Griffin G, Reid S. Can “life writing” be therapeutic in response to trauma? an exploratory research project in Medical Humanities in South Africa. Med Humanit 2024 Feb 22;50(1):162-169. [doi: 10.1136/medhum-2021-012368] [Medline: 37802648]
18. Pennebaker JW, Beall SK. Confronting a traumatic event: toward an understanding of inhibition and disease. J Abnorm Psychol 1986 Aug;95(3):274-281. [doi: 10.1037//0021-843x.95.3.274] [Medline: 3745650]
19. Choi E, Shin LJ, Chen L, Lu Q. Lived experiences of young adult Chinese American breast cancer survivors: a qualitative analysis of their strengths and challenges using expressive writing. Eur J Oncol Nurs 2023 Feb;62:102253. [doi: 10.1016/j.ejon.2022.102253] [Medline: 36628853]
20. Guo L. The delayed, durable effect of expressive writing on depression, anxiety and stress: a meta-analytic review of studies with long-term follow-ups. Br J Clin Psychol 2023 Mar;62(1):272-297. [doi: 10.1111/bjc.12408] [Medline: 36536513]
21. Lai J, Song H, Wang Y, et al. Efficacy of expressive writing versus positive writing in different populations: systematic review and meta-analysis. Nurs Open 2023 Sep;10(9):5961-5974. [doi: 10.1002/nop2.1897] [Medline: 37434395]
22. Abu-Odah H, Su JJ, Wang M, Sheffield D, Molassiotis A. Systematic review and meta-analysis of the effectiveness of expressive writing disclosure on cancer and palliative care patients’ health-related outcomes. Support Care Cancer 2023 Dec 29;32(1):70. [doi: 10.1007/s00520-023-08255-8] [Medline: 38157056]
23. Niles AN, Haltom KEB, Lieberman MD, Hur C, Stanton AL. Writing content predicts benefit from written expressive disclosure: Evidence for repeated exposure and self-affirmation. Cogn Emot 2016;30(2):258-274. [doi: 10.1080/02699931.2014.995598] [Medline: 25650018]
24. Leung YW, Maslej MM, Ho C, et al. Cocreating meaning through expressive writing and reading for cancer caregivers. J Palliat Care 2023 Jul;38(3):307-315. [doi: 10.1177/0825859719871538] [Medline: 31476955]
25. Smyth JM. Written emotional expression: effect sizes, outcome types, and moderating variables. J Consult Clin Psychol 1998 Feb;66(1):174-184. [doi: 10.1037//0022-006x.66.1.174] [Medline: 9489272]
26. Nyssen OP, Taylor SJ, Wong G, et al. Does therapeutic writing help people with long-term conditions? Systematic review, realist synthesis and economic considerations. Health Technol Assess 2016;20(27):1-368. [doi: 10.3310/hta20270]
27. Clifton S. Grieving my broken body: an autoethnographic account of spinal cord injury as an experience of grief. Disabil Rehabil 2014;36(21):1823-1829. [doi: 10.3109/09638288.2013.872202] [Medline: 24382269]
28. Ignagni E, Chandler E, Collins K. Activating the arts in death: what are the cultural implications for MAiD? Int Health Trends & Persp 2021;1(3):336-344. [doi: 10.32920/ihtp.v1i3.1435]
29. Kuppers P. Studying Disability Arts and Culture: An Introduction: Palgrave Macmillan; 2014.
30. Linton S. Claiming Disability: Knowledge and Identity: New York University Press; 1998.

XSL•FO
RenderX

31. Howells V, Zelnik T. Making art: a qualitative study of personal and group transformation in a community arts studio. Psychiatr Rehabil J 2009;32(3):215-222. [doi: 10.2975/32.3.2009.215.222] [Medline: 19136354]
32. Sandelowski M. Whatever happened to qualitative description? Res Nurs Health 2000 Aug;23(4):334-340. [doi: 10.1002/1098-240x(200008)23:4<334::aid-nur9>3.0.co;2-g] [Medline: 10940958]
33. Creswell JW, Plano Clark VL. Designing and Conducting Mixed Methods Research: Sage Publications; 2011.
34. Greening N. Phenomenological research methodology. Sci Res J 2019;VII(V):88-92. [doi: 10.31364/SCIRJ/v7.i5.2019.P0519656]
35. Fischer CT. Bracketing in qualitative research: conceptual and practical matters. Psychother Res 2009 Jul;19(4-5):583-590. [doi: 10.1080/10503300902798375] [Medline: 20183407]
36. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
37. Collins A, Rentschler R, Williams K, Azmat F. Exploring barriers to social inclusion for disabled people: perspectives from the performing arts. J Manage Organ 2022 Mar;28(2):308-328. [doi: 10.1017/jmo.2021.48]
38. Levy S, Young H. Arts, disability and crip theory: temporal re-imagining in social care for people with profound and multiple learning disabilities. Scand J Disabil Res 2020 Mar 10;22(1):68-79. [doi: 10.16993/sjdr.620]
39. Vergo MT, Klassen-Landis M, Li Z, Jiang M, Arnold RM. Getting creative: pilot study of a coached writing intervention in patients with advanced cancer at a rural academic medical center. J Palliat Med 2022 Mar;25(3):472-478. [doi: 10.1089/jpm.2021.0335] [Medline: 34941445]
40. Corner GW, Jhanwar SM, Kelman J, Pessin H, Stein E, Breitbart W. Visible ink: a flexible and individually tailored writing intervention for cancer patients. Palliat Support Care 2015 Apr;13(2):171-178. [doi: 10.1017/S1478951513000710] [Medline: 24138800]
41. Kartikaningsih N, Lawson K, Mayhan M, et al. The impact of an expressive writing and storytelling program on ex-offenders: a pilot study. Int J Offender Ther Comp Criminol 2023 Jul 19;2023:306624X231188228. [doi: 10.1177/0306624X231188228] [Medline: 37464771]
42. Creswell JD, Lam S, Stanton AL, Taylor SE, Bower JE, Sherman DK. Does self-affirmation, cognitive processing, or discovery of meaning explain cancer-related health benefits of expressive writing? Pers Soc Psychol Bull 2007 Feb;33(2):238-250. [doi: 10.1177/0146167206294412] [Medline: 17259584]
43. Chan RCK, Lee PWH, Lieh-Mak F. The pattern of coping in persons with spinal cord injuries. Disabil Rehabil 2000 Jan;22(11):501-507. [doi: 10.1080/096382800413998]
44. Kifley A, Geraghty TJ, Arora M, et al. Complex lived experiences and hidden disability after spinal cord injury: a latent profile analysis of the Australian arm of the International Spinal Cord Injury (Aus-InSCI) Community Survey. Disabil Rehabil 2024 Oct;46(20):4675-4686. [doi: 10.1080/09638288.2023.2283101] [Medline: 38018422]

Abbreviations

SCI: spinal cord injury

Edited by A Mavragani; submitted 10.01.25; peer-reviewed by A Battalova, S Thongchoomsin; revised version received 01.05.25; accepted 11.05.25; published 30.06.25.

Please cite as:
Xie SM, McKenna M, Veach K, Williams S, Jones MG, Vander Kamp E, Green S, Edwards L, Kirklin K, Jones BA, Yuen HK
Assessing the Psychosocial Impact of Expressive Writing on Adults With Spinal Cord Injury: Qualitative Study
JMIR Form Res 2025;9:e71162
URL: https://formative.jmir.org/2025/1/e71162
doi:10.2196/71162

© Shelly M Xie, Molly McKenna, Kendall Veach, Sydney Williams, Mary Grace Jones, Elizabeth Vander Kamp, Salaam Green, Lauren Edwards, Kimberly Kirklin, Benjamin A Jones, Hon K Yuen. Originally published in JMIR Formative Research (https://formative.jmir.org), 30.6.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Evaluating the Effectiveness of Smart Glasses in Reducing Patient Care Time in Emergency Departments: Cohort Study From the Hangzhou Asian Games

Xinwei Jiang[1*], BSc; Bangbo Xia[2*], MD; Mohammad Mostafa Ansari[3], MD; Huiquan Jiang[4], BSc; Jianjiang Qi[5], MSc; Zhongheng Zhang[2], MD, PHD; Sheng Dai[6], MD, PhD; Pingping Zheng[7], MSc; Yang He[2], MSc; Ning Liu[2], MSc; Pengpeng Chen[2], MSc; Ronghua Luo[8], MD; Xuchang Qin[2], MSc; Yansong Miao[2], MSc; Junru Dai[2], MSc; Xiaoyu Zhou[2], MSc; Changliang Wang[9], MD; Hui Chen[10], MD; Wenbin Xu[11], MD; Tao Wu[12], MSc; Qiang Shi[6], BSc; Zhonghua Chen[2], MSc; Liping Zhou[1], BSc; Hao Zhang[2], MSc; Yun Xie[8], MSc; Quan Zhang[8], MD; Bifa Zhou[8], MD; Xiaohong Pan[7], MD; Zixi Chen[7], MD; Libo Zhen[8], MD; Yaqing Sun[8], MD; Zelin Lu[3], BSc; Yihao Loh[3], BSc, MBBS; Shameera Sayer[3], BSc, MBBS; Jennifer Mochtar[3], BSc, MBBS; Pannika Wongpraewit[3], BSc, MBBS; Yifan Wang[13], MD, PhD; Yucai Hong[2], MD, PHD

[1]Nursing Department, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou, China
[10]Department of Cardiology, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou, China
[11]Department of Orthopaedics, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou, China
[12]Department of Rehabilitation Medicine, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou, China
[13]Department of General Surgery, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou, China
[2]Department of Emergency Medicine, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou, China
[3]School of Medicine, Zhejiang University, Road Qingchundong No. 3, Shangcheng District, Hangzhou, China
[4]Medical Department, Organising Committe, Hangzhou Asian Games, Hangzhou, China
[5]Hangzhou Seventh People's Hospital, Hangzhou, China
[6]Department of Medical Affairs, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou, China
[7]Department of Emergency Medicine, Hangzhou First People's Hospital, Hangzhou, China
[8]Hangzhou Red Cross Hospital, Hangzhou, China
[9]Department of Intensive Care Medicine, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou, China
[*]these authors contributed equally

Corresponding Author:
Mohammad Mostafa Ansari, MD
School of Medicine, Zhejiang University, Road Qingchundong No. 3, Shangcheng District, Hangzhou, China

Abstract

Background: Challenges in emergency medicine include overcrowding, insufficient emergency care resources, and extended emergency department (ED) waiting times. These issues contribute to delays in treatment and unfavorable outcomes. This situation worsens in events with large crowds and particularly worsened during the COVID-19 pandemic. The integration of augmented reality (AR) smart glasses could potentially enhance patient care in the ED.

Objective: This study aims to assess the effectiveness of AR smart glasses in reducing patient care time in the ED during the 19th Asian Games and the Fourth Asian Para Games Hangzhou 2022 (HAG2022). The study specifically compares the prepreparation time (PPT) and consult response time (CRT) in patients receiving teleconsultations via AR smart glasses versus those receiving standard care without AR.

Methods: This retrospective study was conducted between September 13, 2023, and October 28, 2023, during HAG2022. The data were gathered from AR smart glasses using 5G technology at the HAG2022 village and electronic health records at Sir Run Run Shaw Hospital, China. The study included 2 groups: the teleconsultation by augmented reality telemedicine system (ARTS) group and the non-ARTS group. The main data assessed were PPT and CRT in ED.

Results: During the research period, 80 patients were divided into 2 cohorts: the ARTS cohort (n=10) and the non-ARTS cohort (n=70). Gender and age demographics showed no significant differences between the cohorts. The ARTS cohort had a significantly lower average PPT of 23 minutes compared to the non-ARTS cohort's 40.3 minutes (P<.001). In addition, CRT in the ARTS

XSL•FO
RenderX

cohort was significantly lower at 15.6 minutes compared to the non-ARTS cohort's 164.8 minutes (*P*=.03). The outcomes suggest that smart glasses are effective in decreasing PPT and CRT.

Conclusions: AR smart glasses have the potential to enhance patient admission efficiency and reduce care time in EDs. However, despite these benefits, further research is needed to confirm their effectiveness, and additional studies are essential to identify the challenges and barriers to their successful implementation in emergency medicine.

Trial Registration: ClinicalTrials.gov NCT06623825; https://clinicaltrials.gov/study/NCT06623825

(JMIR Form Res 2025;9:e65617) doi:10.2196/65617

KEYWORDS

telemedicine; teleconsultation systems; emergency medicine; emergency care; patient care; multidisciplinary treatment; treatment; efficiency; smart glasses; augmented reality; wearable technology; wearable devices; EHR; electronic health record; retrospective study; effectiveness; Chinese population; Asia; Asian Games

Introduction

Emergency departments (EDs) play a pivotal role in providing immediate care for patients with acute illnesses or injuries [1]. EDs face several common problems that hinder their efficiency, including overcrowding and resource limitations. Overcrowding results in prolonged waiting times, length of stay, diminished quality of care, and increased mortality and morbidity rates [2-5]. Furthermore, requiring diagnostic tests and specialist consultations also contribute to extended waiting times and decision-making in EDs [6-8]. The time spent waiting for a specialist to arrive, conduct the consultation, and make a decision constitutes 33% of the length of stay for admitted patients and 54% for discharged patients in ED, highlighting a significant delay that contributes to longer waiting times [9]. Research has revealed that delays in ED care, such as those caused by overcrowding, can significantly impact patient outcomes, particularly for severe conditions like acute stroke [10,11]. A 2021 study reported an 8% increase in 30-day mortality for patients waiting more than 6 - 8 hours in the ED [12].

The 19th Asian Games and Fourth Asian Para Games Hangzhou 2022 (HAG2022), held from September 13 to October 28, 2023, in Hangzhou, Zhejiang province, China, marked the largest sporting event in Asia, with more than 12,000 athletes. While most events occurred in Hangzhou, several other cities in Zhejiang province also hosted competitions [13]. The global pandemic led to the postponement of HAG2022 to 2023 [14], prompting the organizing committee to develop a detailed medical emergency plan to ensure participant safety [15]. Previous experiences have shown that hosting a large-scale event during and after the COVID-19 pandemic posed considerable challenges for medical systems [16]. Such extensive public gatherings impose significant demands on the local emergency medical services (EMS) [17,18]. Therefore early identification of patients most likely requiring consultation, and limiting unnecessary consultations is important to improve EDs services [19] [20].

EMS served as the first line of response during HAG2022. The process began when patients or others contacted the national emergency hotline (120). The medical call center then dispatched ambulances based on the incident's location and urgency. After assessing the patient's condition at the scene, they were either transferred to the polyclinic at the HAG2022 village or, if necessary, to designated hospitals outside the village for advanced care. During this process, crucial information, including the patient's chief complaint, medical history, symptoms, and examination results, was collected and transmitted via 2-way radio to the hospital, allowing for immediate preparation and treatment upon the patient's arrival [21].

Traditional telecommunication methods, such as radio transmissions from the HAG2022 village or an ambulance to the hospital, are time-consuming and hinder continuous patient monitoring. This delay can lead to slower diagnoses and worsening clinical conditions [22].

Telemedicine, enhanced by augmented reality (AR) technology, addresses various challenges by providing visual aids and facilitating specialist communication for remote patient evaluations in outpatient, inpatient, and emergency settings [23]. Smart glasses enhance collaboration for first responders by allowing them to perform triage and treatment without obstruction while enabling physicians in tertiary care to prepare for coming patients [23-25].

This research aims to evaluate the impact of using smart glasses on EDs preparedness and patient care during HAG2022. Key parameters under review include waiting time and time to get specialist consultations in emergency department. To our current knowledge, this research stands as one of the first investigations delving into using smart glasses in ED during a sports event.

Methods

Study Design

This retrospective study was conducted at Sir Run Run Shaw Hospital (SRRSH), China. This study aims to compare the outcomes of 2 groups: the intervention group, using the AR smart glasses, and the control group, adhering to the standard conventional triage process without the AR tool, only using a 2-way radio connection. All calls for emergency care within the HAG2022 village were directed to a central medical call center and subsequently triaged for ambulance transport. Randomization into the teleconsultation by augmented reality telemedicine system (ARTS) and non-ARTS groups was determined based on the availability of AR smart glasses for teleconsultation at the time of each case. Cases were not

XSL•FO
RenderX

pre-randomized; instead, they were categorized retrospectively based on whether the ARTS was used.

Augmented Reality Tool

The AR tool used in this study was a smart glasses manufactured by Rokid AR Studio. It was Rokid glass 2 model released in 2020. The device operated on Android 9.0 and it connected to 5G connection with a bandwidth speed 30-mbps delivering a very low latency. The device supported real-time screen sharing and therefore the data were collected from smart glasses including visual and audio with real-time 2-way communications. Physicians at the hospital can access the information and video via a mobile app, enabling continuous monitoring of patients. A total of 10 AR smart glasses were deployed across various operational units during the study. These included 2 units in the Field of Play, 7 in hotels located outside the Asian Games Village, and one in medical polyclinic within the Asian Games Village. In the end, AR smart glasses were used for teleconsultation in only 1 hotel outside the Asian Games Village and 2 medical polyclinics within the village (see Figure 1). Among the medical polyclinics in the Asian Games Village, one was designated as a standard emergency room, equipped with advanced diagnostic tools such as CT imaging, bedside ultrasound, and basic blood testing capabilities. The remaining polyclinics were operated as general outpatient clinics, staffed with senior attending physicians specializing in general or emergency medicine, and supplied with commonly used medications. Each operational team was composed of trained paramedics, who were responsible for managing the AR smart glasses to enable telemedicine operations.

Figure 1. (A) On-site rescue; (B) management at primary medical stations; (C) comprehensive ambulance transportation; and (D) multidisciplinary teleconsultation at the tertiary hospital.

Participants

The study, conducted between September 13, 2023, and October 28, 2023, included the recruitment of all patients (ages 16 years and older) presenting with acute illnesses who were admitted to the ED of SRRSH from the HAG2022 village polyclinic or the field of injury. These patients included athletes, attendees, and other individuals present in the HAG2022 village during the event. SRRSH served as the primary tertiary referral center for the HAG2022 medical plan, and while other medical facilities were available, this study exclusively focused on cases managed at SRRSH.

Exclusion criteria included cases with incomplete data, technical difficulties with AR smart glasses, or loss of internet connection. All eligible patients presenting to the SRRSH ED during the study period were assigned to either the ARTS (intervention) or non-ARTS (control) groups based on the use or non-use of AR smart glasses for teleconsultation.

Outcomes

The main outcomes assessed in this study were the prepreparation time (PPT) and the consult response time (CRT). PPT refers to the time interval from the patient's arrival at the tertiary hospital to the completion of preparatory steps required for their subsequent management. These steps include the collection of necessary information, initial assessments, diagnostic tests, imaging, and stabilization measures before a decision is made on discharge, admission, or further treatment. For patients and athletes involved in the Asian Games, additional time may be required for registration compared to ordinary patients residing in China. Registration processes often took longer due to factors such as unfamiliarity with international names, particularly distinguishing between surnames and given names, and ensuring the accurate typing of full names, which was critical for medical insurance processing and certification. Furthermore, preparatory steps such as diagnostic tests and imaging often consumed additional time, depending on whether the necessary information was available before or after the patient's arrival at the tertiary hospital. Pre-arrival collection

XSL•FO
RenderX

of patient information had the potential to significantly reduce the waiting time by enabling faster ordering of diagnostic tests and imaging upon arrival. CRT is defined as the time from when the consulting physician is contacted to the moment they provide a response. CRT is critical for time-sensitive conditions such as acute coronary syndrome, stroke, trauma, and septic shock, where delays in consultation can negatively affect outcomes. Traditional consultation methods, like telephone or static video, often lack the ability to capture detailed clinical data. The use of AR glasses improved CRT by providing consultants with a first-person, real-time view of the patient's condition, allowing for the assessment of critical parameters such as pupil response and gait analysis. AR technology enabled up to six consultants to participate simultaneously, using standard smartphones, regardless of location. This innovation can reduce CRT, enhance consultation quality, and demonstrate scalability, particularly in resource-constrained settings or during emergencies. To enhance clarity, a workflow diagram was added to illustrate the emergency care process and the measurement points for PPT and CRT (see Figure 2). The decision to focus on PPT and CRT as primary metrics stems from their direct reflection of triage efficiency and consultant responsiveness, which are critical in emergency medicine. While broader metrics such as length of stay, time to admission, or time to discharge are relevant, they are influenced by multiple systemic factors, including resource availability, hospital workflows, and patient-specific variables. PPT and CRT, on the other hand, specifically measure the timeliness of preconsultation and consultation processes, which are particularly important in optimizing emergency care during large-scale events like the Hangzhou Asian Games.

XSL•FO
RenderX

Figure 2. Comparison of patient admission workflows using the augmented reality telemedicine system (ARTS) versus conventional methods (non-ARTS). HAG2022: 19th Asian Games and Fourth Asian Para Games Hangzhou 2022.

XSL•FO
RenderX

Data Collection

There were specially experienced paramedics and physicians from both HAG2022 polyclinic and SRRSH who volunteered to participate in the study. They underwent thorough training and familiarization sessions to ensure proficient use of the AR tool before its deployment. Specialized teams of trained personnel were used, thereby enabling regular ED staff to continue their routine responsibilities without incurring additional workload. A structured data collection form was developed to record demographic information, triage times, and CRT using electronic health records. The patients were divided into the teleconsultation by the ARTS and the non-ARTS groups. Patients in the ARTS group received teleconsultation with smart glasses before they arrived at the SRRSH. The AR functionalities, including real-time visualization of patient data and interactive guidance, empowered the health care professionals at SRRSH to not only document basic information, chief complaints, and vital signs but also obtain additional information per scenario. Furthermore, they managed to establish a connection with the AR tool at any point during the transportation. Patients in the non-ARTS group underwent standard conventional triage procedures, where the hospital was notified of patient admission via telephone communication. Information such as the chief complaint, vital signs, and general condition was reported to the hospital by paramedics through telephone communication (see Multimedia Appendix 1). Patient privacy, confidentiality, and data protection were strictly maintained throughout the study by assigning numbers to each patient instead of using identification information.

Statistical Analysis

The study's sample size was calculated using the formula: sample size = $(Z\alpha/2+ Z\beta)^2 * \sigma^2 / \Delta^2$, where $Z\alpha/2$ is the critical value corresponding to the significance level, $Z\beta$ is the critical value for the study's power, σ is the estimated SD of the population, and Δ is the minimum detectable effect size. SPSS Statistics software (version 20.0; IBM), was used for statistical analysis and data management in this study. Data analysis used the χ^2 test, Welch *t* test, binary logistic regression analysis, and multiple linear regression analysis, to compare the PPT and CRT between the intervention and control groups.

Ethical Considerations

Ethical approval for this study was obtained from the Research and Ethics Committee at Zhejiang University Affiliated Sir Run Run Shaw Hospital (registry 2023-850-01). Given the retrospective nature of the study and its use of anonymized clinical data, the study was granted an exemption from obtaining additional informed consent. The ethics committee granted a waiver of informed consent, as the study did not interfere with the regular triage procedures or prevent participants from receiving necessary medical treatment. All data were anonymized to ensure privacy and confidentiality, with identifying information replaced by numerical codes. The study adhered to relevant data protection regulations, and access to deidentified data was restricted to authorized personnel only. No compensation was provided to participants for their involvement, as the study was conducted as part of standard clinical care during the HAG2022. Furthermore, no identifiable images of individual participants were included in the study, ensuring the confidentiality of all participants in accordance with ethical standards. Patient privacy is ensured and protected through data deidentification techniques.

The research was further approved by the Ethics Committee of Sir Run Run Shaw Hospital, Zhejiang University School of Medicine (Srrsheca2023-0597) and registered at ClinicalTrials.gov (NCT06623825).

Results

Following the completion of the research period, there were 80 patients categorized into the ARTS cohort (n=10) and the non-ARTS cohort (n=70). Sex and age demographics for the ARTS and non-ARTS cohorts are shown in Table 1. Both cohorts were 70% (ARTS: n=7 and non-ARTS: n=49) male and 30% (ARTS: n=3 and non-ARTS: n=21) female, with no significant gender differences (*P*>.99). The average age for the ARTS cohort was 38.6 (SD 21.0) years and for non-ARTS was 33.3 (SD 12.7) years, with no significant age gap (*P*=.51). Results indicate no significant variations in gender or age between the 2 cohorts.

Table 2 shows a comparison of time intervals for ARTS and non-ARTS cohorts at a tertiary medical center. ARTS group's average PPT was 23 minutes, significantly lower than the non-ARTS group's 40.3 minutes (*P*<.001). In the ARTS cohort, the CRT was 15.6 minutes, significantly lower than the non-ARTS cohort's 164.8 minutes (*P*=.03). The result indicates smart glasses efficiency in reducing PPT and CRT.

Table 3 shows the distribution of patients by diagnosis in the study population. This table provides a detailed breakdown of the number of patients and their corresponding percentages for each diagnosis. Orthopedics or injury emerged as the most frequent diagnosis, with 29 cases, constituting 36.25% (n=26) of the overall study cohort. Following closely was respiratory medicine making up 12.5% (n=10) of the total. This distribution provides insights into the prevalence of different diagnoses within the study population, which can be valuable for understanding the spectrum of conditions during the study (see Table 3)

XSL•FO
RenderX

Table . Comparison of sex distribution and age demographics between the augmented reality telemedicine system (ARTS) group and the non-ARTS group.

Demographics	ARTS	Non-ARTS	*P* value
Sex, n (%)			>.99
Male	7 (70)	49 (70)	
Female	3 (30)	21 (30)	
Age (years)			.51
Mean (SD)	38.6 (21.0)	33.3 (12.7)	
Median (IQR)	31.0 (16 - 81)	30.5 (16 - 71)	

Table . Time intervals statistics by 2 groups: comparison of prepreparation time (PPT) and consult response time (CRT) between the augmented reality telemedicine system (ARTS) group and the non--ARTS group.

	ARTS	Non-ARTS	*P* value
PPT[a] (min)			<.001
Mean (SD)	23.0 (8.8)	40.3 (22.7)	
Median (IQR)	26 (10 - 35)	35 (5 - 150)	
CRT[b] (min)			.03
Mean (SD)	15.6 (4.4)	164.8 (51.8)	
Median (IQR)	14 (10 - 25)	162 (71 - 307)	

[a]PPT: prepreparation time from the patient arrival at the door of tertiary hospital to get crucial exams.

[b]CRT: consult response time from the request to the consultant reaching;

Table . Distribution of patients by diagnosis for patients admitted to the emergency department (ED) during the 19th Asian Games and Fourth Asian Para Games Hangzhou 2022 (HAG2022).

Diagnosis	Total patients, n (%)	ARTS[a], n (%)	Non-ARTS, n (%)
Otolaryngology	5 (6.25)	0 (0)	5 (100)
Obstetrics and gynecology	2 (2.5)	1 (50)	1 (50)
Infectious diseases	2 (2.5)	1 (50)	1 (50)
Anorectal surgery	1 (1.25)	0 (0)	1 (100)
Orthopedics or injury	29 (36.25)	5 (17.24)	24 (82.75)
Respiratory medicine	10 (12.5)	1 (10)	9 (90)
Urology	3 (3.75)	1 (33.33)	2 (66.66)
Dermatology	5 (6.25)	0 (0)	5 (100)
General surgery	6 (7.5)	0 (0)	6 (100)
Neurology	2 (2.5)	1 (50)	1 (50)
Gastroenterology	9 (11.25)	0 (0)	9 (100)
Cardiology	4 (5)	0 (0)	4 (100)
Ophthalmology	2 (2.5)	0 (0)	2 (100)

[a]ARTS: augmented reality telemedicine system.

Discussion

Principal Findings and Comparison With Previous Work

This study was conducted at SRRSH in Hangzhou, China, a tertiary referral center affiliated with Zhejiang University School of Medicine. The hospital has around 1200 beds, including an 80-bed Emergency Department that handles an average of 300 patients per day. With its extensive range of consulting specialties, SRRSH played a pivotal role in emergency care, making it a crucial medical hub during the HAG2022. This study highlights the effectiveness of AR smart glasses in improving patient care efficiency and reducing ED wait times, particularly in resource-constrained settings and during mass

XSL•FO
RenderX

gatherings like large-scale sports events. The research focused on 2 key time intervals: PPT and CRT. The key findings of this study reveal that AR smart glasses significantly reduced both PPT and CRT compared to conventional telecommunication methods. Findings revealed that the ARTS group had significantly shorter PPT (23 minutes) compared to the non-ARTS group (40.3 minutes). Similarly, CRT was markedly reduced in the ARTS group (15.6 minutes) versus the non-ARTS group (164.8 minutes), as shown in Table 2. These results emphasize the efficiency of the AR telemedicine system in expediting consultations and decision-making. The reductions in both PPT and CRT demonstrate the potential of smart glasses to enhance the effectiveness of EMS. By allowing doctors to visually assess and consult with patients remotely before their arrival, the smart glasses enable immediate guidance and preparation of necessary resources. This real-time interaction ensures that the ED team is well-prepared to deliver prompt care and decisions as soon as the patient arrives, potentially leading to improved clinical outcomes.

Previous studies show that the most common use of smart glasses is in medical education and telementoring, with a growing application in surgical and procedural training, highlighting their potential to enhance hands-on learning and remote guidance in medical settings [23]. Studies show smart glasses in EMS offer significant benefits by enabling hands-free communication, real-time data sharing, and seamless documentation, ultimately improving workflow efficiency and enhancing patient care in high-pressure environments. These devices also assist in secondary triage during mass casualty incidents by transmitting live video feeds to hospital physicians, enabling them to make informed decisions based on real-time data. As a result, the triage accuracy rate increased significantly, from 58% in the conventional method to 92% when using the AR triage algorithm, demonstrating a substantial improvement in the quality of triage [21,24,25]. To the best of our knowledge this study is considered the first study to implement AR glasses in real-case scenarios during a large-scale sport game competition to assess PPT and CRT. The significantly lower PPT and CRT of the ARTS group compared to the non-ARTS highlights the value of AR smart glasses in providing a first-person, immersive view that could further enhance real-time assessments and decision-making processes to support clinical decisions and optimize workflow in the ED (see Figure 2).

Limitations

This study was conducted during the HAG2022, with SRRSH serving as the main tertiary referral center. However, other medical facilities in the region also managed emergency cases, and not all patients from the event were admitted to SRRSH. This study encountered several limitations and challenges including the deployment of AR smart glasses was limited by the availability of the smaller size of the ARTS group (n=10) compared to the non-ARTS group (n=70), which reflects logistical challenges during HAG2022, including single-center nature of the study and the limited availability of AR smart glasses to only 3 units. Therefore, necessitating careful case prioritization and the need for trained personnel to operate them. Despite these constraints, appropriate statistical methods were employed to address potential biases caused by unequal group sizes. A key concern of the study was the potential disruption to regular workflows. To address this, the successful implementation of AR smart glasses was supported by dedicated teams of trained personnel, ensuring that regular ED staff could continue their routine tasks without added strain. However, for broader adoption in typical hospital settings, an assessment of resource requirements, including device availability, staff training, and the impact on existing workflows, would be necessary to ensure successful integration into clinical environments. In addition to tasks that benefit from AR, such as trauma and neurological assessments (eg, stroke evaluations), where visual and real-time interaction is essential, AR will also support the maintenance of a multidisciplinary treatment (MDT), enabling physicians from various specialties and departments to collaborate and manage patient care. There are certain simpler tasks like patient registration and basic triage that can be efficiently handled without the need for AR. Nevertheless, even in these cases, AR contributes value by enabling the collection of patient data during transport, which minimizes delays upon arrival and enhances the ED's preparedness (see Figure 3). There was 5G technology used in this study; although it enabled minimal latency and high speed however a subset of patients was excluded from the study due to technical difficulties, such as connectivity issues and device malfunctions. Connectivity problems, including network congestion and wireless interference, were especially prominent in the crowded environment of the games and often disrupted real-time communication. Device-specific malfunctions, such as screen freezes, audio failures, and hardware glitches, further hindered the reliability of the AR smart glasses. In some cases, these challenges necessitated the reassignment of AR-assisted cases to the non-ARTS group. While this reassignment caused minimal delays in PPT and CRT, it did reduce the number of cases in the ARTS group, limiting the dataset size. While these challenges were promptly addressed using backup communication systems, Another limitation of this study is the absence of detailed data on the frequency and impact of technical difficulties, with specific data not being collected on how these issues affected patient care, particularly in time-sensitive conditions such as acute coronary syndrome or stroke. The lack of this data limits our understanding of the reliability of AR technology in critical medical scenarios, which is important for future implementations. Although these disruptions resulted in only minimal delays in PPT and CRT, they did impact the statistical power of the study. Furthermore, differences in patient diagnoses, particularly the higher number of orthopedic injuries in the ARTS group may influence the results. These gaps should be addressed in future studies to assess the overall effectiveness of AR in emergency settings.

XSL•FO
RenderX

Figure 3. This figure emphasizes the tasks that augmented reality (AR) addresses most effectively. AR enhances even simpler tasks, such as patient registration, by enabling data collection during transportation and minimizing delays.

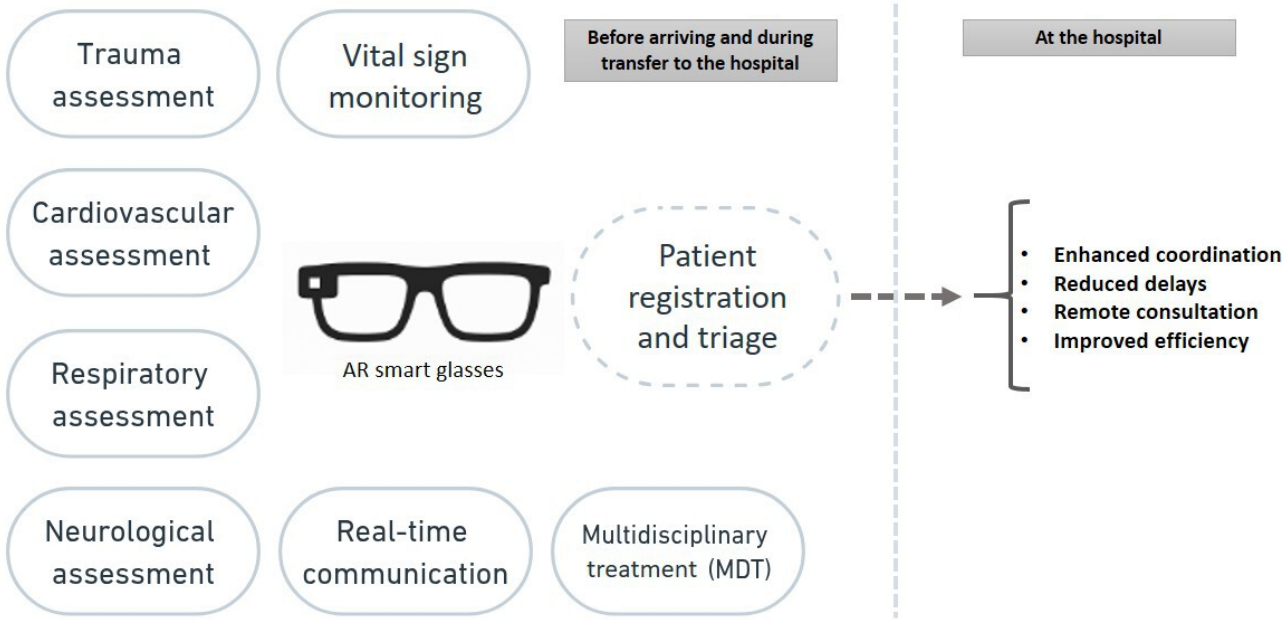

Conclusions

In conclusion, this study demonstrates the potential of AR smart glasses in improving the efficiency of emergency care by reducing prepreparation and consultation response times. The findings suggest that AR technologies could be particularly beneficial in emergency departments, especially in high-demand, resource-constrained environments like those experienced during large public events. To improve the implementation of AR smart glasses in emergency medical settings, future studies should focus on addressing the challenges. Enhancements such as improved network infrastructure, more reliable hardware, and comprehensive troubleshooting training for personnel are essential. In addition, future research with larger population and longer periods is needed to validate these findings and explore the scalability of AR telemedicine in diverse emergency scenarios.

Acknowledgments

This study was funded by the China Fundamental Research Funds for the Central Universities (226-2023-00123) and supported by the Municipal-University Cooperation Project of Taizhou Institute of Zhejiang University (2024TZSX110). The authors would like to express their sincere gratitude to all physicians, nurses, and medical students who contributed to this study.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to privacy and confidentiality concerns; however, they can be obtained from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Establish multilevel system of Asian Games medical support based on an augmented reality (AR) real health care system.
[PDF File, 3022 KB - formative_v9i1e65617_app1.pdf]

References

1. Cheng Y, Cao X, Zhang J, Xu L, Qin L. Comprehensive assessment of emergency departments in county-level public hospitals: a multicenter descriptive cross-sectional study in Henan province, China. Front Public Health 2023;11:1301030. [doi: 10.3389/fpubh.2023.1301030] [Medline: 38035286]
2. Sartini M, Carbone A, Demartini A, et al. Overcrowding in emergency department: causes, consequences, and solutions-a narrative review. Healthcare (Basel) 2022 Aug 25;10(9):1625. [doi: 10.3390/healthcare10091625] [Medline: 36141237]
3. Bernstein SL, Aronsky D, Duseja R, et al. The effect of emergency department crowding on clinically oriented outcomes. Acad Emerg Med 2009 Jan;16(1):1-10. [doi: 10.1111/j.1553-2712.2008.00295.x] [Medline: 19007346]

XSL•FO
RenderX

4. Morley C, Unwin M, Peterson GM, Stankovich J, Kinsman L. Emergency department crowding: a systematic review of causes, consequences and solutions. PLoS ONE 2018;13(8):e0203316. [doi: 10.1371/journal.pone.0203316] [Medline: 30161242]
5. Salway R, Valenzuela R, Shoenberger J, Mallon W, Viccellio A. Emergency department (ED) overcrowding: evidence-based answers to frequently asked questions. Revista Médica Clínica Las Condes 2017 Mar;28(2):213-219. [doi: 10.1016/j.rmclc.2017.04.008]
6. Ataman MG, Sariyer G, Saglam C, Karagoz A, Unluer EE. Factors relating to decision delay in the emergency department: effects of diagnostic tests and consultations. Open Access Emerg Med 2023;15:119-131. [doi: 10.2147/OAEM.S384774] [Medline: 37143526]
7. Li L, Georgiou A, Vecellio E, et al. The effect of laboratory testing on emergency department length of stay: a multihospital longitudinal study applying a cross-classified random-effect modeling approach. Acad Emerg Med 2015 Jan;22(1):38-46. [doi: 10.1111/acem.12565] [Medline: 25565488]
8. Kawano T, Nishiyama K, Hayashi H. Execution of diagnostic testing has a stronger effect on emergency department crowding than other common factors: a cross-sectional study. PLoS ONE 2014;9(10):e108447. [doi: 10.1371/journal.pone.0108447] [Medline: 25310089]
9. Brick C, Lowes J, Lovstrom L, et al. The impact of consultation on length of stay in tertiary care emergency departments. Emerg Med J 2014 Feb;31(2):134-138. [doi: 10.1136/emermed-2012-201908] [Medline: 23353666]
10. Jaffe TA, Goldstein JN, Yun BJ, et al. Impact of emergency department crowding on delays in acute stroke care. West J Emerg Med 2020 Jul 8;21(4):892-899. [doi: 10.5811/westjem.2020.5.45873] [Medline: 32726261]
11. Reznek MA, Murray E, Youngren MN, Durham NT, Michael SS. Door-to-imaging time for acute stroke patients is adversely affected by emergency department crowding. Stroke 2017 Jan;48(1):49-54. [doi: 10.1161/STROKEAHA.116.015131] [Medline: 27856953]
12. Jones S, Moulton C, Swift S, et al. Association between delays to patient admission from the emergency department and all-cause 30-day mortality. Emerg Med J 2022 Mar;39(3):168-173. [doi: 10.1136/emermed-2021-211572] [Medline: 35042695]
13. 19th Asian Games Hangzhou declared open with grand opening ceremony 2023. Olympics. URL: https://olympics.com/en/news/hangzhou-asian-games-opening-ceremony [accessed 2024-08-18]
14. Ansari A. Asian Games 2022 in China postponed to 2023. Olympics. 2022. URL: https://olympics.com/en/news/asian-games-2022-postponed [accessed 2024-08-18]
15. Yuan Y, Jiang H, Qi J, et al. Overall design and implementation of the medical emergency support plan for the 19th Asian Games in Hangzhou. Chin J Emerg Med 2023;32(12):1655-1659. [doi: 10.3760/cma.j.issn.1671-0282.2023.12.013]
16. Kuroki N, Yagishita K, Shimizu K, et al. Emergency department activities at the Athletes' Village during the Tokyo 2020 Olympic and Paralympic Games. Acute Med Surg 2023;10(1):e905. [doi: 10.1002/ams2.905] [Medline: 38020491]
17. Enock KE, Jacobs J. The Olympic and Paralympic Games 2012: literature review of the logistical planning and operational challenges for public health. Public Health 2008 Nov;122(11):1229-1238. [doi: 10.1016/j.puhe.2008.04.016] [Medline: 18619630]
18. Hadjichristodoulou C, Mouchtouri V, Soteriades ES, et al. Mass gathering preparedness: the experience of the Athens 2004 Olympic and Para-Olympic Games. J Environ Health 2005 May;67(9):52-57. [Medline: 15957323]
19. Voaklander B, Gaudet LA, Kirkland SW, Keto-Lambert D, Villa-Roel C, Rowe BH. Interventions to improve consultations in the emergency department: A systematic review. Acad Emerg Med 2022 Dec;29(12):1475-1495. [doi: 10.1111/acem.14520] [Medline: 35546740]
20. Zhang Y, Zhang J, Yuan Y, et al. Practice and discussion on ambulance medical support in the 19th Hangzhou Asian Games. Chin J Emerg Med 2023;32(12):1628-1633 [FREE Full text] [doi: 10.3760/cma.j.issn.1671-0282.2023.12.008]
21. Zhang Z, Ramiya Ramesh Babu NA, Adelgais K, Ozkaynak M. Designing and implementing smart glass technology for emergency medical services: a sociotechnical perspective. JAMIA Open 2022 Dec;5(4):ooac113. [doi: 10.1093/jamiaopen/ooac113] [Medline: 36601367]
22. Dinh A, Yin AL, Estrin D, Greenwald P, Fortenko A. Augmented reality in real-time telemedicine and telementoring: scoping review. JMIR Mhealth Uhealth 2023 Apr 18;11:e45464. [doi: 10.2196/45464] [Medline: 37071458]
23. Broach J, Hart A, Griswold M, et al. Usability and reliability of smart glasses for secondary triage during mass casualty incidents. Proc Annu Hawaii Int Conf Syst Sci 2018 Jan 3;2018:1416-1422. [doi: 10.1002/ams2.807] [Medline: 29398976]
24. Ishikawa K, Yanagawa Y, Ota S, et al. Preliminary study of prehospital use of smart glasses. Acute Med Surg 2022;9(1):e807. [doi: 10.1002/ams2.807] [Medline: 36330311]
25. Follmann A, Ohligs M, Hochhausen N, Beckers SK, Rossaint R, Czaplik M. Technical support by smart glasses during a mass casualty incident: a randomized controlled simulation trial on technically assisted triage and telemedical app use in disaster medicine. J Med Internet Res 2019 Jan 3;21(1):e11939. [doi: 10.2196/11939] [Medline: 30609988]

Abbreviations

AR: augmented reality

XSL•FO
RenderX

ARTS: augmented reality telemedicine system
CRT: consult response time
ED: emergency department
EMS: emergency medical services
HAG2022: 19th Asian Games and Fourth Asian Para Games Hangzhou 2022
PPT: prepreparation time
SRRSH: Sir Run Run Shaw Hospital

Edited by A Mavragani; submitted 26.08.24; peer-reviewed by AL Yin; revised version received 04.05.25; accepted 08.05.25; published 30.06.25.

Please cite as:
Jiang X, Xia B, Ansari MM, Jiang H, Qi J, Zhang Z, Dai S, Zheng P, He Y, Liu N, Chen P, Luo R, Qin X, Miao Y, Dai J, Zhou X, Wang C, Chen H, Xu W, Wu T, Shi Q, Chen Z, Zhou L, Zhang H, Xie Y, Zhang Q, Zhou B, Pan X, Chen Z, Zhen L, Sun Y, Lu Z, Loh Y, Sayer S, Mochtar J, Wongpraewit P, Wang Y, Hong Y
Evaluating the Effectiveness of Smart Glasses in Reducing Patient Care Time in Emergency Departments: Cohort Study From the Hangzhou Asian Games
JMIR Form Res 2025;9:e65617
URL: https://formative.jmir.org/2025/1/e65617
doi:10.2196/65617

© Xinwei Jiang, Bangbo Xia, Mohammad Mostafa Ansari, Huiquan Jiang, Jianjiang Qi, Zhongheng Zhang, Sheng Dai, Pingping Zheng, Yang He, Ning Liu, Pengpeng Chen, Ronghua Luo, Xuchang Qin, Yansong Miao, Junru DAI, Xiaoyu Zhou, Changliang Wang, Hui Chen, Wenbin Xu, Tao Wu, Qiang Shi, Zhonghua Chen, Liping Zhou, Hao Zhang, Yun Xie, Quan Zhang, Bifa Zhou, Xiaohong Pan, Zixi Chen, Libo Zhen, Yaqing Sun, Zelin Lu, Yihao Loh, Shameera Sayer, Jennifer Mochtar, Pannika Wongpraewit, Yifan Wang, Yucai Hong. Originally published in JMIR Formative Research (https://formative.jmir.org), 30.6.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Coaching Patients to Understand and Use Patient-Reported Outcome Data: Intervention Design and Evaluation

Martha Burla[1], MPH; Brocha Z Stern[2], OTR, PhD; Andrew BL Berry[1], PhD; Sarah Pila[1], PhD; Patricia D Franklin[1], MD, MBA, MPH

[1]Department of Medical Social Sciences, Feinberg School of Medicine, Northwestern University, 625 N Michigan Ave, Chicago, IL, United States
[2]Departments of Population Health Science and Policy and Orthopaedics, Icahn School of Medicine at Mount Sinai, New York, NY, United States

Corresponding Author:
Martha Burla, MPH
Department of Medical Social Sciences, Feinberg School of Medicine, Northwestern University, 625 N Michigan Ave, Chicago, IL, United States

Abstract

Background: Providing patients with information about their health and treatment options is important to ensure care that best reflects patient needs, values, and preferences. Patient-reported outcomes (PROs), measures of health status, are regularly collected in clinical contexts and scores can be returned to patients in personalized decision aids. One example of a PRO-based decision aid is the Arthritis care through Shared Knowledge (ASK) report, which shares individual PRO data on knee and hip arthritis–related pain and functional limitations with patients. However, given that the use of such data in clinical consultations is unfamiliar to many patients, support may be required to ensure this information is understood and used as intended.

Objective: This paper describes ASK coaching, an online 1-hour group session designed to ensure patients understood the ASK report, including their PRO scores, and how to use the information in conversations with their clinicians. We present (1) quantitative evaluation results associated with attendance and self-assessment of learning and (2) qualitative evaluation results on motivation to attend, acceptability of the session format, and achievement of session goals.

Methods: The session was designed and refined collaboratively with clinical experts and patient advisers. Patients in one arm of a pragmatic cluster-randomized trial evaluating the ASK report were invited to attend this session. To understand the profile of attendees (N=438) sociodemographic and clinical data were compared with all participants invited to coaching (N=1545) and a patient-reported assessment of self-efficacy was collected on a subset (N=692). In addition, a postsession survey was used to self-assess learning. Qualitative data were synthesized from semistructured postcoaching interviews, paired pre- and postcoaching interviews, and free-text responses to a postsession survey. A qualitative descriptive approach was used for analysis.

Results: Compared with nonattendees, patients reporting higher education, greater health literacy, Medicare insurance, and lower self-efficacy for managing treatments were more likely to attend ASK coaching when invited. Participants' self-assessment of learning showed an improved understanding of current and projected osteoarthritis symptoms and where to find additional information. Qualitatively, patients reported attending coaching to gain information that could benefit their treatment or aid in research. The online group format was generally described as acceptable, and the session goals related to understanding the report and preparing for future conversations with clinicians were met. Suggestions for improvement, such as providing more opportunities for within-group interaction, were also provided.

Conclusions: Our results highlight the value of coaching as an intervention to help patients understand and use novel health information, including PRO data, in conversations with clinicians. Given that it was well-liked by patients, promoting a greater understanding of the PRO-based ASK report, and increased feelings of preparedness for clinical consultation, coaching appears to be a promising intervention to support patients in understanding and using their personal health data.

(JMIR Form Res 2025;9:e65931) doi:10.2196/65931

KEYWORDS

osteoarthritis; shared decision making; patient coaching; patient-reported outcomes; surgical decisions; health status; arthritis; arthritis care; shared knowledge; clinical consultation; quantitative evaluation; self-assessment; self-efficacy; patient empowerment; coaching intervention; health information; decision aid; patient support

XSL•FO
RenderX

Introduction

Patient-reported outcomes (PROs)—standardized assessments of patients' pain, function, and health-related quality of life—are increasingly collected in routine clinical settings [1,2]. PRO scores reflect the patient's perspective on their health and can be used to guide collaborative patient-clinician conversations and care decisions [3,4]. Specifically, PRO scores can be embedded into decision reports used with clinicians at the point of care or accessed via the electronic patient portal and used independently by patients to prepare for a clinical consultation [5,6]. Such uses of PRO data may be particularly important when managing a progressive chronic disease like osteoarthritis. Since patients' gradual decline may lead to unrecognized impairments, PRO information can help them self-reflect on their current health and symptoms [3,7]. Furthermore, incorporating patient-reported data into a clinical encounter has been shown to improve communication between patients and clinicians [8], which is highly relevant for collaborative conversations and care decisions in chronic conditions including osteoarthritis.

The need to refine decision-making in total joint replacement, a common treatment for osteoarthritis, has been specifically acknowledged due to its status as a high-volume elective surgery with known disparities in usage [9-12]. In response, several PRO-based decision aids have been developed to support decision-making for hip and knee osteoarthritis [13-15]. One such report is the Arthritis care through Shared Knowledge (ASK) report, which shares personalized PRO data on current pain and function as well as estimated postoperative symptom information, with patients and clinicians [16]. Detailed information about the content and design of the report, as well as clinician and patient perspectives on report use, are available elsewhere [16-19].

While patients may want to use decision aids, like the ASK report, to engage in conversations about their care, they do not always feel able to do so. Instead, some patients report that they fear upsetting physicians who might perceive patient self-advocacy as a challenge to their authority [20]. In addition, the independent use of personal health data is new to many patients, and in the case of the ASK report specifically, they may experience difficulty understanding what PRO scores represent, and how they can be used in clinical conversations and care decisions [18,21]. Last, in patient interviews completed during the development of the ASK report, patients suggested that additional education may be helpful to ensure patients can independently use personal health data in clinical conversations and feel comfortable doing so.

"Coaching," a form of nondirective support provided by a trained individual, has been discussed as part of supporting patients in combination with decision aids [22]. Coaching can occur outside of clinical consultations and may have utility in supporting patients to optimally understand and use personal health data in conversations with clinicians. As such, ASK coaching was created and embedded within a pragmatic trial of the ASK report to support patients in (1) understanding their personal data, including the PRO scores, and (2) using that data in conversations with clinicians. This paper describes the ASK coaching session design process and presents the results of a multi-method evaluation of the coaching intervention.

Methods

Overview of the Parent Study

The ASK study was a pragmatic cluster-randomized trial with the primary aim of evaluating the effectiveness of a personalized PRO-based report containing estimates of likely outcomes of surgery if elected to treat advanced hip and knee osteoarthritis (ClinicalTrials.gov identifier: NCT03102580). A copy of the ASK report is in Multimedia Appendix 1. The study enrolled patients who scheduled an initial visit with one of 36 orthopedic surgeons to discuss knee and hip arthritis treatment options, including total joint replacement surgery. Aim 1 involved report co-design, with patients and clinicians, aim 2 aimed to compare decision quality for patients at sites randomized to receive or not receive the report, and aim 3 included adding a patient coaching intervention ("ASK coaching") to a randomly assigned subset of study sites. Figure 1 depicts study procedures during aim 3. In this aim, patients completed PROs prior to their initial office-based consultation with a participating orthopedic surgeon. Within the month following this initial consultation, study staff contacted all patients via telephone to invite participation in the study and assess decision quality. Participants receiving care in study sites randomized to coaching were also invited to attend an ASK coaching session at this time.

XSL•FO
RenderX

Figure 1. The figure outlines the primary components of the study procedures for aim 3 (May 2020-February 2022) of a pragmatic cluster-randomized trial for patients with hip and knee osteoarthritis. The noncoaching arm (black box) is not the focus of this paper. The boxes with dashed borders reflect study activities that do not involve data collection. ASK: Arthritis care through Shared Knowledge.

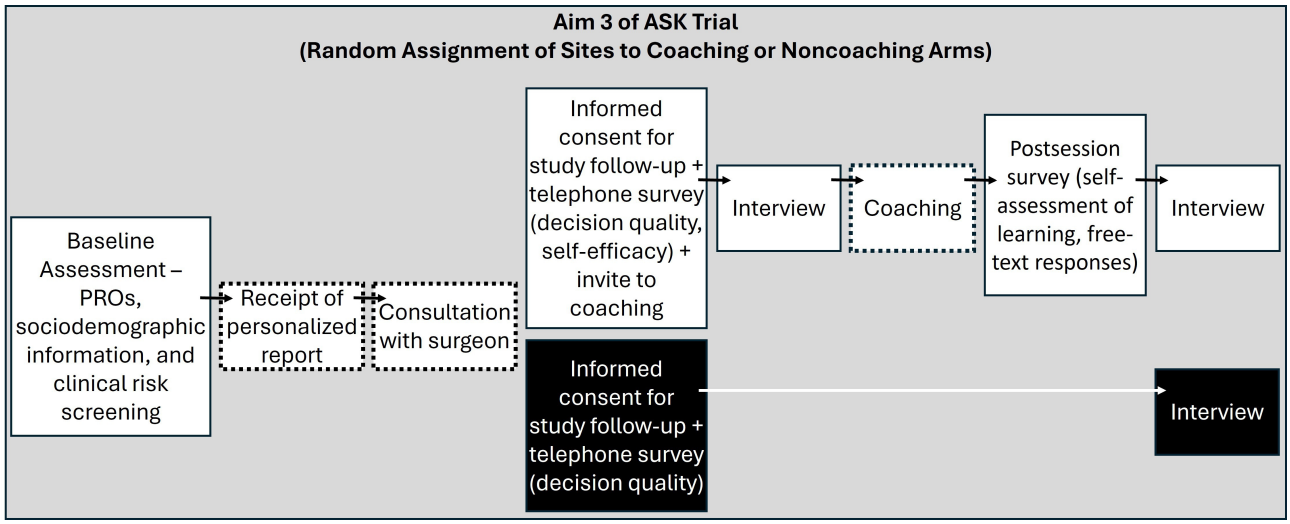

Coaching Intervention Co-Design and Iteration

During the design of the parent study, patient advisers who had recently made decisions regarding the treatment of their hip or knee osteoarthritis were consulted [16]. A subset of those patients expressed that an opportunity to review the content of the ASK report in a group setting would be helpful. The ASK coaching intervention was included in the study design based on this recommendation and, like the ASK report, was developed in collaboration with patient partners and clinical stakeholders.

The research team determined that the coaching intervention would focus on helping participants understand the personal data provided in the ASK report and how to use that data in future conversations with clinicians. The research team then consulted with Citizens for Patient Safety (CPS), an organization that educates patients and clinicians about medical decision-making and navigating the health care system. CPS recommended further helping patients prepare for clinical consultations by encouraging them to be active members of the health care team and helping them develop a list of questions to ask clinicians. They specifically highlighted the Agency for Healthcare Research and Quality's (AHRQ) "Questions are the Answer" materials [23] as a potentially useful tool to guide the session.

A trained health educator (an individual with a Master of Public Health degree and a certification in health education) then developed the hour-long online group coaching session and an osteoarthritis-tailored patient handout based on the AHRQ materials recommended by CPS. This handout, entitled "I Have a Voice" (Multimedia Appendix 2), was used throughout to help patients reflect on their data and how it might inform future conversations with clinicians. The research team also created a website consolidating additional patient resources from reputable sources, including AHRQ, Arthritis Foundation, American Academy of Orthopaedic Surgeons, and American Association of Hip and Knee Surgeons. After completing this initial design, the research team met with clinical experts, including a patient advocacy leader from CPS, a senior orthopedic surgeon, a physical therapist, a health psychologist, and a group of research coordinators with experience recruiting patients with osteoarthritis. Experts attended a mock coaching session and provided input in a semistructured interview. Specific suggestions were extracted from the interview notes and used to refine the coaching intervention organization, content, and support materials. Experts' suggestions primarily focused on the best way to communicate sensitive topics during coaching and the framing of patient predictions of likely outcomes if surgery was elected. For the former, clinicians encouraged the use of positive language, such as referring to comorbidities as "areas for possible improvement to optimize surgical outcomes." For the latter, they suggested emphasizing that a good surgical outcome is one that improves arthritis pain and stiffness, but may not necessarily result in a perfect joint. Experts also offered advice on addressing the needs of nonsurgical patients including the importance of understanding personal data and setting goals regardless of one's treatment plan.

Mock sessions using the refined intervention were then scheduled with ten patient advisers with hip or knee osteoarthritis recruited through the Arthritis Foundation. Patient advisers attended a session and provided feedback in either a one-on-one interview or written format. Generally, their feedback was positive. Most found reviewing the report during coaching to be helpful because they did not understand the PRO data on their own, and almost all mentioned valuing the supportive environment in which to discuss their concerns. Patient advisers were also enthusiastic about the "I Have a Voice" handout. Small changes to language and emphasis were made based on the patient advisor's feedback, but the overarching session goals and content did not change.

The coaching intervention was launched for ASK study participants in early May 2020. Each initial session was staffed by two research team members: (1) the health educator who facilitated the session and (2) a research coordinator who gathered feedback from observations and provided technological support to attendees as needed. At the end of the first month of coaching, minor logistical changes were made, including adding answers to frequently asked questions and clarifying the instructions on joining a meeting via telecommunication

XSL•FO
RenderX

platform Zoom (Zoom Video Communications). No additional changes were made after the first month. An outline of the final coaching session content is presented in Textbox 1.

Textbox 1. Textbox 1. The components of the final Arthritis care through Shared Knowledge (ASK) coaching session embedded within a pragmatic cluster-randomized trial for patients with hip and knee osteoarthritis.

Presession Materials (Mailed and Emailed to Participants):

- Personalized ASK report (Multimedia Appendix 1)
- "I Have a Voice" handout (Multimedia Appendix 2)
- Instructions on how to join the meeting using the telecommunication platform (Zoom)
- Paper copy of the session slides

Session Outline (1 h Online, Group):

- Introductions (facilitator and attendees)
- Review of session goals
- Explanation of the importance of patient self-advocacy
- Detailed review of the ASK report and preparation to use the data in conversations with clinicians
  - Report section: Current pain and function scores
  - Reflection prompt: Is there anything that you might like to tell your clinicians about your pain and function levels?
  - Report section: Patient comorbidities
  - Reflection prompt: What aspects of your health and wellness would you like to address with your health care team, and what questions do you have?
  - Report section: Predictions of postsurgical pain and function
  - Reflection prompt: What are your goals for improvement post osteoarthritis treatment (including nonsurgical treatment)?
  - Report section: Nonsurgical treatment options
  - Reflection prompt: Are there any treatments about which you'd like to request more information?
- Additional strategies for a successful appointment
  - Planning to bring a trusted other
  - Asking to record clinicians' instructions
- Review of additional resources

Recruitment for the Coaching Intervention

All participants in the study's coaching arm were invited to attend the online, group coaching session at the time of the postconsultation telephone survey. Multiple sessions were scheduled each week at varying times to accommodate patient schedules. Those who agreed to participate were scheduled for a session and both mailed and emailed the presession materials (Figure 2). While all study participants in this aim received a copy of their ASK report, only those who agreed to participate in the coaching session received the other materials. Participants scheduled for a coaching session received a reminder the day before, which included the number of a dedicated research coordinator who they could contact for technical support before or during the session.

Coaching consent and attendance rates were tracked throughout. Of all patients invited to coaching (n=1389), 45% consented to participate, and 32% of the total (N=438) attended a scheduled session. Among those who consented, 21% either did not schedule a session at the time of consent or scheduled a session but canceled, and 9% scheduled a session but did not show. Figure 2 summarizes attendance and cancellation rates. Research coordinators recorded reasons for declining to attend coaching, which included participants feeling they either did not need additional information about the report or their arthritis symptoms or were too busy to attend. Sessions had an average of 3 - 4 attendees each, and enrollment extended from May 2020 through February 2022, which overlapped with the peak of the COVID-19 epidemic.

XSL•FO
RenderX

Figure 2. Attendance metrics for ASK coaching from May 2020 through February 2022 within a pragmatic cluster-randomized trial for patients with hip and knee osteoarthritis. ASK: arthritis care through shared knowledge.

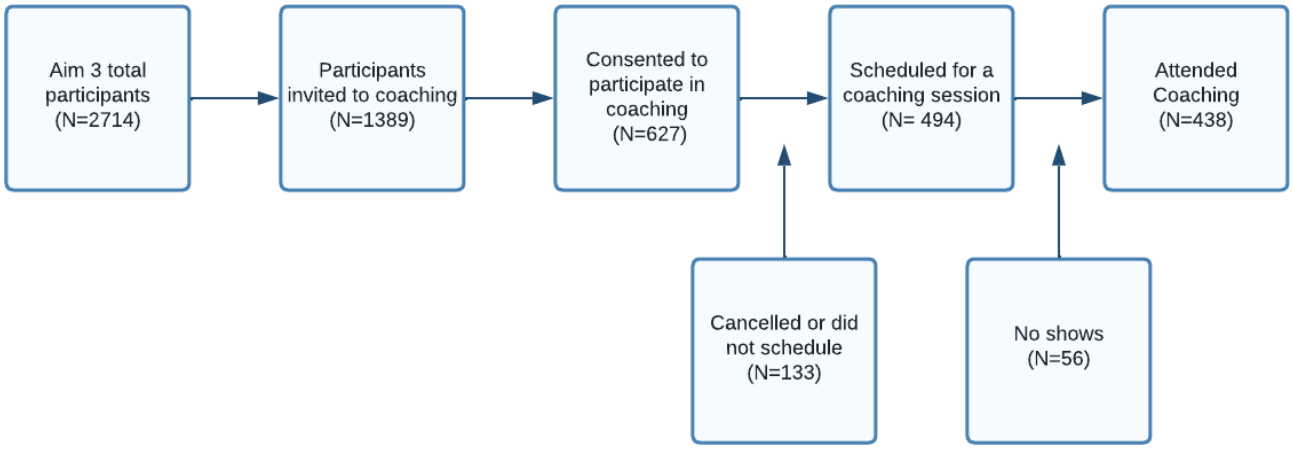

Quantitative Evaluation of the Coaching Intervention

The timing of the quantitative coaching assessments within the broader aim 3 study flow can be found in Figure 1. To understand the profile of patients who attended coaching, sociodemographic information and clinical comorbidity data were obtained from all participants at sites in the study's coaching arm and were descriptively summarized and compared by coaching attendance status using chi-square and 2-sample *t* tests that accounted for possible unequal variance. PROs included the Hip disability and Osteoarthritis Outcome Score (HOOS) or Knee injury and Osteoarthritis Outcome Score (KOOS-12) [24,25], osteoarthritis-specific PROs used broadly in research and clinical care, and the Veterans RAND 12-Item Health Survey (VR-12) [26] that assesses global physical and emotional function. Patients also self-reported pain intensity in other joints and lower back, sociodemographic factors, and medical comorbidities. A question to assess confidence in completing health forms screened for health literacy [27]. In late March 2021, roughly 1 year after the launch of ASK coaching, a patient-reported assessment of self-efficacy was added to the postconsultation telephone survey (ie, the time of the coaching invitation) to further understand the profile of those who did or did not choose to attend coaching sessions. This included 4 treatment-related items from the PROMIS (Patient-Reported Outcomes Measurement Information System) v1.0 Self-Efficacy for Managing Medications and Treatments [28] item bank (items SEMMT011, SEMMT012, SEMMT013, and SEMMT014), and was administered to the last 692 participants invited to coaching. Raw scores on this custom short form were converted to T-scores using the Health Measures Scoring Service. The T-scores were compared between those who attended and did not attend coaching using a Mann-Whitney *U* test as the scores were not normally distributed. A 2-tailed α level of $<.05$ was considered statistically significant for all inferential analyses.

In addition, to quantitatively assess attendee learning from coaching, a postsession survey was added in late March 2021. It included five self-assessment questions related to perceived change in knowledge and was sent via email after the coaching session. Responses to the self-assessment questions were descriptively summarized.

Qualitative Evaluation of the Coaching Intervention

Data were integrated from multiple sources to qualitatively evaluate the ASK coaching intervention. First, questions related to acceptability and perceived benefits of coaching were included in semistructured interviews conducted in April-July 2021 as part of the parent study; 16 of these 25 interview participants had attended a coaching session. Only coaching-related segments from these 16 interviews were included in the present analysis; findings from other qualitative analyses of these interviews have been previously published [17,18]. Second, paired pre- and postcoaching interviews were conducted for 4 additional attendees in February 2022. These interviews provided finer-grained feedback on how coaching influenced participants' understanding and potential use of PROs including in conversations with providers. All interviews were completed by a trained interviewer via Zoom and used a semistructured interview guide; they were audio-recorded and transcribed. Third, qualitative data were extracted from optional free-text responses provided by 47 coaching attendees in the postsession survey described above. Refer to Figure 1 for positioning of the qualitative assessments within the broader aim 3 study flow.

We used a qualitative descriptive approach to analysis, which aligned with our pragmatic and focused qualitative aims to learn more about what motivated participants to attend coaching sessions, assess the acceptability of an online group coaching session, and evaluate whether the session met the goals laid out by the research team [29-32]. After independent reading and preliminary coding of the 4 paired pre- and postcoaching interview transcripts, MB and BZS collaboratively developed a deductive codebook based on the qualitative aims. To increase trustworthiness, all interview transcripts and survey comments were dually coded by MB and BZS, with ongoing meetings to discuss emerging findings, iteratively refine the codebook, and resolve discrepant coding.

Ethical Considerations

To ensure informed consent, all participants in the parent study were provided with a detailed consent form outlining the study's procedures, potential risks and benefits, and their right to withdraw at any time. Participants in the coaching arm received an additional fact sheet with details about the coaching session,

XSL•FO
RenderX

as did those who were invited to participate in an interview. Participants were paid US$ 20 for completing the postconsultation telephone survey as part of the parent study and an additional US$ 20 if they participated in an interview. No additional incentives were offered for participating in coaching or completing the postsession survey. All data collected as part of this study were deidentified and stored on a secure server. All study procedures were approved by the University of Massachusetts Medical School institutional review board (H00012297).

Results

Quantitative Evaluation of the Coaching Intervention

Sociodemographic and clinical profiles of coaching attendees versus nonattendees are summarized in Table 1. Compared with nonattendees, patients reporting higher education and greater health literacy were more likely to attend a coaching session when invited ($P<.001$). In addition, those with Medicare insurance were more likely to attend while those with Medicaid insurance were less likely to attend ($P<.001$). Medicare and Medicaid are US government–run health insurance programs. Medicaid provides insurance to low-income families and individuals. Medicare provides insurance to senior citizens and some individuals with disabilities. While there was a statistically significant between-group difference for age ($P=.002$), the mean difference was less than 2 years between those who attended versus those who did not. Furthermore, while there were no statistically significant differences, those who attended coaching were more likely to be female, have at least one comorbidity, and have slightly less preconsultation pain (Table 1). In addition, the subsample of participants with self-efficacy data, self-efficacy for managing treatments was lower in those who attended coaching (n=166; median T-score 48.6, IQR 41.8-57.3) than those who did not (n=526; median T-score 57.3, IQR 42.7-57.3; $P=.05$).

The postsession self-assessment of learning was completed by 89 coaching attendees (49% of those invited to respond). Approximately, two-thirds of respondents reported greater understanding of their current and projected symptoms. Along with this, three-quarters of participants reported more knowledge of where to find additional osteoarthritis information after coaching (Table 2).

XSL•FO
RenderX

Table . Sociodemographic and clinical profiles, stratified by coaching attendance status, of participants in the arthritis care through shared knowledge (ASK) coaching arm of a pragmatic cluster-randomized trial in patients with hip and knee osteoarthritis (May 2020-February 2022).

Patient characteristics	Total in coaching arm (n=1545)	Attended coaching (n=438)	Invited, but did not attend (n=1107)	*P* value[a]
Age (years), mean (SD)	66.5 (9.4)	67.7 (8.7)	66.1 (9.7)	.002
BMI, mean (SD)	31.0 (8.9)	30.7 (6.7)	31.1 (9.6)	.33
Preconsultation HOOS[b] or KOOS-12[c] Pain Score, mean (SD)	42.6 (16.3)	43.9 (15.2)	42.1 (16.7)	.05
Preconsultation HOOS or KOOS-12 ADL[d] Score, mean (SD)	51.4 (20.4)	52.4 (19.3)	51.0 (20.7)	.20
Preconsultation VR-12[e] Mental Component Summary Score, mean (SD)	55.8 (11.2)	56.4 (10.7)	55.6 (11.4)	.18
Preconsultation VR-12 Physical Component Summary Score, mean (SD)	31.7 (10.1)	32.0 (9.8)	31.5 (10.2)	.42
Sex, n (%)				
Male	572 (37.2)	146 (33.5)	426 (38.7)	.06
Female	966 (62.8)	290 (66.4)	676 (61.3)	
Race and ethnicity, n (%)				
Non-Hispanic White	1265 (85.8)	370 (87.5)	895 (85.2)	.70
Non-Hispanic Black	127 (8.6)	33 (7.8)	94 (8.9)	
Non-Hispanic other	30 (2.0)	7 (1.7)	23 (2.2)	
Hispanic	52 (3.5)	13 (3.1)	39 (3.7)	
Education, n (%)				
College graduate or post-graduate	862 (56.3)	296 (68.4)	566 (51.6)	<.001
Trade, technical school, or some college	363 (23.7)	79 (18.2)	284 (25.9)	
High school or less	263 (17.2)	47 (10.9)	216 (19.7)	
Other	42 (2.8)	11 (2.5)	31 (2.8)	
Confidence filling out medical forms (Health literacy), n (%)				
Not at all confident	14 (0.9)	1 (0.2)	12 (1.2)	<.001
A little bit confident	22 (1.4)	2 (0.5)	20 (1.8)	
Somewhat confident	75 (4.9)	8 (1.8)	67 (6.1)	
Quite a bit confident	253 (16.4)	72 (16.4)	181 (16.4)	
Extremely confident	1178 (76.4)	355 (81.1)	323 (74.5)	
Live with another adult, n (%)				
No, I live alone	424 (27.9)	127 (29.3)	297 (27.3)	.44
Yes, I live with another adult	1098 (72.1)	307 (70.7)	791 (72.7)	
Current marital status, n (%)				
Married	920 (59.8)	260 (59.6)	660 (59.9)	.93
Not married	618 (40.2)	176 (40.4)	442 (40.1)	

XSL•FO
RenderX

Patient characteristics	Total in coaching arm (n=1545)	Attended coaching (n=438)	Invited, but did not attend (n=1107)	*P* value[a]
Health insurance, n (%)				
Private	594 (38.4)	161 (36.8)	433 (39.1)	<.001
Medicaid[f]	138 (8.9)	22 (5.0)	116 (10.5)	
Medicare[f]	717 (46.4)	238 (54.3)	479 (43.3)	
Other, Missing, None	96 (6.2)	17 (3.9)	79 (7.1)	
Sum of comorbidities, n (%)				
No comorbidities	749 (48.6)	193 (44.3)	556 (50.3)	.11
1 comorbid diagnosis	446 (28.9)	144 (33.0)	302 (27.3)	
2 - 5 comorbid diagnoses	337 (21.9)	97 (22.3)	240 (21.7)	
6 - 20 comorbid diagnoses	9 (.06)	2 (0.5)	7 (0.6)	
Low back pain, n (%)				
No or very mild back pain	1,076 (69.7)	304 (69.6)	772 (69.7)	.95
Moderate to worst imaginable pain	468 (30.3)	133 (30.4)	335 (30.3)	
Number of other hip or knee joints with moderate-severe pain, n (%)				
0	774 (50.1)	222 (50.7)	552 (49.9)	.76
1	560 (36.3)	151 (34.5)	409 (37.0)	
2	142 (9.2)	44 (10.1)	98 (8.9)	
3	69 (4.5)	21 (4.8)	48 (4.3)	

[a]Comparison using χ^2 test or two-sample T-test between attendees and nonattendees.

[b]HOOS: Hip disability and Osteoarthritis Outcome Score.

[c]KOOS-12: Knee injury and Osteoarthritis Outcome Score.

[d]ADL: activities of daily living.

[e]VR-12: Veterans RAND 12-Item Health Survey.

[f]Medicare and Medicaid are US government-run health insurance programs. Medicaid provides insurance to low-income families and individuals. Medicare provides insurance to senior citizens and some individuals with disabilities.

XSL•FO
RenderX

Table . Postsession self-assessment of learning from a subsample of arthritis care through shared knowledge coaching attendees in a pragmatic cluster-randomized of patients with hip and knee osteoarthritis (March 2021-February 2022; N=89).

Compared with before attending coaching	Response (responses: n, %)
... I know how my pain and function compare to others like me.[a]	• I know less now (n=1, 1.14%) • I know about the same (n=20, 22.73%) • I know more now (n=67, 76.13%)
... I know how my pain and function might change if I choose surgery.[a]	• I know less now (n=1, 1.14%) • I know about the same (n=18, 20.45%) • I know more now (n=69, 78.11%)
... I know what outcomes I can expect if I choose surgery.[a]	• I know less now (n=0, 0%) • I know about the same (n=23, 26.13%) • I know more now (n=65, 73.87%)
... I know what other aspects of my health impact my arthritis.	• I know less now (n=1, 1.12%) • I know about the same (n=25, 28.09%) • I know more now (n=63, 70.99%)
... I know where to find information about caring for my arthritis.	• I know less now (n=1, 1.12%) • I know about the same (n=9, 10.11%) • I know more now (n=79, 88.76%)

[a] Missing for one respondent.

Qualitative Evaluation of the Coaching Intervention

The qualitative evaluation results are organized into participants' perspectives on motivation for attendance, acceptability of session format, achievement of session goals, and suggestions for improvement. Additional supporting quotes for each subsection are presented in Multimedia Appendix 3.

Motivation for Attendance

Participants' main motivations for attending ASK coaching were to gain information that could benefit their treatment and to aid in research that could help others. Specifically, attendees expressed interest in improving their understanding of the ASK report (including PROs), getting more information about their condition, developing strategies for future conversations with clinicians, and learning from peers.

> *I think because I was so confused by that first page [of PROs in the report], I wanted some more explanation as to if I was understanding it correctly or how to understand it.* [P09, 70 y.o. White, Female]

Acceptability of the Session Format

The online format of the session was generally acceptable to participants although some would have preferred to meet in person had that option been available. Several participants expressed challenges with the online platform and valued the technical support offered by the ASK team.

> *I thought [the session] was very good. I mean even having an IT person on just in case anybody had trouble with Zoom. I mean it feels like you've thought of everything.* [P04, 59 y.o. White, Female]

The group nature of ASK coaching was also generally acceptable. It was even noted as specifically valuable by participants who valued hearing experiences of others, such as those farther along in the treatment process. This finding aligned well with early feedback from the design process positing that a group format would allow for the sharing of experiences. However, there were some divergent opinions, with two participants who had more advanced osteoarthritis noting that differing disease severity and treatment plans among their fellow participants negatively impacted their session experience. One participant highlighted that the small group format was acceptable at her current age but may not have been when she was younger because of social discomfort discussing health. Several participants also pointed out that certain group members negatively impacted group dynamics but did not feel that was a reason to discontinue group offerings.

> *I liked the group format of it, and maybe it's because I am on the earlier end of this condition, and the other participants were all giving me advice. ... You get that little sense of camaraderie. Everybody's rooting for one another.* [P06, 57 y.o. White, Female]

The presence of a facilitator was discussed by many as important to the acceptability of the format, since it provided structure and focus to the session. Participants highlighted specific benefits of having a health educator as the facilitator, including the opportunity to ask health-related questions from someone outside of their immediate health care team.

> *I appreciate the personal, stress-free atmosphere in which I felt totally comfortable asking questions I would never ask my doctor in the interest of time or seeming uninformed.* [P11, 59 y.o. Black, Female]

Achievement of Session Goals

In line with the session goal of helping participants to understand their personal data, coaching attendees reported gaining a better understanding of the ASK report, including their PRO data. Several discussed that coaching participation improved report understanding and clarified misinterpretations of their data even if they had not experienced initial confusion.

XSL•FO
RenderX

> *I'm just arrogant enough to believe that I understood [the report] thoroughly because I read it carefully. ...And I found at the end of the session I was really glad that I had participated. And maybe it's a little bit like when you teach a person to write, you say, 'First tell them what you're gonna say. Then say it. Then tell them what you've said.' So, I think having read it, and then having gone through it more carefully, hearing someone else say it, I was surprised at just how useful [the session] was. I did think I understood [the report] much better.* [P07, 77 y.o. White, Female]

Related to the goal of preparing for future conversations with clinicians, patients reported that participating in ASK coaching helped them feel more confident speaking with clinicians and asking questions. The "I Have a Voice" handout, which was intended to support this, was well-liked and even named as the best part of the session by some participants. Participants noted that the additional resources provided also helped them feel more prepared for such conversations.

> *'I Have a Voice.' Absolutely fabulous. ... And in hearing [the health educator's] presentation, it gave us things to think about to ask. So, it was the most helpful part I think of the talk was that 'I Have a Voice.' Because it's true. ... And [the study surgeon] is great. He asked me some of these questions. But I just, it gave me ideas of things to ask after, when I see him again.* [P09, 70 y.o. White, Female]

> *I do plan on going to the ASK website and to the Arthritis Foundation website either this evening or tomorrow and drilling down a little bit on treatments. ... So, I just want to get more information, so in having that resource, puts me on a more equal footing with my doctor because if before my next appointment if I have more information, I can ask more intelligent questions when I see him and that could improve my treatment. So, it's been good in that way.* [P12, 73 y.o. White, Male]

Suggestions for Improvement

While participants overall described the acceptability and benefits of ASK coaching, they also offered suggestions for improvement. The most frequent request was to provide additional or more structured opportunities for patients to connect with each other inside and outside of the coaching session. Another suggestion was to create coaching groups based on patients' previous experience with joint replacement surgery or stage of disease progression. One participant suggested spreading the material across 2 sessions for better understanding. In addition, given the pragmatic challenges of scheduling the coaching session before the surgeon consultation, another participant suggested having a recorded video to accompany the "I Have a Voice" handout allowing individuals to view it in preparation for the consultation.

Discussion

Overview

This paper describes the iterative design and evaluation of an online, group coaching session offered as a companion to a PRO-based decision aid for patients with hip or knee osteoarthritis who had an initial consultation with an orthopedic surgeon. The collaboratively designed session, referred to as ASK coaching, aimed to support patients in understanding their personal health data, including PRO scores, and using the data in future conversations with clinicians. Evaluation findings highlight differences in attendance based on sociodemographic characteristics and provide insight into motivation for attendance, session format acceptability, and achievement of session goals. Implications for potential design refinement, including priorities for future research, are integrated throughout the discussion.

Principal Findings

Osteoarthritis of the knees and hips is one of the most common causes of chronic pain and disability in the United States. It currently affects more than 21 million Americans, with increasing prevalence as the population ages [33,34]. Osteoarthritis is a chronic condition for which there is no cure, and treatment options vary based on symptom severity and patient preference [35]. Therefore, patients' understanding of their symptoms and ability to communicate with providers are particularly important. In today's clinical practice, patients are accustomed to completing previsit assessments of their demographic, insurance, and health status information. However, receipt of their personal data is new to most patients, as is participating in a group coaching session to support the understanding and use of such data.

While patients routinely completed the previsit PROs, not all of them elected to participate in ASK Coaching, so this paper attempts to understand who this intervention may reach. First, patients reporting higher education and health literacy levels were more likely to join a session. This finding aligns with research showing that patients with more formal education and higher health literacy are more likely to engage in educational programming related to their arthritis care [36], a pattern that can contribute to the exacerbation of inequities if those with the greatest need for support are the least likely to access it. In contrast to high literacy and education, ASK coaching attendees reported lower confidence in managing their treatment as compared to nonattendees, suggesting they may have appropriately self-selected to attend based on their perceived need. Coaching attendees were also less likely to have Medicaid insurance than nonattendees, which may reflect a lack of access to health education or digital interventions among those with lower socioeconomic status [37]. In addition, coaching attendees were slightly older than nonattendees. While this difference in age was likely not clinically meaningful, it provides reassurance that older adults can readily and successfully attend an online session [38]. While there was no significant between-group difference for sex, patients reporting female sex tended to more often attend ASK coaching, which is consistent with existing literature showing women are more likely to seek out health

XSL•FO
RenderX

information than men [39,40]. Patients with at least one comorbidity were more likely to attend a coaching session when invited, which may indicate those managing multiple competing health priorities have a greater need for support in preparing for conversations with clinicians. Interestingly, those with slightly lower preconsultation pain were also more likely to attend, which may indicate pain can be a barrier to engaging with supportive resources [41]. Priorities for future research are identifying for whom this intervention is most beneficial, more deeply understanding barriers and facilitators to attendance in subgroups at risk of intervention-generated inequality [37], and refining the intervention design to address those subgroup needs.

A primary motivation for coaching attendance was to gain additional information. Relatedly, the most common reason for not attending was having adequate information. While not all patients may require a resource like ASK coaching to understand PRO data and engage in conversations about their osteoarthritis care, patients also may not initially recognize gaps in their own knowledge. Qualitative findings reinforced patient-perceived benefits regarding gaining information about their PRO data even when they initially reported understanding. Similarly, previous qualitative research about the ASK report highlighted misinterpretations that were not always recognized by participants [18]. Future research could therefore explore how to best communicate the potential value of participation, even to those who feel confident in their knowledge, through co-design of presession materials. A second reason study participants did not join ASK coaching was that they reported feeling too busy. Time is a common barrier when it comes to engaging participants in health promotion efforts [42,43]. While this research was conducted during the COVID-19 pandemic, a particularly challenging time for all, limited patient time supports further refinement of the session length and the development of an asynchronous option to reduce barriers to attendance.

This paper also explores the acceptability and potential effectiveness of coaching patients to use a PRO-based decision report. Among attendees, the online format, with technological support provided as needed, was generally accepted. While some preferred in-person meetings, if possible, online sessions like this one may be easier to implement as they have been shown to be more cost-effective [44], particularly for health systems or case management programs that serve a large geographic area. Most attendees also endorsed the small group nature of ASK coaching. However, qualitative feedback made it clear that this structure could be improved to maximize the session's benefits by including more unstructured time to ask questions and interact with peers. Given constraints related to clinician availability in many health systems, especially for nonreimbursable educational activities, it is worth noting participants found the nonclinician health educator to be an acceptable professional facilitator. However, adding a peer educator could benefit those interested in peer support, which has been shown to have benefits for other chronic conditions [45,46]. Based on these findings, a variety of design modifications could be evaluated in future research. To encourage participation among those who prefer in-person programs, the intervention design could be refined to improve the experience, including more scalable technological support outside of a research coordinator. Based on feedback about disparate participant experience, groups could also be organized by current treatment plans, such as surgery or nonsurgical treatments, to ensure the discussion is relevant for all. It should be noted, however, that defining groups in this way would limit attendees' ability to learn from people who were further along in the treatment process, something coaching participants appreciated. Additional options for future intervention refinement include splitting the session into multiple shorter sessions or providing information that could be viewed asynchronously before a moderated discussion, both of which would offer greater opportunities for interaction while addressing time constraints.

With regards to session goals, attendees reported that ASK coaching provided clarification of specific PRO graphs and improved overall report comprehension. Previous publications of patients' perspectives of the ASK report highlighted challenges with report comprehension, which support the role of coaching [17,18]. The session goal of preparing patients for future conversations with clinicians was also primarily met, with many patients expressing that they felt more ready to speak up and ask questions, finding the "I Have a Voice" handout particularly helpful. These positive findings related to acceptability and preliminary effectiveness were likely a function of the iterative design process, highlighting the value of collaboration with both clinical experts and patients during design to align the intervention to the users' needs.

Additional implications for design and future research include considerations related to both efficacy and scalability. The timing of the session within the parent study limited patients' ability to immediately use the information and skills from the session. When the study was designed, the return of PRO data directly to patients was not the norm in clinical care. Thus, the parent study proposed that the ASK report would be distributed in the surgeon's office enabling each clinician to review the information with their patients. The online coaching session then followed this surgeon consultation and was intended to provide a second, more detailed review of the report. While the design worked well within the established care structure, it meant that participants had a clinical consultation before attending the session. Given that PROs are more commonly included in clinical care today, and participants felt the session helped increase their confidence when communicating with clinicians, it is likely that a greater benefit would have been achieved had the session taken place prior to the initial surgeon consultation. At a minimum, participants reported that receiving information about the "I Have a Voice" handout in advance of that initial appointment would have been helpful, aligning with the principles of decision guidance—supportive elements embedded within or alongside the decision aid for independent use [22]. In addition, providing this patient preparation resource could benefit clinicians who only have a short time for each appointment by ensuring that patients can quickly share details about their arthritis and its impact on their lives [47]. Combining this handout with a prerecorded asynchronous version of the coaching session would likely also improve this intervention's scalability as it is a one-time cost and does not require additional

XSL•FO
RenderX

clinician time or onsite resources. However, this intervention format may not address all patient needs, such as being able to ask questions of a trained facilitator and interacting with other arthritis patients. In future efforts, the research team will propose a real-time return of the ASK report and coaching before the initial surgeon consultation to support patient preparation for that initial treatment discussion. Additional research in collaboration with clinicians and patient experts will further refine the design of coaching and its implementation, including asynchronous or hybrid versions that align with both patient and health system needs.

Limitations

While the parent study enrolled both English-speaking and Spanish-speaking participants, the coaching session was only offered in English, potentially limiting the generalizability of findings. Also, many participants who chose to attend the session were well-educated about arthritis and treatment options and therefore may have been less likely to benefit from participating. This limits conclusions about the intervention's effectiveness. While there were no significant racial or ethnic differences in coaching attendance, study participants were largely White, reflecting well-documented racial and ethnic disparities in referrals to orthopedists for joint replacement surgery [10]. Coaching acceptability may differ for racial or ethnic minorities and needs to be examined in future research with more diverse samples, given the potential benefits of participating in an intervention of this nature to reduce inequities in osteoarthritis care. Furthermore, patients with a strong preference for a different session format may have simply chosen to not attend, and beyond asking participants who declined to participate why they were not interested, the research team did not seek more detailed feedback. As such, it is difficult to draw generalizable conclusions about session format acceptability. ASK coaching was also part of a complex health intervention that included the PRO-based report and consultation with the surgeon. This complexity makes it challenging to parse out specific benefits of coaching. Moreover, different components of the quantitative and qualitative evaluation were pragmatically added at various points in this multi-method evaluation, limiting the ability to systematically integrate such data using a mixed methods analysis. Advanced data integration may have been helpful to provide additional insight into which patient subgroups might benefit the most from this intervention. While we designed the intervention for efficiency (eg, group visits and online coach), we did not formally evaluate the intervention's costs, which could be a barrier to broad implementation in routine care. Finally, this portion of the study took place during the COVID-19 pandemic, which may have influenced patients' perspectives on the online and group components of the coaching format.

Conclusions

This paper describes the iterative design and evaluation of an online coaching intervention delivered by a health educator to support patients in understanding and using a novel PRO-based report as part of hip or knee osteoarthritis care, including in conversations with clinicians. Among coaching attendees, the acceptability of the session format was overall high, and session goals were achieved. With some modifications related to timing and structure, this intervention appears promising to support patients in using personal health data, including PRO scores, to prepare for conversations with their clinicians. Future research is needed to further refine this intervention and scale implementation, including ensuring reach to diverse patients.

Acknowledgments

Research reported in this work was funded through a Patient-Centered Outcomes Research Institute (PCORI) Award (grant IHS-1507‐31714-IC, PDF). The statements in this work are solely the responsibility of the authors and do not necessarily represent the views of the PCORI, its Board of Governors or Methodology Committee. The authors would like to acknowledge Patty Skolnik and Citizens for Patient Safety for their input during the intervention design process, and the Arthritis Foundation for their assistance recruiting patient advisors. We also thank all patients who contributed their time and shared their experiences with us.

Data Availability

The deidentified datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Copy of the arthritis care through shared knowledge report.
[PDF File, 256 KB - formative_v9i1e65931_app1.pdf]

Multimedia Appendix 2
"I Have a Voice" handout.
[PDF File, 372 KB - formative_v9i1e65931_app2.pdf]

XSL•FO
RenderX

Multimedia Appendix 3
Additional supporting quotes for the qualitative evaluation results from interviews and free-text survey responses from a subsample of arthritis care through shared knowledge (ASK) coaching attendees in the pragmatic cluster-randomized ASK trial in patients with hip and knee osteoarthritis.
[DOCX File, 25 KB - formative_v9i1e65931_app3.docx]

References

1. Makhni EC, Hennekes ME. The use of patient-reported outcome measures in clinical practice and clinical decision making. J Am Acad Orthop Surg 2023 Oct 15;31(20):1059-1066. [doi: 10.5435/JAAOS-D-23-00040] [Medline: 37364243]
2. Dzimitrowicz HE, Blakely LJ, Jones LW, et al. Bridging new technology Into clinical practice with mobile apps, electronic patient-reported outcomes, and wearables. Am Soc Clin Oncol Educ Book 2022 Apr;42:1-6. [doi: 10.1200/EDBK_350550] [Medline: 35522912]
3. Greenhalgh J, Gooding K, Gibbons E, et al. How do patient reported outcome measures (PROMs) support clinician-patient communication and patient care? A realist synthesis. J Patient Rep Outcomes 2018 Dec;2:42. [doi: 10.1186/s41687-018-0061-6] [Medline: 30294712]
4. Porter I, Gonçalves-Bradley D, Ricci-Cabello I, et al. Framework and guidance for implementing patient-reported outcomes in clinical practice: evidence, challenges and opportunities. J Comp Eff Res 2016 Aug;5(5):507-519. [doi: 10.2217/cer-2015-0014] [Medline: 27427277]
5. Mishra VK, Hoyt RE, Wolver SE, et al. Qualitative and quantitative analysis of patients' perceptions of the patient portal experience with OpenNotes. Appl Clin Inform 2019 Jan;10(1):10-18. [doi: 10.1055/s-0038-1676588] [Medline: 30602196]
6. Volk RJ, Llewellyn-Thomas H, Stacey D, et al. Ten years of the international patient decision aid standards collaboration: evolution of the core dimensions for assessing the quality of patient decision aids. BMC Med Inform Decis Mak 2013;13 Suppl 2(Suppl 2):S1. [doi: 10.1186/1472-6947-13-S2-S1] [Medline: 24624947]
7. Smith TO, Purdy R, Lister S, et al. Living with osteoarthritis: a systematic review and meta-ethnography. Scand J Rheumatol 2014;43(6):441-452. [doi: 10.3109/03009742.2014.894569] [Medline: 24882107]
8. Staehelin D, Dolata M, Stöckli L, et al. How patient-generated data enhance patient-provider communication in chronic care: field study in design science research. JMIR Med Inform 2024 Sep 10;12:e57406. [doi: 10.2196/57406] [Medline: 39255481]
9. Shichman I, Roof M, Askew N, et al. Projections and epidemiology of primary hip and knee arthroplasty in medicare patients to 2040-2060. JB JS Open Access 2023;8(1):1. [doi: 10.2106/JBJS.OA.22.00112] [Medline: 36864906]
10. Alvarez PM, McKeon JF, Spitzer AI, et al. Race, Utilization, and Outcomes in Total Hip and Knee Arthroplasty: A Systematic Review on Health-Care Disparities. JBJS Rev 2022 Mar 1;10(3):35231001. [Medline: 35231001]
11. Pacheco-Brousseau L, Poitras S, Charette M, et al. Exploring appropriateness criteria for informing the total knee arthroplasty decision-making process: an interpretive descriptive study. J Eval Clin Pract 2023 Sep;29(6):942-954. [doi: 10.1111/jep.13893] [Medline: 37410789]
12. Ghomrawi HMK, Riddle DL, Hasan MM, et al. Incorporating expected outcomes into clinical decision-making for total knee arthroplasty. Arthritis Care Res (Hoboken) 2023 May;75(5):1132-1139. [doi: 10.1002/acr.24961] [Medline: 35638705]
13. Bansback N, Trenaman L, MacDonald KV, et al. An online individualised patient decision aid improves the quality of decisions in patients considering total knee arthroplasty in routine care: A randomized controlled trial. Osteoarthr Cartil Open 2022 Sep;4(3):100286. [doi: 10.1016/j.ocarto.2022.100286] [Medline: 36474942]
14. Jayakumar P, Moore MG, Furlough KA, et al. Comparison of an artificial intelligence-enabled patient decision aid vs educational material on decision quality, shared decision-making, patient experience, and functional outcomes in adults with knee osteoarthritis: a randomized clinical trial. JAMA Netw Open 2021 Feb 1;4(2):e2037107. [doi: 10.1001/jamanetworkopen.2020.37107] [Medline: 33599773]
15. Lin E, Uhler LM, Finley EP, et al. Incorporating patient-reported outcomes into shared decision-making in the management of patients with osteoarthritis of the knee: a hybrid effectiveness-implementation study protocol. BMJ Open 2022 Feb 21;12(2):e055933. [doi: 10.1136/bmjopen-2021-055933] [Medline: 35190439]
16. Franklin PD, Zheng H, Bond C, et al. Translating clinical and patient-reported data to tailored shared decision reports with predictive analytics for knee and hip arthritis. Qual Life Res 2021 Nov;30(11):3171-3178. [doi: 10.1007/s11136-020-02557-8] [Medline: 32562194]
17. Stern BZ, Pila S, Joseph LI, et al. Patients' perspectives on the benefits of feedback on patient-reported outcome measures in a web-based personalized decision report for hip and knee osteoarthritis. BMC Musculoskelet Disord 2022 Aug 23;23(1):806. [doi: 10.1186/s12891-022-05764-1] [Medline: 35999585]
18. Pila S, Stern BZ, Rothrock NE, et al. Evaluating a web-based personalized decision report for total knee or hip replacement: lessons learned from patients. J Eval Clin Pract 2023 Aug;29(5):844-853. [doi: 10.1111/jep.13887] [Medline: 37316454]
19. Lavallee DC, Rothrock NE, Chen AF, et al. One report, multiple aims: orthopedic surgeons vary how they use patient-reported outcomes with patients. Qual Life Res 2023 Feb;32(2):425-433. [doi: 10.1007/s11136-022-03251-7] [Medline: 36103045]
20. Frosch DL, May SG, Rendle KAS, et al. Authoritarian physicians and patients' fear of being labeled 'Difficult' among key obstacles to shared decision making. Health Aff (Millwood) 2012 May;31(5):1030-1038. [doi: 10.1377/hlthaff.2011.0576]

XSL•FO
RenderX

21. Silveira Bianchim M, Crane E, Jones A, et al. The implementation, use and impact of patient reported outcome measures in value-based healthcare programmes: a scoping review. PLoS ONE 2023;18(12):e0290976. [doi: 10.1371/journal.pone.0290976]
22. Rahn AC, Jull J, Boland L, et al. Guidance and/or decision coaching with patient decision Aids: scoping reviews to inform the International Patient Decision Aid Standards (IPDAS). Med Decis Making 2021 Oct;41(7):938-953. [doi: 10.1177/0272989X21997330] [Medline: 33759626]
23. A. f. H. R. A. quality. Questions Are the Answer. URL: https://www.ahrq.gov/questions/index.html [accessed 2025-03-25]
24. Gandek B, Roos EM, Franklin PD, et al. A 12-item short form of the Hip disability and Osteoarthritis Outcome Score (HOOS-12): tests of reliability, validity and responsiveness. Osteoarthr Cartil 2019 May;27(5):754-761. [doi: 10.1016/j.joca.2018.09.017]
25. Gandek B, Roos EM, Franklin PD, et al. A 12-item short form of the Knee injury and Osteoarthritis Outcome Score (KOOS-12): tests of reliability, validity and responsiveness. Osteoarthr Cartil 2019 May;27(5):762-770. [doi: 10.1016/j.joca.2019.01.011]
26. Selim AJ, Rogers W, Fleishman JA, et al. Updated U.S. population standard for the Veterans RAND 12-item Health Survey (VR-12). Qual Life Res 2009 Feb;18(1):43-52. [doi: 10.1007/s11136-008-9418-2] [Medline: 19051059]
27. Morris NS, MacLean CD, Chew LD, et al. The Single Item Literacy Screener: evaluation of a brief instrument to identify limited reading ability. BMC Fam Pract 2006 Mar 24;7:21. [doi: 10.1186/1471-2296-7-21] [Medline: 16563164]
28. Gruber-Baldini AL, Velozo C, Romero S, et al. Validation of the PROMIS® measures of self-efficacy for managing chronic conditions. Qual Life Res 2017 Jul;26(7):1915-1924. [doi: 10.1007/s11136-017-1527-3] [Medline: 28239781]
29. Sandelowski M. Whatever happened to qualitative description? Res Nurs Health 2000 Aug;23(4):334-340. [doi: 10.1002/1098-240x(200008)23:4<334::aid-nur9>3.0.co;2-g] [Medline: 10940958]
30. Sandelowski M. What's in a name? Qualitative description revisited. Res Nurs Health 2010 Feb;33(1):77-84. [doi: 10.1002/nur.20362] [Medline: 20014004]
31. Colorafi KJ, Evans B. Qualitative descriptive methods in health science research. HERD 2016 Jul;9(4):16-25. [doi: 10.1177/1937586715614171] [Medline: 26791375]
32. Neergaard MA, Olesen F, Andersen RS, et al. Qualitative description - the poor cousin of health research? BMC Med Res Methodol 2009 Jul 16;9:52. [doi: 10.1186/1471-2288-9-52] [Medline: 19607668]
33. Lawrence RC, Felson DT, Helmick CG, et al. Estimates of the prevalence of arthritis and other rheumatic conditions in the United States. Part II. Arthritis Rheum 2008 Jan;58(1):26-35. [doi: 10.1002/art.23176] [Medline: 18163497]
34. Palazzo C, Nguyen C, Lefevre-Colau MM, et al. Risk factors and burden of osteoarthritis. Ann Phys Rehabil Med 2016 Jun;59(3):134-138. [doi: 10.1016/j.rehab.2016.01.006] [Medline: 26904959]
35. Sharma L. Osteoarthritis of the knee. N Engl J Med 2021 Jan 7;384(1):51-59. [doi: 10.1056/NEJMcp1903768] [Medline: 33406330]
36. Duca LM, Helmick CG, Barbour KE, et al. Self-management education class attendance and health care provider counseling for physical activity among adults with arthritis - United States, 2019. MMWR Morb Mortal Wkly Rep 2021 Oct 22;70(42):1466-1471. [doi: 10.15585/mmwr.mm7042a2] [Medline: 34673750]
37. Veinot TC, Mitchell H, Ancker JS. Good intentions are not enough: how informatics interventions can worsen inequality. J Am Med Inform Assoc 2018 Aug 1;25(8):1080-1088. [doi: 10.1093/jamia/ocy052] [Medline: 29788380]
38. Mace RA, Mattos MK, Vranceanu AM. Older adults can use technology: why healthcare professionals must overcome ageism in digital health. Transl Behav Med 2022 Dec 30;12(12):1102-1105. [doi: 10.1093/tbm/ibac070] [Medline: 36073770]
39. Ek S. Gender differences in health information behaviour: a Finnish population-based survey. Health Promot Int 2015 Sep;30(3):736-745. [doi: 10.1093/heapro/dat063] [Medline: 23985248]
40. Ferraris G, Monzani D, Coppini V, et al. Barriers to and facilitators of online health information-seeking behaviours among cancer patients: a systematic review. Digit Health 2023;9:20552076231210663. [doi: 10.1177/20552076231210663] [Medline: 38107979]
41. Tomé-Pires C, Aragonès E, Rambla C, et al. Perceived barriers, facilitators and usefulness of a psychoeducational intervention for individuals with chronic musculoskeletal pain and depression in primary care. Front Psychol 2023;14:1099419. [doi: 10.3389/fpsyg.2023.1099419] [Medline: 37179874]
42. Wolin KY, Bennett GG, McNeill LH, et al. Low discretionary time as a barrier to physical activity and intervention uptake. Am J Health Behav 2008;32(6):563-569. [doi: 10.5555/ajhb.2008.32.6.563] [Medline: 18442336]
43. Strazdins L, Broom DH, Banwell C, et al. Time limits? Reflecting and responding to time barriers for healthy, active living in Australia. Health Promot Int 2011 Mar;26(1):46-54. [doi: 10.1093/heapro/daq060] [Medline: 20952445]
44. Ezeamii VC, Okobi OE, Wambai-Sani H, et al. Revolutionizing healthcare: how telemedicine is improving patient outcomes and expanding access to care. Cureus 2024 Jul;16(7):e63881. [doi: 10.7759/cureus.63881] [Medline: 39099901]
45. Rini C, Austin J, Wu LM, et al. Harnessing benefits of helping others: a randomized controlled trial testing expressive helping to address survivorship problems after hematopoietic stem cell transplant. Health Psychol 2014 Dec;33(12):1541-1551. [doi: 10.1037/hea0000024] [Medline: 24274798]
46. Chen C, Zhou Y, Xu JY, et al. Effect of peer support interventions in patients with type 2 diabetes: a systematic review. Patient Educ Couns 2024 May;122:108172. [doi: 10.1016/j.pec.2024.108172] [Medline: 38320444]

XSL•FO
RenderX

47. Gholamzadeh M, Abtahi H, Ghazisaeeidi M. Applied techniques for putting pre-visit planning in clinical practice to empower patient-centered care in the pandemic era: a systematic review and framework suggestion. BMC Health Serv Res 2021 May 13;21(1):458. [doi: 10.1186/s12913-021-06456-7] [Medline: 33985502]

Abbreviations

AHRQ: Agency for Healthcare Research and Quality
ASK: arthritis care through shared knowledge
CPS: Citizens for Patient Safety
HOOS: Hip disability and Osteoarthritis Outcome Score
KOOS-12: Knee injury and Osteoarthritis Outcome Score
PRO: Patient Report Outcome
PROMIS: Patient-Reported Outcomes Measurement Information System
VR-12: Veterans RAND 12-Item Health Survey

Edited by A Mavragani; submitted 04.09.24; peer-reviewed by A Deschamps, D Staehelin; revised version received 27.03.25; accepted 24.04.25; published 30.06.25.

Please cite as:
Burla M, Stern BZ, Berry ABL, Pila S, Franklin PD
Coaching Patients to Understand and Use Patient-Reported Outcome Data: Intervention Design and Evaluation
JMIR Form Res 2025;9:e65931
URL: https://formative.jmir.org/2025/1/e65931
doi:10.2196/65931

© Martha Burla, Brocha Z Stern, Andrew BL Berry, Sarah Pila, Patricia D Franklin. Originally published in JMIR Formative Research (https://formative.jmir.org), 30.6.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Misrepresentation of Overall and By-Gender Mortality Causes in Film Using Online, Crowd-Sourced Data: Quantitative Analysis

Calla Glavin Beauregard[1,2], MPhil; Christopher M Danforth[1,2,3,4], PhD; Peter Sheridan Dodds[1,2,5,6], PhD

[1]Computational Story Lab, University of Vermont, Burlington, VT, United States
[2]MassMutual Center of Excellence, Vermont Complex Systems Institute, University of Vermont, 89 University Place, Burlington, VT, United States
[3]Department of Mathematics and Statistics, University of Vermont, Burlington, VT, United States
[4]Vermont Advanced Computing Center, University of Vermont, Burlington, VT, United States
[5]Department of Computer Science, University of Vermont, Burlington, VT, United States
[6]Santa Fe Institute, Santa Fe, NM, United States

Corresponding Author:
Calla Glavin Beauregard, MPhil
Computational Story Lab, University of Vermont, Burlington, VT, United States

Abstract

Background: The common phrase "representation matters" asserts that media has a measurable and important impact on civic society's perception of self and others. The representation of health in media, in particular, may reflect and perpetuate a society's disease burden.

Objective: In this study, for the top 10 major causes of death in the United States, we aimed to examine how cinematic representation overall and by-gender mortality diverges from reality.

Methods: Using crowd-sourced data on over 68,000 film deaths from Cinemorgue Wiki, we employ natural language processing techniques to analyze shifts in representation of deaths in movies versus the 2021 National Vital Statistics Survey top 10 mortality causes. We parsed, stemmed, and classified each film death database entry, and then categorized film deaths by gender using a specifically trained gender text classifier.

Results: Overall, movies strongly overrepresent suicide and, to a lesser degree, accidents. In terms of gender, movies overrepresent men and underrepresent women for nearly every major mortality cause, including heart disease and cerebrovascular disease (chi-square test, $P<.001$); 73.6% (477/648) of film deaths from heart disease were men (vs 384,866/695,547, 55.4% in real life) and 69.4% (50/72) of film deaths from cerebrovascular disease were men (vs 70,852/162,890, 43.5% in real life). The 2 exceptions for which women were overrepresented are suicide and accidents (chi-square test, $P<.001$), with 39.7% (945/2382) deaths from suicide in film being women (vs 9825/48,183, 20.4% in real life) and 38.8% (485/1250) deaths from accidents in film being women (vs 75,333/225,935, 33.5% in real life).

Conclusions: We discuss the implications of under- and overrepresenting causes of death overall and by gender, as well as areas of future research.

(JMIR Form Res 2025;9:e70853) doi:10.2196/70853

KEYWORDS

media representation; gender; mortality; natural language processing; NLP; data science

Introduction

Cardiovascular disease is the number one cause of death in both men and women in the United States according to the National Survey of Vital Statistics (NVSS), published annually by the Centers for Disease Control and Prevention (CDC) [1]. However, men and women can experience different risk factors related to cardiovascular disease, as well as present different symptoms when experiencing a cardiovascular event [2]. In recent years, the American Heart Association has highlighted the importance of understanding these differences with calls for an overhaul in how physicians consider gender in the diagnosis of cardiovascular disease [3].

Representation in media is central to how we perceive our society, including health. Media portrayal and representation influence thoughts, behaviors, and actions toward certain groups, ideas, and phenomena, that is, the "[m]edia, in short, [is] central to what ultimately come to represent our social realities" [4]. This connection has been studied across many demographic categories and aspects of social reality. Broadly, narratives delivered by film and other forms of media influence beliefs, which has been theorized as a "dual empathy" pathway wherein viewers process their feelings and thoughts while also

XSL•FO
RenderX

perspective sharing with fictional characters [5]. Salient media stories in both film and books impact identity formation in both middle-age and young adults when relating their own life stories [6].

A subset of narrative theory is transportation theory, which purports that individuals who experience greater narrative transportation in response to fictional and nonfictional media are more likely to change their behaviors and beliefs and identify fewer factual inconsistencies in that media [7]. Research has explored the connection of transportation in film with a variety of health conditions and behaviors; the depiction of smoking in entertainment film increased tobacco use behaviors in adolescents [8], public awareness films improved young adult attitudes toward older adults [9], drunk driving film scenarios reduced self-risk assessment in drunk-driving [10], and a film depicting individuals with taboo chronic gut conditions improved understanding of the conditions and increased likeliness of discussing these conditions with friends and family pre- and postsurvey [11]. In summary, film depictions influence viewer perception and beliefs across areas of health.

Outside of transportation theory, there is ample research regarding the media's representation and impact on health via other pathways. A recent study examined how acute myocardial infarction (eg, a heart attack) was portrayed in 100 popular films, demonstrating the general inaccuracy of its depiction of symptoms as well as its depiction by gender [12]. Aside from that study, in terms of gender demographics, most research tends to focus on negative representation of woman in media, specifically in terms of sexual objectification; "very few studies appear to be available on the relationship between media representation and non-sexual objectification" of women [13]. Similarly, many aspects of social reality, such as economic, health, and well-being outcomes, have been studied with respect to media representation (although not explicitly through the lens of gender). For example, there is a rich history of research on how representation in the media influences the commission of intentional self-harm [14]. In summary, the media typically does not accurately portray official suicide data, and the most comprehensive systematic review on the topic links media reporting and suicidality, with current evidence suggesting increased suicidality trends with increased reporting of suicide [14].

Aside from suicidality, there is a diversity of studies on health issues and their media depiction; a nonexhaustive survey of the literature includes a study on the depiction of obesity in media and mortality in Indigenous Australians [15], a study of how bisexual individuals' mental health is impacted by depictions of bisexual people in media [16], and globally, a scoping review on how media representation of major corporations influence noncommunicable disease risk assessment and public health policy-making related to consumer products [17]. Thus, while there is research on how narrative influences health [5-8], and research on health representation and gender representation in media [18-23], there is little research that explores the overlap between health outcomes and gender representation in media, a gap we aimed to address in this study.

To examine the connection between media and mortality in this study, we downloaded counts by year of all-time cinematic deaths (referred to here as film deaths) from the crowd-sourced Cinemorgue Wiki database [24] and compared them with the 2021 NVSS [1] counts of top 10 causes of mortality by gender. Using natural language processing, we extracted the Cinemorgue counts, classified the deaths by gender, and consolidated these data against the NVSS counts. We then analyzed the comparative ratios by gender between real life and film and considered significant differences in these ratios using the chi-square test. The aim of this study is to understand the extent to which cinematic deaths in Cinemorgue reflect real-world mortality by gender and cause of death.

Methods

There are 2 main data sources for this project: the Cinemorgue Wiki database of film deaths [24] and the 2021 NVSS Leading Causes of Mortality Report [1].

Collecting and Cleaning Data

First, we identified all-time film deaths using the Cinemorgue Wiki [24] which is a crowd-sourced database that allows any members of Fandom to update death by cause by film and film year (1895 - 2023), with a short description of the death. The site is organized by death type or by actor (not by character) and each entry is a short 1 - 2 sentence description of the film death. It is inclusive of any movies publicly entered, and thus not limited to American film deaths only. Conveniently, Cinemorgue has historically provided an .xml download of their entire website (as of August 16, 2023) and other researchers have previously classified cinematic deaths using Federal Bureau of Investigation (FBI) homicide specifications [25]. Refer to Table 1 for a basic description of the database entries.

Using this existing script, we adjusted the classifying portion to identify deaths according to the CDC NVSS's top mortality causes by year for 2021 [1]. Our script parsed the description of each film and reduced it to word stems by film entry. We used generally broad word stems that captured the colloquial terms describing different mortality causes (eg, "heart attack" and "heart failure"). We then categorized film deaths by gender using a reasonably comprehensive and inclusive classifier from PyPi project gender-guesser, which was trained on a representative international database of names that were independently validated by native speakers of many languages [26]. When applying gender-guesser in this study, all 40,000 plus names were included to account for various languages and ethnic backgrounds of actors represented in film (the Cinemorgue Wiki includes movies from many different countries). Furthermore, the code allowed for the classification of mostly male, mostly female, androgynous, and unknown names. Table 2 enumerates the initial name classification.

Based on a random manual review of classification, "mostly male" and "mostly female" seemed to accurately categorize well-known actors, "mostly male" was grouped with male, and "mostly female" was grouped with female. Based on the size of the data and this classifier, we removed androgynous or unknown names (collectively referred to here as "other"),

XSL•FO
RenderX

discussed in the Limitations section. The 2021 top 10 leading causes of death in the United States are depicted in Table 3 [1]. This list actually comprises 13 leading causes of death because the ninth and tenth leading causes of death by gender differ, and a category of "other" is including to reflect the remaining causes. If absent from the official report, these data were queried directly from the CDC WONDER database [27] using the same *International Classification of Diseases, Tenth Revision* (*ICD-10*) parameters from the NVSS [1]. We calculated the ratio of mortality by gender by performing a simple ratio of male or female deaths by mortality cause over total deaths by that cause. It is important to note that the top 10 leading mortality causes differ year-to-year, and 2021 was the second year to include COVID-19 deaths. In addition, the medical terms describing these causes of death are extensive and often highly technical; for ease of understanding, we provide additional notations to specify the colloquial terms for these diseases.

Table . All film death entries by year percentiles as per Cinemorgue (total deaths=68,535).

Percentile	Year
Minimum	1895
25th	1976
50th	1995
75th	2010
Maximum	2023

Table . Results of the gender classifier, gender-guesser, used on all Cinemorgue data, before and after consolidation.

Gender	Count before, n	Count after, n
Women	15,673	17,451
Men	39,627	42,458
Unknown	7383	—[a]
Mostly men	2831	—
Mostly women	1778	—
Androgynous	1243	—
Nonmale or nonfemale[b]	—	8626

[a]Categories collapsed after consolidation.

[b]For the purposes of this table, the terms "male" and "female" are used to be consistent with the name classification package configuration.

XSL•FO
RenderX

Table . Percentage by gender of the National Vital Statistics System 2021 leading causes of mortality in the United States.

Mortality cause, type (women rank, men rank)	Total, n	Female, n (%)	Male, n (%)
Diseases of the heart (1, 1)	695,547	310,661 (44.6)	384,886 (55.4)
Malignant neoplasm, cancer[a] (2, 2)	605,213	286,543 (47.3)	318,670 (52.7)
COVID-19 (3, 3)	416,893	180,283 (43.2)	236,610 (56.8)
Accidents, unintentional injuries (6, 4)	224,935	75,333 (33.5)	149,602 (66.5)
Cerebrovascular disease, stroke[a], brain hemorrhage[a] (4, 5)	162,890	92,038 (56.5)	70,852 (43.5)
Chronic lower respiratory disease, COPD[ab], emphysema[a] (7, 6)	142,342	74,814 (52.6)	67,528 (47.4)
Alzheimer disease (5, 8)	119,399	82,424 (69)	36,975 (31)
Diabetes mellitus (8, 7)	103,294	44,666 (43.2)	56,628 (54.8)
Chronic liver disease and cirrhosis (11, 9)	56,585	20,878 (36.9)	35,707 (63.1)
Nephritis (9, 10)	54,358	25,769 (47.4)	28,589 (52.6)
Suicide, intentional self-harm (15, 8)	48,183	9825 (20.4)	38,358 (79.6)
Essential hypertension (10, 15)	42,816	22,730 (53.1)	20,086 (46.9)
Other	791,776	400,159 (50.5)	391,617 (49.5)

[a]Colloquial term.

[b]COPD: chronic obstructive pulmonary disease.

Analysis of Gender Ratios: Visualization and Chi-Square Testing

In both datasets, we calculated the ratio of mortality by gender by cause by performing a simple ratio of male death (or female death) by mortality cause over total deaths by that cause. Thus, the sum of all ratios should equal 1. From the ratios of mortality by gender in film and NVSS data, we constructed a bar graph visualization. Then, we explored the significance of the difference in gender ratio by cause of death between film and real life using the Pearson chi-square test [28,29].

It is important to underscore how to interpret significant *P* values in this context [30-32]. Accordingly, *P* values <.5 do not "measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone", and are merely "one approach to summarizing the incompatibility between a particular set of data and a proposed model for the data" [30]. That is, if there is a significant *P* value for the chi-square test for any given cause of mortality, this means that the data seem to suggest there is a significant difference between expected and observed frequencies of causes of death by gender between film and real-life but does not establish causality.

Ethical Considerations

All US mortality data used for this study are publicly available, anonymous, and consolidated at the national level. All data have been deidentified. Cinemorgue data comprise fictional, in-film deaths. This study is exempt from ethical approval, as the research involves the analysis of large-scale, publicly available datasets that are legally accessible to the public and do not require any special access or permissions. The data do not include any information through which individuals can be readily identified, either directly or indirectly. Because the data are both publicly available and do not involve interaction or intervention with human participants, nor do they involve the collection of private, identifiable information, this study qualifies for exemption from institutional review board oversight.

Results

In general, the Cinemorgue dataset underrepresents most real-life leading mortality causes as compared with cinematic mortality causes by proportion of deaths in the database. There are 2 exceptions where suicide and accidents are overrepresented. Figure 1 depicts the identity line or line of no

XSL•FO
RenderX

difference at which representation would be equal between both sources.

When considering gender in addition to raw count by database, the differences between film and real life are also apparent. Figure 2 is a subplot panel depicting the relative proportions of 2021 NVSS leading causes of mortality by gender versus all-time film deaths by gender from Cinemorgue. Across almost all causes of death except cancer and chronic liver disease, there are visible and apparent shifts between film representation and real-life representation, with women underrepresented in film across nearly all categories except for suicide. Diabetes and essential hypertension are not represented in the cleaned Cinemorgue dataset. The counts used to construct the ratio shift figure are contained in Table S2 in Multimedia Appendix 1.

In Table 4, we show the significance (if observed frequency >5) of the ratio shift for each cause of death. The ratios by gender of accidents, cerebrovascular disease, diseases of the heart, suicide (intentional self-harm), and the combined category of all other causes of mortality for film deviates significantly from that in real life. As previously mentioned, men are overrepresented in film as compared with women for each of the aforementioned causes of mortality, barring suicide.

XSL•FO
RenderX

Figure 1. Overall number of total deaths in Cinemorgue versus number of recorded deaths in the United States, National Vital Statistics Survey 2021 by cause of mortality. Of total film deaths, suicide and accidents are overrepresented; all other causes of death are underrepresented in film. The identity line depicted the boundary between over- and underrepresentation; causes of death above the line are overrepresented, and causes of death beneath the line are underrepresented in film.

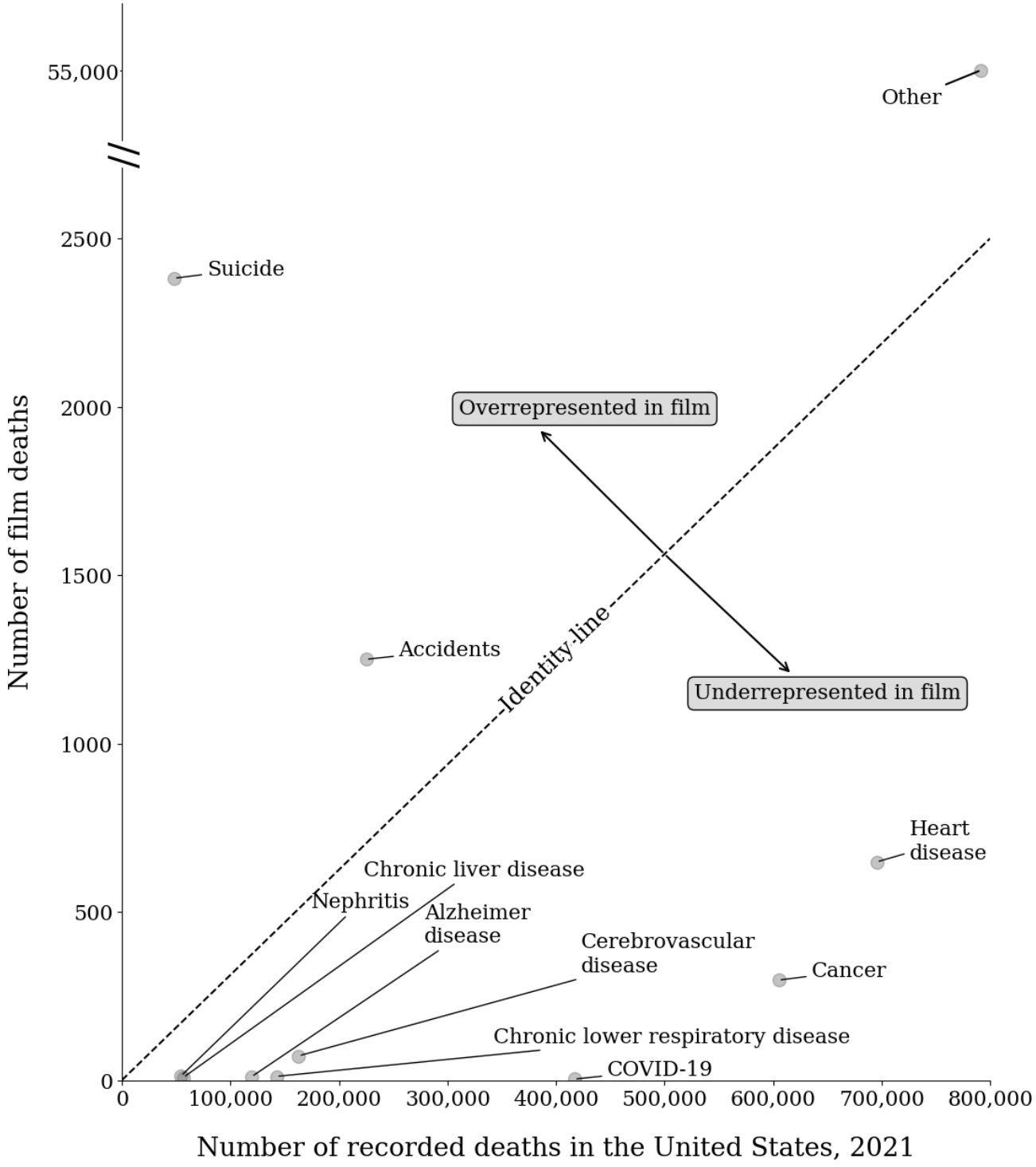

XSL•FO
RenderX

Figure 2. Top 10 mortality causes by gender in film versus real life. Count and proportion of deaths by gender (blue for men, mauve for women) are shown by mortality cause for film and for real life. Women suicide deaths and accidents are overrepresented in film. Men are overrepresented in all other causes of death. We only included causes of death present in the Cinemorgue dataset.

XSL•FO
RenderX

Table . Chi-square test: statistic and *P* values. Each cause of death and its corresponding chi-square test by gender proportion is depicted, with its significance value.

Cause of death	Chi-square (*df*)	*P* value
Accidents[a]	15.81 (1)	<.001
Alzheimer disease	—[b]	—
COVID-19	—	—
Cancer	0.29 (1)	.59
Cerebrovascular disease[a]	19.72 (1)	<.001
Chronic liver disease	—	—
Chronic lower respiratory diseases	—	—
Diseases of heart[a]	87.57 (1)	<.001
Nephritis	—	—
Other[a]	10,832.92 (1)	<.001
Suicide[a]	545.54 (1)	<.001

[a]Significant findings (where film deaths by gender are not representative of real life). There is a significant difference in representation for accidents, cerebrovascular disease, diseases of the heart, suicide, and all other causes.

[b]Cause of death categories in which there were less than 5 observations in films, and thus the chi-square test was not appropriate [25].

Discussion

Principal Findings

Film and real-life mortality differ significantly for some, but not all, leading causes of death. As expected, film is not representative of real life even setting aside gender considerations. From a practical perspective, this makes sense in the context that high-tempo, salacious, or dramatic deaths are likely far more interesting to an audience than everyday causes of mortality. In film, for example, the proportion of "other" deaths represented far outpaces the actual leading cause of death (diseases of the heart) at a factor of over 800. When considering gender, there are significant differences in the ratio of represented death for accidents, cerebrovascular disease, diseases of the heart, suicide (intentional self-harm) and all other deaths not captured using the top 10 leading causes of death classifier, among sufficiently large samples within the Cinemorgue Wiki database.

While our results do not indicate causality, they provide insight to the representation of mortality in popular media. From visualizations and statistical comparison of ratios, there is meaningful descriptive power in the ratios of film versus real-life representation of mortality by gender. In view of recent efforts by the American Heart Association to provide better equality of cardiovascular care for women [3] as well as general research regarding treatment differences by gender [33], the underrepresentation of film deaths by gender demonstrate that death is also lacking in equitable representation on screen. As there is a general consensus in the literature that stereotypes can be exceedingly harmful and contribute to sexism, harassment, and violence against women [13] as well as other gender identities [34], it is reasonable to postulate that a lack of representation in film could exacerbate the public's misunderstanding of disease burden and processes, and could hinder the recognition of gender-specific health issues among not only the public but potentially among health care professionals.

By cause of mortality, some interesting patterns emerge. For the category of "other," men are overrepresented in the Cinemorgue Wiki, which is consistent with their overrepresentation writ large in the Wiki (regardless of cause of death). This could indicate that men are represented more often on screen generally, and represented more often in violent deaths, as previously discussed. For diseases of the heart, men are seemingly overrepresented in film, which agrees with the American Heart Association's recent findings on underrepresentation and lack of understanding of women experiencing heart disease [3]. Likewise, men are overrepresented for death by cerebrovascular disease. However, suicide and accidents are overrepresented for women in film. Based on the suggestibility and risk associated with suicide representations [14], this pattern indicates that future follow-up is needed to understand the relationship between representation and acts of suicide in women.

Limitations

Limitations of this study center on the classification and cleaning of the Cinemorgue film death dataset, as well as the data itself. First, we used only film deaths in this study, on the assumption that films generally have wider ranges of audiences (not confined to certain streaming platforms) and have a similar format. However, the Cinemorgue Wiki database includes all manners of video-represented deaths, including television and video games, which were excluded in this study based on the assumption that video game deaths may overrepresent homicides, due to the violent nature of many video games. Furthermore, we compare the Cinemorgue Wiki database all-time counts with single point-in-time (2021 NVSS) mortality counts; considering that movies are not only watched in the

XSL•FO
RenderX

year in which they are filmed, we consider this a reasonable assumption even though movie popularity changes over time. In addition, in Table 1, we report that 50% (34,268/68,535) of the film deaths in our dataset are from 1995 and later, with 25% (17,132/68,535) occurring between 2010 and 2023, showing that the preponderance of films occur relatively close to the point-in-time mortality counts with many being topically contemporary to the year in which they were filmed.

In addition, our removal of androgynous names may have skewed the resultant data; as illustrated previously in Table 2, 13% of data are effectively "thrown out" of the analysis because it was not classifiable as "female" or "mostly female" or "male" or "mostly male". While using binary categories is consistent with the CDC's current reporting causes by mortality, reducing gender to solely men and women does not capture the diversity of gender identities. Considering the connection between media representation and suicide, as well as the significantly higher instances of suicidal ideation in transgender and nonbinary youth than cisgender youth [35], indicates that further research beyond this study is important for understanding the representation of health outcomes for transgender and nonbinary individuals. In terms of classifying cause of mortality in the Cinemorgue Wiki database, the search terms used on the stemmed text were specific but very broad in order to prevent an undercount of deaths, which could have skewed resultant counts. Intrinsically, the Cinemorgue Wiki database is also subject to the normal risks of crowd-sourced data. Members of the public that are heterogeneous in knowledge, skill, and geographical location comprise the crowd [36] that generates the entries in the Cinemorgue Wiki; biases in any of these characteristics could contribute to a geographically skewed or low-fidelity dataset. While CDC data from the United States were used to calculate gender proportions, every film death in the database was included, even if the film was not English-language. Accordingly, while some films may better represent certain realities than others, research has shown that even explicitly far-fetched or out-of-reality story lines can influence beliefs via transportation [37], so we have retained all films regardless of location of origin or setting.

Future Work

Future research should further detail the corpus of cinematic mortality, and incorporate additional features of each film (eg, age of actors, setting, location, audience scores, and critic reviews) from film encyclopedias and rating websites. This will further elucidate the nature of mortality representation in film relative to viewership. In addition, future work should examine mortality representation across all forms of media, including books, songs, and social media. This would provide a more global perspective on the impact of media—writ large—and capture populations that may not interact with cinematic media. With enough data, the Granger causality test—used to forecast time series from other time series [38]—could be used to understand whether media representation of death by gender lags or leads actual causes of mortality by gender. This would contribute to a better understanding of the interplay between media representation and health outcomes over time. Relatedly, visualizations like a Sankey flow diagram—which provides visualization of time-series-type data in lieu of a formal time series analysis [39]—could provide further opportunities for interpretation of how gender representation has morphed over time. Randomized controlled trials centering on these ideas (perhaps priming viewers with biased images of disease before surveying them on perceived prevalence of disease by gender) could provide support for the case of causality between representation and understanding of society's disease burden.

Conclusions

Our study compared mortality causes by gender in film and real life using Cinemorgue Wiki data on all-time film deaths and NVSS mortality by cause data from 2021 in the United States. Based on the comparison of gender ratios by cause of death, we identified significant differences in representation for accidents, cerebrovascular disease, diseases of the heart, and suicide (intentional self-harm). Considering current struggles in appropriately managing diagnosis and treatment of disease by gender, highlighting how representation of mortality is skewed by gender in films may encourage a more equitable culture surrounding gender and health. Future work should expand this query to include other forms of media to understand the interplay of representation and mortality.

Acknowledgments

The authors are grateful for support from the National Science Foundation (award #2242829) and the Massachusetts Mutual Life Insurance Company. We thank Story Lab members Ashley Fehr and Julia Zimmerman for their feedback on ideas presented in this manuscript. The authors attest that there was no use of generative artificial intelligence technology in the generation of texts, figures, or other informational content of this manuscript.

Data Availability

The datasets generated and analyzed during this study are available in the University of Vermont Story Lab repository [40].

Conflicts of Interest

None declared.

Multimedia Appendix 1
All film counts and percentages for causes of mortality in the Cinemorgue database.

XSL•FO
RenderX

[DOCX File, 24 KB - formative_v9i1e70853_app1.docx]

References

1. Curtin SC, Tejada-Vera B, Bastian BA. Deaths: leading causes for 2021. Natl Vital Stat Rep 2024;73(4) [FREE Full text]
2. Isakadze N, Mehta PK, Law K, Dolan M, Lundberg GP. Addressing the gap in physician preparedness to assess cardiovascular risk in women: a comprehensive approach to cardiovascular risk assessment in women. Curr Treat Options Cardiovasc Med 2019 Jul 29;21(9):47. [doi: 10.1007/s11936-019-0753-0] [Medline: 31359165]
3. Wenger NK, Lloyd-Jones DM, Elkind MSV, et al. Call to action for cardiovascular disease in women: epidemiology, awareness, access, and delivery of equitable health care: a Presidential Advisory from the American Heart Association. Circulation 2022 Jun 7;145(23):e1059-e1071. [doi: 10.1161/CIR.0000000000001071] [Medline: 35531777]
4. Brooks D, Hébert L. Gender, race, and media representation. In: The SAGE Handbook of Gender and Communication: SAGE Publications; 2006. [doi: 10.4135/9781412976053.n16]
5. Dill‐Shackleford KE, Vinney C, Hopper‐Losenicky K. Connecting the dots between fantasy and reality: the social psychology of our engagement with fictional narrative and its functional value. Social & Personality Psych 2016 Nov;10(11):634-646. [doi: 10.1111/spc3.12274]
6. Breen AV, McLean KC, Cairney K, McAdams DP. Movies, books, and identity: exploring the narrative ecology of the self. Qualitative Psychology 2017;4(3):243-259. [doi: 10.1037/qup0000059]
7. Green MC, Brock TC. The role of transportation in the persuasiveness of public narratives. J Pers Soc Psychol 2000;79(5):701-721. [doi: 10.1037/0022-3514.79.5.701]
8. Green MC, Clark JL. Transportation into narrative worlds: implications for entertainment media influences on tobacco use. Addiction 2013 Mar;108(3):477-484. [doi: 10.1111/j.1360-0443.2012.04088.x] [Medline: 22994374]
9. Kubrak T. Impact of films: changes in young people's attitudes after watching a movie. Behav Sci (Basel) 2020 May 2;10(5):86. [doi: 10.3390/bs10050086] [Medline: 32370280]
10. Gebbers A. Transportation into narrative worlds and the motivation to change health-related behavior. International Journal of Communication 2020;14:6977-2203. [doi: 10.31235/1234567]
11. Gomes D, Loth A, Hockley JRF, John Smith E. Gut reaction: the impact of a film on public understanding of gastrointestinal conditions. Scientific Communication and Education 2022. [doi: 10.1101/2021.09.05.459031]
12. Shaw KE, Harris KM, Diedrich A, et al. Portrayal of acute myocardial infarction in popular film: a review of gender, race, and ethnicity. J Am Heart Assoc 2024 Aug 20;13(16):e034222. [doi: 10.1161/JAHA.123.034222] [Medline: 39140282]
13. Santoniccolo F, Trombetta T, Paradiso MN, Rollè L. Gender and media representations: a review of the literature on gender stereotypes, objectification and sexualization. Int J Environ Res Public Health 2023 May 9;20(10):5770. [doi: 10.3390/ijerph20105770] [Medline: 37239498]
14. Sisask M, Värnik A. Media roles in suicide prevention: a systematic review. Int J Environ Res Public Health 2012 Jan;9(1):123-138. [doi: 10.3390/ijerph9010123] [Medline: 22470283]
15. Islam S, Fitzgerald L. Indigenous obesity in the news: a media analysis of news representation of obesity in Australia's Indigenous population. BMC Obes 2016;3:30. [doi: 10.1186/s40608-016-0109-1] [Medline: 27354919]
16. Johnson HJ. Bisexuality, mental health, and media representation. J Bisex 2016 Jul 2;16(3):378-396. [doi: 10.1080/15299716.2016.1168335]
17. Weishaar H, Dorfman L, Freudenberg N, et al. Why media representations of corporations matter for public health policy: a scoping review. BMC Public Health 2016 Aug 30;16(1):899. [doi: 10.1186/s12889-016-3594-8] [Medline: 27577053]
18. Bamman D, Samberg R, So RJ, Zhou N. Measuring diversity in Hollywood through the large-scale computational analysis of film. Proc Natl Acad Sci USA 2024 Nov 12;121(46):e2409770121. [doi: 10.1073/pnas.2409770121]
19. Bleakley A, Jamieson PE, Romer D. Trends of sexual and violent content by gender in top-grossing U.S. films, 1950-2006. J Adolesc Health 2012 Jul;51(1):73-79. [doi: 10.1016/j.jadohealth.2012.02.006] [Medline: 22727080]
20. Anderson H, Daniels M. The largest analysis of film dialogue by gender, ever. The Pudding. 2016. URL: https://pudding.cool/2017/03/film-dialogue [accessed 2024-12-12]
21. Smith SL, Choueiti M, Pieper K. Gender bias without borders: an investigation of female characters in popular films across 11 countries. : Geena Davis Institute on Gender in Media; 2014.
22. Smith SL, Choueiti M, Pieper K, Gillig T, Lee C. Inequality in 700 popular films: examining portrayals of gender, race, and LGBT status from 2007 to 2014. : USC Annenberg School for Communication and Journalism; 2015.
23. Lindner AM, Lindquist M, Arnold J. Million dollar maybe? The effect of female presence in movies on Box Office returns. Sociol Inq 2015 Aug;85(3):407-428. [doi: 10.1111/soin.12081]
24. Cinemorgue Wiki. URL: https://cinemorgue.fandom.com/wiki [accessed 2024-01-24]
25. Lindsay M. Cinemorgue parsing. Kaggle. URL: https://www.kaggle.com/code/michaellindsay/cinemorgue-parsing/notebook [accessed 2024-01-29]
26. Michael J. Gender-guesser 0.4.0: pip install gender-guesser. Python Package Index. URL: https://pypi.org/project/gender-guesser/ [accessed 2024-07-09]
27. CDC WONDER. Centers for Disease Control. URL: https://wonder.cdc.gov [accessed 2024-07-09]

XSL•FO
RenderX

28. Franke TM, Ho T, Christie CA. The chi-square test: often used and more often misinterpreted. Am J Eval 2012;33:448-458. [doi: 10.1177/1098214011426594]
29. McHugh ML. The chi-square test of independence. Biochem Med (Zagreb) 2013;23(2):143-149. [doi: 10.11613/bm.2013.018] [Medline: 23894860]
30. Wasserstein RL, Lazar NA. The ASA statement on p-values: context, process, and purpose. Am Stat 2016 Apr 2;70(2):129-133. [doi: 10.1080/00031305.2016.1154108]
31. Wasserstein RL. Moving to a world beyond "p < 0.05". Am Stat 2019;73:1-19. [doi: 10.1080/00031305.2019.1583913]
32. Greenland S, Senn SJ, Rothman KJ, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol 2016 Apr;31(4):337-350. [doi: 10.1007/s10654-016-0149-3] [Medline: 27209009]
33. Dhawan S, Bakir M, Jones E, Kilpatrick S, Merz CNB. Sex and gender medicine in physician clinical training: results of a large, single-center survey. Biol Sex Differ 2016;7(Suppl 1):37. [doi: 10.1186/s13293-016-0096-4] [Medline: 27785341]
34. Hughto JMW, Pletta D, Gordon L, Cahill S, Mimiaga MJ, Reisner SL. Negative transgender-related media messages are associated with adverse mental health outcomes in a multistate study of transgender adults. LGBT Health 2021 Jan;8(1):32-41. [doi: 10.1089/lgbt.2020.0279] [Medline: 33170060]
35. Suarez NA, Trujillo L, McKinnon II, et al. Disparities in school connectedness, unstable housing, experiences of violence, mental health, and suicidal thoughts and behaviors among transgender and cisgender high school students - Youth Risk Behavior Survey, United States, 2023. MMWR Suppl 2024 Oct 10;73(4):50-58. [doi: 10.15585/mmwr.su7304a6] [Medline: 39378210]
36. Lenart-Gansiniec R, Czakon W, Sułkowski Ł, Pocek J. Understanding crowdsourcing in science. Rev Manag Sci 2023 Nov;17(8):2797-2830. [doi: 10.1007/s11846-022-00602-z]
37. Mulligan K, Habel P. The implications of fictional media for political beliefs. American Politics Research 2013 Jan;41(1):122-146. [doi: 10.1177/1532673X12453758]
38. Shojaie A, Fox EB. Granger causality: a review and recent advances. Annu Rev Stat Appl 2022 Mar;9(1):289-319. [doi: 10.1146/annurev-statistics-040120-010930] [Medline: 37840549]
39. Otto E, Culakova E, Meng S, et al. Overview of Sankey flow diagrams: focusing on symptom trajectories in older adults with advanced cancer. J Geriatr Oncol 2022 Jun;13(5):742-746. [doi: 10.1016/j.jgo.2021.12.017] [Medline: 35000890]
40. Film deaths: misrepresentation of mortality. Computational Story Lab. URL: https://storylab.w3.uvm.edu/deaths-in-films/ [accessed 2025-05-29]

Abbreviations

CDC: Centers for Disease Control and Prevention
***ICD-10*:** *International Classification of Diseases, Tenth Revision*
NVSS: National Survey of Vital Statistics

Edited by A Mavragani; submitted 03.01.25; peer-reviewed by K Shaw, T Riffe; revised version received 10.04.25; accepted 30.04.25; published 24.06.25.

Please cite as:
Beauregard CG, Danforth CM, Dodds PS
Misrepresentation of Overall and By-Gender Mortality Causes in Film Using Online, Crowd-Sourced Data: Quantitative Analysis
JMIR Form Res 2025;9:e70853
URL: https://formative.jmir.org/2025/1/e70853
doi:10.2196/70853

© Calla Glavin Beauregard, Christopher M Danforth, Peter Sheridan Dodds. Originally published in JMIR Formative Research (https://formative.jmir.org), 24.6.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Impact of a Novel Electronic Medical Record–Integrated Electronic Form (Provider Asthma Assessment Form) and Severe Asthma Algorithm in Primary Care: Single-Center, Pre- and Postobservational Study

Matheson L McFarlane[1,2], BScH, MSc; Alison Morra[1,2], BScN, RN; Delanya Podgers[2,3], BScN, RN(EC), MN; David Barber[4,5], BSc, MD; M Diane Lougheed[1,2], MSc, MD

[1]Asthma Research Unit, Kingston Health Sciences Centre, 76 Stuart Street, Kingston, ON, Canada
[2]Department of Medicine, Division of Respirology, Queen's University, Kingston, ON, Canada
[3]Kingston General Hospital, Kingston Health Sciences Centre, Kingston, ON, Canada
[4]Canadian Primary Care Sentinel Surveillance Network, Kingston, ON, Canada
[5]Department of Family Medicine, Queen's University, Kingston, ON, Canada

Corresponding Author:
Matheson L McFarlane, BScH, MSc
Asthma Research Unit, Kingston Health Sciences Centre, 76 Stuart Street, Kingston, ON, Canada

Abstract

Background: Despite national asthma care guidelines, care gaps persist between best-practice and clinical practice, contributing to poor health outcomes. The Provider Asthma Assessment Form (PAAF) is an electronic asthma management and Knowledge Translation tool with an embedded decision support algorithm for severe and/or uncontrolled asthma, designed to support evidence-based asthma management.

Objective: In this study, we aimed to document baseline asthma practice patterns and determine whether the broader intervention of PAAF integration into a primary care electronic medical record (EMR) improves evidence-based asthma diagnosis and management.

Methods: We performed a single-center pre- and postobservational study at an academic Family Health Team in Kingston, Ontario, Canada. Retrospective baseline data were collected for 2 years prior to PAAF implementation from January 2018 to December 2019. Prospective postintervention data were collected from October 2022 to July 2024. A validated adult asthma EMR case definition was applied to EMR data to identify suspected or objectively confirmed asthma cases for both datasets, on which detailed manual chart abstractions were performed. A data extraction was performed for completed PAAFs.

Results: There were 230 patients in the retrospective baseline and 143 patients in the postimplementation cohort. Overall, 31.3% (n=72) of patients at baseline versus 23.8% (n=34) at postimplementation had confirmed asthma. There were significantly more pulmonary function tests requested after the implementation of the PAAF (postimplementation: n=70, 49%; baseline: n=71, 30.9%; $P<.001$). A significantly higher percent of postimplementation patients were on single inhaler controller and reliever therapy (postimplementation: n=31, 21.7%; baseline: n=2, 0.9%; $P<.001$), inhaled corticosteroid/long-acting β-2 agonist therapy (postimplementation: n=36, 25.2%; baseline: n=34, 14.8%; $P=.01$), and inhaled corticosteroid if their asthma was uncontrolled (postimplementation: n=69, 62.2%; baseline: n=100, 43.5%; $P=.002$). Barriers were significantly more commonly addressed after implementation (postimplementation: n=24, 16.8%; baseline: n=11, 4.8%; $P<.001$). A significantly higher average number of asthma control parameters was documented when the PAAF was used (PAAF: mean 5.4, SD 1.9; manual chart abstraction: mean 2.3, SD 1.2; $P<.001$). Care as assessed by key Primary Care—Asthma Performance Indicators showed improvement in the postimplementation cohort, which did not reach statistical significance.

Conclusions: The multifaceted intervention of implementing the PAAF in this primary care practice was associated with improved documentation of diagnosis status and asthma control parameters and improved adherence with evidence-based recommendations for care, such as the use of pulmonary function tests and addressing barriers to effective asthma management. However, uptake was low, and key asthma care gaps were still common. Future directions should involve evaluating the impact of the PAAF on care and outcomes after widespread implementation in primary care settings and investigating methods to increase user uptake of the PAAF.

(JMIR Form Res 2025;9:e74043) doi:10.2196/74043

XSL•FO
RenderX

KEYWORDS

asthma; severe asthma; diagnosis; pulmonary function tests; guideline implementation; digital tools; electronic forms; best practice; knowledge translation

Introduction

Background

Asthma is a common chronic respiratory disease contributing to a significant health care burden on both the individual and society [1,2]. It is defined as an inflammatory condition of the airways, characterized by episodic or persistent symptoms, such as dyspnea, wheezing, and cough, and is associated with variable airflow limitation and airway hyperresponsiveness [3]. Recommendations for best-practice asthma diagnosis and management are disseminated to primary care providers through the publication of clinical guidelines. In Canada, the Canadian Thoracic Society (CTS) evidence-based guidelines state that best-practice diagnosis is based on a compatible clinical history and objective evidence of asthma or a specialist diagnosis [3]. Effective management includes assessment of asthma control, patient self-management education, identifying triggers and environmental controls, and appropriate pharmacological therapy [3].

However, difficulties integrating clinical guidelines into the process of care have contributed to a number of key asthma care gaps in the primary care setting, causing a separation between best-practice and current clinical practice [3-7]. Several key asthma care gaps have been identified, primarily concerning objective diagnosis, continued monitoring with pulmonary function tests (PFTs), addressing asthma control, and delivering patient self-management education [5,8-11]. Furthermore, individuals with severe asthma (SA; 5% - 10% of patients with asthma) account for a disproportionate 50% of asthma-related health care costs [6]. Hence, in a recent position statement, the CTS highlighted a greater need for adequate recognition and management of individuals with SA and/or uncontrolled asthma, including specialist referral and consideration of biologic therapies when appropriate [6].

Consequently, the introduction of Knowledge Translation (KT) initiatives and tools to aid in adopting clinical guidelines in primary care practice is warranted [4]. Electronic medical records (EMRs) provide a unique opportunity at the point-of-care to integrate novel electronic tools (e-Tools), as they facilitate sentinel surveillance, performance benchmarking, and quality improvement [12]. Several validated and beneficial e-Tools have been researched. However, limitations exist with their use, including limited user uptake beyond the scope of research, the absence of standardized data elements, and limited involvement of end users throughout the design process [12-16].

An e-Tool designed for primary care EMRs that addresses the limitations of current e-Tools may have a significant impact on primary care asthma diagnosis and management and prompt adherence to best-practice guidelines. The Provider Asthma Assessment Form (PAAF) is a novel, point-of-care asthma management tool with an embedded SA algorithm and clinical decision support [17]. It was adapted from the Lung Health Foundation's Asthma Care Map for primary care (paper tool), is congruent with current CTS guidelines, and incorporates Primary Care—Asthma Performance Indicators (PC-APIs) and Pan-Canadian Respiratory Standards Initiative for Electronic Health Records (PRESTINE) standardized asthma data elements [3,12,17,18]. It was developed by the Asthma Research Unit at Queen's University in collaboration with primary care providers and has been integrated into the OSCAR primary care EMR at an academic Family Health Team (FHT) in Kingston, Ontario. The ultimate aim of this initiative is to address the major key asthma care gaps in primary care and promote adherence to best-practice guidelines through the introduction of a novel e-Tool to aid primary care providers in their practice.

Objectives

We conducted a comprehensive evaluation of the integration of the PAAF into a primary care EMR to justify the widespread implementation of this form in primary care EMRs and other practice settings. Principally, we aimed to assess the ability of the PAAF's implementation to address several key asthma care gaps in asthma diagnosis, management, and surveillance: underuse of objective measures of lung function to diagnose asthma [19-21], inconsistent documentation of asthma diagnosis in primary care EMRs [22], and underrecognition and suboptimal management of SA and/or uncontrolled asthma [6]. The primary objective was to evaluate the impact of the PAAF's implementation on the use of objective lung function measures to confirm asthma diagnoses, documentation in the EMR of asthma diagnosis status (suspected, confirmed, or excluded), and referral of individuals with suspected or confirmed SA to a specialist.

Methods

Ethical Considerations

The study received ethics approval from the Queen's University Health Sciences & Affiliated Teaching Hospitals' Research Ethics Board (HSREB #6036789). This study involves using existing data from the Canadian Primary Care Sentinel Surveillance Network (CPCSSN) project (HSREB# 19809). CPCSSN has a data-sharing agreement with the center where this study was conducted. Passive consent was used, with all patients having an opportunity to explicitly opt out of the study. All data were deidentified.

Study Design

A pre- and postobservational study was conducted at an FHT in Kingston, Ontario. The study design consisted of 2 cohorts of patients: a retrospective baseline cohort, consisting of patients who had an asthma visit before implementation of the PAAF, and a postimplementation cohort, consisting of patients with a visit for asthma after integration of the PAAF into the OSCAR EMR at an FHT. An asthma visit was defined as an FHT encounter where asthma was suspected, diagnosed, managed, or treated either at an in-person or telephone appointment. Case identification was accomplished by applying the newly validated

XSL•FO
RenderX

adult asthma case definition for primary care EMRs to CPCSSN-deidentified EMR data for the FHT [22]. This case definition identified adult patients (≥18 years of age) with either suspected or confirmed asthma and a documented encounter during the defined study period. This validated case definition has a sensitivity of 81% and specificity of 96% and is based on *International Classification of Diseases, Ninth Revision* (*ICD-9*) codes for asthma billing [22]. A single abstractor completed detailed manual chart abstraction on identified cases. A programmed data extraction from the FHT EMR was subsequently performed for all PAAFs that were used within the study period.

The retrospective baseline cohort included all asthma visits in a 2-year period between January 2018 and December 2019 prior to the COVID-19 pandemic to ensure that baseline data were representative of usual care. The postimplementation period was set as January 2023 to December 2023 (1 calendar year) for manual chart abstractions. Due to barriers faced with implementation (technological difficulties with the form's functionality in the EMR from May to July 2023) and challenges with user uptake, the period for completed PAAF data extraction was delayed until October 2022 to July 2024. The sample derivation methodology is summarized in Figure 1.

Figure 1. Patient cohort identification methodology. FHT: Family Health Team; PAAF: Provider Asthma Assessment Form.

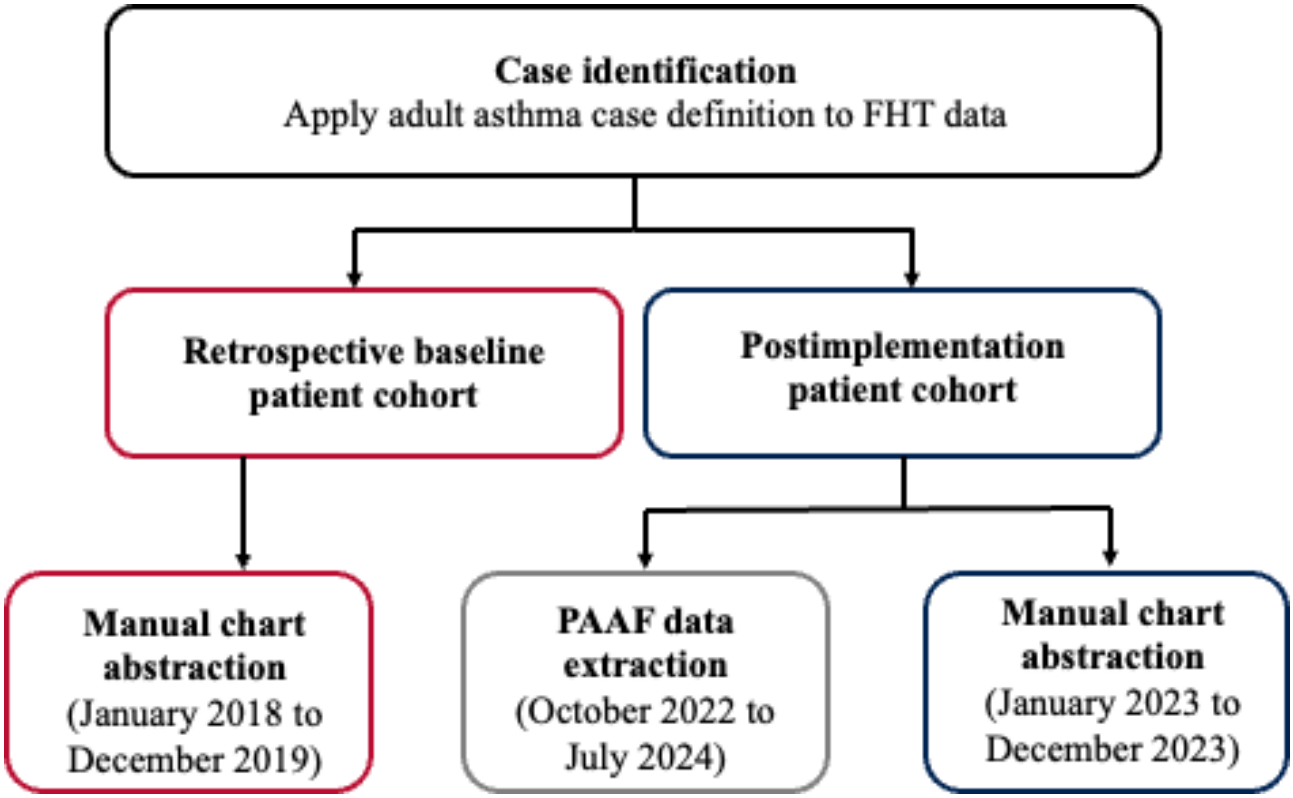

Intervention and Implementation

The PAAF is a web-based electronic form with direct access from the patient dashboard in the OSCAR EMR at the FHT. Patient demographics such as date of birth and chart number are automatically populated when the electronic form is opened in a patient chart. It contains 12 sections pertaining to the diagnosis, family history, smoking history, asthma severity, occupational history, respiratory medications, asthma control, care, management and referrals, asthma action plan, asthma control zone, pulmonary function tests, and assessment tools. Select sections include interactive features, such as a calculation of smoking pack-years and a determination of "yes" or "no" if the patient's asthma is controlled, based on the CTS control criteria. The SA algorithm feature consists of a decision-support prompt appearing at the top of the form once relevant sections have been completed. Decision support prompts include suggesting confirmation of an asthma diagnosis through objective measures, tapering inhaled corticosteroid (ICS) medication if the patient's asthma is in control, addressing reasons for poorly controlled asthma, and considering a referral to a specialist for SA and/or uncontrolled asthma.

Various methods of implementation were used throughout the PAAF study period from October 2022 to July 2024. Initial implementation was conducted by a site champion, a first-year family medicine resident at the FHT. The PAAF was initially introduced by the site champion to attending and resident physicians through a presentation and via email. Subsequent implementation efforts throughout the course of the study period included a reminder poster in the FHT clinic team rooms about the PAAF and how to access it in OSCAR.

A summary email and an attached 1-page handout detailing the purpose of the PAAF and how to use it were also sent to all FHT providers via email twice throughout the study period. Additionally, the study team collaborated with a nursing student community-based project group. The nursing students were given a summary presentation about asthma, the key asthma care gaps, and how the PAAF can be used in clinical practice. They participated in a tutorial and live demonstration of the

XSL•FO
RenderX

PAAF within the OSCAR EMR. The students then used the form during telephone appointments of patients with asthma with supervision from an instructor.

Finally, implementation strategies for the PAAF included educational presentations at Family Medicine Grand Rounds. In both July 2023 and July 2024, the study team gave presentations to incoming first-year family medicine residents during their orientation period. Additional presentations were given throughout the year to upper-year residents on rotation at the FHT.

Manual Chart Abstraction and Data Extraction

Overview

Detailed manual chart abstractions were completed by a single abstractor for adult asthma cases identified in both study cohorts. Chart abstractions were completed using a PAAF manual chart abstraction form consisting of all data elements within the PAAF to ensure consistency and accuracy of data entry. In total, 140 data elements were collected from patient charts. Chart abstractions were internally validated by a registered nurse knowledgeable in asthma management, with guidance from a respirologist. All data collected were deidentified. Data elements included patient demographics, comorbidities, classification of diagnosis status, asthma control, health services use, respiratory medications, smoking status, and care, management and education. Medication classification and ICS dosage rating (low, medium, or high dose) were guided by the Lung Health Foundation's respiratory medication reference document and CTS dosing categories [3]. Additionally, a programed data extraction was performed for all PAAFs partially or fully completed throughout the study duration.

Classification of Confirmed Versus Suspected Asthma

Each patient's asthma diagnosis was classified as confirmed, suspected, or excluded based on whether or not PFTs were ever present in the patient EMR chart, which met the criteria for asthma based on CTS guidelines [3]. Suspected asthma was defined as a compatible clinical history without objective evidence of asthma. Confirmed asthma included a compatible clinical history and objective evidence of asthma on PFTs or a specialist diagnosis (ie, respirologist and allergist). A specialist diagnosis was accepted based on the CTS Severe Asthma Position Statement recommending referral to a specialist when asthma is difficult to confirm based on objective measures (eg, due to the presence of fixed obstruction) [6] and the CTS' PRESTINE expert Delphi panel [17].

Although an empiric trial of controller therapy can be used in the diagnosis of asthma, particularly when access to objective testing is limited or unavailable (eg, during the COVID-19 pandemic), the Choosing Wisely Canada recommendations (consolidated and written by the CTS) emphasize to not continue medications for asthma in individuals with no clear clinical benefit or without confirmation of reversible airflow limitation or airway hyperresponsiveness with objective testing [23]. Following these recommendations, there was insufficient documentation of subjective or objective response to empiric controller therapy within the primary care EMR to support response to empiric therapy as objective proof of confirmed asthma in our study.

Use of PFTs for Diagnosis and Monitoring

To determine the appropriate use of objective lung function measures for diagnosis and monitoring, key recommendations were used from CTS guidelines. Within the current guidelines, objective confirmation of diagnosis status is recommended for all new asthma diagnoses in patients aged 6 years and older [3]. Specific criteria are reversible airflow obstruction on spirometry, variable airflow obstruction over time, or evidence of airway hyperresponsiveness on a challenge test such as methacholine or exercise-induced asthma challenge [3]. Additionally, a key parameter for achieving CTS guideline adherence includes measuring lung function (spirometry or peak expiratory flow) at follow-up visits, as part of ongoing assessment of asthma control [3,24].

Statistical Analysis

Overview

Descriptive statistics, including frequency and percentages, were performed on data elements extracted from manual chart abstraction and PAAF extraction. Univariate inferential statistics were performed on validated primary care performance indicators and data elements pertaining to asthma diagnosis and management [18]. Chi-square tests were performed on categorical variables, and unpaired 2-tailed *t* tests were performed on continuous numerical variables. Statistical significance was set as $P<.05$ for all analyses. All statistical analysis was performed on Microsoft Excel and GraphPad PRISM software.

Sample Size Calculation

To ensure the statistical power of primary outcomes, 2 sample size calculations were performed. For the use of objective lung function measures to confirm an asthma diagnosis, to achieve a power of 80%, at a significance of .05, from a baseline 25% of patients with objectively confirmed asthma [22], and a clinically meaningful effect of 20%, we needed 67 patients in the baseline cohort and 134 in the postimplementation cohort. For the documentation of diagnosis status in EMRs, to achieve a power of 80%, at a significance of .05, from a baseline 10% of patients with documentation diagnosis status [22], and a clinically meaningful effect of 20%, we needed 48 patients in the baseline cohort and 96 in the postimplementation cohort. Therefore, we aimed for over 67 patients in the baseline cohort and 134 in the postimplementation to ensure adequate statistical power.

Results

Sample Derivation and Characteristics

A total of 373 patients were deemed eligible in the final analysis of this study. The sample derivation is displayed in Figure 2. In the retrospective baseline patient cohort, 230 patients were included. In the postimplementation patient cohort, a total of 143 patients were included, of whom the PAAF was used in 12. A total of 27 patients were excluded from manual chart abstraction (retrospective baseline: n=19 and

XSL•FO
RenderX

postimplementation: n=8), as they did not meet eligibility for having an asthma visit within the study period. A total of 6 patients (6 PAAFs) were excluded, as they were identified as either a pediatric patient or a test patient chart (ie, a chart generated for training and programming purposes). There were 7 new patients from the PAAF data extraction, and 5 patients overlapped with the postimplementation chart abstraction.

Table 1 displays the patient characteristics of included patients from the retrospective baseline and postimplementation. No significant differences were found between the 2 groups, except for a significantly higher percentage of patients diagnosed with diabetes in the postimplementation cohort (retrospective baseline: n=14, 6.1% vs postimplementation: n=18, 12.6%; *P*=.04). Both the mean age (retrospective baseline: mean 46.9, SD 17.4; postimplementation: mean 49.7, SD 17.5; *P*=.14) and BMI (kg/m^2) were similar between groups (retrospective baseline: mean 30.0, SD 8.4; postimplementation: mean 31.3, SD 7.7; *P*=.14). There was a high percentage of female patients compared to male patients (sex assigned at birth) in both cohorts (retrospective baseline: n=135, 58.7% female vs postimplementation: n=90, 62.9% female; *P*=.45). There was an observed high percentage of comorbidities with both groups having over 45% of patients with ≥4 comorbidities (retrospective baseline: n=107, 46.5% vs postimplementation: n=65, 45.5%; *P*=.60). The most common comorbidity among both groups was allergies (retrospective baseline: n=187, 81.3% vs postimplementation: n=113, 79%; *P*=.59).

Figure 2. Sample derivation. FHT: Family Health Team; PAAF: Provider Asthma Assessment Form.

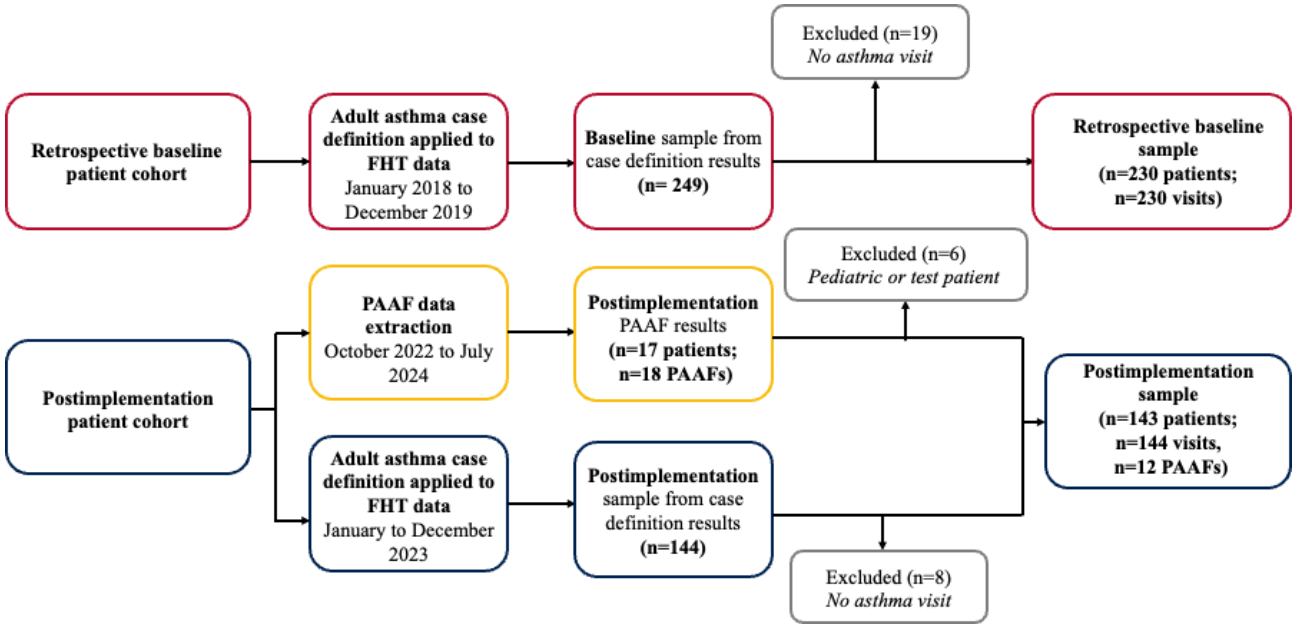

XSL•FO
RenderX

Table . Patient demographics and comorbidities.

Characteristic	Retrospective baseline (n=230)	Postimplementation (n=143)	*P* value
Age (years), mean (SD)	46.9 (17.4)	49.7 (17.5)	.14
Sex assigned at birth, n (%)			.45
Male	95 (41.3)	53 (37.1)	
Female	135 (58.7)	90 (62.9)	
BMI (kg/m^2), mean (SD)	30 (8.4)	31.3 (7.7)	.13
Number of comorbidities, n (%)			.60
0	9 (3.9)	5 (3.5)	
1	29 (12.6)	12 (8.4)	
2	29 (12.6)	23 (16.1)	
3	56 (24.3)	38 (26.6)	
≥4	107 (46.5)	65 (45.5)	
Comorbidities, n (%)			
Allergic rhinoconjunctivitis	118 (51.3)	87 (60.8)	.09
Allergies	187 (81.3)	113 (79)	.59
Anaphylaxis	30 (13)	14 (9.8)	.41
Atopic dermatitis	13 (2.2)	7 (4.9)	.82
Asthma-COPD[a] overlap	5 (2.2)	7 (4.9)	.23
COPD	6 (2.6)	3 (2.1)	.99
Diabetes	14 (6.1)	18 (12.6)	.04[b]
Eczema or hives or uticaria	33 (14.3)	12 (8.4)	.10
Obesity	95 (41.3)	58 (40.6)	.91
GERD[c]	68 (29.6)	46 (32.2)	.64
Psychopathologies	103 (44.6)	69 (48.2)	.67
OSA[d]	52 (22.6)	31 (21.7)	.90
Rhinitis or nasal polyposis or sinusitis	26 (11.3)	18 (12.6)	.74

[a]COPD: chronic obstructive pulmonary disease.
[b]*P*<.05.
[c]GERD: gastroesophageal reflux disease.
[d]OSA: obstructive sleep apnea.

Diagnosis Status and Monitoring With PFTs

In total, only 31.3% (n=72) of patients in the retrospective baseline and 23.8% (n=34) in the postimplementation cohorts were classified as having confirmed asthma (Table 2). The most common method of diagnosis in both groups was evidence of reversible airflow obstruction on spirometry (retrospective baseline: n=36, 50% vs postimplementation: n=21, 61.8%). The second most common method was a positive or borderline methacholine challenge test (retrospective baseline: n=29, 40.3% vs postimplementation: n=6, 17.6%). A higher proportion of confirmed patients with asthma in the postimplementation cohort had their asthma diagnosed by a specialist (respirologist, allergist, or internal medicine) without documented evidence of objective measures (postimplementation: n=4, 11.8% vs retrospective baseline: n=7, 9.7%). A visual representation of suspected versus confirmed asthma can be appreciated in Multimedia Appendix 1.

Table 3 displays results pertaining to monitoring with PFTs. The most noteworthy finding was that PFTs were requested by physicians in a significantly higher proportion of patients in the postimplementation period (postimplementation: n=70, 49% vs retrospective baseline: n=71, 30.9%; *P*<.001). A nonsignificant higher proportion of patients in the retrospective baseline cohort had been monitored with PFTs within 12 months before the documented encounter (retrospective baseline: n=67, 29.1% vs postimplementation: n=28, 19.6%; *P*=.05). However, a nonsignificantly higher proportion of patients had been monitored with PFTs either 12 months before or after the documented encounter (postimplementation: n=50, 35% vs retrospective baseline: n=71, 30.9%; *P*=.43). A visual

XSL•FO
RenderX

representation of PFTs completed versus PFTs requested by providers can be appreciated in Multimedia Appendix 2.

Table . Diagnosis status.

Diagnosis status	Retrospective baseline (n=230), n (%)	Postimplementation (n=143), n (%)	*P* value
Suspected	158 (68.7)	109 (76.2)	.17
Confirmed	72 (31.3)	34 (23.8)	.17
Reversible airflow obstruction (spirometry)[a]	36 (50)	21 (61.8)	
Positive or borderline methacholine challenge[a]	29 (40.3)	6 (17.6)	
Positive exercise challenge[a]	0 (0)	2 (5.9)	
Peak expiratory flow variability[a]	0 (0)	1 (2.9)	
Specialist diagnosis alone[a]	7 (9.7)	4 (11.8)	
Excluded	1 (0.4)	0 (0)	.17

[a]Percentages based on number of confirmed patients.

Table . Pulmonary function testing.

Pulmonary function testing	Retrospective baseline (n=230), n (%)	Postimplementation (n=143), n (%)	*P* value
PFTs[a] requested in the last 12 months	71 (30.9)	70 (49)	<.001[b]
Monitored with PFTs in the last 12 months	67 (29.1)	28 (19.6)	.05
Monitored with PFTs within 12 months before or after documented visit	71 (30.9)	50 (35)	.43

[a]PFT: pulmonary function test.

[b]*P* value with significance <.05.

Medications

Table 4 summarizes the asthma medications documented in the charts of patients at the time of the asthma visit included in the study. Two primary significant differences were found. In the postimplementation group, 21.7% (n=31) of patients were on single inhaler reliever and controller therapy (single inhaler therapy [SIT]) compared to 0.9% (n=2) on SIT in the retrospective baseline cohort (*P*<.001). Similarly, a significantly greater percentage of patients (n=36, 25.2%) in the postimplementation groups were on ICS/long-acting β-2 agonist (LABA) controller therapy compared to the retrospective baseline (n=34, 14.8%; *P*=.01). No other significant differences were found between groups. Nonsignificant trends include a 5.6% decrease in systemic steroid use and a 5% decrease in adherence issues after implementation. A summary of reliever, ICS, second controller, systemic steroid use, and SIT can be found in Multimedia Appendix 3.

XSL•FO
RenderX

Table . Medications.

Medication	Retrospective baseline (n=230), n (%)	Postimplementation (n=143), n (%)	*P* value
Reliever	186 (80.9)	125 (87.4)	.12
SABA[a]	182 (79.1)	116 (81.1)	.69
SAMA[b]	12 (5.2)	4 (2.8)	.31
Controller			
ICS[c]	128 (55.7)	92 (64.4)	.11
Single inhaler therapy	2 (0.9)	31 (21.7)	<.001[d]
Second controller	81 (35.2)	63 (44.1)	.10
ICS/LABA[e]	34 (14.8)	36 (25.2)	.01[d]
ICS/LAMA[f]/LABA	0 (0)	2 (1.4)	.15
LAMA	12 (5.2)	5 (3.5)	.61
LABA	3 (1.3)	2 (1.4)	>.99
LTRA[g]	23 (10)	9 (6.3)	.26
ICS dosage[h]			.49
Low	57 (44.5)	50 (54.3)	
Medium	49 (38.3)	32 (34.8)	
High	26 (20.3)	16 (17.4)	
Biologic	4 (1.7)	2 (1.4)	>.99
Omalizumab[i]	4 (100)	1 (50)	
Tezepelumab[i]	0 (0)	1 (50)	
Systemic corticosteroids used within the last year	58 (25.2)	28 (19.6)	.26
Adherence issues known or suspected	47 (20.4)	22 (15.4)	.27
Asthma medication prescribed or changed at visit	130 (56.5)	78 (54.5)	.75

[a]SABA: short-acting β-2 agonist.
[b]SAMA: short-acting muscarinic antagonist.
[c]ICS: inhaled corticosteroid.
[d]*P* values with significance <.05.
[e]LABA: long-acting β-2 agonist.
[f]LAMA: long-acting muscarinic antagonist.
[g]LTRA: leukotriene receptor antagonist.
[h]ICS percentages based on a number of patients on an ICS.
[i]Biologic percentages based on a number of patients on a biologic.

Asthma Control

At least 1 CTS control parameter was documented in a high percentage of patients in both groups (retrospective baseline: n=226, 98.3% vs postimplementation: n=138, 96.5%; *P*=.31; Table 5). A significantly higher percentage of postimplementation patients were on an ICS if their asthma was uncontrolled (postimplementation: n=69, 62.2% vs retrospective baseline: n=100, 43.5%; *P*=.002). Furthermore, a significantly higher number of CTS asthma control parameters were documented after implementation (postimplementation: mean 2.5, SD 1.3 vs retrospective baseline: mean 2.2, SD 1.2; *P*=.02). A summary of asthma control documentation is represented in Multimedia Appendix 4.

XSL•FO
RenderX

Table . Asthma control: baseline versus postimplementation.

Asthma control	Retrospective baseline (n=230)	Postimplementation (n=143)	*P* value
≥1 Asthma control parameter documented, n (%)	226 (98.3)	138 (96.5)	.31
Asthma controlled at visit, n (%)	30 (13)	28 (19.6)	.11
Asthma uncontrolled at visit, n (%)	195 (84.8)	111 (77.6)	.10
Uncontrolled asthma on an ICS[a,b]	100 (43.5)	69 (62.2)	.002[c]
Number of asthma control parameters documented, mean (SD)	2.2 (1.2)	2.5 (1.3)	.02[c]

[a]Uncontrolled asthma on ICS percentages and *P* value based on a total number of patients with uncontrolled asthma.

[b]ICS: inhaled corticosteroid.

[c]*P* values with significance <.05.

Health Services Use, Asthma Management, and PC-APIs

Table 6 summarizes the health services use and management between cohorts. A significantly higher proportion of barriers were addressed during the documented asthma visit in the postimplementation group compared to the baseline (postimplementation: n=24, 16.8% vs retrospective baseline: n=11, 4.8%; *P*<.001). No other significant differences were found. Although not statistically significant, the proportion of recent emergency department visits in <1 year decreased by 4% in the postimplementation group (postimplementation: n=16, 11.2% vs retrospective baseline: n=35, 15.2%; *P*=.21). An Asthma Action Plan (AAP) was provided or revised or reviewed more often (4.1% more) in the postimplementation period (*P*=.15). Several PC-APIs from both cohorts are represented in Table 7.

XSL•FO
RenderX

Table . Health services use and asthma management.

Health services use and asthma management	Retrospective baseline (n=230), n (%)	Postimplementation (n=143), n (%)	*P* value
Health services use			
ED[a] ever for asthma	75 (32.6)	50 (35)	.65
Recent ED for asthma (<1 year)	35 (15.2)	16 (11.2)	.21
Hospitalized ever for asthma	33 (14.3)	21 (14.7)	>.99
Recent hospitalization for asthma (<1 year)	6 (2.6)	0 (0)	.09
Specialist involved in care (respirologist, internal medicine, or allergist)	64 (27.8)	36 (25.2)	.63
Exacerbation<1 year	109 (47.4)	60 (42)	.34
Smoking status			
Current	34 (14.8)	21 (14.7)	.98
Smoking cessation addressed[b]	19 (55.9)	14 (66.7)	.57
Past	63 (27.4)	41 (28.7)	.98
Care, management, and education			
Device technique assessed	11 (4.8)	8 (5.6)	.81
Triggers addressed	176 (76.5)	109 (76.2)	>.99
Environmental controls discussed	29 (12.6)	12 (8.4)	.24
Barriers addressed	11 (4.8)	24 (16.8)	<.001
Referred to an asthma education program or CRE[c]	8 (3.5)	8 (5.6)	.43
Primary care physician suspected severe asthma	6 (2.6)	7 (4.9)	.26
Referral to a specialist for suspected severe asthma	4 (1.4)	5 (3.5)	.31
Abstractor suspected severe asthma[d]	10 (4.3)	10 (7)	.34
Asthma Action Plan provided or revised or reviewed	13 (5.7)	14 (9.8)	.15

[a]ED: emergency department.

[b]Smoking cessation addressed percentages based on a number of current smokers.

[c]CRE: Certified Respiratory Educator.

[d]Abstractor suspected severe asthma based on Canadian Thoracic Society criteria and the severe asthma algorithm.

XSL•FO
RenderX

Table . Primary Care—Asthma Performance Indicators.

Primary Care—Asthma Performance Indicators	Retrospective baseline (n=230), n (%)	Postimplementation (n=143 patients), n (%)	*P* value
Referrals to an asthma education program or CRE[a]	8 (3.5)	8 (5.6)	.43
Pulmonary function monitoring in the last 12 months	71 (30.9)	50 (35)	.43
Patients with uncontrolled asthma on an ICS[b]	100 (43.5)	69 (62.2)	.002
Well-controlled asthma	30 (13)	28 (19.6)	.11
ED[c] visits for asthma <1 year	35 (15.2)	16(11.2)	.21
Objectively confirmed asthma	72 (31.3)	34 (23.8)	.12
Smoking cessation addressed	19 (8.3)	14 (9.8)	.57
Asthma exacerbations	109 (47.4)	60 (42)	.34
Inhaler technique monitoring	11 (4.8)	8 (5.6)	.81
Asthma Action Plan	3 (15.7)	14 (9.8)	.15

[a]CRE: Certified Respiratory Educator.

[b]ICS: inhaled corticosteroid.

[c]ED: emergency department.

Use of the PAAF Versus Manual Chart Abstraction

Overall, 12 PAAFs were used during asthma encounters with 11 patients total in the postimplementation period. When the PAAF was used, diagnosis status was documented 100% (n=12) of the time. When the PAAF was not used, a detailed manual chart review was needed to determine the diagnosis status.

Asthma control documentation was compared between PAAFs used and manual chart abstraction from EMR encounter notes (Table 8). In terms of asthma control, the average number of CTS asthma control parameters documented was significantly higher when the PAAF was used (PAAF: mean 5.4, SD 1.9 vs manual chart abstraction: mean 2.3, SD 1.2; *P*<.001). Asthma control was documented significantly more often in manual chart abstractions (PAAF: n=7, 58.3% vs manual chart abstraction: n=360, 98.4%; *P*<.001). Finally, there was a significantly higher proportion of patients whose asthma was uncontrolled at the time of the visit in manual chart abstractions (manual chart abstraction: n=302, 82.5% vs PAAF: n=6, 50%; *P*=.01).

Table . Asthma control: manual chart abstractions versus Provider Asthma Assessment Form (PAAF).

Asthma control	Manual chart abstraction (n=366)	PAAF (n=12)	*P* value[a]
≥1 Asthma control parameter documented, n (%)	360 (98.4)	7 (58.3)	<.001
Asthma uncontrolled at visit, n (%)	302 (82.5)	6 (50)	.01
Number of asthma control parameters documented, mean (SD)[b]	2.3 (1.2)	5.4 (1.9)	<.001

[a]All values were significant (*P*<.05).

[b]Average number of asthma control parameters was based on asthma visits where asthma control was documented.

Discussion

Summary of Main Findings

Following a multifaceted intervention of implementing the PAAF in this primary care practice, there were substantial improvements in asthma-specific documentation and adherence to key evidence-based recommendations for care. There was an 18.1% increase in the number of PFTs requested after the implementation of the PAAF. This did not translate to the number of confirmed asthma cases, presumably due to the impacts of the COVID-19 pandemic on PFT wait times. Furthermore, a highly significant and clinically relevant increase was noted in the proportion of patients prescribed SIT after implementation of the PAAF (postimplementation: n=31, 21.7% vs baseline: n=2, 0.9%), and a significantly greater proportion of patients were on ICS/LABA controller therapy after implementation. In general, a higher percentage of patients were on reliever, ICS, and second controller therapy after implementation, and a smaller proportion had received systemic corticosteroids in the last year. Asthma control was highly documented in both cohorts, but significantly more CTS asthma control parameters were documented after implementation, and a higher proportion of patients with uncontrolled asthma were

XSL•FO
RenderX

prescribed ICS therapy. A greater percentage of patients had barriers addressed at the asthma encounter after implementation compared to baseline (postimplementation: n=24, 16.8% vs baseline: n=11, 4.8%). There were no meaningful changes in specialist referrals due to suspected SA. Although it is not possible to discern causality for observed changes based on this study, overall, care as assessed by key PC-APIs showed improvement in the postimplementation cohort.

Interpretation of Main Findings

Diagnosis Status and Monitoring With PFTs

Overall, the proportion of patients in this family practice with confirmed asthma in both cohorts (~25% - 30% of patients) is consistent with previous literature [22]. The finding of PFTs completed or requested within 12 months of the encounter date is consistent with similar literature in Ontario [21]. A study by Gershon et al [21] found that <50% of Ontarian people underwent spirometry within 1 year before and 2.5 years after a new diagnosis of asthma. Our study found that 49% (n=70) of postimplementation patients and 30.9% (n=71) of baseline patients had a PFT requested. Thus, less than 50% of patients in this study had been monitored with a PFT within 12 months of an asthma encounter with their primary care provider. However, it is important to note that our study did not differentiate new diagnoses from previous diagnoses for the purposes of analysis. Additionally, our study found an 18.1% significant increase in PFT requests in the postimplementation cohort versus the retrospective baseline. However, this did not yet translate to PFTs completed in the postimplementation group. The reduced number of patients in postimplementation with confirmed asthma (baseline: n=72, 31.3% vs postimplementation: n=34, 23.8%) may in part be attributed to the backlog of requests and limited access to PFTs due to the COVID-19 pandemic. As previously stated, a greater proportion of PFTs were requested after implementation but not completed 12 months prior to the asthma encounter date. It is evident that a key asthma care gap persists at this family practice, as there is limited confirmation of an asthma diagnosis and monitoring asthma with PFTs. It would be prudent in the future to determine the barriers to confirmation of diagnosis status and potentially incorporate diagnosis status as a required data element to be documented within primary care EMRs, such as including validated PRESTINE data elements for asthma diagnosis [12]. In addition, it would be beneficial to assess and address the current availability, associated costs, and accessibility of PFTs in Canadian primary care settings.

Medications

It was interesting to see the significant increase in SIT in the postimplementation cohort. In recent CTS guidelines, SIT consists of pharmacologic management with budesonide-formoterol inhaler both as a PRN (as needed) reliever and controller therapy [3]. In the Global Initiative for Asthma (GINA) report 2024, SIT (known in the GINA report as maintenance-and-reliever therapy) is recommended treatment for GINA steps 3 - 4 [7]. Hence, there was increased adherence with this guideline recommendation in the postimplementation period. There were also significantly more patients on an ICS/LABA inhaler after implementation. It is a CTS guidelines recommendation that adults with asthma not achieving adequate control on a low-dose ICS be escalated to ICS/LABA therapy [3]. Hence, this can be seen as greater adherence to CTS guidelines as well. Overall, there was increased use of reliever, ICS, and second controller therapy after implementation. There was also a reduction in systemic corticosteroid use in the postimplementation period (5.6% decrease). The reduction in systemic corticosteroid use may be secondary to appropriate increased use of ICS therapies (ICS and ICS/LABA) [3].

Asthma Control

CTS guidelines recommend that asthma control be assessed at each clinical encounter [3]. There was a high degree of compliance with this recommendation in both cohorts, as asthma control was documented in over 96% of asthma encounters. Guidelines also outline 9 asthma control parameters that aid in determining whether asthma is well-controlled at the encounter [3]. In this study, control was documented in a higher average number of encounters in the postimplementation period. However, less than 3 asthma control parameters were documented in both groups. Asthma control would be more accurately assessed during asthma encounters if a greater number of control parameters were assessed [3]. When all control parameters are not documented or assessed, the classification of asthma control could be over- or underestimated.

Severe Asthma

The clinical decision support component of the PAAF includes an embedded SA algorithm, which aids in the recognition and referral of uncontrolled asthma and/or SA. This functionality of the PAAF directly integrates key messaging from a recent CTS position statement on SA and the Pan-Canadian standards for SA in EMRs [6,17]. In primary care, key messaging focuses on improved recognition of uncontrolled asthma and/or SA and increasing referrals for specialized evaluation and asthma education. Thus, the SA outcomes for this study pertained to the recognition of SA by providers and referrals to specialist care.

In terms of identification of SA, primary care physicians suspected SA in 2.6% (n=6) of baseline patients and 4.9% (n=7) of postimplementation patients. In contrast, the abstractor suspected SA based on the SA algorithm [17] in 4.3% (n=10) of baseline and 7% (n=10) of postimplementation. SA was suspected more often in both cohorts based on the SA algorithm compared to primary care providers without the use of the algorithm [17]. This represents a persistent care gap for SA, which could be improved by broader adoption of the PAAF.

Health Services Use, Asthma Education, and PC-APIs

Overall, there was limited adherence to CTS guidelines for best-practice asthma education in both cohorts. The most common aspect of asthma education addressed at each primary care encounter was asthma triggers, as triggers were addressed in over 76% of asthma encounters. However, various core components of effective asthma education were not addressed in this patient population.

XSL•FO
RenderX

CTS guidelines highlight the importance of providing self-management education including a written AAP, discussing environmental controls, and referring patients to asthma education programs when appropriate [3]. AAPs were provided, revised, or reviewed in less than 10% of encounters, and environmental control was discussed in less than 13%. Notably, less than 6% of patients in both cohorts were referred to an asthma education program or Certified Respiratory Educator (CRE), despite having access to an on-site CRE each month as part of the Primary Care Asthma Program [25]. There is strong evidence that community evidence–based asthma care programs improve clinical outcomes, including a decreased risk of asthma exacerbations and urgent health services use [25]. However, this resource was not effectively used in baseline or postimplementation. It would be warranted to further investigate in the future the reasoning behind the lack of use of this resource.

Overall, care as assessed by key PC-APIs showed tendencies toward improvement in the postimplementation cohort, although most did not achieve statistical significance. A significant finding was that an increased proportion of patients with uncontrolled asthma were on an ICS in the postimplementation. Additionally, a positive trend included a 5% reduction in known or suspected adherence issues with pharmacologic therapy after implementation. CTS guidelines [3] and the GINA report [7] emphasize the fundamental importance of appropriate ICS therapy in achieving asthma control for individuals with uncontrolled asthma and preventing future risk of exacerbations. Thus, these findings demonstrated increased adherence with evidence-based guidelines.

Furthermore, in the postimplementation period, a greater proportion of patients received AAPs, had inhaler technique monitoring, smoking cessation addressed, well-controlled asthma, PFT monitoring in the last 12 months, and referrals to an asthma education program or CRE. There were also decreased emergency department visits in the previous year and recent asthma exacerbations after implementation. Overall, the positive trends suggest that implementation of the PAAF, either the tool itself or the implementation strategies used in the study, may be associated with improvements in asthma management and better adherence with best-practice guidelines. However, we did not see a noteworthy change in one of the primary outcomes, specialist referral of patients with suspected SA. Low uptake of the PAAF meant that the study was underpowered to detect significance in specialist referral for suspected SA.

Use of the PAAF

A comparison between the use of the PAAF to document an asthma encounter compared to manual chart abstraction was conducted. When the PAAF was used, asthma diagnosis status (either suspected or confirmed) was documented 100% of the time. Hence, the use of the PAAF would aid in addressing one of the key asthma care gaps in primary care, confirming asthma through objective measures [3,19,26].

Additionally, a significantly higher proportion of asthma control parameters were documented when the PAAF was used compared to encounter notes used for manual chart abstraction. Compared to current EMR primary documentation of asthma management in encounter notes, the use of the PAAF eases access to pertinent medical history as well as documentation of diagnosis status and asthma control.

Provider Uptake

One of the main challenges of this study was the uptake of the PAAF in the clinic and the participation of providers throughout the intervention. This could be due to many factors, such as change management issues (diffusion of innovations theory) [27], the implementation model did not address sufficient determinants of behavior to effect change (Theoretical Domains Framework) [28,29], or insufficient time in the clinic for providers to adopt a new tool. Thus, the perceived utility, provider satisfaction, and barriers and enablers for the use of the form are investigated in another study. However, future studies involving the PAAF could use various other methods to increase uptake in the clinic.

Globally, there is evidence supporting the various benefits of multidisciplinary primary health care teams to meet the changing demands of the health care system [30-32]. Hence, there is an increased occurrence of and need for task shifting from physicians to allied health care professionals [30-32]. Particularly in Canada, there are a growing number of unattached patients with a survey in 2022 finding that ~22% of patients across Canada reported not having a family physician or nurse practitioner [33]. The shift toward multidisciplinary teams would aid in alleviating administrative burden [32], aid in chronic disease management (ie, asthma) [34], and allow for an increased number of patients to be connected to comprehensive primary health care [35]. Therefore, it would be reasonable to design a multifaceted, multidisciplinary intervention to implement the PAAF in primary care practice to increase user uptake. This would be particularly relevant if the PAAF was used as part of a broader asthma intervention involving spirometry testing and asthma education in primary care [34].

Comparison to the Literature

In comparison to other EMR-integrated asthma KT tools, our results were similar. Our findings can be compared to 2 primary care EMR KT tools. The electronic asthma management system was found to be an efficacious tool, particularly increasing AAP delivery, assessment of control, and escalation of pharmacologic therapy after the intervention [15,36]. However, similar to our study, there was limited user uptake beyond the scope of the initial study period [15]. A similar result was found during the evaluation of the breathe (University Health Network) mobile app, with limited user uptake [14]. It is important to distinguish that these e-Tools were patient-facing tools with EMR integration for providers to view results entered by patients. The PAAF, breathe app, and electronic asthma management system were all found to be efficacious tools for asthma management, but long-term engagement with the tools by primary care providers remains low.

Limitations

The main limitation of this study was the limited uptake of the PAAF in the primary care practice. Therefore, the sample size was limited, potentially impacting the power of our study. As

XSL•FO
RenderX

can be seen from the CPCSSN case definition search, there were 136 visits that would have been eligible for the PAAF to be used in the postimplementation time frame. However, it was only used 12 times throughout the study duration. The perceived utility, provider satisfaction, and barriers and enablers to PAAF use are the focus of another study.

Furthermore, it is important to note that the results of this study can only be interpreted in the context of the broader multifaceted intervention encompassing not only the PAAF tool but also the various implementation strategies used as part of the training and orientation to the tool. Further studies are needed to differentiate the specific impact of the tool itself, rather than the global strategies used as part of the implementation of the PAAF in a primary care setting.

It is important to note that it is difficult to discern causality for the changes seen in asthma management practice in this study. As this study was conducted at an academic primary care practice, the resident physicians involved with patient care receive additional instruction and education on effective asthma management during their academic instructional time. As well, there was a concurrent quality improvement project focusing on the Choosing Wisely Canada climate change and inhaler recommendations for patients with asthma [37]. This quality improvement initiative aimed to decrease the number of patients on pressurized metered dose inhalers and increase the number of patients on dry-powder inhalers, which may have influenced the proportion of patients on SIT (which is commonly a dry-powder inhaler). Therefore, it is important to recognize that this initiative may have increased the focus on asthma management in the primary care practice, beyond the implementation efforts for the PAAF.

Additionally, the percentage of postimplementation patients with a PFT in the last 12 months was likely impacted due to a persistent backlog in the Pulmonary Function Lab at Kingston Health Sciences Centre (1 of 2 centers for PFT testing in Kingston) due to the COVID-19 pandemic. Hence, this study may not reflect a complete picture of PFT monitoring in baseline compared to postimplementation. In an effort to address this limitation, the number of requested PFTs within 12 months of the documented encounter was included in the analysis.

It is pertinent to recognize the inherent limitations of manual chart abstraction as well. Data elements collected are limited by the information providers document within the chart and encounter notes; thus, the documentation may not fully capture all aspects of care discussed during an asthma visit.

Conclusions

Following the multifaceted intervention of implementing the PAAF in this primary care practice, there were significant improvements in asthma-specific documentation and adherence to key evidence-based recommendations for care. The use of the PAAF increased asthma visit–specific documentation, such as clearly documenting diagnosis status and asthma control parameters. However, asthma care gaps such as the underuse of PFTs for diagnosis and monitoring, asthma education, and addressing reasons for poor asthma control were still common in this study. Future directions should involve evaluating the impact of the PAAF after widespread implementation in primary care settings and investigating methods to increase user uptake of the PAAF.

Acknowledgments

This work was supported by a Queen’s University Faculty of Health Sciences Internal Grant (William M. Spear Endowment Fund in Pulmonary Research or Richard K. Start Memorial Fund). The authors would like to acknowledge the valued contributions of Rebecca Theal for her work preparing Canadian Primary Care Sentinel Surveillance Network data cuts for this project.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

MDL conceived the study question. MDL, MLM, DB, DP, and AM conceived the study design. MLM completed the data collection, chart abstraction, and analysis, with guidance from MDL and AM. MLM wrote the initial draft, with guidance and revisions from MDL and AM. All authors provided editorial guidance and approved the final submitted version.

Conflicts of Interest

MLM and DB have no conflicts of interest to disclose. MDL has received grants outside the submitted work paid directly to Queen's University from the Canadian Institutes of Health Research (subgrants from Ottawa Health Research Institute and the University of Toronto), Workers Compensation Board of Manitoba, Queen's University, and GlaxoSmithKline, as well as honoraria from AstraZeneca for participation in the PRECISION Program Advisory Board. AM has received an honorarium from AstraZeneca for attendance at a 2021 PRECISION Severe Asthma Summit. DP has received speaker honoraria from AstraZeneca, GlaxoSmithKline, Sanofi Genzyme, the Lung Health Foundation, and Novartis. DP has also participated on advisory boards for Sanofi Genzyme and GlaxoSmithKline.

Multimedia Appendix 1

XSL•FO
RenderX

Confirmed versus suspected asthma.
[PNG File, 59 KB - formative_v9i1e74043_app1.png]

Multimedia Appendix 2
Pulmonary function tests completed versus requested. *P<.001.
[PNG File, 97 KB - formative_v9i1e74043_app2.png]

Multimedia Appendix 3
Summary of prescribed asthma medications. **P<.001.
[PNG File, 63 KB - formative_v9i1e74043_app3.png]

Multimedia Appendix 4
Asthma control documentation.
[PNG File, 35 KB - formative_v9i1e74043_app4.png]

References

1. Wang Z, Li Y, Gao Y, et al. Global, regional, and national burden of asthma and its attributable risk factors from 1990 to 2019: a systematic analysis for the Global Burden of Disease Study 2019. Respir Res 2023 Jun 23;24(1):169. [doi: 10.1186/s12931-023-02475-6] [Medline: 37353829]
2. Teresa To AG, Tassoudji M, Guan J, Wang C, Estrabillo E, Cicutto L. The Burden of Asthma in Ontario: Institute for Clinical Evaluative Sciences; 2006.
3. Yang CL, Hicks EA, Mitchell P, et al. Canadian Thoracic Society 2021 Guideline update: diagnosis and management of asthma in preschoolers, children and adults. Can J Respir Crit Care Sleep Med 2021 Nov 2;5(6):348-361. [doi: 10.1080/24745332.2021.1945887]
4. Boulet LP, Becker A, Bowie D, et al. Implementing practice guidelines: a workshop on guidelines dissemination and implementation with a focus on asthma and COPD. Can Respir J 2006 Mar;13(Suppl A):5-47. [doi: 10.1155/2006/810978] [Medline: 16552449]
5. FitzGerald JM, Boulet LP, McIvor RA, Zimmerman S, Chapman KR. Asthma control in Canada remains suboptimal: the Reality of Asthma Control (TRAC) study. Can Respir J 2006;13(5):253-259. [doi: 10.1155/2006/753083] [Medline: 16896426]
6. FitzGerald JM, Lemiere C, Lougheed MD, et al. Recognition and management of severe asthma: a Canadian Thoracic Society position statement. Can J Respir Crit Care Sleep Med 2017 Oct 2;1(4):199-221. [doi: 10.1080/24745332.2017.1395250]
7. Global strategy for asthma management and prevention. : Global Initative for Asthma; 2024 URL: https://ginasthma.org/wp-content/uploads/2024/12/GINA-Summary-Guide-2024-WEB-WMS.pdf [accessed 2025-05-31]
8. Chung KF, Wenzel SE, Brozek JL, et al. International ERS/ATS guidelines on definition, evaluation and treatment of severe asthma. Eur Respir J 2014 Feb;43(2):343-373. [doi: 10.1183/09031936.00202013] [Medline: 24337046]
9. Gershon AS, Guan J, Wang C, To T. Trends in asthma prevalence and incidence in Ontario, Canada, 1996-2005: a population study. Am J Epidemiol 2010 Sep 15;172(6):728-736. [doi: 10.1093/aje/kwq189]
10. Aaron SD, Vandemheen KL, Boulet LP, et al. Overdiagnosis of asthma in obese and nonobese adults. Can Med Assoc J 2008 Nov 18;179(11):1121-1131. [doi: 10.1503/cmaj.081332]
11. Murray S, Labbé S, Kaplan A, Petrasko K, Waserman S. A multi-stakeholder perspective on asthma care in Canada: findings from a mixed methods needs assessment in the treatment and management of asthma in adults. Allergy Asthma Clin Immunol 2018 Dec;14(1). [doi: 10.1186/s13223-018-0261-x]
12. Lougheed MD, Taite A, ten Hove J, et al. Pan-Canadian asthma and COPD standards for electronic health records: a Canadian Thoracic Society Expert Working Group Report. Can J Respir Crit Care Sleep Med 2018 Oct 2;2(4):244-250. [doi: 10.1080/24745332.2018.1517623]
13. Moloney M, MacKinnon M, Bullock E, et al. Integrating user preferences for asthma tools and clinical guidelines into primary care electronic medical records: mixed methods study. JMIR Form Res 2023 Feb 21;7:e42767. [doi: 10.2196/42767] [Medline: 36809175]
14. Ferrone M, Taite A, Madeley C, Lavigne AS, To T, Yeung M, et al. The evaluation of breathe—a patient mobile health (mHealth) app for adult asthma. Am J Respir Crit Care Med 2016. [doi: 10.1164/ajrccm-conference.2016.193.1_MeetingAbstracts.A1084]
15. Gupta S, Price C, Agarwal G, et al. The Electronic Asthma Management System (eAMS) improves primary care asthma management. Eur Respir J 2019 Apr;53(4):1802241. [doi: 10.1183/13993003.02241-2018] [Medline: 30765503]
16. Taite A, Minard J, Morra A, Lougheed MD. Supporting asthma and COPD best practices in Ontario through primary care electronic medical records. Healthc Q 2020 Jul;23(2):67-74. [doi: 10.12927/hcq.2020.26271] [Medline: 32762824]

XSL•FO
RenderX

17. Lougheed MD, Morra A, Bullock E, et al. Pan-Canadian standards for severe asthma in electronic medical records. Can J Respir Crit Care Sleep Med 2021 Nov 2;5(6):391-399. [doi: 10.1080/24745332.2021.1947758]
18. To T, Guttmann A, Lougheed MD, et al. Evidence-based performance indicators of primary care for asthma: a modified RAND Appropriateness Method. Int J Qual Health Care 2010 Dec;22(6):476-485. [doi: 10.1093/intqhc/mzq061] [Medline: 20978002]
19. Aaron SD, Boulet LP, Reddel HK, Gershon AS. Underdiagnosis and overdiagnosis of asthma. Am J Respir Crit Care Med 2018 Oct 15;198(8):1012-1020. [doi: 10.1164/rccm.201804-0682CI] [Medline: 29756989]
20. Aaron SD, Vandemheen KL, Whitmore GA, et al. Early diagnosis and treatment of COPD and asthma—a randomized, controlled trial. N Engl J Med 2024 Jun 13;390(22):2061-2073. [doi: 10.1056/NEJMoa2401389] [Medline: 38767248]
21. Gershon AS, Victor JC, Guan J, Aaron SD, To T. Pulmonary function testing in the diagnosis of asthma: a population study. Chest 2012 May;141(5):1190-1196. [doi: 10.1378/chest.11-0831] [Medline: 22030804]
22. Moloney M, Morra A, Morkem R, et al. Validation of adult asthma case definitions for primary care sentinel surveillance. Allergy Asthma Clin Immunol 2023;19(1):1-11. [doi: 10.1186/s13223-023-00854-8]
23. Respiratory medicine: eight tests and treatments to question. Choosing Wisely Canada. 2023. URL: https://choosingwiselycanada.org/recommendation/respiratory-medicine [accessed 2025-05-22]
24. Chapman KR, Ernst P, Grenville A, Dewland P, Zimmerman S. Control of asthma in Canada: failure to achieve guideline targets. Can Respir J 2001;8(Suppl A):35A-40A. [doi: 10.1155/2001/245261] [Medline: 11360046]
25. To T, Cicutto L, Degani N, McLimont S, Beyene J. Can a community evidence-based asthma care program improve clinical outcomes?: A longitudinal study. Med Care 2008 Dec;46(12):1257-1266. [doi: 10.1097/MLR.0b013e31817d6990] [Medline: 19300316]
26. Price C, Agarwal G, Chan D, et al. Large care gaps in primary care management of asthma: a longitudinal practice audit. BMJ Open 2019 Jan 29;9(1):e022506. [doi: 10.1136/bmjopen-2018-022506] [Medline: 30696669]
27. Dearing JW, Cox JG. Diffusion of innovations theory, principles, and practice. Health Aff (Millwood) 2018 Feb;37(2):183-190. [doi: 10.1377/hlthaff.2017.1104]
28. Atkins L, Francis J, Islam R, et al. A guide to using the Theoretical Domains Framework of behaviour change to investigate implementation problems. Implement Sci 2017 Jun 21;12(1):77. [doi: 10.1186/s13012-017-0605-9] [Medline: 28637486]
29. Yamada J, Lam Shin Cheung J, Gagne M, et al. Barriers and enablers to objective testing for asthma and COPD in primary care: a systematic review using the theoretical domains framework. Chest 2022 Apr;161(4):888-905. [doi: 10.1016/j.chest.2021.10.030] [Medline: 34740591]
30. Leong SL, Teoh SL, Fun WH, Lee SWH. Task shifting in primary care to tackle healthcare worker shortages: an umbrella review. Eur J Gen Pract 2021 Dec;27(1):198-210. [doi: 10.1080/13814788.2021.1954616] [Medline: 34334095]
31. Paier-Abuzahra M, Posch N, Jeitler K, et al. Effects of task-shifting from primary care physicians to nurses: an overview of systematic reviews. Hum Resour Health 2024 Nov 11;22(1):74. [doi: 10.1186/s12960-024-00956-3] [Medline: 39529012]
32. Saxon RL, Gray MA, Oprescu FI. Extended roles for allied health professionals: an updated systematic review of the evidence. J Multidiscip Healthc 2014;7:479-488. [doi: 10.2147/JMDH.S66746] [Medline: 25342909]
33. Kiran T, Green M, Breton M, et al, editors. Results from the OurCare Survey: Public Perspectives on Primary Care in Canada: American Academy of Family Physicians; 2023. [doi: 10.1370/afm.22.s1.5232]
34. Licskai C, Sands T, Ong M, Paolatto L, Nicoletti I. Using a knowledge translation framework to implement asthma clinical practice guidelines in primary care. Int J Qual Health Care 2012 Oct;24(5):538-546. [doi: 10.1093/intqhc/mzs043] [Medline: 22893665]
35. Nabieva K, McCutcheon T, Liddy C. Connecting unattached patients to comprehensive primary care: a rapid review. Prim Health Care Res Dev 2023;24. [doi: 10.1017/S1463423623000099]
36. Kouri A, Yamada J, Sale JEM, Straus SE, Gupta S. Primary care pre-visit electronic patient questionnaire for asthma: uptake analysis and predictor modeling. J Med Internet Res 2020 Sep 18;22(9):e19358. [doi: 10.2196/19358] [Medline: 32945779]
37. Gupta S, Couillard S, Digby G, et al. Canadian Thoracic Society Position Statement on climate change and choice of inhalers for patients with respiratory disease. Can J Respir Crit Care Sleep Med 2023 Sep 3;7(5):232-239. [doi: 10.1080/24745332.2023.2254283]

Abbreviations

AAP: Asthma Action Plan
CPCSSN: Canadian Primary Care Sentinel Surveillance Network
CRE: Certified Respiratory Educator
CTS: Canadian Thoracic Society
e-Tool: electronic tool
EMR: electronic medical record
FHT: Family Health Team
GINA: Global Initiative for Asthma

XSL•FO
RenderX

ICD-9*: *International Classification of Diseases, Ninth Revision
ICS: inhaled corticosteroid
KT: Knowledge Translation
LABA: long-acting β-2 agonist
PAAF: Provider Asthma Assessment Form
PC-API: Primary Care—Asthma Performance Indicator
PFT: pulmonary function test
PRESTINE: Pan-Canadian Respiratory Standards Initiative for Electronic Health Records
SA: severe asthma
SIT: single inhaler therapy

Edited by A Mavragani; submitted 16.03.25; peer-reviewed by A Kaplan, AF Kalpaklioglu; revised version received 18.04.25; accepted 30.04.25; published 25.06.25.

Please cite as:
McFarlane ML, Morra A, Podgers D, Barber D, Lougheed MD
Impact of a Novel Electronic Medical Record–Integrated Electronic Form (Provider Asthma Assessment Form) and Severe Asthma Algorithm in Primary Care: Single-Center, Pre- and Postobservational Study
JMIR Form Res 2025;9:e74043
URL: https://formative.jmir.org/2025/1/e74043
doi:10.2196/74043

© Matheson L McFarlane, Alison Morra, Delanya Podgers, David Barber, M Diane Lougheed. Originally published in JMIR Formative Research (https://formative.jmir.org), 25.6.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Public Perception of the Brain-Computer Interface Based on a Decade of Data on X: Mixed Methods Study

Mohammed A Almanna[1,2], MBBS; Lior M Elkaim[3], MD; Mohammed A Alvi[4,5,6], MBBS, MSc; Jordan J Levett[7], MD; Ben Li[8,9,10,11], MD; Muhammad Mamdani[10,11,12,13,14,15,16], MA, MPH, PharmD; Mohammed Al-Omran[8,9,10,11,15,17,18], MSc, MD; Naif M Alotaibi[17,19], MSc, MD

[1]College of Medicine, King Saud Bin Abdulaziz University for Health Sciences, Riyadh, Saudi Arabia
[10]Institute of Medical Science, University of Toronto, Toronto, ON, Canada
[11]Temerty Centre for Artificial Intelligence Research and Education in Medicine, University of Toronto, Toronto, ON, Canada
[12]Data Science & Advanced Analytics, Unity Health Toronto, University of Toronto, Toronto, ON, Canada
[13]Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, ON, Canada
[14]Institute for Clinical Evaluative Sciences, University of Toronto, Toronto, ON, Canada
[15]Li Ka Shing Knowledge Institute, St. Michael's Hospital, Unity Health Toronto, Toronto, ON, Canada
[16]Leslie Dan Faculty of Pharmacy, University of Toronto, Toronto, ON, Canada
[17]College of Medicine, Alfaisal University, Riyadh, Saudi Arabia
[18]Department of Surgery, King Faisal Specialist Hospital and Research Center, Riyadh, Saudi Arabia
[19]National Neuroscience Institute, King Fahad Medical City, As Sulimaniyah, Makkah Road, Riyadh 12231, Riyadh, Saudi Arabia
[2]King Abdullah International Medical Research Center, Riyadh, Saudi Arabia
[3]Department of Neurology and Neurosurgery, McGill University, Montreal, QC, Canada
[4]Division of Neurosurgery, Department of Surgery, University of Toronto, Toronto, ON, Canada
[5]Neuro International Collaboration, Toronto, ON, Canada
[6]Department of Neurologic Surgery, Mayo Clinic, Rochester, MN, United States
[7]Faculty of Medicine, University of Montreal, Montreal, QC, Canada
[8]Department of Surgery, University of Toronto, Toronto, ON, Canada
[9]Division of Vascular Surgery, St. Michael's Hospital, Unity Health Toronto, University of Toronto, Toronto, ON, Canada

Corresponding Author:
Naif M Alotaibi, MSc, MD
College of Medicine, Alfaisal University, Riyadh, Saudi Arabia

Abstract

Background: Given the recent evolution and achievements in brain-computer interface (BCI) technologies, understanding public perception and sentiments toward such novel technologies is important for guiding their communication strategies in marketing and education.

Objective: This study aims to explore the public perception of BCI technology by examining posts on X (formerly known as Twitter) using natural language processing (NLP) methods.

Methods: A mixed methods study was conducted on BCI-related posts from January 2010 to December 2021. The dataset included 65,340 posts from 38,962 unique users. This dataset was subject to a detailed NLP analysis including VADER, TextBlob, and NRCLex libraries, focusing on quantifying the sentiment (positive, neutral, and negative), the degree of subjectivity, and the range of emotions expressed in the posts. The temporal dynamics of sentiments were examined using the Mann-Kendall trend test to identify significant trends or shifts in public interest over time, based on monthly incidence. We used the Sentiment.ai tool to infer users' demographics by matching predefined attributes in users' profile biographies to certain demographic groups. We used the BERTopic tool for semantic understanding of discussions related to BCI.

Results: The analysis showed a significant rise in BCI discussions in 2017, coinciding with Elon Musk's announcement of Neuralink. Sentiment analysis revealed that 59.38% (38,804/65,340) of posts were neutral, 32.75% (21,404/65,340) were positive, and 7.85% (5132/65,340) were negative. The average polarity score demonstrated a generally positive trend over the course of the study (Mann-Kendall Statistic=0.266; τ=0.266; P<.001). Most posts were objective (50,847/65,340, 77.81%), with a smaller proportion being subjective (14,393/65,340, 22.02%). Biographic analysis showed that the "broadcasting" group contributed the most to BCI discussions (17,803/58,030, 30.67%), while the "scientific" group, contributing 27.58% (n=16,005), had the highest overall engagement metrics. The emotional analysis identified anticipation (score = 10,802/52,618, 20.52%), trust

XSL•FO
RenderX

(score=9244/52,618, 17.56%), and fear (score=7344/52,618, 13.95%) as the most prominent emotions in BCI discussions. Key topics included Neuralink and Elon Musk, practical applications of BCIs, and the potential for gamification.

Conclusions: This NLP-assisted study provides a decade-long analysis of public perception of BCI technology based on data from X. Overall, sentiments were neutral yet cautiously apprehensive, with anticipation, trust, and fear as the dominant emotions. The presence of fear underscores the need to address ethical concerns, particularly around data privacy, safety, and transparency. Transparent communication and ethical considerations are essential for building public trust and reducing apprehension. Influential figures and positive clinical outcomes, such as advancements in neuroprosthetics, could enhance favorable perceptions. The gamification of BCI, particularly in gaming and entertainment, also offers potential for wider public engagement and adoption. However, public perceptions on X may differ from other platforms, affecting the broader interpretation of results. Despite these limitations, the findings provide valuable insights for guiding future BCI developments, policy making, and communication strategies.

(JMIR Form Res 2025;9:e60859) doi:10.2196/60859

KEYWORDS

brain-computer interface; BCI; Neuralink; sentiment analysis; public perception; technology; innovation; Twitter; social media; natural language processing; NLP; mixed method; semantic; Mann-Kendall; brain-machine interface; decade; data; public perceptions; sentiment; marketing; education

Introduction

Brain-computer interface (BCI) is an emerging technology that allows for direct communication of the brain's signals to external devices. This innovation operates through four sequential stages: signal acquisition, feature extraction, feature translation, and generating the device output [1]. The effectiveness of a BCI system largely depends on its signal acquisition module, which can range from noninvasive methods, such as surface electroencephalography electrodes, to more invasive approaches like endovascular stent-electrode arrays and electrodes implanted on the brain surface. While more invasive methods generally allow for higher signal fidelity and better system performance, overall effectiveness also depends on additional factors such as signal processing and user adaptability [2,3].

BCI has been extensively researched across various fields, including medical rehabilitation [4-6], control of orthotic and prosthetic devices [7,8], assistive technologies [9,10], and video gaming [11]. Additionally, BCI has potential applications in enhancing cognitive functions [12]. The remarkable progress in BCI, along with the involvement of well-recognized institutions and prominent figures like Elon Musk, has brought this technology to the forefront of public awareness. Public perception of BCI technologies is influenced by their representation in the media, with concerns over ethical dilemmas including privacy and mind control, and the invasive nature of certain BCI technologies [13-18].

X (formerly known as Twitter) provides real-time insights into the thoughts, feelings, and conversations of millions of users. Natural language processing (NLP) tools are instrumental in analyzing social media content, offering deeper insights into public perception. NLP methods enable the analysis of public sentiment toward specific topics, the detection of emerging trends, and the identification of demographic groups participating in these discussions. These tools have been extensively used to assess public acceptance of vaccines [19,20], guide economic investments [21], evaluate innovative products [22,23], and more. BCI is an emerging technology with concepts that many may still consider science fiction, leading to polarized opinions among the public. While some individuals might be excited about its potential applications, others may express concerns due to potential complications and the possibility of malicious uses. Understanding public sentiment toward BCI is crucial for guiding ethical frameworks, informing policy decisions, and shaping the direction of future research and development. However, there is a lack of comprehensive studies assessing public perception of BCI through social media discussions. This study aims to bridge the gap by using NLP tools to analyze over a decade of X conversations about BCI. The goals of this study are to quantify sentiments, identify trends in public perception, explore subjectivity, and understand the nature of public discussions related to BCI.

Methods

Data Source and Processing

We used X application programming interface, Twitter application programming interface for academic research, and database to identify posts related to BCI using the search term "brain-computer interface." The acronym "BCI" was not used to avoid including irrelevant posts. The search was conducted from X inception (March 2006) to May 2022, prior to Elon Musk's acquisition of X. The data underwent preprocessing, which involved omitting any mentions, URLs, and hashtags, removing any line breaks, deleting any HTML characters, replacing them with their respective Unicode equivalent, eliminating any special characters or punctuation points except exclamation points (the only punctuation mark relevant for sentiment analysis), and excluding posts from users with fewer than 10 followers to minimize "bot" influence and duplicate entries. We excluded posts before January 2010 due to limited data availability and after December 2021 to maintain the temporal consistency of the dataset, as our data cover only a few months of 2022 (Figure S1 in Multimedia Appendix 1). Detailed individual post data included the text, date and time of posts, the number of reposts, replies, likes, and quote count. Additional data included whether the post included links, media, tagging, or any hashtags. User information included username,

XSL•FO
RenderX

the number of followers, the total number of author posts, user biography, and location.

Biography Analysis

To explore users' demographics, we used the sentiment.ai [24] library match function after preprocessing user biographies and excluding users with empty biographies. Sentiment.ai is a text-based deep machine learning tool that allows for category matching of the most similar phrase and its category, providing a cosine similarity score. The categories (ie, the biographic groups) and phrases (ie, attributes) are shown in Table S1 in Multimedia Appendix 1. We applied categories such as "broadcast," "scientific," "entrepreneurship," and "clinical" based on predefined attributes. We measured the cosine similarity score between profile biographies and the most similar attribute, categorizing them as "others" if the similarity score is less than 0.05 [23] (Figure S2 in Multimedia Appendix 1).

Sentiment Quantification

Sentiment by Valence Aware Dictionary and Sentiment Reasoner

The sentiment polarity was analyzed using the VADER (Valence Aware Dictionary and Sentiment Reasoner) library [25], a lexicon designed for sentiment analysis in social media contexts. It accounts for important elements including emojis, emoticons, slang words, and acronyms or initialisms with sentimental value (eg, "lol") in determining the compound score. The compound score is calculated by adding the sentiment scores of each word, and it is set to be between −1 (negative) and +1 (positive) after being adjusted according to a set of rules. The threshold in our study was a compound score of ≥0.05 for positive sentiment posts, <0.05 and >−0.05 for neutral sentiment posts, and ≤−0.05 for posts expressing negative sentiment (Figure S3 in Multimedia Appendix 1).

Emotions by NRCLex

The NRCLex tool [26] is a Python library that allows for the analysis of the emotional content of text using the National Research Council Canada emotion lexicon, which contains approximately 27,000 words. It provides a simple interface to extract various emotions and sentiments from text. The emotional effects measured include fear, anger, anticipation, trust, surprise, sadness, disgust, and joy. In our study, we used the "raw_emotion_scores" model of the NRCLex tool to count the frequency of words associated with certain emotions in a text. For example, a post containing 3 words associated with anticipation and 1 word associated with trust would have an emotion score of +3 for anticipation and +1 for trust. The primary emotion for each post was determined by identifying the emotion with the highest score, giving equal consideration to multiple emotions if they had equal scores. If no emotion was identified for a post, we used the label "no emotion" and excluded such posts from any emotion analysis. These analyses were applied only to posts with positive or negative sentiment, excluding posts with neutral sentiment, as they are not informative for emotion analysis (Figure S3 in Multimedia Appendix 1).

Subjectivity by TextBlob

TextBlob library [27] was used to classify X posts based on the subjectivity score (ranging from 0 to 1), which detects the degree of personal opinion expressed in the text. Words that are more opinion-based (eg, scary and amazing) have a higher subjectivity score, whereas words that are fact-based (eg, data and communication) have a lower subjectivity score. The threshold used in our study is a subjectivity score of 0.5. If the score is greater than or equal to 0.5, the post is labeled as "subjective," otherwise the post is labeled as "objective." For nontemporal analyses, we performed sensitivity analyses to verify the results by excluding duplicated texts (ie, similar text posted by different users) and conducted the sentiment and emotion analyses on 1000 randomly selected posts (Figures S3 and S4 in Multimedia Appendix 1).

Trends in Public Perception

Changes in sentiments were tracked over time. The Mann-Kendall trend test was applied to identify significant trends in monthly incidence. Temporally weighted analyses were used to identify any changes in users' discussions and comments regarding BCI technologies over time.

Topic Modeling

To better understand the context of topics within the complex discussion regarding BCI in X, we used the BERTopic tool [28] to conduct a topic modeling analysis. BERTopic is a topic modeling technique that uses deep learning and NLP to better understand the context and semantic relationships within text data generating clusters of similar texts, interpreted as topics. It works by embedding text into numerical representations, which are converted into a high-dimensional vector using a pretrained transformer model. Text or sentences related to each other will be close to each other in this vector, while unrelated ones will be away from each other. After dimensionality reduction of the vector, the lower-dimensional embeddings are clustered to group similar documents together. Each cluster represents a potential topic. Finally, each topic is represented by its most representative documents or by extracting keywords that best describe the cluster. Results from this topic modeling analysis include the frequency of certain words within a topic and the probability that given words represent certain topics. We preprocessed the data by removing duplicate posts, then used the "all-MiniLM-L6-v2" embedding model and the SentenceTransformer model to analyze topics in the posts discussing BCI. The top 8 topics discussed in our dataset are visualized in topic word score bar charts, including specific topics for posts expressing positive or negative sentiments alone (Figure S4 in Multimedia Appendix 1)

Statistical Analysis

All analyses were conducted using Python (version 3.8; Python Software Foundation) and R (version 4.4.0; R Foundation for Statistical Computing). The *P* value threshold for statistical significance in this study is set at .05.

Ethical Considerations

All extracted data used and presented in this study were archival, cross-sectional, and observational, obtained from publicly

XSL•FO
RenderX

accessible sources without any interaction with social media users and with their usernames omitted. As such, institutional review board approval was not required.

Results

Post Characteristics

Our study analyzed a total of 65,340 posts, created by 38,962 distinct users (Table 1). These users had a median follower count of 662 (IQR 173-2332). The median number of total posts per user was 8976.5 (IQR 2294-33,677). Most of the posts came from a diverse user base, as only 5.21% (3405/65,340) originated from the top 50 most active contributors. The content of these posts varied, as 60,079 (91.94%) included links, while 5838 (8.93%) featured media. A total of 16,617 (25.43%) posts contained tags, and 18,623 (28.5%) posts contained hashtags. Engagement metrics showed that 17,141 (26.23%) posts received at least 1 like, 5104 (7.81%) posts had at least 1 reply, 12,210 (18.68%) posts were reposted at least once, and 2688 (4.11%) posts contained quote posts. Post characteristics from 1000 randomly selected posts and a text duplicate–free sample are showcased in Table 1.

Table . Overview of post characteristics with complete dataset and validation subsets in BCI discussions on X.

Post characteristics	Total posts (n=65,340)[a], n (%)	Duplicate texts removed (n=39,990)[b], n (%)	Random 1000 posts[c], n (%)
At least 1 repost	12,210 (18.68)	8710 (21.78)	170 (17)
At least 1 reply	5104 (7.81)	4203 (10.51)	78 (7.8)
At least 1 like	17,141 (26.23)	12,640 (31.60)	255 (25.5)
At least 1 quote	2688 (4.11)	1849 (4.62)	32 (3.2)
Contains a link	60,079 (91.94)	35,020 (87.57)	913 (91.3)
Contains a media	5838 (8.93)	4218 (10.45)	80 (8)
Contains a tagging	16,617 (25.43)	12,122 (30.31)	254 (25.4)
Contains a hashtag	18,623 (28.50)	13,336 (33.34)	281 (28.1)

[a]The number of total unique users: 38,962.

[b]The number of unique users of the duplicate-free subset: 25,008.

[c]The number of unique users of the random 1000 posts subset: 959.

User Biography

Biographic analysis (unique users=34,565) showed that “broadcasting” (10,171/34,565, 29.42%) was the highest group, followed by “entrepreneurship” (9359/34,565, 27.02%), “scientific” (9200/34,565, 26.61%), “clinical” (4066/34,565, 11.76%), and “other” (1769/34,565, 5.11%). The contribution to the BCI discussion (total posts=58,030) was the highest from users in the “broadcasting” group with 17,803 (30.67%) posts, followed by “scientific” with 16,005 (27.58%) posts, “entrepreneurship” with 14,008 (24.13%) posts, “clinical” with 7380 (12.71%) posts, and “other” with 2834 (4.88%) posts (Table 2). The “scientific” group had the highest engagement metrics among all the biography groups, while the “entrepreneurship” group had the lowest engagement metrics in our dataset (Table 3).

Table . Distribution of unique users and total posts by biographic group in brain-computer interface discussions on X.

Biographic group	Unique users[a] (n=34,565), n (%)	Total posts[a] (n=58,030), n (%)
Broadcasting	10,171 (29.42)	17,803 (30.67)
Entrepreneurship	9359 (27.02)	14,008 (24.13)
Scientific	9200 (26.61)	16,005 (27.58)
Clinical	4066 (11.76)	7380 (12.71)
Other	1769 (5.11)	2834 (4.88)

[a]Users with empty user biography were excluded from the analysis.

XSL•FO
RenderX

Table . Post characteristics and engagement metrics across biographic groups in brain-computer interface discussions on X.

Post characteristics	Broadcasting[a], n (%)	Entrepreneurship[b], n (%)	Scientific[c], n (%)	Clinical[d], n (%)
At least 1 repost	3549 (19.93)	2456 (17.53)	3675 (22.96)	1601 (21.69)
At least 1 reply	1596 (8.96)	986 (7.03)	1511 (9.44)	545 (7.38)
At least 1 like	4822 (27.08)	3571 (25.49)	5119 (31.98)	2266 (30.70)
At least 1 quote	934 (5.24)	482 (3.44)	831 (5.19)	286 (3.87)
Contains a link	16,307 (91.59)	13,058 (93.21)	14,568 (91.02)	6913 (93.67)
Contains a media	1628 (9.14)	1264 (9.02)	1600 (9.99)	730 (9.89)
Contains a tagging	4128 (23.18)	3757 (26.82)	4766 (29.77)	1834 (24.85)
Contains a hashtag	4990 (28.02)	4655 (33.23)	4687 (29.28)	2289 (31.01)

[a]17,803 (30.67%) posts.
[b]14,008 (24.13%) posts.
[c]16,005 (27.58%) posts.
[d]7380 (12.71%) posts.

Engagement Trends

In 2017, there was a substantial increase in the number of posts, which accounted for approximately 24.52% of the entire dataset (Figure 1). March had the highest number of posts (n=3686), followed by April (n=3094) (Figure 2). This rise in March's posts was primarily due to Elon Musk's announcement of "Neuralink," his BCI company [29]. The words "Neuralink," "Musk," and "Elon" were collectively mentioned 5831 times in March. The number of posts in April can be attributed to the announcement of Facebook's BCI projects, with the term "Facebook" receiving 1944 mentions [30]. A bar chart containing the 25 most frequently mentioned terms, excluding BCI-related terms is included in Figure S5 in Multimedia Appendix 1.

Figure 1. The number of posts shared on X discussing brain-computer interface (BCI) annually from 2010 to 2021. From 2010 to 2014, the number of posts remained relatively steady, fluctuating but less than 4000 per year. A gradual increase began in 2015, peaking sharply in 2017 at around 16,000 posts, marking the highest level of activity in the timeline. This spike coincides with the public announcements of Elon Musk's BCI company, Neuralink, and Facebook's BCI project. Following this peak, there was a substantial drop in 2018, with post numbers returning to earlier levels. From 2019 to 2021, the number of posts showed a fluctuating yet gradual increase.

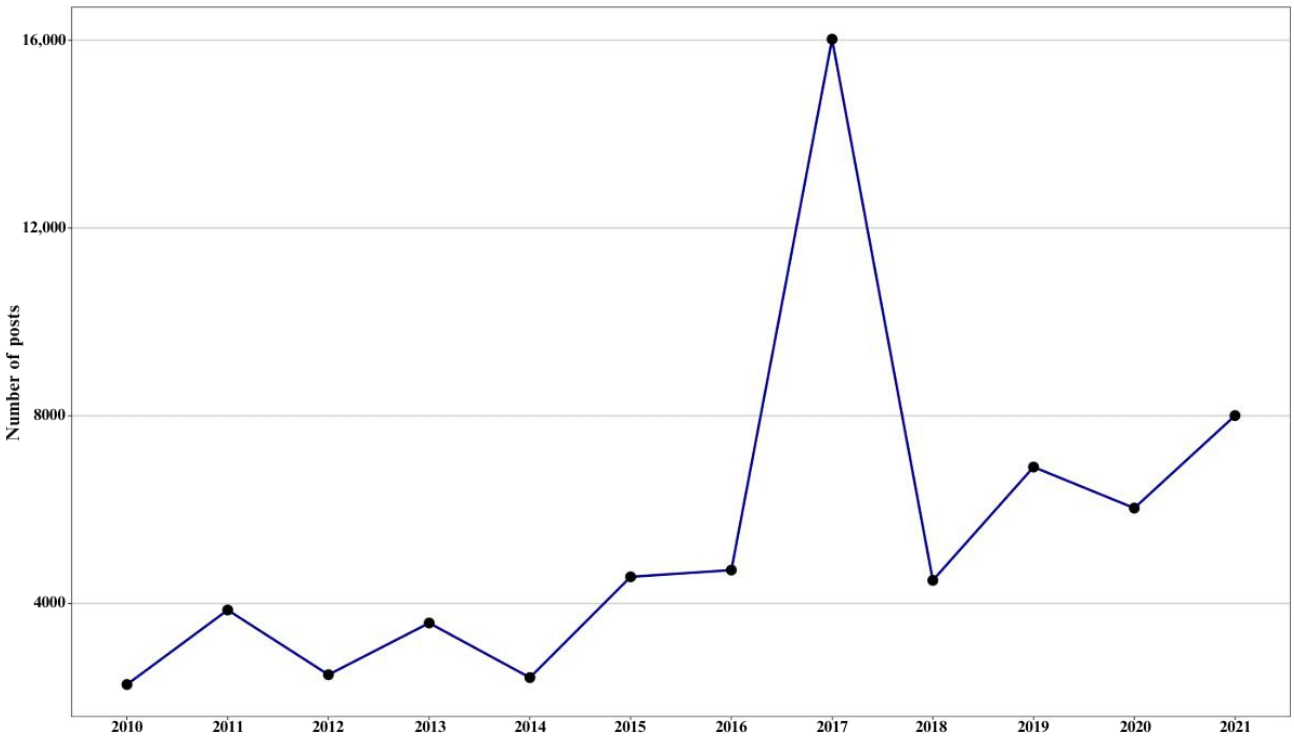

XSL•FO
RenderX

Figure 2. The number of posts shared on X discussing brain-computer interface (BCI) per month of the year 2017. The number of posts increases from January, peaking in March and April at over 3000 posts each month. The peak in March (a) coincides with Neuralink's public announcement, while the peak in April (b) aligns with Facebook's announcement of its BCI project. Following these peaks, there is a sharp decline in May, with the number of posts dropping to the lowest point in June. A slight increase is observed in July, but the number of posts remains lower than the earlier peaks, stabilizing at lower levels from September to December.

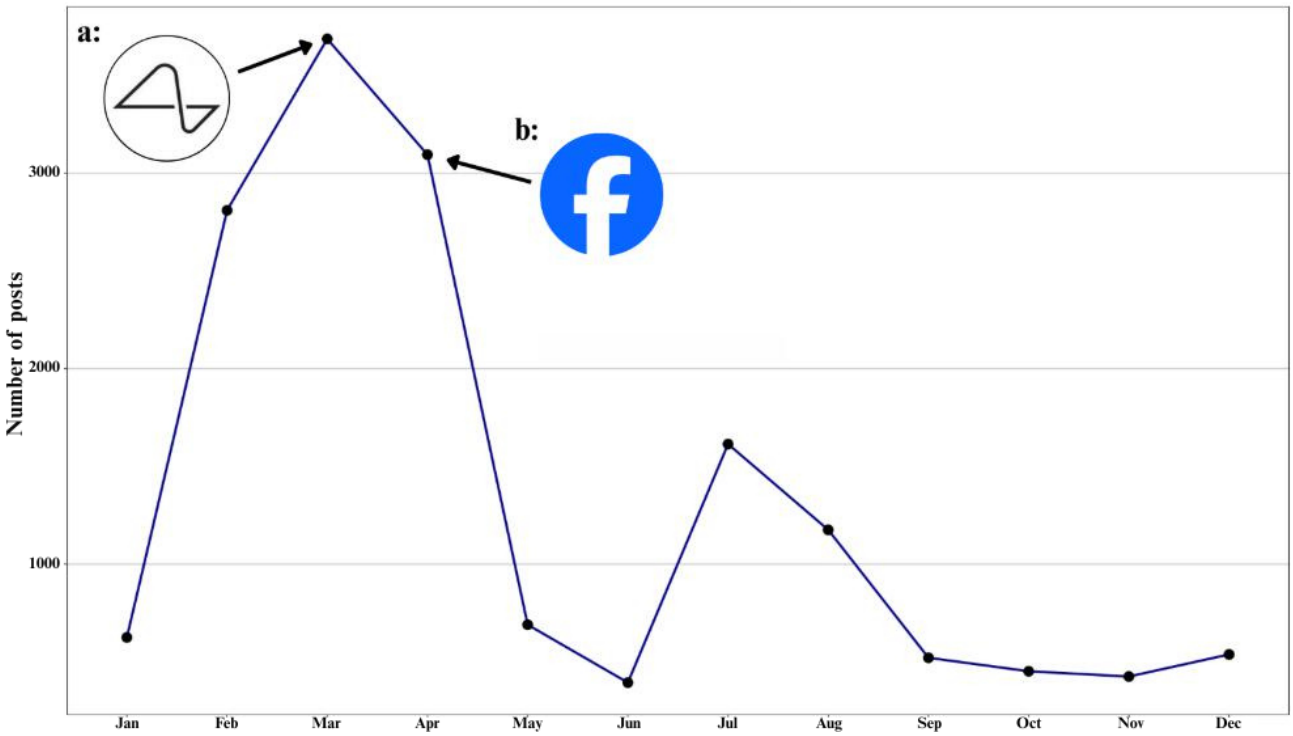

Sentiment Quantification and Trends in Public Perception

Most of the posts tended to express neutral sentiments (38,804/65,340, 59.38%). Approximately one-third of the posts conveyed positive sentiments (21,404/65,340, 32.75%), while a smaller portion expressed negative sentiments (5132/65,340, 7.85%). Deletion of duplicate text resulted in a notable polarization of sentiments, potentially due to the exclusion of posts containing only titles of news articles, which tend to be more neutral in tone (Table 4). A considerable increase in positive posts was observed in February 2017, coinciding with the publication of Stanford-led BCI studies, and in July 2017, following the Defense Advanced Research Projects Agency (DARPA) award for BCI [31-33]. The number of posts expressing negative sentiment remained low, with an average of 428 (8.19%) posts per year throughout the study period (Figure 3; Figure S6 in Multimedia Appendix 1). The average polarity score showed an overall positive trend throughout the study (Mann-Kendall Statistic=0.266; τ=0.266; P<.001; Figure S7 in Multimedia Appendix 1). The mean sentiment score increased substantially upon the announcement of the collaboration between Stanford University and BrainGate in November 2011, in April 2016 following the publication of the study on the use of BCI in restoring functional movement in apatient with quadriplegia, and again in July 2017 with the DARPA-led award [33-36]. The spike in negative sentiment posts in March 2017 was mostly due to discussion related to Musk's involvement in BCI, with an excited yet conservative tone. The word "help" (n=1489 mentions) was common in posts expressing positive sentiment, whereas in posts with negative sentiment, the term "injury" (n=325 mentions) was the most mentioned word. The frequent mentions of Elon Musk and artificial intelligence (AI) in the positive and negative sentiments are suggestive of mixed and polarized opinions about him and his BCI company "Neuralink" (Figure S8 in Multimedia Appendix 1). Most of the posts were objective (50,847/65,340, 77.81%), while fewer were subjective (14,393/65,340, 22.02%; Table 4).

The most prevalent emotion scores observed were "anticipation," "trust," and "fear," accounting for 20.52% (10,802/52,618), 17.56% (9244/52,618), and 13.95% (7344/52,618) of the expressions, respectively. However, "surprise," "joy," and "disgust" were less frequently expressed, constituting 8.77% (4619/52,618), 12.69% (6681/52,618), and 2.79% (1470/52,618), respectively (Table 5). The emotion score of the text duplicate–free dataset and random 1000 posts are in Table 5. Focusing on the 3 most prominent emotions, "anticipation" peaked at 47.49% in 2010, coinciding with the sharing of news regarding the potential gamification of BCI technologies [37]. In July 2017, there was a substantial increase in posts expressing anticipation as their primary emotion reaching as high as 48.07% of posts shared, discussions were mostly related to DARPA's award for BCI, and the use of BCI in the production of music using the "encephalophone" [33,38,39]. "Trust" reached 23.43% of total posts in 2021 and was the lowest in 2016 with 3.57% of total posts. "Fear" displayed a notable increase in posts in 2016, with 25.90%, and was the lowest in 2013 with 3.34% (Figure 4; Figure S9 in Multimedia Appendix 1).

XSL•FO
RenderX

Table . Overview of sentiment and subjectivity in brain-computer interface discussions on X across the complete dataset and validation subsets.

	Total posts (n=65,340), n (%)	Duplicate texts removed (n=39,990), n (%)	Random 1000 posts, n (%)
Sentiment			
Positive sentiment	21,404 (32.75)	15,390 (38.48)	334 (33.4)
Neutral sentiment	38,804 (59.38)	20,924 (52.32)	591 (59.1)
Negative sentiment	5132 (7.85)	3676 (9.1)	75 (7.5)
Subjectivity			
Subjective	14,393 (22.02)	10,571 (26.4)	230 (23)
Objective	50,847 (77.81)	29,419 (73.5)	770 (77)

Figure 3. The annual number of brain-computer interface–related posts on X from 2010 to 2021, categorized by sentiment. Positive sentiment posts (green line) increased significantly from 2016, peaking in 2017 at over 3500 posts, followed by a decline in 2018 and subsequent growth through 2021. Negative sentiment posts (red line) remained relatively steady throughout the period, showing minor fluctuations but consistently lower numbers compared to positive sentiment posts. Overall, positive sentiment posts were more prevalent than negative ones across all years.

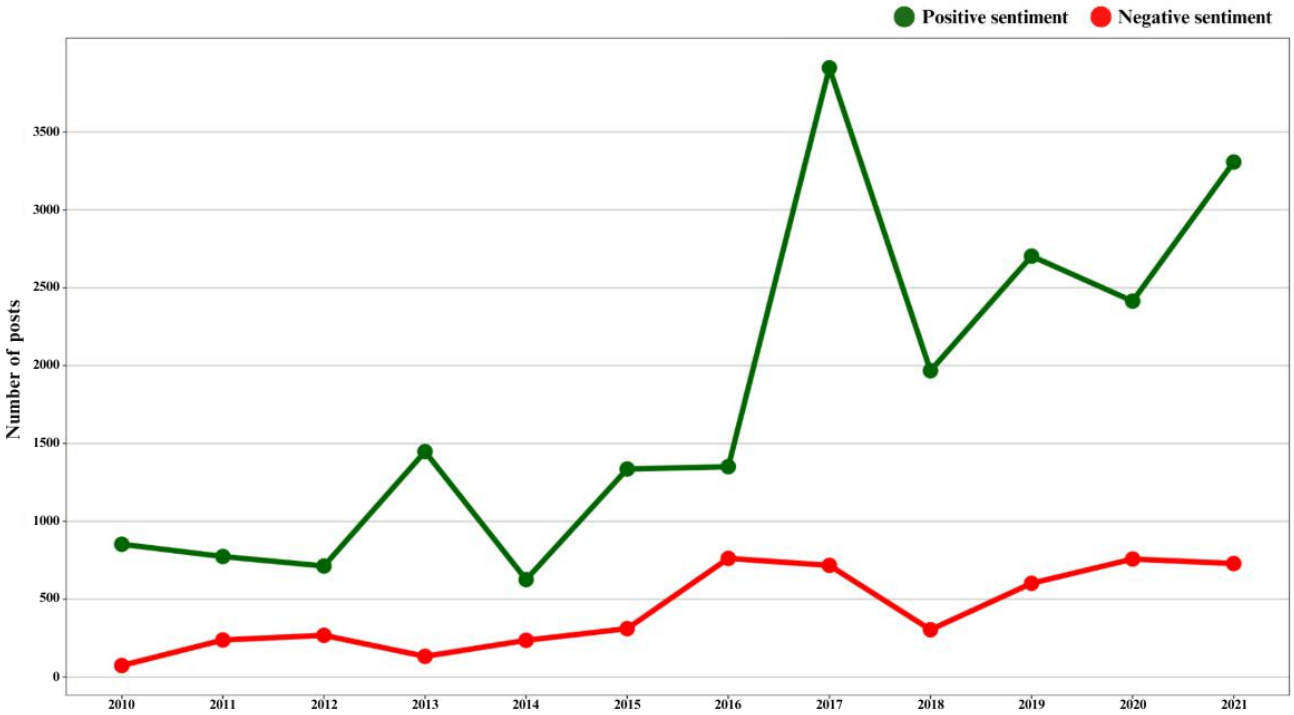

XSL•FO
RenderX

Table . Distribution of positive and negative emotions in brain-computer interface discussions on X across the complete dataset and validation subsets.

Emotion	Total posts, score[a,b] (%)	Duplicated text removed[c,b], score (%)	Random 1000 posts,[b,d] score (%)
Positive emotions			
Anticipation	10,802 (20.52)	7855 (20.52)	454 (22.07)
Trust	9244 (17.56)	7536 (19.69)	386 (18.76)
Joy	6681 (12.69)	5150 (13.45)	265 (12.88)
Surprise	4619 (8.77)	3261 (8.52)	174 (8.45)
Negative emotions			
Fear	7344 (13.95)	5083 (13.28)	274 (13.32)
Sadness	6623 (12.58)	4402 (11.50)	244 (11.86)
Anger	5745 (10.91)	3646 (9.52)	185 (8.99)
Disgust	1470 (2.79)	1317 (3.44)	74 (3.59)

[a]Total score, excluding posts expressing no emotion and neutral sentiment: 52,618

[b]Emotions conveyed equally with similar scores are considered separately.

[c]Total score, excluding posts expressing no emotion and neutral sentiment: 38,268

[d]Total score, excluding posts expressing no emotion and neutral sentiment: 2057

Figure 4. The annual trend of primary emotions—trust, anticipation, and fear—expressed in posts discussing brain-computer interface on X from 2010 to 2021. It indicates fluctuations in the presence of each emotion over time. Anticipation (light blue line) shows varied peaks and troughs throughout the timeline, with prominent peaks in 2010, 2013, and 2017. Trust (dark blue line) displays a more consistent pattern with smaller fluctuations and an increase toward the end of the period. Fear (purple line) remains relatively stable over the years, with slight peaks in 2012, 2015, and 2016, but generally shows lower percentages compared to anticipation and trust.

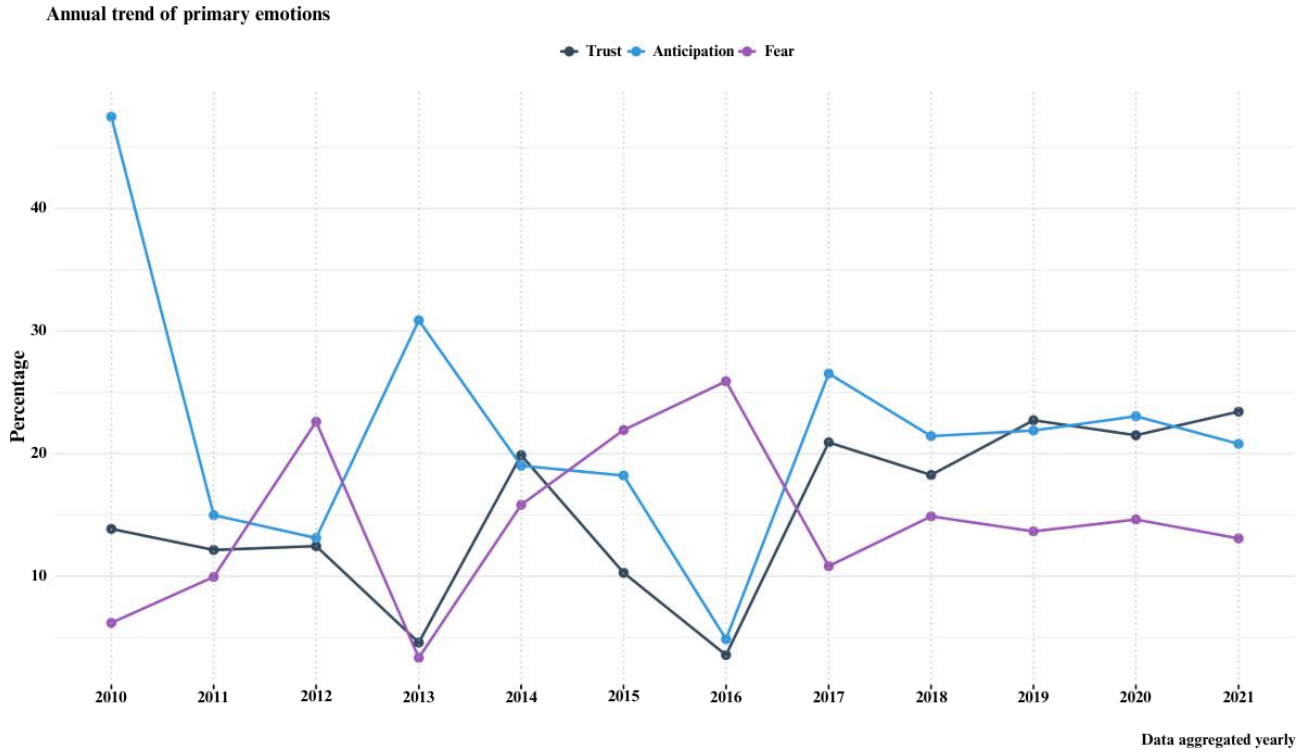

Temporally Weighted Sentiment Analyses

A temporally weighted analysis was conducted to explore the change in public discussion before, during, and after the spike in 2017 (Table 6). The increase in posts expressing a neutral sentiment (11,394/16,023, 71.11%), as well as the rise in posts expressing anticipation (1417/2512, 56.41%) and trust (n=1117/2512, 44.47%) in 2017, demonstrates a public that is excited yet cautious toward this emerging technology. However, after 2017, the public discussion polarized, as positive and negative views increased substantially, from 24.41% (3912/16,023) to 40.86% (10,391/25,430) and from 4.47% (717/16,023) to 9.41% (2393/25,430), respectively. Additionally, there was a notable increase in subjective posts, from 18.95% (3036/16,023) to 26.76% (6804/25,430), after 2017. Fear was substantially expressed before 2017 with 40.97% (1936/4725),

XSL•FO
RenderX

then dropping to below 30% during and after 2017, possibly indicating the public excitement, and trust following the big announcements in the BCI industry in 2017. However, the increase from 23.01% (578/2512) to 28.3% (2237/7904) in fear might indicate unaddressed ethical questions.

Table . Sentiment, subjectivity, and primary emotions over time in brain-computer interface discussions on X.

	Before 2017 (n=23,887), n (%)	During 2017 (n=16,023), n (%)	After 2017 (n=25,430), n (%)
Sentiment			
Positive	7101 (29.73)	3912 (24.41)	10,391 (40.86)
Neutral	14,764 (61.81)	11,394 (71.11)	12,646 (49.73)
Negative	2022 (8.46)	717 (4.47)	2393 (9.41)
Subjectivity			
Objective	19,234 (80.52)	12,987 (81.05)	18,626 (73.24)
Subjective	4653 (19.48)	3036 (18.95)	6804 (26.76)
Emotions [a,b]			
Anticipation	1977 (41.84)	1417 (56.41)	3490 (44.15)
Trust	913 (19.32)	1117 (44.47)	3506 (44.36)
Fear	1936 (40.97)	578 (23.01)	2237 (28.30)

[a]Posts expressing no primary emotion were excluded from the analysis. Emotions conveyed equally with similar scores are considered separately; hence, percentages exceeding 100%.

[b]Before 2017: 4725 posts, during 2017: 2512 posts, and after 2017: 7904 posts.

Topic Modeling

We conducted a topic modeling analysis to understand the semantic and dynamic discussion related to BCI technology. The main topics discussed in our dataset included ideas related to BCI changing the future, Musk's announcement of "Neuralink," Facebook's involvement in BCI technology, and the gamification of BCI technologies (Figure S10 in Multimedia Appendix 1). Topics discussed in posts expressing positive sentiment included the advancement of BCI in research and health care and BCI advancements with virtual reality (VR) and augmented reality (AR; Figure S11 in Multimedia Appendix 1). Discussions in posts conveying negative sentiment were related to conspiracy theories of Musk or "Neuralink," further demonstrating that the involvement of Musk in BCI was although highly influential but also particularly controversial. Moreover, our analysis detected discussions related to the potential use of BCI for malicious reasons including mind control (Figure S12 in Multimedia Appendix 1). Examples of posts related to different topics are included in Table S2 in Multimedia Appendix 1.

Discussion

Principal Findings

Our study highlights the complex and evolving public perception of BCI technology, as expressed on X discussions over the past decade. Sentiment analysis reveals that while a majority of the public holds neutral views, there is a notable undercurrent of anticipation, trust, and fear. The presence of these conflicting emotions suggests that while there is optimism about the potential of BCIs, concerns around ethical issues remain unresolved. Understanding these emotions is essential for the development and acceptance of BCI technology. Key topics that emerged from the discussions include Neuralink and Elon Musk—subjects of both positive and negative conversations—the practical applications of BCIs, and the potential for gamification of BCI technology. In addition to being key market drivers, addressing these factors will influence innovation and drive investments in the BCI space.

Ethical Challenges

BCI is one of the most important technologies in neuroscience, sparking diverse opinions regarding its ethical implications [40-42]. Safety, justice, privacy, and security are the main concerns discussed in ethical literature [43]. Some authors had concerns over the safety of implanting such invasive technologies, which could potentially cause serious health complications [44], while others have criticized the biased narrative in the BCI literature, which tends to overlook the perspectives of individuals with disabilities [45]. There was unease about the handling and storage of sensitive data obtained through BCI, with worries about its potential exploitation for malicious purposes [46]. In our study, we detected discussions reflecting many of these concerns. We specifically highlighted concerns regarding Elon Musk and his company Neuralink, particularly related to safety and animal rights issues [47], as well as the potential for BCI to be used in terrorism, or for brain hacking, which could lead to the leaking of sensitive personal information or even revealing emotions and thoughts. Open and inclusive discussions are essential to guiding the ethical development and use of BCI technology. Moreover, users from the scientific community will play a particularly critical role in this process. Our study found that posts from scientists generated the highest engagement when discussing BCI, underscoring their influence in shaping public understanding. By leveraging their trusted position, scientists can educate and foster informed

XSL•FO
RenderX

discourse on these ethical issues to steer the responsible development and application of BCI technology.

Elon Musk's Involvement

The significant rise in digital discussions about BCI in X, particularly in March 2017, aligns with Elon Musk's announcement of his BCI company, Neuralink. Elon Musk's substantial social media influence is evident, with almost 200 million followers on X. His posts not only are influential but often polarize public opinion [48]. For instance, his comments on cryptocurrencies have demonstrably impacted their market values [49]. Additionally, a sentiment analysis focused on electric vehicles revealed that Musk was a central figure in these discussions [50]. These observations reinforce the significant role of influential figures like Elon Musk in steering the public dialogue and development of cutting-edge technologies.

Most posts conveyed a neutral sentiment, reflecting a combination of anticipation and excitement, yet mixed with fear and anxiety, resulting in a generally doubtful public perception. This perception could potentially be the result of unanswered ethical challenges. Furthermore, Musk's involvement in BCI technologies has significantly influenced public perceptions, often polarizing opinions. The 2023 sentiment analysis conducted on Reddit indicates that neutral and negative views toward Elon Musk outnumber positive sentiments in discussions about him [48]. Additionally, in our study, a common theme in both the negative and positive discussions about BCI frequently centered around Musk or Neuralink. This mixed perception of Elon Musk could be another contributing factor to the sentiment results of our study. Ensuring transparent communication from influential figures like Musk might be key to positively shifting public sentiment toward BCI technologies.

Since acquiring Twitter, now rebranded as X, Elon Musk has become actively involved in geopolitical discourse, leading to conflicting opinions about his involvement in politics. This increased political engagement has further polarized public perception of him [51,52], which might potentially extend to BCI technologies. To mitigate this migration of polarization, it is essential to diversify the voices in BCI communication by encouraging scientists and industry experts to take a more prominent role, thereby reducing the focus on any single individual.

Nonetheless, Musk's involvement in BCI technology has had a significant and positive impact, contributing to greater public awareness and advancements in the field. The surge in BCI-related discussions in 2017, following Musk's announcement of Neuralink, not only increased awareness of the technology and its potential applications but also led to a rise in posts expressing positive sentiments, mixed with anticipation and trust. This suggests a general optimism regarding the potential of BCI technology. Furthermore, this trust and enthusiasm is reflected in the recent growth of BCI-related publications as well as the rapid expansion of the BCI market [53,54]. In summary, even though Musk is considered a controversial figure by many, his influence in the field of BCI has been largely positive. However, it is important to continue making conscious efforts to ensure that this positive impact is sustained and that the technology develops responsibly and ethically.

Bridging Public Sentiment and Practical Applications

We found that anticipation and trust were the most expressed emotions, likely stemming from the involvement of renowned universities in BCI research, including the collaborative BrainGate2 project [34,35]. The widespread media coverage and remarkable BCI outcomes in individuals with severe disabilities have also contributed to this trend [31,32,36]. These technologies enable patients with disability to overcome their disabilities, thereby enhancing their quality of life. The public enthusiasm observed toward these milestones underscores a substantial societal demand for such innovations. This heightened positive sentiment indicates that the integration of BCIs into neuroprosthetics, AI-driven prosthetics, and exoskeleton devices not only represents a significant technological advancement but also fulfills a critical public need for effective solutions to address disabilities. Moreover, BCI technology used for nonrehabilitation activities, such as gaming and music production, led to an increase in anticipation due to its potential appeal and application to everyday consumers.

Fear was the third most prevalent emotion identified. This may be related to ongoing unanswered ethical concerns in BCI development, varying public opinions about public figures, and inherent fear toward novel technologies. Our capture of fear might be an early indicator of potential technophobia, as BCI becomes more commercially available. Difficult to understand and complex technologies are associated with more anxiety and fear of using them [55]. In addition, the concept of "mind control," frequently portrayed in popular culture and science fiction, could potentially heighten fear perceptions. Upcoming depictions and future works that continue to explore this theme may trigger more apprehensive emotions. Such emotion, if not addressed, could foster a conservative mindset, potentially slowing the adoption and hindering the application of innovative technologies such as BCI.

The Gamification of BCI Technologies

One particularly exciting avenue for BCI technology is its integration into gaming. The potential of gamification in BCI technologies is vast, offering applications not only for medical rehabilitation [56,57] but also for entertainment experiences [58-60]. The concept of controlling and fully immersing oneself in a game to create experiences beyond physical limitations is highly appealing to the public, as evidenced by a notable increase in anticipation. Gabe Newell, cofounder and president of Valve, showcased Valve's interest in BCI for gaming, sparking significant enthusiasm [61]. Integrating BCI technologies into gaming could play an important role in further advancing their development, much like the early adoption of VR and AR through gaming platforms such as the HTC Vive (by HTC and Valve Corporation), Meta Quest (by Meta Platforms), and PlayStation VR (by Sony Interactive Entertainment). These technologies, which began in gaming, have since expanded into broader applications in healthcare and education.[62]. Fueled by public enthusiasm, several companies have begun claiming that their products leverage BCI to enhance gaming experiences. However, it remains debatable whether

XSL•FO
RenderX

we have truly reached that point [11]. Numerous attempts have been made to use BCI for gaming, using both invasive and minimally invasive technologies. For example, Neuralink has successfully implemented a BCI in a human with quadriplegia, allowing the individual to play chess and other games by moving a cursor [63]. "I basically have an aimbot in my head," said Noland Arbaugh, the first patient to receive a brain-computer chip from Neuralink. An "aimbot" is a type of cheating software used in video games that automatically aims at opponents at a superhuman speed [64,65]. Similarly, a popular Twitch streamer successfully beat the notoriously difficult game *Elden Ring* using a wearable electroencephalograpy-based BCI device, with thousands of live viewers [66]. With a global video game market size estimated at US $217.06 billion in 2022, these advancements are likely to positively influence public perception of BCI and may attract the attention of major players in the gaming industry, potentially driving further research and development of BCI technologies [67]. Gamification thus presents a significant opportunity for the future of BCIs, acting as a gateway for mainstream consumer adoption, similar to how VR or AR technologies gained traction.

Public Acceptance and Trust: Key Drivers for the Growing BCI Market

The economic outlook for BCIs is promising. In 2023, the global market for BCIs was valued at US $2.0 billion and is projected to grow at a compound annual growth rate of 17.8% from 2024 to 2030. This growth is driven by increasing demand for neuroprosthetic devices and advancements in technology that enhance mobility and communication for patients who are paralyzed as well as expanding applications in gaming and military communication [53]. However, public opinion and acceptance are essential to sustaining this market growth. Greater acceptance of the technology will drive further adoption and development, especially since young investors and future philanthropists are likely to support companies that align with their personal values and contribute to humanitarian goals [68]. Our findings align with this broader market analysis. Most individuals in our study exhibit a neutral sentiment toward BCI technology, reflecting a phase of critical appraisal of this emerging innovation. Additionally, the prominence of emotions such as trust, anticipation, and fear underscores the imperative to educate, demystify, and familiarize the public with BCI technology. Doing so will be key to enhancing acceptance among the public. Additionally, the increase in the number of subjective posts in our analysis reflects increased awareness of BCI as a topic for discussion, and this growing awareness suggests that the public will begin to form their own ideas, opinions, and emotions toward BCI. This awareness might also indicate that people are shifting from passive observers to active participants in BCI discussions, contributing to the evolving conversation and potentially influencing its trajectory. Therefore, a concerted effort to inform and engage the public is crucial for fostering positive sentiment and ensuring the responsible development and integration of BCI technology. This, in turn, will support the continued expansion of the global BCI market.

Limitations

The study has several limitations that may have impacted the accuracy of our results. First, the search excluded the term "BCI" to avoid including irrelevant posts, which may have inadvertently omitted relevant content. Additionally, the analysis was restricted to English-language posts and focused solely on the platform X, potentially overlooking important discussions on other platforms or in different languages. It was not possible to infer some of the important demographics of users missing out on important contextual information. We used existing tools for NLP analyses that were not tailored and validated specifically for the study's topic, which may potentially lead to inaccurate results. Determining the geographical locations of users was not feasible, which would have provided valuable regional insights. Regarding the biography analysis, some overlap between categories is expected, as a single user may fall into multiple classifications. This method of short bioclassification is effective for only a subset of users, and further validation is required in future studies. Finally, the substantial number of posts containing links may indicate that the sample is not representative of general discussions but primarily consists of shared news articles. Future research should address these limitations to enhance the depth and accuracy of the understanding of the public perception toward BCI technologies.

Conclusions

The findings from this NLP-assisted study offer a decade-long overview of public perception of BCI technology. Overall sentiment was mostly neutral, but the emotions most commonly linked to BCI—anticipation, trust, and fear—reflect a complex emotional response, suggesting a cautiously optimistic yet apprehensive attitude toward the advancement of BCI technology. Notably, the presence of fear underscores the importance of addressing ethical concerns and ensuring transparent communication within the BCI field. Resolving these issues is critical for reassuring the public and mitigating apprehensive attitudes. Additionally, the involvement of influential figures and leading institutions, along with reports of positive clinical outcomes, such as advancements in neuroprosthetics and rehabilitation, may foster more favorable public perceptions of BCI technology. The gamification of BCI, particularly its integration into gaming and entertainment, also offers a pathway to increased public engagement and adoption. However, the persistence of fear signals potential resistance that may impede progress if left unaddressed. Prioritizing ethical transparency, expanding public education, and incorporating a more diverse range of voices in the discussion could help drive broader acceptance and responsible use of BCI. Although this study uses advanced AI tools to offer valuable insights into public sentiment, certain limitations, such as potential biases and incomplete demographic data, should be acknowledged. Nonetheless, the findings serve as a valuable reference point for guiding future technological developments, informing policy making, and crafting effective communication strategies within the BCI sector.

XSL•FO
RenderX

Acknowledgments

The authors acknowledge the use of ChatGPT-4o (OpenAI, 2024) for grammatical checks, language refinement, and assistance in writing the codes used for analysis during the preparation of this manuscript. The tool was not used to generate any original text or ideas.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

MA Almanna, LME, JJL, MA Alvi, and NMA contributed to the acquisition, analysis, and interpretation of data. Statistical analysis was performed by MA Almanna and MA Alvi. Administrative, technical, and material support was provided by MA Almanna. NMA supervised the study. All authors were involved in the conception and design of the study as well as in drafting the manuscript and critically revising it. All authors reviewed the final manuscript.

Conflicts of Interest

LME was affiliated with Neuralink, a brain-computer interface company. However, this study was conducted independently, without involvement from Neuralink or its personnel. The other authors report no conflicts of interest.

Multimedia Appendix 1
Additional diagrams, tables, and figures.
[DOCX File, 1106 KB - formative_v9i1e60859_app1.docx]

References

1. Wolpaw JR, Birbaumer N, McFarland DJ, Pfurtscheller G, Vaughan TM. Brain–computer interfaces for communication and control. Clin Neurophysiol 2002 Jun;113(6):767-791. [doi: 10.1016/S1388-2457(02)00057-3]
2. Sun Y, Chen X, Liu B, et al. Signal acquisition of brain-computer interfaces: a medical-engineering crossover perspective review. Fundam Res 2025 Jan;5(1):3-16. [doi: 10.1016/j.fmre.2024.04.011] [Medline: 40166113]
3. He Q, Yang Y, Ge P, et al. The brain nebula: minimally invasive brain-computer interface by endovascular neural recording and stimulation. J Neurointerv Surg 2024 Nov 22;16(12):1237-1243. [doi: 10.1136/jnis-2023-021296] [Medline: 38388478]
4. Mane R, Chouhan T, Guan C. BCI for stroke rehabilitation: motor and beyond. J Neural Eng 2020 Aug 17;17(4):041001. [doi: 10.1088/1741-2552/aba162] [Medline: 32613947]
5. Daly JJ, Cheng RC, Hrovat K, Litinas KH, Rogers JM, Dohring ME. Development and testing of non-invasive BCI FES/robot system for use in motor re-learning after stroke. Presented at: Proceedings of the 13th International Functional Electrical Stimulation Society; Sep 24-28, 2008; Germany URL: https://www.academia.edu/28915386/Development_and_Testing_of_Non_Invasive_BCI_FES_Robot_Sys_tem_For_Use_in_Motor_Re_Learning_After_Stroke?sm=b [accessed 2025-06-10]
6. Daly JJ, Wolpaw JR. Brain-computer interfaces in neurological rehabilitation. Lancet Neurol 2008 Nov;7(11):1032-1043. [doi: 10.1016/S1474-4422(08)70223-0] [Medline: 18835541]
7. Hochberg LR, Serruya MD, Friehs GM, et al. Neuronal ensemble control of prosthetic devices by a human with tetraplegia. Nature New Biol 2006 Jul 13;442(7099):164-171. [doi: 10.1038/nature04970] [Medline: 16838014]
8. Pfurtscheller G, Guger C, Müller G, Krausz G, Neuper C. Brain oscillations control hand orthosis in a tetraplegic. Neurosci Lett 2000 Oct 13;292(3):211-214. [doi: 10.1016/s0304-3940(00)01471-3] [Medline: 11018314]
9. Galán F, Nuttin M, Lew E, et al. A brain-actuated wheelchair: asynchronous and non-invasive brain-computer interfaces for continuous control of robots. Clin Neurophysiol 2008 Sep;119(9):2159-2169. [doi: 10.1016/j.clinph.2008.06.001] [Medline: 18621580]
10. Tanaka K, Matsunaga K, Wang HO. Electroencephalogram-based control of an electric wheelchair. IEEE Trans Robot 2005;21(4):762-766. [doi: 10.1109/TRO.2004.842350]
11. Cattan G. The use of brain–computer interfaces in games is not ready for the general public. Front Comput Sci 2021;3:20. [doi: 10.3389/fcomp.2021.628773]
12. Yeo PS, Nguyen TN, Ng MPE, et al. Evaluation of the implementation and effectiveness of community-based brain-computer interface cognitive group training in healthy community-dwelling older adults: randomized controlled implementation trial. JMIR Form Res 2021 Apr 27;5(4):e25462. [doi: 10.2196/25462] [Medline: 33904819]
13. Musk E, Neuralink. An integrated brain-machine interface platform with thousands of channels. J Med Internet Res 2019 Oct 31;21(10):e16194. [doi: 10.2196/16194] [Medline: 31642810]
14. Kennedy PR, Bakay RA. Restoration of neural output from a paralyzed patient by a direct brain connection. Neuroreport 1998 Jun 1;9(8):1707-1711. [doi: 10.1097/00001756-199806010-00007] [Medline: 9665587]

XSL•FO
RenderX

15. Georgopoulos AP, Schwartz AB, Kettner RE. Neuronal population coding of movement direction. Science 1986 Sep 26;233(4771):1416-1419. [doi: 10.1126/science.3749885] [Medline: 3749885]
16. Laubach M, Wessberg J, Nicolelis MA. Cortical ensemble activity increasingly predicts behaviour outcomes during learning of a motor task. Nature New Biol 2000 Jun 1;405(6786):567-571. [doi: 10.1038/35014604] [Medline: 10850715]
17. Taylor DM, Tillery SIH, Schwartz AB. Direct cortical control of 3D neuroprosthetic devices. Science 2002 Jun 7;296(5574):1829-1832. [doi: 10.1126/science.1070291] [Medline: 12052948]
18. Margalit E, Weiland JD, Clatterbuck RE, et al. Visual and electrical evoked response recorded from subdural electrodes implanted above the visual cortex in normal dogs under two methods of anesthesia. J Neurosci Methods 2003 Mar 15;123(2):129-137. [doi: 10.1016/s0165-0270(02)00345-x] [Medline: 12606062]
19. Lyu JC, Han EL, Luli GK. COVID-19 vaccine-related discussion on Twitter: topic modeling and sentiment analysis. J Med Internet Res 2021 Jun 29;23(6):e24435. [doi: 10.2196/24435] [Medline: 34115608]
20. Monselise M, Chang CH, Ferreira G, Yang R, Yang CC. Topics and sentiments of public concerns regarding COVID-19 vaccines: social media trend analysis. J Med Internet Res 2021 Oct 21;23(10):e30765. [doi: 10.2196/30765] [Medline: 34581682]
21. Swathi T, Kasiviswanath N, Rao AA. Retracted article: an optimal deep learning-based LSTM for stock price prediction using Twitter sentiment analysis. Appl Intell 2022 Sep;52(12):13675-13688. [doi: 10.1007/s10489-022-03175-2]
22. Ding Y, Korolov R, (Al) Wallace W, Wang X(. How are sentiments on autonomous vehicles influenced? An analysis using Twitter feeds. Transportation Research Part C: Emerging Technologies 2021 Oct;131:103356. [doi: 10.1016/j.trc.2021.103356]
23. Kober SE, Buchrieser F, Wood G. Neurofeedback on Twitter: evaluation of the scientific credibility and communication about the technique. Heliyon 2023 Aug;9(8):e18931. [doi: 10.1016/j.heliyon.2023.e18931] [Medline: 37600360]
24. Wiseman B. Sentiment.ai: simple sentiment analysis using deep learning. R Project. 2022. URL: https://cran.r-project.org/web/packages/sentiment.ai/index.html [accessed 2025-05-31]
25. Hutto CJ, Gilbert EE. VADER: a parsimonious rule-based model for sentiment analysis of social media text. ICWSM ;8(1):216-225. [doi: 10.1609/icwsm.v8i1.14550]
26. Mohammad SM, Turney PD. Emotions evoked by common words and phrases: using mechanical turk to create an emotion lexicon. Presented at: Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text; Jun 1-6, 2010; Los Angeles, CA p. 26-34 URL: https://aclanthology.org/W10-0204/ [accessed 2025-06-10]
27. Loria S. TextBlob: simplified text processing. TextBlob. URL: https://textblob.readthedocs.io/ [accessed 2025-05-31]
28. Grootendorst M. BERTopic: neural topic modeling with a class-based TF-IDF procedure. arXiv. Preprint posted online on Mar 11, 2022. [doi: 10.48550/arXiv.2203.05794]
29. Winkler R. Elon musk launches Neuralink to connect brains with computers. The Wall Street Journal. 2017 Mar 27 URL: https://www.wsj.com/articles/elon-musk-launches-neuralink-to-connect-brains-with-computers-1490642652 [accessed 2025-05-18]
30. Strickland E. Facebook announces “typing-by-brain” project. IEEE Spectrum. 2017 Apr 20. URL: https://spectrum.ieee.org/facebook-announces-typing-by-brain-project
31. Pandarinath C, Nuyujukian P, Blabe CH, et al. High performance communication by people with paralysis using an intracortical brain-computer interface. Elife 2017 Feb 21;6:e18554. [doi: 10.7554/eLife.18554] [Medline: 28220753]
32. Goldman B. Brain-computer interface advance allows fast, accurate typing by people with paralysis. Wu Tsai Neurosciences Institute, Stanford University. 2017. URL: https://neuroscience.stanford.edu/news/brain-computer-interface-advance-allows-fast-accurate-typing-people-paralysis [accessed 2025-05-31]
33. Sulleyman A. DARPA to plug computers into brains to let machines talk directly to people. The Independent. 2017. URL: https://www.independent.co.uk/tech/darpa-computer-plug-in-brains-lost-vision-hearing-speech-restore-us-military-r-n-d-research-neural-engineering-system-design-a7835131.html [accessed 2025-04-19]
34. BrainGate2: feasibility study of an intracortical neural interface system for persons with tetraplegia (NCT00912041). ClinicalTrials.gov. 2023. URL: https://clinicaltrials.gov/ct2/show/NCT00912041 [accessed 2023-12-30]
35. Henderson J. Stanford joins BrainGate team developing brain-computer interface to aid people with paralysis. Stanford Medicine News Center. 2011. URL: https://med.stanford.edu/news/all-news/2011/11/stanford-joins-braingate-team-developing-brain-computer-interface-to-aid-people-with-paralysis.html [accessed 2025-05-18]
36. Bouton CE, Shaikhouni A, Annetta NV, et al. Restoring cortical control of functional movement in a human with quadriplegia. Nature New Biol 2016 May 12;533(7602):247-250. [doi: 10.1038/nature17435] [Medline: 27074513]
37. Mulroy J. Brain-computer interface can help you get in the game. PCWorld. 2023. URL: https://www.pcworld.com/article/499690/brain_computer_helps_get_you_in_the_game.html [accessed 2025-04-19]
38. Deuel TA, Pampin J, Sundstrom J, Darvas F. The encephalophone: a novel musical biofeedback device using conscious control of electroencephalogram (EEG). Front Hum Neurosci 2017;11:213. [doi: 10.3389/fnhum.2017.00213] [Medline: 28491030]

XSL•FO
RenderX

39. Chen A. You can play this musical instrument with just your thoughts. The Verge. 2017. URL: https://www.theverge.com/2017/7/13/15965836/encephalophone-musical-instrument-thought-brain-computer-interface-neuroscience [accessed 2025-04-19]
40. Clausen J. Man, machine and in between. Nature New Biol 2009 Feb 26;457(7233):1080-1081. [doi: 10.1038/4571080a] [Medline: 19242454]
41. Nijboer F, Clausen J, Allison BZ, Haselager P. The Asilomar survey: stakeholders' opinions on ethical issues related to brain-computer interfacing. Neuroethics 2013;6(3):541-578. [doi: 10.1007/s12152-011-9132-6] [Medline: 24273623]
42. Demetriades AK, Demetriades CK, Watts C, Ashkan K. Brain-machine interface: the challenge of neuroethics. Surgeon 2010 Oct;8(5):267-269. [doi: 10.1016/j.surge.2010.05.006] [Medline: 20709284]
43. Burwell S, Sample M, Racine E. Ethical aspects of brain computer interfaces: a scoping review. BMC Med Ethics 2017 Nov 9;18(1):60. [doi: 10.1186/s12910-017-0220-y] [Medline: 29121942]
44. Wolpaw JR, Loeb GE, Allison BZ, et al. BCI meeting 2005—workshop on signals and recording methods. IEEE Trans Neural Syst Rehabil Eng 2006 Jun;14(2):138-141. [doi: 10.1109/TNSRE.2006.875583] [Medline: 16792279]
45. Wolbring G, Diep L, Yumakulov S, Ball N, Leopatra V, Yergens D. Emerging therapeutic enhancement enabling health technologies and their discourses: what is discussed within the health domain? Healthcare (Basel) 2013 Jul 25;1(1):20-52. [doi: 10.3390/healthcare1010020] [Medline: 27429129]
46. Li KY. Get out of my head: an examination of potential brain-computer interface data privacy concerns. Boston Coll Intellect Prop Technol Forum 2018;2018:1-12.
47. Levy R. Exclusive: Musk's Neuralink faces federal probe, employee backlash over animal tests. Reuters. 2022. URL: https://www.reuters.com/technology/musks-neuralink-faces-federal-probe-employee-backlash-over-animal-tests-2022-12-05 [accessed 2025-04-19]
48. Bonilla M, Iradukunda S, Thomas P. Sentiment analysis of public perception towards Elon Musk on Reddit (2008-2022). Cardinal Edge 2023;1(3):7.
49. Ante L. How Elon Musk's Twitter activity moves cryptocurrency markets. Technol Forecast Soc Change 2023 Jan;186:122112. [doi: 10.1016/j.techfore.2022.122112]
50. Ruan T, Lv Q. Public perception of electric vehicles on Reddit and Twitter: a cross-platform analysis. Transportation Research Interdisciplinary Perspectives 2023 Sep;21:100872. [doi: 10.1016/j.trip.2023.100872]
51. Metz C, Satariano A, Che C. How Elon Musk became a geopolitical chaos agent. The New York Times. URL: https://www.nytimes.com/2022/10/26/technology/elon-musk-geopolitics-china-ukraine.html [accessed 2025-04-19]
52. Zehn MJ. Elon Musk: a new antagonist in brazilian politics? Brazilian Research and Studies Center. URL: https://bras-center.com/elon-musk-a-new-antagonist-in-brazilian-politics/ [accessed 2025-04-19]
53. Brain computer interface market size, share & trends analysis report by product (invasive, non-invasive), by application (healthcare, communication & control), by end use, by region, and segment forecasts, 2024 - 2030. Grand View Research. URL: https://www.grandviewresearch.com/industry-analysis/brain-computer-interfaces-market [accessed 2025-04-19]
54. Yin Z, Wan Y, Fang H, et al. Bibliometric analysis on brain-computer interfaces in a 30-year period. Appl Intell 2023 Jun;53(12):16205-16225. [doi: 10.1007/s10489-022-04226-4]
55. Osiceanu ME. Psychological implications of modern technologies: "technofobia" versus "technophilia". Procedia - Social and Behavioral Sciences 2015 May;180:1137-1144. [doi: 10.1016/j.sbspro.2015.02.229]
56. de Castro-Cros M, Sebastian-Romagosa M, Rodríguez-Serrano J, et al. Effects of gamification in BCI functional rehabilitation. Front Neurosci 2020;14:882. [doi: 10.3389/fnins.2020.00882] [Medline: 32973435]
57. Gao C, Xia M, Zhang Z, Han Y, Gu Y. Improving the Brain-Computer Interface Learning Process with Gamification in Motor Imagery: A Review. In: IntechOpen eBooks 2022, Vol. 21. [doi: 10.5772/intechopen.105715]
58. Universal brain-computer interface lets people play games with just their thoughts. Cockrell School of Engineering. 2024. URL: https://cockrell.utexas.edu/news/archive/9841-universal-brain-computer-interface-lets-people-play-games-with-just-their-thoughts [accessed 2025-04-19]
59. Kumar S, Alawieh H, Racz FS, Fakhreddine R, Millán JDR. Transfer learning promotes acquisition of individual BCI skills. PNAS Nexus 2024 Feb 1;3(2):2024. [doi: 10.1093/pnasnexus/pgae076]
60. Hubner D, Schall A, Tangermann M. Two player online brain-controlled chess. Annu Int Conf IEEE Eng Med Biol Soc 2019 Jul;2019:3018-3021. [doi: 10.1109/EMBC.2019.8856965] [Medline: 31946524]
61. Porter J. Gabe Newell has big plans for brain-computer interfaces in gaming. The Verge. 2021. URL: https://www.theverge.com/2021/1/25/22248202/gabe-newell-valve-brain-computer-interface-bci-meat-peripherals [accessed 2025-04-19]
62. Hamad A, Jia B. How virtual reality technology has changed our lives: an overview of the current and potential applications and limitations. Int J Environ Res Public Health 2022 Sep 8;19(18):11278. [doi: 10.3390/ijerph191811278] [Medline: 36141551]
63. Musk's Neuralink shows first brain-chip patient playing online chess. Reuters. 2024. URL: https://www.reuters.com/business/healthcare-pharmaceuticals/neuralink-shows-first-brain-chip-patient-playing-online-chess-2024-03-21 [accessed 2025-04-19]
64. Tangermann V. First Neuralink patient says implant has given him incredible gaming skills. Futurism. 2024. URL: https://futurism.com/neoscope/first-neuralink-patient-gaming-skills [accessed 2025-04-19]

XSL•FO
RenderX

65. PowerfulJRE. Joe Rogan experience #2167—Noland Arbaugh. YouTube. URL: https://www.youtube.com/watch?v=Bfo2xIeaOAE [accessed 2025-04-19]
66. Bardhan A. Twitch streamer plays Elden Ring using only her brain. Kotaku. 2023. URL: https://kotaku.com/twitch-streamer-elden-ring-play-brain-eeg-perrikaryal-1850024234 [accessed 2025-04-19]
67. Video game market size, share & trends analysis report by device (console, mobile, computer), by type (online, offline), by region (Asia Pacific, North America, Europe), and segment forecasts, 2023 - 2030. Grand View Research. URL: https://www.grandviewresearch.com/industry-analysis/video-game-market [accessed 2025-04-19]
68. The future of philanthropy. Fidelity Charitable. 2021. URL: https://www.fidelitycharitable.org/content/dam/fc-public/docs/resources/2021-future-of-philanthropy-summary.pdf [accessed 2025-04-19]

Abbreviations

AI: artificial intelligence
AR: augmented reality
BCI: brain-computer interface
DARPA: Defense Advanced Research Projects Agency
NLP: natural language processing
VADER: Valence Aware Dictionary and Sentiment Reasoner
VR: virtual reality

Edited by A Mavragani; submitted 23.05.24; peer-reviewed by D Carvalho, Y Gogna; revised version received 15.11.24; accepted 15.01.25; published 25.06.25.

Please cite as:
Almanna MA, Elkaim LM, Alvi MA, Levett JJ, Li B, Mamdani M, Al-Omran M, Alotaibi NM
Public Perception of the Brain-Computer Interface Based on a Decade of Data on X: Mixed Methods Study
JMIR Form Res 2025;9:e60859
URL: https://formative.jmir.org/2025/1/e60859
doi:10.2196/60859

© Mohammed A Almanna, Lior M Elkaim, Mohammed A Alvi, Jordan J Levett, Ben Li, Muhammad Mamdani, Mohammed Al-Omran, Naif M Alotaibi. Originally published in JMIR Formative Research (https://formative.jmir.org), 25.6.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Changes in Mental State for Help-Seekers of Lifeline Australia's Online Chat Service: Lexical Analysis Approach

Kelly Mazzer[1]; Sonia Curll[1]; Hakar Barzinjy[1]; Roland Goecke[2]; Mark Larsen[3]; Philip J Batterham[4]; Nickolai Titov[5]; Debra Rickwood[1]

[1]Faculty of Health, University of Canberra, 11 Kirinari Street, Bruce, Canberra, Australia
[2]School of Systems & Computing, University of New South Wales, Canberra, Australia
[3]Centre for Big Data Research in Health, University of New South Wales, Sydney, Australia
[4]Centre for Mental Health Research, Australian National University, Canberra, Australia
[5]School of Psychological Sciences, Macquarie University, Sydney, Australia

Corresponding Author:
Kelly Mazzer
Faculty of Health, University of Canberra, 11 Kirinari Street, Bruce, Canberra, Australia

Abstract

Background: Mental health challenges are escalating globally, with increasing numbers of individuals accessing crisis helplines through various modalities. Despite this growing demand, there is limited understanding of how crisis helplines benefit help-seekers over the course of a conversation. Affective computing has the potential to transform this area of research, yet it remains relatively unexplored, partly due to the scarcity of available helpline data.

Objective: This study aimed to explore the feasibility of using lexical analysis to track dynamic changes in the mental state of help-seekers during online chat conversations with a crisis helpline.

Methods: Lexical analysis was conducted on 6618 deidentified online chat transcripts collected by Lifeline Australia between April and June 2023 using the validated Empath lexical categories of Positive Emotion, Negative Emotion, Suffering, and Optimism. Furthermore, 2 context-specific categories, Distress and Suicidality, were also developed and analyzed to reflect crisis support language. Correlation analyses evaluated the relationships between the 6 lexical categories. One-way ANOVAs assessed changes in each lexical category across 3 conversation phases (beginning, middle, and end). Trend analyses using regression modeling examined the direction and strength of changes in lexical categories across 9 overlapping conversation windows (20% size and 50% step overlap).

Results: Significant changes were observed across conversation phases. The context-specific categories showed the strongest improvements from the beginning to end phase of conversation, with a large reduction in Distress (d=0.79) and a moderate reduction in Suicidality (d=0.49). The most frequently occurring terms representing Distress were "hard," "bad," and "down," and for Suicidality were "suicide," "stop," and "hurt." The negatively framed Empath categories also significantly reduced, with moderate effect sizes for Suffering (d=0.49) and Negative Emotion (d=0.39). There were also significant but small reductions in the positively framed Empath categories of Positive Emotion (d=0.15) and Optimism (d=0.07) from the beginning to end phase of conversation. Correlation coefficients indicated the lexical categories captured related but distinct constructs (r=.34 to r=0.82). Trend analyses revealed a consistent downward trajectory across most lexical categories. Distress showed the steepest decline (slope=−0.15, R^2=0.97), followed by Suffering (slope=−0.11, R^2=0.96), Negative Emotion (slope=−0.10, R^2=0.69), and Suicidality (slope=−0.06, R^2=0.88). Positive Emotion showed a slight negative trend (slope=−0.04, R^2=0.54), while Optimism remained relatively stable across the conversation windows (slope=0.01, R^2=0.13).

Conclusions: This study demonstrates the feasibility of using lexical analysis to represent and monitor mental state changes during online crisis support interactions. The findings highlight the potential for integrating affective computing into crisis helplines to enhance service delivery and outcome measurement. Future research should focus on validating these findings and exploring how lexical analysis can be applied to improve real-time support to those in crisis.

(JMIR Form Res 2025;9:e63257) doi:10.2196/63257

KEYWORDS

crisis helpline; lexical analysis; mental health; outcomes; distress; affective computing; Lexical; suicidal; suicide; help-seeker; help-seeking; emotion; chat; mental state; caregivers; digital mental health; digital health; e-health; ANOVA; feasibility study; mental health intervention; crisis support; online communities; support service; online support

XSL•FO
RenderX

Introduction

Background

Crisis helplines are a critical component of mental health care systems, offering immediate, confidential, free, and often 24/7 support to individuals experiencing emotional distress [1]. These services are delivered primarily via telephone, with many now offering text and web-based alternatives. Crisis supporters are usually staff members or volunteers trained in crisis and suicide intervention, with the skills and knowledge to provide support to help-seekers (also known as clients or users) and pathways to further care where needed. With the high prevalence of suicide deaths and emotional distress around the world, crisis helplines offer a cost-effective and scalable way to improve the accessibility and responsiveness of mental health and crisis care [1,2].

Crisis helplines face significant challenges in meeting the growing demand and diversity of help-seekers. Maintaining consistent, high-quality support across telephone and digital services is vital [2], especially as digital services are often used by vulnerable groups including youth and people with disabilities [3,4]. Crisis helplines also need to ensure they remain flexible and adapt quickly to the evolving communication preferences and needs of help-seekers, as highlighted during the recent COVID-19 pandemic [5]. These challenges, compounded by the urgency of crisis support, exacerbate the pressure on a crisis helpline's resources and volunteer workforce [6-8].

Conducting research in crisis helpline settings poses multiple challenges, many of which stem from the anonymous and one-off nature of these services [2,9,10]. Most studies have relied on retrospective self-reports by help-seekers [11], which can provide valuable insight into subjective experiences but are subject to recall biases and fail to capture fluctuations in mental state during the contact. Moreover, low completion rates (eg, 33% completion rate among chatters in one study [3]) indicate potential self-selection bias, whereby help-seekers with more positive outcomes might be more likely to complete a postcontact survey. Alternative approaches, such as crisis supporter assessments, are limited by judgement biases (eg, social desirability concerns) and recall errors, while using external raters tends to be very labor-intensive with consequently small samples [12]. Improved approaches are urgently needed to inform strategies aimed at enhancing service delivery [2,4].

Integrating affective computing approaches within the crisis helpline context presents an opportunity for a transformative shift from traditional research approaches and holds substantial promise for enhancing mental health interventions [13]. Affective computing is concerned with developing systems and devices that can recognize, interpret, process, and simulate human emotions [14]. These systems collect and analyze various data on users' mental states, including text-based emotional cues, vocal tone, and physiological signals. This information is then used by researchers and service providers to understand and enhance user experience in applications such as psychiatry, teaching, and social media [14].

In the helpline context, these advanced computational methods offer several ethical and practical advantages. Affective computing algorithms can be applied to large datasets, with the ability to detect meaningful patterns and insights difficult or impossible to obtain with traditional methods. Automated data collection techniques such as text or voice analysis can provide more representative and objective data than self-report of human-coded data by mitigating human judgment errors and biases (eg, recall, social desirability, and self-selection). They are also unobtrusive, imposing no additional burden on the help-seeker or crisis supporter. Crucially, affective computing techniques facilitate continuous assessment, opening possibilities for dynamic support tools and informing a deeper understanding of emotional responses and outcomes of accessing a crisis helpline [14].

Natural language processing (NLP) and computational linguistics play an important role in affective computing, especially in systems designed to process and understand emotions through text. These techniques can automatically analyze the words people use to provide insight into their mental states and emotions [15,16]. For instance, lexical analysis using existing and widely available lexicon-based software, such as Empath and Linguistic Inquiry and Word Count (LIWC) [17,18], has been applied to electronic health records to detect suicide risk [19] and to text-based transcripts from online therapy to predict depression symptom severity [20].

The past 5 years have seen a rapid growth in studies using NLP for mental health interventions [16,17] and it has now moved into the crisis helpline context, providing initial support for the development of NLP-based tools to provide adjunct assistance to crisis supporters. Recent research has demonstrated the usefulness of NLP to identify and classify self-harm or suicide risk among digital help-seekers [21,22]. Cognitive overload among digital crisis supporters has also been shown to reduce with support from NLP-based tools, including helping to resolve writer's block and providing real-time information based on conversation content [23,24]. Furthermore, NLP can support the efficiency of crisis support, with Althoff et al [25] determining via NLP techniques that more successful online crisis support involved greater time devoted to exploring solutions, as opposed to defining problems.

However, few studies so far have applied NLP to understand help-seeker outcomes from crisis support. This is an important area to explore, as developing innovative ways to monitor and evaluate help-seeker outcomes is essential to maintaining quality service provision and informing service improvements [10,13]. Althoff et al [25] applied LIWIC to explore changes in help-seeker sentiment, being the relative proportion of positive to negative words, they demonstrated a trend toward a more positive perspective over the conversation, with a notable increase at the very end of the conversation. Progressing this area of research requires access to data from service providers, which necessitates sensitively and appropriately navigating protective regulations concerning data confidentiality and ethical considerations of help-seeker privacy [16,26].

XSL•FO
RenderX

Current Study

This study is among the first to apply lexical analysis to explore changes in the frequency and intensity of language associated with mental states used by help-seekers over the duration of a single online chat conversation with a crisis helpline. Implementing lexical analysis overcomes many of the usual barriers to research in helplines. Specifically, it removes the need for human annotation, thereby eliminating the risks of annotator fatigue and bias, avoiding the resource heavy training of annotators, achieving greater efficiency and objectivity, and allowing for the analysis of a much larger amount of data [4,12,27]. This study aims to provide a proof-of-concept for using NLP to monitor the mental state of help-seekers. While some previous research has used NLP to explore broad changes in help-seekers' positive and negative sentiment [eg, 25], in the current study we analyze a wider range of general mental states (Positive Emotion, Negative Emotion, Suffering, and Optimism) and crisis-specific mental states (Distress and Suicidality). In addition, our novel use of Empath allowed us to go beyond frequency (word counts) to examine context and reflect intensity, allowing a richer understanding of changes in help-seeker mental states. Our analyses examined trends across the crisis intervention including over 9 overlapping conversation windows as well as at the beginning, middle and end phases using a large deidentified dataset from Lifeline Australia's online chat service.

Methods

Data

For more than 60 years, Lifeline has been operating as Australia's national, free 24-hour telephone crisis support service. In recent years, Lifeline has also expanded its service delivery to also offer 24-hour digital support via text messaging and online chat services with crisis supporters [6]. Lifeline Australia provided a census of routinely collected data from all contacts made to their online chat service for 3 months from April to June 2023, totaling 20,569 contacts of varying lengths that were answered by a crisis supporter. Consistent with previous research [12], after excluding automated or chatbot messages, all conversations with 10 or more messages from the help-seeker were retained for analysis (N=6618). Data included the date of contact, time of each message, number of messages per conversation, and the content of messages from both the help-seeker and crisis supporter during the conversation. Help-seeker demographics were not available.

Preprocessing

Lifeline Australia does not systematically collect identifiable information; however, any incidental identifying information contained in the messages were scrubbed before access and analysis. Data was divided into help-seeker or crisis supporter messages. Conversations were then split into equal thirds, based on total number of messages in the conversation, to create a beginning, middle, and end phase of each conversation. Conversations were also split into 9 overlapping conversation windows, using a 20% window size with a 50% step overlap, meaning that each subsequent window started at the midpoint of the previous window. Text was converted to lowercase and tokenized into individual words. The Porter Stemmer from The Natural Language Toolkit [28] was used to stem words by removing common morphological affixes and reducing words to their root forms, which enabled the capture of various word forms ("suicid*" to capture "suicide", "suicidal", "suicidality," etc). In addition, n-grams, specifically bigrams, was used to capture meaningful word pairs (2-word phrases) as tokens [29], which provided a more accurate representation of the text's semantic content (eg, "harm myself" as a single Suicidality term). All remaining messages from help-seekers were included in analysis; crisis supporter messages were not analyzed. Table 1 provides descriptive details of the dataset following preprocessing.

Table . Total counts, means, and SDs of help-seeker messages and lexical terms included in the final analyses for each online chat conversation (N=6618) and conversation phase (beginning, middle, and end).

Help-seeker data	Total count, n	Per conversation, mean (SD)	Per third[a], mean (SD)
Messages	201,955	30.52 (20.55)	10.17 (6.86)
Terms[b]	309,628	46.79 (26.67)	15.60 (10.83)

[a]Thirds were created based on number of messages.

[b]Terms represent key words used in conversation, also known as tokens.

Empath Software

Empath is an open-source Python (Python Software Foundation) library and text analysis tool that can perform NLP tasks including lexical analysis. Empath contains around 200 data-driven emotional and topical categories, also known as lexicons, which have been validated through a combination of NLP and human validation. Empath uses a large dataset to evaluate text and assign lexical degree scores based on the presence of predefined categories related to emotions, behaviors, and themes [17]. Empath lexicons are recognized as highly correlated to LIWC's gold-standard categories (r=0.91) [17,18]. Each category has a large list of member terms (words) that represent the category; for example, the category of Optimism includes terms such as "hopeful," "perseverance," and "progress" [17]. Refer to Table 2 for the top 10 frequently occurring member terms for each category used in the current study.

XSL•FO
RenderX

Table . Characteristics of the lexical categories used in the analyses, including source, top 10 terms, total terms, total occurrences, and means and SDs of terms per conversation (N=6618).

Category	Empath or contextual	Top 10 terms[a]	Total terms[b]	Total occurrences[c]	Mean (SD) per conversation
Negative Emotion	Empath	Want, think, see, hard, bad, care, stop, hurt, scary or scar*, and die	94	73,008	11.03 (7.09)
Positive Emotion	Empath	Feel*, friend*, better, keep, family or famili*, care, love, understand, hope, and happi*	75	43,666	6.60 (4.57)
Suffering	Empath	Feel*, bad, hurt, die, long, depress*, wors*, kill, pain*, and cry or cri*	127	42,125	6.37 (4.28)
Optimism	Empath	Feel*, like, will, thank, sure, better, love, hope*, happy or happi*, and appreci*	81	47,442	7.17 (4.20)
Distress	Contextual	Hard, bad, down, hurt, scary or scar*, struggl*, alon*, depress*, stress*, and wors*	137	53,423	8.07 (5.71)
Suicidality	Contextual	Suicid*, stop*, hurt, die, kill, hate, pain*, plan, harm*, and safe*	163	31,428	4.75 (3.98)

[a]Top 10 terms represents 10 most frequently occurring terms in the dataset for each category, ranked in order of frequency.

[b]Total terms are the total number of member terms or words representing a category. Count does not include member terms where a single term represents the name of another Empath category. For example, Death is an Empath category as well as a member term of Suffering; all of Death's 81 member terms would be identified as instances of Suffering, but are not included in the total terms value.

[c]Total occurrences are the total count of occurrences of member terms or words in conversations.

Lexical Categories

Empath Categories

All validated Empath categories were reviewed and 4 were selected for inclusion as they best represented concepts that help-seekers may aim to improve by contacting a crisis helpline like Lifeline. The 4 categories were: Negative Emotion, Positive Emotion, Suffering, and Optimism. Lexical degree scores, ranging from 0 to 1, were assigned to text for the Empath categories of Negative Emotion, Positive Emotion, Suffering, and Optimism. A higher score represents a greater proportion of words present in the text that fall into a category. For example, a degree score of 0.7 for Suffering would indicate a very high occurrence of Suffering-related words in the conversation.

Contextual Categories

There is a risk that lexicons designed to understand general linguistic patterns, such as the Empath categories, may not be appropriate for highly domain specific tasks [30]. Thus, in addition to the 4 Empath categories, 2 new categories were derived for analysis. The categories of Distress and Suicidality are highly context specific; reducing distress and suicidality are 2 of the highest priority outcomes for help-seekers accessing Lifeline services [31].

Lexical categories representing Distress and Suicidality specific to the crisis helpline context were developed using a multistep process with input from various expert groups, including researchers, service providers, and people with lived experience. First, a team (n=8) of experienced crisis supporters from Lifeline Australia reviewed a sample of 200 online chat and text transcripts from a separate dataset not used in the current study. These crisis supporters received a training session from the research team (KM) on how to identify relevant keywords. They were also provided with written guidelines and examples to ensure consistency in their approach. The keywords identified by the crisis supporters were then collated and reviewed by the research team (KM and SC). The resulting list was then shared with several expert groups for further input and refinement, including Lifeline's lived experience advisory group, online chat service team leaders, and Lifeline's clinical practice team. The research team conducted a final review of all inputs, with any disagreements or ambiguous terms resolved through team discussion. Further refinement of keywords was undertaken using manual annotation of a random sample of 100 transcripts used in the current study (SC).

The same Empath lexical degree scoring system could not be applied for Distress and Suicidality as they were not pre-existing Empath categories. Instead, the lexical analysis of the contextual categories, Distress and Suicidality, were based on mean counts of terms present within the text. The mean of Distress

XSL•FO
RenderX

represented the average number of occurrences of Distress member terms in a conversation.

The category of Distress had a total of 137 terms that occurred 53,423 times in the dataset; the top 10 Distress terms accounted for 40.26% (n=21,510) of these occurrences with “hard” being the most frequently occurring term (n=3118), followed by “bad” (n=2884), and “down” (n=2411).The Suicidality category had 163 total terms that occurred 31,428 times; the top 10 Suicidality terms accounted for 56.72% (n=17,825) of occurrences with “suicid*” occurring most frequently (n=2885), followed by “stop*” (n=2306), and “hurt” (n=2174).

Statistical Analysis

Pearson correlations were conducted based on the number of term occurrences to determine the strength of relationship between the 6 lexical categories. Lexical analyses were conducted to examine changes in mental state over the phases (beginning, middle, and end) of an online chat conversation with a help-seeker. One-way ANOVA was performed to determine whether the degree of occurrence of each category was significantly different between the start, middle, and end phases of conversation. Post hoc *t* tests were also conducted to determine which phases of conversation were significantly different from one another. Trend analysis and linear regression were then performed on the mean token counts across 9 overlapping conversation windows to determine the direction (slope) and magnitude (R^2) of change within each lexical category. This method allowed better understanding of the dynamic change across conversations. Figure 1 shows the workflow of analyses and included categories.

XSL•FO
RenderX

Figure 1. Overview of the design and structure of the study, including the lexical categories and conversation phases used in the analyses.

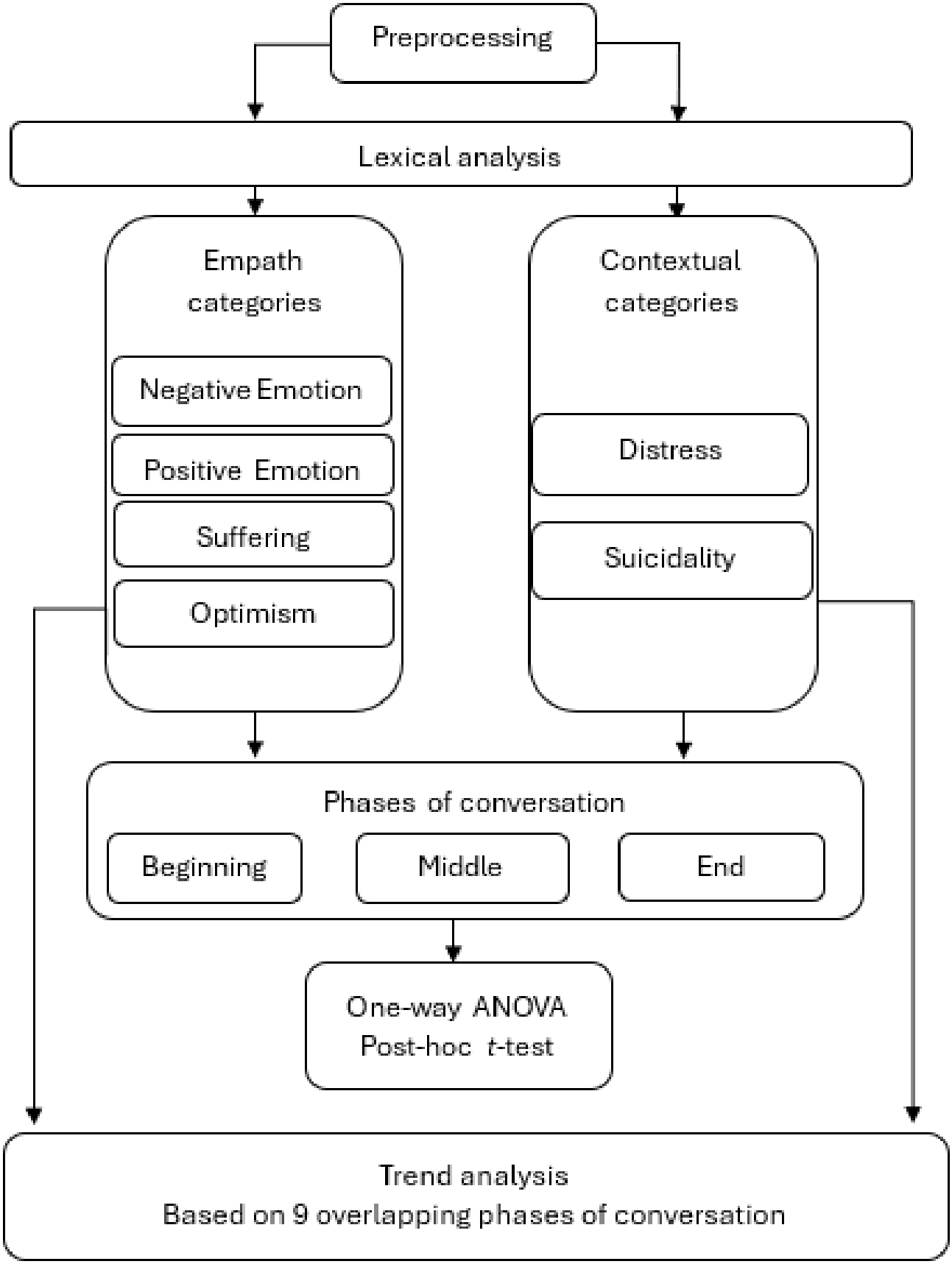

Ethical Considerations

This study was approved by the University of Canberra's Human Research Ethics Committee (approval no. 4673). The data that were analyzed were collected routinely by Lifeline Australia. Due to the anonymity of Lifeline Australia help-seekers, there was no opportunity to obtain specific consent from individuals to use their data. However, all Lifeline Australia help-seekers

XSL•FO
RenderX

are informed that their personal data may be used to conduct research, evaluation, and assurance activities. No identifying information (eg, phone number, email address, or help-seeker name) were provided in the dataset. Deidentified data were stored and analyzed in a secure, Lifeline owned and managed environment.

Results

Overview

Descriptive statistics of conversation length are presented, followed by correlations between all categories. Results for each of the Empath categories of Negative Emotion, Positive Emotion, Suffering and Optimism, are presented as means and SD of lexical degree scores for all conversations separated into the beginning, middle, and end phases of conversation. The contextual categories of Distress and Suicidality are presented as mean occurrences of terms across all conversations in the beginning, middle, and end phases. Tests of significance are reported for changes in each category over the phases of conversation.

Descriptive Statistics

Conversations included in the final analyses had a range of 10 to 382 help-seeker messages (mean 30.52, SD 20.55; Table 1). A total of 309,628 occurrences of terms from all included lexical categories were identified, with a mean of 46.79 (SD 26.67) occurrences of terms per conversation. Negative Emotion had the highest occurrences of terms at 73,008, followed by Distress with 53,423. Note that categories can and do include overlapping member terms; in fact, similarity comparisons are used in Empath's mapping of vocabulary to categories [17]. "Feel*" (feel, feels, feeling, and feelings) was among the most frequently occurring words for many of the emotion-based categories. Table 2 provides descriptions of the categories and their occurrence at the conversation level.

Correlations

Table 3 presents the correlations between mean number of occurrences across the categories. Positive Emotion and Optimism had the strongest relationship (r=0.82), whereby when a help-seeker expresses positive emotions they are also likely to use optimistic language. The next strongest correlations were between Distress with both Suffering (r=0.81) and Negative Emotion (r=0.77). These indicate that help-seekers experiencing high level of distress are also likely to express a lot of suffering and negative emotion. Correlations between all categories revealed significant, P<.001, medium, or strong positive relationships (r=0.34 to 0.82), likely reflecting the shared foundation of emotion and emotion-related terms that each of these categories is defined by. No categories had correlations higher than r=0.82, suggesting each category represented a distinct construct.

Table . A correlation matrix showing the relationships between mean lexical category occurrences in help-seeker online chat messages.

Lexical category[a]	1	2	3	4	5
Negative Emotion					
Positive Emotion	.71				
Suffering	.63	.68			
Optimism	.63	.82	.63		
Distress	.77	.67	.81	.64	
Suicidality	.62	.38	.58	.34	.45

[a]P<.001.

Changes in Emotions by Phase of Conversation

Figure 2 presents the mean occurrence of terms by category for the beginning, middle, and end phases of the conversation. Except for Optimism, all categories revealed a pattern of reduction in the number of term occurrences from the beginning phase of conversation to the end. Surprisingly, this includes the positively framed category of Positive Emotion, which would be expected to increase during a contact with Lifeline's online chat service.

XSL•FO
RenderX

Figure 2. Mean occurrences of lexical category terms across 3 phases of help-seeker online chat conversation (beginning, middle, and end).

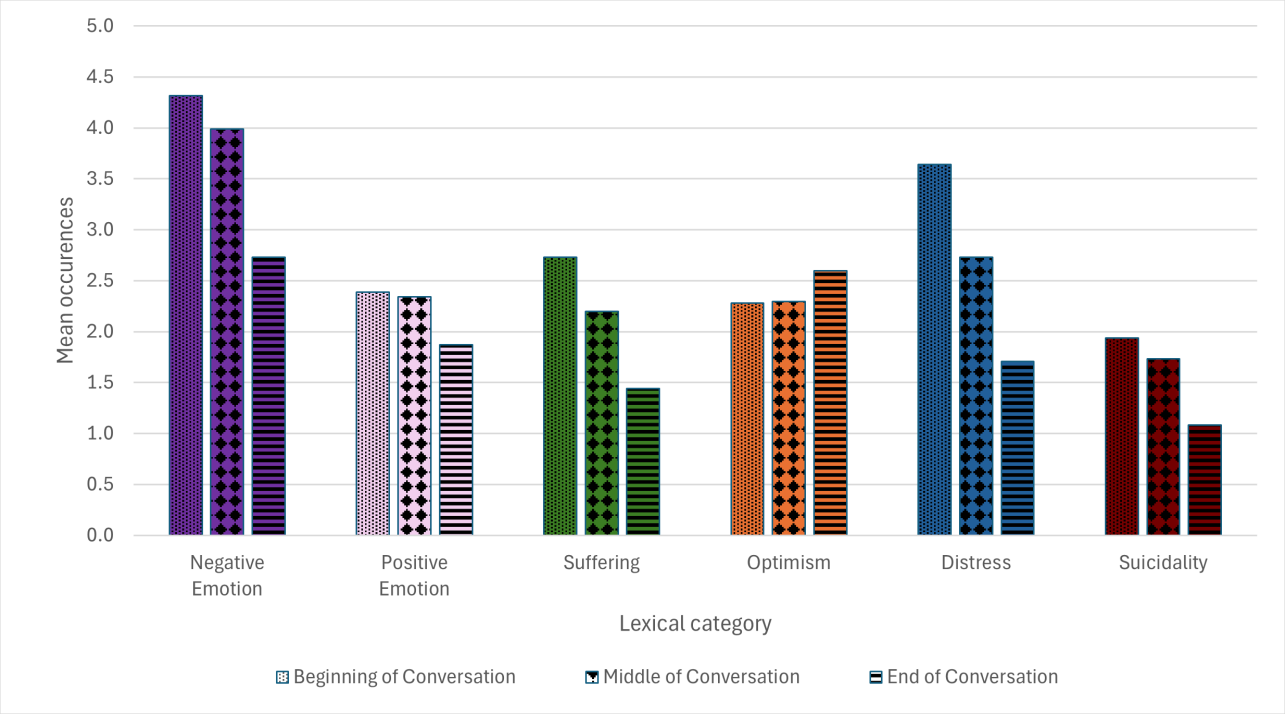

Lexical Analyses

Empath Categories by Conversation Phase

Table 4 presents the lexical degree scores for all 4 Empath categories by phase of conversation. Help-seekers most frequently used terms related to Negative Emotion across all 3 phases. Both negative categories, Negative Emotion ($F_{19851,2}$=273.68, P<.001) and Suffering ($F_{19851,2}$=424.30, P<.001) significantly improved from the beginning, to middle, to end phase of conversation with both demonstrating medium effect sizes from beginning to end of conversation (d=0.39 and d=0.49, respectively). Surprisingly, Positive Emotion ($F_{19851,2}$=54.42, P<.001) and Optimism ($F_{19851,2}$=70.49, P<.001) also significantly reduced in the end phase of conversation compared to the beginning, although only with weak effect sizes (d=0.15 and d=0.07, respectively). Note that the pattern of results based on mean occurrence was similar to the pattern of results based on mean lexical degree score for all categories except Optimism (Figure 2).

Table . One-way ANOVA and post hoc t tests comparing lexical degree scores across conversation phases (beginning, middle, and end) for Empath lexical categories.

Lexical category	F test (df)	Mean lexical degree score (SD)[b]			Cohen d		
		Beginning	Middle	End	Beginning-middle	Middle-end	Beginning-end
Negative Emotion	273.68[a] (19851,2)	0.0248[b] (0.0224)	0.0228[c] (0.0210)	0.0166[d] (0.0198)	0.09	0.31	0.39
Positive Emotion	54.42[a] (19851,2)	0.0129[b] (0.0153)	0.0107[c] (0.0130)	0.0106[c] (0.0147)	0.15	0.01	0.15
Suffering	424.30[a] (19851,2)	0.0134[b] (0.0162)	0.0090[c] (0.0131)	0.0065[d] (0.0118)	0.30	0.20	0.49
Optimism	70.49[a] (19851,2)	0.0077[b] (0.0121)	0.0054[c] (0.0094)	0.0068[d] (0.0120)	0.21	–0.13	0.07

[a]P<.001.

[b]Differing (b,c,d) superscripts represents differences at P<.001 between conversation phases. Where the same superscript is shown across row (b,b), the conversation phases did not differ significantly.

[c]Differing (b,c,d) superscripts represents differences at P<.001 between conversation phases. Where the same superscript is shown (c,c), the conversation phases did not differ significantly.

[d]Differing (b,c,d) superscripts represents differences at P<.001 between conversation phases.

XSL•FO
RenderX

Contextual Categories by Conversation Phase

Table 5 presents changes in Distress and Suicidality based on the mean number of occurrences of terms. Both contextual categories significantly improved from the beginning, to middle, to end phase of conversation (Distress $F_{19851,2}$=1338.07, P<.001 and Suicidality $F_{19851,2}$=421.45, P<.001). Distress had the strongest effect size of any category (Cohen d=0.79), in its reduction from the beginning to the end of the conversation. Suicidality was also among the strongest effects of all categories with Cohen d=0.49 from the beginning to the end of the conversation.

Table . One-way ANOVA and post hoc t tests comparing mean number of occurrences across conversation phases (beginning, middle, and end) for contextual lexical categories.

Category	F[a] test (df)	Mean (SD)[b]			Cohen d		
		Beginning	Middle	End	Beginning-middle	Middle-end	Beginning-end
Distress	1338.07[a] (19851,2)	3.64[b] (2.89)	2.73[c] (2.45)	1.70[d] (1.93)	0.34	0.47	0.79
Suicidality	421.45[a] (19851,2)	1.93[b] (1.98)	1.73[c] (1.81)	1.08[d] (1.46)	0.11	0.40	0.49

[a]P<.001.

[b]Differing (b,c,d) superscripts represents differences at P<.001 between conversation phases. Where the same superscript is shown acros row (b,b), the conversation phases did not differ significantly.

[c]Differing (b,c,d) superscripts represents differences at P<.001 between conversation phases.

[d]Differing (b,c,d) superscripts represents differences at P<.001 between conversation phases.

Trend Analyses

Regression results indicated a consistent downward trajectory for Distress (slope=−0.1471, R^2=0.9741), Suicidality (slope=−0.0636, R^2=0.8821), Negative Emotion (slope=−0.1017, R^2=0.6937), and Suffering (slope=−0.1053, R^2=0.9644), suggesting a steady decline in the expression of these categories over the course of the conversations (Figure 3). Among these, Distress maintained the steepest decline, aligning with findings from previous analyses demonstrating significant reductions in distress-related language.

Figure 3. Trend of mean occurrences for negative lexical categories (Negative Emotion, Distress, Suffering, Suicidality) across 9 overlapping conversation windows.

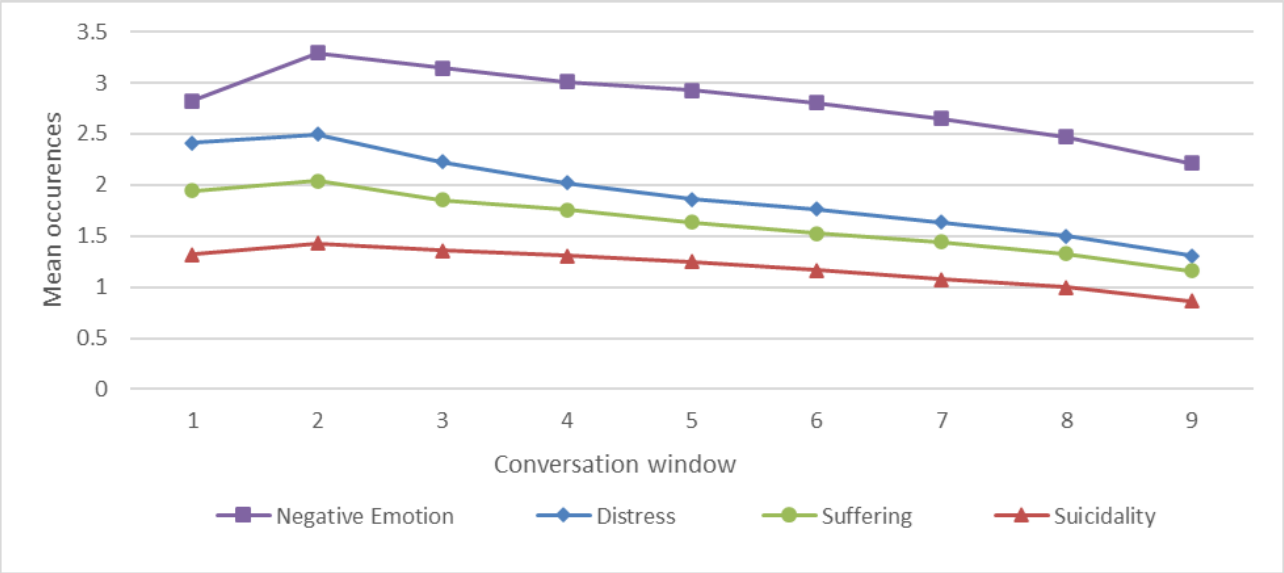

Positive Emotion also displayed a slight negative trend (slope=−0.0359, R^2=0.5372), but its decline was less pronounced compared with the negatively framed categories. On the other hand, optimism remained relatively stable, with a weak positive slope (slope=0.0080, R^2=0.1253), which indicates minimal variation across the conversation windows (Figure 4).

XSL•FO
RenderX

Figure 4. Trend of mean occurrences for positive lexical categories (Positive Emotion, and Optimism) across 9 overlapping conversation windows.

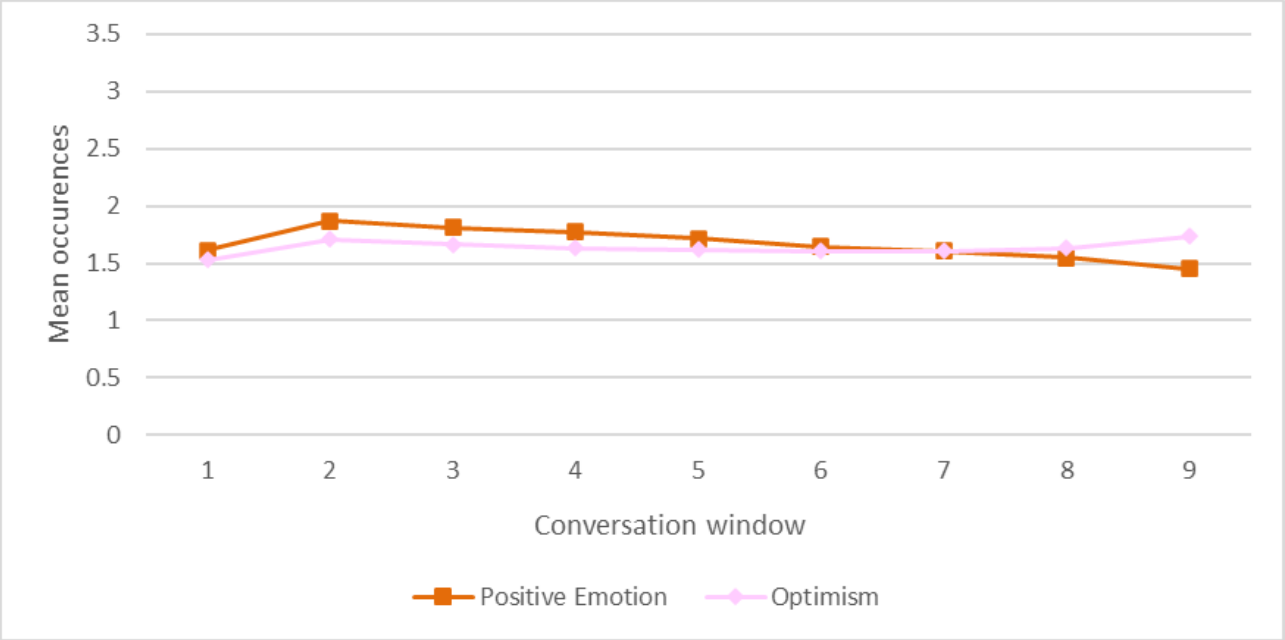

The strength of the regression models varied across categories. Distress, Suicidality, and Suffering, had high R^2 values, indicating strong model fit and predictable changes over time. In contrast, Optimism had the lowest explanatory power, suggesting that its variations may be influenced by additional conversational factors not captured by the linear model. Table S1 in Multimedia Appendix 1 provides means and SD for each category across the 9 windows.

Discussion

Principal Findings

In this study, lexical analysis was combined with traditional statistical techniques to detect changes in language associated with the mental state of help-seekers accessing a chat-based national crisis helpline. Results provide meaningful new insights into how people interact with crisis chat helplines. Crucially, this research is the first to demonstrate the feasibility of a novel methodological approach that may facilitate unobtrusive, objective, and real-time assessment of help-seeker outcomes, with the potential to enhance the effectiveness and efficiency of crisis helplines globally.

Use of negative language reduced across each phase of the crisis chat conversation, from beginning, middle, to end. This pattern was evident across all categories representing negative mental states (Negative Emotion, Suffering, Distress, and Suicidality). Trend analyses across 9 overlapping conversation windows showed a small increase in use of negative language from the first to second window, followed by decreases across the remaining windows. These findings are consistent with meta-analytical evidence for the overall effectiveness of crisis helplines in reducing emotional distress and risk of suicide [9,13], providing validation for our approach. Specific to the chat modality, previous evidence based on help-seeker self-report has found that experiences of distress and feeling suicidal reduced pre-post contact [3]. The current findings strengthen support for this outcome by demonstrating a similar decline in distress and suicidality during the conversation, using an approach unaffected by selection bias (ie, that only help-seekers who felt better or less suicidal completed the postcontact measures). Reducing distress is a core pillar of crisis helplines and has been identified as the most important outcome for help-seekers [31]. Distress had the largest reduction over the conversations in the current study, showing a welcome marker of the effectiveness of crisis helplines.

The reduction in suicidality-related words is also encouraging and aligns with suicide prevention as a key objective of service delivery for crisis helplines [31]. More than 90% of conversations in the current study involved suicidality-related content. Disclosures and discussions about suicide in a helpline context can occur at different stages of the contact depending on the needs of the help-seeker. While any immediate risk of suicide is addressed as soon as it is identified (beginning stage), for other help-seekers the sensitivity of the issue means it may be better discussed after rapport has been established (middle stage) [32]. That the largest reduction in suicidality-related words occurred in the end phase of the conversation in this study supports this notion. The trend analysis further confirms this finding, as Suicidality maintained a downward trajectory across the conversation windows, indicating that the largest reduction in suicidality-related words occurred toward the later stages of the conversation. However, it is important to consider that these aggregate results mask individual variations. The timing of disclosure might also vary between telephone versus digital services, with previous research suggesting greater digital disclosure of sensitive and stigmatized issues among youth [33]. Future studies exploring typical patterns in how and when suicidality is disclosed and discussed across different modalities are needed.

Contrary to expectations, there was a trend toward less expression of positive language across the conversation. Positive Emotion was higher in the beginning, compared with the middle and end phases, although these effects were weak. Findings

XSL•FO
RenderX

based on lexical degree scores indicated that optimism was also highest in the beginning, although average word use was highest in the end phase of the conversation. Trend analyses across 9 overlapping segments also showed a slight decline in Positive Emotion, while Optimism was relatively stable. Unlike simple word occurrence counts, lexical degree scores account for term associations and strengths within a category, and hence these approaches can show some, though not substantive, divergence of results across conversation phases. Rather than indicating ineffectiveness, we suggest the heightened initial positive mental state may reflect hope or relief from deciding to seek help. In addition, the crisis supporter's model of practice is to, first, establish a connection with the help-seeker and then, second, explore their concerns, at which stage an understandable reduction in positive language may occur. Moreover, previous research indicates that most help-seekers do not access crisis support to achieve a positive emotional state but to alleviate an intensely negative one [31,34]. We suggest that while tracking positive mental states may be useful for understanding the complex dynamics of crisis helpline conversations, care should be taken when using positive emotions as outcomes or an indicator of service effectiveness.

Implications for Research and Practice

This proof-of-concept study answers recent calls for innovative approaches to data collection in the crisis helpline context [10,13,25]. The capacity of lexical analysis and advanced affective computing approaches to automatically detect and analyze emotion-based language in large datasets (and potentially in real-time) has the potential to transform crisis helpline research. Affective computing approaches hold promise for improved training instruments, quality assessment, and tools to help crisis supporters understand and respond to help-seekers' needs.

The practical implications of our findings are considerable. By tracking help-seeker mental states in real-time, lexical analysis–based tools may be able to enhance crisis supporter responsiveness, through methods such as visual aids or dashboards that support risk assessment and emotion detection. Such tools might be especially useful in the text and chat contexts where emotion recognition is more challenging due to the absence of vocal cues [35,36]. Similar tools could also be used for professional development and to enhance crisis supporter motivation and satisfaction, thus improving volunteer retention and reducing the risk and impact of negative well-being and burnout. Stronger evidence for crisis helpline effectiveness will facilitate funding opportunities and support more tailored service delivery.

The current findings were based on online chat data and other modalities, like telephone or SMS text message services, may show a different pattern of results. For example, some research shows that help-seekers accessing online crisis helplines have higher levels of suicidality than those accessing telephone helplines [37]. As such, it will be important for future studies to conduct comparative analyses across modalities and to cross-validate NLP methods for monitoring user outcomes with real-time data.

The potential real-time application of lexical analysis-based tools in crisis helplines raises practical and ethical challenges that warrant careful consideration [38,39]. From a practical perspective, ongoing validation of context-specific categories with real-world data will be critical to maintaining accuracy and responsiveness. Language evolves rapidly, particularly in digital environments, with new slang, cultural references, and mental health terminology varying across social contexts [30]. Future research could explore the development of automated methods for continuous learning to detect and incorporate new keywords, enhancing the adaptability of these tools. Ethically, there is a risk of over-reliance on automated tools at the expense of human judgment, potentially comprising the quality of crisis support. Furthermore, the use of these tools must be balanced with the paramount importance of help-seeker confidentiality and privacy. To navigate these ethical challenges, crisis helplines could consider implementing informed consent processes or opt-in mechanisms to provide users with choice and control over the use of artificial intelligence–assisted tools in their interactions. To ensure transparency and maintain the trust of help-seekers, crisis supporters, and the broader community, crisis helplines should develop and publish clear policies on the role of artificial intelligence in crisis interventions, detailing how these tools are used, what data is collected, and how it is protected. Regular evaluation of the risks and benefits, together with robust safeguarding procedures, can help ensure new tools enhance rather than compromise the quality and ethics of crisis intervention delivery. As this field of research grows, it is vital to develop frameworks that ensure research rigor and integrity, including ways to facilitate the safe and ethical sharing of highly sensitive data between service providers and researchers [2,10,40].

Limitations

This investigation revealed several interesting patterns in crisis chat conversations. The use of both pre-existing Empath categories alongside newly developed, context-specific categories, which both showed a similar pattern of results, strengthens the reliability of our findings [41]. However, there are important limitations to note.

A well-known limitation of lexicon-based approaches is their inability to account for context (contextual agnosticism). For example, the use of the word suicide in a chat does not indicate whether the help-seeker is talking about suicide generally or expressing their own suicidality. Similarly, this approach is unable to account for sarcasm, negations, or misspellings. This deficit may be particularly problematic when analyzing crisis online chat conversations, where the language used is often fragmented and ambiguous [4], and may have additional variations associated with computer-mediated communication norms (eg, emojis and abbreviations). To help overcome this limitation, future research could integrate lexical analysis with other approaches, such as sentiment analysis, topic modeling, and deep-learning features such as contextualized embeddings (eg, RoBERTa) [42]. However, more complex approaches would sacrifice the transparency and interpretability of lexicon-based methods [16,43].

XSL•FO
RenderX

Another limitation of lexical-based approaches is the correlation within categories, that is, the overlap between different but related emotional categories. Specifically, the same lexical term, “feel”, represented the top occurring term in multiple categories (Positive Emotion, Suffering, and Optimism). This overlap may hide nuanced shifts in specific mental states, with patterns across categories appearing similar due to their shared expressions [43]. Notably, correlations between the categories included in the current formative study were not exceedingly high and indicated sufficient variance to represent different constructs. Moreover, some overlap in terms and language used to express various emotions is to be expected. Future research may benefit from exploring those relationships between emotions using an approach such as network analysis.

Furthermore, data limitations constrain the generalizability of our findings. Analyzing data from a single helpline (Lifeline Australia), modality (online chat), and time period (3 months) increases risks of bias [44]. A common challenge in crisis helpline research is the lack of access to demographic information about help-seekers due to the anonymous nature of service engagement, which limits understanding of variations across groups [5]. Thus, it remains unclear whether the changes in mental states identified in this study would generalize to other helpline settings, temporal contexts, or service modalities.

Our formative approach of dividing the chats into thirds based on the total number of messages, as well as 9 overlapping windows, in the conversation used 2 methods to enhance comparability across chats of varying pace and length and allowed sufficient data in each segment for meaningful analysis. The use of 9 overlapping windows provides a more nuanced understanding of lexical shifts as it captures subtle fluctuations that may be obscured when dividing conversations into discrete phases (as in beginning, middle, and end). However, it is important to acknowledge that the unique and dynamic nature of each help-seeking process may not be entirely captured by the quantity of text exchanged. Future research may benefit from a dynamic and integrated approach that combines timestamps with message counts to identify more meaningful conversation stages. This could potentially capture both the temporal aspects and content progression of crisis chat interactions.

Finally, our approach assumes that the words used by help-seekers in the conversations reflect key help-seeker outcomes (changes in mental state and suicide risk). While this is likely the case, the validity of using lexical analysis to measure such outcomes needs to be further tested. For example, studies could compare the results of automated text analysis to expert human coding of the same conversations to determine concordance, or triangulate lexical results with self-report measures.

Conclusions

A large dataset of crisis chats from Australia’s national helpline demonstrated how NLP techniques can be used to track language associated with help-seeker mental states. Pending positive results from future validation studies, lexical analysis has the potential to be a valuable tool in monitoring and evaluating outcomes for help-seekers accessing a crisis chat service. The findings should be viewed as a successful test for the feasibility of approach rather than a real-world tool. We hope this formative research and initial step encourages further research toward the development and implementation of tools that can help crisis helplines meet the expanding needs of help-seekers in crisis.

Acknowledgments

This work was conducted by the University of Canberra as part of a National Health and Medical Research Council (NHMRC) Partnership Grant with Lifeline Australia (GNT1153481). The NHMRC had no involvement in the conduct of this research or the preparation of this article. The authors would like to thank Lifeline Australia for their provision of the data analyzed in this study and support for this project.

Data Availability

The datasets generated or analyzed during this study are not publicly available to protect the privacy and maintain the confidentiality of help-seekers but aggregate data are available from the corresponding author on reasonable request. Access to the aggregate data is contingent on obtaining permission from Lifeline Australia.

Authors' Contributions

ML, PJB, NT, RJ, DR, and KM contributed to conceptualizations. SC and KM handled writing–original draft. HB and KM managed formal analysis and data curation. RG, HB, DR, and KM conducted methodology. DR handled funding acquisition. SC, ML, PJB, NT, RJ, DR, and KM performed writing–review and editing.

Conflicts of Interest

ML has collaborated with, and received funding managed by, Lifeline Australia for an unrelated project on preventing suicides in public places.

Multimedia Appendix 1
Means and SDs of token counts for each category across the 9 windows.

XSL•FO
RenderX

[DOCX File, 17 KB - formative_v9i1e63257_app1.docx]

References

1. Preventing suicide: a resource for establishing a crisis line. World Health Organization. 2018. URL: https://apps.who.int/iris/bitstream/handle/10665/311295/WHO-MSD-MER-18.4-eng.pdf [accessed 2024-05-13]
2. Zabelski S, Kaniuka AR, A Robertson R, Cramer RJ. Crisis lines: current status and recommendations for research and policy. Psychiatr Serv 2023 May 1;74(5):505-512. [doi: 10.1176/appi.ps.20220294] [Medline: 36475827]
3. Gould MS, Chowdhury S, Lake AM, et al. National suicide prevention lifeline crisis chat interventions: evaluation of chatters' perceptions of effectiveness. Suicide Life Threat Behav 2021 Dec;51(6):1126-1137. [doi: 10.1111/sltb.12795] [Medline: 34331471]
4. Mokkenstorm JK, Eikelenboom M, Huisman A, et al. Evaluation of the 113 online suicide prevention crisis chat service: outcomes, helper behaviors and comparison to telephone hotlines. Suicide & Life Threat Behav 2017 Jun;47(3):282-296 [FREE Full text] [doi: 10.1111/sltb.12286]
5. Turkington R, Mulvenna M, Bond R, et al. Behavior of callers to a crisis helpline before and during the COVID-19 pandemic: quantitative data analysis. JMIR Ment Health 2020 Nov 6;7(11):e22984. [doi: 10.2196/22984] [Medline: 33112759]
6. Annual report: 2022-2023. Lifeline Australia. 2023. URL: https://www.lifeline.org.au/media/quvfglvm/233-017-lifeline-annual-report-2023-v8-lr-1.pdf [accessed 2024-05-13]
7. Kitchingman TA, Caputi P, Woodward A, Wilson CJ, Wilson I. The impact of their role on telephone crisis support workers' psychological wellbeing and functioning: Quantitative findings from a mixed methods investigation. PLoS ONE 2018;13(12):e0207645. [doi: 10.1371/journal.pone.0207645] [Medline: 30566435]
8. Willems R, Drossaert C, Vuijk P, Bohlmeijer E. Impact of crisis line volunteering on mental wellbeing and the associated factors: a systematic review. Int J Environ Res Public Health 2020 Mar 3;17(5):1641. [doi: 10.3390/ijerph17051641] [Medline: 32138360]
9. Mazzer K, O'Riordan M, Woodward A, Rickwood D. A systematic review of user expectations and outcomes of crisis support services. Crisis 2021 Nov;42(6):465-473. [doi: 10.1027/0227-5910/a000745] [Medline: 33275048]
10. Trail K, Baptiste PJ, Hunt T, Brooks A. Conducting research in crisis helpline settings. Crisis 2022 Jul;43(4):263-269. [doi: 10.1027/0227-5910/a000858] [Medline: 35438001]
11. Tyson P, Law C, Reed S, Johnsey E, Aruna O, Hall S. Preventing suicide and self-harm. Crisis 2016 Sep;37(5):353-360. [doi: 10.1027/0227-5910/a000390] [Medline: 27278572]
12. Lake AM, Niederkrotenthaler T, Aspden R, et al. Lifeline crisis chat: coding form development and findings on chatters' risk status and counselor behaviors. Suicide Life Threat Behav 2022 Jun;52(3):452-466. [doi: 10.1111/sltb.12835] [Medline: 35112387]
13. Hoffberg AS, Stearns-Yoder KA, Brenner LA. The effectiveness of crisis line services: a systematic review. Front Public Health 2019;7(17):399. [doi: 10.3389/fpubh.2019.00399] [Medline: 32010655]
14. Wang Y, Song W, Tao W, et al. A systematic review on affective computing: emotion models, databases, and recent advances. Information Fusion 2022 Jul;83-84(83):19-52. [doi: 10.1016/j.inffus.2022.03.009]
15. Nandwani P, Verma R. A review on sentiment analysis and emotion detection from text. Soc Netw Anal Min 2021;11(1):81. [doi: 10.1007/s13278-021-00776-6] [Medline: 34484462]
16. Malgaroli M, Hull TD, Zech JM, Althoff T. Natural language processing for mental health interventions: a systematic review and research framework. Transl Psychiatry 2023 Oct 6;13(1):309. [doi: 10.1038/s41398-023-02592-2] [Medline: 37798296]
17. Fast E, Chen B, Bernstein M. Empath: understanding topic signals in large-scale text. ArXiv Preprint posted online on February 22, 2023. URL: https://arxiv.org/abs/1602.06979 [accessed 2025-06-11]
18. Pennebaker JW, Francis ME, Booth RJ. LIWC 2015 operator's manual. Linguistic Inquiry and Word Count: LIWC 2015. 2015. URL: https://www.liwc.app/help/psychometrics-manuals [accessed 2024-05-13]
19. Bittar A, Velupillai S, Roberts A, Dutta R. Using general-purpose sentiment lexicons for suicide risk assessment in electronic health records: corpus-based analysis. JMIR Med Inform 2021 Apr 13;9(4):e22397. [doi: 10.2196/22397] [Medline: 33847595]
20. Howes C, Purver M, McCabe R. Linguistic indicators of severity and progress in online text-based therapy for depression. Presented at: Proceedings of the Workshop on Computational Linguistics and Clinical Psychology; Jun 27, 2014; Baltimore, Maryland, USA p. 17145831 URL: http://aclweb.org/anthology/W14-32 [accessed 2025-06-11] [doi: 10.3115/v1/W14-3202]
21. Broadbent M, Medina Grespan M, Axford K, et al. A machine learning approach to identifying suicide risk among text-based crisis counseling encounters. Front Psychiatry 2023;14(14):1110527. [doi: 10.3389/fpsyt.2023.1110527] [Medline: 37032952]
22. Xu Z, Chan CS, Zhang Q, et al. Network-based prediction of the disclosure of ideation about self-harm and suicide in online counseling sessions. Commun Med (Lond) 2022 Dec 6;2(1):156. [doi: 10.1038/s43856-022-00222-4] [Medline: 36474010]
23. Salmi S, Mérelle S, Gilissen R, Brinkman WP. Content-based recommender support system for counselors in a suicide prevention chat helpline: design and evaluation study. J Med Internet Res 2021 Jan 7;23(1):e21690. [doi: 10.2196/21690] [Medline: 33410755]

XSL•FO
RenderX

24. Dinakar K, Chen J, Lieberman H, Picard R, Filbin R. Mixed-initiative real-time topic modeling & visualization for crisis counseling. Presented at: IUI '15: Proceedings of the 20th International Conference on Intelligent User Interfaces; May 18, 2015; Atlanta Georgia USA p. 417-426. [doi: 10.1145/2678025.2701395]
25. Althoff T, Clark K, Leskovec J. Large-scale analysis of counseling conversations: an application of natural language processing to mental health. Trans Assoc Comput Linguist 2016;4(4):463-476. [doi: 10.1162/tacl_a_00111] [Medline: 28344978]
26. Raveau MP, Goñi JI, Rodríguez JF, et al. Natural language processing analysis of the psychosocial stressors of mental health disorders during the pandemic. NPJ Ment Health Res 2023 Oct 5;2(1):17. [doi: 10.1038/s44184-023-00039-6] [Medline: 38609516]
27. Öhman E. The validity of lexicon-based sentiment analysis in interdisciplinary research. Presented at: Proceedings of the Workshop on Natural Language Processing for Digital Humanities; Dec 16-19, 2021; Silchar, Assam, India URL: https://aclanthology.org/2021.nlp4dh-1.2/ [accessed 2025-06-11]
28. Porter MF. An algorithm for suffix stripping. Program 1980 Mar 1;14(3):130-137. [doi: 10.1108/eb046814]
29. Brown PF, Della Pietra VJ, Desouza PV, Lai JC, Mercer RL. Class-based n-gram models of natural language. Comp Linguist 1992;18(4):467-480 [FREE Full text]
30. Hamilton WL, Clark K, Leskovec J, Jurafsky D. Inducing domain-specific sentiment lexicons from unlabeled corpora. Proc Conf Empir Methods Nat Lang Process 2016 Nov;2016:595-605. [doi: 10.18653/v1/D16-1057] [Medline: 28660257]
31. Curll S, Mazzer K, Rickwood D. The development of a core outcome set for crisis helplines: a three-panel Delphi study. Journal of Affective Disorders Reports 2024 Apr;16(16):100763. [doi: 10.1016/j.jadr.2024.100763]
32. Woodward A, Wyllie C. Helplines, tele‐web support services, and suicide prevention. In: O'Connor R, Pirkis J, editors. The International Handbook of Suicide Prevention: Wiley-Blackwell; 2016:490-504. [doi: 10.1002/9781118903223]
33. Rickwood D, Bradford S. Putting young people at the forefront of their mental health care through technology for holistic assessment and routine outcome tracking. iproc 2016;2(1):e40. [doi: 10.2196/iproc.6049]
34. O'Riordan M, Ma JS, Mazzer K, et al. Help-seeker expectations and outcomes of a crisis support service: comparison of suicide-related and non-suicide-related contacts to lifeline Australia. Health Soc Care Community 2022 Nov;30(6):e4535-e4544. [doi: 10.1111/hsc.13857] [Medline: 35676830]
35. Bambling M, King R, Reid W, Wegner K. Online counselling: the experience of counsellors providing synchronous single‐session counselling to young people. Couns and Psychother Res 2008 Jun;8(2):110-116. [doi: 10.1080/14733140802055011]
36. Moylan CA, Carlson ML, Campbell R, Fedewa T. "It's hard to show empathy in a text": developing a web-based sexual assault hotline in a college setting. J Interpers Violence 2022 Sep;37(17-18):NP16037-NP16059. [doi: 10.1177/08862605211025036] [Medline: 34134570]
37. Gilat I, Shahar G. Emotional first aid for a suicide crisis: comparison between telephonic hotline and internet. Psychiatry (Abingdon) 2007;70(1):12-18. [doi: 10.1521/psyc.2007.70.1.12] [Medline: 17492908]
38. Ma JS, O'Riordan M, Mazzer K, et al. Consumer perspectives on the use of artificial intelligence technology and automation in crisis support services: mixed methods study. JMIR Hum Factors 2022 Aug 5;9(3):e34514. [doi: 10.2196/34514] [Medline: 35930334]
39. Hopkins D, Rickwood DJ, Hallford DJ, Watsford C. Structured data vs. unstructured data in machine learning prediction models for suicidal behaviors: a systematic review and meta-analysis. Front Digit Health 2022;4(4):945006. [doi: 10.3389/fdgth.2022.945006] [Medline: 35983407]
40. Pisani AR, Kanuri N, Filbin B, et al. Protecting user privacy and rights in academic data-sharing partnerships: principles from a pilot program at crisis text line. J Med Internet Res 2019 Jan 17;21(1):e11507. [doi: 10.2196/11507] [Medline: 30664452]
41. Czarnek G, Stillwell D. Two is better than one: using a single emotion lexicon can lead to unreliable conclusions. PLoS ONE 2022;17(10):e0275910. [doi: 10.1371/journal.pone.0275910] [Medline: 36240202]
42. Laricheva M, Zhang C, Liu Y, et al. Automated utterance labeling of conversations using natural language processing. ArXiv Preprint posted online on Aug 2, 2022. URL: https://arxiv.org/abs/2208.06525 [accessed 2025-06-11]
43. Lekkas D, Jacobson NC. The hidden depths of suicidal discourse: network analysis and natural language processing unmask uncensored expression. Digit Health 2023;9(9):20552076231210714. [doi: 10.1177/20552076231210714] [Medline: 37928333]
44. Linthicum KP, Schafer KM, Ribeiro JD. Machine learning in suicide science: applications and ethics. Behav Sci Law 2019 May;37(3):214-222. [doi: 10.1002/bsl.2392] [Medline: 30609102]

Abbreviations

LIWC: Linguistic Inquiry and Word Count
NLP: natural language processing

XSL•FO
RenderX

Edited by A Mavragani; submitted 14.06.24; peer-reviewed by P Yip; revised version received 28.03.25; accepted 01.04.25; published 20.06.25.

Please cite as:
Mazzer K, Curll S, Barzinjy H, Goecke R, Larsen M, Batterham PJ, Titov N, Rickwood D
Changes in Mental State for Help-Seekers of Lifeline Australia's Online Chat Service: Lexical Analysis Approach
JMIR Form Res 2025;9:e63257
URL: https://formative.jmir.org/2025/1/e63257
doi:10.2196/63257

© Kelly Mazzer, Sonia Curll, Hakar Barzinjy, Roland Goecke, Mark Larsen, Philip J Batterham, Nickolai Titov, Debra Rickwood. Originally published in JMIR Formative Research (https://formative.jmir.org), 20.6.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Outcomes of an Advanced Epic Personalization Course on Clinician Efficiency through Use of Electronic Medical Records: Retrospective Study

Junye George Chen[1], MBBS; Hao Xing Lai[1], MBBS; Shi Min Wong[2], MScHI; Terry Ling Te Pan[3], MBBS; Er Luen Lim[4], MSHI; Zi Qiang Glen Liau[5], MBBS, MRCS, MMed, MBA

[1]Department of Orthopaedic Surgery, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore

[2]National University Polyclinics Headquarters, Singapore, Singapore

[3]Department of Anaesthesia, Ng Teng Fong General Hospital, Singapore, Singapore

[4]Department of Emergency Medicine, National University of Singapore, Singapore, Singapore

[5]Department of Orthopaedic Surgery, National University Hospital and Alexandra Hospital, NUHS Tower Block, Level 11, 1E Kent Ridge Road, Singapore, Singapore

Corresponding Author:
Zi Qiang Glen Liau, MBBS, MRCS, MMed, MBA
Department of Orthopaedic Surgery, National University Hospital and Alexandra Hospital, NUHS Tower Block, Level 11, 1E Kent Ridge Road, Singapore, Singapore

Abstract

Background: Since Singapore's first migration to *Epic* in 2022, we have been conducting an advanced *Epic* personalization course twice a year for health care professionals with at least 3 months of experience using the system. Electronic medical records education is an under-recognized pillar in reducing health information technology-related stress and clinician burnout.

Objective: The intent of the course is to improve clinician efficiency through customization and personalization of *Epic* interfaces. We hypothesized that compared to their colleagues, trained clinicians would demonstrate significant quantitative improvements in use of the *Epic* system after our course.

Methods: We performed a retrospective analysis from July 2022 to January 2024, including 17 clinicians among 77 individuals who attended our course. Recruitment was done through digital mailers sent out via the local hospital announcement channels. Interested clinicians were able to register for our course via the National University Of Singapore website. Our one-day course involves physical lessons and interprofessional, case-based discussions emphasizing a wide range of high-yield *Epic* functionalities with practice exercises, such as drafting referral letter templates. Three months of pre- and postcourse *Epic* usage statistics of the trained clinicians were retrieved based on aggregate data provided by *Epic* Singapore. Performance metrics included documentation length, time spent in *Epic* functionalities and use of SmartPhrases, order sets and preference lists.

Results: At three months post-course, documentation length decreased by 45.8% (711.8 characters) compared to a 22.2% (126.4 characters) increase among controls. Trained clinicians demonstrated a 2.47-fold increase in use of order sets from 16.2% to 40%, and a 49.9% increase in orders from preference lists after the course, from 35.1% to 52.6%. In total, trained clinicians demonstrated a 1.8-fold increase in combined use of orders from preference lists or order sets after the course, from 51.4% to 92.6%. The number of SmartPhrases created by trained clinicians was 5.64 times higher than among controls, in addition to a 5.57-fold higher use of Quick Filters than controls. Moreover, time in the Chart Review section per day decreased by 29.3% (4.6 min) among trained clinicians versus an increase of 14.6% (2.8 min) among controls. Compared to controls, trained clinicians spent 36.7% (219.6 min) less time in the *Epic* system per day, 56.6% (29 min) less time on Notes per day, and 57.5% (10.7 min) less time on orders per day.

Conclusions: Overall, trained clinicians demonstrated more efficiency in their use of *Epic*, with reduced time spent across various functionalities. Increased use of Smart Tools, including SmartPhrases and Quick Filters was also observed among trained clinicians, which indicated efficiency in reducing total usage time. These findings demonstrated broad improvements across multiple physician efficiency metrics among participating clinicians of our advanced *Epic* personalization course. These gains may contribute to improved mental health among health care professionals and enhanced productivity within health care systems.

(JMIR Form Res 2025;9:e68491) doi:10.2196/68491

XSL•FO
RenderX

KEYWORDS

medical informatics; electronic medical records; EMR; electronic medical records personalization; Epic; electronic health records; EHR; clinician efficiency

Introduction

Background and Significance

Singapore's public healthcare system is organized into three clusters, namely SingHealth, National Healthcare Group (NHG), and the National University Health System. Collectively, these three clusters manage 26 polyclinics, 12 acute general hospitals, and 12 community hospitals, with further expansion in the pipeline to meet the increasing demands of an aging population and chronic diseases [1]. As of 2023, the health workforce includes 17,466 practicing physicians and 38,834 registered nurses [2].

Digitalization of Singapore's health care system is a strategic initiative, with the Ministry of Health collaborating closely with all three regional clusters to establish an integrated digital health ecosystem. Since 2022, the Ministry of Health has gradually implemented the Next Generation Medical Record system, powered by *Epic* Systems, in NHG and National University Health System institutions, with plans for broader adoption. *Epic* Systems is the most widely used electronic medical record (EMR) system globally [3] and was ranked as the number one EMR system in the United States for the 13th consecutive year in 2023 [4]. The adoption of EMRs, including *Epic*, is mandatory in Singapore's public health care institutions to establish standardization, data interoperability, and enhanced patient safety [5].

Epic's EMR platform offers a diverse array of integrated functions that have been shown to enhance clinical workflows, improve patient safety, and support quality improvement efforts across health care systems [6]. For example, the 2023 KLAS Arch Collaborative survey of more than 35,000 clinicians found that *Epic's* automation, real-time reporting, and interoperability features were strongly associated with greater clinician satisfaction and reduced documentation errors [7]. Embedded safety features such as centralized patient data and barcode medication administration help to accelerate patient recoveries and avoid medical errors, thus routinizing safe evidence-based practices and cost-effective preventive medicine [8]. As *Epic* becomes more commonplace [9], emerging technologies including artificial intelligence–powered scribes may further reduce documentation burden and enhance clinician-patient interaction [10,11].

In response to the recent transition to *Epic*, there is a recognized need for enhanced skills-upgrading and peer-based learning initiatives to optimize system integration among health care professionals [12]. A 2025 scoping review of 90 academic studies highlighted that successful EMR implementation requires collaboration, communication, and the development of user competence through structured training and ongoing support [13]. Health care provider buy-in to *Epic* introduces a culture shift, as new EMRs change the way people communicate, prioritize, delegate tasks, and work together [14,15]. Continuous professional development in EMR usage is strongly encouraged and, in many cases, required to ensure clinicians remain proficient and adaptable as the system evolves in Singapore and more hospitals adopt the NGEMR [16].

Reaffirming this commitment, the Yong Loo Lin School of Medicine at the National University of Singapore has introduced a dedicated Continuing Education course titled "How to make the most out of Epic." Being the first course of its kind in Singapore, it is tailored for doctors, nurses, and allied health professionals who have at least three months of experience using the *Epic* system. This advanced personalization course, conducted biannually as a one-day on-campus session, is designed to deepen users' understanding of *Epic*'s functionalities and promote optimal use in clinical practice. The curriculum emphasizes interprofessional collaboration through case-based discussions, allowing participating clinicians to explore real-world scenarios and share evidence-based practices that can drive improvements in patient care.

Objective

The aim of our retrospective study is to evaluate the longitudinal effect of an *Epic* learning course on improving the informatics competency of clinicians in Singapore over a six-month period, with the hypothesis that *Epic* training would translate into greater capacity for efficiency and productivity.

Methods

We followed the STROBE guidelines in conducting this review.

Study Design and Sample Selection

Our course was advertised on the National University Of Singapore website [17] and through electronic mailers via hospital announcement channels. Interested clinicians were able to register for the course via a link on the website.

The "How to make the most out of Epic" official course date was defined as the index date, and data for each clinician was retrospectively collated for the 3 months preceding the index date (ie, preindex period) and for 3 months after index (ie, postindex period). Clinicians were included in the final analysis if they fulfilled the following criteria: (1) ≥21 years of age on the index date; (2) employed as health care clinicians at a health care institution in Singapore (3) ≥3 months of experience using *Epic* as an EMR system at work; and (4) achieved 100% attendance in the "How to make the most out of Epic" course.

A total of 77 individuals including clinicians, nurses, allied health professionals, and administrative executives participated in the course. However, data were available for only 17 clinicians, comprising 15 medical doctors and 2 nurse clinicians. These clinicians were case-controlled against their colleagues from the same clinical departments who did not attend "How to make the most out of Epic" course The analyzed clinicians included attendees from three iterations of the same course, offered biannually, with index dates of October 4, 2022, March 7, 2023, and October 10, 2023, respectively.

XSL•FO
RenderX

Training Contents

The one-day course involving physical lessons and interprofessional case-based discussion sought to "personalize and make *Epic* work for the participants." A wide variety of high-yielding *Epic* functionalities were emphasized, with accompanying practice exercises, including but not limited to performing QuickWin features such as Chart Searches, using keyboard shortcuts, creating and editing SmartPhrases, customizing personal preference lists, setting up order sets, and creating referral letter templates. The course is stackable; completion of three other Informatics courses within a 24-month period contributes to an Executive Certificate in Healthcare Improvement via Informatics, accredited by the National University of Singapore, and recognized across Singapore's health care sector for relevant informatics and quality improvement skills.

Data Source

This retrospective, descriptive study used administrative signals data from *Epic*, collected between July 2022 and January 2024. The Epic database consists of aggregate, system-captured metrics pertaining to *Epic* use.

Ethical Considerations

This study was approved with exemption by our Institutional Review Board (NHG DSRB Ref: 2024‑3739), without the need for informed consent. Privacy and confidentiality protection were maintained by deidentifying the study data. No compensation was provided to the participants during the course of our study. No identification of individual participants was included in any manuscript images or supplementary material.

Study Outcomes and Variables

The analyzed data included the length of typed phrases, clinicians' habits in placing clinical orders via *Epic*, time spent using various *Epic* functions, creation of Quick Filters, and use of Chart Search across both pre- and postindex periods. Quantitative data trends were visualized using line graphs generated with Microsoft Excel 2024.

Results

To put our personalized *Epic* training program into context, it is critical to first understand the architecture and features of the *Epic* EMR system. Within *Epic*, the clinicians used a set of customizable instruments to improve their capacity to document, order, and review patient information. For example, the Chart Search function allows immediate retrieval of desired terms or data points within a patient's record. Quick Filters allow clinicians to efficiently narrow down encounters, notes, or test results, permitting targeted review. Keyboard shortcuts and navigation aids such as the "pick and stick" technique expedite movement through the system and reduce repetitive tasks.

Epic also structures clinical workflows across several modules. Chart Review is a workspace where clinicians can browse a patient's complete medical record, including previous encounters, laboratory results, imaging, and documents in one place. The Clinical Review module is a summary dashboard that shows current clinical information such as problem lists, medications, allergies, and pending tasks, to support timely clinical decision making. The Basket is an integrated communication and task management hub within *Epic*, which hosts notifications, test results, cosignature requests, and other messages requiring clinician attention.

In addition to these core functionalities, *Epic* offers several features specifically designed to improve documentation efficiency and workflow. SmartPhrases are user-defined templates or phrases that can be quickly inserted into clinical notes to standardize and accelerate documentation. Personal preference lists further allow clinicians to tailor frequently used orders or documentation elements for rapid access, while order sets enable the bundling of related orders and documentation templates for common clinical scenarios. Collectively, these features affect the length and structure of typed phrases, clinicians' ordering habits, and time allocation across *Epic* functions. Thus, they are valuable parameters which allow us to evaluate the effectiveness of targeted *Epic* training interventions.

Based on the collective experience of the authors and course instructors, many clinicians initially find the advanced personalized functions such as SmartPhrases, preference lists, and order sets, challenging to master. Consequently, these components were a primary focus of our advanced *Epic* training curriculum. Following the completion of this specialized course, three major themes emerged in the EMR use patterns: improvements in documentation workflow, enhanced time management, and increased efficiency in order management.

First, the trained cases recorded a decreased burden of documentation compared to the control participants, likely due to an enhanced documentation workflow. When comparing the 3 months post- and precourse, respectively, the documentation length, measured by the number of characters, decreased by 45.8% (711.8 characters) among trained clinicians, versus an increase of 22.2% (126.4 characters) among control participants (Figure 1). Similarly, progress note length was also found to decrease by 46.9% (665.7 characters) among trained clinicians, as opposed to an increase of 34.9% (326.1 characters) among control participants. Concomitantly, we observed that the number of SmartPhrases created by trained clinicians was significantly higher at 5.64 times more than that of the controls at 3 months postcourse. SmartPhrases, being precurated abbreviations or sentence templates, reduce manual typing and contribute directly to improved documentation efficiency, contributing to the reduction in documentation length and typing burden.

The second theme revolves around reduced active system use time, possibly as a result of better use of various functionalities of Epic. On average, trained clinicians spent 36.7% (219.6 min) less time using the Epic system per day than controls at 3 months postcourse (Figure 2). Time spent in the Notes module per day was 56.6% (29 min) lower for trained clinicians than controls at 3 months post-course (Figure 3). Additionally, trained clinicians spent 53.4% (19.4 min) less time in Clinical Review and 57.5% (10.7 min) less time on orders per day compared to the controls (Figure 4), while the use of Basket module, time per day remained similar between both groups. Moreover, time

XSL•FO
RenderX

in the Chart Review module per day decreased by 29.3% (4.6 min) for trained clinicians, while control participants showed an increase by 14.6% (2.8 min). These data suggest that advanced training facilitated more efficient navigation and use of *Epics*' module, allowing clinicians to devote less time to routine EMR tasks.

Figure 1. Comparison of documentation length in *Epic* (number of characters) between trained clinicians and control participants, 3 months pre- and posttraining.

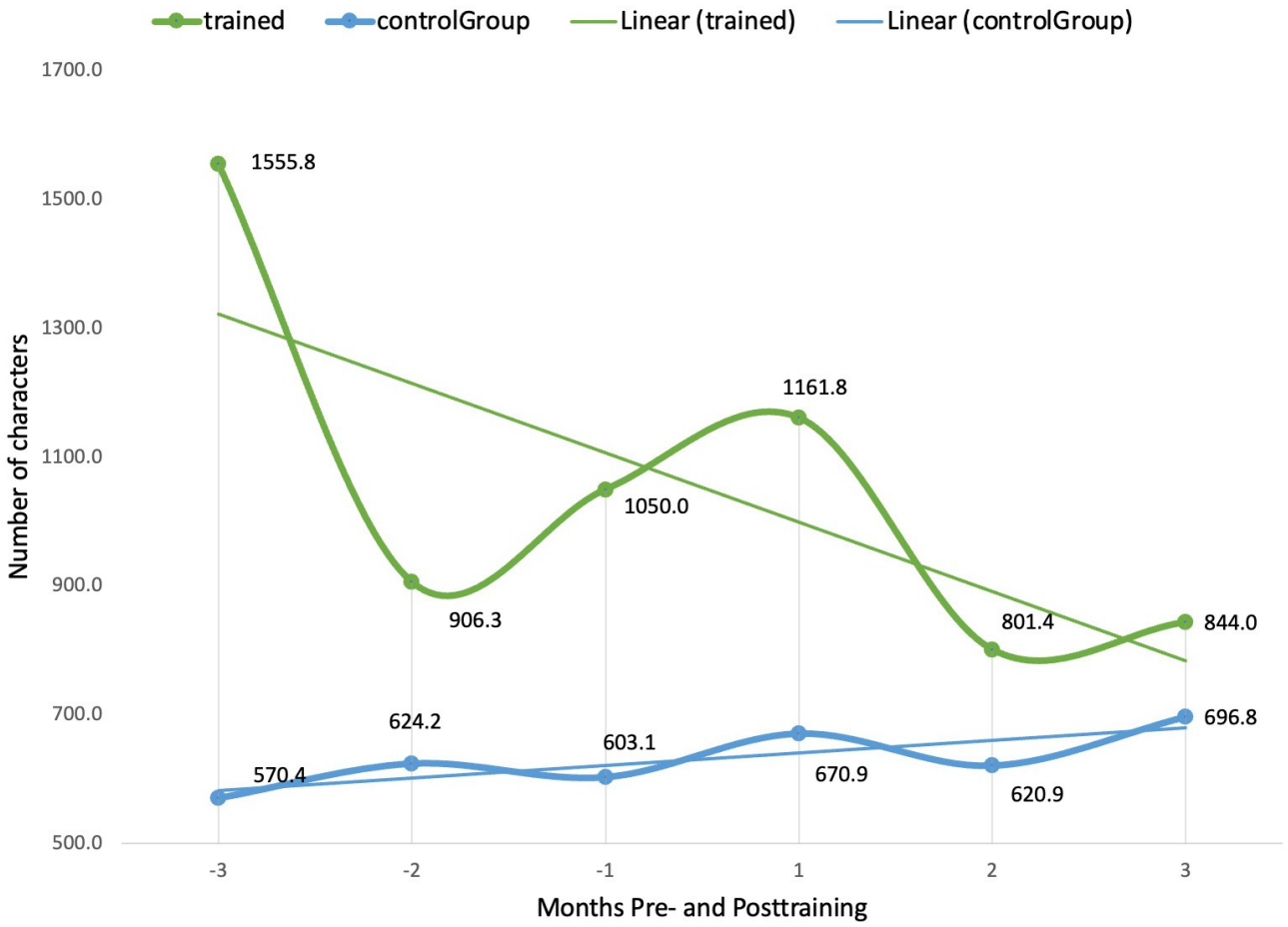

XSL•FO
RenderX

Figure 2. Comparison of time in epic system per day (min) between trained clinicians and controls 3 months pre and posttraining.

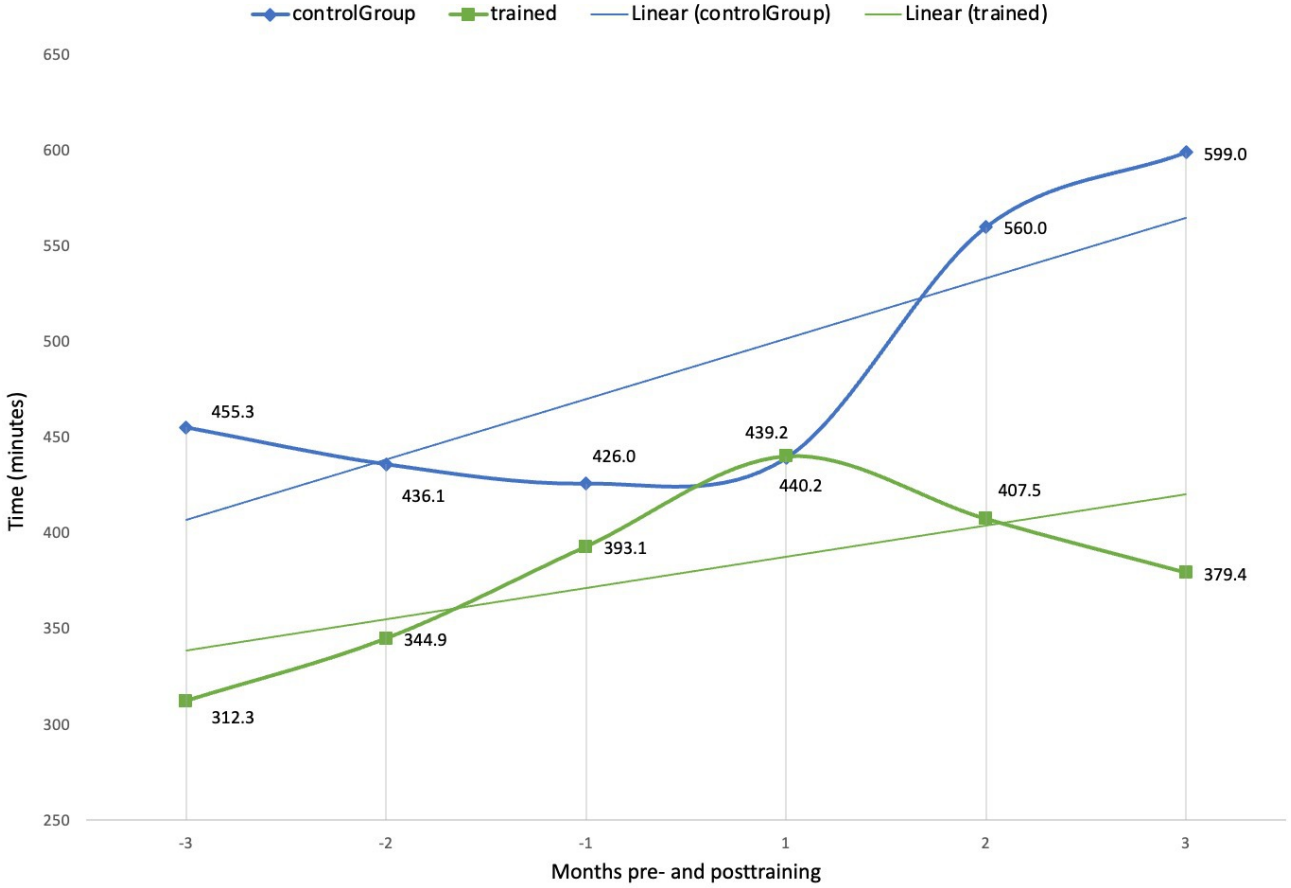

XSL•FO
RenderX

Figure 3. Comparison of time in notes per day (min) between trained clinicians and controls 3 months pre and posttraining.

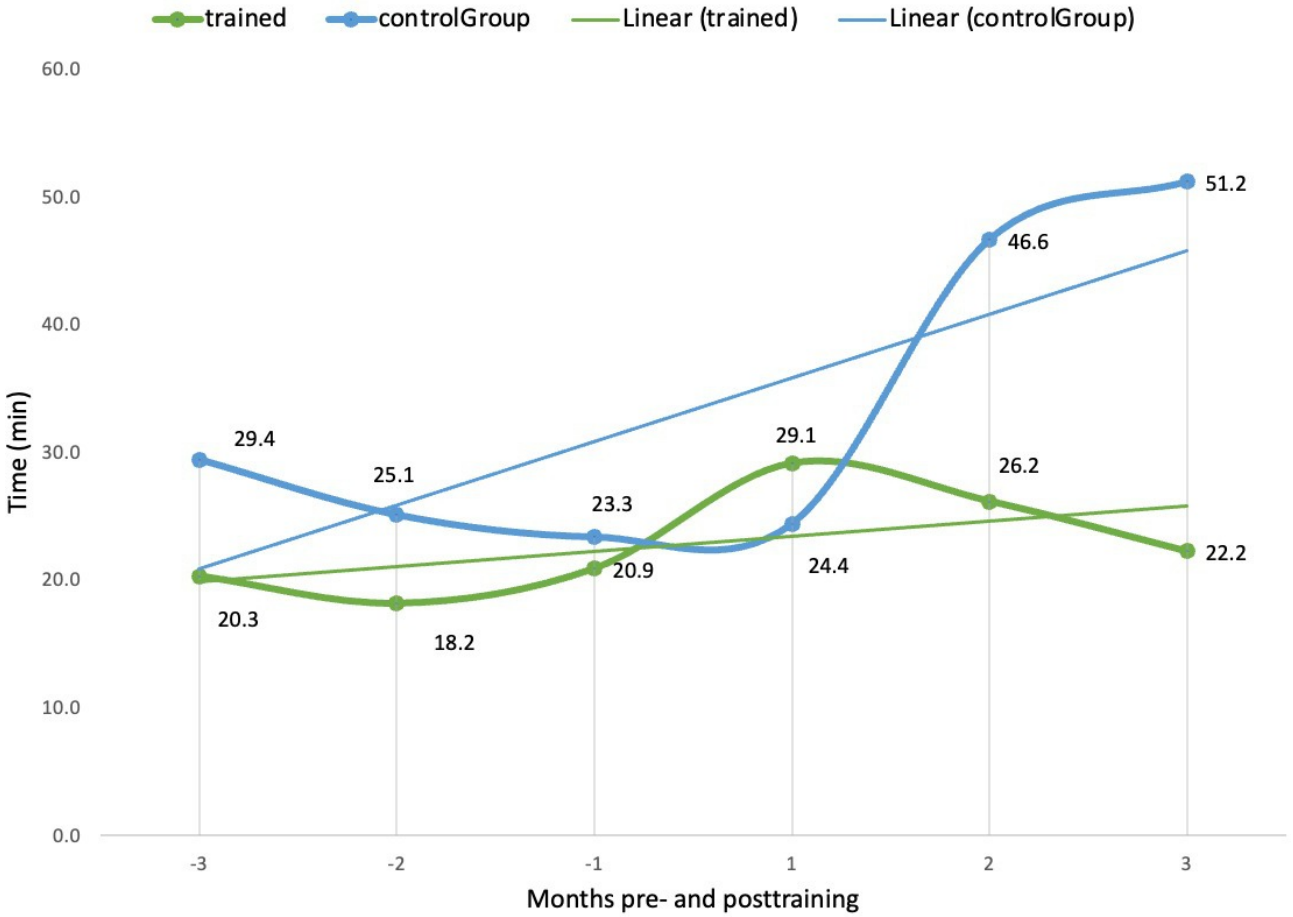

Figure 4. Comparison of time in orders per day (min) between trained clinicians and controls 3 months pre and posttraining.

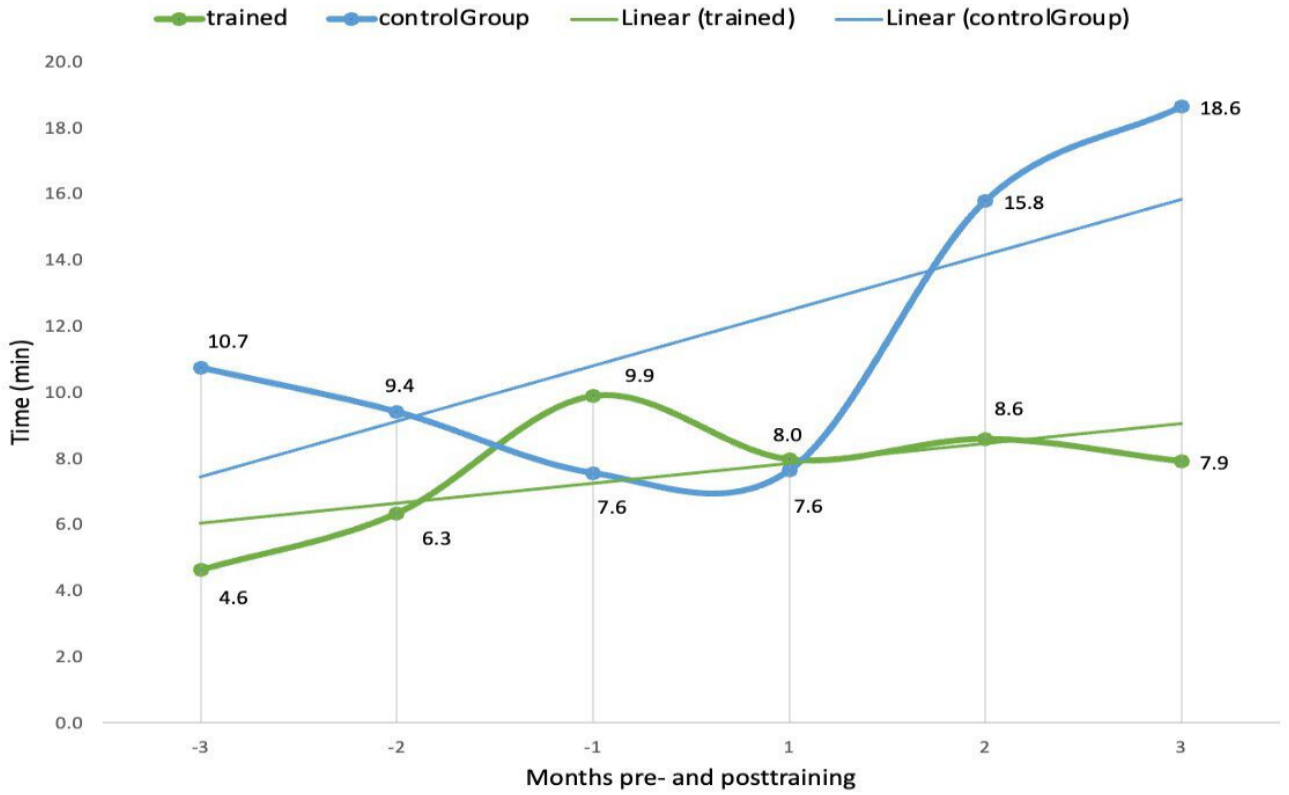

The third key finding centred on increased efficiency in order management, particularly through the customization and use of *Epic's* advanced ordering tools. Trained clinicians demonstrated greater resourcefulness in tailoring their use of preference lists

XSL•FO
RenderX

and order sets. Trained clinicians had a 2.47-fold increased use of order sets from 16.2% to 40% (Figure 5) and a 49.9% increase in orders from preference lists after the course from 35.1% to 52.6% (Figure 6). In total, trained clinicians had a 1.8-fold increased use in orders from preference list or order sets after the course from 51.4% to 92.6%. Additionally, trained clinicians used Quick Filters 5.57 times more compared to controls. Preference lists allow clinicians to save frequently used orders for medications or investigations, while order sets extend this functionality by bundling multiple preference lists. The increased adoption of these customized interfaces significantly streamlined ordering processes and highlights the role of targeted training in reducing both cognitive and time burdens associated with electronic order management.

Figure 5. Comparison of percentage of orders from order set (%) between trained clinicians and controls 3 months pre and posttraining.

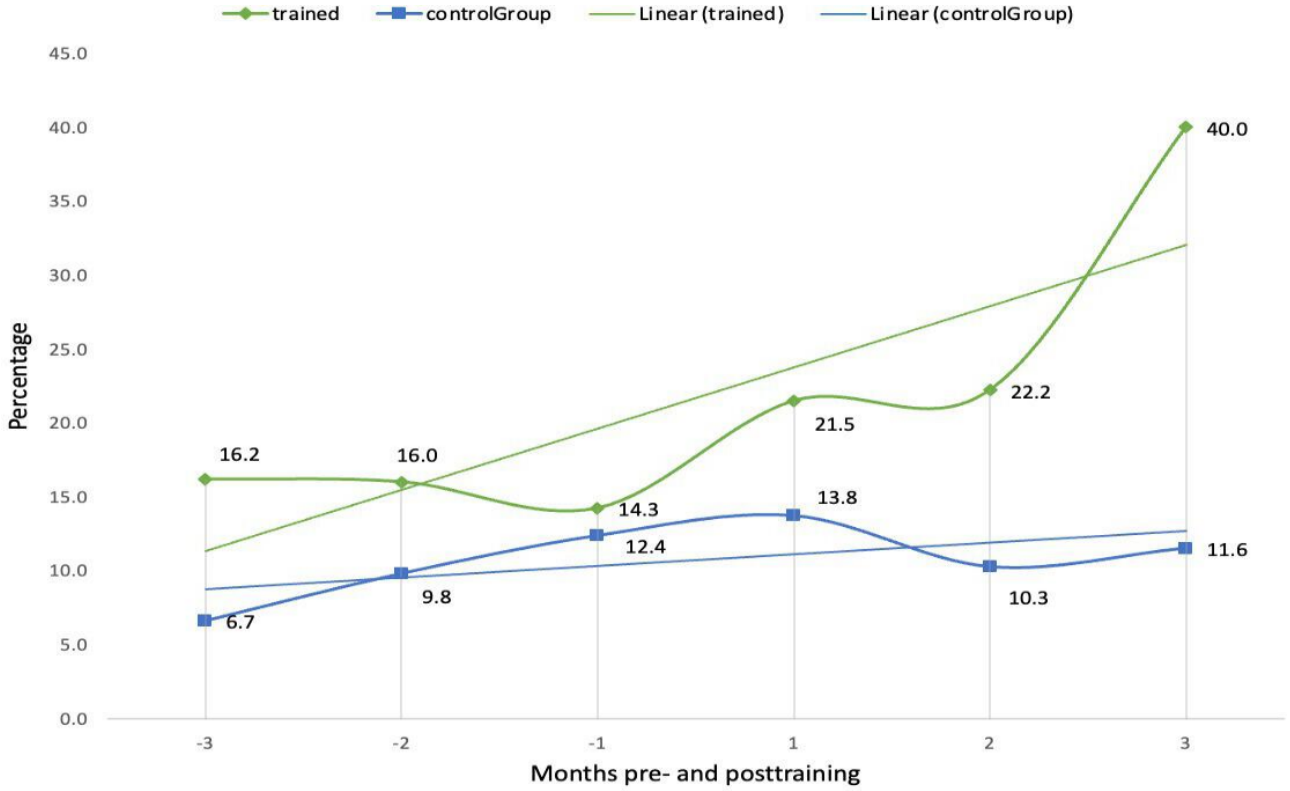

XSL•FO
RenderX

Figure 6. Comparison of orders from preference lists (%) between trained clinicians and controls, 3 months pre- and posttraining. -

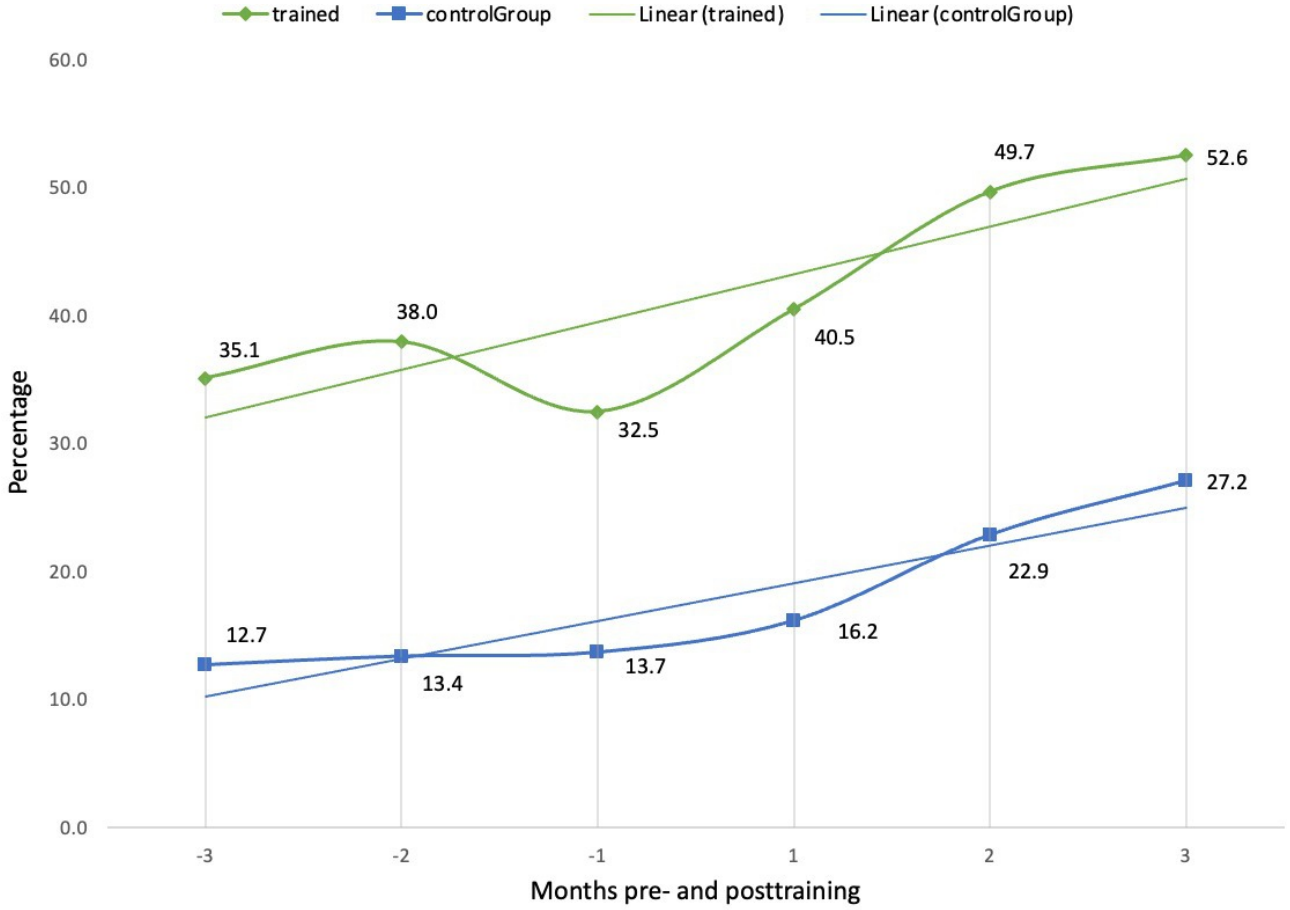

Discussion

Principal Findings

This study aimed to evaluate the impact of an advanced *Epic* EMR training course, "How to make the most out of Epic," on clinician efficiency and workflow in Singapore, within the broader context of competency-based medical education and the growing emphasis on medical informatics [18]. Our findings demonstrate that trained clinicians who completed the course exhibited significant improvements in time management and efficiency across key ambulatory metrics, including documentation, order entry, and clinical review. These outcomes align with the course's objective of bridging interprofessional collaboration and enhancing the practical use of EMR smart features, supporting the integration of informatics competencies into continuing medical education at the Yong Loo Lin School of Medicine, National University of Singapore.

Our results revealed that structured EMR training empowers clinicians to optimize the use of advanced features within *Epic* and harness its potential to improve the quality of health care, consistent with *Epic* being the most studied EMR system for its robust features [19]. Similar training programs conducted in the United States, where *Epic* originated, have shown remarkable gains relating to clinician proficiency [20]. For example, Hollister-Meadows et al [21] found that 133 *Epic*-trained clinicians sustained an average increase of 16 patient consultations per 30-day reporting period, whereas 33 untrained clinicians experienced a decrease of 8 patient consultations [21]. Novel approaches include peer-based training at the Vanderbilt University Medical Center [22], foster a synergistic community to help clinicians overcome the EMR learning curves. Likewise, our course brought together doctors, nurses, and allied health specialists to strengthen interprofessional collaboration workflows using real-world patient examples and scenarios.

Greater mastery of *Epic* provides multifarious benefits for patient care, including standardized aggregation of electronic health information, effective use of decision support tools, and enhanced integration of software buttons with clinical workflows [23]. The findings from our retrospective descriptive study are consistent with that of department-focused "EHR thrive" training conducted by Livingston and Bovi [24], in which trained clinicians demonstrated improved efficiency in time management across documentation, basket management, and clinical orders. Cumulatively, increased utilization of Smart Features within various ambulatory metrics (ie, Notes, Orders, Chart Review, Clinical Review) reduces unnecessary screen time, increases productivity within the health care delivery system, and optimizes integrated care for the patients. As a result, clinicians may devote more time per consultation in direct patient interaction and shared decision-making, rather than "desktop medicine." Further, EMR training and education is often an under-recognized yet important component in reducing Health information technology–related stress and clinician burnout [25], thereby reducing the risk of medical errors. Expanded *Epic* research capabilities into patient-based and

XSL•FO
RenderX

health services databases further contribute to improvement in quality of care [26].

Despite these promising results, several limitations must be acknowledged. Our findings are limited in generalizability due to the retrospective study design and small sample size (n=17), which was affected by recruitment barriers such as mandatory full-day attendance and course fees. Further, only aggregate data was analyzed for confidentiality, which precluded statistical significance testing; demographic information and EMR use duration were unavailable, as these data were not captured by *Epic*. Furthermore, we were unable to assess the provider efficiency profile and *Epic* proficiency scores, which are valuable surrogate measures of EMR training effectiveness as demonstrated by Livingston and Bovi [24]. Inclusion of these metrics would have enabled adjustment for intermonth variability related to clinical workload. Future studies should consider prospective, randomized recruitment, and participant stratification to adjust for confounders such as clinician age, seniority, and previous duration of *Epic* usage.

Conclusion

This preliminary study demonstrates an example of university-led, interprofessional *Epic* training aimed at improving clinician competencies and productivity with EMR interaction. The findings showed significant improvements in efficiency across metrics such as Notes, Chart Review, Clinical Review, and Orders. This evaluation also highlights the benefits of using Smart Tools to improve time management during patient consultation, consistent with literature. Becoming “Epic Smart” and “Computer Smart” extends beyond individual clinician performance. As health care transitions toward digital transformation, training clinicians and students in advanced EMR skills is key to adapting to changing care models and using data-driven decision support. Health care clinicians are encouraged to participate in advanced EMR training programs. Continued investment in informatics education will be crucial to developing a strong workforce capable of meeting the challenges and opportunities of today’s technology-enabled health care.

Acknowledgments

We would like to acknowledge and thank the support of the *Epic* database, without which the project would not have been possible.

Data Availability

The datasets generated during this study are not publicly available due to data confidentiality agreements with the Epic office but are available from the corresponding author on reasonable request. All data generated or analyzed during this study are included in this published article.

Conflicts of Interest

None declared.

Checklist 1
STROBE (Strengthening the Reporting of Observational studies in Epidemiology) Checklist.
[DOCX File, 33 KB - formative_v9i1e68491_app1.docx]

References

1. Healthcare at a glance.: Ministry of Health Singapore; 2025. URL: https://www.moh.gov.sg/healthcare-at-a-glance [accessed 2025-12-05]
2. Health - latest data.: Singapore Department of Statistics; 2025. URL: https://www.singstat.gov.sg/find-data/search-by-theme/society/health/latest-data [accessed 2025-12-05]
3. Miake-Lye IM, Cogan AM, Mak S, et al. Transitioning from one electronic health record to another: a systematic review. J Gen Intern Med 2023 Oct;38(Suppl 4):956-964. [doi: 10.1007/s11606-023-08276-3] [Medline: 37798580]
4. Chishtie J, Sapiro N, Wiebe N, et al. Use of Epic electronic health record system for health care research: scoping review. J Med Internet Res 2023 Dec 15;25:e51003. [doi: 10.2196/51003] [Medline: 38100185]
5. Health information and the NEHR.: Ministry of Health Singapore; 2025. URL: https://healthinfo.gov.sg/overview/health-info-and-nehr/ [accessed 2025-05-12]
6. Kroth PJ, Morioka-Douglas N, Veres S, et al. Association of electronic health record design and use factors with clinician stress and burnout. JAMA Netw Open 2019 Aug 2;2(8):e199609. [doi: 10.1001/jamanetworkopen.2019.9609] [Medline: 31418810]
7. OCHIN Epic earns top recognition from KLAS and Epic systems.: OCHIN; 2025. URL: https://ochin.org/news/klas-epic-recognition/ [accessed 2025-05-12]
8. Johnson RJ. EPIC® and high reliability in healthcare: an evidence based commentary. J Med Inform Decis Mak 2024;1(4):84-96. [doi: 10.14302/issn.2641-5526.jmid-24-4893]
9. Johnson RJ. A comprehensive review of an electronic health record system soon to assume market ascendancy: EPIC®. J Healthc Commun 2016;01(4):36. [doi: 10.4172/2472-1654.100036]

XSL•FO
RenderX

10. Tierney AA, Gayre G, Hoberman B, et al. Ambient artificial intelligence scribes to alleviate the burden of clinical documentation. NEJM Catalyst 2024 Feb 21;5(3). [doi: 10.1056/CAT.23.0404]
11. Ha E, Choon-Kon-Yune I, Luan S, Kishimoto V, Murray L, Hu K, et al. Clinical evaluation of artificial intelligence and automation technology to reduce administrative burden in primary care. : Centre for Digital Health Evaluation, Women's College Hospital Institute for Health System Solutions and Virtual Care; 2024.
12. First healthcare cluster to onboard the next generation electronic medical record.: Synapxe; 2025. URL: https://www.synapxe.sg/blog/national-programmes/first-healthcare-cluster-to-onboard-the-next-generation-electronic-medical-record-ngemr [accessed 2025-05-12]
13. Finnegan H, Mountford N. 25 years of electronic health record implementation processes: scoping review. J Med Internet Res 2025 Mar 3;27:e60077. [doi: 10.2196/60077] [Medline: 40053758]
14. Vanderhout S, Taneja S, Heidebrecht CL, et al. Impacts on quality of care following electronic health record implementation within a large Canadian community hospital: a qualitative study. BMJ Open 2025 Apr 15;15(4):e097646. [doi: 10.1136/bmjopen-2024-097646] [Medline: 40233964]
15. Veenstra GL, Rietzschel EF, Molleman E, Heineman E, Pols J, Welker GA. Electronic health record implementation and healthcare workers' work characteristics and autonomous motivation-a before-and-after study. BMC Med Inform Decis Mak 2022 May 3;22(1):120. [doi: 10.1186/s12911-022-01858-x] [Medline: 35505319]
16. All private hospitals commit to enhancing health data sharing for more seamless care delivery.: Ministry of Health Singapore; 2024. URL: https://www.moh.gov.sg/newsroom/all-private-hospitals-commit-to-enhancing-health-data-sharing-for-more-seamless-care-delivery [accessed 2025-05-12]
17. How to make the most out of EPIC. National University of Singapore – Yong Loo Lin School of Medicine. 2025. URL: https://medicine.nus.edu.sg/continuing-education/course-catalogue/how-to-make-the-most-out-of-epic/ [accessed 2025-06-12]
18. Frank JR, Snell LS, Cate OT, et al. Competency-based medical education: theory to practice. Med Teach 2010 Aug;32(8):638-645. [doi: 10.3109/0142159X.2010.501190]
19. Al Ani M, Garas G, Hollingshead J, Cheetham D, Athanasiou T, Patel V. Which electronic health record system should we use? A systematic review. Med Princ Pract 2022;31(4):342-351. [doi: 10.1159/000525135] [Medline: 35584616]
20. Rajaram A, Hickey Z, Patel N, Newbigging J, Wolfrom B. Training medical students and residents in the use of electronic health records: a systematic review of the literature. J Am Med Inform Assoc 2020 Jan 1;27(1):175-180. [doi: 10.1093/jamia/ocz178] [Medline: 31592531]
21. Hollister-Meadows L, Richesson RL, De Gagne J, Rawlins N. Association between evidence-based training and clinician proficiency in electronic health record use. J Am Med Inform Assoc 2021 Mar 18;28(4):824-831. [doi: 10.1093/jamia/ocaa333] [Medline: 33575787]
22. Users Who Train Each Other Thrive Together.: Corporation ES; 2019. URL: https://www.epic.com/epic/post/users-train-thrive-together/ [accessed 2025-05-12]
23. Virdee J, Thakrar I, Shah R, Koshal S. Going electronic: an Epic move. Br Dent J 2022 Jul;233(1):55-58. [doi: 10.1038/s41415-022-4404-6] [Medline: 35804132]
24. Livingston K, Bovi J. Department-focused electronic health record thrive training. JAMIA Open 2022 Jul;5(2):ooac025. [doi: 10.1093/jamiaopen/ooac025] [Medline: 35474717]
25. Johnson E, Roth E. Improving physician wellness through electronic health record education. Int J Psychiatry Med 2021 Sep;56(5):327-333. [doi: 10.1177/00912174211034633] [Medline: 34292059]
26. Silow-Carroll S, Edwards JN, Rodin D. Using electronic health records to improve quality and efficiency: the experiences of leading hospitals. Issue Brief (Commonw Fund) 2012 Jul;17:1-40 [FREE Full text] [Medline: 22826903]

Abbreviations

EMR: electronic medical record
NHG: National Healthcare Group

Edited by A Mavragani; submitted 07.11.24; peer-reviewed by S Ganesh; revised version received 22.05.25; accepted 25.05.25; published 20.06.25.

Please cite as:
Chen JG, Lai HX, Wong SM, Pan TLT, Lim EL, Liau ZQG
Outcomes of an Advanced Epic Personalization Course on Clinician Efficiency through Use of Electronic Medical Records: Retrospective Study
JMIR Form Res 2025;9:e68491
URL: https://formative.jmir.org/2025/1/e68491
doi:10.2196/68491

XSL•FO
RenderX

© Junye George Chen, Hao Xing Lai, Shi Min Wong, Terry Ling Te Pan, Er Luen Lim, Zi Qiang Glen Liau. Originally published in JMIR Formative Research (https://formative.jmir.org), 20.6.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Investigating Awareness and Acceptance of Digital Phenotyping in Dhaka's Korail Slum: Qualitative Study

Nadia Alam[1], MSc; Domenico Giacco[1], MBBS, PhD; Bulbul Siddiqi[2], PhD; Swaran P Singh[1], MBBS, MD, DM; Sagar Jilka[1], PhD

[1]Warwick Medical School, The University of Warwick, Gibbet Hill Road, Coventry, United Kingdom

[2]Department of Political Science and Sociology, North South University, Dhaka, Bangladesh

Corresponding Author:
Nadia Alam, MSc
Warwick Medical School, The University of Warwick, Gibbet Hill Road, Coventry, United Kingdom

Abstract

Background: Digital phenotyping (DP), the process of using data from digital devices, such as smartphones and wearable technology to understand and monitor people's behavior, health, and daily activities, has shown significant promise in mental health care within high-income countries. However, its application in low- and middle-income countries (LMICs) is limited, particularly among impoverished populations such as slum residents.

Objectives: This study investigates the awareness, knowledge, acceptance, and implementation of DP, including willingness to share data, and concerns regarding privacy and data security, among residents of Dhaka's Korail slum, one of Bangladesh's largest and most densely populated informal settlements. Understanding awareness, acceptance, and privacy concerns surrounding DP in these settings is critical for its effective implementation.

Methods: We conducted 8 focus group discussions with 79% (30/38) of female participants (mean age 37, SD 13.7 years). Participants included 20 individuals diagnosed with serious mental disorders and 18 caregivers. The focus group discussions also included a section explaining what DP is.

Results: Smartphone ownership was reported by 45% (17/38) of the participants, while 92% (35/38) of the participants had access to a smartphone through family members. There was a general lack of awareness about DP among the participants. Initially, 92% (35/38) of participants had no prior knowledge of DP, but after receiving an explanation, they acknowledged its potential applications and benefits. Participants recognized the use of DP for health monitoring, particularly in managing mental health conditions. Participants expressed willingness to share certain types of data, particularly phone usage and location data, provided that content-level information remained private. Despite these perceived benefits, significant concerns about privacy and data security emerged. Participants expressed fears about the potential misuse of their personal information, with some feeling resigned to the idea of already being constantly monitored. Trust in DP tools emerged as a critical factor for adoption, highlighting the need for transparent data protection policies and user control over data sharing. Additionally, participants emphasized the importance of adapting DP tools to local contexts, including cultural considerations and technological literacy.

Conclusions: While DP presents a promising avenue for mental health support in underserved urban populations, its adoption in LMIC slum settings requires targeted educational initiatives, robust privacy safeguards, and community involvement to ensure trust and usability. DP tools should be adapted to fit the cultural context of the target population, possibly involving modifications to the types of data collected or the way data are interpreted. While DP holds potential to improve mental health care in underserved communities, addressing barriers related to awareness, privacy, culture, and usability is crucial. Focusing on educational initiatives, robust data protection, cultural adaptation, user-friendly design, and community engagement, DP can become a valuable tool in bridging the mental health care gap in LMICs.

(JMIR Form Res 2025;9:e65530) doi:10.2196/65530

KEYWORDS

digital phenotyping; mental health; slums; serious mental disorders; lower- and middle-income countries; LMIC; mobile phone

Introduction

Background

Mental health disorders constitute a substantial global health challenge, with approximately 80% of individuals affected residing in low- and middle-income countries (LMICs) [1]. Factors such as rapid urbanization, poverty, and frequent exposure to traumatic events contribute to the high prevalence of mental health problems in these settings. For instance, in the Korail slum of Dhaka, Bangladesh, the combination of poor

XSL•FO
RenderX

living conditions, high population density, and limited access to mental health services exacerbates the prevalence and impact of serious mental disorders (SMDs) such as schizophrenia and major depressive disorder [2-4].

The mental health care gap in LMICs is stark. In Bangladesh, for example, the mental health workforce is critically insufficient, with fewer than 0.5 psychiatrists per 100,000 people [5]. Consequently, many individuals with mental health conditions remain untreated, leading to significant personal and societal costs [1]. This treatment gap is particularly pronounced in urban slums where health care infrastructure is weakest and social determinants of health are most severe.

Korail is one of the largest slums in Dhaka, Bangladesh, with an estimated population of around 200,000 people [6]. It is located in the heart of the city, near affluent neighborhoods and commercial hubs, making it a focal point for discussions about urban inequality [6,7]. The slum is marked by overcrowded and informal housing structures, most of which are constructed from makeshift materials such as bamboo, plastic sheets, and tin [6]. These homes lack basic amenities such as proper sanitation, clean water, and stable electricity, which exacerbates the residents' vulnerability to public health crises and environmental hazards [8].

The population of Korail is highly diverse, consisting largely of internal migrants from rural areas who come to Dhaka in search of employment [8,9]. Many work informally as domestic laborers or small-scale street vendors [6,9]. The socioeconomic diversity within Korail still reflects varying degrees of poverty, with some households relatively better off due to more stable employment or business opportunities, while others remain in extreme deprivation [2,6].

The living conditions in Korail pose significant barriers to health care access, including mental health services [6,8]. The lack of formal health care infrastructure means that residents often rely on informal providers, traditional healers, or nongovernmental organizations for medical help [4]. This makes the introduction of innovative health interventions such as digital phenotyping (DP) both a potential opportunity and a challenge. Given these challenges, implementing digital health solutions requires careful consideration of these unique socioeconomic and cultural factors [3,4].

Korail serves as a microcosm of the broader challenges facing slum communities across LMICs, such as rapid urbanization, extreme poverty, and limited access to essential services [9]. This setting provides a critical backdrop for examining the potential of DP in LMICs, while also highlighting the specific barriers that need to be addressed to make such interventions feasible and effective.

Digital mental health tools could provide a solution, and DP could revolutionize mental health care in LMICs by providing low-cost, scalable methods for early detection, monitoring, and intervention of mental health disorders [10]. The widespread use of smartphones even in impoverished areas offers a unique opportunity to collect valuable data on individuals' behaviors and environments. These data can be used to identify early warning signs of relapse in SMDs, track treatment progress, and personalize interventions [10].

Studies in high-income countries (HICs) have demonstrated the efficacy of DP in various mental health contexts [11]. For instance, passive data from smartphones, such as GPS and call logs, have been used to predict relapse in schizophrenia [12]. These methodologies, however, need to be adapted for LMIC settings, where cultural norms, technological access, and health care infrastructure differ markedly. For instance, communication patterns, social interactions, and even the use of technology can differ widely between populations in HICs and those in LMICs [3]. There is often a lack of understanding and awareness about digital health technologies among the general population and health care providers in LMICs [3]. In the Korail slum, while smartphone usage is common, the concept of DP might be largely unknown. This gap in knowledge necessitates educational initiatives to inform communities about the benefits and implications of DP. These initiatives should aim to demystify the technology, explaining how it works and how it can be used to improve mental health outcomes.

Importance of Context-Specific Research

Understanding the level of awareness, knowledge, and acceptance of DP among residents of LMICs is essential for the successful implementation of digital health interventions [13]. Awareness and acceptance are crucial as limited understanding can hinder adoption and effectiveness [1], with skepticism about sharing personal data posing a significant challenge [14]. Cultural attitudes toward mental health and technology also play a significant role, as stigmatization and cultural beliefs can impact perceptions of DP [15]. Addressing data privacy concerns is critical, as transparency and robust protection measures can enhance trust [16,17]. Engaging the community in the design and implementation of DP tools through participatory approaches can improve usability and relevance, ensuring that interventions meet the specific needs and preferences of the population [18], thereby increasing their sustainability.

While DP has shown promise in HICs, its adoption in LMICs remains limited, and no qualitative research exists [19] to understand people's perceptions and acceptance of this technology in these settings. Understanding these perceptions is vital for developing culturally sensitive and effective DP that can improve mental health outcomes in these underserved populations.

There is currently limited research on the awareness, knowledge, and acceptance of DP in LMICs, particularly among slum populations. To effectively implement DP in LMIC settings, it is crucial to first understand the perspectives of the target population. This study seeks to address this gap by exploring the perceptions of DP among residents of Dhaka's Korail slum. By identifying the key barriers to adoption and the factors that influence acceptance, this study will contribute to bridging the mental health care gap in underserved populations. Through focus group discussions (FGDs) with individuals diagnosed with SMDs and their caregivers, this research aims to provide crucial insights into the potential of DP as a tool for mental health care in resource-poor settings. The findings will inform the development of culturally sensitive, secure, and user-friendly

XSL•FO
RenderX

DP tools adapted to meet the specific needs of slum communities in LMICs. This study aims to explore the awareness, knowledge, and acceptance of DP among residents of the Korail slum in Dhaka, Bangladesh, through FGDs.

Methods

Study Design

This study uses a qualitative research design to explore the awareness, knowledge, and acceptance of DP among residents of the Korail slum in Dhaka, Bangladesh. The qualitative approach was chosen to gather in-depth insights into the participants' perceptions and attitudes toward DP. This study followed the COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist to ensure comprehensive and transparent reporting of the qualitative research process [20]. This study is part of the larger National Institute for Health and Care Research's "Transforming access to care for serious mental disorders in slums" project (TRANSFORM), which outlines the overall workflow [4]. The aim of the TRANSFORM project is to improve access to care for individuals with SMDs living in urban slum settings [4]. Specifically, the study focuses on addressing the treatment gap by implementing and evaluating innovative, community-based interventions. The project seeks to generate evidence on effective, scalable approaches to delivering mental health care in resource-limited urban slum contexts, ultimately aiming to enhance mental health outcomes for underserved populations [4].

Study Setting and Recruitment

Participants were recruited from the Korail slum, one of the largest and most densely populated slums in Dhaka. Participants were selected using purposive sampling. Participants were identified through the TRANSFORM project where community engagement representatives are well known and respected in the area. Given the lack of formal mental health services in the area, recruitment strategies were carefully designed to ensure ethical engagement and inclusivity. A community-based approach was used to identify and enroll participants, leveraging existing partnerships with local stakeholders. We collaborated with community health workers, nongovernmental organizations, and TRANSFORM [4] community representatives, who had well-established relationships with the residents. These trusted individuals facilitated introductions and provided culturally appropriate explanations of the study to potential participants. In addition, community awareness sessions were organized in local field offices, tea stalls, and women's group meetings, where discussions focused on mental health challenges and introduced the concept of DP in an accessible way. The consent process was structured to ensure informed decision-making. Potential participants were provided with a detailed study information sheet in both English and Bengali, outlining the study objectives, procedures, and risks. Participants were given at least 24 hours to consider their involvement before providing written informed consent. For those with limited literacy, the consent form was read aloud, and thumbprint consent was obtained in the presence of a witness. The inclusion criteria were residents of the Korail slum, aged 18 years or older, and either diagnosed with an SMD or caregivers of individuals diagnosed with SMD.

Ethical Considerations

Human Participant Ethics Review Approvals

This study was reviewed and approved by the biomedical and scientific research ethics committee at the University of Warwick (reference: BSREC 100/22‑23). As part of the broader National Institute for Health and Care Research–funded TRANSFORM project [4], the study protocol underwent rigorous ethical scrutiny, including local oversight by institutional partners in Bangladesh, to ensure compliance with country-specific research ethics requirements.

We developed a safeguarding and risk policy to ensure the safety and well-being of participants during their participation in this study. This research involved discussion and questions on mental illness and individual's own or a family member's experience. Such discussions may include distressing events and experiences, which may cause emotional distress. Data collectors NA (a female doctoral candidate from Dhaka based in England) and SR (a male qualitative researcher from Dhaka) were trained in qualitative data collection and had a background in psychology with training in research ethics and conducting research with sensitive population. Our safeguarding policy included actions to respond to events if a participant was uncomfortable or vocalized their distress. This included pausing or stopping data collection, checking in with the participant, and in cases where a participant required professional assistance, the study team, which included local psychiatrists, was available to provide additional support.

Informed Consent

Written informed consent was obtained from all study participants prior to data collection. Participant information sheets and consent forms were provided in both English and Bengali, ensuring clarity and accessibility for the target population. Potential participants were given at least 24 hours to review study materials before consenting. For individuals with lived experience of mental illness, a capacity assessment tool, previously validated in the TRANSFORM study [4], was used to ensure their ability to provide informed consent. In cases where participants were unable to sign, a thumbprint was collected with a witness present. Verbal consent was also recorded at the start of FGDs. Participants were informed that they could withdraw their data without providing a reason.

Privacy and Confidentiality

All data collected during this study were anonymized and stored on encrypted devices in compliance with the General Data Protection Regulation and the UK Data Protection Act (2018). Identifiable data, such as signed consent forms, were securely stored in locked cabinets separate from research data. Transcripts of FGDs were deidentified before analysis, and no personally identifiable information was included in study outputs. All digital data were stored on a secure university server with restricted access limited to the research team.

XSL•FO
RenderX

Compensation

To acknowledge participants' time and effort, a remuneration of 800 Bangladeshi taka (~US $6.60 or £5.30) was provided for each participant. This compensation was determined based on local standards and covered potential costs such as transportation and time off work. The Korail slum is large urban area, which requires lengthy travel, so participants were also remunerated for their cost of local transport (rickshaw) to the field office. Participants were explicitly informed that compensation was not contingent upon completing the study and could be retained even if they chose to withdraw.

Participant Identification in Study Materials

No identifiable personal data of individual participants will be published in the manuscript or supplementary materials. All study data were anonymized before analysis, and participant identifiers were replaced with unique codes.

Procedure

Topic Guide Development

Prior to facilitating FGDs, we codeveloped the interview topic guide in collaboration with members of the Korail community, people with lived experience, and their family members. Community members were engaged through structured discussions to ensure that their perspectives were meaningfully integrated. We discussed the topic of digital mental health and through this discussion, we generated an outline to explain the concept of DP. Alongside this, we also reviewed the literature on DP in LMICs to identify existing themes and gaps in the knowledge base. The initial draft of the topic guide was iteratively refined through feedback sessions with the community, ensuring that it aligned with their priorities and concerns. Where necessary, adjustments were made to improve clarity and relevance. The final guide was designed and generated to facilitate open-ended, meaningful discussions on digital mental health and DP, reflecting both academic insights and lived experiences.

Focus Group Discussions

The FGDs were conducted in the field office of the TRANSFORM Project. The FGDs were structured to ensure a diverse representation of the community, including different genders, ages, and socioeconomic status. Each session was moderated by a trained facilitator, with the assistance of a notetaker. The moderator guided the discussions using a semistructured topic guide, which included questions about participants' understanding of DPs, their current use of digital devices, their willingness to share data, and their concerns regarding privacy and data security (see supplementary information for topic guide in Multimedia Appendix 1), and lasted for a duration of 30 - 45 minutes. Participants were asked about their understanding and views on DP, including its perceived usefulness, comfort levels with data sharing, potential community acceptance, and DP-related concerns.

To ensure that all participants had a similar baseline knowledge of the DP, prior to the start of the FGD, DP was explained in simple Bengali language as follows:

> *Digital phenotyping is a method through which data is collected from smartphones and other smart devices to understand individual's behavior from their regular daily activities. Researchers can use this data to study social cohesion and interactions, behavioral patterns, speech, mobility etc., which are known as digital phenotypes.*
>
> *Data from the phone includes phone usage, location, use of social media (such as Facebook, TikTok). An application is installed on participants' phones after explaining what sort of data will be collected from them, which will allow researchers to collect that data. The data will only be accessible to those involved in the study. The data generated by the app can be used to identify potential mental health problems which can be understood through phone usage patterns.*

Data Collection

The FGDs were audio recorded to ensure accurate capture of the discussions. The recordings were then transcribed verbatim and translated into English. To maintain confidentiality, all personal identifiers were removed during transcription. The translated transcripts were reviewed by the research team to ensure accuracy and fidelity to the original conversations.

Data Analysis

The qualitative data from the FGDs were analyzed using NVivo 14 software. A thematic analysis approach was used to identify patterns and themes within the data. The analysis followed the 5 stages of data analysis [21]: familiarization, identifying a thematic framework, indexing, charting, and mapping and interpretation. Each transcript was read multiple times to enhance familiarity with the data. Initial thoughts, reflections, and preliminary codes were noted. The transcript was then read again and preliminary themes were recorded. Preliminary themes were subsequently grouped into clusters based on common features and meanings. A coding tree was developed to systematically organize these themes, illustrating the hierarchical relationships between overarching themes, subthemes, and specific codes. This structure allowed for a more nuanced understanding of participant experiences and was iteratively refined by the research team. The final coding tree is shown in Figure 1. These themes were validated by cross-checking with the transcript to ensure that they accurately represented the data. This process was repeated for each transcript to maintain consistency. The themes from all transcripts were then compared and combined into master themes to create a comprehensive portrayal of the participants' experiences. The master themes were checked and rechecked against the transcripts to ensure that they were well represented and grounded in the data. Commonalities among the preliminary themes were identified and represented as subthemes, reflecting lower-order aspects of the master themes. To enhance confirmability [22], the findings were grounded in the data by using verbatim quotes from participants to support the identified themes. Dependability [22] was ensured by maintaining a detailed audit trail, including records of coding decisions and analytical memos. The final coding framework was reviewed and refined by 3 researchers

XSL•FO
RenderX

(NA, SR, and SJ) to ensure reliability and validity. Discrepancies were resolved through discussion and consensus.

Reflexivity [23] was a critical component of the analysis process. The researchers' positionalities, including their professional backgrounds and cultural contexts, were acknowledged throughout the study to minimize bias. Regular team discussions were held to reflect on the potential influence of these factors on data interpretation. This reflexive practice ensured that the analysis remained grounded in the participants' narratives. Furthermore, credibility [22] was enhanced through triangulation and peer debriefing. The researchers regularly discussed emerging themes and cross-checked identified themes. Transferability [22] of the study's findings was facilitated through detailed descriptions of the research setting and participant demographics. The narratives from Dhaka's Korail slum enable to assess the applicability of the findings to similar urban slum environments or other resource-limited and impoverished settings. Additionally, the transparency of the thematic analysis process, coupled with the use of participant quotes to support identified themes, strengthens the study's relevance for broader application in comparable contexts. While the findings are rooted in Korail, the themes identified may offer valuable insights for impoverished communities across LMICs.

Figure 1. Coding tree illustrating key themes and subthemes derived from the thematic analysis of focus group discussions on digital phenotyping awareness and acceptance among individuals with serious mental disorders and their caregivers in the Korail slum, Dhaka, Bangladesh. DP: Digital phenotyping.

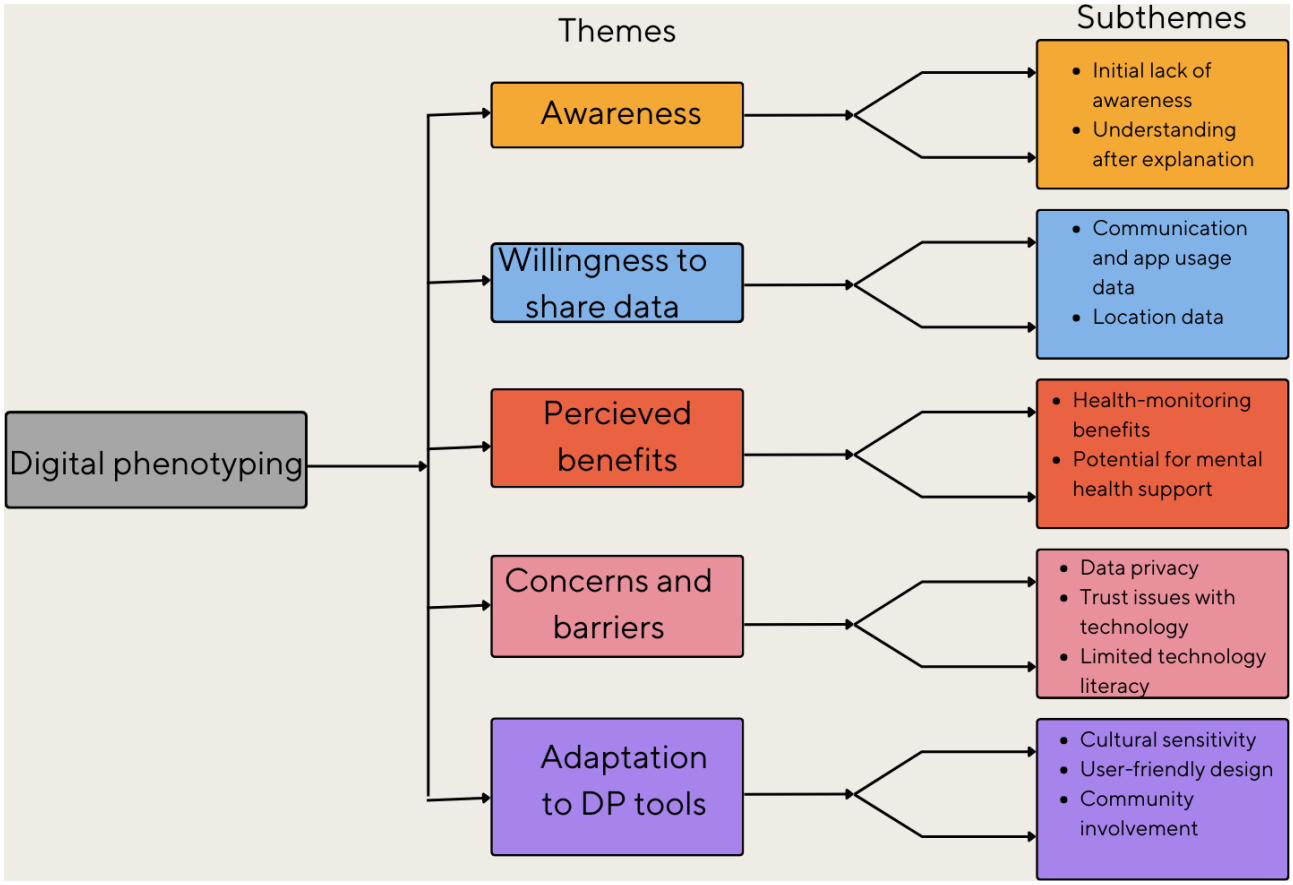

Results

Demographic Information

Eight FGDs were held with homogenous groups of either people with lived experiences or their caregivers and each FGD included 4 - 6 participants. There was a total of 38 participants, with an average age of 37 years (SD 13.7). All participants invited to participate took part in the study and no repeat interviews were carried out. Participants were predominantly women (30/38, 79%) and educated to primary (11/38, 29%) or secondary level (16/38, 42%). Smartphone ownership was roughly split, with 45% (17/38) owning smartphones, but nearly 92% (35/38) had access to a smartphone via family members. See Table 1 for participant demographics.

XSL•FO
RenderX

Table . Participant demographic information (N=38).

Characteristics	Values
Age (years), mean (SD)	37 (13.7)
Sex, n (%)	
Female	30 (79)
Education level, n (%)	
No education	5 (13)
Primary	11 (29)
Middle school	16 (42)
High school	4 (11)
Undergraduate	2 (5)
Marital status, n (%)	
Married	29 (76)
Unmarried	4 (11)
Widowed	4 (11)
Divorced	1 (2)
Employed, n (%)	
Yes	12 (32)
No	26 (68)
Smartphone ownership, n (%)	
Yes	17 (45)
No	21 (55)
Average length of smartphone ownership (years), mean (SD)	3.33 (1.15)
Smartphone ownership in family, n (%)	
Yes	35 (92)
No	3 (8)

Awareness and Understanding of DP

The FGDs revealed a general lack of awareness about DP. Initially, most participants reported to have had no prior knowledge or understanding of the concept and many participants expressed surprise upon learning about the tracking capabilities associated with DP, with responses such as, “I didn’t know someone could track my phone” (caregiver), indicating a widespread unfamiliarity with the concept. Participants initially understood DP as government or police surveillance. For instance, one participant expressed, “Good, you will be able to catch criminals” (caregiver), “It helps in tracking lost phones and finding people in emergencies” (patient), and “knowing the location can help in emergencies” (caregiver). After further discussion, participants began to grasp the notion of DP and its implications. Several participants acknowledged an increased understanding, with comments highlighting their comprehension of both pros and cons of DP, “Now that I understand, it has both pros and cons” (patient), and “I see now how it works; it has its advantages and disadvantages” (caregiver). This shift in understanding highlighted the importance of educational interventions to enhance awareness of DP and its potential applications.

Types of Data Residents Are Willing to Share

Residents showed a willingness to share specific types of data, particularly they talked about communication and app usage. Participants frequently mentioned the use of their phones for making calls, with one stating, “It [phone calls], you can use that data. I use the phone mainly to make calls” (caregiver), and another noting, “I talk to my children and husband on the phone. So, it can be used maybe” (caregiver). Similarly, messaging apps such as WhatsApp were commonly used, with comments such as, “We use messaging apps like WhatsApp, so you can get that data” (patient), indicating a prevalent reliance on these platforms for communication and their opinions regarding collection of those data. Participants directly agreed that they were willing to share communication data with comment such as, “You can see how much we talk” (patient) but did not want their communication content to be seen with one participant saying, “Knowing more than they need [referring to content of communication] to is bad” (caregiver). App usage also emerged as a significant area of data sharing, with participants highlighting their engagement with various apps for communication and entertainment. Examples included, “I use WhatsApp, Facebook, and Bkash,” and “Apps like TikTok and YouTube are frequently used in our home” (caregiver).

XSL•FO
RenderX

Participants revealed that DP can offer insights into personal preferences and behaviors, as one participant noted, “They can know who likes which drama, who listens to music, who likes music. Can understand people’s tastes” (patient).

Additionally, location data were recognized for their use, especially in emergencies, with statements such as, “Tracking [location] can help us...,” (patient). Participants highlighted various practical benefits of location tracking, such as its convenience for navigation and travel. One participant noted, “I keep location on my phone; it is convenient to go somewhere” (caregiver), emphasizing the everyday use of this feature. Overall, participants recognized that location services could significantly aid in navigation and safety. When asked whether it was okay for their location to be used for DP, participants were mostly in agreement with comment, such as “I think it is good. Digital phenotyping is better in this case” (patient). When asked whether it would be perceived dangerous to track people’s location, participants disagreed with comments, such as “No knowing this is not a problem. It’s better for me” (patient) and “In some circumstances it is good [referring to location tracking]” (caregiver).

Perceived Benefits and Risks

Participants identified several benefits of DP, particularly in health monitoring. Participants highlighted the potential of smartphones and apps to improve the management of mental health conditions. Participants made comment such as, “It will be very helpful for monitoring health conditions” (caregiver) and “Health monitoring through apps can provide immediate assistance” (caregiver). They highlighted the convenience and usefulness of receiving counseling and medical treatment through mobile phones, with one participant noting, “If you receive counselling, get treatment and diagnosis through mobile phones, it will be very helpful” (patient). The immediacy and accessibility of health monitoring through apps and opportunities were appreciated, with practical benefits in managing treatment schedules and medication reminders. Overall, the perceived benefits of DP among participants were substantial for improved health monitoring and management, highlighting the potential of digital tools to improve overall well-being.

Despite the general willingness to share data, participants expressed considerable concerns regarding privacy and data security. Many were apprehensive about the potential misuse of their personal information. One participant articulated this fear: “Someone will know your personal information; is that not a bad side of it?” (caregiver). This sentiment was widely shared, with participants expressing fears that their private conversations and activities could be monitored and exploited. A participant highlighted this anxiety, saying, "Fear, because if I go to tell any information to someone, they are recording these things behind me; if my secret is leaked, I may be in danger” (caregiver). Another added, “Some people can harm us with our personal information” (patient), underscoring the potential risks associated with data breaches. There was also a sense of resignation among some participants regarding surveillance. One participant mentioned, “No problem, we are already being watched” (patient), reflecting a belief that privacy invasions are already a part of their reality. This feeling was further echoed by another participant, “We have no problem if you track. We have no problem if you hack” (caregiver), indicating a degree of desensitization to privacy concerns.

Participants also expressed nuanced views about the tracking and monitoring of data. While some acknowledged the necessity of such measures for security purposes, they were wary of their potential for misuse. For instance, one participant remarked, “If they want to track a criminal they can; I didn’t do anything, why should they track mine” (patient), highlighting a concern that innocent individuals might be unjustly monitored. Another participant questioned the motives behind data collection, stating, “I don’t know why. But they take it for their own interests” (caregiver), suggesting skepticism toward the intentions.

The concern extended to the practical usability and accessibility of digital tools, particularly among older family members. One participant pointed out, “It can be challenging to teach older family members how to use health apps” (caregiver), emphasizing the need for user-friendly interfaces and adequate support systems. Another participant noted, “It saves time and travel costs, but not everyone knows how to use the apps properly” (patient), indicating that technological literacy remains a significant barrier.

In summary, while there is a willingness among residents to share data for the benefits they might bring, significant concerns about privacy invasion and data misuse persist. These concerns are coupled with practical challenges related to the usability of digital tools, particularly for older and less tech-savvy individuals.

Adaptation of DP Tools

The discussions underscored the necessity for DP tools to be culturally sensitive and user-friendly to ensure broad acceptance. Participants emphasized that these tools should reflect local customs and practices, as well as provide valuable information relevant to their daily lives. For instance, one participant mentioned, “Health apps could offer valuable information and support” (caregiver), highlighting the potential benefits of culturally tailored digital health solutions and information.

Additionally, adapting data collection methods to local behaviors and technological habits is essential. Given the high use of communication apps such as WhatsApp and social media platforms such as Facebook and TikTok, DP tools could focus on analyzing data from these sources to identify behavioral patterns relevant to mental health. As a participant noted, “We use WhatsApp, Facebook, and YouTube a lot; you can see our activities there” (caregiver).

Furthermore, addressing privacy concerns through robust data protection measures and transparent communication can build trust among users. One participant expressed, “If we know our data is safe and not misused, we will feel more comfortable sharing it” (patient). Incorporating local community feedback into the design of these tools can ensure that they meet the specific needs and concerns of the population. Another participant highlighted, “If you involve us in the process, we can tell you what works best for us” (caregiver).

XSL•FO
RenderX

Discussion

Principal Results

By understanding the local context and involving the community in the research process, this study aims to create sustainable and impactful approach into DP research. The findings from this study highlight both the potential and the challenges associated with implementing DP in LMICs, particularly in underserved communities such as the Korail slum in Dhaka, Bangladesh. The results indicate a general lack of awareness and understanding of DP among residents, though there is significant interest and perceived benefit.

The initial unfamiliarity with DP shows the need for targeted educational interventions. Despite the widespread use of smartphones in the community, most participants had not previously encountered the term or understood the concept. This lack of awareness could hinder the adoption of DP tools, as individuals are unlikely to engage with technologies they do not understand. Educational initiatives should therefore focus on explaining how DP works and its potential benefits in health [1].

Once informed about DP, participants recognized several significant benefits, particularly in the realm of health monitoring. The convenience and use of location tracking in emergencies were noted as important advantages. Participants saw potential in using smartphones and apps to manage mental health conditions. This reflects a broader trend in digital health where mobile health apps are increasingly recognized for their ability to provide immediate assistance, manage treatment schedules, and offer valuable health-related information [10].

One critical barrier to implementing DP in the context of LMICs such as Korail is the low rate of individual smartphone ownership. In this study, only 45% (17/38) of the participants reported owning a smartphone, with many (35/38, 92%) relying on shared devices within their families. This dynamic presents a significant challenge to the feasibility of DP, which typically requires continuous, personal data collection via smartphones.

In many LMICs, it is not uncommon for families to share a single smartphone, which raises questions about the reliability of the data collected [24]. A shared device could distort the individual data necessary for accurate mental health monitoring, especially if different family members use the same phone throughout the day. While participants did not explicitly raise this as a concern in the focus groups, it is essential to consider whether shared use of a “family smartphone” is culturally normalized to the point that participants do not view it as problematic.

To address this challenge, future iterations of DP interventions in LMICs could explore the development of family-centered monitoring systems. Such systems could account for shared device use by developing mechanisms that differentiate between users or adapt to intermittent access by the primary individual being monitored. For example, the technology could include features that identify individual users at the start of each “session” (such as logging the user upon unlocking the phone with their fingerprint) or leverage novel machine learning algorithms that account for shared usage patterns.

The results of this study demonstrate a clear lack of awareness and understanding of DP among the residents of Dhaka’s Korail slum. Most participants had not encountered the concept prior to the FGDs, revealing a significant knowledge gap. While participants expressed willingness to share certain types of data, such as communication and app usage, they were apprehensive about more intrusive forms of data collection, particularly if it involved accessing the content of their communications. This mirrors concerns raised in other LMIC contexts, where distrust in data handling and fear of surveillance are common [24,25]. These concerns must be addressed through transparent communication about how data will be used and protected, as well as the implementation of robust data security measures. Failure to address these issues could significantly impede the adoption of DP tools in slum communities. In many LMICs, the legal infrastructure necessary to safeguard personal data is either underdeveloped or inconsistently enforced, leaving individuals vulnerable to potential misuse of their information. This distrust is compounded by participants’ feelings of resignation regarding surveillance, as many believe that they are already being monitored in some form. This perception represents a substantial hurdle to overcome in implementing any digital health tool. Moreover, the digital divide—characterized by uneven access to technology and varying levels of technological literacy—remains a critical issue. Although smartphone ownership in the Korail slum is relatively high, a significant proportion of residents still rely on family members’ devices.

Comparison With Available Literature and Implications

Despite the perceived benefits, concerns about privacy and data security were prominent. Participants expressed fears that their personal information could be misused, with some feeling resigned to the idea that they are already being watched. This highlights a critical barrier to the adoption of DP tools: the need to ensure robust data protection measures and communicate these effectively to users. Participants in HICs have similar concerns to those in LMICs regarding data sharing, particularly in terms of privacy, data security, and ethical considerations. In HICs, participants express significant concerns about privacy and the potential misuse of their personal data. These concerns include the inadequacy of current regulations to protect sensitive personal information and the potential for data to be sold or analyzed outside the health care system [26].

Transparency, consent, accountability, and fairness are also critical issues. There is a strong emphasis on ensuring that DP tools are developed with robust data protection measures and clear communication about these measures to build trust [27,28]. In both HICs and LMICs, the willingness to share data is often contingent upon understanding how the data will be used and ensuring that robust measures are in place to protect privacy. Addressing these concerns through transparent practices and community engagement is crucial for the acceptance and effectiveness of DP technologies. While privacy is a universal issue, the extent and nature of these concerns can vary

XSL•FO
RenderX

significantly. Participants from HICs generally have greater awareness and more robust expectations of privacy protection mechanisms, often influenced by stringent data protection regulations and a higher level of digital literacy [28]. In contrast, residents of Korail, despite recognizing the importance of privacy, might exhibit a certain resignation toward surveillance, possibly due to their frequent exposure to informal and less regulated data environments. Building trust through transparent data practices and addressing privacy concerns directly will be essential to foster acceptance [16,29].

The necessity for DP tools to be culturally sensitive and user-friendly was also emphasized. Participants stressed that these tools should reflect local customs and practices and be accessible to all, including older and less tech-savvy individuals. This points to the importance of designing digital health tools that are not only technologically effective but also culturally relevant and easy to use. Engaging the community in the design and implementation process can enhance the relevance and acceptance of these tools [30].

The findings suggest several practical implications for the implementation of DP in LMICs. There is a need for comprehensive educational initiatives to raise awareness and understanding of DP. These programs should use local languages and culturally relevant metaphors to explain the technology and its benefits [14]. Robust measures must be in place to protect users' data, and clear communication about these measures can help alleviate fears and build trust among users [17]. DP tools should be adapted to fit the cultural context of the target population, possibly involving modifications to the types of data collected or the way data are interpreted to ensure that they are relevant and acceptable [31]. The usability of digital tools is crucial, particularly for older and less technologically literate individuals. Tools should be designed with these users in mind, providing simple interfaces and adequate support systems [32]. Involving the community in the design, testing, and deployment of DP tools can improve their acceptability and effectiveness. Participatory approaches ensure that interventions are tailored to meet the specific needs and preferences of the community [33].

Strengths and Limitations

Despite the valuable insights gained from this study, several limitations should be acknowledged. First, the sample size of 38 participants, although sufficient for qualitative analysis, may not fully capture the diverse experiences and perspectives of all residents within the Korail slum. This limitation in sample diversity could affect the generalizability of the findings to other slum communities or urban poor populations in different LMIC settings [34]. Second, the reliance on FGDs as the primary data collection method may have influenced participants' responses due to social desirability bias. Participants might have felt pressured to conform to perceived group norms or expectations, potentially limiting the expression of dissenting views or concerns about DP [35]. Another limitation is the potential for translation. DP is an emerging technology [33], and it might be difficult to understand in general, and in the FGD the meaning of DP was explained in Bengali from an English explanation. This process might have introduced nuances or misinterpretations among participants [36]. Ensuring accurate and culturally sensitive translation is critical, but even with meticulous efforts, some meaning can be lost or altered. Moreover, the study's design captures a snapshot of participants' perceptions and attitudes at a single point in time. Longitudinal research could provide a more comprehensive understanding of how awareness, acceptance, and concerns about DP evolve over time, particularly as participants become more familiar with the technology [37]. Technological limitations also played a role. Although smartphone usage is relatively common, the varying degrees of technological literacy among participants could influence their understanding and engagement with DP tools. Future studies should consider incorporating training sessions to ensure that all participants have a baseline understanding of the technology being discussed [38]. Finally, while this study focused on a specific slum in Dhaka, the findings might not be directly applicable to other urban or rural settings in Bangladesh or other LMICs without considering local cultural, social, and economic contexts. Comparative studies across different settings are necessary to validate and expand upon these findings [39].

Finally, integrating DP tools into existing health systems can be challenging. Health systems in LMICs are often underresourced and fragmented, which can complicate the implementation of new technologies. Successful integration requires collaboration with local health authorities and stakeholders to align DP initiatives with national health priorities and infrastructure. Training health care providers to use and interpret DP data is also crucial to ensure that the technology can be effectively used in clinical practice.

Conclusions

This study shows the potential of use of DP to improve mental health care in underserved communities such as the Korail slum. However, for these tools to be effective, it is crucial to address the existing barriers related to awareness, privacy, cultural sensitivity, and usability. By focusing on educational initiatives, robust data protection, cultural adaptation, user-friendly design, and community engagement, DP can become a significant tool in bridging the mental health care gap in LMICs. Further research is needed to develop and implement these solutions in a way that is both effective and sustainable.

Acknowledgments

We would like to acknowledge Sheikh Rasel for his substantial contributions to the study, including assisting in the development of the topic guide, collecting and transcribing data, and for his assistance during data collection phase. We acknowledge Md. Jakaria Habib for his assistance with note-taking during the FGDs. We also acknowledge support of our local community representatives: Aulia Khatun, Md. Abdur Rahman, Ragina Aktar Rumi, Md. Abul Kashem Titu, Raseda, Rashida Begum. We also acknowledge the work of the TRANSFORM Project and TRANSFORM consortium.

XSL•FO
RenderX

Data Availability

Data are available from the corresponding author upon reasonable request. Requests must be in line with the TRANSFORM study data-sharing agreement and will be discussed with the study principal investigator.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Coding tree illustrating key themes and subthemes derived from the thematic analysis of focus group discussions on digital phenotyping awareness and acceptance among individuals with serious mental disorders and their caregivers in the Korail slum, Dhaka, Bangladesh.
[DOCX File, 20 KB - formative_v9i1e65530_app1.docx]

References

1. Rathod S, Pinninti N, Irfan M, et al. Mental health service provision in low- and middle-income countries. Health Serv Insights 2017;10:1178632917694350. [doi: 10.1177/1178632917694350] [Medline: 28469456]
2. Koly KN, Baskin C, Khanam I, et al. Educational and training interventions aimed at healthcare workers in the detection and management of people with mental health conditions in South and South-East Asia: a systematic review. Front Psychiatry 2021;12:741328. [doi: 10.3389/fpsyt.2021.741328] [Medline: 34707524]
3. Gruebner O, Khan MMH, Lautenbach S, et al. Mental health in the slums of Dhaka—a geoepidemiological study. BMC Public Health 2012 Mar 9;12(1):177. [doi: 10.1186/1471-2458-12-177] [Medline: 22404959]
4. Singh SP, Jilka S, Abdulmalik J, et al. Transforming access to care for serious mental disorders in slums (the TRANSFORM Project): rationale, design and protocol. BJPsych Open 2022 Oct 13;8(6):e185. [doi: 10.1192/bjo.2022.584] [Medline: 36226591]
5. Bangladesh WHO special initiative for mental health situational assessment. : World Health Organization; 2020 URL: https://www.who.int/docs/default-source/mental-health/special-initiative/who-special-initiative-country-report---bangladesh---2020.pdf?sfvrsn=c2122a0e_2#:~:text=The%20assessment%20was%20carried%20out,indicators%20across%20the%20life%2Dcourse [accessed 2024-03-13]
6. Ahmed Sinthia S. Analysis of urban slum: case study of Korail slum, Dhaka. Int J Econ Manage Eng 2020 Nov 30;14(11) [FREE Full text]
7. Kanti M, Awlad M, Badrul H, et al. The perceptions of community groups to improve MNCH in urban slums an exploratory case study of Korail slum in Dhaka. : International Centre for Diarrhoeal Disease Research, Bangladesh; 2009 URL: https://www.academia.edu/24662414/The_Perceptions_of_Community_Groups_to_Improve_MNCH_in_Urban_Slums_an_Exploratory_Case_Study_of_Korail_Slum_in_Dhaka [accessed 2025-05-19]
8. Biplob P, Sarker DC, Sarker RC. Assessment of water supply and sanitation facilities for Korail slum in Dhaka city. IJCEE IJENS 2011;11(5):115-118 [FREE Full text]
9. Bashar R, Tonmoy SS, Ria AF, Khan NA. Assessing the real-life socio-economic scenario of established slums in Dhaka: the cases of Korail and Sattola. Eur Online J Nat Soc Sci 2020;9(2) [FREE Full text]
10. Huckvale K, Venkatesh S, Christensen H. Toward clinical digital phenotyping: a timely opportunity to consider purpose, quality, and safety. NPJ Digit Med 2019;2(1):88. [doi: 10.1038/s41746-019-0166-1] [Medline: 31508498]
11. Liang Y, Zheng X, Zeng DD. A survey on big data-driven digital phenotyping of mental health. Inf Fusion 2019 Dec;52:290-307. [doi: 10.1016/j.inffus.2019.04.001]
12. Barnett I, Torous J, Staples P, Sandoval L, Keshavan M, Onnela JP. 151. Relapse prediction in schizophrenia through digital phenotyping. Biol Psychiatry 2018 May;83(9):S61-S62. [doi: 10.1016/j.biopsych.2018.02.169]
13. Labrique A, Vasudevan L, Mehl G, Rosskam E, Hyder AA. Digital health and health systems of the future. Glob Health Sci Pract 2018 Oct 10;6(Suppl 1):S1-S4. [doi: 10.9745/GHSP-D-18-00342] [Medline: 30305334]
14. Onnela JP, Rauch SL. Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. Neuropsychopharmacology 2016 Jun;41(7):1691-1696. [doi: 10.1038/npp.2016.7] [Medline: 26818126]
15. Alloh FT, Regmi P, Onche I, Teijlingen EV, Trenoweth S. Mental Health in low-and middle income countries (LMICs): going beyond the need for funding. Health Prospect 2018 Jun 19;17(1):12-17. [doi: 10.3126/hprospect.v17i1.20351]
16. Martinez-Martin N, Greely HT, Cho MK. Ethical development of digital phenotyping tools for mental health applications: Delphi Study. JMIR Mhealth Uhealth 2021 Jul 28;9(7):e27343. [doi: 10.2196/27343] [Medline: 34319252]
17. Leightley D, Bye A, Carter B, et al. Maximizing the positive and minimizing the negative: social media data to study youth mental health with informed consent. Front Psychiatry 2022;13:1096253. [doi: 10.3389/fpsyt.2022.1096253] [Medline: 36704745]

XSL•FO
RenderX

18. Torous J, Staples P, Barnett I, Sandoval LR, Keshavan M, Onnela JP. Characterizing the clinical relevance of digital phenotyping data quality with applications to a cohort with schizophrenia. NPJ Digit Med 2018;1:15. [doi: 10.1038/s41746-018-0022-8] [Medline: 31304300]
19. Zehra T, Parwani A, Abdul-Ghafar J, Ahmad Z. A suggested way forward for adoption of AI-enabled digital pathology in low resource organizations in the developing world. Diagn Pathol 2023 May 18;18(1):68. [doi: 10.1186/s13000-023-01352-6] [Medline: 37202805]
20. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
21. Pope C, Ziebland S, Mays N. Qualitative research in health care. analysing qualitative data. BMJ 2000 Jan 8;320(7227):114-116. [doi: 10.1136/bmj.320.7227.114] [Medline: 10625273]
22. Ahmed SK. The pillars of trustworthiness in qualitative research. J Med Surg Public Health 2024 Apr;2:100051. [doi: 10.1016/j.glmedi.2024.100051]
23. Braun V, Clarke V. Toward good practice in thematic analysis: avoiding common problems and be(com)ing a knowing researcher. Int J Transgend Health 2023;24(1):1-6. [doi: 10.1080/26895269.2022.2129597] [Medline: 36713144]
24. Ferretti A, Vayena E, Blasimme A. Unlock digital health promotion in LMICs to benefit the youth. PLOS Digit Health 2023 Aug;2(8):e0000315. [doi: 10.1371/journal.pdig.0000315] [Medline: 37540713]
25. Witter S, Sheikh K, Schleiff M. Learning health systems in low-income and middle-income countries: exploring evidence and expert insights. BMJ Glob Health 2022 Sep;7(Suppl 7):e008115. [doi: 10.1136/bmjgh-2021-008115] [Medline: 36130793]
26. Adanijo A, McWilliams C, Wykes T, Jilka S. Investigating mental health service user opinions on clinical data sharing: qualitative focus group study. JMIR Ment Health 2021 Sep 3;8(9):e30596. [doi: 10.2196/30596] [Medline: 34477558]
27. Nebeker C, Lagare T, Takemoto M, et al. Engaging research participants to inform the ethical conduct of mobile imaging, pervasive sensing, and location tracking research. Transl Behav Med 2016 Dec;6(4):577-586. [doi: 10.1007/s13142-016-0426-4] [Medline: 27688250]
28. Takashima K, Maru Y, Mori S, Mano H, Noda T, Muto K. Ethical concerns on sharing genomic data including patients' family members. BMC Med Ethics 2018 Jun 18;19(1):61. [doi: 10.1186/s12910-018-0310-5] [Medline: 29914459]
29. Martinez-Martin N, Insel TR, Dagum P, Greely HT, Cho MK. Data mining for health: staking out the ethical territory of digital phenotyping. NPJ Digit Med 2018;1(1):68. [doi: 10.1038/s41746-018-0075-8] [Medline: 31211249]
30. De Weger E, Van Vooren N, Luijkx KG, Baan CA, Drewes HW. Achieving successful community engagement: a rapid realist review. BMC Health Serv Res 2018 Apr 13;18(1):285. [doi: 10.1186/s12913-018-3090-1] [Medline: 29653537]
31. Juárez-Ramírez R. User-centered design and adaptive systems: toward improving usability and accessibility. Univ Access Inf Soc 2017 Jun;16(2):361-363. [doi: 10.1007/s10209-016-0480-1]
32. Torous J, Kiang MV, Lorme J, Onnela JP. New tools for new research in psychiatry: a scalable and customizable platform to empower data driven smartphone research. JMIR Ment Health 2016 May 5;3(2):e16. [doi: 10.2196/mental.5165] [Medline: 27150677]
33. Khan S, Rathore D, Singh A, Kumari R, Malaviya P. Socio-economic and environmental vulnerability of urban slums: a case study of slums at Jammu (India). Environ Sci Pollut Res 2023 Nov 3;31(12):18074-18099. [doi: 10.1007/s11356-023-30630-5]
34. De Beuckelaer A, Zeeman M, Van Trijp H. Assessment of the cross-national validity of an end-anchored 9-point hedonic product liking scale. Qual Quant 2015 May;49(3):1267-1286. [doi: 10.1007/s11135-014-0049-0]
35. Temple B, Young A. Qualitative research and translation dilemmas. Qual Res 2004 Aug;4(2):161-178. [doi: 10.1177/1468794104044430]
36. Audulv Å, Hall EOC, Kneck Å, et al. Qualitative longitudinal research in health research: a method study. BMC Med Res Methodol 2022 Oct 1;22(1):255. [doi: 10.1186/s12874-022-01732-4] [Medline: 36182899]
37. Venkatesh V, Thong JY, Xu X. Consumer acceptance and use of information technology: extending the Unified Theory of Acceptance and Use of Technology. MIS Q 2012;36(1):157. [doi: 10.2307/41410412]
38. Hanckel B, Petticrew M, Thomas J, Green J. The use of Qualitative Comparative Analysis (QCA) to address causality in complex systems: a systematic review of research on public health interventions. BMC Public Health 2021 May 7;21(1):877. [doi: 10.1186/s12889-021-10926-2] [Medline: 33962595]
39. Bluestone J, Johnson P, Fullerton J, Carr C, Alderman J, BonTempo J. Effective in-service training design and delivery: evidence from an integrative literature review. Hum Resour Health 2013 Oct 1;11(51):51. [doi: 10.1186/1478-4491-11-51] [Medline: 24083659]

Abbreviations:

COREQ: Consolidated Criteria for Reporting Qualitative Research
DP: digital phenotyping
FGD: focus group discussion

XSL•FO
RenderX

HIC: high-income country
LMIC: low- and middle-income country
SMD: serious mental disorder

Edited by A Mavragani; submitted 18.08.24; peer-reviewed by A Thakur, S Erturk; revised version received 12.02.25; accepted 14.02.25; published 23.06.25.

Please cite as:
Alam N, Giacco D, Siddiqi B, Singh SP, Jilka S
Investigating Awareness and Acceptance of Digital Phenotyping in Dhaka's Korail Slum: Qualitative Study
JMIR Form Res 2025;9:e65530
URL: https://formative.jmir.org/2025/1/e65530
doi:10.2196/65530

© Nadia Alam, Domenico Giacco, Bulbul Siddiqi, Swaran P Singh, Sagar Jilka. Originally published in JMIR Formative Research (https://formative.jmir.org), 23.6.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Mobile Apps to Prevent Violence Against Women and Girls (VAWG): Systematic App Research and Content Analysis

Md Hafizur Rahman[1], MSc; Nasimul Ghani Usmani[1], MSc; Pradip Chandra[1], BPH; Ridwana Maher Manna[1], MPH; Anisuddin Ahmed[1], MSc; Mohammad Sohel Shomik[2], MPH; Shams El Arifeen[1], DrPH; Aniqa Tasnim Hossain[1], MSc; Ahmed Ehsanur Rahman[1], PhD

[1]Maternal and Child Health Division, International Centre for Diarrhoeal Disease Research, Bangladesh, 68, Shaheed Tajuddin Ahmed Sarani, Dhaka, Bangladesh

[2]Nutrition Research Division, International Centre for Diarrhoeal Disease Research, Bangladesh, Dhaka, Bangladesh

Corresponding Author:
Md Hafizur Rahman, MSc
Maternal and Child Health Division, International Centre for Diarrhoeal Disease Research, Bangladesh, 68, Shaheed Tajuddin Ahmed Sarani, Dhaka, Bangladesh

Abstract

Background: Numerous reviews have explored specific aspects of violence prevention apps, but given the rapid development of new apps, increased violence during COVID-19, and gaps in understanding functionalities and geographical distribution, an updated review is needed.

Objective: Therefore, we aimed to systematically evaluate the trends, geographical distribution, functional categories, available features, and feature evolution of mobile apps designed to prevent violence against women and girls (VAWG).

Methods: We conducted a systematic search on app reselling platforms and search engines from April 24, 2024 to May 28, 2024, using terms related to VAWG in multiple languages. We included apps meeting our criteria for addressing VAWG, without restrictions on date or language. We conducted content analysis of app and apps were categorized by functionality and feature type. We performed descriptive analyses, trend analysis, co-occurrence network analysis, and geographical mapping.

Results: Out of 432 apps initially identified, 178 were included in the final analysis. Of these, 99 apps were available on both Google Play and the App Store, and 64 were exclusive to Google Play. Most apps were implemented in North America (48/178, 27%), followed by South Asia (31/178, 17%) and Europe and Central Asia (31/178, 17%). Emergency and support apps were most prevalent across regions. Most apps (132/178, 74%) originated from the private sector and were designed for survivor (121/178, 68%), were free without in-app purchases (100/178, 56%), had a website (148/178, 83%), and offered GPS features (142/178, 80%), but only 15% (27/178) provided offline functionality. App releases peaked in 2020 (33/178, 19%), followed by a decline. Regression analysis indicated a significant trend (P=.01) increase in app release, with a 2.40 unit increase per year before 2020 and a 7.01 unit decrease after, showing a post-2020 decline of 4.61 units per year. Apps were primarily categorized as emergency (n=110) or support (n=81), with most emergency apps in the 10,000 to ≥100,000 downloads range. Network analysis showed that emergency services (degree=10, clustering coefficient=0.911), location sharing (degree=10, clustering coefficient=0.911), SOS (Save Our Souls) alerts (degree=10, clustering coefficient=0.911), and educational resources (degree=10, clustering coefficient=0.911) features highly co-occurred in the same app. We found a gradual shift towards more sophisticated and comprehensive safety tools, evolving from basic GPS tracking and SOS alerts to advanced features such as real-time communication, panic buttons, peer support, and group communication, culminating in multifunctional platforms offering personalized safety, community engagement, and proactive risk identification.

Conclusions: Most apps to prevent VAWG emphasize emergency and support functions, and although initial releases increased, there has been a recent decline, with a shift towards integrating more comprehensive safety solutions such as communication, reporting, and community engagement. Future app development should prioritize cross-platform availability, offline functionality, public sector collaboration, and the integration of advanced technologies like artificial intelligence.

Trial Registration: PROSPERO CRD42024500431; https://www.crd.york.ac.uk/PROSPERO/view/CRD42024500431

(JMIR Form Res 2025;9:e66247) doi:10.2196/66247

KEYWORDS

violence against women and girls; mHealth; mobile app; smartphone app; prevention; mobile health

XSL•FO
RenderX

Introduction

Violence against women and girls (VAWG) remains a pervasive global health issue, with profound physical, psychological, and social impacts. The statistics are harrowing: VAWG affects 1 in 3 women in their lifetime, 200 million women experienced genital manipulation, 7% of women experienced sexual assault by a nonpartner, 35% of women experienced physical or sexual violence, and 38% of murders of women are committed by intimate partners [1]. Every 11 minutes, a woman or girl is murdered by a family member or intimate partner, resulting in nearly 50,000 such deaths annually [2]. VAWG results in significant physical and mental health impacts, including injuries, psychological trauma, and an increase in chronic health issues. In addition, the economic costs are substantial, affecting health care systems, productivity, and families, with survivors bearing considerable medical and mental health care expenses [3].

Interventions to prevent VAWG have traditionally included a range of legal, social, community, and technology-based approaches. With the advent of the digital revolution, technology has significantly enhanced and expanded these interventions [4]. Digital technologies have shown immense potential to prevent VAWG [5]. The integration of digital technologies, particularly mobile apps, into VAWG prevention and response strategies represents a relatively nascent and rapidly evolving frontier. Mobile apps offer potential interventions ranging from emergency assistance to educational resources. Besides, mobile apps provide information, support, and intervention services to survivors of violence. These tools range from safety planning and assessment applications to platforms promoting behavior change among perpetrators [6-10]. As a result, mobile apps were found to be useful for women's personal safety, preventing violence, and accessing support [10]. One study reported that mobile apps were perceived to be more effective than pepper spray in ensuring personal safety [11].

Between 2019 and 2022, numerous review studies analyzed mobile apps and focused on specific facets such as personal safety [6,10], antirape [7], domestic violence [12], and intimate partner violence [9]. Most previous reviews on mobile apps often focused on specific forms of violence and overlooked others. The first comprehensive review was conducted by Eisenhut et al [8] in August 2018, provided an initial analysis and functional categorization of mobile apps addressing VAWG through a systematic review . However, in 2024, approximately 1000 new apps are released daily on the Google Play Store alone [13]. The rapid development of new apps in recent years necessitates an updated review to capture recent innovations and emerging trends. Besides, during the COVID-19 pandemic in 2020, there was a significant increase in VAWG [14], underscoring the need to examine the trends in app usage and development during this period.

In addition, previous studies did not include detailed analyses of web-based or offline capabilities and the country or region-wide distribution of mobile app functionalities designed to prevent VAWG. Such knowledge is crucial for understanding key components of app functionalities and their distribution across countries, guiding the development of more targeted and effective interventions to prevent VAWG. Furthermore, previous studies also revealed a significant lack of extensive research on mobile apps intended to prevent VAWG [15,16]. Therefore, we aimed to systematically evaluate the trends, geographical distribution, functional categories, available features, and feature evolution of mobile apps designed to prevent VAWG.

Methods

Study Design and Search

To identify the apps, we conducted a systematic search on app reselling platforms and search engines between April 24, 2024, and May 28, 2024. The reselling platforms included Google Play Store, Apple App Store, and Microsoft Store. The search engines included Google, Bing, Baidu, Yandex, DuckDuckGo, Yahoo, Ecosia, and Ask.com. Key search terms included terms related to VAWG and its different forms. A detailed list of the search terms is provided in Multimedia Appendix 1.

The search was conducted in multiple languages, including English, Spanish, French, Portuguese, Russian, Chinese, Bangla, German, Swedish, Japanese, Korean, Urdu, Hindi, Vietnamese, and Malay. Google Translator was used to translate English search terms into other languages to ensure coverage across different regions and languages. In addition, apps identified in previous systematic reviews up to May 2024 and still available on reselling platforms were also included. The app search was conducted jointly by 2 authors (NGU and PC) to ensure accuracy. We adhered to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines in constructing this review, as outlined in Checklist 1 [17]. A protocol was developed before conducting the review, and the protocol was registered in the International Prospective Register of Systematic Reviews (PROSPERO; CRD42024500431) and there were no deviations from the registered protocol.

Eligibility Criteria

We included all mobile apps found and available on the specified web-based or platform-specific app stores, with no date and language restrictions. The inclusion criteria focused on apps specifically designed to address VAWG, available on Google Play Store, Apple App Store, Microsoft Store, or as Websites. Apps were required to have at least 1 functional feature directly related to the safety, support, or prevention of VAWG. We did not have any date, number of downloads, and language restrictions.

Data Extraction

We conducted a content analysis of the app's features. App content was extracted by installing the app on a mobile device. If an app was not freely available or purchase was required for installation or use, we purchased the app. Besides, contents were also extracted from the "about this app" section on the Google Play Store, the "app support" section on the Apple App Store and from the app's corresponding website. Data extraction was conducted jointly by 2 authors (NGU and PC) to ensure accuracy and completeness. Microsoft Excel was used for the systematic management of app identification, screening, and data extraction.

XSL•FO
RenderX

A structured data extraction form in Excel was used to capture the following variables: app name, app details (description, functionality, and target audience), functional categories, country of initiation, region of initiation, country of implementation, region of implementation, owner institution name, developer institution name, year of first release, year of last update, number of downloads, number of ratings, overall ratings, sector (public, private, or both), target groups (survivors only, nonsurvivors only, or both), primary source of the app (reselling platform or search engine), availability on reselling platforms (web only, App Store, Google Play, Windows Store, or multiple platforms), cost (free without in-app purchases, free with in-app purchases, not free, and unclear), type of VAWG addressed, overall objective of the app, all features, GPS or navigation feature (yes, no), offline mode (yes, no), content language, website availability (yes, no), website link, and whether an external device is required (yes, no, and the name of the external device if applicable). Discrepancies in the app selection process, data extraction, and analysis were resolved through a secondary review by the authors and discussion among authors to ensure accuracy and consistency.

Categorization

Apps were categorized into functional categories based on the functionality of features. The functional categories were adopted from a previous systematic review [8]. A detailed definition of functional categories is given in Multimedia Appendix 1. If an app fell into multiple functional categories, it was assigned to all relevant categories. The functional categories included:

1. Emergency response: Apps providing quick access to emergency contacts and services.
2. Educational: Apps offering information and resources on violence prevention and legal rights.
3. Support services: Apps connecting users to support groups, counseling, and helplines.
4. Reporting tools: Apps enabling users to report incidents of violence anonymously or directly to authorities.
5. Safety planning: Apps assisting in the creation of personalized safety plans.

In addition, features were classified according to their type as follows: emergency and safety alerts, location tracking and sharing, safety tips and guidance, communication and support, education and awareness, reporting and documentation, privacy and security measures, and community engagement and collaboration. If an app's feature belongs to multiple feature types, the app was classified under all applicable types. Details of feature categories and corresponding features are given in Multimedia Appendix 1.

Furthermore, countries of initiation and implementation were categorized into regions according to World Bank regions [18]. Offline mode was defined as the app's ability to operate without an internet connection. Cost was defined as follows:

1. "Free without in-app purchases" indicates the app is free with no additional costs.
2. "Free with in-app purchases" indicates the app is free to download but includes optional purchases.
3. "Not free" indicates the app requires payment for download.
4. "Unclear" indicates insufficient information to determine the cost structure.

Analysis

The analysis of mobile app release trends was conducted using the R (R Core Team) programming language [19], using a structured statistical approach to evaluate temporal patterns from 2010 to May 2024. We began by compiling annual frequency data, which was organized into a data frame for systematic analysis. To visualize these trends, a line plot was created using the *"ggplot2"* package [20], with data points and labels included to enhance interpretability. Cumulative frequency analysis was performed using *"ggplot2"* to illustrate the overall growth trajectory of app releases over time. We used linear regression modeling using *"ggplot2"* to quantify the relationship between the year and the number of app releases, assessing statistical significance through the P value of the regression slope. The regression line and CIs were added to the plot to depict model fit and variability. In addition, segmented regression analysis and plotting were conducted using the "*segmented*" package [21], with a predefined breakpoint in 2020, to identify potential changes in the rate of app releases before and after the onset of the COVID-19 pandemic.

In addition, using the *"igraph"* [22] and *"tidyverse"* [23] packages in R, we constructed and visualized a co-occurrence network to analyze features that are mostly present together in an app. A co-occurrence matrix was developed to represent the frequency of feature pairings, which was subsequently transformed into an *"igraph"* object for network analysis. In the visualization, nodes denoted features, and edges signified the weighted frequency of co-occurrences, with node sizes scaled by the degree to indicate prominence. The network was plotted using the Fruchterman-Reingold layout for enhanced visual separation [24]. Besides, we computed key network statistics, including node degree, betweenness centrality, closeness centrality, eigenvector centrality, clustering coefficient, average path length, and network density, to assess the structural properties and significance of features within the network (Multimedia Appendix 1).

Furthermore, we conducted descriptive analyses using Microsoft Excel for app availability on reselling platforms, feature types, and the most prevalent (features that were available more than once) features by year and percent distribution of sectors, target groups, cost, website availability, GPS functionality, and offline mode. Besides, we analyzed the overall frequency distribution of functional categories, as well as the distributions by sector, number of downloads, and GPS functionality. In functional category plotting, responses were categorized as multicategory classification to account for instances where apps could be classified under more than one functional category. Geographical maps were developed by Adobe Illustrator to present the regional distribution of maps and their functional categories.

XSL•FO
RenderX

Results

App and Reselling Platform

Out of 432 apps initially identified through reseller platforms and search engines, 178 apps were included in the final analysis after duplicate removal and eligibility screening (Figure 1). Of the 178 apps, 99 apps were available on both Google Play and the App Store, 64 apps were available exclusively on Google Play, and 13 apps were exclusive to the App Store (Figure 1).

North America reported the highest implementation rate (48, 27%), with Europe and Central Asia (31, 17%), and South Asia (31, 17%) each contributing 31 apps. In North America, emergency (n=29) and support (n=30) apps were most prevalent. Similarly, in South Asia, Europe, and Central Asia, emergency (n=24 and n=19, respectively) and support apps (n=15 and n=14, respectively) were predominantly implemented (Figure 2).

Figure 1. ofPreferred Reporting Items for Systematic Reviews and Meta-Analyses flow diagram and depiction of the screening and inclusion process. APK: Android Package Kit

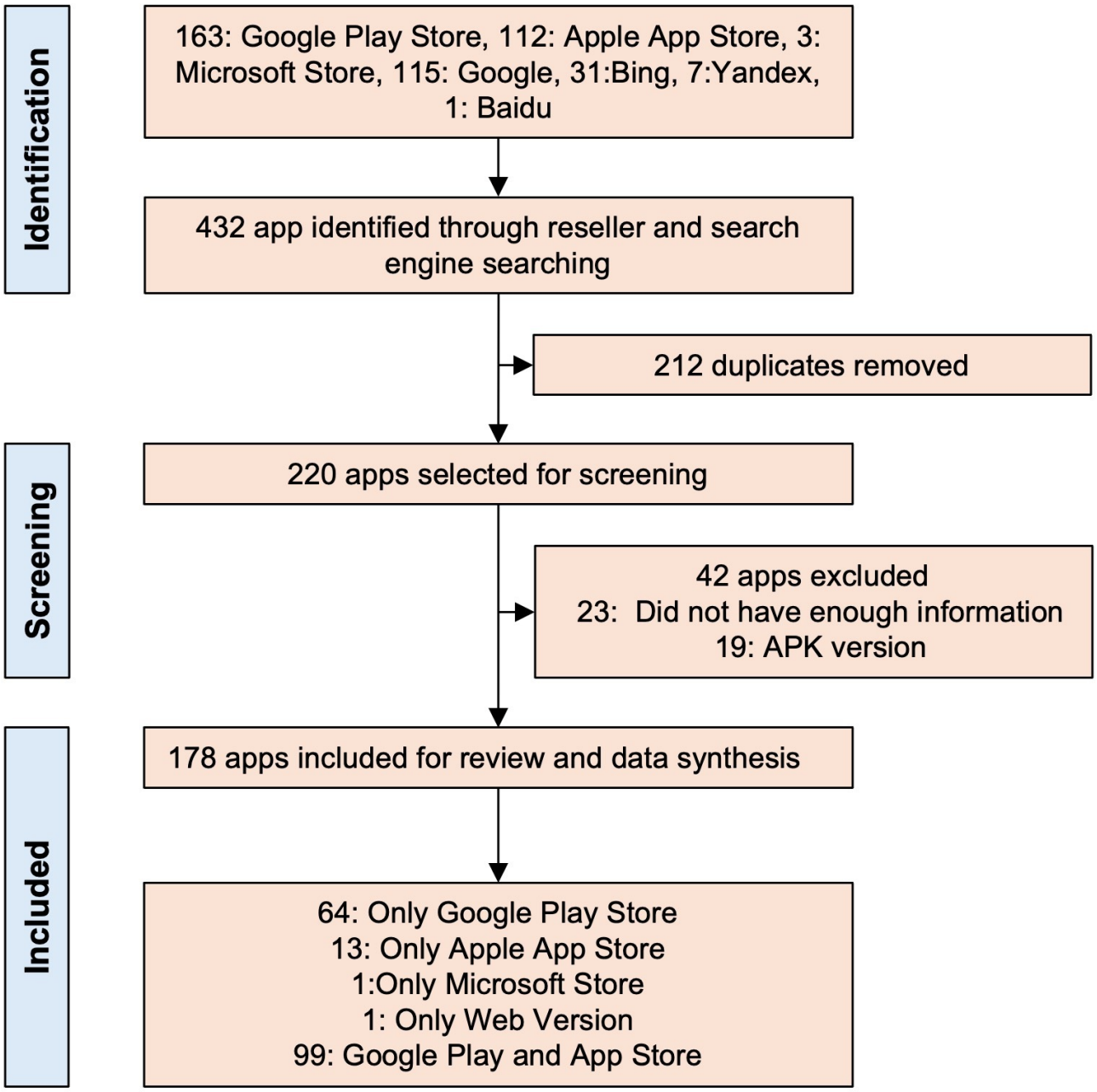

XSL•FO
RenderX

Figure 2. Global distribution of mobile apps: (A) Frequency and percentage by region, (B) Functional category classification across regions in percentage (multicategory classification) (N=178)

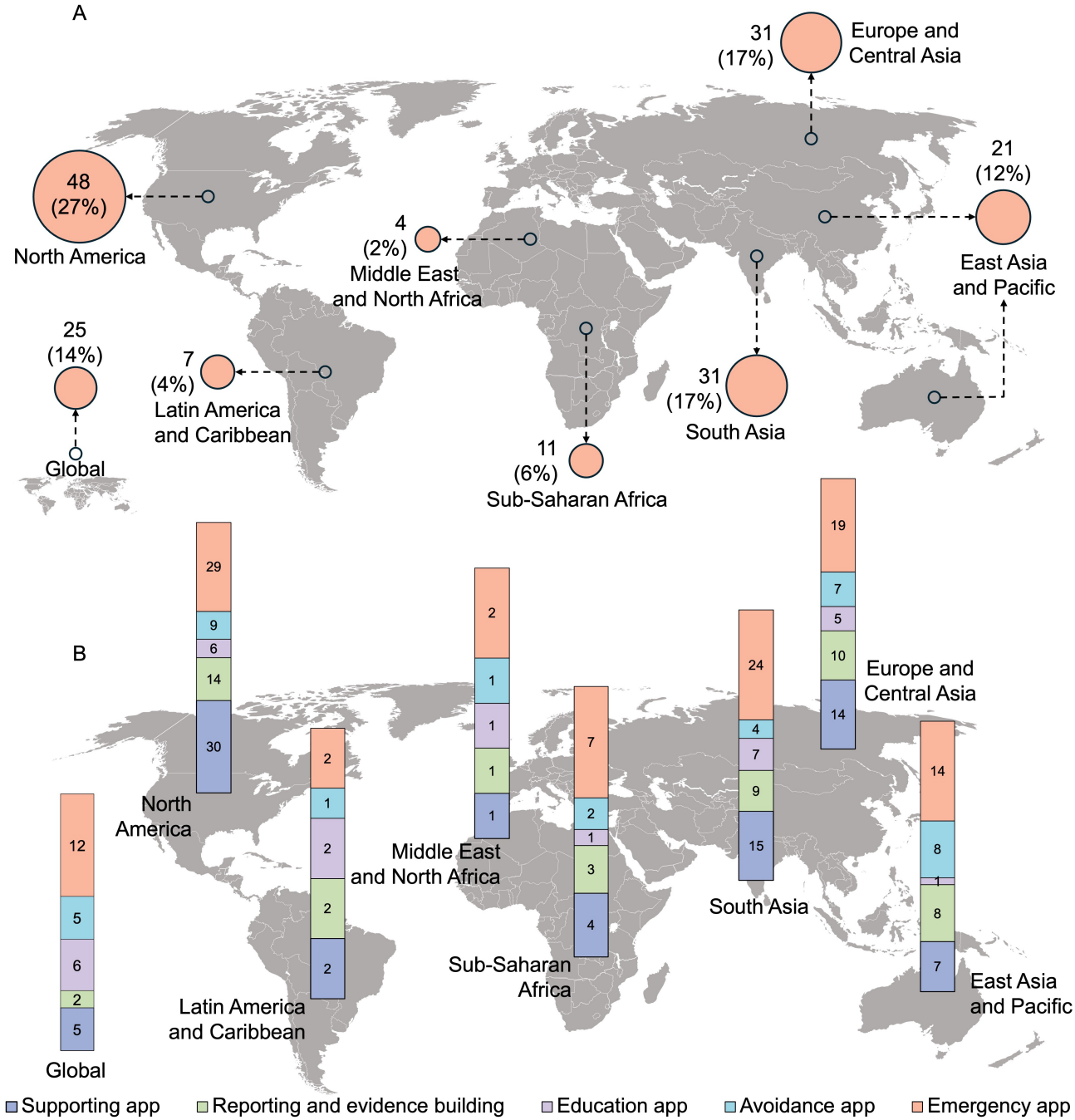

Trends of App Release

Figure 3 illustrates the trend and cumulative number of app releases over the years and the trend analysis of app releases with regression modeling. The trend graph shows a gradual increase in app releases until reaching a peak in 2020 with 33 apps, followed by a significant decline thereafter. Besides, the cumulative graph indicates a steady overall growth from 2 to 178 between 2010 and May 2024. The regression line, with a P=.01, suggests a statistically significant increasing trend in app releases over the years. The shaded area represents the 95% CI for the regression line indicating a good fit for the linear model.

Figure 4 demonstrates the segmented regression analysis, identifying 2020 as a breakpoint. The segmentation line suggests an initial increase in app releases until 2020, followed by a marked decline. Furthermore, the segmented regression model in Multimedia Appendix 1 presents that before the breakpoint in 2020, the number of app releases increased by approximately 2.40 units per year (P<.001). After 2020, the trend shifts, showing a decrease in the rate of change with a reduction of 7.01 units per year, indicating an overall decrease of 4.61 units per year in the postbreakpoint trend.

XSL•FO
RenderX

Figure 3. Trends in mobile app releases over time: (A) cumulative number of app releases by year and (B) annual trend in app releases with linear regression modeling (N=178).

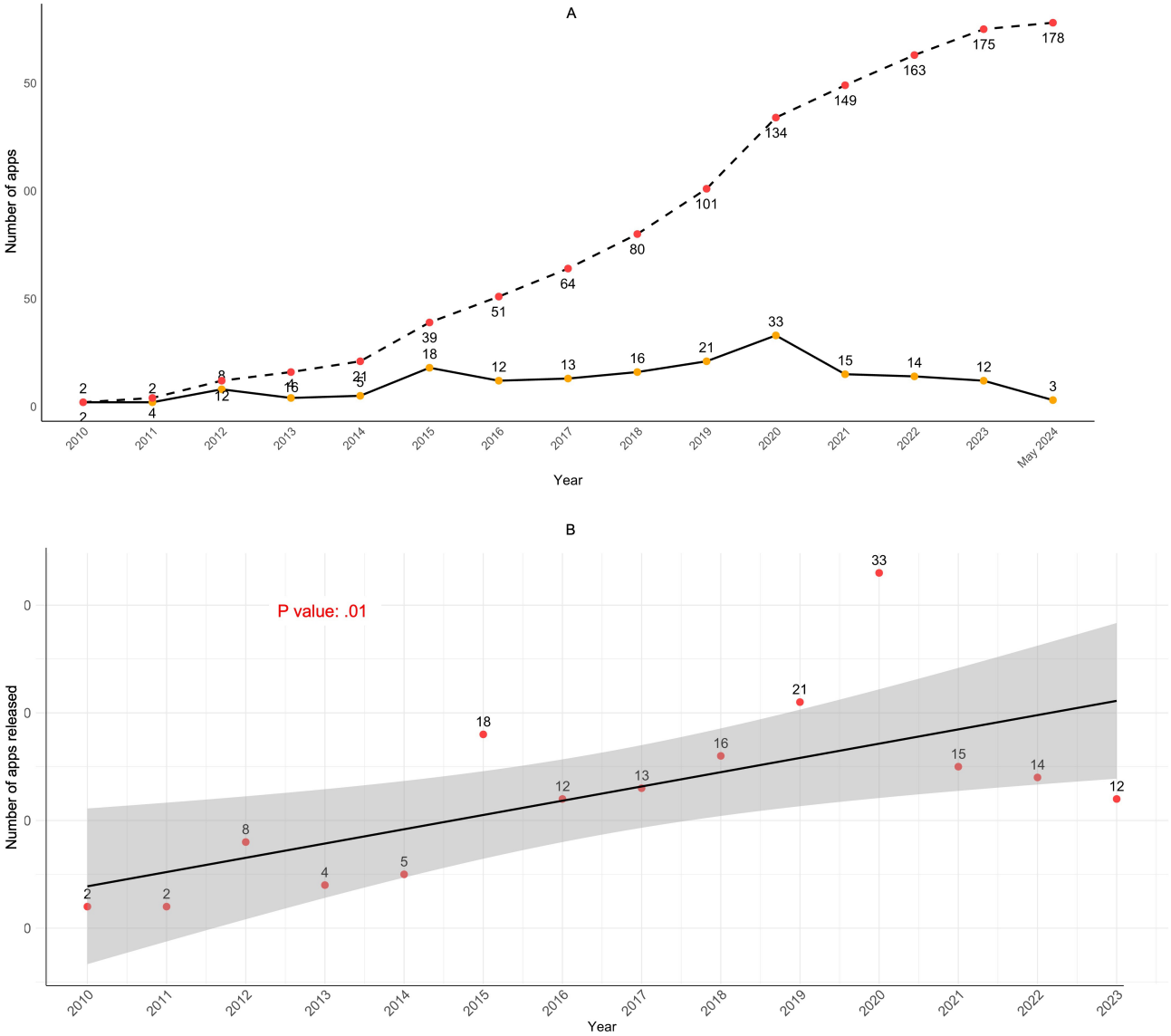

Figure 4. Segmented regression analysis of app releases with 2020 breakpoint (N=178).

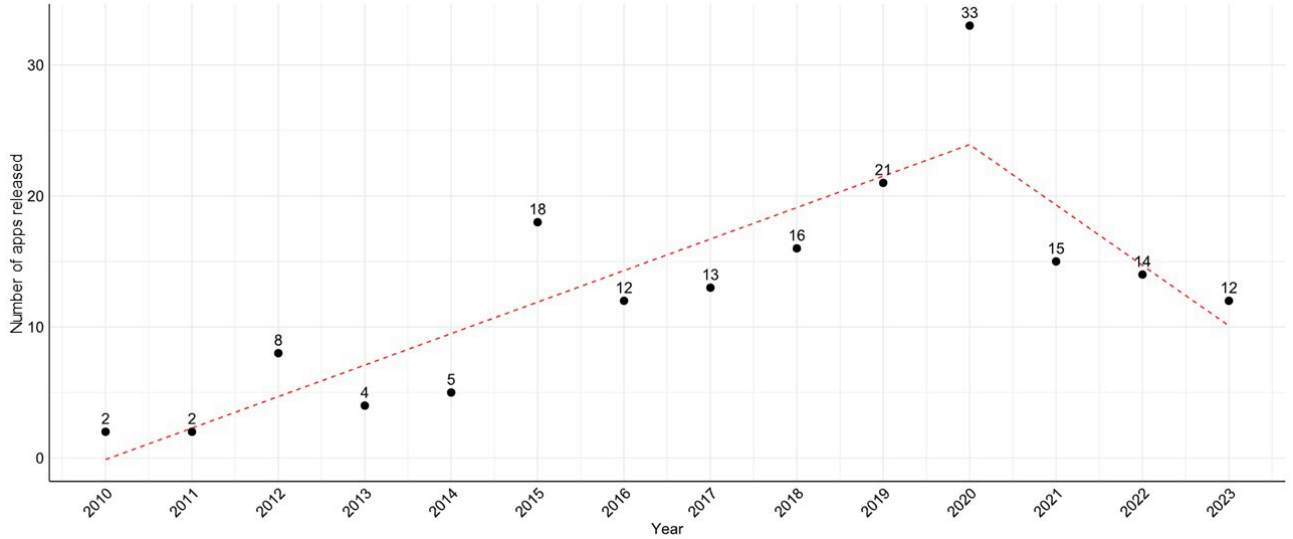

XSL•FO
RenderX

Distribution of Apps

Table 1 presents the frequency and percent distribution of apps by sector, target groups, cost, website availability, GPS availability, and offline mode. In total, 178 apps were included in this analysis. Most apps were from the private sector (n=132, 74%), with fewer from the public sector (n=32, 18%). Most apps were designed specifically for survivors (n=121, 68%), while a smaller proportion cater to both survivors and nonsurvivors (n=57, 32%). In terms of cost, most apps were available for free without in-app purchases (n=100, 56%), while others were free but included in-app purchases (n=39, 22%). Furthermore, most apps had a website (n=148, 83%), and a considerable majority offered GPS or navigation features (n=142, 80%). However, only a small proportion provided offline functionality (n=27, 15%), while the majority did not include this feature (n=152, 85%).

Table 2 presents the frequency distribution of apps by feature type. The most common feature type was emergency and safety alerts, found in 149 apps, followed by location tracking and sharing in 90 apps and safety tips and guidance in 80 apps.

Table 3 presents the trend of prevalent features in mobile apps designed to prevent VAWG from 2010 to May 2024. Initially, the focus was on basic functions like GPS location tracking and SOS alerts (2010 - 2011). Over time, features became more sophisticated, with the introduction of real-time location tracking, emergency calls, and crisis support systems (2012 - 2013).

By 2014 - 2016, additional safety tools like panic buttons, real-time communication, and discreet reporting were developed. From 2017 onward, apps integrated more comprehensive support services, such as educational resources, peer support, and voice activation, evolving into multifunctional platforms offering personalized safety, community engagement, and real-time emergency assistance. In recent years (2019 - 2024), apps have expanded to include group communication, safe journey planning, and risk area identification.

Table . Frequency and percent distribution of apps by sector, target groups, cost, website availability, GPS availability, and offline mode (N=178).

Variable and Category	n (%)
Sectors	
Public	32 (18)
Private	132 (74)
Both	14 (8)
Target groups	
Survivors	121 (68)
Survivors and nonsurvivors	57 (32)
Cost	
Free without in-app purchase	100 (56)
Free with in-app purchase	39 (22)
Not free	35 (20)
Unclear	4 (2)
Website	
Yes	148 (83)
No	30 (17)
GPS	
Yes	142 (80)
No	36 (20)
Offline mode	
Yes	27 (15)
No	151 (85)

XSL•FO
RenderX

Table . Frequency distribution of apps by feature type (multicategory classification) (N=178).

Feature	Frequency of apps
Safety alerts	149
Location tracking and sharing	90
Safety tips and guidance	80
Communication and support	72
Reporting and documentation	54
Privacy and security measures	19
Education and awareness	16
Community engagement and collaboration	14

XSL•FO
RenderX

Table . List of app features available more than once corresponding to year of release.

Year	Prevalent features
2010	• GPS location tracking • SOS[a] alerts
2011	• Location information • Notification services
2012	• Real-time location tracker • Emergency calls or alerts • Contact response team • Crisis support • Reporting systems • Create safe boundaries for added security
2013	• Deliver important safety information • Real-time location tracker • Access to safety resources and phone numbers
2014	• GPS location tracking • Emergency button • Emergency assistance • Tracking maps, journey tracking, or mutual tracking
2015	• GPS location tracking • Panic buttons or send emergency and notification messages • Consolidated information on specialized medical services • Safe space for women • Real-time communication • Reporting resources • Talk or chat with someone at anytime • Advocacy and support services • Provides steps to preserve evidence
2016	• Real-time location tracker • GPS location tracking • Emergency button or panic button • Customize messages and notifications • Discreet reporting • Find nearby essential services
2017	• Panic buttons • Report incident • Live breaking video • Real-time location tracker • Instantly notifies police • Educational resources • Resource directory • Comprehensive support
2018	• Emergency alerts or panic button • GPS location tracking • Real-time location tracker • Send emergency and notification messages • Medical information integration • Call for support • Direct chat support • Volunteer for social projects • Learn about violence, abuse, and rights • Privacy-focused design

XSL•FO
RenderX

Year	Prevalent features
2019	• Emergency SOS alert or panic button • Real-Time location tracker • GPS location tracking • Reporting systems • Peer support program • Empower users • Advocacy and support services • Educational resources • Voice activation
2020	• Instant assistance at your fingertips • Verified emergency response • Real-time location tracker • Instant help from nearest police, hospitals, and fire • In-app messaging or group chat • Access to counseling for emotional support and guidance • Educational resources • Record incidents using multimedia • 24 × 7 Toll-Free helpline
2021	• Panic or emergency response button • Safe location sharing • Real-time location tracker • Reporting system • Quick alerts to emergency or trusted contacts • Safe journey planner • Safety or women-friendly place on map • Helpline numbers • Prompt dispatch of police assistance
2022	• SOS feature • Real-time location tracker • GPS location tracking • Create groups and add contacts • Reporting system • Risk area identification • Community support or access to support
2023	• SOS panic button alerts • Real-time location tracker • Community engagement through alerts and chats • Emergency or trusted contacts • Safety tips • Find nearby safe places • Private space for users
May 2024	• Emergency alert or panic button • GPS location tracking • Real-time location tracker

[a]SOS: Save Our Souls.

Functional Categories

Figure 5 presents the frequency distribution of apps by overall functional categories and the frequency distribution of functional categories of apps by sector, number of downloads, and GPS functionality. The majority of apps are categorized as an emergency (n=110), followed by supporting (n=81), and reporting and evidence building (n=57). In terms of sector distribution, most emergency apps are from the private sector (n=75), while the public sector contributes fewer apps (n=25). Similarly, supporting apps are predominantly from the private sector (n=56), with fewer from the public sector (n=17).

Regarding the number of downloads, most emergency apps fall into the 10,000 to ≥100,000 downloads category (n=71), whereas education apps have the fewest downloads in this category (n=11). For supporting apps, 44 are in the 10,000 to ≥100,000 downloads range. In terms of GPS functionality, the vast majority of emergency apps (n=106) and supporting apps (n=60) include GPS (Figure 5).

Figure 6 presents the network analysis of feature co-occurrence. Location Track and Share, Emergency Services, Education and Awareness, and SOS Alerts have the highest degree (10) and clustering coefficient (0.911), indicating they are well-connected and form tight clusters. However, Community Engagement and

XSL•FO
RenderX

Collaboration and Privacy and Security Measures stand out with higher betweenness centrality (0.585 and 0.407, respectively), serving as critical connectors. Access to Support Information has the highest closeness centrality (0.256), while Location Tracking leads in eigenvector centrality (1.0), highlighting its overall influence in the network (Figure S7 Multimedia Appendix 1).

Figure 5. Distribution of apps by functional categories: (A) overall frequency; (B) distribution by sector; (C) distribution by number of downloads; and (D) distribution by GPS functionality (multi-category classification) (N=178).

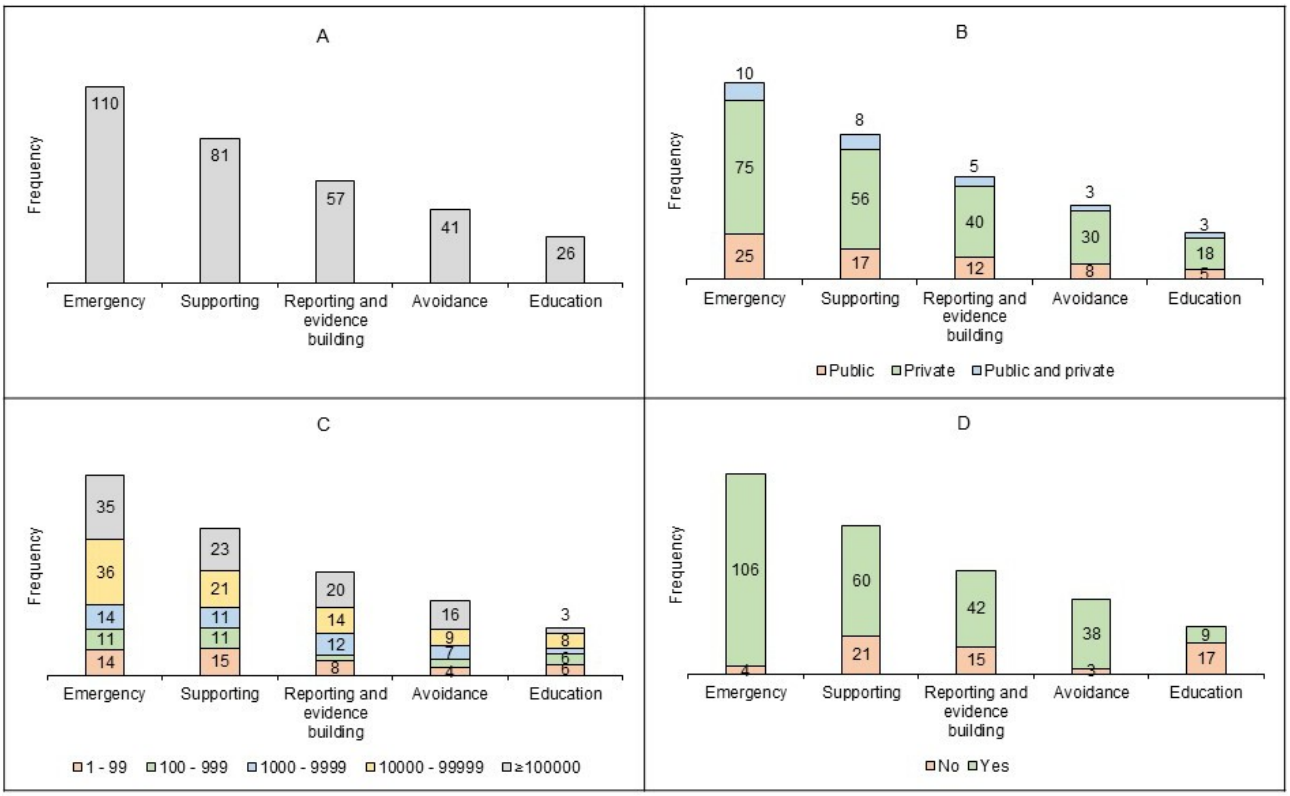

XSL•FO
RenderX

Figure 6. Network analysis of feature co-occurrence. SOS: save our souls.

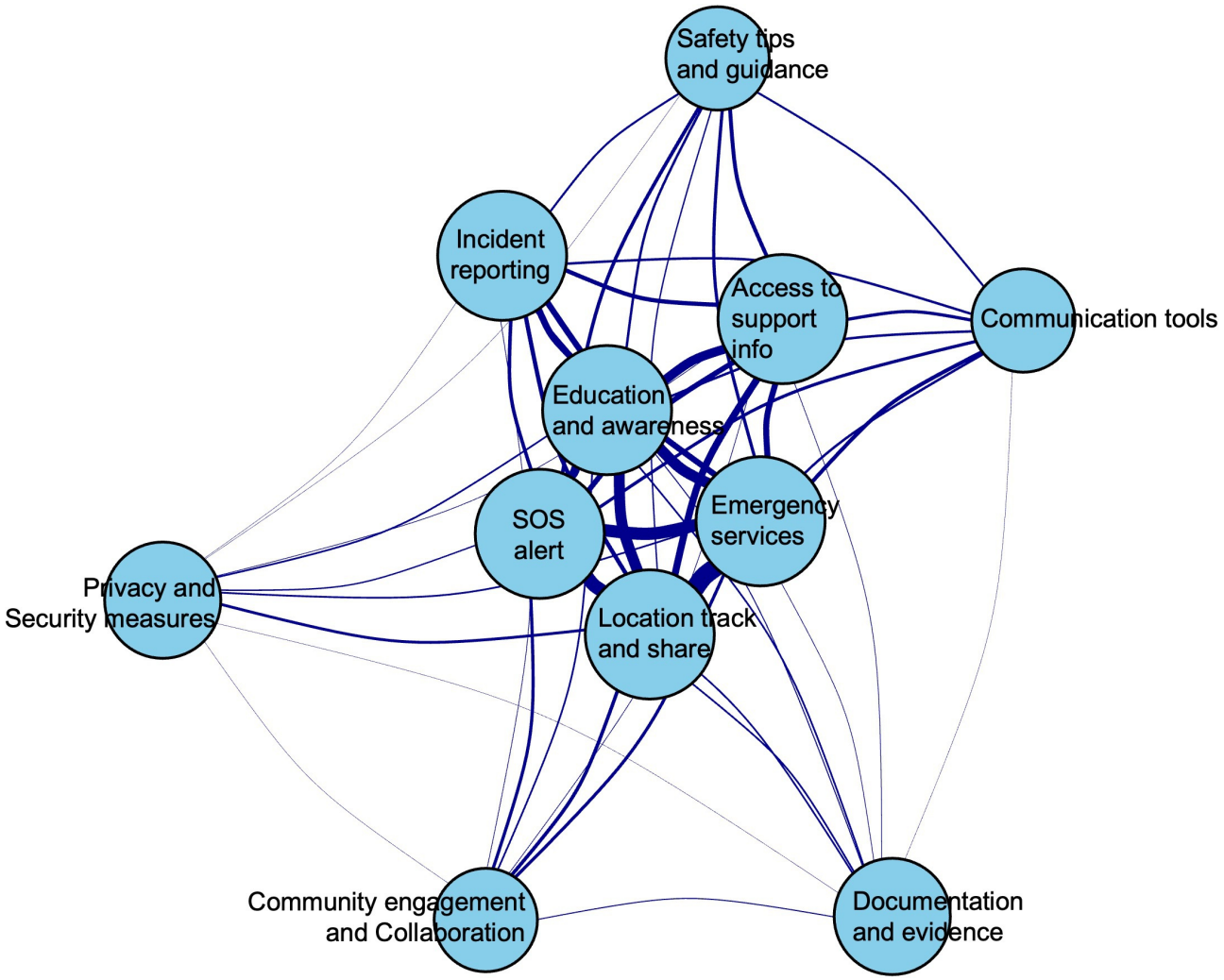

Discussion

Principal Findings and Comparison With Previous Work

Our findings reveal that while the majority of apps designed to prevent VAWG predominantly focus on emergency and support functionalities, there has been a notable shift in recent trends. After an initial surge in app releases, a decline has emerged, accompanied by a growing emphasis on integrating more holistic safety solutions, including features for communication, reporting, and community engagement.

Most apps being available on both Google Play and the App Store suggests that developers recognize the importance of cross-platform compatibility to effectively reach diverse user bases. However, many apps are available exclusively on Google Play, likely due to its open nature and large user base. A similar pattern of app availability was also reported in another systematic review [6]. By not being available on multiple platforms, especially on iOS, app developers miss the opportunity to reach a broader audience, thereby limiting their market reach and potential revenue streams.

The implementation of apps across regions shows a notable concentration in North America, which accounted for the highest number of apps, particularly in emergency and support categories. This trend reflects the region's advanced technological infrastructure and emphasis on leveraging mobile solutions for personal safety and violence prevention. North America leads in the launch and availability of apps compared to the Middle East, which may also explain the higher prevalence of mobile apps specifically aimed at preventing VAWG in this region [25]. Interestingly, a systematic review conducted in 2018 found that most apps were initially developed in South Asia [8], whereas our study focuses on the country of implementation rather than initiation. Our findings reveal that 14 apps were developed in one country and implemented in another. However, understanding the country of implementation instead of the country of development is crucial for grasping the dynamics of the need for such apps in those regions to address VAWG.

The temporal trends in app releases reveal a dynamic market response to societal needs and technological advancements. During the COVID-19 pandemic in 2020, the world experienced a considerable hike in VAWG [14] and the peak in app releases around 2020 reflects a period of intense innovation and response to increased concerns of gender-based violence issues. During the COVID-19 pandemic, digital interventions, especially for domestic and intimate partner violence, became increasingly valuable as social and physical isolation measures intensified,

XSL•FO
RenderX

leaving many survivors sequestered at home with abusive partners [26]. In addition, prepandemic digital tools, such as the myPlan app (Johns Hopkins University) [27], I-DECIDE (University of Melbourne) [28], and iSafe (University of Auckland) [29], which were already recognized for their efficacy in reducing violence and providing confidential support, saw heightened importance [26]. However, the subsequent decline warrants further research into whether a maturation phase in the market, where developers may have reached a saturation point or shifted focus to refining existing solutions rather than introducing new ones. This decline may also suggest the decreased popularity of these apps, which calls for further research.

Our study shows a strong private sector presence in the development of violence prevention apps, consistent with other findings that emphasize the entrepreneurial drive to innovate in this area [8]. However, the limited involvement of public sector entities suggests a missed opportunity for leveraging these technologies within broader social and policy frameworks. Increased public sector engagement could enhance the development and dissemination of apps that are both effective and accessible to diverse populations [30].

The major focus on survivors-centered solutions, coherent with other studies [8], underscores the critical role these apps play in providing immediate safety and support, yet there remains potential for expanding features to encompass broader community engagement and preventive measures. Besides, we found a considerable majority of the apps have an associated website. A web presence can enhance credibility and offer additional resources, serving as a channel for feedback, updates, and community engagement, thus enriching the user experience.

In addition, accessibility remains a key consideration, with the majority of apps available free of charge but often requiring in-app purchases for full functionality, which is coherent with another study [8]. This model may limit access for users with financial constraints, underscoring the importance of developing inclusive strategies that ensure all users can benefit from these technologies [10]. Given that individuals of low socioeconomic status face a higher risk of violence [31], it is crucial for effective apps to be freely accessible.

Furthermore, four-fifths of the apps offer GPS or navigation features, which is a critical function for users seeking immediate help or looking to locate nearby services and resources. Another systematic review also found a similar prevalence of location-sharing features [6,9,10]. GPS functionality can significantly enhance the utility of these apps, particularly in emergency situations where users need quick access to safe locations or services. In addition, the prevalent inclusion of GPS functionality in emergency and supporting apps underscores the critical role of location-based services in increasing the effectiveness of digital interventions. By providing users with real-time assistance and access to nearby resources, GPS-enabled features align with broader research indicating their capacity to improve outcomes in crisis situations [32], thereby enhancing the overall impact and utility of these apps in violence prevention efforts. Despite these positive aspects, only 15% of the apps provide offline functionality. The lack of offline features is a significant limitation, especially in areas with unstable internet connectivity or during situations where internet access may be unavailable. Offline functionality is essential for ensuring that users can access critical information, resources, or services at any time, regardless of their connectivity status.

The landscape of functional categories of mobile apps designed to prevent VAWG is primarily populated by those categorized as emergency and support tools, indicating a societal emphasis on enhancing personal safety and empowerment through immediate digital solutions. This aligns with findings in the literature [6,8-10], which highlight the importance of digital interventions in providing real-time assistance and fostering a sense of security among users. However, contrary to another systematic review that emphasized the prominence of avoidance and education apps, our study reveals that apps can be classified into multiple functional categories, whereas the other review categorized each app into a single category [8].

Features including safety alerts, location tracking and sharing, and safety tips and guidance were prevalent, which are crucial for providing timely and effective support. The widespread inclusion of these features is coherent with other studies [10,33] and underscores the critical role of technology in empowering individuals and communities to respond swiftly to threats and seek help when needed [33]. This trend also reflects broader societal recognition of the need for accessible and reliable safety tools, as research consistently shows that users prioritize functional capabilities that enhance their sense of security and provide tangible support in crisis situations [34].

The high download numbers of emergency apps suggest widespread accessibility and potential interest, but they do not necessarily indicate significant user engagement. The considerable download numbers of the emergency apps may demonstrate a widespread popularity in digital tools that offer immediate assistance and protection. This is supported by research indicating that users are more likely to engage with apps that provide immediate intervention capabilities [35]. Nevertheless, one study found that personal safety apps may reduce fear of crime but do not necessarily decrease vulnerability to victimization, as most apps focus on intervention during or after a criminal event rather than preventing it [6]. However, educational apps, which focus on raising awareness and providing information, have a smaller user base, suggesting that users prioritize functional capabilities that directly enhance safety over purely informational content.

The network analysis reveals that features such as Location Track and Share, Emergency Services, Education & Awareness, and SOS Alerts are frequently integrated together within mobile apps, as indicated by their high degrees and clustering coefficients. This co-occurrence suggests that developers often bundle these features to create comprehensive safety solutions that address multiple aspects of user needs in a cohesive manner. In contrast, features like Community Engagement & Collaboration and Privacy & Security Measures exhibit higher betweenness centrality, serving as critical connectors that bridge different app functionalities and facilitate a seamless user experience. In addition, the prominence of Location Tracking,

XSL•FO
RenderX

as indicated by its leading eigenvector centrality, underscores its overall influence and importance in the network, further highlighting developers' focus on integrating location-tracking features to enhance app efficacy [6,9,10].

Our study also reveals a gradual shift towards more sophisticated and comprehensive safety tools, evolving from basic GPS tracking and SOS alerts to advanced features, such as real-time communication, panic buttons, peer support, and group communication, culminating in multifunctional platforms offering personalized safety, community engagement, and proactive risk identification. This evolution reflects a growing recognition of the importance of holistic approaches that address both immediate threats and long-term support, mirroring broader trends and calls [36,37].

The incorporation of advanced technologies in mobile apps aimed at preventing VAWG underscores a significant evolution in digital safety solutions. Among these innovations, the use of features like feelings trackers exemplifies the integration of nuanced and responsive technology into personal safety apps. Feelings trackers provide users with a method to monitor and record emotional changes, offering valuable insights for both immediate and long-term support. Other notable advancements include the integration of the smart pepper spray, which combines traditional self-defense mechanisms with smart technology, allowing users to trigger alerts and send their location directly from the device. The use of SOS Sirens and Flash features enhances the visibility of emergency signals, while Discreet and Disguised Interfaces provide an added layer of privacy by making the app appear as another application, thus protecting user confidentiality in sensitive scenarios. Voice Activation and Commands, along with features such as Shake Alerts and Fake Calls, contribute to the ease of use and discreet nature of the apps, allowing users to activate safety measures without drawing attention. Automatic Fall Alerts, which detect sudden impacts and send alerts to emergency contacts, further exemplify how technology can pre-emptively address critical situations.

Although United Nations Women advocates for the use of artificial intelligence (AI)-driven technologies—particularly for monitoring hate speech, fact-checking, and countering disinformation on social media, with attention to misogynistic and harmful gendered narratives [38], we found a significant lack of apps using this technology. Future applications could leverage AI and mobile phone sensors to automatically detect and respond to threats without user intervention, thereby transitioning from reactive to proactive safety measures. Such integration also aligns with recent studies emphasizing the role of AI in enhancing predictive and automated safety technologies [39]. Expanding the use of these technologies could significantly improve the effectiveness and responsiveness of safety apps, creating a more comprehensive and anticipatory safety framework.

One limitation of our study is that the search terms used may not have captured all relevant apps available on reselling platforms, as the platform's search functionality itself may not retrieve all apps related to a given term. Besides, Google Translator was used to translate English search terms into other language, which may potentially cause errors or loss of context of the original search terms. Some apps were geo-restricted, limiting our ability to access and analyze details from outside specific countries, potentially leading to gaps in our search and analysis. We relied on diverse and broad search terms guided by previous studies, which may have introduced variability in app selection and limited the focus on apps specifically addressing violence prevention. Although this study included apps in different languages, we were unable to analyze whether language influences app features, which may be a limitation. Furthermore, we were unable to analyze app ratings and user feedback, which are critical components of app evaluation. We did not examine key metrics such as downloads, user satisfaction, user profiles, and valued features, which future research may prioritize, particularly with a focus on the pre- and post-COVID-19 impact on the use and user experience of these apps.

Conclusion

The analysis shows that most apps prioritize emergency and support functions, with notable implementation in the United States, India, and the United Kingdom. Despite initial growth, there has been a decline in new app releases after 2020. Most apps are from the private sector, designed for survivors, and commonly offer GPS functionality and a web presence but lack offline capabilities. Features such as emergency services, location sharing, SOS alerts, and educational resources features highly co-occurred in the same app. Initially, app focused on location sharing and emergency alerts, but it expanded to include communication, reporting, incident management, community engagement, privacy protection, and comprehensive safety solutions.

Future app development should ensure cross-platform availability, including iOS, is crucial to reach a wider audience and maximize impact. Incorporating offline functionality will improve accessibility in areas with limited internet connectivity. Strengthening public sector collaboration can enhance the integration of these technologies into broader social and policy frameworks, making them more accessible. Apps should be designed inclusively, with strategies to provide free or subsidized access for users with financial constraints. In addition, integrating advanced technologies such as AI for automated threat detection and incorporating innovations like smart self-defense mechanisms will enhance the utility and responsiveness of these apps. Furthermore, a further systematic review of the literature is essential to fully understand the implementation outcomes of these apps.

XSL•FO
RenderX

Acknowledgments

This study was funded by the Department of Foreign Affairs, Trade and Development (DFATD), Canada through Advancing Sexual and Reproduction Health and Rights (AdSEARCH), Grant number: SGDE-EDRMS-#9926532, Purchase Order 7428855, Project P007358. Besides, we express our gratitude to the app developers and institutions whose dedicated efforts have contributed to creating applications aimed at addressing violence against women and girls. Additionally, this research study was conducted by icddr,b, receives funding from core donors including the Governments of Bangladesh and Canada who provide unrestricted support to icddr,b for its operations and research. We gratefully acknowledge our core donors for their support and commitment to icddr,b's research efforts.

All authors declared that they had insufficient funding to support open access publication of this manuscript, including from affiliated organizations or institutions, funding agencies, or other organizations. JMIR Publications provided article processing fee (APF) support for the publication of this article.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

MHR conceptualized and designed the study. NGU conducted the database search. NGU and PC screened apps for eligibility. NGU and PC extracted the data from included articles. MHR, NGU, PC, and RMM conducted the analysis. MHR drafted the manuscript. AER, ATH, MSS, AA, and SEA reviewed the manuscript and provided feedback. MHR conducted the final edit of the manuscript. AER and ATH played the role of senior authors, overseeing conduct of the study, the analysis and manuscript drafting process.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary tables and figures.
[DOCX File, 104 KB - formative_v9i1e66247_app1.docx]

References

1. World Health Organisation. Violence against women. 2024. URL: https://www.who.int/en/news-room/fact-sheets/detail/violence-against-women [accessed 2024-06-02]
2. Research, Trend Analysis Branch UNOOD, Crime. Killings of women and girls by their intimate partner or other family members: global estimates 2020. United Nations Office on Drugs and Crime. 2021.
3. Chrisler JC, Ferguson S. Violence against women as a public health issue. Ann N Y Acad Sci 2006 Nov;1087(1):235-249. [doi: 10.1196/annals.1385.009] [Medline: 17189508]
4. UN Women. Innovation and prevention of violence against women: practice brief. United Nation Women. 2024.
5. Storer HL, Nyerges EX, Hamby S. Technology "Feels Less Threatening": The processes by which digital technologies facilitate youths' access to services at intimate partner violence organizations. Child Youth Serv Rev 2022 Aug;139:106573. [doi: 10.1016/j.childyouth.2022.106573]
6. Maxwell L, Sanders A, Skues J, et al. A content analysis of personal safety apps: are they keeping us safe or making us more vulnerable? Violence Against Women 2020 Feb;26(2):233-248. [doi: 10.1177/1077801219832124] [Medline: 30854941]
7. White D, McMillan L. Innovating the problem away? A critical study of anti-rape technologies. Violence Against Women 2020 Aug;26(10):1120-1140. [doi: 10.1177/1077801219856115] [Medline: 31327309]
8. Eisenhut K, Sauerborn E, García-Moreno C, et al. Mobile applications addressing violence against women: a systematic review. BMJ Glob Health 2020;5(4):e001954. [doi: 10.1136/bmjgh-2019-001954] [Medline: 32399255]
9. Draughon Moret J, Todd A, Rose L, et al. Mobile phone apps for intimate partner and sexual violence prevention and response: systematic search on app stores. JMIR Form Res 2022 Feb 8;6(2):e28959. [doi: 10.2196/28959] [Medline: 35133285]
10. Ford K, Bellis MA, Judd N, et al. The use of mobile phone applications to enhance personal safety from interpersonal violence - an overview of available smartphone applications in the United Kingdom. BMC Public Health 2022 Jun 9;22(1):1158. [doi: 10.1186/s12889-022-13551-9] [Medline: 35681167]
11. Cumiskey KM, Brewster K. Mobile phones or pepper spray? Fem Media Stud 2012 Dec;12(4):590-599. [doi: 10.1080/14680777.2012.741893]

XSL•FO
RenderX

12. Sumra M, Asghar S, Khan KS, et al. Smartphone apps for domestic violence prevention: a systematic review. Int J Environ Res Public Health 2023 Mar 23;20(7):5246. [doi: 10.3390/ijerph20075246] [Medline: 37047862]
13. Google play statistics and trends 2025. 42matters. 2024. URL: https://42matters.com/google-play-statistics-and-trends [accessed 2025-05-22]
14. Sánchez OR, Vale DB, Rodrigues L, et al. Violence against women during the COVID-19 pandemic: an integrative review. Int J Gynaecol Obstet 2020 Nov;151(2):180-187. [doi: 10.1002/ijgo.13365] [Medline: 32880941]
15. Doria N, Ausman C, Wilson S, et al. Women's experiences of safety apps for sexualized violence: a narrative scoping review. BMC Public Health 2021 Dec 30;21(1):2330. [doi: 10.1186/s12889-021-12292-5] [Medline: 34969403]
16. Jewkes R, Dartnall E. More research is needed on digital technologies in violence against women. Lancet Public Health 2019 Jun;4(6):e270-e271. [doi: 10.1016/S2468-2667(19)30076-3] [Medline: 31155215]
17. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
18. WB regions and countries. World Bank Group. 2022 Sep 8. URL: https://thedocs.worldbank.org/en/doc/b85c667c7fa4cd060918a2ed3d01aab8-0360012022/wb-regions-and-countries [accessed 2025-05-04]
19. R Core Team R. R: A language and environment for statistical computing. : R foundation for statistical computing Vienna; 2013.
20. Wickham H. ggplot2. WIREs Computational Stats 2011 Mar;3(2):180-185. [doi: 10.1002/wics.147]
21. Muggeo VM. Segmented: an R package to fit regression models with broken-line relationships. R news 2008;8(1):20-25.
22. Csardi MG. Package 'igraph. . 2013(9).
23. Wickham H, Averick M, Bryan J, et al. Welcome to the Tidyverse. J Open Source Softw 2019;4(43):1686. [doi: 10.21105/joss.01686]
24. Gajdoš P, Ježowicz T, Uher V, et al. A parallel Fruchterman–Reingold algorithm optimized for fast visualization of large graphs and swarms of data. Swarm Evol Comput 2016 Feb;26:56-63. [doi: 10.1016/j.swevo.2015.07.006]
25. Leading countries based on mobile app releases in 2017. Statista. 2023 Aug 25. URL: https://www.statista.com/statistics/934739/countries-most-app-releases [accessed 2025-05-04]
26. Emezue C. Digital or digitally delivered responses to domestic and intimate partner violence during COVID-19. JMIR Public Health Surveill 2020 Jul 30;6(3):e19831. [doi: 10.2196/19831] [Medline: 32678797]
27. Alhusen J, Bloom T, Clough A, et al. Development of the MyPlan Safety Decision App with friends of college women in abusive dating relationships. J Technol Hum Serv 2015 Jul 3;33(3):263-282. [doi: 10.1080/15228835.2015.1037414]
28. Hegarty K, Tarzia L, Valpied J, et al. An online healthy relationship tool and safety decision aid for women experiencing intimate partner violence (I-DECIDE): a randomised controlled trial. Lancet Public Health 2019 Jun;4(6):e301-e310. [doi: 10.1016/S2468-2667(19)30079-9] [Medline: 31155223]
29. Koziol-McLain J, Vandal AC, Wilson D, et al. Efficacy of a web-based safety decision aid for women experiencing intimate partner violence: randomized controlled trial. J Med Internet Res 2018 Jan 10;19(12):e426. [doi: 10.2196/jmir.8617] [Medline: 29321125]
30. Mergel I, Edelmann N, Haug N. Defining digital transformation: results from expert interviews. Gov Inf Q 2019 Oct;36(4):101385. [doi: 10.1016/j.giq.2019.06.002]
31. Waters HH, Rajkotia A, Basu Y, et al. The Economic Dimensions of Interpersonal Violence: World Health Organization; 2004.
32. Ranganadh A. Women safety device with GPS tracking and alerts. In: Innovations in Electrical and Electronics Engineering: Proceedings of the 4th ICIEEE 2019: Springer; 2020:797-805 URL: https://link.springer.com/chapter/10.1007/978-981-15-2256-7_74 [accessed 2025-05-04]
33. Banerjee S, Maiti P, Biswas S. Empowering personal safety through mobile apps. In: AI Tools and Applications for Women's Safety: IGI Global; 2024:91-120. [doi: 10.4018/979-8-3693-1435-7.ch006]
34. Malik AS, Acharya S, Humane S. Exploring the impact of security technologies on mental health: a comprehensive review. Cureus 2024 Feb;16(2):e53664. [doi: 10.7759/cureus.53664] [Medline: 38455776]
35. Debnam KJ, Kumodzi T. Adolescent perceptions of an interactive mobile application to respond to teen dating violence. J Interpers Violence 2021 Jul;36(13-14):6821-6837. [doi: 10.1177/0886260518821455] [Medline: 30600761]
36. Emezue C, Chase JAD, Udmuangpia T, et al. Technology-based and digital interventions for intimate partner violence: a systematic review and meta-analysis. Campbell Syst Rev 2022 Sep;18(3):e1271. [doi: 10.1002/cl2.1271] [Medline: 36909881]
37. Ten ways to prevent violence against women and girls. Women UN. 2023. URL: https://www.unwomen.org/en/news-stories/explainer/2023/11/ten-ways-to-prevent-violence-against-women-and-girls [accessed 2025-05-04]
38. Macau UWaUUIi. Artificial Intelligence and the Women, Peace and Security Agenda in South-East Asia. 2024. URL: https://asiapacific.unwomen.org/en/digital-library/publications/2024/05/artificial-intelligence-and-the-women-peace-and-security-agenda [accessed 2025-05-22]
39. Jewani VK, Ajmire PE, Chaurasia S, et al. Artificial intelligence: A smart and empowering approach to women's safety. In: Impact of AI on Advancing Women's Safety: IGI Global: IGI Global; 2024:121-138. [doi: 10.4018/979-8-3693-2679-4.ch008]

XSL•FO
RenderX

Abbreviations

AI: artificial intelligence
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROSPERO : International Prospective Register of Systematic Reviews
SOS: Save our souls
VAWG: violence against women and girls

Edited by A Mavragani; submitted 11.09.24; peer-reviewed by D Li, MRU Zaman Sajib, M Zhang, VP Jover; revised version received 19.02.25; accepted 02.03.25; published 23.06.25.

Please cite as:
Rahman MH, Usmani NG, Chandra P, Manna RM, Ahmed A, Shomik MS, Arifeen SE, Hossain AT, Rahman AE
Mobile Apps to Prevent Violence Against Women and Girls (VAWG): Systematic App Research and Content Analysis
JMIR Form Res 2025;9:e66247
URL: https://formative.jmir.org/2025/1/e66247
doi:10.2196/66247

© Md Hafizur Rahman, Nasimul Ghani Usmani, Pradip Chandra, Ridwana Maher Manna, Anisuddin Ahmed, Mohammad Sohel Shomik, Shams El Arifeen, Aniqa Tasnim Hossain, Ahmed Ehsanur Rahman. Originally published in JMIR Formative Research (https://formative.jmir.org), 23.6.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Cocreating a Mobile Health App Providing Physical Activity Recommendations for Older People Living With Parkinson Disease or Dementia: User-Centered Pilot Study

Ellen Bentlage[1*], PhD; Alberto del Río Ponce[2*], PhD; Mona Ahmed[1], PhD; Pilar Gangas[3,4], PhD; Michael Brach[1], Prof Dr; Jorge Alfonso Kurano[2], MSc; José Manuel Menéndez[2], Prof Dr

[1]Insitute of Sport and Exercise Sciences, University of Münster, Horstmarer Landweg 50, Münster, Germany
[2]Grupo de Aplicación de Telecomunicaciones Visuales, Universidad Politécnica de Madrid, Madrid, Spain
[3]Annexe Offices, International Foundation for Integrated Care, Oxford, United Kingdom
[4]CSIC-UAC Técnico de Asesoramiento Cientifico, CSIC-UAC, Madrid, Spain
[*]these authors contributed equally

Corresponding Author:
Ellen Bentlage, PhD
Insitute of Sport and Exercise Sciences, University of Münster, Horstmarer Landweg 50, Münster, Germany

Abstract

Background: The project "Personalized Integrated Care Promoting Quality of Life for Older People" aimed to develop an integrated care system based on information and communication technology to support older people living with Parkinson disease or dementia disease. One module focuses on physical activity (PA) recommendations.

Objective: The objective of the study is to describe the development process of the PA recommendation system from the behavior-change and technical perspective, followed by its content and satisfaction evaluation.

Methods: This study describes the development of the PA recommendations based on the Health Action Process Approach (HAPA). A first pilot assessed the feasibility of the overall PROCare4Life system (previously reported). In a second pilot, users evaluated the content of the PA recommendations during 40 intervention days. In a third pilot, users evaluated their satisfaction with a mobile health satisfaction questionnaire.

Results: The PA recommendations focused on different aspects of an adapted version of the HAPA model, while they simultaneously approached 3 activation factors: skills, knowledge, and motivation. The content was generally well-received, with most users rating key sections as excellent or good, particularly "benefits and consequences of PA" (34/43, 79%) and "five golden rules of PA" (34/41, 83%). However, less than a third gave high ratings to "PA guidelines of the WHO" (9/36, 25%) and "practical tips for PA" (10/35, 29%). Regarding satisfaction, at least half of the 237 participants found it easy and good to use, with acceptable time spent and clear instructions. Compared with agreement or neutral evaluations, most disagreed with negative statements about it being time-consuming (111/237, 47%) or boring (99/237, 42%). While 41% (97/237) recommended it and 44% (104/237) felt it helped them understand lifestyle benefits, fewer agreed the recommender system helped them set personal goals (78/237, 33%) or motivated change (88/237, 37%). Users found the recommendations understandable, engaging, and practical, though some aspects, such as motivation and goal setting, received criticism. Challenges in pilot 2, particularly related to setup difficulties and limited participation, led to system modifications in pilot 3 that improved usability and data collection.

Conclusions: It has been confirmed that cocreating and iteratively testing the contents on HAPA and approaching the activation factors contributed to increasing the acceptance of the PA. The intervention development was based on user needs and used comparable methodology across user profiles and pilot phases. All in all, users were positive about the content. The research team has identified that digital systems, that provide monitoring functions of the mobile health app and Fitbit wristband are considered advantageous by participants in the cocreation process. Addressing the activation factors can be recommended for researchers and technical developers of other projects. Future adjustments to the design should focus on personalization to encourage the adoption of a healthier lifestyle.

(JMIR Form Res 2025;9:e51831) doi:10.2196/51831

KEYWORDS

healthcare; physical activity; personalized recommendations; ICT; quality of life; codesign process

XSL•FO
RenderX

Introduction

Active and Inactive Lifestyles

The World Health Organization has developed guidelines for different age groups on how much physical activity (PA) is needed for good health [1], whereby adults should get 150 minutes of moderate PA or 75 minutes of vigorous PA per week. Yet, on a validated PA questionnaire answered by over 2000 older adults in the European Union (mean age of 74.9, SD 4.5 years), 42% were found to not meet these global PA recommendations. These findings mirror others indicating that most older adults have a physically inactive and sedentary lifestyle [2-4]. Further, older participants tend to meet the recommendations less often [5]. Notably, people's amount of PA decreased dramatically during the COVID-19 pandemic, when they faced isolation and social distancing. Specifically, home confinement reduced people's total weekly PA time and intensity and increased daily sitting time in a health-threatening direction [6], as these factors increase one's risk of mortality, obesity, sarcopenia, frailty, falls, functional impairments, and neurodegenerative diseases [7,8]. By contrast, regular PA results in a lower risk of falls and mortality as well as an improved overall health status, including better Quality of Life (QoL) [7,9]. PA is known to reduce one's risk of dementia disease (DD) [10], and in Parkinson disease (PD) it is "a vital component to maintain balance, mobility and activities of daily living" [11]. In a web-based survey from May to August 2020, 75% (64/85) of participants with these conditions reported that receiving PA recommendations was desired, and 79% (57/72) of caregivers reported the same. Interviews with patients and workshops with health care professionals support these results [12].

Behavior Change Techniques

Strategies for changing people's behavior to become healthier fall under the general umbrella of behavior change techniques, which might often rely on 3 activation factors: skills, knowledge, and motivation. Research shows that previous interventions that addressed these factors could positively influence individuals' PA behavior [13,14]. Notably, these factors can be addressed by different behavior change techniques; for example, skills can be trained with instructions or demonstrations of a specific exercise [15], and knowledge can be addressed by educating individuals on the benefits of sufficient PA and the consequences of insufficient PA and by comparing their own PA behavior with guidelines [16], and motivation can be targeted by activity monitoring, either by the user, another person or a technical device [16], and through personalized feedback [17]. All of these were incorporated into PROCare4Life mobile app and iteratively tested and cocreated with older people living with PD or dementia. A recent review showed that integrating the abovementioned activation factors results in the most benefits when aiming at influencing PA behavior. Further, addressing the 3 activation factors can be accomplished by using a variety of interfaces [18] and digital interactions, that have been incorporated into PROCare4Life mobile app. Thus, it can be expected according to previous literature that digital health systems incorporating PA recommendations can potentially benefit from incorporating the above-mentioned activation factors and using various modalities to do so. These expectations were tested through the cocreation process of the PROCare4Life mobile app, which includes PA recommendations. The PROCare4Life PA recommendations approached skills and knowledge via notifications and PDFs that were displayed on the users' smartphones, whereas the motivation was approached by personalized gamification of the smartwatch app, which could also be read through the mobile app. In addition, the PA recommendations were developed according to the Health Action Process Approach (HAPA). HAPA was used to guide the development of PROCare4Life PA training and adherence programs from a psychological theory perspective to enhance the efficiency of PA behavioral changes among older people living with PD or dementia (Alzheimer disease and others) [19].

PROCare4Life and Health Recommender Systems

In the European Union, the recent project "Personalized Integrated Care Promoting Quality of Life for Older People" (PROCare4Life) funded by Horizon 2020 (grant agreement number 875221, 2020 - 2023) developed an interactive care system based on digital technologies for people living with neurodegenerative diseases, their caregivers (mobile health app, smartwatch for both profiles), and health care professionals (web interface) [20]. Health recommender systems [19] have become valuable digital tools [21-23], as they provide personalized suggestions and predictions to users by analyzing vast amounts of data and using advanced algorithms [24]. Health recommender systems have shown significant potential to improve patients' and caregivers' healthy behaviors and QoL [25-27]. The cocreation and iterative testing supported the development of the digital interactive care system through several pilot waves. PROCare4Life pilots were implemented at 6 real-life sites including daycare, home, and rehabilitation scenarios, located in 5 different EU (European Union) countries (Germany, Italy, Portugal, Romania, and Spain). The cocreation methodology included several steps:

1. A user requirements study [12,28] in which PROCare4Life users shared their needs linked to receiving personalized PA recommendations.
2. Development of PROCare4Life digital system, including feasibility testing of the digital technologies.
3. Development and content evaluation of the PA recommendations, both by experts, researchers, and representatives of the target future users.
4. Iterative cycles of adaption of the different modules of PROCare4Life. Satisfaction evaluation of the PA recommendations for users [29,30].

The PA recommendations were included in the iterative testing from the start of pilot 2, when the technology was mature enough for testing. They were embedded in the mobile app to support users in creating and maintaining an active lifestyle. The focus of this article is to describe PROCare4Life cocreation and iterative testing processes from the PA recommendations [31,32].

Research Aim

This article focuses on the cocreation process to support the development and iterative testing of the PROCare4Life PA

XSL•FO
RenderX

recommendations and its key results. Based on the specific needs of older people living with PD or dementia, and after assessing iteratively the users' recommendations regarding the PA contents, changes were incorporated into the digital system to improve the personalization of its end users. Overall user satisfaction assessment of the PA recommendations was also included as one of the research aims. The overall aim of PROCare4Life was to increase the QoL of people living with DD or PD, their self-management, and empowerment.

Methods

In this section, the development of the PA recommendations, the functionalities of the PA recommendation system, and methods for pilot 2 and pilot 3 are described.

Development

The HAPA model was used to develop the PA recommendations of the PROCare4Life digital interactive system. It includes an initial motivational phase for goal setting and a subsequent volitional phase focusing on how to implement behavioral changes to pursue personalized goals. Between these 2 phases, individuals using the PROCare4Life system can create an intention to be active [31]. As older people living with PD or dementia, PROCare4Life users needed to have basic knowledge about the PA recommendations, including the adequate volume and intensity levels of activity for each person. Accordingly, the HAPA model was amended to include the construct Awareness of Standards, which was not originally present (Figure 1).

Four HAPA constructs of the motivational phase and 2 constructs of the volitional phase were aligned with 6 performance objectives resulting in the PA recommendation contents with 3 underlying activation factors, as summarized in Table 1 below. The contents were fused into 6 PA recommendation sets. The process to develop the PA recommendations contents is displayed in the table below (Textbox 1).

The goals of the PROCare4Life PA were reinforced by the technical architecture of the digital system, which is described in the following subsection.

Figure 1. Visualization of the Health Action Process Approach.

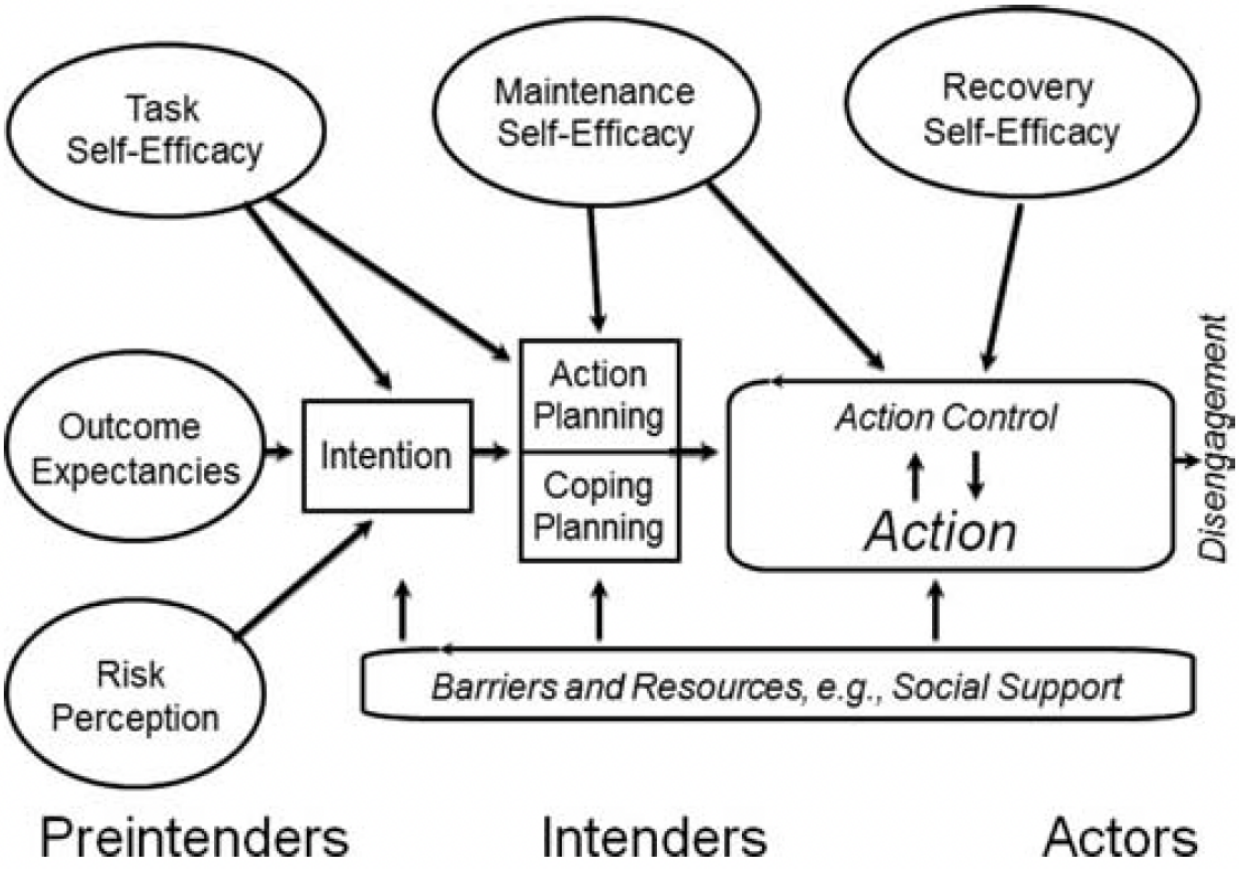

XSL•FO
RenderX

Table . Development of the physical activity (PA) recommendation contents.

HAPA[a] constructs (Activation factor)	Performance objectives	PA recommendation contents	Final PA recommendation sets
Action Self-Efficacy (Skills)	Awareness of ways to be active	Practical tips to be active	Set 1: Benefits of regular and consequences of insufficient physical activity
Outcome Expectancies (Knowledge)	Awareness of benefits of regular physical activity	Benefits of regular physical activity	Set 2: Five golden rules of physical activity
Risk Perception (Knowledge)	Awareness of consequences of insufficient physical activity	Consequences of insufficient physical activity	Set 3: Perceived exertion scale
Awareness of Standards (Knowledge)	Awareness of the physical activity guidelines	Five golden rules of physical activity, Rate of Perceived Exertion Scale, Physical activity guidelines from the World Health Organization	Set 4: Physical activity guidelines of the WHO[b]
Coping Self-Efficacy (Skills)	Experience opportunities to be active	Links for more information and videos about physical activity	Set 5: Practical tips for physical activity
Action Control (Motivation)	To be aware of your own physical activity level	Monitoring of steps via the Fitbit wristband	Set 6: Links for more information and videos about physical activity

[a]HAPA: health action process approach.
[b]WHO: World Health Organization.

Textbox 1. Examples for the different messages of each physical activity (PA) recommendation sets.

Set 1: Benefits of regular and consequences of insufficient PA activity

- Regular physical activity improves your sleeping patterns.
- Insufficient PA increases your risk for joint and back pain.

Set 2: Five golden rules of physical activity

- Reduce your sitting time with regular standing breaks every 30 minutes.
- Every minute of PA counts.

Set 3: Perceived exertion scale

- The Rate of Perceived Exertion Scale is a measure of subjective perceived exertion during PA with different ranges.
- Varies from very light activity (1), light activity (2-3), moderate activity (4-5) to vigorous activity (5-6).

Set 4: Physical activity guidelines of the World Health Organization

- Perform 3 activity-forms per week: Aerobic, muscle strength, and balance training.
- Conduct aerobic exercises, for example walking, 150 - 300 minutes per week.

Set 5: Practical tips for physical activity

- Walking in the fresh air can be a mood booster, it improves your mind and body.
- Conducting yard work, like pulling out weed, will make your arm muscles stronger.
- Practicing this regularly, carrying groceries can be easier.

Set 6: Links for more information and videos about physical activity

- Try this link out for further information on PA: Exercises for people with dementia disease.
- Try this link out for further information on PA: Exercises for people with Parkinson disease.

Technical Architecture

In terms of information collection, processing, and delivery, the PROCare4Life PA technical architecture consists of 2 types of components that transfer the PA recommendations as notifications to users: the low-level subsystem and the high-level subsystem.

The low-level subsystem encompasses the set of devices, sensors, and questionnaires, both active and passive, functions to acquire data from patients. It requires a smartwatch, a

XSL•FO
RenderX

smartphone with the included mobile health app, and questionnaires. Collected data are sent to the other subsystem. The high-level subsystem, includes the recommender engine that extracts the necessary information to generate recommendations according to each user's profile. A notification bus sends and receives notifications involving the exchange of information and messages between the user-facing part and the systems logic based on the RabbitMQ [32] message broker. The recommender system operates on an automated daily routine, sending predefined notifications to patients, independent of their previous data. While this heuristic approach may diverge from traditional recommendation methods, it ensures consistent daily monitoring [33]. At the end of each cycle, a knowledge-based approach was used to deliver customized routines and exercises tailored to each patient's primary disease group. Identity management provides a framework for creating, modifying, and deleting users' data. In the Web interface, health care professionals can check historical notifications from patients and examine different patients' information. The purpose of the clerical and clinical data repository is to store the different modalities of data that are related to users, including clinical and personal information (encrypted to comply with General Data Protection Regulation as well as some preprocessed sensorial summarized data that feeds the different dashboards and reports presented in the PROCare4Life user interface (Figure 2).

Figure 2. PROCare4Life (Personalized Integrated Care Promoting Quality of Life for Older People) physical activity (PA) technical architecture.

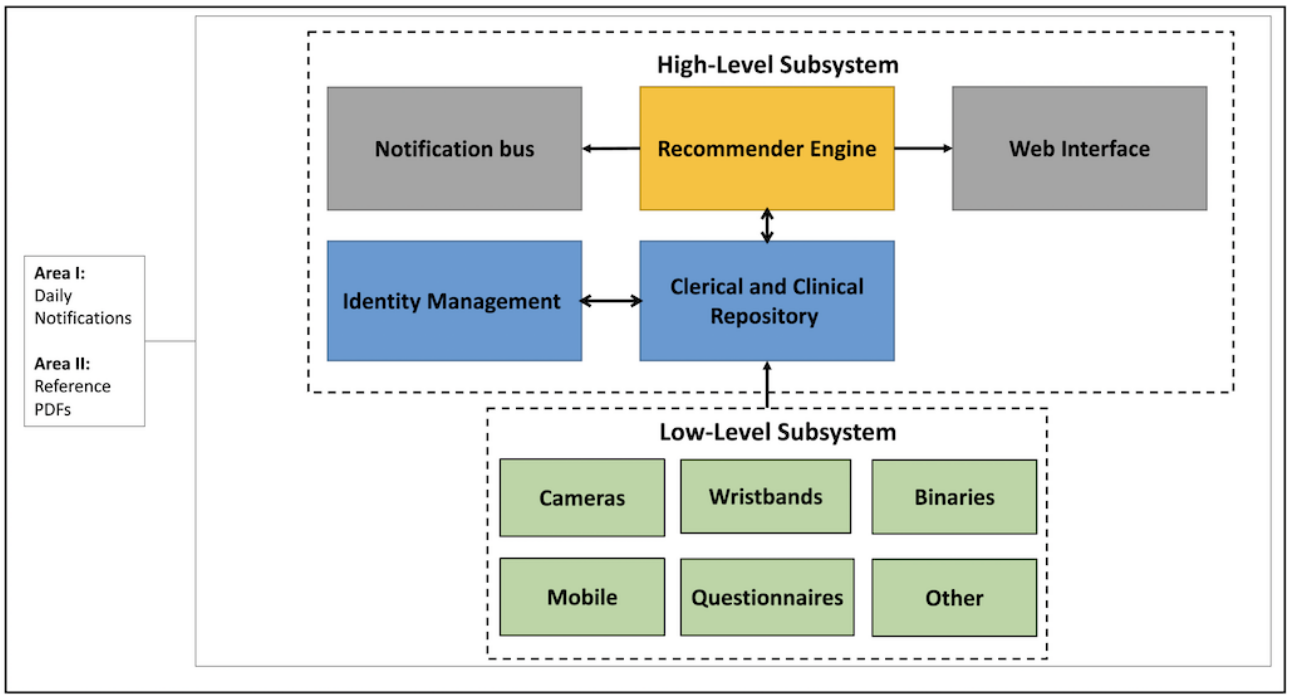

Ethical Considerations

All pilot phases were ethically approved by the local ethical commissions of the 6 pilot sites of PROCare4Life:

- Wohlfahrtswerk für Baden-Württemberg: Approved by the Ethical commission of the University of Münster (2021-15-MB-FA2).
- Asociación Parkinson Madrid: Approved by the Ethical commission of Hospital Clínico San Carlos (21/220-E).
- Casa di Cura Policlinico: Approved by Comitato Etico Milano Area 2 of the Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico (OSMAMI-26/07/2021-0032326-U).
- Campus Neurologico Senior: Approved by the Comissao de Ética Campus Neurologico Senior (Ref number 3-2021).
- Spitalul Universitar de Urgenta Bucuresti and University of Medicine and Pharmacy: Approved by the Comisia de Etica e Cercatarii of the Spitalul Clinic Colentina (Number 24/28.09.2021).

As the research included human participants, the end user pilot sites prepared, within the consortium, a first draft version of the ethical documentation. All relevant information on the ethics of the study was collected. Based on the local regulations of each end user pilot site, the gathered information was integrated into the local documents, including ethical documentation for the ethics commission including the following sections: description and procedures of the study, voluntariness, and anonymity, information about no compensation type for the participation, insurance coverage, scope of data collection and processing, legal basis, option for the withdrawal, contact data of the responsible contact persons. This information was fused into the ethics application and the connected general participant information and consent. In addition, a data handling agreement was established and signed by all parties.

Participants

PROCare4Life second and third pilots were performed in accordance with the following inclusion and exclusion criteria.

The inclusion criteria included individuals who had a clinical diagnosis of DD or PD; were 65 years or older; demonstrated a willingness to participate; had the ability to provide informed consent; and were currently living at home, in a rehabilitation center, or in a daycare center.

XSL•FO
RenderX

By contrast, candidates were excluded from participating in the study in cases of inability to learn and limitations impeding the ability to use the PROCare4Life system or key elements of its technical equipment, such as speech, vision, language, hearing, or psychiatric impairments.

Participants were recruited from the social and health care centers involved in the study (Casa di Cura Policlinico, Asociación Parkinson Madrid, Campus Neurologico Senior, University of Medicine and Pharmacy, Spitalul Universitar de Urgenta Bucuresti, and Wohlfahrtswerk für Baden-Württemberg).

Study Procedure

The study was conducted in 3 different scenarios: neurorehabilitation center, daycare center, and at participants' homes. People living with DD or PD received periodically personalized PA recommendations (Multimedia Appendix 1) over the 40 days that their participation lasted, via their PROCare4Life mobile app. Each day a different notification was sent. All these materials were additionally stored as PDFs (Multimedia Appendix 2) in the app allowing participants to consult them later. Table 2 summarises the PA recommendation sets by the days of their appearance, gradually covering all 6 sets.

Table . PROCare4Life[a] physical activity (PA) recommendation sets and their appearance in the mobile health app.

Sets	Topic	Days
1	Benefits of regular and consequences of insufficient PA	1 - 10
2	Five golden rules of PA	11 - 15
3	Perceived exertion scale	16 - 25
4	PA guidelines of the WHO[b]	26 - 30
5	Practical tips for PA	31 - 35
6	Links for more information and videos about PA	36 - 40

[a]PROCare4Life: Personalized Integrated Care Promoting Quality of Life for Older People.

[b]WHO: World Health Organization.

Pilot 2

The PA recommendations were evaluated by users through intermediate evaluation questions. They appeared in the PROCare4Life mobile app itself once each set of recommendations had been completed. Additionally, the system integrated reminders that prompted patients to respond to some questions linked to each set of notifications. Users were requested to rate the information included in each PA set using a 5-point Likert scale. They were requested to select one of 5 intuitive emojis, referring to "excellent," "good," "average" rated as positive, "poor," and "very poor" rated as negative.

Pilot 3

Over the pilot 3 implementation, participants were also invited to assess the PA recommendation system and contents after the 40-day participation, also via the PROCare4Life app-integrated mobile health satisfaction questionnaire [34]. On 12 items, they were able to select the options "agree," "neutral," or "disagree" (Table 3).

XSL•FO
RenderX

Table . Mobile health satisfaction questionnaire displayed in the PROCare4Life[a] app.

Question and response number	Response
(a) What did you think about using the recommendation system in our PROCare4Life mobile app?	
Q1	It was easy to use
Q2	It was good to use
Q3	The time spent using it was acceptable
Q4	The explanation on how to use it was sufficient
Q5	It was too time-consuming
Q6	It was boring to use
Q7	It was a disturbance
Q8	I can recommend it to others
(b) How did you experience the recommendation system in your PROCare4Life mobile app?	
Q9	It has motivated me to change my lifestyle habits
Q10	It has helped me to understand the benefits of improving my lifestyle habits
Q11	It has helped me to understand how I need to change my lifestyle habits
Q12	It has helped me set personal goals for my lifestyle habits in a way that I could not have done on my own

[a]PROCare4Life: Personalized Integrated Care Promoting Quality of Life for Older People.

Statistical Analysis

The following sociodemographic data were collected for pilot 2 and pilot 3 to support the data analysis while guaranteeing the pseudoanonymization of the data: patient ID; scenarios, birth date; gender; diagnosis, and pilot site. The analysis was performed using Microsoft Excel (version 2.0). Descriptive statistical analysis was performed, identifying frequencies and percentages of responses. The evaluation results were crossed with the sociodemographic variables for comparison across diagnosis, age, pilot site, gender, and scenarios.

Results

Pilot 2: Content Evaluation

Participants' Characteristics

A total of 167 participants were recruited, of which 55 participants were in the home scenario, 94 participants were in the rehabilitation center, and 18 participants were in the daycare center. Out of the total number of participants, 43 (26%) participants evaluated the content of the PA recommendations. Most of them (27/43, 63%) were men. The mean age was 75 (SD 5.9) years. Patients came from Spain (29/43, 67%), Italy (7/43, 16%), and Portugal (7/43, 16%). Regarding living situations, 79% (34/43) were in a rehabilitation center and 19% (7/43) were at home. Continued problems with the recruitment of patients for the daycare centers prevented the collection of significant enough information in that scenario. The most common diagnosis was PD (37/43, 86%).

Results Related to the Content Evaluation

The results from the intermediate evaluation questions are presented in Table 6. The first PA recommendation set evaluations had a 100% completion rate by participants. As time went by and additional sets were delivered, the number of missing evaluations increased from zero to a maximum of 10 people. Therefore, the total number of evaluations, representing 100%, differs from set to set. The first sets were mostly rated as excellent or good: "Benefits and consequences of PA" (34/43, 79.1%), "five golden rules of PA" (34/41, 83%), and "perceived exertion scale" (22/41, 54%). The same appeared by around one-quarter of users for "PA guidelines of the WHO" (9/36, 25%), "practical tips for PA" (10/35, 29%), and "content of links" (14/33, 42%). All in all, it can be clearly stated that all sets were rated by most users as positive (minimum: average). Very few people (n=5) selected the option "poor" or "very poor," in total among all (Table 4).

XSL•FO
RenderX

Table . Results of the intermediate evaluation questions (IEQ).

	IEQ1 PAR[a] set 1: Benefits and consequences of PA[b]	IEQ2 PAR set 2: Five golden rules of PA, n (%)	IEQ3 PAR set 3: Perceived exertion scale, n (%)	IEQ4 PAR set 4: PA guidelines from the WHO, n (%)	IEQ5 PAR set 5: Practical tips for PA, n (%)	IEQ6 PAR set 6: Content of the links, n (%)
Excellent	16 (37)	16 (39)	4 (10)	0 (0)	3 (9)	2 (6)
Good	18 (42)	18 (44)	18 (44)	9 (25)	7 (20)	12 (36)
Average	7 (16)	7 (17)	18 (44)	27 (75)	25 (71)	17 (52)
Poor	0 (0)	0 (0)	1 (2)	0 (0)	0 (0)	2 (6)
Very poor	2 (5)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Not reported	N/A[c] (0)	N/A (2)	N/A (2)	N/A (7)	N/A (8)	N/A (10)
Total	43 (100)	41 (100)	41 (100	36 (100)	35 (100)	33 (100)

[a]PAR: physical activity recommendation.

[b]PA: physical activity.

[c]The data are marked as "not available" because it is unclear whether participants responded to the questions. Due to technical errors, we were unable to determine which responses were provided. As more sets were delivered, the number of missing evaluations increased—from zero initially to a maximum of 10 participants. Consequently, the total number of evaluations, representing 100%, varies across different sets.

Pilot 3: Satisfaction Evaluation

Participants' Characteristics

A total of 273 patients were recruited of which 132 patients were in the home setting, 102 patients were in the rehabilitation center, and 39 patients were in the daycare center. After excluding dropouts (n=36), 237 of 273 (87%) patients provided answers to the mobile satisfaction assessment questionnaire. Patients had a mean age of 72 years. Most patients came from Italy (82/237, 35%), followed by Spain (64/237, 27%), Romania (56/237, 24%), Germany (22/237, 9%), and Portugal (13/237, 5%). Regarding living situations, 120 (52%) people were located at home, 90 (38%) people were in the rehabilitation center, 22 (9%) people were in the daycare center, and 5 (2%) people did not report.

Results Related to the Satisfaction Evaluation

Results were positive in terms of satisfaction for many of the questions linked to the PROCare4Life PA recommendations, although some were not so positive. At least half of the 237 participants agreed to the sentences "Q1-it was easy to use" (121/237, 51%), "Q2-it was good to use" (119/237, 50%), "Q3-time spent using it was acceptable" (123/237, 52%), and "Q4-explanation on how to use it was sufficient" (135/237, 57%). Compared with agreement and neutral evaluation, most people disagreed with negatively phrased worded statements, thus meaning a positive evaluation, namely "Q5-it was too time-consuming" (111/237, 47%), "Q6-it was boring to use" (99/237, 42%), and "Q7-it was a disturbance" (137/237, 58%). Most rated the following quotes positive: "Q8-I can recommend it to others" (97/237, 41%) and "Q10 it has helped me to understand the benefits of improving my lifestyle habits" (104/237, 44%). For the statement "Q12-it has helped me set personal goals for my lifestyle habits in a way that I could not have done on my own," more people agreed (78/237, 33%) as disagreed (71/237, 30%). Most disagreement was expressed for "Q9-it has motivated me to change my lifestyle habits" (88/237, 37%) and "Q11-it has helped me to understand how I need to change my lifestyle habits" (71/237, 30%; Figure 3).

XSL•FO
RenderX

Figure 3. Mobile health satisfaction questionnaire results of 237 participants. Q: question.

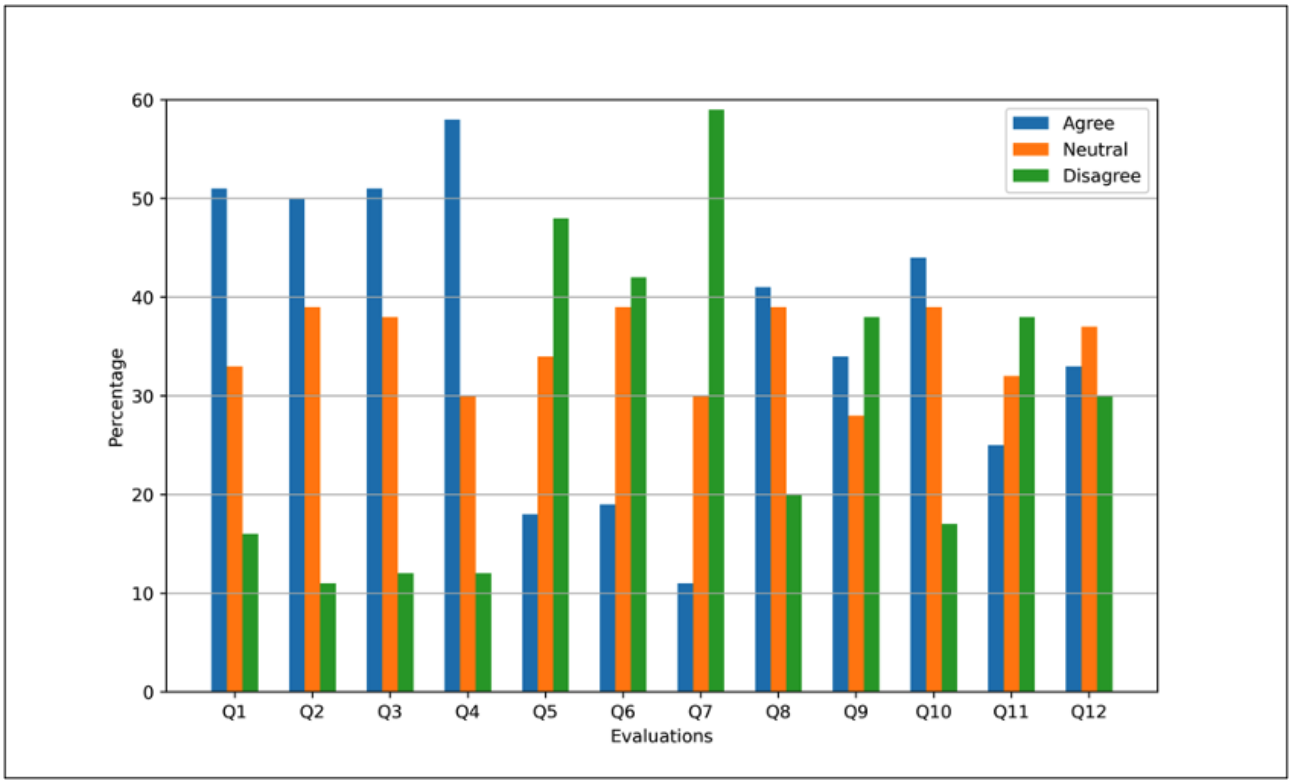

Discussion

Principal Findings

Overview

This report aimed to describe the development process of the PA recommendations according to the HAPA and activation factors, the content of the PA recommendations, and users' satisfaction with the PA recommendation system.

All in all, the results of section 1 (Q1-Q8) of the mobile health satisfaction questionnaire clearly demonstrate that PROCare4Life PA recommendations were found to be easily understandable, interesting, and usable, and users clearly understood the benefits of PA (Q10).

Developing the contents of the HAPA model and approaching all the activation factors in information and communication technology–based interventions supported the intervention development, as stated in previous research [18,35].

Knowledge

The 7 statements within set 1 focusing on the positive influence of regular PA and on the consequences of insufficient PA were delivered to patients' mobile phones. Often, older adults connect exercising directly with potential injuries and express their fear of falling [36]. After reading the first notifications, this incorrect statement can be resolved. Brawley and Latimer [37] stated that pragmatic explanations and supporting messages are necessary to facilitate regular PA. The 5 golden rules helped users to be regularly active, as adherence to the use of the app reflected. The rating of the perceived exertion scale of set 3 introduced users to the possibility of measuring their subjective perceived exertion during PA and to categorize the intensity as very light, light, moderate, or vigorous. Introducing users to the WHO's PA guidelines within set 4, aimed to make users aware of the PA standards. Each form of training was also matched with examples of activities to increase their knowledge.

Skills

Statements of set 5 introduced users to practical tips for being active without specific PA equipment. Further, statements indicating the positive influences of PA on one's daily life within set 1 were sent to the users to perform the PA activities. The final 5 notifications of set 6 gave users access to links for more information and videos about PA based on the diagnosis and further explanations in the local language.

Motivation

The Fitbit wristband helps users in monitoring their number of steps to keep their motivation on a high level.

The PROCare4Life system chose to incorporate a mobile health app and Fitbit because PA recommendations may be better followed and lead to less sedentary lifestyles when they incorporate technology, due to the possibility to personalize the health values [38].

Interpreting the results of Q9-Q12 of the mobile health satisfaction questionnaire shows that the recommendation system is critical on how to motivate to change lifestyle habits, to set personal goals, and to facilitate these. The interpretation of these results is not completely clear to the research team. One possible conclusion was that the app failed to achieve its goal of improving the PA among participants and in changing their lifestyle habits. Another alternative explanation is that the participants were already motivated and understood how to change their lifestyle habits, and thus the reasons behind the

XSL•FO
RenderX

negative assessment of these 2 questions might reflect an already positive starting point for the individuals sharing their disagreement. Further research might be needed in this respect. Features that could counter the criticism are more personalized features that address the remaining constructs of the HAPA that are currently missing: By addressing the construct Action Planning of the HAPA, users could have the option to select a personal goal of steps based upon their measured average number of weekly steps, differentiated into maintenance (average base), low ambition (average base +10%), medium ambition (average base +20%), and high ambition (average base +30%). Lee et al [39] roofed that technology-based goal setting helps people to advance the effectiveness of the number of steps.

By focusing on the construct Action Coping of the HAPA, the system could recommend specific "exergames" via smart TV to the user, focusing on their measured average number of weekly steps, and on days with unsuitable weather for outdoor activity. The findings of Rhodes et al [40] highlight the social aspects of exergaming. Exergames are capable of incorporating more than one user, to keep the motivation on a high level. Addressing the construct Recovery Self-Efficacy of the HAPA, the system could send users an underachievement notification (eg, "This week you did not achieve your personal step goal. You took 900 steps on average. Don't worry, you can try it again or change your personal goal for steps") [41]. Future interventions are encouraged to consider addressing the remains of the HAPA construct, facilitating more personalized PA interventions.

Limitations

The period in which pilot 2 of PROCare4Life was carried out was challenging due to COVID-19 and analysis of its outcomes highlighted some limitations of the system. Specifically, pilot 2 experienced installation and setup issues, including smartphone and the Fitbit. Difficulties with the initial setup prevented the patients from using the system at all, which also implied that PA recommendations were not used nor evaluated. Consequently, one-quarter of the participants were represented in the pilot 2 data.

To counteract changes performed in the solution deployment to reduce the complexity of the setup and improve the user experience. The main change was the elimination of the physical mini-PC, which was the data collection hub for pilot 1 and pilot 2, and at the same time, the main source of installation and operation issues. In pilot 3, most of the patients only had the wristband and the phone. Data collection figures increased significantly in pilot 3. Additional changes were made to the PA recommendation components of the PROCare4Life solution, enabling more data collection points and potentially a more interactive experience for the patients.

This study reported on the theoretical background of behavior change techniques and on their technical development. The results are based on multinational participants, across various settings and health conditions. The multidisciplinary team, which included technical partners (for developing the backend and front end), academic partners (for creating research-related content), an educational network (for producing relevant materials), a security company (for guaranteeing data protection and privacy requirements) as well as clinical pilot sites (for testing the end product), followed an iterative design process for developing the interfaces as recommended by Zaman et al [42].

Conclusions

The PA recommendations were implemented in a systematic way and addressed the activation factors like skills, knowledge, and motivation to focus on relevant behavior change constructs. Based on users' perceptions of the system, the content was rated positively, and users were satisfied with the PA recommendations. Addressing the activation factors can be recommended for researchers and technical developers of other projects. The current implementation of the mobile health app uses daily static notifications and performs limited monitoring of patients' activities. More features need to be implemented to personalize the recommender system to each individual user.

Acknowledgments

We would like to thank all partners of the PROCare4Life consortium, especially the clinical partners who invested a lot of time in patient recruitment and guiding patients through the data collection. We also thank David Linnane from the Integrated Care Foundation for the English language check. This project received funding from the European Union's Horizon 2020 research and innovation program (grant agreement 875221).

Data Availability

The datasets generated or analyzed during this study are available from the shared first authors on reasonable request.

Authors' Contributions

EB and AR were responsible for conceptualization, reference management, and project administration. EB, AR, MA, and JK contributed to the writing of the original draft. PG, MB, and JM provided editing support. Supervision was provided by MB and JA. EB handled the journal submission. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
Physical Activity Recommendations daily messages.
[DOCX File, 35 KB - formative_v9i1e51831_app1.docx]

Multimedia Appendix 2
Physical Activity Recommendations in PDF files.
[DOCX File, 1155 KB - formative_v9i1e51831_app2.docx]

References

1. Bull FC, Al-Ansari SS, Biddle S, et al. World Health Organization 2020 guidelines on physical activity and sedentary behaviour. Br J Sports Med 2020 Dec;54(24):1451-1462. [doi: 10.1136/bjsports-2020-102955] [Medline: 33239350]
2. Taylor D. Physical activity is medicine for older adults. Postgrad Med J 2014 Jan;90(1059):26-32. [doi: 10.1136/postgradmedj-2012-131366] [Medline: 24255119]
3. Davis MG, Fox KR. Physical activity patterns assessed by accelerometry in older people. Eur J Appl Physiol 2007 Jul;100(5):581-589. [doi: 10.1007/s00421-006-0320-8] [Medline: 17063361]
4. Harvey JA, Chastin SFM, Skelton DA. Prevalence of sedentary behavior in older adults: a systematic review. Int J Environ Res Public Health 2013 Dec 2;10(12):6645-6661. [doi: 10.3390/ijerph10126645] [Medline: 24317382]
5. Mattle M, Meyer U, Lang W, et al. Prevalence of physical activity and sedentary behavior patterns in generally healthy European adults aged 70 years and older-baseline results from the DO-HEALTH clinical trial. Front Public Health 2022;10:810725. [doi: 10.3389/fpubh.2022.810725] [Medline: 35493350]
6. Ammar A, Brach M, Trabelsi K, et al. Effects of COVID-19 home confinement on eating behaviour and physical activity: results of the ECLB-COVID19 international online survey. Nutrients 2020 May 28;12(6):1583. [doi: 10.3390/nu12061583] [Medline: 32481594]
7. Cunningham C, O' Sullivan R, Caserotti P, Tully MA. Consequences of physical inactivity in older adults: a systematic review of reviews and meta - analyses. Scandinavian Med Sci Sports 2020 May;30(5):816-827 [FREE Full text] [doi: 10.1111/sms.13616]
8. Izquierdo M, Merchant RA, Morley JE, et al. International Exercise Recommendations in Older Adults (ICFSR): expert consensus guidelines. J Nutr Health Aging 2021;25(7):824-853. [doi: 10.1007/s12603-021-1665-8] [Medline: 34409961]
9. Allender S, Foster C, Boxer A. Occupational and nonoccupational physical activity and the social determinants of physical activity: results from the Health Survey for England. J Phys Act Health 2008 Jan;5(1):104-116. [doi: 10.1123/jpah.5.1.104] [Medline: 18209257]
10. Physical exercise and dementia. Alzheimer's Society. URL: https://www.alzheimers.org.uk/about-dementia/risk-factors-and-prevention/physical-exercise [accessed 2024-09-03]
11. Exercise. Parkinson´s Foundation. 2023 May 8. URL: https://www.parkinson.org/living-with-parkinsons/treatment/exercise [accessed 2024-09-03]
12. Ahmed M, Marín M, How D, et al. End users' and other stakeholders' needs and requirements in the development of a Personalized Integrated Care Platform (PROCare4Life) for older people with dementia or parkinson disease: mixed methods study. JMIR Form Res 2022 Nov 30;6(11):e39199. [doi: 10.2196/39199] [Medline: 36449340]
13. Blakemore A, Hann M, Howells K, et al. Patient activation in older people with long-term conditions and multimorbidity: correlates and change in a cohort study in the United Kingdom. BMC Health Serv Res 2016 Oct 18;16(1):582. [doi: 10.1186/s12913-016-1843-2] [Medline: 27756341]
14. Hibbard JH, Stockard J, Mahoney ER, Tusler M. Development of the Patient Activation Measure (PAM): conceptualizing and measuring activation in patients and consumers. Health Serv Res 2004 Aug;39(4 Pt 1):1005-1026. [doi: 10.1111/j.1475-6773.2004.00269.x] [Medline: 15230939]
15. Bandura A. Self-efficacy: toward a unifying theory of behavioral change. Psychol Rev 1977;84(2):191-215. [doi: 10.1037/0033-295X.84.2.191]
16. DiClemente C, Graydon M. Changing behavior using the transtheoretical model. In: Hagger M, editor. The Handbook of Behavior Change: Cambridge University Press; 2020:136-149.
17. Warner L, French D. Self-efficacy interventions. In: Hagger M, editor. The Handbook of Behavior Change: Cambridge University Press; 2020:461-478.
18. Bentlage E, Nyamadi JJ, Dubbeldam R. The importance of activating factors in physical activity interventions for older adults using information and communication technologies: systematic review. JMIR Mhealth Uhealth 2023 Oct 24;11:e42968. [doi: 10.2196/42968] [Medline: 37933182]
19. Schwarzer R. Health Action Process Approach (HAPA) as a theoretical framework to understand behavior change. AP 2016;30(121):119. [doi: 10.15517/ap.v30i121.23458]
20. Belmonte-Hernández A, Semerci YC, Proença JP, Romera-Giner S. PROCare4Life: an integrated care platform to improve the quality of life of Parkinson's and Alzheimer's patients. 2022 Jun 29 Presented at: PETRA '22; Jun 29 to Jul 1, 2022;

XSL•FO
RenderX

Corfu Greece p. 358-364 URL: https://dl.acm.org/doi/proceedings/10.1145/3529190 [accessed 2025-06-05] [doi: 10.1145/3529190.3534727]
21. Roy D, Dutta M. A systematic review and research perspective on recommender systems. J Big Data 2022 Dec;9(1):59. [doi: 10.1186/s40537-022-00592-5]
22. Singh PK, Choudhury P, Dey AK, Pramanik PKD. Recommender systems: an overview, research trends, and future directions. IJBSR 2021;15(1):14. [doi: 10.1504/IJBSR.2021.10033303]
23. Park DH, Kim HK, Choi IY, Kim JK. A literature review and classification of recommender systems research. Expert Syst Appl 2012 Sep;39(11):10059-10072. [doi: 10.1016/j.eswa.2012.02.038]
24. De Croon R, Van Houdt L, Htun NN, Štiglic G, Vanden Abeele V, Verbert K. Health recommender systems: systematic review. J Med Internet Res 2021 Jun 29;23(6):e18035. [doi: 10.2196/18035] [Medline: 34185014]
25. Casino F, Batista E, Patsakis C, Solanas A. Context-aware recommender for smart health. 2015 Presented at: 2015 IEEE First International Smart Cities Conference (ISC2); Oct 25-28, 2015; Guadalajara, Mexico. [doi: 10.1109/ISC2.2015.7366176]
26. Agapito G, Calabrese B, Guzzi PH, et al. DIETOS: a recommender system for adaptive diet monitoring and personalized food suggestion. Presented at: 2016 IEEE 12th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob); Oct 17-19, 2016; New York, NY. [doi: 10.1109/WiMOB.2016.7763190]
27. Schäfer H, Hors-Fraile S, Karumur RP. Towards health (aware) recommender systems. 2017 Jul 2 Presented at: ACM International Conference Proceeding Series; Jul 2-5, 2017; London United Kingdom. [doi: 10.1145/3079452.3079499]
28. Ahmed M, Marín M, Bouça-Machado R, et al. Investigating users' and other stakeholders' needs in the development of a Personalized Integrated Care Platform (PROCare4Life) for older people with dementia or parkinson disease: protocol for a mixed methods study. JMIR Res Protoc 2021 Jan 12;10(1):e22463. [doi: 10.2196/22463] [Medline: 33433394]
29. Schnall R, Rojas M, Bakken S, et al. A user-centered model for designing consumer mobile health (mHealth) applications (apps). J Biomed Inform 2016 Apr;60:243-251. [doi: 10.1016/j.jbi.2016.02.002] [Medline: 26903153]
30. Bennett JA, Lyons KS, Winters-Stone K, Nail LM, Scherer J. Motivational interviewing to increase physical activity in long-term cancer survivors. Nurs Res 2007 Jan;56(1):18-27. [doi: 10.1097/00006199-200701000-00003]
31. Schwarzer R, Lippke S, Luszczynska A. Mechanisms of health behavior change in persons with chronic illness or disability: the Health Action Process Approach (HAPA). Rehabil Psychol 2011 Aug;56(3):161-170. [doi: 10.1037/a0024509] [Medline: 21767036]
32. RabbitMQ. URL: https://www.rabbitmq.com [accessed 2023-04-17]
33. Lu J, Wu D, Mao M, Wang W, Zhang G. Recommender system application developments: a survey. Decis Support Syst 2015 Jun;74:12-32. [doi: 10.1016/j.dss.2015.03.008]
34. Melin J, Bonn SE, Pendrill L, Trolle Lagerros Y. A questionnaire for assessing user satisfaction with mobile health apps: development using Rasch measurement theory. JMIR Mhealth Uhealth 2020 May 26;8(5):e15909. [doi: 10.2196/15909] [Medline: 32452817]
35. Bakker D, Kazantzis N, Rickwood D, Rickard N. Mental health smartphone apps: review and evidence-based recommendations for future developments. JMIR Ment Health 2016 Mar 1;3(1):e7. [doi: 10.2196/mental.4984] [Medline: 26932350]
36. Costello E, Kafchinski M, Vrazel J, Sullivan P. Motivators, barriers, and beliefs regarding physical activity in an older adult population. J Geriatr Phys Ther 2011;34(3):138-147. [doi: 10.1519/JPT.0b013e31820e0e71] [Medline: 21937904]
37. Brawley LR, Latimer AE. Physical activity guides for Canadians: messaging strategies, realistic expectations for change, and evaluation. Applied Physiology, Nutrition, and Metabolism 2007;32(S2E). [doi: 10.1139/H07-105]
38. Ammar A, Trabelsi K, Brach M, et al. Effects of home confinement on mental health and lifestyle behaviours during the COVID-19 outbreak: insights from the ECLB-COVID19 multicentre study. Biol Sport 2021 Mar;38(1):9-21. [doi: 10.5114/biolsport.2020.96857] [Medline: 33795912]
39. Lee MK, Kim J, Forlizzi J, Kiesler S. Personalization revisited: a reflective approach helps people better personalize health services and motivates them to increase physical activity. In: UbiComp 2015 - Proc 2015 ACM Int Jt Conf Pervasive Ubiquitous Comput: Association for Computing Machinery; 2015:2015. [doi: 10.1145/2750858.2807552]
40. Rhodes RE, Nwachukwu N, Quinlan A. Family exergaming: correlates and preferences. Games Health J 2018 Jun;7(3):188-196. [doi: 10.1089/g4h.2017.0177] [Medline: 29688819]
41. Kunzler F, Kramer JN, Kowatsch T. Efficacy of mobile context-aware notification management systems: a systematic literature review and meta-analysis. 2017 Presented at: 2017 IEEE 13th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob); Oct 9-11, 2017; Rome p. 131-138. [doi: 10.1109/WiMOB.2017.8115839]
42. Zaman SB, Khan RK, Evans RG, Thrift AG, Maddison R, Islam SMS. Exploring barriers to and enablers of the adoption of information and communication technology for the care of older adults with chronic diseases: scoping review. JMIR Aging 2022;5(1):e25251. [doi: 10.2196/25251]

Abbreviations

DD: dementia disease

XSL•FO
RenderX

EU: European Union
HAPA: Health Action Process Approach
PA: physical activity
PD: Parkinson disease
PROCare4Life: Personalized Integrated Care Promoting Quality of Life for Older People
QoL: quality of life

Edited by A Mavragani; submitted 18.08.23; peer-reviewed by A AL-Asadi, T Vanbellingen, X Zhang; revised version received 31.03.25; accepted 01.04.25; published 19.06.25.

Please cite as:
Bentlage E, del Río Ponce A, Ahmed M, Gangas P, Brach M, Kurano JA, Menéndez JM
Cocreating a Mobile Health App Providing Physical Activity Recommendations for Older People Living With Parkinson Disease or Dementia: User-Centered Pilot Study
JMIR Form Res 2025;9:e51831
URL: https://formative.jmir.org/2025/1/e51831
doi:10.2196/51831

© Ellen Bentlage, Alberto del Río Ponce, Mona Ahmed, Pilar Gangas, Michael Brach, Jorge Alfonso Kurano, José Manuel Menéndez. Originally published in JMIR Formative Research (https://formative.jmir.org), 19.6.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Fluctuations in Daily Happiness and Nervousness Based on Depressive and Anxious Symptoms in Adolescents or Young Adults Across 3 Latin American Cities: Experience Sampling Study

Ana L Vilela-Estrada[1], BA; David Villarreal-Zegarra[2,3], MPH; Nikol Mayo-Puchoc[1], BA; Nicola Holt[4], PhD; Ángela Flórez-Varela[5], PGD; Catherine Fung[6], MA; Karen Ariza-Salazar[5], BA; Fernando Luis Carbonetti[7], MD; Sumiko Flores[1], BA; Adriana Carbonel[1], BA; Natividad Olivar[7], MD; Carlos Gomez-Restrepo[5,8,9], PhD; Luis Ignacio Brusco[7], MD; Stefan Priebe[7,10], MD; Francisco Diez-Canseco[1], MPH

[1]CRONICAS Centre of Excellence in Chronic Diseases, Universidad Peruana Cayetano Heredia, Lima, Peru

[10]Unit for Social and Community Psychiatry, East London NHS Foundation Trust, London, United Kingdom

[2]Department of Biomedical Informatics, School of Medicine, University of Utah, Utah, United States

[3]Instituto Peruano de Orientación Psicológica, Lima, Peru

[4]Department of Health and Social Sciences, University of the West of England (UWE), Bristol, United Kingdom

[5]Department of Clinical Epidemiology and Biostatistics, Pontificia Universidad Javeriana, Bogotá, Colombia

[6]Unit for Social and Community Psychiatry, Wolfson Institute of Population Health, Queen Mary University of London, 58 Turner Street, London, United Kingdom

[7]Department of Psychiatry and Mental Health, School of Medicine, University of Buenos Aires, Buenos Aires, Argentina

[8]Department of Psychiatry and Mental Health, Pontificia Universidad Javeriana, Bogotá, Colombia

[9]Hospital Universitario San Ignacio, Bogotá, Colombia

Corresponding Author:
Catherine Fung, MA
Unit for Social and Community Psychiatry, Wolfson Institute of Population Health, Queen Mary University of London, 58 Turner Street, London, United Kingdom

Abstract

Background: Experience sampling methods (ESMs) have been used in clinical research to collect data on emotional and behavioral states in real-life contexts among different populations. Although the use of ESMs in mental health has increased, it has not been applied to larger samples of young people in disadvantaged urban settings.

Objective: This study aimed to determine the extent to which mood status scores (happiness and nervousness) vary during a week, as a function of having or not having symptoms of depression or anxiety, in a sample of adolescents and young adults in the cities of Buenos Aires, Bogotá, and Lima. A secondary objective was to identify factors associated with mood scores, including sociodemographics, quality of life, and daily activities.

Methods: This study was part of the Building Resilience and Resources to Reduce Mental Distress in Young People in Latin America research program, which focuses on mental health resources for young people. Participants (n=143) aged 15 - 24 years completed daily ESM assessments over a week using the mobile app, resulting in 5246 reports. Data were analyzed using descriptive analyses with 2-tailed *t* tests and chi-square tests, and multilevel linear regression was used to examine associations between depressive or anxiety symptoms, mood variability, and factors influencing mean mood. Finally, Spearman correlation assessed the relationship between happiness and nervousness.

Results: The analysis revealed that depressive or anxiety symptoms were not significantly associated with increased variability in mood scores (happiness P=.40 and nervousness P=.84). However, males exhibited greater variability in happiness and nervousness scores (P<.001) than females. Additionally, young people showed higher variability in nervousness than adolescents (P=.02). Regarding average happiness scores, young adults reported higher average happiness than adolescents (β=.604; P=.003). Engaging in structured activities (eg, sports, music lessons, and dance classes) was associated with increased happiness (β=.266; P=.01). In contrast, instrumental activities (eg, cleaning, shopping, meal preparation, or taking medication; β=−.144; P=.02) and work-related tasks (β=−.205; P=.01) were linked to lower happiness and higher nervousness (β=.387; P<.001). Quality of life was positively correlated with happiness (β=.486; P<.001) and negatively correlated with nervousness (β=−.273; P=.005). Finally, as for average scores, a strong negative correlation was found between happiness and nervousness (rs=−0.92; P<.001). The simple multilevel analysis showed that for each point of happiness, nervousness decreased by 0.45 points (95% CI −0.48 to −0.42; t_3=−41.7; P<.001; SE 0.01).

XSL•FO
RenderX

Conclusions: Our study reveals that depressive and anxiety symptoms do not significantly affect the variability in predicted happiness and nervousness scores. However, we observed that demographic factors, such as gender and age, play a role in emotional variability.

(JMIR Form Res 2025;9:e65732) doi:10.2196/65732

KEYWORDS

experience sampling method; mood variability; mood swings; happiness; pediatric; adolescence; teenager; young adult; mental health; mental illness; mental disorder; mental health resources; Latin America; Colombia; Argentina; Peru; mobile phone

Introduction

Background

Adolescence and young adulthood are crucial periods of social, behavioral, and psychological development and change [1]. However, these periods are also associated with increased vulnerability to mental health problems, with most functional mental disorders typically manifesting before the age of 25 years [2]. According to the World Health Organization, approximately 14% of young people worldwide experience mental health problems, with anxiety and depression being the most prevalent [3]. These conditions contribute to approximately 13% of the total health burden in this age group [4], and they are more likely to experience depression and anxiety than older adults [5]. As a result, the World Health Organization has identified the reduction of depression and anxiety in adolescents as a key priority, particularly in low- and middle-income countries [6,7]. This includes Latin America, where adolescents and young people living in large cities are more exposed to risk factors for developing anxiety and depression [8,9]. Some studies in Latin America have estimated that the prevalence of depression or anxiety among young people varies, such as 17% in Colombia [10], 16% in Peru [11], and 20% in Argentina [12].

Evidence suggests that researchers and health professionals should join forces to implement integrated actions to understand the behavior of the incidence and prevalence of mental health problems in young people as one of the first steps in the prevention chain [13]. This should complement retrospective data analysis, which is vulnerable to attrition and recall bias [14], with techniques that explore the triggering of symptoms, considering contextual and social factors and elements of everyday life. In this sense, collecting data on people's experiences and mental states in their daily lives is proposed from psychological and social science research.

The experience sampling method (ESM) is a daily technique of brief assessments that allows data collection in daily life [15]. The ESM uses repeated self-reports through a short questionnaire that participants complete when they receive a signal (eg, phone notifications). The reports assess participants' experiences and behavior in real time, in their natural environment, and at multiple time points throughout their daily lives [15]. The ESM has several methodological advantages: (1) it improves accuracy and ecological validity, (2) it provides an assessment of participants in their natural environment in real time and at multiple points, and (3) it offers a large and dynamic variation in data analysis and reporting [16-18]. Statistically, ESM studies typically generate substantial data, as they can record multiple measures per day over several days. Consequently, ESM data can be used at a descriptive level (eg, estimating the average of a specific variable) as well as at a more complex level (eg, examining variability in the relationships between variables across different groups) [19].

The ESM has been used in clinical research as a tool for a range of psychological, psychophysiological, cognitive, and behavioral data in real-life contexts in diverse populations [18,20,21], including adolescents and young people [22,23]. Compared with traditional techniques (eg, cross-sectional or longitudinal studies), ESM methodology limits recall bias and captures individual variation in mental health indicators in everyday life [18,24]. Multiple measures of an individual using the ESM help to explore affective variability in more detail and identify the context in which adaptive emotional regulation is used [24].

Mood status variability is the changes and fluctuations in how we feel emotionally over time [25]. High emotional variability can disrupt an individual's well-being and trigger emotion dysregulation [25]. In this way, studies frequently focus on the role of basic emotions (eg, happiness, nervousness or anxiety, anger, and sadness) because of their relationship to well-being or mental health problems [26]. This variability tends to occur to a greater extent during adolescence [25]. Capturing these variabilities may aid in the development of future assessments of mental health symptoms. Indeed, longitudinal studies have highlighted the importance of mood variability, given that high levels of "negative" emotions (ie, emotional inertia) may contribute to anxiety and depressive symptoms [26,27]. Research has also found that mood variability is associated with differences in support from others; for example, adolescents reported higher levels of negative mood when they perceived less support from their parents [19].

Moreover, applying this methodology to adolescents highlights the association between mental health indicators and the natural context through the average daily scores. Some studies have reported that adolescents with higher levels of negative mood (eg, depressive symptoms, stress, and negative affect) spend more time alone and at home and have reduced physical activity. Other studies have reported that physical activity [20,24,25], engaging in artistic activities, and being outdoors are associated with higher levels of affective well-being [23,28-30].

Objective

Although the use of the ESM in mental health is becoming more widespread, it has not been applied to larger samples of adolescents and young adults in deprived urban areas of Latin America. Therefore, this study used the ESM to determine the extent to which mood status scores (happiness and nervous) vary during a week, as a function of having or not having

XSL•FO
RenderX

symptoms of depression or anxiety, in a sample of adolescents and young adults in the cities of Buenos Aires (Argentina), Bogotá (Colombia), and Lima (Peru). A secondary aim was to identify factors associated with mood scores, such as sociodemographics, wellness factors (quality of life), and daily activities.

Methods

Study Design

This study is part of the building resilience and resources to reduce mental distress in young people in Latin America (OLA) research program, designed to identify the personal and social resources that help young people prevent and recover from mental distress [31]. We conducted an ESM study in 3 Latin American countries. This study was designed according to the CREMAS (Adapted Strengthening the Reporting of Observational Studies in Epidemiology Checklist for Reporting Ecological Momentary Assessments Studies) checklist [32], which is an adapted STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) Checklist for Reporting Ecological Momentary Assessments.

Ethical Considerations

This study's protocol and all procedures were approved by the Institutional Review Board of Universidad de Buenos Aires (10-02-2020), Pontificia Universidad Javeriana (FM-CIE-1138‑20), and Universidad Peruana Cayetano Heredia (Constancia 581-33-20). Furthermore, it received approval from the Research Ethics Committee of Queen Mary University of London (QMERC2020/02).

In line with ethical guidelines, all young adult participants completed an informed consent form, adolescents completed an informed assent, and their parent or legal guardian provided informed consent. As for the consent procedure, participants were informed about the nature of this study before their participation. They were provided with a detailed explanation of the activity, and those who were willing to participate gave oral or written consent, depending on the specific requirements of this study.

Regarding data privacy and confidentiality, all study data were collected anonymously. No personally identifiable information was linked to the data, and steps were taken to ensure that participants' confidentiality was maintained throughout the research process.

Compensation for participation varied by country. In Peru, participants were offered a voucher redeemable at a chain shop, valued at 50 soles (at an approximate exchange rate of PEN 1=US $0.27). In Colombia, participants received a voucher worth 140,000 Colombian pesos (COP 1=US $0.00021), while in Argentina, compensation was given in the form of cash worth 2000 Argentine pesos (ARS 1=US $0.0053). The compensation was provided to thank participants for their time and involvement in this study.

Participants

The sample consisted of participants between adolescents (15 and 16 y) and young adults (20 to 24 y) from a cohort of the OLA research program from Buenos Aires, Bogotá, and Lima [31]. This research program focused on key developmental stages. This study included 2 groups; the first group comprised individuals aged 15‑16 years, as mental health issues often emerge during this period [33]. The second group included young adults aged 20‑24 years, a stage following adolescence but still relevant for the recurrence of mental health disorders [34]. Participants aged 17‑19 years were deliberately excluded to concentrate on the distinct stages of early adolescence and young adulthood, both of which provide valuable insights into mental health outcomes.

The recruitment of OLA targeted areas within the poorest 50% of neighborhoods, as identified using the Human Development Index [35] in Bogotá and Lima, and the Unsatisfied Basic Needs Index in Buenos Aires [36]. This approach aimed to ensure that the sample was representative of disadvantaged populations in these urban areas. Additionally, the recruitment was carried out through collaborations with service care, community organizations, schools, and youth groups in each of the cities.

This study excluded people with severe mental health problems (such as psychosis, bipolar disorder, and schizophrenia), cognitive impairments, illiteracy, or who were users of Huawei (Huawei Technologies Co, Ltd) phones due to issues that arose during the ESM testing phase. Additionally, participants were excluded if they reported a gender other than male or female and if they did not complete the data on the scales used.

Research assistants initially contacted 359 potential participants in the OLA program, primarily via phone calls and messaging. Of these, 7.8% (28/359) were not interested in participating, 10% (36/359) could not be contacted, 19% (68/359) were excluded due to using Huawei phones, 7.5% (27/359) did not complete the informed consent process, and 13.1% (47/359) did not complete training on app usage.

The average time between participation in the OLA cohort baseline questionnaire and enrollment in this study was 107.6 (SD 29.7) days. For further details, please refer to the main publication regarding this activity [30].

Procedures

Participants who met the inclusion criteria were invited to participate in this study, and informed consent or assent was obtained from those who agreed. Following consent, participants attended a briefing session where the eMoodie (eMoodie Ltd) app was installed on their phones. During this session, they were trained to use the app to complete the Experience Sampling Questionnaire (ESQ).

The eMoodie app, designed specifically for youths, facilitates experience sampling techniques by delivering notifications at random intervals [37]. Over 7 days, participants received notifications to complete ESQs. Notifications for participants aged 15‑16 years were scheduled 5 times daily on school days (excluding school hours) and 8 times daily on weekends. For participants aged 20‑24 years, notifications were sent 8 times daily throughout the 7 days [30]. Their availability was considered to adjust notification times.

XSL•FO
RenderX

Notifications were sent as pop-up alerts on participants' phones, accompanied by a sound alert. Each notification prompted participants to complete an ESQ, which included questions to assess their current mood. Follow-up was conducted throughout this study to ensure compliance. If participants did not respond to notifications during the first half of the day (morning or afternoon), a follow-up call was made to check their progress. If they had not completed the ESQs on day 2, they were contacted again the next day. Additionally, a follow-up call was made if participants completed less than 50% of the ESQs on days 1 and 2. Participants who had responded to more than 50% of the notifications on day 2 were not contacted. Additionally, support was provided as needed, including app troubleshooting and rescheduling of notifications. No extensions were given to participants who missed notifications, as the study adhered to a strict 7-day timeline.

The final dataset was cleaned to remove invalid responses, such as duplicates or ESQs completed in less than 1 second, indicating "skipping," and the study protocol specified that only participants who completed at least 40% of the notifications (compliance rate) were included in the final dataset [31]. After data cleaning, the mean compliance rate was 76%, ranging from 34% to 100%.

Instruments

Basal Measurements

As part of the OLA cross-sectional study [38], all participants completed a questionnaire (online or paper) that included the following measures.

The Patient Health Questionnaire-8 (PHQ-8) was used to measure the severity of depressive symptoms. Eight items are on a 4-point Likert scale (0=not at all, 1=several days, 2=more than half of the days, and 3=almost every day). Participants were considered to have depressive symptoms if they had a cutoff score of 10 or more [39].

The General Anxiety Disorder-7 (GAD-7) was used to measure the severity of anxiety symptoms. It has 7 items on a 4-point Likert scale (0=not at all, 1=several days, 2=more than half of the days, and 3=almost every day). Participants were considered to have anxiety symptoms if they had a cutoff score of 10 or more [40].

Our study categorized participants into 2 groups: those with depressive or anxious symptoms and those without symptoms. They were considered to have depressive or anxious symptoms if their scores on the PHQ-8 or GAD-7 were 10 points or more [39,40], respectively.

The Manchester Short Assessment of Quality of Life (MANSA) was used to measure satisfaction with life, work or unemployment, financial situation, number and quality of friendships, sex life, leisure activities, housing, personal safety, people living with others or living alone, family relationships, physical health, and mental health. Satisfaction is measured using a 7-point Likert scale ranging from 1=totally dissatisfied to 7=totally satisfied [41].

As for sociodemographic characteristics, information was collected on participants' age, gender (male or female), country (Argentina, Colombia, or Peru), main occupation (work, study, housewife, other, or no current occupation), if currently working (no or yes), and education level (primary and high school education, or technical and university education).

ESM Component

The following ESQ questionnaire [30] consisted of the following questions:

1. "What are you doing right now?" for a free description of the main activity being performed. Responses were open text in Spanish and were first organized into simple codes for activities such as walking, eating, bathing, etc. The codes were then grouped into 6 broad categories based on categories identified in the literature: (1) Basic activities: activities related to self-maintenance and self-care. Some examples are eating, dressing, personal care and grooming, mobility, etc [42]. (2) Instrumental activities: these are more complex than previous ones and allow people to adapt to their environment. Examples include managing finances, cleaning the home, shopping, preparing meals, taking medication, etc [42]. (3) Relaxation activities: this category refers to more informal and unstructured activities that young people engage in to relax, unwind, or simply pass the time without a specific purpose. This can include activities such as watching television, listening to music, chatting with friends, playing video games, reading for pleasure, taking a nap, etc [43]. (4) Structured activities: these are organized, planned, and have a specific purpose. Generally, these activities are scheduled in advance and have a set timetable, such as sports practice, music lessons, dance lessons, academic tutoring, etc [44]. The latter type of activity included (5) work and (6) educational activities.
2. "How happy or sad do you feel right now?" answered with a 7-point scale from 1=very happy to 7=very sad.
3. "How nervous or calm are you right now?" answered with a 7-point scale from 1=very nervous to 7=very relaxed and calm). In this case, we reverse the scores so that higher scores represent higher nervousness levels.

Statistical Analyses

Characteristics of the Participants

A descriptive analysis was conducted based on the sociodemographic characteristics and variables of interest of the participants. The analysis was performed for the total sample and stratified by the presence or absence of depressive or anxious symptoms. For numerical variables, a 2-tailed Student *t* test was used to assess differences in sociodemographic characteristics based on the presence of depressive or anxious symptoms. Additionally, for categorical variables, a chi-square test was performed to evaluate the association between depressive or anxious symptoms with categorical variables.

Average Levels of Happiness and Nervousness

Data were reviewed to identify and remove invalid observations and duplicates. The average compliance rate after cleaning was 76% [30]. A multilevel linear regression analysis was conducted to examine the association between having depressive or anxious symptoms (dichotomized exposition 0 and 1) and the dimensions

XSL•FO
RenderX

of happiness (ranging from 1 to 7) and nervousness (ranging from 1 to 7) as criterion variables. We used restricted maximum likelihood. We considered a composite exposure to be the presence of any conditions for the presence of depressive or anxiety symptoms. This analysis considered the nested structure of the data. The first level represented each measurement of happiness and nervousness, averaging 5 to 8 measurements daily. The second level represented each day, with a total of 7 days. Finally, the third level corresponded to each participant, 143 participants (approximately 50 for each country). We consider the presence of depressive or anxious symptoms, gender, age group, quality of life, educational level, and whether the participant is currently working as fixed effects. Additionally, we consider the happiness and nervousness scores as random effects, so we consider these variables using a random intercept and fixed slopes, allowing each participant to have their baseline level of both variables. Further, our analysis controls the time between self-report measures, as we expect measures of happiness and nervousness taken closer together in time to be more correlated than those taken further apart [45].

The use of a multilevel linear regression model accounted for the variability both between measurements taken on each day and between participants, allowing us to capture the hierarchical nature of the data and to examine the association of depressive and anxious symptoms with the dimensions of happiness and nervousness, also adjusting for relevant sociodemographic factors. The effect size was estimated using the adjusted coefficient and a 95% CI. Statistical significance was assessed at a significance level of less than .05. A random effects model was used to account for the nested structure of the data.

Variability in Predicted Happiness and Nervousness Scores (Main Analysis)

Our study assessed the variance of the predicted scores for happiness and nervousness using the previous multilevel linear analysis results. We examined whether there were significant differences in the variances of these predicted scores between different groups, such as gender and age groups, to determine whether the variances were equal or differed significantly between them.

Our ANOVA focuses on a robust test for equality of variances. Significant values mean that the variances of the tested groups are different ($P<.05$), and therefore, the scores of 1 group have more variability than another. The Levene test analysis assumes that the scores are independent [45]. Although our data are nested, we considered it appropriate to use this test as it was applied to the predicted values of happiness and nervousness obtained from the linear multilevel model. These predictions were adjusted for covariates, including both the random and fixed effects of the model. This consideration ensures that the predictions adequately reflect variations in the data adjusted for these covariates.

Relationship Between Happiness and Nervousness

We conducted a correlation analysis using the predicted scores obtained from the full multilevel model to assess the correlation between these 2 variables. The Kolmogorov-Smirnov normality test with Lilliefors correction identified that the normal distribution assumption was unmet. We performed a correlation analysis between predicted happiness and nervousness scores using Spearman correlation coefficient because a nonnormal distribution was assumed, as these are categorical ordinal values. We used the cutoff of the weak (>0.2), moderate (>0.5), and strong correlations (>0.8).

Additionally, our study conducted a simple multilevel analysis to explore the relationship between happiness and nervousness, assuming the same standards as previously mentioned. This analysis adopts a model incorporating a fixed slope, random intercept, and random effects for happiness and nervousness. We do not use any covariates.

Results

Characteristics of the Participants

Initially, 151 individuals participated in the original study; however, 3 participants were excluded for reporting different genders, 4 participants were excluded for not providing MANSA data, and 1 participant was excluded for not reporting education level. Therefore, this study's sample consisted of 143 participants from Argentina (n=48, 33.6%), Colombia (n=46, 32.2%), and Peru (n=49, 34.3%). The mean age of the participants was 18.6 (SD 3.4) years.

The participants were reported over 7 days, and a total of 5246 reports were collected or completed. Regarding the activities performed by the participants, relaxation activities were the most common (n=1904, 36.3%), followed by basic activities (n=1033, 19.7%). Most participants were women (n=85, 59.4%), exhibited depressive or anxious symptoms (n=86, 60.1%), had studied as their primary occupation (n=105, 73.4%), and were not current smokers (n=104, 72.7%). These activities and the sociodemographic characteristics are shown in Table 1.

XSL•FO
RenderX

Table . Sociodemographic characteristics and participant activities in adolescents and young adults across 3 Latin American cities (N=143).

		Overall (n=143)	With depressive or anxious symptoms (n=86)	Without depressive or anxious symptoms (n=57)	*P* value
Numerical variable, mean (SD)[a]					
Age (years)		18.6 (3.4)	18.7 (3.5)	18.5 (3.3)	.75
MANSA[b] score		4.65 (1.1)	4.22 (1)	5.26 (0.9)	<.001
Happiness (weekly average)		4.96 (1.4)	4.78 (1.5)	5.24 (1.3)	<.001
Nervousness (weekly average)		2.95 (1.6)	3.17 (1.6)	2.64 (1.4)	<.001
Categorical variable, n (%)[c]					
Gender					.32
	Male	58 (40.6)	32 (37.2)	26 (45.6)	
	Female	85 (59.4)	54 (62.8)	31 (54.4)	
Country					.98
	Argentina	48 (33.6)	29 (33.7)	19 (33.3)	
	Colombia	46 (32.2)	28 (32.6)	18 (31.6)	
	Peru	49 (34.3)	29 (33.7)	20 (35.1)	
Age group					.81
	Adolescents (15 and 16 y)	71 (49.7)	42 (48.8)	29 (50.9)	
	Young adults (20 to 24 y)	72 (50.3)	44 (51.2)	28 (49.1)	
Main occupation					.96
	Work	19 (13.3)	12 (14)	7 (12.3)	
	Study	105 (73.4)	62 (72.1)	43 (75.4)	
	Housewife	7 (4.9)	5 (5.8)	2 (3.5)	
	Other	3 (2.1)	2 (2.3)	1 (1.8)	
	No current occupation	9 (6.3)	5 (5.8)	4 (7)	
Currently working					.55
	No	104 (72.7)	25 (29.1)	14 (24.6)	
	Yes	39 (27.3)	61 (70.9)	43 (75.4)	
Education level					.70
	Primary and high school education	98 (68.5)	60 (69.8)	38 (66.7)	
	Technical and university education	45 (31.5)	26 (30.2)	19 (33.3)	
Activities carried out[d]					
	Basic activities	1033 (19.7)	588 (18.9)	445 (20.8)	.09
	Relaxation activities	1904 (36.3)	1170 (37.6)	734 (34.3)	.01
	Educational activities	557 (10.6)	284 (9.1)	273 (12.8)	<.001
	Structured activities	129 (2.5)	45 (1.5)	84 (3.9)	<.001
	Instrumental activities	884 (16.8)	547 (17.6)	337 (15.8)	.08

XSL•FO
RenderX

	Overall (n=143)	With depressive or anxious symptoms (n=86)	Without depressive or anxious symptoms (n=57)	*P* value
Work activities	309 (5.9)	215 (6.9)	94 (4.4)	<.001

[a] Student *t* test

[b]MANSA: Manchester Short Assessment of Quality of Life.

[c]Chi-square test.

[d]As the denominator, 5246 evaluations were performed, and only the most relevant activities were reported.

Participants obtained a mean score of 5.05 (SD 1.57) on the happiness scale, which is closer to being happy than to being sad. As for the mean score for nervousness, it was 4.96 (SD 1.42), and this mean score leans more toward nervousness than calmness. Table 1 presents the differences in sociodemographic characteristics between participants with and without depressive or anxious symptoms. Participants with these symptoms showed lower quality of life and happiness scores and higher nervousness scores.

Variability in Predicted Happiness and Nervousness Scores

According to our multilevel model, having depressive or anxiety symptoms did not result in significant differences in the variability of predicted happiness (*P*=.40) and nervousness (*P*=.84) scores compared to not having symptoms. This means that participants with depressive or anxious symptoms do not show greater variability in their scores between the different measurements; that is, there are no large changes from high to low scores from one measurement to another compared to people without such symptoms.

However, we did observe a relationship between the variations in the predicted nervousness scores and some demographic variables, including gender and age group. Specifically, males and individuals aged 20 to 24 years exhibited higher variability in their nervousness scores than females and adolescents, respectively.

Similarly, differences in the variations of the predicted happiness scores were associated with gender. Thus, male participants exhibited the highest variability in their happiness scores (Table 2).

Table . Variability in happiness and nervousness scores in adolescents and young adults across 3 Latin American cities. Our analysis used a robust test for equality of variances. Analysis was performed on 143 participants with a total of 5246 experience sampling method reports.

		Happiness			Nervousness		
		SD	W_{50}	*P* value	SD	W_{50}	*P* value
Gender			107.3	<.001		26	<.001
	Female	0.474			0.373		
	Male	0.561			0.408		
Age group			0.4	.51		5.2	.02
	Adolescents (15 and 16 y)	0.524			0.397		
	Young adults (20 to 24 y)	0.539			0.420		
Depressive or anxious symptoms			0.7	.40		0	.84
	No	0.492			0.354		
	Yes	0.505			0.342		
Day			0.6	.70		0.5	.77
	Monday	0.539			0.418		
	Tuesday	0.536			0.418		
	Wednesday	0.542			0.418		
	Thursday	0.525			0.402		
	Friday	0.527			0.409		
	Saturday	0.527			0.401		
	Sunday	0.555			0.421		

XSL•FO
RenderX

Average Mean Scores of Happiness and Nervousness

Young adults had higher mean scores of happiness than adolescents (β=.604; *P*=.003; 95% CI 0.214 to 0.994). In addition, quality of life scores were associated with higher happiness scores (β=.486; *P*<.001; 95% CI 0.334 to 0.637). Regarding specific activities they engaged in during the assessment, participants who were engaged in structured activities had higher mean happiness scores than those who did not engage in these activities (β=.266; *P*=.01; 95% CI 0.055 to 0.476). On the contrary, participants who were engaged in instrumental activities (β=−.144; *P*=.02; 95% CI −0.263 to −0.026) or work activities (β=−.205; *P*=.01; 95% CI −0.362 to −0.048) had lower mean happiness scores than those who were not engaged in these activities (Table 3).

The mean scores of nervousness were negatively associated with quality of life scores (β=−.273; *P*=.005; 95% CI −0.463 to −0.084) and were higher among those engaged in work activities (β=.387; *P*<.001; 95% CI 0.214 to 0.561) compared to those who did not (Table 4).

Table . Average happiness levels in adolescents and young adults across 3 Latin American cities. Analysis was performed on 143 participants with a total of 5246 experience sampling method reports.

		Mean predicted		Estimate	SE	*t* test (*df*)	*P* value	Lower[a]	Upper[b]
		No	Yes						
Depressive or anxious symptoms		5.19	4.8	0.103	0.16	0.64 (136)	.52	−0.216	0.421
Gender		5.07[c]	4.88[d]	−0.021	0.14	−0.15 (136)	.88	−0.302	0.26
Age group		4.89[e]	5[f]	0.604	0.2	3.06 (136)	.003	0.214	0.994
Activities									
	Basic activities	4.89	5.22	0.103	0.06	1.77 (4254)	.33	−0.641	0.848
	Relaxation activities	4.97	4.94	−0.072	0.05	−1.31 (4254)	.42	−0.767	0.624
	Educational activities	5	4.58	−0.368	0.07	−5.23 (4254)	.12	−1.262	0.527
	Structured activities	4.95	5.37	0.266	0.11	2.47 (4254)	.01	0.055	0.476
	Instrumental activities	4.97	4.88	−0.144	0.06	−2.39 (4254)	.02	−0.263	−0.026
	Work activities	4.97	4.82	−0.205	0.08	−2.56 (4254)	.01	−0.362	−0.048
MANSA[g] score (numeric)		—[h]	—	0.486	0.08	6.34 (136)	<.001	0.334	0.637

[a]Lower: lower bound 95% interval.

[b]Upper: upper bound 95% interval.

[c]Male.

[d]Female.

[e]Adolescents (15 and 16 y).

[f]Young adults (20 to 24 y).

[g]MANSA: Manchester Short Assessment of Quality of Life.

[h]Not applicable.

XSL•FO
RenderX

Table . Average nervousness levels (reverse coded) in adolescents and young adults across 3 Latin American cities. Analysis was performed on 143 participants with a total of 5246 experience sampling method reports.

		Mean predicted		Estimate	SE	*t* test (*df*)	*P* value	Lower[a]	Upper[b]
		No	Yes						
Depressive or anxious symptoms		2.7	3.15	0.173	0.2	0.86 (136)	.39	−0.227	0.573
Gender		2.82[c]	3.07[d]	0.144	0.18	0.81 (136)	.42	−0.208	0.496
Age group		5[e]	5.06[f]	−0.431	0.25	−1.74 (136)	.08	−0.921	0.059
Activities									
	Basic activities	3.02	2.73	−0.092	0.07	−1.4 (4254)	.39	−0.923	0.739
	Relaxation activities	3	2.91	0.009	0.06	0.15 (4254)	.91	−0.763	0.781
	Educational activities	2.9	3.49	0.554	0.08	7.11 (4254)	.09	−0.436	1.545
	Structured activities	2.97	2.59	−0.182	0.12	−1.53 (4254)	.13	−0.415	0.051
	Instrumental activities	2.95	3.01	0.117	0.07	1.75 (4254)	.08	−0.014	0.248
	Work activities	2.94	3.29	0.387	0.09	4.38 (4254)	<.001	0.214	0.561
MANSA[g] score (numeric)		—[h]	—	−0.273	0.1	−2.85 (136)	.005	−0.463	−0.084

[a]Lower: lower bound 95% interval.

[b]Upper: upper bound 95% interval.

[c]Male.

[d]Female.

[e]Adolescents (15 and 16 y).

[f]Young adults (20 to 24 y).

[g]MANSA: Manchester Short Assessment of Quality of Life.

[h]Not applicable.

Relationship Between Happiness and Nervousness

The full multilevel analysis, which includes various covariates, shows a strong and negative correlation between the predicted values of happiness and nervousness (rs=−0.92; P<.001; n=5246). Additionally, the simple multilevel analysis found that for each point of happiness, the nervousness score decreased by an average of 0.45 points (95% CI −0.48 to −0.42; t_3=−41.7; P<.001; SE 0.01).

Discussion

Main Findings and Comparison With Previous Work

This study used the ESM to examine daily mood status variability (happiness and nervousness) in a sample of adolescents and young adults with (n=86/143, 60%), and without (n= 57/143, 40%) anxiety or depressive symptoms in Buenos Aires, Bogotá, and Lima. Our findings revealed that having depressive or anxiety symptoms is not significantly associated with differences in the variability of daily happiness and nervousness scores.

Although some studies have found that mood variability in patients with mental health problems [46-48], mainly in adults with major depressive disorder, was higher than in individuals without mental health problems, there is still much debate about the nature of mood fluctuations in this population. Some studies are consistent with our findings, reporting “emotional inertia,” “emotional inflexibility,” or lower mood variability over time in these individuals. For example, Lamers et al [49] and Nelson et al [50] found no significant differences in mood variability between those with and without mental health issues. Kuppens et al [27] argue that depressive episodes involve high emotional inertia, implying less mood variability, while Shin et al [51] identified that negative experiences correlated with anxiety and depression but did not significantly impact variability.

These mixed findings highlight the complexity of mood variability to mental health, indicating that other factors may also influence these dynamics. Considering this, Pemberton and Tyszkiewicz [52] suggested that the apparent discrepancy between stability and variability in assessments of mood in individuals with mental health problems may be due to the use of different time frames in other studies. It is, therefore, plausible that individuals manifest both short-term mood stability and

XSL•FO
RenderX

variability when examined over a more extended period. The mood swings observed in our study may reflect the "short-term stability" of mood in individuals with mental health problems, and it is possible that over a longer period, fluctuations indicating severity or recovery from emotional distress may become more evident. For example, a study of 482 Dutch adolescents followed up 3 times a year for 5 years to observe how their mood varied and how this related to depressive symptoms, delinquent acts, and alcohol consumption. Most of the adolescents showed emotional stability; however, a small group experienced an increase in mood variability during middle adolescence, showing more persistent depressive symptoms and delinquent behaviors [53]. Similarly, 1 study assessed the daily fluctuation in mood (ie, happiness, anger, sadness, and anxiety) with 54 adolescents (9 - 12 y) during 4 months at a daily reporting interval. Results evidenced "relatively stable" levels of mood during the first months of the COVID-19 pandemic [54]. Another study of 58 early adolescents from the Netherlands was conducted over 7 days to explore associations between social context and adolescents' daily mood state. Participants were enrolled in 5 to 8 daily prompts, using the Ecological Momentary Assessment. Early adolescents showed mood variability when they are alone than in peripheral company [55].

However, our results regarding emotional variability reveal a notable divergence from existing literature on gender differences in adolescents and young adults [25,56]. Specifically, we found that males exhibited greater mood variability in happiness and nervousness than females. This variability was particularly pronounced in happiness scores compared to females. This observation contrasts with the common finding in the literature that females typically report greater emotional variability. However, it aligns with studies suggesting that males might experience more intense emotional responses to stimuli that induce happiness, excitement, or perceived challenges [57,58].

Finally, concerning emotional variability, young adult participants exhibited greater variability than adolescents in nervousness scores. Previous studies have shown that greater emotional instability is common during late adolescence and young adulthood, resulting in wider and more rapid mood swings. However, these are not always largely related to psychological or social maladjustment within each adolescent [59,60]. These fluctuations in mood are usually the result of internal or external processes to which a person is more or less sensitive, but they will also involve individual and contextual aspects such as the affective base (the initial emotional state) and the strength of the attractor (the tendency to return to a specific emotional state) [27].

It was identified that there is a significant and negative relationship between the average scores of happiness and nervousness obtained during the follow-up week, meaning that as happiness increases, nervousness tends to decrease, and vice versa [61]. This finding aligns with the theoretical understanding that these two emotions are inversely related and suggests that intervening in one emotion could impact the other [61]. Recognizing this interplay can inform the development of more effective strategies for emotional well-being, potentially incorporating approaches that address both positive and negative emotions to achieve a more balanced emotional state.

We also found that young adults and those with higher perceived quality of life had higher average happiness scores during the follow-up week. Some studies suggest a reduction in subjective well-being during adolescence, which could be attributed to factors such as social pressure, uncertainty about the future, and other typical changes in this life stage [62-64]. As individuals transition into adulthood, they often develop new identities and relationship patterns, potentially leading to increased cohesion and commitment to personal and social goals [65]. Additionally, previous research indicates that higher subjective well-being is associated with more favorable perceptions of various life aspects, such as health, interpersonal relationships, and the work environment [66]. This supports our findings, suggesting that lower quality of life scores are related to higher nervousness levels.

For daily activities reported in the ESM questionnaire by participants during the follow-up week, engaging in structured activities (such as playing sports, doing art activities, having group activities with others, etc) was positively related to happiness scores. This is consistent with previous literature using the ESM, where adolescents reported feeling happier during structured leisure activities due to the opportunity to develop activities of interest and the positive impact on mood improvement in everyday life, and found that time spent with peers particularly reduces symptoms of anxiety and sadness [67,68].

Finally, doing instrumental activities (eg, managing finances, cleaning the home, shopping, preparing meals, etc) was negatively related to happiness, and performing work activities was related to lower happiness and higher nervousness levels. These results could be explained by the fact that instrumental activities may be perceived as a burden or an obligation rather than as rewarding or satisfying [68-71]. Similarly, work activities can be a source of stress and worry for many people. This may be due to work pressure, job demands, or lack of job satisfaction.

Strengths and Limitations

Using the ESM, this study collects real-time data on emotional and behavioral states in real contexts from young people in urban areas in middle-income countries. The use of the ESM to explore variability in mood was fruitful, as it allowed us to analyze temporal and contextual patterns that, although associative, suggest that engagement in activities (eg, work and instrumental) is important for reducing distress in young people's daily lives. Additionally, our sample size was relatively higher compared to other similar studies [54,55], enabling us to understand the variability in mood status, however, generalizability should be treated with caution.

However, it is important to consider some study limitations when interpreting the results. A limitation of this study is that data collection occurred an average of 107.6 (SD 29.7) days after participants completed the baseline OLA cohort questionnaire. This time lag may have affected the accuracy of characterizing participants, particularly in differentiating

XSL•FO
RenderX

between those with and without depressive or anxiety symptoms, as the data may not fully reflect their emotional states before the start of the week of data collection. Although emotional state was assessed using 2 self-report questions with Likert-type scales to capture different mood extremes (eg, sad-happy or nervous-relaxed), this methodology, although friendly due to the frequency of daily notifications, could have led to participants responding in a way that projected a positive self-image or was aligned with perceived expectations. Furthermore, although the activities performed were assessed, the quality of these experiences (positive or negative) was not considered, nor were other factors that might interact with the activities to influence mood, such as competence or self-esteem.

Finally, although compliance rates were satisfactory, problems with receiving notifications may have affected the participation of some individuals due to issues such as poor or no internet connection during the eMoodie app download, specific technical problems with eMoodie on certain smartphone models, or battery-related problems.

Implications in Public Health

The findings from this study underscore the importance of developing customized approaches for enhancing emotional well-being, considering the interplay between demographic factors, quality of life, and emotional variability. Further research is needed to investigate how elements such as gender, age, and the nature of daily activities affect emotional experiences. For clinical practice, these results suggest that interventions may be more successful if they promote engagement in structured activities and focus on improving overall quality of life. Understanding variations in emotional responses across different demographic groups can inform the creation of more targeted and individualized strategies. While it is well known that adolescents experience variability in mood state, professionals need to acknowledge that mood variability should not relate to psychological or social maladjustment, given that mood swings seem to be an inherent aspect of adolescent lifestyle [60]. From a policy standpoint, ensuring fair access to opportunities and activities that support emotional health is essential. Policies facilitating access to structured and fulfilling activities could enhance emotional well-being and help mitigate differences in emotional experiences among various groups.

Conclusions

Our study reveals that depressive and anxiety symptoms do not significantly affect the variability in predicted happiness and nervousness scores. However, we observed that demographic factors, such as gender and age, play a role in emotional variability, with males and individuals aged 20 to 24 years exhibiting greater nervousness fluctuations and males showing higher variability in happiness. Additionally, young adults reported higher average happiness scores than adolescents, and higher quality of life was associated with greater happiness. Structured activities were linked to increased happiness, whereas instrumental and work-related activities were associated with lower happiness and higher nervousness. These findings highlighted the importance of considering demographic factors and the nature of daily activities when evaluating emotional well-being and suggested that enhancing quality of life and engaging in structured activities could contribute positively to emotional health.

Acknowledgments

The authors would like to thank all participants of this study for sharing their data, experiences, and views with us. This study was funded by the Medical Research Council (MR/S03580X/1).

Data Availability

The dataset analyzed during this study will be available to researchers beginning 9 months and ending 36 months following paper's publication, on reasonable request from Professor Victoria Bird. Proposals should be directed to v.j.bird@qmul.ac.uk to gain access; data requesters will need to sign a data access agreement.

Authors' Contributions

ALV-E, DV-Z, NM-P, SP, and NH handled the conceptualization and did the investigation for this study. AF-V, CF, KA-S, and FLC curated the data. ALV-E and DV-Z worked on the formal analysis. SP, NH, CG-R, LIB, and FD-C acquired the funding. CF administered the project. SP and FD-C supervised this study. NH and DV-Z validated this study. ALV-E, DV-Z, and NM-P wrote the original draft. ALV-E, DV-Z, NM-P, SP, NH, CF, SF, AF-V, KA-S, FLC, AC, NO, LIB, and FD-C reviewed and edited the writing.

Conflicts of Interest

None declared.

References

1. Roach A, Sureshkumar DS, Elliot K, et al. One-year recovery rates for young people with depression and/or anxiety not receiving treatment: a systematic review and meta-analysis. BMJ Open 2023 Jul 21;13(7):e072093. [doi: 10.1136/bmjopen-2023-072093] [Medline: 37479525]

XSL•FO
RenderX

2. Solmi M, Radua J, Olivola M, et al. Age at onset of mental disorders worldwide: large-scale meta-analysis of 192 epidemiological studies. Mol Psychiatry 2022 Jan;27(1):281-295. [doi: 10.1038/s41380-021-01161-7] [Medline: 34079068]
3. Mental health of adolescents. World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/adolescent-mental-health [accessed 2025-05-14]
4. Biswas T, Scott JG, Munir K, et al. Global variation in the prevalence of suicidal ideation, anxiety and their correlates among adolescents: a population based study of 82 countries. EClinicalMedicine 2020 Jul;24:100395. [doi: 10.1016/j.eclinm.2020.100395] [Medline: 33015595]
5. Sikström S, Kelmendi B, Persson N. Assessment of depression and anxiety in young and old with a question-based computational language approach. Npj Ment Health Res 2023 Jul 24;2(1):11. [doi: 10.1038/s44184-023-00032-z] [Medline: 38609578]
6. Zhao H, Li L, Zhang X, et al. Global, regional, and national burden of depressive disorders among young people aged 10-24 years, 2010-2019. J Psychiatr Res 2024 Feb;170:47-57. [doi: 10.1016/j.jpsychires.2023.11.047] [Medline: 38103449]
7. Santomauro DF, Herrera AMM, Shadid J. Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. Lancet 2021 Nov 6;398(10312):1700-1712. [doi: 10.1016/S0140-6736(21)02143-7] [Medline: 34634250]
8. Varese F, Smeets F, Drukker M, et al. Childhood adversities increase the risk of psychosis: a meta-analysis of patient-control, prospective- and cross-sectional cohort studies. Schizophr Bull 2012 Jun;38(4):661-671. [doi: 10.1093/schbul/sbs050] [Medline: 22461484]
9. Thapar A, Collishaw S, Pine DS, Thapar AK. Depression in adolescence. Lancet 2012 Mar 17;379(9820):1056-1067. [doi: 10.1016/S0140-6736(11)60871-4] [Medline: 22305766]
10. Gómez-Restrepo C, Bohórquez A, Tamayo Martínez N, et al. Trastornos depresivos y de ansiedad y factores asociados en la población de adolescentes colombianos, Encuesta Nacional de Salud Mental 2015. Revista Colombiana de Psiquiatría 2016 Dec;45:50-57. [doi: 10.1016/j.rcp.2016.09.009]
11. Y 75 años después una generación optimista lista para transformar el perú y el mundo. UNICEF Perú. URL: https://www.unicef.org/peru/informe/75-anos-despues-generacion-optimista-lista-transformar-peru-mundo [accessed 2025-05-14]
12. Ballarino F. ¿La juventud sin preocupaciones? los problemas de salud mental en la adolescencia UNICEF | para cada infancia. UNICEF Argentina. 2022. URL: https://www.unicef.org/argentina/historias/la-juventud-sin-preocupaciones-los-problemas-de-salud-mental-en-la-adolescencia [accessed 2025-05-14]
13. Colizzi M, Lasalvia A, Ruggeri M. Prevention and early intervention in youth mental health: is it time for a multidisciplinary and trans-diagnostic model for care? Int J Ment Health Syst 2020;14(1):23. [doi: 10.1186/s13033-020-00356-9] [Medline: 32226481]
14. Tzouvara V, Kupdere P, Wilson K, Matthews L, Simpson A, Foye U. Adverse childhood experiences, mental health, and social functioning: a scoping review of the literature. Child Abuse Negl 2023 May;139:106092. [doi: 10.1016/j.chiabu.2023.106092] [Medline: 36907117]
15. Verhagen SJW, Hasmi L, Drukker M, van Os J, Delespaul P. Use of the experience sampling method in the context of clinical trials. Evid Based Ment Health 2016 Aug;19(3):86-89. [doi: 10.1136/ebmental-2016-102418] [Medline: 27443678]
16. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol 2008;4(1):1-32. [doi: 10.1146/annurev.clinpsy.3.022806.091415] [Medline: 18509902]
17. Ebner-Priemer UW, Trull TJ. Ecological momentary assessment of mood disorders and mood dysregulation. Psychol Assess 2009 Dec;21(4):463-475. [doi: 10.1037/a0017075] [Medline: 19947781]
18. Oorschot M, Kwapil T, Delespaul P, Myin-Germeys I. Momentary assessment research in psychosis. Psychol Assess 2009;21(4):498-505. [doi: 10.1037/a0017077]
19. Myin-Germeys I, Kasanova Z, Vaessen T, et al. Experience sampling methodology in mental health research: new insights and technical developments. World Psychiatry 2018 Jun;17(2):123-132. [doi: 10.1002/wps.20513] [Medline: 29856567]
20. Kim H, Kim S, Kong SS, Jeong YR, Kim H, Kim N. Possible application of ecological momentary assessment to older adults' daily depressive mood: integrative literature review. JMIR Ment Health 2020 Jun 2;7(6):e13247. [doi: 10.2196/13247] [Medline: 32484442]
21. Kivelä L, van der Does WAJ, Riese H, Antypa N. Don't miss the moment: a systematic review of ecological momentary assessment in suicide research. Front Digit Health 2022;4:876595. [doi: 10.3389/fdgth.2022.876595] [Medline: 35601888]
22. Mölsä ME, Lax M, Korhonen J, Gumpel TP, Söderberg P. The experience sampling method in monitoring social interactions among children and adolescents in school: a systematic literature review. Front Psychol 2022;13:844698. [doi: 10.3389/fpsyg.2022.844698] [Medline: 35444596]
23. Li YM, Hachenberger J, Lemola S. The role of the context of physical activity for its association with affective well-being: an experience sampling study in young adults. Int J Environ Res Public Health 2022 Aug 23;19(17):10468. [doi: 10.3390/ijerph191710468] [Medline: 36078182]
24. Aldao A. The future of emotion regulation research: capturing context. Perspect Psychol Sci 2013 Mar;8(2):155-172. [doi: 10.1177/1745691612459518] [Medline: 26172497]
25. Maciejewski DF, van Lier PAC, Branje SJT, Meeus WHJ, Koot HM. A 5-year longitudinal study on mood variability across adolescence using daily diaries. Child Dev 2015;86(6):1908-1921. [doi: 10.1111/cdev.12420] [Medline: 26464350]

XSL•FO
RenderX

26. Neumann A, van Lier PAC, Frijns T, Meeus W, Koot HM. Emotional dynamics in the development of early adolescent psychopathology: a one-year longitudinal study. J Abnorm Child Psychol 2011 Jul;39(5):657-669. [doi: 10.1007/s10802-011-9509-3] [Medline: 21494863]
27. Kuppens P, Oravecz Z, Tuerlinckx F. Feelings change: accounting for individual differences in the temporal dynamics of affect. J Pers Soc Psychol 2010 Dec;99(6):1042-1060. [doi: 10.1037/a0020962] [Medline: 20853980]
28. Holt NJ. Using the experience-sampling method to examine the psychological mechanisms by which participatory art improves wellbeing. Perspect Public Health 2018 Jan;138(1):55-65. [doi: 10.1177/1757913917739041] [Medline: 29130867]
29. de Vries LP, Baselmans BML, Bartels M. Smartphone-based ecological momentary assessment of well-being: a systematic review and recommendations for future studies. J Happiness Stud 2021;22(5):2361-2408. [doi: 10.1007/s10902-020-00324-7] [Medline: 34720691]
30. Holt NJ, Flórez-Varela Á, Gómez-Restrepo C, et al. Which personal and social resources help adolescents to recover from negative affect in daily life? An experience sampling study. J Child Adolesc Ment Health 2024 Jul 23:1-22. [doi: 10.2989/17280583.2024.2385308]
31. Priebe S, Fung C, Brusco LI, et al. Which resources help young people to prevent and overcome mental distress in deprived urban areas in Latin America? A protocol for a prospective cohort study. BMJ Open 2021 Sep 13;11(9):e052339. [doi: 10.1136/bmjopen-2021-052339] [Medline: 34518275]
32. Liao Y, Skelton K, Dunton G, Bruening M. A systematic review of methods and procedures used in ecological momentary assessments of diet and physical activity research in youth: an adapted STROBE Checklist for Reporting EMA Studies (CREMAS). J Med Internet Res 2016 Jun 21;18(6):e151. [doi: 10.2196/jmir.4954] [Medline: 27328833]
33. Hankin BL, Young JF, Abela JRZ, et al. Depression from childhood into late adolescence: influence of gender, development, genetic susceptibility, and peer stress. J Abnorm Psychol 2015 Nov;124(4):803-816. [doi: 10.1037/abn0000089] [Medline: 26595469]
34. Benjet C, Albor YC, Bocanegra ES, et al. Incidence and recurrence of depression from adolescence to early adulthood: a longitudinal follow-up of the Mexican Adolescent Mental Health Survey. J Affect Disord 2020 Feb 15;263:540-546. [doi: 10.1016/j.jad.2019.11.010] [Medline: 31744746]
35. Human Development Index. Human development reports. United Nations. URL: https://hdr.undp.org/data-center/human-development-index [accessed 2025-05-27]
36. Santos ME. Measuring Multidimensional Poverty in Latin America: Previous Experience and the Way Forward: Oxford Poverty & Human Development Initiative; 2013.
37. Domoff SE, Banga CA, Borgen AL, et al. Use of passive sensing to quantify adolescent mobile device usage: feasibility, acceptability, and preliminary validation of the eMoodie application. Human Behav and Emerg Tech 2021 Jan;3(1):63-74 [FREE Full text] [doi: 10.1002/hbe2.247]
38. Gómez-Restrepo C, Diez-Canseco F, Brusco LI, et al. Mental distress among youths in low-income urban areas in South America. JAMA Netw Open 2025 Mar 3;8(3):e250122. [doi: 10.1001/jamanetworkopen.2025.0122] [Medline: 40042842]
39. Kroenke K, Strine TW, Spitzer RL, Williams JBW, Berry JT, Mokdad AH. The PHQ-8 as a measure of current depression in the general population. J Affect Disord 2009 Apr;114(1-3):163-173. [doi: 10.1016/j.jad.2008.06.026] [Medline: 18752852]
40. Kroenke K, Spitzer RL, Williams JBW, Monahan PO, Löwe B. Anxiety disorders in primary care: prevalence, impairment, comorbidity, and detection. Ann Intern Med 2007 Mar 6;146(5):317-325. [doi: 10.7326/0003-4819-146-5-200703060-00004] [Medline: 17339617]
41. Priebe S, Huxley P, Knight S, Evans S. Application and results of the Manchester Short Assessment of Quality of Life (MANSA). Int J Soc Psychiatry 1999;45(1):7-12. [doi: 10.1177/002076409904500102] [Medline: 10443245]
42. Edemekong PF, Bomgaars DL, Sukumaran S, Schoo C. Activities of daily living. In: StatPearls: StatPearls Publishing; 2024. [Medline: 29261878]
43. Badura P, Hamrik Z, Dierckens M, et al. After the bell: adolescents' organised leisure-time activities and well-being in the context of social and socioeconomic inequalities. J Epidemiol Community Health 2021 Jul;75(7):628-636. [doi: 10.1136/jech-2020-215319] [Medline: 33402394]
44. Mesurado B. Actividad estructurada vs. actividad desestructurada, realizadas en solitario vs. en compañía de otros y la experiencia óptima. Anales de Psicología 2009;25(2):308-315 [FREE Full text]
45. Caspi A, Moffitt TE, Newman DL, Silva PA. Behavioral observations at age 3 years predict adult psychiatric disorders. Longitudinal evidence from a birth cohort. Arch Gen Psychiatry 1996 Nov;53(11):1033-1039. [doi: 10.1001/archpsyc.1996.01830110071009] [Medline: 8911226]
46. Peeters F, Berkhof J, Delespaul P, Rottenberg J, Nicolson NA. Diurnal mood variation in major depressive disorder. Emotion 2006 Aug;6(3):383-391. [doi: 10.1037/1528-3542.6.3.383] [Medline: 16938080]
47. Thompson RJ, Mata J, Jaeggi SM, Buschkuehl M, Jonides J, Gotlib IH. The everyday emotional experience of adults with major depressive disorder: examining emotional instability, inertia, and reactivity. J Abnorm Psychol 2012 Nov;121(4):819-829. [doi: 10.1037/a0027978] [Medline: 22708886]
48. Bylsma LM, Taylor-Clift A, Rottenberg J. Emotional reactivity to daily events in major and minor depression. J Abnorm Psychol 2011 Feb;120(1):155-167. [doi: 10.1037/a0021662] [Medline: 21319928]

XSL•FO
RenderX

49. Lamers F, Swendsen J, Cui L, et al. Mood reactivity and affective dynamics in mood and anxiety disorders. J Abnorm Psychol 2018;127(7):659-669. [doi: 10.1037/abn0000378]
50. Nelson J, Klumparendt A, Doebler P, Ehring T. Everyday emotional dynamics in major depression. Emotion 2020;20(2):179-191. [doi: 10.1037/emo0000541]
51. Shin KE, Newman MG, Jacobson NC. Emotion network density is a potential clinical marker for anxiety and depression: comparison of ecological momentary assessment and daily diary. Br J Clin Psychol 2022 Jan;61 Suppl 1(Suppl 1):31-50. [doi: 10.1111/bjc.12295] [Medline: 33963538]
52. Pemberton R, Tyszkiewicz MDF. Factors contributing to depressive mood states in everyday life: a systematic review. J Affect Disord 2016 Aug;200:103-110. [doi: 10.1016/j.jad.2016.04.023] [Medline: 27131503]
53. Maciejewski DF, Keijsers L, van Lier PAC, Branje SJT, Meeus WHJ, Koot HM. Most fare well-but some do not: distinct profiles of mood variability development and their association with adjustment during adolescence. Dev Psychol 2019 Feb;55(2):434-448. [doi: 10.1037/dev0000650] [Medline: 30507219]
54. Asscheman JS, Zanolie K, Bexkens A, Bos MGN. Mood variability among early adolescents in times of social constraints: a daily diary study during the COVID-19 pandemic. Front Psychol 2021;12:722494. [doi: 10.3389/fpsyg.2021.722494] [Medline: 34504466]
55. Walsh RJ, van Buuren M, Hollarek M, Sijtsma H, Lee NC, Krabbendam L. Social contexts, momentary mood and affective variability in early adolescence: an exploratory ecological momentary assessment study. J Early Adolesc 2024 Jan;44(1):59-95. [doi: 10.1177/02724316231160147]
56. Flook L. Gender differences in adolescents' daily interpersonal events and well-being. Child Dev 2011;82(2):454-461. [doi: 10.1111/j.1467-8624.2010.01521.x] [Medline: 21410907]
57. Gordillo-León F, Mestas-Hernández L, Pérez-Nieto MA, et al. Diferencias de género en la valoración de la intensidad emocional de las expresiones faciales de alegría y tristeza. Escritos Psicología 2021 Jun;14(1):1-10. [doi: 10.24310/espsiescpsi.v14i1.12675]
58. Acosta FP, Clavero FH. Predictores de la felicidad y la inteligencia emocional en la educación secundaria. Rev colomb psicol 2018 Jan 1;27(1):71-84. [doi: 10.15446/rcp.v27n1.62705]
59. Larson RW, Moneta G, Richards MH, Wilson S. Continuity, stability, and change in daily emotional experience across adolescence. Child Dev 2002;73(4):1151-1165. [doi: 10.1111/1467-8624.00464] [Medline: 12146740]
60. Larson R, Csikszentmihalyi M, Graef R. Mood Variability and the Psycho-Social Adjustment of Adolescents Applications of Flow in Human Development and Education: Springer Netherlands; 2014:285-304. [doi: 10.1007/978-94-017-9094-9_15]
61. Acosta FP, Clavero FH. Ansiedad y felicidad como estados emocionales contrapuestos en la infancia. EJIHPE 2018;8(2):65. [doi: 10.30552/ejihpe.v8i2.245]
62. Orben A, Lucas RE, Fuhrmann D, Kievit RA. Trajectories of adolescent life satisfaction. R Soc Open Sci 2022 Aug;9(8):211808. [doi: 10.1098/rsos.211808] [Medline: 35937913]
63. Smith MD, Wesselbaum D. Well-being inequality among adolescents and young adults. J Adolesc Health 2024 Apr;74(4):703-712. [doi: 10.1016/j.jadohealth.2023.10.015] [Medline: 38069931]
64. Willroth EC, Atherton OE, Robins RW. Life satisfaction trajectories during adolescence and the transition to young adulthood: findings from a longitudinal study of Mexican-origin youth. J Pers Soc Psychol 2021 Jan;120(1):192-205. [doi: 10.1037/pspp0000294] [Medline: 32271086]
65. Marquez J, Long E. A global decline in adolescents' subjective well-being: a comparative study exploring patterns of change in the life satisfaction of 15-year-old students in 46 countries. Child Indic Res 2021 Jun;24(3):1251-1292. [doi: 10.1007/s12187-020-09788-8] [Medline: 34539933]
66. Maes IHL, Delespaul PAEG, Peters ML, et al. Measuring health-related quality of life by experiences: the experience sampling method. Value Health 2015 Jan;18(1):44-51. [doi: 10.1016/j.jval.2014.10.003] [Medline: 25595233]
67. Albertos A, Osorio A, Beltramo C, Albertos A, Osorio A, Beltramo C. Adolescentes y ocio: desarrollo positivo y transición hacia la vida adulta. Educ Educ 2020 Aug 4;23(2):201-210. [doi: 10.5294/edu.2020.23.2.3]
68. Spector WD, Katz S, Murphy JB, Fulton JP. The hierarchical relationship between activities of daily living and instrumental activities of daily living. J Chronic Dis 1987;40(6):481-489. [doi: 10.1016/0021-9681(87)90004-x] [Medline: 3597653]
69. Offer S. Family time activities and adolescents' emotional well‐being. J Marriage Fam 2013 Feb;75(1):26-41 [FREE Full text] [doi: 10.1111/j.1741-3737.2012.01025.x]
70. McHale SM, Blocklin MK, Walter KN, Davis KD, Almeida DM, Klein LC. The role of daily activities in youths' stress physiology. J Adolesc Health 2012 Dec;51(6):623-628. [doi: 10.1016/j.jadohealth.2012.03.016] [Medline: 23174474]
71. Fave AD, Bassi M. The quality of experience in adolescents' daily lives: developmental perspectives. Genet Soc Gen Psychol Monogr 2000 Aug;126(3):347-367. [Medline: 10950201]

Abbreviations

CREMAS: Adapted Strengthening the Reporting of Observational Studies in Epidemiology Checklist for Reporting Ecological Momentary Assessments Studies
ESM: experience sampling method

XSL•FO
RenderX

ESQ: Experience Sampling Questionnaire
GAD-7: General Anxiety Disorder-7
MANSA: Manchester Short Assessment of Quality of Life
OLA: building resilience and resources to reduce mental distress in young people in Latin America
PHQ-8: Patient Health Questionnaire-8
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by A Mavragani; submitted 23.08.24; peer-reviewed by K Aguiar; revised version received 24.02.25; accepted 25.02.25; published 19.06.25.

Please cite as:
Vilela-Estrada AL, Villarreal-Zegarra D, Mayo-Puchoc N, Holt N, Flórez-Varela Á, Fung C, Ariza-Salazar K, Carbonetti FL, Flores S, Carbonel A, Olivar N, Gomez-Restrepo C, Brusco LI, Priebe S, Diez-Canseco F
Fluctuations in Daily Happiness and Nervousness Based on Depressive and Anxious Symptoms in Adolescents or Young Adults Across 3 Latin American Cities: Experience Sampling Study
JMIR Form Res 2025;9:e65732
URL: https://formative.jmir.org/2025/1/e65732
doi:10.2196/65732

© Ana L Vilela-Estrada, David Villarreal-Zegarra, Nikol Mayo-Puchoc, Nicola Holt, Ángela Flórez-Varela, Catherine Fung, Karen Ariza-Salazar, Fernando Luis Carbonetti, Sumiko Flores, Adriana Carbonel, Natividad Olivar, Carlos Gomez-Restrepo, Luis Ignacio Brusco, Stefan Priebe, Francisco Diez-Canseco. Originally published in JMIR Formative Research (https://formative.jmir.org), 19.6.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Association of Technology-Related Skills and Self-Efficacy With Willingness to Participate in Heart Failure Telemonitoring: Cross-Sectional Observational Study

Sharon Cuppen[1,2*], MSc; Mayke van Leunen[3,4*], MD; Tamara Henken[1], MSc; Mayra Goevaerts[3,4], MSc; Martijn Scherrenberg[5,6], MD, PhD; Maarten Falter[5,6], MD, PhD; Paul Dendale[5,6], MD, PhD, Prof Dr Med; Hareld Kemps[3,4], MD, PhD, Prof Dr Med; Willem J Kop[2], MSc, PhD, Prof Dr

[1]Department of Medical Psychology, Máxima Medisch Centrum, Veldhoven, The Netherlands
[2]Department of Medical and Clinical Psychology, Tilburg University, Tilburg, The Netherlands
[3]Department of Industrial Design, Eindhoven University of Technology, Eindhoven, The Netherlands
[4]Department of Cardiology, Máxima Medisch Centrum, De Run 4600, Veldhoven, The Netherlands
[5]Faculty of Medicine and Life Sciences, Hasselt University, Hasselt, Belgium
[6]Department of Cardiology, Jessa Hospital, Hasselt, Belgium
[*]these authors contributed equally

Corresponding Author:
Mayke van Leunen, MD
Department of Industrial Design, Eindhoven University of Technology, Eindhoven, The Netherlands

Abstract

Background: The adoption of telemonitoring in patients with heart failure (HF) is influenced by technology-related skills and self-efficacy, as well as psychological, clinical, and demographic factors. However, the relative importance of these factors with regard to willingness to use telemonitoring is insufficiently understood.

Objectives: This cross-sectional observational study examines the extent to which technology-related skills and self-efficacy are related to willingness to participate in telemonitoring in patients with HF.

Methods: Patients completed questionnaires during hospitalization. Associations of technological skills and self-efficacy with willingness to participate in telemonitoring (dichotomous and continuous scale) were examined using regression models. Mediation-moderation analyses were used to investigate the role of self-efficacy in the association of technological skills with willingness to participate.

Results: This study recruited 61 patients admitted for decompensated HF (mean age 79.9, SD 9.5 years; 24 women). Higher levels of technological skills were associated with higher willingness to participate in telemonitoring (odds ratio [OR] 1.073 per scale unit, 95% CI 1.031-1.117). Technological self-efficacy and learnability were also related to willingness to participate (OR 1.141, 95% CI 1.039-1.252; OR 1.029, 95% CI 1.006-1.052) but did not mediate the association of technological skills with willingness to participate in telemonitoring. Psychological factors (anxiety, depressive symptoms, and perceived social support), age, and cognitive and physical functioning did not moderate the association of technological skills with willingness to participate in telemonitoring.

Conclusions: Technological skills, self-efficacy, and learnability are interrelated factors that need to be considered in patients with HF who are eligible for telemonitoring. Future intervention studies that target these factors could increase patients' willingness and competence in using telemonitoring after admission for HF.

(JMIR Form Res 2025;9:e68992) doi:10.2196/68992

KEYWORDS

heart failure; telemonitoring; participating in telemonitoring; technological skills; technological self-efficacy; technological learnability

XSL•FO
RenderX

Introduction

Background

Heart failure (HF) is a global pandemic leading to poor quality of life and high mortality and morbidity rates, affecting approximately 64.3 million people worldwide [1]. As a result of global population growth, ageing, and improved survival after diagnosis, the prevalence of HF is further increasing leading to a considerable and growing burden on health care costs [2-4]. Innovative strategies are needed to counteract the effects of HF as a considerable socioeconomic and medical burden, especially to lower hospitalization as they account for 50% of total costs and lead to further physical and physiological deconditioning [4,5].

Theoretical Framework

eHealth interventions reduce mortality and readmission rates in patients with HF, while improving their medication adherence and self-care behavior [6]. Telemonitoring is a specific application of eHealth that has the potential to reduce hospitalizations, mortality, and health care costs [7]. Health care professionals use telemonitoring to closely review patient-generated health data to detect early signs of HF progression [7]. As reported in a recent meta-analysis, optimal home telemonitoring reduces HF-related hospitalizations by 15% - 19% and all-cause mortality by 16% [8]. Despite these promising outcomes of telemonitoring, its use is not optimal and requires improvements [9].

Important barriers in using telemonitoring are limited technological skills and low technology-related self-efficacy as well as cognitive and physical limitations that are common in patients with HF [10,11]. Cognitive limitations (eg, memory, concentration, and executive functioning) and physical or functional limitations (eg, impaired vision or hearing and presence of hand tremor) are common in older adults and can interfere with using telemonitoring, since these impairments can affect one's technological capabilities [10-13]. Lack of computer capabilities and insufficient technology onboard training negatively influence the use of technological interventions [14]. In addition, low digital self-efficacy is reported for patients with HF, which might be related to low learnability of digital skills and decreased acceptance and willingness to participate in telemonitoring programs [10]. Furthermore, higher age, female sex, and lower educational level are associated with lower acceptance of telemonitoring [10,15,16]. Evidence also suggests that older adult women are less likely to adopt digital devices than older adult men [17].

For telemonitoring to succeed, patients need to engage in self-management, which requires self-efficacy. Self-efficacy is based on the Social Cognitive Theory of Bandura and refers to the confidence an individual has in his or her own ability to perform a specific behavior [18]. Prior research has shown that experience with computers and telemonitoring devices positively influences the intention to use technological health applications. This relationship is partly and positively mediated by an individual's perceived self-efficacy [19]. There is an inverse association between self-efficacy and several psychological factors, particularly anxiety and depression. A considerable proportion of patients with HF have anxiety and depressive symptoms (with recent prevalence estimates of 30% and 21.5%, respectively) [20,21]. These psychological factors are also associated with poor health outcomes and higher rehospitalization and mortality rates [22]. The negative association between anxiety and depression with self-efficacy might interfere with telemonitoring participation and thereby adversely influence health outcomes.

In contrast to the potential adverse effects of anxiety and depressive symptoms, social support can improve a patient's self-efficacy [23,24]. As HF is more common among older adults and age is directly related to less technological skills [25], it seems plausible that higher levels of social support would enhance a patients' willingness to participate in telemonitoring, particularly in case of limited technological skills or self-efficacy. This perspective is supported by the finding that low social support is a barrier in effectively using medical technology and medical systems [10]. Self-efficacy is a significant mediator in the relationship between social support and self-management behaviors (including treatment adherence) in individuals with HF [23,24].

Objectives

Based on this background, this study investigated the association of technological skills with technological self-efficacy and learnability in the context of willingness to participate in telemonitoring among patients with HF. Second, the role of psychological factors (anxiety, depressive symptoms, and social support) and demographic as well as functional factors in these associations was explored.

Methods

Design and Participants

This study is a single-center cross-sectional observational study among adults with HF who were admitted to the cardiology department for acute decompensated heart failure (ADHF) at Máxima Medical Centre in Veldhoven, the Netherlands. Participants were enrolled from the cardiology department between January and May 2023. Inclusion criteria were a diagnosis of HF according to the European Society of Cardiology guideline definition [26], age of 18 years and older, admission for ADHF, and proficiency in the Dutch language. Patients with a psychiatric disorder or major cognitive impairment (eg, dementia) were excluded.

In total, 119 eligible patients were asked for study participation of whom 61 (51%) gave informed consent. Reasons for nonparticipation (N=58) were not interested in study participation (44, 76%), expected study participation to be too demanding (5, 9%), difficulties understanding the purpose of the study (3, 5%), or death before completing the questionnaires (6, 10%). A graphical abstract of the study can be found in Multimedia Appendix 1.

Ethical Considerations

This study was evaluated by the medical ethics review board of Máxima Medical Centre and received expedited review, given the nature of the assessments and demands on the participating

XSL•FO
RenderX

patients (protocol no. N23.007). Written informed consent was provided to each consecutive patients before inclusion in the study. The study was conducted in accordance to the Declaration of Helsinki. Data were entered in an electronic database using deidentified code numbers. The participants did not receive any form of financial compensation for participating in this trial.

Procedure

Consecutive patients admitted to the cardiology department meeting the eligibility criteria were asked for study participation by the research team (physician or second-year medical psychology master student). All patients with ADHF who were interested in completing the questionnaires regardless of actual telemonitoring participation were enrolled after signing the informed consent document. Participants received a brief written background description on telemonitoring explaining that participation includes daily measuring of blood pressure, heart rate, and body weight. Participants were informed that these measures, along with answers on a brief daily questionnaire on HF-related complaints, are forwarded to the health care professional using an online website in order to enable proper monitoring and, therefore, better cardiovascular care. Participants completed the study questionnaires on paper to avoid bias in participation and responses relevant to technological skills. Data received from the questionnaires were entered in an electronic database using deidentified code numbers.

Measures

The "predictor" measures were technological skills, technological self-efficacy, and learnability. The primary outcome measure was willingness to participate in telemonitoring (assessed dichotomously and continuously). We also documented whether or not patients actually participated in telemonitoring after discharge. Additional predictor measures were psychological factors (anxiety, depressive symptoms, and perceived social support) and background factors (demographics, clinical measures, and cognitive and physical functioning).

Technological Skills and Learnability

The Digital Health Readiness Questionnaire (DHRQ; Dutch version) was used to investigate technological skills, digital literacy, and learnability [27]. This self-report questionnaire, developed at Hasselt University, consists of 20 questions rated on a 5-point Likert scale, with 1 indicating "totally disagree" to 5 indicating "totally agree." The questionnaire includes five subdomains: (1) digital access, (2) use of digital technology, (3) digital literacy, (4) digital health literacy, and (5) digital learnability. Subdomains 1-4 (15 items; total scores ranging from 15 to 75) represent technology-related skills ($DHRQ_{Skills}$), described as "digital readiness" by Scherrenberg et al [27]. Subdomain 5 (5 items; total scores ranging from 5 to 25) represents learnability ($DHRQ_{Learnability}$) [27]. Higher scores indicate higher levels of technological skills or learnability. The psychometric properties of the DHRQ are good with acceptable internal consistency (Cronbach α>0.70 for all subdomains) [27]. In this study, the internal consistency of the DHRQ total score was good (Cronbach α=0.96).

Technological Self-Efficacy

The modified Computer Self-Efficacy Scale was used to measure technological self-efficacy. This self-report questionnaire has 10 statements that respondents answer based on a 10-point Likert scale, from 1 referring to "not at all confident" to 10 referring to "completely confident" [28]. Total scores range from 10 to 100, with higher scores indicating higher levels of computer-related self-efficacy. Within patients in a clinical rehabilitation setting, a Cronbach α value of 0.94 has been found, indicating a high internal consistency [28].

Willingness to Participate in Telemonitoring

Assessment of willingness to participate in telemonitoring was measured with a dichotomous yes or no question as the primary outcome and a continuous measure as a secondary outcome. In addition to reporting whether or not they would participate in telemonitoring (dichotomous outcome), participants were also asked the following question: "Based on a scale from 1 (not at all) to 10 (very likely), how likely is it that you would participate in telemonitoring if you had sufficient technological skills and/or help from friends or family with using the technology?" (continuous outcome).

Psychological Measures

The Generalized Anxiety Disorder-7 (GAD-7) Scale was used to provide information regarding the presence and degree of anxiety symptoms [29]. The Patient Health Questionnaire-9 (PHQ-9) was used to assess the presence and degree of depressive symptoms [30]. Participants were asked to rate symptoms on a 4-point Likert scale, with 0 indicating "not at all" to 3 "nearly every day" for both GAD-7 and PHQ-9. Total scores of answers range from 0 to 21 for the GAD-7 and 0 to 27 for the PHQ-9. Higher scores indicate a higher level of symptoms, with a cutoff score of ≥10 for both scales to indicate levels that are likely to meet diagnostic criteria for anxiety and depression. The GAD-7 and PHQ-9 are recommended for psychological assessment of cardiovascular outpatients [31]. Research has demonstrated good internal consistency (Cronbach α=0.82 and 0.83, respectively) [22,32,33].

The Multidimensional Scale of Perceived Social Support (MSPSS) was used to measure perceived social support. This questionnaire consists of 12 items with 4 items representing each of the 3 subdomains: support from family, friends, or significant others, respectively [34,35]. Questions are based on a 7-point Likert scale, from 1 "very strongly disagree" to 7 "very strongly agree." Total scores range from 12 to 84, with higher scores indicating more perceived social support. The MSPSS has good psychometric properties with a Cronbach α=0.94 for the total instrument [34].

Background Variables

Sociodemographic characteristics regarding age, sex, and educational level were obtained with a self-report questionnaire designed for the purposes of this study. Clinical information relevant to HF was obtained from the medical record, including left ventricular ejection fraction (LVEF).

The self-report Cognitive Failure Questionnaire (CFQ) was used to provide information regarding cognitive functioning,

XSL•FO
RenderX

with 25 items measuring the frequency of everyday cognitive mistakes [36,37]. Items are rated on a 5-point Likert scale, with 0 indicating "never" to 4 indicating "very often." Total scores range from 0 to 100, with higher scores indicating a higher level of cognitive problems. The CFQ has good psychometric properties within the general adult population (Cronbach α=0.89) [34].

The Computer-related Physical Functioning Questionnaire, developed for the purposes of this study, was used to obtain information about 3 domains of physical functioning: vision, hearing, and presence of hand tremors. This self-report inventory includes 6 items, 2 items for each domain measured. Items are rated on a 4-point Likert scale, with 0 indicating "not at all" to 3 indicating "almost all the time" (eg, "I experience problems with my eyesight," and "I am limited in using the computer because of motor problems in my hands."). In this sample, the Cronbach α value was 0.51. Total scores can range from 0 to 18, with higher scores indicating more physical limitations.

Statistical Analyses

Bivariate associations between the variables investigated in this study were examined using Pearson correlations and independent sample 2-tailed *t* tests. To investigate the association of "technological skills" with willingness to participate in telemonitoring (dichotomous outcome), logistic regression analyses were used. Multiple logistic regression analysis was used to adjust for technological self-efficacy and learnability, as well as for demographic (age, sex, and education level) and function-related (cognitive functioning and physical limitations) variables. Results for the continuous outcome of willingness to participate in telemonitoring were analyzed using (multiple) linear regression models and presented as standardized regression coefficients (β). The overall fit of the linear regression models was examined using the total R^2. Participants with missing values were excluded from analyses using a pairwise approach for bivariate models and listwise for multivariable models. Relevant assumptions for logistic and linear regression were evaluated and met prior to conducting the analyses.

To investigate whether technological self-efficacy and learnability played a mediating role in the association between technological skills and willingness to participate in telemonitoring, the PROCESS tool of the Statistical Package for the Social Sciences (SPSS) was used. The overall fit was examined using Nagelkerke R^2 for analyses with the dichotomous outcome variable and R^2 for analyses with the continuous outcome variable. Associations were evaluated using a type I error (2-sided α value) of .05 or a 95% CI. All statistical analyses were conducted using the SPSS software (version 28; IBM Corp).

Statistical Power and Sample Size

The sample size of this study was based on medium to large effect sizes that are likely to be clinically relevant, as no prior research has investigated the research question addressed in this study. The sample of 61 participants enables detection of a bivariate correlation of 0.35 and an OR of 2.2 (assuming that 50% of the participants will have high levels of technological skills), with a power of 0.80 and a 2-sided α value of .05.

Results

Participant Characteristics

The mean age of the sample (N=61) was 79.9 (SD 9.5) years (range: 49-96 years) and 37 (61%) were male (Table 1). Approximately 48% (29/61) had completed an education beyond high school. The mean LVEF was 40 (SD 16%), with 24 patients (39%) being classified as HF with reduced ejection fraction. Four participants (7%) had already participated in telemonitoring prior to study participation.

XSL•FO
RenderX

Table . Descriptive statistics of patients with heart failure during hospital admission (N=61).

Characteristics	Values
Demographic characteristics	
Age (years), mean (SD)	79.9 (9.5)
Sex, male, n (%)	37 (61)
Education, n (%)	
Elementary or high school	31 (51)
Secondary vocational education	14 (23)
Higher vocational education	9 (15)
College or university education	6 (10)
Clinical characteristics	
LVEF[a], mean (SD)	40.3 (15.6)
HF[b] classification, n (%)	
LVEF ≥50%	28 (46)
LVEF 41% - 49%	8 (13)
LVEF ≤40%	24 (39)
BMI (kg/m^2), at discharge, mean (SD)	25.7 (4.9)
History of cerebrovascular accident, n (%)	11 (18)
History of myocardial infarction, n (%)	12 (19)
Percutaneous coronary intervention, n (%)	16 (25)
Coronary artery bypass grafting, n (%)	3 (5)
Peripheral artery disease, n (%)	14 (22)
Arterial fibrillation, n (%)	31 (48)
Diabetes mellitus, n (%)	17 (26)
Hypertension, n (%)	37 (57)
Hypercholesterolemia, n (%)	13 (20)
Psychological and technology-related measure	
Cognitive functioning (CFQ[c]), mean (SD)	28.8 (17.7)
Physical limitations (CPFQ[d]), mean (SD)	4.5 (3.3)
Anxiety (GAD-7[e]), mean (SD)	5.9 (5.1)
Depressive symptoms (PHQ-9[f]), mean (SD)	8.7 (5.6)
Social support (MSPSS[g]), mean (SD)	67.1 (14.6)
Technological skills (DHRQ[h]$_{Skills}$), mean (SD)	38.4 (17.7)
Technological learnability (DHRQ$_{Learnability}$), mean (SD)	14.4 (6.8)
Technological self-efficacy (mCSES[i]), mean (SD)	53.1 (26.4)
Willingness to participate in TM[j], yes, n (%)	40 (66)
Willingness to participate in TM with sufficient technological skills or help, mean (SD)	7.1 (2.5)
Started with TM after hospital discharge, n (%)	26 (44)

[a]LVEF: left ventricular ejection fraction.

[b]HF: heart failure.

[c]CFQ: Cognitive Failure Questionnaire.

XSL•FO
RenderX

[d]CPFQ: Computer-related Physical Functioning Questionnaire.

[e]GAD-7: Generalized Anxiety Disorder-7.

[f]PHQ-9: Patient Health Questionnaire-9.

[g]MSPSS: Multidimensional Scale of Perceived Social Support.

[h]DHRQ: Digital Health Readiness Questionnaire.

[i]mCSES: modified Computer Self-Efficacy Scale.

[j]TM: telemonitoring.

Physical and cognitive functioning were within the normative range, mean Computer-related Physical Functioning Questionnaire 4.5 (SD 3.3) and mean CFQ 28.8 (SD 17.7). Anxiety and depression were above clinical cutoff values (score ≥10) for 14 (23%) and 23 (38%) participants, respectively.

Measures and Statistical Analyses

Willingness to Participate in Telemonitoring

In total, 66% (40/61) of participants were willing to participate in telemonitoring. Patients who affirmed willingness to participate in telemonitoring (N=40) also scored higher on the continuous measure of willingness to participate than those who were not willing to participate (mean 7.9, SD 1.9 vs mean 5.5, SD 2.9; *P*=.002), indicating that both indices assessed the same outcome measure.

Of those 40, a total of 5 were not eligible for telemonitoring after discharge (1 patient moved to another region, 1 died before hospital discharge, and 3 were discharged to a hospice). Of the remaining 35 patients who mentioned wanting to participate in telemonitoring, 24 (69%) actually started with telemonitoring at discharge. In addition, another 10% (2/21) of patients who initially stated that they were not interested in telemonitoring ended up starting with telemonitoring at discharge (total actual participation was 26/61, 43% of the total sample).

Technological Skills and Willingness to Participate in Telemonitoring (Unadjusted Analyses)

Higher levels of technological skills were associated with a higher likelihood of willingness to participate in telemonitoring (OR 1.073, 95% CI 1.031-1.117, per $DHRQ_{Skills}$ scale unit) (Table 2). Patients with an above-mean score (≥39) on the $DHRQ_{Skills}$ (32/61, 53%) were approximately 10 times more likely to be willing to participate in telemonitoring than patients with scores below the mean (OR 9.917, 95% CI 2.752-35.740).

Table . Logistic regression analyses on willingness to participate in telemonitoring in patients with heart failure measured dichotomously during hospital admission.

Variable	Unadjusted analysis			Adjusted analysis[a]		
	OR[b]	95% CI	*P* value	OR	95% CI	*P* value
Technological skills	1.073	1.031-1.117	.001	1.121	1.035-1.215	.01
Age (years)	0.910	0.843-0.983	.02	0.907	0.808-1.018	.10
Sex (0=male; 1=female)	1.083	0.366-3.204	.89	1.084	0.184-6.396	.93
Education level[c]	1.110	0.663-1.858	.69	2.878	1.091-7.591	.03
Physical functioning	0.932	0.788-1.103	.41	1.074	0.819-1.409	.61
Cognitive functioning	1.013	0.981-1.046	.43	0.998	0.947-1.053	.95
Technological learnability	1.141	1.039-1.252	.01	1.171	0.907-1.512	.23
Technological self-efficacy	1.029	1.006-1.052	.01	0.952	0.884-1.024	.19

[a]Variables are adjusted for each other.

[b]OR: odds ratio.

[c]Ranging from college or university education to elementary or high school.

Patients who had higher scores on technological self-efficacy were more likely to be willing to participate in telemonitoring (OR 1.029, 95% CI 1.006-1.052, per modified Computer Self-Efficacy Scale unit). In addition, younger age (OR 0.910, 95% CI 0.843-0.983, per year), more social support (OR 1.045, 95% CI 1.003-1.089, per MSPSS scale unit), and technological learnability (OR 1.141, 95% CI 1.039-1.252, per $DHRQ_{Learnability}$ scale unit) were associated with patients affirming willingness to participate in telemonitoring.

Multivariate analyses (Table 2) showed that the association between technological skills and willingness to participate remained significant when adjusting for age, sex, education level, cognitive functioning, physical limitations, technological

XSL•FO
RenderX

learnability, and technological self-efficacy (OR 1.121, 95% CI 1.035-1.215). Additionally, higher education was indicated as a significant predictor in the multivariate model (Table 2, right columns).

When repeating the analyses using willingness to participate as a continuous outcome variable, a similar pattern of results was found. Higher levels of technological skills were associated with higher levels of the continuous measure of willingness to participate (unadjusted r=0.384, P=.003). The bivariate correlations between willingness to participate (measured as continuous variable) with other variables are displayed in Table S1 in Multimedia Appendix 2. In addition to technological skills, technological self-efficacy (r=0.307; P=.018), social support (r=0.507; P<.001), and technological learnability (r=0.433; P=.001) were associated with the continuous measure of willingness to participate. Older age was correlated with lower levels of technological skills (r=−0.465; P<.001), lower technological self-efficacy (r=−0.484; P<.001), and less learnability (r=−0.334; P=.009) (Table S1 in Multimedia Appendix 2). The association of technological skills with the continuous measure of willingness to participate remained significant when adjusting for covariates (β=.392; P=.044, overall model R^2=0.326) (see Table S2 in Multimedia Appendix 2 for details of the multiple linear regression analysis).

Mediating Role of Technological Self-Efficacy and Learnability in Willingness to Participate in Telemonitoring

Higher levels of technological skills were strongly and positively correlated with higher levels of technological self-efficacy (r=0.74; P<.001) and learnability (r=0.69; P<.001) (Table S1 in Multimedia Appendix 2), 2 constructs that were also both associated to willingness to participate (OR 1.029, 95% CI 1.006-1.052; OR 1.141, 95% CI 1.039-1.252). Table 2 and Table S2 in Multimedia Appendix 2 show that adjusting for these variables did not result in an attenuation of the association between technological skills with willingness to participate in telemonitoring.

To further explore the interplay of pathways involved in participating in telemonitoring, mediation analyses were conducted. These analyses indicated that the positive association between technological skills and willingness to participate was not mediated by technological self-efficacy (B=−0.006, 95% CI −0.078 to 0.035) or learnability (B=0.008, 95% CI −0.035 to 0.047]) (Table 3, left part). Results for the continuous measure of participation also revealed no mediation of technological self-efficacy (B=0.007, 95% CI −0.028 to 0.049) and learnability (B=0.031, 95% CI −0.001 to 0.065) in the association between technological skills and willingness to participate in telemonitoring (Table 3, right part).

XSL•FO
RenderX

Table . The mediating role of technological self-efficacy and learnability, and the moderating role of anxiety, depressive symptoms, and social support in willingness to participate in telemonitoring in patients with heart failure during hospital admission[a].

	Willingness to participate							
	Dichotomous				Continuous			
	B [b]	95% CI	*P* value	Nagelkerke R^2[c]	*B*	95% CI	*P* value	R^2
Mediation effects								
Direct effect of TS[d] on Y[e]	0.076	0.021 to 0.131	.01		0.049	−0.002 to 0.101	.06	
Indirect effect of TS on Y (mediation of technological SE[f])	−0.006	−0.078 to 0.035			0.007	−0.028 to 0.049		
Total model			<.001	.316			.01	.150
Direct effect of TS on Y	0.063	0.013 to 0.113	.01		0.025	−0.022 to 0.072	.29	
Indirect effect of TS on Y (mediation of learnability)	0.008	−0.035 to 0.047			0.031	−0.001 to 0.065		
Total model			<.001	.319			.002	.203
Moderation effects								
TS	0.037	−0.036 to 0.110	.32		0.089	0.020 to 0.157	.01	
Depression	−0.112	−0.365 to 0.141	.39		0.212	−0.055 to 0.479	.12	
TS[g] depression			.35				.35	
Total model			.002	.314			.01	.427
TS	0.081	0.014 to 0.148	.02		0.050	−0.006 to 0.106	.08	
Anxiety	0.062	−0.206 to 0.330	.65		−0.061	−0.344 to 0.222	.67	
TS[g] anxiety			.76				.93	
Total model			.001	.319			.02	.396
TS	−0.089	−0.275 to 0.097	.35		0.058	−0.088 to 0.203	.43	
Social support	−0.030	−0.118 to 0.058	.50		0.084	0.009 to 0.158	.03	
TS[g] Social support			.10				.82	
Total model			<.001	.409			<.001	.580

[a]Moderation models were conducted for each of the 3 psychological variables separately.

[b]Unstandardized coefficients.

[c]Residuals squared.

XSL•FO
RenderX

[d]TS: technological skills.

[e]Y: willingness to participate in telemonitoring.

[f]SE: self-efficacy.

[g]Interaction effect.

Moderating Role of Anxiety, Depressive Symptoms, and Social Support in Willingness to Participate in Telemonitoring

Table 3 (bottom part) indicates that there was no significant moderation (ie, interaction effect) of the variables anxiety (P=.76; P=.93), depression (P=.35; P=.35), and social support (P=.10; P=.82) with technological skills on the dichotomous and continuous outcome measures of willingness to participate in telemonitoring. There were also no significant main effects of these psychological factors, except for social support on the continuous outcome measure (Table 3, bottom part).

Discussion

Summary of the Main Findings

This real-life cohort study shows that the level of technological skills is a significant, independent factor in a patient's willingness to engage in telemonitoring for HF. This association remained significant when taking covariates and moderating factors into account. In addition to technological skills, technological self-efficacy and learnability were found to be important factors in willingness to participate in telemonitoring. Higher technological self-efficacy and learnability levels were associated with both technological skills and willingness to participate in telemonitoring but did not mediate the link between technological skills and willingness to participate. These findings indicate that improving technological skills might help patients with HF to actively adopt remote monitoring options for their clinical care. However, self-efficacy and learnability are additional important target constructs as well since they are strongly and positively associated with technological skills. This study sets the stage for future investigations in which technological self-efficacy and other factors associated with technological skills are addressed with the long-term goal to optimize patient participation in telemonitoring and other forms of telemedicine.

Consistent with previous research, this study indicates that a lack of technological skills is a barrier in participating in telemonitoring among patients with HF [10,11,14,38]. Patients with high technological skills were approximately 10 times more likely to be willing to participate in telemonitoring than patients with low technological skills. These findings suggest that technological skills, self-efficacy, and learnability are strongly and positively intercorrelated, and that these 3 constructs are positively associated with willingness to participate in telemonitoring. In contrast to previous studies, no significant associations were found between age and cognitive dysfunction with willingness to participate in telemonitoring. These discrepancies require further investigations with larger and more heterogeneous samples in terms of education and cultural background.

This study also confirms that technological self-efficacy and learnability play a role in the adoption of health technology [19], but these factors did not mediate the relationship between technological skills and willingness to participate in telemonitoring. This finding might be explained by the high correlation of technological skills with both technological self-efficacy and learnability. Specifically, if one (primary) predictor variable is strongly associated with an outcome measure (in this case technological skills with willingness to participate in telemonitoring), then there is little additional variance left to be explained by other predictors, particularly when these other factors are also highly correlated with the primary predictor. It is also possible that technological self-efficacy and learnability are primary factors in willingness to participate in telemonitoring and that this association is mediated by technological skills, which is the reverse pathway as was investigated in this study.

Limitations and Strengths

The study has limitations that need to be considered when interpreting the findings. Self-report questionnaires are potentially influenced by socially desirable answers and by persons' mental state at the time of completion [39]. This study found that approximately two-thirds of participants who mentioned to be willing to participate actually started telemonitoring. It is not known whether this discrepancy reflects patient, health care, or program-related factors (eg, access to the technology, preenrollment support, or system requirements). Also, exclusively Dutch-speaking participants were included, and therefore results cannot be generalized to other (non-Western) cultures [40]. In addition, given the diverse applications of telemonitoring across various levels, our findings are specific to the methodology used in this sample and cannot be extrapolated to all telemonitoring practices (eg, noninvasive vs invasive). A minimal selection bias, moreover, could not be avoided despite mentioning that study participation did not affect the actual participation. Participations with certain characteristics may be more open to participate in research. As a consequence of the limited sample size and the complex statistical models used in this study, exclusively medium to large effect sizes could be detected with sufficient statistical power. Among the strengths of the study are the assessment of a wide range of relevant covariates, including cognitive functioning, and the fact that the sample of the study is comparable with the typical HF population (eg, high age and representative LVEF groups) [41,42].

Recommendations for Future Research

Future longitudinal and intervention research is needed to better understand the association of technological skills, self-efficacy, and learnability with willingness to participate in telemonitoring and to obtain more robust and generalizable results. Investigations using larger samples and a broader range of participants in terms of demographic characteristics are needed

XSL•FO
RenderX

to increase generalizability of the present study findings. It would also be useful to include "privacy concerns" in future analyses, as this factor might influence the use and willingness to participate in telemonitoring [40,43]. Additionally, future studies are needed to further examine the gap between willingness to participate and actually participating (ie, the intention-behavior gap) and factors influencing the participation duration [44]. The intention-behavior gap in this study was as follows: two-thirds of the patients expressing willingness actually participated and one-third did not, and approximately one-tenth of the patients who initially expressed no willingness ended up participating in telemonitoring. These numbers are consistent with previous research [45]. Future longitudinal studies are also needed to further examine the optimal time sequence of the intervention targeting technological skills, self-efficacy, and learnability on telemonitoring participation. A step-wise approach could potentially be an effective method to train patients in using telemonitoring devices (eg, starting with measuring weight and gradually expanding tasks).

Conclusions

This study shows that higher levels of technological skills, self-efficacy, and learnability are associated with a higher likelihood of willingness to participate in telemonitoring among patients with HF. Therefore, clinical practice can be improved by screening for the degree of technological skills (eg, with the DHRQ or other assessment tools) to identify patients who need onboard training and to further increase participation and efficacy of telemonitoring. It will be important to reduce the gap between willingness to participate and actual participation, and to develop interventions targeting higher adoption of HF telemonitoring.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

SC, MvL, HK, and WK contributed to the concept and design of the study. SC and MvL screened for eligible patients, obtained informed consent, collected the data, performed the data analysis supervised by WK, and drafted the manuscript. MG made the graphical abstract. All authors revised the manuscript. All gave final approval and agreed to be accountable for all aspects of work ensuring integrity and accuracy.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Graphical abstract.
[PNG File, 272 KB - formative_v9i1e68992_app1.png]

Multimedia Appendix 2
Supplementary information on the correlations observed with willingness to participate and other variables, as well as the relationship between technological skills and willingness to participate in telemonitoring.
[DOCX File, 17 KB - formative_v9i1e68992_app2.docx]

References

1. James SL, Abate D, Abate KH. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet 2018 Nov 10;392(10159):1789-1858. [doi: 10.1016/S0140-6736(18)32279-7] [Medline: 30496104]
2. Khan MS, Shahid I, Bennis A, Rakisheva A, Metra M, Butler J. Global epidemiology of heart failure. Nat Rev Cardiol 2024 Oct;21(10):717-734. [doi: 10.1038/s41569-024-01046-6] [Medline: 38926611]
3. Roth GA, Huffman MD, Moran AE, et al. Global and regional patterns in cardiovascular mortality from 1990 to 2013. Circulation 2015 Oct 27;132(17):1667-1678. [doi: 10.1161/CIRCULATIONAHA.114.008720] [Medline: 26503749]
4. Lesyuk W, Kriza C, Kolominsky-Rabas P. Cost-of-illness studies in heart failure: a systematic review 2004-2016. BMC Cardiovasc Disord 2018 May 2;18(1):74. [doi: 10.1186/s12872-018-0815-3] [Medline: 29716540]
5. Pandey A, Gilbert O, Kitzman DW. Physical frailty in older patients with acute heart failure: from risk marker to modifiable treatment target. J Am Geriatr Soc 2021 Sep;69(9):2451-2454. [doi: 10.1111/jgs.17306] [Medline: 34146340]
6. Liu S, Li J, Wan DY, et al. Effectiveness of eHealth self-management interventions in patients with heart failure: systematic review and meta-analysis. J Med Internet Res 2022 Sep 26;24(9):e38697. [doi: 10.2196/38697] [Medline: 36155484]
7. Brahmbhatt DH, Cowie MR. Remote management of heart failure: an overview of telemonitoring technologies. Card Fail Rev 2019 May;5(2):86-92. [doi: 10.15420/cfr.2019.5.3] [Medline: 31179018]
8. Scholte NTB, Gürgöze MT, Aydin D, et al. Telemonitoring for heart failure: a meta-analysis. Eur Heart J 2023 Aug 14;44(31):2911-2926. [doi: 10.1093/eurheartj/ehad280] [Medline: 37216272]

XSL•FO
RenderX

9. Hashemi A, Nourbakhsh S, Tehrani P, Karimi A. Remote telemonitoring of cardiovascular patients: benefits, barriers, new suggestions. ARTRES 2018;22(C):57. [doi: 10.1016/j.artres.2018.04.001]
10. Masterson Creber R, Dodson JA, Bidwell J, et al. Telehealth and health equity in older adults with heart failure: a scientific statement from the American Heart Association. Circ Cardiovasc Qual Outcomes 2023 Nov;16(11):37909212. [doi: 10.1161/HCQ.0000000000000123] [Medline: 37909212]
11. Boyne JJJ, Vrijhoef HJM. Implementing telemonitoring in heart failure care: barriers from the perspectives of patients, healthcare professionals and healthcare organizations. Curr Heart Fail Rep 2013 Sep;10(3):254-261. [doi: 10.1007/s11897-013-0140-1] [Medline: 23666901]
12. Choi NG, Dinitto DM. The digital divide among low-income homebound older adults: internet use patterns, eHealth literacy, and attitudes toward computer/internet use. J Med Internet Res 2013 May 2;15(5):e93. [doi: 10.2196/jmir.2645] [Medline: 23639979]
13. Yazdani-Darki M, Rahemi Z, Adib-Hajbaghery M, Izadi F. Older adults' barriers to use technology in daily life: a qualitative study. Nurs Midwifery Stud 2020;9(4):229-236. [doi: 10.4103/nms.nms_91_19]
14. Stoop AP, van't Riet A, Berg M. Using information technology for patient education: realizing surplus value? Patient Educ Couns 2004 Aug;54(2):187-195. [doi: 10.1016/S0738-3991(03)00211-8] [Medline: 15288913]
15. Czaja SJ, Charness N, Fisk AD, et al. Factors predicting the use of technology: findings from the center for research and education on aging and technology enhancement (CREATE). Psychol Aging 2006 Jun;21(2):333-352. [doi: 10.1037/0882-7974.21.2.333] [Medline: 16768579]
16. Auener SL, van Dulmen SA, Atsma F, et al. Characteristics associated with telemonitoring use among patients with chronic heart failure: retrospective cohort study. J Med Internet Res 2023 Oct 18;25:e43038. [doi: 10.2196/43038] [Medline: 37851505]
17. Kim J, Lee HY, Christensen MC, Merighi JR. Technology access and use, and their associations with social engagement among older adults: do women and men differ? J Gerontol B Psychol Sci Soc Sci 2017 Sep 1;72(5):836-845. [doi: 10.1093/geronb/gbw123] [Medline: 28073816]
18. Bandura A. On the functional properties of perceived self-efficacy revisited. J Manage 2012 Jan;38(1):9-44. [doi: 10.1177/0149206311410606]
19. Balapour A, Reychav I, Sabherwal R, Azuri J. Mobile technology identity and self-efficacy: implications for the adoption of clinically supported mobile health apps. Int J Inf Manage 2019 Dec;49:58-68. [doi: 10.1016/j.ijinfomgt.2019.03.005]
20. Celano CM, Villegas AC, Albanese AM, Gaggin HK, Huffman JC. Depression and anxiety in heart failure: a review. Harv Rev Psychiatry 2018;26(4):175-184. [doi: 10.1097/HRP.0000000000000162] [Medline: 29975336]
21. Easton K, Coventry P, Lovell K, Carter LA, Deaton C. Prevalence and measurement of anxiety in samples of patients with heart failure: meta-analysis. J Cardiovasc Nurs 2016;31(4):367-379. [doi: 10.1097/JCN.0000000000000265] [Medline: 25930162]
22. Crundall-Goode A, Goode KM, Clark AL. What impact do anxiety, depression, perceived control and technology capability have on whether patients with chronic heart failure take-up or continue to use home tele-monitoring services? Study design of ADaPT-HF. Eur J Cardiovasc Nurs 2017 Apr;16(4):283-289. [doi: 10.1177/1474515116657465] [Medline: 27352948]
23. Maeda U, Shen BJ, Schwarz ER, Farrell KA, Mallon S. Self-efficacy mediates the associations of social support and depression with treatment adherence in heart failure patients. Int J Behav Med 2013 Mar;20(1):88-96. [doi: 10.1007/s12529-011-9215-0] [Medline: 22212607]
24. Irani E, Moore SE, Hickman RL, Dolansky MA, Josephson RA, Hughes JW. The contribution of living arrangements, social support, and self-efficacy to self-management behaviors among individuals with heart failure: a path analysis. J Cardiovasc Nurs 2019;34(4):319-326. [doi: 10.1097/JCN.0000000000000581] [Medline: 31058704]
25. Taylor CJ, Ordóñez-Mena JM, Roalfe AK, et al. Trends in survival after a diagnosis of heart failure in the United Kingdom 2000-2017: population based cohort study. BMJ 2019 Feb 13;364:l223. [doi: 10.1136/bmj.l223] [Medline: 30760447]
26. McDonagh TA, Metra M, Adamo M, et al. 2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure. Eur Heart J 2021 Sep 21;42(36):3599-3726. [doi: 10.1093/eurheartj/ehab368] [Medline: 34447992]
27. Scherrenberg M, Falter M, Kaihara T, et al. Development and internal validation of the digital health readiness questionnaire: prospective single-center survey study. J Med Internet Res 2023 Mar 10;25:e41615. [doi: 10.2196/41615] [Medline: 36897627]
28. Laver K, George S, Ratcliffe J, Crotty M. Measuring technology self efficacy: reliability and construct validity of a modified computer self efficacy scale in a clinical rehabilitation setting. Disabil Rehabil 2012;34(3):220-227. [doi: 10.3109/09638288.2011.593682] [Medline: 21958357]
29. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
30. Manea L, Gilbody S, McMillan D. A diagnostic meta-analysis of the patient health questionnaire-9 (PHQ-9) algorithm scoring method as a screen for depression. Gen Hosp Psychiatry 2015;37(1):67-75. [doi: 10.1016/j.genhosppsych.2014.09.009] [Medline: 25439733]

XSL•FO
RenderX

31. Lin Q, Bonkano O, Wu K, Liu Q, Ali Ibrahim T, Liu L. The value of Chinese version GAD-7 and PHQ-9 to screen anxiety and depression in Chinese outpatients with atypical chest pain. Ther Clin Risk Manag 2021;17:423-431. [doi: 10.2147/TCRM.S305623] [Medline: 34040380]
32. Hammash MH, Hall LA, Lennie TA, et al. Psychometrics of the PHQ-9 as a measure of depressive symptoms in patients with heart failure. Eur J Cardiovasc Nurs 2013 Oct;12(5):446-453. [doi: 10.1177/1474515112468068] [Medline: 23263270]
33. Kroenke K, Spitzer RL, Williams JBW, Löwe B. The patient health questionnaire somatic, anxiety, and depressive symptom scales: a systematic review. Gen Hosp Psychiatry 2010;32(4):345-359. [doi: 10.1016/j.genhosppsych.2010.03.006] [Medline: 20633738]
34. Shumaker SC, Frazier SK, Moser DK, Chung ML. Psychometric properties of the multidimensional scale of perceived social support in patients with heart failure. J Nurs Meas 2017 Apr 1;25(1):90-102. [doi: 10.1891/1061-3749.25.1.90] [Medline: 28395702]
35. Zimet GD, Powell SS, Farley GK, Werkman S, Berkoff KA. Psychometric characteristics of the multidimensional scale of perceived social support. J Pers Assess 1990;55(3-4):610-617. [doi: 10.1080/00223891.1990.9674095] [Medline: 2280326]
36. Merckelbach H, Muris P, Rassin E. Fantasy proneness and cognitive failures as correlates of dissociative experiences. Pers Individ Dif 1999 May;26(5):961-967. [doi: 10.1016/S0191-8869(98)00193-7]
37. Merckelbach H, Muris P, Nijman H, de Jong PJ. Self-reported cognitive failures and neurotic symptomatology. Pers Individ Dif 1996 Jun;20(6):715-724. [doi: 10.1016/0191-8869(96)00024-4]
38. Kaufman DR, Patel VL, Hilliman C, et al. Usability in the real world: assessing medical information technologies in patients' homes. J Biomed Inform 2003;36(1-2):45-60. [doi: 10.1016/s1532-0464(03)00056-x] [Medline: 14552846]
39. Stokes JW, Wanderer JP, McEvoy MD. Significant discrepancies exist between clinician assessment and patient self-assessment of functional capacity by validated scoring tools during preoperative evaluation. Perioper Med (Lond) 2016;5(18):18. [doi: 10.1186/s13741-016-0041-4] [Medline: 27413530]
40. Gordon NP, Hornbrook MC. Differences in access to and preferences for using patient portals and other eHealth technologies based on race, ethnicity, and age: a database and survey study of seniors in a large health plan. J Med Internet Res 2016 Mar 4;18(3):e50. [doi: 10.2196/jmir.5105] [Medline: 26944212]
41. van Riet EES, Hoes AW, Wagenaar KP, Limburg A, Landman MAJ, Rutten FH. Epidemiology of heart failure: the prevalence of heart failure and ventricular dysfunction in older adults over time. A systematic review. Eur J Heart Fail 2016 Mar;18(3):242-252. [doi: 10.1002/ejhf.483] [Medline: 26727047]
42. Groenewegen A, Rutten FH, Mosterd A, Hoes AW. Epidemiology of heart failure. Eur J Heart Fail 2020 Aug;22(8):1342-1356. [doi: 10.1002/ejhf.1858] [Medline: 32483830]
43. Pool J, Akhlaghpour S, Fatehi F, Gray LC. Data privacy concerns and use of telehealth in the aged care context: an integrative review and research agenda. Int J Med Inform 2022 Apr;160(104707):104707. [doi: 10.1016/j.ijmedinf.2022.104707] [Medline: 35131698]
44. Sheeran P, Webb TL. The Intention–Behavior Gap. Social & Personality Psych 2016 Sep;10(9):503-518 [FREE Full text] [doi: 10.1111/spc3.12265]
45. Faries MD. Why we don't "just do it": understanding the intention-behavior gap in lifestyle medicine. Am J Lifestyle Med 2016;10(5):322-329. [doi: 10.1177/1559827616638017] [Medline: 30202289]

Abbreviations

ADHF: acute decompensated heart failure
CFQ: Cognitive Failure Questionnaire
DHRQ: Digital Health Readiness Questionnaire
GAD-7 : Generalized Anxiety Disorder-7
HF: heart failure
LVEF: left ventricular ejection fraction
MREC: Medical Research Ethics Committee
MSPSS: Multidimensional Scale of Perceived Social Support
PHQ-9 : Patient Health Questionnaire-9

XSL•FO
RenderX

Edited by A Mavragani; submitted 25.11.24; peer-reviewed by D Chyun, R Murai; revised version received 20.03.25; accepted 21.03.25; published 12.06.25.

Please cite as:
Cuppen S, van Leunen M, Henken T, Goevaerts M, Scherrenberg M, Falter M, Dendale P, Kemps H, Kop WJ
Association of Technology-Related Skills and Self-Efficacy With Willingness to Participate in Heart Failure Telemonitoring: Cross-Sectional Observational Study
JMIR Form Res 2025;9:e68992
URL: https://formative.jmir.org/2025/1/e68992
doi:10.2196/68992

© Sharon Cuppen, Mayke van Leunen, Tamara Henken, Mayra Goevaerts, Martijn Scherrenberg, Maarten Falter, Paul Dendale, Hareld Kemps, Willem J Kop. Originally published in JMIR Formative Research (https://formative.jmir.org), 12.6.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Recommendations for Designing a Digital Health Tool for Blindness Prevention Among High-Risk Diabetic Retinopathy Patients: Qualitative Focus Group Study of Adults With Diabetes

Akua Frimpong[1,2], BS; Alvaro Granados[1], BS; Thomas Chang[1], MD; Julia Fu[1], BS; Shannan G Moore[1,3], BS; Serina Applebaum[1], BS; Bolatito Adepoju[4], MS, DVM; Mahima Kaur[4], MS; Vignesh Hari Krishnan[4], MArch; Amanda Levi[4], MPH; Terika McCall[4,5,6], PhD; Kristen Harris Nwanyanwu[1], MD, MBA, MHS

[1]Department of Ophthalmology and Visual Science, Yale University, 300 George Street, New Haven, CT, United States
[2]Larner College of Medicine, School of Medicine, University of Vermont, Burlington, VT, United States
[3]Department of Ophthalmology, Kaiser Permanente Bernard J. Tyson School of Medicine, Pasadena, CA, United States
[4]Division of Health Informatics, Department of Biostatistics, Yale School of Public Health, New Haven, CT, United States
[5]Department of Biomedical Informatics and Data Science, Yale School of Medicine, New Haven, CT, United States
[6]Center for Interdisciplinary Research on AIDS (CIRA), Yale School of Public Health, New Haven, CT, United States

Corresponding Author:
Kristen Harris Nwanyanwu, MD, MBA, MHS
Department of Ophthalmology and Visual Science, Yale University, 300 George Street, New Haven, CT, United States

Abstract

Background: Diabetic retinopathy (DR) is a leading cause of preventable blindness among working-aged adults. Black, Latine, and low-income individuals are screened less for DR, diagnosed later, treated less often, and go blind more than White individuals.

Objective: This study aimed to engage members to co-design a digital health tool that is accessible, user-friendly, and culturally relevant, through a community-led research approach,.

Methods: Using a qualitative approach, we conducted 4 semistructured focus group interviews with 19 individuals from the Greater New Haven area, aged 18 years or older, and diagnosed with diabetes. We transcribed and coded the focus group interviews and categorized them into themes using affinity mapping. The specific aims were to complete a comprehensive needs assessmen for the development of a community-responsive digital health tool and to increase access to information about DR screening in high-risk populations. We transcribed the focus group interviews, used rapid qualitative analysis to generate themes, and completed affinity mapping to identify content and features for a digital health tool for preventing blindness from DR.

Results: We interviewed 19 individuals (68% [13/19] female, 47% [9/19] Black, 26% [5/19]) Hispanic) in 4 focus groups. Over 80% (15/19) had access to smart devices, including smartphones (17/19, 89%), smartwatches (4/19, 21%), computers (14/19, 74%), and tablets (11/19, 58%). Many participants had access to multiple devices (17/19, 89%). Participants self-reported hemoglobin A1c (mean hemoglobin A1c 6.77, SD 1.93) and age (mean age 58.79, SD 19.54). Education levels among participants varied. Almost half of all the participants (9/19, 47%) completed some college, a little less than a quarter (4/19, 21%) achieved a high school diploma or general education development certificate, and a little less than a quarter (4/19, 21%) completed less than a high school equivalent of education. Household income levels across nearly all participants (14/19, 74%) were below US $50,000, but household size data were not collected. Participants reported extensive experience with diabetes or prediabetes (mean years with diabetes or prediabetes 17.06, SD 17.53). The themes obtained from coding focus group interviews included the mental toll of diabetes, peer support like accountability and local community events, education about diabetes management, barriers to DR screening like long wait times for appointments or cost of medications, and diet-related topics like how to find cost-effective healthy food

Conclusions: DR is one of the leading causes of blindness, and many treatments exist. Despite the existence of treatments, historically marginalized populations experience poor health outcomes, including blindness. Our community-based approach aids in the creation of a culturally responsive digital health tool.

(JMIR Form Res 2025;9:e65893) doi:10.2196/65893

KEYWORDS

diabetes; DM; diabetes mellitus; type 2 diabetes; type 1 diabetes; diabetic retinopathy; needs assessment; blindness; mHealth; mobile health; mobile application; smartphones; applications; digital health; digital technology; digital interventions; Health disparities

XSL•FO
RenderX

Introduction

Diabetic retinopathy (DR) is a leading cause of vision loss and blindness in the United States [1]. Routine eye examinations facilitate early detection and prevent vision loss, as DR is asymptomatic in early stages [2]. The prevalence of DR among people with diabetes was approximately 26% in 2021 [3]. Despite its high prevalence, less than 50% of individuals with diabetes in the United States undergo the recommended annual DR screening [4].

Screening rates vary widely among sociodemographic groups, and most of the groups that remain unscreened and untreated are historically marginalized communities, specifically Black, Latine, and low-income communities [5,6]. These communities are less likely to be aware of DR and experience from reduced engagement in eye care [7]. The barriers to care are complex and include interplays between insurance coverage, knowledge gaps regarding DR and the value of eye health screening [8].

In recent years, developers created mobile apps that focus on delivering diabetes-related education and reducing barriers to health care engagement [9]. However, the majority of studied mobile app interventions lack a culturally informed approach for Black, Latine, and low-income populations, which reduces their generalizability to these communities [10-12]. Studies show that cultural adaptations of health care interventions increase their effectiveness [13,14]. Structural racism, systemic oppression, and ongoing underrepresentation in health research contribute to delays and diminished access to health technology for traditionally underserved populations [15-17]. Patients at high risk of blindness from DR necessitate a culturally sensitive digital health tool that prioritizes their needs. One study tailored the design of a digital tool specifically for low-income Arabic individuals living in Israel, however, their intervention provided only diabetes-related dietary knowledge [18]. There is a clear and urgent need for content and features to improve DR screening rates and provide comprehensive diabetes-related support.

Individuals with diabetes often describe the mental toll of living with diabetes. Adverse psychological experiences are documented in the literature. There is a significant stigma surrounding diabetes, which leads to feelings of shame, guilt, or failure, especially for public management of the disease, like administering insulin [19]. Studies suggest that negative emotions play indirect causal roles, and themselves are outcomes of poor self-management of diabetes [20-22]. Fisher et al [23] showed that even at subclinical levels below the mild-to-moderate depression diagnosis threshold, depressive symptoms and emotional distress are linked to worse diabetes self-management. A systematic review found that psychological conditions, including depression, anxiety disorders, eating disorders, and other severe mental illnesses, are consistently associated with poorer self-management of diabetes in adults [24]. They further noted sources of anxiety related to diabetes, including hypoglycemia and fear of complications and invasive procedures [24]. It is documented in the literature that positive pathways of living with diabetes include support from family, friends, and other informal social networks [25]. Social support networks are often facilitated through the use of digital technology, which provides immediacy in social connections and fosters a sense of community [26]. Our study aims to lay the groundwork for the community-led design of a digital health tool that prioritizes community building to mitigate the mental toll of diabetes.

While various technologies exist that address the disparate needs of individuals with diabetes, there does not exist a single, comprehensive, digital health tool that integrates all the necessary facets. The purpose of this study is to identify the content, features, and key considerations for the community-driven design of a digital health tool to empower and promote marginalized groups' engagement in diabetic eye care and reduce disparities in preventable blindness.

Methods

Study Design

We included focus groups based on the Consolidated Framework for Implementation Research. All focus group sessions were audio recorded and transcribed by a professional service, Atomic Scribe. Subsequently, the focus group interview transcripts were coded and categorized into themes used for affinity mapping.

Recruitment of Participants

The research team (JF, TC, BA, MK, AF, TM, and KN) and members of a Community Advisory Board actively recruited potential participants from the New Haven community. We posted flyers throughout the New Haven community including in the Yale Eye Center, the Dana Eye Center, Yale affiliated primary care physician (PCP) and endocrinologist's offices, recreational centers, community restaurants, churches, and recruited by word-of-mouth. We also reached out to individuals who previously expressed interest in participating in studies at the Sight-Saving Engagement and Evaluation in New Haven (SEEN) Lab. For inclusion in the study, individuals needed to (1) be at least 18 years old, (2) have a diagnosis of diabetes, (3) speak English fluently, and (4) have reliable transportation to attend the in-person focus group.

We screened potential participants via telephone using a screening questionnaire. Each focus group session lasted approximately 90 minutes and included a dinner that accommodated participant food allergies and diet preferences.

Ethical Considerations

The institutional review board (IRB) at Yale School of Medicine approved the study and all participants gave informed consent (IRB #2000034710). This study adhered to the Declaration of Helsinki. Participants were consented before the beginning of any research activities, including the baseline assessment, by the research assistants (JF, TC, AF, BA, and MK), who were trained in the process. The Yale IRB reviewed all materials for ethical standards, and compliance. Privacy and confidentiality standards were maintained through the use of deidentified data for sharing with data analysts, and Protected Health Information was kept digitally on a private, encrypted server, while paper records containing Protected Health Information were kept in a secure, locked area that only provided access to the research

XSL•FO
RenderX

team. Participants who completed the study were each compensated with a gift card containing $50 US dollars.

Data Collection

The research team developed an interview guide Multimedia Appendix 1. We developed the interview guide using a community-based approach by consulting with an advisory board comprised of members from the community. We collaborated with members of our Community Advisory Board to refine the questions, ensuring they accurately addressed specific community needs and contexts (Multimedia Appendix 2). The focus group interview guide included questions about the following topics: (1) participants' experiences with diabetes (2) their eye health awareness, (3) their technology usage, and (4) their preferences for features in a digital tool to learn about eye health and screening. The team refined the interview guide during data collection to include emergent themes and improve question clarity. This iterative process allowed the team to adapt to participant feedback and incorporate prompting questions to elicit comprehensive responses.

Focus Group

We conducted semistructured focus group interviews in 2 community spaces. Results from a study by Guest et al [27] revealed that more than 80% of all themes are discoverable within two to three focus groups, and 90% of themes could be discovered within 3 to 6 focus groups. Therefore, 4 focus group sessions were conducted. Each session was capped at 6 participants to allow for all participants to fully engage in the discussions. We ensured the setting was conducive to discussion, collaboration, and confidentiality. AF (medical student) moderated each session, and multiple notetakers (VR, KN, TM, JF, TC, MK, and BA) attended different sessions to take notes on body language, nonverbal cues, and emerging themes. The moderator and notetakers debriefed after each interview. For the first 30 minutes of the focus group session, notetakers and the moderator ensured that participants completed the demographic surveys and signed the consent forms after demonstrating full understanding of them. We informed participants that all information would be deidentified and used solely for study purposes. The sessions were recorded with an audio recorder and we used a backup recorder on the computer. At the end of the sessions, each participant was compensated with a US $50 gift card. After each focus group, the moderator and notetaker debriefed. We sent all audio recordings to Atomic Scribe for transcription and redaction of identifying information to ensure participant confidentiality. We securely stored all data in an encrypted server.

Data Processing and Analysis

The analytic team was comprised of members from the SEEN Lab (AF, TC, JF, KN, VH, AG, and SA) and the Consumer Health Informatics Lab (TM, BA, MK, AL). Team members included a vitreoretinal surgeon, retina fellow, consumer health informatics expert, user experience and user interface designer, medical students, graduate students with advanced degrees in business, health sciences, and public health and researchers with extensive experiences working with marginalized populations.

Rapid Analysis

We (AF, TC, JF, MK, BA, KN,and TM) conducted a rapid analysis of each focus group session transcript to efficiently process data and share findings with the broader study team, to provide insights for our user-centered design process [28-32]. Analytic team members developed and used a rapid review analysis template to organize and interpret the interview data systematically [28,29]. The template guided the documentation of the analyzer's initial impressions, detailed accounts of participants' experiences with diabetes care, their use of technology in health care and eye care, and the analyzers' summary thoughts on each interview. The initial rapid analyses for each focus group session was led by TM within 1 week after each focus group session. The team recorded emerging themes and quotes.

Interpretation Sessions (Affinity Mapping Analysis of the Focus Group Interviews)

Following the initial analysis, each notetaker presented common themes from focus group transcripts to the broader research team. TM led interpretation sessions using affinity mapping, a type of cluster analysis, in which the team coded the interviews and categorized the comments into themes [33-37]. Interviewers met in 7 sessions between February 2024 and April 2024 to transform raw data from the transcripts and initial rapid analysis into Miro, an online tool for visual collaboration via creating and grouping online "sticky notes." Each sticky note represented a subtheme identified by a single interviewer, coded and colored to correspond to that interviewer. Each interpretation session included the research team (TM, AF, TC, JF, KN, VH, AG, SA, BA, and MK, SA, AG, and AL). The focus was on extracting the most pertinent information from each interview. The interviewers provided important details from information gleaned from the interview and their interactions with potential users. Key points, observations, and challenges in the design process were recorded. Finally, the modelers generated contextual design models to visualize the main data extracted from the focus group sessions (eg, affinity diagram). In the final session, the research team prioritized and categorized the important features and ideas that participants suggested, based on their impact and ease of implementation. After the last session, the study results were presented to the advisory board for reflection and feedback to refine the must have features list.

Results

Demographics and Baseline Characteristics

Thematic saturation was achieved with 4 focus groups (n=19, 2 - 6 participants each focus group session). We recruited 24 individuals, of whom 19 attended one focus group session. Focus group sessions were approximately 90 minutes long (60 min devoted to the semistructured interviews). The study participants included a diverse demographic, with 68% (13/19) female, 47% (9/19) identifying as Black, 26% (5/19) as Hispanic, and 11% (2/19) as indigenous. Over 80% (15/19) of participants had access to smart devices, defined as either smartphones, smartwatches, computers, and tablets.

XSL•FO
RenderX

Most participants, over 70% (13/19), disclosed annual household incomes below US $50,000. Hemoglobin A_{1c} (HbA_{1c}) was self-reported (mean 6.77, SD 1.93). Ages ranged from 35 years old to 72 years old (mean age 58.79, SD 19.54). Approximately half (10/19, 52%) of participants did not pursue higher education (neither partially nor completed a college degree). Over one quarter (5/19, >26%) of participants reported having a diagnosis of DR. Of note, one group consisted of all women, and another group included 3 participants with physical disabilities.

During the initial rapid qualitative analysis, the research team (KN, BA, AF, MK, TC, and VH) presented the main takeaways from the focus group sessions. The main findings highlighted the app should include: (1) education and resources on diabetes, (2) ability to communicate with providers, (3) options to personalize the user experience, (4) features that facilitate peer support, and (5) components that address emotional and mental stress associated with diabetes diagnosis and management. Supporting quotes for the main themes are presented in Table 1.

XSL•FO
RenderX

Table . Supporting quotes for the needs and priorities of participants to inform the development of a digital health tool to improve management of diabetes and diabetic retinopathy.

Themes	Examples
Resources	
Low-cost medications and treatments	• "My problem is paying for the medications. One of the pens that I use costs US $60, and I use three of them." [Speaker 1, Group 4]
Affordable healthy food	• "Because it's hard to buy food.... You got to buy certain foods to keep your diabetes in check. Fruit and vegetables, they're up sky-high." [Speaker 3, Group 2]
Access to providers	• "Sometimes it's hard to get in there. And my appointment is at 2, and I don't be seen until 4:30... If I'm on time, you should be on time." [Speaker 1, Group 2]
Difficulty securing eye care	• "I was just going to do it [go to the appointment] in January. And the other thing is that I moved, so from [Redacted] County to [Redacted] County. So, I had to find all these new doctors." [Speaker 2, Group 4] • "There is a lot of people that really need help. There's a lot of people that really have bad diabetes. They not even have insurance.... And nobody will help them. They not even know where to go." [Speaker 2, Group 2]
Exercises	• "The one that could test your blood sugar. One that could tell you if you've drank enough water. And if you're too high, maybe it'll recommend perhaps get another 2000 steps in or something." [Speaker 1, Group 1] • "I'm in this chair. I really can't exercise, because I can barely stand up. All I do, all I know how to do is get up in the morning and check my sugars and stick myself." [Speaker 1, Group 2]
Eye care	• "Again, I knew, from my mom and because I watched her, that deterioration. She wasn't always blind, but I watched that about I would say 10 or 15 years." [Speaker 2, Group 4] • "I found out when I had double vision, that's when I found out it'll affect your eyes. Too little, too late, but at least I'm on it now." [Speaker 5, Group 3]
Transportation to appointments	• "Sometimes it's hard getting around here in the wintertime, and I don't go [to the appointment]." [Speaker 1, Group 2]
Healthy recipes	• "On Facebook I get one of the [recipe] things in my feed, but they are not diabetic-friendly. They're just the opposite." [Speaker 1, Group 4] "I wonder if a menu planning app would be useful…it tells you what the sugar, what the fiber, what the carbs, all those things are." [Speaker 6, Group 3]
Education	• "I think it's education, too. Even my father having [diabetes], they didn't know why he was so angry all the time, fighting and acting out. They didn't know his blood sugars affected his mood." [Speaker 2, Group 1] • "I don't understand it. Somebody, break it down for me. Just don't everybody keep repeating the same thing to me, I know that a friend of mine has lost his two legs." [Speaker 2, Group 4]
Communication with providers	
Suggested queries for health literacy	• "FAQs[a], because things come up that you didn't think of… If I could just open up an app and say, 'Answer the question', and it would take me to some kind of explanation or answer, that would be useful." [Speaker 6, Group 3] • "[I would like to know] What complications could arise if you don't get your sugar under control." [Speaker 5, Group 3]
MyChart link	• "I would say with the MyChart, I like it because whether I'm in Raleigh, North Carolina, no matter where I'm at, when I go to the emergency room and I couldn't talk, they were able to pull up my MyChart and understand what was going on and what was wrong with me." [Speaker 3, Group 1]
Personalized user experience	

XSL•FO
RenderX

Themes	Examples
Notifications	• "Notifications? Like, yeah, if you're going high, [or] you're going low. But something that could also just balance everything and let you know, "this is where you normally at, and now you're off by this or that." [Speaker 3, Group 2] • "But I don't need an app telling me, 'All right, you got to do this now,' for somebody [like me] that's been living with it for so long." [Speaker 1, Group 4]
Incentives	• "Maybe a point system too with the app and incentive[s]. Maybe after so many points or so, logging in or taking care of yourself or something, [you get] a free meal, [to a] different restaurant. [Laughter]." [Speaker 2, Group 1]
Facilitating peer support	
Accountability	• "Sometimes I find the need for someone to hold me accountable. 'What did you eat today? Why did you do that?'" [Speaker 5, Group 3] • "My family would embarrass me. They'd be like, 'Why are you eating that? Aren't you diabetic?'"[Speaker 4, Group 3]
Local DM[b] events/groups (virtual and in person)	• "Maybe some kind of communication board, 'I have your type of diabetes in your age group.' I don't have a support network. Might be nice to connect with someone online, just to check in with each other." [Speaker 5, Group 2]
Mental impact	
Isolation	• "I get so depressed. I didn't even want to walk outside. Almost all my family died from diabetes. So, I start[ed] thinking about me, my kids and my grandkids, and I make [sic] the decision to try to control my sugar." [Speaker 2, Group 2]
Polypharmacy	• "And so now I am on two different insulin[s] that I take. They're thinking about putting me on Metformin again, so that'll be three. And I am so sick of it. It adds to my depression." [Speaker 4, Group 1]

[a]FAQs: frequently asked questions.

[b]DM: diabetes mellitus.

Resources

Participants expressed a desire for diabetes-specific diet advice, particularly regarding specific food choices and how much to eat as well as finding ways to exercise. Some participants highlighted a need for easy recipes and meal plans for cooking at home, incorporating nutrition information and comprehensive recommendations. When dining out, participants said they would benefit from education regarding what is considered fast food, and what take-out options are better choices. Participants expressed frustration at the obscurity of food labels and wanted education on reading them to better understand their food's nutritional content. Participants understood the importance of exercising but found it difficult to exercise, especially when it came to feeling motivated. Participants expressed wanting resources for all types of skill levels in exercising and having a group event to motivate them to exercise.

> *But that's my whole journey with this diabetes…I'm trying to get my weight down because I realize if my weight is down and I exercise more, [and] I'm moving more, it will go away. I think I can live with a prediabetes.* [Speaker 4, Group 3]
>
> *So, it took a compilation of all the foods, So I think that would be great too, just to have an app that would automatically do that," [Speaker 3, Group 1]. "Something that'll help you with a diet,"* [Speaker 3, Group 2.]
>
> *"Like maybe a meal planner.* [Speaker 5, Group 2]

Education

Participants sought to know what the effects of diabetes are on their health. Many participants desired informative content in the app regarding diabetes. For example, participants requested lists of "do's and don'ts for people with diabetes" and "frequently asked questions." Some participants also requested comprehensive information regarding the complications and side effects of diabetes, results from new diabetes-related research, and information about novel treatments and their effects.

> *When you're not compliant and you're out of control, your sex drive drops. I thought I was [just] getting older… those interpersonal things that nobody wants to talk about… Let's everybody bring that forward.* [Speaker 3, Group 1]
>
> *But a lot of trial and error came from, as I said, watching my mom, watching things with her. And it changed a lot, because I remember going to the hospital and she had that orange. They were teaching her how to stick herself, give herself insulin, which was having an orange, and she was injecting under the skin on the orange. They don't do that anymore.* [Speaker 2, Group 4]
>
> *How [blood sugar] needs to be checked frequently. How [my vision] could blur. About eating different*

XSL•FO
RenderX

things could help my vision. And I said, 'Well, wow, I didn't know that.' [Speaker 1, Group 3]

Eye Health

When it came to understanding how diabetes affects the eyes, some participants shared that they were informed about the risk of vision problems and the importance of eye care at the time of their diagnosis. In contrast, others expressed that they were not told that diabetes can lead to blindness. Some learned about eye care through personal experiences or from family and friends, while others were informed by their health care providers. Many recognized the benefits of regular eye exams, with most having had an exam within the past year to a few months ago, despite some difficulty scheduling appointments. The group had a positive attitude toward diabetes eye care, although experiences varied. Participants also expressed a desire to better understand how diabetes affects the eyes.

So, I had really, really good—have really good vision. But at one point when my diabetes was uncontrolled… I couldn't see the license plate right in front of me. Like everything was blurry and starry at night. So, I couldn't even drive. [My brother] had to drive for me at one point, having my own vehicle. [Speaker 2, Group 1]

When I first became diabetic and I found out, like I said, my sugar was like 740, and I remember everything being blurry. I was so used to being in shape, doing all these sports… it really blindsided me. It's like getting ear-holed by a wide receiver. [Speaker 3, Group 1]

Barriers to Health Care

Participants spoke of the barriers that they face accessing health care. Long wait times to scheduling an appointment with a physician was a common topic, as was limited access to health care. Participants also reported challenges with the sheer amount of medications to take and their associated cost, availability, and side effects. Several participants reported financial barriers, citing the high overall cost of care. Many participants were well-versed in the need to make lifestyle changes but reported barriers to implementing them including transportation, the increasing cost of healthy foods and comparatively high accessibility of fast food.

…sometimes it's hard to go to get an eye exam because they charge so much. It's expensive when you go get an eye exam. And plus, those testers…some of it's out [of] pocket… So that's the main hurdle right there, is try[ing] to afford that test. [Speaker 3, Group 2]

My mother was a diabetic. My father was a diabetic. Her mother was a diabetic. So, it ran in the family. I still don't get this. I can't do certain things. I'm in this chair. I really can't exercise, because I can barely stand up. All I do, all I know how to do is get up in the morning and check my sugars and stick myself. [Speaker 1, Group 2]

Access to Health Information

Participants expressed a desire for a user-friendly digital health tool that is informative without being overwhelming. They suggested features such as a point reward system, customization options, and streamlined content. A key priority was improving communication with healthcare providers, with participants highlighting the need for supplementary materials and guidance to facilitate conversations with doctors. They also showed interest in opportunities to participate in research studies and clinical trials. Some participants spoke about having a digital tool that makes records easy to access, like MyChart.

Participants discussed their preferences for receiving information, including push notifications from the app. They valued notifications about new treatments, support groups, upcoming doctor's appointments, medication reminders, exercise reminders, blood glucose, and blood pressure tracking.

So maybe an app that would say, [if] you start to have blurred vision or something like that…what you should or shouldn't do. And maybe what kind of foods you should eat that will make your eyesight better. [Speaker 4, Group 3]

Support and Engagement

Participants consistently identified family as a source of support. Some participants had older family members with diabetes, so they grew up witnessing the experience of diabetes, forming a unique perspective. Others spoke of friends and community as a resource. Support for these individuals came from community health centers and health care providers. In contrast, other participants described a need for stronger support due to lack of knowledge or commitment from their existing family or community. Participants voiced that they would like to have a platform to build community.

[My friend]'ll say, 'Okay, did you do your shots, your pills and your drugs?' So she reminds me all the time of it, but… she tempts me, "Let's go have sprinkles." [Speaker 1, Group 4]

You were asking about support. she has MyChart, and I have a Hero…A Hero is really good. It dispenses my medications for me. But that's all on her phone as well, my daughter. So, she sees that as well. She says, "Mommy, why haven't you taken such and such? You know you've been taking it late. [Speaker 2, Group 4]

Mental Health Impact

Participants spoke of the emotional toll that their chronic disease has had on them. They described feelings of depression, denial, embarrassment, and psychosocial struggles. They spoke of the emotional journey that living with diabetes engenders and seeking answers to how diabetes complications can affect their mental capacity.

I've been a diabetic for 20 years. It's a nightmare. I remember the day that they told me that I was a full-blown diabetic. I cried my eyeballs, and I went into denial. . . I wouldn't take care of myself. They started me with Metformin. . . and I wouldn't take the

XSL•FO
RenderX

> *pill. I would eat whatever I wanted to eat. . . wouldn't exercise until I got sick. I got really sick, [my blood sugar] went up to 589. I had to be hospitalized.* [Speaker 4, Group 1]

> *I believe I was . . .19, 20 when it came, because I had kind of quit football because I gave up because it gave up on me. I was depressed, so my weight went up to like 375. I was a size 56. And one day, I just could not fix my craving for something to drink. . . my grandmother took me to the hospital. . . I found out my sugar was like 740. . .they told me if I didn't lose at least 150 pounds, I was going to die. So, it took me about two and a half years to get my weight down. It wasn't easy. I had to stop eating a 48-box of Snickers. I had to. . .serve myself less. . . I won't go back to [eating] trash to get it, because I had 5700 calories and I'm starving.* [Speaker 3, Group 1]

Figure 1 presents features organized by impact potential and implementation difficulty. The development of the digital health tool prioritized the must-have features, located to the upper left quadrant. These were considered the most important to the participants and feasible for the research team to execute. The digital health tool will include reliable content about eye care, diabetes, and related health topics, addressing the need for trustworthy information about diagnosis, medications, and treatments. To further promote holistic wellness, it will offer resources for exercise, cooking videos, mental health, and general wellness, and other functions to provide, promote, and support overall well-being. Features will inform users of local events, particularly those related to diabetes, offer resources for clinical trials, established programs, and integrate relevant existing applications for managing patient health data. The advisory board showed consensus on the must-have features. Additional recommendations included adding testimonials and tutorials to the "Frequently Asked Questions" section, and establishing partnerships with local businesses, advocacy organizations, and public works departments, such as the city health department and the National Association for the Advancement of Colored People, to address social needs. In addition, the advisory board highlighted the value of community groups and discussion forums for individuals with diabetes.

Figure 1. Prioritization and categorization of features and ideas for digital health tool design based on impact and ease of implementation.

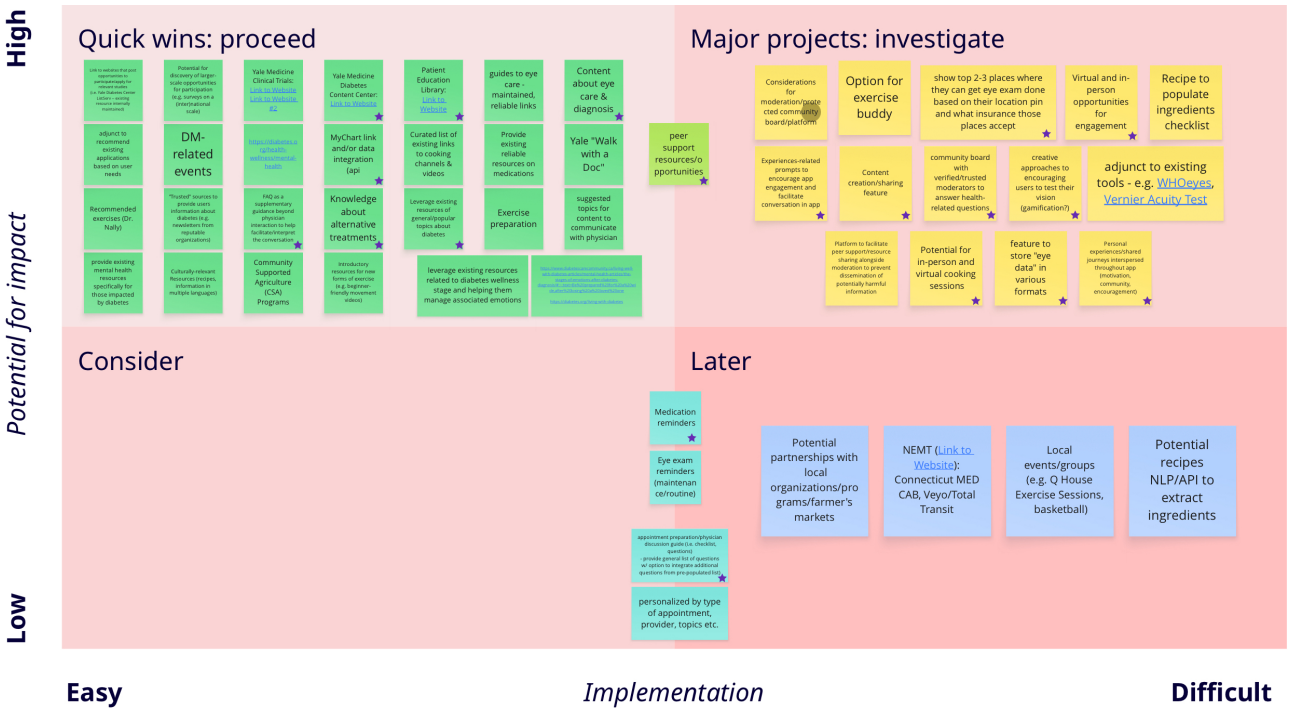

Discussion

Principal Findings

Our findings highlight the need for a culturally responsive digital health tool for individuals with diabetes. This tool will incorporate lifestyle resources, informal support platforms, and electronic health record access while maintaining a well-calibrated notification system. These findings align with previous research on developing smartphone apps for diabetes management, emphasizing the importance of usability, patient-centered resources, and seamless integration with existing health care infrastructures. The consistency with other studies highlights the critical role of tailored digital tools in enhancing diabetes care and patient engagement [9,10,12]. The need for a patient-centered digital health tool that provides access to resources like guided communication with health care providers, eye health, and diabetes education. Key features include personalized user experiences, customizable notifications, peer support, and community-building platforms. Consistent with existing literature, our study confirms that patients with chronic conditions like diabetes prefer technology that supports online community engagement and self-management resources, including exercise routines, healthy recipes, and telehealth options, to enhance their overall health management [38].

XSL•FO
RenderX

Technological Tools for Diabetes Management

Studies show that technological innovations for diabetes management, particularly through mobile devices, can improve health outcomes [10,39]. Low-cost, user-friendly digital tools that centralize information and functionality can facilitate patient autonomy and improve access to health guidance. These technologies can further inform medication management, blood glucose monitoring and subsequent communication with health care professionals [40]. A recent benchmark study evaluated the usability and content of available mobile apps for people with diabetes [41]. In their systematic evaluation, most apps had good usability and potential to facilitate diabetes management. Common app features included nutritional and dietary planning, monitoring glycemic control, and scheduling health appointments. Content areas that were less well represented included mental health, education on diabetes complications, and content aimed at individuals with disabilities. No apps existed with content related to DR or resulting visual impairment [41]. While many digital tools do exist, there are significant content gaps among available resources. There is a widespread need for an app that provides reliable information and guidance for people with diabetes who want to prevent or manage their visual impairment resulting from DR.

In addition to content within an app, the design of an app impacts its success. Specifically, the screen or display size, color, material, menu structure, user-friendliness, accessibility, and simplified operations can influence user adoption and use, particularly among older adults [42,43]. Therefore, when designing interventions or applications designed for older adults with diabetes, health care providers or app developers should focus on aesthetic features that consider the physical, sensory, and cognitive changes experienced by older adults, such as impaired vision, slow hand movement, and decreased response time [44-46].

Mental Health

The impact of chronic disease on mental health is well-documented. Individuals living with type 1 or type 2 diabetes are at heightened risk for depression, anxiety, and eating disorders [47,48]. Many people with chronic diseases, including diabetes, use the internet to access psychological treatments [49-51]. Baldwin et al [52] explored the use of the web-based electronic mental health program, myCompass. MyCompass is both clinically effective and cost-efficient in expanding access to mental health services [50] and has previously demonstrated efficacy in reducing depressive symptoms and diabetes-related distress in individuals with type 2 diabetes. However, a notable limitation of MyCompass is that being primarily a mental health support program, it does not provide support targeting specific diabetes needs [52]. More recent studies incorporate diabetes-specific content [53] alongside existing depression treatments and cognitive behavior therapy [54]. Participants reported significant improvements in health behaviors, such as healthier eating habits, better health monitoring, and incremental progress in overall outcomes [52]. With increasing recognition of the importance of mental health in diabetes care [55], our study acknowledges the need to address mental health in the management of diabetes in the design of our proposed digital health tool.

Community Space

Several studies emphasize the critical role of social support in diabetes management [56-58]. Toobert and Glasgow [59] identified it as a vital factor in self-care adherence, while Marquez et al [60] noted its impact on physical activity and weight loss. Shayeghian et al [61] linked social support to better blood glucose control, and Pereira et al [62] highlighted its importance for self-care behaviors, such as glucose monitoring. Diabetes online communities have emerged as valuable platforms for peer support, fostering communication, improving disease management, and enhancing psychosocial well-being [63].

There are few studies on the role of online peer support in diabetes care. One study used digital observation, applying ethnographic techniques to study interactions on social media and the web to analyze hundreds of diabetes-related posts on platforms like Facebook (Meta), Twitter (X), and YouTube (Google), identifying 6 core themes—humor, diabetic pride, technology use, tips and tricks, community building, and venting [64]. The findings emphasize the importance of online platforms in fostering camaraderie, peer mentorship, and shared experiences often overlooked by health are systems. In addition, it has been shown that integrating mental health services and peer support is beneficial and critical for positive short- and long-term diabetes outcomes [64-66]. Our digital tool aims to replicate these positive outcomes by creating a supportive environment for patients to connect with each other and share their experiences, thereby enhancing their overall care.

Health Literacy

A significant concern for participants was their lack of knowledge about how diabetes could affect them. These gaps included understanding the full range of symptoms, complications, predisposing factors, and psychological impacts of diabetes. Research shows that individuals without diabetes self-management education are less likely to engage in preventative care, leading to worse outcomes [67]. Limited health literacy is recognized as a significant barrier to improving health outcomes. Individuals with low health literacy may have less knowledge about diabetes, have less control of blood glucose levels, engage in less self-management behaviors, and incur higher health care costs [68]. One study demonstrated a direct correlation between health literacy and several key factors in diabetes management, including self-monitoring, nutrition and exercise habits, diabetes knowledge, self-care practices, and social support [69]. Another study recommended that targeted educational and behavioral interventions are crucial in addition to health care service strategies, particularly for older adults and individuals with less formal education [10,65,66].

With the rise of social media use, patients with diabetes are more likely to seek information using technology to help manage their condition [70]. In one study, nearly half of participants sought self-management information online, with dietary planning being the most common focus [71,72]. The term eHealth literacy refers to the critical skills needed to effectively

XSL•FO
RenderX

use technology for seeking, accessing, and understanding health information from electronic sources [73,74]. Previous research demonstrates a significant association between eHealth literacy and health behaviors. Lower eHealth literacy is linked to poor diabetes self-management and worse health outcomes, particularly among older adults with diabetes [75,76].

Incorporating design preferences of intended users into digital health tools may increase participant engagement and empower them to manage their health better [13,14]. This includes trust-centered design elements such as reliable content on eye care, diabetes, social support, and healthy lifestyles [13,14]. Studies suggest that digital tools should be specifically designed to address diabetes-related concerns [58,77,78].

These findings align with the discussed topics that participants acknowledged and the importance of incorporating eHealth literacy and social support into digital health tools.

Eye Health

Some existing apps related to DR include the National Eye Institute's virtual reality app, which simulates vision loss, and the RetinaRisk (Retina Risk) app, which calculates personalized risk for diabetic retinopathy. However, few apps focus extensively on DR content or address visual impairment resulting from the disease. Our participants expressed a strong desire for this type of content. Our digital health tool will fulfill this gap by incorporating trusted resources for education, research opportunities, community events, and mental health support tailored to eye health and diabetes care.

Strengths and Limitations

Although this study has many strengths and advantages, there are several limitations and disadvantages when conducting focus group discussions as part of the methodology. Some challenges include limited control over data collection, the influence of group dynamics on participants' responses, and the difficulty in generalizing findings to a larger population.

Focus groups can sometimes include silent or dominant participants, which may influence the conversation dynamics and impact data collection [79]. To address this, the moderator (AF) and observers (VH, KN, TM, JF, TC, MK, and BA) implemented strategies to ensure that all participants had an opportunity to share their views, such as redirecting the discussion to stay on topic, rephrasing questions to allow wider range of responses, and seeking clarification to encourage the expression of diverse viewpoints and perspectives.

Our focus groups had homogeneous backgrounds (diabetes status and geographic location), while also having heterogeneous traits (age differences, gender, race, and ethnicity) that ensured a diversity of perspectives and greater depth of analysis into the challenges they faced [80]. Although participants in this study shared similar opinions and perspectives, their experiences may not be generalizable to the broader population of individuals with diabetes mellitus, particularly given the geographical restriction of recruitment of the participants, which limits the national representativeness of the sample [81].

Participant recruitment included the participants self-reporting their health measures such as current hemoglobin A_{1c}. This form of providing information can introduce a potential source of error or bias due to self-reporting that cannot be verified [82]. The SEEN tool has the potential to address this limitation by allowing participants to input their information or data more accurately for recall and reporting. For instance, the tool may be interoperable with glucose monitors, allowing participants to integrate their glucose data into the tool, thereby minimizing the likelihood of self-report errors.

This study did not collect household size data, which limits our ability to infer whether individuals met the criteria for poverty. However, the experience of multiple individuals manifesting through direct quotes from them is that medication costs, food costs, and other financial barriers impact access to eye care and screening. Whether or not any participants have experienced financial difficulty with managing their diabetes, their statements point to the universality of financial barriers to health care equity.

Despite these limitations, the study results yielded key insights that will inform the design team on what community needs to prioritize in the development of the digital health tool. Furthermore, our community-based approach lifts the voices of participants who are historically under-represented in medical research, and empowers them with lasting educational benefits and other resources to facilitate their management of DR.

Future Directions

The main findings of this study serve as a guide for the design priorities for a digital health tool tailored to individuals with diabetes or DR.

The next steps, which is currently underway, will be to develop an initial prototype of the digital health tool through 3 phases of production: design, development, and evaluation of the digital health tool. During the design phase, we will create detailed user journey maps, develop health education content for the tool, and develop low-fidelity wireframes and interactive high-fidelity prototypes. All of this will be worked in collaboration with the community members to incorporate feedback and ensure the tool is intuitive and user-friendly. After the design phase, the team will begin the development phase to build the digital health tool by collaborating with developers to build the tool. Finally, the final phase will be to conduct usability testing and implement a pilot study to collect quantitative and qualitative data to test the prototype with application users in order to ensure the tool meets the community's needs.

Future studies can expand the scope and the focus groups. The focus groups can consist of more diverse populations, including different age groups, genders, socioeconomic statuses, and cultural backgrounds, to gather a wide range of perspectives. Researchers could also include participants from various geographic locations as well to understand regional differences in digital health needs and preferences.

Conclusions

Using a user-centered design approach, when developing digital health tools, has the potential to significantly improve adverse health outcomes for individuals living with diabetes through culturally relevant and personalized interventions. The

XSL•FO
RenderX

community-centered approach not only empowers participants from communities which are often unheard to give their voice to health care research, but will also educate them through an educational component, including connecting the participants to resources to help them manage DR.

Acknowledgments

The authors would like to thank the participants for sharing their unique personal experiences. We would like to extend our gratitude to the members of the Community-engaged Research Steering Committee, the Reverend Dr Frederick J Streets, and the advisory board members (Dr Annette Streets, Mr Richard Furlow, Tyisha Walker-Myers, Reverend Andre Gilford). We would also like to thank the director and staff from the Dixwell Community House for providing a space to conduct the focus groups and for providing local vendors and caterers for the sessions, which supported local businesses and reinforced our commitment to keeping the study's impact within the community. We would like to thank the Yale Eye Center for supporting participant recruitment. The authors would like to give special thanks to the Consumer Health Informatics Lab at Yale University for their support in conducting this study. This study received support and funding from the National Center for Advancing Translational Science / The National Institutes of Health / Department of Health and Human Services.

Data Availability

The data sets generated during and/or analyzed during this study are available from the corresponding author upon request.

Conflicts of Interest

Author TM is an associate editor of the Journal of Medical Internet Research.

Multimedia Appendix 1
Focus Group Session Questions.
[DOCX File, 24 KB - formative_v9i1e65893_app1.docx]

Multimedia Appendix 2
Digital Health Tool Pre-Survey/Questionnaire.
[DOCX File, 12 KB - formative_v9i1e65893_app2.docx]

References

1. Cheung N, Mitchell P, Wong TY. Diabetic retinopathy. Lancet 2010 Jul 10;376(9735):124-136. [doi: 10.1016/S0140-6736(09)62124-3] [Medline: 20580421]
2. Morello CM. Etiology and natural history of diabetic retinopathy: an overview. Am J Health Syst Pharm 2007 Sep 1;64(17 Suppl 12):S3-S7. [doi: 10.2146/ajhp070330] [Medline: 17720892]
3. Lundeen EA, Burke-Conte Z, Rein DB, et al. Prevalence of diabetic retinopathy in the US in 2021. JAMA Ophthalmol 2023 Aug 1;141(8):747-754. [doi: 10.1001/jamaophthalmol.2023.2289] [Medline: 37318810]
4. Gibson DM. Frequency and predictors of missed visits to primary care and eye care providers for annually recommended diabetes preventive care services over a two-year period among U.S. adults with diabetes. Prev Med 2017 Dec;105:257-264. [doi: 10.1016/j.ypmed.2017.09.019] [Medline: 28963006]
5. Barsegian A, Kotlyar B, Lee J, Salifu MO, McFarlane SI. Diabetic retinopathy: focus on minority populations. Int J Clin Endocrinol Metab 2017;3(1):034-045. [doi: 10.17352/ijcem.000027] [Medline: 29756128]
6. Ravindranath R, Bernstein IA, Fernandez KS, Ludwig CA, Wang SY. Social determinants of health and perceived barriers to care in diabetic retinopathy screening. JAMA Ophthalmol 2023 Dec 1;141(12):1161-1171. [doi: 10.1001/jamaophthalmol.2023.5287] [Medline: 37971726]
7. Nwanyanwu K, Nunez-Smith M, Gardner TW, Desai MM. Awareness of diabetic retinopathy: insight from the national health and nutrition examination survey. Am J Prev Med 2021 Dec;61(6):900-909. [doi: 10.1016/j.amepre.2021.05.018] [Medline: 34426057]
8. Fairless E, Nwanyanwu K. Barriers to and facilitators of diabetic retinopathy screening utilization in a high-risk population. J Racial Ethn Health Disparities 2019 Dec;6(6):1244-1249. [doi: 10.1007/s40615-019-00627-3] [Medline: 31463812]
9. Salari R, Niakan Kalhori SR, Fatehi F, Ghazisaeedi M, Nazari M. Determining minimum set of features for diabetes mobile apps. J Diabetes Metab Disord 2019 Dec;18(2):333-340. [doi: 10.1007/s40200-019-00417-y] [Medline: 31890658]
10. Doupis J, Festas G, Tsilivigos C, Efthymiou V, Kokkinos A. Smartphone-based technology in diabetes management. Diabetes Ther 2020 Mar;11(3):607-619. [doi: 10.1007/s13300-020-00768-3] [Medline: 31983028]
11. Salari R, R Niakan Kalhori S, GhaziSaeedi M, Jeddi M, Nazari M, Fatehi F. Mobile-based and cloud-based system for self-management of people with type 2 diabetes: development and usability evaluation. J Med Internet Res 2021 Jun 2;23(6):e18167. [doi: 10.2196/18167] [Medline: 34076579]

XSL•FO
RenderX

12. Liu K, Xie Z, Or CK. Effectiveness of mobile app-assisted self-care interventions for improving patient outcomes in type 2 diabetes and/or hypertension: systematic review and meta-analysis of randomized controlled trials. JMIR Mhealth Uhealth 2020 Aug 4;8(8):e15779. [doi: 10.2196/15779] [Medline: 32459654]
13. Liau YW, Cheow C, Leung KTY, et al. A cultural adaptation and validation study of A self-report measure of the extent of and reasons for medication nonadherence among patients with diabetes in Singapore. Patient Prefer Adherence 2019;13:1241-1252. [doi: 10.2147/PPA.S208736] [Medline: 31551654]
14. Hall GCN, Ibaraki AY, Huang ER, Marti CN, Stice E. A meta-analysis of cultural adaptations of psychological interventions. Behav Ther 2016 Nov;47(6):993-1014. [doi: 10.1016/j.beth.2016.09.005] [Medline: 27993346]
15. Brown SD, Lee K, Schoffman DE, King AC, Crawley LM, Kiernan M. Minority recruitment into clinical trials: experimental findings and practical implications. Contemp Clin Trials 2012 Jul;33(4):620-623. [doi: 10.1016/j.cct.2012.03.003] [Medline: 22449836]
16. Wherry K, Zhu C, Vigersky RA. Inequity in adoption of advanced diabetes technologies among medicare fee-for-service beneficiaries. The Journal of Clinical Endocrinology & Metabolism 2022 Apr 19;107(5):e2177-e2185. [doi: 10.1210/clinem/dgab869]
17. Yao R, Zhang W, Evans R, Cao G, Rui T, Shen L. Inequities in health care services caused by the adoption of digital health technologies: scoping review. J Med Internet Res 2022 Mar 21;24(3):e34144. [doi: 10.2196/34144] [Medline: 35311682]
18. Abu-Saad K, Murad H, Barid R, et al. Development and efficacy of an electronic, culturally adapted lifestyle counseling tool for improving diabetes-related dietary knowledge: randomized controlled trial among ethnic minority adults with type 2 diabetes mellitus. J Med Internet Res 2019 Oct 16;21(10):e13674. [doi: 10.2196/13674] [Medline: 31621640]
19. Liu NF, Brown AS, Folias AE, et al. Stigma in people with type 1 or type 2 diabetes. Clin Diabetes 2017 Jan;35(1):27-34. [doi: 10.2337/cd16-0020] [Medline: 28144043]
20. Gonzalez JS, Tanenbaum ML, Commissariat PV. Psychosocial factors in medication adherence and diabetes self-management: implications for research and practice. Am Psychol 2016 Oct;71(7):539-551. [doi: 10.1037/a0040388] [Medline: 27690483]
21. Ducat L, Philipson LH, Anderson BJ. The mental health comorbidities of diabetes. JAMA 2014 Aug 20;312(7):691-692. [doi: 10.1001/jama.2014.8040] [Medline: 25010529]
22. Guerrero Fernández de Alba I, Gimeno-Miguel A, Poblador-Plou B, et al. Association between mental health comorbidity and health outcomes in type 2 diabetes mellitus patients. Sci Rep 2020 Nov 11;10(1):19583. [doi: 10.1038/s41598-020-76546-9] [Medline: 33177607]
23. Fisher L, Gonzalez JS, Polonsky WH. The confusing tale of depression and distress in patients with diabetes: a call for greater clarity and precision. Diabet Med 2014 Jul;31(7):764-772. [doi: 10.1111/dme.12428] [Medline: 24606397]
24. de Groot M, Golden SH, Wagner J. Psychological conditions in adults with diabetes. Am Psychol 2016 Oct;71(7):552-562. [doi: 10.1037/a0040408] [Medline: 27690484]
25. Hill-Briggs F, Adler NE, Berkowitz SA, et al. Social determinants of health and diabetes: a scientific review. Diabetes Care 2020 Nov 2;44(1):258-279. [doi: 10.2337/dci20-0053] [Medline: 33139407]
26. Elnaggar A, Ta Park V, Lee SJ, Bender M, Siegmund LA, Park LG. Patients' use of social media for diabetes self-care: systematic review. J Med Internet Res 2020 Apr 24;22(4):e14209. [doi: 10.2196/14209] [Medline: 32329745]
27. Guest G, Namey E, McKenna K. How many focus groups are enough? building an evidence base for nonprobability sample sizes. Field methods 2017 Feb;29(1):3-22. [doi: 10.1177/1525822X16639015]
28. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. Implement Sci 2009 Aug 7;4:50. [doi: 10.1186/1748-5908-4-50] [Medline: 19664226]
29. Holtzblatt K, Beyer H. Contextual Design 2nd Edition: Design for Life 2017.
30. Hamilton AB, Cohen AN, Glover DL, et al. Implementation of evidence-based employment services in specialty mental health. Health Serv Res 2013 Dec;48(6 Pt 2):2224-2244. [doi: 10.1111/1475-6773.12115] [Medline: 24138608]
31. Vindrola-Padros C, Johnson GA. Rapid techniques in qualitative research: a critical review of the literature. Qual Health Res 2020 Aug;30(10):1596-1604. [doi: 10.1177/1049732320921835] [Medline: 32667277]
32. Nevedal AL, Reardon CM, Opra Widerquist MA, et al. Rapid versus traditional qualitative analysis using the consolidated framework for implementation research (CFIR). Implement Sci 2021 Jul 2;16(1):67. [doi: 10.1186/s13012-021-01111-5] [Medline: 34215286]
33. Hipwell AE, Sturt J, Lindenmeyer A, et al. Attitudes, access and anguish: a qualitative interview study of staff and patients' experiences of diabetic retinopathy screening. BMJ Open 2014 Dec 15;4(12):e005498. [doi: 10.1136/bmjopen-2014-005498] [Medline: 25510885]
34. Alexander RL, Miller NA, Cotch MF, Janiszewski R. Factors that influence the receipt of eye care. Am J Health Behav 2008;32(5):547-556. [doi: 10.5555/ajhb.2008.32.5.547] [Medline: 18241139]
35. Owsley C, McGwin G, Scilley K, Girkin CA, Phillips JM, Searcey K. Perceived barriers to care and attitudes about vision and eye care: focus groups with older African Americans and eye care providers. Invest Ophthalmol Vis Sci 2006 Jul 1;47(7):2797. [doi: 10.1167/iovs.06-0107]

XSL•FO
RenderX

36. Lindenmeyer A, Sturt JA, Hipwell A, et al. Influence of primary care practices on patients' uptake of diabetic retinopathy screening: a qualitative case study. Br J Gen Pract 2014 Aug;64(625):e484-e492. [doi: 10.3399/bjgp14X680965] [Medline: 25071061]
37. Hartnett ME, Key IJ, Loyacano NM, Horswell RL, Desalvo KB. Perceived barriers to diabetic eye care: qualitative study of patients and physicians. Arch Ophthalmol 2005 Mar;123(3):387-391. [doi: 10.1001/archopht.123.3.387] [Medline: 15767483]
38. Ellahham S. Artificial intelligence: the future for diabetes care. Am J Med 2020 Aug;133(8):895-900. [doi: 10.1016/j.amjmed.2020.03.033] [Medline: 32325045]
39. Rai A, Riddle M, Mishra R, Nguyen N, Valine K, Fenney M. Use of a smartphone-based medication adherence platform to improve outcomes in uncontrolled type 2 diabetes among veterans: prospective case-crossover study. JMIR Diabetes 2023 Aug 10;8:e44297. [doi: 10.2196/44297] [Medline: 37561555]
40. Ali Sherazi B, Laeer S, Krutisch S, Dabidian A, Schlottau S, Obarcanin E. Functions of mHealth diabetes apps that enable the provision of pharmaceutical care: criteria development and evaluation of popular apps. Int J Environ Res Public Health 2022 Dec 21;20(1):64. [doi: 10.3390/ijerph20010064] [Medline: 36612402]
41. Lopes ROP, Chagas SR, Gomes EDS, Barbosa JCDA, Silva Í, Brandão MAG. Benchmarking mobile applications for the health of people with diabetes mellitus. Rev Lat Am Enfermagem 2024;32:e4221. [doi: 10.1590/1518-8345.7182.4221] [Medline: 38985044]
42. Kim YW, Han S, Kim KS. Determinants of intention to use digital healthcare service of middle and older users. ISM 2018 Dec 31;19(3):1-23. [doi: 10.52558/ISM.2018.12.19.3.1]
43. Moore K, O'Shea E, Kenny L, et al. Older adults' experiences with using wearable devices: qualitative systematic review and meta-synthesis. JMIR Mhealth Uhealth 2021 Jun 3;9(6):e23832. [doi: 10.2196/23832] [Medline: 34081020]
44. Morey SA, Stuck RE, Chong AW, Barg-Walkow LH, Mitzner TL, Rogers WA. Mobile health apps: improving usability for older adult users. Ergonomics in Design: The Quarterly of Human Factors Applications 2019 Oct;27(4):4-13. [doi: 10.1177/1064804619840731]
45. Czaja SJ, Boot WR, Charness N, Rogers WA. Designing for Older Adults: Principles and Creative Human Factors Approaches, 3rd edition: CRC Press/Routledge/Taylor & Francis Group; 2019.
46. LaMonica HM, Davenport TA, Roberts AE, Hickie IB. Understanding technology preferences and requirements for health information technologies designed to improve and maintain the mental health and well-being of older adults: participatory design study. JMIR Aging 2021 Jan 6;4(1):e21461. [doi: 10.2196/21461] [Medline: 33404509]
47. Pinna F, Suprani F, Deiana V, et al. Depression in diabetic patients: what is the link with eating disorders? results of a study in a representative sample of patients with type 1 diabetes. Front Psychiatry 2022;13:848031. [doi: 10.3389/fpsyt.2022.848031] [Medline: 35782445]
48. Marcus MD, Gandica R, El ghormli L, et al. Longitudinal association of depressive symptoms, binge eating, and quality of life with cardiovascular risk factors in young adults with youth-onset type 2 diabetes: the TODAY2 study. Diabetes Care 2022 May 1;45(5):1073-1081. [doi: 10.2337/dc21-1995]
49. Cunningham PJ. Beyond parity: primary care physicians' perspectives on access to mental health care. Health Aff (Millwood) 2009;28(3):w490-w501. [doi: 10.1377/hlthaff.28.3.w490] [Medline: 19366722]
50. Andrews G, Cuijpers P, Craske MG, McEvoy P, Titov N. Computer therapy for the anxiety and depressive disorders is effective, acceptable and practical health care: a meta-analysis. PLoS ONE 2010 Oct 13;5(10):e13196. [doi: 10.1371/journal.pone.0013196] [Medline: 20967242]
51. Matcham F, Rayner L, Hutton J, Monk A, Steel C, Hotopf M. Self-help interventions for symptoms of depression, anxiety and psychological distress in patients with physical illnesses: a systematic review and meta-analysis. Clin Psychol Rev 2014 Mar;34(2):141-157. [doi: 10.1016/j.cpr.2014.01.005] [Medline: 24508685]
52. Baldwin PA, Sanatkar S, Clarke J, et al. A web-based mental health intervention to improve social and occupational functioning in adults with type 2 diabetes (the springboard trial): 12-month outcomes of a randomized controlled trial. J Med Internet Res 2020 Dec 1;22(12):e16729. [doi: 10.2196/16729] [Medline: 33258790]
53. van Bastelaar KMP, Pouwer F, Cuijpers P, Riper H, Snoek FJ. Web-based depression treatment for type 1 and type 2 diabetic patients: a randomized, controlled trial. Diabetes Care 2011 Feb;34(2):320-325. [doi: 10.2337/dc10-1248] [Medline: 21216855]
54. Newby J, Robins L, Wilhelm K, et al. Web-based cognitive behavior therapy for depression in people with diabetes mellitus: a randomized controlled trial. J Med Internet Res 2017 May 15;19(5):e157. [doi: 10.2196/jmir.7274] [Medline: 28506956]
55. Clarke J, Proudfoot J, Ma H. Mobile phone and web-based cognitive behavior therapy for depressive symptoms and mental health comorbidities in people living with diabetes: results of a feasibility study. JMIR Ment Health 2016 May 31;3(2):e23. [doi: 10.2196/mental.5131] [Medline: 27245948]
56. Koetsenruijter J, van Lieshout J, Lionis C, et al. Social support and health in diabetes patients: an observational study in six european countries in an era of austerity. PLoS ONE 2015;10(8):e0135079. [doi: 10.1371/journal.pone.0135079] [Medline: 26305559]
57. Miller TA, Dimatteo MR. Importance of family/social support and impact on adherence to diabetic therapy. Diabetes Metab Syndr Obes 2013 Nov 6;6:421-426. [doi: 10.2147/DMSO.S36368] [Medline: 24232691]

XSL•FO
RenderX

58. Karimy M, Koohestani HR, Araban M. The association between attitude, self-efficacy, and social support and adherence to diabetes self-care behavior. Diabetol Metab Syndr 2018;10:86. [doi: 10.1186/s13098-018-0386-6] [Medline: 30534204]
59. Toobert DJ, Glasgow RE. Assessing diabetes self-management: the summary of diabetes self-care activities questionnaire. In: Handbook of Psychology and Diabetes: A Guide to Psychological Measurement in Diabetes Research and Practice: Harwood Academic Publishers/Gordon; 1994:351-375.
60. Marquez B, Anderson A, Wing RR, et al. The relationship of social support with treatment adherence and weight loss in Latinos with type 2 diabetes. Obesity (Silver Spring) 2016 Mar;24(3):568-575. [doi: 10.1002/oby.21382] [Medline: 26833676]
61. Shayeghian Z, Aguilar-Vafaie ME, Besharat MA, et al. Self-care activities and glycated haemoglobin in Iranian patients with type 2 diabetes: can coping styles and social support have a buffering role? Psychol Health 2015;30(2):153-164. [doi: 10.1080/08870446.2014.951651] [Medline: 25104430]
62. Pereira MG, Berg-Cross L, Almeida P, Machado JC. Impact of family environment and support on adherence, metabolic control, and quality of life in adolescents with diabetes. Int J Behav Med 2008;15(3):187-193. [doi: 10.1080/10705500802222436] [Medline: 18696312]
63. Hilliard ME, Sparling KM, Hitchcock J, Oser TK, Hood KK. The emerging diabetes online community. Curr Diabetes Rev 2015;11(4):261-272. [doi: 10.2174/1573399811666150421123448] [Medline: 25901500]
64. Bujanda-Sainz de Murieta A, Soto-Ruiz N, García-Vivar C, San Martín-Rodríguez L, Escalada-Hernández P. Use of online communities among people with type 2 diabetes: a scoping review. Curr Diab Rep 2024 May;24(5):96-107. [doi: 10.1007/s11892-024-01538-2] [Medline: 38457015]
65. Litchfield I, Barrett T, Hamilton-Shield J, et al. Current evidence for designing self-management support for underserved populations: an integrative review using the example of diabetes. Int J Equity Health 2023 Sep 11;22(1):188. [doi: 10.1186/s12939-023-01976-6] [Medline: 37697302]
66. Klemme I, Wrona KJ, de Jong IM, Dockweiler C, Aschentrup L, Albrecht J. Integration of the vision of people with diabetes into the development process to improve self-management via diabetes apps: qualitative interview study. JMIR Diabetes 2023 Apr 27;8:e38474. [doi: 10.2196/38474] [Medline: 37104003]
67. Burke SD, Sherr D, Lipman RD. Partnering with diabetes educators to improve patient outcomes. Diabetes Metab Syndr Obes 2014;7:45-53. [doi: 10.2147/DMSO.S40036] [Medline: 24550679]
68. Tseng HM, Liao SF, Wen YP, Chuang YJ. Stages of change concept of the transtheoretical model for healthy eating links health literacy and diabetes knowledge to glycemic control in people with type 2 diabetes. Prim Care Diabetes 2017 Feb;11(1):29-36. [doi: 10.1016/j.pcd.2016.08.005] [Medline: 27595215]
69. Guo XM, Zhai X, Hou BR. Adequacy of health literacy and its effect on diabetes self-management: a meta-analysis. Aust J Prim Health 2020 Dec;26(6):458-465. [doi: 10.1071/PY20079] [Medline: 33296622]
70. McCall T, Threats M, Pillai M, Lakdawala A, Bolton CS. Recommendations for design of a mobile application to support management of anxiety and depression among Black American women. Front Digit Health 2022;4:1028408. [doi: 10.3389/fdgth.2022.1028408] [Medline: 36620185]
71. Eke R, Jannat KT, Yang XT, Parton JM. Disparity and factors associated with internet health information seeking among us adults living with diabetes mellitus: cross-sectional study. J Med Internet Res 2022 May 30;24(5):e32723. [doi: 10.2196/32723] [Medline: 35635741]
72. Wang SY, Yeh HC, Stein AA, Miller ER 3rd. Use of health information technology by adults with diabetes in the United States: cross-sectional analysis of national health interview survey data (2016-2018). JMIR Diabetes 2022 Jan 12;7(1):e27220. [doi: 10.2196/27220] [Medline: 35019844]
73. Kickbusch I, Maag D, Wait S, Alliance for Health and the Future, International Longevity Centre - UK. Navigating Health: The Role of Health Literacy: Alliance for Health and the Future, International Longevity Centre-UK; 2005. [Medline: 101297211]
74. Norman CD, Skinner HA. eHEALS: the eHealth literacy scale. J Med Internet Res 2006 Nov 14;8(4):e27. [doi: 10.2196/jmir.8.4.e27] [Medline: 17213046]
75. Kim KA, Kim YJ, Choi M. Association of electronic health literacy with health-promoting behaviors in patients with type 2 diabetes: a cross-sectional study. Comput Inform Nurs 2018 Sep;36(9):438-447. [doi: 10.1097/CIN.0000000000000438] [Medline: 29742548]
76. Marciano L, Camerini AL, Schulz PJ. The role of health literacy in diabetes knowledge, self-care, and glycemic control: a meta-analysis. J Gen Intern Med 2019 Jun;34(6):1007-1017. [doi: 10.1007/s11606-019-04832-y] [Medline: 30877457]
77. Nam-Gung HK, Kim IH, Chun H. Study on the correlates of digital disparity among older Seoul residents. Journal of Digital Convergence 2017 Apr 28;15(4):73-81. [doi: 10.14400/JDC.2017.15.4.73]
78. Arcury TA, Sandberg JC, Melius KP, et al. Older adult internet use and eHealth literacy. J Appl Gerontol 2020 Feb;39(2):141-150. [doi: 10.1177/0733464818807468] [Medline: 30353776]
79. Dil N, Castiglioni A, Kim K, Aravind N, Torre D. Use and implementation of focus groups. Med Teach 2024 Mar;46(3):317-319. [doi: 10.1080/0142159X.2024.2304451] [Medline: 38271556]
80. Stalmeijer RE, McNaughton N, Van Mook WNKA. Using focus groups in medical education research: AMEE Guide No. 91. Med Teach 2014 Nov;36(11):923-939. [doi: 10.3109/0142159X.2014.917165]

XSL•FO
RenderX

81. Leung FH, Savithiri R. Spotlight on focus groups. Can Fam Physician 2009 Feb;55(2):218-219. [Medline: 19221085]
82. Althubaiti A. Information bias in health research: definition, pitfalls, and adjustment methods. J Multidiscip Healthc 2016;9:211-217. [doi: 10.2147/JMDH.S104807] [Medline: 27217764]

Abbreviations

DR: diabetic retinopathy
HbA_{1c}: Hemoglobin A_{1c}
IRB: institutional review board
NAACP: National Association for the Advancement of Colored People
PCP: primary care physician
SEEN: Sight-Saving Engagement and Evaluation in New Haven

Edited by A Mavragani; submitted 02.09.24; peer-reviewed by WQ Chen; revised version received 23.02.25; accepted 25.02.25; published 13.06.25.

Please cite as:
Frimpong A, Granados A, Chang T, Fu J, Moore SG, Applebaum S, Adepoju B, Kaur M, Hari Krishnan V, Levi A, McCall T, Nwanyanwu KH
Recommendations for Designing a Digital Health Tool for Blindness Prevention Among High-Risk Diabetic Retinopathy Patients: Qualitative Focus Group Study of Adults With Diabetes
JMIR Form Res 2025;9:e65893
URL: https://formative.jmir.org/2025/1/e65893
doi:10.2196/65893

© Akua Frimpong, Alvaro Granados, Thomas Chang, Julia Fu, Shannan G Moore, Serina Applebaum, Bolatito Adepoju, Mahima Kaur, Vignesh Hari Krishnan, Amanda Levi, Terika McCall, Kristen Harris Nwanyanwu. Originally published in JMIR Formative Research (https://formative.jmir.org), 13.6.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Feasibility of Data Collection Via Consumer-Grade Wearable Devices in Adolescent Student Athletes: Prospective Longitudinal Cohort Study

Danielle Ransom[1,2], PsyD; Brant Tudor[1,2], PhD; Sarah Irani[1], MD; Mohamed Rehman[1,2], MD; Stacy Suskauer[2,3], MD; P Patrick Mularoni[1,2], MD; Luis Ahumada[1,2], PhD, MSCS

[1]Johns Hopkins All Children's Hospital, 880 Sixth Street South, Suite 460, St. Petersburg, FL, United States

[2]Johns Hopkins University School of Medicine, Baltimore, MD, United States

[3]Kennedy Krieger Institute, Baltimore, United States

Corresponding Author:
Danielle Ransom, PsyD
Johns Hopkins All Children's Hospital, 880 Sixth Street South, Suite 460, St. Petersburg, FL, United States

Abstract

Background: Recent advancements in sports medicine have been fueled by innovative technologies, particularly consumer-grade wearable devices like Fitbit, Apple Watch, and Garmin. These devices offer physiological and biomechanical data and hold promise for personalized, real-time, and remote assessment of athlete recovery. However, few studies have been conducted with these devices in adolescent student athletes.

Objective: The primary objective of this study was to assess the feasibility of integrating consumer-grade wearable technology into injury recovery monitoring of adolescent student athletes.

Methods: The study included 34 high school student athletes aged 14‑18 diagnosed with either concussion or orthopedic injury, enrolled within 10 days of injury. Participants were equipped with a Fitbit Sense for continuous monitoring of physiological markers, including cardiovascular metrics, physical activity levels, and sleep patterns. Data collection extended 4‑6 weeks beyond injury clearance, during which adherence rates were assessed at both hourly and daily intervals. Hourly adherence was defined as the proportion of participants with at least 1 recorded heart rate data point per hour, while daily adherence was defined as the proportion of participants with at least 1 recorded heart rate data point per 24-hour period.

Results: The study demonstrated high participant adherence to wearing the device. The orthopedic injury cohort exhibited a median adherence rate of 95%, with individual rates ranging from 82% to 100%. Similarly, the concussion cohort demonstrated a median adherence rate of 93%, with adherence rates spanning from 37% to 100%. Notably, the study encountered minimal issues related to device functionality, with only 1 participant necessitating a device replacement.

Conclusions: These findings demonstrate successful integration of wearable technology in data collection for adolescent student athletes recovering from sports-related injuries. However, it is important to consider current limitations, including factors that may influence data accuracy and precision. In conclusion, this feasibility study demonstrates the practicality of using consumer-grade wearable technology for the collection of physiological and biomechanical parameters in adolescent student athletes recovering from sport-related injuries. The high level of adherence highlights the potential applicability of consumer-grade wearable devices in this population. Study findings lay the foundation for future investigations with larger and more diverse cohorts to identify the utility of device metrics in identifying unique patterns of injury-specific recovery (ie, sport-related concussion). Consumer-grade wearable devices offer promise for optimizing assessment and management of injured athletes through wearable technology integration into standard clinical protocols.

(JMIR Form Res 2025;9:e54630) doi:10.2196/54630

KEYWORDS

sports-related injuries; wearable technology; recovery monitoring; adolescent student athletes; concussion; orthopedic injuries; orthopedic; musculoskeletal issues; sport; sports; injury; injuries; athlete; athletes; athletics; recover; monitor; monitoring; students; sensors; Fitbit; data collection; wearable; wearables; adolescent; adolescents; teen; teens; teenager; teenagers

XSL•FO
RenderX

Introduction

Common sport-related injuries, including concussion and orthopedic trauma, have the potential to present significant challenges to athletes beyond restricted participation in sports [1-3]. Concussion, in particular, is a growing concern due to the acute impact of recovery on activities of daily living [4-6]. Orthopedic injuries, on the other hand, can disrupt an athlete's training regimen and potentially lead to long-term musculoskeletal issues if not properly managed. Prolonged recovery from injuries of any type has the potential to contribute to heightened symptoms of depression and anxiety compared with uninjured athletes [2,7].

Complications during recovery from concussion and orthopedic injuries are often minimized with appropriate intervention and management [5,7]. However, disparities exist in access to evidence-based care, with up to 30% of surveyed individuals in the United States reporting that they did not obtain health care for a sports-related injury [8]. Unlike collegiate and professional athletes who are often provided with acute care within a network of affiliated health care specialists, access to optimal care pathways is less clear for adolescent student athletes following sport-related injury [9,10]. Youth rely on the availability of parents or other caregivers to identify, coordinate, and provide transportation to specialists for injury monitoring and rehabilitation while co-managing their academic load and meeting the expectation for full-time school attendance [11]. Student athletes who are unable to access care or adhere to follow-up rehabilitation following sport-related injury are at higher risk for prolonged recovery and residual injury effects on physical and mental health [2,7,12]. Given these challenges, it is imperative to investigate novel, minimally invasive, clinically viable, and readily accessible approaches—such as remote monitoring and data collection in the child's typical environment—to effectively support adolescent student athlete injury recovery.

The field of sports medicine has made notable advancements in recent years through innovative technologies with the potential to improve the assessment and management of injured athletes [13,14]. Among these technologies, consumer-grade wearable devices such as Fitbit, Apple Watch, and Garmin have garnered considerable attention for their ability to provide continuous, real-time physiological and biomechanical data in order to provide comprehensive monitoring of individual athletes' health and performance [15,16]. The integration of these devices with standard clinical protocols offers a promising solution to the challenges of recovery monitoring [17-20]. Consumer-grade wearable devices provide a noninvasive and user-friendly means to collect a multitude of physiological and biomechanical parameters, including heart rate, activity levels, sleep patterns, and movement patterns [18]. Data gleaned through these devices could enable health care providers, coaches, and athletes themselves to make informed decisions regarding training intensity, physiologic adaptation, and overall well-being [21]. However, integrating wearable technology in health care research is not without challenges. Issues such as data accuracy, participant adherence, ethical considerations, and the interpretability of data must be carefully navigated [17,18,21]. Moreover, the diverse nature of injuries and individual variations among athletes necessitate a nuanced approach that considers the unique presentation of each patient [22].

This feasibility study aimed to assess the practicality of integrating consumer-grade wearable technology into the recovery monitoring of adolescent student athletes through an evaluation of device usability and participant adherence across commonly measured physiological and biomechanical parameters, including cardiovascular metrics, physical activity levels, and sleep patterns.

Methods

Participants

The study sample is comprised of 2 age- and sex-matched cohorts of high school student athletes aged 14 - 18 years who were actively participating in their sports season at the time of enrollment (September 2021-June 2022). One cohort included male and female athletes participating in sports who were diagnosed with concussion, as defined by the 2016 Berlin International Consensus Conference for Concussion in Sport [23]. The second cohort consisted of athletes diagnosed with an acute, nonsurgical orthopedic injury. Eligible participants were evaluated and diagnosed with a concussion or acute, nonsurgical orthopedic injury within 10 days of their injury in an ambulatory sports medicine clinic; had access to a reliable internet connection and cell phone or home computer; and were able to read, understand, and comply with study instructions. Participants were additionally required to attend 2 separate study neurocognitive screening visits for study inclusion.

Participants were excluded from the study if they met any of the following criteria: prior traumatic brain injury within the past 6 months; history of moderate to severe traumatic brain injury (Glasgow Coma Scale<13) at any time; brain mass, prior neurosurgery, or central nervous system disorder; moderate to severe cognitive dysfunction or structural brain disease or malformation; or participation in more than 1 sport during the time of enrollment. Additionally, if a participant had not yet recovered from their injury, the study team discontinued protocol participation 180 days post enrollment.

Ethical Considerations

This study was approved by the Johns Hopkins All Children's Hospital institutional review board (IRB00457018 for data collection; IRB00211758 for data analysis and reporting). Informed consent was obtained from participants and their legally authorized representatives, and participation was voluntary. To ensure privacy and confidentiality, data were deidentified and transmitted securely to the Fitbit cloud-based database via a HIPAA (Health Insurance Portability and Accountability Act)-compliant protocol, allowing access to deidentified data through a code link on the Fitabase platform. Participants received compensation, including a Fitbit Sense, which they kept after study participation ended; two US $50 gift cards distributed at enrollment and upon completion of the study protocol; and 6 volunteer hours that could be applied toward high school service requirements.

XSL•FO
RenderX

Wearable Device

The Fitbit Sense (Fitbit Inc) is a wrist-worn wearable device designed for continuous monitoring of various physiological parameters, such as heart rate, resting heart rate, step counts, daily minutes of vigorous activity, and sleep patterns and architecture. This wearable device was chosen because of its long battery life (up to 6 d) and its durability, including waterproofing. The Fitbit Sense is equipped with a range of sensors, including a photoplethysmogram sensor, an optical heart rate sensor, an electrodermal activity sensor, an accelerometer, a gyroscope, and an ambient light sensor. The photoplethysmogram sensor uses green and red light-emitting diodes along with a photodiode to measure blood volume changes in the microvasculature of the skin, allowing for the estimation of heart rate and blood oxygen saturation. The optical heart rate sensor uses a combination of green and infrared light-emitting diodes to capture heart rate data through the measurement of blood flow variations. To detect motion, the accelerometer and gyroscope work in tandem to monitor physical activity and provide information on movement patterns. The accelerometer quantifies linear motion, enabling accurate tracking of steps, distance, and intensity of physical activity. The intensity and duration of these activities are recorded and used to estimate metrics like energy expenditure, active minutes, and overall physical activity levels. The gyroscope measures angular velocity, facilitating the detection of activities like cycling, swimming, and rotational movements. To detect sleep, the Fitbit Sense uses a combination of the accelerometer and heart rate data to discern different sleep stages, including sleep onset and wakefulness, as well as light, deep, and rapid eye movement sleep.

Data Collection

After obtaining written consent from participants and their legally authorized representatives, the study research assistant provided each participant with a Fitbit Sense, along with a device tutorial covering full-time wear, daily syncing with the Fitbit application, and standardized charging times (twice weekly during evening downtime at home). The study team assisted participants with downloading the Fitbit smartphone application, creating an account linked to a study-generated deidentified email address managed by the team, and connecting their provided Fitbit Sense to a smartphone or compatible computer via Bluetooth. To prevent the collection of identifiable protected health information, the study team disabled GPS functionality by default and instructed participants not to alter this setting on their device or the smartphone app for the duration of the study. Once the Fitbit Sense was connected, participants were asked to wear the device for as much of their day as possible, including sleep. Patient-generated data were then transmitted to the Fitbit cloud-based database through the internet-connected application. Fitbit then connected to the data management platform, Fitabase (Small Steps Labs LLC), via HIPAA-compliant protocol to allow investigators to access the deidentified data via code link. The study team monitored the frequency of application syncing and contacted participants via phone or SMS text message to provide troubleshooting if devices were synced less than twice weekly. Data points were gathered across 3 key parameters: cardiovascular, physical activity, and sleep. Cardiovascular data included heart rate (beats/minute) and resting heart rate. Physical activity data included daily step count and daily minutes of physical activity, categorized into vigorous, moderate, light, and sedentary exertion levels. Sleep data included total daily minutes spent asleep, minutes of nighttime wakefulness, and sleep architecture characterization based on time spent per night spent in deep, light, wakeful, and rapid eye movement states.

As part of standard clinical care, participants were monitored by a sports medicine physician in clinic every 1 - 2 weeks until clearance criteria were met. Criteria for concussion clearance included clinical examination by a specialized health care provider to confirm resolution of acute symptoms, return to preinjury symptom levels during rest and activity, tolerance for cognitive and physical exertion, and successful completion of a supervised gradual return-to-play protocol. Length of concussion recovery was defined as the number of days between the date of injury and the date the participant received medical clearance to begin a gradual return-to-play protocol based upon symptom resolution at rest and return-to-learn [24]. Criteria for orthopedic clearance included clinical examination by a sports medicine physician to confirm injury stabilization, satisfactory range of motion and strength, appropriate weight-bearing status (if applicable), demonstrated ability to perform functional activities without significant impairment, well-managed pain levels, absence of indications for immediate surgery, and medical clearance confirming fitness for sport participation. Length of orthopedic recovery was defined as the number of days between the date of injury and the date the participant received medical clearance to participate in sport. Participant adherence was continuously monitored from enrollment to approximately 4 - 6 weeks after meeting injury clearance criteria. At that point, the study team initiated completion procedures, which included transferring the Fitbit account to the participant and discontinuing data tracking. If a participant had not yet recovered from their injury, protocol discontinuation was initiated by the study team 180 days post enrollment.

Statistical Analysis

Statistical analyses were performed using the *Pandas* and *NumPy* Python packages (Python Software Foundation). Continuous variables were summarized with means, SDs, quantiles, and ranges (minimum to maximum). Categorical variables were summarized with counts and percentages. Adherence to device use, the primary feasibility outcome, was assessed through both hourly and daily granularity, based on the presence of recorded heart rate data. *Hourly adherence* was defined as the proportion of participants with at least 1 recorded heart rate data point per hour. If heart rate data were collected during a given hour, a subject was deemed to have complied for that hour. Presented here are the averages for said hour over every day in each 2-week period from enrollment until study completion. For example, with 336 hours in a 2-week period, an overall adherence rate of 1 would indicate that all actively participating subjects in the cohort wore their (functioning) device for at least a portion of each of the 336 hours that comprised that 2-week period. Here, day was defined as 7 AM to 6:59 PM while night was defined as 7 PM to 6:59 AM according to each participant's local time zone. *Daily adherence*

XSL•FO
RenderX

was defined as the proportion of participants with at least 1 recorded heart rate data point per 24-hour period. Patterns of data capture revealed high initial adherence, with missed data points increasing over time. Extending the previous example, a daily adherence rate of 1 for a given 2-week period would indicate that the subject's Fitbit recorded at least 1 heart rate data point during each of the 14 days occurring in the period. If a study participant completed the study during a given 2-week period, a daily adherence rate of 1 would indicate that the subject's Fitbit recorded at least 1 heart rate during each of the days in which they were actively enrolled in the study (eg, 5/5 d). To evaluate feasibility over time, adherence metrics were analyzed across study participation intervals (ie, days [1-28] and [29-41]).

Results

Sample Characteristics

A total of 38 participants were enrolled in the study between September 2021 and April 2022. In total, 4 participants were excluded from analyses due to nonadherence with study procedures, including failure to attend the required second study visit (n=2, orthopedic injury cohort), failure to wear the Fitbit device (n=1, orthopedic injury cohort), and lack of medical clearance by 180 days post enrollment (n=1, concussion cohort). The final study sample consisted of 34 participants (n=17 concussion cohort; n=17, orthopedic injury cohort).

Participants were evenly distributed by sex (female: 61.8%) and matched by age, with a mean age of 15.71 (SD 1.15) years. Participants engaged in a variety of sports, with 71% of the concussion cohort and 58% of the orthopedic injury cohort participating in contact and collision sports. The concussion cohort had a higher proportion of student athletes participating in contact and collision sports (71%) compared with the orthopedic injury cohort (58%). The majority of participants were White and non-Hispanic. Demographic characteristics are presented in Table 1.

XSL•FO
RenderX

Table . Demographic and injury characteristics of adolescent student athlete cohorts enrolled in this prospective longitudinal study.

	Concussion (n=17)	Orthopedic injury (n=17)	Overall (N=34)
Age (years), mean (SD)	15.76 (1.11)	15.65 (1.19)	15.71 (1.15)
Gender, n (%)			
Male	7 (41.2)	6 (35.3)	13 (38.2)
Female	10 (58.8)	11 (64.7)	21 (61.8)
School grade, n (%)			
9	6 (33.3)	5 (29.4)	11 (32.4)
10	3 (16.7)	6 (35.3)	9 (26.5)
11	5 (27.8)	5 (29.4)	10 (29.4)
12	3 (16.7)	1 (5.9)	4 (11.8)
Race, n (%)			
White	14 (82.4)	9 (52.9)	23 (67.6)
Black	1 (5.9)	6 (35.3)	7 (20.6)
Other	2 (11.8)	2 (11.8)	4 (11.8)
Ethnicity, n (%)			
Hispanic	2 (11.8)	5 (29.4)	7 (20.6)
Non-Hispanic	15 (88.2)	12 (70.6)	27 (79.4)
Sport of injury, n (%)			
Soccer	5 (29.4)	6 (35.3)	11 (32.4)
Football	5 (29.4)	1 (5.9)	6 (17.6)
Basketball	1 (5.9)	2 (11.8)	3 (8.8)
Softball	2 (11.8)	1 (5.9)	3 (8.8)
Track	0 (0)	3 (17.6)	3 (8.8)
Volleyball	2 (11.8)	1 (5.9)	3 (8.8)
Baseball	0 (0)	2 (11.8)	2 (5.9)
Cheerleading	1 (5.9)	0 (0)	1 (2.9)
Flag football	0 (0)	1 (5.9)	1 (2.9)
Rugby	1 (5.9)	0 (0)	1 (2.9)
Recovery (days), mean (SD)	25.64 (18.1)	32.12 (20.48)	--

Adherence to Device Use

Overall Trends

Adherence to device wear was examined at both hourly and daily intervals across the study period. Median daily adherence rates were high across cohorts, with a 93% (range 37% - 100%) adherence rate in the concussion cohort and a 95% (range 82% - 100%) adherence rate in the orthopedic injury cohort. The average study duration was 66 days for the concussion cohort and 61 days for the orthopedic injury cohort. The majority of participants (67.6%) actively contributed data for 33 - 61 days (Table 2 and Figure 1). No significant issues with device connectivity or functionality were reported during the study period; however, 1 participant lost the device and required a replacement.

XSL•FO
RenderX

Table . Daily adherence rates to wearable device monitoring by study interval across concussion and orthopedic injury cohorts.

Patient number	Study duration (days)	Total days reporting	Overall daily adherence rate	Daily adherence rate			
				Days 1 - 14	Days 15 - 28	Days 29 - 42	Days 43 - 56
Concussion cohort							
1	52	35	0.67	0.93	0.71	0.36	0.70
2	59	22	0.37	0.93	0.57	0.00	0.00
3	75	75	1	1	1	1	1
4	28	24	0.86	1	0.71	—	—
5	33	28	0.85	0.93	0.71	1	—
6	39	30	0.77	0.86	0.93	0.45	—
7	72	61	0.85	1	0.71	1	1
8	42	42	1	1	1	1	—
9	56	56	1	1	1	1	1
10	154	149	0.97	1	1	1	1
11	74	52	0.70	1	0.36	0.43	0.79
12	49	47	0.96	1	1	0.86	1
13	180	155	0.86	1	0.93	0.79	1
14	43	40	0.93	0.93	0.86	1	1
15	61	59	0.97	1	0.93	1	0.93
16	49	47	0.96	1	0.93	0.93	1
17	56	54	0.96	1	1	0.86	1
Overall	66.0	57.4	0.86	0.98	0.84	0.79	0.88
Orthopedic injury cohort							
1	35	34	0.97	1	1	0.86	—
2	99	82	0.83	1	1	1	0.00
3	59	56	0.95	1	1	1	0.79
4	53	52	0.98	1	1	1	0.91
5	50	43	0.86	1	1	0.50	1
6	95	79	0.83	1	1	1	1
7	47	47	1	1	1	1	1
8	49	48	0.98	1	1	0.93	1
9	118	114	0.97	1	1	1	1
10	56	46	0.82	1	1	0.71	0.57
11	40	39	0.97	1	1	0.92	—
12	40	37	0.93	1	0.86	0.92	—
13	61	57	0.93	1	1	1	0.79
14	81	79	0.98	1	1	0.93	0.93
15	64	56	0.88	1	1	0.79	0.86
16	54	50	0.93	0.93	1	0.86	0.92
17	33	33	1	1	1	1	—
Overall	60.8	56.0	0.93	1.00	0.99	0.91	0.83

XSL•FO
RenderX

Figure 1. Duration of study participation and daily device adherence in concussion and orthopedic injury cohorts.

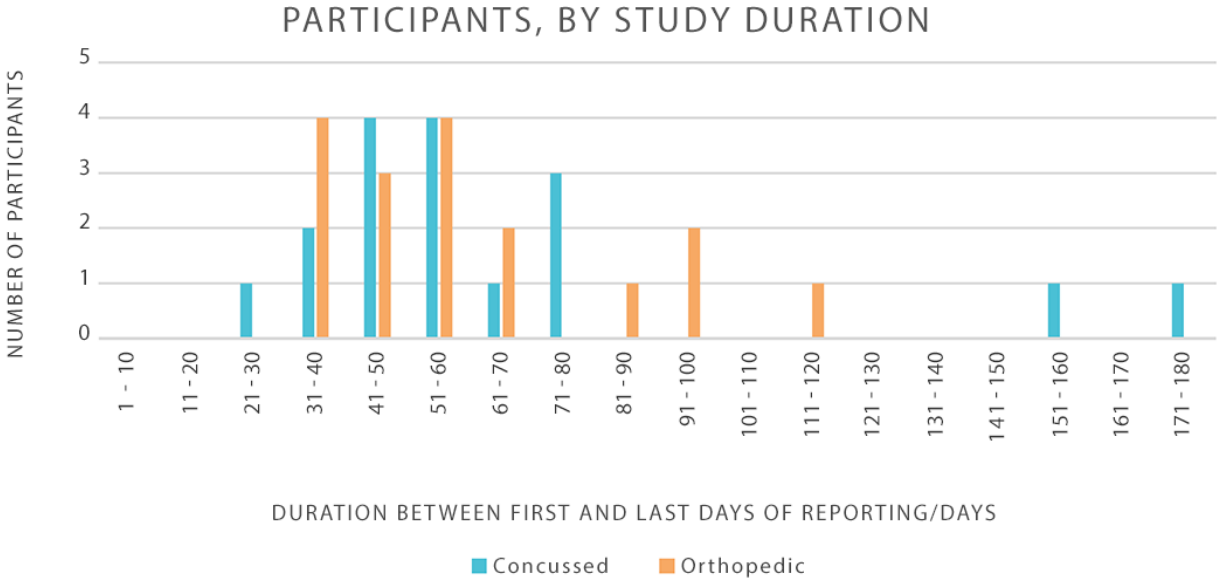

Adherence Over Time

Daily adherence remained high across time intervals, though some variability was observed as participants completed the study at different time points based on individual recovery timelines. Adherence was highest in the initial 2 weeks, with median rates of 98% in the concussion cohort and 100% in the orthopedic injury cohort (Table 2). During the first 2 weeks of participation (days 1 - 14), nearly all participants maintained near-perfect adherence, with rates of 98% in the concussion cohort and 100% in the orthopedic injury cohort. In the following 2-week period (days 15 - 28), adherence remained high (84% concussion, 99% orthopedic injury), though a slight decline was observed as some participants completed the study upon meeting clearance criteria and were no longer enrolled. By days 29 - 42, adherence rates showed a moderate decrease (79% concussion, 91% orthopedic injury), again reflecting study completion rather than disengagement. In the final observed time interval (days 43 - 56), fewer participants remained enrolled due to individual recovery timelines, but those still actively participating sustained strong adherence, with rates of 88% in the concussion cohort and 83% in the orthopedic injury cohort. Rather than disengaging, participants who met clearance criteria exited the study, leading to expected decreases in the number of active participants over time.

Day Versus Night Adherence

Adherence rates were comparable between daytime and nighttime periods (Figure 2), suggesting that participants were comfortable wearing the device overnight and willing to sustain compliance outside of structured daily activities. No significant differences were observed between cohorts.

Figure 2. Hourly adherence rates by time of day (day vs night) and injury cohorts.

		Overall	Day	Night	Day												Night											
					7:00	8:00	9:00	10:00	11:00	12:00	13:00	14:00	15:00	16:00	17:00	18:00	19:00	20:00	21:00	22:00	23:00	0:00	1:00	2:00	3:00	4:00	5:00	6:00
Days 1-14	Concussed	0.88	0.88	0.89	0.84	0.85	0.85	0.86	0.88	0.89	0.90	0.89	0.88	0.89	0.90	0.91	0.92	0.90	0.90	0.91	0.90	0.92	0.91	0.87	0.87	0.86	0.86	0.85
	Orthopedic	0.90	0.91	0.89	0.87	0.87	0.87	0.90	0.91	0.91	0.90	0.95	0.95	0.93	0.95	0.93	0.91	0.91	0.91	0.90	0.91	0.89	0.86	0.89	0.88	0.89	0.89	0.88
Days 15-28	Concussed	0.73	0.74	0.72	0.70	0.71	0.72	0.75	0.77	0.78	0.78	0.76	0.73	0.74	0.74	0.76	0.75	0.75	0.72	0.73	0.71	0.71	0.72	0.72	0.72	0.72	0.70	0.70
	Orthopedic	0.86	0.88	0.85	0.84	0.85	0.85	0.87	0.89	0.87	0.88	0.89	0.89	0.88	0.88	0.88	0.89	0.85	0.86	0.82	0.82	0.82	0.83	0.82	0.85	0.87	0.86	0.85
Days 29-42	Concussed	0.63	0.64	0.61	0.59	0.59	0.59	0.61	0.65	0.67	0.68	0.66	0.66	0.66	0.68	0.67	0.64	0.65	0.65	0.61	0.59	0.60	0.57	0.64	0.59	0.61	0.60	0.60
	Orthopedic	0.71	0.72	0.70	0.64	0.63	0.63	0.68	0.75	0.76	0.75	0.76	0.77	0.76	0.76	0.78	0.77	0.75	0.73	0.73	0.71	0.72	0.70	0.71	0.70	0.66	0.64	0.64
Days 43-56	Concussed	0.72	0.75	0.69	0.72	0.72	0.74	0.74	0.73	0.75	0.75	0.75	0.77	0.75	0.75	0.75	0.65	0.64	0.64	0.61	0.63	0.72	0.75	0.76	0.73	0.72	0.73	0.72
	Orthopedic	0.63	0.64	0.61	0.55	0.57	0.57	0.62	0.65	0.67	0.65	0.67	0.67	0.66	0.68	0.69	0.65	0.62	0.60	0.62	0.62	0.62	0.62	0.63	0.61	0.60	0.58	0.57

0.00 — 1.00

Data Capture and Physiological Monitoring

Participant-generated data provided continuous, real-world monitoring of cardiovascular metrics, physical activity levels, and sleep patterns across the study period. Data completeness was highest in the early weeks, with gradual declines over time as participants completed the study upon recovery clearance. Physiological data collected across time intervals are summarized in Figure 3, demonstrating the range and variability of parameters measured.

XSL•FO
RenderX

Figure 3. Heat map of physiological parameters collected longitudinally via Fitbit Sense during injury recovery. REM: rapid eye movement.

	Mean			Standard Deviation			Minimum			Maximum		
	Days 1-14	Days 15-28	Days 29-42	Days 1-14	Days 15-28	Days 29-42	Days 1-14	Days 15-28	Days 29-42	Days 1-14	Days 15-28	Days 29-42
Resting Heart Rate	63.8	65.8	65.1	7.4	7.2	7.1	49.0	50.3	48.5	88.0	82.0	82.1
Daily Steps	7516.4	8013.8	7669.1	4761.2	5066.4	4560.3	6	5	4	23949	29411	22679
Daily Very Active Minutes	20.2	22.5	21.3	40.3	39.4	43.9	0	0	0	278	256	451
Daily Moderately Active Minutes	22.1	23.2	20.1	32.2	32.9	32.6	0	0	0	213	243	250
Daily Light Activity Minutes	223.8	214.5	210.6	106.3	108.0	118.5	0	0	0	525	570	528
Daily Sedentary Minutes	622.5	647.4	709.4	390.9	406.8	438.9	5	4	7	1440	1440	1440
Minutes Deep Sleep	82.5	81.5	83.9	27.0	24.9	25.9	0	13	16	155	150	176
Minutes Light Sleep	235.0	229.2	237.9	64.6	63.1	55.5	90	83	84	516	478	424
Minutes REM Sleep	92.3	93.8	93.7	31.8	34.5	33.6	6	5	12	208	185	227
Minutes Total Sleep	410.0	404.5	415.5	94.4	90.2	81.2	168	180	177	747	764	660
Minutes Awake During Sleep Event	59.0	56.9	58.9	20.2	19.3	18.3	13	16	14	135	142	126

Min Max

Discussion

Principal Findings

This feasibility study demonstrated the practicality of integrating consumer-grade wearable technology for the collection of physiological and biomechanical parameters in a pilot sample of adolescent student athletes recovering from sport-related injuries. High adherence rates across both cohorts highlight the potential for real-world implementation of wearable devices in clinical and research settings. Factors such as long battery life and ease of use likely contributed to sustained adherence, while the active engagement of student athletes in their sport season may have further motivated participation [25-27].

As expected, daily adherence declined, and greater variability was observed between cohorts as the study progressed, highlighting important considerations for future longitudinal study designs. Specifically, the observed differences in daily adherence rates between the 2 cohorts suggest varying levels of participant adherence with the device in the latter two-thirds of the study. Of note, this time period may have coincided with a gradual return to sport participation. Student athletes with orthopedic injuries demonstrated consistently higher rates of daily adherence compared to those with concussion, indicating a greater willingness or ability to adhere to wearing the Fitbit. For example, most contact and collision sports limit arm and wrist accessories during practice and competition due to safety precautions. If a student athlete was required to remove their device for practice or game play, they may have been less likely to remember to put it back on than they would have for standard charging occurring at home. The slightly lower adherence rates observed for student athletes recovering from concussion may be attributed to the higher number of student athletes from contact and collision sports included in this cohort compared with the orthopedic injury cohort. This suggests that wearable technology may be more readily embraced by certain patient populations and may be influenced by factors such as perceived relevance to their recovery process, sport-specific rules and regulations, or motivational factors. This finding highlights the importance of considering individual student athlete characteristics and needs when implementing such devices in study protocols. Strategies to enhance participant adherence, such as sport-specific procedures for device usage upon return to play and real-time feedback from the research team after a set interval of nonadherence, may be necessary to optimize data capture in studies with adolescent student athletes. Additional domain-specific considerations are presented below.

Cardiovascular Monitoring

Consumer-grade wearable devices offer promise for health care research through real-time monitoring of cardiovascular parameters. This feasibility study provides support for future investigations using these devices for cardiovascular monitoring following sport-related injuries. Variability in resting heart rate over time observed in these data may reflect the impact of limited activity acutely post injury, followed by a transient increase due to heightened sympathetic activity and subsequent adaptation associated with autonomic balance and cardiovascular efficiency as participants returned to sport [28,29]. However, it is important to acknowledge known limitations, including device placement, individual skin type, skin complexion, and user adherence, may influence the accuracy and reliability of the collected data [18]. Continuous heart rate monitoring may not capture transient physiological changes during intense, short-duration activities [30-32]. While these limitations must be considered when interpreting the data, access to continuous cardiovascular patterns provided by consumer-grade wearable devices remains a valuable tool for future study protocols aimed at tailoring rehabilitation interventions, adjusting exercise regimens, and assessing overall recovery progress [33].

Physical Activity Monitoring

This study highlights the broad feasibility of using wearable technology to comprehensively monitor physical activity in adolescent student athletes. Physical activity data, including step counts and exertion levels, gathered from wearable devices have the potential to provide nuanced understanding of participants' activity patterns, which is essential for fine-tuning rehabilitation strategies [34,35]. Tracking changes in step counts over time can serve as a valuable indicator of progress in

XSL•FO
RenderX

regaining mobility and function post injury [36]. However, there are limitations consistent with those discussed within the cardiovascular parameters. The accuracy and precision of consumer-grade wearable devices may vary based on factors such as device calibration, placement, and individual or sport-specific characteristics, which can contribute to discrepancies in the recorded data [16]. In addition, wearable devices may not accurately capture certain activities, particularly those involving nonstandard movements or activities that do not primarily rely on step-based motion [37]. Although participants demonstrated high levels of broad adherence to device use during this study, the consistency in step counts and activity levels over time is unexpected, given that injured student athletes would presumably demonstrate an increase in exertion over time as they returned to sport. As noted earlier, sport-specific rules and regulations may have limited use during practice and game play, which may be reflected in the separate and incongruent finding of increasing sedentary minutes over time. Since physical activity metrics gleaned through the wearable device relied on cardiovascular exertion and step counts, nonuse of the device during primary sport participation could have limited data capture. Of note, resting heart rate is less likely to be impacted by this limitation since data are gathered at night. Monitoring activity levels using wearable devices may offer important insights into the rehabilitation process following sport-related injury but may be limited by adherence to sport-specific regulations during activity participation and should be considered by investigators when developing future study protocols using these devices.

Sleep Monitoring

Sleep data capture using a consumer-grade wearable device may provide valuable insights into participants' sleep patterns during injury recovery. Understanding the quality and duration of sleep is integral to assessing the restorative processes crucial for injury recovery, particularly for athletes recovering from concussion [38,39]. Findings from this study support the feasibility of collecting a range of sleep parameters that could impact injury recovery, including time spent asleep versus awake at night and specific sleep architecture patterns. However, it is imperative to acknowledge limitations associated with gathering sleep data via consumer-grade wearable devices. While this technology has made significant strides in sleep tracking, accuracy may still be influenced by factors such as device positioning, individual sleep habits, and variations in sleep stages [40]. Devices may not capture certain nuances of sleep architecture that are crucial for a comprehensive understanding of sleep quality [40]. Factors like sleep disorders or environmental disturbances may also introduce variability in the recorded data. It is important to note that while wearable-based sleep monitoring yields useful insights, it does not provide the same level of detail as specialized polysomnography assessments, which remain the gold standard in sleep research [40]. The noninvasive approach of consumer-grade devices, however, provides an accessible tool for identifying general trends in sleep pattern disruption that holds promise in health care research [18]. To gain a more comprehensive understanding of sleep quality, complementing wearable device research with information from gold standard clinical assessments is recommended [41]. Recognizing these limitations, future research using consumer-grade wearable devices for sleep monitoring may contribute to interventions aimed at optimizing sleep hygiene and could potentially lead to improved overall recovery outcomes.

Study Limitations and Future Directions

While this pilot study demonstrates significant promise, several considerations and future directions merit attention. The relatively small sample size, the specific focus on adolescent student athletes, and the observational nature of this study limit the generalizability of the findings. A planned follow-up study will explore within- and between-group comparisons, as these were beyond the scope of this feasibility study. In addition, expanding the participant pool to include a broader age range and diverse athletic backgrounds would strengthen the external validity of study findings, as would matching cohorts by sport and including a healthy cohort of uninjured athletes. Additionally, longer-term investigations could shed light on the trajectory of recovery over extended periods and provide insights into the evolving role of consumer-grade wearable technology trends in the rehabilitation process. As noted earlier, the increasing use of consumer-grade wearable technology in athletes, the improving accuracy of devices over time, and the potential benefits of integration into sports medicine research prompt the need for studies to explore the utility of these devices in clinical research. Addressing issues such as data accuracy, participant adherence, and ethical considerations (eg, data security, privacy and confidentiality, and equity in device access) will be crucial in refining methodologies for future research endeavors.

Conclusions

In conclusion, this study highlights that consumer-grade wearable technology is feasible for use in research with adolescent student athletes. These findings, along with their associated limitations, can inform the development of larger scale trials and, eventually, evidence-based protocols that can guide health care providers, coaches, and athletes in leveraging wearable technology to optimize injury recovery. The successful collection of physiological and biomechanical data through an easily accessible and consumer-friendly device lays the foundation for future investigations to explore trends in patient populations of interest with larger and more diverse cohorts. As wearable technology continues to advance and gain popularity, its integration into sports medicine research holds promise for improving the assessment and management of injured athletes.

XSL•FO
RenderX

Acknowledgments

This work was supported in part by a research grant from the Kart 4 Kids Concussion Initiative through the Johns Hopkins All Children's Foundation.

We thank the students and staff at Admiral Farragut Academy and Indian Rocks Christian School for their support of this project. We would also like to thank Mrs Katelyn Heinz for her support with enrollment and the study protocol. We thank Julia Nahman for providing psychometrist support for this study, as well as Dr Lauren Irwin Harper for administering neuropsychological measures. Finally, we thank the Hough Family Foundation and Susan and Robb Hough for supporting our endeavors.

Data Availability

The datasets generated during and analyzed during this study are not publicly available due to restrictions imposed by the institutional review board to protect participant privacy but are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

References

1. Ransom DM, Vaughan CG, Pratson L, Sady MD, McGill CA, Gioia GA. Academic effects of concussion in children and adolescents. Pediatrics 2015 Jun;135(6):1043-1050. [doi: 10.1542/peds.2014-3434] [Medline: 25963014]
2. Chang CJ, Putukian M, Aerni G, et al. Mental health issues and psychological factors in athletes: detection, management, effect on performance, and prevention: American Medical Society for Sports Medicine position statement. Clin J Sport Med 2020 Mar;30(2):e61-e87. [doi: 10.1097/JSM.0000000000000817] [Medline: 32000169]
3. Valovich McLeod TC, Snedden T, Halstead M, et al. Influence of personal and injury-related factors predicting deficits in quality of life domains among pediatric athletes: findings from the sport concussion outcomes in pediatrics study. Clin J Sport Med 2023 Sep 1;33(5):489-496. [doi: 10.1097/JSM.0000000000001140] [Medline: 36858431]
4. Lumba-Brown A, Yeates KO, Sarmiento K, et al. Centers for Disease Control and Prevention Guideline on the diagnosis and management of mild traumatic brain injury among children. JAMA Pediatr 2018 Nov 1;172(11):e182853. [doi: 10.1001/jamapediatrics.2018.2853] [Medline: 30193284]
5. Patricios JS, Schneider KJ, Dvorak J, et al. Consensus statement on concussion in sport: the 6th International Conference on Concussion in Sport–Amsterdam, October 2022. Br J Sports Med 2022;57(11):695-711. [doi: 10.1136/bjsports-2023-107127]
6. Nahman JC, Staples C, Irwin Harper LN, et al. Retracted article: psychological resilience, emotional symptoms, and recovery duration in adolescent sport-related concussion. Appl Neuropsychol Child 2024 Feb 17:1-7. [doi: 10.1080/21622965.2024.2317307] [Medline: 38367962]
7. Maffulli N, Longo UG, Gougoulias N, Loppini M, Denaro V. Long-term health outcomes of youth sports injuries. Br J Sports Med 2010 Jan;44(1):21-25. [doi: 10.1136/bjsm.2009.069526] [Medline: 19952376]
8. Gutierrez G, Sills M, Bublitz CD, Westfall JM. Sports-related injuries in the United States: who gets care and who does not. Clin J Sport Med 2006 Mar;16(2):136-141. [doi: 10.1097/00042752-200603000-00009] [Medline: 16603883]
9. Pierpoint LA, LaBella CR, Collins CL, Fields SK, Dawn Comstock R. Injuries in girls' soccer and basketball: a comparison of high schools with and without athletic trainers. Inj Epidemiol 2018 Jul 16;5(1):29. [doi: 10.1186/s40621-018-0159-6] [Medline: 30009338]
10. Wingerson MJ, Magliato SN, Smulligan KL, Wilson JC, Little CC, Howell DR. Predicting time to evaluation after pediatric concussion: factors affecting specialty concussion care. Orthop J Sports Med 2023 Aug;11(8):23259671231186430. [doi: 10.1177/23259671231186430] [Medline: 37655238]
11. Ransom DM, Caldwell CT, DiVirgilio EK, et al. Pediatric mTBI during the COVID-19 pandemic: considerations for evaluation and management. Child Neuropsychol 2022 Apr;28(3):355-373. [doi: 10.1080/09297049.2021.1985101] [Medline: 34615434]
12. Schmidt JD, Broglio SP, Knight K, et al. Optimizing concussion care seeking: a longitudinal analysis of recovery. Am J Sports Med 2023 Jan;51(1):214-224. [doi: 10.1177/03635465221135771] [Medline: 36412549]
13. Verhagen E, Bolling C. Protecting the health of the @hlete: how online technology may aid our common goal to prevent injury and illness in sport. Br J Sports Med 2015 Sep;49(18):1174-1178. [doi: 10.1136/bjsports-2014-094322] [Medline: 25614537]
14. Hayashi D, Roemer FW, Tol JL, et al. Emerging quantitative imaging techniques in sports medicine. Radiology 2023 Aug;308(2):e221531. [doi: 10.1148/radiol.221531] [Medline: 37552087]
15. Helmer P, Hottenrott S, Rodemers P, et al. Accuracy and systematic biases of heart rate measurements by consumer-grade fitness trackers in postoperative patients: prospective clinical trial. J Med Internet Res 2022 Dec 30;24(12):e42359. [doi: 10.2196/42359] [Medline: 36583938]

XSL•FO
RenderX

16. Roos LG, Slavich GM. Wearable technologies for health research: opportunities, limitations, and practical and conceptual considerations. Brain Behav Immun 2023 Oct;113:444-452. [doi: 10.1016/j.bbi.2023.08.008] [Medline: 37557962]
17. Henriksen A, Haugen Mikalsen M, Woldaregay AZ, et al. Using fitness trackers and smartwatches to measure physical activity in research: analysis of consumer wrist-worn wearables. J Med Internet Res 2018 Mar 22;20(3):e110. [doi: 10.2196/jmir.9157] [Medline: 29567635]
18. Shei RJ, Holder IG, Oumsang AS, Paris BA, Paris HL. Wearable activity trackers-advanced technology or advanced marketing? Eur J Appl Physiol 2022 Sep;122(9):1975-1990. [doi: 10.1007/s00421-022-04951-1] [Medline: 35445837]
19. Cucchiaro G, Ahumada L, Gray G, et al. Feasibility of conducting long-term health and behaviors follow-up in adolescents: longitudinal observational study. JMIR Form Res 2022 Aug 15;6(8):e37054. [doi: 10.2196/37054] [Medline: 35969442]
20. Lonsdale H, Jalali A, Ahumada L, Matava C. Machine learning and artificial intelligence in pediatric research: current state, future prospects, and examples in perioperative and critical care. J Pediatr 2020 Jun;221S:S3-S10. [doi: 10.1016/j.jpeds.2020.02.039] [Medline: 32482232]
21. Benson LC, Räisänen AM, Volkova VG, Pasanen K, Emery CA. Workload a-WEAR-ness: monitoring workload in team sports with wearable technology. a scoping review. J Orthop Sports Phys Ther 2020 Oct;50(10):549-563. [doi: 10.2519/jospt.2020.9753] [Medline: 32998615]
22. Burnham JP, Lu C, Yaeger LH, Bailey TC, Kollef MH. Using wearable technology to predict health outcomes: a literature review. J Am Med Inform Assoc 2018 Sep 1;25(9):1221-1227. [doi: 10.1093/jamia/ocy082] [Medline: 29982520]
23. McCrory P, Meeuwisse W, Dvorak J, et al. Consensus statement on concussion in sport—the 5th international conference on concussion in sport held in Berlin, October 2016. Br J Sports Med 2016 Oct(51):838-847. [doi: 10.1136/bjsports-2017-097699]
24. Putukian M, Purcell L, Schneider KJ, et al. Clinical recovery from concussion-return to school and sport: a systematic review and meta-analysis. Br J Sports Med 2023 Jun;57(12):798-809. [doi: 10.1136/bjsports-2022-106682] [Medline: 37316183]
25. Beukenhorst AL, Howells K, Cook L, et al. Engagement and participant experiences with consumer smartwatches for health research: longitudinal, observational feasibility study. JMIR Mhealth Uhealth 2020 Jan 29;8(1):e14368. [doi: 10.2196/14368] [Medline: 32012078]
26. Keogh A, Dorn JF, Walsh L, Calvo F, Caulfield B. Comparing the usability and acceptability of wearable sensors among older Irish adults in a real-world context: observational study. JMIR Mhealth Uhealth 2020 Apr 20;8(4):e15704. [doi: 10.2196/15704] [Medline: 32310149]
27. Donner E, Devinsky O, Friedman D. Wearable digital health technology for epilepsy. N Engl J Med 2024 Feb 22;390(8):736-745. [doi: 10.1056/NEJMra2301913] [Medline: 38381676]
28. Garcia GL, Porto LGG, da Cruz CJG, Molina GE. Can resting heart rate explain the heart rate and parasympathetic responses during rest, exercise, and recovery? PLoS One 2022;17(12):e0277848. [doi: 10.1371/journal.pone.0277848] [Medline: 36477005]
29. Gonzales TI, Jeon JY, Lindsay T, et al. Resting heart rate is a population-level biomarker of cardiorespiratory fitness: the Fenland Study. PLoS One 2023;18(5):e0285272. [doi: 10.1371/journal.pone.0285272] [Medline: 37167327]
30. Gillinov S, Etiwy M, Wang R, et al. Variable accuracy of wearable heart rate monitors during aerobic exercise. Med Sci Sports Exerc 2017 Aug;49(8):1697-1703. [doi: 10.1249/MSS.0000000000001284] [Medline: 28709155]
31. Cheung CC, Krahn AD, Andrade JG. The emerging role of wearable technologies in detection of arrhythmia. Can J Cardiol 2018 Aug;34(8):1083-1087. [doi: 10.1016/j.cjca.2018.05.003] [Medline: 30049358]
32. Al-Kaisey AM, Koshy AN, Ha FJ, et al. Accuracy of wrist-worn heart rate monitors for rate control assessment in atrial fibrillation. Int J Cardiol 2020 Feb 1;300:161-164. [doi: 10.1016/j.ijcard.2019.11.120] [Medline: 31787389]
33. Fletcher GF, Ades PA, Kligfield P, et al. Exercise standards for testing and training: a scientific statement from the American Heart Association. Circulation 2013 Aug 20;128(8):873-934. [doi: 10.1161/CIR.0b013e31829b5b44] [Medline: 23877260]
34. Lauersen JB, Bertelsen DM, Andersen LB. The effectiveness of exercise interventions to prevent sports injuries: a systematic review and meta-analysis of randomised controlled trials. Br J Sports Med 2014 Jun;48(11):871-877. [doi: 10.1136/bjsports-2013-092538] [Medline: 24100287]
35. Leddy JJ, Burma JS, Toomey CM, et al. Rest and exercise early after sport-related concussion: a systematic review and meta-analysis. Br J Sports Med 2023 Jun;57(12):762-770. [doi: 10.1136/bjsports-2022-106676] [Medline: 37316185]
36. Fallahzadeh R, Verdonk F, Ganio E, et al. Objective activity parameters track patient-specific physical recovery trajectories after surgery and link with individual preoperative immune states. Ann Surg 2023 Mar 1;277(3):e503-e512. [doi: 10.1097/SLA.0000000000005250] [Medline: 35129529]
37. Fuller D, Colwell E, Low J, et al. Reliability and validity of commercially available wearable devices for measuring steps, energy expenditure, and heart rate: systematic review. JMIR Mhealth Uhealth 2020 Sep 8;8(9):e18694. [doi: 10.2196/18694] [Medline: 32897239]
38. Bramley H, Henson A, Lewis MM, Kong L, Stetter C, Silvis M. Sleep disturbance following concussion is a risk factor for a prolonged recovery. Clin Pediatr (Phila) 2017 Dec;56(14):1280-1285. [doi: 10.1177/0009922816681603] [Medline: 29073787]

XSL•FO
RenderX

39. Charest J, Grandner MA. Sleep and athletic performance: impacts on physical performance, mental performance, injury risk and recovery, and mental health: an update. Sleep Med Clin 2022 Jun;17(2):263-282. [doi: 10.1016/j.jsmc.2022.03.006] [Medline: 35659079]
40. de Zambotti M, Cellini N, Goldstone A, Colrain IM, Baker FC. Wearable sleep technology in clinical and research settings. Med Sci Sports Exerc 2019 Jul;51(7):1538-1557. [doi: 10.1249/MSS.0000000000001947] [Medline: 30789439]
41. Menghini L, Cellini N, Goldstone A, Baker FC, de Zambotti M. A standardized framework for testing the performance of sleep-tracking technology: step-by-step guidelines and open-source code. Sleep 2021 Feb 12;44(2):zsaa170. [doi: 10.1093/sleep/zsaa170] [Medline: 32882005]

Abbreviations

HIPAA: Health Insurance Portability and Accountability Act

Edited by A Mavragani; submitted 16.11.23; peer-reviewed by A Khudair, Y Adesida; revised version received 27.02.24; accepted 29.02.24; published 13.06.25.

Please cite as:
Ransom D, Tudor B, Irani S, Rehman M, Suskauer S, Mularoni PP, Ahumada L
Feasibility of Data Collection Via Consumer-Grade Wearable Devices in Adolescent Student Athletes: Prospective Longitudinal Cohort Study
JMIR Form Res 2025;9:e54630
URL: https://formative.jmir.org/2025/1/e54630
doi:10.2196/54630

© Danielle Ransom, Brant Tudor, Sarah Irani, Mohamed Rehman, Stacy Suskauer, P Patrick Mularoni, Luis Ahumada. Originally published in JMIR Formative Research (https://formative.jmir.org), 13.6.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Patient and Public Perceptions of 3D Technologies (Models and Images) to Facilitate Health Care Consultations: Exploratory, Mixed Methods Study

Harleen Kaur Rai[1], BSc, MSc, PhD; Morven Miller[1], BA, MSc, PhD; Steve Leung[2], MB ChB, FRCSEd (Urol), DM; Euan Macleod[3], RMN; Marilyn Lennon[1], BSc, PhD

[1]Department of Computer and Information Sciences, University of Strathclyde, 16 Richmond Street, Glasgow, United Kingdom

[2]Urology Department, Western General Hospital, Edinburgh, United Kingdom

[3]Directorate for Strategic Planning and Transformation, Dumfries and Galloway NHS, Dumfries, United Kingdom

Corresponding Author:
Morven Miller, BA, MSc, PhD
Department of Computer and Information Sciences, University of Strathclyde, 16 Richmond Street, Glasgow, United Kingdom

Abstract

Background: 3D technology, including models and images, can facilitate health care consultations by promoting a better understanding of information by patients and shared decision-making. However, little is yet known about the general public's perspectives about the acceptability of such innovative technology and how it can best be adopted into routine health care consultations. There is a need to explore both public and patient perceptions to avoid the risk of implementing 3D technologies that may not be acceptable or fit-for-purpose.

Objective: This paper aimed to explore the patient and public perceptions of the use of 3D technology during health care consultations.

Methods: This study adopted a citizen science approach using mixed methods to conduct (1) a short web-based survey with members of the public to gather a wide range of opinions regarding the use of various technologies for health care consultations; (2) a longer web-based survey to explore perceived barriers and opportunities people report specifically on the use of 3D technology; and (3) telephone interviews with patients who recently used 3D technology as part of their health care consultations.

Results: A total of 211 participants completed the short survey, of which 25 went on to complete the longer survey. While members of the public were familiar with using various types of technologies during remote consultations, most participants did not have experience with using 3D technology. However, people reported that they could see the potential benefits of such technology to facilitate health care consultations. They expressed positive perceptions toward how this might assist in comprehension of a diagnosis and discussion of alternative treatment plans. They also mentioned potential benefits in relation to communication and shared decision-making either with their health care provider or with their friends and family. These potential benefits were confirmed through telephone interviews with 4 patients who also stressed potential barriers such as emotional distress caused by an overload of information as important considerations for wider implementation. Overall, there was a strong interest and willingness to use 3D technology in future health care consultations.

Conclusions: The use of 3D technology in health care settings is now an option, but there is little research to date on how patients and the wider public might benefit from this. This mixed methods study has shown that people are accepting of 3D technology being used in health care consultations and that there might be real benefits to the patient. These include improved individual and shared decision-making around their treatment through the technology, making disease and treatment options easier to understand for patients. Since 3D technology can still be expensive, the benefits to the patient and health care professionals need to be captured and quantified in terms of reduced travel, efficient use of time, and overall better quality of care and clinical outcomes.

(JMIR Form Res 2025;9:e65235) doi:10.2196/65235

KEYWORDS

3D technology; citizen science; public and patient involvement; health care; public perceptions; health care consultations; consultations; mixed methods study; information; shared decision-making; acceptability; innovative technology; web-based survey; public health; telephone interviews; willingness; qualitative study

XSL•FO
RenderX

Introduction

Patients are often given information in many forms to encourage them to be involved in the decision-making process about their own health. Particularly in surgical disciplines, diagnostic imaging information such as computerized tomography, magnetic resonance imaging, or radiograph (x-ray) scans can be used to illustrate the patient's anatomy and condition [1]. These types of images or information can be hard for most patients to fully understand, especially in cases of low levels of health literacy [2]. The use of more advanced digital technologies might be helpful for sharing this complex information with patients and aiding their overall understanding of their condition and treatment options [3]. This implementation of digital technologies also fits within the recent shift toward remote or virtual consultations in light of the COVID-19 pandemic, a change that arguably has a significant impact on doctor-patient interactions [4]. Many specialties undertook new strategies during this time, such as making further use of telephone clinics and videoconferencing platforms such as Teams or NHS Near Me to conduct clinical meetings [4,5]. There is a clear opportunity to use new technologies, but only if they are acceptable to patients and provide advantages above the "standard" face-to-face consultation in the clinical setting. Clinicians often use 2D images to aid patient discussions of their disease. With the expanding roll-out of 5G telecommunications; however, 3D technologies such as images, models (physical replicas of the disease), and holographic displays are now emerging and, if successful, could significantly improve the consultation experience for patients, carers, and clinicians by reducing costs, travel and waiting times and contributing to the green agenda [6,7]. The latter would allow for expert care to be delivered closer to the patient's home in line with the Scottish Government's delivery plan for Care in the Digital Age [8]. 3D technologies can often be used without a headset or specialized virtual reality eyewear meaning that the patient and clinician are not inhibited by cumbersome virtual reality equipment offering a far less physically and socially restricting environment to view 3D visualizations during a consultation. 3D technologies can convey information of the disease to the patient in novel ways, for example, a 3D image viewed on a monitor can create a depth of field making a complex object more understandable (see Figure 1).

Figure 1. Example of a 3D health care model showcasing detailed views of the skeletal and organ system [9]. Used with permission from Holoxica.

By presenting the information in a novel way, clinicians can effectively explain complex procedures, highlight risk and support the discussion of treatment options potentially leading to a level of joint understanding of the disease state and helping the patient make choices about their future management [2]. In the field of renal cancer surgery for example, Wake et al [2] demonstrated that patient understanding, and education improves with the addition of 3D models in a consultation. Several 3D imaging methods were used but 3D printed models were noted to have the greatest impact on patient understanding of their disease, tumor characteristics and the surgery itself. The ability to see the tumor in 3D can also make the patient feel more involved in their treatment and leads to greater trust in the clinician and their surrounding team [10]. Consequently, 3D models can play a great part in strengthening the shared decision-making process that is rapidly becoming the cornerstone of modern medicine [11,12]. For clinicians, the use of 3D technology provides an additional benefit: it may improve surgical outcomes by allowing surgeons to plan and rehearse their surgery on personalized models of their patients' pathology. This may have the benefits of shorter operative time, more accurate surgery leading to higher rate of curative surgery, and reduced blood loss leading to quicker recovery [13].

With advancing technologies come huge opportunities to enhance the way the public interacts with professionals and health and care systems. However, while there is some research demonstrating the benefits of 3D technologies from a patient perspective, there is not much known about the public's readiness to adopt such technologies as part of standard care. For 3D technologies to be fit-for-purpose and implemented more widely, there is a need to understand what the general public's perceptions and expectations are from these technologies. Without this, there is a risk of adopting technologies which will exclude people rather than improve their interactions and their health outcomes. Therefore, the aim of this study was to explore not just patient but also wider public perceptions of the use of 3D technologies including images and models during health care consultations.

XSL•FO
RenderX

Methods

Design

This was a mixed methods study using surveys and telephone interviews. We adopted a citizen science approach where we set out to actively include voices and experiences of the public as well as people who have had experience of 3D technology in their recent health care interactions. Giving people a chance to provide their lived experiences is important for increased buy in, adoption and scaling of technologies in the future. A real-world service evaluation approach guided this study to gather a range of opinions on an emerging technology and how it might be adopted and received by people in their day to day lives. The aim was not to evaluate or validate the technology itself but to explore whether technology such as this was acceptable and so could be adopted into routine care consultations—specifically from the patients' perspective of potential benefits and barriers. Such evaluations of acceptance prior to adoption and scale are essential to test for change to maximize the chances of the technology working in practice and to avoid human factors issues that might arise if effective technologies are simply rolled out without prior consultation with the public and patients.

Participants

Different user groups were recruited for the various stages of this study. Participants from the public aged 18 years or older were invited to complete a short web-based survey. Participants who had completed the short survey were then asked to share their email address if they were interested in completing a longer survey to elaborate on their previous responses. Telephone interviews were conducted with participants who were aged 18 years or older, who had lived experience of receiving a diagnosis and treatment for kidney cancer and had experience with using 3D technology as part of their care consultations.

Ethical Considerations

Ethical approval was obtained from the University of Strathclyde's Department of Computer and Information Science Ethical Committee (ref: 1797). All participants were provided with an information sheet, included as part of the web-based surveys or via email if participating in telephone interviews. Written or verbal informed consent was obtained from all participants prior to taking part in the surveys or the telephone interviews. Survey participants had the option to share their email addresses, for opting into the long survey. No other personally identifiable data were collected. Data from the telephone interviews were deidentified during analysis. Participants were not compensated for their involvement, but a £50 (US $67.94) gift voucher was awarded to 5 people who completed the short survey and 10 people for the long survey using a digital random number generator.

Recruitment

A convenience sampling approach was used to recruit participants from the public. Invitations to participate in the short survey (see Multimedia Appendix 1) were advertised via social media including personal Twitter accounts of members of the research team as well as sent directly to various charities and carer groups (including Action Kidney Cancer, Fibromyalgia Action UK, Myeloma UK, MacMillan Cancer support, and Carers Scotland). Recruitment for the long survey was facilitated by participants opting in after completing the short survey. Both surveys were completed in the web-based survey tool named Qualtrics. Recruitment for the telephone interviews took place in a primary care setting. A Consultant Urological Surgeon, who was part of the research team, assessed recent patients against the study's eligibility criteria and invited eligible patients from his clinics with lived experiences of a kidney cancer diagnosis to participate in a short telephone interview to evaluate the service. Patients were asked to describe their experiences with 3D technology (eg, images or models) which had been used during their consultation.

Data Collection Short Survey (Survey 1)

A survey including 4 short multiple choice and 4 open-ended questions (see Multimedia Appendix 2) was advertised digitally and shared with third sector organizations to gather a wider range of opinions regarding the use of various technologies (including 3D technology) for health care consultations. The aim was to gather opinions from a more diverse range of people with a range of needs, preferences, and demography. The survey was available for completion for 5 months (March to August 2022).

Long Survey (Survey 2)

A longer survey including a mix of 17 multiple choice and open-ended questions was distributed to people opting in from the short survey. Participants were invited to elaborate on previous responses and to share their stories with us in terms of their experiences of and attitudes toward new 3D technology for health care consultations. The survey was available for completion for one month (September to October 2022).

Case Studies With People With Lived Experience

Four patients with lived experience of kidney cancer surgery participated in individual, telephone-based interviews to explore their actual experiences of using 3D technology such as images or models during consultations for diagnosis or treatment. The research team developed an interview topic guide (see Multimedia Appendix 3) including 5 open-ended questions related to general experience of using 3D technology, benefits, barriers, and future use for such technology. Questions were followed up with probes where appropriate (eg, specific impacts). Each interview was audio-recorded and lasted between 10 and 20 minutes.

Data Analysis

Numerical data from the surveys was presented descriptively using the inbuilt descriptive summary reporting tools in Qualtrics. This included the frequency of responses in each category. Free text responses were examined by 2 members of the researcher team (HKR, MM) to thematically categorize respondents' perceptions. The free text analysis was mostly deductive in nature as our a priori themes were pre-defined by the questions in our survey which were centered around our research aims. Any emerging themes that were additional to our a priori themes were included as additional perceptions. Data from the telephone interviews were not transcribed, rather

XSL•FO
RenderX

the research team (HKR, ML, MM) made extensive notes of the audio-recordings and these were summarized thematically by the lead researcher (HKR).

Results

Quantitative Results From Survey 1 & 2

A total of 253 participants took part in the first, short survey in part of which 211 participants completed it in full. Most participants had experience with taking part in a remote consultation with a health care provider. This was mostly done through phone calls (n=123) or video calls (n=62) (see Table 1).

Table . Results from the short survey showing an overview of remote consultation methods used by members of the public during health care consultations.

Answer	Count, n (%)
Yes—phone call (eg, a landline or your mobile phone)	123 (48.62)
Yes—video call (eg, this includes Skype, Microsoft Teams, Facetime, WhatsApp, Zoom)	62 (24.51)
Yes—email (eg, discussing your care with a health professional)	51 (20.16)
Other—please describe below	1 (0.40)
No	16 (6.32)

Both modalities were rated as “quite good” or “extremely good” by most participants in terms of their usability, performance, and helpfulness but overall, video calls were rated more positively than phone calls on all aspects. In relation to different types of 3D technology, 62 participants had heard of it and were interested in its use.

A total of 25 participants from the short survey went on the completed the long survey in full. None of the participants had any direct experience with using 3D technology in their interactions with a health care consultant. Despite this lack of experience, the majority agreed such technology could potentially be useful (n=17), improve understanding of one’s condition (n=13) and treatment options (n=10), help to understand surgical procedures (n=14) and their associated risks (n=14), impact personal decision-making for treatment (n=11), help with sharing information with family and friends (n=13), and impact the relationship with their health care provider (n=13) .

Qualitative Results From Survey 1 & 2

The free-text results of both surveys were reviewed and are presented below according to the following themes: understanding and sharing information; relationships; barriers and concerns; and benefits and opportunities (current and future).

Understanding and Sharing Information

Within survey 1, respondents acknowledged the usefulness and helpfulness of all technologies (telephone, video call, email, Near Me app) for health and social care consultations. Being able to see the health professional in a video call was felt to promote greater understanding than a simple telephone call. The helpfulness of emails to keep a record of health details and prepare questions to ask of the health professional was noted. Respondents identified specific occasions when technology would be appropriate for sharing information. For example, for routine, non-urgent appointments such as medical reviews, discussing blood results, and when someone is unable to travel. It may also be useful in facilitating conversations to determine whether in person consultations are needed for discussing a change in health status, more serious conditions, sensitive or detailed decisions or mental health issues.

Those who responded to the second survey reported that 3D models would be helpful to increase their understanding of their condition and treatment options primarily because of the increase in clarity and detail of information available. 3D models were seen to make information about a diagnosis more real and relatable. The potential of visualization was also noted in relation to understanding the impact of treatment, for example, comparing between treatments or whether a treatment had been effective. The deeper understanding from 3D models would help people to ask questions about their diagnosis and treatment options, reducing their need to understand medical jargon. Respondents noted that the information from 3D models could help mitigate the potential for misunderstandings as well as being helpful for sharing with family members and friends as, if they understood their condition better, they would be better able to communicate this to others.

Beyond the sharing of information between patients and health professionals within a consultation, the use of 3D models was also seen as a more general educational opportunity for health professionals, patients and for use in educating people in schools.

Relationships

Within survey 1, respondents identified a lack of face-to-face connectedness as a negative outcome of using technology for consultations, in particular losing the subtleties of visual cues and the inability to conduct physical examinations. While emailing a photograph was seen as a workaround by some, it was not thought to be acceptable to others. Within survey 2, the use of 3D models was consistently seen to improve the potential relationship that a patient would have with their health professional by creating a sense of partnership between the two as well as increasing a patient’s confidence and trust in their

XSL•FO
RenderX

health professional. The concept of 3D models helping to overcome the use of medical jargon was raised again in relation to improving relationships between patients and health professionals and this reduced use of jargon was seen to make it easier for patients to ask questions of their health care professional.

Barriers and Concerns

Within survey 1, respondents identified a lack of access to the appropriate equipment and facilities (such as connectivity), as well as how to use them, as a barrier to remote health and social care consultations. Wasted time waiting for a phone or video call was noted as a downside of remote consultations as was waiting for a response to an email.

Within survey 2, some respondents noted that the increased information provided by 3D models may be scary or provide too much information for some people and so increase their sense of anxiety. Visual or cognitive impairment, apathy to technology and advanced age were all identified as factors that may act as barriers to the use of 3D models in practice. Resource and practical issues including associated costs, availability, time taken to produce models and complexity of IT required were also noted to be issues that may constrain the use of 3D models. It was also noted that 3D models would not be helpful for facilitating discussions about all diagnoses (eg, hematological cancers) or treatment options. Another challenge identified was the capability of the technology to be used for remote consultations. The fact that not everyone wants to share this type of information with family members or friends was also highlighted. The environmental impact of producing 3D plastic models was also mentioned. Finally, that care and compassion should not be lost in interactions between patients and health professionals using such technology was an important consideration.

Benefits and Opportunities (Current and Future)

Respondents to survey 1 noted many benefits of technology-facilitated consultations relating to time: saving time in general, being at a convenient time; avoiding the inconveniences and time associated with travelling to appointments; responding to emails at a time that suited them. Remote consultations were seen to be a more convenient and cheaper option than travelling to an appointment. There was an overwhelming sense from respondents to survey 2 that 3D models offered benefits across multiple health care scenarios, for example: diagnosis; presurgical and decision-making consultations. Overall, implementing this technology where possible and appropriate was seen as a positive addition to interactions between patients and health professionals.

Case Studies (Lived Experience)

General Experience

Most participants had experience with both a 3D image and 3D model during successive care consultations. For instance, one participant was shown a 3D image as part of a general consultation and a 3D model was presented at a later stage to provide more details regarding the upcoming surgical procedure. Some of these consultations had to take place remotely due to the COVID-19 pandemic meaning participants were able to see 3D images at home and 3D models were sent by post where needed. An added benefit of these remote consultations included the involvement of family members as they were able to join a Near Me call or could handle the 3D model at home which was perceived to promote shared understanding. However, 2 participants mentioned they preferred seeing a 3D model in a care setting only as they wanted to avoid revisiting issues surrounding their health at their home.

All participants were positive about their experience of using 3D images and 3D models. Most notably, participants felt that such 3D technology provided a better perspective on their diagnosis and as a result felt reassured, less anxious, and more confident fuelled by a better understanding, eg, regarding the size and location of the tumor or treatment procedure:

> *I made my mind up when I'd seen it, that I was not going to die. […] It was very reassuring.* [Interview 1]

Three participants mentioned that conventional scans are not very clear or understandable and that a 3D image on the screen or holding a 3D model were both much clearer:

> *The 3D image that he showed me initially was ten times better than the scan […] it brought everything to life.* [Interview 2]

One participant also mentioned that 'to actually see the image' or 'to feel the model' was better than reading or listening about her condition as she is dyslexic.

Benefits

All participants shared several benefits of which improved understanding of diagnosis and treatment was mentioned most often.

> *I think it takes a lot of the fear away.* [Interview 1]

The technology helped to pull together all the information patients had already received by enabling visualisation of treatment procedures and consequences through images rather than having to understand and remember information that was given verbally or on paper only.

> *Could we say "by enabling me to visualise the procedure and see it presented as images rather than having to remember and try to understand what was said to me."*

This increased understanding also had a positive impact on the relationship with their clinicians, for example, increased trust and confidence. Two participants mentioned they felt that they were receiving more personalized care and were seen as individuals rather than just another patient.

> *They are doing all they can to benefit me.* [Interview 3]

Other benefits included improved information sharing and communication. This was particularly helpful when sharing details surrounding a diagnosis with family members and friends who often themselves were concerned for the participant's well-being. For one participant their own improved understanding helped to understand what was happening to a

XSL•FO
RenderX

family member who was also receiving treatment for kidney cancer. Lastly, when it came to decision-making, for most of the participants, this was not relevant as a decision had been made prior to seeing 3D images and models. However, one participant mentioned that he felt less anxious when making decisions though he could not say for sure whether this was due to a good relationship with his clinician or the 3D technology.

Barriers

While participants did not experience any challenges or difficulties with the 3D images or models, some recognized that some people may feel upset or distressed from being presented with such 3D models. For instance, one participant thought that detailed 3D technology may induce anxiety in some patients due to an overload of detailed information and that improving people's understanding of their diagnosis and/or treatment could lead to increased anxiety as it would 'make things very real'. However, for the same participant this barrier was seen as a benefit as they preferred to be in possession of all the available information:

> *It could give an information overload […] it makes things very real for them but for me this was an absolute benefit.* [Interview 3]

Future Use

When asked about when 3D images or models would be most useful, one participant mentioned they would be helpful at any point during the cancer journey. Others mentioned 3D technology could be most useful when discussing options available for treatment and/or surgery. Some participants felt that it would not be useful at the point of receiving a cancer diagnosis or during a first consultation as patients may need to come to terms with receiving their diagnosis before being presented with additional information through 3D technology as this could be overwhelming.

However, all participants agreed that 3D images and models held a lot of potential and that they should be implemented more widely within health care settings in the future beyond cancer consultations. Some suggestions for future clinical use included showing broken bones, brain injuries, or other types of surgery (not just urology). One participant felt that most patients would like to have a 3D image or model presented to them if it was available and reiterated their usefulness:

> *It's a very positive option to have. I would recommend anybody to do it if they have the option.* [Interview 3]

Discussion

This study set out to explore patient and public perceptions of the use of 3D technologies during health care consultations to further shed light on its potential for adoption as part of standard care. To the best of our knowledge, this is the first work of its kind in which we integrated a citizen science approach with mixed methodology consisting of surveys and case studies. Overall, our findings show a strong interest and willingness to use 3D technology among members of the public which is further supported by positive patient experiences indicating an encouraging potential for 3D technologies to be adopted as part of standard care.

Our results indicate that members of the public were experienced in the use of various technologies during remote consultations. Most people did so through telephone, video calls and emails and found the added convenience of saving time and travel to be the most important benefit. Also, different types of technologies were also associated with different benefits. For instance, video calls were helpful for a greater understanding of information during routine and non-urgent appointments while emails were thought to be useful for keeping a record of health details. Previous research states that the effectiveness of remote consultations and different types of technologies indeed depends on their purpose. For instance, Greenhalgh et al [14] found that remote consultations were perceived to be more appropriate when engaging in more conventional and routine appointments as opposed to appointments for newer and more urgent cases. However, our findings have also highlighted individual barriers including a lack of face-to-face connectedness, appropriate equipment, and facilities as well as a lack of knowledge or skills on how to use these which is backed by previous research [15]. Individual barriers related to health and access may impact the experience of using technologies during remote consultations. In their narrative synthesis, Walthall et al [16] found that internet-based, remote consultations in general practice are more likely to be used by younger, affluent, and educated groups. Researchers stress the importance of considering (health) inequalities in the wider use and implementation of remote consultations and relevant technologies.

The majority of our study participants were also familiar with the existence of 3D technology such as images and displays as part of health care interactions but had no experience of using such technology in practice. However, participants recognised that such technology could have some wide-ranging benefits leading to a positive impact on the overall health care consultation experience. People mentioned an improved understanding and sharing of information regarding their condition and treatment options as one of the main benefits. 3D technology was considered to increase clarity and the level of detail, making the available information more real and relatable. In addition, people also felt that such technology could enhance the relationship with their health care professional through creating a sense of partnership which could lead to increased trust and confidence in the health care professional. These benefits were also confirmed by patients with lived experience. Patients expressed how 3D technology might assist in comprehension of a diagnosis and in discussing alternative treatment plans. They also mentioned potential benefits in communication and shared decision-making either with their health care professional or with their friends and family. It is important to note that there is a large body of literature surrounding patient experiences of remote health care consultations and of more conventional technologies, especially considering the COVID-19 pandemic, but research on the use of 3D technologies in such consultations with patients is still relatively limited. Studies investigating the use of 3D technology from a patient perspective are particularly lacking. A recent

XSL•FO
RenderX

review on the application of 3D technology within health care education highlighted a range of benefits related to optimisation of costs of complex surgical processes as well as reduced time and fewer complexities during operations due to pre-surgical planning using 3D technologies [17]. Similar to our findings, 3D technologies can help to improve the relationship between the patient and health care professional during health care consultations and address any gaps in communication leading to better patient understanding [17]. These benefits are confirmed in an earlier review by Senbekov et al [18] who state that 3D technologies can support patient education and patient-oriented care.

Despite these encouraging benefits, people also felt that for 3D technologies to be implemented widely during health care consultations, some barriers would need to be taken into consideration. For instance, both members of the public and patients with lived experience noted that the increased information provided by 3D technology may lead to an overload of information for some people and so increase their sense of anxiety. Indeed, in an exploratory study with patients with glioma, van de Belt et al [12] found that patients were positive about 3D technologies in terms of improved patient understanding but were also concerned that such technologies could cause emotional distress especially early on in their treatment which was also noted in our research. Members of the public in this study also identified individual barriers such as cognitive impairment and apathy to technology as well as practical and resource barriers including any associated costs and time needed to produce 3D models and their environmental impact that may hinder the wider implementation of 3D technology. As with the use of any technology during remote consultations, people also felt that care and compassion should not be lost in interactions between patients and health care professionals using 3D technology.

This study has several limitations. Given that the surveys were disseminated through selected charities and carers groups, this may have caused selection bias during recruitment leading to our sample lacking full representation of the wider public. Furthermore, we recruited using web-based methods and organized a web-based survey which excluded people without access to the internet further impacting the generalizability of our research findings. In addition, positive experiences regarding both remote consultations in general as well as (the potential of) using 3D technology during such consultations were overrepresented in our data and a more diverse approach to sampling may have helped to capture more nuanced and negative experiences and opinions. However, both the sampling process and the relatively small sample size were appropriate for an exploratory study and allowed for convenient data collection that highlighted potential willingness, positive attitudes and valuable insights toward using 3D technologies during health care consultations that others could find helpful in considering whether the findings may be relevant in their settings. Though we did not capture demographic data in our survey and case studies, this is particularly relevant when considering population groups that face health inequalities (eg, people with lack of access to technology, low levels of literacy, ethnic minority groups) have experiences with remote consultations and 3D technology that are likely to differ. Finally, our case studies did not include detailed transcription and thematic analysis which may have limited the amount of rich and in-depth data we were able to capture.

In conclusion, the use of 3D technology in health care settings is now an option but there is little research to date on patient experiences with such technology and how the wider public might benefit from this. There is an overwhelming sense from our research that 3D technology offers wide-ranging benefits with a potential for application across multiple health care scenarios. People are accepting of 3D technology being used in health care consultations and there may be real benefits in practice including improved patient understanding of diagnosis, treatment options, and risks leading to enhanced individual and shared decision-making as well as a better relationship with the health care professional. More research is needed to better quantify these benefits to the patient and health care professionals but also in terms of cost savings through reduced travel, efficient use of time, and overall better health outcomes. Lastly, with the wider implementation of 3D technologies, barriers surrounding accessibility need to be considered to prevent the exclusion of more vulnerable populations from using 3D technology.

Acknowledgments

We would like to thank the participants for speaking with us freely about their lived experiences and the charities (Action Kidney Cancer, Fibromyalgia Action UK, Myeloma UK, MacMillan Cancer support, and Carers Scotland) for reaching out to their communities with our surveys. We would also like to thank Dumfries and Galloway NHS (National Health Service) and NHS Lothian for supporting this project.

Data Availability

Data from the online surveys are available from the corresponding author (MM) upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1

XSL•FO
RenderX

Recruitment flyer for short survey.
[DOCX File, 199 KB - formative_v9i1e65235_app1.docx]

Multimedia Appendix 2
Short survey.
[DOCX File, 18 KB - formative_v9i1e65235_app2.docx]

Multimedia Appendix 3
Case study interview guide.
[DOCX File, 16 KB - formative_v9i1e65235_app3.docx]

References

1. Elson D, Yang GZ. The principles and role of medical imaging in surgery. In: Athanasiou T, Debas H, Darzi A, editors. Key Topics in Surgical Research and Methodology: Springer; 2010:529-543. [doi: 10.1007/978-3-540-71915-1_39]
2. Wake N, Rosenkrantz AB, Huang R, et al. Patient-specific 3D printed and augmented reality kidney and prostate cancer models: impact on patient education. 3D Print Med 2019 Feb 19;5(1):4. [doi: 10.1186/s41205-019-0041-3] [Medline: 30783869]
3. Norris EC, Halaska C, Sachs PB, Lin CT, Sanfilippo K, Honce JM. Understanding patient experiences, opinions, and actions taken after viewing their own radiology images online: web-based survey. JMIR Form Res 2022 Apr 25;6(4):e29496. [doi: 10.2196/29496] [Medline: 35468086]
4. Khan N, Jones D, Grice A, et al. A brave new world: the new normal for general practice after the COVID-19 pandemic. BJGP Open 2020 Aug;4(3):bjgpopen20X101103. [doi: 10.3399/bjgpopen20X101103] [Medline: 32487520]
5. Evaluation of the Attend Anywhere / Near Me video consulting service in Scotland, 2019-20. Scottish Government. 2020. URL: https://www.gov.scot/binaries/content/documents/govscot/publications/research-and-analysis/2020/07/evaluation-attend-anywhere-near-video-consulting-service-scotland-2019-20-main-report/documents/evaluation-attend-anywhere-near-video-consulting-service-scotland-2019-20/evaluation-attend-anywhere-near-video-consulting-service-scotland-2019-20/govscot%3Adocument/evaluation-attend-anywhere-near-video-consulting-service-scotland-2019-20.pdf [accessed 2025-05-27]
6. Li D. 5G and intelligence medicine-how the next generation of wireless technology will reconstruct healthcare? Precis Clin Med 2019 Dec;2(4):205-208. [doi: 10.1093/pcmedi/pbz020] [Medline: 31886033]
7. Climate emergency & sustainability strategy. Scottish Government. 2022. URL: https://www.gov.scot/binaries/content/documents/govscot/publications/strategy-plan/2022/08/nhs-scotland-climate-emergency-sustainability-strategy-2022-2026/documents/nhs-scotland-climate-emergency-sustainability-strategy-2022-2026/nhs-scotland-climate-emergency-sustainability-strategy-2022-2026/govscot%3Adocument/nhs-scotland-climate-emergency-sustainability-strategy-2022-2026.pdf [accessed 2025-05-27]
8. Care in the digital age: delivery plan 2023-24. Scottish Government. 2023. URL: https://www.gov.scot/publications/care-digital-age-delivery-plan-2023-24/pages/1/ [accessed 2025-05-27]
9. Holoxica. 2023. URL: https://www.holoxica.com/ [accessed 2025-05-27]
10. Bernhard JC, Isotani S, Matsugasumi T, et al. Personalized 3D printed model of kidney and tumor anatomy: a useful tool for patient education. World J Urol 2016 Mar;34(3):337-345. [doi: 10.1007/s00345-015-1632-2] [Medline: 26162845]
11. Elwyn G, Frosch DL, Kobrin S. Implementing shared decision-making: consider all the consequences. Implement Sci 2016 Aug 8;11(1):114. [doi: 10.1186/s13012-016-0480-9] [Medline: 27502770]
12. van de Belt TH, Nijmeijer H, Grim D, et al. Patient-specific actual-size three-dimensional printed models for patient education in glioma treatment: first experiences. World Neurosurg 2018 Sep;117:e99-e105. [doi: 10.1016/j.wneu.2018.05.190] [Medline: 29870846]
13. Wake N, Rosenkrantz AB, Huang R, et al. Impact of 3D printed models on quantitative surgical outcomes for patients undergoing robotic-assisted radical prostatectomy: a cohort study. Abdom Radiol (NY) 2023 Apr;48(4):1401-1408. [doi: 10.1007/s00261-023-03815-2] [Medline: 36749368]
14. Greenhalgh T, Wherton J, Papoutsi C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017 Nov 1;19(11):e367. [doi: 10.2196/jmir.8775] [Medline: 29092808]
15. Greenhalgh T, Rosen R, Shaw SE, et al. Planning and evaluating remote consultation services: a new conceptual framework incorporating complexity and practical ethics. Front Digit Health 2021;3:726095. [doi: 10.3389/fdgth.2021.726095] [Medline: 34713199]
16. Walthall H, Schutz S, Snowball J, Vagner R, Fernandez N, Bartram E. Patients' and clinicians' experiences of remote consultation? A narrative synthesis. J Adv Nurs 2022 Jul;78(7):1954-1967. [doi: 10.1111/jan.15230] [Medline: 35362191]
17. Tasneem I, Ariz A, Bharti D, Haleem A, Javaid M, Bahl S. 3D printing technology and its significant applications in the context of healthcare education. J Ind Intg Mgmt 2023 Mar;08(01):113-130. [doi: 10.1142/S2424862221500159]

XSL•FO
RenderX

18. Senbekov M, Saliev T, Bukeyeva Z, et al. The recent progress and applications of digital technologies in healthcare: a review. Int J Telemed Appl 2020;2020:8830200. [doi: 10.1155/2020/8830200] [Medline: 33343657]

Abbreviations

Edited by A Mavragani; submitted 09.08.24; peer-reviewed by J Walsh; revised version received 16.04.25; accepted 24.04.25; published 18.06.25.

Please cite as:
Rai HK, Miller M, Leung S, Macleod E, Lennon M
Patient and Public Perceptions of 3D Technologies (Models and Images) to Facilitate Health Care Consultations: Exploratory, Mixed Methods Study
JMIR Form Res 2025;9:e65235
URL: https://formative.jmir.org/2025/1/e65235
doi:10.2196/65235

© Harleen Kaur Rai, Morven Miller, Steve Leung, Euan Macleod, Marilyn Lennon. Originally published in JMIR Formative Research (https://formative.jmir.org), 18.6.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Feasibility and Usability of a Web-Based Peer Support Network for Care Partners of People With Serious Illness (ConnectShareCare): Observational Study

Aricca D Van Citters[1], MS; Megan M Holthoff[1], MSHS; Colleen Young[2], BA; Sarah M Eck[3], PhD; Amelia M Cullinan[4,5], MD; Stephanie Carney[6], BA; Elizabeth A O'Donnell[7], MPH; Joel R King[1], BA; Malavika Govindan[8], MD; David Gustafson[9], PhD; Stephanie C Tomlin[8], MS, MPA; Anne B Holmes[8], BA; Ann D Bradley[8], RN; Brant J Oliver[1,10,11], MS, MPH, FNP-BC, PMHNP-BC, PhD; Matthew M Wilson[4,5], MD; Eugene C Nelson[1], DSC, MPH; Amber E Barnato[1,4,5], MD, MPH, MS; Kathryn B Kirkland[1,5], MD

[1]Dartmouth Institute for Health Policy and Clinical Practice, One Medical Center Drive, WTRB Level 5, Lebanon, NH, United States
[10]Departments of Community & Family Medicine, Psychiatry, Geisel School of Medicine at Dartmouth, Lebanon, NH, United States
[11]Office of Care Experience, The Value Institute, Dartmouth Health, Lebanon, NH, United States
[2]Mayo Clinic Connect, Mayo Clinic, Rochester, MN, United States
[3]Scientific Health, Lyme, NH, United States
[4]Palliative Medicine, Geisel School of Medicine at Dartmouth, Lebanon, NH, United States
[5]Section of Palliative Medicine, Dartmouth Health, Lebanon, NH, United States
[6]Dartmouth Health, Lebanon, NH, United States
[7]Alice Peck Day Memorial Hospital, Lebanon, NH, United States
[8]Patient and Family Advisor, United States
[9]College of Engineering, University of Wisconsin - Madison, Madison, WI, United States

Corresponding Author:
Aricca D Van Citters, MS
Dartmouth Institute for Health Policy and Clinical Practice, One Medical Center Drive, WTRB Level 5, Lebanon, NH, United States

Abstract

Background: While it can be rewarding to provide care for a person with serious illness, care partners are often unprepared to manage and cope with the physical and emotional stresses that arise with disease progression and bereavement.

Objective: We aimed to evaluate membership enrollment, engagement, and experiences within a web-based peer support network for active and bereaved care partners of people with serious illness.

Methods: We conducted a formative evaluation of the ConnectShareCare peer-to-peer web-based support network, which targeted care partners of people with serious illness residing in the northeastern United States. Recruitment methods included marketing postcards, flyers, listserv messages, and referrals from community stakeholders, peers, and clinicians. Enrollment occurred through a self-guided, web-based process. Study participants included members enrolled in ConnectShareCare between April 2021 and June 2023. We used the network's analytics dashboard (eg, registration, usage, and notification logs) to evaluate membership enrollment and engagement in discussions. We used surveys of a subset of members to assess experiences, including satisfaction, ability to find meaning by supporting others, and value and opportunities for improvement.

Results: Over 2 years, the network enrolled 250 members, with an average of 9 new members per month. Among 193 members providing information, most (58%, n=112) identified as active care partners, 17% (n=33) identified as bereaved care partners, and 27% (n=52) chose not to specify their role. Two-thirds of the 250 members did not post, 20% (n=50) posted 1 - 10 times, 6% (n=14) posted 11 - 25 times, 6% (n=15) posted 26-100 times, and 3% (n=7) posted more than 100 times. On average, 19 members posted per month resulting in 166 member posts per month. Moderators (1 community manager, 2 volunteer mentors, and 2 project team members) supported members with an average of 111 posts/month. In total, 187 discussion topics were created, including 42% (n=78) started by members and 58% (n=109) started by moderators. Seventy-eight discussion topics had 10 or more posts associated with them. The most frequent discussion topics focused on "check-ins" and "sources of joy and hope." Among 18 care partner members who completed a research survey, 69% (11/16) reported connecting with at least 1 person and 62% (10/16) reported that ConnectShareCare helped them find meaning and purpose by supporting others. Most reported satisfaction with support (12/16, 75%) and information (14/16, 88%) through the network. Although most noted that ConnectShareCare was easy to use (10/17, 59%), respondents were less likely to easily find the information they were seeking

XSL•FO
RenderX

(6/16, 38%). Survey respondents found value in peer connection and support and identified opportunities to improve navigation of resources and engagement of members.

Conclusions: Care partners of people with serious illness can use a web-based peer support network to find meaningful and useful support and information. Additional work is needed to identify the impact of the network on distress, social connectivity, and support programming.

(JMIR Form Res 2025;9:e70206) doi:10.2196/70206

KEYWORDS

serious illness; care partner; caregivers; peer-support; self-help groups; social networking; internet; peer-to-peer

Introduction

In the United States, approximately 38 million unpaid care partners (often family members or friends) provided 36 billion hours of care in 2021, with an estimated value of US $600 billion [1]. While caring for a loved one with a serious illness can bring deep meaning and satisfaction, it can also be challenging [2]. Care partners of people with serious illness are often unprepared to manage and cope with the burden of caregiving and the circumstances that arise with disease progression and bereavement. They often have insufficient communication with clinicians; receive inadequate support and educational materials; and encounter burdens, challenges, and unmet needs [3,4]. Most family care partners work full-time, and many have their own health issues [1]. Addressing care partner needs is a pressing health, economic, and social imperative.

There is a growing literature on web-based peer support networks for care partners of people with serious illness, and such networks typically contain a platform to exchange information and provide peer support [5]. Evidence on the impact of web-based peer support networks is promising, with the potential to improve emotional well-being, mood, burden, quality of life, self-efficacy, decision-making, confidence, and relationship quality [6-13]. Building on this evidence base, we co-designed and implemented ConnectShareCare [14], a web-based peer support network to assist active and bereaved care partners of adults with a serious illness in providing each other with emotional support and exchanging helpful information and resources.

Despite the growth in web-based peer support networks for care partners of people with serious illness, there is limited information describing growth in membership enrollment, engagement of members in discussions, and experiences of members participating in such networks. The primary objective of this study is to describe the feasibility and usability of the ConnectShareCare web-based peer support network for active and bereaved care partners.

Methods

Study Design

We conducted a formative evaluation to examine the feasibility and usability of ConnectShareCare, as defined by growth in membership enrollment, engagement in discussions, and member experiences using the network. The co-design process supporting the development of the web-based peer support network has been described separately by O'Donnell et al [14].

Setting

This study was conducted at a large, rural academic medical center in the northeastern United States. While the web-based peer support network targeted care partners of people with serious illness [15] residing in New Hampshire and Vermont, participation was not restricted by geographic region or by receipt of health care within the academic medical center.

Participants

Network Recruitment and Enrollment

Participants included people enrolled in the ConnectShareCare network ("members") between April 2021 and June 2023. We recruited people to the ConnectShareCare network via multiple methods, including invitations from co-design team members; marketing postcards and flyers posted within the community (eg, community boards, libraries, or senior centers); presentations to community stakeholders (eg, health fairs, networking and learning meetings, or local hospices); and posts to local web-based resources (eg, community listservs or recruitment via peers). We also worked with local health systems to promote referrals from clinicians. We shared referral materials with clinicians and support group providers within the community, health system, and community organizations (eg, hospice care, visiting nursing services, or Veterans Affairs clinics). People enrolled in the network using a self-guided, web-based enrollment process.

Recruitment for Survey on Member Experiences

Among the subset of 181 members who enrolled in the network between April 2021 and August 2022, a total of 55% (n=100) agreed during the enrollment process to be contacted for research. We contacted individuals via email between September and November 2022 to complete an anonymous web-based survey to share their experiences using ConnectShareCare. Potential survey participants received reminder emails 7 and 14 days later. We also posted the anonymous survey link on the ConnectShareCare network and invited people who initially opted out of participating in additional research to complete the survey.

Description of the ConnectShareCare Platform

Peer Connection

The ConnectShareCare network provided participants with the ability to connect with each other using "support group" forums,

XSL•FO
RenderX

categorized as: active caring, grief and loss, and news and announcements. Discussion topics within these forums were often initiated by moderators, but members also established new discussion topics. The network also featured "private messaging" to facilitate writing to another member directly, and "member stories" (including poetry, music, and artwork), which offered a place to share member experiences and learn from one another.

Resources

The network provided members with access to pertinent web-based resource and local and related events and programming. Five categories of resource materials were available, including: planning ahead (eg, state-specific advance directives or care guides); practical resources (eg, housing, insurance, bills, legal, transportation, or food); emotional resources (eg, mental health, finding a therapist, or healing after loss); communication resources (eg, connecting with care providers, interpreters, or finding help); and family resources (eg, managing children, partners, or intimacy). Upcoming events included items such as "Drop-In Mindfulness," "The Parkinson's Workshop," and "Bereavement Hiking Group."

Roles and Responsibilities Within the Network

The network was overseen by a community manager, who set the strategic direction and tactics of the network concerning enrollment and engagement [16]; developed relationships with health care and community partners; recruited, trained, and managed volunteer mentors; and served as the primary moderator of the network. The community manager held a Master of Public Health degree and was supported by an external consultant with significant expertise in community management of web-based peer support networks. Two volunteer mentors were recruited based on their pre-existing skill sets and came to the platform with essential professional and life experiences. Each had formal training in providing complementary health services (eg, Reiki and social work), but were not employed by the health system. In addition to serving as moderators when the community manager was unavailable, volunteer mentors were consistently present on the network, assisting in connecting members, providing resources, and modeling behaviors and posting practices. The network was supported by 2 project team members who were responsible for administrative tasks and served as moderators when the community manager was unavailable.

The moderation structure of the network was modeled after established practices [17]. The community manager signed-in multiple times a week to intentionally nurture and facilitate connection among members with shared experiences; establish new discussion topics, post replies, and ask questions to promote conversation and increase retention of members; and identify or address pressing needs (eg, inappropriate or potentially offensive posts, posts offering unvetted medical advice, etc). When the community manager was unable to fulfill the moderator position (eg, vacation or personal days), volunteer mentors or project team members assumed the moderator role. Community management guidelines (Multimedia Appendix 1) were established and provided to anyone fulfilling the position to clarify expectations and ensure continuity.

Safety Monitoring

The community manager, volunteer mentors, and project team members worked together to monitor all posted discussions for potential safety concerns (eg, reference to self-harm or violence) and noncompliance with community guidelines (eg, kindness; digital safety; avoidance of bullying, marketing, or medical advice; etc; Multimedia Appendix 1). To ensure adequate response time, the community manager or designated team member monitored automated safety notifications. A keyword monitoring function was used, which did not prevent posts from being visible, but triggered an email notification to the community manager and project team members when specific keywords were used. Keywords included terms related to self-harm, suicide, and violence (eg, "suicide," "overdose," "kill," "die," "attack," or "threaten"). Community members could also report discussions, which generated a similar notification. The reporting feature included an option for a free-text response from the flagger, which provided context for why the discussion was reported. To further bolster security and safety, discussion topics were made viewable only to registered members of the ConnectShareCare community, and new members were restricted from posting hyperlinks during their first 10 days.

Ethical Considerations

This study was approved by the Dartmouth Health Institutional Review Board (study #02000907). Individuals who enrolled in the network were informed in the "terms of use" that usage data would be collected and used to improve the network. Those who did not agree to the terms of use were unable to participate in the network. The Institutional Review Board granted a waiver of written documentation of consent for the survey component of this study, and the survey introduction included language describing the voluntary nature of this study, confidentiality, and the absence of identifiable data (Multimedia Appendix 2). No identifiable data was collected, and participants were not compensated for their engagement.

Variables

This study evaluated feasibility via growth in membership enrollment and engagement, and usability via experience of use questions, as defined in Table 1. We used the analytics dashboard provided by the platform vendor, CareHubs, to monitor enrollment and engagement in the network. We used anonymous electronic surveys (Multimedia Appendix 2) to understand usability for a subset of participants. The anonymous survey assessed care partner characteristics, ease of use, value, and opportunities for improvement.

XSL•FO
RenderX

Table . Feasibility and usability metrics used to assess the ConnectShareCare network.

Metric		Definition	Data source
Feasibility			
	Growth in membership enrollment: new members (cumulative and new enrollments per month)	Count of new members enrolled during the month (excluding duplicate entries, test accounts, and members of the project team).	Network analytics: registration log
	Engagement: total posts per month	Total posts per month by enrolled users (stratified by members and moderators).	Network analytics: usage log
	Engagement: total members posting per month	Number of members posting per month (includes members and moderators).	Network analytics: usage log
Usability			
	Experience: satisfaction (with support received and information found)	Proportion of respondents who report "very satisfied" or "somewhat satisfied" with (1) the support received and (2) the information found through the network.	Anonymous survey
	Experience: ability to find meaning by supporting others	Proportion of respondents who "agree" or "strongly agree" that the network helped them find meaning and purpose by supporting others.	Anonymous survey
	Ease of use of the network	Proportion of respondents who "agree" or "strongly agree" to "the website was easy to use" and "I was able to easily find information I was looking for."	Anonymous survey
	Perceived value and opportunities for improvement of the network (from free text responses)	Strengths: (1) What do you like most about the ConnectShareCare support network? (2) Please share an example of a time where ConnectShareCare made a positive impact on your day, or on your ability to connect with others, or manage daily activities. Opportunities: What could we do to improve the: (1) support available through ConnectShareCare? (2) information and resources you found through ConnectShareCare?	Anonymous survey

Statistical Methods

We used statistical process control c Charts [18] to identify changes in enrollment and engagement over time. We used descriptive statistics to summarize categorical information from survey responses. We conducted descriptive statistics with SPSS (version 28; IBM Corp). Subgroup analyses were not conducted due to the small sample size. Missing data were excluded on an analysis-by-analysis basis.

We used thematic analysis to derive qualitative insights into the benefits, limitations, and improvement opportunities identified in free-form text responses to survey questions. Two researchers (JRK and ADVC) independently reviewed all survey responses to familiarize themselves with the data and developed a codebook using inductive methods. Responses were coded by JRK and all coded responses were reviewed by ADVC. JRK and ADVC are experienced qualitative researchers, and JRK has experience in being a care partner for a person with serious illness. Differences in coding decisions were resolved through discussion. To ensure reliability, findings were reviewed with the network moderator and research team. We conducted qualitative analyses using Atlas.ti (version 23, ATLAS.ti Scientific Software Development GmbH) .

Results

Participants

Characteristics of Network Members

During this study, 250 people enrolled in ConnectShareCare. Among members providing additional information about themselves upon enrollment (n=193), most (n=112, 58%) identified as an active care partner, 17% (n=33) identified as a bereaved care partner, and 27% (n=52) chose not to specify their role. Some individuals identified as both active and

XSL•FO
RenderX

bereaved care partners. People most commonly cared for a spouse or partner (n=77, 50%); a child, grandchild, or niece or nephew (n=24, 15%); a parent (n=17, 11%); a sibling (n=8, 5%); a friend or neighbor (n=6, 4%); or another person (n=23, 15%).

Characteristics of Survey Respondents

A subset of care partner members (n=18) completed anonymous surveys on their experience with the network, including 13 engaged members (who signed in to the network at least once in the prior 90 d), 1 inactive member (who had not signed in to the network within the prior 90 d), and 4 members responding from an anonymous link provided on the website. At the time of survey completion, 44% (n=8) had been enrolled in ConnectShareCare for 12 or more months; 39% (n=7) for 4 - 12 months, and 17% (n=3) for 1 - 3 months.

As shown in Table 2, most survey respondents (n=15, 83%) were female, and all spoke English as their primary language. The median age was 68.5 (range 42 - 86, IQR 54-74) years. Most (n=15, 83%) had a college degree.

Over half of the survey respondents (n=11, 61%) identified as bereaved. Most respondents cared for their spouse or life partner (n=13, 72%), with a smaller proportion caring for a child or parent. Respondents were care partners for people with cancer (n=7, 39%), dementia (n=4, 22%), Parkinson disease (n=3, 17%), cardiac conditions (n=2, 11%), lung conditions (n=1, 6%), or did not state the type of condition (n=3, 17%).

XSL•FO
RenderX

Table . Characteristics of care partners responding to an anonymous survey about their experiences with ConnectShareCare (N=18).

	Survey respondents
Gender, n (%)	
Male	3 (17)
Female	15 (83)
Age (years), median (IQR)	68.5 (54-74)
Primary language, n (%)	
English	18 (100)
Education level, n (%)	
High school diploma or equivalent	1 (6)
Some college	2 (11)
College graduate	7 (39)
Master's or doctoral degree	8 (44)
Care partner type, n (%)	
Bereaved	11 (61)
Active	7 (39)
Relationship of care partner to person with serious illness, n (%)	
Spouse or partner	13 (72)
Child	3 (17)
Parent	1 (6)
Not stated	1 (6)
Length of time as a care partner, n (%)	
Active (<12 months)	1 (6)
Active (12 or more months)	4 (22)
Bereaved (<12 months)	7 (39)
Bereaved (12 or more months)	4 (22)
Not stated	2 (11)
Diagnosis of person with serious illness, n (%)	
Cancer	7 (39)
Dementia	4 (22)
Parkinson disease	3 (17)
Heart disease (eg, stroke)	2 (11)
Lung disease (eg, chronic obstructive pulmonary disease or emphysema)	1 (6)
Not stated	3 (17)

Main Results

Growth in Membership Enrollment

As shown in Figure 1, a total of 250 members enrolled during this study period, with an average of 9 members enrolling per month. Enrollment of new members was more rapid during the initial testing phase (April 2021 through July 2021), and a lower rate of enrollment was observed between March 2022 and October 2022.

Most participants (n=142, 57%) were referred by a staff or clinician at the health system. Others were referred by: friends, neighbors, or family members (n=34, 14%); community organizations (n=29, 12%); community listservs (n=23, 9%); flyers or pamphlets (n=11, 4%); or another source (n=20, 8%).

XSL•FO
RenderX

Figure 1. Number of members registered on the ConnectShareCare network each month, between April 2021 and June 2023 (c Chart Statistical Process Control analysis). The mean number of monthly registrations is represented by the turquoise (solid) line. The UCLs and LCLs are represented by the red dash lines. CL: confidence limit; LCL: lower confidence limit; UCL: upper confidence limit.

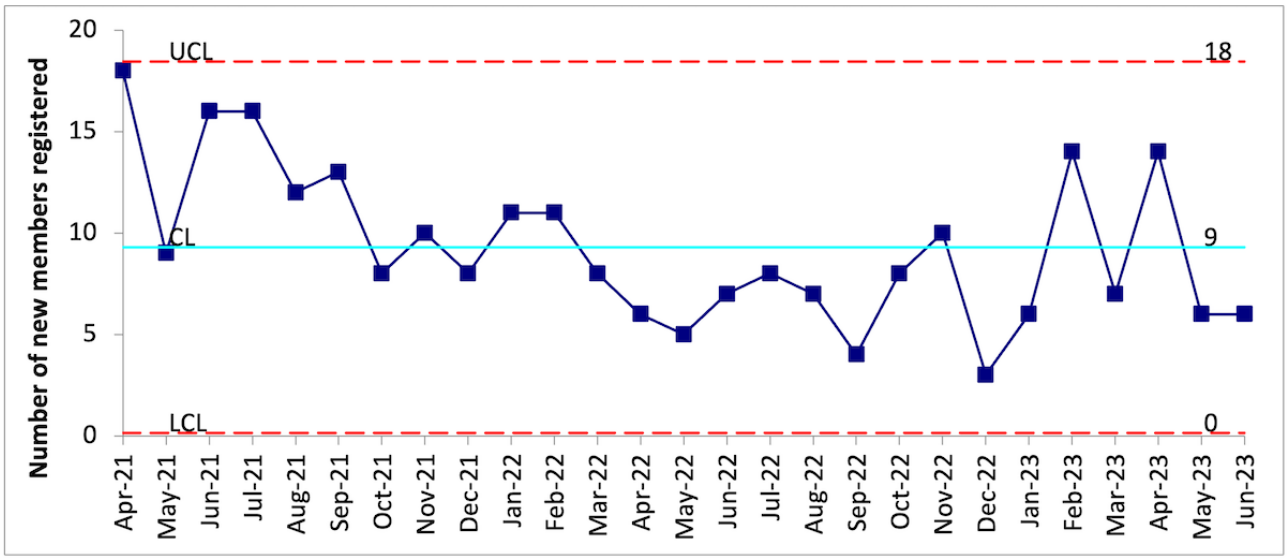

Engagement: Total Posts Per Month and Members Posting Per Month

An average of 166 posts were made by members each month (Figure 2A), whereas moderators (1 community manager, 2 volunteer mentors, and 2 project team members) had an average of 111 posts per month (Figure 2B). There was a moderate direct correlation (correlation coefficient: 0.59) between the number of member posts and moderator posts. On average, 19 members posted comments to the network each month (Figure 2C).

As shown in Figure 2A-C, there was a significant decline in member and moderator posts, as well as in the numbers of people posting between April 2022 and October 2022, which corresponded to a refocused effort by the community manager to define enrollment and engagement [16] strategies and targets for the network. New targets were identified in November 2022, which were approved by institutional leaders and were in place through June 2023.

During this study, 34% (n=86) of members posted on the network. Twenty percent (n=50) posted 1 to 10 times, 6% (n=14) posted 11 to 25 times, 6% (n=15) posted 26 to 100 times, and 3% (n=7) posted more than 100 times. Among those posting, the median number of posts was 6 (IQR 2 - 26.5).

In total, 187 discussion topics were created, including 42% (n=78) started by members and 58% (n=109) started by moderators. Seventy-eight discussion topics each had 10 or more posts associated with them, while 32 discussion topics had only 1 post, including announcements of in-person or web-based support sessions (n=22), caregiving resources (n=4), or general network announcements (n=4).

The most popular discussion topics and respective number of posts included: “Check-in: How are we all doing today?” (n=1753), “Sources of joy and hope” (n=1436), “Can you name your feeling today? (n=649), “Meet fellow members: come introduce yourself” (n=409), “How has your grief changed over time?” (n=237), “Alzheimer Disease and other dementias - come share” (n=231), “In person meet-ups, anyone?” (n=139), “What do you wish you had known about being a care partner?” (n=109), and “What I thought I would never be thankful for...” (n=108).

XSL•FO
RenderX

Figure 2. Number of (A) member posts, (B) moderator posts, and (C) total number of individuals (members and moderators) posting each month on the ConnectShareCare network (c Chart Statistical Process Control analysis). The mean is represented by the turquoise (solid) line. The UCLs and LCLs are represented by the red dash lines. CL: confidence limit; LCL: lower confidence limit; UCL: upper confidence limit.

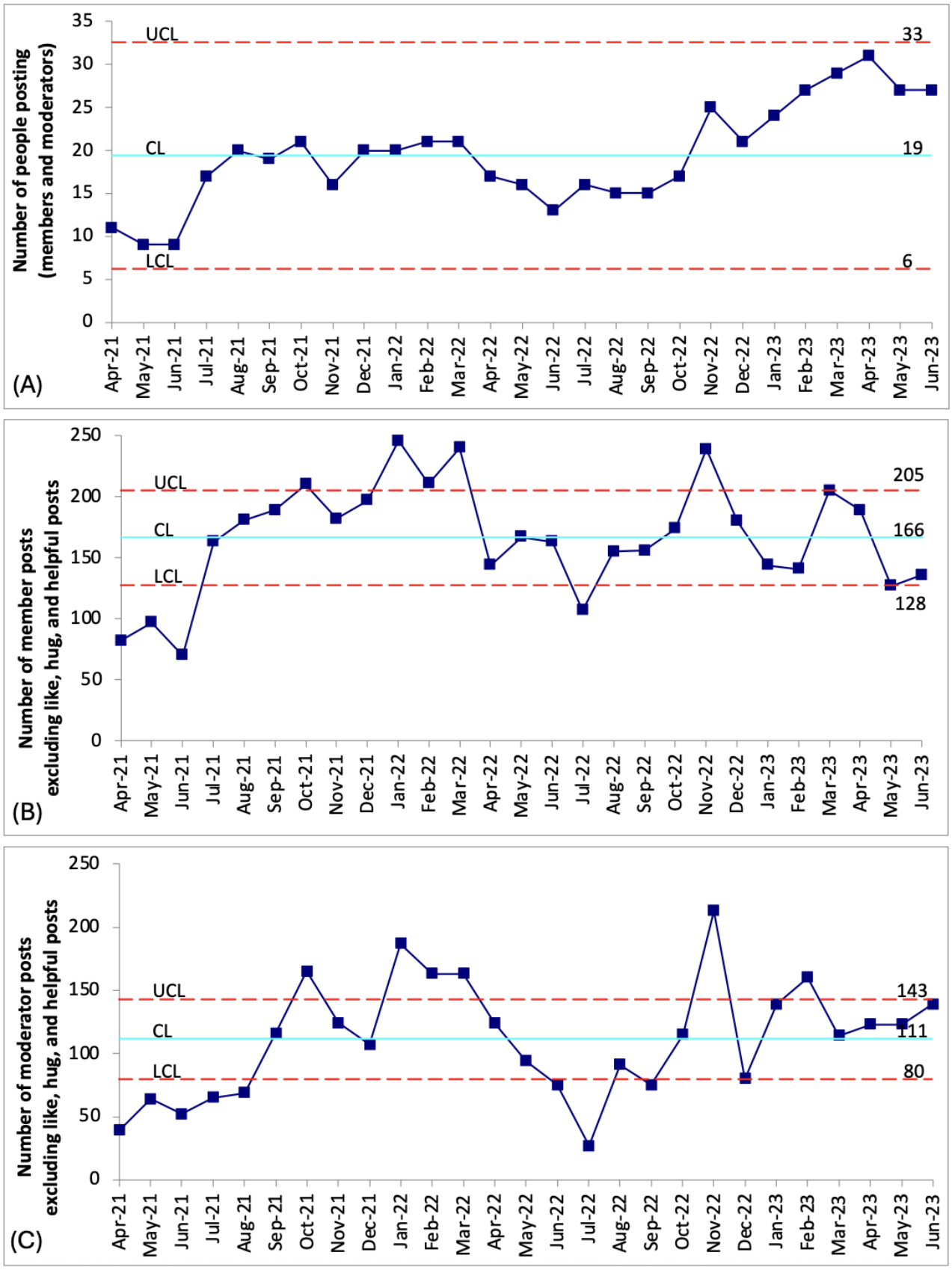

XSL•FO
RenderX

Experience: Satisfaction and Ability to Find Meaning

Care partner experience was assessed via an anonymous survey. Among those seeking support (n=16), most were satisfied with the support they received (n=12, 75%; data not shown). As shown in Figure 3, more than half of the survey respondents seeking support were able to make a connection with at least one other person (n=11, 69%) and believed that ConnectShareCare helped them find meaning and purpose by supporting others (n=10, 62%).

Among those seeking information (n=16), most were satisfied with the information they found (n=14, 88%). Nonetheless, less than half reported being able to easily find information (n=6, 38%) and only 59% (n=10) found the website easy to use.

Figure 3. ConnectShareCare usability, as reported by ConnectShareCare members who responded to the survey about the user experience of the network during this study. Respondents with missing data were excluded from the analysis.

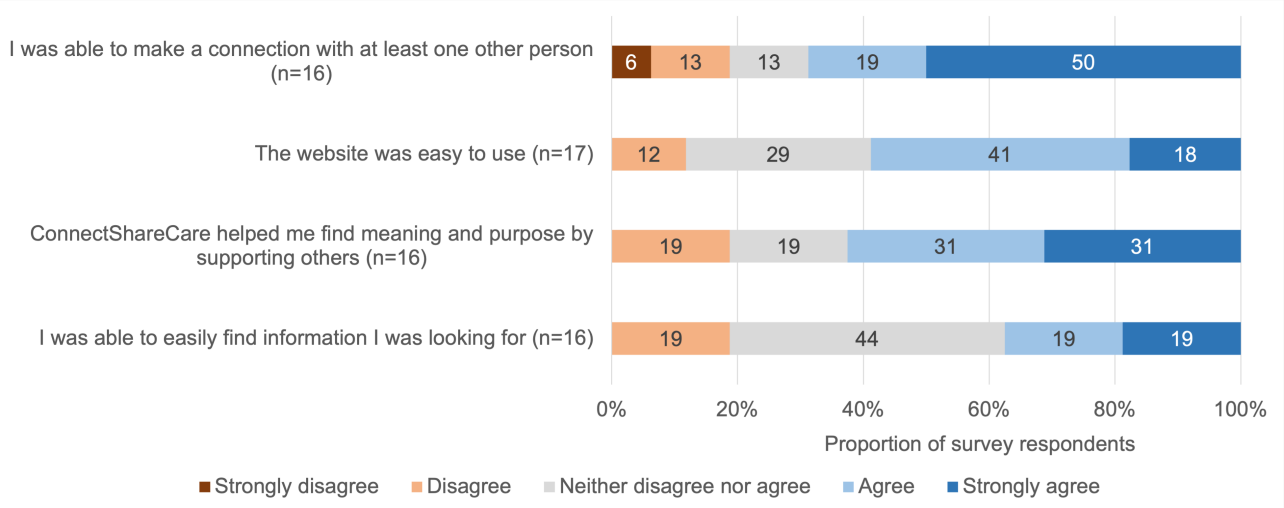

Value and Opportunities for Improvement

Survey respondents most commonly identified peer connection and support as the greatest value of ConnectShareCare. As expressed by a woman aged 66 years actively caring for a loved one: “I was feeling very down and, in response to a post, received ‘hug’ and ‘like’ symbols as well as words of validation and encouragement which helped me turn around my day.” A woman aged 53 years who recently lost a loved one shared the value in: “Feeling connected, talking to people who understand what I might be going through, kindred spirits.” Similarly, a woman aged 51 years who was recently bereaved identified value in: “Getting my feelings out and connecting with people. I think it is important to share stories.” Another woman aged 61 years shared the value as: “The honesty, gratitude, sharing, support that people give of themselves, and offer to others.” Respondents also noted the responsiveness of the network, as shared by this woman aged 66 years caring for a loved one: “I appreciate the caring comments from members to each other and to me. When I ask a question, I get a reply fairly quickly. I appreciate offering support to others.”

Improvement opportunities highlighted both a desire for greater engagement in the network and improved ability to access information or resources. As noted by this bereaved man (age not listed): “Continue to reach out to our regional community. As more individuals and organizations learn about CSC [ConnectShareCare] the more they will add to the CSC [ConnectShareCare] community.” Similarly, a woman aged 56 years who recently lost a loved one shared: “Not sure how you can engage more members to post? Not everyone wants to, and that’s ok. I see you trying!” Improved access to information was highlighted by a woman aged 75 years actively caring for a loved one who stated: “Maybe a little information on how to access things within the group.” The need for additional resources was highlighted by a bereaved woman aged 51 years who commented: “I think we need resources for younger people.”

Discussion

Principal Results

Web-based peer support networks can offer a valuable community for individuals who may not know one another, but who share common experiences. Our study found that it was feasible to enroll and engage care partners of people with serious illness and those in bereavement using a web-based peer support network. Over one-third of ConnectShareCare members posted to the digital network. Respondents reported being able to connect with others and find meaning and purpose through supporting others. While over half indicated the platform was easy to use, less than half found it easy to find the information they were seeking. Overall, the platform offered a “safe” space for people to share.

This study offered real-world testing and implementation of a web-based peer support network in a rural region with a robust moderation structure, including a community manager, volunteer mentors, and project team members. Members were welcomed, connected to others with shared lived experience, offered forum topics and posts to encourage engagement, and monitored to ensure the community rules were respected. An inherent advantage to this support group was the regional nature of the model (as opposed to a national support group), which fostered shared community knowledge, experience, and local opportunities.

XSL•FO
RenderX

Enrollment and engagement levels mirrored the effort put into them. For example, initial enrollment efforts were strong, and in response, the number of enrolled members per month was higher than when community management strategies shifted to focus on member engagement strategies. Engagement of members, as shown by posts per month and members posting per month, also reflected the focus on engagement strategies deployed by the moderators. When the number of posts from the designated community manager declined, the number of member posts decreased. The relationship between the number of moderator posts and member posts highlights the critical role of the moderator in the network, as others have found [19].

Limitations

Our platform and study focused on those who enrolled and engaged. Participation in ConnectShareCare was open to the public, and verification as a care partner was not required, thus, members include not only care partners but also other community members or interested staff and clinicians. Enrollment and usage data reflect real-world implementation data, and patterns of use may differ if members were restricted to care partners only. Future efforts could target and engage those who may be most lonely and may benefit the most from remote interaction (eg, individuals who are homebound). Prior research has indicated a relationship between being homebound and social isolation and loneliness [20,21] and that people with these characteristics are also most likely to engage in web-based peer support communities [22-24]. In this study, all users spoke English as a primary language, and most had college degrees, which may not be representative of the general population or those who may benefit the most from a web-based peer support network. Moreover, our study did not collect racial and ethnic information on network members or survey respondents, and cultural differences may impact enrollment and engagement behaviors. Finally, survey participants ranged in age from 42 to 86 years; however, due to the small sample size, we are unable to assess potential age-related barriers or facilitators of engagement.

As this study required access to the internet and literacy, inherent selection bias existed. Due to the self-enrolling aspect of ConnectShareCare, we lack perspectives of the population of care partners who opted not to participate. Additionally, most of the completed survey responses were received from “engaged” members, and not from those who did not log in at least once in the prior 90 days. While low survey responses are common in anonymous web-based surveys [25], the limited response from those who did not engage likely overinflated positive responses regarding experiences with the network.

Comparison With Prior Work

Members of ConnectShareCare most frequently engaged with fellow community members on discussion topics that offered a place to “check-in” and share how one was doing and feeling at a given time, as well as identifying sources of hope and joy. Other similar networks have found that discussion topics around emotions and sentiments are common among care partners of people with serious illness [26,27].

Like other web-based peer support communities [26,28-30], engagement by members of ConnectShareCare was variable, with the largest proportion (65%) not posting to the network, 32% contributing 1 to 100 posts, and a small proportion (3%) contributing more than 100 posts. While this exceeds the often referenced 90-9-1 rule for “lurkers,” “contributors,” and “super users” of a social network, studies have shown that nonposting members can gain benefits from being “part of” a web-based peer support community [30]. Recent studies have shown that nonposting members exhibit a higher level of perceived functional well-being [28] and demonstrate similar improvements as posters for being better informed, more confident in the relationship with their physician, having improved acceptance of the disease, and feeling more confident about treatment [29].

This study identified some challenges with web-based peer support communities. For example, participation was largely dominated by a small number of individuals. The ConnectShareCare network offered a variety of resources for its members; nonetheless, some survey respondents still struggled to easily find information they were seeking, which may reflect the often overwhelming nature of caring for a loved one [31]. Previous research has similarly found that engagement in web-based peer support networks is difficult for those with limited time and digital information-seeking skills [32]. This observation emphasizes an important role for the community manager to link individuals to specific resources and to optimize members’ abilities to find information within the network’s resources or through the ConnectShareCare community. Notably, following the completion of this evaluation, several enhancements were made to the platform supporting efforts to improve retention, increase findability (eg, search functions), and reduce moderator burden.

The ConnectShareCare network was co-designed to support both active and bereaved care partners, recognizing that while active and bereaved care partners may have different needs, they may benefit from connecting with each other to share information and provide or receive support [14]. Research suggests that care partners may have varying levels of interaction with web-based peer support networks due to factors such as available downtime or emotional isolation [24]. As such, differences in needs and experiences of active and bereaved care partners may impact enrollment, engagement, and care partner comfort within the network. Separate forums for active and bereaved care partners provided the ability for members to seamlessly transition between forums.

Conclusions

This study enhanced our understanding of growth in membership enrollment, engagement, and experiences of using a web-based peer support network, ConnectShareCare, among people in a rural region of the northeastern United States. To maintain the integrity of this model, designated moderators who strive to actively facilitate connection are key to promoting participation in the network. Additional work is needed to understand the types of information that members have difficulty finding and to identify the impact of the network on health and well-being outcomes, such as distress and social connectivity. Future work

XSL•FO
RenderX

in this area may wish to explore the impact of small incentives [33] to encourage people to engage in the network, expansion to non-English speaking populations, and the development of a set of critical success factors to support care partner well-being as well as health care operations. Larger studies with more representative populations of care partners are imperative for determining utility, sustainability, spread, and scalability.

Acknowledgments

The authors wish to thank the many other people who contributed to the design and implementation of the ConnectShareCare network and the many active and bereaved care partners who participated in, and provided feedback on, the network. Specific appreciation is given to Andrea Buccellato, BA, for her integration of the work within the Dartmouth Health system. The authors also wish to thank Martha Bass, MSW; Sarah Berman; Tyler Gooding, BA; Amanda Hoggard, MS; Alexandra Steinbaugh, MSW; Janet Miller Haines, BS; Andy Williams, BS; and Jen Hinson, BS, for their guidance in the design and implementation of the network. Additionally, the authors wish to thank the Upper Valley Elder Forum; Upper Valley Bereavement Partnership Group; Volunteer Services, Office of Care Experience, Dartmouth Hitchcock Medical Center (DHMC); Arts and Humanities, Office of Care Experience, DHMC; Dartmouth Health Palliative Care Program; Dartmouth Health Aging Resource Center; Dartmouth Cancer Center; Visiting Nurse and Hospice of Vermont and New Hampshire; Lake Sunapee VNA & Hospice Care; Bayada Home Care; CommunityCare of Lyme and Community Nurse Connection; Shire Digital; and CareHubs for their support of the ConnectShareCare network. Generative artificial intelligence was not used in the preparation of this paper. This work was supported by the Gordon & Betty Moore Foundation (grant 7485), Couch Family Fund at the Dartmouth Institute for Health Policy and Clinical Practice, and charitable donations supporting the Division of Palliative Medicine at the Dartmouth Hitchcock Medical Center & Clinics and the Dartmouth Cancer Center in Lebanon, New Hampshire. The funders were not involved in the conduct of this study.

Data Availability

Since the data collected in this research includes data from the ConnectShareCare network and anonymous surveys, the nature of the terms of use and privacy policies of the network limit our ability to share network data with others. Data derived from the anonymous survey may be available from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Community management guidelines.
[PDF File, 514 KB - formative_v9i1e70206_app1.pdf]

Multimedia Appendix 2
Anonymous user experience survey.
[DOCX File, 23 KB - formative_v9i1e70206_app2.docx]

References

1. Reinhard SC, Caldera S, Houser A, Choula RB. Valuing the invaluable: 2023 update strengthening supports for family caregivers. AARP Public Policy Institute. 2023. URL: https://www.aarp.org/ppi/info-2015/valuing-the-invaluable-2015-update.html [accessed 2025-05-02]
2. Lewis C, Abrams MK, Schneider EC, Shah T. Coping with serious illness in America: relying on family and friends. The Commonwealth Fund. 2019. URL: https://www.commonwealthfund.org/publications/2019/feb/coping-serious-illness-america [accessed 2025-05-02]
3. Northouse LL, Katapodi MC, Song L, Zhang L, Mood DW. Interventions with family caregivers of cancer patients: meta-analysis of randomized trials. CA Cancer J Clin 2010;60(5):317-339. [doi: 10.3322/caac.20081] [Medline: 20709946]
4. Choi S, Seo J. Analysis of caregiver burden in palliative care: an integrated review. Nurs Forum 2019 Apr;54(2):280-290. [doi: 10.1111/nuf.12328] [Medline: 30737798]
5. Daynes-Kearney R, Gallagher S. Online support groups for family caregivers: scoping review. J Med Internet Res 2023 Dec 13;25:e46858. [doi: 10.2196/46858] [Medline: 38090796]
6. Shin JY, Kang TI, Noll RB, Choi SW. Supporting caregivers of patients with cancer: a summary of technology-mediated interventions and future directions. Am Soc Clin Oncol Educ Book 2018 May 23;38:838-849. [doi: 10.1200/EDBK_201397] [Medline: 30231412]

XSL•FO
RenderX

7. Dam AEH, de Vugt ME, Klinkenberg IPM, Verhey FRJ, van Boxtel MPJ. A systematic review of social support interventions for caregivers of people with dementia: are they doing what they promise? Maturitas 2016 Mar;85:117-130. [doi: 10.1016/j.maturitas.2015.12.008] [Medline: 26857890]
8. Northouse L, Schafenacker A, Barr KLC, et al. A tailored web-based psychoeducational intervention for cancer patients and their family caregivers. Cancer Nurs 2014;37(5):321-330. [doi: 10.1097/NCC.0000000000000159] [Medline: 24945270]
9. DuBenske LL, Gustafson DH, Namkoong K, et al. CHESS improves cancer caregivers' burden and mood: results of an eHealth RCT. Health Psychol 2014 Oct;33(10):1261-1272. [doi: 10.1037/a0034216] [Medline: 24245838]
10. Chih MY, DuBenske LL, Hawkins RP, et al. Communicating advanced cancer patients' symptoms via the internet: a pooled analysis of two randomized trials examining caregiver preparedness, physical burden, and negative mood. Palliat Med 2013 Jun;27(6):533-543. [doi: 10.1177/0269216312457213] [Medline: 22988042]
11. Marziali E, Garcia LJ. Dementia caregivers' responses to 2 internet-based intervention programs. Am J Alzheimers Dis Other Demen 2011 Feb;26(1):36-43. [doi: 10.1177/1533317510387586] [Medline: 21282276]
12. Bass DM, McClendon MJ, Brennan PF, McCarthy C. The buffering effect of a computer support network on caregiver strain. J Aging Health 1998 Feb;10(1):20-43. [doi: 10.1177/089826439801000102] [Medline: 10182416]
13. McKechnie V, Barker C, Stott J. The effectiveness of an internet support forum for carers of people with dementia: a pre-post cohort study. J Med Internet Res 2014 Feb 28;16(2):e68. [doi: 10.2196/jmir.3166] [Medline: 24583789]
14. O'Donnell EA, Van Citters AD, Khayal IS, et al. A web-based peer support network to help care partners of people with serious illness: co-design study. JMIR Hum Factors 2024 May 8;11:e53194. [doi: 10.2196/53194] [Medline: 38717809]
15. Kelley AS, Bollens-Lund E. Identifying the population with serious illness: the "denominator" challenge. J Palliat Med 2018 Mar;21(S2):S7-S16. [doi: 10.1089/jpm.2017.0548] [Medline: 29125784]
16. Young C. Community management that works: how to build and sustain a thriving online health community. J Med Internet Res 2013 Jun 11;15(6):e119. [doi: 10.2196/jmir.2501] [Medline: 23759312]
17. Skousen T, Safadi H, Young C, Karahanna E, Safadi S, Chebib F. Successful moderation in online patient communities: inductive case study. J Med Internet Res 2020 Mar 17;22(3):e15983. [doi: 10.2196/15983] [Medline: 32181743]
18. Oliver BJ, Ogrinc G. Practical measurement for health care improvement. : Joint Commission Resources; 2022.
19. Deng D, Rogers T, Naslund JA. The role of moderators in facilitating and encouraging peer-to-peer support in an online mental health community: a qualitative exploratory study. J Technol Behav Sci 2023;8(2):128-139. [doi: 10.1007/s41347-023-00302-9] [Medline: 36810998]
20. Ankuda CK, Kotwal A, Reckrey J, Harrison KL, Ornstein KA. The experience of homebound older adults during the COVID-19 pandemic. J Gen Intern Med 2022 Apr;37(5):1177-1182. [doi: 10.1007/s11606-021-07361-9] [Medline: 35167063]
21. Cudjoe TKM, Prichett L, Szanton SL, Roberts Lavigne LC, Thorpe RJ. Social isolation, homebound status, and race among older adults: findings from the National Health and Aging Trends Study (2011-2019). J Am Geriatr Soc 2022 Jul;70(7):2093-2100. [doi: 10.1111/jgs.17795] [Medline: 35415872]
22. Zubatsky M. Editorial: virtual groups to address the health of homebound adults during COVID-19: a biopsychosocial framework. J Nutr Health Aging 2021;25(3):281-283. [doi: 10.1007/s12603-021-1584-8] [Medline: 33575716]
23. Szanton SL, Roberts L, Leff B, et al. Home but still engaged: participation in social activities among the homebound. Qual Life Res 2016 Aug;25(8):1913-1920. [doi: 10.1007/s11136-016-1245-2] [Medline: 26883818]
24. Washington KT, Oliver DP, Benson JJ, et al. Factors influencing engagement in an online support group for family caregivers of individuals with advanced cancer. J Psychosoc Oncol 2020;38(3):235-250. [doi: 10.1080/07347332.2019.1680592] [Medline: 31690247]
25. Menon V, Muraleedharan A. Internet-based surveys: relevance, methodological considerations and troubleshooting strategies. Gen Psychiatr 2020;33(5):e100264. [doi: 10.1136/gpsych-2020-100264] [Medline: 32818170]
26. Ni C, Song Q, Malin B, et al. Examining online behaviors of adult-child and spousal caregivers for people living with Alzheimer disease or related dementias: comparative study in an open online community. J Med Internet Res 2023 Nov 17;25:e48193. [doi: 10.2196/48193] [Medline: 37976095]
27. Benson JJ, Oliver DP, Washington KT, et al. Online social support groups for informal caregivers of hospice patients with cancer. Eur J Oncol Nurs 2020 Feb;44(101698):101698. [doi: 10.1016/j.ejon.2019.101698] [Medline: 31816508]
28. Han JY, Hou J, Kim E, Gustafson DH. Lurking as an active participation process: a longitudinal investigation of engagement with an online cancer support group. Health Commun 2014;29(9):911-923. [doi: 10.1080/10410236.2013.816911] [Medline: 24345206]
29. van Uden-Kraan CF, Drossaert CHC, Taal E, Seydel ER, van de Laar M. Self-reported differences in empowerment between lurkers and posters in online patient support groups. J Med Internet Res 2008 Jun 30;10(2):e18. [doi: 10.2196/jmir.992] [Medline: 18653442]
30. Schneider A, von Krogh G, Jäger P. "What's coming next?" Epistemic curiosity and lurking behavior in online communities. Comput Human Behav 2013 Jan;29(1):293-303. [doi: 10.1016/j.chb.2012.09.008]
31. Bijnsdorp FM, Onwuteaka-Philipsen BD, Boot CRL, van der Beek AJ, Pasman HRW. Caregiver's burden at the end of life of their loved one: insights from a longitudinal qualitative study among working family caregivers. BMC Palliat Care 2022 Aug 10;21(1):142. [doi: 10.1186/s12904-022-01031-1] [Medline: 35945558]

XSL•FO
RenderX

32. Yin Z, Stratton L, Song Q, et al. Perceptions and utilization of online peer support among informal dementia caregivers: survey study. JMIR Aging 2024 May 31;7:e55169. [doi: 10.2196/55169] [Medline: 38825836]
33. Ellis JD, Struble CA, Fodor MC, Cairncross M, Lundahl LH, Ledgerwood DM. Contingency management for individuals with chronic health conditions: a systematic review and meta-analysis of randomized controlled trials. Behav Res Ther 2021 Jan;136:103781. [doi: 10.1016/j.brat.2020.103781] [Medline: 33302054]

Abbreviations

Edited by A Mavragani; submitted 18.12.24; peer-reviewed by R Daynes-Kearney; revised version received 11.04.25; accepted 13.04.25; published 11.06.25.

Please cite as:
Van Citters AD, Holthoff MM, Young C, Eck SM, Cullinan AM, Carney S, O'Donnell EA, King JR, Govindan M, Gustafson D, Tomlin SC, Holmes AB, Bradley AD, Oliver BJ, Wilson MM, Nelson EC, Barnato AE, Kirkland KB
Feasibility and Usability of a Web-Based Peer Support Network for Care Partners of People With Serious Illness (ConnectShareCare): Observational Study
JMIR Form Res 2025;9:e70206
URL: https://formative.jmir.org/2025/1/e70206
doi:10.2196/70206

© Aricca D Van Citters, Megan M Holthoff, Colleen Young, Sarah M Eck, Amelia M Cullinan, Stephanie Carney, Elizabeth A O'Donnell, Joel R King, Malavika Govindan, David Gustafson, Stephanie C Tomlin, Anne B Holmes, Ann D Bradley, Brant J Oliver, Matthew M Wilson, Eugene C Nelson, Amber E Barnato, Kathryn B Kirkland. Originally published in JMIR Formative Research (https://formative.jmir.org), 11.6.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Developing and Testing a Framework for Learning Online Collaborative Creativity in Medical Education: Cross-Sectional Study

Shairah Radzi[1], PhD; Joo Seng Tan[2], PhD; Preman Rajalingam[3], PhD; Jennifer Cleland[1], PhD; Sreenivasulu Reddy Mogali[1], PhD

[1]Lee Kong Chian School of Medicine, Nanyang Technological University Singapore, 11 Mandalay road, Singapore, Singapore
[2]Nanyang Business School, Nanyang Technological University Singapore, Singapore, Singapore
[3]Institute of Learning, Mohammed Bin Rashid University of Medicine and Health Sciences, Dubai, United Arab Emirates

Corresponding Author:
Sreenivasulu Reddy Mogali, PhD
Lee Kong Chian School of Medicine, Nanyang Technological University Singapore, 11 Mandalay road, Singapore, Singapore

Abstract

Background: Collaborative creativity (CC) is a social process of generating creative and innovative solutions to real-world problems through collective effort and interaction. By engaging in this process, medical students can develop abilities and mindset for creative thinking, teamwork, interdisciplinary learning, complex problem-solving, and enhanced patient care. However, medical students have demonstrated limited creativity, constrained by existing pedagogical approaches that predominantly emphasize knowledge outcomes. The increasing complexity of health care challenges necessitates a pedagogical framework for medical students to foster CC in a rapidly evolving professional environment.

Objective: This study aimed to develop, test, and evaluate a new Framework for Learning Online Collaborative Creativity (FLOCC).

Methods: FLOCC builds on established pedagogical approaches such as design thinking and integrates sociocultural learning methods (team-based learning [TBL] and problem-based learning [PBL]). It includes 4 individual asynchronous activities (empathy map, frame your challenge, turning insights into how might we questions, and individual brainstorming) and 5 collaborative synchronous activities (bundle ideas, list constraints, final idea, prototyping, and blind testing). In this cross-sectional study, 85 undergraduate medical students participated in 2 separate studies (study 1, n=44; study 2, n=41) involving health care and engineering sustainability problems. Learner acceptability was measured using a 31-item survey (using 7-point Likert scale) consisting of 4 factors (distributed creativity, synergistic social collaboration, time regulation and achievement, and self and emotions) and 3 free text questions. Free-text comments were subjected to the inductive thematic analysis.

Results: Most students were positive about FLOCC, with distributed creativity and synergistic social collaboration factors receiving the highest mean percentages of "'Agree" (78/85, 92% and 75/85, 88%, respectively). These were followed by time regulation and achievement factor (68/85, 80%) and the self and emotions factor (59/85, 70%). Only time regulation and achievement was statistically significant (P=.001) between means of studies 1 and 2. Thematic analysis revealed 4 themes such as learning experiences, collaborative responsibilities, perceived skill development, and technical challenges.

Conclusions: With effective time management, FLOCC shows potential as a framework for nurturing CC in medical students. Medical schools could provide the opportunity and environment that supports creative thinking; therefore, creativity-focused approaches could be integrated into the curriculum to encourage a culture of creativity for breakthrough solutions by future doctors.

(JMIR Form Res 2025;9:e50912) doi:10.2196/50912

KEYWORDS

collaborative creativity; design thinking; medical education; social cultural learning; collaborative learning

Introduction

Background

Many jobs require graduates to engage in collaboration with other professionals, exhibit creativity in solving difficult problems and manage uncertainties in their careers. In medicine, doctors must work in teams to effectively manage patient care. They are also increasingly expected to work together to solve real-life health care challenges such as managing complex

XSL•FO
RenderX

patient problems in a milieu where specialized medical services and manpower are scarce [1].

Over a century has passed since Flexner's influential report [2], and medical education has undeniably experienced substantial changes, however, traditional medical education fail to cultivate the necessary creativity to handle difficult, ill-defined, or "impossible-to-solve" issues or problems which require multiple stakeholders to collaborate and innovate [3]. By default, pedagogical approaches in medical education tend to emphasize knowledge acquisition, retention, and problem-solving within a limited context [4]. In general, learning in medical school includes 2 types of knowledge: factual and procedural knowledge. Mann comprehensive analysis [5] underscores that sociocultural learning theories, rooted in the work of Vygotsky [6], present a valuable theoretical framework for shaping future medical education.

Significant ramifications arise from major modifications in medical education such as changes in our ways of knowing, changes in the discourse of medical education, changes in our view of medical education, which ultimately impacts the practice of medicine. The recent study by Fernández-Rodríguez found that the traditional teaching and assessment are still overused to the detriment of other teaching methods in medical education [7]. These methods typically measure the knowledge via written examinations focusing individual cognitive knowledge [8], not collaboration and creativity capabilities.

According to Guilford and Christensen [9], creativity is the ability and skill to create novel and innovative things. In Stein [10] classical paper, standard definition of creativity referred as "creative work is a novel work that is accepted as tenable or useful or satisfying by a group in some point in time" [10]. There is no single, agreed-upon definition of creativity due to its multidimensional nature [11]. But most agree that it requires the following 4 skills: the capacity for idea production (or "fluency"), attention to detail (or "elaboration"), originality (or "freshness"), and versatility (or "flexibility") [12-14]. Kampylis and Valtanen [15] conducted an analysis of the 42 explicit definitions of creativity. Walia conceptualizes creativity as an ongoing act, irrespective of whether it results in a tangible creation [16].

In the context of education, collaborative learning is broadly defined as students working in groups of two or more sharing responsibility for tasks and products [17]. The fundamental factors for collaborative learning include respectful behavior, constructive feedback, a shared objective, acceptance of roles, engagement, and self-awareness [18]. The collaboration has been shown to foster the critical thinking by allowing students to debate ideas, engaging higher-order cognitive reasoning [19]. This highlights the importance of incorporating creativity and collaboration in medical curricula to address the problems in rapidly evolving health care system [20]. It also teaches students to view mistakes as opportunities for improvement when they are pushed to be creative [20].

Collaborative Creativity

The reference [21] describe collaborative creativity (CC) as a social process that promotes the creative process in the form of relationships in completing group tasks [21]. The research on CC in medical education research is very limited, even though medical practices require teamwork and creative problem solving [22]. The recent systematic review differentiates CC (explicitly generating new ideas) from creative collaboration (focusing on how people work together, with creativity embedded) [12]. Hill et al [23] emphasize that innovation is not about solo genius [23] but is about the collective genius of collaborative problem solving, requiring 3 capabilities: creative abrasion (idea generation via debate), creative agility (rapid experimentation), and creative resolution (integrative decision-making) [24].

Several other factors impact CC, such as diversity of the team [25,26]; team interaction behaviors and conflict resolution [25,26]; task independence [27,28]; and the online or physical work space [29]. Online learning increased during the COVID-19 pandemic, demonstrating how digital platforms offer real-time cocreation and interaction. The information and communication technologies facilitates collaborative content creation through social contact among users and encourages user-to-user communication and cocreation [24]. Gundogdu and Merc [30] review found that web-based and cloud technologies, simulations, and smart tools help support technology-mediated CC. It was found that student performance is comparable to in-person learning [31]. In terms of questioning behavior and project performance, the online group performed better than the face-to-face group on collaborative tasks [32]. In the creative fields such as design and art, frameworks using technology to promote CC have become more common [33-35]. As higher education continues to expand online, medical programs must adapt to remain globally competitive and foster 21st-century competencies, including collective problem-solving, communication, and creativity [34].

While notable progress has been made in medical technology, education still lacks sufficient emphasis on creativity, which is crucial for problem-solving in the medical field [22]. Medical education often uses instructional methods such as traditional lecture-based, problem-based learning (PBL) and team-based learning (TBL). While the evidence indicates that structure of the PBL and TBL linked to learning enhancement, evidence of systematic attempts to foster CC remains limited [36-39]. Some researchers explored other creative methods in medical education such as brainstorming [40], concept maps [41], and storytelling [42] but CC still receives minimal attention. This is a critical gap in the literature, since CC (or lack thereof) can impact the medical professional-patient relationship [43], job engagement, satisfaction, and performance [44], and patient safety [45]. Thus, a need exists to create strategies and frameworks that incorporate the CC into medical education especially by leveraging online learning technologies to equip future medical professionals for the multifaced realities of evolving health care.

Therefore, the aim was to develop and test a novel design-based thinking framework for online collaborative creativity (FLOCC). Our research questions were: (1) What existing pedagogical principles could be utilized to create a novel FLOCC for fostering CC skills in medical students? (2) How could this

XSL•FO
RenderX

FLOCC be operationalized? (3) What are students' perceptions and attitudes toward the FLOCC?

Methods

Overview

This study reports the development and operationalization of a new FLOCC and its preliminary evaluation. The first stage in this process was to consider the conceptual underpinnings and structure of the FLOCC explicitly. The concepts underpinning FLOCC are depicted in Figure 1 and outlined below.

Figure 1. Conceptual underpinnings of the FLOCC: Framework for Learning Online Collaborative Creativity; PBL: problem-based learning; TBL: team-based learning.

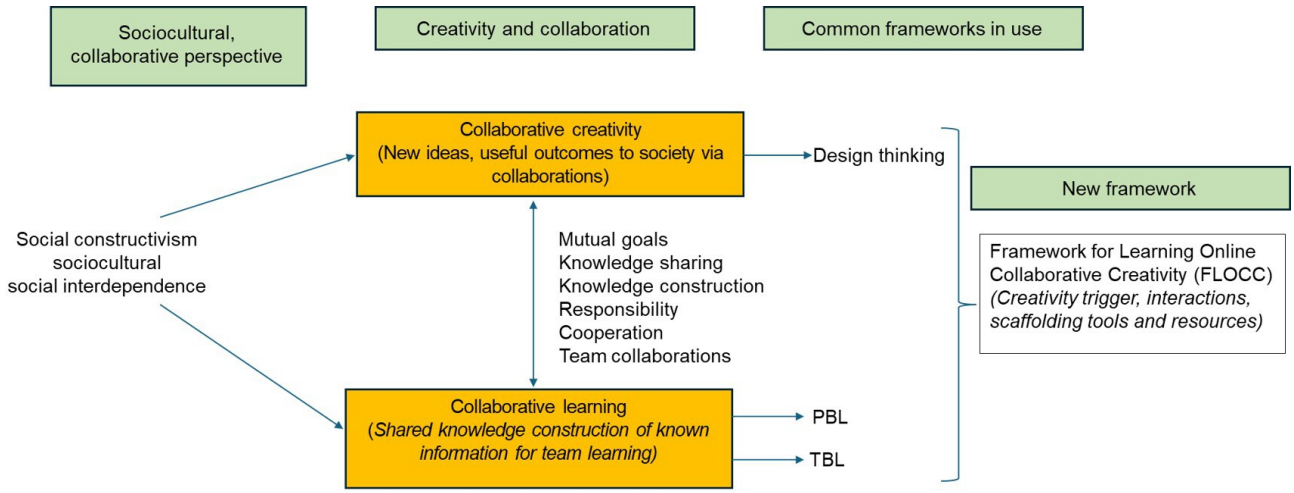

Conceptual Underpinnings of FLOCC

FLOCC integrates the principles of rapid design thinking (RDT) with a sociocultural learning approach, encompassing elements of TBL and PBL. In both TBL and PBL learning takes place through interaction, negotiation, and collaboration in solving clinical problems. Hence, our goal is to identify the convergences of these methodologies within the framework of a collaborative, sociocultural learning grounded in Vygotsky's sociocultural paradigm [6]. Dolmans et al [46] highlights how combining TBL and PBL could synergistically enhance problem solving skills, RDT framework further orients toward innovation and iterative solution creation. The current evidence supports the efficacy of TBL in enhancing the academic performance among the lower-performing medical students [47], while it may be greater demanding [36,38]. Conversely, PBL, has been linked to the development of a broader range of skills such as interpersonal skills, self-directed learning, planning and analytical skills [39]. However, some researchers have shown no difference between PBL and traditional learning in medical education [48] and caution that PBL effectiveness depends on frequent feedback [49].

Design thinking, defined as "a future-oriented method, process, toolbox or mindset" to promote collaborative and creative learning, adopts a human-centered problem-solving approach [50]. While design thinking has seen growing presence in Science, Technology, Engineering, and Mathematics education [51] but its application in medical education remains very limited. A recent review identified only 7 pertinent studies, 3 involving medical students [50,52]. However, while limited in number, these studies and recent opinion pieces suggest design thinking may be an appropriate framework to foster student, patient, and practitioner outcomes such as self-efficacy, learning experiences, and academic development [53,54]. Furthermore, using design thinking opens up a new area for investigation to help educators understand, measure, and assess the experiences of students to promote creativity [55], to create a new product and/or to establish a way of thinking about problems via the development of a novel product [50].

Common characteristics shared by sociocultural learning methods and RDT include goal orientation, team activity, brainstorming, and problem-solving. From these perspectives, learning is impacted at individual and collaborative levels through active engagement and shared goals [56]; building on knowledge through social interaction, constructive arguments and cooperation with others rather than passive learning [57]. From the social constructivist perspective, learning is more efficient in a group setting as knowledge develops as a result of social interaction and language use [58]; hence, students demonstrate more motivation and self-confidence in collaborative learning environments [59]. Through team conversations, students become more adept at identifying problems and pursuing solutions, leading to the process of problem-solving [60]. When individuals are accountable for building their own knowledge and communicating successfully with others, these activities can favorably influence CC [61].

In alignment with the sociocultural learning paradigm and echoing the integrative perspective of Dolmans et al [46], we propose combining TBL, PBL and RDT within FLOCC , harnessing the strengths of each method: TBL's accountability and readiness assurance process [62], PBL's "ill-structured" problem as a stimulus of thinking [63], and RDT's creative process and quick prototyping [64].

XSL•FO
RenderX

Thus, FLOCC can be viewed as a hybrid pedagogical framework that combines key perspectives from PBL, TBL, and design thinking to tackle complex medical problems with CC skills.

Structure of FLOCC

Overview

Process-based frameworks like FLOCC should capture the creative process and promote "mindful idea generation" in a sequential manner [65]. For this, 4 of the investigators (SR, TJS, PR and SRM), all experienced educators in TBL and PBL, adapted the 3 phases of FLOCC (inspiration, ideation, and implementation) from the RDT model, an accelerated version of the Stanford's design thinking model, known as a "people-oriented problem-solving method" [66]. Through iterative discussions and consensus, RDT was chosen as it was especially useful in rapidly changing environments like Covid-19 whilst offering learners similar features to design thinking [64]. FLOCC includes 4 individual asynchronous activities (empathy map, frame your challenge, turning insights into how might we questions, and individual brainstorming) and 5 collaborative synchronous activities (bundle ideas, list constraints, final idea, prototyping and blind testing). These activities are incorporated into the 3 distinct phases of FLOCC, as explained in detail below:

Inspiration

Inspiration refers to defining a problem that can inspire opportunities for creative solution [64]. This phase requires understanding, observing, or listening to people with unmet needs. This means students would requires both communication skills and some domain knowledge [67] to empathize with end users' unmet needs and be open to multiple possibilities. Problems here serve as the creativity trigger [68], akin to PBL, and students may benefit from preparatory materials to support the thought process, similar to pre-TBL preparation. Inspiration uses 2 individual learning scaffolds such as empathy map (I1) and frame your challenge (I2) [69]. These activities help individual learners systematically write down their insights, analyze stakeholders, contexts and think about potential solutions, forming a basis for ideas generation.

Ideation

The ideation phase reinforces observations and experiences recorded during inspiration via the generation of ideas and identification of potential solutions [67,70]. This phase is a dual-purpose scaffold to support individual learners and teams because it requires generating novel ideas or solutions about a topic at an individual level first, then team level. Here, students work individually using the "how might we question" (I3–turning their insights into opportunities for design via divergent thinking) and "Individual Brainstorming" (I4-to write as many ideas as possible based on I3-turning insights into opportunities for design via divergent thinking). Then students do team brainstorming where they combine and filter individual ideas through the team processes such as "bundle ideas" (T1) "list constraints" (T2) and final idea, (T3) and culminating in "prototyping" (T4). Rapid prototyping—a hallmark of design thinking encourages teams to translate their insights into tangible artifacts or solutions [71].

Implementation

Implementation refers to the process where designers test a solution by converting it into a working product and testing its feasibility and market value. This requires problem-solving which can happen via collaboration and mutual feedback [72]. Based on this, the last team-learning scaffold that is, "blind testing" (T5) created where teams assess the usability and feasibility of each other's prototypes and provide feedback to refine them iteratively.

Figure 2 depicts the stages of FLOCC and how it maps onto the characteristics of PBL, TBL and RDT.

XSL•FO
RenderX

Figure 2. Hybrid characteristics of the FLOCC based on the strengths of TBL and PBL, and RDT. FLOCC: Framework for Learning Online Collaborative Creativity; PBL: problem-based learning; RDT: rapid design thinking, TBL: team-based learning.

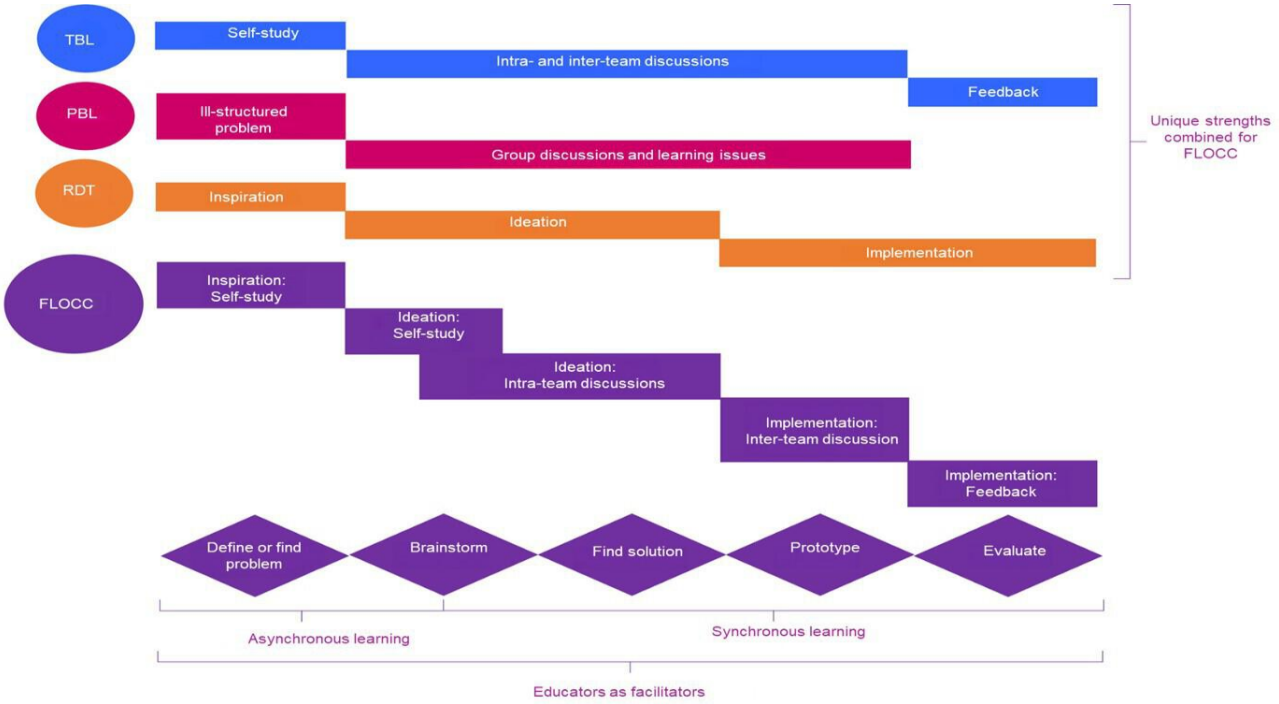

Operationalizing FLOCC

Following conceptualization, our next step was to operationalize FLOCC. The entire study was conducted online due to the COVID-19 pandemic. For operationalizing self-study in individual scaffolds, we provided a project website (Google Sites, Enterprise Standard, Google LLC) hosting explanatory videos about FLOCC and a downloadable Excel spreadsheet for collecting responses (Multimedia Appendix 1) of I1 to I4 individual scaffolds. The individual scaffolds were assigned in an asynchronous online learning format. An overview of the project website and a sample of individual Excel spreadsheets can be found in Multimedia Appendices 1 and 2, respectively.

For team scaffolds (T1 to T4), Google Sheets (Enterprise Standard, Google LLC, USA) was used to collect team responses simultaneously and in real-time. A sample of the Team Google Sheets can be found in Multimedia Appendix 3. In addition, a YouTube (Google) video on prototyping was also included [73].

The overall flow of FLOCC is captured in Figure 3.

XSL•FO
RenderX

Figure 3. The FLOCC, represented by the 3 phases (Inspire, Ideate, and Implement) of the RDT model. Individual and team activities were done via asynchronous and synchronous learning, respectively. FLOCC: Framework for Learning Online Collaborative Creativity; RDT: rapid design thinking.

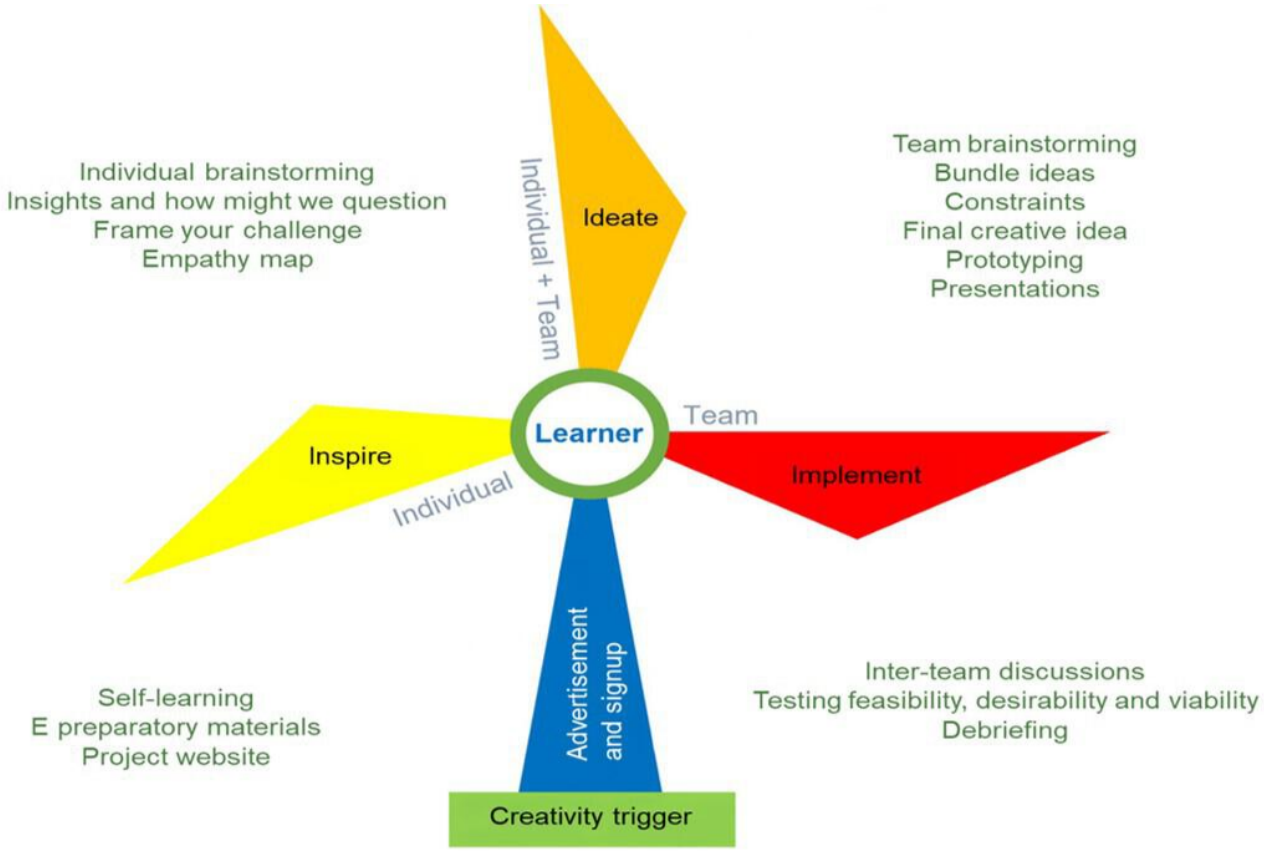

Ethical Considerations

This study was conducted according to the Declaration of Helsinki, approved by Nanyang Technological University institutional review board (IRB 2020-08-036) before its commencement. Students who participated in this study provided full informed consent. The participants were recruited through class announcements and email correspondence. Participation in the study was voluntary, and it was conducted as an extracurricular activity outside of regular school hours. To show their appreciation for taking part in the study, each participant received SGD $50 (US $38.67). The study data are anonymous, and no images of the manuscript or supplementary material identify individual participants.

Evaluation of FLOCC

Study Participants

We carried out a cross-sectional study to evaluate FLOCC using mixed methods to assess student acceptability of FLOCC. Ayala and Elder [74] define acceptability as "how well an intervention will be received by the target population." The participants were undergraduate medical students at Lee Kong Chian School of Medicine, Nanyang Technological University Singapore. The TBL is the main pedagogy at Lee Kong Chian School of Medicine, so some TBL elements in FLOCC might feel familiar to the students. We invited the first 3 years of MBBS students for better representation and to reduce the possibility of selection bias. We also asked them to complete the survey fully to ensure that the data is complete and validate the survey results. The evaluation conducted at 2 points with 2 different creativity triggers: study 1 representing the medical (How might we create sustainable health care practices to cater to the growing population?) and study 2 related to engineering, (How might we create sustainable engineering practices to cater to the growing population?).

Factors to Measure CC

Mavri et al [75] developed the Assessment Scale for Creative Collaboration (ASCC) to measure social CC in blended learning settings. The ASCC measures three factors: (1) synergistic social collaboration (communication and interaction), (2) distributed creativity (divergent thinking and collective innovation), and (3) time regulation and achievement (time management for shared goals.

We adapted 21 items from the ASCC for this study, to quantify the acceptability of the students toward online CC. We added 3 new items to synergistic social collaboration and 2 new items to time regulation and achievement to better reflect the role of facilitation as a potentially motivating factor toward CC, and the level of engagement of students in both individual and team activities used in the study (refer to Multimedia Appendix 4). Finally, in response to literature which indicates that a learner's emotional state may also influence CC [76-78], we also created a new fourth factor called self and emotions, which considered 5 items related to confidence, pride, worry and self-consciousness. In summary, there were a total of 31 items across the 4 factors, all in the format of a 7-point Likert scale (1=strongly disagree to 7=strongly agree). In total, 3 free-text questions were also given to collect students' opinions about

XSL•FO
RenderX

the strengths, weaknesses, and suggestions to improve the FLOCC.

Delivery

Following informed consent, participants received creativity trigger 1 via email. Students were given 1.5 weeks to complete the individual scaffolds (I1 to I4) before engaging in synchronous online team assignments and discussions held in Microsoft Teams for the team activities of FLOCC. The introduction of the study was conducted by SRM and synchronous part of FLOCC was facilitated by PR who explained the ground rules in the main room (eg, respecting the views of others and do not interrupt), the design thinking processes and how to use the scaffolds for each team activity. Another investigator (SR) was also present to support the digital collaborative platforms (Google Workspace and Microsoft Teams) used in the study.

After this, teams (n≤7 per team) were put into breakout rooms to introduce themselves (ice breaker: 7 min), then brought back to the main room for an explanation of each team events. Teams were given weblinks to Google Sheets to complete T1: bundle ideas in 30 minutes. The toggling between the main to breakout rooms were repeated for the rest of the steps in the following order - T2 and T3: list constraints and final idea (30 min), and T4: prototyping (30 min). A team representative was required to talk about their digital products for 5 min using the prototyping section of the Google Sheets. Then, in the Implementation phase, the final team activity blind testing was carried out (10 min). Here, students were asked to complete T5: blind testing in Google Sheets by answering the free text questions and rating the feasibility, viability, and durability of the other team's prototype, as explained in the implementation phase. After study 1, participants received a second trigger (sustainability in engineering) for study 2, following similar procedures. A debrief closed each session. The examples of team's digital products and solutions to triggers 1 and 2 can be found in Multimedia Appendices 5 and 6, respectively.

Data Collection

At the end of each study, the survey (7-point Likert scale items) in Google Forms measured the students' perceptions of the FLOCC. Students were given 20 minutes to complete the survey.

Analysis

Likert-scale questions were analyzed using SPSS (version 28, IBM Corp). Cronbach α was calculated to determine the internal consistency of the survey items within each of the 4 factors (self and emotions, synergistic social collaboration, distributed creativity, and time regulation and achievement). The α value of more than 0.7 was considered satisfactory [78].

After determining the internal consistency of items, the mean and SD, and the percentages of agree, neutral and disagree were consolidated from the 7-point Likert scale at the team, study, and overall cohort levels. This was followed by a Shapiro Wilk and Levene test to determine normality and homogeneity. Data that were normally distributed and equally varied were subjected to an independent samples *t* test and Cohen *d* to determine the statistical significance of means and magnitude (variability) of mean differences [79]. If not normally distributed, data were subjected to the homogeneity of variance for the eligibility of the Mann Whitney U test for statistical calculations based on the mean ranks instead [79]. After that, the eta squared effect size (^{2}) was calculated to determine the magnitude (variability) of the mean ranks. If eta squared values were found to be negative that is, mean scores of study 2 were larger than study 1, they were reported as absolute values for simplicity purposes. A $P<.05$ was considered statistically significant.

The qualitative free-text comments of study 1 were subjected to thematic analysis to gain insights into the FLOCC. The comments selected from the first study were primarily chosen based on factors such as the presence of significant and representative qualitative data, as well as their suitability for thorough analysis. The analysis was done using Nvivo (version 12, QSR International) via inductive coding by 2 coders (SR from the research team and RC external to the study) [80,81]. Any disagreements were resolved by a third coder (SRM).

Results

Demographics

In total, 85 years 1 - 3 MBBS students (38 males, 47 females; aged 18 - 22 y) participated in 2 studies. A total of 44 students participated in study 1 (21 males, 23 females; aged 18 - 22 y), and 41 in study 2 (17 males, 24 females; aged 19 - 22). The mean age in each study and the overall sample population (n=85) was 20.3 (SD 0.9) years. All participants completed the survey.

Internal Consistency of the Questionnaire

Initial Cronbach α for the factors self and emotions, synergistic social collaboration, distributed creativity and time regulation and achievement were 0.42, 0.88, 0.63 and 0.66 respectively. Items 2 "I feel self-conscious when I present my ideas to other people", 21 "We felt pressured to create something original", and 27 "We felt pressured to complete the activity on time" were discarded as their contribution lowered the respective factors Cronbach α <0.70 [82]. Upon their removal, the Cronbach α values improved [83], resulting in a refined 28 item instrument (Multimedia Appendix 4) with α values were 0.75 (self and emotions), 0.88 (synergistic social collaboration), 0.75 (distributed creativity) and 0.74 (time regulation and achievement).

Students' Perceptions on FLOCC

Overall, the study findings (based on refined 28-item instrument) indicated that most students in the total sample (n=85) had positive views of FLOCC across all 4 factors (70% - 92%) (Table 1). Distributed creativity (mean 5.97, SD 0.14) and synergistic social collaboration (mean 5.68, SD 0.2) displayed higher mean scores, followed by time regulation and achievement (mean 5.44, SD 0.22) and self and emotions (mean 5 SD, 0.3). These trends were consistent in both studies (study 1, n=44; study 2, n=41), where distributed creativity and synergistic social collaboration emerged with higher mean scores than the other 2 factors (Table 2).

XSL•FO
RenderX

Table . Summary of percentage frequency results segregated by team, study, and overall sample population levels (n=85) for each of the 4 factors (self and emotions, synergistic social collaboration, distributed creativity, and time regulation and achievement).[a]

	Frequency scores (%)					
	Study 1 (n=44)			Study 2 (n=41)		
Factors and levels	Agree	Neutral	Disagree	Agree	Neutral	Disagree
Self and emotions[b]						
Team 1	66.67	8.33	25	75	10.71	14.29
Team 2	56.25	12.5	31.25	67.86	14.29	17.86
Team 3	78.57	10.71	10.71	79.17	0	20.83
Team 4	67.86	21.43	10.71	70	15	15
Team 5	54.17	8.33	37.5	65	15	20
Team 6	68.75	12.5	18.75	80	20	0
Team 7	46.43	7.14	46.43	83.33	4.17	12.5
Team 8	79.17	8.33	12.5	—[c]	—	—
Study overall	64.77	11.36	23.86	74.39	10.98	14.63
Synergistic social collaboration[d]						
Team 1	91.67	5.56	2.78	94.05	4.76	1.19
Team 2	97.92	2.08	0	90.48	9.52	0
Team 3	83.33	4.76	11.9	84.72	9.72	5.56
Team 4	91.67	8.33	0	90	6.67	3.33
Team 5	76.39	13.89	9.72	90	6.67	3.33
Team 6	89.58	8.33	2.08	90	3.33	6.67
Team 7	84.52	8.33	7.14	77.78	12.5	9.72
Team 8	94.44	2.78	2.78	—	—	—
Study overall	87.88	7.01	5.11	88.21	7.72	4.07
Distributed creativity[e]						
Team 1	94.44	5.56	0	100	0	0
Team 2	95.83	4.17	0	100	0	0
Team 3	88.1	9.52	2.38	94.44	5.56	0
Team 4	92.86	7.14	0	93.33	6.67	0
Team 5	94.44	2.78	2.78	83.33	10	6.67
Team 6	95.83	4.17	0	90	10	0
Team 7	92.86	7.14	0	66.67	25	8.33
Team 8	100	0	0	—	—	—
Study overall	93.94	5.3	0.76	90.24	7.72	2.03
Time regulation and achievement[f]						
Team 1	77.78	11.11	11.11	95.24	2.38	2.38
Team 2	91.67	4.17	4.17	88.1	7.14	4.76
Team 3	69.05	16.67	14.29	80.56	11.11	8.33
Team 4	71.43	21.43	7.14	90	6.67	3.33
Team 5	61.11	13.89	25	70	26.67	3.33
Team 6	87.5	0	12.5	83.33	13.33	3.33
Team 7	64.29	21.43	14.29	86.11	11.11	2.78

XSL•FO
RenderX

	Frequency scores (%)					
	Study 1 (n=44)			Study 2 (n=41)		
Factors and levels	Agree	Neutral	Disagree	Agree	Neutral	Disagree
Team 8	88.89	8.33	2.78	—	—	—
Study overall	74.62	13.64	11.74	85.37	10.57	4.07

[a]Due to the consolidation of responses into the ”Agree,” ”Neutral,” and ”Disagree” categories for each factor, absolute values could not be provided. The reported percentages are based on these consolidated categories.

[b]Population overall—Agree: 69.41; Neutral: 11.18; Disagree: 19.41.

[c]Not applicable.

[d]Population overall—Agree: 88.04; Neutral: 7.35; Disagree: 4.61.

[e]Population overall—Agree: 92.16; Neutral: 6.47; Disagree: 1.37.

[f]Population overall—Agree: 79.80; Neutral: 12.16; Disagree: 8.04.

XSL•FO
RenderX

Table . Summary of mean results segregated by team, study, and overall sample population levels (n=85) for each of the 4 factors (self and emotions, synergistic social collaboration, distributed creativity, and time regulation and achievement).

Factors and levels		Scores (Likert Scale 1 - 7)	
		Study 1 (n=44), mean (SD)	Study 2 (n=41), mean (SD)
Self and emotions[a]			
	Team 1	4.83 (0.63)	5.29 (0.64)
	Team 2	4.81 (0.53)	5.57 (0.53)
	Team 3	5.18 (0.46)	5.33 (0.64)
	Team 4	5.14 (0.18)	4.90 (0.47)
	Team 5	4.38 (0.32)	4.80 (0.58)
	Team 6	4.88 (0.40)	5.65 (0.20)
	Team 7	4.32 (0.12)	5.21 (0.49)
	Team 8	5.17 (0.68)	—[b]
	Study overall	4.84 (0.23)	5.17 (0.40)
Synergistic social collaboration[c]			
	Team 1	5.81 (0.56)	5.82 (0.23)
	Team 2	6.10 (0.33)	5.87 (0.30)
	Team 3	5.43 (0.45)	5.76 (0.26)
	Team 4	5.79 (0.25)	5.67 (0.31)
	Team 5	5.28 (0.38)	5.68 (0.36)
	Team 6	5.44 (0.36)	5.88 (0.51)
	Team 7	5.40 (0.33)	5.39 (0.23)
	Team 8	6.07 (0.37)	—
	Study overall	5.64 (0.23)	5.73 (0.22)
Distributed creativity[d]			
	Team 1	5.94 (0.45)	6.21 (0.16)
	Team 2	6.33 (0.38)	6.38 (0.13)
	Team 3	5.67 (0.17)	6.17 (0.27)
	Team 4	5.93 (0.29)	6.03 (0.42)
	Team 5	5.81 (0.26)	5.70 (0.38)
	Team 6	5.71 (0.32)	6.07 (0.35)
	Team 7	5.93 (0.26)	5.25 (0.29)
	Team 8	6.36 (0.18)	—
	Study overall	5.95 (0.16)	5.99 (0.19)
Time regulation and achievement[e]			
	Team 1	5.17 (0.39)	5.69 (0.25)
	Team 2	5.71 (0.26)	5.69 (0.38)
	Team 3	4.88 (0.31)	5.61 (0.27)
	Team 4	5.24 (0.39)	6.00 (0.27)
	Team 5	4.89 (0.63)	5.37 (0.27)
	Team 6	5.25 (0.32)	5.67 (0.44)
	Team 7	4.93 (0.50)	5.86 (0.33)
	Team 8	5.83 (0.52)	—

XSL•FO
RenderX

Factors and levels	Scores (Likert Scale 1 - 7)	
	Study 1 (n=44), mean (SD)	Study 2 (n=41), mean (SD)
Study overall	5.20 (0.22)	5.70 (0.17)

[a]Population overall: mean 5 (SD 0.30).

[b]Not applicable

[c]Population overall: mean 5.68 (SD 0.20).

[d]Population overall: mean 5.97 (SD 0.14).

[e]Population overall: mean 5.44 (SD 0.22).

The Levene's test showed equality of variances (P>.05) for all factors but the Shapiro Wilk test revealed nonnormally distributed data (P<.05) for 3 factors (distributed creativity in studies 1 and 2 respectively; and synergistic social collaboration in study 2). This meant that statistical calculations related to these nonnormal distributed groups were subjected to the Mann Whitney U test and eta squared effect size (^{2}), and the rest of the normally distributed groups were subjected to the independent samples t test and Cohen d effect size (Table 3).

There was no statistical difference on 3 factors (self and emotions, synergistic social collaboration, and distributed creativity) between the study 1 and study 2 (P=.12, P=.39 and P=.35, respectively; Table 3). However, the factor time regulation and achievement was statistically significant (P=.001; Table 3). Effect size estimates showed larger variability between studies 1 and 2 in self and emotions and time regulation and achievement (^{2}=0.34 and 0.72, respectively) than in synergistic social collaboration and distributed creativity (d=0.01 for both).

Table . Summary of statistical results and effect size between studies 1 and 2, for the four individual factors self and emotions, synergistic social collaboration, distributed creativity, and time regulation and achievement.

Factors and study	Sample size, n	Shapiro wilk (P value)	Levene test (P value)	Independent sample t test or Mann Whitney U test (P value)	Effect size (Cohen d or ^{2})
Self and emotions			.80	.12[a]	−0.34[c]
Study 1	44	.26			
Study 2	41	.31			
Synergistic social collaboration			.18	.39[b]	0.01[d]
Study 1	44	.65			
Study 2	41	<.01			
Distributed creativity			.48	.35[b]	0.01[d]
Study 1	44	.04			
Study 2	41	.01			
Time regulation and achievement			.11	.001[a]	−0.72[c]
Study 1	44	.44			
Study 2	41	.90			

[a]Refer to independent sample t test for normally distributed data.

[b]Refer to Mann Whitney U test for nonnormal distributed data.

[c]Refer to Cohen d values based on magnitude of variability of mean values between studies 1 and 2.

[d]Refer to eta squared (^{2}) values based on magnitude of variability of mean ranks between studies 1 and 2.

Thematic Analysis

Thematic analysis of the free-text responses identified 4 main themes: learning experiences, collaborative responsibilities, perceived skill development, and technical challenges. Quotes are given below to exemplify the open comments.

Learning Experiences

Overall, students reported positive experiences on the structure, and opportunities offered by the FLOCC: "Had a good structure that encouraged participation by all team members of the team" (P#54); "Easy to understand…" (P#02); "Good time allocation" (P#29); and "The challenge we had to solve was interesting and relevant…" (P#43).

They also suggested improvements to the FLOCC, indicating their active involvement and participation: "Maybe include other aspects and not just healthcare…" (P#08); "I think that the original question was rather broad…" (P#23); and "Timings given could be a bit longer..." (P#56).

XSL•FO
RenderX

Collaborative Responsibilities

Students perceived sharing of thinking processes, information, team diversity and team dynamics as key elements for CC: “We could share our thinking process with each other” (P#01); “It gives students an opportunity to work together and exercise teamwork.” (P#24).

There were mixed opinions related to group formation and group diversity: “Interesting mix of people” (P#25); “Forming groups with students with different backgrounds may allow us to have a more fruitful discussion with different points of view…” (P#02).

Perceived Skill Development

Students stated that FLOCC was useful in gaining skills such as design thinking, creativity, communication, and social learning, and self-reflection: “I think that the pre team activity materials were very helpful in helping us understand the design thinking process….” (P#28); “Provides a conducive environment for us to think out of the box and come up with novel ideas for issues we think need to be addressed in the healthcare industry” (P#04); “Team activity allowed me to understand different perspectives I haven’t considered before” (P#35).

Technical Challenges

As mentioned earlier, because of the Covid-19 pandemic, FLOCC was carried out online. Feedback indicated students—who were very used to online learning at this time–would have preferred face-to-face interactions: “Feel that the discussions were slightly abrupt as they were carried out online” (P#46); “New faces+online (more difficult to share ideas with)” (P#37); and “A digital medium makes teamwork a little more challenging” (P#58).

Discussion

Principal Findings

This paper reported the development and evaluation of a novel FLOCC intervention that draws on TBL, PBL and RDT principles to encourage CC in medical students. This paper provided sufficient details of the intervention to ensure clarity and replicability. The evaluation data from 85 participants indicate a high level of acceptability for FLOCC. Student responses also suggest that the FLOCC offers students with the opportunity to acquire, hone, and demonstrate their CC skills given the right tools, flexibility, and appropriate educational support.

We drew on design thinking principles for the foundation of FLOCC, as recent literature has suggested that design thinking may be an appropriate framework to foster student, patient, and practitioners’ outcomes such as self-efficacy, learning experiences, and academic development [50,52-54]. However, unlike previous studies, we did not base FLOCC exclusively on design thinking principles. Instead, we used a blended approach that draws on sociocultural learning (PBL and TBL) principles to develop a robust conceptual framework. The evaluation results suggest that FLOCC was effective in assisting students in fostering CC skills. Furthermore, the adaptation of the ASCC instrument facilitated the measurement the CC outcomes [75].

Other studies highlight that more supervision may be appreciated when working with students from different educational backgrounds [53], and that facilitators play an important role in creating a positive learning environment [84]. Hence, we incorporated a specific survey item to measure students’ perceptions about the quality of the facilitation received from instructors leading the FLOCC. The qualitative data from the open comments indicated that students valued effective facilitation. Similarly, we added a factor “Self and Emotions.” Although feelings and emotions can influence the quality of creative outputs [39,47], these are rarely assesses in design thinking models and CC. The importance of these factors to learning was reflected in the students’ open comments.

To the authors’ knowledge, this is the first study to measure the 4 factors of CC in the context of medical education. This study’s new observations were that students (overall population) perceived higher emphasis on Distributed Creativity (78/85, 92%) and Synergistic Social Collaboration (75/85, 88%) followed by time regulation and achievement (68/85, 80%), and self and emotions (59/85, 70%). This suggests that PBL, TBL, and RDT theoretical perspectives, along with selected technologies of FLOCC, enabled a virtual environment where participants could actively engage in teams to practice collaboration, communication, collective knowledge generation, design thinking, and collaborative problem solving. These observations may come from the ability to relate their educational background and experiences to the creativity trigger, and the perceived proximity from working virtually with other students in the same discipline [85]. On the other hand, previous studies have reported forming interdisciplinary and multidisciplinary teams can be beneficial, depending on the nature of the task [86,87]. Therefore, it would be interesting to provide a trigger that could be beyond students’ educational background and form multidisciplinary teams to measure the effects of the FLOCC process on outcomes of CC.

When comparing 2 studies, results showed no significant differences in any of the variables tested, except hose pertaining to time regulation and achievement. This suggests that FLOCC is successful in fostering teamwork and scaffolding the multifaceted creative process, regardless of problem type, time periods, or study population. Based on these findings, the FLOCC structure is robust and may be safe for use in other contexts. Regarding differences in perceived time regulation and achievement, we believe that these differences are a result of the technical difficulties encountered in study 1 and the fact that certain students may find the online environment challenging. Previous research indicated that challenges associated with technology, students, and facilitators are not necessarily barrier to CC, but rather as opportunities to enhance it through problem solving, being considerate of others, and appreciating the teacher’s effort [30].

Limitations and Future Research Recommendation

There were several limitations in this study. First, it involved volunteer medical students from a single medical school in Singapore. The views of more students at different levels of

XSL•FO
RenderX

medical education, and from other medical schools could be gathered in future research. Second, participants were experienced with TBL, self-selecting and may have had more of an interest in design thinking and collective collaboration than typical medical students. Consequently, participants of this study may have different views on group work and differing group decision-making skills. However, the sample size was reasonable for an exploratory "proof of concept" and acceptability study. As is often the case with online education, it is difficult to separate out student satisfaction with the content and pedagogy of teaching as opposed to the quality of internet connections and technical difficulties. The current evidence is mainly based on the students' self-reported data about their experience with FLOCC. Future research could investigate whether FLOCC is effective and efficient in nurturing CC skills for medical education or practice, as compared to established pedagogies such as PBL and TBL.

Realistically, CC skills cannot be developed in a short time. Therefore, we suggest that the principles underpinning FLOCC should ultimately be used in a curricular (rather than one-off) basis if fostering collective creativity is a priority in medical education.

Implications

FLOCC in the digital environment opened up new possibilities such as the opportunity to understand the "precise mechanisms" needed to acquire and apply new or innovative knowledge. The framework also underpinned the monitoring of individual and team creativity processes in an organized manner. This is important because if the collection of data is unorganized, comprehending and analyzing the stakeholders' reflections would become challenging and almost impossible to measure and to provide feedback [56]. In the context of medical education and practice, this might influence the past, present and future interpretations of appropriate patient care [64]. Educators can leverage FLOCC as a model to design and modify learning activities that further CC and connect the classroom to real-world settings. Testing this framework face-to-face, with other educational intra and interdisciplines (engineering, sciences, business, and others), and a larger sample size may also form part of future studies.

Conclusions

The FLOCC model, which integrates the design thinking and sociocultural learning methods (PBL and TBL), holds considerable promise as a framework for cultivating CC in medical education. Further research is needed to evaluate its effectiveness across intermultidisciplinary teams and diverse creativity triggers, thereby uncovering its full potential.

Acknowledgments

We would like to acknowledge the work of Ms Ramya Chandrasekaran for participating in mock studies to test the digital tools. We also thank the students who took part in the evaluation. This work was supported by the Singapore Ministry of Education, under the grant 'Singapore Ministry of Education Tertiary Education Research Fund' (MOE2019-TRF-027).

Data Availability

Student responses from the survey are not publicly accessible in accordance with ethical requirements by the university. Details of student responses are available upon reasonable request, following ethical practices of the university, from the corresponding author.

Authors' Contributions

SR designed and planned the experiment, collected data, analyzed the results, and wrote the manuscript. JST designed the experiment, conceived analytical methods, and assisted with writing of manuscript. PR designed the experiment, facilitated the study, and assisted with the writing of manuscript. JC contributed to the literature review and data analysis and assisted with the writing of manuscript. SRM conceptualized and designed the experiment, assisted with data analysis and writing of manuscript, and supervised SR.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Hybrid characteristics of the Framework for Learning Online Collaborative Creativity (FLOCC) based on the strengths of team- and problem-based learning (TBL and PBL), and rapid design thinking (RDT).
[PDF File, 619 KB - formative_v9i1e50912_app1.pdf]

Multimedia Appendix 2
Conceptual underpinnings of the Framework for Learning Online Collaborative Creativity (FLOCC).
[PDF File, 75 KB - formative_v9i1e50912_app2.pdf]

XSL•FO
RenderX

Multimedia Appendix 3
Pictorial illustration of website.
[PDF File, 453 KB - formative_v9i1e50912_app3.pdf]

Multimedia Appendix 4
The FLOCC (Framework for Learning Online Collaborative Creativity), represented by the 3 phases (Inspire, Ideate and Implement) of the rapid design thinking (RDT) model. Individual and team activities were done via asynchronous and synchronous learning, respectively.
[DOCX File, 19 KB - formative_v9i1e50912_app4.docx]

Multimedia Appendix 5
Examples of team's digital products and solutions to trigger 1.
[XLSX File, 1666 KB - formative_v9i1e50912_app5.xlsx]

Multimedia Appendix 6
Examples of team's digital products and solutions to trigger 2.
[XLSX File, 666 KB - formative_v9i1e50912_app6.xlsx]

References

1. Hrynchak P, Batty H. The educational theory basis of team-based learning. Med Teach 2012;34(10):796-801. [doi: 10.3109/0142159X.2012.687120] [Medline: 22646301]
2. Flexner A. Medical education in the United States and Canada. Science 1910 Jul 8;32(810):41-50. [doi: 10.1126/science.32.810.41]
3. Ferraro F, Etzion D, Gehman J. Tackling grand challenges pragmatically: robust action revisited. Organization Studies 2015 Mar;36(3):363-390. [doi: 10.1177/0170840614563742]
4. Kamei RK, Cook S, Puthucheary J, Starmer CF. 21st century learning in medicine: traditional teaching versus team-based learning. MedSciEduc 2012 Jun;22(2):57-64. [doi: 10.1007/BF03341758]
5. Mann KV. Theoretical perspectives in medical education: past experience and future possibilities. Med Educ 2011 Jan;45(1):60-68. [doi: 10.1111/j.1365-2923.2010.03757.x] [Medline: 21155869]
6. Vygotsky LS, Cole M. Mind in Society: Development of Higher Psychological Processes: Harvard University Press; 1978.
7. Fernández-Rodríguez CA, Arenas-Fenollar MC, Lacruz-Pérez I, Tárraga-Mínguez R. Teaching methods in medical education: an analysis of the assessments and preferences of students. Sustainability 2023;15(11):9044. [doi: 10.3390/su15119044]
8. Jones RW. Problem-based learning: description, advantages, disadvantages, scenarios and facilitation. Anaesth Intensive Care 2006 Aug;34(4):485-488. [doi: 10.1177/0310057X0603400417] [Medline: 16913347]
9. Guilford JP, Christensen PR. The one - way relation between creative potential and IQ*. J Creat Behav 1973 Dec;7(4):247-252. [doi: 10.1002/j.2162-6057.1973.tb01096.x]
10. Stein MI. Creativity and culture. J Psychol 1953 Oct;36(2):311-322. [doi: 10.1080/00223980.1953.9712897]
11. de Sousa FC. Still the elusive definition of creativity. Int J Psychol Biopsychosoc Approach 2008;2:55-82 [FREE Full text]
12. Daemi H, Moghimi Barforoosh F. Normalization of the creativity test. Adv Cogn Sci 2004;6(3):1-8 [FREE Full text]
13. Dietrich A. The cognitive neuroscience of creativity. Psychonomic Bulletin & Review 2004 Dec;11(6):1011-1026. [doi: 10.3758/BF03196731]
14. Sternberg RJ, Lubart TI. The concept of creativity: prospects and paradigms. In: Handbook of Creativity: Cambridge University Press; 1999. [doi: 10.1017/CBO9780511807916.003]
15. Kampylis PG, Valtanen J. Redefining creativity — analyzing definitions, collocations, and consequences. J Creat Behav 2010 Sep;44(3):191-214. [doi: 10.1002/j.2162-6057.2010.tb01333.x]
16. Walia C. A dynamic definition of creativity. Creat Res J 2019 Jul 3;31(3):237-247. [doi: 10.1080/10400419.2019.1641787]
17. Smith BL, MacGregor JT. What is collaborative learning? : Washington Center for Improving the Quality of Undergraduate Education; 1992 URL: https://teach.ufl.edu/wp-content/uploads/2016/07/WhatisCollaborativeLearning.pdf [accessed 2025-05-22]
18. Iqbal M, Velan GM, O'Sullivan AJ, Balasooriya C. Differential impact of student behaviours on group interaction and collaborative learning: medical students' and tutors' perspectives. BMC Med Educ 2016 Aug 22;16(1):217. [doi: 10.1186/s12909-016-0730-1] [Medline: 27549085]
19. Pervaz Iqbal M, Velan GM, O'Sullivan AJ, Balasooriya C. The collaborative learning development exercise (CLeD-EX): an educational instrument to promote key collaborative learning behaviours in medical students. BMC Med Educ 2020 Mar 2;20(1):62. [doi: 10.1186/s12909-020-1977-0] [Medline: 32122344]
20. Kopel J, Brower G, Culberson JW. Teaching methods fostering enjoyment and creativity in medical education. J Community Hosp Intern Med Perspect 2021;11(6):821-824. [doi: 10.1080/20009666.2021.1979739] [Medline: 34804398]
21. Miell D, Littleton K. Collaborative Creativity: Contemporary Perspectives: Free Association Books; 2004.

XSL•FO
RenderX

22. Baruah J, Paulus PB. Collaborative creativity and innovation in education. In: Creativity Under Duress in Education? Resistive Theories, Practices, and Actions: Springer; 2019:155-177. [doi: 10.1007/978-3-319-90272-2_9]
23. Hill LA, Brandeau G, Truelove E, Lineback K. Collective genius. Harv Bus Rev 2014 Jun;92(6):94-102. [Medline: 25051858]
24. Wheeler S, Waite SJ, Bromfield C. Promoting creative thinking through the use of ICT. Computer Assisted Learning 2002 Sep;18(3):367-378. [doi: 10.1046/j.0266-4909.2002.00247.x]
25. Eichbaum Q. Collaboration and teamwork in the health professions: rethinking the role of conflict. Acad Med 2018 Apr;93(4):574-580. [doi: 10.1097/ACM.0000000000002015] [Medline: 29140919]
26. Moser KS, Dawson JF, West MA. Antecedents of team innovation in health care teams. Creat Innov Manage 2019 Mar;28(1):72-81. [doi: 10.1111/caim.12285]
27. Campion MA, Medsker GJ, Higgs AC. Relations between work group characteristics and effectiveness: implications for designing effective work groups. Pers Psychol 1993 Dec;46(4):823-847. [doi: 10.1111/j.1744-6570.1993.tb01571.x]
28. Farh CIC, Zhang Y, Long L. Designing creative teams from creative members: implications for task interdependence. AMPROC 2015 Jan;2015(1):11535. [doi: 10.5465/ambpp.2015.11535abstract]
29. Hilliges O, Terrenghi L, Boring S, Kim D, Richter H, Butz A. Designing for collaborative creative problem solving. In: Butz A, editor. 2007 Presented at: the 6th ACM SIGCHI conference; Jun 13, 2007; Washington, DC, USA. [doi: 10.1145/1254960.1254980]
30. Gündoğdu B, Merç A. A systematic review of tech-supported collaborative creativity practices in the field of education. J learn teach digit age 2021;7(1):76-89. [doi: 10.53850/joltida.953760]
31. Paul J, Jefferson F. A comparative analysis of student performance in an online vs. face-to-face environmental science course from 2009 to 2016. Front Comput Sci 2019;1:7. [doi: 10.3389/fcomp.2019.00007]
32. Tutty JI, Klein JD. Computer-mediated instruction: a comparison of online and face-to-face collaboration. Education Tech Research Dev 2008 Apr;56(2):101-124. [doi: 10.1007/s11423-007-9050-9]
33. Sykes J. Locating the value and opportunities for online collaborative creativity within advertising. ADCHE 2012 Dec 1;11(2):91-109. [doi: 10.1386/adch.11.2.91_1]
34. Joy KK. A proposed model to increase creativity, collaboration and accountability in the online classroom. IJIET 2015;5(11):873-876. [doi: 10.7763/IJIET.2015.V5.630]
35. Adiloğlu F, Fragiacomo F, Petricone F. Distance artist: building the skills of future creatives. Developing evidence‐based criteria for global virtual team tutoring and management in art and design education. Int J Art Design Ed 2021 Feb;40(1):268-282. [doi: 10.1111/jade.12336]
36. Fatmi M, Hartling L, Hillier T, Campbell S, Oswald AE. The effectiveness of team-based learning on learning outcomes in health professions education: BEME Guide No. 30. Med Teach 2013 Dec;35(12):e1608-e1624. [doi: 10.3109/0142159X.2013.849802] [Medline: 24245519]
37. Ma Y, Lu X. The effectiveness of problem-based learning in pediatric medical education in China: a meta-analysis of randomized controlled trials. Medicine (Baltimore) 2019 Jan;98(2):e14052. [doi: 10.1097/MD.0000000000014052] [Medline: 30633204]
38. Reimschisel T, Herring AL, Huang J, Minor TJ. A systematic review of the published literature on team-based learning in health professions education. Med Teach 2017 Dec;39(12):1227-1237. [doi: 10.1080/0142159X.2017.1340636] [Medline: 28664760]
39. Schmidt HG, Vermeulen L, van der Molen HT. Longterm effects of problem-based learning: a comparison of competencies acquired by graduates of a problem-based and a conventional medical school. Med Educ 2006 Jun;40(6):562-567. [doi: 10.1111/j.1365-2929.2006.02483.x] [Medline: 16700772]
40. Goswami B, Jain A, Koner BC. Evaluation of brainstorming session as a teaching-learning tool among postgraduate medical biochemistry students. Int J Appl Basic Med Res 2017 Dec;7(Suppl 1):S15-S18. [doi: 10.4103/ijabmr.IJABMR_191_17] [Medline: 29344451]
41. Chan ZCY. A qualitative study on using concept maps in problem-based learning. Nurse Educ Pract 2017 May;24:70-76. [doi: 10.1016/j.nepr.2017.04.008] [Medline: 28414998]
42. Liao HC, Wang YH. Storytelling in medical education: narrative medicine as a resource for interdisciplinary collaboration. Int J Environ Res Public Health 2020 Feb 11;17(4):1135. [doi: 10.3390/ijerph17041135] [Medline: 32053911]
43. Macdonald AS. Design for Health: Taylor & Francis Group; 2017:311-327 URL: https://www.taylorfrancis.com/chapters/edit/10.4324/9781315576619-19 [accessed 2025-05-07]
44. Waheed J, Jun W, Yousaf Z, Radulescu M, Hussain H. Towards employee creativity in the healthcare sector: investigating the role of polychronicity, job engagement, and functional flexibility. In: Hussain H, editor. Healthcare (Basel) 2021 Jul 1;9(7):837. [doi: 10.3390/healthcare9070837] [Medline: 34356214]
45. Subeq YM. The importance of cross-disciplinary technology creativity in the field of healthcare. Hu Li Za Zhi 2019 Apr;66(2):4-5. [doi: 10.6224/JN.201904_66(2).01] [Medline: 30924508]
46. Dolmans D, Michaelsen L, van Merriënboer J, van der Vleuten C. Should we choose between problem-based learning and team-based learning? No, combine the best of both worlds!. Med Teach 2015 Apr;37(4):354-359. [doi: 10.3109/0142159X.2014.948828] [Medline: 25154342]

XSL•FO
RenderX

47. Koles PG, Stolfi A, Borges NJ, Nelson S, Parmelee DX. The impact of team-based learning on medical students' academic performance. Acad Med 2010 Nov;85(11):1739-1745. [doi: 10.1097/ACM.0b013e3181f52bed] [Medline: 20881827]
48. Nandi PL, Chan JNF, Chan CPK, Chan P, Chan LPK. Undergraduate medical education: comparison of problem-based learning and conventional teaching. Hong Kong Med J 2000 Sep;6(3):301-306. [Medline: 11025850]
49. Yiou R, Goodenough D. Applying problem-based learning to the teaching of anatomy: the example of Harvard Medical School. Surg Radiol Anat 2006 May;28(2):189-194. [doi: 10.1007/s00276-005-0062-z] [Medline: 16521048]
50. Sandars J, Goh PS. Design thinking in medical education: the key features and practical application. J Med Educ Curric Dev 2020;7:2382120520926518. [doi: 10.1177/2382120520926518] [Medline: 32548307]
51. Li Y, Schoenfeld AH, AA, Graesser AC, Benson LC, English LD, et al. Design and Design Thinking in STEM Education: Springer; 2019:93-104. [doi: 10.1007/s41979-019-00020-z]
52. McLaughlin JE, Wolcott MD, Hubbard D, Umstead K, Rider TR. A qualitative review of the design thinking framework in health professions education. BMC Med Educ 2019 Apr 4;19(1):98. [doi: 10.1186/s12909-019-1528-8] [Medline: 30947748]
53. van de Grift TC, Kroeze R. Design thinking as a tool for interdisciplinary education in health care. Acad Med 2016 Sep;91(9):1234-1238. [doi: 10.1097/ACM.0000000000001195] [Medline: 27097052]
54. Wolstenholme D, Downes T, Leaver J, Partridge R, Langley J. Improving self-efficacy in spinal cord injury patients through "design thinking" rehabilitation workshops. BMJ Qual Improv Report 2014;3(1):u205728. [doi: 10.1136/bmjquality.u205728.w2340]
55. Shively K, Stith KM, Rubenstein LD. Measuring what matters: assessing creativity, critical thinking, and the design process. Gifted Child Today 2018 Jul;41(3):149-158. [doi: 10.1177/1076217518768361]
56. Shimizu I, Kikukawa M, Tada T, Kimura T, Duvivier R, van der Vleuten C. Measuring social interdependence in collaborative learning: instrument development and validation. BMC Med Educ 2020 Jun 1;20(1):177. [doi: 10.1186/s12909-020-02088-3] [Medline: 32487067]
57. Hirtle JSP. Coming to terms: social constructivism. The English Journal 1996 Jan;85(1):91. [doi: 10.2307/821136]
58. Palincsar AS. Social constructivist perspectives on teaching and learning. Annu Rev Psychol 1998;49(1):345-375. [doi: 10.1146/annurev.psych.49.1.345] [Medline: 15012472]
59. Tran VD. Does cooperative learning increase students' motivation in learning? IJHE 2019;8(5):12. [doi: 10.5430/ijhe.v8n5p12]
60. Huang R, Liu D, Tlili A, Yang J, Wang H. Handbook on Facilitating Flexible Learning during Educational Disruption: The Chinese Experience in Maintaining Undisrupted Learning in COVID-19 Outbreak: Smart Learning Institute of Beijing Normal University; 2020.
61. Mahaux M, Nguyen L, Mich L, Mavin A. A framework for understanding collaborative creativity in requirements engineering: empirical validation. Presented at: 2014 IEEE Fourth International Workshop on Empirical Requirements Engineering (EmpiRE); Aug 25, 2014; Karlskrona, Sweden. [doi: 10.1109/EmpiRE.2014.6890116]
62. Schmidt HG, Rotgans JI, Rajalingam P, Low-Beer N. A psychological foundation for team-based learning: knowledge reconsolidation. Acad Med 2019 Dec;94(12):1878-1883. [doi: 10.1097/ACM.0000000000002810] [Medline: 31149922]
63. Laxman K. A conceptual framework mapping the application of information search strategies to well and ill-structured problem solving. Comput Educ 2010 Sep;55(2):513-526. [doi: 10.1016/j.compedu.2010.02.014]
64. Thakur A, Soklaridis S, Crawford A, Mulsant B, Sockalingam S. Using rapid design thinking to overcome COVID-19 challenges in medical education. Acad Med 2021 Jan 1;96(1):56-61. [doi: 10.1097/ACM.0000000000003718] [Medline: 32889947]
65. Howard TJ, Culley SJ, Dekoninck E. Describing the creative design process by the integration of engineering design and cognitive psychology literature. Design Studies 2008 Mar;29(2):160-180. [doi: 10.1016/j.destud.2008.01.001]
66. Auernhammer J, Roth B. The origin and evolution of Stanford University's design thinking: from product design to design thinking in innovation management. J of Product Innov Manag 2021 Nov;38(6):623-644. [doi: 10.1111/jpim.12594]
67. Kirschner PA, Sweller J, Kirschner F, Zambrano R J. From Cognitive Load Theory to Collaborative Cognitive Load Theory. Int J Comput Support Collab Learn 2018;13(2):213-233. [doi: 10.1007/s11412-018-9277-y] [Medline: 30996713]
68. Brenner W, Uebernickel F, Abrell T. Design thinking as mindset, process, and toolbox: experiences from research and teaching at the University of St.Gallen. In: Design Thinking for Innovation: Research and Practice: Springer; 2016:3-21. [doi: 10.1007/978-3-319-26100-3_1]
69. IDEO. The Field Guide to Human-Centered Design: Design Kit; 2015.
70. Brown VR, Paulus PB. Making group brainstorming more effective: recommendations from an associative memory perspective. Curr Dir Psychol Sci 2002 Dec;11(6):208-212. [doi: 10.1111/1467-8721.00202]
71. Beaudouin-Lafon M, Mackay WE. Prototyping Tools and Techniques The Human-Computer Interaction Handbook: CRC Press; 2007:1043-1066. [doi: 10.1201/9781410615862-66]
72. Stahl G. The constitution of group cognition. In: The Routledge Handbook of Embodied Cognition: Routledge; 2014:335-346.
73. BUILD. Student solutions | BUILD's COVID-19 virtual design challenge. YouTube. 2020. URL: https://www.youtube.com/watch?v=6AdzkPK3D3c [accessed 2025-05-07]

XSL•FO
RenderX

74. Ayala GX, Elder JP. Qualitative methods to ensure acceptability of behavioral and social interventions to the target population. J Public Health Dent 2011;71 Suppl 1(1):S69-S79. [doi: 10.1111/j.1752-7325.2011.00241.x] [Medline: 21656958]
75. Mavri A, Ioannou A, Loizides F. The Assessment Scale for Creative Collaboration (ASCC) validation and reliability study. Int J Hum Comput 2020 Jul 2;36(11):1056-1069. [doi: 10.1080/10447318.2019.1709338]
76. Stets J, Turner JH. Handbook of the Sociology of Emotions: Springer; 2006. [doi: 10.1007/978-0-387-30715-2]
77. Tracy JL, Robins RW, Tangney JP. The Self-Conscious Emotions: Theory and Research The Self-Conscious Emotions Theory and Research: Guilford Press; 2007.
78. Bland JM, Altman DG. Statistics notes: Cronbach's alpha. BMJ 1997 Feb 22;314(7080):572-572. [doi: 10.1136/bmj.314.7080.572]
79. Tomczak M, Tomczak E. The need to report effect size estimates revisited. An overview of some recommended measures of effect size. TRENDS in Sport Sciences 2014;1(21):19-25 [FREE Full text]
80. Boyatzis RE. Transforming Qualitative Information: Thematic Analysis and Code Development: Sage Publications; 1998.
81. Jebreen I. Using inductive approach as research strategy in requirements engineering. IJCIT 2012;1(2):162-173.
82. Vaske JJ, Beaman J, Sponarski CC. Rethinking internal consistency in Cronbach's alpha. Leis Sci 2017;39(2):163-173. [doi: 10.1080/01490400.2016.1127189]
83. Hajjar S. Statistical analysis: Internal-consistency reliability and construct validity. IJQQRM 2018;6(1):27-38 [FREE Full text]
84. Ringby B, Duus L. Innovation camp as a method in health education: a study on interdisciplinarity, learning and participation. Eur J Physiother 2017 Oct 31;19(sup1):22-24. [doi: 10.1080/21679169.2017.1381308]
85. Wilson JM, Boyer O'Leary M, Metiu A, Jett QR. Perceived proximity in virtual work: explaining the paradox of far-but-close. Organization Studies 2008 Jul;29(7):979-1002. [doi: 10.1177/0170840607083105]
86. Silva A, Leite M, Vilas-Boas J, Simões R. How education background affects design outcome: teaching product development to mechanical engineers, industrial designers and managers. Eur J Eng Educ 2019 Jul 4;44(4):545-569. [doi: 10.1080/03043797.2018.1465029]
87. Stroeken JHM, De Vries MJ. Learning to deal with social factors as a goal in the education of engineers. Eur J Eng Educ 1995 Jan;20(4):447-456. [doi: 10.1080/0304379950200405]

Abbreviations

ASCC: Assessment Scale for Creative Collaboration
CC: collaborative creativity
FLOCC: Framework for Learning Online Collaborative Creativity
PBL: problem-based learning
RDT: rapid design thinking
TBL: team-based learning

Edited by A Mavragani; submitted 18.07.23; peer-reviewed by IT Isip-Tan, L Selvaratnam, L Michaelson; revised version received 31.03.25; accepted 01.04.25; published 05.06.25.

Please cite as:
Radzi S, Tan JS, Rajalingam P, Cleland J, Mogali SR
Developing and Testing a Framework for Learning Online Collaborative Creativity in Medical Education: Cross-Sectional Study
JMIR Form Res 2025;9:e50912
URL: https://formative.jmir.org/2025/1/e50912
doi:10.2196/50912

© Shairah Radzi, Joo Seng Tan, Preman Rajalingam, Jennifer Cleland, Sreenivasulu Reddy Mogali. Originally published in JMIR Formative Research (https://formative.jmir.org), 5.6.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Personalized Support in Hereditary Breast and Ovarian Cancer After Genetic Counseling by the Chatbot-Based GENIE Mobile App: Proof-of-Concept Wizard of Oz Study

Dominik Wolff[1], PhD; Thomas Kupka[1], PhD; Chiara Reichert[1]; Nils Ammon[2]; Steffen Oeltze-Jafra[1], PhD; Beate Vajen[2], PhD

[1]Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School, Hannover Medical School, Hanover, Germany
[2]Department for Human Genetics, Hannover Medical School, Carl-Neuberg-Str. 1, Hanover, Germany

Corresponding Author:
Beate Vajen, PhD
Department for Human Genetics, Hannover Medical School, Carl-Neuberg-Str. 1, Hanover, Germany

Abstract

Background: The primary aim of genetic counseling at a human genetics center is to empower individuals at risk for hereditary diseases to make informed decisions regarding their health. In Germany, genetic counseling sessions typically last approximately 1 hour and provide highly personalized information by a specialist in human genetics. Despite this, many counselees report a need for additional support following the counseling session.

Objective: This study introduces GENIE, a chatbot-based mobile app designed to assist individuals in the postcounseling phase, with a focus on hereditary breast and ovarian cancer. GENIE delivers expert-curated, personalized information tailored to the user's health and family circumstances. The content is presented through predefined dialogs between the user and the mobile assistant, aiming to extend the benefits of genetic counseling beyond the initial session.

Methods: A Wizard of Oz study was conducted to evaluate a functional prototype of GENIE. A total of 6 patients with breast cancer, at least 2 years postdiagnosis, participated in the study. Participants were given access to the app for a minimum of 1 week. The evaluation was based on their interaction with GENIE, which was personalized using the details of a fictitious patient. Data collection included semistructured interviews and a 45-item questionnaire to assess usability and content quality.

Results: The analysis of the interview and questionnaire data indicated high usability for GENIE, with a mean System Usability Score of 75.33 (SD 4.13). In total, 5 of the 6 participants used the app daily; 3 participants were willing to pay between US $5 and US $45 as a single purchase, while the other 3 participants agreed that the app should be free for the user and the costs should be directly covered by health insurance. Still, opinions on the app's appeal were divided. The layout was seen as moderately professional, a bit crowded, and slightly uninspiring. Nevertheless, participants highlighted the credibility and relevance of the content, noting its alignment with the fictitious patient's scenario. However, areas for improvement were identified, particularly concerning the app's design. All participants would recommend the app to other affected persons.

Conclusions: The findings suggest that a mobile app like GENIE can provide valuable support to individuals in the postcounseling phase of genetic services. GENIE offers distinct advantages over large language models, as the information it provides is carefully curated by human experts, minimizing the risk of inaccuracies or hallucinations and significantly enhancing the system's credibility. This study highlights the need to involve the user group as early as possible in the development of a digital health app. Future work will focus on the implementation of a comprehensive personalization engine, redesign of the user interface, and the execution of a large-scale, 2-arm randomized intervention study to validate GENIE's effectiveness.

(JMIR Form Res 2025;9:e69115) doi:10.2196/69115

KEYWORDS

HBOC; hereditary breast and ovarian cancer; Wizard of Oz study; mobile health; evaluation; hereditary diseases

Introduction

Background

Breast and ovarian cancers rank among the most common types of cancer in low-income countries [1]. It is estimated that 5% - 10% of all patients with breast cancer and up to 25% of all patients with ovarian cancer exhibit a monogenic predisposition to breast and ovarian cancer [2,3]. Pathogenic variants in specific genes substantially elevate the risk for these cancer types and can be inherited in an autosomal dominant manner, thereby posing a risk to offspring. Female carriers of pathogenic variants in BRCA1 or BRCA2—both high-penetrance genes associated with hereditary breast and

XSL•FO
RenderX

ovarian cancer (HBOC)—face a lifetime risk of 50% - 80% for breast cancer, a 60% risk for contralateral breast cancer, and up to 40% for ovarian cancer [1,4,5]. Male carriers of pathogenic BRCA variants have a lower, yet significantly increased, risk of developing breast cancer, estimated between 1% and 7% [6].

Genetic counseling at a recognized human genetics center is a crucial aspect of patient care in Germany, designed to identify individuals at risk for hereditary diseases and provide them with early support. The primary objective of genetic counseling is to empower patients to make informed medical decisions. During the counseling session, patients are assessed for their eligibility for genetic testing of multiple genes associated with HBOC and receive comprehensive, risk-specific information.

In Germany, individuals concerned about a family history of cancer can access genetic counseling. For healthy individuals, genetic counseling is a mandatory prerequisite for genetic testing, as stipulated by the German Genetic Diagnostics Act [7]. Both statutory and private health insurance providers cover the costs of the consultation. Typically, a genetic counseling session lasts approximately 1 hour and is conducted by a medical doctor specialized in human genetics.

The counseling process is highly individualized, taking into account the patient's family history and disease status. For instance, an individual diagnosed with cancer will receive information tailored to treatment options, while a healthy individual with an elevated familial risk may require different guidance. However, the amount of information presented during these sessions can be overwhelming, especially for individuals in high-stress situations, such as immediately following a cancer diagnosis [8]. In a qualitative study, the authors deciphered the needs of patients in the process of genetic counseling and asked how a future mobile assistant could address them [9]. Patients noted that they need (1) support in the time following genetic counseling, but also (2) in the time before genetic counseling by collecting their own and familial medical information, (3) contact options to support services, (4) patient-friendly medical information, and (5) administration-related assistance in a support app. These demands coincide with Germany's Digitalization Strategy for Health and Care, which has one focus on the establishment of people-centric, digitally assisted, cross-sectoral, and cross-professional health care processes [10]. One aim of this strategy is to establish need-based digital assistance and support.

In Germany, there are currently two reliable apps, PINK! Coach and Untire, known as "Digital Health Applications (DiGA)," available to support patients with breast cancer with therapy-related concerns [11]. However, there is currently no app available to support HBOC patients during the genetic counseling process in Germany. Talwar et al [12] have analyzed the characteristics and qualities of genetic mobile apps in English, finding that the majority had reference or resource features with general information about genetics and genetic testing (95.5%), targeted mostly health professional students (86.4%), but did not focus on specific diseases (78.5%). Only 21.6% of the apps were developed by reliable or authoritative agencies. Gasteiger et al [13] have reached a similar conclusion when analyzing patient-oriented genetic and genomic mobile apps in the United Kingdom in 2022. In this study, it became evident that only a few high-quality, genetic patient-oriented apps were available in the United Kingdom. They have emphasized the need for more accessible, culturally sensitive, evidence-based apps to improve genetic literacy within patient populations and specific communities, for example, users for whom English is not their native language.

In this paper, we introduce GENIE, a mobile app designed to support individuals both during and after genetic counseling for HBOC. Additionally, we present the results of a proof-of-principle Wizard of Oz study conducted with 6 patients with HBOC who had completed the acute diagnostic and therapy phase, with their initial diagnosis dating back at least 2 years.

The GENIE Mobile App

The GENIE app is structured in three areas (Compendium, GENIE, and My Profile), which are based on the needs of those receiving human genetic counseling as determined in a prior survey study [9]. We mainly focused on the need for support in the time following genetic counseling by delivering personalized information, and the need for patient-friendly language in communicating medical information. The texts were created taking the Vienna Formula for factual texts into consideration [14]. Average sentence length, word length, and complexity were minimized.

When starting the app, the user is greeted by the personal assistant GENIE. At the first start of the app, the user is introduced to the app's features in the form of a tutorial and asked to provide information in the My Profile area (Figure 1).

XSL•FO
RenderX

Figure 1. Screenshot of My Profile area in the GENIE app. Users can enter data regarding the user's diagnosis and counseling status, physical, and psychological condition. These data are needed for personalization.

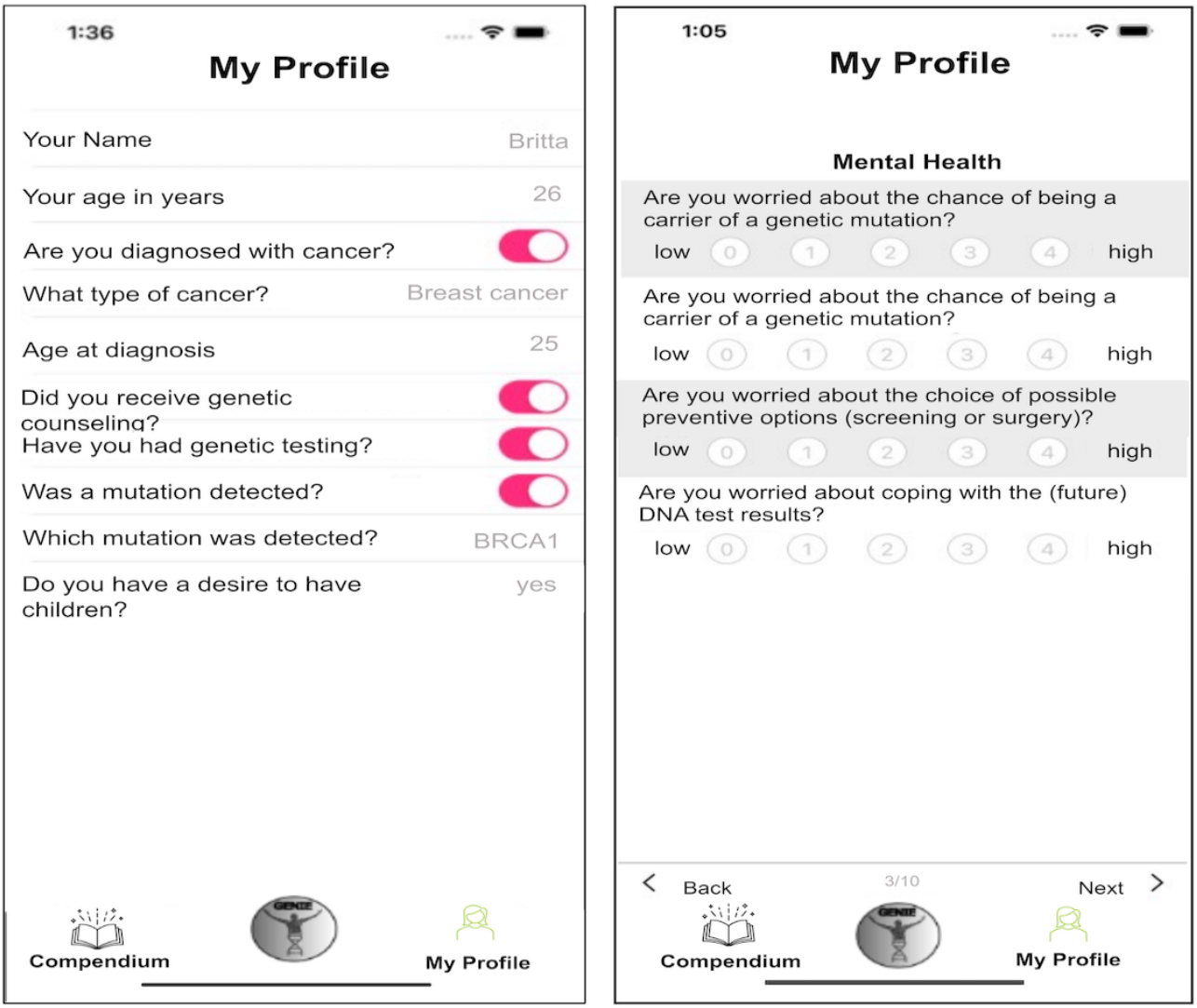

Here, information regarding the user's specific situation is collected, comprising general, counseling-process-related, and disease-status-related information. The collection is based on a mix of validated assessment tools [15,16] and general questions. This information is then used to select topics personalized to the user. Daily, 5 topics are selected and presented in the GENIE area (Figure 2), which is named after our intelligent agent.

XSL•FO
RenderX

Figure 2. Screenshot of the GENIE area in the GENIE app. During the first start of the GENIE mobile app, the user is greeted by the artificial personal assistant. By briefly introducing its aim and different areas, the visualization is kept as simple as possible for the user (left). The personal assistant starts a conversation about information personalized to the situation of the user. The information is prestructured as a chat, and the user can ask predefined questions regarding the topic (right).

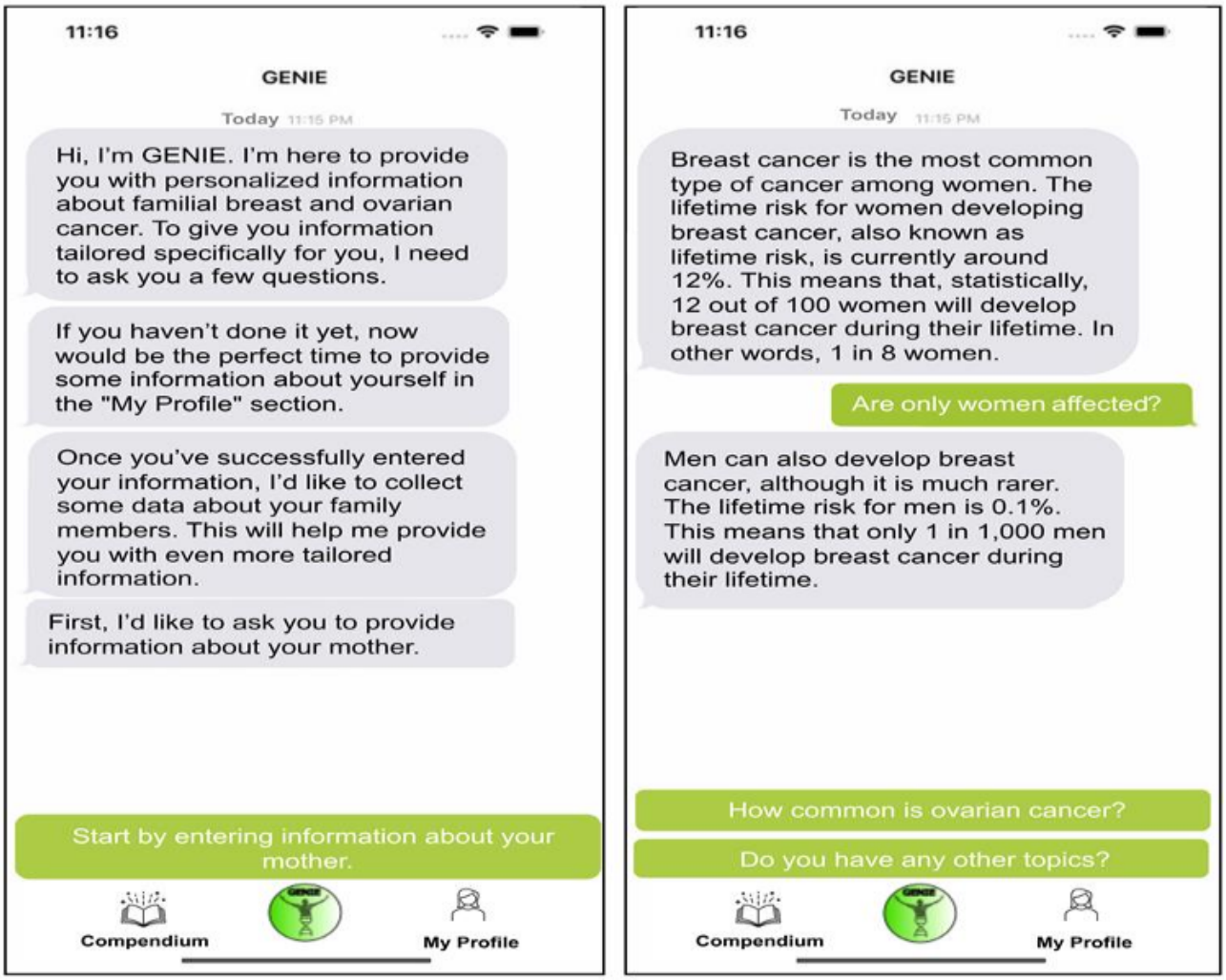

The idea behind the tailored information provision is to minimize excessive information loads. The app's GENIE area is based on a prestructured dialog and adapts the design of widespread short messaging apps in the form of a chat. Chatting with one's personal assistant aims to give a feeling of accompaniment and increase adherence. Instead of typing their own questions or messages, the personal assistant initiates the conversation on the selected topics. The user then selects predefined answers regarding whether the topic is interesting and should be deepened by asking a predefined question or whether to move on to the next topic. For each topic, several consecutive in-depth questions are available. After these, the user is directed by a link in the chat to the Compendium area of the app, where more information on this topic exists. After finishing all 5 daily topics, the assistant says goodbye and states that it will provide new topics tomorrow.

Furthermore, in the Compendium area, the user has unrestricted access to all possible topics the personal assistant can select and provide in the dialog area. This means that besides the personalized topics, additional topics that are potentially not relevant for the specific user can be assessed. Overall, topics are structured by content with communication, family planning, psycho-oncology, and cancer being the main structure (Figure 3).

XSL•FO
RenderX

Figure 3. Screenshot of the Compendium area in the GENIE app. The area is structured along the main topics: communication, family planning, psycho-oncology, and disease (left). The individual topics in the Compendium are structured by the questions asked in the dialog. On the bottom, the user can rate the topic (right).

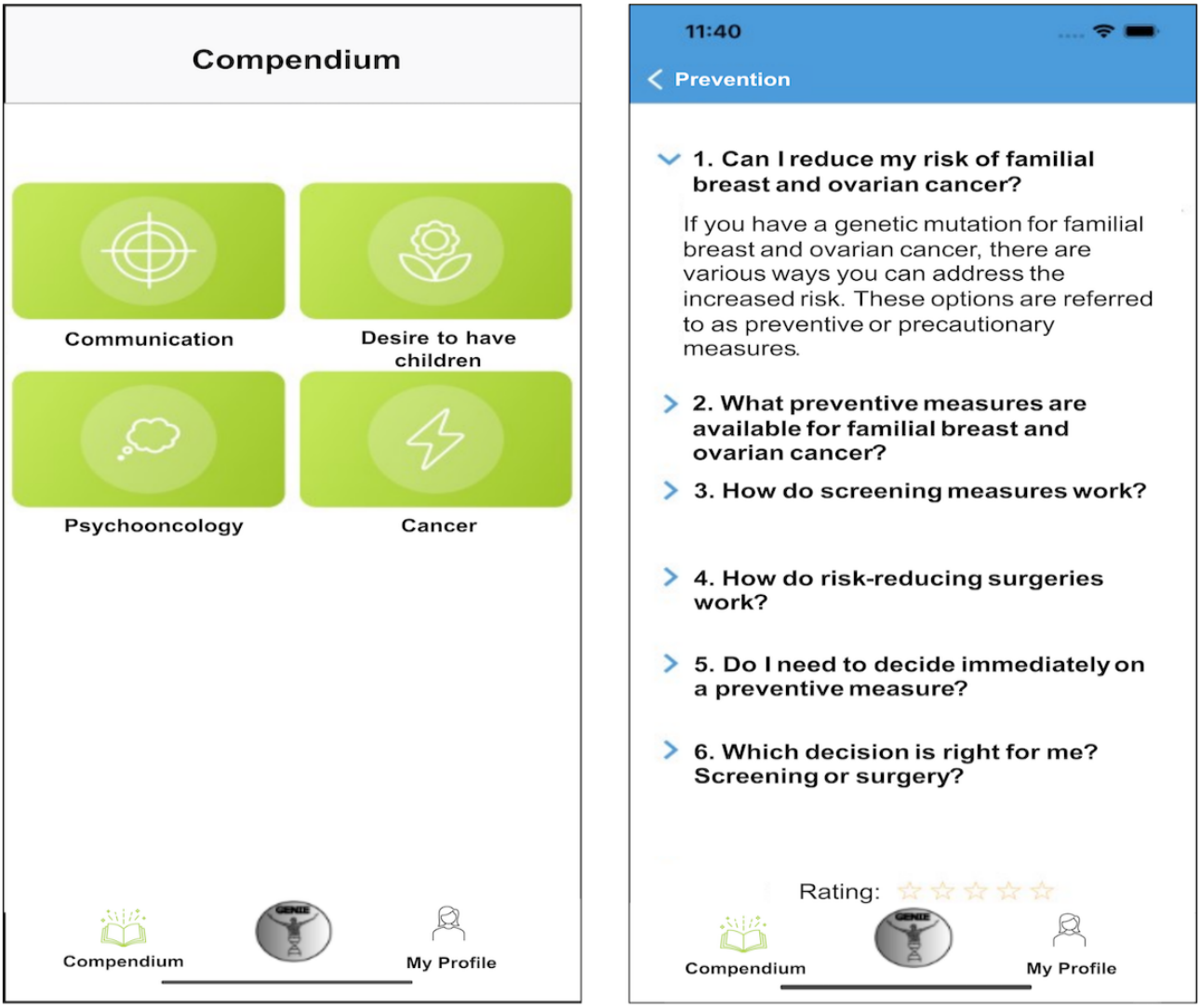

Each individual topic is structured according to questions similar to the dialog. Furthermore, the user can provide a rating for each topic on a 5-star scale.

For the proof-of-principle presented in the study below, we chose a Wizard of Oz study setting, in which 2 human experts created a fictitious character with specific attributions by hand. In the future, a 2-component artificial intelligence (AI) system will select topics from the corpus of the app based on the data entered in the My Profile area. On the one side, explicit expert knowledge will be formalized as an ontology, and on the other side, tacit knowledge will be implemented based on the scoring system proposed by Wolff et al [17].

We plan to further extend the functionality with the possibility to create a pedigree before genetic counseling and an area to upload administration-related components such as medical reports or direct connection to electronic health records.

Methods

Study Design

A partially functional prototype of the GENIE support app was evaluated in the form of a Wizard of Oz study in order to assess user satisfaction. In a Wizard of Oz study, participants believe to interact with an intelligent computer agent, often via a dialog, while the system's intelligence is mocked by a human [18]. The interaction can be direct, more precise, where the participant directly chats with a human, or indirect, which means that the interaction is mostly prestructured. The methodology was successfully applied to telephone information services such as telephone directories, flight or train information, and reservation services. Recently, the Wizard of Oz study design has been used in user experience studies [19-21]. In this study, two nonfunctional parts, the AI-based personalization of topics and the entering of own information in the My Profile area, existed. Instead, a fictitious profile was used for personalization, whose data were deposited in the My Profile area. For simulating GENIE's personalization, two experts in human genetics used this profile description for selecting and ordering topics from

XSL•FO
RenderX

the corpus of the GENIE app, which can be found in Multimedia Appendix 1. The experts' ordering was used to simulate the intelligent behavior of GENIE. The used profile was provided to the participants as text in the German language. An English translation follows:

> *Andrea is 35 years old and was diagnosed with breast cancer some weeks ago. She just finished the genetic counseling session. A genetic test revealed a pathogenic mutation in the BRCA2 gene. Her mother was diagnosed with cancer when she was 40 years old and died shortly after the diagnosis at 41 years due to the disease. Her grandmother on her mother's side also died of breast cancer but was already 75 years old when she was diagnosed. Her maternal grandfather died of a stroke at 83. Her father is 65 and is doing well apart from a slightly high blood pressure. Andrea's paternal grandparents have already passed away and lived to a high age. They were both 92 years old. Andrea herself has a son (5 y) and a daughter (7 y). She has a very good relationship with her 2-year-older brother and his 15-year-old son.*
>
> *Generally, Andrea is a cautious and anxious woman. The cancer diagnosis weighs heavily on her, and she is worried about the future. As a teenager, she felt depressed for several months after the breakup with her first boyfriend. Since then, she has been frightened of being abandoned in her relationships. Due to the diagnosis, Andrea is frightened that her partner Paul might no longer find her attractive, in case of a surgery. She does not know how to communicate the test results to her children. She believes that they are too small to comprehend the situation. Andrea wishes for a third child but is unsure if it is possible with the diagnosis of HBOC. She feels insecure regarding the planning of her family. Therefore, she wishes for more information regarding the preventive options mentioned by the human geneticist during the consultation. She missed that in the conversation due to the huge amount of information explained to her. She would also like to have more information about her breast cancer treatment, as the doctors never really explained it clearly. In view of the overall situation, Andrea would like to receive professional help in the future to cope with her diagnosis psychologically.*

Overall, the iCHECK-DH guideline was used for reporting on the GENIE digital health implementation, where applicable in this publication.

Sample

A total of 6 patients with HBOC evaluated the app after their acute diagnostic and therapy phase. The patient eligibility criteria for the study were: (1) diagnosed HBOC, (2) undertook human genetic counseling, (3) time from first diagnosis at least 2 years (this time gap was required by the ethics committee of Hannover Medical School for the proof of principle study), (4) at least 18 years old, and (5) sufficient German language skills. Recruited participants were given access to the GENIE app for at least a week on a provided iPhone in 2023. Sociodemographic variables were pseudonymized. The patients were selected purposively. All patients presented at the Department of Human Genetics for genetic counseling and testing between 2017 and 2021 and had prior experience in using an iPhone.

The participants' characteristics are shown in Table 1.

In total, 4 participants had a pathogenic variant in BRCA1, and 2 participants had a pathogenic variant in BRCA2. Of the 6 participants, 4 participants had breast cancer. The participants have no background of immigration, few siblings, and few to no children. They are at the lower end of the middle age range (mean 38.7, SD 8.5 years). In total, 2 participants had a lower level of education (general certificate of upper secondary education or below), and 4 participants were highly educated (at least Abitur). All participants were at work during the testing period.

XSL•FO
RenderX

Table . Characteristics of the participants evaluating the GENIE app[a].

ID	Age (years)	Demographics of the center of life	Highest level of education	Immigration background	Number of siblings (f/m)	Number of children (f/m)	Diagnosis	Date of initial diagnosis
1	46	Metropolis	University	No	0/2	0/1	HBOC[b]	June 2020
2	33	Metropolis	University	No	1/0	0/0	HBOC	August 2020
3	40	__[c]	Lower secondary school leaving certificate	No	1/1	0/2	HBOC, breast cancer	February 2020
4	25	Small city	Abitur	No	0/1	0/0	HBOC, breast cancer	May 2021
5	40	Medium-sized city	Secondary school leaving certificate	No	0/1	0/0	HBOC, breast cancer	August 2020
6	48	Small city	Abitur	No	6/1	0/0	HBOC, breast cancer	February 2020

[a]All patients presented at the Department of Human Genetics at Hannover Medical School for genetic counseling and testing between 2017 and 2021 and were recruited via the BRCA network, a nationwide support network for patients with breast cancer in Germany. All patients had the diagnosis of HBOC.

[b]HBOC: hereditary breast and ovarian cancer.

[c]Not applicable.

Evaluation

The evaluation was divided into 2 parts.

Quantitative Study to Evaluate the GENIE App

A 45-item questionnaire in the German language was provided to the participants after the testing of the GENIE mobile app. The questionnaire is structured into the categories: first impression, contents, app layout, usability, use, and willingness to pay. The categories, first impression, contents, and usability consist of an excerpt from Thielsch et al, 2017 [22] and the System Usability Scale by Brooke [23], respectively. The categories, first impression, contents, and app layout consist of a 7-item Likert scale, while the System Usability Scale uses a 5-item Likert scale. The use and willingness to pay are each a binary yes-no question and a free text field for the provision of reasons. The data were analyzed descriptively and visually using a bar plot in Python (Python Software Foundation). Therefore, ratings of negatively formulated questions, such as “the colors do not match,” were transformed to the positive scale by subtracting the negative formulated rating from the scale’s maximum and adding 1 (pos=maxRating–negative+1). The questionnaires, along with the participants’ consent forms, are stored in a fireproof cabinet at the Institute of Human Genetics and will be kept for 10 years.

Qualitative Study to Evaluate the GENIE App

The qualitative study was conducted in the summer of 2023 in Hannover, Germany, to explore the satisfaction of the participants with the GENIE app. All participants were interviewed 2 to 3 weeks after testing the GENIE app. The interviews were performed by NA. NA is a fifth-year medical student at the University Medical Center Freiburg and tested the GENIE app within his PhD study. Due to valuable experience in qualitative analyses in the field of human genetics, BV (research associate at the Department of Human Genetics) provided close supervision of NA. No patient knew the interviewer before. Interviews were conducted as a video call (face-to-face) at home or in the office of the Department of Human Genetics and took place between the researcher and the patient only. The interviewer introduced himself as a medical doctoral student with a particular interest in supporting patients in the process of genetic counseling through mobile support apps. A semistructured interview guide was used, which roughly defined the structure and topics of the interview with open questions, but still allowed sufficient flexibility for new topics.

1. How did you find the layout and structure of the app?
2. How did you find the different sections of the app?
3. How did you find the design of the app?
4. How did you find the information density, clarity, and text length?
5. How did you find the suitability of the texts for Andrea?
6. Did you find the content trustworthy?
7. Did you feel like you were building a relationship with GENIE?
8. Would you recommend the app?

The interviews were audio recorded.

Data Analysis

The audio files of the interviews were transcribed and digitized. Transcription was done according to predefined rules following Kuckartz and Rädiker [24] and Dresing et al [25]. The interviews were evaluated according to structuring content analysis, a subform of qualitative content analysis according to Kuckartz et al [24]. For the evaluation, the software MAXQDA (version 2022; VERBI software) was used. All interviews were included in the analysis.

XSL•FO
RenderX

The following quality criteria, according to Mayring [26], were met during the evaluation:

1. Intercoder agreement—to ensure objectivity, a second person (BV) was involved in the coding process. In case of ambiguity, the researchers discussed the classification of the data into the appropriate categories. The defined category system ensured an intersubjectively comprehensible analysis or category assignment.
2. Intracoder agreement was assured by NA through repeated coding at a later point in time.
3. Procedural documentation—the process of data collection and data analysis was documented in a way that was transparent and comprehensible to others.
4. Rule guidance—the process of data collection and data analysis was systematic and rule-governed (interview protocols, transcription rules, coding rules)
5. Proximity to the object—the data collection took place in the everyday life of the interviewees and was based on the problems of the interviewees.

Ethical Considerations

The study was performed in accordance with the ethical standards of the responsible committee on human experimentation (institutional and national) and with the Helsinki Declaration of 1975, as revised in 2000. The study was approved by the Ethics Committee of Hannover Medical School (10573_BO_K_2022). Informed written consent was obtained from all patients to be included in the study. Participants were informed that participation was voluntary, and that they could withdraw at any time. Furthermore, they were informed about the right to access, rectify, restrict, and delete their data during the study. Patients were informed that nonparticipation would not result in any disadvantages for them. The participants gave their consent for the interview to be recorded and transcribed. The audio files and the data from the questionnaire were stored on a server at the Hannover Medical School (MHH) and are safeguarded according to the principles of Good Scientific Practice of the German Research Foundation. All data will be stored for 10 years on the servers or in locked cabinets at the Institute of Human Genetics at MHH, after which it will be deleted. The interview transcripts and questionnaire data were pseudonymized. In case of withdrawal, there is a list of participants that has been stored separately to allow the identification of individual data. Only project staff from MHH have access to the data. MHH operates in accordance with the provisions of the General Data Protection Regulation, the Federal Data Protection Act of May 25, 2018, and the Lower Saxony Data Protection and Freedom of Information Act of May 16, 2018. No data were shared with third parties. Participants were compensated with US $112 for their efforts. Participants were recruited via email via the BRCA Network, a nationwide support network for patients with breast cancer in Germany. All patients presented at the Department of Human Genetics for genetic counseling and testing between 2017 and 2021.

Results

In the following, we will present the quantitative results from the testing, followed by the qualitative data gathered from the interviews.

Results of the Quantitative Analysis

All participants finished the questionnaire. The questions are provided in Multimedia Appendix 1. Overall, the participants' first impression was that the app is usable (1B), which is further supported by a high mean System Usability Score of 75.33 (SD 4.13). On the other hand, users were discordant regarding the app's appeal (1C). The app's content was interesting on the first impression (1A), as well as after using the app (2A-H). Some participants even wanted to save or print out some of the topics provided (2I). The layout of the app achieved a mean rating of 4.48 (SD 1.36) points on the 7-item Likert scale. On a single-item level, the participants found that the layout is only moderately professional (3O), a bit crowded (3A), and slightly uninspiring (3G), but also pleasantly organized (3C). A complete overview of the results can be found in Figure 4.

Further, 5 of 6 participants answered that they were using the app daily during the evaluation. Regarding the willingness to pay for the mobile app, 3 participants would pay amounts between US $5 and US $45 as a single purchase, while the other 3 expressed that the app should be free for users and financed by health insurance or pharmacy companies.

XSL•FO
RenderX

Figure 4. Participants' ratings by categories and questions (y-axis) with rating scales according to the methods section (x-axis). A total of 6 participants tested the GENIE app for 1 week and rated the first impression by answering 3, the contents by answering 9, and the layout of the GENIE app by answering 18 questions on a 7-item Likert scale, with 1 being the lower (negative) and 7 being the upper (positive) end of the scale. First impression: 4.29 (SD 2.09); contents: 5.01 (SD 1.61); layout: 4.48 (SD 1.36). A glossary of the abbreviations in the y-axis to the corresponding questionnaire items can be found in Multimedia Appendix 1.

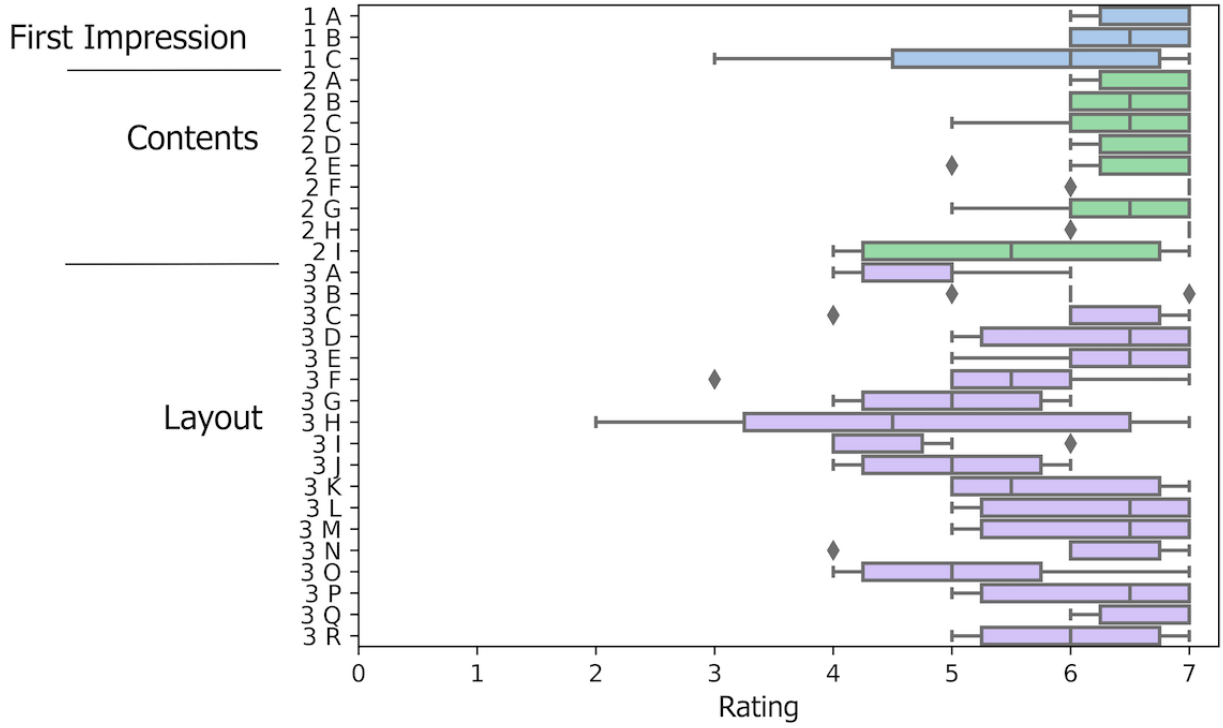

Results of the Qualitative Analysis

The structured interviews primarily addressed the following five points: (1) clarity and text length, (2) credibility, (3) building a relationship level through dialog, (4) personalization of the dialog, and (5) recommendation of the GENIE app.

Clarity and Text Length

All participants rated the clarity of the texts and the text length as very good. It was noted that it is very important for the participants that the texts are also understandable for medical laypersons, which is not always the case in doctor consultations and on the internet.

> *Very easy to understand. Very, very simple to read. Very understandable. This way, you can process multiple topics at once or combine them into one because, when talking to doctors or reading elsewhere, it's often described in very technical terms, which I find to be very specialized. But in the app, I found it very, very pleasant.* [Participant 4]

Credibility

The GENIE app is characterized by evidence-based texts, and source references are stated at the end of the texts. Therefore, we were interested in whether users considered the source references important. Some users immediately noticed the source references and rated them positively.

> *I actually find it very useful and very good, especially when you already have a few points in an area where you have read up, and then realize: Okay, I want to delve deeper into this now. Instead of sitting down and googling everything like crazy, I immediately know: Okay, this information comes from here and there. I can continue to educate myself further and explore it in more depth. It's definitely useful.* [Participant 5]

Building a Relationship Level Through Dialog

Users can access personalized information through the GENIE app in two ways: via a chatbot in the dialog area and via a well-structured knowledge base, such as an encyclopedia. To assess user perceptions of the text delivery strategy, we investigated how the participants experienced these methods. In total, 5 out of 6 participants preferred the dialog area and used the knowledge base only minimally. One participant noted limited use of the dialog feature and reported that she primarily engaged with the texts in the knowledge base, commending their clarity and organization. Additionally, some participants who favored the dialog area developed a sense of rapport with the app, feeling as if they were conversing with a health care professional or another individual with a similar condition. This manifested as an unplanned yet beneficial outcome.

> *Yes, sometimes I had the feeling as if the person was sitting next to me, explaining things to me or reading something aloud, or asking me questions, and I would*

XSL•FO
RenderX

> *answer, or I would ask and they would respond. It felt like, well, like a doctor or a medical student who is also engaging with the topic themselves.* [Participant 2]

> *It's very personal, and that makes it very pleasant, especially in a health context, it's good to have someone by your side. I know this from personal experience as well. It's nice to have someone who gives you the feeling that they've been through it themselves and can help you with a lot of knowledge. And that's exactly how it is (...) with the app. This personal touch is a great idea.* [Participant 4]

Personalization of the Dialogs

One of the key features of the GENIE app is its high level of personalization. The participants were asked to assess how well the content of the texts matched Andrea's situation. All participants stated that the suggested texts fit Andrea's described situation very well.

> *Yes, yes. It was perfect, exactly one-to-one. Nothing was missing. I would say it fit exactly with what Andrea had mentioned about her situation, and the information that came to Andrea was wonderful.* [Participant 2]

Recommendation

At the end of the interview, the participants were asked if they would recommend the app. All participants said they would recommend the app to others affected. In total, 2 participants mentioned that they would have liked to use this app themselves when they received the diagnosis of hereditary breast cancer.

> *I found the app very good, and it was explained superbly. I would recommend it to anyone who has issues with this genetic mutation or human genetics. I would say: Try it out, maybe you'll understand it better, and it's explained in a very simple way.* [Participant 2]

Discussion

Principal Findings

In this study, we were able to demonstrate the benefits of the autonomous mobile support GENIE. The app has the potential to support women during the phase of genetic counseling in the contemplated areas. The qualitative data from the interviews and the data from the questionnaires show the app's high usability and high quality of its content. Still, the design needs to be revised. Overall, the time and location-independent use of the app, while personalizing the topics, satisfied the participants. The simulation of a personal assistant in the genetic counseling process was successful. The only functionality not used was the star rating component. Here, an additional message by the assistant hinting at it could be helpful. In the future, the star rating could be used to further refine the personalization of topics.

Patients diagnosed with HBOC have diverse needs. This group includes women who have already been diagnosed with breast and/or ovarian cancer, as well as healthy women whose cancer risk is merely elevated. Metcalfe et al [27] observed that women with a prior cancer diagnosis indicated a need for more information related to cancer treatment compared to women without a cancer history. Another study demonstrated that women with pathogenic variants in moderate-risk genes for HBOC have distinct informational needs compared to those with variants in high-risk genes [28]. Consequently, Henneman et al [29] suggested that the genetic counseling process should be tailored to meet individual demands and consider the expectations and pre-existing lay knowledge of the counselee. In this evaluation study, we were able to show that most participants preferred the tailored information delivered by the chatbot over the Compendium.

One of the key features is the credibility and clarity of the provided information. The patients reported that the provision of references in the Compendium enhanced the credibility. The high level of credibility and clarity of information provided via the GENIE app is achieved by expert manual curation. However, this process is time-consuming and holds the risk of information not being up to date within the revision's time delay. In the future, we will explore the use of large language models (LLMs) to (partially) automate this process. The provision of information in the form of a dialog resulted in patients perceiving the interaction as if they were speaking with a certified specialist in human genetics. Credibility and perceived competence are well-known factors influencing patient adherence [30,31]. Furthermore, the literature already highlights a mismatch between the information provided during counseling sessions and patients' expectations regarding the desired content, usefulness, and comprehensibility [32-34]. Factors such as the health literacy level of counselees and the appropriateness of language used for lower levels have been identified to have a significant influence on genetic counseling. Additionally, it is widely recognized that verbal medical information provided by health care practitioners, particularly genetic information, is often not accurately recalled. Research on information recall indicates that up to 80% of medical information given by health care practitioners is forgotten, and about half of the information remembered is incorrect [35]. This impacts not only the patient but also family members who receive incorrect or insufficient information. Furthermore, according to Edwards et al [36], emotional and supportive communication is beneficial for the counselors and prevails as the sole benefit of having more information. Although the personal assistant was perceived as a (human-like) specialist in human genetics by some participants, a study conducted prior to the development of the GENIE app emphasized the importance of human interaction with patients. In that study, patients reported that the interpersonal interaction with the genetic specialists was experienced as particularly positive [9]. The importance of human interaction with patients, in general, was also emphasized. Nonetheless, the use of an autonomous chatbot before and after the counseling session can increase the time available for emotional support during the sessions.

A potential risk of autonomous systems such as GENIE or LLMs like ChatGPT in health-related contexts is the potential increase in so-called cyberchondria, which is characterized by excessive web-based health research and heightened, yet unnecessary,

XSL•FO
RenderX

anxiety [37,38]. Multiple aspects of these systems could be associated with cyberchondria: (1) information accessibility, (2) information quality, and (3) credibility. Both GENIE and LLMs serve as low-threshold gateways to information, characterized by easy access and time-independent availability, making repeated and excessive consumption effortless. This, in turn, may potentially foster negative behavior patterns in individuals with cyberchondria.

Nonetheless, GENIE offers several advantages over LLMs. First, the quality of the information is controlled by human experts who manually create and curate the texts provided by the system. Thus, hallucinations, and fictitious or inaccurate information, which can occur in LLMs [39,40], are very unlikely to occur in our agent. In addition, this directly impacts the system's credibility. For GENIE, scientific information sources are provided, which were used to curate the texts. For LLMs, even in the case of using scientific literature as a corpus, it cannot be ruled out at the moment that fictitious publications would be presented by the system as proof of its statements. The same is true for the appropriateness of information for the users. Moreover, in Germany, the Genetic Diagnostics Act prohibits giving direct advice in the genetic counseling session. To the authors' knowledge, there is currently no safe possibility to exclude such behavior in LLMs. On the other hand, the manual curation of texts by experts, including web search, selection of scientific publications, and revision of information in the form of prestructured dialogs, is very laborious and may limit the width of topics that can be recommended by the system.

Limitations

Still, the results of this study are limited by multiple factors. First, the study design as a Wizard of Oz experiment holds the risk that the later system's personalization engine cannot live up to the high personalization standards provided by the experts. Simulating GENIE's AI component and relying on an expert's selection for personalization of the topics holds the risk that the implementation of such a recommender system lacks proper individualization performance. Especially, the formalization of expert knowledge holds the risk of being lossy. Still, in the case of a working personalization system and its use in the real world, the system would be robust regarding changes in the user's profile, which may evolve over time. By updating the profile information in the My Profile area of the app, the system would use the updated information to adapt the selection of topics. Second, the results presented are based on a single fictitious case study, making it possible that the system's topic recommendation fails for other cases. Other limitations are found in the limited study cohort. The number of 6 participants seems rather low. Still, for usability testing, 5 participants are typically able to find the majority of errors [41]. On the other hand, the participants were not in the acute phase of genetic counseling, making it difficult to analyze if the app would stress or even harm participants in this phase. Here, some participants expressed that they wished to have had the app during their acute phase. Although all participants had prior experience using an iPhone, wide market coverage is crucial. We are therefore currently in the process of porting the app to Android while reworking the user interface. Furthermore, all participants are female, and thus, gender aspects are not analyzable. However, the main target audience of the GENIE app is women since they are more prone to HBOC than men. To overcome these limitations, we plan to conduct a two-arm, randomized intervention study with 240 participants at the Institute of Human Genetics at Hannover Medical School, in cooperation with the Center for Familial Breast and Ovarian Cancer and the Department of Biometry at Hannover Medical School. Here, a fully functional version of the GENIE app will be used to provide scientific evidence of the effectiveness of the GENIE app in delivering a positive health care impact for patients with HBOC.

Conclusions

Widespread use of GENIE has the potential to fill existing gaps in the current landscape of mobile solutions for hereditary diseases. Patients reported in interviews that they felt abandoned after genetic counseling. The study presented here demonstrates that GENIE can provide support to patients in this regard. In the next step, we will implement the topic recommendation system simulated in this study by human expertise via the mentioned approaches, precisely an ontology and the scoring system described in Wolff et al [17], by acquiring and formalizing the expert's knowledge for topic personalization. Moreover, we are currently revising the mobile app regarding user interfaces and user experience to improve its usability, together with an expert industry partner. Additionally, we will extend the system's functionality in order to cover further requirements and wishes of patients reported in Ammon et al [9], such as support in creating one's pedigree before genetic counseling and an area to upload administration-related components such as medical reports or direct connection to electronic health records. Furthermore, the intervention study will help to confirm GENIE's positive health care impact on patients with HBOC.

Acknowledgments

The authors would like to acknowledge the patients and their families. Furthermore, the authors express their gratitude for the valuable advice provided by the BRCA-Network e.V. self-help group and their assistance with recruitment. This study was conducted as part of a PhD project and was funded by Else Kröner-Fresenius-Stiftung (Else Kröner Graduate school DigiStrucMed 2020_EKPK.20). Artificial intelligence tools were used to check spelling and grammar.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author upon reasonable request.

XSL•FO
RenderX

Authors' Contributions

BV, TK, and DW designed the study and confirmed that they had full access to all the data in the study and took responsibility for the integrity of the data and the accuracy of the data analysis. NA, BV, DW, TK, and CR were involved in data acquisition, analysis, and interpretation. All authors (BV, DW, NA, TK, SO-J, and CR) were involved in drafting, reviewing, and revising the manuscript and have approved the final version. All authors agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Sorted list of topics relevant to Andrea.
[DOCX File, 15 KB - formative_v9i1e69115_app1.docx]

Checklist 1
Checklist of iCHECK-DH: Guidelines and Checklist for the Reporting on Digital Health Implementations.
[PDF File, 267 KB - formative_v9i1e69115_app2.pdf]

References

1. Nielsen FC, van Overeem Hansen T, Sørensen CS. Hereditary breast and ovarian cancer: new genes in confined pathways. Nat Rev Cancer 2016 Sep;16(9):599-612. [doi: 10.1038/nrc.2016.72] [Medline: 27515922]
2. Campeau PM, Foulkes WD, Tischkowitz MD. Hereditary breast cancer: new genetic developments, new therapeutic avenues. Hum Genet 2008 Aug;124(1):31-42. [doi: 10.1007/s00439-008-0529-1] [Medline: 18575892]
3. Walsh T, Casadei S, Lee MK, et al. Mutations in 12 genes for inherited ovarian, fallopian tube, and peritoneal carcinoma identified by massively parallel sequencing. Proc Natl Acad Sci U S A 2011 Nov 1;108(44):18032-18037. [doi: 10.1073/pnas.1115052108] [Medline: 22006311]
4. Goldberg JI, Borgen PI. Breast cancer susceptibility testing: past, present and future. Expert Rev Anticancer Ther 2006 Aug;6(8):1205-1214. [doi: 10.1586/14737140.6.8.1205] [Medline: 16925486]
5. Antoniou A, Pharoah PDP, Narod S, et al. Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case Series unselected for family history: a combined analysis of 22 studies. Am J Hum Genet 2003 May;72(5):1117-1130. [doi: 10.1086/375033] [Medline: 12677558]
6. Tai YC, Domchek S, Parmigiani G, Chen S. Breast cancer risk among male BRCA1 and BRCA2 mutation carriers. J Natl Cancer Inst 2007 Dec 5;99(23):1811-1814. [doi: 10.1093/jnci/djm203] [Medline: 18042939]
7. Human Genetic Examination Act [Web page in German]. Federal Ministry of Justice and Consumer Protection. URL: https://www.gesetze-im-internet.de/gendg/ [accessed 2024-12-15]
8. Vajen B, Rosset M, Wallaschek H, Baumann E, Schlegelberger B. Psychological distress and coping ability of women at high risk of hereditary breast and ovarian cancer before undergoing genetic counseling—an exploratory study from Germany. Int J Environ Res Public Health 2021 Apr 19;18(8):4338. [doi: 10.3390/ijerph18084338] [Medline: 33921890]
9. Ammon N, Reichert C, Kupka T, et al. Deciphering the needs of patients with hereditary breast and ovarian cancer in the process of genetic counseling to inform the development of a mobile support app: a qualitative study in Germany. J Community Genet 2024 Dec;15(6):603-613. [doi: 10.1007/s12687-024-00727-6] [Medline: 39158769]
10. Digital together: Germany's digitalisation strategy for health and care. : Federal Ministry of Health URL: https://www.bundesgesundheitsministerium.de/fileadmin/Dateien/3_Downloads/D/Digitalisierungsstrategie/Germany_s_Digitalisation_Strategy_for_Health_and_Care.pdf [accessed 2024-12-15]
11. DiGA directory [Web page in German]. Federal Institute for Drugs and Medical Devices. 2024. URL: https://diga.bfarm.de/de/verzeichnis?category=%5B%2274%22%5D [accessed 2025-05-25]
12. Talwar D, Yeh YL, Chen WJ, Chen LS. Characteristics and quality of genetics and genomics mobile apps: a systematic review. Eur J Hum Genet 2019 Jun;27(6):833-840. [doi: 10.1038/s41431-019-0360-2] [Medline: 30809045]
13. Gasteiger N, Vercell A, Davies A, Dowding D, Khan N, Davies A. Patient-facing genetic and genomic mobile apps in the UK: a systematic review of content, functionality, and quality. J Community Genet 2022 Apr;13(2):171-182. [doi: 10.1007/s12687-022-00579-y] [Medline: 35182377]
14. Bamberger R, Vanecek E. Reading–Understanding–Learning–Writing: The Levels of Difficulty of Texts in the German Language [Book in German]: Wien, Jugend & Volk Verlagsgesellschaft; 1984.
15. Tercyak KP, Johnson SB, Roberts SF, Cruz AC. Psychological response to prenatal genetic counseling and amniocentesis. Patient Educ Couns 2001 Apr;43(1):73-84. [doi: 10.1016/s0738-3991(00)00146-4] [Medline: 11311841]
16. Eijzenga W, Bleiker EMA, Hahn DEE, et al. Psychosocial aspects of hereditary cancer (PAHC) questionnaire: development and testing of a screening questionnaire for use in clinical cancer genetics. Psychooncology 2014 Aug;23(8):862-869. [doi: 10.1002/pon.3485] [Medline: 24443031]

XSL•FO
RenderX

17. Wolff D, Behrends M, Gerlach M, Kupka T, Marschollek M. Personalized knowledge transfer for caregiving relatives. Stud Health Technol Inform 2018;247:780-784. [Medline: 29678067]
18. Dahlbäck N, Jönsson A, Ahrenberg L. Wizard of Oz studies—why and how. Knowl Based Syst 1993 Dec;6(4):258-266. [doi: 10.1016/0950-7051(93)90017-N]
19. Weiss A, Bernhaupt R, Schwaiger D, Altmaninger M, Buchner R, Tscheligi M. User experience evaluation with a Wizard of Oz approach: technical and methodological considerations. Presented at: 2009 9th IEEE-RAS International Conference on Humanoid Robots (Humanoids 2009); Dec 7-10, 2009; Paris, France. [doi: 10.1109/ICHR.2009.5379559]
20. Ghosh S, Shah C. Spoken conversational search: evaluating the effect of system clarifications on user experience through Wizard‐of‐Oz study. J Assoc Inf Sci Tech 2025 May;76(5):819-839. [doi: 10.1002/asi.24974]
21. Salber D, Coutaz J. Applying the Wizard of Oz technique to the study of multimodal systems. : Springer Presented at: Human-Computer Interaction: Third International Conference, EWHCI'93; Aug 3-7, 1993; Moscow, Russia p. 219-230. [doi: 10.1007/3-540-57433-6_51]
22. Thielsch MT, Niesenhaus J. User experience, gamification, and performance. In: The Wiley Blackwell Handbook of the Psychology of the Internet at Work: Wiley; 2017:79-101.
23. Brooke J. SUS: a quick and dirty usability scale. In: Usability Evaluation in Industry: Taylor and Francis; 1996:194-189.
24. Kuckartz U, Rädiker S. Qualitative Content Analysis: Methods, Practice, and Computer Support [Book in German], 5th edition: Beltz Juventa; 2022.
25. Dresing T, Pehl T. Practical Guide to Interviews, Transcription & Analysis: Instructions and Rules for Qualitative Researchers [Book in German], 8th edition: Eigenverlag; 2018.
26. Mayring P. Qualitative Content Analysis: Principles and Techniques [Book in German], 13th edition: Julius Beltz GmbH & Co. KG; 2022. [doi: 10.1007/978-3-658-37985-8_43]
27. Metcalfe KA, Liede A, Hoodfar E, Scott A, Foulkes WD, Narod SA. An evaluation of needs of female BRCA1 and BRCA2 carriers undergoing genetic counselling. J Med Genet 2000 Nov;37(11):866-874. [doi: 10.1136/jmg.37.11.866] [Medline: 11073541]
28. Stracke C, Lemmen C, Rhiem K, Schmutzler R, Kautz-Freimuth S, Stock S. Medical knowledge and information needs among women with pathogenic variants in moderate-risk genes for hereditary breast cancer attending genetic counseling at an academic hospital in Germany—a qualitative approach. J Genet Couns 2022 Jun;31(3):698-712. [doi: 10.1002/jgc4.1536] [Medline: 34837291]
29. Henneman L, Timmermans DRM, van der Wal G. Public experiences, knowledge and expectations about medical genetics and the use of genetic information. Community Genet 2004;7(1):33-43. [doi: 10.1159/000080302] [Medline: 15475669]
30. Dai Z, MacDorman KF. The doctor's digital double: how warmth, competence, and animation promote adherence intention. PeerJ Comput Sci 2018;4:e168. [doi: 10.7717/peerj-cs.168] [Medline: 33816821]
31. Polinski JM, Kesselheim AS, Frolkis JP, Wescott P, Allen-Coleman C, Fischer MA. A matter of trust: patient barriers to primary medication adherence. Health Educ Res 2014 Oct;29(5):755-763. [doi: 10.1093/her/cyu023] [Medline: 24838119]
32. Joseph G, Lee R, Pasick RJ, Guerra C, Schillinger D, Rubin S. Effective communication in the era of precision medicine: a pilot intervention with low health literacy patients to improve genetic counseling communication. Eur J Med Genet 2019 May;62(5):357-367. [doi: 10.1016/j.ejmg.2018.12.004] [Medline: 30553023]
33. Kamara D, Weil J, Youngblom J, Guerra C, Joseph G. Cancer counseling of low-income limited English proficient Latina women using medical interpreters: implications for shared decision-making. J Genet Couns 2018 Feb;27(1):155-168. [doi: 10.1007/s10897-017-0132-5] [Medline: 28791541]
34. Cheng JKY, Guerra C, Pasick RJ, Schillinger D, Luce J, Joseph G. Cancer genetic counseling communication with low-income Chinese immigrants. J Community Genet 2018 Jul;9(3):263-276. [doi: 10.1007/s12687-017-0350-4] [Medline: 29197036]
35. Kessels RPC. Patients' memory for medical information. J R Soc Med 2003 May;96(5):219-222. [doi: 10.1177/014107680309600504] [Medline: 12724430]
36. Edwards A, Gray J, Clarke A, et al. Interventions to improve risk communication in clinical genetics: systematic review. Patient Educ Couns 2008 Apr;71(1):4-25. [doi: 10.1016/j.pec.2007.11.026] [Medline: 18207694]
37. Starcevic V, Berle D, Arnáez S. Recent insights into cyberchondria. Curr Psychiatry Rep 2020 Aug 27;22(11):56. [doi: 10.1007/s11920-020-01179-8] [Medline: 32852626]
38. Starcevic V. Cyberchondria: challenges of problematic online searches for health-related information. Psychother Psychosom 2017;86(3):129-133. [doi: 10.1159/000465525] [Medline: 28490037]
39. Perković G, Drobnjak A, Botički I. Hallucinations in LLMs: understanding and addressing challenges. Presented at: 2024 47th MIPRO ICT and Electronics Convention (MIPRO); May 20-24, 2024; Opatija, Croatia. [doi: 10.1109/MIPRO60963.2024.10569238]
40. Hu R, Zhong J, Ding M, Ma Z, Chen M. Evaluation of hallucination and robustness for large language models. Presented at: 2023 IEEE 23rd International Conference on Software Quality, Reliability, and Security Companion (QRS-C); Oct 22-26, 2023; Chiang Mai, Thailand. [doi: 10.1109/QRS-C60940.2023.00089]

XSL•FO
RenderX

41. Nielsen J, Landauer TK. A mathematical model of the finding of usability problems. 1993 Presented at: CHI '93: Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems; Apr 24-29, 1993; Amsterdam, The Netherlands. [doi: 10.1145/169059.169166]

Abbreviations

AI: artificial intelligence
HBOC: hereditary breast and ovarian cancer
LLM: large language model
MHH: Hannover Medical School

Edited by A Mavragani; submitted 22.11.24; peer-reviewed by F Ciorba, S Cummings; revised version received 28.03.25; accepted 07.04.25; published 05.06.25.

Please cite as:
Wolff D, Kupka T, Reichert C, Ammon N, Oeltze-Jafra S, Vajen B
Personalized Support in Hereditary Breast and Ovarian Cancer After Genetic Counseling by the Chatbot-Based GENIE Mobile App: Proof-of-Concept Wizard of Oz Study
JMIR Form Res 2025;9:e69115
URL: https://formative.jmir.org/2025/1/e69115
doi:10.2196/69115

© Dominik Wolff, Thomas Kupka, Chiara Reichert, Nils Ammon, Steffen Oeltze-Jafra, Beate Vajen. Originally published in JMIR Formative Research (https://formative.jmir.org), 5.6.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Usage of a Multipurpose mHealth App Among Adults With Sickle Cell Disease: Randomized Controlled Trial

Robert M Cronin[1], MS, MD; Nives Quaye[1], MPH; Xin Liu[2], PhD; Kristina Landes[1], MPH; Lori E Crosby[3,4], PsyD; Adetola A Kassim[5], MD; Michael R DeBaun[6], MPH, MD; Patrick M Schnell[2], PhD

[1]Department of Internal Medicine, The Ohio State University Wexner Medical Center, 410 W 10th Ave, Columbus, OH, United States

[2]Division of Biostatistics, The Ohio State University, Columbus, OH, United States

[3]Department of Pediatrics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, United States

[4]College of Medicine, University of Cincinnati, Cincinnati, OH, United States

[5]Department of Internal Medicine, Vanderbilt University, Nashville, TN, United States

[6]Department of Pediatrics, Vanderbilt University Medical Center, Nashville, TN, United States

Corresponding Author:
Robert M Cronin, MS, MD
Department of Internal Medicine, The Ohio State University Wexner Medical Center, 410 W 10th Ave, Columbus, OH, United States

Abstract

Background: While mobile health (mHealth) apps have been made for various diseases, including sickle cell disease (SCD), most focus on a single purpose. SCD is a chronic disease that requires knowledge of the disease, self-management, and adherence to treatment plans. While mHealth apps have been made with single features for SCD, there is limited understanding of using an mHealth app with a more comprehensive set of features that could engage adults with SCD, depending on what features they prefer and need to engage and empower them in living with their disease.

Objective: We evaluated the usage of an mHealth app with various features, including pain tracking, quizzes for patient-facing guidelines, pain and asthma action plans, and goal setting.

Methods: Adults with SCD were enrolled at 2 sickle cell centers between 2018 and 2022 as part of a 6-month feasibility randomized controlled trial with participants completing surveys at baseline and 6 months. Participants were randomized into receiving either an mHealth app and booklet with patient-facing guidelines or a booklet with the guidelines alone. The mHealth app comprised web pages with patient-facing guideline material and a Research Electronic Data Capture (REDCap) project. The REDCap project included a personal profile, a pain tracker, goal setting, quizzes about the guidelines, and pain or asthma action plans. The REDCap project also included the ability to send daily text messages at a time they chose, which contained a message they could create and a link to their profile. Outcomes included SCD-specific knowledge and acute health care utilization (emergency room visits and hospitalizations). We evaluated the usage of these different features and relationships with baseline variables, each other, and study outcomes.

Results: Approximately 75% (50/67) of the enrolled and randomized participants completed all the study components, and 100% (26/26) of the participants who were randomized to the mHealth app arm and completed the study used the mHealth app. Further, 15/30 (50%) participants used multiple features. Baseline sickle cell knowledge and female gender were associated with more usage of pain diary (P=.04) and mission (P=.046) features, respectively. While not significant, mission completion was associated with lower hospitalizations (P=.06).

Conclusions: Adults with SCD engaged differently with an mHealth app with multiple features. As this study was not focused on one part of our app, engagement with features in this app was entirely patient-driven, which may demonstrate the expected real-world use of an mHealth app in this population. A multipurpose app can help engage participants in self-management strategies through different features and potentially improve outcomes.

Trial Registration: ClinicalTrials.gov NCT03629678; https://tinyurl.com/3wx5en5r

(JMIR Form Res 2025;9:e67906) doi:10.2196/67906

KEYWORDS

sickle cell disease; mHealth; patient engagement; hospitalizations; consumer health informatics; guidelines

XSL•FO
RenderX

Introduction

Sickle cell disease (SCD) is an inherited hemoglobin disorder that affects approximately 100,000 Americans, most of whom face considerable health care and health care technology disparities [1-4]. Over the last 40 years, SCD has changed from a life-threatening disease of early childhood to a chronic disease with multiple organ system involvement [5]. Individuals with SCD can develop issues with pain (acute and chronic), lung (asthma), heart, kidney, and mental health (depression). These complications lead to multifaceted treatment plans and the significant need for self-management. First, individuals with SCD can have acute vaso-occlusive pain episodes as well as chronic pain. These individuals can manage chronic and acute pain episodes using action plans like asthma. Second, disease-modifying medications need to be taken every day to prevent long-term sequelae and decrease pain. Finally, as with other chronic diseases, evidence-based guidelines advise health care providers about managing SCD and its complications [6]. We made them patient-facing to help individuals with SCD learn more about their disease and decide whether to act per the guidelines.

Many mobile health (mHealth) apps have been developed for SCD, but most focus on one aspect of the disease, primarily pain diaries [7-9]. No apps for SCD exist that combine components to be a multipurpose app that engages participants in multiple facets of their disease and their pain [10-12]. There has been no published evaluation of an mHealth app that encompasses various components of the disease process and allows individuals to tailor their needs to the mHealth app's functionality.

We used the health belief model as our theory for health behavior decision-making to inform the design of our existing mHealth app for this study [13]. Our previous work described evidence-based recommendations that we included in our mHealth app, such as (1) guideline information that is easily accessible, understandable, and actionable; (2) alerts and cues, such as reminders, to promote awareness and adherence to health behaviors; (3) shared decision making including pain diaries, checklists of guidelines, and action plans to discuss with patient care teams; (4) decision support to help them understand the choices available to them in the guidelines, such as pain and asthma action plans, and (5) integration with personal preferences, for example, health goal setting (Figures 1 and 2).

We conducted a feasibility randomized controlled trial to determine the acceptability of our mHealth app [14]. A multipurpose mHealth app can allow patients to monitor and track their pain, but additional features can increase app usage and engagement in their disease. As part of that trial, in this manuscript, we aimed to describe participants' usage patterns and test the hypotheses that different usage would be associated with trial outcomes of knowledge of SCD-specific guidelines and hospitalizations and acute health care utilization.

Figure 1. Screenshots from selected features of the mobile health app for adults with sickle cell disease used in this randomized controlled trial showing the (**A**) guideline, (**B**) checklist, (**C**) quiz, and (**D**) pain tracker.

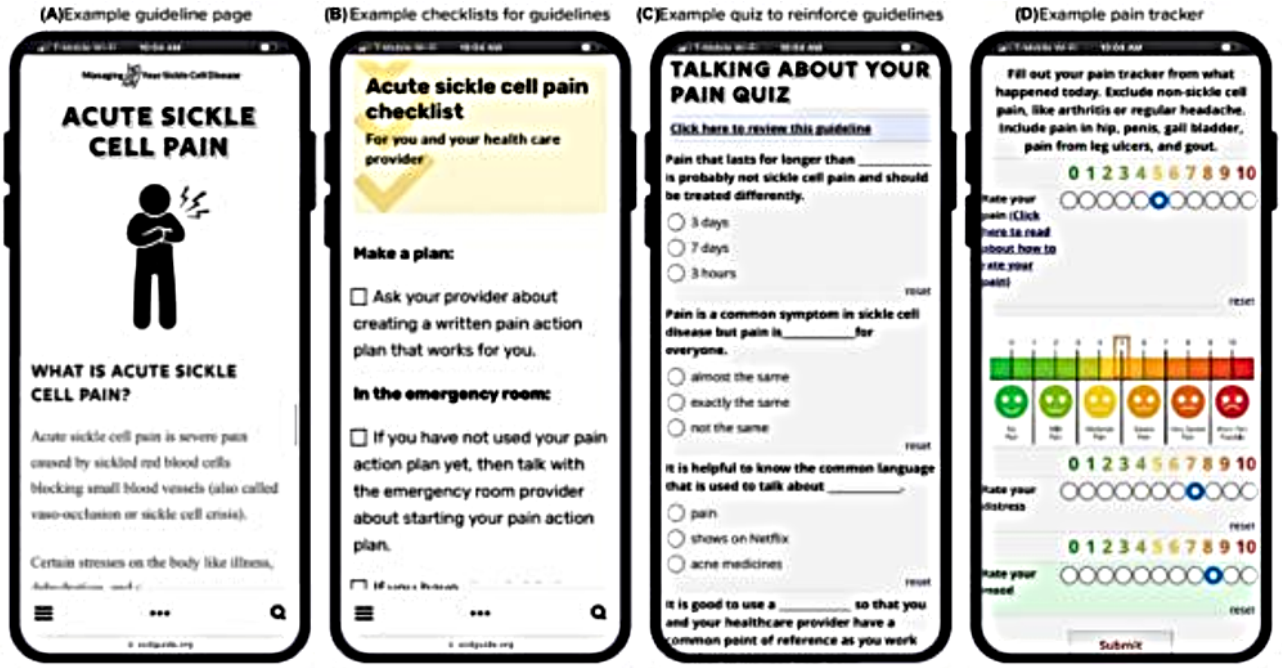

XSL•FO
RenderX

Figure 2. The mHealth app for adults with SCD used in this randomized controlled trial targets knowledge and skills for SCD through its patient-facing guidelines and self-management features. mHealth: mobile health; SCD: sickle cell disease.

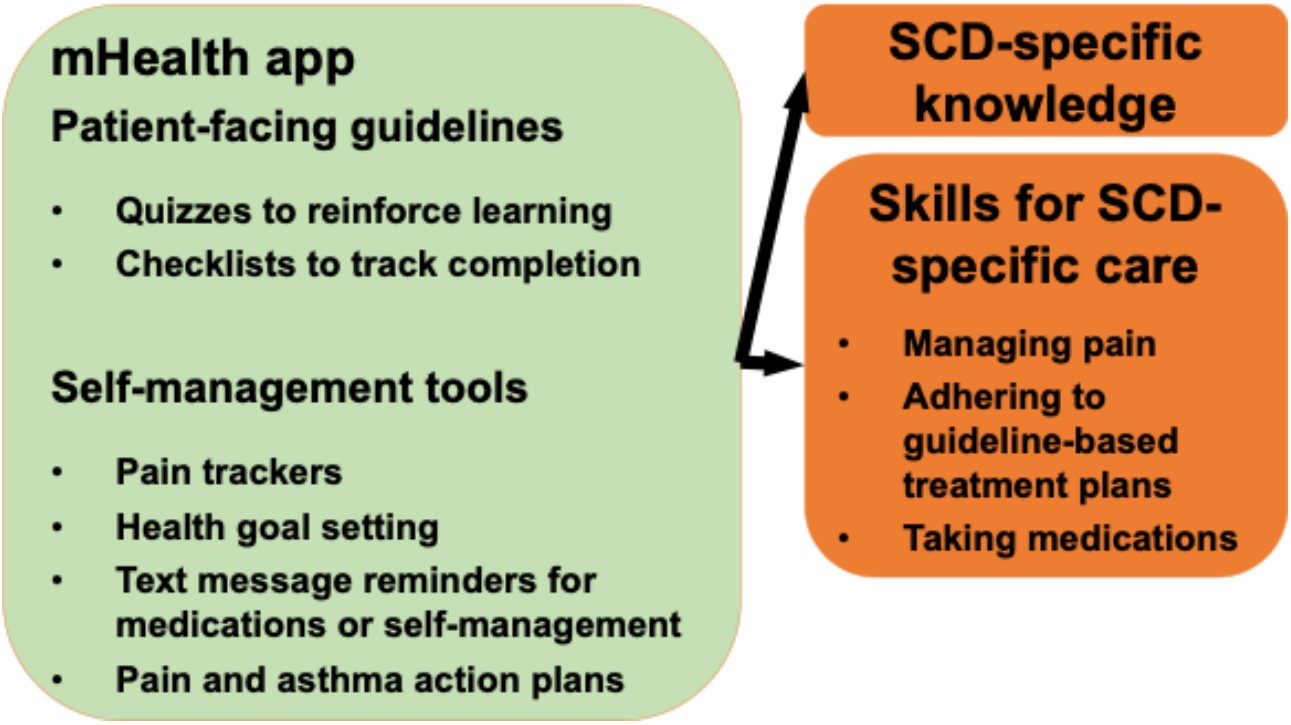

Methods

Study Design

We conducted a parallel randomized controlled trial to assess feasibility as previously described [14] at 2 medical centers—Vanderbilt University Medical Center and the Ohio State University—from 2018 to 2022 (Checklist 1). Inclusion criteria encompassed patients within the SCD clinic with the following genotypes (Hemoglobin SS, Hemoglobin SC, and Hemoglobin Sβ-thal): proficient in spoken and written English, possessing a smartphone or computer, and being 18 to 70 years old. The trial ended due to reaching recruitment goals.

Participants were divided into two arms: (1) the mHealth app arm, where individuals received the mHealth app along with a booklet containing patient-facing guidelines, and (2) the booklet-only arm, where participants received solely a booklet with patient-facing guidelines. Participants were randomized in 1:1 ratio between the 2 arms without any restrictions. Randomization allocation sequence was determined by using a computer-generated random allocation sequence and implemented using the Research Electronic Data Capture (REDCap) randomization module.

The study involved an initial baseline and follow-up visits at 6 months. During the initial visit, after consenting, participants were randomized, and baseline surveys were administered by key study personnel. Participants in the mHealth app arm were introduced to the mHealth app, which took approximately 10 - 15 minutes. At the 6-month follow-up, surveys were repeated, and all survey instruments were available electronically through REDCap, administered via iPads.

Ethical Considerations

Institutional review boards at both centers—Vanderbilt University Medical Center (171036) and the Ohio State University (2020H0468)—approved the trial. Recruitment occurred at SCD clinics, and written informed consent was obtained before randomization. Compensation included a US $25 gift card for the initial visit and US $50 for the 6-month follow-up. Data were anonymized for analysis to help preserve privacy and confidentiality (Multimedia Appendices 1-4).

The mHealth App

The mHealth app comprised 2 main components: a web application featuring patient-facing guideline content and a REDCap project. The web application mirrored the content found in the previously mentioned booklets. It included sections providing guideline-based information on treatments (such as medications, transfusions, and transplants), complications (including acute sickle cell pain, infections, acute chest syndrome, pulmonary hypertension, leg ulcers, strokes, kidney disease, and depression), and wellness topics (women's health, prevention and screening, pain triggers avoidance, stress management, dental care, diet and exercise, as well as tobacco, alcohol, and drug use).

The REDCap project featured a profile and interactive functionalities. Initially, participants could answer questions to generate their profiles, allowing them to access links for reviewing guidelines, editing their profiles, tracking pain, setting goals or missions, taking chapter quizzes related to the guidelines, and creating either a pain or an asthma action plan. The REDCap project also facilitated sending daily text messages at a time of the participant's choosing, containing a customizable message and a link to their profile.

XSL•FO
RenderX

Missions were user-defined missions, goals, or tasks to complete during the trial. Specifically, participants could set a 6-month mission and as many weekly missions as they would like. Some examples of these missions included: “To better manage my pain without pain meds,” “Drink more water,” “No hospital visits,” “To Take all my medication on time each day,” “Recover from the hospital,” “To get a 9 to 5 job,” “Lose weight,” and “Establish and commit to a night routine to relax.”

Participants could document their daily pain experiences within the pain tracker and the strategies used to manage pain, including recording their mood. The goals feature enabled participants to monitor their weekly or 6-month goals. Chapter quizzes were available to reinforce the website’s content and the guidelines. The pain action plan guided individuals through a stepwise approach to managing acute pain at home. In contrast, the asthma action plan was designed for those seeking to develop a strategy for managing their asthma in collaboration with their health care provider. This mHealth app aimed to reinforce key guideline points and promote patient engagement through quizzes and text message reminders.

Measurements of the mHealth App Usage and Outcomes

Primary and secondary outcomes were previously described [14]; in this manuscript, we evaluated mHealth app feature usage and associations with our outcomes. Pain diaries, goal-setting diaries, and quizzes were REDCap surveys, so we measured app feature usage based on the completion of these REDCap surveys. SCD knowledge outcome was measured with a 40-item questionnaire survey covering the guidelines during baseline and follow-up visits. Demographics were self-reported through surveys. Hospitalizations and acute health care utilization (hospitalization, emergency room visits, and day-hospital visits) outcomes were determined by chart review in the first 6 months following baseline and within 6 months prebaseline.

Statistical Analysis

The sample for analysis of app usage was those who were randomized to the mHealth app arm, completed all baseline surveys, and either completed the study up to but not necessarily including the final endpoint survey or withdrew for a reason possibly related to the intervention (ie, not transplant or death). We used descriptive statistics, including median and IQR for continuous data and percentages for categorical data. Linear and generalized linear models evaluated baseline variables and usage differences between groups. Associations between usage of app components (diaries, quizzes, missions, and pain action plan) were evaluated using Wilcoxon test and Spearman correlations, as appropriate.

Results

Demographics and Overall App Usage

A total of 67 participants were randomly allocated; 37 were enrolled in the mHealth app arm and 30 in the booklet alone arm (Figure 3). In total, 30 users who completed baseline surveys and were introduced to the mHealth app were eligible for app usage analysis: 26 completed the entire study, 3 withdrew after completing all baseline surveys but before the final follow-up survey, and 1 did not complete the final follow-up survey. There were no harms or unintended effects. Table 1 shows the baseline characteristics of the participants. All eligible users created a profile and filled out at least one pain diary. While there were no requirements to use the mHealth app, 90% (27/30) participants used at least one additional feature. Features in order of highest usage were missions (26/30, 87% of eligible participants), additional pain diaries (13/30, 43%), quizzes (8/30, 27%), and pain action plans (6/30, 20%). Further, 50% (15/30) participants used multiple features, not including the first pain diary (Table 2).

XSL•FO
RenderX

Figure 3. CONSORT (Consolidated Standards of Reporting Trails) flow diagram of the mobile health app for adults with sickle cell disease for the 6-month randomized controlled trial period.

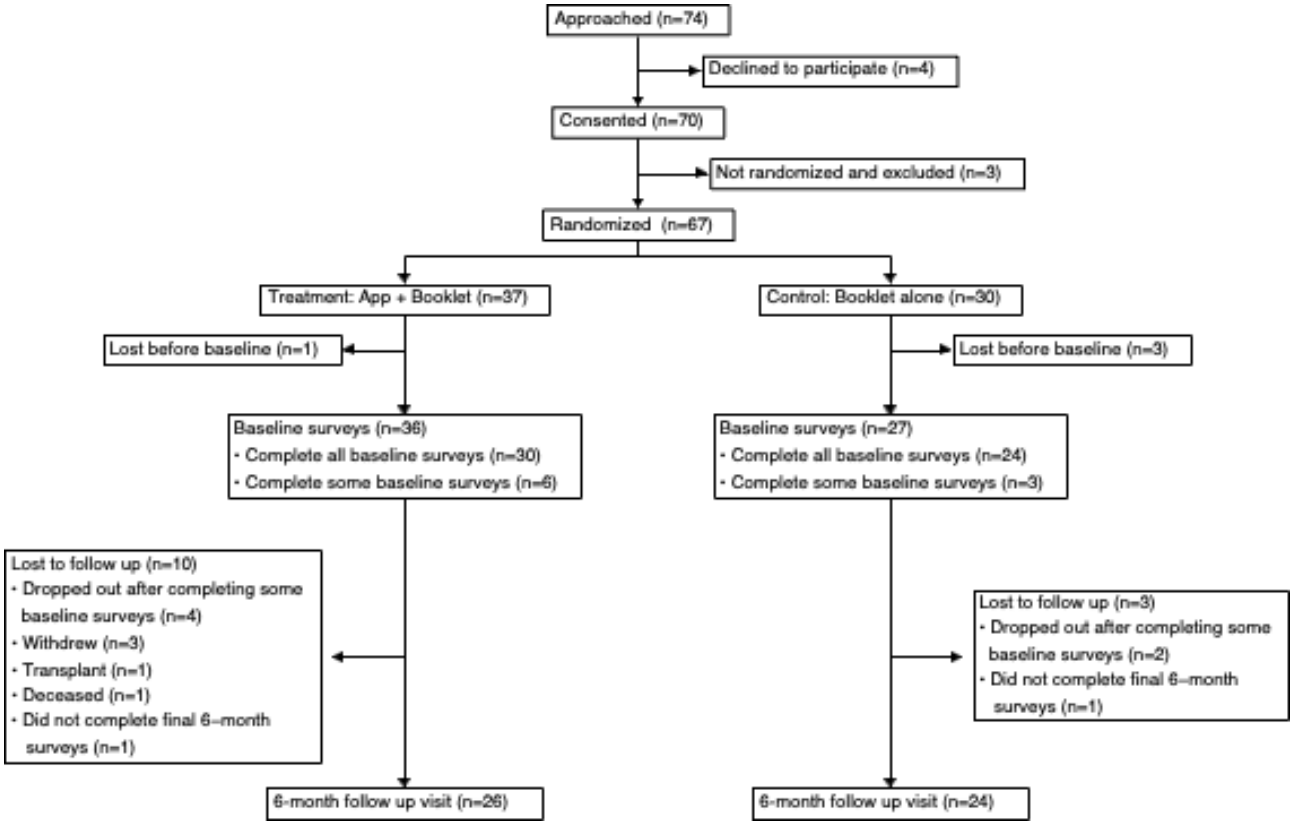

XSL•FO
RenderX

Table . Demographics of the 30 users who completed baseline surveys and were introduced to the mHealth app of this randomized controlled trial in adults with sickle cell disease.

Characteristics	Overall (n=30)
Age (years)	
Mean (SD)	34.5 (10.3)
Median (minimum-maximum)	32.0 [22.0-66.0]
Missing, n (%)	1 (3.3)
Racial identity, n (%)	
Black or African American	29 (96.7)
Missing	1 (3.3)
Ethnicity, n (%)	
Not Hispanic or Latino	25 (83.3)
Unknown or not reported	3 (10.0)
Hispanic or Latino	1 (3.3)
Missing	1 (3.3)
Gender identity, n (%)	
Man	16 (53.3)
Woman	12 (40.0)
Nonbinary	0 (0)
Prefer not to answer	1 (3.3)
Missing	1 (3.3)
Sexual orientation, n (%)	
Lesbian	1 (3.3)
Straight	24 (80.0)
Bisexual	2 (6.7)
Gay	1 (3.3)
Prefer not to answer	1 (3.3)
Missing	1 (3.3)
SCD[a] type, n (%)	
Hemoglobin SS/Sbeta^0 thalassemia	19 (63.3)
Hemoglobin SC/Sbeta$^+$ thalassemia	10 (33.3)
Do not know	0 (0)
Missing	1 (3.3)
History of asthma, n (%)	
Yes	9 (30.0)
No	20 (66.7)
Missing	1 (3.3)
SCD knowledge score	
Mean (SD)	31.4 (4.47)
Median (minimum-maximum)	32.0 (18.0-37.0)
Missing, n (%)	2 (6.7)

[a]SCD: sickle cell disease.

XSL•FO
RenderX

Table . Feature utilization cross-tabulation among participants (adults with sickle cell disease) in this randomized controlled trial who were randomized to the app arm, completed baseline surveys, and were introduced to the mobile health app.

Feature	Total, n (%)	Nothing else, n (%)	Pain action plan, n (%)	Diaries, n (%)	Quizzes, n (%)	Missions, n (%)
Pain action plan	6 (20)	0 (0)	__[a]	5 (17)	6 (20)	6 (20)
Diaries	13 (43)	1 (3)	5 (17)	—	5 (17)	12 (40)
Quizzes	8 (27)	0 (0)	6 (20)	5 (17)	—	8 (27)
Missions	26 (87)	11 (37)	6 (20)	12 (40)	8 (27)	—

[a]Not applicable.

App Function Usage and Association With Baseline Variables and Other App Purposes

Participants Had Different Patterns of Usage of Pain Diaries and Other Functions

While a majority did not use the pain diary more than once (n=17), those who used the pain diary more than once had different patterns of use (Figure 4). Some used it almost daily (Figure 4, usage ranks 1 - 5), measuring their chronic and sometimes acute pain. Others seem to only measure it from time to time (usage ranks 6 - 7), possibly indicating they had no pain some of the time and pain other times, which could indicate days of chronic pain or acute pain exacerbations. Others used it more sporadically (usage ranks 8 - 13), with one using it for about a week and then stopping (usage rank 8). In total, 21 of the 30 participants randomized to the app arm (70%) completed at least one more function after enrollment (Figure 5). Fourteen completed functions at 90 days (the half-way point of the trial), nine within the last month of the trial, but only 3 users completed functions after the trial was over.

Figure 4. Pain diary usage of the mobile health app for adults with sickle cell disease for the 6-month randomized controlled trial period who completed more than one pain diary.

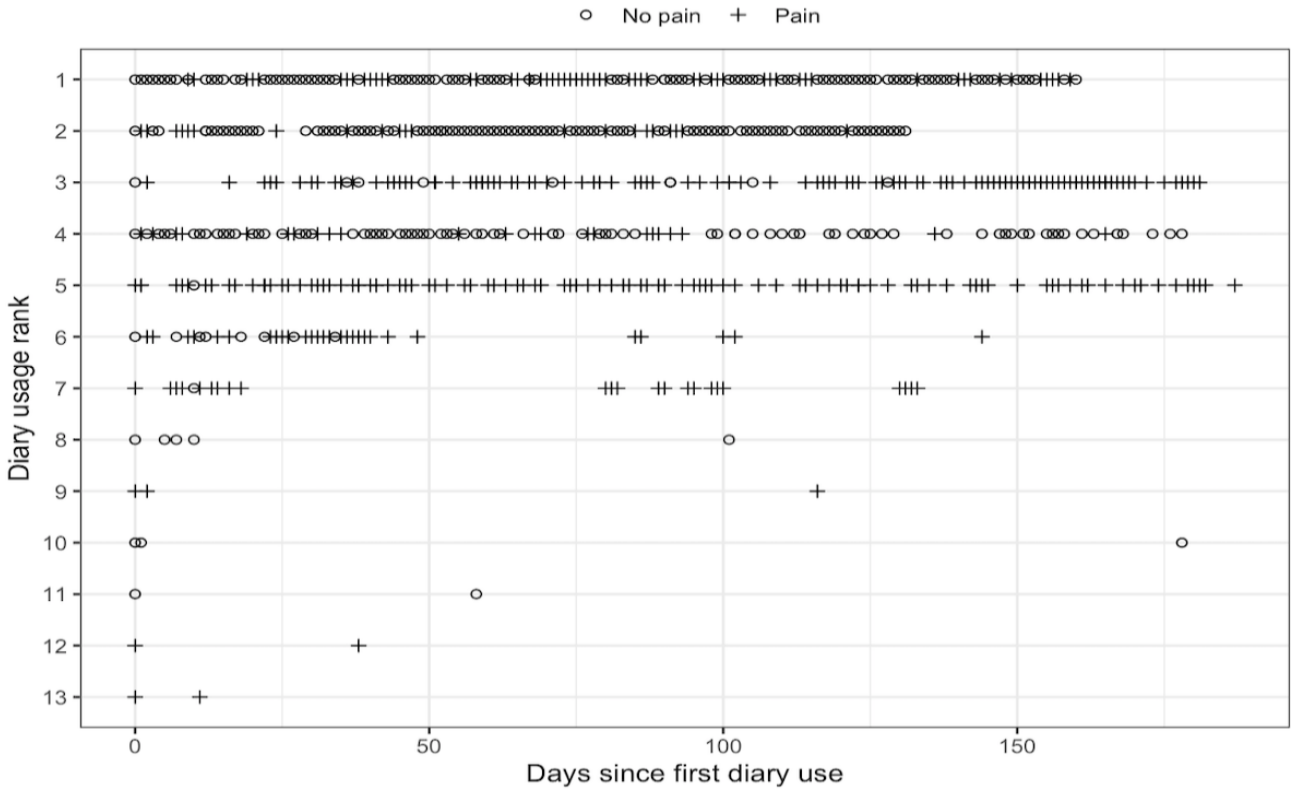

XSL•FO
RenderX

Figure 5. Different app feature usage of the mobile health (mHealth) app for adults with sickle cell disease who were randomized to the app arm, completed baseline surveys, and were introduced to the mHealth app for the 6-month randomized controlled trial period.

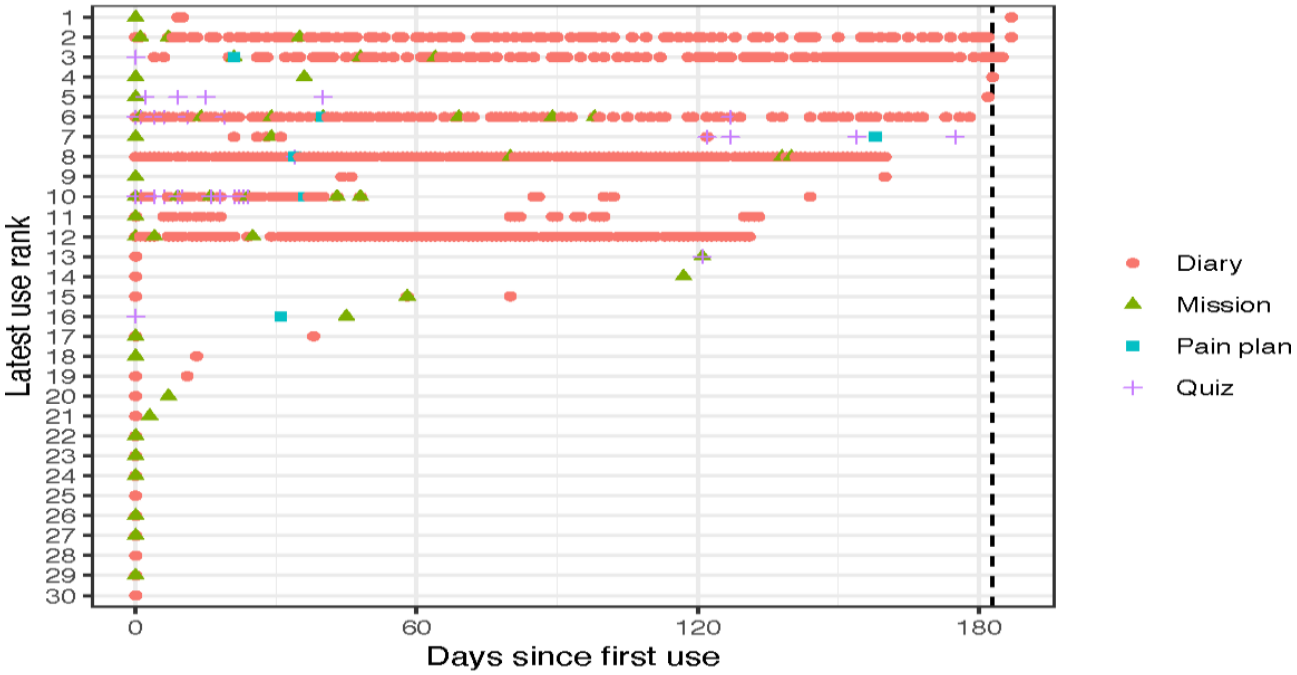

Baseline Knowledge Was Associated With More Pain Diary Completion and Female Gender Identity with More Mission Completion

While baseline age, gender identity, disease type, SCD knowledge, and patient activation were not statistically significantly associated with most functions of the mHealth app, baseline knowledge was significantly associated with diary completion (Table 3). The more knowledge participants had at baseline, the more pain diaries they completed in addition to their first (33% increase per SCD knowledge point, 95% CI 6%-81%; *P*=.04). In addition, female versus male gender identity was statistically significantly associated with completing more missions (mean 3.8, SD 5.9 vs 1.2, SD 0.8; relative rate 317%, 95% CI 114%-1084%; *P*=.05).

Table . Associations between baseline variables and app usage for adults with sickle cell disease (SCD) who were randomized to the app arm, completed baseline surveys, and were introduced to the mobile health app for the 6-month randomized controlled trial period. Measures are relative rate (rr) for binary-count and continuous-count relationships, odds ratio (OR) for binary-binary relationships, and Cohen *d* for binary-continuous relationships.

	Diaries, rr (*P* value)	Pain plan, Cohen *d* or OR (*P* value)	Quizzes, rr (*P* value)	Missions, rr (*P* value)
Age	1.00 (1.00)	1.05[a] (.15)	1.07 (.07)	1.00 (.90)
Gender (female vs male)	2.72 (.22)	3.34[b] (.35)	1.52 (.67)	3.23 (.05)
SCD type[c]	0.35 (.32)	0.32[b] (.63)	0.06 (.22)	0.42 (.27)
SCDK[d]	1.33 (.04)	0.60[a] (.10)	1.09 (.55)	1.07 (.19)

[a]Value represents Cohen *d*.

[b]Value represents OR.

[c]Hemoglobin SC; Hemoglobin Sbeta-thalassemia+ vs Hemoglobin SS; Hemoglobin Sbeta-thalassemia0.

[d]SCDK: sickle cell disease knowledge.

Pain Action Plans, Pain Diaries, Quizzes, and Missions Were All Associated with Each Other

Patients who completed pain action plans completed more quizzes (median 5, IQR 2.5-6.74 vs 0, IQR: 0-0; *P*<.001), missions (median 3, IQR 2.25-5.25 vs 1, IQR 1-1; *P*=.003), and diaries (median 70.5, IQR: 12.8-111 vs 1.0, IQR: 1-2; *P*=.005) compared to those who did not (Table 4). Patients completing more missions tended to complete more diaries (rho=0.69; *P*<.001) and more quizzes (rho=0.47; *P*=.008), and those completing more diaries tended to complete more quizzes (rho=0.37; *P*=.047). Although the rank correlation between diaries and quizzes was positive, the relationship is more informatively described by clusters: most patients (n=23) completed few (0 - 24) diaries and few (0 - 4) quizzes, some (n=5) completed the most (100 - 160 of 180 possible) diaries but few (0 - 7) quizzes, and the remainder (n=2) completed the most (20 and 21 of 21 possible) quizzes but few (36 and 1) diaries. One patient was a notable outlier and completed the most missions (21, second-highest completed 8) and the

XSL•FO
RenderX

second-most diaries (121, highest completed 157 and third-highest completed 113), but no quizzes or pain action plan.

Table . Associations between app usage features for adults with sickle cell disease who were randomized to the app arm, completed baseline surveys, and were introduced to the mobile health app for the 6-month randomized controlled trial period. Measures are Spearman rank correlations (rr) for count-count relationships and difference in medians (dm) for binary-count relationships.

	Diaries, dm or rr (*P* value)	Pain plan, dm or rr (*P* value)	Quizzes, dm or rr (*P* value)	Mission, dm or rr (*P* value)
Diaries	—[a]	69.5[b] (.005)	0.4[c] (.047)	0.7[c] (<.001)
Pain plan	69.5[b] (.005)	—	5.0[b] (<.001)	2.0[b] (.003)
Quizzes	0.4[c] (.047)	5.0[b] (<.001)	—	0.5[c] (.008)
Missions	0.7[c] (<.001)	2.0[b] (.003)	0.5[c] (.008)	—

[a]Not applicable.

[b]Value represents dm.

[c]Value represent rr.

We also evaluated the association of usage of different features with the trial outcomes, knowledge of SCD-specific guidelines and hospitalizations, and acute health care utilization, but the usage was not significantly associated with any outcomes (Table 5). A notable but not statistically significant relationship was observed between missions completed and hospitalizations. Of the 3 participants who had more hospitalizations in the 6 months of the trial than the 6 months before baseline, 2 completed no missions, and one completed 1 mission. In contrast, of the 5 participants with fewer hospitalizations during the trial than before, 3 completed one, 1 completed three, and one completed 21 missions, the highest of any participant. The Spearman rank correlation between missions completed and change in hospitalizations was rho=−0.38 (more missions were associated with fewer hospitalizations than pretrial; *P*=.05). Similar but weaker associations were observed between missions completed and hospitalizations due to pain (rho=−3.7; *P*=.06) and acute care utilization (rho=−0.31; *P*=.13).

Table . Associations between app usage and outcome for adults with sickle cell disease who were randomized to the app arm, completed baseline surveys, and were introduced to the mHealth app for the 6-month randomized controlled trial period. Measures are Spearman rank correlation (rc) for count-count and count-continuous relationships and difference in medians (dm) for binary-count and binary-continuous relationships.

	SCDK[a]	Hosp[b]	Acute care[c]
Diaries, rc (*P* value)	−0.17 (.43)	−0.16 (.43)	0.03 (.88)
Pain plan, dm (*P* value)	−0.16 (.47)	−0.08 (.69)	0.07 (.75)
Quizzes, rc (*P* value)	0.01 (.95)	−0.03 (.89)	0.18 (.39)
Missions, rc (*P* value)	−0.27 (.20)	−0.38 (.06)	−0.31 (.13)

[a]SCDK: sickle cell disease knowledge.

[b]Hosp: hospitalizations.

[c]Acute health care utilization (emergency room visits or hospitalizations).

Discussion

Principal Findings

mHealth apps are a potentially scalable, efficient, and sustainable strategy to improve adult care in SCD. However, there is a knowledge gap regarding patterns of the usage of multipurpose apps in rare diseases, such as SCD. Our study is the first to demonstrate that adults with SCD have different usage patterns of a multipurpose mHealth app for self-management, and certain features may be related to improved patient outcomes. Almost all users used multiple features. Highly engaged participants were more likely to use either quizzes or diaries, but not both. We also saw that completing certain features, such as missions, was associated with lower hospitalizations, although not significant. Complex chronic diseases, such as SCD, could benefit from multipurpose apps that allow users to engage based on preferences and needs.

As this study was not focused on one part of our app, engagement with features in this mHealth app was entirely patient-driven and may demonstrate the expected real-world use of an mHealth app in this population. Users were most engaged with missions but least engaged with action plans. Missions were one of the most straightforward, least time-consuming features that users could fill out independently, whereas action plans require more time and typically would be done with the provider. Interestingly, less than half of the participants used the pain diaries, potentially because they were not the primary function of the mHealth app or trial. This finding contradicts other studies that demonstrated the high use of pain diaries in SCD [8,15,16]. Usage of pain diaries in the general SCD population could be lower when it is not the primary goal of an mHealth app as part of a specific research study. Further study is needed into why some users completed pain diaries, among other mHealth app features, while others did not. While there were associations between the use of one feature of the mHealth app and another, participants tended to engage with

XSL•FO
RenderX

some but not all features. For example, highly engaged participants engaged with either quizzes or diaries, but not both. Pain action plans, pain diaries, quizzes, and missions were all associated with each other. Almost all participants who engaged with one feature besides the missions would typically engage with additional features.

User engagement with mHealth apps is a critical area of study, as it directly influences the effectiveness of these interventions in promoting treatment adherence and self-management among individuals with chronic health conditions. Recent reviews have described challenges in engagement with mHealth apps, including (1) inconsistent terminology and measurements, limiting our ability to compare results across studies, complicating the development of effective engagement strategies [17]; (2) usability issues, where complex interfaces and poor user experiences in mHealth apps can deter individuals with chronic diseases from consistent use [18]; (3) data privacy and security concerns that can deter users from engaging with mHealth apps [19]; and (4) a lack of personalization, where generic content does not cater to individual user needs can lead to reduced engagement [20]. While our mHealth app included user-centered design to overcome usability issues and tailor the intervention to this population's specific needs, is within REDCap, which has enhanced data privacy and security, and has personalization features, including text message reminders and individualized plans, the app's current version had limited functionality to measure engagement robustly. By addressing these barriers and implementing the suggested strategies, developers and healthcare providers can improve user engagement with mHealth applications, leading to better health outcomes. Future research on user engagement could include post-study user interviews or focus groups to obtain more detailed information about why users engaged with certain functionality and not others.

A notable but not statistically significant relationship was observed between missions completed and hospitalizations. Findings from other studies demonstrated improved outcomes associated with apps [21]. However, while many studies show improvements in outcomes like asthma or diabetes control, few have demonstrated improvements in outcomes like hospitalizations. Also, others have shown the opposite effect of increased hospitalization rates with mHealth app usage [22,23]. Also, in adults with SCD, no studies demonstrated improvement in hospitalizations or acute health care utilization using mHealth apps. However, the feasibility study of our app showed a nonstatistically significant improvement in outcomes. A larger trial with more participants is the next logical step that could demonstrate which of the mHealth app's features might improve outcomes.

Baseline knowledge and female gender identity were associated with more feature usage, whereas other variables, such as age, did not show any significant differences. Others have demonstrated that demographics, such as age, are associated with different usage [24] of an mHealth app; however, only female gender identity was associated with one more feature use. Classifying different types of users may be necessary for understanding and improving long-term engagement.

Our study had a low attrition rate; however, sustained engagement varied. Other studies have demonstrated a wide range of attrition rates from 9% to 82% and an average of 43% [25]. Our attrition rate was lower, which could be because our mHealth app used strategies that have led to lower attrition in other studies, including multitouch interventions and tailored messaging [25]. However, another reason for the low attrition could have been because our mHealth app had multiple features that allowed users to engage in the most important features to them and their disease For sustained engagement with the app, we observed that 70% of the users completed more functions after enrollment, with a continued decline throughout the study period, with only a few participants completing functions after trial completion. However, these features were not the primary feature of the trial, but we could not assess all app usage due to the app's limited analytics. Future research with better analytics and a longer follow-up period is needed in understanding sustained usage of this mHealth app.

Since this mHealth app was built primarily in REDCap, the mHealth app can be accessed and used by any of the REDCap Consortium 7178 active partners across 156 countries. Providing a framework for an mHealth app built within REDCap using base features and external modules would demonstrate scalability by allowing researchers in the REDCap Consortium to efficiently adapt the mHealth app to develop, test, and confirm ideal configurations and settings for user engagement and data collection, which impact data quality and quantity and overall significance and applicability of study results. The consortium can download this mHealth app REDCap project and tailor it to their needs and local/national guidelines. Also, the solution would be more sustainable as it would not require maintenance and updates when operating systems change, like native Apple and Android apps, as the institutions typically update their versions of REDCap. However, for those outside of the REDCap consortium, there would be limitations on adapting this app quickly for their purposes, which would limit scalability to that app for nonconsortium partners.

Several limitations occurred in our study. First, the sample size was small, so we could not observe significant associations between mHealth app feature usage and outcomes. A larger randomized controlled trial would allow for further evaluation of the effects of important app features on outcomes. Second, we could not evaluate the interaction between participants and the guidelines feature of the mHealth app. While we measured how to access guidelines and other app features, we could not measure guideline review. This limited functionality also precluded us from being able to measure app usage after participants completed the trial. Our next version of the mHealth app will include analytics to evaluate guideline use. Third, the app's current version did not include the ability to assess usage in terms of time, so we could not evaluate how long it took to perform different functions. Future hybrid effectiveness-implementation studies with better analytical functionality, a larger population, longer follow-up, and an implementation framework could help our understanding of improving outcomes, understanding user engagement, including sustained use after the trial is completed, and implementing this mHealth app.. Finally, this study did not assess other

XSL•FO
RenderX

determinants of app usage (eg, participation in a community-based organization) or other healthcare experiences (eg, discrimination).

Conclusions

Results of our first-ever multipurpose mHealth app for adults with SCD demonstrated that (1) participants will use an mHealth app and use certain features of an mHealth app depending on their needs, even if these features were not the primary purpose of the mHealth app, and (2) that certain features may lead to improvement in outcomes. A multipurpose mHealth app can help engage participants in features they want to use and potentially increase and improve outcomes. This trial's results have important implications for improving the health of a patient population facing significant healthcare and health information technology disparities.

Acknowledgments

We greatly appreciate the help of Chris Simpson, Renee Ashworth, and Alex Minor in recruiting and retaining participants. We also greatly appreciate the input from the Vanderbilt Meharry Sickle Cell Center of Excellence, including Emmanuel Volanakis, the Sickle Cell Foundation of Tennessee, and the Sickle Cell Disease Association of America on the study and interventions. This research was partly funded by the National Heart, Lung, and Blood Institute (NHLBI), part of the National Institutes of Health (NIH), and the National Center for Advancing Translational Sciences (grant UL1TR002733). RMC was supported by the NHLBI (award K23HL141447).

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files (Multimedia Appendices 1-4).

Authors' Contributions

MRD, LEC, and RMC designed the study. NQ, KL, and RMC collected the data; PMS and XL performed the analyses. MRD, PMS, XL, and RMC interpreted the results; RMC and PMS wrote the first version of the manuscript. Before submission, all authors reviewed and edited the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
App usage data.
[XLSX File, 38 KB - formative_v9i1e67906_app1.xlsx]

Multimedia Appendix 2
Data dictionary.
[TXT File, 3 KB - formative_v9i1e67906_app2.txt]

Multimedia Appendix 3
Subject deidentified data.
[XLSX File, 12 KB - formative_v9i1e67906_app3.xlsx]

Multimedia Appendix 4
Prepost sickle cell disease knowledge scores.
[XLSX File, 11 KB - formative_v9i1e67906_app4.xlsx]

Checklist 1
CONSORT (Consolidated Standards of Reporting Trails) checklist.
[PDF File, 128 KB - formative_v9i1e67906_app5.pdf]

References

1. Brousseau DC, Panepinto JA, Nimmer M, Hoffmann RG. The number of people with sickle-cell disease in the United States: national and state estimates. Am J Hematol 2010 Jan;85(1):77-78. [doi: 10.1002/ajh.21570] [Medline: 20029951]
2. Mvundura M, Amendah D, Kavanagh PL, Sprinz PG, Grosse SD. Health care utilization and expenditures for privately and publicly insured children with sickle cell disease in the United States. Pediatr Blood Cancer 2009 Oct;53(4):642-646. [doi: 10.1002/pbc.22069]

XSL•FO
RenderX

3. Steiner CA, Miller JL. Sickle Cell Disease Patients in US Hospitals, 2004: Agency for Healthcare Research and Quality (US); 2006.
4. Yusuf HR, Atrash HK, Grosse SD, Parker CS, Grant AM. Emergency department visits made by patients with sickle cell disease: a descriptive study, 1999-2007. Am J Prev Med 2010 Apr;38(4 Suppl):S536-S541. [doi: 10.1016/j.amepre.2010.01.001] [Medline: 20331955]
5. Chaturvedi S, DeBaun MR. Evolution of sickle cell disease from a life-threatening disease of children to a chronic disease of adults: The last 40 years. Am J Hematol 2016 Jan;91(1):5-14. [doi: 10.1002/ajh.24235] [Medline: 26547630]
6. Murad MH, Liem RI, Lang ES, et al. 2019 sickle cell disease guidelines by the American Society of Hematology: methodology, challenges, and innovations. Blood Adv 2019 Dec 10;3(23):3945-3950. [doi: 10.1182/bloodadvances.2019000931] [Medline: 31794603]
7. Bakshi N, Stinson JN, Ross D, et al. Development, content validity, and user review of a web-based multidimensional pain diary for adolescent and young adults with sickle cell disease. Clin J Pain 2015 Jun;31(6):580-590. [doi: 10.1097/AJP.0000000000000195] [Medline: 25565585]
8. Shah N, Jonassaint J, De Castro L. Patients welcome the Sickle Cell Disease Mobile Application to Record Symptoms via Technology (SMART). Hemoglobin 2014;38(2):99-103. [doi: 10.3109/03630269.2014.880716] [Medline: 24512633]
9. Palermo TM, Lalloo C, Zhou C, et al. A cognitive–behavioral digital health intervention for sickle cell disease pain in adolescents: a randomized, controlled, multicenter trial. Pain 2023;2022:10. [doi: 10.1097/j.pain.0000000000003009]
10. Badawy SM, Cronin RM, Hankins J, et al. Patient-centered eHealth interventions for children, adolescents, and adults with sickle cell disease: systematic review. J Med Internet Res 2018 Jul 19;20(7):e10940. [doi: 10.2196/10940] [Medline: 30026178]
11. Badawy SM, Cronin RM, Liem RI, Palermo TM. Digital behavioural interventions for people with sickle cell disease. Cochrane Database Syst Rev 2021;2021(4):CD014669. [doi: 10.1002/14651858.CD014669] [Medline: 34650329]
12. Hankins JS, Potter MB, Fernandez ME, et al. Evaluating the implementation of a multi-level mHealth study to improve hydroxyurea utilization in sickle cell disease. Front Health Serv 2022;2:1024541. [doi: 10.3389/frhs.2022.1024541] [Medline: 36925803]
13. Utrankar A, Mayo-Gamble TL, Allen W, et al. Technology use and preferences to support clinical practice guideline awareness and adherence in individuals with sickle cell disease. J Am Med Inform Assoc 2018 Aug 1;25(8):976-988. [doi: 10.1093/jamia/ocy036] [Medline: 29741695]
14. Cronin RM, Quaye N, Liu X, et al. A feasibility randomized controlled trial of an mHealth app vs booklets for patient-facing guidelines in adults with SCD. Blood Adv 2023 Oct 24;7(20):6184-6190. [doi: 10.1182/bloodadvances.2023010676] [Medline: 37595029]
15. Jonassaint CR, Shah N, Jonassaint J, De Castro L. Usability and feasibility of an mHealth intervention for monitoring and managing pain symptoms in sickle cell disease: the Sickle Cell Disease Mobile Application to Record Symptoms via Technology (SMART). Hemoglobin 2015;39(3):162-168. [doi: 10.3109/03630269.2015.1025141] [Medline: 25831427]
16. Heath LE, Heeney MM, Hoppe CC, et al. Successful utilization of an electronic pain diary in a multinational phase 3 interventional study of pediatric sickle cell anemia. Clin Trials 2017 Dec;14(6):563-571. [doi: 10.1177/1740774517723307] [Medline: 28743191]
17. Eaton C, Vallejo N, McDonald X, et al. User engagement with mHealth interventions to promote treatment adherence and self-management in people with chronic health conditions: systematic review. J Med Internet Res 2024 Sep 24;26:e50508. [doi: 10.2196/50508] [Medline: 39316431]
18. Li Z, Lu F, Wu J, et al. Usability and effectiveness of eHealth and mHealth interventions that support self-management and health care transition in adolescents and young adults with chronic disease: systematic review. J Med Internet Res 2024 Nov 26;26:e56556. [doi: 10.2196/56556] [Medline: 39589770]
19. Zhou L, Bao J, Watzlaf V, Parmanto B. Barriers to and facilitators of the use of mobile health apps from a security perspective: mixed-methods study. JMIR Mhealth Uhealth 2019 Apr 16;7(4):e11223. [doi: 10.2196/11223] [Medline: 30990458]
20. Borghouts J, Eikey E, Mark G, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. J Med Internet Res 2021 Mar 24;23(3):e24387. [doi: 10.2196/24387] [Medline: 33759801]
21. Forbes A, Keleher MR, Venditto M, DiBiasi F. Assessing patient adherence to and engagement with digital interventions for depression in clinical trials: systematic literature review. J Med Internet Res 2023 Aug 11;25:e43727. [doi: 10.2196/43727] [Medline: 37566447]
22. Yao Y, Guo Y, Lip GYH, mAF-App II Trial investigators. The effects of implementing a mobile health-technology supported pathway on atrial fibrillation-related adverse events among patients with multimorbidity: the mafa-II randomized clinical trial. JAMA Netw Open 2021 Dec 1;4(12):e2140071. [doi: 10.1001/jamanetworkopen.2021.40071] [Medline: 34932104]
23. Marcolino MS, Oliveira JAQ, D'Agostino M, Ribeiro AL, Alkmim MBM, Novillo-Ortiz D. The impact of mHealth interventions: systematic review of systematic reviews. JMIR Mhealth Uhealth 2018 Jan 17;6(1):e23. [doi: 10.2196/mhealth.8873] [Medline: 29343463]
24. Schmidt M, Babcock L, Kurowski BG, Cassedy A, Sidol C, Wade SL. Usage patterns of an mHealth symptom monitoring app among adolescents with acute mild traumatic brain injuries. J Head Trauma Rehabil 2022;37(3):134-143. [doi: 10.1097/HTR.0000000000000768] [Medline: 35125434]

XSL•FO
RenderX

25. Meyerowitz-Katz G, Ravi S, Arnolda L, Feng X, Maberly G, Astell-Burt T. Rates of attrition and dropout in app-based interventions for chronic disease: systematic review and meta-analysis. J Med Internet Res 2020 Sep 29;22(9):e20283. [doi: 10.2196/20283] [Medline: 32990635]

Abbreviations

mHealth: mobile health
REDCap: Research Electronic Data Capture
SCD: sickle cell disease

Edited by A Mavragani; submitted 15.12.24; peer-reviewed by CD Dampier, S Jacobs; revised version received 08.04.25; accepted 10.04.25; published 05.06.25.

Please cite as:
Cronin RM, Quaye N, Liu X, Landes K, Crosby LE, Kassim AA, DeBaun MR, Schnell PM
Usage of a Multipurpose mHealth App Among Adults With Sickle Cell Disease: Randomized Controlled Trial
JMIR Form Res 2025;9:e67906
URL: https://formative.jmir.org/2025/1/e67906
doi:10.2196/67906

© Robert M Cronin, Nives Quaye, Xin Liu, Kristina Landes, Lori E Crosby, Adetola A Kassim, Michael R DeBaun, Patrick M Schnell. Originally published in JMIR Formative Research (https://formative.jmir.org), 5.6.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Collecting Real-Life Psychophysiological Data via Wearables to Better Understand Child Behavior in a Children's Psychiatric Center: Mixed Methods Study on Feasibility and Implementation

Karin Hagoort*, MSc; Kirsten Smeets*, PhD; Saskia Koldijk, PhD; Floortje Scheepers, Prof Dr; Fleur Velders, PhD

Department of Psychiatry, University Medical Center Utrecht, Heidelberglaan 100, Utrecht, The Netherlands

*these authors contributed equally

Corresponding Author:
Karin Hagoort, MSc
Department of Psychiatry, University Medical Center Utrecht, Heidelberglaan 100, Utrecht, The Netherlands

Abstract

Background: In the field of mental health care, the incorporation of wearable devices into routine clinical practice continues to face significant challenges, despite the presence of supporting scientific evidence. Crossing the wasteland between the trial world and the real world is full of obstacles that often only become apparent during the implementation process.

Objective: The objective of this paper was to evaluate the feasibility of using wearables in real-world clinical settings for children with severe developmental problems to help understand and manage disruptive behavior and to gain insights for the development of forthcoming implementation strategies.

Methods: A mixed methods design was used to examine two different aspects of the use of wearables in a clinical setting. The first quantitative part of this study focuses on the feasibility of using wearables to collect reliable data on psychophysiological measures during daily activities in children at a children's psychiatric center. The second qualitative part focuses on the evaluation of the implementation process using the Consolidated Framework for Implementation Research (CFIR) to identify essential steps to successfully incorporate wearable technology in clinical care for children with severe behavioral problems. Empatica E4 wristbands collected data on children's psychophysiological arousal (eg, heart rate [HR] and skin conductance level [SCL]). Staff reported aggressive behavior and daily activities. Data were processed and visualized in a dashboard. User experiences were assessed through interviews with clinical staff. The implementation process was evaluated using the CFIR.

Results: A total of 30 children (27 boys and 3 girls, aged 6 to 14 y; mean age 9.3 y, SD 1.95) wore the wearable for 5 consecutive days. As expected, the children found it easy to wear the device and the clinical staff predominantly expressed positive attitudes toward its use. The data collection proceeded relatively smoothly, and the collected data were of sufficient quality. In total, 315 observations of aggressive behavior were reported, including 54 red incidents (from 18 unique participants) and 261 orange incidents (from 26 unique participants). An exploratory analysis on the association between psychophysiological measures and aggressive behavior revealed that children's HR was significantly higher during aggressive incidents compared to nonaggressive incidents (P=.007). Although not statistically significant, there was a trend suggesting higher peaks per minute during aggressive incidents (P=.07). No significant differences between aggressive and nonaggressive incidents were found for SCL and movement (P=.33 and P=.60). The most challenging CFIR domains in our study were the "characteristics of the intervention" and "the inner setting," reflected in the fact that that the majority of implementation activities were focused on these two domains.

Conclusions: The use of wearables in a real-world study setting is considered feasible and valuable. However, for broader scaling in daily clinical practice, coherent actions on different domains of implementation are required.

(JMIR Form Res 2025;9:e65559) doi:10.2196/65559

KEYWORDS

feasibility study; wearable technology; psychophysiological data; arousal; child psychiatry; aggressive behavior; implementation science; wearable; device; clinical care; feasibility; child; children; psychiatric center; psychiatry; mental health; aggression; observational; aggressive; behavior; wristband; psychophysiological arousal; interview; implementation

Introduction

Technical innovations could be of great benefit in mental health care. Health technology like virtual reality, apps, and wearables have been developed rapidly in highly competitive markets, but only a few are scientifically validated and even fewer ultimately find their way to clinical practice [1]. Wearables have the potential to provide clinicians with valuable and objective real-time data on patients' physiological and emotional states, both within and outside the clinical setting [2]. These devices

XSL•FO
RenderX

enable the measurement of physiological activities, such as electrodermal activity (EDA) and heart rate (HR) [2-5]. This is particularly of interest because elevated psychophysiological measures (ie, HR and EDA) have been linked to emotional and behavioral dysregulation, including aggressive behavior. Given that aggressive behavior is often the main reason of referral for a psychiatric evaluation in a children's psychiatric clinic and imposes a substantial burden on the child, family, and health care staff, these innovative devices could enhance current treatment programs. By objectively monitoring changes in psychophysiological arousal and subsequent behavior in the patient's natural environment, we may gain deeper insights into the onset and regulation of aggressive behavior. Although previous studies have explored these effects, research on children and adolescents remains limited [3,4,6-9].

Furthermore, research involving wearables is predominantly conducted in artificial laboratory settings using experimental designs or in small pilot studies. Consequently, there is limited data on the use of wearables in real-time and naturalistic clinical settings [9-11], where they could significantly impact the overall quality of care. Besides the feasibility of collecting psychophysiological data in child psychiatric care with wearables, it is challenging to effectively implement these technologies into real-world clinical practice [12,13]. This gap between inventive technologies and successful implementation processes is commonly referred to as the "valley of death" and creates a waste of research investments [14]. Implementation can be defined as a series of planned and guided activities aimed at introducing and maintaining technologies within a specific context in order to innovate or improve health care [15]. Common obstacles in implementing technology include lack of funding, increased workload for professionals or resistance to using the technology by health care workers or patients in the long term [16]. Hence, the road to sustainable implementation of health technology is a bumpy one and full of obstacles to overcome. However, few studies focused on the process to effectively implement new technologies in clinical practice especially in complex settings like mental health institutes [12,13,17]. To overcome this valley of death and successfully implement innovative technologies, numerous implementation models and frames have been developed [18]. In the Consolidated Framework of Implementation Research (CFIR), common factors from these models have been brought together within one framework aiming to identify potential barriers and facilitators to implementation and guide the development of strategies to optimize implementation outcomes [19,20].

The two aims of our study are (1) to examine the feasibility of using wearables in inpatient clinical care to collect reliable data on psychophysiological measures as a proxy for arousal and (2) to evaluate the implementation process to successfully incorporate wearable technology in clinical care for children with severe behavioral problems.

Taken together, our hypothesis is that it is feasible to use wearable technology (in this case Empatica E4) in clinical care, aiming to provide children and their environment with more insight into their emotional and physical state and behavior. Using a mixed methods design, this study is composed of a quantitative part in which data is collected using wearables and a qualitative part on user experience and the implementation process. Hence, we focus a priori on outcome and implementation, whereby the intervention is studied while also explicitly observing and gathering information on its implementation [15,21], in an attempt to bridge the aforementioned valley of death.

Methods

Participants and Setting

The study was conducted at the children's psychiatric center of the University Medical Center in Utrecht (UMCU) in The Netherlands, where children aged 6 to 14 years with severe developmental problems are admitted. The center includes an inpatient ward (clinic) for weekday stays, including evenings and nights, and a daycare unit operating from 8 AM to 3 PM on weekdays (daycare). In total, 18 children from the clinic and 12 children from the day care unit were included (see Study Population subsection in the Results section for further details). The majority of the children attended a special education school located adjacent to the treatment center for at least a few hours each day. No inclusion or exclusion criteria were formulated.

Ethical Considerations

The ethics approval was reviewed by the Medical Research Ethics Committee at the UMCU, the Netherlands. The study was evaluated as "no research related medical ethical approval needed" since it imposes a minimal additional burden for the patient with no further interruption or change of their regular clinical program. Participants did not receive compensation. Every Empatica E4 watch has a serial number, which was linked to the personal code of the participant. Only 2 researchers received access to the link between both codes to connect observations to the physiological wearable data. All participants and parents were informed about the goal of the study, data storage, privacy, and the anonymization method of the data and they signed written informed consent. All data will be saved for 15 years after the end of the study.

Design and Procedures

A mixed methods design was used to examine two different aspects of the use of wearables in a clinical setting. The first mainly quantitative part of this study focuses on the feasibility of using wearables to collect reliable data on psychophysiological measures during daily activities in children at a children's psychiatric center.

The second mainly qualitative part focuses on the evaluation of the implementation process using the CFIR to identify essential steps to successfully incorporate wearable technology in clinical care for children with severe behavioral problems. The Good Reporting of A Mixed Methods Study (GRAMMS) recommendations were used to report this study [22] (see Checklist 1).

During the study, children were asked to wear the Empatica E4 wristwatch on their nondominant wrist for 5 consecutive days. This device measured EDA, HR, movement, and skin temperature (see the Psychophysiological Data subsection for

XSL•FO
RenderX

more detailed information on the measurements). The children wore the wristwatch during all normal daily activities, such as school, sports, and treatment programs. At the clinical ward, the children wore the watch from 8 AM to 8 PM, and at the daycare unit, from 8 AM to 3 PM. Over a period of 5 days, the goal was to collect data for 60 hours per child at the clinic and 35 hours per child at the daycare unit.

In addition, behavioral observations with a focus on potential aggressive behavior of the participants were administrated by staff members and 2 independent researchers (see Material and Measurements subsection). At the end of each day, the data of the devices were collected, saved on the computer by the researchers and uploaded into a dashboard. The dashboard with psychophysiological data of 1 day was shown to the clinical staff members to retrospectively evaluate data peaks during the day. After wearing the wearable for 5 consecutive days, children received a diploma and a first insight in the data was shown to their parents. The data were recorded locally on the device. By connecting the wearable to the computer, the accompanying software extracts and synchronizes all recorded data to the Empatica server (as is default). After downloading all data as a zip file containing several .csv files, the data were only locally stored at the UMCU and deleted from the server. All further processing was done locally on UMCU computers. At the end of the study, we asked therapists to participate in an interview for evaluation of the feasibility of using and implementing wearables in daily clinical care. Furthermore, the implementation activities and experiences during the study were carefully registered in meeting notes and a logbook.

Materials and Measurements

Psychophysiological Data

Participants were provided with the Empatica E4 smartwatch [23]. This device was investigated and validated in former research [23]. Sensors in this wristband allow real time measuring of EDA, blood volume pulse (on which HR is based), skin temperature, and x, y, z acceleration measurements (on which movement is based). Since the dashboard provided by Empatica did not contain all the information required to properly assess and interpret the data for clinical practice, a customized dashboard was created. First, the data from the Empatica.csv files were plotted: skin temperature, EDA, HR, and inter-beat interval (IBI). Also, the amount of movement and heart rate variability (HRV; calculated from IBI data) were plotted. Artifact and peak detections were performed on the raw EDA data, which resulted in more detailed parameters such as the skin conductance level (SCL) and the number of peaks per minute (PPM). See the study by de Looff et al [24] and Multimedia Appendix 1, for a detailed explanation of the data processing pipeline, outlining the steps from raw Empatica data to the preprocessed data. To interpret measurements in relation to behavior, a panel was created to display text from observations and corresponding marks were added to the SCL plot. Furthermore, to facilitate interpretability, moving averages and reference lines were added to the plots.

Child Behavior

Child behavior was observed using 2 complementary ways. First, clinical staff recorded child observations for diagnostic purposes in the electronic patient file, following standard procedures. Second, all agitated and potentially aggressive behaviors were documented by independent observers (research interns) or staff (clinical staff and teachers), including the timeframe and specific description of the behavior. Aggressive behavior was coded according to the behavior program coding schema (“traffic light model”) as used in the daycare and clinic (see Table 1). “Green” behavior was classified as normal or very good behavior (eg, receiving a compliment for correct behavior). “Orange” behavior included more disruptive actions, while “red” behavior was considered actual physical aggression and severe disruptive behavior. Clinical staff had no access to the children’s smartwatch data during the day and were unaware of their physiological state of arousal while making their observations.

Table . Traffic light model for behavioral observations.[a]

Color	Behavior	Label
Green	Child follows the rules and behaves in a desired way.	Good behavior
Orange	Swearing, not listening, challenging, arguing, deriding, being agitated, climbing on furniture, and giving the middle finger.	Disruptive behavior
Red	Kicking, hitting, punching, pulling, threatening, pushing, hurting someone, destroying stuff, discriminating, scratching someone else, and spitting.	Physical aggression and severe disruptive behavior

[a]Aggressive behavior was coded according to the behavior program coding schema (“traffic light model”) as used in the daycare and clinic. “Green” behavior was classified as normal or very good behavior, “Orange” behavior included more disruptive actions, while “red” behavior was considered actual physical aggression or severe disruptive behavior.

User Experience and Implementation

Clinical staff were asked to complete an interview after the monitoring period ended. The interview was guided by the Wearable Computer Rating Scale developed by Knight and Baber [25], which is used in other feasibility studies (eg, [26]). To get more insight in the experiences in clinical practice while using the device, questions focused on experiences of the children wearing the device and experiences of the clinical staff regarding the usability of the device in daily practice (what did

XSL•FO
RenderX

you like, what did you not like about it, and did the children wear it reliably) were collected (Multimedia Appendix 2). To evaluate the activities during the implementation process, meeting reports, agenda appointments, email exchanges, and action plans were collected in addition to user experiences.

Data Analysis

Psychophysiological Measures and Child Behavior

To explore the association between psychophysiological measures and aggressive behavior, we analyzed psychophysiological data during instances of aggressive behavior in children (red incidents). Furthermore, 5-minute data frames were extracted at the onset of aggressive behavior for this analysis. The recorded time of the aggression incident is the starting point of the 5-minute data frame per incident. To facilitate within-person comparisons, reference data were extracted for each incident from the same participant, at the same time of day, recorded on other weekdays. Reference days were excluded when aggressive behavior was also observed during the reference episode (35/192 18.2%). The following psychophysiological measures were calculated during aggressive episodes and reference episodes: movement that represents the average levels of movement, as well as HR, SCL, and PPM that refer to nonspecific skin responses. Similar to de Looff et al [4], frames with more than 75% artifacts in the EDA signal were deemed unreliable and were excluded from analysis (354/1723, 20.5%). To determine the extent to which psychophysiological measures during aggressive incidents differ from reference episodes, an exploratory paired-sample *t* tests was performed. A total of 4 hypotheses (movement, HR, SCL, and PPM) were tested with a Bonferroni adjusted alpha level of .013 per test (.05/4). More advanced multivariate analysis was beyond the scope of this feasibility study.

Implementation Process

To be able to identify the key barriers and facilitators retrospectively, all implementation activities, events, and incidents from the multiple sources were chronologically listed and categorized with guidance of the 5 domains of the CFIR [18,19]. Each domain includes multiple constructs (see Figure 1) that help evaluate the implementation process and identify potential barriers or facilitators for implementing wearables in clinical practice. Based on the analysis of the activities and the experiences of the stakeholders, the research team discussed and determined which constructs from the different CFIR domains have been relevant to this study and to what extent these constructs were facilitators or barriers. Furthermore, recommendations for future implementation strategies and clinical uptake are given. We describe the findings of the process evaluation, including the identification of facilitators and barriers to implementation within each domain. In addition, we present the lessons learned for further uptake, which strategies facilitated dealing with identified barriers.

Figure 1. The Consolidated Framework of Implementation Research (CFIR) domains and constructs.

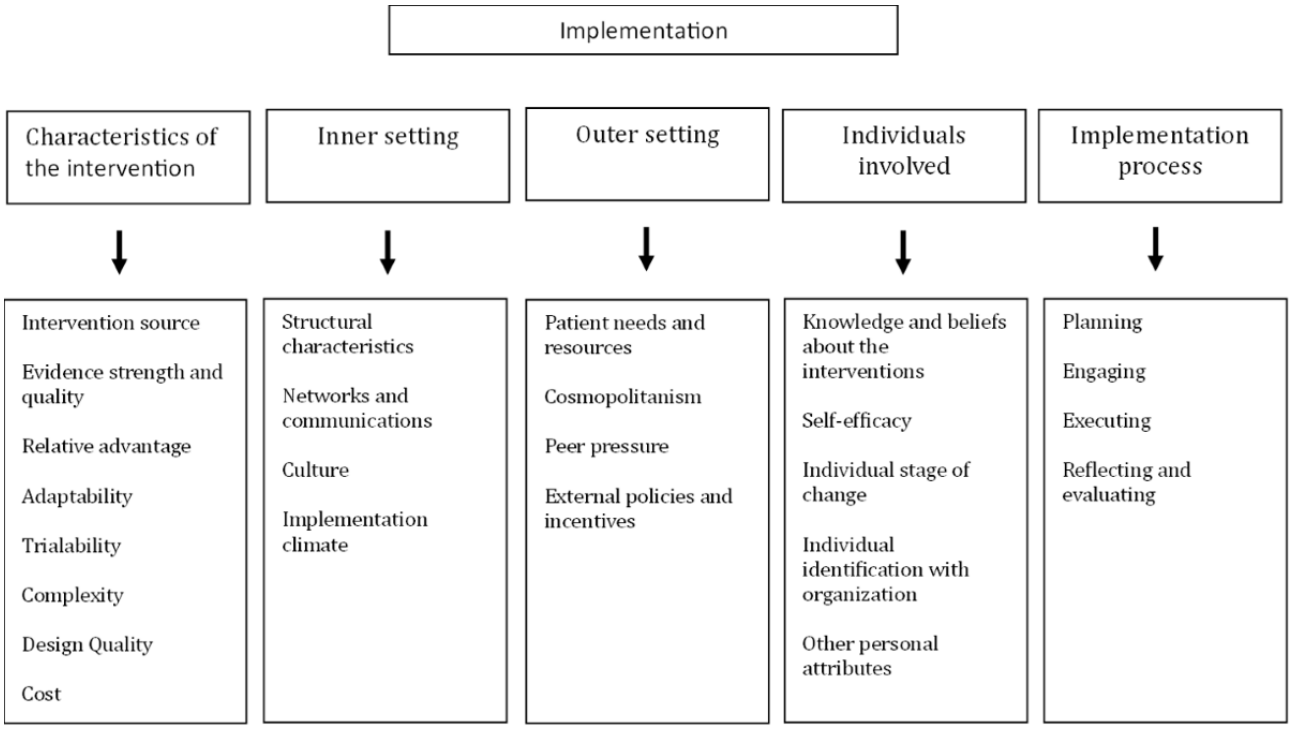

Results

Study Population

In total, 30 children aged between 6 and 14 years old were included and completed the study. The study population mainly consisted of boys (27/30, 90%), admitted at the clinic (18/30, 60%), and diagnosed with a neurodevelopmental disorder (17/30 56.7%). Table 2 summarizes all descriptive characteristics.

XSL•FO
RenderX

Table . Demographic and clinical characteristics of pediatric patients from a psychiatric clinic and daycare unit (n=30).

Characteristics	Results
Age (years), mean (SD)	9.3 (1.95)
Sex (male), n (%)	27 (90)
Clinic (vs daycare), n (%)	18 (60)
Primary diagnosis, n (%)	
Neurodevelopmental disorder	
Autism spectrum disorder (ASD)	11 (37)
Attention deficit hyperactivity disorder	5 (17)
Language disorder	1 (3)
Developmental delay	1 (3)
Trauma- and stressor-related disorders	1 (3)
Obsessive compulsive disorder	1 (3)
Disruptive, impulse-control, and conduct disorders	5 (17)
Other conditions	
Parent-child interaction problems	3 (10)
Sibling Relational Problem	1 (3)
Child neglect	1 (3)

Data Collection

In the clinic, on average, 44 hours of data per child were collected (80% of the aim). In day care treatment, on average, 23 hours of data per child were collected (65% of the aim). Several days show missing data because some of the children did not attend the day care on those days for several reasons. In total, 315 observations of aggressive behavior were reported, including 54 red incidents (from 18 unique participants) and 261 orange incidents (from 26 unique participants).

As described earlier, HR is processed automatically. In general, the measured HR frequencies are within the normal range for children of this age, and no missing data were detected. It was noticed, however, that even when the watch is not on the wrist, the Empatica algorithms determine a (artificially high) HR (Multimedia Appendix 3). This explains the exclusion of 3 participants' outliers. To further assess the reliability of the data, the daily distribution of HR levels for 7 participants was examined. The results showed no significant deviations from the normal range of HR in children, as defined by the Advanced Pediatric Life Support guidelines: 80 - 120 bpm for ages 6 - 12 years and 60 - 100 bpm for ages 12 - 14 years (Figure 2).

Figure 2. The heart rate (HR) distribution of 7 participants on 5 consecutive days for each patient. On the x-axis time is plotted and on the y-axis the HR in beats per minute is plotted. The distribution is plotted in boxplots, showing the minimum score, first quartile, median, third quartile, and maximum score.

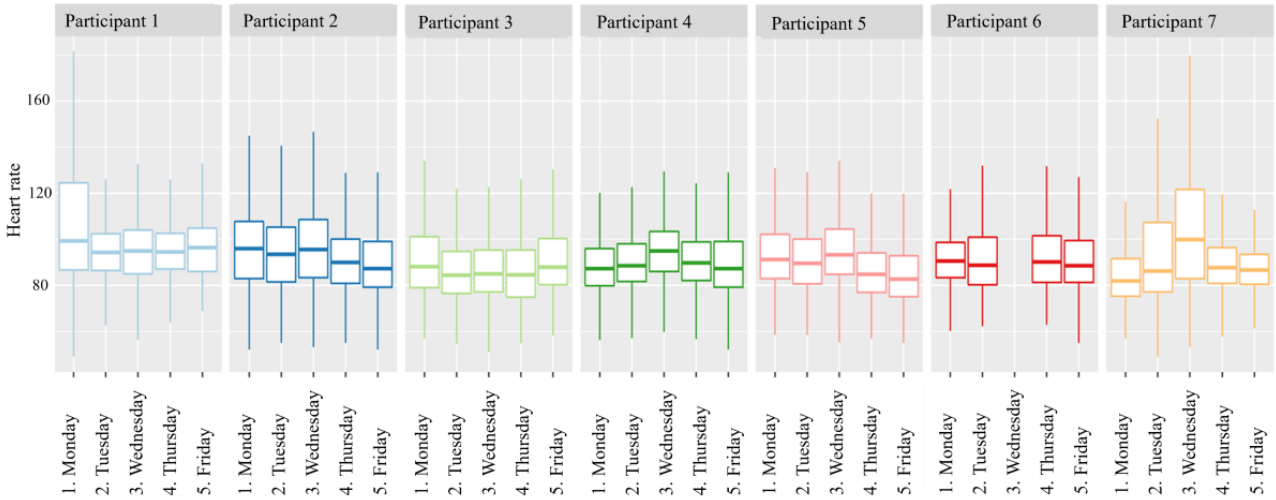

Besides HR, Empatica also provides IBI, which is the distance between 2 peaks. This is only provided in case 2 consecutive

XSL•FO
RenderX

peaks are reliably detected. Only for a small fraction we have valid IBI data, which means that derived measures, like HRV, could not be validly calculated.

For EDA, the Python script "EDA Explorer" was applied to filter the signal ourselves (see Multimedia Appendix 4) and to identify noise. Furthermore, it made visible that most noise is found around the valid signal (see Multimedia Appendix 5). Taking the moving average on the raw signal (as was done in our Dashboard) provides data very similar to the filtered signal. When excluding noise from the EDA data, we distinguish long moments of noise (no valid data for 10 min or the watch was probably not on the wrist) from artifacts in the signal. On average, we collected 39.3 hours of data per participant. After removing long episodes of noise, on average, 34.8 hours of EDA data (−11%) were captured (see Multimedia Appendix 6). After removing artifacts, on average 26.1 hours of EDA data per participant were captured (−25%).

User Experience

Overall, the children were willing to wear the smartwatch. Only 2 of the 34 recruited children did not want to participate beforehand (while parents gave permission) and 2 parents declined to participate. In Multimedia Appendix 1, a summary of the results from the qualitative data of the clinical staff is presented. They reported that the majority of the children were enthusiastic and felt proud wearing the device. On the other hand, children were critical regarding the convenience of wearing the device. According to the clinical staff, most of the children wore the wearable in a reliable way; some older children were playing with it (turning it on and off or removed it when irritated). They reported unanimously that the behavior of the children and clinical staff was not influenced by wearing the device. With regard to integration in clinical practice, the clinical staff reported that assisting the children with the wearables did not require much additional time and could be easily incorporated into the daily routine. Preferably, the observation of behavior is done by others (ie, research interns), as writing down behavioral observations with the exact time point of the incident or event when also helping the child at the same time is challenging. All clinical staff saw future possibilities of using wearables in clinical practice, mainly in prevention by getting a signal when arousal is rising or use it to get more insight in the arousal of the child. Monitoring sleep was mentioned as a third possible feature. However, ethical considerations were issued as well ("Should we always monitoring the patient?" and "Is it always helpful to know moods and intervene?"). The dashboard was found of additional value to get more insight in what is happening "inside" the child and to use as a starting point to have a conversation especially with children who find it difficult to express themselves. During the study, clinical staff and participants (parents and child) themselves were mostly interested in the visualization of the psychophysiological data on the dashboard. Furthermore, 2 examples are presented below showing the dashboard and its potential use in clinical practice (names have been changed to ensure anonymity).

Frank, 10 Years Old

This boy struggled with aggression regulation problems and was previously diagnosed with an autism spectrum disorder. During the data collection week, several red and orange behavioral incidents were reported. Figure 3 shows the data visualization dashboard for the first day when he wore the Empatica E4. Notably, of the 5 orange incidents reported that day, some were associated with a peak in SCL, while others showed no significant change in SCL. The 2 incidents in the morning took place at school. The teacher reported that the boy was removed from class due to misbehavior. Interestingly, no significant peaks in SCL or HR were observed in the dashboard during these 2 incidents. The orange incident around 12:00 noon showed an increase in movement and HR, along with a slight rise in SCL, and occurred while walking from clinic to school. Around 18:30, a substantial peak was observed across different measurements, although no incident or activity was reported at that time.

XSL•FO
RenderX

Figure 3. The psychophysiological data of participant X during day 1. On the x-axis, time is plotted. On the y-axis, different psychophysiological parameters are plotted: movement in *g*, skin temperature in Celsius degrees, skin conductance level (SCL) in micro Siemens, heart rate (HR) in beats per minute and heart rate variability (HRV) in root mean square of successive differences. A higher SCL or HR gives an indication of possible arousal, whereas a lower level of HRV indicates a possible level of arousal.

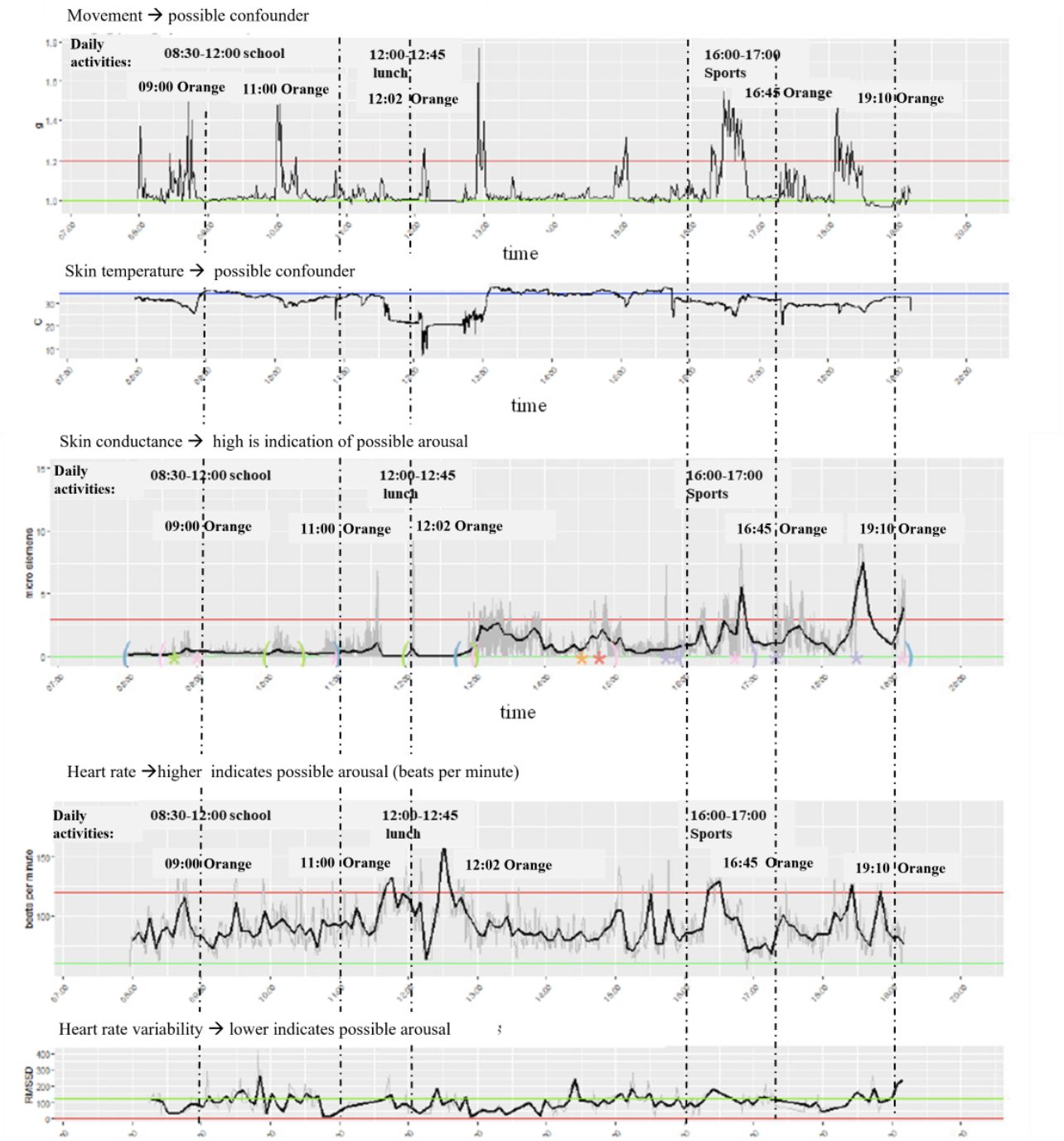

Peter, 12 Years Old

Peter struggled with several behavioral issues. The main problem was compulsive behavior, and among other things, not speaking and eating; leading to many stressful moments during the day, which could turn into orange or red behavioral incidents (see Methods section). Peter wore the Empatica E4 for 5 days, with the visualization of day 3 presented in Figure 4. Interesting patterns were seen on different timeslots. Peaks in SCL data were observed around the Eye Movement Desensitization and Reprocessing treatment session conducted that day. The therapist reported increased tension during the session. Peter reported that his "subjective unit of distress" decreased from level 7 to level 5 during that session. In addition, an orange incident was reported around 15:00 PM when Peter was informed of a change in his daily routine, requiring him to go to bed early. This change caused significant disappointment and annoyance, potentially explaining the minor SCL peak observed at that time. Furthermore, during physical exercise, an increase in both SCL and movement was noted, underscoring the necessity of accurately interpreting SCL peaks. Such peaks should not be misclassified as stress responses but rather as a result of increased sweating during physical activity. Thus, it is crucial

XSL•FO
RenderX

to consider the context and observe behavior and activities during measurements to be able to interpret the data correctly.

Figure 4. The psychophysiological data of participant Y during day 3. On the x-axis, time is plotted. On the y-axis, different psychophysiological parameters are plotted: movement in *g*, skin temperature in Celsius degrees, skin conductance level (SCL) in micro Siemens, heart rate (HR) in beats per minute and heart rate variability (HRV) in root mean square of successive differences. A higher SCL or HR gives an indication of possible arousal, whereas a lower level of HRV indicates a possible level of arousal.

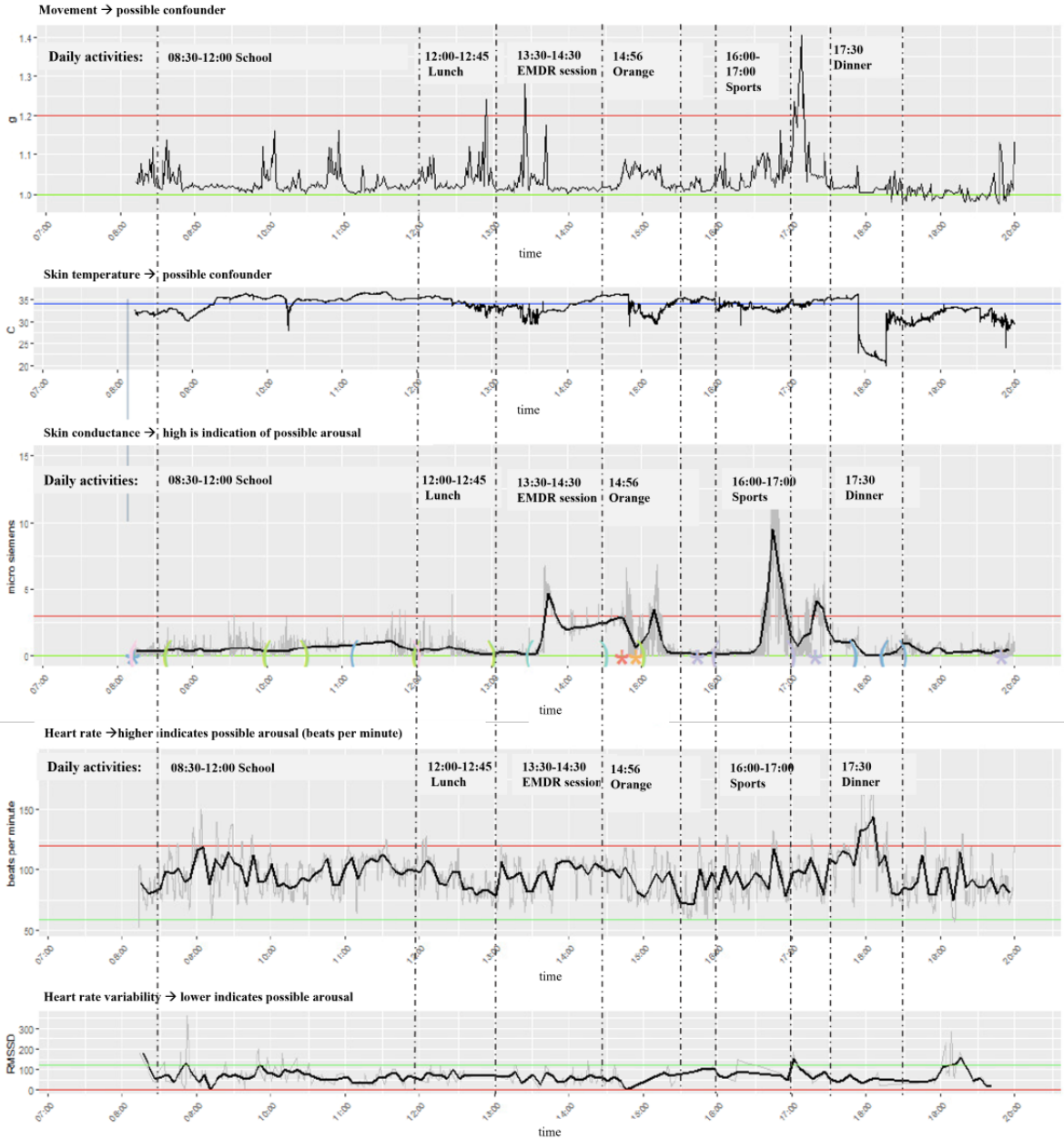

Psychophysiological Measures and Child Aggressive Behavior

An exploratory analysis was performed, in which 4 associations were tested (movement, HR, SCL, and PPM) with paired-sample *t* tests and a Bonferroni adjusted alpha level of .013 per test (.05/4). The HR level of children during aggressive incidents (mean 101.32, SD 19.35) was significantly higher compared to nonaggressive incidents (mean 92.18, SD 8.18; t_{32}=2.866; *P*=.007). For PPM, the conventional threshold for statistical significance was not reached, but the data may suggest a trend toward a higher PPM during aggressive incidents (mean 3.16, SD 2.00) than during nonaggressive incidents (mean 2.59, SD 1.03; t_{32}=1.905; *P*=.07). For SCL and movement, no significant difference was found (*P*=.33 and *P*=.60).

Barriers and Facilitators for Implementation in Clinical Care

The evaluation of the implementation process depicts an intensive and incremental implementation journey to investigate the use of the Empatica E4 in clinical care. Subsequently, key activities and the main facilitating and hindering factors for each CFIR domain were analyzed and discussed below (see Figure 5).

XSL•FO
RenderX

Figure 5. The results of the evaluation of the implementation process using the Consolidated Framework for Implementation Research (CFIR) are presented in terms of factors that had a positive (+), mixed (+/-), or negative (-) impact on the implementation process within each CFIR domain.

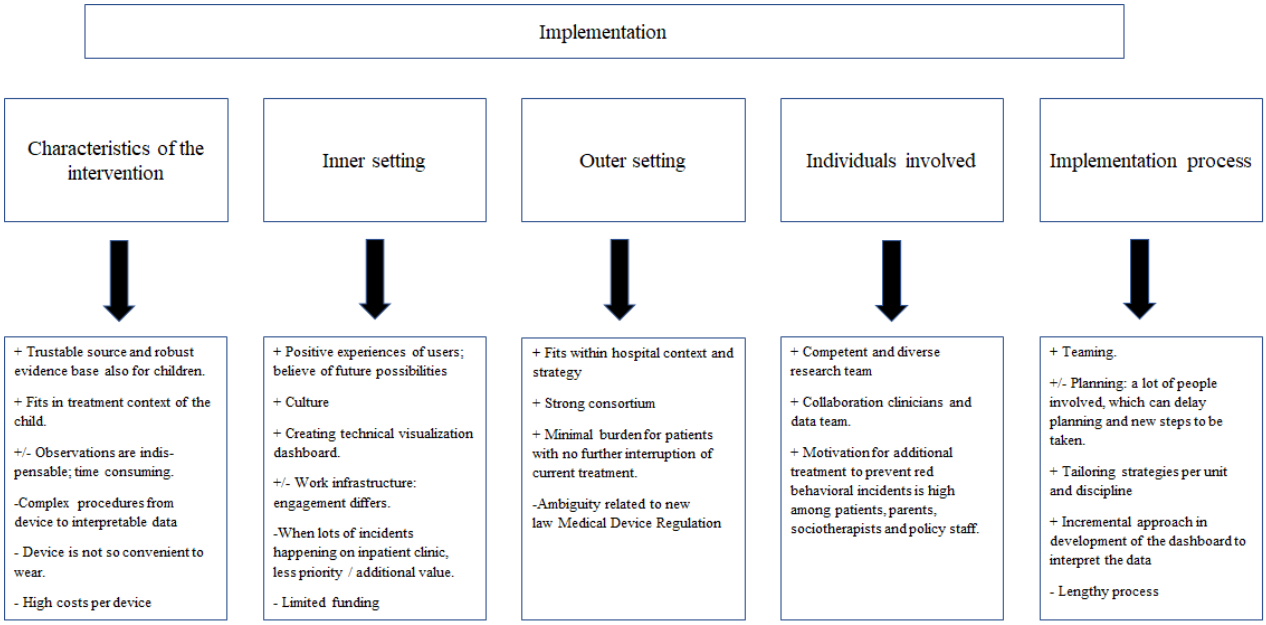

First, the characteristics of the intervention: the intervention encompasses the entire process, involving the wearable device itself, the data collection, and the presentation of data on the dashboard to clinical staff. Positive aspects of the wearable device include its reliability and strong scientific evidence base. User experiences indicate that the devices were easy to operate, although some children reported discomfort while wearing them. The data processing procedure for generating the dashboard was highly customized, with the customization of the dashboard content being perceived as a positive aspect, providing information that was relevant and comprehensible for health care professionals. Simultaneously, the process from data collection to visualization proved to be complex and time-consuming. Downloading the data from the watch, storing, and processing them in the dashboard took approximately 20 minutes per device per day and required the involvement of specific equipment and research staff. For adopting this intervention in daily routine care, the innovation itself should be simplified.

Second, the inner setting: a significant portion of the implementation activities focused on adequately informing, involving, and motivating all stakeholders for participation and cooperation. The extensive coordination required can be attributed to the fact that the study was conducted across two departments, each with its own treatment team, the multidisciplinary nature of the treatment, and the involvement of the school. It was not always evident that information was shared within a team or discipline, often necessitating individual explanations. Throughout the day, many professionals from different disciplines interacted with the children (eg, nurses, psychiatrists, therapists, psychologists, and teachers). Initially, active participation was requested in recording the red, orange, and green incidents, but this proved incompatible with their daily work. This barrier was overcome by deploying dedicated observers during the days the children wore the device. The large number of stakeholders and the necessary coordination with all the different parties stemmed from a culture of high autonomy within disciplines and limited coordination and interdependence. This can also be seen as a barrier within the inner setting that, due to the substantial implementation effort, only minimally hindered the project.

Third, the outer setting: in general, the outer setting can be regarded as a facilitator in this study, as the intervention aligns effectively with the hospital context and its digital strategy. During the study, a new law on Medical Device Regulation was introduced in the European Union. Although the research team could easily approach the legal department in the hospital which is a facilitator, at that moment there was still a lot of uncertainty regarding the impact of the regulation for our study. Due to the inherent ambiguity associated with new legislation, we benefited less from an otherwise facilitating context which causes delay at the start of this study. Furthermore, the research team was integrated into a comprehensive consortium comprising a diverse array of partners with a shared vision that facilitated the provision of wearable devices and contributed to the development of the intervention.

Fourth, the individuals involved: within the domain of “individuals involved” facilitating factors were dominant. The research team which combined the necessary clinical, data science, and implementation expertise, collaborated effectively. It was of significant added value that 2 research team members were also employed at the units of implementation. As a result, they were familiar with the staff, ensuring short communication lines, and enabling them to accurately assess the impact of the innovation and implementation on daily practice. The role of the management of the unit was supportive but in line with the

XSL•FO
RenderX

organization culture not decisive. Due to the positive attitude of the majority of the clinical staff, children, and parents toward the innovation, this has not resulted in any hindrance.

Fifth and last, the implementation process: during the study, the teaming was positive, and the absence of a dedicated project leader was compensated for by a clear distribution of tasks. In this project, we offered tailored implementation strategies based on the needs of the units, and due to short feedback loops, we were able to change strategies if needed; for example, the observation that was done first by clinical staff, but since they experienced a lack of time, this was done additionally by research staff in a later phase. The planning could be seen as a barrier. Due to multiple people involved, the 2 different units, and the informal project structure, the time between data collection was not always effectively planned.

In summary, factors emerging from the domain "characteristics of the intervention" and "inner setting" had the most significant influence on the implementation process. Facilitators for implementation are a positive culture for change, motivation of end users, fruitful collaboration of different disciplines, and detailed results of using wearables on qualitative and quantitative levels. Barriers that have to be kept in mind are the type and design of the wearable and the process for data processing, workload, funding, and further research in how to interpret psychophysiological data.

Discussion

Principal Findings

This feasibility study examined both the use and implementation of smartwatch-based wearables in a children's psychiatric center to help children, parents, and staff to better understand children's behavior.

Using a mixed methods design, the qualitative and quantitative data from this study demonstrate the feasibility of collecting data from this group of children in their natural, yet clinical environment using smartwatch-based wearables. Only by integrating the qualitative and quantitative data, it was possible to thoroughly investigate all aspects of the actual feasibility in clinical care.

With a few exceptions, the children were able to wear the wearable for several days in a row and the data could be collected and processed in an insightful dashboard. No major errors were detected in the psychophysiological data, as collected by the wearables.

Visualizing the data allowed health care professionals to interpret the data and encouraged a more positive attitude toward the use of wearables in clinical care. The results of the single-case explorative visualizations also showed that interpretation of the psychophysiological data is not straightforward and ask for interpretation by the staff based on detailed data on events during the day. In summary, the clinical staff saw opportunities to use psychophysiological data to better understand children's behavior.

With respect to the second part on the implementation process, this study demonstrates that using wearables in this natural setting required substantial additional effort. The most challenging CFIR domains were "characteristics of the intervention" and "inner setting," reflected in the fact that that the majority of implementation activities focused on these 2 domains. Considerable effort was invested in designing the intervention to ensure clinical usability and interpretable results. In particular, the data processing from collection to visualization of the psychophysiological data in the dashboard was complex and time-consuming. Hopefully, this process will be automated in the future, to overcome this barrier (eg, [23]). In addition, extensive coordination was necessary with the staff, not only to inform and involve them but also in ensuring the children to wear the wearable. A thorough and customized implementation strategy has ensured successful navigation of these challenges and resulted in a positive attitude toward the innovation. Many of these barriers and facilitators are likely to be generalizable to similar contexts.

Comparison With Previous Work

Previous research by de Looff et al [4] demonstrated that wearables could be used with adults exhibiting disruptive behavior in a clinical setting. In this study, where we applied the data processing and analysis methods of de Looff et al [4], we showed that it is also possible to measure psychophysiology in hospitalized children with disruptive behavior using wearables.

Even more in line with our study, a small feasibility study (n=10) in a similar group of hospitalized children with disruptive behavior also showed a high adherence rate of using smartwatches in clinical practice [27]. Using other techniques to analyze the data, like the machine learning approach in this study to predict and learn more of the underlying biomarkers of impending disruptive behavior, seems promising and might be the next step to explore in our dataset.

In a scoping review on the use of mobile and wearable artificial intelligence in child and adolescent psychiatry report on 19 articles of which a few also used EDA and HR to predict behavior of children with neurodevelopmental disorders [28]. In addition, 4 out of 19 used similar biosensors and psychophysiological data comparable to our study [9,29-31]. With this study, we address the call in the scoping review of Welch et al [28] for annotated data derived from a naturalistic setting. Our findings also clearly highlighted the challenges and additional effort required to integrate this type of intervention into daily clinical care. This aligns with the challenges mentioned in literature regarding overcoming the gap between the first valley of death and the second valley of death [14]. Bridging the initial gap between evidence-based research in the laboratory and clinical practice is one challenge, but implementing these findings into daily care presents a second gap with additional barriers to overcome.

Limitations

Several limitations need to be addressed. First, it is challenging to ascertain the clinical relevance of increases in EDA or HR, and to establish baseline levels for individual children, especially for EDA. Second, only for a small fraction do we have valid IBI data, which means that derived measures, such as HRV,

XSL•FO
RenderX

could not be validly calculated. This may be due to Empatica's sensitivity to motion and motion artifacts which affect IBI measurements as reported by Schuurmans et al [32], and that especially affects measurements in young active children. Third, we transitioned from clinical staff to research staff for conducting behavioral observations due to the high time demands. This change did not affect the accuracy of timestamping, indicating that even during the brief period during which clinical staff primarily conducted the observations, they were performed with precision. However, this experience indicates that it will be challenging for clinical teams to maintain this level of time-stamping accuracy as part of their regular daily routine once wearables are integrated into daily practice. Finally, we did not use CFIR in the most rigorous manner, since data collection was not prospectively based on CFIR in qualitative interviews combined with a quantitative survey. This would be of added value but was not feasible in this real-world clinical setting. For this reason, we retrospectively evaluated the implementation process based on project documentation and the experiences of the clinical staff, and the research team identified barriers and facilitators from these sources.

Conclusions

In line with our hypothesis, the results show that it is feasible to use wearables in a children's psychiatric center to collect psychophysiological data in children with disruptive behavior. To interpret the data and make it available for clinical staff, numerous steps are required, ranging from data processing to visualizing the data in a dashboard. To help bridge the second gap by actually implementing the use of wearables in daily clinical care, an intervention should require minimal additional time for health care professionals, is intuitive to use, operates without requiring technical support, and is tailored to the clinical context.

Using a dashboard to visualize psychophysiological data and behavioral observations was reported as added value by clinical staff and parents. To optimize the implementation process in other clinical settings, an open access article and code have been made accessible for other researchers and clinicians in the field (eg, [24]). Finally, technical innovations are evolving quickly over time, which is something to keep in mind in the implementation process in the long term (when focusing on one wearable, the other wearable is already in production).

In summary, more research is needed to explore the potential of this innovation to be truly relevant to clinical practice. With these results, we hope to assist organizations in developing an effective tailored implementation strategy for investigating and using wearables in clinical practice, given the evidence that tailored implementation strategies can be effective [33].

Future research should also focus on potential applications for example to gain insight in one's behavior during therapy, live signaling of increased arousal in different environments, and the added value of virtual reality combined with the feedback by wearables. With the insights from the implementation research, we hope to help accelerate the uptake of such technology in clinical practice.

Acknowledgments

Our most sincere thanks to Ernst Eilering of Custodial Institutions Agency for lending us a set of Empatica E4 for this study. We would like to thank Jort Bastiaansen, Pomme de Vogel, Frederieke de Ridder, Floor Hoff, Natascha den Bleijker, Janine van der Staaij, Marilou Bartels, Ellen Hoekstra, and Laura van der Weijden for their contribution to this study.

Data Availability

The data sets generated during and analyzed during this study are not publicly available as participants did not consent to share the data outside of the research team. The data are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview questions that were used for qualitative data collection, based on the Wearable Computer Rating Scale and the summary of the results of the interviews.
[DOCX File, 30 KB - formative_v9i1e65559_app1.docx]

Multimedia Appendix 2
Overview of raw Empatica data, the applied processing and resulting preprocessed dataset.
[PNG File, 120 KB - formative_v9i1e65559_app2.png]

Multimedia Appendix 3
Example of Empatica data as shown in the dashboard. From 16:00 on the watch is not on the wrist, but the Empatica nevertheless provides heart rate (HR measure; measures of skin temperature and electrodermal activity [EDA] do decline and stay at zero).
[PNG File, 169 KB - formative_v9i1e65559_app3.png]

XSL•FO
RenderX

Multimedia Appendix 4
Electrodermal activity (EDA) signal with noise marked in red. Detected by EDA Explorer.
[PNG File, 82 KB - formative_v9i1e65559_app4.png]

Multimedia Appendix 5
Average amount of electrodermal activity (EDA) data per person, from raw to filtered. We distinguish between prolonged noise (>10 min), occurring when the watch was not worn on the wrist, and artifacts in the signal.
[PNG File, 297 KB - formative_v9i1e65559_app5.png]

Multimedia Appendix 6
Electrodermal activity (EDA) signal. First plot: EDA signal including noise; second plot: noise free EDA signal; third plot: moving average of raw EDA signal.
[PNG File, 329 KB - formative_v9i1e65559_app6.png]

Checklist 1
Good Reporting of A Mixed Methods Study (GRAMMS) checklist.
[DOCX File, 20 KB - formative_v9i1e65559_app7.docx]

References

1. Torous J, Bucci S, Bell IH, et al. The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. World Psychiatry 2021 Oct;20(3):318-335. [doi: 10.1002/wps.20883] [Medline: 34505369]
2. de Looff PC, Cornet LJM, de Kogel CH, et al. Heart rate and skin conductance associations with physical aggression, psychopathy, antisocial personality disorder and conduct disorder: An updated meta-analysis. Neurosci Biobehav Rev 2022 Jan;132:553-582. [doi: 10.1016/j.neubiorev.2021.11.003] [Medline: 34774587]
3. Poh MZ, Swenson NC, Picard RW. A wearable sensor for unobtrusive, long-term assessment of electrodermal activity. IEEE Trans Biomed Eng 2010 May;57(5):1243-1252. [doi: 10.1109/TBME.2009.2038487] [Medline: 20172811]
4. de Looff P, Noordzij ML, Moerbeek M, Nijman H, Didden R, Embregts P. Changes in heart rate and skin conductance in the 30 min preceding aggressive behavior. Psychophysiology 2019 Oct;56(10):e13420. [doi: 10.1111/psyp.13420] [Medline: 31184379]
5. Ter Harmsel A, van der Pol T, Swinkels L, Goudriaan AE, Popma A, Noordzij ML. Development of a wearable biocueing app (Sense-IT) among forensic psychiatric outpatients with aggressive behavior: design and evaluation study. JMIR Form Res 2021 Nov 24;5(11):e29267. [doi: 10.2196/29267] [Medline: 34821567]
6. Kuijpers E, Nijman H, Bongers IMB, Lubberding M, Ouwerkerk M. Can mobile skin conductance assessments be helpful in signalling imminent inpatient aggression? Acta Neuropsychiatr 2012 Feb;24(1):56-59. [doi: 10.1111/j.1601-5215.2011.00582.x] [Medline: 25288460]
7. Ter Harmsel JF, Noordzij ML, van der Pol TM, Swinkels LTA, Goudriaan AE, Popma A. Exploring the effects of a wearable biocueing app (Sense-IT) as an addition to aggression regulation therapy in forensic psychiatric outpatients. Front Psychol 2023;14:983286. [doi: 10.3389/fpsyg.2023.983286] [Medline: 36968738]
8. Cabibihan JJ, Javed H, Aldosari M, Frazier TW, Elbashir H. Sensing technologies for autism spectrum disorder screening and intervention. Sensors (Basel) 2016 Dec 27;17(1):46. [doi: 10.3390/s17010046] [Medline: 28036004]
9. Goodwin MS, Mazefsky CA, Ioannidis S, Erdogmus D, Siegel M. Predicting aggression to others in youth with autism using a wearable biosensor. Autism Res 2019 Aug;12(8):1286-1296. [doi: 10.1002/aur.2151] [Medline: 31225952]
10. Nuske HJ, Finkel E, Hedley D, et al. Heart rate increase predicts challenging behavior episodes in preschoolers with autism. Stress 2019 May;22(3):303-311. [doi: 10.1080/10253890.2019.1572744] [Medline: 30822219]
11. Palix J, Akselrod M, Cungi C, Giuliani F, Favrod J. Changes in heart rate variability recorded in natural situation with t-shirt integrated sensors and level of observed behavioral excitation: a pilot study of patients with intellectual disabilities and psychiatric disorders. Front Psychiatry 2017;8:4. [doi: 10.3389/fpsyt.2017.00004] [Medline: 28203208]
12. Graham AK, Lattie EG, Powell BJ, et al. Implementation strategies for digital mental health interventions in health care settings. Am Psychol 2020 Nov;75(8):1080-1092. [doi: 10.1037/amp0000686] [Medline: 33252946]
13. Fernandez ME, Ten Hoor GA, van Lieshout S, et al. Implementation mapping: using intervention mapping to develop implementation strategies. Front Public Health 2019;7:158. [doi: 10.3389/fpubh.2019.00158] [Medline: 31275915]
14. Fernandez-Moure JS. Lost in translation: the gap in scientific advancements and clinical application. Front Bioeng Biotechnol 2016;4:43. [doi: 10.3389/fbioe.2016.00043] [Medline: 27376058]
15. van Gemert-Pijnen JL. Implementation of health technology: directions for research and practice. Front Digit Health 2022;4:1030194. [doi: 10.3389/fdgth.2022.1030194] [Medline: 36440459]
16. Lewy H. Wearable technologies - future challenges for implementation in healthcare services. Healthc Technol Lett 2015 Feb;2(1):2-5. [doi: 10.1049/htl.2014.0104] [Medline: 26609396]

XSL•FO
RenderX

17. Zbukvic I, Rheinberger D, Rosebrock H, et al. Developing a tailored implementation action plan for a suicide prevention clinical intervention in an Australian mental health service: a qualitative study using the EPIS framework. Implement Res Pract 2022;3:26334895211065786. [doi: 10.1177/26334895211065786] [Medline: 37091106]
18. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. Implementation Sci 2009 Dec;4(1):50. [doi: 10.1186/1748-5908-4-50]
19. Breimaier HE, Heckemann B, Halfens RJG, Lohrmann C. The consolidated framework for implementation research (CFIR): a useful theoretical framework for guiding and evaluating a guideline implementation process in a hospital-based nursing practice. BMC Nurs 2015;14(1):43. [doi: 10.1186/s12912-015-0088-4] [Medline: 26269693]
20. Keith RE, Crosson JC, O'Malley AS, Cromp D, Taylor EF. Using the Consolidated Framework for Implementation Research (CFIR) to produce actionable findings: a rapid-cycle evaluation approach to improving implementation. Implement Sci 2017 Feb 10;12(1):15. [doi: 10.1186/s13012-017-0550-7] [Medline: 28187747]
21. Curran GM, Bauer M, Mittman B, Pyne JM, Stetler C. Effectiveness-implementation hybrid designs: combining elements of clinical effectiveness and implementation research to enhance public health impact. Med Care 2012 Mar;50(3):217-226. [doi: 10.1097/MLR.0b013e3182408812] [Medline: 22310560]
22. O'Cathain A, Murphy E, Nicholl J. The quality of mixed methods studies in health services research. J Health Serv Res Policy 2008 Apr;13(2):92-98. [doi: 10.1258/jhsrp.2007.007074] [Medline: 18416914]
23. Garbarino M, Lai M, Tognetti S, Picard R, Bender D. Empatica E3 - a wearable wireless multi-sensor device for real-time computerized biofeedback and data acquisition. Presented at: 4th International Conference on Wireless Mobile Communication and Healthcare - "Transforming healthcare through innovations in mobile and wireless technologies"; Nov 3-5, 2014; Athens, Greece URL: http://eudl.eu/proceedings/MOBIHEALTH/2014 [doi: 10.4108/icst.mobihealth.2014.257418]
24. de Looff P, Duursma R, Noordzij M, et al. Wearables: an R package with accompanying shiny application for signal analysis of a wearable device targeted at clinicians and researchers. Front Behav Neurosci 2022;16:856544. [doi: 10.3389/fnbeh.2022.856544] [Medline: 35813597]
25. Knight JF, Baber C. A tool to assess the comfort of wearable computers. Hum Factors 2005;47(1):77-91. [doi: 10.1518/0018720053653875] [Medline: 15960088]
26. Kleiman E, Millner AJ, Joyce VW, Nash CC, Buonopane RJ, Nock MK. Using wearable physiological monitors with suicidal adolescent inpatients: feasibility and acceptability study. JMIR Mhealth Uhealth 2019 Sep 24;0:e13725. [doi: 10.2196/13725] [Medline: 31586364]
27. Romanowicz M, Croarkin KS, Elmaghraby R, et al. Machine learning identifies smartwatch-based physiological biomarker for predicting disruptive behavior in children: a feasibility study. J Child Adolesc Psychopharmacol 2023 Nov;33(9):387-392. [doi: 10.1089/cap.2023.0038] [Medline: 37966360]
28. Welch V, Wy TJ, Ligezka A, et al. Use of mobile and wearable artificial intelligence in child and adolescent psychiatry: scoping review. J Med Internet Res 2022 Mar 14;24(3):e33560. [doi: 10.2196/33560] [Medline: 35285812]
29. Gay V, Leijdekkers P. Design of emotion-aware mobile apps for autistic children. Health Technol 2014 May;4(1):21-26. [doi: 10.1007/s12553-013-0066-3]
30. Krupa N, Anantharam K, Sanker M, Datta S, Sagar JV. Recognition of emotions in autistic children using physiological signals. Health Technol 2016 Jul;6(2):137-147. [doi: 10.1007/s12553-016-0129-3]
31. Pfeiffer B, Stein Duker L, Murphy A, Shui C. Effectiveness of noise-attenuating headphones on physiological responses for children with autism spectrum disorders. Front Integr Neurosci 2019;13:65. [doi: 10.3389/fnint.2019.00065] [Medline: 31798424]
32. Schuurmans AAT, de Looff P, Nijhof KS, et al. Validity of the Empatica E4 wristband to measure heart rate variability (HRV) parameters: a comparison to electrocardiography (ECG). J Med Syst 2020 Sep 23;44(11):190. [doi: 10.1007/s10916-020-01648-w] [Medline: 32965570]
33. Baker R, Camosso-Stefinovic J, Gillies C, et al. Tailored interventions to address determinants of practice. Cochrane Database Syst Rev 2015 Apr 29;2015(4):CD005470. [doi: 10.1002/14651858.CD005470.pub3] [Medline: 25923419]

Abbreviations

ASD: autism spectrum disorder
CFIR: Consolidated Framework of Implementation Research
EDA: electrodermal activity
GRAMMS: Good Reporting of A Mixed Methods Study
HR: heart rate
HRV: hearth rate variability
IBI: inter-beat interval
PPM: peaks per minute
SCL: skin conductance level
UMCU: University Medical Center in Utrecht

XSL•FO
RenderX

Edited by A Mavragani; submitted 21.10.24; peer-reviewed by D Lindenbach; revised version received 14.02.25; accepted 16.02.25; published 30.05.25.

Please cite as:
Hagoort K, Smeets K, Koldijk S, Scheepers F, Velders F
Collecting Real-Life Psychophysiological Data via Wearables to Better Understand Child Behavior in a Children's Psychiatric Center: Mixed Methods Study on Feasibility and Implementation
JMIR Form Res 2025;9:e65559
URL: https://formative.jmir.org/2025/1/e65559
doi:10.2196/65559

© Karin Hagoort, Kirsten Smeets, Saskia Koldijk, Floortje Scheepers, Fleur Velders. Originally published in JMIR Formative Research (https://formative.jmir.org), 30.5.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Engagement With a Smartphone-Delivered Dietary Education Intervention and Its Relation to Dietary Intake and Cardiometabolic Risk Markers in People With Type 2 Diabetes: Secondary Analysis of a Randomized Controlled Trial

Linnea Sjoblom[1], MSc; Freja Stenbeck[1], MSc; Ylva Trolle Lagerros[2], MD, PhD; Essi Hantikainen[3], PhD; Stephanie E Bonn[1], PhD

[1]Division of Clinical Epidemiology, Department of Medicine, Solna, Karolinska Institutet, T2, Maria Aspmans gata 16, Stockholm, Sweden

[2]Department of Medicine, Huddinge, Karolinska Institutet, Stockholm, Sweden

[3]Institute for Biomedicine, Eurac Research, Bolzano, Italy

Corresponding Author:
Linnea Sjoblom, MSc
Division of Clinical Epidemiology, Department of Medicine, Solna, Karolinska Institutet, T2, Maria Aspmans gata 16, Stockholm, Sweden

Abstract

Background: Mobile health (mHealth) interventions offer a promising way to support healthy lifestyle habits, but effectiveness depends on user engagement. Maintaining high user engagement in app-based interventions is important, yet challenging.

Objective: We aimed to examine the association between user engagement with an app-based dietary education for people with type 2 diabetes and changes in diet quality, dietary intake, and clinical measures.

Methods: In this randomized clinical trial, people with type 2 diabetes were recruited within primary care and randomized 1:1 to a 12-week smartphone-delivered app-based dietary education or control group. Participants were followed up after 3, 6, and 12 months. Dietary intake was assessed using a food frequency questionnaire. The control group received the app at the 3-month follow-up. User engagement was analyzed among all participants. Categories of high (100%), moderate (50% - 99.9%), and low (<50%) user engagement were created based on the percentage of activities completed in the app. We used paired *t* tests to compare mean changes in diet quality, dietary intake, and clinical markers within user engagement groups, and fitted linear regression models to analyze differences in change between groups.

Results: Data from 119 participants (60.5%, 72/119 men) were analyzed. The mean age at baseline was 63.2 (SD 10.3) years and mean BMI was 30.1 (SD 5.1) kg/m^2. User engagement was high with an average of 77.1% of app activities completed. More than half (53.8%, 64/119) of the users showed high user engagement, 21.8% (26/119) moderate, and 24.4% (29/119) low. Directly following the app-based education, a significant difference in change was seen for whole grains (β=20.4, 95%CI 0.57 - 40.3) in participants with high user engagement compared to the low user engagement group who decreased their intake (*P*=.03). At follow-up after 6 to 9 months after completed education, significant differences in change were seen for fiber, wholegrains, carbohydrates, saturated fat, sodium, and total energy in the moderate compared with the low engagement group, and a significant difference in change was seen for carbohydrates in the high, compared with the low, user engagement group.

Conclusions: User engagement was generally high for the smartphone-based dietary education, suggesting that an app targeting dietary habits is feasible to use. Those with higher user engagement seem to maintain healthier dietary behaviours over time, compared to those with low user engagement. Future mHealth interventions should focus on ways to engage those with low interest.

Trial Registration: ClinicalTrials.gov NCT03784612; https://www.clinicaltrials.gov/study/NCT03784612

International Registered Report Identifier (IRRID): RR2-10.2196/24422

(JMIR Form Res 2025;9:e71408) doi:10.2196/71408

KEYWORDS

adherence; dietary change; diabetes mellitus; type 2 diabetes mellitus; healthy diet; mHealth; smartphone app; user engagement; mobile phone

XSL•FO
RenderX

Introduction

Lifestyle interventions, including dietary counseling, are commonly prescribed to people with type 2 diabetes in primary health care, often in combination with medications [1,2]. In recent years, the emergence of new mobile health (mHealth) apps has been proposed as a cost-effective means of delivering nutritional therapy [2-4]. mHealth interventions offer a promising way to support lifestyle changes, but the effect of any intervention is dependent on adherence, ie, the extent to which participants follow an intervention as intended [5-7]. Maintaining user engagement in an app-based intervention is challenging [8,9], yet, important.

Specific intervention features, such as personal tailoring, feedback, and reminders, are associated with higher engagement in digital interventions targeting chronic diseases [10,11]. In a smartphone-delivered intervention targeting healthy lifestyle behaviours, high app user engagement was associated with improvements in dietary quality and reductions in BMI and waist circumference at 1-year follow-up among individuals at increased risk for type 2 diabetes [12]. Furthermore, high user engagement with an app specifically targeting vegetable consumption was also associated with improved vegetable intake among Australian adults [13].

We have previously shown that a smartphone-delivered dietary education had positive effects on dietary fat intake and serum triglyceride levels in persons with type 2 diabetes [14]. In this study, we aimed to examine the association between user engagement with the app-based dietary education and changes in diet quality, dietary intake and clinical variables of BMI, waist circumference, body fat percentage, glycated haemoglobin (HbA_{1c}), serum lipid levels, and blood pressure, directly after completing the education. We also aimed to study the association between user engagement and long-term change in dietary quality and dietary intake 6 to 9 months after the completed education.

Methods

Study Design

We have analyzed data from the HAPPY (Healthy eating using APP technologY) trial, previously described in detail [15]. The main results from the trial have been published [14]. In brief, this was a 2-armed randomized controlled trial comprising women and men with type 2 diabetes that were continuously recruited within 5 primary health care centrs in Stockholm, Sweden. Initial information about the study was provided by health care personnel at their routine health care visit. Patients interested in participating were contacted by study personnel to receive more detailed information. Those who agreed to participate were scheduled for a physical baseline meeting with study personnel. Participants were randomized 1:1 to a 12-week app-based dietary education intervention in addition to regular care, or a control group receiving regular care only. The primary aim of the trial was to examine the effectiveness of the intervention on dietary intake and cardiometabolic risk markers. Study inclusion criteria were: having type 2 diabetes, ≥18 years of age, being able to read and understand Swedish, and having access to and being able to use a smartphone with a personal e-identification. No exclusion criteria were applied. The data collection began in January 2019 and was finalized in August 2023. It was temporary paused during 2020 due to the COVID-19 pandemic. This study is reported according to the CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile HEalth Applications and onLine TeleHealth) checklist, which is developed for eHealth or mHealth interventions (Checklist 1).

Ethical Considerations

The trial was conducted in accordance with the Declaration of Helsinki, and approved by the Ethics Committee of the Regional Ethical Review Board, Stockholm, Sweden (2018/652-31; 2018/1094 - 32; 2018/2393 - 32; 2020 - 00591; 2020 - 07005; and 2022-02557-02). It was also registered at ClinicalTrials.gov (NCT03784612). All study participants received oral and written information about the study before giving their written consent before the study started. The original informed consent allows secondary analysis of data without additional consent. Participants received no compensation for participation in the study and all data were anonymized after data collection.

The intervention group received the app at study start and the control group received the app at the 3-month follow-up. All participants responded to a web-based lifestyle questionnaire at baseline and again after 3 months (directly after completing the dietary education in the intervention group), 6 months (directly after completing the dietary education in the control group), and after 12-months of follow-up. The baseline questionnaire was filled out before participants were informed about their group allocation. We also measured weight (kg), height (cm), waist circumference (cm), body composition, blood pressure, serum levels of glycated haemoglobin (HbA_{1c}), and lipids at baseline, and after 3- and 6 months of follow-up.

Participant Flow and Outcome Assessments

In total, 133 individuals agreed to participate in the trial. Of these, 68 were randomized to the intervention and 65 to the control group. One participant randomized to the intervention group and 3 participants randomized to the control group dropped out before study start, ie, before knowing their group allocation and before providing any baseline data. In total 119 participants, 65 in the intervention group and 54 in the control group, downloaded the app.

In this study, we addressed change in diet quality, dietary intake, and clinical outcomes from baseline to directly after having completed the dietary education, hereafter written as “short-term” follow-up. Furthermore, we investigated the association on diet quality and dietary intake from baseline to 9 months after completing the dietary education for the intervention group and 6 months after completion in the control group, from here on written as “long-term” follow-up.

The HAPPY Trial Intervention and App Content

The smartphone delivered dietary education focused on the overall diet in accordance with current evidence-based guidelines and the Swedish national dietary recommendations

XSL•FO
RenderX

[16]. The app design included the health belief model [17], stages of change model [18], and social cognitive theory [19]. We further included behavior changes techniques such as educational information, goal setting, self-monitoring, feedback, and performance [20]. Participants were encouraged to integrate with the app daily, although it was not a requirement to do so.

The app has six different features, which included (1) educational information, (2) task introduction including a self-set goal of the week, (3) healthy recipes, (4) short fun facts or practical advice, (5) task reminder as a push notification, and (6) task evaluation "how did it go?" with the weekly task. A new topic was introduced each week, for example, "Healthy food patterns" or "Vegetable intake." Figure 1 shows the 12 topics and gives examples of screens (here translated from Swedish to English) from the app. Each week followed the same activity schedule and included either 11 (weeks 3 - 7 and weeks 10 - 12) or 12 (weeks 1, 2, 8, and 9) activities depending on whether there were 4 or 5 recipes. The full 12-week course comprised 136 activities.

Figure 1. Overview of the weekly activity schedule, the topics of the 12 wk education in the randomized controlled HAPPY trial, and some examples of different features of the smartphone application: an educational information text, a weekly task, a recipe and advice.

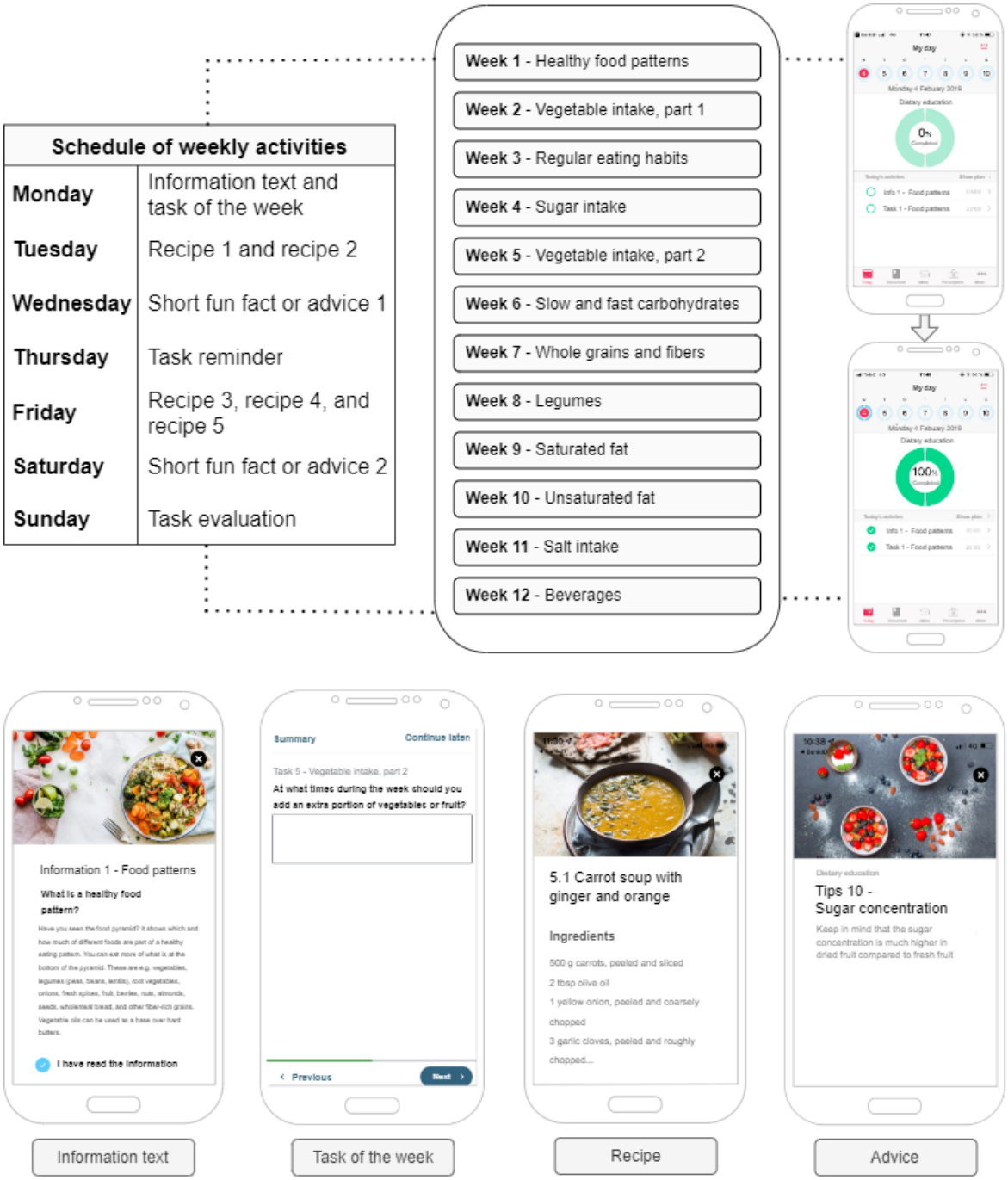

Users were able to follow their progress and track their percentage of completed activities in the app (see Figure 1). Before an activity is completed, 0% is visualized in the green circle, when all activities are completed for the day, the circle fills up and shows 100% completed activities. Uncompleted tasks were moved to the next day until marked as completed

XSL•FO
RenderX

by the user. Information texts, advice, recipes etc, were saved in the app and could be accessed at any time during the study.

Assessment of User Engagement

For the purposes of this study, data on completed activities was extracted from the app. We calculated an overall user engagement score in percent by summarizing all completed activities by a participant divided by the total number of activities. We also calculated the percentage of completed activities within each activity type. Based on the distribution of the overall engagement score, participants were categorized into groups of low (<50%), moderate (50 - 99.9%), and high (100%) user engagement.

Assessment of Dietary Intake

Participants dietary intake were assessed using a validated 95-items semiquantitative food frequency questionnaire (FFQ) [21,22] at baseline and after 3, 6 and 12 months of follow-up. Information on how often each item was consumed (eg once per day, week, or month) was reported by the participants. We obtained standard portions sized from the Swedish Food Composition Database from the National Food Agency [23] to calculate average daily intakes in grams of foods and beverages or grams or micrograms for nutrients.

We assessed overall dietary quality using a Nordic Nutrition Recommendations (NNR)-score based on the food-based dietary guidelines from the latest edition of the NNR 2023 [24]. We have described the NNR-score in detail previously [14]. The NNR-score components and the recommended intakes are presented in Multimedia Appendix 1 (see Table S1). In brief, our NNR-score was calculated based on 10 food based dietary guidelines presented in the NNR including vegetables, fruit and berries, whole grains (cereals), legumes, nuts and seeds, fish and seafood, red meat, vegetable oils, sweets including sugar-sweetened beverages (excluding salty snacks), and alcohol. Proportion scores ranged from 0 - 3 points for each dietary guideline, with adherence complying to the recommendation yielding 3 points. This resulted in a total NNR score from 0 to 30 points. Furthermore, we examined dietary variables corresponding to the weekly topics of the dietary course (Figure 1).

Background Characteristics

Background characteristics included age, sex (female and male), education level (≤12 years and >12 years), diabetes duration (<1 year, 1 - 5 y, and >5 years), smoking status (never or former, and current), self-reported medication for hypertension, hyperlipidaemia, diabetes, and physical activity. Physical activity was assessed by 2 questions asking about time spent exercising and time spent performing daily activities, for example walking, cycling, or gardening, during a typical week [25]. A dichotomous variable (<150 min/week and ≥150 min/week) was created according to current physical activity recommendations [26]. We measured weight (kg), height (cm), and waist circumference (cm) and calculated BMI (kg/m^2). BMI was further categorized as normal weight (<25 kg/m^2), overweight (25.0 - 29.9 kg/m^2), and obesity (≥30.0 kg/m^2) [27]. A dichotomous variable for waist circumference of low risk (<88 cm for women and<102 cm for men) and high risk (≥88 cm for women and ≥102 cm for men) for disease was also created [27]. We included baseline fasting HbA_{1c} (mmol/mol) as well as a dichotomous variable of HbA_{1c} (<52 mmol/mol and ≥52 mmol/mol), which is a general target value in the treatment of type 2 diabetes in Sweden [28].

Statistical Analyses

Descriptive baseline characteristics were presented as mean (SD) continuous variables, as numbers (n), and percentages (%) for categorical variables. User engagement with the smartphone app was assessed by the proportion (%) of activities completed during the intervention period. We conducted analysis of variance to compare differences in continuous variables between user engagement groups, and χ^2 tests for categorical data. Aspects of user engagement were also explored.

Paired *t* tests was used to compare mean changes in diet quality, dietary intake, and clinical markers within user engagement groups. We then fitted linear regression models to analyze differences in changes from baseline to the short-term and from baseline to the long-term follow-up between user engagement groups, with the low user engagement group used as the reference category. Long-term follow-up was not addressed for clinical markers, as these were not measured at the 12-month follow-up. Models were adjusted for baseline values of each dietary variable to account for differences in baseline intakes [29]. Potential confounders were identified using a Directed Acyclic Graph (DAG; see Figure S1 in Multimedia Appendix 1), and we also controlled for age, sex, intervention allocation, education level, BMI, diabetes duration, and HbA_{1c} level. Mean differences in changes for short-term and long-term intakes between groups, represented by β-coefficients with 95% CI, were calculated. Normality of the resulting residuals of each model was visually assessed using histograms with normal curve overlay and Q-Q (quantile-quantile) plots. All were found approximately normally distributed except for the model of sugar-sweetened beverages. This was because almost no one reported intake of sugar-sweetened beverages, which created no variation and thus a violated model assumption. Therefore, we did not run statistical models to analyse difference in change for this variable. Statistical analyses were performed using Stata (version 17.0; Stata Corporation), with the statistical significance level set at $P<.05$.

Results

Figure 2 shows the CONSORT flow diagram of study participants through the trial. The baseline characteristics among all (n=119), and stratified by low (n=29), moderate (n=26), and high (n=64) app user engagement are shown in Table 1. The majority of study participants (60.5%, 72/119) were men, and the average age was 63.2 (SD 10.3) years. The mean BMI was 30.1 (SD 5.1) kg/m^2, and most participants (61.2%, 71/116) had more than 12 years of education. Participants generally reported high levels of physical activity and 76.9% (90/112) met the recommended level of 150 min/week. Only 4/116 (3.5%) were current smokers. The mean HbA_{1c} level was 49.7 (SD 10.6) mmol/mol. Participants in the high user engagement group had

XSL•FO
RenderX

a significantly higher NNR adherence score (mean 14.0, SD 3.4; *P*=.03) compared with low and moderate adherers (mean 12.2, SD 3.4 and mean 12.6, SD 2.8; respectively). Similarly, a statistically higher intake in fruit and vegetables (*P*=.04) was also seen in the high comparted with the other groups, while the moderate group had statistically higher intake of fish and seafood (*P*=.03) compared with the low and high adherers. There were no other statistically significant differences in baseline characteristics between user engagement groups.

Figure 2. Flow diagram of persons with type 2 diabetes recruited from five primary care centers in Stockholm, Sweden, included in the randomized controlled HAPPY trial from recruitment and randomization to baseline, 3, 6, and 12 months follow-up. FFQ: Food Frequency Questionnaire; HAPPY: Healthy eating using APP technologY.

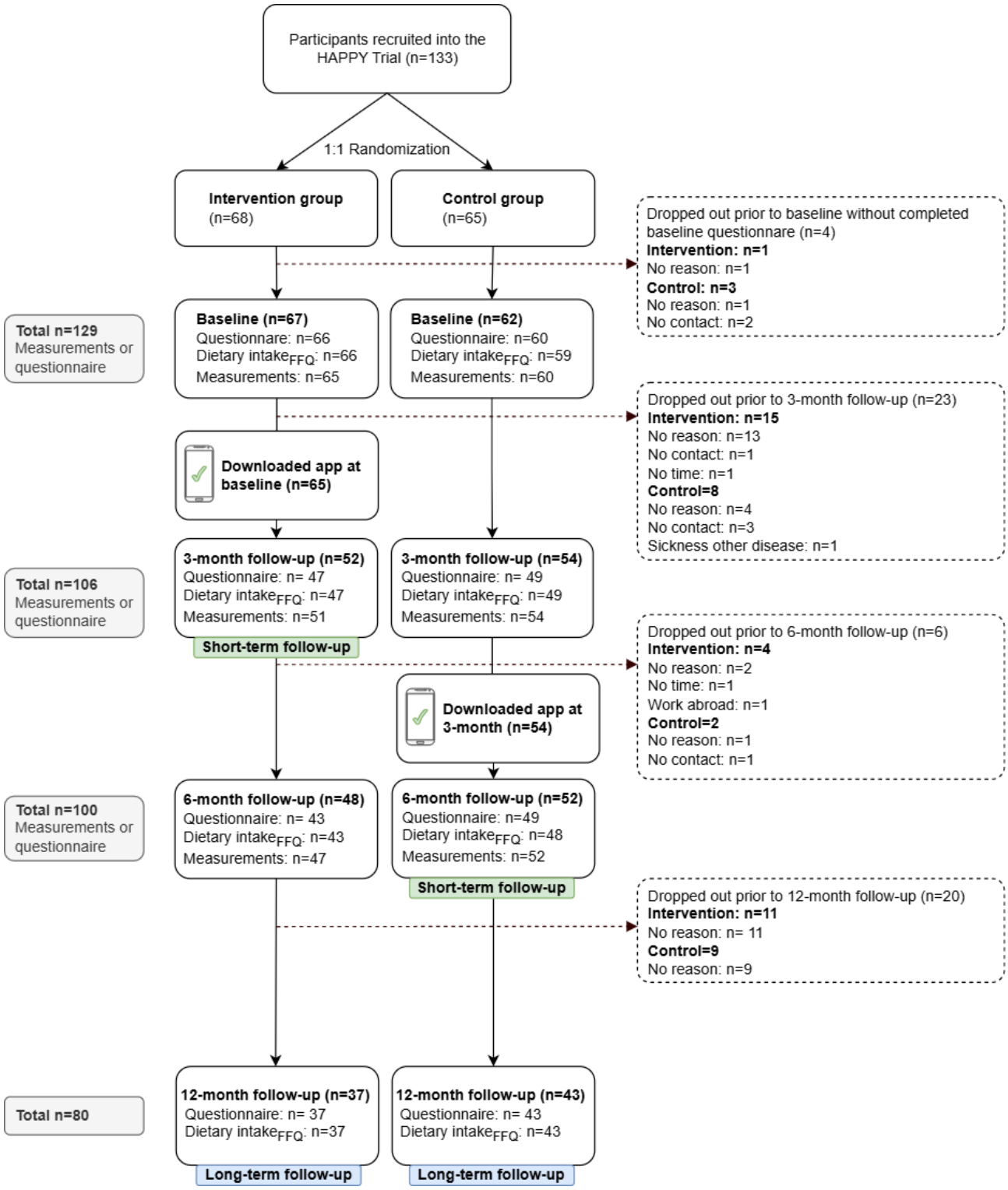

XSL•FO
RenderX

Table . Baseline characteristics of study participants from the randomized controlled Healthy eating using APP technologY (HAPPY) trial that downloaded the app, all and by category of user engagement.

Variable	All (n=119)	User engagement[a]			*P* value[b]
		Low (n=29)	Moderate (n=26)	High (n=64)	
Characteristics, mean (SD)					
Age (years; n=119)	63.2 (10.3)	62.0 (9.5)	62.7 (11.4)	63.9 (10.2)	.37
BMI (kg/m^2; n=119)	30.1 (5.1)	30.5 (5.9)	31.1 (6.1)	29.4 (4.1)	.31
HbA_{1c} (mmol/mol; n=110)	49.7 (10.6)	51.4 (12.5)	50.0 (10.1)	49.0 (10.1)	.66
Physical activity (min/week; n=117)	278.1 (150.9)	275.6 (158.5)	266.0 (152.1)	284.1 (149.4)	.87
Dietary variables (n=117), mean (SD)					
NNR[c] score (0 - 30 points)	13.3 (3.3)	12.2 (3.4)	12.6 (2.8)	14.0 (3.4)	.03
Fruit and vegetables (g/day)	356.3 (236.1)	261.2 (127.8)	356.7 (198.6)	396.2 (273.4)	.04
Legumes/pulses (g/day)	36.4 (36.4)	30.5 (28.6)	30.1 (22.5)	41.5 (43.0)	.25
Total fish and seafood (g/week)	330.9 (201.3)	252.5 (202.2)	393.0 (203.2)	338.7 (192.5)	.03
Red and processed meat (g/week)	592.4 (428.0)	502.1 (317.5)	659.1 (329.3)	603.4 (497.2)	.40
Sugar-sweetened beverages (g/day)	8.8 (50.0)	15.3 (55.2)	23.9 (89.2)	0.0 (0.0)	.09
Fiber (g/day)	27.6 (12.2)	25.4 (10.4)	27.3 (10.6)	28.6 (13.5)	.52
Whole grains (g/day)	62.5 (34.)	60.3 (40.7)	62.4 (33.4)	63.4 (32.8)	.93
Carbohydrates (g/day)	204.1 (75.0)	202.5 (75.0)	210.6 (79.1)	202.1 (74.4)	.88
Saturated fat (g/day)	30.9 (12.2)	29.3 (16.0)	35.2 (8.4)	29.8 (11.5)	.12
Unsaturated fat (g/day)	46.7 (18.9)	42.1 (20.6)	51.8 (13.6)	46.6 (19.8)	.17
Sodium (mg/day)	2623.1 (877.1)	2428.4 (911.9)	2851.2 (884.9)	2612.5 (851.4)	.21
Sucrose (g/day)	30.2 (15.2)	27.1 (13.9)	32.9 (17.2)	30.5 (14.8)	.39
Total energy (kcal/day)	2038.7 (658.2)	1936.0 (769.8)	2215.5 (589.2)	2010.1 (630.2)	.27
Sex, n (%)					.36
Female	47 (39.5)	10 (34.5)	8 (30.8)	29 (45.3)	
Male	72 (60.5)	19 (65.5)	18 (69.2)	35 (54.7)	
Education level, n (%)					.37
≤12 years	45 (38.8)	8 (29.6)	9 (34.6)	28 (44.4)	
>12 years	71 (61.2)	19 (70.4)	17 (65.4)	35 (55.6)	
BMI category (kg/m^2), n (%)					.51
18.5 - 24.9	19 (16)	2 (6.9)	5 (19.2)	12 (18.8)	
25 - 29.9	46 (38.7)	14 (48.3)	8 (30.8)	24 (37.5)	
≥30	54 (45.4)	13 (44.8)	13 (50)	28 (43.8)	

XSL•FO
RenderX

Variable	All (n=119)	User engagement[a]			*P* value[b]
		Low (n=29)	Moderate (n=26)	High (n=64)	
Waist circumference (cm), n (%)					.16
Low	31 (26.1)	11 (37.9)	4 (15.4)	16 (25)	
High	88 (74)	18 (62.1)	22 (84.6)	48 (75)	
HbA_{1c} (mmol/mol), n (%)					.43
Low, <52	71 (64.6)	15 (68.2)	14 (53.9)	42 (67.4)	
High, ≥52	39 (35.5)	7 (31.8)	12 (46.2)	20 (32.3)	
Current smoker (yes), n (%)	4 (3.5)	1 (3.7)	2 (8)	1 (1.6)	.33
Physical activity (min/week), n (%)					.63
<150	22 (23.1)	8 (29.6)	6 (23.1)	13 (20.3)	
≥150	90 (76.9)	19 (70.4)	20 (76.9)	51 (79.7)	
Diabetes duration, n (%)					.90
<1 year	27 (24.1)	6 (23.1)	5 (20)	16 (26.2)	
1 - 5 years	41 (36.6)	10 (38.5)	11 (44)	20 (32.8)	
>5 years	44 (39.3)	10 (38.5)	9 (36)	25 (41)	
Medical use, yes, n (%)	73 (62.4)	22 (81.5)	14 (53.9)	37 (57.8)	.06
Medication for (yes)[d], n (%)					
Hypertension	80 (68.4)	21 (77.8)	19 (73.1)	40 (62.5)	.30
Diabetes, insulin	21 (18)	7 (25.9)	5 (19.2)	9 (14.1)	.40
Diabetes, Metformin	88 (75.2)	23 (85.2)	16 (61.5)	49 (76.6)	.13
Hyperlipidemia	73 (62.4)	22 (81.5)	14 (53.9)	37 (57.8)	.06

[a]User engagement groups, low <50%, moderate 50 - 99.9%, high 100% of the total number of completed activities in the app.

[b]*P* value for difference between user engagement groups using analysis of variance for continuous variables and chi^2 for categorical variables.

[c]NNR, Nordic Nutrition Recommendations.

[d]self-reported.

Overall, user engagement was high, with participants completing an average of 104.9 (77.1%) of the 136 activities in the app. Participants in the low, moderate and high user engagement groups completed on average 20.6 (15.1%), 121.9 (89.6%), and 136 (100%) of all activities, respectively. Among all participants, 88.2% (105/119) completed every activity during the first week, while 62.2% (74/119) completed every activity during the last week (see Figure 3A) and 71.4% (85/119) completed at least one activity per week throughout the 12 weeks (see Figure 3B).

XSL•FO
RenderX

Figure 3. Proportion of persons with type 2 diabetes in the randomized controlled Healthy Eating using APP technologY (HAPPY) trial completing (1) all of the activities per week during the 12-week education and (2) at least one activity per week among all and by the categories of use engagement.

XSL•FO
RenderX

Per definition, participants in the high user engagement group completed 100% of the activities each of the 12 wk (Figure 3A). Among moderate users, the proportion of participants completing 100% of the weekly activities was 100% (26/26) the first and 38.5% (10/26) the last week (Figure 3A). Nevertheless, during the last week, 80.8% (21/26) had completed at least one activity (Figure 3). Among participants with low user engagement, 51.2% (15/29) completed 100% of the weekly activities and 86.2% (25/29) completed at least one activity during the first week. Usage then rapidly decreased, and from week 6 and onwards, no one completed all activities.

Figure 4 shows the user engagement of the different activity types, that is, app features. The activity most often completed during the full 12 weeks was the short fun fact or advice (on average 78.6%, 93.5/119) of participants completed this activity every week), followed by educational information (78.2%, 93/119), the weekly task (77.8%, 92.6/119), the reminder of weekly task (76.2%, 90.7/119), recipes (75.4%, 89.4/119), and the evaluation question linked to the weekly task (74.2%, 88.3/119).

Figure 4. Proportion of persons with type 2 diabetes in the randomized controlled Healthy Eating using APP technologY (HAPPY) trial with completed activities, ie, active users, separated by the 6 different activity types among all participants throughout the 12 weeks.

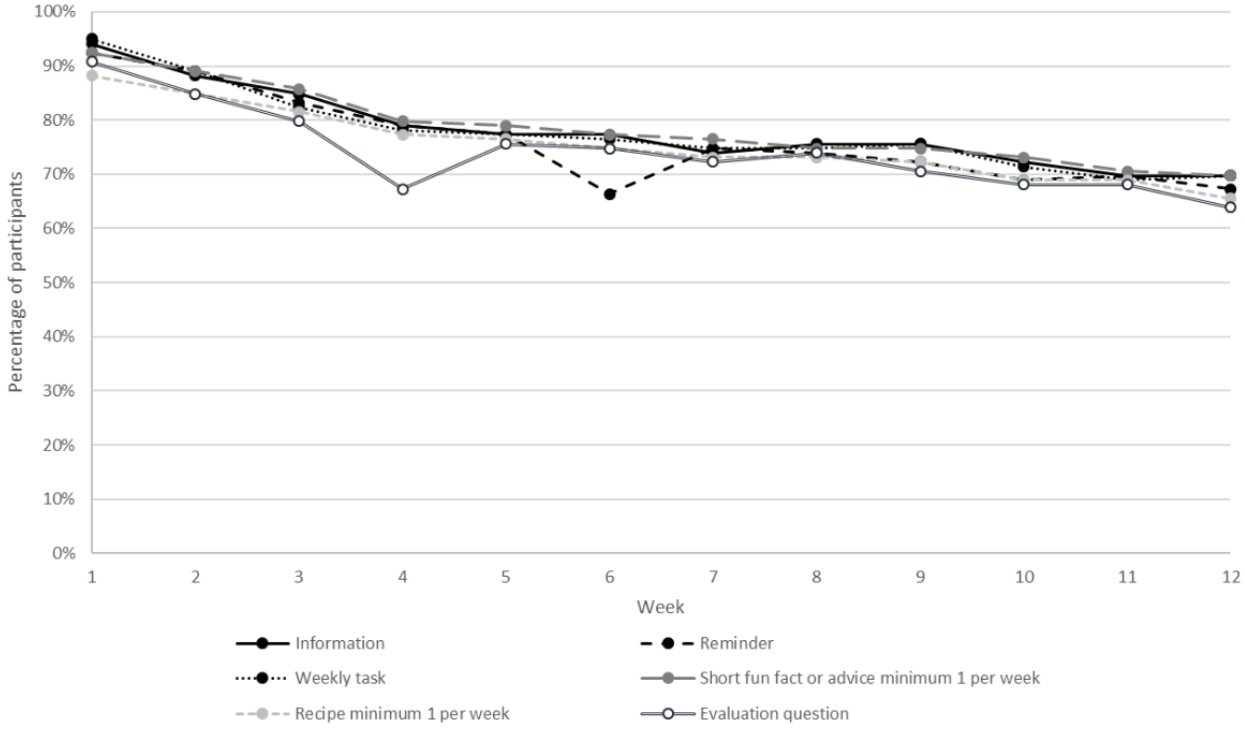

Table 2 presents the comparison of mean dietary intake levels within groups at baseline and the short-term follow-up, that is, directly after completed education, as well as results from linear regression models. The low user engagement group had a statistically significant decrease in whole grains (*P*=.03). Within the moderate user engagement group, a significant increase was seen in the NNR score (*P*=.03), and for fruit and vegetables (*P*=.04) and legumes (*P*=.04).

XSL•FO
RenderX

Table . Mean values and differences of short-term effects (directly after the 12-week dietary education) stratified by low, moderate and high app user engagement and linear regression models presented as difference in changes (β-coefficients and 95% CI) between the user engagement groups in the randomized controlled Healthy eating using APP technologY (HAPPY) trial. User engagement groups, low <50%, moderate 50% - 99%, high 100% of the total number of completed activities in the app.

Variable	Baseline		Short-term follow-up		Within change	Mean difference		Crude[a] Model estimates		Adjusted[b] Model estimates	
	Mean	(SD)	Mean	(SD)	*P* value[c]	Mean	(SD)	β	(95% CI)	β	(95% CI)
NNR[d] (0 - 30 points)											
Low user engagement	13.2	(2.3)	13.2	(2.6)	1.00	0.0	(2.7)	1.00	(reference)	1.00	(reference)
Moderate user engagement	12.2	(2.9)	13.7	(4.2)	.03	1.5	(3.1)	1.3	(-0.8 to 3.3)	1.4	(-0.7 to 3.8)
High user engagement	14.0	(3.3)	14.4	(3.6)	.22	0.5	(2.8)	0.6	(-1.2 to 2.5)	0.3	(-1.7 to 2.3)
Dietary variables (FFQ)[e]											
Fruit and vegetables (g/day)											
Low user engagement	297.8	(121.2)	279.5	(117.9)	.46	-18.2	(77.6)	1.00	(reference)	1.00	(reference)
Moderate user engagement	313.5	(155.7)	373.2	(189.0)	.04	59.7	(126.3)	84.2	(-11.7 to 180.2)	80.0	(-26.2 to 168.2)
High user engagement	391.2	(270.9)	402.5	(202.6)	.62	11.3	(180.5)	67.0	(-18.6 to 152.7)	31.1	(-60.1 to 122.4)
Legumes and pulses (g/day)											
Low user engagement	31.1	(33.2)	31.9	(35.0)	.79	0.8	(9.8)	1.00	(reference)	1.00	(reference)
Moderate user engagement	25.5	(19.4)	38.0	(35.3)	.04	12.5	(26.9)	8.9	(-19.9 to 37.7)	4.9	(-25.9 to 35.6)
High user engagement	42.4	(43.4)	46.8	(47.4)	.49	4.4	(50.4)	9.1	(-16.6 to 34.7)	9.3	(-19.3 to 38.0)
Total fish and seafood (g/week)											
Low user engagement	291.9	(205.8)	276.3	(147.7)	.76	-15.6	(163.6)	1.00	(reference)	1.00	(reference)

XSL•FO
RenderX

Variable	Baseline		Short-term follow-up		Within change	Mean difference		Crude[a] Model estimates		Adjusted[b] Model estimates	
	Mean	(SD)	Mean	(SD)	*P* value[c]	Mean	(SD)	β	(95% CI)	β	(95% CI)
Moderate user engagement	379.9	(210.5)	327.3	(160.0)	.18	-52.6	(176.6)	-5.0	(-131.7 to 121.7)	-29.4	(-161.4 to 102.7)
High user engagement	338.9	(194.8)	348.3	(236.2)	.70	9.5	(191.4)	42.1	(-69.6 to 153.8)	6.1	(-116.9 to 129.2)
Red and processed meat (g/week)											
Low user engagement	602.8	(260.0)	632.3	(460.0)	.80	29.5	(377.5)	1.00	(reference)	1.00	(reference)
Moderate user engagement	661.3	(326.8)	662.3	(396.1)	.99	1.0	(469.4)	12.2	(-217.2 to 241.6)	4.8	(-225.7 to 235.4)
High user engagement	604.5	(502.9)	545.5	(291.5)	.29	-59.0	(438.1)	-87.3	(-290.4 to 115.8)	-90.9	(-308.0 to 126.2)
Sugar-sweetened beverages (g/day)											
Low user engagement	0.0	(0.0)	56.4	(133.7)	.19	56.4	(133.7)	—[f]	—	—	—
Moderate user engagement	28.2	(96.7)	28.2	(96.7)	1.00	0.0	(63.8)	—	—	—	—
High user engagement	0.0	(0.0)	3.3	(26.3)	.32	3.3	(26.3)	—	—	—	—
Fiber (g/day)											
Low user engagement	28.4	(9.7)	24.8	(7.5)	.13	-3.6	(7.2)	1.00	(reference)	1.00	(reference)
Moderate user engagement	25.7	(9.9)	28.8	(10.3)	.06	3.1	(7.2)	5.6	(-0.55 to 11.7)	5.4	(-0.96 to 11.7)
High user engagement	28.4	(13.5)	30.2	(12.0)	.19	1.8	(10.7)	5.4	(-0.04 to 10.8)	4.9	(-1.02 to 10.8)
Whole grains (g/day)											

XSL•FO
RenderX

Variable	Baseline		Short-term follow-up		Within change	Mean difference		Crude[a] Model estimates		Adjusted[b] Model estimates	
	Mean	(SD)	Mean	(SD)	*P* value[c]	Mean	(SD)	β	(95% CI)	β	(95% CI)
Low user engagement	70.0	(37.1)	55.0	(20.8)	.03	-15.0	(18.9)	1.00	(reference)	1.00	(reference)
Moderate user engagement	62.7	(35.7)	63.4	(33.8)	.90	0.8	(28.7)	13.8	(-5.9 to 33.6)	15.6	(-5.5 to 36.7)
High user engagement	63.2	(33.2)	68.5	(39.6)	.16	5.4	(29.3)	18.6	(1.04 to 36.1)	20.4	(0.57 to 40.3)
Carbohydrates (g/day)											
Low user engagement	226.8	(61.8)	202.4	(84.8)	.23	-24.4	(62.7)	1.00	(reference)	1.00	(reference)
Moderate user engagement	207.7	(75.8)	210.3	(73.2)	.82	2.6	(52.2)	22.1	(-18.7 to 62.9)	18.9	(-24.6 to 62.4)
High user engagement	202.3	(74.9)	213.7	(78.1)	.14	11.4	(59.5)	29.5	(-6.7 to 65.7)	23.3	(-17.9 to 64.5)
Saturated fat (g/day)											
Low user engagement	37.3	(17.7)	36.2	(18.7)	.76	-1.1	(11.4)	1.00	(reference)	1.00	(reference)
Moderate user engagement	34.4	(7.8)	34.5	(10.3)	.99	0.0	(12.7)	-0.26	(-7.9 to 7.4)	0.12	(-7.9 to 8.1)
High user engagement	29.5	(11.4)	30.9	(11.0)	.34	1.4	(11.5)	-1.3	(-8.2 to 5.6)	-2.0	(-9.7 to 5.8)
Unsaturated fat (g/day)											
Low user engagement	53.6	(23.4)	51.5	(21.6)	.53	-2.1	(10.6)	1.00	(reference)	1.00	(reference)
Moderate user engagement	49.7	(10.9)	49.8	(12.6)	.98	0.1	(12.2)	0.66	(-9.8 to 11.1)	0.92	(-10.6 to 12.5)
High user engagement	46.2	(19.9)	47.8	(19.3)	.47	1.6	(17.6)	0.90	(-8.4 to 10.2)	2.0	(-9.0 to 13.0)
Sodium (mg/day)											

XSL•FO
RenderX

Variable	Baseline		Short-term follow-up		Within change	Mean difference		Crude[a] Model estimates		Adjusted[b] Model estimates	
	Mean	(SD)	Mean	(SD)	*P* value[c]	Mean	(SD)	β	(95% CI)	β	(95% CI)
Low user engagement	2838.0	(773.5)	2694.8	(692.4)	.38	-143.2	(514.3)	1.00	(reference)	1.00	(reference)
Moderate user engagement	2778.0	(833.8)	2758.2	(713.7)	.90	-19.8	(738.5)	93.5	(-360.4 to 547.4)	35.8	(-447.4 to 518.9)
High user engagement	2603.5	(863.7)	2725.9	(768.4)	.22	122.4	(779.8)	148.9	(-254.7 to 552.5)	44.5	(-418.8 to 507.9)
Sucrose (g/day)											
Low user engagement	31.3	(12.5)	31.6	(19.7)	.94	0.3	(14.7)	1.00	(reference)	1.00	(reference)
Moderate user engagement	31.9	(13.9)	32.5	(13.2)	.82	0.6	(11.9)	0.48	(-7.8 to 8.8)	-1.2	(-9.9 to 7.5)
High user engagement	30.4	(14.8)	31.2	(14.3)	.61	0.8	(12.0)	0.13	(-7.2 to 7.5)	-1.8	(-9.9 to 6.3)
Total energy (kcal/day)											
Low user engagement	2258.5	(739.9)	2096.8	(789.3)	.29	-161.7	(481.3)	1.00	(reference)	1.00	(reference)
Moderate user engagement	2161.3	(533.7)	2147.7	(545.6)	.90	-13.6	(486.6)	111.6	(-240.8 to 464.1)	89.0	(-292.1 to 470.0)
High user engagement	2003.3	(636.0)	2095.6	(608.6)	0.19	92.3	(554.7)	158.2	(-156.4 to 472.8)	106.3	(-258.9 to 471.5)

[a]Results from the linear regression model adjusted for baseline values of the specific outcome variable to account for differences at baseline.

[b]Results from the linear regression model adjusted for baseline values, age, sex, intervention group, education level, BMI categories, diabetes duration, and HbA_{1c}.

[c]Results from the paired *t* test within the user engagement groups.

[d]NNR: Nordic Nutrition Recommendations.

[e]FFQ: food frequency questionnaire.

[f]Not available.

Results from the linear regression models (see Table 2) showed a statistically significant increase in intake of whole grains (β=20.4, 95% CI 0.57-40.3) in the high compared with the low user engagement group. Table 3 shows the results for cardiometabolic risk markers are presented. Statistically significant within group differences from baseline to the short-term follow-up in BMI was seen in the low (*P*=.03) and the high user engagement groups (*P*=.001), and on waist circumference within the moderate (*P*=.006) and high (*P*<.001) user engagement groups. After fitting linear regression models, no statistically significant differences in change were seen between groups for any cardiometabolic risk markers.

XSL•FO
RenderX

Table . Mean values and differences of short-term effects (directly after the 12-week dietary education) stratified by low, moderate or high app user engagement and linear regression models presented as difference in changes (β-coefficients and 95% confidence intervals) between the user engagement groups in the randomized controlled Healthy eating using APP technologY (HAPPY) trial. User engagement groups, low <50%, moderate 50% - 99%, high 100% of the total number of completed activities in the app.

Variable	Baseline		Short-term effect		Within change	Mean difference		Crude[a] Model estimates		Adjusted[b] Model estimates	
	Mean	(SD)	Mean	(SD)	*P* value [c]	Mean	(SD)	β	(95% CI)	β	(95% CI)
Cardiometabolic risk markers											
BMI (kg/m^2)											
Low user engagement	29.3	(4.7)	28.6	(4.8)	.03	-0.7	(1.1)	1.00	(reference)	1.00	(reference)
Moderate user engagement	31.3	(6.2)	31.0	(6.0)	.13	-0.4	(1.1)	.38	(-0.20 to 0.97)	.35	(-0.28 to 0.99)
High user engagement	29.4	(4.1)	29.1	(4.1)	.001	-0.3	(0.)	.37	(-0.13 to 0.88)	.44	(-0.13 to 1.0)
Waist circumference											
Low user engagement	103.7	(11.4)	102.2	(11.6)	.15	-1.5	(3.9)	1.00	(reference)	1.00	(reference)
Moderate user engagement	110.5	(14.5)	107.6	(13.8)	.006	-3.0	(4.8)	-1.02	(-3.60 to 1.55)	-.80	(-3.5 to 1.8)
High user engagement	104.2	(12.5)	102.1	(12.5)	<.001	-2.1	(3.6)	-.53	(-2.7 to 1.70)	.22	(-2.2 to 2.6)
Body fat (%)											
Low user engagement	30.5	(9.8)	29.7	(10.0)	.16	-0.8	(1.8)	1.00	(reference)	1.00	(reference)
Moderate user engagement	33.6	(10.4)	32.5	(10.0)	.15	-1.0	(3.3)	-.19	(-1.95 to 1.57)	-0.07	(-1.9 to 1.8)
High user engagement	32.2	(8.5)	32.7	(8.8)	.12	0.5	(2.3)	1.26	(-0.29 to 2.82)	.96	(-0.76 to 2.7)
HbA_{1c} (mmol/mol)											
Low user engagement	53.0	(10.5)	51.3	(6.2)	.37	-1.7	(5.3)	1.00	(reference)	1.00	(reference)

XSL•FO
RenderX

Variable	Baseline		Short-term effect		Within change	Mean difference		Crude[a] Model estimates		Adjusted[b] Model estimates	
	Mean	(SD)	Mean	(SD)	*P* value [c]	Mean	(SD)	β	(95% CI)	β	(95% CI)
Moderate user engagement	51.8	(9.9)	51.0	(10.5)	.61	-0.8	(7.0)	.62	(-4.00 to 5.24)	1.4	(-3.3 to 6.1)
High user engagement	47.8	(9.7)	47.2	(9.5)	.42	-0.7	(6.1)	-0.20	(-4.46 to 4.05)	-.66	(-4.6 to 3.6)
Triglycerides (mmol/L)											
Low user engagement	1.5	(0.8)	1.5	(0.7)	.83	-0.0	(0.5)	1.00	(reference)	1.00	(ref)
Moderate user engagement	1.7	(1.1)	1.9	(1.6)	.41	0.2	(1.1)	.26	(-0.45 to 0.97)	.39	(-0.37 to 1.1)
High user engagement	1.5	(0.8)	1.6	(1.1)	.34	0.1	(0.8)	.14	(-0.50 to 0.79)	.17	(-0.52 to 0.88)
Total cholesterol (mmol/L)											
Low user engagement	3.5	(0.7)	3.9	(0.7)	.22	0.4	(0.8)	1.00	(reference)	1.00	(reference)
Moderate user engagement	4.4	(1.2)	4.3	(1.3)	.45	-0.1	(0.7)	-.22	(-0.87 to 0.42)	-.19	(-0.87)
High user engagement	4.1	(1.0)	4.2	(1.0)	.39	0.1	(0.9)	-.09	(-0.67 to 0.49)	-.05	(-0‐66 to 0.56)
LDL[d] cholesterol (mmol/L)											
Low user engagement	2.1	(0.4)	2.1	(0.6)	.80	0.0	(0.5)	1.00	(reference)	1.00	(reference)
Moderate user engagement	2.4	(0.7)	2.2	(0.7)	.15	-0.2	(0.5)	-.11	(-0.57 to 0.35)	-0‐19	(-0.67 to 0.29)
High user engagement	2.3	(0.9)	2.2	(0.8)	.73	-0.0	(0.7)	-.01	(-0.43 to 0.40)	-.003	(-0.44 to 0.43)
HDL[e] cholesterol (mmol/L)											

XSL•FO
RenderX

Variable	Baseline		Short-term effect		Within change	Mean difference		Crude[a] Model estimates		Adjusted[b] Model estimates	
	Mean	(SD)	Mean	(SD)	*P* value [c]	Mean	(SD)	β	(95% CI)	β	(95% CI)
Low user engagement	1.1	(0.3)	1.1	(0.3)	.12	0.0	(0.1)	1.00	(reference)	1.00	(reference)
Moderate user engagement	1.2	(0.3)	1.2	(0.3)	.18	0.0	(0.2)	.01	(-0.12 to 0.13)	-.003	(-0.13 to 0.14)
High user engagement	1.3	(0.4)	1.4	(0.4)	.16	0.0	(0.2)	.00	(-0.12 to 0.12)	.005	(-0.12 to 0.13)
Blood pressure (mmHg)											
Systolic											
Low user engagement	135.8	(18.0)	137.0	(13.0)	.77	1.2	(15.8)	1.00	(reference)	1.00	(reference)
Moderate user engagement	138.5	(16.3)	138.4	(14.3)	.95	-0.2	(13.6)	-.11	(-8.4 to 8.19)	1.8	(-7.3 to 10.8)
High user engagement	133.3	(13.7)	130.7	(15.7)	.16	-2.6	(14.3)	-4.96	(-12.2 to 2.30)	-3.6	(-11.9 to 4.6)
Diastolic											
Low user engagement	83.3	(14.1)	85.9	(9.6)	.42	2.6	(12.0)	1.00	(reference)	1.00	(reference)
Moderate user engagement	88.3	(11.3)	87.6	(12.3)	.69	-0.7	(7.9)	-1.28	(-6.99 to 4.42)	-1.0	(-7.5 to 5.4)
High user engagement	82.6	(11.2)	82.2	(10.9)	.76	-0.4	(9.9)	-3.27	(-8.22 to 1.68)	-2.0	(-7.9 to 3.9)

[a]Results from the linear regression model adjusted for baseline values of the specific outcome variable to account for differences at baseline.

[b]Results from the linear regression model adjusted for baseline values, age, sex, intervention group, education level, BMI categories, diabetes duration, and HbA_{1c}.

[c]Results from the paired *t* test within the user engagement groups.

[d]LDL: low-density lipoprotein.

[e]HDL: high-density lipoprotein.

Table 4 shows the results from the long-term follow-up. When comparing within group differences, the moderate user engagement group had a statistically significant higher intake of fiber (*P*=.004) and whole grains (*P*=.01) at the long-term follow-up compared to baseline. No other significant differences were seen within the user engagement groups.

In linear regression models (Table 4), a statistically significant difference in change in fiber (β=7.3, 95% CI 1.2-13.4) and whole grain intake (β=26.6, 95% CI 2.4-50.8) was seen in the moderate engagement group compared with the low user engagement group. Among those with moderate user engagement compared to those with low, statistically significant differences in change were also seen for saturated fat (β=11.7, 95% CI 2.9-20.5), sodium (β=632.0, 95% CI 95.0-1169.0), and total energy (β=568.1, 95% CI 133.0-1003.1).

XSL•FO
RenderX

Table . Mean values and differences of long-term effects (9 months after the 12-week dietary education in the intervention group and 6 months after the education in the control group) stratified by low, moderate and high app user engagement and linear regression models presented as difference in changes (β-coefficients and 95% confidence intervals) between the user engagement groups in the randomized controlled Healthy eating using APP technologY (HAPPY) trial. User engagement groups, low <50%, moderate 50% - 99.9%, high 100% of the total number of completed activities in the app.

Variable	Baseline		Long-term effect		Within change	Mean difference		Crude[a] Model estimates		Adjusted[b] Model estimates	
	Mean	(SD)	Mean	(SD)	*P* value[c]	Mean	(SD)	β	(95% CI)	β	(95% CI)
NNR[d] (0 - 30 points)											
Low user engagement	12.7	(2.9)	13.5	(3.1)	.52	0.7	(3.6)	1.00	(reference)	1.00	(reference)
Moderate user engagement	12.1	(3.2)	13.4	(3.9)	.07	1.4	(2.8)	0.4	(-1.9 to 2.7)	0.8	(-1.7 to 3.3)
High user engagement	13.9	(3.3)	14.0	(3.6)	.72	0.2	(3.1)	-0.2	(-2.1 to 1.8)	-0.9	(-3.2 to 1.4)
Dietary variables (FFQ)[e]											
Fruit and vegetables (g/day)											
Low user engagement	279.1	(129.0)	267.3	(95.3)	.71	-11.8	(103.8)	1.00	(reference)	1.00	(reference)
Moderate user engagement	316.6	(144.7)	375.0	(375.0)	.08	58.5	(125.5)	85.5	(-23.1 to 194.2)	49.0	(-62.5 to 160.6)
High user engagement	390.3	(264.6)	397.8	(208.0)	.77	7.5	(186.5)	64.5	(-28.6 to 157.6)	29.3	(-70.3 to 128.8)
Legumes/pulses (g/day)											
Low user engagement	32.9	(32.6)	27.8	(37.9)	.52	-5.1	(25.6)	1.00	(reference)	1.00	(reference)
Moderate user engagement	24.4	(19.1)	38.2	(45.2)	.18	13.7	(39.4)	14.2	(-8.7 to 37.2)	11.2	(-14.9 to 37.3)
High user engagement	39.2	(37.5)	32.5	(27.3)	.17	-6.7	(34.5)	1.88	(-17.5 to 21.3)	0.57	(-22.3 to 23.5)
Total fish and seafood (g/week)											
Low user engagement	261.3	(212.7)	240.2	(97.1)	.63	-21.1	(139.0)	1.00	(reference)	1.00	(reference)

XSL•FO
RenderX

Variable	Baseline		Long-term effect		Within change	Mean difference		Crude[a] Model estimates		Adjusted[b] Model estimates	
	Mean	(SD)	Mean	(SD)	*P* value[c]	Mean	(SD)	β	(95% CI)	β	(95% CI)
Moderate user engagement	365.0	(225.8)	295.5	(182.8)	.18	-69.5	(195.6)	9.02	(-135.2 to 153.3)	29.8	(-126.2 to 185.8)
High user engagement	348.3	(195.5)	343.4	(223.0)	.88	-4.9	(230.4)	64.4	(-57.6 to 186.4)	67.0	(-71.5 to 205.4)
Red and processed meat (g/week)											
Low user engagement	559.3	(302.3)	576.4	(363.5)	.84	17.0	(266.5)	1.00	(reference)	1.00	(reference)
Moderate user engagement	681.0	(355.7)	682.9	(355.6)	.98	1.9	(259.8)	62.6	(-169.8 to 295.0)	61.2	(-188.8 to 311.2)
High user engagement	595.9	(513.1)	548.8	(329.4)	.47	-47.1	(474.7)	-40.8	(-236.9 to 155.3)	-61.2	(-283.7 to 161.4)
Sugar-sweetened beverages (g/day)											
Low user engagement	0.0	(0.0)	18.8	(62.4)	.34	18.8	(62.4)	—[g]	—	—	—
Moderate user engagement	25.9	(103.4)	25.9	(103.4)	—[f]	0.0	(0.0)	—	—	—	—
High user engagement	0.0	(0.0)	7.8	(56.8)	.32	7.8	(56.8)	—	—	—	—
Fiber (g/day)											
Low user engagement	28.4	(9.9)	24.7	(9.8)	.15	-3.7	(8.0)	1.00	(reference)	1.00	(reference)
Moderate user engagement	25.4	(9.7)	30.4	(13.9)	.004	5.0	(5.8)	7.9	(2.2 to 13.6)	7.3	(1.2 to 13.4)
High user engagement	27.9	(13.5)	29.0	(11.3)	.35	1.1	(8.5)	4.7	(-0.12 to 9.6)	3.9	(-1.5 to 9.3)
Whole grains (g/day)											
Low user engagement	74.2	(35.3)	60.0	(37.2)	.12	-14.1	(27.3)	1.00	(reference)	1.00	(reference)

XSL•FO
RenderX

Variable	Baseline		Long-term effect		Within change	Mean difference		Crude[a] Model estimates		Adjusted[b] Model estimates	
	Mean	(SD)	Mean	(SD)	*P* value[c]	Mean	(SD)	β	(95% CI)	β	(95% CI)
Moderate user engagement	55.3	(28.5)	69.5	(36.7)	.01	14.2	(19.4)	24.4	(3.1 to 45.8)	26.6	(2.4 to 50.8)
High user engagement	60.0	(31.9)	66.0	(36.6)	.15	6.0	(29.5)	17.2	(-0.79 to 35.1)	18.2	(-3.2 to 39.6)
Carbohydrates (g/day)											
Low user engagement	236.9	(78.4)	192.4	(77.9)	.07	-44.5	(71.3)	1.00	(reference)	1.00	(reference)
Moderate user engagement	212.2	(72.2)	230.8	(89.9)	.17	18.7	(52.2)	56.4	(14.1 to 98.7)	60.9	(13.8 to 108.1)
High user engagement	201.4	(79.3)	209.7	(74.1)	.29	8.3	(56.2)	43.1	(7.1 to 79.2)	46.2	(3.6 to 88.7)
Saturated fat (g/day)											
Low user engagement	39.2	(17.2)	31.5	(8.3)	.06	-7.7	(11.8)	1.00	(reference)	1.00	(reference)
Moderate user engagement	35.5	(8.1)	39.2	(14.5)	.36	3.7	(15.6)	9.8	(2.2 to 17.3)	11.7	(2.9 to 20.5)
High user engagement	29.4	(12.0)	30.4	(11.5)	.40	1.0	(9.0)	4.4	(-2.2 to 11.1)	5.9	(-2.3 to 14.1)
Unsaturated fat (g/day)											
Low user engagement	54.0	(24.4)	53.0	(16.0)	.82	-1.1	(15.3)	1.00	(reference)	1.00	(reference)
Moderate user engagement	49.6	(12.4)	54.9	(22.4)	.35	5.3	(21.7)	4.8	(-7.5 to 17.0)	6.9	(-7.2 to 21.0)
High user engagement	45.4	(19.5)	48.0	(19.5)	.23	2.6	(15.8)	0.54	(-9.9 to 11.0)	2.3	(-10.5 to 15.3)
Sodium (mg/day)											
Low user engagement	2837.7	(895.1)	2549.5	(795.8)	.23	-288.2	(742.9)	1.00	(reference)	1.00	(reference)

XSL•FO
RenderX

Variable	Baseline		Long-term effect		Within change	Mean difference		Crude[a] Model estimates		Adjusted[b] Model estimates	
	Mean	(SD)	Mean	(SD)	*P* value[c]	Mean	(SD)	β	(95% CI)	β	(95% CI)
Moderate user engagement	2776.9	(906.3)	3064.4	(1136.0)	.13	287.5	(719.9)	550.9	(58.6 to 1043.1)	632.0	(95.0 to 1169.0)
High user engagement	2595.6	(869.2)	2618.8	(698.7)	.82	23.2	(719.6)	212.4	(-205.8 to 630.6)	210.7	(-277.9 to 699.3)
Sucrose (g/day)											
Low user engagement	31.6	(14.2)	28.0	(13.9)	.18	-3.6	(8.3)	1.00	(reference)	1.00	(reference)
Moderate user engagement	31.7	(13.1)	34.2	(17.0)	.45	2.5	(12.8)	6.1	(-2.2 to 14.4)	4.9	(-4.4 to 14.3)
High user engagement	31.0	(15.5)	31.6	(13.8)	.73	0.6	(11.9)	4.0	(-3.1 to 11.0)	2.6	(-5.9 to 11.1)
Total energy (kcal/day)											
Low user engagement	2370.4	(867.1)	1971.4	(588.6)	.06	-399.0	(628.1)	1.00	(reference)	1.00	(reference)
Moderate user engagement	2181.2	(544.8)	2350.7	(785.2)	.28	169.5	(609.1)	496.2	(114.3 to 878.0)	568.1	(133.0 to 1003.1)
High user engagement	1989.1	(655.4)	2062.9	(596.0)	.30	73.8	(511.2)	327.1	(-1.03 to 655.2)	391.0	(-10.3 to 792.4)

[a]Results from the linear regression model adjusted for baseline values of the specific outcome variable to account for differences at baseline.

[b]Results from the linear regression model adjusted for baseline values, age, sex, intervention group, education level, BMI categories, diabetes duration and HbA_{1c}.

[c]Results from the paired *t* test within the user engagement groups.

[d]NNR: Nordic Nutrition Recommendations.

[e]FFQ: food frequency questionnaire.

[f]Too little variation in data.

[g]Not available.

Discussion

Principal Findings

Our results showed that user engagement of a 12-week app-based dietary education targeting people with type 2 diabetes was high. Those with the highest user engagement followed the NNR guidelines to a greater extent and had higher intake of fruit and vegetables already at baseline. High user engagement was associated with beneficial changes in whole grain intake in the short term, that is, directly following the intervention. We found no associations between user engagement and cardiometabolic risk markers. Long-term participants with moderate user engagement continued to report an increase in whole grain intake, along with additional beneficial changes in fiber intake and an increased intake of carbohydrates 6 to 9 months postintervention in comparison to participants with low user engagement. Interestingly, the moderate group had a higher intake of total energy, saturated fat, and sodium at the long-term follow-up, resulting in a significant difference in change compared with the low user engagement group. Thus, the low user engagement group had a comparatively better diet in terms of saturated fat and sodium intake.

XSL•FO
RenderX

Despite the importance of addressing user engagement, studies commonly do not mention this when presenting their results [9,30-32]. The comparison of studies is further complicated by the lack of a standardized definition. Sittig et al [33] reported high user engagement in a 9-week app-based lifestyle intervention, with participants completing, 75.1% of 177 activities, similar to our 77.1%. Their app, like ours, included written educational content and recipes, but also short educational videos. The length of their intervention was also relatively similar, 9 vs 12 weeks, respectively. In another study by Mummah et al [34], 75% (51/68) of participants engaged at least once with their app targeting vegetable consumption during 8 weeks. Low user engagement was reported by Alonso-Domínguez et al [35], with average app use being only 35 days (corresponding to 38.9% of 90 days) during a 3-month multicomponent diet intervention. Data on user-engagement for longer durations are limited, but Lim et al [36] showed that 62% (61/99) of participants used at least one app feature on at least 75% of days during 6 months. Among motivated participants, maintaining high user engagement for longer time periods may be feasible.

Lavikainen et al [12] identified 4 user engagement patterns during a 12-month app-based lifestyle intervention. They found that 46.9% (904/1926) discontinued use, 38% (731/1926) used the app weekly, 10.8% (208/1926) twice a week, and 4.3% (83/1926) daily. Similar to our findings, high user engagement was associated with improved diet quality, and scoring high on a Healthy Diet Index at baseline increased the odds of belonging to a high user engagement group. This suggests that those motivated to use the app already have a healthier diet. Furthermore, Mummah et al [34] showed that baseline vegetable consumption was a moderator of effect and the effect of their intervention increased with baseline consumption. Hendrie et al [13] showed that baseline vegetable consumption and app use were the best predictors of improved variety and intake of vegetables. Interestingly in this case, those with lower baseline intakes increased their variety most and participants with the highest app use had almost doubled their intake compared to participants with the lowest use. Hence, app effectiveness may relate to both app-use and participant characteristics.

We observed a decrease in completed activities over time for the moderate and low user engagement groups, similar to previous mHealth interventions [37-39]. Nevertheless, the majority, over 70%, of our participants completed at least one activity every week. Somewhat fewer participants engaged with the app during weeks targeting sugar intake and slow and fast carbohydrates. One of the most common topics in diabetes management is carbohydrate calculation [40] and potentially our participants felt confident in this, resulting in less engagement. This is in contrast to previous research showing that user engagement can be improved by focusing on health problems or specific needs [10].

Future studies need to investigate reasons for why people stop engaging with apps. Declining app use has been identified as an early indicator of disengagement in a digital lifestyle intervention, highlighting the importance of closely monitoring and support users during the initial phase [12]. Components to increase user engagement in apps, including commercial apps, include gamification (eg, goal setting and quizzes), a more personal approach to app content, such as individualized and tailored reminders, automated data collection, and human-like features [4,41]. A user-friendly design, culturally and personally tailored content, and caregiver support have also been found to play a role [42]. In our app-design, we used gamification elements, such as allowing the user to set a personal goal for the week and progress-track the percentage of activities that had been completed each day, which has been shown to serve positive reinforcement for behavior changes [42]. This, hopefully, encouraged users to be more active in their own progress. However, our app did not include personalized reminders, for example, through timing where the user can set customized reminders based on their needs, which might have increased engagement.

Previous app experiences and social support within apps, such as networking and experience sharing, have been shown to strengthen user engagement, but while social support may influence behavior change for some, it can also introduce negative social competition [42]. We did not include any social elements in our app. In a qualitative study, people with type 2 diabetes preferred app features that supported them to make informed and independent decisions about their own care, rather than following explicit directives [41]. Factors that could make our app more personal and, thus, more engaging could include allowing participants to choose specific activities in the app, such as only receiving educational information, or receiving recipes suitable for their preferences, such as a vegetarian diet.

Strengths and Limitations

A strength of this study is that we automatically collected detailed data on app usage. However, we do not know how much time users spent with the app. It is a limitation that we did not explore why user engagement dropped over time. The inclusion of both women and men is another strength, especially since men are usually more challenging to recruit in interventions [43]. The sex distribution and other characteristics, such as age, BMI, and HbA_{1c} levels, in our study are similar to the average adult population with type 2 diabetes in Swedish primary care [44]. Among our participants, 84.2% were overweight or obese, compared to 82.7% in primary care. However, our participants were slightly younger (63.2 years vs 68.7 years) and had a lower HbA_{1c} (47.7 mmol/mol vs 52.0 mmol/mol). The fact that participants were recruited in primary care and from areas with relatively high socioeconomic status, and that the app was only available in Swedish, limits the generalizability of our results to other socioeconomic groups.

Additional strengths include use of a validated FFQ [21,22] and addressing both short-term and long-term dietary intake. This allowed us to examine if changes were maintained over time. Furthermore, anthropometrics and cardiometabolic risk markers were objectively measured, but the latter only at short-term follow-up. Different long-term follow-up intervals between participants in the two groups is also a limitation. In addition, power was calculated based on the primary outcome of the trial [15]. The small sample size in this study may have been underpowered to detect statistical significance in secondary analyses. Finally, social desirability bias and recall bias may

XSL•FO
RenderX

have influenced reporting in the FFQ. For example, foods considered healthier may be over-reported and unhealthy foods more under-reported [45]. However, we believe this is likely to be random and not related to a specific user engagement group.

Conclusion

In our study, engagement with a smartphone-based dietary education was high, suggesting that an app targeting dietary habits is feasible to use. Those with high user engagement maintained healthier dietary behaviors over time, compared to those with low user engagement. Future studies should address the challenge of designing targeted mHealth solutions to engage individuals in maintaining app-usage during the full extent of interventions.

Acknowledgments

We would like to express our gratefulness to the personnel at the participating health care centers for the help in recruiting participants to our study. We would also like to thank Anna Dahlgren for her extensive work in the data collection. This study was supported by the Strategic Research Area Health Care Sciences (SFO-V) at Karolinska Institutet (SEB), Karolinska Institutet Doctorial funding (SEB), grants provided by Region Stockholm (NSV project; YTL), and funding from the regional agreement between Stockholm County Council and Karolinska Institutet, clinical research appointment (YTL). The funders had no role in the study design, data collection, analyses, interpretation of data, decision to publish, or preparation of the manuscript. Generative AI tools have not been used in any portion of the presented work or manuscript writing.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to ethical restrictions, but are available from the principal investigator (SEB) on reasonable request.

Authors' Contributions

SEB is the principal investigator of the HAPPY trial and have designed the study together with YTL. LS contributed significantly to the data collection. SEB and LS conceptualized this study. LS and FS prepared the data and performed the statistical analysis. LS and SEB wrote the original draft of the manuscript. FS, YTL, and EH critically reviewed the manuscript. SEB and YTL managed resources and funding acquisition. All the authors have read and approved the finial manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Description of the Nordic Nutrition Recommendations 2023 (NNR) score and a directed acyclic graph (DAG) displaying potential confounders in the association between user engagement and dietary changes.
[DOCX File, 147 KB - formative_v9i1e71408_app1.docx]

Checklist 1
CONSORT-EHEALTH checklist V 1.6.1.
[PDF File, 665 KB - formative_v9i1e71408_app2.pdf]

References

1. Petroni ML, Brodosi L, Marchignoli F, et al. Nutrition in patients with type 2 diabetes: present knowledge and remaining challenges. Nutrients 2021 Aug 10;13(8):2748. [doi: 10.3390/nu13082748] [Medline: 34444908]
2. Evert AB, Dennison M, Gardner CD, et al. Nutrition therapy for adults with diabetes or prediabetes: a consensus report. Diabetes Care 2019 May;42(5):731-754. [doi: 10.2337/dci19-0014] [Medline: 31000505]
3. MHealth - use of appropriate digital technologies for public health. : World Health Organization; 2018 URL: https://apps.who.int/gb/ebwha/pdf_files/WHA71/A71_20-en.pdf [accessed 2024-05-27]
4. Jakob R, Harperink S, Rudolf AM, et al. Factors influencing adherence to mhealth apps for prevention or management of noncommunicable diseases: systematic review. J Med Internet Res 2022 May 25;24(5):e35371. [doi: 10.2196/35371] [Medline: 35612886]
5. Sieverink F, Kelders SM, van Gemert-Pijnen JE. Clarifying the concept of adherence to eHealth technology: systematic review on when usage becomes adherence. J Med Internet Res 2017 Dec 6;19(12):e402. [doi: 10.2196/jmir.8578] [Medline: 29212630]
6. Donkin L, Christensen H, Naismith SL, Neal B, Hickie IB, Glozier N. A systematic review of the impact of adherence on the effectiveness of e-therapies. J Med Internet Res 2011 Aug 5;13(3):e52. [doi: 10.2196/jmir.1772] [Medline: 21821503]

XSL•FO
RenderX

7. Perski O, Blandford A, West R, Michie S. Conceptualising engagement with digital behaviour change interventions: a systematic review using principles from critical interpretive synthesis. Transl Behav Med 2017 Jun;7(2):254-267. [doi: 10.1007/s13142-016-0453-1] [Medline: 27966189]
8. Yardley L, Spring BJ, Riper H, et al. Understanding and promoting effective engagement with digital behavior change interventions. Am J Prev Med 2016 Nov;51(5):833-842. [doi: 10.1016/j.amepre.2016.06.015] [Medline: 27745683]
9. Young C, Campolonghi S, Ponsonby S, et al. Supporting engagement, adherence, and behavior change in online dietary interventions. J Nutr Educ Behav 2019 Jun;51(6):719-739. [doi: 10.1016/j.jneb.2019.03.006] [Medline: 31036500]
10. Schubart JR, Stuckey HL, Ganeshamoorthy A, Sciamanna CN. Chronic health conditions and internet behavioral interventions: A review of factors to enhance user engagement. Comput Inform Nurs 2011 Feb;29(2):81-92. [doi: 10.1097/NCN.0b013e3182065eed] [Medline: 21164337]
11. Kelders SM, Kok RN, Ossebaard HC, Van Gemert-Pijnen J. Persuasive system design does matter: a systematic review of adherence to web-based interventions. J Med Internet Res 2012 Nov 14;14(6):e152. [doi: 10.2196/jmir.2104] [Medline: 23151820]
12. Lavikainen P, Mattila E, Absetz P, et al. Digitally supported lifestyle intervention to prevent type 2 diabetes through healthy habits: secondary analysis of long-term user engagement trajectories in a randomized controlled trial. J Med Internet Res 2022 Feb 24;24(2):e31530. [doi: 10.2196/31530] [Medline: 35200147]
13. Hendrie GA, Hussain MS, Brindal E, James-Martin G, Williams G, Crook A. Impact of a mobile phone app to increase vegetable consumption and variety in adults: large-scale community cohort study. JMIR Mhealth Uhealth 2020 Apr 17;8(4):e14726. [doi: 10.2196/14726] [Medline: 32301739]
14. Sjöblom L, Hantikainen E, Dahlgren A, Trolle Lagerros Y, Bonn SE. The effect of an app-based dietary education on dietary intake and cardiometabolic risk markers in people with type 2 diabetes: results from a randomized controlled trial. Nutr J 2025 Jan 4;24(1):2. [doi: 10.1186/s12937-024-01069-2] [Medline: 39754157]
15. Trolle Lagerros Y, Dahlgren A, Sjöblom L, Bonn SE. Digital support for healthier eating habits among patients with type 2 diabetes: protocol for a randomized clinical trial within primary care (HAPPY trial). JMIR Res Protoc 2020 Nov 16;9(11):e24422. [doi: 10.2196/24422] [Medline: 33196448]
16. Find your way - the swedish dietary advice [hitta ditt sätt - de svenska kostråden]. The Swedish Food Agency [Livsmedelsverket]. 2015. URL: https://www.livsmedelsverket.se/om-oss/publikationer/sok-publikationer/broschyr/hitta_ditt_satt [accessed 2024-05-27]
17. Champion VL, Skinner CS. The health belief model. In: Health Behav Health Educ Theory Res Pract, 4th edition: Jossey-Bass; 2008:45-65.
18. Shaffer JA. Stages-of-change model. In: Gellman MD, Turner JR, editors. Encycl Behav Med: Springer; 2013:1871-1874. [doi: 10.1007/978-1-4419-1005-9_1180]
19. Bandura A. Health promotion from the perspective of social cognitive theory. Psychology & Health 1998 Jul;13(4):623-649. [doi: 10.1080/08870449808407422]
20. National Cancer Institute US. Department of Health and Human Services Theory at A Glance A Guide For Health Promotion Practice, 2nd edition: National Institute of Health; 2012. URL: https://www.sbccimplementationkits.org/demandrmnch/wp-content/uploads/2014/02/Theory-at-a-Glance-A-Guide-For-Health-Promotion-Practice.pdf [accessed 2025-05-14]
21. Eke H, Sjöblom L, Lagerros YT, Bonn SE. A validation study comparing energy and nutrient intake between A web-based food frequency questionnaire and A 4-d dietary record. Nutrition 2024 Apr;120:112332. [doi: 10.1016/j.nut.2023.112332] [Medline: 38295492]
22. Messerer M, Johansson SE, Wolk A. The validity of questionnaire-based micronutrient intake estimates is increased by including dietary supplement use in Swedish men. J Nutr 2004 Jul;134(7):1800-1805. [doi: 10.1093/jn/134.7.1800] [Medline: 15226472]
23. Livsmedelsdatabasen [the swedish food composition database]. Livsmedelsverket [The Swedish Food Agency]. 2023. URL: https://www.livsmedelsverket.se/livsmedel-och-innehall/naringsamne/livsmedelsdatabasen [accessed 2024-05-27]
24. Blomhoff R, Andersen R, Arnesen EK, et al. Nordic Nutrition Recommendations 2023: Nordic Council of Ministers; 2023. [doi: 10.6027/nord2023-003]
25. Olsson SJG, Ekblom Ö, Andersson E, Börjesson M, Kallings LV. Categorical answer modes provide superior validity to open answers when asking for level of physical activity: A cross-sectional study. Scand J Public Health 2016 Feb;44(1):70-76. [doi: 10.1177/1403494815602830] [Medline: 26392418]
26. Physical activity. World Health Organization. 2022. URL: https://www.who.int/news-room/fact-sheets/detail/physical-activity [accessed 2024-05-27]
27. Waist circumference and waist-hip ratio: report of a WHO expert consultation. World Health Organization. 2011. URL: https://apps.who.int/iris/handle/10665/44583 [accessed 2024-05-27]
28. Blood sample: HBA1c [blodprov: HBA1c]. 1177. 2022. URL: https://www.1177.se/undersokning-behandling/undersokningar-och-provtagning/provtagning-och-matningar/blodprov/blodprov-hba1c/ [accessed 2024-05-27]
29. Twisk J, Bosman L, Rijnhart J, Hoekstra T, Welten M, Heymans M. Different ways to estimate treatment effects in randomised controlled trials. Contemp Clin Trials Commun 2018 Jun;10:80-85. [doi: 10.1016/j.conctc.2018.03.008] [Medline: 29696162]

XSL•FO
RenderX

30. Lugones-Sánchez C, Recio-Rodríguez JI, Menéndez-Suárez M, et al. Effect of a multicomponent mHealth intervention on the composition of diet in a population with overweight and obesity-randomized clinical trial EVIDENT 3. Nutrients 2022 Jan 9;14(2):270. [doi: 10.3390/nu14020270] [Medline: 35057451]
31. Martin CK, Miller AC, Thomas DM, Champagne CM, Han H, Church T. Efficacy of SmartLoss SM, a smartphone-based weight loss intervention: results from a randomized controlled trial. Obesity (Silver Spring) 2015 May;23(5):935-942. [doi: 10.1002/oby.21063] [Medline: 25919921]
32. Sun C, Sun L, Xi S, et al. Mobile phone-based telemedicine practice in older Chinese patients with type 2 diabetes mellitus: randomized controlled trial. JMIR Mhealth Uhealth 2019 Jan 4;7(1):e10664. [doi: 10.2196/10664] [Medline: 30609983]
33. Sittig S, Wang J, Iyengar S, Myneni S, Franklin A. Incorporating behavioral trigger messages into a mobile health app for chronic disease management: randomized clinical feasibility trial in diabetes. JMIR Mhealth Uhealth 2020 Mar 16;8(3):e15927. [doi: 10.2196/15927] [Medline: 32175908]
34. Mummah S, Robinson TN, Mathur M, Farzinkhou S, Sutton S, Gardner CD. Effect of a mobile app intervention on vegetable consumption in overweight adults: a randomized controlled trial. Int J Behav Nutr Phys Act 2017 Sep 15;14(1):125. [doi: 10.1186/s12966-017-0563-2] [Medline: 28915825]
35. Alonso-Domínguez R, García-Ortiz L, Patino-Alonso MC, Sánchez-Aguadero N, Gómez-Marcos MA, Recio-Rodríguez JI. Effectiveness of a multifactorial intervention in increasing adherence to the mediterranean diet among patients with diabetes mellitus type 2: a controlled and randomized study (EMID study). Nutrients 2019 Jan 14;11(1):162. [doi: 10.3390/nu11010162] [Medline: 30646500]
36. Lim SL, Ong KW, Johal J, et al. Effect of a smartphone app on weight change and metabolic outcomes in Asian adults with type 2 diabetes: a randomized clinical trial. JAMA Netw Open 2021 Jun 1;4(6):e2112417. [doi: 10.1001/jamanetworkopen.2021.12417] [Medline: 34081137]
37. Gong E, Baptista S, Russell A, et al. My Diabetes Coach, a mobile app-based interactive conversational agent to support type 2 diabetes self-management: randomized effectiveness-implementation trial. J Med Internet Res 2020 Nov 5;22(11):e20322. [doi: 10.2196/20322] [Medline: 33151154]
38. Lee EY, Cha SA, Yun JS, et al. Efficacy of personalized diabetes self-care using an electronic medical record-integrated mobile app in patients with type 2 diabetes: 6-month randomized controlled trial. J Med Internet Res 2022 Jul 28;24(7):e37430. [doi: 10.2196/37430] [Medline: 35900817]
39. Nelson LA, Spieker A, Greevy R, LeStourgeon LM, Wallston KA, Mayberry LS. User engagement among diverse adults in a 12-month text message-delivered diabetes support intervention: results from a randomized controlled trial. JMIR Mhealth Uhealth 2020 Jul 21;8(7):e17534. [doi: 10.2196/17534] [Medline: 32706738]
40. Hood M, Wilson R, Corsica J, Bradley L, Chirinos D, Vivo A. What do we know about mobile applications for diabetes self-management? A review of reviews. J Behav Med 2016 Dec;39(6):981-994. [doi: 10.1007/s10865-016-9765-3] [Medline: 27412774]
41. Baptista S, Wadley G, Bird D, Oldenburg B, Speight J, My Diabetes Coach Research Group. User experiences with a type 2 diabetes coaching app: qualitative study. JMIR Diabetes 2020 Jul 17;5(3):e16692. [doi: 10.2196/16692] [Medline: 32706649]
42. Szinay D, Jones A, Chadborn T, Brown J, Naughton F. Influences on the uptake of and engagement with health and well-being smartphone apps: systematic review. J Med Internet Res 2020 May 29;22(5):e17572. [doi: 10.2196/17572] [Medline: 32348255]
43. Pagoto SL, Schneider KL, Oleski JL, Luciani JM, Bodenlos JS, Whited MC. Male inclusion in randomized controlled trials of lifestyle weight loss interventions. Obesity (Silver Spring) 2012 Jun;20(6):1234-1239. [doi: 10.1038/oby.2011.140] [Medline: 21633403]
44. Eeg-Olofsson K, Åkesson K, Nåtman J, et al. Nationella diabetesregistret, årsrapport 2022 [Article in Swedish]. : Göteborg: Swedish National Diabetes Register; 2022 URL: https://www.ndr.nu/pdfs/Arsrapport_NDR_2022.pdf
45. Warnecke RB, Johnson TP, Chávez N, et al. Improving question wording in surveys of culturally diverse populations. Ann Epidemiol 1997 Jul;7(5):334-342. [doi: 10.1016/s1047-2797(97)00030-6] [Medline: 9250628]

Abbreviations

CONSORT-EHEALTH : Consolidated Standards of Reporting Trials of Electronic and Mobile HEalth Applications and onLine TeleHealth
FFQ: Food Frequency Questionnaire
HAPPY: Healthy Eating using APP technologY
HDL: high-density lipoprotein cholesterol
LDL: low-density lipoprotein cholesterol
mHealth: mobile health
NNR: Nordic Nutrition Recommendations

XSL•FO
RenderX

Edited by A Mavragani; submitted 17.01.25; peer-reviewed by FH Leung; revised version received 09.04.25; accepted 11.04.25; published 30.05.25.

Please cite as:
Sjoblom L, Stenbeck F, Trolle Lagerros Y, Hantikainen E, Bonn SE
Engagement With a Smartphone-Delivered Dietary Education Intervention and Its Relation to Dietary Intake and Cardiometabolic Risk Markers in People With Type 2 Diabetes: Secondary Analysis of a Randomized Controlled Trial
JMIR Form Res 2025;9:e71408
URL: https://formative.jmir.org/2025/1/e71408
doi:10.2196/71408

© Linnea Sjoblom, Freja Stenbeck, Ylva Trolle Lagerros, Essi Hantikainen, Stephanie E Bonn. Originally published in JMIR Formative Research (https://formative.jmir.org), 30.5.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Components of a Digital Storytelling Intervention for Human Papillomavirus and Cancer Prevention Among LGBTQ+ Individuals: Formative Mixed Methods Inquiry

Gabrielle Darville-Sanders[1*], MPH, PhD, CHES; Dominique Munroe[1*], MPH, MD; Emilie Corluyan[1*], MPH, CHES; Utibeabasi Ikoiwak[1*], MPH; Jennifer Nguyen[2*], MPH, PhD; Chad Mandula[3*], PhD; Portia Thomas[4*], MPH, DNSc, RN-CNE; Brandon Sanders[5*], MS

[1]Department of Public Health, Mercer University, 3001 Mercer University Drive, Atlanta, GA, United States
[2]College of Pharmacy, Mercer University, Atlanta, GA, United States
[3]Institute of Higher Education, University of Georgia, Athens, GA, United States
[4]School of Nursing, Yale University, West Haven, CT, United States
[5]Empowerment Resource Center, Atlanta, GA, United States
[*]all authors contributed equally

Corresponding Author:
Gabrielle Darville-Sanders, MPH, PhD, CHES
Department of Public Health, Mercer University, 3001 Mercer University Drive, Atlanta, GA, United States

Abstract

Background: Human papillomavirus (HPV) is one of the most prevalent sexually transmitted infections in the United States; however, vaccination uptake falls far below the goal of 80% of the population set forth by Healthy People 2030. Specifically, within the LGBTQ+ (lesbian, gay, bisexual, transgender, queer/questioning) population, HPV vaccination adherence remains a complex issue. Due to the widespread use of technology within the young adult population, digital health tools such as digital storytelling (DST) have been promoted as an effective way to increase vaccination uptake.

Objective: The purpose of this study was to conduct a formative inquiry into (1) what components should be considered for inclusion in an HPV documentary tailored for sexual and gender minority populations and (2) what dissemination channels would be more effective and impact the uptake and completion of the HPV vaccine among sexual and gender minority populations. Additionally, this study aims to provide insight into perceived HPV risk and its implications on the HPV vaccine uptake within the LGBTQ+ population.

Methods: A mixed methods study was conducted between January 2021 and September 2021 in Atlanta, Georgia. Intake surveys were distributed to individuals identifying as members of the LGBTQ+ community to examine demographic characteristics, barriers to vaccine adherence, and current HPV vaccination status. Perceived HPV risk was assessed using 5 statements on a 1 to 7 Likert scale. Key informant interviews were conducted via Zoom with participants who completed the intake surveys and consented to be interviewed. Transcripts were coded and analyzed using the constant comparison method for emergent themes surrounding components of effective DST campaigns.

Results: Forty-seven individuals completed the intake survey and interview. A total of 13 out of 47 (27.7%) of participants indicated that they were not sure when provided with the statement “I am likely to get HPV”, whereas 12 out of 47 (29.8%) participants strongly disagreed with the statement “I am at high risk for getting HPV” and 13 out of 47 (27.7%) participants indicated that they were not sure when presented with the statement “HPV would be a serious threat to the quality of my life.” A total of 14 out of 47 (29.8%) participants responded that they were not sure to the statement “HPV would be a severe threat to my health” and 13 out of 47 (27.7%) participants strongly agreed that “HPV would be a severe threat to my sex life.” Qualitative analysis indicated a high level of stigma experienced in interactions between the LGBTQ+ population and private practitioners. Major barriers to vaccination hesitancy were concerns about age, perceived reduced risk, and lack of provider recommendation. Participant interviews revealed that “Real Outcomes,” and “Accurate Representation” were the main components that should be considered for inclusion in an HPV documentary tailored for sexual and gender minority populations.

Conclusions: Creation of a DST intervention within the LGBTQ+ population should include information surrounding the real outcomes of HPV and accurate representation.

(JMIR Form Res 2025;9:e58163) doi:10.2196/58163

XSL•FO
RenderX

KEYWORDS

human papillomavirus; vaccination; cancer prevention; digital storytelling; public health

Introduction

Vaccines have proven to be an effective primary public health measure towards preventing disease, disability, and death from infectious diseases [1-3]. Vaccination is not only vital to individual health, but also a crucial component of improving the health of communities by reducing the spread of disease. Human papillomavirus (HPV) is the most prevalent sexually transmitted infection in the United States, with a prevalence of 40% among persons aged 15 - 59 years [4]. It is also responsible for a significant amount of morbidity and mortality [5,6]. Despite disease prevalence, HPV vaccination uptake has fallen short of 80% of the general population, a goal set forth by Healthy People 2030 [7].

The World Health Organization identified complacency, inconvenience, and lack of confidence as several of the various factors that can influence an individual's intention to become vaccinated. These factors, when not addressed, result in vaccine uncertainty or delayed vaccination and the overall undermining of public health efforts due to decreased vaccine coverage and an increased risk of disease outbreaks [8]. Vaccination hesitancy is not a new problem [9-12]and has been recognized by the World Health Organization as a top 10 threat to global health [13-15]. Controversy has consistently surrounded the measles, mumps and rubella vaccine [14], diphtheria, tetanus and pertussis vaccine [14,16], swine flu vaccine [14,17], polio vaccine [14,18], influenza vaccine [19], and most recently the COVID-19 vaccine [9,13,20,21]; affecting adherence of the general population. Similarly, the HPV vaccine has also seen low levels of initial uptake and moderate levels of series completion [22,23]. In 2016, only 60% of 13 - 17-year-olds completed 1 dose of the vaccine series [24], and less than 50% of teens were up to date with the recommended HPV vaccination series [4,24,25]. Furthermore, major disparities surrounding geography [26,27], economic status, educational level [28], and health care provider recommendations [29] have been found to be correlated with HPV vaccination.

Various strategies have been used to address vaccine hesitancy, such as peer mentoring [30-32] and workshops/lectures [33]. One of the more novel ideas, however, has been that of digital storytelling (DST). The participatory nature of DST has significant potential to promote participants' psychosocial health and well-being, revealing hidden stories and initiating community dialogue about issues that are pressing and concerning to research participants [34]. Furthermore, DST as a medium can decrease the time between knowledge generation and implementation [34]. Given the high use of technology among young adults [35], digital health technologies have been touted as an effective way to encourage uptake among this group [36,37]. In DST, participants describe their personal stories to an audience that is specific to a lived experience. On the other side, the audience is influenced by the story through their own relevant feelings or responses [34]. This model has been investigated by researchers to ascertain whether or not it can be effective in regard to the distrust surrounding the COVID vaccine among Black Americans [38]. Kim et al [39] used DST as a means to promote HPV vaccination among Korean American College women and found the intervention to result in improvement in knowledge and attitude at the conclusion of the intervention. Additionally, in their study, the experimental group was twice as likely to receive the HPV vaccine as the comparison group. A similar intervention was conducted by Chen et al [40] in which they found DST to be effective in increasing Korean mothers' intention to vaccinate their children against HPV (from 53% to 74%).

Despite the well-documented success of DST, there remain several limitations to its use. Ethical challenges can arise with the use of DST, specifically regarding the use of participant photographs and voices, which can result in issues with the maintenance of participant confidentiality [34]. There also remains a gap in the use of DST specifically in relation to the LGBTQ+ (lesbian, gay, bisexual, transgender, queer/questioning) population. There is a pattern of discrimination, stigma, and lack of awareness among health care providers, along with issues surrounding access to sex education, health care, and insurance coverage that have impeded LGBTQ+ populations from accessing health care services and thus contributed to major disparities surrounding HPV and HPV vaccination [41-47]. The purpose of this study was to conduct a formative inquiry into (1) what components should be considered for inclusion in an HPV documentary tailored for sexual and gender minority populations, and (2) what dissemination channels would be more effective and impact the uptake and completion of the HPV vaccine among sexual and gender minority populations. Thus, informing the construction of a DST intervention to increase HPV vaccine uptake within the LGBTQ+ population. Additionally, this study aims to provide insight into the influence that perception of HPV risk can have on HPV vaccine uptake.

Methods

Study Design

This mixed method, explanatory sequential study was conducted among individuals who identified as members of the LGBTQ+ community between January 2021 and September 2021 in Atlanta, Georgia.

Ethical Considerations

This study was conducted according to the guidelines 2018 Federal Regulations 21 CFR 56.110(b) and 45 CFR 46.110(b) (for expedited review) and was approved under categories _6, _7 per 63 FR 60364. All procedures involving research study participants were approved by the Mercer University Institutional Review Board on April 2, 2020. Written informed consent was obtained from all participants.

On the opening page of the survey, participants were presented with information about the study's objectives and scope, followed by an informed consent statement. Participants were free to withdraw at any point without consequence. In addition,

XSL•FO
RenderX

participation was anonymous, and no personally identifiable information was collected or used. Informed consent was obtained from all participants involved in the study, in compliance with the H2004105 Mercer University Institutional Review Board protocol. Participants were compensated for their participation via a US $40 Amazon gift card for completion of both the intake survey and participant interview.

Participants

Participants were recruited using convenience sampling to recruit participants for the study. Using statewide organizational connections, community partnerships, web-based or email advertising, and word-of-mouth recruitment, 60 participants were recruited. Eligible participants were those aged 18 years or older, currently residing in Atlanta, Georgia, capable of providing informed consent before starting the survey, and identifying as members of the LGBTQ+ community.

After enrolling in the study via SuperSaaS software (SuperSaaS B.V.), survey participants were emailed the link to the Qualtrics survey with the informed consent and intake survey to complete prior to their interview participation. Participant information was deidentified using a participant identification number assigned to all registrants via the SuperSaaS software [48]. Reminder emails were sent 1 week prior to the session, with follow-up emails sent 3 days before the interview session. Follow-up emails were sent to nonrespondents to encourage participation. If no response was received by the morning of the final day of data collection, those participants were removed from the study, resulting in a cohort of 47.

Procedure

Demographic data were obtained from the 15 initial intake questions distributed after participants had signed up for an interview session. The intake survey was distributed via Qualtrics, took approximately 5 - 10 minutes to complete, and was developed with input from a subject matter expert specializing in LGBTQ populations. Scale items were adapted from a previous HPV study assessing 3 key constructs: risk perception (3 items), self-efficacy (3 items), and behavioral intention (4 items) using a 7-point Likert scale(1=strongly disagree to 7=strongly agree) [49]. The total score was calculated from the sum of the response items, with higher scores indicating a greater perceived risk of HPV. An additional LGBTQ-related demographic question was added, but was not revalidated or pretested.

Due to special precautions associated with the COVID-19 pandemic, all interviews were conducted via Zoom and lasted about 45 minutes. Three student researchers previously trained in qualitative data collection, ethics, and compliance were enlisted to facilitate the interviews. Student researchers verified that the informed consent and intake survey were completed prior to the start of the interview, and using the intake survey, we were able to capture demographic data from our sample. The interview protocol consisted of open-ended questions adapted from a similar study surrounding the initiation of pre-exposure prophylaxis amongst Black men who have sex with men [50]. Two versions of the interview protocol were developed based on vaccination status. If participants indicated that they were vaccinated, their interview focused on the facilitators/motivating factors that encouraged them to do so. On the other hand, if indicated that they were not vaccinated, interview questions focused on the barriers to completing the vaccine series. Following the completion of the interviews, student researchers accessed the Zoom transcript files and cleaned them for data analysis.

Data Analysis

Descriptive statistics were computed for all variables using StataSE (version 14; StataCorp LLC). The total number of responses and percentages were reported for categorical variables; mean and standard deviation were reported for continuous variables.

Using NVivo software (QSR International), 2 researchers trained in qualitative data analysis and NVivo software technology coded the 47 interviews. Using the constant comparative method, a codebook was developed for thematic coding and analysis by the research team [51]. Additionally, to reduce bias and to establish interrater reliability, the researchers developed a schedule for coding and a coding scheme. Each week, the researchers would code 10% of the transcripts independently and then merge coding findings into one dataset. The research team would then meet to run a coding comparison query in NVivo. Any data set sample that reported an agreement less than 90% was discussed to ensure that each individual researcher understood the codes and associated definitions/applications. The data set was coded again by the researcher individually, and the process was replicated until 90% - 100% agreement was achieved. A coding agreement of 90% - 100% means that minimal to no disagreements were reported using kappa coefficient in NVivo and consensus was established [52]. After coding was completed, the data were organized using NVivo software to explore emerging patterns and themes.

Positionality of the Research Team

The diverse racial, cultural, and professional backgrounds of the research team included Black Caribbean, Black US American, White US American, and Asian American. This diversity allowed for study participants to authentically express their thoughts and feel comfortable sharing their lived experiences. Four of the authors hold primary backgrounds in public health, with experiences in health equity, community-centered research, and systems-level changes that shaped the framing of research questions, participant response, and theme prioritization during the qualitative analysis, which enriched the study depth. The secondary author is a foreign-trained medical physician whose clinical experiences in low-resource settings shaped a unique understanding of the structural inequities within health care and the importance of the patient-physician relationship, allowing for a deeper contextual understanding of participant responses.

Furthermore, each author's unique personal experiences shaped how we approached the research process, and we acknowledge that our diverse professional and cultural contexts may have introduced potential biases within the analysis. To avoid speaking for the data, the team used reflexivity and iterative discussions. This included note-taking, which allowed the team

XSL•FO
RenderX

to document and bring to light preconceptions while simultaneously ensuring that our interpretations were grounded in participants' voices. The team sought to ensure that our research findings were both ethically grounded and accurately interpreted through collaborative discussion and critical reflection.

Results

Quantitative Findings

Although initially 60 participants signed up to participate, only 47 total interviews were conducted, resulting in a 78.33% completion rate. Demographic characteristics observed within our cohort were as follows: 51.1% (24/47) of the cohort were 27 years and older, and 48.9% (23/47) of the participants were White (Table 1). The breakdown concerning gender identity were as follows: 44.7% (21/47) of the participants identified as a man, 36.2% (28/47) of the participants identified as a woman, 6.4% (3/47) of the participants were nonbinary, and 4.3% (2/47) of the participants were gender fluid. A total of 29.8% (14/47) of the participants of our sample identified as gay, 19.1% (9/47) of the participants were bisexual, 14.9% (7/47) of the participants were lesbian, 12.8% (6/47) of the participants were pansexual, and 8.5% (4/47) of the participants were queer. In addition, 53.2% (25/47) of the participants indicated that they were single, and 19.1% (9/47) of the participants indicated that they were in a monogamous relationship. Around 74.5% (35/47) of the participants of our sample had been sexually active in the last 6 months, and 14.9% (7/47) of the participants engaged in unprotected sex of some kind. Thirty-six out of 47 (54.4%) participants indicated that their total lifetime partners were between 5 and 29 (54.4%) and 29.8% (14/47) of the participants never had a sexually transmitted disease. Forty-three out of 47 (91.5%) participants had not been diagnosed with HPV, and 28/33 (84.9%) participants tested indicated that they were HIV negative. Additionally, 22/47 participants (46.8%) indicated that they initiated the HPV vaccination.

XSL•FO
RenderX

Table . Demographic characteristics (N=47).

	Values, n (%)
Age (years)	
20	10 (21.3)
21 - 23	8 (17.0)
24 - 26	5 (10.6)
27 and older	24 (51.1)
Race and ethnicity	
Asian	2 (4.3)
Biracial or multiracial	6 (12.8)
Black or African American	14 (29.8)
Hispanic	2 (4.3)
White or Caucasian	23 (48.9)
Assigned sex at birth	
Female	28 (59.6)
Male	19 (40.4)
Gender identity	
Gender fluid	2 (4.3)
Genderqueer	1 (2.1)
Man	21 (44.7)
Nonbinary	3 (6.4)
Other or self-identify	3 (6.4)
Woman	17 (36.2)
Sexuality	
Asexual	2 (4.3)
Bisexual	9 (19.1)
Gay	14 (29.8)
Heterosexual	1 (2.1)
Lesbian	7 (14.9)
Pansexual	6 (12.8)
Queer	4 (8.5)
Questioning	1 (2.1)
Other or self-identify	3 (6.4)
Relationship status	
Cohabitating	4 (8.5)
Married or partnered	9 (19.1)
Monogamous relationship (only dating one person)	9 (19.1)
Single (not in a relationship)	25 (53.2)
Sexually active (Last 6 mo)	
No	12 (25.5)
Yes	35 (74.5)
Sexual behaviors (n=41)	
Protected anal sex	6 (12.8)
Unprotected anal sex	9 (19.1)

XSL•FO
RenderX

	Values, n (%)
Unprotected oral sex	7 (14.9)
Unprotected vaginal sex	12 (25.5)
Protected vaginal sex	7 (14.9)
Lifetime sexual partners	
1 to 4	11 (23.4)
5 to 7	7 (14.9)
8 to 14	7 (14.9)
15 to 29	12 (25.5)
30 or more	10 (21.3)
Have you ever been diagnosed with an STD[a]?	
No	33 (70.2)
Yes	14 (29.8)
Have you ever been tested for HIV?	
No	14 (29.8)
Yes	33 (70.2)
Have you ever been diagnosed with HPV?	
No	43 (91.5)
Yes	4 (8.5)
Have you ever received any doses of the HPV Vaccine?	
No	25 (53.2)
Yes	22 (46.8)
How many doses of the HPV vaccine have you received? (n=22)	
1 dose	2 (9.1)
2 doses	5 (22.7)
All 3 doses	15 (68.2)
HIV Status (n=33)	
Confirmatory positive	5 (10.6)
Negative	28 (84.8)
What kind of insurance do you have?	
Insurance through employer	22 (46.8)
No insurance	4 (8.5)
Other	2 (4.3)
Private insurance	9 (19.1)
Public or government sponsored	8 (17.0)
Unsure	2 (4.3)

[a]STD: sexually transmitted disease.

Of our sample, 68.2% (15/47) of the participants indicated that they had completed the HPV vaccination series (3 doses received) and most 46.8% (22/47) of the participants had health insurance through an employer followed by 19.1% (9/47) of the participants having private insurance and 17.0% (8/47) of the participants being public/government sponsored (Table 1).

Risk for getting HPV (how susceptible they were to contracting the virus) was assessed using 5 statements on a 1 to 7 Likert scale (Table 2). When provided with the statement “I am likely to get HPV,” 27.66% (13/47) of our participants responded not sure, followed by 23.40% (11/47) of the participants who strongly disagreed. Overall, 29.79% (12/47) of our study sample strongly disagreed with the statement “I am at high risk for getting HPV” and 27.66% (13/47) of the participants indicated that they were not sure when presented with the statement “HPV would be a serious threat to the quality of my life.” A total of

XSL•FO
RenderX

29.79% (14/47) of the participants responded not sure to the statement "HPV would be a severe threat to my health and 27.66% (13/47) of the participants strongly agreed that "HPV would be a severe threat to my sex life." Cumulatively, our sample responded with an average score of 18.73 out of a possible 35 for all 5 statements used to measure perceived risk (Table 3).

Table . Human papillomavirus (HPV) perceived risk among sexual and gender minorities (N=47).

	HPV would be a severe threat to my sex life, n (%)	HPV would be a severe threat to my health, n (%)	HPV would be a serious threat to the quality of my life, n (%)	I am at high risk for getting HPV, n (%)	I am likely to get HPV, n (%)
Strongly agree	13 (27.70)	7 (14.90)	11 (23.40)	5 (10.60)	4 (8.51)
Agree	12 (25.50)	10 (21.30)	8 (17.00)	3 (6.40)	4 (8.51)
Slightly agree	8 (17.00)	10 (21.30)	6 (12.80)	4 (8.50)	4 (8.51)
Not sure	9 (19.10)	14 (29.80)	13 (27.70)	9 (19.10)	13 (27.66)
Slightly disagree	1 (2.10)	3 (6.40)	4 (8.50)	4 (8.50)	2 (4.26)
Disagree	3 (6.40)	2 (4.30)	4 (8.50)	10 (21.30)	9 (19.15)
Strongly disagree	1 (2.10)	0 (0.00)	1 (2.10)	12 (25.50)	11 (23.40)

Table . Distribution of perceived risk of human papillomavirus (HPV) scores among participants. Higher scores indicate a greater perceived risk of HPV. Score ranges reflect grouped responses based on participants' self-assessment. Percentages are based on the total sample (N=47).

Range	Values, n (%)
Less than 15	4 (8.5)
16 - 19	15 (32)
20 - 23	14 (29.7)
24 - 27	7 (14.8)
28 - 35	5 (10.6)

Qualitative Findings

Participant interviews revealed that "Real Outcomes," and "Accurate Representation" were the main components that should be considered for inclusion in an HPV documentary tailored for sexual and gender minority populations. Interviewees also revealed insights into their unique "Health Care Experiences" and the "Importance of using social media to disseminate information."

Health Care Experiences

Participants discussed the quality of the health care they have received as well as their exposure to health information. Health care providers see a variety of patients daily and therefore have a higher possibility of exposing individuals to health information [53]. However, according to the accounts relayed by the participants, providers and health systems are doing little to take advantage of that opportunity. As one participant described, they received health information rarely, or only when they asked questions.

> *As far as suburban Atlanta goes, it's much more difficult, right? Because, whenever you go into doctors' clinics or stuff sometimes there are pamphlets, but sometimes there's just not that much information. It's like a, if you ask them, you get information sort of basis. It's not, you can- they'll tell you, without asking.*

Participants' expressions of the quality of care they have received from their health care providers varied (Textbox 1). Participants preferred to use LGBTQ + friendly health care providers and clinics, with some mentioning that the major barrier faced was race, not sexuality. This is exemplified in the quote below:

> *Yeah, I try to go to, I think something basic, like STD testing or like a routine checkup, yeah I try to go to the queer friendly spaces or places that are queer-affiliated. Um, where, as if it goes like, if it seems like I can't handle it in urgent care, um I still try to avoid going to X clinic, because every single time I've been there it's been trash treatment. And just the big bill, so first I'll start with if I can't do it with urgent care or like queer-friendly spaces, I will just try a new hospital. Is my new thing so yeah. Only a stigma I'm working against you know just being black, even in a predominantly black city.*

XSL•FO
RenderX

Textbox 1. Participant quotes on health care experiences (health information exposure and quality of care).

Health information exposure:

- *Both at the clinic, and there's a there's a like a truck that comes by, I guess, where you can also do it certain days of the month, things like that. So, I feel like for on campus people, it's a lot easier to get information, even if you're not necessarily looking for it.*
- *It just kind of seemed a little ridiculous, like um because I don't know much about healthcare, and I don't know much about testing so to have to ask her was like, a lot because I didn't even know what to ask for. I was just like; do you have any other tests? And she was like oh yeah, do you want to like get this test done, I was like yeah, are there any other ones? Like, I'll get all of them done.*
- *I mean I access the, you know, I get messages through like the telehealth portals and things like that around, um, different things that I'm asking about, um. I don't think I really got much. I mean, I remember picking up a couple of brochures here and there about like the HPV vaccine, and different stuff, and reading them at the clinic.*
- *I just felt like she, you know, she really talks about you know how I can prevent certain things like you know here [are] some resources for you to look at you know if you ever needed like that she you know she was the one that pointed me towards the health department if I ever need any like STI testing...overall she's providing a lot of information about different things about you know how to stay healthy, in general...*

Quality of care:

- *I don't like going to the doctor, I very much avoid it, that can be kind of hard to nail down. Just because I've just been rubbed so wrong by going in and just constantly having to either a) explain my whole history again, um b) it's just obvious through their mood or tone that they don't respect me, or see me, as you know. They're just uncomfortable being around me. Like, I was you know it's very hard. It's like, I'm now at the point where it's like, 'okay, the doctor has to prove to me that I can trust them'.*
- *It really depends on where I go and if I get lucky or not.*
- *I think the major differences that doctor is inside the city of Atlanta, so. Most doctors that I see that are inside the city of Atlanta, the likelihood kind of shoots way up that I'm going to have an okay experience, but the further outside of Atlanta I go, the more likely, I am just sort of you know, have to explain things that are unrelated to my being there.*

Many of the participants shared that they avoid visiting their doctor due to past negative experiences, as well as discussed how they felt their health care providers needed to earn their trust before they could be completely open with their health care needs. The participants' experiences highlight the need to improve the consistency of high-quality health care.

Dissemination

The participants indicated the importance of using social media as a channel for disseminating DST products (Textbox 2). The information should be presented in a manner that does not overwhelm viewers and is easy to understand, so that viewers are left satisfied with what they have learned. Participants highlighted the need for information to be just enough that it was not overwhelming and provided key takeaways.

Textbox 2. Participant quotes on dissemination (digital storytelling and use of social media).

Digital storytelling:

- *I will say, those that dive into the more human and emotional side. You know, especially when you're dealing with those [who] are being affected by certain things, and it kind of, kind of pulls you in.*
- *If you have it shorter, you can reach more people, but when you have it longer you have more of a chance to make a higher impact. So, I feel like there are definitely advantages to both, so if you actually did both, then you could use one as a segue into the other.*
- *I think I'd go stay away from doing like the you know, like the really sort of tragic/emotional music, I guess, if you're trying to lessen the stigma um more of just probably something more like trying to normalize it, have people living like their daily lives and stuff maybe like just slice of life type thing.*

Social media:

- *Popular social media websites. Like, the most successful sort of like, videos and stuff I've ended up like, seeing I guess like for like vitalities purposes, like going viral and whatever is like. If I follow people on like, multiple social media platforms, where, even if I don't, and I see that video pop up more than once, or that sort of article or whatever I'm way more likely to see it.*
- *I think it'd be better to if it was shorter too. Somehow get it on social media and just get make it somehow get it to be viral honestly. So, it's got [to be] super catchy.*
- *Honestly, I would say number one is like, the Internet. Um, and like, social media in general, just because um, speaking of the context of growing up in the South, like I, the first place, that I learned to go to look for LGBT like information at all, let alone LGBT healthcare is the Internet.*

Have it be like hey, this is what this is, this is what it can do and just present the facts in an easy to understand manner, where people don't feel overwhelmed by the information and just laying it out for people in a way that they can understand that they don't have to go and do all of their like you do the research for them, and you say like here, it is, and this nice little bundle so you can take it away...

XSL•FO
RenderX

Participants describe how they are drawn towards human and emotional perspectives in what they watch, especially when it comes to stories involving life experiences. The participants' overall suggestions for DST include the importance of presentation, length of the videos or film, and sharing true stories that express emotion and highlight humanity. In relation to modes of dissemination, one participant shared that much of the information they learned relating to the LGBTQ+ community was via web-based mediums. Social media connects people who are like them and provides a safe place for those who do not have a safe place in real life. Because of this reason, it proved to be an optimal channel identified by participants as seen below:

> *I would say, social media is pretty big honestly. Like a lot of people who feel like they don't have voices in real life, I think, are more willing or feel more comfortable reaching out through social media. And I think, from your point of view, as someone who wants to provide more education and information and stuff. I think social media would be a great platform.*

Creating something as simple as a few short videos that provide health information and keep the interest of the viewer, and make it easy to share across multiple platforms, can increase the number of views and potentially lead to the videos going viral.

Representation

Accurate representation of the LGBTQ+ community was another theme that arose during the interviews (Textbox 3). Participants noted that there is a substantial difference between the media's prototypical LGBTQ+ persona and the actual modern-day lives that they live. Accurately capturing their reality would be vital to increasing their engagement. Several of them noted that when they saw themselves being represented in popular sitcoms and television shows, their interest in that show or character increased instantly. Furthermore, it was noted that this representation should also encompass what it looks like to live with HPV, as well as diversity in terms of cultural backgrounds and races, as described by one participant:

> *I would say, you know include all different race of people. Um, and not just that one kind. Um, you know, mix it up a bit. You know, black, white, Afr- I mean, Chinese, Japanese, Korean, Asian you know, go outside the normal box. You know, all sizes um, male, female, transgender. Um, just like ready to go out there, all ages.*

Textbox 3. Participant quotes on representation (race and accurate representation of being LGBTQ+ [lesbian, gay, bisexual, transgender, queer/questioning]).

Representation:

- *We should be represented in a way that includes all races. All, I would say, like all guys and females we're not all the same, and some people, you can't even look at them and know that you know that they're like gay or that they're even sick, you know.*
- *I would say, you know include all different race of people. Um, and not just that one kind. Um, you know, mix it up a bit. You know, black, white, Afr- I mean, Chinese, Japanese, Korean, Asian you know, go outside the normal box. You know, all sizes um, male, female, transgender. Um, just like ready to go out there, all ages.*
- *I think it's definitely important to include some race representation. Because most, most race, they're probably going to be like 'Oh. this isn't for me' and just turn it off or, or not pay attention, but finding a way to connect, to, including ace representation. Um, and why it's important to think about, even if you're not sexually active.*

Doing so would provide a common experience that participants felt would allow them to foster a greater connection with the topic, in addition to providing increased optimism and hope for those living with the disease.

Messaging

Participants also noted the importance of highlighting real outcomes when speaking about HPV. Notable perspectives mentioned the need to demonstrate not only the prevalence of the disease but also to demystify the disease to the general population. Various complications can arise from being infected with HPV; from cervical cancer, anal and penile cancer to more benign problems such as warts.

> *Someone who has lived through it, or is living through it, or had to care for someone who had it, etc, or someone who even had it advance all the way to, to ovarian cancer or you know however it may affect men. That would be definitely attention grabbing.*

According to the participants, including this wide range of problems would be engaging and allow the audience to graphically connect with the importance of being vaccinated. They also mentioned that having caregivers or family members share their viewpoints would also be beneficial to connecting with the intended audience (Textbox 4).

XSL•FO
RenderX

Textbox 4. Participant quotes on messaging (importance of real outcomes).

Real outcomes:

- *Look for somebody that can make a case for getting vaccinated, you know. As well, somebody that- that has it and has had a lot of problems with it. Oh, you know I've had to do, you know, I had HPV and it turned to cervical cancer and I had to you know have my cervix removed... You know, and now I can never have children, I can't you know, whatever because of this, you know get vaccinated. And so, make it a little bit somebody that can talk kind of graphically, just like I just grabbed your attention with just that though.*
- *It's not necessarily separate from you, um, like if you're not sexually active, then I can see how that would be kind of separate from you, but it could still happen to someone that you care about or it could, if you're sexually active it can definitely happen to you.*
- *So, it's like if you can't make that personal connection, if you can't you don't think anyone around you has it, you think like, oh well, it's not a real problem, or like, yeah. So, I think that's just confusing because, obviously, people do have it so there's it's more of like a silent thing, where people just aren't talking about it. Yeah, so kind of connecting that and making people realize that people do have it, they just don't talk about it would be nice, I guess.*

Discussion

Principal Findings

Our goal within this study was to identify what characteristics are essential to include in the development of HPV-related vaccination intervention for an LGBTQ+ audience and to identify dissemination methods that would ensure virality. This study has highlighted 2 key items to include in digital stories: real outcomes and representation.

Comparison to Prior Work

Past research surrounding patient narratives has shown them to be a useful tool to gather information, communication, engagement, persuasion, and health behavior change [54]. DST expands upon the utility of patient narratives through its ability to focus on capturing the experiences of populations to share findings in an engaging manner through digital media [34]. However, the use of DST has some trade-offs with health care professionals arguing that the quality or accuracy of media-based DST is a major concern due to their lack of medical expertise [54]. Thus, the creation of any DST method for any behavioral intervention should allow for the opportunity to collaborate with health care professionals to ensure the provision of reliable, evidence-based medical guidance.

Results from our formative inquiry support the well-reported health disparities that the LGBTQ+ population frequently experiences, such as stigma and discrimination from health care providers [43-46]. Stigma and discrimination in health care settings undermine effective communication, which is essential for delivering quality care and fostering self-confidence. When comparing the communication and care quality between races, researchers found that Hispanic and Asian survivors reported poorer communication than White cancer providers, with Asians also reporting poorer care quality [55].

There is a pattern of discrimination, stigma, and lack of awareness among health care providers that has impeded LGBTQ+ populations from accessing health care services and has had a major impact on their health [41] resulting in less sexually transmitted disease screenings, inappropriate contraceptive counseling, less contraceptive use, and misinformation on sexual education and reproductive promotion. For example, studies have shown that lesbian women have lower use of sexual and reproductive health services, leading to significantly lower rates of pap smear testing [56]. Additionally, the Center for Disease Control has reported that 71% of new HIV infections seen were among gay and bisexual men in the United States in 2022 [57]. The use of DST within this population would allow for increased engagement and connections with others (ie, patients, advocacy groups, caregivers, health care professionals, and policy makers) that share similar lived experiences. Furthermore, through sharing their unique stories via this method, they can alter preconceived notions about their susceptibility/risk, attack misinformation concerning vaccinations, and reduce the health care disparities seen, thus improving the overall health of the LGBTQ+ population.

Strengths and Limitations

Several limitations exist within the study. First, our sampling methodology could be a limitation [58]. We used voluntary convenience sampling to recruit study participants [59]. It is argued that this type of nonprobability sampling is not truly representative of the entire population, as the sample is likely to overrepresent or underrepresent certain groups and mainly targets persons with stronger opinions. Furthermore, this method has a low external validity and a lack of transferability to other populations [60]. Future research should consider partnering with community organizations during the recruitment period, thus diversifying participant groups and perspectives. Second, our cohort was concentrated within the Atlanta area. Atlanta has a unique demographic, cultural, and public health characteristic, which includes a greater exposure to public health campaigns that may influence overall health behaviors and beliefs. DST interventions conducted in other states/ countries may encounter alternative barriers or facilitators to HPV vaccination and thus may discover that other elements may need to be included.

A more in-depth analysis of the resulting themes could have yielded a deeper understanding of participant viewpoints and their implications. While we used a mixed methods research design, we failed to triangulate the data collection due to the practical constraints related to participant recruitment, such as patient and researcher availability and stringent COVID-19 protocols. As a result, the conclusions and insights from the participants were based on a single approach, thus limiting their validity. Additionally, as we did not collect or analyze data regarding differences in findings between vaccination status and other characteristics, which limited our ability to explore

XSL•FO
RenderX

how the perceived risk of HPV may vary across subgroups. Future research should consider analyzing this area in order to not only provide a deeper understanding of the effect that individual factors may have on perceived risk but also provide insight into how to tailor public health interventions.

Conclusions

The LGBTQ+ community is often misrepresented, underrepresented, or not at all represented in media, nor do they often feel safe and comfortable enough to share their experiences with others outside of their community. The use of DST methods involving real outcomes and accurate representation would bridge the gap created by stigmatization in order to improve HPV vaccination adherence and reduce health care disparities within this population.

Acknowledgments

The authors would like to acknowledge the contributions of all undergraduate students (Fatima Alarcon-Tinoco, Zaira Khan, and Danielle Hawk) who assisted in the collection of data for the study. This study was made possible through the financial support of Mercer University's Provost Office Seed Grant.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

All authors agree to be accountable for all aspects of the work. GD-S, BS, UI, JN, CM, and PT were responsible for the concept and design of the study. GD-S, DM, EC, and UI contributed to the acquisition, analysis, or interpretation of the data. All authors participated in drafting the manuscript and provided critical revisions for important intellectual content. Statistical analysis was performed by GD-S, DM, EC, and UI. All authors have read and approved the final version of the manuscript.

Conflicts of Interest

None declared.

References

1. Nandi A, Shet A. Why vaccines matter: understanding the broader health, economic, and child development benefits of routine vaccination. Hum Vaccin Immunother 2020 Aug 2;16(8):1900-1904. [doi: 10.1080/21645515.2019.1708669] [Medline: 31977283]
2. Orenstein WA, Ahmed R. Simply put: vaccination saves lives. Proc Natl Acad Sci U S A 2017 Apr 18;114(16):4031-4033. [doi: 10.1073/pnas.1704507114] [Medline: 28396427]
3. Rodrigues CMC, Plotkin SA. Impact of vaccines; health, economic and social perspectives. Front Microbiol 2020;11:1526. [doi: 10.3389/fmicb.2020.01526] [Medline: 32760367]
4. Hirth J. Disparities in HPV vaccination rates and HPV prevalence in the United States: a review of the literature. Hum Vaccin Immunother 2019;15(1):146-155. [doi: 10.1080/21645515.2018.1512453] [Medline: 30148974]
5. Lewis RM, Laprise JF, Gargano JW, et al. Estimated prevalence and incidence of disease-associated human papillomavirus types among 15- to 59-year-olds in the United States. Sex Transm Dis 2021 Apr 1;48(4):273-277. [doi: 10.1097/OLQ.0000000000001356] [Medline: 33492097]
6. McClung NM, Gargano JW, Park IU, et al. Estimated number of cases of high-grade cervical lesions diagnosed among women - United States, 2008 and 2016. MMWR Morb Mortal Wkly Rep 2019 Apr 19;68(15):337-343. [doi: 10.15585/mmwr.mm6815a1] [Medline: 30998672]
7. Chido-Amajuoyi OG, Jackson I, Yu R, Shete S. Declining awareness of HPV and HPV vaccine within the general US population. Hum Vaccin Immunother 2021 Feb 1;17(2):420-427. [doi: 10.1080/21645515.2020.1783952] [Medline: 32692632]
8. Etowa J, Beauchamp S, Fseifes M, et al. Understanding low vaccine uptake in the context of public health in high-income countries: a scoping review. Vaccines (Basel) 2024 Mar 4;12(3):269. [doi: 10.3390/vaccines12030269] [Medline: 38543903]
9. Alcendor DJ. Targeting COVID vaccine hesitancy in rural communities in Tennessee: implications for extending the COVID-19 pandemic in the south. Vaccines (Basel) 2021 Nov 4;9(11):1279. [doi: 10.3390/vaccines9111279] [Medline: 34835210]
10. Dubé E, Laberge C, Guay M, Bramadat P, Roy R, Bettinger JA. Vaccine hesitancy. Hum Vaccines Immunother 2013 Aug 8;9(8):1763-1773. [doi: 10.4161/hv.24657]
11. Garett R, Young SD. Online misinformation and vaccine hesitancy. Transl Behav Med 2021 Dec 14;11(12):2194-2199. [doi: 10.1093/tbm/ibab128] [Medline: 34529080]

XSL•FO
RenderX

12. Oostvogel PM, van Wijngaarden JK, van der Avoort HG, et al. Poliomyelitis outbreak in an unvaccinated community in The Netherlands, 1992-93. Lancet 1994 Sep 3;344(8923):665-670. [doi: 10.1016/s0140-6736(94)92091-5] [Medline: 7915354]
13. Hou Z, Tong Y, Du F, et al. Assessing COVID-19 vaccine hesitancy, confidence, and public engagement: a global social listening study. J Med Internet Res 2021 Jun 11;23(6):e27632. [doi: 10.2196/27632] [Medline: 34061757]
14. Nuwarda RF, Ramzan I, Weekes L, Kayser V. Vaccine hesitancy: contemporary issues and historical background. Vaccines (Basel) 2022 Sep 22;10(10):1595. [doi: 10.3390/vaccines10101595] [Medline: 36298459]
15. Ten health issues WHO will tackle this year. World Health Organization. URL: https://www.who.int/news-room/spotlight/ten-threats-to-global-health-in-2019 [accessed 2023-09-21]
16. Kulenkampff M, Schwartzman JS, Wilson J. Neurological complications of pertussis inoculation. Arch Dis Child 1974 Jan;49(1):46-49. [doi: 10.1136/adc.49.1.46] [Medline: 4818092]
17. GBS (guillain-barré syndrome) and vaccines | vaccine safety. Center for Disease Control. URL: https://www.cdc.gov/vaccinesafety/concerns/guillain-barre-syndrome.html [accessed 2023-09-21]
18. Historical safety concerns | vaccine safety. Center for Disease Control. URL: https://www.cdc.gov/vaccinesafety/concerns/concerns-history.html [accessed 2023-09-21]
19. Stephens AB, Hofstetter AM, Stockwell MS. Influenza vaccine hesitancy: scope, influencing factors, and strategic interventions. Pediatr Clin North Am 2023 Apr;70(2):227-241. [doi: 10.1016/j.pcl.2022.11.003] [Medline: 36841592]
20. Rittle C. COVID-19 vaccine hesitancy and how to address it. Workplace Health Saf 2022 Feb;70(2):56-62. [doi: 10.1177/21650799211073525] [Medline: 35073804]
21. Yasmin F, Najeeb H, Moeed A, et al. COVID-19 vaccine hesitancy in the United States: a systematic review. Front Public Health 2021;9:770985. [doi: 10.3389/fpubh.2021.770985] [Medline: 34888288]
22. Amantea C, Foschi N, Gavi F, et al. HPV vaccination adherence in working-age men: a systematic review and meta-analysis. Vaccines (Basel) 2023 Feb 15;11(2):443. [doi: 10.3390/vaccines11020443] [Medline: 36851321]
23. Ejezie CL, Osaghae I, Ayieko S, Cuccaro P. Adherence to the recommended HPV vaccine dosing schedule among adolescents aged 13 to 17 years: findings from the national immunization survey-teen, 2019-2020. Vaccines (Basel) 2022 Apr 8;10(4):577. [doi: 10.3390/vaccines10040577] [Medline: 35455325]
24. Bednarczyk RA, Ellingson MK, Omer SB. Human papillomavirus vaccination before 13 and 15 years of age: analysis of national immunization survey teen data. J Infect Dis 2019 Jul 31;220(5):730-734. [doi: 10.1093/infdis/jiy682] [Medline: 30657920]
25. Walker TY, Elam-Evans LD, Singleton JA, et al. National, regional, state, and selected local area vaccination coverage among adolescents aged 13-17 years - United States, 2016. MMWR Morb Mortal Wkly Rep 2017 Aug 25;66(33):874-882. [doi: 10.15585/mmwr.mm6633a2] [Medline: 28837546]
26. Hirth JM, Rahman M, Smith JS, Berenson AB. Regional variations in HPV vaccination among 9–17 year old adolescent females from the BRFSS, 2008–2010. Hum Vaccines Immunother 2014 Dec 2;10(12):3475-3483. [doi: 10.4161/21645515.2014.980202]
27. Rahman M, McGrath CJ, Berenson AB. Geographic variation in human papillomavirus vaccination uptake among 13-17 year old adolescent girls in the United States. Vaccine (Auckl) 2014 May 1;32(21):2394-2398. [doi: 10.1016/j.vaccine.2014.02.097] [Medline: 24637175]
28. Kurani S, MacLaughlin KL, Jacobson RM, et al. Socioeconomic disadvantage and human papillomavirus (HPV) vaccination uptake. Vaccine (Auckl) 2022 Jan 24;40(3):471-476. [doi: 10.1016/j.vaccine.2021.12.003] [Medline: 34916103]
29. Rosen BL, Shepard A, Kahn JA. US health care clinicians' knowledge, attitudes, and practices regarding human papillomavirus vaccination: a qualitative systematic review. Acad Pediatr 2018 Mar;18(2S):S53-S65. [doi: 10.1016/j.acap.2017.10.007] [Medline: 29502639]
30. Esaggoff A, Cohen, S, Chang G, Equils O, Orman SV, Burnett A. 2531. using peer-to-peer education to increase awareness and uptake of HPV vaccine among Chinese international students. Open Forum Infect Dis 2019 Oct 23;6(Supplement_2):S879-S880. [doi: 10.1093/ofid/ofz360.2209]
31. Flowers P, Hart GJ, Williamson LM, Frankis JS, Der GJ. Does bar-based, peer-led sexual health promotion have a community-level effect amongst gay men in Scotland? Int J STD AIDS 2002 Feb;13(2):102-108. [doi: 10.1258/0956462021924721] [Medline: 11839164]
32. Ganczak M, Pasek O, Duda-Duma Ł, Komorzycka J, Nowak K, Korzeń M. A peer-based educational intervention effects on SARS-CoV-2 knowledge and attitudes among Polish high-school students. Int J Environ Res Public Health 2021 Nov 20;18(22):12183. [doi: 10.3390/ijerph182212183] [Medline: 34831939]
33. Tutt M, Begay C, George S, et al. Diné teachings and public health students informing peers and relatives about vaccine education: providing Diné (Navajo)-centered COVID-19 education materials using student health messengers. Front Public Health 2022;10:1046634. [doi: 10.3389/fpubh.2022.1046634] [Medline: 36589984]
34. Rieger KL, West CH, Kenny A, et al. Digital storytelling as a method in health research: a systematic review protocol. Syst Rev 2018 Mar 5;7(1):41. [doi: 10.1186/s13643-018-0704-y] [Medline: 29506568]
35. Villanti AC, Johnson AL, Ilakkuvan V, Jacobs MA, Graham AL, Rath JM. Social media use and access to digital technology in US young adults in 2016. J Med Internet Res 2017 Jun 7;19(6):e196. [doi: 10.2196/jmir.7303] [Medline: 28592394]

XSL•FO
RenderX

36. Lehtimaki S, Martic J, Wahl B, Foster KT, Schwalbe N. Evidence on digital mental health interventions for adolescents and young people: systematic overview. JMIR Ment Health 2021 Apr 29;8(4):e25847. [doi: 10.2196/25847] [Medline: 33913817]
37. Young people and digital health interventions: working together to design better. World Health Organization. URL: https://www.who.int/news/item/29-10-2020-young-people-and-digital-health-interventions-working-together-to-design-better [accessed 2024-03-05]
38. Budhwani H, Maragh-Bass AC, Tolley EE, et al. Tough talks COVID-19 digital health intervention for vaccine hesitancy among Black young adults: protocol for a hybrid type 1 effectiveness implementation randomized controlled trial. JMIR Res Protoc 2023 Feb 13;12(1):e41240. [doi: 10.2196/41240] [Medline: 36689557]
39. Kim M, Lee H, Kiang P, et al. A storytelling intervention in a mobile, web-based platform: a pilot randomized controlled trial to evaluate the preliminary effectiveness to promote human papillomavirus vaccination in Korean American college women. Health Educ Behav 2020 Apr;47(2):258-263. [doi: 10.1177/1090198119894589] [Medline: 31958991]
40. Chen ACC, Kim WS, Todd M, Larkey L. A digital storytelling intervention for Vietnamese American mothers to promote their children's HPV vaccination. Cancer Prev Res (Phila) 2022 Jul 5;15(7):465-472. [doi: 10.1158/1940-6207.CAPR-21-0618] [Medline: 35314854]
41. Carpenter E. "The Health System Just Wasn't Built for Us": queer cisgender women and gender expansive individuals' strategies for navigating reproductive health care. Womens Health Issues 2021;31(5):478-484. [doi: 10.1016/j.whi.2021.06.004] [Medline: 34238669]
42. Fredriksen-Goldsen KI, Simoni JM, Kim HJ, et al. The health equity promotion model: reconceptualization of lesbian, gay, bisexual, and transgender (LGBT) health disparities. Am J Orthopsychiatry 2014 Nov;84(6):653-663. [doi: 10.1037/ort0000030] [Medline: 25545433]
43. George MA. Queering reproductive justice. U Rich L Rev 2020;54(3) [FREE Full text]
44. Loza O, Alvarez CR, Peralta-Torres D. Healthcare and social services providers who serve sexual and gender minorities in a U.S.-Mexico border city. J Immigrant Minority Health 2018 Dec;20(6):1537-1540. [doi: 10.1007/s10903-018-0688-7]
45. Mayer KH, Bradford JB, Makadon HJ, Stall R, Goldhammer H, Landers S. Sexual and gender minority health: what we know and what needs to be done. Am J Public Health 2008 Jun;98(6):989-995. [doi: 10.2105/AJPH.2007.127811] [Medline: 18445789]
46. Russell C. Rights-holders or refugees? do gay men need reproductive justice? Reprod Biomed Soc Online 2018 Nov;7:131-140. [doi: 10.1016/j.rbms.2018.07.001] [Medline: 31061902]
47. Ward BW, Dahlhamer JM, Galinsky AM, Joestl SS. Sexual orientation and health among U.S. adults: national health interview survey, 2013. Natl Health Stat Report 2014 Jul 15(77):1-10. [Medline: 25025690]
48. SuperSaaS. URL: https://www.supersaas.com [accessed 2024-03-07]
49. Katz ML, Kam JA, Krieger JL, Roberto AJ. Predicting human papillomavirus vaccine intentions of college-aged males: an examination of parents' and son's perceptions. J Am Coll Health 2012;60(6):449-459. [doi: 10.1080/07448481.2012.673523] [Medline: 22857137]
50. Thomas PD. A qualitative exploration of pre-exposure prophylaxis (PrEP) initiation decision-making among men who have sex with men (MSM): "It Definitely was A Process". J Natl Black Nurses Assoc 2019 Dec;30(2):10-17. [Medline: 32176964]
51. Glaser BG. The constant comparative method of qualitative analysis. Soc Probl 1965 Apr;12(4):436-445. [doi: 10.2307/798843]
52. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med 2012;22(3):276-282. [doi: 10.11613/BM.2012.031]
53. Technical series on safer primary care: patient engagement. World Health Organization. URL: https://www.who.int/publications-detail-redirect/9789241511629 [accessed 2024-05-03]
54. Park E, Forhan M, Jones CA. The use of digital storytelling of patients' stories as an approach to translating knowledge: a scoping review. Res Involv Engagem 2021 Aug 28;7(1):58. [doi: 10.1186/s40900-021-00305-x] [Medline: 34454604]
55. Jenerette CM, Mayer DK. Patient-provider communication: the rise of patient engagement. Semin Oncol Nurs 2016 May;32(2):134-143. [doi: 10.1016/j.soncn.2016.02.007] [Medline: 27137470]
56. Quinn GP, Sanchez JA, Sutton SK, et al. Cancer and lesbian, gay, bisexual, transgender/transsexual, and queer/questioning (LGBTQ) populations. CA Cancer J Clin 2015;65(5):384-400. [doi: 10.3322/caac.21288] [Medline: 26186412]
57. Fast facts: HIV and gay and bisexual men. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/hiv/data-research/facts-stats/gay-bisexual-men.html [accessed 2024-11-21]
58. Galdas P. Revisiting bias in qualitative research: reflections on its relationship with funding and impact. Int J Qual Methods 2017;16(1). [doi: 10.1177/1609406917748992]
59. Elfil M, Negida A. Sampling methods in clinical research; an educational review. Emerg (Tehran) 2017;5(1):e52. [Medline: 28286859]
60. Jager J, Putnick DL, Bornstein MH. II. More than just convenient: the scientific merits of homogeneous convenience samples. Monogr Soc Res Child Dev 2017 Jun;82(2):13-30. [doi: 10.1111/mono.12296] [Medline: 28475254]

XSL•FO
RenderX

Abbreviations:

DST: digital storytelling
HPV: human papillomavirus
LGBTQ+: lesbian, gay, bisexual, transgender, queer/questioning

Edited by N Cahill; submitted 07.03.24; peer-reviewed by J Loukissas, L Ou; revised version received 10.04.25; accepted 15.04.25; published 30.05.25.

Please cite as:
Darville-Sanders G, Munroe D, Corluyan E, Ikoiwak U, Nguyen J, Mandula C, Thomas P, Sanders B
Components of a Digital Storytelling Intervention for Human Papillomavirus and Cancer Prevention Among LGBTQ+ Individuals: Formative Mixed Methods Inquiry
JMIR Form Res 2025;9:e58163
URL: https://formative.jmir.org/2025/1/e58163
doi:10.2196/58163

© Gabrielle Darville-Sanders, Dominique Munroe, Emilie Corluyan, Utibeabasi Ikoiwak, Jennifer Nguyen, Chad Mandula, Portia Thomas, Brandon Sanders. Originally published in JMIR Formative Research (https://formative.jmir.org), 30.5.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Clinician-Focused Connected Health Requirements Gathering for Attention-Deficit/Hyperactivity Disorder Through Clinical Journey Mapping: Design Science Study

Richard Harris[1], BBus (Hons), MSc; Deirdre Murray[2], MBBS, PhD; Angela McSweeney[3], B Pharm, MBBS, MSc; Frederic Adam[1], MBS, HDR, PhD

[1]Department of Business Information Systems, Cork University Business School, University College Cork, 2.125 O'Rahilly Building, Cork, Ireland
[2]Department of Pediatrics and Child Health, College of Medicine & Health, University College Cork, Cork, Ireland
[3]Health Service Executive, Limerick, Ireland

Corresponding Author:
Richard Harris, BBus (Hons), MSc
Department of Business Information Systems, Cork University Business School, University College Cork, 2.125 O'Rahilly Building, Cork, Ireland

Abstract

Background: Many health care systems globally face severe capacity issues, with lengthening waiting lists and stretched resources. Connected health has been proposed as a game changer for health care. However, the development of connected health apps is difficult and requires multidisciplinary development teams. Patient journey mapping presents an opportunity to streamline the requirements-gathering process for such apps by clearly showing the patient journey to team members who are not familiar with relevant clinical practices. This research project focuses on attention-deficit/hyperactivity disorder (ADHD) as a case study for using clinical journey mapping to represent the "gold standard" care pathway for ADHD treatment; the Dundee Clinical Care Pathway. This pathway was analyzed in detail and was further explored in discussions with stakeholders to produce a patient journey map.

Objective: The objective of this paper is to answer three research questions: (1) visualizing the Dundee ADHD clinical care pathway using integrated patient journey mapping and exploring how its use benefits multidisciplinary development teams; (2) optimizing the integrated patient journey map arising from the Dundee Clinical Care Pathway, in line with the underlying clinical realities of Child and Adolescent Mental Health Service in Ireland; and (3) proposing areas where connected health integration can deliver efficiency and substantial gains for Child and Adolescent Mental Health Services.

Methods: This study uses a design science approach where a sample artifact is presented to a relevant audience for review and feedback and is then leveraged to work iteratively toward an improved, final artifact. This paper presents the feedback collected from both information systems and clinical professionals at each iteration of the map.

Results: This research delivers a comprehensive clinical patient journey map based on the Dundee clinical care pathway. Using unified modeling language concepts and color coding, multiple patient personas are mapped onto a streamlined diagram, allowing the diagram, at an abstract level, to cover the most typical clinical scenarios.

Conclusions: Clinical journey mapping provides a way for team members to get up to speed on clinical practices, while also presenting a way for development teams to identify key gaps where connected health systems can be embedded in clinical pathways to optimize the use of clinical resources and ultimately deliver better patient outcomes.

(JMIR Form Res 2025;9:e53617) doi:10.2196/53617

KEYWORDS

connected health; ADHD; Dundee clinical care pathway; integrated patient journey mapping tool; neurodevelopmental; design science; journey; map; mapping; diagram; visualization; attention deficit; information system; care pathway; design; implementation; integration; attention-deficit/hyperactivity disorder

Introduction

Connected health is defined as "a conceptual model for health management where devices, services, or interventions are designed around the patient's needs, and health-related–data is shared, in such a way that the patient can receive care in the most proactive and efficient manner possible" [1]. It incorporates telehealth and telemedicine, which describe situations where health care is provided without patients being necessarily face-to-face with the clinician, using technologies

XSL•FO
RenderX

such as voice and video chat to provide services. Connected health systems have been shown to produce positive outcomes in terms of both treatment and participant satisfaction in many areas, including mental health services [2].

Ensuring that young patients have timely access to high-quality mental health services is critical, as adverse childhood experiences have a disproportionate impact when left untreated [3]. Yet, Child and Adolescent Mental Health Service (CAMHS) is struggling to meet demand. Reports indicate an increasing number of adolescents presenting to CAMHS for treatment. Figures from the United Kingdom show that the number of children with a diagnosed "mental disorder" is growing [4-7], with waiting lists for CAMHS running in months [8]. Connected health systems could play a key role in supporting CAMHS. As part of its action plan to modernize CAMHS in Ireland, the Maskey report [9] proposes to "explore the options available in Information and communication technology to improve the governance, effectiveness, efficiency, and accessibility of CAMHS," namely by implementing electronic health records, electronic questionnaires, and generally, telemedicine on a broader scale [9]. This highlights the opportunity to embed connected health within CAMHS to achieve their digital transformation.

A critical step in ensuring the success of a connected health system is having a clear understanding of the experiences of patients during their treatment, the "pain points" they face, and where connected health systems can have an impact [10]. To capture the patient journey and the digitization opportunities within it, a visual representation is beneficial [11,12]. The "integrated patient journey mapping tool" (IPJM) [13], designed to support "multidisciplinary practitioners in designing health care solutions that meet the demands of existing constraints, performance improvement, and patient experience" [13,14] has been adopted in many studies. This paper builds on IPJM, using a design science approach to capture the Dundee Clinical Care Pathway [15], which is commonly used to manage attention-deficit/hyperactivity disorder (ADHD). ADHD is a common neurodevelopmental disorder, diagnosed in approximately 5% of children [16,17].

A key goal of this paper is, through patient journey mapping, to highlight areas where connected health systems could be embedded within the Dundee pathway. The Dundee pathway, originally developed in Scotland, creates a "clearly structured, evidence-based clinical pathway for the assessment and management of children and adolescents with ADHD" [15]. It has been identified in a systems review [9] as the preferred construct for the delivery of ADHD care and it is used in this paper as an initial template.

Methods

Research Questions

This paper proposes three research questions to effectively leverage the IPJM within the context of the Dundee pathway:

- Visualize the Dundee ADHD clinical care pathway using IPJM and explore how the use of IPJM benefits multidisciplinary development teams.
- Optimize the integrated patient journey map arising from the Dundee Clinical Care Pathway, in line with the underlying clinical realities of CAMHS in Ireland.
- Propose areas where connected health integration can deliver efficiency and substantial gains for CAMHS services.

The first research question justifies the use of IPJM to map the Dundee pathway, in the specific context of collaborative connected health systems development. The second question helps fine-tune the pathway, based on the accounts and experiences of key stakeholders toward improving its legibility [14], and the final research question looks to select specific steps and activities within the pathway where connected health can be leveraged to improve its efficiency.

Design Science Process

To populate the IPJM with personas of individuals based on the Dundee Clinical Care Pathway and to validate it, it is important to use a robust and rigorous approach that includes input from clinicians with various backgrounds, as well as academics from relevant disciplines, to ensure that the proposed visualization is correct and respects the guidelines of the model pathway. To achieve the desired result, a design science approach [18] was used to create and iterate through successive versions of the patient journey map artifact (PJMA). Design science is an approach involving "the creation of an artifact and/or design theory as a means to improve the current state of practice as well as existing research knowledge" [19]. Design science focuses on the development of a new artifact (eg, a diagram, app, or graph), as well as the study of the use of the newly created artifact. The research output for this project consists of this research paper which highlights the design of a diagram to visually illustrate the Dundee ADHD Clinical Care pathway, and the artefactual output consisting of the final diagram presented in Figure 1.

XSL•FO
RenderX

Figure 1. Third iteration of the PJMA, implementing feedback received from iteration two. Color coding changed to improve contrast and discharge point added, among other changes.

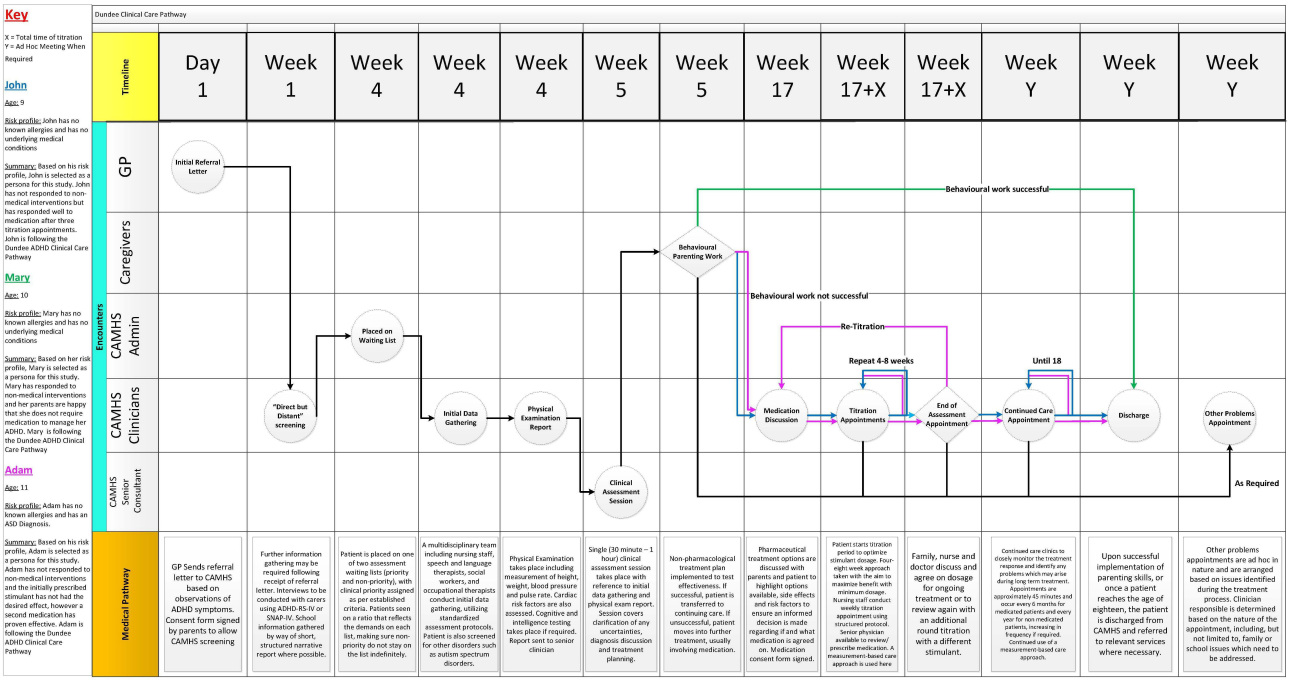

In rigorous application of the design science method, the design stages of the artifact, including clear illustrations of all iterations, with the feedback of the participants and the resulting alterations, are presented in detail. This ensures that the final design of the PJMA has been validated through a rigorous process and that the PJMA is a suitable representation of the Dundee ADHD Clinical Care Pathway.

Design science was chosen as the methodology of choice because (1) it is a generally accepted methodology within the information systems (IS) field that has proven to produce academically rigorous results, (2) it promotes an iterative process that permits feedback from various stakeholders to be used to refine the artifact design, similar to other research methodologies (eg, the Delphi study method [20]), and (3) it offers transparency in the development of the PJMA as each version is documented and presented as part of the final publication [21].

Interviews were conducted with individuals whose professional experience was relevant to the research area, these include information technology consultants, CAMHS clinicians, and IS researchers. Table 1 gives details of the professionals interviewed during the process. As the area of connected health within ADHD is a novel one, the use of qualitative data gathering and analysis in parallel with design science was viewed as a robust choice to build the ADHD case study, and subsequently, propose a novel theory [22].

Table . Total number of interviews by category.

Interview category	Number of interviews
CAMHS[a] clinician	3
CAMHS nurse	1
Dundee pathway expert	1
Social worker	1
Connected health researcher	2
Connected health consultant	2
Total	10

[a]CAMHS: Child and Adolescent Mental Health Service.

The results of each step within the design science research methodology [18] process for the development of the PJMA can be found in Figure 2, starting with the identification of the problem and objectives of the solution. Each iteration receives its own design and development, demonstration, and evaluation subsection, in line with design science research principles [18,19].

XSL•FO
RenderX

Figure 2. Design science research methodology, highlighting problem statement, solution objectives, and iterative development approach. DSRM: design science research methodology.

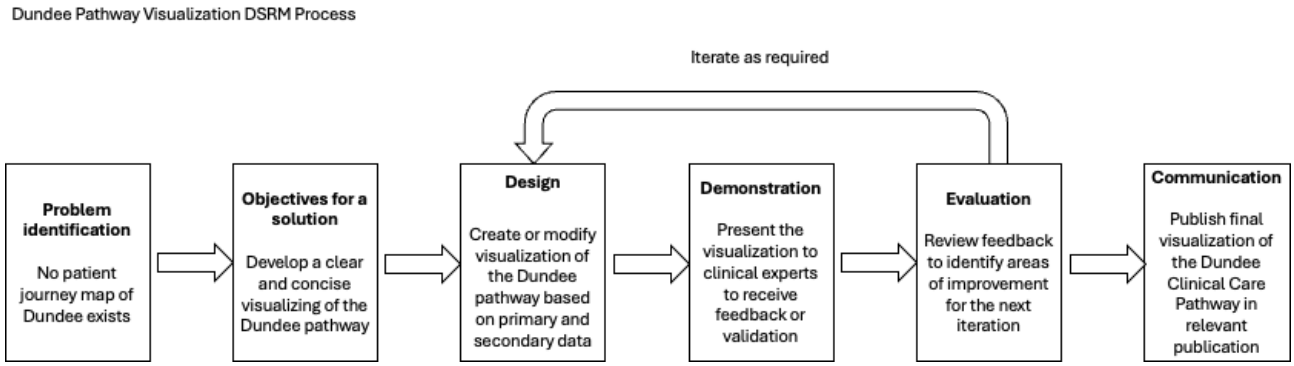

The objective of this research project is to create an accurate visual representation of the Dundee ADHD Clinical Care Pathway, which can be used as a visual aid for multidisciplinary teams who develop connected health apps to support clinicians in the treatment of ADHD. For the visual aid to be successful in conveying the Dundee pathway to viewers, it must do the following: (1) be true and accurate to the pathway as described by Coghill and Seth [15], (2) effectively convey the clinical journey through the pathway to individuals who have little to no clinical experience in treating or working with individuals diagnosed with ADHD, and (3) create an evaluation of IPJM using IS design principles to create a PJMA.

The three iterations conducted as part of this research project are presented in the following sections.

Ethical Considerations

Approval was sought for this research project from University College Cork's Social Research and Ethics Committee which approved both application Log 2022‐134 entitled "Connected Health Avenues for Adolescent Mental Health" on July 19, 2022, and subsequent amendment "2022‐134A" Informed consent was given by all participants prior to the commencement of the interviews. This was achieved by way of an information leaflet highlighting the study rationale, how findings would be used, and how the data would be stored and processed. Participants were then asked to sign a consent form to ensure they had read the information leaflet and were happy to participate. Due to ethical concerns, patients were not interviewed at this point of the project. CAMHS deals with adolescents and children who are receiving mental health care and therefore are classed as extremely vulnerable. This research has focused on mapping the pathway from a clinician's view.

Results

Iteration One

Design and Development (Iteration One)

In order to create the initial visualization, two unstructured interviews were conducted with a CAMHS clinician and nurse with experience in implementing the Dundee pathway. These interviews allowed the gathering of key data to assist in the development of the first iteration. Data collected in these interviews, along with a review of the literature, led to the creation of Figure 3, the first step in the process of developing a robust and verified PJMA.

XSL•FO
RenderX

Figure 3. First iteration of the PJMA, visualizing the Dundee clinical care pathway as per interview and review of literature. Images under "patient journey" provided by Microsoft Office stock images, for illustrative purposes only. PJMA: patient journey map artifact.

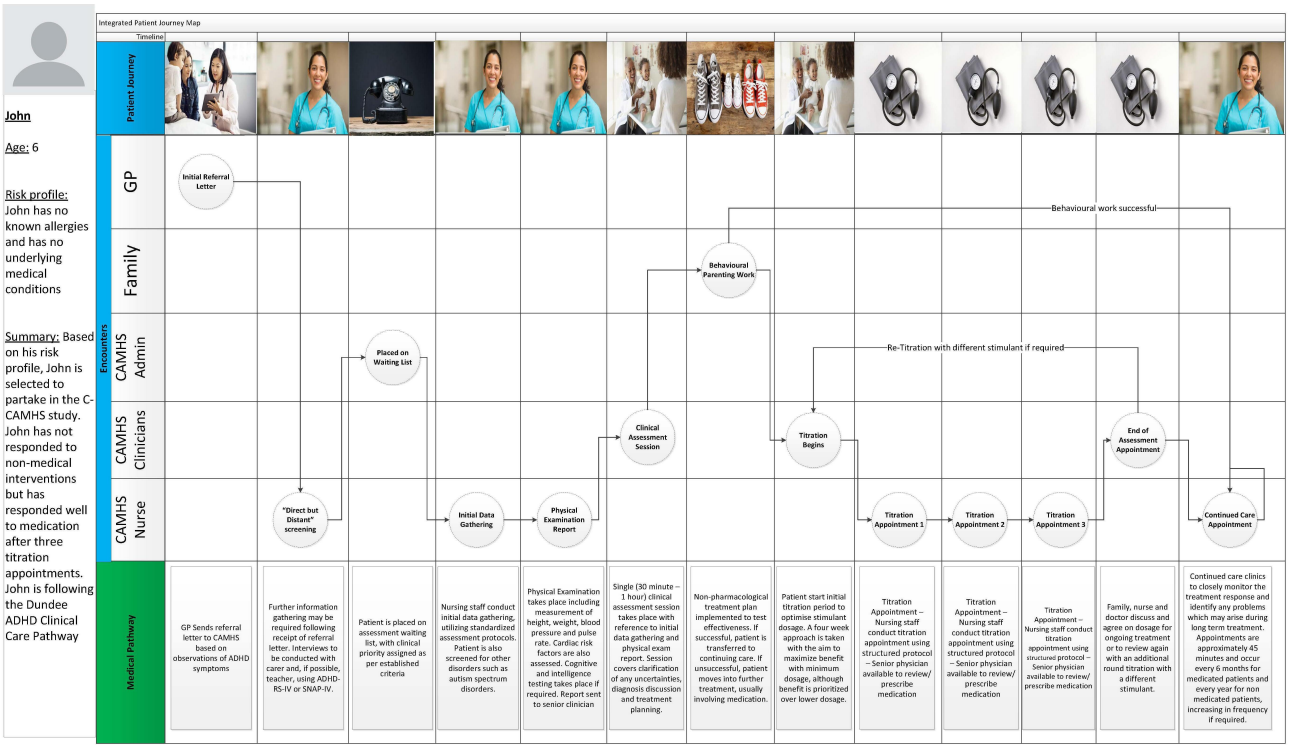

The PJMA consists of a number of key areas:

- The left section of the pathway shows the patient persona, including the persona name, age, a brief risk profile, and a summary of the patient's journey.
- The center section of the journey map highlights encounters that the patient will have with different actors in the pathway and each point along the journey represents an interaction.
- The section at the bottom highlights details for each interaction to provide greater context.

Each interaction leads to another, with two lines arising from "Behavioral Parenting work" to highlight that, if this intervention is successful, the patient may jump straight to the "Continuing Care Appointment" step. If unsuccessful, the pathway continues to titration [15]. The continuing care appointment represents a continuous cycle where periodic check-ins are conducted to ensure the patient remains happy with their treatment. A fork marks the "end of assessment appointment," noting if the prescribed medication does not have the desired effect, additional rounds of titration are conducted with alternative medications.

Iteration one shows a single persona, but several personas were identified from interviews with CAMHS clinicians, with the persona in iteration one "John" representing a fictitious individual who would engage with every stage of the Dundee pathway. John has no known allergies and no underlying medical issues. John is therefore a suitable persona to illustrate the Dundee pathway. As John has no other medical concerns, it is highly likely that John can progress through the Dundee Clinical Care Pathway as described in the literature, without the need for additional interventions. John does not respond to behavioral techniques but does respond well to the first stimulant tested during titration.

Demonstration or Evaluation (Iteration One)

Demonstration of the first iteration of the PJMA, shown in Figure 3, was completed by presenting it to an expert of the original Dundee publication [15]. It offered an excellent opportunity to ensure the first iteration of the map was true to the original publication. Valuable feedback was received by the expert and feedback was generally positive.

Based on the feedback, it was decided that a second iteration was needed and that the feedback should be incorporated into it. General feedback highlights that the PJMA should be streamlined, pictures removed, and ad hoc appointments should be illustrated in subsequent iterations. Table 2 lists the interviewees for this iteration.

Table . Interview participant from iteration one by category.

Interview category	Number of interviews
Senior CAMHS[a] clinician	1
CAMHS nurse	1
Dundee pathway expert	1

[a]CAMHS: Child and Adolescent Mental Health Service.

XSL•FO
RenderX

Iteration Two

Design and Development (Iteration Two)

Creating the second iteration of the PJMA presented the opportunity to include multiple personas in the diagram by implementing unified modeling language (UML) principles to use the encounters as divergence points, or decision nodes, for each persona. Three personas are presented in the second iteration, representing three common patient scenarios, John remains as per iteration one, with Mary and Adam added as new personas.

- Mary differs from John as she has responded to nonpharmacological interventions and her parents agree that she does not require medication to manage her ADHD.
- Adam differs from John as he has not responded to nonpharmacological interventions and the initially prescribed stimulant has not had the desired effect, however, a second stimulant is effective.

To differentiate between each of the personas, color coding is used to represent each individual, with John represented in blue, Mary in pink, and Adam in green. Lines in black represent all three personas as the first number of encounters are the same for each persona.

Mary's divergence point is during the "Behavioral work" encounter, as this was deemed effective for her, and she moved straight to "Continued Care." Adam and John continued to titration where John was successful with his first stimulant. Adam required a second round of titration as the first stimulant did not have the desired outcome. This is illustrated by "Retitration." All three personas finally reach "Continued Care," which is illustrated by a circular arrow as this is a stage where patients are monitored continually until the age of 18 years.

At all relevant stages, an "as required" branch extends out into an "Other Problems Appointment" illustrating ad hoc appointments that may be necessary to discuss problems during treatment which may or may not be directly related to the treatment plan but may still be required for the patient's well-being.

For iteration two, images were removed from the diagram and replaced with a timeline section. This was identified during the semistructured interviews and represents a best-case scenario for the pathway. Two "wildcards" are included in the timeline. X represents the total number of weeks an individual is undergoing titration, which varies for each patient, and Y represents an unspecified as or when-required timeframe. Encounters have also been updated as per feedback from clinicians. Figure 4 highlights iteration two in detail.

Figure 4. Second iteration of the PJMA, implementing feedback received from iteration one. Removal of images and ad hoc appointments added, among other changes. PJMA: patient journey map artifact.

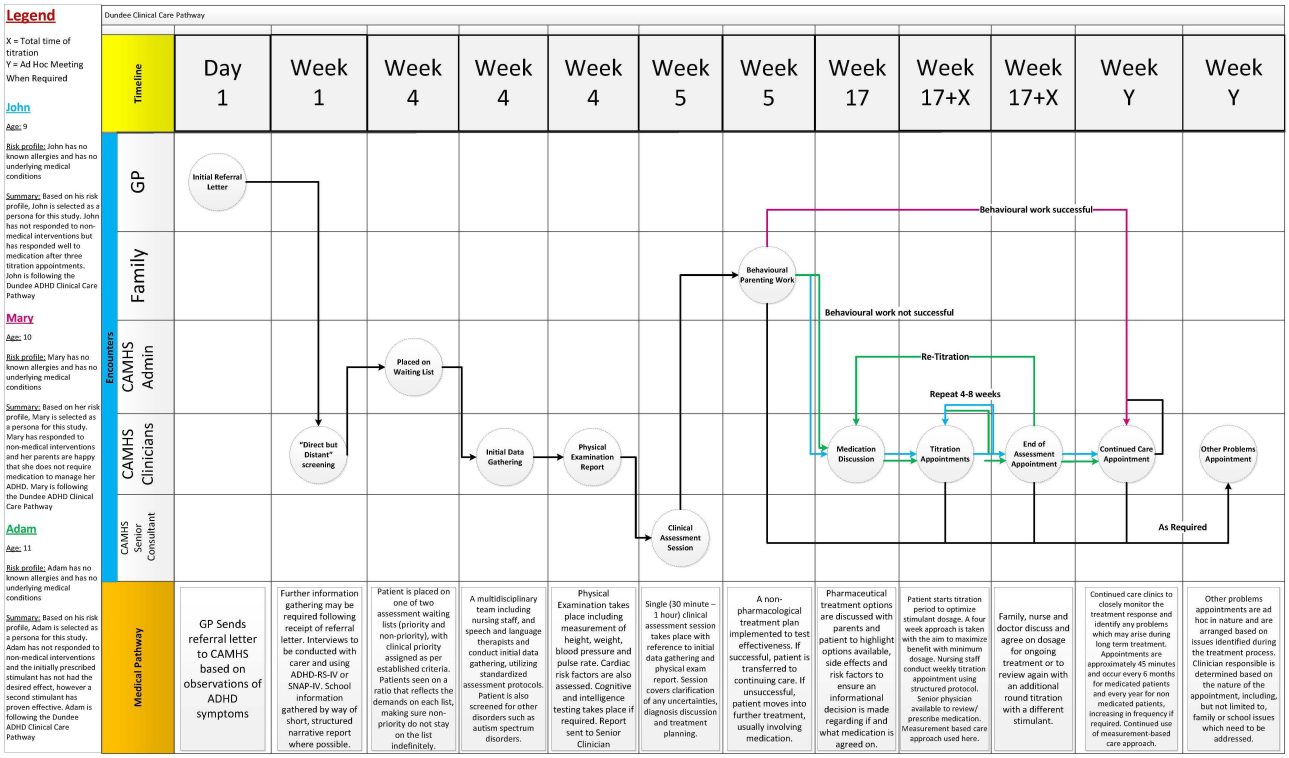

Evaluation—Iteration Two

Evaluation of iteration two took place in two phases. Phase one was clinical verification to ensure the updated visualization was still an accurate representation of the pathway, with phase two involving four interviews of connected health consultants or academics to discuss the merits and drawbacks of the visualization for use as part of a requirements-gathering toolset. Table 3 lists the interviewees for this iteration.

XSL•FO
RenderX

Table . Interview participant from iteration two by category.

Interview category	Number of interviewees
CAMHS[a] clinician	2
Social worker	1
Connected health researcher	2
Connected health consultant	2

[a]CAMHS: Child and Adolescent Mental Health Service.

Clinical accuracy validation of iteration two came in the form of interviews with two CAMHS clinicians, and a social worker, all of whom have experience working with or treating those with ADHD. Overall feedback on the diagram was very positive, with clinicians praising its accuracy and clarity. A key theme that was mentioned throughout the interviews was the multidisciplinary nature of CAMHS teams. Several clinicians, including nurses, consultants, occupational therapists, speech and language therapists, and social workers are involved in the screening and data-gathering process. It is important that these stakeholders are represented in the diagram to ensure their needs are met by future connected health apps.

The management of consent is another key concern. It is critical that informed consent is received at the point of referral to ensure that the CAMHS team is given permission to begin the screening process as this involves contacting multiple stakeholders in the patient's life, including but not limited to, teachers, carers, parents, or any other relevant caregiver. This consent form will need to be well documented in any connected health app. A medication consent form will also need to be produced and signed by parents at the "Medication Discussion" stage as this protects all parties involved in the treatment process. This medication consent form needs to be captured in the diagram.

When discussing "behavioral parenting work," it was noted that many parents do not decide to go down this treatment route as they may have already attempted this and concluded that this approach is not a suitable solution. As a result, parents have attended their general practitioner for a referral to CAMHS with the view of beginning pharmacological treatment. This is not, however, the case for all patients, and behavioral parenting work may still be a very effective solution when parents are provided with appropriate teaching materials and literature.

When discussing the "Other Problems Appointment" encounter, clinicians advised that this appointment generally occurs during treatment, either pharmacological or nonpharmacological but may occur at any stage of the pathway. Risk is the key driver of these appointments with the type and severity of the risk dictating who the patient is referred to and how quickly. The health and safety of the patient is always the key priority.

It was noted that, if "Behavioral Parenting Work" was successful, the patient is discharged from the services instead of being placed on a continued care plan. When discharged in this manner, patients can be referred to the service again, subject to a shorter waiting period.

It was noted that patients presenting to CAMHS often have comorbidities such as autism spectrum disorder (ASD), which may inhibit the effectiveness of stimulant medications.

Clinicians are also advised on the positioning of encounters based on their own clinical practice, with some screening or examinations taking place before or after others depending on the system in place within each specific service. It was, however, proposed that the relevant examinations would be identified within the diagram based on this feedback.

Following on from clinical verification of the diagram, the diagram was reviewed by four experts, consisting of connected health consultants and academics to provide feedback and help to identify possible areas for improvement. This yielded some very positive feedback, as well as very useful suggestions.

The colors for Adam and John's personas were judged to be too hard to distinguish as the shades of blue and green were quite similar. The importance of keeping the visualization to a one-page size was emphasized, as a "one-pager" was deemed to be optimal from an information conveyance viewpoint. It was noted that some regulatory concerns pertaining to connected health apps were missing from the diagram. It was suggested that the map could be turned into a multilayered system, where additional information is overlaid, similar to "acetate sheets." It was also noted that this iteration was quite clinically focused, and there was no mapping of the patient's "emotional journey" through the service. The use of UML shapes was suggested to highlight where decisions are being made that may change the course of a patient's journey. The benefits of visualizing the pathway to identify gaps where connected health can be embedded were emphasized. It was finally noted that images may be beneficial in visualizing who is responsible for each encounter.

Iteration Three

Design and Development

Based on the feedback received from the evaluation of iteration two, it was decided that a third iteration of the visualization would be developed. This iteration, shown in Figure 1, has been updated to include the following changes.

The Adam persona has been updated to include a comorbid ASD diagnosis. This reflects the commonality of this comorbid diagnosis and the likelihood that ASD will impact stimulant performance. The colors of Adam and Mary have been switched to increase the contrast between John and Adam as both occupy similar locations on the diagram. A diamond shape is now used to illustrate where an individual patient's journey may diverge

XSL•FO
RenderX

based on clinical decision-making. A discharge encounter is added, and an "until 18" note is added to the continued care loop to illustrate that patients are discharged after effective behavioral work, or when patients become 18 years old. The medical pathway section was also updated to include relevant forms and the multidisciplinary nature of the data-gathering process. The legend was made more visible to highlight its importance. "Family" encounters were updated to "Caregivers" to include teachers, childminders, etc.

Demonstration or Evaluation (Iteration Three)

Demonstration of iteration three took the form of an email debrief containing Figure 1 to ensure that all feedback had been taken on board and that participants had the opportunity to provide final feedback. Overall feedback on the third iteration was very positive and participants highlighted their anticipation of the full publication of the work. This feedback gave the research team the confidence that the next step for this PJMA is to move on to the "Communication" phase of the design science methodology noted above and compile all iterations, feedback, and results into a research piece for consumption by the wider community. Table 4 notes the breakdown of individuals receiving a debrief email.

Table . Table of debriefed participants by category.

Participant category	Number of interviewees
Senior CAMHS[a] clinician	1
CAMHS nurse	1
Dundee pathway expert	1
CAMHS clinician	2
Social worker	1
Connected health researcher	2
Connected health consultant	2
Total	10

[a]CAMHS: Child and Adolescent Mental Health Service.

Discussion

Main Findings

Feedback from clinicians and IS professionals suggests that using the IPJM mapping tool in Figure 1, effectively conveys the Dundee pathway to nonclinical audiences, enabling them to assimilate the information and subsequently facilitating more in-depth and focused conversations with clinicians when identifying pain points and opportunities for connected health adoption. This can then inform policy making in the health care area.

Based on the feedback from Clinicians and IS professionals, the modified IPJM diagram was well received by both groups. One may infer that visualizing the Dundee pathway is of benefit to stakeholders who are of a nonclinical background, as a visual aid to bring them up to speed on how the Dundee Clinical Care Pathway operates. The multidisciplinary nature of connected health apps requires input and contributions of individuals outside the clinical field, therefore, ensuring all team members working on a connected health project are aware of the standard care pathway. Figure 1 illustrates the Dundee pathway on an integrated patient journey map, the accuracy of which is verified by clinicians working in the field. This journey map may serve as a starting point for future research in connected health app design and development by effectively illustrating the Dundee pathway to nonclinical connected health team members.

As part of our research, we considered how the IPJM tool can be modified to enhance its accessibility. Using UML techniques, as reported by Curry et al [12], along with color coding, has allowed multiple personas to be layered onto one map, facilitating a more condensed, yet detailed, view of the pathway. UML also enables the presentation of clear focal points where clinical decisions divert individual persona's trajectory on the pathway based on typical individual patient needs or circumstances.

Those interviewed noted that the one-page nature of the diagram is extremely powerful when conveying the details of the Dundee pathway. It reflects the full complexity of care to anyone working on a potential connected health design and implementation project. For instance, the single-page nature of the diagram supports robust conversations between clinicians and IS consultants to effectively identify possible areas for connected health adoption. One critical element of this is the added confidence it gives participants in terms of "sign of," before substantial development begins. McCarthy et al [23] have identified the frequent changes in the path that occur during the development of IT solutions that address wicked problems in clinical care. These path changes are often disruptive and divisive for the development team as participants blame each other for misunderstanding the requirements. The robustness and ambiguity-removing effect of IPJM can secure a more reliable capture of requirements to support app development.

Figure 1 identifies clear areas where connected health can be effectively embedded. The "Discreet but Distant Screening" phase presents a clear opportunity to move from a paper-based approach to a software-assisted approach. The paper-based screening process takes a substantial amount of time, time which

XSL•FO
RenderX

could be saved so clinicians can see more patients whilst a portal or an app does the computations of the scores automatically for them and provides full visibility on all cases in the waiting list, particularly the most acute ones.

During the "behavioral parenting work" phase, a connected health app could provide caregivers with relevant information in the form of text, audio, and video supports in sync with each relevant phase of the treatment. An app may also be interactive and collect feedback from caregivers and feed this information back to clinicians, who can see which techniques yield positive results and those that do not. This would allow clinicians to receive a clear and detailed picture of what is effective and subsequently allow them to make more informed decisions and provide more pertinent, faster advice.

The "Continued Care Appointments" phase presents an opportunity to use remote patient monitoring using Bluetooth-connected devices such as blood pressure monitors and connected scales. These devices may be connected to the app provided to each patient (eg, loaded on their phone or that of their caregiver) which will subsequently send the information to the clinician portal, eliminating the need for manual entry, and thus, eliminating the risk of human error when inputting results. Similar apps have been developed in the obstetrics area which showed great potential for the early detection of such conditions as preeclampsia [13,14].

Figure 5 proposes a generic architecture model for potential future connected health apps in the ADHD area, which abstracts and generalizes the above recommendations. The top layer of this diagram illustrates the interface layer used by patients and caregivers, either by way of a web portal or a mobile-connected health app. The second layer represents the operational areas for connected health apps, specifically within the home, in the community, or in a ward setting. The third level highlights two distinct touchpoints: the first being the patient-facing app, which gathers vital information from both home and community settings and feeds this data into the connected health architecture; the second touchpoint being the clinician support system, operating in community and ward environments, which assists medical professionals in making informed decisions based on data collected by and stored in the connected health architecture. Finally, the base layer in the model houses the model base (eg, algorithms) and database of the connected health app. This layer includes digital health records, questionnaire results, blood pressure readings, and patient information resources. Some rule-based engines and artificial intelligence components could also be located there to provide a higher grade of decision support to clinicians.

Figure 5. An architecture model for potential connected health apps in the ADHD area (adapted from Harris et al [24]). ADHD: attention-deficit/hyperactivity disorder.

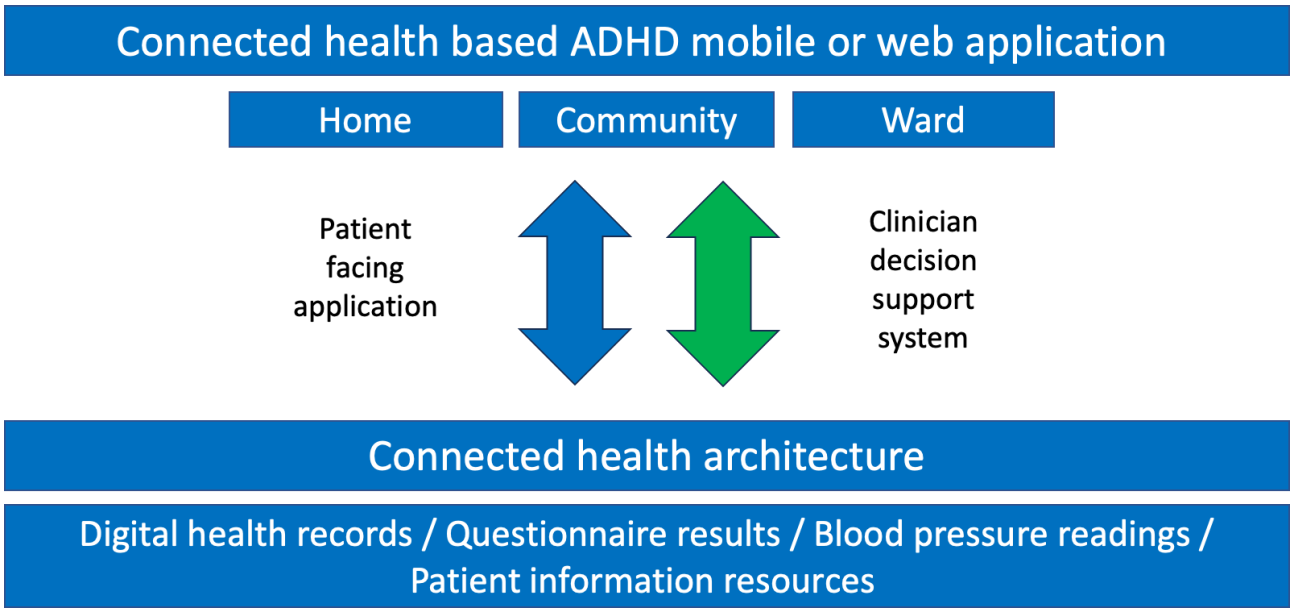

Limitations and Future Research

The PJMA diagram is a clinical pathway representation, and as such, focuses on the pathway from the point of view of the clinician. There is certainly scope for future research focusing on the pathway from the patient's perspective. Individuals who are using CAMHS services for ADHD treatment fall under multiple vulnerable categories as they are children or adolescents who are receiving mental health services. Substantial ethical concerns exist with interviewing such individuals, which is why we focused on the clinician and information technology side of the process. Our research was formative in nature, but further work in this domain from the point of view of the patient is required which will deliver additional insights into connected health apps for CAMHS.

Conclusions

The multidisciplinary nature of connected health projects requires tools and techniques to be made available to allow nonclinical members of the team to effectively get acquainted with the clinical requirements of each particular project. Clinical journey mapping provides a way for team members to get acquainted with clinical practices, and also present their teammates with ways to identify key gaps where connected health systems can be embedded in clinical pathways in order

XSL•FO
RenderX

to facilitate better use of clinical resources and ultimately deliver better patient outcomes.

Continued research in this area will help deliver continued progress in youth mental health, as well as health care in general, toward the well-being of future generations. Specifically, research into the design and implementation of connected health systems that can become embedded into the Dundee ADHD treatment pathway will be critical for the continued development of, and success of, Youth mental health services serving those with ADHD around the world. Research on software-based screening questionnaires, remote vital sign monitoring, and software-based teaching materials has the potential to support those with ADHD and their families.

Acknowledgments

This work has been written with the assistance of AudioPen (AudioPen Inc; account type: prime; 2022 - 2024), Grammarly (Grammarly Inc; 2022 - 2024), and ChatGPT (OpenAI; account type: plus; 2025). Original transcripts from ChatGTP are available in Multimedia Appendix 1.

Data Availability

Data sharing is not applicable to this article as no datasets were generated or analyzed during this study.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Original ChatGPT Transcripts
[DOCX File, 16 KB - formative_v9i1e53617_app1.docx]

References

1. Caulfield BM, Donnelly SC. What is Connected Health and why will it change your practice? QJM 2013 Aug;106(8):703-707. [doi: 10.1093/qjmed/hct114] [Medline: 23676416]
2. Proudfoot J, Ryden C, Everitt B, et al. Clinical efficacy of computerised cognitive-behavioural therapy for anxiety and depression in primary care: randomised controlled trial. Br J Psychiatry 2004 Jul;185:46-54. [doi: 10.1192/bjp.185.1.46] [Medline: 15231555]
3. Asmussen K, McBride T, Waddell S. The potential of early intervention for preventing and reducing ACE-related trauma. Soc Policy Soc 2019 Jul;18(3):425-434. [doi: 10.1017/S1474746419000071]
4. NHS Digital. Mental health of children and young people in england, 2017—summary of key findings. : NHS Digital; 2018 URL: https://files.digital.nhs.uk/A6/EA7D58/MHCYP%202017%20Summary.pdf [accessed 2025-04-08]
5. NHS Digital. Mental health of children and young people in england, 2020—wave 1 follow up to the 2017 survey. : NHS Digital; 2020 URL: https://digital.nhs.uk/data-and-information/publications/statistical/mental-health-of-children-and-young-people-in-england
6. NHS Digital. Mental health of children and young people in england, 2020—wave 2 follow up to the 2017 survey. : NHS Digital; 2021 URL: https://digital.nhs.uk/data-and-information/publications/statistical/mental-health-of-children-and-young-people-in-england/2021-follow-up-to-the-2017-survey#
7. Newlove-Delgado T, Marcheselli F, Williams T, et al. Mental health of children and young people in england, 2022. : NHS Digital; 2022 URL: https://digital.nhs.uk/data-and-information/publications/statistical/mental-health-of-children-and-young-people-in-england/2022-follow-up-to-the-2017-survey
8. Child and adolescent mental health services annual report 2012-2013. : Health Service Executive; 2013 URL: https://www.hse.ie/eng/services/publications/mentalhealth/camhs12,13.pdf
9. Maskey S. Report on the look-back review into child & adolescent mental health services county MHS area a. : HSE; 2022 URL: https://www.hse.ie/eng/services/news/newsfeatures/south-kerry-camhs-review/report-on-the-look-back-review-into-camhs-area-a.pdf [accessed 2025-04-08]
10. Carroll N. Key success factors for smart and connected health software solutions. Computer (Long Beach Calif) 2016;49(11):22-28 [FREE Full text] [doi: 10.1109/MC.2016.340]
11. Simonse L, Albayrak A, Starre S. Patient journey method for integrated service design. Des Health 2019 Jan 2;3(1):82-97. [doi: 10.1080/24735132.2019.1582741]
12. Curry J, McGregor C, Tracy S. A communication tool to improve the patient journey modeling process. Presented at: 2006 International Conference of the IEEE Engineering in Medicine and Biology Society; Aug 30 to Sep 3, 2006; New York, NY, USA. [doi: 10.1109/IEMBS.2006.4398507]

XSL•FO
RenderX

13. McCarthy S, O'Raghallaigh P, Woodworth S, Lim YL, Kenny LC, Adam F. An integrated patient journey mapping tool for embedding quality in healthcare service reform. J Decis Syst 2016 Jun 10;25(sup1):354-368. [doi: 10.1080/12460125.2016.1187394]
14. McCarthy S, O'Raghallaigh P, Woodworth S, Lim YY, Kenny LC, Adam F. Embedding the pillars of quality in health information technology solutions using "Integrated Patient Journey Mapping" (IPJM): case study. JMIR Hum Factors 2020 Sep 17;7(3):e17416. [doi: 10.2196/17416] [Medline: 32940610]
15. Coghill D, Seth S. Effective management of attention-deficit/hyperactivity disorder (ADHD) through structured re-assessment: the Dundee ADHD Clinical Care Pathway. Child Adolesc Psychiatry Ment Health 2015;9:52. [doi: 10.1186/s13034-015-0083-2] [Medline: 26587055]
16. World Health Organization. International statistical classification of diseases and related health problems (11th ed.). World Health Organization. 2019. URL: https://icd.who.int/en/ [accessed 2022-12-26]
17. Polanczyk G, de Lima MS, Horta BL, Biederman J, Rohde LA. The worldwide prevalence of ADHD: a systematic review and metaregression analysis. Am J Psychiatry 2007 Jun;164(6):942-948. [doi: 10.1176/ajp.2007.164.6.942] [Medline: 17541055]
18. Peffers K, Tuunanen T, Rothenberger MA, Chatterjee S. A design science research methodology for information systems research. J Manage Inf Syst 2007 Dec;24(3):45-77. [doi: 10.2753/MIS0742-1222240302]
19. Vaishnavi V, Kuechler B. Design research in information systems. Association for Information Systems. 2021. URL: https://www.desrist.org/design-research-in-information-systems/ [accessed 2022-12-26]
20. Keeney S, McKenna H, Hasson F. The Delphi Technique in Nursing and Health Research: John Wiley & Sons, Ltd; 2010.
21. Hevner AR, March ST, Park J, Ram S. Design science in information systems research. MIS Q 2004;28(1):75. [doi: 10.2307/25148625]
22. Eisenhardt KM. Building theories from case study research. Acad Manage Rev 1989;14:532-550. [doi: 10.2307/258557]
23. McCarthy S, O'Raghallaigh P, Li Y, Adam F. Control enactment in context: understanding the interaction of controlee and controller perceptions in inter‐organisational project teams. Inf Syst J 2023 Sep;33(5):1029-1084. [doi: 10.1111/isj.12434]
24. Harris R, Dempsey E, Murray D, Woodworth S, O'Raghallaigh P, Adam F. Towards a blueprint for decision support in connected health: scenarios in maternal and child health. J Decis Syst 2022 Dec 15;31(sup1):60-76. [doi: 10.1080/12460125.2022.2083049]

Abbreviations

ADHD: attention-deficit/hyperactivity disorder
ASD: autism spectrum disorder
CAMHS: Child and Adolescent Mental Health Services
IPJM: integrated patient journey mapping
IS: information systems
PJMA: patient journey map artifact
UML: unified modeling language

Edited by A Mavragani; submitted 12.10.23; peer-reviewed by D Chrimes, N Carroll; revised version received 02.12.24; accepted 02.12.24; published 26.05.25.

Please cite as:
Harris R, Murray D, McSweeney A, Adam F
Clinician-Focused Connected Health Requirements Gathering for Attention-Deficit/Hyperactivity Disorder Through Clinical Journey Mapping: Design Science Study
JMIR Form Res 2025;9:e53617
URL: https://formative.jmir.org/2025/1/e53617
doi:10.2196/53617

© Richard Harris, Deirdre Murray, Angela McSweeney, Frederic Adam. Originally published in JMIR Formative Research (https://formative.jmir.org), 26.5.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Application of the Bidirectional Encoder Representations from Transformers Model for Predicting the Abbreviated Injury Scale in Patients with Trauma: Algorithm Development and Validation Study

Jun Tang[1*], MEng; Yang Li[2,3*], MD; Keyu Luo[4], MD; Jiangyuan Lai[5], BM; Xiang Yin[4], MD; Dongdong Wu[1], PhD

[1]Department of Information, Daping Hospital, Army Medical University, No.10 Daping Changjiang Branch Road, Yuzhong District, Chongqing, China

[2]Department of Emergency Medicine, Medical Center of Trauma and War Injury, Daping Hospital, Army Medical University, Chongqing, China

[3]National Key Laboratory of Trauma and Chemical Poisoning, Chongqing, China

[4]Department of Orthopedics, Daping Hospital, Army Medical University, No.10 Daping Changjiang Branch Road, Yuzhong District, Chongqing City, Chongqing, China

[5]Department of Traumatic Surgery, School of Basic Medicine, Army Medical University, Chongqing, China

[*]these authors contributed equally

Corresponding Author:
Dongdong Wu, PhD
Department of Information, Daping Hospital, Army Medical University, No.10 Daping Changjiang Branch Road, Yuzhong District, Chongqing, China

Abstract

Background: Deaths related to physical trauma impose a heavy burden on society, and the Abbreviated Injury Scale (AIS) is an important tool for injury research. AIS covers injuries to various parts of the human body and scores them based on the severity of the injury. In practical applications, the complex AIS coding rules require experts to encode by consulting patient medical records, which inevitably increases the difficulty, time, and cost of evaluation of patient and also puts higher demands on the workload of information collection and processing. In some cases, the sheer number of patients or the inability to access detailed medical records necessary for coding further complicates independent AIS codes.

Objective: This study aims to use advanced deep learning techniques to predict AIS codes based on easily accessible diagnostic information of patients to improve the accuracy of trauma assessment.

Methods: We used a dataset of patients with trauma (n=26,810) collected by the Chongqing Daping Hospital between October 2013 and June 2024. We mainly selected the patient's diagnostic information, injury description, cause of injury, injury region, injury types, and present illness history as the key feature inputs. We used a robust optimization Bidirectional Encoder Representations from Transformers (BERT) pretraining method to embed these features and constructed a prediction model based on BERT. This model aims to predict AIS codes and comprehensively evaluate its performance through a 5-fold cross-validation. We compared the BERT model with previous research results and current mainstream machine learning methods to verify its advantages in prediction tasks. In addition, we also conducted external validation of the model using 244 external data points from the Chongqing Emergency Center.

Results: The BERT model proposed in this paper performs significantly better than the comparison model on independent test datasets with an accuracy of 0.8971, which surpassed the previous study by 10 % points. In addition, the area under the curve (AUC value of the BERT model is 0.9970, and the F_1-score is 0.8434. In the external dataset, the accuracy, AUC, and F_1-score results of the model are 0.7131, 0.8586, and 0.6801, respectively. These results indicate that our model has high generalization ability and prediction accuracy.

Conclusions: The BERT model we proposed is mainly based on diagnostic information to predict AIS codes, and its prediction accuracy is superior to previous investigations and current mainstream machine learning methods. It has a high generalization ability in external datasets.

(JMIR Form Res 2025;9:e67311) doi:10.2196/67311

KEYWORDS

trauma; abbreviated injury scale; deep learning; diagnostic information; transformer model; validation study

XSL•FO
RenderX

Introduction

Background

With the frequent occurrence of traffic crashes and the intensification of natural disasters, injuries have become the main cause of morbidity and mortality worldwide. According to the World Health Organization's (WHO) 2022 report [1], approximately 4.4 million people die, and tens of millions endure from nonfatal injuries every year due to such incidents.

The Abbreviated Injury Scale (AIS) [2] is the most widely used injury severity coding system, developed and periodically refined by the AIS Committee under the Association for the Advancement of Automotive Medicine (AAAM). AIS serves as the foundation for several severity scoring systems, such as the Injury Severity Score (ISS) [3], the Maximum Abbreviated Injury Score [4], and the New Injury Severity Score [5]. Since 2008, the AIS score or ISS score has been used as a criterion for evaluating trauma centers in various countries and has now developed into a globally recognized trauma scoring system.

However, the AIS coding system is a highly refined and complex scoring system that covers injuries to various parts of the human body and scores them based on the severity of the injury. In practical applications, AIS codes often rely on the subjective judgment and rich clinical experience of medical professionals, which may lead to certain coding differences between different medical institutions or personnel. While advances in trauma care have improved overall outcomes, significant disparities persist across sociodemographic groups. Low-income populations experience 38% longer prehospital delays for penetrating injuries compared with high-income counterparts [6], potentially biasing AIS severity assessments due to delayed clinical documentation. This dual challenge of subjective variability and systemic bias further increases the difficulty of accurate AIS code prediction.

The application of artificial intelligence (AI) models in medicine is increasing and many are based on AIS codes to predict mortality and prognosis outcomes [7,8]. While few studies have used diagnostic-related information to predict AIS codes. Although the neural machine translation (NMT) [9] model uses International Classification of Diseases (ICD) codes and other relevant information to predict AIS codes, the accurate acquisition of ICD codes itself necessitates substantial coding effort and also depends on detailed medical records and other clinical information during the patient's diagnostic process. Therefore, this also puts higher demands on the workload of information collection and processing. To overcome these shortcomings, we hope to use advanced deep learning (DL) techniques to directly predict AIS codes based on easily accessible diagnostic information, thereby improving the accuracy of trauma assessment for patients.

Therefore, we aim to use patient with trauma data from Chongqing Daping Hospital from October 2013 to June 2024 to construct a Bidirectional Encoder Representations from Transformers (BERT) [10] model based on DL for predicting the AIS codes corresponding to specific trauma. In this model, we use the patient's diagnostic information as the main input feature and compare it with the NMT model from previous research.

Related Work

In recent years, AI technology has been frequently used to discover complex correlations between various features in medical applications [11-13], such as individual injuries and mortality [14]. Lee et al [15] developed an ensemble model based on deep neural networks, incorporating the ICD, triage scale, procedure codes, and other clinical features as inputs to predict in-hospital mortality among patients with physical trauma. This model achieved an area under the curve (AUC) of up to 0.9507, outperforming advanced predictive models such as AdaBoost and XGBoost. Kang et al [8] created an AI algorithm grounded in DL models, leveraging the AIS codes to predict in-hospital mortality. By comparing their model with conventional ISS and New Injury Severity Score systems, they demonstrated the superior accuracy and AUC value of their proposed model. Tran et al [16] used ICD-10 codes and machine learning (ML) algorithms to develop a mortality prediction model via the National Trauma Data Bank. A comparison of its performance with that of logistic regression, ISS, and Trauma Mortality Prediction Model (TMPM-ICD10) validated that their XGBoost–based ML model exhibited superior performance. In terms of AIS code prediction, Hartka et al [9] proposed the use of an NMT model to convert ICD codes into AIS codes and compared its accuracy in assessing injury severity with that of two established conversion methods: the ICD-AIS map [17] and the ICD Programs for Injury Classification in R (ICDPIC-R) package [18]. Their results demonstrated that the NMT model achieved the highest accuracy across all injury severity classifications.

In the past few years, advanced pretrained language representation models such as BERT, Robustly Optimized BERT Pretraining Approach (RoBERTa), and HFL (a Chinese BERT pretraining model) have made remarkable breakthroughs in the field of natural language processing, demonstrating significant performance gains in various tasks such as text classification, sentiment analysis, and question answering [19]. However, to our current knowledge, despite the increasingly widespread application of DL techniques in the field of medical information processing, DL methods with pretrained language representation models have not yet been widely used for predicting AIS codes. Although the NMT model has shown some accuracy in predicting AIS codes, the AIS coding system, as a standardized tool for assessing the severity of injuries, occupies an important position in trauma medicine and emergency medicine. It can provide objective and comparable injury assessment for clinical doctors based on the specific injury situation of patients, which is of great significance for guiding treatment decisions and evaluating prognosis.

Methods

Patients and Dataset

The Chongqing Daping Hospital Trauma Database contains data about patients' diagnostic information, injury description, age, sex, place of injury, cause of injury, external cause code 1 (ECode1), external cause code 2 (ECode2), injury region, injury

XSL•FO
RenderX

types, present illness history, and AIS codes, where the AIS codes are based on the AIS2015 version [20], provided by professionally trained doctors according to the specific injury situation of the patient. To ensure the accuracy of the coding, the hospital has adopted a dual coding system: one doctor is responsible for preliminary coding, while the other doctor conducts follow-up checks.

The Daping Hospital Trauma Database contains data from 26,810 patients registered between October 1, 2013 and June 30, 2024 with the exclusion criteria of (1) patients transferred to another hospital, (2) patients who died in the emergency department before admission to the ICU or general ward, (3) data in the Daping Hospital Trauma Database with a feature loss rate ≥30%, (4) samples with any missing AIS codes or diagnostic information, and (5) data with less than 30 AIS code categories. According to the exclusion criteria, 13,216 pieces of data met the requirements. In addition, we also used an external dataset of 244 Chongqing Emergency Centers that met the inclusion and exclusion criteria.

We divided the Daping Hospital Trauma Database dataset into training data and testing data. The training dataset includes data from October 1, 2013 to December 31, 2022, which is used to train our model to learn the mapping relationship from input features to target variables. The test dataset includes data from January 1, 2023 to June 30, 2024. This partitioning dataset method has multiple benefits: the test dataset uses the most recent data, which can better evaluate the model's adaptability. By having the model learn historical data during the training phase and then face different but relevant data during the testing phase, it can encourage the model to learn more generalized features, which helps improve the accuracy of predicting future unknown data.

The number of training, testing, and external datasets is 10,827, 2389, and 244, respectively, with 337 types of AIS codes included in the training dataset, 332 types in the testing dataset, and 83 in the external dataset. In the training dataset, the number of AIS code 853161.3 is the highest, reaching 475 (accounting for 4.38%), while the number of AIS code 854221.2 is the lowest, at 25 (accounting for 0.23%). In the test dataset, the number of AIS code 853161.3 is also the highest, at 119 (accounting for 4.98%), while the number of AIS code 910200.1 is the lowest, at 6 (accounting for 0.25%). In the external dataset, the number of AIS code 853161.3 is also the highest, reaching 10 (accounting for 4.10%), while the number of AIS code 856151.2 is the lowest, at 1 (accounting for 0.41%). The injuries covered by our dataset are mainly concentrated in areas such as the skin, limbs, and head and neck, especially those types of injuries that are most common in practical work and have a direct impact on clinical decision-making and treatment plans.

The testing dataset (n=2389) was used only to independently test our developed model and not for training or internal validation. We first performed a 5-fold cross-validation using the training data to prevent overfitting. The training dataset (n=10,827) was randomly shuffled and stratified into 5 equal groups with 4 groups used for training and 1 group used for validation. This process was repeated 5 times by shifting the internal validation group. Then, the overall performance of the model was evaluated through independent testing data. Finally, the generalization ability of the model was validated through multicenter validation using external data.

BERT Prediction Model Development

BERT Model Architecture

As shown in Figure 1, we developed a DL–based BERT model for predicting AIS codes. The model uses the masked language modeling technique of the pretrained model RoBERTa [10] to learn contextual information about data of patient with trauma. This approach can capture both forward and backward contextual information of the input sequences to achieve a deeper understanding of input textual data.

To provide further clarity on the data used in our model, we have included a Multimedia Appendix 1 with examples of model inputs and outputs. These examples illustrate the structure of the data and how our model processes it to generate AIS scores.

Figure 1. The Bidirectional Encoder Representations from Transformers model architecture. AIS: Abbreviated Injury Scale.

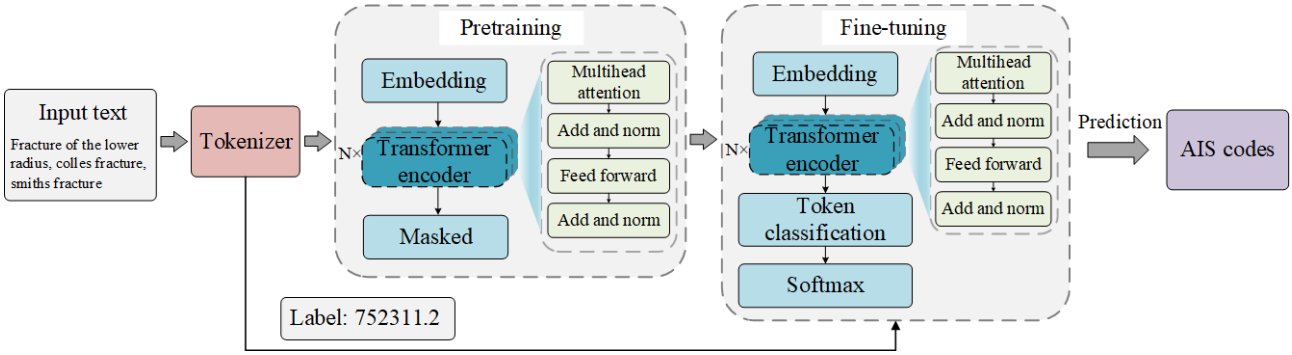

BERT Pretraining

Our BERT model is a pretraining model based on a modified version of the RoBERTa model, and we similarly carry out in-depth optimizations of the BERT model, including the implementation of dynamically tuned masking strategies and enhanced text encoding processing. The significant advantage of these optimizations is that by randomly performing masking operations on the input data, different masking patterns are used for the same training data in different training epochs, thus effectively increasing data diversity during model training without the need to expand the training dataset. In addition, by adopting a larger batch size, more training data, and longer training time, the BERT model can learn richer linguistic

XSL•FO
RenderX

features and potential patterns in the data, which significantly improves the prediction performance of the model.

Specifically, the vocabulary list used for training includes the patient's diagnostic information, injury description, age, sex, place of injury, cause of injury, procedure codes (ECode1 and ECode2), injury region, injury types, and present illness history, as well as the necessary special tokens (eg,<CLS_TOKEN>,<SEP_TOKEN>,<PAD_TOKEN>, and <MASK_TOKEN>). The entire text sequence is treated as a sentence with the sequence being identified by the start token <CLS_TOKEN> and the end token <SEP_TOKEN>;<PAD_TOKEN> is used to pad the sequence to a uniform length, and <MASK_TOKEN> is used for masking. We chose the masked language modeling method for training to accommodate the need for contextual understanding in AIS code prediction. During training, we randomly perturbed 15% of the trauma diagnostic information elements in the input sequences, similar to the RoBERTa setup, where 80% of the tokens of these 15% selected elements were replaced with <MASK_TOKEN>, requiring the model to predict the correct AIS codes at these <MASK_TOKEN> locations during training, 10% of the tokens were replaced with other trauma diagnostic information randomly selected to increase the difficulty of training and generalization, and the remaining 10% remained unchanged and served as positive samples for model learning. During pretraining, the model predicts what kind of residue it is in the masked position. For each batch, the loss is defined as:

$$\text{Loss}=-1|\text{batch}|\sum\text{seq}\in\text{batch}\sum\text{i}\in\text{masklogp}\hat{}(\text{si}|\text{S})$$

where p^(si|S\mask) represents the probability that the model predicts the element si at the ith masked position, given all the sequence information except for the masked positions.

BERT Fine-Tuning

We consider the task of AIS code prediction as a multivariate labeling classification task, where the BERT model is used to predict the AIS codes for each instance in the input sequence. To accomplish this, we map AIS code categories to unique integer labels, which are used as supervised learning objectives. We add a multivariate classification header on top of the pretrained model, whose output dimension matches the number of AIS codes categories. During training, the BERT model receives input sequences and extracts contextually relevant feature representations. These features are then passed to the multivariate classification header to generate predicted probabilities corresponding to each AIS code category. We use a cross-entropy loss function to compute the difference between the predicted probability distribution and the true AIS code label. The loss during fine-tuning is defined as:

$$L=-1N\sum i=1N\log(y\hat{}i,yi)$$

where yi is the true category index of the ith sample, and y^i,yi is the probability that the ith sample predicted by the model belongs to category yi.

BERT Model Configuration

After a careful hyperparameter search, we determined the optimal model configuration: an 8-layer BERT architecture including an input layer, six 384-unit hidden layers, and an output layer, which together form the encoder-decoder transformer components with 5 transformer blocks. In the process of determining the 8-layer architecture, we conducted subsequent experiments and tried different configurations of BERT models with 4, 8, and 12 layers and comprehensively evaluated their performance. The results showed that the 8-layer architecture demonstrated excellent performance on multiple evaluation metrics, so we ultimately chose this architecture for our research. The model is also configured with 12 attention heads and an embedding size of 128. During optimization, we used the Adam optimizer to adjust the model weights with a learning rate of 0.0001, a batch size of 64, and the Gaussian Error Linear Unit as the activation function. To further reduce the number of parameters and computational costs, we implemented low-rank factorization techniques in the embedding layer.

The BERT model was trained via early stopping with training concluding after the validation loss did not improve after 10 epochs. The model was trained on a computing cluster with 48 GB of memory and one Graphics Processing Unit (NVIDIA RTX A6000). We implemented the model through the Pytorch framework.

Evaluation Methods

We use accuracy, AUC, and F_1-scores as evaluation metrics. Because our prediction task is a multiclass problem and all categories of AIS codes are considered equally important, we chose the macro averaging method when calculating AUC and F_1-scores to ensure that the model's performance across all AIS code categories is comprehensively reflected.

Comparison Method

To validate the performance of our BERT model, we selected a series of representative methods for comparison. Specifically, we used previous research (NMT) and current mainstream ML methods, including K-nearest neighbors (KNN), multilayer perceptron (MLP), XGBoost, AdaBoost, and decision tree (DT). The following is a detailed introduction to these methods:

Neural Machine Translation

The NMT model proposed by Hartka et al [9] is a DL technique commonly used for human language translation. The model is implemented using OpenNMT, an open-source toolkit developed by the Harvard NLP team and SYSTRAN for NMT, to convert ICD codes into exact AIS codes. This paper shares the same goals and tasks as their work. However, accurately obtaining ICD codes not only requires a lot of coding work but also relies on detailed medical records and other clinical information during the patient's diagnostic process. In contrast, the BERT model mainly relies on easily accessible diagnostic information to predict specific AIS codes. The NMT model and the BERT model both adopt a similar Transformer architecture. In their experimental configuration, the NMT model includes 6 hidden layers with 512 units, 8 attention heads, a loss rate of 0.1, weights adjusted by the Adam optimizer, learning rate decay determined by Noam decay, and classification cross entropy used as the training loss function.

XSL•FO
RenderX

Machine Learning

For ML methods, we use Word2Vec word embedding technology to convert text data into a format suitable for ML algorithm processing.

K-Nearest Neighbor

KNN is a simple but effective classification algorithm. It is based on distance metrics such as Euclidean distance to find the k samples in the training set that are most similar to the test samples and predicts the category of the test samples based on the categories of these neighbors. In our experiment, the value of k was set to 3.

Multilayer Perceptron

MLP is a feedforward neural network consisting of an input layer, a hidden layer, and an output layer. It approximates complex functional relationships through multilayer nonlinear transformations. In our experiment, MLP used 2 hidden layers with the first layer having 20 neurons and the second layer having 50 neurons.

XGBoost

XGBoost is an ensemble learning method based on gradient boosting, which constructs strong classifiers by combining multiple weak classifiers. XGBoost has achieved significant performance improvements in multiple fields, especially when dealing with large-scale datasets and high-dimensional features. In our experiment, we used multiclass log loss as an evaluation metric, which is a commonly used choice in multiclass classification problems.

AdaBoost

AdaBoost is an adaptive boosting algorithm that constructs strong classifiers by adjusting the weights of each weak classifier. AdaBoost has demonstrated strong performance in handling classification tasks, especially when dealing with imbalanced datasets. In our experiment, a DT stump (with a depth of 1) was used as the weak classifier.

Decision Tree

DT is an intuitive classification and regression method. It generates decision paths through a series of conditional judgments, thereby achieving classification or regression prediction of samples. In our experiment, the default Gini impurity was used as a measure of splitting quality, and the model was constructed by recursively segmenting the feature space.

Ethical Considerations

This study is an observational study; the data used were reviewed and approved by the Internal Review Board of Chongqing Daping Hospital (approval number: 2024_219), and informed consent from patients was exempted.

Results

BERT Model Training

The Daping Hospital dataset used in this study contains a total of 13,216 records, including 10,827 in the training dataset and 2389 in the independent testing dataset. In addition, we obtained 244 external data from the Chongqing Emergency Center. Table 1 provides a detailed list of the demographic characteristics and injury status of these datasets. In both the training and testing datasets, the IQR of age is 33 years. In terms of gender distribution, males accounted for 62.5%, 59.3%, and 63.9% of the training, testing, and external datasets, respectively. In terms of injury causes, the most common were falls (accounting for 57.2%, 57.9%, and 41.0% in the 3 datasets) and traffic crashes (accounting for 14.4%, 16.6%, and 43.9%, respectively). To reflect the distribution of severity data for single injuries, we presented the severity distribution of individual injuries based on the data after the decimal point of the AIS code. Our dataset mainly includes data for mild, moderate, and severe injuries. It is worth noting that in all 3 datasets, moderate injuries account for the highest proportion, at 45%, 51.5%, and 49.6%, respectively.

Table . Demographic and injury characteristics of patients with trauma in the datasets (N=13460).

Variables	Training dataset (2013/10 - 2022/12)	Testing dataset (2023/01 - 2024/06)	External dataset (2023/09 - 2023/10)
Total number of patients	10,827	2389	244
Age range (years), IQR	1 - 102 (33)	3 - 98 (33)	14 - 97 (31)
Males (%)	6771 (62.5)	1419 (59.3)	156 (63.9)
Mechanism of injury, n (%)			
Traffic crash	1557 (14.4)	397 (16.6)	107 (43.9)
Falls	6189 (57.2)	1382 (57.9)	100 (41.0)
Blunt	123 (1.1)	80 (3.3)	22 (9.0)
Sports injury	329 (3.0)	140 (5.9)	10 (4.1)
Other	2629 (24.3)	390 (16.3)	5 (2.0)
Severity of injury, n (%)			
Mild injury	2468 (22.8)	392 (16.4)	40 (16.4)
Moderate injury	4870 (45.0)	1231 (51.5)	121 (49.6)
Severe injury	3489 (32.2)	766 (32.1)	83 (34.0)

XSL•FO
RenderX

Overall Predictive Performance of the BERT Model

We first compared the performance of the BERT model with the NMT model and several advanced ML models, including KNN, MLP, XGBoost, AdaBoost, and DT. Comparison results on our independent test dataset are shown in Table 2. We used accuracy, AUC, and F_1-scores as evaluation metrics. It is worth noting that since the NMT model only provides prediction results for AIS code accuracy, we only present the accuracy of NMT in Table 2. For other metrics not provided by the NMT model, we uniformly use the symbol "—" for annotation.

The performance of our proposed BERT model is significantly better than all comparison models across all indicators. Specifically, the accuracy of the BERT model is as high as 0.8971, while the accuracy of the NMT model is only 0.7380 with a difference of over 10 % points between the two. In addition, the AUC value of the BERT model is 0.9970, and the F_1-score is 0.8434. Among all the ML methods compared, the DT method achieved excellent performance in accuracy and AUC of 0.8506 and 0.9945, respectively, and the XGBoost method achieved the best results in the F_1 index at 0.7586, but they still failed to surpass the performance of our BERT model. These comparative experiments fully demonstrate that our BERT model has high prediction accuracy.

Figure 2 shows the training curve of the BERT model with the (**A**) and (**B**) graphs displaying their results on the training and testing datasets, respectively. The x-axis represents epochs and is set to 50, whereas the y-axis represents the values of accuracy, AUC, and F_1-score. The figure shows that the BERT model tends to be stable in all the metrics when the training or test dataset reaches epoch 26. The test dataset shows that the model performs well in terms of overall prediction accuracy and accurately classifies samples into the correct categories. For both the training and testing datasets, the AUC is close to 1, and high AUC values further demonstrate the model's strong ability to distinguish between positive and negative samples, maintaining excellent performance at almost all possible classification thresholds.

To demonstrate the performance advantage of our fine-tuned BERT model, we compared it with the base version of BERT (BERT-base) and the pretrained model HFL/chinese-roberts-wwm-ext [21]. The comparison results are shown in Table 3. Compared to the BERT-base and HFL models, the BERT model exhibits higher performance in accuracy, reaching 0.8971, while also significantly leading in the F_1-score, at 0.8434. The results of these two indicators are 2.99% and 7.54% higher than the HFL model, respectively, indicating that our BERT model achieves a better balance between accuracy and recall in classification tasks. Although the BERT is not significantly different from HFL and BERT-base on AUC, its advantages in key evaluation metrics highlight BERT's superior overall performance in handling specific tasks in this paper.

Table . Prediction results of the Bidirectional Encoder Representations from Transformers model and comparative model in the test dataset.

Model	Accuracy	AUC[a]	F^1-scores
NMT[b]	0.7380	NA[c]	NA
KNN[d]	0.7935	0.9414	0.6879
MLP[e]	0.8064	0.9886	0.6194
XGBoost[f]	0.8374	0.9937	0.7586
AdaBoost[g]	0.8506	0.9860	0.7050
DT[h]	0.8506	0.9945	0.7049
BERT[i]	0.8971	0.9970	0.8434

[a]AUC: area under the curve.
[b]NMT: neural machine translation.
[c]NA: not available.
[d]KNN: K-Nearest Neighbor.
[e]MLP: multilayer perceptron.
[f]XGBoost: Extreme Gradient Boosting.
[g]AdaBoost: adaptive boosting
[h]DT: decision tree.
[i]BERT: Bidirectional Encoder Representations from Transformers

XSL•FO
RenderX

Figure 2. The visualization curve of Bidirectional Encoder Representations from Transformers model's predictive performance on training and testing datasets. AUC: area under the receiver operating characteristic curve.

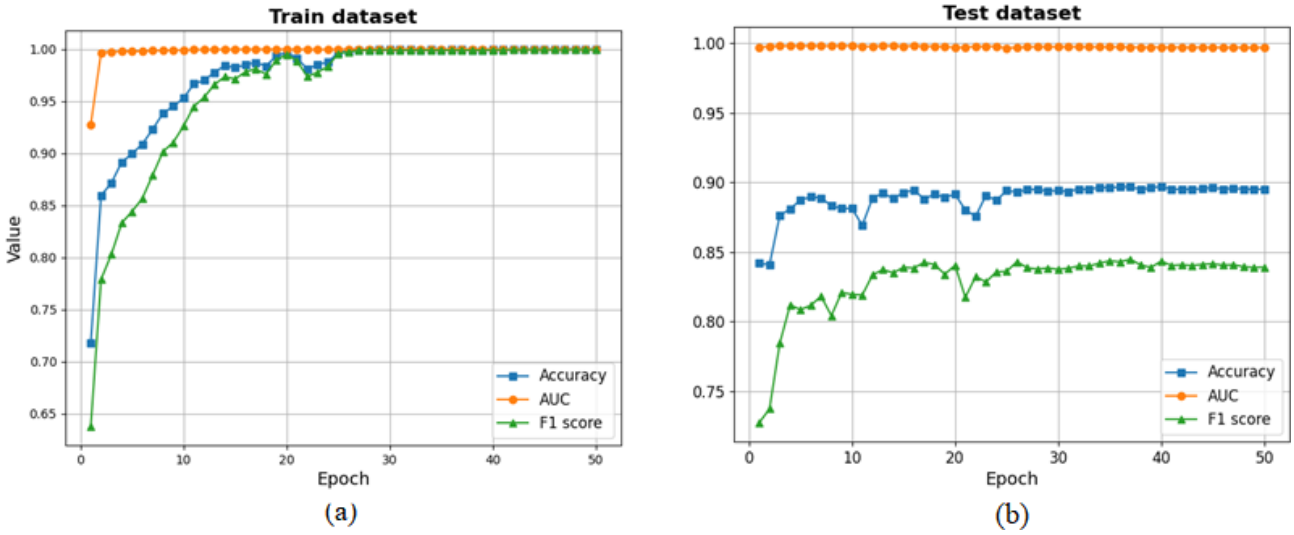

Table . Comparison between our Bidirectional Encoder Representations from Transformers model and other pretrained models in the test dataset.

Model	Accuracy	AUC[a]	F_1-scores
BERT-base[b]	0.8559	0.9971	0.7284
[c]HFL	0.8672	0.9973	0.7680
BERT	0.8971	0.9970	0.8434

[a]AUC: area under the curve.

[b]BERT: Bidirectional Encoder Representations from Transformers.

[c]HFL: a Chinese BERT pretraining model.

Ablation Study

To verify the maximum contribution of specific input data feature combinations to the model, we designed a series of ablation studies. The input features of the original experiment include patients' diagnostic information, age, sex, injury description, place of injury, cause of injury, procedure codes (ECode1 and ECode2), injury region, injury types, and present illness history.

First, we conducted a single-factor ablation study to observe the impact of removing each feature item one by one on the performance of the model. The experimental results show that diagnostic information and injury description are the most significant data types that affect the performance of the model. Based on the findings of the single-factor ablation study, we further designed the following multifactor ablation study to explore the importance of diagnostic information and injury description in feature combination. The specific results are shown in Table 4.

As shown in Table 4, after removing diagnostic information, the accuracy and F_1-score of the model significantly decreased. Although the AUC remained high, it also decreased, indicating that diagnostic information features have a significant impact on the performance of the model. After removing the injury description, the performance of the model also decreased, but the decrease was smaller compared to removing diagnostic information. The ablation study demonstrated that diagnostic information contributed the most to the model, followed by the injury description, and the feature combination we used achieved the best results.

XSL•FO
RenderX

Table . Results of ablation study in the test dataset.

Model	Accuracy	AUC[a]	F_1-scores
Diagnostic information removed[b]	0.6033	0.9523	0.4668
Injury description removed[c]	0.8888	0.9875	0.8047
Diagnostic information and injury description removed[d]	0.6014	0.9438	0.4699
BERT[e]	0.8971	0.9970	0.8434

[a]AUC: area under the curve.

[b]Diagnostic information removed: Remove diagnostic information based on all basic features.

[c]Injury description removed: Remove the injury description based on all basic features.

[d]Diagnostic information and injury description removed (simultaneously remove diagnostic information and injury description): Based on all basic features, simultaneously remove diagnostic information and injury description.

[e]BERT: Bidirectional Encoder Representations from Transformers.

External Validation

External validation of the constructed BERT model was conducted using data from 244 patients with trauma at Chongqing Emergency Center. This external dataset has a similar data structure to the training dataset, containing a total of 83 AIS code categories. The experimental results of the external dataset are shown in Table 5 with accuracy, AUC, and F_1 of 0.7131, 0.8586, and 0.6801, respectively. Compared with the test dataset, the performance of the BERT model slightly decreases on external datasets, which may be due to differences in data distribution between different medical institutions. But the overall performance is still satisfactory, indicating that the BERT model has strong generalization ability.

Table . Validation results of Bidirectional Encoder Representations from Transformers model on external datasets.

Dataset	Accuracy	AUC[a]	F_1-scores
Test dataset	0.8971	0.9970	0.8434
External dataset	0.7131	0.8586	0.6801

[a]AUC: area under the curve.

Discussion

Principal Findings

In this study, we successfully constructed a BERT–based DL model using data of patient with trauma from Chongqing Daping Hospital to predict AIS codes, achieving an accuracy of 89.71%. This model, leveraging patient diagnostic information as primary input features, demonstrated superior performance compared to existing advanced AIS prediction models, including the previously studied NMT framework [9]. Additionally, we validated the model's high generalization ability using data from an external center, thereby fulfilling our objective of enhancing trauma assessment through DL.

Key innovations such as dynamic masking strategies, low-rank embedding decomposition, and bidirectional contextual modeling enabled the model to capture nuanced clinical semantics while maintaining computational efficiency. Notably, the model directly outputs complete AIS codes—a critical advancement over rule-based methods like the ICD-AIS map [17] and the ICDPIC-R package [18] tools, which lack granular code prediction capabilities. These results underscore the potential of transformer-based architectures to enhance trauma assessment workflows, particularly in scenarios requiring rapid, large-scale injury coding.

Our findings align with emerging evidence supporting transformer models in clinical text processing [10], yet extend prior work by addressing the unique challenges of AIS coding. Unlike NMT models that process sequential tokens independently via recurrent mechanisms [9], the BERT model's bidirectional attention dynamically links contextual elements of injury descriptions. In addition, the pretrained biomedical embeddings provided higher precision in rare injury terminology recognition compared to NMT's task-specific training.

The model's external validation performance further reinforces its clinical utility. While annual variations in trauma patterns typically degrade conventional models, our temporal split testing revealed stable predictive accuracy. This robustness suggests the framework could adapt to shifting trauma trends without frequent retraining. Moreover, direct AIS code generation eliminates the multistep mapping required by ICD-based tools [9,17,18], reducing error propagation risks in mass casualty scenarios where rapid triage coding is critical.

These findings suggest that BERT may become a powerful tool for injury research. Although independent coding of AIS injuries by trained medical professionals and comprehensive medical data remains the gold standard in this field, in some cases, such as when the number of patients is large or detailed medical records are difficult to obtain, independent coding becomes impractical. At this point, given input features that satisfy the model, our BERT model can automatically provide prediction

XSL•FO
RenderX

results for AIS codes, providing highly accurate AIS code predictions for individual patients.

Limitations

Our research has several limitations. First, during the data collection and processing phase, a large amount of data was excluded due to the lack of information on injury description and present illness history, which had a significant impact on the integrity and representativeness of the final dataset. Second, while low-rank decomposition improved efficiency, BERT's inherent sequence length restrictions (≤512 tokens) may truncate complex trauma descriptions. Third, the Chinese-language training data raises questions about cross-lingual applicability, given known variations in medical terminologies across languages [22]. Finally, as with most DL systems, the model's black-box nature limits clinical interpretability.

Future studies should address these gaps by (1) integrating multimodal data (eg, imaging reports) to compensate for text incompleteness, (2) benchmarking against large language models with superior few-shot learning capacities, and (3) developing hybrid systems that combine BERT's predictive power with large language model–driven explainability features.

Conclusions

The BERT model we propose is mainly based on diagnostic information to predict AIS codes, and its prediction accuracy is superior to existing methods. These findings highlight the potential of advanced AI techniques to enhance clinical decision-making processes and improve the efficiency and accuracy of AIS code prediction.

By automating a task that traditionally requires hours of expert review per case, our framework could democratize high-quality trauma registries in resource-limited settings. Crucially, the model does not seek to replace human coders but provides a scalable adjunct for high-volume scenarios—a balance increasingly advocated in AI-augmented health care [23]. As trauma systems worldwide adopt electronic health records, such tools may transform retrospective coding into a prospective clinical decision aid, ultimately bridging the gap between injury documentation and precision trauma care. Future iterations incorporating multi-institutional data and explainability interfaces could further establish BERT-derived models as indispensable tools in computational trauma epidemiology.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (72374206), the Artificial Intelligence Medical Research Project of Daping Hospital in Chongqing, China (ZXAIYB015), the Basic Research Project of the Army Characteristic Medical Center (ZXJCYJ202414), and the Chongqing Science and Health Joint Medical Research Project (2024MSXM084).

Data Availability

The dataset used in this study is not publicly available but is available from the corresponding author on reasonable request.

The algorithm code is currently not publicly available. Researchers can apply for academic permission to use the model code by contacting the corresponding author via email.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Model input and output examples.
[DOCX File, 16 KB - formative_v9i1e67311_app1.docx]

References

1. World Health Organization. Global Status Report on Violence Prevention 2022. World Health Organization 2022 [FREE Full text]
2. Rating the severity of tissue damage: I. The abbreviated scale. JAMA 1971 Jan 11;215(2):277-280. [doi: 10.1001/jama.1971.03180150059012] [Medline: 5107365]
3. Baker SP, O'Neill B, Haddon W Jr, Long WB. The injury severity score: a method for describing patients with multiple injuries and evaluating emergency care. J Trauma 1974 Mar;14(3):187-196. [Medline: 4814394]
4. Baker SP, O'Neill B. The injury severity score: an update. J Trauma 1976 Nov;16(11):882-885. [doi: 10.1097/00005373-197611000-00006] [Medline: 994270]
5. Kuo SCH, Kuo PJ, Chen YC, et al. Comparison of the new exponential injury severity score with the Injury Severity Score and the New Injury Severity Score in trauma patients: a cross-sectional study. PLoS ONE 2017;12(11):e0187871. [doi: 10.1371/journal.pone.0187871] [Medline: 29121653]
6. Haider AH, Weygandt PL, Bentley JM, et al. Disparities in trauma care and outcomes in the United States: a systematic review and meta-analysis. J Trauma Acute Care Surg 2013 May;74(5):1195-1205. [doi: 10.1097/TA.0b013e31828c331d] [Medline: 23609267]

XSL•FO
RenderX

7. Peng HT, Siddiqui MM, Rhind SG, et al. Artificial intelligence and machine learning for hemorrhagic trauma care. Mil Med Res 2023 Feb 16;10(1):6. [doi: 10.1186/s40779-023-00444-0] [Medline: 36793066]
8. Kang WS, Chung H, Ko H, et al. Artificial intelligence to predict in-hospital mortality using novel anatomical injury score. Sci Rep 2021 Dec 7;11(1):23534. [doi: 10.1038/s41598-021-03024-1] [Medline: 34876644]
9. Hartka T, Chernyavskiy P, Glass G, et al. Evaluation of neural machine translation for conversion of international classification of disease codes to the Abbreviated Injury Scale. Accid Anal Prev 2023 Oct;191:107183. [doi: 10.1016/j.aap.2023.107183] [Medline: 37418869]
10. Liu Y, Ott M, Goyal N, et al. RoBERTa: a robustly optimized BERT pretraining approach. Arxiv. Preprint posted online on Jul 26, 2019. [doi: 10.48550/arXiv.1907.11692]
11. Korndorffer JR Jr, Hawn MT, Spain DA, et al. Situating artificial intelligence in surgery: a focus on disease severity. Ann Surg 2020 Sep 1;272(3):523-528. [doi: 10.1097/SLA.0000000000004207] [Medline: 33759839]
12. Swinckels L, Bennis FC, Ziesemer KA, et al. The use of deep learning and machine learning on longitudinal electronic health records for the early detection and prevention of diseases: scoping review. J Med Internet Res 2024 Aug 20;26:e48320. [doi: 10.2196/48320] [Medline: 39163096]
13. Lammers D, Marenco C, Morte K, et al. Machine learning for military trauma: novel massive transfusion predictive models in combat zones. J Surg Res 2022 Feb;270(369-375):369-375. [doi: 10.1016/j.jss.2021.09.017] [Medline: 34736129]
14. Xie P, Wang H, Xiao J, et al. Development and validation of an explainable deep learning model to predict in-hospital mortality for patients with acute myocardial infarction: algorithm development and validation study. J Med Internet Res 2024 May 10;26:e49848. [doi: 10.2196/49848] [Medline: 38728685]
15. Lee S, Kang WS, Seo S, et al. Model for predicting in-hospital mortality of physical trauma patients using artificial intelligence techniques: nationwide population-based study in Korea. J Med Internet Res 2022 Dec 13;24(12):e43757. [doi: 10.2196/43757] [Medline: 36512392]
16. Tran Z, Zhang W, Verma A, et al. The derivation of an International classification of diseases, tenth revision–based trauma-related mortality model using machine learning. J Trauma Acute Care Surg 2022 Mar 1;92(3):561-566. [doi: 10.1097/TA.0000000000003416] [Medline: 34554135]
17. Loftis KL, Price JP, Gillich PJ, et al. Development of an expert based ICD-9-CM and ICD-10-CM map to AIS 2005 update 2008. Traffic Inj Prev 2016 Sep;17 Suppl 1(1-5):1-5. [doi: 10.1080/15389588.2016.1191069] [Medline: 27586094]
18. Clark DE, Black AW, Skavdahl DH, Hallagan LD. Open-access programs for injury categorization using ICD-9 or ICD-10. Inj Epidemiol 2018 Apr 9;5(1):11. [doi: 10.1186/s40621-018-0149-8] [Medline: 29629480]
19. Radford A, Narasimhan K, Salimans T, et al. Improving language understanding with unsupervised learning. OpenAI 2018 Jun 11;URL [FREE Full text]
20. Jihong Z, Ao Y, Xing L, et al. Analysis of the main revisions of the 2015 version of the simplified injury score. Chin J Traumatol 2022;38(8):747-749. [doi: 10.3760/cma.j.cn501098-20220729-00263]
21. Cui Y, Che W, Liu T, et al. Pre-training with whole word masking for Chinese BERT. IEEE/ACM Trans Audio Speech Lang Process 2021;29:3504-3514. [doi: 10.1109/TASLP.2021.3124365]
22. Névéol A, Dalianis H, Velupillai S, et al. Clinical natural language processing in languages other than English: opportunities and challenges. J Biomed Semantics 2018 Mar 30;9(1):12. [doi: 10.1186/s13326-018-0179-8] [Medline: 29602312]
23. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]

Abbreviations

AAAM: Association for the Advancement of Automotive Medicine
AI: artificial intelligence
AIS: Abbreviated Injury Scale
AUC: area under the curve
BERT: Bidirectional Encoder Representations from Transformers
DL: deep learning
DT: decision tree
ECode: external cause code
ICD: International Classification of Diseases
ISS: Injury Severity Score
KNN: K-nearest neighbor
ML: machine learning
MLP: multilayer perceptron
NMT: neural machine translation
RoBERTa: Robustly Optimized BERT Pretraining Approach
WHO: World Health Organization

XSL•FO
RenderX

Edited by A Mavragani; submitted 08.10.24; peer-reviewed by F Williamson, M Wang, T Hartka, W Feng; revised version received 23.03.25; accepted 07.04.25; published 29.05.25.

Please cite as:
Tang J, Li Y, Luo K, Lai J, Yin X, Wu D
Application of the Bidirectional Encoder Representations from Transformers Model for Predicting the Abbreviated Injury Scale in Patients with Trauma: Algorithm Development and Validation Study
JMIR Form Res 2025;9:e67311
URL: https://formative.jmir.org/2025/1/e67311
doi:10.2196/67311

© Jun Tang, Yang Li, Keyu Luo, Jiangyuan Lai, Xiang Yin, Dongdong Wu. Originally published in JMIR Formative Research (https://formative.jmir.org), 29.5.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Comparative Efficacy of MultiModal AI Methods in Screening for Major Depressive Disorder: Machine Learning Model Development Predictive Pilot Study

Donghao Chen[1], PhD; Pengfei Wang[2,3], PhD; Xiaolong Zhang[2,3], PhD; Runqi Qiao[1], PhD; Nanxi Li[2,3], PhD; Xiaodong Zhang[2,3], BD; Honggang Zhang[1], PhD; Gang Wang[2,3], PhD, MD

[1]School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China

[2]Beijing Key Laboratory of Mental Disorders, National Clinical Research Center for Mental Disorders & National Center for Mental Disorders, Beijing Anding Hospital, Capital Medical University, Beijing, China

[3]Advanced Innovation Center for Human Brain Protection, Capital Medical University, Beijing, China

Corresponding Author:
Gang Wang, PhD, MD
Beijing Key Laboratory of Mental Disorders, National Clinical Research Center for Mental Disorders & National Center for Mental Disorders, Beijing Anding Hospital, Capital Medical University, Beijing, China

Abstract

Background: Conventional approaches for major depressive disorder (MDD) screening rely on two effective but subjective paradigms: self-rated scales and clinical interviews. Artificial intelligence (AI) can potentially contribute to psychiatry, especially through the use of objective data such as objective audiovisual signals.

Objective: This study aimed to evaluate the efficacy of different paradigms using AI analysis on audiovisual signals.

Methods: We recruited 89 participants (mean age, 37.1 years; male: 30/89, 33.7%; female: 59/89, 66.3%), including 41 patients with MDD and 48 asymptomatic participants. We developed AI models using facial movement, acoustic, and text features extracted from videos obtained via a tool, incorporating four paradigms: conventional scale (CS), question and answering (Q&A), mental imagery description (MID), and video watching (VW). Ablation experiments and 5-fold cross-validation were performed using two AI methods to ascertain the efficacy of paradigm combinations. Attention scores from the deep learning model were calculated and compared with correlation results to assess comprehensibility.

Results: In video clip-based analyses, Q&A outperformed MID with a mean binary sensitivity of 79.06% (95%CI 77.06% - 83.35%; *P*=.03) and an effect size of 1.0. Among individuals, the combination of Q&A and MID outperformed MID alone with a mean extent accuracy of 80.00% (95%CI 65.88% - 88.24%; *P*= .01), with an effect size 0.61. The mean binary accuracy exceeded 76.25% for video clip predictions and 74.12% for individual-level predictions across the two AI methods, with top individual binary accuracy of 94.12%. The features exhibiting high attention scores demonstrated a significant overlap with those that were statistically correlated, including 18 features (all *Ps*<.05), while also aligning with established nonverbal markers.

Conclusions: The Q&A paradigm demonstrated higher efficacy than MID, both individually and in combination. Using AI to analyze audiovisual signals across multiple paradigms has the potential to be an effective tool for MDD screening.

(JMIR Form Res 2025;9:e56057) doi:10.2196/56057

KEYWORDS

major depressive disorder; artificial intelligence; computational psychiatry; facial action unit; multimodal analysis; multiparadigm analysis; MDD

Introduction

Depressive disorder is a common mental disorder affecting approximately 322 million patients in the world, with major depressive disorder (MDD) as one of its two main subcategories, which can significantly affect all aspects of life, including performance at school, productivity at work, and relationships with family and friends [1]. The primary methods to assess depression encompass mental status examinations and assessment scales. However, mental status examinations such as the Hamilton Rating Scale for Depression (HAMD) necessitate direct, in-person interviews conducted by clinicians, which can result in processes that are both time-consuming and labor-intensive [2]. Self-Report Symptom Inventories (SRSI) such as the Beck Depression Inventory [3] and the Patient Health Questionnaire-9 (PHQ-9) [4] are time-efficient but can be influenced by subjective biases, which allows for individual

XSL•FO
RenderX

variability [5]. Therefore, the outcomes are susceptible to both intentional and unintentional subjective influences [6] and more approaches are needed to improve efficiency and accuracy.

In recent years, artificial intelligence (AI) has garnered attention for its application in signal analysis across various modalities. For instance, support vector machines have been used to analyze functional magnetic resonance imaging (fMRI) data [7] and a convolutional neural network (CNN) has been applied to an electroencephalogram (EEG) [8] to detect depression. While physiological signals such as fMRI and EEG are unaffected by subjective factors and directly reflect the participants' physical states, they involve complex procedures and high costs. In contrast, noncontact signals, including text, audio, visual content, and scale information are more accessible for analysis.

In the text modality, hidden Markov models and random forest models were developed to predict depression and posttraumatic stress disorder based on frequency of Twitter usage and content [9]. By aggregating weighted words using lexicons, the sentiment score derived from text messages demonstrated a positive association with the severity of depression as measured by the self-rated Patient Health Questionnaire-8 (PHQ-8) [10-12]. Among the audio modalities, speech patterns such as a narrowed pitch range and reduced phonemes within the vowel space have emerged as important objective indicators for assessing depressive states [13,14]. Along with prosodic features, mel frequency cepstral coefficients (MFCC) [15], detailed spectral features [16], and deep-learned acoustic characteristics [17] have also been used to identify the presence of depressive symptoms, achieving binary accuracy of up to 79% or F_1-score of 0.890.

For the visual or multimodal domain, several open datasets are available. One notable example is the Audio/Visual Emotion Challenge and Workshop [18], which focuses on the detection of depression and uses an audio-visual dataset that includes image features extracted from original images and audio recordings and transcribed text from Google Cloud, paired with the PHQ-8 scores. Facial action units (AU), as outlined in the Facial Action Coding System [19], serve as the foundation for facial expressions and constitute essential image features in the Audio/Visual Emotion Challenge and Workshop. Commonly observed AUs correspond to a range of expressions such as smiling and frowning (see Table S1 in Multimedia Appendix 1). A higher overall frowning (Action Unit 4, ie AU4) and head-down posture were identified in a study by Fiquer et al [20], while a lower overall AU12 and a markedly higher overall AU14 were identified in a study by Girard et al [21]. This indicated that the distribution of AUs differs significantly between depressed and nondepressed persons. Facial Action Coding System has also been employed in the analysis of stress [22], anxiety [23], and Parkinson's disease [24].

There are several other existing datasets, including Mundt-35 [25], BlackDog [26], and MODMA [27]. Most of these datasets contain a single paradigm, primarily relying on interviews such as HAMD, or targeting the scores of SRSIs such as the PHQ-8. Additionally, current multimodal AI methods mainly extract local features from utterances or sentences for video clip predictions [28,29]. At the same time, we believe the screening and diagnosis of MDD should include the entire process, similarly to the process of clinical practice.

For other paradigm options, mental imagery description (MID) [30] can manifest across different sensory modalities, encompassing visual [31,32], auditory [33], and textual information, and tends to evoke stronger emotional responses than verbal processing [30].

Thus, aiming to evaluate the efficacy of different paradigms, we aggregated them in a tool, namely the Electronic Tool for Depression (ETD), and used a state-of-the-art (SOTA) method using audiovisual signals to validate their efficacies. We propose the ETD to be a nonsubjective and easy-use MDD screening tool. The SOTA method generates predictions on video clips, and two of the four paradigms contain only visual signals; therefore, we implemented a voting mechanism for individual predictions and proposed a global feature method for the remaining vision-only paradigms. This pilot study underscores our primary contributions, which can be summarized as follows: (1) to validate the efficacy of the paradigms via AI on audiovisual signals and aggregate them within a tool for MDD screening and (2) to propose a global feature method and explore its efficacy and interpretability.

Methods

Design of the Task and Building the Tool

The ETD consists of four paradigms, aggregated into an application designed for an 11.5-inch tablet featuring an 8-MP front-facing camera and a 44.8 kHz sample rate microphone. Before using the ETD, clinicians adjusted the tablet to ensure that the participant's head is aligned with the device at an appropriate distance (approximately 50 centimeters) for effective face capture. The ETD structure and app design are depicted in Figure 1. Paradigm 1 uses a conventional self-rated scale, specifically the PHQ-9. Paradigm 2 encompasses a question-and-answering (Q&A) paradigm simulating psychiatric examinations. Paradigm 3 requires participants to describe images with the hint words [30-33]. Paradigm 4 presents three video clips of varying emotional sentiment scores [34,35] in a positive, neutral, and negative sequence (41 - 73 seconds, average 60.33 seconds). Participants sequentially respond to these components, with recordings capturing their reactions during both the viewing and responding phases, including the PHQ-9 selections; the entire process takes approximately 5 minutes. It is essential to clarify that the scale was only used to elicit reactions, and its scores did not contribute to the predictions.

XSL•FO
RenderX

Figure 1. Components of the Electronic Tool for Depression (ETD). PHQ-9: Patient Health Questionnaire-9; Q&A: question-and-answer.

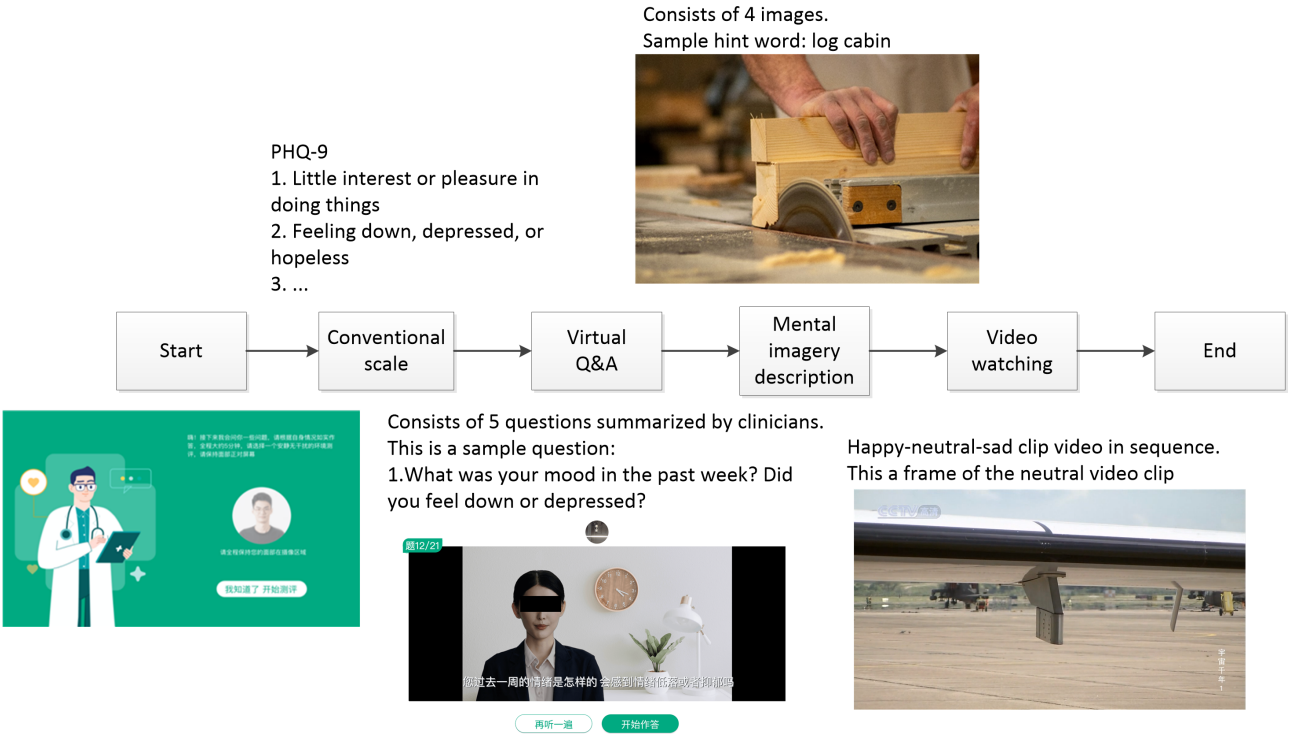

Recruitment

In this study, 89 participants were recruited from April 2022 to December 2022 in Beijing. Among these, 51 were recruited from the Beijing Anding Hospital Inpatient Department and were all diagnosed with MDD by experienced psychiatrists according to the International Code of Diseases, tenth revision (*ICD-10*) [36]. All participants met the inclusion criteria, which were as follows: (1) age 18 - 65 years, (2) proficiency in standard Chinese, (3) educational level of primary school or above, and (4) ability to understand and cooperate with the research protocol.

Exclusion criteria included (1) diagnosis of schizophrenia, schizoaffective disorder, or other mental disorders and (2) history of organic brain disease. The remaining 38 participants were recruited openly from the general population (employees and college students) who were not experiencing depression-related symptoms.

Participants from the hospital completed two steps: the first involved using the ETD app, and the second included assessment using the HAMD-17 scale by clinicians. Community participants only completed the ETD test, and all were confirmed to have no depressive symptoms based on the PHQ-9 assessment. Finally, the asymptomatic and the healthy control groups formed the nonMDD group (48 participants), while the mild group and the moderate or severe group were collectively referred to as the MDD group (41 participants). For ease of explanation, the mild group was designated as MDD-sub1, and the moderate or severe group was designated as MDD-sub2. Sex was compared using the X^2 test; age and HAMD scores were compared using the Mann-Whitney *U* test. There were no significant differences in sex ratio or age between the groups, while the MDD group had significantly higher PHQ-9 scores than the nonMDD group.

Ethical Considerations

This study was approved by the Ethics Committee of Beijing Anding Hospital Capital Medical University. No compensation fee was paid to participants, with written informed consent obtained for the data usage of research analysis. Data were deidentified and all analyses followed data privacy guidelines.

Model Training

All recorded videos underwent a manual verification process to ensure that the image ratio of a complete head, face, and eyes exceeded the empirical 95% threshold. We adopted the MFCC-based recurrent neural network (RNN) [29] as the validation model, which used a multimodal method that integrated MFCC and AU features and achieved a SOTA accuracy of 95.6% in binary classification of depression on the DAIC-WOZ (Distress Analysis Interview Corpus) dataset [18]. We pretrained the RNN on RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song dataset) [37], aggregated the AU features, and fine-tuned the model on our dataset. We developed a CNN model for AU detection using EfficientNet [38] on BP4D [39], achieving a mean F_1-score of 0.76 on selected AUs. The sample size of video clips for the RNN was 11,075, comprising 6826 normal, 2933 mild, and 1316 moderate or severe instances. We established a clip-voting ratio to represent the individual results. While the RNN simultaneously processed local audio and visual data, it did not incorporate conventional scale and video watching. To address this limitation, we proposed a global feature extraction method (depicted in Figure 2) to derive global features and build AI models. For the vision modality, we used Gaze360 [40] and Dlib [41] along with AU features to estimate gaze and head orientation. For the audio modality, we extracted MFCC-based features and the pure audio duration of the human voice. For the text modality, we calculated sentiment scores using the *pyltp*

XSL•FO
RenderX

package [42]. The features were concatenated in the order of visual, audio, and text features. Additionally, we incorporated statistical characteristics such as mean and variance to enhance their global representation. Normalization and bias adjustments were applied to ensure that all the features were positive for later attention computation (complete feature list is provided in Table S2 in Multimedia Appendix 1).

Figure 2. The global feature extraction method architecture. MFCC: mel frequency cepstral coefficients; MLP: multilayer perceptron.

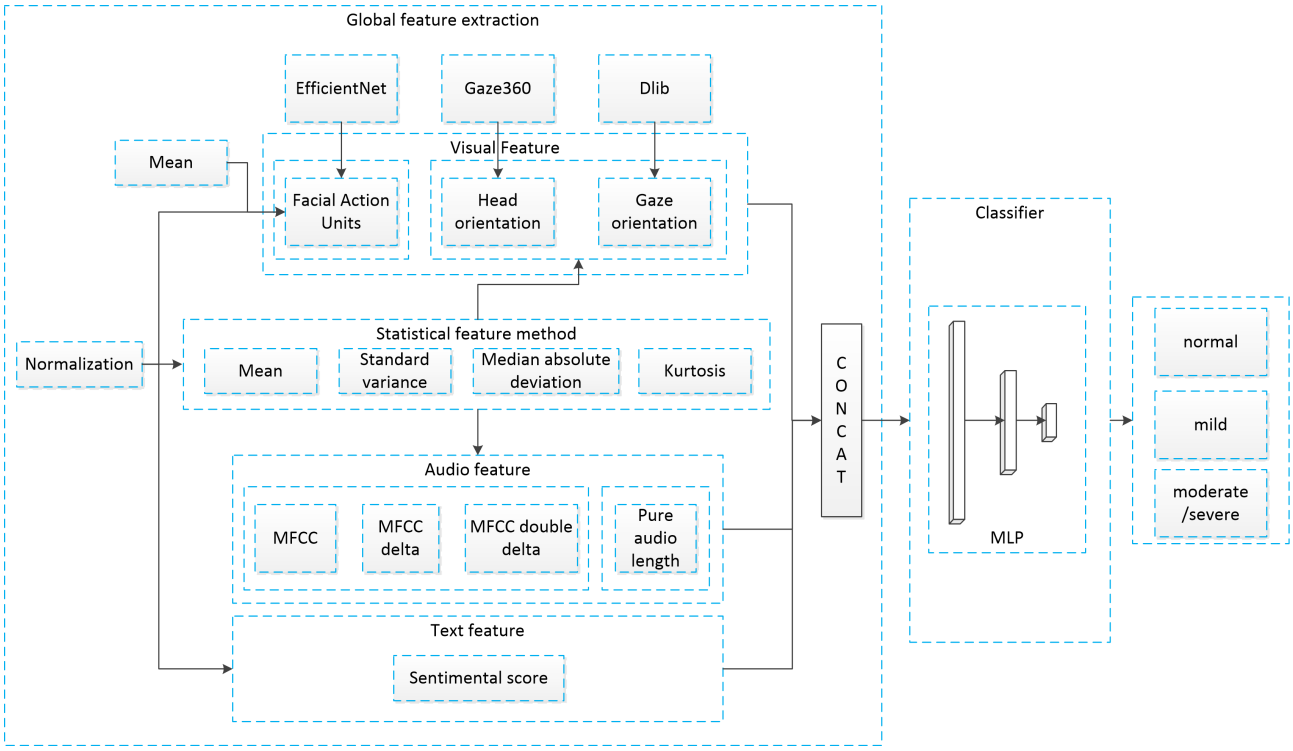

We adopted a multilayer perceptron (MLP) as our classifier for global predictions, which is identical to the RNN. The MLP comprises layers with 512, 1024, 128, and 3 neurons, incorporating batch normalization and a 0.2 dropout rate to mitigate overfitting; a softmax layer was added as the prediction. The Adam optimizer was used with a base learning rate of 1e-3, β_1 of 0.9, β_2 of 0.999, and ε of 1e-8. Given that deep learning methods are often considered "dark magic," we sought to enhance comprehensibility by employing Grad-Cam [43] to visualize the attention scores of the MLP's best-performing model across each feature. These results were then compared with Spearman and Kendall correlation coefficients computed using scikit-learn [44].

Statistical Analysis

Ablation experiments were conducted on various paradigm combinations. The models predicted three levels of severity, and binary performance was assessed to distinguish between depressed or nondepressed states. Sensitivity, specificity, accuracy, and area under the curve (AUC) were measured for binary results. Accuracy was specifically calculated for severity predictions. The five-fold performances underwent the Friedman test, followed by the posthoc Nemenyi test and Cliff δ effect size. A 95% CI was computed using bootstrapping, with the exception of single fold clip prediction AUC, which used normal approximation.

Results

Demographic characteristics are shown in Table 1 and the findings of clip prediction and the Friedman test are presented in Table 2. The difference among Q&A, MID, and QI (combination of Q&A and MID) is significant in binary sensitivity (P=.02), with a large effect (ε^2=0.47). The differences in binary AUC and extent accuracy are close to significant (P=.07 and P=.09, respectively), with large effects (ε^2=0.26 and ε^2=0.23, respectively). The differences in binary specificity and binary accuracy were not significant and exhibited small effects. Posthoc Nemenyi test results for sensitivity are detailed in Table 3, revealing that Q&A outperformed MID (P= .03) with a large effect size (Cliff δ=1.0). The difference between QI and MID is close to significant (P=.06) with a large effect (Cliff δ=1.0).

The results of individual prediction and the Friedman test are presented in Table 4. In the RNN voting analysis, the differences among Q&A, MID, and QI were significant in terms of binary sensitivity (P<.01) with a large effect (ε^2=0.61). The differences in binary accuracy and binary AUC were nonsignificant (P =.13 and P=.09, respectively) but showed large effects (ε^2=0.18 and ε^2=0.23, respectively). Posthoc Nemenyi test results on extent accuracy are presented in Table 4.

XSL•FO
RenderX

Table . Demographic characteristics.

Factors	MDD[a]	nonMDD[b]	*P* value
Sex, n (%)			.13[c]
Male	12 (29.3)	18 (37.5)	
Female	29 (70.3)	30 (62.5)	
Age (years), mean (SD)	38.41 (15.12)	35.98(12.37)	.50[d]
HAM-D[e], mean (SD)	14.51 (4.66)	–[f]	–
PHQ-9[g], mean (SD)	13.05 (6.00)	4.17 (2.71)	<.001[d]

[a]MDD: major depressive disorder.
[b]nonMDD: non-major depressive disorder.
[c]Chi-square test was used to derive the *P* value.
[d]Mann-Whitney U test was used to derive the *P* value.
[e]HAM-D: Hamilton rating scale for Depression.
[f]Not applicable.
[g]PHQ-9: PHQ-9: Patient Health Questionnaire-9.

Table . Clip prediction results of the MFCC-based[a] RNN[b] [29] for paradigm combinations.

Paradigm/statistics/performance	Sensitivity (%), mean (95% CI)	Specificity (%), mean (95% CI)	SA[c] (%), mean (95% CI)	AUC[d] (%), mean (95% CI)	EA[e] (%), mean (95% CI)
Q&A[f]	79.06 (77.06 - 83.35)	85.71 (73.30 - 90.19)	83.01 (74.43 - 86.10)	78.12 (66.11 - 82.35)	88.70 (83.15 - 91.33)
MID[g]	56.99 (41.78 - 63.36)	85.10 (76.36 - 89.81)	76.25 (69.90 - 80.30)	70.40 (65.09 - 73.60)	81.43 (76.42 - 85.35)
QI[h]	80.22 (75.88 - 84.72)	85.61 (77.06 - 89.50)	84.41 (81.98 - 86.44)	80.36 (76.45 - 82.78)	90.37 (88.81 - 91.64)
P value[i]	.02	.55	.25	.07	.09
Effect size (ε^2)	0.47	0.00	0.07	0.27	0.23

[a]MFCC: mel frequency cepstral coefficients.
[b]RNN: recurrent neural network.
[c]SA: screen accuracy.
[d]AUC: area under the curve.
[e]EA: extent accuracy.
[f]Q&A: question-and-answer.
[g]MID: mental imagery description.
[h]QI: combination of Q&A and MID.
[i]Friedman test was used to calculate the *P* value.

XSL•FO
RenderX

Table . Posthoc test results of the MFCC-based[a] RNN[b] clip prediction sensitivity between pairs of Q&A[c], MID[d], and QI[e].

Paradigm statistic item	*P* value[f]	Cliff δ (effect size)
Q&A-MID[g]	.03	1.0 (large)
Q&A-QI[h]	.90	−0.04 (negligible)
QI-MID[i]	.06	1.0 (large)

[a]MFCC: mel frequency cepstral coefficients.
[b]RNN: recurrent neural network.
[c]Q&A: question-and-answer.
[d]MID: mental imagery description.
[e]QI: combination of Q&A and MID.
[f]Nemenyi test.
[g]Q&A-MID: comparison between Q&A and mental imagery description.
[h]Q&A-QI: comparison between Q&A and combination of Q&A and mental imagery description.
[i]QI-MID: combination of Q&A and mental imagery description, and single paradigm mental imagery description.

Table . Individual prediction results of the MFCC-based[a] RNN[b] voting for paradigm combinations.

Paradigm performance	Sensitivity (%), mean (95% CI)	Specificity (%), mean (95% CI)	SA[c] (%), mean (95% CI)	AUC[d] (%), mean (95% CI)	EA[e] (%), mean (95% CI)
Q&A[f]	75.00 (57.50 - 87.50)	82.22 (68.89 - 86.67)	78.82 (72.94 - 83.53)	90.83 (84.17 - 95.83)	70.59 (57.65 - 77.65)
MID[g]	60.00 (42.50 - 67.50)	88.89 (77.78 - 95.56)	75.29 (62.35 - 80.00)	85.00 (81.94 - 90.45)	65.88 (57.65 - 71.77)
SQI[h]	75.00 (50.00 - 87.50)	88.89 (80.00 - 93.33)	82.36 (69.42 - 89.42)	92.50 (86.11 - 96.11)	80.00 (65.88 - 88.24)
P value	.29	.26	.13	.09	.009
Effect size	0.04	0.06	0.18	0.23	0.61

[a]MFCC: mel frequency cepstral coefficients.
[b]RNN: recurrent neural network.
[c]SA: screen accuracy.
[d]AUC: area under the curve.
[e]EA: extent accuracy.
[f]Q&A: question-and-answer.
[g]MID: mental imagery description.
[h]SQI: combination of conventional questionnaire, Q&A, and mental imagery description.

Table 5. QI outperformed MID ($P<.05$) with a substantial effect (Cliff δ=0.64). The difference between Q&A and QI was nonsignificant ($P=.14$) but indicated a large effect (Cliff δ=−0.48). In the global feature MLP analysis, differences among the paradigms were insignificant and exhibited small effect sizes, with results in Table S3 in Multimedia Appendix 1.

The best fold performance is shown in Table 6. The global feature, SQIV (combination paradigm of CS, Q&A, MID, and VW) MLP achieved a peak individual binary accuracy of 94.12%. Notably, the RNN voting SQI model also achieved a top accuracy of 94.12%, but with a higher extent accuracy of 94.12%, and an AUC of 0.99.

XSL•FO
RenderX

Table . Post-hoc statistic test results of the RNN[a] voting individual prediction extent accuracy between pairs of Q&A,[b] MID[c], and QI[d].

Paradigm statistic item	*P* value[e]	Effect size (Cliff δ)
Q&A-MID[f]	.60	0.32 (small)
Q&A-QI[g]	.14	−0.48 (large)
QI-MID[h]	.01	0.64 (large)

[a]RNN: recurrent neural network.
[b]Q&A: question-and-answer.
[c]MID: mental imagery description.
[d]QI: combination of Q&A and MID.
[e]Nemenyi test.
[f]Q&A-MID: comparison between Q&A and mental imagery description.
[g]Q&A-QI: comparison between Q&A and combination of Q&A and mental imagery description.
[h]QI-MID: combination of Q&A and mental imagery description, and single paradigm mental imagery description.

Table . Performance of the best fold of the global feature MLP[a] SQIV[b] model, the MFCC-based[c] RNN[d] SQI[e] clip model, and the MFCC-based RNN SQI[e] voting model.

Method	Paradigm performance	Sensitivity %, (95% CI)	Specificity %, (95% CI)	SA[f] %, (95% CI)	AUC[g] (95% CI)	EA[h] %(95% CI)
MLP[a]	SQIV[b]	100.0 (63.06 - 100.0)	88.89 (51.75 - 99.72)	94.12 (71.31 - 99.85)	0.97 (0.87 - 1.0)	76.47 (50.10 - 93.19)
RNN[d]	SQI[e]	80.69 (77.74 - 83.64)	89.93 (88.60 - 91.26)	87.36 (86.09 - 88.63)	0.91 (0.90 - 0.92)	83.03 (81.60 - 84.46)
RNN[d] voting	SQI[e]	100.0 (63.06 - 100.0)	88.89 (51.75 - 99.72)	94.12 (71.31 - 99.85)	0.99 (0.91 - 1.0)	94.12 (71.31 - 99.85)

[a]MLP: multilayer perceptron.
[b]SQIV: combination of conventional scale, Q&A, mental imagery description, and video watching.
[c]MFCC: mel frequency cepstral coefficients.
[d]RNN: recurrent neural network.
[e]SQI: combination of conventional scale, Q&A, and mental imagery description
[f]SA: screen accuracy
[g]AUC: area under the curve
[h]EA: extent accuracy

The test results of comparison between the RNN voting and the proposed global feature method can be found in Table S4 in Multimedia Appendix 1. The input data are the same in Q&A and MID; for paradigm combinations, we compared RNN-voting QI and the global feature method (ie, combination paradigm of CS, Q&A, and MID; SQI), as they use the most comparable input data. No statistically significant differences were identified between the two methods across all three paradigms. Figures S1-S4 in Multimedia Appendix 1 illustrate the collective and individual learnings of the best MLP SQIV model. The mean attention scores of the features are sequenced by the nonMDD group, the MDD-sub1 group, and the MDD-sub2 group. The complete attention scores with Spearman correlation scores are mentioned in Table S5 in Multimedia Appendix 1, sorted in descending order for the MDD-sub2 group. Previously analyzed nonverbal markers in studies by Fiquer et al [20] and Girard et al [45] can be found in Table S6 in Multimedia Appendix 1. As shown in Figure S1 in Multimedia Appendix 1, the different groups exhibit varying levels of attention to specific features. The Spearman and Kendall correlation coefficients for each feature relative to the target extent of depression are available in Table S5 in Multimedia Appendix 1, where 18 features demonstrated a *P* value<.05.

Discussion

Principal Findings

We aimed to evaluate the efficacy of different paradigms via AI on audiovisual signals. We aggregated the four paradigms within the ETD and held 5-fold cross-validation on the two AI models among the paradigm combinations. Our findings show that there are differences in paradigm efficacies, and the AI model learns knowledge consistent with prior human experience.

For the single paradigm with the MFCC-based RNN, Q&A outperformed MID in identifying patients but performed equally in distinguishing extent levels. The difference between Q&A and MID in clip sensitivity was significant, but nonsignificant in individual extent accuracy. This makes Q&A more precise in identifying MDD patients.

XSL•FO
RenderX

For paradigm combinations with the MFCC-based RNN, integrating MID with Q&A slightly decreased clip sensitivity significance but significantly improved individual extent accuracy significance compared with MID. Considering that the difference between QI and Q&A was nonsignificant in either clip sensitivity and extent accuracy, and the differences among Q&A, MID, and QI were nonsignificant in the other performance indexes, we conclude that Q&A demonstrated higher efficacy than MID and suggest that paradigm combinations perform better than a single paradigm. As known, Q&A is a simplified version of a clinical interview, and the questions are all symptom-related, which makes it the most relevant paradigm for MDD and the most important one.

In the individual prediction of the global feature MLP, no significant differences were observed across the paradigm combinations. When fixing the paradigms, no notable differences were found between the MFCC-based RNN voting and the global feature MLP, which validated the global features' effectiveness. Some large effect sizes were noted, particularly in binary AUC and extent accuracy for Q&A and QI, which may be attributable to feature granularity. The RNN feature integrates both local and global information, with local features benefiting from transfer learning, which enhances performance—achieving a top binary accuracy of 94.12%, a top mean binary accuracy of 82.36%, and a mean extent accuracy of 80.00%. In contrast, the global features may be coarse at the granular level. Even so, the MLP still achieved a mean binary accuracy ranging from 74.12% to 85.88%, with a 95% CI spanning 69.41% to 91.77%, and a top binary accuracy of 94.12%. Overfitting might exist and could result in wide CIs.

Compared with support vector machine models [7], which showed a mean binary accuracy of 78.95% on event-related fMRI [46] and 85.00% on block-related fMRI [47] and CNN [8], which achieved a mean binary accuracy of 85.62% on EEG data, the ETD demonstrated equivalent performance while relying on much more readily accessible daily data. Compared to SRSIs, the audiovisual data are more objective and easier to use. Compared to interview-based assessments such as the HAMD [2], the ETD required approximately 5 minutes, saving about 83% of the time. The ETD's performance and efficiency support its potential to objectively, accurately, and efficiently screen for MDD.

In the visualization, it is noteworthy that the high- and low-attention features did not intersect, particularly between the MDD group and the nonMDD groups, indicating that participants in different groups exhibited diverse behavioral patterns. Almost all 16 features mentioned by Fiquer et al [20] and Girard et al [21] exhibited high attention scores, with the lowest score being 0.68. Among these, 15 features ranked in the top 20%, except for "head down" which ranked 31st in MDD-sub1, demonstrating consistency with prior studies. Additionally, "head motion velocity" was not included in this study. When comparing with correlation results, 18 features emerged as having significant positive or negative relationships with MDD extent, both in feature items and in correlation trends observed via Spearman and Kendall methods—differing only in specific weights. Of these, 11 aligned with the correlation trends; 5 showed patterns in which the attention scores for the MDD group were either higher or lower than those of the nonMDD group, and only 2 showed no clear trends. For instance, the mean attention score of AU4, interpreted as "frown" by Fiquer et al [20], increased with the extent of MDD, and the Spearman correlation for AU4 was positive (P=.005). These high-attention score and correlation-consistent features may serve as urgently needed objective markers and should be further investigated.

The MLP leverages global features representing statistical values throughout the process, sacrificing some detail at the granular level while maintaining low model and computational complexity. Despite potential overfitting, the alignment of the visualization results with correlation findings indicates that the MLP has acquired knowledge consistent with medical prior knowledge, supporting its performance and underscores its potential as a valuable tool. For the inconsistent elements, the neural network introduces significant nonlinearity and captures relationships in high-dimensional spaces. In contrast, Spearman and Kendall correlations are limited to assessing relationships between single inputs and targets. We propose that a trained model can reveal complex multi-input–target relationships that are difficult to define manually. Furthermore, results may vary with the accumulation of additional data.

The ETD's efficiency—requiring less time and energy— and its objectivity and accuracy make it a flexible and practical tool to be applied across diverse medical scenes that prioritize lightweight and quietness, particularly in screening and health monitoring scenes. Multimodal analysis may produce better results; for instance, AUC of binary depression status increased from 0.72 to 0.76 with networked smartphone sensors combining to text messages [12], and the binary accuracy increased from 76.27% to 95.60% when AU features were added to acoustic features in the baseline MFCC-based RNN [29]. As wearable devices continue to gain popularity, easily obtainable physical signals such as ECG and photoplethysmography can be integrated as additional modalities to enhance clinical outcomes. In our work, we currently use visual, acoustic, and text information jointly, which we believe may be a key point in the high performance observed and should receive more attention in future studies. As audiovisual features are also related to other conditions such as anxiety disorder and schizophrenia [48], or to distinguish between MDD or bipolar depression [49], aggregating multiple paradigms may further improve efficacy.

Conclusions

The Q&A method showed greater efficacy compared to MID, and combining paradigms may yield better results than using individual paradigms alone. Visualization interpretation showed that the AI method acquired knowledge that aligns with medical expertise and identified several potentially significant markers. By applying AI to multimodal audiovisual signals, these findings position the ETD as a valuable, objective tool for screening MDD and show potential for applications across a broader spectrum of psychiatric disorders with various data modalities.

Limitations

The efficacy of the modalities remains inadequately explored. Automatically detected AUs may not achieve the reliability of

XSL•FO
RenderX

human-labeled results. Additionally, conclusions drawn from current analyses may require revision as sample sizes increase, particularly in deep learning frameworks.

Acknowledgments

The study was supported by Sci-Tech Innovation 2030 – Major Project of Brain science and brain-inspired intelligence technology (2021ZD0200600) and the Beijing Municipal Administration of Hospitals Incubating Program (PX2019068).

Data Availability

The datasets analyzed during this study are not publicly available as the agreement is only between the participants and researchers but are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Definitions, individual prediction results, and attention scores.
[DOCX File, 564 KB - formative_v9i1e56057_app1.docx]

References

1. Depression and other common mental disorders: global health estimates. World Health Organization. 2017. URL: https://www.who.int/publications/i/item/depression-global-health-estimates [accessed 2025-03-21]
2. HAMILTON M. A rating scale for depression. J Neurol Neurosurg Psychiatry 1960 Feb;23(1):56-62. [doi: 10.1136/jnnp.23.1.56] [Medline: 14399272]
3. Beck AT, Steer RA, Ball R, Ranieri W. Comparison of Beck Depression Inventories -IA and -II in psychiatric outpatients. J Pers Assess 1996 Dec;67(3):588-597. [doi: 10.1207/s15327752jpa6703_13] [Medline: 8991972]
4. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613. [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
5. Pichot P. Self-report inventories in the study of depression. In: Hippius H, Klerman GL, Matussek N, editors. New Results in Depression Research 1986:53-58. [doi: 10.1007/978-3-642-70702-5_7]
6. Ben-Porath YS. Assessing personality and psychopathology with self-report inventories. In: Handbook of Psychology 2003:553-577. [doi: 10.1002/0471264385]
7. Rosa MJ, Portugal L, Hahn T, et al. Sparse network-based models for patient classification using fMRI. Neuroimage 2015 Jan 15;105:493-506. [doi: 10.1016/j.neuroimage.2014.11.021] [Medline: 25463459]
8. Li X, La R, Wang Y, et al. EEG-based mild depression recognition using convolutional neural network. Med Biol Eng Comput 2019 Jun;57(6):1341-1352. [doi: 10.1007/s11517-019-01959-2]
9. Reece AG, Reagan AJ, Lix KLM, Dodds PS, Danforth CM, Langer EJ. Forecasting the onset and course of mental illness with Twitter data. Sci Rep 2017 Oct 11;7(1):13006. [doi: 10.1038/s41598-017-12961-9] [Medline: 29021528]
10. Schwartz HA, Eichstaedt J, Kern ML, et al. Towards assessing changes in degree of depression through facebook. Presented at: Proceedings of the Workshop on Computational Linguistics and Clinical Psychology; Jun 2014; Baltimore, Maryland, USA. [doi: 10.3115/v1/W14-3214]
11. Chandra Guntuku S, Buffone A, Jaidka K, Eichstaedt JC, Ungar LH. Understanding and measuring psychological stress using social media. ICWSM 2019;13(1):214-225. [doi: 10.1609/icwsm.v13i01.3223]
12. Liu T, Meyerhoff J, Eichstaedt JC, et al. The relationship between text message sentiment and self-reported depression. J Affect Disord 2022 Apr 1;302:7-14. [doi: 10.1016/j.jad.2021.12.048] [Medline: 34963643]
13. Cummins N, Scherer S, Krajewski J, Schnieder S, Epps J, Quatieri TF. A review of depression and suicide risk assessment using speech analysis. Speech Commun 2015 Jul;71:10-49. [doi: 10.1016/j.specom.2015.03.004]
14. Scherer S, Morency LP, Gratch J, Pestian J. Reduced vowel space is a robust indicator of psychological distress: a cross-corpus analysis. Presented at: ICASSP 2015 - 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); Apr 19-24, 2025; South Brisbane, Queensland, Australia. [doi: 10.1109/ICASSP.2015.7178880]
15. Logan B. Mel frequency cepstral coefficients for music modeling. Proc of Ismir 2000 [FREE Full text]
16. Cummins N, Epps J, Breakspear M, Goecke R. An investigation of depressed speech detection: features and normalization. Presented at: INTERSPEECH 2011; Aug 27-31, 2011; Florence, Italy URL: https://www.isca-archive.org/interspeech_2011 [doi: 10.21437/Interspeech.2011-750]
17. Kim AY, Jang EH, Lee SH, Choi KY, Park JG, Shin HC. Automatic depression detection using smartphone-based text-dependent speech signals: deep convolutional neural network approach. J Med Internet Res 2023 Jan 25;25:e34474. [doi: 10.2196/34474] [Medline: 36696160]

XSL•FO
RenderX

18. Gratch J, Artstein R, Lucas G. The distress analysis interview corpus of human and computer interviews. 2019 Oct 15 Presented at: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14); Reykjavik, Iceland p. 3123-3128 URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/508_Paper.pdf [accessed 2025-03-21]
19. Ekman P, Friesen WV. Facial action coding system (FACS): a technique for the measurement of facial actions. APA PsycTests. URL: https://doi.org/10.1037/t27734-000 [accessed 2025-03-21]
20. Fiquer JT, Moreno RA, Brunoni AR, Barros VB, Fernandes F, Gorenstein C. What is the nonverbal communication of depression? Assessing expressive differences between depressive patients and healthy volunteers during clinical interviews. J Affect Disord 2018 Oct 1;238:636-644. [doi: 10.1016/j.jad.2018.05.071] [Medline: 29957481]
21. Girard JM, Cohn JF, Mahoor MH, Mavadati S, Rosenwald DP. Social risk and depression: evidence from manual and automatic facial expression analysis. Presented at: 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG); Apr 22-26, 2013 p. 1-8. [doi: 10.1109/FG.2013.6553748] [Medline: 24598859]
22. Gavrilescu M, Vizireanu N. Predicting depression, anxiety, and stress levels from videos using the facial action coding system. Sensors (Basel) 2019 Aug 25;19(17):3693. [doi: 10.3390/s19173693] [Medline: 31450687]
23. Melfsen S, Osterlow J, Florin I. Deliberate emotional expressions of socially anxious children and their mothers. J Anxiety Disord 2000;14(3):249-261. [doi: 10.1016/s0887-6185(99)00037-7] [Medline: 10868983]
24. Smith MC, Smith MK, Ellgring H. Spontaneous and posed facial expression in Parkinson's disease. J Int Neuropsychol Soc 1996 Sep;2(5):383-391. [doi: 10.1017/s1355617700001454] [Medline: 9375163]
25. Mundt JC, Snyder PJ, Cannizzaro MS, Chappie K, Geralts DS. Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology. J Neurolinguistics 2007 Jan;20(1):50-64. [doi: 10.1016/j.jneuroling.2006.04.001] [Medline: 21253440]
26. McIntyre G, Gocke R, Hyett M, Green M, Breakspear M. An approach for automatically measuring facial activity in depressed subjects. Presented at: 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops (ACII 2009); Sep 10-12, 2009; Amsterdam p. 1-8. [doi: 10.1109/ACII.2009.5349593]
27. Cai H, Yuan Z, Gao Y, et al. A multi-modal open dataset for mental-disorder analysis. Sci Data 2022 Apr 19;9(1):178. [doi: 10.1038/s41597-022-01211-x] [Medline: 35440583]
28. Ray A, Kumar S, Reddy R, Mukherjee P, Garg R. Multi-level attention network using text, audio and video for depression prediction. Presented at: AVEC '19: Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop; Oct 15, 2019; Nice France p. 81-88. [doi: 10.1145/3347320.3357697]
29. Rejaibi E, Komaty A, Meriaudeau F, Agrebi S, Othmani A. MFCC-based recurrent neural network for automatic clinical depression recognition and assessment from speech. Biomed Signal Process Control 2022 Jan;71:103107. [doi: 10.1016/j.bspc.2021.103107]
30. Holmes EA, Mathews A, Mackintosh B, Dalgleish T. The causal effect of mental imagery on emotion assessed using picture-word cues. Emotion 2008 Jun;8(3):395-409. [doi: 10.1037/1528-3542.8.3.395] [Medline: 18540755]
31. Weßlau C, Cloos M, Höfling V, Steil R. Visual mental imagery and symptoms of depression - results from a large-scale web-based study. BMC Psychiatry 2015 Dec 2;15:308. [doi: 10.1186/s12888-015-0689-1] [Medline: 26631081]
32. Andrade J, May J, Deeprose C, Baugh SJ, Ganis G. Assessing vividness of mental imagery: The Plymouth Sensory Imagery Questionnaire. Br J Psychol 2014 Nov;105(4):547-563. [doi: 10.1111/bjop.12050] [Medline: 24117327]
33. Tiggemann M, Kemps E. The phenomenology of food cravings: the role of mental imagery. Appetite 2005 Dec;45(3):305-313. [doi: 10.1016/j.appet.2005.06.004] [Medline: 16112776]
34. Ge Y, Zhao G, Zhang Y, Houston RJ, Song J. A standardised database of Chinese emotional film clips. Cogn Emot 2019 Aug;33(5):976-990. [doi: 10.1080/02699931.2018.1530197] [Medline: 30293475]
35. Liu YJ, Yu M, Zhao G, Song J, Ge Y, Shi Y. Real-Time movie-induced discrete emotion recognition from EEG Signals. IEEE Trans Affective Comput 2018;9(4):550-562. [doi: 10.1109/TAFFC.2017.2660485]
36. International Advisory Group for the Revision of ICD-10 Mental and Behavioural Disorders. A conceptual framework for the revision of the ICD-10 classification of mental and behavioural disorders. World Psychiatry 2011 Jun;10(2):86-92. [doi: 10.1002/j.2051-5545.2011.tb00022.x] [Medline: 21633677]
37. Livingstone SR, Russo FA. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 2018 May 16;13(5):e0196391. [doi: 10.1371/journal.pone.0196391]
38. Tan M, Le Q. EfficientNet: rethinking model scaling for convolutional neural networks. 2019 Presented at: Proceedings of the 36th International Conference on Machine Learning p. 6105-6114 URL: https://proceedings.mlr.press/v97/tan19a.html [accessed 2025-03-21]
39. Yin L, Chen X, Sun Y, Worm T, Reale M. A high-resolution 3D dynamic facial expression database. Presented at: 2008 8th IEEE International Conference on Automatic Face & Gesture Recognition; Sep 17-19, 2008; Amsterdam, Netherlands. [doi: 10.1109/AFGR.2008.4813324]
40. Kellnhofer P, Recasens A, Stent S, Matusik W, Torralba A. Gaze360: physically unconstrained gaze estimation in the wild. Presented at: 2019 IEEE/CVF International Conference on Computer Vision (ICCV); Oct 27 to Nov 2, 2019; Seoul, Korea (South p. 6912-6921. [doi: 10.1109/ICCV.2019.00701]

XSL•FO
RenderX

41. King DE. Dlib-ml: A machine learning toolkit. J Mach Learn Res 2009;10(3):1755-1758. [doi: 10.1145/1577069.1755843]
42. Che W, Feng Y, Qin L, Liu T. N-LTP: an open-source neural language technology platform for chinese. Presented at: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing; Nov 2021; Online and Punta Cana, Dominican Republic. [doi: 10.18653/v1/2021.emnlp-demo.6]
43. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. Presented at: 2017 IEEE International Conference on Computer Vision (ICCV); Oct 22-29, 2017; Venice p. 618-626. [doi: 10.1109/ICCV.2017.74]
44. Pedregosa F, Varoquaux G, Gramfort A. Scikit-learn: Machine learning in Python. JMLR 2011;12:2825-2830 [FREE Full text]
45. Girard JM, Cohn JF, Mahoor MH, Mavadati SM, Hammal Z, Rosenwald DP. Nonverbal social withdrawal in depression: evidence from manual and automatic analysis. Image Vis Comput 2014 Oct;32(10):641-647. [doi: 10.1016/j.imavis.2013.12.007] [Medline: 25378765]
46. Fu CHY, Williams SCR, Cleare AJ, et al. Attenuation of the neural response to sad faces in major depression by antidepressant treatment: a prospective, event-related functional magnetic resonance imaging study. Arch Gen Psychiatry 2004 Sep;61(9):877-889. [doi: 10.1001/archpsyc.61.9.877] [Medline: 15351766]
47. Hahn T, Marquand AF, Ehlis AC, et al. Integrating neurobiological markers of depression. Arch Gen Psychiatry 2011 Apr;68(4):361-368. [doi: 10.1001/archgenpsychiatry.2010.178] [Medline: 21135315]
48. Abbas A, Hansen BJ, Koesmahargyo V, et al. Facial and vocal markers of schizophrenia measured using remote smartphone assessments: observational study. JMIR Form Res 2022 Jan 21;6(1):e26276. [doi: 10.2196/26276] [Medline: 35060906]
49. Ruihua M, Meng Z, Nan C, et al. Differences in facial expression recognition between unipolar and bipolar depression. Front Psychol 2021;12:619368. [doi: 10.3389/fpsyg.2021.619368] [Medline: 34335353]

Abbreviations

AI: artificial intelligence
AU: action unit
AUC: area under the curve
CNN: conventional neural network
CS: conventional scale
DL: deep learning
EEG: electroencephalogram
ETD: electronic tool for depression
fMRI: functional magnetic resonance imaging
HAMD: Hamilton rating scale for depression
MDD: major depressive disorder
MFCC: mel frequency cepstral coefficients
MID: mental imagery description
PHQ-8: Patient Health Questionnaire-8
PHQ-9: Patient Health Questionnaire-9
Q&A: question and answering
QI: combination of Q&A and MID
RNN: recurrent neural network
SOTA: state-of-the-art
SQ: combination paradigm of CS and Q&A
SQIV: combination paradigm of CS, Q&A, MID, and VW
SRSI: Self-Report Scales and Inventories
VW: video-watching paradigm

Edited by A Mavragani; submitted 23.02.24; peer-reviewed by ST Arasteh, Z Shen; revised version received 02.05.25; accepted 06.05.25; published 30.05.25.

Please cite as:
Chen D, Wang P, Zhang X, Qiao R, Li N, Zhang X, Zhang H, Wang G
Comparative Efficacy of MultiModal AI Methods in Screening for Major Depressive Disorder: Machine Learning Model Development Predictive Pilot Study
JMIR Form Res 2025;9:e56057
URL: https://formative.jmir.org/2025/1/e56057
doi:10.2196/56057

XSL•FO
RenderX

© Donghao Chen, Pengfei Wang, Xiaolong Zhang, Runqi Qiao, Nanxi Li, Xiaodong Zhang, Honggang Zhang, Gang Wang. Originally published in JMIR Formative Research (https://formative.jmir.org), 30.5.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Comparison of ChatGPT and Internet Research for Clinical Research and Decision-Making in Occupational Medicine: Randomized Controlled Trial

Felix A Weuthen[1]; Nelly Otte[1], MD; Hanif Krabbe[2], MD; Thomas Kraus[1], MD; Julia Krabbe[1,3], MD

[1]Institute of Occupational, Social and Environmental Medicine, Medical Faculty, Rheinisch-Westfälische Technische Hochschule Aachen University, Aachen, Germany

[2]Department of Vascular Surgery, St. Josef Hospital Bochum, Katholisches Klinikum Bochum, Medical Faculty, Ruhr University Bochum, Bochum, Germany

[3]Institute for Prevention and Occupational Medicine of the German Social Accident Insurance, Medical Faculty, Ruhr University Bochum, Bürkle de la Camp Platz 1, Bochum, Germany

Corresponding Author:
Julia Krabbe, MD
Institute of Occupational, Social and Environmental Medicine, Medical Faculty, Rheinisch-Westfälische Technische Hochschule Aachen University, Aachen, Germany

Abstract

Background: Artificial intelligence is becoming a part of daily life and the medical field. Generative artificial intelligence models, such as GPT-4 and ChatGPT, are experiencing a surge in popularity due to their enhanced performance and reliability. However, the application of these models in specialized domains, such as occupational medicine, remains largely unexplored.

Objective: This study aims to assess the potential suitability of a generative large language model, such as ChatGPT, as a support tool for medical research and even clinical decisions in occupational medicine in Germany.

Methods: In this randomized controlled study, the usability of ChatGPT for medical research and clinical decision-making was investigated using a web application developed for this purpose. Eligibility criteria were being a physician or medical student. Participants (N=56) were asked to work on 3 cases of occupational lung diseases and answer case-related questions. They were allocated via coin weighted for proportions of physicians in each group into 2 groups. One group researched the cases using an integrated chat application similar to ChatGPT based on the latest GPT-4-Turbo model, while the other used their usual research methods, such as Google, Amboss, or DocCheck. The primary outcome was case performance based on correct answers, while secondary outcomes included changes in specific question accuracy and self-assessed occupational medicine expertise before and after case processing. Group assignment was not traditionally blinded, as the chat window indicated membership; participants only knew the study examined web-based research, not group specifics.

Results: Participants of the ChatGPT group (n=27) showed better performance in specific research, for example, for potentially hazardous substances or activities (eg, case 1: ChatGPT group 2.5 hazardous substances that cause pleural changes versus 1.8 in a group with own research; P=.01; Cohen r=–0.38), and led to an increase in self-assessment with regard to specialist knowledge (from 3.9 to 3.4 in the ChatGPT group vs from 3.5 to 3.4 in the own research group; German school grades between 1=very good and 6=unsatisfactory; P=.047). However, clinical decisions, for example, whether an occupational disease report should be filed, were more often made correctly as a result of the participant's own research (n=29; eg, case 1: Should an occupational disease report be filed? Yes for 7 participants in the ChatGPT group vs 14 in their own research group; P=.007; odds ratio 6.00, 95% CI 1.54 - 23.36).

Conclusions: ChatGPT can be a useful tool for targeted medical research, even for rather specific questions in occupational medicine regarding occupational diseases. However, clinical decisions should currently only be supported and not made by the large language model. Future systems should be critically assessed, even if the initial results are promising.

Trial Registration: German Clinical Trials Registry DRKS00036492; https://drks.de/search/en/trial/DRKS00036492/entails

(JMIR Form Res 2025;9:e63857) doi:10.2196/63857

KEYWORDS

occupational diseases; artificial intelligence; internet research; medicine; occupational medicine; ChatGPT; AI; decision-making; algorithms; algorithm; large language models; LLMs; physicians; medical students; occupational lung diseases; occupational lung disease

XSL•FO
RenderX

Introduction

The application of artificial intelligence in the field of medicine has a long history, dating back to the mid-20th century. Initially used in research, its use in clinical medicine emerged in the 1970s. The MYCIN computer expert system was used at Stanford University for the purpose of diagnosing and treating infectious diseases with antibiotics [1]. Even though the diagnoses produced by the system at that time exhibited remarkably high success rates, it was not accepted at that time. The recent rapid innovation of large language models (LLMs) has led to the emergence of ChatGPT, which is the first LLM to provide the data basis and performance to support or carry out medical research and clinical decisions. Nevertheless, the clinical application is currently viewed with a degree of skepticism, as ChatGPT, especially in the earlier versions 2 and 3, demonstrated a marked tendency to "confabulate," to fabricate statements and even references [2]. This phenomenon is frequently referred to as "hallucinating" in the literature [3]. Following the upgrade to ChatGPT 4, which includes a Bing internet connection, the tendency to hallucinate has been reported to have decreased significantly, allowing well-founded statements to be made. However, it is essential to subject LLM information to rigorous scrutiny and verification for accuracy.

ChatGPT has been used in patient care for some time, for example, for the creation of informational pamphlets [4], the evaluation of educational videos [5], and also in radiological [6] or dermatological [7] diagnostics. Nevertheless, there are currently significant discrepancies between the assessments of the relevant experts, and therefore, the use of this technology without proper oversight is not recommended [4,5,7]. Occupational medicine is a small specialty at the interface between work and medicine. One main field of activity is the prevention or early detection of occupational diseases. In Germany, an occupational disease can only be recognized officially if there is a disease "that insured employees suffer as a consequence of the occupational activities they perform in the course of their jobs and that are listed in the Ordinance on Occupational Diseases in Germany [8]. Almost 80% of annual deaths are caused by occupational diseases of the lungs in Germany [9]. However, particularly nonoccupational physicians are frequently uncertain as to whether the clinical presentation of patients and their occupational history justify the reporting of an occupational disease. Furthermore, the results of an internet search are often inconclusive, particularly with regard to German occupational disease law.

This project originated from the clinical experience and routine of an occupational medicine institute at a university hospital, as well as interactions with colleagues during consultations. An application that facilitates targeted research or even indicates whether there is a justified suspicion of an occupational disease could have a positive impact on the daily work of doctors and on the probable high number of unreported occupational illnesses that are not reported.

The objective of this study was to assess the potential suitability of generative LLM, such as ChatGPT, as a support tool for medical research and even clinical decisions in occupational medicine in Germany. In particular, the first insights into the potential for such technology to provide assistance with questions pertaining to occupational disease law and the practice of daily medical care should be provided. Physicians and medical students were invited to work on 3 occupational medicine cases regarding occupational lung diseases within a web-based application. One randomly selected group was prompted to use an integrated chat application with input in ChatGPT, whereas the other was instructed to use their customary research instruments, including web-based search engines such as Google and, in Germany frequently used medical information websites such as Amboss [10] or DocCheck [11]. The responses provided were subjected to quantitative assessment based on the number of correct answers.

Methods

Participants

In this randomized controlled study, medical students and doctors were recruited via announcements on notice boards and personal contact. Flyers with information about the study and a QR code to the study website were shared in web-based and analog notice boards, as well as student messenger groups at the university hospital. Flyers were also distributed in other hospitals and rescue helicopter stations via contact persons who received the flyers and hung them on the notice boards. The web-based study was conducted in German via a web-based application. The inclusion criterion was the indication of current medical studies and semester or practice as a physician and specialty. An exclusion criterion was the use of ChatGPT in the group, which should use their own research tools. To avoid unconscious influence toward the use of ChatGPT, the participants were not informed of this at the start of the study. No specific sample size calculation was carried out beforehand on the assumption that recruitment per se would be rather difficult due to the effort involved. A minimum of 50 participants was set as a minimum number in a 3-month interval.

With a sum of 3 weeks of lectures, occupational medicine is a very small part of the study program at medical universities. Physicians and medical students both were expected to have a similar level of knowledge in relation to occupational medicine, since occupational health aspects play only a very small role, if any, in specialist medical training. Thus, both groups were included in the study.

The study process is depicted in Figure 1. As the first step, demographic data and a self-assessment of occupational medical knowledge were requested according to German school grades between 1 and 6, with 1=very good being the best grade and 6=unsatisfactory being the worst. The respondents were then asked 6 questions on occupational diseases, which they were asked to answer from memory without any research.

They were then randomized into one of two groups: (1) research with the integrated chat window, which enabled a query with ChatGPT, or (2) research with the research tools familiar to the person. The person was free to choose which tools were used; they were only asked to name these tools after the case studies. A digital weighted coin was flipped for the group assignment.

XSL•FO
RenderX

If there was an unequal distribution of doctors and students, the next group assignment with the next coin toss was more likely to be allocated to the other group. Afterward, the participants were immediately referred to case processing.

The group assignment was not blinded in the traditional way, since group membership was indicated by the presence or absence of the chat window. However, the participants did not learn the exact group characteristics in advance only that web-based research was to be examined.

Figure 1. Study procedure. After the sociodemographic and knowledge questionnaire, participants were assigned to one of the 2 groups. After processing of case 1, participants could decide if they work on case 2 or proceed to the final questionnaire. The same choice was given after case 2.

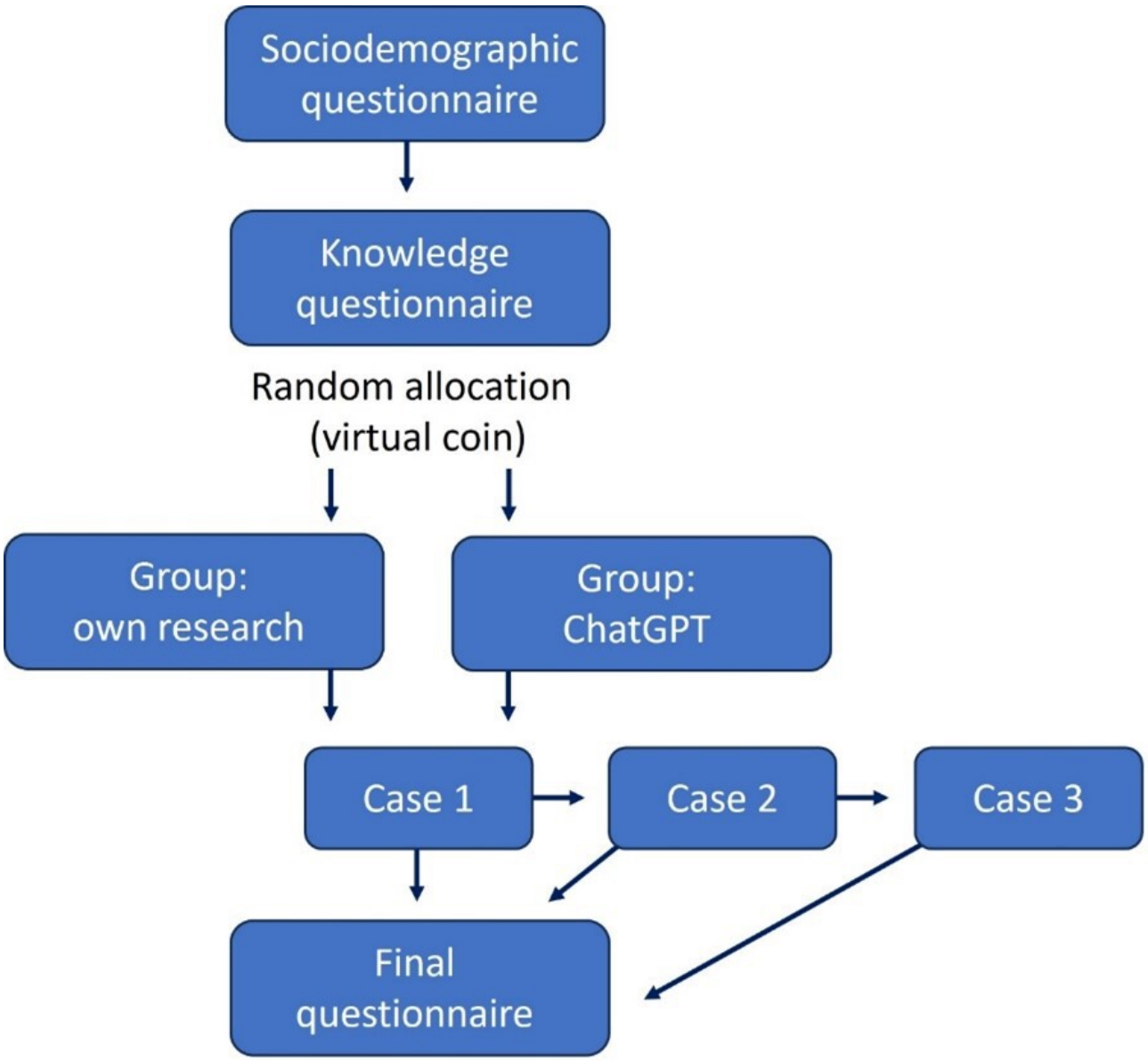

Questions and Cases

The study was conducted between March 1, 2024, and May 31, 2024, as planned for a 3-month period. The English translation of the survey conducted in German can be found in Multimedia Appendix 1. Three cases were offered, each with 6 questions. All 3 cases were based on real patients in occupational medicine practice and only slightly altered for the study. Cases were then examined for accuracy by 2 occupational health specialists who specialize in occupational lung diseases. The correct answers were determined in advance. A pilot test with the setup was tested by 3 pilot testers, a medical student, a nonoccupational physician, and an occupational health specialist. According to their comments, slight adjustments were made, for example, copying the questions directly into the chat window was disabled.

Case 1 is based on the case of an outdoor worker who was treated for cholangiocellular carcinoma. As an incidental finding, asbestos-associated changes in the pleura were found in the computed tomography of the thorax. An occupational disease report was made with a justified suspicion of an occupational disease according to the Occupational Diseases Ordinance in Germany. Case 2 is based on the case of a young woman who worked in galvanization and developed a sensitization and allergy to a metal sulfate. Case 3 was based on a former dental technician with a recognized occupational disease (berylliosis).

XSL•FO
RenderX

Exactly 6 questions were asked for each case, which were to be answered with yes or no, multiple choice, or free-text options. For each question, there was also a “don’t know” option. The questions can be viewed in Multimedia Appendix 1. Three questions were always presented on the screen at the same time as the introductory case vignette. Next to it was either a window for integrated ChatGPT input (Figure 2) or the indication that the usual means should be used freely. Each question had to be answered or the “don’t know” option had to be checked. After completing the 3 questions, the user moved on to the next 3. It was then no longer possible to go back and change the answers, and rule out learning effects from the later questions. In order to record how exactly the respondents entered their answers, it was also not possible to copy the questions or the case vignette and paste them into the chat window.

The number of correct answers or “don’t know” answers was counted in the evaluation. For questions that were offered for answering without research before the group assignment (before), a comparison of before and after was also carried out.

After each case, it was possible to choose whether another case should be processed or whether the respondent should be forwarded to the final questions. This was to prevent the final questions from not being processed because the respondents did not want to work on the other cases.

In the final questionnaire, respondents were again asked to assess their occupational medicine expertise and were also asked which research tools were ultimately used. In addition to a rating of the experience of the research method, positive and negative comments were recorded.

The primary outcome was the case performance as indicated by a number of (right) answers recorded from case processing. The secondary outcome was a change in (right) answers for certain questions that were asked before without any support and again during case processing with the group-assigned method and self-assessment of occupational medicine expertise recorded before and after case processing.

Figure 2. View of the ChatGPT group for solving case 2 before (A) and after (B) input. ChatGPT 4 was connected to the input window..

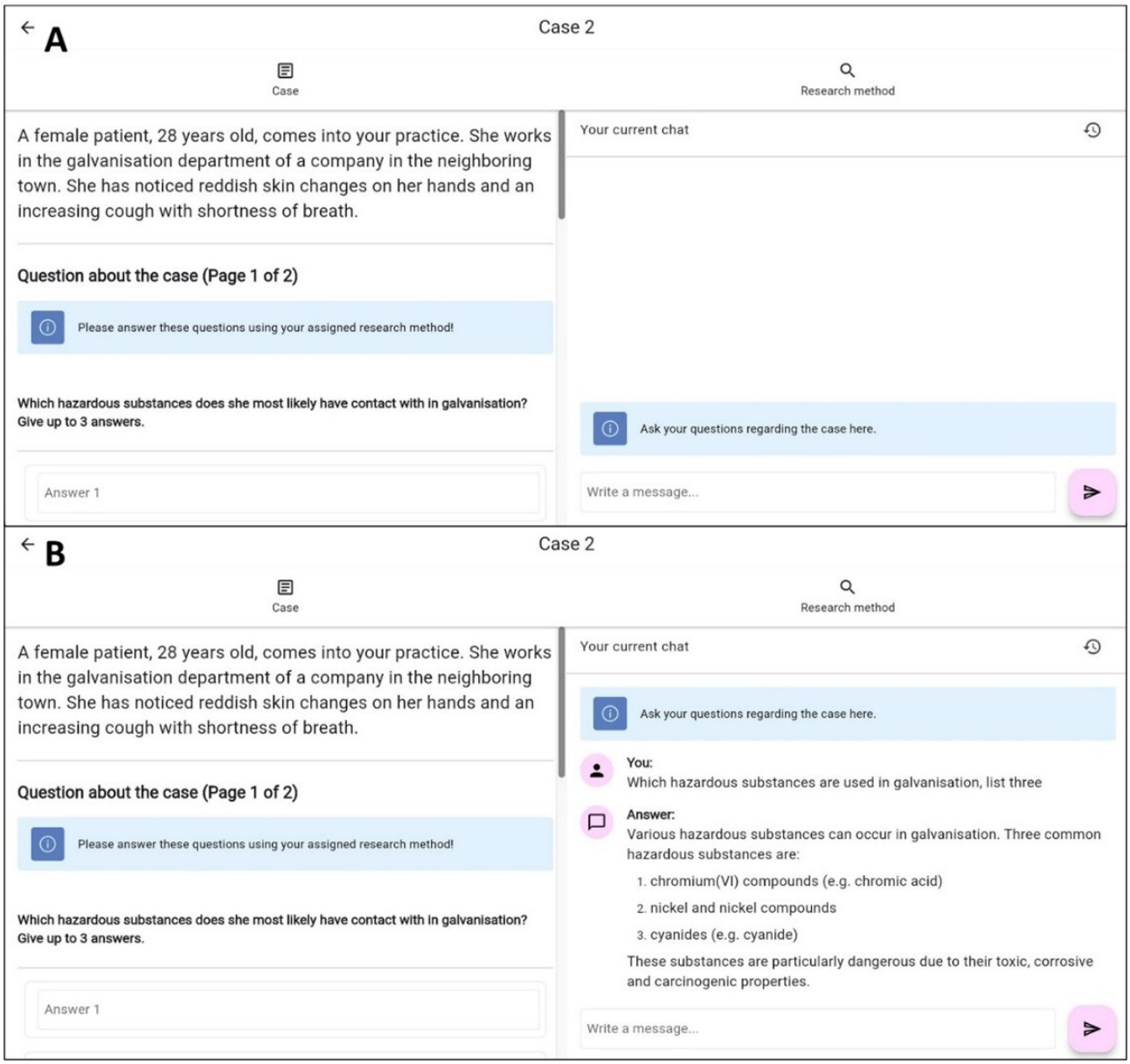

XSL•FO
RenderX

Application

A web-based application with an integrated environment for answering questions, subsequent assignments to a group, and a case view was developed for the study. In addition, an integrated chat window was built into the case view for the ChatGPT cohort. This should look as similar as possible to the ChatGPT user interface. A screenshot can be seen in Figure 2. The development of this application allowed for user-friendly data collection of chat entries and responses during unsupervised study participation. All data was stored in a structured query language database. The LLM was integrated via the Chat Completions application programming interface, a developer interface from OpenAI, the developers of GPT-4 and ChatGPT [12]. Communication with OpenAI took place via our own servers and was carried out in such a way that OpenAI did not receive any information such as the IP addresses of the participants aside from the actual input. The participants' input and, in the case of ongoing chat, the previous input and responses were used as context for the LLM. The model's response was already loaded ("streamed") into the application in sections during generation to enable the response to be displayed earlier. The most up-to-date and powerful model from OpenAI between March 2024 and May 2024 was used: gpt-4‑0125-preview. Each conversation contained a system prompt. This is a command that gives the model context and instructions for the conversation. The system prompt contained, among other things, an explanation that it is an assistant for doctors and students to investigate occupational disease and should be helpful. The entire system prompt read as follows: "You are a helpful assistant who helps with questions regarding occupational diseases in Germany. You communicate with medical students or doctors. You check your information for accuracy. If there are ambiguities, for example, with abbreviations, ask what exactly is meant. Answer specifically and be brief. If there are any uncertainties, explain them."

ChatGPT Input

The inputs made by the participants and the outputs generated by ChatGPT were examined by 2 individuals separately and then compared. Differences in the assessment were discussed between the 2 raters and resolved by consensus. In addition to the number of words entered and output, the type of communication on the participants' side was also recorded, for example, was ChatGPT addressed, was the input in complete sentences or as keywords?

Ethical Considerations

This study was carried out with the approval of the responsible ethics committee (Ethics Committee of the Medical Faculty of RWTH Aachen University, EK 24‑065) in accordance with the Declaration of Helsinki in its current version and with national law. Informed consent was obtained from all participants. Study data were obtained anonymously without any personal information. Participants received no compensation. Three €25 vouchers (approx. 28 US$) were raffled off among all participants. They could voluntarily provide an email address, without this being linked to the other data. This email address was only used to draw and send the vouchers.

Statistical Analysis

The data were analyzed using GraphPad Prism (version 10.2.3) and SPSS (version 29.0.0.0; IBM Corp).

No sample size analysis was performed beforehand. The recruitment period was set at 3 months. The aim was to recruit as many patients as possible during this time with at least 50 participants being included.

The data are given as number and percentage, number of answers or correct answers, or mean (SD). Group differences between the ChatGPT group and the group with its own research were examined either by Mann-Whitney *U* test or chi-square test for group sizes of at least 5 persons, otherwise by Fisher exact test . All statistical tests were 2-sided with $P<.05$ as the significance level.

Results

Demographics of the Participants

A total of 70 respondents made entries in the web-based questionnaire (Figure 3). A total of 10 (10/70, 14%) had already dropped out before being assigned to a group. One person (1/70, 1%) did not state whether they had studied medicine or worked as a doctor and 3 people (3/70, 4%) stated after case processing that they had used ChatGPT or similar LLM but had been assigned to the group with research without ChatGPT. They had to be excluded after participation in the study since participants were not informed of this exclusion criterion beforehand to avoid unconscious influence toward the use of ChatGPT. After exclusion, the data of 56 respondents were evaluated (Table 1).

A total of 27 participants, 17 (17/27, 63%) female and 21 (21/27, 78%) students, were assigned to the ChatGPT group. A total of 29 participants, 24 (24/27, 83%) female, and 15 (15/29, 52%) students, used their own research tools. The ChatGPT group had significantly more students than the research group. Overall, the average age was 28.5 (SD 7.2) years and students were on average (SD) in their 9th (1.8) semester. Both groups rated themselves similarly in terms of occupational medicine expertise. The ChatGPT group gave themselves a 3.9 (SD 1) according to school grades (almost a 4, which corresponds to the grade "sufficient"). The research group gave themselves a 3.8 (SD 0.9).

The number of participants was recorded again before each substep; only 28 of 56 participants (28/56, 50%) answered the last case. The same decrease was observed in both groups, meaning that a "loss-to-follow-up" bias is unlikely here due to the research method.

XSL•FO
RenderX

Figure 3. CONSORT (Consolidated Standards of Reporting Trials) flow diagram. After assessment of 70 participants for eligibility, 59 were randomized in 2 groups. 27 participants were allocated to the ChatGPT group of which all were subsequently included in the analysis. 32 participants were allocated to the group with own research. Here, three participants had to be excluded for analysis as they used ChatGPT as research tool.

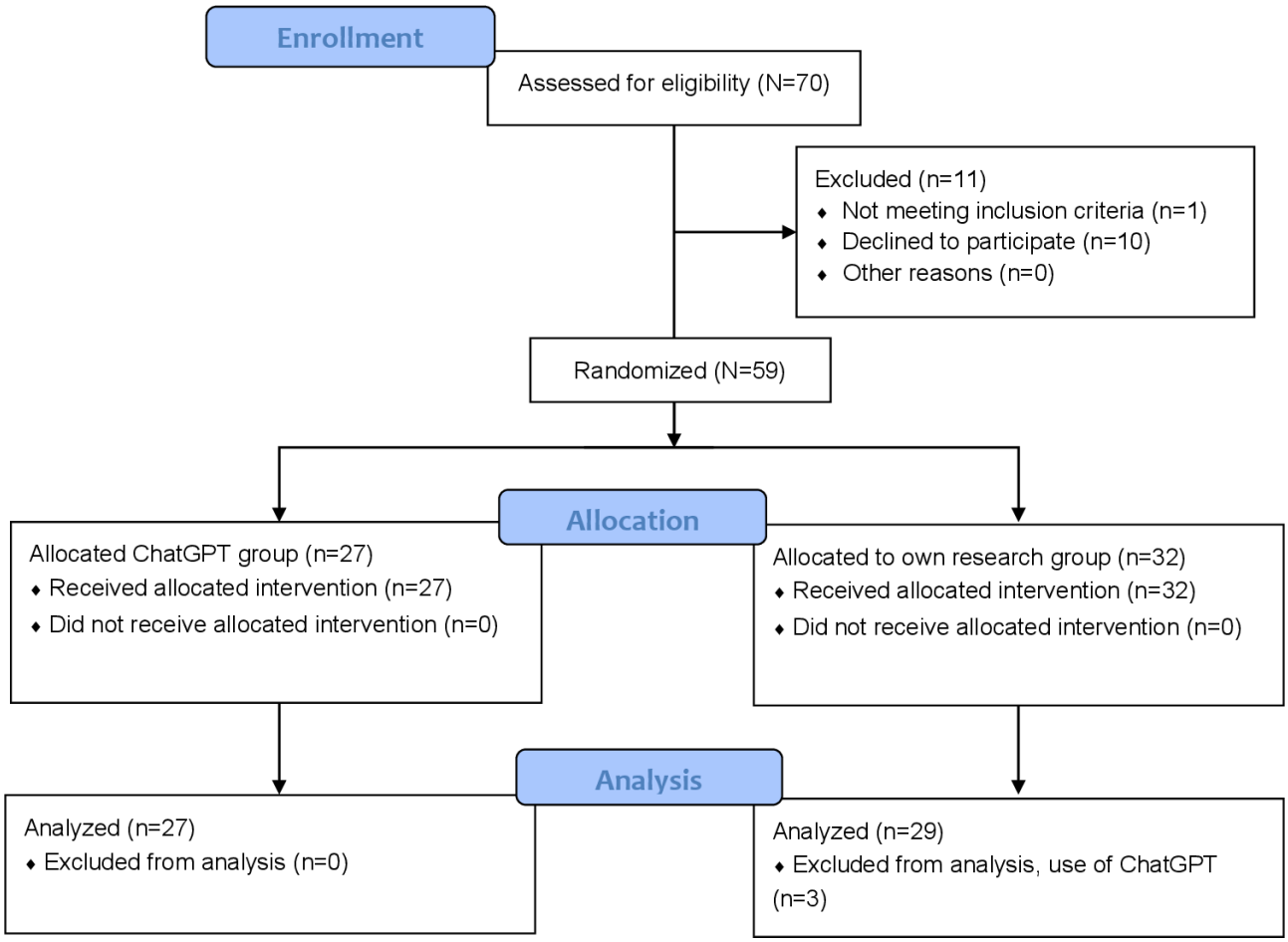

XSL•FO
RenderX

Table . Demographics of the participants. After completing the sociodemographic and knowledge questionnaire, participants were randomly allocated to the ChatGPT group or to conduct their own research. Data are presented as n (%) or mean value (SD).

	Total	ChatGPT	Own research	*P* value
Sex, n (%)	56	27 (48.2)	29 (51.8)	—[a]
Female	41	17 (63.0)	24 (82.8)	.10[b]
Male	15	10 (37.0)	5 (17.2)	—
Status, n (%)				
Students	36	21 (77.8)	15 (51.7)	.046[bc]
Physicians	20	6 (22.2)	14 (48.3)	—
In training	12	3 (50.0)	9 (64.3)	—
Specialists	7	2 (33.3)	5 (35.7)	—
Attendings	1	1 (16.7)	0 (0)	—
Occupational medicine	5	2 (33.3)	3 (21.4)	—
Participation, n (%)				
Case 1	41	22 (81.5)	19 (65.5)	—
Case 2	32	18 (66.7)	14 (48.3)	—
Case 3	28	15 (55.6)	13 (44.8)	—
Concluding questions	29	16 (59.3)	13 (44.8)	—
Age (years), mean (SD)	28.5 (7.2)	27.3 (7.3)	29.6 (7)	.07[d]
Semester (students), mean (SD)	9 (1.8)	8.9 (1.7)	9.2 (1.9)	.51[d]
Self-assessment (school grades), mean (SD)	3.8 (0.9)	3.9 (1)	3.8 (0.9)	.74[d]
Research tools commonly used, n (%)				
Amboss [10]	45	23 (51.1)	22 (48.9)	—
Doccheck [11]	48	24 (50.0)	24 (50.0)	—
Google	44	20 (45.5)	24 (54.5)	—
UptoDate [13]	8	4 (50.0)	4 (50.0)	—
Wikipedia	26	11 (42.3)	15 (57.7)	—
Thieme eRef [14]	12	5 (41.7)	7 (58.3)	—
Via Medici [15]	3	1 (33.3)	2 (66.7)	—

[a]—: not applicable.

[b]Chi-square test.

[c]$P<.05$. For the number students

[d]Mann-Whitney *U* test for unpaired samples.

Case 1: Asbestos-Associated Changes in the Pleura

In terms of the number of correct answers, there was a significantly higher number of correct answers to the question about hazardous substances that can cause pleural changes in the ChatGPT group than after their own research (Tables 2 and 3) which corresponded to a small effect (Cohen $r<0.5$). Only the ChatGPT group was able to significantly increase the number of correct answers to the question about 3 types of cancers caused by asbestos compared to before the group assignment.

With regard to the question of whether a cholangiocellular carcinoma can be recognized as an occupational disease in Germany, there were no group differences, neither between ChatGPT and own research nor between before and after group assignment. Only half as many participants (7/21, 33%)) were able to make the (correct) decision that an occupational disease should be reported in this case with the support of ChatGPT than with their own research (14/21, 67%). Application of ChatGPT was associated with a 6 times higher probability of indicating to report the occupational disease than with own research.

XSL•FO
RenderX

Table . Case 1: Gardener with asbestos-associated pleural lesions—free-text questions. Participants were asked 3 questions, listing 3 answers in free text (number of responses). Free-text questions included the instruction to list hazardous substances that can cause pleural changes, materials that can induce them, and types of cancers caused by asbestos exposure. Either ChatGPT or the research method of their own choosing was used, depending on group allocation for research.

Case 1 (n=41)	Total (mean, SD)	ChatGPT (mean, SD)	Own research (mean, SD)	*P* value[a]	Cohen *r*
Hazardous substances with pleural changes, mean (SD)	2.2 (0.8)	2.5 (0.8)	1.8 (0.8)	.01[b]	−0.38
Materials pleural changes, mean (SD)	1.5 (1.2)	1.7 (1.3)	1.3 (1.1)	.35	−0.28
Types of cancers asbestos, mean (SD)	2.8 (0.5)	2.9 (0.5)	2.7 (0.6)	.37	−0.25
Types of cancers before, mean (SD)	2 (0.8)	1.8 (0.8)	2.2 (0.8)	.18	−0.22
Types of cancers (comparison before vs after, *P* value)[c]	<.001[d]	<.001[d]	.06	—[e]	—

[a]Mann-Whitney *U* test for unpaired samples.

[b]*P*<.05.

[c]Wilcoxon test for paired samples.

[d]*P*<.001.

[e]—: not applicable.

Table . Case 1: Gardener with asbestos-associated pleural lesions—multiple-choice questions. Participants were asked 3 questions choosing the right answers from multiple choice. Multiple-choice questions asked whether a CCC[a] could be officially recognized as an OD[b] in Germany, which OD number corresponds to asbestos-induced pleural changes, and whether an official OD report should be filed. Either ChatGPT or the research method of their own choosing was used, depending on group allocation for research.

Case 1 (n=41)	Total	ChatGPT	Own research	*P* value[c]	OR[d] (95% CI)
CCC as OD	30	16	14	.95	1.05 (0.26 - 4.20)
CCC as OD before	19	10	9	.90	1.08 (0.32 - 3.70)
CCC as OD (comparison before versus after as difference)	11	6	5	—[e]	—
CCC as OD “Don’t know”	3	1	2	—	2.47 (.21 - 29.63)
CCC as OD “Don’t know” before	18	9	9	.68	1.30 (0.38 - 4.49)
CCC as OD “Don’t know” (comparison before versus after as difference)	–15	–8	–7	—	—
Multiple-choice OD (pleural changes asbestos)	36	20	16	.51	0.53 (0.08 - 3.59)
OD report	21	7	14	.007[f]	6.00 (1.54 - 23.36)

[a]CCC: cholangiocellular carcinoma.

[b]OD: occupational disease.

[c]Chi-square test or Fisher exact test if fewer than 5 per group.

[d]OR: odds ratio.

[e]—: not applicable.

[f]*P*<.01.

XSL•FO
RenderX

Case 2: Galvanization-Associated Occupational Allergy

There were no relevant differences between groups for the correct listing of hazardous substances in galvanization, only a general learning effect in contrast to the time before group assignment (Tables 4 and 5). Participants in the ChatGPT group named the correct next diagnostic steps significantly more often (medium effect; Cohen $r<0.8$). They were able to list more correct fields of activity in which occupational asthma can develop (medium effect; Cohen $r<0.8$), although they were able to state fewer from memory than the other group in the query before group assignment (medium effect; Cohen $r<0.8$).

There were no group differences in the questions about the suspected diagnosis, the correct occupational disease according to the German ordinance on occupational diseases, and whether a report should be made, although the proportion of correct answers was rather low in both groups.

Table . Case 2: Allergy in galvanization—free-text questions. Participants were asked 3 questions, listing 3 answers in free text (number of responses) or choosing the right answers from multiple choice. Free-text questions included the instruction to list hazardous substances workers could be exposed to in galvanization, the next steps in the diagnostic procedure for the patient in case 2, and in which occupational fields occupational asthma could occur. Either ChatGPT or the research method of their own choosing was used, depending on group allocation for research.

Case 2 (n=32)	Total (mean, SD)	ChatGPT (mean, SD)	Own research (mean, SD)	*P* value[a]	Cohen *r*
Hazardous substances galvanization	2.9 (0.4)	3 (0)	2.8 (0.6)	.18	–0.41
Hazardous substances galvanization before	0.7 (1.1)	0.5 (1)	1 (1.2)	.18	—[b]
Hazardous substances galvanization (comparison before versus after; p[§])	<.001[c]	<.001[c]	<.001[c]	—	—
Diagnostic procedure	2.2 (0.6)	2.5 (0.5)	1.9 (0.6)	.02[d]	–0.62
Occupational fields allergic asthma	2.9 (0.4)	3 (0)	2.7 (0.5)	.03[d]	–0.56
Occupational fields allergic asthma before	2.1 (0.9)	1.8 (1.1)	2.5 (0.5)	.049[d]	–0.54
Occupational fields allergic asthma (comparison before versus after; *P* value)[e]	<.001[c]	.002	.25	—	—

[a]Mann-Whitney *U* test for unpaired samples.

[b]—: not applicable.

[c]$P<.001$.

[d]$P<.05$.

[e]Wilcoxon test for paired samples.

Table . Case 2: Allergy in galvanization—multiple-choice questions. Participants were asked 3 questions choosing the right answers from multiple choice. Multiple-choice questions asked for the suspected diagnosis, the corresponding occupational disease, and whether an official occupational disease report should be filed. Either ChatGPT or the research method of their own choosing was used, depending on group allocation for research.

Case 2 (n=32)	Total, n	ChatGPT, n	Own research, n	*P* value[a]	OR[b] (95% CI)
Suspected diagnosis	15	9	6	.69	0.75 (0.18 - 3.06)
Occupational disease	14	8	6	.93	0.94 (0.23 - 3.84)
Occupational disease report	28	15	13	.42	2.6 (0.24 - 28.15)

[a]Chi-square test for the number of correct answers..

[b]OR: odds ratio.

Case 3: Dental Technician With Berylliosis

In both groups, approximately the same number of hazardous substances could be listed to which the patient was exposed as a dental technician (Tables 6 and 7). There were also no group differences between the ChatGPT and own research or before and after group assignment with regard to occupational fields for potentially causing berylliosis. Both groups learned more by using their research tools.

The participants in both groups also answered the question of whether sarcoidosis can be recognized as an occupational

XSL•FO
RenderX

disease equally often correctly, even when comparing before and after group assignment. Participants in the group with their own research were able to correctly identify the case as occupational disease no. 1110 according to the German ordinance on occupational diseases ("diseases caused by beryllium or its compounds") more often than users of ChatGPT. Given the asymmetrical distribution of the groups among the responses, no realistic estimate of the odds ratio could be obtained. The lymphocyte transformation test and the need to report this occupational disease were reported by the same number of participants in both groups, albeit few.

Table . Case 3: Dental technician with berylliosis—free-text questions. Participants were asked 2 questions listing 3 answers in free text (number of responses. Free-text questions included the instruction to list hazardous substances dental technicians could be exposed to and occupational fields where occupational asthma could occur. Either ChatGPT or the research method of their own choosing was used, depending on group allocation for research.

Case 3 (n=28)	Total (mean, SD)	ChatGPT (mean, SD)	Own research (mean, SD)	*P* value[a]	Cohen *r*
Hazardous substances: Dental technician	2.8 (0.5)	2.9 (0.5)	2.8 (0.4)	.31	–0.51
Occupational fields berylliosis	1.6 (0.7)	1.4 (0.8)	1.8 (0.6)	.21	–0.55
Occupational fields berylliosis before	0.3 (0.9)	0.2 (0.8)	0.5 (1)	.31	–0.52
Occupational fields berylliosis (comparison before versus after; *P* value)[b]	<.001[c]	<.001[c]	<.001[c]	—[d]	—

[a]Mann-Whitney *U* test for unpaired samples.

[b]Wilcoxon test for paired samples.

[c]*P*<.001.

[d]—: not applicable.

Table . Case 3: Dental technician with berylliosis—multiple-choice questions. Participants were asked 4 questions choosing the right answers from multiple choice. Multiple-choice questions asked whether sarcoidosis could be officially recognized as an OD[a] in Germany, which OD number berylliosis corresponds to in Germany, which diagnostic tests could be used, and whether an official occupational disease report should be filed. Either ChatGPT or the research method of their own choosing was used, depending on group allocation for research

Case 3 (n=28)	Total, n	ChatGPT, n	Own research, n	*P* value[b]	OR[c] (95% CI)
Sarcoidosis as OD	23	13	10	.50	0.51 (0.07 - 3.68)
Sarcoidosis as OD before	11	4	7	.14	3.2 (0.66 - 15.59)
Sarcoidosis as OD (comparison before versus after as difference)	12	9	3	—[d]	—
Sarcoidosis as OD "Don't know"	0	0	0	—	—
Sarcoidosis as OD "Don't know" before	5	4	1	.19	0.23 (0.02 - 2.38)
Sarcoidosis as OD "Don't know" (comparison before versus after as difference)	–5	–4	–1	—	—
OD berylliosis	8	1	7	.006[e]	16.33 (1.63 - 163.44)
Diagnostic test	13	5	8	.14	3.20 (0.68 - 15.07)
OD report	20	9	11	.15	3.21 (0.66 - 15.59)

[a]OD: occupational disease.

[b]Chi-square test or Fisher's exact test if fewer than 5 per group.

[c]OR: odds ratio.

[d]—: not applicable.

[e]*P*<.01.

XSL•FO
RenderX

Concluding Questions

Even though there was a tendency toward higher satisfaction in the ChatGPT group, there were no significant differences between the 2 groups (Table 8). The ChatGPT group showed a significantly greater increase in self-assessment than the group with their own research.

The participants in the group with their own research primarily used DocCheck and Google, with a few also using Amboss and Wikipedia. Participants who indicated ChatGPT here were completely excluded from the analysis.

Table . Concluding questions—survey on satisfaction and self-assessment before and after the cases. Participants were asked about their satisfaction with the research method, as well as their self-assessment after case processing. For participants who provided a self-assessment in the final questionnaire, their corresponding self-assessment before case processing was compared. Either ChatGPT or the research method of their own choosing was used, depending on group allocation for research. Participants in the own research group were asked which tools they used.

Concluding questions (n=29)	Total	ChatGPT	Own research	*P* value[a]	Cohen *r*
Satisfaction with research method, mean (SD)	3.1 (1.2)	2.8 (1.3)	3.4 (0.9)	.14	−0.57
Self-assessment OME[b] before, mean (SD)	3.7 (.8)	3.9 (.9)	3.5 (0.8)	.12	−0.56
Self-assessment OME after, mean (SD)	3.4 (1)	3.4 (1.1)	3.4 (0.9)	.92	−0.44
Self-assessment OME (comparison before vs after), *P* value[c]	.10	.047[e]	.99	—[d]	—
Used research tools, n (%)					
Amboss	5	0 (0)	5 (100)	—	—
DocCheck	10	0 (0)	10 (100)	—	—
Google	13	0 (0)	13 (100)	—	—
Wikipedia	2	0 (0)	2 (100)	—	—

[a]Mann-Whitney *U* test for unpaired samples.

[b]OME: occupational medicine expertise.

[c]Wilcoxon test for paired samples.

[d]—: not applicable.

[e]$P<.05$

Evaluations of the Participants

After using the tool, participants were able to voluntarily leave positive and negative feedback. In the

ChatGPT group, the simple and practical use, as well as the possibility of follow-up questions, was noted positively several times. It was also praised that the answer was displayed in detail with reasons so that something additional was learned and there was the opportunity to ask very specific questions. On the negative side, the time it took to provide a complete answer was mentioned several times, as was the lack of references.

There were no positive comments in the group on their own research; negative comments were that the internet search was laborious and time-consuming.

In general, many participants requested a solution for each question with the right answer. This was not given so that there could be no “contamination” of the study population, for example, within a semester.

Analysis of ChatGPT Input

In the ChatGPT group, there were 2 participants who only entered keywords, similar to a Google search. The rest corresponded with ChatGPT like a person and gave explicit instructions in full sentences.

The average number of entered instructions to ChatGPT increased from 5.8 (2.7 SD) messages in case 1 to 6.3 (SD 2.3) to 10.5 (SD 4.2) messages.

Discussion

This study investigated how a generative LLM such as ChatGPT can support medical research and clinical decision-making and how it performs in comparison to conventional research. The subject of occupational lung diseases was selected in accordance with the stipulations of German occupational disease legislation. Approximately 80% of deaths from occupational diseases in Germany are caused by lung diseases [16]. In the processing of the cases, it was demonstrated that ChatGPT facilitated the participants‘ ability to conduct targeted research, such as identifying potential hazardous substances or activities, and

XSL•FO
RenderX

enhanced their self-assessment of their specialist knowledge. However, clinical decisions, such as determining whether an occupational disease report should be filed, were more frequently made correctly through the participants' independent research.

This project was initiated in response to the recognition that numerous medical professionals encounter difficulties in navigating the complexities of occupational disease, including uncertainty about its existence and the appropriate reporting procedures. During the course of everyday clinical practice, there was a clear indication of a need for a dedicated digital resource to address these challenges. In Germany, there is a web-based search tool provided by the German Social Accident Insurance [17]. One can enter the diagnosed disease according to the ICD-10 code and receive possible occupational diseases. However, it is solely based on the presenting disease and has no possibility to put in certain chemicals or exposures. From experience in our university hospital, the majority of medical students and physicians in other fields than occupational medicine do not know it. With the introduction of ChatGPT 4, an enhanced LLM with a more substantial data foundation and augmented performance, coupled with its internet connectivity via Bing, which should markedly curtail the proclivity for confabulation, has made a corresponding function via ChatGPT a realistic prospect for the first time [18,19]. The exclusion of 3 participants due to the use of ChatGPT illustrates the growing prevalence of ChatGPT as a search engine and database. This phenomenon appears to be particularly pronounced among younger individuals [20]. The 3 individuals excluded from our study were all students under the age of 30.

Similar to the results of this study, other studies that used ChatGPT to answer medical questions showed that ChatGPT can answer many correctly. In 1 study, ChatGPT 4 was even able to answer more ophthalmology questions correctly than ChatGPT 3.5 and a human comparison group [21]. ChatGPT answered questions about diagnoses and differential diagnoses for medical case vignettes with an acceptable but not yet good level of accuracy [22].

In regard to clinical decision-making, ChatGPT has been demonstrated to be less effective than medical experts in the present context. In this study, the group that used their customary research techniques was significantly more successful in making accurate clinical decisions than the ChatGPT group. It seems probable that the discrepancy can be attributed to implicit considerations on the part of the participants, given that ChatGPT provides a definitive "yes" or "no" response to the posed question, whereas the participants in their own research are required to arrive at a decision independently. Probably, considerations such as the fact that imaging with ionizing radiation should be avoided in a young woman in case 2 were not given sufficient weight by ChatGPT. A comparable outcome was observed in the study conducted by Zaboli et al [23], which compared the triage decisions made by specialized triage nurses in the emergency department with those generated by ChatGPT. Here, the triage nurses demonstrated a significantly superior performance compared to ChatGPT. In a separate study, ChatGPT was tasked with answering questions from specialist orthopedic examinations in the United States. The applicable knowledge demonstrated by ChatGPT 4 was found to be comparable to that of an individual in their third year of orthopedic training (residency) [24]. Furthermore, ChatGPT has demonstrated the capacity to respond to queries and offer diagnoses in urological matters at the level of medical practitioners in training, while human expertise is more proficient in more complex scenarios [25]. Interestingly, ChatGPT 4 also exhibits inferior performance compared to human experts in radiological disciplines, despite the frequent assertion that it is particularly well-suited for use in this domain. For instance, the proficiency of radiology professionals in advanced training was not reached in severe neuroradiological cases [26]. In the United States Medical Licensing Examination, ChatGPT performed comparably to a third-year medical student [27,28].

In terms of standardized tasks, however, ChatGPT appears to be highly functional. In the evaluation of standard ECGs, ChatGPT outperformed both emergency physicians and cardiologists. For more challenging questions, it demonstrated a level of proficiency comparable to that of cardiologists and surpassed the performance of emergency physicians [29]. Similarly, the generation of information letters for patients yielded comparable outcomes. In this context, the letters created by ChatGPT were rated more highly by patients and physicians than those designed by surgeons [30]. Furthermore, ChatGPT 4 is capable of providing satisfactory responses to radiology-related patient queries [31,32].

The quality of the LLM's output is contingent upon the quality of the training data. It seems reasonable that in the context of occupational medicine, and particularly in the specific case of German occupational disease law, the training data available for the purpose of training the models was likely limited. Specific training with information relevant to occupational medicine could enhance the application but was not available at the time. Information on German occupational diseases are typically accessible on the web, though predominantly in German. Specially developed LLM for the medical context, such as Med-Palm from Google [33,34] represent potential improvements but were not considered in this study due to a lack of accessibility.

In this study, ChatGPT demonstrated significant difficulties in finding a solution to case 3, a dental technician with berylliosis. This case was selected for this study due to its rarity in Germany, where it is not frequently recognized or reported [9]. During case processing, confabulations and false statements were recorded in the output. In a multiple-choice question in which the potential occupational disease was to be selected from five different ones with the initial digit 11 (prefix for occupational diseases caused by metals and metalloids), ChatGPT regularly stated that none of the offered options were correct. Furthermore, it proposed occupational diseases numbered 4103 (asbestos dust lung disease, also known as asbestosis) or 4104 (lung cancer, laryngeal cancer, or ovarian cancer caused by asbestos dust). Furthermore, ChatGPT 4 even invented sources that did not exist.

As with any innovation, its use in a medical context must be subjected to rigorous scrutiny and it is to be expected that errors

XSL•FO
RenderX

in the LLM will occur. The responses to the use of ChatGPT were noteworthy. The participants exhibited an enhancement in their self-assessment of their own specialist knowledge. Nevertheless, there was a tendency for participants in the group who conducted their own research to report a higher level of self-assessment. In conclusion, both groups awarded themselves an identical rating following the case processing, with an average grade of 3.4 (satisfactory). Additionally, there was a tendency toward greater satisfaction with the use of ChatGPT. This is evidenced by the positive ratings given to the ease and practicality of its use, as well as the possibility of posing follow-up queries. Conversely, the negative ratings assigned to the working time of ChatGPT were attributed to the system requirements at the time. However, with ChatGPT 4o, a significantly faster system with comparable quality is now available.

This study evaluated the performance in only 3 cases with as little as 6 questions per case. It was designed as a pilot study to assess whether ChatGPT can be used for occupational medicine cases. As a further limitation, only occupational disease cases related to the lungs were included. Whether the usability could be extended for the entire field of occupational diseases cannot be assessed in this study and should be addressed in future studies.

In addition to the number of cases, the lack of monitoring of the group with its own research represents a further limiting factor in this study. Neither the entries nor the sources were subjected to any form of verification. For example, it is possible that a greater number of individuals may have used ChatGPT without disclosing this information. This study was designed with the explicit intention of testing against the conventional research methodology. Consequently, the typical applications and devices were also employed to ensure that the approach was as realistic as possible. With regard to the relatively small number of participants, recruitment proved challenging, and only approximately half of the participants engaged with the study until the conclusion of the final case. The lengthy processing time was identified as the primary reason for withdrawal. Unfortunately, processing time was not recorded. Thus, it cannot be assessed in detail if and how the processing time affected dropout. In some cases, the participants expressed a desire to conduct further research until all questions had been answered correctly. This could also have resulted in a positive distortion of performance. The number of correct answers could have been lower had the participants been prompted to provide answers after their initial brief research. As both groups were equally decimated over the cases, a “lost-to-follow-up bias” seems unlikely due to the research method. The two groups were formed by randomization using a digital weighted coin. In the case of an unequal proportion of students and doctors, the coin was weighted in such a way that it was then more likely to be assigned to the other group, but never simply assigned. Nevertheless, there was a certain unequal distribution of doctors across the groups, so they were more represented in the group with their own research. Physicians can be assumed to have greater expertise and experience, but the proportion of colleagues working in occupational medicine was evenly distributed. Nevertheless, a certain bias may have taken place here, so that regardless of the research method, the performance of the group with more physicians was greater, for example, with regard to clinical decisions. In future surveys, this should be taken into account even more and a higher number of participants should generally be included. Nevertheless, this study is one of the few that examines the realistic use of ChatGPT with the input of end users and does not only assess data or questions as input in ChatGPT. The use of ChatGPT by precisely these users, namely medical students and doctors, is conceivable and is already happening to some extent today in the clinical routine. Although no significant differences between the 2 groups could be observed, female participants made up the majority of participants in the study overall. There is a sex disparity in the use of generative LLM tools usually favoring men. A survey conducted by the Bank for International Settlements stated that 50% of men reported using generative artificial intelligence over the previous 12 months, compared to only 37% of women [35]. However, women engage with chatbots in a more relational and exploratory manner than men, for example, asking follow-up questions and seeking clarification [36]. A recent large meta-analysis even reported a female lead in digital knowledge and skills among students [37]. However, the manner in which e-learning and LLM tools are used is shaped not only by sex but also by age, background, and prior experience [38-40]. The effects of those factors could not be investigated in this setup but should be addressed and examined in future studies.

In conclusion, it can be stated that ChatGPT 4 is a valuable tool for targeted medical research, even for highly specific questions in occupational medicine concerning occupational diseases in Germany. However, it is imperative that clinical decisions are not based on the output of the LLM, but also assessed by qualified medical professionals. The potential of the LLM to perform well after training with relevant data or overviews regarding occupational diseases and specialized occupational medicine instructions (prompts) remains uncertain and requires further investigation.

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability

The data sets generated during and/or analyzed during this study are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Authors' Contributions

FAW and JK were primarily responsible for the conceptualization of the study, the design of the methodology, and the formal analysis of the data. They also curated the dataset used for the study. The investigation was carried out by FAW, NO, HK, and JK, with FAW and NO also contributing significantly to the visualization of results. Resources for the study were provided by TK. Supervision of the project was undertaken by TK and JK. FAW wrote the original version of the manuscript, all authors contributed to reviewing and editing the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Socio-demographic questionnaire.
[DOCX File, 29 KB - formative_v9i1e63857_app1.docx]

Checklist 1
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File, 12703 KB - formative_v9i1e63857_app2.pdf]

References

1. Shortliffe EH. Computer-based medical consultations: MYCIN. J Clin Eng 1976;388. [doi: 10.1097/00004669-197610000-00011]
2. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare 2023;11(6):887. [doi: 10.3390/healthcare11060887] [Medline: 36981544]
3. Smith AL, Greaves F, Panch T. Hallucination or confabulation? Neuroanatomy as metaphor in large language models. PLOS Digit Health 2023;2(11):e0000388. [doi: 10.1371/journal.pdig.0000388] [Medline: 37910473]
4. Kooraki S, Hosseiny M, Jalili MH, et al. Evaluation of ChatGPT-generated educational patient pamphlets for common interventional radiology procedures. Acad Radiol 2024;31(11):4548-4553. [doi: 10.1016/j.acra.2024.05.024] [Medline: 38839458]
5. Yüce A, Yerli M, Misir A. Can Chat-GPT assist orthopedic surgeons in evaluating the quality of rotator cuff surgery patient information videos? J Shoulder Elbow Surg 2025;34(1):141-146. [doi: 10.1016/j.jse.2024.04.021] [Medline: 38852711]
6. Cesur T, Güneş YC. Optimizing diagnostic performance of ChatGPT: the impact of prompt engineering on thoracic radiology cases. Cureus 2024;16(5):e60009. [doi: 10.7759/cureus.60009] [Medline: 38854352]
7. Stoneham S, Livesey A, Cooper H, Mitchell C. ChatGPT versus clinician: Challenging the diagnostic capabilities of artificial intelligence in dermatology. Clin Exp Dermatol 2024;49(7):707-710. [doi: 10.1093/ced/llad402] [Medline: 37979201]
8. Occupational diseases. Federal Institute for Occupational Safety and Health. URL: https://www.baua.de/EN/Topics/Prevention/Physical-health/Occupational-diseases/Occupational-diseases_node [accessed 2025-05-02]
9. DGUV statistics 2022 figures and long-term trends. : German Social Accident Insurance; 2023 URL: https://publikationen.dguv.de/detail?sArticle=4768 [accessed 2025-05-02]
10. AMBOSS. URL: https://www.amboss.com/int [accessed 2025-04-06]
11. DocCheck. URL: https://www.doccheck.com/welcome [accessed 2025-04-06]
12. Chat completions API. OpenAI Platform. URL: https://platform.openai.com/docs/guides/text-generation/chat-completions-api [accessed 2025-04-06]
13. UpToDate: trusted, evidence-based solutions for modern healthcare. Wolters Kluwer. URL: https://www.wolterskluwer.com/en/solutions/uptodate [accessed 2025-04-06]
14. eRef. URL: https://eref.thieme.de/favorites/organization?lang=en [accessed 2025-04-06]
15. Via Medici. URL: https://viamedici.thieme.de [accessed 2025-04-06]
16. Safety and health at work—reporting year 2022: report on accident prevention at work. : Federal Institute for Occupational Safety and Health; 2023 URL: https://www.baua.de/EN/Service/Publications/Report/Suga-2022 [accessed 2025-05-02]
17. German Social Accident Insurance. URL: https://www.dguv.de/bk-info/index.jsp [accessed 2025-04-06]
18. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese Medical Licensing Examination: comparison study. JMIR Med Educ 2023 Jun 29;9:e48002. [doi: 10.2196/48002] [Medline: 37384388]
19. He N, Yan Y, Wu Z, et al. Chat GPT-4 significantly surpasses GPT-3.5 in drug information queries. J Telemed Telecare 2025 Feb;31(2):306-308. [doi: 10.1177/1357633X231181922] [Medline: 37350055]
20. statista. Global user demographics of chatgpt in 2023, by age and gender. 2023. URL: https://www.statista.com/statistics/1384324/chat-gpt-demographic-usage [accessed 2025-05-02]
21. Moshirfar M, Altaf AW, Stoakes IM, Tuttle JJ, Hoopes PC. Artificial intelligence in ophthalmology: a comparative analysis of GPT-3.5, GPT-4, and human expertise in answering StatPearls questions. Cureus 2023 Jun;15(6):e40822. [doi: 10.7759/cureus.40822] [Medline: 37485215]

XSL•FO
RenderX

22. Rao A, Pang M, Kim J, et al. Assessing the utility of ChatGPT throughout the entire clinical workflow: development and usability study. J Med Internet Res 2023 Aug 22;25:e48659. [doi: 10.2196/48659] [Medline: 37606976]
23. Zaboli A, Brigo F, Sibilio S, Mian M, Turcato G. Human intelligence versus Chat-GPT: who performs better in correctly classifying patients in triage? Am J Emerg Med 2024 May;79(44-47):44-47. [doi: 10.1016/j.ajem.2024.02.008] [Medline: 38341993]
24. Hofmann HL, Guerra GA, Le JL, et al. The rapid development of artificial intelligence: GPT-4's performance on orthopedic surgery board questions. Orthopedics 2024;47(2):e85-e89. [doi: 10.3928/01477447-20230922-05] [Medline: 37757748]
25. Xv Y, Peng C, Wei Z, Liao F, Xiao M. Can Chat-GPT A substitute for urological resident physician in diagnosing diseases?: A preliminary conclusion from an exploratory investigation. World J Urol 2023 Sep;41(9):2569-2571. [doi: 10.1007/s00345-023-04539-0] [Medline: 37505265]
26. Horiuchi D, Tatekawa H, Oura T, et al. Comparing the diagnostic performance of GPT-4-based ChatGPT, GPT-4V-based ChatGPT, and radiologists in challenging neuroradiology cases. Clin Neuroradiol 2024 Dec;34(4):779-787. [doi: 10.1007/s00062-024-01426-y] [Medline: 38806794]
27. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The Implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 8;9:e45312. [doi: 10.2196/45312] [Medline: 36753318]
28. Cherif H, Moussa C, Missaoui AM, Salouage I, Mokaddem S, Dhahri B. Appraisal of ChatGPT's aptitude for medical education: comparative analysis with third-year medical students in a pulmonology examination. JMIR Med Educ 2024 Jul 23;10:e52818. [doi: 10.2196/52818] [Medline: 39042876]
29. Günay S, Öztürk A, Özerol H, Yiğit Y, Erenler AK. Comparison of emergency medicine specialist, cardiologist, and chat-GPT in electrocardiography assessment. Am J Emerg Med 2024 Jun;80(51-60):51-60. [doi: 10.1016/j.ajem.2024.03.017] [Medline: 38507847]
30. Lockie E, Choi J. Evaluation of a chat GPT generated patient information leaflet about laparoscopic cholecystectomy. ANZ J Surg 2024 Mar;94(3):353-355. [doi: 10.1111/ans.18834] [Medline: 38131402]
31. Gordon EB, Towbin AJ, Wingrove P, et al. Enhancing patient communication with Chat-GPT in radiology: evaluating the efficacy and readability of answers to common imaging-related questions. J Am Coll Radiol 2024 Feb;21(2):353-359. [doi: 10.1016/j.jacr.2023.09.011] [Medline: 37863153]
32. Scheschenja M, Viniol S, Bastian MB, Wessendorf J, König AM, Mahnken AH. Feasibility of GPT-3 and GPT-4 for in-depth patient education prior to interventional radiological procedures: a comparative analysis. Cardiovasc Intervent Radiol 2024 Feb;47(2):245-250. [doi: 10.1007/s00270-023-03563-2] [Medline: 37872295]
33. Med-Palm: A large language model from google research, designed for the medical domain. Google Research. URL: https://sites.research.google/med-palm [accessed 2025-04-06]
34. Cao X, Xu X, Rekik I, editors. Machine Learning in Medical Imaging: Springer Nature Switzerland; 2024.
35. Aldasoro I, Armantier O, Doerr S, Gambacorta L, Oliviero T. The gen AI gender gap. Econ Lett 2024 Aug;241:111814. [doi: 10.1016/j.econlet.2024.111814]
36. Bastiansen MHA, Kroon AC, Araujo T. Female chatbots are helpful, male chatbots are competent? Publizistik 2022 Nov;67(4):601-623. [doi: 10.1007/s11616-022-00762-8]
37. Campos DG, Scherer R. Digital gender gaps in students' knowledge, attitudes and skills: an integrative data analysis across 32 countries. Educ Inf Technol 2024 Jan;29(1):655-693. [doi: 10.1007/s10639-023-12272-9]
38. Gonzales AL. Disadvantaged minorities' use of the internet to expand their social networks. Communic Res 2017 Jun;44(4):467-486. [doi: 10.1177/0093650214565925]
39. Tarhini A, Hone K, Liu X, Tarhini T. Examining the moderating effect of individual-level cultural values on users' acceptance of E-learning in developing countries: a structural equation modeling of an extended technology acceptance model. Interactive Learning Environments 2017 Apr 3;25(3):306-328. [doi: 10.1080/10494820.2015.1122635]
40. Bünning M, Schlomann A, Memmer N, Tesch-Römer C, Wahl HW. Digital gender gap in the second half of life Is declining: changes in gendered internet use between 2014 and 2021 in Germany. J Gerontol B Psychol Sci Soc Sci 2023 Aug 2;78(8):1386-1395. [doi: 10.1093/geronb/gbad079] [Medline: 37218293]

Abbreviations

LLM: large language model

XSL•FO
RenderX

Edited by A Mavragani; submitted 07.08.24; peer-reviewed by MV Garrido, YS Kiyak; revised version received 19.01.25; accepted 20.01.25; published 20.05.25.

Please cite as:
Weuthen FA, Otte N, Krabbe H, Kraus T, Krabbe J
Comparison of ChatGPT and Internet Research for Clinical Research and Decision-Making in Occupational Medicine: Randomized Controlled Trial
JMIR Form Res 2025;9:e63857
URL: https://formative.jmir.org/2025/1/e63857
doi:10.2196/63857

© Felix A Weuthen, Nelly Otte, Hanif Krabbe, Thomas Kraus, Julia Krabbe. Originally published in JMIR Formative Research (https://formative.jmir.org), 20.5.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Parents', Teachers', and Sledders' Acceptability of a Virtual Reality Game for Sledding Safety Education: Cross-Sectional Study

Meggy Hayotte[1], PhD; Jonas Kreiner[2], BSc; Bernhard Hollaus[2], PhD

[1]Université Côte d'Azur, LAMHESS, France, Campus STAPS, 261 boulevard du Mercantour, Nice, France

[2]Management Center Innsbruck (MCI), Department of Medical, Health & Sports Engineering, Innsbruck, Austria

Corresponding Author:
Meggy Hayotte, PhD
Université Côte d'Azur, LAMHESS, France, Campus STAPS, 261 boulevard du Mercantour, Nice, France

Abstract

Background: Sledding is a common recreational activity in the Alpine region, practiced by families, friends, and at school, but it is rarely considered to cause serious injuries. Current actions to promote safe sledding are limited to announcements at the start of dedicated tracks or to didactic sheets presented by teachers in schools. However, these actions are currently limited and do not allow the development of piloting technique skills. Virtual reality has the potential to develop piloting skills, although the development of a virtual reality for sledding safety education needs to be guided by its acceptability.

Objective: The purpose of this study was to (1) examine the acceptability of the virtual reality game for sledding safety education from the perspective of different potential user profiles (ie, parents, teachers, and sledders) based on the Unified Theory of Acceptance and Use of Technology 2 enriched by the health locus of control (HLC), and (2) understand participants' preferences and needs in terms of features to be integrated to the technology.

Methods: Three profiles of participants (ie, parents, teachers, or sledders) were recruited through email diffusion from elementary schools, university, and ski resort announcements (Tyrol, Austria). They completed a series of questionnaires measuring: (1) demographic and general information, (2) acceptability of virtual reality sledding game for safety education (named VRodel), (3) perception of HLC, and (4) preferences and needs for VRodel.

Results: A total of 122 participants (15 teachers, 43 parents or legal guardians, and 64 sledders) completed the survey. Findings reveal that the constructs of hedonic motivation (β=.43, P<.001), price sensitivity (β=.28, P<.001), and habit (β=.36, P<.001) explained 65% of the variance in behavioral intention to use VRodel. Two acceptability clusters were identified (low and high), but no differences between the acceptability clusters emerged based on age, status, gender, or previous virtual reality use. Internal HLC was positively correlated with all dimensions of acceptability except social influence and facilitating conditions. Some correlations between acceptability constructs were also shown with powerful others' HLC. Participants highlighted the need to include realistic visual details and realistic interactions in the virtual environment for development.

Conclusions: The acceptability of a virtual reality game for sledding safety education was quite high, and relationships with HLC were shown in the expected directions. Based on participants' preferences, developers are advised to promote immersion in the game.

(JMIR Form Res 2025;9:e63813) doi:10.2196/63813

KEYWORDS

virtual reality; VR; sledding; safety education; Unified Theory of Acceptance and Use of Technology 2; UTAUT2; technology acceptability; cross sectional study

Introduction

Background

Sledding or tobogganing is a common recreational activity that is relatively low-cost and accessible, particularly developed in the alpine space (eg, Austria, Germany, or Switzerland). We can count more than 80 toboggan runs in Tyrol, Austria [1]. This physical activity, which consists of descending a mountain on a dedicated track with the use of a sled directed by a weight shift or by applying pressure on the runners, is widely practiced at all ages by families, friends, and at school during "snow class." However, sledding is rarely considered to cause serious injuries and occasional death [2]. Hospital-treated tobogganing injury cases recorded by the Victorian Injury Surveillance Unit between 2004 and 2006 involved all age groups, from 5 - 9 to 65 - 69, with peaks in ages 10 - 14, 20 - 24, and 40 - 44 years [3]. All age groups can therefore be considered to be concerned for potential sledding injuries. The report by the Swiss Council for Accident Prevention BFU (Beratungsstelle für Unfallverhütung), also highlights that all age groups are concerned by injuries, although 60% of injuries involve children

XSL•FO
RenderX

and youth [4]. It is often assumed that sledding injuries occur mainly in children, however an analysis conducted in Switzerland over 13 years demonstrated that 45% of injuries were reported in the ≤20 age group, 33% in the 20 - 40 age group, and 22% in the over-40 age group [5]. Consequently, the sledding injury prevention needs to address all age groups, which led us to consider adults first, as they can take on different roles as sledders, or as parents or teachers. As with road safety on bicycles, adults have an important role to play as parents and teachers in promoting safe sledding for children. The risk factors for sledding injury are (1) lack of hazard awareness and self-regulation, (2) lack of knowledge, and (3) insufficient skill in piloting techniques [4]. Developing self-regulation, knowledge, and skill in piloting techniques would promote better control of the sled, which would be a protective factor in avoiding injury. The need to develop sledding skills has become more pressing as the popularity tends to grow and the number of accidents increases [6]. Current actions to promote safe sledding are limited to announcements at the beginning of dedicated tracks, or didactic sheets presented by teachers in schools (eg, reports from the Swiss Council for Accident Prevention BFU). However, these actions are currently limited and do not allow the development of piloting technique skills.

Virtual Reality for Unintentional Injury Prevention

Unintentional injury prevention (like sledding injury prevention) could be improved by the use of technology-based interventions. Among the technology used for injury prevention, virtual reality programs are the most commonly used [7]. In this systematic review, interventions targeted various populations (eg, children, parents, or school staff). The results highlighted that virtual reality programs have improved behaviors, but showed little gains in knowledge. Another systematic review on virtual reality for unintentional injury prevention training with children showed no significant differences in knowledge gained between virtual reality groups and control groups [8]. Therefore, we might expect that a virtual reality sledding game for safety education (VRodel) would not support knowledge development, but would be an effective strategy for improving an individual's sledding skills, thus reducing the rate of luge injuries.

Virtual reality has been used in unintentional injury prevention to teach cognitive and motor skills (ie, decision-making or problem-solving skills, perceptual motor or psychomotor skills, procedural skills, and spatial skills) [8]. However, the transferability of these cognitive and motor skills to real-world environments was still mentioned as a critical gap in the literature [8]. Few studies have examined this gap nowadays. It has notably been shown in 1 study that similar behaviors were observed in real and virtual environments, even if speed was slightly overestimated in virtual environments [9]. Although virtual reality–based interventions are effective in developing skills in safety education contexts [7-9], it is necessary to examine them in the specific context of sledding safety education. However, before pursuing the long and expensive process of developing a virtual reality game for sledding safety, it is necessary to understand its potential reach. Although the emphasis needs to be on children and youth, all ages are concerned by sledding safety. It is therefore advisable to develop a game suitable for all potential user profiles (ie, sledders of any age, parents, or teachers). In particular, the acceptability of this type of technology and potential related factors needs to be examined in the light of recognized theoretical models.

Acceptability of Virtual Reality

Technology acceptability refers to the a priori more or less positive mental representation that a user has before using a tool, which reflects their explicit willingness to use it [10]. Several models have been developed among theories of technology acceptability [10]. A recent systematic literature review on the acceptability of augmented reality in the field of training and education [11] highlighted that the Technology Acceptance Model [12,13], the Unified Theory of Acceptance and Use of Technology [14], and their variants are the most commonly used models in this context. Another systematic review that examined the acceptability of immersive virtual technologies in the context of exercise promotion in older adults failed to highlight theoretical models of acceptability [15]. In fact, in this review, the acceptability was reduced to an indicator of satisfaction. Satisfaction is considered as a determinant of use, which refers to the enjoyment of use [10]. The concept of hedonic motivation (HM), which is defined as "the fun or pleasure derived from using a technology" [16], is therefore close to the perceived satisfaction with the tool. As the Unified Theory of Acceptance and Use of Technology 2 (UTAUT2) [12] is today considered to be the most integrative, parsimonious, and predictive model of technology acceptability [17,18] and for its potential and previous uses in the context of virtual reality [11,19,20], we choose to adopt this theoretical framework.

The UTAUT2 combined eight theoretical models into seven constructs: (1) performance expectancy (PE), defined as "the degree to which using a technology will provide benefits to consumers in performing certain activities"; (2) effort expectancy (EE), defined as "the degree of ease associated with consumers' use of technology"; (3) social influence (SI), defined as "the extent to which consumers perceive that important others (eg, family and friends) believe they should use a particular technology"; (4) facilitating conditions (FCs), defined as "consumers' perceptions of the resources and support available to perform a behavior"; (5) price value, defined as the "consumers' cognitive tradeoff between the perceived benefits of the applications and the monetary cost of using them"; (6) HM; and (7) habit (HT), defined as "the extent to which an individual believes the behavior to be automatic" [16]. These 7 constructs directly, or indirectly predict the future use of technology through behavioral intention (BI) [16]. Sex, age, and experience moderate these relationships in the model. For a new technology, without information on the actual price or the price of a similar technology, the relevance of the "price value" variable of the UTAUT2 could be questioned. Faced with this problem, Kapser and Abdelrahman [21] replaced "price value" by "price sensitivity" (PS). The PS, in the context of the UTAUT2, refers to the consumers' willingness to pay for a specific technology.

Previous studies have recently applied the UTAUT2 to virtual reality in various contexts. For example, Boel et al [19] showed that PE, SI, FCs, and HM are significant predictors of BI of

XSL•FO
RenderX

teachers to use immersive virtual reality in education. Bower et al [20], who studied the preservice teachers' intention to use immersive virtual reality in education, revealed that UTAUT2 provided a suitable model in this context. They showed a wide variety of ratings among the UTAUT2 dimensions, with HM receiving the highest scores and HT scoring the lowest. It has already been critically argued in the literature that all the UTAUT2 constructs did not consistently predict BI to use a technology [22,23]. Therefore, the application of the UTAUT2 in the specific context of virtual reality for sledding safety education needs to be examined.

Extensions of the UTAUT2

Several extensions have been made to the UTAUT2 into four directions: (1) new endogenous mechanisms, (2) new exogenous mechanisms, (3) new moderation mechanisms, or (4) new outcome mechanisms [23]. The UTAUT2 is robust and could be considered as a baseline model [23,24], however, other mechanisms could be considered regarding the specificities of the technology studied.

In the specific context of education and health promotion, we can consider the locus of control (LOC) or the health locus of control (HLC) as a potential extension of the UTAUT2. The HLC, originally conceptualized as a unidimensional construct, refers to the people's belief that their health (or other events for the LOC) is or is not determined by their behavior, with at one end "health-externals" who believe they have little control, and at the other end "health-internals" who believe that their health (or other events for the LOC) is the result of their behavior [25]. Later, the concept was refined into three differentiated dimensions: (1) the internal HLC, which refers to an active role in one's own health and taking responsibility toward health; (2) the powerful others external HLC, which refers to perceptions that other individuals, such as their physicians, control their health; and (3) the chance external HLC, which refers to the fact that health outcome is determined by luck, fate, and chance [26].

LOC has been previously reported to have relationships with acceptability constructs [27-29]. Internal LOC has been shown to be positively related to perceived usefulness, perceived ease of use, and perceived behavioral control in the context of learning [28,29]. Other authors [27] have shown that internal HLC was positively related to willingness to use health apps and willingness to use online trackers, as well as powerful others' HLC. Meanwhile, another study [30], showed no direct relationship between HLC and BI to use mobile health, but an indirect relationship through PE, EE, and SI. However, to the best of our knowledge, no study has been conducted in the specific context of safety education, which would be particularly appropriate given the positioning of safety education as a cross between education and health promotion. Therefore, the relationships between HLC and the UTAUT2 need to be further investigated, especially in this specific context.

Study Objectives and Hypotheses

This study will examine the acceptability of teachers, parents, and sledders of a virtual reality game for sledding safety education to guide the future development of the game. This study aimed to (1) characterize the acceptability of the virtual reality game for sledding safety education by parents, teachers, and sledders from the perspective of the UTAUT2-16 enriched by the HLC [26], and (2) understand participants' preferences and needs in terms of features to be integrated to the technology.

A first series of hypotheses directly relied on the UTAUT2. As it has been demonstrated in the original model [16], it can be hypothesized that PE, EE, SI, FCs, HM, and PS would be positively related to BI to use the VRodel.

A second series of hypotheses was concerned with the relationships between acceptability constructs and HLC. Based on previous studies from the literature demonstrating positive relationships between the acceptability constructs and internal HLC [27-30]. Positive correlations are therefore expected in our study. Concerning the relationships between powerful others' HLC, to the best of our knowledge, only 1 study demonstrated a positive correlation with willingness to use health apps and willingness to use online trackers [27]. As technology acceptability reflects users' willingness to use a technology [10], we expected positive correlations between the UTAUT2 constructs and powerful others' HLC.

Methods

Ethical Considerations

This study obtained ethical approval from the MCI (Management Center Innsbruck) Ethics Committee (number: 20230301). The administration of the questionnaires met the criteria of free participation, anonymity, and confidentiality of the responses, and was carried out online using LimeSurvey software (version 5.6.14; LimeSurvey, CE) with no possibility of missing data. All participants who voluntarily agreed to participate gave their electronic consent by clicking on "I agree to participate" in the online survey after having read the participant information sheet. The information sheet states that this acceptability study was conducted before the game's development, and that the results will guide future development. No compensation was offered for completing the survey. Participants were informed that they can interrupt this study at any time for any reason without enduring any disadvantage. All survey responses were collected anonymously, and the data were stored on a secure server (MCI servers).

Recruitment

All volunteer adults were eligible to participate in this study. Three participant profiles were sought: parents of children attending elementary schools; elementary school teachers; and sledders (individuals who go sledding at least once a year). Elementary schools in Tyrol (Austria) have been contacted by phone by JK to participate in the project by diffusing the link to the web-based questionnaire to parents and teachers of their schools. Additionally, posters were also distributed to ski resorts (Tyrol, Austria) to encourage participation in this study by sledders over the age of 18 years to ensure their free consent, as the study was only conducted online. The link to the survey was also publicly posted on social media by the research team members and project partners.

XSL•FO
RenderX

Measures

Demographic and General Information

Participants completed a questionnaire gathering demographic data including age, gender, education level, current working status, their frequency per year of sledding, and their use of virtual reality. They were also asked to choose the perspective from which they were responding to the survey: (1) teacher, (2) parents or legal guardians, or (3) sledder.

Acceptability of VRodel

Participants read an illustrated written description of a game based on virtual reality for sledding safety education, named VRodel (Multimedia Appendix 1). After, they were invited to complete the following subscales of the German version of the UTAUT-2 scale [31]: PE (4 items), EE (4 items), SI (3 items), FCs (4 items), HM (3 items), HT (4 items), and BI (3 items). Given the current development of the VRodel proof of concept, no price could be set in the description. Therefore, we replace the price value by the PS, as done by Kapser and Abdelrahman [21]. The PS was measured through 5 items [21,32]. All items were measured on a 7-point scale ranging from 1=strongly disagree to 7=strongly agree.

The tense of the items was changed from present to conditional, as VRodel was only a proof of concept, and to our knowledge, no similar technology was commercially available. Items were slightly adjusted to the specific context of sledding, and "Pokémon Go" was replaced by "VRodel." Therefore, we pretested the clarity of the material. The clarity of the VRodel description was assessed with 2 items (ie, "The VRodel description is clear and understandable"; "The description allows me to project myself in the use of VRodel"), and 1 item to examine the need for illustration (ie, "I will need an illustration to better project myself in the use of VRodel") on a 7-point Likert scale from 1=strongly disagree to 7=strongly agree. Participants rated the modified instrument by scoring the clarity (ie, "to what extent do you think the statement is clear and understandable?") of each item on a 7-point Likert scale ranging from 1=strongly disagree to 7=strongly agree.

Participants were also asked to justify and propose a rewording of low-rated items through comments. A first clarity survey was conducted with a convenience sample of 13 participants (4 females, 9 males; mean age (M_{age}) 25.6, SD 3.8 y). Results showed acceptable clarity scores (mean 5.3/7, SD 0.7). However, 11 items were below 5/7 and were rephrased. Even the VRodel description was clear and allowed the projection into usage (mean 6/7, SD 1), based on participants' comments, we have slightly rephrased the VRodel description and added a VRodel illustration (ie, requested at mean 6.4/7, SD 0.9). A second clarity survey was conducted with a sample of 11 participants from the target population (7 females, 4 males; 6 sledders, 3 teachers, and 2 parents; M_{age} 30.7, SD 9.0 y). The results showed acceptable clarity scores (mean 5.9/7, SD 0.5). All items were above 5/7 except 1 item (ie, HT3). This item was rephrased and examined in a third clarity survey with a sample of 11 participants from the target population (2 females, 9 males; 8 sledders, 2 teachers, and 1 parent; M_{age} 27.3, SD 6 y) and scored a mean of 6.2/7 (SD 0.8).

Because several adaptations were made, we examined the factor structure of the scale. The bifactorial confirmatory model with 27 items showed a good fit to the data: χ^2_{277}=425.23, P<.001, CFI=.92, TLI=.90, RMSEA=.067. Cronbach α ranged from 0.76 to 0.91 and were considered good [33], except for HT (α=0.45), which is lower and should be considered with caution.

Perception of Health Prevention

The German HLC questionnaire [34] was completed by participants. This scale contains three dimensions: (1) internal HLC (7 items), (2) powerful others' HLC (7 items), and (3) chance HLC (7 items). All items were measured on a 7-point scale ranging from 1=strongly disagree to 7=strongly agree. Cronbach α ranged from 0.69 to 0.77 and were considered reasonable [33].

Preferences and Needs for the Development of VRodel

Some questions were designed to measure preferences and needs for VRodel to define future development directions. Participants were asked for different locations (eg, school, ski resorts, hotels or lodges, tourist information centers, or at home) and to what extent they think VRodel should be implemented, on a scale from 1=strongly disagree to 7=strongly agree. Needs in terms of virtual reality immersive experience in VRodel (eg, realism or gamification) were also asked of participants to define priority development directions. Twelve affirmations were designed by the team members involved in the VRodel project: (1) realistic visual details (trees, terrain, etc); (2) realistic sounds; (3) simulation of accidents; (4) increase of vibrations with increasing speed; (5) ideal driving line marked as arrows or lines; (6) game elements (rewards, badges, tokens on the track to collect, etc); (7) information about safe behavior on the toboggan run (how to deal with other people, proper sitting position while tobogganing, or how to get up to the toboggan run); (8) other sledders who are also on the same toboggan run; (9) multiplayer mode with interaction with other players; (10) race mode (high scores based on your own time); (11) bonus items (speed enhancers or traction enhancers); (12) changes in visual details (night mode, fog, or poor visibility). For each item, participants were asked to rate on a 7-point scale ranging from 1=strongly disagree to 7=strongly agree their level of agreement with the following statement: "I think VRodel should include…"

Statistical Analysis

Analyses were performed using SPSS AMOS (version 23; IBM Corporation). The data was first checked to remove unusable surveys, and missing data was analyzed. The skewness ranged from −1.2 to 2, and the kurtosis ranged from −1.1 to 4, which can be considered a normal distribution [35].

Multiple regression analyses were used to examine the explained variance and the main contributors to BI to use VRodel. A cluster analysis, based on the K-means method, was then conducted to examine the acceptability groups. This approach has been used in several studies for similar purposes [36]. The optimal number of groups was determined based on the elbow method [37] and with a principle of parsimony. Following this method, 1 to 5 clusters were iterated, and the within-cluster sum of squares (ie, distance from the centroid of the cluster) was

XSL•FO
RenderX

calculated for each k-cluster. Then, the graph representing the within-cluster sum of squares on the y-axis and the k values (ie, number of clusters) on the x-axis was plotted. The within-cluster sum of squares decreases by adding more clusters. The optimal number of clusters was determined by identifying the k-cluster, before adding a new cluster, which results in only a marginal decrease in the within-cluster sum of squares. We labeled the groups regarding their characteristics of acceptability of VRodel. ANOVA was performed to assess potential age differences between groups. Chi-squared tests were used to assess potential differences in status (ie, parents, teachers, or sledders), gender, and previous virtual reality use between groups.

Correlations between acceptability constructs and HLC dimensions were computed. Then, a multivariate ANOVA was conducted to examine whether HLC differed between acceptability groups.

Participants were also invited to express their preferences and needs for VRodel through specific questions. These questions were summarized using descriptive statistics and presented according to the acceptability of VRodel clusters, if appropriate.

As the main analyses were the examination of relationships, the rule of thumb recommended a minimum of 50 participants [38]. For multivariate analyses, larger sample sizes are recommended. Therefore, a minimum of 100 participants was expected.

Results

Participants' Characteristics

Data was collected from March to May 2023. A total of 163 surveys were completed, of which 41 were excluded (ie, refusal of consent, n=3; incomplete survey without at least the items of acceptability of VRodel completed, n=26; abnormally short or long time to complete the survey, n=1; or individuals with the role of sledder with not at least 1 report of sledding per year, n=11).

The final sample of 122 participants (Table 1) was composed of 15 teachers (10 females, M_{age} 41.7, SD 11.2 y), 43 parents or legal guardians (23 females, M_{age} 46.6, SD 12.4 y), and 64 sledders (18 females, M_{age} 28.8, SD 16.6 y). Missing data (ie, HLC) are present for 12 individuals who stopped completing the survey before the end but they were considered in analyses.

XSL•FO
RenderX

Table . Characteristics of participants evaluating the acceptability of a virtual reality game for sledding safety education (N=122).

Variable and level		Value
Gender, n (%)		
	Male	71 (58.2)
	Female	51 (41.8)
Age (years), mean (SD)		36.8 (14.5)
Participant's profile, n (%)		
	Parent or legal guardian	43 (35.2)
	Teacher	15 (12.3)
	Sledder	64 (52.5)
Employment status, n (%)		
	Full-time employed	55 (45.1)
	Part-time employed	19 (15.6)
	Retired	7 (5.7)
	Studying	41 (33.6)
Level of education (years), n (%)		
	<12	14 (11.5)
	12	49 (40.2)
	15	30 (24.6)
	≥15	29 (23.8)
Frequency of sledding, n (%)		
	≤1 time per year	36 (29.5)
	2 to 4 times per year	44 (36.1)
	5 to 9 times per year	17 (13.9)
	≥10 times per year	25 (20.5)
Frequency of virtual reality use, n (%)		
	Never	30 (24.6)
	Sometimes	70 (57.4)
	Often to very often	22 (18.0)

Acceptability of VRodel

Acceptability constructs were rated as moderate to high (M_{PE} 4.32, SD_{PE} 1.29; M_{EE} 5.01, SD_{EE} 0.84; M_{SI} 3.77, SD_{SI} 1.40; M_{FC} 4.31, SD_{FC} 4.08; M_{HM} 5.22, SD_{HM} 1.24; M_{BI} 3.64, SD_{BI} 1.35), except for HT and PS which are perceived quite low (M_{HT} 2.40, SD_{HT} 1.24; M_{PS} 2.74, SD_{PS} 1.13). The acceptability constructs explained 65% of the variance in BI to use VRodel ($F_{7,114}$=32.82, P<.001). HM (β=.43, P<.001), HT (β=.36, P<.001), and PS (β=.28, P<.001) are significant contributors to the BI to use VRodel. Other acceptability constructs (ie, PE β=−.05, P=.53; EE β=.06, P=.36; SI β=.10, P=.21; FCs β=−.03, P=.62) were not significant contributors to the BI to use VRodel.

According to Figure 1, we can see an "elbow" [37] for the 2-cluster option. This solution was selected, and the clusters were examined for each acceptability dimension. The first cluster, composed of 53 participants, is characterized by low levels of all acceptability constructs, whereas the second cluster, composed of 69 participants, is characterized by high levels of all acceptability constructs (Figure 2). ANOVA showed no age differences between the acceptability clusters ($F_{1,120}$=0.98, P=.32). Chi-squared tests showed no difference between clusters for status (χ^2_2=0.79, P=.67), gender (χ^2_1=1.37, P=.24), or previous virtual reality use (χ^2_2=0.60, P=.74).

XSL•FO
RenderX

Figure 1. Within-cluster sum of squares reduction (percentage) for k-means clusters (ranging from 1 to 5) of acceptability of a virtual reality game for safety sledding education.

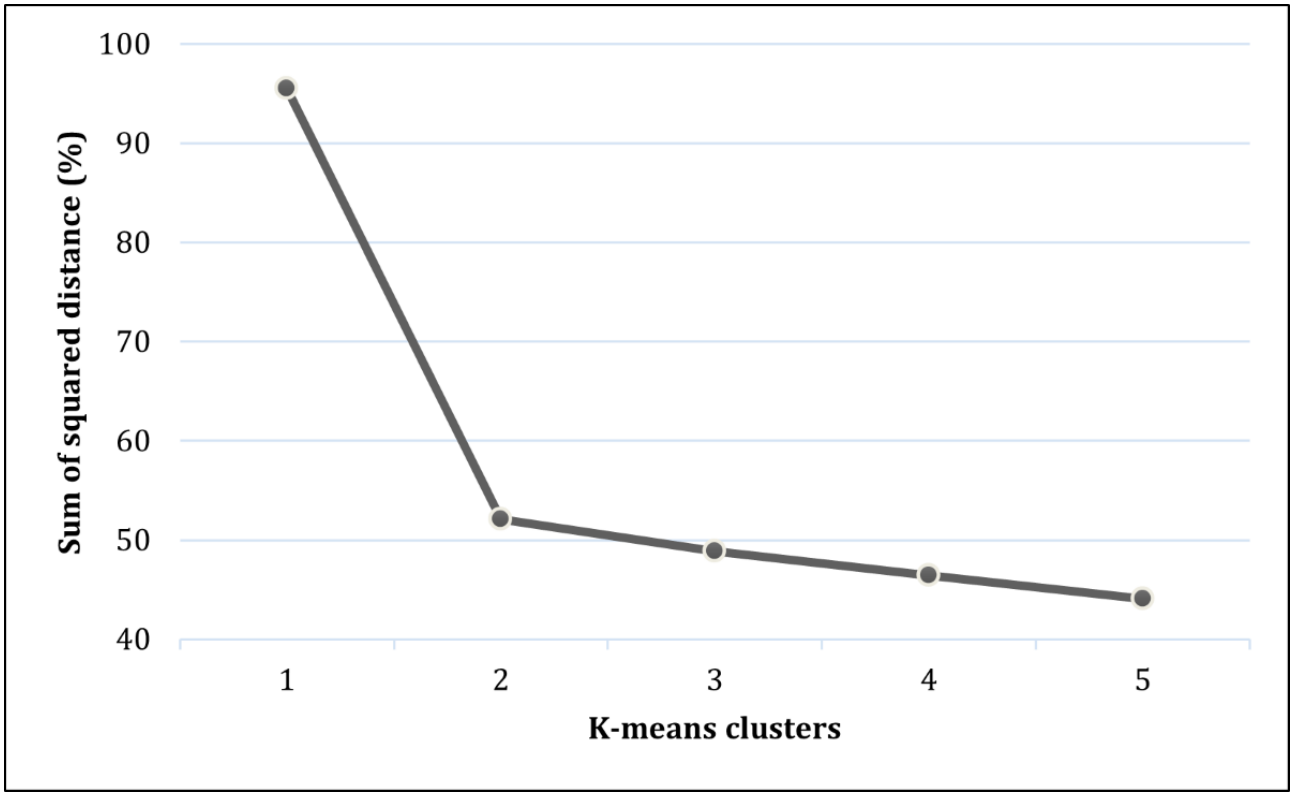

Figure 2. Z-scores of the acceptability of a virtual reality game for safety sledding education constructs in the low-level cluster (1; n=53) and the high-level cluster (2; n=69).

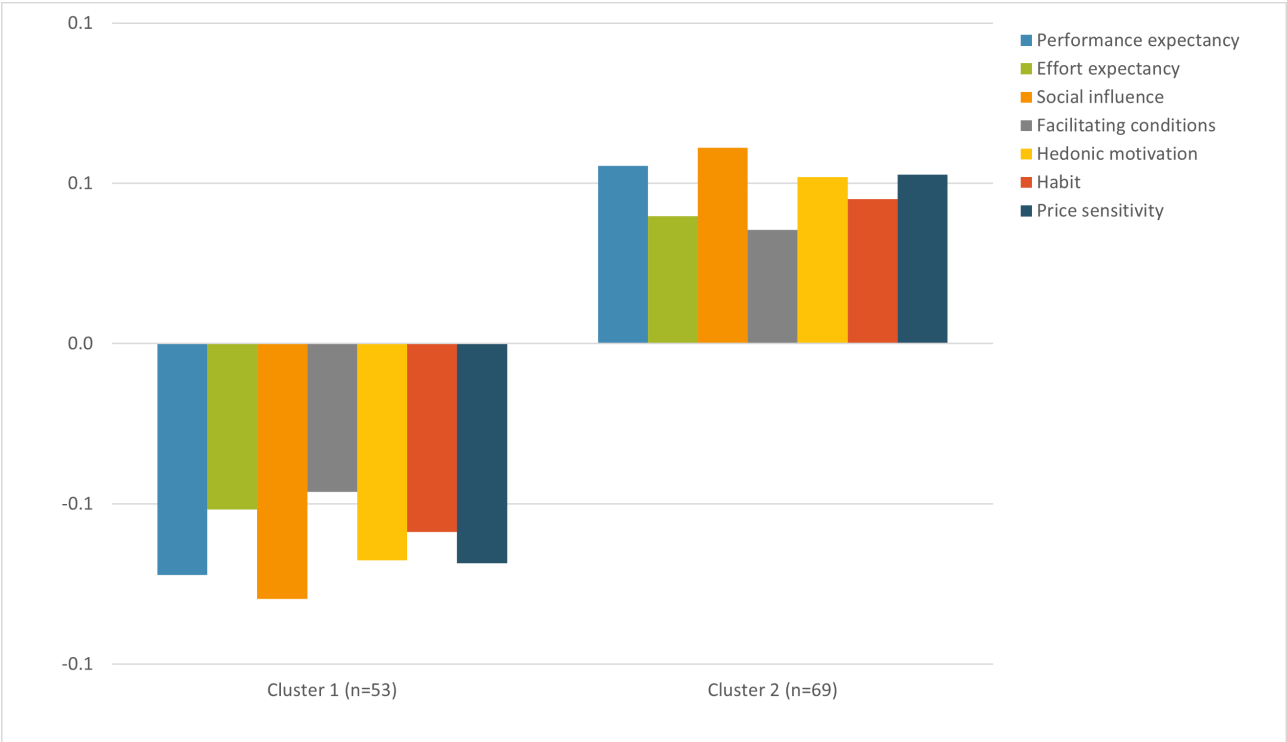

Relationships Between Acceptability of VRodel and HLC

Internal HLC was positively correlated to PE (r=0.29, P=.003), EE (r=0.30, P=.002), HM (r=0.31, P=.001), HT (r=0.25, P=.01), PS (r=0.20, P=.04), and BI to use VRodel (r=0.27, P=.004). Powerful others' HLC was positively correlated to PE (r=0.30, P=.003), SI (r=0.31, P=.001), HT (r=0.41, P<.001), PS (r=0.38, P<.001), and BI to use VRodel (r=0.22, P=.02). Chance HLC was positively correlated to HT (r=0.22, P=.02).

XSL•FO
RenderX

A 1-way multivariate ANOVA revealed a main effect of cluster membership on HLC (Wilks =.86, $F_{3,106}$=5.96, P=.001, η_p^2=.14). Participants of the 2 acceptability of VRodel clusters exhibited differences in terms of internal HLC ($F_{1,108}$=7.17, P=.009, η_p^2=.06) and powerful others' HLC ($F_{1,108}$=12.89, P<.001, η_p^2=.11). Participants in the low acceptability of VRodel cluster (n=47) tend have lower levels of internal HLC ($M_{\text{Internal HLC}}$ 4.45, $SD_{\text{Internal HLC}}$ 0.76) and powerful others' HLC ($M_{\text{Powerful others' HLC}}$ 3.47, $SD_{\text{Powerful others' HLC}}$ 0.78) than participants in the high acceptability of VRodel cluster (n=63; $M_{\text{Internal HLC}}$ 4.85, $SD_{\text{Internal HLC}}$ 0 .79; $M_{\text{Powerful others' HLC}}$ 3.98, $SD_{\text{Powerful others' HLC}}$ 0.71). No differences were found between clusters for chance HLC.

Preferences and Needs for the Development of VRodel

Regarding perceptions of where VRodel should be integrated, no differences emerged between the acceptability clusters, therefore, the results were presented for the overall sample of participants. VRodel implementation locations have been ranked in the following order: hotels or lodges (mean 5.64/7, SD 1.07), tourist information centers (mean 5.35/7, SD 1.42), ski resorts (mean 5.13/7, SD 1.64), at home (mean 4.76/7, SD 1.73), and at school (mean 4.28/7, SD 1.76).

Regarding the priority of features to be implemented in VRodel, some differences emerged between the clusters, therefore, the results were presented for each cluster (Table 2). All features presented to the participants scored higher than 4/7, except for the bonus items in the low acceptability of VRodel cluster, meaning that all features could be interesting to implement. However, realistic visual details, information about safe behaviors, and realistic interactions in the virtual environment through speed-dependent vibration were the top 3 features requested by all participants.

Table . Evaluation of the features to implement in VRodel from the high- and low-level clusters of acceptability of a virtual reality game for safety sledding education.

Features to implement in VRodel	Low-level cluster of acceptability of VRodel[a] (n=53), mean (SD)	High-level cluster of acceptability of VRodel[a] (n=69), mean (SD)	Priority
Realistic visual details	6.2 (1.2)	6.4 (1)	1
Information about safe behavior on the toboggan run	6.1 (1.2)	6.3 (1.1)	2
Increase in vibrations with increasing speed	5.8 (1.2)	6.3 (0.9)	3[b]
Changes in visual details (night mode, fog, or poor visibility)	5.7 (1.3)	6.1 (1.2)	4
Realistic sounds	5.4 (1.4)	6 (1.2)	5[b]
Other tobogganers who are also on the same toboggan run	5.7 (1.2)	5.8 (1.3)	6
Simulation of accidents	5.5 (1.5)	5.8 (1.3)	7
Ideal driving line marked as arrows or lines	5.4 (1.4)	5.7 (1.2)	8
Game elements (rewards, badges, tokens on the track to collect, etc)	4.9 (1.9)	5.6 (1.3)	9[b]
Multiplayer mode with interaction with other players	5 (1.7)	5.5 (1.4)	10
Race mode (high scores based on your own time)	5.2 (1.6)	5.3 (1.7)	11
Bonus items (speed enhancers and traction enhancers)	3.6 (1.8)	4.3 (1.8)	12[b]

[a]Scores are expressed on a 7-point scale ranging from 1=strongly disagree to 7=strongly agree and presented as the mean and SD for each cluster.
[b]Significant difference (P<.05) between the two clusters.

Discussion

Principal Results

This study carried out an important first step in the development process of a new tool, the evaluation of its acceptability. A strength of our study is that acceptability characterization was based on strong theoretical foundations with the perspective of the UTAUT2 [16] enriched by the HLC framework [26]. The results showed moderate to high scores of acceptability, except for HT and PS, which are perceived as quite low. HM was the construct of acceptability rated as the highest, whereas HT was the lowest. HM (β=.43, P<.001), HT (β=.36, P<.001), and PS (β=.28, P<.001) were significant predictors of BI to use VRodel (65% of explained variance). These results suggest that VRodel should be well accepted, as long as it will integrate gamified

XSL•FO
RenderX

components, will not be too expensive, and will be based on the actual low virtual reality usage of future users. Two clusters emerged with low and high levels of acceptability, respectively. No differences were found between the characteristics of participants in the 2 clusters (ie, age, participant's profile, gender, or previous use of virtual reality). Therefore, it seems that a common game could be first drafted for all users and then declined depending on its effectiveness, which would be examined in a future study.

Internal and powerful others' HLC were higher in the high acceptability of the VRodel cluster than in the low acceptability of the VRodel cluster. The acceptability of a virtual reality game for sledding safety education constructs was correlated with internal and powerful others' HLC. To the best of our knowledge, this is the first study to demonstrate relationships between HLC and UTAUT2 for virtual reality in an educational setting. It is therefore important to note that for virtual reality with a health educational purpose, the HLC is suitable to enrich the UTAUT2.

"Realistic visual details" was the most requested feature. This feature can be achieved through photogrammetry. The use of photogrammetry for the design of virtual environments is not novel [39], but it tends to become more widespread for sports purposes [40]. It is possible to create a virtual world that is perceived as real from the ground up, but it is necessary to record high-quality video with drones (eg, Phantom 4 or Mavic 3 from Da Juang Innovation [DJI] Enterprise). The position measurement of the drone could be improved by using a ground station. This technique could also be combined with the Unity Terrain Tools software. All features of VRodel have been well evaluated, so we can recommend the integration of all of them according to their priority level established in this study. Finally, the implementation of VRodel was not positioned primarily for diffusion and integration in schools, as team members had assumed. In fact, the school appeared to be the least appropriate place to implement this technology. It would thus be appropriate to work further with tourist sites (ie, hotels or lodges, tourist information centers, or ski resorts) to establish the technical specifications for the development of the VRodel game. The findings of this study have significant implications to help developers and engineers build an immersive virtual reality game for sledding safety education.

Limitations

This study nevertheless has some limitations. First, participants only read an illustrated written description of a game based on virtual reality for sledding safety education. The description was short which could lead to unrealistic expectations for the game. Therefore, the acceptability scores cannot be generalized, and another study of acceptability with actual testing of the game would be necessary.

Another limitation is the cross-sectional design of this study, where all measures were collected at the same time. Methodological precautions were taken by using validated questionnaires in a specific order, and testing of their validity and reliability after small modifications. Participation in this study was voluntary with no incentive, and slow responses were excluded, so we could expect a sufficient level of motivation from participants to complete the online survey. However, a common method bias could still occur [41]. In future studies of the acceptability of VRodel with actual testing of the game, it would be interesting to obtain measures of the predictor and criterion variables from different sources and including a temporal separation between them [41].

Although no differences emerged between the different profiles of participants, we recognize that the number of teachers was small. Additional studies would be needed to better understand teachers' perceptions if VRodel was implemented in school for teaching sledding safety (although this type of implementation was not recognized as a priority). The representativeness of the sample is therefore limited, and although the results have led to technical specifications for the development of a functional VRodel prototype, it would be necessary to continue involving future potential users with different profiles in the following stages of VRodel development.

Comparison With Prior Work

Hypotheses that acceptability constructs would predict the BI to use VRodel, as per the original model [16], were partially validated. Our study demonstrated that only HM, HT, and PS were significant contributors to the BI to use VRodel. Therefore, in this specific context of virtual reality for safety sledding, not all acceptability constructs predict BI. This is in line with other studies demonstrating the differences regarding the specificities of the contexts and technologies considered [22,23]. Per Boel et al [19] and with Bower et al [20] in the context of immersive virtual reality in education, HM also predicts BI. PS, which is related to price value, was shown to be a predictor of BI, as in the study by Tseng et al [42]. Although HT was found to be a predictor of BI, no differences emerged in terms of previous experience with virtual reality between the acceptability of VRodel clusters. Surprisingly, PE was not a predictor of BI as in other similar studies [19,42]. A potential explanation could be the lack of knowledge about the effectiveness of virtual reality for safety education, resulting from the presentation of a proof of concept of the VRodel to be developed, rather than the real game with prior interaction with it. Few other studies have already shown no relationship between PE and BI in the education context [43]. It would be necessary to further investigate this relationship after usage of the game.

Internal and powerful others' HLC were positively correlated to certain constructs of the UTAUT2, in line with our hypotheses. In line with Bennett et al [27], internal and powerful others' HLC were positively correlated with BI to use VRodel. Our results were also partially in line with Ahadzadeh et al [30] and showed that PE, EE, and SI were related to HLC. Although no relationship was expected between chance HLC and UTAUT2 constructs, a small correlation was found between HT and chance HLC. This result should be interpreted with caution, given that the internal consistency of the HT subscale was rather low (ie, α=0.45). While the relationships between UTAUT2 and the 3 subscales of HLC have been poorly investigated in the literature, these findings are congruent with other similar studies [27-29].

XSL•FO
RenderX

Conclusions

This study characterized the acceptability of the virtual reality game for sledding safety education by parents', teachers', and sledders' perspectives by applying the UTAUT2 in conjunction with the HLC framework. It is therefore the first study to demonstrate that HLC is suitable for enriching the UTAUT2 in the context of virtual reality for unintentional injury prevention. Important directions for the effective development of a virtual reality game for sledding safety education have been highlighted and prioritized. Future studies would be needed to examine the effectiveness of such a technology and the transferability of knowledge and skills acquired in a virtual environment to the real-world environment of a sled track.

Acknowledgments

This work was supported by the "Innovationsförderung" from the Land Tyrol. MH has received financial support from the Ulysseus European University Alliance and the IDEX UCA[JEDI] to come to the Department of Medical, Health & Sport Engineering in the Management Center Innsbruck (Austria) for a research exchange from February to May 2023. This study was conducted as part of this research exchange. The funding sources had no role in the design, conduct, analysis, or reporting of the study. The authors sincerely thank all the participants. They also thank the team members of the Department of Medical, Health & Sport Engineering, in Management Center Innsbruck (Innsbruck, Austria); the management staff, teachers, and social workers of the schools in Tyrol, and the sport centers for their aid in diffusing the survey. Rodel Austria Federation; Martin Abentung, MSc; and Markus Moser, MBA, were partners in the project, and we thank them for their involvement. No generative artificial intelligence tools were used to create the original content for publication.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Original German version and translated English version description of the virtual reality sledding game for safety education.
[DOCX File, 201 KB - formative_v9i1e63813_app1.docx]

Checklist 1
RATE-XR checklist. RATE-XR: Reporting for the Early-Phase Clinical Evaluation of Applications Using Extended Reality.
[PDF File, 175 KB - formative_v9i1e63813_app2.pdf]

References

1. Tobogganing in Tyrol. Tyrol. URL: https://www.tyrol.com/activities/sport/tobogganing [accessed 2025-05-05]
2. Shorter NA, Mooney DP, Harmon BJ. Snowboarding injuries in children and adolescents. Am J Emerg Med 1999 May;17(3):261-263. [doi: 10.1016/s0735-6757(99)90121-3] [Medline: 10337886]
3. Ashby K, Cassell E. Injury in snow and ice sports. : Victorian Injury Surveillance Unit; 2007 URL: https://www.monash.edu/__data/assets/pdf_file/0004/218479/haz66.pdf [accessed 2025-05-05]
4. Bianchi G, Brügger O, Kerschbaumer D. Sicherheitsanalyse des schlittelns und rodelns in der schweiz: unfall- und risikofaktorenanalyse sowie präventionsempfehlungen [Report in German]. : Beratungsstelle für Unfallverhütung; 2012 URL: https://www.researchgate.net/profile/Othmar-Bruegger/publication/260822491_Sicherheitsanalyse_des_Schlittelns_und_Rodelns_in_der_Schweiz/links/00b7d532706e65021f000000/Sicherheitsanalyse-des-Schlittelns-und-Rodelns-in-der-Schweiz.pdf [accessed 2025-05-05]
5. Heim D, Altgeld K, Hasler RM, Aghayev E, Exadaktylos AK. A 13-year analysis from Switzerland of non-fatal sledging (sledding or tobogganing) injuries. Injury 2014 Jan;45(1):338-341. [doi: 10.1016/j.injury.2012.08.027] [Medline: 23010070]
6. Clos M, Evain JN, Wroblewski I, Bouzat P, Mortamet G. Serious sledding injuries in children dramatically increased during the COVID-19 pandemic. Acta Paediatr 2022 Feb;111(2):389-390. [doi: 10.1111/apa.16097] [Medline: 34491604]
7. Omaki E, Rizzutti N, Shields W, et al. A systematic review of technology-based interventions for unintentional injury prevention education and behaviour change. Inj Prev 2017 Apr;23(2):138-146. [doi: 10.1136/injuryprev-2015-041740]
8. Santiago A, Matthews B, Chen WS, Keane T, Kaufman J. Immersive virtual reality for unintentional injury prevention training with children: a systematic review. Comput Educ: X Reality 2024 Dec;5:100072. [doi: 10.1016/j.cexr.2024.100072]
9. Bhagavathula R, Williams B, Owens J, Gibbons R. The reality of virtual reality: a comparison of pedestrian behavior in real and virtual environments. Proc Hum Factors Ergono Soc Annu Meet 2018 Sep;62(1):2056-2060. [doi: 10.1177/1541931218621464]

XSL•FO
RenderX

10. Alexandre B, Reynaud E, Osiurak F, Navarro J. Acceptance and acceptability criteria: a literature review. Cogn Tech Work 2018 May;20(2):165-177. [doi: 10.1007/s10111-018-0459-1]
11. Graser S, Böhm S. A systematic literature review on technology acceptance research on augmented reality in the field of training and education. arXiv. Preprint posted online on Nov 21, 2024. [doi: 10.48550/arXiv.2411.13946]
12. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989 Sep;13(3):319. [doi: 10.2307/249008]
13. Davis FD, Bagozzi RP, Warshaw PR. User acceptance of computer technology: a comparison of two theoretical models. Manage Sci 1989 Aug;35(8):982-1003. [doi: 10.1287/mnsc.35.8.982]
14. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view. MIS Q 2003;27(3):425. [doi: 10.2307/30036540]
15. Doré B, Gaudreault A, Everard G, et al. Acceptability, feasibility, and effectiveness of immersive virtual technologies to promote exercise in older adults: a systematic review and meta-analysis. Sensors (Basel) 2023 Feb 24;23(5):2506. [doi: 10.3390/s23052506] [Medline: 36904709]
16. Venkatesh V, Thong JYL, Xu X. Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology. MIS Q 2012;36(1):157. [doi: 10.2307/41410412]
17. Alkhwaldi A, Kamala M. Why do users accept innovative technologies? A critical review of models and theories of technology acceptance in the information system literature. J Multidiscip Eng Sci 2017;4(8):7962-7971 [FREE Full text]
18. Chang A. UTAUT and UTAUT 2: a review and agenda for future research. The Winners 2012 Sep 28;13(2):10. [doi: 10.21512/tw.v13i2.656]
19. Boel C, Rotsaert T, Valcke M, Rosseel Y, Struyf D, Schellens T. Are teachers ready to immerse? Acceptance of mobile immersive virtual reality in secondary education teachers. Res Learn Technol 2023;31:1-14. [doi: 10.25304/rlt.v31.2855]
20. Bower M, DeWitt D, Lai JWM. Reasons associated with preservice teachers' intention to use immersive virtual reality in education. Brit J Educational Tech 2020 Nov;51(6):2215-2233. [doi: 10.1111/bjet.13009]
21. Kapser S, Abdelrahman M. Acceptance of autonomous delivery vehicles for last-mile delivery in Germany – extending UTAUT2 with risk perceptions. Transp Res Part C: Emerging Technol 2020 Feb;111:210-225. [doi: 10.1016/j.trc.2019.12.016]
22. Tamilmani K, Rana N, Dwivedi Y. Mobile application adoption predictors: systematic review of UTAUT2 studies using weight analysis. Presented at: 17th Conference on e-Business, e-Services, and e-Society, I3E 2018; Oct 30 to Nov 1, 2018; Kuwait City, Kuwait. [doi: 10.1007/978-3-030-02131-3_1]
23. Venkatesh V, Thong JYL, Xu X. Unified theory of acceptance and use of technology: a synthesis and the road ahead. JAIS 2016;17(5):328-376 [FREE Full text] [doi: 10.17705/1jais.00428]
24. Tamilmani K, Rana NP, Wamba SF, Dwivedi R. The extended Unified Theory of Acceptance and Use of Technology (UTAUT2): a systematic literature review and theory evaluation. Int J Inf Manage 2021 Apr;57:102269. [doi: 10.1016/j.ijinfomgt.2020.102269]
25. Wallston BS, Wallston KA, Kaplan GD, Maides SA. Development and validation of the Health Locus of Control (HLC) Scale. J Consult Clin Psychol 1976 Aug;44(4):580-585. [doi: 10.1037/0022-006X.44.4.580] [Medline: 939841]
26. Wallston KA, Wallston BS, DeVellis R. Development of the multidimensional Health Locus of Control (MHLC) Scales. Health Educ Monogr 1978;6(2):160-170. [doi: 10.1177/109019817800600107] [Medline: 689890]
27. Bennett BL, Goldstein CM, Gathright EC, Hughes JW, Latner JD. Internal health locus of control predicts willingness to track health behaviors online and with smartphone applications. Psychol Health Med 2017 Dec;22(10):1224-1229. [doi: 10.1080/13548506.2017.1317354] [Medline: 28415852]
28. Hsia JW. The effects of locus of control on university students' mobile learning adoption. J Comput High Educ 2016 Apr;28(1):1-17. [doi: 10.1007/s12528-015-9103-8]
29. Hsia JW, Chang CC, Tseng AH. Effects of individuals' locus of control and computer self-efficacy on their e-learning acceptance in high-tech companies. Behav Inf Technol 2014 Jan 2;33(1):51-64. [doi: 10.1080/0144929X.2012.702284]
30. Ahadzadeh AS, Wu SL, Ong FS, Deng R. The mediating influence of the Unified Theory of Acceptance and Use of Technology on the relationship between internal health locus of control and mobile health adoption: cross-sectional study. J Med Internet Res 2021 Dec 29;23(12):e28086. [doi: 10.2196/28086] [Medline: 34964718]
31. Harborth D, Pape S. German translation of the Unified Theory of Acceptance and Use of Technology 2 (UTAUT2) questionnaire. SSRN J 2018. [doi: 10.2139/ssrn.3147708]
32. Goldsmith RE, Kim D, Flynn LR, Kim WM. Price sensitivity and innovativeness for fashion among Korean consumers. J Soc Psychol 2005 Oct;145(5):501-508. [doi: 10.3200/SOCP.145.5.501-508] [Medline: 16201674]
33. Taber KS. The use of Cronbach's alpha when developing and reporting research instruments in science education. Res Sci Educ 2018 Dec;48(6):1273-1296. [doi: 10.1007/s11165-016-9602-2]
34. Lohaus A, Schmitt GM. Kontrollüberzeugungen zu krankheit und gesundheit (KKG): bericht über die entwicklung eines testverfahrens [Report in German]. : Universitätsbibliothek Bielefeld; 1989.
35. Hair JF, Black WC, Babin BJ, Anderson RE. Multivariate Data Analysis: Pearson Education Limited; 2014.
36. Hellec J, Hayotte M, Chorin F, Colson SS, d'Arripe-Longueville F. Applying the UTAUT2 model to smart eyeglasses to detect and prevent falls among older adults and examination of associations with fall-related functional physical capacities: survey study. J Med Internet Res 2023 May 12;25:e41220. [doi: 10.2196/41220] [Medline: 37171835]

XSL•FO
RenderX

37. Kodinariya TM, Makwana PR. Review on determining of cluster in k-means clustering. Int J Adv Res Comput Sci Manag Stud 2013 Nov;1(6):90-95.
38. Van Voorhis CRW, Morgan BL. Understanding power and rules of thumb for determining sample sizes. TQMP 2007 Sep 1;3(2):43-50. [doi: 10.20982/tqmp.03.2.p043]
39. Statham W. Use of photogrammetry in video games: a historical overview. Games Cult 2020 May;15(3):289-307. [doi: 10.1177/1555412018786415]
40. Kreiner J, Hollaus B, Heyer Y, Spörk V. Photogrammetry and how to make backcountry skiing safer. Presented at: EASM; Sep 5-8, 2022; Innsbruck, Austria.
41. Podsakoff PM, Podsakoff NP, Williams LJ, Huang C, Yang J. Common method bias: it's bad, it's complex, it's widespread, and it's not easy to fix. Annu Rev Organ Psychol Organ Behav 2024 Jan 22;11(1):17-61. [doi: 10.1146/annurev-orgpsych-110721-040030]
42. Tseng TH, Lin S, Wang YS, Liu HX. Investigating teachers' adoption of MOOCs: the perspective of UTAUT2. Interact Learn Environ 2022 Apr 3;30(4):635-650. [doi: 10.1080/10494820.2019.1674888]
43. Acosta-Enriquez BG, Farroñan EVR, Zapata LIV, et al. Acceptance of artificial intelligence in university contexts: a conceptual analysis based on UTAUT2 theory. Heliyon 2024 Oct 15;10(19):e38315. [doi: 10.1016/j.heliyon.2024.e38315] [Medline: 39430455]

Abbreviations

BFU: Beratungsstelle für Unfallverhütung
BI: behavioral intention
EE: effort expectancy
FC: facilitating condition
HLC: health locus of control
HM: hedonic motivation
HT: habit
LOC: locus of control
MCI: Management Center Innsbruck
PE: performance expectancy
PS: price sensitivity
SI: social influence
UTAUT2: Unified Theory of Acceptance and Use of Technology-2
VRodel: virtual reality sledding game for safety education

Edited by A Mavragani; submitted 30.06.24; peer-reviewed by D Patel, MT Goi; revised version received 07.04.25; accepted 08.04.25; published 21.05.25.

Please cite as:
Hayotte M, Kreiner J, Hollaus B
Parents', Teachers', and Sledders' Acceptability of a Virtual Reality Game for Sledding Safety Education: Cross-Sectional Study
JMIR Form Res 2025;9:e63813
URL: https://formative.jmir.org/2025/1/e63813
doi:10.2196/63813

© Meggy Hayotte, Jonas Kreiner, Bernhard Hollaus. Originally published in JMIR Formative Research (https://formative.jmir.org), 21.5.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Co-Designed Online Training Program for Worry Management: The Role of Young People With Lived Experience of Worry in Program Development

Jessica Steward[1], MSc; Michelle L Moulds[2], PhD; Colette R Hirsch[1,3], PhD

[1]Institute of Psychiatry, Psychology and Neuroscience, King's College London, 16 De Crespigny Park, London, United Kingdom

[2]School of Psychology, UNSW Sydney, Sydney, New South Wales, Australia

[3]South London and Maudsley NHS Foundation Trust, London, United Kingdom

Corresponding Author:
Colette R Hirsch, PhD
Institute of Psychiatry, Psychology and Neuroscience, King's College London, 16 De Crespigny Park, London, United Kingdom

Abstract

Background: Many young people report high levels of worry, highlighting the need for interventions that teach strategies to help them shift focus away from worry. To maximize uptake by this population, interventions should be brief and accessible; to maximize dissemination, they should have potential for delivery at scale. We produced a multisession, online training program, Shift Focus, co-designed with young people with lived experience of worry. The online training program was accessed via a mobile app. In this paper, we describe how Lived Experience Advisory Panel (LEAP) members were involved in each stage of the process of developing the Shift Focus online training program, from refining session content through to designing and testing the online training program prototype.

Objective: We aimed to engage with young people with lived experience of worry, to help refine, further develop, and tailor a new online training program designed to help shift focus away from worry.

Methods: We recruited LEAP members (aged 16‑25 y) with lived experience of worry from diverse backgrounds across the United Kingdom. We used a highly iterative participatory design process, such that LEAP members provided input during all 4 phases of program development: refining and further developing session content, piloting sessions, developing user experience design, and testing the online training program prototype.

Results: Feedback from LEAP members during each phase of the online platform development informed key decisions regarding the platform content, functionality, and the interface design to ensure it suited our target population. In phase 1, we learned that the platform needed to be simple and aesthetically pleasing, personalized to individual needs and preferences, accessible to all, track progress, and provide individuals with a sense of community with others with similar lived experiences. In phase 2, we learned that the platform also needed to provide further guidance on how to apply the Shift Focus techniques to daily life, using personalized reminder settings. In phase 3, we additionally learned that ease of navigation and interactivity were key to maintaining user engagement. The importance of program tracking was reiterated, as well as the need for accessibility settings to support all learning styles. In phase 4, we identified that technical problems with the online platform were a barrier to engagement. The inclusion of future iterations (eg, reward systems) to help promote engagement was suggested by LEAP members in multiple phases.

Conclusions: LEAP members brought unique expertise and made key contributions to the development of the Shift Focus online training program and were highly valued members of the team. A highly iterative participatory design process enabled continuous feedback from LEAP members throughout, ensuring that their input was meaningful and that their key messages and ideas were incorporated into the final program.

(JMIR Form Res 2025;9:e66461) doi:10.2196/66461

KEYWORDS

digital intervention; mental health; patient and public involvement; young people; adolescent; experts by experience; participatory; anxiety; development; design; user experience

XSL•FO
RenderX

Introduction

Many young people experience high levels of uncontrollable worry [1], which can lead to great distress and be linked to psychological disorders such as generalized anxiety disorder (GAD). GAD is characterized by excessive and uncontrollable worries about everyday events or activities that are out of proportion to the inherent risk [2]. Uncontrollable worry takes up limited capacity resources, and such resources are needed for individuals to shift their focus from worry to other tasks or streams of thinking [3]. Given this, high worriers may benefit from cognitive techniques designed to shift attention away from worry.

Experimental studies provide preliminary support for this possibility. Eagleson et al [4] trained individuals presenting with GAD to shift focus away from worry by thinking about positive outcomes of their worries in either imagery or verbal based form, while participants in the control group were trained to shift focus away from worry by generating an unrelated positive image which was not associated with their worries. Participants practiced their designated technique for 1 week. In all 3 conditions, participants reported fewer negative intrusive thoughts and lower scores on self-reported measures of worry and anxiety at follow-up. Given this, although originally conceived as a control condition, generating an unrelated positive image appeared to be helpful for adults presenting with worry. Sawjani et al [5] extended this study and demonstrated that high worriers could be trained (in a single session) to generate an unrelated positive image to stop worry in the short term.

Given that worry is a difficulty experienced by many young people, we built on and extended the single session training developed by Sawjani et al [5] to develop a multisession online training program (Shift Focus), specifically tailored for young people with high levels of worry. Shift Focus is comprised of 8 brief sessions, which include psychoeducation about worry, guided imagery exercises, and within-session exercises practicing shifting focus away from worry and teaching ways to apply the skills in daily life.

We reasoned that an intervention for young people should be brief and accessible, and also have scope for flexible delivery. Digital health interventions (DHIs) offer such flexibility, providing the opportunity for this population to access evidence-based psychological interventions at a time and place of their choosing. Moreover, mobile apps promote independent skills practice in real time, enabling the application of techniques when they are most needed in day-to-day life. Accordingly, we developed an online training program that was accessed via a mobile app, as a means by which to teach young people cognitive strategies to shift their focus away from worry when they want to.

Despite the abovementioned benefits of DHI, recent reviews have found that while effective, they typically have poor uptake and service user engagement [6,7]—limiting their capacity to deliver longer-term mental health benefits. Engaging stakeholders in the development and tailoring of DHI is essential to maximize engagement and improve acceptability. Working closely with individuals with lived experience is a crucial element in developing new acceptable DHI. Accordingly, we recruited and engaged a panel of young people with lived experience of worry (a Lived Experience Advisory Panel [LEAP]) to obtain their feedback on Shift Focus, and leveraged their invaluable insights to refine and further develop the program.

Patient and public involvement (PPI) in research has become increasingly common internationally [8], and refers to contributions from people with lived experience of a psychological or physical health condition under study, including input into how research is designed, conducted, and disseminated [9]. Different forms of knowledge and expertise can then be brought together, eradicating traditional divides between researchers and service users that in the past may have been grounded in assumptions which differentiate “expert” researchers from people with lived experience, traditionally considered nonexpert. There has been some criticism of PPI initiatives because they are too narrow and tokenistic in nature (eg, [10,11]), and in turn, a call for PPI members to be more empowered within research settings [12]. Encouragingly, there is evidence that meaningful PPI engagement improves the quality of research, ensuring that it is relevant to the needs of the community, thus enabling more positive impact [13].

Our goal in this paper is to outline the range of ways in which our LEAP members contributed to the process of developing, co-designing, and testing the acceptability of Shift Focus, an online, standalone, multisession training program (delivered via a mobile app) for young people (16 - 25 y) who report difficulty stopping worry. Specifically, we wanted to determine whether the online training program was acceptable to our target population of young people. To do so, we sought feedback on LEAP members’ satisfaction with the program and willingness to engage, and on their perceptions of the program’s capacity to help them manage worry. We obtained feedback on the content, look, and feel of the online training program over different phases of development, with the goals of maximizing potential user satisfaction and engagement of the target population when they are offered the finalized version of the online training.

LEAP members’ involvement was made up of 4 phases. The aim of phase 1 was to develop session content. Specifically, LEAP members provided feedback on the sessions produced by the researchers to further develop and refine session content. Additionally, we sought feedback on the acceptability of the intervention and sought feedback to help identify and address any barriers to young people’s understanding of session content. Informed by this feedback, we modified the intervention with the goal of building sessions that were tailored to young people’s needs and preferences.

In phase 2 we explored feasibility; that is, the online training program’s practicality and the ease with which young people could implement it in their daily lives. Specifically, we were interested in whether young people had the time and resources to engage in the program, as well as their views on whether it was feasible that they would use the techniques taught in the program in day-to-day life. We also obtained LEAP feedback

XSL•FO
RenderX

on the modified content (ie, changes made in response to feedback in phase 1) and additions (eg, push notifications), and explored acceptability.

The aim of phase 3 was to develop user experience (UX) design. This involved engaging in a user journey mapping workshop with UX designers to identify what users see, think, feel, and do when they interact with the online platform, and to obtain feedback on any difficulties that they experienced. The purpose of the workshop was to ensure that the UX designers had a clear understanding of the user group. Phase 3, therefore, involved an iterative process with LEAP members and UX designers, so that the online platform would be suited to young people's needs and preferences.

The final phase (phase 4) involved an online training program testing. LEAP members provided final feedback on the online platform. Our goal was to identify and resolve any final technical issues and obtain further feedback on LEAP members' experiences in reviewing the sessions.

Methods

Overview

LEAP consultation and online training program development took place from September 2022 to December 2023 (Figure 1). There were 4 phases of LEAP involvement: developing and refining session content (phase 1), exploring the feasibility of completing the training and obtaining feedback on modified content (phase 2), developing UX design (phase 3), and testing the online program prototype (phase 4). The highly iterative participatory design process enabled continuous feedback from LEAP members during all phases of program development.

LEAP members were consulted online. New LEAP members joined at the beginning of phases 1, 2, and 4. This was to ensure we had diverse perspectives and insights from different groups of young people, enriching the development process. Phase 3 LEAP members were a subset of those from phase 1. We invited this subgroup to provide feedback on the UX design of the online platform in phase 3 on the basis that their strong familiarity with the session content would enrich their feedback during the UX design phase.

Figure 1. Phases of development of the multisession online training program. UX: user experience.

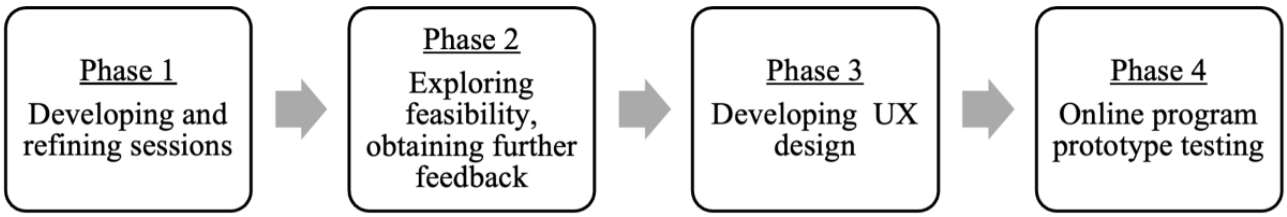

Ethical Considerations

The research project was approved by King's College Research Ethics Committee (HR-19/20 - 14855). LEAP members were consulted to co-design the Shift Focus online training program, and as such, were not research participants per se; therefore, they did not need to provide informed consent. Nonetheless, LEAP members fully understood their role in the development of the project and the reason that they were being consulted and asked to provide feedback; hence, they were fully informed regarding their involvement. The information LEAP members provided is stored in an anonymized form on King's College London servers. LEAP members were reimbursed £20 (US$ 22.55) per hour for their time.

Phase 1: Developing Sessions

Methods

Ten LEAP members (aged 16 - 25 y; female n=8, male n=2) who self-identified as worriers were recruited from Child and Adolescent Mental Health Services (CAMHS) in South London and Maudsley NHS Foundation Trust (SLaM) and a Digital Research Advisory Group in Greater Manchester Mental Health NHS Foundation Trust. JS and CRH presented the LEAP opportunity at the Digital Research Advisory Group meeting and asked young people to make contact via email if they were interested in contributing to the project. The opportunity to be part of the LEAP was featured in the SLaM monthly newsletter, which is distributed to SLaM clinicians, who then shared the poster with appropriate service users in local CAMHS services. With consent, the clinicians provided the service users' email addresses to JS, who made contact with potential LEAP members to explain what the role would involve. We also recruited from the general community via posters (eg, placed on local bulletin boards in South London) and social media paid advertisements (eg, posted on Facebook [Meta] and Instagram [Meta]).

Most LEAP members were currently accessing or had previously accessed CAMHS. LEAP members were from a diverse range of ethnic backgrounds (groupings in this paper were based on free text descriptions of their ethnic identity, which were then mapped to Office for National Statistics Census Categories [14][Asian or British Asian background: n=3; Black African background n=1; White background: n=6]). LEAP members completed the Penn State Worry Questionnaire (PSWQ) [15] to indicate their level of worry.

Shift Focus is comprised of 8 brief sessions developed by the research team and programmed on the online survey platform, Qualtrics. JS met with LEAP members individually via Microsoft Teams (Microsoft Corp) and provided a brief overview of the intervention. LEAP members accessed the intervention sessions through a Qualtrics link, shared their screens with JS, and provided feedback on the sessions. LEAP members were given the option of either sharing their thoughts about the sessions as they worked through them or providing

XSL•FO
RenderX

feedback at the end of the session. JS recorded the time it took each LEAP member to complete a given session (mean=10 min and 4 s; SD=13.7 s). LEAP members answered questions on the acceptability and feasibility of completing each session (Multimedia Appendix 1). Low scores on the scales indicated poor acceptability or feasibility of use, which was then followed up with detailed discussion.

In parallel with phase 1 and phase 2, we collaborated with a trainee clinical psychologist at Oxford University to create animations to explain key psychological concepts within sessions. We prioritized psychological concepts that were difficult to understand, as well as concepts that LEAP members particularly recommended that we represent in animation form. LEAP members provided ideas for the visuals and feedback on drafts of the animations before they were finalized.

Results

Overview

Table 1 presents examples of some of the feedback obtained from LEAP members in phase 1. We note whether feedback was implemented; when it was not implemented, we provide our rationale. This feedback led to major session revisions and requests for additional LEAP input. Feasible changes were implemented before phase 2.

Key learnings from phase 1 LEAP consultations were that the online training program should: (1) be simple and aesthetically pleasing, (2) be personalized, (3) support all learning styles, (4) track progress, (5) provide a sense of community, (6) provide guidance on postsession activities, and (7) have a clear rationale, aims, and objectives. LEAP members also suggested changes to the sessions to make them more accessible.

Table . Examples of feedback provided by LEAP[a] members during phase 1.

LEAP members	Feedback	Implementation
LEAP member 1	"Recap would be great because even if the sessions are not too long it still feels like a lot of information to take in. This would ensure I don't feel overwhelmed with the new information."	Yes (January 30, 2023). Recaps were implemented throughout sessions.
LEAP member 1	"It would be good to have more categories added to the worry themes. Sometimes people have very specific worries, so would be good to have fewer generic worries. Other themes could include anxieties about leaving the house, worries about the future or how successful their career is (there is lots of pressure on young people to achieve highly)."	Yes (March 3, 2023). Feedback added to session content and noted on UX[b] design document for phase 3.
LEAP member 2	"Maybe try to be more specific about identifying certain situations where we are worrying less. For example, if I'm at school, I could be prompted to think about times when I am worrying less. If it is more general in the session content, I might not think about it."	No (March 8, 2023). JS discussed this with the research team. While we understood the motivation for the suggestion, we had concerns about priming an episode of worry. For example, if we remind young people when they are not worrying, they might start to worry. Therefore, we opted to leave it up to young people to notice times when they are not worrying in their daily life, and allowed them to add activities to the online platform if they chose.

[a]LEAP: Lived Experience Advisory Panel.

[b]UX: user experience.

Simple and Aesthetically Pleasing

LEAP members strongly emphasized that the online platform interface should be simple to navigate and that text should be easy to comprehend. It was clear that an online platform that was not easy to use during skills practice would not be used. They also requested an aesthetically pleasing design with distinct colors for different sections of the platform. Accordingly, we condensed the text, simplified psychological constructs, and noted their ideas for the UX design in phase 3.

Personalized to Individual Needs and Preferences

LEAP members indicated that the online platform should be tailored to individual needs and feel personalized. We added features for users to set a personal username on the homepage, reflect on their own worries, and upload personal images for exercises in sessions during phase 3. Budget constraints prevented us from implementing personal profiles and mood tracking features. LEAP members also requested gender-specific voice options for audio exercises, reasoning that some participants may feel more at ease with a specific gender. We noted these items for future directions in a larger study.

LEAP members also requested personal reminder settings (ie, being able to choose when the reminders from the online platform arrived). LEAP members reported that this would increase engagement because reminders would not arrive at inconvenient times, and it would allow them to feel autonomous over their skill development. We implemented this idea during phase 3.

XSL•FO
RenderX

Support All Learning Styles

LEAP members emphasized the importance of the session content being accessible to all, particularly neurodiverse individuals. They suggested that this could be achieved by including multimodal formats (written text, videos, audios, and visuals) to support all learning styles. We decided to create animations to explain key psychological concepts, add bold text for visual learners, and use an audio format instead of text format when possible. We also added optional recap animations at the start of sessions to help consolidate key psychoeducation concepts. LEAP members also requested an audio book feature and the ability to review previous session content, however, this could not be implemented due to budget constraints.

The repetitive nature of exercises within sessions was identified as a significant barrier to engagement. To address this, we minimized identical content and created an animation to explain why repetition of session exercises was important. We also improved session pacing to help users absorb the material, reduce the repetitiveness of exercises, and avoid overwhelming with session content. We introduced a 48-hour break between sessions 4, 5, 6, and 7. Users were asked to practice the skills from sessions during these intervals to embed the skills they had learned.

Tracks Progress

Tracking progress was identified as essential for engagement. We created a session tracking page with a clear timeline, added progress bars, and end-of-session summaries that outlined how the future session would build upon the previous session. Several progress tracking features were out of scope. Most LEAP members requested a reward system within a personal profile on the online platform (eg, collecting points or rewards if they engaged with the mobile app regularly). This feature was noted as essential for a larger study.

Provides a Sense of Community and Connection to Others

LEAP members emphasized a need for a sense of community on the online platform and requested a space to engage with other young people who experience anxiety. Due to limited capacity for moderation, the addition of a community section was not feasible in this study. With permission, we added examples based on LEAP members' experiences within sessions to validate their experience.

LEAP members requested personal audio exercises for each session, referencing other platforms on which familiar figures provided friendly introductions at the beginning of sessions to create a welcoming atmosphere. They also noted that this could reduce the repetitiveness of exercises across sessions. Based on this feedback, JS provided tailored introductions for each session's audio exercises.

LEAP members also sought validation for their progress and acknowledgment that difficulty working through the content and exercises was normal. To address this, we provided validation and normalized challenges through written text and audio files in the sessions.

Guidance on Postsession Activities

LEAP members requested guidance on postsession activities, raising concerns that they may forget to resume their daily tasks after a session by staying on the online platform or continuing to worry. In response, we created an interactive activity panel at the end of sessions during phase 3, which encouraged users to choose an engaging activity to complete after the session to help them return to daily life and step away from their electronic devices.

Clear Rationale, Aims, and Objectives

LEAP members requested a more detailed rationale for the session content, including clear aims and objectives. In response, we added additional psychoeducation, a detailed rationale for the skills taught, and more information about research on the topic of managing worry.

Phase 2: Exploring Feasibility of Completing the Training and Obtaining Feedback on Modified Content

Methods

Ten new LEAP members aged 16 - 25 years (female n=16 and male n=4) were recruited through the King's College London Research Circular Advertisement. LEAP members self-selected by completing a screening questionnaire (eligible if they scored ≤28 on the PSWQ) and completed a study interest call with a member of the research team. LEAP members were from a diverse range of ethnic backgrounds (Asian or Asian British background n=6, Black African background n=1, White background n=10, mixed or multiple ethic background n=1, other ethnic background n=1, and unknown n=1).

LEAP members were given access to the Shift Focus sessions and asked to complete them within 4 weeks. The intervention was comprised of eight 10-minute sessions delivered via Qualtrics, which participants accessed via their mobile phones. JS manually granted session access and sent text message reminders to simulate future platform notifications. LEAP members also participated in online exit interviews to provide feedback on the sessions.

Results

Overview

The key messages from LEAP consultations in phase 2 was that the program should (1) provide guidance on how to apply the skills to daily life, and further, that we should: (2) reframe homework tasks, (3) adapt the personal reminder settings and (4) make further tweaks to the Shift Focus techniques. Most feedback was implemented, except in cases where research budget constraints meant that the suggested changes were not feasible.

Guidance on How to Apply the Skills to Daily Life

LEAP members reported needing additional support on how to apply session skills in daily life, especially in nonquiet spaces such as public transport. In response, we created a ninth and

XSL•FO
RenderX

final session to provide guidance on how to apply the skills in everyday situations.

Reframe Homework Tasks

Most LEAP members understood the purpose of homework exercises between sessions. Some LEAP members completed the homework exercises while others did not. A few LEAP members reported that they found it difficult to find the motivation to engage with the homework exercises. Our goal was to prevent users from viewing homework as a chore. We carefully chose our language and introduced new terminology to replace the word "homework," and designed a new part of the online training platform called the Shift Focus Training Hub. Users were asked to visit the Hub to complete 2 different between-session exercises. LEAP members also reported that they found it less motivating to complete homework exercises because they could not track their homework completion. It was clear that integrating a reward system and progress tracker into the Shift Focus Training Hub would be essential to increase engagement in a future study.

Adapt Personal Reminder Settings

We discussed the notion of "alarm fatigue" through over-notification with software developers because this was identified as a barrier to engagement in phase 2. In response, users could choose reminder frequency (1 - 3 daily) and timing (8 AM, 3 PM, and 8 PM) for homework exercises. This idea was implemented during phase 3.

Modifications to the Shift Focus Techniques

We made several changes to the Shift Focus techniques following LEAP feedback in phase 2. One technique involved asking LEAP members to identify activities that they felt were particularly absorbing, and thus less impacted by worry. They appreciated reflecting on this and requested more integration of this concept into the sessions. We also iteratively adapted the final session content and significantly modified the audio exercises based on feedback. LEAP members also helped identify aspects of the techniques that could temporarily increase worry, leading to content adjustments to minimize this risk.

In addition to these new recommendations, LEAP members reiterated ideas expressed by phase 1 LEAP members, namely: the online platform must be personalized to individual needs and preferences (including personal reminder settings) and include progress tracking features. This reinforced the importance of implementing these ideas during phase 3, within the constraints of our research budget.

Phase 3: Developing UX Design

Methods

Six LEAP members aged 16 - 25 years (female n=4, male n=2) from phase 1 (see above for recruitment approach) provided further feedback in phase 3. LEAP members were from an Asian or Asian British background (n=3) and a White background (n=3).

The UX design team created wireframes (Figure 2) following the user journey mapping workshop with CRH and JS. Wireframing is a process of creating black-and-white images of the online platform on screens and arranging them to determine how users would navigate the platform. It focuses on the functionality and layout of the online platform.

JS presented the wireframes to LEAP members individually via Microsoft Teams. JS explained a scenario in which the online platform was recommended to them by a friend, and they were using it for the first time. Using the think-aloud approach [16], LEAP members shared their screens, navigated through the wireframes, and voiced their thoughts. JS noted their feedback and identified navigation issues. The responses were then collated, and feasible feedback was incorporated by the UX designers before moving onto the user interface (UI) design phase.

The designers presented the initial UI designs (look and feel) of the online platform, including functionality, layout, and color schemes. JS repeated the process with LEAP members, gathering feedback on the UI design. The designers incorporated feasible feedback and then presented final designs to the research team for the last amendments.

XSL•FO
RenderX

Figure 2. Examples of wireframes from the online training program.

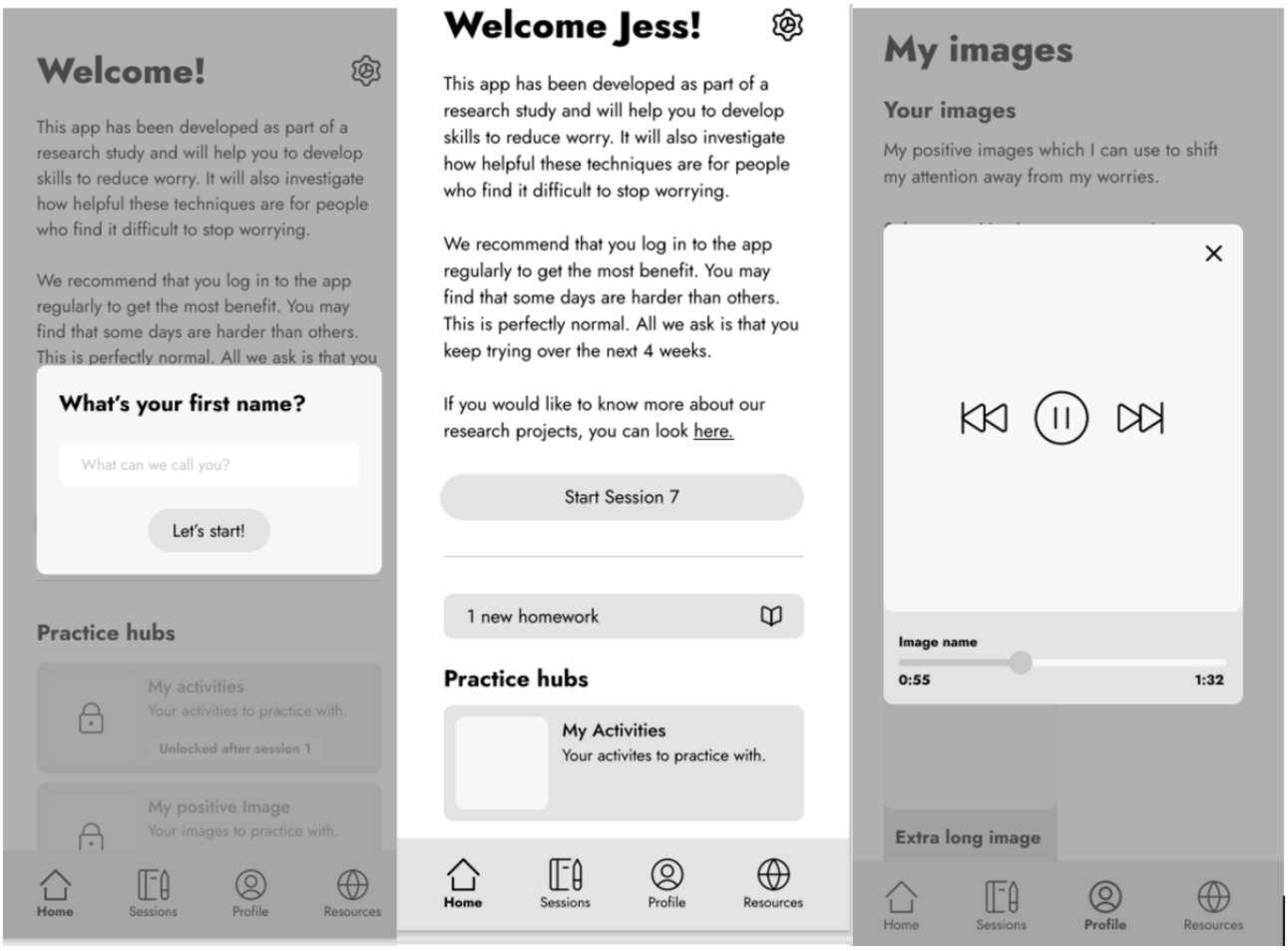

Results

Overview

Key messages from LEAP consultations in phase 3 were that the online platform should be (1) simple to navigate, (2) interactive and engaging, and (3) track progress. LEAP members requested that the platform should have: (4) accessibility settings and (5) provide a sense of community. LEAP members additionally (6) provided extensive feedback on the look and feel of the online platform and (7) noted key changes to session content.

Simple to Navigate

A think-aloud approach was crucial to determine how LEAP members found navigating the online platform. LEAP members reported areas that they found difficult to navigate or confusing. It was clear that an online platform that was simple to use and navigate was essential for engagement.

Interactive and Engaging

LEAP members requested more interactive and engaging features on the platform. For example, LEAP members liked the reminder buttons, which were designed with different colors and icons for morning (yellow with a sun icon) and evening (dark blue with a moon icon) to enhance user engagement with personal reminder settings. LEAP members also liked being able to upload personal images for tasks that could be seen as interactive tiles which they could flip around to add their own descriptions of what it meant to them.

Track Progress

LEAP members requested a progress tracking bar within sessions to motivate users to finish their sessions. LEAP members also wanted to track how many times they had completed exercises on the online platform. We added a progress bar and session timeline page, but could not implement exercise tracking due to budget constraints. Progress tracking and a reward system were noted as essential for a larger study.

Accessibility Settings

LEAP members reinforced the need for the online platform to be accessible to all and suggested including features such as adjustable text size and dark or light modes so that participants could reduce the brightness if needed. Although these suggestions were noted for a larger study, they could not be implemented due to budget constraints. In phase 1, LEAP members requested transcripts for audio clips to improve accessibility. However, in phase 3, LEAP members found this particular multimodal format unhelpful and requested that the transcripts be removed, which we did.

Sense of Community

LEAP members requested a space on the platform to engage with other young people who experience worry. As noted previously, we did not have the capacity for a moderator, so

XSL•FO
RenderX

this could not be implemented. As an alternative, we proposed a community section that reported the number of young people who had completed the Shift Focus program on a given day. LEAP members reported that this could create a sense of competition, resulting in them worrying about how quickly they should be completing the program. Consequently, we decided not to implement tracking of program completion in a community section.

Look and Feel

LEAP members provided extensive feedback on the color schemes for the online platform. LEAP members did not find our first color scheme for the platform aesthetically pleasing. For example, they found the color white too bright, and they did not like the orange-colored reminder boxes because it felt was like an amber warning (ie, made them feel as though they had done something wrong). LEAP members also requested clearer headers and page structure. In response, we revised the colors and page layout (Figures 3 and 4). Budget constraints prevented us from allowing users to choose their own background color; this feature was noted for inclusion in a larger future study.

Figure 3. Images comparing a wireframe, preliminary design, and final design of the homepage that incorporated LEAP feedback. LEAP: Lived Experience Advisory Panel.

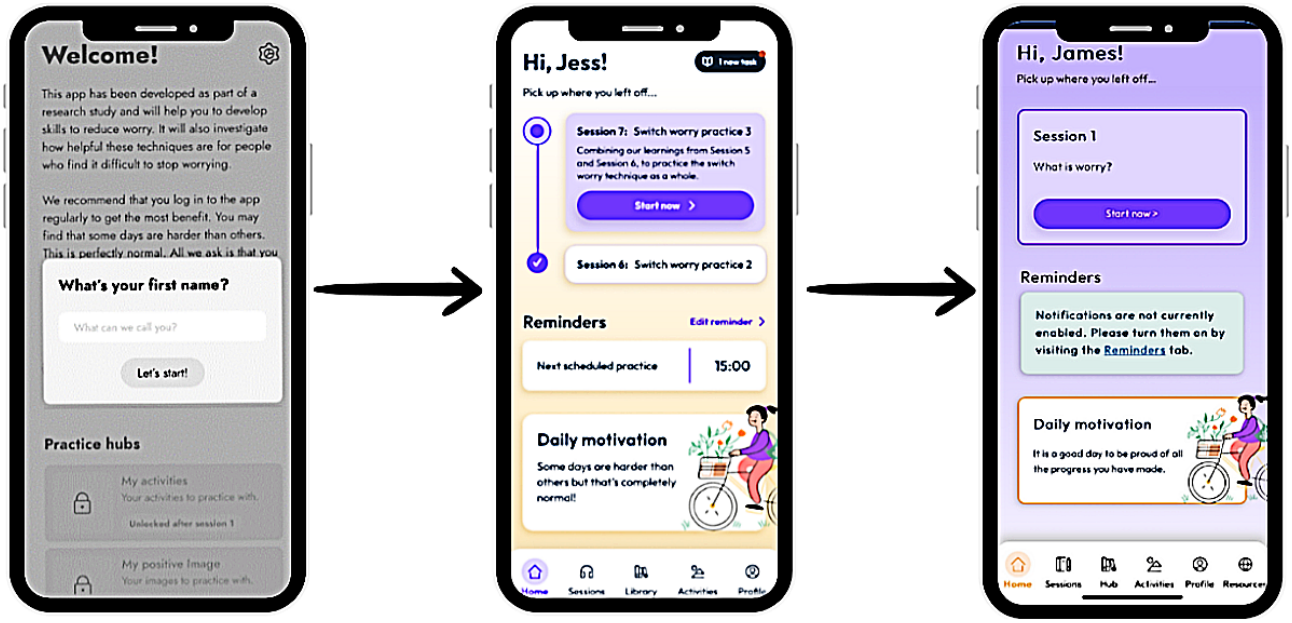

XSL•FO
RenderX

Figure 4. Images comparing an original session design and a final design that incorporated LEAP feedback. LEAP: Lived Experience Advisory Panel.

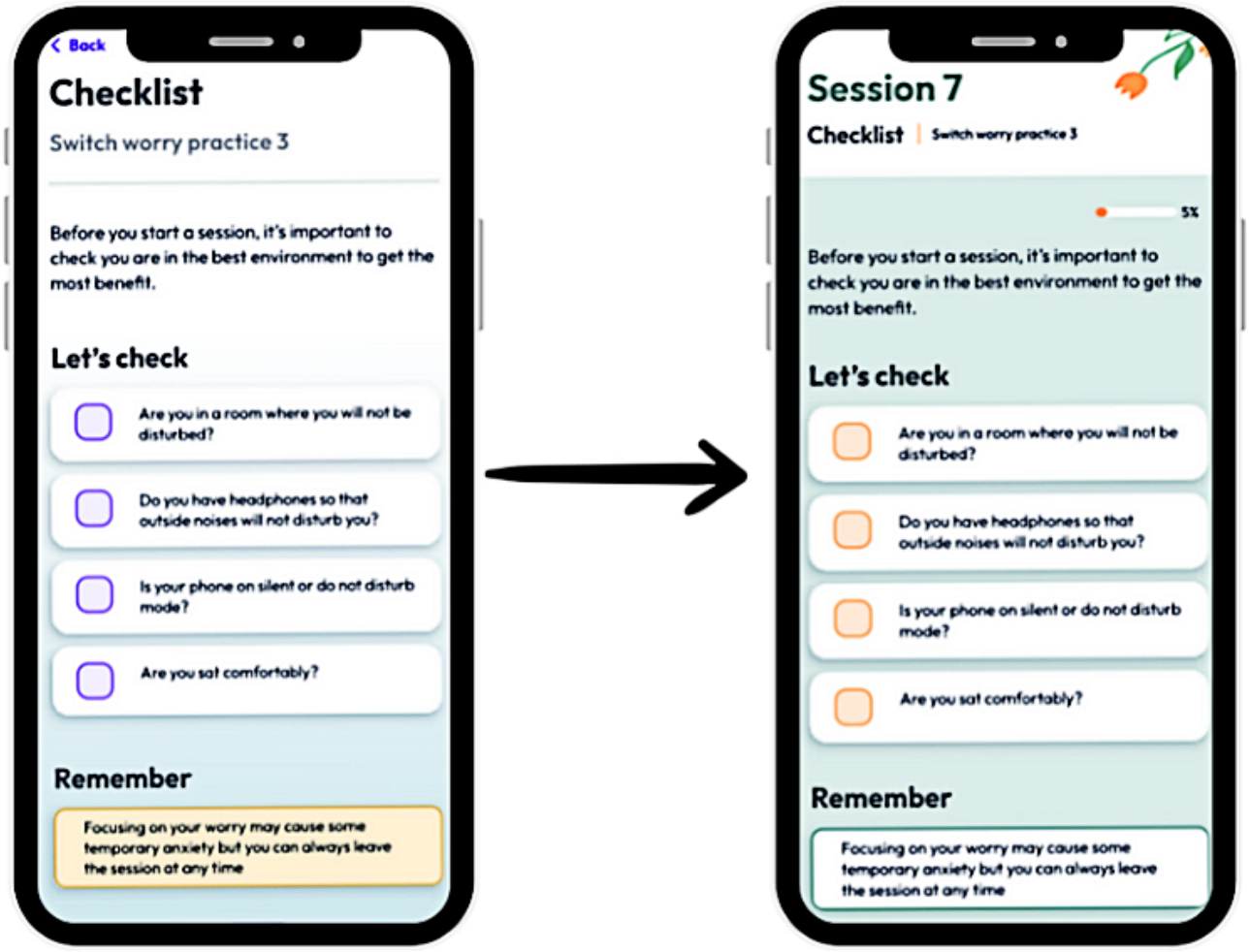

Key Changes to Session Content

LEAP members highlighted session content they did not like or found confusing. For example, LEAP members reported that overuse of the phrase "well done!" felt patronizing. We revised the session content in response to this feedback.

Phase 4: Online Training Program Testing

Methods

Eight new LEAP members aged 16 - 25 years (female n=6, male n=2) were recruited through the Shift Focus social media page on Instagram to provide feedback on the online training program prototype. All LEAP members self-selected as being worriers from the social media advertisement. The group was predominantly White (n=7) with 1 member from another ethnic background (n=1). In phase 4, we used the Repetitive Thinking Questionnaire to measure repetitive negative thinking, as the PSWQ is for those aged 18 years and older and so does not cover the full age range for those who will be ultimately offered the online training program. The Repetitive Thinking Questionnaire is validated in both adult and adolescent samples [17,18].

LEAP members accessed the Shift Focus sessions over 4 weeks. The intervention was comprised of nine 10-minute sessions delivered via the online platform. LEAP members had the option of setting personal reminders for homework exercises and received push notifications for new sessions or inactivity. LEAP members were interviewed after completing the four weeks of the online training program to identify any technical problems or engagement issues before study launch. Feedback that could not be implemented due to time constraints was noted for future studies.

Results

Overview

The key messages from LEAP consultations in phase 4 were: (1) technical problems hindered engagement, (2) users may have data security concerns, and (3) the need for a platform tutorial. LEAP members also suggested (4) further changes to session content in a future study.

Technical Problems

LEAP consultations in phase 4 identified technical problems with the online platform, including malfunctioning buttons and missed reminders. Technical glitches were a barrier to engagement as LEAP members who experienced technical problems needed more reminders from the research team to complete sessions. Piloting the prototype to identify and fix technical issues before study launch was essential for this reason.

XSL•FO
RenderX

Data Security Concerns

A LEAP member reported that it felt strange uploading their own personal images to the online platform and highlighted the importance of data security concerns among users. The research team subsequently addressed this by providing clear information about data security and storage in the study information sheet. In any future study, we will explicitly highlight these details on the online platform to alleviate such concerns.

Platform Tutorial

LEAP members found it confusing that they had access to all sections of the online platform from the outset, particularly given that some areas of the platform were not needed until later sessions. LEAP members suggested there was a video tutorial on platform navigation at the start, rather than introducing areas of the platform as they worked through sessions. This will be implemented in a future study.

Session Content Changes

LEAP members also proposed further changes we could make to the session content in a future study. This included adding more guidance on how to apply the skills to daily life, a clearer emphasis on the long-term benefits of using the skills, more varied sessions with less repetition, and additional resources for independent learning (eg, signposting articles). This was invaluable feedback that will be implemented in a future study.

Discussion

Principal Findings

The Shift Focus online platform was developed, designed, and tested with psychologists, software developers, and young people with lived experience of worry. We aimed to engage with young people with lived experience of worry, to refine and further develop a new online training program designed to help young people shift focus away from worry. We adopted a highly iterative participatory design process across 4 phases—an approach that ensured LEAP members were integral to all stages of the Shift Focus platform development, and that the online platform was tailored and personalized to young people's needs and preferences.

In phase 1, we learned that the online platform needed to be simple to use, interactive, aesthetically pleasing, and personalized to individual needs and preferences. This echoes existing research which emphasizes the importance of aesthetics [19], simplicity [20], interactivity [21], and personalization [22] when designing DHI's. It was also important that the platform was accessible, tracked progress, and provided a sense of community with others with shared lived experience of worry.

In phase 2, we learned that the platform needed to provide further guidance on how to apply the Shift Focus techniques to daily life and have personalized reminder settings. LEAP members also suggested (in multiple phases) that we should include enhanced progress tracking features and a reward system in a future iteration of the Shift Focus online training program to promote engagement. This is consistent with findings that suggest that gamification, such as points or incentives, can motivate consistent use and adherence to DHIs [23] .

In phase 3, LEAP members provided invaluable feedback on the appearance and functionality of the online platform, which played a crucial role in shaping the "look and feel" of the online training platform. Similarly to phase 1, LEAP members re-emphasized the importance of a community section on the platform to connect with other young people with shared lived experience of worry. Clearly, a sense of belonging and social connection is a crucial consideration when developing DHIs for young people.

In phase 4, we identified that technical problems with the online platform were barriers to engagement. Technical issues have been widely reported as a barrier to engagement in DHI research [24] . Additionally, data security concerns with the online training platform were raised by a LEAP member. This highlights the importance of clearly outlining how data is stored to DHI users to establish user trust and engagement.

We put the voice of lived experience at the center of the project by ensuring key messages from LEAP were incorporated from phases 1-4. LEAP members were very effective at helping the research team implement gentle shifts in language and refine explanations of psychological concepts to make them more accessible to our target population. Our LEAP members ranged in age from 16 to 25 years, and we observed significant differences in the preferences of the younger (16-18 years) and older (23-25 years) members. For example, the younger members preferred simplified text with bullet points, while the older members expressed a preference for more reflective questions and detailed session rationales. This large age gap was challenging at times due to the different preferences that were sometimes diametrically opposed. In the future, we will explore the possibility of developing age-specific versions of the online platform to address these varied needs and preferences.

Power differentials and research hierarchies have been reported as barriers to working collaboratively and developing meaningful working relationships with LEAP members [25,26]. We found that encouraging LEAP members to be open and critical of our research was important to gain the most from their insights and reduce power dynamics. We regularly emphasized to LEAP members that they brought unique expertise to the team and how valuable it was for them to challenge our research ideas and ways of thinking. We also explained how any negative feedback provided was very helpful because it highlighted our blind spots and areas in which we could improve. Our goal was to create a space for LEAP members to be critical of our work and session content, to enable them to feel that their expertise was valued.

LEAP initiatives have faced criticism for being too narrow and tokenistic [10,11], leading to calls for greater empowerment of LEAP members within research settings [12]. Therefore, it is essential for research teams to implement research processes that reduce the possibility of tokenism. We created a table to log our rationale for not implementing ideas suggested by LEAP members in phase 1. This kept the research team accountable and ensured that we did not just implement suggestions from LEAP consultations based on our own personal preferences.

XSL•FO
RenderX

Additionally, it was important that we were transparent with LEAP members when we could not implement an idea due to financial constraints within our research project. Managing expectations was key, and being clear about what was out of scope was important to ensure that LEAP members did not think their ideas were ignored. If we were unable to implement an idea, we worked collaboratively with the LEAP member to consider possible ways we could address the issue raised, even in part, within the scope of our budget. LEAP members valued our honesty and had a stronger understanding of the research team's decision-making processes, which we believe increased engagement.

It was important to ask LEAP members whether they felt like valued members of the research team to ensure that we learned from the process and could change our practices to be more inclusive if necessary. Focusing only on the impact LEAP has on research, without considering the potential benefits or costs the work has on the LEAP members involved, overlooks an essential part of the process [27,28]. Therefore, we sought feedback on this via informal discussions and anonymous feedback questionnaires.

During phase 3, the UX designers and the LEAP members generated many excellent ideas for the design of the online platform. This sometimes resulted in competing ideas, which made decision-making during the design phase challenging. We found it helpful to clarify with LEAP members why they favored certain UX design features over others, rather than simply being informed of their likes and dislikes. This ensured the design of the online platform was flexible for UX designers, but importantly, also met the needs of young people.

This study has limitations. We invited any young people aged 16‑25 years to participate if they classified themselves as being a high worrier. With the exception of phase 2, we did not use a cutoff on a self-report measure of worry to determine eligibility. Furthermore, in phases 1 and 3 (same individuals in both phases), six LEAP members had moderate but not high levels of worry, and in phase 4, one LEAP member did not have high levels of repetitive thinking on the standardized questionnaire. As a result, it is possible that feedback from these individuals did not reflect the views of high worriers who are the target population. That said, they were aware that they were providing feedback for an online training program designed to help high worriers. Additionally, JS created a table to document all the feedback from LEAP members in phase 1 and whether the feedback was implemented. This process was not used in phases 2-4. Future studies should plan to generate feedback implementation tables consistently across all phases. Furthermore, we did not present the phase 1 feedback implementation table to the LEAP members, and we suggest that this be done in future work to enhance their understanding of our decision-making processes.

Conclusion

LEAP members brought unique expertise grounded in their lived experience of worry. LEAP members were highly valued members of the team and were encouraged to give critical feedback at all stages. A highly iterative participatory co-design process enabled continuous feedback from LEAP members throughout, ensuring their input was meaningful and key messages and ideas were incorporated into the online platform. Study exit interviews from the experiment that will use the Shift Focus online program delivered via a mobile app will inform us as to whether the process led to us developing an engaging online training platform that worked well for young people (JS et al, unpublished data, 2025). Our Shift Focus study is one example of how having meaningful LEAP involvement in the development of DHIs benefits the research and intervention greatly.

Acknowledgments

The authors would like to acknowledge all the Lived Experience Advisory Panel (LEAP) members across the United Kingdom who have contributed their very valuable expertise and feedback during the development of the Shift Focus online platform. The research would not have been possible without them. We also extend our thanks to Dr Aika Hui, who developed the animations for the sessions, and Hayley O'Sullivan, MSc, for her contributions to the original session content. This paper represents independent research partly funded by Mental Health and Quality of Life (MQ) Transforming Mental Health (MQF20/24) and the National Institute for Health and Care Research (NIHR) Maudsley Biomedical Research Centre at South London and Maudsley National Health Service (NHS) Foundation Trust and King's College London. The views expressed are those of the authors and not necessarily those of the NIHR, MQ, or the Department of Health and Social Care.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional materials—questions asked in phase 1 regarding the content of each session.
[PDF File, 260 KB - formative_v9i1e66461_app1.pdf]

References

1. Gonçalves DC, Byrne GJ. Who worries most? Worry prevalence and patterns across the lifespan. Int J Geriatr Psychiatry 2013 Jan;28(1):41-49. [doi: 10.1002/gps.3788] [Medline: 22374912]

XSL•FO
RenderX

2. Rowa K, Waechter S, Hood HK, Antony MM. Generalized anxiety disorder. In: Psychopathology: History, Diagnosis, and Empirical Foundations, 3rd edition: John Wiley & Sons, Inc; 2017:149-186. [doi: 10.1002/9781394258949]
3. Hirsch CR, Mathews A. A cognitive model of pathological worry. Behav Res Ther 2012 Oct;50(10):636-646. [doi: 10.1016/j.brat.2012.06.007] [Medline: 22863541]
4. Eagleson C, Hayes S, Mathews A, Perman G, Hirsch CR. The power of positive thinking: pathological worry is reduced by thought replacement in generalized anxiety disorder. Behav Res Ther 2016 Mar;78:13-18. [doi: 10.1016/j.brat.2015.12.017] [Medline: 26802793]
5. Sawjani Y, Spyropoulou S, Bardi J, Bougas G, Moulds M, Hirsch CR. Training Positive Mental Imagery to Reduce Worry: Association of Cognitive Bias Modification; 2024.
6. Lipschitz JM, Pike CK, Hogan TP, Murphy SA, Burdick KE. The engagement problem: a review of engagement with digital mental health interventions and recommendations for a path forward. Curr Treat Options Psychiatry 2023 Sep;10(3):119-135. [doi: 10.1007/s40501-023-00297-3] [Medline: 38390026]
7. Baumel A, Muench F, Edan S, Kane JM. Objective user engagement with mental health apps: systematic search and panel-based usage analysis. J Med Internet Res 2019 Sep 25;21(9):e14567. [doi: 10.2196/14567] [Medline: 31573916]
8. Biddle MSY, Gibson A, Evans D. Attitudes and approaches to patient and public involvement across Europe: a systematic review. Health Soc Care Community 2021 Jan;29(1):18-27. [doi: 10.1111/hsc.13111] [Medline: 32705752]
9. Public involvement. NHS. URL: https://www.hra.nhs.uk/planning-and-improving-research/best-practice/public-involvement/ [accessed 2025-04-26]
10. Reynolds J, Ogden M, Beresford R. Conceptualising and constructing “diversity” through experiences of public and patient involvement in health research. Res Involv Engagem 2021;7:53. [doi: 10.1186/s40900-021-00296-9] [Medline: 34294162]
11. Ocloo J, Matthews R. From tokenism to empowerment: progressing patient and public involvement in healthcare improvement. BMJ Qual Saf 2016 Aug;25(8):626-632. [doi: 10.1136/bmjqs-2015-004839]
12. Russell J, Greenhalgh T, Taylor M. Patient and Public Involvement in NIHR Research 2006–2019: Policy Intentions, Progress and Themes: National Institute for Health Research; 2019.
13. Brett J, Staniszewska S, Mockford C, et al. A systematic review of the impact of patient and public involvement on service users, researchers and communities. Patient 2014;7(4):387-395. [doi: 10.1007/s40271-014-0065-0] [Medline: 25034612]
14. List of ethnic groups. GOV.UK. URL: https://www.ethnicity-facts-figures.service.gov.uk/style-guide/ethnic-groups [accessed 2025-04-26]
15. Meyer TJ, Miller ML, Metzger RL, Borkovec TD. Development and validation of the Penn State Worry Questionnaire. Behav Res Ther 1990;28(6):487-495. [doi: 10.1016/0005-7967(90)90135-6] [Medline: 2076086]
16. Someren M, Barnard YF, Sandberg J. The Think Aloud Method: A Practical Approach to Modelling Cognitive: AcademicPress; 1994, Vol. 11.
17. McEvoy PM, Salmon K, Hyett MP, et al. Repetitive negative thinking as a transdiagnostic predictor of depression and anxiety symptoms in adolescents. Assessment 2019 Mar;26(2):324-335. [doi: 10.1177/1073191117693923] [Medline: 29214855]
18. Spinhoven P, van Hemert AM, Penninx BW. Repetitive negative thinking as a predictor of depression and anxiety: a longitudinal cohort study. J Affect Disord 2018 Dec 1;241:216-225. [doi: 10.1016/j.jad.2018.08.037] [Medline: 30138805]
19. Denison-Day JL, Muir S, Newell C, Appleton KM. The role of aesthetics in intentions to use digital health interventions. PLOS Digit Health 2023 Jun;2(6):e0000274. [doi: 10.1371/journal.pdig.0000274] [Medline: 37347727]
20. Bakker D, Kazantzis N, Rickwood D, Rickard N. Mental health smartphone apps: review and evidence-based recommendations for future developments. JMIR Ment Health 2016 Mar 1;3(1):e7. [doi: 10.2196/mental.4984] [Medline: 26932350]
21. Alqahtani F, Orji R. Insights from user reviews to improve mental health apps. Health Informatics J 2020 Sep;26(3):2042-2066. [doi: 10.1177/1460458219896492] [Medline: 31920160]
22. Matthews P, Rhodes-Maquaire C. Personalisation and recommendation for mental health apps: a scoping review. Behav Inf Technol 2024:1-16. [doi: 10.1080/0144929X.2024.2356630]
23. Saleem M, Kühne L, De Santis KK, Christianson L, Brand T, Busse H. Understanding engagement strategies in digital interventions for mental health promotion: scoping review. JMIR Ment Health 2021 Dec 20;8(12):e30000. [doi: 10.2196/30000] [Medline: 34931995]
24. Borghouts J, Eikey E, Mark G, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. J Med Internet Res 2021 Mar 24;23(3):e24387. [doi: 10.2196/24387] [Medline: 33759801]
25. Egid BR, Roura M, Aktar B, et al. “You want to deal with power while riding on power”: global perspectives on power in participatory health research and co-production approaches. BMJ Glob Health 2021 Nov;6(11):e006978. [doi: 10.1136/bmjgh-2021-006978] [Medline: 34764147]
26. Allen D, Cree L, Dawson P, et al. Exploring patient and public involvement (PPI) and co-production approaches in mental health research: learning from the PARTNERS2 research programme. Res Involv Engagem 2020 Dec;6(1):1-11. [doi: 10.1186/s40900-020-00224-3]
27. Staley K, Barron D. Learning as an outcome of involvement in research: what are the implications for practice, reporting and evaluation? Res Involv Engagem 2019;5(1):14. [doi: 10.1186/s40900-019-0147-1] [Medline: 30915234]

XSL•FO
RenderX

28. Jackson T, Pinnock H, Liew SM, et al. Patient and public involvement in research: from tokenistic box ticking to valued team members. BMC Med 2020 Apr 13;18(1):79. [doi: 10.1186/s12916-020-01544-7] [Medline: 32279658]

Abbreviations

CAMHS: Child and Adolescent Mental Health Services
DHI: digital health intervention
GAD: generalized anxiety disorder
LEAP: Lived Experience Advisory Panel
NHS: National Health Service
PPI: patient and public involvement
PSWQ: Penn State Worry Questionnaire
SLaM: South London and Maudsley NHS Foundation Trust
UI: user interface
UX: user experience

Edited by A Mavragani; submitted 26.09.24; peer-reviewed by B Cliffe, E Nsubuga; revised version received 21.02.25; accepted 02.03.25; published 21.05.25.

Please cite as:
Steward J, Moulds ML, Hirsch CR
Co-Designed Online Training Program for Worry Management: The Role of Young People With Lived Experience of Worry in Program Development
JMIR Form Res 2025;9:e66461
URL: https://formative.jmir.org/2025/1/e66461
doi:10.2196/66461

© Jessica Steward, Michelle L Moulds, Colette R Hirsch. Originally published in JMIR Formative Research (https://formative.jmir.org), 21.5.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Necessity of Regulating Drinking Scenes on Social Media Platforms Focusing on YouTube Sulbang Videos: Public Opinion From Surveys and YouTube Content Analysis

HyoRim Ju[1,2], MD; HyeWon Lee[3], RN, BSN; Juyoung Choi[3], RN, BSN; EunKyo Kang[3,4,5], MD

[1]Department of Family Medicine, Dankook University Hospital, Cheonan, Republic of Korea

[2]Department of Family Medicine, College of Medicine, Dankook University, Cheonan, Republic of Korea

[3]National Cancer Control Institute, National Cancer Center, 323 Ilsan-ro, Ilsandong-gu, Gyeonggi-do, Goyang, Republic of Korea

[4]National Cancer Center Graduate School of Cancer Science and Policy, National Cancer Center, Goyang, Republic of Korea

[5]Department of Family Medicine, National Cancer Center, Goyang, Republic of Korea

Corresponding Author:
EunKyo Kang, MD
National Cancer Control Institute, National Cancer Center, 323 Ilsan-ro, Ilsandong-gu, Gyeonggi-do, Goyang, Republic of Korea

Abstract

Background: Alcohol consumption is a major risk factor for diseases and social burdens worldwide. Despite this, depictions of alcohol use continue to rise across various social media platforms, increasing concerns about their potential impact, particularly on adolescents. While some guidelines exist to regulate alcohol portrayals in media, they remain largely advisory and lack legal enforcement. As alcohol-related content becomes more widespread on social media, the need for stronger regulatory measures is growing.

Objective: This study aimed to analyze the content of sulbang (broadcasts featuring alcohol consumption) on YouTube and to assess public opinions regarding the regulation of alcohol-related broadcasts on social media platforms such as YouTube.

Methods: To evaluate public attitudes toward appropriate regulations on alcohol depictions in web-based media, a survey was conducted with 1500 adults (aged 20 - 74 years) residing in South Korea. Participants were recruited through stratified multistage sampling, with a 21.8% (n=1500) response rate from 6880 invitations. The survey included Likert-scale and rank-ordered questions, with reliability assessed using Cronbach α. Additionally, a content analysis of 318 YouTube (sulbang) videos was conducted based on the Korean government's media alcohol scene guidelines. Two trained coders independently analyzed the videos, achieving high intercoder reliability (Cohen κ=0.92).

Results: This study found that exposure to sulbang content was significantly higher among individuals with higher education levels (n=33, 26.2% graduate degree holders), lower income groups (P<.001), and women. Younger individuals and heavy drinkers were also more likely to engage with such content, with heavy drinkers showing a significantly higher likelihood (P<.001). Regarding public opinion, 83.1% (n=1247) of respondents supported some form of regulation on sulbang content. However, heavy drinkers were less inclined to agree (coefficient: −0.3652; P<.001). Age was positively associated with stronger support for regulation (coefficient: 0.21984; P<.001), while women were significantly more likely than men to advocate for stricter restrictions (coefficient: 0.37827; P<.001). Exposure frequency also had the strongest correlation with support for regulation (coefficient: 1.0278; P<.001). The analysis of 318 YouTube videos revealed an average Like ratio of 97.9% (range: 32.7 - 100.0), indicating predominantly positive viewer responses, with a median Video Power Index of 939.6 (range: 10.4-84,821.7). Content analysis based on the Media Drinking Scene Guidelines showed that 89.0% (n=283) of the videos glorified drinking, often portraying alcohol as a stress reliever or a source of recovery. Additionally, 92.8% (n=295) of the videos depicted binge drinking or drunkenness, and 27.7% (n=88) of the videos featured celebrities or notable figures consuming alcohol. Furthermore, 42.8% (n=136) of the videos presented distorted drinking norms, such as glorifying high tolerance or linking alcohol to sexual advances. In contrast, only 0.6% (n=2) of the videos were age-restricted, and 31.1% (n=99) included any warning message.

Conclusions: Given the potential influence of alcohol-related content on drinking perceptions and behaviors, regulatory measures should be explored to mitigate possible risks. Strengthening content guidelines and increasing awareness could help address concerns about alcohol-related social media exposure.

(JMIR Form Res 2025;9:e65162) doi:10.2196/65162

XSL•FO
RenderX

KEYWORDS

alcohol; drinking scenes; social media; YouTube; sulbang; drinking scenes of social medial; public health; social media platforms; drinking regulate; alcohol policy

Introduction

Alcohol is a beverage made from ethanol and is one of the most widely used psychoactive substances worldwide. The World Health Organization classifies alcohol as a psychoactive substance due to its ability to alter brain function by affecting cognition, behavior, and consciousness [1]. This classification highlights the potential risks associated with alcohol consumption and underscores the need for public health intervention [2]. However, as of 2019, 43.8% of the global population aged 15 years and older consumed alcohol, with 38.0% of them engaging in heavy or binge drinking (defined as consuming at least 60 grams of alcohol on one or more occasions in the past month). Notably, about one in 5 adolescents aged 15 - 19 years currently consume alcohol [3]. In South Korea, as of 2022, 58.7% of adults and 13.0% of adolescents reported drinking alcohol. This rate decreased during the COVID-19 pandemic but started to increase again in 2022 [4].

Alcohol use is a primary contributor to the global burden of disease and can lead to substantial health repercussions [5,6]. In particular, the link between alcohol and liver disease, such as liver cancer and cirrhosis is well-documented in numerous academic studies [7,8], while alcohol consumption is also associated with various nonliver cancers and heart conditions, including atrial fibrillation [9-11]. Moreover, alcohol use leads to societal problems such as drunk driving and violence, negatively affecting not only the drinkers themselves but also their families and society as a whole [12,13].

However, alcohol is often perceived as a social facilitator and stress reliever, associated with relaxation, social bonding, and celebration. This positive perception contributes to the normalization of alcohol use across various cultures [14]. Such perceptions are reinforced by media portrayals where celebrities are frequently shown consuming alcohol in television shows, movies, and advertisements. These depictions have the potential to normalize and glamorize alcohol consumption, influencing viewers' attitudes and behaviors toward drinking [14-16]. Bandura's Social Learning Theory provides a framework for understanding this phenomenon, suggesting that individuals, particularly adolescents, learn behaviors through observing others, especially when those behaviors are modeled by influential figures or portrayed positively in media. This theory underscores how repeated exposure to such portrayals can reinforce positive perceptions of drinking and normalize alcohol consumption [17,18]. Supporting this, previous research demonstrates that media exposure to drinking scenes is associated with the development of favorable attitudes toward alcohol and increased consumption among adolescents [18-20]. Accordingly, the role of media is emphasized in strategies aimed at reducing harmful levels of alcohol consumption, which include behaviors such as binge drinking, underage drinking, and excessive consumption that increase health risks and societal harm [2]. However, the majority of countries worldwide still do not have effective rules governing depictions of alcohol consumption in media and advertisements [3]. In 2017, South Korea introduced 10 guidelines aimed at minimizing the risks of media exposure to alcohol-related content. These guidelines sought to discourage positive portrayals of drinking and limit depictions of binge drinking. To address the evolving media landscape, including platforms like YouTube, the guidelines were revised and expanded to 12 in 2023 [21,22]. However, these guidelines remain advisory and lack legal enforcement. According to recent surveys by the Ministry of Health and Welfare and the Korea Health Promotion Institute, 66.3% of respondents reported watching alcohol-related scenes on TV or YouTube in the past year, with 27.6% being exposed to such scenes at least once a week [23]. Furthermore, the popularity of sulbang (alcohol-drinking broadcasts) on YouTube, a prominent social media platform, has increased exposure to alcohol-related content, which has become more provocative over time [24,25]. In addition, social media platforms like YouTube lack age restrictions for viewing alcohol-related broadcasts, allowing children and adolescents to easily access these sulbang videos, potentially influencing their positive perceptions of drinking [14-16]. Therefore, in the current context of increasing alcohol-related content on media platforms, there is a heightened need for the introduction of stronger regulatory policies for drinking-related content in media, including legal enforcement measures.

This study aims to analyze sulbang content on YouTube, one of the most prominent social media platforms, and to investigate public opinion on regulating alcohol-related broadcasts on social media. Additionally, it seeks to determine whether current (sulbang) content adheres to existing alcohol broadcasting guidelines and to provide evidence for the necessity of regulatory policies, such as age restrictions, on future alcohol-related videos. By addressing the gap between existing guidelines and the current media landscape, this study highlights the importance of implementing practical regulatory measures.

Methods

Study Design

Our study has two main components:

1. A cross-sectional survey was conducted among 1500 adults aged 20 - 74 years residing in Korea to assess public attitudes toward appropriate levels of restrictions on drinking alcohol depictions in social media platforms.
2. To analyze the prevalence of "sulbang" on YouTube, a content analysis was performed on 318 YouTube videos with over 10,000 views identified through a search using the keyword "sulbang."

Participants and Procedures

We conducted a nationwide web-based survey targeting 1500 adults aged between 20 and 74 years. Our research sampling design was based on a stratified multistage sampling design

XSL•FO
RenderX

using population data categorized by region, age, and sex provided by the National Statistical Office of Korea, thereby ensuring that survey participants represented the entire national population. Those who clicked on the link received an explanation about the study's purpose.

Adequate explanations were provided to all participants, and informed consent was obtained before they participated in the survey. Among those who were invited, 21.8% of the participants consented to the research, resulting in a total of 1500 responses that were used for analysis.

Data Collection and Questionnaire

The sociodemographic factors of survey participants collected in the study included age, gender, education level, household income, marital status, and residential area. Using a structured questionnaire, the participants' frequency of alcohol consumption, the quantity consumed per drinking session, and the frequency of exposure to drinking alcohol scenes in social media, including press and web-based platforms, were assessed over the previous year. Additionally, to assess opinions on appropriate levels of restrictions on alcohol-related social media, we posed the following question: "Do you believe that alcohol scenes on social media platforms such as YouTube should be restricted? If you think it's necessary, to what extent do you think it should be limited?" The response comprised the following options: (1) do not believe any restrictions on viewing are necessary; (2) limitations to a warning level cautioning about scenes featuring alcohol before viewing; (3) restriction for viewing only upon adult certification; (4) restrictions such as pixelation even with adult certification; and (5) restrictions prohibiting the inclusion of drinking alcohol scenes in videos.

Sample Size and Data Analysis

Our target sample size was 1500 participants, which would allow for the description of responses with a 2.5% margin of error at a 95% confidence level. A descriptive analysis of the participants' baseline characteristics as well as each item in the questionnaire was conducted, using frequencies and percentages. In addition, the study analyzed the relationship between sociodemographic factors and the frequency of exposure to scenes depicting alcohol consumption on social media platforms. It also conducted a regression analysis on perceptions of appropriate restriction levels for drinking scenes in social media and sociodemographic factors.

YouTube Search Strategy and Data Collection

Using YouTube, the keywords "sulbang (alcohol-drinking shows)," "alcohol," "drinking broadcast," and "drinking" were searched. Results were sorted using YouTube's default search option of "relevance-based ranking." An analysis by an internet search engine revealed that over 90% of users typically select from the first 3 pages of search results, so 300 videos were chosen based on keywords collected through relevance-based rankings. A total of 318 videos were analyzed after excluding unrelated overlapping videos (eg, news) and videos with less than 10,000 views. Permission from YouTube was not required to conduct this study, as all the data used in our paper was publicly available, and no special access was needed for data collection.

Extracted Variables

Data collected for each video included video ID, upload date, views, likes, dislikes, comments, and video length. Video Power Index was calculated as ([Likes * 100 / [Likes+Dislikes]] * [Views/Day] / 100). Video content analysis was evaluated based on the Media Drinking Scene Guidelines developed by the Ministry of Health and Welfare and the Korea Health Promotion Institute.

The guidelines were first developed in 2017 and revised in 2023. They include the 12 recommendations set out below.

1. Scenes involving drinking should only be included if absolutely necessary.
2. Positive portrayals of drinking should be avoided.
3. Illegal behavior or actions related to drinking alcohol that disrupt public order should not be portrayed as normal.
4. Videos must refrain from depicting risky behaviors such as violence, suicide, or sensational behavior related to drinking.
5. Videos should not depict teenagers drinking alcohol and should be extremely cautious about depicting teenagers together with adults consuming alcohol.
6. Scenes depicting celebrities or public figures consuming alcohol must be portrayed with caution in the video, considering their impact.
7. Videos should avoid depicting harmful drinking behaviors, such as binge drinking or drunkenness.
8. Drinking scenes should not be used as a means of advertising alcohol-related products.
9. Scenes that disregard the right to self-determination regarding drinking should be avoided.
10. A negative drinking culture should not be depicted as a typical situation.
11. Content that excessively highlights or glorifies drinking should be restricted to minimize access by children and adolescents through age restrictions, etc.
12. In scenes that excessively highlight or glorify drinking, the harmful effects of alcohol consumption must be communicated through warning signs.

The final YouTube video content coding was independently analyzed by 2 trained coders. The intercoder reliability was assessed using Cohen kappa (κ=0.92), yielding a value above 0.8, which indicates a high level of agreement. The coders underwent training sessions to ensure consistency and adherence to the coding protocol.

Ethical Considerations

This study was approved by the National Cancer Center Ethics Board (project number: 2023-0284-0004). Before participation, all respondents were informed about the study's purpose, procedures, potential risks, and benefits. Informed consent was obtained online before the survey began, and participation was entirely voluntary. No compensation was provided to participants. Access to the data was restricted to authorized researchers only. To ensure privacy and confidentiality, all collected data were fully anonymized and deidentified before analysis.

XSL•FO
RenderX

Results

Sociodemographic Factors and the Frequency of Exposure to Alcohol-Drinking Videos

Table 1 shows the relationship between sociodemographic variables and the frequency of exposure to depictions of alcohol consumption on social media platforms. Although no statistically significant variances based on gender were observed, a higher percentage of women than men reported viewing social media content depicting alcohol consumption more than 4 times a week. There were no significant differences in age or marital status, but the younger age groups tended to view more social media containing alcohol consumption. Conversely, a significant difference was observed for educational level. Specifically, the group with higher education levels had the highest proportion of individuals viewing alcohol-related content more than 4 times a week (n=60, 16.9% for up to high school graduates; n=194, 19.0% for college graduates; and n=33, 26.2% for those with graduate degrees or higher; $P<.001$).

XSL•FO
RenderX

Table . The relationship between sociodemographic factors and the frequency of exposure to scenes depicting alcohol consumption in social media including web-based platforms: analysis of a nationwide survey of South Korean adults (N=1500).

	Total (N=1500), n (%)	The frequency of exposure to scenes depicting alcohol consumption in social media						*P* value
		None (n=220), n (%)	≤1 time/moment (n=205), n (%)	≤1 time/week (n=327), n (%)	2 - 3 times/week (n=461), n (%)	4 - 6 times/week (n=209), n (%)	Every day (n=78), n (%)	
Gender								.11
Men	761 (50.7)	111 (14.6)	102 (13.4)	188 (24.7)	228 (30.0)	95 (12.5)	37 (4.9)	
Women	739 (49.3)	109 (14.8)	103 (13.9)	139 (18.8)	233 (31.5)	114 (15.4)	41 (5.6)	
Age (years)								.07
20 - 29	242 (16.1)	30 (12.4)	26 (10.7)	53 (21.9)	71 (29.3)	47 (19.4)	15 (6.2)	
30 - 39	246 (16.4)	35 (14.2)	40 (16.3)	42 (17.1)	78 (31.7)	39 (15.9)	12 (4.9)	
40 - 49	307 (20.5)	42 (13.7)	34 (11.1)	67 (21.8)	107 (34.9)	38 (12.4)	19 (6.2)	
50 - 59	331 (22.1)	43 (13.0)	45 (13.6)	82 (24.8)	105 (31.7)	37 (11.2)	19 (5.7)	
60 - 69	288 (19.2)	53 (18.4)	50 (17.4)	68 (23.6)	73 (25.4)	34 (11.8)	10 (3.5)	
70+	86 (5.7)	17 (19.8)	10 (11.6)	15 (17.4)	27 (31.4)	14 (16.3)	3 (3.5)	
Marital status								.10
Single	538 (35.9)	81 (15.1)	74 (13.8)	114 (21.2)	152 (28.3)	84 (15.6)	33 (6.1)	
Married	888 (59.2)	121 (13.6)	118 (13.3)	201 (22.6)	293 (33.0)	113 (12.7)	42 (4.7)	
Separated, divorced, or widowed	74 (4.9)	18 (24.3)	13 (17.6)	12 (16.2)	16 (21.6)	12 (16.2)	3 (4.1)	
Educational status								<.001
≤ High school	355 (23.7)	77 (21.7)	57 (16.1)	71 (20.0)	90 (25.4)	44 (12.4)	16 (4.5)	
Bachelor's degree	1,019 (67.9)	127 (12.5)	134 (13.2)	230 (22.6)	334 (32.8)	143 (14.0)	51 (5.0)	
≥ Graduate degree	126 (8.4)	16 (12.7)	14 (11.1)	26 (20.6)	37 (29.4)	22 (17.5)	11 (8.7)	
Household income (US $)								<.001
<2000	288 (19.2)	68 (23.6)	42 (14.6)	59 (20.5)	58 (20.1)	44 (15.3)	17 (5.9)	
2000 - 3999	628 (41.9)	81 (12.9)	101 (16.1)	129 (20.5)	198 (31.5)	88 (14.0)	31 (4.9)	
4000 - 5999	310 (20.7)	29 (9.4)	35 (11.3)	76 (24.5)	109 (35.2)	49 (15.8)	12 (3.9)	
≥6000	274 (18.3)	42 (15.3)	27 (9.9)	63 (23.0)	96 (35.0)	28 (10.2)	18 (6.6)	
Self-reported health status								.53
Excellent to very good	486 (32.4)	62 (12.8)	73 (15.0)	110 (22.6)	156 (32.1)	57 (11.7)	28 (5.8)	
Good	827 (55.1)	126 (15.2)	112 (13.5)	182 (22.0)	246 (29.7)	121 (14.6)	40 (4.8)	
Fair to poor	187 (12.5)	32 (17.1)	20 (10.7)	35 (18.7)	59 (31.6)	31 (16.6)	10 (5.3)	
Smoking								<.001
Never-smoker	458 (30.5)	45 (9.8)	38 (8.3)	103 (22.5)	171 (37.3)	73 (15.9)	28 (6.1)	
Current smoker	282 (18.8)	29 (10.3)	44 (15.6)	73 (25.9)	89 (31.6)	35 (12.4)	12 (4.3)	
Ex-smoker	760 (50.7)	146 (19.2)	123 (16.2)	151 (19.9)	201 (26.4)	101 (13.3)	38 (5.0)	
Drinking alcohols								<.001

XSL•FO
RenderX

	Total (N=1500), n (%)	The frequency of exposure to scenes depicting alcohol consumption in social media						*P* value
		None (n=220), n (%)	≤1 time/moment (n=205), n (%)	≤1 time/week (n=327), n (%)	2 - 3 times/week (n=461), n (%)	4 - 6 times/week (n=209), n (%)	Every day (n=78), n (%)	
Nondrinking (Abstainer)	272 (18.1)	90 (33.1)	41 (15.1)	45 (16.5)	49 (18.0)	33 (12.1)	14 (5.1)	
Social drinker	660 (44.0)	87 (13.2)	95 (14.4)	146 (22.1)	214 (32.4)	92 (13.9)	26 (3.9)	
Heavy drinker	568 (37.9)	43 (7.6)	69 (12.1)	136 (23.9)	198 (34.9)	84 (14.8)	38 (6.7)	

Significant differences in exposure to alcohol-related content in media were observed based on income levels. An analysis focusing on the frequency of individuals who viewed social media featuring alcohol consumption more than 4 times a week indicated that the lowest income group (US $2000 per month) exhibited a higher tendency to see content associated with drinking (*P*<.001). The study did not find a significant association between self-reported health status and exposure to alcohol-related content on social media; however, a significant correlation was observed between smoking and alcohol consumption. Specifically, heavy drinkers were significantly more likely to view alcohol-related content on social media platforms (*P*<.001).

Opinion on Regulating Scenes Depicting Alcohol Consumption in Social Media Platforms

Figure 1 illustrates the opinions regarding the regulation of scenes depicting alcohol consumption on social media platforms, including web-based platforms such as YouTube. Among the respondents, 16.9% (n=253) said that viewing restrictions were unnecessary, while 29.9% (n=449) said that they believed cautionary warnings about alcohol-related scenes before viewing were essential. Additionally, a significant proportion (376/1500, 25.1%) felt that restrictions requiring adult certification for viewing were necessary, while 8.3% (n=124) of respondents advocated for stricter measures, such as pixelation, even with adult certification. Finally, 19.9% (n=298) of respondents supported restrictions that would prohibit the inclusion of alcohol consumption scenes in videos. Notably, women expressed a stronger need for restrictions compared with men, with 22.2% (n=164) of women believing that restrictions on drinking videos were necessary, compared with 17.6% (n=134) of men. The data also revealed that 21.4% (n=163) of men and 12.2% (n=90) of women believed that there should be no restrictions on drinking videos, further indicating that women were more in favor of restrictions on such videos.

XSL•FO
RenderX

Figure 1. Opinion on regulating scenes depicting alcohol consumption in social media platforms: Analysis of a nationwide survey of South Korean adults (N=1500).

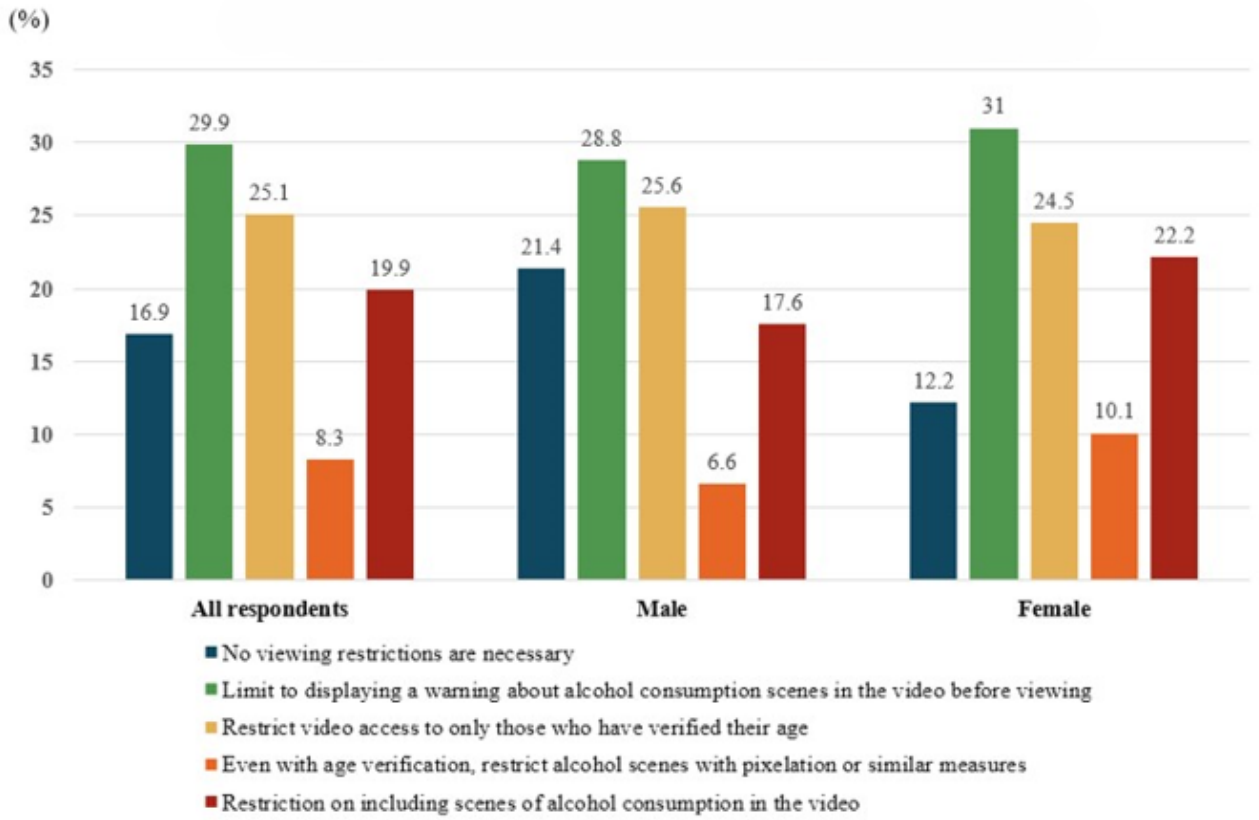

Regression Analysis of Perceptions of the Appropriate Level of Restriction for Drinking Scenes

Table 2 shows the results of an analysis conducted on the factors influencing perceptions of restrictions on drinking scenes. The analysis considered demographic factors such as age and gender, as well as the association between the amount of alcohol consumed and the level of exposure to drinking scenes on social media with perceptions of restrictions on drinking scenes. The perception of the level of restrictions on drinking scenes in social media was found to be positively associated with age, with a coefficient of 0.21984, and a *P* value of <.001. Accordingly, the analysis confirmed that older age groups believe that scenes depicting drinking on social media should have greater restrictions.

Regarding gender differences, women were significantly more likely than men to advocate for stricter limitations on depictions of drinking on social media (coefficient: 0.37827; *P*<.001), although no significant variances were observed for education level, marital status, household income, or self-reported health status.

Table . Regression analysis of perceptions of the appropriate level of restriction for drinking scenes in web-based social media and sociodemographic factors: analysis of a nationwide survey of South Korean adults (N=1500).

Variable	Coefficient	Standard error	*Z* value	*P* value
Age	0.21984	0.03993	5.505	<.001
Gender	0.37827	0.09352	4.045	<.001
Educational status	0.03793	0.08976	0.423	.67
Marital status	−0.0661	0.10786	−0.613	.54
Household Income	−0.02224	0.0493	−0.451	.65
Self-reported health status	0.02753	0.06323	0.435	.66
Drinking amount of alcohols	−0.3652	0.06605	−5.53	<.001
Exposure to scenes depicting alcohol consumption in web-based social media	1.0278	0.16492	6.23	<.001

However, in relation to the quantity of alcohol consumed, there was a stronger tendency to support less stringent restrictions on drinking scenes as alcohol consumption increased (coefficient: −0.3652; *P*<.001). The exposure level to drinking scenes showed

XSL•FO
RenderX

the strongest correlation (coefficient: 1.0278; $P<.001$), with individuals who were more frequently exposed to drinking scenes tending to hold stronger opinions regarding the necessity for increased restrictions on these scenes.

Descriptive Features and Video Content Analysis of "Sulbang"

The analysis encompassed 318 YouTube videos, with a Like ratio of 97.9% (range: 32.7 - 100.0), indicating a prevalent positive reception among viewers. The median Video Power Index value stood at 939.6 (ranging from 10.4 to 84,821.7; Table 3). Examination of alcohol-related content within drinking scenes, in accordance with the Media Drinking Scene Guidelines, revealed that approximately 89.0% (n=283) of the content glorified drinking (Table 4). Videos showing positive portrayals often suggest that alcohol helps with stress relief and can rejuvenate a fatigued body after just one drink. Videos featuring celebrities or notable figures consuming alcohol constituted 27.7% (n=88) of the total, while 92.8% (n=295) depicted either binge drinking or drunkenness. Videos that undermined the autonomy of others and promoted drinking accounted for 15.7% (n=50) of the total, with 42.8% (n=136) depicting a distorted drinking culture, often emphasizing high alcohol tolerance or portraying alcohol as a catalyst for sexual advances. Conversely, just 0.6% (n=2) of the videos were age-restricted to viewers younger than 19 years, and only 31.1% (n=98) included warnings.

Table . Characteristics of YouTube (sulbang) videos related to alcohol consumption (n=318).

Variable	Median (range)
Views	298,556 (10,404-17,790,441)
Video length (minutes)	14.8 (0.9 - 153.1)
Time on YouTube (days)	362.0 (126-2,505)
Comments	450.5 (8-23,845)
Likes (thumbs up)	5481 (10-832,067)
Dislikes (thumbs down)	150 (0 - 13,875)
Like ratio	97.9 (32.7 - 100.0)
View ratio	1030.5 (10.5 - 85,944.2)
Video power index	939.6 (10.4 - 84,821.7)

Table . Content analysis of YouTube (Sulbang) videos based on the 2023 Korean media drinking scene guidelines (n=318).

Media drinking scene guidelines	Not following the guidelines, n (%)
1. Scenes involving drinking should be minimized and only included if absolutely necessary.	3 (0.9)
2. Positive portrayals of drinking should be avoided.	283 (89.0)
3. Illegal behavior or actions related to drinking alcohol that disrupt public order should not be portrayed as normal.	5 (1.6)
4. Videos must refrain from depicting risky behaviors such as violence, suicide, or sensational behavior related to drinking.	4 (1.3)
5. Videos should not depict teenagers drinking alcohol and should be very careful about depicting teenagers together with adults consuming alcohol.	0 (0.0)
6. Scenes depicting celebrities or public figures consuming alcohol must be portrayed with caution in the video, considering their impact.	88 (27.7)
7. Videos should avoid depicting harmful drinking behaviors, such as binge drinking or drunkenness.	295 (92.8)
8. Drinking scenes should not be used as a means of advertising alcohol products.	37 (11.6)
9. Scenes that disregard the right to self-determination regarding drinking should be avoided.	50 (15.7)
10. A negative drinking culture should not be depicted as a typical situation.	136 (42.8)
11. Contents that excessively highlights or glorifies drinking should be restricted to minimize access by children and adolescents through age restrictions, etc	2 (0.6)
12. In scenes that excessively highlight or admire drinking, the harmful effects of alcohol consumption must be communicated through warning signs	98 (31.1)

XSL•FO
RenderX

Discussion

Principal Findings

Alcohol consumption is a major global health risk factor and social burden [26,27], yet it has been positively perceived as a means of stress relief and fostering social bonds in everyday life [14]. This perception has normalized alcohol use across various cultures, and exposure to drinking scenes has increased, particularly due to the rise in social media platform usage, such as YouTube, following the COVID-19 pandemic [28,29]. In South Korea, based on previous studies that suggested repeated exposure to drinking scenes could lead to the normalization of alcohol use and the glorification of its effects, 10 media drinking scene guidelines were introduced in 2017 to regulate exposure to drinking scenes in media. This effort was further enhanced in 2023 with updates that included age restrictions and warning labels, especially considering the pronounced impact on adolescents and young adults [21,22]. However, these guidelines remain legally unenforceable.

This study aims to emphasize the importance of implementing practical regulatory measures for social media platforms by analyzing sulbang (alcohol-themed broadcasting) content on YouTube and examining public opinion on regulating alcohol-related broadcasts on social media. The study explores the correlations between sociodemographic factors, frequency of exposure to drinking scenes on YouTube, and the global scope of regulations on such scenes. It also evaluates whether the increased sulbang content on YouTube aligns with the media drinking scene guidelines. The findings reveal that women, individuals with higher education levels, or those with lower incomes were more frequently exposed to drinking content. Although there was no significant difference between age groups, younger individuals tended to watch more alcohol-related social media content. Additionally, individuals with higher alcohol consumption reported greater engagement in alcohol-related content on social media platforms. Regarding opinions on regulating YouTube drinking content, women and older adults showed a stronger tendency to support stricter regulations. There was also a positive correlation between exposure to drinking scenes and belief in the necessity of regulation (greater exposure was associated with stronger support for regulation) and a negative correlation between alcohol consumption and belief in the necessity of regulation (higher consumption was associated with lower support for regulation). This may suggest that heavy drinkers perceive drinking as more normalized due to repeated exposure, which aligns with Bandura's social learning theory, explaining how observed behaviors can shape perceptions and habits [17,18]. Finally, an evaluation of YouTube (sulbang) content based on the media drinking scene guidelines showed that many videos portrayed alcohol consumption positively or emphasized binge drinking and distorted drinking cultures. However, relatively few videos included age restrictions or warning labels. These findings demonstrate a lack of enforcement mechanisms in web-based media regulations and highlight the need for future legally binding regulations.

Previous studies have widely demonstrated that exposure to drinking scenes in media increases positive perceptions of alcohol consumption and drinking behavior [30-32]. This study aligns with those findings, particularly emphasizing the tendency of binge drinkers to watch alcohol-related content more frequently and to have lower awareness of the need for regulation. This suggests that repeated exposure may reinforce positive attitudes toward drinking and lead to resistance against regulation [33]. According to 2021 high-risk drinking statistics in South Korea, while the high-risk drinking rate among men has decreased, it has increased among women, particularly among men aged 40-50 years and women aged 20-30 years [34]. This supports previous findings that women are more emotionally influenced by drinking scenes in media [35,36]. These results underline the need for tailored regulatory and educational strategies targeting specific demographic groups.

Currently, most countries enforce offline alcohol regulations, such as taxation and age restrictions on sales, but regulations for alcohol-related content on social media remain nonbinding recommendations. Considering the negative impacts of exposure to drinking scenes and the necessity of regulation, strong legal measures, such as age restrictions, broadcasting time regulations, and warning messages before broadcasts, should be swiftly implemented to effectively control the exposure to alcohol content on social media.

Limitations

This study has several limitations. First, the analysis focused solely on YouTube as the research platform. However, YouTube is the most widely used video-sharing platform worldwide and records the highest viewership among social media platforms. Moreover, it provides videos of sufficient length to analyze drinking behaviors. Therefore, analyzing only YouTube allowed for the collection of a sufficiently large data set. Second, there is no validated psychometric tool specifically designed to assess exposure to alcohol-related content and regulatory attitudes. As a result, a separate questionnaire was designed for this study. Although the survey instrument was not a previously validated tool, considerable effort was made to ensure the accuracy and reliability of the study results. To improve representativeness, a stratified sampling method was adopted based on region, age, and gender. Additionally, structured response options and clear definitions of key concepts were provided to minimize ambiguity and enhance response reliability. Another limitation is that this study indirectly assessed exposure to drinking scenes through a survey rather than directly measuring actual exposure. To address this, the survey items related to alcohol consumption were carefully structured to be as specific as possible. Finally, since this study is a cross-sectional study, it has limitations in explaining causal relationships. These limitations have been explicitly acknowledged in the conclusion, and future research should use data analytics and experimental methodologies to achieve more objective and direct evaluations.

Conclusions

This study aimed to analyze sulbang (alcohol-themed broadcasting) content on YouTube and investigate public opinion regarding the regulation of alcohol-related broadcasts on social media platforms like YouTube. The findings indicated

XSL•FO
RenderX

that individuals with higher tendencies toward alcohol consumption were more likely to frequently watch sulbang content and generally opposed regulatory measures. However, the majority of survey participants agreed on the necessity of regulating (sulbang) content. Additionally, an evaluation of sulbang content based on media drinking scene guidelines showed that most videos portrayed alcohol consumption positively or included binge drinking, while very few videos featured age restrictions or warning labels. Considering the potential influence of such content in shaping attitudes toward alcohol consumption and the possible impacts of increased exposure, it is essential to consider regulatory approaches for alcohol-related content on social media.

This study highlights the potential association between positively portrayed alcohol consumption content and public attitudes but cannot confirm causal relationships due to its reliance on cross-sectional data. To address this limitation, further research using longitudinal or experimental methodologies is needed. Such studies would play a crucial role in assessing the long-term effects of exposure and evaluating the effectiveness of proposed regulatory measures.

Acknowledgments

This study was supported by a grant from the National Cancer Center, South Korea (grant 2310651 - 2).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

EKK contributed to the study design. All authors collected and analyzed the research data, and interpreted the data. EKK and HRJ wrote the manuscript, and HL and JC reviewed the manuscript. All authors approved the final version of the manuscript for submission.

Conflicts of Interest

None declared.

References

1. 11th revision of the International Classification of Diseases: classes of psychoactive substances. : World Health Organization; 2024.
2. Global strategy to reduce the harmful use of alcohol. : World Health Organization; 2010.
3. Global status report on alcohol and health and treatment of substance use disorders. : World Health Organization; 2024.
4. Korea Health Promotion Institution. Alcohol statistics data collection. : World Health Organization; 2023.
5. GBD 2016 Alcohol and Drug Use Collaborators. The global burden of disease attributable to alcohol and drug use in 195 countries and territories, 1990-2016: a systematic analysis for the global burden of disease study 2016. Lancet Psychiatry 2018 Dec;5(12):987-1012. [doi: 10.1016/S2215-0366(18)30337-7] [Medline: 30392731]
6. GBD 2016 Alcohol Collaborators. Alcohol use and burden for 195 countries and territories, 1990-2016: a systematic analysis for the global burden of disease study 2016. Lancet 2018 Sep 22;392(10152):1015-1035. [doi: 10.1016/S0140-6736(18)31310-2] [Medline: 30146330]
7. Rehm J, Taylor B, Mohapatra S, et al. Alcohol as a risk factor for liver cirrhosis: a systematic review and meta-analysis. Drug Alcohol Rev 2010 Jul;29(4):437-445. [doi: 10.1111/j.1465-3362.2009.00153.x] [Medline: 20636661]
8. Turati F, Galeone C, Rota M, et al. Alcohol and liver cancer: a systematic review and meta-analysis of prospective studies. Ann Oncol 2014 Aug;25(8):1526-1535. [doi: 10.1093/annonc/mdu020] [Medline: 24631946]
9. Csengeri D, Sprünker NA, Di Castelnuovo A, et al. Alcohol consumption, cardiac biomarkers, and risk of atrial fibrillation and adverse outcomes. Eur Heart J 2021 Mar 21;42(12):1170-1177. [doi: 10.1093/eurheartj/ehaa953] [Medline: 33438022]
10. Habtemichael LH, Djekic D, Rosengren AR, et al. Alcohol consumption in young men and risk of heart failure and all-cause mortality- a cohort study. Eur Heart J 2021 Oct 12;42(Supplement_1):ehab724. [doi: 10.1093/eurheartj/ehab724.0830]
11. Rumgay H, Shield K, Charvat H, et al. Global burden of cancer in 2020 attributable to alcohol consumption: a population-based study. Lancet Oncol 2021 Aug;22(8):1071-1080. [doi: 10.1016/S1470-2045(21)00279-5] [Medline: 34270924]
12. Hels T, Bernhoft IM, Lyckegaard A, et al. Risk of injury by driving with alcohol and other drugs. DTU Orbit. 2011. URL: https://orbit.dtu.dk/en/publications/risk-of-injury-by-driving-with-alcohol-and-other-drugs [accessed 2024-08-07]
13. Parker RN. Alcohol and violence: connections, evidence and possibilities for prevention. J Psychoactive Drugs 2004 May;Suppl 2:157-163. [doi: 10.1080/02791072.2004.10400051] [Medline: 15279128]
14. Brown KG, Stautz K, Hollands GJ, Winpenny EM, Marteau TM. The cognitive and behavioural impact of alcohol promoting and alcohol warning advertisements: an experimental study. Alcohol Alcohol 2016 May;51(3):354-362. [doi: 10.1093/alcalc/agv104] [Medline: 26391367]

XSL•FO
RenderX

15. Alcohol Advertising and Promotion, National Academies of Sciences, Engineering, and Medicine. Reducing Underage Drinking: A Collective Responsibility: The National Academies Press; 2004. [doi: 10.17226/10729.2004]
16. Park K, Jo H, Na S, Lee H. Effects of exposure to drinking scenes in media on drinking problems in adults: dual mediating effects of positive alcohol expectancy and drinking motivation. Korea J Health Edu Promot 2021 Dec 31;38(5):33-43. [doi: 10.14367/kjhep.2021.38.5.33]
17. Hanewinkel R, Sargent JD, Poelen EAP, et al. Alcohol consumption in movies and adolescent binge drinking in 6 European countries. Pediatrics 2012 Apr;129(4):709-720. [doi: 10.1542/peds.2011-2809] [Medline: 22392174]
18. Anderson P, de Bruijn A, Angus K, Gordon R, Hastings G. Impact of alcohol advertising and media exposure on adolescent alcohol use: a systematic review of longitudinal studies. Alcohol Alcohol 2009;44(3):229-243. [doi: 10.1093/alcalc/agn115] [Medline: 19362797]
19. Snyder LB, Milici FF, Slater M, Sun H, Strizhakova Y. Effects of alcohol advertising exposure on drinking among youth. Arch Pediatr Adolesc Med 2006 Jan;160(1):18-24. [doi: 10.1001/archpedi.160.1.18] [Medline: 16389206]
20. Jackson KM, Janssen T, Gabrielli J. Media/marketing influences on adolescent and young adult substance abuse. Curr Addict Rep 2018 Jun;5(2):146-157. [doi: 10.1007/s40429-018-0199-6] [Medline: 30393590]
21. Media drinking scene guidelines. : Ministry of Health and Welfare, South Korea; 2017.
22. Media drinking scene guidelines (revision). : Ministry of Health and Welfare, South Korea; 2023.
23. A public survey on policy and promotion of preventing alcohol-related harms in 2021. : Ministry of Health and Welfare & Korea Health Promotion Institute; 2021.
24. Giannakodimos I, Lykouras D, Lagiou O, Tsakas S, Karkoulias K. The presence of smoking and alcohol imaginary in popular movies available in Netflix online streaming platform. J Community Health 2022 Aug;47(4):710-715. [doi: 10.1007/s10900-022-01101-8] [Medline: 35633424]
25. Kang S, Ryu JY, Yoo SC. Why do people watch ‘Sulbang’?: an exploratory study of social media sulbang watching motivation and content attitude. JSS 2021 Aug 31;60(2):407-454. [doi: 10.22418/JSS.2021.8.60.2.407]
26. Rehm J, Mathers C, Popova S, Thavorncharoensap M, Teerawattananon Y, Patra J. Global burden of disease and injury and economic cost attributable to alcohol use and alcohol-use disorders. Lancet 2009 Jun 27;373(9682):2223-2233. [doi: 10.1016/S0140-6736(09)60746-7] [Medline: 19560604]
27. Rumgay H, Shield K, Charvat H, et al. Global burden of cancer in 2020 attributable to alcohol consumption: a population-based study. Lancet Oncol 2021 Aug;22(8):1071-1080. [doi: 10.1016/S1470-2045(21)00279-5]
28. Comparison of video content usage patterns and perceptions by age group. : Hankook Research; 2024.
29. Alfayad K, Murray RL, Britton J, Barker AB. Content analysis of Netflix and Amazon Prime instant video original films in the UK for alcohol, tobacco and junk food imagery. J Public Health (Oxf) 2022 Jun 27;44(2):302-309. [doi: 10.1093/pubmed/fdab022]
30. Siegfried N, Pienaar DC, Ataguba JE, et al. Restricting or banning alcohol advertising to reduce alcohol consumption in adults and adolescents. Cochrane Database Syst Rev 2014 Nov 4;2014(11):CD010704. [doi: 10.1002/14651858.CD010704.pub2] [Medline: 25369459]
31. Moreno MA, Whitehill JM. Influence of social media on alcohol use in adolescents and young adults. Alcohol Res 2014;36(1):91-100. [Medline: 26259003]
32. Henehan ER, Joannes AE, Greaney L, Knoll S, Wong QW, Ross CS. Youth cognitive responses to alcohol promotional messaging: a systematic review. J Stud Alcohol Drugs Suppl 2020 Mar(s19):26-41. [doi: 10.15288/jsads.2020.s19.26]
33. Smith LA, Foxcroft DR. The effect of alcohol advertising, marketing and portrayal on drinking behaviour in young people: systematic review of prospective cohort studies. BMC Public Health 2009 Feb 6;9:1-11. [doi: 10.1186/1471-2458-9-51] [Medline: 19200352]
34. The Korea Disease Control and Prevention Agency. Trends in the prevalence of high-risk drinking, during 2012–2021. Public Health Wkly Rep 2023;16(12). [doi: 10.56786/PHWR.2023.16.12.3]
35. Doo Syen K, Kihwun C. Gender difference in the effect of sympathy and empathy with female drinking in TV drama on drinking intention: focusing on involvement as a mediator. Korea J Cons Advertising Psychol 2011 May;12(2):349-378. [doi: 10.21074/kjlcap.2011.12.2.349]
36. Patrick ME, Terry-McElrath YM, Lanza ST, Jager J, Schulenberg JE, O’Malley PM. Shifting age of peak binge drinking prevalence: historical changes in normative trajectories among young adults aged 18 to 30. Alcohol Clin Exp Res 2019 Feb;43(2):287-298. [doi: 10.1111/acer.13933] [Medline: 30645773]

XSL•FO
RenderX

Edited by A Mavragani; submitted 07.08.24; peer-reviewed by A Hartanto, C Najera; revised version received 10.03.25; accepted 11.03.25; published 21.05.25.

Please cite as:
Ju H, Lee H, Choi J, Kang E
The Necessity of Regulating Drinking Scenes on Social Media Platforms Focusing on YouTube Sulbang Videos: Public Opinion From Surveys and YouTube Content Analysis
JMIR Form Res 2025;9:e65162
URL: https://formative.jmir.org/2025/1/e65162
doi:10.2196/65162

© HyoRim Ju, HyeWon Lee, Juyoung Choi, EunKyo Kang. Originally published in JMIR Formative Research (https://formative.jmir.org), 21.5.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Self-Stigma in Adults Living With Chronic Skin Disease: Development of the HautKompass Web-Based Program and Pilot Test of Its Usability, Acceptability, and Feasibility

Juliane Traxler[1], MSc; Caroline F Z Stuhlmann[1], PhD; Neuza da Silva Burger[1], PhD; Christian Stierle[2,3], Dipl Psych, Prof Dr; Vahid Djamei[4], MSc; Anna Darzina[4], BSc; Marie Rudnik[1], MSc; Rachel Sommer[1], PhD

[1]German Center for Health Services Research in Dermatology, Institute for Health Services Research in Dermatology and Nursing, University Medical Center Hamburg-Eppendorf, Martinistraße 52, Hamburg, Germany

[2]Psychology School, Hochschule Fresenius, Hamburg, Germany

[3]Health Psychology and Paedagogy, Riga Stradiņš University, Riga, Latvia

[4]Swiss4ward, Zurich, Switzerland

Corresponding Author:
Juliane Traxler, MSc
German Center for Health Services Research in Dermatology, Institute for Health Services Research in Dermatology and Nursing, University Medical Center Hamburg-Eppendorf, Martinistraße 52, Hamburg, Germany

Abstract

Background: Self-stigma is common among people with chronic skin disease and can substantially impair quality of life and psychosocial well-being. Few interventions targeting skin disease–related self-stigma are available, especially in Germany.

Objective: This pilot study aims to develop a web-based self-guided program to reduce self-stigma among people with chronic skin disease, and test its usability, acceptability, and feasibility.

Methods: We developed the HautKompass program based on 2 systematic literature reviews and the expertise from psychodermatologists and patients. Its usability, acceptability, and feasibility were tested among adults with psoriasis, atopic dermatitis, hidradenitis suppurativa, alopecia areata, and vitiligo. After completing the program, participants provided feedback on each session and on their overall experience using the Client Satisfaction Questionnaire and study-specific feedback items. The program was considered feasible if the dropout rate was below 40% and participants spent 45 minutes or less per session. Data were analyzed descriptively.

Results: HautKompass is grounded in compassion-focused therapy and cognitive behavioral therapy and consists of 8 self-guided sessions. Of the 41 persons who provided informed consent and filled in the screening questionnaire, 29 were eligible for participation. A total of 27 participants started the program and 20 completed all sessions and the posttest questionnaire. Results indicated high usability (mean 26.12, SD 6.13; on a scale ranging from 8 - 32) and acceptability (mean 17.41, SD 3.12; on a scale ranging from 5 - 20). Users rated the program as helpful, the psychoeducation and exercise instructions as comprehensible, and the extent of the program as adequate. Criticism concerned the length of some sessions, the electronic voice used in exercises, and some of the examples being too “general” or “cliché.” Regarding feasibility, the program’s extent was deemed adequate, participants spent substantially less time on the individual sessions (mean 16.9, SD 4.4 minutes) than the predefined criterion and few participants dropped out after starting the program (n=7, 26%), indicating low barriers. Importantly, the majority of users dropped out within the first 2 sessions, possibly due to the slightly longer duration and focus on theory, or due to unmet expectations, highlighting areas for improvement.

Conclusions: Overall, HautKompass was found to be a usable, acceptable program with feasible implementation. Limitations of this pilot study include not testing accessibility for people with disabilities and the small, relatively young, and mostly female sample, which limits the generalizability of the findings. The feedback obtained was used to revise the program and the recruitment strategy prior to testing its effectiveness in a randomized controlled trial. If HautKompass is found to be effective in reducing skin disease–related self-stigma, it will be made widely available to improve psychosocial care for people with chronic skin disease and could be adapted for other skin diseases and visible differences in general.

Trial Registration: ClinicalTrials.gov NCT06324695; https://clinicaltrials.gov/study/NCT06324695

(JMIR Form Res 2025;9:e70290) doi:10.2196/70290

KEYWORDS

intervention; psychosocial care; self-stigmatization; stigma; dermatology; visible difference; eHealth

XSL•FO
RenderX

Introduction

People with visible differences, such as skin disease, often experience both external (or public) and internalized (or self-) stigma [1,2], which can have a range of serious consequences: low self-worth, avoidance of social situations, interference with goal achievement, and ultimately poor health outcomes and quality of life [1,3-5]. Patients, their relatives, and health care providers consider it the biggest burden in several life domains, including work and relationships [1,4,6].

To alleviate this debilitating impact, psychosocial interventions reducing self-stigma in persons with chronic skin disease are urgently needed. Systematic reviews by Topp et al [7] and Traxler et al [8] reported that most interventions including counseling, self-help, social skills training, and cognitive behavioral therapy (CBT), generally had positive effects on self-stigma and related constructs. Importantly, a range of limitations and gaps was identified in the current literature: heterogeneous study quality, insufficient description of rationale and development process, few studies on inflammatory skin diseases, and limited availability across countries and cultures. Consequently, the development of new, ideally skin-generic, psychosocial interventions and a thorough evaluation of their effectiveness are warranted.

Most interventions identified by Topp et al [7] and Traxler et al [8] require in-person contact, which has the disadvantages of potentially long waiting lists, higher costs, and greater demand from patients in terms of travel and their schedules. Many people with a dermatological condition are already burdened by frequent medical appointments and high direct and indirect costs of their disease [9-11]. Furthermore, severely affected patients in particular may not feel comfortable attending face-to-face meetings for 45 minutes or more and in the presence of a “therapist” who can be seen as a potential “stigmatizer” [12], due to their symptoms making it difficult to sit for long periods of time or due to social anxiety. Online interventions and self-guided programs can be a useful, easily accessible alternative and have the potential to reach far more individuals than face-to-face interventions [13]. Computer and web-based treatments have been studied for a variety of psychological conditions and proved effective, cost-effective, and not inferior to face-to-face therapy [13,14]. However, few online interventions are available for dermatology [15] and self-stigma, none of which are available in German.

To fill this gap, this project aimed to develop a free web-based self-guided program to reduce skin disease–related self-stigma and pilot test its usability, acceptability, and feasibility.

Methods

Overview

The HautKompass project, funded by the German Ministry of Education and Research (01GY2105), consists of 4 phases: (1) 2 systematic literature reviews, (2) development and pilot testing of the web-based intervention HautKompass, (3) a randomized controlled trial (RCT) to evaluate the effectiveness of the intervention (trial registration: ClinicalTrials.gov NCT06324695), and (4) implementation and dissemination. This study specifically addresses phase 2. For the pilot test, usability was defined as the extent to which HautKompass could be used as intended and with ease by patients with chronic skin disease; acceptability was defined as whether users found the HautKompass program, its content, interface, navigation, and design appealing; and feasibility was defined as whether HautKompass could be easily and conveniently implemented. Specifically, implementation was considered feasible if (1) the dropout rate from the study was below 40% [16] and (2) participants spent 45 minutes or less per session. Definitions are based on Rothstein et al [17] and the World Health Organization [18].

HautKompass Development

Initially, 2 systematic reviews were conducted to identify (1) psychosocial predictors and mechanisms of self-stigma [19], and (2) existing interventions targeting self-stigma in chronic skin disease [8]. The first review included 28 studies, which found the experience of social stigma, lack of social support, and coping strategies, such as acceptance, to be important, modifiable psychosocial determinants of self-stigma. Based on these results, a conceptual model of (self-) stigma related to chronic skin disease originally developed by Germain et al [20] was refined to be used as a therapeutic framework for HautKompass (Figure 1) [19].

XSL•FO
RenderX

Figure 1. Theoretical model of determinants of stigma used in the HautKompass pilot study in German adults with chronic skin diseases. QoL: quality of life.

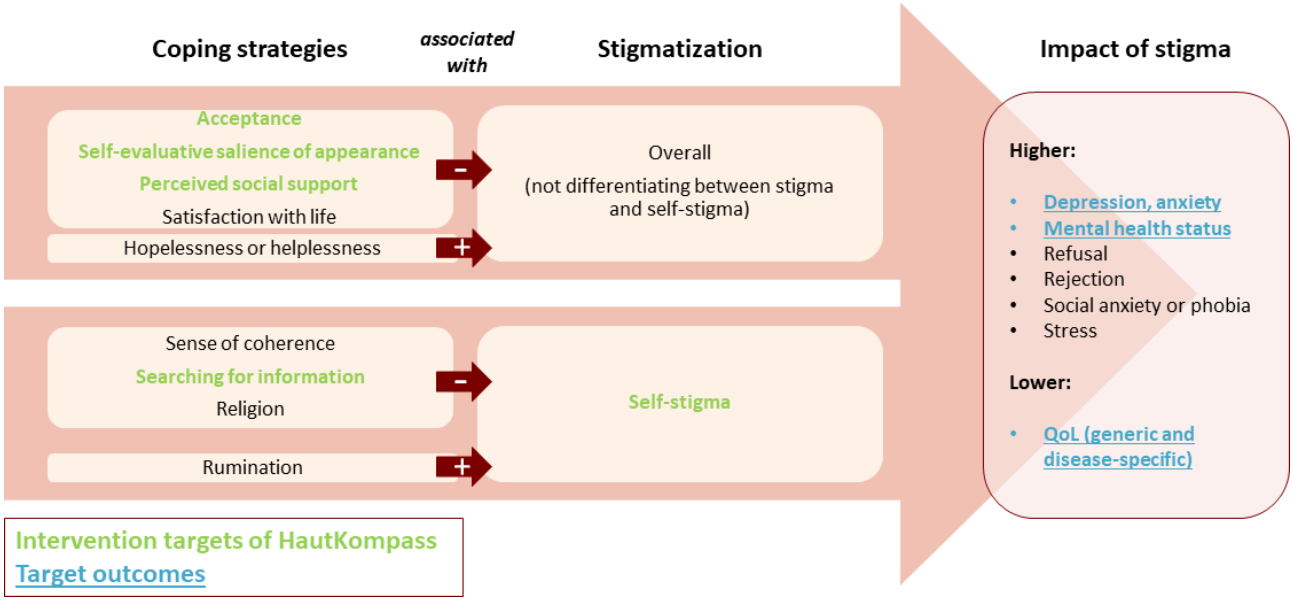

To address these psychosocial determinants, HautKompass is grounded in CBT and compassion-focused therapy [21], both of which have been found to be viable and effective when delivered online [22,23]. The available literature provides the best quality evidence for CBT approaches to effectively reduce self-stigma in dermatology [24,25]. Self-compassion interventions have received little attention in the field of dermatology. However, their successful application in the field of weight-related self-stigma [26-29], their effectiveness in reducing shame and self-criticism in patients with psoriasis [30], and the association of self-compassion with adaptive coping [31] is promising for other appearance-related concerns. Furthermore, Norman et al [32] suggest that some people with visible differences, especially those who have experienced severe discrimination, might feel more comfortable with compassion-focused therapy than when purely relying on challenging unhelpful thoughts, which is the focus of CBT.

The structure and content were developed based on the 2 systematic reviews [8,19] and in consultation with a patient advisory board (n=6) as well as health care professionals with expertise in psychodermatology (n=6). Specifically, at 2 time points, patients were asked to evaluate design and content, plus to provide unstructured feedback via an online form. Health care professionals were consulted repeatedly in meetings and through email contact. According to the web-based interventions [24,33-35] identified in the systematic review [8] and expert discussion, we aimed to develop a focused program with a few, relatively short sessions so as not to exhaust the users' attention, but with a higher frequency and flexibility than typical face-to-face therapy. Thus, the program consists of 8 sessions of approximately 15-25 minutes each, ideally completed at least once a week, which users can access from any computer and work through independently at their own pace.

As the program was designed to be self-guided and unsupervised, it consists predominantly of psychoeducation and relatively simple exercises, including a series of breathing and imagery exercises, attention training, writing exercises, and goal setting. All content has been tailored for people with chronic skin disease. Five patient characters and 2 therapist characters have been created to guide users and to share example stories. An outline of the session topics and an illustration of the website can be found in Figure 2. Sessions and individual exercises can be repeated as needed. Users are encouraged to practice the learned techniques in their daily lives and to take breaks between sessions. Email reminders to continue with or to complete sessions were sent automatically after 2, 5, and 7 days of inactivity in the program.

XSL•FO
RenderX

Figure 2. Information on the HautKompass program as used in the pilot study in German adults with chronic skin diseases. (A) Outline of the HautKompass session topics. (B) HautKompass session navigation pane. The sessions build on each other and are unlocked after completion of the previous one. Completed sessions can be reviewed at any time and selected exercises can be repeated under “Übungen” (“exercises”).

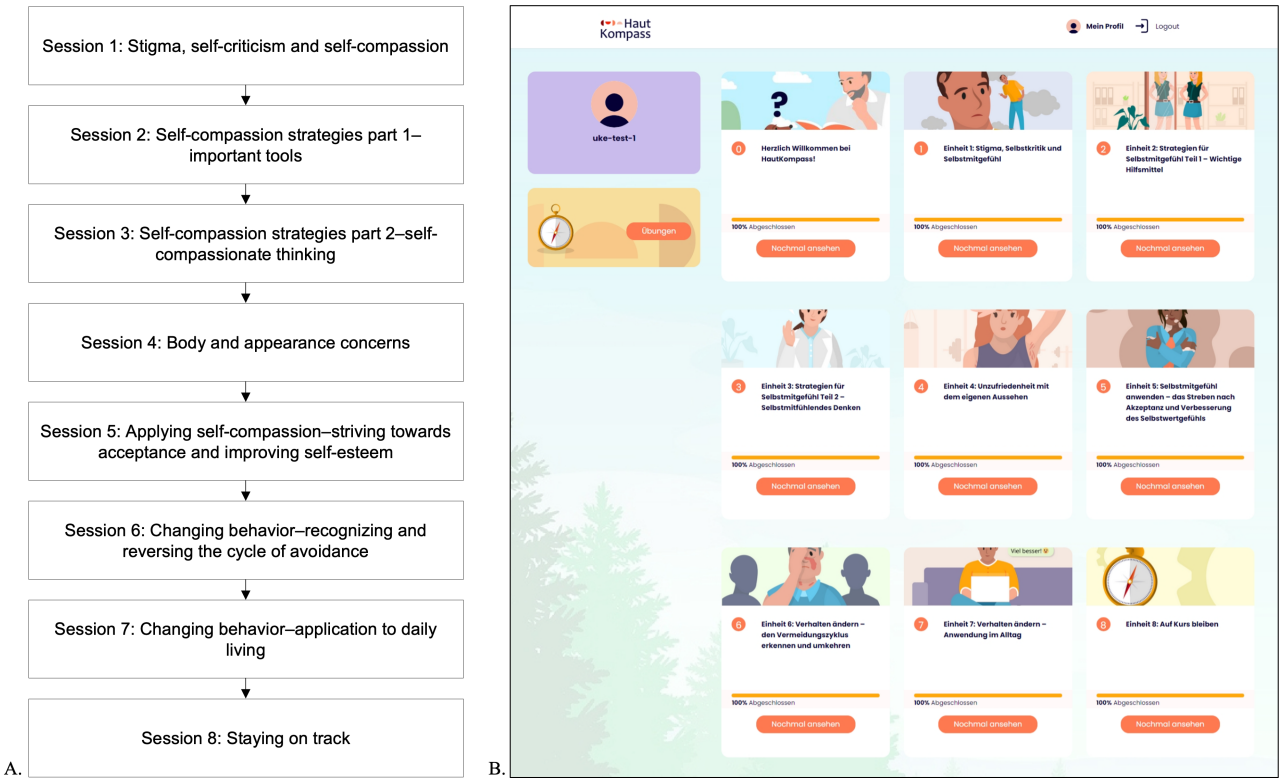

Participants and Recruitment

The target population was adults with alopecia areata, atopic dermatitis, hidradenitis suppurativa, psoriasis, or vitiligo. This selection combines indications with high disease burden and either high prevalence or particular clinical importance in dermatology. A convenience sample was recruited from the Institute for Health Services Research in Dermatology and Nursing, the dermatology outpatient clinic at University Medical Center Hamburg-Eppendorf. Adults with one of the target skin diseases were informed about the study by clinical staff and student assistants and were handed a flyer with a link and QR code to the registration page and contact details of the study team. Interested individuals could register online and receive detailed study information. This recruitment strategy may have led to selection biases, which need to be considered when interpreting the results.

Eligibility criteria were assessed through self-report using a screening form and included being 18 years or older, having one of the specified conditions, access to a stable internet and computer, proficiency in German, and the ability to consent autonomously. Participants were excluded if they received psychiatric or psychotherapeutic treatment within 12 months prior to the study, which could confound the assessment of the effectiveness of the program, or if they scored high on depression or anxiety in the screening questionnaire (see Outcome Measures section), as severely burdened individuals are recommended to seek more tailored professional help. People deemed ineligible for participation received information on how to find psychological help if needed, irrespective of the reason for which they were excluded. Enrollment closed once 5 participants per diagnosis had filled in the baseline questionnaire. Participants spending less than 5 minutes per session for more than half of the sessions were excluded post hoc from the analyses, as their feedback was deemed potentially unreliable. Participants who completed at least 1 session were considered for the analyses of the completed sessions.

Procedure

The pilot study was conducted entirely web-based between October 2023 and December 2023. Upon enrollment, participants filled in a first questionnaire on demographics and clinical data and then received unique log-in credentials for the HautKompass website. Participants could work through the program at their own pace but were encouraged to complete it within 14 days. At the end of each session, they were asked to fill in a short postsession questionnaire concerning that specific session and could enter free text feedback. After completion of the last session, participants were asked to provide more extensive and global feedback in a posttest questionnaire.

Outcome Measures

Screening Questionnaire

Patient Health Questionnaire-4 (PHQ-4) [36,37] screener consists of 4 items: 2 items that tap into depressive symptoms and 2 items for anxiety. Participants rate the degree to which they have experienced the respective complaints (eg, “feeling nervous, anxious or on edge”) over the last 2 weeks on a scale from 0=not at all to 3=nearly every day. Scores range from 0 to 12, with higher scores reflecting greater symptomatology. To screen persons before participation, a sum score of ≥5 on either the 2 anxiety items or the 2 depression items was used

XSL•FO
RenderX

as a cutoff for inclusion. First, this criterion was chosen as the program was not designed for people with severe mental health problems but rather as a first low-barrier service for mildly to moderately affected individuals. Second, learning and practicing self-compassion is often difficult for individuals with clinical levels of depression and anxiety [38], especially without therapist guidance.

Postsession Questionnaire

Items and response scales can be found in Table 1. Usability and acceptability were each assessed by means of 4 self-developed items (eg, “The exercises in this session are useful”). Feasibility was assessed by recording dropouts and using website analytics, namely the time taken to complete each session.

XSL•FO
RenderX

Table . Postsession feedback questionnaire ratings, number of participants completing each session, and average session completion time in the HautKompass pilot study in German adults with chronic skin diseases.[a]

Item	Response scale	Values								
		S1	S2	S3	S4	S5	S6	S7	S8	all S
Usability, mean (SD)										
The session was easy to understand.	1=do not agree at all to 5=fully agree	4.70 (0.47)	4.65 (0.79)	4.59 (0.51)	4.35 (0.93)	4.76 (0.44)	4.71 (0.47)	4.63 (0.62)	4.56 (0.63)	4.58 (0.42)
The session was helpful for me personally.	1=do not agree at all to 5=fully agree	3.85 (1.09)	3.72 (1.41)	3.88 (1.36)	3.53 (1.66)	4.41 (1.00)	4.12 (1.27)	4.19 (1.17)	4.19 (1.11)	3.73 (1.10)
The exercises in this session are useful.	1=does not apply to me to 4=totally applies to me	3.20 (0.89)	3.22 (1.06)	3.12 (0.99)	2.88 (1.05)	3.65 (0.79)	3.59 (0.80)	3.50 (0.63)	3.53 (0.83)	3.19 (0.85)
The exercise instructions in this session are easy to understand.	1=does not apply to me to 4=totally applies to me	3.68 (0.58)	3.56 (0.62)	3.82 (0.39)	3.44 (0.63)	3.82 (0.39)	3.82 (0.39)	3.63 (0.50)	3.44 (0.52)	3.58 (0.45)
Acceptability, mean (SD)										
The amount of text was …	1=far too little to 5=far too much	3.20 (0.52)	3.24 (0.44)	3.41 (0.51)	3.71 (0.69)	3.00 (0.35)	3.29 (0.59)	3.44 (0.63)	3.75 (0.78)	3.37 (0.40)
The length of the session was …	1=far too short to 5=far too long	3.10 (0.45)	3.29 (0.47)	3.29 (0.47)	3.59 (0.71)	3.06 (0.24)	3.18 (0.39)	3.38 (0.5)	3.69 (0.79)	3.34 (0.38)
The session was interesting.	1=do not agree at all to 5=fully agree	4.25 (0.91)	4.17 (0.99)	4.06 (1.20)	3.71 (1.53)	4.47 (1.01)	4.18 (1.13)	4.19 (1.17)	4.38 (1.15)	4.06 (0.98)
The examples used are appropriate and helped me better understand the content of this session.	1=does not apply to me to 4=totally applies to me	3.70 (0.57)	3.17 (0.99)	3.24 (0.97)	3.00 (0.79)	3.41 (0.87)	3.59 (0.62)	3.56 (0.63)	3.50 (0.82)	3.31 (0.60)
Feasibility										
Completed, n (%)	—[b]	22 (92)	20 (83)	18 (75)	18 (75)	18 (75)	18 (75)	17 (71)	17 (71)	—
Completion time (minutes), mean (SD)	—	21.13 (25.68)	21.66 (21.70)	21.17 (19.13)	19.58 (9.39)	11.95 (5.49)	14.48 (8.75)	11.70 (4.60)	13.53 (6.90)	16.9 (4.10)

[a]S1–S8: Sessions 1–8; all S: averaged across all 8 sessions.

[b]Not applicable.

XSL•FO
RenderX

Posttest Questionnaire

Items and response scales can be found in Table 2. Usability was assessed by means of 5 self-developed items, 3 of which were based on those developed by Lara et al [39]. Acceptability was assessed through 2 items from Lara et al [39], 2 self-developed items, and the Client Satisfaction Questionnaire Adapted to Internet-Based Interventions (CSQ) [40]. The CSQ is an 8-item instrument (eg, "The training has met my needs") assessing participants' global satisfaction with internet-based interventions on a 4-point Likert scale ranging from 1=does not apply to me to 4=does totally apply to me. To compute the total score, item scores are summed and can range from 8 to 32, with higher scores indicating greater satisfaction. Feasibility was assessed through one self-developed item in the posttest questionnaire.

Table . Descriptives of posttest feedback questionnaire items (N=17)[a] in the HautKompass pilot study in German adults with chronic skin diseases.

Item	Response scale	Values
Usability, mean (SD)		
Overall, the extent of the program was …	1=far too short to 5=far too long	3.47 (0.72)
I felt comfortable writing about my personal thoughts and experiences.	1=not true to 4=completely true	3.35 (0.70)
The website is user-friendly.[b]	1=not true to 4=completely true	3.76 (0.56)
HautKompass has met my expectations.[b]	1=not true to 4=completely true	3.35 (0.93)
HautKompass seems useful for dealing with skin disease–related self-stigma.[b]	1=not true to 4=completely true	3.53 (0.62)
Acceptability		
Is there anything you would want to change about the format of HautKompass? n (%)	Yes	5 (29)
Is there anything you would want to change about the content of HautKompass? n (%)	Yes	1 (6)
The website is persuasive, mean (SD)[b]	1=not true to 4=completely true	3.29 (0.99)
The page layout of the sessions is inviting, mean (SD)[b]	1=not true to 4=completely true	3.47 (0.72)
CSQ[c]	1=not true to 4=completely true	26.12 (6.13)
Feasibility		
Have you practiced any of the exercises and learned techniques outside the session? n (%)	Yes	14 (82)

[a]The posttest feedback questionnaire was completed by 20 participants, of whom 3 were excluded post hoc

[b]Items based on Lara et al [39].

[c]CSQ: Client Satisfaction Questionnaire Adapted to Internet-Based Interventions.

Statistical Analysis

Data were processed using SPSS (version 27.0; IBM Corp). Descriptive statistics were computed for sociodemographic and clinical data as well as for the CSQ, the acceptability questionnaire, and all feedback questions.

Ethical Considerations

The study was carried out in compliance with the Helsinki Declaration. Ethical approval was obtained from the local ethics committee at University Medical Center Hamburg-Eppendorf (approval reference: LPEK-0568). Informed consent was obtained from all participants through a written consent form, which provided details about the nature and goals of the study as well as potential harms and benefits. Participation in the study was entirely voluntary and participants retained the right to withdraw at any time without any disadvantages. Data were anonymized. Participants received a US $77.35 voucher as compensation.

Results

Sample Characteristics

Figure 3 depicts the participant flow. A total of 27 participants (n=18, 75% female) started the HautKompass program, of which 20 (74%) completed all sessions and the posttest questionnaire. Three participants completed more than half of the sessions in less than 5 minutes and, therefore, were excluded from the analysis. The demographics and clinical characteristics are presented in Table 3.

XSL•FO
RenderX

Figure 3. Participant flow in the HautKompass pilot study in German adults with chronic skin diseases. Three participants were excluded post-hoc from the analyses. PHQ-4: Patient Health Questionnaire-4.

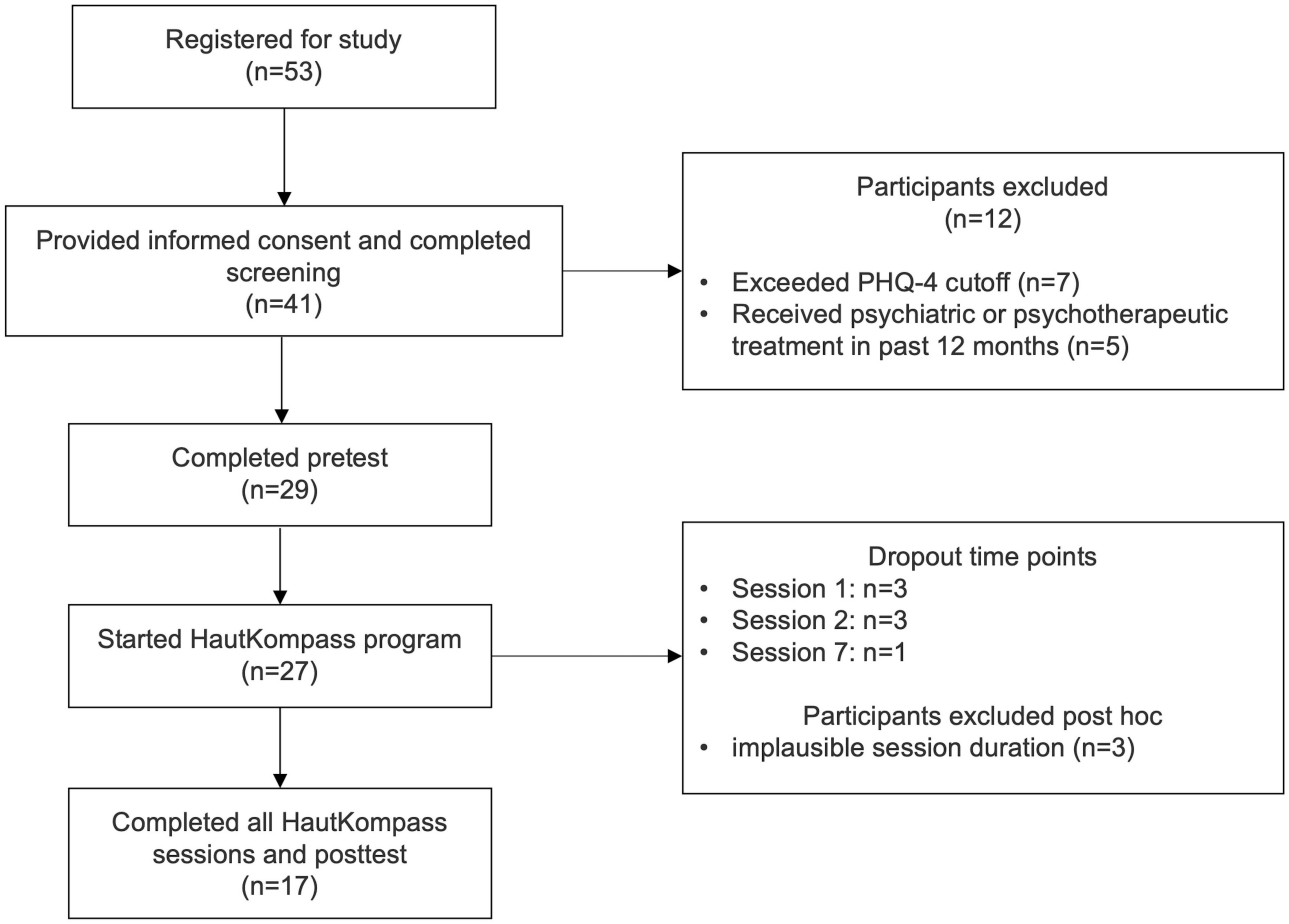

XSL•FO
RenderX

Table . Participant demographics and clinical characteristics (n=24) in the HautKompass pilot study in German adults with chronic skin diseases.

		Values
Age (years), mean (SD; range)		32.2 (12.25; 18 - 63)
Sex, n (%)		
	Female	18 (75)
	Male	6 (25)
Marital status, n (%)		
	Single	7 (29)
	In relationship	11 (46)
	Married	5 (21)
	Divorced	1 (4)
Education, n (%)		
	Secondary school	9 (38)
	Vocational training	3 (12)
	Bachelor's degree	9 (38)
	Master's degree	3 (12)
Employment status, n (%)		
	Student	9 (38)
	Employed	14 (58)
	Unemployed	1 (4)
Skin condition, n (%)		
	Alopecia areata	1 (4)
	Atopic dermatitis	7 (29)
	Hidradenitis suppurativa	9 (38)
	Psoriasis	6 (25)
	Vitiligo	1 (4)
Current treatment, n (%)		
	Yes	21 (88)
	No	3 (12)
Duration skin condition (years), n (%)		
	Less than 1	1 (4)
	1 - 5	5 (21)
	5 - 10	3 (12)
	More than 10	15 (63)
Disease severity, n (%)		
	Mild	11 (46)
	Moderate	11 (46)
	Severe	2 (8)

Usability

Tables 1 and 2 show the participants' responses to the postsession and posttest feedback questionnaires, respectively. The extent of HautKompass was perceived to be adequate and participants felt overall comfortable writing about their personal thoughts and experiences. Furthermore, the website was rated as very user friendly, useful for reducing skin disease–related self-stigma, and has mostly met users' expectations. The individual sessions were overall easy to understand (mean 4.56, SD 0.42), and rather helpful for users personally (mean 3.73, SD 1.10). The exercises were perceived as mostly useful (mean 3.19, SD 0.85), and the instructions as easy to understand (mean 3.58, SD 0.45).

XSL•FO
RenderX

Acceptability

Acceptability of HautKompass was generally high. Both layouts of the website and individual sessions were rated as likable. User satisfaction assessed with the CSQ was high, with an average score of mean 26.12 (SD 6.13, range: 9 - 32). Six participants (35%) suggested making some changes to the format or content of the program, such as to "also address people who don't think negatively of themselves all day long" and to make the examples "*less general*" or "*less cliché*." Regarding session-specific feedback, participants rated the sessions as interesting (mean 4.06, SD 0.98) and considered the examples appropriate and helpful (mean 3.31, SD 0.60). The amount of text (mean 3.37, SD 0.40) and length of each session (mean 3.34, SD 0.38) were both rated as "just right" on average, although 2 sessions were rated as a bit too long and text-heavy, which also reflected in the open text feedback (Textbox 1). The open text feedback revealed some additional criticisms and suggestions but many positive features were highlighted as well.

Textbox 1. Examples of free text feedback in the HautKompass pilot study in German adults with chronic skin diseases.

Criticism

- "The feedback from the system does not address your specific case, but is very general, which is somewhat frustrating after you have put energy into the exercise." [Session 1]
- "It would be more pleasant if a real voice were used instead of a computerized voice." [Session 2]
- "Going through the same concept three times, even if it's a different situation, is a bit long." [Session 8]
- "Apart from the length, the unit was really good." [Session 8]

Positive comments

- "I think the session is really very good and everything is right in terms of content. It's easy to understand and certainly helps everyone in general. It's also fun and motivates you to do the next session." [Session 3]
- "It was good to take some active time and meditate / relax to calm down." [Session 2]
- "The calendar with the planned activities is great. I have decided to have a fixed day where I do things that I like to do as a relaxation day." [Session 3]
- "I found it very helpful to set a precise goal which I work through step by step." [Session 7]

Feasibility

In total, 53 persons registered for the study (female: n=34, 64%; male: n=18, 34%; diverse: n=1, 2%), of which 41 (female: n=28, 68% and male: n=13, 32%) signed the informed consent form and completed the screening questionnaire. The nonparticipation rate was higher for males (n=5, 28%) than females (n=6, 17%). At this stage, 12 persons were excluded, either due to high scores on the PHQ-4 (n=7) or due to receiving some form of psychotherapeutic treatment within the past 12 months (n=5). All of the eligible 29 persons filled in the pretest questionnaire, and 27 started working through the program. Six of them dropped out within the first 2 sessions, while only 1 participant dropped out at a later stage (dropout rate=25.93%). In total, 20 users completed all sessions and the posttest questionnaire.

On average, participants spent 16.90 minutes per session, with session 7 being the shortest (11.70 minutes) and session 2 being the longest (21.66 minutes). Furthermore, 14 participants (82.35%) practiced some of the exercises and learned techniques in their daily lives.

Discussion

Principal Results

The aim of this pilot study was to assess the usability, acceptability, and feasibility of the newly developed HautKompass web-based program for people experiencing skin disease–related self-stigma. Overall, all 3 criteria were found to be largely satisfactory: users found HautKompass to be acceptable and usable, and suggested a few changes, which were addressed where possible. The predefined feasibility criteria, namely a dropout rate below 40% and a duration of individual sessions below 45 minutes, were met.

Overall, HautKompass was rated as highly userfriendly and useful. In light of the relatively young sample (aged 18 - 63 years), it needs to be noted that these results were obtained in a small sample of 20 individuals and may not apply to older individuals, many of whom are less adept and less confident in using online programs and might experience barriers that were not detected in this study.

The pilot users found the program largely acceptable and the overall appearance was appealing to most. A few suggestions for changes were made, which were addressed to enhance the acceptability of HautKompass: the computerized voice was replaced by real voice recordings, and some examples were slightly rephrased to be more appealing or relatable. One user indicated that receiving nonpersonalized feedback was frustrating, which aligns with previous findings that users of eHealth programs value personalized content and feedback [41,42]. Unfortunately, offering tailored feedback was not feasible within the scope of this project. In addition and as described earlier, the results may not be generalizable to male and older users. While we have deliberately designed both male and female characters of different age groups and aimed to include examples that may be relevant in different stages in life, we cannot rule out that the design and overall appearance of HautKompass appeal most to younger females. In fact, 2 users (female, 24 years and male, 33 years) raised concerns that the

XSL•FO
RenderX

program seemed to be designed for a relatively young audience and that older users might not feel equally well represented. The data also suggest that it is feasible to conduct an RCT evaluating the effectiveness of HautKompass and subsequently implementing the program for broader use. It is worth noting that only half of those who registered for the study actually started the program. Twelve participants (23%) were excluded at screening, which may indicate that people with poorer mental health in particular seek this kind of help, and the exclusion criteria were too strict. Although the program was not specifically designed for individuals with severe anxiety or depression, a recent meta-analysis showed that self-compassion can alleviate depressive and anxiety symptoms also in clinical populations, even using web-based formats [43]. It was thus decided to omit the PHQ-4 criterion for the upcoming RCT. A further 12 of those who registered did not give informed consent, possibly because they were only given more detailed information about the study at this point. It is conceivable that they were either discouraged by the time commitment required for this pilot study or did not feel that the program was for them, both of which underline the need for more targeted recruitment. However, the dropout rate from the program itself (25.93%) was comparable to or even lower than in other digital interventions [16,44,45]. This shows that, once participants complete the screening, they encounter few barriers and after working through the first sessions are unlikely to dropout. The finding that the sessions during which most participants did dropout (sessions 1 and 2) were also the longest ones and may indicate the participants' preference for short and concise sessions, which is in line with previous research [46]. Furthermore, the information provided prior to participation may have been insufficient such that some participants held expectations they realized were not going to be met when starting the program [47]. Another potential explanation is that the first session in particular was largely theoretical and offered little concrete advice for daily life. To address these issues, we shortened some sessions by omitting challenging and less useful exercises, avoiding repetitions, and reducing the amount of text, and we refined the recruitment information to better reflect the program's content and the study procedures in order to give potential participants realistic expectations. Altogether, the short session duration, low dropout rate, and willingness of users to revise exercises still demonstrate the feasibility of implementing HautKompass as a self-guided flexible, web-based program.

Strengths and Limitations

An important strength of the development process is its participatory approach. A patient advisory board was consulted repeatedly to tailor the program to their needs to guarantee that HautKompass would be both relevant and appealing to the target group. Furthermore, the sample covered a wide range of educational backgrounds.

This study also has several limitations that need to be addressed. The sample is rather small and not fully representative of the general population; for example, men were underrepresented and although the age range went up to 63 years, the samples were relatively young. This may indicate that men and older people do feel less addressed by the HautKompass program and the recruitment strategy. While some studies testing the effectiveness of eHealth interventions have not found substantial gender differences, men are typically less likely to seek help [39] and more likely to drop out from web-based interventions [48]. Evidence suggests older people are still more reluctant to use eHealth interventions, although the proportion of elderly users is increasing [40]. Consequently, although half of the participants in the advisory board were male and we included their perspective in the refinement of the program, the data from this pilot study did not allow us to check whether the program appeals to men and older individuals, whether they feel adequately represented and their needs are met. Hence, extra efforts to recruit these groups might be needed, for example, using stratified sampling (ie, balanced numbers of women and men, and younger and older participants) or highlighting the ease of use to less techsavvy people.

Another limitation is that we did not test accessibility for people with learning disabilities, or cognitive and motor impairments. HautKompass has many visual features, including exercises that require vision, so it may not be well-suited for people with visual impairments. While we reduced the amount of text and aimed to keep its difficulty low, wider user testing is indicated.

Conclusions

This pilot study shows that HautKompass is a highly usable and acceptable web-based intervention, and the data suggests its implementation to be feasible. The results have highlighted several barriers and areas for improvement, which have been addressed where possible. Currently, the program's effectiveness in reducing self-stigma is being tested in an RCT with a 6 months follow-up with 500 participants (n=100 per target diagnosis). Participants are being recruited in hospitals, clinics, and dermatology practices across Germany as well as through patient organizations. We expect HautKompass to reduce self-stigma by improving self-compassion and acceptance in people with various chronic skin diseases, compared to a waitlist control group. If proven effective, we aim to make HautKompass freely available and disseminate it nationwide through patient organizations and over 5000 practicing dermatologists. The program is the first of its kind available in the German language. As psychodermatological services are hardly available in Germany so far, this would be a major advance in psychosocial care with the potential to significantly improve patients' quality of life and well-being.

In addition, various adaptations, extensions, and new applications of HautKompass are possible. As the program has been developed to be skin-generic, it should be tested for other skin diseases. It could also be adapted for visible differences in general with relatively few modifications. In addition, a mobile app would be of great value to give participants more flexibility in time and place, and easier access to do the exercises more frequently. Finally, the current version of HautKompass is aimed at adults, but children and adolescents are still developing their self-image and self-esteem and maybe even more vulnerable to the detrimental effects of self-stigma. Therefore, there is great potential in developing a version for younger patients, parents, or relatives of children affected by skin disease.

XSL•FO
RenderX

Acknowledgments

This study was funded by the German Ministry of Education and Research (01GY2105).

The authors cordially thank all participants of the patient advisory board and the study for their time and valuable input. The authors would also like to thank swiss4ward for developing and designing the HautKompass website. The authors furthermore thank Catharina Braren-von Stülpnagel and Valentina Paucke for proofreading intervention materials and providing feedback as well as for statistical support, Regina von Spreckelsen for sharing her clinical expertise, and the Scientific Communication Team of the Institute for Health Services Research in Dermatology and Nursing, in particular Merle Twesten, for copyediting. Furthermore, we acknowledge financial support from the Open Access Publication Fund of UKE–Universitätsklinikum Hamburg-Eppendorf and DFG–German Research Foundation. No artificial intelligence tools or similar technologies were used in this study or in the preparation of this manuscript.

Data Availability

The datasets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

JT was responsible for conceptualization, project administration, methodology, resources, formal analysis, visualization, and writing–original draft. CFZS and NdSB were responsible for conceptualization, methodology, resources, and writing–review & editing. CS was responsible for conceptualization, methodology, and writing–review & editing. VD and AD were responsible for software, resources, data curation, and writing–review & editing. MR was responsible for methodology, investigation, and writing–review & editing. RS was responsible for funding acquisition, conceptualization, supervision, methodology, and writing–review & editing.

Conflicts of Interest

None declared.

References

1. Sommer R, Augustin M, Mrowietz U, Topp J, Schäfer I, von Spreckelsen R. Perception of stigmatization in people with psoriasis-qualitative analysis from the perspective of patients, relatives and healthcare professionals. Hautarzt 2019 Jul;70(7):520-526. [doi: 10.1007/s00105-019-4411-y] [Medline: 31134287]
2. Alpsoy E, Polat M, FettahlıoGlu-Karaman B, et al. Internalized stigma in psoriasis: a multicenter study. J Dermatol 2017 Aug;44(8):885-891. [doi: 10.1111/1346-8138.13841] [Medline: 28407292]
3. Corrigan PW, Rao D. On the self-stigma of mental illness: stages, disclosure, and strategies for change. Can J Psychiatry 2012 Aug;57(8):464-469. [doi: 10.1177/070674371205700804] [Medline: 22854028]
4. Temel AB, Bozkurt S, Alpsoy E. 029 Internalized stigma in acne vulgaris, vitiligo and alopecia areata. J Investig Dermatol 2017 Oct;137(10):S197. [doi: 10.1016/j.jid.2017.07.125]
5. Krüger C, Schallreuter KU. Stigmatisation, avoidance behaviour and difficulties in coping are common among adult patients with vitiligo. Acta Derm Venereol 2015 May;95(5):553-558. [doi: 10.2340/00015555-1981] [Medline: 25269389]
6. van Beugen S, Maas J, van Laarhoven AIM, et al. Implicit stigmatization-related biases in individuals with skin conditions and their significant others. Health Psychol 2016 Aug;35(8):861-865. [doi: 10.1037/hea0000404] [Medline: 27505208]
7. Topp J, Andrees V, Weinberger NA, et al. Strategies to reduce stigma related to visible chronic skin diseases: a systematic review. J Eur Acad Dermatol Venereol 2019 Nov;33(11):2029-2038. [doi: 10.1111/jdv.15734] [Medline: 31177601]
8. Traxler J, Stuhlmann CFZ, Graf H, Rudnik M, Westphal L, Sommer R. Interventions to reduce skin-related self-stigma: a systematic review. Acta Derm Venereol 2024 Sep 10;104:adv40384. [doi: 10.2340/actadv.v104.40384] [Medline: 39254291]
9. Jungen D, Augustin M, Langenbruch A, et al. Cost-of-illness of psoriasis—results of a German cross-sectional study. J Eur Acad Dermatol Venereol 2018 Jan;32(1):174-180. [doi: 10.1111/jdv.14543] [Medline: 28857297]
10. Kullab J, Schielein MC, Stuhlmann CFZ, et al. Out-of-pocket costs in alopecia areata: a cross-sectional study in German-speaking countries. Acta Derm Venereol 2023 Jan 4;103:adv00838. [doi: 10.2340/actadv.v103.4441] [Medline: 36598280]
11. Mohr N, Naatz M, Zeervi L, et al. Cost-of-illness of atopic dermatitis in Germany: data from dermatology routine care. J Eur Acad Dermatol Venereol 2021 Jun;35(6):1346-1356. [doi: 10.1111/jdv.17203] [Medline: 33651457]
12. Carlbring P, Andersson G, Cuijpers P, Riper H, Hedman-Lagerlöf E. Internet-based vs. face-to-face cognitive behavior therapy for psychiatric and somatic disorders: an updated systematic review and meta-analysis. Cogn Behav Ther 2018 Jan;47(1):1-18. [doi: 10.1080/16506073.2017.1401115] [Medline: 29215315]

XSL•FO
RenderX

13. Barak A, Hen L, Boniel-Nissim M, Shapira N. A comprehensive review and a meta-analysis of the effectiveness of internet-based psychotherapeutic Interventions. J Technol Hum Serv 2008 Jul 3;26(2-4):109-160. [doi: 10.1080/15228830802094429]
14. White V, Linardon J, Stone JE, et al. Online psychological interventions to reduce symptoms of depression, anxiety, and general distress in those with chronic health conditions: a systematic review and meta-analysis of randomized controlled trials. Psychol Med 2022 Feb;52(3):548-573. [doi: 10.1017/S0033291720002251] [Medline: 32674747]
15. Bundy C, Pinder B, Bucci S, Reeves D, Griffiths CEM, Tarrier N. A novel, web-based, psychological intervention for people with psoriasis: the electronic Targeted Intervention for Psoriasis (eTIPs) study. Br J Dermatol 2013 Aug;169(2):329-336. [doi: 10.1111/bjd.12350] [Medline: 23551271]
16. Moshe I, Terhorst Y, Philippi P, et al. Digital interventions for the treatment of depression: a meta-analytic review. Psychol Bull 2021 Aug;147(8):749-786. [doi: 10.1037/bul0000334] [Medline: 34898233]
17. Rothstein JD, Jennings L, Moorthy A, et al. Qualitative assessment of the feasibility, usability, and acceptability of a mobile client data app for community-based maternal, neonatal, and child care in rural Ghana. Int J Telemed Appl 2016:2515420. [doi: 10.1155/2016/2515420] [Medline: 28070186]
18. Monitoring and evaluating digital health interventions: a practical guide to conducting research and assessment. : World Health Organization; 2016 978-92-4-151176-6.
19. Stuhlmann CFZ, Traxler J, Paucke V, da Silva Burger N, Sommer R. Predictors and mechanisms of self-stigma in five chronic skin diseases: a systematic review. J Eur Acad Dermatol Venereol 2025 Mar;39(3):622-630. [doi: 10.1111/jdv.20314] [Medline: 39247975]
20. Germain N, Augustin M, François C, et al. Stigma in visible skin diseases—a literature review and development of a conceptual model. J Eur Acad Dermatol Venereol 2021 Jul;35(7):1493-1504. [doi: 10.1111/jdv.17110] [Medline: 33428316]
21. Gilbert P. The origins and nature of compassion focused therapy. Br J Clin Psychol 2014 Mar;53(1):6-41. [doi: 10.1111/bjc.12043] [Medline: 24588760]
22. Hudson MP, Thompson AR, Emerson LM. Compassion-focused self-help for psychological distress associated with skin conditions: a randomized feasibility trial. Psychol Health 2020 Sep;35(9):1095-1114. [doi: 10.1080/08870446.2019.1707829] [Medline: 31880167]
23. Ruwaard J, Lange A, Schrieken B, Emmelkamp P. Efficacy and effectiveness of online cognitive behavioral treatment: a decade of interapy research. Stud Health Technol Inform 2011;167(9-14):9-14. [Medline: 21685634]
24. Bessell A, Brough V, Clarke A, Harcourt D, Moss TP, Rumsey N. Evaluation of the effectiveness of Face IT, a computer-based psychosocial intervention for disfigurement-related distress. Psychol Health Med 2012 Oct;17(5):565-577. [doi: 10.1080/13548506.2011.647701]
25. Papadopoulos L, Bor R, Legg C. Coping with the disfiguring effects of vitiligo: a preliminary investigation into the effects of cognitive-behavioural therapy. Br J Med Psychol 1999 Sep;72(3):385-396. [doi: 10.1348/000711299160077] [Medline: 10524722]
26. Braun TD, Olson K, Panza E, et al. Internalized weight stigma in women with class III obesity: a randomized controlled trial of a virtual lifestyle modification intervention followed by a mindful self-compassion intervention. Obes Sci Pract 2022 Dec;8(6):816-827. [doi: 10.1002/osp4.616] [Medline: 36483124]
27. Hopkins CM. Reduction of internalized weight bias via mindful self-compassion: theoretical framework and results from a randomized controlled trial [Dissertation]. : Duke University; 2022.
28. Forbes YN, Moffitt RL, Van Bokkel M, Donovan CL. Unburdening the weight of stigma: findings from a compassion-focused group program for women with overweight and obesity. J Cogn Psychother 2020 Nov 1;34(4):336-357. [doi: 10.1891/JCPSY-D-20-00015] [Medline: 33372127]
29. Palmeira L, Cunha M, Pinto-Gouveia J. Processes of change in quality of life, weight self-stigma, body mass index and emotional eating after an acceptance-, mindfulness- and compassion-based group intervention (Kg-Free) for women with overweight and obesity. J Health Psychol 2019 Jul;24(8):1056-1069. [doi: 10.1177/1359105316686668] [Medline: 28810375]
30. Muftin Z, Gilbert P, Thompson AR. A randomized controlled feasibility trial of online compassion-focused self-help for psoriasis. Br J Dermatol 2022 Jun;186(6):955-962. [doi: 10.1111/bjd.21020] [Medline: 35041766]
31. Ewert C, Vater A, Schröder-Abé M. Self-Compassion and coping: a meta-analysis. Mindfulness 2021 May;12(5):1063-1077. [doi: 10.1007/s12671-020-01563-8]
32. Norman A, Veale J, Williamson H. Assessing the usability and acceptability of Face IT@home: an online CBT intervention for people with visible differences. tCBT 2022;15. [doi: 10.1017/S1754470X22000381]
33. Adkins KV, Overton PG, Thompson AR. A brief online writing intervention improves positive body image in adults living with dermatological conditions. Front Med (Lausanne) 2022;9:1064012. [doi: 10.3389/fmed.2022.1064012] [Medline: 36619619]
34. Iliffe LL, Thompson AR. Investigating the beneficial experiences of online peer support for those affected by alopecia: an interpretative phenomenological analysis using online interviews. Br J Dermatol 2019 Nov;181(5):992-998. [doi: 10.1111/bjd.17998] [Medline: 30972732]

XSL•FO
RenderX

35. Krasuska M, Millings A, Lavda AC, Thompson AR. Compassion-focused self-help for skin conditions in individuals with insecure attachment: a pilot evaluation of acceptability and potential effectiveness. Br J Dermatol 2018 Feb;178(2):e122-e123. [doi: 10.1111/bjd.15934] [Medline: 28869774]
36. Kroenke K, Spitzer RL, Williams JBW, Löwe B. An ultra-brief screening scale for anxiety and depression: the PHQ-4. Psychosomatics 2009;50(6):613-621. [doi: 10.1176/appi.psy.50.6.613] [Medline: 19996233]
37. Löwe B, Wahl I, Rose M, et al. A 4-item measure of depression and anxiety: validation and standardization of the Patient Health Questionnaire-4 (PHQ-4) in the general population. J Affect Disord 2010 Apr;122(1-2):86-95. [doi: 10.1016/j.jad.2009.06.019] [Medline: 19616305]
38. Pauley G, McPherson S. The experience and meaning of compassion and self-compassion for individuals with depression or anxiety. Psychol Psychother 2010 Jun;83(Pt 2):129-143. [doi: 10.1348/147608309X471000] [Medline: 19785933]
39. Lara MA, Patiño P, Tiburcio M, Navarrete L. Satisfaction and acceptability ratings of a web-based self-help intervention for depression: retrospective cross-sectional study from a resource-limited country. JMIR Form Res 2022 Apr 4;6(4):e29566. [doi: 10.2196/29566] [Medline: 35377324]
40. Boß L, Lehr D, Reis D, et al. Reliability and validity of assessing user satisfaction with web-based health interventions. J Med Internet Res 2016 Aug 31;18(8):e234. [doi: 10.2196/jmir.5952] [Medline: 27582341]
41. Kuijpers W, Groen WG, Loos R, et al. An interactive portal to empower cancer survivors: a qualitative study on user expectations. Support Care Cancer 2015 Sep;23(9):2535-2542. [doi: 10.1007/s00520-015-2605-0] [Medline: 25876157]
42. Hurmuz MZM, Jansen-Kosterink SM, van Velsen L. How to prevent the drop-out: understanding why adults participate in summative eHealth evaluations. J Healthc Inform Res 2023 Mar;7(1):125-140. [doi: 10.1007/s41666-023-00131-8] [Medline: 36910916]
43. Han A, Kim TH. Effects of self-compassion interventions on reducing depressive symptoms, anxiety, and stress: a meta-analysis. Mindfulness 2023 Jun 5;2023:1-29. [doi: 10.1007/s12671-023-02148-x] [Medline: 37362192]
44. Pedersen DH, Mansourvar M, Sortsø C, Schmidt T. Predicting dropouts from an electronic health platform for lifestyle interventions: analysis of methods and predictors. J Med Internet Res 2019 Sep 4;21(9):e13617. [doi: 10.2196/13617] [Medline: 31486409]
45. Melville KM, Casey LM, Kavanagh DJ. Dropout from Internet-based treatment for psychological disorders. Br J Clin Psychol 2010 Nov;49(Pt 4):455-471. [doi: 10.1348/014466509X472138] [Medline: 19799804]
46. Christensen H, Griffiths KM, Farrer L. Adherence in internet interventions for anxiety and depression. J Med Internet Res 2009 Apr 24;11(2):e13. [doi: 10.2196/jmir.1194] [Medline: 19403466]
47. Jakob R, Harperink S, Rudolf AM, et al. Factors influencing adherence to mHealth apps for prevention or management of noncommunicable diseases: systematic review. J Med Internet Res 2022 May 25;24(5):e35371. [doi: 10.2196/35371] [Medline: 35612886]
48. Van der Mispel C, Poppe L, Crombez G, Verloigne M, De Bourdeaudhuij I. A self-regulation–based eHealth intervention to promote a healthy lifestyle: investigating user and website characteristics related to attrition. J Med Internet Res 2017 Jul 11;19(7):e241. [doi: 10.2196/jmir.7277] [Medline: 28698168]

Abbreviations

CBT: cognitive behavioral therapy
BMBF: Bundesministerium für Bildung und Forschung (German Federal Ministry for Education and Research)
CSQ: Client Satisfaction Questionnaire Adapted to Internet-Based Interventions
PHQ-4: Patient Health Questionnaire-4
RCT: randomized controlled trial

Edited by A Mavragani; submitted 19.12.24; peer-reviewed by F Balieva, M Boarini; revised version received 27.02.25; accepted 02.03.25; published 22.05.25.

Please cite as:
Traxler J, Stuhlmann CFZ, da Silva Burger N, Stierle C, Djamei V, Darzina A, Rudnik M, Sommer R
Self-Stigma in Adults Living With Chronic Skin Disease: Development of the HautKompass Web-Based Program and Pilot Test of Its Usability, Acceptability, and Feasibility
JMIR Form Res 2025;9:e70290
URL: https://formative.jmir.org/2025/1/e70290
doi:10.2196/70290

© Juliane Traxler, Caroline F Z Stuhlmann, Neuza da Silva Burger, Christian Stierle, Vahid Djamei, Anna Darzina, Marie Rudnik, Rachel Sommer. Originally published in JMIR Formative Research (https://formative.jmir.org), 22.5.2025. This is an open-access

XSL•FO
RenderX

article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Evaluating the Acceptability and Utility of a Personalized Wellness App (Aspire2B) Using AI-Enabled Digital Biomarkers: Engagement Enhancement Pilot Study

Calissa J Leslie-Miller[1], MS; Shellen R Goltz[2], PhD, RD; Pamela L Barrios[2], PhD; Christopher C Cushing[1,3], PhD; Teena Badshah[2], MS, DDPHN; Corey T Ungaro[2], PhD; Shankang Qu[4], PhD; Yulia Berezhnaya[2], PhD; Tristin D Brisbois[2], PhD, MBS

[1]Department of Clinical Child Psychology, University of Kansas, Lawrence, KS, United States
[2]Life Sciences, Global R&D, PepsiCo Inc, 700 Anderson Hill Rd, Purchase, NY, United States
[3]Life Span Institute, University of Kansas, Lawrence, Kansas, United States
[4]Data Science and Analytics, R&D, PepsiCo Inc, Valhalla, NY, United States

Corresponding Author:
Tristin D Brisbois, PhD, MBS
Life Sciences, Global R&D, PepsiCo Inc, 700 Anderson Hill Rd, Purchase, NY, United States

Abstract

Background: There is significant global interest in promoting wellness, with digital solutions like mobile health apps being broadly downloaded; yet, there is a challenge in maintaining engagement for long-term behavior change. Developing a widely accepted mobile wellness app is imperative for advancing personalized wellness interventions.

Objective: The primary objective of this study was to evaluate the Aspire2B wellness app (powered by Salus Optima), designed to exceed industry standards for participant engagement by incorporating evidence-based behavior change strategies and to assess its acceptability (eg, liking the face scan) and utility (eg, willing to use the face scan technology for other health insights) as a digital health solution.

Methods: Participants aged 18-65 years, who were smartphone and fitness tracker users, were recruited in the United States during March-May 2022. Participants received US $5 compensation for downloading the app, with no further incentive for usage. Following completion of onboarding (ie, survey questions about lifestyle behaviors), participants were placed in either a nutrition, sleep, or fitness 4-week challenge. During the challenge, participants used various app features at their own will, such as a facial scan for wellness insights (eg, heart rate and biological age), recipes, and workout videos. These interactions with the app were cumulatively evaluated as engagement metrics. Participants were also asked to answer offboarding questions to evaluate any changes to lifestyle behaviors and experience using the app features (eg, acceptability of face scan experience).

Results: Out of the 398 people who created an account, 85.9% (342/398) completed onboarding and a face scan. Following this, 74.9% (298/398) of users completed additional survey questions about current wellness behaviors. Notably, interaction with the app was relatively stable from week 2 to 4 (173/398, 43.5%), outperforming industry standards by roughly 3×. In addition, on average, participants completed 2.1 - 2.7 face scans per week, with approximately 7% (24/342) of participants maintaining regular use of face scan technology for 4 weeks. In users who completed offboarding questions, 88.8% (111/125) found Aspire2B credible, 64.8% (81/125) liked the face scan experience, 7.2% (9/125) disliked the face scan experience, and 83.2% (104/125) said they would use face scan technology for other insights into their health.

Conclusions: These findings highlight strong initial engagement with Aspire2B, followed by significant sustained user engagement over a 4-week period. Furthermore, users indicated high levels of credibility and willingness to use face scan technology for wellness insights. These findings collectively demonstrate the capability of a personalized wellness app using AI-enabled digital biomarkers and evidence-supported behavior change techniques to yield positive user perception and provide long-term engagement.

Trial Registration: ClinicalTrials.gov NCT06191822; https://clinicaltrials.gov/study/NCT06191822

(JMIR Form Res 2025;9:e63471) doi:10.2196/63471

KEYWORDS

wellness application; development protocol; AI-enabled digital biomarkers; acceptability; nutrition; sleep; fitness; mobile phone

XSL•FO
RenderX

Introduction

Background

Around the world, there is substantial and widespread interest in the promotion of wellness. Digital solutions are a common wellness tool for consumers, with over 300 million people using a mobile health app in 2023 [1]. This trend presents an important opportunity for managing population health as there continues to be suboptimal uptake of physical activity, healthy diet, and adequate sleep, which are the most important health behaviors for wellness and longevity [2]. Encouragingly, there exists substantial evidence that basic behavior change strategies can effectively increase the uptake of wellness behaviors [3]. However, it is clear that the majority of wellness app users quickly disengage and stop using the app with novel features providing only a small boost to engagement and retention long-term [3].

Furthermore, for mobile apps to have a meaningful effect on well-being by driving behavior change, it is crucial to pinpoint enduring technologies that will foster engagement over an extended period, allowing sufficient time for behavior shifts to take place [4]. Mobile apps contain 2 classes of features, those that are designed to change behavior and those that are designed to promote engagement. The most common and best supported strategies for behavior change are self-monitoring of performance, tools for facilitating setting effective goals, and feedback mechanisms [5,6]. It is less clear which app features are useful for promoting engagement (trophies, badges, or points within the app), and there is insufficient evidence that they lead to long-term engagement (the field lacks high-quality randomized controlled trials of app engagement strategies). One way to potentially increase engagement is by delivering novel health information using technologies that are intriguing or surprising to users.

Over the last 50 years, a body of evidence has emerged that characterizes the biological factors underpinning the aging process. The notion of biological age is captivating as it offers individuals an estimation of their aging status that transcends chronological age and perceived age [7]. Providing consumers with a biological age that differs from their chronological age has the potential to stimulate behavioral shifts [8]. Specifically, if information reveals that one's biological age is advancing faster than their chronological age, these individuals may be motivated to adopt behavior changes aimed at bringing the two into closer alignment.

Historically, the science of biological age was limited to high throughput omics research and was mainly accessible to researchers or specialized health care providers [9]. However, it is now possible to estimate an individual's biological age by using computer vision combined with artificial intelligence models [10]. This advancement brings a cutting-edge scientific concept directly to consumers, who are already familiar with activities such as using their phone's camera for tasks like facial scans to unlock their iPhone (Apple Inc) [11]. As barriers to accessing personal information about biological age are reduced, the field needs a series of rigorous studies to assess the feasibility and acceptability of apps that leverage the appeal of biological age to drive and maintain engagement in wellness programs, ultimately leading to positive changes in wellness behavior.

Objectives

Our study aimed to address several key questions. First, we explored who uses the app by examining participants' characteristics. We also investigated whether users find the Aspire2B app, including its content and features, credible and valuable for providing wellness insights. Another focus was determining whether users remain engaged with the app over time. Finally, we examined which participant characteristics were related to long-term engagement with the app. Given that the app was grounded in behavior change science, we hypothesized that the majority of participants would find Aspire2B credible and enjoyable, use it for wellness insights, and recommend the app to others. Furthermore, we hypothesized Aspire2B retention rates would exceed industry standards, such that more than 20% of users that create an account, complete onboarding, more than 9% interact with the app in week 2, and more than 4% interact with the app in week 4 [12]. In addition, we predicted that participants who were informed of a biological age value higher than their chronological age would be more likely to engage in the app (ie, serving as a nudge or "wake-up" call to engage in more healthful lifestyle behaviors). Finally, we were interested in exploring participant characteristics that would predict higher engagement with the app without a priori predictions, as well as whether improvements in wellness metrics could be detected following program completion.

Methods

Participant Recruitment

Participant recruitment was led by Human8, who used one of their partner's panels to achieve the desired target sample in the United States during March-May 2022. The partner panel consists of willing participants, older than 18 years, who have regular internet access, and have opted in to be part of a database that allows them to voice their opinions with the opportunity to be incentivized. These panelists were invited to complete a recruitment questionnaire to determine eligibility for this study. Additional inclusion criteria required participants to be 18 - 65 years old and report that they were smartphone and fitness tracker users. Those who met the inclusion criteria were provided with a link to download the Aspire2B app.

Ethical Considerations

This study was reviewed and approved by the Sterling institutional review board. Participants were presented information about informed consent, after downloading the app, through a screen that occurred before onboarding. Participants were compensated $5 for downloading the application, with no further incentives for usage (refer to Participant Recruitment section). Images used in this paper do not identify real participants and data analysis was completed on deidentified data.

XSL•FO
RenderX

Intervention

Overview

Aspire2B, powered by Salus Optima, is a personalized wellness app (Figure 1) that is rooted in evidence-supported techniques that facilitate positive behavioral shifts (eg, improved sleep quality, nutrition, and exercise). After downloading the app, users create a profile (refer to Required Onboarding Questions section), identify their personal wellness goals, and complete face scans that immediately display wellness metrics (ie, biological age, heart rate, blood pressure, and stress index). Biological age score was calculated through the Gero AI algorithm, a machine learning framework [10], and their blood pressure, heart rate, and stress index were estimated using DeepAffex technology (developed by NuraLogix) [13]. To further personalize the experience of Aspire2B, participants were encouraged to answer more questions covering aspects like sleep quality, physical activity, and dietary intake, although these were not mandatory (refer to Optional Onboarding Questions section). From these responses, subjects were assigned to one of three challenges, that are (1) nutrition, (2) movement, (3) sleep based on set criteria (eg, those reporting poor sleep quality were placed in sleep challenge). Those who skipped these questions were prompted to complete them later, with the default challenge assigned as nutrition. Each participant set 28-day goals to guide their progress. Furthermore, Aspire2B streamlines the user experience through integration with Apple Health and Google Fit, enhancing personalized feedback and progress tracking.

Figure 1. User flow diagram for participants completing a 4-week pilot study using Aspire2B, a personalized wellness app.

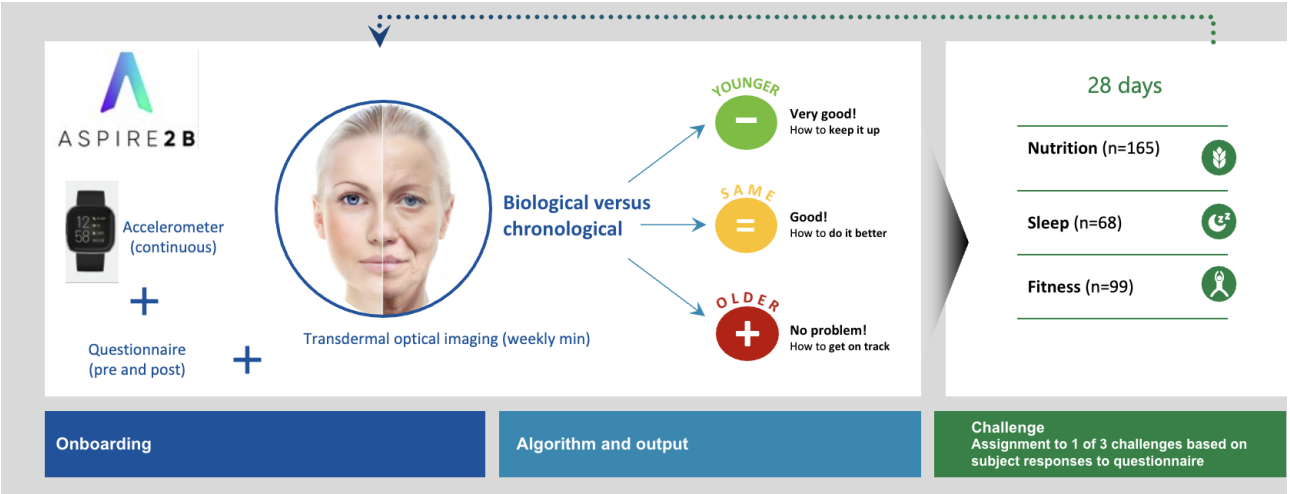

Educational Content

The app content included a variety of tools to help users work toward their personal wellness goals. Participants could access educational tools at any time, such as exercise videos, articles, tips, and healthy recipes. All articles and tips were grounded in science and supported by peer-reviewed publications (Multimedia Appendix 1).

Wellness Tracking

Participants had constant access in the app to a face scan that could provide them with a biological age score, heart rate, blood pressure, and stress index. The DeepAffex Anura face scan technology uses transdermal optimal imaging via photoplethysmography to detect facial blood flow patterns. With a conventional video camera and a cloud-based Affective AI engine, it is possible to predict various health metrics from a scan. The mean difference between the Anura measurement and its paired auscultatory reference measurement has been found to be within acceptable limits [14,15]. Participants could also monitor their step count in the app, which was automatically imported from the user's device into Aspire2B, and sleep time and quality via self-report (ie, number of hours and sleep quality on a 5-point scale with emojis as anchors). Participants were also able to rate their mood on a 9-point emoji scale (adapted from study by Desmet et al [16]) to monitor changes in mood. Participants were assigned a wellness score, which was generated from participant progress on step count, sleep (time and quality), and mood. Each component was weighted in its contribution to the wellness score and required meeting the expert identified targets (eg, 10,000 steps per day).

Notifications and Awards

Aspire2B attempts to take a gamified approach to acknowledge achievements and encourage sustained effort. Digital badges are used to recognize the attainment of subgoals or the unlocking of new levels. To promote user engagement, Aspire2B uses a tailored reminder system that allows users to set or customize reminders. While these features are present, they were not refined or optimized for this trial as the intention of Aspire2B was to serve as a minimum viable product to test the interest in the biological age concept and acceptance of face scan technology.

Measures

Required Onboarding Questions

Demographics

Participants were asked demographic questions, such as age, gender, race, and income.

BMI

Participants were asked their height and weight to be able to calculate BMI.

XSL•FO
RenderX

General Health

Participants were asked how old they felt, to rate their health on a 5-point scale from excellent to poor, and to list how many days their physical health was not good over the last 30 days.

Physical Activity

Participants were asked to indicate whether they have engaged in moderate and vigorous physical activities in the past 30 days.

Oat Intake

Participants were asked how many servings of oats they ate per week.

Mood

Participants were asked to indicate feelings of little energy and little pleasure over the past 2 weeks on a scale from 1 (not at all) to 4 (nearly every day).

Optional Onboarding Questions

Sleep Quality

Participants were asked to rate their sleep quality over the past 7 days on a scale from 0 to 10.

Eating Behavior

Participants were asked to indicate how many servings per day they had vegetables, fruit, sugar-sweetened beverages, and whole grains, as well as how many servings they had per week of beans or lentils, fish, nuts, seeds or nut butter, and red or processed meat.

Movement

Participants were asked to describe their usual daily activities with respect to sitting, standing, lifting, and stair climbing.

Alcohol Consumption

Participants were asked to indicate how many drinks they have in a typical week.

Smoking

Participants were asked to indicate if they currently smoke cigarettes and if yes, whether they smoked more than 1 pack per day.

Optional Offboarding Questions

Health Behavior

Participants were asked to repeat questions from the onboarding including body weight, general health, physical activity and movement, mood, sleep, eating behavior including oat intake, and alcohol and cigarette use.

Credibility

Participants were asked to respond “yes” or “no” to the question: “Do you find Aspire2B credible?”

General Experience

Participants were asked to indicate how likely they would be to recommend Aspire2B to others on a scale of 1 (not recommend) to 10 (strongly recommend).

Face Scan Experience

Participants were asked if they liked the face scan experience on a scale of 1 (strongly disagree) to 4 (strongly agree).

Face Scan Insights

Participants were asked to answer “yes” or “no” for if they would use the face scan technology to obtain other insights into their health.

Data Analyses

All data were analyzed using IBM SPSS Statistics (Version 28), JMP Pro Software (Version 17.2) or Jamovi (Version 2.3.21.0). When not specified otherwise, analyses used the full dataset available, with baseline demographic variables (eg, age) derived from enrollment data and engagement metrics analyzed as cumulative measures independent of specific time point. Descriptive statistics were used to summarize and describe the data. Comparisons between Aspire2B participants and the general US population were completed qualitatively using data from the National Health and Nutrition Examination Survey (NHANES) and Behavioral Risk Factor Surveillance System (BRFSS) survey, using the most recently collected data available at the time of analysis (2017 - 2018 for NHANES [17] and 2021 for BRFSS [18]). BRFSS was used for comparisons of age group, gender, race or ethnicity, US census region, employment status, education, income, BMI, smoking status, and general health. NHANES was used for alcohol, physical activity, and days of poor physical or mental health. For BRFSS, analysis was restricted to individuals reporting owning a cell phone and who were in the same age range as Aspire2B participants. For NHANES, data on cell phone ownership were not available, but the population was restricted to the same age range. To identify participant characteristics related to engagement, Pearson correlations were conducted. In order to inform future data driven algorithms, additional exploratory general linear models were conducted to identify if the 4-week program led to any lifestyle changes for participants, and whether there was an interaction with age.

Results

Participant Characteristics

Participants were aged 20 - 65 (mean 40.15, SD 9.59) years. The majority of participants were White females with a bachelor’s degree or above (Table 1). Based on participant responses to onboarding questions, we found that, as compared with the BRFSS [19] or the NHANES [20], participants had similar BMI distributions, but were more likely to consume alcohol frequently. Participants were also more likely to report feeling tired and little pleasure in doing things and were less likely to report “excellent” self-rated health. However, participants were more likely to report engaging in moderate or vigorous physical activities and were less likely to be smokers. At onboarding, participants in each challenge (nutrition, movement, and sleep) were similar regarding height, weight, and BMI, but were statistically different in terms of age, such that those in the sleep challenge had significantly lower chronological age (P=.02).

XSL•FO
RenderX

Table . Demographics for participants who downloaded the Aspire2B app. Participant ages have been grouped in this table for clarity and ease of interpretation, but they are not grouped in the analyses. In addition, demographic data are missing for 16 people.

Characteristics		All participants, n (%)	Nutrition challenge, n (%)	Movement challenge, n (%)	Sleep challenge, n (%)
Age group (y)					
	18 - 26	19 (5)	7 (1.8)	4 (1)	7 (1.8)
	26 - 35	112 (29.3)	43 (11.3)	33 (8.6)	21 (5.5)
	36 - 45	142 (37.2)	64 (16.8)	33 (8.6)	31 (8.1)
	46 - 55	79 (20.7)	36 (9.4)	22 (5.8)	7 (1.8)
	56 - 65	30 (7.9)	15 (3.9)	9 (2.4)	3 (0.8)
Sex					
	Male	105 (27.5)	42 (11)	21 (5.5)	21 (5.5)
	Female	277 (72.5)	123 (32.2)	80 (20.9)	48 (12.6)
Race or ethnicity					
	White or Caucasian	277 (72.5)	115 (30.1)	74 (19.4)	53 (13.9)
	Black or African American	37 (9.7)	14 (3.7)	9 (2.4)	7 (1.8)
	Hispanic or Latinx	35 (9.2)	17 (4.5)	8 (2.1)	7 (1.8)
	Native American or Alaska Native	1 (0.3)	0 (0)	1 (0.3)	0 (0)
	Multiracial or Biracial	12 (3.1)	7 (1.8)	3 (0.8)	1 (0.3)
	A race or ethnicity not listed here	20 (5.2)	12 (3.1)	6 (1.6)	1 (0.3)
Employment status					
	Employed full-time	225 (58.9)	90 (23.6)	59 (15.4)	44 (11.5)
	Employed part-time	41 (10.7)	25 (6.5)	8 (2.1)	4 (1)
	Self-employed	21 (5.5)	3 (0.8)	10 (2.6)	5 (1.3)
	Unemployed	21 (5.5)	11 (2.9)	3 (0.8)	7 (1.8)
	Homemaker	57 (14.9)	29 (7.6)	15 (3.9)	8 (2.1)
	Retired	7 (1.8)	4 (1)	2 (0.5)	0 (0)
	Student	10 (2.6)	3 (0.8)	4 (1)	1 (0.3)
Education					
	High-school degree or GED	34 (8.9)	15 (3.9)	6 (1.6)	8 (2.1)
	Some college but no degree	67 (17.5)	28 (7.3)	18 (4.7)	13 (3.4)
	Associate degree (eg, AA and AS)	47 (12.3)	25 (6.5)	11 (2.9)	8 (2.1)
	Bachelor's degree (eg, BA, BBA, and BS)	164 (42.9)	69 (18.1)	47 (12.3)	28 (7.3)
	Master's degree (eg, MA, MS, and MEng)	53 (13.9)	22 (5.8)	13 (3.4)	11 (2.9)
	Doctorate (eg, PhD and EdD)	7 (1.8)	3 (0.8)	0 (0)	1 (0.3)
	Professional degree (eg, MD, DDS, and JD)	10 (2.6)	3 (0.8)	6 (1.6)	0 (0)
Income (US $)					
	Less than 29,000	22 (5.8)	11 (2.9)	3 (0.8)	6 (1.6)

XSL•FO
RenderX

Characteristics		All participants, n (%)	Nutrition challenge, n (%)	Movement challenge, n (%)	Sleep challenge, n (%)
	30,000-49,999	54 (14.1)	19 (5)	16 (4.2)	11 (2.9)
	50,000-74,999	88 (23)	35 (9.2)	22 (5.8)	21 (5.5)
	75,000-99,999	76 (19.9)	34 (8.9)	24 (6.3)	12 (3.1)
	100,000-150,000	92 (24.1)	41 (10.7)	21 (5.5)	17 (4.5)
	Over 150,000	50 (13.1)	25 (6.5)	15 (3.9)	2 (0.5)

Acceptability

The majority (111/125, 88.8%) of users who responded to the offboarding questions found Aspire2B credible. In addition, on a scale of 1 to 10, participants would generally recommend Aspire2B to others (mean 6.91, SD 2.45). Only 7.2% (9/125) of users disliked the face scan experience, with 64.8% (81/125) who liked the face scan experience and 83.2% (104/125) said they would use face scan technology for other insights into their health.

Retention

Out of the 398 people who created an account (Figure 2), 85.9% (342/398) completed onboarding and a face scan. Following this, 74.9% (298/398) of users completed the additional survey questions. Notably, interaction with the app was the same during weeks 2 and 4 (final week) of the program (173/398, 43.5%), outperforming industry standards [12] by roughly 3× (Table 2). The majority of participants were placed in the nutrition challenge (n=165), followed by the movement challenge (n=99) and sleep challenge (n=68). Similarly, retention rates were highest for the nutrition intervention (93/165, 56.3%), followed by exercise (49/99, 49.5%) and sleep interventions (27/68, 39.7%), though differences between groups were not significantly different. On average, participants completed 2.1 - 2.7 face scans per week, with approximately 7% (24/342) of participants maintaining regular use of face scan technology for 4 weeks.

XSL•FO
RenderX

Figure 2. CONSORT (Consolidated Standards of Reporting Trials) flow diagram.

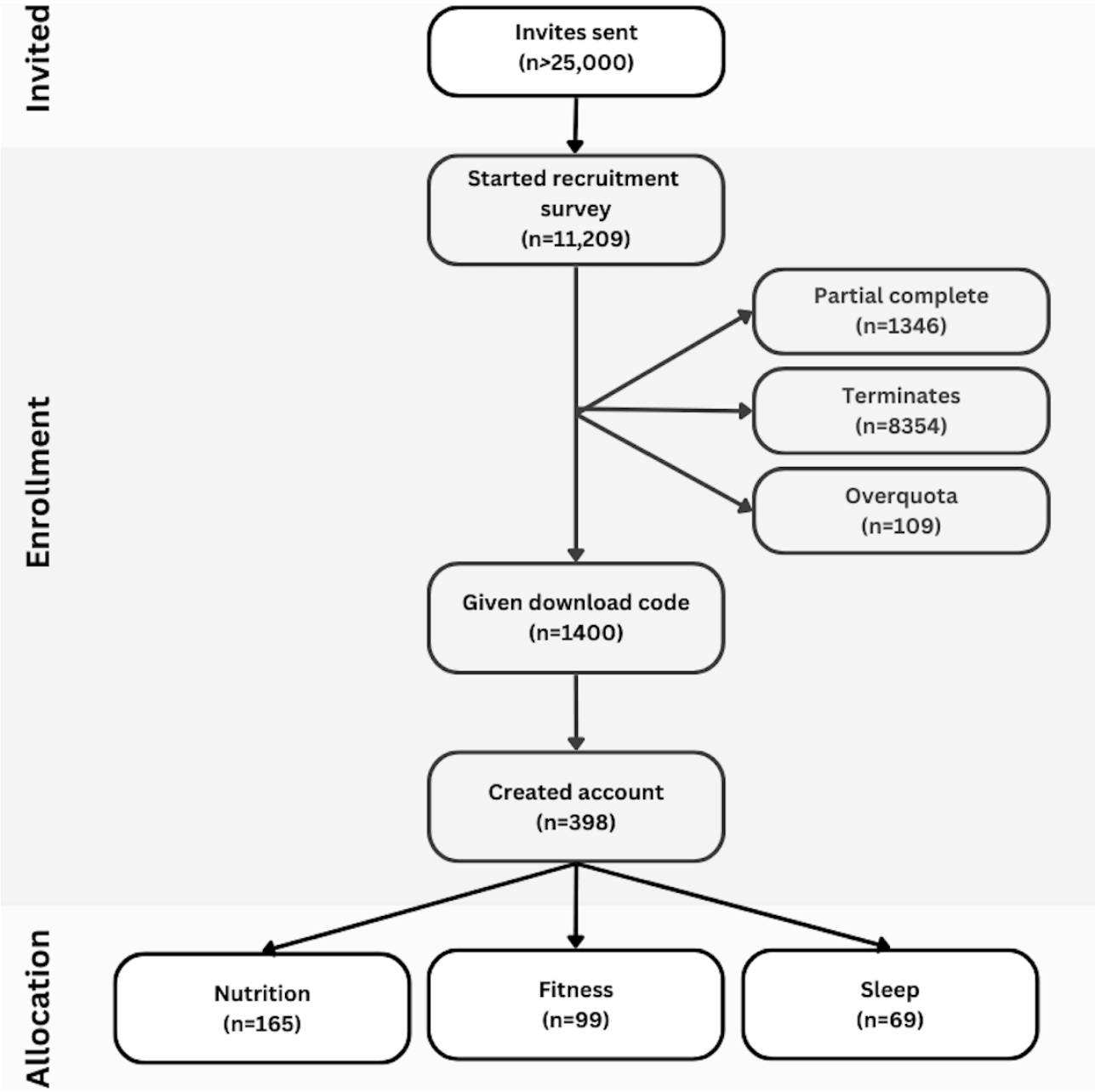

Table . Retention rates throughout a 4-week pilot study evaluating acceptability and utility of Aspire2B, a personalized wellness app, as compared with industry standards for wellness apps [12]. Week 2 and 4 interaction for Aspire2B overall percentage engagement were determined by comparing those who interacted with the app during that week with those who created an account. For the individual nutrition, fitness, and sleep challenges, percent engagement was determined by comparing those who interacted with the app during a specific week with those who completed onboarding and were assigned to a challenge. In addition, completing onboarding includes completion of the face scan for Aspire2B.

Enrollment stage	Industry standards, %	Aspire2B overall, %	Nutrition challenge, %	Fitness challenge, %	Sleep challenge, %
Created account	__[a]	398	__[b]	__[b]	__[b]
Completed onboarding	20	86	41	25	17
Week 2 interaction	9	43	60	58	47
Week 4 interaction	4	43	56	49	40
Offboarding completed	__[a]	33	47	38	26

[a]Not available.

[b]Not applicable.

XSL•FO
RenderX

Participant Characteristics Related to Engagement

Age was correlated with engagement across many domains. More specifically, older participants were more likely to log their sleep (r_{382}=0.21, P<.001) and mood (r_{382}=0.22, P<.001), as well as open tips (r_{382}=0.21, P<.001), view their biological age (r_{382}=0.18, P<.001), complete face scans (r_{382}=0.14, P=.005), and finish app challenges (r_{382}=0.15, P=.004).

Exploratory Analyses

It is noteworthy that we found an interaction between chronological age and biological age as a predictor of the number of challenges completed (P=.02) and total workouts (P=.04). Interaction probing [21] suggests that for older individuals, a higher discrepancy for biological age (biological>chronological) is predictive of overall challenge completion and total workout number. On the other hand, for younger adults a similar higher discrepancy is related to fewer challenges completed and total workouts. While other engagement metrics are not predicted by this discrepancy, there is potential to encourage behavior change through interest in this concept. Furthermore, biological age and wellness scores did improve for participants who remained engaged over the 4-week program (P=.02 and P=.001, respectively). There appeared to be greater improvements in wellness scores for older participants (P<.001). Lifestyle factors such as oats consumption and sleep quality improved over the 4-week program (P<.001).

Discussion

Principal Results

Participants who used Aspire2B were 20 - 65 years old (with 36 - 45 y being the largest group), predominantly White females with a bachelor's degree or higher. Most participants found Aspire2B credible and were interested in using the app for health insights. Furthermore, the majority of participants completed the onboarding process, including the face scan. As hypothesized, engagement remained relatively high throughout the study. The majority of participants joined the nutrition challenge, which also had the highest retention rates, followed by the movement and sleep challenges. On average, participants completed a face scan multiple times per week, with a small but consistent group maintaining regular use throughout the 4 weeks. Interestingly, age was correlated with engagement, such that older participants were more engaged in the app across many domains.

Retention rates significantly outperformed industry standards [12]. Participants were placed in 1 of 3 challenges based on their needs. Since even modest levels of personalization of programs [22] have shown increased engagement, it is plausible that the personalized challenge impacted overall engagement. This study also revealed high acceptability among users, with the majority finding Aspire2B credible and would recommend it to others. This suggests that users perceive Aspire2B as a trustworthy and credible platform, which is essential for ensuring user engagement and satisfaction [23]. Furthermore, users self-reported improvements in various lifestyle aspects, offering encouragement for the Aspire2B platform and face scan technology as effective nudging tools.

Biological age itself appears to be a construct that participants find engaging and relevant, as evidenced by their interaction with this feature, particularly among older adults. Notably, the objective was not to evaluate the algorithm for accuracy, but rather to test the concept to determine whether it is appealing and effective in influencing behavior. Older participants revisited biological age more frequently. This is useful information for future work as it indicates a subpopulation of individuals for whom biological age may be salient and may be receptive to setting behavioral goals linked to biological age. For these individuals, a discrepancy between biological age and chronological age may also be motivating for engaging in health behaviors. Importantly, technology (digital application and face scan technology) was found acceptable for older users, offering a viable solution to nudge individuals toward better lifestyle choices.

In addition, patterns were found regarding predictors of user engagement. Higher chronological age correlated with greater engagement across various domains. Notably, those in the sleep arm had significantly lower chronological age, which may explain the higher attrition rate. Older users were more likely to log their sleep and mood, as well as open tips, view their biological age, complete face scans, and finish app challenges. It is unclear whether these behaviors are linked to attempts to manage biological age directly or whether older users are more likely to engage with Aspire2B generally, a concept requiring further testing.

Limitations and Future Directions

This study is not without limitations. Employees of consumer package goods companies were not excluded from the study. Although it was not openly disclosed that the app was developed by a multinational food and beverage corporation, this information was available in the terms and conditions of the app, which may have introduced bias into the study results. In addition, this study was US-specific and may not generalize to global apps. Future investigation into the feasibility and effectiveness of Aspire2B in a global context, considering cultural and regional variations, is necessary. Furthermore, selection bias may be present, as participants who opted to participate in this study may differ systematically from those who declined, but we cannot assess the extent of this bias since we do not have data from nonparticipants. Markedly, the wellness app in this study serves as a rudimentary platform for hypothesis testing and may be used to improve engagement required in features for a commercial launch.

It may also be beneficial to investigate the impact of increased personalization through the integration of machine learning techniques to determine if they can lead to higher and long-term user engagement. Further personalization by segmenting consumers based on their values and behaviors may also offer increased tailored experiences that improve overall success rates. There may also be interesting differences in user engagement and retention between apps that offer a continuous wellness program and those that follow a finite trial mode, such as the study completed here. Finally, the addition of more

XSL•FO
RenderX

features, such as more physiological results revealed through face scan, may also increase user acquisition and retention. Future studies may also benefit from trying to tease out if the face scan technology is the draw or if high engagement was due to attempting to improve one's biological age through the set program. These future directions can help refine the strategies used in wellness apps, ultimately promoting long-term engagement and effectively fostering healthier behaviors in users.

Conclusions

Overall, this study highlights the potential of Aspire2B as a credible and engaging wellness app, particularly for older adults, as engagement levels were positively correlated with age across various features. This study provides a foundation for further exploration of personalized challenges, biological age feedback, and user-friendly face scan technology to drive engagement, and therefore, potentially improvements in lifestyle behavior. Importantly, the high retention rates indicate a promising model for future wellness technologies.

Acknowledgments

To improve readability, we used AI tools exclusively for grammar and style review.

Data Availability

Deidentified data from this study are not available in a public archive. Deidentified data from this study will be made available (as allowable according to IRB) by emailing the corresponding author.

Authors' Contributions

PLB and TDB were responsible for the concept, design of the study, and IRB approval. PLB, YAB, TB, and TDB were responsible for Aspire2B application content. CJL-M, SQ, and CTU were responsible for data analysis. CJL-M, SRG, CCC, and TDB drafted the original manuscript. All authors have reviewed and approved the final manuscript.

Conflicts of Interest

SRG, PLB, YAB, TB, CTU, and TDB, are employees of PepsiCo, Inc. CJL-M receives funds from PepsiCo in the form of a Graduate Research Assistantship. CCC receives consulting funds from PepsiCo. The views expressed in this manuscript are those of the authors and do not necessarily reflect the position or policy of PepsiCo, Inc.

Multimedia Appendix 1
Application educational content.
[DOCX File, 28 KB - formative_v9i1e63471_app1.docx]

References

1. Health app revenue and usage statistics. Business of Apps. 2024. URL: https://www.businessofapps.com/data/health-app-market/ [accessed 2024-05-16]
2. Promoting health for adults | CDC. URL: https://www.cdc.gov/pcd/collections/pdf/Policy-Collection-W3_508c.pdf [accessed 2022-09-08]
3. Baumel A, Muench F, Edan S, Kane JM. Objective user engagement with mental health apps: systematic search and panel-based usage analysis. J Med Internet Res 2019 Sep 25;21(9):e14567. [doi: 10.2196/14567] [Medline: 31573916]
4. König LM, Attig C, Franke T, Renner B. Barriers to and facilitators for using nutrition apps: systematic review and conceptual framework. JMIR Mhealth Uhealth 2021 Apr 1;9(6):e20037. [doi: 10.2196/20037] [Medline: 34254938]
5. Yang CH, Maher JP, Conroy DE. Implementation of behavior change techniques in mobile applications for physical activity. Am J Prev Med 2015 Apr;48(4):452-455. [doi: 10.1016/j.amepre.2014.10.010] [Medline: 25576494]
6. Brannon EE, Cushing CC. A systematic review: is there an app for that? Translational science of pediatric behavior change for physical activity and dietary interventions. J Pediatr Psychol 2015 May;40(4):373-384. [doi: 10.1093/jpepsy/jsu108] [Medline: 25502745]
7. Jylhävä J, Pedersen NL, Hägg S. Biological age predictors. EBioMedicine 2017 Jul;21:29-36. [doi: 10.1016/j.ebiom.2017.03.046] [Medline: 28396265]
8. Levine ME. Modeling the rate of senescence: can estimated biological age predict mortality more accurately than chronological age? J Gerontol A Biol Sci Med Sci 2013 Jun;68(6):667-674. [doi: 10.1093/gerona/gls233] [Medline: 23213031]
9. Rutledge J, Oh H, Wyss-Coray T. Measuring biological age using omics data. Nat Rev Genet 2022 Dec;23(12):715-727. [doi: 10.1038/s41576-022-00511-7] [Medline: 35715611]
10. Hacking complex diseases and aging. GERO.AI. URL: https://gero.ai [accessed 2024-01-25]
11. Use face ID on your iphone or ipad pro. Apple Support. URL: https://support.apple.com/en-us/108411 [accessed 2024-11-21]

XSL•FO
RenderX

12. Ceci L. Mobile app user android retention rate worldwide Q3 2024, by category. Statista. URL: https://www.statista.com/statistics/259329/ios-and-android-app-user-retention-rate/ [accessed 2025-04-24]
13. DeepAffexTM cloud API. Apiary. URL: https://dfxapiversion10.docs.apiary.io [accessed 2024-03-19]
14. Luo H, Yang D, Barszczyk A, et al. Smartphone-based blood pressure measurement using transdermal optical imaging technology. Circ Cardiovasc Imaging 2019 Aug;12(8):e008857. [doi: 10.1161/CIRCIMAGING.119.008857] [Medline: 31382766]
15. Yang D, Xiao G, Wei J, Luo H. Preliminary assessment of video-based blood pressure measurement according to ANSI/AAMI/ISO81060-2: 2013 guideline accuracy criteria: Anura smartphone app with transdermal optimal imaging technology. Blood Press Monit 2020 Oct;25(5):295-298. [doi: 10.1097/MBP.0000000000000467] [Medline: 32842022]
16. Desmet PMA, Vastenburg MH, Romero N. Mood measurement with Pick-A-Mood: review of current methods and design of a pictorial self-report scale. JDR 2016;14(3):241. [doi: 10.1504/JDR.2016.079751]
17. NHANES 2017-2018 overview. Centers for Disease Control and Prevention. URL: https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/overview.aspx?BeginYear=2017 [accessed 2025-01-07]
18. BRFSS survey data and documentation. Centers for Disease Control and Prevention. 2023. URL: https://www.cdc.gov/brfss/annual_data/annual_2021.html [accessed 2025-01-07]
19. Behavioral risk factor surveillance system. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/brfss/index.html [accessed 2023-08-09]
20. National Health and Nutrition Examination Survey. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/nchs/nhanes/?CDC_AAref_Val=https://www.cdc.gov/nchs/nhanes/index.htm [accessed 2023-08-09]
21. Bauer DJ, Curran PJ. Probing interactions in fixed and multilevel regression: inferential and graphical techniques. Multivariate Behav Res 2005;40(3):373-400. [doi: 10.1207/s15327906mbr4003_5] [Medline: 26794689]
22. Kim S, Baek TH, Kim YK, Yoo K. Factors affecting stickiness and word of mouth in mobile applications. JRIM 2016 Aug 8;10(3):177-192. [doi: 10.1108/JRIM-06-2015-0046]
23. Diana F, Bahry S, Masrom M, Masrek MN. Website credibility and user engagement: a theoretical integration. Presented at: 2016 4th International Conference on User Science and Engineering (i-USEr); Aug 23-25, 2016; Melaka, Malaysia p. 216-221. [doi: 10.1109/IUSER.2016.7857963]

Abbreviations

BRFSS: Behavioral Risk Factor Surveillance System
NHANES: National Health and Nutrition Examination Survey

Edited by A Mavragani; submitted 20.06.24; peer-reviewed by Anonymous; revised version received 19.02.25; accepted 19.02.25; published 14.05.25.

Please cite as:
Leslie-Miller CJ, Goltz SR, Barrios PL, Cushing CC, Badshah T, Ungaro CT, Qu S, Berezhnaya Y, Brisbois TD
Evaluating the Acceptability and Utility of a Personalized Wellness App (Aspire2B) Using AI-Enabled Digital Biomarkers: Engagement Enhancement Pilot Study
JMIR Form Res 2025;9:e63471
URL: https://formative.jmir.org/2025/1/e63471
doi:10.2196/63471

© Calissa J Leslie-Miller, Shellen R Goltz, Pamela L Barrios, Christopher C Cushing, Teena Badshah, Corey T Ungaro, Shankang Qu, Yulia Berezhnaya, Tristin D Brisbois. Originally published in JMIR Formative Research (https://formative.jmir.org), 14.5.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Leveraging Technology to Engage Supplemental Nutrition Assistance Program Consumers With Children at Farmers Markets: Qualitative Community-Engaged Approach to App Development

Callie Ogland-Hand[1,2], BA; Jillian Schulte[1,3], MA, MPH; Owusua Yamoah[1,2], MA, PhD; Kathryn Poppe[1], MPH, RDN; Timothy H Ciesielski[1,2], ScD, MD, MPH; Regan Gee[1,3], MSc; Ana Claudia Zubieta[4], PhD; Darcy A Freedman[1,2], MPH, PhD

[1]Mary Ann Swetland Center for Environmental Health, School of Medicine, Case Western Reserve University, 6700 Euclid Ave, Cleveland, OH, United States

[2]Department of Population and Quantitative Health Sciences, School of Medicine, Case Western Reserve University, Cleveland, OH, United States

[3]Department of Anthropology, Case Western Reserve University, Cleveland, OH, United States

[4]College of Food, Agricultural, and Environmental Sciences, The Ohio State University, Columbus, OH, United States

Corresponding Author:
Darcy A Freedman, MPH, PhD
Mary Ann Swetland Center for Environmental Health, School of Medicine, Case Western Reserve University, 6700 Euclid Ave, Cleveland, OH, United States

Abstract

Background: Fruit and vegetable consumption is lower than national trends among people receiving Supplemental Nutrition Assistance Program (SNAP) benefits due to economic and physical access barriers. Monetary nutrition incentive programs at farmers markets aim to reduce these barriers to improve diet quality among SNAP consumers. We leveraged community-engaged methods to collaboratively design a mobile app to increase the use of both nutrition incentive programs and farmers markets among SNAP households with children. This population represents about 35% of all SNAP households providing the dual benefit of improving diet for both adults and children.

Objective: In this paper, we share the iterative, community-engaged development process used to design a technology intervention that encourages the integration of farmers markets into the food shopping routines of SNAP consumers with children.

Methods: Our qualitative community-engaged approach was informed by human-centered design, following the inspiration and ideation phases of this framework. In the “inspiration” phase, we worked with community nutrition experts to define both the goal of and target audience for the app (ie, SNAP households with children). In the subsequent “ideation” phase, we completed 3 stages of data collection. We developed 2 interface prototypes and received feedback from end users on design and usability preferences before selecting a baseline model. Additional feedback gathered from qualitative interviews with 20 SNAP consumers with children was incorporated into the app’s version 1 (V1) development. We then shared V1 with SNAP consumers, children, and farmers market managers to test the app’s functionality, design, and utility.

Results: In the “inspiration” phase, the community nutrition partners identified SNAP consumers with children younger than 18 years as the target population for the app. In the “ideation” phase, we successfully created V1 through 3 stages of a qualitative, community-engaged process. First, about 75% (n=3) of SNAP consumers and all farmers market managers selected a grocery shopping design option for the layout of the app. Second, we integrated features identified by SNAP consumers with children into the app design, such as market information (ie, location with GPS address links, hours, website), likely available market inventory, market events, and grocery shopping checklists. Finally, we obtained recommendations for future versions of the app, including real-time changes in market hours, additional notification options, and grocery list personalization during a demonstration of V1. Both SNAP consumers and farmers market managers expressed interest in the app’s launch and utility.

Conclusions: It is feasible for community nutrition researchers to successfully design a community-engaged mobile app with the assistance of software developers. The community-engaged approach was key to us integrating potential end users’ preferences in the design of V1. Future work will assess the app’s impact on low-income families’ use of local farmers markets and nutrition incentive programs, as well as fruit and vegetable consumption.

(JMIR Form Res 2025;9:e70104) doi:10.2196/70104

XSL•FO
RenderX

KEYWORDS

farmers market; fruit and vegetable consumption; nutrition incentive; SNAP; community-engaged research; mobile app; app development

Introduction

Background

Most Americans (≈90%) do not meet recommendations for daily fruit and vegetable intake and the trends are even worse for those with low income [1]. People receiving Supplemental Nutrition Assistance Program (SNAP) benefits consume fewer fruits and vegetables compared with people not receiving these benefits [2], and low-income food-insecure households have higher rates of chronic disease [3]. Accordingly, increasing fruit and vegetable consumption is a national priority to reduce chronic disease risk among Americans [1], and this is especially important for low-income households.

Children tend to eat insufficient fruits and vegetables [4-6]. Child fruit and vegetable consumption is impacted by food availability at home, neighborhood food environment, and parents' perceived produce availability [7]. Relatedly, increased income is directly associated with increased fruit and vegetable consumption, putting children from families with low income at particular risk of not meeting dietary recommendations [4]. Low-income parents with young children represent about 35% of all SNAP households [8]. Financial constraints, limited time, and child food preferences are particularly burdensome for this population [9-12].

Monetary nutrition incentive programs address cost barriers to purchasing fruits and vegetables by augmenting SNAP benefits with additional "incentives" to lower food costs [13-15]. These programs are often available at local farmers markets, and they have successfully improved access to fruits and vegetables, improved fruit and vegetable intake, and reduced food insecurity, particularly for low-income households [16-19]. Despite the benefits of nutrition incentives at farmers markets, lack of awareness about these opportunities limits their use. A recent study in Ohio found only 45% of SNAP consumers with access to nutrition incentive programming within 5 miles of their home were using the program; most of the nonusers (>80%) were unaware that the incentive program existed [20]. Yet, 77% of the nonusers said that they were likely to use the program in the next 6 months after learning about the program [20]. Other research has found lack of awareness is a modifiable barrier to the use of nutrition programs [21,22]. These findings highlight how a lack of awareness of nutrition incentive programs at farmers markets can impede use.

Design Framework

In line with calls to translate scientific evidence into interventions that increase fruit and vegetable access and consumption, particularly among low-income families [23], our team set out to design a mobile app for SNAP consumers with children. Mobile apps are a popular approach in translational science in behavioral health interventions [24,25], demonstrating some success in changing dietary patterns [26]. Effective dissemination, implementation, and translational science, however, hinges on community engagement, where the target population is engaged in all processes from problem identification to design and implementation of interventions [27]. Community participation in technology development often occurs through the use of "Human-Centered Design"—a collaborative design process that includes multiple stakeholders [28,29]. Human-centered design is a community-engaged framework for technology development [30], and its use can help create buy-in, trust, and use of tools among target populations [31]. Our community-engaged approach was inspired by the iterative process of human-centered design, and our work in this paper will focus on 2 process phases and 3 stages of data collection.

Goal of This Study

We sought to raise awareness of nutrition incentive programs among SNAP consumers with children by developing a mobile app. To do this we leveraged a qualitative community-engaged process. We chose a community-engaged approach because the technology that is desired and designed by end users should be more useful to them, and it can lead to rapid adoption [30,31]. Ultimately, this approach should identify mobile app features that increase awareness of farmers markets and nutrition incentive programs and provide value to customers and managers of farmers markets.

Methods

Overview

The development of the app consisted of three community-engaged phases informed by human-centered design as follows: (1) inspiration, (2) ideation, and (3) implementation [28], outlined in Figure 1. Applying community-engaged concepts to app development has the potential to improve health behaviors, health knowledge, and awareness and use of health benefit programs [32]. Consequently, community-based methods and human-centered design principles have been applied to the field of public health and health care technology tools and services, ranging from physical activity in organ transplant recipients [33], telehealth services [31,34], and the Special Supplemental Nutrition Program for Women, Infant, and Children (WIC) programs [35,36]. Because we wanted to create an app with the aforementioned benefits that would be widely useful, human-centered design was embraced as a guiding framework for this community-engaged work. The inspiration and ideation phases occurred between May 2021 and October 2022, and the implementation phase is still ongoing.

XSL•FO
RenderX

Figure 1. Our development process for the app was an iterative, qualitative, and community-engaged process with 3 phases inspired by human-centered design principles of inspiration, ideation, and implementation [30]. SNAP: Supplemental Nutrition Assistance Program.

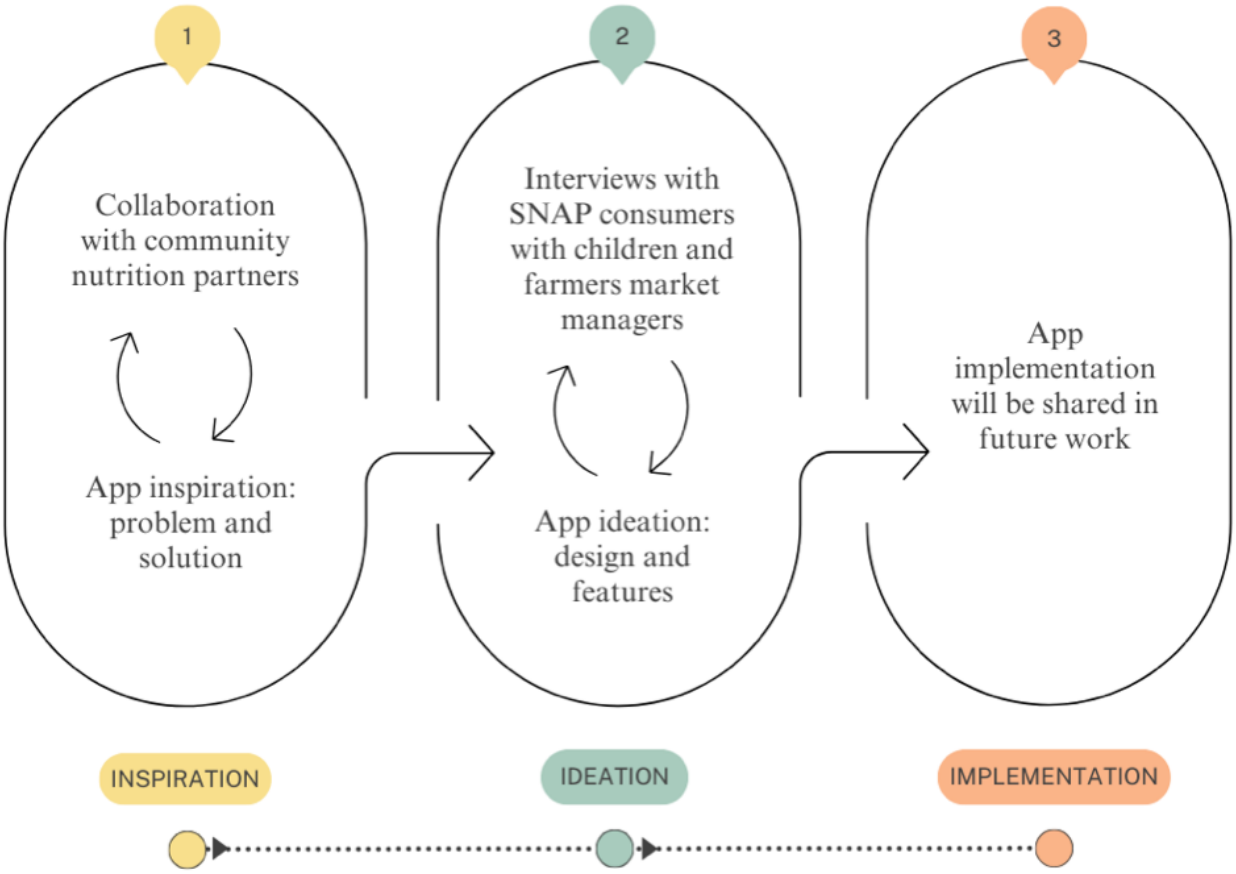

This study was designed and conducted by researchers at Case Western Reserve University in consultation with community nutrition partners at Ohio SNAP-Ed and Produce Perks Midwest, who have expertise in public health, federal food assistance programs (ie, SNAP, WIC), nutrition incentive programs, and farmers markets.

Phase 1: Inspiration

The "inspiration" phase included collaboration with community nutrition partners to identify the problem and brainstorm possible solutions.

Problem Identification

For over a decade, our research team has focused on healthy food access among low-income populations with a focus on farmers markets. Financial costs and physical access barriers limit the consumption of fruit and vegetables, especially among households with low income. While nutrition incentive programs aim to curb financial burdens and encourage farmers market use, awareness of local farmers markets and existing nutrition incentive programs is low among SNAP consumers [20,37-39]. The lack of repeat SNAP consumers at farmers markets is a national issue. Of the SNAP households shopping at a farmers market in 2021, roughly half made only one purchase throughout the year [40]. This underuse may be amplified by additional barriers to farmers market, including location, hours of operation, and transportation [41]. Face-to-face interventions aimed at improving engagement with farmers markets among low-income populations were not widely adopted because of the human capital investment required to support implementation [42], so technology may be an avenue to increase farmers market use.

Identifying Solutions

Building on this evidence, researchers from our team along with community nutrition partners tapped into lessons we have learned from our ongoing use of technology to improve community nutrition [22,43,44]. Our goal was to explore technology-based solutions to bridge the awareness gap among SNAP populations related to the availability of nutrition incentive programming.

Our previous work found that the majority of people hear about food procurement opportunities (like farmers markets) and nutrition incentive programs through word-of-mouth [20,45]. We hypothesized that technology could provide a unique opportunity to spur word-of-mouth dissemination [46,47]. Additionally, in support of a technology-based intervention, several studies have found significant improvement in diets and other health behaviors through app use [48-51]. Thus, the idea for a mobile app was identified as a solution to the underuse of farmers markets and nutrition incentive programs among low-income populations.

Prior Experience With Digital Solutions for Farmers Markets

Support for a digital technology intervention like the app was strengthened by our research team's previous successful

XSL•FO
RenderX

development and deployment of technology-based solutions to improve community nutrition, including the Produce Path Manager Portal. Produce Path Manager Portal is an app that allows farmers market managers to track transactions, nutrition incentive use, vendors, and general information about their farmers market [52]. In use at over 500 farmers markets across 7 states, Produce Path Manager Portal is designed for market managers and state-level organizations to track and report on nutrition incentive programming. Farmers market managers create individual accounts for their markets, but market information is not accessible publicly for market customers.

Proposed Solution: Consumer-Facing Portal

The creation of the new app enhances the market manager features (of Produce Path) as a manager-facing portal and merges those features with a customer-facing portal for market customers. Given their different perspectives and expertise, both managers and consumers were included in our design and development process. We contracted a software development team to help design the prototype for the app and build version 1 (V1). Through a bidding and request-for-proposal process, we invited 2 companies for interviews with our research team and community nutrition partners. We selected one company based on their described skills, previous app development, and ideas for the app.

Phase 2: Ideation

Overview

The "ideation" phase included development and feedback iterations (3 stages) to inform prototype design and learn about interests in using the app as well as features and design preferences. Qualitative feedback was collected over 16 months (May 2021-October 2022) with key end users: SNAP consumers with children and farmers market managers. Data came through feedback sessions and structured web-based interviews conducted via Zoom. Informed consent was obtained from participants involved in structured interviews. A summary of the data collection process and goals is outlined in Table 1.

Table . Our qualitative, community-engaged process of developing an app, informed by human-centered design: data collection, participants, and goals.

Design phase and goals		Data collection and participants
Phase 1: Inspiration		
	• 1. Problem identification • 2. Identifying solutions	• Consultation with community nutrition partners (n=7)
Phase 2: Ideation		
	• 1. Determining prototype design and app interest	Stage 1 Interviews • Feedback sessions with SNAP[a] consumers (n=4) • Feedback sessions with farmers market managers (n=3)
	• 2. Determining features and design preferences	Stage 2 Interviews • Formal interviews with SNAP consumers (n=20)
	• 3. Feedback on V1	Stage 3 Interviews • Feedback sessions with SNAP consumers (n=7) • Feedback sessions with farmers market managers (n=6)
Phase 3: Implementation		
	—[b]	—

[a]SNAP: Supplemental Nutrition Assistance Program.

[b]Not available (future work).

Stage 1 Interviews (November 2021 to February 2022): Determining Prototype Design and App Interest

Similar to approaches adapted by Kruzan et al [53] and Chandler et al [54] we worked with software developers to create 2 initial prototype interfaces. One was a shopping-centered design with a grocery list option and the second was a game-centered design with rewards and points for using the app. Both prototypes featured recipes with produce from farmers markets, a map of farmers markets, and a digital wallet for managing SNAP benefits, as illustrated in Figures 2 and 3. The shopping-centered design included market inventory and the ability to add produce to a grocery list. The gamification design distributed points to users for using the app and its features, and accumulated points could be traded in for in-app rewards, such as ribbons and badges.

XSL•FO
RenderX

Figure 2. The grocery shopping prototype design, one of 2 app design options designed by our software developers and presented to Supplemental Nutrition Assistance Program (SNAP) consumers for feedback.

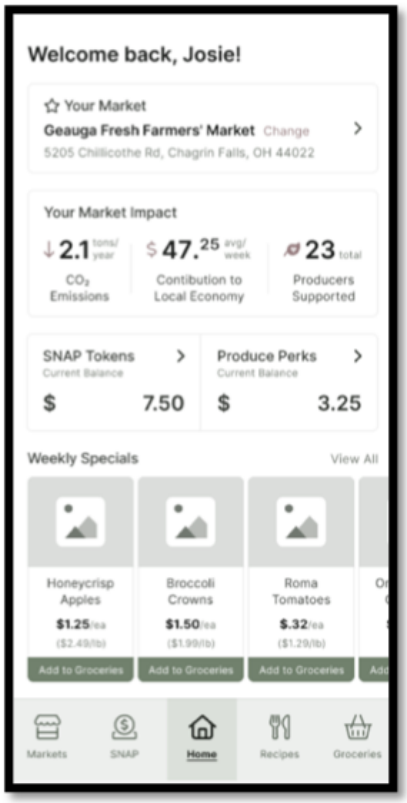

Figure 3. The gamification model prototype design, one of 2 app design options designed by our software developers and presented to Supplemental Nutrition Assistance Program (SNAP) consumers for feedback.

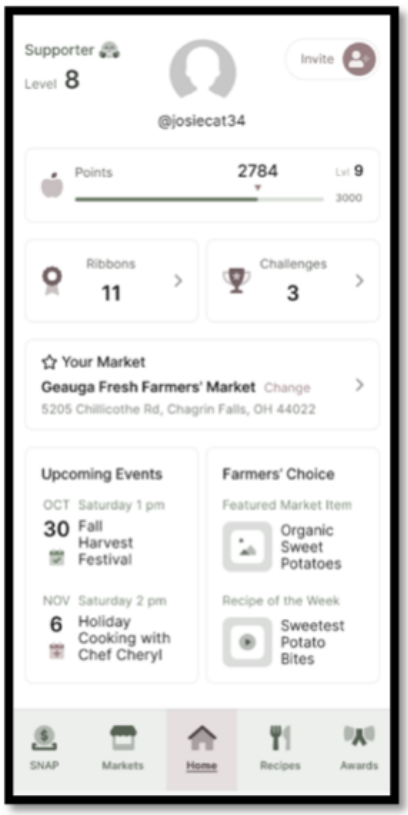

We held web-based feedback sessions via Zoom with SNAP consumers with children who had previous farmers market experience (n=4) and farmers market managers (n=3) to prioritize the prototype designs and understand the desired features of the front-facing app. These participants were purposively recruited by word-of-mouth by our community

XSL•FO
RenderX

nutrition partners. During the individual interviews, all participants were presented with simple, low-fidelity versions of the 2 app models and shown the content of the app and how they would function. Participants were asked about their overall impressions of both prototypes (ie, "Were there specific features of this model that you liked? Why?") and what features would motivate them to keep using the app. Participants were also asked if they thought either of the prototypes could potentially improve their farmers market experiences. Interviews were conducted by researchers COH and KP, and they lasted between 30 and 60 minutes and were not recorded or transcribed.

Stage 2 Interviews (July to September 2022): Determining Features and Design Preferences

Following the prototype selection in stage 1, we conducted structured qualitative interviews with 20 SNAP consumers with children younger than 18 years to understand their healthy food shopping routines and features of the app that might align with those routines. These interviews were conducted digitally via Zoom by researchers COH, JS, OY, RG, and KP, and they lasted between 45 and 60 minutes. We collected information about their likes and dislikes of mobile app designs in general, as well as apps that they specifically used in their cooking and shopping routines. These participants were recruited from prior research studies and across the state of Ohio through phone calls and flyers at farmers markets, grocery stores, and program offices (ie, WIC, Head Start). Eligible participants had used SNAP in the last 12 months, had at least 1 child younger than 18 years at home half of the time, did at least half of the household grocery shopping, and had experience using a smartphone app.

The interview guide and our engagement with the potential end users were rooted in the diffusion of innovation (DOI) theory. DOI is one of the first social science theories to understand behavior change [55]. This theory considers both diffusion, meaning the passive spread of an idea or product over time, and innovation, which includes novel ideas, products, interventions, etc [56]. Recently, DOI has been applied to health-related technology and mobile apps to understand their adoption and scalability [57-61]. The interview questions were shaped by 3 DOI attributes [56]: (1) relative advantage (ie, "Can you tell me about any apps or websites that you use for preparing and planning for shopping and cooking for your family?"), (2) compatibility (ie, "How would you rate this statement on a scale of 1, meaning not a good fit, and 10, a perfect fit: Using an app that tells me what foods are available at the farmers market every week could help me add more fresh and healthy foods to my shopping list?"), and (3) complexity (ie, What do you dislike about the apps that you use, or have used in the past?"). Attending to these constructs of innovations can aid in adoption and diffusion [55,62].

Following each interview, the interviewer and notetaker met to debrief and recorded key features of interest and any feedback relevant to the app development on a debrief form. Preliminary rapid analysis of debrief forms and discussions among the researchers who were involved in the interviews revealed features and design ideas mentioned by participants throughout the interviews. Informed by participants, research team members, community partners, and software developers, some suggested features were incorporated into the development of V1 in real time based on feasibility, available resources, and time. All interviews were recorded and transcribed, and formal qualitative analysis will be presented in future work.

Stage 3 Interviews (October to November 2022): Feedback on Version 1

Finally, individual feedback sessions were held with SNAP consumers (n=7) and farmers market managers (n=6) regarding V1 of the app. All SNAP consumers in these feedback sessions had previously participated in the stage 2 interviews. Two farmers market managers had previously participated in stage 1 interviews, and the remaining were 4 were purposively recruited by word-of-mouth by our community nutrition partners. SNAP users were selected based on the relevance of their responses in the stage 2 interviews related to app development and app use experience. During the debrief sessions after each interview in stage 2, the interviewer and note taker rated the relevance of participants' responses to the research goals on a scale of 0 (low) to 10 (high). Participants who scored 8 or higher were invited to participate in stage 3 feedback sessions. In these feedback sessions, V1 was shared to test the functionality, design, and utility of the app.

During these individual interviews, researchers demonstrated the app; SNAP consumers were shown the consumer-facing portal and market managers were shown the manager-facing portal to market managers. After the demonstration, participants were given time to practice using the app on their own (ie, SNAP consumers added certain items to their grocery list and market managers created a new market event). All participants were asked about their likes and dislikes (ie, "How well will the app support farmers market use and experience for you and other families with children who shop with SNAP?"), pains and gains (ie, "What feature(s) was confusing and why? What will help make it less confusing?"), and ideas for additional features (ie, "From your perspective, what kind of information should be included in a marketing video about farmers markets and shopping with SNAP at farmers markets?"). SNAP consumers were asked about the functionality of the shopping list and market selection, whereas market managers were asked about the functionality of updating market inventory, events, and vendors. The goal of these pilot interviews was to inform future app updates.

Phase 3: Implementation

The final stage of this work, as it is guided by human-centered design concepts, consists of iterations of measuring and learning. The implementation research testing the app will be presented in future work, along with pilot results.

Ethical Considerations

This study was approved by the Case Western Reserve University Institutional Review Board (STUDY20220359). All participants consented verbally to the study, and no specific quotations or identifiable details have been shared in this paper. SNAP consumers in the stage 2 (formal semistructured qualitative) interviews were compensated with a US$ 30 e-gift card to a superstore for their time.

XSL•FO
RenderX

Results

Phase 1: Inspiration (Problem Identification)

To improve awareness of and access to farmers markets among SNAP consumers, we needed to answer the following questions: (1) how is information about farmers markets and nutrition incentives shared with SNAP consumers? (2) What information do SNAP consumers need and want? Collaboration with our community partners led us to further focus this research question on a specific population—low-income parents—who face additional challenges and barriers to healthy food shopping and engagement with local farmers markets [9-12].

Phase 2: Ideation

Stage 1 Interviews: Determining Prototype Design and App Interest

Feedback sessions in stage 1 determined the prototype design. Core features of both prototypes included: market home pages with the address, hours, and website in addition to inventory, vendors, and events tabs; a map of nearby markets with the ability to "star" a home market; a recipe bank that could be filtered by ingredient; and, a web-based SNAP wallet that tracked both EBT and nutrition incentive use. The grocery shopping model uniquely included a customizable grocery list and the recipe ingredients could be directly added to the grocery list. On the other hand, the gamification version included an awards page to encourage app use, including challenges to visit multiple farmers markets and make recipes with leftovers, and in-app engagement to favorite recipes and review markets.

About 75% (n=3) of the SNAP consumers interviewed were highly in favor of a grocery model design compared with the gamification model (Figure 2 can be used for comparison). Their choice was largely informed by the utility of a grocery list, and not feeling motivated by in-app rewards or points in the game version. One participant for example shared that earning badges in the gamification version would not motivate them to use the app since they could not use those rewards in "real life." Another shared that they felt farmers market users would like the grocery shopping model since it provides exactly what they need. Thus, we moved forward with the grocery model for the app development. Participants expressed general interest in the app, stating that app-available information could help them eat healthier, and they emphasized the potential for the app to save them time and money while grocery shopping.

All farmers market managers were similarly interested in the app and were shown the customer-facing portal. They particularly liked the feature that listed the available produce at their markets, along with the option to indicate if a product was limited in quantity at their market. They felt the app would help them engage with market customers, and that it would help customers prepare for visits to the market.

There was mixed feedback on the web-based SNAP wallet feature and the feasibility of linking EBT cards through the app: some SNAP consumers expressed distrust in digital finances, and some market managers were concerned about market vendors not having the technology to accept digital payments. Thus, we decided to not pursue this feature in V1 development. Additionally, participants expressed a desire for a clean, non-cluttered design. Participants were distracted by having too much information on the home screen.

Stage 2 Interviews: Determining Features and Design Preferences

Formal qualitative interviews with SNAP consumers in stage 2 provided feedback on the desired features and design of the app, as well as overall utility. Participants were mostly Black (n=10, 50%), women (n=19, 95%) with children between 0 and 5 years, and lived in urban counties (n=14, 70%) in Ohio (Table 2 contains demographic characteristics of participants). About 80% (n=16) of the participants used mobile apps daily in their routine. Many suggested features and functions from stage 2 interviews were implemented in real-time development based on participant consensus and feasibility from our software design partners, as illustrated in Figures 4-6. Cost and time prevented us from incorporating all suggested features, such as features to manage nutrition incentive balances.

XSL•FO
RenderX

Table . Characteristics of stage 2 interview participants (Supplemental Nutrition Assistance Program [SNAP] consumers, n=20).

	Participants (n=20)
Gender, n (%)	
Female	19 (95)
Male	1 (5)
Age (years), mean (range)	37.8 (24-54)
Race and ethnicity, n (%)	
Black	10 (50)
White	6 (30)
Latinx	2 (10)
Unidentified	2 (20)
Parent or caretaker of children aged 0 - 5 years, n (%)	14 (70)
Number of children aged 0 - 5 years in these households, mean (range)	1.8 (1-4)
Parent or caretaker of children aged 6 - 11 years, n (%)	13 (65)
Number of children aged 6 - 11 years in these households, mean (range)	1.5 (1-4)
Parent or caretaker of children aged 12 - 17 years, n (%)	7 (35)
Number of children aged 12 - 17 years in these households, mean (range)	1.3 (1-3)
County of residence[a] n (%)	
Urban counties	14 (70)
Rural counties	6 (30)
Employment status, n (%)	
Employed full-time	4 (20)
Employed part-time	6 (30)
Currently out of work	3 (15)
Homemaker	6 (30)
Unable to work	1 (5)
2021 total household income (US $), n (%)	
Less than $25,000	13 (65)
Between $25,000 and $34,999	3 (15)
Between $35,000 and $49,999	3 (15)
Between $50,000 and $74,999	1 (5)
In a typical month, how involved are you in food shopping for your household?, n (%)	
I do more than half of the shopping	1 (5)
I do all the shopping	19 (95)
Have you ever used Produce Perks[b]?, n (%)	
Yes	9 (45)
No	10 (50)
I don't know	1 (5)
In a typical month, how often do you use an app on your smartphone?, n (%)	
Never	1 (5)
Weekly	3 (15)
Daily	16 (80)

XSL•FO
RenderX

[a]Urban and rural county designations are based on the Ohio Department of Health classification. Rural and partially rural counties are classified as rural in this analysis.

[b]Produce Perks is the SNAP-based fruit and vegetable incentive program in Ohio.

Figure 4. Example features that were incorporated from human-centered design phase 2, stage 2 interviews with Supplemental Nutrition Assistance Program (SNAP) consumers: a map of markets based on location or inputted address.

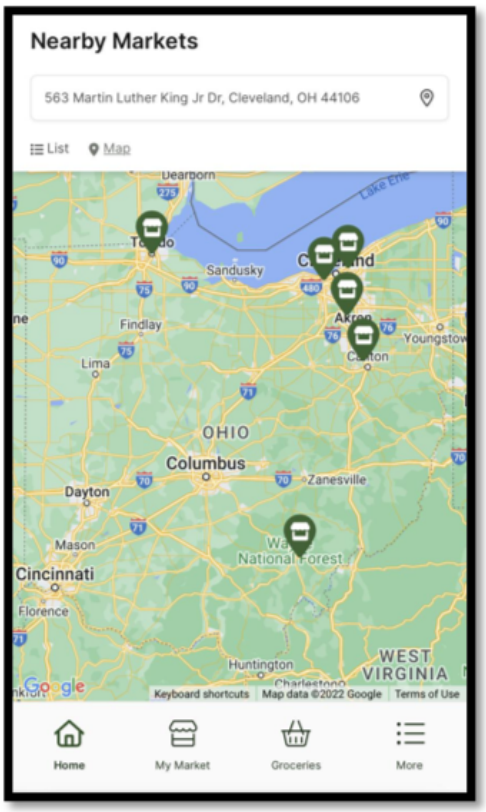

XSL•FO
RenderX

Figure 5. Example features that were incorporated from human-centered design phase 2, stage 2 interviews with Supplemental Nutrition Assistance Program (SNAP) consumers: hyperlinked market address that links to a navigation app.

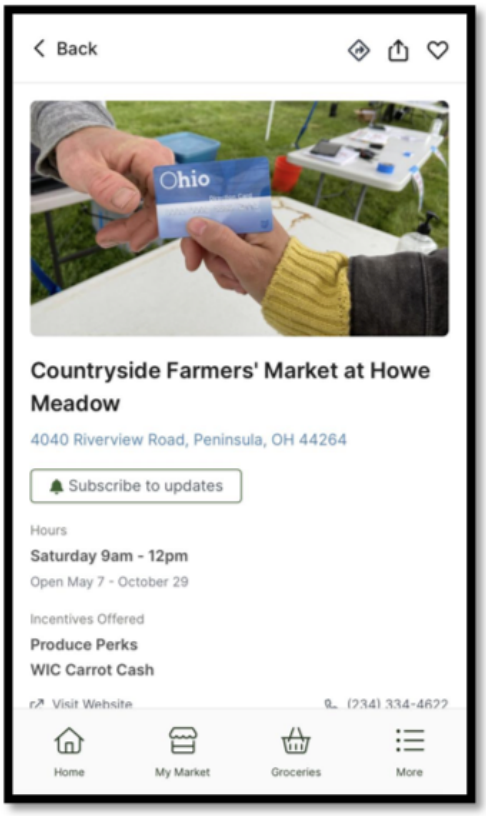

Figure 6. Example features that were incorporated from human-centered design phase 2, stage 2 interviews with Supplemental Nutrition Assistance Program (SNAP) consumers: likely available inventory at a specific market with the option to add to groceries by clicking on the basket icon.

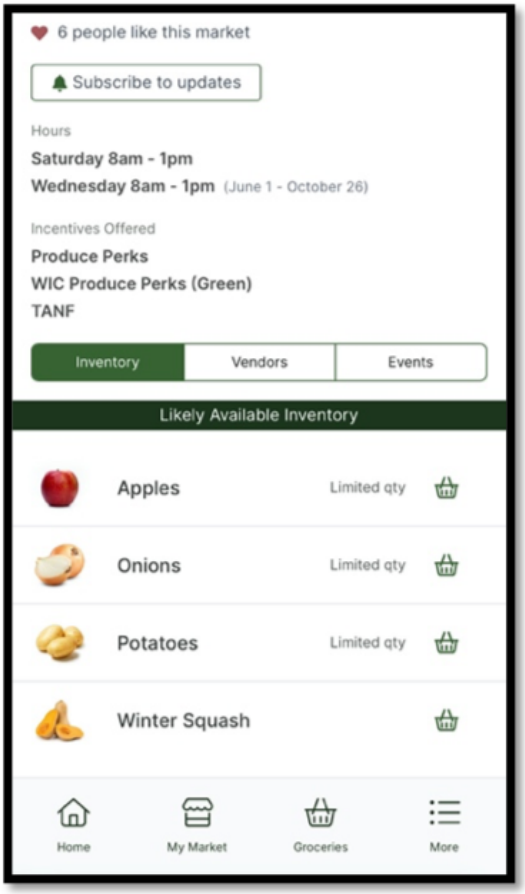

XSL•FO
RenderX

Consumer suggestions included in V1:

- List of likely available inventory.
- List of farmers market vendors.
- Highlight markets that accept incentive programs.
- Map of nearby farmers markets based on user location.
- Hyperlink market address to external GPS app.
- Share farmers market events with the ability to add them to the external calendar app.
- Include recipes with produce from the market.
- Grocery list function with the ability to cross items off.
- Instructions on how to use nutrition incentives.
- Up-to-date content.

Stage 3 Interviews: Feedback on Version 1

After incorporating the previously mentioned suggestions from stage 2 interviews, we finalized the app V1 development and informally shared it with 7 SNAP consumers from the stage 2 interviews and 6 farmers market managers representing markets in rural, urban, and diverse racial populations. The SNAP consumers were all women, 43% (n=3) identified as Black and another 43% (n=3) identified as White, about 14% (n=1) were Hispanic, and 57% (n=4) were from urban counties in Ohio. SNAP consumers thought the app was clear and straightforward. They provided additional features and design suggestions to inform the app version 2 and future app updates, including:

- Add links to vendors' websites or social media handles.
- List nonproduce inventory at the market, such as baked goods, craft items, and dairy products; then, add filter options by item category.
- Add custom items to the grocery list, in addition to the inventory items at the market, and the ability to undo crossing an item off the list.
- Rethink the home screen to include users' preferred farmers market.
- App branding incorporated throughout the app.

Farmers market managers found the manager-facing portal of the app V1 simple, user-friendly, easy to navigate, and visually appealing. The manager-facing portal was created from Produce Path Manager Portal, which they were familiar with, and offered enhanced features. Market managers appreciated the messaging feature and discussed that it could replace some market newsletters and offer more streamlined communication between markets and customers. Farmers market managers expressed the following suggested updates:

- Increase customization features to add more information about vendors, inventory, and general market details.
- Add analytics on customers who use the app, such as what items customers are adding to their grocery list to track product demand.
- Ability to add real-time open and closed dates and times for their market.

Discussion

Principal Findings

By using a qualitative community-engaged approach to app development, SNAP consumers and farmers market managers were involved throughout the design phases of the app to raise awareness about nutrition incentive programs and farmers markets in general. We found SNAP consumers and farmers market managers were interested in the grocery shopping model compared with the gamified version of the app. Through the interactive feedback and design process, participants suggested key features for the app including listing likely available inventory including nonproduce items, mapping nearby markets based on location, providing information about nutrition incentive programs, and connecting market addresses and events to outside apps (maps and calendars).

Centering partnerships between researchers, community members, stakeholders, and community partners are essential to creating equitable research [63]. The collaborative nature of community-engaged research allows community members and the target end users to make key design decisions and ensure that the developed technology meets their needs and desires. In this way, community engagement is a recommended method to work toward equity [29,64]. It is important to note that translational research requires the diverse partnerships that community-engagement and human-centered design promote. Translation, or the process of the widespread use of evidence-based programs, practices, and policies, has been a long-time priority of the Centers for Disease and Control to help prevent chronic disease [65]. Successful translation of products, such as a mobile app, necessitates input from diverse participants involved in product development [65]. Our research is strengthened by the collaborations we forged with OSU SNAP-Ed, Produce Perks Midwest, farmers markets throughout Ohio, SNAP consumers, and our contracted software developers.

Researchers aiming to develop mobile apps through similar processes should be aware that the collaborative and iterative nature of community-engaged methods is time- and resource-intensive. Data collection to inform app design and development occurred over the course of 16 months. For the app to remain useful and functioning well, consistent maintenance and a long-term contract with our app developers are essential. Indeed, work to improve and update the app is ongoing to date, and there are future plans for planning and piloting studies to test efficacy, as well as implementation research to assess app adoption.

This research aligns with growing interest in community-based approaches and human-centered design within public health [66], and the prior research on app development specifically for WIC participants [35,36,67-69]. This research illuminated that WIC participants prefer an app that allows them to locate nearby WIC offices, scan WIC foods at stores, provide appointment reminders, generate shopping lists, share recipe galleries, and can be navigated with minimal difficulty [35,67,68].

There is growing research on the need for more "human centric mobile apps" for federal assistant programs, such as SNAP [70,71]. The Ohio Department of Job and Family Services launched the mobile app ConnectEBT, but this primarily functions as a digital wallet and does not include other features relevant to food shopping for EBT users [72]. This is a significant gap in research, as one study found that WIC participants using a WIC-related app had higher rates of full

XSL•FO
RenderX

redemption, thus maximizing their food benefits [73]. Together, these prior studies demonstrate the potential for technology to encourage healthy eating behaviors among low-income populations. These prior findings motivate our approach and support the use of end user input in mobile app development.

Limitations

One primary limitation of our approach is the number of SNAP consumers (n=24 unique participants) and market managers (n=7 unique participants) that were involved, and this might impact the accuracy of our development. It is also possible that these individuals do not represent consumers and managers outside of Ohio, and this will affect our precision. Thus, we do not claim that this app is optimally tuned for all SNAP consumers or farmers market managers, but these issues can be addressed in future work. Some regional or demographic tailoring may be needed if the app is applied in other settings, as well as language translation if someone has limited English proficiency. Another major limitation of all technology research, including our developed app, is that it requires a level of digital literacy, which may render it inaccessible for some, given the national digital divide [74]. The SNAP consumers and farmers market managers involved in our research had varying levels of comfort using technology and familiarity with apps in general. A new focus on these "digital determinants of health" is increasingly important as health apps become more widely available [75]. Additionally, our research team did not have software design expertise, so the physical development of the app was contracted outside of our institution. This process requires financial and time investments, and the nature of app development requires long-term maintenance. Finally, this app was designed to improve awareness of farmers markets and nutrition incentive programs to help low-income families integrate more fruits and vegetables into their food shopping, but the app features cannot alleviate all access barriers, such as transportation to/from the markets and hours of operation.

Conclusions

Community-engaged methods, informed by human-centered design principles, are both a viable and uniquely beneficial approach to apply to health intervention app development. Both SNAP consumers and farmers market managers were excited by the prospect of the app and expressed desire for its launch and utility in its functions, highlighting the importance of engaging potential end users in technology development and centering their needs. Our goal is that the continued involvement of SNAP consumers and farmers market managers in the development process will promote buy-in and encourage the iterative development of an equitable, accessible app. Pilot studies of the app V1 and development of version 2 will be presented in future work.

Acknowledgments

Thank you to the participants who contributed to this research, our community nutrition partners, and collaborators at Produce Perks for guiding this work, especially Pat Bebo and Ana Bird. Thank you also to our software development partners TaaZaa and SkyCatchFire who designed and built the app. This project was supported by a grant from the US Department of Agriculture (USDA; grant G-2223-17-0850) and the Clinical and Translational Science Collaborative of Northern Ohio which is funded by the National Center for Advancing Translational Sciences (NCATS) of the National Institutes of Health (NIH; grant UM1TR004528). The content is solely the responsibility of the authors and does not necessarily represent the official views of the USDA or NIH.

Data Availability

Data sharing is not applicable to this article as no data sets were generated or analyzed during this study.

Authors' Contributions

COH and JS were involved in the methodology, investigation, formal analysis, and writing of the original draft, with COH also handling visualization. OY contributed to the methodology, investigation, and manuscript review and editing. KP contributed to the methodology, investigation, resources, supervision, and review and editing. THC contributed to the conceptualization, methodology, investigation, data curation, and manuscript review and editing. RG handled the methodology, investigation, formal analysis, and manuscript review and editing. ACZ contributed to the methodology and review and editing. DAF was responsible for the conceptualization, methodology, resources, project administration, and funding acquisition. All authors have read and approved the final manuscript for publication.

Conflicts of Interest

DAF, KP, and COH invented the intellectual property being evaluated for commercialization. Case Western Reserve University is the owner of the invention being commercialized with outside entities, and as such, could have financial benefit, such as royalties or other income. All other authors have no conflicts of interest to report.

References

1. Lee SH, Moore LV, Park S, Harris DM, Blanck HM. Adults meeting fruit and vegetable intake recommendations - United States, 2019. MMWR Morb Mortal Wkly Rep 2022 Jan 7;71(1):1-9. [doi: 10.15585/mmwr.mm7101a1] [Medline: 34990439]

XSL•FO
RenderX

2. Wolfson JA, Bleich SN. Fruit and vegetable consumption and food values: national patterns in the United States by supplemental nutrition assistance program eligibility and cooking frequency. Prev Med 2015 Jul;76:1-7. [doi: 10.1016/j.ypmed.2015.03.019]
3. Gregory CA, Reed-Jones M. Chronic disease risk increased with U.S. household food insecurity. USDA Economic Research Service. 2024. URL: http://www.ers.usda.gov/data-products/chart-gallery/gallery/chart-detail/?chartId=108913 [accessed 2024-10-18]
4. Hamner HC, Dooyema CA, Blanck HM, et al. Fruit, vegetable, and sugar-sweetened beverage intake among young children, by state - United States, 2021. MMWR Morb Mortal Wkly Rep 2023 Feb 17;72(7):165-170. [doi: 10.15585/mmwr.mm7207a1] [Medline: 36795611]
5. Wambogo E, Ansai N, Ahluwalia N, Ogden C. Fruit and vegetable consumption among children and adolescents in the United States, 2015–2018. : Centers for Disease Control and Prevention, National Center for Health Statistics; 2020 URL: https://www.cdc.gov/nchs/products/databriefs/db391.htm [accessed 2024-09-04]
6. Grimm KA, Kim SA, Yaroch AL, Scanlon KS. Fruit and vegetable intake during infancy and early childhood. Pediatrics 2014 Sep;134(Suppl 1):S63-S69. [doi: 10.1542/peds.2014-0646K] [Medline: 25183758]
7. Moffat LF, Ritchie LD, Gosliner W, Plank KR, Au LE. Perceived produce availability and child fruit and vegetable intake: the healthy communities study. Nutrients 2021 Oct 20;13(11):3681. [doi: 10.3390/nu13113681] [Medline: 34835936]
8. Monkovic M. Characteristics of Supplemental Nutrition Assistance Program households: fiscal year 2022. : U.S. Department of Agriculture, Food and Nutrition Service, Office of Policy Support; 2024 URL: https://www.fns.usda.gov/research/snap/characteristics-fy22 [accessed 2024-09-04]
9. Daniel C. Is healthy eating too expensive?: How low-income parents evaluate the cost of food. Soc Sci Med 2020 Mar;248:112823. [doi: 10.1016/j.socscimed.2020.112823] [Medline: 32062569]
10. Andress L, Fitch C. Juggling the five dimensions of food access: perceptions of rural low income residents. Appetite 2016 Oct 1;105:151-155. [doi: 10.1016/j.appet.2016.05.013] [Medline: 27208595]
11. Ravikumar D, Spyreli E, Woodside J, McKinley M, Kelly C. Parental perceptions of the food environment and their influence on food decisions among low-income families: a rapid review of qualitative evidence. BMC Public Health 2022 Jan 5;22(1):9. [doi: 10.1186/s12889-021-12414-z] [Medline: 34983469]
12. Mushi-Brunt C, Haire-Joshu D, Elliott M. Food spending behaviors and perceptions are associated with fruit and vegetable intake among parents and their preadolescent children. J Nutr Educ Behav 2007;39(1):26-30. [doi: 10.1016/j.jneb.2006.06.004] [Medline: 17276324]
13. Supplemental Nutrition Assistance Program (SNAP). USDA Food and Nutrition Service. URL: https://www.fns.usda.gov/snap/supplemental-nutrition-assistance-program [accessed 2021-12-04]
14. Gus Schumacher Nutrition Incentive Program. USDA Food and Nutrition Service. 2023. URL: https://www.fns.usda.gov/snap/gusnip-grant-program [accessed 2024-03-22]
15. About incentive projects. Nutrition Incentive Hub. URL: https://www.nutritionincentivehub.org/about/about-incentive-projects/ [accessed 2024-03-22]
16. Jilcott Pitts SB, Wu Q, Demarest CL, et al. Farmers' market shopping and dietary behaviours among supplemental nutrition assistance program participants. Public Health Nutr 2015 Sep;18(13):2407-2414. [doi: 10.1017/S1368980015001111] [Medline: 25895894]
17. Pitts SBJ, Gustafson A, Wu Q, et al. Farmers' market use is associated with fruit and vegetable consumption in diverse southern rural communities. Nutr J 2014 Dec;13(1):1. [doi: 10.1186/1475-2891-13-1]
18. Hu X, Clarke LW, Zendehdel K. Farmers' market usage, fruit and vegetable consumption, meals at home and health–evidence from Washington, DC. Sustainability 2021;13(13):7437. [doi: 10.3390/su13137437]
19. Moran A, Thorndike A, Franckle R, et al. Financial incentives increase purchases of fruit and vegetables among lower-income households with children. Health Aff (Millwood) 2019 Sep 1;38(9):1557-1566. [doi: 10.1377/hlthaff.2018.05420]
20. Vargo L, Ciesielski TH, Embaye M, Bird A, Freedman DA. Understanding SNAP recipient characteristics to guide equitable expansion of nutrition incentive programs in diverse food retail settings. Int J Environ Res Public Health 2022 Apr 20;19(9):4977. [doi: 10.3390/ijerph19094977] [Medline: 35564371]
21. Cohen AJ, Oatmen KE, Heisler M, et al. Facilitators and barriers to supplemental nutrition assistance program incentive use: findings from a clinic intervention for low-income patients. Am J Prev Med 2019 Apr;56(4):571-579. [doi: 10.1016/j.amepre.2018.11.010] [Medline: 30799161]
22. Freedman DA, Lee E, Ohri-Vachaspati P, et al. Predictors of farmers' market shopping among people receiving supplemental nutrition assistance program benefits. Am J Community Psychol 2018 Jun;61(3-4):488-499. [doi: 10.1002/ajcp.12245] [Medline: 29659024]
23. McCartney PR. The translational science spectrum. J Obstet Gynecol Neonatal Nurs 2021 Nov;50(6):655-658. [doi: 10.1016/j.jogn.2021.09.007] [Medline: 34648752]
24. Austin CP. Opportunities and challenges in translational science. Clin Transl Sci 2021 Sep;14(5):1629-1647. [doi: 10.1111/cts.13055] [Medline: 33982407]

XSL•FO
RenderX

25. Mannino RG, Arconada Alvarez SJ, Greenleaf M, Parsell M, Mwalija C, Lam WA. Navigating the complexities of mobile medical app development from idea to launch, a guide for clinicians and biomedical researchers. BMC Med 2023 Mar 23;21(1):109. [doi: 10.1186/s12916-023-02833-7] [Medline: 36959646]
26. Payne HE, Lister C, West JH, Bernhardt JM. Behavioral functionality of mobile apps in health interventions: a systematic review of the literature. JMIR Mhealth Uhealth 2015 Feb 26;3(1):e20. [doi: 10.2196/mhealth.3335] [Medline: 25803705]
27. Meissner P, Cottler LB, Eder MM, Michener JL. Engagement science: the core of dissemination, implementation, and translational research science. J Clin Trans Sci 2020 Jun;4(3):216-218. [doi: 10.1017/cts.2020.8] [Medline: 32695491]
28. Ergonomics of human-system interaction — part 210: human-centred design for interactive systems. International Organization for Standardization (ISO). 2019. URL: https://www.iso.org/standard/77520.html [accessed 2023-10-10]
29. Holeman I, Kane D. Human-centered design for global health equity. Inf Technol Dev 2019 Sep 29;26(3):477-505. [doi: 10.1080/02681102.2019.1667289] [Medline: 32982007]
30. Johnson T, Das S, Tyler N. Design for health: human-centered design looks to the future. Glob Health Sci Pract 2021 Nov 29;9(Suppl 2):S190-S194. [doi: 10.9745/GHSP-D-21-00608] [Medline: 34845042]
31. Idris MY, Korin M, Araya F, et al. Including the public in public eHealth: the need for community participation in the development of state-sponsored COVID-19-related mobile apps. JMIR Mhealth Uhealth 2022 Mar 9;10(3):e30872. [doi: 10.2196/30872] [Medline: 35113793]
32. Zhang Q, Zhang J, Park K, Tang C. Association between usage of an app to redeem prescribed food benefits and redemption behaviors among the special supplemental nutrition program for women, infants, and children participants: cross-sectional study. JMIR Mhealth Uhealth 2020 Oct 14;8(10):e20720. [doi: 10.2196/20720] [Medline: 33052133]
33. Handler L, Jaloul P, Clancy J, et al. A qualitative study of the perspectives of healthcare professionals on features of digital health interventions to support physical activity in solid organ transplant recipients. Prog Transplant 2023 Mar;33(1):43-49. [doi: 10.1177/15269248221145039] [Medline: 36537126]
34. Sánchez Antelo V, Szwarc L, Le Pera A, et al. Ten steps to design a counseling app to reduce the psychosocial impact of human papillomavirus testing on the basis of a user-centered design approach in a low- and middle-income setting. JCO Glob Oncol 2022 Oct;8:e2200168. [doi: 10.1200/GO.22.00168] [Medline: 36252163]
35. Weber SJ, Dawson D, Greene H, Hull PC. Mobile phone apps for low-income participants in a public health nutrition program for women, infants, and children (WIC): review and analysis of features. JMIR Mhealth Uhealth 2018 Nov 19;6(11):e12261. [doi: 10.2196/12261] [Medline: 30455172]
36. Biediger-Friedman L, Crixell SH, Silva M, Markides BR, Smith KS. User-centered design of a Texas WIC app: a focus group investigation. Am J Health Behav 2016 Jul;40(4):461-471. [doi: 10.5993/AJHB.40.4.8] [Medline: 27338993]
37. Garner JA, Coombs C, Savoie-Roskos MR, Durward C, Seguin-Fowler RA. A qualitative evaluation of double up food bucks farmers' market incentive program access. J Nutr Educ Behav 2020 Jul;52(7):705-712. [doi: 10.1016/j.jneb.2019.11.010] [Medline: 31924558]
38. Gosliner W, Hewawitharana SC, Strochlic R, Felix C, Long C. The California nutrition incentive program: participants' perceptions and associations with produce purchases, consumption, and food security. Nutrients 2022 Jan;14(13):2699. [doi: 10.3390/nu14132699]
39. Freedman DA, Flocke S, Shon EJ, et al. Farmers' market use patterns among supplemental nutrition assistance program recipients with high access to farmers' markets. J Nutr Educ Behav 2017 May;49(5):397-404. [doi: 10.1016/j.jneb.2017.01.007] [Medline: 28274790]
40. 2021 year end summary: SNAP retailer data. : USDA Food and Nutrition Service; 2021 URL: https://www.fns.usda.gov/snap/retailer/data [accessed 2025-04-23]
41. Freedman DA, Vaudrin N, Schneider C, et al. Systematic review of factors influencing farmers' market use overall and among low-income populations. J Acad Nutr Diet 2016 Jul;116(7):1136-1155. [doi: 10.1016/j.jand.2016.02.010] [Medline: 27021526]
42. Freedman D, Lee E, Flocke S, et al. Evaluation of a peer-to-peer approach to improve the reach of farmers' markets among low-income populations. Health Behav Res 2019 Mar 4;2(1). [doi: 10.4148/2572-1836.1034]
43. Lee E, Dalton J, Ngendahimana D, et al. Consensus modeling to develop the farmers' market readiness assessment and decision instrument. Transl Behav Med 2017 Sep;7(3):506-516. [doi: 10.1007/s13142-017-0504-2] [Medline: 28730383]
44. Yamoah O, Schulte J, Osborn L, Ogland-Hand C, Zubieta AC, Freedman DA. Pantry clients and supplemental nutrition assistance program-education practitioners' perspectives on factors influencing healthy eating policy, system and environmental interventions in food pantries. J Nutr Sci 2023;12:e81. [doi: 10.1017/jns.2023.64] [Medline: 37528826]
45. Monteban MF, Bess KD, Walsh CC, et al. People and places shaping food procurement among recipients of supplemental nutrition assistance program (SNAP). Health Place 2018 Sep;53:155-163. [doi: 10.1016/j.healthplace.2018.08.003]
46. Kauff J, Sama-Miller E, Makowsky E. Promoting public benefits access through web-based tools and outreach: a national scan of efforts volume I: background, efforts in brief, and related initiatives. Office of the Assistant Secretary for Planning and Evaluation (ASPE). 2011. URL: http://aspe.hhs.gov/reports/promoting-public-benefits-access-through-web-based-tools-outreach-national-scan-efforts-volume-i-0 [accessed 2025-02-11]

XSL•FO
RenderX

47. Headrick G, Ruth A, White SA, et al. Integration and coordination across public benefit programs: insights from state and local government leaders in the United States. Prev Med Rep 2023 Feb;31:102077. [doi: 10.1016/j.pmedr.2022.102077] [Medline: 36483579]
48. Zhou M, Fukuoka Y, Mintz Y, et al. Evaluating machine learning-based automated personalized daily step goals delivered through a mobile phone app: randomized controlled trial. JMIR Mhealth Uhealth 2018 Jan 25;6(1):e28. [doi: 10.2196/mhealth.9117] [Medline: 29371177]
49. Walsh JC, Corbett T, Hogan M, Duggan J, McNamara A. An mHealth intervention using a smartphone app to increase walking behavior in young adults: a pilot study. JMIR Mhealth Uhealth 2016 Sep 22;4(3):e109. [doi: 10.2196/mhealth.5227] [Medline: 27658677]
50. Mummah S, Robinson TN, Mathur M, Farzinkhou S, Sutton S, Gardner CD. Effect of a mobile app intervention on vegetable consumption in overweight adults: a randomized controlled trial. Int J Behav Nutr Phys Act 2017 Sep 15;14(1):125. [doi: 10.1186/s12966-017-0563-2] [Medline: 28915825]
51. Clarke P, Evans SH, Neffa-Creech D. Mobile app increases vegetable-based preparations by low-income household cooks: a randomized controlled trial. Public Health Nutr 2019 Mar;22(4):714-725. [doi: 10.1017/S1368980018003117] [Medline: 30472970]
52. Freedman DA, Hunt AR, Merritt K, Shon EJ, Pike SN. Dissemination of technology to evaluate healthy food incentive programs. Am J Prev Med 2017 Mar;52(3 Suppl 3):S309-S314. [doi: 10.1016/j.amepre.2016.08.036] [Medline: 28215387]
53. Kruzan KP, Reddy M, Washburn JJ, Mohr DC. Developing a mobile app for young adults with nonsuicidal self-injury: a prototype feedback study. Int J Environ Res Public Health 2022 Dec 2;19(23):16163. [doi: 10.3390/ijerph192316163] [Medline: 36498234]
54. Chandler R, Hernandez N, Guillaume D, Grandoit S, Branch-Ellis D, Lightfoot M. A community-engaged approach to creating a mobile HIV prevention app for black women: focus group study to determine preferences via prototype demos. JMIR Mhealth Uhealth 2020 Jul 24;8(7):e18437. [doi: 10.2196/18437] [Medline: 32706723]
55. Haider M, Kreps GL. Forty years of diffusion of innovations: utility and value in public health. J Health Commun 2004;9 Suppl 1:3-11. [doi: 10.1080/10810730490271430] [Medline: 14960400]
56. Rogers EM. Diffusion of Innovations, 5th edition: Simon and Schuster; 2003.
57. Vamos CA, Griner SB, Kirchharr C, et al. The development of a theory-based eHealth app prototype to promote oral health during prenatal care visits. Transl Behav Med 2019 Nov 25;9(6):1100-1111. [doi: 10.1093/tbm/ibz047]
58. Coyte PC, Holmes D. Health care technology adoption and diffusion in a social context. Policy Polit Nurs Pract 2007 Feb;8(1):47-54. [doi: 10.1177/1527154407299646] [Medline: 17470771]
59. Hernandez JPT. Network diffusion and technology acceptance of a nurse chatbot for chronic disease self-management support : a theoretical perspective. J Med Invest 2019;66(1.2):24-30. [doi: 10.2152/jmi.66.24] [Medline: 31064947]
60. Coons JC, Patel R, Coley KC, Empey PE. Design and testing of Medivate, a mobile app to achieve medication list portability via fast healthcare interoperability resources. J Am Pharm Assoc (2003) 2019 Mar;59(2):S78-S85. [doi: 10.1016/j.japh.2019.01.001]
61. Nwosu AC, Mason S. Palliative medicine and smartphones: an opportunity for innovation? BMJ Support Palliat Care 2012 Mar;2(1):75-77. [doi: 10.1136/bmjspcare-2011-000151] [Medline: 24653504]
62. Iqbal M, Zahidie A. Diffusion of innovations: a guiding framework for public health. Scand J Public Health 2022 Jul;50(5):533-537. [doi: 10.1177/14034948211014104] [Medline: 34058897]
63. Viswanathan M, Ammerman A, Eng E, et al. Key question 1: definition of community-based participatory research. In: Community-Based Participatory Research: Assessing the Evidence: Agency for Healthcare Research and Quality (US); 2004. URL: https://www.ncbi.nlm.nih.gov/sites/books/NBK11852 [accessed 2025-04-23]
64. Marrone NL, Nieman CL, Coco L. Community-based participatory research and human-centered design principles to advance hearing health equity. Ear Hear 2022;43(Suppl 1):33S-44S. [doi: 10.1097/AUD.0000000000001183] [Medline: 35724253]
65. Wilson KM, Brady TJ, Lesesne C, on NWG. An organizing framework for translation in public health: the Knowledge to Action Framework. Prev Chronic Dis 2011 Mar;8(2):A46. [Medline: 21324260]
66. Chen E, Neta G, Roberts MC. Complementary approaches to problem solving in healthcare and public health: implementation science and human-centered design. Transl Behav Med 2021 May 25;11(5):1115-1121. [doi: 10.1093/tbm/ibaa079]
67. Friedman LB, Silva M, Smith K. A focus group study observing maternal intention to use a WIC education app. Am J Health Behav 2018 Nov 1;42(6):110-123. [doi: 10.5993/AJHB.42.6.11] [Medline: 30158006]
68. Weber SJ, Shearer E, Mulvaney SA, et al. Prioritization of features for mobile apps for families in a federal nutrition program for low-income women, infants, and children: user-centered design approach. JMIR Form Res 2021 Jul 30;5(7):e30450. [doi: 10.2196/30450] [Medline: 34328432]
69. Bensley RJ, Hovis A, Horton KD, et al. Accessibility and preferred use of online web applications among WIC participants with internet access. J Nutr Educ Behav 2014;46(3 Suppl):S87-S92. [doi: 10.1016/j.jneb.2014.02.007] [Medline: 24810002]
70. Jo J, Xie J, Carroll JM. Food with dignity: public values in the Supplemental Nutrition Assistance Program mobile applications. Presented at: Proceedings of the 2023 ACM Conference on Information Technology for Social Good; Sep 6-8, 2023; Lisbon, Portugal. [doi: 10.1145/3582515.3609541]

XSL•FO
RenderX

71. Patel S, Nair A, Rajpura NH, Jogadia DD. Food for thought: greenbasket's approach to helping SNAP families and local farmers towards self-reliance. Presented at: CHI '24: CHI Conference on Human Factors in Computing Systems; May 11-16, 2024; Honolulu, HI, United States. [doi: 10.1145/3613905.3647976]
72. ODJFS mobile app: ConnectEBT. Ohio Benefits Self Service Portal Core. 2023. URL: https://benefits.ohio.gov/home/news-and-events/all-news/odjfs-mobile-app [accessed 2024-06-05]
73. Zhang Q, Zhang J, Park K, Tang C. App usage associated with full redemption of WIC food benefits: a propensity score approach. J Nutr Educ Behav 2021 Sep;53(9):779-786. [doi: 10.1016/j.jneb.2021.03.002] [Medline: 34175218]
74. van Dijk J, Hacker K. The digital divide as a complex and dynamic phenomenon. Inf Soc 2003 Sep;19(4):315-326. [doi: 10.1080/01972240309487]
75. Philpot LM, Ramar P, Roellinger DL, Njeru JW, Ebbert JO. Individual-level digital determinants of health and technology acceptance of patient portals: cross-sectional assessment. JMIR Form Res 2024 Jun 10;8:e56493. [doi: 10.2196/56493] [Medline: 38695754]

Abbreviations

DOI: diffusion of innovation
SNAP: Supplemental Nutrition Assistance Program
V1: version 1
WIC: Women, Infants, and Children

Edited by A Mavragani; submitted 15.12.24; peer-reviewed by E Chen, R Strochlic; revised version received 20.03.25; accepted 21.03.25; published 16.05.25.

Please cite as:
Ogland-Hand C, Schulte J, Yamoah O, Poppe K, Ciesielski TH, Gee R, Zubieta AC, Freedman DA
Leveraging Technology to Engage Supplemental Nutrition Assistance Program Consumers With Children at Farmers Markets: Qualitative Community-Engaged Approach to App Development
JMIR Form Res 2025;9:e70104
URL: https://formative.jmir.org/2025/1/e70104
doi:10.2196/70104

© Callie Ogland-Hand, Jillian Schulte, Owusua Yamoah, Kathryn Poppe, Timothy H Ciesielski, Regan Gee, Ana Claudia Zubieta, Darcy A Freedman. Originally published in JMIR Formative Research (https://formative.jmir.org), 16.5.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Perception of Empathy in Mental Health Care Through Voice-Based Conversational Agent Prototypes: Experimental Study

Ruvini Sanjeewa[1], BSc; Ravi Iyer[1], PhD; Pragalathan Apputhurai[1], PhD; Nilmini Wickramasinghe[2], PhD; Denny Meyer[1], PhD

[1]Department of Health Science and Biostatistics, School of Health Sciences, Swinburne University of Technology, Hawthorn, Australia
[2]School of Computing, Engineering & Mathematical Sciences, La Trobe University, Melbourne, Australia

Corresponding Author:
Ruvini Sanjeewa, BSc
Department of Health Science and Biostatistics, School of Health Sciences, Swinburne University of Technology, Hawthorn, Australia

Abstract

Background: Empathy is a critical component of effective mental health care communication. Positive perceptions of empathy in conversational agents (CAs) operating in the health care domain are therefore needed to enhance the quality of care provided by these emerging technologies. However, research on how users perceive empathy in CAs is limited, particularly in voice-based prototypes.

Objective: The objective of this study is to identify to what extent perceptions of empathy in CA prototypes correspond with the engineered empathy levels for these voice-based prototypes. In addition, as a secondary aim, this study investigates how the demographic characteristics of participants affect their perception of empathy in a mental health helpline service context.

Methods: Swinburne University first-year psychology students (N=306) were presented with 9 CA prototypes engineered to portray low, medium, or high empathy levels, and their perceptions of empathy were collected via an electronic survey. Perceptions of empathy were rated using the Perceived Emotional Intelligence (PEI) Scale and the Raters' Scale (RS10).

Results: Most participants were female (233/306, 76%) with a mean age of 30 (SD 10.69) years, while a majority (194/306, 63%) were of Australian and New Zealand background. A strong positive correlation between the PEI and RS10 ratings was observed (r=0.829, P<.001). The empathy ratings across the 3 engineered empathy levels showed significant differences when using both PEI (χ^2_2=11.865, P=.003) and RS10 (χ^2_2=19.737, P<.001) measures. A linear mixed model for PEI showed significantly higher ratings for high rather than low engineered empathy levels (t_8=−2.34, P=.048). RS10 ratings were also significantly higher for high rather than low engineered empathy levels (t_8=−2.45, P=.04). However, no significant differences were detected between the CAs with engineered medium-level empathy and the CAs with low or high engineered empathy levels. The linear mixed model for PEI showed significantly higher ratings for participants of the Asian and Other ethnic categories compared to the Oceanic category (t_{285}=2.54, P=.01 and t_{286}=2.25, P=.03 respectively). The RS10 ratings were also significantly higher for the Other category rather than for the Oceanic category (t_{284}=2.24, P=.03). Women showed significantly higher RS10 ratings than men (t_{283}=1.94, P=.05).

Conclusions: Recognizing empathy levels in CA prototypes proved challenging, highlighting possible complexities involved with voice-based empathy detection. The perception of empathy may also be affected by different ethnic and gender-based factors. The study findings emphasize the importance of personalized communications by CAs, with expressions of empathy tailored to key demographic characteristics of users. Future studies in a similar context would benefit from the inclusion of participants who are end users of a mental health care service with more balanced gender and age distributions. Multimodal interactions could also be considered for CA prototype development.

(JMIR Form Res 2025;9:e69329) doi:10.2196/69329

KEYWORDS

perceived empathy; mental health care; helpline service; conversational agent prototypes; voice interactions; demographic information

XSL•FO
RenderX

Introduction

Empathy is a psychological construct that includes emotional as well as cognitive components. Empathy refers to the detection and perception of another person's emotional experience, that is, feeling that emotional experience as one's own. While empathy is critical for human interaction and the development of trust and safety [1], it is a complex and multidimensional construct [1]. Empathy requires an imagining of another's experience, coupled with an interpretation of what they may be feeling and the reasons underlying these feelings [2].

Several factors moderate how individuals perceive empathy [3]. Women appear to perceive empathy more adeptly than men [4-6], while sensitivity to the emotional experiences of other people declines with age [6]. Cultural and ethnic backgrounds also influence how people perceive empathy [7]. Thus, social norms strongly inform how empathy is perceived and portrayed [8].

Conversational agents (CAs) with empathic capabilities offer advantages for mental health care [9], particularly in addressing current shortfalls in the current workforce [10]. If CAs can be shown to engage with users empathically, this technology may be leveraged to facilitate certain health care interactions, including patient triage, particularly via telephone and at scale.

In this experimental study, we engineer several vocal CA prototypes, displaying varying levels of empathy, which might feature in a triage context in a mental health helpline service setting. This study attempts to identify how perceptions of empathy in CA communications differ between individuals listening to staged conversations between a CA and a suicide helpline caller.

Although many studies have examined the concept of empathy [11,12], few studies have focused on how perceptions of empathy vary between individuals. An investigation of the perception of empathy by consumers of a customer service chatbot used the 3-item SERVQUAL scale to capture perceived levels of empathy. Based on feedback from participants recruited through electronic platforms, it was evident that empathy in chatbots leads to customer satisfaction at a level comparable with that achieved by human call-responders [13]. A second study proposed a new framework for evaluating empathy in dialogic systems using the Empathy Scale for Human-Computer Communication (EHSCC) in an attempt to standardize the concept of empathic assessment [14]. This framework is intended for a broader audience, not limited to clinical settings, where empathy assessments are made using text-based interactions or using transcriptions of voice-based interactions. A third study developed a Perceived Empathy of Technology Scale (PETS) to measure the perceived empathy of users with technologies such as chatbots [15].

This study differs from the aforementioned studies by using voice-based CA prototypes with varying levels of empathy, developed to reflect a mental health helpline service context. Swinburne University first-year psychology students, reflective of younger members of the broader community, were asked to evaluate the level of empathy perceived within each prototype using a range of assessment measures. Our study, therefore, extends the focus of previous research to the mental health care domain.

The primary objective of this study is to evaluate the extent to which ratings of empathy correspond with the levels of empathy engineered for each of the 9 CAs. The secondary objective was to determine how participant characteristics, specifically gender identity (sex assigned at birth), age, ethnic background, Aboriginal and Torres Strait Islander status, Index for Relative Socio-Economic Advantage and Disadvantage (IRSAD), and home language, may moderate perceptions of empathy.

Therefore, by investigating how engineered empathy in CA voice prototypes is perceived and whether the demographics of the participants impact this decision, the study offers insights into designing more effective empathic CAs for mental health care settings.

Methods

Data Collection

In this study, we engineered standardized vocal responses by 9 CA prototypes to short statements made by an artificial real-world caller to an Australian mental health telephone counseling helpline service. Short statements were identified by the research team from a large collection of call recordings obtained as part of a previous study [16]. Data collection was conducted through an electronic questionnaire developed in Qualtrics (see Multimedia Appendix 1), inviting participants to rate the level of empathy (eg, low, medium, or high) portrayed by each of 3 randomly assigned CAs.

Participants

Participants were recruited from a Research Experience Program (REP) offered to first-year psychology students at Swinburne University of Technology, Australia. As a foundational unit of study provided at the undergraduate level within multiple disciplines (eg, science, humanities, and business), it is unlikely that the students would have any practical psychology experience, suggesting similar perceptions of empathy to the broader community. The advertisement for the study was posted on the internal REP platform housed on the university SONA system, which is accessible to this student group. Once interested students confirmed their decision to participate in the survey, they were provided with the survey link by the research administrator. Only students who were living in Australia and aged 18 years and older were eligible for participation.

Survey Procedure

The survey was conducted between February 27, 2024, and June 2, 2024. Consenting students were electronically directed to the Qualtrics survey. Demographic information was collected from each student, including gender identity, age, residential postcode, Aboriginal and Torres Strait Islander status, ethnic background, and home language. Then, a set of 3 audio recordings was assigned to each participant as shown in Figure 1. These recordings were created as prototype CAs to be used within a mental health helpline service context. Each participant rated the level of empathy of each CA using two scales: the

XSL•FO
RenderX

Perceived Emotional Intelligence (PEI) Scale [17] and the Raters' Scale (RS10) [18].

Figure 1. Outline of how the voice-based CA prototypes were assigned to participants in the study. CA: conversational agent; R: randomization. (reproduced with permission from Media Medic [https://www.mediamedic.studio] - All Rights Reserved.)

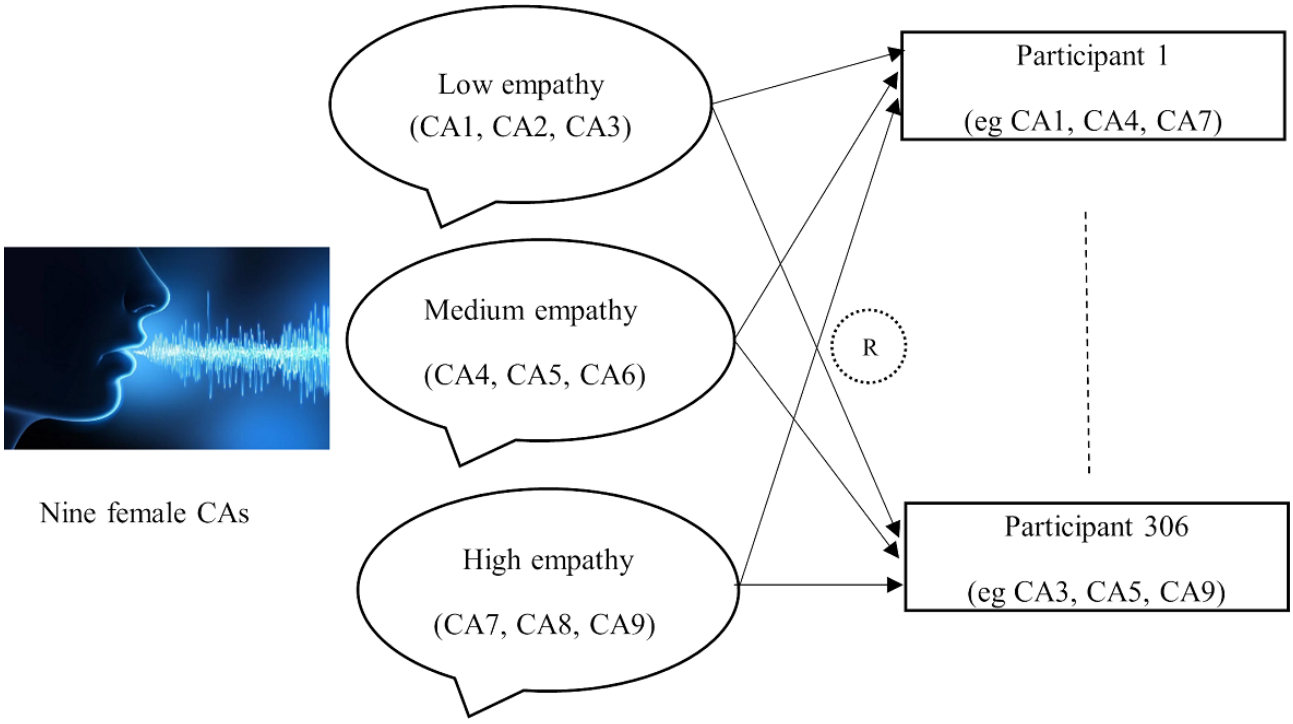

This study has been reported in accordance with the Checklist for Reporting Results of Internet E-Surveys (CHERRIES) [19] shown in Checklist 1.

CA Prototypes

The recordings for 9 prototype CAs were developed using the NaturalReader (NaturalSoft Ltd) [20], a speech synthesis software. Prototype responses were developed following an intensive analysis of counselor responses to callers of On The Line Australia (OTLA), a national telephone helpline counseling service. A range of only female voices was chosen, in part to simplify the experiment but also based upon focus group responses with OTLA staff, suggesting that female voices convey greater variation in empathy than male voices (blinded for review).

NaturalReader software allowed each voice to be modulated to accommodate either low, medium or high levels of conveyed empathy. The level of empathy chosen for each prototype was validated by research team members (blinded for review), drawing upon substantial experience working in mental health. The CA prototypes used in the study are mentioned under data availability.

Measures

The level of empathy was collected using two measures: the Perceived Emotional Intelligence Scale (PEI) and the Rater's Scale (RS10). The PEI is a 7-item scale with items rated on a 10-point sliding scale from 1=low empathy to 10=highest level of empathy. The RS10 is a single-item 10-point analog scale from 0=poor empathy to 10=highest level of empathy. This has proven reliable for researchers when minimizing the burden on participants to obtain responses within a short time.

A summary of the demographic features of participants and measures for perceived empathy used are attached in Multimedia Appendix 2.

Statistical Analysis

RStudio (Posit, PBC, version 2024.04.2, Build 764) was used for all analyses. Participant characteristics and empathy ratings were summarized with appropriate descriptive statistics. Spearman rank correlation was used to assess the relationship between PEI and RS10 empathy ratings. Kruskal-Wallis tests were performed to evaluate the differences between the engineered empathy levels in CA prototypes in terms of participant ratings using the PEI and RS10 measures [21].

Empathy ratings were regarded as nested within individual participants. A random intercept model was fitted with the PEI rating as the response variable and the RS10 rating as a fixed effect to more accurately model the relationship between the PEI and RS10 ratings. A linear mixed model was fitted separately for PEI and RS10 ratings, with the participant demographic information and engineered empathy levels as the fixed effects. Random intercepts for each participant and each of the 9 CA prototypes were added. These models were designed to show how participant demographic information and engineered empathy levels impacted the empathy ratings provided by participants.

Ethical Considerations

The study was approved by the Swinburne Human Research Ethics Committee (Ref: 20237339 - 16599) and conducted in accordance with all aspects of the World Medical Association Declaration of Helsinki. The online survey clearly stated information about the study, its objective, what participation

XSL•FO
RenderX

involved, and its benefits to participants, along with contact details of support services if a participant feels a low degree of discomfort due to the nature of the conversation included in the CA prototypes.

Consent was obtained as a part of the survey before proceeding to the questions. It was communicated to the interested students that participation was voluntary, allowing them to withdraw from the survey at any point. In addition, all responses were kept anonymous. The extracted survey results were stored in One Drive for Business with encryption accessible only to the research team. The students were not compensated for their participation. However, they received half a credit point for taking part in this 30-minute survey, which was a part of their course requirements. The credits were offered at the end of each week when the survey was closed for students who signed up that week to avoid any influence of the credit points on their engagement and responses. The study project was unrelated to first-year psychology course content, and survey completion was not a condition of receiving course credit; the REP program is a staple component of first-year psychology, in which most students choose to participate. Thus, we believe that any possible response bias was minimal and that the responses provided are not dissimilar to those that would be received from members of the general population within a similar age range. However, an attempt was made to validate the veracity of participant responses by checking the correlation between the PEI and RS10 ratings.

The details of the survey, including the consent form, are included in Multimedia Appendix 1.

Results

Participant Characteristics and Empathy Ratings

Data were collected from 306 Swinburne University of Technology students. Participant characteristics are provided in Table 1. Ages ranged between 18 and 61 (mean 29.65, SD 10.69) years. A mean IRSAD decile (7.12, SD 2.62) was derived using participant postcodes [22]. Figures 2 and 3 show the distributions for the RS10 and PEI empathy ratings according to engineered empathy levels.

Table . Summary of participant characteristics (N=306): demographic information.

Question	Participants, n (%)
Gender identity	
Woman	233 (76.1)
Man	58 (19)
Other	7 (2.29)
Missing	8 (2.61)
Do you identify as Aboriginal or Torres Strait Islander, or both?	
Yes	5 (1.63)
No	300 (98)
Missing	1 (0.33)
Is English your home language?	
Yes	258 (84.3)
No	47 (15.4)
Missing	1 (0.33)
Ethnic background	
Oceania (including Australia and New Zealand)	194 (63.4)
Asia	47 (15.4)
Europe	46 (15)
Other (America, Africa, and Middle East)	18 (5.88)
Missing	1 (0.33)

XSL•FO
RenderX

Figure 2. Distribution of RS10 empathy ratings by engineered empathy levels. RS10: Rater's Scale.

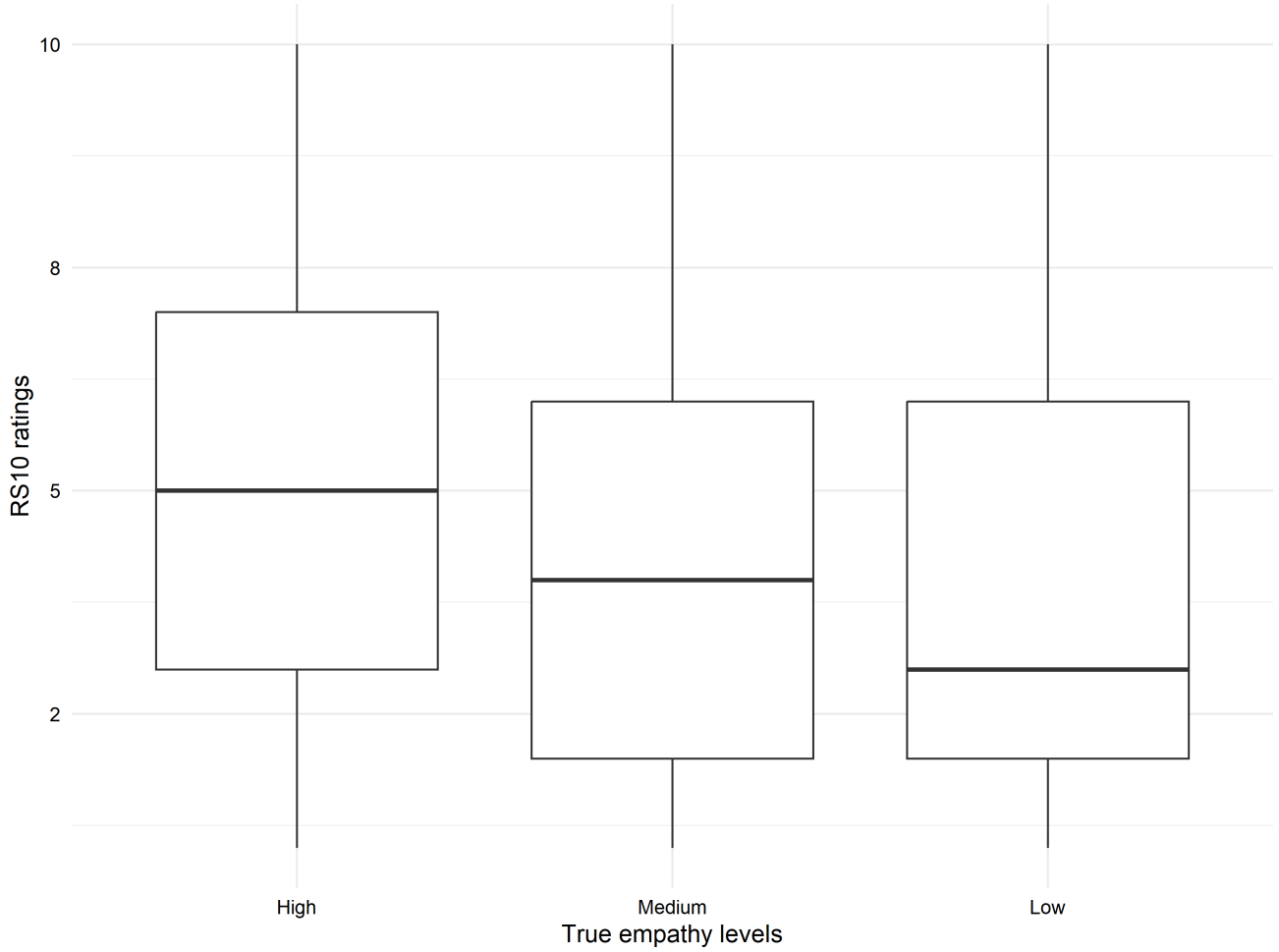

XSL•FO
RenderX

Figure 3. Distribution of Perceived Emotional Intelligence empathy ratings by engineered empathy levels. The two dots are outliers that are higher than the rest of the data points. PEI: Perceived Emotional Intelligence.

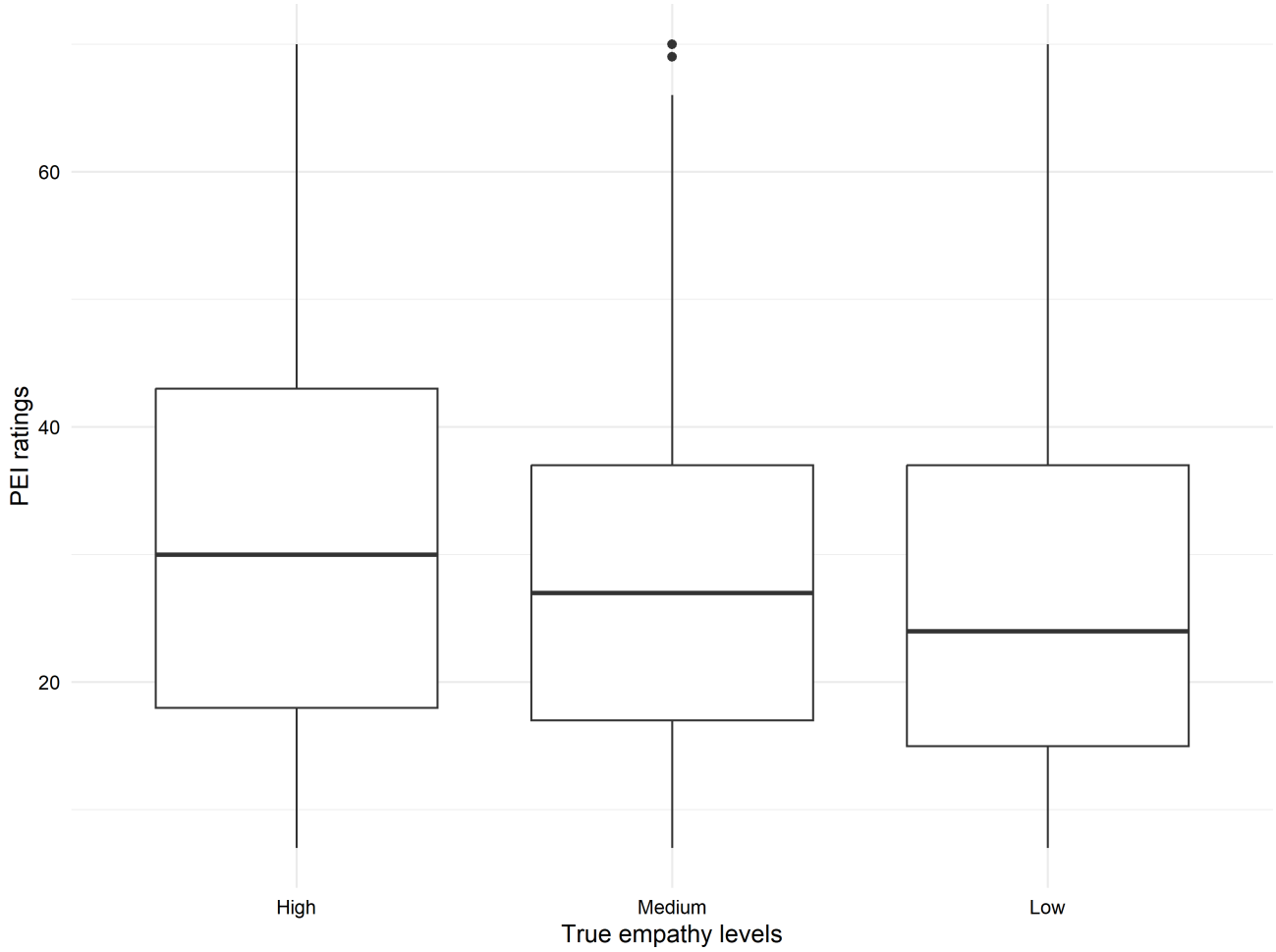

A strong positive correlation between the PEI and RS10 ratings was observed (r=0.829, P<.001), supporting the veracity of responses. Further validation was provided by a linear mixed model predicting PEI values from RS10 ratings (β=4.69, t_{863}=45.8, P<.001). Ratings across engineered empathy levels revealed significant differences in both PEI (χ^2_2=11.86, P=.003) and RS10 (χ^2_2=19.74, P<.001) with higher ratings observed for the high empathy level compared to the low and medium levels.

Linear Mixed Model Analyses

In the linear mixed model analyses, the percentages of variation attributed to participants were 63% and 46% for the PEI and RS10 ratings, respectively, while the percentages of variation attributed to the CA prototypes were only 1.5% and 2.5% for the PEI and RS10 ratings, respectively.

A linear mixed model was used to predict PEI ratings from engineered prototype empathy levels and participant characteristics (Table 2). PEI ratings were significantly higher for high rather than low prototype empathy levels (β=3.89, t_8=−2.34, P=.048), as were PEI measures for participants of Asian rather than Oceanic ethnicity (β=6.53, t_{285}=2.54, P=.01) and participants of the Other ethnicity category rather than Oceanic ethnicity (β=8.23, t_{286}=2.25, P=.03).

XSL•FO
RenderX

Table . Linear mixed model for Perceived Emotional Intelligence ratings.

Fixed effects	Coefficient (β)	SE	*t* test (*df*)	*P* value
Intercept	26.68	4.67	5.72 (276.03)	<.001
Engineered empathy levels				
Low versus high	−3.89	1.66	−2.34 (7.79)	.048
Medium versus high	−2.65	1.66	−1.59 (7.79)	.15
Age	−0.06	0.07	−0.8 (286.71)	.43
Gender identity (woman vs man)	2.64	1.9	1.36 (285.34)	.17
Ethnicity				
Asia versus Oceania	6.53	2.57	2.54 (285.23)	.01
Europe versus Oceania	2.99	2.24	1.33 (285.72)	.18
Other[a]	8.23	3.66	2.25 (285.88)	.03
Is English your home language? (yes vs no)	0.58	2.63	0.22 (285.51)	.83
IRSAD[b] decile	0.23	0.30	0.76 (287.01)	.45

[a]Sub-Saharan African, North African, the Middle East, and Peoples of Americas versus Oceania.

[b]Index for Relative Socio-Economic Advantage and Disadvantage.

A linear mixed model was also used to predict RS10 ratings from engineered prototype empathy levels and participant characteristics (Table 3). RS10 measures were significantly higher for high rather than low engineered empathy prototype levels (β=.82, t_8=−2.45, *P*=.04), as were the RS10 measures for other ethnicities versus Oceania origin (β=1.18, t_{284}=2.24, *P*=.03). In addition, women provided significantly higher RS10 values than men.

Table . Linear mixed model for Rater's Scale ratings.

Fixed effects	Coefficient	SE	*t* test (*df*)	*P* value
Intercept	4.05	0.69	5.85 (218.18)	<.001
Engineered empathy levels				
Low versus high	−0.82	0.33	−2.45 (8.17)	.04
Medium versus high	−0.49	0.33	−1.46 (8.17)	.18
Age	0.01	0.01	−0.98 (284.94)	.33
Gender identity (woman vs man)	0.54	0.28	1.94 (283)	.05
Ethnicity				
Asia versus Oceania	0.65	0.37	1.75 (282.84)	.08
Europe versus Oceania	0.20	0.32	0.63 (283.6)	.53
Other[a]	1.18	0.53	2.24 (283.87)	.03
Is English your home language? Yes versus no	0.16	0.38	0.43 (283.34)	.67
IRSAD[b] decile	0.03	0.04	0.76 (285.21)	.45

[a]Sub-Saharan African, North African, the Middle East, and Peoples of Americas versus Oceania.

[b]Index for Relative Socio-Economic Advantage and Disadvantage.

XSL•FO
RenderX

Discussion

Key Findings

This study aimed to investigate the complex issue of empathy perception and the factors that influence these perceptions. This novel study used experimental prototypes of voice-based CAs, engineered to reflect low, medium, or high levels of empathy in their vocals. The perception of empathy provided by study participants was compared against pre-established empathy levels. In addition, several demographic characteristics of the participants were evaluated to identify how these affected their perceptions of empathy.

The results indicate that perceptions of empathy do not always match with pre-established ("engineered") empathy levels, and it appears that it is difficult for people to identify small differences in empathy levels (eg, between low and medium empathy and between medium and high empathy) using only vocal cues. In addition, it has been shown that demographic characteristics impact the way that people perceive empathy. In particular, ethnic background and to some extent gender have been shown to affect how people perceive empathy in this study.

The evaluation of empathy was conducted using two scales: the PEI scale, which is comprised of multiple factors catering to the broader concept of emotional intelligence, and the RS10 scale, a single-factor scale providing a more direct evaluation of empathy levels (0 - 10). Most of the results from the linear mixed modeling were similar for the two scales, which provides more confidence in the conclusions. However, the participant gender effect was only marginally significant in the case of the RS10 rating and not significant in the case of the PEI rating, suggesting, perhaps, that the simpler RS10 scale is more sensitive to gender differences in perceptions of empathy.

Other studies have found that women have higher empathic ability compared to men [23], perhaps making them more attuned to the emotional status of others and, therefore, more likely to perceive empathy in others. Their higher empathic ability perhaps makes women more sensitive to empathic cues in others, leading them to provide more precise perception ratings. By far, most participants in this study were women, making it unfeasible for separate sex-based models to be developed. Future studies should try to obtain a more balanced representation of male and female participants so that an in-depth exploration of gender differences in empathy perception can be studied. In addition, a useful extension of this study would be to consider how nonbinary and other gender identities also perceive empathy, given that these minorities often suffer a greater mental health burden.

The PEI rating was perhaps more sensitive to the detection of ethnic differences than the RS10, because it detected significantly higher perceptions of empathy for participants with an Asian rather than Oceanic background, while the RS10 did not. As a measure of perceived emotional intelligence, the PEI measures a broader concept of empathy than the RS10. While empathy is the ability to put oneself in the place of another to understand their feelings [24], emotional intelligence is a set of skills that help us to effectively understand and express ourselves, understand and relate to others and cope with difficult situations [25]. The PEI scale's multifactorial nature, therefore, evaluates a wider range of attributes than the RS10, perhaps explaining its greater sensitivity for detecting ethnic differences.

The ethnic background of the study participants had a significant impact on the perceived empathy ratings for both the PEI scale and the RS10 rating scale. Most participants were of Oceanic background (ie, Australia and New Zealand), which was also the reference group for the linear mixed model. When the PEI ratings were considered, participants from Asia and the Other category (Sub-Saharan Africa, North Africa, the Middle East and Peoples of Americas) provided significantly higher ratings than those of Oceanic ethnicity. However, only the Other category showed significantly higher ratings when using the RS10 scale. The demonstration of empathy is strongly informed by cultural background that reflects either individualist or collectivist values [26]. Western backgrounds are cited as more individualist, respecting and prioritizing individual rights and well-being more than non-Western backgrounds and less likely to see empathy in others. However, social desirability holds less importance in Western countries [27], perhaps resulting in more critical ratings (lower ratings) when it comes to perceptions of empathy compared to collectivistic countries [28], suggesting that differences in perceived empathy may be due to differences in social desirability.

These findings raise challenges for researchers trying to understand the complex phenomenon of empathy perception when designing an empathic CA. How does one design a CA that exhibits empathy for all end users when perceptions of empathy are so subjective, depending on gender and ethnic background? Also, what does this say about "true" or "engineered" empathy ratings? Maybe this helps to explain the lack of accuracy in the perceived empathy ratings for medium empathy levels that was discovered in this study and the absence of research addressing this crucial question.

Limitations of the Study

The main limitation of the study was the poor generalizability of the sample of 306 participants who contributed to the electronic survey. The students from the REP were predominantly women (233/306, 76%). Most students in the sample were of Caucasian ethnicity from Australia or New Zealand (194/30,63%) and European background (46/306, 15%). In addition, only 5 of the 306 (1.6%) were of Aboriginal or Torres Strait Islander origin. Further, the students recruited for this study through the REP sample were mostly young adults, with 64% (197/306) within the age range of 18 - 32 years, also having a relatively higher socioeconomic status with the average IRSAD decile. Therefore, future researchers are encouraged to extend the sample to individuals with a wider age range and level of socioeconomic status. Also, it is important that participants with more varied backgrounds are considered (not just psychology students), also allowing for more variation in ethnicity. It is also recommended that a suitable measure of social desirability will be included in future surveys, which will allow for adjustment of social desirability bias in the perceived ratings of empathy.

XSL•FO
RenderX

In addition, the sample consisted entirely of first-year psychology students. Future studies might benefit from a broader representation of the population that includes direct users of a mental health helpline service. It was not possible to access the direct users of a mental health helpline service within the scope of the current project. Instead, this project aimed to obtain a broad perspective on the perception of empathy in voice-based CA prototypes. The student participants were sourced from a range of fields including business, humanities, and science-based undergraduate courses and were therefore reflective of a broad range of life experiences and interests. Furthermore, the study population was aged between 18 and 32 years; among the most vulnerable group to experience mental health emergencies [29].

The perception of empathy is often based on more than just vocal cues. Nonverbal cues such as facial expressions and mannerisms can provide valuable supplementary information that conveys intention and context. It may be that particular ethnic groups and cultural backgrounds require more than just the CA vocalizations used in the present study.

Future Opportunities

This study was focused on the design of a CA for deployment in a mental health environment, but these results hold broader implications because empathic communications are a key component of mental health care [30]. However, future research needs to acknowledge that the perception of empathy is a complex issue.

The significance of ethnicity as a predictor of perceived empathy suggests possible benefits in tailoring the training of health care professionals in the expression of empathy from a cross-cultural lens. Training guides for health care practitioners, teaching them how to adapt, demonstrate empathy, and personalize their communication styles based on gender and ethnic backgrounds, are also possibilities raised by this study. It is recommended that the development of these training materials be undertaken using a cocreation process involving relevant stakeholders.

The identification of CA empathy levels remains problematic. While our PEI and RS10 scales showed good agreement with each other and exhibited an ability to differentiate between low and high engineered empathy levels, they were unable to identify medium levels of empathy correctly. This might suggest a role for empathy education: promoting a greater understanding of how empathy is defined and reflected in speech. In addition, this study demonstrates that the display of empathy by CAs needs to be tailored to end-user characteristics. This will ensure that CAs used in various telephone mental health care applications can provide a more human-like user experience.

Acknowledgments

The research was enabled by Swinburne University of Technology as part of a postgraduate research award.

Data Availability

The voice-based conversational agent prototypes generated during the study are available in the Figshare repository [31], as is the dataset summary of participant demographic information [32].

Authors' Contributions

RS wrote the first draft of the manuscript, developed the CA prototypes, online survey and conducted the statistical analysis, which was guided by DM. DM and RI collaborated in the development of the CA prototypes and survey structure. All authors reviewed and provided revisions leading to the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Online survey for evaluating conversational agent prototypes.
[DOCX File, 31 KB - formative_v9i1e69329_app1.docx]

Multimedia Appendix 2
Summary of demographic features of participants and measures used for perceived empathy.
[DOCX File, 18 KB - formative_v9i1e69329_app2.docx]

Checklist 1
Checklist for Reporting Results of Internet E-Surveys (CHERRIES).
[DOCX File, 23 KB - formative_v9i1e69329_app3.docx]

References

XSL•FO
RenderX

1. Moreno-Poyato AR, Rodríguez-Nogueira Ó, MiRTCIME.CAT working group. The association between empathy and the nurse-patient therapeutic relationship in mental health units: a cross-sectional study. J Psychiatr Ment Health Nurs 2021 Jun;28(3):335-343. [doi: 10.1111/jpm.12675] [Medline: 32657511]
2. Khanjani Z, Mosanezhad jeddi E, Hekmati I, et al. Comparison of cognitive empathy, emotional empathy, and social functioning in different age groups. Aust Psychol 2015 Feb 1;50(1):80-85. [doi: 10.1111/ap.12099]
3. Artinger F, Exadaktylos F, Koppel H. In others' shoes: do individual differences in empathy and theory of mind shape social preferences? PLoS One 2014;9(4):e92844. [doi: 10.1371/journal.pone.0092844] [Medline: 24743312]
4. Cundiff NL, Nadler JT, Swan A. The influence of cultural empathy and gender on perceptions of diversity programs. J Leadersh Organ Stud 2009 Aug;16(1):97-110. [doi: 10.1177/1548051809334193]
5. Löffler CS, Greitemeyer T. Are women the more empathetic gender? The effects of gender role expectations. Curr Psychol 2023 Jan;42(1):220-231. [doi: 10.1007/s12144-020-01260-8]
6. Rosen JB, Brand M, Kalbe E. Empathy mediates the effects of age and sex on altruistic moral decision making. Front Behav Neurosci 2016;10:67. [doi: 10.3389/fnbeh.2016.00067] [Medline: 27147990]
7. Kapıkıran NA. Sources of ethnocultural empathy: personality, intergroup relations, affects. Curr Psychol 2023 May;42(14):11510-11528. [doi: 10.1007/s12144-021-02286-2]
8. Clarke E, DeNora T, Vuoskoski J. Music, empathy and cultural understanding. Phys Life Rev 2015 Dec;15:61-88. [doi: 10.1016/j.plrev.2015.09.001] [Medline: 26419700]
9. Inkster B, Sarda S, Subramanian V. An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: real-world data evaluation mixed-methods study. JMIR Mhealth Uhealth 2018 Nov 23;6(11):e12106. [doi: 10.2196/12106] [Medline: 30470676]
10. Kearns WR, Bertram J, Divina M, et al. Bridging the skills gap: evaluating an AI-assisted provider platform to support care providers with empathetic delivery of protocolized therapy. AMIA Annu Symp Proc 2023;2023:436-445. [Medline: 38222441]
11. Jami PY, Walker DI, Mansouri B. Interaction of empathy and culture: a review. Curr Psychol 2024 Jan;43(4):2965-2980. [doi: 10.1007/s12144-023-04422-6]
12. Olderbak S, Wilhelm O. Emotion perception and empathy: an individual differences test of relations. Emotion 2017 Oct;17(7):1092-1106. [doi: 10.1037/emo0000308] [Medline: 28358563]
13. Markovitch DG, Stough RA, Huang D. Consumer reactions to chatbot versus human service: an investigation in the role of outcome valence and perceived empathy. J Retail Consum Serv 2024 Jul;79:103847. [doi: 10.1016/j.jretconser.2024.103847]
14. Concannon S, Tomalin M. Measuring perceived empathy in dialogue systems. AI Soc 2024 Oct;39(5):2233-2247. [doi: 10.1007/s00146-023-01715-z]
15. Schmidmaier M, Rupp J, Cvetanova D, Mayer S. Perceived Empathy of Technology Scale (PETS): measuring empathy of systems toward the user. Presented at: CHI '24: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems; May 11-16, 2024; Honolulu, HI URL: https://dl.acm.org/doi/proceedings/10.1145/3613904 [accessed 2025-04-25] [doi: 10.1145/3613904.3642035]
16. Iyer R, Nedeljkovic M, Meyer D. Using voice biomarkers to classify suicide risk in adult telehealth callers: retrospective observational study. JMIR Ment Health 2022 Aug 15;9(8):e39807. [doi: 10.2196/39807] [Medline: 35969444]
17. Mayer JD, Salovey P, Caruso DR, et al. Measuring emotional intelligence with the MSCEIT V2.0. Emotion 2003 Mar;3(1):97-105. [doi: 10.1037/1528-3542.3.1.97] [Medline: 12899321]
18. Davey HM, Barratt AL, Butow PN, et al. A one-item question with a Likert or Visual Analog Scale adequately measured current anxiety. J Clin Epidemiol 2007 Apr;60(4):356-360. [doi: 10.1016/j.jclinepi.2006.07.015] [Medline: 17346609]
19. Eysenbach G. Improving the quality of Web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34. [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
20. NaturalReader. URL: https://www.naturalreaders.com/webapp.html [accessed 2023-05-25]
21. McKight PE. The Corsini Encyclopedia of Psychology: John Wiley & Sons; 2010.
22. Socio-economic indexes for areas (SEIFA), Australia. Australian Bureau of Statistics. 2023. URL: https://www.abs.gov.au/statistics/people/people-and-communities/socio-economic-indexes-areas-seifa-australia/latest-release [accessed 2023-04-27]
23. Pang C, Li W, Zhou Y, et al. Are women more empathetic than men? Questionnaire and EEG estimations of sex/gender differences in empathic ability. Soc Cogn Affect Neurosci 2023 Feb 28;18(1):nsad008. [doi: 10.1093/scan/nsad008] [Medline: 36807483]
24. Seïler NR, Craig P. Empathetic technology. In: Emotions, Technology, and Design: Elsevier 2016:55-81. [doi: 10.1016/B978-0-12-801872-9.00004-1]
25. Bar-On R. The Bar-On model of emotional-social intelligence (ESI). Psicothema 2006;18 Suppl:13-25. [Medline: 17295953]
26. Chopik WJ, O'Brien E, Konrath SH. Differences in empathic concern and perspective taking across 63 countries. J Cross Cult Psychol 2017 Jan;48(1):23-38. [doi: 10.1177/0022022116673910]
27. Zheng S, Masuda T, Matsunaga M, et al. Cultural differences in social support seeking: the mediating role of empathic concern. PLoS One 2021;16(12):e0262001. [doi: 10.1371/journal.pone.0262001] [Medline: 34969056]

XSL•FO
RenderX

28. Fang J, Prybutok V, Wen C. Shirking behavior and socially desirable responding in online surveys: a cross-cultural study comparing Chinese and American samples. Comput Human Behav 2016 Jan;54:310-317. [doi: 10.1016/j.chb.2015.08.019]
29. Adolescent and young adult health. World Health Organization. 2024. URL: https://www.who.int/news-room/fact-sheets/detail/adolescents-health-risks-and-solutions [accessed 2025-04-25]
30. Benster LL, Swerdlow NR. Pathways to empathy in mental health care providers. Adv Health Behav 2020;3(1):125-135. [doi: 10.25082/AHB.2020.01.004]
31. Sanjeewa R, Meyer D, Iyer R, Apputhurai P, Wickramasinghe N. Voice-based conversational agent prototypes. Swinburne figshare. URL: https://figshare.swinburne.edu.au/articles/media/Voice-based_conversational_agent_prototypes/28378010 [accessed 2025-04-25]
32. Sanjeewa R, Meyer D, Iyer R, Apputhurai P, Wickramasinghe N. Dataset summary of participant demographic information. Swinburne figshare. URL: https://figshare.swinburne.edu.au/articles/dataset/Dataset_summary_of_participant_demographic_information/28377959?file=52481534 [accessed 2025-04-25]

Abbreviations

CA: conversational agent
CHERRIES: Checklist for Reporting Results of Internet E-Surveys
EHSCC: Empathy Scale for Human-Computer Communication
IRSAD: Index for Relative Socio-Economic Advantage and Disadvantage
OTLA: On the Line Australia
PEI: Perceived Emotional Intelligence
PETS: Perceived Empathy of Technology Scale
REP: Research Experience Program
RS10: Raters' Scale

Edited by A Mavragani; submitted 27.11.24; peer-reviewed by R Chen; revised version received 25.02.25; accepted 02.03.25; published 07.05.25.

Please cite as:
Sanjeewa R, Iyer R, Apputhurai P, Wickramasinghe N, Meyer D
Perception of Empathy in Mental Health Care Through Voice-Based Conversational Agent Prototypes: Experimental Study
JMIR Form Res 2025;9:e69329
URL: https://formative.jmir.org/2025/1/e69329
doi:10.2196/69329

© Ruvini Sanjeewa, Ravi Iyer, Pragalathan Apputhurai, Nilmini Wickramasinghe, Denny Meyer. Originally published in JMIR Formative Research (https://formative.jmir.org), 7.5.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Real-World Effectiveness of Glucose-Guided Eating Using the Data-Driven Fasting App Among Adults Interested in Weight and Glucose Management: Observational Study

Michelle R Jospe[1], PhD; Martin Kendall[2], BEng; Susan M Schembre[1], PhD; Melyssa Roy[3], MBChB, PhD

[1]Georgetown Lombardi Cancer Center, Georgetown University, 3800 Reservoir Rd NW, Washington, DC, United States

[2]Optimising Nutrition LLC, Brisbane, Australia

[3]Department of Medicine, University of Otago, Dunedin, New Zealand

Corresponding Author:
Michelle R Jospe, PhD
Georgetown Lombardi Cancer Center, Georgetown University, 3800 Reservoir Rd NW, Washington, DC, United States

Abstract

Background: The Data-Driven Fasting (DDF) app implements glucose-guided eating (GGE), an innovative dietary intervention that encourages individuals to eat when their glucose level, measured via glucometer or continuous glucose monitor, falls below a personalized threshold to improve metabolic health. Clinical trials using GGE, facilitated by paper logging of glucose and hunger symptoms, have shown promising results.

Objective: This study aimed to describe user demographics, app engagement, adherence to glucose monitoring, and the resulting impact on weight and glucose levels.

Methods: Data from 6197 users who logged at least 2 days of preprandial glucose readings were analyzed over their first 30 days of app use. App engagement and changes in body weight and fasting glucose levels by baseline weight and diabetes status were examined. Users rated their preprandial hunger on a 5-point scale.

Results: Participants used the app for a median of 19 (IQR 9-28) days, with a median of 7 (IQR 3-13) weight entries and 52 (IQR 25-82) glucose entries. On days when the app was used, it was used a median of 1.8 (IQR 1.4-2.1) times. A significant inverse association was observed between perceived hunger and preprandial glucose concentrations, with hunger decreasing by 0.22 units for every 1 mmol/L increase in glucose (95% CI −0.23 to −0.21; $P<.001$). Last observation carried forward analysis resulted in weight loss of 0.7 (95% CI −0.8 to −0.6) kg in the normal weight category, 1 (95% CI −1.1 to −0.9) kg in the overweight category, and 1.2 (95% CI −1.3 to −1.1) kg in the obese category. All weight changes nearly doubled when analyzed using a per-protocol (completers) analysis. Fasting glucose levels increased by 0.11 (95% CI 0.09-0.12) mmol/L in the normal range and decreased by 0.14 (95% CI −0.16 to −0.12) mmol/L in the prediabetes range and by 0.5 (95% CI −0.58 to −0.42) mmol/L in the diabetes range. Per-protocol analysis showed fasting glucose reductions of 0.26 (SD 4.7) mg/dL in the prediabetes range and 0.94 (16.9) mg/dL in the diabetes range.

Conclusions: The implementation of GGE through the DDF app in a real-world setting led to consistent weight loss across all weight categories and significant improvements in fasting glucose levels for users with prediabetes and diabetes. This study underscores the potential of the GGE to facilitate improved metabolic health.

(JMIR Form Res 2025;9:e65368) doi:10.2196/65368

KEYWORDS

precision health; digital health; metabolic health; personalized nutrition; blood glucose self-monitoring; biological feedback; glucose; blood glucose; monitoring; self monitoring; dietary intervention; diet; app engagement; glucose monitoring

Introduction

Self-monitoring glucose levels before eating can help people recognize their physiological cues to eat, improve glycemic control, and help modify eating behaviors. This method has previously been referred to as hunger recognition [1] and hunger training [2], and is now termed glucose-guided eating (GGE) [3,4]. Early clinical trials confirmed that self-monitoring of glucose could be a feasible and effective approach to promote weight loss and insulin sensitivity in adults without diabetes [2,5,6]. Other subsequent trials of GGE in similar populations of adults without diabetes have demonstrated equally intriguing effects on weight loss, improved insulin resistance, and reduced glycemic variability [3,4,7]. Despite the promise, little is known about the real-world effectiveness and adherence to GGE.

XSL•FO
RenderX

GGE is designed as a short-term, intensive training program to help individuals develop long-lasting self-regulation skills in response to physiological hunger. By encouraging users to eat only when preprandial glucose levels are near or below a personalized threshold, GGE aims to enhance mindful eating habits, improve metabolic health, and reduce reliance on external cues for eating. Hunger is a natural physiological response that drives eating behavior, and GGE uses near-fasting glucose levels as a personalized proxy for hunger, reflecting short-term energy availability. Clinical trials implementing GGE typically involve a structured training period of up to 30 days, during which participants monitor their preprandial glucose levels using portable glucometers or wearable continuous glucose monitors. Previous research has demonstrated that eating when glucose is low can reduce glycemic variability, fasting insulin, and Homeostatic Model Assessment of Insulin Resistance (HOMA-IR) [3,4], with these improvements potentially lowering the risk of chronic diseases such as cardiovascular disease and certain cancers [8,9]. This initial training period has also been shown to result in benefits that persist for months beyond the structured intervention [10].

GGE also has a psychological and behavioral basis. Research has shown that eating when glucose is below a specified threshold helps people distinguish between eating driven by physiological hunger and eating driven by hedonic, emotional, or habitual factors [11]. Our previous results have shown that following GGE can reduce emotional eating, increase awareness of hunger and satiety, reduce the number of eating occasions, and help individuals explore the effects of different types of foods on their glucose levels [7,12].

The principles of GGE have been incorporated into Data-Driven Fasting (DDF; Optimising Nutrition LLC), an app that has been commercially available since 2021 [13]. Similar to clinical trials of GGE, DDF helps users learn to eat in response to physiological hunger by self-monitoring their preprandial glucose. Users are encouraged to eat only when their preprandial glucose levels are below their personalized threshold, as described by GGE [2]. However, the method for establishing personalized glucose thresholds in DDF differs from traditional GGE applications. Traditional GGE typically uses average morning fasting glucose levels to set a single static threshold that remains constant throughout the experience. In contrast, DDF uses preprandial glucose levels and continuously updates these thresholds throughout the challenge period. DDF also provides nutritional recommendations based on pre- and postprandial glucose levels to improve the glycemic quality of the diet and limit postprandial hyperglycemia at subsequent meals, an approach not previously used in GGE. Furthermore, DDF offers flexibility in adhering to glucose thresholds by providing an alternative threshold for times when users struggle to stay below their established threshold. In addition, self-monitoring of glucose in research settings has historically been done with paper-and-pen logs, whereas DDF uses a mobile app that features progress tracking and educational resources.

The increasing availability of glucose monitoring technology enhances the relevance and potential impact of DDF. Glucometers are readily available to the general population, and the accessibility of continuous glucose monitors is broadening, particularly with the release of over-the-counter models in several countries. Glucose monitoring technology provides valuable and immediate biological feedback about short-term energy availability and the metabolic consequence of carbohydrate intake. Consequently, it represents a feasible and practical intervention tool that can help people improve their metabolic health. While clinical trials are promising, it is unknown whether GGE, when facilitated by a commercially available mobile app and implemented in a free-living environment, can be adhered to or produce clinically meaningful effects on weight loss or improvements in blood glucose. Therefore, the goal of this study is to describe the demographic characteristics of DDF users, their engagement with the app and adherence to preprandial glucose monitoring, and the subsequent impact on weight and glucose levels, providing insights into the real-world implementation of GGE and its potential to improve metabolic health outcomes.

Methods

Study Design

This was an observational study of DDF app users from January 1, 2021, to December 8, 2023. Participants in this study were not recruited but were existing users of the commercially available DDF app. Data were collected from users who voluntarily engaged with the app as part of their normal use, and all data analyzed were deidentified entries made during the first 30 days of app use.

DDF App

DDF is a freely accessible web-based app with a subscription option [13]. Users could pay a small fee for a structured 30-day program that includes daily lessons, weekly live question and answer sessions, and community support. Users are encouraged to purchase a glucometer for glucose tracking, with the Contour Next (Ascensia Diabetes Care) being the recommended model. DDF has an interactive interface designed to help users follow GGE and provides nutritional guidance based on recorded preprandial (premeal) and postprandial (postmeal) self-monitored glucose levels. Perceived preprandial hunger ratings were recorded in the app using a 5-point scale, with 1 indicating “not hungry” (Figure 1).

XSL•FO
RenderX

Figure 1. Screenshots of the Data-Driven Fasting web-based app.

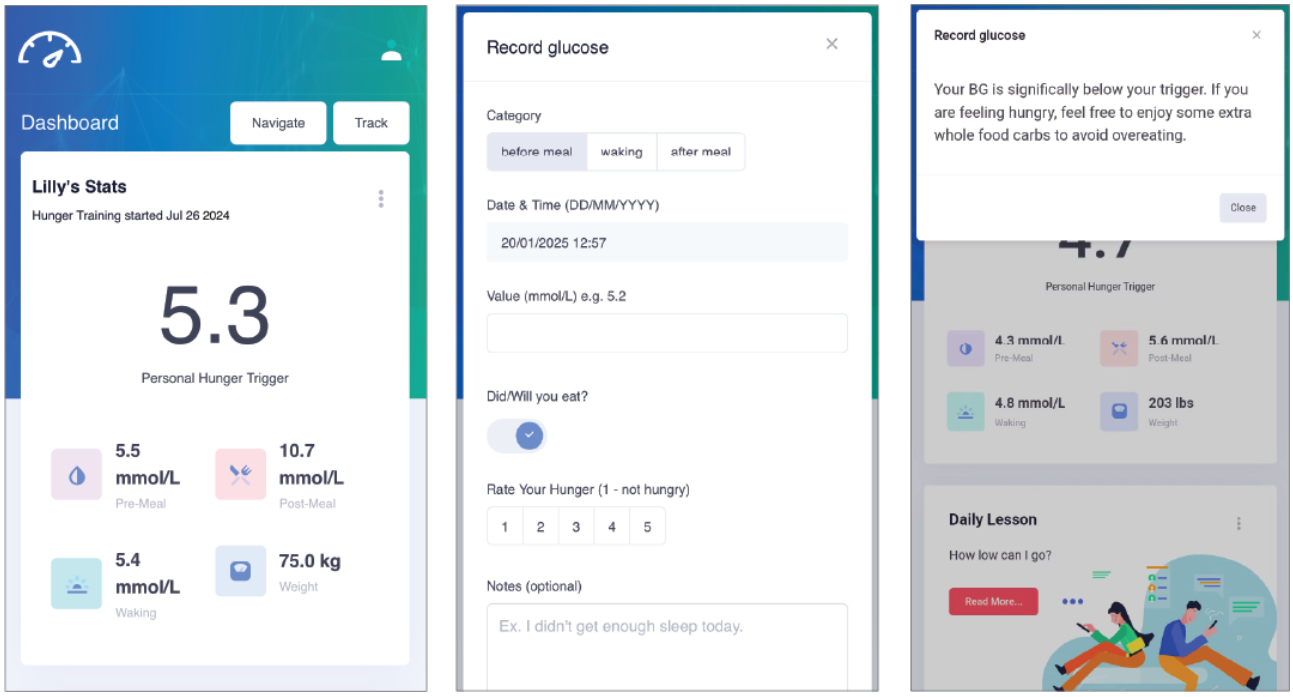

DDF Initiation

During the first 3 days of app use, participants are instructed to eat normally but to test their glucose levels before every eating occasion. Reported preprandial glucose levels during this time are averaged to establish a baseline glucose threshold. After this initial phase, users are encouraged to observe their hunger sensations and test their preprandial glucose levels only when hungry, and to delay eating until their glucose levels fall below their personalized threshold (ie, glucose-guided eating). After 5 days of glucose-guided eating, the threshold is adjusted daily by averaging reported preprandial glucose over the previous 7 days, with a maximum decrease of 0.5% per day. During times marked by higher-than-usual glucose levels due to conditions other than recent eating (eg, high stress, fatigue, pain, and intense exercise) users can opt to use an alternative, higher threshold (+0.6mmol/L or +10 mg/dL above their current threshold).

DDF Nutritional Guidance

The DDF app provides automated nutritional guidance based on recorded preprandial and postprandial glucose levels (relative to the current glucose threshold). When preprandial glucose is slightly above the threshold (≤0.4mmol/L or 7mg/dL), users are encouraged to prioritize meals with a higher protein composition. When glucose is just below the threshold, users are encouraged to eat “a nutritious meal.” If preprandial glucose is considerably below the threshold (>0.4mmol/L or 7mg/dL), users are encouraged to prioritize “whole food carbs” to quickly raise glucose into a normal range. Users who choose to also report postprandial glucose levels also receive feedback. Specifically, in the event postprandial glucose levels are more than 1.6 mmol/L or 30 mg/dL above the current threshold, the app suggests they “consider eating less of it (or even avoiding) that meal in the future.”

Outcome Measures

In addition to monitoring pre- and postprandial glucose levels and perceived hunger levels, the DDF app also allows users to track their fasting glucose levels and body weight. For this study, we extracted all available data entered into the DDP app by users during the first 30 days of use with day 0 representing the first day a preprandial glucose level was recorded. Implausible weight (<30 to >300 kg, and other unlikely values based on neighboring weights), height (<120 to >230 cm), and glucose values (<2.8 to >11.1 mmol/L) were excluded from the dataset.

Statistical Analysis

In an effort to examine the effectiveness of the DDF app, we aimed to assess the changes in body weight (primary outcome) and fasting glucose levels (secondary outcome) between day 0 and day 30 of using the DDF app. The analysis included users who had at least 2 days of recorded entries. A total of 6519 people used the DDF app, with 95.1% (n=6197) using it at least twice in the first 30 days and thus were included in the analytical sample. To ensure the robustness of the findings, a sensitivity analysis was conducted to compare results when including users with 2 or more days of data versus 7 or more days. The findings were consistent across both criteria, confirming that the inclusion of users with fewer days of data did not affect the outcomes. Descriptive results, such as app usage data, were summarized using appropriate measures of central tendency (eg, means or medians) and variability (eg, SDs or IQRs), depending on the data distribution. Participants were categorized based on their baseline BMI and fasting glucose levels using established clinical thresholds [14,15]. Changes in weight and fasting glucose levels between the first and last weeks of app use were assessed using paired comparisons with 95% CIs. For participants who did not have weight or glucose measures in the last week, the last observation carried forward (LOCF) method was applied. A linear mixed-effects model was used to examine the relationship between preprandial glucose and

XSL•FO
RenderX

perceived hunger, with preprandial glucose included as a fixed effect and a random intercept for participants to account for repeated measures within individuals. This approach allowed for the estimation of the overall association between preprandial glucose and hunger while accounting for variability in baseline hunger levels across users. Model estimates were reported with 95% CIs, and the significance of the fixed effect was assessed. Analyses were done with R (v4.2.2, R Foundation for Statistical Computing).

Ethical Considerations

This study was approved by the University of Otago Human Ethics Committee (reference: HD22/064). Participants were users of a commercially available app who consented to their anonymized data being used for research purposes through the app's privacy policy. All data were deidentified before analysis. No compensation was provided for participation.

Results

On average, participants in the analytical sample had a BMI in the obese range (≥30 kg/m^2), fasting glucose in the prediabetes range (5.6−6.9 mmol/L), and were predominantly female (Table 1). Participants primarily resided in North America (4590/6197, 74.1%), Europe (732/6197, 11.8%), and Australia and New Zealand (652/6197, 10.5%). No age data were collected, and height data were missing for 1152/6197 users (18.6%).

Those with only 1 day of entries (n=322) were excluded from the analysis and were similar to the analytical sample in terms of baseline weight (P=.08) and fasting glucose (P=.41) but had a lower BMI (24.5kg/m², P=.01) and included significantly more men and nonbinary individuals (P<.001).

The app was used by the analytical sample for a median of 19 (IQR 9-28) days, with 7 (IQR 3-13) weight entries and 52 (IQR 25-82) glucose entries, which were primarily preprandial glucose entries (Table 2). On days when the app was used, it was used a median of 1.8 (IQR 1.4-2.1) times. Using at least one preprandial glucose entry as a proxy for daily use, app engagement decreased from 100% at baseline to 47% by day 15% and 24% by day 30. Daily engagement patterns, broken down by fasting glucose category, exhibited similar trends, with consistent decreases across all groups. In contrast, weekly engagement trends show that the majority of users engaged with the app across all 4 weeks (Figure 2). Users rated their preprandial hunger on a 5-point scale. A significant inverse association was observed between perceived hunger and preprandial glucose concentrations, with hunger decreasing by 0.22 units for every 1 mmol/L increase in glucose (95% CI −0.23 to −0.21, P<.001). This effect corresponds to a negative linear relationship, analogous to an inverse Pearson correlation, while accounting for individual variability in baseline hunger levels (random intercept SD 0.63).

The analysis of weight changes for the total analytical sample and by baseline BMI category showed consistent reductions across all groups (Table 3). There were initially 4 users with a BMI in the underweight category, but none provided a weight in the last week of the intervention and were therefore excluded from the analysis. Using LOCF analysis, those with a BMI in the normal weight, overweight, and obese categories lost 0.7 kg, 1 kg, and 1.2 kg, respectively. All weight changes were nearly doubled when analyzed per protocol (completers) analysis (only considering those who provided weight measurements at week 0 and week 4).

Figure 3 shows the weight change by percentage of baseline weight using the last observation carried forward. Over 4 weeks, those with a BMI in the normal, overweight, and obese ranges lost 1.5% (95% CI 1.4%-1.6%), 1.7% (95% CI 1.6%-1.8%), and 1.7% (95% CI 1.6%-1.7%) of weight, respectively.

The analysis of fasting glucose changes for the total analytical sample and by baseline fasting glucose category revealed significant changes across all groups (Table 4 and Figure 4). Using LOCF, those with fasting glucose levels in the prediabetes and diabetes ranges had a reduction in fasting glucose of 0.14 mmol/L (2.5 mg/dL) and 0.5 mmol/L (9 mg/dL), respectively. Like with weight changes, the difference in fasting glucose was nearly doubled when only considering those with fasting glucose data at week 0 and week 4, with those in the prediabetes and diabetes ranges reducing their fasting glucose by 0.26 mmol/L (4.7 mg/dL) and 0.94 mmol/L (16.9 mg/dL), respectively. Conversely, those with fasting glucose levels in the normal range had an increase in glucose of 0.11 mmol/L (2 mg/dL) with the LOCF analysis and 0.23 mmol/L (4.1 mg/dL) for those with fasting glucose values at week 0 and week 4. Preprandial glucose decreased for all fasting glucose categories (Figure 5). Over 4 weeks, those with a fasting glucose in the normal, prediabetes, and diabetes ranges had a mean reduction in preprandial glucose of 0.16 (95%CI 0.13-0.19) mmol/L, 0.29 (95% CI 0.25-0.33) mmol/L, and 0.62 (95% CI 0.45-0.8) mmol/L, respectively.

XSL•FO
RenderX

Table . Baseline characteristics of Data-Driven Fasting app users from January 2021 to December 2023 (n=6197).

Characteristics	Values
Weight (kg) (n=4465), mean (SD)	83.9 (18.9)
BMI (kg/m^2) (n=3308), mean (SD)	30.1 (6.3)
Underweight (<18), n (%)	4 (0.1)
Normal weight (≥18->25), n (%)	690 (20.9)
Overweight (≥25), n (%)	1162 (35.1)
Obesity (≥30), n (%)	1452 (43.9)
Fasting glucose (mmol/L) (n=6092), mean (SD)[a]	5.8 (1.1)
Normal (≤5.5), n (%)	2723 (44.7)
Prediabetes (5.6 - 6.9), n (%)	2763 (45.4)
Diabetes (≥7), n (%)	606 (10)
Gender (n=6197), n (%)	
Women	5130 (82.8)
Men	1052 (17)
Nonbinary	15 (0.2)

[a] Based on American Diabetes Association classifications.

Table . Data-Driven Fasting app engagement (n=6249).

Variable	Total number in 30-day period, median (IQR)
Days used	19 (9-28)
Weight entries	7 (3-13)
Glucose entries	52 (25-82)
Fasting glucose entries	11 (5-21)
Preprandial glucose entries	27 (12-43)
Postprandial glucose entries	10 (5-20)

Figure 2. Data-Driven Fasting app engagement: (A) daily engagement by fasting glucose category and (B) weekly engagement.

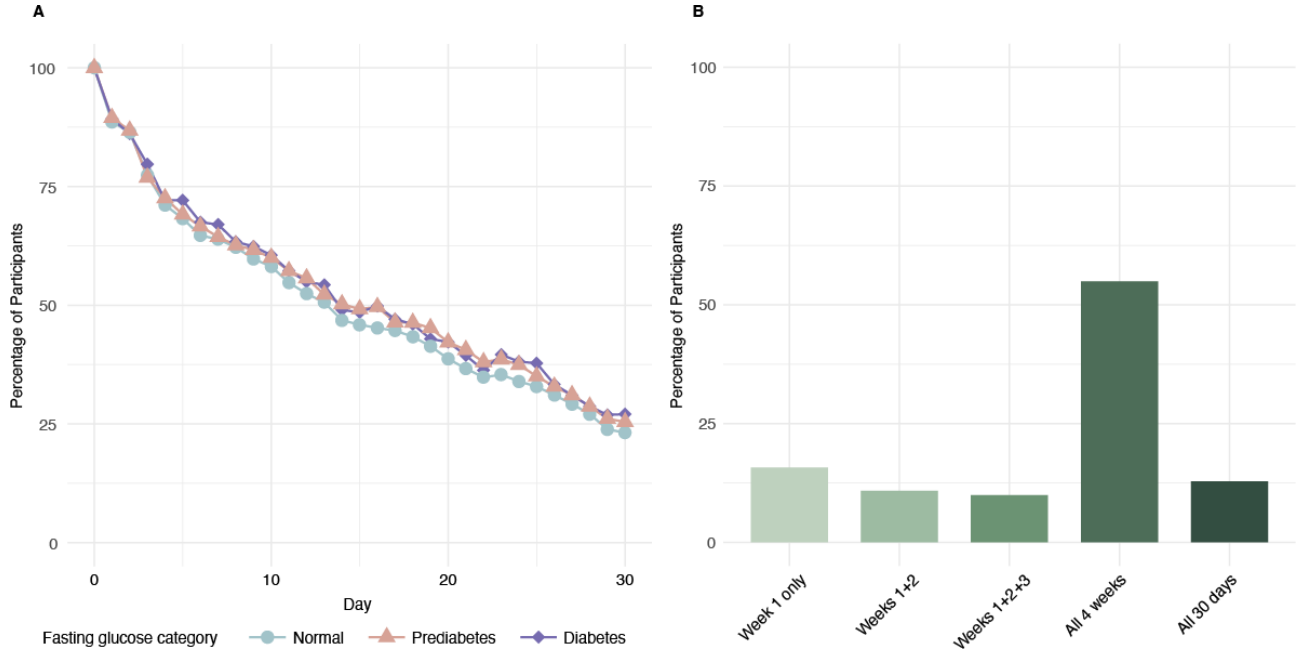

XSL•FO
RenderX

Table . Weight changes by BMI category.

Participant categories	Week 0		Week 4		Per protocol (those with weight at week 0 and 4)		Last observation carried forward	
	n[a]	Weight (kg), mean (SD)	n[a]	Weight (kg), mean (SD)	n[a]	Weight change (kg), mean (95% CI)	n[a]	Weight change (kg), mean (95% CI)
All participants	3304	83.9 (18.9)	2102	81.7 (17.7)	2102	−2.2 (−2.3 to −2.1)	3304	−0.9 (−0.9 to −0.8)
Normal weight	690	64.3 (7.3)	365	63.3 (6.8)	365	−1.4 (−1.5 to −1.2)	690	−0.7 (−0.8 to −0.6)
Overweight	1162	77 (9.3)	635	75.4 (9)	635	−1.8 (−1.9 to −1.6)	1162	−1.0 (−1.1 to −0.9)
Obesity	1452	98.7 (16.3)	757	95.7 (15.6)	757	−2.3 (−2.4 to −2.1)	1452	−1.2 (−1.3 to −1.1)

[a]Number of participants.

Figure 3. Weight trajectories by baseline BMI using last observation carried forward (n=3304).

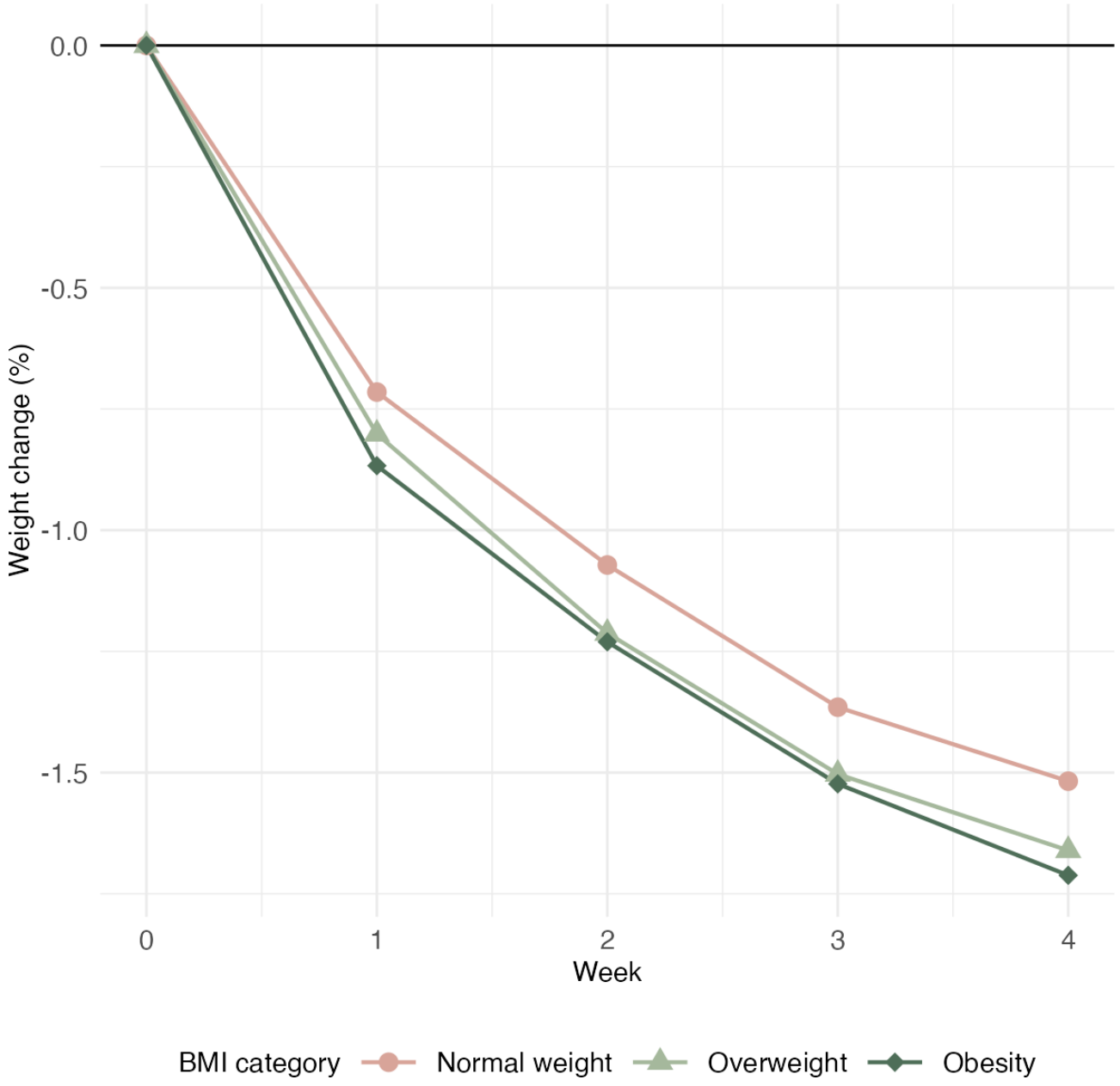

XSL•FO
RenderX

Table . Fasting glucose changes by fasting glucose category.

Participant categories	Week 0		Week 4		Per protocol (those with glucose at week 0 and 4)		Last observation carried forward	
	n[a]	Fasting glucose, mean (SD)	n[a]	Fasting glucose, mean (SD)	n[a]	Fasting glucose change (mmol/L), mean (95% CI)	n[a]	Fasting glucose change (mmol/L), mean (95% CI)
All participants	6027	5.8 (1.1)	3066	5.7 (1)	3022	−0.13 (−0.16 to −04.1)	6092	−0.06 (−0.08 to −0.05)
Normal	2698	5 (0.4)	1282	5.3 (0.6)	1268	0.23 (0.19 to 0.26)	2723	0.11 (0.09 to 0.12)
Prediabetes	2739	6.1 (0.4)	1458	5.8 (0.7)	1441	−0.26 (−0.3 to −0.23)	2763	−0.14 (−0.16 to −0.12)
Diabetes	590	8.1 (1.1)	326	7.1 (1.5)	313	−0.94 (−1.07 to −0.81)	606	−0.50 (−0.58 to −0.42)

[a]Number of participants.

XSL•FO
RenderX

Figure 4. Fasting glucose trajectories by baseline fasting glucose (n=6092).

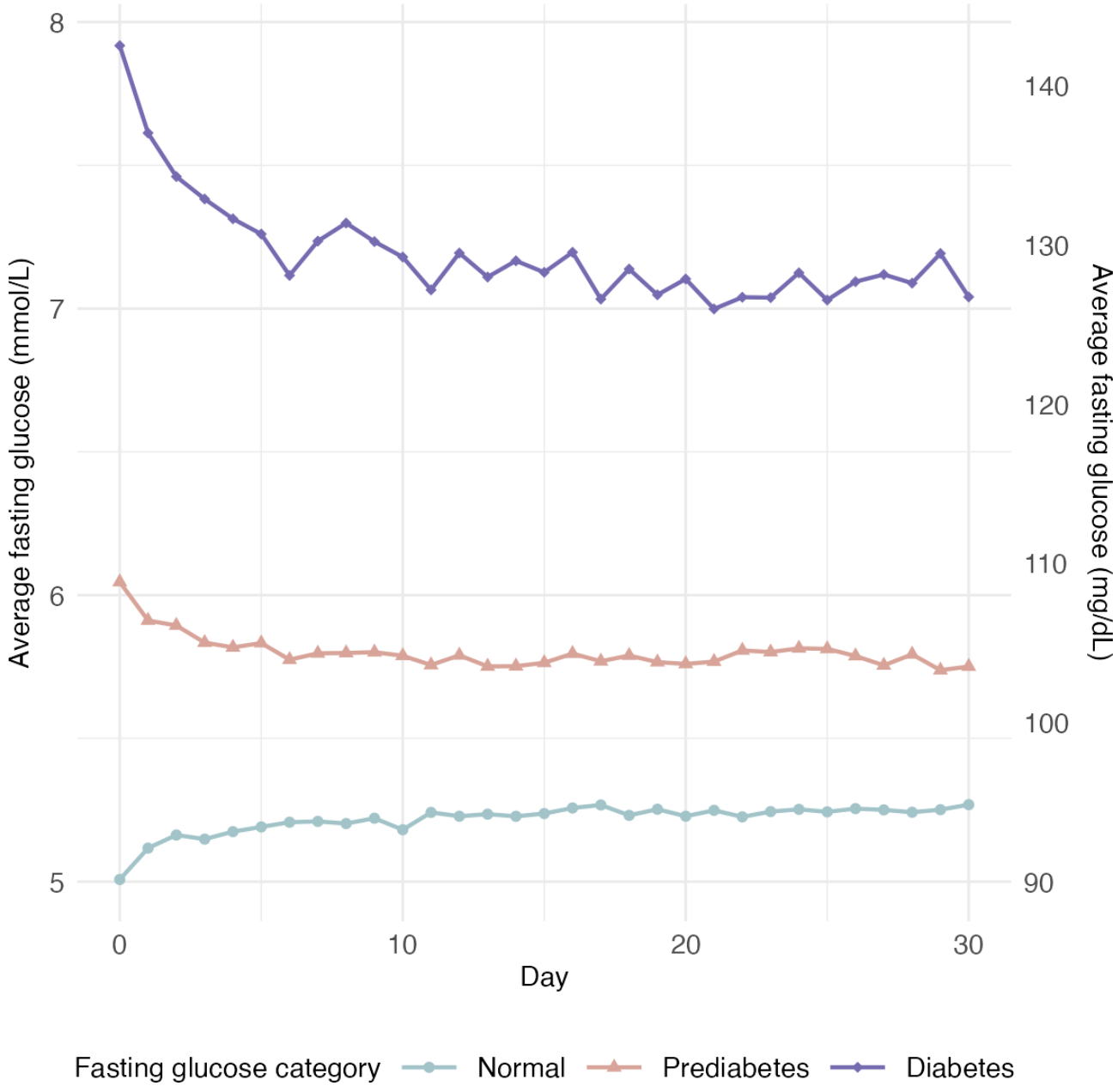

XSL•FO
RenderX

Figure 5. Preprandial glucose trajectories by baseline fasting glucose (n=6092).

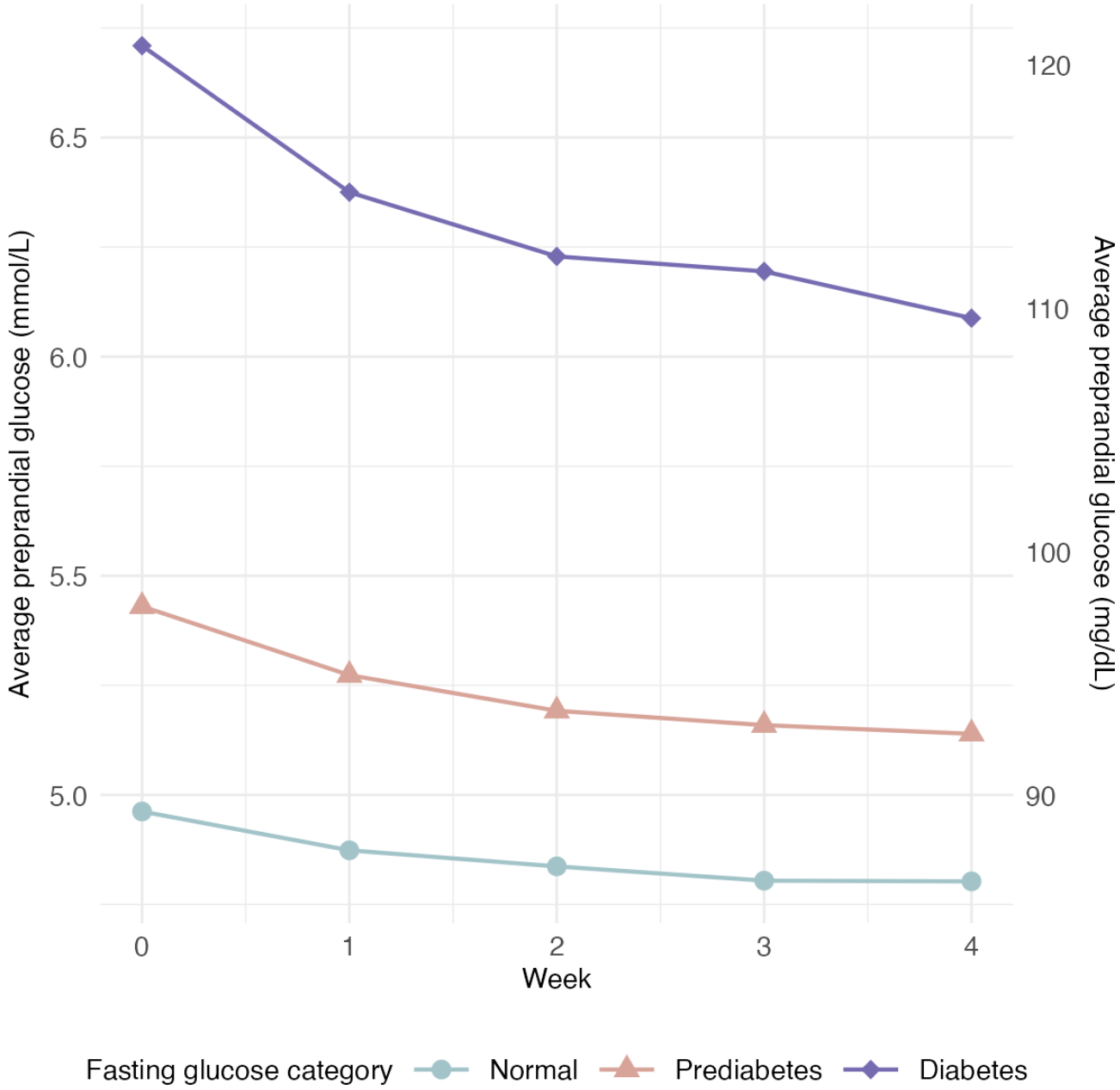

Discussion

Principal Findings

Previous research indicates that in experimental settings, GGE results in meaningful improvements in body weight, theory-based eating behaviors, and metabolic health outcomes, particularly those related to glucose metabolism. DDF, a commercially available mobile app that supports GGE, represents a scalable, personalized public health intervention with a growing evidence base. This study aimed to evaluate DDF app engagement and the effectiveness of DDF on changes in body weight and fasting glucose levels. The findings indicate that DDF produces modest improvements in body weight, particularly among users with higher weights at baseline, and clinically meaningful reductions in fasting glucose, particularly among users with fasting glucose in the diabetes range at baseline. These results support the effectiveness of DDF despite a steady decline in app use over the 30-day period.

DDF produced an average weight loss of 2.2 kg in the completer analysis, comparable with the results of studies using GGE in more controlled research settings of different durations. Average weight loss in the published clinical trials ranges from 1.5 kg in a 2-week feasibility trial for participants with overweight or obesity [2] to 4‑5 kg over 6 months [7,10]. Despite its simplicity and reliance on intermittent glucose monitoring, DDF achieved 30-day weight losses comparable with that reported in a study evaluating the January AI app, which uses continuous glucose monitoring and artificial intelligence–driven personalized dietary feedback [16]. This comparison underscores the effectiveness of GGE, which relies on intermittent preprandial glucose measurements and simpler personalized thresholds, in achieving comparable real-world weight loss. The modest inverse association between hunger and glucose observed in this study aligns closely with findings from previous

XSL•FO
RenderX

work conducted with participants across a range of BMI categories [2]. Results from this commercial implementation of GGE are reassuring, confirming the real-world potential of this approach to improve body weight.

In contrast, this is the first GGE study to report a clinically meaningful effect on fasting glucose [17]. The results here are similar to a recent 8-week clinical trial of time-restricted eating in a sample of adults with obesity, which resulted in a reduction in fasting blood glucose of 0.4 mmol/L with greater improvements among participants who had elevated blood glucose levels at baseline [18]. Interestingly, among users who had normal fasting blood glucose at baseline, use of the DDF app resulted in a slight increase in fasting glucose levels over the 30-day period (0.11 mmol/L to 0.23 mmol/L or 2 mg/dL to 4 mg/dL), although still within the normal range (<5.6 mmol/L or <100 mg/dL). Future studies should aim to explore this phenomenon, if replicated. Despite subtle fluctuations of fasting glucose among normoglycemic users of the DDF app, the majority of users experienced clinically meaningful improvements in glycemia. Future clinical trials should aim to replicate the beneficial effects of GGE on markers of glucose metabolism (and insulin resistance), particularly among people with prediabetes or diabetes.

The strengths of this study include the real-world application of GGE and a very large sample size that included a diverse range of baseline weights and fasting glucose levels. The results demonstrate the outcomes of the DDF method when implemented under real-world conditions. Data collected from people in free-living conditions are more representative of effectiveness and can be better generalized to the wider population. The DDF app provides an accessible and structured method for glucose-guided eating, leveraging user-friendly tools like glucometers to promote self-awareness and behavioral changes.

Despite these strengths, there was a lack of available data describing users in terms of age, ethnicity, and health status; however, the app could be easily modified to collect these data for future research purposes. In addition, our analyses examining the effect of DDF on body weight and fasting glucose levels were limited by the use of self-reported data. Adherence to the program could only be inferred from interaction with the DDF app and the absence of granular tracking tools, such as Google Analytics, limits the ability to fully evaluate user interaction with the app. It is possible that users continued to follow GGE without recording all measures or used the app incorrectly. While the median app usage of 19 days reflects real-world engagement patterns, it may not capture the long-term impact of sustained use. However, previous research has demonstrated that the benefits of GGE can persist for up to 5 months following the initial training period [10]. The DDF app focuses on helping users establish effective self-regulation strategies within the first 30 days, which may contribute to lasting benefits. Future research should investigate the durability of these outcomes over extended periods. The study design did not include a control group, which limits our ability to definitively conclude that changes in our outcomes were due to the use of DDF. While the use of the glucometer alone (and their companion apps) might have induced added benefits, research to date suggests that these benefits are likely to be minor without a structured program [19,20]. Compared with gold-standard measures like laboratory fasting glucose or HbA_{1c}, preprandial glucose monitoring with a glucometer offers a more accessible yet less precise approach [21]. This trade-off highlights the practicality of the DDF app for real-world settings, particularly for individuals seeking to manage weight and glucose levels without requiring expensive or complex equipment.

For users, the DDF app provides a simple and structured framework to improve metabolic health by promoting self-awareness of hunger and glucose patterns. For health care professionals, the app provides a unique opportunity to observe real-world glucose patterns and weight trends, offering insights that could inform personalized dietary and metabolic health interventions. Adherence and outcomes in the real world are likely to be affected by a range of external factors, such as social support, food security, family structure, and food preferences, which were not possible to measure in this setting. Nevertheless, these preliminary findings are positive, particularly in light of the real-world research setting, and suggest a more rigorous examination of the long-term potential benefit of DDF is warranted.

Conclusions

GGE, supported by the DDF app, is a simple intervention that could be easily implemented at scale with minimal support services. Glucometers are widely available and inexpensive, and continuous glucose monitors are available over-the-counter in several countries. Results from this study suggest that the DDF app may be most beneficial for people with overweight or obesity and those who have dysglycemia. The DDF app could be used alongside other lifestyle interventions to support behavior change among these populations. Future randomized controlled trials are warranted to confirm the metabolic benefits of DDF and GGE in specific populations, such as individuals with prediabetes or diabetes, and to assess the sustainability of the behavioral and metabolic improvements achieved through the app over the long term.

Acknowledgments

The authors thank Dr Jill Haszard for their expert guidance with the statistical analysis. This manuscript was prepared with language editing assistance from ChatGPT-4o (OpenAI).

Data Availability

The data that support the findings of this study are not publicly available due to commercialization restrictions.

XSL•FO
RenderX

Conflicts of Interest

MK is an employee of Optimising Nutrition, which facilitates Data-Driven Fasting and provided the data. MRJ reports former consultation to ZOE.

References

1. Ciampolini M, Bianchi R. Training to estimate blood glucose and to form associations with initial hunger. Nutr Metab (Lond) 2006 Dec 8;3(1):42. [doi: 10.1186/1743-7075-3-42] [Medline: 17156448]
2. Jospe MR, Brown RC, Roy M, Taylor RW. Adherence to hunger training using blood glucose monitoring: a feasibility study. Nutr Metab (Lond) 2015;12(1):22. [doi: 10.1186/s12986-015-0017-2] [Medline: 26075007]
3. Jospe MR, Liao Y, Giles ED, Hudson BI, Slingerland JM, Schembre SM. A low-glucose eating pattern is associated with improvements in glycemic variability among women at risk for postmenopausal breast cancer: an exploratory analysis. Front Nutr 2024;11(1301427):1301427. [doi: 10.3389/fnut.2024.1301427] [Medline: 38660060]
4. Schembre SM, Jospe MR, Giles ED, et al. A low-glucose eating pattern improves biomarkers of postmenopausal breast cancer risk: an exploratory secondary analysis of A randomized feasibility trial. Nutrients 2021 Dec 16;13(12):4508. [doi: 10.3390/nu13124508] [Medline: 34960058]
5. Ciampolini M, Lovell-Smith D, Bianchi R, et al. Sustained self-regulation of energy intake: initial hunger improves insulin sensitivity. J Nutr Metab 2010;2010(4):1-7. [doi: 10.1155/2010/286952] [Medline: 20721291]
6. Ciampolini M, Lovell-Smith D, Sifone M. Sustained self-regulation of energy intake. Loss of weight in overweight subjects. Maintenance of weight in normal-weight subjects. Nutr Metab (Lond) 2010 Jan 19;7(1):4. [doi: 10.1186/1743-7075-7-4] [Medline: 20148113]
7. Jospe MR, Taylor RW, Athens J, Roy M, Brown RC. Adherence to hunger training over 6 months and the effect on weight and eating behaviour: secondary analysis of a randomised controlled trial. Nutrients 2017 Nov 17;9(11):1260-1213. [doi: 10.3390/nu9111260] [Medline: 29149038]
8. Levitan EB, Song Y, Ford ES, Liu S. Is nondiabetic hyperglycemia a risk factor for cardiovascular disease? A meta-analysis of prospective studies. Arch Intern Med 2004 Oct 25;164(19):2147-2155. [doi: 10.1001/archinte.164.19.2147] [Medline: 15505129]
9. Wang M, Yang Y, Liao Z. Diabetes and cancer: epidemiological and biological links. World J Diabetes 2020 Jun 15;11(6):227-238. [doi: 10.4239/wjd.v11.i6.227] [Medline: 32547697]
10. Jospe MR, de Bruin WE, Haszard JJ, Mann JI, Brunton M, Taylor RW. Teaching people to eat according to appetite - does the method of glucose measurement matter? Appetite 2020 Aug 1;151:104691. [doi: 10.1016/j.appet.2020.104691] [Medline: 32246953]
11. Schembre SM, Liao Y, Huh J, Keller S. Using pre-prandial blood glucose to assess eating in the absence of hunger in free-living individuals. Eat Behav 2020 Aug;38:101411. [doi: 10.1016/j.eatbeh.2020.101411] [Medline: 32711350]
12. de Bruin WE, Ward AL, Taylor RW, Jospe MR. “Am I really hungry?” A qualitative exploration of patients’ experience, adherence and behaviour change during hunger training: A pilot study. BMJ Open 2019 Dec 31;9(12):e032248. [doi: 10.1136/bmjopen-2019-032248] [Medline: 31892654]
13. Data Driven Fasting App. URL: https://app.datadrivenfasting.com [accessed 2025-01-28]
14. American Diabetes Association Professional Practice Committee. 2. Classification and diagnosis of diabetes: Standards of Medical Care in Diabetes-2022. Diabetes Care 2022 Jan 1;45(Suppl 1):S17-S38. [doi: 10.2337/dc22-S002] [Medline: 34964875]
15. Weir CB, Jan A. BMI classification percentile and cut off points. In: StatPearls: StatPearls Publishing; 2014. [Medline: 31082114]
16. Veluvali A, Dehghani Zahedani A, Hosseinian A, et al. Impact of digital health interventions on glycemic control and weight management. NPJ Digit Med 2025 Jan 9;8(1):20. [doi: 10.1038/s41746-025-01430-7] [Medline: 39789102]
17. Schembre SM, Jospe MR, Bedrick EJ, et al. Hunger training as a self-regulation strategy in a comprehensive weight loss program for breast cancer prevention: a randomized feasibility study. Cancer Prev Res (Phila) 2022 Mar 1;15(3):193-201. [doi: 10.1158/1940-6207.CAPR-21-0298] [Medline: 34893531]
18. Peeke PM, Greenway FL, Billes SK, Zhang D, Fujioka K. Effect of time restricted eating on body weight and fasting glucose in participants with obesity: results of a randomized, controlled, virtual clinical trial. Nutr Diabetes 2021 Jan 15;11(1):1-11. [doi: 10.1038/s41387-021-00149-0]
19. Mannucci E, Antenore A, Giorgino F, Scavini M. Effects of structured versus unstructured self-monitoring of blood glucose on glucose control in patients with non-insulin-treated type 2 diabetes: a meta-analysis of randomized controlled trials. J Diabetes Sci Technol 2018 Jan;12(1):183-189. [doi: 10.1177/1932296817719290] [Medline: 28697625]
20. Parkin CG, Buskirk A, Hinnen DA, Axel-Schweitzer M. Results that matter: structured vs. unstructured self-monitoring of blood glucose in type 2 diabetes. Diabetes Res Clin Pract 2012 Jul;97(1):6-15. [doi: 10.1016/j.diabres.2012.03.002] [Medline: 22464874]

XSL•FO
RenderX

21. Sacks DB, Arnold M, Bakris GL, et al. Guidelines and recommendations for laboratory analysis in the diagnosis and management of diabetes mellitus. Diabetes Care 2023 Oct 1;46(10):e151-e199. [doi: 10.2337/dci23-0036] [Medline: 37471273]

Abbreviations

DDF: Data-Driven Fasting
GGE: glucose-guided eating
HOMA-IR: Homeostatic Model Assessment of Insulin Resistance
LOCF: last observation carried forward

Edited by A Mavragani; submitted 14.08.24; peer-reviewed by JE Teixeira, M Grady; revised version received 20.02.25; accepted 21.02.25; published 08.05.25.

Please cite as:
Jospe MR, Kendall M, Schembre SM, Roy M
Real-World Effectiveness of Glucose-Guided Eating Using the Data-Driven Fasting App Among Adults Interested in Weight and Glucose Management: Observational Study
JMIR Form Res 2025;9:e65368
URL: https://formative.jmir.org/2025/1/e65368
doi:10.2196/65368

© Michelle R Jospe, Martin Kendall, Susan M Schembre, Melyssa Roy. Originally published in JMIR Formative Research (https://formative.jmir.org), 8.5.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Boosting Digital Health Engagement Among Older Adults in Hong Kong: Pilot Pre-Post Study of the Generations Connect Project

Aaron Wan Jia He[1], MPT, PhD; Runqi Yuan[1], MEd; Tzu Tsun Luk[2], PhD; Kelvin Man Ping Wang[2], PhD; Sophia Siu Chee Chan[1], PhD

[1]School of Public Health, Li Ka Shing Faculty of Medicine, University of Hong Kong, Rm 520, 5/F, Academic Building, No. 3 Sassoon Road, Pokfulam, Hong Kong, China (Hong Kong)

[2]School of Nursing, Li Ka Shing Faculty of Medicine, University of Hong Kong, Hong Kong, China (Hong Kong)

Corresponding Author:
Sophia Siu Chee Chan, PhD
School of Public Health, Li Ka Shing Faculty of Medicine, University of Hong Kong, Rm 520, 5/F, Academic Building, No. 3 Sassoon Road, Pokfulam, Hong Kong, China (Hong Kong)

Abstract

Background: Older adults' utilization of digital health care remains low despite a high demand for regular health services. Easily accessible eHealth interventions designed for older adults are needed.

Objective: This study aimed to examine the feasibility and effectiveness of an intergenerational, home-based eHealth literacy intervention package on older adults in Hong Kong.

Methods: In this study, 101 older adults (n=64, 63.4% female) with a median age of 80 (IQR 77-85) years received an intergenerational, home-based eHealth literacy intervention package, delivered by trained university student interventionists. The intervention (median 60, IQR 40.8-70 minutes) included personalized guidance on using mobile health apps, QR code scanners and instant messaging, and access to online health information, along with recommendations for physical and mental well-being. Following the intervention, a daily health-coaching message was sent to older adults via WhatsApp for 14 days. eHealth literacy, health, and lifestyle were assessed at baseline and at a 2-week follow-up using paired *t* tests.

Results: Retention rate for the 2-week follow-up was 70.3% (71/101). Compared to baseline, eHealth literacy scores increased by 2.39 points (P=.11; Cohen d=0.20), and daily smartphone use rose by 0.45 hours (P=.07; Cohen d=0.05). Participants self-reported increased physical activity (50/71, 70%), more frequent viewing of health videos (43/70, 61%), and improved handwashing practices (39/71, 55%). The intervention achieved a high satisfaction rating of 4.32 out of 5.

Conclusions: The intergenerational, home-based eHealth literacy intervention package was feasible and acceptable, showing promise for increasing older adults' engagement with digital health care resources and promoting healthy behaviors. Future studies should explore longer-term effects and ways to further improve the intervention.

(JMIR Form Res 2025;9:e69611) doi:10.2196/69611

KEYWORDS

older adults; eHealth literacy; home-based intervention; intergenerational learning; health promotion

Introduction

Hong Kong is undergoing a rapid demographic shift; the proportion of older adults has increased from 13.3% in 2011 to 19.6% in 2021 [1], and it is expected to reach 36.0% by 2046 [2]. Unfortunately, the rise in life expectancy comes with an increase in unhealthy years of life even among industrialized states [3], creating significant challenges for health care systems. Globally, health care expenditure has doubled in the past 2 decades, reaching 9.8% of global gross domestic product (GDP) in 2019 [4]. In 2022, Hong Kong spent 7.3% of its GDP on health care, one of the highest expenditures on health care in Asia [5]. Traditionally, people have preferred seeking care at large public hospitals even for simple health problems due to limited publicly funded primary care facilities [6]. This model of health care is unsustainable as tertiary care providers become overloaded while primary health care remains underutilized [7].

Chronic conditions and noncommunicable diseases (NCDs) affect over 70% of Hong Kong's older adults [8], who require frequent medical care [9]. Additionally, mental health concerns such as depression are prevalent [10-12], exacerbated by social isolation [13] and loneliness [9,14,15], particularly during the COVID-19 pandemic. In response to these chronic health and mental health challenges, and in the context of primary health care reform in Hong Kong, improving older adults' ability to access and engage with digital health care services has become essential.

XSL•FO
RenderX

eHealth literacy, defined as the "ability to seek, find, understand, and appraise health information from electronic sources" [16], has emerged as a key factor in improving health outcomes. It is thus imperative to identify and support population groups at risk of low eHealth literacy [17]. Older adults, despite their high demand for health care, often struggle to adopt digital health care tools due to a lack of confidence, interest, or familiarity with technology [18-20]. These challenges could widen the digital divide, leaving many older adults disconnected from essential eHealth resources [21]. Tailored interventions are needed to promote eHealth literacy and engage older adults more effectively with digital health care systems.

Existing studies have shown that targeted interventions can enhance eHealth literacy, promote healthy behaviors, and sustain mental health. However, these interventions often rely on convenient samples recruited from a limited pool of participants and are implemented in centralized locations, such as classrooms in community senior centers [22-25]. Such design limits accessibility for underprivileged or home-bound older adults, who may be excluded from traditional classroom-based interventions. Additionally, quantitative evidence remains scarce for home-based interventions designed to address the unique challenges faced by these populations [26-29].

This pilot study explored the effect of an intergenerational, home-based intervention package focused on improving older adults' engagement with mobile eHealth resources, while also aiming to better their lifestyles and overall well-being. The intervention is grounded in the Information-Motivation-Behavioral Skills model, a behavior-oriented theory widely used in health promotion programs [30,31]. The Information-Motivation-Behavioral Skills model consists of 4 domains: information, motivation, behavioral skills, and behavior, positing that information, motivation, and behavioral skills directly and indirectly influence health behavior [31]. This model was chosen as it provides a comprehensive framework to address the complex factors influencing eHealth literacy in older adults. The intergenerational and home-based components of the intervention are designed to enhance its effectiveness. Intergenerational approaches have been shown to improve digital literacy and reduce social isolation among older adults by leveraging the technological proficiency of younger generations [32]. Home-based interventions are particularly beneficial for older adults as they provide a familiar and comfortable environment, reducing barriers to participation and increasing engagement [33]. By incorporating an intergenerational approach, the intervention leverages the tech-savviness of the younger generation, often referred as "digital natives," to enhance older adults' self-efficacy in technology use, a critical component of eHealth interventions [20]. Trained student interventionists visited older adults in their homes, providing practical knowledge on utilizing eHealth resources and adopting health-promoting behaviors. The home visits also aimed to bridge the generational gap and reduce loneliness among older adults [33]. This study aims to examine the feasibility and effectiveness of a nursing student–led, home-based eHealth intervention to better serve older adults across Hong Kong.

Methods

This pilot study is part of the Generations Connect Project [34], a large-scale, community-based initiative aiming to train over 1000 nursing students providing community service and behavioral coaching to 10,000 older adults in Hong Kong.

Study Design

This is a pilot, single group, pre-post study to evaluate the feasibility and effectiveness of intergenerational, home-based eHealth interventions on eHealth literacy among older adults in Hong Kong. Data collection occurred both in-person (baseline) and via telephone (2-week follow-up) in November 2022. Consolidated Standards of Reporting Trials (CONSORT) reporting guidelines and the extension to randomized pilot and feasibility trials were followed [35] (Checklist 1).

Participants

Older adults were recruited through collaborations with local nongovernment organizations covering 18 districts across Hong Kong. Older adults were eligible if they were (1) 65 years or older, (2) spoke Chinese, and (3) were physically, mentally, and cognitively capable of engaging in simple physical tasks and (4) understanding survey questions. Older adult participants were told the availability of home-based eHealth and health behavior coaching sessions, which remove physical barriers commonly seen in centralized interventions.

Recruitment involved partnerships with community centers, senior clubs, and health care facilities. Information sessions introduced the study, followed by screening interviews for interested individuals. Those meeting the inclusion criteria were scheduled for baseline assessments. Home visits were arranged for those unable to attend group sessions.

Ethical Considerations

This study was conducted in accordance with ethical guidelines for research involving human participants. The research protocol was approved by the Institutional Review Board of the University of Hong Kong (UW 22 - 693). Informed consent was ensured through comprehensive information sheets and verbal explanations, with consent forms signed before any procedures. The consent form included information about the study's purpose, procedures, risks, and benefits, as well as the participant's right to withdraw at any time without penalty.

As a token of appreciation for their time and participation, older adults enrolled in the study received HKD $50 (US $6.44; HKD $1=US $0.13) gift bags as an incentive during the initial student visits.

All data collected during the study were deidentified to protect participants' privacy. Data were initially recorded on paper and subsequently digitized using Qualtrics, with all identifying information removed. Access to the data was restricted to the research team, and all digital data were stored securely.

Training Student Interventionists

Student interventionists were current full-time students enrolled in the Bachelor of Nursing or Master of Science in Nursing programs at the University of Hong Kong, who were fluent in

XSL•FO
RenderX

Cantonese. Students were recruited through internal university emails and on-campus promotional materials such as posters. Recruited students were trained by research and clinical faculty members from the School of Nursing before attending home visits. A special-designed curriculum was developed with learning materials covering (1) the aging population and primary health care's new journey in Hong Kong, (2) effective communication techniques with older adults, (3) assistance with digital devices and mobile-based eHealth services, (4) knowledge on NCD prevention by promoting healthy lifestyles, (5) COVID-19 preventive measures, (6) strategies for building positive mental health and well-being, and (7) home visit protocol and general safety practice guidelines.

Measurements

Participants' data were collected by a self-administered survey at baseline and a telephone survey at 2 weeks after the intervention. The baseline questionnaire covered demographics (9 items), physical health (4 items), exercise habits (8 items), dietary practices (5 items), smoking (3 items), alcohol consumption (1 item), psychosocial health (16 items), eHealth literacy and mobile phone and app usage (17 items), and COVID-19 preventive measures (13 items). The primary outcome was eHealth literacy assessed by the 8-item eHealth Literacy Scale (eHEALS) [30], ranging from 8 to 40, with higher scores indicating better eHealth literacy. This scale is widely regarded as reliable and is frequently used in studies involving Chinese older adults, including those in Hong Kong [36-38].

Secondary outcomes included mental well-being, evaluated by the World Health Organization–Five Well-Being Index (WHO-5 [39]; multiplied by 4 to create a convenient scale of 0 to 100; higher scores indicate better well-being), and the 3-item University of California, Los Angeles (UCLA) Loneliness Scale [40] (3 to 9; lower scores indicate lower loneliness). Physical well-being was measured by a single item (1 to 5; higher scores indicate better physical well-being). Physical activities were assessed using a modified version of the International Physical Activity Questionnaire–Short Form [41].

Feasibility outcomes included (1) the recruitment rate (number of participants enrolled/number of eligible participants) and (2) the retention rate (number of participants who completed the follow-up/number of participants enrolled). Two additional feasibility outcomes were assessed: (1) the time required to conduct assessments and deliver the intervention package and (2) student interventionists' compliance with the intervention. Acceptability was assessed using a 5-point Likert scale measuring satisfaction with the home-based interventions and the 14-day digital follow-up message, ranging from 1 (strongly dissatisfied) to 5 (strongly satisfied).

The 2-week follow-up survey included similar items (demographics excluded) with 4 additional questions on perceived behavioral changes that occurred compared to previsit and 2 more questions on their satisfaction with the home-based interventions and 14-day digital follow-up messages.

Home-Based Assessment, Intervention Package, and 2-Week Follow-Up

Trained student interventionists conducted assessments and delivered interventions during home visits (median 60, IQR 40.8-70 minutes). Written consent was obtained from participants before the assessment began.

The assessment process was flexible, adapting to older adults' responses. Student interventionists conducted the baseline assessment in a semistructured manner, engaging participants in natural conversations while covering questionnaire items. Questions were read aloud to older adults, and responses were given verbally.

The intervention was similarly semistructured, covering four major themes:

1. eHealth Literacy: instructions on using WhatsApp to contact health professionals and connect with family or friends, using QR code scanners to acquire health information from printed materials, and downloading and using mobile health service apps, such as HA Go and Electronic Health Record Sharing System (eHealth).
2. NCD prevention: advice on healthy lifestyle changes, including diet, smoking cessation, alcohol moderation, and physical activities.
3. Psychosocial health: promotion of positive thinking, social engagement (adhering to COVID-19 regulations), and relaxation techniques.
4. COVID-19 prevention: education on hygiene, handwashing techniques, vaccination benefits, mask-wearing, and rapid antigen tests.

Health education leaflets with QR codes linking to online resources were provided and stored in a small blue folder. Student interventionists explained the leaflet content and tailored discussions to each older adult's health needs. They demonstrated smartphone use, including scanning QR codes and installing health apps, ensuring participants understood and could perform these tasks independently.

At the end of the intervention, student interventionists created dedicated one-on-one WhatsApp group chats with the older adults, sending daily health-related messages for the next 14 days. Participants were encouraged to engage with the content and share their opinions, using simple responses like emojis or voice messages.

A follow-up survey was conducted via telephone by the same student interventionist 2 weeks after the visit.

Sample Size

Given the exploratory nature of this pilot study, a sample size of 100 older adults was deemed appropriate to provide preliminary insights into the feasibility and potential effectiveness of the intergenerational, home-based eHealth literacy intervention. This sample size allowed for a meaningful assessment of the intervention's impact on eHealth literacy, health behaviors, and overall well-being, while also informing the design of future, larger-scale studies.

XSL•FO
RenderX

Data Analysis

Data were initially recorded on paper and subsequently digitized using Qualtrics. Data were analyzed using IBM SPSS Statistics (version 29.0). Descriptive statistics were calculated to summarize the data, using means, standard deviations, and standard errors or frequencies and percentages as appropriate. Paired *t* tests assessed pre- to posttest changes, and Wilcoxon tests were used for nonnormal data.

To assess potential biases due to loss to follow-up, baseline characteristics of participants who completed the follow-up (n=71) were compared with those who did not (n=30) using independent sample *t* tests for continuous variables (eg, age, eHEALS) and chi-square tests for categorical variables (eg, gender, education). Missing data were addressed using multiple imputation (MI) with 50 imputations, as missing completely at random (MCAR) tests did not reject the missing at random assumption (P=.85 for all variables; P=.08 for *t* test variables) [42,43]. The MI method was chosen for its ability to reduce bias under the assumption that data are missing at random and to provide more accurate standard errors and confidence intervals [44].

The imputation model included demographic variables (eg, age, education) and several auxiliary variables, such as self-rated happiness, weekly fruit and vegetable consumption, received COVID-19 vaccine doses, and frequencies of going outside. All reported pre- and posttest values are based on MI to ensure robust findings and mitigate potential biases associated with missing data. Complete case analyses were included as sensitivity analysis. Statistical significance was set at P<.05.

Results

Recruitment and Baseline Demographics

From the beginning of October to November 2022, 104 participants were screened as eligible; 3 (2.9%) were excluded due to being younger than 65 years, and the remaining 101 (97.1%) participants consented to the study (Figure 1). The baseline demographic characteristics of the 101 older adults are shown in Table 1. The median age was 80 (IQR 77-85) years, with the oldest participant being 93 years old. Nearly two-thirds (64/101, 63.4%) were female. Only 4.0% (4/101) reported receiving tertiary education such as college. Out of the 101 participants, 41 (40.6%) had lost their spouse, whereas 55 (54.5%) were married. Fifty-one (51.0%) participants reported living with their family. Ninety-eight (97.0%) participants reported having less than HKD $25,000 (US $3225.73) per month of household income, 42 (42.0%) reported claiming Comprehensive Social Security Assistance, and all older adults reported living in government-subsidized public housing. Out of all participants, 95 (94.1%) were retired. Physical well-being, mental well-being, and lifestyle habits of older adult participants, along with their medical history, socialization patterns, and habits, are detailed in Tables S1–S3 in (Multimedia Appendix 1). The retention rate at 2 weeks was 70.3% (71/101).

XSL•FO
RenderX

Figure 1. Flow diagram of participant enrollment and study procedures in the intergenerational, home-based eHealth intervention for Hong Kong's older adult population (October-November 2022).

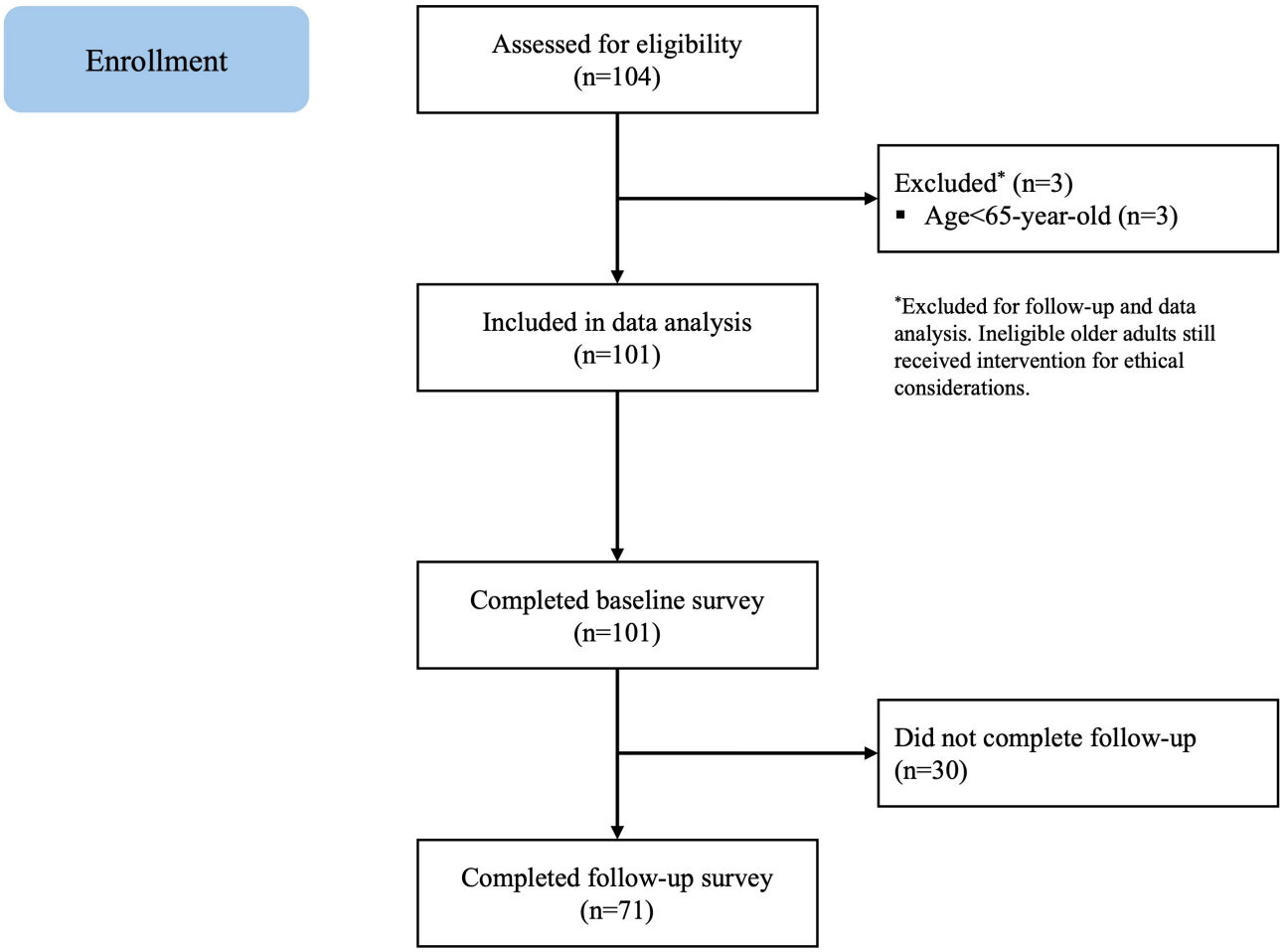

XSL•FO
RenderX

Table . Demographic and baseline characteristics of 101 older adults participating in the home-based eHealth intervention study conducted across Hong Kong's 18 districts (October-November 2022).

Characteristic	Value
Age (years), median (IQR)	80 (77, 85)
65-74, n (%)	10 (9.9)
75-84, n (%)	64 (63.4)
≥85, n (%)	27 (26.7)
Gender, n (%)	
Male	37 (36.6)
Female	64 (63.4)
Education level (n=100; missing=1), n (%)	
Primary or below	60 (60)
Secondary	36 (36)
Tertiary	4 (4)
Marital status, n (%)	
Single	1 (1.0)
Married	55 (54.5)
Divorced	4 (4.0)
Widowed	41 (40.6)
Household income (HKD $)[a], n (%)	
No income	62 (61.4)
<25,000	36 (35.6)
>25,000	3 (3.0)
Claiming Comprehensive Social Security Assistance (n=100; missing=1), n (%)	
No	58 (58)
Yes	42 (42)
Employment status, n (%)	
Part-time	1 (1.0)
Retired	95 (94.1)
Housewife or homemaker	5 (4.9)
Housing type	
Public Housing	101 (100)
Currently living with family (n=100; missing=1), n (%)	
Yes	49 (49)
No or no family	51 (51)
eHealth Literacy Scale (eHEALS) score (n=100; missing=1), mean (SD)	21.36 (7.85)
Owned smartphone with internet access (n=100; missing=1), n (%)	
Yes	89 (89)
No	11 (11)
Smartphone screen time per day (n=88; missing=1), mean (SD)	1.90 (1.38)
<1 hour, n (%)	55 (63)
1 - 2 hours, n (%)	11 (13)
2 - 3 hours, n (%)	7 (8)
3 - 4 hours, n (%)	6 (7)

XSL•FO
RenderX

Characteristic	Value
>4 hours, n (%)	9 (10)
Reason for using smartphone[b] (n=88; missing=1), n (%)	
Staying in touch with family and friends	73 (83)
Watching videos	32 (36)
Browsing news	15 (17)
Listening to music	14 (16)
Gaming	10 (11)
Reading	4 (5)
Online shopping	1 (1)
Acquired or received health care–related information from online sources (n=86; missing=3), n (%)	
Yes	25 (29)
No	61 (71)
eHealth mobile apps used[b] (n=87; missing=2), n (%)	
LeaveHomeSafe (安心出行)	84 (97)
eHealth (醫健通)	26 (30)
HA Go	9 (10)
TouchMed (e藥通)	3 (3)
None of the above	2 (2)

[a]HKD $1=US $0.13.

[b]Multiple-response questions (participant may select more than 1 choice).

Efficacy of the Intervention

The changes in measures between baseline and the 2-week follow-up among older adults are shown in Table 2. Results from Wilcoxon rank tests (Table S4 in Multimedia Appendix 1) were consistent with the parametric *t* tests.

In terms of digital literacy and daily smartphone use, the mean eHEALS score was 21.41 (SE 0.79) at baseline and 23.80 (SE 1.41) at follow-up. This increase of 2.39 points was not statistically significant (*P*=.11). Daily smartphone usage increased by 0.45 hours, from 2.09 (SE 0.16) to 2.54 (SE 0.22) hours, which was also not significant (*P*=.07).

Among mental health measures, there was a nonsignificant decrease of 1.87 in adjusted WHO-5 scores, from 73.15 (SE 2.17) at baseline to 71.28 (SE 3.72) at follow-up (*P*=.61). The 3-item UCLA Loneliness Scale scores showed a nonsignificant increase of 0.21, from 3.92 (SE 0.13) to 4.13 (SE 0.22; *P*=.39).

For physical well-being and activity measures, self-rated physical well-being had a nonsignificant change from 3.51 (SE 0.09) at baseline to 3.67 (SE 0.19) at follow-up (*P*=.42). Changes in exercise variables were not significant either; strenuous exercise increased from 47.41 (SE 12.66) to 65.88 (SE 17.91; *P*=.36) minutes per week, moderate exercise decreased from 171.12 (SE 26.77) to 153.94 (SE 29.02; *P*=.61) minutes per week, and light exercise (walking) increased from 525.92 (SE 58.92) to 773.00 (SE 153.80; *P*=.13) minutes per week. Sitting, measured in minutes per day, had a nonsignificant increase from 367.32 (SE 26.14) to 398.09 (SE 52.93; *P*=.57).

All complete case results were similar to those obtained using MI, except for self-rated physical well-being. Complete case analyses showed a significant increase in physical well-being ratings, from 3.61 (SE 0.11) to 3.82 (SE 0.09; *P*=.04).

At the 2-week follow-up, 70% (50/71) of older adults reported exercising more frequently, 61% (43/71) reported watching health-related videos more often, and 55% (39/71) washed their hands more regularly (Table S5 in Multimedia Appendix 1).

XSL•FO
RenderX

Table . Primary and secondary outcomes of the intergenerational, home-based eHealth intervention for Hong Kong's older adult population (October-November 2022).

Variables	Participants, n	Baseline, mean (SE)	2-week follow-up, mean (SE)	*P* value	Cohen *d*
Digital health literacy and smartphone use					
eHEALS[a]					
MI[b]	101	21.41 (0.79)	23.80 (1.41)	.11	—[d]
CC[c]	70	21.53 (0.93)	23.40 (1.02)	.10	0.20
Screen time[e]					
MI	101	2.09 (0.16)	2.54 (0.22)	.07	—
CC	64	2.06 (0.18)	2.13 (0.15)	.71	0.05
Additional measures					
WHO-5[f]					
MI	101	73.15 (2.17)	71.28 (3.72)	.61	—
CC	71	73.80 (2.68)	73.80 (2.51)	≥.99	0.00
UCLA-3[g]					
MI	101	3.92 (0.13)	4.13 (0.22)	.39	—
CC	70	3.99 (0.16)	3.90 (0.19)	.69	–0.05
Physical well-being[h]					
MI	101	3.51 (0.09)	3.67 (0.19)	.42	—
CC	71	3.61 (0.11)	3.82 (0.09)	.04	0.25
Strenuous exercise[i]					
MI	101	47.41 (12.66)	65.88 (17.91)	.36	—
CC	68	45.81 (12.76)	38.81 (11.72)	.61	–0.06
Moderate exercise[i]					
MI	101	171.12 (26.77)	153.94 (29.02)	.61	—
CC	70	168.86 (31.41)	123.70 (22.07)	.08	–0.21
Walking[i]					
MI	101	525.92 (58.92)	773.00 (153.80)	.13	—
CC	68	496.01 (64.36)	548.38 (88.38)	.63	0.06
Sitting[j]					
MI	101	367.32 (26.14)	398.09 (52.93)	.57	—
CC	68	381.18 (32.23)	358.24 (30.97)	.46	–0.09

[a]eHEALS: eHealth Literacy Scale.
[b]MI: multiple imputation.
[c]CC: complete case.
[d]Not applicable.
[e]Time (categorical hours) spent using smartphone per day.
[f]WHO-5: World Health Organization–Five Well-Being Index.
[g]UCLA-3: 3-item University of California, Los Angeles Loneliness Scale.
[h]Self-reported single score, ranging from 1 to 5 (1= very poor, 5=very good).
[i]Minutes per week.
[j]Minutes per day.

XSL•FO
RenderX

Feasibility and Acceptability of the Intervention

At baseline, 57 university student interventionists delivered material (n=101, 100%) on eHealth literacy, NCD prevention, psychosocial health, COVID-19 prevention, and educational leaflets. They demonstrated QR code scanning, encouraged independent practice, assisted with installing health apps and registering for services, and created one-on-one WhatsApp group chats with older adults after the intervention (n=101, 100%).

At the 2-week follow-up, participants self-reported their mean satisfaction with the intervention and the 14-day WhatsApp messaging period as 4.32 (SD 0.53) and 3.87 (SD 0.72), respectively, out of 5.

Other Differences

No significant differences were found between participants who completed the follow-up (n=71) and those who did not (n=30) in age, gender, income, or living arrangement. A significant difference was observed in education level, but it should be noted that this observation was made among a lower number of participants (n<5) at times (Table S7 in Multimedia Appendix 1).

Discussion

Principal Findings

This pilot study indicates that an intergenerational, home-based eHealth intervention package is a feasible and acceptable method of improving older adults' engagement with digital health resources while promoting healthy behaviors. These findings address our objective to examine the feasibility and effectiveness of such an intervention in Hong Kong's older adult population.

The intervention showed promising, though modest, improvements in eHealth literacy and smartphone usage among older adults. The mean eHEALS scores increased by 2.39 (21.41, SE 0.79 to 23.80, SE 1.41) points, and daily smartphone use rose by 0.45 hours (2.09 to 2.54 hours), though these changes were not statistically significant. Despite the lack of statistical significance, the positive trends suggest potential benefits that warrant further investigation with larger sample sizes and longer follow-up periods. The high satisfaction rating (4.32 out of 5) indicates strong acceptability among participants. Furthermore, most participants reported improved health behaviors, including increased physical activity (50/71, 70%), more frequent viewing of health videos (43/71, 61%), and improved handwashing practices (39/71, 55%), indicating the intervention positively impacted their daily routines and health.

Comparisons with existing literature reveal interesting patterns. Our findings align with previous studies [21,22], but the comparisons suggest that longer durations may be required to achieve more substantial improvements in digital literacy. For example, Xie's study [21] involved 2 weekly classes over a total of 4 weeks and showed significant improvements in participants' computer knowledge (P<.01). Additionally, our sample was noticeably older (mean age: 80.42, SD 5.72 years) with lower baseline eHEALS scores (mean 21.41, SE 0.79) compared to that of Chang et al [22], whose participants (mean age: 74.64 years) started with higher baseline eHEALS scores (mean 28.31) and achieved statistically significant improvements 2 months postintervention (P<.001). These comparisons suggest that the older age and limited digital experience of our participants may have constrained the impact of the intervention within the 2-week timeframe. Previous literature [45,46] suggests that longer follow-up time periods may be required to observe noticeable changes as older adults require time to learn and put knowledge to use.

The home-based design of the intervention was effective in reaching underprivileged older adults, a population often overlooked by traditional classroom-based programs. Only 4.0% (4/101) of participants had tertiary education, and just 3.0% (3/101) had a household income above the local median (HKD $25,000 per month). Many participants had limited prior exposure to eHealth technologies; fewer than 30% had ever searched for health information online, and most had not used mobile health apps, except for the government-mandated COVID-tracking app, LeaveHomeSafe. Conducting the intervention in participants' homes allowed them to learn comfortably at their own pace, free from the fear of judgment by peers, an essential factor given their low baseline eHEALS scores (mean 21.41, SE 0.79). The home-based setting also helped minimize stress and boost confidence, enhanced by positive language and a personalized pace set by student interventionists. This approach removed several key barriers commonly identified in older adults' learning of digital technologies [20].

We also recruited 57 nursing students (Table S6 in Multimedia Appendix 1) with varying levels of study and background knowledge, all of whom received specialized training covering primary health care, aging populations in Hong Kong, health assessments, intervention delivery, and communication skills with older adults. This experience not only provided student interventionists with an extracurricular opportunity to develop practical skills but also enhanced their ability to communicate with older adults, an important component of geriatric care. Programs like the Generations Connect Project demonstrate the potential of intergenerational interventions to support older adults while enriching health care students' education and professional development [14].

This study had 3 limitations. First, the short follow-up period may have limited our ability to observe significant changes in health-related indicators and eHealth literacy. Future studies with longer follow-up periods are needed to better assess the sustained effects of interventions. Second, the follow-up surveys were conducted by telephone due to privacy concerns, differing from the home-based baseline assessments, which may have affected consistency. Third, the low educational attainment of participants (60/101, 60.0% having received only primary school education or below) may have hindered their engagement with the intervention content, reducing its effectiveness. These limitations highlight the need for materials tailored to the educational level of older adults and for more standardized intervention protocols in future research.

XSL•FO
RenderX

Conclusions

This pilot study showed that an intergenerational, home-based eHealth intervention may be an effective approach to improving older adults' engagement with digital health care resources while simultaneously supporting their physical and psychological well-being. The findings suggest that such interventions can help bridge the digital divide for older adult populations and promote healthier behaviors.

Future research should consider extending the intervention duration, increasing the frequency of home visits, and developing more tailored educational materials. Additionally, collaborations with local health care agencies could help address financial barriers by providing affordable mobile devices and internet access. The broader implications of this research extend to health care policy and practice, highlighting the potential of community-based, intergenerational approaches to address the growing health care needs of aging populations while preparing health care students for geriatric care through practical, community-oriented training experiences.

Acknowledgments

Special thanks to the Sino Group and Ng Teng Fong Foundation for their generous funding to support the Generations Connect Project.

Data Availability

The deidentified dataset used in the pilot study of the Generations Connect Project is available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary tables on additional participant characteristics and demographic information.
[DOCX File, 40 KB - formative_v9i1e69611_app1.docx]

Checklist 1
CONSORT 2010 checklist.
[DOC File, 234 KB - formative_v9i1e69611_app2.doc]

References

1. Hong Kong 2021 Population Census - thematic report: older persons. : Census and Statistics Department; 2023 URL: https://www.censtatd.gov.hk/en/EIndexbySubject.html?scode=600&pcode=B1120118 [accessed 2025-05-07]
2. Hong Kong population projections 2022-2046. : Census and Statistics Department; 2023 URL: https://www.censtatd.gov.hk/en/data/stat_report/product/B1120015/att/B1120015092023XXXXB01.pdf [accessed 2025-05-07]
3. Choi M, Sempungu JK, Lee EH, Lee YH. Living longer but in poor health: healthcare system responses to ageing populations in industrialised countries based on the findings from the Global Burden of Disease Study 2019. BMC Public Health 2024 Feb 22;24(1):576. [doi: 10.1186/s12889-024-18049-0] [Medline: 38388412]
4. Global expenditure on health: public spending on the rise. : World Health Organization; 2021 URL: https://www.who.int/publications/i/item/9789240041219 [accessed 2025-05-07]
5. Hong Kong country commercial guide. International Trade Adminstration. 2024. URL: https://www.trade.gov/country-commercial-guides/hong-kong-healthcare [accessed 2025-05-02]
6. Yip WCM, Hsiao WC, Chen W, Hu S, Ma J, Maynard A. Early appraisal of China's huge and complex health-care reforms. Lancet 2012 Mar 3;379(9818):833-842. [doi: 10.1016/S0140-6736(11)61880-1] [Medline: 22386036]
7. Du X, Patel A, Anderson CS, Dong J, Ma C. Epidemiology of cardiovascular disease in China and opportunities for improvement: JACC International. J Am Coll Cardiol 2019 Jun 25;73(24):3135-3147. [doi: 10.1016/j.jacc.2019.04.036] [Medline: 31221263]
8. Coping with chronic illness. Elderly Health Service Department of Health. 2023. URL: https://www.elderly.gov.hk/english/healthy_ageing/mental_health/coping_with_chronic_illness.html
9. Kim J, Kim Y, Ha J. Changes in daily life during the COVID-19 pandemic among South Korean older adults with chronic diseases: a qualitative study. Int J Environ Res Public Health 2021 Jun 24;18(13):6781. [doi: 10.3390/ijerph18136781] [Medline: 34202534]
10. Petrova NN, Khvostikova DA. Prevalence, structure, and risk factors for mental disorders in older adults. Adv Gerontol 2021;34(1):152-159. [Medline: 33993676]

XSL•FO
RenderX

11. Sinha P, Hussain T, Boora NK, et al. Prevalence of common mental disorders in older adults: results from the National Mental Health Survey of India. Asian J Psychiatr 2021 Jan;55:102463. [doi: 10.1016/j.ajp.2020.102463] [Medline: 33212298]
12. Zenebe Y, Akele B, W/Selassie M, Necho M. Prevalence and determinants of depression among old age: a systematic review and meta-analysis. Ann Gen Psychiatry 2021 Dec 18;20(1):55. [doi: 10.1186/s12991-021-00375-x] [Medline: 34922595]
13. Vernooij-Dassen M, Verhey F, Lapid M. The risks of social distancing for older adults: a call to balance. Int Psychogeriatr 2020 Oct;32(10):1235-1237. [doi: 10.1017/S1041610220001350] [Medline: 32576306]
14. Choi EY, Farina MP, Wu Q, Ailshire J. COVID-19 social distancing measures and loneliness among older adults. J Gerontol B Psychol Sci Soc Sci 2022 Jul 5;77(7):e167-e178. [doi: 10.1093/geronb/gbab009] [Medline: 33428753]
15. Ernst M, Niederer D, Werner AM, et al. Loneliness before and during the COVID-19 pandemic: a systematic review with meta-analysis. Am Psychol 2022;77(5):660-677. [doi: 10.1037/amp0001005] [Medline: 35533109]
16. Norman CD, Skinner HA. eHealth literacy: essential skills for consumer health in a networked world. J Med Internet Res 2006 Jun 16;8(2):e9. [doi: 10.2196/jmir.8.2.e9] [Medline: 16867972]
17. Berkowsky RW. Exploring predictors of eHealth literacy among older adults: findings from the 2020 CALSPEAKS survey. Gerontol Geriatr Med 2021;7:23337214211064227. [doi: 10.1177/23337214211064227] [Medline: 34926723]
18. Frishammar J, Essén A, Bergström F, Ekman T. Digital health platforms for the elderly? Key adoption and usage barriers and ways to address them. Technol Forecast Soc Change 2023 Apr;189(122319):122319. [doi: 10.1016/j.techfore.2023.122319]
19. Hunsaker A, Hargittai E. A review of internet use among older adults. New Media & Society 2018 Oct;20(10):3937-3954. [doi: 10.1177/1461444818787348]
20. Pourrazavi S, Kouzekanani K, Bazargan-Hejazi S, et al. Theory-based e-health literacy interventions in older adults: a systematic review. Arch Public Health 2020;78:72. [doi: 10.1186/s13690-020-00455-6] [Medline: 32793345]
21. Xie B. Improving older adults' e-health literacy through computer training using NIH online resources. Libr Inf Sci Res 2012 Jan 1;34(1):63-71. [doi: 10.1016/j.lisr.2011.07.006] [Medline: 22639488]
22. Chang SJ, Yang E, Lee KE, Ryu H. Internet health information education for older adults: a pilot study. Geriatr Nurs 2021;42(2):533-539. [doi: 10.1016/j.gerinurse.2020.10.002] [Medline: 33092906]
23. Chiu CJ, Hu YH, Lin DC, Chang FY, Chang CS, Lai CF. The attitudes, impact, and learning needs of older adults using apps on touchscreen mobile devices: results from a pilot study. Comput Human Behav 2016 Oct;63:189-197. [doi: 10.1016/j.chb.2016.05.020]
24. Opdenacker J, Boen F, Coorevits N, Delecluse C. Effectiveness of a lifestyle intervention and a structured exercise intervention in older adults. Prev Med 2008 Jun;46(6):518-524. [doi: 10.1016/j.ypmed.2008.02.017] [Medline: 18405960]
25. Turner J, Greenawalt K, Goodwin S, Rathie E, Orsega-Smith E. The development and implementation of the Art of Happiness intervention for community-dwelling older adults. Educ Gerontol 2017;43(12):1-11. [doi: 10.1080/03601277.2017.1380894]
26. Arthanat S, Vroman KG, Lysack C, Grizzetti J. Multi-stakeholder perspectives on information communication technology training for older adults: implications for teaching and learning. Disabil Rehabil Assist Technol 2019 Jul;14(5):453-461. [doi: 10.1080/17483107.2018.1493752] [Medline: 30044659]
27. Geerts N, Schirmer W, Vercruyssen A, Glorieux I. Bridging the 'instruction gap': how ICT instructors help older adults with the acquisition of digital skills. Int J Lifelong Educ 2023 Mar 4;42(2):195-207. [doi: 10.1080/02601370.2023.2174197]
28. LoBuono DL, Leedahl SN, Maiocco E. Teaching technology to older adults: modalities used by student mentors and reasons for continued program participation. J Gerontol Nurs 2020 Jan 1;46(1):14-20. [doi: 10.3928/00989134-20191118-02] [Medline: 31895957]
29. Pihlainen K, Ehlers A, Rohner R, et al. Older adults' reasons to participate in digital skills learning: an interdisciplinary, multiple case study from Austria, Finland, and Germany. Studies in the Education of Adults 2023 Jan 2;55(1):101-119. [doi: 10.1080/02660830.2022.2133268]
30. Chang SJ, Choi S, Kim SA, Song M. Intervention strategies based on Information-Motivation-Behavioral Skills model for health behavior change: a systematic review. Asian Nurs Res (Korean Soc Nurs Sci) 2014 Sep;8(3):172-181. [doi: 10.1016/j.anr.2014.08.002] [Medline: 25030493]
31. Fisher WA, Fisher JD, Harman J. The Information-Motivation-Behavioral Skills model: a general social psychological approach to understanding and promoting health behavior. In: Social Psychological Foundations of Health and Illness: Blackwell Publishing; 2003:82-106. [doi: 10.1002/9780470753552]
32. Phang JK, Kwan YH, Yoon S, et al. Digital intergenerational program to reduce loneliness and social isolation among older adults: realist review. JMIR Aging 2023 Jan 4;6:e39848. [doi: 10.2196/39848] [Medline: 36598801]
33. Bae S, Kim D. Improving home-like environments in long-term care units: an exploratory mixed-method study. Sci Rep 2024 Jun 9;14(1):13243. [doi: 10.1038/s41598-024-62328-0] [Medline: 38853152]
34. Generations Connect. HKU Bulletin. 2024 Nov 25. URL: https://bulletin.hku.hk/knowledge-exchange/generations-connect [accessed 2025-05-02]
35. Eldridge SM, Chan CL, Campbell MJ, et al. CONSORT 2010 statement: extension to randomised pilot and feasibility trials. BMJ 2016 Oct 24;355:i5239. [doi: 10.1136/bmj.i5239] [Medline: 27777223]

XSL•FO
RenderX

36. Kim S, Chow BC, Park S, Liu H. The usage of digital health technology among older adults in Hong Kong and the role of technology readiness and ehealth literacy: path analysis. J Med Internet Res 2023 Apr 12;25:e41915. [doi: 10.2196/41915] [Medline: 37043274]
37. Xie L, Mo PKH. Comparison of eHealth Literacy Scale (eHEALS) and Digital Health Literacy Instrument (DHLI) In assessing electronic health literacy in Chinese older adults: a mixed-methods approach. Int J Environ Res Public Health 2023 Feb 13;20(4):3293. [doi: 10.3390/ijerph20043293] [Medline: 36833987]
38. Jiang X, Wang L, Leng Y, et al. The level of electronic health literacy among older adults: a systematic review and meta-analysis. Arch Public Health 2024 Nov 7;82(1):204. [doi: 10.1186/s13690-024-01428-9] [Medline: 39511667]
39. Topp CW, Østergaard SD, Søndergaard S, Bech P. The WHO-5 Well-Being Index: a systematic review of the literature. Psychother Psychosom 2015;84(3):167-176. [doi: 10.1159/000376585] [Medline: 25831962]
40. Russell DW. UCLA Loneliness Scale (version 3): reliability, validity, and factor structure. J Pers Assess 1996 Feb;66(1):20-40. [doi: 10.1207/s15327752jpa6601_2] [Medline: 8576833]
41. Craig CL, Marshall AL, Sjöström M, et al. International physical activity questionnaire: 12-country reliability and validity. Med Sci Sports Exerc 2003 Aug;35(8):1381-1395. [doi: 10.1249/01.MSS.0000078924.61453.FB] [Medline: 12900694]
42. Kang H. The prevention and handling of the missing data. Korean J Anesthesiol 2013 May;64(5):402-406. [doi: 10.4097/kjae.2013.64.5.402] [Medline: 23741561]
43. Little RJA. A test of missing completely at random for multivariate data with missing values. J Am Stat Assoc 1988 Dec;83(404):1198-1202. [doi: 10.1080/01621459.1988.10478722]
44. Woods AD, Gerasimova D, Van Dusen B, et al. Best practices for addressing missing data through multiple imputation. Infant Child Dev 2024 Jan;33(1):e2407. [doi: 10.1002/icd.2407]
45. Göransson C, Wengström Y, Hälleberg-Nyman M, Langius-Eklöf A, Ziegert K, Blomberg K. An app for supporting older people receiving home care - usage, aspects of health and health literacy: a quasi-experimental study. BMC Med Inform Decis Mak 2020 Sep 15;20(1):226. [doi: 10.1186/s12911-020-01246-3] [Medline: 32933500]
46. Nahm ES, Zhu S, Bellantoni M, et al. The effects of a theory-based patient portal e-learning program for older adults with chronic illnesses. Telemed J E Health 2019 Oct;25(10):940-951. [doi: 10.1089/tmj.2018.0184] [Medline: 30431393]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
eHEALS: eHealth Literacy Scale
GDP: gross domestic product
MCAR: missing completely at random
MI: multiple imputation
NCD: noncommunicable disease
UCLA: University of California, Los Angeles

Edited by A Mavragani; submitted 04.12.24; peer-reviewed by J Chen, R Kwan; revised version received 28.03.25; accepted 07.04.25; published 08.05.25.

Please cite as:
He AWJ, Yuan R, Luk TT, Wang KMP, Chan SSC
Boosting Digital Health Engagement Among Older Adults in Hong Kong: Pilot Pre-Post Study of the Generations Connect Project
JMIR Form Res 2025;9:e69611
URL: https://formative.jmir.org/2025/1/e69611
doi:10.2196/69611

© Aaron Wan Jia He, Runqi Yuan, Tzu Tsun Luk, Kelvin Man Ping Wang, Sophia Siu Chee Chan. Originally published in JMIR Formative Research (https://formative.jmir.org), 8.5.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Using a Robot to Address the Well-Being, Social Isolation, and Loneliness of Care Home Residents via Video Calls: Qualitative Feasibility Study

Lise Birgitte Holteng Austbø[1,2], MHS; Ingelin Testad[1,3], MNS, PhD, Prof Dr; Martha Therese Gjestsen[1,2], BSN, MHS, PhD

[1]Centre for Age-Related Medicine, Stavanger University Hospital, Stavanger, Norway
[2]Department of Clinical Medicine, University of Bergen, Postboks 7804, Bergen, Norway
[3]University of Exeter Medical School, Exeter, United Kingdom

Corresponding Author:
Lise Birgitte Holteng Austbø, MHS
Centre for Age-Related Medicine, Stavanger University Hospital, Stavanger, Norway

Abstract

Background: About 40,000 people are living in Norwegian care homes, where a majority are living with a dementia diagnosis. Social isolation and loneliness are common issues affecting care home residents' quality of life. Due to visitation restrictions during the pandemic, residents and family members started using digital solutions to keep in contact. There is no framework or guidelines to inform the uptake and use of technologies in the care home context, and this often results in non-adoption and a lack of use after the introduction phase. Hence, there is a great need for research on the feasibility of a robot that can facilitate video communication between residents and family members.

Objective: This study aimed to (1) introduce video communication through a robot to address social isolation and loneliness in a care home during a period of 6 weeks and (2) identify elements central to the feasibility concerning testing and evaluating the use of the robot.

Methods: Three focus group interviews were undertaken: 1 with family members (n=4) and 2 with care staff (n=2 each). The informants were purposely selected to ensure that they had the proper amount of experience with the robot to have the ability to inform this study's objectives. The focus group interviews were tape-recorded and transcribed verbatim, then subsequently analyzed using systematic text condensation.

Results: The data analysis of focus group interviews and individual interviews resulted in three categories: (1) organizing the facilitation of video calls, (2) using a robot in dementia care, and (3) user experience with the robot.

Conclusions: Video communication in care homes is a feasible alternative to face-to-face interactions, but it depends on organizational factors such as information flow, resources, and scheduling. In dementia care, the user-friendly robot supports person-centered care through tailored social interaction. Both family members and staff express enthusiasm for video calls as an option and see its potential for future use.

(JMIR Form Res 2025;9:e59764) doi:10.2196/59764

KEYWORDS

care home; geriatric; aging; dementia; user experience; social isolation; loneliness; feasibility; robot; video calls; mobile phone

Introduction

Background

About 40,000 people are living in Norwegian care homes [1], and about 84% of them are living with dementia [2]. Most residents also have medical comorbidity along with complex care needs due to a combination of cognitive, functional, and communication impairments and neuropsychiatric symptoms [3]. Social isolation and loneliness are common problems [4-6], and these aspects were exacerbated because of the infection control measures (eg, visitor restrictions) undertaken in response to the COVID-19 pandemic [7,8]. As a consequence of the social distancing, families started using digital solutions to keep in touch with their loved ones residing in care homes [9-11]. In the present literature, there is a large number of small-scale studies that explore the use of technologies for care home residents, their families, and staff. For example, digital communication solutions (eg, tablets with video and sound; telepresence robots) for interactions between patients and family [12,13]. There is a general notion that the "digital revolution" in health care also includes long-term care, but there is no framework or are no guidelines to inform the uptake and use of technologies in the care home context [14-16]. Hence, there is

XSL•FO
RenderX

a need for systematic research on the feasibility of such solutions in terms of factors affecting the uptake and use, and also to inform future research in this field [17].

Informing a Digital Intervention in Care Homes

Despite the rhetoric associated with the benefits of digital solutions in health care, the uptake and use have not progressed at the pace and scale anticipated [18]. Limited fidelity of technology recommendation to a person's needs may be one reason, but it is increasingly recognized in the research literature that the health care personnel's acceptance of the technological application itself remains a key challenge in adopting an intervention [19,20]. This underlines the vital importance that the involved stakeholders (eg, researchers, policy makers, health care personnel, patients, and carers) can judge the value of technologies in their own right. Conversely, until we develop solutions that are considered to be useful and fit for purpose by the actual users, we will repeat what has been observed, analyzed, and conceptualized by Greenhalgh and Abimbola [21], in their Nonadoption, Abandonment, Difficulties in Scaling Up, Spread, and Sustainability framework: problems in technology projects usually occur because they are too complex and because the complexity is suboptimally handled.

Independently from the COVID-19 pandemic, but capitalizing on the experience accumulated, the application of digital solutions in a care home context has the potential to improve the quality of care as well as the quality of life for each resident. A recent review by Knapp et al [22], provides an overview of digital solutions that appear most ready for use in the next 5 years: digital care for tailored strategies for carers; mobile technologies for supporting self-care and daily activities; touchscreen and multimedia interventions and activities to improve mood, engagement, and behaviors; and ICT-based technologies for social connection.

During the periods of COVID-19 lockdown in care homes in Norway, staff described the need for a solution for residents and their family and friends to keep in contact while visits to the care homes were restricted. Through a collaboration with a cluster of technology vendors (Norwegian Smart Care Cluster), we identified a secure digital solution that could be tested for this purpose.

The digital solution entails using a robot to address social isolation and loneliness. The robot is developed in collaboration with care home staff and residents, and family members, and consists of an iPad mounted on a wheeled stand or a tripod suitable for tabletop placement. The technology enables video communication similar to platforms such as FaceTime (Apple Inc) or Skype (Microsoft Corp), allowing for 2-way communication between residents and family members who have the family member application compatible with the robot downloaded on their smartphone. Family members can maneuver the robot from their phone; however, this feature was not used in this study. The robot is not equipped to interact with residents through voice prompts. The technology is integrated and developed on a secure health platform developed and provided by PatientSky (EG Norge) to ensure patient privacy.

The study applies the New Medical Research Council guidance for developing and evaluating complex interventions [23] in the process of developing an intervention entailing the use of a robot in care homes. The New Medical Research Council's framework is recommended for the development of interventions containing several interacting components, which aligns well with the multitude of stakeholders and number of components involved when a digital solution is tested in primary care. This study reported in this paper pertains to the first steps in the framework, which are development and piloting. Development phases and feasibility or pilot studies are recommended before conducting larger evaluation studies of complex interventions, to explore the procedures and applicability as well as participants' experiences with the intervention [24,25]. This can facilitate improvements and inform the design of potential future confirmatory studies.

Study Aims

This study aimed to (1) introduce video communication through a robot to address social isolation and loneliness in a care home during 6 weeks and (2) identify elements central to the feasibility in terms of testing and evaluating the use, that is, how recruitment and inclusion can be optimized, which outcome measures related to the intervention are feasible, and how care staff and family members experienced using the robot as a video communication tool.

Experiences and findings related to the first 2 elements (ie, how recruitment and inclusion can be optimized, which outcome measures related to the intervention are feasible) are applied when we are currently planning a randomized controlled trial using the same intervention. This paper reports on the care staff and family member perspectives and experiences with using the robot in video communication. The findings can help inform other researchers who potentially face similar challenges when developing digital interventions in a similar setting.

Methods

Context

This study is a result of the collaboration between researchers or authors, and members in a care home research network called Fokus, geographically located in Western Norway. The Fokus network supports research activity within care homes, with the overall aim of improving the lives of people in care homes, with particular focus on individuals with dementia. Designed to be mutually beneficial, the Fokus network aims to enhance the quantity and quality of practice-adjacent research within care home settings, increase the accessibility of research to primary care, but also vice versa: primary care stakeholders are vital to researchers per developing knowledge that is relevant in their day-to-day practice. Fokus currently consists of 20 care homes in the southern part of the Western Norway Regional Health Authority. As part of this collaborative network, this feasibility study was conducted in 2 units in one of the Fokus care homes; a dementia care unit and a short-term care unit.

Intervention

The robot was introduced to the care staff and patients, and was used for 6 weeks. Care staff and family members were

XSL•FO
RenderX

introduced to the robot and received training before starting the project. Patients and family members communicated through the robot regularly, but the robot did not replace actual visits. It is estimated that the robot was used 1 - 2 times a week per patient, and the conversations lasted for approximately 15 - 30 minutes and were facilitated by care staff to ensure no technological difficulties. The entire staff at the 2 units was encouraged to interact with the robot as much as possible to gain experience.

Study Design

This is a 6-week feasibility study to test and describe the use of a robot for video calls in a care home setting. This study will help inform how recruitment and inclusion can be optimized and assess the feasibility and practicality of measuring the outcomes of the intervention in this context. We used trusted, validated questionnaires from the residents for this purpose, however this article will focus on the care staff and family member perspective and experience with using the robot in video communication. Data was collected using focus group interviews [26].

Recruitment

As this is a feasibility study, no formal sample size calculation was required. We purposely recruited 2 care home units, 1 dementia care unit, and 1 short term care unit in a care home from the Fokus network. Care staff were recruited from the network members, who further identified residents and family members from their respective units and determined their eligibility to use the robot. Recruitment was conducted in May and June of 2020. Inclusion and exclusion criteria was designed to be as inclusive as possible, and only to exclude individuals who would not be able to use the intervention: (1) all individuals residing in participating care homes who have severe dementia corresponding to the score of “3” or greater on the Clinical Dementia Rating scale, (2) not able to undertake activities in daily living, and (3) any resident from whom consent or the advice of a consultee could not be obtained. A total of 5 residents used the robot for video communication with their respective family members. Four staff members were designated users and facilitators of the communication.

Data Collection

Three focus group interviews were undertaken: 1 with family members (n=4) and 2 with care staff (n=2 each), respectively. The informants were purposely selected by the head of the care home units, to ensure that they had the proper amount of experience with the robot to have the ability to inform this study’s objectives. Although 5 residents used the robot, 4 family members participated in the interviews, as 1 family member had a conflicting schedule at the time of the interview. The interviews were conducted by a moderator (LBHA) and comoderator (MTG). The administrative coordinator of the Fokus network participated in one of the staff interviews. All interviews were based on a semistructured interview guide focusing on the staff’s and family members’ direct experience with the robot, as well as their take on promoting and hindering factors to the implementation and use of such a measure in a care home setting. The interview with family members and the first staff interview were conducted through Zoom and the second interview with staff was conducted in the care home; all 3 were audio recorded. All interviews lasted for about 60 - 70 minutes.

Analysis

The focus group interviews were transcribed verbatim immediately after (by LBHA) and analyzed by all authors according to systematic text condensation as described by Malterud [26]. This method includes 4 steps to convey the participants’ experience with a phenomenon. First, all authors read the transcripts to form an overall impression and identify main themes. Second, meaning units were derived from the transcripts and coded into subgroups based on the themes from the first step. Third, the meaning units were written into artificial quotes called condensates. As the last step of the analysis, the condensates were written into analytical text and represented as results. The analytical process is demonstrated in Figure 1.

Figure 1. Analytical process.

Subtheme	Meaning unit	Artificial quote condensate	Analytical text
Resources impacted the use of the robot.	*“It takes time from other tasks, and in the unit where we worked, things can be calm for a second, and suddenly there are 5 people who need help.”*	I noticed that it took time away from other tasks I had in the care home unit. This was particularly true in the dementia unit where situations changed quickly, and many people need help simultaneously	According to staff, using the robot was time-consuming and at times very challenging, especially in the dementia care unit where the situation can change rapidly from calm to chaotic. Many residents need assistance simultaneously, and these situations appear quickly and unpredictably.

XSL•FO
RenderX

Data Saturation

Data saturation was discussed during analysis and was considered obtained when no more information could be attained and further coding of the material was not feasible.

Ethical Considerations

This study was reviewed by and obtained ethical approval from the Regional Ethics Committee South East D (#15405). Informants provided a written consent and were informed that they could withdraw from this study at any point and without reason. Qualitative data from the interviews were transcribed verbatim and anonymized by replacing informants' names with a number. All data were collected and stored per data protection regulations; they were stored electronically on computers, which were access-controlled via passwords. Hard copies of transcripts were securely stored in locked filing cabinets in offices that were accessible only to research staff. Data were deleted at the end of this study. The participants were not compensated for their participation.

Results

Cohort Characteristics

All care staff and family members who had experience using the robot were invited to focus group interviews, and 8 participants consented to participate. Care staff and family members were interviewed separately; 2 staff members from a dementia care unit participated in the first interview, 2 staff members from a short-term care unit participated in the second interview, and 4 family members participated in the third interview. The interviews were conducted in the fall of 2020. Care staff included nurses, and nursing and medical doctor students. The family members were children and spouses of the residents who had been using the robot. There were no exclusions or losses of participants after consent to participate in the interviews.

The data analysis of focus group interviews resulted in three categories: (1) organizing the facilitation of video calls, (2) using a robot in dementia care, and (3) user experience with the robot. Content from step (4) in the analysis (recontextualization) is presented as analytical text with category headings, respectively, and assembled with quotes that are representative of the category.

Organizing the Facilitation of Video Calls

Overview

Findings from the focus group interviews showed that organizational factors affected the use of the robot for video calls. Adequate information was paramount to involve staff, resource barriers impacted the use of the robot, and structuring the video calls to accommodate routines was essential to the facilitation. Subthemes within this theme are related to information, resources, and scheduling.

Care Staff Information Needs and Superusers

Care staff highlighted the need for extensive information and orientation before implementing the robot in their daily routine and care. Findings demonstrate that a well-planned introduction focusing on organizational leadership and motivated staff is crucial to successful use.

> *It is very important that everyone who will be involved and use the robot is involved from the very beginning and receives information.* [Staff from the dementia care unit, focus group 1]

At first, all staff were motivated and engaged with the project. However, after the introduction phase, 2 superusers in each unit were responsible for using the robot and facilitating video communication. This resulted in a lack of enthusiasm from the rest of the staff. Care home leaders must drive the implementation and use of digital solutions to ensure optimal uptake.

Resources Impacting the Use of the Robot

According to staff, using the robot was time-consuming and at times very challenging, especially in the dementia care unit, where the situation can change rapidly from calm to chaotic. Many residents need assistance simultaneously, and these situations appear quickly and unpredictably.

> *If you do not have the time or resources […], it highly depends on your work shift. During the daytime, I never have the time.* [Staff from the dementia care unit, focus group 1]

Staff explained that assisting residents using the robot took time away from other responsibilities. Yet, they acknowledged that the robot would be a great relief if residents could use it independently. They also highlighted that clarifying the care home unit's capacity concerning time and technology knowledge is key before implementing such a solution.

Scheduled Use of the Robot

Staff concurred during the interviews that facilitating video communication with residents' family members without a scheduled time was time-consuming. Establishing regular communication with family members was emphasized as key to using the robot beneficially.

> *Many people have one-to-one care, and staff members always have to be present, so it is more about making everything work with the daily tasks.* [Staff from the dementia care unit, focus group 1]

Making video call appointments allowed staff to plan and prioritize resources to have enough capacity to support the resident during the calls.

> *Communication with family members and scheduling calls. It was time consuming when the participants could not use the robot by themselves.* [Staff from the short-term care unit, focus group 2]

[Staff from short-term care unit, focus group 2]

This was also a great help in ensuring the robot was used regularly. Both family members and staff expressed that they, as care partners, need to be involved in planning to ensure that using a robot is not seen as a burden.

XSL•FO
RenderX

Using a Robot in Dementia Care

Overview

Introducing a robot intended to support person-centered care and social interaction for people with dementia requires thorough evaluation and assessment. This will uncover the benefits and consequences of using the robot, and help understand how to fully use it. In this main theme, the subthemes are people living with dementia using the robot and a user-friendly robot in dementia care.

People Living With Dementia Using the Robot

Informants reveal that using the robot for people with a dementia diagnosis pinpoints new opportunities and challenges. Care staff observed that their residents enjoyed seeing their family members through video and that the experience brought instant pleasure and joy.

> *I could see that the resident was smiling, the instant joy was most definitely present, and he did not become agitated afterwards.* [Staff from the dementia care unit, focus group 1]

In addition to enjoying the actual video conversation with their family members, the robot seemed appealing to residents with a previous interest in computers. Despite having advanced dementia and a reduced understanding of the concept of conversation, some residents were still fascinated with the technology.

Although care staff saw benefits for their residents, they stressed the importance of evaluating cognitive and functional status before using the robot for residents with dementia. This was to identify residents who may benefit from using it. One resident became agitated and uneasy after using the robot to talk to their family, thus requiring follow-up from the staff, who expressed the importance of a well-planned routine when ending a video conversation.

User-Friendly Robot in Dementia Care

Care staff found the robot easy to use and very user-friendly. Being part of a project using a digital solution was a good experience, and they said that the robot was a great tool in dementia care. However, implementing technology brought some skepticism, especially regarding technology replacing human contact and in-person visits.

> *You do not want to replace visits with a screen because it is not the same. However, I did not feel that that changed.* [Staff from the short-term care unit, focus group 2]

In the first weeks of this study, the care home used the robot on a stand with wheels. However, they experienced some challenges using this in a dementia unit and switched to using the robot on a stand for tabletops. There were minor issues with the internet while trying to connect to family members, yet this did not affect the usage.

User Experiences With the Robot

Overview

Family members were enthusiastic about using the robot and saw it as a good solution for communication during the pandemic. Staff expressed a positive attitude after using the robot for 6 weeks and highlighted essential aspects for practical implementation and use. The subthemes under this main theme are family member experience and care staff experience with video calls, respectively.

Family Member Experience With Video Calls

There was a general agreement that the family members appreciated the chance to see their loved one. Hearing their voices and seeing their facial expressions gave family members a safe feeling and made it easier for them to get a sense of the residents' state of health.

> *Because phone calls are just a conversation, but when you can see them as well, you feel more of a connection.* [Family member, focus group 3]

> *For my mom, I think this went well, and I thought it was great to see her, and she could see us. Other family members came onto the video, saw her and waved at her. We found much joy in it.* [Family member, focus group 3]

They also pinpointed the robot's usefulness in communicating with distant relatives, allowing the whole family to gather.

Care Staff Experience With Facilitating Video Calls

Introducing the robot to care home staff created both skepticism and enthusiasm. Staff were motivated and excited to try new technology and saw it as a possible solution for relief on a busy day. However, some staff experienced a mixed reception and acceptance among their colleagues. As some residents became agitated and uneasy after using the robot, staff lacked the motivation to use it and saw it as a burden. On the contrary, some did not see it as a burden, but it could be demanding to familiarize themselves with the robot and incorporate it into their daily tasks:

> *Using a solution like this is something that cannot be forced. There has to be a need for it.* [Staff from the short-term care unit, focus group 2]

A well-planned introduction and adequate information during start-up are key. Sufficient information in the beginning and frequent follow-up during the 6 weeks is crucial to staff feeling motivated to use the robot. The robot will only be used to its full potential if they have a sense of ownership of it.

> *Information and follow up of the care staff is very important when using the robot [...] if not, it is easy to think "Well, this does not entail me. I have enough to deal with, with my own tasks and patients.* [Staff from the dementia care unit, focus group 1]

During the pandemic, the informants pointed out that they saw a clear need for a solution like the robot in this study. Care home staff had to assume a role as a link between residents and family members. They had to care for residents unable to receive visits from their families and simultaneously take care of family

XSL•FO
RenderX

members who did not have the opportunity to follow up with their loved ones.

> *We had to take on the role of a family member and caregiver simultaneously, as well make sure to update the residents' family [...] It became very busy for us [...] it would be great to have this opportunity, and great if everyone had one (a robot) in their room.* [Staff from short-term care unit, focus group 1]

There was a general agreement that technology acceptance might be influenced by generation. Some are worried about using digital solutions to deliver care and the distance between caregiver and resident that might arise:

> *They are anxious to lose the well-established human contact in care.* [Staff from the short-term care unit, focus group 1]

They suggested that future care home staff may have different perspectives on the integration of technology into their work.

Discussion

Principal Findings

In this study, using a robot for video communication between care home residents and family members was feasible and provided an alternative to face-to-face communication during visitation restrictions. However, the feasibility was substantially influenced by organizational factors such as information flow, available resources and scheduling video calls with family members. Using a robot in dementia care can support person-centered care by facilitating social interactions, as long as the use is adapted to individual needs. The robot was user-friendly, and both family members and staff expressed enthusiasm to have video calls as an option. This paper provides new insight into using a robot in a care home context, based on the user perspective. This study bridges an existing gap in the literature and has the potential to inform future research aiming to measure the effectiveness of technology on quality of life and loneliness among care home residents.

In the care home context, the introduction of the robot elicited information-related needs from the care staff to maintain encouragement and motivation to use the robot. Novel technology in health care services is often associated with a lack of use after the introduction phase [27,28]. This became evident in this study as facilitation tasks were predominantly obtained by a few individuals in each care home unit after the start-up. Ko et al [14] states that designated superusers appear to be in a better position to support technology implementation, compared to the rest of the staff. However, little is known about the selection of these superusers to optimize uptake. In our study, the designated staff were preselected and received training in facilitating video calls through the robot to involve the care home unit. They, however, seemed to carry out the task mostly by themselves. An issue that might be solved with a continuing flow of information to keep the remaining staff motivated.

Even though a positive attitude toward the robot was evident, available resources and capacity to fully engage and incorporate the solution into everyday care emerged as an obstacle. Despite the understanding that lack of use is often due to unmet expectations or lack of clear value to the user [18], external promoting and hindering factors can contribute. This was evident in our study as the staff clearly stated that the capacity to implement a robot needs to be evaluated at the organizational meso level to ensure optimal use.

Facilitating video communication in daily life unveiled both opportunities and challenges, especially finding the time to call family members alongside daily routines. Staff started to engage family members and schedule calls to accommodate resource challenges. The agreement between family members and staff about involvement in setting up calls enforced collaboration and was seen as a benefit. Revealing the engagement on both sides, we can argue that the robot brought forward an enthusiasm, regardless of skepticism and attitudes toward new technology. Unlike our findings, previous research reports that scheduling calls was a major source of frustration, as this made the opportunity for communication more rigid [9]. This further emphasized the need for family involvement to ensure that using a robot is not seen as a burden.

The introduction of a robot for care home residents with dementia displayed both benefits and disadvantages for the residents. It necessitates a comprehensive evaluation that takes into account the unique needs of each resident to complement person-centered care. Aligned with other research, staff reported that some residents had adverse reactions after the video call and became uneasy and confused [11]. A study [11] found that the appearance of the device caused confusion and anxiety, however, this stopped once the conversation started, and the resident could see their family. Additionally, a sudden disconnect in the technology, where the conversation ended abruptly left the resident anxious and upset. Similar to this, our results showed that the agitation started when the conversation came to an end. It seems that the transitional phase between the start and end of a video call, and the appearance and disappearance of family members on the screen, is grounds for distress. This makes it evident that new technology has to be fit for purpose and its intended users [21,29].

The robot's user-friendliness resulted in positive experiences when facilitating video calls. Yet, some were still skeptical and concerned that the technology would replace human care. These are aspects that are highly important to take forward when planning implementation and use of technology in health care systems, as these are features highlighted as attributes affecting adoption and future use of new technology [29].

Having video calls as an alternative during visitation restrictions revealed a notable level of enthusiasm among family members and a positive attitude among care staff regarding the robot's role in facilitating communication during the pandemic. These findings underscore the importance of investigating both user perspectives to comprehensively assess the feasibility and effectiveness of such technology in health care settings.

While dealing with the reality of not being able to visit residents in care homes, family members expressed gratitude toward the given opportunity to maintain social interactions through the robot. Restrictions and lack of social interaction were a major issue in care homes during the pandemic, and video calls were

XSL•FO
RenderX

care home residents' main source of social enrichment during the pandemic [9]. The ability to hear their voices and observe their facial expressions through the use of technology engendered a profound sense of safety among family members. Kelly et al [9] found that adding a visual aspect to the communication made a substantial difference to their experience. Moreover, this gave family members a better understanding of the resident's overall state of health.

Various emotions and reactions arose from the staff; some were enthusiastic, and some expressed skepticism toward the robot. The enthusiasm was grounded in excitement about integrating technology into their existing routine and saw it as an addition to high-quality care. This is consistent with a finding from a review by Ko et al [14]; care homes are using technology as supplements and not replacements for "human care."

New technology needs to have clear value for staff, but the most important evident value is for the residents [27,28]. As a consequence of some residents becoming agitated after a video call, certain staff members were discouraged from using the robot and perceived it as a potential burden. This observation underscores the importance of individualizing solutions in the field of care to prevent any adverse effects on the user, and for staff to observe the value of the robot in its own right [18].

Care home staff found themselves assuming a dual role of caregivers for residents who could not receive visits from their families, and facilitators of communication between residents and their loved ones. Most residents in Norwegian care homes have a dementia diagnosis, which is already associated with burnout and depression in care home staff, as caregiving for this group is demanding [30]. This added workload intensified the demand for such technological tools. Notably the informants in this study were interviewed while the Norwegian government still had imposed guidelines and restrictions in terms of social distancing for individuals with increased risk. This is important in the consideration of this paper and the generalizability of the results.

Strengths and Limitations

This study contributes novel information about the feasibility of a robot in a care home setting through the exploration of the user perspective. The results will emphasize an unmet knowledge need in the field of technology in dementia care, as well as inform future research and guide the implementation of similar digital solutions. Although the total number of informants was small (n=9), the group of participants varied with a wide range of working experience and experience from being a family member to a person residing in a care home, making the group representative of stakeholders of a care home.

The studies reliability was maintained through an accurate and well-described performance of interviews and the analysis. All researchers sought to approach the data with an open mind to investigate the phenomenon purely from the perspective of the participants. Our methodological approach was suited to the aim of the study, to explore the feasibility of the robot based on the user perspective, ensuring the validity of our results. Interviews with care staff and family members were based on similar interview guides, only to differ in wording to fit their role toward the patients, and were performed by the same researchers to ensure internal validity. Preliminary results were distributed and presented to the informants for discussion and validation.

Implications

This study provides novel information and insight about the use of a robot for video communication in a care home context, as described by care staff and family members. Evidence based on technology implementation in health care mostly concerns non-adoption, abandonment, and poor uptake, making our results paramount in future research and implementation planning. One of the main challenges with existing implementation processes is the "top-down" perspective. Implementation strategies are commonly described by policymakers and researchers, but rarely in collaboration with the users. Stakeholders are key partners in technology implementation, and their acceptance and perceptions are of utmost importance to successful use.

Our results suggest that using video communication in a care home context has the potential to impact loneliness, social isolation by providing a communication alternative when face-to-face visits are not an option. According to previous research, being able to maintain social interaction with family increases quality of life [31]. By studying and describing the feasibility of using a robot for video communication, we can ensure innovative and more efficient pathways for planning and implementing technology in a care home context without disturbing well-established routines, ultimately benefiting patient care and staff satisfaction.

This paper emphasizes the importance of further research, both investigating the effect of using a robot in a care home context, focusing on the reduction of social isolation and loneliness and further describe factors promoting and hindering implementation.

This paper highlights the complexity of introducing a robot in care home context. It underscores the critical roles of leadership, resource assessment, structured scheduling, and person-centered evaluation in ensuring a successful and meaningful integration. As the landscape of care continues to evolve, understanding what factors influence the feasibility of implementation and use of a robot in care homes will be pivotal in enhancing the quality of care provided to individuals living with dementia. Further research and ongoing adaptation will be necessary to address the evolving needs and expectations of care home residents and their families.

Conclusion

This study suggests that using a robot for video communication in care homes is a feasible alternative to face-to-face interactions during visitation restrictions. Feasibility depends on organizational factors such as information flow, resources, and scheduling. In dementia care, the user-friendly robot supports person-centered care through tailored social interaction. Both family members and staff express enthusiasm for video calls as an option.

XSL•FO
RenderX

Acknowledgments

This study was funded by the Centre for Age-Related Medicine, Stavanger University Hospital, and Foundation Dam, through their Extra program related to the coronavirus pandemic. The authors would like to thank the care homes that are members in the Fokus network and all the informants for their invaluable contribution to this study. The authors would also like to thank Hege Eiklid, MSc, Innocom, Marit Hagland, MBA, and Karoline Blikra Mokleiv, MSc, Norwegian Smart Care Cluster for a productive collaboration.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Authors' Contributions

LBHA contributed to this study's design, the development of data collection tools, and data analysis, and drafted this paper. MTG contributed to this study's design, was responsible for the development of data collection tools, and contributed to data analysis and the drafting of this paper. IT planned this study's design and contributed to data analysis and the drafting of this paper. All authors have read and approved the final version of this paper.

Conflicts of Interest

None declared.

References

1. Figures for 2022 on users of care services. Statistics Norway. 2023. URL: https://www.ssb.no/en/pleie/ [accessed 2025-04-24]
2. Helvik AS, Engedal K, Benth J, Selbæk G. Prevalence and severity of dementia in nursing home residents. Dement Geriatr Cogn Disord 2015;40(3-4):166-177. [doi: 10.1159/000433525] [Medline: 26138271]
3. Røen I, Selbæk G, Kirkevold Ø, Engedal K, Testad I, Bergh S. Resourse Use and Disease Couse in Dementia - Nursing Home (REDIC-NH), a longitudinal cohort study; design and patient characteristics at admission to Norwegian nursing homes. BMC Health Serv Res 2017 May 22;17(1):365. [doi: 10.1186/s12913-017-2289-x] [Medline: 28532443]
4. Neves BB, Sanders A, Kokanović R. “It’s the worst bloody feeling in the world”: experiences of loneliness and social isolation among older people living in care homes. J Aging Stud 2019 Jun;49(49):74-84. [doi: 10.1016/j.jaging.2019.100785] [Medline: 31229221]
5. Slettebø Å. Safe, but lonely: living in a nursing home. Nord J Nurs Res 2008 Mar;28(1):22-25. [doi: 10.1177/010740830802800106]
6. Brownie S, Horstmanshof L. The management of loneliness in aged care residents: an important therapeutic target for gerontological nursing. Geriatr Nurs (Lond) 2011 Sep;32(5):318-325. [doi: 10.1016/j.gerinurse.2011.05.003]
7. Ho KHM, Mak AKP, Chung RWM, Leung DYL, Chiang VCL, Cheung DSK. Implications of COVID-19 on the loneliness of older adults in residential care homes. Qual Health Res 2022 Jan;32(2):279-290. [doi: 10.1177/10497323211050910] [Medline: 34855529]
8. Huber A, Seifert A. Retrospective feelings of loneliness during the COVID-19 pandemic among residents of long-term care facilities. Aging Health Res 2022 Mar;2(1):100053. [doi: 10.1016/j.ahr.2022.100053] [Medline: 35018357]
9. Kelly RM, Xing Y, Baker S, Waycott J. Video calls as a replacement for family visits during lockdowns in aged care: interview study with family members. JMIR Aging 2023 Jun 12;6(6):e40953. [doi: 10.2196/40953] [Medline: 37191951]
10. Lemaire C, Humbert C, Sueur C, Racin C. Use of digital technologies to maintain older adults’ social ties during visitation restrictions in long-term care facilities: scoping review. JMIR Aging 2023 Feb 10;6(6):e38593. [doi: 10.2196/38593] [Medline: 36599164]
11. Zamir S, Hennessy CH, Taylor AH, Jones RB. Video-calls to reduce loneliness and social isolation within care environments for older people: an implementation study using collaborative action research. BMC Geriatr 2018 Mar 2;18(1):62. [doi: 10.1186/s12877-018-0746-y] [Medline: 29499659]
12. Tsai HH, Tsai YF. Changes in depressive symptoms, social support, and loneliness over 1 year after a minimum 3-month videoconference program for older nursing home residents. J Med Internet Res 2011 Nov 15;13(4):e93. [doi: 10.2196/jmir.1678] [Medline: 22086660]
13. Bemelmans R, Gelderblom GJ, Jonker P, de Witte L. Socially assistive robots in elderly care: a systematic review into effects and effectiveness. J Am Med Dir Assoc 2012 Feb;13(2):114-120. [doi: 10.1016/j.jamda.2010.10.002] [Medline: 21450215]
14. Ko M, Wagner L, Spetz J. Nursing home implementation of health information technology: review of the literature finds inadequate investment in preparation, infrastructure, and training. INQUIRY 2018;55(55):46958018778902. [doi: 10.1177/0046958018778902] [Medline: 29888677]

XSL•FO
RenderX

15. Koumakis L, Chatzaki C, Kazantzaki E, Maniadi E, Tsiknakis M. Dementia care frameworks and assistive technologies for their implementation: a review. IEEE Rev Biomed Eng 2019;12(12):4-18. [doi: 10.1109/RBME.2019.2892614] [Medline: 30640629]
16. Krick T, Huter K, Domhoff D, Schmidt A, Rothgang H, Wolf-Ostermann K. Digital technology and nursing care: a scoping review on acceptance, effectiveness and efficiency studies of informal and formal care technologies. BMC Health Serv Res 2019 Jun 20;19(1):400. [doi: 10.1186/s12913-019-4238-3] [Medline: 31221133]
17. Trainum K, Tunis R, Xie B, Hauser E. Robots in assisted living facilities: scoping review. JMIR Aging 2023 Mar 6;6(6):e42652. [doi: 10.2196/42652] [Medline: 36877560]
18. Van Velthoven MH, Cordon C. Sustainable adoption of digital health innovations: perspectives from a stakeholder workshop. J Med Internet Res 2019 Mar 25;21(3):e11922. [doi: 10.2196/11922] [Medline: 30907734]
19. Gjestsen MT, Wiig S, Testad I. Health care personnel's perspective on potential electronic health interventions to prevent hospitalizations for older persons receiving community care: qualitative study. J Med Internet Res 2020 Jan 2;22(1):e12797. [doi: 10.2196/12797] [Medline: 31895045]
20. Harst L, Wollschlaeger B, Birnstein J, Fuchs T, Timpel P. Evaluation is key: providing appropriate evaluation measures for participatory and user-centred design processes of healthcare IT. Int J Integr Care 2021 Jun 21;21(2):24. [doi: 10.5334/ijic.5529] [Medline: 34220388]
21. Greenhalgh T, Abimbola S. The NASSS framework - a synthesis of multiple theories of technology implementation. Stud Health Technol Inform 2019 Jul 30;263(263):193-204. [doi: 10.3233/SHTI190123] [Medline: 31411163]
22. Knapp M, Shehaj X, Wong G. Digital interventions for people with dementia and carers: effective, cost-effective and equitable? Neurodegener Dis Manag 2022 Oct;12(5):215-219. [doi: 10.2217/nmt-2022-0025] [Medline: 35833456]
23. Skivington K, Matthews L, Simpson SA, et al. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. BMJ 2021 Sep 30;374(374):n2061. [doi: 10.1136/bmj.n2061] [Medline: 34593508]
24. Craig P, Dieppe P, Macintyre S, et al. Developing and evaluating complex interventions: the new Medical Research Council guidance. BMJ 2008 Sep 29;337(337):a1655. [doi: 10.1136/bmj.a1655] [Medline: 18824488]
25. Campbell M, Fitzpatrick R, Haines A, et al. Framework for design and evaluation of complex interventions to improve health. BMJ 2000 Sep 16;321(7262):694-696. [doi: 10.1136/bmj.321.7262.694] [Medline: 10987780]
26. Malterud K. Systematic text condensation: a strategy for qualitative analysis. Scand J Public Health 2012 Dec;40(8):795-805. [doi: 10.1177/1403494812465030] [Medline: 23221918]
27. Gjestsen MT, Wiig S, Testad I. What are the key contextual factors when preparing for successful implementation of assistive living technology in primary elderly care? A case study from Norway. BMJ Open 2017 Sep 7;7(9):e015455. [doi: 10.1136/bmjopen-2016-015455] [Medline: 28882908]
28. Catwell L, Sheikh A. Evaluating eHealth interventions: the need for continuous systemic evaluation. PLoS Med 2009 Aug;6(8):e1000126. [doi: 10.1371/journal.pmed.1000126] [Medline: 19688038]
29. Greenhalgh T, Wherton J, Papoutsi C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017 Nov 1;19(11):e367. [doi: 10.2196/jmir.8775] [Medline: 29092808]
30. Duffy B, Oyebode JR, Allen J. Burnout among care staff for older adults with dementia: the role of reciprocity, self-efficacy and organizational factors. Dementia (London) 2009;8(4):515-541. [doi: 10.1177/1471301209350285] [Medline: 29172697]
31. Ballard C, Orrell M, Sun Y, et al. Impact of antipsychotic review and non-pharmacological intervention on health-related quality of life in people with dementia living in care homes: WHELD-a factorial cluster randomised controlled trial. Int J Geriatr Psychiatry 2017 Oct;32(10):1094-1103. [doi: 10.1002/gps.4572] [Medline: 27640872]

Abbreviations

Edited by A Mavragani; submitted 22.04.24; peer-reviewed by J Farmer, W Tang; revised version received 21.03.25; accepted 01.04.25; published 08.05.25.

Please cite as:
Austbø LBH, Testad I, Gjestsen MT
Using a Robot to Address the Well-Being, Social Isolation, and Loneliness of Care Home Residents via Video Calls: Qualitative Feasibility Study
JMIR Form Res 2025;9:e59764
URL: https://formative.jmir.org/2025/1/e59764
doi:10.2196/59764

XSL•FO
RenderX

© Lise Birgitte Holteng Austbø, Ingelin Testad, Martha Therese Gjestsen. Originally published in JMIR Formative Research (https://formative.jmir.org), 8.5.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Evaluating the Use of a Note-Taking App by Japanese Resident Physicians: Nationwide Cross-Sectional Study

Taiju Miyagami[1], MD, PhD; Yuji Nishizaki[1,2], MPH, MD, PhD; Taro Shimizu[3], MSc, MBA, MPH, MD, PhD; Yu Yamamoto[4], MD; Kiyoshi Shikino[5,6], MHPE, MD, PhD; Koshi Kataoka[2], MMSc, PhD; Masanori Nojima[7], MPH, MD, PhD; Gautam A Deshpande[1], MD; Toshio Naito[1], MD, PhD; Yasuharu Tokuda[8,9], MPH, MD, PhD

[1]Department of General Medicine, Faculty of Medicine, Juntendo University, Bunkyo-ku Hongo 3-1-3, Tokyo, Japan
[2]Division of Medical Education, School of Medicine, Juntendo University, Tokyo, Japan
[3]Department of Diagnostic and Generalist Medicine, Dokkyo Medical University Hospital, Tochigi, Japan
[4]Division of General Medicine, Center for Community Medicine, Jichi Medical University, Tochigi, Japan
[5]Department of General Medicine, Chiba University Hospital, Chiba, Japan
[6]Department of Community-Oriented Medical Education, Chiba University Graduate School of Medicine, Chiba, Japan
[7]Center for Translational Research, The Institute of Medical Science, The University of Tokyo, Tokyo, Japan
[8]Muribushi Okinawa Center for Teaching Hospitals, Okinawa, Japan
[9]Tokyo Foundation for Policy Research, Tokyo, Japan

Corresponding Author:
Yuji Nishizaki, MPH, MD, PhD
Department of General Medicine, Faculty of Medicine, Juntendo University, Bunkyo-ku Hongo 3-1-3, Tokyo, Japan

Abstract

Background: Note-taking is a method that has long been used to optimize studying. Recent innovations have seen the introduction of digital note-taking using software apps. Although the current state of digital note-taking has been verified mainly among students, the use and efficacy of digital note-taking by physicians in actual clinical practice remain unknown. Therefore, we sought to understand the characteristics of note-taking residents using a note-taking app and determine whether there is a difference in basic medical knowledge compared to that of nondigital note-taking residents.

Objective: This study investigated the use of a digital note-taking app by Japanese resident physicians.

Methods: This analytical cross-sectional study was conducted in resident physicians during the General Medicine In-Training Examination (GM-ITE), a clinical competency examination for resident physicians. The GM-ITE is a multiple-choice test with a maximum score of 80 points. Using a structured questionnaire, we collected data on the sociodemographic characteristics (sex, age, postgraduate year [PGY], or others), clinical training, GM-ITE scores, and the use of an app for note-taking to record case experience. The GM-ITE evaluated the scores by dividing them into 4 groups (groups 1 - 4), in order from the lowest to the highest. We conducted a multivariate analysis of sociodemographic, clinical training, and GM-ITE score variables to determine the independent predictors of the use of a digital note-taking app.

Results: This study included 3833 participants; 1242 (32.4%) were female, 1988 (51.8%) were PGY 1 residents, 2628 (68.6%) were training in a rural area, 3236 (84.4%) were in community-based hospitals, and 1750 (45.3%) were app users. The app users were more likely to be in their PGY 2, to work in a community-based hospital, to have general internal medicine rotation experience, to use online medical resources more frequently, and to have more time for self-study. The results showed that the app users group had a higher GM-ITE score than the nonapp users group (adjusted odds ratio 0.74, 95% CI 0.25 to 1.22; P=.003).

Conclusions: To the best of our knowledge, this is the first study to investigate note-taking by physicians in Japan using apps. The app users group had a higher GM-ITE score than the nonapp users, suggesting that they may have higher clinical skills. In the future, we would like to conduct more in-depth research on the facts of note-taking using apps, based on our results.

(JMIR Form Res 2025;9:e55087) doi:10.2196/55087

KEYWORDS

note-taking; applications; digital notes; resident physicians; medical education

XSL•FO
RenderX

Introduction

Note-taking during learning has long been a mainstay of educational practice and data over the last 60 years, demonstrating its contribution to improved learning and test scores [1,2]. In 1995, a study on effective note-taking among students reported that free note-taking by learners was a particularly effective process [3]. A 2002 report noted that the most effective way for medical students to perform well was to take written notes on materials prepared in advance by the teachers [4]. Since 2010, the use of mobile technology in education has increased in medical and pharmacy schools and is correlated with high student satisfaction [5,6]. Note-taking is also useful to improve understanding during the class and to retrospectively reflect on learned material [7].

In recent years, medical students have been using digital devices extensively for learning and literature searches [8]. Moreover, in medical education, digital methods are often used mainly as a replacement for traditional methods [9]. These days, medical students have increasingly encouraged incorporating digital tools, such as the iPad, for note-taking [10]. It has been suggested that digital note-taking may be better than nondigital note-taking concerning readability and search capabilities during revision of the notes [11]. Furthermore, students tend to have a positive attitude toward learning the use of digital tools [12]. They are reportedly able to use mobile devices as online information resources, which leads to improved learning efficiency in the classroom [13]. Multiple mobile devices are also being used in clinical settings. They are used for telemedicine, and have been developed to support clinical decision-making, providing timely feedback to residents in addition to improving clinical skills [14,15].

However, there are also disadvantages to the use of digital tools such as apps. One of the disadvantages of using apps in clinical settings is the challenge of using them well in multiple electronic tools [8]. First, medical students are not used to digital technology, and a tendency exists for them to prefer paper-based materials to digital ones [5], which is supported by a study that reported the digital group performing worse than the group that took notes using paper and pen [16]. Multitasking may have been a factor in the poor performance [16]. Additionally, a study in Mexico has also raised the issue of the difficulty of using online content when the internet connection is poor, and whether or not the instructor is willing to allow the use of online content [17].

However, we are unaware of the type of individuals using a note-taking app and the background factors, such as the extent to which clinical skills differ between app users and nonapp users.

Therefore, we sought to understand the characteristics of note-taking residents using a note-taking app and determine whether there is a difference in the basic medical knowledge compared to that of nondigital note-taking residents.

Methods

Design and Participants

This was a nationwide analytical cross-sectional survey of resident physicians in Japan. Since 2004 in Japan, newly graduated resident physicians have undergone mandatory cross-disciplinary 2-year training (including rotations in internal medicine, emergency medicine, pediatrics, gynecology, psychiatry, surgery, and community medicine) after passing the national medical examination during their final year of medical school [18]. The departments for rotational training and duration and timing during the 2 years of training vary from hospital to hospital and are not set in stone. Resident physicians in their first (postgraduate year [PGY] 1) and second (PGY 2) years take the same General Medicine In-Training Examination (GM-ITE), and examination participants complete questionnaires that survey the actual status of resident physicians immediately after the examination. Participants were gathered by an announcement from the training supervisor at each teaching hospital, and resident physicians participated voluntarily. This study was conducted between January 17, 2023, and January 30, 2023, and followed STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines [19].

Variables

The GM-ITE assesses general clinical knowledge and its practical relevance by the medical training guidelines of the Ministry of Health, Labor, and Welfare (MHLW) of Japan. This exam provides feedback to individual resident physicians and evaluates training programs and facilities. The GM-ITE comprises 80 multiple-choice questions, with optional tests that are answered by approximately 50% of all resident physicians in Japan. The GM-ITE is a test organized by the Japan Association for Medical Education Program (JAMEP). JAMEP is a nonprofit organization established with the aim of contributing to the promotion of Japanese medicine through support for training and education to improve the quality of medical care, as a third-party organization that checks the quality of medical education in Japan. The authors received the data from the person in charge at JAMEP. In line with the MHLW's goals for resident physicians, the 2022 GM-ITE consisted of a total of 80 structured questions, including (1) eight questions on medical history taking and professionalism, (2) eighteen questions on symptomatology and clinical reasoning, (3) eighteen questions on physical examination and clinical skills, and (4) thirty-six questions on basic clinical knowledge, including disease knowledge. The GM-ITE is answered on the computer, and after the test is finished, the participant decides whether or not to agree to participate in questionnaires on the computer. Thereafter, they answer the following questions. Baseline characteristics included resident physicians' age, gender, year in training, and hospital location (urban or rural; the 20 cities designated by the Ministry of Internal Affairs and Communications and the 23 wards of Tokyo were defined as urban areas, and all other cities were defined as rural areas), general hospital type (community-based or university hospital), number of beds, whether rotation in a general medicine department was completed, number of emergency department

XSL•FO
RenderX

shifts per month, average number of inpatients cared for daily, use of medical online resources, self-study time per day, and duty hours per week. The ages of the resident physicians were divided into 3 groups. Group 1 consisted of resident physicians aged 24 and 25 years, group 2 consisted of physicians aged 26 and 27 years, and group 3 consisted of physicians aged 28, 29, and 30 years and older. The main outcome question regarding app use was asked as follows. First, participants were asked whether they kept records of the cases they had experienced, and those who did not were excluded. They were then asked whether they used the application software to record the cases they had experienced, and they responded with either yes or no. The relationship between the total GM-ITE score and app use was also investigated. The GM-ITE is taken by residents under supervision at hospitals. However, due to the recent COVID-19 pandemic, some residents took the examination from their homes. The data of residents who took the exam at home were excluded from the analysis. Participants who did not collect cases for self-study, either on paper or through an app, were excluded. Additionally, those with missing data were excluded.

Ethical Considerations

This study was approved by the JAMEP Ethics Committee (No. 22‐31). The study was conducted in accordance with the principles of the Declaration of Helsinki [20]. All the methods were performed following the Ethical Guidelines for Medical and Health Research Involving Human Subjects.

All participants reviewed this study document detailing data anonymization, voluntariness, and the dissemination of research outcomes before involvement. Only participants who provided informed consent were included in this study. The data are being carefully managed by the JAMEP staff in a locked file. The data were anonymized at the time of passing them to the authors; thus, it is not possible to identify the participants. In particular, the questions in this study are limited to minor invasive questions. The participants answered the questions on a completely voluntary basis, and no compensation, including rewards, was provided.

Data Analysis

The results are presented as means (SD) for continuous variables or as prevalence (%) for categorical variables. During data analysis, the subgroup analysis (app users vs app nonusers) were performed to determine the factors associated with the app use. Comparisons between the 2 groups were performed using the chi-square test. Univariate or multivariate analyses with a mixed logistic regression model were performed using a note-taking app to collect cases as the objective variable.

Other fixed covariates were sex, age, PGY, hospital location, hospital type, bed number, general medicine department rotation, number of emergency department duties per month, average number of inpatients cared for daily, use of medical online resources, self-study time per day, and duty hours per week. The GM-ITE scores were divided into 4 quartiles from top to bottom. Group 1 had the lowest score groups and group 4 had the highest score.

The total GM-ITE score was used as the objective variable for the analysis of secondary outcomes. Univariate or multivariate analyses with a linear mixed model were conducted using the GM-ITE total score as the objective variable, with a random effect of the training hospital. The fixed covariate was the presence or absence of app use, in addition to the previously mentioned variables.

The 95% CIs were defined using the mixed logistic regression and linear mixed models. All calculations were performed using SAS (version 9.4; SAS Institute). *P* values <.05 were considered statistically significant. The sample size was not calculated for this exploratory study.

Results

A total of 9011 residents from 662 teaching hospitals participated in the 2022 GM-ITE. Participants who did not provide consent (n=2375), took the GM-ITE at home (n=573), did not record their cases for self-study purposes (n=1366), and had at least 1 missing data variable (n=864) were excluded. A total of 3833 participants were included in the final analysis (Figure 1).

Most (n=2591, 67.6%) of the participants were female, 32.4% (n=1242) were PGY-1 resident physicians, and 52.4% (n=2009) were of age group 2. Most (n=2628, 68.6%) of them were training in rural areas, and 84.4% (3236) practiced in community-based hospitals, while less than half (n=1772, 46.2%) had completed a general medicine rotation. Emergency department duty: most (n=2766, 72.2%) of the participants had 3‐5 duties per month in the emergency department. Almost half (n=1896, 49.5%) of the participants had less than 60 duty hours per week. Nearly half (n=1933, 50.4%) reported caring for 5‐9 inpatients daily. Furthermore, less than one-third of the participants (n=1095, 28.6%) were using online medical resources, while 41.4% (n=1586) and 41% (n=1570) spent 0‐30 minutes and 31‐60 minutes for self-study time per day, respectively (Table 1). The participants' GM-ITE scores ranged from 12 to 71 points.

A substantial proportion (n=1750, 45.3%) of participants were using the note-taking app.

Among this study's participants, PGY-2 residents used the app more frequently than PGY-1 residents (n=896, 48.6% vs n=843, 42.4%, respectively; adjusted odds ratio [aOR] 1.10, 95% CI 1.03 to 1.41; *P*=.02). Participants in community-based hospitals were more likely to use the app than those in university hospitals (n=1513, 46.8% vs n=226, 37.9%, aOR 1.42, 95% CI 1.04 to 1.94; *P*=.03). Those who worked in hospitals with a larger number of beds were also more likely to use the app (aOR 1.08, 95% CI 1.02 to 1.13; *P*=.004). Those who had rotated through the Department of General Medicine were more likely to use the app than those who had not (n=881, 49.7% vs n=858, 41.6%, aOR 1.19, 95% CI 1.02 to 1.39; *P*=.03). The number of inpatients cared for per day was most frequently 10‐14 among app users, which showed a significant difference (aOR 1.62, 95% CI 1.18 to 2.21; *P*=.003). Participants in the app user group also used online medical resources more frequently (n=1365, 49.9% vs n=374, 34.2%, aOR 1.52, 95% CI 1.3 to 1.79; *P*<.001; Table 2).

XSL•FO
RenderX

The GM-ITE total score was used for evaluation. The following variables showed significant differences in the GM-ITE total score in the multivariate analysis. Female scores tended to be lower than male scores (adjusted difference [aD] −1.18, 95% CI −1.69 to −0.68; *P*<.001). PGY-2 scores tended to be higher than PGY-1 scores (aD 2.45, 95% CI 1.92 to 2.99; *P*<.001). The age of the resident physician tended to be lower in both group 2 and group 3 than in group 1, with lower GM-ITE scores (group 2 [aD −2.55, 95% CI −3.17 to −1.92; *P*<.001] and group 3 [aD -5.88, 95%CI −6.62 to −5.13; *P*<.001]). Predictors of the high GM-ITE scores were working at community-based hospitals (aD 3.64, 95% CI 2.44 to 4.84; *P*<.001), hospitals with more beds (aD 0.36, 95% CI 0.17 to 0.55; *P*<.001), high number of emergency department duties per month (aD 1.98, 95% CI 0.15 to 3.80; *P*=.03), high number of inpatient (aD 1.40, 95% CI 0.32 to 2.47; *P*=.01), using medical online resources (aD 1.45, 95% CI 0.9 to 1.99; *P*<.001), and high number of self-study time (aD 1.41, 95% CI 0.13 to 2.69; *P*=.03). The GM-ITE score tended to be higher for the group working 60 to 79 hours per week than for the group working 0 to 59 hours per week (aD 0.97, 95% CI 0.45 to 1.59; *P*<.001). However, for the group working 80 hours or more per week, there was no difference in score from the group working 0 to 59 hours per week, suggesting that working 60 to 79 hours per week is optimal for improving the GM-ITE score. Moreover, app users had significantly higher scores (aD 0.74, 95% CI 0.25 to 1.22; *P*=.003; Table 3).

Figure 1. Participant characteristics. GM-ITE: General Medicine In-Training Examination.

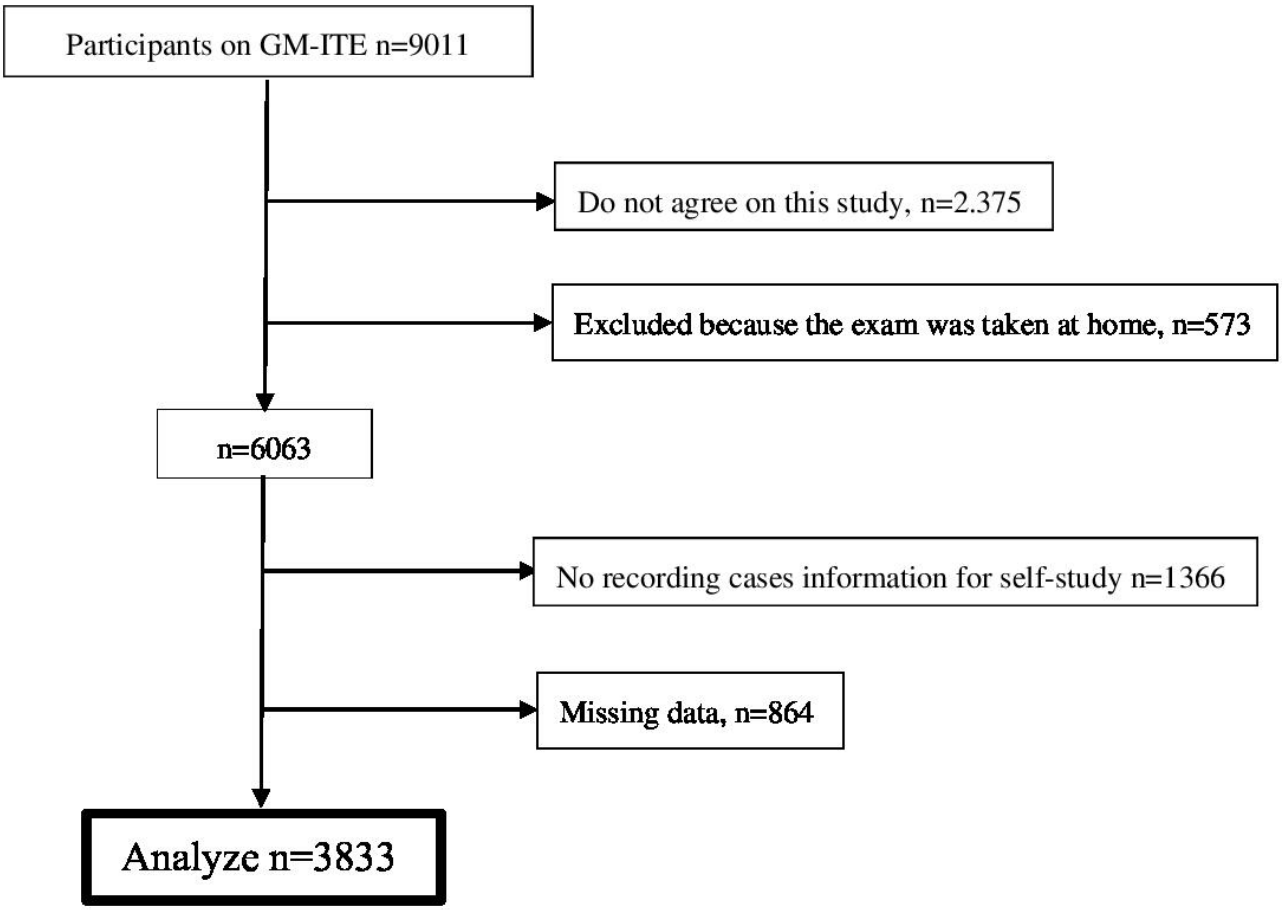

XSL•FO
RenderX

Table . Participant characteristics (N=3833).

Participants' profile	Total
Sex, n (%)	
Female	1242 (32.4)
Male	2591 (67.6)
Grade, n (%)	
PGY-1[a]	1.8 (51.9)
PGY-2	1845 (48.1)
Age, n (%)	
Group 1	891 (23.3)
Group 2	2009 (52.4)
Group 3	933 (24.3)
Hospital location, n (%)	
Urban	1205 (31.4)
Rural	2628 (68.6)
Hospital type	
Community-based hospital, n (%)	3236 (84.4)
University hospital, n (%)	597 (15.6)
Bed number, mean (SD)	547 (220)
General medicine department rotation, n (%)	
Yes	1772 (46.2)
No	2061 (53.7)
Emergency department duty, n (%)	
0 per month	94 (2.5)
1 - 2 per month	615 (16)
3 - 5 per month	2766 (72.2)
>6 per month	347 (9.1)
Unknown	11 (0.3)
Number of cared for inpatients, n (%)	
0 - 4	1481 (38.6)
5 - 9	1933 (50.4)
10 - 14	260 (6.7)
15 and over	76 (2)
Unknown	83 (2.2)
Using medical online resources, n (%)	
Yes	1095 (28.6)
No	2738 (71.4)
Self-study time per day, n (%)	
0 - 30 minutes	1586 (41.4)
31 - 60 minutes	1570 (41)
61 - 90 minutes	482 (12.6)
91 minutes or more	143 (3.7)
None	52 (1.4)
Duty hour (h/wk), n (%)	

XSL•FO
RenderX

Participants' profile	Total
0 - 59	1896 (49.5)
60 - 79	1401 (36.6)
>80	536 (14)
GM-ITE[b] score total, mean (SD)	45.9 (8.2)
Group 1, range	12 - 40
Group 2, range	41 - 46
Group 3, range	47 - 52
Group 4, range	53 - 71

[a]PGY: postgraduate year.

[b]GM-ITE: General Medicine In-Training Examination.

XSL•FO
RenderX

Table . Comparison of app-using and nonusing groups. We conducted a logistic regression analysis comparing people who used the app with those who did not, using those who did not use the app as the reference.

		App use (n=1750), n (%)	Nonapp use (n=2133), n (%)	OR[a]	Univariable (95% CI)	*P* value	aOR[b]	Multivariable (95% CI)	*P* value
Gender									
	Male	1175 (45.3)	1416 (54.7)	Reference	N/A[c]	N/A	N/A	N/A	N/A
	Female	564 (45.4)	678 (54.6)	1	0.88 to 1.15	.96	1.10	0.95 to 1.28	.20
Grade									
	PGY-1[d]	843 (42.4)	1145 (57.6)	Reference	N/A	N/A	N/A	N/A	N/A
	PGY-2	896 (48.6)	949 (51.4)	1.28	1.13 to 1.46	<.001	1.20	1.03 to 1.41	.02
Resident age									
	Group 1	407 (45.7)	484 (54.3)	Reference	N/A	N/A	N/A	N/A	N/A
	Group 2	944 (47)	1065 (53)	1.05	0.9 to 1.24	.51	0.97	0.8 to 1.16	.71
	Group 3	388 (41.6)	545 (58.4)	0.85	0.7 to 1.02	.08	0.88	0.7 to 1.1	.27
Hospital location									
	Urban	545 (45.2)	660 (54.8)	Reference	N/A	N/A	N/A	N/A	N/A
	Rural	1194 (45.4)	1434 (54.6)	1.01	0.88 to 1.16	.85	1.03	0.85 to 1.24	.80
Hospital type									
	University Hospital	226 (37.9)	371 (62.1)	Reference	N/A	N/A	N/A	N/A	N/A
	Community-based hospital	1513 (46.8)	1723 (53.2)	1.42	1.21 to 1.72	<.001	1.42	1.04 to 1.94	.03
Bed number		555 (219)[e]	539 (221)[e]	1.03	1 to 1.06	.02	1.08	1.02 to 1.13	.004
General medicine department rotation									
	No	858 (41.6)	1203 (58.4)	Reference	N/A	N/A	N/A	N/A	N/A
	Yes	881 (49.7)	891 (50.3)	1.37	1.21 to 1.58	<.001	1.19	1.02 to 1.39	.03
Emergency department duty									
	0 per month	32 (34)	62 (66)	Reference	N/A	N/A	N/A	N/A	N/A
	1 - 2 per month	250 (40.7)	365 (59.3)	1.33	0.84 to 2.1	.22	1.16	0.71 to 1.91	.56
	3 - 5 per month	1276 (46.1)	1490 (53.9)	1.66	1.08 to 2.56	.02	1.32	0.82 to 2.14	.26
	>6 per month	177 (51)	170 (49)	2.02	1.25 to 3.25	.004	1.6	0.93 to 2.75	.09
	Unknown	4 (36.3)	7 (63.6)	1.11	0.3 to 4.06	.88	0.81	0.2 to 3.29	.77
Number of cared for inpatients									
	0 - 4	623 (42.1)	858 (57.9)	Reference	N/A	N/A	N/A	N/A	N/A
	5 - 9	903 (46.7)	1030 (53.3)	1.21	1.05 to 1.38	.007	1.03	0.88 to 1.21	.70
	10 - 14	149 (59.6)	111 (40.4)	1.85	1.42 to 2.41	<.001	1.62	1.18 to 2.21	.003
	15 and over	33 (43.4)	43 (56.6)	1.06	0.66 to 1.68	.82	0.92	0.54 to 1.56	.74
	Unknown	31 (42.5)	52 (57.5)	0.82	0.52 to 1.30	.41	0.77	0.47 to 1.26	.30
Using medical online resources									
	No	374 (34.2)	721 (65.8)	Reference	N/A	N/A	N/A	N/A	N/A
	Yes	1365 (49.9)	1373 (50.2)	1.92	1.66 to 2.22	<.001	1.52	1.3 to 1.79	<.001
Self-study time per day									

XSL•FO
RenderX

		App use (n=1750), n (%)	Nonapp use (n=2133), n (%)	OR[a]	Univariable (95% CI)	*P* value	aOR[b]	Multivariable (95% CI)	*P* value
	0 - 30 min	625 (39.4)	961 (60.6)	Reference	N/A	N/A	N/A	N/A	N/A
	31 - 60 min	743 (47.3)	827 (52.7)	1.38	1.2 to 1.59	<.001	1.23	1.06 to 1.44	.008
	61 - 90 min	275 (57.1)	207 (42.9)	2.04	1.66 to 2.51	<.001	1.79	1.43 to 2.24	<.001
	91 min or more	82 (57.3)	61 (42.7)	2.07	1.46 to 2.92	<.001	1.63	1.12 to 2.37	.01
	None	14 (26.9)	38 (73.1)	0.57	0.31 to 1.06	.07	0.59	0.31 to 1.14	.12
Duty hour (h/wk)									
	0 - 59	803 (42.4)	1093 (57.6)	Reference	N/A	N/A	N/A	N/A	N/A
	60 - 79	661 (47.2)	740 (52.8)	1.22	1.06 to 1.4	.006	0.99	0.85 to 1.15	.88
	>80	275 (51.3)	261 (48.7)	1.43	1.18 to 1.74	<.001	1.12	0.9 to 1.39	.31
GM-ITE[f] score									
	Group 1[g]	370 (37.8)	609 (62.2)	Reference	N/A	N/A	N/A	N/A	N/A
	Group 2	375 (42.1)	515 (57.9)	1.2	1 to 1.44	.06	1.06	0.87 to 1.29	.58
	Group 3	449 (45.9)	529 (54.1)	1.4	1.17 to 1.67	<.001	1.15	0.94 to 1.40	.18
	Group 4	545 (55.3)	441 (44.7)	2.03	1.7 to 2.44	<.001	1.42	1.15 to 1.75	.001

[a]OR: odds ratio.

[b]aOR: adjusted odds ratio.

[c]N/A: not applicable.

[d]PGY: postgraduate year.

[e]Mean (SD).

[f]GM-ITE: General Medicine In-Training Examination.

[g]The GM-ITE group was classified into four quartiles in order of score, from group 1.

XSL•FO
RenderX

Table . Evaluation based on total GM-ITE[a] score. Analysis using multiple regression analysis.

		Difference	Univariable (95% CI)	*P* value	Adjusted difference	Multivariable (95% CI)	*P* value
Gender							
	Female (vs male)	−0.86	−1.41 to −0.31	.002	−1.18	−1.69 to −0.68	<.001
Grade							
	PGY-2[b] (vs PGY-1)	1.4	0.88 to 1.92	<.001	0.18	−0.04 to 0.4	.11
Resident age							
	Group 1	N/A[c]	N/A	N/A	N/A	N/A	N/A
	Group 2	−1.38	−2.01 to −0.75	<.001	−2.55	−3.17 to −1.92	<.001
	Group 3	−5.29	-6.02 to −4.55	<.001	−5.88	−6.62 to −5.13	<.001
Hospital location							
	Rural (vs urban)	0.06	−0.5 to 0.63	.82	0.49	−0.25 to 1.22	.19
Hospital type							
	Community-based hospital (vs university hospital)	4.58	3.88 to 5.28	<.001	3.64	2.44 to 4.84	<.001
Bed number		−0.07	−0.19 to 0.05	.28	0.36	0.17 to 0.55	<.001
General medicine department rotation		1.5	0.98 to 2.02	<.001	0.6	0.06 to 1.15	.03
Emergency department duty							
	0 per month	N/A	N/A	N/A	N/A	N/A	N/A
	1 - 2 per month	1.72	−0.04 to 3.5	.06	0.33	−1.32 to 1.99	.69
	3 - 5 per month	4.18	2.51 to 5.86	<.001	1.23	−0.38 to 2.83	.13
	>6per month	5.17	3.31 to 7.03	<.001	1.98	0.15 to 3.8	.03
	Unknown	3.1	−1.99 to 8.19	.23	2.45	−2.25 to 7.15	.31
Number of cared for inpatients							
	0 - 4	N/A	N/A	N/A	N/A	N/A	N/A
	5 - 9	2.09	1.54 to 2.64	<.001	1.48	0.95 to 2.02	<.001
	10 - 14	2.06	1 to 3.14	<.001	1.4	0.32 to 2.47	.01
	15 and over	1.53	−0.35 to 3.41	.11	0.51	−1.32 to 2.34	.59
	Unknown	0.57	−1.24 to 2.37	.54	0.31	−1.33 to 1.95	.71
Using medical online resources		3.11	2.9 to 3.68	<.001	1.45	0.9 to 1.99	<.001
Self-study time per day							
	0 - 30 min	N/A	N/A	N/A	N/A	N/A	N/A
	31 - 60 min	1.87	1.3 to 2.43	<.001	0.87	0.35 to 1.39	.001
	61 - 90 min	2.46	1.62 to 3.29	<.001	0.99	0.23 to 1.76	.01
	91 min or more	3.54	2.15 to 4.93	<.001	1.41	0.13 to 2.69	.03
	None	−1.04	−3.29 to 1.21	.37	−0.7	−2.72 to 1.31	.49
Duty hour (h/wk)							
	0 - 59	N/A	N/A	N/A	N/A	N/A	N/A
	60 - 79	2.27	1.7 to 2.83	<.001	0.97	0.45 to 1.49	<.001
	>80	2.06	1.28 to 2.85	<.001	0.29	−0.45 to 1.04	.44

XSL•FO
RenderX

	Difference	Univariable (95% CI)	*P* value	Adjusted difference	Multivariable (95% CI)	*P* value
Using application	2.11	1.59 to 2.63	<.001	0.74	0.25 to 1.22	.003

[a]GM-ITE: General Medicine In-Training Examination.

[b]PGY: postgraduate year.

[c]N/A: not applicable.

Discussion

Principal Findings

To our knowledge, this is the first study to compare the characteristics of note-taking app users with those of nonusers among resident physicians actively engaged in clinical practice in Japan. The results showed that app users, compared to nonapp users, tend to be more likely to use online medical resources, have rotated in general medicine, in community-based hospitals, be more experienced trainees, and engage in more self-study. GM-ITE scores were also higher among those who used the app.

Resident physicians who used online resources tended to use the apps more frequently. The use of online resources and apps is considered to be associated with digital literacy. A previous report stated that one of the disadvantages of digital note-taking was a limited period during which students were not able to use the apps easily [21]. Students found apps to be difficult to use if they were using an app that was unfamiliar to them [21]. Another study reported that higher information and digital literacy were associated with satisfactory information retrieval among health care professionals [22]. In a study comparing the effectiveness of information literacy versus digital literacy among family physicians in the Middle East, those with higher information literacy were more effective in searching for information relevant to their practice [23]. Similar to previous reports, both information literacy and digital literacy were involved and affected the frequency of app use.

Participants who used apps tended to have rotated through the Department of General Medicine compared to nonapp users. Japan has an increasingly aging population, and Japanese general medicine doctors are often tasked with treating older adults and are required to deal with multimorbidity [24,25]. Additionally, there have been reports that general medicine doctors based in clinics often use apps to consult with patients during consultations; thus, general practice is highly compatible with app use [26]. Furthermore, as general medicine doctors are keen propagators of education, they could encourage their trainees to use the app, which could have contributed to the results of this study [27].

Our results also indicate that resident physicians in community-based hospitals work in an environment with more app use than in university hospitals. A study of Japanese resident physicians reported that they experienced more cases, spent more time on self-study, worked more hours, and tended to have higher GM-ITE scores in community-based hospitals than university hospitals [28]. Therefore, it is possible that residents in busy community-based hospitals need to compile a large amount of information for self-study, thereby using digital aides to collect and organize this information more efficiently and effectively.

Compared with PGY-1 residents, PGY-2 residents showed a higher rate of app use. There are 2 possible reasons for this. First, PGY-2 residents may have exchanged information with each other on the efficiency of using apps, such as those used for case collection, during their initial year of training and thus have been influenced to use the technology preferentially. In the Japanese training environment, PGY-1s often spend most of their time on mandatory rotations, while PGY-2 students are provided more opportunities for elective rotations. Almost all physicians in Japan proceed to a single specialty department by their PGY-3 year [29]. Therefore, it is possible that PGY-2 residents were motivated to study hard to prepare for their specialty training.

It may be unsurprising that total GM-ITE scores were higher among app users, given that those using digital note-taking have the advantage of being able to organize what they have learned more efficiently than paper note-taking. In the previous studies, high-achieving medical students were aware of the need to manage their time efficiently [30,31]. Increased efficiency is one of the critical concepts of app use. A 2014 report indicated that paper-based systems were more likely to be retained in short-term memory and contribute to higher conceptual awareness; however, a 2022 report showed that short-term memory retention was not significantly different between paper and digital note-taking [32,33]. No reports have compared the facts of paper and digital note-taking on long-term memory; however, digital note-taking may be superior to traditional paper note-taking per efficiency and the amount of information that can be recorded and retained without loss or degradation [11,33]. Thus, the results of this study also support the results of previous reports.

With respect to the GM-ITE score, a positive correlation was observed with being a community-based hospital, self-study time, working hours, the number of emergency shifts, the number of inpatients, and the year of residency, which was consistent with the results of previous reports [27,28]. This suggests that the GM-ITE is a test that is related to clinical practice. The fact that male doctors tended to score higher may have been influenced by the gender hierarchy in Japan [34].

Finally, the app user group spent more time in self-study. App users may have been more motivated, resulting in a more positive effect on clinical performance. Previous studies have also reported that being self-motivated was associated with higher rates of graduation from graduate school [35]. From the results of this study, we consider the possibility that being motivated may have led to an interest in various things, including app use and improved clinical performance.

XSL•FO
RenderX

Limitations

This study has several limitations. First, our methodology did not specify detailed note-taking procedures, leaving these potentially important specifics unknown. In particular, the survey did not examine the motivation of residents or their ideas about personal efficiency. In this questionnaire survey, the answer item is "30 years old or older"; hence, the exact average age could not be calculated. In addition, this study did not examine the specific product names of the apps, descriptions of data collection methods, frequency of use, satisfaction, or challenges; we wish to investigate these variables in future studies. Also, as this was a cross-sectional study, causal relationships remain unclear. Further, a sample size calculation was not performed because this was an exploratory study. This may have affected the interpretation and effectiveness of the research results.

Conclusions

This pioneering study on app-based note-taking among physicians revealed that app users are often efficient and motivated, offering valuable insights for medical education. Building on these findings, future research should further explore app-based note-taking and its potential integration into clinical practice.

Acknowledgments

The authors express their gratitude to the Japan Association for Medical Education Program (JAMEP) secretariat for the excellent assistance provided for our research. This work was supported by the Ministry of Health, Labor, and Welfare (MHLW) Policy Grants of Research on Regional Medicine (21IA2004). The MHLW was not involved in this study's design, conduct, data analysis and interpretation, paper preparation, review, approval, or the decision to submit this paper for publication.

Data Availability

The datasets used for this study are available from the corresponding author upon reasonable request.

Authors' Contributions

TM and YN had full access to all of the data in the study, and they take responsibility for the integrity of the data and the accuracy of the data analysis. TM, YN, TS, YY, KS, GAD, and YT conceptualized this study. TM and YN did the data curation. TM, YN, KK, and MN conducted the formal analysis. YN, TN, and YT acquired the funding. TM, YN, YY, KS, and YT handled the investigation. TM, YN, TS, YY, and YT worked on the methodology. TM, YN, TN, and YT administered the project. TM, YN, KK, and MN collected resources. TM, YN, KK, and MN assisted with the software. GAD, TN, and YT supervised. TM, YN, KK, and MN validated and visualized this study. TM and YN wrote the original paper. All authors reviewed and edited this paper.

Conflicts of Interest

YN received an honorarium from the Japan Association for Medical Education Program (JAMEP) as the General Medicine In-Training Examination (GM-ITE) project manager. YT is the director of the JAMEP. KS received an honorarium from the JAMEP as a reviewer of the GM-ITE. KS, TS, and YY received honoraria from the JAMEP as examination preparers of the GM-ITE. All other authors do not have any competing interests to declare.

References

1. Berliner D. Effects of test-like events and note-taking on learning from lecture instruction—technical report. : Center for Educational Research, University of Massachusetts; 1970.
2. Di Vesta FJ, Gray GS. Listening and note taking. J Educ Psychol 1972 Feb;63(1):8-14. [doi: 10.1037/h0032243] [Medline: 5007747]
3. Kiewra KA, Benton SL, Kim SI, Risch N, Christensen M. Effects of note-taking format and study technique on recall and relational performance. Contemp Educ Psychol 1995 Apr;20(2):172-187. [doi: 10.1006/ceps.1995.1011]
4. Morrison EH, McLaughlin C, Rucker L. Medical students' note-taking in a medical biochemistry course: an initial exploration. Med Educ 2002 Apr;36(4):384-386. [doi: 10.1046/j.1365-2923.2002.01167.x] [Medline: 11940180]
5. George P, Dumenco L, Doyle R, Dollase R. Incorporating iPads into a preclinical curriculum: a pilot study. Med Teach 2013;35(3):226-230. [doi: 10.3109/0142159X.2012.735384] [Medline: 23425119]
6. DiVall MV, Hayney MS, Marsh W, et al. Perceptions of pharmacy students, faculty members, and administrators on the use of technology in the classroom. Am J Pharm Educ 2013 May 13;77(4):75. [doi: 10.5688/ajpe77475] [Medline: 23716743]
7. Bauer A, Koedinger K. Pasting and encoding: note-taking in online courses. Presented at: Sixth IEEE International Conference on Advanced Learning Technologies (ICALT'06); Jul 5-7, 2006; Kerkrade, The Netherlands p. 789-793. [doi: 10.1109/ICALT.2006.1652559]
8. Fan S, Radford J, Fabian D. A mixed-method research to investigate the adoption of mobile devices and Web2.0 technologies among medical students and educators. BMC Med Inform Decis Mak 2016 Apr 19;16:43. [doi: 10.1186/s12911-016-0283-6] [Medline: 27094813]

XSL•FO
RenderX

9. Grainger R, Liu Q, Gladman T. Learning technology in health professions education: realising an (un)imagined future. Med Educ 2024 Jan;58(1):36-46. [doi: 10.1111/medu.15185] [Medline: 37555302]
10. Pyörälä E, Mäenpää S, Heinonen L, Folger D, Masalin T, Hervonen H. The art of note taking with mobile devices in medical education. BMC Med Educ 2019 Apr 2;19(1):96. [doi: 10.1186/s12909-019-1529-7] [Medline: 30940152]
11. Kim K, Turner SA, Pérez-Quiñones MA. Requirements for electronic note taking systems: a field study of note taking in university classrooms. Educ Inf Technol 2009 Sep;14(3):255-283. [doi: 10.1007/s10639-009-9086-z]
12. Wallace S, Clark M, White J. 'It's on my iPhone': attitudes to the use of mobile computing devices in medical education, a mixed-methods study. BMJ Open 2012;2(4):e001099. [doi: 10.1136/bmjopen-2012-001099]
13. Davies BS, Rafique J, Vincent TR, et al. Mobile Medical Education (MoMEd) - how mobile information resources contribute to learning for undergraduate clinical students - a mixed methods study. BMC Med Educ 2012 Jan 12;12:1. [doi: 10.1186/1472-6920-12-1] [Medline: 22240206]
14. Dimond R, Bullock A, Lovatt J, Stacey M. Mobile learning devices in the workplace: "as much a part of the junior doctors" kit as a stethoscope'? BMC Med Educ 2016 Aug 17;16(1):207. [doi: 10.1186/s12909-016-0732-z] [Medline: 27530343]
15. Okazaki S, Castañeda JA, Sanz S. Clinicians' assessment of mobile monitoring: a comparative study in Japan and Spain. Med 2 0 2013;2(2):e11. [doi: 10.2196/med20.2874] [Medline: 25075234]
16. Wood E, Zivcakova L, Gentile P, Archer K, De Pasquale D, Nosko A. Examining the impact of off-task multi-tasking with technology on real-time classroom learning. Comput Educ 2012 Jan;58(1):365-374. [doi: 10.1016/j.compedu.2011.08.029]
17. Rodríguez-Ríos A, Espinoza-Téllez G, Martínez-Ezquerro JD, Rendón-Macías ME. Information and communication technology, mobile devices, and medical education. J Med Syst 2020 Mar 16;44(4):90. [doi: 10.1007/s10916-020-01559-w] [Medline: 32173765]
18. Guideline for clinical training of medical practitioners-fiscal year 2020 edition. Ministry of Health Law. URL: https://www.mhlw.go.jp/content/10800000/ishirinsyokensyu_guideline_2020.pdf [accessed 2025-04-16]
19. von Elm E, Altman DG, Egger M, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. BMJ 2007 Oct 20;335(7624):806-808. [doi: 10.1136/bmj.39335.541782.AD] [Medline: 17947786]
20. WMA Declaration of Helsinki – ethical principles for medical research involving human participants. World Medical Association. URL: https://www.wma.net/policies-post/wma-declaration-of-helsinki/ [accessed 2025-04-16]
21. Gatwood J, Hohmeier K, Kocak M, Chisholm-Burns M. Acceptance of productivity software as a course management and collaboration tool among student pharmacists. Curr Pharm Teach Learn 2021 Apr;13(4):361-367. [doi: 10.1016/j.cptl.2020.11.019] [Medline: 33715797]
22. Kostagiolas PA, Kourouthanassis PE, Martzoukou K, Korfiatis N, Niakas D. Information seeking behavioural paths of physicians for diabetes mellitus care: a qualitative comparative analysis of information needs, sources, and barriers. Health Syst (Basingstoke) 2018;7(1):13-28. [doi: 10.1080/20476965.2017.1390050] [Medline: 31214336]
23. Antoun J, Lapin J, Beck D. Efficacy and efficiency of information retrieval of community family physicians at the point of care: exploring the associations with information and computer literacy. J Med Libr Assoc 2023 Jul 10;111(3):677-683. [doi: 10.5195/jmla.2023.1539] [Medline: 37483368]
24. Aoki T, Yamamoto Y, Ikenoue T, Onishi Y, Fukuhara S. Multimorbidity patterns in relation to polypharmacy and dosage frequency: a nationwide, cross-sectional study in a Japanese population. Sci Rep 2018 Feb 28;8(1):3806. [doi: 10.1038/s41598-018-21917-6] [Medline: 29491441]
25. Miyagami T, Shimizu T, Kosugi S, et al. Roles considered important for hospitalist and non-hospitalist generalist practice in Japan: a survey study. BMC Prim Care 2023 Jul 7;24(1):139. [doi: 10.1186/s12875-023-02090-w] [Medline: 37420166]
26. Harada T, Morikawa T, Furuya H, et al. Use of internet/application-based e-consults by general practitioners in Japan to resolve patients' problems: a descriptive observational study. Cureus 2024 Dec;16(12):e76690. [doi: 10.7759/cureus.76690] [Medline: 39742168]
27. Nishizaki Y, Shimizu T, Shinozaki T, et al. Impact of general medicine rotation training on the in-training examination scores of 11, 244 Japanese resident physicians: a nationwide multi-center cross-sectional study. BMC Med Educ 2020 Nov 13;20(1):426. [doi: 10.1186/s12909-020-02334-8] [Medline: 33187497]
28. Nishizaki Y, Nozawa K, Shinozaki T, et al. Difference in the general medicine in-training examination score between community-based hospitals and university hospitals: a cross-sectional study based on 15,188 Japanese resident physicians. BMC Med Educ 2021 Apr 15;21(1):214. [doi: 10.1186/s12909-021-02649-0] [Medline: 33858403]
29. Changes in junior resident clinical training system. Ministry of Health Law. 2018. URL: https://www.mhlw.go.jp/topics/bukyoku/isei/rinsyo/hensen [accessed 2025-04-16]
30. Bin Abdulrahman KA, Khalaf AM, Bin Abbas FB, Alanazi OT. Study habits of highly effective medical students. Adv Med Educ Pract 2021;12:627-633. [doi: 10.2147/AMEP.S309535] [Medline: 34135654]
31. Abdulghani HM, Al-Drees AA, Khalil MS, Ahmad F, Ponnamperuma GG, Amin Z. What factors determine academic achievement in high achieving undergraduate medical students? A qualitative study. Med Teach 2014 Apr;36 Suppl 1:S43-S48. [doi: 10.3109/0142159X.2014.886011] [Medline: 24617784]
32. Mueller PA, Oppenheimer DM. The pen is mightier than the keyboard: advantages of longhand over laptop note taking. Psychol Sci 2014 Jun;25(6):1159-1168. [doi: 10.1177/0956797614524581] [Medline: 24760141]

XSL•FO
RenderX

33. Wiechmann W, Edwards R, Low C, Wray A, Boysen-Osborn M, Toohey S. No difference in factual or conceptual recall comprehension for tablet, laptop, and handwritten note-taking by medical students in the United States: a survey-based observational study. J Educ Eval Health Prof 2022;19:8. [doi: 10.3352/jeehp.2022.19.8] [Medline: 35468666]
34. Watari T, Gupta A, Kataoka H. Representation of gender and postgraduate experience among professional medical society boards in Japan. JAMA Netw Open 2022 Dec 1;5(12):e2247548. [doi: 10.1001/jamanetworkopen.2022.47548] [Medline: 36534403]
35. Chang JC, Wu YT, Ye JN. A study of graduate students' achievement motivation, active learning, and active confidence based on relevant research. Front Psychol 2022;13:915770. [doi: 10.3389/fpsyg.2022.915770] [Medline: 35756232]

Abbreviations:

aD: adjusted difference
aOR: adjusted odds ratio
GM-ITE: General Medicine In-Training Examination
JAMEP: Japan Association for Medical Education Program
MHLW: Ministry of Health, Labor, and Welfare
PGY: postgraduate year
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by A Mavragani; submitted 02.12.23; peer-reviewed by J Kaswija, MX Yang, Y Ohira; revised version received 19.02.25; accepted 16.03.25; published 09.05.25.

Please cite as:
Miyagami T, Nishizaki Y, Shimizu T, Yamamoto Y, Shikino K, Kataoka K, Nojima M, Deshpande GA, Naito T, Tokuda Y
Evaluating the Use of a Note-Taking App by Japanese Resident Physicians: Nationwide Cross-Sectional Study
JMIR Form Res 2025;9:e55087
URL: https://formative.jmir.org/2025/1/e55087
doi:10.2196/55087

© Taiju Miyagami, Yuji Nishizaki, Taro Shimizu, Yu Yamamoto, Kiyoshi Shikino, Koshi Kataoka, Masanori Nojima, Gautam A Deshpande, Toshio Naito, Yasuharu Tokuda. Originally published in JMIR Formative Research (https://formative.jmir.org), 9.5.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Evaluating Feasibility and Acceptability of the "My HeartHELP" Mobile App for Promoting Heart-Healthy Lifestyle Behaviors: Mixed Methods Study

Jina Choo[1,2], PhD, DrPH; Songwhi Noh[1,2], MSN; Yura Shin[1,2], MSN

[1]College of Nursing, Korea University, 145, Anam-ro, Seongbuk-gu, Seoul, Republic of Korea

[2]Transdisciplinary Major in Learning Health Systems, Department of Health Sciences, Korea University Graduate School, Korea University, 145, Anam-ro, Seongbuk-gu, Seoul, Republic of Korea

Corresponding Author:
Jina Choo, PhD, DrPH
College of Nursing, Korea University, 145, Anam-ro, Seongbuk-gu, Seoul, Republic of Korea

Abstract

Background: Few mobile apps have strategies for self-monitoring multiple heart-healthy behaviors simultaneously, as well as automated and tailored feedback on individual behavioral outcomes for cardiovascular health. An app named "My HeartHELP" was developed for the general adult population to promote 6 heart-healthy lifestyle behaviors—physical activity, nonsedentary behaviors, healthy eating behaviors, nonsmoking, no alcohol binge drinking, and self-assessment of body weight. Three behavioral strategies were used: (1) text messaging the users for information on cardiovascular health, (2) self-monitoring of 6 heart-healthy behaviors to fill out the blanks of behavioral items, and (3) automated and tailored feedback messaging to users for behavioral outcomes obtained from self-monitoring.

Objectives: This study aimed to evaluate the feasibility and acceptability of the "My HeartHELP" app.

Methods: The participants were 29 community residents in Seoul, South Korea, who met at least 1 criterion of metabolic syndrome. To evaluate the feasibility, we assessed 3 records, which are as follows: First, the "record for self-monitoring" was determined as feasible if an average percentage for each of the 6 behaviors over 4 weeks was 75% or higher based on percentages of participants who completed to record each of 6 heart-healthy behaviors. Second, the "record for access to the app" was determined as feasible if users accessed at least once a day on average per week. Third, "records for behavioral changes" over 4 weeks were collected via a self-reported questionnaire. To evaluate acceptability, we used an assessment tool comprising 12 items that included subscales for comprehensibility, ease, health benefits, technical completeness, overall satisfaction, and recommendation to others on a 5-point Likert scale. Acceptability was determined as acceptable if the average scores for the total scale and each subscale were 3.5 points or greater. Second, qualitative data were collected through 2 focus groups, each consisting of 14 or 15 participants. All data were collected in June and July 2022.

Results: During the 4 weeks, 95.6% (range: 85.8%-97.4%) of the participants adhered to more than 75% of "completion of daily self-monitoring of each heart-healthy behavior," having met the criterion. The participants accessed the app on average 1.8 (SD 1.70) times per day, meeting the criteria. Participants had positive behavioral changes in all 6 behaviors, of which nonsedentary behavior (10%-28%; χ^2_1=1.76; P<.001) and non–fast-food intake were especially statistically significant (72%-93%; χ^2_1=5.64; P=.03) over 4 weeks. Participants reported 3.8 points for a total score of acceptability and more than 3.5 points for all subscales, which met the criterion. Qualitative data obtained from focus groups indicated that automated and tailored feedback messages motivated participants to promote healthy lifestyles.

Conclusions: The "My HeartHELP" app may be a feasible and acceptable mobile app to promote self-monitoring and possibly behavioral changes in heart-healthy lifestyle behaviors.

(JMIR Form Res 2025;9:e66108) doi:10.2196/66108

KEYWORDS

healthy lifestyle; primary prevention; cardiovascular diseases; community health nursing; mobile app

Introduction

Cardiovascular disease is preventable through the promotion of heart-healthy behaviors, such as physical activity, nonsmoking, nonsedentary, and healthy eating behaviors, refraining from alcohol binge drinking, and self-assessment of body weight [1,2]. Each additional behavioral modification may reduce the risk of cardiovascular disease [1,3]. However, only a small percentage of the population is known to adhere to these

XSL•FO
RenderX

behaviors [4,5]. In this context, strategies are needed to effectively change multiple heart-healthy behaviors that are accessible to everyone, even before the onset of heart disease, and in community settings.

An app of mobile health technologies that allow community-dwelling individuals to monitor their own health behaviors and promote health behaviors in their daily lives may be needed to effectively change heart-healthy behaviors [6]. Mobile apps have the advantages of self-directedness, instant connectivity, ubiquity, personalization, and interactive learning through social engagement [7], which may enhance self-monitoring and practice of health behaviors without time and space constraints. However, few mobile apps are equipped with strategies for a 1-stop approach to monitor multiple heart-healthy behaviors that may correlate with each other [8] and target individuals without cardiovascular disease in a community setting. However, evidence indicating that mobile health technologies can effectively improve heart-healthy behaviors is lacking. This may be due to a lack of empathy and ineffective educational effects through active and persistent interactions with cardiovascular experts [9]. In this context, mobile technologies should be designed to enhance the self-monitoring of their own behaviors and interactions between mobile app users and experts in the context of cardiovascular prevention and promotion.

Several mobile apps have been developed to monitor health behavior changes [10,11] in the context of disease management for hypertension, diabetes, cardiovascular disease, cancer, and obesity [12-15]. However, the monitoring strategies used in each study were inconsistent [16]. In addition, the design of such mobile apps for health behavioral interventions has been confined to the process of self-reporting individual health behaviors in the form of a checklist [17]. Finally, few health-promoting mobile apps have automated feedback text messages for individual behavioral outcomes achieved with the self-monitoring of multiple lifestyle behaviors in the context of cardiovascular health.

The Information-Motivation-Behavioral Skills model (IMB model) is widely used as a theoretical framework for predicting health behavior changes [18]. The IMB model addresses the core constructs of information, motivation, and behavioral skills as prerequisites for health behavior change and assumes that when an individual acquires sufficient information related to health behavior, is motivated to change behavior, and improves behavioral skills, their behavior change and maintenance are promoted, which in turn can improve health levels [18]. In this theoretical background, we designed a mobile app to be embedded with behavioral strategies to encourage individuals to be informed, motivated, and skilled in behavior change by using cognitive-behavioral strategies addressed by the American Heart Association [19].

One such app is the “My HeartHELP” app, which was developed for the general adult population to promote heart-healthy lifestyle behaviors based on the constructs of information, motivation, and behavioral skill in the IMB model [20]. The “My HeartHELP” app is an abbreviation derived from the Heart-Healthy Lifestyle-Promoting app [21]. This app targets 6 heart-healthy lifestyle behaviors, such as physical activity, nonsedentary behaviors, healthy eating behaviors, nonsmoking, refraining from alcohol binge drinking, and self-assessment of body weight [22,23]. Of the 6 behaviors, 5 were selected based on the rationale that the US Centers for Disease Control and Prevention emphasizes key preventive behaviors for heart health, including healthy food and beverage choices, maintaining a healthy weight, engaging in regular physical activity, and nonsmoking [23]. An additional behavior, nonsedentary behavior, was incorporated due to the established association between sedentary behavior and increased cardiovascular risk [1]. The “My HeartHELP” app was designed with three behavioral strategies : (1) text messaging users for information on cardiovascular health, (2) self-monitoring of 6 lifestyles, and (3) automated and tailored feedback text messaging to users for behavioral outcomes obtained from self-monitoring. To increase the construct of information in the IMB model, texts on knowledge about cardiovascular prevention [24] was provided via text messaging. To increase the constructs of motivation and behavioral skills in the IMB model, goal setting [25,26], self-reflection via self-monitoring of behaviors [27], and positive reinforcement and tailored real-time feedback of behavioral performance via text messages [28,29] were installed in the “My HeartHELP” app. In particular, providing feedback is a well-established strategy in behavioral interventions. A Cochrane review underscored its importance in facilitating behavior change, particularly in contexts where substantial improvement is needed [30]. This aligns with the American Heart Association’s cognitive-behavioral strategies, which also emphasize feedback as a key component in sustaining behavioral improvements [19].

In this context, we aimed to assess the feasibility and acceptability of the “My HeartHELP” app in this study to conduct a future randomized controlled trial (ISRCTN83643383).

Methods

Study Design and Sample

This study used a mixed methods design [31] with a single-group approach to evaluate the feasibility and acceptability of the “My HeartHELP” app using both quantitative and qualitative data. This study design was chosen to incorporate qualitative data, in addition to quantitative data, to facilitate a more comprehensive evaluation of app users’ perceptions of its acceptability, including its health benefits and overall usefulness. The app was implemented over a 4-week period. Data were collected at both pretest and posttest stages. At each stage, a questionnaire was administered, and at the posttest stage, focus group discussions were conducted.

The participants were 29 adults. They were recruited by the following ways: First, we posted a recruitment notice on the bulletin boards of the university affiliated with the researchers. Second, we additionally posted it on a community-based mobile commercial and life platform (ie, Karrot app). In this way, the Karrot automatically displays the notice to residents in 174 neighborhoods across 5 municipal districts surrounding the university in Seoul, South Korea. Individuals interested in

XSL•FO
RenderX

participating in the study contacted the researchers via mobile communication platforms (eg, the KakaoTalk app) or by phone, after which an eligibility screening was conducted based on the following criteria. The inclusion criteria were adults (1) aged between 20 and 64 years who had at least one of the metabolic syndrome components (ie, abdominal obesity, hypertension, dyslipidemia, or glucose intolerance). The exclusion criteria were individuals diagnosed with (1) cardiovascular disease and undergoing treatment, (2) major depression or anxiety disorder, (3) physical impairment, (4) cognitive impairment, and (5) difficulties in using a mobile app.

The sample size was determined to be 30 adults based on a study by Symer et al [32] that tested the validity of a mobile app. Of the 30 participants, 1 lost contact during the study period of 4 weeks and 29 were included in the final analysis of quantitative and qualitative data.

Overview of the "My HeartHELP" App

The "My HeartHELP" app was designed with 3 behavioral strategies, that is, behavioral information, behavioral self-monitoring, and behavioral performance/feedback. The structure and content of the app are described in detail in Figure 1 and Table 1.

Figure 1. Overview of the "My HeartHELP" app. This mobile app was developed in Korean and designed based on 3 behavioral strategies: behavioral information, behavioral self-monitoring, and behavioral performance/feedback. In this study, the app was applied over a 4-week period to 29 participants residing in Seoul, South Korea, who were at risk for cardiovascular disease. Pre- and posttests were performed to evaluate the feasibility and acceptability of the app.

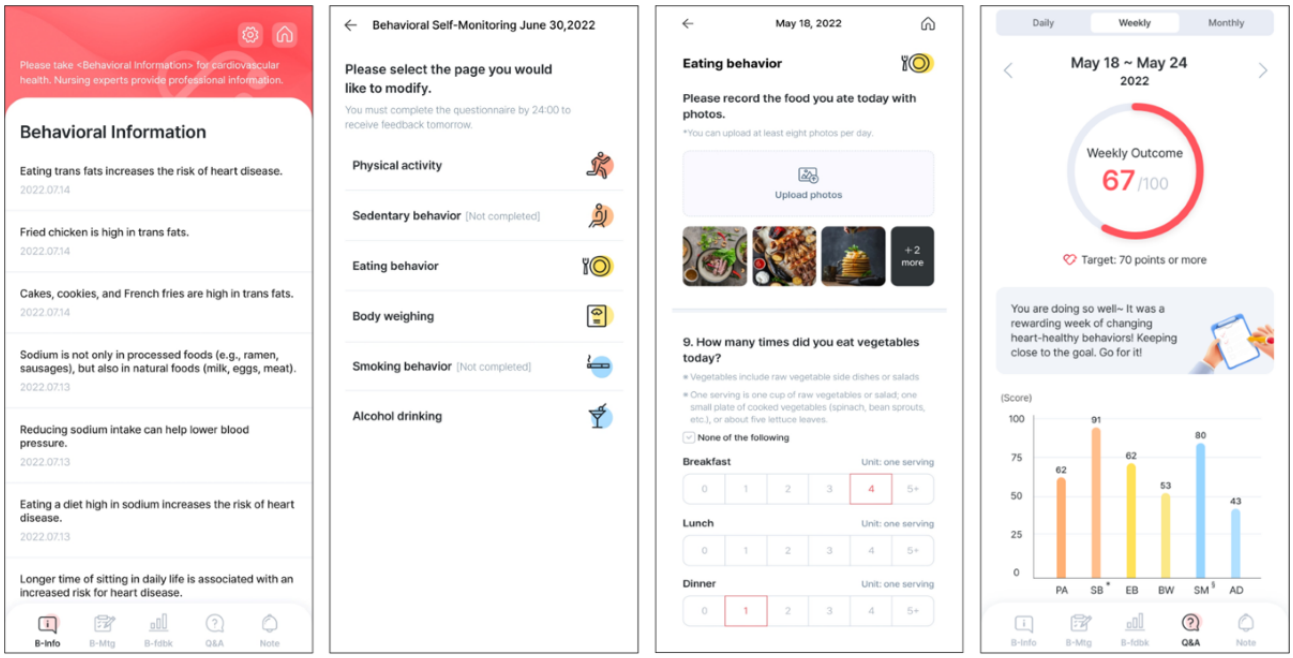

XSL•FO
RenderX

Table . The structure of the "My HeartHELP" app: its strategies and contents.

Strategies	Contents
Behavioral information	• Behavioral knowledge • A total of 43 text messages for symptoms and characteristics of cardiovascular disease, cardiovascular risk factors, and heart-healthy behaviors • 8 - 11 texts delivered per week to participants • From Monday to Friday at 1:00 PM over 4 weeks
Behavioral self-monitoring	• Ensuring that the participants recorded the extent to which they practiced each of the 6 heart-healthy behaviors during the entire day • Physical activity (steps of walking and types and minutes of moderate-intensity exercise) • Sedentary behavior (time, hours) • Eating behavior (fruit and vegetable, sugar-sweetened beverage, fast foods, foods high in saturated fat or trans fat, whole grain, and sodium intake) • Self-measurement of body weight (kg per day) • Smoking behavior (cigarettes per day) • Alcohol-drinking behavior (drinks per day)
Behavioral performance/feedback	• Interface displays according to the level of heart-healthy behavior achieved over the entire day, 7 days (a week), and 30 days (a month) • Based on the algorithm for behavioral target goals that the authors had installed for each of 6 heart-healthy behaviors • Automated and tailored feedback text messaging • Daily at 9:30 AM from Monday to Friday • Weekly at 9:30 AM on Monday • Monthly at 9:30 AM on the first Tuesday • Example: "You did a good job today" or "You almost reached your goal today. Just a little more effort, and you will reach your goal. Cheer up!"

Behavioral Information

"Behavioral information" refers to text messaging of heart-health–related information on the characteristics of cardiovascular diseases and heart-health–related behaviors (Figure 1 and Table 1). Text messages were delivered as push notifications 5 times per week at 1:00 PM over a course of 4 weeks.

Behavioral Self-Monitoring

"Behavioral self-monitoring" refers to the participants' recording the extent to which they achieved 6 heart-healthy behaviors (Figure 1 and Table 1). Physical activity was recorded into 2 types: daily step counts for walking and daily time duration for moderate-intensity aerobic exercise; sedentary behavior recorded screen time (hours); and eating behavior recorded frequencies and types of whole grain intake, fruit and vegetable intake, sugar-sweetened beverage intake, sodium intake, fast-food intake, and intake of foods high in saturated fat. The participants were instructed to weigh body weights to measure and record. Participants were asked to daily record whether they smoked cigarettes and, if so, how many cigarettes they smoked. The participants were requested to record the frequency of alcohol consumption with a standard glass [33,34].

Behavioral Performance/Feedback

"Behavioral performance/feedback" refers to automatized and tailored interface displays including diagram and texts (Figure 1 and Table 1) according to outcomes adhered to 6 heart-healthy behaviors achieved over 1 day, 7 days (a week), and 30 days (a month), based on the algorithm for behavioral target goals that the authors had installed for the 6 heart-healthy behaviors.

The frequency of feedback delivery was structured as daily, weekly, and monthly, based on the rationale that frequent feedback may enhance motivation, self-monitoring, and goal attainment for behavioral change [35], and that daily, weekly, and monthly self-monitoring may support positive health outcomes [35-37]. Push notifications were delivered to request access to the interface display.

Daily outcomes are displayed as achieved scores out of 10 points (eg, 6/10) (Figure 1), weekly outcomes as achieved scores out of 100 points and bar graphs per behavior (Figure 1), and monthly outcomes as achieved scores out of 100 points and the symbol of the heart if achieved. Specifically, the daily outcomes were obtained by achieved scores from the components of 6 behaviors. Physical activity was evaluated by awarding 1 point for achieving a minimum of 8000 steps per day [38] or engaging in at least 30 minutes of moderate-intensity exercise per day [39]. Nonsedentary behavior was assessed by awarding 1 point for limiting sedentary time to less than 2 hours per day [40]. Eating behaviors were evaluated across multiple dietary components. Fruit and vegetable intake was assessed by awarding 1 point for consuming a minimum of 5 servings per day [41]. Similarly, 1 point was awarded for abstaining from sugar-sweetened beverages [42], avoiding fast-food consumption [43], refraining from consuming foods high in trans fats [44], and maintaining sodium intake below 1500 mg per day [45]. Self-monitoring of body weight was assessed by awarding 1 point for weighing oneself daily [46,47]. Smoking behavior was

XSL•FO
RenderX

evaluated by awarding 1 point for complete abstinence from smoking [48]. Finally, alcohol consumption was assessed by awarding 1 point for limiting intake to fewer than 5 drinks per day for men and fewer than 4 drinks per day for women [49].

Weekly outcomes were derived from the average scores obtained over 7 days, which were then converted to a 100-point scale and categorized into 3 ranges: ≥70, 50 - 69, and <50. Feedback texts were generated based on these score ranges. Similarly, a monthly outcome was calculated from the average scores over the previous 4 weeks, converted to a 100-point scale, and classified into 2 categories: ≥70 and <70. Corresponding feedback texts were generated accordingly.

To account for users' baseline characteristics and minimize redundant data entry, participants who self-identified as nonsmokers and nondrinkers were provided with a tailored interface that excluded smoking- and alcohol-related inputs. In addition, they received automated feedback reinforcing their adherence to these behaviors.

Study Procedure and Data Collection

The study procedure and data collection were conducted in 3 sequential stages between June and July 2022 in Seoul, South Korea. Specifically, the pretest was conducted on June 16 - 17 (stage 1), followed by the 4-week implementation of the "My HeartHELP" app from June 20 to July 17 (stage 2). The posttest and focus group sessions took place on July 21 - 22 (stage 3).

Stage 1: The pretest took place on the same day using a self-reported questionnaire on general characteristics and heart-healthy behaviors. After the pretest, the pre-education session for informing the usage of the "My HeartHELP" app was conducted as a face-to-face group session. At this session, the principal investigator (JC) presented a lecture to inform the characteristics of heart disease and the importance of promoting heart-healthy behaviors for 20 minutes. This lecture aimed to provide background knowledge on the variables measured by the mobile app and to facilitate understanding of the app's usage. Following the lecture, the researchers provided a detailed explanation of the structure and usage of the "My HeartHELP" app, along with an overview of the overall research process. Stage 2: The implementation of the "My HeartHELP" app took place for 4 weeks without contact with the researchers in between. Stage 3: Postsurvey and focus group session took place on the same day after the end of the 4-week implementation. The postsurvey session was administered using the self-reported questionnaire with heart-healthy behaviors (ie, repeated measures) and the acceptability of the app. Focus group sessions, each lasting 60 minutes, were conducted with 2 groups comprising of 14 - 15 participants per group to assess the acceptability of "My HeartHELP" app. The principal investigator (JC) facilitated the discussion by using open-ended questions, probing questions, and laddering techniques [50] to encourage participants to freely express their thoughts and provide in-depth feedback. Key topic areas included overall impressions, strengths and weaknesses, comprehensibility, perceived health benefits, and usability (ie, ease of use and convenience in navigation and operation). In addition, challenges such as system errors, data input difficulties, and technical issues were explored. Participants also suggested improvements related to system errors, data input difficulties, and readability concerns. All sessions were audio-recorded and transcribed verbatim for analysis.

Study Variables and Metrics

General characteristics included sex, age, college-educated status, employment status, exposure to heart health education, history of chronic diseases (ie, hypertension, diabetes mellitus, or dyslipidemia), and self-reported height and body weight.

The Feasibility of "My HeartHELP"

To evaluate the feasibility of "My HeartHELP," we assessed the following 3 behavioral records: First, the "record for self-monitoring" was determined based on the data of the percentage of participants who completed recording each of 11 behavioral components (Figure 1) of the 6 heart-healthy behaviors per day, and their average percentage over 4 weeks was assessed. We determined that this was feasible if the average percentage was 75% or higher [32]. Second, the "record for access to the app" was determined based on the data of a mean of daily access counts per week. Next, we determined whether it was feasible for the users to access the app at least once a day on average per week. Third, the "record for behavioral changes" was determined as feasible if participants showed improvements in each of the 6 behaviors when comparing pretest and posttest. Among the 6 behaviors, the following behaviors were included in the analysis : Physical activity (2 items) was assessed as moderate-intensity aerobic exercise for 150 minutes per week [39]. Walking (2 item) was also assessed for 150 minutes per week [38]. Nonsedentary behavior (1 item) was assessed daily, defined as having less than 2 hours of screen time (screen time included activities such as watching TV and using a mobile phone, computer, or tablet, excluding work- or study-related activities) [40]. The intake of fruits and vegetables (1 item) was assessed as 2 or more servings per day [41]. Non—sugar-sweetened beverage intake (1 item) was assessed none per day [42]. Non—fast-food intake (1 item) was assessed as consumption less than 2 times per week [43]. Whole grain intake (1 item) was assessed as consuming 2 or more servings per day, where 1 serving equals 1 slice of whole grain bread [51]. High-risk alcohol consumption (2 items) was defined as 5 or more drinks per session for men (including 4 or more drinks per session for women) [49]. Maintaining healthy body weight (1 item) was assessed by having participants report their current weight.

The Acceptability of "My HeartHELP"

To evaluate the acceptability of "My HeartHELP," we used quantitative data from an assessment tool and qualitative data from focus groups. First, we developed the assessment tool with 12 items that consisted of the following subscales: comprehensibility (2 items), ease (2 items), health benefits (5 items), technical completeness (1 item), overall satisfaction (1 item), and recommendation to others (1 item). Health benefits were categorized into "behavioral information," "behavioral monitoring," and "behavioral performance/feedback." The items were scored on a 5-point Likert scale (1=not at all, 2=quite not, 3=somewhat, and 4=very much). The subscales were considered acceptable if the mean score for each subscale was 3.5 points

XSL•FO
RenderX

or greater. Second, qualitative data obtained from focus groups were used. The following questions [52] were addressed: "What is your overall impression of the My HeartHELP?" "Please tell us three pros and three cons of the My HeartHELP." "Please tell us about My HeartHELP based on (1) comprehensibility (2) health benefits (3) usability (4) challenges." The qualitative data were considered acceptable if researchers identified meaningful quotations highlighting the benefits and usefulness of the "My HeartHELP" app in facilitating positive heart-healthy behavioral changes and improving health outcomes. While the quantitative data provided structured acceptability scores, they were limited in capturing the depth of users' perceptions and the specific challenges encountered during app usage. To address this limitation, qualitative findings from focus group discussions were integrated, offering a more nuanced understanding of user perceptions regarding comprehensibility, perceived benefits, and challenges. This approach enabled a more comprehensive evaluation that extends beyond numerical assessments alone.

Data Analysis

SPSS/WIN (version 28.0; SPSS, Inc) was used [53]. The significance level was set at less than .05. Participants' baseline measures (ie, general characteristics and heart-healthy behaviors) were analyzed using descriptive statistics (frequencies, percentages, means, and SDs). To compare the changes in heart-healthy behaviors between pre- and postsurvey, McNemar's chi-square test for dichotomous variables and the 2-tailed paired *t* test for continuous variables were conducted. The measures of feasibility and acceptability were analyzed using descriptive statistics for the quantitative data and content analysis for the qualitative data. Qualitative data obtained from focus groups were analyzed using Graneheim and Lundman's qualitative content analysis [54]. The interview recordings were repeatedly reviewed, transcribed, and thoroughly examined to capture participants' perspectives comprehensively. The transcribed data were then segmented into meaningful units, systematically coded, and organized into categories and subcategories to ensure a structured interpretation. Continuous comparisons were conducted across data points to refine themes and establish thematic classifications. The 2 sets of qualitative data were independently analyzed by 3 authors (JC, SN, and YS) and consolidated into meaningful themes and categories [55,56].

Ethical Considerations

This study was approved by the institutional review board of Korea University (KUIRB-2022-0147-01). All participants provided written informed consent for their participation in the study, including permission for audio recording during the focus group discussions. All procedures were performed in accordance with the ethical standards of the institutional research committee and the 2013 Declaration of Helsinki. To ensure the privacy and confidentiality of participants, all data were deidentified and securely stored on encrypted servers, with access restricted to authorized members of the research team. Participants who successfully completed all stages of the study received a monetary compensation of KR ₩200,000 (US $137) [57]. Participants who completed only the pretest received a KR ₩10,000 (US $7) gift voucher. All details related to compensation were fully disclosed to participants during the informed consent process, prior to their enrollment in the study.

Results

Participant Characteristics

The participants (N=29) had a mean age of 41.1 (SD 13.03) years (Table 2). Of all the participants, 62% (18/29) were women; 79% (23/29) were college-educated; 66% (19/29) were employed; 45% (13/29) had a previous history of hypertension, diabetes mellitus, or dyslipidemia; and 69% (20/29) fell into the category of more than 25 kg/m^2 of BMI. Moreover, 7% (2/29) of the participants had undergone a heart-health educational session.

XSL•FO
RenderX

Table . General characteristics of participants (N=29).

Participant characteristics	Values
Age (years), mean (SD)	41.1 (13.03)
Gender, n (%)	
Women	18 (62)
Men	11 (38)
College-educated, n (%)	
Yes	23 (79)
No	6 (21)
Employed, n (%)	
Yes	19 (66)
No	10 (34)
Having diseases[a], n (%)	
Yes	13 (45)
No	16 (55)
BMI, kg/m^2, n (%)	
≥25	20 (69)
<25	9 (31)
Exposed to heart-health educational sessions, n (%)	
Yes	2 (7)
No	27 (93)

[a]Having diseases indicate hypertension, diabetes mellitus, or dyslipidemia.

The Feasibility of "My HeartHELP"

The feasibility for the "record of self-monitoring" was determined to be feasible. The record of self-monitoring for 11 behavioral components of the 6 heart-healthy lifestyle behaviors is shown in Multimedia Appendix 1. The 4-week percentages of participants who adhered to more than 75% of the completion of daily self-monitoring for each heart-healthy behavior are reported in Multimedia Appendix 1. The average percentage was 95.6%, ranging from a minimum of 85.8% (recording body weight) to a maximum of 97.4% (recording physical activity).

The feasibility for "record of access to the app" was determined as feasible. Mean daily access counts per week showed 2.2 (SD 2.10) at week 1, 1.7 (SD 1.58) at week 2, 1.6 (SD 1.34) at week 3, and 1.6 (SD 1.62) at week 4. The grand mean was 1.8 (SD 1.70), which exceeded the feasibility criteria (Figure 2).

XSL•FO
RenderX

Figure 2. Feasibility of the "My HeartHELP" app: behavioral record of access to the app. The box plot represents the mean values of daily app access counts per week over a 4-week period based on data from 29 participants at risk for cardiovascular disease residing in Seoul, South Korea. The feasibility criterion was set at an average of at least 1 access per week.

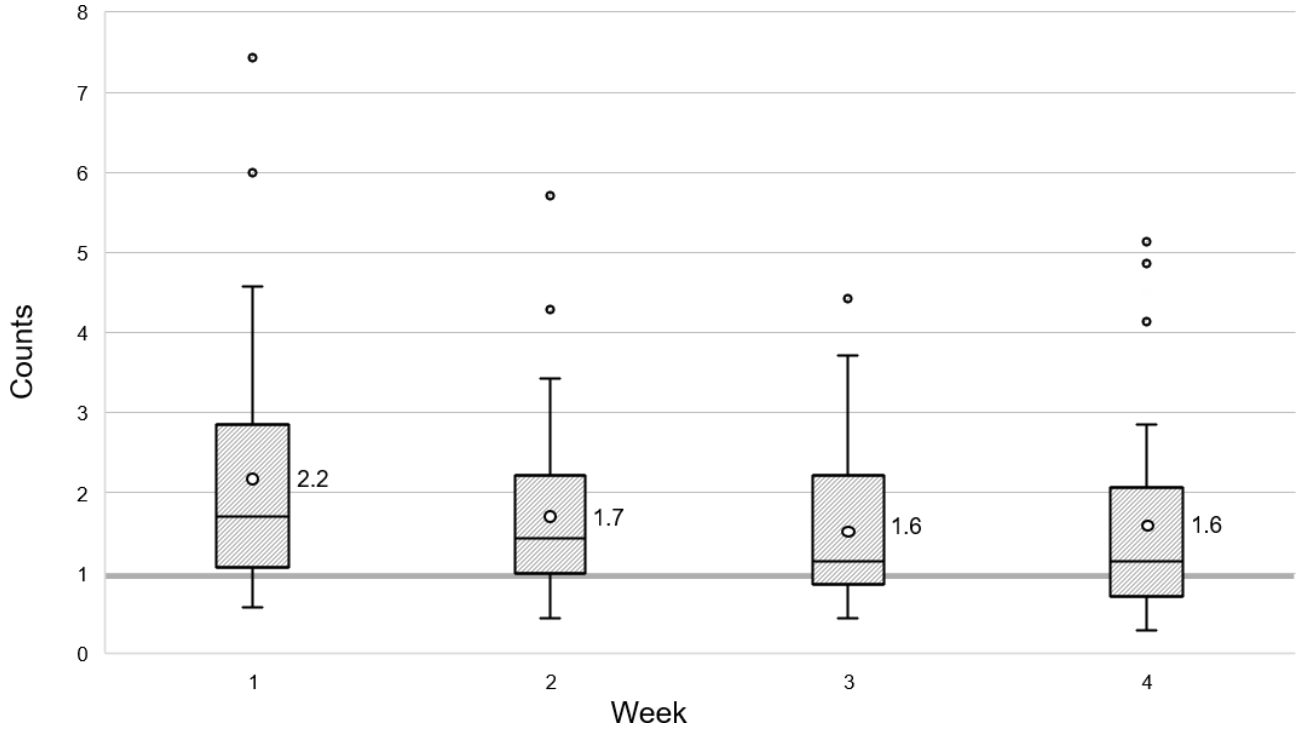

The feasibility for "record for behavioral changes" is shown in Table 3. Numbers of participants for physical activity increased from 24% (7/29) to 34% (10/29) for sufficient moderate-intensity aerobic exercise, from 59% (17/29) to 62% (18/29) for sufficient walking exercise, and from 10% (3/29) to 28% (8/29) for nonsedentary behavior (χ^2_1=1.76; *P*<.001). Regarding eating behavior, numbers of participants increased from 7% (2/29) to 21% (6/29) for intake of fruits and vegetables, from 86% (25/29) to 90% (26/29) for the intake of non–sugar-sweetened beverages, from 72% (21/29) to 93% (27/29) for the intake of non–fast foods (χ^2_1=5.64; *P*=.03), and from 62% (18/29) to 66% (19/29) for intake of whole grains. The number of high-risk alcohol drinkers decreased from 38% (11/29) to 34% (10/29). Finally, body weight changed from 76.2 kg to 76.0 kg.

XSL•FO
RenderX

Table . Changes in heart-healthy behaviors after the application of "My HeartHELP" (N=29).

Characteristics	Pretest	Posttest	Chi-square or *t* test (*df*)	*P* value
Physical activity				
Sufficient moderate-intensity aerobic exercise, n (%)			10.72 (1)[a]	.38
Yes	7 (24)	10 (34)		
No	22 (76)	19 (66)		
Sufficient walking exercise (≥150 minutes), n (%)			14.78 (1)[a]	.99
Yes	17 (59)	18 (62)		
No	12 (41)	11 (38)		
Nonsedentary behavior (<2 hours/day), n (%)			1.76 (1)[a]	<.001
Yes	3 (10)	8 (28)		
No	26 (90)	21 (72)		
Fruit and vegetable intake (≥2 servings/day), n (%)			1.13 (1)[a]	.22
Yes	2 (7)	6 (21)		
No	27 (93)	23 (79)		
Non–sugar-sweetened beverage intake (less than once per day), n (%)			7.87 (1)[a]	.99
Yes	25 (86)	26 (90)		
No	4 (14)	3 (10)		
Non–fast foods (<2 servings/week), n (%)			5.64 (1)[a]	.03
Yes	21 (72)	27 (93)		
No	8 (28)	2 (7)		
Whole grains (≥2 servings/day), n (%)			11.47 (1)[a]	.99
Yes	18 (62)	19 (66)		
No	11 (38)	10 (34)		
High-risk alcohol drinking[b], n (%)			24.98 (1)[a]	.99
Yes	11 (38)	10 (34)		
No	18 (62)	19 (66)		
Body weight (kg), mean (SD)	76.2 (15.36)	76.0 (15.21)	1.22 (28)[c]	.23

[a]Chi-square test.

[b]High-risk drinking indicates ≥5 drinks per session for men (≥4 drinks per session for women).

[c]*t* test.

The Acceptability of "My HeartHELP"

Acceptability using quantitative data was determined as acceptable if the scores for all the subscales of acceptability were 3.5 points (the criteria of acceptability) or greater (Table 4). The scores ranged from a minimum of 3.6 points (health benefits for information, health benefits for performance/feedback, and technical completeness) to a maximum of 4.1 points (comprehensibility).

XSL•FO
RenderX

Table . Evaluation of the acceptability for the My HeartHELP (N=29).

	Values, mean (SD)
Total score of the acceptability	3.8 (0.61)
Comprehensibility	4.1 (0.81)
Ease	3.9 (0.84)
Health benefits	
Overall	3.8 (0.87)
Behavioral information	3.6 (1.02)
Behavioral monitoring	3.8 (0.90)
Behavioral performance/feedback	3.6 (0.70)
Completeness, technical	3.6 (0.87)
Satisfaction, overall	3.8 (0.73)
Recommendation to others	3.8 (0.87)

The acceptability using the qualitative data may be determined as acceptable, as shown in Table 5. Participants mentioned that reflecting on their own heart-healthy behaviors via self-monitoring, adjusting their own heart-healthy behaviors via receiving feedback messages, and becoming motivated for weight control were health benefits. The participants mentioned that easily understood phrases were acceptable in terms of the use of easy language in the app. However, the participants suggested the following challenges for the structure and interface: differentiating between good and bad behaviors by color, including options for a variety of dietary behavior options, and installing additional reminder alarms (Table 5).

XSL•FO
RenderX

Table . Themes and quotations extracted using focus group discussion (N=29).

Themes and categories	Quotations
Comprehensibility	
Understanding phrases easily	"Overall, I think it was easy to understand. The language used was everyday language that we encounter in our daily lives, so it was not difficult."
Benefits	
Reflecting on their own heart-healthy behaviors via self-monitoring	"I habitually consume large yogurt drinks several times a day. However, as I continued to record them under the sweetened beverage category, I realized that just two of these large yogurts exceeded my daily sugar intake limit, leading me to stop consuming them."
	"I used to drink alcohol quite often even when dining alone, but after seeing the records I entered, I found myself not ordering alcohol even when getting delivery food."
Adjusting their own heart-healthy behaviors via receiving feedback messages	"When I woke up each day, I first checked the daily achievement scores in 'My HeartHELP' After that, I started paying more attention to what I ate and definitely did more exercise throughout the day."
Getting motivated for weight control	"I'm maintaining my 3 kg weight loss because I have been using it continuously to motivate me to lose weight."
Challenges	
Differentiating good and bad behaviors by colors	"It would be good to distinguish unhealthy foods, like salty ones, by marking them with red or orange, and to mark healthy foods with yellow."
Including options of a variety of dietary behavior options	"I think it would have been better if I could enter other foods not listed. Or, it would be nice if there was an explanation that allowed me to include foods not presented based on my own judgment."
Installing additional reminder alarms	"I think it would be good if push notifications could come several times a day. Feedback messages arrive during the day, and I want to enter them all at once in the evening, but sometimes I forget and it is already past midnight."

Discussion

Principal Findings

We found that the "My HeartHELP" app may be feasible and acceptable for individuals at risk of cardiovascular disease to be informed of their cardiovascular health, to self-monitor their heart-healthy behaviors, and to adhere to heart-healthy behaviors. Moreover, our data revealed that positive changes in heart-healthy behaviors may be achieved using the "My HeartHELP" app when included as a strategy for heart-healthy behavioral lifestyle interventions.

The "My HeartHELP" app met the criteria of feasibility and acceptability for the validity test of a mobile app. The vast majority of participants (95.6%) completed daily self-monitoring of their heart-healthy behaviors and accessed the app more than once per day (1.8 times). Moreover, the participants perceived the "My HeartHELP" app as comprehensible, easy to use, beneficial to their health, highly usable, and technically complete. Overall, the participants were satisfied with the My HeartHELP app and agreed to recommend it to others. Meanwhile, monitoring one's own behavior may be directly linked to adherence to healthy behavior [58,59] through the behavioral actions of participants, such as opening the app and entering their daily behavioral outcome for the items. Logging food in real time on mobile devices appears to encourage individuals to become more aware of their food consumption, which may have supported their efforts to achieve dietary goals and weight loss [60,61]. In this context, the action of accessing the app itself may induce motivation to monitor one's own behavior, which can lead to self-awareness of one's own behavior and thus to desirable heart-healthy behavior [61].

Our findings revealed that heart-healthy behaviors changed positively over 4 weeks, although significant changes were not verified for all 6 behaviors. Mobile health interventions have

XSL•FO
RenderX

revealed inconsistent results for improving heart-healthy behaviors and have reported positive effects limited to a single behavior, that is, physical activity. In this regard, the "My HeartHELP" app is equipped with behavioral strategies that allow individuals to monitor and practice complex and multiple heart-healthy behaviors using a 1-stop approach. Six heart-healthy behaviors via the app were positively changed over the course of 4 weeks, with the limitation of a short duration. These positive results may be explained by installing behavioral strategies for promoting motivation and self-efficacy based on the IMB model [18], that is, goal setting, self-monitoring, and feedback provision. A systematic review reported that behavioral counseling had a small to moderate effect on behavioral interventions when tailored to advice, goal setting, and written material provision [62]. Therefore, we can determine the feasibility and acceptability of the "My HeartHELP" app as a critical strategy for heart-healthy behavioral lifestyle interventions.

We found that the health benefits scores (behavioral information and performance/feedback) and technical completeness score in the acceptability presented the lowest score (3.6 points each), although the acceptability criteria were met (3.5 points). Linked to this finding, qualitative data from the focus group revealed that the participants had requested an additional alarm that could remind them to practice heart-healthy behaviors. Moreover, direct interaction with experts is necessary beyond self-monitoring and the performance interface of healthy behavioral performance [63]. In the future, feedback text messaging in which experts can provide sympathy for and support individual behavioral performance is needed to initiate intervention studies. Finally, the participants mentioned that the information messages were perceived as not individualized to an individual's lifestyle and situation. This finding indicates that the use of mHealth alone may have limitations in providing information tailored to an individual's situation and that face-to-face education is needed to provide information tailored to specific individual needs.

Limitations

This study outlined a few limitations. First, the generalizability of the results to ethnic groups other than Koreans is limited. In addition, to apply it to more people, the app developed in Korean needs to be translated and developed in other languages for linguistic inclusiveness. Second, because it is based on the diet of Koreans, in terms of monitoring eating behavior, its scope of use may be limited. However, because the behavioral strategies of "My HeartHELP" are key skills for behavioral change, the application of "My HeartHELP" to other languages may be necessary. Third, although "My HeartHELP" establishes behavioral goal setting based on standard criteria for adults [38-43,49,51], future studies should further incorporate tailored goal-setting strategies that promote individuals' autonomy, as emphasized in self-determination theory. Enhancing autonomy in goal setting may foster intrinsic motivation, which is essential for sustaining long-term behavior change [29,64]. Fourth, the 4-week follow-up period may have been too short to capture sustained behavior changes, highlighting the need for longer validation studies. Finally, the use of focus group discussions instead of anonymous methods may introduce social desirability bias, as participants may adjust their feedback based on group dynamics rather than expressing independent perspectives [65].

Conclusions

"My HeartHELP," which includes information, self-monitoring, and real-time feedback messaging strategies that promote the practice of complex heart-healthy behaviors, could be feasible and acceptable as an app for promoting cardiovascular health and could contribute to achieving optimal heart health for populations at risk for cardiovascular disease in the community. Practically, community settings, unlike hospital settings, present challenges and limitations in providing intensive professional care for comprehensive heart-healthy lifestyle modification. Therefore, the utilization of the "My HeartHELP" app in health promotion programs aimed at cardiovascular disease prevention at community settings may offer a potential way to solve the limitations.

Acknowledgments

This research was supported by the National Research Foundation of Korea (NRF) grant from the Korean Government (NRF-2019R1A2C1004116) and NRF grant from the Ministry of Science and ICT, Korean Government (RS-2024 - 00336847).

Data Availability

The data analyzed in this study may be available from the corresponding author upon reasonable request.

Authors' Contributions

JC conceptualized and designed this study and prepared the original manuscript draft; JC, SN, and YS were involved in data collection; SN and YS performed the data curation and analyses; and JC provided expert consultation on data interpretation and helped draft the manuscript. All authors were involved in the review and revision of the manuscript and gave approval for the final version to be published.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
Feasibility of the "My HeartHELP" app: Record of self-monitoring for 11 behavioral components of the 6 heart-healthy lifestyle behaviors. A total of 29 participants at risk for cardiovascular disease residing in Seoul, South Korea, took part in the study. The feasibility of the self-monitoring record was assessed based on participant adherence, which was measured by calculating the percentage of participants who adhered to at least 75% of daily self-monitoring completion for each heart-healthy behavior over 4 weeks.
[PPTX File, 48 KB - formative_v9i1e66108_app1.pptx]

Checklist 1
Good reporting of a mixed-methods study (GRAMMS) checklist.
[DOCX File, 18 KB - formative_v9i1e66108_app2.docx]

References

1. Lavie CJ, Ozemek C, Carbone S, Katzmarzyk PT, Blair SN. Sedentary behavior, exercise, and cardiovascular health. Circ Res 2019 Mar;124(5):799-815. [doi: 10.1161/CIRCRESAHA.118.312669] [Medline: 30817262]
2. Spring B, Ockene JK, Gidding SS, et al. Better population health through behavior change in adults: a call to action. Circulation 2013 Nov 5;128(19):2169-2176. [doi: 10.1161/01.cir.0000435173.25936.e1] [Medline: 24100544]
3. Barbaresko J, Rienks J, Nöthlings U. Lifestyle indices and cardiovascular disease risk: a meta-analysis. Am J Prev Med 2018 Oct;55(4):555-564. [doi: 10.1016/j.amepre.2018.04.046] [Medline: 30241617]
4. Nielsen JB, Leppin A, Gyrd-Hansen DE, Jarbøl DE, Søndergaard J, Larsen PV. Barriers to lifestyle changes for prevention of cardiovascular disease - a survey among 40-60-year old Danes. BMC Cardiovasc Disord 2017 Sep 12;17(1):245. [doi: 10.1186/s12872-017-0677-0] [Medline: 28899356]
5. van Trier TJ, Mohammadnia N, Snaterse M, Peters RJG, Jørstad HT, Bax WA. Lifestyle management to prevent atherosclerotic cardiovascular disease: evidence and challenges. Neth Heart J 2022 Jan;30(1):3-14. [doi: 10.1007/s12471-021-01642-y] [Medline: 34762283]
6. Gold N, Yau A, Rigby B, Dyke C, Remfry EA, Chadborn T. Effectiveness of digital interventions for reducing behavioral risks of cardiovascular disease in nonclinical adult populations: systematic review of reviews. J Med Internet Res 2021 May 14;23(5):e19688. [doi: 10.2196/19688] [Medline: 33988126]
7. Goundar MS, Kumar BA. The use of mobile learning applications in higher education institutes. Educ Inf Technol 2022 Jan;27(1):1213-1236. [doi: 10.1007/s10639-021-10611-2]
8. Noh S, Choo J. Gender-specific clustering of lifestyle behaviors and its impacts on cardiovascular health. J Cardiovasc Nurs 2024;39(4):E115-E125. [doi: 10.1097/JCN.0000000000001006] [Medline: 37249529]
9. Svavarsdóttir MH, Sigurðardóttir Á, Steinsbekk A. How to become an expert educator: a qualitative study on the view of health professionals with experience in patient education. BMC Med Educ 2015 May 13;15(1):87. [doi: 10.1186/s12909-015-0370-x] [Medline: 25966846]
10. Frederix I, Van Driessche N, Hansen D, et al. Increasing the medium-term clinical benefits of hospital-based cardiac rehabilitation by physical activity telemonitoring in coronary artery disease patients. Eur J Prev Cardiol 2015 Feb;22(2):150-158. [doi: 10.1177/2047487313514018] [Medline: 24249840]
11. Burke LE, Sereika SM, Parmanto B, et al. Effect of tailored, daily feedback with lifestyle self-monitoring on weight loss: the SMARTER randomized clinical trial. Obesity (Silver Spring) 2022 Jan;30(1):75-84. [doi: 10.1002/oby.23321] [Medline: 34898011]
12. Kumar N, Khunger M, Gupta A, Garg N. A content analysis of smartphone-based applications for hypertension management. J Am Soc Hypertens 2015 Feb;9(2):130-136. [doi: 10.1016/j.jash.2014.12.001] [Medline: 25660364]
13. García-Gómez JM, de la Torre-Díez I, Vicente J, Robles M, López-Coronado M, Rodrigues JJ. Analysis of mobile health applications for a broad spectrum of consumers: a user experience approach. Health Informatics J 2014 Mar;20(1):74-84. [doi: 10.1177/1460458213479598]
14. Pandey A, Hasan S, Dubey D, Sarangi S. Smartphone apps as a source of cancer information: changing trends in health information-seeking behavior. J Cancer Educ 2013 Mar;28(1):138-142. [doi: 10.1007/s13187-012-0446-9] [Medline: 23275239]
15. Ubhi HK, Kotz D, Michie S, et al. Comparative analysis of smoking cessation smartphone applications available in 2012 versus 2014. Addict Behav 2016 Jul;58:175-181. [doi: 10.1016/j.addbeh.2016.02.026] [Medline: 26950256]
16. McKay FH, Cheng C, Wright A, Shill J, Stephens H, Uccellini M. Evaluating mobile phone applications for health behaviour change: a systematic review. J Telemed Telecare 2018 Jan;24(1):22-30. [doi: 10.1177/1357633X16673538] [Medline: 27760883]
17. Widmer RJ, Collins NM, Collins CS, West CP, Lerman LO, Lerman A. Digital health interventions for the prevention of cardiovascular disease: a systematic review and meta-analysis. Mayo Clin Proc 2015 Apr;90(4):469-480. [doi: 10.1016/j.mayocp.2014.12.026] [Medline: 25841251]

XSL•FO
RenderX

18. Fisher WA, Fisher JD, Harman J. The information-motivation-behavioral skills model: a general social psychological approach to understanding and promoting health behavior. In: Suls J, Wallston KA, editors. Social Psychological Foundations of Health and Illness: Blackwell Publishing; 2003:82-106. [doi: 10.1002/9780470753552.ch4]
19. Artinian NT, Fletcher GF, Mozaffarian D, et al. Interventions to promote physical activity and dietary lifestyle changes for cardiovascular risk factor reduction in adults: a scientific statement from the American Heart Association. Circulation 2010 Jul 27;122(4):406-441. [doi: 10.1161/CIR.0b013e3181e8edf1] [Medline: 20625115]
20. Choo J, Noh S, Shin Y. Abstract P640: evaluating feasibility and acceptability of “My HeartHELP” mobile application for promoting heart-healthy lifestyle behaviors. Circulation 2023 Feb 28;147(Suppl_1). [doi: 10.1161/circ.147.suppl_1.P640]
21. Choo J, Noh S, Shin Y. Abstract P147: a community-based, heart-healthy lifestyle-promoting program using hybrid behavioral strategies: study protocol of a three-group randomized controlled trial. Circulation 2023 Feb 28;147(Suppl 1). [doi: 10.1161/circ.147.suppl_1.P147]
22. Li Y, Pan A, Wang DD, et al. Impact of healthy lifestyle factors on life expectancies in the US population. Circulation 2018 Jul 24;138(4):345-355. [doi: 10.1161/CIRCULATIONAHA.117.032047] [Medline: 29712712]
23. Preventing heart disease. Centers for Disease Control and Prevention. 2024. URL: https://www.cdc.gov/heart-disease/prevention/index.html [accessed 2025-04-09]
24. Choo J, Noh S, Moon J, Shin Y. Development and psychometric testing of the Heart-Healthy Information Questionnaire. Eur J Cardiovasc Nurs 2023 Apr 12;22(3):299-310. [doi: 10.1093/eurjcn/zvac055] [Medline: 35766172]
25. Mann T, de Ridder D, Fujita K. Self-regulation of health behavior: social psychological approaches to goal setting and goal striving. Health Psychol 2013 May;32(5):487-498. [doi: 10.1037/a0028533] [Medline: 23646832]
26. Miller CK, Bauman J. Goal setting: an integral component of effective diabetes care. Curr Diab Rep 2014 Aug;14(8):509. [doi: 10.1007/s11892-014-0509-x] [Medline: 24907180]
27. Western MJ, Standage M, Peacock OJ, Nightingale T, Thompson D. Supporting behavior change in sedentary adults via real-time multidimensional physical activity feedback: mixed methods randomized controlled trial. JMIR Form Res 2022 Mar 2;6(3):e26525. [doi: 10.2196/26525] [Medline: 35234658]
28. Prestwich A, Kellar I, Parker R, et al. How can self-efficacy be increased? Meta-analysis of dietary interventions. Health Psychol Rev 2014;8(3):270-285. [doi: 10.1080/17437199.2013.813729] [Medline: 25053214]
29. Self-determination theory. Wikipedia. 2025. URL: https://en.wikipedia.org/wiki/Self-determination_theory [accessed 2025-04-09]
30. Ivers N, Jamtvedt G, Flottorp S, et al. Audit and feedback: effects on professional practice and healthcare outcomes. Cochrane Database Syst Rev 2012 Jun 13;2012(6):CD000259. [doi: 10.1002/14651858.CD000259.pub3] [Medline: 22696318]
31. Fetters MD, Curry LA, Creswell JW. Achieving integration in mixed methods designs-principles and practices. Health Serv Res 2013 Dec;48(6 Pt 2):2134-2156. [doi: 10.1111/1475-6773.12117] [Medline: 24279835]
32. Symer MM, Abelson JS, Milsom J, McClure B, Yeo HL. A mobile health application to track patients after gastrointestinal surgery: results from a pilot study. J Gastrointest Surg 2017 Sep;21(9):1500-1505. [doi: 10.1007/s11605-017-3482-2] [Medline: 28685388]
33. What is a standard drink? National Institute on Alcohol Abuse and Alcoholism. 2025. URL: https://www.niaaa.nih.gov/alcohols-effects-health/overview-alcohol-consumption/what-standard-drink [accessed 2025-04-09]
34. Guidelines for alcohol consumption in cancer prevention. National Cancer Center Korea. 2023. URL: https://www.cancer.go.kr/download.do?uuid=68379447-9a18-4ac0-a601-8e32ee918674.pdf [accessed 2025-04-09]
35. Ambeba EJ, Ye L, Sereika SM, et al. The use of mHealth to deliver tailored messages reduces reported energy and fat intake. J Cardiovasc Nurs 2015;30(1):35-43. [doi: 10.1097/JCN.0000000000000120] [Medline: 24434827]
36. Alshagrawi S, Abidi ST. Efficacy of an mHealth behavior change intervention for promoting physical activity in the workplace: randomized controlled trial. J Med Internet Res 2023 Apr 27;25:e44108. [doi: 10.2196/44108] [Medline: 37103981]
37. Elbert SP, Dijkstra A, Oenema A. A mobile phone app intervention targeting fruit and vegetable consumption: the efficacy of textual and auditory tailored health information tested in a randomized controlled trial. J Med Internet Res 2016 Jun 10;18(6):e147. [doi: 10.2196/jmir.5056] [Medline: 27287823]
38. Lee LL, Mulvaney CA, Wong YKY, Chan ES, Watson MC, Lin HH. Walking for hypertension. Cochrane Database Syst Rev 2021 Feb 24;2(2):CD008823. [doi: 10.1002/14651858.CD008823.pub2] [Medline: 33630309]
39. Jabbarzadeh Ganjeh B, Zeraattalab-Motlagh S, Jayedi A, et al. Effects of aerobic exercise on blood pressure in patients with hypertension: a systematic review and dose-response meta-analysis of randomized trials. Hypertens Res 2024 Feb;47(2):385-398. [doi: 10.1038/s41440-023-01467-9] [Medline: 37872373]
40. Choo J, Kim HJ, Park S. Neighborhood environments: links to health behaviors and obesity status in vulnerable children. West J Nurs Res 2017 Aug;39(8):1169-1191. [doi: 10.1177/0193945916670903] [Medline: 27753629]
41. Wang X, Ouyang Y, Liu J, et al. Fruit and vegetable consumption and mortality from all causes, cardiovascular disease, and cancer: systematic review and dose-response meta-analysis of prospective cohort studies. BMJ 2014 Jul 29;349:g4490. [doi: 10.1136/bmj.g4490] [Medline: 25073782]

XSL•FO
RenderX

42. Huang Y, Chen Z, Chen B, et al. Dietary sugar consumption and health: umbrella review. BMJ 2023 Apr 5;381:e071609. [doi: 10.1136/bmj-2022-071609] [Medline: 37019448]
43. Jiao J, Moudon AV, Kim SY, Hurvitz PM, Drewnowski A. Health implications of adults' eating at and living near fast food or quick service restaurants. Nutr Diabetes 2015 Jul 20;5(7):e171. [doi: 10.1038/nutd.2015.18] [Medline: 26192449]
44. Sacks FM, Lichtenstein AH, Wu JHY, et al. Dietary fats and cardiovascular disease: a presidential advisory from the American Heart Association. Circulation 2017 Jul 18;136(3):e1-e23. [doi: 10.1161/CIR.0000000000000510] [Medline: 28620111]
45. How much sodium should I eat per day? American Heart Association. 2024. URL: https://www.heart.org/en/healthy-living/healthy-eating/eat-smart/sodium/how-much-sodium-should-i-eat-per-day [accessed 2025-04-09]
46. Sakane N, Suganuma A, Domichi M, et al. The effect of a mHealth App (KENPO-app) for specific health guidance on weight changes in adults with obesity and hypertension: pilot randomized controlled trial. JMIR Mhealth Uhealth 2023 Apr 12;11:e43236. [doi: 10.2196/43236] [Medline: 37043287]
47. Burke LE, Wang J, Sevick MA. Self-monitoring in weight loss: a systematic review of the literature. J Am Diet Assoc 2011 Jan;111(1):92-102. [doi: 10.1016/j.jada.2010.10.008] [Medline: 21185970]
48. Tobacco. World Health Organization. 2023. URL: https://www.who.int/news-room/fact-sheets/detail/tobacco [accessed 2025-04-09]
49. Alcohol's effects on health. National Institute on Alcohol Abuse and Alcoholism. 2024. URL: https://www.niaaa.nih.gov/alcohol-health/overview-alcohol-consumption/moderate-binge-drinking [accessed 2025-04-09]
50. Drury VB, Saw SM, Finkelstein E, Wong TY, Tay PK. A new community-based outdoor intervention to increase physical activity in Singapore children: findings from focus groups. Ann Acad Med Singap 2013 May;42(5):225-231. [Medline: 23771109]
51. Aune D, Keum N, Giovannucci E, et al. Whole grain consumption and risk of cardiovascular disease, cancer, and all cause and cause specific mortality: systematic review and dose-response meta-analysis of prospective studies. BMJ 2016 Jun 14;353:i2716. [doi: 10.1136/bmj.i2716] [Medline: 27301975]
52. Vorrink S, Huisman C, Kort H, Troosters T, Lammers JW. Perceptions of patients with chronic obstructive pulmonary disease and their physiotherapists regarding the use of an eHealth intervention. JMIR Hum Factors 2017 Sep 19;4(3):e20. [doi: 10.2196/humanfactors.7196] [Medline: 28928110]
53. IBM SPSS statistics for Windows. Version 28.0 ed. : IBM Corp; 2021.
54. Graneheim UH, Lundman B. Qualitative content analysis in nursing research: concepts, procedures and measures to achieve trustworthiness. Nurse Educ Today 2004 Feb;24(2):105-112. [doi: 10.1016/j.nedt.2003.10.001] [Medline: 14769454]
55. McDowell JRS, McPhail K, Halyburton G, Brown M, Lindsay G. Perceptions of a service redesign by adults living with type 2 diabetes. J Adv Nurs 2009 Jul;65(7):1432-1441. [doi: 10.1111/j.1365-2648.2009.05003.x]
56. Russell CK, Gregory DM. Evaluation of qualitative research studies. Evid Based Nurs 2003 Apr;6(2):36-40. [doi: 10.1136/ebn.6.2.36] [Medline: 12710415]
57. Currency converter. OANDA. 2025. URL: https://www.oanda.com/currency-converter/en/?from=KRW&to=USD&amount=1 [accessed 2025-04-09]
58. Kassavou A, Wang M, Mirzaei V, Shpendi S, Hasan R. The association between smartphone app-based self-monitoring of hypertension-related behaviors and reductions in high blood pressure: systematic review and meta-analysis. JMIR Mhealth Uhealth 2022 Jul 12;10(7):e34767. [doi: 10.2196/34767] [Medline: 35819830]
59. MacPherson MM, Merry KJ, Locke SR, Jung ME. How can we keep people engaged in the behavior change process? An exploratory analysis of two mHealth applications. J Technol Behav Sci 2022 Sep;7(3):337-342. [doi: 10.1007/s41347-022-00253-7]
60. Sevick MA, Zickmund S, Korytkowski M, et al. Design, feasibility, and acceptability of an intervention using personal digital assistant-based self-monitoring in managing type 2 diabetes. Contemp Clin Trials 2008 May;29(3):396-409. [doi: 10.1016/j.cct.2007.09.004] [Medline: 17997364]
61. Burke LE, Styn MA, Sereika SM, et al. Using mHealth technology to enhance self-monitoring for weight loss: a randomized trial. Am J Prev Med 2012 Jul;43(1):20-26. [doi: 10.1016/j.amepre.2012.03.016] [Medline: 22704741]
62. Patnode CD, Redmond N, Iacocca MO, Henninger M. Behavioral counseling interventions to promote a healthy diet and physical activity for cardiovascular disease prevention in adults without known cardiovascular disease risk factors: updated evidence report and systematic review for the US preventive services task force. JAMA 2022 Jul 26;328(4):375-388. [doi: 10.1001/jama.2022.7408] [Medline: 35881116]
63. Jakob R, Harperink S, Rudolf AM, et al. Factors influencing adherence to mHealth apps for prevention or management of noncommunicable diseases: systematic review. J Med Internet Res 2022 May 25;24(5):e35371. [doi: 10.2196/35371] [Medline: 35612886]
64. Ryan RM, Deci EL. Intrinsic and extrinsic motivations: classic definitions and new directions. Contemp Educ Psychol 2000 Jan;25(1):54-67. [doi: 10.1006/ceps.1999.1020] [Medline: 10620381]
65. Smithson J. Using and analysing focus groups: limitations and possibilities. Int J Soc Res Methodol 2000 Jan;3(2):103-119. [doi: 10.1080/136455700405172]

XSL•FO
RenderX

Abbreviations

IMB: Information-Motivation-Behavioral Skills

Edited by A Mavragani; submitted 04.09.24; peer-reviewed by D Lauver, E Deja; revised version received 10.03.25; accepted 11.03.25; published 02.05.25.

Please cite as:
Choo J, Noh S, Shin Y
Evaluating Feasibility and Acceptability of the "My HeartHELP" Mobile App for Promoting Heart-Healthy Lifestyle Behaviors: Mixed Methods Study
JMIR Form Res 2025;9:e66108
URL: https://formative.jmir.org/2025/1/e66108
doi:10.2196/66108

© Jina Choo, Songwhi Noh, Yura Shin. Originally published in JMIR Formative Research (https://formative.jmir.org), 2.5.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Characterizing Patient-Reported Fatigue Using Electronic Diaries in Neurodegenerative and Immune-Mediated Inflammatory Diseases: Observational Study

Adrien Bennetot[1], PhD; Rana Zia Ur Rehman[2], PhD; Robbin Romijnders[3], PhD; Zhi Li[4], PhD; Victoria Macrae[5,6]; Kristen Davies[7], MD; Wan-Fai Ng[6,7,8], PhD; Walter Maetzler[3], MD; Jennifer Kudelka[3], MD; Hanna Hildesheim[3], Dr; Kirsten Emmert[3], PhD; Emma Paulides[9], MD; C Janneke van der Woude[9], Dr; Ralf Reilmann[10,11,12], Dr; Svenja Aufenberg[10]; Meenakshi Chatterjee[13], PhD; Nikolay V Manyakov[14], PhD; Clémence Pinaud[1]; Stefan Avey[4], PhD

[1]Let it Care, 11 avenue Marquise du Deffand, Antony, France
[10]George-Huntington-Institute, University of Münster, Münster, Germany
[11]Department of Neurodegenerative Diseases, Hertie Institute for Clinical Brain Research , University of Tübingen, Tübingen, Germany
[12]Department of Clinical Radiology, University of Münster, Münster, Germany
[13]Johnson & Johnson, Cambridge, MA, United States
[14]Johnson & Johnson, Beerse, Belgium
[2]Johnson & Johnson, Buckinghamshire, United Kingdom
[3]Department of Neurology, University Hospital Schleswig-Holstein Campus Kiel, Kiel University, Kiel, Germany
[4]Johnson & Johnson, Spring House, PA, United States
[5]NIHR Newcastle Biomedical Research Centre, Newcastle Upon Tyne Hospitals NHS Foundation Trust, Newcastle Upon Tyne, United Kingdom
[6]NIHR Newcastle Clinical Research Facility, Newcastle Upon Tyne Hospitals NHS Foundation Trust, Newcastle Upon Tyne, United Kingdom
[7]Translational and Clinical Research Institute, Faculty of Medical Sciences, Newcastle University, Newcastle Upon Tyne, United Kingdom
[8]HRB Clinical Research Facility, University College Cork, Cork, Ireland
[9]Department of Gastroenterology and Hepatology, Erasmus MC, Rotterdam, The Netherlands

Corresponding Author:
Adrien Bennetot, PhD
Let it Care, 11 avenue Marquise du Deffand, Antony, France

Abstract

Background: Fatigue is a prevalent and debilitating symptom in many chronic conditions, including immune-mediated inflammatory diseases (IMIDs) and neurodegenerative diseases (NDDs). Fatigue often fluctuates significantly within and between days, yet traditional patient-reported outcomes (PROs) typically rely on recall periods of a week or more, potentially missing these short-term variations. The development of digital tools, such as electronic diaries (eDiaries), offers a unique opportunity to collect granular, real-time data. However, the feasibility, adherence, and comparability of eDiary-based assessments to established PROs require further investigation.

Objective: This study aimed to evaluate the feasibility and acceptability of using a high-frequency eDiary to capture intraday variability in fatigue and to compare eDiary data with scores obtained from the Functional Assessment of Chronic Illness Therapy-Fatigue (FACIT-F), a validated weekly recall PRO.

Methods: Data were collected from 159 participants enrolled in the IDEA-FAST (Identifying Digital Endpoints to Assess Fatigue, Sleep and Activities in Daily Living in Neurodegenerative Disorders and Immune-Mediated Inflammatory Diseases) feasibility study; a 4-week prospective observational study conducted at 4 European centers. Participants included individuals with NDDs (n=39), IMIDs (n=78), and healthy volunteers (n=42). Participants used an eDiary to report their physical and mental fatigue levels up to 4 times daily on a 7-point Likert scale (0=low and 6=high). Adherence was calculated as the proportion of completed eDiary entries relative to the total expected entries. Correlations between averaged eDiary scores and weekly FACIT-F scores were analyzed.

Results: Adherence to the eDiary protocol was 5505/8880 (61.99%) overall, varying by cohort, with the highest adherence (1117/1200, 93.07%) observed in the primary Sjögren syndrome cohort and the lowest adherence in the Parkinson disease (410/960, 42.7%) and Huntington disease (320/720, 44.4%) cohorts. The average adherence was 430/1680 (43.45%) in the NDD cohorts and 3367/4560 (73.84%) in the IMID cohorts. Fatigue levels showed clear diurnal variation, with significantly higher fatigue reported in the evening compared to the morning (P<.001). A moderate correlation (Spearman=0.46, P<.001) was observed

XSL•FO
RenderX

between eDiary fatigue scores and FACIT-F scores, with stronger cohort-specific associations for certain FACIT-F items. These results indicate that eDiaries provide complementary insights to weekly PROs by capturing intraday fluctuations in fatigue.

Conclusions: This study demonstrates the feasibility, acceptability, and validity of using high-frequency eDiaries to assess fatigue in chronic conditions. By effectively detecting intra- and interday fatigue variations, eDiaries complement traditional PROs such as FACIT-F, offering a more nuanced understanding of fatigue patterns. Future research should explore optimized eDiary protocols to balance participant burden with data granularity.

(JMIR Form Res 2025;9:e65879) doi:10.2196/65879

KEYWORDS

chronic disease; fatigue; neurodegenerative diseases; immune-mediated inflammatory diseases; diary, patient-reported outcomes; electronic diaries; digital technologies; digital health; eHealth; mobile phone

Introduction

Fatigue poses a significant personal and financial burden to individuals living with immune-mediated inflammatory diseases (IMID) [1,2] and neurodegenerative diseases (NDD) [3] as it profoundly affects the individuals' overall quality of life and daily activities [4]. The importance of fatigue is underscored by research priorities that emphasize the urgent need to unravel the intricate factors contributing to fatigue and to develop effective solutions [2].

Fatigue is a multifaceted phenomenon that extends beyond physical exhaustion, impacting cognitive, emotional, and motivational aspects [2,4]. Its complexity makes it challenging to assess and comprehend fully, particularly due to its subjective nature and variations among individuals, often not addressed during routine medical visits. Considering this, continuous monitoring of fatigue could provide useful information to facilitate fatigue assessment. By continuously observing and studying this complex phenomenon, we may also gain valuable insights into its variability and underlying mechanisms of fatigue.

Conventionally, participant-reported outcomes (PROs) have been the primary method of measuring fatigue in clinical trials and research studies. These PROs rely on self-reports to gauge the frequency, duration, and severity of fatigue [5,6]. However, PROs have several limitations, including recall bias and insensitivity to changes over time [7,8].

To address this need, an Innovative Medicines Initiative funded project was created called IDEA-FAST (Identifying Digital Endpoints to Assess Fatigue, Sleep and Activities in Daily Living in Neurodegenerative Disorders and Immune-Mediated Inflammatory Diseases) [9]: identifying digital end points to assess fatigue, sleep, and activities of daily living using a combination of digital health technologies to comprehensively assess fatigue in the real world. These digital measures hold the potential to provide objective assessments of fatigue that may complement self-reported fatigue to provide a more holistic view of fatigue. However, the identification and evaluation of objective digital measures of fatigue necessitates comparisons with perceived fatigue at relatively frequent intervals, preferably multiple times a day [10,11].

As such, this study focuses on evaluating the properties of a 4-times-a-day digital questionnaire. This study's objective is to lay the foundation for the development and evaluation of digital measures of fatigue by examining their properties and comparing them with participant-reported fatigue.

In summary, this study aims to:

1. Determine whether there is any effect of time of the day (eg, morning, afternoon, and evening) on self-reported fatigue, which would argue in favor of fatigue assessments more often than once a day.
2. Determine if 4-times-a-day self-reported fatigue captures the overall trend of fatigue when compared to a validated questionnaire, such as the Functional Assessment of Chronic Illness Therapy-Fatigue (FACIT-F) [12-15] which has a 1-week recall period.
3. Determine the burden of participants to respond to an electronic diary (eDiary) for the assessment of fatigue 4 times a day for 4 weeks.

Methods

Ethical Considerations

The studies involving human participants were reviewed and approved by the Ethical Committee of the Medical Faculty, Kiel University, Kiel, Germany; Health Research Authority and Health and Care Research Wales, United Kingdom; Ethik-Kommission der Ärztekammer Westfalen-Lippe und der Westfälischen Wilhelms-Universität, Münster, Germany; and the Medical Ethics Review Committee, Erasmus MC, Netherlands. The patients or participants provided their written informed consent to participate in this study. Ethical approval was first granted by the Ethical Committee of the Medical Faculty of Kiel University (D491/20) in June 2020 and then by the Research Ethics Committees of all other study sites: Newcastle upon Tyne Hospitals National Health Service (NHS) Foundation Trust/Newcastle University in August 2020 (20/PR/0185), Erasmus University Medical Centre in Rotterdam in November 2020, and George-Huntington-Institute in Münster in September 2020 (2020-654-b-S). This study was registered with the German Clinical Trial Registry (DRKS00021693) and was conducted according to the principles of the Declaration of Helsinki (version of 2013). Data were pseudonymized in the clinical trial database before analysis.

Participants were recruited through multiple channels across the 4 study centers. Recruitment lists of individuals who had previously consented to be contacted for research purposes were used. Additionally, patients currently receiving inpatient or

XSL•FO
RenderX

outpatient treatment at the respective clinics were informed about this study and invited to participate. Healthy control participants were recruited by approaching relatives of individuals with study-relevant diagnoses.

All participants received compensation for their travel expenses, which included coverage of costs incurred using public transport. For participants using their own car, reimbursement was provided based on the standard local cost per kilometer of travel. Importantly, no financial compensation was offered for participation in this study itself.

Study Demographics

The feasibility study (FS) of the IDEA-FAST project aims to identify candidate digital measures of fatigue and sleep disturbances in 6 disease populations which include Parkinson disease (PD), Huntington disease (HD), inflammatory bowel disease (IBD), primary Sjögren syndrome (PSS), rheumatoid arthritis (RA), and systemic lupus erythematosus (SLE) as well as healthy volunteers.

Participants from the 7 cohorts (healthy, PD, HD, IBD, SLE, RA, and PSS) were recruited in 4 European centers: Universitätsklinikum Schleswig-Holstein Kiel, Newcastle University, Erasmus University Medical Centre Rotterdam, and George-Huntington-Institute Münster. A total of 159 volunteers were recruited for this study. Among them, 111 participants filled at least 2 eDiary entries and 128 participants completed at least 1 “home visit,” during which a staff nurse would visit them at home or call them on the telephone to make them fill in the weekly PROs. Participants at the intersection of these 2 sets are used for the analysis (103 participants) unless otherwise indicated.

Demographic and clinical aspects of the participants are summarized in Table 1.

XSL•FO
RenderX

Table . Demographics and clinical aspects of the recruited participants.

Variable and category		Participants		
		with clinical data (n=159)	with eDiary[a] data (n=111)	with weekly PROs[b] (n=128)
Age (years), mean (SD)		53 (16)	53 (16)	54 (15)
BMI, mean (SD)		26 (5)	26 (5)	26 (5)
Time since diagnosis (years), mean (SD)		11 (8)	11 (7)	11 (8)
Site, n (%)				
	Erasmus University Medical Centre Rotterdam	28 (18)	17 (15)	16 (13)
	George-Huntington-Institute	18 (12)	12 (11)	18 (14)
	Universitätsklinikum Schleswig-Holstein Kiel	56 (35)	34 (31)	41 (32)
	Newcastle University	57 (35)	48 (43)	53 (41)
Cohort, n (%)				
	HD[c]	14 (9)	7 (6)	12 (9)
	Healthy	42 (27)	30 (27)	38 (29)
	IBD[d]	18 (11)	13 (11)	9 (7)
	PD[e]	25 (16)	16 (15)	18 (15)
	PSS[f]	18 (11)	16 (15)	18 (15)
	RA[g]	24 (15)	16 (15)	17 (13)
	SLE[h]	18 (11)	13 (11)	16 (12)
Gender, n (%)				
	Female	102 (64)	74 (66)	82 (64)
	Male	57 (36)	37 (34)	46 (36)
Comorbidities, n (%)				
	At least 1 comorbidity	40 (25)	26 (23)	28 (21)
Physical activity (Godin[i]), n (%)				
	Active	90 (56)	70 (63)	79 (61)
	Insufficiently active	23 (14)	16 (14)	20 (15)
	Moderately active	24 (15)	19 (17)	22 (17)

[a]eDiary: electronic diary.
[b]PRO: participant-reported outcome.
[c]HD: Huntington disease.
[d]IBD: inflammatory bowel disease.
[e]PD: Parkinson disease.
[f]PSS: primary Sjögren syndrome.
[g]RA: rheumatoid arthritis.
[h]SLE: systemic lupus erythematosus.
[i]Godin: Godin-Shephard leisure-time physical activity questionnaire.

Protocol and Data Collection

Participants attended a scheduled study visit at the recruitment center at the beginning of this study, during which their basic demographics and medical data were collected. Each participant underwent a longitudinal evaluation over 4 weeks as described in Figure 1.

XSL•FO
RenderX

Figure 1. Data collection protocol for this study, composed of 4 home periods of 5 days spaced between 2 and 4 days. During each period participants reported their physical and mental fatigue 4 times a day. After each period, participants completed a FACIT-F questionnaire. FACIT-F: Functional Assessment of Chronic Illness Therapy-Fatigue.

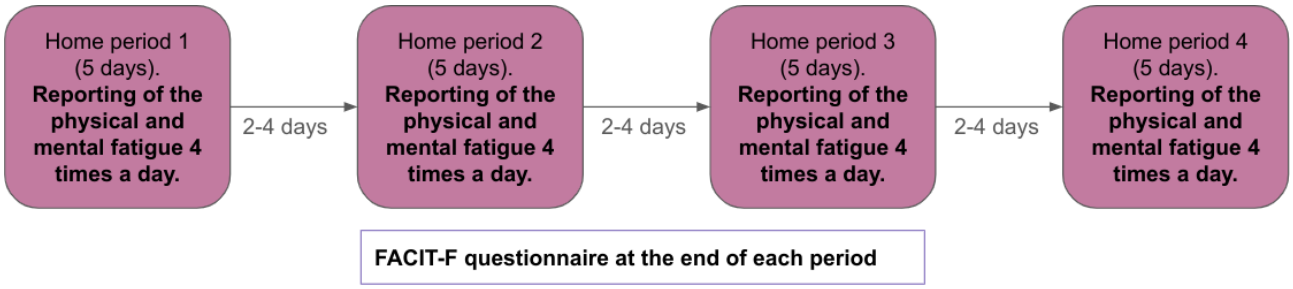

During this study, 2 types of data were collected:

- After each device use period, participants were asked to complete PROs, including the FACIT-F questionnaire, to assess their fatigue level during the past 7 days during a home visit or a phone call. The FACIT-F is a 13-item scale measuring fatigue, validated for several chronic illnesses including RA [14]. For each item, participants select a response from a 5-point Likert scale. Total scores range from 0 to 52 with lower scores indicating more fatigue [16].
- During the 4 weeks of this study, participants were asked to report their mental and physical levels of fatigue 4 times a day, by answering the question “At the moment I feel” with a 7 item Likert scale from “low” (0) to “high” (6). Participants must answer 2 times, once for physical fatigue and once for mental fatigue. Online questionnaires in the FS were facilitated by an app running on the participants’ smartphones. We refer to those questionnaires as the eDiary.

Participant Adherence to the Protocol

Participant adherence to the 4-times-a-day eDiary was defined as the number of completed physical fatigue diary entries received during the first 28 days of the study divided by the total number of expected fatigue diary entries across all patients and days. The expected number of diary entries was calculated per patient as 80; four times a day for four 5-day home reporting periods. This method was chosen because the exact dates of the home period start and end were not recorded and varied by each participant. Adherence was calculated in the set of 111 participants with at least 1 eDiary entry. Adherence was first summarized at the participant level and then further aggregated by cohort.

Effect of Time of the Day on Fatigue

The relationship between the level of fatigue and time of the day was investigated using linear mixed-effects models (LMMs), which can accommodate missing data [17], on the 4 times a day eDiary answers. Either physical or mental fatigue was set as the dependent variable, time of the day and cohort, and their interaction term, were modelled as fixed effects, and the subject was modelled as a random effect. As the eDiary responses were timestamped at the whole hour (eg, 9 am or 1 pm), the time of the day was modelled as a categorical variable (morning, afternoon, evening, and night). LMMs were calculated with JASP (Jeffreys’s Amazing Statistics Program), and an effect was considered significant if the corresponding *P* value of the likelihood ratio test was below .05.

Daily eDiary Association With FACIT-F

Physical and mental fatigue from daily eDiary were compared with the FACIT-F overall score and its items. To examine the relationship between fatigue levels measured by the FACIT-F questionnaire and daily diary scores across the 4 home visits, a repeated-measures correlation analysis was conducted between FACIT-F and its items scores with a separately daily eDiary, where the eDiary score was averaged across each home period before FACIT-F assessment. This analysis accounts for the repeated nature of the data, which involved measuring fatigue at multiple time points for each participant.

Additionally, an association between each home visit and FACIT-F was also explored separately, where Spearman correlation coefficients were calculated to assess the strength and direction of the relationship between the FACIT-F questionnaire items and the average daily diary scores for physical and mental fatigue for each home visit separately. The significance level was set at $P<.05$.

Due to the exploratory nature of our analysis, no adjustments in multiple comparisons were performed.

Results

Participant Adherence

Overall, the adherence was acceptable given the high frequency of administration (4 times per day) with 5505/8880 (61.99%) of expected diary entries completed over all days. The percentage of daily responses completed by cohort and overall is shown in Figure 2. Across all participants, 88/111 (79.3%) of days had at least 1 response, 72/111 (64.9%) has at least 2 responses, 55/111 (49.5%) had at least 3 responses, and 32/111 (28.8%) had all 4 expected responses. The adherence varied dramatically between cohorts with the highest adherence (1117/1200, 93.08%) observed in the PSS cohort and the lowest adherence in the PD (410/960, 42.7%) and HD (320/720, 44.4%) cohorts. The average adherence was 730/1680 (43.45%) in the NDD cohorts and 3367/4560 (73.84%) in the IMID cohorts.

XSL•FO
RenderX

Figure 2. Heat map showing the number of completed eDiary physical fatigue questions per day (0 - 4) with each study participant represented by 1 row. Note that there were 3 scheduled breaks of 2 - 4 days between each of the four 5-day home periods during which no responses were expected. eDiary: electronic diary; HD: Huntington disease; PD: Parkinson disease; IBD: inflammatory bowel disease; IMID: immune-mediated inflammatory diseases; NDD: neurodegenerative diseases; PSS: primary Sjögren syndrome; RA: rheumatoid arthritis; SLE: systemic lupus erythematosus.

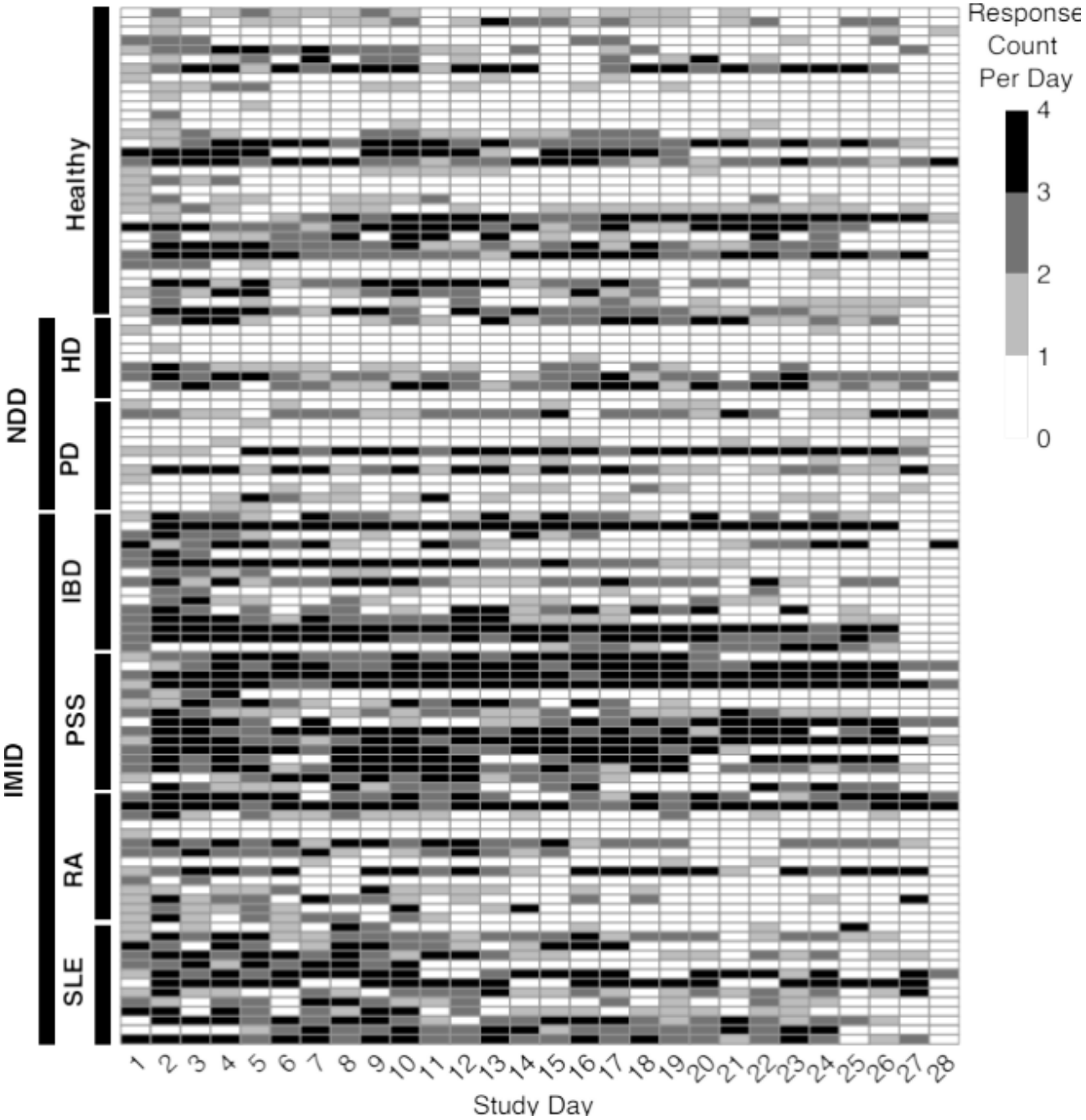

For the following analyses we retained only those participants who responded to at least 1 FACIT-F questionnaire and at least 1 eDiary during 5 days before the FACIT-F. For the LMM, the likelihood ratio test showed that there were significant associations between time of the day and both physical fatigue and mental fatigue (Table 2). Visual inspection suggests that participants experienced more physical and mental fatigue at later times of the day (Figure 3).

XSL•FO
RenderX

Table . Results of LMMs[a] used to assess the relationship between self-reported physical and mental fatigue levels and the time of day across 7 participant cohorts in the IDEA-FAST[b] feasibility study.[c]

Effect		Chi-square (*df*)	*P* value
Physical fatigue			
	Cohort	17.3 (6)	.008
	Time of the day	54.2 (3)	<.001
	Cohort * time of the day	17.1 (18)	.51
Mental fatigue			
	Cohort	11.7 (6)	.07
	Time of the day	46.02 (3)	<.001
	Cohort * time of the day	20.4 (18)	.31

[a]LMM: linear mixed-effects model.

[b]IDEA-FAST [9]: Identifying Digital Endpoints to Assess Fatigue, Sleep and Activities in Daily Living in Neurodegenerative Disorders and Immune-Mediated Inflammatory Diseases.

[c]Fatigue levels were measured using a 7-point Likert scale (0=low, 6=high) with 4 daily eDiary (electronic diary) entries (morning, afternoon, evening, and night) over a 4-week observation period. The models included time of day and participant cohort as fixed effects and individual participants as random effects to account for repeated measures. The 7 cohorts included healthy volunteers and participants with Parkinson disease (PD), Huntington disease (HD), inflammatory bowel disease (IBD), primary Sjögren syndrome (PSS), rheumatoid arthritis (RA), and systemic lupus erythematosus (SLE). The table highlights significant diurnal variations in fatigue and differences between cohorts.

XSL•FO
RenderX

Figure 3. Scatterplot of subject-averaged physical and mental fatigue for different times of the day. Each dot represents the average self-reported fatigue for all responses for a single participant, during a respective period of a day. The cohort-level averages show increased levels of both physical (top) and mental fatigue (bottom) at later times of the day. The different cohorts are HD, healthy volunteers, IBD, PD, PSS, RA, and SLE. HD: Huntington disease; IBD: inflammatory bowel disease; PD: Parkinson disease; PSS: primary Sjögren syndrome; RA: rheumatoid arthritis; SLE: systemic lupus erythematosus.

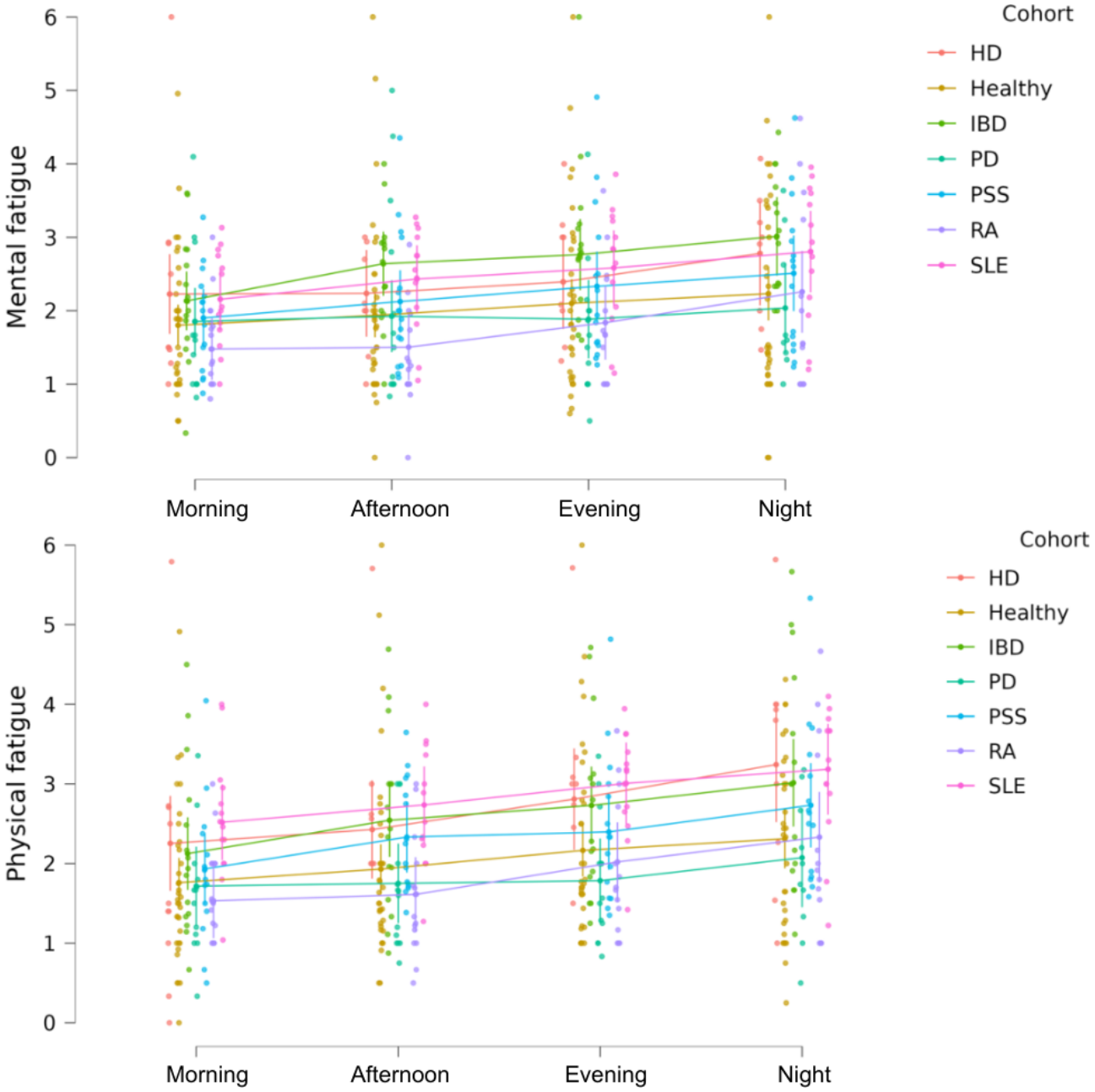

Daily eDiary Association With FACIT-F

The distributions of the physical and mental fatigue questions from daily eDiary and FACIT-F overall scores are shown in Figure 4A-C. Daily eDiary responses and FACIT-F scores are clustered around low fatigue levels.

XSL•FO
RenderX

Figure 4. Distribution of responses. (A) Overall (total) FACIT-F score distribution, (B) physical fatigue from eDiary, and (C) mental fatigue from eDiary. Lower scores on the eDiary questions and higher scores on the FACIT-F questions reflect less fatigue. eDiary: electronic diary; FACIT-F: Functional Assessment of Chronic Illness Therapy-Fatigue; HD: Huntington disease; IBD: inflammatory bowel disease; PD: Parkinson disease; PSS: primary Sjögren syndrome; RA: rheumatoid arthritis; SLE: systemic lupus erythematosus.

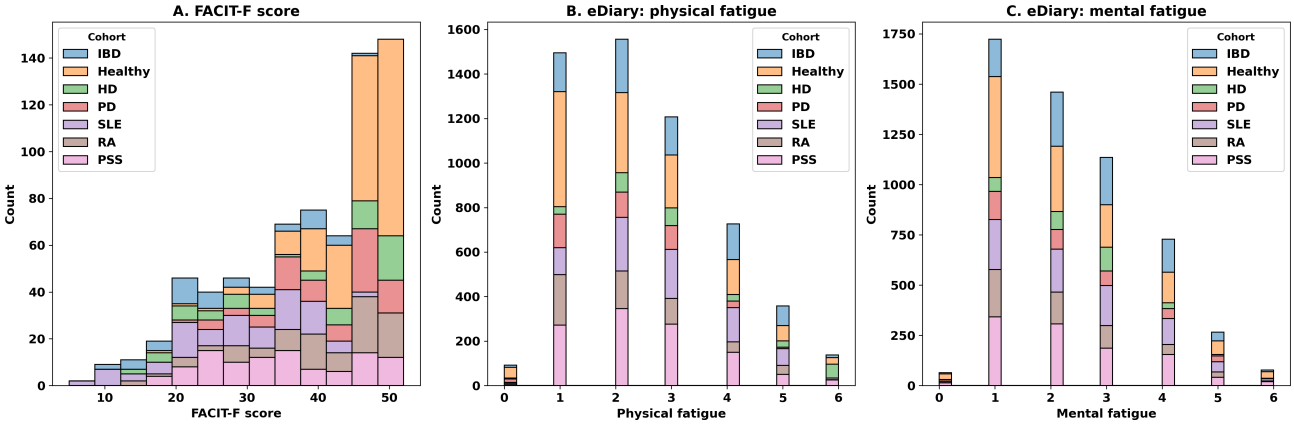

To observe the general trends in the eDiary and FACIT-F responses related to fatigue across various home visits within different cohorts, we plotted the average response of the FACIT-F score (Figure 5A) and the FACIT-F item "I feel fatigued" (Figure 5B) across each home visit as an example. This item "I feel fatigued" from the FACIT-F questionnaire (Figure 5B) is selected for the similarity of its wording to that of the eDiary. It exhibits a similar pattern within various cohorts (IBD, healthy, HD, PD, SLE, and RA) to the physical and mental fatigue eDiary 7-days average (Figure 5C, Figure 5D) before each home visit (corresponding to the end of home periods), where the higher scores in both eDiary and FACIT-F indicate higher fatigue levels. Among the cohorts studied, healthy volunteers demonstrated the lowest levels of fatigue, followed by the PD and HD cohorts.

Figure 5. Responses from FACIT-F and daily diary across 4 home visits (bars indicate the average value while whiskers indicate the SD). (A) Bar plot of FACIT-F overall score (mean 39.57, SD 10.70), (B) bar plot of items of FACIT-F score: I feel fatigued (mean 1.42, SD 1.17), (C) daily dairy physical fatigue response average across 7 days before each home visit (mean 2.18, SD 1.06), and (D) mental fatigue response from daily diary average across 7 days before each home visit (mean 2.07, SD 0.96). The different cohorts are HD, healthy volunteers, IBD, PD, PSS, RA, and SLE. FACIT-F: Functional Assessment of Chronic Illness Therapy-Fatigue; HD: Huntington disease; IBD: inflammatory bowel disease; PD: Parkinson disease; PSS: primary Sjögren syndrome; RA: rheumatoid arthritis; SLE: systemic lupus erythematosus.

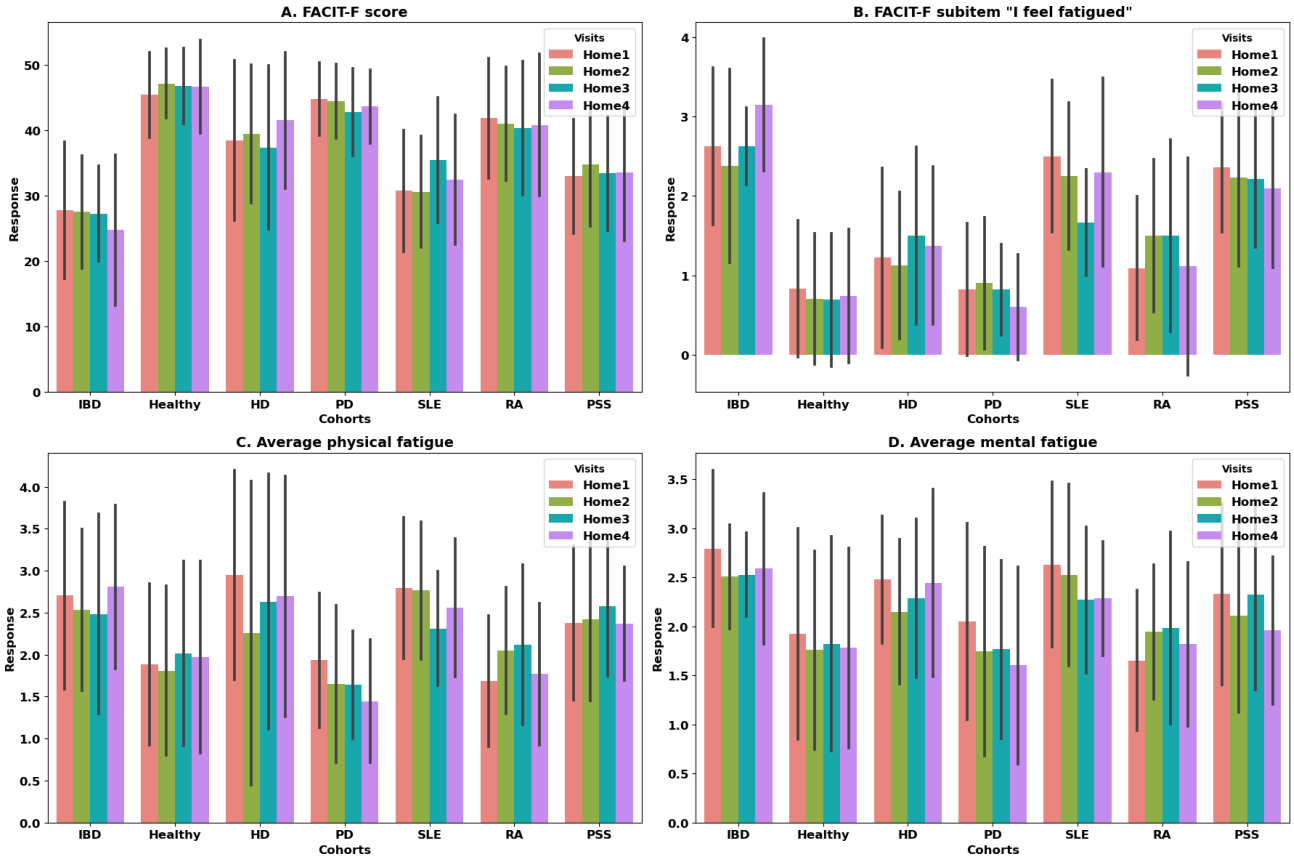

XSL•FO
RenderX

Along with the total FACIT-F score, its item correlations with the eDiary were also investigated. Positive correlations are expected between FACIT-F items and eDiary but negative correlations are expected between the total FACIT-F score and the eDiary, as some items are inverted for the calculation of the total FACIT-F score [15]. Overall, a moderate repeated measure correlation was observed between several items of the FACIT-F questionnaire and the average daily eDiary scores for physical fatigue (Figure 6) and mental fatigue (not shown). However, only correlations between specific FACIT-F items and eDiary score reached statistical significance. For instance, in the PSS cohort, a positive correlation was found between the item "I feel weak all over" and both physical and mental fatigue. Other significant correlations are marked by the black boxes.

Figure 6. Repeated measure correlation between average daily physical fatigue measure with eDiary and FACIT-F overall score and its items. Values reaching statistical significance (P<.05) squared with a black box. Numbers in brackets, for example, "HD (10)," represent the number of different participants involved in the correlation calculation. The different cohorts are HD, healthy volunteers, IBD, PD, PSS, RA, and SLE. eDiary: electronic diary; FACIT-F: Functional Assessment of Chronic Illness Therapy-Fatigue; HD: Huntington disease; IBD: inflammatory bowel disease; PD: Parkinson disease; PSS: primary Sjögren syndrome; RA: rheumatoid arthritis; SLE: systemic lupus erythematosus.

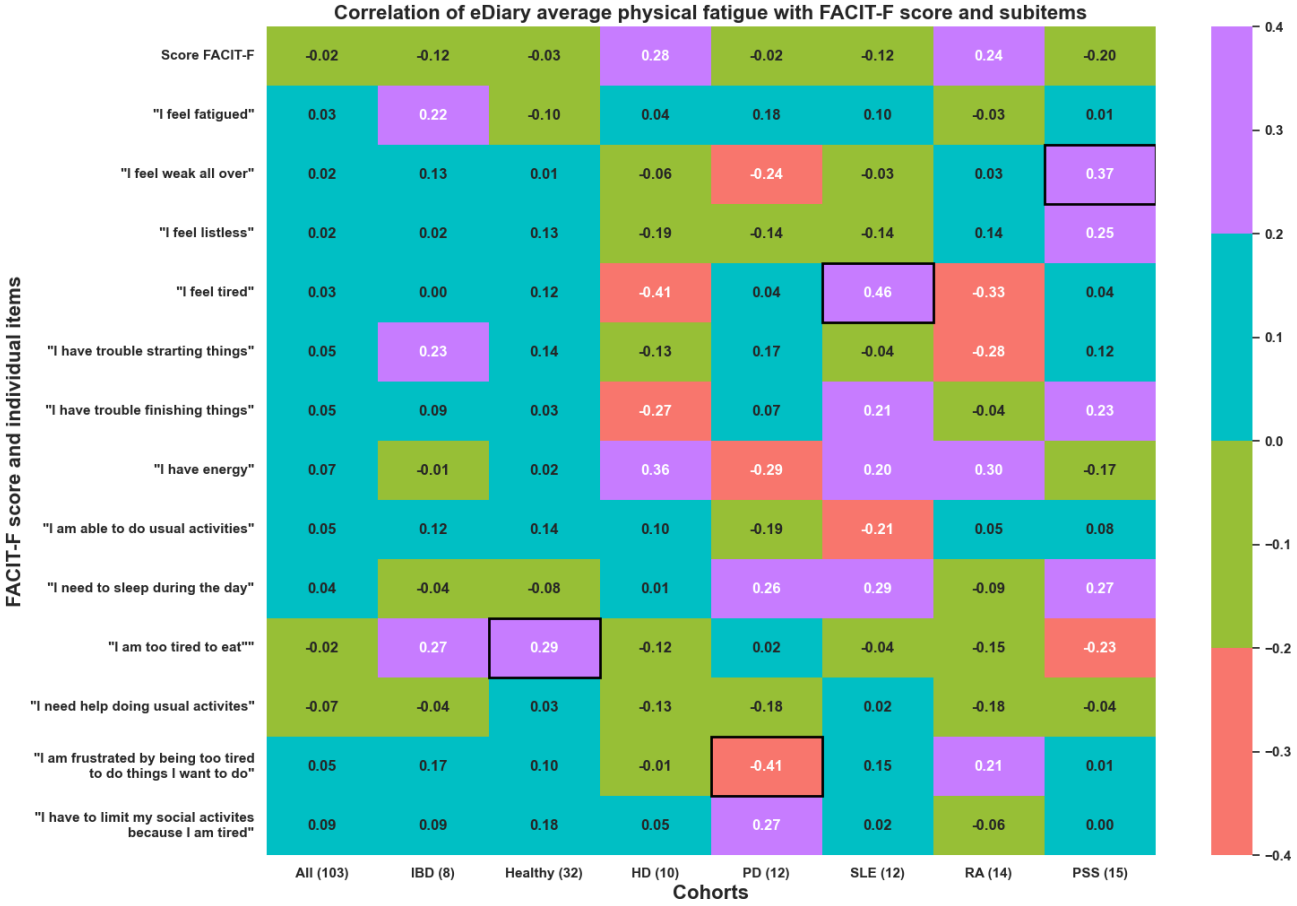

When combining all the cohorts, a negative insignificant correlation was observed between the overall FACIT-F score and both physical and mental fatigue averaged across 7 days before each home visit. Furthermore, certain items, including "feeling fatigued," "feeling weak all over," "feeling listless," "feeling tired," "experiencing trouble in starting and finishing tasks," "having low energy," and others, each exhibited a positive insignificant correlation with both the average physical and mental fatigue scores obtained from the daily eDiary.

Furthermore, the association between the eDiary at different time points (morning, early afternoon, late afternoon, and evening, which were averaged across the week before each home visit), and the FACIT-F score was explored. The heat maps in Figures 7 and 8 illustrate the significant correlations observed in these weekly averaged measures in the morning and evening for physical fatigue (other periods of the day and mental fatigue in Multimedia Appendices 1-3). Importantly, the items having a significant correlation in the morning differed from those in the evening, both for physical and mental fatigue. These findings suggest that fatigue is a multifactorial phenomenon that varies throughout the day as well.

XSL•FO
RenderX

Figure 7. Repeated measures correlation between FACIT-F overall score and its items with 5-days averaged morning assessment of physical fatigue with the eDiary. The different cohorts are HD, healthy volunteers, IBD, PD, PSS, RA, and SLE. eDiary: electronic diary; FACIT-F: Functional Assessment of Chronic Illness Therapy-Fatigue; HD: Huntington disease; IBD: inflammatory bowel disease; PD: Parkinson disease; PSS: primary Sjögren syndrome; RA: rheumatoid arthritis; SLE: systemic lupus erythematosus.

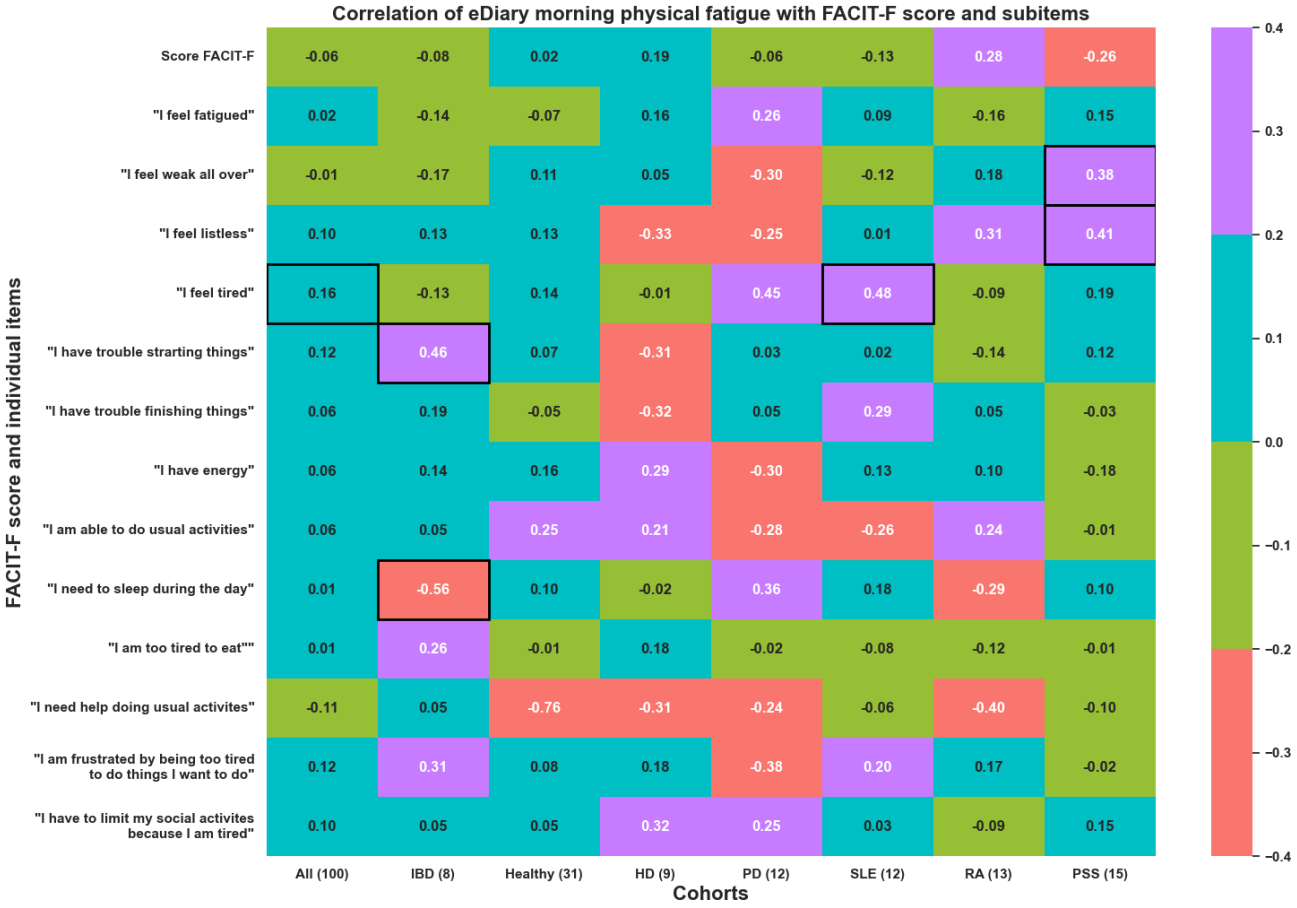

XSL•FO
RenderX

Figure 8. Repeated measures correlation between FACIT-F overall score and its items with the 5-days averaged evening assessment of physical fatigue with the eDiary. The different cohorts are HD, healthy volunteers, IBD, PD, PSS, RA, and SLE. eDiary: electronic diary; FACIT-F: Functional Assessment of Chronic Illness Therapy-Fatigue; HD: Huntington disease; IBD: inflammatory bowel disease; PD: Parkinson disease; PSS: primary Sjögren syndrome; RA: rheumatoid arthritis; SLE: systemic lupus erythematosus.

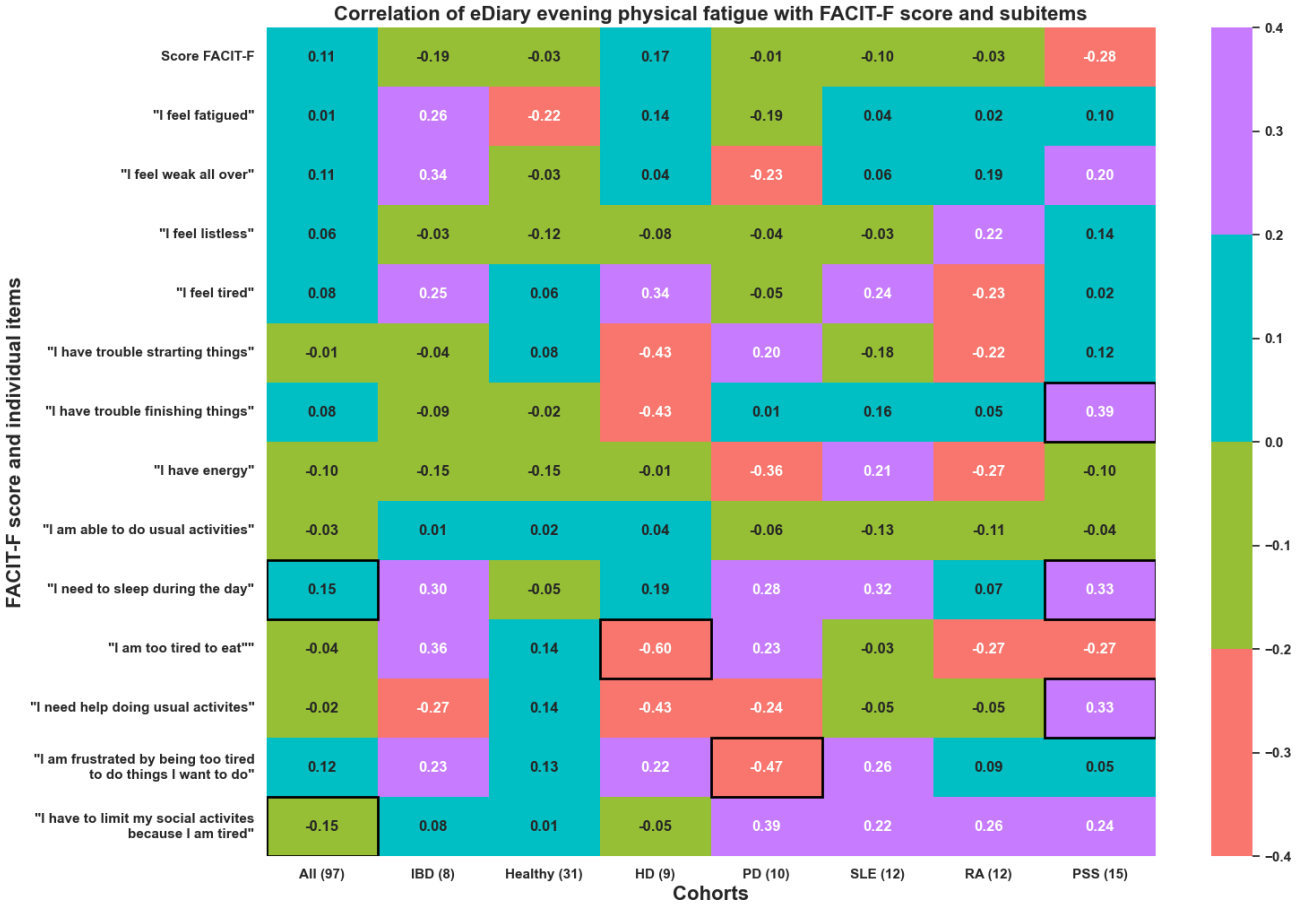

A significant negative Spearman correlation between the FACIT-F overall score and average eDiary score across each home period was observed separately for each visit. The correlation strength was in a similar range (from −0.4 to −0.6) across home visits for various cohorts. One example of the correlation is shown for home period 2 (Figure 9). Other visits are shown in Multimedia Appendices 1-3.

Figure 9. Scatter plot of the individual fatigue scores, with regression lines based on the Pearson correlation between the FACIT-F overall score and the average daily physical (left) and mental (right) fatigue for the home period 2. FACIT-F: Functional Assessment of Chronic Illness Therapy-Fatigue.

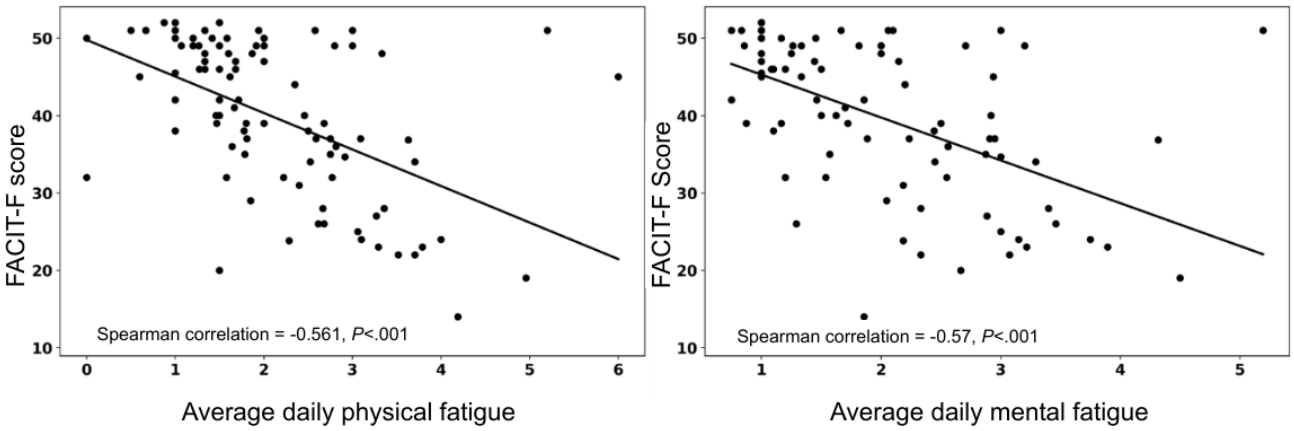

Discussion

This study explored the feasibility of using a high-frequency eDiary for capturing intraday fluctuations in physical and mental fatigue among individuals with IMIDs, NDDs, and healthy volunteers. Our findings indicate that participants adhered to 5505/8880 (61.99%) of the expected diary entries over 4 weeks, with adherence varying across disease cohorts. Fatigue levels, both physical and mental, increased throughout the day, reflecting the impact of cumulative daily activity. This study demonstrated a moderate correlation between eDiary-reported fatigue scores and the FACIT-F, with stronger associations observed for specific cohorts and items, such as “I feel weak

XSL•FO
RenderX

all over" in the PSS group. While overall significance across all groups was not achieved, the findings highlight the complementary role of eDiaries in capturing real-time fatigue patterns that align with weekly recall measures. The ability of eDiaries to detect intraday variability supports their use in fatigue modelling and personalized interventions, bridging subjective self-reports with broader clinical insights.

Overall, the average adherence (number of completed diary entries vs total number of expected diary entries) over the 4 weeks of the study was 5505/8880 (61.99%). Notably, the PSS cohort exhibited the highest percentage of completed eDiary responses, while the HD and PD groups had the lowest. A systematic review of the literature summarizing results from 12 studies in clinical contexts administering a mobile-collected ecological momentary assessment 4 - 5 times/day for an average monitoring period of 12 days found that cohort compliance (defined as the percentage of participants in the cohort who were compliant) was approximately 75% - 85% [18]. This definition of adherence is different from our study, making a direct comparison challenging. It is important to consider that the participants' involvement during our study extended beyond responding to the daily diary because participants of the FS were asked to use several wearable sensors, nonwearable sensors, and mobile phone apps. The complex protocol and participant burden of simultaneous device wearing might have contributed to the lower adherence observed in this study. Future studies should consider reducing the frequency to 3 times per day to retain the ability to capture intraday granularity while improving the probability of adherence. Future studies should also seek to understand how many times per day and over how many days fatigue should be measured to reliably assess fatigue symptoms before and after interventions. To investigate whether the time of the day influences self-reported levels of physical and mental fatigue, participants filled out an eDiary 4 times a day to report their level of fatigue on a scale from 0 (low fatigue) to 6 (high fatigue). Statistical analysis of the reported values by using an LMM suggests that physical and mental fatigue are positively correlated with the time of the day (P value<.001; Table 2). Physical fatigue and mental fatigue are considered 2 separate entities [19]. Physical fatigue is a form of tiredness caused by repeated muscle movements [20]. In contrast, mental fatigue is defined as a psychobiological state caused by prolonged periods of demanding cognitive activity [21,22]. The difference between physical and mental fatigue was clearly explained to the participants of our study and, as such, we can trust that they were able to discriminate them while filling in responses. The finding that at later times of the day higher levels of physical and mental fatigue are observed fits well with these definitions, as it can be expected that people have been both physically and mentally active during the day. The trend that at later times of the day more fatigue is experienced, is at least partly comparable with previous findings showing that tiredness fluctuates throughout the day in both a healthy and major depression disease cohort [23]. Similarly in healthy individuals, subjective tiredness was high in the morning, reached a trough at midday, and then rose steadily until the late evening time to resemble a V-shaped pattern [24,25]. The results presented in our study do not show a V-shape but a continuous increase in fatigue during the day. This difference could possibly be explained by that we are not modelling tiredness but fatigue, which are not the same entities [20,21,26], and the 7-point Likert scale has a limited range of possible values for fatigue. Overall, all these studies argue in favor of the presence of fatigue fluctuations throughout the day, supporting the need for continuous monitoring to capture intraday variability of fatigue.

One of the aims of this study was to assess whether self-reported fatigue reported 4 times a day accurately reflects the overall pattern of fatigue when compared to the validated FACIT-F questionnaire, which asks about fatigue over 1 week. The correlations between the daily eDiary entries and FACIT-F scores were examined in several participant cohorts, such as IBD, healthy, HD, SLE, RA, and PSS. Generally, persons who are in good health tend to have higher scores on the FACIT-F scale, which indicates that they experience less fatigue. On the other hand, participants with chronic diseases such as SLE and RA tend to have lower scores, indicating that they experience higher degrees of fatigue. The fatigue distribution exhibits a differentiation between the groups of individuals with good health and those with chronic illnesses. Concerning physical fatigue, individuals in good health report the lowest levels, while participants with SLE and RA experience higher levels. Participants with IBD and PSS exhibit moderate degrees of fatigue, but those with HD display varied amounts. Similarly, in the case of mental fatigue, those who are in good health report the lowest levels, while participants with SLE and RA report higher levels. Participants with HD, IBD, and PSS exhibit a wider range of levels, with many expressing considerable mental exhaustion. A moderate repeated measure correlation was found between specific items of the FACIT-F questionnaire and the average daily eDiary ratings for physical and mental fatigue. However, only few items achieved statistical significance. In the PSS cohort, a positive association was observed between the item "I feel weak all over" with both physical and mental exhaustion. When all groups were considered together, the overall association between FACIT-F scores and daily eDiary scores did not show statistical significance. However, specific items showed positive but insignificant relationships. This indicates that although eDiaries can accurately record patterns of fatigue, the degree of association may differ depending on the particular items and groups of participants.

These findings are consistent with previous research that shows the effectiveness of eDiaries in managing chronic illnesses [27] compared the FACIT-F with a simpler visual analogue scale ranging from 0 to 10 to evaluate fatigue. This study revealed a robust association between the 2 measures, indicating that frequent self-reports can effectively capture fatigue patterns observed in weekly evaluations. A separate investigation conducted by Cella et al [28] examined the extent to which the FACIT-F accurately measures self-reported fatigue ratings and found a strong level of consistency across several assessment techniques. Additionally, a study conducted by Rao et al [11] showed a good correlation between digital measures and daily self-reported exhaustion. This finding supports the reliability of frequent self-assessment in accurately capturing patterns of fatigue. There is a moderate correlation between daily eDiary entries and FACIT-F scores, suggesting eDiary could be used to monitor fatigue in near real time. This capacity improves

XSL•FO
RenderX

patient care by allowing daily fatigue reports-based interventions. Four times a day fatigue data accurately represented fatigue trends, which may be compared to the FACIT-F questionnaire's weekly recall period. The consistency of this pattern suggests that frequent self-reporting can provide reliable data for fatigue modelling using digital tools. Given the multifaceted nature of fatigue, a holistic approach to assessing the relevant physiological signals is essential, where the eDiary data captured during the day can act as a label.

Several limitations warrant consideration. First, this study's complexity, involving eDiaries, wearable sensors, and multiple home visits, likely contributed to participant burden and reduced adherence. Simplified protocols could mitigate this issue. Second, the use of a 7-point Likert scale for fatigue assessment may have constrained variability, underestimating true intraday fluctuations. Third, this study's duration of 4 times once a week may not reflect long-term patterns or the impact of seasonal changes. Additionally, participants frequently missed completing 1 or more fatigue questionnaires on a given day. This underscores the importance of optimizing assessment frequency to achieve a balance between maintaining data quality and ensuring participant adherence. Lastly, cohort-specific differences in adherence and correlations between eDiary and FACIT-F scores highlight the need for personalized approaches in future studies.

This study demonstrates the feasibility and utility of using high-frequency eDiaries for capturing intraday variations in physical and mental fatigue across diverse populations, including individuals with neurodegenerative and IMIDs. The observed adherence rates and correlations with the FACIT-F underscore the complementary value of eDiaries alongside established weekly recall-based measures. Importantly, eDiaries allow for the detection of nuanced diurnal patterns, providing a richer understanding of fatigue dynamics that traditional tools may overlook.

The integration of eDiaries into clinical research offers several broader implications. By enabling real-time monitoring, eDiaries can inform personalized interventions. These insights are particularly valuable for individuals with chronic conditions, where managing fatigue is a critical component of improving quality of life. From a research perspective, eDiaries can enhance the precision of clinical trials, allowing for better characterization of treatment effects on fatigue. Their ability to capture intraday fluctuations also opens avenues for exploring the interplay between fatigue and physiological markers captured by wearable sensors. This multimodal approach can deepen our understanding of fatigue's multifactorial nature, bridging subjective experiences with objective measurements. However, the broader adoption of eDiaries necessitates attention to usability and adherence, particularly in populations with severe disease burdens or impairments. Simplified protocols, optimized frequency of assessments, and seamless integration with digital health ecosystems are critical for ensuring widespread adoption.

In conclusion, this study underscores the transformative potential of high-frequency digital tools in fatigue assessment. By complementing traditional measures with real-time insights, eDiaries have the potential to enhance fatigue management in clinical and research settings, paving the way for more personalized, data-driven approaches to chronic disease care.

Acknowledgments

The authors thank the IDEA-FAST (Identifying Digital Endpoints to Assess Fatigue, Sleep and Activities in Daily Living in Neurodegenerative Disorders and Immune-Mediated Inflammatory Diseases) project consortium members [29] for their contributions to the project, facilitating this study. The presented work was conducted as part of the IDEA-FAST project, which has received funding from the Innovative Medicines Initiative 2 Joint Undertaking (grant agreement 853981). This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation program and EFPIA (European Federation of Pharmaceutical Industries and Associations) and associated partners. This communication reflects the view of the authors and neither Innovative Medicines Initiative nor the European Union and EFPIA are liable for any use that may be made of the information contained herein.

Data Availability

The datasets presented in this paper are not readily available because of the sensitive nature of the data. This study's data will be made available upon request for validation purposes, subject to the signing of a suitable data sharing agreement. Validation can only take place on the IDEA-FAST (Identifying Digital Endpoints to Assess Fatigue, Sleep and Activities in Daily Living in Neurodegenerative Disorders and Immune-Mediated Inflammatory Diseases) data platform which supports all necessary statistical software tools. Requests to access the datasets should be directed to the IDEA-FAST website's contact page [30].

Conflicts of Interest

WFN has provided consultations for Amgen, Bristol Myers Squibb, Novartis, Johnson & Johnson, Sanofi, Argenx, IQVIA, Quotients, Resolves Therapeutics, EQT & Veloxis. MC, SA, and NVM are employees of Johnson & Johnson and hold company stocks or stocks options. The remaining authors declare no conflicts of interest.

Multimedia Appendix 1

XSL•FO
RenderX

Scatter plot of the Spearman correlation between the FACIT-F overall score and the average daily (left) physical and (right) mental fatigue for period 1. FACIT-F: Functional Assessment of Chronic Illness Therapy-Fatigue.
[PNG File, 106 KB - formative_v9i1e65879_app1.png]

Multimedia Appendix 2
Scatter plot of the Spearman correlation between the FACIT-F overall score and the average daily (left) physical and (right) mental fatigue for period 3. FACIT-F: Functional Assessment of Chronic Illness Therapy-Fatigue.
[PNG File, 98 KB - formative_v9i1e65879_app2.png]

Multimedia Appendix 3
Scatter plot of the Spearman correlation between the FACIT-F overall score and the average daily (left) physical and (right) mental fatigue for period 4. FACIT-F: Functional Assessment of Chronic Illness Therapy-Fatigue.
[PNG File, 96 KB - formative_v9i1e65879_app3.png]

References

1. Dures E, Fraser I, Almeida C, et al. Patients' perspectives on the psychological impact of inflammatory arthritis and meeting the associated support needs: open-ended responses in a multi-centre survey. Musculoskeletal Care 2017 Sep;15(3):175-185. [doi: 10.1002/msc.1159] [Medline: 27605495]
2. Davies K, Dures E, Ng WF. Fatigue in inflammatory rheumatic diseases: current knowledge and areas for future research. Nat Rev Rheumatol 2021 Nov;17(11):651-664. [doi: 10.1038/s41584-021-00692-1] [Medline: 34599320]
3. Kluger BM, Krupp LB, Enoka RM. Fatigue and fatigability in neurologic illnesses: proposal for a unified taxonomy. Neurology (ECronicon) 2013 Jan 22;80(4):409-416. [doi: 10.1212/WNL.0b013e31827f07be] [Medline: 23339207]
4. Antikainen E, Njoum H, Kudelka J, et al. Assessing fatigue and sleep in chronic diseases using physiological signals from wearables: a pilot study. Front Physiol 2022;13:968185. [doi: 10.3389/fphys.2022.968185] [Medline: 36452041]
5. Hewlett S, Dures E, Almeida C. Measures of fatigue: Bristol Rheumatoid Arthritis Fatigue Multi-Dimensional Questionnaire (BRAF MDQ), Bristol Rheumatoid Arthritis Fatigue Numerical Rating Scales (BRAF NRS) for severity, effect, and coping, Chalder Fatigue Questionnaire (CFQ), Checklist Individual Strength (CIS20R and CIS8R), Fatigue Severity Scale (FSS), Functional Assessment Chronic Illness Therapy (Fatigue) (FACIT-F), Multi-Dimensional Assessment of Fatigue (MAF), Multi-Dimensional Fatigue Inventory (MFI), Pediatric Quality Of Life (PedsQL) Multi-Dimensional Fatigue Scale, Profile of Fatigue (ProF), Short Form 36 Vitality Subscale (SF-36 VT), and Visual Analog Scales (VAS). Arthritis Care Res (Hoboken) 2011 Nov;63 Suppl 11(S11):S263-S286. [doi: 10.1002/acr.20579] [Medline: 22588750]
6. Gawron VJ. Overview of self-reported measures of fatigue. Int J Aviat Psychol 2016 Oct;26(3-4):120-131. [doi: 10.1080/10508414.2017.1329627]
7. Stone AA, Shiffman S, Schwartz JE, Broderick JE, Hufford MR. Patient non-compliance with paper diaries. BMJ 2002 May 18;324(7347):1193-1194. [doi: 10.1136/bmj.324.7347.1193] [Medline: 12016186]
8. Rosenman R, Tennekoon V, Hill LG. Measuring bias in self-reported data. Int J Behav Healthc Res 2011 Oct;2(4):320-332. [doi: 10.1504/IJBHR.2011.043414] [Medline: 25383095]
9. IDEA-FAST. Innovative Health Initiative. URL: https://www.imi.europa.eu/projects-results/project-factsheets/IDEA-FAST [accessed 2025-03-19]
10. Holmberg L, Ohlander EM, Byers T, et al. A search for recall bias in a case-control study of diet and breast cancer. Int J Epidemiol 1996 Apr;25(2):235-244. [doi: 10.1093/ije/25.2.235] [Medline: 9119547]
11. Rao C, Di Lascio E, Demanse D, Marshall N, Sopala M, De Luca V. Association of digital measures and self-reported fatigue: a remote observational study in healthy participants and participants with chronic inflammatory rheumatic disease. Front Digit Health 2023;5:1099456. [doi: 10.3389/fdgth.2023.1099456] [Medline: 37426890]
12. Butt Z, Lai JS, Rao D, Heinemann AW, Bill A, Cella D. Measurement of fatigue in cancer, stroke, and HIV using the Functional Assessment of Chronic Illness Therapy - Fatigue (FACIT-F) scale. J Psychosom Res 2013 Jan;74(1):64-68. [doi: 10.1016/j.jpsychores.2012.10.011] [Medline: 23272990]
13. Cella D, Lai JS, Chang CH, Peterman A, Slavin M. Fatigue in cancer patients compared with fatigue in the general United States population. Cancer 2002 Jan 15;94(2):528-538. [doi: 10.1002/cncr.10245] [Medline: 11900238]
14. Cella D, Yount S, Sorensen M, Chartash E, Sengupta N, Grober J. Validation of the Functional Assessment of Chronic Illness Therapy Fatigue Scale relative to other instrumentation in patients with rheumatoid arthritis. J Rheumatol 2005 May;32(5):811-819. [Medline: 15868614]
15. Cella D, Lai JS, Stone A. Self-reported fatigue: one dimension or more? Lessons from the Functional Assessment of Chronic Illness Therapy--Fatigue (FACIT-F) questionnaire. Support Care Cancer 2011 Sep;19(9):1441-1450. [doi: 10.1007/s00520-010-0971-1] [Medline: 20706850]
16. Tennant KF. Assessment of fatigue in older adults: the FACIT Fatigue Scale (version 4). : Hartford Institute for Geriatric Nursing; 2019 URL: https://hign.org/sites/default/files/2020-06/Try_This_General_Assessment_30.pdf [accessed 2025-03-19]

XSL•FO
RenderX

17. Magezi DA. Linear mixed-effects models for within-participant psychology experiments: an introductory tutorial and free, graphical user interface (LMMgui). Front Psychol 2015;6:2. [doi: 10.3389/fpsyg.2015.00002] [Medline: 25657634]
18. Williams MT, Lewthwaite H, Fraysse F, Gajewska A, Ignatavicius J, Ferrar K. Compliance with mobile ecological momentary assessment of self-reported health-related behaviors and psychological constructs in adults: systematic review and meta-analysis. J Med Internet Res 2021 Mar 3;23(3):e17023. [doi: 10.2196/17023] [Medline: 33656451]
19. Kunasegaran K, Ismail AMH, Ramasamy S, Gnanou JV, Caszo BA, Chen PL. Understanding mental fatigue and its detection: a comparative analysis of assessments and tools. PeerJ 2023;11:e15744. [doi: 10.7717/peerj.15744] [Medline: 37637168]
20. Mizuno K, Tanaka M, Yamaguti K, Kajimoto O, Kuratsune H, Watanabe Y. Mental fatigue caused by prolonged cognitive load associated with sympathetic hyperactivity. Behav Brain Funct 2011 May 23;7:17. [doi: 10.1186/1744-9081-7-17] [Medline: 21605411]
21. Tanaka M, Ishii A, Watanabe Y. Neural effects of mental fatigue caused by continuous attention load: a magnetoencephalography study. Brain Res 2014 May 2;1561:60-66. [doi: 10.1016/j.brainres.2014.03.009] [Medline: 24642273]
22. Boksem MAS, Tops M. Mental fatigue: costs and benefits. Brain Res Rev 2008 Nov;59(1):125-139. [doi: 10.1016/j.brainresrev.2008.07.001] [Medline: 18652844]
23. Crowe E, Daly M, Delaney L, Carroll S, Malone KM. The intra-day dynamics of affect, self-esteem, tiredness, and suicidality in major depression. Psychiatry Res 2019 Sep;279:98-108. [doi: 10.1016/j.psychres.2018.02.032] [Medline: 29661498]
24. Dockray S, Grant N, Stone AA, Kahneman D, Wardle J, Steptoe A. A comparison of affect ratings obtained with ecological momentary assessment and the day reconstruction method. Soc Indic Res 2010 Nov;99(2):269-283. [doi: 10.1007/s11205-010-9578-7] [Medline: 21113328]
25. Stone AA, Schwartz JE, Schkade D, Schwarz N, Krueger A, Kahneman D. A population approach to the study of emotion: diurnal rhythms of a working day examined with the day reconstruction method. Emotion 2006 Feb;6(1):139-149. [doi: 10.1037/1528-3542.6.1.139] [Medline: 16637757]
26. Craik FIM. Effects of distraction on memory and cognition: a commentary. Front Psychol 2014;5:841. [doi: 10.3389/fpsyg.2014.00841] [Medline: 25120527]
27. Frade-Sosa B, Morlà R, Sapena N, Sanmartí R, Gómez-Puerta JA. Pos1474-hpr fatigue is an important comorbidity for ra patients. How can we measure it easily in clinical practice? Ann Rheum Dis 2022 Jun;81:1083. [doi: 10.1136/annrheumdis-2022-eular.1297]
28. Cella D, Wilson H, Shalhoub H, et al. Content validity and psychometric evaluation of Functional Assessment of Chronic Illness Therapy-Fatigue in patients with psoriatic arthritis. J Patient Rep Outcomes 2019 May 20;3(1):30. [doi: 10.1186/s41687-019-0115-4] [Medline: 31111255]
29. The IDEA-FAST investigators. IDEA-FAST. URL: https://idea-fast.eu/the-idea-fast-investigators/ [accessed 2025-03-20]
30. Contact. IDEA-FAST. URL: https://idea-fast.eu/contact/ [accessed 2025-03-20]

Abbreviations

eDiary: electronic diary
FACIT-F: Functional Assessment of Chronic Illness Therapy-Fatigue
FS: feasibility study
HD: Huntington disease
IBD: inflammatory bowel disease
IDEA-FAST: Identifying Digital Endpoints to Assess Fatigue, Sleep and Activities in Daily Living in Neurodegenerative Disorders and Immune-Mediated Inflammatory Diseases
IMID: immune-mediated inflammatory disease
JASP: Jeffreys's Amazing Statistics Program
LMM: linear mixed-effects model
NDD: neurodegenerative disease
PD: Parkinson disease
PRO: patient-reported outcome
PSS: primary Sjogren syndrome
RA: rheumatoid arthritis
SLE: systemic lupus erythematosus

XSL•FO
RenderX

Edited by A Mavragani; submitted 05.09.24; peer-reviewed by R Marshall; revised version received 24.01.25; accepted 25.01.25; published 05.05.25.

Please cite as:
Bennetot A, Zia Ur Rehman R, Romijnders R, Li Z, Macrae V, Davies K, Ng WF, Maetzler W, Kudelka J, Hildesheim H, Emmert K, Paulides E, van der Woude CJ, Reilmann R, Aufenberg S, Chatterjee M, Manyakov NV, Pinaud C, Avey S
Characterizing Patient-Reported Fatigue Using Electronic Diaries in Neurodegenerative and Immune-Mediated Inflammatory Diseases: Observational Study
JMIR Form Res 2025;9:e65879
URL: https://formative.jmir.org/2025/1/e65879
doi:10.2196/65879

© Adrien Bennetot, Rana Zia Ur Rehman, Robbin Romijnders, Zhi Li, Victoria Macrae, Kristen Davies, Wan-Fai Ng, Walter Maetzler, Jennifer Kudelka, Hanna Hildesheim, Kirsten Emmert, Emma Paulides, C Janneke van der Woude, Ralf Reilmann, Svenja Aufenberg, Meenakshi Chatterjee, Nikolay V Manyakov, Clémence Pinaud, Stefan Avey. Originally published in JMIR Formative Research (https://formative.jmir.org), 5.5.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

School-Based Virtual Reality Programming for Obtaining Moderate-Intensity Exercise Among Children With Disabilities: Pre-Post Feasibility Study

Byron Lai[1], PhD; Ashley Wright[1*], MS; Bailey Hutchinson[1*], BS; Larsen Bright[1*]; Raven Young[1*], BS; Drew Davis[1*], MD; Sultan Ali Malik[2*], MBBS; James H Rimmer[3*], PhD; Pelham High Community Engagement Group[4*]

[1]Division of Pediatric Rehabilitation Medicine, Department of Pediatrics, University of Alabama at Birmingham, 1600 7th Avenue South, Birmingham, AL, United States

[2]CMH Kharian Medical College, National University of Medical Sciences, Rawalpindi, Pakistan

[3]Dean's Office, School of Health Professions, University of Alabama at Birmingham, Birmingham, AL, United States

[4]

[*]these authors contributed equally

Corresponding Author:
Byron Lai, PhD
Division of Pediatric Rehabilitation Medicine, Department of Pediatrics, University of Alabama at Birmingham, 1600 7th Avenue South, Birmingham, AL, United States

Abstract

Background: Children have busy daily schedules, making school an ideal setting for promoting health-enhancing exercise behavior. However, children with mobility disabilities have limited exercise options to improve their cardiorespiratory fitness and cardiometabolic health.

Objective: This study aims to test the feasibility of implementing a virtual reality (VR) exercise program for children with mobility disabilities in a high school setting.

Methods: A pre- to posttrial single-group design with a 6-week exercise intervention was conducted at a high school. The study aimed to enroll up to 12 students with a disability. Participants were given the option of exercising at home or school. The exercise prescription was three 25-minute sessions per week at a moderate intensity, using a head-mounted VR display. School exercise sessions were supervised by research staff. Home exercise sessions were performed autonomously. Several implementation metrics of feasibility were recorded, including exercise attendance, volume, adverse events or problems, and benefits related to health-related fitness (walking endurance and hand-grip strength). The study also included a qualitative evaluation of critical implementation factors and potential benefits for participants that were not included in the study measures. Outcomes were descriptively analyzed, and 2-tailed *t* tests were used as appropriate.

Results: In total, 10 students enrolled in the program and 9 completed the study (mean age 17, SD 0.6 y). In total, 5 (56%) participants exercised at school, and 4 (44%) exercised at home; 1 participant dropped out prior to exercise. The mean attendance for all 9 completers was 61.1% (11/18 sessions). The mean exercise minutes per week was 35.5 (SD 22) minutes. The mean move minutes per session was 17.7 (SD 11) minutes. The mean minutes per session was 18 (SD 1.4) minutes for school exercisers and 17 (SD 18) minutes for home exercisers, indicating variable responses from home exercisers. The mean rating of perceived exertion per exercise session was 4.3 (SD 2), indicating a moderate intensity that ranged from low to hard intensity. No adverse events or problems were identified. No improvements in walking endurance or hand-grip strength were observed. School exercisers achieved a higher attendance rate (83%) than home exercisers (27%; $P<.001$) and seemingly had a 2-fold increase in the volume of exercise achieved (school: mean 279, SD 55 min; 95% CI 212 - 347; home: mean 131, SD 170 min; 95% CI –140 to 401; $P=.10$). Qualitative themes relating to implementation factors and benefits to participant well-being were identified.

Conclusions: This study identified factors to inform an optimal protocol for implementing a high school–based VR exercise program for children with disabilities. Study findings demonstrated that moderate exercise at school is feasible in VR, but simply providing children with VR exergaming technology at home, without coaching, will not successfully engage them in exercise.

Trial Registration: ClinicalTrials.gov NCT05935982; https://clinicaltrials.gov/study/NCT05935982

(JMIR Form Res 2025;9:e65801) doi:10.2196/65801

XSL•FO
RenderX

KEYWORDS

disability; adapted physical activity; leisure-time exercise; exercise; telehealth; tele-exercise; cerebral palsy; pediatric rehabilitation; intellectual disability; developmental disability; child; high school; exercise; mobility; mobility disability; cardio; cardiorespiratory; cardiometabolic; feasibility; virtual reality; controlled trial; t test; VR; exergame; mixed method

Introduction

Children with disabilities generally participate in low volumes of physical activity [1,2], which places them at high risk for cardiometabolic disorders (eg, cardiovascular and metabolic diseases) [3-6]. Cardiometabolic disorders among people with disabilities seem to onset during the early adult years and manifest around the mid-30s at higher prevalences than people without disabilities [6-8]. Age-related trends in disease onset are believed to be linked with gradual reductions in health-enhancing doses of physical activity and accelerated physical deconditioning that is unique to people with disabilities [5,8,9]. Reduced participation in physical activity is suggested to be primarily caused by a lack of accessible, usable, and available opportunities and support within the community [10]. Given that children who participate in physical activity are more likely to participate as adults [11,12], there is a strong need to engage children with disabilities in regular physical activity behavior as a means of health prevention.

Outside of formal rehabilitation from physical and occupational therapy, the high school setting is often the last structured opportunity for physical activity in a person's lifetime. In the United States, physical activity opportunities in high school are mandated by federal law to be provided to all children with disabilities. Nevertheless, mandated opportunity does not necessarily mean that a child will have accessible and usable exercises that can lead to meaningful improvements in health. A systematic review has demonstrated that school-based exercise studies generally have not elicited favorable effects on cardiometabolic health among children with disabilities, but improvements can be observed in cardiorespiratory and muscular fitness and other components of health-related physical fitness [13]. These findings are also supported by more overarching reviews of disability exercise studies outside of the school setting [2]. Children with mobility disabilities who cannot walk, run, or cycle for prolonged periods have very few evidenced aerobic exercise modalities for improving their cardiometabolic health and cardiorespiratory fitness [2,14,15].

Active video gaming is one of the few evidenced aerobic exercise modalities that can improve health among children with disabilities [14,15]. Past active video gaming studies typically used console-based devices such as the Nintendo Wii or X-Box Connect. However, those devices have been discontinued by their manufacturer, and the market is dominated solely by virtual reality (VR) head-mounted displays (HMDs). HMDs emerged during the COVID-19 pandemic in 2019, with the release of the Oculus Quest (now referred to as Meta Quest). The Quest was the first all-in-one HMD that allowed a high-quality gaming experience, and it could be purchased at any major retailer at a price comparable to gaming consoles. In 2019, the first study to use a VR HMD among children with disabilities demonstrated that 2 children with spina bifida could obtain health-enhancing doses of moderate-intensity exercise at home with behavioral telehealth coaching. Since then, HMDs have been used in a variety of contexts, such as pediatric rehabilitation, pain management, and mindfulness training [16-19]. One trial is currently testing the preliminary efficacy of VR HMD exercise training on cardiometabolic health in children with cerebral palsy [20]. However, to the best of our knowledge, few studies have tested the implementation of a VR HMD exercise program for children with disabilities in high school, particularly using the latest HMD, the Meta Quest 3.

The purpose of this 1-year feasibility study was to work with a community engagement group to develop and test the implementation of a VR HMD exergaming program for children with special needs at a high school. The overall purpose of this study was to identify a replicable protocol that could be implemented within the daily operations of a high school through the following research questions:

- How much exercise can students feasibly fit within a school-based program?
- How well would students attend the program?
- How many students can be anticipated to be enrolled from a single school?
- What payment incentive would be required to satisfy enrollment targets and ensure adherence to data collection?
- What were the potential effect estimates of the program on physical fitness tests?
- How do outcomes compare between exercising in school or exercising at home by using a loaned headset by the school?
- Qualitatively, what critical factors should be considered in designing a larger trial?

Methods

Design

This feasibility study was a community-based participatory research design with 2 phases, namely, a development phase and an implementation phase. The development phase included a 4-month preparation phase, where study staff worked with a community engagement group to modify a working VR randomized controlled trial home-based protocol to a local public high school setting. The 9-month implementation phase used a "learn-as-you-go" approach with a single group pre- to posttrial design to test the implementation of the developed VR HMD school-based protocol. The study was implemented at a public high school within the United States, which had adapted education support for children with special needs.

Community Engagement Group

The community engagement group included a case manager of the high school, a young adult with cerebral palsy (male; 19 y of age) with mild intellectual disability and mobility disability who was a student at the school, and his caregiver. The group was involved in the study, starting from the design in the

XSL•FO
RenderX

development phase to the completion of the study in the implementation phase. The group met weekly with the research staff during the development phase, then biweekly during the first half of the implementation phase, and monthly thereafter. The purpose of the community engagement group was to make protocol modifications to intervention and data collection procedures of an existing clinical intervention [20] that would best fit in real-world, daily operations at the school.

Participants

This feasibility study aimed to recruit a convenient sample of 12 participants, which is a general recommendation for exploratory studies targeting implementation process metrics [21], as opposed to group-effect comparisons that would need larger samples [22]. Eligibility criteria are listed below. Participants were recruited through word-of-mouth and flyers by the community engagement group (case manager and young adult).

Inclusion criteria for the study included (1) between 10 and 19 years of age, (2) enrolled in school special education or self-reported disability, (3) a note from a physician permitting participation in moderate-intensity exercise or cleared for sports participation through physical examination that was on file by the school, and (4) ability to communicate in English (or a caregiver).

Exclusion criteria for the study included (1) complete blindness or deafness and (2) conditions that may make participation unsafe.

Supplies

Participants used a VR HMD for exercise (Quest 3, Meta, United States), which came with 2 hand-held controllers. Active video games were purchased through the Meta Oculus digital cloud server. The games were purchased for a standard Meta account created for the study. Games included Beat Saber, Les Mills Body Combat, Creed: Rise to Glory, Smash Drums, Power Beats VR, and Thrill of the Fight.

Intervention Development and Protocol

During the development phase of the study, the community engagement group worked with the research team to modify a home-based VR HMD clinical trial protocol [20] to be more easily implemented in a high school setting. In summary, the original intervention prescription included 150 minutes of moderate-intensity activity per week using the Quest 3 at home, for a total of 12 weeks. The home exercise was supported by theory-driven behavioral telecoaching calls [20]. In that study, caregivers were the primary intervention agents who managed their child's exercise schedules and study activities.

Through the 4-month development phase for this study, meetings with the community engagement group resulted in the following finalized exercise protocol to test in high school:

- Frequency: 3 times per week.
- Intensity: moderate intensity exercise as indicated by a rating of perceived exertion of 5 to 7 on the Borg 0 - 10 scale.
- Time: as many minutes of exercise as possible that could be fit within students' school schedules (eg, either within their physical education period or a study hall or gap period).
- Type: VR exercise at home or in the school setting, at the choice of the caregiver and participant. At enrollment, caregivers were asked whether they preferred the child to perform the exercise at school or home with a borrowed headset. Home-based participants were not given more attention than those who exercised at school. For example, no behavioral coaching calls were given to home participants as is done in the protocol from which this study was modified [20].

Intervention Procedures

Participants and their caregivers who were interested in the program were referred to study staff by the community engagement case manager. Participants were contacted by research staff and screened through phone calls. Eligible participants were directed to a web-based electronic data capture platform (REDCap [Research Electronic Data Capture]; Vanderbilt University) to provide informed consent and assent, participant demographics, and complete the physical activity survey. Next, they were scheduled for an on-site visit at the school to complete the physical fitness tests. Next, study staff scheduled the exercise sessions within a gap period or physical education period for participants who chose to exercise at school. Exercise sessions were performed at the school for 6 weeks, and all sessions were supervised by at least 1 research team member. For participants who chose to exercise at home, they were provided with a Quest 3 with the preinstalled games in a convenient carrying case. They were provided with the same exercise prescription as those who chose to exercise in the school. Home-based participants were not provided with additional attention or behavioral coaching. Quest 3 headsets were returned to the school once the home-based participant completed the 6-week program.

Measures

Participant Characteristics

Age, sex, height, weight, and ethnicity were descriptively reported for the enrolled participants. The use of a mobility device and wheelchair was recorded. As a gauge for how physically active participants were upon joining the study, physical activity level was measured using the National Institutes of Health Patient-Reported Outcomes Measurement Information System Pediatric Physical Activity Short Form 8-item survey [23]. The survey results in a total T-score. T-scores range from 20 to 80, with a mean of 50 (SD 10). A higher T-score indicates high levels of participation in physical activity. Scores that were 0.5 - 1 SD below the mean indicate "mild" difference or impairment; scores 1.0 - 2.0 SD below the mean indicate "moderate" impairment; and scores ≥2.0 SD below the mean indicate "severe" impairment.

Research Question 1: Exercise Output

Exercise volume was measured weekly by 2 variables, namely, mean "move" minutes and rating of perceived exertion. Move minutes were automatically recorded by the Quest 3. The Quest

XSL•FO
RenderX

3 includes a built-in fitness tracker app, referred to as Meta Quest Move. The app tracks calories and move minutes, and these data are stored in cloud-based Meta servers through end-to-end encryption. Meta does not provide specifics on how these data are calculated, but the calculations are reported by Meta to be estimated through an algorithm that includes basal metabolic rate along with the headset, controller, and hand movements. At school, the maximum possible number of active minutes that a participant could exercise each week was 75 minutes because of designated 3-hour class periods. Home-based exercise participants were asked to match the school prescription (3 sessions per week for a total of 75 min). Mean move minutes were averaged each week across the 6-week program. To provide a measure of the intensity of the exercise, participants were asked immediately after exercise to provide a single rating of perceived exertion score that summarized each session, using the Borg 0 - 10 scale [24]. In summary, a score of 0 represents "nothing at all," a score of 3 is indicative of moderate intensity, a score of 5 indicates "strong" exercise, a score of 7 indicates "very strong," and 10 indicates "extremely strong." The mean calories were also used as an indicator of exercise output. Adverse events or problems were reported and documented in accordance with the policies established by the university.

Research Question 2: Student Attendance

Student attendance was expressed as a percentage value, measured by the number of exercise sessions attended, divided by the total sessions prescribed. The exercise prescription was a total of 18 sessions, broken up into 3 sessions per week for 6 weeks.

Research Questions 3 and 4: Student Enrollment and Payment Incentives

Student enrollment was measured by the number of students enrolled per school semester. The program lasted a total of 2 school semesters and was completed prior to summer break. Payment incentives were planned to be increased if the target sample size of 12 was not met from the first school semester. Payment incentives started at US $50 for each of the 2 data collection visits, for a total of US $100. Because of inadequate enrollment in the first semester, the payment incentive for the second semester was increased to US $100 per data collection visit, for a total of US $200.

Research Question 5: Potential Effect Estimates of the Program on Physical Fitness Tests

Physical fitness was measured by walking endurance and hand-grip strength. Walking endurance was measured using a 6-minute walk test (6MWT). A 6-minute push test was planned to be used for students who used wheelchairs, but this was never used. The 6MWT measures the distance walked within 6 minutes around a designated course. The course was a long unobstructed hallway that allowed the participant to freely walk for the duration of the test. When participants reached the end of the hall, they were instructed to turn and continue down the hallway. The rationale for choosing this location was that the study staff found there was limited space to set up and conduct a larger circular track during school hours, particularly without interference from other students. The 6MWT is suggested to be valid and reliable among children without disabilities [25]. However, the 6MWT has conflicting evidence regarding validity for various pediatric clinical groups [26,27]. Hand-grip strength was measured by the mean value of 3 maximal grip trials on a hand-held dynamometer (Camry Digital Hand Dynamometer, Camry, United States), while the participant was in a seated position. Grip-strength has evidence to support its use among children with intellectual disabilities [27].

Research Question 6: Ideal Exercise Environment

To identify the ideal method for how schools can implement the exercise program for students in a future trial, outcomes were compared between exercise settings (home vs school). Compared outcomes included enrollment, exercise volume, attendance, and fitness tests.

Research Question 7: Qualitative Evaluation of Factors for Future Trials

The qualitative component of this study used a mini-ethnographic design, whereby the interventionists were immersed in the culture at school with participants. At postintervention, the interventionists were interviewed (semistructured) to gain insight into feasibility issues and possible programs that should be considered in a future trial. Specific questions were: (1) What were critical issues that impeded or facilitated the implementation of the program at school? and (2) What potential health or fitness-related benefits did participants experience throughout the program (particularly, benefits that were not measured)? When considering these questions, the interventionists were asked to recall and talk through their experience with each of the on-site exercise students. Interventionists were asked to take notes throughout the study. Interviews were audio recorded and analyzed by the lead interventionist, a qualitative analyst (BL), who had conducted more than 400 interviews related to disability and health.

Analyses

Descriptive statistics, including means, SDs, and CIs, were obtained for all study variables. Changes in means were compared primarily using descriptive comparisons with 95% CIs, with 2-tailed *t* tests as appropriate. Analyses were performed using SPSS Statistics (version 29; IBM Corp). Qualitative data were analyzed using a thematic analysis approach [19].

Ethical Considerations

The study was approved by the University of Alabama at Birmingham Institutional Review Board for Human Use prior to enrolling participants (FWA00005960). Participants had to provide informed consent and assent documentation prior to joining the study. The principal of the high school approved and supported the study. Participants were compensated for participation in the data collection (US $100 per data collection, total incentive of US $200). All data were deidentified.

XSL•FO
RenderX

Results

Participant Characteristics

A total of 10 participants enrolled in the program over the 2 semesters. One participant lost interest in joining the program and dropped out prior to starting the program. A total of 9 participants completed the data collection. A total of 5 (56%) participants chose to exercise at school, and 4 (44%) participants chose to exercise at home. Characteristics for the 9 completers are provided in Table 1. The mean age of completers was 17 (SD 0.6) years (6 male and 3 female participants). The mean total Patient-Reported Outcomes Measurement Information System physical activity T-score was 41.9 (SD 7.2), indicating that physical activity levels were mild to moderately below average. Only 1 participant used a mobility device, ankle-foot orthotics. None of the participants used a wheelchair.

Table . Participant characteristics.

ID	Age (years)	Sex	Hispanic or Latino	Ethnicity	Height (inches)	Weight (lbs)	Primary disability	Physical activity (T-score)	Exercise setting
1	17	Female	No	White	56	225	Down syndrome	43.1	School
2	18	Male	No	White	60	190	Autism	28.8	School
3	17	Male	No	White	68	195	Brain tumor	48.8	School
4	16	Male	No	White	61	145	Autism or ADHD[a]	44.6	School
5	16	Female	No	White	62	105	Unknown	43.1	School
6	17	Male	No	White	68	130	Unknown	40.4	Home
7	17	Female	No	White	62	125	ADHD	48.3	Home
8	17	Male	Yes	White	68	126	ADHD	31.8	Home
9	17	Male	No	White	67	114	Cerebral palsy	48.4	Home

[a]ADHD: attention-deficit/hyperactivity disorder.

Implementation Observations

At the high school, all classes within students' schedules were at the longest, 50-minute periods, and the shortest, 25-minute periods. All students exercised during their study hall period. The community engagement group and research staff chose to structure a 25-minute exercise session within the students' 1-hour study hall to allow the students to have 25 minutes to spend on completing academic work. Students did not have physical education built into their class schedule, because they had completed the 1-year physical education requirement prior to joining the study. The education guidelines of the state required only 1 year of physical education across the students' 4 years of education. The exercise setting was typically conducted in an available classroom, which sometimes required desks and chairs to be moved to allow adequate space for gameplay. The school information technology staff created a secure Wi-Fi for the headsets to log in to. School Wi-Fi was generally protected to prevent inappropriate internet use. The VR headsets required Meta accounts with an active email address. Research staff created the accounts and preinstalled games onto the accounts prior to the intervention.

Exercise Volume Results

From the 9 completers (people who completed the data collections), the total mean move minutes across the 6 weeks was 213 (SD 136), ranging from 0 to 377 minutes. The mean move minutes per week was 35.5 (SD 22) minutes. The mean move minutes per session was 17.7 (SD 11) minutes. Excluding a person from the home exercise group who attended no sessions, the mean minutes per session was 19.9 (SD 11) minutes. The mean rating of perceived exertion per exercise session was 4.3 (SD 2), indicating that most exercise sessions were performed at a moderate intensity, which ranged from low to hard intensity. The mean calories expended per attended session was 90.9 (SD 53.8), ranging from 33 to 196. No adverse events or problems (eg, seizures, injuries, or incidental behavioral issues between participants and staff) occurred throughout the study.

Attendance and Enrollment Results

The mean attendance for all 9 completers was 61.1% (mean of 11 sessions attended out of 18 total prescribed sessions). One completer attended zero sessions but completed the pre- and postdata collections. Two participants were enrolled during the first school semester, with a payment incentive of US $50 per data collection (total incentive of US $100). The remaining 8 participants were enrolled during the second school semester after the payment incentive was increased to US $100 per data collection (total incentive of US $200). Recruitment efforts by the research team and community engagement group were consistent between semesters.

Physical Fitness Results

Mean 6MWT distances for the 9 completers were 447 (SD 94; range: 305 - 567 meters; 95% CI 375 - 518) and 469 (SD 78; range: 350 - 568 meters; 95% CI 408 - 529) at pre- and postintervention, respectively. The mean difference in 6MWT

XSL•FO
RenderX

distance from pre to post was 22 (SD 52; t_8=1.3; *P*=.24; 95% CI −17.7 to 61.8), indicating a possible trend toward improvement but no statistically significant change across time. Left grip strength at pre and postintervention was 24.9 (SD 8; range: 9 - 35; 95% CI 19 - 32) and 24 (SD 9; range: 8 - 39; 95% CI 17 - 31), respectively. Right grip strength at pre and postintervention was 26.7 (SD 8; range: 12 - 40 kg; 95% CI 20 - 33) and 27.7 (SD 10; range: 11 - 47 kg; 95% CI 20 - 35), respectively. These data indicated no difference in grip strength from pre- to postintervention for both hands.

Comparing School and Home Settings

There did not appear to be a difference in baseline levels of physical activity participation between groups (school exercisers: mean T-score 41, SD 8; n=5; home exercisers: mean T-score 42.2, SD 7; n=4). Intervention attendance was higher among those who exercised at school (83% attendance; mean sessions attended 15, SD 2; 95% CI 13 - 18) compared with home exercisers (27% attendance; mean sessions attended 4.8, SD 4; 95% CI –1 to 11), and 95% CIs indicated a difference between settings of 6 - 15 sessions that were statistically significant (*P*<.001; mean difference=10.7 sessions). Reasons for nonattendance were mostly due to illness and detention, in response to poor behavior at school. The mean total move minutes among the school exercisers (279, SD 55 minutes; 95% CI 212 - 347) was approximately a 2-fold increase over the home exercisers (131, SD 170 min; 95% CI –140 to 401). The 95% CIs indicated a difference between settings of –39 to 337 minutes in favor of the home exercisers, but this difference was not statistically significant (*P*=.10; mean difference=149 min). The mean minutes per session was 18 (SD 1.4) minutes for school exercisers and 17 (SD 18) minutes for home exercisers, indicating variable responses from home exercisers. There did not appear to be any visual or statistical changes between the home and school settings for physical fitness tests.

Qualitative Evaluation of School Program Benefits

Given the small number of on-site school participants, only 4 themes resulted from thematic analysis. Two themes were related to program benefits, and 2 themes were related to implementation. First, gaming at school aided well-being. Students reported that an exergaming break amidst a full school schedule made them feel better. Specifically, gaming provided a mental break from schoolwork and social interaction stressors (eg, negatively perceived interactions with peers without disabilities). Second, participation in moderate-intensity exergaming increased energy levels throughout the program. For example, halfway through the program, 1 participant who was heavily engaged in the exercise was notably less exhausted when walking in between classrooms, and the participant took less rest in between exercise bouts. This benefit was observed halfway through the 6-week program. Moreover, participants noticeably increased their exercise intensity and gaming skills across the program.

Third, interventionists reported that supervision was necessary for the exercise sessions. Students sometimes require setup and management of their exercise space. For example, some students exercised in a classroom, which required assistance in moving around chairs to make sufficient space for gaming. In 1 case, a participant who was highly engaged in the exercise needed to be prevented from overexertion. The participant was demonstrating overexertion symptoms (eg, face turning red, breathing heavily, and balance becoming unstable), and this required interventionists to urge the participant to sit down and slow down on the exergaming movements. Fourth, designated school resources are critical factors for successful implementation. Space for exercise was limited within the high school. The school gymnasium was often not a suitable place for exergaming, considering that the space often had verbal and physical interference (accidental balls and objects rolling or flying into the play space) from other students. Empty classrooms were often the primary exercise scenarios. Interventionists noted that a designated space for exergaming would allow more time for exercise by reducing the time needed to set up a play space. Moreover, having a designated wireless internet connection specifically for the headsets aided implementation. The school staff assisted with creating a wireless network access log-in that was made specifically for the project. Although internet access was not necessary to play the games in the Quest 3 headset, games often required software updates that would alleviate bugs or issues in gameplay.

Discussion

Principal Findings

This study developed and piloted a school-modified VR active video gaming program in a high school among students with disabilities. Several key findings were identified that will inform a larger clinical trial. First, this study identified a feasible dose of exercise that can be obtained at school. Without making substantial changes to a student's school schedule, 30 minutes in a single day is likely the most that can be inserted into a student's schedule, specifically in their study hall period. This dose is similar to other feasibility studies attempting to insert external exercise programs into school settings [13,28,29]. Most schools in the United States require physical education, but requirements vary widely. The high school in this study required only 1 year of physical education, in accordance with requirements by the state. Considering these constraints and exercise setup (eg, equipping the headset, clearing space to play, and starting a game) and cooldown allowed students at school to achieve 18 minutes of moderate-intensity aerobic exercise each session. It is important to note that 18 minutes, done 3 times a week or even daily, is far lower than the national guidelines of 60 minutes of moderate to vigorous activity per day for children or even 150 minutes of moderate aerobic exercise per week for adults [30]. Nevertheless, 18 minutes of moderate exercise done 3 times per week may elicit health benefits among people with disabilities. A review of meta-analyses demonstrated that people with disabilities can obtain health benefits at far lower doses of exercise than those noted in adult guidelines [2].

Another notable finding was that the school was an ideal setting for both recruitment and exercise. Obtaining 9 students from a single high school is encouraging, considering that the mean sample size for randomized controlled trials of exercise among children with disabilities is approximately 30 children [14]. To

XSL•FO
RenderX

foster strong recruitment, study findings suggest a minimum of a US $200 payment incentive for a total of 2 data collections. The US $200 payment incentive resulted in 77% (7/9) of participants from a single school semester, whereas a US $100 incentive resulted in only 2 enrolled participants for a semester. Moreover, students who exercised with supervision at school achieved strong attendance (83%; 15 out of 18 sessions), particularly when compared with nonsupervised home-based programs (27%; 4.8 out of 18 sessions). Although supervision places a greater burden on school or research staff, the differences in attendance likely justify the burden. Study findings demonstrate that simply providing children with VR exergaming technology will not successfully engage them in exercise volumes that can likely lead to health benefits.

Limitations

Regarding health and well-being benefits, a limitation of this study was the inclusion of physical performance tests instead of more concrete measures of health. Results on physical performance tests can vary due to mobility, intellectual, and behavioral differences among children with disabilities. Health assessments with stronger psychometric properties will be necessary to further investigate the health benefits of VR active gaming in school settings. Examples of more robust assessments could include a graded exercise test to measure peak oxygen consumption as an indicator of aerobic fitness, or the inclusion of biomarkers (eg, blood spot tests or blood draws) to measure changes in cardiometabolic health. Measures to support the qualitative findings could include surveys or biomarkers of well-being, stress, and anxiety. Interviews could have been conducted directly with students for more in-depth data. Another study limitation was the short 6-week duration of the intervention. Although 6 weeks of structured exercise may be sufficient to improve cardiorespiratory fitness, changes in cardiometabolic health are much harder to influence from exercise and generally require higher durations lasting several months with longer volumes of weekly exercise [2,13,29]. Future trials aiming to incorporate longer intervention durations in school settings should start participants in exercise as early as possible in a school semester, which typically lasts approximately 18 weeks. Another limitation was the small sample size and quasi-experimental design. Study findings warrant confirmation in larger trials that are randomized.

Conclusions

This study identified a feasible VR program dose that could be fit within the daily activities of a public high school in the Southeast United States. Future research is needed to identify optimal exercise doses that can improve the physical and cardiometabolic health of children with disabilities at school, given the structural and logistical constraints of implementation.

Acknowledgments

This study was supported by grant funding provided by the Forge AHEAD Center and the Center for Engagement in Disability Health and Rehabilitation Sciences (CEDHARS), both of which were located at the University of Alabama at Birmingham. Research reported in this publication was supported by the National Institute on Minority Health and Health Disparities of the National Institutes of Health under award P50MD017338. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We would like to thank the community engagement group for their contribution to this project and give special thanks to Pelham High School for supporting the study. No artificial intelligence was used in the writing or development of this manuscript.

The Pelham High Community Engagement Group includes Dorothy Coogan, Jason Harchuck, and Larsen Bright.

Data Availability

The datasets generated during or analyzed during this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

References

1. Carlon SL, Taylor NF, Dodd KJ, Shields N. Differences in habitual physical activity levels of young people with cerebral palsy and their typically developing peers: a systematic review. Disabil Rehabil 2013 Apr;35(8):647-655. [doi: 10.3109/09638288.2012.715721] [Medline: 23072296]
2. Ginis KAM, van der Ploeg HP, Foster C, et al. Participation of people living with disabilities in physical activity: a global perspective. Lancet 2021 Jul 31;398(10298):443-455. [doi: 10.1016/S0140-6736(21)01164-8] [Medline: 34302764]
3. Messiah SE, Vidot DC, Somarriba G, et al. Obesity and cardiometabolic disease risk factors among US adolescents with disabilities. World J Diabetes 2015 Feb 15;6(1):200-207. [doi: 10.4239/wjd.v6.i1.200] [Medline: 25685291]
4. Dhanasekara CS, Ancona D, Cortes L, et al. Association between autism spectrum disorders and cardiometabolic diseases: a systematic review and meta-analysis. JAMA Pediatr 2023 Mar 1;177(3):248-257. [doi: 10.1001/jamapediatrics.2022.5629] [Medline: 36716018]

XSL•FO
RenderX

5. Ryan JM, Crowley VE, Hensey O, Broderick JM, McGahey A, Gormley J. Habitual physical activity and cardiometabolic risk factors in adults with cerebral palsy. Res Dev Disabil 2014 Sep;35(9):1995-2002. [doi: 10.1016/j.ridd.2014.03.051] [Medline: 24864052]
6. Heyn PC, Tagawa A, Pan Z, Thomas S, Carollo JJ. Prevalence of metabolic syndrome and cardiovascular disease risk factors in adults with cerebral palsy. Develop Med Child Neuro 2019 Apr;61(4):477-483. [doi: 10.1111/dmcn.14148]
7. Peterson MD, Kamdar N, Hurvitz EA. Age-related trends in cardiometabolic disease among adults with cerebral palsy. Dev Med Child Neurol 2019 Apr;61(4):484-489. [doi: 10.1111/dmcn.13777] [Medline: 29704244]
8. McPhee PG, MacDonald MJ, Cheng JL, Dunford EC, Gorter JW. Emerging evidence for accelerated ageing and cardiovascular disease in individuals with cerebral palsy. J Rehabil Med 2019 Jul 8;51(7):525-531. [doi: 10.2340/16501977-2564] [Medline: 31120542]
9. Rimmer JH, Schiller W, Chen MD. Effects of disability-associated low energy expenditure deconditioning syndrome. Exerc Sport Sci Rev 2012 Jan;40(1):22-29. [doi: 10.1097/JES.0b013e31823b8b82] [Medline: 22016146]
10. Ginis KAM, Ma JK, Latimer-Cheung AE, Rimmer JH. A systematic review of review articles addressing factors related to physical activity participation among children and adults with physical disabilities. Health Psychol Rev 2016 Dec;10(4):478-494. [doi: 10.1080/17437199.2016.1198240] [Medline: 27265062]
11. Dohle S, Wansink B. Fit in 50 years: participation in high school sports best predicts one's physical activity after age 70. BMC Public Health 2013 Dec 1;13:1100. [doi: 10.1186/1471-2458-13-1100] [Medline: 24289060]
12. Telama R, Yang X, Viikari J, Välimäki I, Wanne O, Raitakari O. Physical activity from childhood to adulthood: a 21-year tracking study. Am J Prev Med 2005 Apr;28(3):267-273. [doi: 10.1016/j.amepre.2004.12.003] [Medline: 15766614]
13. Manojlovic M, Roklicer R, Trivic T, et al. Effects of school-based physical activity interventions on physical fitness and cardiometabolic health in children and adolescents with disabilities: a systematic review. Front Physiol 2023;14:1180639. [doi: 10.3389/fphys.2023.1180639] [Medline: 37362446]
14. Lai B, Lee E, Wagatsuma M, et al. Research trends and recommendations for physical activity interventions among children and youth with disabilities: a review of reviews. Adapt Phys Activ Q 2020 Apr 1;37(2):211-234. [doi: 10.1123/apaq.2019-0081] [Medline: 32131052]
15. Lai B, Lee E, Kim Y, et al. Leisure-time physical activity interventions for children and adults with cerebral palsy: a scoping review. Dev Med Child Neurol 2021 Feb;63(2):162-171. [doi: 10.1111/dmcn.14751] [Medline: 33241561]
16. Tobaiqi MA, Albadawi EA, Fadlalmola HA, Albadrani MS. Application of virtual reality-assisted exergaming on the rehabilitation of children with cerebral palsy: a systematic review and meta-analysis. J Clin Med 2023 Nov 14;12(22):7091. [doi: 10.3390/jcm12227091] [Medline: 38002703]
17. Lai B, Powell M, Clement AG, Davis D, Swanson-Kimani E, Hayes L. Examining the feasibility of early mobilization with virtual reality gaming using head-mounted display and adaptive software with adolescents in the pediatric intensive care unit: case report. JMIR Rehabil Assist Technol 2021 May 27;8(2):e28210. [doi: 10.2196/28210] [Medline: 34042602]
18. Lai B, Davis D, Narasaki-Jara M, et al. Feasibility of a commercially available virtual reality system to achieve exercise guidelines in youth with spina bifida: mixed methods case study. JMIR Serious Games 2020 Sep 3;8(3):e20667. [doi: 10.2196/20667] [Medline: 32880577]
19. Lai B, Davis D, Young R, et al. Group telegaming through immersive virtual reality to improve mental health among adolescents with physical disabilities: pre- and posttrial protocol. JMIR Res Protoc 2022 Oct 13;11(10):e42651. [doi: 10.2196/42651] [Medline: 36194864]
20. Lai B, Davis D, Young R, et al. The effects of virtual reality tele-exergaming on cardiometabolic indicators of health among youth with cerebral palsy: protocol for a pilot randomized vontrolled trial. JMIR Res Protoc 2022 Aug 17;11(8):e40708. [doi: 10.2196/40708] [Medline: 35976192]
21. Julious SA. Sample size of 12 per group rule of thumb for a pilot study. Pharm Stat 2005 Oct;4(4):287-291. [doi: 10.1002/pst.185]
22. Teresi JA, Yu X, Stewart AL, Hays RD. Guidelines for designing and evaluating feasibility pilot studies. Med Care 2022 Jan 1;60(1):95-103. [doi: 10.1097/MLR.0000000000001664] [Medline: 34812790]
23. Tucker CA, Bevans KB, Becker BD, Teneralli R, Forrest CB. Development of the PROMIS pediatric physical activity item banks. Phys Ther 2020 Aug 12;100(8):1393-1410. [doi: 10.1093/ptj/pzaa074] [Medline: 32313952]
24. Borg E, Kaijser L. A comparison between three rating scales for perceived exertion and two different work tests. Scand J Med Sci Sports 2006 Feb;16(1):57-69. [doi: 10.1111/j.1600-0838.2005.00448.x] [Medline: 16430682]
25. Li K, Cardoso C, Moctezuma-Ramirez A, Elgalad A, Perin E. Heart rate variability measurement through a smart wearable device: another breakthrough for personal health monitoring? Int J Environ Res Public Health 2023 Dec 6;20(24):7146. [doi: 10.3390/ijerph20247146] [Medline: 38131698]
26. Bartels B, de Groot JF, Terwee CB. The six-minute walk test in chronic pediatric conditions: a systematic review of measurement properties. Phys Ther 2013 Apr;93(4):529-541. [doi: 10.2522/ptj.20120210] [Medline: 23162042]
27. Wouters M, Evenhuis HM, Hilgenkamp TIM. Systematic review of field-based physical fitness tests for children and adolescents with intellectual disabilities. Res Dev Disabil 2017 Feb;61:77-94. [doi: 10.1016/j.ridd.2016.12.016] [Medline: 28064026]

XSL•FO
RenderX

28. Bellamy J, Broderick C, Hardy LL, et al. Feasibility of a school-based exercise intervention for children with intellectual disability to reduce cardio-metabolic risk. J Intellect Disabil Res 2020 Jan;64(1):7-17. [doi: 10.1111/jir.12690] [Medline: 31625658]
29. Wallén EF, Müllersdorf M, Christensson K, Marcus C. A school-based intervention associated with improvements in cardiometabolic risk profiles in young people with intellectual disabilities. J Intellect Disabil 2013 Mar;17(1):38-50. [doi: 10.1177/1744629512472116] [Medline: 23322342]
30. Piercy KL, Troiano RP, Ballard RM, et al. The physical activity guidelines for Americans. JAMA 2018 Nov 20;320(19):2020-2028. [doi: 10.1001/jama.2018.14854] [Medline: 30418471]

Abbreviations

6MWT: 6-minute walk test
HMD: head-mounted display
REDCap: Research Electronic Data Capture
VR: virtual reality

Edited by A Mavragani; submitted 26.08.24; peer-reviewed by K Yuenyongchaiwat, N Charususin, S DeLuca; revised version received 12.03.25; accepted 14.03.25; published 25.04.25.

Please cite as:
Lai B, Wright A, Hutchinson B, Bright L, Young R, Davis D, Malik SA, Rimmer JH, Pelham High Community Engagement Group School-Based Virtual Reality Programming for Obtaining Moderate-Intensity Exercise Among Children With Disabilities: Pre-Post Feasibility Study
JMIR Form Res 2025;9:e65801
URL: https://formative.jmir.org/2025/1/e65801
doi:10.2196/65801

© Byron Lai, Ashley Wright, Bailey Hutchinson, Larsen Bright, Raven Young, Drew Davis, Sultan Ali Malik, James H Rimmer, Pelham High Community Engagement Group. Originally published in JMIR Formative Research (https://formative.jmir.org), 25.4.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Discussions of Antibiotic Resistance on Social Media Platforms: Text Mining and Mixed Methods Content Analysis Study

Jocelyne Arquembourg[1*], PhD; Philippe Glaser[2*], PhD; France Roblot[3*], PhD, MD; Isabelle Metzler[4], MSc; Mélanie Gallant-Dewavrin[5], MSc; Hugues Feutze Nanguem[6,7], MPA; Adel Mebarki[8], MBA; Paméla Voillot[8*], MSc; Stéphane Schück[8], MD

[1]Laboratoire l3, Telecom Paris, Telecom Paris, Palaiseau, France
[2]Unité EERA Institut Pasteur, CNRS UMR3525, Université de Paris, Paris, France
[3]CHRU Poitiers, INSERM U1070, CHRU Poitiers, Poitiers, France
[4]Association France Spondyloarthrites, Association France Spondyloarthrites, Tulle, France
[5]Association HTaPFrance, Association HTaPFrance, Meursault, France
[6]CREDIMI: Center for Research on International Market and Investment Law, University of Burgundy, Dijon, France
[7]Pfizer, Paris, France
[8]Kap Code, 146 Rue Montmartre, Paris, France
[*]these authors contributed equally

Corresponding Author:
Paméla Voillot, MSc
Kap Code, 146 Rue Montmartre, Paris, France

Abstract

Background: With the increasing popularity of web 2.0 apps, social media has made it possible for individuals to post messages on antibiotic ineffectiveness. In such online conversations, patients discuss their quality of life (QoL). Social media have become key tools for finding and disseminating medical information.

Objective: To identify the main themes of discussion, the difficulties encountered by patients with respect to antibiotic ineffectiveness and the impact on their QoL (physical, psychological, social, or financial).

Methods: A noninterventional retrospective study was carried out by collecting social media posts in French language written by internet users mentioning their experience with antibiotics, and the impact of their ineffectiveness on their QoL. Messages posted between January 2014 and July 2020 were extracted from French-speaking publicly available online forums.

Results: A total of 3773 messages were included in the analysis corpus after extraction and filtering. These messages were posted by 2335 individual web users, most of them being women around 35 years of age. Inefficacy of treatment options and the lack of information regarding the use of antibiotics were among the most discussed topics. QoL was discussed in 63% of the 3773 messages posted. The most common is the physical impact (78%). Patients discussed the persistence of symptoms and adverse effects. The second kind of impact is psychological (65%), characterized by feelings of anxiety or despair about the situation.

Conclusions: This social media analysis allowed us to identify a strong impact of the perceived ineffectiveness of antibiotic therapy on patients' daily life particularly in terms of physical and psychological consequences. These results provide health care experts information directly generated by patients regarding their own experiences. Social media studies constitute a complementary source of evidence that could be used to optimize messages to the public about appropriate use of antibiotics.

(JMIR Form Res 2025;9:e37160) doi:10.2196/37160

KEYWORDS

antibiotic ineffectiveness; antibiotic resistance; health-related quality of life; real-world; social media; quality of life; quantitative; qualitative; app; application; online; medical information; French; users; antibiotic; social media use

Introduction

When antibiotics were first introduced in the middle of the 20th century, they were hailed as wonder drugs. Patients and physicians alike were amazed at the almost miraculous effect of these drugs on serious bacterial infections. For the past 70 years, physicians have come to expect that antibiotics would cure almost all their patient's bacterial infections, and patients expect that the miracle drugs will still work wonders. Despite an increasing rate of antibiotic resistance globally, they remain

XSL•FO
RenderX

one of the most cost-effective, life-saving medicines contributing to an extended lifespan [1].

Commonly prescribed antibiotics are becoming less effective in everyday populations over time as a result of inappropriate prescriptions, unnecessary use, or their overuse in both humans and animals. At the same time, research and development on new antibiotics has slowed down [2]. It is to note that almost all the new antibiotics that have been brought to market in recent decades are variations of antibiotic drugs classes that had been discovered by the 1980s.

The European Surveillance of Antimicrobial Consumption Network showed that large variations in antibiotic use and antibiotic resistance rates exist across Europe, with higher use in Southern Europe including France and lower in Northern Europe [3].

In general, social media refers to forms of electronic communication (such as websites for social networking and microblogging) through which users create online communities to share information, opinions, personal messages, photos, videos, and other contents within internet apps [4]. Laudon and Traver [5] describe online social networking as an online social area for people who share common ties that can interact with one another. Organizations are able to use the internet to connect with consumers in the health care field. Consumers heavily rely on information found online and use internet to gather health care information and connect with other patients to garner support and learn about similar conditions. Individuals will use social media to post reviews or other comments that support or possibly deter others from choosing that type of health care in the future. It is essential for providers to be active on social media and provide accurate information and connect with readers. Certain groups use online channels more often depending on the topic they are searching for. And people responsible for others, such as parents or caregivers, search health topics even more. Almost 90% of adult users accessed popular social media sites to find and share health information [6]. People aged 18 to 24 years are twice as likely to use social media to discuss health issues than people aged between 45 to 54 years [7].

The digitizing of health care data, as well as advancements in computer processing and data storage, has enabled the development of advanced algorithms in the form of artificial intelligence (AI) [8]. AI tools are strongly related with data mining and AI is nowadays ranked among the top-10 technology, whichever the application [9]. Despite their limitations, AI tools and techniques that are still in their infancy already provide substantial benefits in providing in-depth knowledge on individuals' health and predicting population health risks. Their use for medicine and public health is expected to increase substantially in the near future [10].

Health is a multifaceted concept and there is no single indicator that can adequately assess its impact in relation to quality of life (QoL). The framework for measuring the QoL includes, among others, information on the share of the population with unmet needs for medical and dental care and a contrasting set of health determinant indicators [11]. In this study, we collected on the web and sought to analyze the messages and complaints of patients with antibiotic treatment failure using innovative tools, without prejudging the cause of the failure (misuse or antibiotic resistance) in order to identify the main topics of discussion, and to identify the difficulties encountered by patients regarding antibiotic ineffectiveness and the impact on their quality of life (physical, psychological, social, or financial).

Methods

Study Design and Population

This was a noninterventional retrospective study using a text mining approach to retrieve information from social media posts (data available in the public domain) written by French speaking internet users between 2014 and 2020 (Figure 1). The study was conducted in 2 phases, data collection using the published Detec't webcrawler [12,13] developed by Kap Code (Paris, France) to collect antibiotic resistance-related posts and quantitative and qualitative analyses to identify trends and characterize key topics discussed by French speaking users.

Figure 1. Study framework.

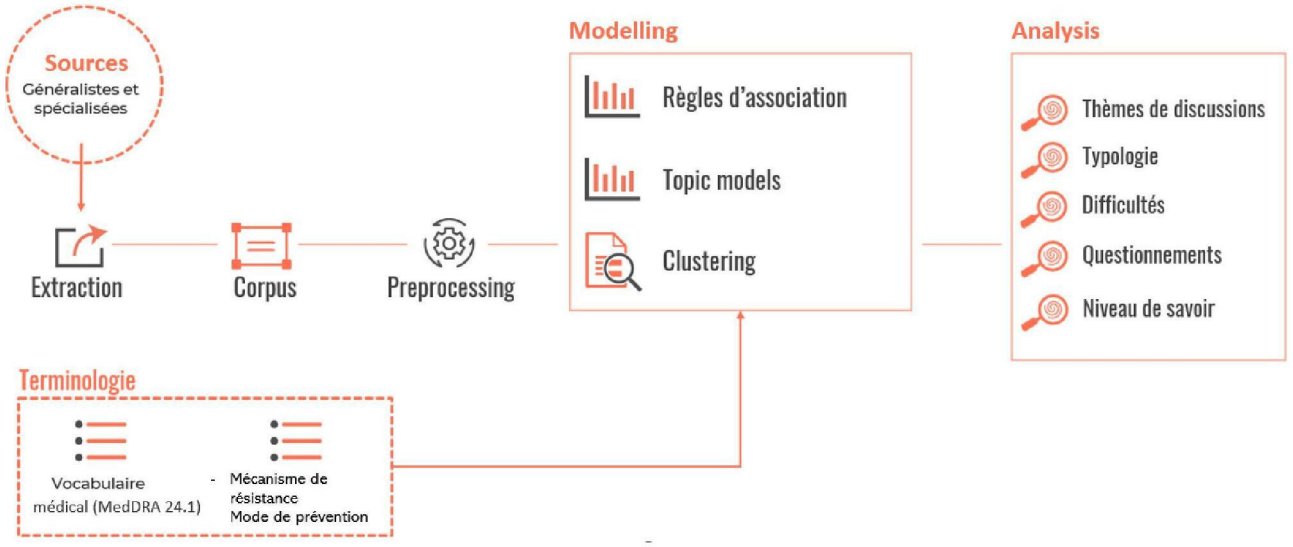

XSL•FO
RenderX

Data Extraction

A web crawler is an engine that browses through hyperlinks and stores them for future download of associated web pages (identified by the visited hyperlinks) [14]. Scraping of messages was performed according to the HTML structure of each forum. All discussions containing at least one of the keywords or one of their synonyms, coupled with the search for a lexical field of inefficiency within the same message (eg, inefficient or does not work), were automatically retrieved with all the associated metadata and anonymized before being stored in a study-specific database.

Messages associated with antibiotic's supposed ineffectiveness were retrieved from 46 French sources (such as Doctissimo [Reworld Media], Twitter [rebranded as X], and listed in Table 1) dating from 2014 to 2020.

Table . Data sources. Messages about lack of efficacy of antibiotics and antibiotic resistance were retrieved from 46 French sources.

Forums	Posts, n	Users, n
Twitter (X)	888	475
Doctissimo	885	621
Atoute	396	131
Futura-Sciences	260	135
Au féminin	171	126
Meamedica	125	124
Les impatientes	123	54
Psychologies	94	43
Onmeda	94	85
Journal des femmes	79	62
Maladies rares	77	67
Thyroide	72	46
YouTube	71	68
Psychoactif	68	42
Lymphome espoiri	61	41
Vinted	53	40
Vulgaris Medical	49	38
Ligue contre le cancer	47	43
Andlil	39	18
Alarm.asso	38	15
Maman pour la vie	37	29
Pedia Blog	13	7
Forum HardWare	10	7
Other sites	23	18
Total	3773	2335

The analysis corpus (Figure 2) consisted of the corpus filtered after the removal of messages containing predetermined keywords written in a language other than French, post containing animal-related vocabulary, and messages containing at least 1 of the study-specific exclusion words listed in Multimedia Appendix 1.

XSL•FO
RenderX

Figure 2. Flowchart presenting the steps for creating the analysis corpus.

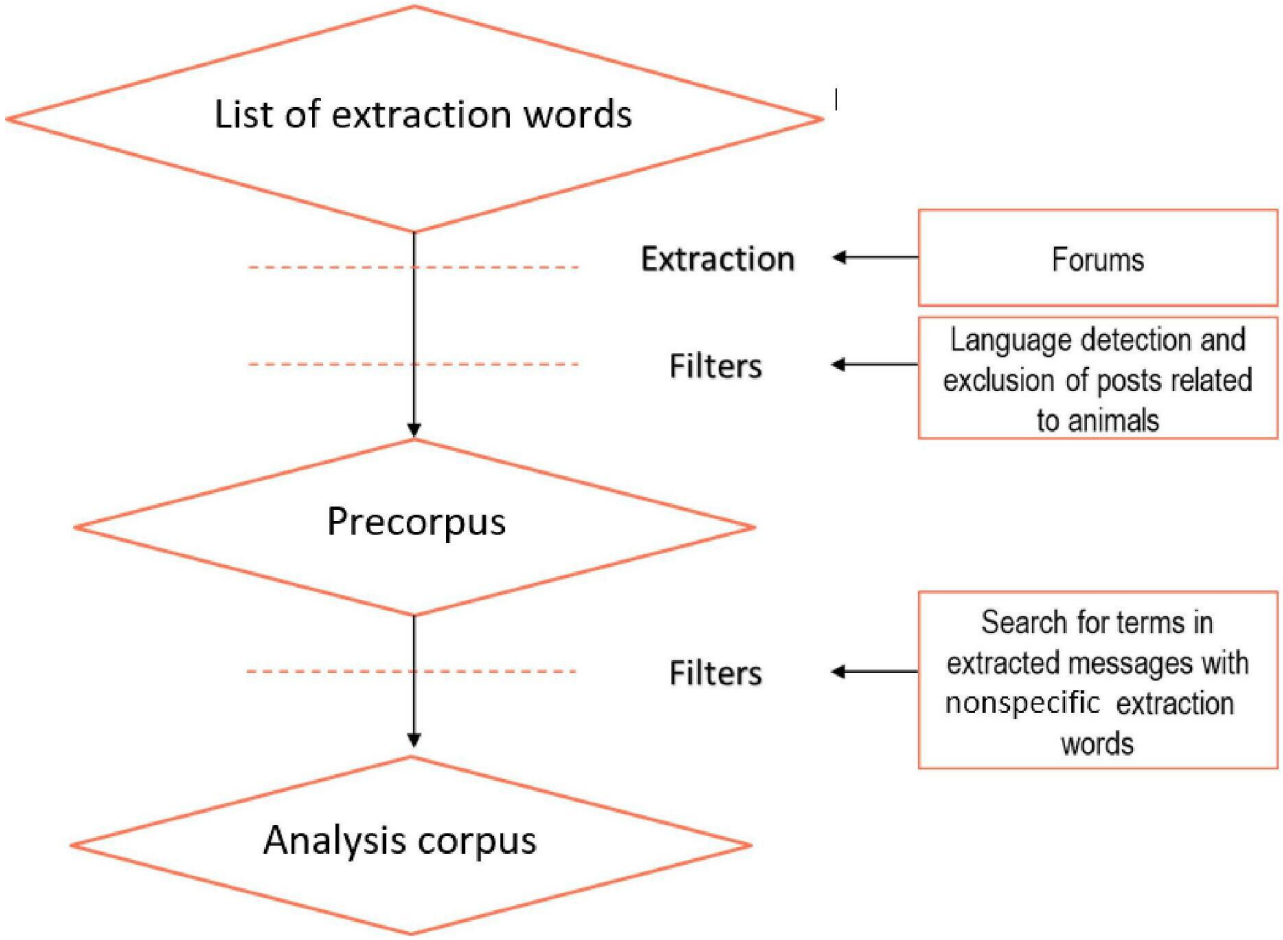

Data Analysis

Age and Gender

Web users' gender is determined through the identification of regular expressions for each gender. First names and gender-associated suffix and prefix are first searched in the username. Then, content of all available messages is screened for gender-specific lexical fields and gender agreements of adjectives and verbs. Finally, a score is computed for each gender and a prediction for the user is obtained by comparing them.

Web users' ages are identified by a 2-fold method. First, regular expressions of age are identified in the total of the user posts. Then, if no expressions were found, a probabilistic machine learning model predicts an age based on several features. Among them, the model considers syntactic aspects of posts as well as expressed feeling and the source on which the user expresses himself.

Topic Model

A topic model was applied to identify the themes addressed in each message. Topic models consist of text mining approach aiming to automatically identify the abstract themes addressed in a collection of documents. Such models are based on the hypothesis that each document in the corpus corresponds to a distribution of several topics. A biterm topic model (BTM) was used to identify the topics without previous knowledge. A topic is defined as a subject of discussion, which amounts to tokens that frequently appear together in a corpus. The BTM considers the whole corpus as a mixture of topics, where each posts' co-occurring pair of tokens (the biterm) is drawn from a specific topic independently, and modeled topics are probability distributions over the biterms [15].

Topics being probability distributions over tokens of the corpus of study, they can be characterized by the highest per-topic probability tokens. Weighting these probabilities through term-frequency inverse document frequency weighting allows to allocate a higher importance to topic-specific tokens. In this case, the per-topic probability of a token was weighted by the inverse of the probabilities of this token in other topics. For each topic, tokens were ranked from highest to lowest weighted probabilities. The first 15 tokens are designated as the set of characteristic tokens and used to manually name the topic.

QoL

QoL impacts were identified; thanks to an algorithm previously developed [16] on the basis of 2 standardized questionnaires measuring health-related QoL (EQ5D and SF36). The algorithm is 2-fold. First, it indicates if an impact is expressed and, second, thanks to 5 specific models, it indicates the nature of the impact (physical, psychic, activity-related, relational, or financial). The 6 models (1 for the first step and 5 for the second) are extreme gradient boosting models trained on annotated patients posts related various pathologies. Features involved in both steps of

XSL•FO
RenderX

the algorithm describe expressed feelings, grammar, conjugation, and lexical fields of health-related QoL-related features, integrating semantic fields specific to the social media domain, allowing it to capture the patient experience as closely as possible. This algorithm was used to automatically create groups of posts as per type of impact, which were subsequently manually reviewed.

Ethical Considerations

This study included data from publicly available sources; private groups and web pages were thus excluded from our data extraction process. We did not seek approval as users automatically grant their consent for the reuse of their data when they post on public platforms. Furthermore, the results of this study do not contain any identifiable information and are presented in aggregate. Information such as the name, username or handle, geographic locations, or any other sensitive data were not included.

Results

Description of the Population and Posts

After filtering and formatting, the analysis corpus contained a total of 3773 messages corresponding to 2335 different web users, with a mean of 1.62 posts per user (Figure 3).

Figure 3. Methodology.

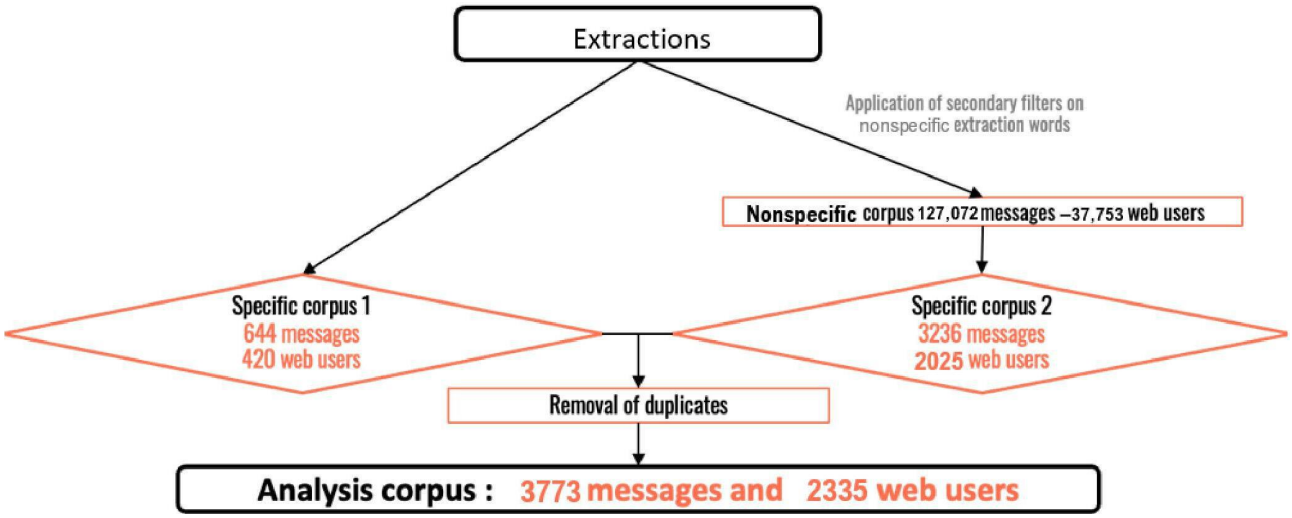

Data showed that there is no specific forum for the sharing of information on activities relating to lack of efficacy of antibiotics and antibiotic resistance and a disparity of sources with a significant proportion of extracted data coming from X and Doctissimo with 888 (23.5%) and 885 (23.5%) posts out of 3773, respectively (Table 1). Furthermore, there was no connection between sources, suggesting a lack of structuring and of visibility of information networks.

We observed fluctuations in the evolution of the volume of messages between January 2014 and July 2020 (Figure 4). We observe a seasonality effect mainly during the “cold” and flu periods, from November to March.

Figure 4. Fluctuations in the evolution of the volume of messages between January 2014 and July 2020 .

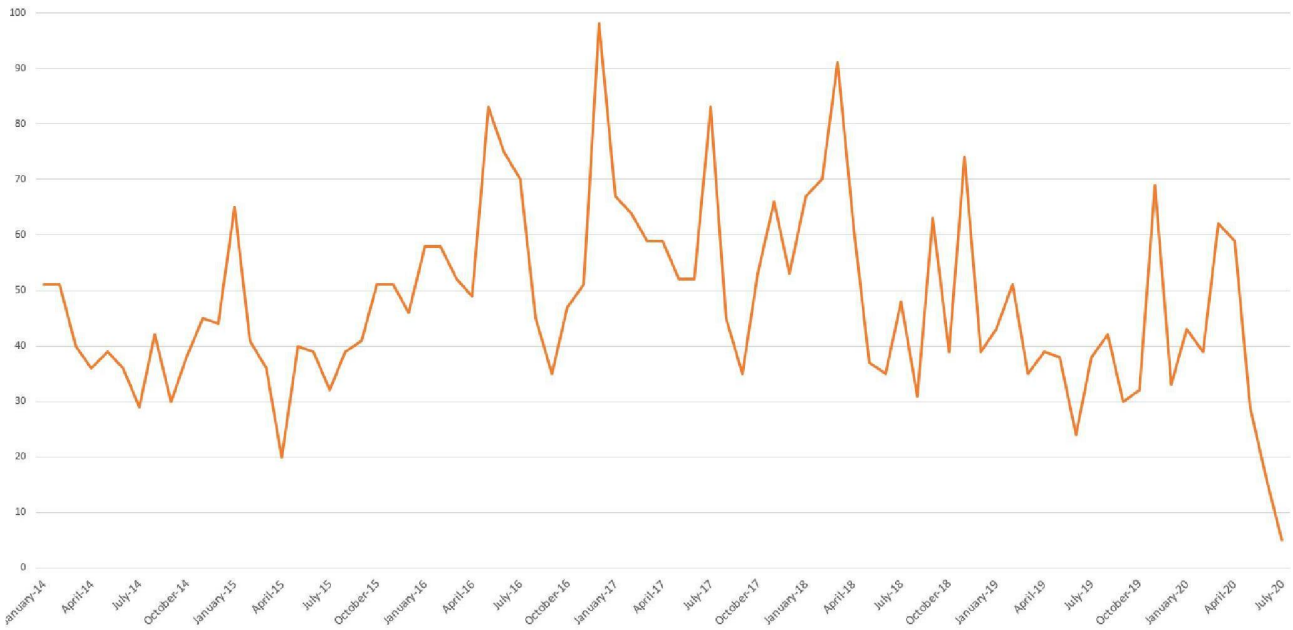

XSL•FO
RenderX

Most users were determined as being women (2015/2337, 86.22%), and the average age was estimated of about 35 years (Figure 5). This is consistent with the results of many studies that point out that women express more personal issues in social networks [17,18]. Furthermore, mothers are more likely to seek advice and help about their infants' health from multiple sources, including social media [19].

Figure 5. Age distribution.

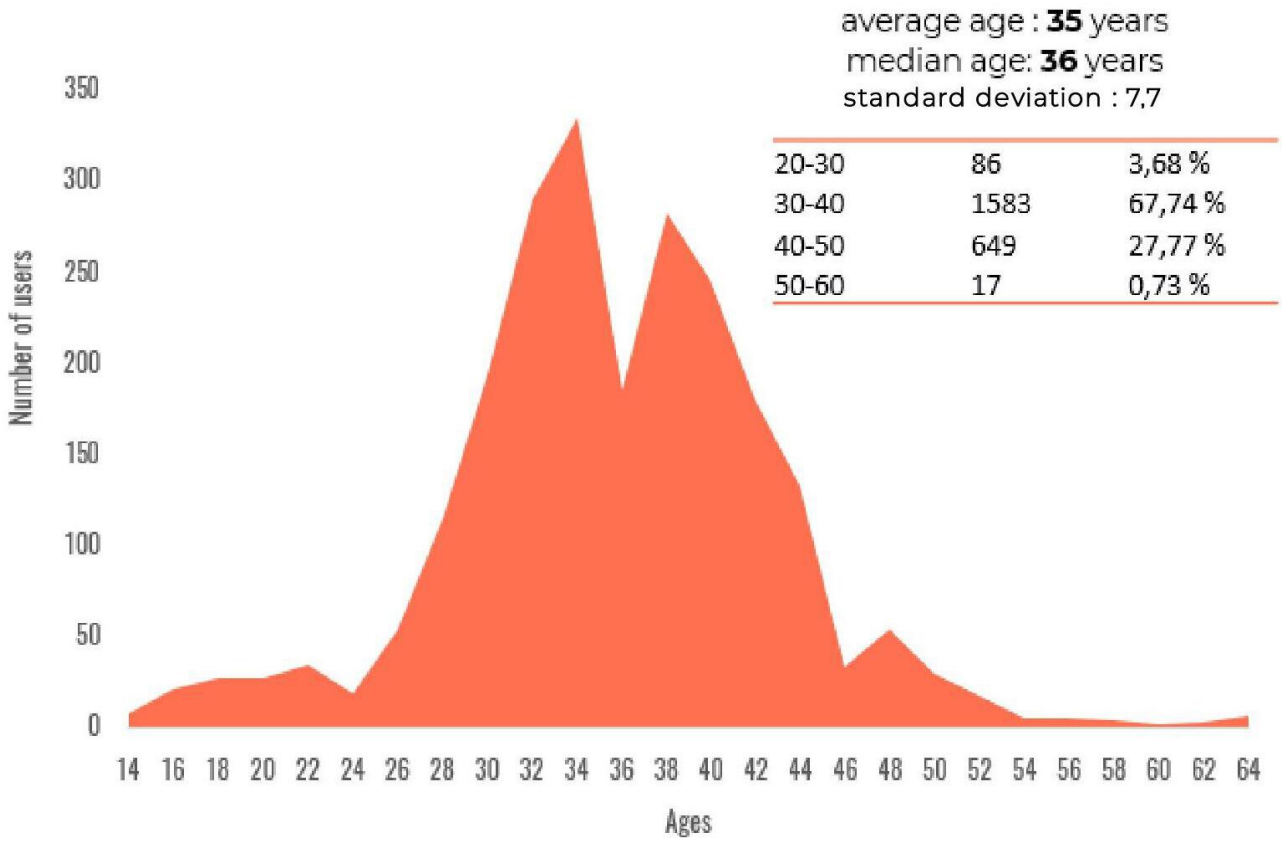

Discussion Themes and Topics

As shown in Table 2, the BTM identified 13 main themes of discussion. The largest category of topics was related to the inefficacy of treatment options ("ineffective antibiotics" [597/3770, 15.84%]), especially for ambulatory diseases such as cystitis (356/3770, 9.44%), acne (226/3770, 5.99%), or dental disorders (109/3770, 2.89%), followed closely by the lack of information regarding their use ("antibiotics and pregnancy," 402/3770, 10.66%). Furthermore, our findings showed a certain degree of superficiality in the knowledge regarding antibiotics resistance (351/3770, 9.31%) and antibiotics overuse (301/3770, 7.98%).

XSL•FO
RenderX

Table . Distribution of posts by overall topic category. Application of a biterm topic model that automatically identifies the different topics addressed within the message corpus. Each theme is associated with a set of characteristic words which allows to define them. A message can only be associated with 1 theme.

Topics	n (%)
Ineffective antibiotics	597 (15.84)
Antibiotics and ear, nose, and throat disorder	461 (12.23)
Antibiotics and pregnancy	402 (10.66)
Persistent urinary infections	356 (9.44)
Mechanisms of resistance	351 (9.31)
Antibiotics overuse	301 (7.98)
Inappropriate use of antibiotics	272 (7.21)
Medical data exchanges	238 (6.31)
Antibiotics and acne	226 (5.99)
Gastrointestinal impact of antibiotics	137 (3.63)
Dental disorders and antibiotics	109 (2.89)
Sensitivity analysis antibiogram	77 (2.04)
Nosocomial infection	66 (1.75)

QoL

QoL is an important criterion for evaluating the effects of a disease and of treatment interventions. In this study, the personal experience of the users in terms of impact on QoL was assessed as physical well-being, social well-being, financial well-being, emotional or psychological well-being, and day-to-day activities. Among the 3773 messages posted, 63% (2384) contained at least 1 impact related to the ineffectiveness of antibiotic treatment on their daily life (Table 3). The most reported impact is physical (1866, 78% of messages), with patients discussing the persistence of symptoms (pain and fatigue), and adverse effects (diarrhea and joint pain). Psychological impact (1551, 65% of messages) is characterized by feelings of anxiety, fear, or despair, with some messages even mentioning depression. Current activities (745, 31% of messages) were also impacted, with sick leave, professional difficulties, or a decrease in regular physical activities. Social impact (602, 25% of messages) was marked by family or sexual problems and feelings of social isolation. Financial impact was less discussed in only 1% (28) of the messages, with messages essentially related to unnecessary expenses or care coverage.

Table . Impact of antibiotic treatment ineffectiveness on patients' quality of life.

Impact	n (%)
Physical	1866 (39)
Psychological	1551 (32)
Day-to-day activities	745 (16)
Social	602 (13)
Financial	28 (1)

Discussion

Principal Findings

Using a web-based tool, we have been able to assess retrospectively how (web user) patients perceive lack of efficacy of antibiotics in France over the last 7 years. Most users were women, and the average age was of about 35 years (Figure 5). Discussions were mainly about therapeutic wandering, without a strong knowledge about antibiotics mechanism of action and antibiotic resistance. This situation might be due to some lack of knowledge and interest on the subject, highlighting the wandering without addressing the essential points at the origin of the lack of efficacy like prescription of antibiotics in case of viral infection. Furthermore, we showed that there is a strong impact of perceived ineffectiveness of antibiotic therapy on the QoL of the patients, especially a strong physical and psychological impact for diverse infections and situations.

Limitations of the Study

This study is to our knowledge the first analysis of messages posted on social media and related to antibiotic resistance in human medicine. It also shows several limitations. First, by focusing only on terms in the French language, we will have underreported the total volume of messages related to antibiotics resistance, and thus, these results are not generalizable at a worldwide scale. Furthermore, using social media to analyze patients' reactions excludes patients who do not have access to the internet or who are not familiar with the use of online discussions.

XSL•FO
RenderX

Extraction bias and a limited amount of data compared with similar studies related to other patient groups (such as breast cancer) were the second limitation. Even if a reasonable sized set of keywords was used, keyword selection in social media studies can induce varying levels of extraction bias, accentuated by a low volume of data.

Another limitation of using social media is the lack of complete information about individual cases. There is also the problem of discovering demographic information—only limited or no information regarding individual user demographics (such as age and origin) may be available.

Despite these limitations, social media represent an ideal place where patients can freely and spontaneously discuss their experiences along with their therapy, thus providing valuable information on their QoL difficult to access by other means. Nevertheless, these observations should be interpreted cautiously, since social media data may include a higher frequency of erroneous information, and patients posting on social media forums may not be representative of the wider patient population.

Implications and Future Research

Messages published on social networks should be integrated into the assessment of patient's QoL, as they can help to characterise the patient's experience in a more individualized and spontaneous way. Furthermore, it seems important to further explore certain subpopulations (eg, pregnant women, children, or even older adults via their parents or caregivers).

To the extent that the most dominant topics can be interpreted as unmet informational needs, our study highlights the refinement of practical implications such as the improvement of existing tools, and further reinforcement of the use of available tools (such as antibiotic susceptibility testing or rapid detection tests), and as well as the communication around them. The ability of internet users to attribute this perceived ineffectiveness to inadequate prescribing or antibiotic resistance is not perceptible in this study and could be the subject of further analysis.

This study suggested to carry out a study segmented by year to examine the evolution of the different subjects (QoL, topics, and treatments) over the years and to observe the global evolutions. Furthermore, similar studies in another country, such as Germany or the United Kingdom, should be planned to compare messages posted between three countries with similar way of life and growth domestic product.

Conclusion

Given that health information is shared extensively on social networks, such services can potentially be used to gather important real-time health data and may provide a venue to identify potential misuse or misunderstanding of antibiotics, promote positive behavior change, and disseminate valid information related to QoL and to bacterial and nonbacterial infections.

One of the most important challenge for health care professionals is to educate people about antibiotics, their adverse effects, and to encourage them to stop the misuse of such drugs. Our study illustrates the current situation concerning the lack of efficacy of antibiotics and the paucity of valuable information and treatment options for wandering patients. In our case, nothing allows us to affirm that these are therapeutic failures related to antibiotic resistance, inappropriate prescriptions or misuse by patients. However, it can provide health care experts with direct patient input on their experience in real-life conditions. This web analysis made it possible to collect, for the first time, the expression of patients who have failed antibiotic therapy. While the extent of these failures is difficult to demonstrate, these patients' experiences tell us how they feel. This research deserves to be continued to better understand the evolution of this phenomenon.

Acknowledgments

Medical writing support was provided by Nathalie Baudry of the microenterprise NB-Pharmaconsult. The study was funded by Pfizer SAS.

Conflicts of Interest

LO has disclosed that he is employee of Pfizer. FNH, PV, and SS have disclosed that they are employees of Kap Code, a CRO, which contracted with Pfizer to carry out this study.
PG, JA, IM, and MG-D have nothing to declare.

Multimedia Appendix 1
List of exclusion words used to filter the extraction of messages.
[DOCX File, 25 KB - formative_v9i1e37160_app1.docx]

References

1. Sengupta S, Chattopadhyay MK, Grossart HP. The multifaceted roles of antibiotics and antibiotic resistance in nature. Front Microbiol 2013;4:47. [doi: 10.3389/fmicb.2013.00047] [Medline: 23487476]
2. Ventola CL. The antibiotic resistance crisis: part 1: causes and threats. P T 2015 Apr;40(4):277-283. [Medline: 25859123]

XSL•FO
RenderX

3. European Centre for Disease Prevention and Control. Antimicrobial resistance in the EU/EEA (EARS-net) - annual epidemiological report for 2020. 2022.
4. von Muhlen M, Ohno-Machado L. Reviewing social media use by clinicians. J Am Med Inform Assoc 2012;19(5):777-781. [doi: 10.1136/amiajnl-2012-000990] [Medline: 22759618]
5. Laudon KC, Traver CG. E-Commerce: Business, Technology, Society, 4th edition: Pearson; 2008.
6. Tennant B, Stellefson M, Dodd V, et al. eHealth literacy and Web 2.0 health information seeking behaviors among baby boomers and older adults. J Med Internet Res 2015 Mar 17;17(3):e70. [doi: 10.2196/jmir.3992] [Medline: 25783036]
7. Harger howe advertising. Top healthcare recruitment social media secrets. URL: https://inbound.hargerhowe.com/blog/top-healthcare-recruitment-social-media-secrets [accessed 2025-03-03]
8. AHRQ. Medal.org (Medical Algorithms): Agency for Healthcare Research and Quality, Innovation Exchange. 2011. URL: https://innovations.ahrq.gov/qualitytools/medalorg-medical-algorithms [accessed 2025-03-10]
9. 10 Tech & IT Buzzwords For 2021 You Won't Be Able To Avoid BI Blog | Data Visualization & Analytics Blog | datapine. 2020 Dec 1. URL: https://cloudwars.com/reporting/10-tech-it-buzzwords-for-2021-you-wont-be-able-to-avoid/ [accessed 2021-02-10]
10. Shaban-Nejad A, Michalowski M, Buckeridge DL. Health intelligence: how artificial intelligence transforms population and personalized health. NPJ Digit Med 2018;1:53. [doi: 10.1038/s41746-018-0058-9] [Medline: 31304332]
11. URL: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Archive:Quality_of_life_indicators_-_health.
12. Abdellaoui R, Schück S, Texier N, Burgun A. Filtering entities to optimize identification of adverse drug reaction from social media: how can the number of words between entities in the messages help? JMIR Public Health Surveill 2017 Jun 22;3(2):e36. [doi: 10.2196/publichealth.6577] [Medline: 28642212]
13. Kürzinger ML, Schück S, Texier N, et al. Web-based signal detection using medical forums data in France: comparative analysis. J Med Internet Res 2018 Nov 20;20(11):e10466. [doi: 10.2196/10466] [Medline: 30459145]
14. Najork M. Web crawler architecture. In: LIU L, ÖZSU MT, editors. Encyclopedia of Database Systems: Springer US; 2009:3462-3465. [doi: 10.1007/978-0-387-39940-9_457]
15. Yan X, Guo J, Lan Y, Cheng X. A biterm topic model for short texts. Presented at: WWW '13 Proceedings of the 22nd international conference on World Wide Web; May 13-17, 2013; Rio de Janeiro Brazil p. 1445-1456 URL: https://dl.acm.org/doi/proceedings/10.1145/2488388
16. Renner S, Marty T, Khadhar M, et al. A new method to extract health-related quality of life data from social media testimonies: algorithm development and validation. J Med Internet Res 2022 Jan 28;24(1):e31528. [doi: 10.2196/31528] [Medline: 35089152]
17. Rollero C, Daniele A, Tartaglia S. Do men post and women view? The role of gender, personality and emotions in online social activity. Cyberpsychology (Brno) 2019;13(1). [doi: 10.5817/CP2019-1-1]
18. Social Media Today. Gender-Specific Behaviors on Social Media and What They Mean for Online Communications. URL: https://www.socialmediatoday.com/social-networks/gender-specific-behaviors-social-media-and-what-they-mean-online-communications [accessed 2021-02-17]
19. Duggan M, Lenhart A, Lampe C, Ellison N. Parents and social media. Pew Research Center. 2015 URL: https://www.pewinternet.org/2015/07/16/parents-and-social-media/

Abbreviations

AI: artificial intelligence
BTM: biterm topic model
QoL: quality of life

Edited by A Mavragani; submitted 24.02.22; peer-reviewed by J Cai, T Delory; revised version received 26.09.22; accepted 27.09.22; published 25.04.25.

Please cite as:
Arquembourg J, Glaser P, Roblot F, Metzler I, Gallant-Dewavrin M, Nanguem HF, Mebarki A, Voillot P, Schück S
Discussions of Antibiotic Resistance on Social Media Platforms: Text Mining and Mixed Methods Content Analysis Study
JMIR Form Res 2025;9:e37160
URL: https://formative.jmir.org/2025/1/e37160
doi:10.2196/37160

© Jocelyne Arquembourg, Philippe Glaser, France Roblot, Isabelle Metzler, Mélanie Gallant-Dewavrin, Hugues Feutze Nanguem, Adel Mebarki, Paméla Voillot, Stéphane Schück. Originally published in JMIR Formative Research (https://formative.jmir.org), 25.4.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License

XSL•FO
RenderX

(https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Motivations of Citizens to Attend an eHealth Course in the Public Library: Qualitative Interview Study

Lucille Standaar[1,2*], MSc; Adriana Margje Israel[3*], MD; Rosalie van der Vaart[4], PhD; Brigitta Keij[1], PhD; Frank J van Lenthe[3], Prof Dr; Roland Friele[2,5], Prof Dr; Mariëlle A Beenackers[3], PhD; Lilian Huibertina Davida van Tuyl[5], MBA, PhD

[1]Centre for Public Health, Healthcare and Society, Department of National Public Health and Healthcare, National Institute for Public Health and the Environment, Postbus 1, Bilthoven, The Netherlands

[2]Department of Tranzo, School of Social and Behavioral Sciences, Tilburg University, Tilburg, The Netherlands

[3]Department of Public Health, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands

[4]Research Group Technology for Healthcare, The Hague University of Applied Sciences, The Hague, The Netherlands

[5]Netherlands Institute for Health Services Research, Utrecht, The Netherlands

[*]these authors contributed equally

Corresponding Author:
Lucille Standaar, MSc
Centre for Public Health, Healthcare and Society, Department of National Public Health and Healthcare, National Institute for Public Health and the Environment, Postbus 1, Bilthoven, The Netherlands

Abstract

Background: There is worldwide recognition of the potential increase of digital health inequity due to the increased digitalization of health care systems. Digital health skill development may prevent disparities in eHealth access and use. In the Dutch context, the public library has started to facilitate support in digital health skill development by offering public eHealth courses. Understanding the motivations of people to seek support may help to further develop this type of public service.

Objective: This is a qualitative study on the motivations of citizens participating in an eHealth course offered by public libraries. The study aimed to explore why citizens were motivated to seek nonformal support for eHealth use.

Methods: A total of 20 semistructured interviews with participants who participated in an eHealth course were conducted in 7 public libraries across the Netherlands. The interviews were conducted between April and June 2022. Purposive sampling took place in the public library during the eHealth course. The interviews covered participants' motivations, attitudes, and experiences with eHealth use and their motivations to seek help with eHealth use. Interviews were audio-recorded and transcribed. Themes were identified via a comprehensive thematic data analysis.

Results: The participants were 51 to 82 years of age (average 73.5, SD 6.6 y) and 14 (70%) participants were female. Three motivational themes were identified: (1) adapting to an increasingly digital society, (2) sense of urgency facilitated by prior experience in health care, and (3) a need for self-reliance and autonomy. Additionally, participants expressed a general desire for social contact and lifelong learning. A lack of adequate informal support by friends and family for digital skills and positive experiences with formal support from public libraries stimulated the participants to seek formal support for eHealth use.

Conclusions: We show that the participants had a feeling of urgency that sparked their motivation to seek nonformal support in the public library. Motivations to participate in the eHealth course stemmed from the need to adapt to the digital society, being a patient or a caregiver, or the need or wish to be independent from others. Participants of the study were mainly older female adults who had native language abilities, up-to-date digital devices, and time. It is likely that other populations experience similar feelings of urgency but have other support needs. Future research should explore the needs and attitudes of nonusers and other users of digital health toward seeking support in eHealth access and use.

(JMIR Form Res 2025;9:e60612) doi:10.2196/60612

KEYWORDS

digital health; help-seeking behavior; socioeconomic factors; community health services; older people; eHealth; older adults; geriatric; support; eHealth literacy; interview; Netherlands; thematic data analysis; health literacy; mobile phone

XSL•FO
RenderX

Introduction

With the growing demand for health care services, eHealth is increasingly recognized as a critical tool for efficiently managing and delivering health care [1]. Through the adoption of patient portals, telemonitoring tools, and smartphone apps, patients can experience improved access to health information, shared decision-making, and self-management [2,3]. However, not everyone within society is able to access and use digital health care [4-12]. The increased implementation of digital tools in health services is linked to the widening disparities in health [13].

The need for more emphasis on equity in the access and use of digital health care systems is recognized worldwide. Research shows that disparities in the use of digital health are common in member states of the European Union [14,15]. The causes of digital health inequities derive from unequal access to resources, language, and digital health skills required for effective eHealth utilization [8-10,16-22]. World Health Organization and Organization for Economic Co-operation and Development encourage countries to adopt digital health literacy policies on citizen and health system levels [23,24]. Yet, only 27 out of 52 European Union member states have developed policies on digital health literacy (2023) [23]. Digital health literacy is defined by Norman and Skinner [25] as "the ability to seek, find, understand, and appraise health information from electronic sources and apply the knowledge gained to addressing or solving a health problem." Enhancement of digital health skills is one of the avenues for achieving a more equitable digital health care system [23,24,26].

In The Netherlands, the COVID-19 pandemic created awareness about the lack of digital health literacy which resulted in a collaborative effort between the government and the public library organization to facilitate support and education in the development of digital skills, including digital health skills [27-29]. As a result of this collaboration, the public library in the Netherlands has evolved into an organization that aims to support "those who need help" [30,31]. The public library offers a nonformal free-of-charge eHealth course, designed to enhance digital health skills [31]. In the literature, 3 categories of support are defined; formal support: institutionalized education with officially recognized awards, nonformal support: institutionalized support and education without officially recognized awards, and informal support: support via social networks [32]. Support via the public library is a form of nonformal support. As the public library is freely accessible to the whole population, costs or resources would not deter the use of this type of support. However, the first step to participate remains to be taken by the individual and depends on their motivation to improve their digital health skills [17,19,31].

A better understanding of the motivations of people who seek nonformal support can create useful insights into the sentiments that drive eHealth adoption and support-seeking behavior. These insights can be used to improve eHealth implementation and the design of eHealth support strategies. This study aims to explore the motivations that drive participants in a freely accessible eHealth course to seek nonformal support for eHealth use. Building on our prior research and other findings, which demonstrated that education was associated with motivation to use eHealth [18], this study's findings will also be explored through the lens of the educational background of people who seek support.

Methods

Study Design

This study aims to understand and explore the motivations of participants to seek support with digital health skill development. A qualitative study design with semistructured interviews was deemed most appropriate as this study aims to explore and describe the motivations to seek help in the eHealth use of participants of an eHealth course [33]. Due to the explorative nature of the research question, this study follows the interpretivism approach [34]. As this study has an explorative nature with the aim to create a deep understanding of the motivations of the participants to seek support, the results of this study will be unique to the study setting and the study sample. This paper adhered to the SRQR (Standards for Reporting Qualitative Research) checklist (Checklist 1) [35].

Ethical Considerations

The study was declared to fall outside the scope of the Dutch Medical Research Involving Human Subjects Act by the Clinical Expertise Center of the Dutch National Institute for Public Health and the Environment (VPZ-559). Transcripts of the interviews and the informed consent forms are stored and protected at the Dutch National Institute of Public Health and the Environment. All participants provided their informed consent prior to conducting the interviews. Participants received a 15€ (US $16.047) voucher for their participation. Transcripts were deidentified.

Setting

Digital health care is widely implemented within the Dutch health care system and involves data storage, interaction between (health care) professionals, interaction between health care professionals and patients, and freely accessible health information websites and apps [36]. Implementation of digital health care has been experimented with over the last decades within the Dutch health care system and is stimulated by the government [37-39]. The Netherlands is one of the best-connected countries in the European Union, with 98% of its citizens connected to a type of internet connection [40]. The history of experimentation with eHealth implementation, governmental support, and the wide availability of the internet is likely to support the digitalization of the health care system further.

The eHealth course "DigiVitaler" is part of the web-based educational material from the National Public Library and available to all public libraries since 2021 [41]. In 2022, 60 out of 137 public libraries offered the eHealth course [41]. This course is structured into multiple classes, each covering a specific eHealth topic, such as the patient portal, video consultations, and web-based health information. Participants are introduced to eHealth by providing hands-on experience with eHealth apps in simulated digital environments. The course

XSL•FO
RenderX

consists of 1 up to 8 classes, each with a duration of 2 - 3 hours, and accommodation of 2 - 12 participants. Public libraries are free to choose the number of classes they offer. Public libraries used printed, word-of-mouth, and social media advertisement strategies to announce the course, disseminated via libraries' own platforms, local newspapers, and local organizations.

Recruitment

Purposive sampling of participants who had experience with the eHealth course was performed by 3 researchers (LS, AMI, and Anita Suijkerbuijk) in the context of 7 public libraries. Researchers aimed to recruit a representative sample of all participants of the eHealth course. Through contact with the developer (Stichting DigiSterker), the researchers retrieved contact information of the libraries offering the eHealth course. The course developer was not involved in this study. Libraries were contacted via email and received information about the study with the request to allow the researchers to be present during the eHealth course for participant recruitment. A total of 7 out of the 26 contacted libraries agreed to participate. Reasons for libraries to decline were course cancellations due to a low number of course participants, no course dates in the data collection period, no response, or unwillingness to participate in the study. After obtaining consent from the librarians, the researchers (LS, AMI, and Anita Suijkerbuijk) attended 9 classes in 7 libraries located in rural and urban areas in the Netherlands. At the start of each class, potential participants were introduced to the study through a presentation and a flyer. With the approval of the course trainer, researchers would play an active role during the class, providing support and answering the course participants' questions about the course material. At the end of the class, participants were invited for an interview. Participants who expressed interest were provided with detailed study information, a consent form, and the opportunity to ask questions. Recruitment and interview procedures were designed to limit the participants' potential efforts to participate in the study. Procedures were conducted in a familiar context for the participants and required limited travel time and no digital skills. All participants were interviewed at the library, either immediately following the course or at a preferred later date. Telephone interviews were an option but were not requested by the participants. Digital interview formats were not considered as it was highly likely that the participants had limited digital skills, which would potentially pose a barrier to recruitment.

Participants

Participants were eligible to participate if they were older than 18 years of age and participated in the eHealth course. Educational background was categorized according to the International Standard Classification of Education (ISCED), where a low educational background is categorized as ISCED 0 - 2, intermediate as ISCED 3 - 4, and high as ISCED 5 - 8 [42]. The rurality of the participants' living area was derived from the area where libraries were situated. Rurality was determined via the rurality classification from Statistics Netherlands [43]. A total of 48 participants attended the classes visited by the researchers, of which 20 (20/48, 42%) participants participated in the study. Reasons for nonparticipation were lack of time or interest or the participant's own perception of lack of experience with eHealth.

Data Collection

In this study, motivations related to seeking help and the use of eHealth were studied. The themes of the interview guide covered: the demographics of the participant; the motivations, self-efficacy, and attitudes for eHealth use; motivations regarding seeking and receiving support for eHealth use; and social context of support and eHealth use. The interview guide can be found in Multimedia Appendix 1.

Multiple theories describe the transition from motivation to behavior [44-46]. The 3 concepts, attitude, social norms, and self-efficacy, are often described to result in motivation [44-46]. In this study, these theories were used as a means to develop the interview guide that covers the main concepts of motivation. Attitude refers to positive and negative beliefs about the consequences of certain behaviors [44-46]. Social norms include the beliefs of others about the behavior [44-46]. Self-efficacy expectations refer to the beliefs of one's personal capacities to actually change behavior [44-46]. Questions that cover these concepts entailed questions such as: "What or who caused you to seek support with eHealth use and why?" (attitude and social norm); "What was the role of family; friends; health care professionals; the library; or the municipality to use eHealth?" (social norm); "Would you be able to use eHealth yourself after the course?" (self-efficacy). The interview guide was developed by LS and AMI and reviewed by LT, AS, and MAB.

Semistructured interviews were conducted face-to-face by researchers LS, Anita Suijkerbuijk, and AMI, all experienced in conducting qualitative research. None of the researchers knew the participants prior to this study. The data collection period lasted from April until June 2022. All interviews were audio-recorded with participants' permission and lasted 30 - 60 minutes. After each cycle of 6 interviews, the content of the interviews was discussed by LS and AMI. After 20 interviews, no new information regarding motivation to seek help with eHealth use was heard. At this point, it was concluded that data saturation was reached.

Data Analysis

Deidentified verbatim transcriptions of the interviews were analyzed with the "Codebook Thematic Analysis" approach [47]. This approach allows for a structured process of coding, theme development, and conceptualization in a multidisciplinary team while maintaining the open, inductive approach of reflective thematic analysis to explore the motivations of the participants in-depth [48]. First, transcripts were read and summarized for data familiarization. A deductive codebook was derived from the interview guide. The first 2 interviews were double-coded by LS, AS, LT, and AMI, who then collaboratively defined and fine-tuned the codebook. After consensus, the remaining interviews were coded and cross-checked by LS and AMI, using MaxQDA (VERBI Software) for data management. This deductive analysis was performed to create an overview of the data gathered, followed by an inductive analysis to identify underlying patterns. The following topics were analyzed inductively: motivation for

XSL•FO
RenderX

eHealth use, motivation for seeking support in eHealth use, and motivation to participate in the eHealth course. The first 20% of the transcript segments per topic were inductively coded by LS and AMI separately. After reaching a consensus, the rest of the transcripts were coded and cross-checked by LS and AMI. The inductive codes were arranged into candidate themes. Candidate themes were then further defined through discussion between all researchers to determine the final themes. After analysis, all the transcript summaries were reread and the presence of the themes for every participant was evaluated through the lens of their life course with the aim of identifying possible differences according to educational background. Quotes were translated using the translation software DeepL (DeepL SE) and these were checked by a native English speaker with limited Dutch fluency.

Results

Overview

A total of 5 men and 15 women with an average age of 73.5 (SD 6.6; range: 51 - 82) years participated in this study. All participants were White and fluent in the Dutch language. Most participants (n=13) lived in an urban area. Seven participants were categorized with a low educational background (LE), 6 participants with an intermediate (IE), and 7 participants with a high educational background (HE). All participants made use of 1 up to 4 digital devices and 14 participants had prior experience with digital health services. The study sample's (n=20) demographic overview can be found in Multimedia Appendix 2. Participants mentioned three types of support to develop their digital skills: (1) family, friends, and acquaintances; (2) the public library and other public or social organizations; and (3) commercial resources. Participants were either introduced to the eHealth course by library staff during participation in other courses offered by the library, advertisements in the local newspaper, or via their church, social organization, or computer club.

General Experience With the Digitalizing Society

The participants' experiences with and attitudes toward the digital world varied, yet all expressed some degree of difficulty using digital devices and services due to a perceived lack of adequate digital skills. Many pointed out that it is difficult for "their generation" to keep up with the latest developments within an ever-changing field. "Their generation" pertains to those who did not grow up with computers or the internet and only learned this later in life, even though some had multiple years of digital experience.

> *I think it's a shame that there's no choice anymore. There's obviously a generation of us that needs to have a transition that's a little bit smooth, and not like: bam, next year everything's just different. I think a lot of people my age have problems with that.* [Participant 5, female, 72 years of age, LE]

For most, this digital experience was work-related where they were trained in specific, mostly simple computer programs useful for their jobs. However, this was different from their experience in this day and age where every aspect of life has become infused with digitalization and there often is little to no alternative left.

Themes

Overview

Three themes were identified to describe individuals' motivations to seek help for eHealth use: (1) adapting to an increasingly digital society, (2) a sense of urgency facilitated by health care experience, and (3) a need for self-reliance and autonomy. Additionally, a desire for social contact and lifelong learning were general reasons to participate in courses offered by the public library. An overview of the themes, the topics, and the number of codes linked to a topic is presented in Multimedia Appendix 3.

Theme 1—Adapting to an Increasingly Digital Society

Overview

The majority of the participants experienced that society is becoming more digital and they feel the urgency to try and keep up with the digital transition to be able to function as participating citizens. Three aspects encompass this motivation: (1) realizing the lack of one's own digital skills, (2) feeling forced by society to keep up with digitalization, and (3) a desire to feel included in society and be prepared for the future. This theme was identified among all participants, regardless of educational level. It seems in this respect that among the participants of this course, their age and the fact they have little experience with digital technology played a bigger role than their educational background.

A Lack of Digital Skills

The experience and expectation of lacking digital skills created a sense of urgency for most participants to learn how to use eHealth. Participants found it uneasy to navigate through the digital world in general. Many participants reported that developing and maintaining digital skills requires continuous effort.

> *But it's my experience, if you don't do it regularly, it disappears again. Then you think, "Oh dear, how did I do this again?" And just figuring that out, it's very frustrating.* [Participant 11, female, 73 years of age, HE]

Some participants experienced fear of using the computer since they were afraid to make (irreversible) mistakes.

> *The fear of buttons for me is that I think I'm pressing such wrong buttons that everything'll disappear.* [Participant 15, female, 74 years of age, HE]

Some participants tried to use eHealth but failed due to a lack of digital skills. Others found that their lack of digital skills limited them in using eHealth to its full potential or impeded their preferred option in health care services. This motivated them to start learning.

Feeling Forced by Society to Keep Up With Digitalization

As services and goods become increasingly digitalized without alternatives, participants felt compelled to adapt and commit to them.

XSL•FO
RenderX

> *Then I think: something is being shoved down my throat again. I don't want that at all. But still, I get it: you have to, whether you want to or not.* [Participant 19, female, 74 years of age, LE]

Anxiety about falling behind and becoming isolated from society or losing part of their autonomy was common. Feelings of insecurity, shame, or feeling scared using new digital technology were no exception, as previous experiences with or news warnings about scamming or loss of privacy raised their concerns.

> *A copy of my passport, I find that very uncomfortable when asked for it. […] I don't really dare to [upload] ... because I'm afraid, yes it's very farfetched but it does happen. I'm still afraid of being taken over by cybercriminals.* [Participant 17 (female, 51 years of age, HE]

Many participants foresaw that health care services would further digitalize. They feared diminished access to care if they were unable to use eHealth, which resulted in beliefs that they needed to forcefully commit. However, the barriers participants experienced for the use of eHealth were numerous. Negative experiences contributed to the reluctance participants felt to use eHealth: lack of uniformity between eHealth apps, difficulty identifying appropriate health information, and difficulty in phrasing health problems in digital form.

> *Too complicated for us at this age [asking questions via e-consultation]. You have to do too many steps. You have to log in, you have to type, you have to explain what it is, and everything has to be short.* [Participant 10, female, 78 years of age, HE]

Desire to Feel Included in Society and Be Prepared for the Future

The desire to keep up with digital developments led to an intrinsic motivation to learn how to use eHealth. The participants felt that by keeping up, they stay part of society and are prepared for possible (future) encounters with eHealth. Some participants felt the urgency to learn how to use eHealth now because they foresaw that this would become harder as they age.

> *It is the future where we are headed, soon you have to do a lot by computer and by smartphone. Then it's really nice if I know how to use it, because if I'm approaching 70 and I still have to learn, well then it is too late.* [Participant 6, female, 66 years of age, LE]

These participants noticed that they were the early adopters of eHealth in their social network. Friends and relatives did not share the same interests or understand why the participant was willing to learn digital skills to use eHealth. This is part of the reason why nonformal support in the form of the eHealth course is a particularly attractive option, as the informal network of the participants is not as knowledgeable.

Theme 2—Sense of Urgency Facilitated by Prior Health Care Experience

Overview

The second theme depicts that the participants felt motivated to seek help to use eHealth through their prior health care experience. Three aspects encompass this theme: (1) the responsibility as a patient or caregiver, (2) experience with the benefits of eHealth, and (3) the attitude of health care professionals toward eHealth. This theme was mentioned by all participants from the low-educated group, a few participants in the intermediate-education group, and half of the high-educated group. Regardless of their educational background, the participants who have had experience in the health care system felt more urgency to use eHealth.

The Responsibility as a Patient or Caregiver

Many participants had some form of experience with health care, although not all with eHealth. Being a patient or caregiver to their partner or child stimulated a feeling of responsibility to take an active role in the care process. For these participants, this active role encompassed the ability to use eHealth. Participants illustrated this with examples of preparing hospital appointments or looking up information in their electronic patient records. They are able to take more time to think about the medical information before a consultation to prepare questions, and also afterward to reread the information given during a consultation. eHealth allowed participants to better process the information and use this to their advantage.

> *My husband was in the hospital with a heart attack [...] and I saw everything was digital there. I had already looked up his medical history once and printed it out because my husband forgets a lot. [...] And I think, now they are going to start something else [a new digital system] in the hospital again. I have to learn this from the start, and I think my husband has to go along with that.* [Participant 13, female, 70 years of age, LE]

A few participants expressed that they feel this type of active role is expected of them by medical professionals to be able to partake in shared decision-making. These experiences motivated the participants to learn more about eHealth when they were presented with the opportunity.

Experience With the Benefits of eHealth

Prior experience with health care created perceptions that eHealth can be beneficial and can be a necessity in health care. The benefits participants mentioned were less travel time, the opportunity for digital health care appointments at home, improved exchange of health information between health professionals and organizations, and an improved overview and insight into their personal health information.

> *The next time I visit my doctor and I need information, then beforehand I'll read [my file]. [...] I'd rather know what's on the record about me first. Maybe I can ask another sensible question. Because otherwise you're just a clueless patient.* [Participant 14, male, 76 years of age, IE]

XSL•FO
RenderX

Some participants also believed that eHealth could contribute to the efficiency of the health care system, which they deemed positive for themselves and society. Most participants were not familiar with the possibilities of eHealth before their health care experience. After experiencing these benefits, participants felt eager to learn about the possibilities of digital care. As a result, they sought nonformal support.

Attitude of Health Care Professionals Toward eHealth

A positive attitude toward eHealth or regular use of eHealth by health professionals motivated participants to use eHealth and learn skills that enable eHealth use. Some participants decided to do the course after a health professional suggested or used eHealth for their care.

> *Yes, Thuisarts.nl [a website that provides easily readable patient information on primary health care issues, issued by the Dutch College of General Practitioners], I consult that often. [...] That's what I agreed with the general practitioner at the time. He said, "Did you look on Google again?" [...] And then, "If you go to Thuisarts.nl, that's very reliable." And I just do that.* [Participant 12, female, 80 years of age IE]

Most participants indicated they never received any information about the possibilities of eHealth via their general practitioner or other health care professionals. Some participants argued that better promotion of eHealth is needed to spread awareness and stimulate use.

Theme 3—Need for Self-Reliance and Autonomy

Overview

Participants' motivation to seek help for eHealth use was also tied to their wish for independence. The need for self-reliance and autonomy was described by three aspects: (1) limited informal support, (2) discomfort with dependency on others, and (3) discomfort with loss of their own autonomy. This theme was identified equally among all participants, regardless of educational level. Thus, again, the social circumstances such as having children and the existence or absence of a social network seemed to affect the participants more directly than their educational background.

Limited Informal Support

Limited or inadequate informal support resulted in participants feeling forced to become independent users of eHealth. In a few cases, limited support meant that participants were childless, whereas they felt that society often wrongly assumed that their family would take care of them. This can create a sense of loneliness, but also motivation to learn these skills.

> *I do feel more or less obliged [to become independent in eHealth use]. [...] because I have nobody to fall back on. Everyone says: oh, you just ask your children. Well, I don't have any.* [Participant 1, female, 73 years of age, HE]

Other participants expressed they never felt a need to learn because their spouse would take care of all things digital. However, after their partner falls ill or has passed away, they are on their own and feel they have no other choice but to learn these skills themselves. In addition to the absence of informal support, some participants also experienced inadequate informal support, when the support they receive does not match their needs and wishes. This is especially true for the help the participants receive from their children or grandchildren, who are digital natives, and proceed to quickly resolve any issues. This is often done in such a way that the participants are left clueless about what the solution entails and are not stimulated to learn how to tackle these problems themselves. They continue to be dependent on others, which frustrates them. To be able to develop these skills on their own, they feel they need more structured explanations, time to practice, and patient teachers which the course offers them.

> *Otherwise, I always have to ask one of the grandchildren. […] They'll help me, but it's all quick, quick, quick. Well, that doesn't work. […] That's why I'm here taking this course.* [Participant 6, female, 66 years of age, LE]

Discomfort With Dependency on Others

To lean on others for help in eHealth use was uncomfortable for some participants due to their desire for independence, leading to their wish for independent eHealth use. Usually requesting help with small or acute matters within close ties such as family is acceptable, but depending on others outside of the family such as neighbors or friends feels uncomfortable and like they are burdening others. Some participants expressed that even asking family for help can feel like they are asking too much of their children, especially if they feel their children are busy with their own household, young children to take care of, or demanding jobs. The course offers a welcome alternative, where the participants do not feel like a burden.

> *Because I wanted to deal with it [eHealth] more easily. I do have children who want to help, but everyone still has their own thing. I don't want to become dependent on my kids.* [Participant 10, female, 78 years of age, HE]

Some also reported that they are afraid or find it difficult to ask for help from formal institutions such as the municipality. This, in turn, also has to do with a lack of trust and safety. Participants feel discouraged to ask for support from these types of organizations because they think they will not get sufficient guidance. Participants found commercial services to be too expensive, incompatible with their own equipment, insufficiently informative, or too time-consuming.

Discomfort With the Loss of Their Own Autonomy

Asking for help or being unable to use eHealth meant for some participants that they could not manage their own life and health. Their discomfort lies in their dependency on others rather than being uncomfortable burdening others with their need for help. It means the loss of autonomy which many of the participants experienced their whole lives through their employment, role in the household, and management of daily life. But as they get older, they feel they lose this freedom as it becomes harder to keep up with digital innovation. Many participants feel they want to be able to solve problems on their own, without asking

XSL•FO
RenderX

for help. Attending the course feel like they are in control and actively solving the problem by learning new skills and by doing something themselves. They experience this as more empowering than asking someone else to solve their problems for them.

> *Of course, I always wanted to do everything myself. Now you have to [ask for help] every time... That's the thing. Every time you have to give up a piece of your personality. That's how it feels.* [Participant 19, female, 74 years of age, LE]

In addition to these 3 themes related specifically to the motivation for eHealth use support, the participants expressed a general desire for social contact and life-long learning as a reason to participate in the eHealth course. With regards to social contact, participants reported finding solace and support to tackle digital (health) skills together with likeminded people, where they are not the only ones experiencing difficulty with new technology and have the opportunity to learn from each other's questions and mistakes without feeling like the odd one out. This feeling of kindred spirits can contribute to alleviating feelings of shame that participants may feel due to their lack of adequate digital skills. With regard to life-long learning, some participants proposed their curious personality as a reason to participate in all kinds of different courses offered to them by the public library. Most of these participants initially started on a path to improve their digital skills and this course was a natural continuation of that process to develop these further. Especially those participants who already had positive experiences with formal support at the public library were stimulated to continue with this type of support.

Discussion

Principal Findings

In this qualitative study, including participants of older age and with varying educational backgrounds, three motivational themes were identified for seeking help with eHealth use: (1) adapting to an increasingly digital society, (2) a sense of urgency facilitated by health care experience, and (3) a need for self-reliance and autonomy. A lack of adequate informal support from friends and family and positive experiences with nonformal support stimulated the participants to seek nonformal support in the public library for eHealth use. In addition to the 3 themes, participants expressed a more general desire for social contact with peers and lifelong learning.

The first theme "adapting to an increasingly digital society" explores the motivation to seek support driven by a wish or need for digital inclusion and societal participation in a digitalized society. It shows that the exponential development of digital technology can seem daunting and therefore insurmountable for some older adults who have little to no prior experience with this [49,50]. However, for others, it acts as a stimulus to stay engaged with societal changes, fueling the desire to adopt new technology [49,51]. In this study, this latter attitude also emerged as an important reason to seek support for eHealth use. While computer anxiety and low computer self-efficacy have been named important barriers to technology and eHealth adoption in older adults [52-54], this study shows that awareness of one's own limited self-efficacy in eHealth use can be an important incentive to seek support when individuals see the potential usefulness for themselves as "preparation for the future." Associated concepts such as healthy aging, aging in place (eg, living longer at home), and older person's empowerment have all been positively linked to the use of digital (health) technology [55-58] and are reflected in the third theme.

The second theme "a sense of urgency facilitated by health care experience" revolves around participants' prior experience with eHealth, health care, and health professionals. In this study, some participants experienced this as a necessity, while for others, this stimulated a positive attitude toward the potential benefits. Both acted as a facilitator for seeking support. Ware et al [55] equally found that a desire for ownership, access, and individual responsibility for personal medical information were motivators to use eHealth. Other research also underscores the importance of perceived eHealth benefits, positive prior eHealth experiences, and endorsement of health care professionals to motivate individual use [53,54]. This "window of opportunity" could be used to motivate others to seek nonformal support for their digital skills, for example, when introduced by trusted health professionals. However, this type of experience as a motivator for older adults to seek nonformal support for the development of digital (health) skills seems to be a novel finding.

The third theme "a need for self-reliance and autonomy" emphasizes that participants want to be self-sufficient in their abilities. This attitude toward self-reliance and autonomy illustrated how participants want to live and manage their health independently. It can support aging in place, encouraging older adults to live independently for as long as possible, a desire of most older adults and stimulated by policy makers [59,60]. Informal support by family and friends can both be a facilitator or barrier to the adoption of new digital technology [51,53,54]. In this study, part of the participants showed a dislike for dependency on family and friends and did not want to burden others turning informal support into a barrier, a sentiment also observed by others [51,53]. In this study, the help offered by family members did not lead to the development of more digital skills, as the participants desired. Therefore, they experienced limited and inadequate informal support. This particular phenomenon is seen elsewhere in informal support for internet use among older adults [61]. In addition, our findings on limited informal support are in line with others who find that intergenerational support can lack sufficient patience and understanding for the level of digital skills of older adults [51,53]. This may act as a barrier if it creates a learning environment characterized by judgmental delivery, a fast pace, and the use of unfamiliar jargon [54].

Theoretical reflection from the perspective of Maslow's Hierarchy of Needs provides an opportunity to create a deeper understanding of the origin of the identified motivations. Maslow's [62] theory focuses on human needs, theorizing 5 different types of needs. When these needs are not met, motivations to change behavior can emerge [62]. Furthermore, the theory implies a hierarchy of the 5 needs [62]. The basic needs are physiological needs (eg, water and food) and safety

XSL•FO
RenderX

needs (eg, a roof over one's head and the heating on) [62]. Then there are psychological needs, starting with the social need for love and belonging and moving up to the need for self-esteem [62]. Finally, there are the self-fulfillment needs, or the need for self-actualization, for example, developing one's full potential [62]. Themes 1 and 2 both reflect a safety need. The urgency to adapt to an increasingly digital society and the urgency of being a patient or caregiver show a need for security in access to (digital) health care now or in the future. The motivations from theme 3 can be interpreted as esteem needs, as this theme can be interpreted as a need for independence in access and use of the digital world and independence from support. The motivation stemming from the desire for lifelong learning can be interpreted as a need for self-actualization. The need for social connection can be interpreted as a social need. The perspective of Maslow's theory provides the insight that independent access to digital health care was perceived as a basic safety need and an esteem need by the participants.

Motivation to use eHealth and to seek support in eHealth use are 2 aspects of many when discussing digital health inequity. The digital divide model by van Dijk [10] describes how personal and contextual factors are related to the resources one has available to have motivation, physical access, use, and benefit from digital media, or in this case, digital health. Resources to improve the likelihood of accessing digital media involve cultural, social, mental, material, and time-related factors [10]. The study sample expressed a lack of digital skills and knowledge about digital health but also expressed access to up-to-date devices, language skills, time, and feeling comfortable in a public library setting. Our prior research and other research found that other populations experience barriers to seeking support in the factors that were facilitative for this study population [63-65]. Standaar et al [64], Choudhary and Bansal [63], and Goedhart et al [65] reported that other populations that need and potentially seek help in digital skills development experience barriers in language skills, digital skills, stigma, time, logistics, and feelings of belonging in the context of institutions. A literature review by Choudhary and Bansal [63] allowed for the categorization of the barriers to participate in training programs on the administrative level, training level, learner level, and community level. This categorization acknowledges that barriers are not only individual-bound but also bound to the training content, training context, and community in which training is provided [63]. Our findings suggest that access to up-to-date devices, language skills, time, and feeling comfortable in a public library setting were required to perform support-seeking behavior in the public library context. These reflections suggest that access to support for digital health skill development is also dependent on personal factors, contextual factors, and resources.

We explored the motivations for seeking eHealth support among participants. Additionally, we explored if differences in educational backgrounds led to differences in motivation. We found that other life events like losing connection with (digital) society, becoming a patient or caregiver, and becoming dependent on others later in life appeared to be more important in the emergence of their motivations to participate in an eHealth course. Even prior computer experience was not always an indication of the well-developed digital skills that are needed in the current digital age. Furthermore, exploration of the results through the perspective of education is complex in this older, mostly female sample as it may not clearly indicate educational differences due to historical cultural norms and educational opportunities for women in the 20th century [66]. Therefore, the results of this study provide very limited and nongeneralizable insights on different motivations to seek support in eHealth use in people from different educational backgrounds.

Strengths and Limitations

This study has several strengths. First, participants come from diverse backgrounds, including rural and urban areas enabling exploration of a wide range of experiences. A large percentage of the potential participants (20/48, 42%) participated in the study, ensuring a broad representation of the study population. Second, interviews continued until data saturation was reached, ensuring robust findings. Third, the eHealth course was not part of a research program, which allowed us to evaluate participants' motivations in a real-world setting. The on-site visits enhanced the understanding of the eHealth course context and allowed the interviews to take place in familiar surroundings for the participants. Our participant selection resulted in a culturally nondiverse older group with access to digital devices, motivation, and the capability to seek nonformal support, as a result of this specific research context. Next to this, it is likely that the people attending an eHealth course have some prior knowledge or experience with eHealth, the majority (n=14) of this study sample had some experience with eHealth. Other research in the same context confirms from digital health skill development trainers' perspective that the population attending the eHealth course is homogenous and similar to this study sample [64]. Asmar et al [67] refer to this specific group as "community supported." However, their and other research indicates that older adults are far from homogenous in their use of digital technology or support-seeking behavior [67-69]. Therefore, it is likely that in other settings, different typologies and characteristics of older adults might result in other motivations.

Implications and Future Research

This study highlights the importance of nonformal support in using eHealth, for those lacking adequate informal support. Kebede et al [54] found that attainment of digital competence among older adults depends on awareness of existing technology and the availability of support and instruction. They stated that personalized training and opportunities for need-based learning in a safe environment facilitate skill acquisition and initial digital engagement if the learning environment is accessible and inclusive for older adults [54]. Our findings indicate that public libraries can provide this type of support for digital health skill development. The results shed light on the motivations of participants, mainly older adults, to engage in nonformal support for eHealth use, advancing digital health literacy. While existing literature has focused on facilitators and barriers of digital (health) technology adoption by older adults, measuring eHealth or digital health literacy and the effectiveness of such interventions, often in research settings [53,54,70-72], this study

XSL•FO
RenderX

reveals motivations for participation in the real world. It highlights how motivations are associated with their specific needs for support and training, a need for security in access, and a desire for independent use of digital health care.

It would be interesting to explore the group of nonusers and users who do not seek nonformal support. Earlier reflections on our results suggested that this study population possessed certain characteristics, resources, and needs that align with seeking support in the public library context. Research shows that eHealth nonusers are common among other age and socioeconomic groups and that their needs for support and eHealth use are diverse [4,8,64,65,73-75]. Scholars find that populations rely on different types of support, it being formal, nonformal, and informal support, to gain access to digital skill development and eHealth [76-78]. Findings from Nygren and Hayat underline the importance of informal support to overcome digital literacy barriers. However, other findings indicate that digital inequality can manifest itself in informal support, where those who experience the most difficulties in accessing and use of digital tools are also those who have fewer resources for high-quality informal support [65,74]. Furthermore, van Deursen et al [79] show that in comparison to formal support, informal support might be insufficient to overcome all skill and content barriers for beneficial internet use. The barriers of limited access and poor quality of informal support were also found in this study, additionally, we found that discomfort with being dependent on others was also important. Smit et al [78] showed that Dutch citizens with low literacy overcome digital and literacy barriers by support from family, friends, community, and volunteers and base decisions on who they ask for support based upon the level of privacy surrounding a topic, the affective dimension involved, and the depth of knowledge possessed by the supportive actors [78]. These and our findings implicate that different populations are likely to seek support from different sources that fit their needs and personal circumstances. Further research is needed to understand the motivations, experiences, and needs in support of eHealth use by both digital technology users and nonusers. Findings can provide insights into what qualities eHealth support should entail to ensure equitable access to support and consequently to eHealth.

Conclusions

This study indicates that participants seeking nonformal support for eHealth use were driven by a need to secure independent access and use of digital health care. This need is derived from the necessity to adapt to a digital society, the expected or current need to use digital health, and a desire to independently engage in a digital society. In general, participants were motivated to develop digital health skills they feel are necessary for societal participation. Additionally, the eHealth course met the need for social connection and lifelong learning. The study sample was a predominantly homogenous group of older female adults with language skills, access to up-to-date devices, and time. It is likely that the need for support to gain independent access and use of digital health care is present in other populations. Future research should explore the needs and attitudes of users and nonusers of digital health toward support in eHealth access.

Acknowledgments

We would like to thank Anita Suijkerbuijk for her contributions during the set-up and execution of the study. We would also like to thank Sumit Mehra for his feedback and contributions to finalizing the article. This study was financially supported by the Dutch Ministry of Health for scientific publication and the Netherlands Organization for Health Research and Development (ZonMw project 555002013).

Data Availability

The datasets generated during or analyzed during this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview guide used to conduct semistructured face-to-face interviews with participants of the eHealth course DigiVitaler in the context of Dutch public libraries about their motivations to seek support in eHealth use.
[DOCX File, 16 KB - formative_v9i1e60612_app1.docx]

Multimedia Appendix 2
Demographics of the study participants interviewed between April and June 2022. Participants were categorized in educational level based upon the international standard classification of education level (ISCED). Rurality of the public library the participants attended the eHealth course was determined based upon zip code of the library and the indexed rurality by Netherlands Statistics.
[DOCX File, 19 KB - formative_v9i1e60612_app2.docx]

Multimedia Appendix 3

XSL•FO
RenderX

Overview of inductive themes and inductive topics and number of codes per topic derived from twenty interviews that were conducted with participants from an eHealth course in Dutch public library settings.
[DOCX File, 14 KB - formative_v9i1e60612_app3.docx]

Checklist 1
Standards for Reporting Qualitative Research checklist
[DOCX File, 16 KB - formative_v9i1e60612_app4.docx]

References

1. da Fonseca MH, Kovaleski F, Picinin CT, Pedroso B, Rubbo P. E-health practices and technologies: a systematic review from 2014 to 2019. Healthcare (Basel) 2021;9(9):1192. [doi: 10.3390/healthcare9091192]
2. Shaw T, McGregor D, Brunner M, Keep M, Janssen A, Barnet S. What is eHealth (6)? Development of a conceptual model for eHealth: qualitative study with key informants. J Med Internet Res 2017 Oct 24;19(10):e324. [doi: 10.2196/jmir.8106] [Medline: 29066429]
3. Abernethy A, Adams L, Barrett M, et al. The promise of digital health: then, now, and the future. NAM Perspect 2022;2022. [doi: 10.31478/202206e] [Medline: 36177208]
4. Goedhart NS, Verdonk P, Dedding C. “Never good enough.” A situated understanding of the impact of digitalization on citizens living in a low socioeconomic position. Policy Internet 2022 Dec;14(4):824-844. [doi: 10.1002/poi3.315]
5. Goedhart NS, Zuiderent-Jerak T, Woudstra J, Broerse JEW, Betten AW, Dedding C. Persistent inequitable design and implementation of patient portals for users at the margins. J Am Med Inform Assoc 2021 Feb 15;28(2):276-283. [doi: 10.1093/jamia/ocaa273] [Medline: 33463691]
6. Chesser A, Burke A, Reyes J, Rohrberg T. Navigating the digital divide: a systematic review of eHealth literacy in underserved populations in the United States. Inform Health Soc Care 2016;41(1):1-19. [doi: 10.3109/17538157.2014.948171] [Medline: 25710808]
7. Xu RH, Zhou LM, Wong EY, Wang D. The association between patients’ eHealth literacy and satisfaction with shared decision-making and well-being: multicenter cross-sectional study. J Med Internet Res 2021 Sep 24;23(9):e26721. [doi: 10.2196/26721] [Medline: 34559062]
8. Deursen A, Helsper EJ. The third-level digital divide: who benefits most from being online? In: Communication and Information Technologies Annual: Emerald Group Publishing Limited; 2015, Vol. 10:29-52. [doi: 10.1108/S2050-206020150000010002]
9. Al-Dhahir I, Reijnders T, Faber JS, et al. The barriers and facilitators of eHealth-based lifestyle intervention programs for people with a low socioeconomic status: scoping review. J Med Internet Res 2022 Aug 24;24(8):e34229. [doi: 10.2196/34229] [Medline: 36001380]
10. Van Dijk J. The Digital Divide: John Wiley & Sons; 2020.
11. Livingood WC, Bautista MAB, Smotherman C, et al. Comparative study of different SES neighborhood clinics for health literacy and internet access. Digital Health 2022;8:20552076221123715. [doi: 10.1177/20552076221123715] [Medline: 36081750]
12. Lythreatis S, Singh SK, El-Kassar AN. The digital divide: a review and future research agenda. Technol Forecast Soc Change 2022 Feb;175:121359. [doi: 10.1016/j.techfore.2021.121359]
13. Latulippe K, Hamel C, Giroux D. Social health inequalities and eHealth: a literature review with qualitative synthesis of theoretical and empirical studies. J Med Internet Res 2017 Apr 27;19(4):e136. [doi: 10.2196/jmir.6731] [Medline: 28450271]
14. Alvarez-Galvez J, Salinas-Perez JA, Montagni I, Salvador-Carulla L. The persistence of digital divides in the use of health information: a comparative study in 28 European countries. Int J Public Health 2020 Apr;65(3):325-333. [doi: 10.1007/s00038-020-01363-w] [Medline: 32277246]
15. Elena-Bucea A, Cruz-Jesus F, Oliveira T, Coelho PS. Assessing the role of age, education, gender and income on the digital divide: evidence for the European union. Inf Syst Front 2021 Aug;23(4):1007-1021. [doi: 10.1007/s10796-020-10012-9]
16. Hilbert M. The end justifies the definition: the manifold outlooks on the digital divide and their practical usefulness for policy-making. Telecomm Policy 2011 Sep;35(8):715-736. [doi: 10.1016/j.telpol.2011.06.012]
17. van Deursen AJ, van Dijk JA. The first-level digital divide shifts from inequalities in physical access to inequalities in material access. New Media Soc 2019 Feb;21(2):354-375. [doi: 10.1177/1461444818797082] [Medline: 30886536]
18. Standaar L, van Tuyl L, Suijkerbuijk A, Brabers A, Friele R. Differences in eHealth access, use, and perceived benefit between different socioeconomic groups in the Dutch context: secondary cross-sectional study. JMIR Form Res 2025:e49585. [doi: 10.2196/49585] [Medline: 39773883]
19. Faber JS, Al-Dhahir I, Reijnders T, et al. Attitudes toward health, healthcare, and eHealth of people with a low socioeconomic status: a community-based participatory approach. Front Digital Health 2021;3:690182. [doi: 10.3389/fdgth.2021.690182] [Medline: 34713165]

XSL•FO
RenderX

20. Frishammar J, Essén A, Simms C, Edblad R, Hardebro V. Older individuals and digital healthcare platforms: usage motivations and the impact of age on postadoption usage patterns. IEEE Trans Eng Manage 2023;70(8):2903-2919. [doi: 10.1109/TEM.2022.3187792]
21. Freeman T, Fisher M, Foley K, et al. Barriers to digital health services among people living in areas of socioeconomic disadvantage: research from hospital diabetes and antenatal clinics. Health Promot J Austr 2022 Jul;33(3):751-757. [doi: 10.1002/hpja.540] [Medline: 34510601]
22. Scheerder A, van Deursen A, van Dijk J. Determinants of internet skills, uses and outcomes. A systematic review of the second- and third-level digital divide. Telemat Inform 2017 Dec;34(8):1607-1624. [doi: 10.1016/j.tele.2017.07.007]
23. The ongoing journey to commitment and transformation: digital health in the WHO European region. World Health Organization. 2023. URL: https://www.who.int/europe/publications/m/item/digital-health-in-the-who-european-region-the-ongoing-journey-to-commitment-and-transformation [accessed 2025-01-23]
24. OECD skills outlook 2023: skills for a resilient green and digital transition. : OECD; 2023. [doi: 10.1787/27452f29-en]
25. Norman CD, Skinner HA. eHealth literacy: essential skills for consumer health in a networked world. J Med Internet Res 2006;8(2):e9. [doi: 10.2196/jmir.8.2.e9]
26. Azzopardi-Muscat N, Sørensen K. Towards an equitable digital public health era: promoting equity through a health literacy perspective. Eur J Public Health 2019 Oct 1;29(Supplement_3):13-17. [doi: 10.1093/eurpub/ckz166] [Medline: 31738443]
27. Advice stimulating appropriate care and digital care [Web page in Dutch]. Rijksoverheid. 2022. URL: https://www.rijksoverheid.nl/documenten/rapporten/2020/07/16/advies-stimuleren-van-passende-zorg-en-digitale-zorg [accessed 2025-01-22]
28. Bibliotheekconvenant 2020-2023 [Web page in Dutch]. Bibliotheekconvenant. 2023. URL: https://www.bibliotheeknetwerk.nl/sites/default/files/documents/Bibliotheekconvenant%202020-2023_dt.pdf [accessed 2023-10-13]
29. Digital government information point [Web page in Dutch]. informatiepunt. 2024. URL: https://www.informatiepuntdigitaleoverheid.nl/?lang=en [accessed 2023-10-13]
30. Giest S, Samuels A. Administrative burden in digital public service delivery: the social infrastructure of library programs for e‐inclusion. Rev Policy Res 2023 Sep;40(5):626-645. [doi: 10.1111/ropr.12516]
31. van de Burgt A, Klaren M. Public library services related to basic skills for adults [Web page in Dutch]. Bibliotheeknetwerk. 2022. URL: https://www.bibliotheeknetwerk.nl/publicaties/basisvaardigheden-voor-volwassenen-2021 [accessed 2023-10-13]
32. Malcolm J. Informality and Formality in Learning: A Report for the Learning and Skills Research Centre: Learning and Skills Research Centre; 2003. [doi: 10.13140/RG.2.1.1297.3924]
33. Dicicco-Bloom B, Crabtree BF. The qualitative research interview. Med Educ 2006 Apr;40(4):314-321. [doi: 10.1111/j.1365-2929.2006.02418.x] [Medline: 16573666]
34. Bhattacherjee A. Social science research: principles, methods, and practices, global text project. Digital Commons @ University of South Florida. 2012. URL: https://digitalcommons.usf.edu/oa_textbooks/3 [accessed 2025-01-30]
35. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014 Sep;89(9):1245-1251. [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]
36. Keij B, Versluis A, Alblas EE, Keuper JJ, van Tuyl LHD, van der Vaart R. E-healthmonitor 2023 stand van zaken digitale zorg [Report in Dutch]. : Bilthoven, Rijksinstituut voor Volksgezondheid en Milieu; 2023. [doi: 10.21945/RIVM-2024-0008]
37. Ministery of Economic Affairs and Climate Policy. Dutch Digitisation Strategy 2.0. Nederland Digitaal. 2019. URL: https://www.nederlanddigitaal.nl/documenten/publicaties/2019/11/13/english-version-of-the-dutch-digitalisation-strategy-2.0 [accessed 2023-05-05]
38. Regeerprogramma hoofdstuk 6a [Web page in Dutch]. Rijksoverheid.: Zorg; 2024. URL: https://www.rijksoverheid.nl/regering/regeerprogramma/6a-zorg [accessed 2025-01-30]
39. Peeters JM, Krijgsman JW, Brabers AE, Jong JDD, Friele RD. Use and uptake of eHealth in general practice: a cross-sectional survey and focus group study among health care users and general practitioners. JMIR Med Inform 2016 Apr 6;4(2):e11. [doi: 10.2196/medinform.4515] [Medline: 27052805]
40. Digital Economy and Society Index (DESI) 2021. European Commission. 2021. URL: https://ec.europa.eu/newsroom/dae/redirection/document/80493 [accessed 2023-10-31]
41. Basic skills dashboard for adults [Web page in Dutch]. Bibliotheek. URL: https://www.bibliotheeknetwerk.nl/dashboard/dashboard-basisvaardigheden-voor-volwassenen [accessed 2024-11-21]
42. Institute for Statistics International Standard Classification of Education: ISCED 2011. UNESCO. 2012. URL: https://uis.unesco.org/sites/default/files/documents/international-standard-classification-of-education-isced-2011-en.pdf [accessed 2024-05-16]
43. Population on January 1 and average; gender, age and region [Web page in Dutch]. Centraal Bureau voor de Statistiek. 2024. URL: https://www.cbs.nl/nl-nl/cijfers/detail/03759ned?dl=39E0B [accessed 2024-06-19]
44. de Vries H, Dijkstra M, Kuhlman P. Self-efficacy: the third factor besides attitude and subjective norm as a predictor of behavioural intentions. Health Educ Res 1988;3(3):273-282. [doi: 10.1093/her/3.3.273]
45. Fishbein M, Ajzen I. Predicting and Changing Behavior: The Reasoned Action Approach: Psychology Press; 2011.
46. de Vries H. An integrated approach for understanding health behavior; the I-change model as an example. Psychol Behav Sci Int J 2017;2(2):555-585. [doi: 10.19080/PBSIJ.2017.02.555585]

XSL•FO
RenderX

47. Braun V, Clarke V. Reflecting on reflexive thematic analysis. Qual Res Sport Exerc Health 2019 Aug 8;11(4):589-597. [doi: 10.1080/2159676X.2019.1628806]
48. Braun V, Clarke V. Conceptual and design thinking for thematic analysis. Qual Psychol 2022;9(1):3-26. [doi: 10.1037/qup0000196]
49. Köttl H, Gallistl V, Rohner R, Ayalon L. “But at the age of 85? Forget it!”: internalized ageism, a barrier to technology use. J Aging Stud 2021 Dec;59:100971. [doi: 10.1016/j.jaging.2021.100971] [Medline: 34794716]
50. Schwab K. The Fourth Industrial Revolution: Crown Currency; 2017.
51. Luijkx K, Peek S, Wouters E. “Grandma, you should do it—It’s cool” older adults and the role of family members in their acceptance of technology. Int J Environ Res Public Health 2015 Dec 5;12(12):15470-15485. [doi: 10.3390/ijerph121214999] [Medline: 26690188]
52. Czaja SJ, Charness N, Fisk AD, et al. Factors predicting the use of technology: findings from the Center for Research and Education on Aging and Technology Enhancement (CREATE). Psychol Aging 2006 Jun;21(2):333-352. [doi: 10.1037/0882-7974.21.2.333] [Medline: 16768579]
53. Wilson J, Heinsch M, Betts D, Booth D, Kay-Lambkin F. Barriers and facilitators to the use of e-health by older adults: a scoping review. BMC Public Health 2021 Aug 17;21(1):1556. [doi: 10.1186/s12889-021-11623-w] [Medline: 34399716]
54. Kebede AS, Ozolins LL, Holst H, Galvin K. Digital engagement of older adults: scoping review. J Med Internet Res 2022 Dec 7;24(12):e40192. [doi: 10.2196/40192] [Medline: 36477006]
55. Ware P, Bartlett SJ, Paré G, et al. Using eHealth technologies: interests, preferences, and concerns of older adults. Interact J Med Res 2017 Mar 23;6(1):e3. [doi: 10.2196/ijmr.4447] [Medline: 28336506]
56. Ollevier A, Aguiar G, Palomino M, Simpelaere IS. How can technology support ageing in place in healthy older adults? A systematic review. Public Health Rev 2020 Nov 23;41(1):26. [doi: 10.1186/s40985-020-00143-4] [Medline: 33292707]
57. Robbins TD, Lim Choi Keung SN, Arvanitis TN. E-health for active ageing: a systematic review. Maturitas 2018 Aug;114:34-40. [doi: 10.1016/j.maturitas.2018.05.008] [Medline: 29907244]
58. Mcmellon CA, Schiffman LG. Cybersenior empowerment: how some older individuals are taking control of their lives. J Appl Gerontol 2002 Jun;21(2):157-175. [doi: 10.1177/07364802021002002]
59. Wiles JL, Leibing A, Guberman N, Reeve J, Allen RES. The meaning of “aging in place” to older people. Gerontologist 2012 Jun;52(3):357-366. [doi: 10.1093/geront/gnr098] [Medline: 21983126]
60. Longer at home program [Report in Dutch]. : Den Haag: Ministerie van Volksgezondheid, Welzijn en Sport; 2018 URL: https://www.rijksoverheid.nl/documenten/rapporten/2018/06/15/programma-langer-thuis [accessed 2025-04-20]
61. Helsper EJ, van Deursen A. Do the rich get digitally richer? Quantity and quality of support for digital engagement. Inf Commun Soc 2017 May 4;20(5):700-714. [doi: 10.1080/1369118X.2016.1203454]
62. Maslow AH. A theory of human motivation. Psychol Rev 1943;50(4):370-396. [doi: 10.1037/h0054346]
63. Choudhary C, Nidhi B. Barriers affecting the effectiveness of digital literacy training programs (DLTPs) for marginalised populations: a systematic literature review. J Tech Educ Train 2022 Jun 30;14(1):110-127. [doi: 10.30880/jtet.2022.14.01.010]
64. Standaar LMB, Israel AMC, van der Vaart R, et al. Support in digital health skill development for vulnerable groups in a public library setting: perspectives of trainers. Front Digital Health 2024;6:1519964. [doi: 10.3389/fdgth.2024.1519964] [Medline: 39872241]
65. Goedhart NS, Broerse JEW, Kattouw R, Dedding C. ‘ Just having a computer doesn’t make sense’: the digital divide from the perspective of mothers with a low socio-economic position. New Media Soc 2019 Nov;21(11-12):2347-2365. [doi: 10.1177/1461444819846059]
66. Leyenaar MH. From pioneers to players: women in the netherlands. women and politics around the world. In: Gelb J, Palley ML, editors. A Comparative History and Surve: Bloomsbury Academic; 2009, Vol. 2:485-502.
67. Asmar A, Van Audenhove L, Mariën I. Social support for digital inclusion: towards a typology of social support patterns. Soc Incl 2020 May;8(2):138-150. [doi: 10.17645/si.v8i2.2627]
68. Quan-Haase A, Williams C, Kicevski M, Elueze I, Wellman B. Dividing the grey divide: deconstructing myths about older adults’ online activities, skills, and attitudes. Am Behav Sci 2018 Aug;62(9):1207-1228. [doi: 10.1177/0002764218777572]
69. Petrovčič A, Reisdorf BC, Grošelj D, Prevodnik K. A typology of aging internet users: exploring digital gradations in internet skills and uses. Soc Sci Comput Rev 2023 Oct;41(5):1921-1940. [doi: 10.1177/08944393221117753]
70. Karnoe A, Kayser L. How is eHealth literacy measured and what do the measurements tell us? A systematic review. Knowl Manag E-Learn 2015 Dec 15;7(4):576-600. [doi: 10.34105/j.kmel.2015.07.038]
71. Kim H, Xie B. Health literacy in the eHealth era: a systematic review of the literature. Patient Educ Couns 2017 Jun;100(6):1073-1082. [doi: 10.1016/j.pec.2017.01.015] [Medline: 28174067]
72. Xie L, Zhang S, Xin M, Zhu M, Lu W, Mo PH. Electronic health literacy and health-related outcomes among older adults: a systematic review. Prev Med 2022 Apr;157:106997. [doi: 10.1016/j.ypmed.2022.106997]
73. Helsper EJ, Eynon R. Distinct skill pathways to digital engagement. Eur J Commun 2013 Dec;28(6):696-713. [doi: 10.1177/0267323113499113]
74. Courtois C, Verdegem P. With a little help from my friends: an analysis of the role of social support in digital inequalities. New Media Soc 2016 Sep;18(8):1508-1527. [doi: 10.1177/1461444814562162]

XSL•FO
RenderX

75. Kaihlanen AM, Virtanen L, Buchert U, et al. Towards digital health equity—a qualitative study of the challenges experienced by vulnerable groups in using digital health services in the COVID-19 era. BMC Health Serv Res 2022 Feb 12;22(1):188. [doi: 10.1186/s12913-022-07584-4] [Medline: 35151302]
76. Nygren H, Nissinen K, Hämäläinen R, De Wever B. Lifelong learning: formal, non‐formal and informal learning in the context of the use of problem‐solving skills in technology‐rich environments. Brit J Educ Tech 2019 Jul;50(4):1759-1770. [doi: 10.1111/bjet.12807]
77. Hayat TZ, Brainin E, Neter E. With some help from my network: supplementing eHealth literacy with social ties. J Med Internet Res 2017 Mar 30;19(3):e98. [doi: 10.2196/jmir.6472] [Medline: 28360024]
78. Smit A, Swart J, Broersma M. Bypassing digital literacy: marginalized citizens' tactics for participation and inclusion in digital societies. New Media Soc 2024;0. [doi: 10.1177/14614448231220383]
79. van Deursen AJAM, Courtois C, van Dijk JAGM. Internet skills, sources of support, and benefiting from internet use. Int J Hum Comput Interact 2014 Apr 3;30(4):278-290. [doi: 10.1080/10447318.2013.858458]

Abbreviations

ISCED: International Standard Classification of Education
SRQR: Standards for Reporting Qualitative Research

Edited by A Mavragani; submitted 17.05.24; peer-reviewed by N Goedhart, T Huang; revised version received 30.01.25; accepted 05.03.25; published 28.04.25.

Please cite as:
Standaar L, Israel AM, van der Vaart R, Keij B, van Lenthe FJ, Friele R, Beenackers MA, van Tuyl LHD
The Motivations of Citizens to Attend an eHealth Course in the Public Library: Qualitative Interview Study
JMIR Form Res 2025;9:e60612
URL: https://formative.jmir.org/2025/1/e60612
doi:10.2196/60612

© Lucille Standaar, Adriana Margje Israel, Rosalie van der Vaart, Brigitta Keij, Frank J van Lenthe, Roland Friele, Mariëlle A Beenackers, Lilian Huibertina Davida van Tuyl. Originally published in JMIR Formative Research (https://formative.jmir.org), 28.4.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Adolescent Emoji Use in Text-Based Messaging: Focus Group Study

Matt Minich, PhD; Bradley Kerr, MS; Megan Moreno, MD, MPH, MSED

Department of Pediatrics, School of Medicine and Public Health, University of Wisconsin-Madison, 2870 University Ave, Madison, WI, United States

Corresponding Author:
Matt Minich, PhD
Department of Pediatrics, School of Medicine and Public Health, University of Wisconsin-Madison, 2870 University Ave, Madison, WI, United States

Abstract

Background: Adolescents increasingly communicate through text-based messaging platforms such as SMS and social media messaging. These are now the dominant platforms for communication between adolescents, and adolescents use them to obtain emotional support from parents and other adults. The absence of nonverbal cues can make it challenging to communicate emotions on these platforms, however, so users rely on emojis to communicate sentiment or imbue messages with emotional tone. While research has investigated the functions of emojis in adult communication, less is known about adolescent emoji use.

Objective: This study sought to understand whether the pragmatic functions of adolescent emoji use resemble those of adults, and to gain insight into the semantic meanings of emojis sent by adolescents.

Methods: Web-based focus groups were conducted with a convenience sample of adolescents, in which participants responded to questions about their use and interpretation of emojis and engaged in unstructured interactions with one another. Two trained coders analyzed transcripts using a constant comparative coding procedure to identify themes in the discussion.

Results: A total of 6 focus groups were conducted with 31 adolescent participants (mean age 16.2, SD 1.5 years). Discussion in the groups generally fell into 4 themes: emojis as humorous or absurd, emokis as insincere or complex expressions of setiment, emojis as straightforward experssions of sentiment, and emojis as having context-dependent meanings. Across themes, participants often described important differences between their own emoji use and emoji use by adults.

Conclusions: Adolescent focus group participants described patterns of emoji use that largely resembled those observed in studies of adults. Like adults, our adolescent participants described emojis' semantic meanings as being highly flexible and context-dependent. They also described both phatic and emotive functions of emoji use but described both functions in ways that differed from the patterns of emoji use described in adult samples. Adolescents described their phatic emoji use as absurd and described their emotive emoji use as most often sarcastic. These findings suggest that emoji use serves similar pragmatic functions for both adolescents and adults, but that adolescents see their emoji use as more complex than adult emoji use. This has important implications for adults who communicate with adolescents through text-based messaging and for researchers interested in adolescents' text-based communication.

(JMIR Form Res 2025;9:e59640) doi:10.2196/59640

KEYWORDS

communication; text messaging; smartphones; emoji; focus groups; adolescent; teen; youth; teenagers; text; phone; messaging; text communication; emotion

Introduction

Overview

Adolescent communication is increasingly mediated by text-based communication platforms like SMS or social media messaging. About 95% of US adolescents aged 13 - 17 years have a smartphone for personal use [1], and adolescents in this age group send and receive an average of 67 text messages each day [2]. Text messaging is now the dominant mode of communication between adolescents [2] and is increasingly important to adolescents' relationships with parents and other adults [3]. For example, text-based communication is an effective way for healthcare practitioners to reach young or adolescent clients [4] and can help therapists discuss sensitive topics with younger clients [5]. Many parents also surveil the text exchanges of their adolescent children—in a nationwide survey, 64% of parents said they regularly checked the contents of their child's phone, including text message logs [6]. Thus, many adolescents' most important relationships are affected in some way by the features of text-based messaging platforms.

One feature that applies to text-based messaging of all kinds is the absence of nonverbal cues, such as body language or tone

XSL•FO
RenderX

of voice, that people use to convey emotional context [7]. Users often address this deficit with emojis, which can represent emotional states through pictograms [8]. However, emojis can create confusion because users often interpret them differently [9]. Past research has addressed this confusion both by identifying patterns in emojis' pragmatic functions [10] and by reducing ambiguity around emojis' semantic meanings [11,12], but that work has largely relied on samples of adult emoji users. Because emoji use and interpretation differ across age groups [13,14], it is unclear whether the insights from this research can be extended to adolescents. Given the emerging importance of text-based messaging to adolescent communication, this study explores adolescents' perspectives on their own emoji use. Its aims are twofold: to understand whether the pragmatic functions of adolescent emoji use resemble those of adults and to gain insight into the semantic meanings of emojis sent by adolescents.

Pragmatic Functions of Adult Emoji Use

The earliest known use of emotive pictograms in computer-mediated communication was by Carnegie Mellon Professor Scott Fahlman, who proposed in 1982 that users of an intradepartmental message board distinguish serious posts from humorous ones by marking them with the symbols :-) or :-([8]. Though similar symbols (called emoticons) are still used, they have since been largely supplanted by the Unicode emoji library, which gives users access to a set of more than 3600 pictographs that are mostly standardized across devices [15]. The use of emojis is now nearly ubiquitous—more than 5 billion emojis are exchanged each day on Facebook alone [15], and the "face with tears of joy" emoji ( ) was chosen as the Oxford English Dictionary's Word of the Year in 2015.

When asked about their own emoji use, adults say they use the pictograms both to express emotions directly and to modify the tone of messages. For example, surveyed English-speaking adults said they understand emojis as tools for tone modification, in which the emoji modifies text and clarifies its interpretation [16]. US adults have reported using emojis primarily to express sentiment, to strengthen expressions, or to adjust the tone of messages [17]. Thus, the utility of emojis for adults largely overlaps with that of emoticons, which adults also use to communicate feelings and reduce the ambiguity of text [18]. For emojis and emoticons alike, surveyed adults have said they send pictograms primarily to express their feelings, to strengthen the content of a message, or to imbue a message with an element of comedy or fun [19]. Notably, adults' reasons for sending emojis differ slightly across age groups, with younger adults reporting a more diverse set of motives (including, for example, a desire to make a message ironic or sarcastic) than older adults [19]. Research has also found that older adults use a less diverse emoji vocabulary than younger adults and are generally less likely to use emojis in text-based communication [20].

Theorists have suggested adult emoji use can be broadly categorized as serving 1 of 2 pragmatic functions: a phatic function or an emotive function [10]. The phatic function of emojis describes their use as a fill-in for "small talk:" semiverbal gestures that maintain connection and keep communication light and friendly but lack specific semantic meanings. For example, a user might send a simple smile emoji ( ) to open or close a conversation or to fill an uncomfortable period of silence. This aligns with research finding that emojis are usually associated with positive sentiments [21], and that they can be used to facilitate connections in a way that is agnostic to a conversation's emotional content [22]. The emotive function of emojis describes their use to convey the sender's emotional state or to imbue a message with emotional context. This may be done by using emojis in place of words describing emotions (for example, by sending in place of the words "I'm angry"), or by adding emojis to complete statements (eg, "I saw a car accident today ").

Semantic Ambiguity of Emojis

Though users often send emojis to reduce ambiguity about messages' emotional context, emojis can create confusion because receivers interpret them differently. For example, when experimental participants were asked to interpret expressive face emojis, they disagreed on even basic sentiment assignments (positive, negative, or neutral) 25% of the time [23]. This confusion persisted even when the emojis were accompanied by relevant text. Similarly, experimental participants tended to rate text messages paired with face emojis as more emotionally ambiguous than messages paired with nonface emojis [21]. When experimenters asked participants to interpret emojis contained within text messages, participants offered a wide variety of interpretations. For example, some participants understood the loudly crying emoji ( ) to signify sadness while others perceived this emoji as sarcastic or exaggerated [24].

Research has shown that the interpretation of emojis can be affected by the ages of the sender and receiver. For example, older adults were found more likely to interpret emojis as having positive meanings [14], and younger Chinese adults were more likely to use "positive emojis" (such as the simple smile) to express negative sentiments [25]. Older adults have also proven less likely than younger adults to recognize emojis as indicators of sarcasm [26], and more likely to express confusion about emojis or choose literal interpretations of their meaning [13]. Older adults have reported less confidence in their interpretations of emojis and perceptions that emojis are difficult to use correctly [20].

Understanding Adolescent Emoji Use

Among adults, research shows that both the pragmatic functions and semantic interpretations of emojis differ as a factor of age. Older adults have reported less diverse motivations for using emojis [19] than younger adults, for example, and more difficulty interpreting emojis they receive in messages from others [13,20]. This suggests that adolescent emoji use might differ from adult emoji use in meaningful ways, which can create challenges both for adults seeking to communicate with adolescents through text-based messaging and for researchers interested in text-based communication between adolescents.

Understanding how adolescents use and interpret emojis is particularly important because text-based messaging has emerged as an essential platform for adolescent communication. Text messaging behavior peaks around 11th grade [27] and slows in young adulthood [28], leading some scholars to suggest that the affordances of this medium may make it particularly

XSL•FO
RenderX

well-suited for adolescents' developmental needs. Indeed, the asynchronous and socially distanced nature of text messaging may facilitate adolescents' desires for autonomy, identity development, and the formation of intimate relationships [27]. Text-based messaging is also increasingly important to communication between adolescents and adults. Adolescents often use these platforms to request emotional support from parents, for example, and have reported improvements in their offline relationships with parents on days when they exchange more text messages with them [3]. Adolescents have expressed desires to use text-based messaging platforms for communication with adults about emotional support [4] and have said text-based platforms provide them with a greater sense of authority and control in interactions with counselors and therapists [29].

Emotive pictographs like emojis have long been used to communicate emotions through text [8], but age-related differences in the function and interpretation of emojis could facilitate miscommunication. This could exacerbate existing problems with communication between adolescents and adults, such as communication between adolescents and their parents. Parents and adolescents often fail to accurately understand one another's cognitive and affective states even when communicating face-to-face [30,31]. This miscommunication has been described as a failure of empathic accuracy [32], or the accurate prediction of a conversation partner's thoughts and feelings. Adults often use emojis to improve empathic accuracy [33] in text-based exchanges, but it is unclear whether adolescents use them for a similar purpose. Thus, this study seeks to better understand how adolescents use emojis in these exchanges, and how they understand the semantic meanings of those emojis. Given the persistence of emotive pictographs over decades of computer-mediated communication [8], these insights should remain useful even as communication platforms evolve.

Methods

Participants

Adolescent focus group participants were recruited through a node distribution method, in which contacts of the research team (staff at youth-serving organizations and participants in previous studies) used word-of-mouth to inform potential participants about the study. All participants provided assent and parental consent and then completed a demographic survey that assessed participants' age, gender, race, and ethnicity.

A total of 31 adolescents participated in the focus groups. A total of 48% of participants were female, and the average participant age was 16.2 (SD 1.5) years. Complete demographic information is presented in Table 1 below.

Table . Participant demographics. Self-reported demographic data of participants in 6 web-based focus groups conducted to identify themes in adolescents' understandings of their own use of emojis in text-based messaging.

Characteristic	Values, n (%)
Gender	
Cisgender female	14 (45.2)
Cisgender male	10 (32.3)
Non-binary	2 (6.5)
Transgender female	1 (3.2)
Transgender male	1 (3.2)
Other	1 (3.2)
Did not disclose	2 (6.5)
Ethnicity	
Non-Hispanic or non-Latino	26 (83.9)
Hispanic or Latino	3 (9.7)
Did not disclose	2 (6.5)
Race	
Caucasian or White	16 (51.6)
Asian or Asian-American	6 (19.4)
American Indian or Alaska Native	1 (3.2)
Black or African-American	1 (3.2)
Native Hawaiian or other Pacific Islander	0 (0.0)
Other	0 (0.0)
More than one race	5 (16.1)
Did not disclose	2 (6.5)

XSL•FO
RenderX

Data Collection Procedure

Adolescent perspectives were collected through 6 web-based focus groups, which allowed participants to express their experiences without restriction and encouraged them to build on one another's contributions. Participants completed a brief demographic questionnaire before joining the group. Groups were conducted using a digital conference call service by 2 trained facilitators (the first author and second author) and lasted up to 90 minutes. Both facilitators were White males employed at a large midwestern research institution, who both possessed MS degrees at the time of data collection. In focus groups, facilitators initially asked open-ended questions about emoji use and interpretation (eg, "What are some ways you use emojis in your communication?"), as well as specific questions about styles of emoji use (eg, "Do you usually use emojis to replace words or phrases, or to add emphasis?"). Participants were also shown several emojis, including the simple smile ( ) and 5 randomly selected face emojis (in the iPhone iOS style) asked to describe their interpretations of those emojis and the circumstances in which they might share them.

Ethical Considerations

The procedure used to collect data for this study was approved by the University of Wisconsin-Madison Health Sciences Institutional Review Board (#2019‑0839), and no adverse events were reported. Informed consent from parents or guardians and informed assent from adolescent participants were obtained through a digitally delivered form before focus group participation. Informed consent documents clarified that deidentified focus group data could be made available to investigators outside the original study team. Privacy and confidentiality of participant data were protected by redacting all identifying information from focus group transcripts before analysis. All documents containing identifiable data (eg, informed consent documents and original focus group recordings) were stored on a secure server only accessible by members of the study team. Participants were compensated US $40 for participation in the study, regardless of whether they completed their focus group participation (no participants chose to leave focus groups early).

Analytical Approach

Analysis was performed using a constant comparative procedure, in which 2 trained coders applied inductive reasoning to identify themes [34]. Originally developed for use in Grounded Theory methodologies [35], this procedure allows analysts to identify themes from complex qualitative data sets without relying on a priori hypotheses. This approach was appropriate for this study both because of the dearth of past research into adolescent emoji use and because of the relatively unstructured nature of the data, which included not only responses to researchers' prompts but also unplanned interactions between participants.

After 3 focus groups had been conducted, coders began reviewing focus groups for theoretical saturation: a point at which new ideas or themes no longer emerged from discussions. Coders identified theoretical saturation after the completion of 6 focus groups, signaling the end of data collection and the beginning of thematic analysis.

Thematic analysis was conducted in 3 stages. First, coders independently read transcripts from 2 focus groups, each developing their own draft codebook. The coders then compared and discussed these codebooks, consolidating them into a single codebook that both coders agreed upon. This codebook contained a set of superordinate "branch" codes, each of which encompassed a set of more specific, subordinate "leaf" codes.

In the second stage of thematic analysis, coders refined the codebook by independently coding the same 2 additional transcripts and then meeting to compare results. Coders identified and discussed any discrepancies in their coding of these transcripts, then agreed on codebook revisions to resolve them. When coders agreed that all areas of confusion had been addressed, they divided and independently coded the remaining transcripts, so that each transcript was coded by a single coder. While coding this final set of transcripts, coders held regular meetings to discuss any further need for codebook revisions and to reach an agreement on the coding of any ambiguous statements.

In the third stage of thematic analysis, coders met to review the coded transcripts and identify overarching themes and subthemes. Coders reread excerpts assigned to each superordinate "branch" and subordinate "leaf" code, then independently collapsed these codes into larger "theme" groups. Coders then met to discuss the similarities and differences of the 2 sets of themes, consolidating them into a single set of shared themes. These themes were then reviewed by an independent youth advisory board consisting of 6 adolescents, who affirmed these themes as a reasonable summary of transcript content.

Finally, after thematic analysis was complete, coders considered how the agreed-upon themes aligned with their a priori areas of theoretical interest: identifying pragmatic functions of emoji use (phatic and emotive) and resolving the semantic ambiguity of emojis.

Results

Overview

Coders identified 4 emergent themes in focus group transcripts. One described the phatic function of emojis: "emojis as humorous or absurd", and 2 described their emotive functions: "emojis as insincere or complex expressions of sentiment and emojis as straightforward expressions of sentiment". Finally, one theme spoke to the semantic ambiguity of emojis: "emojis as having context-dependent meanings".

Notably, some statements made by participants were classified as pertaining to more than one theme. The extent of this crossover, as well as the total number of statements pertaining to each theme, is recorded in Table 2 below.

XSL•FO
RenderX

Table . Distribution of participant statements across themes. A matrix denoting the number of statements made by adolescent focus group participants coded for each of the 4 themes identified during analysis. Intersecting columns and rows represent the number of statements coded as 2 themes (eg, column 1 of row 2 describes the number of statements coded as both theme 1 and theme 2).

	Theme 1	Theme 2	Theme 3	Theme 4
Theme 1	39	6	2	11
Theme 2	6	81	4	11
Theme 3	2	4	57	5
Theme 4	11	11	5	46

Theme 1: Emojis as Humorous or Absurd

Overview

When participants described phatic uses of emojis (ie, use of emojis to maintain connection without giving them semantic meaning), they described doing so in the context of humor or absurdity. In these cases, it was understood that the sender or receiver found humor in the act of using emojis itself.

Theme 1a: Emojis as Innately Humorous

Several participants said they understood emojis as being inherently humorous and described sending emojis for the intention only of communicating humor or social connection. The emojis themselves in these communications were either divorced from their original semantic meanings or were assumed to have no semantic meaning. In most cases, participants described the absence of semantic meaning as essential to the emoji's communicative function.

> *I think, like there's one, which is just like, as a red background and like a B [ ], and they were, I mean, they'd seemingly like have no meaning. And they still don't really like, take on that big of a meaning, but they were like the randomness for like, used for, um, like comedic purpose or for like an inside joke kind of thing.*
>
> *…The sparkle [ ], or the pixie, dust, I think was originally used to like enhance the meaning of something, or like making it aesthetic in some way. But now I'm just using it for random words that it doesn't really make sense in—but it's like funny in a way (laughs). It's kind of like ironically…*
>
> *…there was like the chair emoji [ ], which like people would just use it to be funny and like the point of it was that it had no meaning, but like it was just supposed to be like a random emoji with no meaning.*

Theme 1b: Emoji Spam

Participants also described engaging in a specific pattern of absurd emoji use called "Emoji spamming," in which they sent long strings of randomly selected emojis. Again, participants described these strings as not only lacking any specific semantic meaning but described their communicative function as being innately tied to their absence of meaning.

> *…Emoji spamming often just signifies, like, chaos. Like, I think it's often used to, like, maybe you saw something outside and you wanna say something about it sarcastically, so you take a picture of it... Maybe, like, let's say you take a picture of a car crash, then you're like, "Oh, this is such a pretty car," and then you emoji spam. It's just, like, used to be sarcastic and funny.*
>
> *…I know definitely some of my friends do this, like, if they're typically experiencing, like, a negative emotion, they'll like emoji spam and do a bunch of random, like, really random ones…*
>
> *…People, like, sometimes, like, spam emojis because they're, like, angry that they lost [a mobile game] or they're excited that they won. I really think it goes both ways…*

Theme 2: Emojis as Insincere or Complex Expressions of Sentiment

Overview

Participants tended to describe emojis as inappropriate for communicating sincere emotions and mood, though some did acknowledge using them for this purpose (see theme 3). Rejection of sincere emoji use was generally framed as (1) a broad aversion to the use of emojis, (2) a preference for the ironic or sarcastic use of emojis, or (3) the use of apparently positive emojis to communicate negative emotions or moods.

Theme 2a: Aversion to the Use of Emojis

Many participants said they did not use emojis frequently, and they often specified that they were less likely to use emojis to communicate negative emotions or in emotionally negative situations, when they described emojis as socially inappropriate:

> *…When I'm, like, in a serious conversation I don't usually use emojis because I feel, like, it's not as serious if I were to use emojis.*
>
> *I don't really use emojis all that much. So if I am feeling those [negative] emotions, I'll probably just verbalize and like text them instead of using emojis for it, since it's kind of probably more of a serious conversation.*

Theme 2b: Preference for Ironic or Sarcastic Use of Emojis

Many participants said they typically used emojis in ironic or sarcastic ways so that their intended meaning was the opposite of a literal interpretation. For example, participants might use a laughing emoji to communicate that a joke had failed to land or might use a "cool" emoji [ ] to lampoon a person or message:

XSL•FO
RenderX

> *..With the sunglasses, I kind of use that sarcastically, kind of, like, saying, "Oh, I'm so cool I did that." Like, "Oh, that's so cool I'm doing something."*
>
> *I am one of those people that use this emoji [ ] sarcastically because, like, I feel like if I wanted to show that I was happy through emoji I would put, like, the closed eye smiling [ ] or something but this one seems, like, kind of creepy when you look at it so I usually use it sarcastically.*
>
> *…Just people in general, like, they'll use an emoji and me and my friends will, like, use that emoji [ ] to make fun of the people who use them, like, genuinely. Like [we'll] use it as, like, satire I guess you could say because we think it's funny that people actually use them.*

Mentions of ironic/sarcastic emoji use were notably more frequent in most focus groups than mentions of sincere emoji use.

Theme 2c: Communicating Negative Moods With Positive Emoji

When prompted to describe their use of traditionally positive emojis, participants often said they used these emojis to communicate negative emotions. Participants described positive emojis as adding layers of irony or insincerity to negative emotions, or to communicate a reluctant acceptance of negative circumstances.

> *I don't know if ironic is the right word, but I would use it [ ] in the opposite sense. Like, you're smiling through the pain or something. I know it's terrible, but that's just how I would see them.*
>
> *a lot of times, you could use that emoji [ ] as, like, I don't know, like, you're smiling through the pain, or, like, used, like, sarcastically, saying, like, "Oh, I'm so happy I messed up on something."*
>
> *I guess non-adolescents that don't really like use emojis ironically would definitely think that like [ ] is extremely like happy and positive. But, I mean, really, it's not. It's pretty lifeless.*
>
> *I feel like, it also goes with like a sarcastic tone too in kind of a joking manner. Like I texted my best friend last night and I was like "It's 5:00a.m. and I'm awake "*

Theme 3: Emojis as Straightforward Expressions of Sentiment

Though participants generally agreed that adults were more likely to use emojis to communicate sincere emotions, some described instances where they used emojis without irony or humor. In many such cases, they described the emoji as referring to a nameless, specific circumstance or emotional state:

> *Maybe I, I dropped my like pencil in the hallway and I had to like pick it up, like I would use that emoji [ ] like being like, "Oh yeah, so stressed about it."*
>
> *I'll probably just use like an emoji, like a heart or something, or a smiley. Or like something like that to just, you know, convey the message that, you know, it's how I'm feeling.*
>
> *I use emojis when I don't want to write out like everything. Like if I think something's funny I can just do like a laughing emoji [ ] and yeah, it's just easier to get the point across without having to write.*

Theme 4: Emojis as Having Context-Dependent Meanings

Finally, participants described emojis as having flexible meanings that shifted to suit the audience or conversational context. They cautioned against assigning specific meanings to emojis and said that they anticipated audience responses when choosing emojis.

> *... For the most part, there's no real universal meaning for most emojis. Like, I'd say there's a few that, like, most people always use to signify a few things, but for the most part, they all mean different things for different people in different scenarios.*
>
> *You have to think about and interpret how another person is going to receive that emoji. And so, when you are with your friends, or are texting your friends and you're using the emojis like ironically, you know that they're going to interpret those emojis ironically. But like with adults, I feel like, um, especially with adolescents, we, um, you know—that they're not going to interpret them the same way we do, so we take it more literally.*

Many participants stressed differences between adult and adolescent styles of emoji use, usually emphasizing how these groups differ in their preferences for sincerity or authenticity. Specifically, participants said they were more likely to use sarcasm or humor when sending emojis to peers, while they expected adults to interpret any emoji as a sincere expression of sentiment.

> *…Speaking of like older generations who use, who use like messaging or whatever, I like use the emojis in a different way than I do with my friends. Like with older people, I'll just use them in like an actual way of what they're like theoretically supposed to mean. Like a smiley face will actually mean a smiley face…*

Participants also said that certain emojis held meanings within their social groups that they knew were not shared by other groups. Though this was sometimes described as extending to entire geographic regions, these meanings were most often described as specific to groups of friends:

> *Almost everyone has different emojis that mean something to them with, like, another friend, like, an inside joke or, like, another meaning.*

Discussion

Principal Findings

This focus group study explored the ways adolescents use and interpret emojis, which has received less scholarly attention than adults' emoji use. By collecting data through focus groups and applying a coding strategy informed by Grounded Theory

XSL•FO
RenderX

[35], this work was the first to our knowledge to examine adolescent emoji use through adolescents' own words. The 4 themes that emerged in our analysis suggest that adolescent emoji use has many similarities with adult use but that it differs in some important ways. First, adolescents described the phatic functions of emojis that were informed by a sense of humor or absurdity (theme 1). Second, adolescents confirmed the emotive function of emojis, which they described as sometimes being achieved through a layer of sarcasm or irony (themes 2 and 3). Finally, they described emojis as having highly variable semantic meanings that are often understood differently by adolescents and adults (theme 4). Taken together, these findings suggest adolescents and adults use emojis for similar reasons (ie, phatic and emotive uses), but that adolescent emoji use is informed by humor, irony, and social dynamics that may make these messages difficult for adults to interpret.

Adolescents' Phatic Emoji Use is Humorous and Absurd

Focus group participants described emoji uses that seem to take advantage of emojis' phatic functions (ie, to maintain connection without providing any specific semantic meaning), but their use of this function was distinct from phatic use by adults. For example, adolescents said they frequently sent emojis that were meant to have no semantic meaning at all (theme 1a) or sent strings of randomly selected "emoji spam," or single emojis with no meaning outside of their repeated use (theme 1b). These phatic uses are notably different from the phatic uses observed in adult messaging [10], which include opening a conversation with a simple smile ( ) or closing a conversation with a wave ( ). Still, they fulfill the core phatic function of emojis: shoring up social connections without contributing substantial content.

Adolescents' Emotive Emoji Use is Often Sarcastic

Adolescents also described patterns of emoji use consistent with the emotive functions of emojis, but these patterns were informed by a preference for humor and a complex understanding of emojis' semantic meanings. For example, adolescents said they preferred not to use emojis in text communication (theme 2a) and were particularly averse to using emojis to express negative emotions. This aligns adolescent use with some research on adult use, which has found that adults tend to associate emoji use with expressions of more positive emotions [21] and sometimes see emojis as inappropriate in negative emotional contexts [36]. Though adolescents said they did sometimes use emojis to communicate sincere emotions (theme 3), much of the content from our focus groups suggested that adolescents understand the emotive functions of emojis differently than adults. While adults often use emojis to reduce emotional ambiguity [33] or to complement the emotional valence of text messages [18], adolescents reported that they usually sent emojis as sarcastic or ironic expressions of emotions (theme 2b) and used positive emojis to communicate negative emotions (theme 2c). Adolescents did not describe this sarcastic use of emojis as a source of confusion in communication with peers, however—instead such use was widely understood to communicate emotions in a sardonic or self-effacing tone. Thus, our results suggest adolescents also use emojis to reduce emotional ambiguity, but that they do so with the assumption that others can discern between straightforward and sarcastic expressions.

Adolescents Believe Adults Use Emojis Differently

Our overall interpretation of these findings—that adolescents use emojis for similar reasons as adults but perceive their use as more nuanced—is further supported by the fact that adolescents understood their own emoji use as different from use by adults (theme 4). Importantly, our focus group participants understood adults as favoring literal interpretations of emojis, which they often find risible (theme 2b). Participants said they anticipate these differences when communicating with adults, often suggesting adults would misunderstand the emojis they shared in peer-to-peer communication. Importantly, participants never said they were confused by the emoji choices made by adults. This suggests that adolescents do not understand adult emoji use as its own complex style of communication but instead view it as less developed than their own use.

Thus, adolescents in our focus group believed their use of emojis in their peer-to-peer communication was at least partially inaccessible to their parents and other adults. This might suggest that adolescents use emojis to facilitate one of the core "tasks" of their development: establishing autonomy from parents [37]. Coconstruction theory suggests that adolescents use digital media to achieve developmental tasks [38], and past research suggests that adolescents might gravitate to text messaging and other forms of text-based communication because these platforms help them develop relationships with peers while establishing independence and autonomy [27]. The apparent complexity of adolescent emoji use might serve a similar purpose—just as adolescents have long used slang, fashion, and other cultural codes to distinguish themselves from adults in physical spaces, they may use novel styles of emoji use to assert autonomy within text-based digital spaces.

Limitations

Study findings are limited in 3 important ways. First, while our purposeful sampling procedure recruited a diverse sample of focus group participants, it is likely that important subpopulations were not included in our sample. Thus, it may be that adolescents of certain demographic groups or from certain regions use emojis differently than the participants in this study. Second, the insights of this study depend on self-reflection by focus group participants who may not have been aware of or may have chosen not to disclose certain important aspects of their emoji use. Observation of real-world adolescent emoji use might reveal different patterns. Third, this study's comparisons between adolescent and adult emoji use were seen strictly from the perspective of adolescents because this study did not assess adults' motivations for emoji use. Future work could make stronger comparisons by assessing the emoji use of both adolescents and adults.

Practical Implications

These findings support the assertion [10] that emojis largely fill both emotive and phatic functions in interpersonal communication and suggest that adolescents see their emoji choices as informed by a complex set of shared norms they believe are inaccessible to adults. In addition to their

XSL•FO
RenderX

contribution to the theoretical understanding of adolescent emoji use, these findings also have implications for parents, who increasingly interact with adolescents through text-based platforms [3] and sometimes observe adolescents' text-based communications with peers [6]. These results suggest that parents should exercise caution when interpreting emojis sent by adolescents. While adolescents did say they often adjust their emoji use when communicating with adults, this may not be the case when other adolescents are also involved in the conversation (such as in a group chat), or if adolescents do not have the energy or motivation to adjust their habits for clarity of communication. These results also suggest that adults are likely to misinterpret the emojis used in adolescent-to-adolescent communication, which is an important consideration for parents, school administrators, or practitioners who may find themselves reading such conversations.

Acknowledgments

The authors confirm that generative artificial intelligence tools were not used at any point during the conceptualization or implementation of this study or the writing of this manuscript.

Data Availability

The datasets and generated analyzed during this study are not currently publicly available because they are currently in use in other ongoing projects. Updates and additional information are available from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

References

1. Anderson M, Jiang J. Teens, social media & technology. Analysis & Policy Observatory. 2018. URL: https://apo.org.au/node/175701 [accessed 2025-04-21]
2. Lenhart A. Teens, social media & technology overview. Pew Research Center. 2015. URL: https://www.pewresearch.org/internet/2015/04/09/teens-social-media-technology-2015/ [accessed 2025-04-21]
3. Jensen M, George MJ, Russell MA, Lippold MA, Odgers CL. Daily parent-adolescent digital exchanges. Res Child Adolesc Psychopathol 2021 Sep;49(9):1125-1138. [doi: 10.1007/s10802-020-00765-x] [Medline: 33821369]
4. Ersahin Z, Hanley T. Using text-based synchronous chat to offer therapeutic support to students: a systematic review of the research literature. Health Educ J 2017 Aug;76(5):531-543. [doi: 10.1177/0017896917704675]
5. Gibson K, Cartwright C. Young people's experiences of mobile phone text counselling: balancing connection and control. Child Youth Serv Rev 2014 Aug;43:96-104. [doi: 10.1016/j.childyouth.2014.05.010]
6. Lenhart A, Ling R, Campbell S, Purcell K. Teens and mobile phones: text messaging explodes as teens embrace it as the centerpiece of their communication strategies with friends. ERIC - Institute of Education Sciences. 2010. URL: https://eric.ed.gov/?id=ED525059 [accessed 2025-04-21]
7. Archer D, Akert RM. Words and everything else: verbal and nonverbal cues in social interpretation. J Pers Soc Psychol 1977;35(6):443-449. [doi: 10.1037//0022-3514.35.6.443]
8. Bai Q, Dan Q, Mu Z, Yang M. A systematic review of emoji: current research and future perspectives. Front Psychol 2019 Sep 15. [doi: 10.3389/fpsyg.2019.02221]
9. Pfeifer VA, Armstrong EL, Lai VT. Do all facial emojis communicate emotion? The impact of facial emojis on perceived sender emotion and text processing. Comput Human Behav 2022 Jan;126:107016. [doi: 10.1016/j.chb.2021.107016]
10. Danesi M. The Semiotics of Emoji: The Rise of Visual Language in the Age of the Internet: Bloomsbury Publishing; 2017. URL: https://www.bloomsbury.com/us/semiotics-of-emoji-9781474282000/ [accessed 2025-04-24]
11. Rodrigues D, Prada M, Gaspar R, Garrido MV, Emoji LDL, Database E. Norms for emoji and emoticons in seven evaluative dimensions. Behav Res Methods 2018 Feb;50(1):392-405. [doi: 10.3758/s13428-017-0878-6]
12. Kaneko D, Toet A, Ushiama S, Brouwer AM, Kallen V, van Erp JBF. EmojiGrid: a 2D pictorial scale for cross-cultural emotion assessment of negatively and positively valenced food. Food Res Int 2019 Jan;115:541-551. [doi: 10.1016/j.foodres.2018.09.049] [Medline: 30599977]
13. Herring SC, Dainas AR. Gender and age influences on interpretation of emoji functions. Trans Soc Comput 2020 Jun 30;3(2):1-26. [doi: 10.1145/3375629]
14. Weiß M, Bille D, Rodrigues J, Hewig J. Age-related differences in emoji evaluation. Exp Aging Res 2020;46(5):416-432. [doi: 10.1080/0361073X.2020.1790087] [Medline: 32662319]
15. Emoji statistics. Emojipedia. URL: https://emojipedia.org/stats [accessed 2025-04-21]
16. Dainas AR, Herring S. C. Interpreting emoji pragmatics. In: Approaches to Internet Pragmatics: Theory and Practice: John Benjamins Publishing Company; 2021:107-144. [doi: 10.1075/pbns.318.04dai]
17. Hu T, Guo H, Sun H, Nguyen TV, Luo J. Spice up your chat: the intentions and sentiment effects of using emojis. ICWSM 2017;11(1):102-111. [doi: 10.1609/icwsm.v11i1.14869]

XSL•FO
RenderX

18. Derks D, Bos AER, von Grumbkow J. Emoticons in computer-mediated communication: social motives and social context. Cyberpsychol Behav 2008 Feb;11(1):99-101. [doi: 10.1089/cpb.2007.9926] [Medline: 18275321]
19. Thompson D, Filik R. Sarcasm in written communication: emoticons are efficient markers of intention. J Comput-Mediat Comm 2016 Mar;21(2):105-120. [doi: 10.1111/jcc4.12156]
20. Boutet I, Goulet-Pelletier JC, Sutera E, Meinhardt-Injac B. Are older adults adapting to new forms of communication? A study on emoji adoption across the adult lifespan. Comput Hum Behav Rep 2024 Mar;13:100379. [doi: 10.1016/j.chbr.2024.100379]
21. Riordan MA. Emojis as tools for emotion work: communicating affect in text messages. J Lang Soc Psychol 2017 Oct;36(5):549-567. [doi: 10.1177/0261927X17704238]
22. Kelly R, Watts L. Characterising the inventive appropriation of emoji as relationally meaningful in mediated close personal relationships. Presented at: Experiences of Technology Appropriation: Unanticipated Users, Usage, Circumstances and Design; Sep 20, 2015; Oslo, Norway.
23. Miller H, Kluver D, Thebault-Spieker J, Terveen L, Hecht B. Understanding emoji ambiguity in context: the role of text in emoji-related miscommunication. ICWSM 2017 May 3;11(1):152-161. [doi: 10.1609/icwsm.v11i1.14901]
24. Völkel ST, Buschek D, Pranjic J, Hussmann H. Understanding emoji interpretation through user personality and message context. Presented at: MobileHCI '19; Oct 1-4, 2019; Taipei, Taiwan. [doi: 10.1145/3338286.3340114]
25. An J, Li T, Teng Y, Zhang P. Factors influencing emoji usage in smartphone mediated communications. Presented at: Transforming Digital Worlds; Mar 25-28, 2018; Sheffield, UK. [doi: 10.1007/978-3-319-78105-1_46]
26. Cui J, Colston HL, Jiang G. Is that a genuine smile? Emoji-based sarcasm interpretation across the lifespan. Metaphor Symb 2024 Jul 2;39(3):195-216. [doi: 10.1080/10926488.2024.2314595]
27. Ehrenreich SE, George MJ, Burnell K, Underwood MK. Importance of digital communication in adolescents' development: theoretical and empirical advancements in the last decade. J Res Adolesc 2021 Dec;31(4):928-943. [doi: 10.1111/jora.12643] [Medline: 34820954]
28. Ling R, Bertel T, Sundsøy P. Texting among same-aged individuals: an analysis of traffic data. Presented at: International Communication Association Pre-Conference on Mobile Communication; Jun 21, 2010; Singapore.
29. Dhesi M, Sefi A, Donati M, Hayes J, Cooper M. Helpful and unhelpful elements of synchronous text‐based therapy: a thematic analysis. Couns Psychother Res 2022 Mar;22(1):157-165. [doi: 10.1002/capr.12414]
30. McLaren RM, Pederson JR. Relational communication and understanding in conversations about hurtful events between parents and adolescents. J Commun 2014 Feb;64(1):145-166. [doi: 10.1111/jcom.12072]
31. Sillars A, Smith T, Koerner A. Misattributions contributing to empathic (in)accuracy during parent-adolescent conflict discussions. J Soc Pers Relat 2010 Sep;27(6):727-747. [doi: 10.1177/0265407510373261]
32. Ickes W. Measuring empathic accuracy. In: Interpersonal Sensitivity: Lawrence Erlbaum Associates Publishers; 2001:219-241.
33. Kaye LK, Wall HJ, Malone SA. "Turn that frown upside-down": a contextual account of emoticon usage on different virtual platforms. Comput Human Behav 2016 Jul;60:463-467. [doi: 10.1016/j.chb.2016.02.088]
34. Glaser BG. The constant comparative method of qualitative analysis. Soc Probl 1965 Apr;12(4):436-445. [doi: 10.2307/798843]
35. Glaser BG, Strauss AL, Strutzel E. The discovery of grounded theory; strategies for qualitative research. Nurs Res 1968 Jul;17(4):364. [doi: 10.1097/00006199-196807000-00014]
36. Rodrigues D, Lopes D, Prada M, Thompson D, Garrido MV. A frown emoji can be worth a thousand words: perceptions of emoji use in text messages exchanged between romantic partners. Telemat Inform 2017 Dec;34(8):1532-1543. [doi: 10.1016/j.tele.2017.07.001]
37. Luyckx K, Goossens L, Soenens B, Beyers W. Unpacking commitment and exploration: preliminary validation of an integrative model of late adolescent identity formation. J Adolesc 2006 Jun;29(3):361-378. [doi: 10.1016/j.adolescence.2005.03.008] [Medline: 15878619]
38. Subrahmanyam K, Smahel D, Greenfield P. Connecting developmental constructions to the internet: identity presentation and sexual exploration in online teen chat rooms. Dev Psychol 2006 May;42(3):395-406. [doi: 10.1037/0012-1649.42.3.395] [Medline: 16756432]

Edited by A Mavragani; submitted 17.04.24; peer-reviewed by IA Lane, J Cui; revised version received 05.03.25; accepted 05.03.25; published 28.04.25.

Please cite as:
Minich M, Kerr B, Moreno M
Adolescent Emoji Use in Text-Based Messaging: Focus Group Study
JMIR Form Res 2025;9:e59640
URL: https://formative.jmir.org/2025/1/e59640
doi:10.2196/59640

XSL•FO
RenderX

© Matt Minich, Bradley Kerr, Megan Moreno. Originally published in JMIR Formative Research (https://formative.jmir.org), 28.4.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Impact of the Kidney Score Platform on Communication About and Patients' Engagement With Chronic Kidney Disease Health: Pre–Post Intervention Study

Delphine Tuot[1], MAS, MDCM; Susan Crowley[2,3], MD, MBA; Lois Katz[4], MD; Joseph Leung[4], MD; Delly Alcantara-Cadillo[2,3], MD; Christopher Ruser[2,3], MD; Elizabeth Talbot-Montgomery[5], BS; Joseph Vassalotti[5,6], MD

[1]Division of Nephrology, University of California, San Francisco, Zuckerberg San Francisco General Hospital, 1001 Potrero Ave, Bldg 100, Room 342, San Francisco, CA, United States

[2]Veterans Administration Connecticut Healthcare System, New Haven, CT, United States

[3]Department of Medicine, Yale University, West Haven, CT, United States

[4]Veterans Administration NY Harbor Healthcare System, New York, NY, United States

[5]National Kidney Foundation, New York, NY, United States

[6]Department of Medicine, Icahn School of Medicine, Mt Sinai, New York, NY, United States

Corresponding Author:
Delphine Tuot, MAS, MDCM
Division of Nephrology, University of California, San Francisco, Zuckerberg San Francisco General Hospital, 1001 Potrero Ave, Bldg 100, Room 342, San Francisco, CA, United States

Abstract

Background: Chronic kidney disease (CKD) affects 14% of the US adult population, yet patient knowledge about kidney disease and engagement in their kidney health is low despite many CKD education programs, awareness campaigns, and clinical practice guidelines.

Objective: We aimed to examine the impact of the Kidney Score Platform (a patient-facing, risk-based online tool that provides interactive health information tailored to an individual's CKD risk plus an accompanying clinician-facing Clinical Practice Toolkit) on individual engagement with CKD health and CKD communication between clinicians and patients.

Methods: We conducted a pre-post intervention study in which English-speaking veterans at risk for CKD in two primary care settings interacted with the Kidney Score platform's educational modules and their primary care clinicians were encouraged to review the Clinical Practice Toolkit. The impact of the Kidney Score on the Patient Activation Measure (the primary outcome), knowledge about CKD, and communication with their clinician about kidney health was determined with paired *t* tests. Multivariable linear and logistic models were used to determine whether changes in outcomes after versus before intervention were influenced by age, race or ethnicity, sex, and diabetes status, accounting for baseline values.

Results: The study population (n=76) had a mean (SD) age of 64.4 (8.2) years, 88% (67/76) was male, and 30.3% (23/76) self-identified as African-American. Approximately 93% (71/76) had hypertension, 36% (27/76) had diabetes, and 9.2% (7/76) had CKD according to the laboratory criteria but without an *ICD-10* (International Classification of Diseases, 10th Edition) diagnosis. Patient interaction with the Kidney Score did not change the mean Patient Activation Measure (preintervention: 40.7%, postintervention: 40.2%, P=.23) but increased the mean CKD knowledge score (preintervention: 40.0%, postintervention 51.1%, P<.01), and changed the percentage of veterans who discussed CKD with their clinician (preintervention: 12.3%, postintervention: 31.5%, P<.01). Changes did not differ by age, sex, race, or diabetes status. Results were limited by the small sample size due to low recruitment and minimal clinician engagement with the Clinical Practice Toolkit during the COVID-19 pandemic.

Conclusions: One-time web-based tailored education for patients can increase CKD knowledge and encourage conversations about kidney health. Increasing patient activation for CKD management may require multilevel, longitudinal interventions that facilitate ongoing conversations about kidney health between patients and clinician teams.

(JMIR Form Res 2025;9:e56855) doi:10.2196/56855

KEYWORDS

chronic kidney disease; CKD; CKD communication; CKD knowledge; patient activation; kidney; kidney score platform; kidney health; United States; US; adult; aging; clinical practice; awareness campaign; health information; clinician; primary care; longitudinal intervention; web-based; mobile health; mHealth

XSL•FO
RenderX

Introduction

Chronic kidney disease (CKD) affects 37 million Americans [1] and is associated with high risks of emergency department visits, hospitalizations, cardiovascular events, and early mortality [2-5]. Yet, as many as half of all individuals with laboratory manifestations of kidney disease [6] and those at the highest risk of CKD progression to kidney failure [7] are unaware that they have kidney disease. CKD is usually asymptomatic. Individuals cannot readily know their disease status or risk for disease without risk recognition, testing, detection, and communication by clinicians [8]. A clinician diagnosis of CKD has been associated with increased delivery of evidence-based care, as well as increased patient awareness of their kidney disease [9-11]. However, clinician detection of CKD and communication about kidney disease in the United States are suboptimal [12,13].

Interventions to enhance patient CKD knowledge, increase patient engagement in kidney health, and promote effective communication about CKD between patients and clinicians are critical to facilitate the identification of high-risk populations that would benefit the most from aggressive management of CKD and nephrology referral. They are also important for individuals at a lower risk of CKD progression, to ensure accurate drug dosing, avoid nephrotoxic medications, and encourage the optimal management of diabetes and hypertension, the two most common etiologies of kidney failure in the United States [14].

Existing education programs, awareness campaigns, and clinical practice guidelines have minimally improved CKD awareness in the US population [15,16]. One reason may be because interventions have targeted either the patient or the provider, but rarely both at the same time. To bridge this gap, the National Kidney Foundation developed the Kidney Score Platform, which includes a patient-facing, risk-based online tool that provides interactive health information tailored to an individual's CKD risk as well as an accompanying clinician-facing Clinical Practice Toolkit [17]. The patient-facing elements aim to increase individual awareness of CKD and encourage patients with and at risk for CKD to initiate discussions about kidney disease with their clinicians. The Clinical Practice toolkit assists clinicians in discussing CKD with individuals with and at risk for kidney disease. In this manuscript, we describe the impact of the Kidney Score Platform on individual engagement with their CKD health and CKD communication between clinicians and patients.

Methods

Study Design, Population, and Settings

This was a pre-post–intervention study that examined the impact of the Kidney Score on CKD knowledge and participation in CKD self-management. The Kidney Score is a web-based educational platform designed to improve awareness and understanding of kidney disease among individuals at risk and living with CKD. The study took place in the primary care settings of the Veterans Affairs Medical Centers (VA New York Harbor Healthcare System and VA Connecticut Healthcare System) between September 2022 and March 2023. Eligible participants included English-speaking veterans between the ages of 18 and 75 years with diabetes or hypertension defined in the electronic health record (EHR). The exclusion criteria included an estimated glomerular filtration rate (eGFR) <15 ml/min/1.73 m^2 or individuals with end-stage kidney disease or kidney transplant recipients, as the Kidney Score is not geared towards individuals with severe CKD. Veterans enrolled in hospice services, those with vision impairment, and those with severe dementia identified in the EHR or during the consent process were also excluded from the study, as CKD awareness is much less important for this population's overall health.

Ethical Considerations

Eligible veterans with an upcoming primary care visit at one of two participating Veterans Affairs Medical Centers (VA New York Harbor Healthcare System and VA Connecticut Healthcare System) were identified by the study team, leveraging the EHR. These veterans were invited to participate in the study by mail. Individuals who did not opt out were contacted by phone; those that provided telephone assent were mailed/emailed an information sheet describing the program. During a subsequent phone call, verbal informed consent was obtained and participants were asked to complete an online preintervention survey administered by the study team. Study participants were given a unique study ID; all collected data were deidentified. Participants were given a US $30 gift card for participating. The study was approved by the Institutional Review Boards of the VA Connecticut Healthcare System (#02290) and VA NY Harbor Healthcare System (#01705).

Study Processes and Intervention

The Kidney Score is a free, online website that uses a rule engine and risk predictive analytics to provide interactive health information tailored to an individual's CKD risk and health status. In an anonymous fashion, individuals enter risk factor information as well as laboratory findings in the Kidney Score Platform's online interface and receive educational programming tailored to their clinical status and risk for CKD development or progression. Development of the Kidney Score has been previously described in detail and leveraged the Behavioral Change Wheel framework [18]. It was subsequently found to be acceptable to patients with and at risk for kidney disease [17].

After providing informed consent and baseline data, study participants were invited by the study team to engage with the Kidney Score Platform, which included educational modules tailored to their CKD risk as determined by the Kidney Score's rule engine. Veterans were given the option to print out questions to ask their providers about their kidney health and risk for kidney disease. Participants could access the tool in any location of their preference (ie, home, waiting room, or community center). Within a week after completing their subsequent primary care appointment, participants were asked to complete a postintervention closed survey electronically to assess the impact of the Kidney Score on their understanding of CKD and the quality of their conversation with their clinician about kidney health.

XSL•FO
RenderX

The pre- and postintervention surveys consisted of 4 instruments and 1 additional question. The question order was fixed for all the study participants; the surveys spanned 5 screens and did not include any adaptive components. Survey responses were all voluntary; there was no "completeness check" and participants were allowed to "go back" to change survey responses prior to submitting the final data.

Outcomes

The primary outcome was the change in self-efficacy for CKD management, ascertained by the Patient Activation Measure [19] after the participant engaged with the Kidney Score and completed a visit with their primary care clinician compared to preintervention. Secondary outcomes included self-reported changes in CKD knowledge [20], perceived risk of kidney disease [21], and communication quality with providers [22], all ascertained by validated instruments. The change in the percentage of individuals who discussed kidney disease with their provider was also assessed using the question "At your last visit with your primary care clinician, did you discuss your risk for chronic kidney disease or testing for kidney disease?"

Measurements

Patient demographic data (ie, age, sex, and race or ethnicity) were self-reported by participants. Prevalent comorbid conditions (diabetes, hypertension, and CKD) and the baseline eGFR were ascertained from the EHR.

Statistical Analysis

Power calculations performed before study activities suggested that 103 individuals would need to participate in the study to detect a clinically meaningful change in the Patient Activation Measure (the primary outcome). Due to challenges with study operations with the transition towards virtual care delivery during the COVID-19 pandemic, the study was stopped after the full participation of 76 individuals. We examined the characteristics of these study participants using descriptive statistics. Changes in the Patient Activation Measure, communication quality with providers, and kidney knowledge were determined with paired *t* tests. The McNemar test for binary paired data was used to ascertain changes in the perceived risk of kidney disease and the concern for developing kidney disease. Multivariable linear models (for continuous outcomes) and logistic models (for dichotomous outcomes) were used to determine whether changes in outcomes post versus preintervention were influenced by age, race or ethnicity, sex, and diabetes status, after accounting for baseline values. Statistical analyses were performed using Stata/SE, version 14.0 (StataCorp).

Results

Recruitment letters were sent to 1115 eligible veterans, of whom 87 (7.8%) opted out of receiving further communication. The study team was able to contact 206 veterans by phone, of whom 185 agreed to participate (Figure 1). Reasons for declining participation included being uncomfortable with online technology, mistrust of the study team, mistrust in the Veterans Affairs treatment center, not being interested in kidney disease, and not having the time to participate in a study. Approximately 41% (76/185) of the recruited veterans completed all aspects of the study, including completing the previsit survey, reviewing the web-based Kidney Score Platform, attending a scheduled visit with their primary care clinician, and completing the postvisit survey. The study population had a mean (SD) age of 64.4 (8.2) years and 88% (67/76) were male. Nearly one-third (23/76) of the study participants self-identified as African-American and 5% (4/76) self-identified as Native Hawaiian, Pacific Islander, or Asian. Approximately 93% (71/76) had hypertension and 36% (27/76) had diabetes; 9.2% (7/76) of the population had CKD defined by an eGFR 15 - 60 ml/min/1.73 m^2 but none had a diagnosis of CKD in their clinical record (Table 1). Characteristics of the study population were qualitatively similar to those describing the overall eligible primary care population in the participating clinics (Multimedia Appendix 1) .

XSL•FO
RenderX

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) diagram for this implementation study assessing the impact of the kidney score on chronic kidney disease knowledge and participation in self-management.

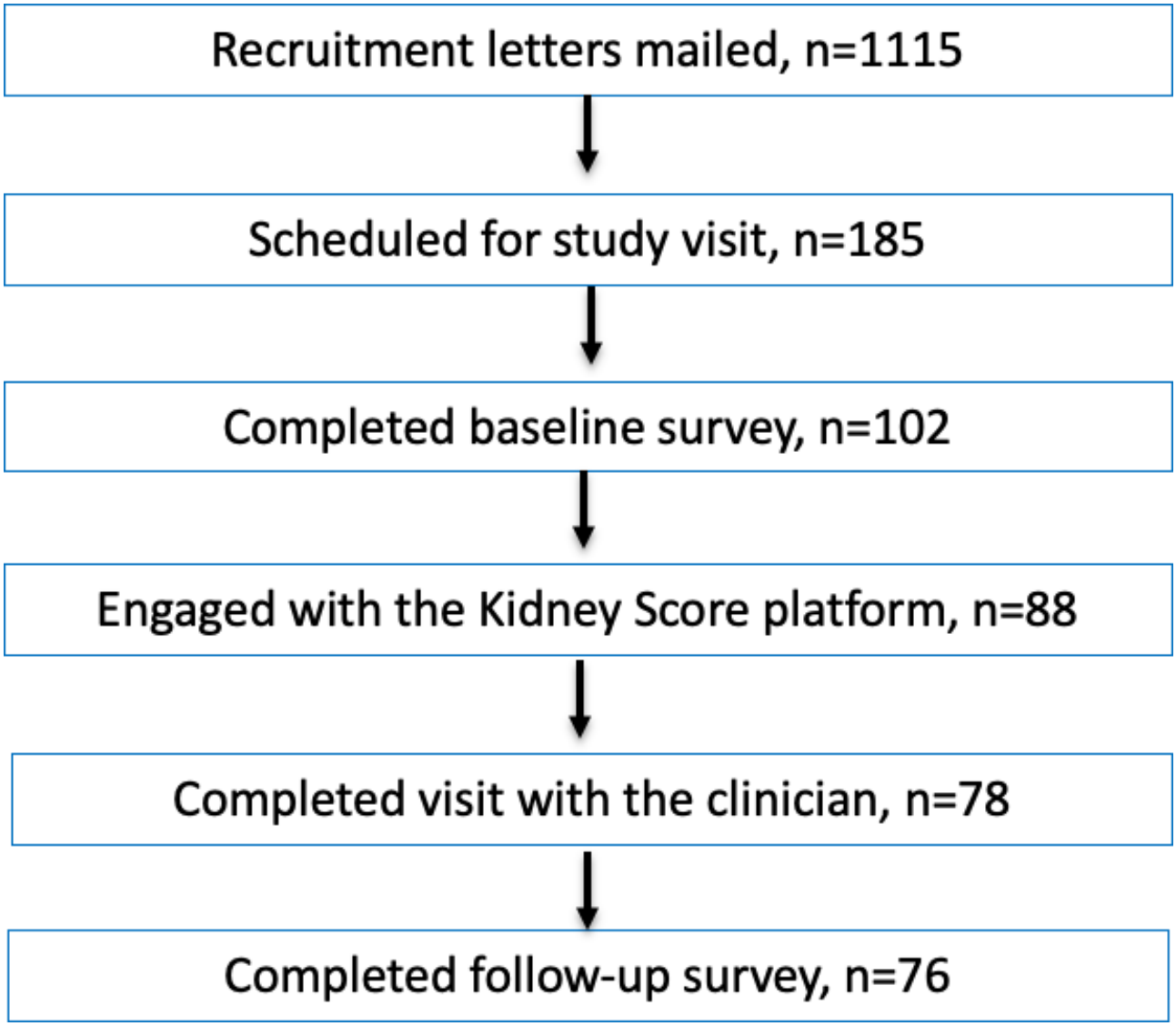

XSL•FO
RenderX

Table . Characteristics of veterans with and at risk for chronic kidney disease (CKD) who participated in this study that examined the impact of the kidney score on CKD knowledge and participation in CKD self-management.

Characteristics	N=76
Age, years, mean (SD)	64.4 (8.2)
Male sex, n (%)	67 (88.2)
Race, n (%)	
American Indian or Alaskan Native	1 (1.3)
Asian	2 (2.6)
Black	23 (30.3)
Native Hawaiian or Pacific Islander	2 (2.63)
Declined to answer	5 (6.6)
White	43 (56.6)
Ethnicity, n (%)	
Hispanic or Latino	4 (5.3)
Not Hispanic or Latino	71 (93.4)
Declined to answer	1 (1.3)
Hypertension,[a] n (%)	71 (93.4)
Diabetes,[a] n (%)	27 (35.5)
Chronic kidney disease[b]	7 (9.2)

[a]Hypertension and diabetes are defined by the *ICD-10* (International Classification of Diseases, 10th Edition) code.

[b]Chronic kidney disease is defined by an estimated glomerular filtration rate of 15 - 60 ml/min/1.73 m^2.

Among veterans with complete data, the mean (SD) Patient Activation Measure score at baseline was 40.7% (5.32%). After interacting with the Kidney Score Platform and attending a visit with their provider, the mean Patient Activation Measure score was 40.2% (5.48%), which was largely unchanged (*P*=.23). Similarly, the self-reported communication quality with providers did not change after the interaction with the Kidney Score Platform, from a mean (SD) baseline score of 2.77 (1.06) out of 3.00 to a mean score of 2.84 (0.94; *P*=.51). On multivariable regression models, age, sex, race or ethnicity, and diabetes status were not associated with changes in the Patient Activation Measure scores nor the communication quality.

The mean (SD) CKD knowledge score increased after the intervention from a baseline of 40% (25) to 51.1% (26.5), representing a statistically significant difference (*P*<.01). Similarly, the percentage of veterans who discussed CKD with their clinician in their prior visit increased from 12.3% at baseline to 31.5% after interacting with the Kidney Score Platform (*P*<.01). In multivariable regression models, change in CKD awareness and communication about CKD did not differ by age, sex, race, or diabetes status (Figure 2).

Consistent with the increased communication about CKD in a low-risk primary care population, the proportion of individuals with substantial concern for developing kidney problems or kidney disease in the next 10 years did not change significantly after participating in the study. At baseline, an estimated 52.1% of individuals reported feeling moderately or very concerned for developing CKD; after the study, this decreased to an estimated 42.3% (*P*=.06). Similarly, the estimated percentage of individuals who reported a moderate or high likelihood of developing kidney disease in the next 10 years did not change substantially before versus after the intervention (29.6% vs 31.0%; *P*=.7). Age, sex, race, and diabetes were not associated with greater concern or perceived risk of developing kidney disease.

XSL•FO
RenderX

Figure 2. Change in individual behaviors potentially associated with self-efficacy for chronic kidney disease (CKD) management after engagement with the kidney scope online platform.

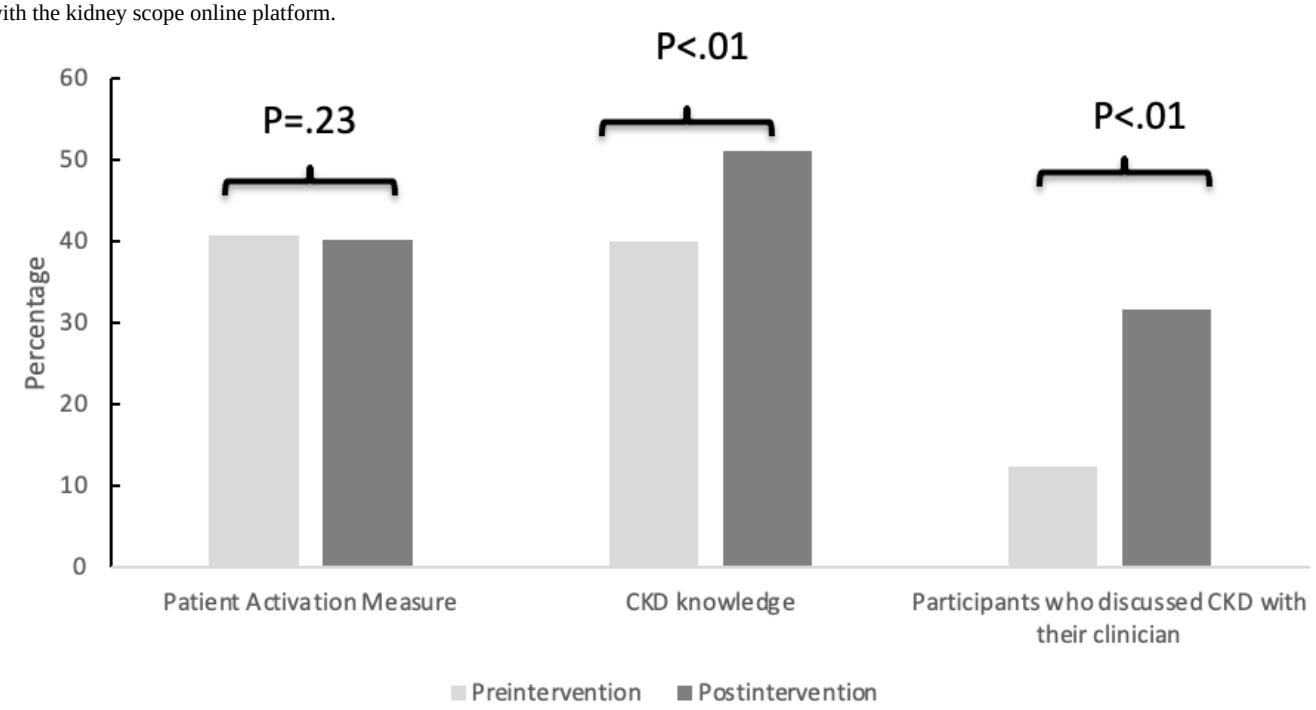

Discussion

Principal Findings and Comparison With Previous Work

The Kidney Score Platform is one of the first interventions that has successfully increased patient knowledge about CKD and communication about kidney health between primary care clinicians and patients. This success can be attributed to its development based on the Behavioral Change Wheel [23], a well-recognized individual behavioral change theory. Theory-informed interventions are more likely to be effective and sustainable within the context of health care delivery than those that are not. The development of the Kidney Score Platform was based on extensive formative work, including qualitative interviews with clinicians and patients [18], as well as proto-typing, beta-testing, and usability testing with various stakeholders, culminating in a refined educational patient experience [17]. This process identified the importance of a tool with which both patients and clinicians could interact and strengthen their communication skills around kidney disease and kidney health.

Despite an increase in CKD conversations and CKD knowledge, the Kidney Score Platform did not increase patient activation for CKD management, which is a hallmark of self-efficacy. The baseline mean (SD) Patient Activation Measure score was 40.1% (5.3%), representing level 1 activation, characterized by individuals who may not believe that the patient's role is important in chronic disease management. This reflects a large opportunity for improvement, rendering this study's null primary outcome results disappointing. We speculate that the lack of change in the Patient Activation Measure score can be attributed in part to the suboptimal sample size, as well as the characteristics of the study population and the intervention itself. Veterans who participated in this study were primarily at risk for CKD (69/76) with only 9.2% (7/76) having laboratory evidence of CKD considering the eGFR. Engaging with the Kidney Score may be more motivating among those with CKD versus those at risk for CKD. Similarly, clinician discussions about CKD and how individuals can participate in self-management are likely more impactful for individuals with laboratory evidence of kidney disease compared to those at risk. The absence of a change in the perceived risk of kidney disease among study participants supports this lack of activation among those at risk for CKD. While we did not have the statistical power to assess changes in patient activation by CKD status, we suspect that a study population enriched with veterans with CKD (as opposed to those at risk for CKD) would have led to a greater change in patient activation.

The nature of the intervention may have also contributed to a lack of change in patient activation. While a handful of patient interactions with the Kidney Score Platform and one patient-initiated discussion with a clinician during a routine primary care visit may be sufficient to increase the broad understanding of kidney disease among individuals with or at risk for CKD, it is likely insufficient to motivate the need for behavior change to optimize kidney health. More than one conversation about kidney disease is likely needed to enhance participation in healthy behaviors and promote self-confidence in other aspects of CKD self-management. With that in mind, it is possible that patient activation was remeasured too early after the time-limited intervention. While speculative, it is plausible that an initial interaction with the Kidney Score Platform may have prompted future discussions about kidney disease and thus future changes in patient activation. Future interventions leveraging the Kidney Score Platform or other patient-facing educational modules will need to include multiple opportunities for discussions about kidney health over time, while also being mindful to not exacerbate potential health anxiety associated with greater knowledge about kidney disease.

XSL•FO
RenderX

This is consistent with Kolb's theory of experiential learning, which posits that most adult learning is gained through experience and engagement over time [24].

The application of Kolb's theory has been demonstrated for the CKD population with moderate to severe CKD depicted by an eGFR 15 - 60 ml/min/1.73 m^2. For example, Medicare's Kidney Disease Education program is designed to be delivered in 6 sessions for those with an eGFR 15 - 30 ml/min/1.73 m^2 and includes modules about knowledge about risk factor control, how the kidneys function, medication management, nutrition, and treatments for kidney failure [25]. Two recent studies demonstrated improvements in access to home dialysis and improved preparation for hemodialysis for the population that received kidney disease education versus eligible controls, although the education delivery was low, among approximately 1% (3469/369,968) of eligible patients [26,27]. A systematic review of educational interventions for patients with CKD with an eGFR range similar to our study participants found that more frequent interventions, (ie, weekly or monthly) were more effective [28]. The elements of the Kidney Score such as the website interaction could easily be scaled for more frequent interventions in future educational investigations.

Additionally, while this study was originally designed as a multilevel intervention directed at patients and clinicians, challenges during the COVID-19 pandemic mitigated the clinician focus of this study. Clinician input refined the Kidney Score Platform, but engagement with the Clinical Practice Toolkit was minimal due to clinician burn out and an emphasis on COVID-19 during the study period. The intervention was ultimately patient-focused with little emphasis on changing clinician behaviors with respect to discussing CKD with individuals with and at risk for kidney disease. Given the strong consensus that multilevel interventions are needed to address health communication [29], future research is needed on what types of tools can be integrated into clinician workflows without exacerbating burn out.

Strengths and Limitations

The strengths of this investigation include theory-informed patient-focused education to improve multiple domains of patient knowledge and engagement assessed using validated instruments. The limitations include low recruitment that may have biased results to the null, necessary study design refinement during the COVID-19 pandemic, absence of a quantitative measure identifying the number of times a participant engaged with the Kidney Score Platform, and limited generalizability to nonveterans considering the administration of the health care delivery systems.

Conclusions

Educating, engaging, and empowering individuals at risk for and living with CKD are critical. The findings of this study suggest that web-based tailored education for patients can increase CKD knowledge and encourage conversations about kidney health. Increasing patient activation and self-efficacy for CKD management will require multilevel, longitudinal interventions that facilitate ongoing conversations about kidney health between patients and clinician teams.

Acknowledgments

Gary Bronz and Anne Cordato provided operational and administrative support. Erin Madden provided primary care clinic demographic information. This study was funded by the Veteran Administration Center for Innovation (VA118-17-R-2144).

Data Availability

Datasets generated or analyzed during this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Characteristics of the primary care population potentially eligible for the study, March 1, 2020 to February 28, 2022.
[DOCX File, 17 KB - formative_v9i1e56855_app1.docx]

References

1. Centers for Disease Control and Prevention. Chronic kidney disease in the united states. : US Department of Health and Human Services, Centers for Disease Control and Prevention; 2019 URL: https://www.cdc.gov/kidney-disease/php/data-research/index.html [accessed 2020-02-26]
2. United States Renal Data System. 2018 USRDS annual data report: epidemiology of kidney disease in the united states. National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases. 2018 URL: https://www.ajkd.org/article/S0272-6386(19)30009-5/fulltext [accessed 2020-02-26]
3. Go AS, Chertow GM, Fan D, McCulloch CE, Hsu CY. Chronic kidney disease and the risks of death, cardiovascular events, and hospitalization. N Engl J Med 2004 Sep 23;351(13):1296-1305. [doi: 10.1056/NEJMoa041031] [Medline: 15385656]

XSL•FO
RenderX

4. Sarnak MJ, Greene T, Wang X, et al. The effect of a lower target blood pressure on the progression of kidney disease: long-term follow-up of the modification of diet in renal disease study. Ann Intern Med 2005 Mar 1;142(5):342-351. [doi: 10.7326/0003-4819-142-5-200503010-00009] [Medline: 15738453]
5. Inker LA, Astor BC, Fox CH, et al. KDOQI US commentary on the 2012 KDIGO clinical practice guideline for the evaluation and management of CKD. Am J Kidney Dis 2014 May;63(5):713-735. [doi: 10.1053/j.ajkd.2014.01.416] [Medline: 24647050]
6. Tuot DS, Plantinga LC, Hsu CY, et al. Chronic kidney disease awareness among individuals with clinical markers of kidney dysfunction. Clin J Am Soc Nephrol 2011 Aug;6(8):1838-1844. [doi: 10.2215/CJN.00730111] [Medline: 21784832]
7. Chu CD, McCulloch CE, Banerjee T, et al. CKD Awareness Among US adults by future risk of kidney failure. Am J Kidney Dis 2020 Aug;76(2):174-183. [doi: 10.1053/j.ajkd.2020.01.007] [Medline: 32305206]
8. Vassalotti JA, Centor R, Turner BJ, et al. Practical approach to detection and management of chronic kidney disease for the primary care clinician. Am J Med 2016 Feb;129(2):153-162. [doi: 10.1016/j.amjmed.2015.08.025] [Medline: 26391748]
9. Rothberg MB, Kehoe ED, Courtemanche AL, et al. Recognition and management of chronic kidney disease in an elderly ambulatory population. J Gen Intern Med 2008 Aug;23(8):1125-1130. [doi: 10.1007/s11606-008-0607-z] [Medline: 18443883]
10. Allen AS, Forman JP, Orav EJ, Bates DW, Denker BM, Sequist TD. Primary care management of chronic kidney disease. J Gen Intern Med 2011 Apr;26(4):386-392. [doi: 10.1007/s11606-010-1523-6] [Medline: 20922494]
11. Szczech LA, Stewart RC, Su HL, et al. Primary care detection of chronic kidney disease in adults with type-2 diabetes: the ADD-CKD Study (awareness, detection and drug therapy in type 2 diabetes and chronic kidney disease). PLoS ONE 2014;9(11):e110535. [doi: 10.1371/journal.pone.0110535] [Medline: 25427285]
12. Plantinga LC, Tuot DS, Powe NR. Awareness of chronic kidney disease among patients and providers. Adv Chronic Kidney Dis 2010 May;17(3):225-236. [doi: 10.1053/j.ackd.2010.03.002] [Medline: 20439091]
13. Greer RC, Cooper LA, Crews DC, Powe NR, Boulware LE. Quality of patient-physician discussions about CKD in primary care: a cross-sectional study. Am J Kidney Dis 2011 Apr;57(4):583-591. [doi: 10.1053/j.ajkd.2010.08.027] [Medline: 21131116]
14. United States Renal Data System. 2022 USRDS annual data report: epidemiology of kidney disease in the United States. In: National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases 2022.
15. Plantinga LC, Boulware LE, Coresh J, et al. Patient awareness of chronic kidney disease: trends and predictors. Arch Intern Med 2008 Nov 10;168(20):2268-2275. [doi: 10.1001/archinte.168.20.2268] [Medline: 19001205]
16. Management of Chronic Kidney Disease (CKD).: US Department of Veterans Affairs/Department of Defense URL: https://www.healthquality.va.gov/guidelines/CD/ckd [accessed 2020-02-26]
17. Tuot DS, Crowley ST, Katz LA, et al. Usability testing of the kidney score platform to enhance communication about kidney disease in primary care settings: qualitative think-aloud study. JMIR Form Res 2022 Sep 28;6(9):e40001. [doi: 10.2196/40001] [Medline: 36170008]
18. Tuot DS, Crowley ST, Katz LA, et al. The kidney score platform for patient and clinician awareness, communication, and management of kidney disease: protocol for a mixed methods study. JMIR Res Protoc 2020 Oct 19;9(10):e22024. [doi: 10.2196/22024] [Medline: 33074162]
19. Hibbard JH, Mahoney ER, Stockard J, Tusler M. Development and testing of a short form of the patient activation measure. Health Serv Res 2005 Dec;40(6 Pt 1):1918-1930. [doi: 10.1111/j.1475-6773.2005.00438.x] [Medline: 16336556]
20. Wright JA, Wallston KA, Elasy TA, Ikizler TA, Cavanaugh KL. Development and results of a kidney disease knowledge survey given to patients with CKD. Am J Kidney Dis 2011 Mar;57(3):387-395. [doi: 10.1053/j.ajkd.2010.09.018] [Medline: 21168943]
21. Boulware LE, Carson KA, Troll MU, Powe NR, Cooper LA. Perceived susceptibility to chronic kidney disease among high-risk patients seen in primary care practices. J Gen Intern Med 2009 Oct;24(10):1123-1129. [doi: 10.1007/s11606-009-1086-6] [Medline: 19711135]
22. Lorig K, Stewart A, Ritter P, Gonzalez V, Lauren D, Lynch J. Outcome Measures for Health Education and Other Health Care Interventions: Sage Publications; 1996. [doi: 10.4135/9781452232966]
23. Michie S, van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. Implement Sci 2011 Apr 23;6:42. [doi: 10.1186/1748-5908-6-42] [Medline: 21513547]
24. Kolb DA. Experiential Learning: Experience as the Source of Learning and Development, 2nd edition: Pearson Education, Inc; 2015.
25. Centers for Medicare & Medicaid Services (CMS) H. Kidney Disease Education. URL: https://www.medicare.gov/coverage/kidney-disease-education [accessed 2024-01-06]
26. Shukla AM, Bozorgmehri S, Ruchi R, et al. Utilization of CMS pre-ESRD Kidney Disease Education services and its associations with the home dialysis therapies. Perit Dial Int 2021 Sep;41(5):453-462. [doi: 10.1177/0896860820975586] [Medline: 33258420]
27. Ruchi R, Bozorgmehri S, Chamarthi G, et al. Provision of kidney disease education service is associated with improved vascular access outcomes among US incident hemodialysis patients. Kidney360 2022 Jan 27;3(1):91-98. [doi: 10.34067/KID.0004502021] [Medline: 35368570]

XSL•FO
RenderX

28. Lopez-Vargas PA, Tong A, Howell M, Craig JC. Educational interventions for patients with ckd: a systematic review. Am J Kidney Dis 2016 Sep;68(3):353-370. [doi: 10.1053/j.ajkd.2016.01.022] [Medline: 27020884]
29. Oh AY, Rising CJ, Gaysynsky A, et al. Advancing multi-level health communication research: A Delphi study on barriers and opportunities. Transl Behav Med 2022 Dec 30;12(12):1133-1145. [doi: 10.1093/tbm/ibac068] [Medline: 36378100]

Abbreviations

CKD: chronic kidney disease
eGFR: estimated glomerular filtration rate
EHR: electronic health record

Edited by A Mavragani; submitted 28.01.24; peer-reviewed by N Alqurini, T Markossian; revised version received 06.03.25; accepted 07.03.25; published 29.04.25.

Please cite as:
Tuot D, Crowley S, Katz L, Leung J, Alcantara-Cadillo D, Ruser C, Talbot-Montgomery E, Vassalotti J
Impact of the Kidney Score Platform on Communication About and Patients' Engagement With Chronic Kidney Disease Health: Pre–Post Intervention Study
JMIR Form Res 2025;9:e56855
URL: https://formative.jmir.org/2025/1/e56855
doi:10.2196/56855

© Delphine Tuot, Susan Crowley, Lois Katz, Joseph Leung, Delly Alcantara-Cadillo, Christopher Ruser, Elizabeth Talbot-Montgomery, Joseph Vassalotti. Originally published in JMIR Formative Research (https://formative.jmir.org), 29.4.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Effect of a Digital Health Exercise Program on the Intention for Spinal Surgery in Adult Spinal Deformity: Exploratory Cross-Sectional Survey

Marsalis Christian Brown[1*], MD; Christopher Quincy Lin[2*], MS; Christopher Jin[3], BS; Matthew Rohde[4*], BS; Brett Rocos[5], MD; Jonathan Belding[1], MD; Barrett I Woods[6*], MD; Stacey J Ackerman[7*], PhD

[1]Department of Orthopaedic Surgery, MetroHealth Medical Center, 2500 MetroHealth Dr, Cleveland, OH, United States
[2]Department of Medicine, The Larner College of Medicine, The University of Vermont, Burlington, VT, United States
[3]Department of Medical, Medical College of Wisconsin, Milwaukee, WI, United States
[4]The Department of Orthopaedics, Zucker School of Medicine at Hofstra/Northwell, Hempstead, NY, United States
[5]Duke Spine Center, Duke University Health System, Durham, NC, United States
[6]Department of Spine Surgery, Rothman Orthopaedic Institute, Thomas Jefferson University Rothman Institute, Philadelphia, PA, United States
[7]Department of Biomedical Engineering, Johns Hopkins University, San Diego, CA, United States
[*]these authors contributed equally

Corresponding Author:
Marsalis Christian Brown, MD
Department of Orthopaedic Surgery, MetroHealth Medical Center, 2500 MetroHealth Dr, Cleveland, OH, United States

Abstract

Background: Adult spinal deformity (ASD) is a prevalent condition estimated at 38%. Symptomatic ASD is associated with substantial health care costs. The role of nonoperative interventions in the management of ASD remains elusive. The National Scoliosis Clinic's (NSC) scoliosis realignment therapy (SRT) is a personalized digital health exercise program for the nonoperative management of ASD.

Objective: This exploratory study had two objectives: (1) to evaluate the effect of the SRT program on users' intention of having spinal fusion; and (2) from a US payer perspective, to estimate the annual cost savings per 100,000 beneficiaries by averting spinal surgery.

Methods: Individuals were enrolled in the SRT study from October 1, 2023 to September 1, 2024. Participants completed a web-based, cross-sectional survey about their history of prior scoliosis surgery and intent of having surgery before and after use of SRT (on a 4-point Likert scale, where 1 = "No Intent for Surgery" and 4 = "High Intent for Surgery"). Intent for surgery before and after participation in SRT was compared using a nonparametric Wilcoxon signed-rank test for paired data. Annual cost savings per 100,000 beneficiaries by averting spinal fusions were estimated separately for commercial payers and Medicare using published literature and public data sources. Payer expenditures were inflation-adjusted to 2024 US dollars using the Hospital Services component of the Consumer Price Index.

Results: A total of 62 NSC members (38.8%) responded to the survey and were enrolled in the SRT program for an average (SD) of 17 (12) weeks. The mean (SD) age was 65.3 (13.5) years, and the majority were female (47/48, 98%) and White (45/46, 98%). Among the SRT users who did not have prior scoliosis surgery (n=56), 14% (8/56) reported a decrease in intent for surgery (that is, a lower Likert score) with the use of SRT. The mean (SD) intent for surgery scores before compared to after SRT were 1.29 (0.53) and 1.14 (0.35), respectively (mean difference 0.15 [P=.006]). Participants with "No Intent for Surgery" pre- versus postuse of SRT (42/56 versus 48/56, respectively) corresponded to an absolute risk reduction of 11% and a number needed to treat of 9 to avert one spinal fusion. Among the 6 participants who transitioned to "No Intent" for spinal surgery with the use of SRT, 3 were aged <65 years and 3 were ≥65 years of age. The annual cost savings from averted spinal surgeries were estimated at US $415,000 per 100,000 commercially-insured beneficiaries and US $617,000 per 100,000 Medicare beneficiaries.

Conclusions: SRT is a personalized, scoliosis-specific digital health exercise program with the potential for averting 1 spinal surgery for every 9 participants, resulting in a substantial reduction in payer expenditures while improving the quality of care for commercial payers and Medicare beneficiaries.

(JMIR Form Res 2025;9:e66889) doi:10.2196/66889

KEYWORDS

adult spinal deformity; scoliosis; nonoperative management; spinal realignment therapy; digital health; program evaluation

XSL•FO
RenderX

Introduction

Adult spinal deformity (ASD) is a common condition with an estimated prevalence of 38% for primary (de novo) degenerative scoliosis [1,2]. ASD is a 3-dimensional deformity of the spine defined by a major curve magnitude angle of ≥10°, and women are more likely to have the diagnosis [1-4]. Patients with symptomatic ASD report increased back pain as well as lower health-related quality of life as demonstrated by worse SF-36 (36-Item Short Form Health Survey) scores compared to the general population [5,6].

As the US population continues to age, there is growing interest in understanding the most effective means to manage ASD. While nonoperative modalities are typically the first-line treatment for patients with symptomatic ASD, the role of nonoperative management has been questioned, with some studies reporting 2-year costs ranging from US $2041 to US $14,022 without improvements in patient outcomes [7-9]. In contrast, several studies have reported that patients who receive operative interventions for ASD have significantly reduced disability and pain and better improvement in clinical outcomes compared to nonoperative treatment [9-12]. However, spinal fusion is costly (Medicare reimbursement of $60,269 [13]) and carries significant risks; a multicenter database study of nearly 1000 adult surgical patients with ASD with 2-year follow-up reported an overall complication rate of 67.4% [14], and a separate meta-analysis demonstrated complication rates ranging from 17.0% to 71.5% [12]. Considering the substantial economic burden and complication rate associated with operative interventions for ASD—and noting that, other digital care programs in orthopedics (low back pain and chronic knee pain) have demonstrated significant reductions in surgical intent [15-17]—this study sought to further elucidate the role of a digital health program in the nonoperative management of ASD.

The National Scoliosis Clinic's (NSC) Scoliosis Realignment Therapy (SRT) is a remote, exercise-based, ScolioPilates® therapy specifically designed for individuals with scoliosis. ScolioPilates® is a scoliosis-specific exercise program using elongation, corrective breathing, strengthening, and activities of daily living to address pain associated with scoliosis. Additionally, members of SRT are enrolled in a supportive community with virtual group sessions, education, and AI-driven technology to assess spinal curvature with personalized therapy. While Clohisy et al [18] reported that the majority of patients crossover from nonoperative to operative treatment due to perceived worsening of symptoms, Rhode et al [19] recently reported that the use of SRT for 6 weeks resulted in significant improvement in pain as measured using the Scoliosis Research Society Health-Related Quality of Life Questionnaire (SRS-22r).

This exploratory study had two objectives: (1) to evaluate the effect of the SRT program on users' intention of having spinal fusion; and (2) from a US payer perspective, to estimate the annual cost savings per 100,000 beneficiaries by averting spinal surgery.

Methods

Recruitment

Individuals with scoliosis were recruited to the NSC through multiple mechanisms, including social media and the NSC website. NSC members were enrolled nationwide in the SRT study between October 1, 2023 and September 1, 2024. All NSC members using SRT during the aforementioned time period were invited by email to participate in the closed, web-based, cross-sectional survey about surgical intention. A formal sample size calculation was not performed, given that this was an exploratory study. In addition to demographics and clinical characteristics, the survey asked, "Has a surgeon ever told you you need scoliosis surgery, or offered you surgery?", "Have you ever intended to have scoliosis surgery?" and "Since you have started working with the NSC SRT, what is your current intention to have scoliosis surgery?" Responses to the questions about surgical intent were graded on a 4-point Likert scale, where 1=No Intent for Surgery, 2=Low Intent, 3=Moderate Intent, and 4=High Intent for Surgery. Given that this was an exploratory web-based survey, radiographic parameters were not collected.

Statistical Analysis

The primary outcome measure was the intent to pursue surgical correction of scoliosis pain. A nonparametric Wilcoxon signed-rank test for paired data was conducted to evaluate the difference in intention scores before and after SRT. The Wilcoxon test is appropriate for small samples that are not normally distributed. All analyses were performed using R software (version 4.4.1; R Foundation for Statistical Computing).

For potentially averted spinal surgeries, annual cost savings per 100,000 beneficiaries were estimated separately for US commercial payers and Medicare using published literature and multiple public data sources (Table 1). Payer expenditures were inflation-adjusted to US dollars in 2024 by using the Hospital Services component of the Consumer Price Index.

XSL•FO
RenderX

Table . Parameter values, data sources, and estimated cost savings from potentially averted spinal fusions among individuals with adult spinal deformity in the United States (2024, $US) [20].

Parameter	Value		Source
	Commercial	Medicare	
Medical resource use for scoliosis by adults, %	0.502	1.531	[20,21]
Adult scoliosis patients per 100,000 beneficiaries, n	502	1531	Derived
Convert from nonoperative care to spinal fusion (annual), %	21.7	21.7	[22]
Conversions to spinal fusion (annual), n	109	332	Derived
Absolute risk reduction in intent for surgery, %	11	11	SRT survey
SRT[a] participation rate (users), %	28	28	[23]
Spinal fusions averted annually per 100,000 beneficiaries, n	3.4	10.2	Derived
Spinal fusion insurer payment, US $[bc]	123,551[d]	60,269	[13,24,25]
Annual cost savings per 100,000 beneficiaries, $	414,534	616,504	Derived

[a]SRT: scoliosis realignment therapy.

[b]Weighted average of diagnosis-related group (DRG) 456 (spinal fusion with spinal curvature with major complication, 25%), DRG 457 (spinal fusion with spinal curvature with complication, 59%), and DRG 458 (spinal fusion with spinal curvature without complication or major complication, 16%).

[c]Inflated from 2022 to 2024 (half) US $ using the Consumer Price Index for hospital services; multiplier 1.097.

[d]Calculated using a Medicare-to-commercial payment multiplier of 2.05 for hospital inpatient services.

Data Exclusion

Users with prior spinal fusion surgery were excluded from the surgical intent analysis. Participants were given 2 weeks to respond to the survey, with subsequent reminders. Users who failed to respond to the surgical intent questions were excluded.

Ethical Considerations

This web-based survey evaluated a wellness program, not a medical treatment, and study data were deidentified for privacy and confidentiality protection. The datasets generated and analyzed during this study were restricted to the researchers conducting the analysis. NSC members who participated in the web-based survey study were provided US $50 gift cards for the completion of the survey.

Per the Office of Human Research Protections, under the US Department of Health and Human Services, research that (1) involves only survey procedures of adults and (2) is collected in a deidentified fashion is exempt from institutional review board (IRB) review and informed consent can be waived [26]. A retrospective IRB exemption was obtained for this study, confirming that both conditions are met, and therefore, this study was IRB exempt, and informed consent was not indicated.

Results

User Statistics

In total, 160 NSC members were invited to participate in the survey; 62 (38.8%) individuals nationwide voluntarily completed the survey. Six participants were excluded from the intention to pursue surgery analysis: 5 who had prior fusion surgery and 1 who did not respond. Among survey respondents who reported age (N=47), the mean (SD) age was 65.3 (13.5) years (age <65 years, 40%; age ≥65 years, 60%). The mean (SD) SRT duration of use was 118 (86.6 days); the majority were female (47/48, 98%) and White (45/46, 98%; Table 2).

XSL•FO
RenderX

Table . Demographics and clinical characteristics of scoliosis realignment therapy (SRT) cross-sectional survey participants with adult spinal deformity (n=62).

Characteristic	Value
Age in years, mean (SD)	65.3 (13.5)
Age <65 years, n (%)	19 (30.6)
Age ≥65 years, n (%)	28 (45.2)
Did not respond	15 (24.2)
Gender, n (%)	
Female	47 (75.8)
Male	1 (1.6)
Did not respond	14 (22.6)
Race, n (%)	
White	45 (72.6)
American Indian and Caucasian	1 (1.6)
Did not respond	16 (25.8)
SRT Duration of Use in days, mean (SD)	118 (86.6)
Offered scoliosis surgery in the past, n (%)	
Yes	15 (24.2)
No	26 (41.9)
Did not respond	21 (33.9)
Prior fusion surgery for scoliosis, n (%)[a]	
Yes	5 (8.1)
No	56 (90.3)
Did not respond	1 (1.6)

[a]Individuals with either prior fusion surgery or no response excluded from surgical intent analysis.

Evaluation Outcomes

Surgical Intent

Overall, 14% of participants (8/56) reported a decrease in intent for scoliosis surgery (ie, a lower Likert score) with the use of SRT, and no participants reported an increase in surgical intent. The mean (SD) intent for surgery scores before compared to after SRT were 1.29 (0.53) and 1.14 (0.35), respectively (mean difference 0.15 [*P*=.006]; Table 3). Participants with "No Intent" for spinal surgery pre- versus postuse of SRT (42/56 versus 48/56, respectively) corresponded to an absolute risk reduction of 11% and a number needed to treat of 9 to potentially avert 1 spinal fusion (1 divided by 0.11). Among the 6 participants who transitioned to "No Intent" for spinal surgery with the use of SRT, 3 were aged <65 years (ages 30, 62, and 64 years, respectively) and 3 were aged ≥65 years (ages 73, 73, and 82 years, respectively).

XSL•FO
RenderX

Table . Intentions for surgical intervention for adult spinal deformity prior to and after using scoliosis realignment therapy (SRT; cross-sectional survey; n=56).

Statement	Percentage of respondents				Mean (SD)[a]
	No intent	Low	Moderate	High	
Have you ever intended to have scoliosis surgery?	75.0	21.4	3.6	0.0	1.29 (0.530)
Since you have started working with the NSC[b] SRT, what is your current intention to have scoliosis surgery?	85.7	14.3	0.0	0.0	1.14 (0.353)

[a]Four-point Likert scale, where 1=No Intent for Surgery, 2=Low Intent, 3=Moderate Intent, and 4=High Intent. *P*=0.006, Wilcoxon signed-rank test.
[b]NSC: National Scoliosis Clinic.

Estimated Cost Savings by Averting Spinal Fusion

The percentage of adult patients using medical services for scoliosis was 0.502% (age 18 - 64 years) and 1.531% (age ≥65 years) for commercial and Medicare beneficiaries, respectively, or, 502 and 1531 per 100,000 beneficiaries, respectively (Table 1). An estimated 21.7% (41/189) of these patients converted from nonoperative care to spinal fusion annually. Applying the SRT-related 11% absolute risk reduction and an assumed 28% SRT participation rate based on digital health wellness intervention programs without incentives, a total of 3.4 fusions would potentially be averted annually per 100,000 commercially insured beneficiaries and 10.2 fusions potentially avoided annually per 100,000 Medicare beneficiaries. Applying a Medicare payment of US $60,269 for hospital inpatient spinal fusion (weighted average of diagnosis-related groups [DRGs] 456 - 458 in 2024 in US $) and multiplier for Medicare-to-commercial payment of 2.05 (ie, commercial payment of US $123,551), the annual cost savings from averted surgeries were estimated at US $415,000 per 100,000 commercially insured beneficiaries, and US $617,000 per 100,000 Medicare beneficiaries (Table 1).

Discussion

Principal Findings

This web-based, cross-sectional survey study explored whether the use of SRT, a personalized digital therapy program for adults with scoliosis, resulted in a reduction in intent for spinal surgery. After using SRT for an average of 17 weeks, an 11% absolute risk reduction in intent for surgery was observed, which corresponds to a number needed to treat of 9, suggesting that one spine surgery could potentially be averted for every 9 individuals using SRT. Hence, SRT offers a promising nonoperative therapy that may decrease spinal fusions with its associated risks and economic burden.

Comparison With Prior Work

The reduced intent in pursuing surgery with the use of SRT is consistent with other digital care programs in orthopedics, recognizing that scoliosis is a unique condition with its own clinical challenges. For a 12-week digital care program in low back pain, Shebib et al [15] demonstrated a significant reduction in surgical interest (*P*=.01). Similarly, Smittenaar and colleagues [16] reported a decrease in surgery intent at 3 months (*P*<.001) with a 12-week digital care program for chronic knee pain. Mecklenburg et al [17] also found that the self-reported likelihood of having surgery decreased over 1 year with the use of a digital care program for chronic knee pain (*P*=.01).

Investigators have reported conversion rates from nonoperative to operative treatment ranging from 12.7% (24/189) within 1 year to 31% (42/135) after 6 months [18,22]. Clohisy et al [18] further noted that the majority (90%) of patients crossed over from nonoperative to operative treatment due to perceived worsening of symptoms, and the remainder because the patients believed that the nonsurgical therapies were ineffective. Rhode et al [19] recently reported outcomes following 6 weeks of SRT using SRS-22r. Specifically, a significant improvement was observed in the pain (*P*<.001), self-image (*P*=.05), and mental health (*P*<.001) subdomains of the SRS-22r, with the improvement in the pain subdomain exceeding the minimal clinically important difference threshold. In addition, these investigators reported high satisfaction with the SRT program (9.5 out of 10, where 1=Extremely Dissatisfied and 10=Extremely Satisfied) [19]. Given that the SRT cohort reported in the study by Rhode et al (n=23) [19] is a subset of those reported herein (n=62), the statistically significant decrease in intent for surgery is consistent with the previously reported perceivable improvement in pain with the use of SRT. These consistent findings across SRT studies support the validity of using the question on surgical intent to extrapolate to the estimated cost savings.

In light of the aging population among other factors, Wadhwa and colleagues [27] reported that, among both commercial and Medicare beneficiaries, the rate of fusion for adult spinal deformity surgery doubled from 2007 to 2015. In addition, Passias et al [9] reported that the mean 2-year cost of nonoperative treatment was US $2041, and the mean 2-year cost for operative treatment was US $66,860 based on Medicare reimbursements. As such, the estimated cost savings for averted surgeries reported in the present study are likely underestimated

XSL•FO
RenderX

for multiple reasons, specifically, the percentage of beneficiaries seeking medical care for scoliosis is based on data from 2013 [20], the DRG reimbursement amounts reflect only the spinal surgery hospitalization (including the global period) [13], and incentives could increase the SRT participation rate [23].

The cost analyses are based on the best available published data, recognizing that more recent trends in insurer reimbursement and technological advancements may not be reflected. The cost-savings estimates rely on several assumptions, including a 28% participation rate for intervention-focused digital wellness programs. Mattke et al [23] described the relationship between incentive structure and participation rate, noting that the use of incentives increased wellness program participation. Our analysis assumed no incentives with an intervention-focused participation rate of 28% (versus 30% if any incentives). Others have reported a 32% participation rate for physical well-being programs [28]. Hence, assuming the absence of incentives offers a conservative estimate of the participation rate.

Limitations

There are several limitations of this formative research study. First, participants used SRT for an average of 17 weeks, which begs the question about whether the observed decrease in surgical intent would be sustained. That said, Smittenaar et al [16] found that the reduction in surgery interest continued for 6 months following the initiation of a digital health program for chronic knee pain ($P<.001$), and Mecklenburg et al [17] reported that the decreased likelihood of knee surgery with the use of a digital care program was sustained over 5 years ($P=.002$). Second, the self-reported survey data are subjective and participants completed both the pre- and post-SRT questions about surgical intention after initiating SRT, which potentially may be subject to recall bias. Third, the pre-post study design may be subject to confounding because it lacks an independent control or comparison group and is less rigorous than a randomized control trial; notwithstanding, each participant served as their own control. Furthermore, study participants were not blinded in that they were aware of using SRT, which may have introduced the possibility of performance bias. Lastly, although participants represented a nationwide sample of adults with scoliosis, the study included a relatively small cohort that did not permit the exploration of nuanced effects and was not ethnically diverse; therefore, the results may not be generalizable. Future studies will explore the durability of SRT use on surgical intent with a larger, more diverse cohort.

Conclusions

SRT is a personalized, scoliosis-specific digital health exercise program with the potential for averting 1 spinal fusion for every 9 participants, resulting in a substantial reduction in payer expenditures while improving the quality of care for commercial payer and Medicare beneficiaries. The SRT remote, digital care program holds promise as a feasible strategy to transform the nonoperative paradigm in adult spinal deformity. Digital care delivery is becoming more common with currently available health technology, and can increase patient agency and engagement, while improving overall outcomes and decreasing health care spending.

Acknowledgments

Generative AI was not used in any portion of the manuscript writing.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

MCB was involved in the conceptualization, data analysis, interpretation of findings, and manuscript preparation. CQL was involved in the conceptualization, interpretation of findings, and manuscript preparation. CJ was involved in the conceptualization, data acquisition, data analysis, and manuscript preparation. MR was involved in the conceptualization and manuscript preparation. BR, JB, and BIW were involved in the conceptualization, manuscript preparation, and critical manuscript review. SJA was involved in the conceptualization, data analysis, interpretation of findings, manuscript preparation, and critical manuscript review.

Conflicts of Interest

SJA declares to have an advisory contract with NSite Medical.
MCB declares consulting for NSite Medical.
All other authors have no conflicts of interest to disclose.

References

1. Bunnell WP. The natural history of idiopathic scoliosis before skeletal maturity. Spine (Phila Pa 1986) 1986 Oct;11(8):773-776. [doi: 10.1097/00007632-198610000-00003]
2. McAviney J, Roberts C, Sullivan B, Alevras AJ, Graham PL, Brown BT. The prevalence of adult de novo scoliosis: a systematic review and meta-analysis. Eur Spine J 2020 Dec;29(12):2960-2969. [doi: 10.1007/s00586-020-06453-0] [Medline: 32440771]
3. Schwab F, Dubey A, Gamez L, et al. Adult scoliosis: prevalence, SF-36, and nutritional parameters in an elderly volunteer population. Spine (Phila Pa 1986) 2005 May;30(9):1082-1085. [doi: 10.1097/01.brs.0000160842.43482.cd]

XSL•FO
RenderX

4. Francis RS. Scoliosis screening of 3,000 college-aged women. The Utah study--phase 2. Phys Ther 1988 Oct;68(10):1513-1516. [Medline: 3174833]
5. Jackson RP, Simmons EH, Stripinis D. Incidence and severity of back pain in adult idiopathic scoliosis. Spine (Phila Pa 1976) 1983 Oct;8(7):749-756. [doi: 10.1097/00007632-198310000-00011] [Medline: 6229884]
6. Bess S, Line B, Fu KM, et al. The health impact of symptomatic adult spinal deformity: comparison of deformity types to United States population norms and chronic diseases. Spine (Phila Pa 1976) 2016 Feb;41(3):224-233. [doi: 10.1097/BRS.0000000000001202] [Medline: 26571174]
7. Glassman SD, Carreon LY, Shaffrey CI, et al. The costs and benefits of nonoperative management for adult scoliosis. Spine (Phila Pa 1986) 2010 Mar;35(5):578-582. [doi: 10.1097/BRS.0b013e3181b0f2f8]
8. Teles AR, Mattei TA, Righesso O, Falavigna A. Effectiveness of operative and nonoperative care for adult spinal deformity: systematic review of the literature. Global Spine J 2017 Apr;7(2):170-178. [doi: 10.1177/2192568217699182] [Medline: 28507887]
9. Passias PG, Ahmad W, Dave P, et al. Economic burden of nonoperative treatment of adult spinal deformity. J Neurosurg Spine 2023 Dec 1;39(6):751-756. [doi: 10.3171/2023.7.SPINE23195] [Medline: 37728175]
10. Bridwell KH, Glassman S, Horton W, et al. Does treatment (nonoperative and operative) improve the two-year quality of life in patients with adult symptomatic lumbar scoliosis. Spine (Phila Pa 1986) 2009 Sep;34(20):2171-2178. [doi: 10.1097/BRS.0b013e3181a8fdc8]
11. Liu S, Schwab F, Smith JS, et al. Likelihood of reaching minimal clinically important difference in adult spinal deformity: a comparison of operative and nonoperative treatment. Ochsner J 2014;14(1):67-77. [Medline: 24688336]
12. Choi SH, Son SM, Goh TS, Park W, Lee JS. Outcomes of operative and nonoperative treatment in patients with adult spinal deformity with a minimum 2-year follow-up: a meta-analysis. World Neurosurg 2018 Dec;120:e870-e876. [doi: 10.1016/j.wneu.2018.08.179] [Medline: 30189313]
13. Data.CMS.gov centers for medicare and medicaid services. Medicare Inpatient Hospitals - by Geography and Service Data Dictionary | CMS Data. URL: https://data.cms.gov/resources/medicare-inpatient-hospitals-by-geography-and-service-data-dictionary-0 [accessed 2024-09-17]
14. Lafage R, Bass RD, Klineberg E, et al. Complication rates following adult spinal deformity surgery: evaluation of the category of complication and chronology. Spine (Phila Pa 1976) 2024;49(12):829-839. [doi: 10.1097/BRS.0000000000004969]
15. Shebib R, Bailey JF, Smittenaar P, Perez DA, Mecklenburg G, Hunter S. Randomized controlled trial of a 12-week digital care program in improving low back pain. NPJ Digit Med 2019;2(1):1. [doi: 10.1038/s41746-018-0076-7] [Medline: 31304351]
16. Smittenaar P, Erhart-Hledik JC, Kinsella R, Hunter S, Mecklenburg G, Perez D. Translating comprehensive conservative care for chronic knee pain into a digital care pathway: 12-week and 6-month outcomes for the hinge health program. JMIR Rehabil Assist Technol 2017 Apr 5;4(1):e4. [doi: 10.2196/rehab.7258] [Medline: 28582253]
17. Mecklenburg G, Smittenaar P, Erhart-Hledik JC, Perez DA, Hunter S. Effects of a 12-week digital care program for chronic knee pain on pain, mobility, and surgery risk: randomized controlled trial. J Med Internet Res 2018 Apr 25;20(4):e156. [doi: 10.2196/jmir.9667] [Medline: 29695370]
18. Clohisy JCF, Smith JS, Kelly MP, et al. Failure of nonoperative care in adult symptomatic lumbar scoliosis: incidence, timing, and risk factors for conversion from nonoperative to operative treatment. J Neurosurg Spine 2023 Oct 1;39(4):498-508. [doi: 10.3171/2023.5.SPINE2326] [Medline: 37327144]
19. Rhode M, Lin C, Jin C, et al. Does a digital care program for adult scoliosis improve outcomes? A pilot study. J Spine 2024;13(4). [doi: 10.2196/preprints.66889]
20. Daubs MD, Watkins-Castillo SI. Orthopedics research society (ORS), United States bone & joint initiative. Table 2B22: Healthcare visits for conditions contributing to curvature of spine disorders by age, United States. 2013. URL: https://www.bmus-ors.org/docs/bmus_e4_T2B.2.2.pdf [accessed 2024-09-08]
21. Projected population by five-year age group and sex for the united states. : U.S. Census Bureau, Population Division; 2023 Nov URL: https://www.census.gov/data/tables/2023/demo/popproj/2023-summary-tables.html [accessed 2024-09-08]
22. Passias PG, Jalai CM, Line BG, et al. Patient profiling can identify patients with adult spinal deformity (ASD) at risk for conversion from nonoperative to surgical treatment: initial steps to reduce ineffective ASD management. Spine J 2018 Feb;18(2):234-244. [doi: 10.1016/j.spinee.2017.06.044] [Medline: 28688984]
23. Mattke S, Kapinos KA, Caloyeras JP, et al. Workplace Wellness Programs: Services Offered, Participation, and Incentives.: RAND Corporation; 2015. URL: https://www.rand.org/pubs/research_reports/RR724.html [accessed 2024-09-17]
24. U.S. bureau of labor statistics. Consumer Price Index for All Urban Consumers, Hospital Services. URL: https://data.bls.gov/pdq/SurveyOutputServlet [accessed 2024-09-08]
25. Commercial reimbursement benchmarking: Commercial payment rates for medical services as percentage of Medicare fee-for-service rates. URL: https://www.milliman.com/en/insight/commercial-reimbursement-benchmarking-payment-rates-medicare-fee-for-service [accessed 2024-09-17]
26. U.S. department of health and human services. Human Subject Regulations Decision Charts: 2018 Requirements. URL: https://www.hhs.gov/ohrp/regulations-and-policy/decision-charts-2018/index.html#c1 [accessed 2025-04-22]

XSL•FO
RenderX

27. Wadhwa H, Leung C, Sklar M, et al. Utilization trends, cost, and payments for adult spinal deformity surgery in commercial and medicare-insured populations. Neurosurgery 2022 Dec 1;91(6):961-968. [doi: 10.1227/neu.0000000000002140] [Medline: 36136402]
28. Valencia C. How to get employees to (actually) participate in well-being programs. Harvard Business Review Published online October. 2021. URL: https://hbr.org/2021/10/how-to-get-employees-to-actually-participate-in-well-being-programs [accessed 2024-11-05]

Abbreviations

ASD: adult spinal deformity
DRG: diagnosis-related group
IRB: institutional review board
NSC: National Scoliosis Clinic
SF-36: 36-Item Short Form Health Survey
SRT: scoliosis realignment therapy

Edited by A Mavragani; submitted 25.09.24; peer-reviewed by RS Punekar; revised version received 08.04.25; accepted 10.04.25; published 29.04.25.

Please cite as:
Brown MC, Lin CQ, Jin C, Rohde M, Rocos B, Belding J, Woods BI, Ackerman SJ
Effect of a Digital Health Exercise Program on the Intention for Spinal Surgery in Adult Spinal Deformity: Exploratory Cross-Sectional Survey
JMIR Form Res 2025;9:e66889
URL: https://formative.jmir.org/2025/1/e66889
doi:10.2196/66889

© Marsalis Christian Brown, Christopher Quincy Lin, Christopher Jin, Matthew Rohde, Brett Rocos, Jonathan Belding, Barrett I Woods, Stacey J Ackerman. Originally published in JMIR Formative Research (https://formative.jmir.org), 29.4.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Assessing the Utilization of Electronic Consultations in Genetics: Seven-Year Retrospective Study

Sawona Biswas[1,2], MS; Joyce So[2], MD, PhD; Robert Wallerstein[2], MD; Ralph Gonzales[3], MD; Delphine Tout[4], MD; Lisa DeAngelis[5], BA; Aleksandar Rajkovic[1,6], MD, PhD

[1]Department of Pathology, University of California San Francisco, 2340 Sutter Street, San Francisco, CA, United States
[2]Department of Pediatrics, University of California, San Francisco, San Francisco, CA, United States
[3]Department of Medicine, University of California, San Francisco, San Francisco, CA, United States
[4]Division of Nephrology, University of California, San Francisco, San Francisco, CA, United States
[5]Ambulatory Patient Care Access, UCSF Health, San Francisco, CA, United States
[6]Department of Obstetrics, Gynecology & Reproductive Sciences, University of California, San Francisco, CA, United States

Corresponding Author:
Sawona Biswas, MS
Department of Pathology, University of California San Francisco, 2340 Sutter Street, San Francisco, CA, United States

Abstract

Background: Patient and health care provider access to genetic subspecialists is challenging owing to limited number of genetics experts across the United States. The University of California San Francisco (UCSF) Genetics electronic consultation (e-Consult) service was implemented along with the usual referral pathway to improve access to timely genetic expertise through robust asynchronous provider-to-provider communication.

Objectives: This study examined the impact of the UCSF Genetics e-Consult service on patient access to genetics expertise.

Methods: A retrospective chart review of 622 e-Consult requests was conducted. Data pertinent to e-Consult completion rates, provider response times, consultation content, and adherence to geneticist recommendations were abstracted.

Results: From October 2016 to March 2024, the UCSF Genetics e-Consult service received a total of 622 consultation orders, with yearly volumes increasing from 34 in 2017 to 144 in 2023. A total of 360/622 (57.8%) consultations were completed, of which 197/360 (54.6%) were resolved without requiring a specialty care visit. Of the 262/622 (42.1%) e-Consult orders declined by the geneticist reviewer, 184/262 (70.2%) were scheduled for a synchronous genetics visit due to case complexity precluding an appropriate e-Consult response and 29.8% (78/262) were recommended to be referred to a different and more appropriate specialty. Geneticists responded to 83.9% (522/622) of e-Consults within 3 days, with most spending between 5 and 20 minutes on their e-Consult response. Nearly half of the genetics e-Consult requests (69/144; 47.9%) came from primary care providers and pediatricians. Among the 144 e-Consult requests in 2023, 50.6% (73/144) were about diagnostic queries, 17% (25/144) were on symptom management, and 11% (16/144) were about test interpretation. Provider adherence to geneticists' recommendations was observed in 84% (116/144) of cases.

Conclusions: The UCSF Genetics e-Consult service has demonstrated efficiency in providing timely genetic consultations, with a high rate of provider adherence to recommendations. These findings support the potential of e-Consult frameworks as a viable strategy for enhancing access to genetic health care services.

(JMIR Form Res 2025;9:e63028) doi:10.2196/63028

KEYWORDS

genomic; e-Consult; genetic; utility; retrospective; assessment; effectiveness; electronic consultation; healthcare providers; genetic experts; university; consultations; e-Consult frameworks; accessibility; genetic testing; patient care

Introduction

The field of genetics and genomics has experienced significant advancements in recent years, leading to increased integration of genetic testing and considerations into various medical disciplines. However, the number of genetic subspecialists remains limited, resulting in challenges for patients and health care providers seeking timely access to genetics expertise [1,2]. The shortage of genetic specialists, coupled with the growing demand for their services, has led to prolonged waiting times for appointments and potential delays in patient care [3].

As genetic testing is increasingly incorporated into the diagnostic and treatment processes across medical specialties, there is a pressing need to improve communication and collaboration between genetic specialists and a wide range of health care providers [4]. Primary care clinicians, specialists in

XSL•FO
RenderX

other disciplines, and allied health professionals often require guidance and support when interpreting genetic test results, understanding their implications, and supporting patient informed decisions regarding their management [5].

To address these challenges, health care systems have explored innovative solutions, such as electronic consultations (e-Consults), to facilitate prompt access to specialist expertise [6]. e-Consults have emerged as a promising approach to bridge the gap between the limited number of genetic specialists and the growing demand for their service. Within genetics and genomics programs, e-Consult initiatives have demonstrated their effectiveness in addressing basic questions without the need for an in-person visit and providing educational opportunities for generalists to enhance their knowledge and skills in genetics care [7-10]. As genetics becomes increasingly integrated into mainstream medical practice, e-Consults serve as a valuable tool to foster collaboration, ensure knowledge sharing, and provide high-quality, genetics-informed care to patients.

The e-Consult service at the University of California San Francisco (UCSF) started in 2012 [11]. Initially, the program was grant funded and then financially supported by the health system. By 2015, the UCSF entered into agreements with three payers to receive reimbursement for specialist e-Consults, and by 2020, the Centers for Medicare and Medicaid Services (CMS) and other payers started providing reimbursement for e-Consults that are completed and not declined. The introduction of Genetics e-Consults occurred in June 2017, positioning the UCSF as an early adopter in the United States. Herein, we describe the establishment of our Genetics e-Consult service for providers within our institution and its utilization.

Methods

Study Structure

This study analyzed e-Consults submitted through the UCSF Health's electronic medical record system. Providers across multiple specialties, including primary care, pediatrics, and oncology, initiated consultations. The analysis included all Genetics e-Consults from October 2016 to March 2024, capturing utilization trends and consultation characteristics.

Process of e-Consult Services

Ordering providers within UCSF Health can initiate an e-Consult request through our hospital's electronic medical record (Epic Hyperspace 2023; developed by Epic Systems Corporation, Verona, WI, USA). Providers can choose from the Adult Genetics, Pediatrics Genetics, or Cancer Risk Genetics service. After selecting the service, the ordering providers will be prompted to choose one of the four diagnosis-specific SmartText templates: (1) Query genetic condition or syndrome, (2) Known genetic condition or syndrome, (3) Family history of genetic condition, or (4) Other. Each template will guide the provider through different open-ended questions that capture the reason for the e-Consult, relevant test results, family history information, and any other pertinent information. After submission, the e-Consult request is routed to the designated pool and the e-Consultant on service reviews the details and responds.

The Genetics e-Consultants are a group of board-certified geneticists practicing in the Pediatric or Adult Genetics service at the UCSF with more than 5 years of experience. The e-Consultant after accepting the referral summarizes or refines the consultative questions, offers an individualized recommendation along with supporting evidence or current best practices, and outlines a contingency plan (including guidelines for when to refer the patient for an in-person or telehealth specialty visit). The standard response time for an e-Consult is set at 3 business days. In cases where the clinical complexity exceeds the scope of an e-Consult, the e-Consultant may opt to recommend a traditional new patient appointment. The e-Consultant will cc: their specialty staff on the e-Consult response who then convert the e-Consult order to a referral order for routing to the appropriate scheduling team (upon submittal, the requester indicates okay to convert to referral or not okay to convert). Alternatively, if the consultation falls outside the consultant's area of expertise and is better suited for another specialty, the e-Consultant will decline the e-Consult with that recommendation. The e-Consultant self-reports the time spent in the electronic medical record before marking the consultation as complete. The median of these recorded times was then calculated. Within our department, a dedicated team of geneticists is responsible for addressing e-Consults. Once the e-Consult is completed or declined, it is returned to the referring provider's Epic In Basket (Epic Systems Corporation) an integrated electronic messaging and task management system within the Epic electronic health record, with either the clinical recommendations or reason for decline. A short 8-question provider survey was conducted to assess the utilization, satisfaction, and impact of the e-Consult service.

Retrospective Chart Review

A retrospective chart review was conducted to explore the details of e-Consults ordered from October 2016 to March 2024. The review did not require institutional review board approval at the UCSF. Our group characterized the ordering providers, departments, referral groups, descriptions of the consult, primary referral diagnosis, consult status, response times, e-Consult recommendations, number of related referrals to Genetics, and tests performed. Data were collected from EPIC reports and individual chart review, and descriptive statistical analysis was completed using R software (version 4.4.3; R Foundation for Statistical Computing). and Microsoft XML. Statistical reporting included all e-Consult activity originating from an order. Per UCSF Health e-Consult EPIC workflow, all e-Consult orders result in a completed and closed encounter for reporting, whether completed with a clinical recommendation or declined for the reason indicated.

Ethical Considerations

This retrospective quality control and quality improvement study involved a chart review of deidentified data from the electronic medical records at the UCSF. The study was conducted solely for the purpose of evaluating and improving the existing Genetics e-Consult service within the institution. As this project did not involve primary data collection and was

XSL•FO
RenderX

intended for internal quality improvement with no impact on patient care decisions or patient rights, it did not constitute human subjects research under the purview of the Department of Health and Human Services. Accordingly, this study did not require Institutional Review Board review or approval per UCSF's internal policies and federal regulations. All study data were deidentified and securely stored within UCSF's electronic health records system. No financial compensation was provided to study participants, as the research involved retrospective data analysis. No images or materials include identifiable patient information. This not a funded project. All authors are UCSF employees.

Results

From October 2016 to March 2024, the UCSF Genetics e-Consult service received a total of 622 consultation orders, with yearly volumes increasing from 34 in fiscal year (FY) 2017 to 144 in FY 2023 (Figure 1).

Of the 622 Genetics e-Consults, 360/622 (57.8%) were completed, while 262/622 (42.1%) were either declined or deferred to another specialty (Figure 2). Responses were provided on the same day for 263/622 consults (42.3%), within 3 days for 522/622 consults (83.9%), and within 7 days for 562/622 consults (90.4%; Multimedia Appendix 1).

A comprehensive case-level analysis was conducted for the 144 e-Consults ordered during the FY 2023 (June 2022 to July 2023), as this was the first year in which the e-Consult service was fully integrated, with standardized workflows, defined practices, and distinct divisions for Adult, Pediatric, and Cancer Genetics. Prior years lacked the necessary structure and documentation for a comparable analysis. Adult Genetics received the most consults (109/144, 75.7%), followed by Pediatric Genetics (25/144, 17.4%) and Adult Cancer Risk Genetics (10/144, 6.9%). A total of 88/144 (61.1%) of these e-Consults were completed with detailed responses, 46/144 (31.9%) were declined and converted to in-person Genetics appointments, and 10/144 (6.9%) were declined for other reasons such as insufficient information or referral to another specialty (Figure 3).

Nearly half (69/144, 47.9%) of the FY 2023 e-Consults originated from general practices such as primary care, general medicine, and women's health. The median time to answering an e-Consult was 15 (IQR 10 - 25) minutes. e-Consultants spent less than 5 minutes in all the declined cases and an average of 11 - 20 minutes on completed cases and mostly 5 - 10 minutes on cases that were declined and scheduled with Genetics (Figure 3). The most common types of e-Consult requests were queries about genetic diagnoses (73/144, 50.6%), assistance with test interpretation (16/144, 11.1%), and management recommendations for patients with known genetic diagnoses (13/144, 9%) or unknown genetic diagnoses (12/144, 8.3%; Multimedia Appendix 1). e-Consultants provided a range of services, with the most frequent being scheduling patients for outpatient visits (61/144, 42.3%), providing genetic counseling and management recommendations (35/144, 24.3%), making management recommendations alone (27/144, 18.5%), and offering guidance on genetic testing (17/144, 11.8%) and recommended genetics referrals (5/144, 3.4%; Multimedia Appendix 1). A chart review of 134 e-Consults (88 completed and 46 scheduled) revealed that 65/134 patients (48.5%) were referred to and seen by Genetics, 14/134 (12%) were referred with a pending appointment, 5/134 (3.7%) were referred but canceled their appointments, 31/134 (23.1%) had their care informed by the e-Consult recommendations, 5/134 (3.7%) had genetic testing ordered by their referring provider, 12/134 (9%) did not get a Genetics referral, and 2/134 (1.5%) did not have any genetic testing performed (Figure 3).

XSL•FO
RenderX

Figure 1. Total number of e-Consults received by genetics (Adult Genetics, Pediatrics Genetics, or Cancer Risk Genetics) at the University of California, San Francisco from Oct 01, 2016 till 01 March, 2024.

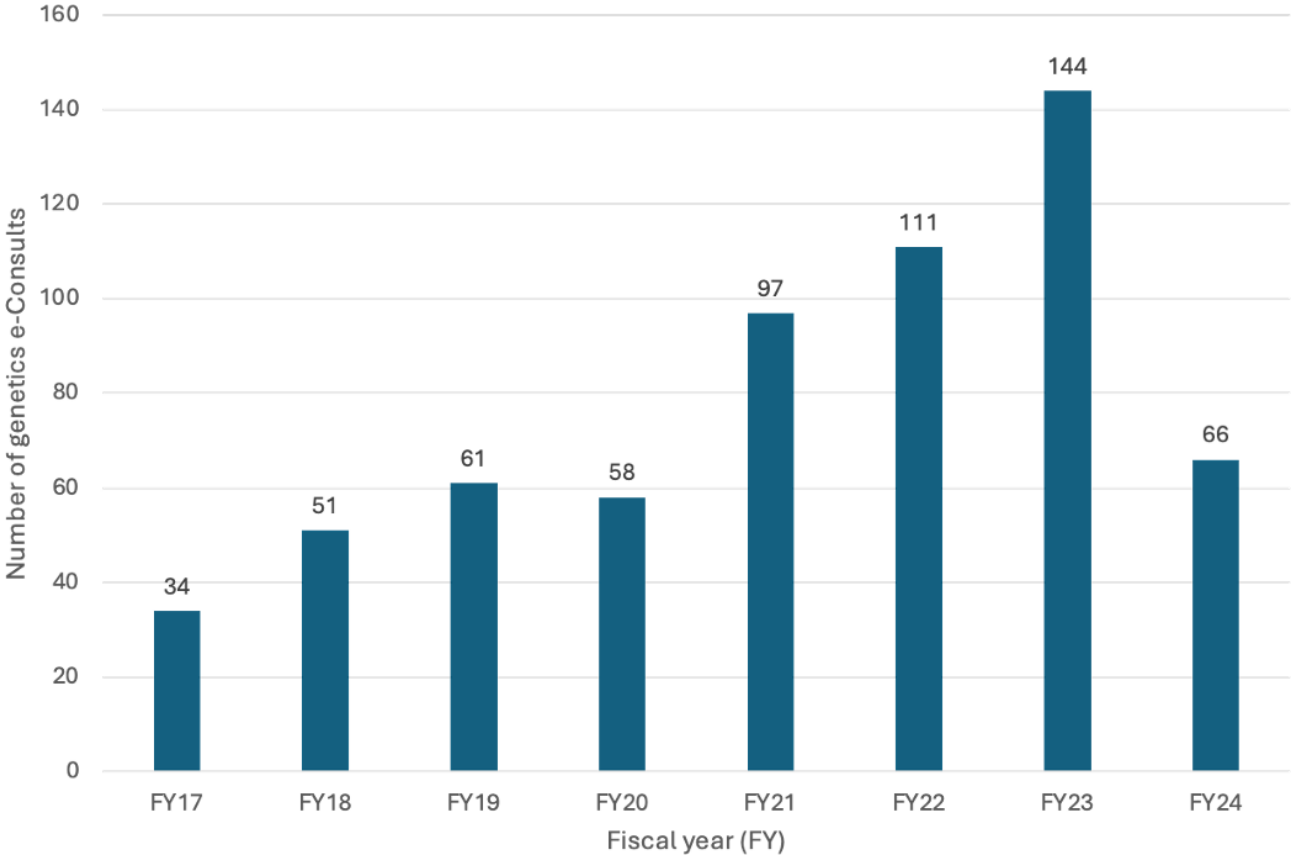

Figure 2. Flowchart and breakdown of e-Consults received and responded 2017 - 2024. UCSF: University of California, San Francisco

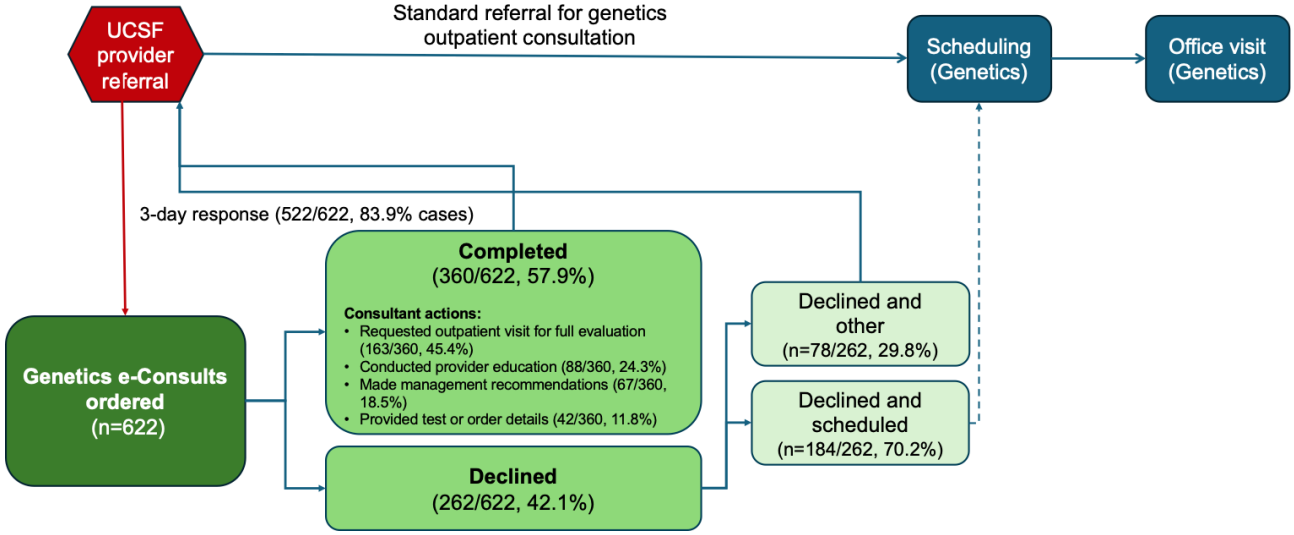

XSL•FO
RenderX

Figure 3. Breakdown of e-Consults received in the fiscal year (FY) 2023.

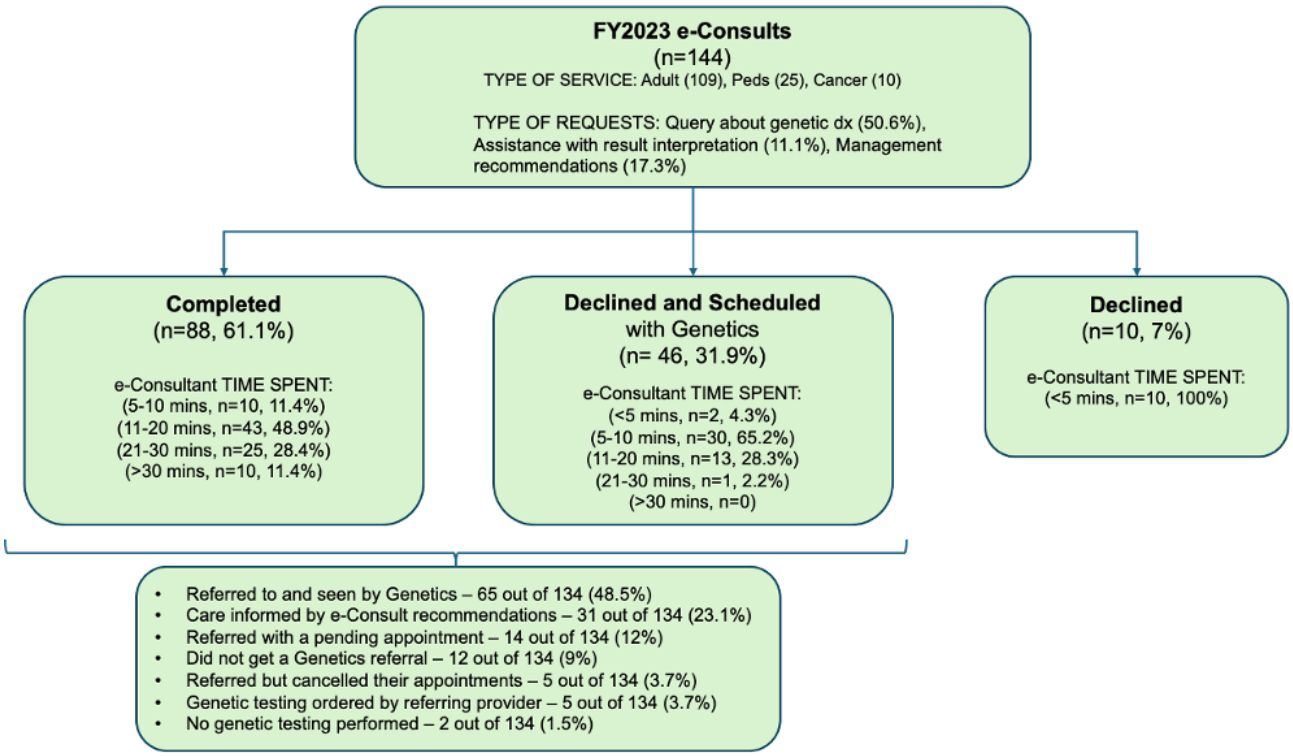

Discussion

Our study demonstrates the successful implementation and growth of an e-Consult service that connects general providers and specialists with genetics experts at a large academic medical center. The UCSF Genetics e-Consult service experienced a steady increase in volume from 34 consults in FY 2017 to 144 in FY 2023 (Figure 1), highlighting the growing demand for accessible genetics expertise.

The e-Consult service provided an efficient means to triage genetics questions, with 57.8% (360/622) of the submitted consults completed and 42.1% (262/622) declined, primarily to schedule in-person evaluations instead (184/262, 70.2% of declined cases). The average turnaround time of 3 days demonstrates the responsiveness of the service in addressing clinical questions.

An analysis of 144 e-Consults from FY 2023 revealed that Adult Genetics handled the majority 109/144 (75.7%), followed by Pediatric Genetics 25/144 (17.4%) and Adult Cancer Risk Genetics 10/144 (6.9%). While 88/144 (61.1%) of these consults were completed electronically, 46/144 (31.9%) were declined and converted to in-person visits (Figure 3), suggesting that e-Consults can help determine when face-to-face evaluations are needed. Nearly half (69/144, 47.9%) of FY 2023 e-Consults came from generalists in primary care, general medicine, and women's health, indicating broad adoption across nongenetics specialties. e-Consultants provided prompt responses, spending 20 minutes or less on 75% of the cases (Figure 3).

The most common reasons for e-Consults were questions about genetic diagnoses (73/144, 50.6%), test interpretations (16/144, 11.1%), and patient management (Multimedia Appendix 1). e-Consultants frequently recommended clinic visits (61/144, 42.3%), provided counseling and management advice (35/144, 24.3%), and offered guidance on genetic testing (17/144, 11.8%;). Referral outcomes data demonstrated that e-Consults resulted in 62% (84/134) of patients being seen by or scheduled with appointments at Genetics, while for 20% of patients (31/134), e-Consult advice influenced their care without necessitating a formal referral to Genetics.

These findings align with and expand upon the limited existing literature on Genetics e-Consult programs. Similar to studies by Bhola et al [9] and Folkerts et al [10], we observed efficient turnaround times, a wide range of clinical topics addressed, and high rates of guideline-concordant care when e-Consults provided actionable recommendations. Our higher e-Consult volume likely reflects the longer duration and wider scope of our program.

The high proportion of e-Consults completed electronically (360/622, 57.9%) highlights the potential for e-Consults to optimize genetics care delivery by addressing straightforward questions, triaging complex cases for in-person evaluation, and guiding appropriate previsit workups. The 84% adherence rate among providers to e-Consult recommendations underscores the educational impact of specialist guidance.

The limitations of our study include the single-center design, lack of long-term outcome data, and absence of referring provider and patient perspectives. Future research should compare e-Consult programs across diverse health care settings; incorporate feedback from key stakeholders; and evaluate the impact on care access, quality, and costs.

In conclusion, the UCSF Genetics e-Consult service demonstrates a scalable and efficient model for expanding access to genetics expertise. By providing timely, individualized guidance to nongenetics providers, e-Consults can help optimize

XSL•FO
RenderX

resource utilization, provider education, and patient care as genetics becomes increasingly integrated into routine clinical practice.

Acknowledgments

No Generative AI was used in any portion of the manuscript.

Data Availability

All data generated or analyzed during this study are included in this published article and in the supplementary information files. Any additional data can be requested from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary tables.
[DOCX File, 21 KB - formative_v9i1e63028_app1.docx]

References

1. Kieke MC, Conta JH, Riley JD, Zetzsche LH. The current landscape of genetic test stewardship: a multi-center prospective study. J Genet Couns 2021 Aug;30(4):1203-1210. [doi: 10.1002/jgc4.1403] [Medline: 33728729]
2. Maiese DR, Keehn A, Lyon M, Flannery D, Watson M, Working Groups of the National Coordinating Center for Seven Regional Genetics Service Collaboratives. Current conditions in medical genetics practice. Genet Med 2019 Aug;21(8):1874-1877. [doi: 10.1038/s41436-018-0417-6] [Medline: 30686822]
3. Dragojlovic N, Borle K, Kopac N, et al. The composition and capacity of the clinical genetics workforce in high-income countries: a scoping review. Genet Med 2020 Sep;22(9):1437-1449. [doi: 10.1038/s41436-020-0825-2] [Medline: 32576987]
4. Rubanovich CK, Cheung C, Mandel J, Bloss CS. Physician preparedness for big genomic data: a review of genomic medicine education initiatives in the United States. Hum Mol Genet 2018 Aug 1;27(R2):R250-R258. [doi: 10.1093/hmg/ddy170] [Medline: 29750248]
5. Salari K, Karczewski KJ, Hudgins L, Ormond KE. Evidence that personal genome testing enhances student learning in a course on genomics and personalized medicine. PLoS ONE 2013;8(7):e68853. [doi: 10.1371/journal.pone.0068853] [Medline: 23935898]
6. Carrillo de Albornoz S, Sia KL, Harris A. The effectiveness of teleconsultations in primary care: systematic review. Fam Pract 2022 Jan 19;39(1):168-182. [doi: 10.1093/fampra/cmab077] [Medline: 34278421]
7. Vidgen ME, Fowles LF, Istiko SN, et al. Evaluation of a genetics education program for health interpreters: a pilot study. Front Genet 2021;12:771892. [doi: 10.3389/fgene.2021.771892] [Medline: 35186003]
8. Vimalananda VG, Orlander JD, Afable MK, et al. Electronic consultations (e-consults) and their outcomes: a systematic review. J Am Med Inform Assoc 2020 Mar 1;27(3):471-479. [doi: 10.1093/jamia/ocz185] [Medline: 31621847]
9. Bhola PT, Liddy C, Afkham A, Keely E, Graham GE. A pilot eConsultation service in Eastern Ontario: bridging clinical genetics and primary care. Eur J Hum Genet 2019 Jul;27(7):1026-1032. [doi: 10.1038/s41431-019-0342-4] [Medline: 30778171]
10. Folkerts EK, Pelletier RC, Chung DC, et al. A pooled electronic consultation program to improve access to genetics specialists. Genet Med Open 2023;1(1):100819. [doi: 10.1016/j.gimo.2023.100819] [Medline: 39669243]
11. Gleason N, Prasad PA, Ackerman S, et al. Adoption and impact of an eConsult system in a fee-for-service setting. Healthc (Amst) 2017 Mar;5(1-2):40-45. [doi: 10.1016/j.hjdsi.2016.05.005] [Medline: 27469441]

Abbreviations

CMS: Centers for Medicare and Medicaid Services
e-Consult: electronic consultation
FY: fiscal year
UCSF: University of California San Francisco

XSL•FO
RenderX

Edited by A Mavragani; submitted 07.06.24; peer-reviewed by LE Hull, M Katapodi, M Jani; revised version received 10.03.25; accepted 11.03.25; published 30.04.25.

Please cite as:
Biswas S, So J, Wallerstein R, Gonzales R, Tout D, DeAngelis L, Rajkovic A
Assessing the Utilization of Electronic Consultations in Genetics: Seven-Year Retrospective Study
JMIR Form Res 2025;9:e63028
URL: https://formative.jmir.org/2025/1/e63028
doi:10.2196/63028

© Sawona Biswas, Joyce So, Robert Wallerstein, Ralph Gonzales, Delphine Tout, Lisa DeAngelis, Aleksandar Rajkovic. Originally published in JMIR Formative Research (https://formative.jmir.org), 30.4.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Predicting Transvaginal Surgical Mesh Exposure Outcomes Using an Integrated Dataset of Blood Cytokine Levels and Medical Record Data: Machine Learning Approach

Mihyun Lim Waugh[1], PhD; Tyler Mills[2], BS; Nicholas Boltin[1], PhD; Lauren Wolf[1], PhD; Patti Parker[3], BSN; Ronnie Horner[4], PhD; Thomas L Wheeler II[5], MD, MSPH; Richard L Goodwin[1], PhD; Melissa A Moss[1], PhD

[1]Department of Biomedical Engineering, University of South Carolina, 301 Main St, Rm 2C02, Columbia, SC, United States

[2]University of South Carolina School of Medicine, Columbia, SC, United States

[3]Prisma Health, Greenville, SC, United States

[4]Department of Health Services Research and Administration, University of Nebraska Medical Center, Omaha, NE, United States

[5]Department of Obstetrics and Gynecology, Spartanburg Regional Healthcare, Spartanburg, SC, United States

Corresponding Author:
Melissa A Moss, PhD
Department of Biomedical Engineering, University of South Carolina, 301 Main St, Rm 2C02, Columbia, SC, United States

Abstract

Background: Transvaginal insertion of polypropylene mesh was extensively used in surgical procedures to treat pelvic organ prolapse (POP) due to its cost-efficiency and durability. However, studies have reported a high rate of complications, including mesh exposure through the vaginal wall. Developing predictive models via supervised machine learning holds promise in identifying risk factors associated with such complications, thereby facilitating better informed surgical decisions. Previous studies have demonstrated the efficacy of anticipating medical outcomes by employing supervised machine learning approaches that integrate patient health care data with laboratory findings. However, such an approach has not been adopted within the realm of POP mesh surgery.

Objective: We examined the efficacy of supervised machine learning to predict mesh exposure following transvaginal POP surgery using 3 different datasets: (1) patient medical record data, (2) biomaterial-induced blood cytokine levels, and (3) the integration of both.

Methods: Blood samples and medical record data were collected from 20 female patients who had prior surgical intervention for POP using transvaginal polypropylene mesh. Of these subjects, 10 had experienced mesh exposure through the vaginal wall following surgery, and 10 had not. Standardized medical record data, including vital signs, previous diagnoses, and social history, were acquired from patient records. In addition, cytokine levels in patient blood samples incubated with sterile polypropylene mesh were measured via multiplex assay. Datasets were created with patient medical record data alone, blood cytokine levels alone, and the integration of both data. The data were split into 70% and 30% for training and testing sets, respectively, for machine learning models that predicted the presence or absence of postsurgical mesh exposure.

Results: Upon training the models with patient medical record data, systolic blood pressure, pulse pressure, and a history of alcohol usage emerged as the most significant factors for predicting mesh exposure. Conversely, when the models were trained solely on blood cytokine levels, interleukin (IL)-1β and IL-12 p40 stood out as the most influential cytokines in predicting mesh exposure. Using the combined dataset, new factors emerged as the primary predictors of mesh exposure: IL-8, tumor necrosis factor-α, and the presence of hemorrhoids. Remarkably, models trained on the integrated dataset demonstrated superior predictive capabilities with a prediction accuracy as high as 94%, surpassing the predictive performance of individual datasets.

Conclusions: Supervised machine learning models demonstrated improved prediction accuracy when trained using a composite dataset that combined patient medical record data and biomaterial-induced blood cytokine levels, surpassing the performance of models trained with either dataset in isolation. This result underscores the advantage of integrating health care data with blood biomarkers, presenting a promising avenue for predicting surgical outcomes in not only POP mesh procedures but also other surgeries involving biomaterials. Such an approach has the potential to enhance informed decision-making for both patients and surgeons, ultimately elevating the standard of patient care.

(JMIR Form Res 2025;9:e59631) doi:10.2196/59631

XSL•FO
RenderX

KEYWORDS

cytokines; inflammatory response; medical record; pelvic organ prolapse; polypropylene mesh; supervised machine learning models; polypropylene; mesh surgery; surgical outcome; cost-efficiency; risk factor; efficacy; health care data; female; informed decision-making; patient care; digital health

Introduction

Patients experiencing difficulty from symptoms of pelvic organ prolapse (POP) are prescribed nonsurgical option as a first-line treatment, including pelvic floor physical therapy to strengthen the muscles supporting the pelvic organs and the use of pessaries that are inserted into the vagina to provide structural support and alleviate symptoms. Lifestyle modifications are also suggested, such as weight management and avoidance of heavy lifting, to reduce symptoms. However, conservative treatments often fail to provide adequate relief, leading many patients to pursue surgical interventions. Notably, patients who undergo surgical treatments are more likely to report complete satisfaction 1 year after the procedure compared to those who opt for nonsurgical options [1]. The choice of surgical approach depends largely on the type of prolapse. Among these, transvaginal mesh surgery was once a widely used treatment when reinforcement of the pelvic floor structures was deemed necessary to treat severe or recurrent prolapse, though its use was later restricted following safety concerns raised by the Food and Drug Administration. This paper focuses on transvaginal mesh surgery to explore strategies for identifying patient populations that are more likely to experience favorable surgical outcomes, with the aim of minimizing the number of patients who face less successful results.

Surgical treatment via transvaginal mesh implantation uses polypropylene mesh due to its low cost, nontoxicity, and mechanical resilience. Its superior tensile strength and elasticity compared to native pelvic tissue enabled it to withstand the pressure changes within the pelvic cavity [2]. Despite these advantageous properties, the usage of polypropylene mesh in transvaginal repair of POP has been prohibited in multiple countries due to postsurgical complications, notably mesh exposure through the vaginal wall [2,3]. In fact, studies have reported mesh exposure rates ranging from 4% to 12.3% in POP patients following surgery [4-6]. Conversely, however, nearly 90% of patients experience successful POP treatment via transvaginal mesh. Consequently, accurately identifying risk factors and developing personalized predictive models can greatly benefit both patients and surgeons in making informed surgical decisions.

Building a machine learning model to predict postsurgical mesh exposure holds considerable promise, as it offers a robust framework that harnesses large datasets, identifies relevant patterns, and provides quantitative predictions that can be used as clinical decision-making tools. In fact, supervised machine learning models have demonstrated efficacy in predicting various surgical outcomes, including patient survival, symptom improvement, and disease progression [7-9]. This efficacy is particularly evident when using datasets that integrate patient health care information and biological assessments. For instance, Jung et al [7] developed an individualized predictive model incorporating patient health care information from medical records as well as pre- and postoperative laboratory measurements (eg, total bilirubin and sodium levels) to anticipate graft failure in pediatric liver transplant surgery. Similarly, Chowdhury et al [8] applied biomarkers, such as mucus cytokines, along with patient demographics and clinical characteristics to machine learning algorithms to predict chronic rhinosinusitis improvement when treated with endoscopic sinus surgery. Kawakita et al [9] developed a model to predict delayed graft function with kidney transplantation surgery by integrating donor and recipient health care characteristics (eg, BMI, race, age, diabetes diagnosis) and clinical laboratory data (eg, blood urea nitrogen levels, serum glutamic pyruvic transaminase). Despite these successful examples of incorporating demographic, clinicopathological, therapeutic, and biomolecular data into personalized clinical decision-making tools aimed at predicting poor clinical outcomes, such methods have yet to gain acceptance in the current clinical practice of mesh surgery.

In this preliminary study, we explored the benefit of integrating distinct datasets when developing supervised machine learning models to predict mesh exposure in patients who have undergone POP repair surgery involving transvaginal insertion of polypropylene mesh. Models were trained with 3 different datasets: patient medical record data (such as demographics, social history, and medical history), biomaterial-induced blood cytokine levels, and the integration of both datasets. Models trained with these 3 different datasets were able to identify pertinent health risk factors as well as important predictors among the cytokines. Moreover, models trained with the integrated dataset relative to individual datasets demonstrated superior performance. These outcomes illustrate the potential of leveraging integrated datasets with machine learning models to accurately predict transvaginal mesh implantation outcomes and identify crucial risk factors associated with such outcomes. Such predictive models offer invaluable guidance to patients and surgeons. This approach holds potential to enhance the decision-making process for both transvaginal and other mesh surgeries, leading to improved patient care and better surgical outcomes.

Methods

Study Population

As described previously [10], recruited patients with surgical intervention for POP using transvaginal polypropylene mesh included both those with mesh exposure, a postsurgical complication where the implanted mesh protrudes through the vaginal wall, and those without. A cohort of 20 healthy, nonpregnant female patients, divided evenly between the 2 categories, was determined necessary for statistical observation of a difference in blood cytokine biomarkers with a 10% chance of error. Briefly, this statistical analysis used a 1-tailed t test based on a priori power analysis, a Cohen d of 1, and a conservative study power of 0.80.

XSL•FO
RenderX

Participants were selected retrospectively from a list of patients who had undergone surgical intervention for POP using polypropylene mesh at Prisma Health Greenville Memorial Hospital, as detailed in Waugh et al [10]. The patients were randomly identified and recruited via phone call until 10 patients with mesh exposure and 10 without were enrolled for this study. Participants with POP recurrence or those taking medications that could affect inflammatory responses were excluded from the study.

Ethical Considerations

The study protocol was approved by the Institutional Review Board of Prisma Health (number 1984257). Informed consent was waived due to the retrospective nature of the data. All collected data were deidentified and stored in a secure, password-protected folder with restricted access. As this was a retrospective study with no subject participation, no compensation was provided.

Patient Medical Record Data Collection

Medical record data were collected from 20 female patients who had experienced prior surgical intervention for POP using transvaginal polypropylene mesh, selected retrospectively and randomly as described earlier. The data categories include vital signs (systolic blood pressure [BP], diastolic BP, and pulse pressure), previous medical history (hypertension, diabetes, renal disease, hyperlipidemia, hemorrhoids, endometriosis, gastroesophageal reflux, and hysterectomy), social history (alcohol and tobacco usage, sexual activity, marital status, and education level), and other relevant health statistics (age at surgery, BMI, BMI change since surgery, POP stage diagnosis at surgery, and number of births). Mean and SD values were calculated for continuous variables (systolic and diastolic BP, pulse pressure, age at surgery, BMI, and BMI changes) along with the range of the variables (Table 1). Unpaired *t* tests were conducted on each variable to determine significance ($P<.05$).

XSL•FO
RenderX

Table . Demographics of the patients with pelvic organ prolapse (POP) study population (N=20 subjects).

Variables	Mesh exposure (n=10 subjects)	No mesh exposure (n=10 subjects)
Vital signs		
Systolic blood pressure		
Mean (SD)	127.4 (17.9)	123.3 (19.3)
Range	102 - 162	98 - 152
Diastolic blood pressure		
Mean (SD)	77.9 (11.3)	76.9 (10.7)
Range	58 - 98	60 - 92
Pulse pressure		
Mean (SD)	49.5 (21.3)	46.4 (11.9)
Range	24 - 104	
Previous medical history, n (%)		
Hypertension	4 (40)	5 (50)
Diabetes	0 (0)	2 (20)
Renal disease	2 (20)	5 (50)
Hyperlipidemia	3 (30)	3 (30)
Hemorrhoids	3 (30)	2 (20)
Endometriosis	3 (30)	4 (40)
Gastroesophageal reflux	5 (50)	6 (60)
Hysterectomy	8 (80)	8 (80)
Social history, n (%)		
Alcohol	6 (60)	2 (20)
Tobacco	2 (20)	1 (10)
Sexual activity	9 (90)	5 (50)
Marital status		
Single or lives alone	0 (0)	1 (10)
Married	8 (80)	6 (60)
Separated or divorced	1 (10)	1 (10)
Widowed	1 (10)	2 (20)
Highest education level		
Primary	0 (0)	1 (10)
Secondary	2 (20)	2 (20)
College or higher	8 (80)	7 (70)
Other relevant health statistics		
Age at surgery (years)		
Mean (SD)	56.8 (1.2)	59.4 (8.0)
Range	43 - 77	47 - 75
BMI		
Mean (SD)	29.3 (7.0)	28.8 (6.5)
Range	19.3-42.3	21.5 - 44.9
BMI change, mean (SD)	−1.6 (4.5)	−1.1 (3.5)
POP stage preoperation, n (%)		
1	1 (10)	1 (10)

XSL•FO
RenderX

Variables	Mesh exposure (n=10 subjects)	No mesh exposure (n=10 subjects)
2	4 (40)	2 (20)
3	4 (40)	7 (70)
4	1 (10)	0 (0)
Number of births, n (%)		
1 - 2	6 (60)	5 (50)
3+	4 (40)	5 (50)

Patient Blood Cytokine Measurements

Patient blood samples were collected from the same 20 female patients and incubated with sterile polypropylene mesh (2 cm × 2 cm), as detailed in Waugh et al [10]. Plasma layers collected following centrifugation (1500xg, 4 °C) of blood samples were analyzed via multiplex assay to quantify cytokine levels, as in Waugh et al [10]. These cytokines include interleukin (IL)-1α, IL-1β, IL-2, IL-4, IL-6, IL-8, IL-10, IL-12 p40, IL-12 p70, IL-17A, interferon-γ, tumor necrosis factor-α (TNF-α), and granulocyte-macrophage colony-stimulating factor. In total, 3 independent measurements were observed to detect the cytokine levels, with each sample evaluated in duplicate.

Data Analysis

Raw datasets, including blood cytokine levels and various medical record data, were analyzed using the statistical programming language R (RStudio, Inc). The imported data contained 60 observations (20 subjects x 3 independent cytokine measurements) and 35 total variable fields (21 medical record variables+13 cytokines+1 target variable). The target variable of exposure referred to the presence or absence of surgical mesh exposure through the vaginal wall following POP surgery. Univariate and multivariate analysis was implemented to explore the dataset, including identifying missing values, analyzing outliers, and standardizing categorical variables.

Supervised machine learning models were created using the caret package (version 6.0 - 90) in the R programming language. The 4 models trained were decision tree (DT), logistic regression (LR), Naïve Bayes (NB), and artificial neural network (ANN). This approach focused on three different datasets: (1) patient medical record data, (2) blood cytokine levels following incubation with surgical mesh, and (3) an integrated dataset of both medical record data and blood cytokine levels. Prior to creating the predictive models, the original data were split using an industry standard of 70% for training and 30% for testing. Each group contained an equal distribution of subjects who did or did not experience postsurgical mesh exposure through the vaginal wall, the prediction target for each model. Each model was trained using the 70% (42/60) subset and a cross-validation training control. A 10-fold cross-validation with 25% (15/60) left out replicated 3 times was used on each model to avoid bias and overfitting. From this, training accuracies are reported. Additional testing was performed for the prediction accuracy of each model using the 30% (18/60) test data. Prediction accuracies are reported along with sensitivity and specificity for the prediction of subjects to experience postsurgical mesh exposure. After computing predictive statistics for each dataset, analysis using the varImp function was conducted to identify which variables contributed most to the predictive outcome within each model. This analysis allowed for visualization of individual variable importance regarding each predictive model.

Results

Descriptive Analysis of Patient Medical Record Data

Health care data of patients with POP were collected from electronic medical records and analyzed in Table 1. The cohort included 20 patients with POP, among whom were the 10 that experienced vaginal mesh exposure and 10 that did not. Overall, no significant differences ($P<.05$) in systolic and diastolic BP, pulse pressure, age at surgery, BMI, and BMI changes were noted between patients with or without mesh exposure. Higher numbers of patients with no mesh exposure had previous medical diagnoses of diabetes and renal diseases. In contrast, higher numbers of patients with mesh exposure reported social histories of alcohol usage as well as sexual activity.

Predictive Analysis

Each of 4 supervised machine learning models, DT, NB, LR, and ANN, was implemented on 3 separate datasets: patient medical record data (21 variables), blood cytokine levels following incubation with surgical mesh (13 variables), and an integrated dataset of both medical data and blood cytokine levels (34 variables). The 5 most important factors that predict mesh exposure in each model are presented in Figure 1.

XSL•FO
RenderX

Figure 1. Four supervised machine learning models were used to identify important factors capable of yielding effective predictive analysis: decision tree (DT), Naïve Bayes (NB), logistic regression (LR), and artificial neural network (ANN). Each model was performed with patient medical record data (left), blood cytokine levels following incubation with surgical mesh (center), or an integrated dataset of both patient medical record data and blood cytokine levels (right). The importance of the 5 most important factors to predict mesh exposure in each model is indicated as a percentage. BP: blood pressure; GERD: gastroesophageal reflux disease; hx: history; IL: interleukin; TNF-α: tumor necrosis factor-α; GM-CSF: granulocyte-macrophage colony-stimulating factor.

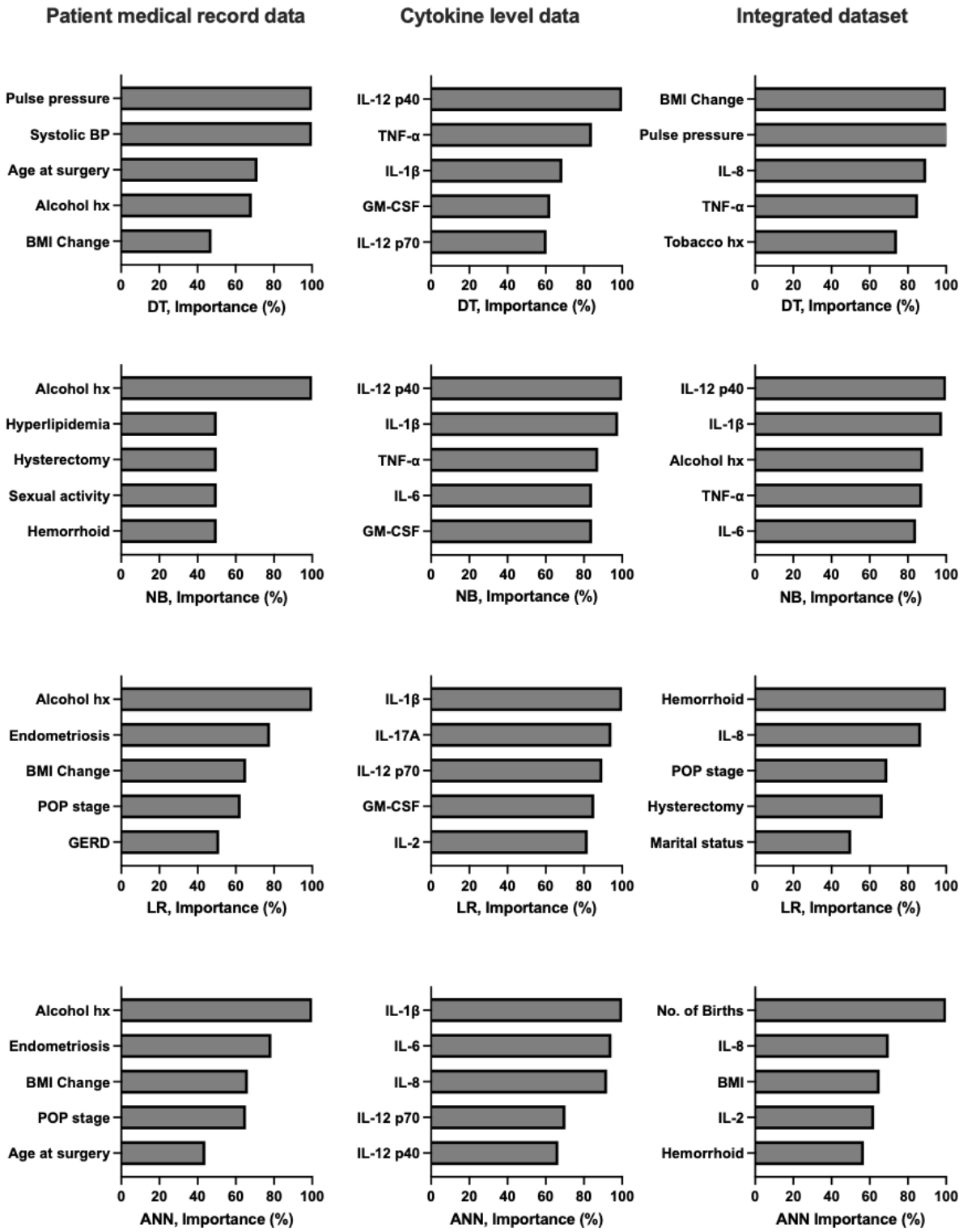

When the models were trained with patient medical record data alone, history of alcohol usage was the most important factor for NB, LR, and ANN models, and it also accounted for greater than 60% of the importance in the DT model. BMI change was among the 5 most relevant factors in 3 of the models, DT, LR, and ANN. In total, 3 additional factors were among the 5 most relevant factors in 2 of the 4 models: medical history of endometriosis (LR and ANN, >77% importance), POP stage (LR and ANN, >60% importance), and age at surgery (DT and ANN, >40% importance).

XSL•FO
RenderX

When models were trained with blood cytokine levels alone, IL-1β was the most important cytokine to predict mesh exposure in LR and ANN models and among the 5 most important cytokines in NB and DT models, while IL-12 p40 was the most important cytokine in DT and NB models and among the 5 most important cytokines in the ANN model. In total, 2 cytokines exhibited>60% importance in 3 models: granulocyte-macrophage colony-stimulating factor in DT, NB, and LR models and IL-12 p70 in DT, LR, and ANN models. In total, 2 cytokines exhibited>80% importance in 2 models: TNF-α in DT and NB models and IL-6 in NB and ANN models.

Finally, the predictive models were also performed on an integrated dataset of both medical record data and blood cytokine levels. IL-8 was among the 5 most important factors (>70% importance) in 3 models, DT, LR, and ANN. In total, 2 factors were among the 5 most important in 2 models: TNF-α in DT and NB models and hemorrhoids in LR and ANN models. Factors of highest importance in all 4 models included a combination of health care characteristics and blood cytokine levels. For the NB model, the 5 most important factors in the integrated dataset matched those identified to be of highest importance in the individual datasets. In the DT model, 3 factors identified to be of high importance in individual datasets were also of the highest importance in the integrated dataset, while 2 previously unidentified factors rose to high importance within the integrated dataset. In contrast, ANN and LR models identified four unique factors to be of highest importance within the integrated dataset.

Statistics were generated for the performance of each supervised machine learning model (Table 2). Each model achieved the highest prediction accuracy when using the integrated dataset, ranging from 78% (14/18) to 94% (17/18), compared to either medical record data or blood cytokine levels alone, which scored ranges of 33% (6/18) to 50% (9/18) and 50% (9/18) to 83% (15/18), respectively. Similarly, the highest training accuracy for each model was observed when using the integrated dataset, as opposed to either medical record data or blood cytokine levels alone. When comparing different models, both LR and ANN achieved the highest training and prediction accuracies with the integrated dataset. In fact, when the LR and ANN models were trained with the integrated dataset to predict mesh exposure, a training accuracy of 91% (38/42) and a prediction accuracy of 94% (17/18) were observed in combination with sensitivity and specificity ranging from 89% (16/18) to 100% (18/18).

Table . Summary of supervised machine learning model statistics. Patient medical record data (Med), blood cytokine levels following incubation with surgical mesh (Cyto), and an integrated dataset of both patient medical record data and blood cytokine levels (Int) were used to predict the presence or absence of pelvic organ prolapse (POP) postsurgical mesh exposure; 70% (42/60) of observations were used for training, and 30% (18/60) of observations were used for testing.

Model and dataset		Training accuracy, n (%)	Prediction accuracy, n (%)	95% CI	Sensitivity, n (%)	Specificity, n (%)	Prediction κ
Naïve Bayes							
	Med	16 (37)	9 (50)	0.118-0.882	0 (0)	18 (100)	0
	Cyto	26 (62)	11 (61)	0.357-0.827	6 (33)	16 (89)	0.222
	Int	31 (73)	16 (89)	0.653-0.986	14 (78)	18 (100)	0.778
Decision tree							
	Med	18 (43)	6 (33)	0.0433-0.777	12 (67)	0 (0)	0.333
	Cyto	27 (64)	11 (61)	0.57-0.827	6 (33)	16 (89)	0.222
	Int	28 (67)	14 (78)	0.524-0.936	12 (67)	16 (89)	0.556
Logistic regression							
	Med	11 (27)	6 (33)	0.043-0.777	6 (33)	6 (33)	0.333
	Cyto	31 (73)	9 (50)	0.26-0.740	10 (56)	8 (44)	0.000
	Int	38 (90)	17 (94)	0.727-0.999	18 (100)	16 (89)	0.889
Artificial neural network							
	Med	17 (40)	9 (50)	0.118-0.882	6 (33)	12 (67)	0.000
	Cyto	33 (79)	15 (83)	0.586-0.964	14 (78)	16 (89)	0.667
	Int	38 (90)	17 (94)	0.727-0.999	16 (89)	18 (100)	0.889

Discussion

Summary

Machine learning is gaining acceptance as a tool for predicting medical conditions, including treatment prognosis, disease progression, and surgical outcomes [7,9,11,12]. In particular, literature reports have illustrated the utility of predictive models that incorporate both health care data and biomarkers [13,14]. These successes lay the groundwork for a pioneering method that integrates patient medical records and biomarker measurements for predicting mesh-related postsurgical outcomes. Patients with POP, in particular, could benefit from such a personalized approach. These patients currently make surgical decisions based solely on risk factors derived from

XSL•FO
RenderX

demographic statistics of the general population of surgical patients with POP, while individual medical history and personalized responses to the biomaterial used in the surgery could better inform their decision. Our previous study demonstrated the efficacy of predictive models that use biomaterial-induced blood cytokine-expression profiles to accurately predict mesh exposure following surgical intervention for POP using transvaginal polypropylene mesh [10]. In this exploratory study, we examined the benefit of combining continuous and categorical data from health care records and biomarker measurements to build supervised machine learning models and develop more inclusive and efficient predictive models with a higher rate of prediction accuracy, rendering such models more favorable as a novel, noninvasive, personalized clinical decision tool.

Principal Findings and Comparison to Prior Work

Comparative statistical analysis of medical record data of patients with POP with transvaginal mesh is illustrated in Table 1. Even though this study had a small population of 20 patients, the median outcomes in mesh-exposure and nonexposure groups were comparable to those reported in other studies with a larger study cohort [15,16]. The current dataset revealed that sexually active patients were more prone to experiencing mesh exposure, a trend similarly observed by Kaufman et al [4]. This observation could be attributed to the mechanical challenges posed by sexual activity, potentially leading to friction-induced issues with the mesh implant [17]. In contrast, within the current dataset, variables such as BP, medical diagnoses, education levels, BMI, age at surgery, and parity numbers were not significant risk factors when comparing patients with or without mesh exposure. Similarly, Long et al [18] found no disparity in BMI between the 2 groups. Additionally, Chavez et al [16] noted that risk factors including obesity, diabetes, and smoking status were not associated with vaginal mesh exposure, aligning with our result.

While significant associations were not evident within many individual health care data categories, implementing supervised machine learning models on such data both reinforced groups identified via comparative statistics and revealed new groups associated with patients who experienced mesh exposure. For instance, the history of alcohol usage, indicated as important via comparative statistics, was also considered the most important factor in NB, LR, and ANN models, while BP variables were newly identified as important in the DT model. From a clinical perspective, alcohol use could contribute to malnutrition and inflammation, which may affect the healing process and influence the outcomes of mesh surgery [19]. Similarly, changes in BP and pulse pressure could indicate cardiovascular or systemic issues that may impact surgical planning or postoperative recovery [20]. However, using supervised machine learning models on patient medical record data alone yielded the lowest prediction accuracy rate when compared to models trained with either cytokine levels or an integrated dataset (Table 2). Similarly, low predictive power was reported in Taneja et al [21] when health care data alone was employed in their predictive model [21]. Thus, relying solely on medical record data to construct predictive models does not provide a comprehensive representation. Given an individual's varying inflammatory response [10,22], as well as their race and genetic components [23,24], merely identifying commonalities among medical record data may not accurately represent predictors of the mesh exposure postsurgery.

Previous studies have indicated a relationship between elevated levels of pro-inflammatory cytokines and postsurgical complications [25-27]. When creating predictive models trained with 13 biomaterial-induced cytokines known to influence wound healing, the models predicted with a higher prediction accuracy of 50% (9/18) to 83% (15/18) compared to models trained with patient medical record data (Table 2). These supervised machine learning models demonstrated that IL-1β was one of the most important cytokines to predict mesh exposure across all four models. Elevated levels of IL-1β have been associated with various diseases, including obesity, cardiovascular diseases, cancer, and periodontitis [28-30]. Moreover, IL-1β was increased in a murine model following hernia polypropylene mesh implantation [31], mirroring our findings. Additionally, IL-12 p40 played a pivotal role in DT and NB models. This pro-inflammatory cytokine, which serves as a critical link between the innate and adaptive immune response, is observed in inflammatory conditions, such as pulmonary sarcoidosis and inflammatory bowel diseases [32-34].

All 4 supervised machine learning models achieved their highest prediction accuracies when trained with an integrated dataset comprising both patient medical record data and biomaterial-induced blood cytokine levels. This integrated dataset proved particularly effective when used in LR and ANN models, achieving 94% (17/18) prediction accuracy. The approach also exceeds predictive health care models for other surgical outcomes that used only cytokine-based [35] or electronic health record-based approaches [36]. In parallel, Taneja et al [21] demonstrated that their sepsis predictive models had higher predictive power when combined data were used, rather than either electronic medical record or biomarker data alone. Interestingly, this integrated approach also identifies new factors of high importance. In contrast to the models trained on cytokine expression alone, models trained with integrated data observed that IL-8 was among the important predictors in DT, LR, and ANN models, and TNF-α in DT and NB models. IL-8 and TNF-α are 2 prominent pro-inflammatory cytokines elevated in patients postsurgery, including mesh implantation and pelvic surgery [37,38]. In contrast to the models trained on medical record data alone, hemorrhoids were considered one of the important factors in LR and ANN models trained with the integrated dataset. Models trained with the integrated dataset also outperformed those trained with either dataset alone in terms of training accuracy, sensitivity, and specificity across all four models (Table 2).

Limitations and Future Directions

Despite employing rigorous research methodologies, it is imperative to acknowledge certain limitations inherent in this study. As a pilot study, the population size was limited to 20 patients within a single hospital system. Even though it was on a small scale, this study yielded a prediction accuracy as high as 94% (17/18), surpassing comparable predictive models that ranged from 71% to 86% in prediction accuracy across a

XSL•FO
RenderX

population size of 198-467 [35,39], thus affirming the validity of our approach. Future research will involve a larger participant pool drawn from multiple hospitals to further confirm these results. To mitigate potential confounders, patients with POP recurrence or taking medication that would alter inflammatory response were excluded. Future studies with a larger population will also enable matching participants with respect to these and other potentially confounding variables. In the analysis of the integrated dataset, medical record data were repeated 3 times to match the structure of the cytokine dataset, which contained 3 repeated measurements. This approach was employed to balance the contribution of both data types. This may represent to a potential limitation, although the resulting variable importance graphs for the integrated dataset demonstrated a relatively equal contribution from both the cytokine and medical record data, mitigating concerns that the repetition of the medical record data led to overfitting. Despite these limitations, the results of this pilot study highlight the importance of integrating inflammatory biomarkers with medical record data in the prediction of postsurgical mesh exposure.

Conclusions

Supervised machine learning models demonstrated significantly higher prediction accuracy when trained with an integrated dataset comprising both patient medical record data and biomaterial-induced blood cytokine levels, surpassing the performance of models trained using either dataset alone. These results demonstrate for the first time the advantage of an approach integrating medical record data and blood biomarkers. This innovative approach to predicting surgical outcomes following POP repair via transvaginal insertion of polypropylene mesh holds promise for both patients and surgeons. Additionally, such a method could be applied to the study of other mesh-related surgeries, such as those used to repair POP transabdominally or to treat hernia and stress urinary incontinence. By presenting a more personalized decision-making tool, this approach offers a more accurate depiction of potential outcomes, thereby enhancing informed decision-making and improving patient care.

Acknowledgments

This work was supported by a Transformative Research Grant from Greenville Health System Health Sciences Center to MAM, TLW II, RH, and RLG, by a South Carolina IDeA Networks of Biomedical Research Excellence (INBRE) Student Initiated Research Project (SIRP) to MLW, and by a South Carolina Research Authority Academic Matching Grant to MAM and NB. The authors wish to extend their sincere gratitude to Dr. Matthew Hermes, who contributed to the research insights.

Data Availability

The datasets generated or analyzed during this study are openly available in a public repository [40].

Conflicts of Interest

None declared.

References

1. Hullfish KL, Bovbjerg VE, Gurka MJ, Steers WD. Surgical versus nonsurgical treatment of women with pelvic floor dysfunction: patient centered goals at 1 year. J Urol 2008 Jun;179(6):2280-2285. [doi: 10.1016/j.juro.2008.01.147] [Medline: 18423762]
2. Seifalian A, Basma Z, Digesu A, Khullar V. Polypropylene pelvic mesh: what went wrong and what will be of the future? Biomedicines 2023 Mar 1;11(3):741. [doi: 10.3390/biomedicines11030741] [Medline: 36979721]
3. Syed KK, Consolo MJ, Gousse AE, Repair A. Anterior vaginal wall prolapse repair and the rise and fall of transvaginal mesh. Did we come full circle? A historical perspective. Urology 2021 Apr;150:110-115. [doi: 10.1016/j.urology.2020.08.015] [Medline: 32827535]
4. Kaufman Y, Singh SS, Alturki H, Lam A. Age and sexual activity are risk factors for mesh exposure following transvaginal mesh repair. Int Urogynecol J 2011 Mar;22(3):307-313. [doi: 10.1007/s00192-010-1270-6] [Medline: 20882269]
5. Akyol A, Akca A, Ulker V, et al. Additional surgical risk factors and patient characteristics for mesh erosion after abdominal sacrocolpopexy. J Obstet Gynaecol Res 2014 May;40(5):1368-1374. [doi: 10.1111/jog.12363] [Medline: 24754851]
6. Artsen AM, Liang R, Meyn L, Bradley MS, Moalli PA. Dysregulated wound healing in the pathogenesis of urogynecologic mesh complications. Sci Rep 2023 Dec 5;13(1):21437. [doi: 10.1038/s41598-023-48388-8] [Medline: 38052928]
7. Jung S, Park K, Ihn K, et al. Predicting graft failure in pediatric liver transplantation based on early biomarkers using machine learning models. Sci Rep 2022 Dec 27;12(1):22411. [doi: 10.1038/s41598-022-25900-0] [Medline: 36575218]
8. Chowdhury NI, Li P, Chandra RK, Turner JH. Baseline mucus cytokines predict 22-item sino-nasal outcome test results after endoscopic sinus surgery. Int Forum Allergy Rhinol 2020 Jan;10(1):15-22. [doi: 10.1002/alr.22449] [Medline: 31645085]
9. Kawakita S, Beaumont JL, Jucaud V, Everly MJ. Personalized prediction of delayed graft function for recipients of deceased donor kidney transplants with machine learning. Sci Rep 2020 Oct 27;10(1):18409. [doi: 10.1038/s41598-020-75473-z] [Medline: 33110142]

XSL•FO
RenderX

10. Waugh ML, Boltin N, Wolf L, et al. Prediction of pelvic organ prolapse postsurgical outcome using biomaterial-induced blood cytokine levels: machine learning approach. JMIR Perioper Med 2023 May 31;6:e40402. [doi: 10.2196/40402] [Medline: 37256676]
11. Li B, Shaikh F, Zamzam A, Syed MH, Abdin R, Qadura M. A machine learning algorithm for peripheral artery disease prognosis using biomarker data. iScience 2024 Mar;27(3):109081. [doi: 10.1016/j.isci.2024.109081]
12. Thorwarth RM, Scott DW, Lal D, Marino MJ. Machine learning of biomarkers and clinical observation to predict eosinophilic chronic rhinosinusitis: a pilot study. Int Forum Allergy Rhinol 2021 Jan;11(1):8-15. [doi: 10.1002/alr.22632] [Medline: 32472743]
13. Reddy BK, Delen D, Agrawal RK. Predicting and explaining inflammation in Crohn's disease patients using predictive analytics methods and electronic medical record data. Health Informatics J 2019 Dec;25(4):1201-1218. [doi: 10.1177/1460458217751015] [Medline: 29320910]
14. Ryan CT, Zeng Z, Chatterjee S, et al. Machine learning for dynamic and early prediction of acute kidney injury after cardiac surgery. J Thorac Cardiovasc Surg 2023 Dec;166(6):e551-e564. [doi: 10.1016/j.jtcvs.2022.09.045] [Medline: 36347651]
15. Niu K, Lu YX, Shen WJ, Zhang YH, Wang WY. Risk factors for mesh exposure after transvaginal mesh surgery. Chin Med J (Engl) 2016 Aug 5;129(15):1795-1799. [doi: 10.4103/0366-6999.186631] [Medline: 27453227]
16. El-Khawand D, Wehbe SA, O'Hare PGI, Arunachalam D, Vakili B. Risk factors for vaginal mesh exposure after mesh-augmented anterior repair. Female Pelvic Med Reconstr Surg 2014;20(6):305-309. [doi: 10.1097/SPV.0000000000000095]
17. Kowalik CR, Lakeman MME, de Kraker AT, Roovers JPWR. Effects of mesh-related complications in vaginal surgery on quality of life. Int Urogynecol J 2019 Jul;30(7):1083-1089. [doi: 10.1007/s00192-018-3680-9] [Medline: 29909555]
18. Long CY, Lo TS, Wang CL, Wu CH, Liu CM, Su JH. Risk factors of surgical failure following transvaginal mesh repair for the treatment of pelvic organ prolapse. Eur J Obstet Gynecol Reprod Biol 2012 Apr;161(2):224-227. [doi: 10.1016/j.ejogrb.2012.01.009] [Medline: 22326614]
19. Daher GS, Choi KY, Wells JW, Goyal N. A systematic review of oral nutritional supplement and wound healing. Ann Otol Rhinol Laryngol 2022 Dec;131(12):1358-1368. [doi: 10.1177/00034894211069437] [Medline: 35043693]
20. Mogoanta SS, Paitici S, Mogoanta CA. Postoperative follow-up and recovery after abdominal surgery. In: Zaghal A, Rifai AE, editors. Abdom Surg: IntechOpen; 2021. [doi: 10.5772/intechopen.97739]
21. Taneja I, Reddy B, Damhorst G, et al. Combining biomarkers with EMR data to identify patients in different phases of sepsis. Sci Rep 2017 Sep 7;7(1):10800. [doi: 10.1038/s41598-017-09766-1] [Medline: 28883645]
22. Pos Z, Selleri S, Spivey TL, et al. Genomic scale analysis of racial impact on response to IFN-alpha. Proc Natl Acad Sci U S A 2010 Jan 12;107(2):803-808. [doi: 10.1073/pnas.0913491107] [Medline: 20080756]
23. Hoffmann SC, Stanley EM, Cox ED, et al. Ethnicity greatly influences cytokine gene polymorphism distribution. Am J Transplant 2002 Jul;2(6):560-567. [doi: 10.1034/j.1600-6143.2002.20611.x] [Medline: 12118901]
24. Gökçe İ. Effects of human genetic factors (ethnicity and race) on clinical severity of SARS-CoV-2 (COVID-19). JEB Med Sci 2021;1(3):147-158. [doi: 10.5606/jebms.2020.75631]
25. Mathe Z, Serban RC, Pintilie I, Somkereki C, Hutanu A, Scridon A. Postoperative interleukin-8 levels are related to the duration of coronary artery bypass grafting surgery and predict in-hospital postsurgical complications. Rev Rom Med Lab 2018 Jul 1;26(3):293-303. [doi: 10.2478/rrlm-2018-0021]
26. Paruk F, Chausse JM. Monitoring the post surgery inflammatory host response. J Emerg Crit Care Med 2019;3:47-47. [doi: 10.21037/jeccm.2019.08.06]
27. Tsilimigras DI, Sigala F, Karaolanis G, et al. Cytokines as biomarkers of inflammatory response after open versus endovascular repair of abdominal aortic aneurysms: a systematic review. Acta Pharmacol Sin 2018 Jul;39(7):1164-1175. [doi: 10.1038/aps.2017.212] [Medline: 29770795]
28. Khafagy R, Dash S. Obesity and cardiovascular disease: the emerging role of inflammation. Front Cardiovasc Med 2021;8:768119. [doi: 10.3389/fcvm.2021.768119] [Medline: 34760952]
29. Pani P, Tsilioni I, McGlennen R, et al. IL-1B(3954) polymorphism and red complex bacteria increase IL-1β (GCF) levels in periodontitis. J Periodontal Res 2021 Jun;56(3):501-511. [doi: 10.1111/jre.12850] [Medline: 33638191]
30. Tulotta C, Ottewell P. The role of IL-1B in breast cancer bone metastasis. Endocr Relat Cancer 2018 Jul;25(7):R421-R434. [doi: 10.1530/ERC-17-0309] [Medline: 29760166]
31. Heymann F, von Trotha KT, Preisinger C, et al. Polypropylene mesh implantation for hernia repair causes myeloid cell-driven persistent inflammation. JCI Insight 2019 Jan 24;4(2):e123862. [doi: 10.1172/jci.insight.123862] [Medline: 30674727]
32. Abdi K. IL-12: the role of p40 versus p75. Scand J Immunol 2002 Jul;56(1):1-11. [doi: 10.1046/j.1365-3083.2002.01101.x] [Medline: 12100467]
33. Shigehara K, Shijubo N, Ohmichi M, et al. Increased circulating interleukin-12 (IL-12) p40 in pulmonary sarcoidosis. Clin Exp Immunol 2003 Mar 24;132(1):152-157. [doi: 10.1046/j.1365-2249.2003.02105.x]
34. Verstockt B, Salas A, Sands BE, et al. IL-12 and IL-23 pathway inhibition in inflammatory bowel disease. Nat Rev Gastroenterol Hepatol 2023 Jul;20(7):433-446. [doi: 10.1038/s41575-023-00768-1]

XSL•FO
RenderX

35. Chen Z, Chen L, Yao G, Yang W, Yang K, Xiong C. Novel blood cytokine-based model for predicting severe acute kidney injury and poor outcomes after cardiac surgery. J Am Heart Assoc 2020 Nov 17;9(22):e018004. [doi: 10.1161/JAHA.120.018004] [Medline: 33131359]
36. Weller GB, Lovely J, Larson DW, Earnshaw BA, Huebner M. Leveraging electronic health records for predictive modeling of post-surgical complications. Stat Methods Med Res 2018 Nov;27(11):3271-3285. [doi: 10.1177/0962280217696115] [Medline: 29298612]
37. Parpoudi S, Mantzoros I, Gkiouliava A, et al. Effect of N-acetyl-L-cysteine on inflammation after intraperitoneal mesh placement in a potentially contaminated environment: an experimental study in the rat. Asian J Surg 2022 Nov;45(11):2191-2196. [doi: 10.1016/j.asjsur.2021.11.001] [Medline: 34801356]
38. Shi JY, Bergerat-Thompson AM, Mitchell CM. The impact of the vaginal microbiome on vaginal epithelial wound healing. Research Square. Preprint posted online on Aug 25, 2023. [doi: 10.21203/rs.3.rs-3282980/v1]
39. Chu CS, Lee NP, Adeoye J, Thomson P, Choi SW. Machine learning and treatment outcome prediction for oral cancer. J Oral Pathol Med 2020 Nov;49(10):977-985. [doi: 10.1111/jop.13089] [Medline: 32740951]
40. IntegratedDataset. GitHub. URL: https://github.com/tomhjlim/IntegratedDataset [accessed 2025-03-17]

Abbreviations

ANN: artificial neural network
BP: blood pressure
DT: decision tree
IL: interleukin
LR: logistic regression
NB: Naïve Bayes
POP: pelvic organ prolapse
TNF-α: tumor necrosis factor alpha

Edited by A Mavragani; submitted 12.06.24; peer-reviewed by B Zhao, MF Kinci; revised version received 22.01.25; accepted 23.01.25; published 01.05.25.

Please cite as:
Waugh ML, Mills T, Boltin N, Wolf L, Parker P, Horner R, Wheeler II TL, Goodwin RL, Moss MA
Predicting Transvaginal Surgical Mesh Exposure Outcomes Using an Integrated Dataset of Blood Cytokine Levels and Medical Record Data: Machine Learning Approach
JMIR Form Res 2025;9:e59631
URL: https://formative.jmir.org/2025/1/e59631
doi:10.2196/59631

©Mihyun Lim Waugh, Tyler Mills, Nicholas Boltin, Lauren Wolf, Patti Parker, Ronnie Horner, Thomas L Wheeler II, Richard L Goodwin, Melissa A Moss. Originally published in JMIR Formative Research (https://formative.jmir.org), 1.5.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Integration of a Mental Health App (e-MICHI) Into a Blended Treatment of Depression in Adolescents: Single-Group, Naturalistic Feasibility Trial

Lena Lincke[1,2*], MSc; Tim Martin-Döring[1,2*], MSc; Andrea Daunke[1,2], Dipl-Psych; Antonia Sadkowiak[1]; Daria Alexandra Nolkemper[3,4], MA; Nina Sproeber-Kolb[5], Prof Dr; Stefanie Bienioschek[3,4,6], Dr Med; Olaf Reis[1,2], PhD, PD; Michael Kölch[1,2], Prof Dr Med

[1]Department of Child and Adolescent Psychiatry, Neurology, Psychosomatics and Psychotherapy, Rostock University Medical Center, Gehlsheimer Str. 20, Rostock, Germany

[2]Partner Site Greifswald/Rostock, German Center for Child and Adolescent Health (DZKJ), Rostock, Germany

[3]Brandenburg Medical School Theodor Fontane (MHB), Neuruppin, Germany

[4]Department of Child and Adolescent Psychiatry, Psychotherapy and Psychosomatics, University Hospital Ruppin-Brandenburg (ukrb), Neuruppin, Germany

[5]School of Social Work, Baden-Wuerttemberg Cooperative State University, Stuttgart, Germany

[6]Faculty of Health Sciences, Joint Faculty of the University of Potsdam, the Brandenburg University of Technology Cottbus-Senftenberg and the Brandenburg Medical School Theodor Fontane, Potsdam, Germany

[*]these authors contributed equally

Corresponding Author:
Tim Martin-Döring, MSc
Department of Child and Adolescent Psychiatry, Neurology, Psychosomatics and Psychotherapy, Rostock University Medical Center, Gehlsheimer Str. 20, Rostock, Germany

Abstract

Background: Major depressive disorder is a common diagnosis among adolescents. Cognitive behavioral therapy is recommended as the first line of treatment. Digital health interventions, such as apps, could contribute to treatment. Advantages could be easy accessibility and availability, reduced time for face-to-face therapy, and the ability to intensify therapy by incorporating it into the patients' everyday lives. Challenges such as low adherence rates are common in digital health interventions. Therefore, they need to undergo rigorous testing for feasibility and effectiveness.

Objective: An evaluated, cognitive behavioral therapy–based face-to-face therapy program for depression in adolescents was transformed into an app called e-MICHI. This study examined its feasibility and efficacy for use in blended therapy in outpatient settings.

Methods: Adolescents aged 12 to 18 years with major depressive disorder receiving outpatient care were recruited from 2 university hospitals (n=36 included in analysis). The e-MICHI intervention combined daily app engagement over 6 weeks with 3 face-to-face sessions with a therapist. Feasibility was measured using various variables, including an adherence score (0=no or little patient engagement to 3=excellent engagement) and engagement rates (number of modules completed, number of messages sent by participants via the in-app messenger), satisfaction ratings from both participants and therapists, as well as participants' ratings of the usefulness of the antidepressant strategies covered in the app and the transfer of these strategies to everyday practice. Trends of efficacy were evaluated from multiple perspectives (participant self-rating, independent rater, or therapist), using the Beck Depression Inventory-II, the Children's Depression Rating Scale-Revised, and the Clinical Global Impressions-Severity Scale. Feasibility metrics were assessed by analyzing their central tendency and dispersion, efficacy data were analyzed using a repeated measures ANOVA.

Results: e-MICHI was positively evaluated by both participants and therapists (participants: mean 7.3, SD 1.2 and therapists: mean 7.3, SD 1.1, on a scale from 0=bad to 10=excellent). Participants demonstrated high adherence rates (nearly 80%, n=25, received a "good" or "excellent" adherence score) and showed overall good engagement (app modules completed [maximum 6]: mean 5.03, SD 1.27 and messages sent via messenger: mean 23, SD 22.1). Participants rated the psychoeducational content of the app as particularly useful and reported consistent practice of the e-MICHI strategies in everyday life. Use of the app was associated with a significant reduction of depressive symptoms (before app use vs 3-month follow-up, Beck Depression Inventory-II: mean −6.76, SD 11.49, P=.01; Children's Depression Rating Scale-Revised: mean −16.45, SD 16.76, P<.001; Clinical Global Impressions-Severity Scale: mean −1.1, SD 1.24, P<.001).

XSL•FO
RenderX

Conclusions: While acknowledging its limitations, such as the small number of participants and the limited validity concerning efficacy, this study confirms the feasibility of e-MICHI for treating adolescent depression in outpatient settings.

Trial Registration: German Clinical Trial Register DRKS00022420; https://drks.de/search/en/trial/DRKS00022420/details

(JMIR Form Res 2025;9:e58427) doi:10.2196/58427

KEYWORDS

adolescents; major depressive disorder (MDD); cognitive behavioral therapy (CBT); digital health interventions (DHIs); feasibility; adherence; outpatient setting; mobile application; accessibility; availability; efficacy; mobile phone

Introduction

Major depressive disorder (MDD) is a prevalent psychiatric diagnosis among adolescents and the primary cause of inpatient treatment of adolescents in Germany [1,2]. Cognitive behavioral therapy (CBT) is the first-line treatment [3,4].

Digital health interventions (DHIs) could contribute to facilitated access to psychotherapeutic treatment [5]. Adolescents and young adults show the greatest openness and acceptance toward mental e-health interventions [6]. However, there is a scarcity of e-health interventions for depression in adolescents. Furukawa et al [7] found that all forms of internet-based CBT were superior to a waiting list control condition. Similarly, meta-analytical evidence indicates that DHIs are moderately to considerably effective (standardized mean difference 0.16 to 0.76) when compared to primarily inactive control groups [8]. Numerous studies have reported high dropout rates and low adherence levels for nonguided e-health interventions [9,10]. DHIs that include some form of personalized support, for example, from psychotherapeutic professionals, seem to be more efficacious due to higher adherence rates [11]. In addition to adherence, transferability presents a further challenge. The transfer of the contents and skills learned in e-health interventions to everyday practice by users seems to be poor: within the SPARX-R program, an intervention for adolescent depression, only around 40% of users practiced few, if any, of the techniques taught in the program [12]. Incorporating engaging elements such as text, audiovisual aids, and gamification, alongside the real-world implementation and involvement of parents and professionals could enhance this transfer [13].

There is an increasing availability of various DHIs for adolescents [14]. Due to the lack of scientific evaluation, Christ et al [15] called for further research and long-term follow-up assessments of these interventions. Furthermore, Seiferth et al [13] established guidelines for evidence-based DHIs including individuals with suicidal thoughts, emphasizing robustness, comparability, crisis planning, and access to additional treatments. This study aimed to assess the feasibility and indicators for efficacy of a blended DHI application, based on an evaluated face-to-face therapy program for depression in adolescents. The target population was adolescents with clinically verified depressive symptoms who were receiving child and adolescent psychiatric and psychotherapeutic care.

Methods

Procedure

This study used a single-group design to investigate the feasibility of a 6-week blended DHI for adolescents with depression. To this end, treatment adherence and the effects on symptom severity during the intervention period and a 90-day follow-up were analyzed.

This study was conducted at the Departments of Child & Adolescent Psychiatry at Rostock University Medical Center and University Hospital Ruppin-Brandenburg, in association with Brandenburg Medical School, Neuruppin. Adolescents aged between 12 and 18 years who met the criteria for a MDD, according to the *ICD-10* (*International Statistical Classification of Diseases and Related Health Problems, 10th Revision*) [16] as diagnosed by a certified child and adolescent psychiatrist or psychotherapist and were receiving outpatient care were included in the trial. Patients who met these inclusion criteria were informed about this study and asked whether they wanted to participate. If they and their parents agreed to participate, MDD diagnoses were confirmed by independent raters using the Beck Depression Inventory-II (BDI-II) [17] and the Children's Depression Rating Scale-Revised (CDRS-R) [18]. Additionally, a comprehensive diagnostic assessment was performed (Table 1).

XSL•FO
RenderX

Table . Overview of assessments conducted at different time points in a feasibility trial integrating the e-MICHI mental health app into a blended treatment of adolescent depression.

	Screening or recruitment	Pretreatment (T1)	Interim (T2)	Posttreatment (T3)	Follow-up (T4)
Trial days	–7	1	21	42	T3 +90
Diagnosis of depressive disorder	✓	N/A[a]	N/A	N/A	N/A
Inclusion and exclusion criteria	✓	N/A	N/A	N/A	N/A
Patient information and consent	✓	N/A	N/A	N/A	N/A
Demographics	N/A	✓	N/A	N/A	N/A
Previous and concomitant medication	N/A	✓	N/A	N/A	N/A
Concomitant diseases	N/A	✓	N/A	N/A	N/A
Treatment expectancy	N/A	✓	N/A	N/A	N/A
Adherence score	N/A	N/A	N/A	N/A	✓
Satisfaction and usability[b]	N/A	N/A	N/A	✓	N/A
BDI-II[c]	N/A	✓	✓	✓	✓
CDRS-R[d]	N/A	✓	N/A	✓	✓
CGI-S[e]	N/A	✓	✓	✓	✓
Adverse events	N/A	N/A	✓	✓	✓

[a]N/A: not applicable.

[b]Once per participant and once per therapist.

[c]BDI-II: Beck Depression Inventory-II [17].

[d]CDRS-R: Children's Depression Rating Scale-Revised (clinician rating) [18].

[e]CGI-S: Clinical Global Impressions-Severity Scale [19].

Additional psychopharmacological antidepressant treatment was allowed, and treatment as usual was provided as needed, to ensure that this study's population was representative of "real world" patients in routine care. Exclusion criteria for participation in the trial were acute suicidality requiring acute inpatient treatment, severe other psychiatric diagnoses such as schizophrenia and bipolar-disorder, and IQ below 85 or inadequate German language skills, as this would hinder the use of the program. Assessment for these criteria was based on clinical judgement following the Kinder-DIPS diagnostic interview for mental disorders [20].

Intervention

The e-therapy was delivered through a newly developed app called "e-MICHI" (Figure 1). e-MICHI is based on the MICHI therapy program ("Manualized Intervention to Cope With Depressive Symptoms, Help Strengthen Resources and Improve Emotion Regulation") [21], which has been shown to be effective for the treatment of depression in adolescents [22]. MICHI incorporates the proven effective elements of CBT for depression: psychoeducation, behavioral activation, cognitive restructuring, emotion regulation, enhancement of self-esteem, problem-solving, and coping with sudden crises [23]. The e-MICHI app consisted of 6 modules containing psychoeducational texts, exercises, quizzes, and a total of 27 video clips (Table 2). Participants were encouraged to complete each module within 1 week. In addition to the modules, e-MICHI included other features such as a mood diary and an in-app messenger. The messenger enabled communication between participant and therapist via text message. Once a participant had completed a module, the therapist received an automatic notification via messenger (eg, "I completed lesson 1!") and subsequently contacted the participant to address any queries or challenges related to the module. Participants could contact their therapist via messenger at any time if they had difficulty understanding the content of the modules or if they experienced a decrease in depressive symptoms. The app was hosted by Zone 35, a digital solutions provider [24].

XSL•FO
RenderX

Figure 1. Screenshots of the e-MICHI app interface, illustrating the home screen (providing an overview of progress and tasks) and the module start screen (introducing structured activities and content).

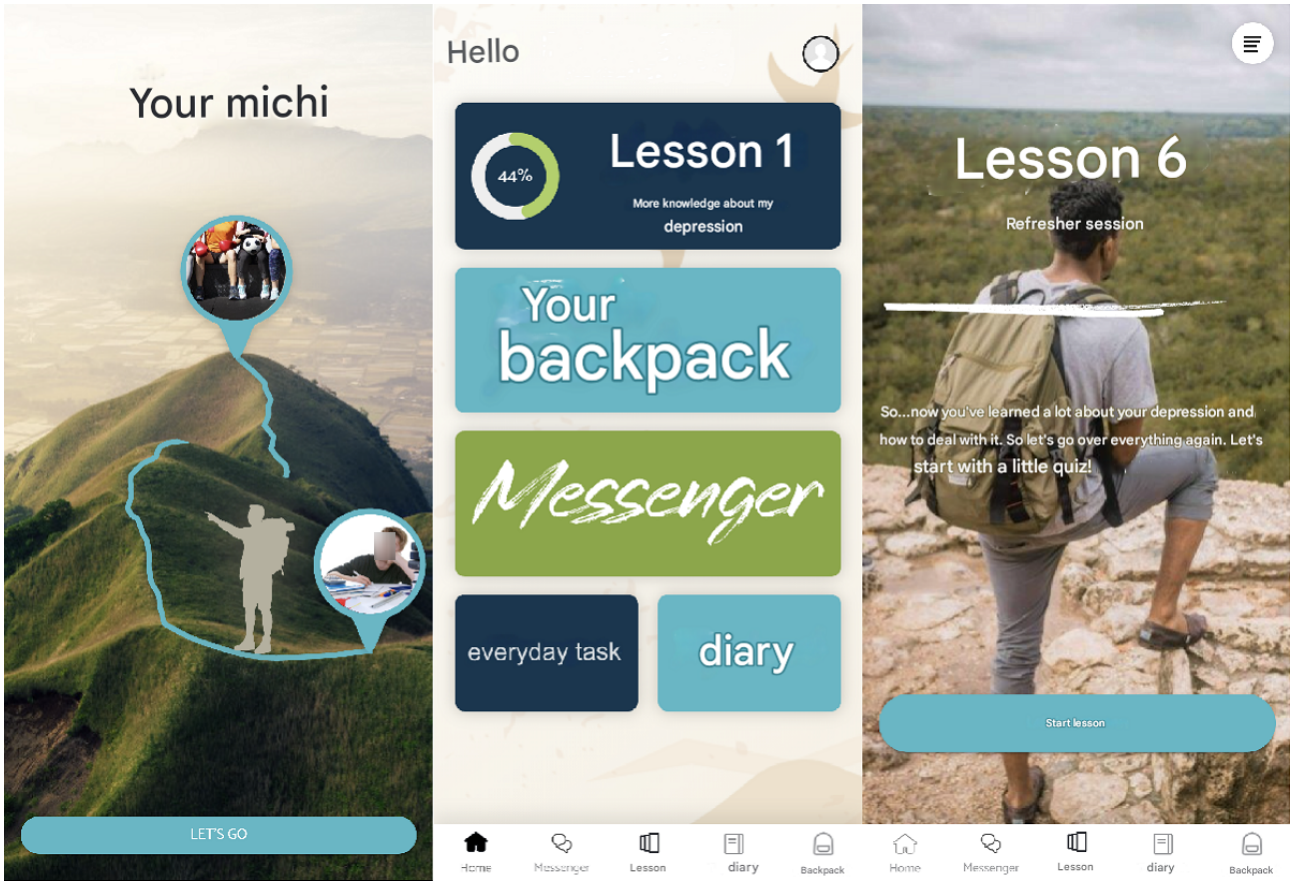

Table . Detailed summary of e-MICHI modules included in the feasibility trial. Each module's title, focus area, and corresponding therapeutic techniques are outlined.

	Content and psychoeducation	Exercises	Homework
1: "Knowledge of depression"	· Symptoms · Suicidality assessment	· Note personal takeaway[a] · Three quiz questions[a]	· Mood diary[b]
2: "Feelings"	· Depression causation · Mood improvement	· Personal disease model · Select pleasant activities · Identify strengths · Mind journey exercise	· Feedback on strengths · Pleasant activities · Mood diary[b] (including activities) · Feedback on strengths
3: "Influence on feelings"	· Positive self-esteem · Thoughts-behavior-feelings link	· Note compliments · Record thoughts, behaviors, or feelings in situations	· Mood diary[b] (behavior, thoughts, or feelings) · Practice complimenting
4: "Change thoughts"	· Crisis management · Cognition restructuring	· Note advice, link strategies to problems · Differentiate thoughts	· Personal emergency plan · Reflect on strategies with a therapist
5: "Support and health"	· Problem-solving skills	· Problem-solving exercise	· Appoint a "person of trust" · Share learned strategies
6: Refresher	· Quiz-based repetition	· Quiz on learned content	· No specific homework

[a]Repeats for lessons 1 - 5.

[b]7 days, 3 times per day.

Face-to-face contact was mandatory at baseline and after 3 and 6 weeks of using the app. Additional contact with the therapist was possible via messenger, phone, or as face-to-face contact, depending on the participant's needs. Every contact was recorded in a protocol.

XSL•FO
RenderX

As suicidality is a common symptom of depression, especially in adolescents, the first module included information about suicidal thoughts and how to seek immediate help at the hospital in case of acute suicidality. Additionally, when consent was obtained, information was provided about procedures for serious suicidal behavior and suicidal behavior was clarified as part of the diagnostic process (see "Ethical Considerations").

Measures

Numerous assessments were made at baseline (T1), 3 weeks (T2), and 6 weeks (T3), as well as at the 3-month follow-up (T4). A full list is provided in Table 2. The following estimates have been used as a basis for the assessment of feasibility: adherence was assessed by recording the number of modules completed and the number of messages sent by participants to their therapist via the in-app messenger. Therapists documented their weekly interactions with the patient, including any additional contacts initiated by the patient and conducted between the patient and therapist (via phone, video call, or in-person) as well as adherence, among other things. An adherence score was created, based on the therapists' weekly study documentation. For this purpose, 4 categories were used to rate each module, ranging from 0 ("little or no engagement") to 3 ("excellent engagement," meaning successful lesson completion, including the accomplishment of the everyday task). This rating was conducted by 2 independent raters; initially, each rater provided a rating per module, followed by reaching a consensus together, and then averaging over all evaluable modules.

Patient satisfaction was measured using a self-report questionnaire developed by the authors of the SPARX study [25], which was subsequently modified and translated for the purposes of this study. The questionnaire contained 37 quantitative and 5 qualitative items, which addressed various markers of user experience. The following section will focus exclusively on the items that were evaluated in the context of this paper. For an overall evaluation of the app the item "How would you rate the app overall on a scale of 1 - 10?" was asked (1=bad and 10=excellent). The impact of the app was measured by rating the usefulness of 7 antidepressant strategies addressed in the app using a 5-point Likert scale ranging from 0 ("very useless") to 4 ("very useful"). Transferability to daily life was measured by rating the frequency of use of 6 strategies on a 5-point Likert scale ranging from 0 ("never") to 4 ("nearly every day"). Total scores for the scales "usefulness of the strategies" and "frequency of use of the strategies" were then calculated by summing the individual items within each scale.

Therapists' overall satisfaction with the app was assessed using the item "How satisfied are you overall with the MICHI app?" on a scale from 1=not satisfied at all to 10=very satisfied. This item was part of a more extensive survey of therapists regarding their experiences with the app. A thorough evaluation of the therapists' perspectives on the app will be presented elsewhere [26].

Trends of efficacy were measured by changes in BDI-II [17], CDRS-R [18], and Clinical Global Impressions-Severity Scale [19] scores. The cutoff value of 29 in the BDI-II rating at T1 was used to classify the severity of depressive symptoms as mild or moderate versus severe [17].

Adverse events (AE) were coded according to MedDRA (Medical Dictionary for Regulatory Activities) terminology. Detailed information collected for each AE included a description of the event, duration, whether the AE was serious, intensity, relationship to trial treatment, action taken, and clinical outcome. Summary tables presented the number of participants observed with AEs by MedDRA System Organ Class and Preferred Term, along with the corresponding percentages. Additional subcategories were based on event intensity and relationship to trial treatment.

Ethical Considerations

The institutional review board of the Medical Faculty of the University of Rostock approved this study (registration number: A 2020 - 0093). Before enrollment, participants and parents were verbally informed about the objective and methodology of this study and provided with separate written information documents. Written consent was then obtained from both parents and participants. Participants were informed that they could withdraw from the study at any time without negative consequences for their treatment. Participants aged 18 years did not require parental consent. Participants were not financially compensated for their participation in this study, as they were allowed to take part in an innovative new treatment program.

Suicidality was assessed at baseline and in the first module of the e-MICHI app. Participants were queried about whether they experience thoughts of death. They were informed that if they do have thoughts of death, their primary therapist would be notified via messenger and would reach out to them.

All data were deidentified for storage and subsequent evaluation, ensuring participant confidentiality throughout the research process.

Procedures Against Bias

To reduce bias, independent raters, namely psychology students and research psychologists, assessed inclusion criteria and outcome measures, like the CDRS-R [18]. The raters underwent training in the administration of the diagnostic interviews, but were not acquainted with the participants' psychiatric and psychotherapeutic histories. Furthermore, the assessment of adherence scores was conducted independently by 2 authors (LL and AS). Discrepancies between their findings were discussed and resolved.

Data Analysis

Feasibility metrics were assessed by analyzing their central tendency and dispersion. Additionally, their association with the severity of depression was examined using appropriate statistical tests such as Pearson correlation coefficients, Mann-Whitney *U* tests, and 2-sample *t* tests.

Concerning trends of efficacy, 1-factor ANOVAs were conducted with 3 repeated measurements of the dependent variable (BDI-II and CDRS-R scores) within 1 sample, provided the prerequisite conditions were met. If the necessary conditions were not met, a nonparametric method was used for statistical

XSL•FO
RenderX

analysis. Previous research has shown that using DHIs to treat depression in children and teenagers has mild to moderate positive effects [27]. As a result, we expected a moderate effect size of Cohen f=0.25 for this investigation. With a priori of 28 participants in total and a 5% probability of type I error, as well as a power of 0.8 and a 20% probability of type II error, the sample size was required to observe significant changes in the sample as per the protocol. The G*Power (Heinrich-Heine-Universität Düsseldorf) tool was used to carry out the sample size calculation [28]. Considering an attrition rate of approximately one-third than that was reported in previous trials involving smartphone app interventions for depressive symptoms [29], it was deemed as a suitable anticipated dropout rate for the investigation, ending up with a total of 36 enrolled participants. This sample size enabled us to explore the impact of covariates and associations between different variables. Furthermore, violations beyond the per-protocol sample of 28 participants could be recorded.

The main analysis was conducted with the per-protocol group, consisting of all participants from the intention-to-treat population who did not violate the study protocol, by failing to meet all inclusion criteria, meeting at least 1 exclusion criterion or being noncompliant to therapy.

Subgroup analyses were performed to identify potential variances among predictive factors, such as adherence, comorbid disorders, and treatment expectancy. The statistical test interpretation was exploratory, except for the trends of efficacy outcomes for which this study was powered.

Results

Participants and Dropout

A total of 61 participants were assessed for eligibility in this study. While informed consent was obtained from 57 individuals, 6 (11%) participants did not commence with the program. A further 15 (26%) participants dropped out (Figure 2). Dropouts were defined as participants who completed at least 1 lesson within the app but stopped using it for a variety of reasons, such as becoming unreachable for further contact, starting external outpatient psychotherapy, or being admitted to a day clinic or inpatient facility. For demographic data on the participants, please refer to Table 3. There was no significant difference in age or severity of depressive symptoms at T1 between the dropout group and the included participants. Further dropout analysis was conducted using a chi-square test to examine differences by study center, gender, school type, education, treatment expectancy, and previous psychiatric or psychotherapeutic experience. The results show no significant differences between dropouts and other participants on these variables. Thus, dropouts were deemed irrelevant for further analyses.

XSL•FO
RenderX

Figure 2. Participant flowchart depicting the recruitment, eligibility screening, and retention rate in the feasibility trial integrating the e-MICHI mental health app into a blended treatment of adolescent depression. BDI-II: Beck Depression Inventory-II.

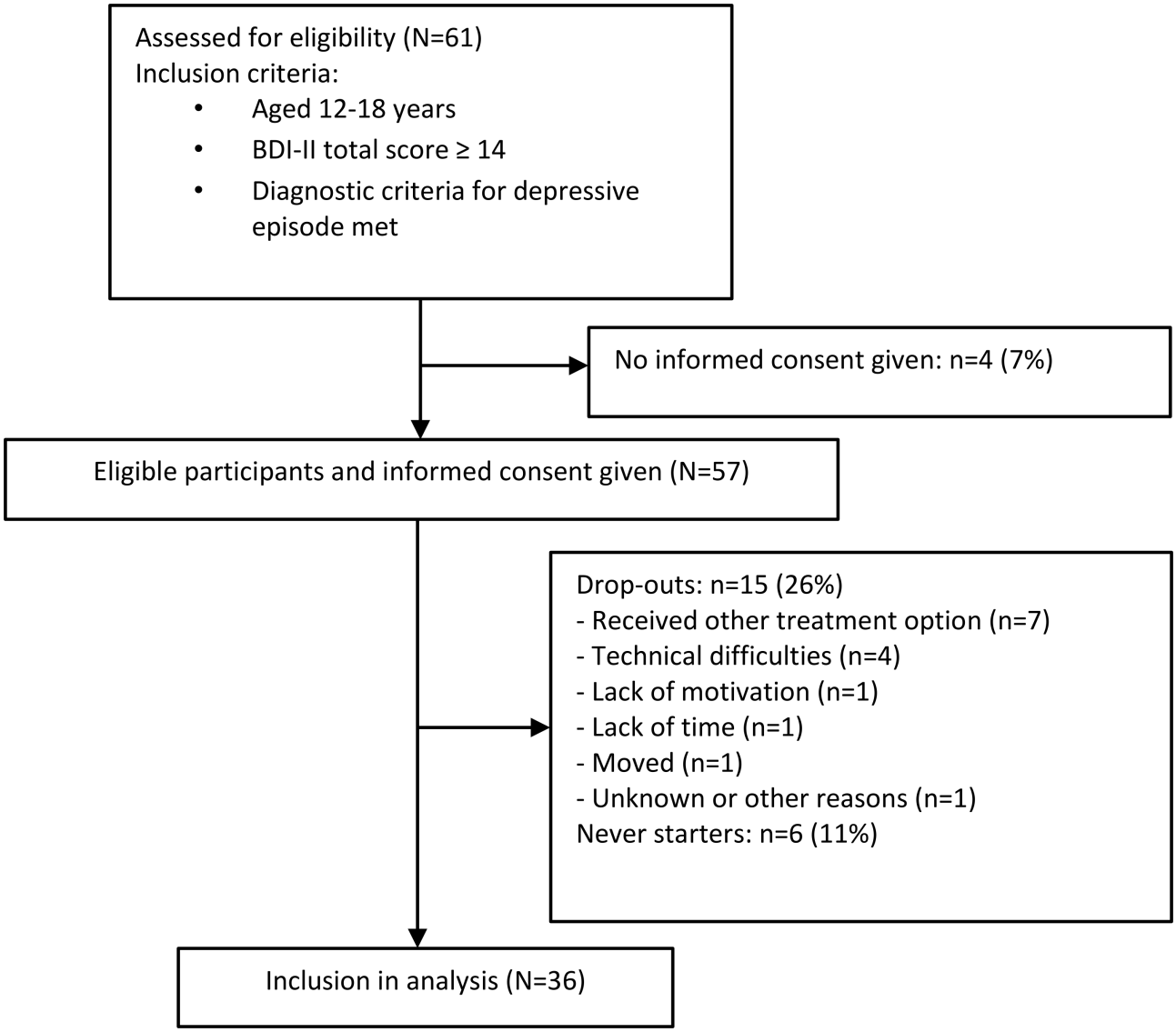

XSL•FO
RenderX

Table . Baseline demographic and clinical characteristics of participants (N=36) in the feasibility trial, including age, gender, distribution across outpatient centers, baseline depression severity, educational backgrounds, and prior treatment experience.

Characteristic		Values
Age, mean (SD)		
	Years	15 (1)
	Months	6 (7)
Age in years, n (%)		
	12	1 (3)
	13	7 (19)
	14	6 (17)
	15	10 (28)
	16	3 (8)
	17	8 (22)
	18	1 (3)
Gender, n (%)		
	Female	29 (83)
	Male	4 (11)
	Diverse	2 (6)
Center, n (%)		
	Neuruppin	17 (47)
	Rostock	19 (53)
Severity of depressive symptoms, n (%)		
	BDI-II<29	6 (17)
	BDI-II≥29	29 (83)
Education, n (%)		
	Grammar school	16 (46)
	Middle school	8 (23)
	Comprehensive school	4 (11)
	Vocational school	2 (6)
	Orientation stage	1 (3)
	Not attending school	4 (11)
Previous psychiatric or psychotherapy experience, mean (SD)		
	Years	1 (2)
	Months	8 (0)
Previous psychiatric or psychotherapy experience, n (%)		
	No previous experience	9 (26)
	≤1 year treatment	15 (43)
	>1 year treatment	11 (31)

Adherence and Overall Satisfaction

While 50% (n=17) of participants successfully completed all modules, 74% (n=25) of participants completed at least the first 5 modules covering all therapeutic content (Table 4). Regarding patient communication via the app's messaging feature, an average of 23 (SD 22.1) messages were sent per participant, with a high variance. According to the adherence score, almost 80% (n=25) of all participants showed good or excellent adherence (Table 4). On average, less than 1 additional contact initiated by the patient was documented by the therapist per week (mean 0.7, SD 0.31). There was no statistically significant difference in adherence score and messaging activity between participants with severe depressive symptoms and mild or moderate depressive symptoms (Table 4) and there was no

XSL•FO
RenderX

correlation between the number of messages sent by participants and the number of lessons completed (r=.33, P=.06).

Table . Feasibility measures, including completion rates for app modules, adherence to the intervention, engagement via messages, additional contacts, and relationships between feasibility outcomes and depression severity.

Variables	Values	*P* value
Modules completed (N=34), mean (SD)	5.03 (1.27)	N/A[a]
Modules completed, n (%)		N/A
1	34 (100)	
2	34 (100)	
3	31 (91)	
4	30 (88)	
5	25 (74)	
6	17 (50)	
Adherence score (N=32), n (%)		N/A
0 (no or little engagement)	0 (0)	
1 (moderate engagement)	7 (22)	
2 (good engagement)	10 (31)	
3 (excellent engagement)	15 (47)	
Messages sent (N=34), mean (SD)	23 (22.1)	
Additional contact per week, mean (SD)	0.7 (0.31)	
Severity of depression, *t* test *(df)*		
Perceived usefulness	0.732 (25)	.47
Strategy implementation	0.808 (24)	.43
Severity of depression, Mann-Whitney *U*		
Modules complete	56.5	.76
Messages sent	63	.73
Patient app rating	40	.1

[a]N/A: not applicable.

Further analysis also showed no correlation between adherence score and therapy expectation (r=.17, P=.34), the overall satisfaction with the app (r=–.07, P=.73), and changes in depressive symptoms, as measured by both BDI-II (r=.18, P=.40) and CDRS-R (r=−.24, P=.19) scales. There was no statistically significant difference between the adherence scores of the group with previous psychiatric or psychotherapeutic experience and the group with no previous experience (t_{30}=−1.22, P=.23).

Participants rated their overall satisfaction with the intervention as 7.3 on average (SD 1.2). Therapists indicated similar rates of overall satisfaction (mean 7.3, SD 1.1). Qualitative feedback from the therapists indicated that they particularly appreciated the inclusion of a mood diary due to its potential to monitor and reflect on emotional change.

Impact of the App and Transferability to Daily Life

A detailed analysis on the perceived usefulness of the therapeutic techniques addressed in the e-MICHI app is presented in Table 5. In general, a high rate of participants indicated using therapeutic techniques in everyday life, with variance between different techniques (Table 6).

XSL•FO
RenderX

Table . Participant feedback on the perceived usefulness of various therapeutic techniques addressed in the e-MICHI app based on self-reported relevance to managing depression symptoms (N=27).

	Very useless or useless, n (%)	Unsure, n (%)	Very useful or useful, n (%)
Psychoeducation	1 (4)	6 (22)	20 (74)
Behavioral change or activation	7 (26)	5 (19)	15 (56)
Identification of negative thoughts[a]	3 (11)	12 (44)	12 (44)
Changing negative thinking[a]	6 (22)	10 (37)	11 (41)
Problem solving	5 (19)	9 (33)	13 (48)
Feel better about myself[a]	10 (37)	6 (22)	11 (41)
Emotion regulation	6 (22)	9 (33)	12 (44)

[a]As part of cognitive restructuring.

Table . Participant-reported frequency of practicing therapeutic techniques addressed in the e-MICHI app at the third assessment point (T3), reflecting the integration of these strategies into participants' daily routines.

	N	Never, n (%)	A few times, n (%)	About once per week, n (%)	Several times per week, n (%)	Nearly every day, n (%)
Paid attention to basic needs	27	1 (4)	6 (22)	7 (26)	6 (22)	7 (26)
Spent time for activities that improve mood	27	1 (4)	3 (11)	7 (26)	11 (41)	5 (19)
Applied techniques of problem-solving	27	6 (22)	7 (26)	4 (15)	7 (26)	3 (11)
Identified positive things about yourself	27	2 (7)	6 (22)	9 (33)	7 (26)	3 (11)
Recognized negative automatic thoughts	27	3 (11)	6 (22)	6 (22)	9 (33)	3 (11)
Tried to change negative thoughts	26	2 (8)	7 (27)	6 (23)	9 (35)	2 (8)

A positive correlation was found between changes in BDI-II scores, reflecting a perceived decrease in depressive symptoms, and both the perceived overall usefulness of the techniques (r=.54, P=.004) and their total frequency of use (r=.62, P=.001). There was also a significant positive correlation between participants' overall satisfaction with the app and the perceived usefulness of the techniques (r=.64, P<.001), as well as the frequency of their implementation (r=.70, P<.001).

Based on the severity of depressive symptoms, the analysis revealed no significant differences in the perceived utility of the techniques or their implementation frequency (see Table 4).

A more detailed investigation was carried out to explore whether the perceived effectiveness of e-MICHI in "tackling problems or issues causing difficulty" was related to the severity of depression. Using a Mann-Whitney U test, no significant difference in the evaluation of the item's usefulness between participants with lower BDI-II scores and those with higher BDI-II scores was found (U=52.00, Z=−0.34, P=.76).

Trends of Efficacy of e-MICHI

Depressive symptoms decreased significantly during the blended therapy, regardless of the measure used (BDI-II, CDRS-R, and Clinical Global Impressions-Severity Scale scores), and remained stable at follow-up 3 months after using the app (Table 7).

XSL•FO
RenderX

Table . Efficacy measures evaluating changes in self-reported depression severity, clinician-rated changes in depressive symptoms, and severity of illness from baseline (T1) to follow-up (T4) in the feasibility trial of e-MICHI.

Variables		Values
BDI-II[a] (N=21)		
	T1, mean (SD)	37.71 (10.77)
	T2, mean (SD)	32.52 (12.89)
	T3, mean (SD)	30.62 (13.21)
	T4, mean (SD)	30.95 (15.38)
	F test (*df*)	4.10 (3, 60)
	P value	.01
CDRS-R[b] (N=27)		
	T1, mean (SD)	58.93 (11.51)
	T3, mean (SD)	48.37 (15.28)
	T4, mean (SD)	42.48 (16.61)
	F test (*df*)	16.39 (2, 52)
	P value	<.001
CGI-S[c] (N=29)		
	T1, mean (SD)	4.48 (0.79)
	T2, mean (SD)	3.97 (0.87)
	T3, mean (SD)	3.48 (0.95)
	T4, mean (SD)	3.38 (1.32)
	Chi-square (*df*)	33.45 (3)
	P value	<.001

[a]BDI-II: Beck Depression Inventory-II.

[b]CDRS-R: Children's Depression Rating Scale-Revised (clinician rating).

[c]CGI-S: Clinical Global Impressions-Severity Scale.

About AEs

There were 6 AEs of nonsuicidal self-injury, 1 of which was a serious adverse event (SAE) that required inpatient admission for crisis intervention. Thoughts of death were reported on 12 occasions, but there were no indications of acute suicidality. Another AE involved an outpatient procedure for a nonrelated medical condition. None of the 7 AEs, including the SAE, were deemed related to the intervention and this study's therapy continued unchanged.

Of the 7 dropouts who received other treatment options, 3 were inpatients and 2 were day patients (planned partial hospitalization), both categories coded as SAE. However, these SAEs were not considered to be related to the intervention. Additionally, there were 2 dropouts where study treatment was discontinued due to other outpatient treatment options.

Discussion

Principal Findings

This study aimed to investigate the feasibility of the DHI e-MICHI as a blended intervention to treat mild, moderate, and severe depression among adolescents. We found that adolescents adhered well to the intervention, indicated general satisfaction as well as a reduction in depressive symptoms from both patient and therapist perspectives. The use of the app was feasible in a population of adolescents with clinically significant depression and significant effects were observed in child and adolescent psychiatric outpatients. Improvement of depressive symptoms measured at the end of the intervention remained stable at the 3-month follow-up.

Crucial points in e-mental health applications are adherence and transferability to everyday practice [13]. We observed high adherence rates, which are comparable to other blended interventions [30] and substantially higher than unguided e-health applications [31,32]. In this study, participant satisfaction and reported transfer to daily practice were high. This differs from results of other studies, such as the SPARX-R program [12], which reported lower perceived usefulness and transfers in daily practice in adolescents excluded from mainstream education. A high transfer of psychotherapeutic exercises and skills into everyday life could be one of the most important benefits of using apps in psychotherapeutic practice. Apps could intensify therapy and with more individualized feedback, for example, based on ecological momentary

XSL•FO
RenderX

assessment or generated by AI, a tailored personalized psychotherapeutic intervention could be provided.

Concerning adherence, a substantial proportion of participants, about 3-quarters, completed the core modules of e-MICHI. Reasons for high adherence may include aesthetic design [33], relevant content [34], or the perceived immediacy of support through the app's messenger as stated by participants. These findings align with the principles highlighted by Domhardt et al [8]. Adolescents also reported applying the strategies they learned from the app in their daily lives, a promising sign of real-life transfer, even among those who rated their depressive symptoms as severe. The high adherence may be influenced by the fact that our study population consisted of adolescents seeking psychiatric-psychotherapeutic treatment who were clinically significantly depressed and impaired by their disorder. Interestingly, our findings indicate no significant differences in adherence between participants with long or short-term and no prior therapeutic experience. This highlights the possibility that even individuals with prior experience in therapy may derive benefits from e-MICHI suggesting the intervention could serve as a valuable adjunct to traditional therapeutic approaches, enhancing treatment outcomes and promoting sustained engagement across diverse patient populations. Additionally, the absence of an existing therapeutic relationship for adequate adherence would not limit the use of DHIs, for example, to bridge the gap for face-to-face psychotherapy or as a primary intervention. Notably, adolescents appreciated the close contact with therapists provided by the messaging function, highlighting the added benefit in blended approaches [35]. DHIs may therefore be more beneficial when embedded in conventional care settings.

The acceptance of the app is evidenced by the high satisfaction ratings from participants and therapists alike. The high acceptance rate, especially among adolescents, may be partly due to their general openness to digital interventions [6]. In combination with the therapeutic relationship and the facilitated way of communicating directly with the therapist via the app, e-MICHI, or DHIs in general, seem to be an added value in psychiatric-psychotherapeutic treatment for adolescents.

The inclusion of a mood diary, as appreciated by the therapists, could enhance therapeutic engagement and outcomes due to its potential to track and reflect on emotional changes [36]. This consensus, together with the high level of satisfaction, supports the usefulness of the app as a complementary tool in psychotherapy and its potential to strengthen the therapeutic alliance, as also shown by Tremain et al [37].

In our study, additional contacts beyond the planned 3 sessions were rare, indicating efficient management of patient care with reduced direct contact frequency compared to our clinical experience, where patients in outpatient settings are typically seen every 1 to 2 weeks. Additionally, compared to treatment as usual, messenger contacts helped provide feedback on mood diary entries and suggested steps for e-MICHI daily task implementation.

This finding aligns with the objective of providing more accessible care [38] as it can reduce both waiting times to start psychotherapy and the need to travel long distances to receive care. It must be noted, that technical problems and difficulties with devices or limited network coverage or access to free Wi-Fi may be barriers to the use of therapeutic apps, especially for adolescents. In this study, 4/57 (7%) of originally included participants dropped out due to technical problems. Access to free Wi-Fi seems to be essential for adolescents to be able to use the app in their living environments. Ways to use the app offline would be desirable to make it accessible to even more patients.

Regarding trends of efficacy, there was a noticeable decrease in depressive symptoms during the intervention, as reported by both participants and external evaluators. This significant improvement highlights the potential of e-MICHI in the treatment of depression, warranting further validation through randomized controlled trials (RCTs) to more robustly establish its clinical efficacy.

Given the sample, which included participants with severe depression, drop out due to referral to other psychiatric treatment services was not surprising. The advantage of the blended approach is evident in the case of a patient who was admitted for crisis intervention and was able to resume the intervention after discharge providing continuity of care and additional therapist support, consequently increasing adherence.

Regarding safety, this study presents an optimistic perspective on the e-MICHI blended CBT intervention for adolescents with depression, as evidenced by a small number of unrelated AEs and no suicide attempts. Comparatively, Rasing et al [39] observed 2 suicide attempts in a similar blended CBT intervention, but found no significant differences in AEs between face-to-face and blended treatment conditions. A recent review indicated no increased risk of suicidality with blended treatment for adolescents with depression [40]. Concerning self-harm, Fleming et al [41] reported a comparable number of 2 incidents in a similar sample size, neither of which were related to the intervention. The absence of suicide attempts during the e-MICHI intervention, along with the findings from Rasing et al [39] and Fleming et al [41], suggests that blended interventions may not substantially increase the risk of severe AEs, such as suicide attempts and self-harm. Nonetheless, Domhardt et al [8] emphasize the need for continued vigilance and monitoring in blended mental health interventions, especially for high-risk populations.

Contrasting the results of e-MICHI with similar blended interventions, e-MICHI allowed participants to use their own devices, incorporating beneficial strategies such as text messaging for homework reminders and self-monitoring between sessions [42], and included participants with severe depressive symptoms, addressing the criticism regarding the exclusion of such cases in previous research [43]. As noted, we found no significant difference in adherence rates in participants with severe depressive symptoms. While this is consistent with previous findings [44], this result may also be due to the small sample size.

Comparing e-MICHI to unguided internet-based CBT [40] highlights the potential benefits of a blended approach. The higher adherence rates and satisfaction scores suggest that integrating therapist support into an e-health platform may

XSL•FO
RenderX

enhance the perceived usefulness of these interventions, as well as how well adolescents are able to apply the therapeutic strategies they learn to their daily lives.

Limitations

The non-RCT design of this study limits the generalizability of our findings on efficacy. As a feasibility study we wanted to detect trends of efficacy; the trial was not designed for efficacy testing. An RCT will be the next step in generating evidence for the e-MICHI intervention. Most of this study's population had a history of psychotherapy. Therefore, our participants might be more familiar with therapeutic contents than adolescents who use such a therapeutic app as a first-time psychotherapeutic intervention. The adherence scoring method faced challenges due to several therapy documentations missing or lacking sufficient information for rating. Given the small sample size, it is possible that these factors could have influenced the findings. For the development of therapeutic interventions according to precision medicine, it would be of interest to know about precise times of use of specific contents (eg, in the morning, evening, etc), direct effects on symptoms (eg, mood), etc. Unfortunately, due to technical and privacy restrictions, we were unable to evaluate this data.

Conclusion

Due to its easy application, the e-MICHI intervention seems to be a promising tool to enhance face-to-face treatment of depression in adolescents. This study identified areas for further development of the app based on user feedback and technical issues, to improve its effectiveness and user experience in the future. It contributes to filling the existing evidence gap on the feasibility and efficacy of e-therapy interventions via apps for adolescents with depression in the German health care system. According to the results of this study, the potential benefits of blended therapy for depression by intensifying therapy and facilitating patient-therapist communication are evident. As digital interventions continue to evolve, further research is imperative to address challenges, for example, accommodating a broad range of individuals, crisis intervention, and high dropout rates, ensuring that such platforms can effectively serve as viable therapeutic options for adolescents with depression.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

MK receives royalties from publishing companies as an author of books. He served as a scientific advisor for Janssen. He received grants from the German Ministry of Research and Education and the Ministry of Social Affairs, Health and Sports Mecklenburg-Vorpommern. This work is unrelated to the above grants and relationships. LL, TM-D, AD, AS, DAN, NS-K, SB, and OR declare no competing interests.

References

1. Kölch M, Reis O, Ulbrich L, Schepker R. Mental disorders in minors during the COVID-19 pandemic-an analysis of nationwide hospital treatment data. Dtsch Arztebl Int 2023 May 19;120(20):362-363. [doi: 10.3238/arztebl.m2023.0010] [Medline: 37493050]
2. Witte J, Zeitler A, Batram M, Diekmannshemke J, DAK-Kinder HL. Und jugendreport 2022: DAK im dialog. 2022 URL: https://caas.content.dak.de/caas/v1/media/45734/data/21b30080687f14a8c09d7a54370f8860/kinder-und-jugendreport-2022-sammelband.pdf [accessed 2025-03-27]
3. Dolle K, Schulte-Körne G. Behandlung von depressiven Störungen bei Kindern und Jugendlichen: Evidenz- und konsensbasierte Leitlinie. Prax Kinderpsychol Kinderpsychiatr 2013;63(3):237-263. [doi: 10.13109/prkk.2014.63.3.237]
4. Depression in children and young people: identification and management (NG134). National Institute for Health and Care Excellence. 2019. URL: https://www.nice.org.uk/guidance/ng134 [accessed 2025-03-26]
5. Allgaier K, Schmid J, Hollmann K, Reusch PA, Conzelmann A, Renner TJ. Times are changing: digitalisation in child and adolescent psychotherapy. Eur Child Adolesc Psychiatry 2021 Nov;30(11):1667-1670. [doi: 10.1007/s00787-020-01610-8] [Medline: 32737600]
6. Lincke L, Ulbrich L, Reis O, et al. Attitudes toward innovative mental health treatment approaches in Germany: e-mental health and home treatment. Front Psychiatry 2022;13:889555. [doi: 10.3389/fpsyt.2022.889555] [Medline: 35911231]
7. Furukawa TA, Suganuma A, Ostinelli EG, et al. Dismantling, optimising, and personalising internet cognitive behavioural therapy for depression: a systematic review and component network meta-analysis using individual participant data. Lancet Psychiatry 2021 Jun;8(6):500-511. [doi: 10.1016/S2215-0366(21)00077-8] [Medline: 33957075]
8. Domhardt M, Steubl L, Baumeister H. Internet- and mobile-based interventions for mental and somatic conditions in children and adolescents. Z Kinder Jugendpsychiatr Psychother 2020 Jan;48(1):33-46. [doi: 10.1024/1422-4917/a000625] [Medline: 30422059]
9. Achilles MR, Anderson M, Li SH, Subotic-Kerry M, Parker B, O'Dea B. Adherence to e-mental health among youth: considerations for intervention development and research design. Digit Health 2020;6:2055207620926064. [doi: 10.1177/2055207620926064] [Medline: 32547775]

XSL•FO
RenderX

10. Richards D, Richardson T. Computer-based psychological treatments for depression: a systematic review and meta-analysis. Clin Psychol Rev 2012 Jun;32(4):329-342. [doi: 10.1016/j.cpr.2012.02.004] [Medline: 22466510]
11. Lehtimaki S, Martic J, Wahl B, Foster KT, Schwalbe N. Evidence on digital mental health interventions for adolescents and young people: systematic overview. JMIR Ment Health 2021 Apr 29;8(4):e25847. [doi: 10.2196/25847] [Medline: 33913817]
12. Kuosmanen T, Fleming TM, Newell J, Barry MM. A pilot evaluation of the SPARX-R gaming intervention for preventing depression and improving wellbeing among adolescents in alternative education. Internet Interv 2017 Jun;8(40-47):40-47. [doi: 10.1016/j.invent.2017.03.004] [Medline: 30135827]
13. Seiferth C, Vogel L, Aas B, et al. How to e-mental health: a guideline for researchers and practitioners using digital technology in the context of mental health. Nat Mental Health 2023;1(8):542-554. [doi: 10.1038/s44220-023-00085-1]
14. Grist R, Porter J, Stallard P. Mental health mobile apps for preadolescents and adolescents: a systematic review. J Med Internet Res 2017 May 25;19(5):e176. [doi: 10.2196/jmir.7332] [Medline: 28546138]
15. Christ C, Schouten MJ, Blankers M, et al. Internet and computer-based cognitive behavioral therapy for anxiety and depression in adolescents and young adults: systematic review and meta-analysis. J Med Internet Res 2020 Sep 25;22(9):e17831. [doi: 10.2196/17831] [Medline: 32673212]
16. The ICD-10 classification of mental and behavioural disorders: clinical descriptions and diagnostic guidelines. World Health Organization. 1992. URL: https://cdn.who.int/media/docs/default-source/classification/other-classifications/9241544228_eng.pdf [accessed 2025-03-27]
17. Beck AT, Steer RA, Brown G. Manual for Beck Depression Inventory-II: Psychological Corporation; 1996.
18. Keller F, Grieb J, Ernst M, Spröber N, Fegert JM, Kölch M. Children's Depression Rating Scale-Revised (CDRS-R): development of a German version and psychometric properties in a clinical sample. Z Kinder Jugendpsychiatr Psychother 2011 May;39(3):179-185. [doi: 10.1024/1422-4917/a000090] [Medline: 21563109]
19. Guy W. ECDEU Assessment Manual for Psychopharmacology: Government Printing Office; 1976.
20. Schneider S, Pflug V, Margraf J, In-Albon T. Kinder-DIPS: Diagnostic Interview for Mental Disorders in Children and Adolescents [Diagnostisches Interview Bei Psychischen Störungen Im Kindes- Und Jugendalter]: Mental Health Research and Treament Center, Ruhr-Universität Bochum; 2017.
21. Spröber N, Fegert JM, Kölch M, Straub J. Depressionen Im Jugendalter: MICHI ; Manual Für Die Gruppentherapie: Beltz; 2012.
22. Straub J, Plener PL, Keller F, Fegert JM, Spröber N, Kölch MG. MICHI–eine Gruppen-Kurzzeitpsychotherapie zur Behandlung von Depressionen bei Jugendlichen. Kindheit Entwickl 2015 Jul;24(3):189-198. [doi: 10.1026/0942-5403/a000175]
23. McCarty CA, Weisz JR. Effects of psychotherapy for depression in children and adolescents: what we can (and can't) learn from meta-analysis and component profiling. J Am Acad Child Adolesc Psychiatry 2007 Jul;46(7):879-886. [doi: 10.1097/chi.0b013e31805467b3] [Medline: 17581452]
24. Zone35 individual. innovative. digital. Zone35. 2020. URL: https://zone35.de/#home [accessed 2025-03-26]
25. Merry SN, Stasiak K, Shepherd M, Frampton C, Fleming T, Lucassen MFG. The effectiveness of SPARX, a computerised self help intervention for adolescents seeking help for depression: randomised controlled non-inferiority trial. BMJ 2012 Apr 18;344:e2598. [doi: 10.1136/bmj.e2598] [Medline: 22517917]
26. Lincke L, Kilb H, Martin-Döring T, et al. Blended care for adolescents with depression from the perspective of treatment providers - feasibility study in an outpatient setting [Blended care für jugendliche mit depressiven erkrankungen aus der sicht von behandelnden - machbarkeitsstudie in einem ambulanten setting]. Kindheit Entwickl .
27. Hollis C, Falconer CJ, Martin JL, et al. Annual research review: digital health interventions for children and young people with mental health problems - a systematic and meta-review. J Child Psychol Psychiatry 2017 Apr;58(4):474-503. [doi: 10.1111/jcpp.12663] [Medline: 27943285]
28. Faul F, Erdfelder E, Lang AG, Buchner A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods 2007 May;39(2):175-191. [doi: 10.3758/bf03193146] [Medline: 17695343]
29. Linardon J, Fuller-Tyszkiewicz M. Attrition and adherence in smartphone-delivered interventions for mental health problems: a systematic and meta-analytic review. J Consult Clin Psychol 2020 Jan;88(1):1-13. [doi: 10.1037/ccp0000459] [Medline: 31697093]
30. Schuster R, Leitner I, Carlbring P, Laireiter AR. Exploring blended group interventions for depression: randomised controlled feasibility study of a blended computer- and multimedia-supported psychoeducational group intervention for adults with depressive symptoms. Internet Interv 2017 Jun;8(63-71):63-71. [doi: 10.1016/j.invent.2017.04.001] [Medline: 30135830]
31. Melville KM, Casey LM, Kavanagh DJ. Dropout from Internet-based treatment for psychological disorders. Br J Clin Psychol 2010 Nov;49(Pt 4):455-471. [doi: 10.1348/014466509X472138] [Medline: 19799804]
32. Waller R, Gilbody S. Barriers to the uptake of computerized cognitive behavioural therapy: a systematic review of the quantitative and qualitative evidence. Psychol Med 2009 May;39(5):705-712. [doi: 10.1017/S0033291708004224] [Medline: 18812006]
33. Denison-Day JL, Muir S, Newell C, Appleton KM. The role of aesthetics in intentions to use digital health interventions. PLOS Digit Health 2023 Jun;2(6):e0000274. [doi: 10.1371/journal.pdig.0000274] [Medline: 37347727]

XSL•FO
RenderX

34. Beatty L, Binnion C. A systematic review of predictors of, and reasons for, adherence to online psychological interventions. Int J Behav Med 2016 Dec;23(6):776-794. [doi: 10.1007/s12529-016-9556-9] [Medline: 26957109]
35. Lindhiem O, Bennett CB, Rosen D, Silk J. Mobile technology boosts the effectiveness of psychotherapy and behavioral interventions: a meta-analysis. Behav Modif 2015 Nov;39(6):785-804. [doi: 10.1177/0145445515595198] [Medline: 26187164]
36. Matthews M, Doherty G, Coyle D, Sharry J. Designing mobile applications to support mental health interventions. In: Lumsden J, editor. Handbook of Research on User Interface Design and Evaluation For Mobile Technology 2008:635-656. [doi: 10.4018/978-1-59904-871-0.ch038]
37. Tremain H, McEnery C, Fletcher K, Murray G. The therapeutic alliance in digital mental health interventions for serious mental illnesses: narrative review. JMIR Ment Health 2020 Aug 7;7(8):e17204. [doi: 10.2196/17204] [Medline: 32763881]
38. Bundespsychotherapeutenkammer. Wartezeiten in Der Psychotherapie: BPtK Spezial; 2011, Vol. 6:1-2.
39. Rasing SPA, Stikkelbroek YAJ, den Hollander W, et al. Pragmatic quasi-experimental controlled trial evaluating the outcomes of blended CBT compared to face-to-face CBT and treatment as usual for adolescents with depressive disorders. Int J Environ Res Public Health 2021 Mar 17;18(6):3102. [doi: 10.3390/ijerph18063102] [Medline: 33802913]
40. Rasing SPA. Blended treatment for depressive disorders in youth: a narrative review. J Cogn Ther 2021 Mar;14(1):47-58. [doi: 10.1007/s41811-020-00088-1]
41. Fleming T, Dixon R, Frampton C, Merry S. A pragmatic randomized controlled trial of computerized CBT (SPARX) for symptoms of depression among adolescents excluded from mainstream education. Behav Cogn Psychother 2012 Oct;40(5):529-541. [doi: 10.1017/S1352465811000695] [Medline: 22137185]
42. Kobak KA, Mundt JC, Kennard B. Integrating technology into cognitive behavior therapy for adolescent depression: a pilot study. Ann Gen Psychiatry 2015;14(37):37. [doi: 10.1186/s12991-015-0077-8] [Medline: 26535048]
43. Sethi S, Campbell AJ, Ellis LA. The use of computerized self-help packages to treat adolescent depression and anxiety. J Technol Hum Serv 2010 Aug 31;28(3):144-160. [doi: 10.1080/15228835.2010.508317]
44. De Vos R, Tromp N, Bodden D, Stikkelbroek Y. Therapeut onmisbaar bij ‘blended therapie’ tegen depressie. Kind Adolesc Prakt 2017 Jun;16(2):6-13. [doi: 10.1007/s12454-017-0014-x]

Abbreviations

AE: adverse event
BDI-II: Beck Depression Inventory-II
CBT: cognitive behavioral therapy
CDRS-R: Children’s Depression Rating Scale-Revised
DHI: digital health intervention
***ICD-10*:** *International Statistical Classification of Diseases and Related Health Problems, 10th Revision*
MDD: major depressive disorder
MedDRA: Medical Dictionary for Regulatory Activities
MICHI: Manualized Intervention to Cope With Depressive Symptoms, Help Strengthen Resources, and Improve Emotion Regulation
RCT: randomized controlled trial
SAE: serious adverse event

Edited by A Mavragani; submitted 18.03.24; peer-reviewed by K Sprange; revised version received 14.01.25; accepted 15.01.25; published 01.05.25.

Please cite as:
Lincke L, Martin-Döring T, Daunke A, Sadkowiak A, Nolkemper DA, Sproeber-Kolb N, Bienioschek S, Reis O, Kölch M
Integration of a Mental Health App (e-MICHI) Into a Blended Treatment of Depression in Adolescents: Single-Group, Naturalistic Feasibility Trial
JMIR Form Res 2025;9:e58427
URL: https://formative.jmir.org/2025/1/e58427
doi:10.2196/58427

© Lena Lincke, Tim Martin-Döring, Andrea Daunke, Antonia Sadkowiak, Daria Alexandra Nolkemper, Nina Sproeber-Kolb, Stefanie Bienioschek, Olaf Reis, Michael Kölch. Originally published in JMIR Formative Research (https://formative.jmir.org), 1.5.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium,

XSL•FO
RenderX

provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Identification of Major Bleeding Events in Postoperative Patients With Malignant Tumors in Chinese Electronic Medical Records: Algorithm Development and Validation

Hui Li[1], PhD; Haiyang Yao[2], MD; Yuxiang Gao[2], BA; Hang Luo[2], MD; Changbin Cai[3], MD; Zhou Zhou[3], MD; Muhan Yuan[3], MD; Wei Jiang[4], MD

[1]Department of Thoracic Surgery, Beijing Chao-Yang Hospital, Capital Medical University, No.8 South Road of Workers' Stadium, Chaoyang District, Beijing, China

[2]Department of Technology, Shanghai Palan DataRx Co., Ltd, Shanghai, China

[3]Sanofi China Medical Affairs, Sanofi China, Shanghai, China

[4]Pharmaceutical Business Division, Basebit (Shanghai) Information Technology Co., Ltd, Shanghai, China

Corresponding Author:
Hui Li, PhD
Department of Thoracic Surgery, Beijing Chao-Yang Hospital, Capital Medical University, No.8 South Road of Workers' Stadium, Chaoyang District, Beijing, China

Abstract

Background: Postoperative bleeding is a serious complication following abdominal tumor surgery, but it is often not clearly diagnosed and documented in clinical practice in China. Previous studies have relied on manual interpretation of medical records to determine the presence of postoperative bleeding in patients, which is time-consuming and laborious. More critically, this manual approach severely hinders the efficient analysis of large volumes of medical data, impeding in-depth research into the incidence patterns and risk factors of postoperative bleeding. It remains unclear whether machine learning can play a role in processing large volumes of medical text to identify postoperative bleeding effectively.

Objective: This study aimed to develop a machine learning model tool for identifying postoperative patients with major bleeding based on the electronic medical record system.

Methods: This study used data from the available information in the National Health and Medical Big Data (Eastern) Center in Jiangsu Province of China. We randomly selected the medical records of 2,000 patients who underwent in-hospital tumor resection surgery between January 2018 and December 2021 from the database. Physicians manually classified each note as present or absent for a major bleeding event during the postoperative hospital stay. Feature engineering involved bleeding expressions, high-frequency related expressions, and quantitative logical judgment, resulting in 270 features. Logistic regression (LR), K-nearest neighbor (KNN), and convolutional neural network (CNN) models were developed and trained using the 1600-note training set. The main outcomes were accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) for each model.

Results: Major bleeding was present in 4.31% (69/1600) of the training set and 4.75% (19/400) of the test set. In the test set, the LR method achieved an accuracy of 0.8275, a sensitivity of 0.8947, a specificity of 0.8241, a PPV of 0.2024, an NPV of 0.9937, and an F_1-score of 0.3301. The CNN method demonstrated an accuracy of 0.8900, sensitivity of 0.8421, specificity of 0.8924, PPV of 0.2807, NPV of 0.9913, and an F_1-score of 0.4211. While the KNN method showed a high specificity of 0.9948 and an accuracy of 0.9575 in the test set, its sensitivity was notably low at 0.2105. The C-statistic for the LR method was 0.9018 and for the CNN method was 0.8830.

Conclusions: Both the LR and CNN methods demonstrate good performance in identifying major bleeding in patients with postoperative malignant tumors from electronic medical records, exhibiting high sensitivity and specificity. Given the higher sensitivity of the LR method (89.47%) and the higher specificity of the CNN method (89.24%) in the test set, both models hold promise for practical application, depending on specific clinical priorities.

(JMIR Form Res 2025;9:e66189) doi:10.2196/66189

KEYWORDS

machine learning; electronic medical record; postoperative patients with malignant tumors; postoperative bleeding; tumor surgery; abdominal

XSL•FO
RenderX

Introduction

Bleeding events are frequent complications encountered in postoperative clinical settings and can stem from the use of anticoagulant or antiplatelet drugs, invasive surgical treatment, or patient-related conditions and the existence of comorbidities, which are associated with increased morbidity, mortality, and health care costs [1-4]. Patients experiencing gastrointestinal bleeding while in the intensive care unit face a fourfold increase in mortality risk compared to those without bleeding issues, along with an additional eight-day stay within the unit [4]. Ample clinical evidence supports the correlation between intraoperative blood loss, especially excessive blood loss and adverse effects on the tumor prognosis. On the other hand, concern about postoperative bleeding may become the main reason why clinicians might be overly cautious in using medications to prevent venous thrombosis (VTE), despite VTE being the second leading cause of death among patients with tumors. While the mortality rate of patients with tumors and VTE is three times higher than that of other patients, timely detection of bleeding risks through progress notes and balanced selection of anticoagulants are crucial for postoperative patients with malignant tumors [5,6].

Currently, there is a growing number of retrospective studies on postoperative bleeding risk in cancer patients. It is also worth noting the lack of fully established risk prediction schemes or risk assessment tools for postoperative bleeding. However, a major challenge in conducting these studies is determining whether patients experienced bleeding events during their past treatment processes. In the clinical setting in China, current approaches to identifying bleeding episodes predominantly rely on diagnostic records, which often lack precision due to inconsistent descriptions and missing details. The task of pinpointing bleeding events from a patient's medical history can be particularly challenging, especially when dealing with extensive records [7,8]. Despite the transition to electronic medical records, these valuable sources of data are frequently underused. Within these records, details within course notes, including physical examination reports and discharge summaries, often contain firsthand accounts of bleeding incidents or clear indications of their absence. Nonetheless, manual identification of these events can be both time-consuming and arduous, raising concerns about accuracy and practicality. Thus, emphasis should be placed on the imperative need to develop methodologies geared toward effectively identifying bleeding events within existing medical records.

In the realm of clinical research, the accurate identification of bleeding events within large clinical datasets holds paramount importance. Regrettably, the current landscape lacks automated and scalable machine learning (ML) methodologies tailored for this objective. This is a significant unmet need, especially considering the clinical need for early detection of postoperative bleeding to improve patient outcomes. ML has been regarded as a method for developing models that depict intricate nonlinear systems and handling a vast array of potential variables found in contemporary electronic medical records. ML techniques have found application in various health care scenarios such as forecasting cancer susceptibility, automatically categorizing clinical images, and predicting post-transplant prognosis. This is enabled by the abundance of real-world longitudinal datasets derived from the extensive integration of electronic health record (EHR) [9-11]. For health care challenges involving large and complex datasets, especially those with unstructured data, ML methods have demonstrated advantages over traditional statistical regression methods [12,13]. At present, mainstream text recognition methods such as support vector machine (SVM) or random forest (RF) are mostly suitable for large sample size data with high positivity rates, and the recognition of bleeding events may be a process of searching for sporadic positive events in large sample sizes [14,15]. The aim of this study was to develop a machine learning model tool for identifying postoperative patients with major bleeding events based on the electronic medical record system.

Methods

Population

This study used a retrospective design. The data for this study were obtained from the National Health and Medical Big Data (Eastern) Center in Jiangsu Province, which is maintained by the Jiangsu Provincial Health Commission. This database includes clinical data from hospitals in Jiangsu province, with sensitive and identifiable information removed to protect privacy. Retrospective electronic medical records data, including clinical notes were used for identifying the study population. We specifically focused on patients with malignant tumor, who underwent surgical procedures between January 2018 and December 2021, reflecting actual clinical scenarios. To more accurately identify high-risk individuals for bleeding, we chose to focus on patients with chest, abdominal and gynecological malignant tumor instead of urinary system tumors. This decision was based on the fact that surgeries for chest, abdominal, and gynecological tumors involve greater surgical trauma, making postoperative bleeding more likely. Additionally, secondary tumors were excluded because they were typically palliatively resected, resulting in a lower risk of bleeding compared to primary tumors.

The primary objective was to develop a classifier capable of recognizing significant cases of bleeding with clinical relevance. To achieve this, we randomly allocated 1600 notes for training purposes and 400 notes for testing the model.

The inclusion criteria in the study population are as follows: (1) patients were diagnosed with primary chest, abdominal, or gynecological malignant tumors; (2) patients underwent surgical resection of the malignant tumor; and (3) patients were ≥18 years old. The exclusion criteria are as follows: (1) patients were diagnosed with urinary system tumors; (2) patients were diagnosed with a secondary tumor; (3) patients only underwent endoscopic surgery (ie, gastroscopy, enteroscopy, cystoscopy); (4) patients diagnosed with bleeding before surgery or who underwent surgery due to bleeding events; and (5) hospitalization course record for the given visit is missing.

Each note was classified as major bleeding present, indicating that clinically relevant bleeding was referenced in the note, or

XSL•FO
RenderX

major bleeding absent, indicating that clinically relevant bleeding was not referenced in the note. Major bleeding was defined as fatal or symptomatic bleeding in a critical area or organ or bleeding causing a fall in hemoglobin level of ≥2 g/dL or transfusion of ≥2 units of erythrocyte concentrate, according to the definition of the International Society of Thrombosis and Haemostasis [16].

We included medical history and physical examination notes, progress notes, and discharge summaries that were signed by an attending physician. All investigators with direct access to the data completed a data use agreement. An independent medical professional was selected to read the medical and surgical records and label whether there was bleeding, and the bleeding classification (ie, major bleeding or nonmajor bleeding). Then, another independent medical professional was designated to review the labeled cases and extract or highlight the text content in the medical records where bleeding events occurred for reference in developing bleeding-related regular expressions. Any disagreements or unclear content in the text reading between the two medical professionals were recorded and discussed in an external expert workshop. In the workshop, questionable clinical scenarios were confirmed by two experienced medical experts.

Feature Engineering

Given the low positivity rate and limited sample size of major bleeding events in the existing data, the methodology for identifying major bleeding events during manual annotation was carefully considered when crafting features. A significant portion of these features was extracted through the application of natural language processing (NLP) techniques like Jieba (for Chinese word splitting) and regular expressions (for text and value extraction). These features are categorized as follows:

The first category: These features were generated from the content of manually annotated major bleeding events in the training set. This set includes 20 key features strongly linked to postoperative bleeding, such as terms like “postoperative,” “ hematochezia,” “ hemorrhagic fluid,” and “treatment with hemostasis surgery.” Furthermore, 241 additional features were compiled by segmenting factors associated with major bleeding events.

The second category: These features were derived by tokenizing the patient’s course texts from the training set using Jieba and arranging them based on their frequency of occurrence. These features represent words that exhibit some relevance to major bleeding events but were not present in the first part. Features in this category were chosen based on a minimum frequency threshold of 900, aiming to complement the initial features and mitigate any shortcomings due to manually crafted features.

The third category: This feature reflects insights from clinical experts and primarily revolves around indicators like preoperative, intraoperative, and postoperative hemoglobin levels, bleeding volume, and transfusion volume. They are instrumental in establishing the logic for identifying postoperative major bleeding for each patient. Features in this segment heavily rely on quantitative logical assessments to fill in any gaps present in the preceding two feature groups. For instance, transfusion volumes extracted from structured surgical records and transfusion documentation are analyzed in conjunction with the standardized decision-making flowchart to determine major bleeding status. A detailed, structured process identification diagram is presented in Figure 1.

XSL•FO
RenderX

Figure 1. Structured process identification diagram for feature selection.

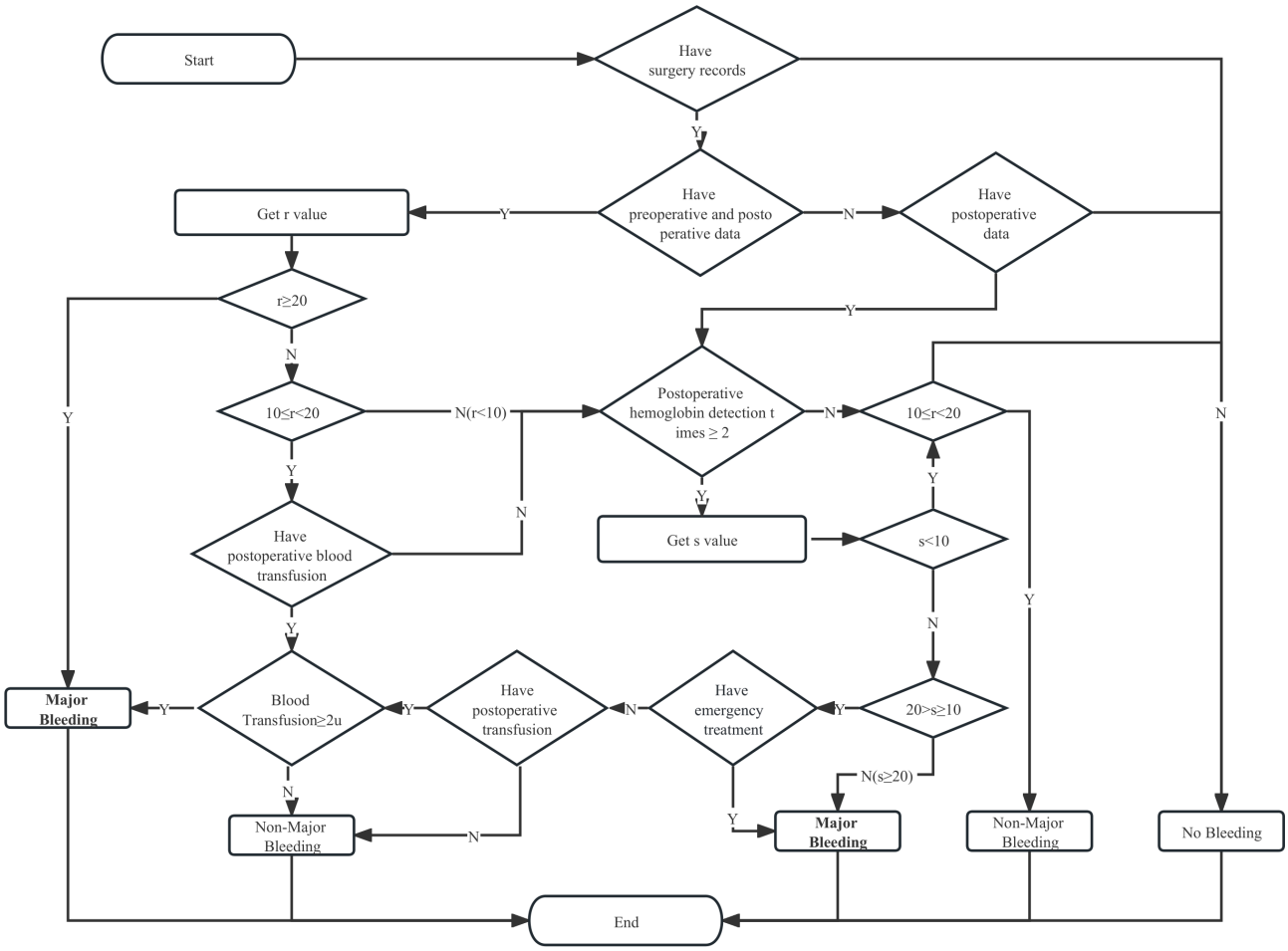

Model Training

A supervised learning approach was adopted, where a dataset comprising 2000 patients was divided into a training set of 1600 patients and a test set of 400 patients. Using the features generated through feature engineering and the corresponding labels assigned during manual review, three distinct ML models were successfully trained using different algorithms: logistic regression (LR), k-nearest neighbors (KNN), and convolutional neural networks (CNN).

Here is a breakdown of how each model was implemented:

1. LR:

- Model implementation: developed using Keras
- Architectural details: constructed with a single dense layer
- Activation function: sigmoid function
- Loss function: SigmoidFocalCrossEntropy
- Optimizer: Adam optimizer
- Validation split: 20% of the training dataset is set aside for validation

1. KNN:

- Model implementation: Used the KNeighborsClassifier class from Scikit-Learn.

1. CNN:

- Model implementation: implemented using Keras
- Architectural details: featured 3 hidden layers - a 1-dimensional convolutional layer, a max pooling layer, and a dense layer
- Activation functions: Rectified Linear Unit for the convolutional layer and sigmoid for the output layer
- Output layer: a single dense unit
- Validation split: Again, 20% of the training data is reserved for validation.

All three ML models were trained on the training set consisting of 1600 cases to learn the underlying patterns between the extracted features and the corresponding labels. The use of LR, KNN, and CNN allows for a diverse exploration of the dataset with different algorithms to harness their unique strengths in handling and learning from the data.

Model Evaluation

Following the training phase, the evaluation of the three ML models—LR, KNN, and CNN—was conducted on the test set containing 400 cases. Various performance metrics were calculated for each model to assess their effectiveness in making predictions. These metrics included accuracy, sensitivity,

XSL•FO
RenderX

positive predictive value (PPV), F_1-score, negative predictive value (NPV), and specificity of each model.

After calculating these metrics, the ML models with the highest sensitivity and specificity were chosen for further consideration due to their critical role in the diagnostic accuracy required by the specific use case. By prioritizing both sensitivity and specificity, you ensured a balanced assessment of the models' performance regarding true positive and true negative rates.

In addition to the confusion matrices, which provide a detailed overview of true positives, false negatives, false positives, and true negatives at the default prediction threshold, we used the receiver operating characteristic curve and the corresponding area under the curve for each model. This comprehensive analysis of the receiver operating characteristic curves and area under the curve values allowed for a deeper understanding of the trade-offs between true positive and false positive rates across different thresholds and enabled the selection of the best-performing models based on a more nuanced evaluation beyond conventional accuracy metrics.

Ethical Considerations

The study protocol was conducted in accordance with the Declaration of Helsinki and approved by the institutional review board at Shanghai Ethics Committee for Clinical Research (Approval number: SECCR/2023-119-01). The data used in this study were deidentified. Informed consent was waived by the Ethics Committee owing to the use of deidentified data.

Results

Baseline Characteristics

The training set represented 1600 patients, of whom 48.5% (776) were female, the mean age was 62.86 (SD 9.16) years, with a total of 33654 course records. The test set represented 400 patients, of whom 49.3% (n=197) were female; the mean age was 62.81 (SD 9.21) years, with a total of 8491 course records. The ratio of postoperative major bleeding was 4.31% (n=69) in the training set and 4.75% (n=19) in the test set (Table 1).

Table . Population characteristics of training and test sets.

Characteristics	Training set (n=1600)	Test set (n=400)
Gender (female) n (%)	776 48.5	197 49.3
Age (years), mean (SD)	62.86 (9.16)	62.81 (9.21)
Course records, n	33,654	8491
Proportion of major bleeding, n (%)	69 (4.31)	19 (4.75)

Note-Based Feature Selection

A total of 270 features were selected; first, 261 features related to postoperative bleeding were selected by regular manual disassembly. Second, Jieba was used to segment the disease course text and select 8 features according to word frequency classification. Furthermore, the last feature was created by a logical recognition graph generated from expert opinion. More details are provided in the feature engineering section. The frequency of the top 20 features is shown in Table 2.

XSL•FO
RenderX

Table . The frequency of the top 20 features selection.

Features (description or regularization in English)	Frequency
Postoperative	1592
(chest | abdomen | pelvis) cavity | rectum. {0,2} depression | posterior fornix | anastomosis | pancreas | digestive tract | stomach | vagina	1579
(out | lose) blood (stop | break into)?| hematoma | subcutaneous Ecchymosis | congestion	1033
Bleeding	1002
Bloody fluid | red.{0,2} drainage fluid	995
Hemostasis	528
Laparotomy	513
Dark red	347
Emergency	343
(To | Perform).{0,15} to stop bleeding	303
Swelling	273
(To | Perform).{0,5} to stop bleeding	253
(((chest | abdomen | pelvis) cavity | rectum.{0,2} depression | posterior fornix | anastomosis | pancreas | digestive tract | stomach).{0,10}bleeding | vagina.{0,2}large .{0,2}(out | lose)blood)	251
Blood loss	246
Pelvic cavity.{0,5}Drainage.{0,5}(bloody liquid | fluid dark red)	225
(?: inject | give).{0,20}(?:suspended less white (?:red blood cell | erythrocyte blood transfusion volume | red blood cell)|whole blood | red blood cell | less white suspended red blood cell | red suspension | suspended less white Red).{1,20}(?:U | ML | CC | ml | u | ml | mL | cc | infusion | unit) | (?: in | out | to).{0,5}(? :U |ML |CC | milliliter | u | ml | mL | cc | infusion | unit).{0,5}(?: suspended less white(?:cell red blood cell | red blood cell transfusion amount | red blood cell) | whole blood | Red blood cells | less white suspension red blood cells | red suspension | suspension less white red)	220
((out | loss) blood | introduction | bloody Effusion).{0,15}(U | ML | CC | ml | u | ml | mL | cc | unit)	220
Lose.{0,20}(red blood cells | whole blood | red suspension)	220
Structured recognition results	206
(exist | has).{0,5} bleed	178

Comparison and Verification of the Efficiency of Three Machine Learning Models

In the identification of major bleeding events within the test set, the LR method had an accuracy of 0.8275, sensitivity of 0.8947, specificity of 0.8241, PPV of 0.2024, NPV of 0.9937, and F_1-score of 0.3301. The accuracy of the CNN method in the testing set was 0.8900, the sensitivity was 0.8421, the specificity was 0.8924, the PPV was 0.2807, the NPV was 0.9913, and the F_1-score was 0.4211. The KNN method had an accuracy of 0.9575, sensitivity of 0.2105, specificity of 0.9948, PPV of 0.6667, NPV of 0.9619, and F_1-score of 0.3200 (Table 3). The C-statistic was higher in the LR method (C=0.9018), followed by the CNN method (C=0.8830) (Figure 2).

XSL•FO
RenderX

Table . The performance of each model in identifying major bleeding events.

Sets and models	Accuracy	Sensitivity	Positive predictive value	F_1-score	Negative predictive value	Specificity
Training set (n=1600)						
LR[a]	0.8231	1.0000	0.1960	0.3278	1.000	0.8152
CNN[b]	0.9056	0.9710	0.3102	0.4702	0.9986	0.9027
KNN[c]	0.9643	0.1884	0.9286	0.3132	0.9647	0.9993
Test set (n=400)						
LR	0.8275	0.8947	0.2024	0.3301	0.9937	0.8241
CNN	0.8900	0.8421	0.2807	0.4211	0.9913	0.8924
KNN	0.9575	0.2105	0.6667	0.3200	0.9619	0.9948

[a]LR: logistic regression.

[b]CNN: convolutional neural network.

[c]KNN: K-nearest neighbor.

Figure 2. Receiver operating characteristic curves for identifying major bleeding from clinical notes using the LR and CNN methods. AUC: area under the curve; LR: logistic regression; CNN: convolution neural network.

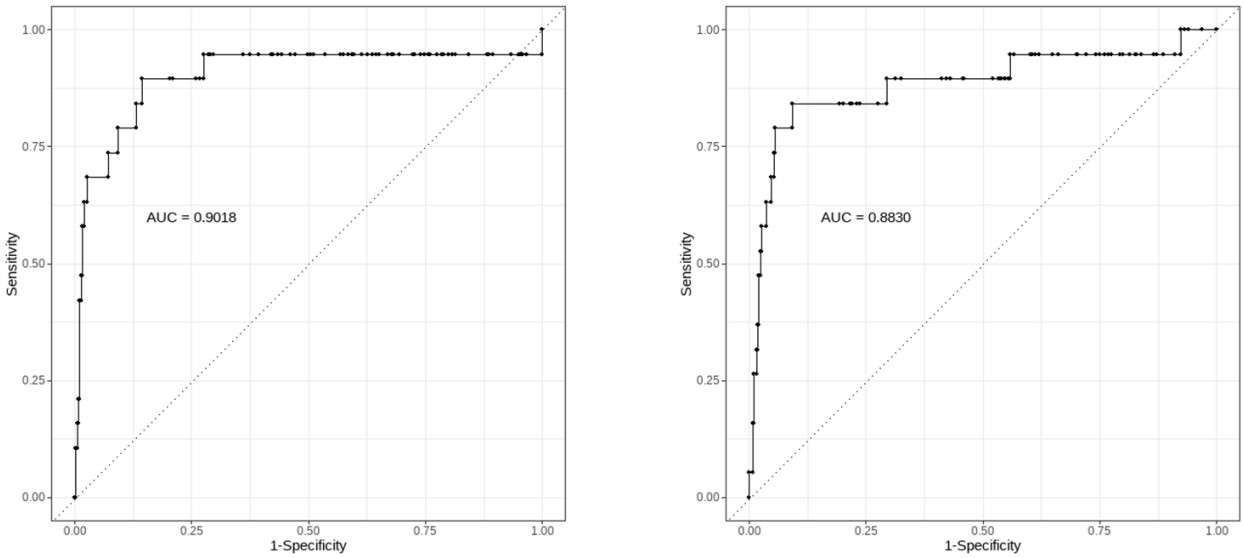

LR CNN

Discussion

Principal Findings

In this study, we implemented LR, CNN, and KNN algorithms for ML detection of bleeding events within electronic medical record systems. Notably, CNN was specifically applied to categorize radiology free-text reports and exhibited commendable accuracy [17]. The CNN approach demonstrated a sensitivity of 84% and a specificity of 89% in detecting recorded bleeding events in patients at a threshold of 0.50, while the LR method showcased a sensitivity of 89% and a specificity of 82% at the same threshold. The KNN algorithm can be negatively impacted by high dimensionality in feature matrices. These findings underscore the viability of employing algorithms to pinpoint patients with bleeding events within extensive textual disease records.

Evaluation Metrics and Application Context

F_1-scores are frequently used in the evaluation of classifiers. The F_1-score can be considered the harmonic mean between the precision and recall (sensitivity) [18]. Thus, it symmetrically represents both precision (how accurately the model identifies only true positive instances) and sensitivity (how well the classifier identifies all actual positive instances) in one metric. Although the three machine learning models in this investigation did not attain high F_1-scores on the test set, it must be clarified that the goal of the investigation was to detect bleeding events as accurately as possible in practical applications for subsequent manual assessment. Therefore, greater attention is paid to the sensitivity and specificity (ie, how well the classifier identifies all actual negative instances) of the model, the high values of which imply that the model can detect hemorrhagic events while avoiding false alarms, which is crucial for this investigation. Consequently, priority is placed on the models' sensitivity and the attainment of greater specificity. Hence, we maintain that

XSL•FO
RenderX

the models developed in this study remain valuable for identifying bleeding events within disease records and represent a promising avenue for future research endeavors.

Comparison to Prior Work

The integration of NLP and ML has proven successful across various domains. For instance, in the context of health care, predictive models leveraging ML and statistical methods have demonstrated the ability to forecast occurrences such as postpartum hemorrhage upon labor admission with reasonable discriminatory power. A prior study demonstrated the effectiveness of a Hybrid CNN-LSTM Autoencoder model for the detection of bleeding events within EHR data. This was accomplished through the integration of a supervised CNN with a pretrained, unsupervised Bidirectional Long Short-Term Memory autoencoder. The primary objective was to accurately predict the presence of a bleeding event within a given English sentence from an EHR record [19].

While prior research primarily focused on English EHR data and general bleeding event detection, such as the study by Li et al that used a Hybrid CNN-LSTM Autoencoder for sentence-level bleeding detection, and a more recent work that applied retrieval augmented generation with large language models for detecting nonsurgical major bleeding events in English EHRs [20], our investigation centers on the identification of major bleeding events within Chinese EHR. Chinese NLP presents inherent complexities, especially in word segmentation within EHR text, posing a significant hurdle. Furthermore, the shift in focus to major bleeding, which exhibits a considerably lower incidence rate compared to general bleeding events, substantially reduces the availability of positive samples, thereby intensifying the challenge for robust predictive modeling. Building upon these complexities, unlike the prior studies' sentence-level analysis, our research takes a holistic approach to identifying major bleeding events within the entire patient visit context. This broader, patient-centric perspective further complicates the analysis and requires a more comprehensive understanding of the clinical narrative. These critical distinctions—language specificity, event granularity, and analytical scope—collectively highlight the greater challenges of our study compared to existing work in the field.

Strengths and Limitations

A key strength of this study lies in the strategic approach to model selection and feature engineering. Initially, SVM and RF algorithms were explored. However, recognizing their reliance on effective word segmentation and the challenges posed by the heterogeneous nature of the text data, the focus was deliberately shifted to LR, KNN, and CNN. To further enhance model performance, term frequency-inverse document frequency (TF-IDF) scoring was effectively used to compute feature representations, addressing the complexities of free-text data. Moreover, to overcome the limitation of a relatively small sample size, particularly the scarcity of positive samples (fewer than 100), and the resulting dispersed TF-IDF features, a targeted feature engineering approach was strategically adopted. This involved selecting predefined features associated with bleeding based on expert medical domain knowledge. This approach allowed the ML models to concentrate on clinically relevant features, represented as binary values (0 and 1), rather than solely relying on traditional TF-IDF metrics applied to the entire text. This targeted feature engineering proved crucial for enhancing performance in a limited data setting. As a result of these strengths, including strategic model selection, targeted feature engineering, manual annotation, and training with limited samples, the LR and CNN models demonstrated superior performance compared to SVM and RF models based on TF-IDF features for detecting major bleeding events. This aligns with findings from similar research, as evidenced by a study that underscored the enhanced performance of CNNs when expert annotation of text data is incorporated [19], further validating the efficacy of CNNs in text classification tasks within the medical domain.

However, manual feature selection has certain limitations. First, engineers must dedicate time to writing code for feature extraction based on the definitions provided by the medical field, which is a more time-intensive process compared to the automatic feature computation performed by machines. To partially mitigate this, we prioritized high-impact features identified through clinical expert consultations, but future studies could integrate semiautomated pipelines to balance efficiency and domain specificity. Second, as previously discussed in related research [21], the identification of postoperative bleeding and feature selection depends on manual procedures, thereby increasing the likelihood of bias and oversight. While dual annotations by multiple clinicians were used to reduce subjectivity, discrepancies were resolved through consensus rather than quantitative metrics, potentially affecting reproducibility. Future work should adopt standardized annotation protocols with inter-rater reliability assessments. Third, the chosen features encapsulate domain-specific knowledge, sometimes being closely tied to specific hospital departments with varying requirements and documentation practices in electronic medical records. Broader applicability requires multicenter collaboration to harmonize feature definitions across institutions. Moreover, our study included a retrospective analysis with a relatively limited number of specimens, a factor to be considered for broader applicability in ML analyses. Similar challenges have been observed in other comparable studies [19]. Consequently, not only does this constrained universality impede application in diverse fields, but also when extending to other illnesses, hospitals, and departments, there is often a need to reimagine and recreate the features.

Conclusions

Based on our new text feature selection method, both the LR and CNN methods perform well in identifying major bleeding occurring in postoperative patients with malignant tumors from electronic medical records.

XSL•FO
RenderX

Acknowledgments

Medical writing support, under the direction of the authors, was provided by Hao Hu (Shanghai Palan DataRx Co.,Ltd) and funded by Sanofi. This study was sponsored by Sanofi.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to patient privacy concerns but are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: H Li, HY, YG, H Luo, CC, ZZ, MY
Data curation: WJ
Formal analysis: H Luo, HY, YG
Methodology: H Li, HY, YG, H Luo, CC, ZZ, MY
Supervision: H Li
Writing – original draft: HY, YG, H Luo, WJ
Writing – review & editing: H Li, CC, ZZ, MY

Conflicts of Interest

The authors individually and collectively are responsible for all content and editorial decisions and received no payment from Sanofi related to the development/presentation of this publication. CC, ZZ are Sanofi employees and may hold shares and/or stock options in the company. MY was previously a Sanofi employee and is now employed at Beijing Baheal Chengchuang Pharmaceutical Investment Co., Ltd. The other authors declare that there is no conflict of interest.

References

1. Tasu JP, Vesselle G, Herpe G, et al. Postoperative abdominal bleeding. Diagn Interv Imaging 2015;96(7-8):823-831. [doi: 10.1016/j.diii.2015.03.013] [Medline: 26078019]
2. Maxwell MJ, Wilson MJA. Complications of blood transfusion. Continuing Education in Anaesthesia Critical Care & Pain 2006 Dec;6(6):225-229. [doi: 10.1093/bjaceaccp/mkl053]
3. Berger JS, Bhatt DL, Steg PG, et al. Bleeding, mortality, and antiplatelet therapy: results from the Clopidogrel for High Atherothrombotic Risk and Ischemic Stabilization, Management, and Avoidance (CHARISMA) trial. Am Heart J 2011 Jul;162(1):98-105. [doi: 10.1016/j.ahj.2011.04.015]
4. Cook DJ, Griffith LE, Walter SD, et al. The attributable mortality and length of intensive care unit stay of clinically important gastrointestinal bleeding in critically ill patients. Crit Care 2001 Dec;5(6):368-375. [doi: 10.1186/cc1071] [Medline: 11737927]
5. Khorana AA, Mackman N, Falanga A, et al. Cancer-associated venous thromboembolism. Nat Rev Dis Primers 2022 Feb 17;8(1):11. [doi: 10.1038/s41572-022-00336-y] [Medline: 35177631]
6. Anderson DR, Morgano GP, Bennett C, et al. American Society of Hematology 2019 guidelines for management of venous thromboembolism: prevention of venous thromboembolism in surgical hospitalized patients. Blood Adv 2019 Dec 10;3(23):3898-3944. [doi: 10.1182/bloodadvances.2019000975]
7. Taggart M, Chapman WW, Steinberg BA, et al. Comparison of 2 natural language processing methods for identification of bleeding among critically ill patients. JAMA Netw Open 2018 Oct 5;1(6):e183451. [doi: 10.1001/jamanetworkopen.2018.3451] [Medline: 30646240]
8. Corey KM, Kashyap S, Lorenzi E, et al. Development and validation of machine learning models to identify high-risk surgical patients using automatically curated electronic health record data (Pythia): a retrospective, single-site study. PLoS Med 2018;15(11):e1002701. [doi: 10.1371/journal.pmed.1002701]
9. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med 2018;1:18. [doi: 10.1038/s41746-018-0029-1] [Medline: 31304302]
10. Weller GB, Lovely J, Larson DW, Earnshaw BA, Huebner M. Leveraging electronic health records for predictive modeling of post-surgical complications. Stat Methods Med Res 2018 Nov;27(11):3271-3285. [doi: 10.1177/0962280217696115] [Medline: 29298612]
11. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J 2015;13:8-17. [doi: 10.1016/j.csbj.2014.11.005]
12. Murdoch TB, Detsky AS. The inevitable application of big data to health care. JAMA 2013 Apr 3;309(13):1351-1352. [doi: 10.1001/jama.2013.393] [Medline: 23549579]
13. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature New Biol 2015 May 28;521(7553):436-444. [doi: 10.1038/nature14539] [Medline: 26017442]

XSL•FO
RenderX

14. Delate T, Jones AE, Clark NP, Witt DM. Assessment of the coding accuracy of warfarin-related bleeding events. Thromb Res 2017 Nov;159:86-90. [doi: 10.1016/j.thromres.2017.10.004] [Medline: 29035718]
15. Wilchesky M, Tamblyn RM, Huang A. Validation of diagnostic codes within medical services claims. J Clin Epidemiol 2004 Feb;57(2):131-141. [doi: 10.1016/S0895-4356(03)00246-4] [Medline: 15125622]
16. Schulman S, Angerås U, Bergqvist D, et al. Definition of major bleeding in clinical investigations of antihemostatic medicinal products in surgical patients. J Thromb Haemost 2010 Jan;8(1):202-204. [doi: 10.1111/j.1538-7836.2009.03678.x] [Medline: 19878532]
17. Chen MC, Ball RL, Yang L, et al. Deep learning to classify radiology free-text reports. Radiology 2018 Mar;286(3):845-852. [doi: 10.1148/radiol.2017171115] [Medline: 29135365]
18. Rai T, Morisi A, Bacci B, et al. Keeping pathologists in the loop and an adaptive F1-score threshold method for mitosis detection in canine perivascular wall tumours. Cancers (Basel) 2024 Feb 2;16(3):644. [doi: 10.3390/cancers16030644] [Medline: 38339394]
19. Li R, Hu B, Liu F, et al. Detection of bleeding events in electronic health record notes using convolutional neural network models enhanced with recurrent neural network autoencoders: deep learning approach. JMIR Med Inform ;7(1):e10788. [doi: 10.2196/10788]
20. Kaplinsky P, Singh R, Fusillo TF, Leader A, Zwicker JI, Mantha S. Retrieval augmented generation for the detection of major bleeding events in the electronic health record. Blood 2024 Nov 5;144(Supplement 1):2263-2263. [doi: 10.1182/blood-2024-203911]
21. Jing X, Wang X, Zhuang H, Fang X, Xu H. Multiple machine learning approaches based on postoperative prediction of pulmonary complications in patients with emergency cerebral hemorrhage surgery. Front Surg 2021;8:797872. [doi: 10.3389/fsurg.2021.797872] [Medline: 35127804]

Abbreviation

CNN: Convolutional neural network
KNN: K-nearest neighbor
LR: Logistic regression
ML: Machine learning
NLP: natural language processing
NPV: negative predictive value
PPV: positive predictive value
RF: random forest
SVM: support vector machine
VTE: venous thrombosis

Edited by A Mavragani; submitted 05.09.24; peer-reviewed by NH Le, Y Zhang; revised version received 25.03.25; accepted 07.04.25; published 01.05.25.

Please cite as:
Li H, Yao H, Gao Y, Luo H, Cai C, Zhou Z, Yuan M, Jiang W
Identification of Major Bleeding Events in Postoperative Patients With Malignant Tumors in Chinese Electronic Medical Records: Algorithm Development and Validation
JMIR Form Res 2025;9:e66189
URL: https://formative.jmir.org/2025/1/e66189
doi:10.2196/66189

© Hui Li, Haiyang Yao, Yuxiang Gao, Hang Luo, Changbin Cai, Zhou Zhou, Muhan Yuan, Wei Jiang. Originally published in JMIR Formative Research (https://formative.jmir.org), 1.5.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Mobile Health App for Adolescent Asthma Self-Management: Development and Usability Study of the Pulmonary Education and Knowledge Mobile Asthma Action Plan

Xing He[1,2], PhD; Jiang Bian[1,2], PhD; Ariel Berlinski[3,4], MD; Yi Guo[5], PhD; A Larry Simmons[3,6], MD; S Alexandra Marshall[7], MPH, PhD; Carolyn J Greene[8], PhD; Rita Hudson Brown[3,4], BA, ACRP; Jessica Turner[4], MPH; Tamara T Perry[3,4], MD

[1]Department of Biostatistics and Health Data Science, Indiana University, Indianapolis, IN, United States
[2]Regenstrief Institute, Indianapolis, IN, United States
[3]Department of Pediatrics, University of Arkansas for Medical Sciences, Little Rock, AR, United States
[4]Arkansas Children's Research Institute, 13 Children's Way, Little Rock, AR, United States
[5]Department of Health Outcomes & Biomedical Informatics, University of Florida, Gainesville, FL, United States
[6]Arkansas Children's Hospital, Little Rock, AR, United States
[7]Department of Health Behavior & Health Education, University of Arkansas for Medical Sciences, Little Rock, AR, United States
[8]United States Department of Veterans Affairs, Mann-Grandstaff VA Medical Center, Spokane, WA, United States

Corresponding Author:
Tamara T Perry, MD
Department of Pediatrics, University of Arkansas for Medical Sciences, Little Rock, AR, United States

Abstract

Background: Adolescents with asthma are vulnerable to poor asthma outcomes due to inadequate self-management skills and nonadherence to medications. Mobile health (mHealth) apps have shown promise in improving asthma control, medication adherence, and self-efficacy. However, existing mHealth asthma apps lack personalization and real-time feedback and are not tailored for at-risk adolescents.

Objective: This study aimed to design, develop, and test a smartphone-based mHealth Asthma Action Plan for adolescents, called Pulmonary Education and Knowledge Mobile Asthma Action Plan (PEAK-mAAP), in preparation for a large-scale randomized controlled trial.

Methods: We employed user-centered design principles to develop our app, leveraging our previous work and following guidelines from the National Heart, Lung, and Blood Institute. The app consists of a patient-facing mobile app and a provider-facing portal. A convenience sample of 13 adolescents (aged 12 - 20 years) was recruited from the Arkansas Children's Research Institute database or direct health care provider referrals. Participants underwent a task-based usability assessment followed by the System Usability Scale assessment to measure user satisfaction, interface effectiveness, and overall system usability.

Results: PEAK-mAAP integrates 7 core modules supporting personalized asthma self-management, symptom monitoring, medication tracking, and real-time feedback. The mean System Usability Scale score was 83/100 (SD 5.54), indicating high user satisfaction and system usability. Notably, older adolescents (>17 years) reported higher usability scores (87.5) than younger users (77.5), suggesting potential age-related differences in app navigation and engagement.

Conclusions: The results demonstrate that PEAK-mAAP is a feasible and user-friendly mHealth intervention for adolescent asthma self-management. While the high usability score reflects a positive user experience, some participants encountered initial usability challenges, highlighting the need for minor refinements and user training materials. The integration of personalized self-management tools and real-time feedback distinguishes PEAK-mAAP from existing asthma apps, addressing key barriers to adherence and engagement. Moving forward, an ongoing randomized controlled trial will assess its clinical effectiveness, long-term engagement, and impact on asthma outcomes, providing further insights into its potential as a scalable solution for adolescent asthma care.

(JMIR Form Res 2025;9:e64212) doi:10.2196/64212

KEYWORDS

asthma self-management; mobile health; user-centered design; usability; adolescents; mobile phone

XSL•FO
RenderX

Introduction

According to the Global Asthma Report 2022 [1], asthma impacts 262 million individuals globally across all ages. The report highlighted that 1 in 10 children have asthma symptoms, and half have poorly controlled symptoms. Like most chronic diseases, a proactive way to mitigate the cost and adverse effects caused by asthma is to equip patients with effective self-management strategies [2]. Earlier investigations of mobile health (mHealth) asthma management apps show a positive effect on asthma control, medication adherence, and self-efficacy [3,4]. mHealth technology that combines personalized health management into patients' daily routines can effectively enhance patients' asthma self-management [4]. For instance, Farzandipour et al [5] designed a smartphone-based asthma self-management app with educational content, data recording, reminders, and physician communication features, which was rated as "good" in usability and positively impacted patients' self-management knowledge. Kosse et al [6] found that an mHealth intervention significantly improved medication adherence in adolescents with asthma who had poor adherence rates, suggesting that health care providers should consider a tailored mHealth approach in their treatment. Fedele et al [7] conducted a randomized controlled trial (RCT) to test the feasibility and efficacy of a mHealth intervention to enhance caregiver support and asthma self-management in early adolescents, revealing clinically significant improvements in asthma control. Nevertheless, existing mHealth asthma apps lack personalization and real-time feedback, and are not tailored for at-risk adolescents. Furthermore, recent review highlights a significant lack of large-scale RCTs evaluating the efficacy and sustainability of eHealth and mHealth interventions in adolescents and young adults with chronic diseases, including asthma [8].

Adolescents are at high risk of adverse asthma outcomes due to multiple factors, such as (1) their self-management skills are often insufficient, which may lead to nonadherence to medication; and (2) adolescence is typically when there is a transition from dependence on parental supervision to medical autonomy [9]. Current practices should ensure adolescents have the correct knowledge, beliefs, and positive attitudes to self-manage their illness [10]. National Heart, Lung, and Blood Institute (NHLBI) recommends using Asthma Action Plans (AAPs) to facilitate asthma self-management [11]. In our previous studies [12,13], we developed a smartphone-based mobile health Asthma Action Plan (mAAP) tailored for adolescents. Our findings demonstrated its feasibility as a self-management tool, significantly benefiting adolescents with uncontrolled asthma. Leveraging the insights from these feasibility studies, we adopted a user-centered design (UCD) [14] approach, collaborating with a diverse group of stakeholders (ie, adolescents with asthma, their parents, clinicians, nurses, and researchers) to iteratively redesign and re-engineer the mAAP and named it, Pulmonary Education and Knowledge Mobile Asthma Action Plan (PEAK-mAAP). Critically, we designed PEAK-mAAP to address specific barriers to adopting health behaviors based on the well-established health belief model (HBM) [15]. Built on the Apache Cordova (Adobe Systems) [16] mobile app framework, PEAK-mAAP has cross-platform compatibility to meet the needs of Android (Google), iOS (Apple Inc), and web users. Complementing PEAK-mAAP, we introduced a web portal for health providers and guardians, allowing them to oversee the adolescents' engagement with PEAK-mAAP. In this paper, we detail the design, development, and usability assessment of PEAK-mAAP.

Methods

Design of PEAK-mAAP

We designed the functionalities of PEAK-mAAP based on the HBM to affect behavior change (ie, asthma self-management based on prescribed AAP). Figure 1 depicts how the design of PEAK-mAAP addresses the different constructs of the HBM. Notably, the HBM integrates considerations of individual demographics and psychological traits—termed modifying variables. These variables indirectly influence health-related actions by impacting factors like health motivation, perceived risks, severity, benefits, and obstacles.

After the preliminary design of PEAK-mAAP, we engaged in iterative enhancements by involving all stakeholders through a UCD approach. We convened a community advisory board with our intended end users, that is, adolescent (ie, teens and emerging adults aged between 12 and 20) patients with asthma, their caregivers, asthma care providers, UCD specialists, and researchers. community advisory board members provided feedback on the features of mAAP and iterative feedback and beta testing of the final app.

XSL•FO
RenderX

Figure 1. Design of Pulmonary Education and Knowledge Mobile Asthma Action Plan according to the health belief model.

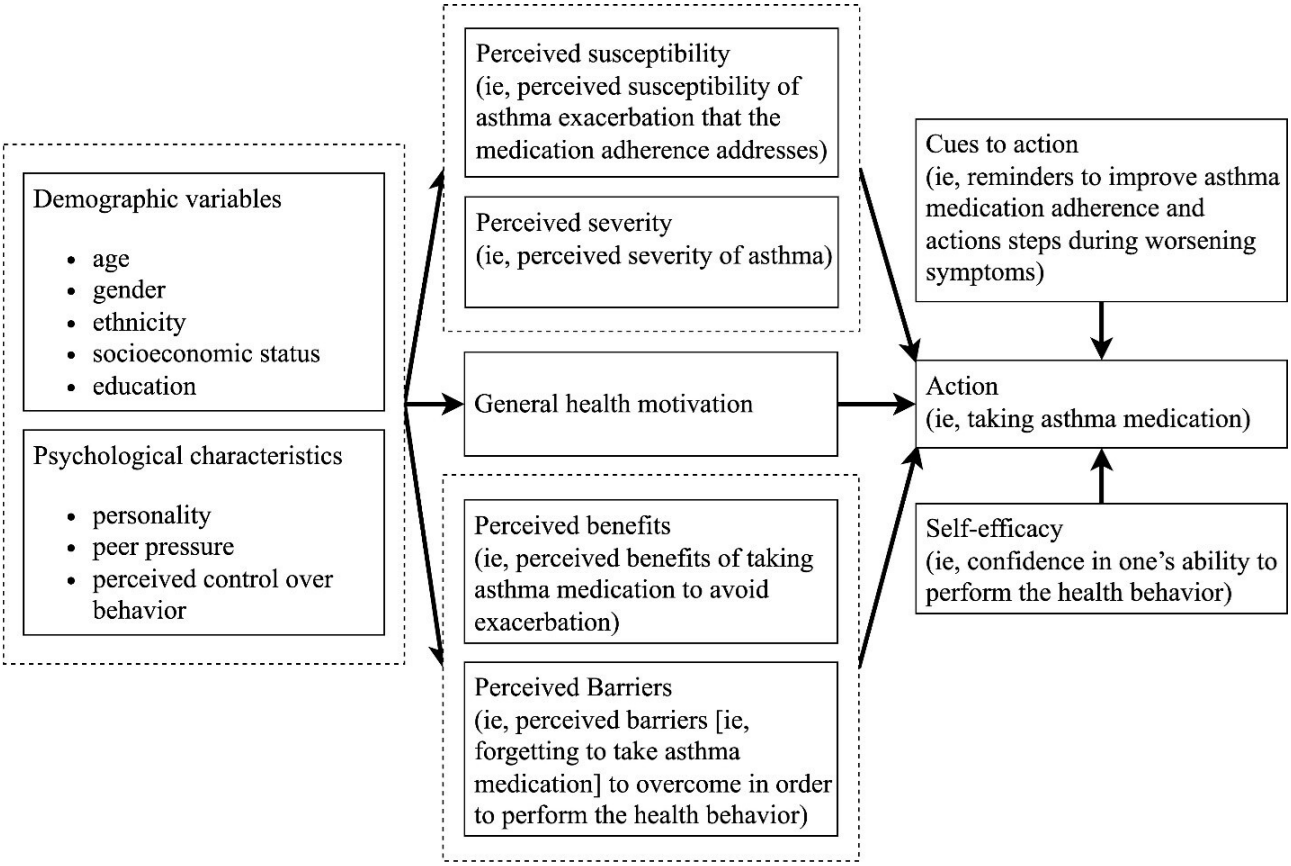

Implementation of PEAK-mAAP

PEAK-mAAP is a hybrid mobile app developed using standard web technologies and housed within a native container—specifically Apache Cordova—for adolescents. The web portal, designed for health care providers and guardians of adolescents, is built with Ruby on Rails [17]. Figure 2 illustrates the 5-component system architecture that supports the functionality of the app and the web portal.

Figure 2. The 5-component system architecture which supports Pulmonary Education and Knowledge Mobile Asthma Action Plan and the web portal. API: application programming interface; REDCap: Research Electronic Data Capture.

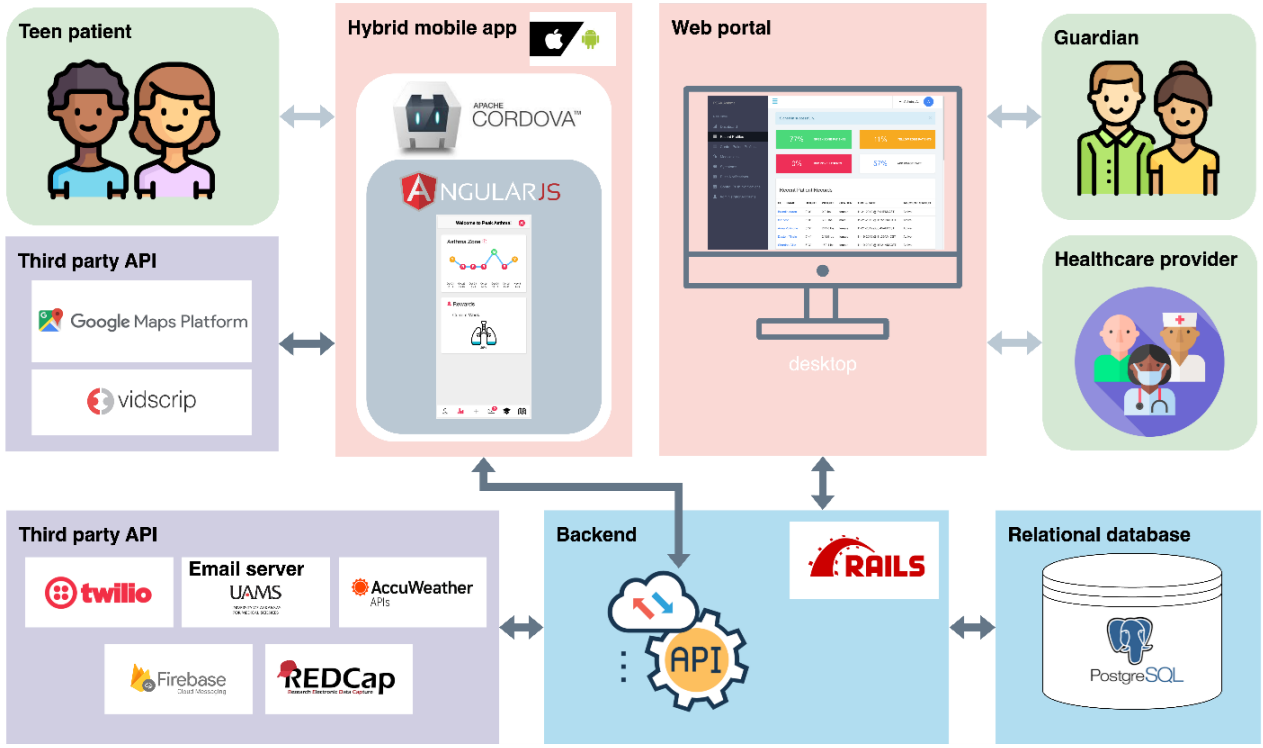

XSL•FO
RenderX

First, a relational database, PostgreSQL [18], an advanced open-source object-relational database system, is employed to manage and categorize a range of data, including patient demographic details, guardian and provider profiles, medication records, and specific activity logs such as peak flow recordings. In addition, the database houses preset content, like educational tips and instructional videos. To safeguard against any potential data loss, regular backups are performed.

Second, our backend, built on the robust Ruby on Rails framework, offers RESTful application programming interfaces (APIs) to bridge our hybrid mobile app with its backend database. It also integrates with 5 external APIs, like weather, to enrich the mobile app's features. Besides, adhering to web development best practices, this backend employs the model-view-controller design pattern to support the web portal. The web portal is tailored for adolescent health care providers such as doctors, nurses, and researchers, as well as guardians like parents and grandparents, enabling them to monitor PEAK-mAAP usage effectively.

Third, we implemented PEAK-mAAP as a hybrid mobile app to aid adolescents in managing their asthma. This approach involved the development of a cross-platform web app, encapsulated within a native app container. The hybrid mobile app was crafted using Apache Cordova and AngularJS (Google) [19] (ie, an open-source JavaScript framework primarily used for building dynamic single-page web apps). This allows programmers to use standard web technologies such as HTML5, CSS3, and JavaScript for cross-platform development.

Fourth, we engineered a model-view-controller–based web app to facilitate monitoring adolescents' mobile app usage by both health care providers and guardians. Health care providers can access a comprehensive view of all patients' information, mobile app usage, and asthma-related activities through a specialized web portal. Concurrently, guardians are provided a tailored view allowing them to observe their children's information and interactions with the PEAK-mAAP, reinforcing transparency and personalized care.

Fifth, to enhance our mobile app for adolescents' asthma self-management, we integrated 7 third-party APIs. Directly, we incorporated the Google Maps [20] API for locating hospitals and pharmacies and Vidscrip (Clear.md Inc) [21] for educational asthma videos. Indirectly, through our Ruby on Rails backend, we added the Twilio [22] API for alert calls and texts during severe asthma exacerbation; an email server hosted by the University of Arkansas for Medical Sciences to notify health care providers and guardians; the AccuWeather [23] API for real-time weather and pollen levels; the Firebase Cloud Messaging [24] API for cross-platform notifications; and the REDCap (Research Electronic Data Capture; Vanderbilt University) [25,26] API for automated monthly survey distributions. These integrations ensure a holistic approach to asthma management for adolescents.

Usability Assessment

The PEAK-mAAP usability assessment was conducted over two months, spanning January and February 2019, before the formal enrollment of research participants in the subsequent RCT. Usability assessment is a widely applied method across various industries, including aviation, automotive, and software development, aimed at enhancing system functionality by assessing an app's effectiveness and efficiency. Effectiveness refers to the extent to which users can successfully accomplish tasks and achieve their goals, while efficiency evaluates the time and effort required to complete a given task.

Participants were selected using a convenience sampling method and a total of 13 individuals, ranging from 12 to 20 years old in age, were recruited. All participants had a confirmed clinical diagnosis of asthma, as documented in their medical records. Recruitment efforts targeted potential candidates through 2 primary channels: the Arkansas Children's Research Institute database, which provided access to a broad pool of pediatric patients with existing asthma diagnoses, and direct referrals from health care providers familiar with eligible patients. To qualify for the usability assessment, participants were required to meet specific eligibility criteria designed to ensure they could effectively engage with the PEAK-mAAP system. Inclusion criteria consisted of (1) a current diagnosis of asthma, verified by a health care professional; (2) previous exposure to or familiarity with AAPs, to facilitate comparative feedback on the mobile app; (3) proficiency in English to ensure clear communication during the assessment and accurate comprehension of the app's content; and (4) the absence of any visual or auditory impairments that could hinder the ability to interact with the app's interface. Individuals were excluded if they had significant respiratory conditions other than asthma, such as cystic fibrosis or chronic obstructive pulmonary disease, which could introduce variability in the management strategies assessed. In addition, participants with developmental delays or cognitive impairments that could interfere with their ability to comprehend instructions, navigate the app, or provide reliable feedback were not eligible for the study. The sample size of 13 was deemed sufficient, adhering to the standard usability principle that even 5 participants can identify the most severe usability issues and 80% of all usability issues [27].

To ensure consistency, all participants were provided with the same iPad (Apple Inc) device for testing and watched a prerecorded demonstration video of the PEAK-mAAP app. Following the video, participants were instructed to explore the app using the study iPad to familiarize themselves with its features. Subsequently, they were asked to complete eight tasks using the app: (1) logging Peak Flow, (2) logging Symptoms, (3) setting a reminder for logging their Peak Flow or symptoms, (4) reading daily tips and viewing educational videos, (5) reading notifications, (6) describing directions or locations of the hospital or pharmacy displayed on the map, (7) describing the weather for patient's location, and (8) reporting number of days they logged and rewards within the app. A stopwatch was used to time each user while performing each task or scenario. Any extra steps performed during the scenarios were recorded to determine if there were common mistakes or additional steps taken for any of the app's features. During this process, users were asked to adopt the think-aloud [28] process by verbalizing any trouble they had with the app, features they found exciting or beneficial, as well as features that were confusing or not useful. This approach provided evaluators with deeper insights

XSL•FO
RenderX

beyond those obtainable through standard cognitive walkthroughs. There were at least 2 evaluators in the room during each usability test session. While observing the user, evaluators tracked issues and comments, and then aggregated the information from all usability tests. Observed errors were categorized using Nielsen-Shneiderman heuristics [29] (Table 1), a set of principles that provide guidelines and best practices for user interface design based on established usability heuristics.

During the study, evaluators assigned a severity rating to each identified issue, categorizing them into (1) cosmetic, (2) minor, (3) major, and (4) catastrophic. A cosmetic issue refers to a superficial concern that does not affect users' ability to complete a task, such as formatting inconsistencies or color preferences; these issues are typically of lower priority for resolution. A minor issue represents a problem with an available workaround, such as restricting numerical input to whole numbers rather than allowing decimals. Major issues, however, can significantly impact data quality, user satisfaction, or system functionality. Examples include the inability to access a function by clicking an icon, necessitating the use of a menu bar instead, or encountering difficulty when attempting to exit a screen. These issues require prioritization. Finally, a catastrophic severity rating indicates a critical system failure that demands immediate remediation. For instance, a user's inability to locate or input essential information constitutes a catastrophic issue.

XSL•FO
RenderX

Table . Nielsen-Shneiderman heuristics.

Error	Heuristic	Description
1	Consistency	Consistency and standards: Users should not have to wonder whether different words, situations, or actions mean the same thing. Standards and conventions in product design should be followed (*eg, font and layout and position*).
2	Visibility	Visibility of system state: Users should be informed about what is going on with the system through appropriate feedback and display of information (eg, what is current state of the system and what changes are made after an action).
3	Match	Match between system and the world: The image of the system perceived by users should match the model the users have about the system (eg, user model matches system image, and objects on the system should match objects of the task).
4	Minimalist	Any extraneous information is a distraction and a slow-down (eg, less is more and simple is efficient).
5	Memory	Minimize memory load: Users should not be required to memorize a lot of information to carry out tasks. Memory load reduces users' capacity to carry out the main tasks (eg, recognition vs concrete examples such as DD/MM/YY, eg, 10/20/15).
6	Feedback	Informative feedback: Users should be given prompt and informative feedback about their actions (eg, response time, concrete and specific, and not abstract and general).
7	Flexibility	Flexibility and efficiency: Users always learn and users are always different. Give users the flexibility of creating customization and shortcuts to accelerate their performance (eg, shortcuts for experienced users, templates, and bookmarks).
8	Message	Good error messages: The messages should be informative enough such that users can understand the nature of errors, learn from errors, and recover from errors (eg, precise not vague or general and polite messages).
9	Error	Prevent errors: It is always better to design interfaces that prevent errors from happening in the first place (eg, avoid text wraps, avoid slips or mistakes, and interfaces that make errors impossible).
10	Closure	Clear closure: Every task has a beginning and an end. Users should be clearly notified about the completion of a task (eg, clear beginning, middle and end, and clear feedback that goal was achieved).
11	Undo	Reversible actions: Users should be allowed to recover from errors. Reversible actions also encourage exploratory learning (eg, reverse actions in a single action and prevent serious errors).
12	Language	Use users' language: The language should be always presented in a form understandable by the intended users (eg, use standard meaning, language for specialized group, and users' perspective).

XSL•FO
RenderX

Error	Heuristic	Description
13	Control	Users in control: Do not give users that impression that they are controlled by the systems (eg, users are initiators not responders to actions, and avoid surprising actions and tedious sequences).
14	Documentation	Help and documentation: Always provide help when needed.

User Satisfaction Assessment

Satisfaction refers to the user's contentment and attitude toward the system. After completion of the think-aloud process, users were asked to complete a System Usability Scale (SUS) [30] survey. The SUS is a short, 10-question survey tool with high reliability and validity that measures user satisfaction. The survey employs a Likert scale ranging from 1 (strongly disagree) to 4 (strongly agree) to generate a single composite score representing the app's overall usability. Participants recorded their responses on paper, which were later digitized by the study team. At the end of the assessment, users were encouraged to share any concerns or recommendations.

Ethical Considerations

This study was approved by the University of Arkansas for Medical Sciences Institutional Review Board (protocol #206110). Informed consent (Multimedia Appendix 1) was obtained from all participants after explaining the study's purpose and procedures. Participant data were deidentified to ensure privacy and confidentiality. Each participant received US $50 as compensation for their time and involvement.

Results

Design of PEAK-mAAP

Using the HBM to explain current behaviors within our population, we hypothesize that the "cues to action" provided by PEAK-mAAP will be an essential factor in triggering behavior change. The app provides asthma educational video modules to address the "perceived susceptibility and severity" of asthma. The personalized mobile app features also provide additional "cues to action" during worsening symptoms to address "perceived benefits and barriers" (ie, give a clear action plan for the users to manage current asthma symptoms). Based on the education provided combined with self-management assistance from PEAK-mAAP, we hypothesize that self-efficacy and adherence will improve and result in improved asthma control for PEAK-mAAP users. To prove our hypotheses, we are conducting an RCT. The results of the RCT will be published upon completion of the trial.

Final Implementation of PEAK-mAAP

PEAK-mAAP is structured around 7 key modules.

First, as illustrated in Figure 3A, the patient information module encompasses a table presenting patient and health care provider details; a red "MEDICATIONS" button linking to the patient's asthma medication plan; a blue "REPORT CONTROLLER MEDICATION" button for logging medications; a "REQUEST CLINCARD BALANCE" button for users to request their ClinCard's balance, a reloadable debit card for research participant incentive; customizable reminder settings; and a "VIEW PRIVACY POLICY AND TERMS OF SERVICE" button to access the PEAK-mAAP's privacy terms. Figure 3B highlights a patient's personalized AAP, guiding actions including controller medication and rescue medication use based on their asthma zone, whether green, yellow, or red.

Second, a dashboard of patient activities, as seen in Figure 3C; the "Asthma Zone" allows visualization of the patient's recent asthma zones through a line chart, where each zone is depicted by uniquely colored circles. For color-blind users' ease, each circle also carries a letter mark indicating the logged zone. This intuitive design lets patients swiftly gauge their recent asthma status. The "Survey" card indicates the days leading up to the RCT's monthly survey due date. The "Rewards" card tracks weekly progress toward participant logging goals. As part of the PEAK-mAAP RCT trial, users earn nominal monetary incentives for active app engagement each week. The first daily record of a Peak Flow value or symptom will begin to fill the lung icon, signifying task completion and eligibility for the day's incentive. Completion of tasks each day will increase filling the lungs until the week's goals (eg, logging daily or completing the RCT's monthly survey) are met and the lungs fill completely. Such tangible rewards can bolster users' adherence to the app.

XSL•FO
RenderX

Figure 3. (**A**) The patient information tab; (**B**) the patient medication page with the patient's personalized Asthma Action Plan; and (**C**) the asthma zone record line chart, the survey countdown module, and the reward progress tracker.

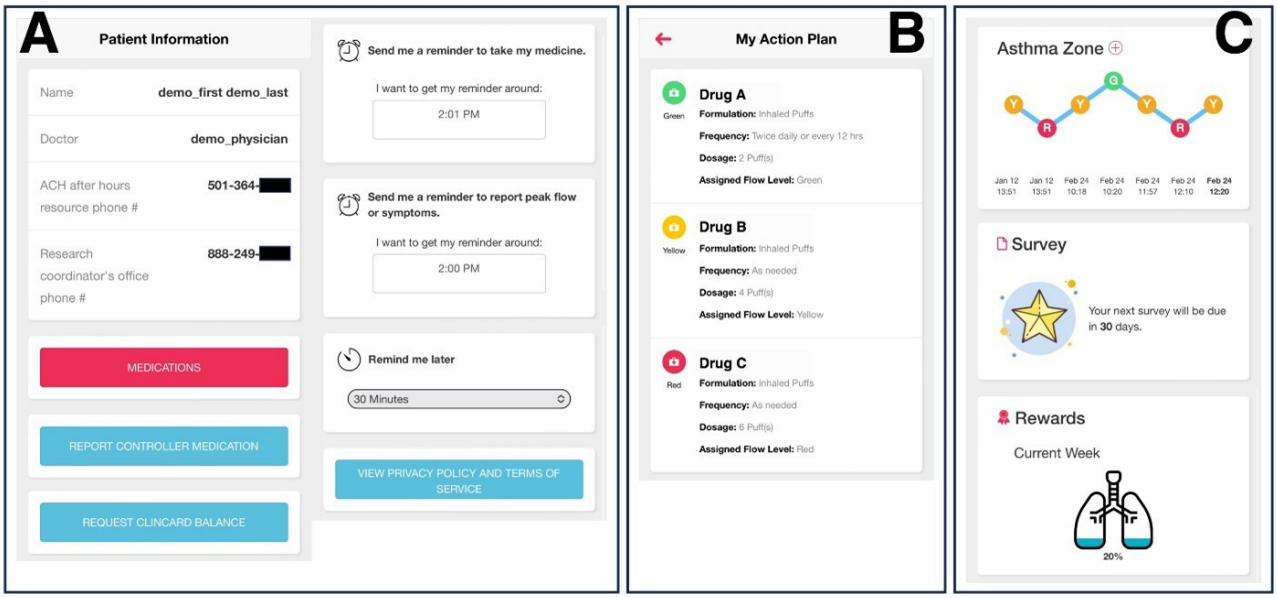

Third, a Peak Flow or asthma symptom record module: based on self-recorded Peak Flow value or asthma symptoms, PEAK-mAAP will classify the patient into their personalized asthma zones (ie, green, yellow, and red zones) and give zone-specific instructions (ie, action plan) such as taking emergency medications for acute symptoms. As shown in Figure 4A, before recording asthma-related symptoms or Peak Flow value, the system will first ask the participant "How are you feeling today?" The options are (1) "I'm feeling Great," (2) "I feel OK," and (3) "I feel horrible." Afterward, as shown in Figure 4B, the patient is asked to select how they prefer to determine their asthma zone, either Peak Flow or asthma symptoms. Figure 4C depicts the "Record your Peak Flow" screen, in which the patient can swipe on the blue bar to change the Peak Flow value. When the patient wants to report symptoms, a screen shown in Figure 4D will pop up, and the patient selects from a panel of 6 groups of symptoms to report.

Fourth, a notification module, as seen in Figure 5A, which shows a list of example notifications. The most recent messages are on the top of the list, and the unread and read notifications are separated by an "OLD MESSAGE" line separator. Figure 5B is an example of the in-app notification and Figure 5C is an example of the push notification to mobile phone native notification system.

XSL•FO
RenderX

Figure 4. Pulmonary Education and Knowledge Mobile Asthma Action Plan navigates the patient through a series of simple steps to log daily asthma status and during any episodes of increased symptoms to determine their asthma zone: (**A**) "How are you feeling today?"; (**B**) "Determine your Asthma Zone"; (**C**) "Record your Peak Flow"; and (**D**) symptoms recording page. Patients can elect to enter either Peak Flow or symptoms at their preference.

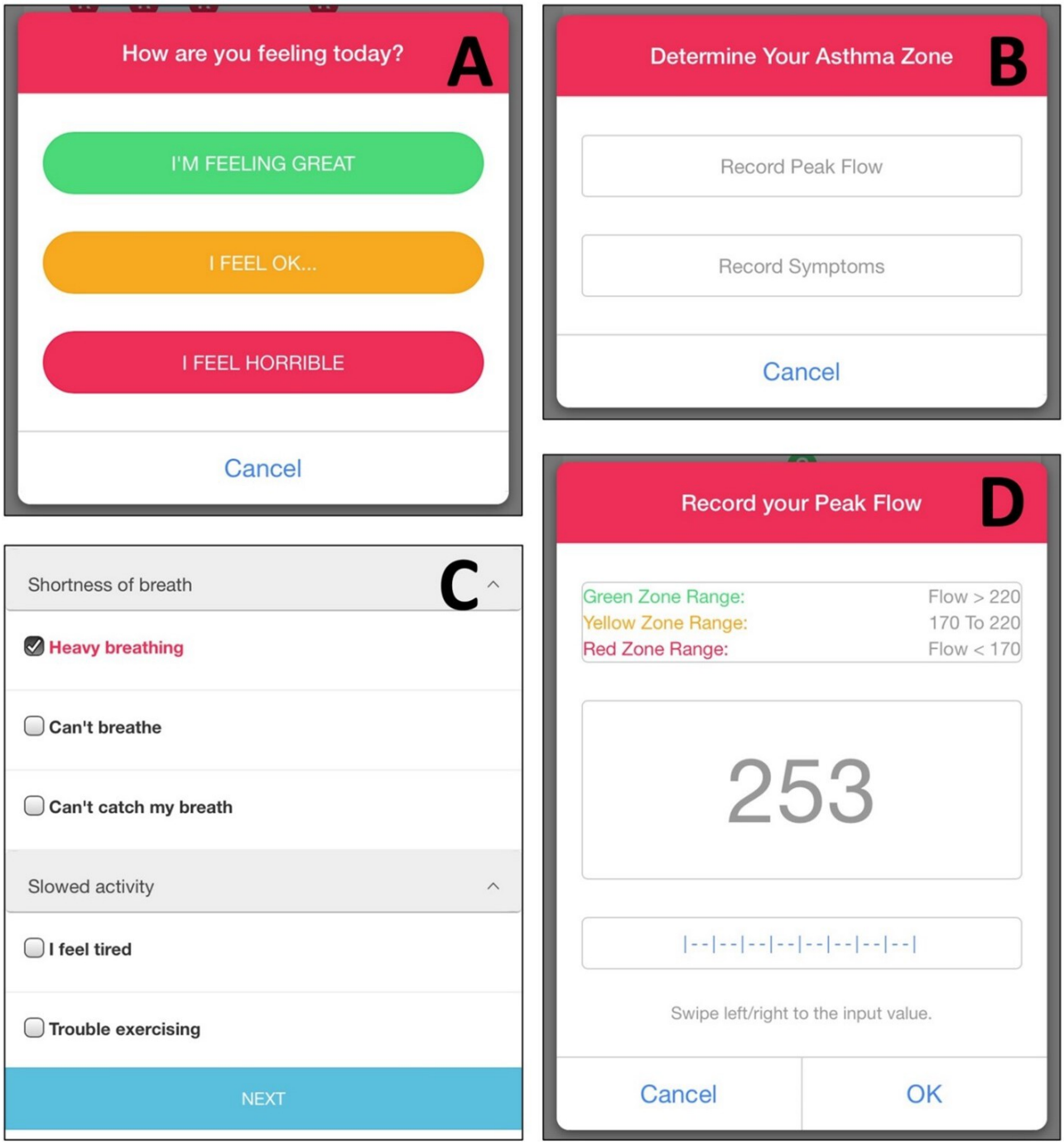

XSL•FO
RenderX

Figure 5. (**A**) The history of notifications that were sent to the patient; (**B**) the in-app notification screen; and (**C**) the mobile system's native notifications.

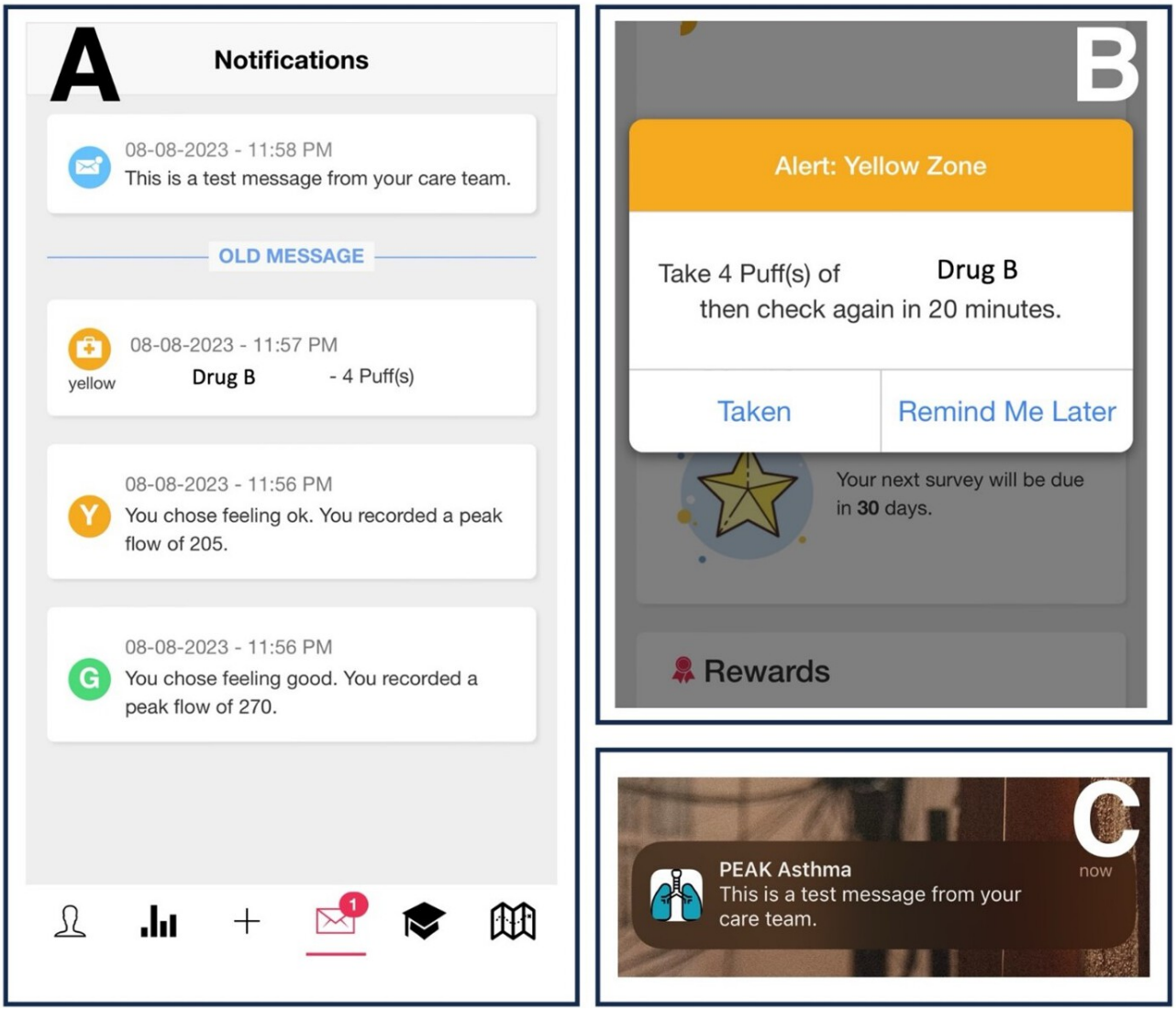

Fifth, an education module, as shown in Figure 6A, which consists of (1) an asthma educational tip card and (2) a list of asthma educational videos. The educational tip card will be refreshed weekly and the tips will also be pushed to the patients through the notification function. The asthma educational videos were created by the research team including asthma clinicians, asthma educators, and respiratory therapists using the Vidscrip platform. These resources aim to heighten patient awareness and empower them to manage their asthma more effectively.

Sixth, a nearby medical help map, as seen in Figure 6B, displaying a Google Map pinpointing nearby medical facilities, including hospitals and pharmacies. Users can click on markers for more detailed information about a facility and also receive directions via the Google Maps mobile app.

Seventh, a weather information module, as shown in Figure 6C, consisting of a weather information card, which shows real-time weather data and pollen level information that may affect patients' asthma outcomes.

XSL•FO
RenderX

Figure 6. (**A**) The education module; (**B**) the nearby medical help map; and (**C**) the weather information module.

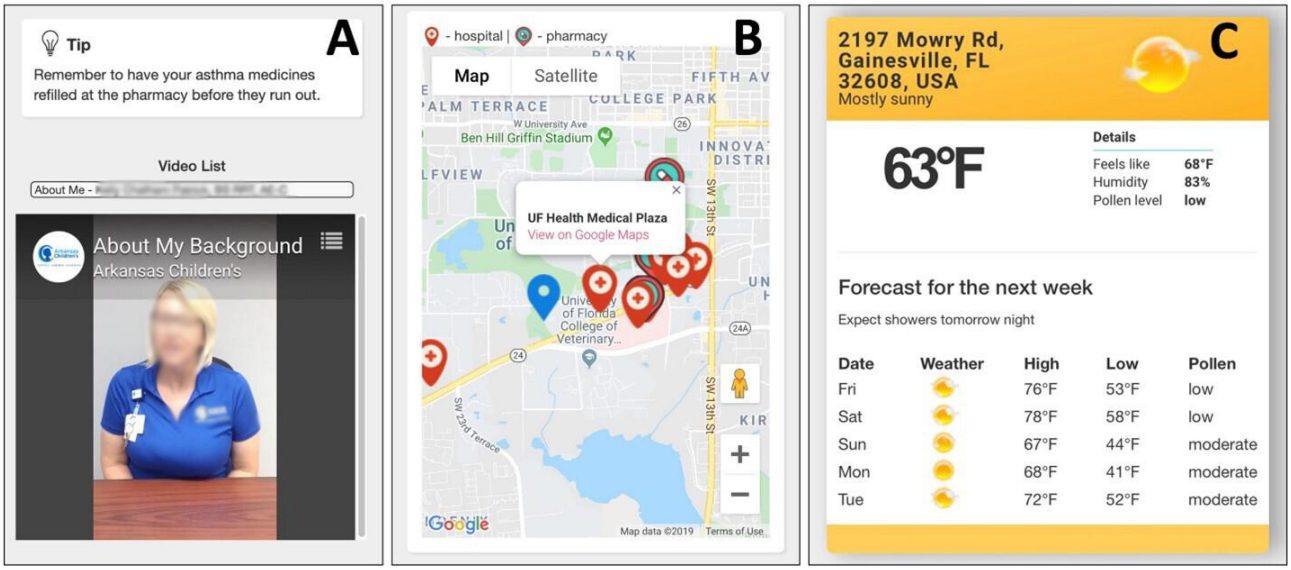

Web Portal

We developed a web portal for the health care providers and guardians of the patients to enable monitoring of the patient's usage of PEAK-mAAP and keep track of patients' asthma status. The health care providers are granted full access, which means they can view all the patients' information and see the statistics related to the use of PEAK-mAAP. Furthermore, health care providers can create patient accounts, edit the information of the patients, and send notifications to all patients or a specific one. On the other hand, guardians are only allowed to access limited information specifically related to the patient they are associated with.

Effectiveness and Efficiency of PEAK-mAAP

During the think-aloud process, evaluators recorded a total of 39 instances in which participants encountered difficulties in correctly performing a task, referred to as errors. These errors corresponded to violations of 9 out of the 14 heuristics under evaluation. The most frequently violated heuristic principles were match (n=12) and visibility (n=8). These were followed by memory (n=5), control (n=4), and language (n=3). Additional violations were observed in the closure, minimalist, document, and error heuristics. Examples of errors related to the match principle included confusion between the notification tab and the reminder tab, as well as difficulties in navigating the map to locate hospitals and pharmacies. The visibility principle was the second most frequently violated, with errors such as difficulty distinguishing between old and new messages within the app and challenges in locating and logging symptoms that could trigger relevant notifications. A memory error example was that a participant could not remember how to log their activity for the week to receive a reward. An example of a control error was that the participant had trouble recording their peak flow using the dial to select the correct number. Not understanding the icons used on the map was an example of a language error. There were 5 principles that were not violated, that are, undo, message, consistency, feedback, and flexibility. Of the 39 total errors, 26% (n=10) were rated as minor, 69% (n=27) as major, and 5% (n=2) as catastrophic.

Satisfaction Assessment of PEAK-mAAP

Participant satisfaction with the PEAK-mAAP mobile app was assessed using the SUS, yielding an overall mean score of 83 (SD 5.54), indicating high usability satisfaction. Older participants (≥17 years old) reported higher usability scores, with a mean of 87.5 (SD 1.71), compared with younger participants (12 - 16 years old), who had a mean score of 77.5 (SD 2.36). These findings suggest that older users found the app more user-friendly. The highest agreeability was found in survey question 7 stating, "*I would imagine that most people would learn to use the app very quickly.*" In total, 12 participants reported that they strongly agreed with this statement, and the remaining responses indicated that they agreed with this statement. The greatest variability was seen in survey question 10 stating, "*I needed to learn a lot of things before I could get going with the app*" (strongly disagree=6, disagree=2, neither or neutral=2, agree=1, strongly agree=2).

Discussion

Principal Findings

The primary objective of this study was to develop and evaluate PEAK-mAAP, an innovative mHealth app designed to improve asthma self-management among adolescents. Findings indicate that PEAK-mAAP, rooted in the HBM and developed using a UCD approach, enhances adolescents' engagement in asthma management. The usability assessment provided positive feedback on the app's ease of use and its potential for supporting behavior change. These results align with our initial hypotheses that a well-integrated mHealth platform—combining educational content, symptom tracking, and behavior modification strategies—can effectively promote self-management practices.

UCD and Usability Assessment

A key factor in PEAK-mAAP's potential success is its UCD approach, which ensures the app is intuitive and user-friendly. By incorporating user feedback early in development, the app was designed to meet the needs and expectations of adolescents, thereby encouraging consistent use. When users feel that an app

XSL•FO
RenderX

is tailored to their needs, they are more likely to engage with it over time. The usability assessment results reinforced this; participants rated the app highly in usability, achieving an impressive usability score. Overall user satisfaction with PEAK-mAAP was high. The SUS scores averaged 83 (SD 5.54), with older adolescents (>17 years) rating it higher (87.5, SD 1.71) than younger participants (77.5, SD 2.36). This trend may indicate that older adolescents are generally more comfortable using mobile apps, or that certain app features are more intuitive for them. Regardless of age, participants agreed that most users would be able to learn how to use the app quickly.

PEAK-mAAP performed well in many heuristic categories. Users encountered no issues with consistency, feedback, flexibility, message, or undo. However, errors in "match" (12/39, 31%) and "visibility" (8/39, 21%) posed significant challenges. The "match" errors stemmed from inconsistencies between user expectations and actual system functions, particularly regarding reminders and navigation tools. Meanwhile, "visibility" issues were especially problematic, as some users struggled to distinguish new messages from old ones. Notably, 2 visibility errors were classified as catastrophic, as they significantly impacted user interaction with notifications.

However, despite strong usability performance, testing identified critical issues that could impact the user experience if left unaddressed. For example, users had difficulty identifying the correct icon to set Peak Flow logging reminders—many mistakenly selected the notification icon instead. In addition, the dial interface for recording Peak Flow was not user-friendly, making it challenging for users to input accurate data. This issue was particularly significant because the peak flow value determines the personalized asthma management guidance provided by the app. Younger participants also faced challenges with reading the map, likely due to their limited experience with navigation tools.

Identified Usability Issues and Modifications

To enhance usability and ensure a seamless experience, several modifications were made before the RCT. The reminder icon was redesigned to align with users' expectations, and the notification icon was changed from a bell to an envelope to prevent confusion. Similarly, the dial's sensitivity was adjusted to allow for more precise Peak Flow input, reducing the risk of inaccurate data entries. To address visibility concerns, new messages on the notification screen were color-coded (blue for unread messages, gray for read ones), and a dividing line was added to separate them. In addition, on the login screen, a password reset option was implemented to improve accessibility. Another significant modification was made to the educational content delivery. Initially, weekly asthma education videos under the education tab played consecutively, which could be overwhelming. This feature was adjusted to allow users to control playback, preventing automatic continuous play. For motivation and engagement, the app incorporated an age-appropriate reward system, represented by a progressively filling blue lung icon that visually tracked the user's daily logging progress. Finally, usability feedback also informed the development of training materials to better introduce participants to the app's features, ensuring that users fully understood its capabilities.

Advantages of the Hybrid Mobile App Approach

The decision to develop PEAK-mAAP as a hybrid mobile app brings several benefits. Hybrid apps work across different platforms, meaning more people can use the app without compatibility issues. This approach is also cost-effective since developers do not need to create separate versions for different platforms. Plus, hybrid apps can easily integrate with a phone's built-in features, enhancing the user experience. Compared with a native implementation, where the app must be developed separately for Android and iOS platforms, the hybrid approach significantly shortened the development time frame. It facilitated a rapid reach to a broader user base and proved to be more cost-effective in construction and maintenance. Furthermore, the hybrid mobile app can run on browsers, ensuring that adolescents remain engaged with PEAK-mAAP even if access to the mobile platform is unavailable, such as in the case of a lost phone.

Limitations

Despite these promising outcomes, several limitations must be acknowledged. Some users experienced an initial learning curve, which could impact short-term engagement. Longitudinal studies are needed to assess sustained usage and long-term effectiveness. Furthermore, as the efficacy of PEAK-mAAP has not yet been tested through an RCT, conclusions regarding its impact on clinical outcomes remain speculative at this stage.

Conclusion

This study presents the design, development, and usability evaluation of PEAK-mAAP, a hybrid mHealth intervention designed to support asthma self-management in adolescents. Rooted in the HBM and developed through a UCD approach, PEAK-mAAP integrates educational content, symptom tracking, medication adherence tools, and real-time feedback to enhance self-management behaviors. Usability testing revealed high user satisfaction, with strong SUS scores and positive engagement. While some challenges—particularly in navigation, reminders, and visibility—were identified, subsequent modifications improved the user experience before launching the RCT. The hybrid mobile app approach ensures cross-platform compatibility, cost efficiency, and accessibility, positioning PEAK-mAAP as a scalable and adaptable solution for adolescent asthma management. Although usability results are promising, further research should assess long-term effectiveness and sustainability in improving asthma control through a rigorous RCT. Addressing usability concerns and incorporating ongoing user feedback will be critical to fostering widespread adoption and adherence among adolescents. Ultimately, PEAK-mAAP represents a significant advancement in mHealth technology, bridging gaps in adolescent asthma care with a personalized and evidence-based approach to self-management.

XSL•FO
RenderX

Acknowledgments

This work was supported by the National Institutes of Health (NIH; grant 1R01NR015988-01A1). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. We did not use generative artificial intelligence in any portion of the manuscript writing.

Data Availability

The data generated and analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

TTP and JB designed methods and experiments. All coauthors contributed to design the app. XH developed and maintained the app. XH, JB, and TTP drafted the main manuscript. All coauthors provided critical edits of the early draft and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Usability testing agreement.
[PDF File, 87 KB - formative_v9i1e64212_app1.pdf]

References

1. The Global Asthma Report 2022. int J tuberc lung dis 2022 Nov 25;26(1):1-104. [doi: 10.5588/ijtld.22.1010]
2. Farzandipour M, Nabovati E, Sharif R, Arani MH, Anvari S. Patient self-management of asthma using mobile health applications: a systematic review of the functionalities and effects. Appl Clin Inform 2017 Oct;8(4):1068-1081. [doi: 10.4338/ACI-2017-07-R-0116] [Medline: 29241254]
3. Alquran A, Lambert KA, Farouque A, et al. Smartphone applications for encouraging asthma self-management in adolescents: a systematic review. Int J Environ Res Public Health 2018 Oct 29;15(11):2403. [doi: 10.3390/ijerph15112403] [Medline: 30380692]
4. Song T, Yu P, Zhang Z. Design features and health outcomes of mhealth applications for patient self-management of asthma: a systematic review: mhealth apps for asthma self-management. Presented at: ACSW '22: Proceedings of the 2022 Australasian Computer Science Week; Feb 14-18, 2022; Brisbane, Australia. [doi: 10.1145/3511616.3513110]
5. Farzandipour M, Nabovati E, Heidarzadeh Arani M, Akbari H, Sharif R, Anvari S. Enhancing asthma patients' self-management through smartphone-based application: design, usability evaluation, and educational intervention. Appl Clin Inform 2019 Oct;10(5):870-878. [doi: 10.1055/s-0039-1700866] [Medline: 31724144]
6. Kosse RC, Bouvy ML, de Vries TW, Koster ES. Effect of a mHealth intervention on adherence in adolescents with asthma: a randomized controlled trial. Respir Med 2019 Mar;149:45-51. [doi: 10.1016/j.rmed.2019.02.009] [Medline: 30803885]
7. Fedele DA, Thomas JG, McConville A, et al. Using mobile health to improve asthma self-management in early adolescence. Journal of Adolescent Health 2021;69(6):1032-1040. [doi: 10.1016/j.jadohealth.2021.06.011]
8. Li Z, Lu F, Wu J, et al. Usability and effectiveness of eHealth and mHealth interventions that support self-management and health care transition in adolescents and young adults with chronic disease: systematic review. J Med Internet Res 2024 Nov 26;26(1):e56556. [doi: 10.2196/56556] [Medline: 39589770]
9. Nanzer AM, Lawton A, D'Ancona G, Gupta A. Transitioning asthma care from adolescents to adults: severe asthma series. Chest 2021 Oct;160(4):1192-1199. [doi: 10.1016/j.chest.2021.05.019] [Medline: 34019863]
10. Holley S, Morris R, Knibb R, et al. Barriers and facilitators to asthma self-management in adolescents: a systematic review of qualitative and quantitative studies. Pediatr Pulmonol 2017 Apr;52(4):430-442. [doi: 10.1002/ppul.23556] [Medline: 27717193]
11. Asthma NAEAPP Third Expert Panel on the Diagnosis and Management of Asthma. Expert panel report 3: guidelines for the diagnosis and management of asthma. : National Heart, Lung, and Blood Institute (US); 2007.
12. Burbank AJ, Lewis SD, Hewes M, et al. Mobile-based asthma action plans for adolescents. J Asthma 2015;52(6):583-586. [doi: 10.3109/02770903.2014.995307] [Medline: 25494553]
13. Perry TT, Marshall A, Berlinski A, et al. Smartphone-based vs paper-based asthma action plans for adolescents. Ann Allergy Asthma Immunol 2017 Mar;118(3):298-303. [doi: 10.1016/j.anai.2016.11.028] [Medline: 28111110]
14. Abras C, Maloney-Krichmar D, Preece J. User-centered design. In: Bainbridge W, editor. Encyclopedia of Human-Computer Interaction: Sage Publications; 2004:445-456.
15. Rosenstock IM, Strecher VJ, Becker MH. Social learning theory and the Health Belief Model. Health Educ Q 1988;15(2):175-183. [doi: 10.1177/109019818801500203] [Medline: 3378902]
16. Apache Cordova. URL: https://cordova.apache.org/ [accessed 2023-08-09]

XSL•FO
RenderX

17. Ruby on Rails. URL: https://rubyonrails.org/ [accessed 2023-08-09]
18. PostgreSQL. URL: https://www.postgresql.org/ [accessed 2023-08-09]
19. AngularJS. URL: https://angularjs.org/ [accessed 2023-08-09]
20. Google Maps platform. Google for Developers. URL: https://developers.google.com/maps [accessed 2023-08-09]
21. Vidscrip. URL: https://www.vidscrip.com/ [accessed 2023-08-09]
22. Twilio. URL: https://www.twilio.com/en-us [accessed 2023-08-09]
23. Accuweather. URL: https://www.accuweather.com/ [accessed 2023-08-09]
24. Moroney L. Firebase cloud messaging. In: Moroney L, editor. The Definitive Guide to Firebase: Build Android Apps on Google's Mobile Platform: Apress; 2017:163-188. [doi: 10.1007/978-1-4842-2943-9_9]
25. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research Electronic Data Capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381. [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
26. Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: Building an international community of software platform partners. J Biomed Inform 2019 Jul;95:103208. [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]
27. Nielsen J. Why you only need to test with 5 users. Nielsen Norman Group. URL: https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/ [accessed 2023-08-09]
28. Charters E. The use of think-aloud methods in qualitative research an introduction to think-aloud methods. Brock Education Journal 2003;12(2). [doi: 10.26522/brocked.v12i2.38]
29. Zhang J, Johnson TR, Patel VL, Paige DL, Kubose T. Using usability heuristics to evaluate patient safety of medical devices. J Biomed Inform 2003;36(1-2):23-30. [doi: 10.1016/s1532-0464(03)00060-1] [Medline: 14552844]
30. Brooke J. SUS: a "quick and dirty" usability. In: Usability Evaluation in Industry 1996:189-194 URL: https://books.google.com/books?hl=en&lr=&id=ujFRDwAAQBAJ&oi=fnd&pg=PA189&dq=SUS+A+quick+and+dirty+usability+scale+&ots=ZaTJvZ-TAy&sig=wuNAJAYIO9uNOKFTJwcMMB8L0M [accessed 2025-04-15]

Abbreviations

AAP: Asthma Action Plan
API: application programming interface
HBM: health belief model
mAAP: mobile health Asthma Action Plan
mHealth: mobile health
PEAK-mAAP: Pulmonary Education and Knowledge Mobile Asthma Action Plan
RCT: randomized controlled trial
REDCap: Research Electronic Data Capture
SUS: System Usability Scale
UCD: user-centered design

Edited by A Mavragani; submitted 11.07.24; peer-reviewed by E Nabovati, M Praena-Crespo; revised version received 04.03.25; accepted 14.03.25; published 24.04.25.

Please cite as:
He X, Bian J, Berlinski A, Guo Y, Simmons AL, Marshall SA, Greene CJ, Brown RH, Turner J, Perry TT
Mobile Health App for Adolescent Asthma Self-Management: Development and Usability Study of the Pulmonary Education and Knowledge Mobile Asthma Action Plan
JMIR Form Res 2025;9:e64212
URL: https://formative.jmir.org/2025/1/e64212
doi:10.2196/64212

© Xing He, Jiang Bian, Ariel Berlinski, Yi Guo, A Larry Simmons, S Alexandra Marshall, Carolyn J Greene, Rita Hudson Brown, Jessica Turner, Tamara T Perry. Originally published in JMIR Formative Research (https://formative.jmir.org), 24.4.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Tailoring Rheumatoid Arthritis Visit Timing Based on mHealth App Data: Mixed Methods Assessment of Implementation and Usability

Robert S Rudin[1], PhD; Leah M Santacroce[2], MS; Ishani Ganguli[3,4], MD, MPH; Daniel H Solomon[2,4,5], MD, MPH

[1]RAND, 20 Park Plaza, Boston, MA, United States

[2]Division of Rheumatology, Inflammation, and Immunity, Brigham and Women's Hospital, Boston, MA, United States

[3]Division of General Internal Medicine and Primary Care, Brigham and Women's Hospital, Boston, MA, United States

[4]Harvard Medical School, Boston, MA, United States

[5]Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital, Boston, MA, United States

Corresponding Author:
Robert S Rudin, PhD
RAND, 20 Park Plaza, Boston, MA, United States

Abstract

Background: Visits to medical subspecialists are common, with follow-up timing often based on heuristics rather than evidence. Unnecessary visits contribute to long wait times for new patients. Specialists could enhance visit timing and reduce frequency by systematically monitoring patients' symptoms between visits, especially for symptom-driven conditions like rheumatoid arthritis (RA). We previously designed an intervention using a mobile health (mHealth) app to collect patient-reported outcomes (PRO). One of several aims of the app was to assist rheumatologists in determining visit timing for patients with RA. The intervention did not reduce visit frequency.

Objective: To explore possible reasons for the lack of association between the intervention and visit frequency, we describe app usage, assess usability, and identify barriers and facilitators for using between-visit PRO data to reduce visits when patients' symptoms are stable.

Methods: We analyzed patients' use of the app by reporting adherence (percent of PRO questionnaires completed during the 12-month study) and retention (use in the last month of the study). To examine rheumatologists' experiences, we summarized views of the electronic health record (EHR)–embedded PRO dashboard and EHR inbox messages suggesting early or deferred visits. We assessed app usability using the interactive mHealth App Usability Questionnaire for Ease of Use and Usefulness for patients and the System Usability Scale for rheumatologists. We assessed rheumatologist-level effects of intervention usage using Kruskal-Wallis rank sum and equality of proportion tests. We identified barriers and facilitators through interviews and surveys.

Results: The analysis included 150 patients with RA and their 11 rheumatologists. Patients answered a median of 53.3% (IQR 34.1%-69.2%) of PRO questionnaires; this proportion varied by rheumatologist (range 40.7% - 67%). Over half of the patients used the app during the final month of the study (56%, range 51% - 65%, by rheumatologists); the median number of months of use was 12 (IQR 9-12). Rheumatologists viewed the dashboard 78 times (17.6% of 443 visits) with significant differences in viewing rates by rheumatologist (range 10% - 66%; $P<.01$). There were 108 generated messages sent to rheumatologists suggesting a deferred visit (24.4% of 443 visits) with significant differences in message counts received per visit by rheumatologist (range 10.8% - 22.6%; $P=.03$). Rheumatologists' reported barriers to offering visit deferrals included already scheduling as far out as they were comfortable and rescheduling complexities for staff. Based on 39 patient interviews and 44 surveys, patients reported 2 main barriers to app usage: questionnaire frequency not being tailored to them and reduced motivation after not discussing PRO data with their rheumatologist. A total of 5 interviewed patients received the option to defer their visits, of which 3 elected to defer the appointment and 2 chose to keep it.

Conclusions: While an mHealth app for reporting RA PROs was used frequently by patients, using these data to reduce the frequency of unneeded visits was not straightforward. Better engagement of clinicians may improve the use of PRO data.

(JMIR Form Res 2025;9:e60854) doi:10.2196/60854

KEYWORDS

mobile health; patient-reported outcomes; visit timing; visit frequency; rheumatoid arthritis; rheumatology

XSL•FO
RenderX

Introduction

Visits to medical specialists represent a substantial component of health care services. About half of patient visits to physicians in the United States are to specialists, totaling roughly 500 million visits per year at a cost of billions of dollars [1,2]. The current approach to care delivery inefficiently allocates specialty care resources: visits are scheduled largely in preset intervals (eg, 3, 6, or 12 months) that are not evidence-based and have a weak relationship to clinical need [3,4]. Sometimes these schedules are required for insurance purposes, that is, to have medications covered. Evidence for optimal visit timing is scant [3,5]. As a result, specialists' schedules fill with potentially unneeded visits, creating inefficiencies, wasteful spending, and patient burdens [6], and contribute to longer wait times for other patients who need specialist care; these access issues are especially pronounced for rural, minority, and other vulnerable populations [7].

We previously designed a mobile health (mHealth) intervention to improve patient care in rheumatology [8]. It was well accepted by patients, and we became interested in whether one additional value of the app could be to reduce visit follow-up frequency when clinically appropriate [8,9]. The rationale was that a reasonable proportion of visits were for stable patients where remote symptom reporting could help clinicians defer visits. We focused on rheumatology because this specialty has a great need for optimizing visit timing due to projected workforce shortages in the next decade [10]. Furthermore, prior works suggest a potential for reducing visit frequency, in part, because a substantial number of visits, for rheumatoid arthritis (RA) in particular, do not result in a change in treatment, suggesting limited clinical use [8,11-13]. The mHealth intervention involved patients using a smartphone app to record their symptoms using patient-reported outcome (PRO) questionnaires; based on these data, their rheumatologists received notifications that either suggested the need for an earlier visit if symptoms were acutely worsening or the potential to defer a scheduled visit if symptoms were stable (see Methods for details).

Our initial clinical trial of this intervention found no association with reduced visit frequency. To explore possible reasons for this finding and inform further intervention development, we examined quantitative and qualitative data on patient and rheumatologist engagement and data from surveys and interviews. We describe the app intervention usage, assess usability, and identify possible barriers and facilitators to improve visit follow-up frequency using between-visit PRO data.

Methods

Setting and Study Population

The prior study was conducted at 2 rheumatology practices affiliated with Brigham and Women's Hospital, which is part of Mass General Brigham, a large academic medical center in Boston, MA, United States. All clinics within this health system share an electronic health record (EHR) system (Epic Systems, Inc). The intervention was implemented by engaging 11 clinically focused rheumatologists who previously provided input on the intervention design [8,9]. As described previously, we identified eligible patients by searching upcoming patient visit schedules for the 11 rheumatologists [8,9]. We used prior diagnosis codes to identify patients with RA. Only established patients were included, and therefore all prescheduled visits were follow-ups. We recruited these patients via patient portal messages, phone calls, and in-person prior to scheduled visits.

mHealth Intervention

The intervention was designed with extensive input from patients and rheumatologists through several iterations as part of a user-centered design process [8,9,14,15]. Briefly, the intervention included a patient-facing mHealth app that used push notifications to prompt patients to complete one of four types of validated PRO questionnaires regarding pain interference, function, fatigue, and RA disease activity [8]. Each questionnaire was repeated every 8 days in a staggered fashion so that patients completed one questionnaire every 2 days. Rheumatologists were able to view all results within the patient's EHR chart. In addition, the PRO data could trigger 3 types of EHR inbox messages to the rheumatologists. First, all rheumatologists received a message that data were available for viewing 48 hours before a visit with a patient who had entered PRO data. Second, if PRO data suggested several weeks of worsening symptoms, the rheumatologist received a message with a suggestion to offer an earlier visit. Finally, if PRO data suggested stable symptoms during the visit interval from the prior visit until 2 weeks prior to a scheduled visit, rheumatologists received a message suggesting that the patient be offered a deferred visit. The specific logic that determined when these messages are sent has been described previously [8].

Study Design and Data Sources

As noted, this initial trial of the mHealth app found no change in visit frequency attributable to the PRO data and EHR inbox messages [9]. We sought to explore potential reasons for these findings and inform further intervention development by (1) describing use of the mHealth app by patients and use of the PRO data by rheumatologists, (2) assessing mHealth app and EHR-embedded PRO data usability, and (3) identifying barriers and facilitators. Patient and rheumatologist perspectives were assessed through surveys during and after the study period. Data sources included the EHR (for demographic and clinical characteristics of study patients and visit timing), mHealth app and dashboard usage logs (for patient adherence and retention, rheumatologist views of the PRO data, and messages sent suggesting early or deferred visits), interviews with patients (for barriers and facilitators), and surveys of patients and rheumatologists (for usability and barriers and facilitators).

For analyses of intervention usage and usability, rheumatologists with 15 or more patients in the study were analyzed as individuals, and the rest were aggregated. Because only one of the instances in which an earlier visit was suggested resulted in an earlier visit, we focused the analysis primarily on aspects related to deferred visits [9].

XSL•FO
RenderX

Intervention Usage Analysis

We used descriptive statistics to summarize patient adherence and retention rates with the mHealth app. We used the Kruskal-Wallis rank sum test to assess differences in median patient PRO completion percentage across rheumatologists. We used descriptive statistics to summarize volumes and timing of EHR inbox messages sent to the rheumatologists suggesting the potential for an early or deferred visit and view of the EHR-embedded PRO data. For both measures, we tested for equality of proportions across rheumatologists.

Usability Analysis

We summarized results from mHealth App Usability Questionnaire (MAUQ) items, which were administered to patients after they completed the study [16]. The MAUQ is a validated instrument measuring mHealth app usability. Specifically, we used the interactive versions of the ease of use and satisfaction (MAUQ_E) and usefulness (MAUQ_U) subscales of the questionnaire because they were most relevant to this intervention. The MAUQ scales range from 1 (worst, ie, strongly disagree with all questions) to 7 (best, ie, strongly agree with all questions), with 4 meaning neutral.

For rheumatologists, we summarized the results of the System Usability Scale (SUS) administered after all the rheumatologists' patients had completed the study [17,18]. The SUS is a widely used usability measure with a range from 0 (worst) to 100 (best), with an average score across evaluations of 68 - 70.

Barriers and Facilitators Analysis

We invited patients who had received a suggested early or deferred visit to participate in an interview. We conducted those interviews virtually and transcribed the recordings. Using conventional content analysis, one research team member (RSR) created a codebook and summarized the results using the framework method, and a second research team member (DHS) reviewed a portion of the results to confirm findings [19,20]. We used similar methods to analyze free-text responses from patient survey results administered at study completion. When presenting qualitative results from interviews, we used "some" to indicate a finding was present in fewer than 5 responses, "many" to denote greater than 5 responses, and "most" to denote greater than half of the responses.

For rheumatologists, we summarized data from a survey administered after study completion about how often they offered patients the option to defer a visit and the reasons for not offering that option.

All quantitative analyses were conducted in R (version 4.2.2; R Foundation for Statistical Computing).

Ethical Considerations

This study was reviewed and approved by the institutional review board at Mass General Brigham (number 2021P000790). All written and audiovisual data (which were not anonymized) were stored on secure servers accessible only to the research team. Written consent was obtained for all patients to participate in the trial and for the use of their medical data in analyses, verbal consent was obtained for interviews, and consent information was presented before all electronic survey questions. Participants were compensated with a gift card for the trial and a US $25 check for the interviews.

Results

Participants

Characteristics of the 150 intervention patients included in the study are shown in Table 1. Patients were mostly female and White.

XSL•FO
RenderX

Table . Characteristics of patients using the mHealth app for rheumatoid arthritis (N=150).

Characteristics	Values
Sex, n (%)	
Female	125 (83.3)
Male	25 (16.7)
Age (years), median (IQR)	62 (52-69)
Race, n (%)	
Asian	4 (2.7)
Black	5 (3.3)
White	136 (90.7)
Declined or others	5 (3.3)
Seropositivity status, n (%)	
Missing	36 (24)
0	39 (26)
1	75 (50)
CRP[a,b], median (IQR)	1.4 (0.7 - 4.5)
DMARD[c,d] use, n (%)	
csDMARD[e]	93 (62)
btsDMARD[f]	90 (60)
Other medications, n (%)	
NSAIDs[g]	30 (20)
Corticosteroid	55 (36.7)
Opioids	25 (16.7)
Comorbidities, n (%)	
Hypertension	26 (17.3)
Obesity	15 (10)
Depression	7 (4.7)
Osteoporosis	19 (12.7)

[a]CRP: C-reactive protein.
[b]Two values were missing for CRP.
[c]DMARD: disease-modifying antirheumatic drug.
[d]It is common for patients to use more than 1 DMARD.
[e]csDMARD: conventional synthetic disease-modifying antirheumatic drug.
[f]btsDMARD: biologic or targeted synthetic disease-modifying antirheumatic drug.
[g]NSAID: nonsteroidal anti-inflammatory drug.

Intervention Usage

Four rheumatologists each had at least 15 patients in the intervention and were analyzed as individual observations; the other 7 rheumatologists were aggregated and treated as one observation. The median percentage of PRO questionnaires completed was 53.3% across all patients (Table 2). This completion rate varied by rheumatologists, ranging from 40.7% (the rheumatologist with the most patients) to 67% (the rheumatologist with the third most patients). Differences in completion rates between all rheumatologists or groups were not statistically significant based on the Kruskal-Wallis rank sum test (P=.50). The median number of months patients answered questionnaires was 12 (IQR 9-12), with little variation by rheumatologist; 56% (84/150) of patients completed at least one questionnaire in the final month of the study. See app usage by rheumatologists in Table 2 and over time in Figure 1.

XSL•FO
RenderX

Table . App usage by rheumatoid arthritis patients grouped by their rheumatologist.

Rheumatologist[a]	Patients, n	Patient visits, n	Percent PRO[b] questionnaires completed,[c] median (IQR)	Months with answered questionnaires,[d] median (IQR)	Patients who completed at least 1 PRO in the final month of the study, n (%)
All rheumatologists	150	443	53.3 (34.1-69.2)	12 (9-12)	84 (56)
Four rheumatologists with the greatest volume of study patients					
1st highest	51	148	40.7 (25.8-63.7)	11 (8-12)	26 (51)
2nd highest	25	68	48.9 (34.1-64.8)	12 (9-12)	13 (52)
3rd highest	20	53	67 (49.5-78.6)	12 (11-12)	13 (65)
4th highest	15	49	46.7 (44-61)	12 (9-12)	8 (53)
All other rheumatologists					
5th-11th highest	39	125	58.8 (47.3-72)	12 (11-12)	24 (62)

[a]Ordered by volume of study patients who receive care from that rheumatologist.

[b]PRO: patient-reported outcome.

[c]The total possible PRO questionnaires offered were 182.

[d]Number of months within the 12-month study in which patients answered at least 1 PRO questionnaire.

Figure 1. Patient questionnaire completion rates within the rheumatoid arthritis app over the 12-month study period. PRO: patient-reported outcome; RADAI: Rheumatoid Arthritis Disease Activity Index.

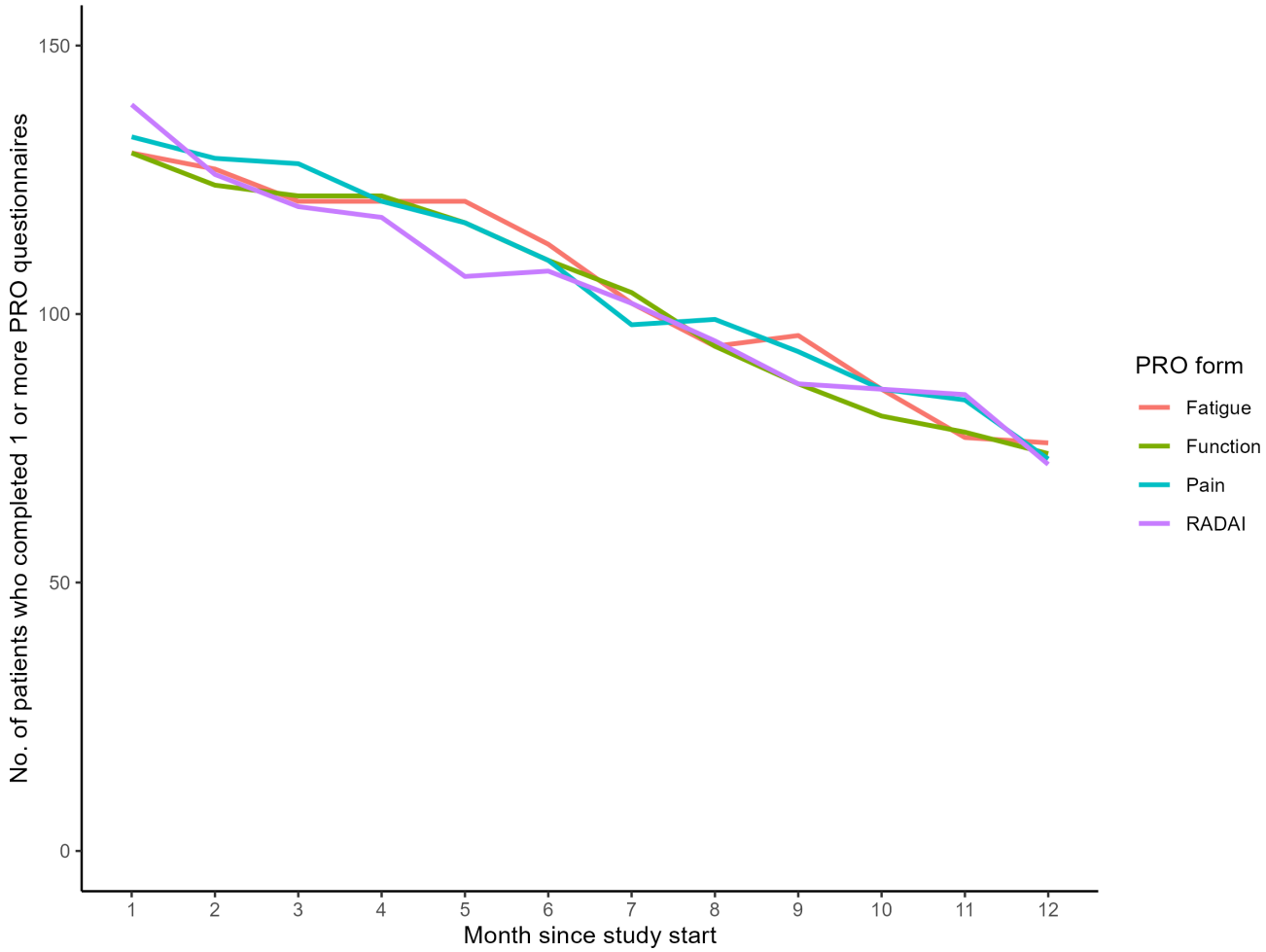

There was variation by rheumatologist in the percentage of their received messages that suggested a deferred visit, ranging from 18.9% of patient visits (the rheumatologist with the most patients) to 37.7% (the rheumatologist with the third most patients) (Table 3). In a 6-sample test for equality of proportions, there was a statistically significant difference in the percent of deferred visit suggestions received by rheumatologists (P=.03). There were no significant differences among early visits. For rheumatologist usage of the PRO data as measured by the number of views, there was variation by rheumatologist ranging

XSL•FO
RenderX

from 10.1% of patient visits (rheumatologist with the most patients) to 66% (rheumatologist with the third most patients). In a 6-sample test for equality of proportions, the difference was statistically significant ($P<.01$), which means that at least one value (number of views of PRO data per patient visit) differed significantly from the others.

Table . Usage of rheumatologist-facing intervention information generated by the rheumatoid arthritis app, available within the electronic health record.

Rheumatologist[a]	Patients, n	Patient visits, n	Messages sent to rheumatologist suggesting visit deferral, n (%)	Messages sent to rheumatologist suggesting early visit, n (%)	Rheumatologist views of PRO[b] data in EHR[c], n (%)
All rheumatologists	150	443	108 (24.4)	31 (6.9)	78 (17.6)
Four rheumatologists with the greatest volume of study patients					
1st highest	51	148	28 (18.9)	8 (5.4)	15 (10.1)
2nd highest	25	68	13 (19.1)	5 (7.4)	11 (16.2)
3rd highest	20	53	20 (37.7)	5 (9.4)	35 (66)
4th highest	15	49	18 (36.7)	3 (6.1)	7 (14.3)
All other rheumatologists					
5th-11th highest	39	125	29 (23.2)	10 (8)	10 (8)

[a]Ordered by volume of study patients who receive care from that provider.

[b]PRO: patient-reported outcome.

[c]EHR: electronic health record.

Usability

Of the 150 intervention patients, 88 completed the MAUQ surveys with a median MAUQ_E of 6.3 (IQR 5.6-6.9) and median MAUQ_U of 4.8 (IQR 4.1-5.6) (Table 4 and Figure 2). Based on Figure 2, most patients found the app easy to use across all questions. However, the usefulness of the app was less clear for patients. Most patients agreed that the app "would be useful for my health and well-being," but about half of patients were neutral about its usefulness on the other questions. There was a small range across rheumatologists of the aggregated MAUQ_E (6.2 - 6.9) and a more substantial range for the MAUQ_U (4.6 - 6.3).

Of 11 participating rheumatologists, 10 completed the SUS to assess the EHR-embedded dashboard and reported a mean of 68.25 (range 45 - 82.5). (The average score for the SUS across a wide range of apps is 68 (SD 12.5) [21].)

Table . Patient usability scores of the rheumatoid arthritis app at the completion of the 12-month study.

Rheumatologist[a]	Patients, n	MAUQ_E[b], median (IQR)	MAUQ_U[b], median (IQR)
All rheumatologists	88	6.3 (5.6-6.9)	4.8 (4.1-5.6)
Four rheumatologists with the greatest volume of study patients			
1st highest	31	6.2 (5.5-6.6)	4.6 (3.4-5.3)
2nd highest	17	6.3 (6.2-7.0)	4.6 (4.2-5.0)
3rd highest	12	6.7 (6.4-7.0)	5.2 (4.3-6.3)
4th highest	4	6.9 (6.2-7.0)	6.3 (5.7-6.6)
All other rheumatologists			
5th-11th highest	24	6.3 (5.6-6.9)	4.8 (3.8-5.3)

[a]Ordered by volume of study patients who receive care from that provider.

[b]The mHealth App Usability Questionnaire for Ease of Use (MAUQ_E) and Usefulness (MAUQ_U) were rescaled from 1 - 5 to 1 - 7 (worst to best) to be consistent with their standard use, with 1 indicating strongest disagreement with all questions and 7 indicating strongest agreement with all questions.

XSL•FO
RenderX

Figure 2. Rheumatoid arthritis app usability as measured by the mHealth App Usability Questionnaire for Ease of Use (MAUQ_E) and Usefulness (MAUQ_U) for clinically integrated apps, administered at the completion of the 12-month study.

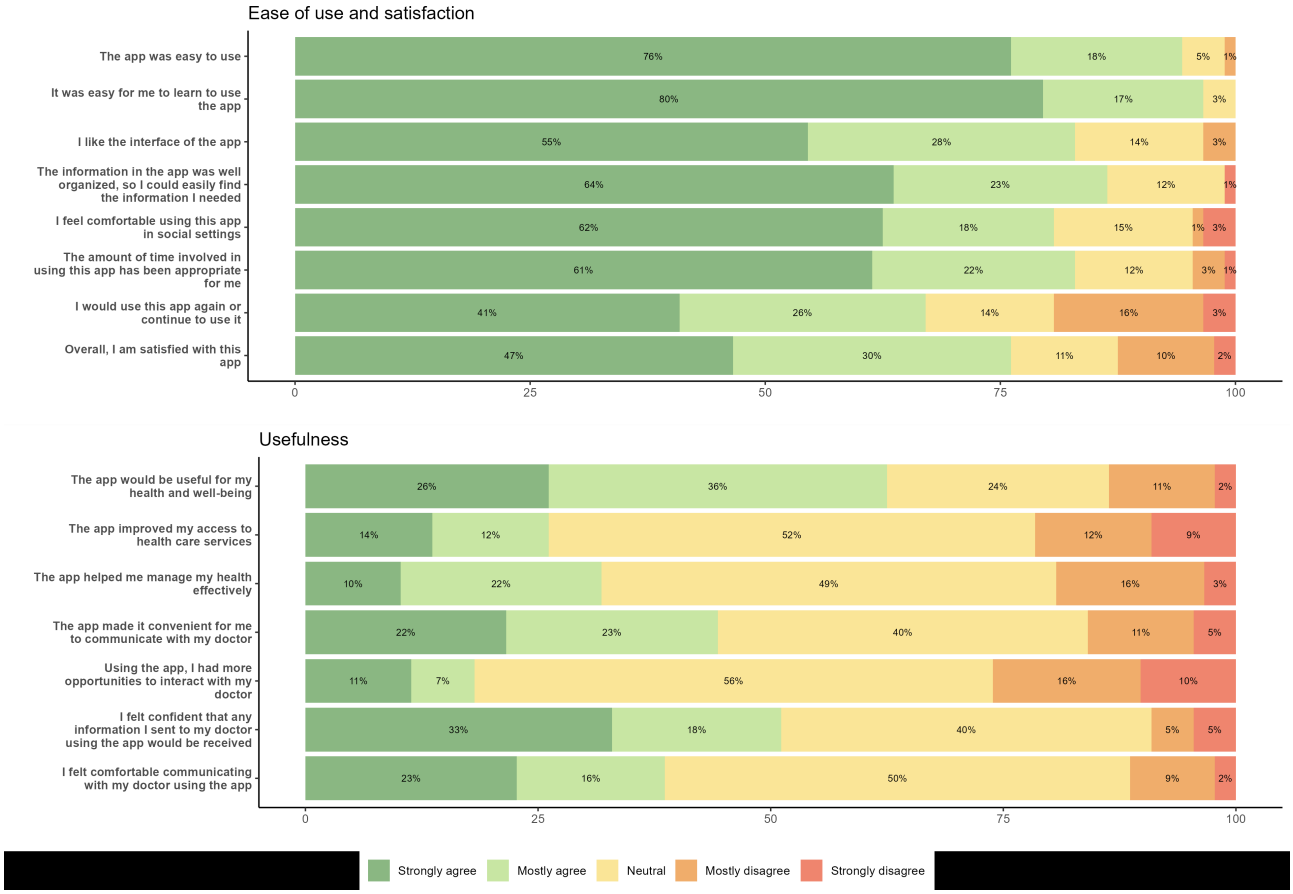

Barriers and Facilitators

We interviewed 39 patients shortly after their responses triggered a message to their rheumatologist suggesting a possible early or deferred visit. We also assessed free text comments from 44 patients who completed the final study survey. Facilitators and barriers are summarized in Table 5. The most cited benefit of the intervention mentioned by patients was that the app improved awareness of their symptoms: “I really enjoyed using this app and felt like it gave me a great way to check in with myself on my symptoms and take a pulse on how my RA was” [Patient 78]. Patients described benefits both from being aware of when their symptoms were worse and from recognizing when they were feeling well, such as when a medication change was working (“I’m in the green every time, it’s amazing!” [Patient 141]).

XSL•FO
RenderX

Table . Facilitators and barriers to rheumatoid arthritis patient engagement by intervention component based on interviews with users and open-ended survey responses.

Intervention component	Facilitators	Barriers
PRO[a] questionnaires	Increased awareness of symptoms was perceived as a benefit; graph of responses helped assess the impact of medications	• Frequency: Too often for those with well-controlled or stable symptoms, and too little for those with changing symptoms, reduced motivation to complete questionnaires • Type: Questions did not always capture the nature of symptoms; many wanted to record free text notes • Timing: Some had trouble remembering symptoms during the prior 8 days
Discussing PRO scores with rheumatologists	Discussion motivated patients to continue using the app and be aware of their symptoms	• Lack of discussion reduced motivation to continue using the app
Offers for deferred visits	Perceived time saved from avoiding unneeded visits was viewed positively	• Lack of knowledge that the app would inform visit timing reduced motivation to use the app; perception that laboratory testing needs would dictate visit timing impeded the use of deferrals; receiving care from multiple rheumatologists complicated visit deferral decisions; living in different locations during the year reduced opportunities for visit deferrals; concern about the ability to find a visit slot in a suitable time frame reduced interest in visit deferrals
Offers for early visits	Potential for early visits helped make patients feel looked after; patients trust in rheumatologist's judgment made them likely to accept an earlier visit	• Lack of clarity over whether symptoms were RA-related[b] caused confusion over whether an early visit was needed
User experience with technology	App was viewed as easy to use	• Technical issues, the need for questionnaires to be available earlier in the morning, and the lack of SMS notifications were challenges to using app

[a]PRO: patient-reported outcome.

[b]RA: rheumatoid arthritis.

Although most of the patients we spoke with did not discuss the app with their rheumatologists, many did, and most of those patients were appreciative and motivated to use the app because of it: "She said 'well I can see on this graph'...it was definitely comforting because I'm in the good part...so it was good she brought it up" [Patient 28]; "I did discuss the app data with my doctor...It was very helpful to review them with her" [Patient 180]. In a few cases, patients described examples in which decisions regarding medications were made based on discussion of the app data. In 2 cases, a patient described how the app data helped inform the decision during a visit to further increase visit intervals. Some of the patients who discussed the app data with their rheumatologists said the app added little to their care: "I was cognizant that [my doctor] looked at the app and has referred to it in discussing my condition, but she is a sufficiently good and attentive communicator that I didn't notice much in the way of changed communications with her (which I considered excellent to begin with)" [Patient 145].

However, many patients were not aware that their rheumatologist has access to or viewed the data: "My doctor and I have never discussed the use of this app. I'm not even sure she knows about the app" [Patient 144]. This reduced motivation to use the app: "I was disappointed that the app was not used at all, really, in my care. I felt that the information just went into cyberspace and was never heard from again" [Patient 119]; "I don't see any connection between the app and the care I receive with my particular rheumatologist...I found it interesting that she didn't bring it up so that gave me the message that she doesn't find it helpful in her decision making" [Patient 30]. One patient noted a disconnect between her impression of the primary purpose of the app (for her rheumatologist to oversee her PRO responses and reach out if needed) and her rheumatologist's view of the app's main purpose (for patient self-management).

Of the 5 interviewed patients who reported receiving a call offering them the option to defer their visit because of stable symptoms they reported in the app, 3 accepted the offer to defer the visit: "Not having to take the trip down to Boston from Maine, as long as [my doctor] thought was fine, we said okay and we rescheduled it [from September until November]" [Patient 200]. One of those patients had a symptom flare shortly after electing to defer the visit: "the following week I had a bad rheumatoid attack and I thought 'oh shoot I should have kept that appointment' but I think it was okay" [Patient 190]. Two

XSL•FO
RenderX

decided to keep their appointment: “when she called me I was in the middle of...about 5 different joints that were inflamed and so I said to her that I’d like to keep the visit” [Patient 145]. When asked about the potential to receive such a call, almost all patients interviewed had a positive view: “If I’m feeling well and it’s just kind of a check-in, then I wouldn’t need to drive into Boston for it” [Patient 70]; “I’d do a cartwheel...it is such a hassle [to drive into Boston]...traffic can be brutal” [Patient 141].

The only patient we interviewed who received a call from the clinic offering them the option for an earlier visit based on their PRO data believed that they did not need to come in early for their RA. When asked about the potential to receive such a call, almost all patients interviewed had a positive view: “when you have to call in...you’re hesitant to do that because you know you’re always wondering ‘oh is this something that’s just going to pass in a little while’...whereas if you guys could actually see the graph changing and call...it is comforting” [Patient 28]; “I think that would be great!...I’d feel like I’m not just spinning the wheel for nothing, answering the questions” [Patient 70]. However, a few believed they were already in good communication with their rheumatologist and would have contacted them if they needed help: “Very likely I would’ve already been there...I would’ve already thought to give her a call and let her know I’m not feeling great” [Patient 30].

About half of patients interviewed or who provided free text comments on the survey mentioned challenges related to the app questionnaire: some found them too frequent (“I did not see the value and found the number of surveys excessive and as I am in remission” [Patient 84]); some found them not frequent enough (“Every day would be better than every other day because you know if you do it every day then you know when you miss a day” [Patient 103]); and some found them inadequate for capturing their relevant symptoms (“This app did NOT capture the total wellbeing of the patient. It is strictly about joints. My RA involves my lungs” [Patient 3]).

Of the 11 rheumatologists, 10 completed the final survey (Table 6). Responses showed that the major barrier for rheumatologists not offering deferred visits was that they were already scheduling patients out as far as they were comfortable with. The complexity of rescheduling visits and the need for patients to come in for labs anyway were also notable barriers.

XSL•FO
RenderX

Table . Rheumatologist survey of reasons for not offering patients the option to defer a visit, administered after completion of the 12-month study.[a]

Reasons for not offering patients with stable PROs[b] the option to defer a visit[c]	Top ranked, n[d]	In top three, n[d]	Indicated at least once
I am too busy to spend time deciding if the visit could be delayed	0	1	Yes
Most of my patients should come in even if their PROs are stable	0	1	Yes
Most of my patients need to come in for labs anyway	1	5	Yes
Rescheduling visits makes scheduling more complicated for me and my staff	2	6	Yes
Rescheduling visits increases my workload by removing easier visits	0	0	No
I am already scheduling patients as far out as I am comfortable with	3	8	Yes
If the open slots from a delayed visit are not filled, there will be lost revenue	0	0	Yes
I do not trust the PRO data to inform the decision to delay a visit	0	0	No
Delaying a visit could complicate prior authorizations	0	2	Yes
My time needed to determine if the visit could be delayed and not reimbursed	0	1	Yes
Other[e]	0	0	Yes

[a]N=10 responses; 1 rheumatologist had left the clinic, and all others responded.

[b]PRO: patient-reported outcome.

[c]Exact wording for questions: You received in-basket messages suggesting delayed visits during the study about patients with stable PROs who had visits scheduled in 2 weeks. The messages suggested that you consider asking your assistant to call the patient and give them the option to delay the visit. (1) Of these in-basket messages, please estimate how many for which you gave the patient the option to delay the visit (choose 1 response): none, some, all. (2) For patients for whom you did not offer a delayed visit, what do you think were the primary reasons? Rank all factors that influenced you (1=most influential and so on). (3) Please add any other thoughts you have about offering patients the option to delay a visit based on their PROs. (Write as much or as little as you'd like.)

[d]"Top ranked" means the respondent listed the item as their first choice. "In top three" means the respondent listed the item as their first, second, or third choice. Some respondents did not rank their responses numerically but rather selected options using an 'x'—we did not count those in the "top ranked" and only included them in the "top three" if they selected three or fewer.

[e]One rheumatologist wrote, "some patients seem to want to come in anyway, even if they don't have to come."

Discussion

Principal Results

Our examination of intervention usage, usability, and barriers and facilitators provides some potential reasons—from both rheumatologist and patient perspectives—for why our prior study did not find an effect of an intervention designed to improve visit follow-up frequency.

From the rheumatologist perspective, although usability of the EHR-embedded dashboard was satisfactory, use of the dashboard varied substantially across rheumatologists, with most using it in a small percentage of visits. Many participating rheumatologists did not offer patients deferred visits because they were not comfortable increasing visit intervals any further, believed rescheduling visits was complicated for them and their staff, or allowed the laboratory testing schedule to determine visit timing.

From the patient perspective, most continued to use the app in the final month of the study and found the app easy to use as measured by the MAUQ_E. The MAUQ_U scores were less favorable, suggesting that patients found the app less useful than easy to use. Relatively few disagreed that it was useful, but many were neutral on its usefulness. While some interviewed patients found the app useful to improve self-awareness of their symptoms, many found the questionnaires too frequent, and one patient found that the lack of questionnaires about nonarticular symptoms was problematic. Some patients were not aware that their rheumatologist would use their reported data to determine visit timing (which was correct in some cases) or that their rheumatologist even had access to these data, which

XSL•FO
RenderX

reduced their motivation to complete the questions. However, patients were largely enthusiastic about the idea of reducing the need for visits because of the time it could save them. Although some patients liked the idea of being called for earlier visits, many were skeptical that the data they reported in the app would be useful to their rheumatologists to help inform that decision.

Implications and Comparison With Prior Work

Although we identified several rheumatologist- and patient-facing barriers, the participants who engaged in the intervention provided a proof of concept that this intervention has the potential to improve the efficiency of face-to-face visits by facilitating visit deferrals when patients have stable symptoms. Between-visit symptom monitoring has other benefits, such as helping patients be more aware of their symptoms so they can get help sooner and facilitating discussions of symptoms during visits [14,15,22-26]. This work provides early evidence that improving visit efficiency can be added to that list of potential benefits. Qualitative data and the wide range of dashboard usage by rheumatologists suggest that clinician engagement is likely the critical factor that determines the extent to which visit deferrals will be offered and patients find the app useful. Other clinically integrated digital interventions have also encountered this challenge [27-29].

Clinician-facing enhancements may include generating evidence of the benefits of further spacing out visits to improve buy-in, simplifying the patient rescheduling process, and making patients preferences for longer visit intervals available to them. Patient-facing enhancements may include personalizing the timing and topics of the questionnaires based on patient responses or specified preferences, allowing the patients to add notes to elaborate on their symptoms when the questionnaires do not adequately capture them, assessing patient interest in changing visits, and informing patients within the user interface of how their rheumatologists will use the data. These enhancements require additional design and user testing. This approach may also prove promising in other medical specialties such as pulmonology and cardiology, which will also require additional design effort.

To our knowledge, this is the first attempt to use data collected between visits to inform visit timing in the United States. Prior work in rheumatology has attempted to free up appointments using scheduled nurse phone calls but did not base decisions on longitudinally collected patient data [13]. A recent rheumatology trial in the Netherlands found that app-supported patient-initiated care achieved comparable clinical results with fewer rheumatology consults for patients with stable low disease activity [12].

Limitations

This work has several limitations. First, the intervention was implemented in one academic health center with 11 rheumatologists who lacked direct incentives to be more efficient with visit timing. Other settings with different incentives may have had different experiences with the intervention. Second, the clinical logic for when an EHR inbox message was sent to the rheumatologists suggesting early or deferred visits was determined based on the expert opinion of one rheumatologist (DHS) with review and agreement by the participating study rheumatologists rather than through empirical evaluation. Therefore, the logic may have suggested changes to visit timing that were not appropriate or missed opportunities to change visits that were appropriate. Third, we were unable to comprehensively attribute an actual early or deferred visit to an app-generated message because we did not have data for when a provider initiated a patient outreach to offer an early or deferred visit in response to a study-generated message (that data is not recorded reliably or systematically in the EHR). Finally, responses to the rheumatologist survey may be influenced by social desirability bias [30]. Specifically, it is notable that none indicated that removing easier visits would be a barrier, and few selected barriers related to revenue and reimbursement, which may nonetheless be important factors to some clinicians.

Conclusions

Informing visit timing systematically based on data collected from patients between specialty visits would, among other benefits, represent a fundamental advance in health care delivery—a move from one-size-fits-all visit timing to data-informed determinations that personalize care to individual patient needs. We examined how a PRO-focused mHealth app for RA might contribute to deferring unneeded follow-up visits. The app suggested many opportunities to defer visits, some of which resulted in rescheduling visits, though many did not. Based on our quantitative and qualitative results, clinicians had inconsistent buy-in, which patients perceived. The experience of the rheumatologists who did engage in the intervention and the patients who received the option to defer a visit provide a proof of concept that this intervention has the potential to succeed in reducing visit frequency if further enhanced based on user-centered design to improve clinician engagement and patient-perceived usefulness. It is also possible that in different health care settings, that is, nonacademic practices, or when used by clinicians practicing under different incentives (eg, capitation and performance measures relevant to access), such an app that allows for rescheduling patients may achieve greater uptake and produce significant improvements in access. Directly addressing clinician buy-in is likely the key factor and requires experimentation with different kinds of implementation strategies, such as stronger leadership support, audit feedback, and direct incentives. If successfully adopted, this type of intervention may reduce per-patient visit frequency, improve patient satisfaction, and improve access.

Acknowledgments

This study was supported by the Rheumatology Research Foundation and VERITY (P30-AR072577). RSR carried out this work independently, outside of his employment at RAND.

XSL•FO
RenderX

Data Availability

The datasets generated and analyzed during this study are not publicly available to protect the privacy of the research participants.

Conflicts of Interest

RSR has received consulting fees from the Assistant Secretary for Technology Policy at the US Department of Health and Human Services unrelated to this work. DHS receives separate research funding from CorEvitas, Janssen Pharmaceuticals, and Novartis. He also receives royalty payments from UpToDate on unrelated content. IG receives consulting fees from FPrime unrelated to this work.

References

1. Santo L, Kang K. National hospital ambulatory medical care survey: 2019 national summary tables. : Centers for Disease Control and Prevention; 2023 Jan 20. [doi: 10.15620/cdc:123251]
2. Martin AB, Hartman M, Lassman D, Catlin A, National Health Expenditure Accounts Team. National health care spending in 2019: steady growth for the fourth consecutive year. Health Aff (Millwood) 2021 Jan;40(1):14-24. [doi: 10.1377/hlthaff.2020.02022] [Medline: 33326300]
3. Javorsky E, Robinson A, Boer Kimball A. Evidence-based guidelines to determine follow-up intervals: a call for action. Am J Manag Care 2014;20(1):17-19. [Medline: 24512163]
4. Ganguli I, Wasfy JH, Ferris TG. What is the right number of clinic appointments? Visit frequency and the accountable care organization. JAMA 2015 May 19;313(19):1905-1906. [doi: 10.1001/jama.2015.3356] [Medline: 25844726]
5. Jiang Y, Rudin RS, Solomon DH. Clinical visit frequencies in rheumatology: a systematic literature review. Arthritis Care Res (Hoboken) 2023 Oct;75(10):2054-2062. [doi: 10.1002/acr.25106] [Medline: 36807719]
6. Ganguli I, Chant ED, Orav EJ, Mehrotra A, Ritchie CS. Health care contact days among older adults in traditional Medicare: a cross-sectional study. Ann Intern Med 2024 Feb;177(2):125-133. [doi: 10.7326/M23-2331] [Medline: 38252944]
4. Sequist TD. Ensuring equal access to specialty care. N Engl J Med 2011 Jun 9;364(23):2258-2259. [doi: 10.1056/NEJMe1103390]
8. Solomon DH, Dalal AK, Landman AB, et al. Development and testing of an electronic health record-integrated patient-reported outcome application and intervention to improve efficiency of rheumatoid arthritis care. ACR Open Rheumatol 2022 Nov;4(11):964-973. [doi: 10.1002/acr2.11498] [Medline: 36099161]
9. Solomon DH, Altwies H, Santacroce L, et al. A mobile health application integrated in the electronic health record for rheumatoid arthritis patient–reported outcomes: a controlled interrupted time - series analysis of impact on visit efficiency. Arthritis Rheumatol 2024 May;76(5):677-683. [doi: 10.1002/art.42774]
10. Battafarano DF, Ditmyer M, Bolster MB, et al. 2015 American College of Rheumatology Workforce Study: supply and demand projections of adult rheumatology workforce, 2015–2030. Arthritis Care Res (Hoboken) 2018 Apr;70(4):617-626. [doi: 10.1002/acr.23518]
11. Hewlett S, Kirwan J, Pollock J, et al. Patient initiated outpatient follow up in rheumatoid arthritis: six year randomised controlled trial. BMJ 2005 Jan 22;330(7484):171. [doi: 10.1136/bmj.38265.493773.8F] [Medline: 15546895]
12. Seppen B, Wiegel J, Ter Wee MM, et al. Smartphone-assisted patient-initiated care versus usual care in patients with rheumatoid arthritis and low disease activity: a randomized controlled trial. Arthritis Rheumatol 2022 Nov;74(11):1737-1745. [doi: 10.1002/art.42292] [Medline: 35818342]
13. Butt S, Newman E, Smith N. Nurse scheduled telephone visit: the right rheumatology care for the right patient at the right time. Arthritis Rheumatol 2016;68(Suppl 10) [FREE Full text]
14. Colls J, Lee YC, Xu C, et al. Patient adherence with a smartphone app for patient-reported outcomes in rheumatoid arthritis. Rheumatology (Sunnyvale) 2021 Jan 5;60(1):108-112. [doi: 10.1093/rheumatology/keaa202]
15. Lee YC, Lu F, Colls J, et al. Outcomes of a mobile app to monitor patient-reported outcomes in rheumatoid arthritis: a randomized controlled trial. Arthritis Rheumatol 2021 Aug;73(8):1421-1429. [doi: 10.1002/art.41686] [Medline: 33559338]
16. Zhou L, Bao J, Setiawan IMA, Saptono A, Parmanto B. The mHealth App Usability Questionnaire (MAUQ): development and validation study. JMIR Mhealth Uhealth 2019 Apr 11;7(4):e11500. [doi: 10.2196/11500] [Medline: 30973342]
17. Bangor A, Kortum PT, Miller JT. An empirical evaluation of the System Usability Scale. Int J Hum Comput Interact 2008 Jul 29;24(6):574-594. [doi: 10.1080/10447310802205776]
18. Lewis JR. The System Usability Scale: past, present, and future. Int J Hum Comput Interact 2018 Jul 3;34(7):577-590. [doi: 10.1080/10447318.2018.1455307]
19. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005 Nov;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
20. Gale NK, Heath G, Cameron E, Rashid S, Redwood S. Using the framework method for the analysis of qualitative data in multi-disciplinary health research. BMC Med Res Methodol 2013 Sep 18;13(1):117. [doi: 10.1186/1471-2288-13-117] [Medline: 24047204]
21. Hyzy M, Bond R, Mulvenna M, et al. System Usability Scale benchmarking for digital health apps: meta-analysis. JMIR Mhealth Uhealth 2022 Aug 18;10(8):e37290. [doi: 10.2196/37290] [Medline: 35980732]

XSL•FO
RenderX

22. Rudin RS, Fanta CH, Qureshi N, et al. A clinically integrated mHealth app and practice model for collecting patient-reported outcomes between visits for asthma patients: implementation and feasibility. Appl Clin Inform 2019 Oct;10(5):783-793. [doi: 10.1055/s-0039-1697597] [Medline: 31618782]
23. Rudin RS, Fanta CH, Predmore Z, et al. Core components for a clinically integrated mHealth app for asthma symptom monitoring. Appl Clin Inform 2017 Oct;8(4):1031-1043. [doi: 10.4338/ACI-2017-06-RA-0096] [Medline: 29241243]
24. Sulca Flores JA, Dalal AK, Sousa J, et al. Evaluation of a primary care-integrated mobile health intervention to monitor between-visit asthma symptoms. Appl Clin Inform 2024 Aug;15(4):785-797. [doi: 10.1055/s-0044-1788978] [Medline: 39357877]
25. Meirte J, Hellemans N, Anthonissen M, et al. Benefits and disadvantages of electronic patient-reported outcome measures: systematic review. JMIR Perioper Med 2020 Apr 3;3(1):e15588. [doi: 10.2196/15588] [Medline: 33393920]
26. Bartlett SJ, De Leon E, Orbai AM, et al. Patient-reported outcomes in RA care improve patient communication, decision-making, satisfaction and confidence: qualitative results. Rheumatology (Oxford) 2020 Jul 1;59(7):1662-1670. [doi: 10.1093/rheumatology/kez506] [Medline: 31665477]
27. Rudin RS, Thakore N, Mulligan KL, Ganguli I. Addressing the drivers of medical test overuse and cascades: user-centered design to improve patient-doctor communication. Jt Comm J Qual Patient Saf 2022 Apr;48(4):233-240. [doi: 10.1016/j.jcjq.2022.01.005] [Medline: 35177360]
28. Ganguli I, Mulligan KL, Chant ED, et al. Effect of a peer comparison and educational intervention on medical test conversation quality: a randomized clinical trial. JAMA Netw Open 2023 Nov 1;6(11):e2342464. [doi: 10.1001/jamanetworkopen.2023.42464] [Medline: 37943557]
29. Greenhalgh T, Wherton J, Papoutsi C, et al. Analysing the role of complexity in explaining the fortunes of technology programmes: empirical application of the NASSS framework. BMC Med 2018 May 14;16(1):66. [doi: 10.1186/s12916-018-1050-6] [Medline: 29754584]
30. Kreuter F, Presser S, Tourangeau R. Social desirability bias in CATI, IVR, and web surveys: the effects of mode and question sensitivity. Public Opin Q 2008 Dec 1;72(5):847-865. [doi: 10.1093/poq/nfn063]

Abbreviations

EHR: electronic health record
MAUQ: mHealth App Usability Questionnaire
MAUQ_E: mHealth App Usability Questionnaire for Ease of Use
MAUQ_U: mHealth App Usability Questionnaire for Usefulness
mHealth: mobile health
PRO: patient-reported outcome
RA: rheumatoid arthritis
SUS: System Usability Scale

Edited by A Mavragani; submitted 02.12.24; peer-reviewed by J Knitza, M Bass; revised version received 28.01.25; accepted 28.01.25; published 21.04.25.

Please cite as:
Rudin RS, Santacroce LM, Ganguli I, Solomon DH
Tailoring Rheumatoid Arthritis Visit Timing Based on mHealth App Data: Mixed Methods Assessment of Implementation and Usability
JMIR Form Res 2025;9:e60854
URL: https://formative.jmir.org/2025/1/e60854
doi:10.2196/60854

© Robert S Rudin, Leah M Santacroce, Ishani Ganguli, Daniel H Solomon. Originally published in JMIR Formative Research (https://formative.jmir.org), 21.4.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Developing an Online Community Advisory Board (CAB) of Parents From Social Media to Co-Design an Human Papillomavirus Vaccine Intervention: Participatory Research Study

Regan M Murray[1], MPH; Shawn C Chiang[2], MPH, PhD; Ann C Klassen[3], PhD; Jennifer A Manganello[4], MPH, PhD; Amy E Leader[5], MPH, DrPH; Wen-Juo Lo[6], MA, PhD; Philip M Massey[7], MPH, PhD

[1]Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, United States

[2]Department of Health Behavior, Texas A&M University School of Public Health, College Station, TX, United States

[3]Department of Community Health and Prevention, Drexel University, Philadelphia, PA, United States

[4]Department of Health Policy, Management and Behavior, University at Albany, State University of New York, Albany, NY, United States

[5]Division of Population Science, Department of Medical Oncology, Thomas Jefferson University, Philadelphia, PA, United States

[6]Department of Counseling, Leadership, and Research Methods, University of Arkansas, Fayetteville, AR, United States

[7]Department of Community Health Sciences, University of California, Los Angeles, 650 Charles E Young Drive S, 4th Fl, Los Angeles, CA, United States

Corresponding Author:
Philip M Massey, MPH, PhD
Department of Community Health Sciences, University of California, Los Angeles, 650 Charles E Young Drive S, 4th Fl, Los Angeles, CA, United States

Abstract

Background: Social media health interventions have grown significantly in recent years. However, researchers are still developing innovative methods to meaningfully engage online communities to inform research activities. Little has been documented describing this approach of using online community advisory boards (CABs) to co-create health communication interventions on social media.

Objective: This study describes the formation, engagement, and maintenance of an online CAB focused on co-creating a health education intervention for parents regarding the human papillomavirus (HPV) vaccine. The study provides guiding principles for public health researchers implementing such CABs in future digital health interventions.

Methods: In May 2020, Twitter was used to recruit parents of children aged 9‑14 years, who were active users of the platform and were interested in serving on a CAB focused on child health and online programs. The recruitment campaign included Twitter (rebranded as X in 2023) advertising tools (eg, “interests” and “audience look-a-likes”). A total of 17 parents completed a screening survey and 6 completed a follow-up phone interview. Following phone interviews, 6 parents were invited to join the CAB, where they committed to a 1-year involvement. The CAB participated in eleven 1-hour online meetings in the first year, contributing to monthly feedback through participatory workbooks. Long-term engagement was sustained through icebreakers and casual online interactions, as well as providing real-time updates to demonstrate CAB feedback integration. An anonymous midterm evaluation was conducted at the end of the project’s first year to assess processes and identify future growth opportunities.

Results: A total of 6 parents (5 females and 1 male) with children aged 9-14 years from diverse racial and ethnic backgrounds (African American, South Asian American, and White) across 6 states in the United States, representing urban, suburban, and rural areas, agreed to serve as CAB members. All 6 CAB members committed to 1 year of service beginning in July 2020 with 4 extending their participation into a second year (August 2021-August 2022). The CAB provided expert insights and feedback to co-develop the intervention, including character development, narrative content creation, study recruitment, survey development, and intervention delivery. The midterm evaluation showed 100% (6/6) satisfaction among CAB members, who valued the connections with other parents and their contribution to research. While all members felt confident discussing HPV, 83% (5/6) suggested diversifying the group and increasing informal bonding to enhance engagement and inclusivity, especially for differing vaccination views.

Conclusions: This study demonstrates that online CABs are a highly effective model for co-creating and informing online health communication interventions. The engagement of parents from diverse backgrounds and the structured use of online tools (eg, interactive workbooks) creates a constructive and thoughtful environment for incorporating parent contributions to research. This study highlights guiding principles to forming, engaging, and maintaining an online CAB to enhance health research and practice.

(JMIR Form Res 2025;9:e65986) doi:10.2196/65986

XSL•FO
RenderX

KEYWORDS

online community advisory boards; community engagement; social media; digital health; digital health intervention; HPV vaccine; human papillomavirus; HPV; parent health; child health

Introduction

Community advisory boards (CABs) are a widely integrated methodology in participatory health research to understand community needs, advise research, and ultimately improve health behaviors or health outcomes [1]. Comprising community members, leaders, stakeholders, and others, CABs are united by a common health interest, identity, status, or experience [2]. In the field of public health, CABs are integral to research programs to provide unique perspectives and guidance related to proposed research questions and interventions. Researchers often implement participatory strategies and inspire innovative discussions that lead to novel approaches to address health-related issues that are aligned with community needs [3]. Over time, as community needs evolve and change, CABs can also shape the development and adaptation of interventions to better align with those identified needs. Engaging community members early and often in the research process is an effective strategy to increase efficacy of the designed intervention, enhance transparency, and reduce population-level disparities [4].

In the context of child health, advisory boards consisting of parents or caregivers are essential for ensuring that public health research and interventions are aligned with family values and needs [5-7]. CABs comprising parents and caregivers have addressed a wide range of public health topics including but not limited to asthma [8], Early Head Start programs [9], childhood obesity [10], and childhood vaccination [11]. For parents, opportunities to engage in research focused on child health have been shown to be driven by altruistic motivations, previous experience in research or health care, and desire to share personal perspectives [12]. Parents are often the ultimate decision makers when it comes to their child's health; incorporating their input and understanding their perspectives provides important insights into facilitating the adoption of health behaviors among children.

As public health research and practice are incorporated more into the social media and online environment, the need for online CABs is essential. For parents, online and internet-delivered health interventions targeting child health continue to grow [13-15] and CABs will be an important approach to engage and work alongside diverse communities. Already, online opportunities to engage parents in research have allowed participation of varied perspectives across diverse geographies and facilitated ease of involvement by working around demanding parent schedules [12]. By shifting to the online environment, research teams can engage diverse populations through flexible and accessible means. Heightened flexibility allows researchers to swiftly adapt to the evolving needs of the community, as well as the ever-evolving nature of social media platforms and the online environment.

Engaging with diverse populations with varied perspectives is crucial to fostering trust, developing culturally relevant and accepted interventions, and encouraging participation in health behavior interventions [16]. This is further bolstered by expanding into the sphere of community-based participatory research through online platforms. Furthermore, as elements of everyday life increasingly transition to online platforms, the need for robust online CABs becomes paramount. This aligns with the broader trend of hybrid and remote work environments following the pandemic [17], offering flexibility that accommodates the diverse schedules and locations of members, similar to how many jobs and careers have adapted to remote work.

While the need to integrate online CABs into public health research is apparent, particularly as social media becomes an increasingly influential health intervention tool, there is a gap in evidence regarding the process of developing and maintaining online CABs. Addressing this gap is essential to promote thoughtful and systematically replicable online CABs in future public health efforts. Guiding frameworks play an important role in the development of effective CABs by providing researchers with structure and strategies for engaging CAB members from creation to evaluation [18]. Scholars interested in evaluating processes related to community-engaged advisory strategies have adapted earlier frameworks to meet the needs of systematically evaluating CABs [2]. The framework by Newman et al [2] identifies 3 primary domains that support the necessary functions of community advisory boards: formation, operation, and maintenance. Other researchers have found utility in following this guidance as well [19]. This study uses the framework developed by Newman et al [2] with some slight modifications in terminology; we refer to the "operation" domain labeled by Newman as "engagement."

This case study contributes to limited evidence on approaches and procedures for developing online CABs. We provide a case study, focusing on the development of human papillomavirus (HPV) vaccine messages for parents who use Twitter (rebranded as X in 2023), a microblogging social media platform, to share evidence and findings to support the development of future online CAB creation. Briefly, the HPV vaccine is an important topic for child and adolescent health, and parents are critical stakeholders in deciding whether their child receives the HPV vaccine. The HPV vaccine is routinely recommended for males and females between the ages of 9 and 12 years and can also be administered up to the age of 26 years and older in some cases [20]. The vaccine protects against certain strains of HPV that are associated with various cancers, including cervical, anal, throat, and penile cancers, as well as genital warts [20]. Parent involvement in online CABs is supported by previous studies showing that the inclusion of parents in HPV vaccine communication interventions resulted in improved vaccination rates [21], decreased vaccine hesitancy [22], as well as increased education and knowledge [23].

In this study, we describe how we assembled an online CAB consisting of parents from across the United States to serve in

XSL•FO
RenderX

an advisory role in the development of an online intervention. The goal of the CAB was to co-create a social media intervention on Twitter to serve as an information source for parents when deciding to get the HPV vaccine for their child. We used the framework by Newman et al [2] of formation, engagement, and maintenance to organize our Methods, Results, and Discussion sections. The purpose of this study is to describe promising strategies to be used and adapted for the creation of online CABs, ultimately to continue to support the development of participatory online health research and practice.

Methods

Formation

To initiate the creation of the online CAB, the study team first clearly defined the purpose, roles, expected functions, and membership and recruitment strategies tailored to an online format [2]. It was determined the CAB would assume an advisory role to provide support, feedback, and review of digital materials based on the project's needs. Given the study's focus on developing an online digital health intervention delivered through Twitter, recruitment was purposefully conducted on the same social media platform to align with the target audience. To recruit potential CAB members, the study team leveraged Twitter Ads Manager's Promoted Ad feature to create and disseminate targeted advertisements aimed at parents with children aged 9 - 14 years in the United States. At the time this study took place, the platform allowed extensive targeting capabilities and enabled the team to specify demographic filters and audience interests for purposive recruitment. Figure 1 shows an example of these recruitment ads.

In May 2020, parents were recruited using Twitter's "Promoted Ad" feature, which allowed the team to target broad user groups by specifying specific demographics, interests, and online behaviors and connections [24], making it a useful tool for online CAB recruitment as well as study participant recruitment more broadly. All advertisements were tailored based on age, gender, location, and language. To maximize outreach and engagement within the online environment, the study team optimized targeting by filtering for specific "keywords," "interests," "conversation topics," and "follower look-alikes" to reach parents with children aged 9 - 14 most likely to engage with the intervention platform. Table 1 provides a detailed breakdown of ad parameters, including follower look-alikes targeting, where the ads targeted parents with similar interests to followers of popular parent-related accounts on Twitter.

Interested parents were directed to an online screening survey linked within the promoted ads. The survey collected demographic information, Twitter and other social media usage patterns, and HPV vaccine awareness. Multimedia Appendix 1 provides the full list of screening questions. Parents who met the selection criteria (ie, used Twitter at least once a week, had a child aged 9 - 14 years, and agreed to be contacted by the study team) were then invited to online interviews conducted through videoconferencing by a member of the study team.

The interview guide, provided in Multimedia Appendix 2, collected information about the individual's background, interest in the project topic (ie, HPV vaccination), and assessed their fit as potential CAB members. The interviews also served as an opportunity for the study team to provide more detailed information about the project, the online nature of the CAB, member responsibilities, compensation, and next steps. CAB responsibilities described included the anticipated time commitment (ie, 10 online meetings over one year beginning in August 2020).

Figure 1. Recruitment advertisement graphic used on Twitter to recruit parents to the online community advisory board in May 2020.

XSL•FO
RenderX

Table . Twitter ad parameters used for recruiting parents to the community advisory board in May 2020.

Parameter	Definition	Ads for mothers	Ads for fathers
Sex	__[a]	Female	Male
Age range (years)	—	25-54	25-54
Location	—	United States	United States
Keywords	Target or exclude people who searched for, tweeted with, or engaged with keywords.	• Mom • Parent • Family • Health • Parenting	• Dad • Parent • Health • Parenting • Family
Interests	Choose broad or narrow interests to reach the right people	• Parenting teens • Health news and general info	• Parenting teens • Health news and general info
Conversation topics	Target people who tweet about or engage with specific conversation topics.	• Parenting teens	• Parenting teens
Follower look-alikes	Target people with interests similar to an account's followers. For example, enter @TwitterAds to target people likely to be interested in advertising on Twitter.	• @ScaryMommy • @AmerAcadPeds • @MomCentral • @RookieMoms	• @FatherlyHQ • @LifeofDadShow • @AmerAcadPeds

[a]Not available.

Engagement

CAB members collaborated with the study team to establish key activities, leadership structure, and operating principles. The group collectively determined meeting frequency, refined member roles and responsibilities, and reviewed the online research processes, including obligations to the University's Institutional Review Board, as well as the study design and procedures. The remote nature of the CAB allowed for us to adopt a flattened hierarchy to ensure equal responsibility and input among members without a formal leadership structure.

Recognizing the importance of capturing meaningful, systematic feedback in an online setting, the study team developed interactive online workbooks that included questions, examples, and open-ended prompts to elicit parent input on project materials. These workbooks were shared digitally with CAB members one week before each scheduled monthly meeting, as members indicated this to be the optimal time frame for preparation. Completed workbooks were analyzed by the study team to identify themes for discussion during meetings. The workbooks served as a structured and accessible tool for gathering organized feedback on topics such as parenting experiences and HPV vaccine perceptions. Multimedia Appendix 3 provides an example workbook used to solicit input and feedback on developing non-narrative messages for the intervention.

For virtual meeting logistics, each session began with a short icebreaker to build rapport among members, workbooks were then discussed at length, giving CAB members an opportunity to provide real-time feedback on workbook topics. All meetings were recorded, and meeting minutes were distributed digitally to both the study team and CAB members, ensuring those unable to attend (which was rare) could stay updated. This online approach enabled flexibility and sustained engagement, accommodating busy schedules and parental responsibilities.

Feedback from the CAB was incorporated iteratively into the study. The study team maintained transparency through regular online updates, presenting how CAB suggestions influenced the intervention design and delivery. This iterative process strengthened trust and collaboration in a fully online setting.

Maintenance

To evaluate the success and identify challenges encountered by CAB members during the first year, an anonymous midterm was conducted in July 2021, using an online survey and videoconference discussions at the end of the first year. The evaluation included measures examining the overall member satisfaction, meeting logistics, quality of activities and involvement, and what they envisioned for the future of the CAB. Multimedia Appendix 4 provides the full list of midterm evaluation questions.

To encourage sustained engagement, multiple digital strategies were used, including compensating members for their time and participation (ie, US $40 for workbook preparation and meeting attendance), conducting regular online meetings (ie, monthly during the first year) through videoconferencing, and sharing annual progress presentations to highlight study progress and CAB contributions to the study. To further celebrate their involvement, CAB members were provided a certificate of membership at the conclusion of the 1-year commitment. CAB members were invited to continue their participation for another year, extending their advisory role until August 2022. After completing the midterm evaluation, CAB activities were adjusted based on feedback collected and the remaining activities in the project lifecycle. Figure 2 provides an overview of the CAB activity timeline.

XSL•FO
RenderX

Figure 2. Project timeline of community advisory board activities (May 2020-August 2022). CAB: community advisory board.

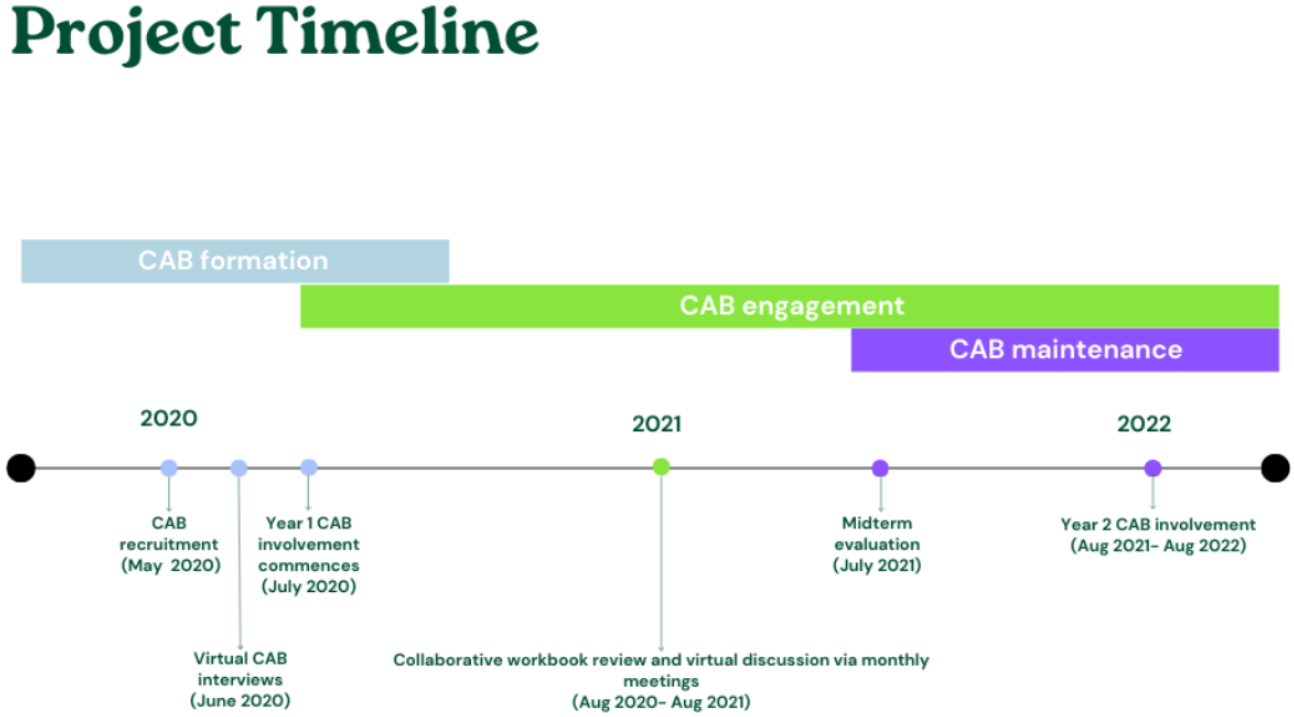

Ethical Considerations

Evaluation activities were deemed exempt by the University of Arkansas Institutional Review Board (protocol 2108351667). All results presented in this manuscript are anonymous or deidentified, with no identification of individual CAB members included in the paper or supplementary materials. In addition, CAB members were compensated for their time and participation in the project, receiving US $40 in e-gift cards per meeting (US $20 for completing a workbook and US $20 for attending the meeting).

Results

Formation

In May 2020, as a part of a targeted social media recruitment campaign on Twitter, 17 parents completed the CAB screening survey. Of the 17, 6 responded to requests for an interview through videoconferencing. Following these interviews, all 6 parents were invited to join the parent advisory board, with all 6 accepting the offer. Table 2 details the CAB membership consisting of 5 mothers and 1 father, reflecting diversity across several key dimensions. The CAB represented diversity in terms of geography (Texas, Illinois, Pennsylvania, Indiana, Georgia, and Washington DC), community-type (urban, suburban, and rural), vaccination experience (half have vaccinated their child, though none were against getting the HPV vaccine for their child), and race and ethnicity (Black or African American, South Asian, Hispanic, and White).

Parents committed to 1-year involvement, which included participation in monthly online meetings. At the conclusion of the first year, 2 members stepped down due to personal commitments, while 4 members elected to remain active advisors to the project for an additional year, until the project's completion.

Table . Community advisory board demographic composition (deidentified), July 2020.

Parental role	Race and ethnicity	Location	Community type	Vaccination perspective
Mother	Black	Washington, DC	Urban	Vaccine advocate
Mother	White	Indiana	Rural	Vaccine advocate
Mother	White	Georgia	Suburban	Vaccine hesitant
Mother	Hispanic	Pennsylvania	Urban	Vaccine advocate
Mother	South Asian	Maryland	Suburban	Vaccine advocate
Father	Black	Texas	Urban	Vaccine advocate

Engagement

During the first year of the study, the CAB actively participated in 11 1-hour meetings conducted through Zoom (Zoom Communications). Meetings were held in the evening, typically 7 PM Eastern Time, to best accommodate parents' schedules across time zones. As the study transitioned into its second year and moved into the implementation phase, meeting frequency was reduced to six 1-hour meetings over one year. This adjustment reflected the reduced need for iterative feedback.

XSL•FO
RenderX

The CAB provided valuable feedback on a range of HPV vaccine topics informed by their lived experiences as parents. Key contributions included identifying scientifically relevant messaging for the HPV vaccine, developing personas and refining the storylines used in the narrative communication intervention [25], co-creating Twitter messages, and reviewing study recruitment ads. Table 3 describes how CAB input informed various aspects of the intervention, as well as survey refinement and pilot testing of materials.

CAB involvement ultimately led to the development of a 2-week HPV vaccine intervention delivered on Twitter. The intervention included eight topics related to the HPV vaccine: (1) character introductions, (2) vaccine normative beliefs, (3) vaccine knowledge and awareness, (4) vaccine accessibility, (5) vaccine safety, (6) HPV vaccine for boys, (7) vaccine disparities and equity, and (8) cancer prevention. Multimedia Appendix 5 provides examples of the first two topics, or "chapters," of the program, entitled "Meet the parents!" and "Parents are getting their kids the HPV vaccine."

XSL•FO
RenderX

Table . Examples of community advisory board's impact on quality, reach, and outcomes of the intervention (July 2020-August 2022).

Topic	Timeframe	CAB[a] activity goal	CAB feedback and integration
CAB formation and project integration	July 2020-December 2021	• Create CAB operating procedures • Inform study activities with CAB input and feedback.	• "Having the full study team present is not a problem-it is actually an advantage to be able to put names and faces together" • "I really enjoy the breakout rooms and feel we can speak about the topic but also get to know the other members a bit better." • "I like the mix of face to face zoom meetings and pdf workbooks."
Persona development	January 2021-May 2021	• Alter personas to be more relevant and relatable.	• "Many elements of this [persona] description remind me of myself or some in my closest circle." • "As I was reading [the persona descriptions], I could name specific people in my life who fit the characteristics of each character."
Intervention content creation	April 2021-August 2021	• Identify key themes for presenting information in the intervention. • Review narrative intervention content.	• "Knowledge and awareness has to come first to introduce the HPV vaccine into the general conversation. With information about accessibility, the audience can learn that [HPV is] not something that only certain people can get…" • "Love the images and think they will be effective for communities of color in particular and all parents in general." • "The use of gifs and memes is great."
Survey development	April 2021-December 2021	• Ensure questions are relevant to parents and their vaccine decision making.	• "People have very different relations with their providers and that plays a role in their decision making, making it important to include."
Study recruitment	November 2021-August 2022	• Develop recruitment ads with relevant language and images.	• "I'd like more graphics to share for recruitment"
Intervention delivery	December 2021-August 2022	• Advise on the various ways parents use Twitter. • Advise on the logistics of implementing the intervention on Twitter.	• "The order of tweets showing up in the feed is confusing and out of order." • "Using [Twitter] threads [to deliver the intervention], makes the most sense."

[a]CAB: community advisory board.

Maintenance

Table 4 shows the summarized themes from the midterm evaluation, and the full list of evaluation questions is provided in Multimedia Appendix 4.

XSL•FO
RenderX

Table . Qualitative themes of the anonymous community advisory board (CAB) midterm evaluation (July 2021).

Theme	Subtheme	Illustrative finding
Function of the CAB[a]	• Frequency • Convenience	• Online meetings after normal business hours work best. • Parents are happy with the meeting time and meeting frequency, and some would like to meet more often.
CAB research engagement	• Confidence • Knowledge building	• All parents expressed confidence in describing the project and purpose of the advisory board. • Everyone feels a great deal more knowledgeable about HPV[b] and the HPV vaccine since joining the advisory board. • All parents thought their input was well incorporated into the project.
Quality of the CAB	• Comfort • Collaboration	• Parents largely feel safe and comfortable sharing ideas and opinions. • Some discomfort sharing unpopular opinions (eg, vaccine hesitancy). • Parents really enjoy working with one another and appreciate the connection they have with other board members.
Next steps for the CAB	• Increased diversity	• All parents feel that it would be beneficial to recruit additional, more diverse people to join the advisory board. • Parents have an interest in participating in additional research activities

[a]CAB: community advisory board.

[b]HPV: human papilloma virus.

The midterm evaluation highlighted a high level of satisfaction among CAB members regarding their experience as advisory board members. Members particularly valued the connections they formed with other parents and their ability to contribute meaningfully to research in real time. All 6 CAB members reported that the monthly meeting frequency, timing, and format were well-suited to their needs as parents. However, 2 members suggested that additional meetings could further strengthen their connections with other members.

> *[Meeting logistics are] just right. It would be nice to maybe have a meet-up that isn't HPV-related so we can bond more. I really enjoy getting to know everyone.* [CAB member 4]
>
> *I don't know that anything can be done to improve the format. I think the team has done a very good job of facilitating the meetings.*
>
> *[CAB member 1]*
>
> *I love the communication! Everyone is so nice and fun to speak with. Even when we disagree it is never an issue. Maybe we should have a group chat.* [CAB member 5]

CAB members unanimously expressed confidence in describing the project to other parents and indicated that they gained more knowledge related to HPV and the HPV vaccine since joining the advisory board.

> *I personally felt like I was knowledgeable about HPV and the vaccine but this experience helps me understand it on a more deeper and informative level.* [CAB member 4]
>
> *The presentation we had regarding the HPV vaccine helped give me a clear understanding of the history, efficacy and uses for the vaccine. I have used some of the information that I learned and have used it while talking to friends and family about the HPV vaccine.* [CAB member 3]
>
> *I think I just took for granted that I knew enough about HPV and the HPV vaccine to understand it as being one more required shot kids need. I assumed it to be a simple STD, but now appreciate the wider repercussions of not getting vaccinated against it.* [CAB member 5]

The most enjoyed aspects or activities of the CAB members were expressed as: being part of a group that is improving the lives of their children, learning about HPV vaccination, learning about other parent perspectives, and seeing the project grow, being a part of a parent-led group, contributing to scientific research, and learning about HPV more broadly.

XSL•FO
RenderX

> *That I am actually a part of something that will help my son and generations to come.* [CAB member 5]
>
> *It has been really enjoyable to watch the project grow from the beginning to where it is now. I've also enjoyed hearing everyone's different perspectives.* [CAB member 3]
>
> *Our group was able to create an authentic appreciation of the importance for work like this. It's not just another social media survey, but real scientific research.* [CAB member 4]

Despite the positive feedback, areas of improvement were identified. While all 6 members expressed comfort with sharing their input with the full project team and CAB members, 5 felt more comfortable doing so during the smaller breakout sessions. This underscores the need to create a safe and inclusive space for parents to voice diverse or unpopular opinions, including vaccine hesitancy. In addition, members emphasized the importance of recruiting more parents with diverse backgrounds and perspectives to enrich group discussion, particularly those with differing views on vaccination.

> *I do think we need new members to diversify the group and add new voices and ideas.*
>
> *[CAB member 3]*
>
> *I feel that most everyone on the team is in complete support of the vaccination and is not really open to my questioning of the vaccinations.*
>
> *[CAB member 2]*
>
> *It always takes me a little while to warm up to new groups, but a tone of openness and support was established early on that made our group a safe space!* [CAB member 6]

Discussion

Principal Findings

This study describes strategies used for forming, engaging, and maintaining an online CAB for a parent-focused HPV vaccine intervention developed for social media. We strategically leveraged social media tools to recruit parents from the same social media platform where the intervention would be delivered (ie, Twitter). To engage with the CAB, we implemented interactive workbooks to ensure that every member had opportunities to provide input and feedback on intervention content and study direction. These digital tools streamlined the collection of data and facilitated meaningful discussions despite geographic barriers. An anonymous midterm evaluation was conducted to assess the first year's outcomes and build upon successes, with 4 of the 6 CAB members continuing to advise on the project during its final second-year phase.

By using social media–based strategies to establish the CAB, we successfully increased the relevance of health materials for parents, improved the interventions reach, and enhanced its overall quality. The recommendations from the CAB significantly influenced the development and delivery of the intervention, while the online format enabled greater flexibility and accessibility for participants. To guide this process, we applied and adapted the Newman et al (2011) framework for community advisory boards, synthesizing best practices of community-based participatory research to focus on three key domains: formation, engagement, and maintenance. These domains provided an essential structure for successful operations, ensuring that the virtual CAB we formed effectively represented community interests and contributed effectively to the research process [2].

Online CAB Formation

We found the need to clearly define the unique purpose, roles, functions, and responsibilities of an online advisory board to be paramount and supported by previous work [26]. There was great utility in applying Newman's framework, which was originally designed for in-person CABs, to the development of online CAB [2]. While it can be challenging to build trust, transfer accurate knowledge, and maintain communication in in-person interactions, these challenges can be further amplified in the online environment as other researchers have also noted [19,27,28]. To address this, our approach emphasized establishing clear digital communication strategies, fostering a sense of community that grew beyond the project, and ensuring that all members felt equally involved and respected. Our study builds on these principles and details the promising approaches we encountered so that future research can replicate our processes of building a supportive environment and collaborative spirit, as online interventions and online advisory boards continue to expand.

Social media played a critical role in the recruitment and formation process of the CAB and this finding adds to the growing body of evidence supporting this important and effective tool [27]. By using Twitter's targeted advertising tools, in addition to our recruitment survey and interviews, we reached a population of parents with diverse thoughts, backgrounds, and parental experiences. We prioritized the recruitment of vaccine naïve parents who were interested in learning about the HPV vaccine. Notably, intention to vaccinate their child was not a factor in determining who was interviewed or selected for membership, allowing us to include a wide range of perspectives. This decision draws from research demonstrating the need to be intentional in building an advisory board with diverse experiences and perspectives [29]. By embracing this diverse, flexible, and forward-thinking approach, we were able to create an online CAB that was resilient, adaptable, and invested in the project's success.

Recruiting advisory board members in May 2020, at the onset of the COVID-19 pandemic, presented unique challenges to research also experienced in other studies [28]. While most of society was just beginning to adapt to changes in work and communication structures, this project had already committed to forming an online CAB to align with the online nature of the intervention. This early commitment to the online CAB and the parents' intention to participate in a child-health focused research topic, likely contributed to strong engagement with the project tasks despite the quickly changing nature of the pandemic, especially in early 2020. In addition, the digital nature of the CAB and intervention allowed the project to move forward seamlessly during periods of mandatory lockdowns

XSL•FO
RenderX

and rapidly changing guidance. This decision to prioritize an online framework also highlighted the usability and effectiveness of online CAB models in health research.

Online CAB Engagement

Flexibility was vital for effective online CAB engagement. For parents, various factors such as work, holidays, long weekends, school events, personal obligations, or religious observations can affect their ability to respond timely, attend meetings, and complete tasks. By anticipating these factors, researchers can enhance participation and sustain engagement. To accommodate these needs, we used strategies such as asynchronous activities, including digital workbooks, to collect feedback systematically and reliably. This approach builds from the successes of other studies, demonstrating the effectiveness of flexibility and asynchronous activities [30]. These workbooks allowed CAB members to provide input at their convenience and served as a reliable tool for integrating their feedback into the intervention.

During online meetings, the use of videoconferencing tools facilitated dynamic discussions and collaborative decision-making. This approach reinforced expectations of ongoing participation and capitalized on the momentum of active collaboration. Meetings were recorded and made available to those who could not attend in real time (though this was rare), further enhancing flexibility and engagement.

Engaging parents through Twitter also provided an important avenue for buy-in and oversight of research activities. Parents' familiarity with the platform facilitated a sense of ownership and connection to the intervention, which was ultimately delivered on the same social media platform. Their breadth of experiences and depth of understandings provided important grounding in the study's health messaging element and enriched its development. This approach demonstrated the ability to use social media not only as a recruitment tool but also as a means for ongoing engagement with community stakeholders.

Online CAB Maintenance

Maintaining and sustaining an online CAB required a strong emphasis on diversity, transparency, and active engagement to ensure the board's long-term effectiveness and participation. The small membership size (n=6), allowed for personalized interactions and the cultivation of strong connections among members. "Icebreakers" and informal conversations during meetings were particularly effective for building camaraderie, uncovering common interests, and strengthening the sense of community within the CAB, an approach that has demonstrated effectiveness across studies and populations [2,26,28].

Frequent meetings were instrumental in explaining project activities and tasks, the purpose tasks (ie, workbooks), and demonstrating how CAB contributions were integrated into the intervention. This "give-back" approach, fostered a strong sense of ownership and commitment among members, as they could see the tangible impact of their participation. To sustain engagement between meetings, members were encouraged to stay connected through online platforms, such as LinkedIn, or private group cats, creating additional opportunities for collaboration and relationship building.

The digital-first format also allowed us to celebrate member contributions in meaningful ways. For example, certificates of membership were shared electronically, and members received regular updates on how their feedback had influenced the intervention's development. These findings, demonstrating the effectiveness of real-time updates, adds to existing evidence to support the approach [19,26]. These practices emphasized the value of their input and strengthened their motivation to continue participating with the project.

Limitations

Our approach had several limitations worth noting. First, we began recruiting CAB members during the height of the COVID-19 pandemic, a period marked by widespread skepticism of online information and heightened vaccine-related distress. These factors may have limited the number of people interested in the project, particularly for those hesitant about engaging in a vaccine-related initiative. Furthermore, as the pandemic continued, fatigue and continued experienced by parents and the broader public may have contributed to a slight drop-off in CAB membership between years 1 and 2.

Second, while social media platforms provided effective tools for recruitment, reliance on Twitter alone may have excluded individuals or limited diversity. Expanding recruitment efforts to include additional platforms could diversify participant pools and enhance inclusivity in future studies. Furthermore, increasing the diversity of the CAB, whether in terms of age, race, gender, disability status or viewpoints, would further enrich discussions, better represent the community, and ultimately better inform the research. For instance, incorporating parents who are not digitally native or who face barriers to social media engagement could offer valuable perspectives often overlooked in online-only recruitment strategies.

We also recognized the underrepresentation of fathers on the CAB (n=1). Although Twitter's targeting tools allowed for gender-based recruitment, leveraging specific social media groups or communities tailored to fathers, such as online parenting forms or father-focused groups, may help address this imbalance. Leveraging connections to recruit more fathers could be a valuable strategy for future studies that seek to engage parents, as increased representation and diversity only broadens the range of perspectives and ensures the CAB truly reflects the community it serves. Importantly, while underrepresentation of fathers may be viewed as a limitation, our formative research found that fathers were more likely to defer to the mother regarding child health decisions, perhaps explaining their underrepresentation in this study.

While the online format offers flexibility and accessibility, some CAB members noted challenges related to ensuring that all voices, particularly those with differing opinions, were fully incorporated into the project. Online group discussions may sometimes discourage dissenting opinions due to the absence of in-person cues or the fast-paced nature of digital conversations. Using structured feedback mechanisms, such as anonymous surveys or one-on-one digital interviews, may help future online CABs better capture a broad range of perspectives and ensure that all voices are equally heard. The online model reflected how the study was developed and was essential for

XSL•FO
RenderX

navigating the constraints of the pandemic; however, it inherently posed limitations for fostering the depth of connection and collaboration that may arise in face-to-face interactions. Future studies could consider incorporating hybrid models, blending online meetings with occasional in-person events, to balance flexibility of online formats with the relational benefits of in-person collaboration.

Despite these limitations, we believe that an online CAB offers unique opportunities to engage diverse populations, facilitate inclusive participation, and provide meaningful contributions to health programs delivered online. By iterating on these approaches and addressing the highlighted limitations, future research can further refine online CAB methodologies and amplify their impact in public health initiatives.

Conclusion

This study demonstrates the feasibility and effectiveness of using online and social media–based CABs in health research. Textbox 1 shows a list of guiding principles for developing online CABs based on findings from this study and supported by previous research [26-30]. Future research should explore additional methods for integrating digital tools to further enhance engagement and maintain long-term participation in online CABs.

Online and digital health interventions have become influential tools for public health, providing innovative means to address health disparities and influence health behaviors on a broad scale. As online interventions continue to expand, so must our efforts to strategically and effectively engage online advisors within the community. Social media and online platforms serve as powerful avenues for recruiting diverse and representative CAB members, fostering active engagement in research processes, and generating quality feedback that can significantly enhance research programs. These platforms not only enable the inclusion of voices from geographically dispersed and underserved populations, but also create opportunities for dynamic, real-time collaboration that aligns with the evolving digital landscape. By leveraging these approaches, researchers can maximize the potential of online CABs as flexible, scalable, and inclusive mechanisms for participatory research. This model holds promise for driving meaningful health outcomes, reducing inequities, and fostering a deeper connection between research initiatives and the communities they aim to serve.

Textbox 1. Guiding principles for developing online community advisory boards.

Forming the online community advisory board:

- Clearly define the purpose, roles, functions, and responsibilities expected of the online community advisory board members, ensuring responsibilities align with the format and digital engagement.
- Leverage social media platforms and tools, such as targeted ads and other adjustable parameters to recruit diverse members from specific demographics.
- Prioritize diversity in geography, online behaviors, and parental perspectives by using digital methods to reach underrepresented groups and those with varied social media experience.

Engaging the online community advisory board:

- Use digital tools like interactive workbooks and online surveys to collect feedback systematically, ensuring ease of access and usability in an online environment.
- Build a sense of community and connection by facilitating collaborative discussions during video calls and leverage shared online spaces to connect members meaningfully, as whole people, despite geographic distance.
- Ensure flexibility in scheduling and participation, leveraging asynchronous tools such as recorded meetings, digital materials, and follow-up communication to accommodate members' busy online and offline schedules.

Maintaining the online community advisory board:

- Enhance inclusivity by using social media and other online platforms to recruit members from diverse online communities, ensuring broad representation across digital spaces and perspectives.
- Foster a sense of belonging by incorporating icebreakers and encouraging casual online interactions, such as LinkedIn connections or private text or chat groups to sustain engagement between meetings and outside of the project.
- Continuously demonstrate the impact of members' input by providing real-time updates through online platforms and integrating feedback into the digital intervention to maintain motivation and accountability.

Acknowledgments

This study was supported by the National Cancer Institute of the National Institutes of Health (award R01CA229324). The content was solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

XSL•FO
RenderX

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Screening questions for potential community advisory board members, May 2020.
[DOCX File, 15 KB - formative_v9i1e65986_app1.docx]

Multimedia Appendix 2
Community advisory board interview guide.
[DOCX File, 16 KB - formative_v9i1e65986_app2.docx]

Multimedia Appendix 3
Sample workbook for community advisory board (CAB) members.
[PDF File, 2357 KB - formative_v9i1e65986_app3.pdf]

Multimedia Appendix 4
Parent advisory board year 1 midterm evaluation.
[DOCX File, 16 KB - formative_v9i1e65986_app4.docx]

Multimedia Appendix 5
Human papillomavirus (HPV) vaccine content for Twitter co-created by community advisory board (CAB) members and study team.
[PDF File, 20554 KB - formative_v9i1e65986_app5.pdf]

References

1. Barkin S, Schlundt D, Smith P. Community-engaged research perspectives: then and now. Acad Pediatr 2013;13(2):93-97. [doi: 10.1016/j.acap.2012.12.006] [Medline: 23498079]
2. Newman SD, Andrews JO, Magwood GS, Jenkins C, Cox MJ, Williamson DC. Community advisory boards in community-based participatory research: a synthesis of best processes. Prev Chronic Dis 2011 May;8(3):A70. [Medline: 21477510]
3. Concannon TW, Fuster M, Saunders T, et al. A systematic review of stakeholder engagement in comparative effectiveness and patient-centered outcomes research. J Gen Intern Med 2014 Dec;29(12):1692-1701. [doi: 10.1007/s11606-014-2878-x] [Medline: 24893581]
4. Connors E, Selove R, Canedo J, et al. Improving community advisory board engagement in precision medicine research to reduce health disparities. J Health Dispar Res Pract 2019;12(6):80-94. [Medline: 32832256]
5. Alexander R, Estabrooks P, Brock DJP, Hill JL, Whitt-Glover MC, Zoellner J. Capacity development and evaluation of a parent advisory team engaged in childhood obesity research. Health Promot Pract 2021 Jan;22(1):102-111. [doi: 10.1177/1524839919862251] [Medline: 31409144]
6. Hartling L, Elliott SA, Buckreus K, Leung J, Scott SD. Development and evaluation of a parent advisory group to inform a research program for knowledge translation in child health. Res Involv Engagem 2021 Jun 7;7(1):38. [doi: 10.1186/s40900-021-00280-3] [Medline: 34099069]
7. Curran JA, Bishop A, Chorney J, MacEachern L, Mackay R. Partnering with parents to advance child health research. Healthc Manage Forum 2018 Mar;31(2):45-50. [doi: 10.1177/0840470417744568] [Medline: 29400092]
8. Halladay JR, Donahue KE, Sleath B, et al. Community advisory boards guiding engaged research efforts within a clinical translational sciences award: key contextual factors explored. Prog Community Health Partnersh 2017;11(4):367-377. [doi: 10.1353/cpr.2017.0044] [Medline: 29332850]
9. McAllister CL, Green BL, Terry MA, Herman V, Mulvey L. Parents, practitioners, and researchers: community-based participatory research with early head start. Am J Public Health 2003 Oct;93(10):1672-1679. [doi: 10.2105/ajph.93.10.1672] [Medline: 14534219]
10. Brock DJP, Estabrooks PA, Hill JL, et al. Building and sustaining community capacity to address childhood obesity: a 3-year mixed-methods case study of a community-academic advisory board. Fam Community Health 2019;42(1):62-79. [doi: 10.1097/FCH.0000000000000212] [Medline: 30431470]

XSL•FO
RenderX

11. Schoeppe J, Cheadle A, Melton M, et al. The immunity community: a community engagement strategy for reducing vaccine hesitancy. Health Promot Pract 2017 Sep;18(5):654-661. [doi: 10.1177/1524839917697303] [Medline: 28398837]
12. Engster SA, Fascetti C, Mykita A, Pompa K, Reis EC. Bringing parent voices into a pediatric research network through a virtual parent panel. J Am Board Fam Med 2020;33(5):665-674. [doi: 10.3122/jabfm.2020.05.200050] [Medline: 32989061]
13. Liverpool S, Mota CP, Sales CMD, et al. Engaging children and young people in digital mental health iterventions: systematic review of modes of delivery, facilitators, and barriers. J Med Internet Res 2020 Jun 23;22(6):e16317. [doi: 10.2196/16317] [Medline: 32442160]
14. Oh C, Carducci B, Vaivada T, Bhutta ZA. Digital interventions for universal health promotion in children and adolescents: a systematic review. Pediatrics 2022 May 1;149(Suppl 5):e2021053852H. [doi: 10.1542/peds.2021-053852H] [Medline: 35503335]
15. Brigden A, Anderson E, Linney C, et al. Digital behavior change interventions for younger children with chronic health conditions: systematic review. J Med Internet Res 2020 Jul 31;22(7):e16924. [doi: 10.2196/16924] [Medline: 32735227]
16. Shea L, Pesa J, Geonnotti G, Powell V, Kahn C, Peters W. Improving diversity in study participation: patient perspectives on barriers, racial differences and the role of communities. Health Expect 2022 Aug;25(4):1979-1987. [doi: 10.1111/hex.13554] [Medline: 35765232]
17. Iogansen X, Malik JK, Lee Y, Circella G. Change in work arrangement during the COVID-19 pandemic: a large shift to remote and hybrid work. Transp Res Interdiscip Persp 2024 May;25:100969. [doi: 10.1016/j.trip.2023.100969]
18. Cargo M, Mercer SL. The value and challenges of participatory research: strengthening its practice. Annu Rev Public Health 2008;29:325-350. [doi: 10.1146/annurev.publhealth.29.091307.083824] [Medline: 18173388]
19. Cramer ME, Lazoritz S, Shaffer K, Palm D, Ford AL. Community advisory board members' perspectives regarding opportunities and challenges of research collaboration. West J Nurs Res 2018 Jul;40(7):1032-1048. [doi: 10.1177/0193945917697229] [Medline: 28367677]
20. Guideline for human papillomavirus (HPV) vaccine use. American Cancer Society. 2024. URL: https://www.cancer.org/health-care-professionals/american-cancer-society-prevention-early-detection-guidelines/hpv-guidelines.html [accessed 2024-08-30]
21. Lockhart S, Dempsey AF, Pyrzanowski J, O'Leary ST, Barnard JG. Provider and parent perspectives on enhanced communication tools for human papillomavirus vaccine-hesitant parents. Acad Pediatr 2018;18(7):776-782. [doi: 10.1016/j.acap.2018.05.012] [Medline: 30177028]
22. Beavis AL, Meek K, Moran MB, Fleszar L, Adler S, Rositch AF. Exploring HPV vaccine hesitant parents' perspectives on decision-making and motivators for vaccination. Vaccine X 2022 Dec;12:100231. [doi: 10.1016/j.jvacx.2022.100231] [Medline: 36337836]
23. Petagna CN, Perez S, Hsu E, et al. Facilitators and barriers of HPV vaccination: a qualitative study in rural Georgia. Res Sq 2024 Mar 8:rs.3.rs-3979079. [doi: 10.21203/rs.3.rs-3979079/v1] [Medline: 38496559]
24. What are promoted ads? Twitter. URL: https://business.twitter.com/en/help/overview/what-are-promoted-ads.html [accessed 2023-05-31]
25. Massey PM, Chiang SC, Rose M, et al. Development of personas to communicate narrative-based information about the HPV vaccine on twitter. Front Digit Health 2021;3:682639. [doi: 10.3389/fdgth.2021.682639] [Medline: 34713151]
26. Matthews AK, Anderson EE, Willis M, Castillo A, Choure W. A community engagement advisory board as a strategy to improve research engagement and build institutional capacity for community-engaged research. J Clin Transl Sci 2018 Apr;2(2):66-72. [doi: 10.1017/cts.2018.14] [Medline: 31660220]
27. Gelinas L, Pierce R, Winkler S, Cohen IG, Lynch HF, Bierer BE. Using social media as a research recruitment tool: ethical issues and recommendations. Am J Bioeth 2017 Mar;17(3):3-14. [doi: 10.1080/15265161.2016.1276644] [Medline: 28207365]
28. Bosak J, Drainoni ML, Christopher M, et al. Community advisory board members' perspectives on their contributions to a large multistate cluster RCT: a mixed methods study. J Clin Transl Sci 2024;8(1):e1. [doi: 10.1017/cts.2023.673] [Medline: 38384918]
29. Greenwood DA, Litchman ML, Ng AH, et al. Development of the intercultural diabetes online community research council: codesign and social media processes. J Diabetes Sci Technol 2019 Mar;13(2):176-186. [doi: 10.1177/1932296818818455] [Medline: 30614252]
30. Greenwood MJ, Heninger JR. Structured communication for patient safety in emergency medical services: a legal case report. Prehosp Emerg Care 2010;14(3):345-348. [doi: 10.3109/10903121003760788] [Medline: 20438344]

Abbreviations

CAB: community advisory board
HPV: human papillomavirus

XSL•FO
RenderX

Edited by A Mavragani; submitted 31.08.24; peer-reviewed by D Carvalho, N Godwin; revised version received 10.03.25; accepted 11.03.25; published 16.04.25.

Please cite as:
Murray RM, Chiang SC, Klassen AC, Manganello JA, Leader AE, Lo WJ, Massey PM
Developing an Online Community Advisory Board (CAB) of Parents From Social Media to Co-Design an Human Papillomavirus Vaccine Intervention: Participatory Research Study
JMIR Form Res 2025;9:e65986
URL: https://formative.jmir.org/2025/1/e65986
doi:10.2196/65986

© Regan M Murray, Shawn C Chiang, Ann C Klassen, Jennifer A Manganello, Amy E Leader, Wen-Juo Lo, Philip M Massey. Originally published in JMIR Formative Research (https://formative.jmir.org), 16.4.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Machine Learning Approach to Identifying Empathy Using the Vocals of Mental Health Helpline Counselors: Algorithm Development and Validation

Ruvini Sanjeewa[1], BSc; Ravi Iyer[1], PhD; Pragalathan Apputhurai[1], PhD; Nilmini Wickramasinghe[2], PhD; Denny Meyer[1], PhD

[1]School of Health Sciences, Swinburne University of Technology, Hawthorn, PO Box 218, John Street, Melbourne, Australia

[2]School of Computing, Engineering & Mathematical Sciences, La Trobe University, Melbourne, Australia

Corresponding Author:
Ruvini Sanjeewa, BSc
School of Health Sciences, Swinburne University of Technology, Hawthorn, PO Box 218, John Street, Melbourne, Australia

Abstract

Background: This research study aimed to detect the vocal features immersed in empathic counselor speech using samples of calls to a mental health helpline service.

Objective: This study aimed to produce an algorithm for the identification of empathy from these features, which could act as a training guide for counselors and conversational agents who need to transmit empathy in their vocals.

Methods: Two annotators with a psychology background and English heritage provided empathy ratings for 57 calls involving female counselors, as well as multiple short call segments within each of these calls. These ratings were found to be well-correlated between the 2 raters in a sample of 6 common calls. Using vocal feature extraction from call segments and statistical variable selection methods, such as L1 penalized LASSO (Least Absolute Shrinkage and Selection Operator) and forward selection, a total of 14 significant vocal features were associated with empathic speech. Generalized additive mixed models (GAMM), binary logistics regression with splines, and random forest models were used to obtain an algorithm that differentiated between high- and low-empathy call segments.

Results: The binary logistics regression model reported higher predictive accuracies of empathy (area under the curve [AUC]=0.617, 95% CI 0.613 - 0.622) compared to the GAMM (AUC=0.605, 95% CI 0.601 - 0.609) and the random forest model (AUC=0.600, 95% CI 0.595 - 0.604). This difference was statistically significant, as evidenced by the nonoverlapping 95% CIs obtained for AUC. The DeLong test further validated these results, showing a significant difference in the binary logistic model compared to the random forest (D=6.443, *df*=186283, *P*<.001) and GAMM (Z=5.846, *P*<.001). These findings confirm that the binary logistic regression model outperforms the other 2 models concerning predictive accuracy for empathy classification.

Conclusions: This study suggests that the identification of empathy from vocal features alone is challenging, and further research involving multimodal models (eg, models incorporating facial expression, words used, and vocal features) are encouraged for detecting empathy in the future. This study has several limitations, including a relatively small sample of calls and only 2 empathy raters. Future research should focus on accommodating multiple raters with varied backgrounds to explore these effects on perceptions of empathy. Additionally, considering counselor vocals from larger, more heterogeneous populations, including mixed-gender samples, will allow an exploration of the factors influencing the level of empathy projected in counselor voices more generally.

(JMIR Form Res 2025;9:e67835) doi:10.2196/67835

KEYWORDS

vocal features; voice characteristics; empathy; mental health care; crisis helpline service

Introduction

Empathy is defined as experiencing the emotions (emotional empathy) and cognitions (cognitive empathy) of others and responding to them appropriately [1]. Empathy is especially important for patient care, where the lived experience of the patient must be understood by responding health care professionals, while also conveying this understanding in conjunction with a desire to help the patient [2]. The effectiveness of physician empathy has been shown to improve patient satisfaction and commitment to recovery while reducing anxiety and distress levels, leading to better clinical results [3]. Furthermore, empathic behavior by mental health (MH) care providers reduces their own risk of burnout [4].

XSL•FO
RenderX

Telephone helpline services offer an effective means of supporting those who need immediate MH care [5]. The demand for such services has increased dramatically since the outbreak of the COVID-19 pandemic [6], increasing the expectations of counseling staff to provide support for people with complex MH concerns [7]. As a basic counseling skill, empathy is key to successful engagement with patients in the context of complex psychosocial needs.

Besides emotional and cognitive empathy in understanding the status of a patient, contextual awareness is equally important for therapeutic engagement [8]. This means that empathic responses need to be contextually appropriate by considering environmental cues, culture, demographic factors, and the specific circumstances of the patient to understand the broader context of their MH status [9,10]. This allows counselors to tailor responses, based on context, to engage in effective communication with distressed patients, thereby delivering better outcomes [3].

Verbal cues and tone of voice are crucial when communicating empathy [11]. For example, reduced speech rate and lower pitch are perceived as more empathic by patients when receiving bad news from health care providers in an oncology setting [12] and while actively listening to telephone callers, nurses have been found to express empathy through their choice of words, voice and intonation, projection of compassion and warmth, as well as “tuning in” to the caller’s story and identifying with the caller’s emotions [13].

Unfortunately, the global demand for MH support is not being met by the existing workforce [14]. This service gap is leading to growing interest in alternative digital technological solutions. Technological innovations, such as the design of conversational agents, have demonstrated potential in facilitating effective and immediate patient care. However, to optimize upon end user acceptance, conversational agents need to display empathy [15,16]. However, we have yet to identify the precise vocal features most associated with an empathic human response, and it is also not known if it is possible to categorize empathy levels using such vocal features. Thus, the aim of this paper is to (1) identify the vocal features significantly associated with empathy in a large collection of telephone helpline counseling call recordings; and to (2) evaluate the accuracy of a machine learning algorithm to correctly designate short segments of each recording to categories of low and high empathy.

This study has been reported in accordance with the TRIPOD+AI (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis+Artificial Intelligence) checklist as shown in Checklist 1.

Methods

Data Collection

Recordings of telephone helpline calls (n=57) were obtained from On The Line, Australia, a suicide helpline counseling service. Participants were counselors for a Suicide Call-Back Service (a national helpline service coordinated by On The Line, Australia). Calls were randomly sampled from July 1, 2019, to June 30, 2021, stratified by organizationally determined suicide risk level (high or low). The level of suicide risk of each caller had been previously assessed by counselors using the Columbia Suicide Severity Rating Scale (C-SSRS) to differentiate between calls featuring high suicide risk (with C-SSRS ratings of 6‐7) and calls with low risk of suicide (with C-SSRS ratings of 1‐2; please refer to Iyer et al [17] for further details [18]). Only the counselors’ voice recordings were used in this study. No information was provided that could be used to identify the callers or counselors for any of these calls.

Ethical Considerations

This study was approved by the Swinburne Human Research Ethics Committee (Ref:20226835‐11907). The application was given a waiver of consent from the ethics committee for the use of nonidentifiable secondary research data. However, the research team also signed a confidentiality agreement that restricted the discussion of the contents of the recordings only within the research team. The secondary data and annotation results were saved on OneDrive for Business (Microsoft Corp) and were only accessible to the research team. No compensation was paid to participants in this study because they could not be identified by the research team.

Annotation of Call Segments and Overall Call Empathy for Counselors

Annotations of counselor empathy were conducted by 2 independent researchers (Inge Gnatt and Sarah Dunning) using RStudio (version 2024.04.2, build 764; R Foundation) [19]. The call annotators Inge Gnatt and Sarah Dunning were recruited via research team networks. Both had extensive experience working as counselors for a MH helpline service. Inge Gnatt had experience working as an annotator on a similar project. Segments of the counselor voices were selected from each call using Audacity (version 3.5.1, CMake Release Build; Muse Group & contributors) [20], ensuring that overlaps between the caller and responding counselor voices were minimized. Empathy displayed by the counselor within each call segment was rated using the Carkhuff and Truax Empathy (CTE) scale [21]. A weekly project team meeting, attended by a clinical psychologist (Maja Nedeljkovic), was used to reconcile any disparities in ratings. A Qualtrics web-based questionnaire was used to also collect data on the overall level of empathy displayed by the responding counselor during each call and to evaluate caller distress at the commencement and conclusion of each call. Figure 1 shows the flow of the voice analysis process during this study.

XSL•FO
RenderX

Figure 1. Overview of the voice analysis process for classifying empathy in counselor voices. AELS: Active-Empathic Listening Scale; AUC: area under the curve; CTE: Carkhuff and Truax Empathy; LASSO: Least Absolute Shrinkage and Selection Operator; LOOCV: Leave One Out Cross Validation; GAMM: generalized additive mixed model; N/A: not available; OTLA: On the Line Australia; PCM-f321e: pulse code modulation float; PEIS: Perceived Emotional Intelligence Scale; RS7: Rating Scale 7 (7-item).

XSL•FO
RenderX

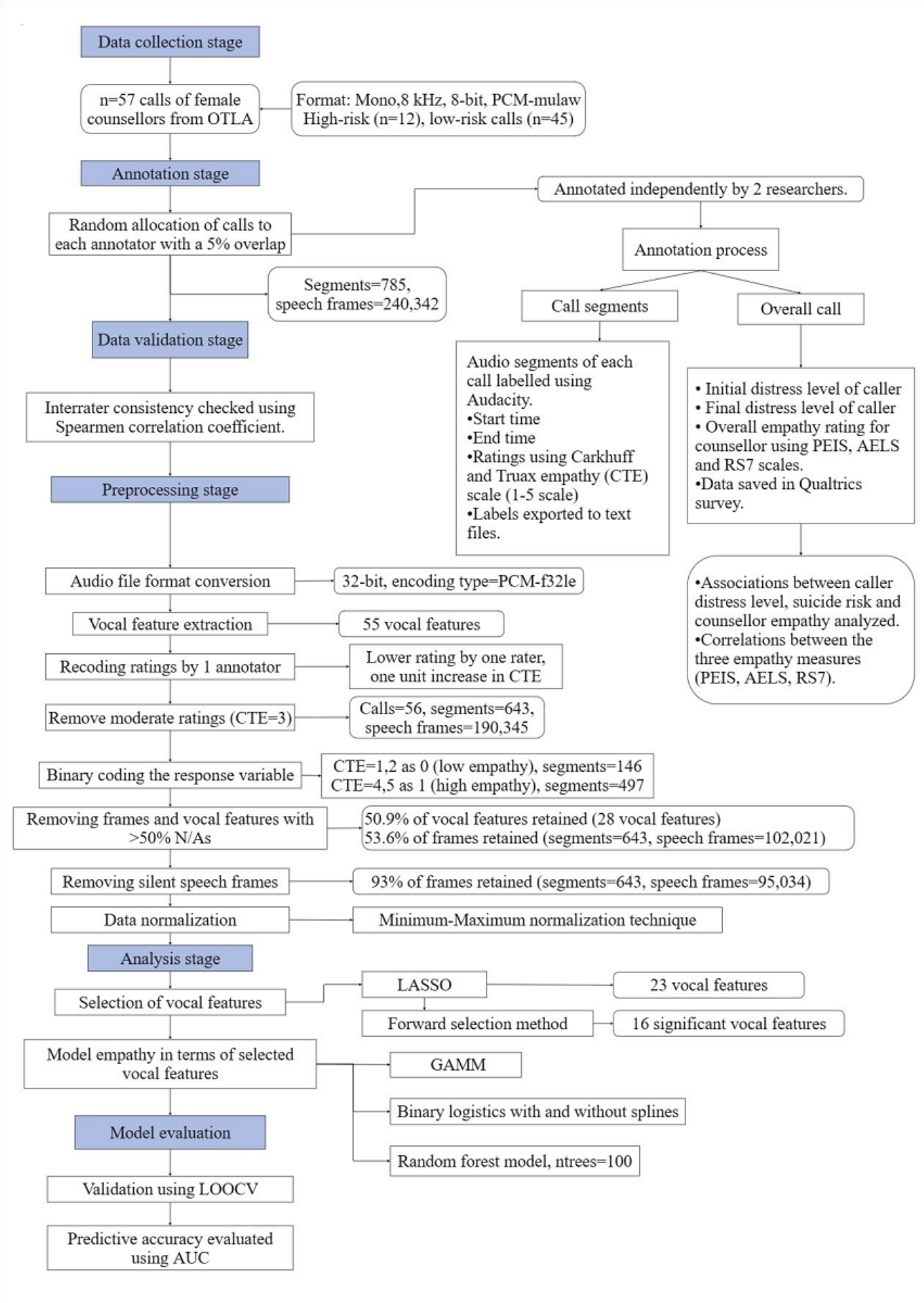

XSL•FO
RenderX

Measures of Counselor Empathy and Caller Distress

The CTE scale was used to rate the audio segments selected from each call on a 5-point Likert-style scale (1="low empathy" to 5="high empathy") [21]. Additionally, 3 measures were used to assess overall counselor empathy for each call. These 3 scales were modified to suit counselor-caller conversations through an iterative process in which members of the research team provided independent feedback to achieve the final questionnaire. Examples were developed by the annotators for each item included in these scales, ensuring clarity and consistency of ratings. The details for these scales are provided in Multimedia Appendix 1. The measurement scales included:

1. The Perceived Emotional Intelligence (PEI) Scale [22] identified variations in PEI in the counselor's vocals. The PEI is a 20-item scale with each item scored with 1="never or almost never true" and 7="almost or almost always true."
2. The Active-Empathic Listening (AEL) Scale [23] was modified appropriately to produce 10 items measuring empathic listening using a 7-point Likert-style scale with 1="never or almost never true" and 7="almost or almost always true."
3. Rating Scale 7 (RS7): A single item 7-point Likert Scale [24] was used to rate overall empathy with 1="low empathy" and 7="high empathy."

Finally, at the start and end of each call the annotators assessed the level of caller distress using the distress thermometer [25], a visual analogue 11-point scale (0="no distress" to 10="extreme distress").

Data Validation Through Interrater Consistency Check

Six (10%) calls were chosen at random to measure interrater reliability. The empathy ratings of the more experienced rater (Inge Gnatt) were used as the reference, against which the SD ratings were compared. Spearman correlation [26] was calculated for each of the 3 scales used to rate overall counselor empathy for each call. The Mann-Whitney *U* test was then used to check for the significance of differences between the ratings provided by the annotators.

Relationships Between Perceptions of Empathy and Call Context

The associations between perceived counselor empathy and call context were explored using a combination of empathy ratings, caller distress at the beginning and end of the call, and caller suicide risk. Caller distress and suicide risk were correlated with perceived counselor empathy to evaluate the relationship between level of empathy and caller disposition.

Preprocessing Stage: Audio File Format Conversion and Vocal Features Extraction

The input call recordings were obtained as 8 kHz sample rate, 8-bit depth .wav files. The encoding type of the files were transformed to PCM float format with 32-bit depth to ensure compatibility with RStudio for analysis. Vocal features (n=55) were extracted per 30-millisecond speech frames (50% overlap; Blackman windows) within each annotated segment using RStudio (version 2.7.0, *Soundgen* package [27]).

Removing Moderate Ratings and Binary Coding Empathy Level

Overview

The vocal segments that scored a rating of 3 out of 5 on the CTE scale were removed (n=142, 18%) from further analysis because of their neutral empathic character. A binary response variable was then created for each of the 643 remaining segments (190,345 speech frames of 30 ms) with an empathy rating of 4 - 5 coded as high empathy (n=146 segments) and a rating of 1 - 2 coded as low empathy (n=497 segments).

Removing Missing Values in Speech Frames and Vocal Features

Vocal features and 30-millisecond speech frames with more than 50% missing values were removed to maintain the quality of data and improve the overall accuracy of the results. The resulting data retained 50.9% (n=28) of the original vocal features and 53.6% (n=102,021) of the original speech frames.

Removing Silent Speech Frames and Normalization

The silent speech frames were also removed from each of the 643 segments leaving 95,034 speech frames in the final analysis sample. Finally, the minimum and maximum normalizing technique [28] was applied to reduce the influence of background noise.

Analysis Stage

Selection of Vocal Features

Variable selection was performed to identify vocal features that were strongly associated with empathy. L1 penalized LASSO (Least Absolute Shrinkage and Selection Operator) regression was used to select the most relevant variables by shrinking the coefficients of the least relevant variables to 0 [29]. Tenfold cross-validation was used to optimize the tuning parameter, lambda. Further refinement per the selected vocal features was then conducted using a forward stepwise regression model.

Models for Identifying Empathy Level With Selected Vocal Features

Three methods were used to classify low and high empathy segments based on this final set of selected vocal features. A generalized additive mixed model (GAMM) included vocal features as fixed effects, with each call treated as a random effect. Spline functions for the selected vocal features were used to account for nonlinearity [30]. The GAM function [31] of package *mgcv* [32] in RStudio was used for the analysis.

Random forest classification also allowed for nonlinear relationships using step functions while more efficiently processing large datasets [33]. This has been a prominent classification model used in studies involving vocal analysis [34,35]. The binary logistic regression model was the third model considered, again accounting for nonlinearity using splines, and was used in our study as the baseline model [36].

Model Evaluation

Probabilistic predictions for high versus low empathy levels were obtained for each segment using Leave One Caller Out

XSL•FO
RenderX

Cross Validation. Based on these probabilities, receiver operating characteristic curves were created, and areas under the curves (AUCs) [37,38] were used to compare the reliability of these models. The Youden index [39] was used to decide the optimal cut point for classifying segments based on their estimated high versus low empathy probabilities.

Results

Overview

The following results were obtained from the 57 calls for female counselors. This sample of calls included 12 (12/57, 21%) calls at high risk of suicide and 45 (45/57,79%) calls at low risk of suicide.

Annotation of Call Segments

Using the CTE scale, 146 (18.6%) segments showed low empathy, 142 (18.1%) showed medium empathy, and 497 (63.3%) segments demonstrated high empathy.

Overall Call Empathy Ratings

Multimedia Appendix 2 provides descriptive statistics for the overall empathy ratings for the 57 calls using the 3 scales. Excellent reliability is observed from both the raters, Inge Gnatt and Sarah Dunning, for the PEI and AEL scales, with Cronbach α values above 0.9. The descriptive statistics and the Spearman correlation statistics between the annotators are shown in Multimedia Appendix 3. A strong agreement between raters was observed with the PEI measure and the RS7 empathy measure approaching statistical significance.

Relationships Between Perceptions of Empathy and Call Context

The relationship of counselor empathy ratings with caller's distress at the start and end of the call and suicide risk are shown in Table 1. While the correlations between the initial distress and the 3 empathy measures were not significant, a moderate, statistically significant negative correlation between the final distress of the caller and the empathy of the counselor was observed for both the PEI and RS7 ($P<.01$ and $P<.001$, respectively). The suicide risk of callers as measured using the C-SSRS, had a statistically significant but weak positive correlation with counselor empathy across the PEI and AEL measures. Strong statistically significant correlations among the 3 empathy measures were found, validating the empathy measurement process.

Table . Spearman correlation coefficients for caller distress and suicide risk with counselor empathy ratings[a].

	Caller context			Ratings for counselor empathy		
Empathy rating	Initial distress	Final distress	Suicide risk	PEIS[b]	AELS[c]	RS7[d]
PEIS	–0.011	–0.414**	0.310*	1		
AELS	0.102	–0.233	0.308*	.822***	1	
RS7	–0.024	–0.443***	0.055	.847***	.851***	1

[a]***$P<.001$ (2-tailed), **$P<.01$ (2-tailed), and *$P<.05$ (2-tailed).

[b]PEIS: Perceived Emotional Intelligence Scale.

[c]AELS: Active-Empathic Listening Scale.

[d]RS7: Rating Scale 7 (7-item).

Analysis Stage

The selection of vocal features using LASSO to predict high versus low empathy used a cross-validation of the training dataset to reveal an optimum Log (Lambda) parameter=–8.542. The relationship between this parameter and the binomial deviance is shown in Multimedia Appendix 4. Using this lambda value, 23 vocal features were retained. These 23 vocal features were then passed on to the forward selection binary logistic regression model, which identified 16 significant vocal features as shown in Table 2.

XSL•FO
RenderX

Table . Results of the forward binary logistic regression model for vocal feature selection for identifying high versus low empathy in counselor voices.

Vocal features	Coefficient	*Z* value	*P* value
Depth of amplitude	–1.468	–15.994	<.001
Frequency of amplitude (Hz)	0.068	2.524	.01
Frequency of amplitude (Hz) via MS[a]	–0.462	–12.350	<.001
Purity of amplitude via MS	–0.834	–9.596	<.001
Amplitude (dB)	–3.391	–51.688	<.001
Dominant frequency (Hz)	0.985	3.136	.002
Entropy	3.47	20.843	<.001
Shannon entropy	–6.016	–25.27	<.001
Epoch	0.289	5.04	<.001
First formant frequency (Hz)	1.244	5.08	<.001
First formant width (Hz)	0.194	4.181	<.001
Second formant frequency (Hz)	–0.191	–2.808	.005
Second formant width (Hz)	–0.063	–1.802	.07
Third formant frequency (Hz)	–0.099	–1.521	.13
Third formant width (Hz)	0.004	0.118	.91
Spectral flux	0.11	1.435	.15
HNR[b] (dB)	–0.37	2.869	.004
Spectral novelty	0.035	0.715	.47
Peak frequency (Hz)	–0.293	–1.827	.06
25th percentile frequency (Hz)	–1.311	–4.720	<.001
50th percentile frequency (Hz)	–0.262	–1.586	.11
Spectral centroid (Hz)	5.782	15.717	<.001
Spectral slope (Hz)	–3.944	–17.881	<.001

[a]MS: modulation spectrum.

[b]HNR: harmonics-to-noise ratio.

Feature Extraction and Classification

The results of the GAMM are shown in Multimedia Appendix 5. Of the 16 selected vocal features, the GAMM used only 14. Based on the effective *df* values, 2 of the vocal features (Shannon entropy and the 25th percentile frequency) show a linear relationship in the GAMM. Figure 2 illustrates the nonlinear nature of all other relationships. An AUC value of 0.605 was obtained for the GAMM [40].

XSL•FO
RenderX

Figure 2. The smoothed relationship between the selected vocal features and the standardized predicted level of empathy displayed in counselor speech with 95% CIs. MS: modulation spectrum.

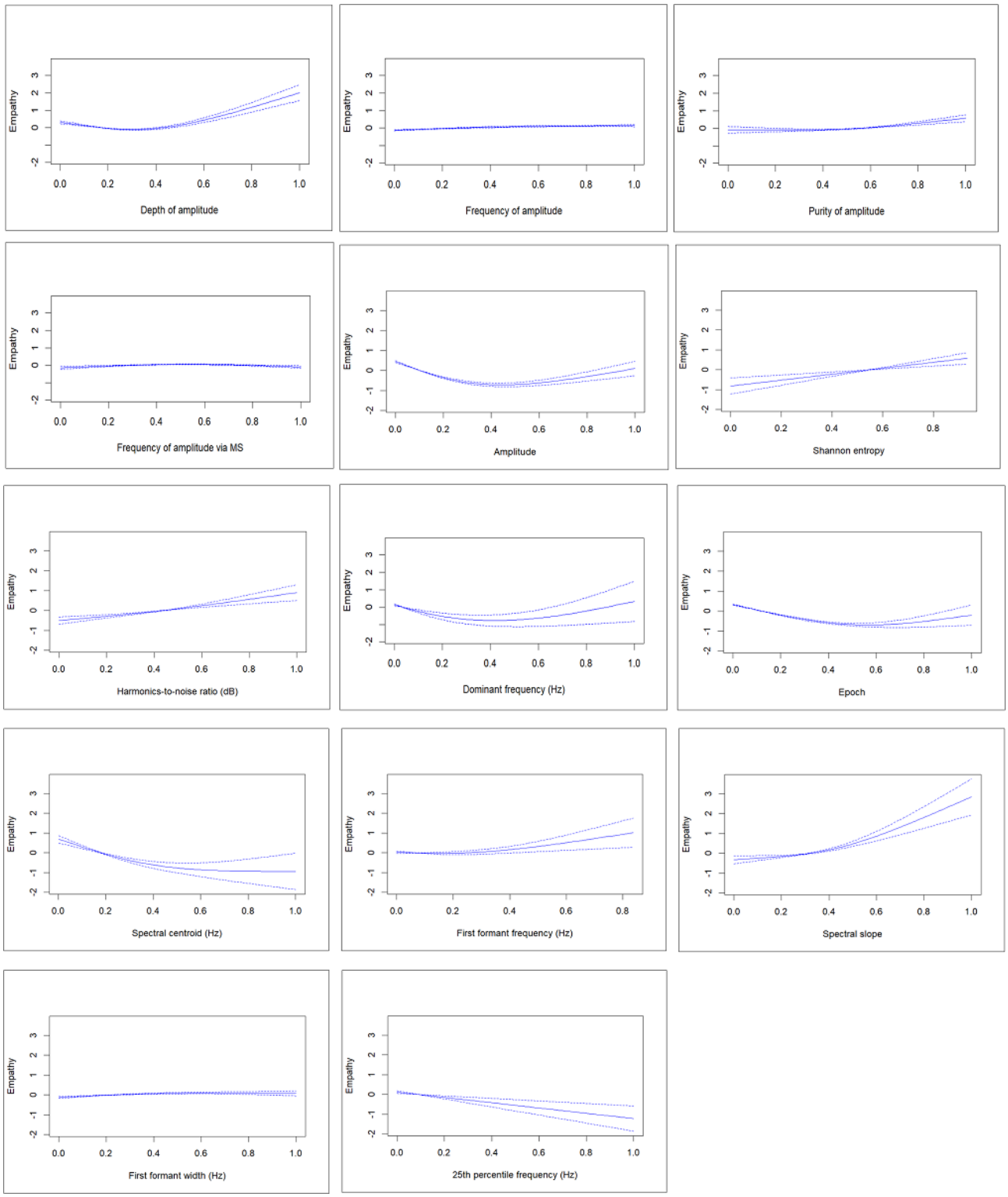

The binary logistic classification model showed higher AUC values and nonoverlapping CIs at 95%. The DeLong test confirmed that the binary logistic model outperformed both the random forest (D=6.443, *df*=186283, *P*<.001) and GAMM (Z=5.846, *P*<.001) models [41]. However, this model achieved a lower classification accuracy than the other 2 methods, as shown in Table 3, suggesting that the probability cut point used for classification purposes was not ideal.

XSL•FO
RenderX

Table . Performance comparison of the 3 classification models for identifying high versus low counselor empathy.

Classification model	Accuracy (%)	AUC[a] value	95% CIs
GAMM[b]	75	0.605	0.601 - 0.609
Random forest	74	0.6	0.595 - 0.604
Binary logistic regression	69	0.617	0.613 - 0.622

[a]AUC: area under the curve.

[b]GAMM: generalized additive mixed model.

As illustrated in Figure 2, higher empathy was associated with higher values for the first formant frequency, dominant frequency, Shannon entropy, spectral slope, and harmonic-to-noise ratio. In contrast, higher empathy was associated with lower values for the 25th percentile frequency and spectral centroid. Finally, lower empathy was associated with intermediate values for depth of amplitude, amplitude, dominant frequency, epoch, and first formant frequency of speech.

The GAMM was able to differentiate between low and high 30-millisecond segments of speech to a classification accuracy of 75%. This was superior to both random forest and binary logistic regression models (74% and 69%, respectively).

Epoch (39%), amplitude (22%), and depth of amplitude (7%) were the top 3 vocal features contributing to empathic speech in the GAMM. The vocal features that contribute the most toward the identification of empathic speech vary across the 3 methods used, as shown in Multimedia Appendix 5.

For purposes of validation, the voice algorithm was also evaluated on a synthesized dataset of female voices created from a text-to-speech application. This approach yielded significant differences between high and low empathic voices in the validation dataset for the GAMM, random forest, and binary logistic regression classification models.

Discussion

Principal Findings

This study was undertaken to identify a range of vocal features that can predict the level of empathy exhibited in the recordings of female counselors and to accurately classify short segments of each recording according to low or high empathy ratings. We were successful in identifying 14 unique vocal features that significantly distinguished between low and high empathy ratings in the GAMM. Furthermore, we were able to successfully classify short segments of speech using these vocal features to an accuracy level of 75% using this model. Although this study considered only 57 calls, each call featured multiple segments annotated for the level of empathy, and each of these segments was further divided into 30-millisecond speech frames that were the observational points considered during the modeling process. Reducing the number of vocal features in the model led to lower AUC values, suggesting that the original model featuring 14 vocal features was not overfitted. Furthermore, the 95% AUC CIs for the binary logistic regression did not overlap those for the other 2 methods, confirming that the binary logistic regression model produced superior results.

Empathy is an important component of human interactions and social connections that promotes general well-being [42]. It is an essential component of MH care support, required to enhance therapeutic alliance and rapport building. Empathy embodies the ability to understand and compassionately reflect the range of feelings and experiences communicated by others. Empathic communication relies upon verbal (words and vocal features) and nonverbal means such as body language and facial expression [43]. However, it is only verbal expressions of empathy through vocal features that are the subject of this study.

The detection of empathy is traditionally based upon subjective human perception captured using standardized empathy scales or questionnaires [44]. Research on detecting empathic speech through voice feature extraction has been the focus of at least 2 recent studies, both showing similar accuracies to what we have found. For example, the first study by Chen et al [45] analyzed the acoustic prosodic features of speech recorded in YouTube (Google LLC) videos used for empathy training purposes. The use of these features resulted in classification accuracies of 59% (F_1-score) when differentiating between empathy and neutral categories. A second study by Alam et al [46] explored acoustic and lexical features of empathic speech using annotations sourced from Italian call center conversations. In this study, classification using support vector machines yielded accuracies of 68.1%.

However, our study is unique in identifying a range of vocal features that identify the level of empathy using an ecologically valid dataset of counselor-caller conversations obtained from On the Line Australia, an Australian helpline service. Machine learning techniques are involved in both vocal feature selection and empathy classification. The 2 annotators chosen for empathy labeling purposes had a psychological background to strengthen the validity of this process. This led to the development of an algorithm that detects the human vocal features that are associated with empathic speech. This algorithm has the potential to enhance the training of counselors in the use of empathic speech and offers valuable insights into effective human communication in the MH care domain.

Call Context and Counselor Empathy in a Crisis Helpline Setting

The analysis revealed that there is a strong negative correlation between the final distress level of the caller and the counselor's level of empathy, suggesting that empathic communication with a caller can lower their level of distress by the end of the call. Higher levels of empathy allow the counselors to build rapport and trust with their patients, allowing for effective emotional support during a crisis. These findings align with the existing

XSL•FO
RenderX

literature about the benefits of empathic interactions. The suicide risk of the caller was positively related to counselor empathy. This is an indication that counselors effectively recognize crisis situations, exhibiting higher levels of empathy when speaking to callers with high risks of suicide. It also confirms that the level of empathy displayed by counselors is adapted to the situation of the caller. These relationships of counselor empathy with caller distress and suicide risk confirm the importance of counselor empathy in the context of crisis helpline services.

Characteristics of Empathic Vocals

This study has identified several vocal features associated with empathy. The depth of amplitude in speech reflects varying levels of loudness, emphasizing the expressiveness and dynamic nature of the human vocals apparent in empathic speech. A stable, more consistent emotional delivery during speech (purity of amplitude) also helps to convey empathy. Quieter vocals (amplitude and lower tonal frequency or dominant frequency) are also associated with higher empathy.

Higher first formant frequencies are associated with “a” vowel sound, which corresponds with high ratings of empathy. Additionally, a higher harmonic-to-noise ratio, indicating greater clarity and more pleasant-sounding vocals, is also associated with greater empathy. The spectral slope has a strong positive relationship to empathy, while the spectral centroid shows the opposite relationship. This indicates that a lower spectral centroid, with more low-frequency components, makes a speaker sound more empathic.

Based on these findings, it is evident that empathy in vocals is provided by a combination of multiple human vocal features, and variations in each of the features exert a different impact on empathy. In particular, the way that a specific threshold of loudness in the vocals decides the delivery of perceived empathy in the context of effective counseling provides compelling evidence. This further suggests that the right balance of each of these vocal features is needed, where stability, energy, and clarity play a pivotal role.

However, the study of empathy in vocals is a complex topic and has challenges. Especially the subjective nature of empathy perception is an area that requires further study. This study relied on empathy ratings provided by 2 psychology-trained female raters of English heritage. Different results may have been obtained if raters with different cultural, social, and educational backgrounds had been included [47].

Vocal Feature Extraction and Empathy Classification

Three methods were used to study the association between empathy and relevant vocal features. A GAMM, a random forest, and a logistic regression approach with splines were fitted and compared using AUC values and using the Leave One Caller Out Cross Validation method for evaluating these classifiers. The accuracy of the empathy level classifications achieved was similar (75%, 74%, and 69%, respectively) when Youden index was used to choose the probability cut point, as were their AUC values (0.605, 0.600, and 0.617, respectively). The GAMM and binary logistic regression with splines required significantly more computational time compared to the random forest, which used 100 trees. Despite the significant improvement in AUC value with the logistic regression with splines in comparison to the random forest model, the AUC results do not seem to be dissimilar. However, the AUC 95% CIs for the binary logistic regression do not overlap with those for the other 2 methods, confirming that the logistic regression model provides a better fit to the data. This was further validated by the DeLong test results showing significant differences in AUC values of the binary logistic model compared to the other 2 methods. The partial overlap between the AUC CIs of the random forest and binary GAMMs indicates a significant difference in model fit. Therefore, the binary logistic classification model outperforms the other 2 methods in its ability to distinguish high-empathic speech from low-empathic speech. However, the slightly lower classification accuracy (69%) for this model suggests that the Youden method used to determine the probability cut point used for classification purposes may not be optimal.

These relatively low AUC values, also seen in other related research projects, can perhaps be partly attributed to the difficulties encountered in providing accurate ratings of empathy. The algorithms developed for detecting empathy from vocal features were reliant on the quality of the input data provided for the empathy ratings. A larger sample of raters and a larger sample of calls might have produced more reliable data, and this is recommended for future research in this area.

However, the multimodal approaches commonly used for empathy recognition through the use of words, vocals, visual signs, and psychological signals reflect the multifaceted nature of empathy [48]. The importance of facial expressions for recognizing empathy is particularly emphasized in this literature, where factors such as observation time and the type of emotion expressed significantly influence the accuracy of identification [49]. These multimodal approaches suggest that a higher accuracy in detecting empathy can be achieved when all these factors are collectively considered, rather than vocal approaches on their own.

Limitations

The inherent differences of empathy perception among the annotators of this study were a concern in this study. An additional analysis was conducted to explore this further, incorporating a third annotator without psychological expertise and from a different cultural background (Multimedia Appendix 6). The findings from this analysis produced an even lower model performance, confirming that perceptions of empathy do vary between individuals and cultures. This analysis included both female and nonfemale counselor vocals, which may also have contributed to this poorer performance. While this finding underscores the complexity of recognizing empathy, it also highlights how cultural differences, personal experience, and psychological knowledge of individuals contribute to subjective perceptions. Therefore, future research would benefit from accommodating these differences in empathy perception within the model by including multiple annotators with varied backgrounds.

Another limitation of this study was due to the very small number of available male counselor recordings (n=13). The main analysis could therefore only be conducted on female

XSL•FO
RenderX

vocals, limiting the generalization of our findings. Ideally it would have been possible to develop separate models for empathy in male and female counselors and to test whether there were significant differences between these models. Larger samples of counselor vocals would also have been preferable, providing more diversity in the data. Unfortunately, only a very small sample of male counselor calls was obtained. A recent study that focused on the vocal characteristics of distressed adults using machine learning techniques identified a significant difference between male and female vocal behaviors. This suggested that it would not be appropriate to use a single model to describe the voices of both male and female counselors [50]. For this reason, only female counselor voices were considered in this study. Recent research on machine learning applications has explored synthesized data as an avenue to address class imbalance, and this was considered as a possible remedy to address the lower representation of male voices in this study [51]. However, using synthetic speech, such as through text-to-speech applications, faces the challenge of reflecting the diversity inherent in natural human voices [52]. So, this option was not considered for boosting the number of male counselor calls, so as to make an analysis of these data possible.

However, this option was used when the algorithm developed in this study was validated on an external synthesized dataset consisting of female recordings generated using a text-to-speech application. These data also showed successful outcomes for the algorithm in differentiating between high and low empathic voices in female voices. These results highlight synthetic voice augmentation as a promising future research direction in machine learning applications. Alternatively, a balanced representation of gender, as well as cultural background, empathy levels, and individual characteristics, are necessary considerations for the counselor recordings used in future research of this nature.

A further limitation is the source of the calls used for this study. The context of a suicide helpline service is very specific, and it may be that more algorithmic success would have been possible in a less stressed environment. Additionally, for most of the calls, the level of empathy was assessed by a single rater. As mentioned above, it would have been preferable if a greater duplication of ratings could have been used to provide the dependent variable for the models that have been used to identify the level of empathy in counselor vocals. However, this pilot study, using machine learning techniques to identify vocal features related to empathy in female counselor voices, has shown promise. Therefore, further explorations of this approach with increases in the call sample size and a more balanced gender representation, and with more annotators to allow more duplication of ratings and more variation in annotator background while also including calls from more than 1 helpline service, will benefit future research on this topic.

Implications of This Study

The importance of empathy in reducing the distress of callers confirms the need for the incorporation of empathic communication skills in training programs for counselors. Additionally, a statistically significant positive correlation was found between suicide risk and counselor empathy ($P<.05$). This suggests that counselors tend to be more empathic toward high-risk callers. This perhaps highlights the need for counselors to adhere to a more caller-centered approach, ensuring that empathy is consistently exhibited for both low-risk and high-risk callers. Resources should perhaps be allocated equally for all callers rather than having a crisis intervention strategy that is tailored to prioritize callers needing emergency support.

To the best of the authors' knowledge, this is the first study that has identified the unique features of human vocals that are associated with the communication of empathy in a MH care setting. The results of this study have implications for the training of counselors and psychologists working for MH-related telephone helpline services. Additionally, these findings can serve as a training resource for MH professionals more broadly, enhancing the quality of care provided. The engineering of empathic chatbots, especially within a triage capacity, is another significant area of research that would benefit from the findings of this study [53].

However, collecting the vocal data of individuals for research purposes raises important ethical concerns. This research needs to prioritize user consent and caller privacy. It is recommended that people with lived experience in telephone counseling and MH be asked to assist with the co-design and coproduction of such research to ensure that any resulting training programs or monitoring systems are acceptable to users and meet consumer needs.

Acknowledgments

This research was funded by Swinburne University of Technology. The data call recordings were provided by On The Line Australia who was the main research partner of this study. We acknowledge Dr Jakqui Barnfield for her support from On The Line, Australia. We also thank Inge Gnatt, BPsych (Hons), and Sarah Dunning, BPsych, for their contributions as research annotators and Dr Maja Nedeljkovic for her supervision of the annotation process.

Data Availability

All data analyzed in this study are included in summary form under Multimedia Appendices. Our ethics approval does not allow the reporting of individual call data.

XSL•FO
RenderX

Authors' Contributions

RS developed the first draft of this study's manuscript and conducted a voice analysis, including the creation of the algorithm. RI contributed to developing the algorithm for detecting voice features that relate to empathy. RS contributed to the call annotation process, where variations in counselor empathy within the calls were identified. DM and RI contributed their supervision during the entire annotation process, confirming the well-being of the researchers involved. DM, RI, PA, and NW were involved in this study's manuscript revisions.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Online questionnaires used for data collection.
[DOCX File, 30 KB - formative_v9i1e67835_app1.docx]

Multimedia Appendix 2
Descriptive statistics of the analysis.
[DOCX File, 16 KB - formative_v9i1e67835_app2.docx]

Multimedia Appendix 3
Results of consistency between the raters.
[DOCX File, 16 KB - formative_v9i1e67835_app3.docx]

Multimedia Appendix 4
The results of LASSO regression. LASSO: Least Absolute Shrinkage and Selection Operator.
[DOCX File, 23 KB - formative_v9i1e67835_app4.docx]

Multimedia Appendix 5
Detailed results of the significance of the vocal features through GAMM, random forest, and binary logistics regression models. GAMM: Generalized Additive Mixed Model.
[DOCX File, 28 KB - formative_v9i1e67835_app5.docx]

Multimedia Appendix 6
Impact of expanded dataset and rater diversity on voice analysis results.
[DOCX File, 29 KB - formative_v9i1e67835_app6.docx]

Checklist 1
TRIPOD+AI (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis+Artificial Intelligence) statement for reporting the modeling process.
[PDF File, 620 KB - formative_v9i1e67835_app7.pdf]

References

1. Smith A. Cognitive empathy and emotional empathy in human behavior and evolution. Psychol Rec 2006 Jan;56(1):3-21. [doi: 10.1007/BF03395534]
2. Hojat M. Empathy in Health Professions Education and Patient Care: Springer Cham; 2016.
3. Derksen F, Bensing J, Lagro-Janssen A. Effectiveness of empathy in general practice: a systematic review. Br J Gen Pract 2013 Jan;63(606):e76-e84. [doi: 10.3399/bjgp13X660814] [Medline: 23336477]
4. Sturzu L, Lala A, Bisch M, Guitter M, Dobre D, Schwan R. Empathy and burnout - a cross-sectional study among mental healthcare providers in France. J Med Life 2019;12(1):21-29. [doi: 10.25122/jml-2018-0050] [Medline: 31123521]
5. Peppou LE, Economou M, Skali T, Papageorgiou C. From economic crisis to the COVID-19 pandemic crisis: evidence from a mental health helpline in Greece. Eur Arch Psychiatry Clin Neurosci 2021 Mar;271(2):407-409. [doi: 10.1007/s00406-020-01165-4] [Medline: 32666279]
6. Betancourt JA, Rosenberg MA, Zevallos A, Brown JR, Mileski M. The impact of COVID-19 on telemedicine utilization across multiple service lines in the United States. Healthcare (Basel) 2020 Oct 1;8(4):380. [doi: 10.3390/healthcare8040380] [Medline: 33019667]

XSL•FO
RenderX

7. Pavlova A, Scarth B, Witt K, Hetrick S, Fortune S. COVID-19 related innovation in Aotearoa/New Zealand mental health helplines and telehealth providers - mapping solutions and discussing sustainability from the perspective of service providers. Front Psychiatry 2022;13:973261. [doi: 10.3389/fpsyt.2022.973261] [Medline: 36111308]
8. Gibbons SB. Understanding empathy as a complex construct: a review of the literature. Clin Soc Work J 2011 Sep;39(3):243-252. [doi: 10.1007/s10615-010-0305-2]
9. Angus L, Kagan F. Empathic relational bonds and personal agency in psychotherapy: implications for psychotherapy supervision, practice, and research. Psychotherapy (Chic) 2007 Dec;44(4):371-377. [doi: 10.1037/0033-3204.44.4.371] [Medline: 22122315]
10. Pugh MA, Vetere A. Lost in translation: an interpretative phenomenological analysis of mental health professionals' experiences of empathy in clinical work with an interpreter. Psychol Psychother 2009 Sep;82(Pt 3):305-321. [doi: 10.1348/147608308X397059] [Medline: 19208292]
11. Tan L, Le MK, Yu CC, et al. Defining clinical empathy: a grounded theory approach from the perspective of healthcare workers and patients in a multicultural setting. BMJ Open 2021 Sep 14;11(9):e045224. [doi: 10.1136/bmjopen-2020-045224] [Medline: 34521657]
12. McHenry M, Parker PA, Baile WF, Lenzi R. Voice analysis during bad news discussion in oncology: reduced pitch, decreased speaking rate, and nonverbal communication of empathy. Support Care Cancer 2012 May;20(5):1073-1078. [doi: 10.1007/s00520-011-1187-8] [Medline: 21573770]
13. Gustafsson SR, Wahlberg AC. The telephone nursing dialogue process: an integrative review. BMC Nurs 2023;22(1):345. [doi: 10.1186/s12912-023-01509-0]
14. Mental health atlas 2020: review of the Eastern Mediterranean region. World Health Organization. 2022. URL: https://applications.emro.who.int/docs/9789292740443-eng.pdf?ua=1 [accessed 2025-03-27]
15. Callejas Z, Griol D. Conversational agents for mental health and wellbeing. dialog systems: a perspective from language, logic and computation. In: Dialog Systems: Springer, Cham; 2021:219-244. [doi: 10.1007/978-3-030-61438-6_11]
16. Pelau C, Dabija DC, Ene I. What makes an AI device human-like? The role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence in the service industry. Comput Human Behav 2021 Sep;122:106855. [doi: 10.1016/j.chb.2021.106855]
17. Iyer R, Nedeljkovic M, Meyer D. Using voice biomarkers to classify suicide risk in adult telehealth callers: retrospective observational study. JMIR Ment Health 2022 Aug 15;9(8):e39807. [doi: 10.2196/39807] [Medline: 35969444]
18. Posner K, Brown GK, Stanley B, et al. The Columbia-Suicide Severity Rating Scale: initial validity and internal consistency findings from three multisite studies with adolescents and adults. Am J Psychiatry 2011 Dec;168(12):1266-1277. [doi: 10.1176/appi.ajp.2011.10111704] [Medline: 22193671]
19. RStudio. RStudio IDE (Version 2024.04.2+764). 2024. URL: https://posit.co/products/open-source/rstudio/ [accessed 2025-04-09]
20. Audacity 3.5.1. Audacity Support. 2024. URL: https://support.audacityteam.org/additional-resources/changelog/older-versions/audacity-3.5/audacity-3.5.1 [accessed 2025-03-27]
21. Heck EJ, Davis CS. Differential expression of empathy in a counseling analogue. J Couns Psychol 1973;20(2):101-104. [doi: 10.1037/h0034171]
22. Mayer JD, Salovey P, Caruso DR, Sitarenios G. Measuring emotional intelligence with the MSCEIT V2.0. Emotion 2003 Mar;3(1):97-105. [doi: 10.1037/1528-3542.3.1.97] [Medline: 12899321]
23. Bodie GD. The Active-Empathic Listening Scale (AELS): conceptualization and evidence of validity within the interpersonal domain. Commun Q 2011 Jul;59(3):277-295. [doi: 10.1080/01463373.2011.583495]
24. Joshi A, Kale S, Chandel S, Pal DK. Likert Scale: explored and explained. BJAST 2015;7(4):396-403. [doi: 10.9734/BJAST/2015/14975]
25. Ownby KK. Use of the distress thermometer in clinical practice. J Adv Pract Oncol 2019 Mar;10(2):175-179. [Medline: 31538028]
26. Bikker AP, Fitzpatrick B, Murphy D, Mercer SW. Measuring empathic, person-centred communication in primary CARE nurses: validity and reliability of the Consultation and Relational Empathy (CARE) measure. BMC Fam Pract 2015 Oct 23;16:149. [doi: 10.1186/s12875-015-0374-y] [Medline: 26493072]
27. Anikin A. Soundgen: an open-source tool for synthesizing nonverbal vocalizations. Behav Res Methods 2019 Apr;51(2):778-792. [doi: 10.3758/s13428-018-1095-7] [Medline: 30054898]
28. Casale S, Russo A, Scebba G, Serrano S. Speech emotion classification using machine learning algorithms. Presented at: Proceedings of the 2th IEEE International Conference on Semantic Computing (ICSC 2008); Aug 4-7, 2008; Santa Monica, CA. [doi: 10.1109/ICSC.2008.43]
29. Rasmussen MA, Bro R. A tutorial on the LASSO approach to sparse modeling. Chemometr Intell Lab Syst 2012 Oct;119:21-31. [doi: 10.1016/j.chemolab.2012.10.003]
30. Sóskuthy M. Generalised additive mixed models for dynamic analysis in linguistics: a practical introduction. arXiv. Preprint posted online on Mar 15, 2017. [doi: 10.48550/arXiv.1703.05339]
31. Wood SN, Pya N, Säfken B. Smoothing parameter and model selection for general smooth models. J Am Stat Assoc 2016 Oct 1;111(516):1548-1563. [doi: 10.1080/01621459.2016.1180986]

XSL•FO
RenderX

32. Wood SN. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. J R Stat Soc Series B Stat Methodol 2011 Jan 1;73(1):3-36. [doi: 10.1111/j.1467-9868.2010.00749.x]
33. Noroozi F, Sapiński T, Kamińska D, Anbarjafari G. Vocal-based emotion recognition using random forests and decision tree. Int J Speech Technol 2017 Jun;20(2):239-246. [doi: 10.1007/s10772-017-9396-2]
34. Guo CY, Chen F, Chang YJ, Yan JT. Applying random forest classification to diagnose autism using acoustical voice-quality parameters during lexical tone production. Biomed Signal Process Control 2022 Aug;77:103811. [doi: 10.1016/j.bspc.2022.103811]
35. Raahul A, Sapthagiri R, Pankaj K, Vijayarajan V. Voice based gender classification using machine learning. IOP Conf Ser Mater Sci Eng 2017 Nov 1;263:042083. [doi: 10.1088/1757-899X/263/4/042083]
36. Jacob A. Modelling speech emotion recognition using logistic regression and decision trees. Int J Speech Technol 2017 Dec;20(4):897-905. [doi: 10.1007/s10772-017-9457-6]
37. Tafiadis D, Kosma EI, Chronopoulos SK, et al. Voice handicap index and interpretation of the cutoff points using receiver operating characteristic curve as screening for young adult female smokers. J Voice 2018 Jan;32(1):64-69. [doi: 10.1016/j.jvoice.2017.03.009] [Medline: 28392085]
38. Rice ME, Harris GT. What does it mean when age is related to recidivism among sex offenders? Law Hum Behav 2014 Apr;38(2):151-161. [doi: 10.1037/lhb0000052] [Medline: 23876093]
39. Uloza V, Latoszek BBV, Ulozaite-Staniene N, Petrauskas T, Maryn Y. A comparison of Dysphonia Severity Index and Acoustic Voice Quality Index measures in differentiating normal and dysphonic voices. Eur Arch Otorhinolaryngol 2018 Apr;275(4):949-958. [doi: 10.1007/s00405-018-4903-x] [Medline: 29442165]
40. Carter JV, Pan J, Rai SN, Galandiuk S. ROC-ing along: evaluation and interpretation of receiver operating characteristic curves. Surgery 2016 Jun;159(6):1638-1645. [doi: 10.1016/j.surg.2015.12.029] [Medline: 26962006]
41. Zou L, Choi YH, Guizzetti L, Shu D, Zou J, Zou G. Extending the DeLong algorithm for comparing areas under correlated receiver operating characteristic curves with missing data. Stat Med 2024 Sep 20;43(21):4148-4162. [doi: 10.1002/sim.10172] [Medline: 39013403]
42. Gerdes KE, Segal E. Importance of empathy for social work practice: integrating new science. Soc Work 2011 Apr;56(2):141-148. [doi: 10.1093/sw/56.2.141] [Medline: 21553577]
43. Regenbogen C, Schneider DA, Finkelmeyer A, et al. The differential contribution of facial expressions, prosody, and speech content to empathy. Cogn Emot 2012 Sep;26(6):995-1014. [doi: 10.1080/02699931.2011.631296]
44. Concannon S, Tomalin M. Measuring perceived empathy in dialogue systems. AI Soc 2024 Oct;39(5):2233-2247. [doi: 10.1007/s00146-023-01715-z]
45. Chen R, Chen H, Kulkarni A, et al. Detecting empathy in speech. Presented at: Interspeech 2024; Sep 1-5, 2024; Kos, Greece URL: https://www.isca-archive.org/interspeech_2024 [accessed 2025-04-09] [doi: 10.21437/Interspeech.2024-347]
46. Alam F, Danieli M, Riccardi G. Annotating and modeling empathy in spoken conversations. Comput SPEECH Lang 2018 Jul;50:40-61. [doi: 10.1016/j.csl.2017.12.003]
47. Zhao Q, Neumann DL, Cao Y, et al. Culture-sex interaction and the self-report empathy in Australians and mainland Chinese. Front Psychol 2019;10:396. [doi: 10.3389/fpsyg.2019.00396] [Medline: 30914986]
48. Hasan MR, Hossain MZ, Ghosh S, Krishna A, Gedeon T. Empathy detection from text, audiovisual, audio or physiological signals: task formulations and machine learning methods. arXiv. Preprint posted online on Dec 17, 2024. [doi: 10.48550/arXiv.2311.00721]
49. Besel LDS, Yuille JC. Individual differences in empathy: the role of facial expression recognition. Pers Individ Dif 2010 Jul;49(2):107-112. [doi: 10.1016/j.paid.2010.03.013]
50. Iyer R, Nedeljkovic M, Meyer D. Using vocal characteristics to classify psychological distress in adult helpline callers: retrospective observational study. JMIR Form Res 2022 Dec 19;6(12):e42249. [doi: 10.2196/42249] [Medline: 36534456]
51. Thölke P, Mantilla-Ramos YJ, Abdelhedi H, et al. Class imbalance should not throw you off balance: choosing the right classifiers and performance metrics for brain decoding with imbalanced data. Neuroimage 2023 Aug 15;277:120253. [doi: 10.1016/j.neuroimage.2023.120253] [Medline: 37385392]
52. Minixhofer C, Klejch O, Bell P. Evaluating and reducing the distance between synthetic and real speech distributions. 2023 Presented at: Interspeech 2023; Aug 20-24, 2023; Dublin, Ireland URL: https://www.isca-archive.org/interspeech_2023/ [accessed 2025-03-27] [doi: 10.21437/Interspeech.2023-1978]
53. Sanjeewa R, Iyer R, Apputhurai P, Wickramasinghe N, Meyer D. Empathic conversational agent platform designs and their evaluation in the context of mental health: systematic review. JMIR Ment Health 2024 Sep 9;11:e58974. [doi: 10.2196/58974] [Medline: 39250799]

Abbreviations

AEL: Active-Empathic Listening
AUC: area under the curve
C-SSRS: Columbia Suicide Severity Rating Scale
CTE: Carkhuff and Truax Empathy

XSL•FO
RenderX

GAMM: generalized additive mixed model
LASSO: Least Absolute Shrinkage and Selection Operator
MH: mental health
PEI: Perceived Emotional Intelligence
RS7: Rating Scale 7
TRIPOD+AI: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis+Artificial Intelligence

Edited by A Mavragani; submitted 22.10.24; peer-reviewed by R Gore, S Chowdhury; revised version received 16.02.25; accepted 18.02.25; published 16.04.25.

Please cite as:
Sanjeewa R, Iyer R, Apputhurai P, Wickramasinghe N, Meyer D
Machine Learning Approach to Identifying Empathy Using the Vocals of Mental Health Helpline Counselors: Algorithm Development and Validation
JMIR Form Res 2025;9:e67835
URL: https://formative.jmir.org/2025/1/e67835
doi:10.2196/67835

© Ruvini Sanjeewa, Ravi Iyer, Pragalathan Apputhurai, Nilmini Wickramasinghe, Denny Meyer. Originally published in JMIR Formative Research (https://formative.jmir.org), 16.4.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Acoustic and Natural Language Markers for Bipolar Disorder: A Pilot, mHealth Cross-Sectional Study

Cristina Crocamo[1], PhD; Riccardo Matteo Cioni[1], MD; Aurelia Canestro[1], MD; Christian Nasti[1], MD; Dario Palpella[1], MD; Susanna Piacenti[1], MD; Alessandra Bartoccetti[1], MD; Martina Re[1], MD; Valentina Simonetti[2], MSc; Chiara Barattieri di San Pietro[2,3], PhD; Maria Bulgheroni[2], MSc; Francesco Bartoli[1], MD, PhD; Giuseppe Carrà[1], MD, MSc, PhD

[1]School of Medicine and Surgery, University of Milano-Bicocca, via Cadore 48, Monza, Italy

[2]Ab.Acus, Milan, Italy

[3]Laboratory of Neurolinguistics and Experimental Pragmatics (NEP), University School for Advanced Studies IUSS, Pavia, Italy

Corresponding Author:
Cristina Crocamo, PhD
School of Medicine and Surgery, University of Milano-Bicocca, via Cadore 48, Monza, Italy

Abstract

Background: Monitoring symptoms of bipolar disorder (BD) is a challenge faced by mental health services. Speech patterns are crucial in assessing the current experiences, emotions, and thought patterns of people with BD. Natural language processing (NLP) and acoustic signal processing may support ongoing BD assessment within a mobile health (mHealth) framework.

Objective: Using both acoustic and NLP-based features from the speech of people with BD, we built an app-based tool and tested its feasibility and performance to remotely assess the individual clinical status.

Methods: We carried out a pilot, observational study, sampling adults diagnosed with BD from the caseload of the Nord Milano Mental Health Trust (Italy) to explore the relationship between selected speech features and symptom severity and to test their potential to remotely assess mental health status. Symptom severity assessment was based on clinician ratings, using the Young Mania Rating Scale (YMRS) and Montgomery-Åsberg Depression Rating Scale (MADRS) for manic and depressive symptoms, respectively. Leveraging a digital health tool embedded in a mobile app, which records and processes speech, participants self-administered verbal performance tasks. Both NLP-based and acoustic features were extracted, testing associations with mood states and exploiting machine learning approaches based on random forest models.

Results: We included 32 subjects (mean [SD] age 49.6 [14.3] years; 50% [16/32] females) with a MADRS median (IQR) score of 13 (21) and a YMRS median (IQR) score of 5 (16). Participants freely managed the digital environment of the app, without perceiving it as intrusive and reporting an acceptable system usability level (average score 73.5, SD 19.7). Small-to-moderate correlations between speech features and symptom severity were uncovered, with sex-based differences in predictive capability. Higher latency time (ρ=0.152), increased silences (ρ=0.416), and vocal perturbations correlated with depressive symptomatology. Pressure of speech based on the mean intraword time (ρ=–0.343) and lower voice instability based on jitter-related parameters (ρ ranging from –0.19 to –0.27) were detected for manic symptoms. However, a higher contribution of NLP-based and conversational features, rather than acoustic features, was uncovered, especially for predictive models for depressive symptom severity (NLP-based: R^2=0.25, mean squared error [MSE]=110.07, mean absolute error [MAE]=8.17; acoustics: R^2=0.11, MSE=133.75, MAE=8.86; combined: R^2=0.16; MSE=118.53, MAE=8.68).

Conclusions: Remotely collected speech patterns, including both linguistic and acoustic features, are associated with symptom severity levels and may help differentiate clinical conditions in individuals with BD during their mood state assessments. In the future, multimodal, smartphone-integrated digital ecological momentary assessments could serve as a powerful tool for clinical purposes, remotely complementing standard, in-person mental health evaluations.

(JMIR Form Res 2025;9:e65555) doi:10.2196/65555

KEYWORDS

digital mental health; remote assessment; mHealth; speech; NLP; natural language processing; acoustic; symptom severity; machine learning; markers; mental health; bipolar disorders; app; applications; multimodal; mobile health; voice; vocal; bipolar; verbal; emotion; emotional; psychiatry; psychiatric; mental illness

XSL•FO
RenderX

Introduction

Bipolar disorder (BD) is a lifelong, episodic illness characterized by mood recurrences, including manic or hypomanic, depressive, and mixed episodes [1-3]. The burden associated with BD, affecting families, carers, and mental health care systems, is heavy [4]. Community services often struggle in delivering regular monitoring of BD treatment needs, resulting in relapses that seem difficult to predict [4-6].

Language disturbances are among the core symptoms of acute episodes in BD, since speech patterns are modulated by the emotional and neurophysiological status [7,8]. Therefore, language may play a key role in the assessment of an individual's current experiences, emotions, thought patterns, and symptoms. While content analysis may reveal grandiosity associated with elevated mood, impulsivity, or changes in goal-directed activities, natural language may provide insights into mood fluctuations, cognitive processes, and behavioral patterns [9]. In particular, changes in the rate of speech are likely to indicate mood oscillations, including pressure of speech and increased verbosity during manic episodes [10] and poverty of speech and increased pause times during depressive episodes [11-13]. Clinicians are trained to recognize variations in language and voice, along with gestures and facial expressions, implicitly assessing both coherence and organization of speech and natural language features. However, this process is inevitably vulnerable to inconsistencies and biases.

Recent research in mental health and computer science has put forward computational approaches for speech analysis across a variety of mental disorders, proposing automated methods to assess and monitor the individual's mental state through speech patterns [14-18]. Promising techniques in speech acoustic signal processing [10,11,17,19-21], using mobile health (mHealth) technology, can bridge subjective and objective components across various stages, such as prediction of illness onset, diagnostic processes, assessment of severity, and forecast of treatment outcomes [22-25]. Indeed, natural language processing (NLP) techniques, exploring language resources (eg, lexical choices, syntax, and semantics) both qualitatively and quantitatively (eg, topic modeling, clustering, and classification), may produce deeper insights across different clinical conditions [9,26]. For example, observable linguistic traits (eg, increased use of both first-person pronouns and negative emotion expressions) can be identified among people with BD [23]. However, although linguistic features are informative, they are context-dependent and inferred according to word transcriptions [27]. Thus, speech analyses combining acoustic-dependent features (eg, speech prosody and voice quality) with NLP-based measures appear more promising in terms of model predictions, possibly providing a more accurate mental health assessment [23,27,28].

Indeed, research has shown that acoustic features are markers of emotional states in BD [29], and that quantifiable speech differences can predict the scores of scales such as the Young Mania Rating Scale (YMRS) and the Montgomery-Åsberg Depression Rating Scale (MADRS) [13,27]. On the other hand, recent evidence has shown how smartphone-based voice data [30] can enhance BD monitoring in real time, detecting possible mood changes [31,32]. Thus, speech-based systems embedded in smartphones might be useful tools for complementary, continuous assessments of BD clinical states. We therefore built an app-based tool, jointly using acoustic and NLP-based features from the speech of people with BD who delivered a narrative, and carried out a pilot study aimed at testing its feasibility and performance to remotely assess the individual clinical status. Continuous, uninterrupted spoken accounts, as supplied by individuals, provided the unique opportunity to combine communication style information from an in-depth set of acoustic features and NLP-based scores as potential digital markers of symptom severity in speech. We rigorously chose to test the tool's performance against standard psychometric assessments of mania and depression in order to explore its potential for remote, complementary assessments.

Methods

The report of this study adheres to the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) statement (checklist presented in Multimedia Appendix 1) [33].

Study Design and Sampling Strategies

We conducted a pilot, cross-sectional study involving adult participants (aged 18 years or older) from the caseload of the Nord Milano Mental Health Trust (Italy). The Trust includes 2 psychiatric intensive care units, with a total of 27 beds, and also provides community mental health care for the same 280,000 inhabitants of the northern area of the Metropolitan City of Milan through 4 community mental health teams with multidisciplinary staff. The relevant catchment area comprises highly urbanized, both deprived and affluent, districts.

Inclusion criteria comprised a diagnosis of BD and the willingness to participate in the study. People with physical impairments affecting their acoustic capabilities were excluded. Based on inclusion and exclusion criteria, eligible individuals were identified among individuals consecutively admitted to the Trust. Then, they were approached by the research team, explaining the purpose of the study and, if any, potential risks.

Ethical Considerations

Recruitment efforts were carried out in accordance with ethical guidelines to ensure the well-being and safety of all participants. Study participants signed a written informed consent and were not compensated for their involvement. The study received ethical approval (protocol number 172 - 17032023) from the local ethical committee. To maintain participant privacy and confidentiality, all study data were pseudonymized prior to analysis. No individual participants are identifiable in any images included in this manuscript or Multimedia Appendices.

Procedures

Acoustic data were retrieved by asking participants to self-administer verbal performance tasks through a mobile app on their smartphones (SPEAKapp; [34]). Clinical testing and app usage took place on the same day in the study setting (inpatient and outpatient services). Then, the System Usability Scale (SUS), a short 10-item questionnaire based on a 5-point

XSL•FO
RenderX

Likert scale, was administered to assess the usability [35] of the app.

Verbal performance in terms of prose recall was based on the Babcock test [36], for which participants were asked to listen to a short story characterized by graphic and intense contents (eg, a death in a car crash) and then to repeat what she or he remembered from this narrative. This enabled to capture speech timing patterns based on sustained speech samples.

The app gathered participants' verbal production by using the smartphone-integrated microphone, recording and processing participants' speech by leveraging Google Speech-To-Text APIs [37] and Python libraries (eg, Parselmouth for the Praat software [38]). Recordings involved the use of one audio channel based on the participant's voice in a controlled environment with minimal acoustic conditions. Both the raw audio data and the transcribed text content were processed to extract acoustic and NLP-based features from speech outputs. NLP and acoustic signal models were embedded in the backend part of the mobile app.

Measures

Consistent with recent evidence, we assumed speech as verbal behavior, the spoken output of the mental system underlying the language [39]. Through speech recognition, acoustic and linguistic features were extracted. Then, based on both NLP and acoustic features, we considered a multidimensional framework in order to generate appropriate discriminative information for the potential use of speech patterns as digital markers in BD [27,31]. A full description of selected features is provided in Table S1 in Multimedia Appendix 2.

NLP-Based, Semantic, and Conversational Indices

NLP-based scores were computed according to distributional semantic models, encompassing vectorial representations for the meaning of words in a multi-dimensional space.

Standard linguistic scores included both the number of words, indicative of poverty of speech, and the number of words produced that matched the story text. On the other hand, novel NLP-based scores integrated mean intraword time, estimating the average time taken to articulate or pronounce subsequent words, as an indicator of processing speed, as well as word mover's distance (WMD), capturing both lexical overlap and semantic similarity. In particular, WMD was estimated as the minimum cumulative distance between words required to exactly match the point cloud of the text of the full correct story (ie, the content distance between the full correct story and the story narrative produced by the participant), thus incorporating the semantic similarity between individual word pairs into the word distance metric [40]. In addition, latency time was calculated as a novel NLP-based score, taking into account the delay between the initiation of a spoken utterance or action and the production of its intended outcome or response when starting the task (ie, the first word).

Additional objective information was extracted from speech data. These quantitative measures included (1) speech duration, (2) speaking time (ie, phonation), (3) silence, (4) ratios of speaking time to speech duration as well as of silence to speaking time, and (5) speech rate.

Acoustic Indices From Vocal Signals (Prosodic Cues Indices)

Measures for prosodic cues (acoustic indices quantifying how people talk during conversations) were based on the signal's frequency and energy or amplitude. These were assumed to contribute to conveying paralinguistic meaning [41]. Based on nontextual data, acoustic components of speech were defined as the key phonetic elements, that is, objectively and reproducibly quantified speech sounds [27,42]. Fundamental frequency (F0) was measured by the frequency of phonation [43]. The short-term instability of the vibration of the vocal cords during phonation (ie, jitter-related indices) was also extracted (Table S1 in Multimedia Appendix 2). Higher jitter values indicated speech patterns likely characterized by irregularities or hesitations, thus mirroring potential underlying psychological distress or emotional instability. Furthermore, microperturbations of the ampleness of the signal (ie, how variable acoustic peaks refer to the period-to-period variability of the signal peak-to-peak amplitude) were identified as small fluctuations in the intensity of vocal sound waves by shimmer-related measures, with higher values indicating greater variability or instability, while lower ones suggesting more stable vocal intensity (ie, smoother and more regular speech production).

Since both periodic and nonperiodic sound waves may characterize the voice, the mean harmonics-to-noise ratio was used to measure the relationship between harmonic and nonharmonic voice elements. Noisier, more raucous voices (ie, not smooth or clear) were expected to show lower harmonics-to-noise ratios, indicating vocal cord tension or irritation, possibly suggesting emotional distress.

Psychometric Measures

Diagnosis of BD was confirmed by the Structured Clinical Interview for *DSM-5* (*Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition*; SCID-5). Based on clinician-rated assessments, depressive symptom severity was measured by the MADRS [44], while YMRS was used to assess manic symptoms [45]. Scores ranged from 0 to 60 for both MADRS [44] and YMRS [45]. In addition, cutoffs for severe mood symptoms were either a YMRS score ≥20 [46,47] or a MADRS score ≥19 [48].

Statistical Analyses

First, we summarized participants' characteristics, providing standard statistics for continuous and categorical variables. For both MADRS and YMRS, continuous scores were used. However, a supplementary analysis was performed based on clinically meaningful thresholds for symptom severity. A bivariate analysis was then carried out to measure the strength of the potential association between speech indices and psychometric measures. Features' summary statistics were plotted, and correlation coefficients (Pearson and Spearman, according to assumptions on data distribution, eg, normality) were estimated. Color gradient heat plots were also generated for data visualization. Taking into account potential sex

XSL•FO
RenderX

differences in speech acoustic indices [49-51], subgroup analyses were performed. Statistical significance was set at $P<.05$.

Second, based on state-of-the-art algorithms, NLP and acoustic features extracted from natural language and audio streams (Table S1 in Multimedia Appendix 2) were used to train machine-learning models to detect depressive and manic states by means of scores from MADRS and YMRS. Data were randomly split using a 5-fold nested cross-validation approach for training and testing in order to provide an unbiased evaluation of the model's performance. In particular, random forest (RF) models, with the potential to handle both linear and nonlinear relationships between features and the target variable, were implemented. The supervised learning algorithm, with no assumptions about the distribution of the target variable, was based on the ensemble learning method of different decision trees, whose predictions were aggregated using the scikit-learn library in Python. Exploiting the bagging techniques, building multiple decision trees, RF contributed to minimizing overfitting issues by randomizing the feature selection during each tree split. This was assumed to reduce sensitivity to noise and to make decision trees less correlated through the use of a unique subset of the initial data for every base model. Moreover, we deemed features scaling unnecessary due to both the properties of the RF model and the performance metrics of comparisons. Relevant models were trained to test final performance by metrics (ie, mean squared error [MSE], mean absolute error [MAE], and R-squared [R^2]). These tested overall performance, even controlling for sex. Shapley Additive Explanations values, showing features' impact, were plotted. Data were analyzed using Stata release 18 and Python (version 3.10.9).

Results

Sample Characteristics

We included 32 subjects with BD (mean age 49.6, SD 14.3 years; 50% [16/32] females). The mean (SD) age at onset was 24.4 (10) years. As a whole, participants experienced more manic (median 4, IQR 8) than depressive episodes (median 2, IQR 5). About 40% (12/32) of participants reported a previous mood episode within 1 year before study enrollment. The MADRS median (IQR) score was 13 (21), while the YMRS median (IQR) score was 5 (16). Considering the app usage, participants reported high SUS scores on average (mean 73.5, SD 19.7). Demographic and clinical details are fully provided in Table 1.

XSL•FO
RenderX

Table . Sample[a] characteristics.

Characteristics	BD[b] (N=32)
Sex, n (%)	
Female	16 (50)
Male	16 (50)
Age (years), mean (SD)	49.6 (14.3)
Marital status, n (%)	
In a relationship	12 (37)
Family situation, n (%)	
Living alone	11 (34)
Education, n (%)[c]	
Elementary	1 (3)
Middle	12 (37)
High	13 (41)
University or superior	5 (16)
Employment, n (%)[c]	
Employed	13 (41)
Setting, n (%)	
Outpatient	13 (41)
Inpatient	19 (59)
Polarity of first episode, n (%)[c]	
Depressive	12 (38)
Hypomaniac or maniac	13 (41)
Unknown	1 (3)
Age of onset[c] (years), mean (SD)	24.4 (10)
Family history[c]	11 (34%)
Hospitalizations, median (IQR)	
Lifetime	3 (7.5)
12 months	1 (2)
Suicide attempts (lifetime), n (%)	10 (31)
Alcohol use disorder (lifetime), n (%)	3 (9)
Substance use disorder (lifetime), n (%)	6 (19)
Medication, n (%)	
FGA[d]	6 (19)
SGA[e]	28 (87)
Mood stabilizer	26 (81)
Antidepressant	8 (25)
Benzodiazepine	16 (50)
Psychometric assessment	
Depressive symptoms (MADRS[f]), median (IQR)	13 (21)
MADRS <19, n (%)	17 (53)
MADRS ≥19, n (%)	15 (47)

XSL•FO
RenderX

Characteristics	BD[b] (N=32)
Manic symptoms (YMRS[g]), median (SD)	5 (16)
YMRS <20, n (%)	26 (81)
YMRS ≥20, n (%)	6 (19)
SUS[h] score, mean (SD)	73.5 (19.7)

[a]The sample is for a pilot, cross-sectional study in Italy.

[b]BD: bipolar disorder.

[c]Missing values: education (1), employment (2), age of onset (10), polarity of first episode (6), family history (10), alcohol use disorder (2), substance use disorder (1), FGA (2), SGA (1), mood stabilizer (2), antidepressant (3), benzodiazepine (4).

[d]FGA: first-generation antipsychotics.

[e]SGA: second-generation antipsychotics.

[f]MADRS: Montgomery-Åsberg Depression Rating Scale.

[g]YMRS: Young Mania Rating Scale.

[h]SUS: System Usability Scale (range 0 - 100).

Associations Between Symptom Severity and Speech Features

For descriptive purposes, NLP-based, conversational, and acoustic features are summarized in Figures S1A-S1D and S2A-S2D in Multimedia Appendix 3 by depressive and manic symptom severity, respectively.

In particular, grouping data into 2 categories (Multimedia Appendix 3), statistically significant differences by depressive symptoms' severity were found for many NLP-based and conversational-like measures, including word number, phonation (also as percentage over the speech duration), and mean intraword time. Correlation analyses, based on Spearman nonparametric analysis of symptom severity continuous scores, are displayed in Figures 1A-C and 2A-C. These showed that both the total number of words and the length of phonation, as well as the related percentage out of segment duration, were negatively correlated (coefficients=−0.35, −0.32, and −0.42) to depressive symptoms (Figure 1A). Consistent results were observed for the ratio between silence and phonation (coefficient=0.42), as well as for mean intraword time, which was positively correlated to depressive (coefficient=0.53) and negatively to manic (coefficient=−0.34) symptoms. Among items for depressive symptoms assessment, this correlation was particularly clear between acoustic features and suicidal thoughts (coefficients ranging from 0.18 to 0.51). In addition, latency time also showed a moderate, though obviously opposite, correlation with manic and depressive symptoms, respectively (coefficients=−0.28 and 0.15).

Subgroup analyses for NLP-based and conversational features revealed more pronounced relationships in females (Figure 1C) as compared with males (Figure 1B), showing a high correlation between depressive symptoms and mean intraword time (coefficient=0.75), phonation percentage (coefficient=−0.56), and, consequently, the silence-phonation ratio (coefficient=0.56). Similarly, latency time was negatively correlated to manic symptoms among females (coefficient=−0.60).

XSL•FO
RenderX

Figure 1. Correlation heatmap of NLP-based, semantic and conversational features in people with bipolar disorder. (A) Overall sample; (B) Male subgroup; (C) Female subgroup. MADRS: Montgomery-Åsberg Depression Rating Scale, YMRS: Young Mania Rating Scale.

XSL•FO
RenderX

On the other hand, a small positive correlation was uncovered between depressive symptoms and higher values of instability in speech patterns (jitter-related indices, with coefficients ranging from 0.10 to 0.16). In contrast, small-to-moderate negative correlations were observed between manic symptoms and lower values of instability (jitter-related indices, with coefficients ranging from −0.19 to −0.27). Small estimates were found for F0, respectively (coefficient=0.16 and −0.18; Figure 2A). Except for shimmer_apq11 (manic symptoms coefficient=−0.22), we did not find any substantial relationship between shimmer-related indices (describing stable and unstable vocal intensity and speech production) and symptomatology.

Subgroup analyses suggested a role for sex also in influencing acoustic features. In particular, we found deeper connections in males as compared with females, especially in terms of F0 and jitter-related indices (Figure 2B and C).

XSL•FO
RenderX

Figure 2. Correlation heatmap of acoustic features in people with bipolar disorder. (A) Overall sample; (B) Male subgroup; (C) Female subgroup. MADRS: Montgomery-Åsberg Depression Rating Scale, YMRS: Young Mania Rating Scale.

XSL•FO
RenderX

Predictive Models From Speech Features

Considering depressive symptoms, performance metrics showed a contribution of NLP-based and conversational features higher than what was attributable to acoustic ones (Table 2). In particular, mean intraword time, silence-phonation ratio, ppq5 jitter (ie, perturbations in F0), WMD, and percentage of phonation over duration all ranked high in terms of relative importance.

Including sex into the analysis, a differential contribution of various features (NLP-based and conversational vs acoustics) to the predictive models for depressive (Figure 3A) and manic (Figure 3B) symptoms can be found. However, as for manic symptoms, although a relative contribution of different NLP-based and acoustic (eg, F0 SD) features was recorded, we could not find any reliable estimates for the relevant model, even including sex. Table 2 shows detailed estimated performance metrics for testing for the trained RF regressors, even controlling for sex.

XSL•FO
RenderX

Table . Performance estimates for random forest regression models in people with bipolar disorder.

Performance[a]	Depressive symptoms		Manic symptoms	
	Unadjusted	Adjusted[b]	Unadjusted	Adjusted[b]
NLP[c]				
R^2 average	0.26	0.25	—[d]	—
Fold 1	0.10	−0.55	−0.54	0.18
Fold 2	0.48	0.53	−0.13	0.02
Fold 3	0.06	0.37	0.25	0.12
Fold 4	0.54	0.64	0.23	−0.42
Fold 5	0.13	0.26	0.01	147.98
Mean squared error average	105.46	110.07	153.78	147.98
Fold 1	136.73	259.26	223.06	92.05
Fold 2	33.02	46.49	121.79	156.74
Fold 3	137.64	104.25	135.66	121.60
Fold 4	79.32	33.00	134.85	167.15
Fold 5	140.61	107.35	153.32	202.35
Mean absolute error average	8.08	8.17	10.58	10.13
Fold 1	9.58	13.64	12.47	7.79
Fold 2	3.36	5.57	9.28	10.90
Fold 3	10.31	8.56	10.40	9.26
Fold 4	7.59	4.34	9.82	9.71
Fold 5	9.26	8.74	10.96	13.00
Acoustics				
R^2 average	—	0.11	—	—
Fold 1	0.29	–0.22	0.002	–0.22
Fold 2	–0.83	–0.10	–0.02	–0.15
Fold 3	–0.59	0.03	–0.14	–0.14
Fold 4	0.23	0.18	–0.38	–0.44
Fold 5	0.36	0.64	–0.28	0.01
Mean squared error average	161.64	133.75	162.86	163.51
Fold 1	47.97	222.18	68.9	125.14
Fold 2	333.00	200.40	160.47	122.80
Fold 3	202.17	85.04	185.63	175.34
Fold 4	128.54	148.52	272.25	225.23
Fold 5	96.49	12.62	127.06	170.30
Mean absolute error average	10.02	8.86	10.35	10.73
Fold 1	5.27	11.76	7.09	9.77
Fold 2	16.43	13.8	10.94	8.82
Fold 3	11.77	6.00	12.48	11.70
Fold 4	9.69	10.13	14.05	12.34

XSL•FO
RenderX

Performance[a]	Depressive symptoms		Manic symptoms	
	Unadjusted	Adjusted[b]	Unadjusted	Adjusted[b]
Fold 5	6.95	2.62	7.21	11.01
Combined				
R^2 average	0.05	0.16	—	—
Fold 1	0.32	0.60	–0.56	–0.13
Fold 2	–0.09	0.11	0.24	0.07
Fold 3	–0.29	0.07	0.08	–0.54
Fold 4	0.10	0.04	0.06	0.18
Fold 5	0.22	0.22	–0.41	0.14
Mean squared error average	120.90	118.53	135.94	140.03
Fold 1	87.51	34.13	158.54	183.39
Fold 2	183.67	111.11	60.67	122.32
Fold 3	184.71	164.45	192.81	126.43
Fold 4	47.71	148.83	178.84	112.36
Fold 5	100.91	134.10	88.83	155.68
Mean absolute error average	8.65	8.68	9.61	10.00
Fold 5	6.69	4.37	11.21	11.50
Fold 5	11.49	7.33	6.95	8.86
Fold 5	11.04	11.29	11.40	10.30
Fold 5	5.57	10.15	11.19	8.26
Fold 5	8.46	10.26	7.28	11.08

[a]Metrics for testing based on a nested cross-validation approach (pilot, cross-sectional study, N=32). Range for symptom scores: 0 - 60.

[b]Including sex.

[c]NLP: natural language processing.

[d]Not available.

XSL•FO
RenderX

Figure 3. Individual features contribution to depressive and manic symptoms predictions in sex-adjusted models among people with bipolar disorder.

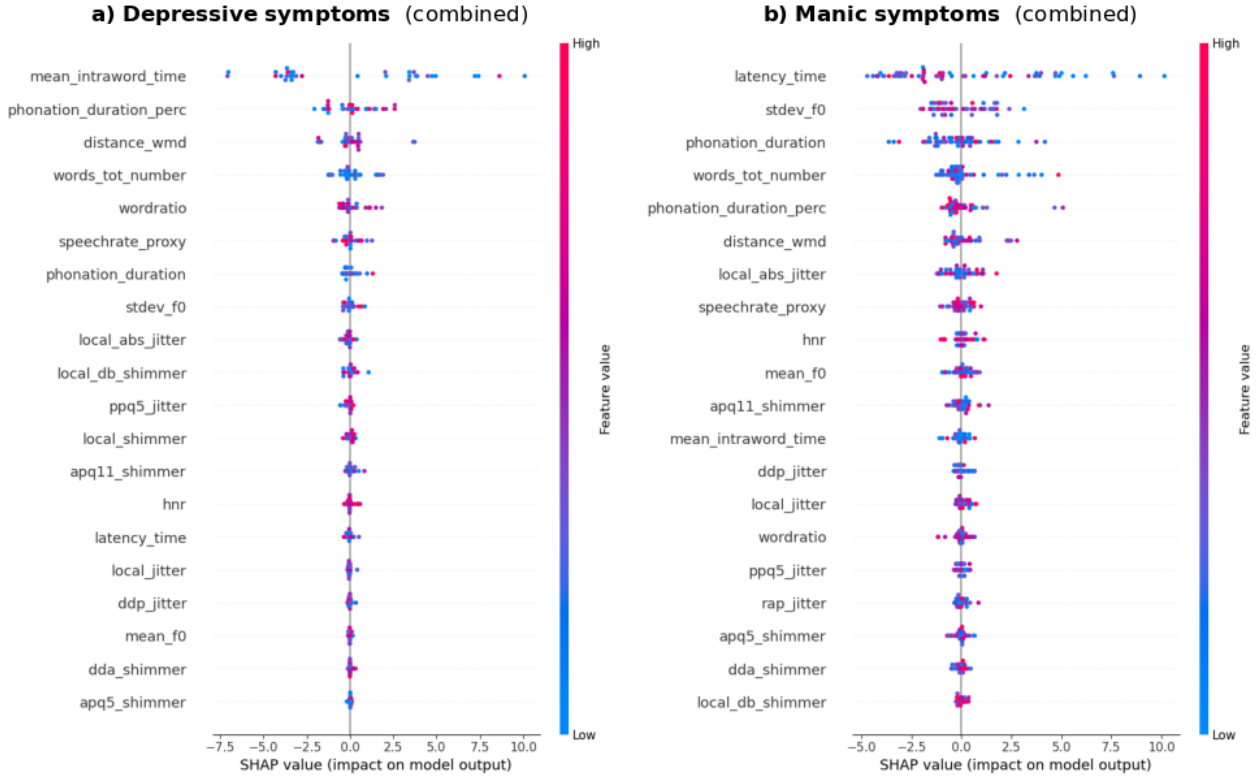

Discussion

Main Findings

This study aimed at piloting the simultaneous use of speech acoustics, as well as natural language features, to glean insights into BD depressive and manic symptoms. Our findings corroborate evidence on the relationships between symptom severity and speech features, supporting the potential predictive role for clinical purposes of digital mental health applications, embedded in a mHealth integrated system.

First, the speech of participants with BD showed that vocal perturbations (eg, higher instability and hesitations considering voice quality), latency time, and increased silences and pauses over time speaking all correlated to depressive symptoms. Consistently, increased depressive symptoms resulted in NLP-based features such as a smaller number of words and longer mean intraword time, with lower pressure of speech. In our exploratory study, this relationship was particularly clear among females. This effect was corroborated by the predictive model, showing a contribution of NLP-based and conversational features higher than for acoustic ones. This finding aligns with prior evidence, advocating that text-based features contribute more to model accuracy than audio parameters [18]. However, also the latter component (ie, fundamental frequency, jitter- and shimmer-related indices) deserves a careful assessment, since our findings show that these indices might have an impact at least among males to predict future episodes. Indeed, recent evidence from healthy populations sheds light on sex differences in speech markers (eg, prosodic features) with different acoustic cues conveying various emotions [50]. A combination of inherent biological dissimilarities, socialization processes, influences of the social environment, and cultural expectations might contribute to these differences in both expression and perception of related emotional prosody [52,53]. Moreover, individuals may modulate their speech to align with the dominant pitch range within a specific linguistic community [54], and similar modulation may occur in conversational dialogues versus monologues and in spontaneous versus elicited speech. Thus, this criterion should be taken into account when designing apps with speech recognition and processing tasks for people with BD [31].

Second, voice instability and hesitations, as well as mean intraword time, were negatively correlated to manic symptoms. However, the interpretation of the relationship between manic features and vocal abnormalities is not straightforward. Mixed findings emerged on the relationships between speech features and manic symptoms, preventing us from supporting our original hypothesis. One plausible explanation may stem from the sample characteristics. Indeed, our participants were more likely to report depressive symptoms, and just a few had severe manic features.

However, the overall moderate correlations between speech markers and symptom severity were consistent with previous work that used speech smartphone data to discriminate between different mood states [20,21]. It has been argued that speech features may be useful to detect a trait [55] rather than a state [56] in BD. However, alterations in voice perturbations have been observed when assessing vocal markers of suicidal ideation [57], and this makes further research for vocal features reasonable, at least for depressive conditions.

XSL•FO
RenderX

Smartphone-Based Applications

Consistent with previous research on smartphone-based applications designed to record and analyze speech patterns in real time, our findings emphasize the feasibility of a simple, yet clinically useful, application of digital technology [13]. In particular, we developed the frontend of the app as a basic digital environment, freely managed by participants on their own smartphones. Participants reported a high level of engagement with the tool, showing an acceptable system usability level as assessed by SUS [35], without perceiving intrusiveness of the recording of both elicited and spontaneous conversations.

Comparisons of the vocal performance of people with BD with unaffected relatives and healthy controls have shown a clear speech "fingerprint" of the clinical condition [58], suggesting the utility of multilevel inputs [59]. However, there is also the need for a wider understanding of fluctuations in symptom severity and mood states in this population [60]. The major strength of our study consists in the usefulness of different speech data (eg, linguistic, conversational, acoustics) to differentially identify symptoms of BD. Thus, for relapse prevention purposes, future research should possibly explore systems combining smartphone-based generated objective acoustics data with additional information, such as from facial expressions and gestures [61]. This would ultimately improve BD state prediction, even considering classification tasks [21,62-64].

Clinical Implications: Interdisciplinary Perspective

This pilot study represents a step forward in the identification and utilization of digital biomarkers for BD from natural language and audio streams, with implications for personalized mental health care and early intervention strategies. Our approach holds promise for complementary, remote assessments enhancing depressive and partly manic states prediction by exploiting participants' speech. This would have significant implications, especially considering BD fluctuating symptomatology. Nonetheless, leveraging live speech recordings as a predictive tool, repeated assessments are needed to identify individuals at risk of transitioning to depressive and manic states.

Despite promising findings from automated assessments, mental health care heavily relies on participant interviews, yet with often subjective reports, cognitive limitations, and stigma [18]. Integrated systems, aiming at taking advantage of candidate digital markers from speech recognition, would possibly boost a care approach in which digital technology enhances, but does not replace, existing models from clinical assessment [30]. Indeed, automated assessment does not inherently lead to adherence and engagement of individuals with BD [65].

Finally, clinical, hypothesis-driven research on BD should not be dismissed, since algorithms may not be considered a black-box replacement for traditional data modeling, but they rather integrate with other systems, embedding a substantial clinical validation [66,67].

Limitations and Future Directions

We should acknowledge some limitations of this study. Analyzing speech and natural language in individuals with BD implies a challenge due to the nature of the disorder and to ethical considerations.

First, properties of chosen machine-learning models may hamper identification of unknown patterns based on values that fall outside the training set. Effective NLP and supervised learning models may require high-quality, annotated datasets. While exploratory in nature, the study's limited sample size may have constrained the model's statistical power and the ability to capture the full complexity of the underlying data distribution, thereby hindering meaningful subgroup comparisons. Our preliminary findings should be replicated and extended in a larger, more diverse sample of people with BD to mitigate the risks associated with overfitting. Furthermore, future research should address classification approaches based on severity thresholds for both MADRS and YMRS. Accordingly, there is potential for alternative modeling approaches for regression tasks (eg, splines) that might be implemented in the future. While still considering the number of predictors, these may possibly enable a better understanding of the nature of the existing relationships and nonlinear patterns.

Consistently, the lack of standardized (linguistic and acoustic) markers represents a barrier when studying relationships with mood states. Indeed, the model may still learn to overfit to irrelevant or noisy features the data may contain, especially if they are informative in the training set by chance.

Furthermore, the speaker's identity may show a possible confounding role in a between-subject design. Therefore, studies with a longitudinal design (ie, within-subjects) should be recommended, deploying Ecological Momentary Assessment approaches [24,68]. In addition, speech patterns may generate misinterpretations if individual cultural and linguistic factors are not accounted for [69]. Similarly, speech during manic episodes may exhibit circumstantiality or tangentiality, where individuals provide excessive details or veer off-topic. Rapid speech, tangential thinking, or unconventional language use pose challenges for automatic speech recognition systems. Analyzing such complex speech patterns requires a deep evaluation of language and context, achieving appropriate understanding of an individual's usual way of communicating in order to distinguish changes associated with BD episodes.

Furthermore, in our study, speech features were averaged over relevant duration, thus constraining the role of temporal variations across related measures in predicting symptom severity. Future research should endeavor to integrate dynamic aspects of speech on mood states transitioning.

Finally, other clinical variables, not investigated in our sample, are likely to influence the individual's speech. For instance, it should be noted that anxiety and anxious distress, often co-occurring with bipolar depression [70], may significantly influence speech features [71], as well as medication prescribed [72-74] and drug or alcohol comorbid conditions [75].

XSL•FO
RenderX

Conclusions

Speech patterns, underlying both linguistic and acoustic features, are able to yield quantifiable differences, thus embodying digital markers of symptom severity. Multimodal, smartphone-integrated digital assessments could serve as powerful tools for clinical purposes to remotely complement standard mental health evaluations, potentially contributing to distinguish clinical conditions in people with BD. Feasibility of similar systems seems promising, though issues related to privacy, intrusiveness, and clinical therapeutic relationships should be carefully considered.

Acknowledgments

This research was supported by the FSE REACT-EU Competitive Research Grant Axis-IV DM 1062/2021: “Natural Language Processing in Digital Mental Health.” The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data Availability

The datasets including audio data streams supporting the conclusions of this article are not publicly available as the original source has not granted permission to share that information but may be available from the corresponding author on reasonable request.

Authors' Contributions

CC, FB, and GC handled conceptualization. RMC, AC, CN, DP, SP, AB, and MR performed investigation. CC, FB, VS, CB, and MB contributed to methodology. VS and MB assisted with software. CC conducted formal analysis. GC performed supervision. CC contributed to writing – original draft. CC, RMC, AC, CN, DP, SP, AB, MR, VS, CB, MB, FB, and GC contributed to writing – review and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Checklist.
[DOC File, 87 KB - formative_v9i1e65555_app1.doc]

Multimedia Appendix 2
Features.
[DOCX File, 16 KB - formative_v9i1e65555_app2.docx]

Multimedia Appendix 3
Supplementary analyses.
[DOCX File, 329 KB - formative_v9i1e65555_app3.docx]

References

1. Bartoli F, Crocamo C, Carrà G. Clinical correlates of DSM-5 mixed features in bipolar disorder: a meta-analysis. J Affect Disord 2020 Nov 1;276:234-240. [doi: 10.1016/j.jad.2020.07.035] [Medline: 32697704]
2. Grande I, Berk M, Birmaher B, Vieta E. Bipolar disorder. Lancet 2016 Apr 9;387(10027):1561-1572. [doi: 10.1016/S0140-6736(15)00241-X] [Medline: 26388529]
3. McIntyre RS, Berk M, Brietzke E, et al. Bipolar disorders. Lancet 2020 Dec 5;396(10265):1841-1856. [doi: 10.1016/S0140-6736(20)31544-0] [Medline: 33278937]
4. Karambelas GJ, Filia K, Byrne LK, Allott KA, Jayasinghe A, Cotton SM. A systematic review comparing caregiver burden and psychological functioning in caregivers of individuals with schizophrenia spectrum disorders and bipolar disorders. BMC Psychiatry 2022 Jun 23;22(1):422. [doi: 10.1186/s12888-022-04069-w] [Medline: 35733174]
5. Fajutrao L, Locklear J, Priaulx J, Heyes A. A systematic review of the evidence of the burden of bipolar disorder in Europe. Clin Pract Epidemiol Ment Health 2009 Jan 23;5:3. [doi: 10.1186/1745-0179-5-3] [Medline: 19166608]
6. Ogilvie AD, Morant N, Goodwin GM. The burden on informal caregivers of people with bipolar disorder. Bipolar Disord 2005;7 Suppl 1(Suppl 1):25-32. [doi: 10.1111/j.1399-5618.2005.00191.x] [Medline: 15762866]
7. Goodwin FK, Jamison KR. Manic-Depressive Illness: Bipolar Disorders and Recurrent Depression: Oxford University Press, USA; 2007.

XSL•FO
RenderX

8. Karam ZN, Provost EM, Singh S, et al. Ecologically valid long-term mood monitoring of individuals with bipolar disorder using speech. Proc IEEE Int Conf Acoust Speech Signal Process 2014 May;2014:4858-4862. [doi: 10.1109/ICASSP.2014.6854525] [Medline: 27630535]
9. Harvey D, Lobban F, Rayson P, Warner A, Jones S. Natural language processing methods and bipolar disorder: scoping review. JMIR Ment Health 2022 Apr 22;9(4):e35928. [doi: 10.2196/35928] [Medline: 35451984]
10. Birnbaum ML, Abrami A, Heisig S, et al. Acoustic and facial features from clinical interviews for machine learning-based psychiatric diagnosis: algorithm development. JMIR Ment Health 2022 Jan 24;9(1):e24699. [doi: 10.2196/24699] [Medline: 35072648]
11. Gideon J, Provost EM, McInnis M. Mood state prediction from speech of varying acoustic quality for individuals with bipolar disorder. Proc IEEE Int Conf Acoust Speech Signal Process 2016 Mar;2016:2359-2363. [doi: 10.1109/ICASSP.2016.7472099] [Medline: 27570493]
12. Guidi A, Schoentgen J, Bertschy G, Gentili C, Scilingo EP, Vanello N. Features of vocal frequency contour and speech rhythm in bipolar disorder. Biomed Signal Process Control 2017 Aug;37:23-31. [doi: 10.1016/j.bspc.2017.01.017]
13. Maxhuni A, Muñoz-Meléndez A, Osmani V, Perez H, Mayora O, Morales EF. Classification of bipolar disorder episodes based on analysis of voice and motor activity of patients. Pervasive Mob Comput 2016 Sep;31:50-66. [doi: 10.1016/j.pmcj.2016.01.008]
14. Matton K, McInnis MG, Provost EM. Into the wild: transitioning from recognizing mood in clinical interactions to personal conversations for individuals with bipolar disorder. 2019 Presented at: Interspeech 2019; Sep 15-19, 2019; Graz, Austria p. 1438-1442 URL: https://www.isca-archive.org/interspeech_2019 [doi: 10.21437/Interspeech.2019-2698]
15. Arevian AC, Bone D, Malandrakis N, et al. Clinical state tracking in serious mental illness through computational analysis of speech. PLoS ONE 2020;15(1):e0225695. [doi: 10.1371/journal.pone.0225695] [Medline: 31940347]
16. Girard JM, Vail AK, Liebenthal E, et al. Computational analysis of spoken language in acute psychosis and mania. Schizophr Res 2022 Jul;245:97-115. [doi: 10.1016/j.schres.2021.06.040] [Medline: 34456131]
17. Low DM, Bentley KH, Ghosh SS. Automated assessment of psychiatric disorders using speech: a systematic review. Laryngoscope Investig Otolaryngol 2020 Feb;5(1):96-116. [doi: 10.1002/lio2.354] [Medline: 32128436]
18. Malgaroli M, Hull TD, Zech JM, Althoff T. Natural language processing for mental health interventions: a systematic review and research framework. Transl Psychiatry 2023 Oct 6;13(1):309. [doi: 10.1038/s41398-023-02592-2] [Medline: 37798296]
19. Cummins N, Baird A, Schuller BW. Speech analysis for health: current state-of-the-art and the increasing impact of deep learning. Methods 2018 Dec 1;151:41-54. [doi: 10.1016/j.ymeth.2018.07.007] [Medline: 30099083]
20. Faurholt-Jepsen M, Busk J, Frost M, et al. Voice analysis as an objective state marker in bipolar disorder. Transl Psychiatry 2016 Jul 19;6(7):e856. [doi: 10.1038/tp.2016.123] [Medline: 27434490]
21. Faurholt-Jepsen M, Rohani DA, Busk J, et al. Discriminating between patients with unipolar disorder, bipolar disorder, and healthy control individuals based on voice features collected from naturalistic smartphone calls. Acta Psychiatr Scand 2022 Mar;145(3):255-267. [doi: 10.1111/acps.13391] [Medline: 34923626]
22. Daus H, Bloecher T, Egeler R, De Klerk R, Stork W, Backenstrass M. Development of an emotion-sensitive mHealth approach for mood-state recognition in bipolar disorder. JMIR Ment Health 2020 Jul 3;7(7):e14267. [doi: 10.2196/14267] [Medline: 32618577]
23. Dikaios K, Rempel S, Dumpala SH, Oore S, Kiefte M, Uher R. Applications of speech analysis in psychiatry. Harv Rev Psychiatry 2023;31(1):1-13. [doi: 10.1097/HRP.0000000000000356] [Medline: 36608078]
24. Dunster GP, Swendsen J, Merikangas KR. Real-time mobile monitoring of bipolar disorder: a review of evidence and future directions. Neuropsychopharmacology 2021 Jan;46(1):197-208. [doi: 10.1038/s41386-020-00830-5] [Medline: 32919408]
25. Marzano L, Bardill A, Fields B, et al. The application of mHealth to mental health: opportunities and challenges. Lancet Psychiatry 2015 Oct;2(10):942-948. [doi: 10.1016/S2215-0366(15)00268-0] [Medline: 26462228]
26. Le Glaz A, Haralambous Y, Kim-Dufor DH, et al. Machine learning and natural language processing in mental health: systematic review. J Med Internet Res 2021 May 4;23(5):e15708. [doi: 10.2196/15708] [Medline: 33944788]
27. Farrús M, Codina-Filbà J, Escudero J. Acoustic and prosodic information for home monitoring of bipolar disorder. Health Informatics J 2021;27(1):1460458220972755. [doi: 10.1177/1460458220972755] [Medline: 33438502]
28. Gong Y, Poellabauer C. Topic modeling based multi-modal depression detection. 2017 Oct 23 Presented at: MM '17; Oct 23-27, 2017; Mountain View California USA p. 2017 URL: https://dl.acm.org/doi/proceedings/10.1145/3133944 [doi: 10.1145/3133944.3133945]
29. Muaremi A, Gravenhorst F, Grünerbl A, Arnrich B, Tröster G. Assessing bipolar episodes using speech cues derived from phone calls. In: Lect Notes Inst Comput Sci Soc Inform Telecommun Eng 2014:103-114. [doi: 10.1007/978-3-319-11564-1_11]
30. Bond RR, Mulvenna MD, Potts C, O'Neill S, Ennis E, Torous J. Digital transformation of mental health services. Npj Ment Health Res 2023 Aug 22;2(1):13. [doi: 10.1038/s44184-023-00033-y] [Medline: 38609479]
31. Flanagan O, Chan A, Roop P, Sundram F. Using acoustic speech patterns from smartphones to investigate mood disorders: scoping review. JMIR Mhealth Uhealth 2021 Sep 17;9(9):e24352. [doi: 10.2196/24352] [Medline: 34533465]

XSL•FO
RenderX

32. de Oliveira L, Portugal LCL, Pereira M, et al. Predicting bipolar disorder risk factors in distressed young adults from patterns of brain activation to reward: a machine learning approach. Biol Psychiatry Cogn Neurosci Neuroimaging 2019 Aug;4(8):726-733. [doi: 10.1016/j.bpsc.2019.04.005] [Medline: 31201147]
33. von Elm E, Altman DG, Egger M, et al. Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. BMJ 2007 Oct 20;335(7624):806-808. [doi: 10.1136/bmj.39335.541782.AD] [Medline: 17947786]
34. Ab.acus Srl. URL: https://www.ab-acus.eu/index.php/portfolio-items/speakapp/ [accessed 2024-03-19]
35. Hyzy M, Bond R, Mulvenna M, et al. System Usability Scale benchmarking for digital health apps: meta-analysis. JMIR Mhealth Uhealth 2022 Aug 18;10(8):e37290. [doi: 10.2196/37290] [Medline: 35980732]
36. Italian standardization and classification of neuropsychological tests. The Italian Group on the Neuropsychological Study of Aging. Ital J Neurol Sci 1987 Dec;Suppl 8(1-120):1-120. [Medline: 3330072]
37. Google speech-to-text apis. Google Cloud. URL: https://cloud.google.com/speech-to-text [accessed 2025-04-09]
38. Jadoul Y, Thompson B, de Boer B. Introducing Parselmouth: a Python interface to Praat. J Phon 2018 Nov;71:1-15. [doi: 10.1016/j.wocn.2018.07.001]
39. de Boer JN, Brederoo SG, Voppel AE, Sommer IEC. Anomalies in language as a biomarker for schizophrenia. Curr Opin Psychiatry 2020 May;33(3):212-218. [doi: 10.1097/YCO.0000000000000595] [Medline: 32049766]
40. Kusner M, Sun Y, Kolkin N, Weinberger K. From word embeddings to document distances. 2015 Presented at: Proceedings of the 32nd International Conference on Machine Learning, PMLR; Jul 5-7, 2015; Lille, France p. 957-966.
41. Nadeu M, Prieto P. Pitch range, gestural information, and perceived politeness in Catalan. J Pragmat 2011 Feb;43(3):841-854. [doi: 10.1016/j.pragma.2010.09.015]
42. Smith SW. Digital Signal Processing: A Practical Guide for Engineers and Scientists: California Technical Publishing; 2002:978. [doi: 10.1016/B978-0-7506-7444-7.X5036-5]
43. Ververidis D, Kotropoulos C. Emotional speech recognition: resources, features, and methods. Speech Commun 2006 Sep;48(9):1162-1181. [doi: 10.1016/j.specom.2006.04.003]
44. Montgomery SA, Asberg M. A new depression scale designed to be sensitive to change. Br J Psychiatry 1979 Apr;134:382-389. [doi: 10.1192/bjp.134.4.382] [Medline: 444788]
45. Young RC, Biggs JT, Ziegler VE, Meyer DA. A rating scale for mania: reliability, validity and sensitivity. Br J Psychiatry 1978 Nov;133:429-435. [doi: 10.1192/bjp.133.5.429] [Medline: 728692]
46. Lukasiewicz M, Gerard S, Besnard A, et al. Young Mania Rating Scale: how to interpret the numbers? Determination of a severity threshold and of the minimal clinically significant difference in the EMBLEM cohort. Int J Methods Psychiatr Res 2013 Mar;22(1):46-58. [doi: 10.1002/mpr.1379] [Medline: 23526724]
47. Samara MT, Levine SZ, Leucht S. Linkage of Young Mania Rating Scale to Clinical Global Impression Scale to enhance utility in clinical practice and research trials. Pharmacopsychiatry 2023 Jan;56(1):18-24. [doi: 10.1055/a-1841-6672] [Medline: 35896419]
48. Thase ME, Harrington A, Calabrese J, Montgomery S, Niu X, Patel MD. Evaluation of MADRS severity thresholds in patients with bipolar depression. J Affect Disord 2021 May 1;286:58-63. [doi: 10.1016/j.jad.2021.02.043] [Medline: 33677183]
49. Eichhorn JT, Kent RD, Austin D, Vorperian HK. Effects of aging on vocal fundamental frequency and vowel formants in men and women. J Voice 2018 Sep;32(5):644. [doi: 10.1016/j.jvoice.2017.08.003] [Medline: 28864082]
50. Ertürk A, Gürses E, Kulak Kayıkcı ME. Sex related differences in the perception and production of emotional prosody in adults. Psychol Res 2024 Mar;88(2):449-457. [doi: 10.1007/s00426-023-01865-1] [Medline: 37542581]
51. Mendoza E, Valencia N, Muñoz J, Trujillo H. Differences in voice quality between men and women: use of the long-term average spectrum (LTAS). J Voice 1996 Mar;10(1):59-66. [doi: 10.1016/s0892-1997(96)80019-1] [Medline: 8653179]
52. Chaplin TM. Gender and emotion expression: a developmental contextual perspective. Emot Rev 2015 Jan;7(1):14-21. [doi: 10.1177/1754073914544408] [Medline: 26089983]
53. Lin Y, Ding H, Zhang Y. Gender differences in identifying facial, prosodic, and semantic emotions show category- and channel-specific effects mediated by encoder's gender. J Speech Lang Hear Res 2021 Aug 9;64(8):2941-2955. [doi: 10.1044/2021_JSLHR-20-00553] [Medline: 34310173]
54. Aung T, Puts D. Voice pitch: a window into the communication of social power. Curr Opin Psychol 2020 Jun;33:154-161. [doi: 10.1016/j.copsyc.2019.07.028] [Medline: 31445439]
55. Zhang J, Pan Z, Gui C, et al. Analysis on speech signal features of manic patients. J Psychiatr Res 2018 Mar;98:59-63. [doi: 10.1016/j.jpsychires.2017.12.012] [Medline: 29291581]
56. Guidi A, Salvi S, Ottaviano M, et al. Smartphone application for the analysis of prosodic features in running speech with a focus on bipolar disorders: system performance evaluation and case study. Sensors (Basel) 2015 Nov 6;15(11):28070-28087. [doi: 10.3390/s151128070] [Medline: 26561811]
57. Ozdas A, Shiavi RG, Silverman SE, Silverman MK, Wilkes DM. Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk. IEEE Trans Biomed Eng 2004 Sep;51(9):1530-1540. [doi: 10.1109/TBME.2004.827544] [Medline: 15376501]

XSL•FO
RenderX

58. Faurholt-Jepsen M, Rohani DA, Busk J, Vinberg M, Bardram JE, Kessing LV. Voice analyses using smartphone-based data in patients with bipolar disorder, unaffected relatives and healthy control individuals, and during different affective states. Int J Bipolar Disord 2021 Dec 1;9(1):38. [doi: 10.1186/s40345-021-00243-3] [Medline: 34850296]
59. Torous J, Bucci S, Bell IH, et al. The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. World Psychiatry 2021 Oct;20(3):318-335. [doi: 10.1002/wps.20883] [Medline: 34505369]
60. Balcombe L, De Leo D. Digital mental health challenges and the horizon ahead for solutions. JMIR Ment Health 2021 Mar 29;8(3):e26811. [doi: 10.2196/26811] [Medline: 33779570]
61. Soenksen LR, Ma Y, Zeng C, et al. Integrated multimodal artificial intelligence framework for healthcare applications. NPJ Digit Med 2022 Sep 20;5(1):149. [doi: 10.1038/s41746-022-00689-4] [Medline: 36127417]
62. Grünerbl A, Muaremi A, Osmani V, et al. Smartphone-based recognition of states and state changes in bipolar disorder patients. IEEE J Biomed Health Inform 2015 Jan;19(1):140-148. [doi: 10.1109/JBHI.2014.2343154] [Medline: 25073181]
63. Osmani V. Smartphones in mental health: detecting depressive and manic episodes. IEEE Pervasive Comput 2015;14(3):10-13. [doi: 10.1109/MPRV.2015.54]
64. Cohen J, Richter V, Neumann M, et al. A multimodal dialog approach to mental state characterization in clinically depressed, anxious, and suicidal populations. Front Psychol 2023;14:1135469. [doi: 10.3389/fpsyg.2023.1135469] [Medline: 37767217]
65. Or F, Torous J, Onnela JP. High potential but limited evidence: using voice data from smartphones to monitor and diagnose mood disorders. Psychiatr Rehabil J 2017 Sep;40(3):320-324. [doi: 10.1037/prj0000279] [Medline: 28891659]
66. Garcia-Ceja E, Riegler M, Nordgreen T, Jakobsen P, Oedegaard KJ, Tørresen J. Mental health monitoring with multimodal sensing and machine learning: a survey. Pervasive Mob Comput 2018 Dec;51:1-26. [doi: 10.1016/j.pmcj.2018.09.003]
67. McCoy LG, Brenna CTA, Chen SS, Vold K, Das S. Believing in black boxes: machine learning for healthcare does not need explainability to be evidence-based. J Clin Epidemiol 2022 Feb;142:252-257. [doi: 10.1016/j.jclinepi.2021.11.001] [Medline: 34748907]
68. Yerushalmi M, Sixsmith A, Pollock Star A, King DB, O'Rourke N. Ecological momentary assessment of bipolar disorder symptoms and partner affect: longitudinal pilot study. JMIR Form Res 2021 Sep 2;5(9):e30472. [doi: 10.2196/30472] [Medline: 34473069]
69. Clark EL, Easton C, Verdon S. The impact of linguistic bias upon speech-language pathologists' attitudes towards non-standard dialects of English. Clin Linguist Phon 2021 Jun 3;35(6):542-559. [doi: 10.1080/02699206.2020.1803405] [Medline: 32781853]
70. Bartoli F, Bachi B, Callovini T, et al. Anxious distress in people with major depressive episodes: a cross-sectional analysis of clinical correlates. CNS Spectr 2024 Feb;29(1):49-53. [doi: 10.1017/S1092852923002377] [Medline: 37489522]
71. Malgaroli M, Hull TD, Calderon A, Simon NM. Linguistic markers of anxiety and depression in somatic symptom and related disorders: observational study of a digital intervention. J Affect Disord 2024 May 1;352:133-137. [doi: 10.1016/j.jad.2024.02.012] [Medline: 38336165]
72. Bartoli F, Crocamo C, Clerici M, Carrà G. Allopurinol as add-on treatment for mania symptoms in bipolar disorder: systematic review and meta-analysis of randomised controlled trials. Br J Psychiatry 2017 Jan;210(1):10-15. [doi: 10.1192/bjp.bp.115.180281] [Medline: 27856422]
73. Bartoli F, Cavaleri D, Bachi B, et al. Repurposed drugs as adjunctive treatments for mania and bipolar depression: a meta-review and critical appraisal of meta-analyses of randomized placebo-controlled trials. J Psychiatr Res 2021 Nov;143:230-238. [doi: 10.1016/j.jpsychires.2021.09.018] [Medline: 34509090]
74. Bartoli F, Cavaleri D, Nasti C, et al. Long-acting injectable antipsychotics for the treatment of bipolar disorder: evidence from mirror-image studies. Ther Adv Psychopharmacol 2023;13:20451253231163682. [doi: 10.1177/20451253231163682] [Medline: 36994116]
75. Carrà G, Scioli R, Monti MC, Marinoni A. Severity profiles of substance-abusing patients in Italian community addiction facilities: influence of psychiatric concurrent disorders. Eur Addict Res 2006;12(2):96-101. [doi: 10.1159/000090429] [Medline: 16543745]

Abbreviations

BD: bipolar disorder
DSM-5: Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition
MADRS: Montgomery-Åsberg Depression Rating Scale
MAE: mean absolute error
mHealth: mobile health
MSE: mean squared error
NLP: natural language processing
RF: random forest
SCID-5: Structured Clinical Interview for DSM-5
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

XSL•FO
RenderX

WMD: word mover's distance
YMRS: Young Mania Rating Scale

Edited by A Mavragani; submitted 20.08.24; peer-reviewed by D Meyer, V Martin; revised version received 29.01.25; accepted 12.02.25; published 16.04.25.

Please cite as:
Crocamo C, Cioni RM, Canestro A, Nasti C, Palpella D, Piacenti S, Bartoccetti A, Re M, Simonetti V, Barattieri di San Pietro C, Bulgheroni M, Bartoli F, Carrà G
Acoustic and Natural Language Markers for Bipolar Disorder: A Pilot, mHealth Cross-Sectional Study
JMIR Form Res 2025;9:e65555
URL: https://formative.jmir.org/2025/1/e65555
doi:10.2196/65555

© Cristina Crocamo, Riccardo Matteo Cioni, Aurelia Canestro, Christian Nasti, Dario Palpella, Susanna Piacenti, Alessandra Bartoccetti, Martina Re, Valentina Simonetti, Chiara Barattieri di San Pietro, Maria Bulgheroni, Francesco Bartoli, Giuseppe Carrà. Originally published in JMIR Formative Research (https://formative.jmir.org), 16.4.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Using the Healthy Native Youth Implementation Toolbox to Provide Web-Based Adolescent Health Promotion Decision Support to American Indian and Alaska Native Communities: Implementation Study

Amrita Sidhu[1*], MPH, MDS; Ross Shegog[1*], PhD; Stephanie Craig-Rushing[2*], PhD; Nicole Trevino[2*], MA; Michelle Singer[2*], BSc; Cornelia Jessen[3*], MA; Gwenda Gorman[4*], BSc; Sean Simpson[5*]; Melissa Peskin[1], PhD; Belinda Hernandez[1], PhD; Christine Markham[1*], PhD

[1]Department of Health Promotion and Behavioral Sciences, School of Public Health, UTHealth Houston, 7000 Fannin St, Houston, TX, United States

[2]Northwest Portland Area Indian Health Board, Portland, OR, United States

[3]Alaska Native Tribal Health Consortium, Anchorage, AK, United States

[4]Inter Tribal Council of Arizona, Inc, Phoenix, AZ, United States

[5]Good Medicine Tribal Public Health Consulting, Madison, WI, United States

[*]these authors contributed equally

Corresponding Author:
Amrita Sidhu, MPH, MDS
Department of Health Promotion and Behavioral Sciences, School of Public Health, UTHealth Houston, 7000 Fannin St, Houston, TX, United States

Abstract

Background: American Indian and Alaska Native (AI/AN) youth experience numerous health inequities, including those in sexual, reproductive, and mental health. Implementation of culturally relevant, age-appropriate evidence-based programs may mitigate these inequities. However, numerous barriers limit the adoption and implementation of evidence-based adolescent health promotion programs in AI/AN communities.

Objective: This study examines user reach and engagement from 2022 to 2024 of web-based decision support (the Healthy Native Youth [HNY] website and the embedded HNY Implementation Toolbox), designed to increase the implementation of evidence-based adolescent health promotion programming in AI/AN communities.

Methods: Promotional strategies were designed for optimal geographic reach to Tribal organizations, opinion leaders, federal decision makers, and funders. Promotional channels included grassroots, community, and professional networks. We used Google Analytics to examine the uptake of the HNY website and HNY Implementation Toolbox from January 2022 to January 2024. The Toolbox provides culturally relevant tools and templates to help users navigate through 5 phases of program adoption and implementation: Gather, Choose, Prepare, Implement, and Grow. User reach was estimated by demographic characteristics and geographic location; user engagement was estimated by visit frequency and duration, bounce rates, and frequency of page and tool access.

Results: Over the study period, page views of the HNY website and HNY Toolbox increased 10-fold and 27-fold, respectively. Over the 2-year evaluation period since the Toolbox "go live" date, approximately 1 in 8 users of the HNY website visited the Toolbox. The majority of HNY website users were located in Washington (n=1515), California (n=1290), and Oregon (n=1019) and were aged between 18 and 24 (n=1559, 21.7%) and 25 - 34 (n=1676, 23.29%) years. Toolbox users were primarily located in California (n=1238), Washington (n=1142), and Oregon (n=986), mostly aged between 35 and 44 years (n=444, 35%). Both website and Toolbox users were primarily female, who accessed the site and Toolbox via desktop computers. The most frequently accessed phase pages within the Implementation Toolbox were Gather, Choose, Implement, and Prepare, as supported by bounce rates and average time on page. The most viewed phase was the "Gather" phase, with 3278 views. The most frequently downloaded tools within the Toolbox were Gather: Community Needs and Resource Assessment, with 136 downloads. The phases and tools accessed may have differed based on the user's goal or stage of implementation.

Conclusions: Findings indicate positive initial reach and engagement of the HNY website and HNY Implementation Toolbox among AI/AN educators that has consistently increased over the 2 years. The provision of web-based decision support that guides AI/AN users through the adoption, implementation, and maintenance of culturally relevant, age-appropriate, evidence-based

XSL•FO
RenderX

adolescent health promotion programs in their communities may help increase the implementation of effective adolescent health promotion programs to ultimately increase health equity among AI/AN youth.

(JMIR Form Res 2025;9:e67885) doi:10.2196/67885

KEYWORDS

implementation; culturally relevant program; evidence-based health promotion; user engagement; reach; decision support system; American Indian; Alaska; native communities; youth; adolescent; decision support; Alaska native; health inequities; sexual; reproductive; mental health; AI/AN; Tribal organization; Google Analytics; toolbox

Introduction

In the United States, as of 2022, the federal government recognizes 574 distinct American Indian and Alaska Native (AI/AN) Tribes across 37 states [1]. Together, the combined AI/AN population comprises 9.7 million people, representing 2.9% of the US population in 2020 [1]. The AI/AN population is young, with 30% being up to 18 years old, compared with 24% of the US total population [2], making adolescent health an important topic in native communities. However, AI/AN youth face many health inequities compared with other racial/ethnic groups, including a disproportionate burden of adverse mental, sexual, and reproductive health, and violence-related outcomes [3-5].

Evidence-based programs (EBPs) that promote adolescent health may help mitigate these inequities among AI/AN teens [6]. An increasing number of EBPs incorporating the strengths and cultural teachings of native communities have been developed or adapted for AI/AN youth [7-13]. Despite efforts to increase access to culturally relevant EBPs, barriers exist that limit their adoption and implementation by AI/AN educators [14-16]. Native communities may lack implementation readiness to discuss and approach adolescent health, experience difficulty navigating Tribal review and school board approval processes [16], have inadequate funding or program integration that compromises program sustainability, or be geographically isolated with limited access to programs [15,17]. Lack of administrative and parental support, lack of trained personnel, and lack of program integration with cultural values represent additional barriers to the implementation of EBPs in this population [15].

The Healthy Native Youth (HNY) website was developed in 2016 in collaboration with AI/AN community advisors to mitigate barriers to access and to enhance the adoption and implementation of culturally relevant adolescent health EBPs in native communities to promote health equity [6,14,18,19] (Figure 1). Digital platforms are an accepted delivery channel to enable greater reach of evidence-based interventions to underserved populations at lower costs [20-23]. Digital health interventions for AI/AN youth have been reported to contribute to enhanced health metrics such as mental health, substance use, and sexual risk reduction [10,24]. Prior to the HNY Website, such programs need to be located and accessed individually, without structured dissemination and implementation decision support. The HNY website provides a “one-stop-shop” for Tribal health advocates to access engaging, culturally relevant, age-appropriate health curricula and programs for AI/AN youth within a single website structured to provide implementation decision support [14]. The website comprises 5 sections: “Implementation Toolbox” (guidance for adoption and implementation of EBPs), “Curricula” (information about culturally informed EBPs), “For Caring Adults” (empowerment for nurturing adult-youth communication), “Community” (networking for advocates of adolescent health), and “Resources” (links and materials for youth advocates). The Curricula section of the HNY website contains comparative information on the content, specifications, and access to 19 culturally relevant, age-appropriate evidence-based for AI/AN youth focused on multiple topics, including mental health, sexual health, substance use, and suicide prevention. The user can filter and compare curricula on multiple dimensions to determine “best-fit” for their community and includes implementation materials at no cost.

The HNY Implementation Toolbox, housed within the HNY website is maintained and owned by the Northwest Portland Area Indian Health Board (NPAIHB) and provides stepped web-based decision support designed to facilitate the adoption and implementation of adolescent health EBPs in native communities (Figure 1). The Toolbox was developed on WordPress, an open-source content management system using an iterative design process informed by native practitioners and academicians to adapt an existing web-based decision support system (iCHAMPSS), which is described elsewhere [17,25]. Digital decision support can provide web-based expert guidance to facilitate complex tasks [26]. Findings from a 6-month user experience study with 25 AI/AN educators rated the Toolbox as credible, easy to use, and more helpful than current resources [17]. In the 6-month postsurvey, participants indicated increased knowledge, advocacy skills, and self-efficacy to adopt, implement, and maintain culturally relevant adolescent health promotion programs in their community [17]. The Toolbox launched in January 2022 and the sustainability of the program is through ongoing grant funding through the aforementioned NPAIHB. National Institutes of Health R21 funds (approximating US $440,000 total costs; US $164,000 indirect costs) were sufficient to develop and formatively assess the HNY Website Toolbox.

XSL•FO
RenderX

Figure 1. Roadmap of the Healthy Native Youth website landing page (top) and landing pages for each section (bottom) for access by American Indian and Alaska Native (AI/AN) communities.

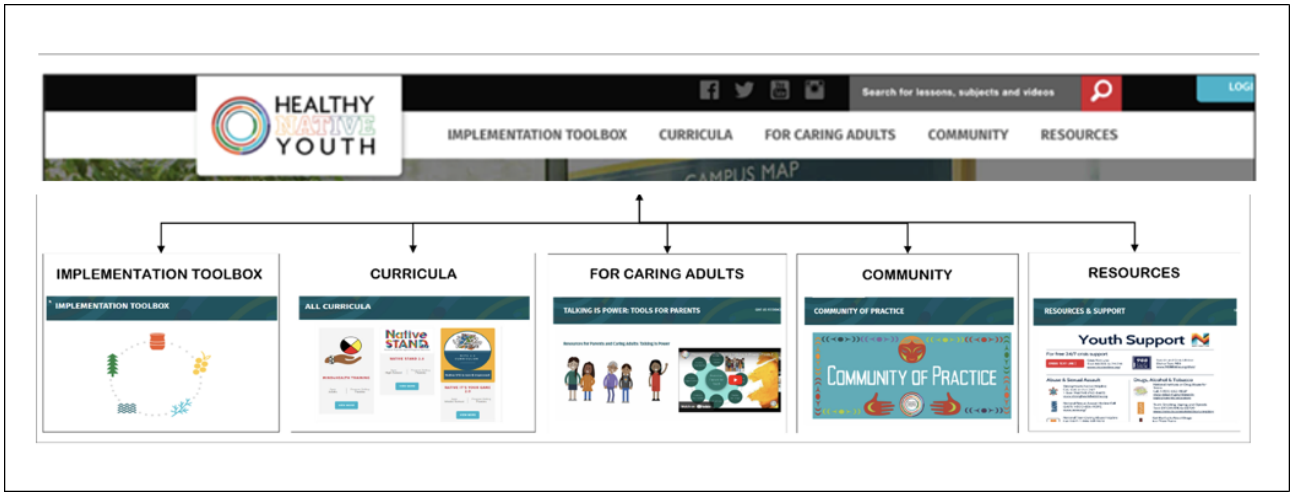

The Toolbox is designed to guide users through five culturally adapted dissemination and implementation phases: (1) Gather, (2) Choose, (3) Prepare, (4) Implement, and (5) Grow. (Figure 2) [17]. The "Gather" phase guides users to identify the health priorities and interests of youth in their community. "Choose" guides users to select a culturally relevant and age-appropriate evidence-based health program for their community. "Prepare" guides users to implement the selected program, including facilitating support from key decision makers. "Implement" guides users to deliver their selected program while emphasizing listening, reflection, and feedback. Finally, "Grow" guides users to sustain their selected program [17]. Each phase includes strategies to generate support for community members, elders, and youth throughout the planning process. The Toolbox comprises a "resource tools library" with over 20 phase-specific support tools, templates, helpful tips, links to web-based resources, and stories from the field (testimonials from experienced native health educators) [17]. All rights, including copyright, for this website and content on this Site are owned by or licensed to the NPAIHB.

The purpose of this study was to examine user reach and engagement with the HNY website and the embedded HNY Implementation Toolbox during an initial 2-year roll-out of the Toolbox from January 2022 to January 2024. Study findings of the reach of these web-based decision support strategies can inform further strategies to increase dissemination and implementation and provide a foundation for future hybrid effectiveness-implementation studies.

XSL•FO
RenderX

Figure 2. The 5 culturally adapted dissemination and implementation phases and associated tasks in the Healthy Native Youth Implementation Toolbox.

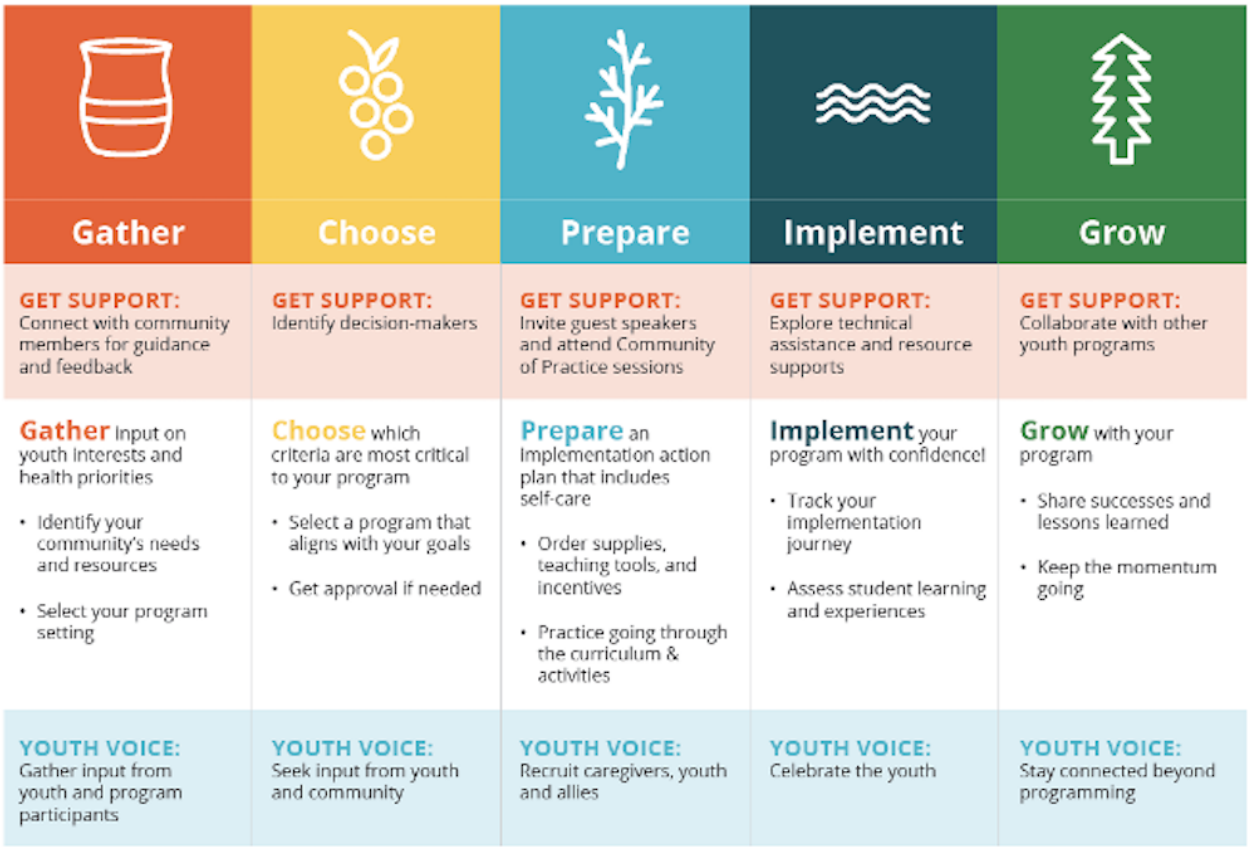

Methods

HNY Website and Implementation Toolbox Dissemination

Outreach dissemination strategies over the 2-year period encompassed channels through grassroots, community, and professional networks. These information channels comprised monthly AI/AN Community of Practice Network meetings highlighting the HNY Website and Toolkit for professionals who plan, manage, and facilitate healthy youth programs in Indian country, social media posts through existing networks in the AI/AN community devoted to health messages and event promotion, field studies with AI/AN practitioners and community members and related manuscript publication, presentations at professional conferences (eg, American Public Health Association, Department of Health and Human Services Office of Population Affairs Teen Pregnancy Prevention) and Tribal meetings (eg, The Tribal Public Health Conference) reporting formative, outcome, and process data from the HNY Toolbox, and presentations on varied health topics that promoted the HNY website as a resource. Rollouts were planned to ensure optimal breadth (geographically) and depth of exposure (including local Tribal organizations, opinion leaders, national Federal decision makers, and funders).

Data Extraction and Analysis

Data from the HNY website and the embedded HNY Implementation Toolbox were assessed using Google Analytics collected from January 2022 to January 2024. Google Analytics is a web analytics service that tracks and reports website traffic and user behavior [27,28]. This tool is used extensively by digital marketers and businesses to test and improve digital marketing performance, assess customer behavior, and identify strategies to optimize and improve traffic to a website by providing insight into user behavior.

Data were collected to assess reach, user demographics, and user engagement through a real-time, interactive dashboard accessible through a registered account on the Google Analytics website [29].

Reach is the extent to which an intervention or program is delivered to the priority population and was assessed as the number of users from each US state [29]. It is a necessary antecedent for community exposure to intervention content and any opportunity to establish effectiveness [29,30]. A user was defined as a visitor to the sites who had an assigned unique identifier with one or more associated sessions [29]. User demographics were assessed by geographic location, user gender and age, and the type of device used to access the sites. User engagement was assessed by the number of page views, number of pages viewed per session, number of "tools" downloaded, average duration of sessions, bounce rate, and exit rate. Page views were assessed as the number of separate views each page received. A session was defined as a user visit to, and interaction with, the webpage (eg, button clicks, downloads, and page navigation). Pages per session were assessed as the number of web pages of the site that the user viewed in a single session [29]. Duration of a session was assessed as the average length of time in minutes that users spent on a website/webpage during

XSL•FO
RenderX

a session [29]. Bounce rate, or nonengagement rate, was assessed as the percentage of single-page sessions a specific webpage received. If a user loaded a page but did not interact a second time (by viewing another page) within 30 minutes, this was considered a bounce [29]. For example, a bounce rate of 50% means that half the users engaged with the website on the visited page only. Exit rates were assessed as the percentage of users that left the website from a webpage (ie, the webpage was the last page visited by them before they exited the website) [29]. For example, an exit rate of 20% means that 20% of users left the website from this page. Reporting bounce rates and exit rates assess if the website/tool met the expectations of a user. If these are lower than industry standards then the site is considered to be engaging [29]. Downloads were assessed by user selections of a download option for a given resource template.

Ethical Considerations

This study was reviewed and approved by the Committee for the Protection of Human Subjects at the University of Texas Health Science Center at Houston (institutional review board number: HSC-SPH-18 - 0958). The HNY website also outlines its privacy policy in the footer section of the website as a means of informed consent from users and describes how user data are collected and may be used for research purposes. All data used are deidentified and anonymous. This manuscript, including related images and supplementary material, does not allow the identification of individual participants or users. No compensation was offered to participants as this paper reports analytics of a web-based resource that is accessible free of cost.

Results

HNY Website Analytics

From January 2022 to January 2024 the HNY website had 10,237 total users. Most users were located in Washington (n=1515), California (n=1290), and Oregon (n=1019) (Figure 3A). The majority of users identified as female (n=5217, 70.0%), were mainly 25 - 34 years (n=1676, 23.0%) or 18 - 24 (n=1559, 22.0%) years of age, and accessed the website primarily via desktop (80.0%) and mobile (19.0%) devices (Table 1). The most frequently accessed sections (in decreasing order) were Curricula (9902 views), Resources (6397 views), HNY Implementation Toolbox home (4364 views), Community (4054 views), and For Caring Adults (3766 views) (Table 2). Page duration ranged from 1.7 minutes (Implementation Toolbox) to 3.06 minutes (Curricula), bounce rate from 50.2% (Curricula) to 65.5% (Community), and exit rates from 15.9% (Curricula) to 60.0% (For Caring Adults). Access to the HNY website increased from 2479 page views in the first quarter of 2022 to a cumulative total of 33,004 page views by the fourth quarter of 2023 (Figure 4).

Figure 3. Geographic distribution (top 10 US states) and number of users by state who visited the HNY Website (left column) and HNY Implementation Toolbox (right column) from January 1, 2022, to January 1, 2024.

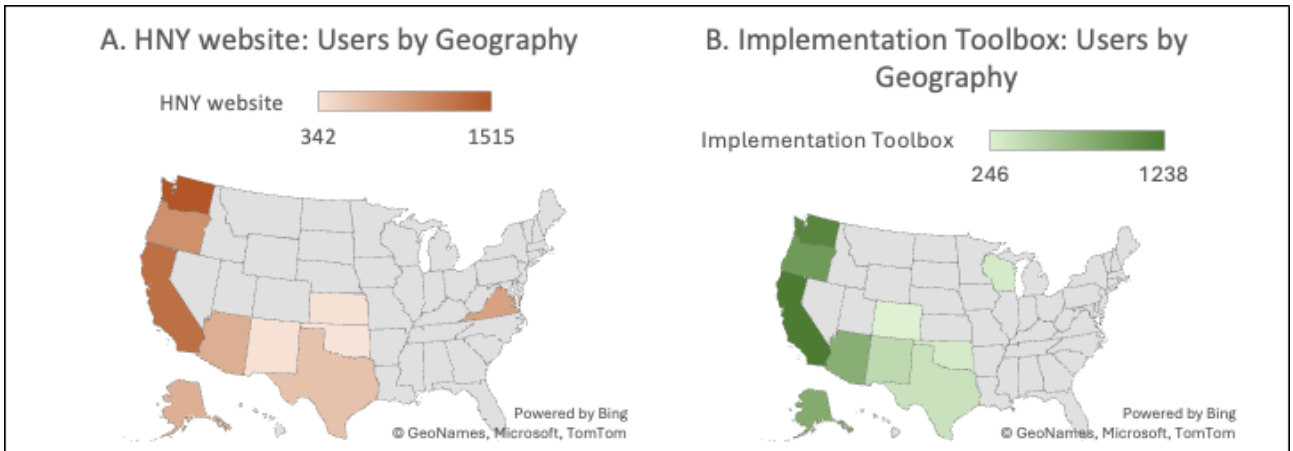

State of origin	User (n)
Washington	1515
California	1290
Oregon	1019
Virginia	884
Arizona	779
Alaska	764
Texas	617
New Mexico	366
Kansas	362
Oklahoma	342

State of origin	User (n)
California	1238
Washington	1142
Oregon	986
Alaska	847
Arizona	815
New Mexico	456
Texas	382
Oklahoma	307
Wisconsin	307
Colorado	246

XSL•FO
RenderX

Table . User demographics (gender, age, and device used) for the Healthy Native Youth (HNY) website (left column) and HNY Implementation Toolbox (right column) from January 2022 to January 2024.

	HNY website	HNY Implementation Toolbox
Users by gender, n (%)		
Male	2213 (30.0)	244 (34.0)
Female	5217 (70.0)	472 (66.0)
Users by age (years), n (%)		
18 - 24	1559 (22.0)	253 (20.0)
25 - 34	1676 (23.0)	319 (25.0)
35 - 44	1638 (23.0)	444 (35.0)
45 - 54	1216 (17.0)	135 (11.0)
55 - 64	690 (9.0)	82 (6.0)
65 and older	416 (6.0)	39 (3.0)
Users by device (%)[a]		
Mobile	19.0	19.0
Tablet	1.0	0
Desktop	80.0	81.0

[a] During the time period of this study, the number of users by device was not a part of the analytic dashboard. Therefore, only percentage values have been reported.

Table . Most frequently accessed sections in the Healthy Native Youth (HNY) website from January 2022 to January 2024 (Google Analytic data from January 2022 to January 2024).

Section[a]	Total users, n	Page views, n	Page views/session[b]	Page duration[b] (min)	Bounce rate[c], %	Exit rate, %
HNY website home	10,237	38,009	2.4	2.75	40.7	19.6
Curricula	3494	9902	3.8	3.06	50.2	15.9
Resources	2849	6397	2.51	2.83	64.6	40.3
Implementation Toolbox home	2566	4364	1.8	1.71	59.8	35.4
Community	2486	4054	1.8	1.3	57.4	60.0
For caring adults	1727	3766	3.64	2.0	65.5	36.6

[a]Corresponds to top-level sections in the HNY website (Figure 1).

[b]Average.

[c]Nonengagement rate.

XSL•FO
RenderX

Figure 4. Cumulative user frequency over time by quarter from Jan 2022 to Jan 2024 and associated dissemination strategies (outreach events). AI/AN: American Indian and Alaska Native; HNY: Healthy Native Youth; NA: not available; Q: quarter.

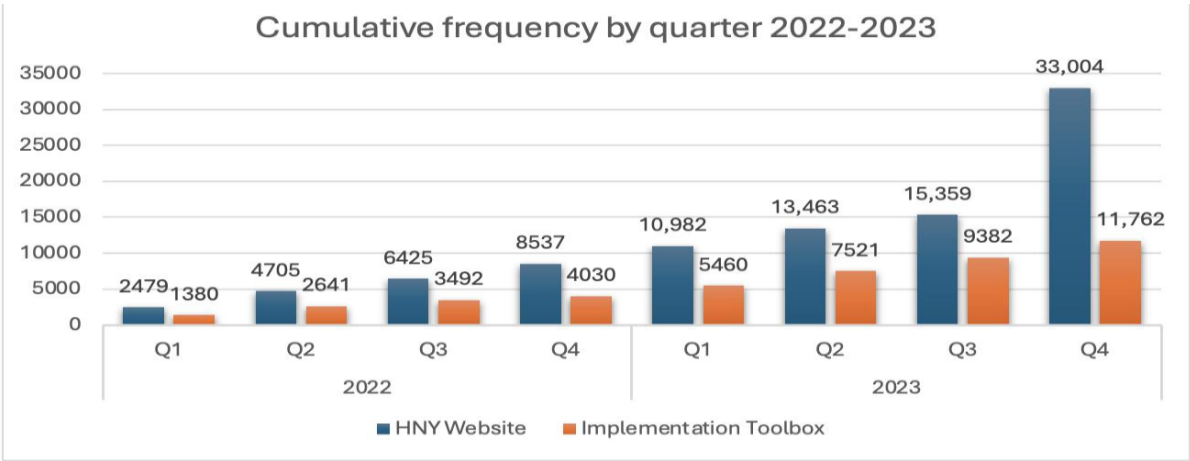

Year of evaluation	2022				2023			
Quarter	1	2	3	4	1	2	3	4
Months	Jan - Mar	Apr - Jun	Jul - Sept	Oct - Dec	Jan - Mar	Apr - Jun	Jul - Sept	Oct - Dec
Website page views (n)	2479	2226	1918	2112	2395	2,531	1,896	17,645
Website page views cumulative	2479	4705	6425	8537	10,932	13,463	15,359	33,004
Toolbox page views	1380	1261	851	538	1430	2061	1861	2380
Toolbox page views cumulative	1380	2641	3492	4030	5460	7521	9382	11,762
Dissemination strategy								
AI/AN Community of Practice Network Meetings (Date, Attendees)	NA	NA	NA	NA	NA	NA	Sept. (n=5)	Nov. (n=30) Oct. (n=13)
Social media (Instagram) posts.	14	4	5	3	6	6	3	6
HNY presentations promoting the HNY Website and HNY Toolbox at professional meetings and Tribal gatherings	1	0	4	4	5	3	0	1
Field study - Formative and usability studies.	x		x	x				
Publication				x				

HNY Implementation Toolbox Analytics

From January 2022 to January 2024 the Implementation Toolbox had a total of 2566 total users (Table 3) and 11,762 page views (Figure 4). Most users were located in California (n=1238), Washington (n=1142), and Oregon (n=986) (Figure 3B). The majority of users identified as female (n=472, 66%), were mainly 35 - 44 years (n=444, 35%) or 25 - 34 (n=319, 25%) years of age and accessed the Toolbox mainly via desktop (81%) and mobile (19%) devices (Table 1). Phase page views (in decreasing order) were Gather (3278 views), Choose (1275 views), Prepare (907 views), Implement (843), and Grow (618 views) (Table 3). Page duration ranged from 1.67 minutes (Grow) to 2.08 minutes (Choose), bounce rate from 41.8% (Gather) to 72.9% (Prepare), and exit rates from 22.1% (Prepare) to 34.5% (Grow). The most frequently downloaded tools within the Toolbox were from the Gather phase and included the "Community needs and resource assessment" factsheet (136 views), the "Identify youth advocates and community partners" factsheet (111 views), and a "Bingo data collection" activity guide (83 views) (Table 4). Access to the HNY Implementation Toolbox increased from 1380 page views in the first quarter of 2022 to a cumulative total of 11,762 page views by the fourth quarter of 2023 in conjunction with dissemination activities of scientific presentations, community of practice sessions, social media posts, and webinars (Figure 4).

XSL•FO
RenderX

Table . Phases of the Healthy Native Youth (HNY) Implementation Toolbox accessed from January 2022 to January 2024 across the United States by user metrics.

Implementation phase	Total users, n	Page views, n	Page views/session[a]	Page duration[a] (min)	Bounce rate, %	Exit rate, %
Implementation Toolbox landing page	2566	4364	1.8	1.71	60.3	35.5
Gather	1863	3278	16.7	2.07	41.8	33.9
Choose	876	1275	12.3	2.08	49.5	22.3
Prepare	575	907	27.1	1.40	72.9	22.1
Implement	488	843	7.8	1.7	64.0	30.8
Grow	353	618	5.7	1.66	57.8	34.5

[a]Average.

Table . Top 10 most frequently downloaded tools from the Healthy Native Youth (HNY) Implementation Toolbox by phase, type, and name from January 2022 to January 2024 across the United States.

#	Implementation phase	Tool type	Tool name	Downloads, n
1	Gather	Factsheet	Community needs and resource assessment	136
2	Gather	Factsheet	Identify youth advocates and community partners	111
3	Gather	Activity sheet	Bingo data collection activity	83
4	Prepare	Template	Curriculum implementation plan template	76
5	Gather	Checklist	Site selection	72
6	Implement	Log	Session reflection log template	44
7	Prepare	Inventory list	Example list of materials to order for native STAND	40
8	Implement	Log	Class attendance sheet template	30
9	Choose	Factsheet	Identify youth advocates and community partners	29
10	Gather	Factsheet	Alaska Tribal Health Consortium adolescent health action plan	28

Discussion

Principal Findings

The HNY website and embedded HNY Implementation Toolbox provide web-based resources to support the adoption, implementation, and maintenance of culturally relevant, age-appropriate evidence-based adolescent health promotion programs in AI/AN communities. The HNY website is a “one-stop-shop” for Tribal health advocates to access culturally relevant and age-appropriate health curricula, programs, and related resources for AI/AN youth [14]. The embedded HNY Implementation Toolbox provides guidance to enable them to successfully implement these curricula in their communities. Over the 2-year evaluation period since the Toolbox’s “go live” date, approximately 1 in 8 users of the HNY website visited the Toolbox. This, and expanded geographic reach, corroborate the previously reported utility of the Toolbox [17]. The differential traffic and reach between the HNY website and the Toolbox may reflect the comparative maturity of the HNY website compared with the relatively newer embedded Toolbox. The greater reach in the western states is expected because the HNY website was developed and is housed in the NPAIHB located in Washington State, and early promotion was among affiliated Tribes in Washington State, Oregon, and Idaho. Demographically, the Toolbox users were older, with 35% (n=444) being 35 - 44 years of age. This is likely a reflection of the demographics of community health advocates who are evaluating community needs, facilitating decision-making on the curricula best addressing these needs, and actively engaging community support to adopt and implement these curricula in their communities. This, and whether users had attended Tribal

XSL•FO
RenderX

health conferences or Community of Practice gatherings in which the Toolbox was promoted could not be verified from the Google data.

Over the evaluation period, page views of the HNY Website and the HNY Toolbox increased 10-fold and 9-fold, respectively. Page views to the Toolbox witnessed a steady increase from 2022 to 2023. The reason for this is uncertain but is possibly attributable to the Toolbox being showcased in monthly Community of Practice meetings that gather dozens of advocates for Tribal youth in Indian country. This suggests the potential for embedded practitioner networks as a particularly utilitarian strategy to disseminate youth programs [27,28]. Given that the HNY website and Implementation Toolbox coexist, the promotion of one likely promotes traffic to the other. This is the strength of a collaborative approach that provides multiple resources that address a particular theme, whether it be a priority population, setting, or health domain. The cumulative page views over time are also suggestive of the need to provide a persistent and longitudinal promotional strategy. We will continue to build awareness nationally via presentations, workshops, and round table discussions at Tribal, education, and public health conferences.

The Toolbox's average bounce rate (57.7%) was lower than that of the HNY website (62.1%) and also the industry average (62.3%); the average number of Toolbox pages visited per session was almost 5 times than that of the industry average (11.9 vs 2 pages) [14,31,32]. Further, the duration of use on both sites (2.4 and 1.8 respectively) are close to averages on comparable websites (2.58) [33]. These findings suggest that the Toolbox is "sticky," and that users are finding material commensurate with their needs. This is also supportive of previously reported positive usability ratings for the site from AI/AN community decision makers.

User preferences within the Toolbox were for the Gather Phase and tools related to this phase that enable needs assessment and identification of community advocates. This may be indicative of Tribal communities being in this early phase of adopting and implementing adolescent health programs, that these tools are particularly useful for the work of health educators, or simply that community members are likely to start at the first step in their review and exploration of a novel support tool like the Toolbox. Future studies can investigate if progression through stages occurs over time, the association of phase use with "in-situ" community implementation, as well as challenges to implementation that can be a function of differing state or local policies, regional internet accessibility, funding, community readiness to adopt programs, and logistics of obtaining school board approvals [16].

Study findings need to be interpreted in the context of several methodologic limitations and biases due to data collection. Google Analytics was originally intended for marketing and sales purposes and is limited largely to descriptive data [34,35]. Guidelines for the use and interpretation of Google Analytics do not exist. Discrepancies in the data are possible due to a new client ID being assigned every time a user deletes cookies or signs in from a different device or browser [36]. This may produce an inaccurate count of "new" users because different IDs could be allocated to the same user [34]. Additionally, deletion of cookies may degrade data captured via Google Analytics which relies on user cooperation for successful data collection [20]. Google Analytics assigns a value of 0 seconds if a user visits a webpage until they navigate to a second page, which can lead to inaccurate page duration assessments [34]. Google Analytics does not collect demographic and contextual data on ethnicity, Tribe, or the context of the application of the HNY website and Toolbox. The gender of users captured by Google Analytics may only reflect a subset of users that log into their Google accounts when visiting the page and it is difficult to determine how the demographics align with the target population [34]. Page views and resource downloads do not necessarily equate to the actual adoption and implementation of programs. There is a risk of selection bias wherein the analytics reported may only reflect a subset of users that log into their Google accounts The purpose of use, whether in the context of seeking guidance on program implementation, review and feedback for research, or simply to assuage curiosity, was not ascertained in this study and will need to be if the utility of the HNY Website and Toolbox is to be fully understood. To address the above biases and limitations, reporting using Google Analytics may be applied in combination with other field data collection (qualitative or quantitative) to triangulate data sources. This study reported on an array of promotional strategies to achieve increased dissemination. Future studies, perhaps using a mixed methods design, are recommended to investigate their comparative effectiveness.

Despite limitations, the use of Google Analytics to assess and gather feedback on web-based decision support and to track user behavior on multiple devices is well accepted [20]. Advantages are that it can be used for self-comparison over time and provide a comparison to industry standards. This can provide a positive feedback loop wherein higher website traffic builds positive normative perceptions that generate further user traffic through "word of mouth." An assessment of reach represents an early phase in understanding program dissemination and implementation that provides the foundation for future hybrid effectiveness-implementation evaluation, including network analysis to assess the dynamics of community diffusion.

Conclusions

The HNY website and HNY Implementation Toolbox provide AI/AN educators and youth advocates with step-by-step guidance to support the adoption, implementation, and maintenance of culturally relevant, age-appropriate evidence-based adolescent health promotion programs in their communities. Google Analytics provides insight into the use and reach of the HNY website and Implementation Toolbox among AI/AN youth advocates, which will help inform ongoing dissemination efforts. Future studies are needed to assess the Toolbox's actual impact on the uptake and implementation of adolescent health promotion EBPs in native communities.

XSL•FO
RenderX

Acknowledgments

Funding for this study was received from the National Institutes of Health Office of Minority Health (1R21MD01396001A1) and the Indian Health Service HIV Program and the Minority AIDS Initiative Fund (contract number 248-96-0011).

Data Availability

The datasets generated or analyzed during this study are not publicly available because the data could contain identifiable information linked to the users, but are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

References

1. Federally recognized American Indian tribes and Alaska Native entities. USA.gov. URL: https://www.usa.gov/indian-tribes-alaska-native [accessed 2025-04-02]
2. Facts for features: American Indian and Alaska Native heritage month: November 2021. US Census Bureau. 2022. URL: https://www.census.gov/newsroom/facts-for-features/2021/aian-month.html [accessed 2022-02-27]
3. Trends in teen pregnancy and childbearing. Office of Population Affairs. URL: https://opa.hhs.gov/adolescent-health/reproductive-health-and-teen-pregnancy/trends-teen-pregnancy-and-childbearing [accessed 2024-03-17]
4. Allen J, Wexler L, Rasmus S. Protective factors as a unifying framework for strength-based intervention and culturally responsive American Indian and Alaska Native suicide prevention. Prev Sci 2022 Jan;23(1):59-72. [doi: 10.1007/s11121-021-01265-0] [Medline: 34169406]
5. Clayton HB, Kilmer G, DeGue S, et al. Dating violence, sexual violence, and bullying victimization among high school students -youth risk behavior survey, United States, 2021. MMWR Suppl 2023 Apr 28;72(1):66-74. [doi: 10.15585/mmwr.su7201a8] [Medline: 37104527]
6. Morrison-Beedy D, Mazurek Melnyk B. Making a case for integrating evidence-based sexual risk reduction and mental health interventions for adolescent girls. Issues Ment Health Nurs 2019 Nov 2;40(11):932-941. [doi: 10.1080/01612840.2019.1639087]
7. Spencer LM, Schooley MW, Anderson LA, et al. Seeking best practices: a conceptual framework for planning and improving evidence-based practices. Prev Chronic Dis 2013 Dec 12;10:E207. [doi: 10.5888/pcd10.130186] [Medline: 24331280]
8. Craig Rushing S, Gaston A, Kaufman C, Markham C, Jessen C, Gorman G, et al. Using technology to promote health and wellbeing among American Indian and Alaska Native teens and young adults. In: Dyson LE, Grant S, Hendricks M, editors. Indigenous People and Mobile Technologies: Taylor & Francis Group; 2016:163-178.
9. Tingey L, Chambers R, Patel H, et al. Prevention of sexually transmitted diseases and pregnancy prevention among Native American youths: a randomized controlled trial, 2016–2018. Am J Public Health 2021 Oct;111(10):1874-1884. [doi: 10.2105/AJPH.2021.306447]
10. Shegog R, Craig Rushing S PhD, MPH, Jessen C MA, et al. Native IYG: improving psychosocial protective factors for HIV/STI and teen pregnancy prevention among youth in American Indian/Alaska Native communities. J Appl Res Child 2017;8(1). [doi: 10.58464/2155-5834.1320]
11. Hafner SP, Craig Rushing S. Sexual health, STI and HIV risk, and risk perceptions among American Indian and Alaska Native emerging adults. Prev Sci 2019 Apr;20(3):331-341. [doi: 10.1007/s11121-018-0920-7] [Medline: 30006906]
12. Skye M, McCoy T, Kelley A, et al. Effectiveness of Native STAND: a five-year study of a culturally relevant sexual health intervention. Adolescents 2021;1(3):321-334. [doi: 10.3390/adolescents1030024]
13. Kaufman CE, Whitesell NR, Keane EM, et al. Effectiveness of circle of life, an HIV-preventive intervention for American Indian middle school youths: a group randomized trial in a Northern Plains tribe. Am J Public Health 2014 Jun;104(6):e106-e112. [doi: 10.2105/AJPH.2013.301822]
14. Craig Rushing S, Stephens D, Shegog R, et al. Healthy Native Youth: improving access to effective, culturally-relevant sexual health curricula. Front Public Health 2018;6:225. [doi: 10.3389/fpubh.2018.00225]
15. Sacca L, Shegog R, Hernandez B, et al. Barriers, frameworks, and mitigating strategies influencing the dissemination and implementation of health promotion interventions in indigenous communities: a scoping review. Implement Sci 2022 Feb 21;17(1):18. [doi: 10.1186/s13012-022-01190-y] [Medline: 35189904]
16. Markham C, Torres J, Rushing SC, Gorman G, Jessen C, Gaston A, et al. Usability and psychosocial impact of decision support to increase sexual health education in American Indian and Alaska Native communities. J Health Dispar Res Pract 2018;11(1) [FREE Full text]
17. Markham CM, Rushing SC, Manthei J, et al. The Healthy Native Youth Implementation Toolbox: using implementation mapping to adapt an online decision support system to promote culturally-relevant sexual health education for American Indian and Alaska Native youth. Front Public Health 2022;10:889924. [doi: 10.3389/fpubh.2022.889924] [Medline: 36388328]

XSL•FO
RenderX

18. Peskin MF, Hernandez BF, Markham C, et al. Sexual health education from the perspective of school staff: implications for adoption and implementation of effective programs in middle school. J Appl Res Child 2011;2(2). [doi: 10.58464/2155-5834.1046]
19. Hernandez BF, Peskin M, Shegog R, et al. Choosing and maintaining programs for sex education in schools: the CHAMPSS model. J Appl Res Child 2011;2(2). [doi: 10.58464/2155-5834.1047]
20. Marcu G, Ondersma SJ, Spiller AN, Broderick BM, Kadri R, Buis LR. Barriers and considerations in the design and implementation of digital behavioral interventions: qualitative analysis. J Med Internet Res 2022 Mar 30;24(3):e34301. [doi: 10.2196/34301] [Medline: 35353043]
21. Marcu G, Ondersma SJ, Spiller AN, Broderick BM, Kadri R, Buis LR. The perceived benefits of digital interventions for behavioral health: qualitative interview study. J Med Internet Res 2022 Mar 30;24(3):e34300. [doi: 10.2196/34300] [Medline: 35353045]
22. Linardon J, Cuijpers P, Carlbring P, Messer M, Fuller-Tyszkiewicz M. The efficacy of app-supported smartphone interventions for mental health problems: a meta-analysis of randomized controlled trials. World Psychiatry 2019 Oct;18(3):325-336. [doi: 10.1002/wps.20673] [Medline: 31496095]
23. Nelson LA, Spieker AJ, Mayberry LS, McNaughton C, Greevy RA. Estimating the impact of engagement with digital health interventions on patient outcomes in randomized trials. J Am Med Inform Assoc 2021 Dec 28;29(1):128-136. [doi: 10.1093/jamia/ocab254] [Medline: 34963143]
24. Toombs E, Kowatch KR, Dalicandro L, McConkey S, Hopkins C, Mushquash CJ. A systematic review of electronic mental health interventions for Indigenous youth: results and recommendations. J Telemed Telecare 2021 Oct;27(9):539-552. [doi: 10.1177/1357633X19899231] [Medline: 31937199]
25. Peskin MF, Hernandez BF, Gabay EK, et al. Using intervention mapping for program design and production of iCHAMPSS: an online decision support system to increase adoption, implementation, and maintenance of evidence-based sexual health programs. Front Public Health 2017;5:203. [doi: 10.3389/fpubh.2017.00203] [Medline: 28848729]
26. Bhargava HK, Power DJ, Sun D. Progress in web-based decision support technologies. Decis Support Syst 2007 Aug;43(4):1083-1095. [doi: 10.1016/j.dss.2005.07.002]
27. Noar AP, Jeffery HE, Subbiah Ponniah H, Jaffer U. The aims and effectiveness of communities of practice in healthcare: a systematic review. PLoS One 2023;18(10):e0292343. [doi: 10.1371/journal.pone.0292343] [Medline: 37815986]
28. Bodison SC, Sankaré I, Anaya H, et al. Engaging the community in the dissemination, implementation, and improvement of health-related research. Clin Transl Sci 2015 Dec;8(6):814-819. [doi: 10.1111/cts.12342] [Medline: 26546337]
29. Balasubramanian BA, Fernald D, Dickinson LM, et al. REACH of interventions integrating primary care and behavioral health. J Am Board Fam Med 2015;28 Suppl 1(Suppl 1):S73-S85. [doi: 10.3122/jabfm.2015.S1.150055] [Medline: 26359475]
30. Reilly KL, Kennedy S, Porter G, Estabrooks P. Comparing, contrasting, and integrating dissemination and implementation outcomes included in the RE-AIM and implementation outcomes frameworks. Front Public Health 2020;8:430. [doi: 10.3389/fpubh.2020.00430] [Medline: 32984239]
31. Average bounce rate by industry 2024 [benchmark]. Capturly. URL: https://capturly.com/blog/average-bounce-rate-by-industry-2023-benchmark [accessed 2024-07-22]
32. Pages per session: a measure of engagement. Analyzify. URL: https://analyzify.com/hub/pages-per-session-a-measure-of-engagement [accessed 2024-07-22]
33. Average session duration: industry benchmarks. First Page Sage. URL: https://firstpagesage.com/reports/average-session-duration-by-industry [accessed 2024-07-22]
34. Analytics tools & solutions for your business. Google Marketing Platform. URL: https://marketingplatform.google.com/about/analytics [accessed 2024-07-22]
35. Henström M, Duncanson K, Collins CE, Ashton LM, Davidson E, Ball R. Online reach and engagement of a child nutrition peer-education program (PICNIC): insights from social media and web analytics. BMC Public Health 2022 Apr 26;22(1):836. [doi: 10.1186/s12889-022-13252-3] [Medline: 35473610]
36. Client ID vs user ID. Analytics Mania. URL: https://www.analyticsmania.com/post/google-analytics-client-id-vs-user-id [accessed 2024-07-22]

Abbreviations

AI/AN: American Indian and Alaska Native
EBP: evidence-based program
HNY: Healthy Native Youth
NPAIHB: Northwest Portland Area Indian Health Board

XSL•FO
RenderX

Edited by A Mavragani; submitted 23.10.24; peer-reviewed by TF Ojo; revised version received 08.01.25; accepted 09.01.25; published 16.04.25.

Please cite as:
Sidhu A, Shegog R, Craig-Rushing S, Trevino N, Singer M, Jessen C, Gorman G, Simpson S, Peskin M, Hernandez B, Markham C Using the Healthy Native Youth Implementation Toolbox to Provide Web-Based Adolescent Health Promotion Decision Support to American Indian and Alaska Native Communities: Implementation Study
JMIR Form Res 2025;9:e67885
URL: https://formative.jmir.org/2025/1/e67885
doi:10.2196/67885

© Amrita Sidhu, Ross Shegog, Stephanie Craig-Rushing, Nicole Trevino, Michelle Singer, Cornelia Jessen, Gwenda Gorman, Sean Simpson, Melissa Peskin, Belinda Hernandez, Christine Markham. Originally published in JMIR Formative Research (https://formative.jmir.org), 16.4.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Evaluation of a Curriculum-Based Nutrition Education Intervention Protocol in Elementary Schools: Nonrandomized Feasibility Study

Jacqueline Marie Brown[1], MHSc; Nicholas Rita[2], MEd; Beatriz Franco-Arellano[1], PhD; Ann LeSage[2], PhD; Joanne Arcand[1], PhD, RD

[1]Faculty of Health Sciences, Ontario Tech University, 2000 Simcoe Street North, Oshawa, ON, Canada

[2]Mitch and Leslie Frazer Faculty of Education, Ontario Tech University, Oshawa, ON, Canada

Corresponding Author:
Joanne Arcand, PhD, RD
Faculty of Health Sciences, Ontario Tech University, 2000 Simcoe Street North, Oshawa, ON, Canada

Abstract

Background: Improving children's food literacy through school-based interventions can support developing healthy eating habits. However, teachers lack appropriate resources, time, and training to provide nutrition education in schools. Serious games, which are games designed for a purpose other than entertainment, have been demonstrated to improve children's food literacy and dietary intake and can address the barriers teachers face in providing nutrition education. Foodbot Factory (Arcand Lab) is a nutrition education intervention that is aligned with curricula and uses a serious game to provide nutrition education to students. Further evidence is needed to understand how serious games, including Foodbot Factory, can be researched in schools to support nutrition education.

Objective: The objective of this study was to evaluate the feasibility of a research study protocol that implements the curriculum-based nutrition education intervention Foodbot Factory into a real-world classroom setting. The evaluation of the protocol included study processes, resources, and management feasibility outcomes, as well as a preliminary assessment of scientific outcomes relevant to the intervention.

Methods: A nonrandomized study determined the feasibility of intervention implementation. Grade 4 and 4/5 classrooms were assigned to have nutrition education lessons for 5 days with either the Foodbot Factory or a control intervention. Outcomes were assessed in 4 feasibility domains of study processes (eg, recruitment and attrition rates), resources (eg, time taken to deliver the intervention), and management (eg, challenges with intervention delivery), and a preliminary assessment of scientific outcomes pertaining to the acceptability and impacts of the interventions. These outcomes were captured in semistructured field notes completed by study staff and a Nutrition Attitudes and Knowledge questionnaire and acceptability questionnaire completed by participants. Data were analyzed descriptively and using a paired *t* test to assess within-group changes in nutrition knowledge.

Results: In total, 4 classrooms participated in the feasibility study, with varying recruitment rates for schools (3/20, 15%), classrooms (4/4, 100%), parents (54/102, 53%), and children (49/54, 91%). The time required to implement the research protocol, including data collection and lesson plans, was sufficient and management of the intervention implementation was overall successful. Some challenges were experienced with classroom management during data collection, specifically with electronic data collection. After the intervention, participants reported a positive affective experience (26/41, 63%) and learning something new about healthy eating (31/41, 76%). Participants in both study groups improved their nutrition knowledge, but the changes were not statistically significant. The Foodbot Factory group had a statistically significant improvement in their knowledge of vegetables and fruit (P=.04) and protein foods (P=.03).

Conclusions: These findings indicate that the study protocol is feasible to implement and evaluate Foodbot Factory in a representative sample with select modifications to improve recruitment and data collection procedures.

(JMIR Form Res 2025;9:e69242) doi:10.2196/69242

KEYWORDS

nutrition education; serious games; children; food literacy; school nutrition intervention; feasibility

Introduction

One way to support children in acquiring and maintaining healthy eating patterns is by developing their food literacy. Food literacy describes the set of interrelated food and nutrition knowledge (ie, understanding food groups and the nutrients in foods), food skills (ie, ability to prepare foods and read recipes), and attributes (eg, self-efficacy and confidence) that interact with our broader socioecological environment to shape our dietary behaviors [1]. Research has demonstrated that children

XSL•FO
RenderX

and adolescents with higher levels of nutrition knowledge, a core component of food literacy, are more likely to have a higher quality dietary pattern [2,3]. Improving food literacy is especially relevant in Canada where the average child exceeds recommended intakes for saturated fat, sodium and free sugars, and consumes 21% - 25% of their daily caloric intake from foods that are not recommended by dietary guidelines [4,5]. This dietary pattern can increase the future risk of noncommunicable diseases such as cardiovascular disease and type 2 diabetes [6].

School curriculum, policies, and programming are an important way to support the development of child food literacy and healthy eating behaviors. All jurisdictions in Canada have food and nutrition as a core component of elementary health curricula [7], however, teachers face several barriers to implementing nutrition education in their classrooms due to a lack of training and dedicated time and resources for nutrition [8-10]. Published literature shows that well-designed curriculum-based interventions can effectively improve nutrition knowledge and behaviors [11]. A meta-analysis of school- and curriculum-based nutrition education interventions found that experiential learning approaches, such as school gardens and cooking classes, had the greatest impact on child nutrition knowledge and dietary intake [12]. While these experiential teaching approaches enhance the student learning experience, they are resource- and time-intensive and fail to address the needs of teachers. Thus, alternative experiential learning approaches are needed that can effectively support children's food literacy development and address the teacher-reported barriers to providing nutrition education.

Technology-based nutrition interventions can address some of these teacher-reported barriers, due to high access and ease of use in the classroom [13]. Serious games, which are games designed for a primary reason other than entertainment, have emerged as a leading technology-based educational platform as they use an experiential learning approach [14]. Research on nutrition-focused serious games has found that they can improve vegetable and fruit intake and nutrition knowledge among children [15-17]. Unfortunately, these nutrition-focused games are often not available to the public [18], do not consistently align with relevant nutrition curriculum, and are underinvestigated in classrooms as a resource to support curriculum implementation. Research and resources are required to support teachers in providing nutrition education and create opportunities for children to develop their nutrition knowledge and food literacy in the classroom.

The Foodbot Factory intervention was developed by an interdisciplinary team of dietitians, game developers, teachers, and researchers, to support elementary teachers with curriculum-based nutrition education and to improve children's nutrition knowledge [19]. The intervention includes a serious game for students played on a mobile app, with lesson plans for teachers. The content is designed for children ages 8 - 12 years (Grades 4 and 5) and aligns with the 2019 Ontario Health and Physical Education curriculum and Canada's Food Guide [20,21]. Our previous proof-of-concept research demonstrated that the Foodbot Factory serious game significantly improved children's nutrition knowledge compared with a control food-themed game [22]. However, this study was conducted in a controlled setting, not in the intended classroom environment. The lesson plans developed have also not yet been tested in classrooms, which is critical to understand their implementation [23]. Before embarking on a larger randomized trial to evaluate the impact of Foodbot Factory in classrooms, a setting that comes with unique implementation challenges, a feasibility assessment of an intended research protocol is warranted. Therefore, the objective of this research was to determine the feasibility of a research protocol to evaluate the Foodbot Factory intervention and a control intervention with children in the classroom setting. Study processes, resources, management, and scientific outcomes were the feasibility elements assessed. Such data will inform future research protocols to evaluate the efficacy of Foodbot Factory as part of a randomized trial.

Methods

Study Design

This was a nonrandomized study to assess the feasibility of implementing a trial protocol for a nutrition education intervention (Foodbot Factory) and a control intervention among Grade 4, 4/5, and 5 elementary school classrooms, over a 5-day period. Classrooms were assigned in a 1:1 ratio to one of two groups: (1) the Foodbot Factory group receiving nutrition education using the Foodbot Factory serious game and lesson plans or (2) the control group receiving nutrition education using nontechnology based learning activities. This study protocol was cocreated with several school board partners. This approach increased the feasibility of the study protocol from the school board's perspective, particularly in relation to the length of time required for the study and the use of a study teacher to provide the intervention (described in the Interventions: Overview section).

The primary objective was to assess the feasibility, or ability to be successful in implementing the study protocol as planned using feasibility outcomes that are evaluated across 4 domains, that are study processes, resources, management, and a preliminary assessment of scientific outcomes [24]. The specific outcomes within each domain were based on established guidelines for conducting feasibility research [24]. Study process, resource, and management outcomes were documented throughout the study period. The preliminary assessment of scientific outcomes was assessed by having participants complete questionnaires on their nutrition knowledge (day 1 and day 5) and intervention acceptability (day 5; Figure 1).

XSL•FO
RenderX

Figure 1. Feasibility study overview.

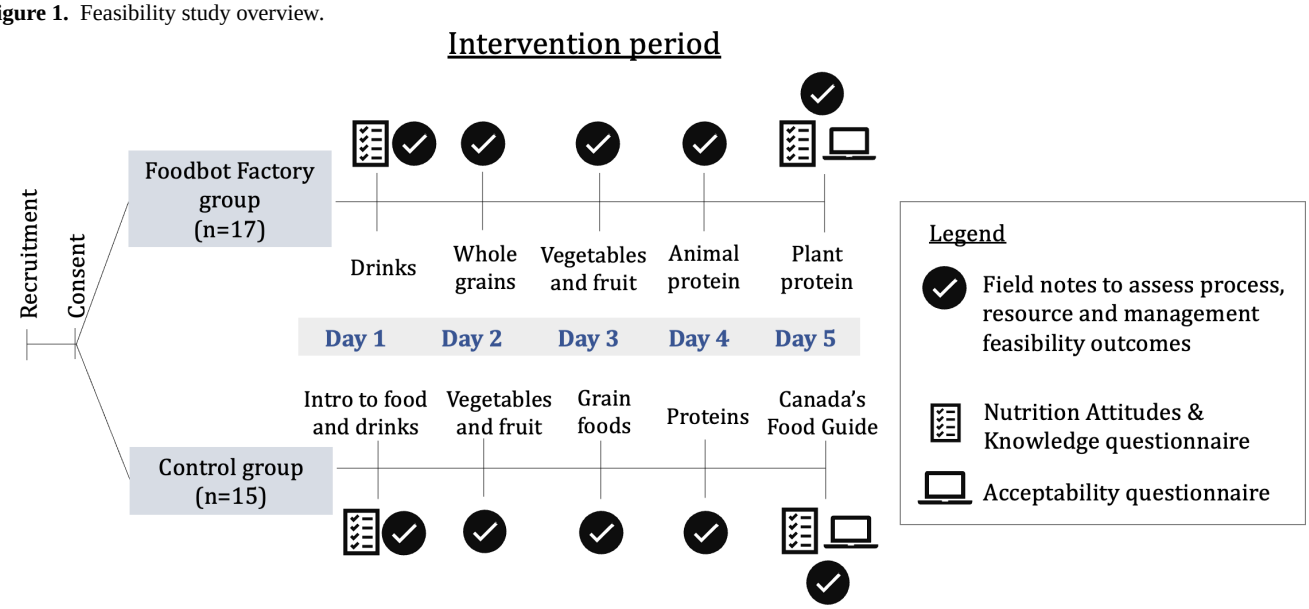

Ethical Considerations

This study was approved by the Ontario Tech University Research Ethics Board (#16930) and by the participating school board. Informed consent (Multimedia Appendix 1) was obtained from participants' parents or guardians, along with informed assent from child participants. All study data are in a deidentified format. Classroom teachers received a $50 CAD (US $37.50) gift card as an acknowledgment for having their classroom participate in the research. All children in the classroom (regardless of participation in the study) received an Ontario Tech University water bottle as a token of appreciation.

Participants

Grade 4, 4/5, and 5 classrooms in a Greater Toronto Area school board were eligible to participate if they had not yet covered the healthy eating component of the curriculum. Furthermore, 4 classrooms from 2 elementary schools participated in the study. The research team randomly selected schools and contacted them via email to gauge interest in participation. Principals who expressed interest subsequently invited classroom teachers to participate in the study. In each school we recruited 2 classrooms, assigning 1 to the Foodbot Factory group and the other to the control group. Classrooms, rather than individuals, were assigned to the group as the intervention is designed for classrooms and it is not logistically feasible to assign individuals within the same classroom to different study groups. In the first school, 1 classroom was randomly assigned to their group and the second classroom was assigned to the opposite group. In the second school, classrooms were assigned to the study groups based on enrollment to ensure a comparable number of individual participants in each group. Before the study, classroom teachers emailed parents the web-based consent form. Child assent was obtained on day 1 of the study (Monday) from those with parental consent. For participants who were absent on day 1, assent was collected on day 5 for the purposes of collecting data on intervention acceptability only. All children in a classroom participated in the intervention, but only those with parental consent participated in data collection and analysis.

Interventions

Overview

Both interventions were 5 consecutive days in duration and classrooms received nutrition education for 35 - 40 minutes per day [23]. The interventions were provided by a hired certified teacher, who was part of the research team ("study teacher"). Due to the nature of their role, the study teacher was not blinded to the intervention group. The study teacher was trained to follow standardized operating procedures for the delivery of the interventions to reduce sources of bias and ensure a high level of fidelity. The study teacher was accompanied by an observer who was trained in the same procedures. The research team originally estimated that data collection would take a maximum of 20 minutes, and that lesson plan delivery would take 40 minutes. Classroom teachers were not involved in teaching the lessons, but they were present in the classroom during the study and supported classroom management.

Foodbot Factory Intervention

Classrooms in the Foodbot Factory group received nutrition education using the Foodbot Factory serious game and corresponding curriculum-based lesson plans [19,25]. The serious game is based on experiential learning theory [26], while the lesson plans are rooted in constructivist learning theory, to help children connect knowledge learned in the game to their lived experiences [27]. The Foodbot Factory intervention consists of 5 nutrition education lessons, with additional instructions for teachers on the importance of using food-neutral language, suggestions to incorporate cultural foods into each lesson and modifications for different learning needs. Each lesson follows the 3-phase lesson structure, which is an effective format for structuring a lesson [28]. The first phase of the lesson, "Minds On", introduces children to the lesson topic and establishes expectations (5 - 10 min). In the second phase, "Exploration", children play through 1 module of the Foodbot Factory serious game on a tablet provided by the research team (10 - 15 min). For the final phase, "Consolidation", discussion and teacher-led activities allow children to connect their new

XSL•FO
RenderX

knowledge from the serious game to their previous knowledge and lived experiences (10 min).

Control Intervention

Classrooms in the control group received nutrition education that covers the same topics as the Foodbot Factory intervention but used nontechnology based learning materials (eg, activity sheets). The control intervention also consists of 5 nutrition education lessons, following the same 3-phase lesson structure (Table 1). However, in place of the Foodbot Factory serious game, the lesson plans incorporated pre-existing resources that were sourced from a popular website repository of educational materials [29]. These resources were carefully reviewed and selected by the research team to closely match the Foodbot Factory intervention learning goals and ensure their alignment with curriculum and quality. Unlike clinical practice, where there are often existing guidelines for usual care that may serve as a control group in research, there is great variation in the strategies teachers implement for nutrition education. However, a consistent control intervention is required for comparative research purposes. The approach for the control intervention in this study was informed by qualitative interviews and focus groups conducted with Canadian elementary school teachers who frequently reported using pre-existing resources found online for nutrition [10].

XSL•FO
RenderX

Table . Foodbot Factory and control intervention learning topics and activities.

Daily learning topic and study group		Phase 1: "Minds On" introductory activities and set expectations (approximately 10 min)	Phase 2: "Action" main lesson activity (approximately 10-15 min)	Phase 3: "Consolidation" summarize and review lesson (approximately 10 min)
Day 1—drinks (Foodbot Factory) and introduction to food and drinks (control)				
	Foodbot Factory	• Introductory slide show on food and drinks • Teacher-led class activity • Class discussion	• Play "Drinks" module in Foodbot Factory serious game	• Complete phase 1 class activity, adding to it based on what was learned • Class discussion
	Control	• Introductory slide show • Complete activity sheet on food groups and drinks independently • Class discussion	• Complete activity sheet on different food groups and drinks independently	• Think-pair-share activity sheet • Take up answers from phase 2 activity • Class discussion
Day 2/3[a]—whole grain foods (Foodbot Factory day 2 and control group day 3)				
	Foodbot Factory	• Introductory slide show • Class discussion	• Play "Whole Grain Foods" module in Foodbot Factory serious game	• Small-group activity and take up answers • Class discussion
	Control	• Introductory slide show • Class discussion	• Complete activity sheets on whole grain foods independently or in pairs	• Take up answers from phase 2 activity • Class discussion
Day 2/3—vegetables and fruit (Foodbot Factory day 3 and control group day 2)				
	Foodbot Factory	• Introductory slide show • Complete activity sheet on vegetables and fruit independently • Class discussion	• Play "Vegetables & Fruit" module in Foodbot Factory serious game	• Complete phase 1 activity sheet, adding to it based on what was learned • Class discussion
	Control	• Introductory slide show • Class discussion	• Complete activity sheets on vegetables and fruit in pairs or small groups	• Take up answers from phase 2 activity • Complete activity sheet in pairs or small groups
Day 4—animal protein (Foodbot Factory) and protein foods (control)				
	Foodbot Factory	• Introductory slide show • Class discussion	• Play "Animal Protein" module in Foodbot Factory serious game	• Teacher-led class activity • Class discussion
	Control	• Introductory slide show • Class discussion	• Complete activity sheets on animal and plant protein foods independently or in pairs	• Take up answers from phase 2 activity • Small group activity sheet • Class discussion
Day 5—plant protein (Foodbot Factory) and Canada's Food Guide (control)				
	Foodbot Factory	• Introductory slide show • Class discussion	• Play "Plant Protein" module in Foodbot Factory serious game	• Teacher-led class activity • Class discussion
	Control	• Introductory slide show • Class discussion	• Complete activity sheet on Canada's Food Guide independently	• Take up answers from phase 2 activity • Class discussion

[a]Study groups cover both the topics of "Whole Grain Foods" and "Vegetables and Fruit" on different days.

XSL•FO
RenderX

Outcomes

We holistically determined if each study outcome was feasible for a research protocol by considering the context of each individual outcome, and if the outcome would reasonably facilitate a fully powered study.

Study Processes

Feasibility of study processes refers to the main elements of a study that are necessary for its success, including recruitment and retention of participants [24]. We assessed 7 variables related to study processes including school recruitment, classroom recruitment, parent recruitment, child recruitment, attrition rate, data collection, and instrument collection completion rates. Recruitment rates were calculated as the number of recruited school or participants over the total number of contacted or eligible participants (eg, school recruitment rate=schools successfully recruited/schools contacted). The attrition rate was calculated as the number of children recruited to the study with data available from 2 data collection questionnaires, a baseline Nutrition Attitudes and Knowledge (NAK) questionnaire and postintervention NAK, over the total number of child participants. Furthermore, 2 data collection completeness rates were calculated. The first was calculated as the number of child participants with 2 fully complete NAK questionnaires over the total number of child participants who completed the study. The second data collection completion rate was determined by dividing the number of structured field notes completed by study staff by the total number of expected field notes. The instrument collection completeness rate was calculated as the number of child participants with all study questionnaires completed (2 NAKs and the acceptability questionnaire) over the total number of child participants.

Study Resources

Feasibility of study resources refers to the time and resources required to run the study [24]. Furthermore, 3 variables related to study resources were assessed including time taken to deliver the intervention, time taken to collect data, and impacts on study personnel who were the study teacher and outcome assessor (ie, what was the qualitative experience for those conducting the study). Variables related to time were taken with a stopwatch and impacts on study personnel were assessed qualitatively. This study did not assess outcomes related to material resources as all necessary resources were brought into classroom by the study team (eg, tablets to play the Foodbot Factory game, activity sheets, and writing supplies).

Study Management

Feasibility of study management assesses possible issues with study implementation [24]. In this feasibility study, we collected data on 2 variables related to study management, which included challenges with data collection and challenges with intervention delivery. These outcomes were assessed qualitatively.

Preliminary Assessment of Scientific Outcomes

A preliminary assessment of the scientific outcomes of the study included an evaluation of the impacts of the intervention on nutrition knowledge and its safety. These data are intended to inform a fully powered future study, not to determine the effectiveness of an intervention [24]. In total, 3 variables were assessed in this category including acceptability of the intervention, adverse events, and impacts of the intervention on nutrition knowledge. An acceptability questionnaire assessed the perceived acceptability of the intervention by participants. The acceptability questionnaire was based on the theoretical framework of acceptability [30]. The questionnaire was adapted from the generic theoretical framework of acceptability questionnaire to be specific to the nutrition education interventions and ensure readability for children [31]. It consisted of 11 items on a numerical 5-point Likert scale questions defined by level of agreement (1=strongly disagree to 5=strongly agree). Adverse events were documented in study field notes and a self-report question on the acceptability questionnaire asking participants about feelings of stress due to the intervention. Nutrition knowledge was measured using the validated NAK questionnaire, which was developed and validated by our research team and is sensitive to detecting changes in nutrition knowledge [22,32]. The NAK questionnaire consists of 20 questions to assess a child's overall nutrition knowledge, and 4 subscores of nutrition knowledge (5 questions per subscore) to assess knowledge of a specific food group (eg, drinks, whole grain foods, vegetables and fruit, and protein foods). NAK questions are guided by the recommendations in Canada's Food Guide [20], which also are part of the Foodbot Factory intervention.

Outcome Assessment

Feasibility outcomes were collected throughout the study from (1) structured daily field notes that were created and completed by the research team to document the aforementioned study process, resource, and management outcomes; (2) assessments of nutrition knowledge using the NAK questionnaire [32]; and (3) an acceptability questionnaire. The 2 questionnaires were used to provide a preliminary assessment of the scientific impacts of the intervention.

Field notes were completed each day of the study by the study teacher and observer (days 1 through 5). The NAK questionnaire was administered to participants by an outcome assessor on day 1, before the first nutrition education lesson, and again on day 5, after the final lesson. The acceptability questionnaire was also administered by the outcome assessor and completed on day 5, after participants completed the NAK questionnaire. The questionnaires were completed independently by participants at their desks in the classroom with the outcome assessor circulating to monitor progress. The outcome assessor was present in each classroom throughout the intervention period as an observer, thus they were not blinded to the intervention group. In the first and second classrooms, students completed electronic forms using Qualtrics (Silver Lake); however, there were several issues with electronic data collection (described in Results) and the remaining 2 classrooms completed paper-based forms.

Sample Size

Our research team previously conducted a pilot study of the Foodbot Factory serious game, which provided us with an understanding of the distribution and effects sizes for the outcome of nutrition knowledge [22]. Thus, our objective with

XSL•FO
RenderX

this feasibility study was pragmatic in that we needed to understand the logistics of recruitment, data collection, and intervention delivery in the classroom setting [33]. For this type of feasibility study, 12 to 30 participants per group is suggested by other researchers [34,35]. Therefore, we aimed to recruit 4 classrooms as the feasibility challenges in this study were likely to arise related to both the classroom environment and individual participants. This resulted in 102 children eligible to participate and allowed us to assess the feasibility of implementing the interventions in different neighborhoods and understand the pragmatic elements required to scale the intervention up in a fully powered study.

Data Analysis

Outcomes related to study processes were analyzed using descriptive statistics, frequencies, and percentages. Resource outcomes were assessed using descriptive statistics and impacts on study personnel were assessed narratively. Outcomes related to management were assessed narratively. The scientific outcome of acceptability was analyzed using descriptive statistics. On all Likert scale questions, a response of 1 or 2 was considered as disagreement, a response of 3 was considered neutral, and a response of 4 or 5 was considered as agreement. Intervention safety was analyzed using both descriptive statistics and narratively. The scientific outcome of intervention impacts was assessed using descriptive statistics but was not considered in the final determination of protocol feasibility as the sample size would not have sufficient power. Within-group analysis for changes in overall nutrition knowledge and nutrition knowledge subscores was completed using a paired *t* test, after confirming the data were normally distributed through the Shapiro-Wilk test. Participants were excluded from the analysis of the overall knowledge score if there were any missing data. However, participants were included in nutrition knowledge subscore analyses if they had complete data for a particular subscore. All statistical analyses were completed using R Statistical Software version 4.2.0 (R Core Team) [36].

Results

Overview

In total, 4 classrooms participated from 2 schools between February and May 2023 (Figure 2). Both schools were in urban areas in culturally and linguistically diverse neighborhoods. The first school was in a higher socioeconomic-status neighborhood while the second school was in a neighborhood with a higher proportion of newcomers to Canada and government-subsidized housing. Participating classrooms included one Grade 4 and three Grade 4/5 split classes.

XSL•FO
RenderX

Figure 2. Flowchart of participant progress through the study.

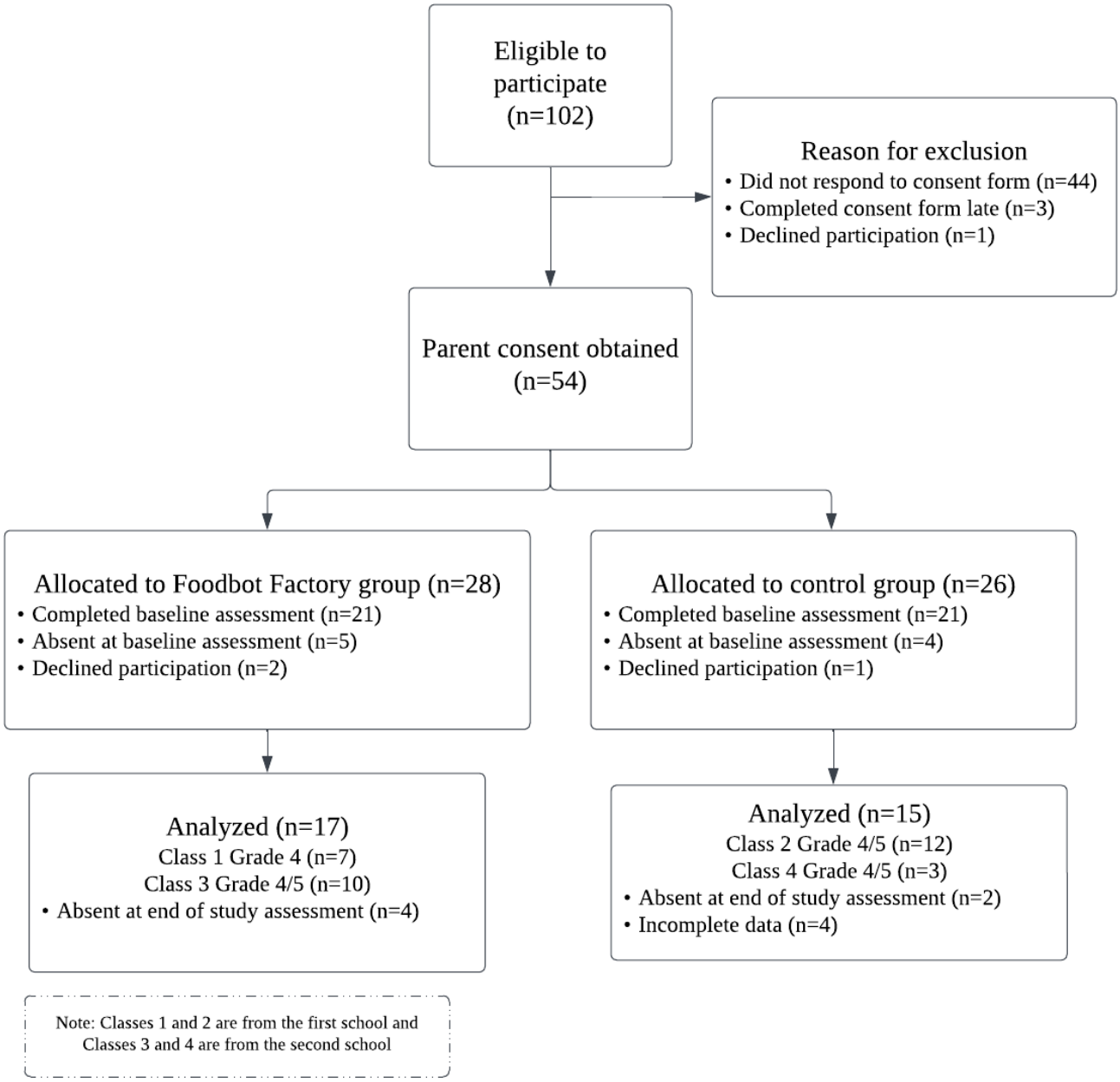

Study Process Outcomes

The school recruitment rate was low (3/20, 15%). Of the 20 principals who were contacted by email, 70% (n=13) did not respond, 10% (n=2) declined participation, and 5% (n=1) replied with a question about the study but lost to follow-up. To counter the low email response rate, the research used phone calls for recruitment and was able to recruit 1 school after 3 phone calls. Once a principal agreed to have their school participate, 100% of classroom teachers in those schools agreed to have their classroom participate. In total, 4 classrooms from 2 of the 3 recruited schools were scheduled for participation, with a total of 102 children. While a third school was recruited to participate, they ultimately were not scheduled to be in the study, due to a staffing shortage among the research team. Parent recruitment rate was low (54/102, 53%); 43% (44/102) did not respond to the consent form, 3% (3/102) consented after the study start date, and 1% (1/102) declined participation. To improve parent recruitment, the research team requested that classroom teachers send a reminder email. Reminders were sent by teachers in 2 classrooms, resulting in a 77% parent recruitment rate compared with 40% in classrooms with no reminders. Anecdotally, 1 classroom with lower parent recruitment had a high number of parents who did not speak English as a first language, and the classroom teacher noted this was likely a barrier to participating in the study. Most children with parental consent (49/54, 91%) agreed to participate in the study, with 24 in the intervention group and 25 in the control group. The average age of participants was 9.6 (SD 0.68) years. Participants self-reported their gender as a boy (24/49, 49%), girl (18/49, 37%), or a self-defined identity (2/49, 4%), with the remaining participants preferring not to answer (5/49, 8%). In the intervention group, 17 participants completed the study; however, only 10 had a complete set of data. In the control group, 15 participants completed the study, with 9 having a complete set of data. The attrition rate was 19%. For participants, the instrument collection completion rate was moderate (73%) and the data collection completion rate for the NAK questionnaire was low (59%). The

XSL•FO
RenderX

data collection completion rate of the structured field notes by study staff was 100%.

Study Resource Outcomes

The average time required for child assent was 6.5 minutes, lesson plan delivery was 34 minutes (8 min for minds on, 14 min for exploration, and 12 min for consolidation), acceptability questionnaire completion was 5.5 minutes, and NAK questionnaire completion was 13 minutes; all consistent with original estimates. Study personnel, including the study teacher and outcome assessor, reported an overall positive experience as they enjoyed interacting with the students. In one classroom, classroom management was a significant challenge due to a handful of students who frequently interrupted the lessons and outcome assessments. This was a minor psychological stressor for the study teacher. There were challenges with retaining the hired study teachers on the project due to career changes and an unanticipated emergency, requiring the research team to reschedule 2 classrooms to a later date.

Study Management Outcomes

Data collection procedures and delivery of the intervention were executed as planned, although some challenges were experienced. Tablets were originally used for participants to complete data collection forms, but this posed a few unanticipated challenges. For example, participants intentionally or unintentionally “refreshed” or navigated away from the web-based NAK questionnaire, leading to the loss of completed responses. In addition, once a NAK questionnaire was submitted, study personnel were unable to verify if all questions had been completed. This resulted in incomplete data for some participants. Several participants struggled to focus on completing the data collection forms due to distractions in the classroom. The most significant distraction occurred when other children who were not participating in the study were allowed by the classroom teacher to engage in other activities, namely the use of electronic devices. Furthermore, as participants completed the questionnaire independently, those who finished early then moved on to a different activity. Participants who needed more time to complete the questionnaire became increasingly distracted as others around them finished, which may have contributed to them rushing to finish and skipping questions.

Based on the implementation of the interventions in the first 2 classrooms, some changes were made to the intervention delivery that were implemented in the subsequent 2 classrooms. The changes clarified instructions, modified select discussion questions to improve wording and added slideshows with the daily learning objectives instead of writing them on the board. Overall, these changes were relatively minor and were intended to improve the clarity and accessibility of the lesson delivery. In the study, classroom management was particularly challenging in 1 classroom due to some children repeatedly interrupting the study teacher. Study personnel found that having the classroom teacher and educational assistants support classroom management was helpful. In 2 classrooms, the intervention schedule was modified to occur over 4 days instead of 5, due to a conflicting school-wide track and field event. On one of the 4 days, 2 intervention lessons were provided consecutively instead of having 1 lesson each day. While it was logistically feasible to provide 2 lessons back-to-back in 1 day, student engagement was lower during the second lesson. Another challenge occurred in the Foodbot Factory intervention group, where some children struggled with the augmented reality components in the Foodbot Factory serious game. The research team responded to this by improving instructions about how the augmented reality features work. In addition, some children struggled to transition away from using the tablet to play the Foodbot Factory serious game to the next learning activity as they found the game engaging and wanted to keep playing.

Preliminary Assessment of Scientific Outcomes

Based on observations, participants enjoyed participating in both the Foodbot Factory and control intervention lessons and were engaged in learning about nutrition through the activities. Overall, intervention acceptability among participants was moderate to high, with most participants reporting that they had a positive experience (26/41, 63%), understood the goals of the nutrition education lessons (37/41, 90%), and that they learned something new (31/41, 76%; Figure 3). No participants reported feeling sad or stressed about food and nutrition from the lessons on the acceptability questionnaire nor did study personnel observe any verbal or nonverbal signs of concern from children in the classroom.

Scores on the NAK questionnaire were normally distributed as determined by the Shapiro-Wilk test. Both groups demonstrated improvements in their overall nutrition knowledge (Foodbot Factory group: mean 10.8, SD 2.1 to mean 12.5, SD 3.5; control group: mean 11.8, SD 1.9 to mean 12.7, SD 1.8), although these improvements were not statistically significant. For subscores of nutrition knowledge, participants in the intervention group demonstrated statistically significant improvements in knowledge of vegetables and fruit and protein foods. Participants in the control group demonstrated statistically significant improvements in knowledge of grain foods (Table 2).

XSL•FO
RenderX

Figure 3. Acceptability of the nutrition education interventions as rated by participants (n=41).

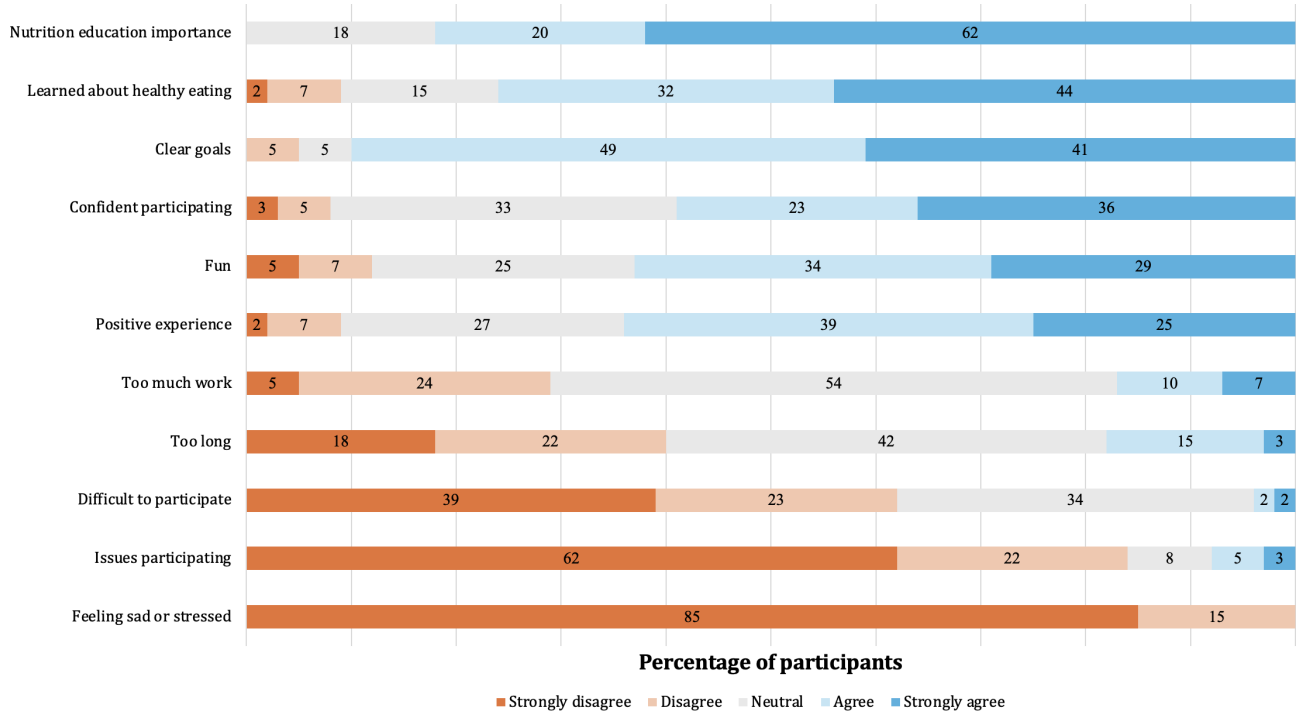

Table . Changes in overall and subscores of nutrition knowledge within the Foodbot Factory and control groups[a].

Nutrition knowledge	Foodbot Factory group				Control group			
	Score, n	Baseline, mean (SD)	End of study, mean (SD)	*P* value	Score, n	Baseline, mean (SD)	End of study, mean (SD)	*P* value
Overall nutrition knowledge	10	10.8 (2.1)	12.5 (3.5)	.13	9	11.8 (1.9)	12.7 (1.8)	.09
Drinks	17	3.94 (0.66)	3.76 (0.56)	.33	15	3.93 (0.70)	3.93 (0.59)	≥.99
Grain foods	16	2.31 (1.2)	2.75 (1.1)	.17	15	2.87 (1.1)	3.40 (1.1)	.03
Vegetables and fruit	15	2.27 (0.88)	3.00 (0.93)	.04	14	2.86 (0.86)	3.07 (1.0)	.61
Protein foods	12	1.92 (0.90)	2.75 (1.6)	.03	10	2.60 (1.1)	1.70 (0.95)	.05

[a]Data presented as means and SDs with paired *t* tests used to assess changes in knowledge from baseline to the end-of-study within groups. Sample sizes vary as only participants with a complete set of responses were included in the analysis.

Discussion

Principal Findings

Our research study determined the feasibility of implementing the Foodbot Factory nutrition education intervention as part of a research protocol in Grade 4 and 4/5 classrooms, identifying 9 facilitators to support a larger study and 4 risks that will require mitigation. Elements of the study that were deemed feasible included classroom recruitment, child recruitment, the attrition rate, time taken to deliver the intervention and collect data, the impacts of the study on personnel, management of delivering the intervention, lack of adverse events, and acceptability of the intervention. However, strategies will be needed to improve school recruitment, parent recruitment, data collection completeness, and the management of data collection procedures. In this feasibility study, we made minor modifications to the protocol on implementation to improve recruitment and intervention implementation (eg, conducting recruitment phone calls and modifying intervention instructions for the classroom). With further modifications to mitigate the aforementioned risks, our findings largely support the feasibility of this research protocol to evaluate the Foodbot Factory nutrition education intervention as part of a fully powered randomized controlled trial.

The primary strength of this study was our ability to better understand how to collaborate with schools, teachers, parents, and children in our future research. Key takeaways learned from this study that will be valuable to other researchers, include the importance of consulting with school boards before conducting research, ensuring recruitment materials and methods are relevant for the target audience, engaging classroom teachers and testing data collection methods before their implementation. We discussed our methodology and objectives with staff from several school boards before seeking school board approval to

XSL•FO
RenderX

conduct the research. From the outset, this approach resulted in a protocol that would be more feasible and acceptable to implement in classrooms and ensure value to our participants and the school board. For example, the initial version of our study protocol requested 5 hours of total class time, which was perceived as too much time taken away from other curriculum. Not only did we modify the protocol to reduce the total time by 1 hour, we also updated our recruitment materials to highlight alignment of the study with curriculum. This was a critical change as teachers have very limited time to cover content that is outside of the curriculum. Ultimately, the resulting protocol is not only acceptable to school boards, but the duration of the intervention now better aligns with the time teachers typically allocate to a lesson and clearly communicates the value of the study to possible participants. These features increase real-world acceptability of the intervention. Researchers should consult with their local school boards and teachers to establish a collaborative relationship from the beginning to develop interventions and research protocols that are suitable and valuable.

When recruiting individual schools, classroom teachers, and parents, recruitment materials should be succinct and highlight the value of study participation. In this study, phone calls were effective at recruiting schools as they enable direct consultation on the practical aspects of the study and allow the school to quickly evaluate if the study will work for them. Low recruitment of schools was still seen in this study and has been reported as a challenge in other school-based health studies due to low interest, poor timing, and other commitments [37,38]. During this study, engagement of the classroom teacher was critical for parent recruitment, as we were not permitted to contact parents directly. We found low parent recruitment was driven by parents not returning consent forms, rather than them explicitly declining participation, which has been observed in other studies [37]. We also observed significant improvements in recruitment when the classroom teacher sent reminders to parents about the study, highlighting the important role of classroom teachers in engaging parents in study participation. In this study, classroom teachers were asked to contact all parents and guardians to share the consent form and reminders to complete it. However, classroom teachers may introduce selection bias if they selectively reach out to parents and guardians. Instructions for sending out parent consent forms to classroom teachers should be clear and concise to avoid this bias.

In this study, we show the importance of testing data collection methods before implementation in a classroom setting. First, an important takeaway from this study is that electronic data collection presented several challenges. Although other studies have used electronic data collection methods, it may present more challenges if the technology has not been used before in that context [39]. In this study, training was not provided to participants on how to complete the electronic data collection form before data collection as we did not have sufficient time allotted for such training. Training on data collection methods could minimize issues like participants accidentally refreshing the page and skipping questions [39]. In this study, questions on the electronic data collections forms were not made mandatory for submission as study participation was voluntary. Second, participants completed data collection independently, requiring them to use their literacy skills to read and respond to each question. This data collection approach reduced accessibility and led to participants finishing at varying time points and the classroom environment becoming increasingly distracting for participants who needed more time. Missing data from both missed responses and missing surveys in other school-based studies is a common phenomenon ranging from 11% - 51% at a given outcome assessment time point [40,41]. In a future protocol, data collection could be modeled on the procedures classrooms use when administering standardized tests. This would align research data collection with procedures that are already familiar to children, provide clear instructions, increase accessibility, allow for verification of data collection completeness to encourage complete responses and directly guide the classroom as a group through the questions so all participants finish at the same time.

This study also examined the impacts of the developed nutrition education interventions on children's nutrition knowledge. Statistically insignificant improvements in overall nutrition knowledge were observed in both groups, and participants in the Foodbot Factory group significantly improved their knowledge of vegetables and fruit and protein foods. The results within-groups at the individual level indicate improvements in nutrition knowledge but do not have the same effect size as a larger proof-of-concept study of the Foodbot Factory serious game, where significant improvements in children's overall nutrition knowledge were observed both within and between groups [22]. This indicates that the findings in this feasibility study for nutrition knowledge are inconclusive, due to the insufficiently powered sample size, and should be used primarily for descriptive purposes and informing improvements to the intervention [42]. Some changes were made to the lesson plans during the study to improve clarity and accessibility (eg, using a slideshow to share learning objectives instead of writing them on the board) as the process of conducting feasibility studies and developing research protocols is recognized to be iterative and adaptive [43]. Since these changes were relatively minor and were focused on delivery, rather than intervention content, we do not anticipate they influenced the outcomes of interest.

This study had several strengths and limitations that can inform future research. This study included a small sample size, which is primarily driven by low parent recruitment. Recruitment may be a primary feasibility challenge moving forward to a fully powered study. Including more classrooms may have revealed additional feasibility issues; however, the classrooms were intentionally sampled from schools in neighborhoods with known differences in their sociodemographic profiles. While the challenges in this study are similar to those observed in international research [37,38,40,41], the results may be most applicable for those conducting school-based research in the Greater Toronto Area, or other urban environments in Ontario, Canada. In this study, we chose not to randomize classrooms as we were unsure of how successful our recruitment would be and this approach would allow us to ensure balance between study groups. Nonrandomized feasibility studies are common for research in earlier phases of preparation for a trial and when

XSL•FO
RenderX

there are significant unknowns [44]. To address the lack of randomization and balance school- and neighborhood-level characteristics, the research team assigned 1 classroom at each school to a different group; however, we acknowledge the potential of selection bias and an uneven balance of characteristics between groups that may have been introduced. In future research, we will recruit 2 classrooms per school and use pair-matching randomization to both reduce bias and ensure groups are balanced on school and neighborhood level variables that can impact our outcomes of interests. Due to the study design, where classrooms rather than individuals were assigned to each treatment arm, we were unable to assess differences between groups as the data should be analyzed at the classroom cluster level, which this study was not powered to do. In future research, an adequately powered randomized controlled trial that randomizes classrooms as clusters would be more appropriate to assess the impact of the interventions in a pragmatic fashion. In this study, we provided all material resources necessary, including tablets for students to use the Foodbot Factory serious game. This can be considered as a strength, as it increased the feasibility of conducting the research, but also a limitation as we did not evaluate possible material resource limitations (and the knowledge and processes required to implement them) for conducting technology-based research in schools. Researchers should carefully consult with school boards to understand the availability of resources to ensure success or invest in the required technology to avoid material resource barriers as we have done in this study. We also chose to use a certified study teacher who was trained to implement the lesson plans as it increases feasibility from the school board. We will use this approach in a future efficacy study. Future research should aim to also evaluate the effectiveness of the research when the lesson plans are implemented by classroom teachers. At this stage, our research team did not assess retention of knowledge or examine other variables that may be impacted by a nutrition education intervention, including dietary intake and behaviors. We also did not assess acceptability of the study protocol from the perspective of classroom teachers, which was not the focus of this study. Retention of knowledge and dietary intake will be assessed in future research and qualitative interviews will also be conducted with classroom teachers to inform implementation needs and strategies for the Foodbot Factory intervention. A current limitation of the Foodbot Factory intervention is that it predominantly covers nutrition knowledge and does not address all elements of food literacy. The serious game is also only available for use on mobile devices. Our future work aims to make a web-accessible version of the Foodbot Factory serious game and expand the content to different age groups and food literacy concepts.

Conclusions

In conclusion, the majority of feasibility outcomes assessing study processes, resources, management, and intervention acceptability suggest that the study protocol was feasible [45]. The data offer several insights to inform future studies and, when the protocol is implemented into a fully powered trial, the successes identified can be leveraged, and the risks associated with school recruitment, parent recruitment, and data collection can be averted with targeted mitigation strategies. These lessons learned have been applied to our protocol for a fully powered cluster randomized controlled trial evaluating the efficacy of Foodbot Factory in classrooms [46]. The development and evaluation of the protocol in this study was an important step in understanding how to best evaluate technology-based nutrition education interventions in Canadian classrooms. The lessons learned in this study will also support the rising number of researchers embarking on school-based nutrition and health interventions in school and classroom settings both in Canada and abroad.

Acknowledgments

Funding for this research was provided by Ontario Tech University's Dean's Discretionary Fund Small Research Grants Program (JA), a Social Sciences and Humanities Research Council (SSHRC) Explore Grant (BFA), an Ontario Graduate Scholarship (JMB), and a Canadian Institutes of Health Research (CIHR) Canada Graduate Doctoral Scholarship (JMB). The authors would like to thank the principals, teachers, parents, and children for being part of this research.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Informed consent form.
[DOCX File, 54 KB - formative_v9i1e69242_app1.docx]

References

1. Thomas H, Azevedo Perry E, Slack J, et al. Complexities in conceptualizing and measuring food literacy. J Acad Nutr Diet 2019 Apr;119(4):563-573. [doi: 10.1016/j.jand.2018.10.015] [Medline: 30670348]
2. Thakur S, Mathur P. Nutrition knowledge and its relation with dietary behaviour in children and adolescents: a systematic review. Int J Adolesc Med Health 2022 Dec 1;34(6):381-392. [doi: 10.1515/ijamh-2020-0192] [Medline: 33594848]
3. Vaitkeviciute R, Ball LE, Harris N. The relationship between food literacy and dietary intake in adolescents: a systematic review. Public Health Nutr 2015 Mar;18(4):649-658. [doi: 10.1017/S1368980014000962] [Medline: 24844778]

XSL•FO
RenderX

4. Hack S, Jessri M, L'Abbé MR. Nutritional quality of the food choices of Canadian children. BMC Nutr 2021 May 29;7(1):16. [doi: 10.1186/s40795-021-00422-6] [Medline: 34049592]
5. Ahmed M, Praneet Ng A, L'Abbe MR. Nutrient intakes of Canadian adults: results from the Canadian Community Health Survey (CCHS)-2015 Public Use Microdata File. Am J Clin Nutr 2021 Sep 1;114(3):1131-1140. [doi: 10.1093/ajcn/nqab143] [Medline: 34020449]
6. Afshin A, Sur PJ, Fay KA, et al. Health effects of dietary risks in 195 countries, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. The Lancet 2019 May;393(10184):1958-1972. [doi: 10.1016/S0140-6736(19)30041-8]
7. School nutrition: support for providing healthy food and beverage choices in schools. Canadian Pediatric Society. 2020. URL: https://cps.ca/documents/position/school-nutrition-support#:~:text=Schools%20provide%20important%20opportunities%20to,healthier%20dietary%20choices%20and%20behaviours [accessed 2022-02-01]
8. Jones AM, Zidenberg-Cherr S. Exploring nutrition education resources and barriers, and nutrition knowledge in teachers in California. J Nutr Educ Behav 2015;47(2):162-169. [doi: 10.1016/j.jneb.2014.06.011] [Medline: 25257845]
9. Perera T, Frei S, Frei B, et al. Improving nutrition education in U.S. elementary schools: challenges and opportunities. JEP 2015;6(30):41-50 [FREE Full text]
10. Franco-Arellano B, Vogel E, Stein M, et al. A qualitative study exploring elementary school teachers' experiences teaching nutrition in Canada. Presented at: International Society for Behavioral Nutrition and Physical Activity Annual Meeting; Jun 14-17, 2023; Uppsala, Sweden p. 354 URL: https://annualmeeting.isbnpa.org/wp-content/uploads/2023/07/ISBNPA-2023-Abstract-Book-web-final.pdf [accessed 2025-04-02]
11. Graziose MM, Koch PA, Wang YC, Lee Gray H, Contento IR. Cost-effectiveness of a nutrition education curriculum intervention in elementary schools. J Nutr Educ Behav 2017 Sep;49(8):684-691. [doi: 10.1016/j.jneb.2016.10.006] [Medline: 27843129]
12. Cotton W, Dudley D, Peralta L, Werkhoven T. The effect of teacher-delivered nutrition education programs on elementary-aged students: an updated systematic review and meta-analysis. Prev Med Rep 2020 Dec;20:101178. [doi: 10.1016/j.pmedr.2020.101178] [Medline: 32944494]
13. Education Indicators in Canada: An International Perspective: Statistics Canada; 2020.
14. Laamarti F, Eid M, El Saddik A. An overview of serious games. International Journal of Computer Games Technology 2014;2014:1-15. [doi: 10.1155/2014/358152]
15. Baranowski T, Ryan C, Hoyos-Cespedes A, Lu AS. Nutrition education and dietary behavior change games: a scoping review. Games Health J 2019 Jun;8(3):153-176. [doi: 10.1089/g4h.2018.0070] [Medline: 30339086]
16. Thompson D, Bhatt R, Vazquez I, et al. Creating action plans in a serious video game increases and maintains child fruit-vegetable intake: a randomized controlled trial. Int J Behav Nutr Phys Act 2015 Mar 18;12(1):39. [doi: 10.1186/s12966-015-0199-z] [Medline: 25890060]
17. Hermans RCJ, van den Broek N, Nederkoorn C, Otten R, Ruiter ELM, Johnson-Glenberg MC. Feed the Alien! The effects of a nutrition instruction game on children's nutritional knowledge and food intake. Games Health J 2018 Jun;7(3):164-174. [doi: 10.1089/g4h.2017.0055] [Medline: 29634366]
18. Brown JM, Franco-Arellano B, Froome H, Siddiqi A, Mahmood A, Arcand J. The content, quality, and behavior change techniques in nutrition-themed mobile apps for children in Canada: app review and evaluation study. JMIR Mhealth Uhealth 2022 Feb 16;10(2):e31537. [doi: 10.2196/31537] [Medline: 35171100]
19. Brown JM, Savaglio R, Watson G, et al. Optimizing child nutrition education with the Foodbot Factory mobile health app: formative evaluation and analysis. JMIR Form Res 2020 Apr 17;4(4):e15534. [doi: 10.2196/15534] [Medline: 32301743]
20. Canada's food guide. Government of Canada. 2019. URL: https://food-guide.canada.ca/en [accessed 2024-10-24]
21. Ontario Ministry of Education. Health and Physical Education Grades 1-8: Government of Ontario: Toronto; 2019.
22. Froome HM, Townson C, Rhodes S, et al. The effectiveness of the Foodbot Factory mobile serious game on increasing nutrition knowledge in children. Nutrients 2020 Nov 6;12(11):3413. [doi: 10.3390/nu12113413] [Medline: 33172094]
23. Brown JM, Rita N, Franco-Arellano B, LeSage A, Arcand J. Co-design and refinement of curriculum-based Foodbot Factory intervention to support elementary school nutrition education. Nutrients 2024 Nov 2;16(21):3769. [doi: 10.3390/nu16213769] [Medline: 39519601]
24. Thabane L, Ma J, Chu R, et al. A tutorial on pilot studies: the what, why and how. BMC Med Res Methodol 2010 Jan 6;10(1):1. [doi: 10.1186/1471-2288-10-1] [Medline: 20053272]
25. Franco-Arellano B, Brown JM, Daggett Q, et al. Updating the Foodbot Factory serious game with new interactive engaging features and enhanced educational content. Appl Physiol Nutr Metab 2024 Jan 1;49(1):52-63. [doi: 10.1139/apnm-2023-0214] [Medline: 37905542]
26. Kiili K. Digital game-based learning: towards an experiential gaming model. Internet High Educ 2005 Jan;8(1):13-24. [doi: 10.1016/j.iheduc.2004.12.001]
27. Gordon M. Toward a pragmatic discourse of constructivism: reflections on lessons from practice. Educ Stud 2009 Feb 3;45(1):39-58. [doi: 10.1080/00131940802546894]
28. Walle JA, Bay-Williams JM, Karp KS. Elementary and Middle School Mathematics: Teaching Developmentally, 9th edition: Pearson; 2016.

XSL•FO
RenderX

29. About us. Teachers Pay Teachers. 2022. URL: https://www.teacherspayteachers.com/About-Us [accessed 2022-03-06]
30. Sekhon M, Cartwright M, Francis JJ. Acceptability of healthcare interventions: an overview of reviews and development of a theoretical framework. BMC Health Serv Res 2017 Jan 26;17(1):88. [doi: 10.1186/s12913-017-2031-8] [Medline: 28126032]
31. Sekhon M, Cartwright M, Francis JJ. Development of a theory-informed questionnaire to assess the acceptability of healthcare interventions. BMC Health Serv Res 2022 Mar 1;22(1):279. [doi: 10.1186/s12913-022-07577-3] [Medline: 35232455]
32. Franco-Arellano B, Brown JM, Froome HM, LeSage A, Arcand J. Development and pilot testing of the Nutrition Attitudes and Knowledge Questionnaire to measure changes of child nutrition knowledge related to the Canada's Food Guide. Appl Physiol Nutr Metab 2021 Dec;46(12):1495-1501. [doi: 10.1139/apnm-2021-0170] [Medline: 34289315]
33. Chan CL, Taljaard M, Lancaster GA, Brehaut JC, Eldridge SM. Pilot and feasibility studies for pragmatic trials have unique considerations and areas of uncertainty. J Clin Epidemiol 2021 Oct;138:102-114. [doi: 10.1016/j.jclinepi.2021.06.029] [Medline: 34229091]
34. Julious SA. Sample size of 12 per group rule of thumb for a pilot study. Pharm Stat 2005 Oct;4(4):287-291. [doi: 10.1002/pst.185]
35. Johanson GA, Brooks GP. Initial scale development: sample size for pilot studies. Educ Psychol Meas 2010 Jun;70(3):394-400. [doi: 10.1177/0013164409355692]
36. R Core Team. R: a language and environment for statistical computing. : R Foundation for Statistical Computing; 2022.
37. Braga-Pontes C, Simões-Dias S, Lages M, Guarino MP, Graça P. Nutrition education strategies to promote vegetable consumption in preschool children: the Veggies4myHeart project. Public Health Nutr 2022 Apr;25(4):1061-1070. [doi: 10.1017/S1368980021004456] [Medline: 34702425]
38. Epstein DS, Enticott JC, Larson HJ, Barton C. Recruiting for research on sensitive topics in schools: an experience with Vaxcards, a collectable vaccine card game. Trials 2021 May 3;22(1):320. [doi: 10.1186/s13063-021-05288-4] [Medline: 33941247]
39. Hodges A, Cordier R, Joosten A, Bourke-Taylor H, Chen YW. Evaluating the feasibility, fidelity, and preliminary effectiveness of a school-based intervention to improve the school participation and feelings of connectedness of elementary school students on the autism spectrum. PLOS ONE 2022;17(6):e0269098. [doi: 10.1371/journal.pone.0269098] [Medline: 35648779]
40. Mabungane S, Ramroop S, Mwambi H. Analysis of missing data in progressed learners: the use of multiple imputation methods. African Journal of Research in Mathematics, Science and Technology Education 2023 May 4;27(2):112-122. [doi: 10.1080/18117295.2023.2193496]
41. van de Weijer-Bergsma E, Langenberg G, Brandsma R, Oort FJ, Bögels SM. The effectiveness of a school-based mindfulness training as a program to prevent stress in elementary school children. Mindfulness (N Y) 2012. [doi: 10.1007/s12671-012-0171-9]
42. Tickle-Degnen L. Nuts and bolts of conducting feasibility studies. Am J Occup Ther 2013;67(2):171-176. [doi: 10.5014/ajot.2013.006270] [Medline: 23433271]
43. Eldridge SM, Lancaster GA, Campbell MJ, et al. Defining feasibility and pilot studies in preparation for randomised controlled trials: development of a conceptual framework. PLoS ONE 2016;11(3):e0150205. [doi: 10.1371/journal.pone.0150205] [Medline: 26978655]
44. Lancaster GA, Thabane L. Guidelines for reporting non-randomised pilot and feasibility studies. Pilot Feasibility Stud 2019;5(1):114. [doi: 10.1186/s40814-019-0499-1] [Medline: 31608150]
45. Teresi JA, Yu X, Stewart AL, Hays RD. Guidelines for designing and evaluating feasibility pilot studies. Med Care 2022 Jan 1;60(1):95-103. [doi: 10.1097/MLR.0000000000001664] [Medline: 34812790]
46. Brown JM, Tahir S, Franco-Arellano B, et al. Efficacy of the Foodbot Factory digital curriculum-based nutrition education intervention in improving children's nutrition knowledge, attitudes and behaviours in elementary school classrooms: protocol for a cluster randomised controlled trial. BMJ Open 2025 Jan 28;15(1):e092426. [doi: 10.1136/bmjopen-2024-092426] [Medline: 39880422]

Abbreviations

NAK: Nutrition Attitudes and Knowledge

XSL•FO
RenderX

Edited by A Mavragani; submitted 25.11.24; peer-reviewed by L Aljerf, MCE Battjes-Fries; revised version received 05.03.25; accepted 18.03.25; published 16.04.25.

Please cite as:
Brown JM, Rita N, Franco-Arellano B, LeSage A, Arcand J
Evaluation of a Curriculum-Based Nutrition Education Intervention Protocol in Elementary Schools: Nonrandomized Feasibility Study
JMIR Form Res 2025;9:e69242
URL: https://formative.jmir.org/2025/1/e69242
doi:10.2196/69242

© Jacqueline Marie Brown, Nicholas Rita, Beatriz Franco-Arellano, Ann LeSage, Joanne Arcand. Originally published in JMIR Formative Research (https://formative.jmir.org), 16.4.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Evaluating the Acceptability of a Brief Web-Based Alcohol Misuse Prevention Program Among US Military Cadets: Mixed Methods Formative Evaluation

Emily Schmied[1,2], MPH, PhD; Lauren Hurtado[3,4], BS; W Ken Robinson[5], DMin; Cynthia M Simon-Arndt[4], MBA, MA; Richard Moyer III[6], PsyD; Leslie Wilson[7], PhD; Mark Reed[1], PhD; Shannon M Blakey[8], PhD; Marni Kan[8], PhD

[1]School of Public Health, San Diego State University, San Diego, CA, United States
[2]Institute for Behavioral and Community Health, San Diego, CA, United States
[3]Leidos Inc, San Diego, CA, United States
[4]Psychological Health and Readiness Department, Military Population Health Directorate, Naval Health Research Center, San Diego, CA, United States
[5]United States Air Force Academy, CO, United States
[6]San Diego State University Research Foundation, San Diego, CA, United States
[7]Counseling and Psychological Services, San Diego State University, San Diego, CA, United States
[8]RTI International, Research Triangle Park, NC, United States

Corresponding Author:
Emily Schmied, MPH, PhD
School of Public Health, San Diego State University, San Diego, CA, United States

Abstract

Background: As alcohol misuse remains pervasive within the military, evidence-based prevention programs that are feasible to implement and appropriately tailored to meet the needs and norms of military personnel are critically needed. Further, programs that target future military leaders, such as trainees, recruits, and cadets, may be especially impactful. eCHECKUP TO GO is a web-based, evidence-based brief alcohol intervention designed to reduce alcohol misuse through education and personalized feedback that may be suitable for military trainees. However, because it was developed for civilian students, efforts to adapt the content for military settings are needed.

Objective: This study aimed to evaluate the acceptability of a military version of eCHECKUP TO GO, tailored to include military-specific terminology and alcohol use statistics.

Methods: US Air Force Academy cadets were recruited to participate in a single-arm, mixed methods study. Following the completion of eCHECKUP TO GO, participants completed a survey that assessed satisfaction with specific aspects of the user experience, including ease of use, design, and relevance of the information and personalized feedback (range: 1=strongly disagree to 7=strongly agree). A subset of cadets also participated in a focus group to expound on the survey responses.

Results: Survey participants included 22 cadets (n=12, 55% male; mean age 19.6, SD 1.8 years). In addition, 6 (27%) cadets participated in the focus group. Participants were satisfied with the program overall (mean 5.8, SD 0.9) and gave the highest ratings to ease of use (mean 6.6, SD 0.7), site design (mean 6.5, SD 0.6), and site interactivity (mean 6.4, SD 1.0). Items pertaining to tailoring, relevance, and amount of content specific to cadets scored lowest (mean 5.8, SD 1.4; mean 5.6, SD 1.4; and mean 5.5, SD 1.5, respectively). Most (n=15, 68%) participants said they would act upon the information they were provided. Focus group participants made suggestions for improved tailoring, such as increasing content on social aspects of drinking and military-specific risks of alcohol misuse (eg, Uniform Code of Military Justice violations).

Conclusions: Although the acceptability of eCHECKUP TO GO was high, continued efforts are needed to ensure the content accurately reflects the experiences of cadets. Researchers who design military health promotion interventions need to consider the varied contexts within the force and rigorously evaluate the acceptability of all content before implementation.

(JMIR Form Res 2025;9:e67637) doi:10.2196/67637

KEYWORDS

alcohol misuse prevention; digital interventions; formative research; military health; acceptability; alcohol prevention; US military; military; United States; formative evaluation; alcohol use; evidence-based prevention; alcohol intervention; mixed methods study; survey; alcohol use disorder; alcohol misuse; heavy drinker; educational web-based intervention; web-based intervention

XSL•FO
RenderX

Introduction

Alcohol misuse is pervasive within the US military [1-4]. Data from the 2018 Department of Defense (DoD) Health Related Behaviors Survey indicated that binge drinking and heavy drinking [5] are highly prevalent (34% and 9.8%, respectively). For many service members, alcohol misuse may lead to serious adverse health, career, and social outcomes [6-8]. For instance, among all active duty service members, alcohol use disorder was among the top 10 causes of outpatient medical visits for both men and women in 2022 [2,3]. During the same period, alcohol-related issues were also a leading cause of inpatient hospitalization [3]. It has also been associated with earlier separation from service and disciplinary action [6,7]. Further, it has been noted that alcohol misuse during service may persist following separation [9,10], potentially contributing to the widely documented health disparities between veterans and civilians [11,12].

Alcohol misuse can pose serious threats to the health and performance of service members, thus interventions designed to prevent it before it causes lasting harm are urgently needed. Prior research has shown that the implementation of alcohol misuse prevention programs in the military is complicated by numerous factors, including the size, diversity, and wide geographic distribution of the force [11]. One possible approach is to target individuals in the earliest stages of their careers who likely already consume alcohol but have yet to develop problematic alcohol habits [13-15], namely those enrolled in military service academies or training settings. Another promising strategy is to use brief alcohol interventions, or BAIs, which often consist of singular sessions and thus require fewer resources to implement [16]. BAIs are endorsed for use among service members who screen positive for unhealthy alcohol use in the current Department of Veterans Affairs and DoD joint clinical practice guidelines for the Management of Substance Use Disorders [17].

BAIs have been used in military settings with varying results, though most reports are from studies conducted in veteran or general active duty populations [18]. The available evidence from training settings shows promise. For example, Klesges et al [19] evaluated the Alcohol Misconduct Prevention Program (AMPP) in Air Force technical training [20,21]. AMPP, which consisted of a combination of a single-session, group-administered BAI, and random breathalyzer tests, was associated with a significant reduction in alcohol-related risk incidents (eg, alcohol-related disciplinary infractions). While AMPP was found to be cost-effective [22], the continued implementation of in-person group trainings may not be sustainable in all settings. Further, because alcohol use is prohibited in military training settings [11], participant concerns about confidentiality may limit the impact of such methodologies [23]. Web-based BAIs that give personalized feedback to individual program participants may provide a solution to these challenges.

Prior web-based trainings (including BAIs) evaluated in active duty personnel and veterans have been shown to be acceptable and efficacious, but few have been evaluated in service academies or training settings, leaving a critical knowledge gap [24-27]. This study addresses this gap by assessing the acceptability of a web-based, individualized BAI among military cadets. eCHECKUP TO GO (alcohol-specific version; San Diego State University) is an evidence-based, theoretically driven web-based intervention originally developed for civilian adolescents and young adults. It is designed to reduce alcohol misuse by targeting cognitive risk factors (eg, peer normative behaviors) and increasing positive behavioral strategies (eg, increasing time between drinks) to minimize drinking-related risks by providing personalized feedback to participants. The efficacy of eCHECKUP TO GO has been repeatedly demonstrated in young adults [28-32]. Given the purpose, format, and prior success of eCHECKUP TO GO, it may be suitable for decreasing alcohol misuse in military academies. However, before implementing eCHECKUP TO GO, it is important to determine whether the format and existing content are acceptable among potential program participants. Specifically, while there are demographic similarities between civilian young adults and military cadets and midshipmen, the program may need to be tailored to account for cultural and contextual differences between the communities. The results of this study can be used to determine whether eCHECKUP TO GO may be appropriate for use in military academies or training settings.

Methods

Study Design and Procedures

Participants in this study were a convenience sample of 22 cadets enrolled at the US Air Force Academy (USAFA). Prospective participants were recruited through third-party announcements distributed by local study partners by email and word-of-mouth. Interested cadets attended a recruitment event where the study purpose and procedures were explained by members of the research team and those who provided written consent to participate immediately began their participation. A mixed methods design was used in which all consenting participants completed the eCHECKUP TO GO program (alcohol version) and an electronically administered, anonymous feedback survey. Immediately after completing the feedback survey, a subset of participants volunteered to participate in a 20-minute focus group discussion designed to expound on the content of the survey. The focus group was facilitated by two members (ES and LH) of the research team and was audio-recorded and transcribed for analysis. Data collection took place in January 2023.

Description of Intervention

eCHECKUP TO GO is a brief, educational, web-based intervention designed to minimize alcohol misuse by providing personalized feedback. The program includes two components. First, participants complete a self-assessment, in which they are prompted to report their demographic characteristics (eg, age, gender, height, and weight), current alcohol consumption and behaviors (eg, frequency and quantity of drinking and driving while drinking), and alcohol-related cognitions (eg, reasons for drinking, such as “alcohol helps me reduce stress”). Participants also describe their goals and priorities, such as the importance

XSL•FO
RenderX

of their fitness levels and relationships with others. The assessment also prompts participants to estimate the alcohol consumption behaviors of individuals in their peer groups (ie, people of the same age, gender, and occupation). Once the assessment is completed, participants are directed to a series of modules that provide personalized feedback based on the responses provided during the assessment. The feedback modules include a range of information. For instance, one module addresses perceived peer norms by showing how the participants' alcohol consumption compares with that of their peers. Other modules show participants how much money they may spend on alcohol, their risk of becoming a problem drinker, and the potentially negative consequences of continued drinking based on their current use. Additional modules provide suggestions for how participants might change their habits and cognitions to reduce their risk for alcohol misuse and better achieve their self-reported goals (eg, alternate alcoholic drinks with water and set a drinking limit before going out). The version of eCHECKUP TO GO used in this study was adapted for use in military populations by incorporating service-specific terminology and imagery. Additionally, the peer normative feedback provided to participants was based on alcohol use data reported by Airmen in the most recent DoD Health Related Behaviors Survey [1], a trusted source used to inform military programming and resource allocation decisions. The program takes 20‑30 minutes to complete. Figures 1 and 2 show examples of program feedback.

Figure 1. eCHECKUP TO GO example alcohol misuse risk feedback.

XSL•FO
RenderX

Figure 2. eCHECKUP TO GO example peer normative feedback.

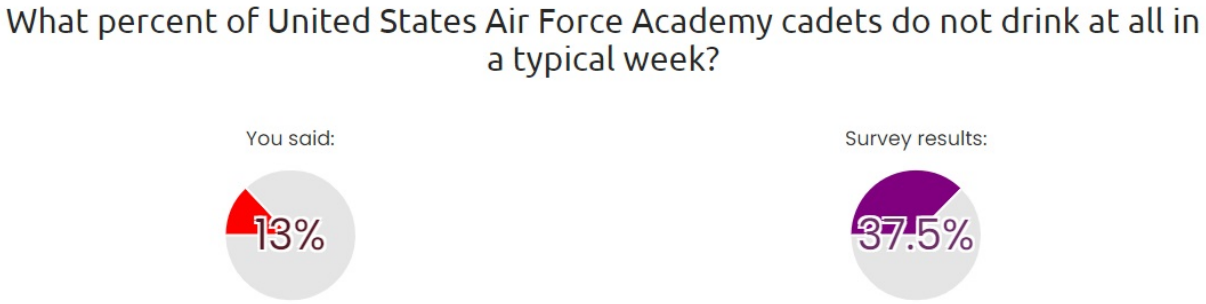

Ethical Considerations

All study procedures and measurement tools were reviewed and approved by the Institutional Review Board at San Diego State University (protocol HS-2022 - 0009) and the USAFA Human Research Protection Program and Cadet Wing. Participants provided written consent to participate, and only those who did so were invited to complete the study activities. Survey data were collected without personal identifiers and only anonymized data were used for the study; all data were stored in a secured drive with access limited to approved study personnel. Participants received no compensation for participating.

Measures

Survey

The self-report survey assessed participant demographic characteristics and satisfaction with various aspects of the eCHECKUP TO GO program, including the intended utility of the information. Sociodemographic characteristics included gender, ethnoracial identity, and age.

Satisfaction With eCHECKUP TO GO Content and User Experience

An abbreviated 11-item version of the general Web trust questionnaire [33-35] assessed perceptions of aspects of the program known to be important to individuals' reviews of web-based health information: credibility of information (ie, "The advice seemed credible" and "I trusted the information on the site"), quality of the information and content (ie, "The site had a professional design"), and personalization of information (ie, "It felt like the advice was tailored to me personally"). Response options ranged from 1=strongly disagree to 7=strongly agree. Three original items designed by the researchers were added to obtain more specific information about participants' perceptions of the program content. Two were included to assess satisfaction with the alcohol-related content (eg, "the feedback I received about alcohol use"), and another to assess satisfaction with the military content (ie, "amount of information about cadets or military service members"). Participants rated their degree of satisfaction on a scale from 1=completely dissatisfied to 7=completely satisfied. In addition to the rating scales, participants were asked to respond to two open-ended questions: "What did you like about the program?" and "How do you think the program could be improved to best help others?"

Use of Information

The intended outcome of eCHECKUP TO GO is to inspire behavioral changes related to alcohol, so participants were also asked at the conclusion of the program whether they intended to act on the advice they received related to alcohol consumption (yes or no). If yes, they were asked to explain in a free-response field which of the actions or strategies they intended to use, thereby providing an additional indicator of program satisfaction and the perceived relevance of the information [33].

Focus Group Guide

The focus group guide included nine questions that assessed participants' perceptions of the eCHECKUP TO GO program. Participants were prompted to provide their opinions on the aesthetics, functional features, and content of the program (eg, "Tell me what you thought about the look and feel of the program" and "How relevant did you find the recommendations were to you?"). Participants were also asked for general feedback (ie, "What else can we do to improve the eCHECKUP TO GO program?").

Analysis

Quantitative survey data and qualitative data were first analyzed separately. Descriptive statistics, including means and frequencies, were computed for all close-ended survey items to assess the distribution and completeness of responses. Open-ended survey questions and focus group data were combined and analyzed using the rigorous and accelerated data reduction (RADaR) technique [36]. The RADaR technique was selected to provide expedited feedback about necessary program adaptations to be implemented before making eCHECKUP TO GO more widely available to cadets [37]. Using RADaR, two coauthors (ES and LH) engaged in an iterative process of data reduction and coding to identify pertinent themes related to the following study research questions: (1) What aspects of eCHECKUP TO GO were participants most satisfied with? and (2) How could eCHECKUP TO GO be improved for use among cadets? Once a codebook was finalized, two analysts (ES and LH) independently reviewed all study data and compared results to identify potential discrepancies. Few discrepancies were identified and all were resolved through discussion. Study analyses were completed by integrating quantitative and qualitative data to examine their convergence and complementarity, with results shown in a joint display [38,39]. Data collection and analyses were led by researchers without

XSL•FO
RenderX

firsthand military service experience. The lead analyst (ES) is a former military beneficiary and the second analyst is a civilian (LH); both have experience conducting military health research.

Results

Overview

The survey participants were 22 cadets. A slight majority (n=12, 54%) were male. The mean age was 19.64 (SD 1.86; range 18 - 26) years, with 6 (27%) participants aged 21 years or older. Approximately, 40% (n=9) of participants identified as non-Hispanic White, approximately one-fifth (n=4, 18%) of participants identified as Hispanic or Latinx or multiracial, 14% (n=3) of participants identified as Black or African American, and 4% (n=1) of participants identified as Asian Pacific Islander or Asian. When compared to the cadet population at large, this sample includes proportionally more female and individual participants from minoritized ethnoracial groups [40]. About one-quarter (n=6, 27%) also participated in the focus group. Demographic characteristics were not recorded for focus group participants to preserve their anonymity; thus, it is not possible to determine whether they significantly differed from the larger group. Further, participants were not asked to report their current or prior alcohol use behaviors because alcohol use is not permitted in this population or setting.

Results of the survey and focus group are jointly displayed in Table 1. Through analysis of the qualitative data, the research team identified three overarching themes addressed by participants: design and functionality, credibility of content, and tailoring and relevance of content. Further, it was observed that participant satisfaction with these aspects of eCHECKUP TO GO differed, with more positive responses provided about the design and functionality than the other themes. A review of the quantitative survey responses revealed a parallel pattern, such that items pertaining to design and functionality were rated highest, while items pertaining to tailoring and relevance of content were rated lowest. A summary of the feedback provided by participants is presented here by theme.

XSL•FO
RenderX

Table . Joint display of quantitative and qualitative responses demonstrating participant perception of the eCHECKUP TO GO program.

Theme	Survey responses and exemplifying quotations from open-ended survey items and focus groups
Design and functionality	• Exemplifying quotations: satisfaction • "The program was very accessible and easy to understand in terms of direction." [SP[a] #19] • "I liked how interactive the overall site was and how it gave me information based on my inputs." [SP #1] • Exemplifying quotations: suggestions for improvement • "I think the program is overall very good. Some of the questions I misinterpreted so having like a small description below with a better or more explanation of the question could probably help." [SP #7]
Credibility of content	• Exemplifying quotations: satisfaction • "I appreciated the stats." [SP #21] • "I like that it gave a lot of helpful information on drinking and how to be a safer drinker." [SP #16] • "Yeah, I think it was very feasible. I guess, having the discipline to do it." [FGP[b] #3] • Exemplifying quotations: suggestions for improvement • "Provide more information and stats on alcohol use in the military, specifically cadets." [SP #21]
Tailoring and relevance of content	• Exemplifying quotations: satisfaction • "Yeah, I liked how if you didn't put yourself down as drinking. It would, it was like, here's things you can do to recommend to others instead of you need to do these things." [FGP #2] • "I think by asking us to input what we thought the statistics were and what was important to us made it much more personal." [SP #12] • Exemplifying quotations: Suggestions for improvement • "You had all this stuff like with the airmen and cadets, but when you went to the consequences, it was largely just like social and physical consequences, maybe like help emphasize like the military aspect of it ... Just to reinforce that like drinking can ruin your career, which is what we're all here to set up for." [FGP #2] • "Provide more military-tailored advice ([it's a] different drinking culture here than civilian schools)." [SP #10]

[a]SP: survey participant.

[b]FGP: focus group participant.

Design and Functionality

On the survey, participants were most satisfied with the ease of use (on the 1 - 7 scale: mean 6.64, SD 0.73), how easy it was to understand the information (mean 6.45, SD 0.91), and the design and interactivity (mean 6.55, SD 0.60; and mean 6.45, SD 1.01, respectively). The qualitative data additionally elucidated the design features of the program most liked by participants. Several survey respondents commented on the features of the program that made it easy to use. For example, one participant wrote, "I thought it was clean and easy to follow and did not have unnecessary filler" [survey participant (SP) #17]. In the focus group, a participant who appreciated the embedded skip logic that allowed participants to advance throughout the assessment based on their responses said, "I liked in the parts where it was like kind of the yes or no that would just automatically move down to the next question instead of like having to scroll after answering each one" [focus group participant (FGP) #4]. Others commented on the interactivity of the program ("I liked how interactive the overall site was and how it gave me information based on my inputs" [SP #3]).

When it came to suggesting improvements in the design and functionality, most participants commented on the

XSL•FO
RenderX

self-assessment component of the program. Several participants described difficulty completing the drinking history questions, which are structured in the format of a calendar where respondents enter the number and type of drinks they consumed on different days of the week and during certain hours. In the focus group, two participants described difficulty with the calendar, stating that "the calendar was too stressful" [FGP #2] and "I don't pay attention to, like, how many drinks I had an hour or in the specific hours, which like, how long I've been drinking" [FGP #3].

Credibility of Content

Survey responses revealed a high degree of trust in the information and suggestions provided in eCHECKUP TO GO, as indicated by the high scores for credibility of information (mean 6.36, SD 0.66) and feasibility of alcohol misuse prevention strategies provided (mean 6.41, SD 0.79). The qualitative results helped to explain the highest ratings, as the participant responses revealed an appreciation for data, which are frequently presented to eCHECKUP TO GO users. One respondent noted, "I appreciated the stats" [SP #21], while others remarked, "I liked how it showed graphs and other ways for people to see how those products affect themselves and their peers" [SP #4], and "[I liked the] statistics at the end, as well as the estimation questions to display knowledge" [SP #10]. Others noted the specific information they recalled, such as "I really liked the part where it told you about like the calories and like how it can impact muscle mass, especially because physicality is such a big thing here that if, like, if people are drinking, they just don't realize the calories and how it can hurt you in your muscles" [FGP #2]. Further showing the importance of data to participants, the suggestions provided related to the credibility of content were to add even more data ("Maybe more data and statistics" [SP #12]).

Tailoring and Relevance of Content

As shown in Table 1, the scores for tailoring and relevance of eCHECKUP TO GO content were relatively lower (ie, amount of information about cadets: mean 5.48, SD 1.57). Qualitative responses provided context for these lower scores, showing that while many participants appreciated the personalized feedback they received, they felt that the content did not include enough information tailored to the unique experiences of military cadets. In terms of the personalized feedback, one participant noted that what they liked best about the program was it was "very personally [sic] to me, showed me the statistics, how much I am at risk" [SP #22], while another said, "I think personally it made me more aware of how much I was drinking or how much others are drinking ... it can make us more aware of our habits and our spending" [FGP #3].

When describing opportunities to improve the program tailoring, focus group participants discussed a need for more information on the social aspect of drinking. One participant explained:

> *There could be ... a lengthy section for the social aspect of drinking ... a lot of cadets just go out and drink together. And I don't think the survey emphasized that really enough. I think it was individually tailored, which I think is good. But also, it didn't go into the, too much, in my opinion, to the causes of like why people go out and drink, maybe just because of like the social pressure, because again, especially being cadets, there's always that implied social pressure.* [FGP #2]

Another participant also suggested adding more content on peer pressure and drinking culture, saying, "I feel like there should be a section on ... how much of the overall culture, overall climate of drinking contribute to your own and like do others pressure you to drink and stuff like that" [FGP #4]. Relatedly, another cadet requested content about "how to help friends and notice if they have an alcohol problem" [SP #22].

Participants recommended improving content tailoring to address other aspects of cadet or military life as well. Some felt that the existing content of eCHECKUP TO GO did not consider the unique schedules of cadets. They explained that there are periods of time when they are forced to completely abstain from drinking and do not have access to alcohol and then periods where it is almost expected that they would indulge, making it hard to answer the self-assessment items that asked about behaviors in the past week or a "typical" month. According to one participant, "We don't get off base much. So, our drinking habits may look very irrational, making it hard to say I drink x on a typical week" [SP #16]. Other suggestions pertained to the personalized feedback provided about the consequences of drinking. It was suggested that the military-specific consequences be included, which previously were not in the program:

> *When you went to the consequences, it was largely just like social and physical consequences. Maybe, like help emphasize, like, the military aspect of it. [Such as] like this many people a year get put on alcohol probation here [or] like in the actual Air Force, like this many people get in trouble for drinking, and like drinking can lead to these Uniform Code of Military Justice violations and has been linked to them or something like that. Just to reinforce that like drinking can ruin your career, which is what we're all here to set up for.* [FGP #2]

Another noted that the eCHECKUP TO GO program provided feedback about how much money they spent on alcohol compared with a rent payment, which cadets do not pay, adding, "So, we don't pay rent or at least we don't see it ... So, I guess if it's specific to cadets, maybe just don't add that" [FGP #3]. Finally, some requested "more military-tailored advice (different drinking culture here than civilian schools)" [SP #21].

Future Use of Alcohol Misuse Prevention Strategies

Over two-thirds of participants (n=15, 68%) said they intended to use at least one alcohol misuse prevention strategy they learned from eCHECKUP TO GO. Those who answered affirmatively were asked to describe which, if any, of the alcohol misuse prevention strategies learned from eCHECKUP TO GO they would use in the future. Through qualitative analysis, it was determined that responses could be categorized into two salient themes: strategies to reduce the amount consumed and strategies to protect others. Several participants planned to reduce their intake by spacing out their drinks, saying, "I intend

XSL•FO
RenderX

to use the 'spacing drinks over time' strategy" [SP #3] and "When reaching the legal drinking age, to practice safe drinking habits like spacing out drinks and setting limits for myself" [SP #20]. Additionally, numerous participants planned to use strategies to protect their peers, such as one participant who explained, "Even if I am not drinking and don't have to worry about putting myself in danger, I can look out for my friends when they are drinking and make sure they are making good decisions regarding alcohol" [SP #1]. Others indicated that they would act as the designated driver for their friends, stating, "I will likely be the DD for my group on more occasions" [SP #17].

Discussion

Principal Findings

Addressing the longstanding issue of alcohol misuse in the military requires creative solutions designed to mitigate the challenges inherent to promoting health in large, diverse communities [1,2,4,11,41]. This study sought to determine whether a military-tailored version of eCHECKUP TO GO is an acceptable option for use among military cadets, recruits, or trainees. Study results identified aspects of the program that were well received by participants, including the web-based format and interface, and the personalized feedback about alcohol risk and protective strategies. Importantly, no concerns were expressed by participants about confidentiality. These findings suggest that this internet-based, individually completed BAI may be acceptable in this community, providing a possible solution to the logistical issues and privacy concerns that often impede the implementation and efficacy of health promotion programs in military settings.

Although eCHECKUP TO GO was largely endorsed by participants, opportunities for improvement were identified. Though participants liked the self-assessment and personalized feedback, many felt that these modules did not sufficiently acknowledge the broader social context impacting alcohol use, particularly cultural norms and peer influence. Heavy alcohol use is widely known to be a norm within the military and among young adults, and research has shown that consuming alcohol with peers may be viewed as an opportunity to build cohesion [41-46]. Participants touched on these influences when they described situations in which they felt expected to drink or chose to drink to avoid feeling left out. These findings align with a study by Meadows et al [43] that found positive associations between perceived alcohol culture within respondents' command or unit and binge and heavy drinking. While alcohol use is an individual behavior, it cannot be disentangled from the social context in which it occurs when working with military cadets or trainees [43]. Thus, future alcohol misuse prevention interventions may be more acceptable and efficacious if they directly address the impact of peer influence and provide concrete strategies for navigating social situations in which alcohol may be present.

Study results also highlight another cultural norm that should be incorporated into eCHECKUP TO GO and other alcohol misuse prevention programming: service before self. "Service before self" is a core value of the Air Force [47], and the other service branches uphold similar maxims. Feedback from study participants indicates that their ingrained commitment to service should be used as a cornerstone of alcohol misuse prevention content. For example, participants who intended to act upon the advice received as part of the program largely intended to use strategies to protect their comrades, and some asked for more content on how to help recognize alcohol misuse in others. Further, participants felt the program could be more impactful if it included information about career consequences of alcohol misuse, such as the Uniform Code of Military Justice violations. These findings parallel results from a study by Aycock et al [48], which also noted that alcohol misuse prevention messaging for Airmen should emphasize adverse career impacts. Framing alcohol misuse interventions as a means of protecting comrades and fulfilling a commitment to serve could bolster buy-in and work to dismantle longstanding alcohol-related cultural norms [41-43].

Limitations

Methodological limitations must be considered when interpreting the results of this study. Due to the relatively small sample size and unique population, the results may not generalize to the USAFA community or other military service academies. Additional formative research may be needed to finalize intervention refinement before subsequent efficacy trials. In particular, more data may be needed to ensure the experiences and preferences of participants from minoritized communities who have historically been underrepresented in research are considered. Additionally, while no concerns about anonymity were expressed by participants, it remains possible that participants did not feel comfortable discussing their opinions about alcohol use out of fear of accidental disclosure of their identity. However, the responses received likely reflect participants' opinions and behaviors because participants were given the option of skipping questions they did not wish to answer. Finally, as this was a formative evaluation, the results cannot be used to estimate the efficacy of eCHECKUP TO GO completion on alcohol misuse among cadets. The next steps in this line of research should include an efficacy trial powered sufficiently to determine the impact of eCHECKUP TO GO completion on attitudes toward alcohol misuse, alcohol-related protective behavioral strategies, and alcohol consumption among military cadets.

Conclusions

The results of this study indicate that a military version of eCHECKUP TO GO may be suitable for use in military training settings, though additional tailoring is needed to account for the unique norms and values of this community. Study findings have important implications for the development and implementation of military alcohol misuse prevention programs, but the lessons learned can also be applied to programs that address other health behaviors. In communities that value service above self, program content should emphasize the ways in which improving individual health behaviors can benefit the community at large.

XSL•FO
RenderX

Acknowledgments

This work was sponsored by the US Government under Other Transactions number W81XWH-21-9-0011.012, Award W81XWH-21-9-0011.012. The US Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the US Government. WKR and CMSA are employees of the US Government. This work was prepared as part of their official duties. Title 17, U.S.C. §105 provides that copyright protection under this title is not available for any work of the US Government. Title 17, U.S.C. §101 defines US Government work as work prepared by a military service member or employee of the US Government as part of that person's official duties. This work was supported by the Defense Health Agency Medical Technology Enterprise Consortium under work unit 61135. The views expressed in this work are those of the authors and do not necessarily reflect the official policy or position of the Department of the Navy, Department of Defense, or the US Government.

Data Availability

Data are not available for public use because participants did not provide consent for the data to be used for additional research purposes.

Authors' Contributions

MK, SMB, ES, WKR, MR, LW, and RM conceptualized the project. ES, LH, and CMSA collected all project data. ES and LH analyzed project data. ES led the development of the manuscript; all other authors contributed to editing.

Conflicts of Interest

LW and RM may earn royalties for the use of the eCHECKUP TO GO Program.

References

1. Meadows SO, Engel CC, Collins RL, et al. 2015 department of defense health related behaviors survey (HRBS). Rand Health Q 2018;8:5. [Medline: 30323988]
2. Armed Forces Health Surveillance Division. Ambulatory visits among active component members, U.S. Armed Forces, 2022. MSMR 2023 Jun 20;30(6):19-25. [Medline: 37540151]
3. Armed Forces Health Surveillance Division. Hospitalizations among active component members, U.S. Armed Forces, 2022. MSMR 2023 Jun 20;30(6):12-18. [Medline: 37540152]
4. Armed Forces Health Surveillance Division. Absolute and relative morbidity burdens attributable to various illnesses and injuries, active component, U.S. Armed Forces, 2021. MSMR 2022 Jun 1;29(6):2-9. [Medline: 36250737]
5. Drinking levels defined. National Insitute on Alcohol Abuse and Alcoholism. 2023. URL: https://www.niaaa.nih.gov/alcohol-health/overview-alcohol-consumption/moderate-binge-drinking#:~:text=NIAAA%20defines%20binge%20drinking%20as,)%2C%20in%20about%202%20hours [accessed 2024-06-01]
6. Mattiko MJ, Olmsted KLR, Brown JM, Bray RM. Alcohol use and negative consequences among active duty military personnel. Addict Behav 2011 Jun;36(6):608-614. [doi: 10.1016/j.addbeh.2011.01.023] [Medline: 21376475]
7. Porter B, Rodriguez LM, Woodall KA, Pflieger JC, Stander VA. Alcohol misuse and separation from military service: a dyadic perspective. Addict Behav 2020 Nov;110:106512. [doi: 10.1016/j.addbeh.2020.106512] [Medline: 32623237]
8. Waller M, McGuire ACL, Dobson AJ. Alcohol use in the military: associations with health and wellbeing. Subst Abuse Treat Prev Policy 2015 Jul 28;10:27. [doi: 10.1186/s13011-015-0023-4] [Medline: 26216215]
9. Norman SB, Schmied E, Larson GE. Predictors of continued problem drinking and substance use following military discharge. J Stud Alcohol Drugs 2014 Jul;75(4):557-566. [doi: 10.15288/jsad.2014.75.557] [Medline: 24988254]
10. Derefinko KJ, Hallsell TA, Isaacs MB, et al. Substance use and psychological distress before and after the military to civilian transition. Mil Med 2018 May 1;183(5-6):e258-e265. [doi: 10.1093/milmed/usx082] [Medline: 29420791]
11. Haibach JP, Haibach MA, Hall KS, et al. Military and veteran health behavior research and practice: challenges and opportunities. J Behav Med 2017 Feb;40(1):175-193. [doi: 10.1007/s10865-016-9794-y] [Medline: 27678001]
12. Hoerster KD, Lehavot K, Simpson T, McFall M, Reiber G, Nelson KM. Health and health behavior differences: U.S. military, veteran, and civilian men. Am J Prev Med 2012 Nov;43(5):483-489. [doi: 10.1016/j.amepre.2012.07.029] [Medline: 23079170]
13. Shirvani M, Reed MB, Clingan S. The relationship between emerging adult alcohol consumption and military enlistment. Mil Med 2017 Sep;182(9):e1731-e1737. [doi: 10.7205/MILMED-D-16-00312] [Medline: 28885929]
14. Derefinko KJ, Klesges RC, Bursac Z, Little MA, Hryshko-Mullen A, Talcott GW. Alcohol issues prior to training in the United States Air Force. Addict Behav 2016 Jul;58:142-148. [doi: 10.1016/j.addbeh.2016.02.028] [Medline: 26945450]

XSL•FO
RenderX

15. Taylor JE, Haddock K, Poston WSC, Talcott WG. Relationship between patterns of alcohol use and negative alcohol-related outcomes among U.S. Air Force recruits. Mil Med 2007 Apr;172(4):379-382. [doi: 10.7205/milmed.172.4.379] [Medline: 17484307]
16. Brief alcohol interventions for unhealthy alcohol use. : Department of Defense Psychological Health Center of Excellence; 2023 URL: https://health.mil/Reference-Center/Publications/2021/04/26/PHCoE-Evidence-Brief-Alcohol-Interventions-for-Unhealthy-Alcohol-Use-508 [accessed 2025-03-28]
17. Va/dod clinical practice guidelines for the management of substance use disorder (2021). : U.S. Department of Veterans Affairs; 2021.
18. Doherty AM, Mason C, Fear NT, Rona R, Greenberg N, Goodwin L. Are brief alcohol interventions targeting alcohol use efficacious in military and veteran populations? A meta-analysis. Drug Alcohol Depend 2017 Sep 1;178:571-578. [doi: 10.1016/j.drugalcdep.2017.05.029] [Medline: 28750345]
19. Klesges RC, Talcott W, Ebbert JO, et al. Effect of the Alcohol Misconduct Prevention Program (AMPP) in air force technical training. Mil Med 2013 Apr;178(4):445-451. [doi: 10.7205/MILMED-D-12-00400] [Medline: 23707832]
20. Little MA, Klesges RC, Mallawaarachchi I, et al. Prevention of alcohol-related incidents in the U.S. Air Force: results from a cluster randomized trial. Mil Med 2023 Jul 22;188(7-8):e1874-e1881. [doi: 10.1093/milmed/usab489]
21. Derefinko KJ, Linde BD, Klesges RC, et al. Dissemination of the brief alcohol intervention in the United States Air Force: study rationale, design, and methods. Mil Behav Health 2018 Jan 2;6(1):108-117. [doi: 10.1080/21635781.2017.1397569]
22. Li T, Waters TM, Kaplan EK, et al. Economic analyses of an alcohol misconduct prevention program in a military setting. Mil Med 2017 Jan;182(1):e1562-e1567. [doi: 10.7205/MILMED-D-16-00098] [Medline: 28051974]
23. Griffin KW, Williams C, Travis W, Tharp A. Prevention of unwanted sexual contact among cadets at the United States Air Force academy: a brief small-group intervention. Am J Public Health 2021 Apr;111(4):672-674. [doi: 10.2105/AJPH.2020.306050] [Medline: 33476231]
24. Pedersen ER, Parast L, Marshall GN, Schell TL, Neighbors C. A randomized controlled trial of A web-based, personalized normative feedback alcohol intervention for young-adult veterans. J Consult Clin Psychol 2017 May;85(5):459-470. [doi: 10.1037/ccp0000187] [Medline: 28287799]
25. Hurtado S, Shaffer RA, Schuckit MA, et al. Evaluation of an alcohol misuse prevention program in a military population. : Naval Health Research Center; 2003 URL: https://apps.dtic.mil/sti/tr/pdf/ADA421218.pdf [accessed 2025-03-28]
26. Pemberton MR, Williams J, Herman-Stahl M, et al. Evaluation of two web-based alcohol interventions in the U.S. military. J Stud Alcohol Drugs 2011 May;72(3):480-489. [doi: 10.15288/jsad.2011.72.480] [Medline: 21513685]
27. Hester RK, Squires DD, Delaney HD. The Drinker's Check-up: 12-month outcomes of a controlled clinical trial of a stand-alone software program for problem drinkers. J Subst Abuse Treat 2005 Mar;28(2):159-169. [doi: 10.1016/j.jsat.2004.12.002] [Medline: 15780546]
28. Doumas DM, Esp S, Turrisi R, Hausheer R, Cuffee C. A test of the efficacy of a brief, web-based personalized feedback intervention to reduce drinking among 9th grade students. Addict Behav 2014 Jan;39(1):231-238. [doi: 10.1016/j.addbeh.2013.10.011] [Medline: 24148137]
29. Doumas DM, Haustveit T, Coll KM. Reducing heavy drinking among first year intercollegiate athletes: a randomized controlled trial of web-based normative feedback. J Appl Sport Psychol 2010 Jul 23;22(3):247-261. [doi: 10.1080/10413201003666454]
30. Doumas DM, Mastroleo NR. Reducing alcohol use among high school athletes: a randomized controlled trial of a brief, web-based personalized normative feedback intervention. Sport Psychol 2021 Mar;35(1):55-63. [doi: 10.1123/tsp.2020-0022] [Medline: 38962423]
31. Ganz T, Braun M, Laging M, Schermelleh-Engel K, Michalak J, Heidenreich T. Effects of a stand-alone web-based electronic screening and brief intervention targeting alcohol use in university students of legal drinking age: a randomized controlled trial. Addict Behav 2018 Feb;77:81-88. [doi: 10.1016/j.addbeh.2017.09.017] [Medline: 28985586]
32. Doumas DM, Esp S, Turrisi R, Bond L. A randomized controlled trial of the eCHECKUP to GO for high school seniors across the academic year. Subst Use Misuse 2021;56(13):1923-1932. [doi: 10.1080/10826084.2021.1958862] [Medline: 34347564]
33. Briggs P, Burford B, De Angeli A, Lynch P. Trust in online advice. Soc Sci Comput Rev 2002;20(3):321-332. [doi: 10.1177/08939302020003009]
34. Harris PR, Sillence E, Briggs P. Perceived threat and corroboration: key factors that improve a predictive model of trust in internet-based health information and advice. J Med Internet Res 2011 Jul 27;13(3):e51. [doi: 10.2196/jmir.1821] [Medline: 21795237]
35. Sillence E, Blythe JM, Briggs P, Moss M. A revised model of trust in internet-based health information and advice: cross-sectional questionnaire study. J Med Internet Res 2019 Nov 11;21(11):e11125. [doi: 10.2196/11125] [Medline: 31710297]
36. Watkins DC. Rapid and rigorous qualitative data analysis: the "RADaR" technique for applied research. Int J Qual Methods 2017;16(1). [doi: 10.1177/1609406917712131]

XSL•FO
RenderX

37. Gale RC, Wu J, Erhardt T, et al. Comparison of rapid vs in-depth qualitative analytic methods from a process evaluation of academic detailing in the Veterans Health Administration. Implement Sci 2019 Feb 1;14(1):11. [doi: 10.1186/s13012-019-0853-y] [Medline: 30709368]
38. Palinkas LA, Aarons GA, Horwitz S, Chamberlain P, Hurlburt M, Landsverk J. Mixed method designs in implementation research. Adm Policy Ment Health 2011 Jan;38(1):44-53. [doi: 10.1007/s10488-010-0314-z] [Medline: 20967495]
39. Guetterman TC, Fàbregues S, Sakakibara R. Visuals in joint displays to represent integration in mixed methods research: a methodological review. Methods Psychol 2021 Dec;5:100080. [doi: 10.1016/j.metip.2021.100080]
40. Demographic profile of the incomine USAFA class of 2025. Department of Defense, United States Air Force Academy. URL: https://www.usafa.edu/app/uploads/CL2025-Class-Profile.pdf [accessed 2025-02-07]
41. Osborne AK, Wilson-Menzfeld G, McGill G, Kiernan MD. Military service and alcohol use: a systematic narrative review. Occup Med (Lond) 2022 Jul 11;72(5):313-323. [doi: 10.1093/occmed/kqac045] [Medline: 35674143]
42. Teachman J, Anderson C, Tedrow LM. Military service and alcohol use in the United States. Armed Forces Soc 2015 Jul;41(3):460-476. [doi: 10.1177/0095327X14543848]
43. Meadows SO, Beckman R, Engel CC, Jeffery DD. The culture of alcohol in the U.S. military: correlations with problematic drinking behaviors and negative consequences of alcohol use. Armed Forces Soc 2023 Apr;49(2):531-555. [doi: 10.1177/0095327X211069162]
44. Knox J, Schneider J, Greene E, Nicholson J, Hasin D, Sandfort T. Using social network analysis to examine alcohol use among adults: a systematic review. PLoS ONE 2019;14(8):e0221360. [doi: 10.1371/journal.pone.0221360] [Medline: 31437257]
45. Ames GM, Duke MR, Moore RS, Cunradi CB. The impact of occupational culture on drinking behavior of young adults in the U.S. Navy. J Mixed Methods Res 2009 Apr;3(2):129-150. [doi: 10.1177/1558689808328534]
46. Besse K, et al. How soldiers perceive the drinking environment in communities near military installations. J Alcohol Drug Educ 2018;62(1):71-90 [FREE Full text]
47. Vision: it's our promise to protect. US Air Force. URL: https://www.airforce.com/vision [accessed 2024-06-15]
48. Aycock CA, Mallawaarachchi I, Wang XQ, et al. Developing a text messaging intervention to prevent binge and heavy drinking in a military population: mixed methods development study. JMIR Form Res 2024 Mar 19;8:e55041. [doi: 10.2196/55041] [Medline: 38502165]

Abbreviations

AMPP: Alcohol Misconduct Prevention Program
BAI: brief alcohol intervention
DoD: Department of Defense
FGP: focus group participant
RADaR: rigorous and accelerated data reduction
SP: survey participant
USAFA: US Air Force Academy

Edited by A Mavragani; submitted 16.10.24; peer-reviewed by M Choi, T Bakas; revised version received 10.02.25; accepted 11.02.25; published 16.04.25.

Please cite as:
Schmied E, Hurtado L, Robinson WK, Simon-Arndt CM, Moyer III R, Wilson L, Reed M, Blakey SM, Kan M
Evaluating the Acceptability of a Brief Web-Based Alcohol Misuse Prevention Program Among US Military Cadets: Mixed Methods Formative Evaluation
JMIR Form Res 2025;9:e67637
URL: https://formative.jmir.org/2025/1/e67637
doi:10.2196/67637

© Emily Schmied, Lauren Hurtado, W Ken Robinson, Cynthia M Simon-Arndt, Richard Moyer III, Leslie Wilson, Mark Reed, Shannon M Blakey, Marni Kan. Originally published in JMIR Formative Research (https://formative.jmir.org), 16.4.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Impact of a Virtual Reality Video ("A Walk-Through Dementia") on YouTube Users: Topic Modeling Analysis

Xiaoli Li[1], PhD; Xiaoyu Liu[2], PhD; Cheng Yin[3], PhD; Sandra Collins[1], PhD; Eman Alanazi[4], PhD

[1]School of Health Sciences, Southern Illinois University Carbondale, 1365 Douglas Drive, ASA Building - Mail Code 6615, Carbondale, IL, United States

[2]College for Public Health & Social Justice, Saint Louis University, Saint Louis, MO, United States

[3]Department of Rehabilitation and Health Services, University of North Texas, Denton, TX, United States

[4]Health Informatics Department, College of Health Sciences, Saudi Electronic University, Riyadh, Saudi Arabia

Corresponding Author:
Xiaoli Li, PhD
School of Health Sciences, Southern Illinois University Carbondale, 1365 Douglas Drive, ASA Building - Mail Code 6615, Carbondale, IL, United States

Abstract

Background: Emerging research has highlighted the potential of virtual reality (VR) as a tool for training health care students and professionals in care skills for individuals with Alzheimer disease and related dementias (ADRD). However, there is limited research on the use of VR to engage the general public in raising awareness about ADRD.

Objective: This research aimed to examine the impact of the VR video “A Walk-Through Dementia” on YouTube users by analyzing their posts.

Methods: We collected 12,754 comments from the VR video series “A Walk-Through Dementia,” which simulates the everyday challenges faced by individuals with ADRD, providing viewers with an immersive experience of the condition. Topic modeling was conducted to gauge viewer opinions and reactions to the videos. A pretrained Bidirectional Encoder Representations from Transformers (BERT) model was used to transform the YouTube comments into high-dimensional vector embeddings, allowing for systematic identification and detailed analysis of the principal topics and their thematic structures within the dataset.

Results: We identified the top 300 most frequent words in the dataset and categorized them into nouns, verbs, and adjectives or adverbs using a part-of-speech tagging model, fine-tuned for accurate tagging tasks. The topic modeling process identified eight8 initial topics based on the most frequent words. After manually reviewing the 8 topics and the content of the comments, we synthesized them into 5 themes. The predominant theme, represented in 2917 comments, centered on users’ personal experiences with the impact of ADRD on patients and caregivers. The remaining themes were categorized into 4 main areas: positive reactions to the VR videos, challenges faced by individuals with ADRD, the role of caregivers, and learning from the VR videos.

Conclusions: Using topic modeling, this study demonstrated that VR applications serve as engaging and experiential learning tools, offering the public a deeper understanding of life with ADRD. Future research should explore additional VR applications on social media, as they hold the potential to reach wider audiences and effectively disseminate knowledge about ADRD.

(JMIR Form Res 2025;9:e67755) doi:10.2196/67755

KEYWORDS

Alzheimer disease; Alzheimer disease and related dementias; ADRD; dementia; public awareness; text mining; older adult; health care student; training; health care professional; simulation; digital world; virtual environment; virtual tour; walk-through dementia; virtual reality; VR; VR video; VR application; topic modeling; YouTube; Bidirectional Encoder Representations from Transformers; BERT; social media comments; experiential learning tool

Introduction

With a rapidly aging population, the number of older adults living with chronic conditions is on the rise. Among these conditions, Alzheimer disease and related dementias (ADRD) have become leading causes of disability and dependency worldwide, characterized by memory loss and other cognitive impairments [1]. In the later stages, individuals with ADRD may lose the ability to interact with their surroundings, making care management increasingly complex both within health care settings and at home [2]. This condition not only affects the quality of life for those diagnosed but also places a significant burden on health care systems and families. While most older adults prefer to stay in their homes as long as possible, many eventually require higher levels of care, often transitioning to nursing homes or memory care facilities [3]. Caregivers—whether family members or professionals—must

XSL•FO
RenderX

have a thorough understanding of ADRD to provide the best care possible. In addition, various policy initiatives, such as those promoting the creation of "dementia-friendly communities" [4], underscore the importance of raising public awareness about dementia. These communities aim to foster environments where people understand ADRD and where individuals with dementia feel included, involved, and empowered to maintain control over their daily lives [5]. Such initiatives emphasize the need for widespread dementia awareness and education within communities to improve the support and care provided to those affected by ADRD.

However, there is a lack of public awareness and understanding of what ADRD is and its impact on individuals' abilities, well-being, and capacity to live meaningful lives [6]. As a result, misunderstanding and discrimination among people living with ADRD are common. For instance, there are false beliefs that people with ADRD cannot experience quality of life, are entirely dependent on others, and have lost all autonomy and dignity [6]. Such stigma and misunderstandings can have negative effects on the psychosocial well-being, self-esteem, and overall quality of life for both those living with ADRD and their caregivers. In addition, stigma can contribute to social exclusion, isolation, and even self-stigmatization, further exacerbating the challenges faced by individuals with ADRD and their families [7].

Disseminating reliable information about ADRD is especially crucial because addressing modifiable factors can help reduce misunderstandings, provide support to caregivers, and raise awareness about the condition [8,9]. Social media platforms, particularly YouTube, have become powerful tools for disseminating health-related information to audiences that are not easily accessible [10]. With 73% of middle-aged adults and 45% of older adults in the United States using social media such as Facebook (Meta) and YouTube, these platforms are vital tools for raising awareness, educating communities, and deepening public understanding of ADRD [11]. They not only facilitate the sharing of personal stories but also provide educational content, making them well-positioned to increase knowledge and combat the stigma associated with ADRD [12].

Previous studies have explored various topics involving YouTube and ADRD, such as assessing the quality of ADRD caregiving information on the platform [13], describing social media activities for aging and ADRD education [14], developing internet-based ADRD training modules [15], examining YouTube's effectiveness in dementia education to a target community [16], and exploring the role of personal stories in raising awareness about dementia [17]. In addition, research has examined the role of personal stories in raising awareness about dementia and how social media mobilizes knowledge about pain management in dementia [18]. However, these studies focused on traditional videos as educational tools, and no research has examined the use of virtual reality (VR) videos for public ADRD education.

VR videos provide immersive, 360-degree environments that allow users to engage with simulated worlds using devices like enclosed headsets or body suits [19,20]. In the field of ADRD education, unlike traditional media, VR offers a unique opportunity to simulate the experiences of individuals living with dementia, providing a more visceral understanding of their challenges. While VR has primarily been used to educate health care students and the caregiving workforce about ADRD [21-23], its potential for raising public awareness has not been fully explored. This gap is particularly important, as VR can deliver an engaging and experiential form of learning that may influence attitudes towards ADRD.

This study aims to explore the impact of a free YouTube VR video series titled "A Walk-through Dementia" (AWTD), which simulates the experiences of individuals living with dementia [20]. AWTD consists of 3 videos, each lasting 3 to 5 minutes, and was developed by Alzheimer's Research United Kingdom to enhance understanding of the challenges faced by people with dementia. This research seeks to understand the general public's experience with VR videos on ADRD. By analyzing YouTube comments, the study will explore whether these VR videos can enhance public awareness of ADRD and serve as an effective tool for addressing misunderstandings and stigma surrounding ADRD.

Methods

Overview

We followed a 3-step exploratory research approach to conduct the study. First, we collected comprehensive data by gathering the most viewed VR video series on YouTube. Second, we performed topic modeling in the comments section to identify the main themes discussed. Third, we used Thematic analysis to conduct these themes into concise, easily digestible summaries. The workflow of the methods is outlined in Figure 1 below.

XSL•FO
RenderX

Figure 1. The workflow of the methods. It includes data collection via the YouTube API, preliminary analysis, and topic modeling using BERTopic. The final stage involves interpreting topics by reviewing keywords and synthesizing them into major themes. API: application programming interface; BERT: Bidirectional Encoder Representations from Transformers.

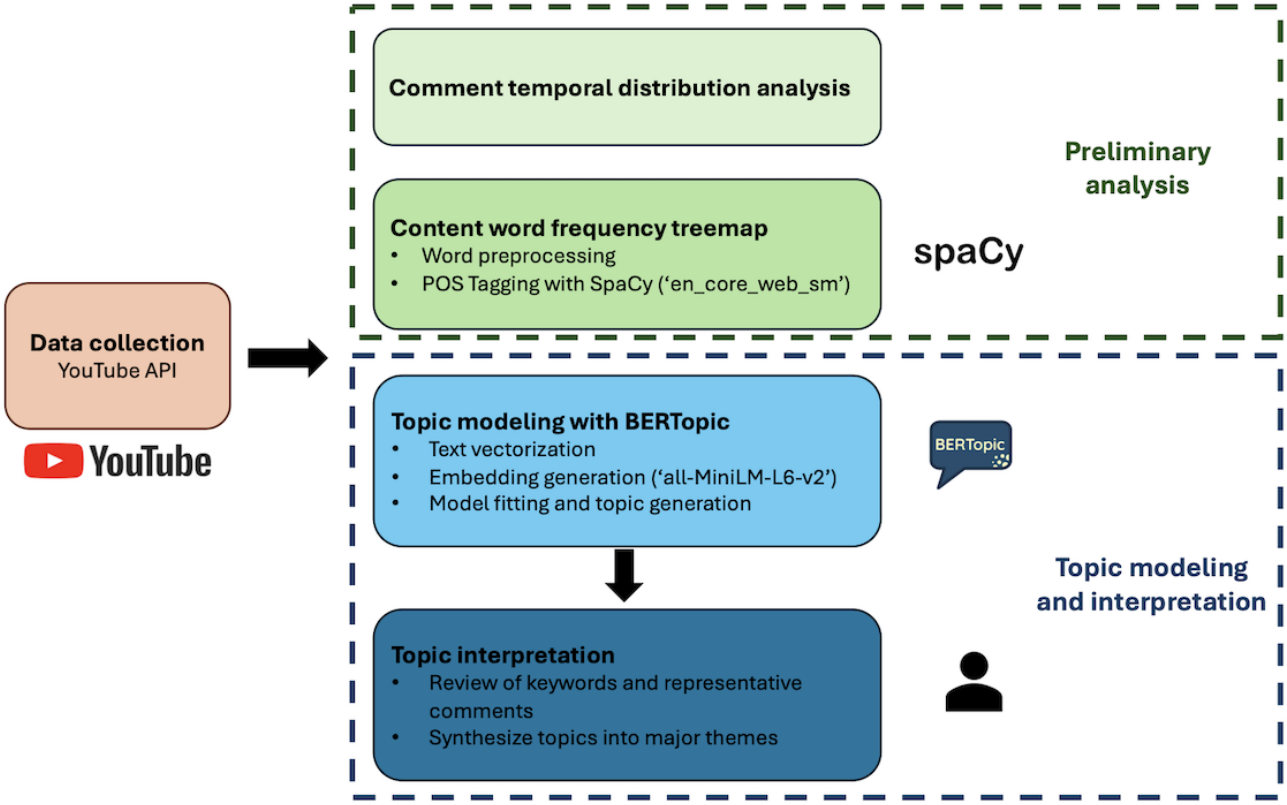

Datasets

The analyzed video comments are from a series of 360° videos titled "A Walk-Through Dementia," the most viewed videos under the search terms "dementia" and "Alzheimer disease." Developed by Alzheimer's Research United Kingdom, this series aims to raise awareness about the impact of dementia on individuals' lives by engaging the public, health care professionals, and caregivers. The series comprises 3 main scenarios depicted in separate videos (Textbox 1) .

Textbox 1. Three main scenarios in "A Walk-Through Dementia."

- At the supermarket: the video demonstrates how dementia affects tasks such as reading shopping lists, dealing with short-term memory loss, changes in food preferences, and managing social anxiety at checkout counters.
- Walking home: the video highlights the difficulties in navigating busy streets, recognizing familiar places and people, and managing visual and spatial misperceptions.
- At home: the video illustrates challenges within a familiar environment, such as difficulty remembering instructions, issues with hand-eye coordination, and managing multi-step tasks like making a cup of tea.

Compared to other VR videos retrieved using the same search terms, the AWTD series has garnered the highest view counts (walking home [WH]: 3.2 million views; at the supermarket [AS]: 2.7 million views; at home [AH]: 0.12 million views) relative to the most viewed videos by other creators (eg, "Coping with Alzheimer's, Together and Apart" with 67 thousand views). In addition, this series has attracted the most comments. To maintain content consistency and consider the low number of comments in other VR videos (eg, "Coping with Alzheimer's, Together and Apart" with 66 comments), we decided to focus solely on the comment sections of the 3 videos in this series.

Preliminary Analysis

To collect the comments for this study, we used the official YouTube application programming interface, retrieving both primary comments and their corresponding replies. This resulted in a total of 12,754 comments. After removing duplicate and irrelevant entries, the final dataset included 11,543 unique comments.

Before conducting topic modeling, we performed a series of preliminary analyses to gain an initial understanding of the content across the 3 videos. These analyses involved examining the temporal distribution of comment postings and calculating word frequencies. The purpose of this step was to identify key content words essential for interpreting the overarching themes in the comments.

To ensure focus on the most meaningful words, we restricted our analysis to content words, specifically nouns, main verbs, adjectives, and most adverbs. We identified the top 300 most frequent words in the dataset and categorized them into nouns, verbs, and adjectives or adverbs using a part-of-speech tagging

XSL•FO
RenderX

model, fine-tuned for accurate tagging tasks. In addition, we manually corrected any miscategorized words during the automated tagging process. The frequencies of these words were calculated and used to create a treemap visualization, highlighting the distribution and prominence of these content words across the comments.

Topic Modeling With BERTopic

Topic modeling is widely used in health informatics and related disciplines for text mining large datasets, such as comments or tweets, to extract main topics from extensive texts. For our study, we used BERTopic (Bidirectional Encoder Representations from Transformers), a state-of-the-art topic modeling technique that leverages transformer-based deep learning models to identify topics in large text collections. Its powerful capabilities have been demonstrated in various studies [24-26]. It consistently outperforms traditional topic modeling methods, such as latent Dirichlet allocation (LDA) and nonnegative matrix factorization (NMF), with a particularly pronounced advantage in analyzing short and unstructured texts, where embedding-based models show greater potential for generating meaningful insights [27]. To validate this choice, we conducted empirical comparisons between BERTopic, LDA, and NMF using topic diversity scores, which measure the uniqueness of generated topics. It is calculated as the ratio of the number of unique words to the product of the number of top words per topic and the total number of topics [28]. BERTopic achieved a significantly higher score (0.8750) compared to LDA (0.2375) and NMF (0.25), indicating its superior ability to produce distinct and less redundant topics. These results align with previous research, which highlights BERTopic's effectiveness in capturing contextual meaning and semantic nuances often missed by traditional models [27,29]. All analyses and visualizations were conducted using Python in Colab, using the BERTopic package (version 0.16.0).

We began by preprocessing the text data and removing special characters, symbols, and punctuation. Then, we used the vectorization technique CountVectorizer to convert the preprocessed text into numerical representations. These numerical representations capture the frequency or importance of words in each document and can be directly input into topic modeling algorithms.

Subsequently, we conducted topic modeling with BERTopic. Based on a preliminary review of the comments, we estimated the number of topics to range from 5 to 10. After training models with various predefined topic numbers, we selected 8 as the final number because it provided a balance between maintaining model coherence and capturing meaningful thematic distinctions within the comments and retained only topics that included more than 50 comments. As part of the topic modeling process, we analyzed both unigrams and bigrams to generate topics. While unigrams offer a broad sense of the topics by identifying the most frequent individual terms, bigrams provide a more fine-grained perspective by highlighting word pairs that are often used together [30]. Analyzing both unigrams and bigrams helped create a more comprehensive understanding of the topics discussed.

Topic modeling is a common approach to analyzing social media content, such as Facebook (Meta) posts, tweets, and Reddit communities, related to public health. It identifies key themes, including awareness campaigns, personal caregiving experiences, and scientific advancements [31]. This approach provides a framework for understanding how structured health campaigns influence public discourse on social media [32]. Therefore, topic modeling is suitable for this study to examine comments on VR videos on YouTube, contributing to a more comprehensive view of ADRD in social media contexts.

Ethical Considerations

This study consisted of observations of public behaviors using secondary research data. The study data are anonymous and deidentified. Therefore, ethical approval was not required.

Results

Overview of the Video Comments

We extracted all comments from the 3 VR videos in the AWTD series. Both WH and AH were posted on June 2, 2016, while AS was posted the following year on March 21, 2017. As of September 2024, when the data was collected, AS had the highest number of comments, with 6572 (including replies), WH had 5954 comments, and AH had 122 comments.

The treemap visualization illustrates the most frequently used words in the comments across the 3 videos. Each rectangle represents a word, with its size and color intensity corresponding to the word's frequency in the comments for each video. The word "dementia" is the most prominent in the noun category, indicating that it is a central topic in the discussions. Other significant nouns include "family," "home," "memory," "people," "video," "mom," and "disease." These nouns suggest that the comments often focus on personal experiences, caregiving, videos as a tool, and the impact of dementia on individuals and families.

Among the verbs, words including "know," "remember," "think," and "lost" are frequent. This shows that the comments are action-oriented, reflecting emotions, actions taken, or care experiences of the commenters. The presence of verbs like "forgot" and "care" suggests a focus on ADRD symptom and caregiving actions.

In the adjectives or adverbs category, "even," "really," "much," "old," and "still" are prominent. These words typically modify other words to express degree, quality, or timing, indicating that commenters are emphasizing the severity, persistence, or specific qualities of the experiences or events they are discussing. In addition, words like "sorry," "bad," "good," "best," "sad," and "hard" reflect emotional responses and value judgments, indicating that many comments convey personal sentiments or evaluations of the situations being described (Figure 2).

XSL•FO
RenderX

Figure 2. Top 300 words frequency treemap categorized by part-of-speech. The visualization groups the most frequently used words from YouTube comments into verbs (green), nouns (blue), and adjectives or adverbs (orange). The size of each word reflects its frequency.

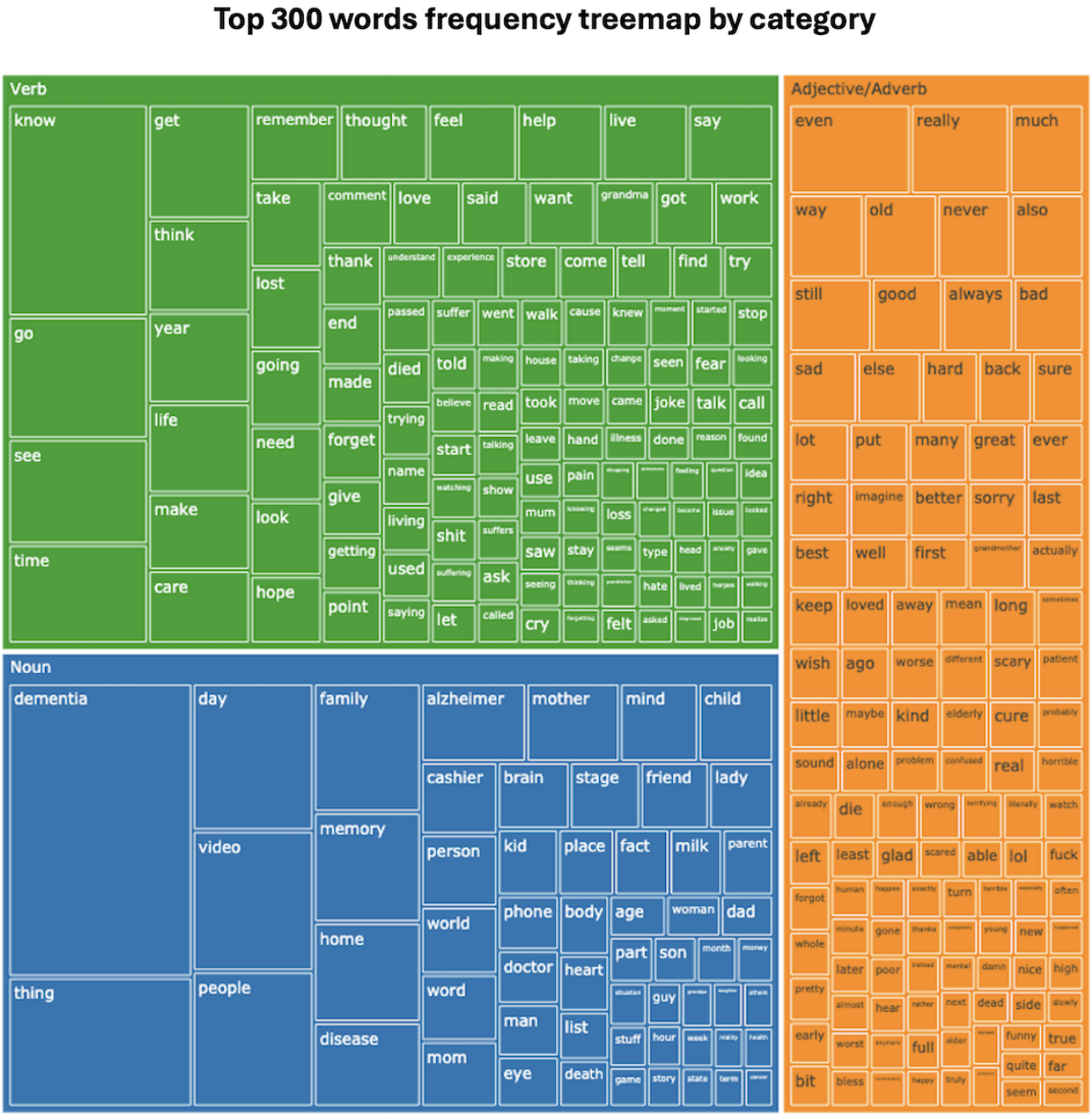

Topics of the Video Comments

The topic modeling process identified 8 initial topics. The 2D dense clusters of these topics, visualized in Figure 3, depict the text embeddings. Each cluster, represented by a different color, corresponds to a unique topic derived from the dataset. Each point signifies a document, with spatial proximity indicating similarity between comments. The topic size decreases as the topic number increases, with Topic 0 being the largest, comprising 2040 comments. This topic is labeled "dementia_like_just" by BERTopic. The most representative words characterizing each topic are shown in Figure 4. For instance, the top 5 words in Topic 7 are "walls," "god," "live," "jesus," and "christ," indicating that this topic revolves around religious discussions.

XSL•FO
RenderX

Figure 3. Topic distribution and clustering of "A Walk-Through Dementia" comments. Comments are distributed into 8 distinct topics, each represented by a different color. The topics are ranked from 0 to 7 based on their size, with Topic 0 ("dementia_just_like") being the largest and Topic 7 ("walls_god_live") the smallest.

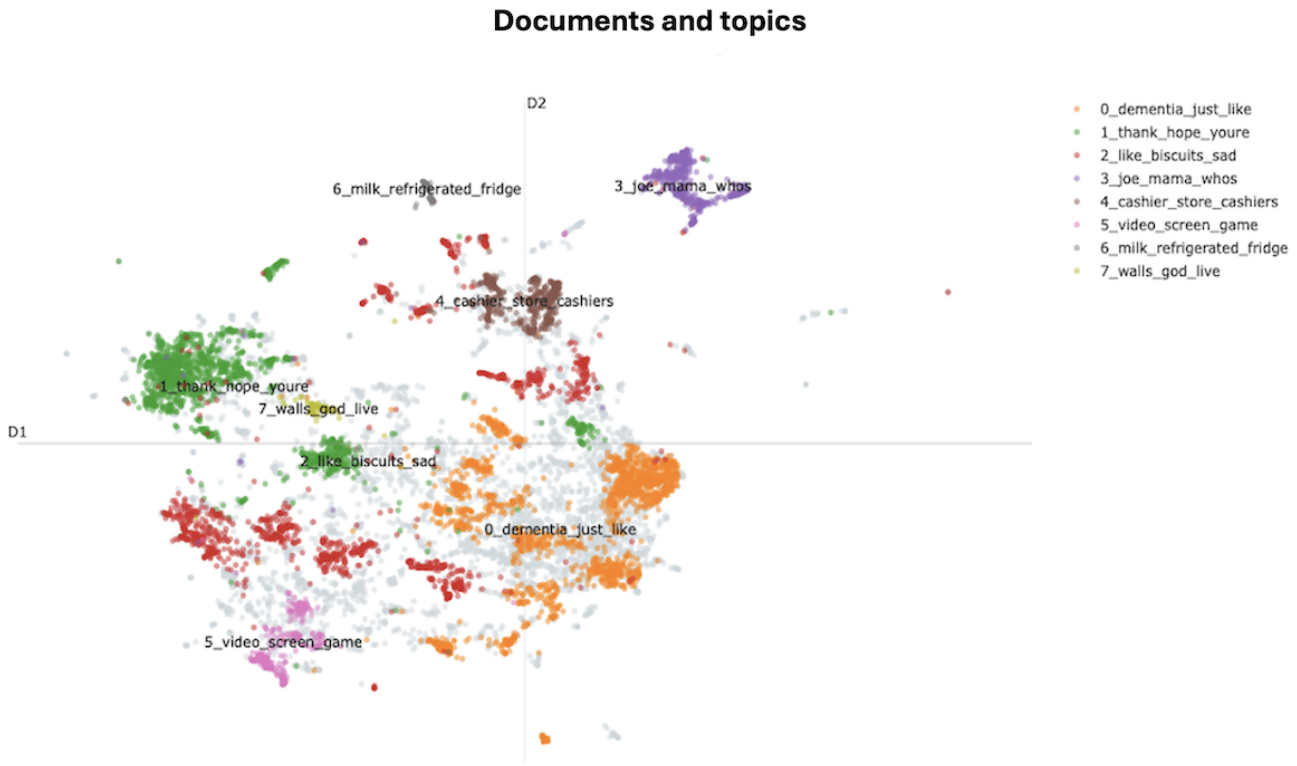

Figure 4. Top words by topic from "A Walk-Through Dementia" (AWTD) comments. It presents the most relevant words for each of the 8 identified topics in AWTD YouTube comments. The length of the bars after each word indicates the word's importance within that topic. For example, the word "dementia" is the most significant word in Topic 0.

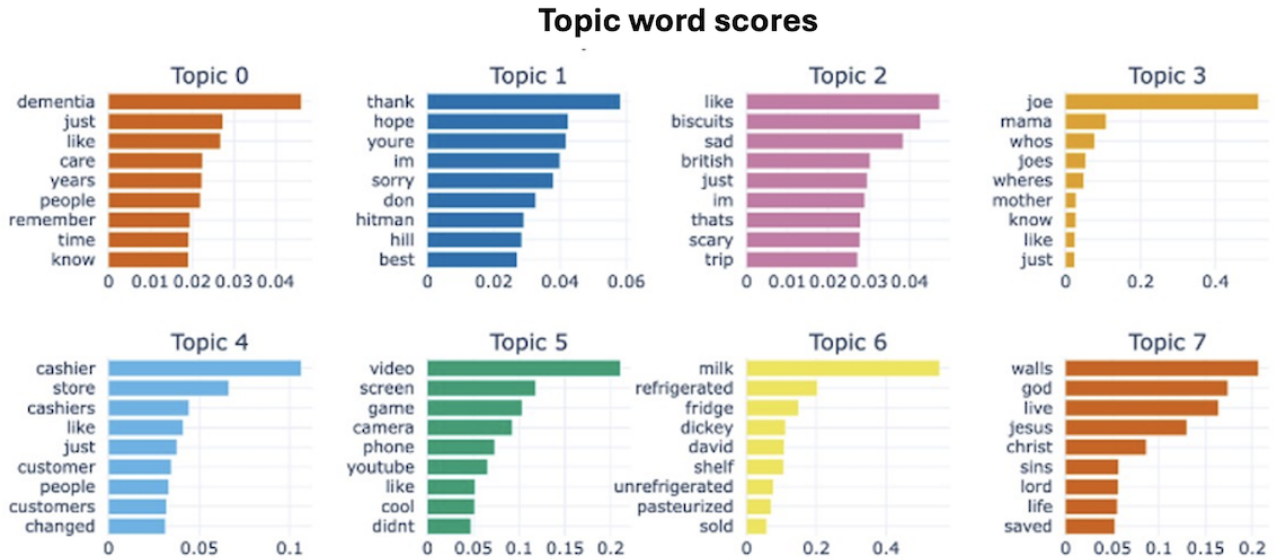

After manually reviewing the 8 topics and content of comments, we synthesized them into five cohesive themes reflecting viewers' thoughts and needs: (1) the personal experience of ADRD care and expressions of empathy and support, (2) positive reactions to the VR videos, (3) challenges faced by individuals with ADRD, (4) caregivers' role, and (5) learning from VR videos. Notably, the majority of comments reflected a deep emotional connection to the content, underscoring the significant impact of personal experiences related to ADRD. Many viewers expressed a newfound understanding and empathy toward those living with dementia and the challenges they encounter in daily life, highlighting the educational value of the immersive experience. The reviewers also discussed the roles and responsibilities of a caregiver of a person with ADRD and what the members of the general public learned from AWTD. Table 1 illustrates our topic analysis and theme development process, listing the initial topics, keywords, representative comments, and summarized themes.

XSL•FO
RenderX

Table . Overview of topic analysis and theme development. The table presents 8 topics identified by BERTopic, along with associated keywords and representative comments.

Topic ID	Keywords	Representative comment	Summarized theme
0	dementia, just, like, care, years, people, home, time, family, know	• "My mom died of Alzheimer's. It is hard when someone you love forgets where they live, who you are, what's going on." • "That would break my heart. I guess some people get used to it and live with this sickness of one of their family members, but for an outsider, like me, reading your comment was just a tough read."	Personal experiences of ADRD[a] care, empathy and support
1	thank, your, I'm, hope, sorry, don, like, don't	• "I hope you are doing well, your Strength and ability to push on every day is so admirable and inspiring. Thank you for being such an amazing human being…" • "I am a member of a search team, and we are called upon to search urban areas, forests, swamps and such all the time for elderly people with dementia who 'wandered off'"	Personal experiences of ADRD care, empathy and support
2	biscuits, like, sad, British, trip, tea, I'm, scary, sugar, that's	• "Like a bad trip you have no control over" • "I love the subtle details, like how there was an orange car parked by her house, but not the same type she saw earlier." • "When a person mistakes a puddle for a hole in the ground, don't just accept that as normal behavior, look into it"	Challenges faced by individuals with ADRD
3	joe, mama, who's, son, joes, where's, just, like, voice, know	• "who's Joe?" "Old lady's son" "where's Joe?" • "For me, the most terrifying part of the whole video was when Joe closed the door, leaving his mom all alone. She's now isolated in a big home with tall stairs and couldn`t easily walk out." • "Children, you should make a cup of tea for your mum" • "Dementia is often overlooked and extremely underestimated."	Caregivers' role

XSL•FO
RenderX

Topic ID	Keywords	Representative comment	Summarized theme
4	cashier, store, cashiers, like, just, customer, people, customers, changed, rude	• "The cashier clearly has no understanding. She is rushing the poor women to pay. So rude. It's super devastating to see this because we know they can't stand up for themselves because of how much they struggle." • "When we put ourselves in someone's shoes, we learn so much about others and conditions we never really paid much thought to" • "I DEFINITELY need to be a more patient person with my aging mother! Thanks for posting this." • "When I was working in the nursing home, I always loved to just hold their hand, now I understand why they never refuse."	Learning from A Walk-Through Dementia
5	video, screen, camera, game, phone, YouTube, cool, didn't, like, realize	• "This is a great 360 video that puts us right in the middle of someone's illness" • "What a powerful video. Did anyone realize that if you move your phone, it pans the camera? Very cool!" • "This is so helpful, thank you! People without dementia are challenged to empathize because they don't have the insight this video provides." • "This is so much more informative than any government add or script" • "I could feel the anxiety realizing she was home alone. What a powerful message." • "I sit here reflecting on this incredible powerful and educational tool. I (My Eyes, My Heart, My Soul, My Feelings, My Spirit, My Brain) have just experience." • "I find the 360 aspect is really cool, the video is of great help to the general public in understanding the life difficulties of dementia cases."	Positive reactions to the VR[b] videos
6	milk, refrigerated, fridge, dickey, David, shelf, unrefrigerated, pasteurized, sold, isle		Challenges faced by individuals with ADRD

XSL•FO
RenderX

Topic ID	Keywords	Representative comment	Summarized theme
		• "Milk in all, 2spoon, 1s, 1/2s.; the milk wasn't refrigerated; Why is the milk not in the fridge" • "I would need to write it down for sure" • "Guests can serve themselves. Or else make a list for the mother, and tape it on a cupboard door for the mother to make it easier" • "They should really get her a place with no stairs, easy access to beds and rooms, cushions and pads in case she slips. Music and things to keep her distracted."	
7	god, Jesus, Christ, sins, life, saved, lord, John, repent, loves	• I'm taking care of my demented grandmother… alone, 24/7. Jesus Christ it is difficult, I don't know how long I can keep on doing this" • "God bless them. I take care of dementia patients all the time at the hospital, but this sheds new light."	Personal experiences of ADRD care, empathy, and support

[a]ADRD: Alzheimer disease and related dementias.

[b]VR: virtual reality.

Discussion

Principal Findings

The study evaluated AWTD, a series of free YouTube VR videos designed to raise public awareness of ADRD. At the time of data collection, AWTD was the most viewed video for the search terms "dementia" and "Alzheimer's disease" on YouTube, boasting over 6 million views. This series also garnered the highest number of comments, totaling 12,648, compared to a similar VR video, "Coping with Alzheimer's: Together and Apart," which had 67,000 views and only 66 comments. The comments from viewers predominantly came from individuals with a personal connection to someone with ADRD, professionals working with affected individuals, or general audience members. This suggests that the YouTube VR videos were particularly effective in disseminating information about ADRD to a diverse and broader audience.

Our analysis identified five key themes from the comments: (1) positive reactions to the VR videos, (2) personal experience of ADRD along with expressions of empathy and support, (3) the role of caregivers, (4) challenges faced by individuals with ADRD, and (5) learning from AWTD. These findings suggest that VR has significant potential to raise public awareness of ADRD and deepen understanding of the condition, particularly through immersive and emotionally engaging experiences.

Positive Reactions to the VR Experience and Immersive Learning

The immersive nature of VR proved highly effective in engaging viewers, allowing them to experience the challenges of ADRD firsthand, significantly enhancing their understanding of the condition. Many commenters described it as a convincing and immersive experience, both insightful and evocative, providing a powerful educational tool for ADRD care. Its appeal stemmed in part from its novelty in the field of dementia education, particularly for YouTube users, including caregivers, family members, and professionals with extensive experience in dementia care. Commenters noted that they had never encountered anything like this in their personal or professional lives. This is unsurprising, as academic literature recognizes VR as an emerging area of interest in dementia education [23,33], highlighting the need for further exploration of its applications and potential impact.

Our findings suggest that the success of the VR video lies in its ability to immerse viewers in the perspective of someone living with dementia. By wearing the VR headset, users are fully absorbed in the experience, allowing them to focus entirely on each scenario and draw connections to their personal lives or work experiences. Commenters noted how the VR video placed them directly in the midst of the illness, enabling them to feel the anxiety of a person with ADRD who is home alone. One commenter remarked, "My eyes, my heart, my soul, my feelings, my spirit, my brain have just experienced this," emphasizing how immersion creates a profound sense of presence, making

XSL•FO
RenderX

viewers feel as if they were in the shoes of someone with ADRD, inspiring empathy and support.

These findings are consistent with research from other scholars who emphasize the benefits of VR dementia education for medical students and professionals [34-36], as well as broader VR literature [37], which highlights the critical role of immersion and presence in virtual scenarios. In addition, AWTD stands out as a free VR video available on YouTube, unlike most VR content, which is typically developed and hosted on developers' websites, requiring purchase or offering limited access [38,39]. By being widely accessible, AWTD effectively disseminates information and knowledge about ADRD to the general public, offering significant benefits for public education on the condition. This aligns with Lupton's findings, which state that social media platforms like YouTube provide opportunities for researchers to step beyond traditional academic approaches and use innovative, free methods to share knowledge with new and diverse audiences [40].

Role of Caregivers and Observed Gaps

The VR experience prompted viewers to reflect on the crucial role of caregivers, with many commenters noting gaps in the videos, which demonstrates how VR can stimulate critical thinking and awareness of caregiving responsibilities in ADRD. Some commenters discussed the noticeable absence of caregivers in the 3 VR videos. They expressed concern that caregivers, particularly family members, should have been more present and attentive in the depicted scenarios. Comments like "Who is Joe?" (son of an individual with ADRD), "Where is Joe?", and "The daughter should make tea for her mother (individual with ADRD)," reflected the viewers' belief that caregivers were missing in moments when they were needed most. In these scenarios, family members, who play a crucial role in caregiving, appeared to overlook the early signs of ADRD, such as confusion, memory lapses, and difficulty completing daily tasks. By neglecting these key symptoms, caregivers failed to provide necessary support, which may have contributed to negative outcomes for individuals with ADRD, including diminished psychosocial well-being, lowered self-esteem, and a reduced overall quality of life [41]. The VR videos can help family caregivers become more aware of their roles and responsibilities in caring for individuals with ADRD, potentially encouraging further learning and training on dementia care. This heightened awareness can lead to more proactive approaches to caregiving, allowing family members to better recognize and respond to the needs of their loved ones, ultimately improving the quality of care provided to individuals living with ADRD.

Personal Connections and Empathy

By simulating the experiences of individuals with ADRD, VR video fostered a deeper emotional connection, encouraging viewers to express empathy and reflect on how the condition affects both patients and their families. Scenarios depicted individuals struggling to live with ADRD, allowing viewers to face challenges that they may have been previously unaware of, which were vividly portrayed through the VR experience. Commenters specifically noted moments like navigating around a puddle, which the person with ADRD believes to be a hole in the ground (street scenario), highlighting their disorientation. Other scenes—such as struggling to retrieve items from shelves and feeling rushed by a cashier (supermarket scenario) or becoming confused when asked to make tea for her daughter's guests (AH scenario)—underscored the daily difficulties faced by individuals with ADRD. Experiencing these challenges from the perspective of someone with ADRD seemed to provide viewers with a deeper insight and better understanding of the condition's impact.

These experiences prompted viewers to recognize how negative attitudes, impatience, or misunderstanding from others exacerbate the challenges faced by individuals with ADRD. Building a dementia-friendly or ADRD-aware community is essential to support people with ASRD to live well and carry out their daily activities. One potential approach to achieving this is to incorporate the VR video into a broader public training program, where the scenarios are reflected upon and discussed by the learners. As some commenters stated:

> *The cashier clearly has no understanding of ADRD. It's super devastating to see this because we know they can't stand up for themselves because of how much they struggle.*
>
> *When we put ourselves in someone's shoes, we learn so much about others and conditions we never really paid much thought to.*
>
> *I definitely need to be a more patient person with my aging mother! Thanks for posting this.*
>
> *When I was working in the nursing home, I always loved to just hold their hand, now I understand why they never refuse.*

VR video is a powerful way to raise awareness of ADRD, helping the public shift their perception of the condition and learn essential skills for caring for individuals with ADRD. Similar findings regarding VR as a learning tool have been noted for medical students and professional caregivers [42,43]. The VR video therefore has a fundamental role to play in helping to create equal, inclusive communities for all, improving the actions people undertake and the language they use when interacting with those with ADRD [21].

The research findings underscore the importance of immersive learning in fostering empathy and understanding for individuals with ADRD. To build on this success, there is a clear need for more VR tools that can be used in conjunction with follow-up discussions featuring educational takeaway messages tailored for diverse audiences with varying levels of knowledge and experience regarding ADRD [44]. Creating VR programs that cater to different linguistic and cultural contexts could also address gaps in accessibility. The existing scenarios might oversimplify the condition and fail to capture the wide variability in ADRD experiences. A more diverse set of scenarios covering different stages and severities of the condition would likely provide a more nuanced understanding. The insights gained from this research can guide these discussions and aid in the development of the format and learning materials for a future VR educational program based on AWTD.

XSL•FO
RenderX

Social Media Influence and VR in Health Communication

To contextualize the impact of VR as an educational tool, we ground our study within social media influence theory [45], which posits that social media platforms and their users shape public discourse, influence attitudes, and drive behavioral change through content creation, sharing, and interaction. Social media platforms serve as powerful conduits for disseminating health-related content, leveraging both algorithmic curation and social interactions to amplify messages. VR, as an immersive medium, engages user through both central (cognitive processing of health messages) and peripheral (emotional and experiential engagement) routes. This dual-pathway influence enhances message retention and can drive behavioral change, making VR an effective medium for public health communication.

The interactive nature of social media ensures that VR content does not exist in isolation but is actively discussed, shared, and reframed within online discourse. Algorithms prioritize highly engaging and emotionally resonant content, increasing its visibility and potential impact. User-generated discussions, as seen in the YouTube comments analyzed in this study, serve as mechanisms through which social media influence is both reinforced and coconstructed. When users share personal stories, express emotions, or react to the VR experience, they contribute to a collective conversation that reinforces common beliefs and introduces new perspectives on ADRD. This aligns with Social Media Influence theory's assertion that online interactions shape public discourse and personal identities, ultimately contributing to the normalization or transformation of health-related attitudes and behaviors.

These theoretical perspectives allow us to interpret the thematic findings within a broader framework of digital intervention effectiveness and audience engagement. By integrating topic modeling and BERT-based analysis with media influence theory, our study not only examines the immediate impact of VR-based health communication but also contributes to a deeper understanding of how digital interventions shape public awareness and health behaviors. The methodological framework established here can be adapted and extended to different populations and varying digital health content, enabling broader applications of VR in public health communication and education.

Limitations

One limitation of this study is the absence of demographic information about the commenters, particularly regarding their ethnicity, socioeconomic status, and professional roles. This information could provide insights into how social factors affect the experience of VR videos. In addition, as all VR video content was delivered in English, non-English-speaking populations may have been excluded, limiting the representativeness of the research sample in relation to the general public. Furthermore, the comments analyzed in this study constitute only a small portion of the total posts made in the forum, potentially failing to capture the full range of discussions surrounding ADRD. Further research could focus on gathering a more comprehensive dataset across platforms and including different languages to capture a broader perspective. Finally, this study focused exclusively on one social media platform, YouTube, and one VR video. Future research could broaden its scope by incorporating data from multiple social media platforms and a wider array of VR videos to enhance our understanding of the topic.

Conclusion

This study used topic modeling to highlight the potential of VR videos as a powerful and immersive educational tool, offering the public a more profound understanding of the experiences of individuals living with ADRD. By fostering emotional and cognitive engagement, this approach has the potential to cultivate greater empathy and community support for those affected by ADRD. Beyond its immediate findings, this research underscores the transformative role of VR in reshaping how complex health conditions are communicated and understood. Future studies should expand on this foundation by exploring the broader applications of VR on social media platforms, such as Dementia VR Film, Beatriz Lab VR Module, and virtual reality scenarios. These tools not only have the capacity to reach wider audiences but also to drive meaningful societal change by enhancing awareness, reducing stigma, and promoting inclusive support systems for individuals with ADRD and their caregivers.

Acknowledgments

We would like to express our gratitude to the creators and uploaders of the publicly available content on YouTube, which served as the foundation for this secondary data analysis. Their contributions enabled us to conduct this research and explore the potential of virtual reality videos as an educational tool for Alzheimer disease and related dementias.

Data Availability

All data generated or analyzed during this study are included in this published article.

Conflicts of Interest

None declared.

References

XSL•FO
RenderX

1. Dementia. World Health Organization. 2023 Mar 15. URL: https://www.who.int/news-room/fact-sheets/detail/dementia [accessed 2024-05-15]
2. Shu S, Woo BK. Use of technology and social media in dementia care: current and future directions. World J Psychiatry 2021 Apr 19;11(4):109-123. [doi: 10.5498/wjp.v11.i4.109] [Medline: 33889536]
3. Prince M, Wimo A, Guerchet M, Ali GC, Wu YT, Prina M. The global impact of dementia: an analysis of prevalence, incidence, cost and trends. : Alzheimer's Disease International; 2015 URL: https://www.alzint.org/u/WorldAlzheimerReport2015.pdf [accessed 2024-05-15]
4. Dementia friendly communities: global developments. Alzheimer's Disease International. 2017. URL: https://www.alzint.org/u/dfc-developments.pdf [accessed 2024-05-15]
5. Green G, Lakey L. Building dementia-friendly communities: a priority for everyone. Act on Alzheimer's. 2013. URL: https://actonalz.org/sites/default/files/2023-01/Dementia_friendly_communities_full_report.pdf [accessed 2024-05-15]
6. Milne A. The "D" word: reflections on the relationship between stigma, discrimination and dementia. J Ment Health 2010 Jun;19(3):227-233. [doi: 10.3109/09638231003728166] [Medline: 20441486]
7. Mukadam N, Livingston G. Reducing the stigma associated with dementia: approaches and goals. Aging health 2012 Aug;8(4):377-386. [doi: 10.2217/ahe.12.42]
8. Cheng ST, Lam LCW, Chan LCK, et al. The effects of exposure to scenarios about dementia on stigma and attitudes toward dementia care in a Chinese community. Int Psychogeriatr 2011 Nov;23(9):1433-1441. [doi: 10.1017/S1041610211000834] [Medline: 21729424]
9. Cheston R, Hancock J, White P. A cross-sectional investigation of public attitudes toward dementia in Bristol and South Gloucestershire using the approaches to dementia questionnaire. Int Psychogeriatr 2016 Oct;28(10):1717-1724. [doi: 10.1017/S1041610216000843] [Medline: 27353013]
10. Chen J, Wang Y. Social media use for health purposes: systematic review. J Med Internet Res 2021 May 12;23(5):e17917. [doi: 10.2196/17917] [Medline: 33978589]
11. Auxier B, Anderson M. Social media use in 2021. Pew Research Center. 2021. URL: https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/ [accessed 2024-05-15]
12. Moorhead SA, Hazlett DE, Harrison L, Carroll JK, Irwin A, Hoving C. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. J Med Internet Res 2013 Apr 23;15(4):e85. [doi: 10.2196/jmir.1933] [Medline: 23615206]
13. Wu S, Foster K, Ramirez M, Jin H, Joe E. How well do videos on YouTube meet information needs of ADRD family caregivers? A qualitative analysis. Innovation in Aging 2020 Dec 16;4(Supplement_1):273-274. [doi: 10.1093/geroni/igaa057.875]
14. Teano AL, Scott A, Gipson C, Albert M, Pettigrew C. Social media programs for outreach and recruitment supporting aging and alzheimer disease and related dementias research: longitudinal descriptive study. JMIR Aging 2024 Jul 9;7:e51520. [doi: 10.2196/51520] [Medline: 38981112]
15. Dassel K, Butler J, Telonidis J, Edelman L. Development and evaluation of alzheimer's disease and related dementias (ADRD) best care practices in long-term care online training program. Educ Gerontol 2020 Mar 3;46(3):150-157. [doi: 10.1080/03601277.2020.1717079]
16. Lam NHT, Woo BKP. YouTube as a new medium for dementia education among Chinese Americans. Community Ment Health J 2020 Apr;56(3):435-439. [doi: 10.1007/s10597-019-00493-7] [Medline: 31641910]
17. Heward M, Palfreman-Kay J, Innes A. In their words: how television and visual media can raise awareness of dementia and other health conditions that carry stigma, including disabilities. Journal of Popular Television, The 2015 Oct 1;3(2):229-242. [doi: 10.1386/jptv.3.2.229_1]
18. Castillo LIR, Hadjistavropoulos T, Brachaniec M. The effectiveness of social media in the dissemination of knowledge about pain in dementia. Pain Med 2021 Nov 26;22(11):2584-2596. [doi: 10.1093/pm/pnab157] [Medline: 34009395]
19. Jiang H, Vimalesvaran S, Wang JK, Lim KB, Mogali SR, Car LT. Virtual reality in medical students' education: scoping review. JMIR Med Educ 2022;8(1):e34860. [doi: 10.2196/34860]
20. Kamińska D, Sapiński T, Wiak S, et al. Virtual reality and its applications in education: survey. Information 2019;10(10):318. [doi: 10.3390/info10100318]
21. Hicks B, Konovalova I, Myers K, Falconer L, Board M. Taking 'A walk through dementia': exploring care home practitioners' experiences of using a virtual reality tool to support dementia awareness. Ageing Soc 2023 May;43(5):1042-1067. [doi: 10.1017/S0144686X21000994]
22. A walk through dementia. Alzheimer's Research UK. 2016. URL: https://www.alzheimersresearchuk.org/campaigns/awtd/ [accessed 2024-05-15]
23. Li X, Liu X, Collins S, Mutthina R, Chiu C. The virtual reality intervention programs and dementia care education: a systematic review. Educ Gerontol 2025:1-17. [doi: 10.1080/03601277.2025.2466131]
24. Lee H, Hong Lee S, Re Lee K, Hyun Kim J. ESG discourse analysis through BERTopic: comparing news articles and academic papers. Computers, Materials & Continua 2023;75(3):6023-6037. [doi: 10.32604/cmc.2023.039104]
25. Wang Z, Chen J, Chen J, Chen H. Identifying interdisciplinary topics and their evolution based on BERTopic. Scientometrics 2024 Nov;129(11):7359-7384. [doi: 10.1007/s11192-023-04776-5]

XSL•FO
RenderX

26. Hristova G, Netov N. Media coverage and public perception of distance learning during the COVID-19 pandemic: a topic modeling approach based on bertopic. Presented at: 2022 IEEE International Conference on Big Data (Big Data); Dec 17-20, 2022; Osaka, Japan. [doi: 10.1109/BigData55660.2022.10020466]
27. Egger R, Yu J. A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify twitter posts. Front Sociol 2022;7:886498. [doi: 10.3389/fsoc.2022.886498] [Medline: 35602001]
28. Asnawi MH, Pravitasari AA, Herawan T, Hendrawati T. The combination of contextualized topic model and MPNet for user feedback topic modeling. IEEE Access 2023;11:130272-130286. [doi: 10.1109/ACCESS.2023.3332644]
29. Chong M, Chen H. Racist framing through stigmatized naming: a topical and geo‐locational analysis of #Chinavirus and #Chinesevirus on Twitter. Proceedings of the Association for Information Science and Technology 2021 Oct;58(1):70-79. [doi: 10.1002/pra2.437]
30. Nokel M, Loukachevitch N. A method of accounting bigrams in topic models. In: Proceedings of the 11th Workshop on Multiword Expressions: Association for Computational Linguistics; 2015:1-9. [doi: 10.3115/v1/W15-0901]
31. Paul MJ, Dredze M. Discovering health topics in social media using topic models. PLoS ONE 2014;9(8):e103408. [doi: 10.1371/journal.pone.0103408] [Medline: 25084530]
32. Kong D, Chen A, Zhang J, et al. Public discourse and sentiment toward dementia on Chinese social media: machine learning analysis of weibo posts. J Med Internet Res 2022 Sep 2;24(9):e39805. [doi: 10.2196/39805] [Medline: 36053565]
33. Hirt J, Beer T. Use and impact of virtual reality simulation in dementia care education: a scoping review. Nurse Educ Today 2020 Jan;84:104207. [doi: 10.1016/j.nedt.2019.104207]
34. Adefila A, Graham S, Clouder L, Bluteau P, Ball S. myShoes – the future of experiential dementia training? The Journal of Mental Health Training, Education and Practice 2016 May 9;11(2):91-101. [doi: 10.1108/JMHTEP-10-2015-0048]
35. Morganti F, Palena N, Savoldelli P, Greco A. Take the first-person perspective to become dementia-friendly: the use of 360° video for experiencing everyday-life challenges with cognitive decline. Front Psychol 2020;11:1117. [doi: 10.3389/fpsyg.2020.01117]
36. Wijma EM, Veerbeek MA, Prins M, Pot AM, Willemse BM. A virtual reality intervention to improve the understanding and empathy for people with dementia in informal caregivers: results of a pilot study. Aging Ment Health 2018 Sep;22(9):1115-1123. [doi: 10.1080/13607863.2017.1348470] [Medline: 28691861]
37. Falconer L. Situated learning in virtual simulations: researching the authentic dimension in virtual worlds. J Interact Learn Res 2013;24(3):285-300.
38. Bard JT, Chung HK, Shaia JK, Wellman LL, Elzie CA. Increased medical student understanding of dementia through virtual embodiment. Gerontol Geriatr Educ 2023;44(2):211-222. [doi: 10.1080/02701960.2022.2067850] [Medline: 35451921]
39. Sung HC, Su HF, Lee WL, Yamakawa M, Wang HM. Effects of a dementia virtual reality-based training with peer support for home care workers: a cluster randomized controlled trial. Int J Geriatr Psychiatry 2022 Sep;37(9). [doi: 10.1002/gps.5799] [Medline: 35996760]
40. Lupton D. 'Better understanding about what's going on': young Australians' use of digital technologies for health and fitness. Sport Educ Soc 2020 Jan 2;25(1):1-13. [doi: 10.1080/13573322.2018.1555661]
41. Li X, Dorstyn D, Mpofu E, et al. Nursing assistants and resident satisfaction in long-term care: a systematic review. Geriatr Nurs (Lond) 2021 Nov;42(6):1323-1331. [doi: 10.1016/j.gerinurse.2021.08.006]
42. Yamaguchi Y, Ryuno H, Fukuda A, et al. Effects of a virtual reality intervention on dementia care education among acute care nurses in Japan: a non-randomised controlled trial. Geriatr Nurs 2022;48:269-273. [doi: 10.1016/j.gerinurse.2022.10.013] [Medline: 36332442]
43. Stargatt J, Bhar S, Petrovich T, Bhowmik J, Sykes D, Burns K. The effects of virtual reality-based education on empathy and understanding of the physical environment for dementia care workers in Australia: a controlled study. J Alzheimers Dis 2021;84(3):1247-1257. [doi: 10.3233/JAD-210723] [Medline: 34633323]
44. Board M, Phipps L, Mitchell R, Murphy J. Evaluating the impact of the virtual reality app 'A walk through dementia' on year one health care students clinical practice. Presented at: The Alzheimer's Show; Jun 7-8, 2019; Olympia, London URL: https://eprints.bournemouth.ac.uk/33667/ [accessed 2024-05-15]
45. Rachmad Y. Social Media Influence Theory: Phoenix Cactus Book Publishing; 2023.

Abbreviations

ADRD: Alzheimer disease and related dementias
AH: at home
AS: at the supermarket
AWTD: A Walk-Through Dementia
BERT: Bidirectional Encoder Representations from Transformers
LDA: latent Dirichlet allocation
NMF: nonnegative matrix factorization
VR: virtual reality

XSL•FO
RenderX

WH: walking home

Edited by A Mavragani; submitted 20.10.24; peer-reviewed by Y Sutherland, Z Wang; revised version received 17.02.25; accepted 18.02.25; published 16.04.25.

Please cite as:
Li X, Liu X, Yin C, Collins S, Alanazi E
Impact of a Virtual Reality Video ("A Walk-Through Dementia") on YouTube Users: Topic Modeling Analysis
JMIR Form Res 2025;9:e67755
URL: https://formative.jmir.org/2025/1/e67755
doi:10.2196/67755

© Xiaoli Li, Xiaoyu Liu, Cheng Yin, Sandra Collins, Eman Alanazi. Originally published in JMIR Formative Research (https://formative.jmir.org), 16.4.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Feasibility, Acceptability, and Effectiveness of a Smartphone App to Increase Pretransplant Vaccine Rates: Usability Study

Amy G Feldman[1,2,3], MD, PhD; Brenda L Beaty[1,3], MSPH; Susan L Moore[1,3,4], MSPH, PhD; Sheana Bull[4], MPH, PhD; Kumanan Wilson[5,6,7], MSc, MD, FRCP; Katherine M Atkinson[8], PhD; Cameron Bell[7], BEng; Kathryn M Denize[7], MSc; Allison Kempe[1,3,9], MPH, MD

[1]Section of Pediatric Gastroenterology, Hepatology, and Nutrition, University of Colorado, 13123 E 16th Ave, Aurora, CO, United States
[2]Digestive Health Institute, Children's Hospital Colorado, Aurora, CO, United States
[3]Adult and Child Center for Outcomes Research and Delivery Science, Aurora, CO, United States
[4]Colorado School of Public Health, Aurora, CO, United States
[5]Department of Medicine, University of Ottawa, Ottawa, ON, Canada
[6]Bruyère Health Research Institute, Ottawa, ON, Canada
[7]CANImmunize Inc, Ottawa, ON, Canada
[8]Department of Global Public Health, Karolinska Institutet, Stockholm, Sweden
[9]Children's Hospital Colorado, Aurora, CO, United States

Corresponding Author:
Amy G Feldman, MD, PhD
Section of Pediatric Gastroenterology, Hepatology, and Nutrition, University of Colorado, 13123 E 16th Ave, Aurora, CO, United States

Abstract

Background: Vaccine-preventable infections result in significant morbidity, mortality, and costs in pediatric transplant recipients. Despite intensive medical care in the pretransplant period, less than 20% of children are up to date for age-appropriate vaccines at the time of transplant. Mobile health apps have the potential to improve pretransplant vaccine rates.

Objective: This paper aimed to perform phase 2 beta testing of the smartphone app, Immunize PediatricTransplant, to determine (1) if it was effective in achieving up-to-date vaccine status by the time of transplant in a cohort of children awaiting transplants and (2) if the app was feasible and acceptable to parent and transplant provider users.

Methods: We recruited 25 dyads of parents and providers of a child awaiting a liver, kidney, or heart transplant at Children's Hospital Colorado, Ann and Robert H. Lurie Children's Hospital, and the Children's Hospital of Philadelphia. Parents and providers filled out an entry questionnaire before app use to gather baseline information. A research team member entered the child's vaccine records into the app. The parent and provider downloaded and used the app until the transplant to view vaccine records, read vaccine education, communicate with team members, and receive overdue vaccine reminders. After the transplant (or on April 1, 2024, the conclusion of the study), the parent and provider filled out an exit questionnaire to explore feasibility and acceptability of the app. The child's vaccine records were reviewed to determine if the child was up to date on vaccines at the time of transplant.

Results: Twenty-five parent and provider dyads were enrolled; 56% (14/25) had a child awaiting a liver transplant, 28% (7/25) had a child awaiting a kidney transplant, and 16% (4/25) had a child awaiting a heart transplant. At the conclusion of the study, 96% (24/25) of the children were up to date on vaccines. Of the 36 parents and providers who filled out an exit questionnaire, 97% (n=35) agreed or strongly agreed that they felt knowledgeable about pretransplant vaccine use and 86% (n=31) agreed or strongly agreed that communication around vaccines was good after using the app. Further, 91% (20/22) of parents and 79% (11/14) of providers recommended the app to future parents and providers of transplant candidates. Parents and providers suggested that in the future the app should connect directly to the electronic medical record or state vaccine registries to obtain vaccine data.

Conclusions: The overwhelming majority of children whose parents and providers used the Immunize PediatricTransplant app were up to date on vaccines at the time of transplant. The majority of app users felt the app was feasible and acceptable. In future iterations of the app and subsequent clinical trials, we will explore whether application programming interfaces might be used to extract vaccine data from the electronic medical record. If implemented broadly, this app has the potential to improve pretransplant vaccine rates, resulting in fewer posttransplant infections and improved posttransplant outcomes.

(JMIR Form Res 2025;9:e68855) doi:10.2196/68855

XSL•FO
RenderX

KEYWORDS

mobile app; vaccines; immunizations; transplantation; children; beta test; mobile health; mHealth

Introduction

Transplant recipients are on life-long immunosuppressive medications to prevent graft rejection. As a result of these medications, they are at increased risk for life-threatening infections including vaccine-preventable infections (VPIs) [1-3]. In the first 5 years posttransplant, VPIs occur in pediatric solid organ transplant recipients at rates up to 87 times higher than in the general pediatric population [4,5]. VPIs result in significant morbidity, mortality, and increased hospitalization costs [4,5]. Pretransplant vaccines are an important strategy to decrease the risk of VPIs posttransplant. Unfortunately, despite intensive medical management in the pretransplant period, less than 20% of pediatric liver transplant recipients have received all age-appropriate vaccines by the time of transplant [6]. Barriers to pretransplant vaccination include gaps in knowledge about pretransplant vaccines, lack of communication between team members regarding vaccines, difficulty remembering when vaccines are due amidst other acute medical problems, and lack of an easily accessible centralized vaccine record [7].

Mobile health (mHealth) apps can assist with medical education, team communication, and patient adherence [8-12]. Specifically, mHealth apps have demonstrated utility in facilitating vaccine delivery in the general population [13-20]. In a previous study, our team developed a novel mHealth app, Immunize PediatricTransplant, to help improve vaccine rates in the pretransplant period [21]. Specifically, the content of the app (1) provides information about vaccines in the pretransplant period; (2) houses a cloud-based, easily accessible vaccine record; (3) includes a chat and communication feature to facilitate communication between parents and transplant providers about vaccines; and (4) provides text or email reminders for parents and providers when vaccines are overdue [21].

According to the mHealth Agile Development and Evaluation Lifecycle (Figure 1) [22], after a functional mobile tool is developed, phase 2 beta testing is needed to test the mHealth product amongst external users to understand if the prototype meets the needs of users in a way that will lead to sustained adoption. Therefore, we designed this study to (1) evaluate the effectiveness of the Immunize PediatricTransplant app in achieving up-to-date vaccine status by the day of transplant (or conclusion of the study on April 1, 2024) and (2) explore app feasibility and acceptability amongst parent and provider users.

Figure 1. mHealth Agile Development and Evaluation Lifecycle (adapted from Wilson et al [22]). mHealth: mobile health.

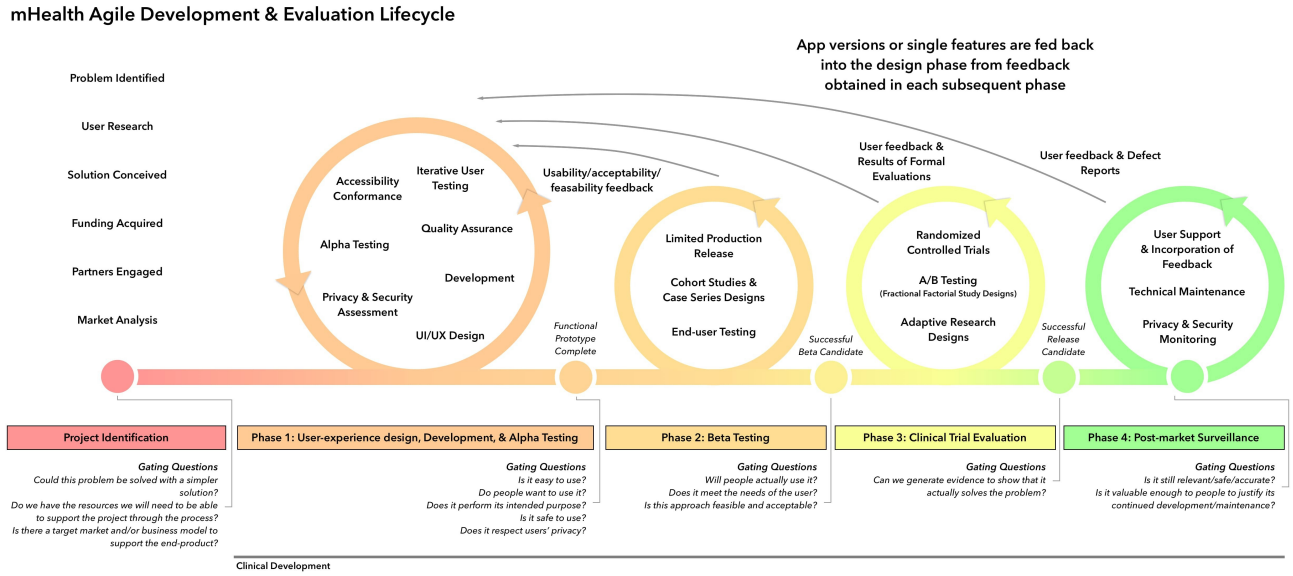

Methods

Study Design, Participants, and Recruitment

This prospective cohort study was conducted with enrollment between November 1, 2021, and July 31, 2023. Dyads of parents and transplant providers caring for a child awaiting a heart, liver, or kidney transplant were recruited from the Children's Hospital Colorado, Ann and Robert H. Lurie Children's Hospital, and the Children's Hospital of Philadelphia via an in-clinic, email, or electronic medical record (EMR) invitation. Parents of a child being transplanted for acute organ failure or retransplantation were excluded. Parents who refused vaccines for religious or ethical reasons, who were not fluent in English, or who did not have access to a mobile phone were excluded. Certain transplant providers had more than 1 patient enrolled in the study and used the app for all enrolled patients. All participants received a US $100 gift card for participation.

The primary outcome of this study was up-to-date vaccine status of the child on the day of transplant (or end of the study) according to standard Centers for Disease Control (CDC) vaccine recommendations [23]. Secondary outcomes included feasibility and acceptability of the app amongst parent and provider app users.

XSL•FO
RenderX

Study Procedures and Data Collection

Getting to Know You Entry Questionnaire

Each parent and transplant provider were given a baseline survey to gather information about them.

Immunize PediatricTransplant App

After the Getting to Know You questionnaire was completed, the parent and provider both downloaded the app. A research team member entered the child's vaccine records (using state registries, the EMR, and physician and parental records) to initiate the app to send text and email reminders to the parents and providers when a vaccine was due. Parents and providers were instructed to log into the app as often as they wished to view their child's vaccine records, read information about vaccines, or communicate with team members about vaccines. Vaccine reminders were sent to parents and transplant providers (via text, email, and app notification) when vaccines were due starting at enrollment and going until the time of transplant or end of the study.

Exit Questionnaire

After the transplant, the parent and provider filled out a survey with questions assessing app feasibility (How many months did you use the app for? Did you experience technical difficulties with the app?) and acceptability (Would you recommend the app to others? Did you feel communication between families and providers around vaccines was good after using the app? Did you feel knowledgeable about pretransplant vaccines after app use?). The exit questionnaire included an open-ended question asking for suggestions for future app improvement. The exit questionnaire was sent via email after the transplant was completed. Weekly reminders were sent until the questionnaire was complete.

Vaccine Rates at the Time of Transplant

A research team member collected vaccine records at the time of transplant to evaluate whether the child was up to date on age-appropriate vaccines at the time of transplant.

Statistical Analysis

Effectiveness

We calculated the percentage of transplant recipients who were completely up to date on age-appropriate vaccines at the time of transplant according to standard CDC vaccine recommendations [23].

Feasibility

Data from feasibility questions in the exit questionnaire were summarized through the proportion of the feasibility outcome.

Acceptability

Data from acceptability questions in the exit questionnaire were summarized through the proportion of the acceptability outcome.

Ethical Considerations

The study was approved by the Colorado Multiple Institutional Review Board (CO-18‑0045). All participants provided consent for study participation. Participation was voluntary. An email and password were required to log in to the study app. Participants could access all study content through a mobile app available on both the iOS and Android platforms. All study data were encrypted at rest and in transit. Only authorized researchers were provided access to the participant data. The application's data was stored exclusively on Amazon Web Services in the United States. Industry standard information security practices were leveraged in the development of the app. All data were deidentified to ensure participant privacy and confidentiality. Participants who completed the full study received US $100 compensation. No artificial intelligence was used in any portion of the manuscript writing.

Results

Participants

We enrolled 25 dyads of parents and guardians of heart, liver, and kidney transplant recipients and their transplant provider. The majority (15/25, 60%) of children were <5 years old (Tables 1 and 2).

XSL•FO
RenderX

Table . Parent and guardian demographic characteristics (n=25).

Characteristic	Value, n (%)
Gender identity	
Female	20 (80)
Male	5 (20)
Ethnicity	
Caucasian	18 (72)
African American	2 (8)
Asian	2 (8)
Hispanic-Latino	1 (4)
Other or prefer not to answer	2 (8)
Center of care	
Children's Hospital Colorado	19 (76)
Children's Hospital of Philadelphia	3 (12)
Ann and Robert H. Lurie Children's Hospital	3 (12)
How far do you live from your transplant center?	
<1 hour drive	8 (32)
1 - to 3-hour drive	10 (40)
>3-hour drive	3 (12)
We are in a different state than our center and need to fly to our center	4 (16)
How would you describe the town you live in?	
Rural	7 (28)
Suburban	13 (52)
Urban	5 (20)
How old is your child?	
0 - 11 months	4 (16)
12 months to 5 years	11 (44)
6 - 11 years	4 (16)
12 - 18 years	6 (24)
What organ is your child listed for?	
Heart	4 (16)
Liver	14 (56)
Kidney	7 (28)

XSL•FO
RenderX

Table . Provider demographic characteristics (n=15; there were providers who had multiple patients in the study).

Characteristic	Value, n (%)
Role on transplant team	
Cardiologist, hepatologist, or nephrologist	4 (27)
Transplant fellow	1 (7)
Transplant nurse coordinator	10 (67)
What transplant center do you work with?	
Children's Hospital Colorado	10 (67)
Children's Hospital of Philadelphia	2 (13)
Ann and Robert H. Lurie, Children's Hospital	3 (20)
How long have you worked as an attending transplant physician, transplant fellow, or transplant nurse coordinator?	
<1 year	3 (20)
1 - 5 years	5 (36)
6 - 10 years	3 (20)
>10 years	3 (20)
Gender identity	
Female	14 (93)
Male	1 (7)
Ethnicity	
Caucasian	11 (73)
African American	0 (0)
Asian	1 (7)
Hispanic-Latino	1 (7)
Other or prefer not to answer	2 (13)

Effectiveness of the App

Vaccine records were obtained on the day of transplant; 96% (24/25) of children were up to date on all age-appropriate vaccines. The 1 patient who was not up to date was a 15-year-old who was missing a second human papillomavirus (HPV) shot. The patient's team had specifically chosen to delay the second HPV vaccine until after transplant because the patient was on high-dose steroids for autoimmune hepatitis and was unlikely to mount a strong immune response to a pretransplant vaccine.

Feasibility of the App

App Usage

Parents and providers were instructed to use the Immunize PediatricTransplant app until the day of their child's transplant (or until the end of the study on April 1, 2024, whichever came first). No children were transplanted in less than a month, 27% (6/22) were transplanted in 1 - 3 months, 23% (5/22) in 4 - 6 months, and 50% (11/22) in 7 or more months.

Technical Difficulties While Using the App

Only 23% (5/22) of parents and 21% (3/14) of providers reported technical difficulties with the app. Difficulties described included initial log on (n=5), entry of historical vaccines or vaccines given while using the app (n=2), and communication with team members through the portal (n=1).

Acceptability of the App

Of the 25 parents and guardians and 15 providers, 88% (22/25) and 93% (14/15), respectively, completed an exit questionnaire. Of these, 97% (35/36) agreed or strongly agreed that they felt knowledgeable about vaccine use in the pretransplant period after using the app and 86% (31/36) agreed or strongly agreed that communication around vaccines between families and providers was good after using the app. Users rated the following features as extremely or moderately useful: the centralized vaccine record (n=30, 83%), the automated vaccine reminders (n=24, 67%), and the written educational information (n=21, 58%; Table 3). Further, 91% (20/22) of parents and 79% (11/14) of providers recommended the app to future families and transplant teams with children awaiting a transplant.

XSL•FO
RenderX

Table . Acceptability of the various features of the app (n=36; parents, guardians, and providers combined).

	Extremely or moderately useful, n (%)	Somewhat useful, n (%)	Not useful, n (%)
Written educational information in the app	21 (58)	12 (33)	3 (8)
Communication tool (chat feature)	18 (50)	9 (25)	9 (25)
Automated vaccine reminders	24 (67)	8 (22)	4 (11)
Centralized vaccine record	30 (83)	2 (6)	4 (11)

Suggestions for App Improvement

The open-ended responses to the question of "Do you have any suggestions for app improvement?" are listed in Textbox 1. The most common suggestion (n=6) was to populate the vaccine history from the EMR or other vaccine databases such as immunization information systems (IISs).

Textbox 1. Open-ended responses in the exit questionnaire to "Do you have any suggestions for app improvement?"

Responses

- Pull vaccine records from electronic medical record
- Automatically populate the vaccination history
- When entering vaccines, it would be nice to enter more than one at a time
- Scan vaccine documents or have a professional health care provider enter in the vaccine data, not the family, there's too much room for error
- Allow downloads of vaccine records from state immunization information systems (IIS)
- Connect the app to a vaccine data base
- Space for additional vaccines (extra COVID doses or repeat series)
- Make the live chat a more noticeable feature
- Improve ease of use for vaccine entry
- Connect primary care provider with the app
- Have a way to submit option of family refusal to avoid getting reminders about a vaccine your family is refusing
- Tech support for users
- Create a way for medical providers to silence alerts for vaccines that family is refusing
- App appearance
- If we could have the primary care pediatrician participate, that would be great
- Encourage primary care pediatricians to use it for communication
- Integration with vaccine databases such as the Colorado Department of Public Health and Environment

Discussion

Principal Findings

We sought to beta test the Immunize PediatricTransplant mHealth app to understand whether (1) it could help address the problem of underimmunization of pediatric solid organ transplant candidates and (2) it would be feasible and acceptable to parent and transplant provider users. The most significant finding of the study is the success of the app in getting nearly all children up to date on age-appropriate vaccines by the time of transplant. Historically, the majority of pediatric transplant candidates have not been up to date on age-appropriate vaccines at the time of transplant. In a study of 281 children who underwent a pediatric liver transplant across the United States between August 2017 and August 2018, only 19% were up to date on age-appropriate vaccines on the day of transplant and amongst those who were not up to date, 51% were missing 4 or more immunizations [6]. Increasing immunization rates pretransplant could significantly improve posttransplant outcomes by decreasing posttransplant VPIs and the associated morbidity, mortality, and increased medical costs [4,5].

While the app was overall feasible and acceptable to users (31/36, 86% of users recommended the app to future parents and providers), in the open-ended question in the exit questionnaire asking for suggestions for app improvement we learned that users did not want to have to enter vaccine records themselves into the app (in real-world use, there would be no research coordinator to enter vaccines). Multiple users recommended that the app connect directly with the EMR or IISs to retrieve accurate vaccine data.

There are several apps currently available in the App Store to promote childhood vaccination [24]. To our knowledge,

XSL•FO
RenderX

Immunize PediatricTransplant is the first app designed specifically for a group of children with complex medical needs at a higher risk for acquiring VPIs. In a 2020 systemic review of 28 studies evaluating 25 immunization apps, 4 studies described significant benefits of implementation of the app and 4 studies reported a significant impact on knowledge and vaccine decision-making [25]. Our study adds additional evidence that an mHealth app can improve vaccine rates in a high-risk population of children. In addition to providing vaccine education, record keeping, and reminder systems, Immunize PediatricTransplant also includes a communication portal for families and providers to directly communicate in real time about vaccine questions.

Future Directions

Based on user feedback, in the future we will explore utilizing application programming interfaces (APIs) like Health Level 7 FHIR (Fast Healthcare Interoperability Resources) [26] and SMART on FHIR [27] to allow the app to directly connect and communicate with the EMR and IISs [28]. While APIs are legal according to the 21st Century Cures Act [29], it would be critical to ensure compliance with Health Insurance Portability and Accountability Act (HIPAA) regulations protecting patient privacy. We will then need to evaluate the tool in a clinical trial (phase 3 of the mHealth Agile Development and Evaluation Lifecycle) [22] to see if it results in up-to-date immunization status at the time of transplant for a diverse population of pediatric transplant candidates across the country. As this is a "low-risk app" that does not pose any risk to user safety if it does not function (providers should still be reviewing vaccines as part of standard medical care), an alternative to a formal clinical trial would be ongoing A/B testing where a group of users trial a new version of the app that contains minor variations to a specific component of the app [22]. Based on user feedback, ongoing iterations of the app could be made. In future studies, we will modify the primary end point to not classify patients with deliberately missed vaccines (as instructed by the medical team) as not up to date.

After clinical trial or A/B testing, postmarket surveillance will be important to gather ongoing feedback about if the app remains useful and effective and whether further improvements are needed to keep it relevant as new vaccines come into the market [22]. In future research, it will be important to expand the qualitative sections of the exit questionnaire or perform postintervention interviews or focus groups to gain a deeper understanding of the users' experiences, likes, and dislikes about the app and recommendations for future app users including primary care physicians.

Limitations

Multiple limitations exist in this study beta testing the Immunize PediatricTransplant app. Our sample size was small (only 25 dyads of parents and transplant providers, which equated to 36 app users) and the majority of participants came from 1 large academic pediatric transplant center. Additionally, the app was only trialed by English-speaking parents and providers, limiting generalizability to all transplant parents and providers. Participation in the study was voluntary, which could introduce selection bias where people who were more interested in or comfortable with mHealth technology were more likely to participate. Similarly, there could be social desirability bias, where people were more likely to report finding the app to be beneficial in an attempt to please researchers. Finally, it is possible that simply by having parents and providers participate in an intervention-based research study, it increased attention and awareness to vaccines, in turn increasing the amount of transplant recipients with up-to-date vaccine statuses. As this was not a randomized trial, there was no control group to compare with. These limitations are all inherent to beta testing, and the app will need to go through future phase 3 clinical trial evaluation.

Conclusions

Each year, the United States invests $1.2 billion dollars in pediatric solid organ transplants [30]; however, we fail to protect our investment by not ensuring that these children are fully immunized. A novel mobile app, Immunize PediatricTransplant, has now been developed to help overcome barriers to pretransplant vaccines. The app was feasible and acceptable to parent and provider users and successfully resulted in nearly all children being up to date on age-appropriate vaccines by the time of transplant. However, the app was reliant on a research team member entering historical vaccine information, which is not practical for real-world use. In future iterations, we will need to explore whether APIs could be leveraged to extract vaccine information from the EMR or IISs and directly input these data into the app.

Acknowledgments

The authors would like to acknowledge all transplant stakeholders who participated in this research. AGF is supported by the Agency for Healthcare Research and Quality (grant K08 HS026510). Contents are the authors' sole responsibility and do not necessarily represent official National Institutes of Health (NIH) views.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

KW and CB are cofounders of CANImmunize Inc. KW is a paid member of the independent data safety board for the Medicago vaccine trial.

XSL•FO
RenderX

References

1. Burroughs M, Moscona A. Immunization of pediatric solid organ transplant candidates and recipients. Clin Infect Dis 2000 Jun;30(6):857-869. [doi: 10.1086/313823] [Medline: 10852737]
2. Kaplan LJ, Daum RS, Smaron M, McCarthy CA. Severe measles in immunocompromised patients. JAMA 1992 Mar 4;267(9):1237-1241. [doi: 10.1001/jama.1992.03480090085032] [Medline: 1538561]
3. McGregor RS, Zitelli BJ, Urbach AH, Malatack JJ, Gartner JC Jr. Varicella in pediatric orthotopic liver transplant recipients. Pediatrics 1989 Feb;83(2):256-261. [Medline: 2536479]
4. Feldman AG, Sundaram SS, Beaty BL, Kempe A. Hospitalizations for respiratory syncytial virus and vaccine-preventable infections in the first 2 years after pediatric liver transplant. J Pediatr 2017 Mar;182:232-238. [doi: 10.1016/j.jpeds.2016.12.021] [Medline: 28088400]
5. Feldman AG, Beaty BL, Curtis D, Juarez-Colunga E, Kempe A. Incidence of hospitalization for vaccine-preventable infections in children following solid organ transplant and associated morbidity, mortality, and costs. JAMA Pediatr 2019 Mar 1;173(3):260-268. [doi: 10.1001/jamapediatrics.2018.4954] [Medline: 30640369]
6. Feldman AG, Sundaram SS, Beaty BL, Torres R, Curtis DJ, Kempe A. Immunization status at the time of liver transplant in children and adolescents. JAMA 2019 Nov 12;322(18):1822-1824. [doi: 10.1001/jama.2019.14386] [Medline: 31714979]
7. Feldman AG, Marsh R, Kempe A, Morris MA. Barriers to pretransplant immunization: a qualitative interview study of pediatric solid organ transplant stakeholders. J Pediatr 2020 Dec;227:60-68. [doi: 10.1016/j.jpeds.2020.07.038] [Medline: 32681988]
8. Krishna S, Boren SA. Diabetes self-management care via cell phone: a systematic review. J Diabetes Sci Technol 2008 May;2(3):509-517. [doi: 10.1177/193229680800200324] [Medline: 19885219]
9. Scherr D, Kastner P, Kollmann A, et al. Effect of home-based telemonitoring using mobile phone technology on the outcome of heart failure patients after an episode of acute decompensation: randomized controlled trial. J Med Internet Res 2009 Aug 17;11(3):e34. [doi: 10.2196/jmir.1252] [Medline: 19687005]
10. Kao CK, Liebovitz DM. Consumer mobile health apps: current state, barriers, and future directions. PM R 2017 May;9(5S):S106-S115. [doi: 10.1016/j.pmrj.2017.02.018] [Medline: 28527495]
11. Domin A, Spruijt-Metz D, Theisen D, Ouzzahra Y, Vögele C. Smartphone-based interventions for physical activity promotion: scoping review of the evidence over the last 10 years. JMIR Mhealth Uhealth 2021 Jul 21;9(7):e24308. [doi: 10.2196/24308] [Medline: 34287209]
12. Tabi K, Randhawa AS, Choi F, et al. Mobile apps for medication management: review and analysis. JMIR Mhealth Uhealth 2019 Sep 11;7(9):e13608. [doi: 10.2196/13608] [Medline: 31512580]
13. Wilson K, Atkinson KM, Penney G. Development and release of a national immunization app for Canada (ImmunizeCA). Vaccine 2015 Mar;33(14):1629-1632. [doi: 10.1016/j.vaccine.2015.02.022] [Medline: 25704801]
14. Wilson K, Atkinson KM, Westeinde J. Apps for immunization: leveraging mobile devices to place the individual at the center of care. Hum Vaccin Immunother 2015;11(10):2395-2399. [doi: 10.1080/21645515.2015.1057362] [Medline: 26110351]
15. Wilson K, Atkinson KM, Deeks SL, Crowcroft NS. Improving vaccine registries through mobile technologies: a vision for mobile enhanced Immunization information systems. J Am Med Inform Assoc 2016 Jan;23(1):207-211. [doi: 10.1093/jamia/ocv055] [Medline: 26078414]
16. CANImmunize. Apple App Store. URL: https://itunes.apple.com/ca/app/immunize-canada/id574951915?mt=8 [accessed 2025-04-02]
17. Feldman AG, Atkinson K, Wilson K, Kumar D. Underimmunization of the solid organ transplant population: an urgent problem with potential digital health solutions. Am J Transplant 2020 Jan;20(1):34-39. [doi: 10.1111/ajt.15605] [Medline: 31553135]
18. Zaidi S, Shaikh SA, Sayani S, et al. Operability, acceptability, and usefulness of a mobile app to track routine immunization performance in rural Pakistan: interview study among vaccinators and key informants. JMIR Mhealth Uhealth 2020 Feb 13;8(2):e16081. [doi: 10.2196/16081] [Medline: 32053115]
19. Mekonnen ZA, Gelaye KA, Were M, Tilahun B. Effect of mobile phone text message reminders on the completion and timely receipt of routine childhood vaccinations: superiority randomized controlled trial in northwest Ethiopia. JMIR Mhealth Uhealth 2021 Jun 15;9(6):e27603. [doi: 10.2196/27603] [Medline: 34128813]
20. Mekonnen ZA, Gelaye KA, Were MC, Gashu KD, Tilahun BC. Effect of mobile text message reminders on routine childhood vaccination: a systematic review and meta-analysis. Syst Rev 2019 Jun 28;8(1):154. [doi: 10.1186/s13643-019-1054-0] [Medline: 31253186]
21. Feldman AG, Moore S, Bull S, et al. A smartphone app to increase immunizations in the pediatric solid organ transplant population: development and initial usability study. JMIR Form Res 2022 Jan 13;6(1):e32273. [doi: 10.2196/32273] [Medline: 35023840]
22. Wilson K, Bell C, Wilson L, Witteman H. Agile research to complement agile development: a proposal for an mHealth research lifecycle. NPJ Digit Med 2018;1:46. [doi: 10.1038/s41746-018-0053-1] [Medline: 31304326]

XSL•FO
RenderX

23. Wodi AP, Murthy N, McNally V, Cineas S, Ault K. Advisory Committee on Immunization Practices Recommended Immunization Schedule for Children and Adolescents Aged 18 Years or Younger - United States, 2023. MMWR Morb Mortal Wkly Rep 2023 Feb 10;72(6):137-140. [doi: 10.15585/mmwr.mm7206a1] [Medline: 36757872]
24. Vaccination. Google Play Store. URL: https://play.google.com/store/search?q=vaccination&c=apps&gl=GB [accessed 2024-12-10]
25. de Cock C, van Velthoven M, Milne-Ives M, Mooney M, Meinert E. Use of apps to promote childhood vaccination: systematic review. JMIR Mhealth Uhealth 2020 May 18;8(5):e17371. [doi: 10.2196/17371] [Medline: 32421684]
26. What is HL7 FHIR? Office of the National Coordinator for Health Information Technology. URL: https://www.healthit.gov/sites/default/files/page/2021-04/What%20Is%20FHIR%20Fact%20Sheet.pdf [accessed 2024-12-10]
27. SMART on FHIR. SMART Health IT. URL: https://docs.smarthealthit.org [accessed 2024-12-10]
28. What is an API? Amazon Web Services. URL: https://aws.amazon.com/what-is/api/ [accessed 2024-01-11]
29. 21st Century Cures Act, HR 34, 114th Cong (2015). URL: https://www.congress.gov/bill/114th-congress/house-bill/34/text [accessed 2025-02-18]
30. Bentley TS, Phillips SJ. 2017 U.S. organ and tissue transplant cost estimates and discussion. : Milliman; 2017 URL: https://www.milliman.com/en/insight/2017-us-organ-and-tissue-transplant-cost-estimates-and-discussion [accessed 2025-04-02]

Abbreviations:

API: application programing interface
CDC: Centers for Disease Control
EMR: electronic medical record
FHIR: Fast Healthcare Interoperability Resources
HIPAA: Health Insurance Portability and Accountability Act
HPV: human papillomavirus
IIS: immunization information system
mHealth: mobile health
VPI: vaccine-preventable infection

Edited by A Mavragani; submitted 16.12.24; peer-reviewed by IA Eide; revised version received 18.02.25; accepted 20.02.25; published 15.04.25.

Please cite as:
Feldman AG, Beaty BL, Moore SL, Bull S, Wilson K, Atkinson KM, Bell C, Denize KM, Kempe A
Feasibility, Acceptability, and Effectiveness of a Smartphone App to Increase Pretransplant Vaccine Rates: Usability Study
JMIR Form Res 2025;9:e68855
URL: https://formative.jmir.org/2025/1/e68855
doi:10.2196/68855

© Amy G Feldman, Brenda L Beaty, Susan L Moore, Sheana Bull, Kumanan Wilson, Katherine M Atkinson, Cameron Bell, Kathryn M Denize, Allison Kempe. Originally published in JMIR Formative Research (https://formative.jmir.org), 15.4.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Development of an eHealth Mindfulness-Based Music Therapy Intervention for Adults Undergoing Allogeneic Hematopoietic Stem Cell Transplantation: Qualitative Study

Sara E Fleszar-Pavlovic[1], PhD; Blanca Noriega Esquives[1], MD, PhD; Padideh Lovan[1,2], PhD, RD, LDN; Arianna E Brito[1], BS; Ann Marie Sia[3]; Mary Adelyn Kauffman[1], DMA, MT-BC, NMT; Maria Lopes[4], BA; Patricia I Moreno[1,5], PhD; Tulay Koru-Sengul[1,5], PhD, MHS; Rui Gong[1,6], PhD; Trent Wang[1,6], DO, MPH; Eric D Wieder[1,6], PhD; Maria Rueda-Lara[1,7], MD; Michael Antoni[1,8], PhD; Krishna Komanduri[9], MD; Teresa Lesiuk[1,10], PhD, MT-BC; Frank J Penedo[1,8], PhD

[1]Sylvester Comprehensive Cancer Center, University of Miami Miller School of Medicine, Miami, FL, United States
[10]Frost School of Music, University of Miami, Coral Gables, FL, United States
[2]School of Nursing and Health Studies, University of Miami Miller School of Medicine, Miami, FL, United States
[3]Department of Undergraduate Research, University of Miami, Coral Gables, FL, United States
[4]Department of Psychology, College of Arts and Sciences, University of Miami, Coral Gables, FL, United States
[5]Department of Public Health Sciences, University of Miami Miller School of Medicine, Miami, FL, United States
[6]Department of Medicine, University of Miami Miller School of Medicine, Miami, FL, United States
[7]Department of Psychiatry and Behavioral Sciences, University of Miami Miller School of Medicine, Miami, FL, United States
[8]Departments of Psychology and Medicine, University of Miami Miller School of Medicine, 1120 NW 14th Street, Miami, FL, United States
[9]Department of Medicine, Helen Diller Family Comprehensive Cancer Center, University of California, San Francisco, San Francisco, CA, United States

Corresponding Author:
Frank J Penedo, PhD
Sylvester Comprehensive Cancer Center, University of Miami Miller School of Medicine, Miami, FL, United States

Abstract

Background: Allogeneic hematopoietic stem cell transplantation (allo-SCT) is an effective treatment for various hematologic cancers, though it often results in severe side effects and psychological distress, which can negatively impact health outcomes. Integrative therapies like mindfulness-based stress reduction (MBSR), mindfulness meditation (MM), and music therapy (MT) yield promising results in enhancing both psychosocial outcomes (eg, reducing anxiety and depression) and physiological adaptation (eg, decreasing inflammation) in cancer patients.

Objective: We developed and refined, using focus groups and environmental and field testing, an eHealth-delivered mindfulness-based music therapy (eMBMT) intervention aimed at improving health-related quality of life, symptom burden (ie, pain, fatigue, and sleep), disease activity (ie, chronic graft-versus-host disease, cytomegalovirus activation, and infections) and psychosocial (ie, depression, anxiety, and cancer-specific distress) and physiological adaptation (ie, inflammation and immune reconstitution) tailored to adults receiving allo-SCT.

Methods: eMBMT intervention content is grounded in MT, MM, and MBSR, developed by a multidisciplinary team, and adapted for adults undergoing allo-SCT. eMBMT content was refined through focus groups and usability and field testing. Focus groups used a semistructured interview guide, while field testing used the “think aloud” method. Usability was evaluated using the 30-item Usefulness, Satisfaction, and Ease of Use (USE) questionnaire. Descriptive statistics analyzed the USE questionnaire and participant characteristics, while rapid qualitative analysis was applied to focus groups and field-testing sessions. Survivors eligible to participate in the focus groups and usability and field testing were adults (>18 years old) who received an allo-SCT (<36 months) for myelodysplastic syndrome, acute myeloid leukemia, or chronic myeloid leukemia, and were in remission for greater than 3 months.

Results: During the focus groups, participants (n=11; mean age 43.6, SD 17.8 years) provided qualitative feedback highlighting the shock of diagnosis, challenges during hospitalization, and coping strategies posttreatment. The eMBMT platform received positive evaluations for usefulness (mean 6.47, SD 0.29), ease of use (mean 6.92, SD 0.60), and satisfaction (mean 6.16, SD 0.82). Key themes from field testing highlighted the significance of social support, hope, and maintaining an active lifestyle. Suggestions for improvement included incorporating more representative content, reducing text, enhancing guidance, offering diverse music options, and streamlining blood sample collection.

XSL•FO
RenderX

Conclusions: The eMBMT intervention is a comprehensive, user-friendly eHealth tool tailored to the unique needs of allo-SCT patients. The positive feedback and identified areas for improvement underscore its potential to enhance well-being, symptom management, and overall quality of life for cancer survivors. A future pilot randomized controlled trial will further evaluate the feasibility, acceptability, and preliminary efficacy of the eMBMT intervention in improving health-related quality of life, symptom burden, disease activity, and psychosocial and physiological adaptation.

Trial Registration: ClinicalTrials.gov NCT05968963; https://clinicaltrials.gov/study/NCT05968963

(JMIR Form Res 2025;9:e65188) doi:10.2196/65188

KEYWORDS

allogeneic stem cell transplantation; hematologic malignancy; bone marrow transplant; mindfulness-based music therapy; mindfulness; music therapy; eHealth; music therapy intervention; adult; adolescence; allogeneic; stem cell transplantation; stem cell; transplantation; qualitative study; treatment; hematologic cancers; psychological distress; side effects; mindfulness-based stress reduction; stress reduction; anxiety; depression; diagnosis; blood sample collection; eHealth tool; quality of life; cancer survivors

Introduction

In the United States, hematologic malignancies account for approximately 10% of all new cancer diagnoses and 9.4% of cancer-related deaths annually [1]. It is projected that over 180,000 new cases of malignant hematologic cancers will be diagnosed in 2024 [2]. Allogeneic hematopoietic stem cell transplantation (allo-SCT) is an effective therapy for various hematologic cancers, including acute myeloid leukemia, chronic myeloid leukemia, and myelodysplastic syndrome [3]. Allo-SCT is a multiphase therapy starting with a pretransplant evaluation and finding a compatible donor. This is followed by intensive conditioning (chemotherapy, radiation, or both) to destroy cancer cells, suppress the immune system, and prepare for stem cell engraftment. The patient then receives the donor's stem cells intravenously and enters the posttransplant phase, focusing on engraftment and monitoring for complications like graft-versus-host disease (GVHD) and infections [3].

Despite its curative potential, allo-SCT may be accompanied by negative side effects (eg, nausea, vomiting, pain, fatigue, mood disturbance, and risk of infections) [4]. In addition, poor psychosocial functioning and emotional distress (eg, anxiety, depression, and social isolation) can persist well beyond active treatment [5-7]. These negative psychological factors, common among allo-SCT recipients [5,6], are associated with higher mortality rates and less favorable health outcomes in the first-year posttreatment [8-10], including greater rates of GVHD and infection, longer hospitalization stays, and higher readmission rates [11]. Furthermore, research indicates that psychological distress can affect physiological responses to treatment such as increased inflammation [12,13]. For example, in patients receiving allo-SCT, pretransplant anxiety, depression, and stress have been associated with delayed white blood cell recovery, slower immune function recovery, and shorter survival. In contrast, pretransplant optimism, positive affect, and social support correlate with a shorter time to engraftment and reduced incidence of GVHD [14-17]. Therefore, interventions targeting psychological distress and physical burden in allo-SCT recipients may enhance health-related quality of life (HRQoL), improve patient-reported outcomes, and support immune recovery and overall health in cancer survivors.

A growing body of research suggests that mindfulness-based interventions, including mindfulness-based stress reduction (MBSR) and mindfulness meditation (MM) yield promising results in enhancing both psychosocial outcomes (eg, reducing anxiety and depression) and physiological adaptation (eg, decreasing inflammation) in cancer patients [18-21]. The core principle of mindfulness involves maintaining nonjudgmental awareness of the present moment, which includes recognizing and accepting one's thoughts, emotions, and bodily sensations as they arise. Practicing mindfulness fosters a greater sense of self-awareness and emotional regulation, which can contribute to improved mental and physical well-being [22]. Although highly underused, mindfulness-based interventions have proven effective in alleviating cancer symptoms, cancer-related distress, and depression in various patient populations. Mindfulness-based interventions also enhance self-efficacy, coping skills, and HRQoL in cancer patients [12,20,23-27] and individuals with other chronic diseases [28-31].

In addition to mindfulness, music therapy (MT) is a well-recognized approach for addressing various psychosocial and physiological needs in clinical populations [32-44]. It has demonstrated effectiveness in reducing stress and enhancing emotional well-being among cancer patients [35,45-49]. MT offers numerous benefits for cancer patients by using both interactive (eg, singing) and receptive techniques (eg, listening to music and music imagery) to improve mood, reduce stress, pain, and anxiety, as well as promote relaxation [50]. Although most findings suggest psychological improvement associated with MT, a recent meta-analysis of controlled trials provided evidence that MT can significantly improve physical symptoms (eg, fatigue) and reduce pain in cancer patients [51,52]. MT can also serve as a foundation for planning effective rehabilitation programs and improving overall quality of life [50]. Using MT is known to be a low-risk, minimally invasive, and cost-effective approach to managing patients' well-being particularly in the often-stressful environment of a cancer clinic [53].

Although previous research has demonstrated that practices related to mindfulness and MT can positively affect fatigue, pain, anxiety, and HRQoL in cancer patients [54,55], there is a lack of well-powered randomized controlled trials and reproducible interventions specific for allo-SCT patients. In addition, significant gaps persist in our understanding of how

XSL•FO
RenderX

specific components of mindfulness and MT impact psychosocial and physiological adaptation, HRQoL, and overall health outcomes in this population. Therefore, we designed an eHealth-delivered mindfulness-based music therapy (eMBMT) aimed at reducing symptom burden and improving HRQoL in patients undergoing allo-SCT. We selected an eHealth intervention based on its proven advantages, including patient accessibility and convenience, improved communication, and reduced burden on health care systems. Overall eHealth interventions offer a valuable approach to complement traditional cancer care [56]. This study aims to refine and finalize, using focus groups and environmental and field testing, the newly developed eMBMT intervention.

Methods

Intervention Content Development

The eMBMT content was developed by a multidisciplinary team (eg, board-certified music therapists [MT-BC], psychologists, oncologists, and psychiatrists) and is based on MT, MM, MBSR [57], and Lesiuk's [58,59] MBMT pilot trial with breast cancer survivors. eMBMT was further adapted for patients undergoing allo-SCT and structured to be delivered through BrightOutcome's SmartManage: Tools for Health Living [60,61] platform. SmartManage is a patient-centered web application that is easy to use, Health Insurance Portability and Accountability Act (HIPAA) compliant, and accessible through multiple platforms (ie, tablets and smartphones). SmartManage serves as the primary platform for hosting the MBMT content, organized into 8 modules corresponding to sessions led by the MT-BC. SmartManage enables participants to engage in real time during sessions and supports independent practice by providing access to session materials, guided exercises, and additional resources for extended learning. Participants were encouraged to engage with SmartManage outside of scheduled sessions, using its resources to review session materials and independently practice intervention components at their own pace, thereby reinforcing the therapeutic process. The eMBMT program consists of eight 60-minute sessions led by a MT-BC. The sessions integrate MT and mindfulness attitudes (ie, nonjudging, patience, letting go, acceptance, trust, beginners mind, and nonstriving) developed by Kabat-Zinn [57]. MT-BCs guide patients in engaging with video and audio materials, practicing active and receptive music therapy exercises, incorporating music and journaling, and participating in video conferencing, all aimed at fostering mindfulness attitudes through music therapy. eMBMT's 8 tailored sessions span the allo-SCT treatment and recovery journey beginning upon hospital admission (approximately 7 days before transplant) and ending postdischarge (approximately 75 - 115 days posttransplant).

eMBMT begins with an orientation session to provide an overview of the program and its aims. The following 7 sessions are aligned with the patient's treatment trajectory, each emphasizing a specific mindfulness attitude to be practiced throughout their allo-SCT treatment: (1) nonjudging encourages evaluating stimuli without habitual reactions, whether from internal dialogue, sensations, or external events; (2) patience helps manage worry by allowing events to unfold over time; (3) letting go involves accepting things as they are and being open to change; (4) acceptance focuses on viewing one's experience without comparisons and allowing recovery to unfold at its own pace; (5) trust emphasizes believing in oneself and empowering self-reliance over thoughts; (6) beginner's mind encourages seeing newness in familiar environments, fostering a sense of wonder, and shedding negative perceptions; and (7) nonstriving promotes the willingness to try new experiences and suspend critical judgments, particularly through new music experiences. MT-BCs focus on guiding patients through each session to first learn the attitude through music exercises, then how to apply the attitude to potentially improve their treatment journey.

Intervention Refinement

Following the initial content development, we used focus groups and usability and field-testing sessions to refine and finalize the eMBMT content and intervention.

Sampling and Recruitment

Participants were recruited by (a) physician referral or (b) the University of Miami's consent-to-contact database to participate in a focus group and a subsequent one-on-one usability and field-testing session. Patients at the University of Miami Health System can voluntarily enroll and consent to be contacted about clinical study opportunities for which they qualify through the Consent to Contact Registry managed by the University of Miami Clinical and Translational Science Institute. Potential participants were identified using the Clinical and Translational Science Institute Consent-to-Contact Registry through the generation authorized patient lists based on International Classification of Diseases (ICD-10) codes. Allo-SCT survivors were eligible to participate if they were (1) aged >18 years; (2) received an allo-SCT for myelodysplastic syndrome, acute myeloid leukemia, or chronic myeloid leukemia <36 months; and (3) were in remission >3 months. Survivors who (1) had a history of severe psychiatric illness (eg, psychosis, active suicidality, or inpatient treatment) in the past 12-months, (2) severe cognitive impairment (per the Short Portable Mental Status Questionnaire) [62], or (3) hearing impairment were excluded from the study. Patients who were referred by their physician or were identified by the consent-to-contact list were contacted by telephone by study staff to assess interest in participating and eligibility. Of the survivors (n=64) screened for eligibility, 34 were ineligible, 11 declined screening, and 7 did not consent to participate. A total of 12 survivors agreed to participate and provided informed consent; however, 1 participant was ill and unable to attend the focus group. The study team conducted 2 focus groups (focus group 1: n=6 and focus group 2: n=5) and 9 one-on-one usability and field-testing sessions.

Procedures

Before attending a focus group, participants completed a survey assessing sociodemographic (eg, age, gender, race, and income) and medical characteristics (eg, cancer type and time in remission). Focus groups were conducted through Zoom (Zoom Communications) and facilitated by 2 MT-BCs and study staff

XSL•FO
RenderX

trained in qualitative methods. The aim of the focus groups was to understand the needs and burdens of allo-SCT patients and gather qualitative feedback to inform eMBMT content and platform design. Facilitators followed a semistructured interview guide including open-ended questions assessing: (1) challenges associated with receiving an allo-SCT (eg, uncertainty of the disease course, treatment-related side effects, and management of emotional and physical symptoms), (2) use of eHealth-based delivery for mindfulness, music therapy, and mindfulness-based music exercises, and (3) preferences for eMBMT length and detail of content. Facilitators also presented 2 modules from the eMBMT SmartManage platform and participants were asked about design and layout preferences (eg, theme, style, colors, and fonts), content detail and type (eg, text, images, and videos), and functionality features (eg, navigation, accessibility, and user interaction). Each focus group was facilitated by 2 experienced moderators (ie, "facilitators") with training in qualitative research methods (SEF-P, BNE, TL, and MAK). To ensure balanced participation and that all voices were heard, the facilitators used strategies such as directly inviting quieter participants to share their thoughts, using prompts to encourage diverse perspectives, and managing dominant voices to maintain an inclusive environment [63,64]. Focus groups were used to capture a diverse range of perspectives and facilitate interactive discussions among participants, allowing for the exploration of shared experiences and unique viewpoints related to the intervention. This approach was chosen because participants could build on each other's ideas, providing deeper insights into coping strategies and preferences. By leveraging the group dynamic, focus groups enabled the identification of common challenges and individual needs, ensuring the intervention was informed by both collective and personal experiences. Focus groups lasted approximately 3 hours and were audio recorded and detailed notes were taken to supplement the recordings.

Participants were contacted 1 week post–focus group participation to complete a usability and field-testing session. The usability assessment evaluated the ease of learning and identified platform design issues. Field testing evaluated both technical (eg, technical problems) and functional (eg, application use and integrating eMBMT into daily life) reliability. Sessions were conducted in-person or through Zoom and facilitated by study staff trained in usability and field testing of eHealth platforms. For usability testing, participants completed the 30-item USE questionnaire [65] which contains 4 subscales assessing usefulness (eg, "It helps me be more effective"), ease of use (eg, "It is easy to use"), ease of learning (eg, "I learned to use it quickly"), and satisfaction (eg, "I am satisfied with it") on an 8-point Likert scale (1=strongly disagree to 8=strongly agree). For field testing, we used the "think aloud" method [66], which encourages participants to vocalize any thoughts or comments while performing or immediately after performing a task, such as reading or watching a video. Participants were asked to share their thoughts and reactions as they explored the eMBMT and platform. Finally, participants were prompted with questions addressing technical reliability (eg, technology problems) and functional reliability (eg, application use and integrating eMBMT into daily life). The usability and field-testing sessions lasted approximately 3 hours and were audio-recorded.

Data Analysis

The Usefulness, Satisfaction, and Ease of Use (USE) questionnaire [65] and sociodemographic and medical characteristics were analyzed with descriptive statistics. We applied rapid qualitative analysis (RQA) to analyze the focus groups and usability and field-testing sessions following Watkins' guidelines [67]. Relative to traditional qualitative analysis methods, such as thematic analysis and grounded theory, which demand extensive time for tasks like multiple transcript readings and detailed coding, RQA is a more efficient and less time-demanding approach. RQA methods involve team members individually summarizing key points from qualitative data, and combining individual data by categorizing related key points to produce relevant themes [68]. RQA is a pragmatic, valid, and reliable method, which is crucial for guiding the refinement of digital interventions [68]. Study staff were trained on RQA, developed a coding matrix, individually coded the transcripts, and finally, met to complete a comprehensive final matrix summarizing data to facilitate the identification of themes. Themes were developed through iterative coding of the interview transcripts, with the team working to identify patterns and nuances in participant responses. While some themes aligned with the interview prompts, further analysis captured additional insights into coping mechanisms and patient preferences, which were critical for tailoring the intervention.

Ethical Considerations

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. We complied with the following ethical considerations. First, we recruited and obtained informed consent from all participants before the focus groups and usability and field-testing sessions, as approved by the University of Miami's Institutional Review Board (eProst 20230726). Participants were informed that their participation was voluntary, the focus groups and usability and field-testing sessions would be recorded, the collected information would be deidentified and securely stored, and they would receive compensation of US $50 for completing the focus group and US $50 for completing the usability and field-testing session, for a total of US $100. Second, we ensured that all transcript data were deidentified by replacing participant names with a unique identifier. All collected data are securely stored with access restricted to authorized researchers. Third, all researchers completed research compliance training through the Collaborative Institutional Training Initiative, ensuring adherence to best practices and ethical considerations in working with participants. This project is funded by the National Cancer Institute (1R61CA263335-01A1) and registered on ClinicalTrials.gov (NCT05968963).

Results

Participant Characteristics

Table 1 presents participant demographics and medical characteristics. Participants (N=11; mean age 43.6, SD 17.8; range 19 - 76 years) were female (55%, n=6), White (91%,

XSL•FO
RenderX

n=10), Hispanic (73%, n=8), with a history of acute myeloid leukemia (100%, n=11), and a mean time of 1.64 (SD 0.81) years post-SCT at the time of the focus groups.

Table . Demographics and medical characteristics of allogeneic stem cell transplantation (allo-SCT) survivors who participated in focus groups as part of the refinement of the eHealth-based mindfulness-based music therapy (eMBMT) intervention.

Demographics & medical characteristics	Total sample (N=11)	Focus group 1 (n=6)	Focus group 2 (n=5)
Age (years), mean (SD)	43.6 (17.8)	42.1 (15.9)	45.9 (18.4)
Gender, n (%)			
Female	6 (54.5)	4 (66.7)	2 (40)
Male	5 (45.5)	2 (33.3)	3 (60)
Race, n (%)			
White or Caucasian (including White Hispanic)	10 (91)	5 (83.3)	5 (100)
Black, African American, or Afro-Caribbean Black	1 (9)	1 (16.7)	0 (0)
Ethnicity, n (%)			
Hispanic	8 (72.7)	4 (66.7)	4 (80)
Non-Hispanic	3 (27.3)	2 (33.3)	1 (20)
Employment status, n (%)			
Full-time	5 (45.5)	3 (50)	2 (40)
Part-time	3 (27.3)	1 (16.7)	2 (40)
Unemployed with disability	2 (18.2)	2 (33.3)	0 (0)
Unemployed without disability	1 (9)	0 (0)	1 (20)
Education, n (%)			
High school or GED[a]	2 (18.2)	2 (33.3)	0 (0)
Some college	1 (9)	1 (16.7)	0 (0)
Graduated from a 2-year college	1 (9)	1 (16.7)	0 (0)
Graduated from a 4-year college	4 (36.5)	1(16.7)	3 (60)
Master's degree	3 (27.3)	1 (16.7)	2 (40)
Cancer type, n (%)			
Acute myeloid leukemia	11 (100)	6 (100)	5 (100)
Treatment type, n (%)			
Allogeneic stem cell transplant	11(100)	6 (100)	5 (100)
Time since diagnosis (months), n	26	26.7	25.3

[a]General Education Development diploma.

Focus Group Findings

Focus groups were conducted between April and May 2023. Qualitative analyses identified 4 themes.

The Initial Shock of a Cancer Diagnosis and Distress Over Unfavorable Outcomes

Most participants described the diagnosis of cancer as a profoundly shocking experience because they were diagnosed when they were feeling well and without any noticeable symptoms ("I didn't know I was sick. I had regular blood work done April of 2021 and everything was normal. And then June 1st I had a lump like a lymph node in my neck that was a little enlarged, but I wasn't feeling sick."). In addition, some participants hesitated to seek care because they were worried about exposure to COVID-19, resulting in delayed diagnosis and treatment ("I had previous history of asthma. So, my concern was: I go to the hospital, I'm gonna die… it's gonna hit people with respiratory problems, so I delayed it."). Once diagnosed, they grappled with uncertainty, the fear of missing out on meaningful activities, and the pressure to make tough treatment decisions (eg, fertility-related choices). After not responding to the initial treatment and given the option to receive an allo-SCT, participants faced frustration and anxiety related to finding a stem cell match, coping with chemotherapy's

XSL•FO
RenderX

physical side effects, and age-related issues (eg, hearing and vision loss, reduced mobility).

Challenges Associated With Hospitalization, Loss of Normalcy, and Adjusting to Life After Treatment

During hospitalization, participants experienced psychological distress due to prolonged social isolation and lack of personal interactions with their family and friends (eg, "...So, my isolation turned into 60, 70, 90 days, so that made it even worse"). They had feelings of loneliness, anxiety, boredom, and frustration surrounding missing major life events such as graduation, prom activities, or their children's events (eg, sport or school events). The COVID-19 pandemic restrictions negatively impacted the process of connecting patients to support services. Most participants reported a lack of access to support services during their hospital stay, which exacerbated their feeling of loneliness (eg, "I couldn't see anyone because it was COVID. So, for 2, 3 months there was nothing. Nobody offered me anything"). Following the conditioning phase and the stem cell infusion, participants experienced multiple physical side effects including weight loss, hair loss, nausea, vomiting, diarrhea, mouth sores, and fatigue. These symptoms negatively impacted their daily activity levels during their hospital stay and after being discharged. They also indicated several psychosocial side effects including, fear of recurrence, anxiety, depression, feeling very different (eg, "I'm not even close to who I was before this happened to me"), as well as a substantial impact on body image and its associated stigma (eg, "I don't want to be called the cancer kid at school"). They also noted that their families and friends were impacted by their disease (eg, "Another thing that I think it's forgotten is about the children. Most of us, we have children, and they have seen us in the darkest moments of our lives… and you think that they wouldn't remember, but they remember").

Coping Strategies for Living With a Hematologic Cancer

Participants developed various coping strategies for managing their disease. Participants vocalized several perceived useful strategies, such as (1) seeking social support from family, friends, health care staff, and other patients; (2) denial (eg, "I don't have cancer, I'm not going to admit it"); (3) spirituality and faith; (4) positive thinking and having an optimistic outlook; (5) goal setting; and (6) staying active and busy (eg, exercising, journaling, or watching television). In contrast, receiving adult coloring books and health communication materials was perceived as unhelpful or burdensome because they experienced changes in their ability to concentrate or memory lapses (eg, "We don't read the green pamphlet because it's too much, too overwhelming, and you don't have the time or the mind to think about that, you're just literally surviving"). Support services were generally limited during their hospital stay due to COVID-19. However, when patients received supportive services, they valued the interaction with the therapist, provider, or both during those difficult times (eg, art therapy, music therapy, massage therapy, and physical therapy). In addition, although they mentioned that telehealth and videoconferencing were very convenient for maintaining continuity of care, they also experienced "screen fatigue." Participants also noted the lack of follow-up after being discharged and that they had to request support services rather than being automatically referred to them. A participant mentioned that "there's a waiting period of a year to see a sexologist in the gynecology department."

Usefulness of Mindfulness and Music Therapy to Cope With Cancer

Participants perceived mindfulness as useful for staying positive, strengthening their spirituality, and coping with uncertainty. Although participants were open to practicing mindfulness through online means, they preferred in-person sessions and the ability to share their experiences with other patients. Participants perceived music therapy as helpful to change their mood, and bring joy, confidence, and faith. In contrast, a participant described herself as "not a music person" and that she "could go in life without music and would be fine." Most participants were open to participating in online music therapy sessions; however, they underscored the need to account for patients' physical and psychological symptoms, which could potentially impact their participation.

Feedback on MBMT Prototype

Most participants agreed with the proposed number of sessions (n=8) and duration of the intervention (approximately 60 minutes/session). However, participants recommended implementing more sessions pretransplant to better help them be emotionally and psychologically prepared for the allo-SCT. They also suggested more sessions postdischarge to help them reintegrate into their day-to-day activities. Participants agreed with providing blood samples, but also suggested streamlining the process and obtaining samples along with already scheduled blood work. Participants described the eMBMT prototype within the SmartManage platform as "easy to read," "easy to navigate," "very straightforward," and "very appealing." Participants endorsed the order in which the mindfulness attitudes were presented (eg, "there is definitely a pattern, I think…There's sort of a journey to it"). They believed that the in-person interaction with the music therapist would be very helpful. Participants recommendations included (1) incorporating diverse music genres(eg, "disco when you need to move" or "hip hop when you need to walk"), (2) allowing music selection based on patient's mood, (3) recording audio or video clips to reduce the amount of text, (4) adding a sleep sounds feature, (5) reflecting the patient population within the content (eg, images specific allo-SCT concerns), (6) involving caregivers, (7) exploring a partnership with meditation and mindfulness apps, and (8) providing the necessary tools (eg, iPads or headphones).

Usability and Field-Testing Session Findings

Overview

Usability and field-testing sessions were completed between June and July 2023. The 2 participants who completed the focus groups were unable to participate in usability and field-testing sessions. Participants (n=9) agreed that the eMBMT platform was useful (mean 6.47 SD 0.29), easy to use (mean 6.92, SD 0.60) and learn (mean 7.39, SD 0.75), and were satisfied with the platform (mean 6.16, SD 0.82).

Usability and field-testing sessions were completed between June and July 2023. Qualitative analysis identified 2 themes.

XSL•FO
RenderX

Participants' Opinions About the MBMT Intervention

Overall, participants considered the program "helpful," "easy to read," and easy to navigate. They indicated that if the program had been provided during their hospitalization, they would have used it. Participants also liked the "concept" of the program (eg, holistic approach, accessibility, mindfulness meditation, and music to calm and distract from treatment). Regarding the aesthetics, participants perceived that the program was pleasant and appreciated personalized features (eg, their name on the front page). They also liked the colors, text size, layout, and images; although, 2 participants noted that the pictures did not represent the patient population accurately (eg, no photos of cancer patients without hair or racially diverse). Furthermore, some participants mentioned that the amount of text was overwhelming, section titles were confusing because they were very similar, and the platform lacked navigation guidance. They also mentioned that the exercises were repetitive with a lack of music options and that they could not have done some of the proposed activities during their hospital stay because of the physical symptoms (eg, mouth sores or extreme fatigue). Participants also experienced a few technical issues, such as some of the features not working (ie, audio), poor audio quality (ie, "white noise"), and slow internet connection.

Suggestions for Improving the eMBMT Intervention

Participants provided specific suggestions related to the content, including (1) reducing the amount of text by recording video and audio clips, (2) stating a specific goal at the beginning of each session (eg, this session will help you reduce your anxiety), (3) adding more relatable content (eg, "missing being outside," "you may be feeling fatigued today. Here is what you can do") and more practical knowledge (eg, "How will this help me?" and "Is there any section that helps you control pain by breathing more?") rather than definitions (eg, "what is mindfulness?"), (4) connecting the mindfulness attitudes to patients' experiences, (5) adding more guidance throughout the modules, (6) embedding YouTube videos with an introduction slide rather than links, (7) adding different music genres and the ability to choose, and (8) including more representative images. Participants also suggested new features such as (1) implementing a checkmark to indicate sections that have been reviewed, (2) allowing them to interact with other participants (eg, enabling a chat feature and support groups), (3) a journaling or writing section and a feedback box, (4) reducing the amount of clicking and scrolling, and (5) creating a Spotify playlist with the music from the program.

Discussion

Principal Findings

Overview

This study refined and finalized, using focus groups and environmental and field testing, an eMBMT intervention aimed at reducing symptom burden and improving HRQoL in patients undergoing allo-SCT. The focus group findings of this study underscore the significant emotional and physical challenges faced by allo-SCT patients, including the distress of diagnosis, hospitalization, and posttransplant recovery. These challenges align with previous research indicating that allo-SCT patients frequently experience psychological distress, including anxiety, depression, and feelings of isolation, which can negatively impact recovery and long-term survival [5,6,9]. Focus group findings also highlight that allo-SCT survivors had an appreciation for the potential positive impacts of the mindfulness and music therapy intervention components. In addition, the iterative refinement process, informed by usability and field-testing feedback, led to key modifications including enhanced content representativeness, reduced text with audio options, personalized music selection, and adjustments to session timing to better support patients pre- and posttransplant.

The development and refinement of the novel eMBMT intervention represents a significant step towards providing integrative interventions addressing HRQoL, symptom burden, disease activity, and psychosocial and physiological adaptation of adults undergoing allo-SCT. Grounded in theory and well-established therapeutic techniques (ie, MT, MM, and MBSR), the eMBMT intervention provides a structured delivery of in-person and eHealth sessions facilitated by an MT-BC, and is designed to be accessible and user-friendly. An eHealth intervention may be especially well suited for this population as shifting treatment phases and unexpected periods of isolation make it difficult to coordinate in-person support services.

The iterative process of refining the eMBMT content through focus groups and usability and field-testing sessions was crucial for its refinement. Engaging allo-SCT survivors in the development process allowed for the collection of valuable data that directly informed the content, design, and functionality of the intervention. Overall, participants reported positive experiences with the platform and described the content as useful and most participants expressed that they would have participated in such an intervention had it been available at the time of their treatment.

The qualitative analysis identified key themes that highlight the complex emotional and physical challenges faced by allo-SCT patients. These include the initial shock of a cancer diagnosis, the distress associated with hospitalization and treatment, and the ongoing struggle to regain normalcy posttransplant. Participants also shared coping strategies, underscoring the importance of social support, positive thinking, and maintaining an active lifestyle. The perceived usefulness of mindfulness and music therapy further validates the inclusion of these components in the eMBMT intervention.

The use of rapid qualitative analysis enabled the timely synthesis of participant feedback, which was essential for adapting the intervention to the unique needs of patients undergoing allo-SCT. Rapid qualitative analysis has been increasingly recognized as a valid and efficient method for evaluating eHealth interventions and has been successfully used in previous studies to inform digital health tools [67,68]. While a traditional qualitative approach might have provided additional granularity, the chosen method ensured that the findings were actionable within the project's timeline. We acknowledge that rapid qualitative analysis may result in themes that closely reflect the structure of the interview prompts. To address this, the research team used iterative coding and thematic refinement to ensure

XSL•FO
RenderX

the themes captured meaningful nuances and extended beyond the initial prompts, thereby providing deeper insights into participants' experiences and needs.

The feedback from usability and field-testing sessions was overwhelmingly positive, with participants rating the platform highly in terms of usefulness, ease of use, and satisfaction. However, several areas for improvement were identified, which were invaluable for further refining the intervention to better meet the needs of patients. Below we discuss the key areas of improvement and subsequent modifications made to the eMBMT content and intervention.

Enhance the Representativeness of the Content

More representative images of the patient population were included within the eMBMT platform content (eg, an image of a cancer patient in hospital robe listening to music, patients without hair, more racially and ethnically diverse).

Reduce the Amount of Text and Reading

We reduced the amount of text and provided an option to listen to content. All intervention content was professionally audio-recorded and integrated into the respective section of the platform.

Provide Headphones

All participants will now receive a pair of personal headphones for ease of listening to content.

Add More Guidance Throughout the Modules

Additional instructions guiding participants through the modules were added.

Modify the eMBMT Session Delivery Timeline

The eMBMT session delivery timeline was modified such that more sessions are provided pretransplant to aid with the emotional and psychological preparation for the allo-SCT. In addition, more sessions are provided postdischarge to help with reintegration into their day-to-day activities. Table 2 presents the sessions, content topics, and the delivery timeline.

Table . Overview of the 8-session eHealth-based mindfulness-based music therapy (eMBMT) intervention for allogeneic stem cell transplantation (allo-SCT) patients, detailing session content, mindfulness principles, and time relative to the transplant.

Session	SCT[a] treatment phase	Mindfulness attitude	MT[b] component	Attitude carry over after eMBMT
1	Admission	MBMT[c] Orientation	Breathing to music	Focus on the present
2	Conditioning	Nonjudging	Music listening and journaling	Unpleasant and pleasant experiences and writing
3	Early post transplant	Patience	Focused music and environmental sound listening, simple body movements to music	Awareness and experience of patience
4	Early post transplant	Letting go	Music-assisted relaxation and scripts	Discussion and journaling of letting go
5	Engraftment to discharge	Acceptance	Music-assisted relaxation and body scan	Nonjudgmental awareness and acceptance of physical sensations, thoughts, and feelings
6	Postdischarge	Trust	Vocal improvisation	Trusting oneself
7	Postdischarge	Beginner's mind	Novel instrument sounds	Discussion and journaling of beginner's mind
8	Postdischarge	Nonstriving	Music improvisation	Focus on the process, letting go of preconceived goals with what "should be"

[a]SCT: stem cell transplantation.

[b]MT: music therapy.

[c]MBMT: mindfulness-based music therapy.

Include Different Genres of Music and Preferred Music

Participants communicated the desire for a more varied genre of music and the ability to choose their preferred music. Thus, more diverse samples of music were included as well as allowing participants to choose their preferred music to complete eMBMT activities.

Provide a Journaling Section

Modules now provide sections within each module where participants can journal, write their thoughts, or take notes.

Streamline Blood Sample Collection

The blood sample collection was streamlined such that at baseline, blood is collected only after the patient has received their port and when the patient is scheduled for a follow-up appointment post-discharge.

XSL•FO
RenderX

Address Technical Issues

All technical issues (eg, audio was not functioning properly) were resolved. Furthermore, audio that was deemed of poor quality (ie, “white noise”) was professionally re-recorded.

Feedback from the future pilot randomized controlled trial will shed light on potential further modifications to the eMBMT content and intervention as well as provide examples of how participants successfully implement mindfulness attitudes in their own lives throughout their allo-SCT treatment.

Limitations

This study has several limitations. First, the sample size was small and primarily consisted White and Hispanic individuals, which may limit the generalizability of the findings to more diverse populations undergoing allo-SCT. Second, while the focus groups and usability and field testing provided valuable qualitative insights, the results may be influenced by selection bias, as participants who opted into the study may have been more motivated or receptive to mindfulness and music therapy interventions. Third, the rapid qualitative analysis approach, while efficient, may have limited the depth of thematic exploration compared to more traditional qualitative methodologies. Finally, while usability testing identified technical issues and design preferences, future iterations should consider additional real-world implementation challenges, including potential barriers to patient engagement with digital health interventions.

Conclusion

The development and refinement of the eMBMT intervention through focus groups and usability and field testing has resulted in an eHealth intervention for use with allo-SCT patients. The study findings underscore the importance of patient involvement in intervention design, ensuring that content is both relevant and engaging. As digital interventions become increasingly prevalent in health care, eMBMT can serve as a model for leveraging technology to deliver complementary, evidence-based psychosocial support. Future research should explore its long-term impact, scalability, and integration into standard transplant care to optimize patient outcomes and reduce health care burdens. A future pilot randomized controlled trial will evaluate the feasibility, acceptability, and preliminary efficacy of the eMBMT on health-related quality of life, symptom burden, disease activity, and psychosocial and physiological adaptation [69].

Acknowledgments

This study was supported by National Cancer Institute (NCI) 5R61CA263335-02 (FJP and TL were the principal investigators). SEFP and BNE are funded by the Ruth L Kirschstein NRSA Institution Research Training Grant (T32; 5T32CA251064-03) in Cancer Training in Disparities and Equity (C-TIDE). PIM is supported by an NCI career development award (K01CA258955).

Data Availability

The datasets generated or analyzed during this study are not publicly available due to them containing information that could compromise patient privacy but are available from the corresponding author on reasonable request.

Authors' Contributions

SEFP, FJP, TL, KK, TW, MA, MRL, EDW, TKS, MAK, and BNE contributed to the study’s conception and design. TL, MAK, BNE, SEFP, and AEB conducted the interviews. BNE, SEFP, and AEB conducted the data analysis. SEFP, BNE, PL, and AEB wrote the first draft of the manuscript. All authors read, provided suggestions, and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Cancer facts & figures 2024. American Cancer Society. 2024. URL: https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/2024-cancer-facts-figures.html [accessed 2025-03-26]
2. Zhang C. Blood cancer. National Foundation for Cancer Research. 2024. URL: https://www.nfcr.org/cancer-types/cancer-types-blood-cancer/ [accessed 2025-03-26]
3. Gyurkocza B, Rezvani A, Storb RF. Allogeneic hematopoietic cell transplantation: the state of the art. Expert Rev Hematol 2010 Jun;3(3):285-299. [doi: 10.1586/ehm.10.21] [Medline: 20871781]
4. Bevans MF, Mitchell SA, Marden S. The symptom experience in the first 100 days following allogeneic hematopoietic stem cell transplantation (HSCT). Support Care Cancer 2008 Nov;16(11):1243-1254. [doi: 10.1007/s00520-008-0420-6] [Medline: 18322708]
5. Mosher CE, Redd WH, Rini CM, Burkhalter JE, DuHamel KN. Physical, psychological, and social sequelae following hematopoietic stem cell transplantation: a review of the literature. Psychooncology 2009 Feb;18(2):113-127. [doi: 10.1002/pon.1399] [Medline: 18677717]
6. Chang G, Orav EJ, McNamara TK, Tong MY, Antin JH. Psychosocial function after hematopoietic stem cell transplantation. Psychosomatics 2005;46(1):34-40. [doi: 10.1176/appi.psy.46.1.34] [Medline: 15765819]

XSL•FO
RenderX

7. Malard F, Holler E, Sandmaier BM, Huang H, Mohty M. Acute graft-versus-host disease. Nat Rev Dis Primers 2023 Jun 8;9(1):27. [doi: 10.1038/s41572-023-00438-1] [Medline: 37291149]
8. Prieto JM, Atala J, Blanch J, et al. Patient-rated emotional and physical functioning among hematologic cancer patients during hospitalization for stem-cell transplantation. Bone Marrow Transplant 2005 Feb;35(3):307-314. [doi: 10.1038/sj.bmt.1704788] [Medline: 15580279]
9. Prieto JM, Atala J, Blanch J, et al. Role of depression as a predictor of mortality among cancer patients after stem-cell transplantation. JCO 2005 Sep 1;23(25):6063-6071. [doi: 10.1200/JCO.2005.05.751]
10. Park JE, Kim KI, Yoon SS, et al. Psychological distress as a negative survival factor for patients with hematologic malignancies who underwent allogeneic hematopoietic stem cell transplantation. Pharmacotherapy 2010 Dec;30(12):1239-1246. [doi: 10.1592/phco.30.12.1239] [Medline: 21114391]
11. Majhail NS, Omondi NA, Denzen E, Murphy EA, Rizzo JD. Access to hematopoietic cell transplantation in the United States. Biol Blood Marrow Transplant 2010 Aug;16(8):1070-1075. [doi: 10.1016/j.bbmt.2009.12.529] [Medline: 20036337]
12. Antoni MH, Dhabhar FS. The impact of psychosocial stress and stress management on immune responses in patients with cancer. Cancer 2019 May 1;125(9):1417-1431. [doi: 10.1002/cncr.31943] [Medline: 30768779]
13. Vogelzangs N, de Jonge P, Smit JH, Bahn S, Penninx BW. Cytokine production capacity in depression and anxiety. Transl Psychiatry 2016 May 31;6(5):e825. [doi: 10.1038/tp.2016.92] [Medline: 27244234]
14. Knight JM, Lyness JM, Sahler OJZ, Liesveld JL, Moynihan JA. Psychosocial factors and hematopoietic stem cell transplantation: potential biobehavioral pathways. Psychoneuroendocrinology 2013 Nov;38(11):2383-2393. [doi: 10.1016/j.psyneuen.2013.06.016] [Medline: 23845514]
15. Amonoo HL, Brown LA, Scheu CF, et al. Positive psychological experiences in allogeneic hematopoietic stem cell transplantation. Psychooncology 2019 Aug;28(8):1633-1639. [doi: 10.1002/pon.5128] [Medline: 31128072]
16. McGregor BA, Syrjala KL, Dolan ED, Langer SL, Redman M. The effect of pre-transplant distress on immune reconstitution among adult autologous hematopoietic cell transplantation patients. Brain Behav Immun 2013 Mar;30 Suppl(Suppl):S142-S148. [doi: 10.1016/j.bbi.2012.07.020] [Medline: 22910186]
17. Lee SJ, Loberiza FR, Rizzo JD, Soiffer RJ, Antin JH, Weeks JC. Optimistic expectations and survival after hematopoietic stem cell transplantation. Biol Blood Marrow Transplant 2003 Jun;9(6):389-396. [doi: 10.1016/s1083-8791(03)00103-4] [Medline: 12813447]
18. Duncan M, Moschopoulou E, Herrington E, et al. Review of systematic reviews of non-pharmacological interventions to improve quality of life in cancer survivors. BMJ Open 2017 Nov 28;7(11):e015860. [doi: 10.1136/bmjopen-2017-015860] [Medline: 29187408]
19. Cramer H, Lauche R, Klose P, Lange S, Langhorst J, Dobos GJ. Yoga for improving health-related quality of life, mental health and cancer-related symptoms in women diagnosed with breast cancer. Cochrane Database Syst Rev 2017 Jan 3;1(1):CD010802. [doi: 10.1002/14651858.CD010802.pub2] [Medline: 28045199]
20. Oberoi S, Yang J, Woodgate RL, et al. Association of mindfulness-based interventions with anxiety severity in adults with cancer: a systematic review and meta-analysis. JAMA Netw Open 2020 Aug 3;3(8):e2012598. [doi: 10.1001/jamanetworkopen.2020.12598] [Medline: 32766801]
21. Antoni MH, Moreno PI, Penedo FJ. Stress management interventions to facilitate psychological and physiological adaptation and optimal health outcomes in cancer patients and survivors. Annu Rev Psychol 2023 Jan 18;74:423-455. [doi: 10.1146/annurev-psych-030122-124119] [Medline: 35961041]
22. Crane RS, Brewer J, Feldman C, et al. What defines mindfulness-based programs? The warp and the weft. Psychol Med 2017 Apr;47(6):990-999. [doi: 10.1017/S0033291716003317] [Medline: 28031068]
23. Elkins G, Fisher W, Johnson A. Mind-body therapies in integrative oncology. Curr Treat Options Oncol 2010 Dec;11(3-4):128-140. [doi: 10.1007/s11864-010-0129-x] [Medline: 21116746]
24. Johns SA, Brown LF, Beck-Coon K, Monahan PO, Tong Y, Kroenke K. Randomized controlled pilot study of mindfulness-based stress reduction for persistently fatigued cancer survivors. Psychooncology 2015 Aug;24(8):885-893. [doi: 10.1002/pon.3648] [Medline: 25132206]
25. Lin LY, Lin LH, Tzeng GL, et al. Effects of mindfulness-based therapy for cancer patients: a systematic review and meta-analysis. J Clin Psychol Med Settings 2022 Jun;29(2):432-445. [doi: 10.1007/s10880-022-09862-z] [Medline: 35249176]
26. Liu H, Gao X, Hou Y. Effects of mindfulness-based stress reduction combined with music therapy on pain, anxiety, and sleep quality in patients with osteosarcoma. Braz J Psychiatry 2019;41(6):540-545. [doi: 10.1590/1516-4446-2018-0346] [Medline: 31116262]
27. Zhang Q, Zhao H, Zheng Y. Effectiveness of mindfulness-based stress reduction (MBSR) on symptom variables and health-related quality of life in breast cancer patients-a systematic review and meta-analysis. Support Care Cancer 2019 Mar;27(3):771-781. [doi: 10.1007/s00520-018-4570-x] [Medline: 30488223]
28. Cash E, Salmon P, Weissbecker I, et al. Mindfulness meditation alleviates fibromyalgia symptoms in women: results of a randomized clinical trial. Ann Behav Med 2015 Jun;49(3):319-330. [doi: 10.1007/s12160-014-9665-0] [Medline: 25425224]
29. Hilton L, Hempel S, Ewing BA, et al. Mindfulness meditation for chronic pain: systematic review and meta-analysis. Ann Behav Med 2017 Apr;51(2):199-213. [doi: 10.1007/s12160-016-9844-2] [Medline: 27658913]

XSL•FO
RenderX

30. Vieten C, Laraia BA, Kristeller J, et al. The mindful moms training: development of a mindfulness-based intervention to reduce stress and overeating during pregnancy. BMC Pregnancy Childbirth 2018 Jun 1;18(1):201. [doi: 10.1186/s12884-018-1757-6] [Medline: 29859038]
31. Victorson D, Kentor M, Maletich C, et al. Mindfulness meditation to promote wellness and manage chronic disease: a systematic review and meta-analysis of mindfulness-based randomized controlled trials relevant to lifestyle medicine. Am J Lifestyle Med 2014;9(3):185-211. [doi: 10.1177/1559827614537789]
32. van der Steen JT, van Soest-Poortvliet MC, van der Wouden JC, Bruinsma MS, Scholten RJ, Vink AC. Music-based therapeutic interventions for people with dementia. Cochrane Database Syst Rev 2017 May 2;5(5):CD003477. [doi: 10.1002/14651858.CD003477.pub3] [Medline: 28462986]
33. Lemieux AF, Fisher JD, Pratto F. A music-based HIV prevention intervention for urban adolescents. Health Psychol 2008 May;27(3):349-357. [doi: 10.1037/0278-6133.27.3.349] [Medline: 18624599]
34. Wachi M, Koyama M, Utsuyama M, Bittman BB, Kitagawa M, Hirokawa K. Recreational music-making modulates natural killer cell activity, cytokines, and mood states in corporate employees. Med Sci Monit 2007 Feb;13(2):CR57-CR70. [Medline: 17261984]
35. Archie P, Bruera E, Cohen L. Music-based interventions in palliative cancer care: a review of quantitative studies and neurobiological literature. Support Care Cancer 2013 Sep;21(9):2609-2624. [doi: 10.1007/s00520-013-1841-4] [Medline: 23715815]
36. Fancourt D, Ockelford A, Belai A. The psychoneuroimmunological effects of music: a systematic review and a new model. Brain Behav Immun 2014 Feb;36(15-26):15-26. [doi: 10.1016/j.bbi.2013.10.014] [Medline: 24157429]
37. Bittman B, Poornima I, Smith MA, Heidel RE. Gospel music: a catalyst for retention, engagement, and positive health outcomes for African Americans in a cardiovascular prevention and treatment program. Adv Mind Body Med 2020;34(1):8-16. [Medline: 32277749]
38. Magee WL, Clark I, Tamplin J, Bradt J. Music interventions for acquired brain injury. Cochrane Database Syst Rev 2017 Jan 20;1(1):CD006787. [doi: 10.1002/14651858.CD006787.pub3] [Medline: 28103638]
39. Lee JH. The effects of music on pain: a meta-analysis. J Music Ther 2016;53(4):430-477. [doi: 10.1093/jmt/thw012] [Medline: 27760797]
40. Dileo C. Effects of music and music therapy on medical patients: a meta-analysis of the research and implications for the future. J Soc Integr Oncol 2006;4(2):67-70. [doi: 10.2310/7200.2006.002] [Medline: 19442338]
41. Leggieri M, Thaut MH, Fornazzari L, et al. Music intervention approaches for Alzheimer's disease: a review of the literature. Front Neurosci 2019;13:132. [doi: 10.3389/fnins.2019.00132] [Medline: 30930728]
42. Haase JE, Robb SL, Burns DS, et al. Adolescent/young adult perspectives of a therapeutic music video intervention to improve resilience during hematopoietic stem cell transplant for cancer. J Music Ther 2020 Feb 25;57(1):3-33. [doi: 10.1093/jmt/thz014] [Medline: 31802124]
43. Robb SL, Burns DS, Stegenga KA, et al. Randomized clinical trial of therapeutic music video intervention for resilience outcomes in adolescents/young adults undergoing hematopoietic stem cell transplant: a report from the Children's Oncology Group. Cancer 2014 Mar 15;120(6):909-917. [doi: 10.1002/cncr.28355] [Medline: 24469862]
44. Russ KA, Holochwost SJ, Perkins SM, et al. Cortisol as an acute stress biomarker in young hematopoietic cell transplant patients/caregivers: active music engagement protocol. J Altern Complement Med 2020 May;26(5):424-434. [doi: 10.1089/acm.2019.0413] [Medline: 32073877]
45. Bradt J, Dileo C, Myers-Coffman K, Biondo J. Music interventions for improving psychological and physical outcomes in people with cancer. Cochrane Database Syst Rev 2021 Oct 12;10(10):CD006911. [doi: 10.1002/14651858.CD006911.pub4] [Medline: 34637527]
46. Gramaglia C, Gambaro E, Vecchi C, et al. Outcomes of music therapy interventions in cancer patients-a review of the literature. Crit Rev Oncol Hematol 2019 Jun;138(241-254):241-254. [doi: 10.1016/j.critrevonc.2019.04.004] [Medline: 31121392]
47. Dóro CA, Neto JZ, Cunha R, Dóro MP. Music therapy improves the mood of patients undergoing hematopoietic stem cells transplantation (controlled randomized study). Support Care Cancer 2017 Mar;25(3):1013-1018. [doi: 10.1007/s00520-016-3529-z] [Medline: 27942868]
48. Bates D, Bolwell B, Majhail NS, et al. Music therapy for symptom management after autologous stem cell transplantation: results from a randomized study. Biol Blood Marrow Transplant 2017 Sep;23(9):1567-1572. [doi: 10.1016/j.bbmt.2017.05.015] [Medline: 28533058]
49. Cassileth BR, Vickers AJ, Magill LA. Music therapy for mood disturbance during hospitalization for autologous stem cell transplantation: a randomized controlled trial. Cancer 2003 Dec 15;98(12):2723-2729. [doi: 10.1002/cncr.11842] [Medline: 14669295]
50. Stanczyk MM. Music therapy in supportive cancer care. Rep Pract Oncol Radiother 2011 Jun 8;16(5):170-172. [doi: 10.1016/j.rpor.2011.04.005] [Medline: 24376975]
51. Gutgsell KJ, Schluchter M, Margevicius S, et al. Music therapy reduces pain in palliative care patients: a randomized controlled trial. J Pain Symptom Manage 2013 May;45(5):822-831. [doi: 10.1016/j.jpainsymman.2012.05.008] [Medline: 23017609]

XSL•FO
RenderX

52. Köhler F, Martin ZS, Hertrampf RS, et al. Music therapy in the psychosocial treatment of adult cancer patients: a systematic review and meta-analysis. Front Psychol 2020;11(2095):651. [doi: 10.3389/fpsyg.2020.00651] [Medline: 32373019]
53. Harper FWK, Heath AS, Moore TF, Kim S, Heath EI. Using music as a tool for distress reduction during cancer chemotherapy treatment. JCO Oncol Pract 2023 Dec;19(12):1133-1142. [doi: 10.1200/OP.22.00814] [Medline: 37433094]
54. Tsai HF, Chen YR, Chung MH, et al. Effectiveness of music intervention in ameliorating cancer patients' anxiety, depression, pain, and fatigue: a meta-analysis. Cancer Nurs 2014;37(6):E35-E50. [doi: 10.1097/NCC.0000000000000116] [Medline: 24662723]
55. Lee JH. Erratum to: the effects of music on pain: a meta-analysis. J Music Ther 2021 Aug 24;58(3):34286339. [doi: 10.1093/jmt/thab012] [Medline: 34286339]
56. Penedo FJ, Oswald LB, Kronenfeld JP, Garcia SF, Cella D, Yanez B. The increasing value of eHealth in the delivery of patient-centred cancer care. Lancet Oncol 2020 May;21(5):e240-e251. [doi: 10.1016/S1470-2045(20)30021-8] [Medline: 32359500]
57. Kabat-Zinn J. Full Catastrophe Living: Using the Wisdom of Your Body and Mind to Face Stress, Pain, and Illness, 15th edition: Bantam Dell; 2005.
58. Lesiuk T. The effect of mindfulness-based music therapy on attention and mood in women receiving adjuvant chemotherapy for breast cancer: a pilot study. Oncol Nurs Forum 2015 May;42(3):276-282. [doi: 10.1188/15.ONF.276-282] [Medline: 25901379]
59. Lesiuk T. The development of a mindfulness-based music therapy (MBMT) program for women receiving adjuvant chemotherapy for breast cancer. Healthcare (Basel) 2016 Aug 9;4(3):27517966. [doi: 10.3390/healthcare4030053] [Medline: 27517966]
60. Puccinelli M, Seay J, Otto A, et al. An adapted cognitive behavioral stress and self-management intervention for sexual minority men living with HIV and cancer using the SmartManage ehealth platform: protocol and study design. JMIR Res Protoc 2022 Jul 18;11(7):e37822. [doi: 10.2196/37822] [Medline: 35849435]
61. BrightOutcome. URL: https://www.brightoutcome.com/index.php [accessed 2023-11-09]
62. Pfeiffer E. A short portable mental status questionnaire for the assessment of organic brain deficit in elderly patients. J Am Geriatr Soc 1975 Oct;23(10):433-441. [doi: 10.1111/j.1532-5415.1975.tb00927.x] [Medline: 1159263]
63. Krueger RA, Casey MA. Focus Groups: A Practical Guide for Applied Research, 5th edition: Sage Publications; 2014.
64. McQuarrie EF, Krueger RA. Focus groups: a practical guide for applied research. J Mark Res 1989 Aug;26(3):371. [doi: 10.2307/3172912]
65. Lund AM. Measuring usability with the USE questionnaire. Usability Interface 2001;8(2):3-6 [FREE Full text]
66. Fonteyn ME, Kuipers B, Grobe SJ. A description of think aloud method and protocol analysis. Qual Health Res 1993 Nov;3(4):430-441. [doi: 10.1177/104973239300300403]
67. Watkins DC. Rapid and rigorous qualitative data analysis: the "RADaR" technique for applied research. Int J Qual Methods 2017;16:1. [doi: 10.1177/1609406917712131]
68. St George SM, Harkness AR, Rodriguez-Diaz CE, Weinstein ER, Pavia V, Hamilton AB. Applying rapid qualitative analysis for health equity: lessons learned using "EARS" with Latino communities. Int J Qual Methods 2023;22:38463016. [doi: 10.1177/16094069231164938] [Medline: 38463016]
69. Fleszar-Pavlovic SE, Esquives BN, Brito AE, et al. eHealth mindfulness-based music therapy for patients undergoing allogeneic hematopoietic stem cell transplantation: A pilot randomized controlled trial protocol. Contemp Clin Trials 2024 Jul;142(107577). [doi: 10.1016/j.cct.2024.107577] [Medline: 38763308]

Abbreviations

allo-SCT: allogeneic stem cell transplantation
eMBMT: eHealth-based mindfulness-based music therapy
GVHD: graft-versus-host disease
HIPAA: Health Insurance Portability and Accountability Act
HRQoL: health-related quality of life
ICD: *International Classification of Diseases*
MBMT: mindfulness-based music therapy
MBSR: mindfulness-based stress reduction
MM: mindful meditation
MT: music therapy
MT-BC: board-certified music therapist
RQA: rapid qualitative analysis
USE: Usefulness, Satisfaction, and Ease of Use

XSL•FO
RenderX

Edited by A Mavragani; submitted 08.08.24; peer-reviewed by MT Knobf, R Knoerl; revised version received 12.02.25; accepted 13.02.25; published 11.04.25.

Please cite as:
Fleszar-Pavlovic SE, Noriega Esquives B, Lovan P, Brito AE, Sia AM, Kauffman MA, Lopes M, Moreno PI, Koru-Sengul T, Gong R, Wang T, Wieder ED, Rueda-Lara M, Antoni M, Komanduri K, Lesiuk T, Penedo FJ
Development of an eHealth Mindfulness-Based Music Therapy Intervention for Adults Undergoing Allogeneic Hematopoietic Stem Cell Transplantation: Qualitative Study
JMIR Form Res 2025;9:e65188
URL: https://formative.jmir.org/2025/1/e65188
doi:10.2196/65188

© Sara E Fleszar-Pavlovic, Blanca Noriega Esquives, Padideh Lovan, Arianna E Brito, Ann Marie Sia, Mary Adelyn Kauffman, Maria Lopes, Patricia I Moreno, Tulay Koru-Sengul, Rui Gong, Trent Wang, Eric D Wieder, Maria Rueda-Lara, Michael Antoni, Krishna Komanduri, Teresa Lesiuk, Frank J Penedo. Originally published in JMIR Formative Research (https://formative.jmir.org), 11.4.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Evaluating User Experience and Satisfaction in a Concussion Rehabilitation App: Usability Study

Michael G Hutchison[1,2,3], PhD; Alex P Di Battista[1,4], PhD; Kyla L Pyndiura[1,3], MSc

[1]Faculty of Kinesiology and Physical Education, University of Toronto, 100 Devonshire Pl, Toronto, ON, Canada

[2]David L. MacIntosh Sport Medicine Clinic, Faculty of Kinesiology & Physical Education, University of Toronto, Toronto, ON, Canada

[3]Centre for Sport-Related Concussion Research, Innovation, and Knowledge, University of Toronto, Toronto, ON, Canada

[4]Defence Research and Development Canada, Toronto Research Centre, Toronto, ON, Canada

Corresponding Author:
Michael G Hutchison, PhD
Faculty of Kinesiology and Physical Education, University of Toronto, 100 Devonshire Pl, Toronto, ON, Canada

Abstract

Background: Evidence-based guidelines support the use of structured exercise to facilitate concussion recovery. Despite the growing number of mobile health (mHealth) apps aimed at managing concussions, few focus on delivering exercise rehabilitation protocols. Therefore, a mobile app was developed to provide personalized rehabilitation programs based on evidence-based exercise principles designed to cater to individuals recovering from concussions.

Objective: This study aimed to evaluate the usability and user experience of a mobile app designed to deliver an evidence-based rehabilitation program to individuals recovering from concussions.

Methods: A two-week prospective single-arm pilot study was conducted among adults with a physician-diagnosed concussion. Participants engaged in home-based rehabilitation exercises through a mobile app. Usability was assessed using a combination of the mHealth App Usability Questionnaire (MAUQ) and five custom questions evaluating confidence in recommendations, exercise flow, clarity of voice commands, and usability of the exercise report feature. Following the two-week period, participants rated each question on a 7-point Likert scale ranging from “strongly agree” to “strongly disagree”.

Results: Twenty-six participants consented and were enrolled in the study, with 23 participants (82%) completing all study components at the end of two weeks. The majority of participants were women, aged 26-38 years, and on average, approximately at three months postconcussion. Responses to both the MAUQ and custom questions were overwhelmingly positive. Overall, seven MAUQ questions received 100% positive responses, with no single question scoring below 83% positive responses. In the “ease of use and satisfaction” category, 100% of users responded positively to questions on ease of learning, usability, interface likeness, and comfort in social settings, while 83%-96% (19-22) of users responded positively to the remaining four questions. In the “system information arrangement” category, 100% (N=23) of users rated screen navigation, function usability, and health care service acceptability positively, with 87%-96% of users approving action acknowledgment, error recovery, and expected functions. Under the “usefulness” category, 96% (n=22) of users found the app beneficial for health and well-being, and 91% (n=21) users felt it effectively managed their health. For the five custom questions, 100% (N=23) users responded positively to voice command clarity, exercise awareness, ease of following exercises, and report understandability, with a single unfavorable response noted for confidence in app recommendations.

Conclusions: The findings of this study indicate that the mobile app is a user-friendly platform for delivering evidence-based exercise rehabilitation to individuals recovering from concussions. Positive user feedback, particularly in the areas of recommendation confidence, ease of exercise flow, and clarity of voice commands, highlights the app’s potential to support concussion recovery. Future iterations of the app will aim to improve time efficiency and streamline error recovery processes to further enhance the user experience.

(JMIR Form Res 2025;9:e67275) doi:10.2196/67275

KEYWORDS

mild traumatic brain injury; recovery; mHealth; app; digital health; smartphone; eHealth; digital; technology; usability; concussion rehabilitation; brain injury; rehabilitation protocols; evidence-based exercise; single-arm pilot study; home-based rehabilitation; user-friendly; questionnaire; telehealth; telemedicine

XSL•FO
RenderX

Introduction

Concussion is a common form of traumatic brain injury (TBI) that can occur in a variety of settings such as sports, motor vehicle accidents, workplaces, and daily life. Symptoms range from physical and cognitive disruptions to emotional and sleep disturbances [1], underscoring the highly individualized nature of the injury. Following an initial 24‐ to 48-hour period of cognitive and physical rest, research has demonstrated the benefits of aerobic exercise in enhancing concussion recovery. This approach not only alleviates symptom severity but can also expedite the recovery process [2-4].

The recommendation of aerobic exercise as an intervention for concussion often involves establishing a personalized heart rate (HR) threshold set by trained professionals [5,6]. More recently, the determination of exercise intensity has evolved to include the use of age-predicted maximum HR formulas [3], offering a more accessible and individualized approach to concussion management. However, to enhance the impact of these advancements, it is crucial to integrate exercise protocols into clinical practices beyond sports and exercise medicine. This can ensure that a wider patient population benefits from the most effective, evidence-based strategies for concussion management, thus achieving a more comprehensive and equitable standard of care.

Mobile health (mHealth) has emerged as a widely accessible medium for individuals to obtain medical care. This technology is particularly beneficial for individuals in remote areas or those who face challenges in accessing health care facilities. Consequently, there has been a surge in the development of health care–focused mobile apps in recent years, demonstrating significant benefits. Notably, mHealth interventions in mental health have shown promise in reducing symptoms of depression and anxiety across diverse population groups [7,8]. Additionally, mHealth apps have been effectively used in various health care domains, including cardiovascular disease prevention and diabetes management, showcasing their versatility and impact on modern health care [9-12].

The application of mHealth is increasingly recognized for its potential benefits in managing TBI. Various mobile apps aimed at providing education about TBI, symptom tracking, daily reminders for patients with TBI, and aiding in the screening or assessment of TBI are now available [13-15]. Despite these advancements, a notable gap exists in apps specifically tailored for the management of TBIs [13]. Moreover, many of these apps have yet to achieve widespread user adoption; a significant concern is that many available apps lack evidence-based interventions, making it difficult for health care providers to recommend them and for patients to use them with confidence [13]. In concussion care, the primary focus of mobile apps has been on symptom reporting, monitoring activity levels, and concussion recognition targeted at parents, coaches, athletic trainers, and athletes [16-18]. However, current apps do not focus on interventions involving exercise or physical activity, indicating an area for further development in mHealth solutions for TBI and concussion management.

Hutchison et al [19] conducted a study evaluating the feasibility of a mobile app designed to deliver resistance exercises using minimal equipment for individuals recovering from concussions. The app provides an aerobic stimulus by guiding users to achieve a percentage of their age-predicted maximal HR through tailored exercise routines. This innovative tool includes instructional videos to assist users in their rehabilitation process. Initial results demonstrated that most participants were able to reach the target HR zones by following the app-based exercise videos [19]. Building on these preliminary findings, which focused on the app's ability to elicit the desired physiological responses [19], the present study aims to expand on this research by assessing the app's overall usability. Specifically, we sought to identify potential areas for improvement, gauge user satisfaction, and evaluate the app's effectiveness in supporting concussion recovery.

Methods

Study Design

The usability study consisted of a two-week, community-oriented, prospective single-arm observational pilot study involving adults aged 18 years and older with a physician-diagnosed concussion. Participants were engaged in an active, home-based rehabilitation program accessed through the mobile app. Individuals were recruited through word-of-mouth, flyers, and social media platforms (eg, X, formerly known as Twitter, Facebook) and were provided detailed information regarding study requirements and procedures. Participant enrollment was conducted remotely via Zoom, and individuals were screened by a research assistant to confirm inclusion criteria consisting of a concussion diagnosis by a physician, aged at least 18 years, and access to an iPhone. Furthermore, individuals with coexisting conditions such as musculoskeletal or soft-tissue injuries, vestibular disorders, or neurological conditions that could impede physical activity, were excluded from the study. Twenty-six individuals consented and enrolled in the study. Upon enrollment, participants were mailed an Apple Watch Series 6 to use with their iPhones and access to the Rhea Health Inc. mobile app. Ultimately, 23 participants completed the usability questionnaire after using the mobile app for two weeks.

Ethical Considerations

This study received ethical approval from the Health Sciences Research Ethics Board at the University of Toronto (protocol #39143). To ensure comprehensive understanding and informed decision-making, prospective participants were given an information sheet by a member of the research team. This sheet detailed the study's eligibility criteria, purpose, methodologies, and any potential risks involved. Before participants gave their consent, their eligibility was confirmed via email. A Zoom videoconference was then conducted to clarify any queries regarding the study and reinforce the voluntary nature of participation and the freedom to withdraw at any point without adverse consequences. Informed consent was secured through an electronic process in which participants reviewed and signed a digital document that outlined the study's purpose, procedures, and potential risks. This consent form was then submitted using

XSL•FO
RenderX

the REDCap (Research Electronic Data Capture) platform, hosted by the University of Toronto.

Data collected via the mobile app and Apple Watch were securely stored on the Amazon Web Services' DynamoDB database, located in Montreal, Quebec, Canada. This storage solution employed Amazon Web Services' proprietary encryption and restricted access system to ensure data security. Although participants were required to use their emails and passwords for accessing the mobile app and syncing data from their Apple Watch to Apple Health, personally identifiable information was stored in a separate database table and was not required for data analysis.

Upon the study's completion, participants were compensated CAD $50 (equivalent to approximately US $37.16) for their participation and were entered into a lottery for a chance to win an Apple Watch Series 6. This gesture not only acknowledged their contribution but also encouraged engagement and retention in the study.

The Mobile App

The rehabilitation protocol in this study was delivered through the Rhea mobile app (Rhea Health Inc.) (Figure 1), a rehabilitation platform that devises personalized programs based on an individual's symptoms and time elapsed since injury. The app integrates symptom data, including 22 items from the Sports Concussion Assessment Tool (SCAT-5) [20], along with seven additional questions related to vision and balance. Participants assessed their symptom severity using a 4-point Likert Scale, ranging from 0 (none) to 3 (severe). Following a staged approach endorsed by consensus statements on concussion management, the Rhea rehabilitation protocol employed a six-stage model that progressively increased in intensity and complexity. These stages extend from Initial Rest (Stage 1) to Return (Stage 6). Participants commenced their rehabilitation at Stage 2, completing three sessions at this level before progressing. One of the Rhea app's unique features is its dynamic adjustment of the rehabilitation plan that is based on participant feedback, which categorizes symptomology as "same," "better," or "worse." This feedback-driven mechanism ensures that the participants either stay at their current stage, advance to a more challenging stage, or revert to a less intensive stage, as required. The primary objective of this adaptive approach is to fine-tune the intensity and complexity of exercises to avoid exacerbating symptoms, thus promoting a safe and effective journey through the recovery stages. In addition to the six core stages, Rhea's program provided participants with exercises specifically tailored to neck, vision, and balance rehabilitation. These exercises are designed to address each participant's unique symptom profile, further personalizing the recovery experience. This approach not only ensures adherence to the rehabilitation protocol but also maximizes the potential for effective recovery tailored to individual needs. To monitor adverse events, participants were instructed to inform a member of the research team of any adverse events (eg, injuries, medical conditions) that were experienced while using the mobile app.

Figure 1. Screenshots of the Rhea mobile app from left to right display (A) the home page of a participant's plan, (B) a symptom checklist during onboarding that assists with exercise prescription, (C) an exercise preview screen providing description and cues for correct form, and (D) an exercise session screen with a timer and details of the next exercise for the user.

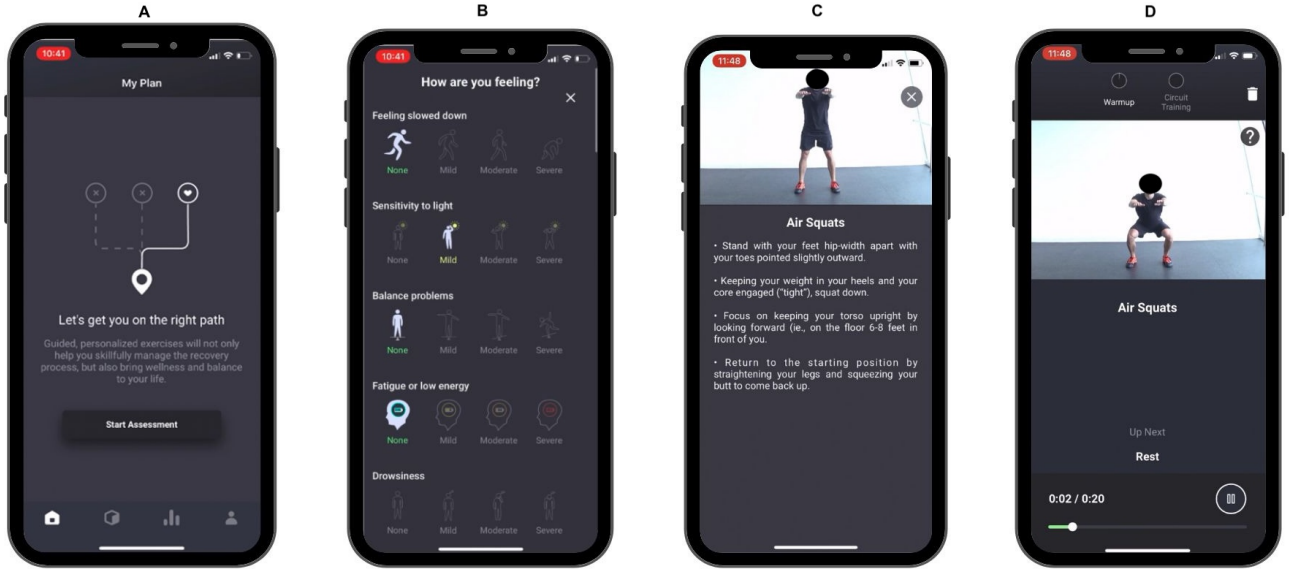

Usability Questionnaire

The mHealth App Usability Questionnaire (MAUQ) [21] contains 21 questions rated on a 7-point Likert scale, ranging from strongly agree to strongly disagree, under three categories: ease of use and satisfaction, system information arrangement, and usefulness. Studies have identified the MAUQ to be a reliable and valid tool for assessing mHealth app usability [21]. In this study, we used 16 questions from the MAUQ and omitted the 5 questions from the usefulness category, as they assess aspects related to sharing information with health care providers and clinical utility. Due to the preliminary nature of the research and the use of a research-grade tool that lacked standards and approval for clinical use, we did not ask or expect participants to engage with health care professionals using the tool. Additionally, we included five custom questions to evaluate the app's critical features related to the delivery of exercise rehabilitation, as the app was based on evidence-based exercise

XSL•FO
RenderX

rehab principles. These custom questions were designed to assess user confidence in the app's recommendations, clarity of exercise instructions, the flow of the exercise content, and the usability of the app's voice commands. More specifically, given the central role of accurate exercise prescription in rehabilitation, it was important to gauge whether users trusted the app's recommendations; therefore the question, "I had confidence in the recommendations I received from the app" was included. Since effective rehabilitation relies on the clarity and ease of exercise guidance, we included two questions targeting the flow and clarity of the exercise sessions: "The flow of the exercise sessions on the app is easy to follow" and "I knew exactly which exercises I had to do during an exercise session." These questions allowed us to determine whether participants found the session progression intuitive and well-structured. We also added a question about voice commands: "The in-app voice commands are clear and understandable." This was a critical feature for users who might benefit from auditory guidance, particularly those experiencing cognitive fatigue or visual sensitivity due to their concussion. Clear voice commands would support a hands-free interaction and reduce the cognitive load associated with reading instructions on a screen. Lastly, we evaluated the app's reporting functions through the question, "The exercise report section is easy to follow and understand." A clear and accessible exercise report is vital for tracking progress and ensuring participants understand their rehabilitation. These custom questions were specifically tailored to capture the nuances of delivering exercise rehabilitation through a digital platform, ensuring that the app was not only usable but also effective in guiding users through their recovery process.

Data Analysis

The purpose of the current study was to summarize and report participant feedback following a two-week period of using the mobile app. Therefore, the Likert-scale responses from the MAUQ and additional customized app usability questions were summarized and reported as raw counts and percentages. For additional continuous variables, the median (IQR) were calculated.

Results

Participant Characteristics

Study participant characteristics can be seen in Table 1. Briefly, the sample comprised predominantly of female participants (n=18, 78%), aged 26 - 38 years. On average, the participants were nearly three months post-concussion, with 48% (n=11) sustaining injury during sports or physical activity.

XSL•FO
RenderX

Table . Demographic and clinical characteristics of participants, including age, sex, mechanism of injury, and time since concussion.

Characteristics	Participants (N=23)
Sex, n (%)	
Female	18 (78)
Male	5 (22)
Age, median (IQR)	28.4 (25.1-37.3)
Days from injury, median (IQR)	82 (25-461)
Mechanism of injury, n (%)	
Sports/PA[a]	11 (48)
MVC[b]	1 (4.3)
Work	1 (4.3)
Fall	2 (8.7)
Other	8 (35)
History of concussion, n (%)	13 (59)
High performance athlete, n (%)	5 (22)
Exercise program, n (%)	8 (35)
Physical activity (times/week)	
0	3 (13)
1	2 (8.7)
2	4 (17)
3	3 (13)
4	2 (8.7)
5	4 (17)
6	3 (13)
≥7	2 (8.7)

[a]PA: physical activity.
[b]MVC: Motor vehicle collision.

Concussion Characteristics

Participants entered the study with an average of 14 symptoms and a symptom severity score of 29. On average, they reported feeling at approximately 70% of their usual health, and 43% (n=10) of the participants stated that they were getting better following their injury at first, but later experienced symptom worsening. Please see Table 2 for the complete symptom profiles of the participants.

Table . Symptom profile at onboarding, including symptom severity, perceived normalcy, and symptom progression trends.

Characteristics	Participants (N=23)
Symptom number, median (IQR)	14 (11-18)
Symptom severity, median (IQR)	29 (19-44)
Perceived percent normalcy, median (IQR)	70 (50-75)
Trends in symptoms since injury, n (%)	
No change	1 (4.3)
Steadily better	8 (35)
Steadily worse	1 (4.3)
Worse, then better	3 (13)
Better, then worse	10 (43)

XSL•FO
RenderX

MAUQ Findings

Participants' responses to the MAUQ can be seen in Figure 2. Across the 16 MAUQ questions, "agree" or "strongly agree" were the most common responses. Under the "ease of use and satisfaction" category, all 23 participants selected "agree" or "strongly agree" for the question "easy to learn." Twenty participants (87%) selected "agree" or "strongly agree" for both "easy to use" and "well organized" questions. The question with the least favorable response in this category was "appropriate time involved," where 4 participants (17%) selected "disagree" or "somewhat disagree."

Under the category "system information arrangement," 20 (87%) participants selected "agree" or "strongly agree" for the question "all functions usable," and 19 (87%) for the question "screen navigation consistent." The most unfavorably responded question for this category was "ease of recovery following mistake," where 3 participants (13%) chose either "strongly disagree," "somewhat disagree," or "disagree."

In the "usefulness" category, 20 (87%) participants responded with "agree" or "strongly agree" for "health and well-being usefulness," and 21 participants (87%) reported similarly for "helped manage health." Please see Multimedia Appendix 1 for a summary of participant responses to each question of the MAUQ.

Figure 2. Responses from participants to each of the 16 mHealth app usability questionnaire (MAUQ). Participants ranked each question using a 7-point Likert scale ranging from strongly agree (dark blue) to strongly disagree (red). Each horizontal bar indicates the breakdown of each question's responses as percentages.

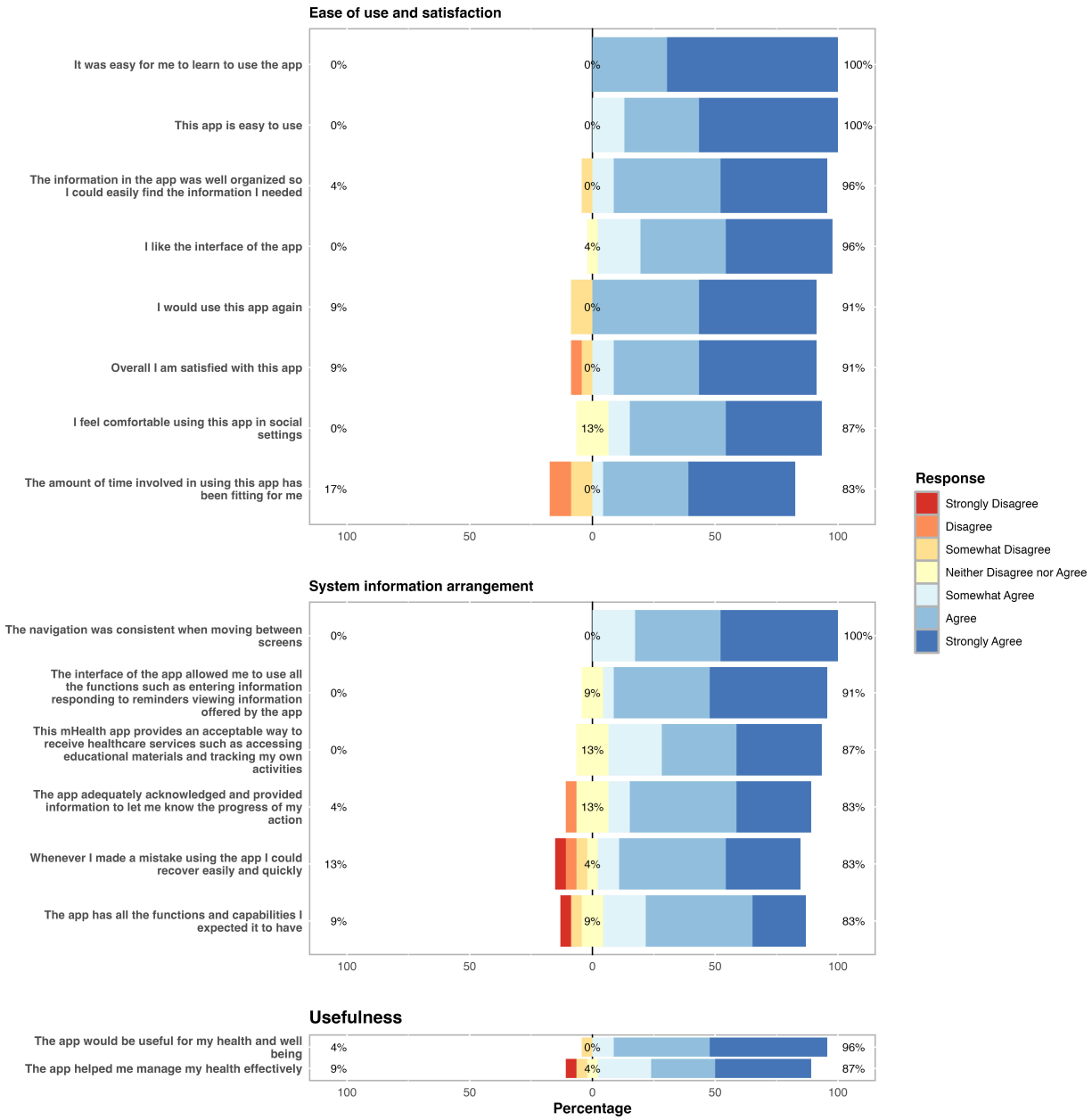

XSL•FO
RenderX

Additional Customized Questions

The additional questions specific to the app yielded similar responses to the MAUQ, with "agree" or "strongly agree" being the most common response for all five questions (Figure 3); only a single negative response was reported across all questions. The most favorable response was for "voice command clarity" where all but one participant (n=22, 95%) reported either "agree" or "strongly agree"; the other participant selected "somewhat agree." Please see Multimedia Appendix 2 for a summary of the participant responses for the additional questions.

Figure 3. Participants' responses to five additional custom questions. Participants ranked each question using a 7-point Likert scale ranging from strongly agree to strongly disagree; however, only 5 responses were chosen by participants, ranging from strongly agree (dark blue) to somewhat disagree (orange). Each horizontal bar indicates the breakdown of each question's responses as percentages.

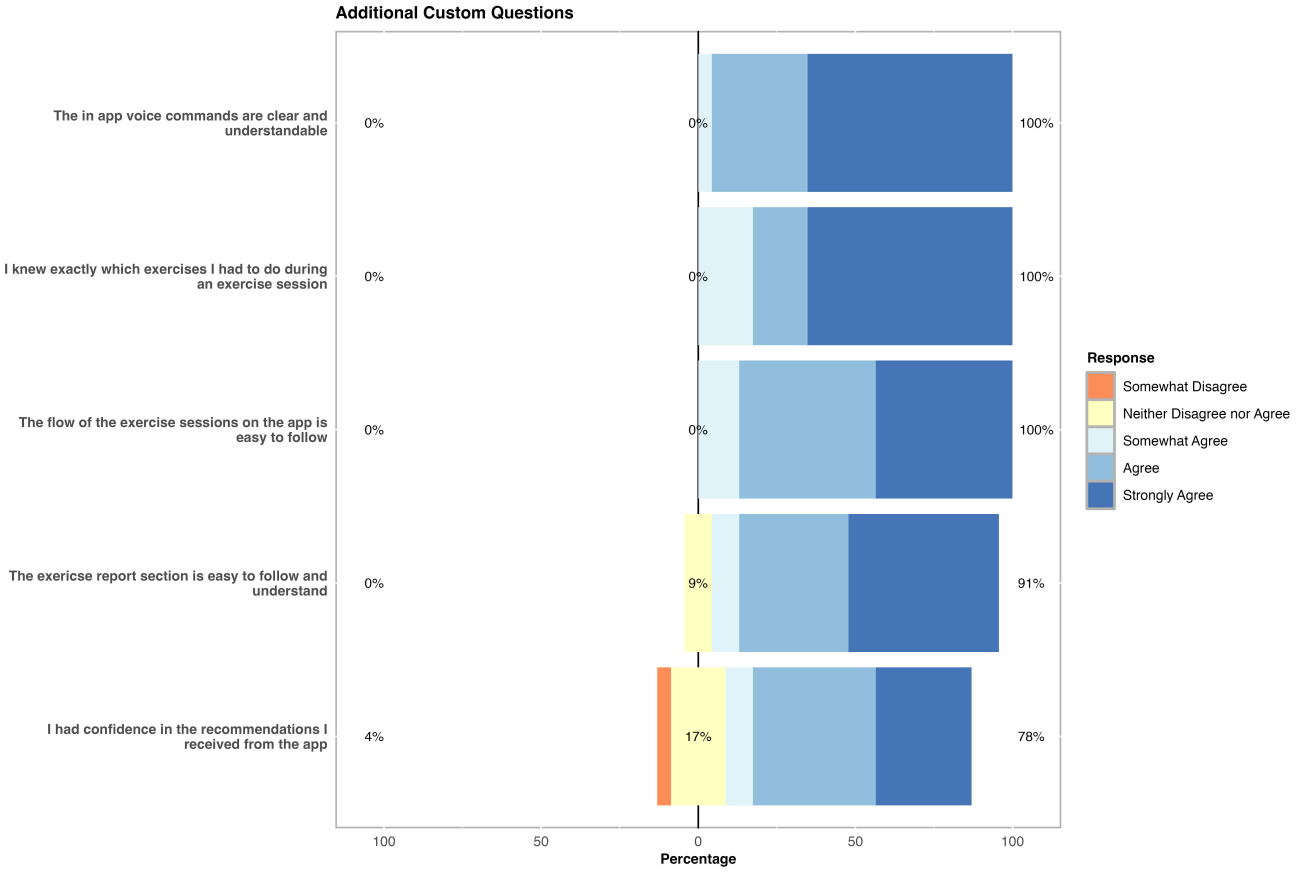

Discussion

This study aimed to assess the usability of a mobile app that provides exercise rehabilitation for individuals with persistent concussion symptoms. Usability was assessed using a combination of the MAUQ and additional custom questions. Overall, the responses were strongly positive, indicating that participants found the app both useful and easy to navigate.

The app's core features were well received by users, as indicated by the absence of disagreements for seven of the MAUQ questions. Responses to these questions were either neutral or positive among participants, which strongly supports the overall user satisfaction with the app. The most favorable responses were primarily in the "ease of use and satisfaction" category, followed closely by the "system information arrangement" category. This positive feedback is particularly significant for individuals recovering from concussions, as symptoms such as cognitive complaints (eg, difficulty concentrating) and sensitivity to screens and light are commonly reported [22-24]. Providing a mobile app that is easy to navigate and well organized is essential for enhancing adoption and encouraging consistent use throughout the recovery process.

The strongly positive feedback based on the custom questions provides evidence that the mobile app effectively meets the needs of individuals recovering from concussions. Participants reported high confidence in the app's recommendations, which is particularly significant given that the app is grounded in evidence-based exercise rehabilitation principles. Confidence in the accuracy and safety of the recommendations is critical in concussion recovery, as missteps in physical activity intensity can lead to symptom exacerbation or worsening of neurocognitive performance [25]. The intuitive flow of exercise sessions and the clarity in guiding participants on which exercises to perform suggest that the app successfully translates complex rehabilitation protocols into an easy-to-follow format. Additionally, including clear voice commands helps reduce cognitive load, making the app accessible even to those with cognitive fatigue or visual sensitivity. Furthermore, the ease of understanding the exercise report section implies that users could track their progress effectively, promoting engagement and adherence to the rehabilitation protocol. These findings underscore the importance of mHealth solutions that provide

XSL•FO
RenderX

evidence-based care and inspire user confidence, ultimately enhancing the recovery experience for concussion patients.

While participants generally found the app easy to use, our study revealed concerns regarding the time commitment and error recovery processes, both of which require further attention. Specifically, some participants found the app's required time commitment to be excessive, and others experienced difficulty with the app's response to user errors. To address these concerns, future developments of the app will aim to find an optimal balance between screen time and cognitive load. This balance is crucial for supporting individuals with concussions, who may be particularly sensitive to long periods of screen interaction and cognitive stress. Enhancements will include more efficient navigation and task completion processes, as well as integrated educational features that prompt users to take necessary breaks and provide guidance on managing screen time effectively. Moreover, improving how the app handles mistakes—making error recovery more intuitive and less stressful—is essential for fostering a supportive and rehabilitative environment. These adjustments will not only enhance the user experience but also play a pivotal role in the app's functionality, ensuring it meets the unique needs of its users during their recovery.

There are several limitations to this study. First, the sample comprised predominantly female participants, which may limit the generalizability of the findings to male participants. Future studies should aim for more balanced demographics to validate the app's usability across diverse populations. Additionally, our sample consisted of individuals with persistent symptoms, leaving the usability of the app in the acute and subacute phases of concussion recovery still unexplored. Larger-scale studies with a more diverse population are necessary to further confirm the app's usability and effectiveness across different recovery stages. It is also important to note that we intentionally excluded five questions from the original MAUQ due to the early iteration of this tool. Specifically, we chose not to assess aspects related to sharing information with health care providers and clinical utility, as practitioners were not expected to engage with a research-grade tool at this stage. As the app evolves, future research should include these additional questions to better understand how the app complements health care practitioners and facilitates integration into clinical practice.

In conclusion, the Rhea app represents a promising mHealth solution for concussion rehabilitation. The overwhelmingly positive usability feedback supports the potential for broader adoption. However, continued refinement based on user feedback and further research with larger, more diverse populations is essential to ensure the app's long-term success in facilitating concussion recovery.

Acknowledgments

This research received funding from the Connaught Fund Innovation Award, the University of Toronto and the XSeed Funding Program, Faculty of Applied Science and Engineering, University of Toronto.

Data Availability

The datasets generated or analyzed during this study are available in the public GitHub repository [26].

Conflicts of Interest

Rhea Health Inc is a mobile health tech start-up developed by the University of Toronto, and MGH and the University of Toronto own equity shares in Rhea Health Inc. APDB is an unpaid data analytic lead for Rhea Health Inc.

Multimedia Appendix 1
Summary of participants' responses to MAUQ questions.
[DOCX File, 18 KB - formative_v9i1e67275_app1.docx]

Multimedia Appendix 2
Summary of participants' responses to additional custom questions.
[DOCX File, 16 KB - formative_v9i1e67275_app2.docx]

References

1. Patricios JS, Schneider KJ, Dvorak J, et al. Consensus statement on concussion in sport: the 6th International Conference on Concussion in Sport-Amsterdam, October 2022. Br J Sports Med 2023 Jun;57(11):695-711. [doi: 10.1136/bjsports-2023-106898] [Medline: 37316210]
2. Lawrence DW, Richards D, Comper P, Hutchison MG. Earlier time to aerobic exercise is associated with faster recovery following acute sport concussion. PLoS ONE 2018;13(4):e0196062. [doi: 10.1371/journal.pone.0196062] [Medline: 29668716]
3. Hutchison MG, Di Battista AP, Lawrence DW, Pyndiura K, Corallo D, Richards D. Randomized controlled trial of early aerobic exercise following sport-related concussion: progressive percentage of age-predicted maximal heart rate versus usual care. PLoS ONE 2022;17(12):e0276336. [doi: 10.1371/journal.pone.0276336] [Medline: 36548338]

XSL•FO
RenderX

4. Micay R, Richards D, Hutchison MG. Feasibility of a postacute structured aerobic exercise intervention following sport concussion in symptomatic adolescents: a randomised controlled study. BMJ Open Sport Exerc Med 2018;4(1):e000404. [doi: 10.1136/bmjsem-2018-000404] [Medline: 30018795]
5. Leddy JJ, Hinds AL, Miecznikowski J, et al. Safety and prognostic utility of provocative exercise testing in acutely concussed adolescents: a randomized trial. Clin J Sport Med 2018 Jan;28(1):13-20. [doi: 10.1097/JSM.0000000000000431] [Medline: 29257777]
6. Leddy J, Baker JG, Haider MN, Hinds A, Willer B. A physiological approach to prolonged recovery from sport-related concussion. J Athl Train 2017 Mar;52(3):299-308. [doi: 10.4085/1062-6050-51.11.08] [Medline: 28387557]
7. Mohr DC, Kwasny MJ, Meyerhoff J, Graham AK, Lattie EG. The effect of depression and anxiety symptom severity on clinical outcomes and app use in digital mental health treatments: meta-regression of three trials. Behav Res Ther 2021 Dec;147:103972. [doi: 10.1016/j.brat.2021.103972] [Medline: 34600398]
8. Bendtsen M, Müssener U, Linderoth C, Thomas K. A mobile health intervention for mental health promotion among university students: randomized controlled trial. JMIR Mhealth Uhealth 2020 Mar 20;8(3):e17208. [doi: 10.2196/17208] [Medline: 32196462]
9. Burke LE, Ma J, Azar KMJ, et al. Current science on consumer use of mobile health for cardiovascular disease prevention: a scientific statement from the American Heart Association. Circulation 2015 Sep 22;132(12):1157-1213. [doi: 10.1161/CIR.0000000000000232] [Medline: 26271892]
10. Klimis H, Thakkar J, Chow CK. Breaking barriers: Mobile health interventions for cardiovascular disease. Can J Cardiol 2018 Jul;34(7):905-913. [doi: 10.1016/j.cjca.2018.02.012] [Medline: 29887218]
11. Gandhi S, Chen S, Hong L, et al. Effect of mobile health interventions on the secondary prevention of cardiovascular disease: systematic review and meta-analysis. Can J Cardiol 2017 Feb;33(2):219-231. [doi: 10.1016/j.cjca.2016.08.017] [Medline: 27956043]
12. de Ridder M, Kim J, Jing Y, Khadra M, Nanan R. A systematic review on incentive-driven mobile health technology: as used in diabetes management. J Telemed Telecare 2017 Jan;23(1):26-35. [doi: 10.1177/1357633X15625539] [Medline: 26888421]
13. Christopher E, Alsaffarini KW, Jamjoom AA. Mobile health for traumatic brain injury: a systematic review of the literature and mobile application market. Cureus 2019 Jul 10;11(7):e5120. [doi: 10.7759/cureus.5120] [Medline: 31523551]
14. Juengst SB, Hart T, Sander AM, Nalder EJ, Pappadis MR. Mobile health interventions for traumatic brain injuries. Curr Phys Med Rehabil Rep 2019 Dec;7(4):341-356. [doi: 10.1007/s40141-019-00240-9]
15. Schmidt M, Babcock L, Kurowski BG, Cassedy A, Sidol C, Wade SL. Usage patterns of an mHealth symptom monitoring app among adolescents with acute mild traumatic brain injuries. J Head Trauma Rehabil 2022;37(3):134-143. [doi: 10.1097/HTR.0000000000000768] [Medline: 35125434]
16. Huber DL, Thomas DG, Danduran M, Meier TB, McCrea MA, Nelson LD. Quantifying activity levels after sport-related concussion using actigraph and mobile (mHealth) technologies. J Athl Train 2019 Sep;54(9):929-938. [doi: 10.4085/1062-6050-93-18] [Medline: 31411898]
17. Lee H, Sullivan SJ, Schneiders AG, et al. Smartphone and tablet apps for concussion road warriors (team clinicians): a systematic review for practical users. Br J Sports Med 2015 Apr;49(8):499-505. [doi: 10.1136/bjsports-2013-092930] [Medline: 24668048]
18. Worthen-Chaudhari L, McGonigal J, Logan K, Bockbrader MA, Yeates KO, Mysiw WJ. Reducing concussion symptoms among teenage youth: evaluation of a mobile health app. Brain Inj 2017;31(10):1279-1286. [doi: 10.1080/02699052.2017.1332388] [Medline: 28665690]
19. Hutchison MG, Di Battista AP, Loenhart MM. A continuous aerobic resistance exercise protocol for concussion rehabilitation delivered remotely via a mobile app: feasibility study. JMIR Form Res 2023 Jun 19;7:e45321. [doi: 10.2196/45321] [Medline: 37335605]
20. Echemendia RJ, Meeuwisse W, McCrory P, et al. The sport concussion assessment tool 5th Edition (SCAT5): background and rationale. Br J Sports Med 2017 Jun;51(11):848-850. [doi: 10.1136/bjsports-2017-097506] [Medline: 28446453]
21. Zhou L, Bao J, Setiawan IMA, Saptono A, Parmanto B. The mHealth app usability questionnaire (MAUQ): development and validation study. JMIR Mhealth Uhealth 2019 Apr 11;7(4):e11500. [doi: 10.2196/11500] [Medline: 30973342]
22. Stillman AM, Madigan N, Torres K, Swan N, Alexander MP. Subjective cognitive complaints in concussion. J Neurotrauma 2020 Jan 15;37(2):305-311. [doi: 10.1089/neu.2018.5925] [Medline: 31407632]
23. Mansur A, Hauer TM, Hussain MW, et al. A nonliquid crystal display screen computer for treatment of photosensitivity and computer screen intolerance in post-concussion syndrome. J Neurotrauma 2018 Aug 15;35(16):1886-1894. [doi: 10.1089/neu.2017.5539] [Medline: 30074876]
24. Abusamak M, Alrawashdeh HM. Post-concussion syndrome light sensitivity: a case report and review of the literature. Neuroophthalmology 2022;46(2):85-90. [doi: 10.1080/01658107.2021.1983612] [Medline: 35273410]
25. Majerske CW, Mihalik JP, Ren D, et al. Concussion in sports: postconcussive activity levels, symptoms, and neurocognitive performance. J Athl Train 2008;43(3):265-274. [doi: 10.4085/1062-6050-43.3.265] [Medline: 18523563]
26. Hutchison-et-al-2025-RHEA-usability. GitHub. 2025. URL: https://github.com/dibatti5/Hutchison-et-al-2025-RHEA-Usability [accessed 2025-04-03]

XSL•FO
RenderX

Abbreviations

HR: heart rate
MAUQ: mHealth app usability questionnaire
mHealth: mobile health
SCAT-5: sports concussion assessment tool
TBI: traumatic brain injury

Edited by A Mavragani; submitted 07.10.24; peer-reviewed by J Greenberg; revised version received 10.02.25; accepted 20.02.25; published 11.04.25.

Please cite as:
Hutchison MG, Di Battista AP, Pyndiura KL
Evaluating User Experience and Satisfaction in a Concussion Rehabilitation App: Usability Study
JMIR Form Res 2025;9:e67275
URL: https://formative.jmir.org/2025/1/e67275
doi:10.2196/67275

© Michael G Hutchison, Alex P Di Battista, Kyla L Pyndiura. Originally published in JMIR Formative Research (https://formative.jmir.org), 11.4.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Oxidative Stress Markers and Prediction of Severity With a Machine Learning Approach in Hospitalized Patients With COVID-19 and Severe Lung Disease: Observational, Retrospective, Single-Center Feasibility Study

Olivier Raspado[1], Dr Med; Michel Brack[2], Dr Med; Olivier Brack[3]; Mélanie Vivancos[4]; Aurélie Esparcieux[1], Dr Med; Emmanuelle Cart-Tanneur[5], Dr Bio; Abdellah Aouifi[1], Dr Med

[1]Infirmerie Protestante, 1 Chemin du Penthod, Caluire-et-Cuire, France

[2]Oxidative Stress College, La Garenne-Colombes, France

[3]Statistique Industrielle Khi² Consulting (KSIC), Bayet, France

[4]Clinical Research and Innovation Department, Infirmerie Protestante, Caluire-et-Cuire, France

[5]Eurofins Biomnis Laboratory, Lyon, France

Corresponding Author:
Olivier Raspado, Dr Med
Infirmerie Protestante, 1 Chemin du Penthod, Caluire-et-Cuire, France

Abstract

Background: Serious pulmonary pathologies of infectious, viral, or bacterial origin are accompanied by inflammation and an increase in oxidative stress (OS). In these situations, biological measurements of OS are technically difficult to obtain, and their results are difficult to interpret. OS assays that do not require complex preanalytical methods, as well as machine learning methods for improving interpretation of the results, would be very useful tools for medical and care teams.

Objective: We aimed to identify relevant OS biomarkers associated with the severity of hospitalized patients' condition and identify possible correlations between OS biomarkers and the clinical status of hospitalized patients with COVID-19 and severe lung disease at the time of hospital admission.

Methods: All adult patients hospitalized with COVID-19 at the Infirmerie Protestante (Lyon, France) from February 9, 2022, to May 18, 2022, were included, regardless of the care service they used, during the respiratory infectious COVID-19 epidemic. We collected serous biomarkers from the patients (zinc [Zn], copper [Cu], Cu/Zn ratio, selenium, uric acid, high-sensitivity C-reactive protein [hs-CRP], oxidized low-density lipoprotein, glutathione peroxidase, glutathione reductase, and thiols), as well as demographic variables and comorbidities. A support vector machine (SVM) model was used to predict the severity of the patients' condition based on the collected data as a training set.

Results: A total of 28 patients were included: 8 were asymptomatic at admission (grade 0), 14 had mild to moderate symptoms (grade 1) and 6 had severe to critical symptoms (grade 3). As the first outcome, we found that 3 biomarkers of OS were associated with severity (Zn, Cu/Zn ratio, and thiols), especially between grades 0 and 1 and between grades 0 and 2. As a second outcome, we found that the SVM model could predict the level of severity based on a biological analysis of the level of OS, with only 7% misclassification on the training dataset. As an illustrative example, we simulated 3 different biological profiles (named A, B, and C) and submitted them to the SVM model. Profile B had significantly high Zn, low hs-CRP, a low Cu/Zn ratio, and high thiols, corresponding to grade 0. Profile C had low Zn, low selenium, high oxidized low-density lipoprotein, high glutathione peroxidase, a low Cu/Zn ratio, and low glutathione reductase, corresponding to grade 2.

Conclusions: The level of severity of pulmonary damage in patients hospitalized with COVID-19 was predicted using an SVM model; moderate to severe symptoms in patients were associated with low Zn, low plasma thiol, increased hs-CRP, and an increased Cu/Zn ratio among a panel of 10 biomarkers of OS. Since this panel does not require a complex preanalytical method, it can be used and studied in other pathologies associated with OS, such as infectious pathologies or chronic diseases.

(JMIR Form Res 2025;9:e66509) doi:10.2196/66509

KEYWORDS

oxidative stress; COVID-19; SARS-CoV-2; coronavirus; respiratory; infectious; pulmonary; respiration disorders; hospitalization; machine learning; ML; biomarker; lung; severity; prediction

XSL•FO
RenderX

Introduction

Oxidative stress (OS) is an imbalance between the production of free radicals and the body's ability to neutralize them, leading to damage to cells, proteins, and DNA [1-3]. This is a natural physiological process that is not harmful if it remains balanced, but it can be involved in the onset or worsening of chronic diseases such as cancer [4], metabolic disorders [5], atherosclerosis, [6] and cardiovascular diseases [7,8]. OS can also contribute to nerve cell death and deterioration of brain function, leading to neurodegenerative diseases [9,10].

Healthy lifestyle habits can help reduce OS and its harmful effects, for example, learning to manage emotional stress [11]; performing regular physical activity to promote the production of endogenous antioxidant enzymes [12]; and adopting a balanced diet that is able to provide antioxidants such as vitamins C and E, beta-carotene, and polyphenols and is rich in fruits, vegetables, whole grains, and lean protein sources [13]. Avoiding or reducing exposure to environmental toxins, such as toxic chemicals, cigarette smoke, and air pollutants, can also reduce OS [14].

Furthermore, many respiratory viral infections, including COVID-19, cause death of the infected cells, activation of the innate immune response, and secretion of inflammatory cytokines. All these processes are associated with the development of OS, which makes an important contribution to the pathogenesis of viral infections. COVID-19 is a complex disease in which interaction of the virus with target cells, action of the immune system, and the body's systemic response to these events are closely intertwined [15].

COVID-19 infection has various levels of pulmonary involvement, with asymptomatic, benign, serious, or even fatal forms, especially when the infected patient has comorbidities such as a history of cancer or malignant hematological diseases; concurrent chemotherapy treatment; severe chronic kidney disease, with or without dialysis; a history of solid organ transplantation or allogeneic hematopoietic stem cell transplantation; chronic polypathologies such as diabetes, high blood pressure, and obesity; or 2 or more organ failures [16]. All of these associated and chronic pathologies will influence the response of the immune system to COVID-19 infection and inflammation, as well as the level of OS present before and during the infection. Recent observational studies have reported that increasing COVID-19 severity may be responsible for a worse prognosis [17] among diabetes patients [18] and increased severity of lung disease [19]. Two meta-analyses showed that supplementation with antioxidants (in the form of vitamins and trace elements) in critically ill patients was associated with decreased mortality and counteracted OS damage [20,21].

The identification of OS biomarkers is essential for early diagnosis of chronic disease, for evaluating treatment efficacy, for monitoring lifestyle interventions, and for understanding the mechanisms involved. The decrease in the blood concentration of thiol proteins is one of the most relevant markers of OS [22]. However, clusters of evidence are emerging from the recent literature showing that levels of biomarkers of oxidative damage in biological fluids can be used for the prediction of measured concentrations of a limited number of exogenous and endogenous antioxidants [23].

Since March 2020, our health care and medical team began to hypothesize that antioxidant supplementation for patients in the intensive care unit (ICU) could improve their prognosis. We were heavily involved in treatment on a daily basis and tried to understand this serious and complex respiratory pathology and treat it as effectively as possible. Our patients benefit from a precise clinical assessment and comprehensive blood biological assays during their initial care and during their hospital stay. Other medical teams have been able to establish scores or panels composed of different blood biological assays to evaluate certain markers of OS and the inflammatory response [16,18], but the measurement and analysis of these biomarkers require complex preanalytical and analytical processes that cannot be carried out during hospitalization in the acute phase, that is, at the patient's bedside. To identify relevant OS biomarkers, we designed the OXYCOVID study as a feasibility study with the primary objective of testing whether the methodological protocol would be appropriate and feasible in the context of the ICU and ICU admission. The secondary objectives of this study were to identify relevant OS biomarkers associated with the severity of hospitalized patients and to identify whether there is a correlation between the clinical status of hospitalized patients with COVID-19 and severe lung disease at admission using artificial intelligence tools, such as a support vector machines (SVMs). SVMs are used to solve classification problems (eg, category assignment), discrimination problems, and regression problems. They require only a few parameters. The performance of the SVM family of machine learning algorithms is generally as good as, or even better than, neural networks or Gaussian models. These biological assays, if they can be used to determine an OS level, could become a routine element of care, allowing a more precise assessment of the severity of lung lesions in the event of COVID-19 infection.

Methods

Study Design

The OXYCOVID study is an observational, retrospective, single-center study based on medical data collected during hospitalization. All adult patients (ie, aged >18 years) hospitalized with COVID-19 at the Infirmerie Protestante (Lyon, France) were included, regardless of the care service they used. Patients were not included if they did not consent to the collection and processing of their personal data, if they were pregnant or breastfeeding, or if they had a protected status (ie, curatorship or other legal protection; deprivation of liberty by judicial or administrative decision). The data were collected from the period from February 9, 2022, to May 18, 2022. The final sample for analysis included 28 patients meeting the inclusion and exclusion criteria. All biological data were recorded in each patient's digital medical record.

Data Collection

The data were first collected and then sorted specifically for each patient after checking for the absence of errors and for missing data. After verification, no discrepancies were found and the clinical and biological data were judged to be usable.

XSL•FO
RenderX

The data are fully anonymized. The data included sociodemographics (sex and age), comorbidities, and risk factors (weight, height, smoking status, hypertension, diabetes, chronic renal disease, severe chronic respiratory disease, autoimmune disease, immune deficiency, and cancer). COVID-19 vaccines were also recorded, as were data related to the current COVID-19 infection (date of the latest polymerase chain reaction test from a nasopharyngeal swab, variant, and date of first symptoms) and the reasons for hospitalization.

At admission, clinical symptoms were recorded, including the presence of fever (temperature >38°C), cough (yes or no), dyspnea (felt by the patient during the initial questioning), myalgia (felt by the patient during the initial questioning), fatigue (felt by the patient during the initial questioning), diarrhea, oxyge saturation (SpO_2), oxygenation system used (nasal oxygen therapy, mask oxygen therapy, or Optiflow system), and pulmonary embolism (determined by a thoracic computed tomography angiogram). Biological data were taken from venous blood samples during hospitalization and included complete blood count; levels of urea, creatinine, C-reactive protein (CRP), troponin, D-dimer, ferritin, lactase dehydrogenase, and prothrombin; activated partial thromboplastin time; levels of total bilirubin, gamma-GT, transaminase, albumin, and interleukin-6; and an OxyCheck panel including the following 10 markers: zinc (Zn), copper (Cu), Cu/Zn ratio, selenium, uric acid, high sensitivity CRP (hs-CRP), oxidized low-density lipoprotein, glutathione peroxidase, glutathione reductase, and thiols.

Patients were triaged at admission into 3 mutually exclusive grades. Grade 0 represented asymptomatic patients. Grade 1 represented patients with mild to moderate clinical signs (SpO_2 >94% in ambient air or with oxygen [maximum 6 L/min], fever, cough, dyspnea, and respiratory rate between 20 and 30 respirations/min). Grade 2 represented the most severe patients, that is, those with severe pneumonia (respiratory rate >30 respirations/min; arterial oxygen saturation <90% despite fraction of inspired oxygen >40% [>6 L/min]) or a rapid increase in oxygen needs (tachycardia [heart rate >120 beats/min]), sweating, agitation, a need for noninvasive ventilatory support with Optiflow (flow rate <60 L/min at 60% fraction of inspired oxygen, ROX index [oxygen saturation/respiration rate index] >3). Grade 2 also included critical patients with severe pneumonia or acute respiratory distress syndrome (index ROX <3 [if the ROX index was <2.5, the patient was intubated]; Optiflow flow rate >60 L/min at 60% fraction of inspired oxygen [if the Optiflow flow rate was >80 L/min at 100% of fraction of inspired oxygen, the patient was intubated]; partial pressure of oxygen in the arterial blood/fraction of inspired oxygen ≤300 mm Hg) and patients with septic shock; acute, life-threatening organ dysfunction related to a known or suspected infection; an altered mental state (delirium or confusion); oliguria; skin mottling; laboratory evidence of coagulopathy, thrombocytopenia, acidosis, or elevated lactic acid; or hyperbilirubinemia.

Statistical Analysis

Baseline characteristics are reported as counts and percentages for categorical variables and medians and IQR for continuous variables. Each patient was assigned only one grade, so the observations are considered independent. We tested for significant outliers among all the biomarkers, determined if distributions followed a normal distribution using the Shapiro-Wilk test, and used the Levene test for homogeneity of variances. Due to the small sample size, the criteria for outlier values were not strictly applied in the ANOVA. Tukey post-hoc tests were used for multiple pairwise comparisons between severity groups when significant differences were observed in the ANOVA. Finally, an SVM was used to predict assignment for each grade, taking into account the entire analysis profile. Statistical analysis was performed with R (version 4.1.1; R Project for Statistical Computing) and the *tidyverse*, *ggpubr*, and *rstatix* libraries. The level of significance was set at $P<.05$.

Considering that collinearities within our dataset induced by obvious links between biological parameters disrupted traditional methods (eg, logistic regression or the classification method), we used an SVM to predict assignment to each grade for each patient, considering the entire analysis profile. The SVM algorithm is a supervised learning algorithm that uses training data to build a model to predict or classify new observations. SVMs are a machine learning method for nonlinear regression classification [24-26]. Our approach consisted in using the biomarker database as a training dataset for the SVM model (eg, the SVM training dataset). This resulted in a model capable of predicting the probability that a patient would be assigned a specific grade based on the results of a biological analysis, provided that the values did not exceed those used in the initial dataset; that is, the sum of the 3 probabilities was equal to 1. This method resulted in a profile in the form of a graph establishing the link between a combination of biomarkers and the probability of being assigned to the different severity grades. The SVM analysis was performed using JMP Pro (version 17.2; JMP Statistical Discovery).

Ethical Considerations

Participation in this study was voluntary, and all participants received an individual notice at admission to the ICU, which included informed consent (the individual notice is included in Multimedia Appendix 1). This study was based on data available in the medical record and did not require any additional medical visits or examinations; it also did not lead to any changes in the management or treatment of patients. The data were irreversibly deidentified for the statistical analysis. Participants did not receive any compensation. This study was approved by the ethics committee of Infirmerie Protestante of Lyon (CE-23‐01).

Results

Baseline Characteristics of Patients

A total of 28 patients met the inclusion criteria. The median age was 75 (IQR 68.75-80) years, and the patients were mostly male (n=17, 60.71%; Table 1). At admission, 8 patients were asymptomatic (grade 0), 14 had mild- to moderate-severity illness (grade 1) and 6 had severe to critical symptoms (grade 3). The most common reported symptoms were cough and fever.

XSL•FO
RenderX

Table . Baseline characteristics of the 28 patients hospitalized for COVID-19, including hospitalized patients; socio-demographics, vaccination status, clinical symptoms at admission, and comorbidities.

	All patients (N=28)	Asymptomatic patients (grade 0; n=8)	Mild to moderate patients (grade 1; n=14)	Severe to critical patients (grade 2; n=6)
Sociodemographics				
Sex ratio	1.54	0.6	1.8	5
Male, n (%)	17 (61)	3 (38)	9 (64)	5 (83)
Age (years), median (IQR)	75 (68.75-80)	71.5 (65.5-75.25)	75.5 (68.75-80.25)	71.0 (65.5-78.75)
COVID-19 vaccination, n (%)				
Vaccinated (primary series)	5 (18)	1 (12)	3 (21)	1 (17)
Unvaccinated	3 (11)	0 (0)	2 (14)	0 (0)
Unknown	20 (71)	7 (87)	9 (64)	5 (83)
Clinical symptoms at admission, n (%)				
Alteration of the general state	12 (43)	0 (0)	9 (64)	3 (50)
Fever (temperature >38°C)	12 (43)	0 (0)	8 (57)	4 (67)
Cough	10 (36)	0 (0)	9 (63)	1 (17)
Dyspnea (>20 respirations/min)	10 (36)	0 (0)	8 (57)	2 (33)
Myalgia	2 (7)	2 (25)	0 (0)	0 (0)
Diarrhea	3 (11)	0 (0)	3 (21)	0 (0)
Pulmonary embolism	1 (4)	0 (0)	1 (7)	0 (0)
Comorbidities, n (%)				
Diabetes	11 (39)	3 (37)	5 (36)	3 (50)
Hypertension	14 (50)	3 (37)	8 (57)	3 (50)
Cancer	13 (46)	1 (12)	9 (64)	3 (50)

Descriptive Analysis of the OS Biomarkers

Results for the detection of outliers, the normality test, and the homogeneity of variance are provided in Table 2. The OS biomarkers that met the criteria for the ANOVA were Zn (P=.002), Cu/Zn ratio (P=.009), hs-CRP (P=.02), and thiols (P=.004).

XSL•FO
RenderX

Table . Verification of the conditions for the ANOVA (outlier detection, verification of the normality of values between groups, and verification of homogeneity of variance between groups) and results of a 1-way ANOVA.

	Outlier detection, n			Shapiro-Wilk test [a]			Levene test[b]	ANOVA
	Grade 0	Grade 1	Grade 2	Grade 0	Grade 1	Grade 2		
Zinc	1	0	0	.97	.97	.66	.35	0.002
Copper	0	0	0	.98	.31	.80	.78	0.343
Copper/zinc ratio	0	1	0	.68	.008	.24	.52	0.009
Selenium	0	0	0	.50	.46	.045	.63	0.303
Uric acid	1	1	0	.02	.19	.60	.89	0.528
High-sensitivity C-reactive protein	0	0	0	<.001	.04	.72	.08	0.017
Oxidized low-density lipoprotein	1	0	0	.18	.70	.33	.55	0.872
Glutathione peroxydase	0	0	1	.18	.51	.03	.99	0.380
Glutathione reductase	0	0	0	.23	.88	.19	.03	0.194
Thiols	1	0	0	.34	.76	.07	.63	0.004

[a]If the data are normally distributed, the *P* value is greater than .05.

[b]If the variances are homogeneous, the *P* value is greater than .05.

Statistical differences were observed between the grade 0 and grade 1 patients and between the grade 0 and grade 2 patients for the 3 OS biomarkers, but the difference was nonsignificant between the grade 1 and 2 patients (Table 3).

Table . Comparison between the severity grades depending on the 3 OS biomarkers using the Tukey post-hoc test.

	Grade 0 versus grade 1	Grade 0 versus grade 2	Grade 1 versus grade 2
Zinc	0.005	0.003	0.609
Copper/zinc ratio	0.012	0.029	0.981
Thiols	0.004	0.019	1

Predictive Model Using SVM

Table 4 shows the results obtained from the SVM (Multimedia Appendix 2 provides detailed results). The clinical classification and the SVM prediction differed in only 2 of the 28 cases. One of these cases was a patient who was clinically classified as grade 1; he was symptomatic but did not meet the criteria for intubation. The biological analysis profile was abnormal (low thiols [4.1], low zinc, very high Cu/Zn ratio [2.2]). The SVM model classified this patient as grade 2 with a probability of 70.7%. He remained hospitalized for 1 month with active therapy, made good progress, and was discharged. The other patient was initially classified as grade 2; he was strongly symptomatic with high dependence on high-flow oxygen therapy. The biological analysis profile was abnormal (Cu/Zn ratio=3.57). The SVM model classified this case as grade 1 with a probability of 65.8%, probably due to slightly decreased plasma thiol and zinc levels. The clinical course of this patient quickly became favorable.

With a misclassification rate of only 7%, we estimated that the training of the SVM model was good enough and that it was able to predict the grade of patients with COVID-19 infection based on the results of the initial specific biological assays (upon arrival).

Table . Comparison of the clinically observed grade and the most likely grade (eg, a higher probability of belonging to the grade) predicted by the support vector machine (SVM).

Clinically observed grade	Most likely grade predicted by SVM model (patients), n		
	Grade 0	Grade 1	Grade 2
Grade 0	8	0	0
Grade 1	0	13	1
Grade 2	0	1	5

XSL•FO
RenderX

Profiles of Biological Results as Illustrative Examples

As an illustrative example, we created 3 biological result profiles (named A, B, and C). These profiles were submitted to the SVM model to determine the probability that they belonged to grade 0, 1, or 2 (Table 5). Profile B had a significantly high Zn level, low hs-CRP, a low Cu/Zn ratio, and high thiols and was predicted to be grade 0. At the opposite end, profile C had low Zn, low selenium, high oxidized LDL, high glutathione peroxidase, a low Cu/Zn ratio, and low glutathione reductase, corresponding to grade 2. Figure 1, Figure 2, and Figure 3 illustrate the predicted probabilities per biomarker and per profile for profiles A to C.

Table . Examples of 3 different biological profiles and the probability that they belonged to each grade of severity, as determined by the trained support vector machine (Multimedia Appendix 2 provides complete results).

Profile	Zinc (µg/L)	Copper (µg/L)	Selenium (µg/L)	Uric acid (µmol/L)	High-sensitivity C-reactive protein (mg/L)	Oxidized low-density lipoprotein (U/L)	Glutathione peroxidase (U/L)	Copper/zinc ratio	Glutathione reductase (U/g of hemoglobin)	Thiols (µmol/g of protein)	P[a] grade 0	P grade 1	P grade 2
A	584.68	1300.82	55.96	329.82	134.34	58.75	403.54	2.35	9.80	4.58	12.8%	73.9%	13.3%
B	850.00	850.00	70.00	350.00	2.00	30.00	400.00	1.00	9.00	7.00	99%	0.7%	0.3%
C	595.00	1313.00	75.00	266.00	206.00	91.00	600.00	2.20	9.70	4.10	5.9%	23.9%	70.2%

[a]P: probabililty.

Figure 1. Predicted probabilities of belonging to grade 0, 1, or 2 per biomarker for profile A. Values shown in red indicate the specific values for the profile. P: probability.

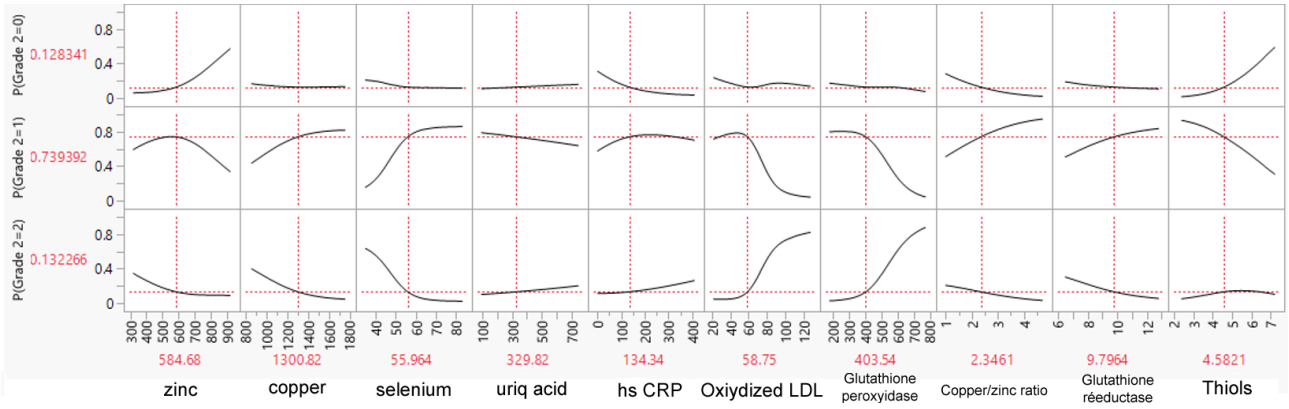

Figure 2. Predicted probabilities of belonging to grade 0, 1, or 2 per biomarker for profile B. Values shown in red indicate the specific values for the profile. P: probability.

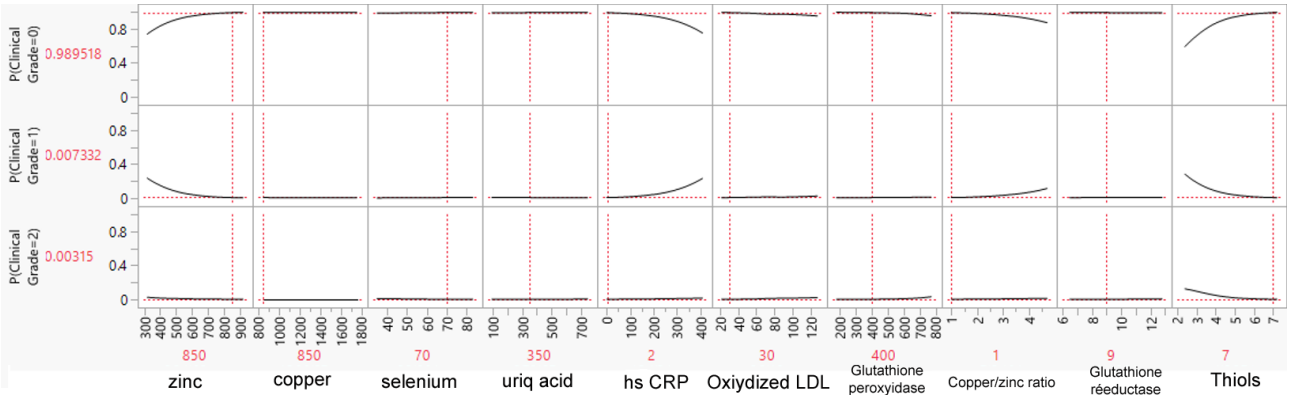

XSL•FO
RenderX

Figure 3. Predicted probabilities of belonging to grade 0, 1, or 2 per biomarker for profile C. Values shown in red indicate the specific values for the profile P: probability.

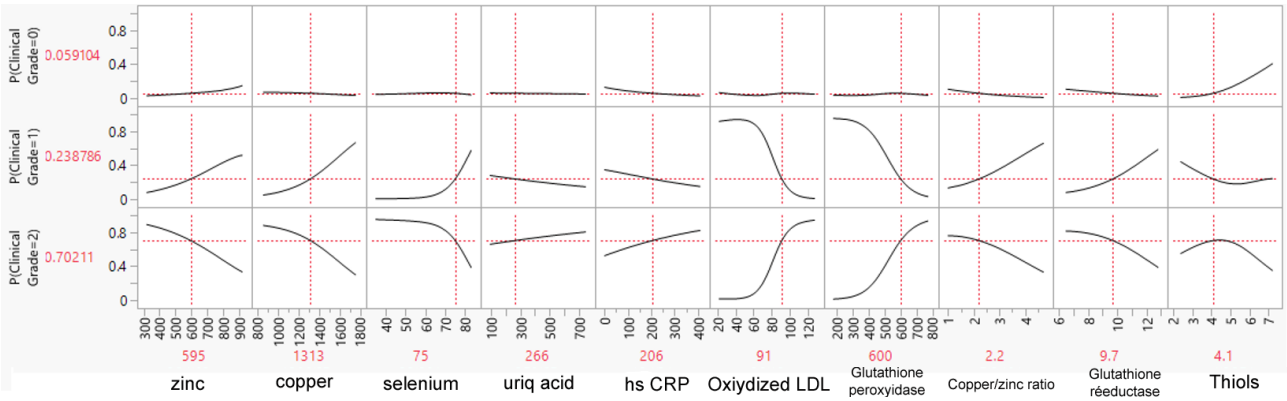

Discussion

Main Findings

Initially, we assumed that antioxidant supplementation for patients in the ICU would reduce the severity of clinical signs and improve their prognosis. The main challenge of this feasibility study was performing a preanalytical and analytical process for measuring OS biomarkers. As a first observational finding, this study identified 3 biomarkers of OS (Zn, Cu/Zn ratio, and thiols) associated with the severity of patients, especially patients with a severity grade of 0 (ie, asymptomatic) or 1 (ie, mild to moderate severity). As a second analytical finding, we found that the SVM model could provide a prediction of the level of severity based on a biological analysis of the level of OS in a relatively limited cohort and during an epidemic of respiratory infectious disease, with only 7% misclassified cases in the training dataset.

The originality of this work is the determination of the grade of clinical severity through the analysis of specific OS biomarkers using a machine learning model. What distinguishes an SVM algorithm from other classical approaches is that when the response is continuous, the fitted models are called a support vector regression (SVR) model. In a typical regression problem, the goal is to fit a model that minimizes the error between a predicted response and the actual response. In an SVR problem, the goal is to fit a model such that the error between a predicted response and the actual response lies in a range of $-\varepsilon$ to ε. This allows for more flexible fitting. In JMP Pro, ε is equal to 0.1. The SVR algorithm doubles the data by creating 2 classes, $Y + \varepsilon$ and $Y - \varepsilon$. Then, the same algorithm used for the classification problem is also used for the prediction problem (ie, SVR).

Maximization in SVM algorithms is done by solving a quadratic programming problem. In JMP Pro, the algorithm used by the SVM platform is based on the sequential minimal optimization (SMO) algorithm introduced by John Platt in 1998 [27]. Typically, the SVM quadratic programming problem is very large. The SMO algorithm divides the overall quadratic programming problem into a series of smaller problems. The smaller problems are solved analytically rather than numerically, which means that they produce closed-form solutions. Therefore, the SMO algorithm is more efficient than solving the overall quadratic programming problem. While this univariate analysis did not find any differences between grade 1 and grade 2, the trained SVM model estimated the severity grade from the same OS biomarkers; this could have a significant impact in the management of patients at risk of sudden respiratory decompensation. Although the power of the model would benefit from being improved with clinical and biological data from a larger cohort, this approach could be applied to other types of epidemics and other biological assays, as well as to the investigation of other diseases.

Comparison With Prior Work

We know that moderate and severe forms of COVID-19 are associated with high OS and a low total antioxidant level, which is measured with complex preanalytical and analytical methods [28,29] that are not generally used in all laboratories, have long delays in receiving results, and are not compatible with management in ICUs [30]. Moderate and severe forms of COVID-19 are accompanied by an increase in inflammatory biomarkers, such as CRP, procalcitonin, the neutrophil to lymphocyte ratio, and hemostasis parameters (activated partial thromboplastin time, prothrombin time, D-dimers, and fibrinogens) [29]. The natural correction of low levels of endogenous antioxidants is slow, with a tendency to recover 3 months after hospital discharge [28].

High OS is not only specific to severe disease or infection, but positively influences their severity. A basic infectiology study with a porcine animal model of infection with the avian-type H1N1 European swine influenza virus has shown that prior infection of animals with mycoplasma generates a high level of OS, increasing the severity of influenza and reducing animal performance; host responses could be influenced by diet [31].

The systemic OS status, determined using many biomarkers, was still significantly increased in recovered COVID-19 patients during the recovery phase. Nonhospitalized individuals with COVID-19 presented signs of systemic OS, which is longitudinally associated with the development of post–COVID-19 condition. Regular assessment of plasma thiol levels as a monitoring biomarker and supplementation in cases of deficiency might be potential therapeutic targets at the onset of post–COVID-19 condition [32].

From the beginning of the epidemic, plasma thiol assays at admission were a promising tool to predict ICU admission in patients with COVID-19 [19]. The principles of nutrition for patients in intensive care follow validated recommendations

XSL•FO
RenderX

with a high level of evidence [33]. Faced with inflammatory syndrome and significant catabolism, caloric and protein intake must be sufficient to combat muscle wasting and the risk of infection. In products intended for parenteral or enteral nutrition, specific micronutrients or macronutrients intended to modulate inflammatory processes and optimize the body's immunological or metabolic responses are added, such as the antioxidants β-carotene, vitamin A, vitamin C, or vitamin E, in order to combat OS [34], which plays a central role in the development of major inflammatory states and visceral failure through the production of reactive oxygen species (free radicals), which influence the consumption, distribution, and level of antioxidants [35]. Lowered levels of antioxidants are common and are associated with an increase in morbidity and mortality [36]. Two meta-analyses showed that supplementation with antioxidants (eg, vitamins and trace elements) among ICU patients is associated with a reduction in mortality [20,21] and a reduction in mechanical ventilation time [20], but two more recent meta-analyses showed that the evidence for these benefits remained questionable because they were measured in small study populations and with highly variable cocktails of antioxidant nutrients [23,37]. Currently, it is not recommended to provide massive doses of antioxidants to patients treated in intensive care.

Strengths and Limitations

The main limitation of this study is the sample size, with less than 10 patients in the grade 0 and grade 2 groups. Despite the small number of patients, we observed a statistical difference between grade 0 and grade 1 patients in terms of biomarkers associated with severity. The small sample size was caused by a short study inclusion period due to the hospital context and the ongoing crisis, which was not suited to the inclusion of more patients or to the inclusion of other hospitals. The medical and nursing team has continued to collect relevant clinical and biological data concerning patients treated for COVID-19 infection since the end of the study. These new data could be aggregated with those collected during this study and analyzed with an SVM in order to improve the performance of the severity stage prediction. As a beneficial side effect, this study allowed us to test the ability of the clinical research team to study OS biomarkers and identify covariates and risk factors for further investigation after using propensity scoring to adjust for comorbidities and sociodemographic variables of patients.

As early as 2020, applying the SVM method to COVID-19 was reported in the scientific literature, but it was mainly used for diagnosis and prediction of severity in a radiological manner and as a scanner [26]. The prediction of severity stage by SVM has been extended to the analysis of biomarkers classically used in inflammatory and infectious pathologies [27,28] but has not provided a consensus or an infallible score to predict the level of severity. To our knowledge, this study is the first to show the SVM method being used for predicting the severity stage of COVID-19 with pulmonary involvement.

An important point to consider before generalizing the use of prediction scores based on this panel of blood biological OS analyses is the dosage method and standardization of reference values for each analysis and between each laboratory.

Conclusions

In patients with COVID-19 infection, moderate to severe symptoms were correlated with a lowered Zn level, a lowered plasma thiol level, an increased hs-CRP level, and an increased Cu/Zn ratio in a panel of 10 OS biomarkers not requiring complex preanalysis (Zn, Cu, Cu/Zn ratio, selenium, uric acid, hs-CRP, oxidized LDL, glutathione peroxidase, glutathione reductase, and thiols). Applying a machine learning method (an SVM) to the results of this same panel to evaluate the OS level in a retrospective cohort allowed prediction of the level of severity of pulmonary damage in hospitalized patients with COVID-19. Since this panel does not require a complex preanalytical method, it can be used and studied in other pathologies associated with OS, such as infectious pathologies or chronic diseases. Intensive care teams are accustomed to monitoring and managing major inflammatory symptoms, and although it is likely that they need to augment their practices to manage OS and the supply of antioxidants, the biological exploration of OS requires highly complex preanalytical and analytical processes that, to date, are highly limiting factors in the development and dissemination of relevant OS assessments. The fragility and thermolability of certain biomarkers means that any redox reactions that might take place in the sample must be halted; thus, the mandatory use of −80 °C conditions throughout the preanalytical phase makes analysis complex and in some cases even technically unfeasible. We have therefore developed an OS panel with 10 biomarkers that overcomes all these constraints without compromising scientific relevance by performing an assessment of OS based on a series of biomarkers that do not require complex preanalysis (simple freezing at −20 °C); these biomarkers were selected with a machine learning approach in 2009 [23]. The results from this panel showed real significance in estimating OS with robust and reproducible results. Iterative training of this SVM model with a biomarker panel in a larger cohort of patients with OS would give it more power and precision. By determining the ideal weighting of each of the biomarkers with a scientifically validated algorithm, we could obtain a score to estimate the level of OS. A Belgian intensive care team assessed total antioxidant capacity with an electrochemical method (eg, the Total Antioxidant Power technology) rather than by the analysis of biological samples (blood, saliva, or urine), finding that it had potential as a less expensive and much faster alternative to the individual analysis of biomarkers linked to pro-oxidants [30]. This would allow intensive care teams to monitor and respond in real time to systemic OS in their patients.

The World Health Organization has already announced the very probable risk of the appearance of a new epidemic caused by new types of severe acute respiratory syndrome or by the "humanization" of the avian influenza A (H5N1) virus in the years to come [38]. Given the contagiousness and the risk of decompensation, particularly respiratory decompensation, or even death from these viral infections, considering systemic OS in addition to the inflammatory syndrome could be a determining factor in the effectiveness of care in the future. The early and systematic use of machine learning models such as SVMs in clinical, radiological, and biological assessments at the start of an epidemic could provide a better understanding of the

XSL•FO
RenderX

functioning and mechanisms of novel pathogens and reduce human loss of life. Sharing of results between different care centers with dedicated databases would accelerate learning and improve the accuracy of these statistical tools.

From a public health perspective, individuals are subject to a range of exposures through the environment (air pollution, solar radiation, and temperature variations) and lifestyle (lack of sleep, psychological stress, tobacco, and poor diet); this is summarized in the term "exposome." This directly impacts the OS level and therefore the body's ability to defend itself in the event of infection. Taking action to maintain a healthy environment and lifestyle would help mitigate the severity of diseases such as COVID-19 [39].

Acknowledgments

This study was funded by the Infirmerie Protestante of Lyon. Preanalytical and analytical processes were performed by Biomnis. EpiComet SAS provided scientific writing assistance in the preparation and development of the manuscript in accordance with the Good Publication Practice guidelines.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Individual notice information.
[DOCX File, 51 KB - formative_v9i1e66509_app1.docx]

Multimedia Appendix 2
Support vector machine results from 28 patients, including observed oxidative stress biomarkers and the probability of patients being assigned severity grades 0, 1, or 2.
[DOCX File, 23 KB - formative_v9i1e66509_app2.docx]

References

1. Forman HJ, Zhang H. Targeting oxidative stress in disease: promise and limitations of antioxidant therapy. Nat Rev Drug Discov 2021 Sep;20(9):689-709. [doi: 10.1038/s41573-021-00233-1] [Medline: 34194012]
2. Pizzino G, Irrera N, Cucinotta M, et al. Oxidative stress: harms and benefits for human health. Oxid Med Cell Longev 2017;2017:8416763. [doi: 10.1155/2017/8416763] [Medline: 28819546]
3. Ji LL, Yeo D. Oxidative stress: an evolving definition. Fac Rev 2021;10:13. [doi: 10.12703/r/10-13] [Medline: 33659931]
4. Jelic MD, Mandic AD, Maricic SM, Srdjenovic BU. Oxidative stress and its role in cancer. J Cancer Res Ther 2021;17(1):22-28. [doi: 10.4103/jcrt.JCRT_862_16] [Medline: 33723127]
5. Rezzani R, Franco C. Liver, oxidative stress and metabolic syndromes. Nutrients 2021 Jan 21;13(2):301. [doi: 10.3390/nu13020301] [Medline: 33494242]
6. Li M, Xin S, Gu R, et al. Novel diagnostic biomarkers related to oxidative stress and macrophage ferroptosis in atherosclerosis. Oxid Med Cell Longev 2022;2022:8917947. [doi: 10.1155/2022/8917947] [Medline: 36035208]
7. Lorenzon Dos Santos J, Quadros AD, Weschenfelder C, Garofallo SB, Marcadenti A. Oxidative stress biomarkers, nut-related antioxidants, and cardiovascular disease. Nutrients 2020 Mar 3;12(3):682. [doi: 10.3390/nu12030682] [Medline: 32138220]
8. Gtif I, Abdelhedi R, Ouarda W, et al. Oxidative stress markers-driven prognostic model to predict post-discharge mortality in heart failure with reduced ejection fraction. Front Cardiovasc Med 2022;9:1017673. [doi: 10.3389/fcvm.2022.1017673] [Medline: 36419488]
9. Ton AMM, Campagnaro BP, Alves GA, et al. Oxidative stress and dementia in Alzheimer's patients: effects of synbiotic supplementation. Oxid Med Cell Longev 2020;2020:2638703. [doi: 10.1155/2020/2638703] [Medline: 32411323]
10. Ionescu-Tucker A, Cotman CW. Emerging roles of oxidative stress in brain aging and Alzheimer's disease. Neurobiol Aging 2021 Nov;107:86-95. [doi: 10.1016/j.neurobiolaging.2021.07.014] [Medline: 34416493]
11. Salim S. Oxidative stress: a potential link between emotional wellbeing and immune response. Curr Opin Pharmacol 2016 Aug;29:70-76. [doi: 10.1016/j.coph.2016.06.006] [Medline: 27400336]
12. El Assar M, Álvarez-Bustos A, Sosa P, Angulo J, Rodríguez-Mañas L. Effect of physical activity/exercise on oxidative stress and inflammation in muscle and vascular aging. Int J Mol Sci 2022 Aug 5;23(15):8713. [doi: 10.3390/ijms23158713] [Medline: 35955849]

XSL•FO
RenderX

13. Ávila-Escalante ML, Coop-Gamas F, Cervantes-Rodríguez M, Méndez-Iturbide D, Aranda-González II. The effect of diet on oxidative stress and metabolic diseases-clinically controlled trials. J Food Biochem 2020 May;44(5):e13191. [doi: 10.1111/jfbc.13191] [Medline: 32160647]
14. Gangwar RS, Bevan GH, Palanivel R, Das L, Rajagopalan S. Oxidative stress pathways of air pollution mediated toxicity: recent insights. Redox Biol 2020 Jul;34:101545. [doi: 10.1016/j.redox.2020.101545] [Medline: 32505541]
15. Chernyak BV, Popova EN, Prikhodko AS, Grebenchikov OA, Zinovkina LA, Zinovkin RA. COVID-19 and oxidative stress. Biochemistry (Mosc) 2020 Dec;85(12):1543-1553. [doi: 10.1134/S0006297920120068] [Medline: 33705292]
16. Huang C, Wang Y, Li X, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 2020 Feb 15;395(10223):497-506. [doi: 10.1016/S0140-6736(20)30183-5] [Medline: 31986264]
17. Suhail S, Zajac J, Fossum C, et al. Role of oxidative stress on SARS-CoV (SARS) and SARS-CoV-2 (COVID-19) infection: a review. Protein J 2020 Dec;39(6):644-656. [doi: 10.1007/s10930-020-09935-8] [Medline: 33106987]
18. Paul AK, Hossain MK, Mahboob T, et al. Does oxidative stress management help alleviation of COVID-19 symptoms in patients experiencing diabetes? Nutrients 2022 Jan 13;14(2):321. [doi: 10.3390/nu14020321] [Medline: 35057501]
19. Ducastel M, Chenevier-Gobeaux C, Ballaa Y, et al. Oxidative stress and inflammatory biomarkers for the prediction of severity and ICU admission in unselected patients hospitalized with COVID-19. Int J Mol Sci 2021 Jul 12;22(14):7462. [doi: 10.3390/ijms22147462] [Medline: 34299080]
20. Manzanares W, Dhaliwal R, Jiang X, Murch L, Heyland DK. Antioxidant micronutrients in the critically ill: a systematic review and meta-analysis. Crit Care 2012 Dec 12;16(2):R66. [doi: 10.1186/cc11316] [Medline: 22534505]
21. Heyland DK, Dhaliwal R, Suchner U, Berger MM. Antioxidant nutrients: a systematic review of trace elements and vitamins in the critically ill patient. Intensive Care Med 2005 Mar;31(3):327-337. [doi: 10.1007/s00134-004-2522-z] [Medline: 15605227]
22. Kalem AK, Kayaaslan B, Neselioglu S, et al. A useful and sensitive marker in the prediction of COVID-19 and disease severity: thiol. Free Radic Biol Med 2021 Apr;166:11-17. [doi: 10.1016/j.freeradbiomed.2021.02.009] [Medline: 33588050]
23. de la Villehuchet AM, Brack M, Dreyfus G, et al. A machine-learning approach to the prediction of oxidative stress in chronic inflammatory disease. Redox Rep 2009;14(1):23-33. [doi: 10.1179/135100009X392449] [Medline: 19161675]
24. Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995 Sep;20(3):273-297. [doi: 10.1007/BF00994018]
25. Vapnik VN. The support vector method. In: Gerstner W, Germond A, Hasler M, Nicoud JD, editors. Artificial Neural Networks — ICANN'97 ICANN 1997 Lecture Notes in Computer Science, Vol 1327: Springer; 1997. [doi: 10.1007/BFb0020166]
26. Hsu CW, Chang CC, Lin CJ. A practical guide to support vector classification. National Taiwan University. URL: https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf [accessed 2025-03-27]
27. Platt JC. Fast training of support vector machines using sequential minimal optimization. In: Burges CJC, Schölkopf B, Smola AJ, editors. Advances in Kernel Methods: Support Vector Learning: MIT Press; 1998:41-65. [doi: 10.7551/mitpress/1130.003.0016]
28. Jindal M, Lehl SS, Jaswal S, Bhardwaj N, Gupta M. Total antioxidant status and oxidative stress in patients with COVID-19 infection. J Assoc Physicians India 2024 Aug;72(8):36-39. [doi: 10.59556/japi.72.0575] [Medline: 39163060]
29. Djordjevic J, Ignjatovic V, Vukomanovic V, et al. Laboratory puzzle of oxidative stress, parameters of hemostasis and inflammation in hospitalized patients with COVID-19. Biomedicines 2024 Mar 13;12(3):636. [doi: 10.3390/biomedicines12030636] [Medline: 38540248]
30. Pincemail J, Rousseau AF, Kaux JF, et al. A pilot study on oxidative stress during the recovery phase in critical COVID-19 patients in A rehabilitation facility: potential utility of the PAOT® technology for assessing total anti-oxidative capacity. Biomedicines 2023 Apr 28;11(5):1308. [doi: 10.3390/biomedicines11051308] [Medline: 37238982]
31. Deblanc C, Robert F, Pinard T, et al. Pre-infection of pigs with Mycoplasma hyopneumoniae induces oxidative stress that influences outcomes of a subsequent infection with a swine influenza virus of H1N1 subtype. Vet Microbiol 2013 Mar 23;162(2-4):643-651. [doi: 10.1016/j.vetmic.2012.11.028] [Medline: 23266108]
32. Vlaming-van Eijk LE, Bulthuis MLC, van der Gun BTF, et al. Systemic oxidative stress associates with the development of post-COVID-19 syndrome in non-hospitalized individuals. Redox Biol 2024 Oct;76:103310. [doi: 10.1016/j.redox.2024.103310] [Medline: 39163767]
33. Principes de nutrition en reanimation. Société française d'anesthésie et de réanimation. URL: https://sfar.org/wp-content/uploads/2018/10/14-Principes-de-nutrition-en-reanimation.pdf [accessed 2025-04-09]
34. Jones NE, Heyland DK. Pharmaconutrition: a new emerging paradigm. Curr Opin Gastroenterol 2008 Mar;24(2):215-222. [doi: 10.1097/MOG.0b013e3282f4cdd8] [Medline: 18301274]
35. Fuhrman MP. Antioxidant supplementation in critical illness: what do we know? Nutrition 2000 Jun;16(6):470-471. [doi: 10.1016/s0899-9007(00)00360-9] [Medline: 10869911]
36. Forceville X, Vitoux D, Gauzit R, Combes A, Lahilaire P, Chappuis P. Selenium, systemic immune response syndrome, sepsis, and outcome in critically ill patients. Crit Care Med 1998 Sep;26(9):1536-1544. [doi: 10.1097/00003246-199809000-00021] [Medline: 9751590]

XSL•FO
RenderX

37. Gudivada KK, Kumar A, Shariff M, et al. Antioxidant micronutrient supplementation in critically ill adults: a systematic review with meta-analysis and trial sequential analysis. Clin Nutr 2021 Mar;40(3):740-750. [doi: 10.1016/j.clnu.2020.06.033] [Medline: 32723509]
38. Avian influenza A(H5N1) - Cambodia. World Health Organization. URL: https://www.who.int/emergencies/disease-outbreak-news/item/2024-DON533 [accessed 2025-03-27]
39. Morales JS, Valenzuela PL, Castillo-García A, et al. The exposome and immune health in times of the COVID-19 pandemic. Nutrients 2021 Dec 22;14(1):24. [doi: 10.3390/nu14010024] [Medline: 35010900]

Abbreviations

CRP: C-reactive protein
Cu: copper
hs-CRP: high-sensitivity C-reactive protein
ICU: intensive care unit
OS: oxidative stress
ROX index: oxygen saturation/respiration rate index
SMO: sequential minimal optimization
SpO_2: oxygen saturation
SVM: support vector machine
SVR: support vector regression
Zn: zinc

Edited by A Mavragani; submitted 15.09.24; peer-reviewed by A Akomolafe, N Idowu; revised version received 19.02.25; accepted 20.02.25; published 11.04.25.

Please cite as:
Raspado O, Brack M, Brack O, Vivancos M, Esparcieux A, Cart-Tanneur E, Aouifi A
Oxidative Stress Markers and Prediction of Severity With a Machine Learning Approach in Hospitalized Patients With COVID-19 and Severe Lung Disease: Observational, Retrospective, Single-Center Feasibility Study
JMIR Form Res 2025;9:e66509
URL: https://formative.jmir.org/2025/1/e66509
doi:10.2196/66509

© Olivier Raspado, Michel Brack, Olivier Brack, Mélanie VivancosI, Aurélie Esparcieux, Emmanuelle Cart-Tanneur, Abdellah Aouifi. Originally published in JMIR Formative Research (https://formative.jmir.org), 11.4.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Association of Social Media Recruitment and Depression Among Racially and Ethnically Diverse Metabolic and Bariatric Surgery Candidates: Prospective Cohort Study

Jackson M Francis[1], MPH; Sitapriya S Neti[2,3], MPH; Dhatri Polavarapu[2,3], MPH; Folefac Atem[2,3], PhD; Luyu Xie[1], PharmD, PhD; Olivia Kapera[4], MPH; Matthew S Mathew[1], MS; Elisa Marroquin[5], MS, PhD; Carrie McAdams[6], MD, PhD; Jeffrey Schellinger[7], MCH, RD, CSOWM, LD; Sophia Ngenge[1], MPH; Sachin Kukreja[1], MD; Benjamin E Schneider[8], MD; Jaime P Almandoz[7], MBA, MD; Sarah E Messiah[1,9], MPH, PhD

[1]Peter O'Donnell Jr. School of Public Health, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd, Dallas, TX, United States
[2]School of Public Health, University of Texas Health Science Center at Houston (UTHealth), Dallas, TX, United States
[3]Department of Biostatistics & Data Science, School of Public Health, University of Texas Health Science Center at Houston, Dallas, TX, United States
[4]School of Public Health, The University of Texas at Austin, Austin, TX, United States
[5]Department of Nutritional Sciences, Texas Christian University, Fort Worth, TX, United States
[6]Department of Psychiatry, University of Texas Southwestern Medical Center, Dallas, TX, United States
[7]Division of Endocrinology, Department of Internal Medicine, University of Texas Southwestern Medical Center, Dallas, TX, United States
[8]Department of Surgery, University of Texas Southwestern Medical Center, Dallas, TX, United States
[9]Department of Pediatrics, University of Texas Southwestern Medical Center, Dallas, TX, United States

Corresponding Author:
Jackson M Francis, MPH
Peter O'Donnell Jr. School of Public Health, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd, Dallas, TX, United States

Abstract

Background: Due to the widespread use of social media and the internet in today's connected world, obesity and depression rates are increasing concurrently on a global scale. This study investigated the complex dynamics involving social media recruitment for scientific research, race, ethnicity, and depression among metabolic and bariatric surgery (MBS) candidates.

Objective: This study aimed to determine (1) the association between social media recruitment and depression among MBS candidates and (2) racial and ethnic differences in social media recruitment engagement.

Methods: The analysis included data from 380 adult MBS candidates enrolled in a prospective cohort study from July 2019 to December 2022. Race and ethnicity, recruitment method (social media: yes or no), and depression status were evaluated using χ^2 tests and logistic regression models. Age, sex, and ethnicity were adjusted in multivariable logistic regression models.

Results: The mean age of the candidates was 47.35 (SD 11.6) years, ranging from 18 to 78 years. Participants recruited through social media (n=41, 38.32%) were more likely to report past or current episodes of depression compared to nonsocial media–recruited participants (n=74, 27.11%; *P*=.03), with a 67% increased likelihood of depression (odds ratio [OR] 1.67, 95% CI 1.04 - 2.68, *P*=.03). Further analysis showed that participants with a history of depression who were below the mean sample age were 2.26 times more likely to be recruited via social media (adjusted OR [aOR] 2.26, 95% CI 1.03 - 4.95; *P*=.04) compared to those above the mean age. Hispanic (n=26, 38.81%) and non-Hispanic White (n=53, 35.10%) participants were significantly more likely to be recruited via social media than non-Hispanic Black (n=27, 18.37%) participants (*P*<.001). After adjusting for covariates, non-Hispanic Black participants were 60% less likely than non-Hispanic White participants to be recruited via social media (aOR 0.40, 95% CI 0.22 - 0.71; *P*=.002).

Conclusions: We found that individuals recruited through social media channels, especially younger participants, were more likely to report past or current episodes of depression compared to those recruited through nonsocial media. The study also showed that non-Hispanic Black individuals are less likely to engage in social media recruitment for scientific research versus other racial and ethnic groups. Future mental health–related studies should consider strategies to mitigate potential biases introduced by recruitment methods to ensure the validity and generalizability of research findings.

(JMIR Form Res 2025;9:e58916) doi:10.2196/58916

XSL•FO
RenderX

KEYWORDS

social media; social media recruitment; depression; depressed; major depressive disorder; MDD; depressive disorder; depressive; race; racial; racial differences; ethnic; ethnic differences; ethnicity; metabolic surgery; bariatric surgery; obesity; obese; online health; ethics; mental health; eHealth; internet; digital health; digital mental health; online interventions; public engagement; public health

Introduction

Over the past 2 decades, obesity (ie, BMI ≥30 kg/m^2) prevalence among adults in the United States has risen from 30.5% to 42.4% [1], and has nearly tripled worldwide [2]. Further, the prevalence of severe obesity (BMI ≥40 kg/m^2) has doubled from 4.7% to 9.2% [3]. Non-Hispanic Black (NHB) adults are disproportionately impacted, with an age-adjusted obesity prevalence of 49.9%, followed by Hispanic (45.6%), non-Hispanic White (NHW) (41.4%), and non-Hispanic Asian (16.1%) individuals [3].

Various obesity treatment options, including diet control, exercise, cognitive and family-based behavioral therapy, pharmacological therapy, and metabolic and bariatric surgery (MBS) have been explored to address this critical issue. Of these, MBS is a safe and effective treatment option, particularly for individuals with severe obesity who are resistant to other interventions [4]. A 2014 meta-analysis of 164 studies involving 160,000 patients found that BMI decreased by at least 11.8 kg/m² one-year post MBS, with sustained weight loss observed five years post surgery [5]. Most individuals who undergo MBS experience a reduction or complete elimination of obesity-related health issues, including high blood pressure, heart disease, diabetes, asthma, and osteoarthritis [6,7]. Additionally, long-term outcomes of MBS in adolescents aged 16 - 21 years demonstrated sustained weight reduction and improvement in comorbidities, including reduced depression prevalence over a decade post surgery [7]. Given the increasing prevalence of obesity and the proven safety, effectiveness, and long-term success of MBS, there has been an increase in the number of MBS procedures being performed, resulting in a parallel increase in MBS-related research [8,9]. However, the COVID-19 pandemic led to a temporary decrease in MBS surgeries due to delays in elective procedures [10]. Research studies conducted during the pandemic faced unique challenges, prompting a shift toward innovative recruitment methods. Researchers have increasingly used social media for participant recruitment, survey administration, and trend analysis. Although literature reviews underscore the efficacy of social media as a recruitment methodology, its use comes with challenges and drawbacks [11,12].

This study acknowledges the dual nature of social media, which, while a potent recruitment tool, has also been implicated in mental health and body image concerns. Celebrities and influencers on social media platforms have contributed to rising negative body image perceptions [13], which have been associated with higher depressive symptoms, loss of confidence, loneliness, and risk of developing eating disorders [13,14]. Studies have shown that viewing images that have been heavily edited can negatively impact an individual's body image and that editing selfies—regardless of whether they are posted—was linked to self-objectification and dissatisfaction with facial features [15]. Additionally, overall time spent on image-based social media platforms has been associated with weight-bias internalization [16]. A meta-analysis examining the relationship between body image and eating disorders found that both body checking and body image avoidance were associated with eating disorder pathology [17].

Acknowledging the gaps in our understanding of social media's impact on mental health among individuals with severe obesity and who are considering MBS, this study aimed to determine (1) the association between social media recruitment and depression among MBS candidates and (2) racial and ethnic differences in social media recruitment engagement. It was hypothesized that depression prevalence would be higher in individuals recruited through social media and that no significant racial and ethnic differences would exist between social media– and nonsocial–based media recruitment.

Methods

Study Design and Eligibility Criteria

The Bariatric Health Study (NIH/NIMHD 5R01MD011686) is an ongoing prospective cohort research study enrolling participants who qualify and have been referred for MBS based on established medical criteria. The American Society for Metabolic and Bariatric Surgery recommends MBS for individuals with a BMI >35 kg/m^2, regardless of the presence or absence of comorbidities; for individuals with type 2 diabetes and a BMI >30 kg/m^2; or for individuals with a BMI between 30 - 34.9 kg/m^2 who have not achieved adequate weight loss or improvement in comorbidities with nonsurgical obesity treatment methods [18]. Eligibility criteria mandated that participants be at least 18 years old, medically eligible for MBS, and referred for surgery by a clinical provider. To manage and collect study data securely and in compliance with HIPAA (Health Insurance Portability and Accountability Act of 1996) regulations, we used REDCap (Research Electronic Data Capture; Vanderbilt University), a web-based database management platform [19,20].

Ethical Considerations

The study was approved by the University of Texas Health System Institutional Review Board (IRB# HSC-SPH-18 - 0850) and the Committee for the Protection of Human Subjects. Informed consent was obtained from all study participants. Participant data are stored in a HIPAA-compliant, password-protected, encrypted database with limited access. The data used for this analysis were deidentified. Study participants were compensated with a US $50 gift card upon completion of the initial Mini International Neuropsychiatric Interview (MINI) and surveys. Upon surgery completion, participants were compensated at 6, 12, and 24 months with US

XSL•FO
RenderX

$100, US $125, and US $150 gift cards for completing the additional MINI interviews and surveys.

Participant Recruitment

Recruitment for this study was open from August 2019 to October 2022. Recruitment sites for this prospective cohort study consisted of MBS specialty clinics in the North Texas region. These included Minimally Invasive Surgical Associates, Lee Bariatrics Clinic, City Hospital at White Rock, Dallas, TX, and the UT Southwestern Weight Wellness Subspecialties Clinic, Dallas, TX. Before the COVID-19 pandemic, eligible MBS patients were recruited by research coordinators during mandated educational seminars held by MBS team members before surgery. Research coordinators provided interested candidates with information and a sign-up sheet at these seminars. In addition, the institutional review board–approved brochures and flyers were placed in the MBS clinic waiting rooms. Interested participants completed a web-based eligibility screener and were contacted by research coordinators to complete informed consent and baseline measurements. However, these educational seminars transitioned to virtual platforms due to the pandemic. Additionally, social media channels were added as a recruitment source to meet our recruitment targets, as elective surgeries at our academic center were paused on several occasions due to surges in COVID-19. Specifically, web-based social media recruitment was conducted through Facebook and Instagram and was managed by the study's research coordinators. Targeted advertisements on these platforms directed outreach toward individuals meeting our inclusion criteria. Interested individuals completed the study's online eligibility screener to ensure that the inclusion criteria were met. Following this, informed consent was completed, and research coordinators performed baseline measurements. To sustain study engagement, biweekly informative videos and posts addressing topics related to MBS were created.

Main Outcome

The primary outcome measure was self-reported depression, measured by the MINI [21], a validated tool for evaluating major psychiatric disorders, comprising dichotomous questions with follow-up questions for positive responses [21]. Previous studies have demonstrated that the MINI accurately diagnoses psychiatric conditions [22]. The interview primarily features "yes" or "no" questions, with additional clarifications provided to participants when necessary. Following the completion of questions for each psychiatric disorder, research coordinators (interviewers) tallied the "yes" and "no" responses to identify if the interviewee screened positively for a variety of potential psychiatric disorders. In this study, we assessed current, past, and recurrent major depressive episodes, posttraumatic stress disorder, agoraphobia, social anxiety disorder, general anxiety disorder, and panic disorder. Following the initial interview, research coordinators determined the presence or absence of depression based on participants' responses. Participants who screened positive for a current, past, or recurrent major depressive episode during the MINI interview were classified as having a history of depression. Conversely, if none of these conditions were selected (current, past, or recurrent), participants were categorized as not having a history of depression.

Main Exposure

We assessed each participant's recruitment method through the enrollment survey. Participants responded to the question, "How did you learn about this study?" They could choose from options such as "Lee Bariatrics, Texas Presbyterian Hospital, UT Southwestern Medical Center, Minimally Invasive Surgical Associates, UTSW Weight Wellness Program, City Hospital at White Rock, Flyer, Brochure or Pamphlet, Social media (ie, Facebook, Instagram), or Other." We classified participants as positive for social media recruitment if they chose "Social media (ie, Facebook, Instagram)" or mentioned social media recruitment in the "Other" section. We categorized participants as negative for social media recruitment if they selected other options.

Covariates

Covariates including age, sex, race and ethnicity, BMI, education level, and perceived financial well-being were obtained through self-report. Age was treated as a continuous variable, representing the participants' age in years. Sex was a categorical variable, categorized as male, female, or other. Race and ethnicity were also categorical variables, encompassing diverse categories such as NHW, NHB, Hispanic, and Other. BMI was calculated via self-reported height and weight and verified via medical records provided by participants when available. Education was a categorical variable, categorized as none, elementary school, high school, college, and graduate or professional degree. In our survey, we did not have a direct measure of income. However, as a proxy measure, we asked the question, "In the last two weeks, have you had enough money to meet your needs?" This variable was categorical, with answers being "not at all, a little, moderately, mostly, and completely." These covariates served as essential demographic factors for our analysis, providing a comprehensive understanding of their potential influence on the investigated variables.

Statistical Analysis

We analyzed descriptive variables, including sex, age, race and ethnicity (ie, NHB, NHW, Hispanic), BMI, and history of depression, comprising both categorical and continuous variables. To explore the associations between recruitment methods, social media or nonsocial media, and depression history (yes/no), we conducted Pearson χ^2 tests. Crude odds ratios were calculated using univariable logistic regression models to assess the relationship between recruitment method and depression history. Furthermore, we performed multivariable logistic regression by adjusting for age, race and ethnicity, and sex to calculate the adjusted odds ratios for depression in relation to the recruitment method used. There were less than 3% missing data for demographic variables (ie, age, sex, and ethnicity); no data was missing for the outcome variable. We performed all analyses using STATA (v.17.1, Stata Corp LP), with statistical significance set at a *P* value below 5%.

XSL•FO
RenderX

Results

Participant Characteristics

Participant characteristics (N=380) stratified by recruitment method, social media (n=107), and nonsocial media (n=273) are presented in Table 1. Participants recruited through social media had a mean age of 47.27 (SD 1.02) years, whereas those in the nonsocial media group had a mean age of 47.38 (SD 0.73) years (*P*=.93). Sex distribution in the social media group was 9 (8.41%) males and 98 (91.59%) females, while the nonsocial media group included 41 (15.02%) males and 232 (84.98%) females, *P*=.09. The mean BMI for the social media group was 44.92 (SD 1.05) kg/m^2 and the nonsocial media group was 45.68 (SD 0.64) kg/m^2 (*P*=.54). Regarding racial and ethnic backgrounds, the social media group had 49.53% (n=53) of participants identifying as NHW, 25.23% (n=27) as NHB, 24.3% (n=26) as Hispanic, and 0.93% (n=1) as other. In the nonsocial media group, 35.9% (n=98) were NHW, 43.96% (n=120) were NHB, 15.02% (n=41) were Hispanic, and 5.13% identified as Other (*P*<.001).

Table . Baseline characteristics of metabolic and bariatric surgery candidates recruited via social media and traditional methods in a prospective cohort study (N=380, 2019 - 2022).

Variables	Participants (N=380)		*P* value
	Social media (n=107)	Nonsocial media (n=273)	
Age (years), mean (SD)	47.27 (1.02)	47.38 (0.73)	.93
Sex, n (%)			.09
Males (n=50, 13.16%)	9 (8.41)	41 (15.02)	
Female (n=330, 86.84%)	98 (91.59)	232 (84.98)	
BMI (kg/m^2), mean (SD)	44.92 (1.04)	45.68 (0.64)	.54
Race/ethnicity, n (%)			<.001[a]
Non-Hispanic White (n=151)	53 (49.53)	98 (35.9)	
Non-Hispanic Black (n=147)	27 (25.23)	120 (43.96)	
Hispanic (n=67)	26 (24.3)	41 (15.02)	
Other (n=15)	1 (0.93)	14 (5.13)	

[a]Significant.

Depression Prevalence by Recruitment Method

Table 2 presents the prevalence of depression categorized by recruitment method and race and ethnicity among all participants (N=380). Of the participants, 115 (30.26%) had a history of depression, while 265 (69.74%) had no history of depression. Our findings showed a difference between the prevalence of depression and the recruitment methods (*P*=.03).

In the nonsocial media group (n=273, 71.84%), the prevalence rates of depression (n=74, 27.11%) varied by race and ethnicity. In the NHW subgroup (n=98), 26 (26.53%) participants had a history of depression, while 72 (73.47%) did not. In the NHB subgroup (n=120), 28 (23.33%) participants had a history of depression, whereas 92 (76.67%) had no history. Among the Hispanic subgroup (n=41), 15 (36.59%) participants had a history of depression, while 26 (63.41%) had no history. In the "Other" subgroup (n=14), 5 (35.71%) participants had a history of depression, while 9 (64.29%) had no history.

Within the social media group (n=107, 28.16%), subgroups based on race and ethnicity exhibited varied depression prevalence rates (n=41, 38.32%). In the NHW subgroup (n=53), 22 (41.51%) had a history of depression, while 31 (58.49%) had no history. Among the NHB subgroup (n=27), 7 (25.93%) had a history of depression, while 20 (74.07%) had no history. In the Hispanic subgroup (n=26), 11 (42.31%) individuals had a history of depression, whereas 15 (57.69%) had no history. A single participant in the "Other" subgroup had a history of depression.

XSL•FO
RenderX

Table . Prevalence of depression among metabolic and bariatric surgery candidates by recruitment method and race and ethnicity in a prospective cohort study (2019 - 2022).

Variable and group	Participants (N=380), n (%)		*P* value
	History of depression	No history of depression	
Overall prevalence	115 (30.26)	265 (69.74)	.03[a]
Social media group (n=107)	41 (38.32)	66 (61.68)	—[b]
NHW[c] (n=53)	22 (41.51)	31 (58.49)	
NHB[d] (n=27)	7 (25.93)	20 (74.07)	
Hispanic (n=26)	11 (42.31)	15 (57.69)	
Other (n=1)	1 (100)	0 (0)	
Nonsocial media group (n=273)	74 (27.11)	199 (72.89)	—
NHW (n=98)	26 (26.53)	72 (73.47)	
NHB (n=120)	28 (23.33)	92 (76.67)	
Hispanic (n=41)	15 (36.59)	26 (63.41)	
Other (n=14)	5 (35.71)	9 (64.29)	

[a]There was a significant difference between depression prevalence in the social media group versus nonsocial media group.

[b]Not applicable.

[c]NHW: non-Hispanic White.

[d]NHB: non-Hispanic Black.

Recruitment Method by Race and Ethnicity

Table 3 presents the odds of recruitment via social media, adjusting for age, sex, and BMI, stratified by race and ethnicity. In the unadjusted model, representing the crude odds, Hispanic participants had odds of 1.17 (95% CI 0.24 - 0.71, *P*=.60) for being recruited via social media, while NHB participants had odds of 0.42 (95% CI 0.65 - 2.12, *P*=.001), with NHW participants serving as the reference group.

In the adjusted model, adjusted for age, sex, BMI, education, and financial status, Hispanic participants had odds of 1.22 (95% CI 0.63 - 2.37, *P*=.55) for being recruited via social media, while NHB participants had odds of 0.40 (95% CI 0.22 - 0.72, *P*=.002). These findings indicate that after adjusting for demographic variables, NHB participants demonstrated significantly lower odds of recruitment via social media compared to NHW participants.

Table . Odds of being recruited via social media by race and ethnicity among metabolic and bariatric surgery candidates, adjusting for age, sex, BMI, education, and financial status.

Variables	Odds ratio (95% CI)	*P* value
Race/ethnicity		
Model 1[a] (crude odds ratio)		
Hispanic	1.17 (0.24 - 0.71)	.60
NHB[b]	0.42 (0.65 - 2.12)	.001[c]
NHW[d]	1.0 (ref)	—[e]
Model 2[f] (adjusted odds ratio)		
Hispanic	1.22 (0.63 - 2.37)	.55
NHB	0.40 (0.22 - 0.72)	.002[c]
NHW	1.0 (ref)	—

[a]Model 1: crude (social media as primary outcome).

[b]NHB: non-Hispanic Black.

[c]Significant.

[d]NHW: non-Hispanic White.

[e]Not applicable.

[f]Model 2: adjusted for age, sex, BMI, education, and financial status (social media as primary outcome).

XSL•FO
RenderX

History of Depression and Social Media Recruitment

Univariate logistic models showed that individuals with a history of depression were 67% more likely (OR 1.67, 95% CI 1.04 - 2.68, *P*=.03) to be recruited via social media than those with no history of depression (Table 4). However, after adjustment for age, sex, race and ethnicity, BMI, education, and financial status, the association between depression history and recruitment method was no longer significant. This indicated that one or more of these covariates influenced the relationship between our outcome variables, recruitment method, and depression history, warranting further subanalysis.

Table . Odds of depression for social media recruits, adjusting for age among metabolic and bariatric surgery candidates, race and ethnicity, sex, BMI, education, and financial status.

Variables	Odds ratio (95% CI)	*P* value
Depression history		
Model 1[a] (crude odds ratio)		
Depression	1.67 (1.04 - 2.68)	.03[b]
No depression	1.0 (ref)	__[c]
Model 2[d] (adjusted odds ratio)		
Depression	1.51 (0.88 - 2.62)	.13
No depression	1.0 (ref)	—

[a]Model 1: crude (social media as primary outcome).

[b]Significant.

[c]Not applicable.

[d]Model 2: adjusted for age, race/ethnicity, sex, BMI, education, and financial status (social media as primary outcome).

Age, Depression, and Recruitment Method

After adjusting for sex, race and ethnicity, BMI, education, and financial status, participants with a history of depression who were below the mean age were over twice as likely (OR 2.26, 95%CI, 1.03 - 4.95) to be recruited via social media compared to those with no history of depression (Table 5). However, with the same adjustments, no significant association was found among participants above the mean sample age.

Table . Relationship between age and recruitment method on odds of depression among metabolic and bariatric surgery candidates, after adjusting for sex, race and ethnicity, BMI, education, and financial status.

Variables	Adjusted odds ratio (95% CI)	*P* value
Depression history		
Model 1[a]		
Depression	2.26 (1.03 - 4.95)	.04[b]
No depression	1.0 (ref)	__[c]
Model 2[d]		
Depression	1.13 (0.47 - 2.73)	.79
No depression	1.0 (ref)	—

[a]Model 1: age below mean, adjusted for sex, race/ethnicity, BMI, education, and financial status (social media as primary outcome).

[b]Significant.

[c]Not applicable.

[d]Model 2: age above mean, adjusted for sex, race/ethnicity, BMI, education, and financial status (social media as primary outcome).

Discussion

Principal Findings

This study aimed to investigate the relationship between social media recruitment, depression, race, and ethnicity among individuals seeking MBS. Our analysis of 380 MBS candidates showed that participants recruited via social media exhibited a 67% higher likelihood of depression history compared to nonsocial media recruits. Furthermore, participants with a history of depression below the mean sample age were over twice as likely to be recruited via social media. Our findings also revealed significant disparities in social media recruitment among racial and ethnic groups. Hispanic and NHW participants were more likely to be recruited through social media than NHB. Specifically, adjusted models demonstrated a 60% lower likelihood of social media recruitment among NHB participants compared to NHW participants. These findings underscore the

XSL•FO
RenderX

significance of exploring the interplay between social media, mental health disorders, and demographic factors in this population.

Our first key finding reveals a significant association between social media recruitment and a history of depression, particularly among younger individuals below the sample mean age, who were twice as likely to report such a history. Extensive studies have explored the relationship between social media usage and depression, emphasizing the heightened risk of depression, anxiety, and psychological distress associated with addictive or problematic platform use [23]. A systematic examination of 11 studies focusing on young individuals corroborates this, revealing a modest yet statistically significant association between social media usage and symptoms of depression [24]. The observation that younger individuals with a history of depression are more likely to be recruited via social media underscores the growing concern about the impact of social media on mental health, particularly in younger populations.

Other contributing factors between social media and depression include sleep impairment and sedentary behavior [25]. These factors, combined with the other potentially negative effects of problematic social media use, defined as a "nonsubstance–related disorder by which detrimental effects occur as a result of preoccupation and compulsion to excessively engage in social media platforms despite negative consequences" [26], may contribute to the higher prevalence of depression observed in our sample of MBS candidates. Addressing these issues through increased physical activity could offer protective benefits, as regular exercise has been shown to reduce depressive symptoms, support healthy weight management, and improve sleep quality [25,27-32]. While both depression and obesity are multifaceted conditions influenced by various biological, psychological, and social determinants, promoting physical activity represents a practical starting point for improving overall mental and physical health.

Our second key finding was that NHB participants were 60% less likely to be recruited via social media. This finding is particularly concerning given that NHB individuals bear a disproportionate burden of obesity and its health-related consequences. According to the National Center for Health Statistics, NHB women represent the demographic group with the highest risk for overweight, obesity, and severe obesity, followed closely by NHB men [33]. However, this underrepresentation is not surprising, given the significant history of distrust due to unethical practices that have occurred when enrolling this population in previous scientific studies. Prior research has indicated that NHB individuals are less inclined to participate in scientific studies overall [34]. Our results corroborate this disparity and underscore the challenges of using social media for recruitment in scientific research.

The efficacy of social media as a tool for recruiting diverse samples remains a topic of debate [11]. The success of such methods often depends on a comprehensive understanding of the target audience's social media habits [35]. Recent research has shown promising results, with one study demonstrating a substantial increase in racial and ethnic diversity among participants recruited through social media for a clinical trial [36]. Other studies have raised concerns about potential biases, noting an overrepresentation of younger NHW individuals with higher education and income levels [37]. Our findings align with these observations, as we also reported a higher likelihood of recruiting Hispanic and NHW individuals through social media compared to NHB individuals. This result emphasizes the need for proactive measures to address potential biases in social media recruitment.

Additionally, obesity has been consistently associated with an increased risk of depression [38]. However, the significance of social support, whether received in person or through web-based platforms, cannot be overstated in optimizing positive outcomes for patients undergoing MBS. Previous research has underscored the positive correlation between heightened social support and decreased depression, along with improved weight loss among both pre- and post-MBS groups [39]. In one study, Facebook was used as a platform to deliver weight management program materials, resulting in documented weight loss among participants [40]. Furthermore, a 2016 study examining MBS-related Facebook support groups identified various recurring themes in online discussions, ranging from eating recommendations and postsurgery changes to anxiety, depression, and body image concerns [41]. The study concluded that online groups are valuable for fostering support, sharing experiences, and discussing challenges for pre- and post-MBS patients.

Overall, social media as a tool for research recruitment has several advantages, including a broad reach, cost-effectiveness, and the potential to foster social support related to weight management, eating disorders, and mental health. However, researchers must approach social media recruitment cautiously or risk recruiting homogenous samples and should be aware of social media's association with depression. Combining more traditional recruitment methods with online recruitment can help ensure diverse study samples and enhance the validity, generalizability, and equity of findings.

Future research should explore culturally sensitive recruitment methods that include social media and other innovative techniques while continuing to examine the relationship between mental health and social media usage. Improved recruitment methods should consist of partnerships with community organizations, tailored outreach campaigns, and evaluations of effectiveness. As our digital world continues to evolve, enhancing recruitment diversity is critical to reducing health inequities and advancing the quality and fairness of clinical and public health research.

Limitations and Strengths

Despite providing valuable insights, this study has limitations. One limitation of our study is that the sample was drawn exclusively from a single geographic region, which may impact the generalizability of the findings to a broader population. However, it is crucial to highlight that our sample includes a diverse participant population. Furthermore, our sample consisted specifically of MBS candidates, and we examined social media recruitment among this group and not social media use itself. All results should be interpreted with these key factors and limitations in mind.

XSL•FO
RenderX

Conclusion

In conclusion, our findings show that MBS candidates recruited through social media had a higher prevalence of depression, particularly among younger individuals, and were more likely to be NHW or Hispanic. Therefore, our findings underscore the significance of using various recruitment methods to prevent bias and optimize and ensure sample diversity in research.

Acknowledgments

This work was funded by the National Institutes of Health, National Institute on Minority Health and Health Disparities (Grant R01MD011686). The funding organizations had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; or the decision to submit the manuscript for publication.

Data Availability

The data supporting this study's findings are not openly available due to reasons of sensitivity and are available from the corresponding author upon reasonable request. Data are in controlled access data storage at UTHealth School of Public Health.

Authors' Contributions

Formal analysis: FA, JMF
Investigation: BES, CMA, JS, JPA, SK
Supervision: SEM
Writing – original draft: DP, JMF, SSN
Writing – review & editing: DP, EM, JMF, LX, MSM, OK, SEM, SN, SSN

Conflicts of Interest

None declared.

References

1. Boutari C, Mantzoros CS. A 2022 update on the epidemiology of obesity and a call to action: as its twin COVID-19 pandemic appears to be receding, the obesity and dysmetabolism pandemic continues to rage on. Metab Clin Exp 2022 Aug;133:155217. [doi: 10.1016/j.metabol.2022.155217]
2. Obesity and overweight. World Health Organization. 2021. URL: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight [accessed 2023-12-01]
3. Stierman B, Afful J, Carroll MD, et al. National health and nutrition examination survey 2017-March 2020 prepandemic data files-development of files and prevalence estimates for selected health outcomes. Natl Health Stat Report 2021 Jun 14(158). [doi: 10.15620/cdc:106273] [Medline: 39380201]
4. Valerio G, Maffeis C, Saggese G, et al. Diagnosis, treatment and prevention of pediatric obesity: consensus position statement of the Italian Society for Pediatric Endocrinology and Diabetology and the Italian Society of Pediatrics. Ital J Pediatr 2018 Jul 31;44(1):88. [doi: 10.1186/s13052-018-0525-6] [Medline: 30064525]
5. Chang SH, Stoll CRT, Song J, Varela JE, Eagon CJ, Colditz GA. The effectiveness and risks of bariatric surgery: an updated systematic review and meta-analysis, 2003-2012. JAMA Surg 2014 Mar;149(3):275-287. [doi: 10.1001/jamasurg.2013.3654] [Medline: 24352617]
6. Haase CL, Lopes S, Olsen AH, Satylganova A, Schnecke V, McEwan P. Weight loss and risk reduction of obesity-related outcomes in 0.5 million people: evidence from a UK primary care database. Int J Obes 2021 Jun;45(6):1249-1258. [doi: 10.1038/s41366-021-00788-4]
7. de la Cruz-Muñoz N, Xie L, Quiroz HJ, et al. Long-term outcomes after adolescent bariatric surgery. J Am Coll Surg 2022 Oct 1;235(4):592-602. [doi: 10.1097/XCS.0000000000000325] [Medline: 36102560]
8. Messiah SE, Xie L, de la Cruz-Muñoz N, Lipshultz SE. Use of metabolic and bariatric surgery among US youth. JAMA Pediatr 2023 Aug 1;177(8):856-857. [doi: 10.1001/jamapediatrics.2023.0803] [Medline: 37252718]
9. Alsuhibani A, Thompson JR, Wigle PR, et al. Metabolic and bariatric surgery utilization trends in the United States: evidence from 2012 to 2021 National Electronic Medical Records Network. Ann Surg Open 2023 Dec;4(4):e317. [doi: 10.1097/AS9.0000000000000317] [Medline: 38144499]
10. Messiah SE, Xie L, Mathew MS, et al. Impact of the COVID-19 pandemic on metabolic and bariatric surgery utilization and safety in the United States. Obesity Surg 2022 Jul;32(7):2289-2298. [doi: 10.1007/s11695-022-06077-x] [Medline: 35499637]
11. Topolovec-Vranic J, Natarajan K. The use of social media in recruitment for medical research studies: a scoping review. J Med Internet Res 2016 Nov 7;18(11):e286. [doi: 10.2196/jmir.5698] [Medline: 27821383]

XSL•FO
RenderX

12. Darko EM, Kleib M, Olson J. Social media use for research participant recruitment: integrative literature review. J Med Internet Res 2022 Aug 4;24(8):e38015. [doi: 10.2196/38015] [Medline: 35925655]
13. Pedalino F, Camerini AL. Instagram use and body dissatisfaction: the mediating role of upward social comparison with peers and influencers among young females. Int J Environ Res Public Health 2022 Jan 29;19(3):1543. [doi: 10.3390/ijerph19031543] [Medline: 35162562]
14. Yazdani N, Siedlecki KL. Mediators of the relationship between cognition and subjective well-being. J Happiness Stud 2021 Oct;22(7):3091-3109. [doi: 10.1007/s10902-021-00357-6]
15. Vandenbosch L, Fardouly J, Tiggemann M. Social media and body image: recent trends and future directions. Curr Opin Psychol 2022 Jun;45:101289. [doi: 10.1016/j.copsyc.2021.12.002] [Medline: 35030460]
16. Moufawad M, Hoque A, Kells M, Sonneville KR, Hahn SL. Social media use and weight bias internalization: association moderated by age and weight perception. J Eat Disord 2024 Jun 18;12(1):84. [doi: 10.1186/s40337-024-01043-7] [Medline: 38890765]
17. Walker DC, White EK, Srinivasan VJ. A meta-analysis of the relationships between body checking, body image avoidance, body image dissatisfaction, mood, and disordered eating. Int J Eat Disord 2018 Aug;51(8):745-770. [doi: 10.1002/eat.22867] [Medline: 29659039]
18. Is metabolic and bariatric surgery right for you? American Society for Metabolic and Bariatric Surgery. URL: https://asmbs.org/patients/is-metabolic-and-bariatric-surgery-right-for-you [accessed 2024-03-19]
19. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381. [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
20. Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: building an international community of software platform partners. J Biomed Inform 2019 Jul;95:103208. [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]
21. Sheehan DV, Lecrubier Y, Sheehan KH, et al. The International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. J Clin Psychiatry 1998;59(20):22-33. [Medline: 9881538]
22. Pettersson A, Modin S, Wahlström R, Af Winklerfelt Hammarberg S, Krakau I. The Mini-International Neuropsychiatric Interview is useful and well accepted as part of the clinical assessment for depression and anxiety in primary care: a mixed-methods study. BMC Fam Pract 2018 Jan 24;19(1):19. [doi: 10.1186/s12875-017-0674-5] [Medline: 29368585]
23. Keles B, McCrae N, Grealish A. A systematic review: the influence of social media on depression, anxiety and psychological distress in adolescents. Int J Adolesc Youth 2020 Dec 31;25(1):79-93. [doi: 10.1080/02673843.2019.1590851]
24. McCrae N, Gettings S, Purssell E. Social media and depressive symptoms in childhood and adolescence: a systematic review. Adolescent Res Rev 2017 Dec;2(4):315-330. [doi: 10.1007/s40894-017-0053-4]
25. Lederman O, Ward PB, Firth J, et al. Does exercise improve sleep quality in individuals with mental illness? A systematic review and meta-analysis. J Psychiatr Res 2019 Feb;109:96-106. [doi: 10.1016/j.jpsychires.2018.11.004] [Medline: 30513490]
26. Shannon H, Bush K, Villeneuve PJ, Hellemans KG, Guimond S. Problematic social media use in adolescents and young adults: systematic review and meta-analysis. JMIR Ment Health 2022 Apr 14;9(4):e33450. [doi: 10.2196/33450] [Medline: 35436240]
27. Krittanawong C, Maitra NS, Qadeer YK, et al. Association of depression and cardiovascular disease. Am J Med 2023 Sep;136(9):881-895. [doi: 10.1016/j.amjmed.2023.04.036] [Medline: 37247751]
28. Popovic D, Bjelobrk M, Tesic M, et al. Defining the importance of stress reduction in managing cardiovascular disease - the role of exercise. Prog Cardiovasc Dis 2022;70:84-93. [doi: 10.1016/j.pcad.2022.01.008] [Medline: 35131232]
29. Ozemek C, Lavie CJ, Kaminsky LA. Review of recent cardiac rehabilitation research related to enrollment/adherence, mental health, and other populations. J Cardiopulm Rehabil Prev 2021;41(5):302-307. [doi: 10.1097/HCR.0000000000000649]
30. Boidin M, Grégoire CA, Gagnon C, et al. Effects of variation in exercise training load on cognitive performances and neurotrophic biomarkers in patients with coronary artery disease. J Appl Physiol (1985) 2024 Nov 1;137(5):1158-1167. [doi: 10.1152/japplphysiol.00636.2023] [Medline: 38961824]
31. Lavie CJ, Zhang I, Yang D, Liu M. Improving fitness through exercise will improve our heart and mind. Heart and Mind 2023;7(1):49-51. [doi: 10.4103/hm.hm_59_22]
32. Banno M, Harada Y, Taniguchi M, et al. Exercise can improve sleep quality: a systematic review and meta-analysis. PeerJ 2018;6:e5172. [doi: 10.7717/peerj.5172] [Medline: 30018855]
33. Welsh LK, Luhrs AR, Davalos G, et al. Racial disparities in bariatric surgery complications and mortality using the MBSAQIP data registry. Obesity Surg 2020 Aug;30(8):3099-3110. [doi: 10.1007/s11695-020-04657-3] [Medline: 32388704]
34. Milani SA, Swain M, Otufowora A, Cottler LB, Striley CW. Willingness to participate in health research among community-dwelling middle-aged and older adults: does race/ethnicity matter? J Racial Ethn Health Disparities 2021 Jun;8(3):773-782. [doi: 10.1007/s40615-020-00839-y] [Medline: 32808194]
35. Khatri C, Chapman SJ, Glasbey J, et al. Social media and internet driven study recruitment: evaluating a new model for promoting collaborator engagement and participation. PLoS One 2015;10(3):e0118899. [doi: 10.1371/journal.pone.0118899] [Medline: 25775005]

XSL•FO
RenderX

36. Strauch J, Agnew A, Meenaghan E, Miller KK, Haines MS. Recruitment strategies to increase racial and ethnic diversity in anorexia nervosa clinical research. J Eat Disord 2023 Jul 15;11(1):118. [doi: 10.1186/s40337-023-00844-6] [Medline: 37454157]
37. Whitaker C, Stevelink S, Fear N. The use of Facebook in recruiting participants for health research purposes: a systematic review. J Med Internet Res 2017 Aug 28;19(8):e290. [doi: 10.2196/jmir.7071] [Medline: 28851679]
38. Luppino FS, de Wit LM, Bouvy PF, et al. Overweight, obesity, and depression: a systematic review and meta-analysis of longitudinal studies. Arch Gen Psychiatry 2010 Mar;67(3):220-229. [doi: 10.1001/archgenpsychiatry.2010.2] [Medline: 20194822]
39. Conceição EM, Fernandes M, de Lourdes M, Pinto-Bastos A, Vaz AR, Ramalho S. Perceived social support before and after bariatric surgery: association with depression, problematic eating behaviors, and weight outcomes. Eat Weight Disord 2020 Jun;25(3):679-692. [doi: 10.1007/s40519-019-00671-2] [Medline: 30859467]
40. Napolitano MA, Hayes S, Bennett GG, Ives AK, Foster GD. Using Facebook and text messaging to deliver a weight loss program to college students. Obesity (Silver Spring) 2013 Jan;21(1):25-31. [doi: 10.1002/oby.20232] [Medline: 23505165]
41. Koball AM, Jester DJ, Domoff SE, Kallies KJ, Grothe KB, Kothari SN. Examination of bariatric surgery Facebook support groups: a content analysis. Surg Obes Relat Dis 2017 Aug;13(8):1369-1375. [doi: 10.1016/j.soard.2017.04.025] [Medline: 28600115]

Abbreviations

aOR: adjusted odds ratio
HIPAA: Health Insurance Portability and Accountability Act
MBS: metabolic and bariatric surgery
MINI: Mini International Neuropsychiatric Interview
NHB: non-Hispanic Black
NHW: non-Hispanic White
OR: odds ratio
REDCap: Research Electronic Data Capture

Edited by A Mavragani; submitted 12.04.24; peer-reviewed by C Lavie, H Hecht, K Phankitnirundorn, MB Ellur; revised version received 06.02.25; accepted 07.02.25; published 10.04.25.

Please cite as:
Francis JM, Neti SS, Polavarapu D, Atem F, Xie L, Kapera O, Mathew MS, Marroquin E, McAdams C, Schellinger J, Ngenge S, Kukreja S, Schneider BE, Almandoz JP, Messiah SE
Association of Social Media Recruitment and Depression Among Racially and Ethnically Diverse Metabolic and Bariatric Surgery Candidates: Prospective Cohort Study
JMIR Form Res 2025;9:e58916
URL: https://formative.jmir.org/2025/1/e58916
doi:10.2196/58916

© Jackson M Francis, Sitapriya S Neti, Dhatri Polavarapu, Folefac Atem, Luyu Xie, Olivia Kapera, Matthew S Mathew, Elisa Marroquin, Carrie McAdams, Jeffrey Schellinger, Sophia Ngenge, Sachin Kukreja, Benjamin E Schneider, Jaime P Almandoz, Sarah E Messiah. Originally published in JMIR Formative Research (https://formative.jmir.org), 10.4.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Encouraging the Voluntary Mobilization of Mental Resources by Manipulating Task Design: Explorative Study

Lina-Estelle Louis[1], PhD; Saïd Moussaoui[1], Prof Dr; Sébastien Ravoux[2], ING; Isabelle Milleville-Pennel[1], Prof Dr

[1]École Centrale Nantes, Centre National de la Recherche Scientifique (CNRS), Laboratoire des Sciences du Numérique de Nantes (LS2N), UMR 6004, Nantes Université, Nantes, France

[2]Entreprise Onepoint, Nantes, France

Corresponding Author:
Isabelle Milleville-Pennel, Prof Dr
École Centrale Nantes, Centre National de la Recherche Scientifique (CNRS), Laboratoire des Sciences du Numérique de Nantes (LS2N), UMR 6004, Nantes Université, Nantes, France

Abstract

Background: Cognitive training is increasingly being considered and proposed as a solution for several pathologies, particularly those associated with aging. However, trainees need to be willing to invest enough mental effort to succeed and make progress.

Objective: In this study, we explore how gamification in a narrative context (ie, the addition of visual game-like elements [GLEs] embedded in real-world contexts) could contribute increase in perceived playfulness (PP) and voluntary mental effort allocated to a cognitive task. In such context, narrative elements and GLEs can be designed to align with a commonly relatable scenario (like simulating fishing or gardening activity) to ground the task in familiar, real-world contexts. We also consider if the supposed effect of GLEs on PP and voluntary mental effort could endure while manipulating an intrinsic variable of the task (ie, by increasing cognitive solicitation).

Methods: In total, 20 participants (average age 33.6, SD 8.6 y) took part in 3 cognitive tasks proposed in a numerical format: a classic version of the Corsi test (Classic Corsi, a spatial memory task), a playful version of the Classic Corsi test (Playful Corsi), with added visual GLEs in a narrative context, and a playful version of the Classic Corsi test with added cognitive solicitation, that is, mental motor inhibition (Playful Corsi Multi). We assessed the impact of visual GLEs and cognitive solicitation on PP (1 question) and mental workload (MWL) using NASA-Task Load Index (NASA-TLX) and workload profile (WP) questionnaires.

Results: Results showed that PP was not influenced by interface's playful characteristics (Classic Corsi [mean 62.4, SD 8.8] vs Playful Corsi [mean 66, SD 8.8]; W=77; *P*=.30) but decreased the time necessary to complete the task (Classic Corsi [mean 10.7, SD 2.1 s] vs Playful Corsi [mean 6.8, SD 1.6 s]; W=209; *P*<.001) as well as performance (Classic Corsi [mean 92.4, SD 9.1] vs Playful Corsi [mean 88.2, SD 11.3]; W=140.5; *P*=.02). So, possibly, visual GLEs could raise the stakes of the task slightly and implicitly encourage people to go a bit faster. Furthermore, visual GLEs increased MWL regarding attentional resources (assessed by WP: Classic Corsi [mean 52.4, SD 10.9] vs Playful Corsi [mean 65.8, SD 10.9]; W=27.5; *P*=.04), while manipulating cognitive solicitation impacted MWL when linked to task requirements (assessed by NASA-TLX: Playful Corsi [mean 54.2, SD 9.4] vs Playful Corsi Multi [mean 67.5, SD 9.4]; W=35.5; *P*=.01) without impacting the performance to the task (Playful Corsi [mean 83.8, SD 13.9] vs Playful Corsi Multi [mean 94, SD 5.5]; W=27; *P*=.007). Thus, working on the way cognitive functions are solicited would be wiser than adding visual GLEs to improve users' voluntary mental effort while preserving performance.

Conclusion: These results offer valuable insights to improve users' experience during gamified cognitive tasks and serious games.

(JMIR Form Res 2025;9:e63491) doi:10.2196/63491

KEYWORDS

visual game-like elements; gamification; multiplicity of cognitive functions; cognitive tasks; perceived playfulness; mental workload; performance; cognitive training; aging; mental effort; cognitive function; cognitive skills; cognitive tests

Introduction

Cognitive training is increasingly being considered and proposed as a solution for several pathologies, particularly those associated with aging. However, trainees must be willing to invest enough mental effort to succeed and make progress. During the execution of an activity, we have the possibility of voluntarily activating our mental resources to reach a level of performance set by the task or by ourselves. This voluntary modulation of the quantity of mental resources invested when performing a task corresponds to the notion of mental effort [1-3]. Voluntary mental effort is a dimension of mental workload (MWL). The latter is a broader concept that also encompasses

XSL•FO
RenderX

notions such as the perception of our own performance, or emotional states such as frustration. Voluntary mental effort should be distinguished from mental effort imposed by the task properties and undergone by the participant. Indeed, voluntary mental effort must be seen as a "positive" mental effort as it corresponds to a voluntary deployment of resources with positive consequences for the participant (high perceived performance and low frustration). Thus, in this context it should be considered regarding the corresponding perceived performance and frustration when performing the task. Positive mental effort should be accompanied by high perceived performance and low frustration. To make the notion of positive mental effort operational, we return to the definition of this concept, specifying that it is part of the MWL construct. In the scientific literature, NASA-Task Load Index (NASA-TLX) [2] and workload profile (WP) [4] questionnaires permit to measure MWL linked respectively to task requirements and the mobilization of attentional resources (since the WP employs the multiple resource theory of [5-7] as its foundation). NASA-TLX is particularly interesting for addressing the notion of positive mental effort, as it considers various dimensions of MWL, including mental effort, perceived performance and frustration, that we considered as crucial in the definition of positive mental effort.

Mental effort has been studied in various contexts since it is a key element for enhancing performance. However, few studies have focused on the notion of positive mental effort during cognitive training, which aims to enhance general cognitive skills such as memory, attention, or processing speed through repeated interventions using cognitive tasks or intellectually demanding activities [8-10]. Our aim is thus to study the factors that influence voluntary mobilization of mental effort when confronted to cognitive tasks used in a cognitive training process. To do that, we need to consider a task allowing to assess voluntary mobilization of mental effort but without the confounding effects associated with longitudinal cognitive training paradigms (reinforcement, error feedback, and repetition). Thus, in this study, our aim is not to design interventions for cognitive training, but rather to propose isolated cognitive tasks as controlled experimental tools to investigate the role of some factors. These factors could then be integrated in tasks developed for cognitive training. So, we wondered how we could encourage individuals to voluntarily mobilize greater mental effort to achieve better performance. To this end, we wanted to explore one possibility, that is, the gamification of a cognitive task.

Gamification is defined as the use of game-like elements (GLEs) in nongame contexts. GLEs are elements typically found in games such as visual elements (icons, colored patterns, evocative images, or playful animations), badges systems, narration, or leaderboards [9,11]. This definition of gamification differentiates it from serious games. Indeed, according to Vermeir et al [9], serious games use full-fledged games in nongame contexts, whereas gamification uses elements of a game integrated into real-world contexts (for instance, a cognitive task that uses colored patterns).

Furthermore, gamification can be linked to cognitive tasks, but its introduction may affect the validity of cognitive tests by potentially introducing new cognitive demands. Therefore, the gamification of cognitive tasks is most applied within the context of cognitive training [9,11-13]. The reason is because cognitive tasks are perceived as demotivating and frustrating given their repetitive and effortful aspects (Lumsden et al [14]) which may impact on perceived playfulness (PP). PP is the sense of being focused on an activity that is viewed as inherently enjoyable, fun, and interesting [15,16]. Higher PP results in higher user intention to perform a task [17], a stronger attitude toward using a particular technology [16], a better performance, and a higher effective response to computer training tasks [18]. Thus, we predict that the increasing of PP would impact the voluntary mobilization of mental resources and thus positive mental effort. There would therefore be a real interest in increasing the PP of a cognitive task during cognitive training. Integrating GLEs into cognitive tasks could be one solution, as their inclusion is supposed to increase PP [11,19].

Among all GLEs, performance feedback via a reward points system (the unit of measurement quantifying a user progression) is one of the most widely used [19] and therefore the most studied in the literature. Nevertheless, this system provides feedback on performance and thus contributes to its improvement. While it does indeed increase motivation and PP, it also introduces a bias in the sense that it induces a form of learning. In this study, we aimed to understand the factors that encourage the voluntary mobilization of mental resources in a cognitive task (positive mental effort), with the hypothesis that this mobilization will have a positive impact on performance. Introducing reinforcement would induce a sort of learning process that could mask the effect of other variables allowing to increase positive mental effort. Thus, we provide no reinforcement through performance feedback.

Consequently, only a selected few GLEs, specifically visual ones, which have no impact on performance feedback, should be incorporated. Visual GLEs can be defined as the deliberate arrangement of several visual variables, such as position and form [20], integrated into a particular context. Visual GLEs are contextualized with narrative elements (short story explaining the task and a visual surrounding corresponding to the story) to make them meaningful for the participant and add ecological aspect to the cognitive task. By ecological aspect, we mean a less arbitrary situation. The task has an objective that can be linked to a game that could be encountered in real life. For example, the narrative elements can be designed to align with commonly relatable scenarios, (like aiming at targets, arranging objects to solve a puzzle, navigating a maze, or simulating a fishing or gardening activity) to ground the task in familiar, real-world contexts. These elements serve two purposes, that is, (1) the short story provides a framework for participants to understand the task's purpose, making the activity less abstract and (2) the visual surroundings (eg, buoys, and water effects) enhance immersion, reinforcing the narrative and fostering a sense of engagement. Together, these components aim to reduce the perceived arbitrariness of the cognitive task, encouraging participants to connect the task with a recognizable and meaningful activity from their own experiences.

Nevertheless, despite showing that GLEs increase PP [19], no study has specifically investigated the contribution of visual

XSL•FO
RenderX

GLEs integrated with a narrative context within cognitive tasks and the impact on both PP and positive mental effort. Thus, our first aim (hypothesis 1) is to ascertain if visual GLEs integrated with a narrative context (which will be called a playful task) are enough to increase PP and then positive mental effort (characterized by higher perceived mental effort and performance but lower perceived frustration) when compared with a classic design. And therefore, we predicted that this increase in voluntary mental effort could also increase real performances.

Nevertheless, in everyday life, an individual must involve various amounts of cognitive functions to meet the requirements of an activity. The number of cognitive functions solicited is a variable intrinsic to task requirements which would have a direct impact on mental demands. Mental demands is the dimension of MWL directly related to the amount of mental activities required [2]. This could also impact mental effort imposed by the task itself. We can thus assume that this effect could also play a role in positive mental effort, notably via perceived performance and frustration. Thus, an increase in mental effort required by the task could lead to a decrease in positive mental effort. To our knowledge, no study has investigated the impact of increasing the solicitation of cognitive functions on PP, positive mental effort, or the resulting actual performances in a narrative ecological context. Thus, our second aim (hypothesis 2) is to determine if the multiplicity of cognitive solicitation of the playful task might decrease PP, positive mental effort (characterized by higher perceived effort and frustration but fewer perceived performance), and actual performances when compared with the simple playful task.

Hence, this study will explore how extrinsic features (visual GLEs within a narrative context) impact PP and positive mental effort, and whether this effect persists when manipulating intrinsic demands of the task (eg, the multiplicity of cognitive solicitation in a narrative context). To this end, we have set up a protocol containing various cognitive tasks. In the first step, we will consider the impact of adding GLEs to a standard cognitive task. Then we will investigate further the impact of adding cognitive solicitation to this task. The tasks and the logic behind their design are detailed in the Methods section.

Methods

Ethical Considerations

This study has received a positive decision from the Northwest III Ethics Commitee for the Protection of Persons, with the reference number n° 21.04612.000059, in accordance with the ethical guidelines for human participants research and the Declaration of Helsinki. Informed consent was obtained from all participants. All data were anonymized to ensure privacy and were securely stored on a protected cloud platform accessible only to the research team. No identifiable participant information is included in the manuscript or supplementary materials. At the end of the study, each participant received a €50 (US $54.69) gift voucher as compensation.

Participants

In total, 20 (9 women, 11 men) took part in this experiment. Their average age was 33.6 (SD 8.6) years. All of them had at least a bachelor's degree and 85% (17/20) of them had at least a bac+3 level. Furthermore, 70% (14/20) worked for Onepoint company, 20% (4/20) came from social network LinkedIn (Microsoft) and the others were acquaintances of the experimenters or from Nantes University Hospital.

Recruitment

Participants were recruited by mailing an information letter inviting interested people to contact the experimenter if they wanted to participate in the study. At the reception of the mail, the experimenter contacted the participant in order to confirm (by oral statement) that he or she were healthy French-speaking volunteers residing in Nantes region that met the inclusion criteria (fluent in French, right-handed, having normal or corrected-to-normal vision, and having normal or corrected-to-normal hearing). Then the participant was informed about the experimental duration and procedure. After acceptance by the participant, an appointment was made for the experimental tests. The consent document was signed at the participant's arrival, before the beginning of the experiment.

Tasks and Procedure

Overview

The experimentation took place at Nantes University Hospital. Participants performed 3 experimental tasks in a random order on a screen positioned in front of them. All tasks were devised using Unity3D development software (version 2019.4.30f1; Unity Technologies). A link to the videos of all tasks is available in Multimedia Appendix 1.

To determine the level of complexity of the tasks, we considered several factors. Primarily, the complexity of a simple task (only soliciting 1 cognitive function) should not be arduous. Indeed, a simple task should accommodate the integration of an additional cognitive function in a condition that involves multiple cognitive functions. On the other hand, we did not want the simple task to appear too simplistic. Thus, we relied on the intermediate complexity levels from the study conducted by Louis et al [21], which offers a compromise considering the aforementioned conditions. Indeed, Louis et al [21] compared several cognitive tasks with different complexity levels, and the 5-item Corsi test induced an intermediate MWL (based on NASA-TLX) while still allowing high performance. Thus, the 5-item Corsi test is a sufficiently cognitively demanding task without being difficult. Consequently, to verify hypothesis 1 (visual GLEs integrated with a narrative context increase PP, positive mental effort, and performances) and 2 (the multiplicity of cognitive solicitation might decrease PP, positive mental effort, and performances), 3 digitized versions of the Corsi test were proposed, that are Classic Corsi [22], Playful Corsi, and Playful Corsi Multi.

Classic Corsi

This task [22] is a visual-spatial memory span task involving remembering a sequence of cubes pointed by the computer and reproducing the sequence in the same order by clicking with

XSL•FO
RenderX

the mouse on screen elements. Thus, the cubes have been selected by a mouse click; in the following sections, we will use the term “clicked cubes” to designate this action. The material (Figure 1A) used is consistent with the study by Kessels et al [23] that proposed for each cube a number which is not visible by the participants (Figure 1B). Furthermore, 5 trials involving the memorization of 5 cubes (intermediate level as adopted from a study by Louis et al [21]) are proposed. The five sequences of cubes were (1) 5-2-1-8-6, (2) 4-2-7-3-1, (3) 3-9-2-4-8, (4) 7-8-2-9-4, and (5) 4-2-6-8-1 (Figure 1B).

The sequence of cubes for the training session was 2 - 6.

Performance dimensions include clicked cubes (number of cubes that are clicked by the participant), correct cubes (the cubes correctly recalled whatever the order), false cubes (the cubes incorrectly recalled), omissions (failure to press a cube), total time, and expected responses (common performance indicator for all tasks) corresponds to this calculation: 100×(25–(False cubes+Omissions)÷25) and exact sequences (number of cubes clicked in the expected order and position) which can be compared with the terminology of memory or block span, corresponding to the length of the last correctly repeated sequence, in a traditional Corsi test [23]. However, our task did not aim to determine a memory span, as is typically the case in Corsi tests. In a memory span assessment, the goal is to identify the maximum number of items a participant can correctly retain and reproduce. In contrast, our task focused on exact sequences, and the instructions provided to participants did not explicitly emphasize memory span as a performance objective. Therefore, we have chosen to retain the term exact sequences to better reflect the specific nature and goals of our task.

Figure 1. (A) Visual environment of the digitized version of the Classic Corsi test. (B) Positions of cubes in the Classic Corsi test from study by Kessels et al [23]. (C) Playful version of the digitized version of the Classic Corsi test: cubes are replaced by 9 buoys with different colors integrated in a narrative context (throwing balls towards buoys floating in a pond).

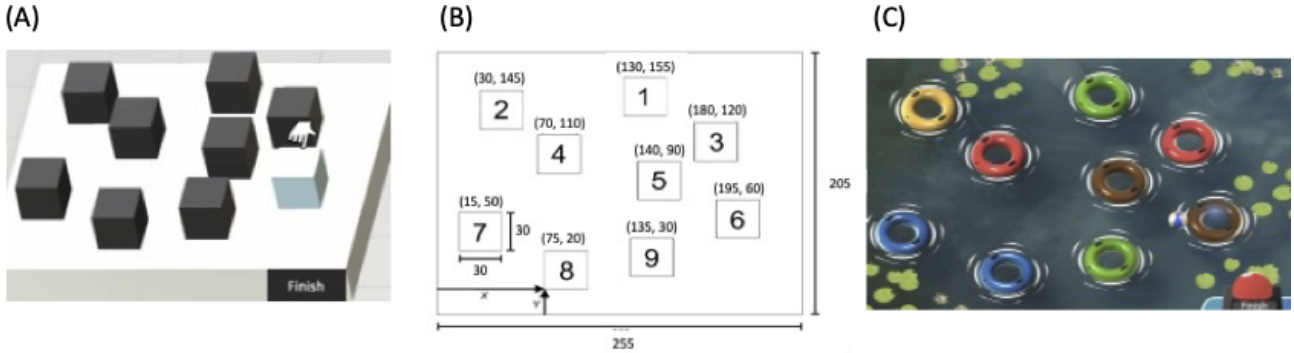

Playful Corsi

This task is identical to the classic version, but cubes are replaced by 9 buoys with different colors integrated in a narrative context (throwing balls toward buoys floating in a pond; Figure 1C). This task uses the same performance metrics as Classic Corsi.

To verify hypothesis 2 (the multiplicity of cognitive solicitation might decrease PP, positive mental effort, and performances), we needed a task that solicited another cognitive function. We called it “Playful Corsi Multi.”

Playful Corsi Multi

Following several pretests in which various cognitive functions (such as mental flexibility or working memory updating) were compared, it was observed that mental motor inhibition, as required in tasks such as go no-go [24], was the most suitable cognitive function to integrate with the Corsi test. This choice minimized the impact on the extrinsic characteristics of the task while maintaining its core objectives.

The Playful Corsi Multi uses the same material as Playful Corsi (Figure 1). Participants memorize a sequence of 5 ball throws, reproducing it by clicking twice on buoys. During sequence reproduction, flowers may appear on buoys that must be clicked. In this case, participants do not have to make the second click on the gray buoy with a flower (no-go), and the response is validated with a single click (Figure 2). Furthermore, 4 pointing sequences with 2 no-go targets (gray buoy with flower) are proposed. In all sequences, a “trap” buoy with a flower on a nongray buoy is present and must be clicked twice, as similarly required for a buoy (of any color) without a flower. The sequence of 5 pointed buoys for the training session was with 1 no-go and 1 ‘’trap’’ buoy with a flower on a nongray buoy.

Performance dimensions encompass false buoys, false alarms (2 clicks instead of 1 upon the appearance of a no-go target), omissions (failure to perform a second click on a buoy), clicked buoys, exact buoys (number of buoys clicked in the expected order and position), expected responses (correspond to the accuracy and are calculated as 100×[25−(False alarms+omissions)÷25], *exact sequences* (number of buoys clicked in the expected order and position), reaction time for correct responses (the buoys correctly recalled), reaction time for all items combined, and total time (between the appearance of the red buzzer [finish button] and pressing it).

XSL•FO
RenderX

Figure 2. Schematic figures of the Playful Corsi Multi. (A) Ball displacement corresponding to memorization instructions given to participants; (B) actions required by the participants to respond to the instructions (2 clicks if there is no flower on the buoys and 1 click if there is a flower on the buoys).

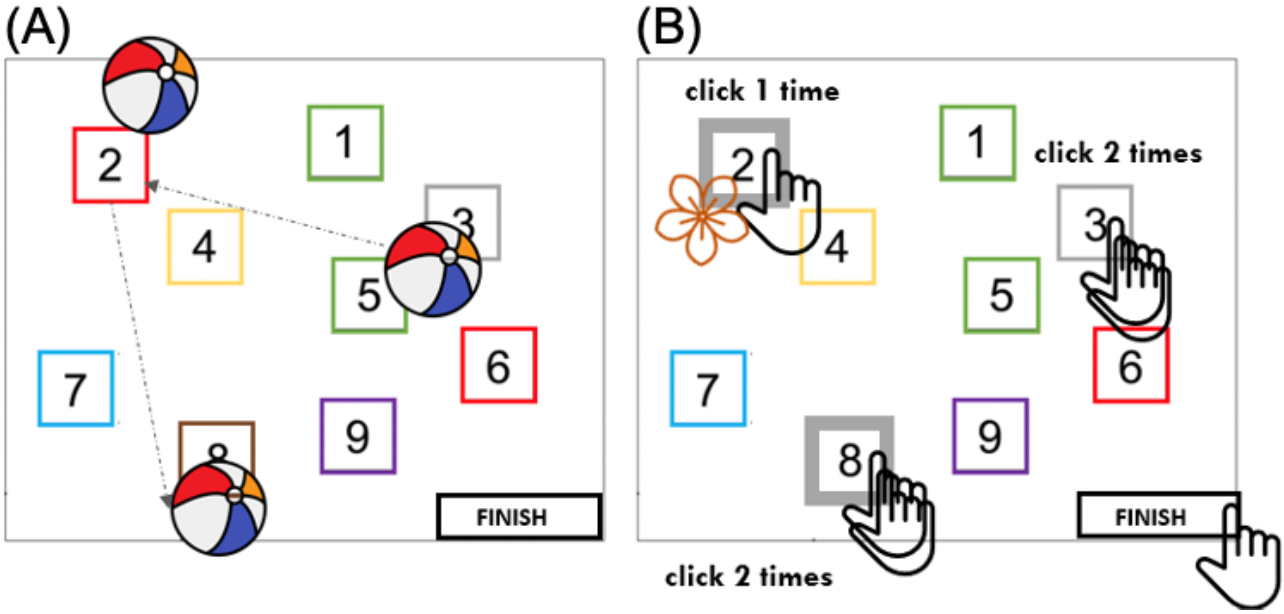

Subjective Measures

Studies by Rubio et al [25] and Paxion [26] indicate that WP and NASA-TLX questionnaires complement each other. Furthermore, our study aimed to assess MWL related to attentional resources (WP questionnaire) and cognitive functions, emphasizing mental demands and mental effort (NASA-TLX questionnaire).

NASA-TLX Questionnaire

NASA-TLX [2] evaluates perceived workload across 6 distinct subscales. In total, 3 dimensions relate to activity, that are mental demands (requirements for mental and perceptual activity), physical demands (requirements for physical activity), and temporal demands (sense of time pressure), 2 dimensions relate to strategies, that are performance (feeling of success in achieving the objectives set by the experimenter or oneself) and effort (the mental or physical effort required to reach the performance level targeted), and 1 dimension relate to emotional state, that is frustration (irritation, stress, insecurity, and discouragement). Following each task level, participant score each dimension from 0 (no demand) to 100 (maximum demand). In our study, we used the unweighted version of the questionnaire by calculating an average of the 6 dimensions to derive a raw task load index validated by Byers et al [27].

WP Questionnaire

The WP questionnaire [4] asked the participants to provide the proportion of attentional resources across 8 subscales. Indeed, the dimensions of workload in this method were aligned with those in the multiple resource theory (described in Introduction). According to Tsang and Velazquez [4], 2 dimensions of WP questionnaire relate to “stages of processing”: perceptual or central processing (attentional resources required for activities like perceiving, remembering, problem-solving, and decision making) and response selection and execution (attentional resources required for response selection and execution; eg, the selection of the appropriate pedal to stop an automobile). Furthermore, 2 other dimensions are related to “processing codes”: spatial processing (according to Tsang and Velazquez [4], for tasks that are spatial in nature) and verbal processing (for tasks that are verbal in nature). In addition, 2 other dimensions are related to “input modality”: visual processing (for tasks that are performed based on the visual information received) and auditory processing (for tasks that are performed based on the auditory information received). And finally, 2 dimensions are related to “output modalities”: manual output (for tasks that required considerable attention for producing manual response), and speech output (for other tasks requiring considerable attention for producing speech responses).

Participants score each dimension from 0 (no demand) to 100 (maximum demand) representing the proportion of allocated attentional resources.

To reduce the time spent on questionnaires and given the absence of problem solving but also considering that the NASA-TLX’s mental demand dimension also questioned this source of MWL, the dimension perceptual or central processing was not considered. For the same reasons, because of the absence of actioners’ selection, auditory and speech tasks in our study, the dimensions response selection and execution, auditory processing, and speech output were not considered. Thus, only 4 of the 8 dimensions were evaluated. Indeed, as demonstrated by Tsang and Velazquez [4], the dimensions can be considered independently and rated in a 1D way. Therefore, we decided to focus on the spatial processing (WP3) and visual processing (WP5) as Corsi test is a visuo-spatial task, verbal processing (WP4) as the participants mentally verbalize the go no-go items (eg, the buoys colors), and manual output (WP7) dimensions, in relation to the participant’s movement of the mouse to click on the various items.

PP Measure

Regarding PP, we assessed it using a question presented on a Likert scale ranging from 0 to 100: “To what extent did you find the task you have just completed playful and fun?”

XSL•FO
RenderX

Experimental Procedure

After receiving approval from the Northwest III Ethics Committee for the Protection of Persons, the recruitment started. Potential participants received a study summary via email, followed by individual meetings held 7 days before the Nantes University Hospital study. Then, participants signed a consent letter. The experimental phase proceeded as follows: Participants were welcomed at Nantes University Hospital. Their eligibility and nonopposition to participation in the study were verified and they completed a pretest questionnaire (age, gender, and level of education). Then, they began the experiment. Each task had a training session (refer to Task and Procedure section) and after each task level, participants received NASA-TLX and WP questionnaires, and PP measure. Including questionnaires, each Corsi test (classic and playful one) has a total duration of 11 min and Playful Corsi Multi has a total duration of 16 min.

Data Analysis

Raw data is available in Multimedia Appendix 2. Subjective measures (NASA-TLX, WP, and PP) were normalized through Cousineau correction [28]. This normalization allows for more accurate estimation of within - participants CIs for subjective data in the case of experimental designs with several conditions. Cousineau correction was described as follows. Let y_{ij} be the ith participant's score in the jth condition (i=1, ..., N; j=1, …, M). Then, the normalized observations z_{ij} were defined as follows:

$$z_{ij}=y_{ij}-\frac{1}{N}\sum_{j=1}^{M}y_{ij}+\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}y_{ij}$$

Furthermore, some distributions of our data were not normal, and variances were not homogeneous. Thus, we used nonparametric statistics (Wilcoxon signed-rank tests and Conover post hoc comparisons) with the JASP (Jeffrey's Amazing Statistics Program) software (version JASP 0.16; University of Amsterdam).

Finally, for the P value, the lower error threshold of .05 was considered significant.

Results

Comparisons Between Classic Corsi and Playful Corsi

Performance Measures

For the Corsi tests, 2 relevant performance dimensions appeared: expected responses and total time. The other Corsi performances (clicked cubes, correct cubes, false cubes, omissions, and exact sequences) were included within expected responses dimension (refer to Methods section).

Between the 2 tasks, expected responses and total time were significantly higher for the Classic Corsi than the Playful Corsi (Table 1; Figure 3).

Table . Statistical comparisons between Classic Corsi and Playful Corsi concerning performances data and mental workload questionnaires.

	Classic Corsi, mean (SD)	Playful Corsi, mean (SD)	W	*P* value
Expected responses	92.4 (9.1)	88.2 (11.3)	140.5	.02
Total time	10.7 (2.1)	6.8 (1.6)	209	<.001
Perceived playfulness	62.4 (8.8)	66 (8.8)	77	.30
Overall	187 (52.9)	186.3 (52.8)	65	.23
Mental demands	46.6 (13.8)	54.2 (13.9)	53	.16
Physical demands	13.6 (5.9)	13.4 (5.2)	36	.82
Temporal demands	35.6 (11.3)	38.9 (15.4)	93	.95
Effort	39.5 (11.7)	43.8 (13.6)	71	.54
Perceived performance	23.7 (10.4)	30.4 (9.9)	55.5	.19
Frustration	27.8 (16.8)	35.8 (16.8)	64.5	.23
WP3[a]	52.4 (12.1)	59.1 (12.1)	85.5	.48
WP5[b]	52.4 (10.9)	65.8 (10.9)	27.5	.04
WP7[c]	15.3 (12.8)	23.6 (12.8)	42.5	.30

[a]WP3: workload profile on spatial processing.

[b]WP5: workload profile on visual processing.

[c]WP7: workload profile on manual output.

XSL•FO
RenderX

Figure 3. Comparisons of mental demands, effort, perceived performance, frustration, perceived playfulness, and expected responses between Classic Corsi and Playful Corsi and between Playful Corsi and Playful Corsi Multi.

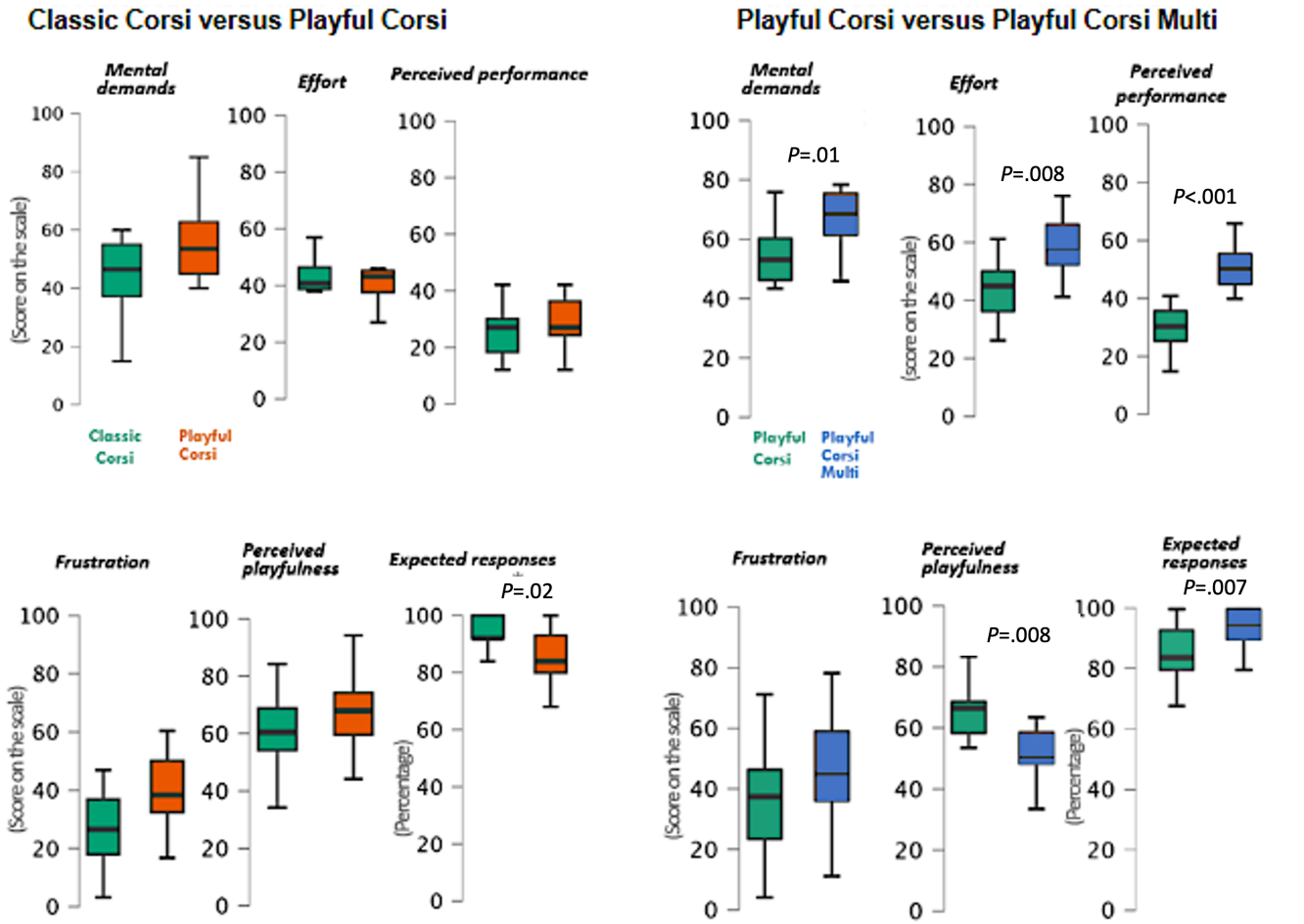

NASA-TLX Questionnaire

Comparing Classic Corsi and Playful Corsi, the overall NASA-TLX score (corresponding to the sum of the 6 NASA-TLX dimensions which are on a scale of 600) showed no significant difference (Table 1), also for mental demands, physical demands, temporal demands, effort, performance, and frustration.

WP Questionnaire

We did not consider the WP4-verbal dimension since it was not a solicited dimension by the Corsi test, a visuo-spatial task (but it was solicited by Multi Playful Corsi test that contained mental inhibition based on the item colors; described in more detail in the Multicolored Playful Corsi vs Playful Corsi Multi section). Between the Classic Corsi and the Playful Corsi, only the WP5-visual dimension was significantly higher for the Playful Corsi than for the Classic Corsi (Table 1). But, between the 2 conditions, WP3-spatial and WP7-manual showed no significant difference (Table 1).

PP Measure

Between the 2 tasks, PP showed no significant difference (Table 1).

Multicolored Playful Corsi vs Playful Corsi Multi

Performance Measures

Total time was not considered, as in the Playful Corsi Multi, there are 2 clicks per response, instead of 1 for the Multicolored Playful Corsi. Therefore, a trial-by-trial comparison was not possible. Concerning expected responses between the 2 tasks, they were significantly better for Playful Corsi Multi than Playful Corsi (Table 2; Figure 3).

XSL•FO
RenderX

Table . Statistical comparisons between Playful Corsi and Playful Corsi Multi concerning performances data and mental workload questionnaires.

	Playful Corsi, mean (SD)	Playful Corsi Multi, mean (SD)	W	*P* value
Expected responses	83.8 (13.9)	94 (5.5)	27	.007
Perceived playfulness	66 (11.9)	50.9 (11.9)	161	.008
Overall	216.6 (40.5)	284.4 (40.5)	25.5	.003
Mental demands	54.2 (9.4)	67.5 (9.4)	35.5	.01
Physical demands	13.4 (8.9)	15.4 (8.9)	33	.40
Temporal demands	38.9 (12.4)	46.6 (12.4)	56	.20
Effort	43.8 (9.9)	58.1 (9.9)	29	.008
Perceived performance	30.4 (7)	50.2 (7)	1	<.001
Frustration	35.8 (17.7)	46.5 (17.7)	63	.20
WP3[a]	59.1 (12.6)	64.4 (12.6)	81.5	.40
WP4[b]	10.2 (11.7)	16.3 (11.7)	14	.30
WP5[c]	65.8 (7.4)	68.3 (7.4)	71	.50
WP7[d]	23.6 (14.7)	21.1 (14.7)	93	.80

[a]WP3: workload profile on spatial processing.

[b]WP4: workload profile on verbal processing.

[c]WP5: workload profile on visual processing.

[d]WP7: workload profile on manual output.

NASA-TLX Questionnaire

The overall MWL was significantly higher for Playful Corsi Multi than Playful Corsi (Table 2; Figure 3), as were the mental demands, effort, and performance. However, physical demands, temporal demands, and frustration showed no significant difference between the 2 tasks.

WP Questionnaire

Between the tasks, the WP3, WP4, WP5, and WP7 showed no significant difference (Table 2; Figure 3).

PP Measure

PP was significantly higher for the Playful Corsi compared to the Playful Corsi Multi (Table 2; Figure 3).

Discussion

Principal Objectives

In this study we wanted to explore (1) how extrinsic variable (visual GLEs within a narrative context) impacts PP, positive mental effort, and performances at the very first stage of facing with a task devoted to cognitive training, and (2) if this effect endures while manipulating intrinsic variable (multiplicity of cognitive solicitation in a narrative context).

Impact of Visual GLEs on PP and Positive Mental Effort

Our first hypothesis that visual GLEs integrated with a narrative context increase PP and positive mental effort is invalidated for various reasons. First, we observed no difference in terms of PP between Classic Corsi and its playful version. In the classic version of the Corsi, the PP is not poor but falls short of optimal. Consequently, we could have potentially improved the PP by manipulating the visual GLEs, but this was not the case. In this way, the PP did not seem to depend on the visual aspect of the task.

Concerning the supposed action of visual GLEs on positive mental effort, it appears that mental effort, perceived performance, and frustration did not differ between the 2 tasks. Thus, visual GLEs are not enough increasing positive mental effort.

On the contrary, MWL related to the mobilization of visual resources increased when we manipulated the extrinsic variables to task requiring (presence of visual GLEs). This is not surprising as, in the Playful Corsi, there are additional visual elements, such as waves or water lilies, which increase the MWL based on visual processing. Nevertheless, no other difference was noted about spatial processing and manual output, indicating that the tasks are relatively similar on the other dimensions.

Concerning the impact on effective performances, visual GLEs decreased the total time compared with a classic design, potentially attributed to the inherent challenge of game-like environments. There are no speed instructions for this task. Speed is therefore neither a performance criterion nor an objective to be achieved. The increase in speed observed in the playful condition may be due to an implicit incentive to accelerate, given the game-like design of the task, where speed is often an inherent goal. Concerning the expected responses (corresponded to the accuracy), our result contradicted the study by Redlinger et al [11] which specified that visual GLEs did not affect performance. Indeed, in the study by Redlinger et al

XSL•FO
RenderX

[11], the presence of backgrounds led to slight reductions in accuracy but not significantly. In our study, this result is to consider in light to the observed impact on the response time. Indeed, we can assume that going faster leads to more errors. This can be linked to the well-known speed-accuracy trade-off [29], that is, the accuracy of a response varies with the time taken to produce it. Furthermore, given that in our study, the visual GLEs included changes to the background, additionally to the object on which participants had to interact (cubes or buoys), this could have induced greater distraction for participants and therefore significant differences in performance in favor of the Classic Corsi.

In conclusion, our results were in line with the meta-analysis of Vermeir et al [9], indicating that gamified tasks are perceived as significantly more demanding, at least considering the perceived demand. Furthermore, this gamification is not accompanied with a voluntary mental effort improvement nor a better PP. A deterioration in performance has even been observed. So, adding visual GLEs with narrative context is not always beneficial for the task.

Impact of Increasing Cognitive Demanding

Our second hypothesis that the multiplicity of cognitive demanding of the playful task might decrease PP and positive mental effort is partially confirmed. PP is negatively influenced by intrinsic variables (solicitation of multiple cognitive functions) but there is every reason to believe that this manipulation has increased the positive mental effort. This is supported by a higher perceived mental effort in conjunction with a higher perceived performance (confirmed by better effective performances) but no increase in the perceived frustration.

Concerning the decrease in PP, increasing the solicitation of cognitive processing increases mental demand, which could have a negative impact on PP. Indeed, Fang et al [17] and Donovan et al [30] showed that PP had a significant and negative correlation with task complexity. In our study, this could also stem from the way we question the notion of PP: “To what extent did you find the task you have just completed playful and fun?” As the task becomes more cognitively demanding, it can be perceived as more “serious,” which could reduce the feeling of “playful.” Also, we acknowledge the limitation of combining 2 descriptors (“playful” and “fun”) in a single question, as they may not hold the same meaning for all participants. “Playful” could refer to the task’s design or its gamified elements, while “fun” might reflect the participant’s subjective enjoyment or emotional response. This overlap introduces potential ambiguity, as participants may weigh one descriptor more heavily than the other. For the next studies, we will consider separating these descriptors into distinct questions to capture more nuanced insights into PP and enjoyment, thereby reducing interpretive variability.

Nevertheless, even though the augmentation of the complexity of a cognitive task decreases the PP, this manipulation increases mental effort and perceived performance without increasing frustration. This is the opposite of what we expected to see. This can stem from the fact that “serious” is not antagonistic to “pleasant,” “enjoyable,” or “motivating.” It just means that what would help induce greater voluntary mental effort, would correspond more to stronger cognitive stimulation (agreed) than just the “fun” characteristics of an activity. This means that, to encourage the voluntary mobilization of mental resources, one possibility is to manipulate the cognitive demanding of a task rather than its design only. Here, by “cognitive demanding,” we are talking about cognitive challenge, not difficulty. In fact, the actual performance, which is better in this condition, shows us that the task is not more complex to perform. But everything seems to indicate that the participant is more determined to perform. Such a performance boost, resulting from cognitive function multiplicity, could be a drawback when performing cognitive assessment but is likely to be desirable in a cognitive training scenario [14]. Indeed, in cognitive training scenarios, the goal is to enhance overall cognitive functioning by engaging multiple cognitive processes simultaneously since it mimics the complexity and demands of real-life situations where multiple cognitive functions are often required at once.

Therefore, to encourage the voluntary mobilization of mental resources, based on our results, a possible approach could be to manipulate the cognitive demanding of a cognitive task (increasing mental effort and performance without increasing frustration) rather than its design but allowing it to maintain a high level of performance.

Limitations

Finally, it is important to specify that this study should be extended to include a larger number of participants to strengthen statistical significance. Given that the experiments were carried out at the Nantes Hospital, which is far from the city center and has few medical staff available, we were constrained by logistical considerations. In addition, the choice of terms used for measuring PP was not without ambiguity and may require refinement in future studies.

Conclusion and Implications for Research

This study can provide cognitive testing practitioners or even social game practitioners with valuable information on specific game-like guidelines to impact users’ PP, while taking into consideration the impact on MWL. We showed that the PP is a skillful balance between design and the solicitation of cognitive functions. Possibly, visual GLEs could raise the stakes of the task slightly and implicitly encourage people to go a bit faster (as observed with the Corsi test), but working on the way cognitive functions are solicited would be wiser to improve users’ voluntary mental effort than adding GLEs. Future research could explore further this last hypothesis as well as the influence of additional GLEs, such as sounds, variations in narrative context, or performance feedback.

XSL•FO
RenderX

Acknowledgments

The authors want to thank all participants for voluntarily taking part in this study and Sébastien Maire who implemented the task battery. The authors also thank Nantes Hospital, where the experiments were conducted, and more particularly, Aurélien Van Langhenhove and Vincent Roualdes for their hospitality, expertise, and discussions.

Data Availability

The data that support the findings of this study are available from the corresponding author (IMP), upon reasonable request.

Authors' Contributions

LEL contributed to the conceptualization, methodology, investigation, data curation, formal analysis, writing – original draft, visualization, and project administration. IMP contributed to the conceptualization, methodology, formal analysis, writing – review and editing, project administration, and supervision. SM handled the resources and supervision. SR handled the supervision and funding acquisition.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Video demonstration of the tasks used in the study.
[MP4 File, 107500 KB - formative_v9i1e63491_app1.mp4]

Multimedia Appendix 2
Raw data on how to encourage the voluntary mobilization of mental resources in a cognitive task.
[XLSX File, 36 KB - formative_v9i1e63491_app2.xlsx]

References

1. Kahneman D. Attention and Effort: Prentice-Hall; 1973.
2. Hart SG, Staveland LE. Development of NASA-TLX (task load index): results of empirical and theoretical research. In: Hancock PA, Meshkati N, editors. Human Mental Workload: North-Holland Press:139-183. [doi: 10.1016/S0166-4115(08)62386-9]
3. Delignières D, Deschamps T. L'effort mental [Book in French]. In: Delignières D, editor. L'effort: Revue EPS; 2000:25-40.
4. Tsang PS, Velazquez VL. Diagnosticity and multidimensional subjective workload ratings. Ergonomics 1996 Mar;39(3):358-381. [doi: 10.1080/00140139608964470] [Medline: 8849491]
5. Wickens CD. Processing resources in attention. In: Parasuraman R, Davis DR, editors. Varieties of Attention: Academic Press; 1984:63-101.
6. Wickens CD. Multiple resources and performance prediction. Theor Issues Ergon Sci 2002 Jan;3(2):159-177. [doi: 10.1080/14639220210123806]
7. Wickens CD. Multiple resources and mental workload. Hum Factors 2008 Jun;50(3):449-455. [doi: 10.1518/001872008X288394] [Medline: 18689052]
8. Rabipour S, Raz A. Training the brain: fact and fad in cognitive and behavioral remediation. Brain Cogn 2012 Jul;79(2):159-179. [doi: 10.1016/j.bandc.2012.02.006] [Medline: 22463872]
9. Vermeir JF, White MJ, Johnson D, Crombez G, Van Ryckeghem DML. The effects of gamification on computerized cognitive training: systematic review and meta-analysis. JMIR Serious Games 2020 Aug 10;8(3):e18644. [doi: 10.2196/18644] [Medline: 32773374]
10. Gobet F, Sala G. Cognitive training: a field in search of a phenomenon. Perspect Psychol Sci 2023 Jan;18(1):125-141. [doi: 10.1177/17456916221091830] [Medline: 35939827]
11. Redlinger E, Glas B, Rong Y. Impact of visual game-like features on cognitive performance in a virtual reality working memory task: within-subjects experiment. JMIR Serious Games 2022 Apr 28;10(2):e35295. [doi: 10.2196/35295] [Medline: 35482373]
12. Ninaus M, Pereira G, Stefitz R, et al. Game elements improve performance in a working memory training task. IJSG 2015;2(1). [doi: 10.17083/ijsg.v2i1.60]
13. Mohammed S, Flores L, Deveau J, et al. The benefits and challenges of implementing motivational features to boost cognitive training outcome. J Cogn Enhanc 2017 Dec;1(4):491-507. [doi: 10.1007/s41465-017-0047-y] [Medline: 30221244]
14. Lumsden J, Edwards EA, Lawrence NS, Coyle D, Munafò MR. Gamification of cognitive assessment and cognitive training: a systematic review of applications and efficacy. JMIR Serious Games 2016 Jul 15;4(2):e11. [doi: 10.2196/games.5888] [Medline: 27421244]

XSL•FO
RenderX

15. Moon JW, Kim YG. Extending the TAM for a World-Wide-Web context. Information & Management 2001 Feb;38(4):217-230. [doi: 10.1016/S0378-7206(00)00061-6]
16. Hung SY, Tsai JCA, Chou ST. Decomposing perceived playfulness: a contextual examination of two social networking sites. Information & Management 2016 Sep;53(6):698-716. [doi: 10.1016/j.im.2016.02.005]
17. Fang X, Chan S, Brzezinski J, Xu S. A study of task characteristics and user intention to use handheld devices for mobile commerce. Presented at: The 2nd Annual Workshop on HCI Research in MIS; Dec 11-13, 2003; Seattle, WA, United States.
18. Lin CS, Wu S, Tsai RJ. Integrating perceived playfulness into expectation-confirmation model for web portal context. Information & Management 2005 Jul;42(5):683-693. [doi: 10.1016/j.im.2004.04.003]
19. Codish D, Ravid G. Academic course gamification: the art of perceived playfulness. IJELL 2014;10(1):131-151. [doi: 10.28945/2066]
20. Bertin J. Semiology of Graphics: Diagrams, Networks, Maps: University of Wisconsin Press:983.
21. Louis LEL, Moussaoui S, Van Langhenhove A, et al. Cognitive tasks and combined statistical methods to evaluate, model, and predict mental workload. Front Psychol 2023;14:1122793. [doi: 10.3389/fpsyg.2023.1122793] [Medline: 37251030]
22. Corsi PM. Human memory and the medial temporal region of the brain [PhD thesis]. : McGill University; 1972.
23. Kessels RP, van Zandvoort MJ, Postma A, Kappelle LJ, de Haan EH. The Corsi Block-Tapping Task: standardization and normative data. Appl Neuropsychol 2000;7(4):252-258. [doi: 10.1207/S15324826AN0704_8] [Medline: 11296689]
24. Vidal J, Mills T, Pang EW, Taylor MJ. Response inhibition in adults and teenagers: spatiotemporal differences in the prefrontal cortex. Brain Cogn 2012 Jun;79(1):49-59. [doi: 10.1016/j.bandc.2011.12.011] [Medline: 22325785]
25. Rubio S, Díaz E, Martín J, Puente JM. Evaluation of subjective mental workload: a comparison of SWAT, NASA - TLX, and workload profile methods. Applied Psychology 2004 Jan;53(1):61-86. [doi: 10.1111/j.1464-0597.2004.00161.x]
26. Paxion J. Complexite´ des situations, expe´rience, tension et vigilance: quels impacts sur la charge de travail et les performances de jeunes conducteurs. sciences cognitives [Thesis in French]. : Aix-Marseille Université; 2014.
27. Byers JC, Bittner AC, Hill SG. Traditional and raw task load index (TLX) correlations: are paired comparisons necessary? In: Mital A, editor. Advances in Industrial Ergonomics and Safety: Taylor and Francis; 1989:481-485.
28. Morey RD. Confidence intervals from normalized data: a correction to Cousineau (2005). TQMP 2008;4(2):61-64. [doi: 10.20982/tqmp.04.2.p061]
29. Heitz RP. The speed-accuracy tradeoff: history, physiology, methodology, and behavior. Front Neurosci 2014;8:150. [doi: 10.3389/fnins.2014.00150] [Medline: 24966810]
30. Donovan JJ, Hafsteinsson LG, Lorenzet SJ. The interactive effects of achievement goals and task complexity on enjoyment, mental focus, and effort. J Applied Social Pyschol 2018 Mar;48(3):136-149. [doi: 10.1111/jasp.12498]

Abbreviations

GLE: game-like element
JASP: Jeffrey's Amazing Statistics Program
MWL: mental workload
NASA-TLX: NASA-Task Load Index
PP: perceived playfulness
WP: workload profile
WP3: workload profile on spatial processing
WP4: workload profile on verbal processing
WP5: workload profile on visual processing
WP7: workload profile on manual output

Edited by A Mavragani; submitted 24.06.24; peer-reviewed by LK Chua, P Harvey; revised version received 07.02.25; accepted 08.02.25; published 10.04.25.

Please cite as:
Louis LE, Moussaoui S, Ravoux S, Milleville-Pennel I
Encouraging the Voluntary Mobilization of Mental Resources by Manipulating Task Design: Explorative Study
JMIR Form Res 2025;9:e63491
URL: https://formative.jmir.org/2025/1/e63491
doi:10.2196/63491

© Lina-Estelle Louis, Saïd Moussaoui, Sébastien Ravoux, Isabelle Milleville-Pennel. Originally published in JMIR Formative Research (https://formative.jmir.org), 10.4.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction

XSL•FO
RenderX

in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Support of Home-Based Structured Walking Training and Prediction of the 6-Minute Walk Test Distance in Patients With Peripheral Arterial Disease Based on Telehealth Data: Prospective Cohort Study

Fabian Wiesmüller[1,2,3], MSc; Andreas Prenner[4], Dr; Andreas Ziegl[5], PhD; Gihan El-Moazen[3], BSc; Robert Modre-Osprian[5], PhD; Martin Baumgartner[1,3], PhD; Marianne Brodmann[4], MD; Gerald Seinost[4], MD; Günther Silbernagel[4], MD; Günter Schreier[1], PhD; Dieter Hayn[1,2], PhD

[1]Center for Health & Bioresources, AIT Austrian Institute of Technology GmbH, Graz, Austria

[2]Ludwig Boltzmann Institute for Digital Health and Prevention, Ludwig Boltzmann Gesellschaft, Salzburg, Austria

[3]Institute of Neural Engineering, Graz University of Technology, Graz, Austria

[4]Division of Angiology, Department of Internal Medicine, Medical University of Graz, Graz, Austria

[5]telbiomed Medizintechnik und IT Service GmbH, Graz, Austria

Corresponding Author:
Fabian Wiesmüller, MSc
Center for Health & Bioresources, AIT Austrian Institute of Technology GmbH, Graz, Austria

Abstract

Background: Telehealth has been effective in managing cardiovascular diseases like stroke and heart failure and has shown promising results in managing patients with peripheral arterial disease. However, more work is needed to fully understand the effect of telehealth-based predictive modeling on the physical fitness of patients with peripheral arterial disease.

Objective: For this work, data from the Keep Pace study were analyzed in depth to gain insights on temporal developments of patients' conditions and to develop models to predict the patients' total walking distance at the study end. This could help to determine patients who are likely to benefit from the telehealth program and to continuously provide estimations to the patients as a motivating factor.

Methods: This work analyzes continuous patient-reported telehealth data, in combination with in-clinic data from 19 Fontaine stage II patients with peripheral arterial disease who underwent a 12-week telehealth-based walking program. This analysis granted insights into the increase of the total walking distance of the 6-minute walk tests (6MWT) as a measure for physical fitness, the steady decrease in the patients' pain, and the positive correlation between well-being and the total walking distance measured by the 6MWT.

Results: This work analyzed trends of and correlations between continuous patient-generated data. Findings of this study include a significant decrease of the patients' pain sensation over time (P=.006), a low but highly significant correlation between pain sensation and steps taken on the same day (r=−0.11; P<.001) and the walking distance of the independently performed 6MWTs (r=−0.39; P<.001). Despite the reported pain, adherence to the 6MWT measurement protocol was high (85.53%). Additionally, patients significantly improved their timed-up-and-go test times during the study (P=.002). Predicting the total walking distance at the study end measured by the 6MWT worked well at study baseline (root mean squared error of 30 meters; 7.04% of the mean total walking distance at the study end of 425 meters) and continuously improved by adding further telehealth data. Future work should validate these findings in a larger cohort and in a prospective setting based on a clinical outcome.

Conclusions: We conclude that the prototypical trend estimation has great potential for an integration in the telehealth system to be used in future work to provide tailored patient-specific advice based on these predictions. Continuous data from the telehealth system grant a deeper insight and a better understanding of the patients' status concerning well-being and level of pain as well as their current physical fitness level and the progress toward reaching set goals.

Trial Registration: ClinicalTrials.gov Identifier: NCT05619835; https://tinyurl.com/mrxt7y9u

(JMIR Form Res 2025;9:e65721) doi:10.2196/65721

XSL•FO
RenderX

KEYWORDS

mHealth; telehealth; peripheral arterial disease; home-based structured walking training; trend estimation; predictive modeling; continuous data; walking; walking training; prediction; prediction model; cardiovascular disease; stroke; heart failure; physical fitness; telehealth system

Introduction

Background

Peripheral arterial disease is a significant global health concern with an estimated 202 million people affected worldwide [1]. Prevalence in Europe is reported at 5.3%, increasing to 15% - 20% among individuals over the age of 70 years [2]. Peripheral arterial disease is characterized by insufficient oxygen supply due to arterial vessel alterations in the extremities. This leads to pain when walking, forcing patients to stop walking and thus affecting mobility and quality of life. Pain may occur even at rest. Furthermore, the healing of wounds is slowed down, and peripheral arterial disease may ultimately lead to necrosis. Peripheral arterial disease is classified into distinct grades according to the Rutherford system, which includes 6 categories, and the Fontaine classifications, which include 4 stages. Evidence-based guidelines for the management of peripheral arterial disease are provided by organizations such as the European Society of Cardiology, European Society of Vascular Medicine, and American Heart Association [3-5].

Treatment for Fontaine stage I is typically limited to risk factor management (eg, nicotine abstention and blood pressure management), which remains a relevant factor throughout all following stages. In Fontaine stage II, medication regimen and structured walking training (SWT) become additional treatment options. While medication is also recommended throughout stages III and IV, SWT is only prescribed in stages II and III. Stages III and IV have surgical interventions as an additional treatment option [3,4].

SWT can more than double the pain-free walking distance of patients with peripheral arterial disease within a 12-week program [6,7], and achieve long-term outcomes comparable to surgical intervention (revascularization) while being less expensive and non-invasive [4,8]. To further foster SWT in patients with peripheral arterial disease, the Trans-Atlantic Inter-Society Consensus-II group has defined a training protocol for patients. This recommendation suggests performing walking exercises at an intensity that induces pain for 3 - 5 minutes in total within a 30 - 60-minute session. Such a session should be conducted 3 times per week for 3 months [9]. To evaluate the progress of SWT and to assess the physical fitness of patients with peripheral arterial disease, standardized tests have been developed.

The state-of-the-art method for measuring the physical fitness of patients with peripheral arterial disease is the 6-minute walk test (6MWT), during which a patient walks as far as possible within 6 minutes on an incline-free surface [10]. Usually, the 6MWT is performed under the supervision of a health care professional (HCP) who measures the total walking distance after the completion of the test [11]. Another test to measure physical fitness is the timed-up-and-go (TUG) test. The TUG test starts with the participant sitting on a chair. The participant is then instructed to stand up, walk for 3 meters, turn around, walk back, and sit back down. The time the participant needs to complete this sequence of tasks is measured [12]. Even though TUG tests are not primarily used for patients with peripheral arterial disease, they provide information about the patients' general mobility and frailness. Additionally, a previous publication identified a significant correlation between the 6MWTs and TUG tests in patients with peripheral arterial disease [13].

Telehealth is defined as the delivery of health care or rehabilitation services to patients remotely using digital technologies [14], which, over the last years, has proven to be effective in the field of cardiovascular diseases, particularly heart failure and stroke [15,16]. Telehealth and telerehabilitation solutions have recently also shown promising results in the field of peripheral arterial disease [17,18]. Telehealth can support early detection of (post-operative) complications, with the potential to improve functional capacity, claudication onset time, and quality of life for patients with peripheral arterial disease [19].

Despite these studies, the benefits and the potential of telehealth in the setting of peripheral arterial disease are not yet fully explored. Aspects like the influence of continuously measured parameters and digital tools on the physical fitness of patients with peripheral arterial disease must be further investigated.

Keep Pace Study Protocol

During the Keep Pace study, 19 patients with peripheral arterial disease underwent a 12-week program with digital assistance to evaluate the impact of telehealth solutions on patients with peripheral arterial disease with a focus on the decision of undergoing surgery after the program [20]. The participants for the Keep Pace study were recruited at the outpatient clinics of the Department of Angiology of the University Hospital Graz, Austria. Patients who were shortlisted for getting an angioplasty were contacted and asked to participate in the Keep Pace study. Patients who gave informed consent and fit the inclusion criteria were included in this study. The inclusion criteria were as follows: age between 18 and 80 years, peripheral arterial disease in Fontaine stage IIa or IIb, ownership of an Android mobile device, and capability to use telehealth services. Patients lacking compliance to perform SWT were excluded. At the onset of the study, an initial in-clinic visit was performed, during which the patient's medical history and subjective limitations in daily activities, including pain-free and total walking distance, were assessed. All patients were diagnosed with peripheral arterial disease based on imaging diagnostics before study entrance. Because the location and grade of the occlusion were not part of the analysis, all diagnostic imaging options, such as duplex ultrasound, computer tomography, and magnetic resonance angiography, were accepted, which did not provide information about the location and grade of the occlusion for all patients. In addition to the imaging peripheral arterial disease criteria,

XSL•FO
RenderX

claudication had to be present for patients to be eligible for this study. In the following 8 weeks participants were telemedically supervised by a nurse. In the subsequent remaining 4 weeks, patients performed their SWT therapy independently without regular telehealth consultations. Thereafter, an in-clinic study end visit was performed [20] .

During the Keep Pace study, the collected telehealth data have been used to continuously monitor the patients' status. They might also be used for predictive modeling applications, eg, to detect already at early stages, which patients will benefit from the training and which patients might require additional attention. It is expected that such a prediction could help to achieve a more patient-specific therapeutic regime, tailored exercise support, and expectation management. However, currently, there is no evidence to support this hypothesis. This work takes an in-depth look at the telehealth data recorded during the Keep Pace study [20] by focusing on analyzing trends in patient-reported timeseries data.

Objective

The aim of this investigation was to test the feasibility of predicting the total 6-minute walk test distance at the study end in patients with peripheral arterial disease, which can be used to determine patients who will likely benefit more from the telehealth program than others and additionally provide patient-specific feedback and training regime adaptions throughout the patients' time in the telehealth system.

Methods

Dataset

For this work, data of 19 patients with peripheral arterial disease in the Fontaine stage II were analyzed, which were collected over a 12-week study period in the Keep Pace study [20]. While the study protocol only specified that participants had to be patients with peripheral arterial disease in Fontaine stage II, the actual patient cohort was unanimously in Fontaine stage IIb. The characteristics of the patients are shown in Table 1.

Data from 2 sources were collected during Keep Pace and further processed in this analysis: (1) data assessed in-clinic by HCPs at the beginning and at the end of the study (Textbox 1) and (2) patient-reported data collected with a telehealth system during the 12-week study period (Textbox 2).

Table . Patient characteristics by gender and in total (SD) [20].

Characteristics	Female (n=7)	Male (n=13)	Total (n=20)
Age, years, mean (SD)	64.14 (8.68)	62.46 (5.39)	63.05 (6.55)
Weight, kg, mMean (SD)	70.14 (9.5)	81.67 (11.95)	77.64 (12.27)
Height, m, mean (SD)	1.65 (0.03)	1.76 (0.05)	1.72 (0.07)
BMI, kg/m²	25.79 (3.44)	26.33 (4.51)	26.15 (4.08)
Subjective pain-free walking distance, m, mean (SD)	73.57 (50.22)	81.54 (37.16)	78.75 (41.04)
Fontaine stage IIa, n	0 (0%)	0 (0%)	0 (0%)
Fontaine stage IIb, n	7 (100%)	13 (100%)	20 (100%)
Diabetics, n (%)	1 (14%)	2 (15%)	3 (15%)
Hypercholesterolemia, n (%)	7 (100%)	13 (100%)	20 (100%)
Smokers, n (%)	2 (28%)	10 (77%)	12 (60%)
Coronary artery disease, n (%)	3 (43%)	4 (31%)	7 (35%)

Textbox 1. In-clinic data.

Patient characteristics: sex, height, age, weight, disease-related data like the date of diagnosis, cardiovascular risk factors, current medication, diabetes and smoking assessment and year of diagnosis

6MWTs: manual measurement by an HCP and automated measurement with a dedicated 6MWT app [21]

TUG test data: automated measurement with a TUG test device as described in [21,22], including the total TUG time and sub-task times (stand-up, walk-forward, turnaround, walk-back, sit-down, total walk time, and mobility time)

Textbox 2. Patient-reported telehealth data.

6MWTs: automated measurement with a 6MWT app [21] — recommended once per week

Steps per day: automatically measured by the mobile phone

Subjective well-being: recommended once per day

Pain: recommended once per day

XSL•FO
RenderX

The telehealth data described in (Textbox 2) were gathered by the patients using a dedicated mobile application. Patients did not receive an incentive for completing the 6MWTs and the SWT apart from the inherent incentive of improving their physical fitness and quality of life by adhering to the measurement and training regime. Subjective well-being was assessed on a 3-item scale (good, medium, and bad) and pain was reported on a scale (0=no pain to 10=maximum pain). The steps per day were recorded with the smartphones' native step counter, and patients could perform unsupervised 6MWTs using an integrated previously developed GPS-based 6MWT-app [21].

Predictive Modeling

We investigated different models for predicting the total walking distance at the study end right at the onset of the program (baseline) and at week *n* of the program, only based on data recorded until week *n*. A leave-one-out setting was applied, which excluded 1 patient from the training process and used this patient as a validator for the developed algorithm, resulting in 19 different data splits per tested model. This ensures that each patient is used as the validator once per model. The quality of the prediction as compared to the measured data at the study end was calculated using the root mean squared error (RMSE). The following predictive models were applied.

Distance at Baseline

This approach was mainly used as a reference for other models. The 6MWT distance at the study end was estimated as equal to the 6MWT distance at baseline (ie, this model assumed no improvement during the study).

Average Improvement Based on All Patients

This model assumed that all patients had similar improvements. Therefore, for estimating the 6MWT distance at the study end of patient *i* in week *n* ($d_{av,n,i}$), the mean change of all other patients from week *n* to the study end was calculated and added to the respective patient's 6MWT distance $d_{n,i}$ in week *n* (equation 1).

(1)dav,n,i= dn,i+ 118∑j=1j≠i19(dend, j−dn,j)

Logarithmic Interpolation Based on an Individual Patient

Logarithmic interpolation was used to fit a logarithmic function as defined by the parameters $a_{n,i}$ and $b_{n,i}$ into all measured 6MWT distances $d_{k,j}$ performed at week $x_{k,j}$ until week *n* by the respective patient *i*. The estimated distance at the study end as derived in week *n* for patient *i* ($d_{log_ind,n,i}$) was calculated according to equation 2.

(2)dlog_ind,n,i=(an,i*loge x+bn,i)|x=12

with[an,i,bn,i]=logfit(xk,j,dk,j) |∀k≤n, j=i

Logarithmic Interpolation Based on All Patients

Assuming that all patients improved similarly, logarithmic interpolation was also applied to data from all patients when defining parameters a_n and b_n in week *n*. Therefore, all measurements $d_{k,j}$ were first corrected for the mean of each patient (dk,j−) (see equation 3).

(3)dlog_all,n,i=(an,i*loge x+bn,i)|x=12

with[an,bn]=logfit(xk,j,dk,j−) |∀k≤n,∀j

Linear Regression

All models described earlier are based on 6MWT distances only. With the linear regression approach, also other data assessed during the program were considered. Feature candidates applied for predictions in week *n* were derived from all data recorded during the in-clinic visit at baseline (see In-clinic data in the Methods section) and from telehealth data recorded until week *n*. Telehealth data contained the feature candidates, as shown in Textbox 3:

All possible combinations of up to 8 features were evaluated. Therefore, each feature was first tested on its own, and subsequently features were added iteratively to test all possible combinations. However, due to the small sample size, a maximum number of 4 features was the goal.

Textbox 3. Feature candidates from telehealth data.

Mean 6MWT distance up to week *n*: this parameter represents all total walking distance measurements from the 6MWT app from the baseline up to the week of the prediction.
Mean 6MWT distance in week *n*: this parameter considers only measurements of 6MWTs performed in the week of the prediction.
Logarithmic prediction in week *n*: this feature is derived from the trend-based analysis, where a predicted outcome is available for each week.

Statistical Analyses

All statistical analyses were implemented in MATLAB (The MathWorks) and all predictive models were either implemented in MATLAB or Python. Wherever reasonable, pre- and post-study comparisons were performed on data assessed at the in-clinic visits at baseline and the study end, using paired tests. Continuously measured parameters were tested for correlation with one another. Unless stated otherwise, correlations were tested using the Pearson's correlation coefficient with a previous Anderson-Darling test to test for a normal distribution of the data. Trend analyses were performed for all the continuous measurements mentioned earlier for which the data were aggregated on a weekly basis by calculating mean values per parameter. Predictive models were evaluated by calculating the root mean square error between the estimated 6MWT total walking distance at the study end and the actually measured distance. For this work, a *P* value <.05 was considered to be significant.

Ethical Considerations

Ethical approval for the study was obtained from the Ethics Committee of the Medical University of Graz (34‑127 ex 21/22 1566‑2021, ClinicalTrials.gov Identifier: NCT05619835). All analyses performed during this work are covered by this ethics approval. All participants received oral

XSL•FO
RenderX

and written information prior to the study entry and provided written informed consent to participate in this study. This informed consent also covers the analysis of secondary study outcomes performed in this paper. All data were recorded in a pseudonymized way and anonymity was ensured for all participants during the presentation of the findings. Patients did not receive any financial incentives for participating in the study.

Results

Data Characteristics and Trend Analysis

The 6-Minute Walk Test

The study protocol recommended 12 6MWTs. 57.89% (11 out of 19) patients performed 12 or more tests and were 100% adherent to the measurement protocol. The 8 remaining patients performed, on average, 7.88 (SD 2.36) 6MWTs and hence achieved a mean adherence of 65.62% (SD 19.64%) of the 12 recommended tests. On average, patients performed 13.37 (SD 8.90) 6MWTs and overall adherence to the measurement routine was 85.53% (SD 21.31%).

Figure 1 illustrates the 6MWT total walking distance aggregated on a weekly basis for all patients. The total walking distance increased over time, following a logarithmic trend.

As can be seen in Table 2, a significant correlation was found between the 6MWT total walking distance at baseline versus the distance at the study end. However, no significant correlation between either the number of tests and the distance at baseline, or the number of tests and the distance at the study end, or the number of tests and the change in walking distance during the program, or between the distance at baseline and the change in walking distance during the program has been identified.

Figure 1. Course over time of weekly aggregated 6-minute walk test (6MWT) total walking distances for all patients, illustrated as lines per patient and boxplots. Data are normalized for the patient-individual mean. Dashed lines illustrated interpolations over weeks with missing values. In-clinic measurements of the first and last visit are shown in dark blue; tests measured with the 6MWT app are shown in azure. The number n of patients who performed at least 1 6MWT in the respective week is shown on the bottom.

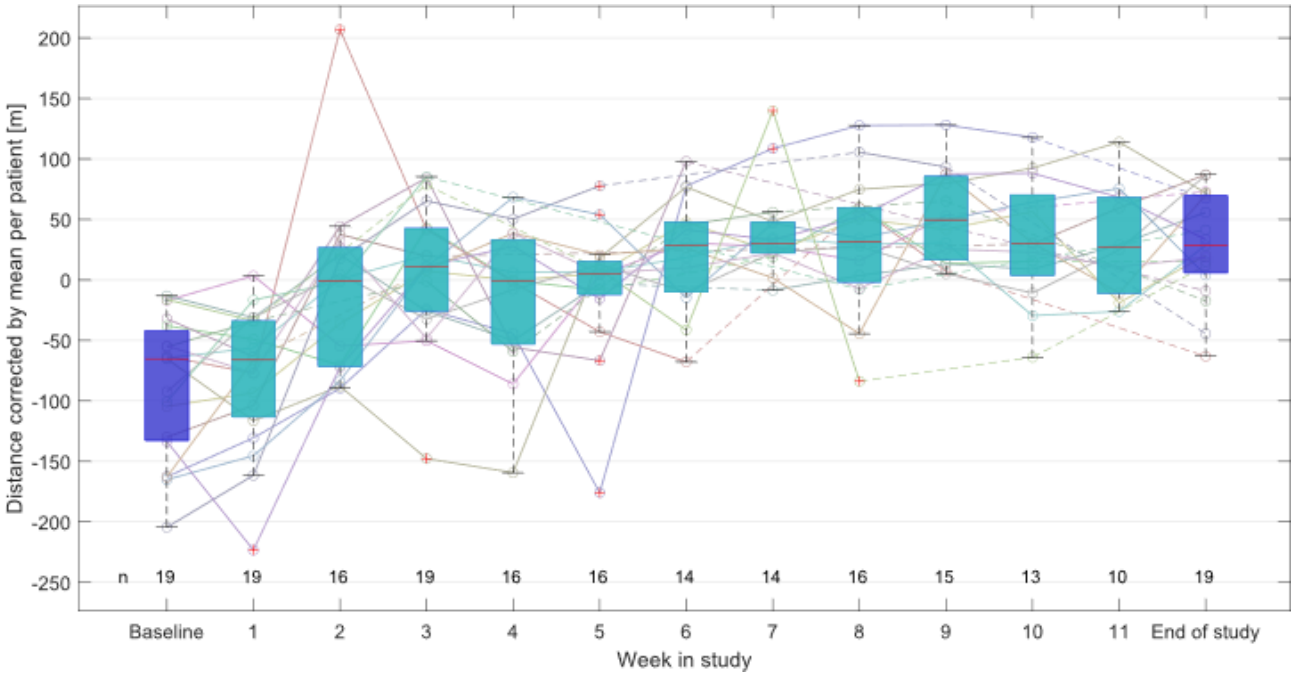

Table . Correlation between the final walking distance and the number of tests and the change in the walking distance between baseline and the study end (R: correlation coefficient; PC: Pearson correlation coefficient; S: Spearman correlation coefficient).

	Distance at the study end	Number of tests	Change in walking distance
Distance at baseline	*r*=0.78, *P*<.001, S	*R*=–0.14, *P*=.56, S	*r*=0.02, *P*=.93, S
Distance at the study end	N/A	*r*=−0.17, *P*=.48, S	*r*=0.71, *P*<.001, S
Number of tests	*r*=−0.13, *P*=.60, PC	N/A	*r*=0.02, *P*=.93, S

Timed-Up-and-Go Test

Table 3 summarizes all TUG subtask measures at baseline and the study end. Table 3 shows that the sit-down time significantly decreased from 2.21 to 1.89 seconds (*P*=.038), and the total walk time significantly decreased from 6.96 to 6.57 seconds (*P*=.003). Additionally, the overall TUG time significantly decreased from 10.6 seconds at baseline to 9.89 seconds at the study end (*P*=.002). Also, all other parameters decreased but did not reach statistical significance.

Figure 2 shows the decrease of the total TUG time.

XSL•FO
RenderX

Table . TUG subtask times (in seconds; median (IQR)) at baseline and the study end with the *P* values calculated with the 2-tailed Wilcoxon test for matched pairs.

	Baseline, median (IQR)	Study end, median (IQR)	*P* value
Stand-up	2.67 (2.34-4.44)	2.54 (2.10-3.00)	.09
Walk-forward	1.95 (1.70-2.21)	1.76 (1.69-2.10)	.06
Turnaround	1.82 (1.43-2.34)	1.56 (1.15-2.08)	.11
Walk-back	2.09 (1.95-2.35)	1.95 (1.82-2.21)	.06
Sit-down	2.21 (2.10-2.87)	1.89 (1.56-2.60)	.04
Total walk time	6.96 (5.98-9.78)	6.57 (5.33-7.42)	.003
Mobility time	4.10 (3.77-4.43)	3.64 (3.51-4.04)	.08
Total TUG time	10.6 (9.63-12.25)	9.89 (8.58-11.59)	.002

Figure 2. Improvement of the timed-up-and-go test time at baseline versus the study end (n: number of observations; MED: median; IQR: interquartile range).

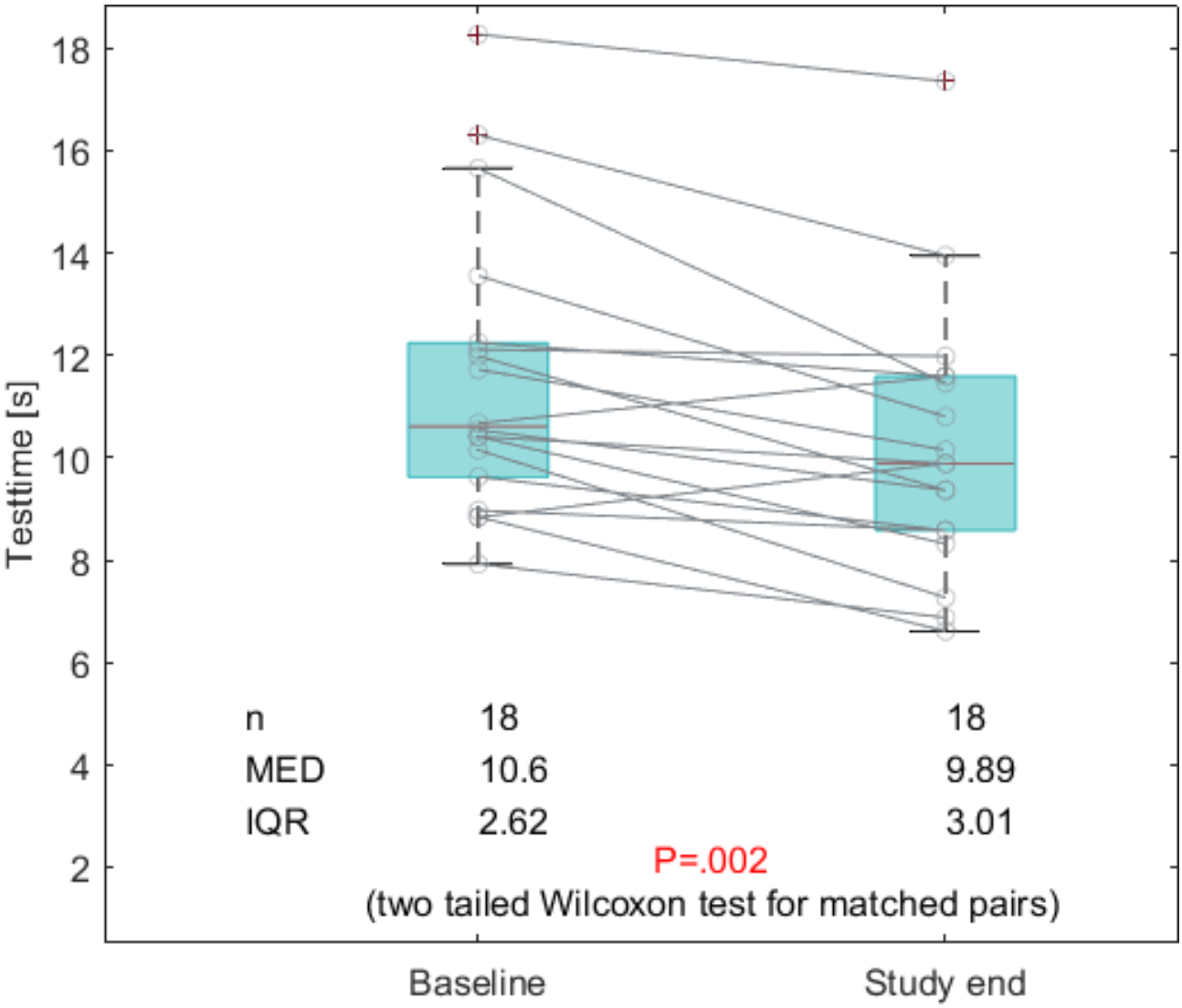

Steps Per Day

The mean number of steps per day as measured by the smartphone's native app was 4176.58 (SD 2742.73). No significant change in steps per day was found when aggregating over all patients. However, 3 patients showed a significant reduction of steps, and 4 patients significantly improved. No relation between the change of recorded steps and the decision of undergoing surgery after the program was observed.

Pain

Overall, patient-reported pain was significantly reduced from a mean of 4.67 (SD 1.90) at baseline to 3.60 (SD 2.39) at the study end ($P<.05$). Additionally, pain showed a very low but

XSL•FO
RenderX

highly significant correlation with steps per day (r=−0.11; P<.001) and the distance of the 6MWTs (r=−0.39; P<.001) on days when patients submitted both measurements. Figure 3 illustrates the course over time of weekly pain values corrected for the mean per patient.

Figure 3. Course over time of weekly aggregated pain measurements for all patients, illustrated as lines per patient and boxplots. The level of pain is normalized for the patient-individual mean. Dashed lines illustrated interpolations over weeks with missing values. The number of patients who reported at least 1 pain level in the respective week is shown on the bottom.

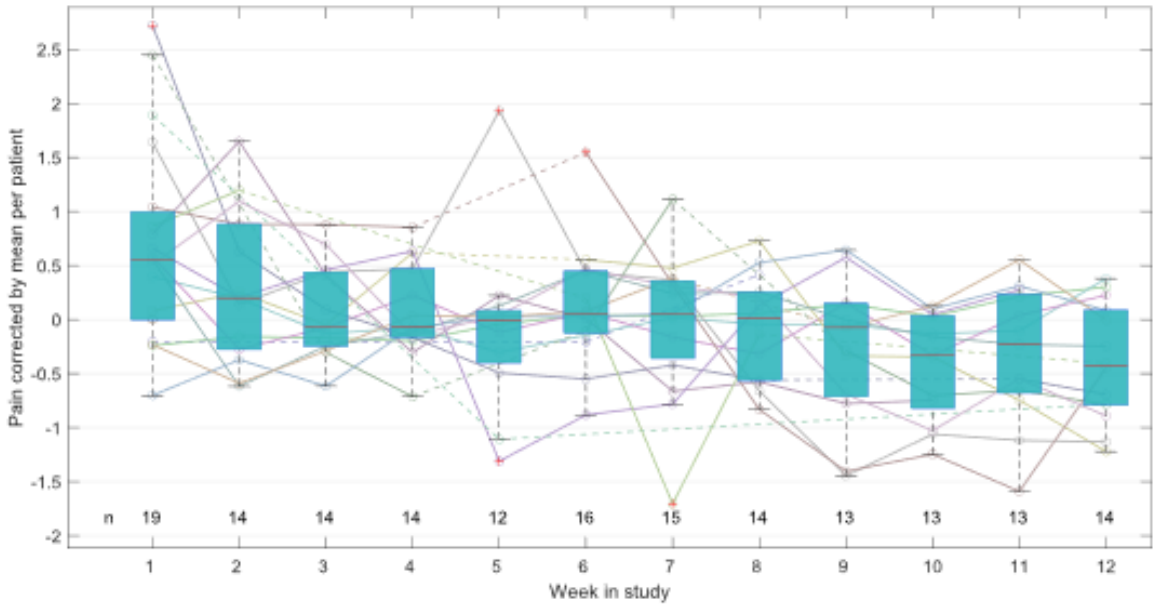

Subjective Well-Being

Out of 1597 subjective well-being data points, 41 (2.63%) were reported as bad, 686 (43.0%) as medium, and 870 (54.5%) as good. No trends or patterns in the development of subjective well-being and no significant change from baseline to the study end were identified. On 5 days with bad, 160 with medium, and 106 with good well-being, 6MWTs were performed by the respective patient. On days with a submitted subjective well-being and a 6MWT, the number of bad well-being was 1.88%, which is in the same magnitude as the overall bad well-being submissions. Subjective well-being showed a significant correlation with the 6MWT total walking distance (r=0.26, P<.001, Spearman). On days with subjective well-being and step count data available (40 bad, 657 medium, and 835 good), a weak but highly significant correlation between subjective well-being and steps per day was found (r=0.19; P<.001).

Trend Prediction and Predictive Modeling

Table 4 shows the RMSEs between estimated and measured 6MWT total walking distance at the study end (week 12), depending on the week of estimation and the respective model. The distance at the baseline model, which estimated the distance at the study end equal to the distance at baseline, revealed an error of 136 meters. The average improvement in the week n method achieved an error of 109 meters at baseline and a continuous decrease of the error up to week 7 with an error of 43 meters. The logarithmic interpolation with a patient individual fit was not able to outperform the improvement in the week n method. However, the logarithmic fit on all patients was able to outperform the improvement in the week n method from week 1 to week 6. The linear regression model used data from the baseline of the study in combination with the time-dependent features, described in the Methods ection under the subsection Predictive modeling. The linear regression model continuously decreased the prediction error from week to week since the time-dependent features became more proximate to the outcome. For this analysis, the number of features was set to be between 1 and 4, whichever combination performed the best according to Table 4. The linear regression was able to outperform the logarithmic interpolation in every week. The error of 30 meters achieved by the linear regression model at baseline was 7.04% the mean total walking distance at the study end of 425 meters. The progress of the estimation errors over time is shown in Figure 4.

XSL•FO
RenderX

Table . Root mean square error (m) between the estimated 6-minute walk test (6MWT) total walking distance at the study end and measured distance, as achieved with different models based on the available data in the corresponding week.

Model	Baseline	Week										
		1	2	3	4	5	6	7	8	9	10	11
Distance at baseline	136	—[a]	—	—	—	—	—	—	—	—	—	—
6MWT data only												
Average improvement in week n	109	112	87	69	59	73	52	43	54	63	51	9
Logarithmic interpolation for individual fit	—	233	141	91	87	82	67	62	61	65	58	60
Logarithmic interpolation for all patient fit	—	50	59	50	50	48	48	48	49	50	50	49
6MWT + other data												
Linear regression	30	30	29	29	29	29	25	25	22	22	14	14

[a]Not applicable.

XSL•FO
RenderX

Figure 4. Course over time of the root mean square error (RMSE) between estimated and measured 6-minute walk test total walking distance, depending on the week of estimation, per model.

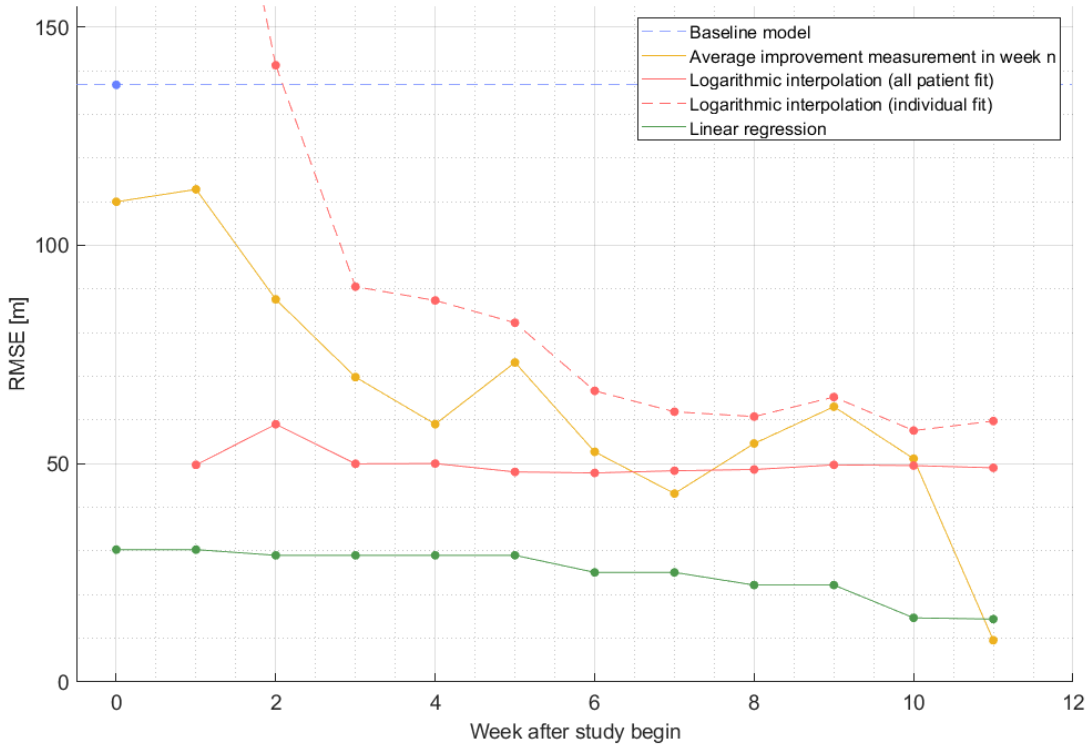

Table 5 shows the RMSE of models using data assessed in-clinic at the baseline of the study. Only the RMSE of the best-performing feature or feature combination is shown in the table. The ideal number of features for the linear regression was 4 (distance at the onset of the study, the patients' sex, the difference between subjectively estimated distance until the onset of pain and the actual distance, and the ratio between the 2 TUG tests, both performed at the onset).

Table . Root mean square errors (m) of the estimated 6-minute walk test total distance as compared to the actually measured distance in week 12, based on the number of features at the baseline.

Model	Number of features							
	1	2	3	4	5	6	7	8
Linear regression	52.08	38.89	31.10	30.31	30.63	30.59	31.90	33.23

Discussion

Principal Results

In the current work, an in-depth analysis of previously not analyzed data recorded during the Keep Pace Study was performed, in which 19 patients with peripheral arterial disease performed a 12-week walking program with telemonitoring assistance. In addition to the improved pain-free walking distance as previously published in [20], we identified significant improvements in physical activity scores during the study, based on both the TUG test and the 6MWT. The regular recording of 6MWTs by the patients at home in combination with the telehealth service allowed HCPs to monitor the individual success of the SWT in real time. Based on these data (Figure 1), we identified that the majority of the 6MWT improvement happened in the first 3 weeks of the program, while saturating thereafter, which is in good accordance with the training effect in other studies, even though this stagnation of improvement happens later in healthy individuals [23]. We also identified a significant reduction of pain during the program. In contrast to the training effect, this improvement did not show any saturation but spread until the end of the program.

Despite the pain that was reported by the patients, the overall adherence to the 6MWT measurement protocol was high (85.53%). Unfortunately, we were only able to measure the adherence to the test protocol and not the adherence to the actual SWT, which was not recorded during the Keep Pace study. Since we had no insight into whether the patients continuously wore their smartphones with the step count, we could not use the daily steps to draw conclusions about the adherence to the WT. However, based on the high adherence to the measurement protocol and the overall improvement in walking distance, we assume that most patients were adherent to the training program, too. Neither adherence nor improvement during the program

XSL•FO
RenderX

was dependent on the initial 6MWT distance, which indicates the program's potential across all baseline fitness levels.

While both the 6MWT distance and the TUG time improved during the program, no significant increase in the number of steps per day, as another measurement of the daily physical activity, was observed. However, since steps were tracked by the smartphones' native apps only, and since smartphones are not expected to be carried continuously, the number of steps might not be representative.

The TUG test is currently not a primary assessment tool for physical fitness in patients with peripheral arterial disease. However, TUG times and 6MWT distances are highly correlated in patients with peripheral arterial disease. Since the TUG test takes only approximately 20 - 30 seconds (as compared to the 6 minutes of the 6MWT), it is much less burdensome and time-consuming. Further studies are required to analyze whether parts of the 6MWT can be replaced by TUG tests or if TUG data can provide additional insights in combination with 6MWTs to optimize peripheral arterial disease management.

The data showed that the physical activity and fitness levels decreased with an increase in the pain sensation, as to be expected since pain is a limiting factor for physical activity. Similarly, the general well-being correlated with the outcome of the 6MWTs. However, due to the highly unequal distribution of subjective well-being on days on which a 6MWT was performed, the correlation between well-being and 6MWT total distance must be interpreted with care. Additionally, using a scale from 1 to 3 for measuring the well-being might result in inaccurate depictions due to a lack of granularity in the reporting compared to a continuous scale from 0 to 10 which was used for the measurement of pain. This lack of granularity might explain why no significant change in the well-being during the study has been identified.

The developed predictive models highlighted that—based on telehealth data—the outcome can be predicted with an error of 30 meters at baseline, which corresponds to 7.04% of the mean total walking distance at the study end. Whether this prediction is accurate enough to help adjust the training to the individual patient or to apply patient-specific measures, such as additional phone calls, must be evaluated separately in a prospective setting. With the modeling approaches, we have shown 2 potential use cases, where (1) a prediction at study baseline can be done accurately to identify patients with great potential for the system and (2) continuous, weekly predictions to give patient-specific feedback. A prediction of the final walking distance as in case (1) could be used to not only identify patients with high potential to benefit from the system but also to motivate them to join the system by providing them with a probable result if they were adherent to the program. The predictions of (2) could be used to provide patient-specific feedback and training regime adaptions.

Limitations

The results of this study must be interpreted with care since our dataset contained only 19 patients. Additionally, the study population was very homogeneous with only patients with peripheral arterial disease of Fontaine stage IIb since this stage benefits the most from SWT. While this work showed that predictive modeling and trend estimation work well for patients with Fontaine stage IIb, future work must validate these findings for other stages of peripheral arterial disease. Additionally, with a rather short follow-up time of 90 days, the effects of the telehealth solution on the long-term adherence to the walking program could not be evaluated.

These limitations apply especially for the predictive models. Therefore, we applied a leave-one-out train-test split to maximize the available training data and limited the number of features for the models. In general, the trend estimation worked well compared to the distance-at-baseline method but showed great differences in the RMSE between the methods.

As expected, telehealth data can improve the accuracy of models that predict 6MWT distances at the study end. Simple models showed high instability when applied on data of the first weeks only, especially when based on individual patients' data only. However, even these models improved when telehealth data from further weeks were included. The lowest error of all models was achieved with the average improvement model applied in week 11. However, this approach performed worse in previous weeks, and predictions in week 11 might be of limited interest in a real-world application. Additionally, we assume that this extraordinarily good result might have been achieved by chance, due to the small sample size. The algorithms used for predicting the total walking distance at the study end were low in complexity due to the small study size of only 19 patients. A random forest was also tested as a possible approach but was outperformed by linear regression in every instance. More complex models like the random forest or support vector machines, or even (deep) neuronal networks, might perform even better. However, much larger datasets will be required to train these models, even if transfer learning is applied.

Future Work

Future work should validate the effect of predictive modeling in supporting home-based structured walking training in patients with peripheral arterial disease with (a) a larger sample size, (b) prospectively, and (c) based on clinical outcomes compared to a control group only using the telehealth system without continuous predictive modeling support. Finally, our results need to be integrated into real-world telehealth settings, ideally in a prospective study to evaluate the impact of informed decision-making by the HCP based on recorded telehealth data and on the results of the predicted outcomes. This approach might not only help HCPs provide patient-specific care but could also help the patients with their motivation by seeing the progress and the prediction of what they could achieve if they stayed adherent to the walking program.

Conclusions

This paper shows that the prototypical trend estimation has great potential for an integration in the telehealth system to be used in future work to provide tailored patient-specific advice based on these predictions. Additionally, the continuous data from the telehealth system grant a deeper insight and a better understanding of the patients' status concerning well-being and

XSL•FO
RenderX

level of pain as well as their current physical fitness level and the progress towards reaching set goals.

Acknowledgments

Supported by TU Graz Open Access Publishing Fund.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to privacy concerns and based our ethics approval but are available from the corresponding author on reasonable request.

Authors' Contributions

FW had the main responsibility for preparing the final manuscript and for the statistical analyses. AP was responsible for preparing and conducting the KEEP PACE study and interpretation of the results. AZ contributed to the study design, study preparation, statistical analyses, and interpreting the results. GEM contributed to conducting the study preparation and interpretation of the results. RMO contributed to the study preparations and interpretation of the results. MB contributed to the interpretation of the results. MB contributed to the study design. GS contributed to the study preparation and helped conducting the study. GS contributed to the study preparation and helped conducting the study. GS contributed to preparing the final manuscript, interpreting the results, and the statistical analyses. DH contributed to preparing the final manuscript and the statistical analyses. All authors discussed the results, supervised the findings, and contributed to the article and approved the submitted version. No funding was requested for this project. It was a self-financed project of Meduni Graz and AIT within the framework of a dissertation. Generative artificial intelligence has been used to improve grammar and language during the writing of this paper.

Conflicts of Interest

None declared.

References

1. Song P, Rudan D, Wang M, Chang X, Rudan I. National and subnational estimation of the prevalence of peripheral artery disease (PAD) in China: a systematic review and meta-analysis. J Glob Health 2019 Jun;9(1):010601. [doi: 10.7189/jogh.09.010601] [Medline: 30873278]
2. Olinic DM, Spinu M, Olinic M, et al. Epidemiology of peripheral artery disease in Europe: VAS Educational Paper. Int Angiol 2018 Aug;37(4):327-334. [doi: 10.23736/S0392-9590.18.03996-2] [Medline: 29936722]
3. Aboyans V, Ricco JB, Bartelink M, et al. 2017 ESC Guidelines on the Diagnosis and Treatment of Peripheral Arterial Diseases, in collaboration with the European Society for Vascular Surgery (ESVS): document covering atherosclerotic disease of extracranial carotid and vertebral, mesenteric, renal, upper and lower extremity arteries. Endorsed by: the European Stroke Organization (ESO)The Task Force for the Diagnosis and Treatment of Peripheral Arterial Diseases of the European Society of Cardiology (ESC) and of the European Society for Vascular Surgery (ESVS). Eur Heart J 2018 Mar 1;39(9):763-816. [doi: 10.1093/eurheartj/ehx095] [Medline: 28886620]
4. Frank U, Nikol S, Belch J, et al. ESVM Guideline on peripheral arterial disease. Vasa 2019 Sep;48(Suppl 102):1-79. [doi: 10.1024/0301-1526/a000834] [Medline: 31789115]
5. Gerhard-Herman MD, Gornik HL, Barrett C, et al. 2016 AHA/ACC guideline on the management of patients with lower extremity peripheral artery disease: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. J Am Coll Cardiol 2017 Mar 21;69(11):e71-e126. [doi: 10.1016/j.jacc.2016.11.007] [Medline: 27851992]
6. Murphy TP, Cutlip DE, Regensteiner JG, et al. Supervised exercise versus primary stenting for claudication resulting from aortoiliac peripheral artery disease: six-month outcomes from the claudication: exercise versus endoluminal revascularization (CLEVER) study. Circulation 2012 Jan 3;125(1):130-139. [doi: 10.1161/CIRCULATIONAHA.111.075770] [Medline: 22090168]
7. Murphy TP, Cutlip DE, Regensteiner JG, et al. Supervised exercise, stent revascularization, or medical therapy for claudication due to aortoiliac peripheral artery disease: the CLEVER study. J Am Coll Cardiol 2015 Mar 17;65(10):999-1009. [doi: 10.1016/j.jacc.2014.12.043] [Medline: 25766947]
8. Nicolaï SPA, Teijink JAW, Prins MH, Exercise Therapy in Peripheral Arterial Disease Study Group. Multicenter randomized clinical trial of supervised exercise therapy with or without feedback versus walking advice for intermittent claudication. J Vasc Surg 2010 Aug;52(2):348-355. [doi: 10.1016/j.jvs.2010.02.022] [Medline: 20478681]
9. Norgren L, Hiatt WR, Dormandy JA, et al. Inter-Society Consensus for the Management of Peripheral Arterial Disease (TASC II). Eur J Vasc Endovasc Surg 2007;33 Suppl 1:S1-75. [doi: 10.1016/j.ejvs.2006.09.024] [Medline: 17140820]

XSL•FO
RenderX

10. Nordanstig J, Behrendt CA, Baumgartner I, et al. Editor's Choice -- European Society for Vascular Surgery (ESVS) 2024 Clinical Practice Guidelines on the Management of Asymptomatic Lower Limb Peripheral Arterial Disease and Intermittent Claudication. Eur J Vasc Endovasc Surg 2024 Jan;67(1):9-96. [doi: 10.1016/j.ejvs.2023.08.067] [Medline: 37949800]
11. Shu J, Santulli G. Update on peripheral artery disease: epidemiology and evidence-based facts. Atherosclerosis 2018 Aug;275(379-81):379-381. [doi: 10.1016/j.atherosclerosis.2018.05.033] [Medline: 29843915]
12. Podsiadlo D, Richardson S. The timed "Up & Go": a test of basic functional mobility for frail elderly persons. J Am Geriatr Soc 1991 Feb;39(2):142-148. [doi: 10.1111/j.1532-5415.1991.tb01616.x] [Medline: 1991946]
13. Wiesmüller F, Prenner A, Ziegl A, El-Moazen G, Modre-Osprian R, Seinost G, et al. Correlation between timed-up-and-go test and 6-minute walk test in patients with peripheral artery disease. : IEEE; 2024 Presented at: 2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).
14. Brennan DM, Mawson S, Brownsell S. Telerehabilitation: enabling the remote delivery of healthcare, rehabilitation, and self management. Stud Health Technol Inform 2009;145(231-48):231-248. [Medline: 19592797]
15. Poelzl G, Egelseer-Bruendl T, Pfeifer B, et al. Feasibility and effectiveness of a multidimensional post-discharge disease management programme for heart failure patients in clinical practice: the HerzMobil Tirol programme. Clin Res Cardiol 2022 Mar;111(3):294-307. [doi: 10.1007/s00392-021-01912-0] [Medline: 34269863]
16. Zerna C, Jeerakathil T, Hill MD. Telehealth for remote stroke management. Can J Cardiol 2018 Jul;34(7):889-896. [doi: 10.1016/j.cjca.2017.12.025] [Medline: 29459240]
17. Cucato G, Snowden C, McCone E, et al. Evaluating the feasibility and acceptability of an exercise and behaviour change intervention in socioeconomically deprived patients with peripheral arterial disease: the textpad study protocol. PLoS One 2022;17(6):e0269999. [doi: 10.1371/journal.pone.0269999] [Medline: 35749440]
18. Schwamm LH, Chumbler N, Brown E, et al. Recommendations for the implementation of telehealth in cardiovascular and stroke care: a policy statement from the American Heart Association. Circulation 2017 Feb 14;135(7):e24-e44. [doi: 10.1161/CIR.0000000000000475] [Medline: 27998940]
19. Haveman ME, Kleiss SF, Ma KF, et al. Telemedicine in patients with peripheral arterial disease: is it worth the effort? Expert Rev Med Devices 2019 Sep;16(9):777-786. [doi: 10.1080/17434440.2019.1649595] [Medline: 31379218]
20. Prenner A, Ziegl A, Wiesmüller F, et al. Usability of a telehealth-nurse supported home-based walking training for peripheral arterial disease - The Keep Pace! pilot study. Vasa 2024 Jul;53(4):246-254. [doi: 10.1024/0301-1526/a001127] [Medline: 38808475]
21. Ziegl A, Rzepka A, Kastner P, et al. mHealth 6-minute walk test - accuracy for detecting clinically relevant differences in heart failure patients. Annu Int Conf IEEE Eng Med Biol Soc 2021 Nov;2021:7095-7098. [doi: 10.1109/EMBC46164.2021.9630118] [Medline: 34892736]
22. Ziegl A, Kastner P, Modre-Osprian R, Schreier G. Automated measurement and subtask analysis of the timed up-and-go test in the field of geriatrics. Annu Int Conf IEEE Eng Med Biol Soc 2018 Jul;2018(1526-9):1526-1529. [doi: 10.1109/EMBC.2018.8512464] [Medline: 30440683]
23. Scharhag-Rosenberger F, Meyer T, Walitzek S, Kindermann W. Time course of changes in endurance capacity: a 1-yr training study. Med Sci Sports Exerc 2009 May;41(5):1130-1137. [doi: 10.1249/MSS.0b013e3181935a11] [Medline: 19346973]

Abbreviations

6MWT: 6-minute walk test
HCP: health care professional
RMSE: root mean squared error
SWT: structured walking training
TUG: timed-up-and-go

Edited by A Mavragani; submitted 23.08.24; peer-reviewed by EL Garcia; revised version received 04.12.24; accepted 05.12.24; published 10.04.25.

Please cite as:
Wiesmüller F, Prenner A, Ziegl A, El-Moazen G, Modre-Osprian R, Baumgartner M, Brodmann M, Seinost G, Silbernagel G, Schreier G, Hayn D
Support of Home-Based Structured Walking Training and Prediction of the 6-Minute Walk Test Distance in Patients With Peripheral Arterial Disease Based on Telehealth Data: Prospective Cohort Study
JMIR Form Res 2025;9:e65721
URL: https://formative.jmir.org/2025/1/e65721
doi:10.2196/65721

XSL•FO
RenderX

© Fabian Wiesmüller, Andreas Prenner, Andreas Ziegl, Gihan El-Moazen, Robert Modre-Osprian, Martin Baumgartner, Marianne Brodmann, Gerald Seinost, Günther Silbernagel, Günter Schreier, Dieter Hayn. Originally published in JMIR Formative Research (https://formative.jmir.org), 10.4.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Web-Based Explainable Machine Learning-Based Drug Surveillance for Predicting Sunitinib- and Sorafenib-Associated Thyroid Dysfunction: Model Development and Validation Study

Fan-Ying Chan[1], MSc; Yi-En Ku[1], BSc; Wen-Nung Lie[2], PhD; Hsiang-Yin Chen[1,3], PharmD

[1]Department of Clinical Pharmacy, College of Pharmacy, Taipei Medical University, 250 Wuxing St, Xinyi Dist, Taipei, Taiwan

[2]Department of Electrical Engineering, National Chung Cheng University, Chiayi, Taiwan

[3]Department of Pharmacy, Wan Fang Hospital, Taipei Medical University, Taipei, Taiwan

Corresponding Author:
Hsiang-Yin Chen, PharmD
Department of Clinical Pharmacy, College of Pharmacy, Taipei Medical University, 250 Wuxing St, Xinyi Dist, Taipei, Taiwan

Abstract

Background: Unlike one-snap data collection methods that only identify high-risk patients, machine learning models using time-series data can predict adverse events and aid in the timely management of cancer.

Objective: This study aimed to develop and validate machine learning models for sunitinib- and sorafenib-associated thyroid dysfunction using a time-series data collection approach.

Methods: Time series data of patients first prescribed sunitinib or sorafenib were collected from a deidentified clinical research database. Logistic regression, random forest, adaptive Boosting, Light Gradient-Boosting Machine, and Gradient Boosting Decision Tree were used to develop the models. Prediction performances were compared using the accuracy, precision, recall, F1-score, area under the receiver operating characteristic curve, and area under the precision-recall curve. The optimal threshold for the best-performing model was selected based on the maximum F1-score. SHapley Additive exPlanations analysis was conducted to assess feature importance and contributions at both the cohort and patient levels.

Results: The training cohort included 609 patients, while the temporal validation cohort had 198 patients. The Gradient Boosting Decision Tree model without resampling outperformed other models, with area under the precision-recall curve of 0.600, area under the receiver operating characteristic curve of 0.876, and F1-score of 0.583 after adjusting the threshold. The SHapley Additive exPlanations analysis identified higher cholesterol levels, longer summed days of medication use, and clear cell adenocarcinoma histology as the most important features. The final model was further integrated into a web-based application.

Conclusions: This model can serve as an explainable adverse drug reaction surveillance system for predicting sunitinib- and sorafenib-associated thyroid dysfunction.

(JMIR Form Res 2025;9:e67767) doi:10.2196/67767

KEYWORDS

thyroid dysfunction; machine learning; cancer; sunitinib; sorafenib; TKI; tyrosine kinase inhibitor

Introduction

Background

Sunitinib- and sorafenib-associated thyroid dysfunction are time-varying and underreported adverse drug reactions (ADR). Despite the efficacy of multitargeted tyrosine kinase inhibitors TKI as first- or second-line therapies for solid and hematologic cancers, thyroid dysfunction—especially hypothyroidism—may complicate treatment regimens using sunitinib and sorafenib. Although not life-threatening, such adverse events can lead to a suboptimal quality of life, physical discomfort, or the need for thyroid dysfunction treatment [1,2]. Incidences of thyroid dysfunction vary from 10%- to 85% for sunitinib and from 6.3% to 27% for sorafenib [3-6]. Hypothyroidism develops late and is prolonged, with onset ranging from 5 to 20 months [1,6-8]. Due to its high variability and difficulty in predicting, close monitoring and proactive ADR surveillance with machine learning (ML) models may be warranted to manage sunitinib- and sorafenib-associated thyroid dysfunction.

Incorporating real-time laboratory data can enhance the performance of ML models in predicting thyroid adverse events. Previous studies have demonstrated the effectiveness of ML in predicting thyroid dysfunction, such as amiodarone-induced dysfunction using time-series data collection methods with a robust performance [9]. In contrast, predicting immune checkpoint inhibitor-induced adverse thyroid events without continuous laboratory testing resulted in an area under the precision-recall curve (AUPRC) of 0.510 [10]. Another study

XSL•FO
RenderX

developed a predictive model for immune checkpoint inhibitor-induced adverse thyroid events without detailed time-point data collection, achieving an area under the receiver operating characteristic curve (AUROC) of 0.885 with thyroid-related features [11]. However, no studies have focused on predicting thyroid adverse events caused by TKIs such as sunitinib and sorafenib. Adopting timely data collection with detailed clinical biochemical tests to develop thyroid dysfunction surveillance systems can help clinicians timely adjust TKI treatment by balancing risks and benefits.

Applying model interpretation and web-based applications can alleviate the black-box drawback of ML models and enhance understanding of ADRs. The SHapley Additive exPlanations (SHAP) analysis is a mathematical method that has become increasingly popular for explaining ML models [12-14]. It is based on game theory concepts that calculate feature importance and contributions to predicted outcomes at both population and individual levels [15]. Model deployment in web-based or smartphone applications were constructed in recent years to strengthen the scalability of predicting specific diseases or ADRs [13,16,17]. Through a user-friendly interface and explainable model, applications can provide clinicians with a greater understanding and actionable preventive remedies for their patients. These techniques successfully have increased the interpretability and accessibility of ML models to support clinical decision-making.

The landscape of drug surveillance has experienced notable changes, shifting from passive reporting systems or rules-based alerting systems to active identification of adverse drug events, driven by advancements in electronic health records and artificial intelligence. Traditional methods often rely on spontaneous reporting or rules-based alerting, which may lead to underreporting and delays in identifying ADRs. The ML models, particularly tree-based algorithms, can analyze large-scale clinical data in real time, proactively detecting ADRs and enhancing medication safety at the point of care [18]. By continuously collecting and analyzing patient data, ML-driven drug surveillance based on time-series data extraction enables dynamic risk assessment, facilitating early intervention and personalized treatment adjustments. Given the underreported and time-varying nature of thyroid dysfunction, this progressive approach can serve as a valuable tool for predicting ADRs in patients undergoing TKI therapy.

Objective

The objective of this study was to develop and validate progressive ML predictive models for sunitinib- and sorafenib-induced thyroid dysfunction. Five algorithms—logistic regression (LR), random forest (RF), Gradient Boosting Decision Tree (GBDT), Light Gradient Boosting Machine (LGBM), and Adaptive Boosting (AdaBoost)—were used to construct the models. The specific aims of this study were (1) to use time-series data collected at the baseline, and at 1, 2, 3, 4, 5, 6, 9, 12, 18, 24, 30, and 36 months after the index date to generate predictive models using the five algorithms; (2) to select the best model by comparing the accuracy, precision, recall, F1-score, AUROC, and AUPRC, after adjusting for the optimal threshold; (3) to interpret the best-performing model with a SHAP analysis to analyze feature importance levels and contributions, comparing findings with a previous amiodarone-induced thyroid dysfunction predictive model; and (4) to deploy the best-performing model by constructing a web-based application.

Methods

Ethical Considerations

This retrospective study used the deidentified Clinical Research Database (CRD) which includes data from Taipei Medical University Hospital, Wan Fang Hospital, and Shuang Ho Hospital of the TMU health care system. Ethical review for this study was waived by the TMU-Joint Institutional Review Board (approval no.: N202202053). As the data had been deidentified, the requirement for informed consent was waived. No compensation was provided to the participants.

Study Design and Patient Cohort

Data from patients who were first prescribed sunitinib or sorafenib from 2013 to 2019 were collected as the derivation (training) cohort, whereas data from patients treated between 2010 and 2012 were collected as a temporal validation (testing) cohort. The data-splitting time point was chosen considering a common proportion of 7:3 for the training and testing cohorts and similar incidences of thyroid dysfunction in both cohorts for better model development and validation. More recent data were used for model training to better reflect current clinical practice, allowing models to be developed with up-to-date treatment patterns and patient characteristics. Patients were excluded if they were younger than 18 years, pregnant, with a history of hypothyroidism, had a diagnosis of thyroid carcinoma, had undergone thyroidectomy, used levothyroxine, or had abnormal thyroid-stimulating hormone (TSH) levels within 1 year before the index date, which was the first day of sunitinib or sorafenib use. Patients were also excluded if they were lost to follow up. Each individual was followed up for 3 years, and data were collected until the end of the study period, loss to follow-up, death, or occurrence of thyroid dysfunction. Figure 1 demonstrates the study design of the proposed models. This study followed the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis + Artificial Intelligence (TRIPOD + AI) reporting guideline [19]. Registration was not required, and no separate study protocol was prepared.

XSL•FO
RenderX

Figure 1. Study design of the proposed models. The study process included data collection, data splitting, data preprocessing, feature selection, resampling, model development, model validation, threshold adjustment, model interpretation, and development of a web-based application. AdaBoost: Adaptive Boosting; AUPRC: area under the precision recall curve; AUROC: area under the receiver operating characteristic curve; B-SMT: Oversampling with the Borderline Synthetic Minority Oversampling Technique; CRD: clinical research database; GBDT: Gradient Boosting Decision Tree; LGBM: Light Gradient Boosting Machine; RFECV: Recursive Feature Elimination with Cross-Validation; SMT: Synthetic Minority Oversampling Technique; TMU: Taipei Medical University.

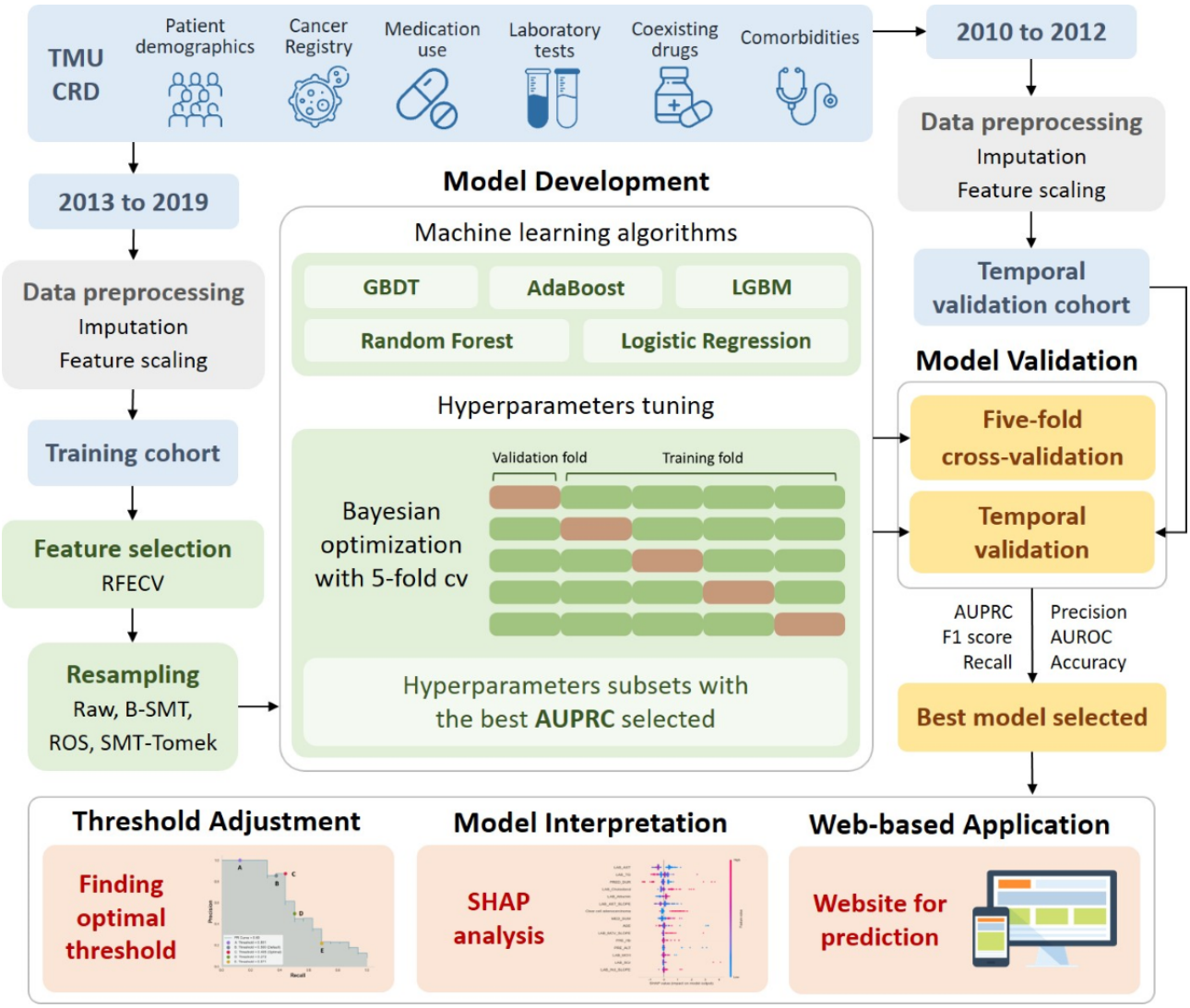

Time-Series Data Collection and Preprocessing

Variables including patient demographics, cancer-related information, medication use, laboratory tests, coexisting drugs, and comorbidities were collected from the CRD. Patient demographics including age, gender, and cancer-related information were collected once at the baseline. Laboratory tests, medication use, comorbidities, coexisting drugs, weight, and body mass index (BMI) were collected at the baseline, and at 1, 2, 3, 4, 5, 6, 9, 12, 18, 24, 30, and 36 months after the index date (Multimedia Appendix 1). Reengineering features included calculating the cumulative dose, summed days of medication, duration of sunitinib or sorafenib, and the slope of laboratory tests at recent and previous data collection time points. The number of coexisting thyroid-related drugs was calculated to determine whether the aggregated use of these medications increased the risk of thyroid dysfunction.

Data preprocessing including missing value imputation and feature scaling was performed to ensure data quality. Features with over 80% missing values, except for the TSH level, were first excluded from the model. Missing values were imputed using multivariate imputation by chained equations [20]. If previous data were traceable, missing values, using time-series data collection, were precisely imputed using the last observation carried forward. Missing values in laboratory test slopes were imputed with zero. A feature-scaling method—the robust scaler—was used to reduce the effect of extreme values of continuous variables. Features were first selected through a literature review and clinical domain knowledge to identify thyroid-related variables. Recursive feature elimination, a feature selection technique that recursively removes the least important features with 5-fold cross-validation was then used, and the AUROC was considered to select optimal features for each model. The codebook and missing rates of each variable in the training and testing set are described in Multimedia Appendix 2.

XSL•FO
RenderX

Prediction Outcome

The predictive outcome of this study was the occurrence of sunitinib- or sorafenib-associated thyroid dysfunction. Patients were defined as a TSH level ≥4 mIU/L, a diagnosis of hypothyroidism with International Classification of Diseases codes (*ICD-9-CM*; *243* or *244* or *ICD-10*; *E02* or *E03)*, or use of levothyroxine (Anatomical Therapeutic Chemical code: H03AA02). These criteria were defined following previous studies that focused on drug-induced thyroid dysfunction [9,21,22].

Model Development and Validation

The study used three resampling strategies and raw data applied to the five algorithms to develop 20 machine learning models. The five algorithms included LR, RF, AdaBoost [23], LGBM [24], and GBDT [25]. Resampling strategies used were random over-sampling (ROS), borderline synthetic minority oversampling technique (B-SMT), and a combination of over- and under-sampling using SMT and Tomek links (STMK). Bayesian optimization with 5-fold cross-validation and the AUPRC were considered for hyperparameter tuning [26]. For boosting models, hyperparameters such as learning_rate and n_estimators were optimized to balance convergence speed and model complexity. Specifically, in LGBM, subsample and colsample_bytree were fine-tuned to mitigate overfitting. Key parameters including max_depth, min_samples_split, and min_samples_leaf were adjusted to improve generalization and prevent overfitting for RF. Table S3.1 in Multimedia Appendix 3 lists the ranges of hyperparameters used for tuning in each algorithm. Pseudocodes of the model development process are listed in Multimedia Appendix 4.

A stratified 5-fold cross-validation was first used for internal validation of the training dataset, and the performances of the 20 models were evaluated using the temporal validation cohort. The accuracy, precision, recall, F1-score, AUROC, and AUPRC were the metrics used for evaluating model performance. Model performance was compared and evaluated following the sequence of the highest AUPRC, AUROC value, and F1-score to select the best-performing model.

Threshold Adjustment, Model Interpretation, and Web-Based Application

To optimize the performance of the best-performing model, different thresholds were selected on the PRC based on different percentages of outcome predictions. Five cutoff points were chosen: (1) a high-risk threshold for identifying a lower number of patients potentially at risk; (2) a default threshold of 0.50; (3) an optimal threshold based on the maximum F1-score; (4) an equal threshold where precision, recall, and F1-score were equal, and (5) the low-risk threshold for identifying a higher number of patients potentially at risk. The F1-score, recall, precision, and accuracy were then compared for these 5 thresholds. The best model was further examined by SHAP analysis to explain feature importance and contribution at both the population and individual levels [27]. At the population level, SHAP summary plots were used to demonstrate feature importance and how the top-ranked features impacted outcome predictions. On the other hand, SHAP force plots were used to visualize how the features contributed to outcome predictions in specific patients.

The model was further integrated into a web-based application to increase the accessibility of the model. The Python Flask framework was used to develop the application programming interface. The value of each feature served as input data that were fed back into the model to generate a predictive probability of thyroid dysfunction. The consistency of feature scaling between input data and training data was ensured by applying the same scaling transformation. The input variables with missing values were imputed with the last observation or the same multivariate imputation by chained equations transformation applied during model development to ensure consistency in predictions.

Statistical Analysis

Patient characteristics were evaluated with independent *t* tests or the Wilcoxon rank-sum test for continuous variables, and the χ^2 test or Fisher exact test for categorical variables. A two-sided *P* value <.05 was considered statistically significant. Data were analyzed using SAS (version 9.4; SAS Institute), Python (version 3.9.5; Python Software Foundation), and R software (version 4.2.2; R Foundation for Statistical Computing). The statistical significance of the AUPRC was calculated using MedCalc software (version 22.001).

Results

Patient Characteristics and Multivariate Analysis

This study enlisted data from 900 patients prescribed sunitinib or sorafenib from initial screening from the CRD. A STROBE (Strengthening The Reporting of Observational Studies in Epidemiology) flowchart of patient selection is described in Multimedia Appendix 5. After applying the exclusion criteria, 807 patients remained. Patient characteristics and multivariate analysis of the training and temporal validation cohorts are listed in Multimedia Appendix 6. The training cohort contained 609 patients, with 52 (8.5%) patients developing thyroid dysfunction, while 16 (8.1%) patients experienced thyroid dysfunction in the temporal validation cohort of 198 patients. There were no significant differences in age or gender in the derivation or temporal validation cohort. Cancer type, histology, aspartate aminotransferase (AST), and bilirubin levels significantly differed (*P*<.001) between patients with and those without thyroid dysfunction in the two cohorts.

Model Development and Validation

The total number of features selected by recursive feature elimination were 20, 18, 15, 40, and 20 in GBDT, AdaBoost, LGBM, RF, and LR, respectively. The most frequently selected features were recent laboratory tests including AST, alanine transaminase (ALT), bilirubin, cholesterol, and triglyceride levels. Patient demographics such as age and BMI summed days of medication use, and follow-up duration were also among the top-ranked selected features (Multimedia Appendix 7). The optimal hyperparameter subsets selected by Bayesian optimization are listed in Table S3.2 in Multimedia Appendix 3.

XSL•FO
RenderX

Figure 2 shows the model performance based on the AUPRC, AUROC, F1-score, precision, recall, and accuracy in 20 ML models of the temporal validation cohort. Most of the GBDT and AdaBoost models had higher AUPRC values and F1-scores, while recall was higher in the RF and LR models. Of all the ML models, the GBDT without resampling (GBDT_RAW) outperformed the other models, with an AUPRC of 0.600, AUROC of 0.876, and an F1-score of 0.522. The AUPRC of GBDT_RAW (0.600, 95% CI 0.350-0.798) was significantly higher than those of the GBDT_ROS (0.388, 95% CI 0.175 - 0.618; $P<.05$), GBDT_BMST (0.300, 95% CI 0.124 - 0.549; $P<.05$), and GBDT_STMK (0.346, 95% CI 0.147 - 0.582; $P<.05$). Multimedia Appendix 8 lists model performances of 5-fold cross-validation with the training cohort and the statistical significance tests of the AUPRC of all the models.

Figure 2. Model performance on temporal validation. The model performance evaluated by the six metrics is demonstrated in a heatmap. Blue and green colors represent higher and lower values, respectively. ROS: random oversampling; BSMT: Oversampling with Borderline Synthetic Minority Oversampling Technique; STMK: Synthetic Minority Oversampling Technique-Tomek Links; GBDT: Gradient Boosting Decision Tree; AdaBoost: Adaptive Boosting; LGBM: Light Gradient Boosting Machine; RF: random forest; LR: logistic regression; AUPRC: area under the precision-recall curve; AUROC: area under the receiver operating characteristic curve.

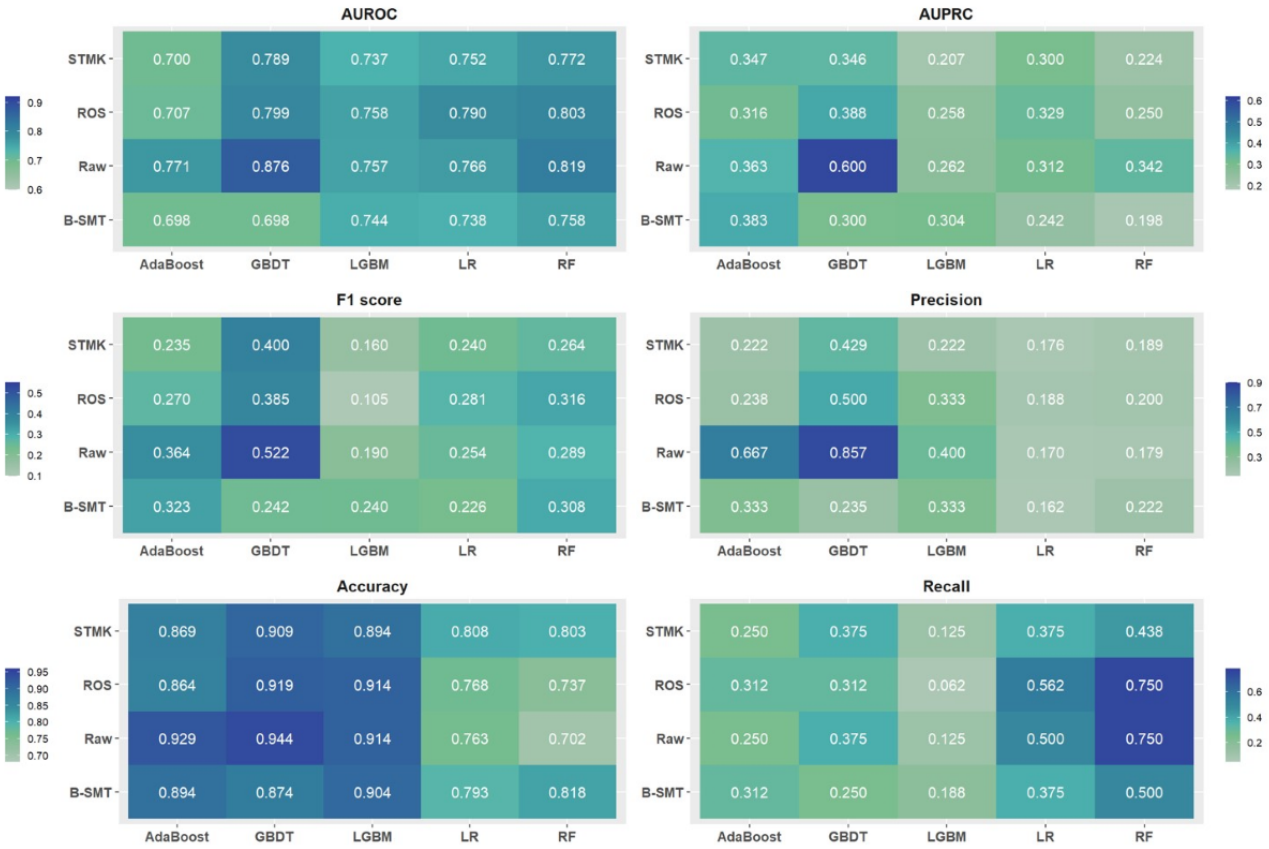

Threshold Adjustment

Figure 3 shows the five cutoff points on the PRC and the precision, recall, and F1-score at different thresholds of the GBDT_RAW model. Based on different percentages of outcome predictions, cutoff points were 0.851 for identifying the top 1% of patients (2 predictive positive cases, point A — high-risk threshold),0.500 for the top 3.5% (7 predictive positive cases, point B — default threshold), 0.495 for the top 4% (8 predictive positive cases, point C — optimal threshold based on the maximum F1-score), 0.272 for the top 8% (16 predictive positive cases, point D — equal threshold), and 0.071 for the top 26% of patients (51 predictive positive cases, point E — low-risk threshold). When moving the threshold from 0.500 to 0.851 (point A), the precision increased from 0.875 to 1.000, while recall significantly decreased to 0.125. In contrast, recall significantly increased to 0.750 and the precision decreased to 0.231 when changing the threshold to 0.066 (point E). The precision, recall, and F1-score reached 0.500 when the threshold was adjusted to 0.272 (point D). The optimal threshold of 0.495 (point C) for the GBDT_RAW model was selected based on the maximum F1-score, with a precision of 0.875, recall of 0.438, and F1-score of 0.583.

XSL•FO
RenderX

Figure 3. Threshold adjustment for the GBDT_RAW model. The five cutoff points on the precision-recall curve (A) represent different percentages of outcome predictions, precision, recall, and F1-scores based on different thresholds; (B) ,The X-axis represents the value of the threshold and the Y-axis shows values of the precision, recall, and F1-score.

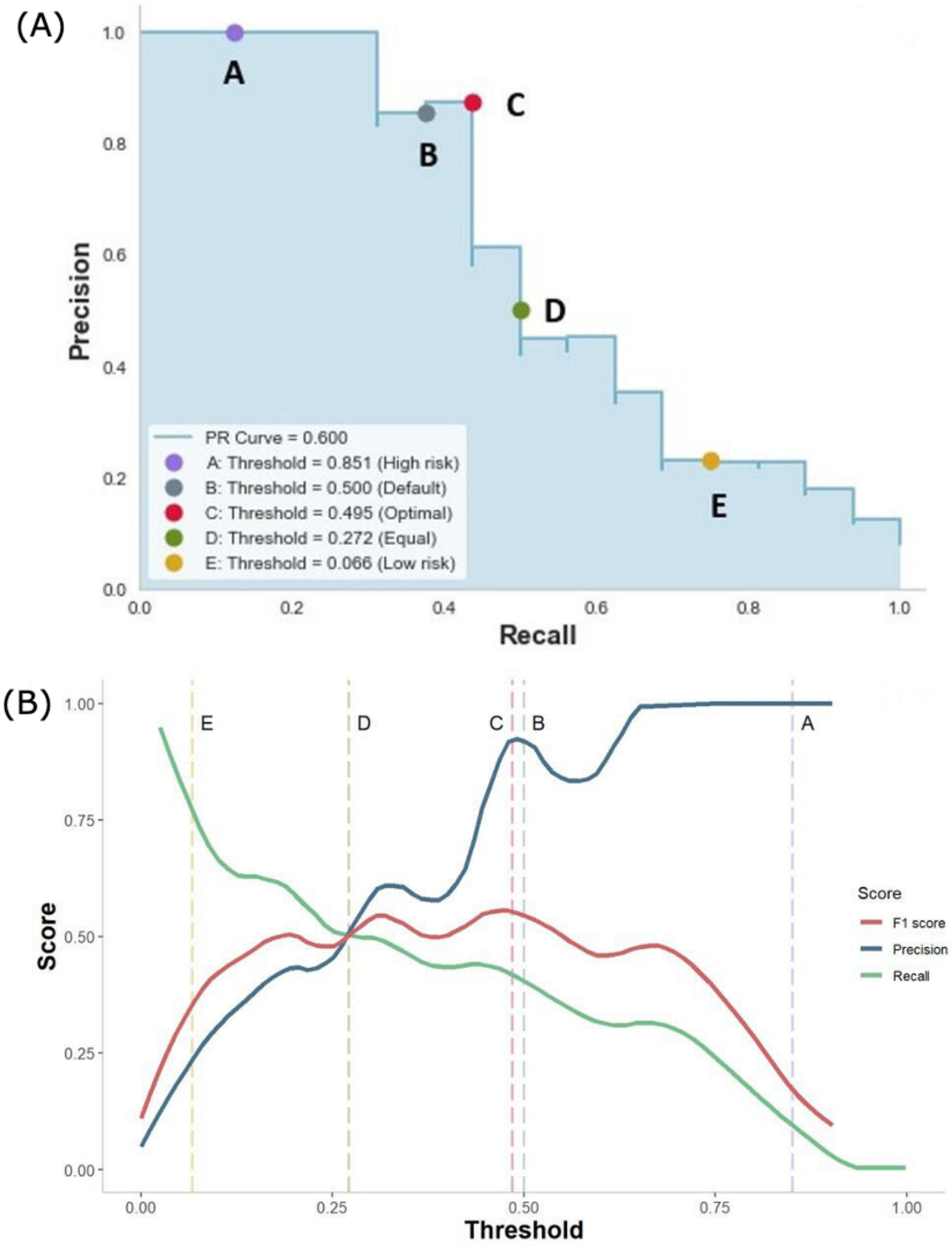

XSL•FO
RenderX

Model Interpretation

Figure 4 shows model interpretation implemented with the SHAP analysis on the best-performing GBDT_RAW model. The summary plot in Figure 4A displays that the top five important predictors were the AST level, TG level, follow-up duration, cholesterol level, and albumin level. Figure 4B shows that patients with lower AST levels, higher cholesterol levels, higher albumin levels, longer durations of medication use, and clear cell adenocarcinoma histology had higher SHAP values for predicting the occurrence of thyroid dysfunction. At the individual level, force plots showed the impacts of features on predicting thyroid dysfunction in two specific patients. Figure 4C shows a patient with a predictive probability of 0.02, resulting from his AST level of 69 IU/L and the summed days of medications of 62 days, which contributed most negatively to outcome predictions. For a patient with a predictive probability of 0.82 in Figure 4D, the medication duration of 708 days and clear cell adenocarcinoma served as the most important factors that positively impacted outcome predictions.

Figure 4. Model interpretation with a SHAP analysis for the GBDT_RAW model. At the population level, the summary bar plot (A) Mean SHAP values of all features, and a summary dot plot; (B) shows how each feature impacted the outcome prediction. The x-axis shows SHAP values of all features, and the colors represent the feature values, with red and blue respectively demonstrating higher and lower values. At the individual level, force plots in (C) and (D) show how features contributed to the model output value in each patient. Red and blue arrows represent positive and negative contributions of features, and the length of the arrow shows the magnitude of the impact on outcome predictions. LAB_AST: recent aspartate aminotransferase level; LAB_TG: recent triglyceride level; PRED_DUR: follow-up days; LAB_Cholesterol: recent cholesterol level; LAB_Albumin: recent albumin level; LAB_AST_SLOPE: slope of aspartate aminotransferase level; MED_SUM: sum days of medication; LAB_MCV_SLOPE: slope of mean corpuscular volume; PRE_Hb: previous hemoglobin level; PRE_ALT: previous alanine transaminase level; LAB_MCH: recent mean corpuscular haemoglobin level; LAB_SCr: recent serum creatinine level; LAB_Hct_SLOPE: slope of hematocrit level; MED_DUR: duration of medication; PRE_MCHC: previous mean corpuscular hemoglobin concentration level; LAB_SCr_SLOPE: slope of serum creatinine level; LAB_MCHC_SLOPE: slope of mean corpuscular hemoglobin concentration level.

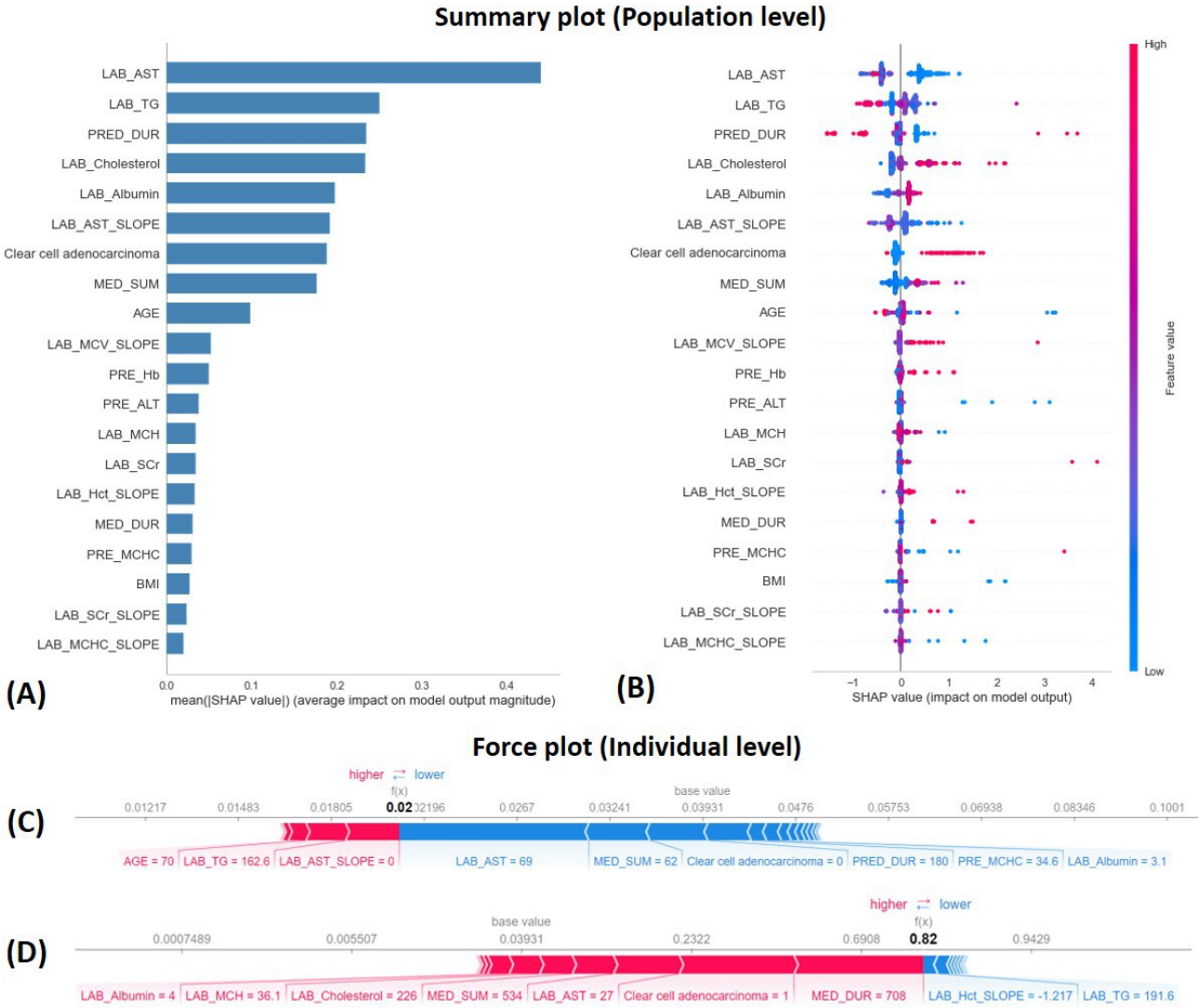

Web-Based Application

A web-based application for predicting sunitinib- and sorafenib-associated thyroid dysfunction was developed using the GBDT_RAW model (Figure 5). The application provides a user-friendly interface by showing 20 features selected by the GBDT model. Values of the features can be filled in with the appropriate units shown in each input box. The model then generates a predictive probability for the risk of sunitinib- or

XSL•FO
RenderX

sorafenib-associated thyroid dysfunction. Ultimate predictions for the occurrence of thyroid dysfunction were determined based on thresholds adopted by clinical users.

Figure 5. Web-based application with a user-friendly interface. A web-based application developed using the GBDT_RAW model showing 20 selected features selected and the predictive probability.

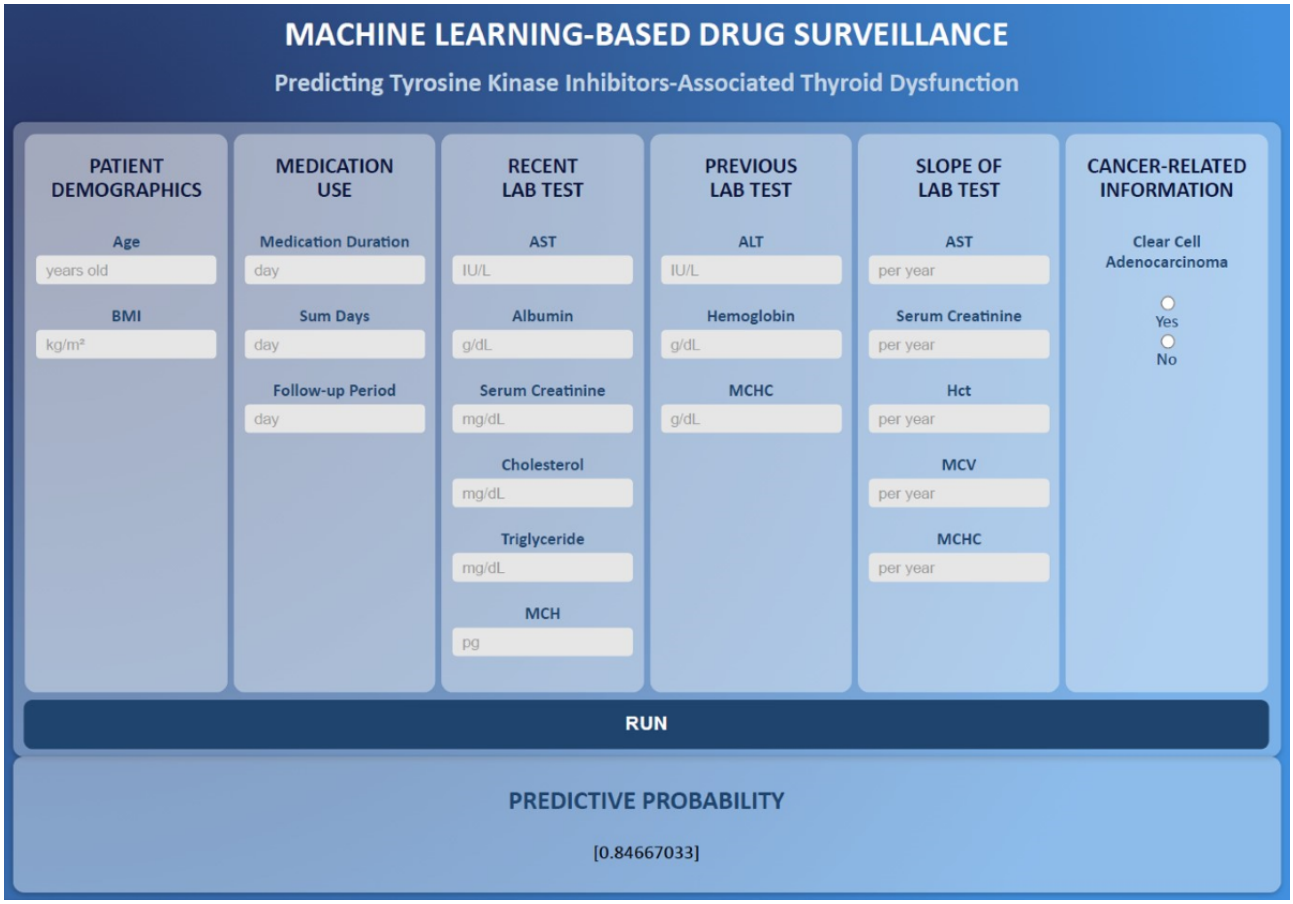

Discussion

Principal Findings

This study developed explainable ML models by collecting time-series data to predict sunitinib- and sorafenib-associated thyroid dysfunction. The present model allowed ongoing predictions according to the rapidly shifting status of the disease for patients undergoing long-term TKI treatment. The best-performing GBDT without resampling model was optimized by threshold moving strategies to achieve a maximum F1-score. This study further used a SHAP analysis that provided interpretability at both the population and individual levels, identifying key predictors such as AST, cholesterol, and albumin levels. The feature selection process revealed similarities with prior research on amiodarone-induced thyroid dysfunction, supporting the model's potential applicability to other drug-induced thyroid dysfunctions [9]. Integrating the model into a web-based application demonstrated its practical utility by allowing real-time risk estimation based on patient-specific data.

Time-Series Data Collection

Time-series data collection in this study had a few advantages for thyroid dysfunction predictions on TKI users as illustrated in Figure 6. Clinical data at multiple time points in a long-term follow-up period were collected for model building. Time-series data allowed reengineering features by calculating rates of change of laboratory tests at recent and previous time points. The last observation carried forward method using the last value to replace missing data in subsequent time points ensured that missing values were imputed with plausible estimates [28]. This time-series data collection closely monitored patients newly prescribed TKI users and those on long-term TKI therapy. The time-series model captured the critical moment of an approaching adverse event for an individual patient, providing oncologists with invaluable information to treat patients on TKIs.

Identifying thyroid dysfunction was reported to potentially serve as a prognostic indicator for treatment outcomes. Sunitinib and sorafenib respectively demonstrated progression-free survival (PFS) periods of 11 and 5.5 months in patients with renal cell carcinoma [29,30]. Interestingly, studies indicated that patients with thyroid dysfunction had a longer PFS compared to patients with normal thyroid function among sunitinib (11.9 vs 8.8 months) and sorafenib users (19.3 vs 5.5 months) [4,31]. These findings suggest that thyroid dysfunction may not only represent an adverse event but could serve as a marker of enhanced therapeutic response. Leveraging a time-series data collection approach, this study can predict thyroid dysfunction risk while simultaneously identifying patients who may derive greater

XSL•FO
RenderX

clinical benefits from sustained TKI therapy. Unlike one-snap data collection which arbitrarily differentiates high-risk patients from a low-risk group, this time-series approach allows high-risk patients to use lifesaving TKIs for a relatively longer grace period which can significantly enhance their survival times. Models with time-series data collection balance the risks and benefits of survival and adverse drug events, bestowing precious survival periods for end-stage cancer patients with reasonable medication safety and improved quality of life.

Figure 6. Graphic comparison of one-snap and time-series data collection methods. Time-series and one-snap data collection methods were compared. The dashed line represents possible changes in variables during a long-term follow-up period. LOCF: Last Observation Carried Forward

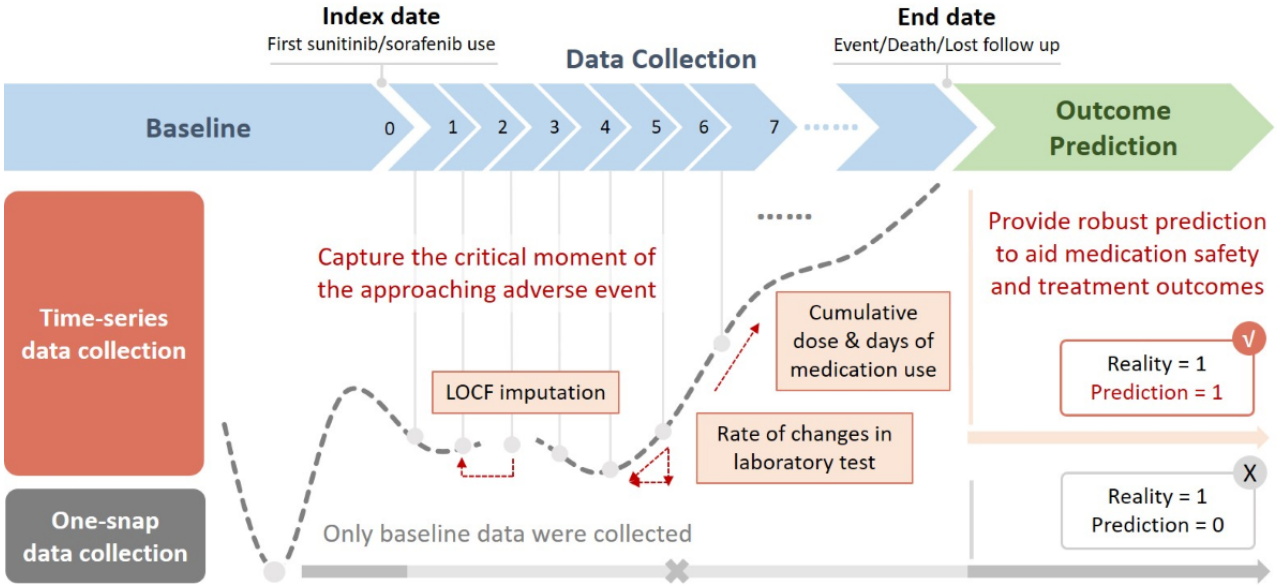

Best-Performing Model With Threshold Adjustment

The GBDT_RAW model outperformed other machine learning models in this study, with thresholds adjusted to optimize model performance. The gradient boosting method analyzed nonlinear relationships and handled imbalanced data to predict diagnoses, hospital readmissions, and clinical outcomes [32-34]. During the hyperparameter tuning process, the AUPRC was considered, rather than the F1-score, without adjusting the threshold for imbalanced data [35,36]. The model without resampling outperformed resampling approaches by maintaining high precision, while the sparsity of the minority class in this study probably caused resampling to introduce noise or unrealistic synthetic data, failing to improve recall. The final model then used five cutoff points on the PRC to decide optimal thresholds and percentages of positive predictions. A relatively low threshold can identify all patients at risk but may increase the number of false positives, overwhelming clinicians with numerous warnings to review. On the contrary, choosing a relatively high threshold reduces the number of false positives but increases the chances of missing true positive cases. The selected threshold increased the precision, recall, and F1-score of the model's performance, while considering the relative costs of false positives and false negatives to meet clinical needs [37].

Model Interpretation and Potential for Model Expansion

Similar features for predicting drug-induced thyroid dysfunction were selected by the presence of TKIs and a previous amiodarone model built by our research team [9]. The feature selection process identified age, BMI, medication duration, summed days of medication, rate of change in serum creatinine levels, and recent mean corpuscular hemoglobin, AST, and cholesterol levels as key predictors of thyroid dysfunction in this study. Interestingly, these same features were also selected in our previous amiodarone model, suggesting their potential relevance across different drug-induced thyroid dysfunction models. Similar features highlight the potential for model expansion to increase generalizability. Top-ranked features identified in this study and their contributions calculated by the SHAP analysis were aligned with clinical domain knowledge. Higher cholesterol levels were found to positively impact predictions, which can be explained by the effect of thyroid function on lipoprotein metabolism [38]. The present study also found that longer summed days of medication increased the risk of thyroid dysfunction. This finding was confirmed by a long-term safety study of sunitinib, which showed the frequency of thyroid dysfunction, unlike other adverse events, gradually increased over time [39]. The AST level and histology of clear cell carcinoma served as important predictors, which may have resulted from a relatively lower incidence of thyroid dysfunction in patients with hepatocellular carcinoma and renal cell carcinoma.

Clinical Implications of Web-Based Drug Surveillance

ML-based drug surveillance provides a promising tool for predicting and managing thyroid dysfunction caused by TKIs. Traditional clinical decision support systems (CDSS) rely on rule-based alerts for drug interactions or contraindications but often fail to capture the multifactorial nature of drug-induced thyroid dysfunction, which depends on thyroid-related factors [40]. AI-driven CDSSs, leveraging ML algorithms, can overcome these limitations by identifying complex patterns in patient data, providing personalized risk assessments, and generating timely alerts for thyroid function monitoring. Deploying such ML models on cloud-based platforms alongside

XSL•FO
RenderX

rule-based CDSSs could enhance real-time risk stratification, enabling early detection, intervention, and potential adjustments in therapy to prevent endocrine complications. Further exploration of CDSSs incorporating ML predictions are warranted to improve clinical practice and medication safety [40].

Limitations

There are a few limitations of this study. Data were from a single healthcare system and included limited numbers of sunitinib- or sorafenib-treated patients. There were missing data in the diagnosis of comorbidities and medication use. The nature of a retrospective study introduces underestimations of the incidence of thyroid dysfunction. As a result, different features selected by the present model and our previous amiodarone model mainly resulted from features with higher missing rates. Multicenter and multicountry studies for improving model extrapolation are needed before clinical application. Future qualitative research and prospective studies with the involvement of physicians should also be conducted to assess the usability and accessibility of the model for real-world evidence.

Conclusions

This study applied time-series data collection to capture the critical moment of sunitinib- and sorafenib-associated thyroid dysfunction for ADR surveillance. The optimal threshold can balance the precision and recall meeting clinical needs. Feature importance was explained at the population and individual levels. The web-based application increased the model accessibility, allowing clinical users to receive real-time predictions. The comparison of features with amiodarone-induced thyroid dysfunction highlighted the potential for future model expansion.

Acknowledgments

This work was supported by the National Science and Technology Council (NSTC) [grant no. NSTC-113-2320-B-038-057].

Data Availability

The data were obtained from the Taipei Medical University clinical research database; this was approved by the Institutional Review Board of Taipei Medical University (TMU-JIRB-N202202053). The data and web-based application are available from the corresponding author upon reasonable requests, subject to appropriate request mechanisms and additional Institutional Review Board approval for personal information protection.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Time-series data collection.
[DOCX File, 364 KB - formative_v9i1e67767_app1.docx]

Multimedia Appendix 2
Codebook and missing rates of variables.
[DOCX File, 24 KB - formative_v9i1e67767_app2.docx]

Multimedia Appendix 3
Hyperparameters tuning.
[DOCX File, 21 KB - formative_v9i1e67767_app3.docx]

Multimedia Appendix 4
Pseudocodes of the model developing process.
[DOCX File, 18 KB - formative_v9i1e67767_app4.docx]

Multimedia Appendix 5
Patient selection flowchart.
[DOCX File, 141 KB - formative_v9i1e67767_app5.docx]

Multimedia Appendix 6
Patient characteristics and multivariate analysis.
[DOCX File, 23 KB - formative_v9i1e67767_app6.docx]

XSL•FO
RenderX

Multimedia Appendix 7
Feature selection with recursive feature elimination.
[DOCX File, 21 KB - formative_v9i1e67767_app7.docx]

Multimedia Appendix 8
Model performance of five-fold cross-validation.
[DOCX File, 24 KB - formative_v9i1e67767_app8.docx]

References

1. Torino F, Corsello SM, Longo R, Barnabei A, Gasparini G. Hypothyroidism related to tyrosine kinase inhibitors: an emerging toxic effect of targeted therapy. Nat Rev Clin Oncol 2009 Apr;6(4):219-228. [doi: 10.1038/nrclinonc.2009.4] [Medline: 19333228]
2. Abdel-Rahman O, Fouad M. Risk of thyroid dysfunction in patients with solid tumors treated with VEGF receptor tyrosine kinase inhibitors: a critical literature review and meta analysis. Expert Rev Anticancer Ther 2014 Sep;14(9):1063-1073. [doi: 10.1586/14737140.2014.929501] [Medline: 24927771]
3. Walko CM, Aubert RE, La-Beck NM, et al. Pharmacoepidemiology of clinically relevant hypothyroidism and hypertension from sunitinib and sorafenib. Oncologist 2017 Feb;22(2):208-212. [doi: 10.1634/theoncologist.2016-0233] [Medline: 28167571]
4. Vasileiadis T, Chrisofos M, Safioleas M, Kontzoglou K, Papazisis K, Sdrolia A. Impact of sunitinib-induced hypothyroidism on survival of patients with metastatic renal cancer. BMC Cancer 2019 Apr 30;19(1):407. [doi: 10.1186/s12885-019-5610-8] [Medline: 31039771]
5. Wu S, Chen JJ, Kudelka A, Lu J, Zhu X. Incidence and risk of hypertension with sorafenib in patients with cancer: a systematic review and meta-analysis. Lancet Oncol 2008 Feb;9(2):117-123. [doi: 10.1016/S1470-2045(08)70003-2] [Medline: 18221915]
6. Sabatier R, Eymard JC, Walz J, et al. Could thyroid dysfunction influence outcome in sunitinib-treated metastatic renal cell carcinoma? Ann Oncol 2012 Mar;23(3):714-721. [doi: 10.1093/annonc/mdr275] [Medline: 21653681]
7. Clemons J, Gao D, Naam M, Breaker K, Garfield D, Flaig TW. Thyroid dysfunction in patients treated with sunitinib or sorafenib. Clin Genitourin Cancer 2012 Dec;10(4):225-231. [doi: 10.1016/j.clgc.2012.08.002] [Medline: 23017335]
8. Jannin A, Penel N, Ladsous M, Vantyghem MC, Do Cao C. Tyrosine kinase inhibitors and immune checkpoint inhibitors-induced thyroid disorders. Crit Rev Oncol Hematol 2019 Sep;141:23-35. [doi: 10.1016/j.critrevonc.2019.05.015] [Medline: 31202955]
9. Lu YT, Chao HJ, Chiang YC, Chen HY. Explainable machine learning techniques to predict amiodarone-induced thyroid dysfunction risk: multicenter, retrospective study with external validation. J Med Internet Res 2023 Feb 7;25:e43734. [doi: 10.2196/43734] [Medline: 36749620]
10. Kim W, Cho YA, Kim DC, Jo AR, Min KH, Lee KE. Factors associated with thyroid-related adverse events in patients receiving PD-1 or PD-L1 inhibitors using machine learning models. Cancers (Basel) 2021 Oct 30;13(21):21. [doi: 10.3390/cancers13215465] [Medline: 34771631]
11. Wang Q, Wu T, Zhao R, et al. Development and validation of a prediction model for thyroid dysfunction in patients during immunotherapy. Endocr Pract 2024 Oct;30(10):943-950. [doi: 10.1016/j.eprac.2024.07.006] [Medline: 39004306]
12. Deng H, Eftekhari Z, Carlin C, et al. Development and validation of an explainable machine learning model for major complications after cytoreductive surgery. JAMA Netw Open 2022 May 2;5(5):e2212930. [doi: 10.1001/jamanetworkopen.2022.12930] [Medline: 35612856]
13. Hu J, Xu J, Li M, et al. Identification and validation of an explainable prediction model of acute kidney injury with prognostic implications in critically ill children: a prospective multicenter cohort study. EClinicalMedicine 2024 Feb;68:102409. [doi: 10.1016/j.eclinm.2023.102409] [Medline: 38273888]
14. Jiang Z, Bo L, Wang L, et al. Interpretable machine-learning model for real-time, clustered risk factor analysis of sepsis and septic death in critical care. Comput Methods Programs Biomed 2023 Nov;241:107772. [doi: 10.1016/j.cmpb.2023.107772] [Medline: 37657148]
15. Slack D, Hilgard S, Jia E, Singh S, Lakkaraju H. Fooling LIME and SHAP: adversarial attacks on post hoc explanation methods. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society 2020:180-186. [doi: 10.1145/3375627.3375830]
16. Ma M, Wan X, Chen Y, et al. A novel explainable online calculator for contrast-induced AKI in diabetics: A multi-centre validation and prospective evaluation study. J Transl Med 2023 Jul 31;21(1):517. [doi: 10.1186/s12967-023-04387-x] [Medline: 37525240]
17. Mosquera-Lopez C, Ramsey KL, Roquemen-Echeverri V, Jacobs PG. Modeling risk of hypoglycemia during and following physical activity in people with type 1 diabetes using explainable mixed-effects machine learning. Comput Biol Med 2023 Mar;155:106670. [doi: 10.1016/j.compbiomed.2023.106670] [Medline: 36803791]

XSL•FO
RenderX

18. Syrowatka A, Song W, Amato MG, et al. Key use cases for artificial intelligence to reduce the frequency of adverse drug events: a scoping review. Lancet Digit Health 2022 Feb;4(2):e137-e148. [doi: 10.1016/S2589-7500(21)00229-6] [Medline: 34836823]
19. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 2024 Apr 16;385:e078378. [doi: 10.1136/bmj-2023-078378] [Medline: 38626948]
20. van Buuren S, Oudshoorn CG. Multivariate Imputation by Chained Equations: Leiden: TNO; 2000. URL: https://stefvanbuuren.name/publications/MICE%20V1.0%20Manual%20TNO00038%202000.pdf
21. Kornelius E, Chiou JY, Yang YS, Peng CH, Lai YR, Huang CN. Iodinated contrast media increased the risk of thyroid dysfunction: a 6-year retrospective cohort study. J Clin Endocrinol Metab 2015 Sep;100(9):3372-3379. [doi: 10.1210/JC.2015-2329] [Medline: 26168278]
22. Feldt S, Schüssel K, Quinzler R, et al. Incidence of thyroid hormone therapy in patients treated with sunitinib or sorafenib: a cohort study. Eur J Cancer 2012 May;48(7):974-981. [doi: 10.1016/j.ejca.2012.01.036] [Medline: 22382202]
23. Freund Y, Schapire RE. Experiments with a new boosting algorithm. ICML'96: Proceedings of the Thirteenth International Conference on International Conference on Machine Learning 1996 Jul:148-156 [FREE Full text]
24. Ke G, Meng Q, Finley T, et al. Lightgbm: A highly efficient gradient boosting decision tree [abstract]. 2017 Presented at: 31st Conference on Neural Information Processing Systems (NIPS 2017); Dec 4-9, 2017 URL: https://papers.nips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html [accessed 2025-03-31]
25. Friedman JH. Greedy function approximation: A gradient boosting machine. Ann Statist 2001;29(5):1189-1232. [doi: 10.1214/aos/1013203451]
26. Snoek J, Larochelle H, Adams RP. Practical bayesian optimization of machine learning algorithms [abstract]. 2012 Presented at: Advances in neural information processing systems; Dec 3-6, 2012; Lake Tahoe, Nevada URL: https://papers.nips.cc/paper_files/paper/2012/hash/05311655a15b75fab86956663e1819cd-Abstract.html
27. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. 2017 Presented at: 31st Conference on Neural Information Processing Systems (NIPS 2017); Dec 4-9, 2017; Long Beach, CA URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf [accessed 2025-03-31]
28. Lachin JM. Fallacies of last observation carried forward analyses. Clin Trials 2016 Apr;13(2):161-168. [doi: 10.1177/1740774515602688] [Medline: 26400875]
29. Escudier B, Eisen T, Stadler WM, et al. Sorafenib in advanced clear-cell renal-cell carcinoma. N Engl J Med 2007 Jan 11;356(2):125-134. [doi: 10.1056/NEJMoa060655] [Medline: 17215530]
30. Motzer RJ, Hutson TE, Tomczak P, et al. Overall survival and updated results for sunitinib compared with interferon alfa in patients with metastatic renal cell carcinoma. J Clin Oncol 2009 Aug 1;27(22):3584-3590. [doi: 10.1200/JCO.2008.20.1293] [Medline: 19487381]
31. Schmidinger M, Vogl UM, Bojic M, et al. Hypothyroidism in patients with renal cell carcinoma: blessing or curse? Cancer 2011 Feb 1;117(3):534-544. [doi: 10.1002/cncr.25422] [Medline: 20845482]
32. Silva NCD, Albertini MK, Backes AR, Pena GDG. Machine learning for hospital readmission prediction in pediatric population. Comput Methods Programs Biomed 2024 Feb;244:107980. [doi: 10.1016/j.cmpb.2023.107980] [Medline: 38134648]
33. Duan H, Zhang Y, Qiu H, et al. Machine learning-based prediction model for distant metastasis of breast cancer. Comput Biol Med 2024 Feb;169:107943. [doi: 10.1016/j.compbiomed.2024.107943] [Medline: 38211382]
34. May SB, Giordano TP, Gottlieb A. Generalizable pipeline for constructing HIV risk prediction models across electronic health record systems. J Am Med Inform Assoc 2024 Feb 16;31(3):666-673. [doi: 10.1093/jamia/ocad217]
35. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE 2015;10(3):e0118432. [doi: 10.1371/journal.pone.0118432] [Medline: 25738806]
36. Lipton ZC, Elkan C, Naryanaswamy B. Optimal thresholding of classifiers to maximize F1 measure. Mach Learn Knowl Discov Databases 2014;8725:225-239. [doi: 10.1007/978-3-662-44851-9_15] [Medline: 26023687]
37. Rizos CV, Elisaf MS, Liberopoulos EN. Effects of thyroid dysfunction on lipid profile. Open Cardiovasc Med J 2011;5:76-84. [doi: 10.2174/1874192401105010076] [Medline: 21660244]
38. Porta C, Gore ME, Rini BI, et al. Long-term safety of sunitinib in metastatic renal cell carcinoma. Eur Urol 2016 Feb;69(2):345-351. [doi: 10.1016/j.eururo.2015.07.006] [Medline: 26215605]
39. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med 2020;3:17. [doi: 10.1038/s41746-020-0221-y] [Medline: 32047862]
40. Flynn A. Using artificial intelligence in health-system pharmacy practice: finding new patterns that matter. Am J Health Syst Pharm 2019 Apr 17;76(9):622-627. [doi: 10.1093/ajhp/zxz018] [Medline: 31361834]

Abbreviations

AdaBoost: Adaptive Boosting

XSL•FO
RenderX

ADR: adverse drug reaction
ALT: alanine transaminase
AST: aspartate aminotransferase
AUPRC: area under the precision recall curve
AUROC: area under the receiver operating characteristic curve
CDSS: clinical decision support systems
CRD: clinical research database
GBDT : Gradient Boosting Decision Tree
LGBM: Light Gradient Boosting Machine
LR: logistic regression
ML: machine learning
RF: random forest
TKI: tyrosine kinase inhibitor
TMUH: Taipei Medical University Hospital
TRIPOD+AI: Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis + artificial intelligence
TSH: thyroid-stimulating hormone

Edited by A Mavragani; submitted 23.10.24; peer-reviewed by AP Vit, L Guo; revised version received 03.03.25; accepted 04.03.25; published 10.04.25.

Please cite as:
Chan FY, Ku YE, Lie WN, Chen HY
Web-Based Explainable Machine Learning-Based Drug Surveillance for Predicting Sunitinib- and Sorafenib-Associated Thyroid Dysfunction: Model Development and Validation Study
JMIR Form Res 2025;9:e67767
URL: https://formative.jmir.org/2025/1/e67767
doi:10.2196/67767

© Fan-Ying Chan, Yi-En Ku, Wen-Nung Lie, Hsiang-Yin Chen. Originally published in JMIR Formative Research (https://formative.jmir.org), 10.4.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Assessment of Gait Parameters Using Wearable Sensors and Their Association With Muscle Mass, Strength, and Physical Performance in Korean Older Adults: Cross-Sectional Study

Jinyoung Shin[1], MD, PhD; Hyuk Jung Kweon[2], MD, PhD; Jaekyung Choi[1], MD, PhD

[1]Department of Family Medicine, Konkuk University Medical Center, Konkuk University School of Medicine, Seoul, Republic of Korea

[2]Department of Family Medicine, Konkuk University Medical Center, Chungju Hospital, Konkuk University School of Medicine, 268 Chungwon-daero, Chungju, Republic of Korea

Corresponding Author:
Hyuk Jung Kweon, MD, PhD
Department of Family Medicine, Konkuk University Medical Center, Chungju Hospital, Konkuk University School of Medicine, 268 Chungwon-daero, Chungju, Republic of Korea

Abstract

Background: Gait speed indicates the onset or decline of physical performance in sarcopenia. However, real-time measurements of other gait parameters, such as step length, stride length, step width, and support time, are limited. The advent of wearable technology has facilitated the measurement of these parameters, necessitating further investigation into their potential applications.

Objective: This study aimed to investigate the relationship between gait parameters measured using wearable sensors and muscle mass, strength, and physical performance in community-dwelling older adults.

Methods: In a cross-sectional study of 91 participants aged ≥65 years, gait parameters, such as step count, step length, cadence, single and double support times, vertical oscillation, and instantaneous vertical loading rate (IVLR), measured using a wireless earbud device, were analyzed on the basis of the appendicular skeletal muscle mass index (SMI), calf circumference, handgrip strength, 5-time chair stand test, short physical performance battery (SPPB), and the SARC-F (strength, assistance with walking, rise from a chair, climb stairs and fall frequency) questionnaire. This study was conducted from July 10 to November 1, 2023, at an outpatient clinic of a university hospital in Seoul, Korea. Multiple regression analysis was performed to investigate independent associations after adjusting for age, sex, BMI, and comorbidities.

Results: Among 91 participants (45 men and 46 women; mean age 74.1 years for men and 73.6 years for women), gait speed and vertical oscillation showed negative associations with their performance in the 5-time chair stand test ($P<.001$) and SARC-F and positive associations with their performance in the SPPB ($P<.001$). Vertical oscillations were also associated with grip strength ($P=.003$). Single and double support times were associated with performance in the 5-time chair stand test and SPPB ($P<.001$). In addition, double support time was associated with SARC-F scores ($P<.001$). Gait speed, support time, vertical oscillation, and IVLR showed independent associations with performance in the 5-time chair stand test and SPPB ($P<.001$), both related to muscle strength or physical performance. Gait speed, double support time, and vertical oscillation all had significant associations with SARC-F scores.

Conclusions: This study demonstrated a significant association between gait monitoring using wearable sensors and quantitative assessments of muscle strength and physical performance in older people. Furthermore, this study substantiated the extensive applicability of diverse gait parameters in predicting sarcopenia.

(JMIR Form Res 2025;9:e63928) doi:10.2196/63928

KEYWORDS

gait analysis; sarcopenia; wearable electronic devices; muscle mass; physical performance; older adults; geriatric; cross-sectional study; outpatient clinic; Korea; mHealth; mobile health

Introduction

Sarcopenia is an age-related condition characterized by the progressive loss of skeletal muscle mass and strength, as well as physical performance [1], which increases the risk of disability, falls, cognitive decline, loss of independence, and mortality among community-dwelling older adults [2-5]. The International Clinical Practice Guidelines for Sarcopenia recommend annual clinical assessments or more frequent evaluations after major health events, such as hospitalization, to detect early signs of frailty [6,7]. According to the 2019 Asian Working Group for Sarcopenia (AWGS) consensus update, the diagnosis of sarcopenia should be based on changes in muscle strength, physical performance, and muscle mass, with case

XSL•FO
RenderX

findings leading to a comprehensive assessment and diagnosis in primary care or community preventive settings [8-10].

Gait monitoring, particularly gait speed, has been identified as a valid indicator of sarcopenia and is associated with health outcomes such as disability and mortality in older adults [11,12]. Although various gait parameters such as step length, stride length and width, and single support time during walking can be detected, their use in research has been limited owing to challenges in real-time measurement and result standardization [13]. Wearable sensors have made it easier to conduct real-time evaluations. A meta-analysis of 82 studies addressing over 100 gait parameters outcomes discovered high validity and reliability, particularly for spatiotemporal parameters such as step time and stride time [14].

Because smartphone apps enable the measurement of spatiotemporal outcomes while older adults walk, the clinical utility of real-time monitoring of changes in mobility among older adults using wearable sensors has been highlighted [15]. Daily walking count guidance using wearable sensors improves physical activity and reduces sarcopenia in older adults in Taiwan [16]. A study examining men older than 50 years in Korea using a wearable smart belt for 4 weeks discovered significant correlations between gait speed and sarcopenia, as defined by grip strength and dual-energy x-ray absorptiometry [17]. Several studies measured gait parameters using watches or footwear [18]. However, few studies have investigated the association between various gait parameters other than gait speed or step length and sarcopenia diagnosis based on muscle strength, muscle mass, and physical performance using an earbud device [19,20]. This study aimed to investigate the relationship between various gait parameters measured using wearable sensors (particularly earbud-type sensors) and sarcopenia, based on muscle mass, strength, and physical performance, among community-dwelling older adults.

Methods

Study Design and Participants

This is a cross-sectional study. The primary objective of this study was to investigate the association between gait parameters and various sarcopenia measurements such as muscle mass, strength, and physical performance. The secondary objective was to determine if newly usable gait parameters, which can be easily measured using wearable sensors, were significantly associated with sarcopenia. In the initial study, the target sample size for both men and women was 45 individuals in an initial study [21]. This analysis was performed as a secondary study, using data from a comparison study of a portable device for muscle mass and bioelectrical impedance analysis (BIA). At an older adult community center, we recruited 91 participants aged 65 years or older who could perform daily activities independently. We displayed a recruitment poster at the community center, inviting older adults interested in the study to apply to participate. This study was performed at the outpatient clinic of a university hospital in Seoul, Korea, between July 10 and November 1, 2023. The exclusion criterion was ongoing treatment for uncontrolled conditions such as malignancy, acute stroke, or dementia.

Muscle Mass, Strength, and Physical Performance Measurements

Muscle mass was determined using the appendicular skeletal muscle mass index (SMI) and calf circumference. The SMI is calculated by dividing the total lean mass of both arms and legs by the square of the height (kg/m^2) as measured using BIA (InBody 770; Cerritos). Calf circumference was measured at the thickest point on both sides, with moderate to high sensitivity and specificity for detecting low skeletal muscle mass [10,22]. Muscle strength was determined by the grip strength measured 3 times at 30-second intervals for each hand in a standing position, alternating between the dominant and nondominant sides, using a Smedley handheld dynamometer [23]. The 5-time chair stand test, which assesses both physical performance and muscle strength, requires participants to stand and sit 5 times as quickly as possible from a straight-backed chair with no arm support [24]. A short physical performance battery (SPPB) was used to assess balance, gait, and lower limb strength. This included standing with feet side-by-side, semitandem, and tandem positions; walking 8 feet; and rising from a chair to return to a seated position 5 times [25]. The SARC-F (strength, assistance with walking, rise from a chair, climb stairs and fall frequency) questionnaire was used to assess sarcopenia risk based on its 5 components, with fall frequency ranging 0-10. A total score of ≥4 indicates sarcopenia. The Korean version of the SARC-F was previously validated [26,27]. According to the AWGS consensus report in 2014, we used the cutoff values for SMI (less than 7.0 kg/m^2 for men and less than 5.7 kg/m^2 for women using BIA) and calf circumference (<34 cm for men and <33 cm for women) [1]. Low muscle strength was defined as grip strength of <28 kg for men and <18 kg for women. The cutoff value for the 5-time chair stand test was 12 seconds or more, while that of the SPPB was ≤9 (range 0 - 12) points [10].

Gait Parameters

Gait characteristics were measured using a wireless earbud (Beflex) device (Figure 1) equipped with a BiomechEngine chip for accurate accelerometer data analysis and gait parameter estimation. In young adults, the earbud device's validity was confirmed by measuring walking and running parameters [28]. Participants were instructed to insert the device into both ears and walk for a minimum of 30 seconds to record various gait metrics such as step count, step length, cadence, gait speed, times of single and double support, vertical oscillation, and instantaneous vertical loading rate (IVLR). Step count is defined as the number of steps analyzed within a 2-second interval, where a step from the left foot to the right foot is considered a single cycle of 2 steps. Step length was calculated as the distance between the heel of the front foot and heel of the back foot during the walking cycle. Gait speed and cadence were automatically calculated. Single support time was defined as the average duration of one foot supporting the body while walking. Double support time was defined as the average duration of both feet's support and vertical oscillation, which was defined as the average distance of the body's center moved vertically during walking [28]. The IVLR measures the steepest slope of the vertical ground reaction force at the points of left- and right-foot contact. The algorithms used to estimate these

XSL•FO
RenderX

gait parameters are not currently disclosed; however, they are commercially available. Gait parameters were measured a few minutes after walking.

Figure 1. Wearable sensors for gait monitoring (5.7 g, 17.1 × 20.2 × 22 mm^3).

Ethical Considerations

The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the institutional review board of Konkuk University Medical Center (KUMC 2022-11-033). Informed consent was obtained from all subjects enrolled in the study. Written informed consent was obtained from the patients for publication of this paper. The authors attest that there was no use of generative artificial intelligence technology in the generation of text, figures, or other informational content in this manuscript.

Statistical Analysis

Descriptive statistics were used to summarize participants' characteristics based on sex. Independent samples 2-tailed *t* tests were used to compare continuous variables, whereas chi-square tests were used for categorical variables. Pearson correlation analysis was used to determine the relationship between gait parameters and sarcopenia. We performed a logarithmic transformation to measure gait, muscle, and physical performance to adjust for a skewed distribution. After adjusting for age, sex, BMI, and comorbidities, multiple regression analysis was used to investigate the independent associations between sarcopenia and specific gait parameters for the secondary end point. Participants reported their age, sex, and comorbidities, as well as any daily life difficulties caused by low vision or hearing impairment ("yes" or "no"). Height and weight were measured to calculate the BMI (kg/m^2) and waist circumference at the narrowest point. Bonferroni correction for the evidential threshold was performed based on the number of exposures for the 6 sarcopenia indices; therefore, statistical significance was set at $P<.05$ divided by 6 (ie, a threshold of $P<.0083$ per exposure). All statistical analyses were conducted using IBM SPSS Statistics (version 27.0).

Results

This study included 91 participants (45 men and 46 women). The participants' characteristics are summarized in Table 1. The mean age of the study participants was 73.9 years, with more than half of them aged between 65 and 74 years. According to Asian criteria for overweight and obesity, the participants had a normal BMI and waist circumference [29].

While there were differences in grip strength, calf circumference, and SMI between men and women, there were no differences in their performance in the 5-time chair stand test, SPPB, or SARC-F (Table 2). The prevalence of sarcopenia varied according to the definition used for each measurement, ranging from 4.3% (SARC-F) to 82.6% (5-time chair stand test). The prevalence of sarcopenia according to grip strength was higher in women than in men. However, the prevalence of sarcopenia, defined by the SMI, was higher in men than in women. There were no in the other indicators between men and women.

Table 3 lists the association between various gait parameters and sarcopenia measurements. Gait speed, single and double support times, vertical oscillation, and IVLR demonstrated significant correlations with performance on the 5-time chair stand test and the SPPB. Additionally, gait speed, double support time, and vertical oscillation were associated with performance on the SARC-F. Vertical oscillation positively correlated with grip strength, and step count negatively correlated with performance on the 5-time chair stand test. However, after Bonferroni correction, step length and cadence did not show

XSL•FO
RenderX

any significant correlations with sarcopenia. Calf circumference and SMI were not associated with gait parameters.

After adjusting for confounding factors, gait speed, double support time, and vertical oscillation were significantly correlated with the 5-time chair stand test, SPPB, and SARC-F scores (Table 4). Single support time and IVLR were significantly correlated with performance in the 5-time chair stand test and SPPB.

Table . Characteristics of participants.

		Total (n=91)	Men (n=45)	Women (n=46)
Age (years), mean (SD)		73.9 (5.5)	74.1 (5.0)	73.6 (6.1)
Age groups (years), n (%)				
	65 - 74	51 (56.0)	24 (53.3)	27 (58.7)
	75 - 84	38 (41.8)	21 (46.7)	17 (37.0)
	≥85	2 (2.2)	0 (0)	2 (4.3)
BMI (kg/m^2), mean (SD)		24.4 (2.8)	24.6 (2.5)	24.2 (3.1)
Waist circumference (cm), mean (SD)		80.0 (8.9)	83.0 (7.4)	77.0 (9.3)
Number of comorbidities, mean (SD)		1.46 (1.07)	1.58 (0.97)	1.35 (1.16)
Comorbidities, n (%)				
	None	17 (18.7)	5 (11.1)	12 (26.1)
	Cardiovascular disease	69 (75.8)	40 (88.9)	29 (63.0)
	Musculoskeletal disease	14 (15.4)	2 (4.4)	12 (26.1)

Table . Diagnostic measurements of sarcopenia in study participants.

	Measurements			Sarcopenia		
	Men (n=45), mean (SD)	Women (n=46), mean (SD)	*P* value	Men (n=45), n (%)	Women (n=46), n (%)	*P* value
5-time chair stand test[a]	15.7 (5.5)	15.7 (4.6)	.96	33 (73.3)	38 (82.6)	.29
SPPB[b]	9.0 (1.5)	8.9 (1.5)	.73	29 (64.4)	29 (63.0)	.89
SARC-F[c]	0.66 (1.51)	0.87 (1.34)	.50	3 (6.7)	2 (4.3)	.63
Grip strength[d]	28.8 (5.6)	18.1 (4.8)	<.001	13 (28.9)	23 (50.0)	<.001
Calf circumference[e]	34.2 (2.8)	32.1 (2.5)	<.001	20 (44.4)	25 (54.3)	.35
Skeletal muscle index[f]	7.7 (0.71)	6.2 (0.6)	<.001	7 (15.6)	5 (10.9)	.02

[a]Cutoff was ≥12 seconds.

[b]SPPB: short physical performance battery; cutoff was ≤9 points.

[c]SARC-F: strength, assistance with walking, rise from a chair, climb stairs and fall frequency; cutoff was ≥4 points.

[d]Cutoff was <28 kg for men and <18 kg for women.

[e]Cutoff was <34 cm for men and <33 cm for women.

[f]Cutoff was <7.0 kg/m^2 for men and <5.7 kg/m^2 for women.

XSL•FO
RenderX

Table . Associations between gait parameters and sarcopenia measurements (n=91). All variables were subjected to correlation analysis after log transformation and Bonferroni correction was applied (*P*<.05/6 [.0083]).

		5-time chair stand test	SPPB[a]	SARC-F[b]	Grip strength	Calf circumference	SMI[c]
Step count							
	Coefficient	−0.322	0.163	−0.137	−0.071	−0.198	−0.204
	P value	.002	.12	.20	.50	.06	.053
Step length							
	Coefficient	−0.221	0.237	−0.209	−0.162	−0.176	−0.162
	P value	.04	.02	.047	.13	.10	.13
Cadence							
	Coefficient	−0.018	−0.036	−0.191	0.057	−0.041	0.040
	P value	.86	.75	.07	.59	.70	.71
Gait speed							
	Coefficient	−0.437	0.381	−0.281	−0.176	−0.196	−0.230
	P value	<.001	<.001	.007	.10	.06	.03
Single support time							
	Coefficient	0.489	−0.350	0.176	0.167	0.158	0.255
	P value	<.001	.001	.10	.11	.14	.02
Double support time							
	Coefficient	0.604	−0.541	0.371	−0.007	0.083	0.151
	P value	<.001	<.001	<.001	.95	.43	.15
Vertical oscillation							
	Coefficient	−0.387	0.518	−0.467	0.306	0.107	0.198
	P value	<.001	<.001	<.001	.003	.31	.06
IVLR[d]							
	Coefficient	−0.530	0.449	−0.270	−0.032	−0.128	−0.177
	P value	<.001	<.001	.01	.76	.23	.09

[a]SPPB: short physical performance battery.

[b]SARC-F: strength, assistance with walking, rise from a chair, climb stairs and fall frequency.

[c]SMI: appendicular skeletal muscle mass index.

[d]IVLR: instantaneous vertical loading rate.

Table . Multiple regression analysis for sarcopenia and gait indexes in older adults, having adjusted for age, sex, waist circumference, BMI, and comorbidities. Bonferroni correction was applied (*P*<.05/6 [.0083]).

	5-time chair stand test		SPPB[a]		SARC-F[b]	
	Coefficient	*P* value	Coefficient	*P* value	Coefficient	*P* value
Gait speed	−0.393	<.001	0.331	.001	−0.328	.003
Single support time	0.545	<.001	−0.384	.001	0.222	.055
Double support time	0.558	<.001	−0.452	<.001	0.389	<.001
Vertical oscillation	−0.306	.008	0.399	<.001	−0.528	<.001
IVLR[c]	−0.496	<.001	0.392	<.001	−0.267	.02

[a]SPPB: short physical performance battery.

[b]SARC-F: strength, assistance with walking, rise from a chair, climb stairs and fall frequency.

[c]IVLR: Instantaneous vertical loading rate.

XSL•FO
RenderX

Discussion

Principal Findings

Gait measurements using wearable sensors show a strong correlation with physical performance and muscle strength. This study confirmed that the gait speed, double support time, and vertical oscillation measured using wearable sensors were reliable walking indicators of sarcopenia in community-dwelling older adults. Furthermore, using real-time gait monitoring measurements, we found that single support time and IVLR were good indicators of sarcopenia.

Age-related alterations in gait, such as shorter steps, increased double support time, lower cadence, or a wider step width, indicate a decline in gait speed [30]. These changes are attributed to a combination of sensorimotor factors, including motor unit reduction, impaired muscular activation, fiber-type substitution, decreased cutaneous sensation, and slower nerve and reaction speeds [12]. Furthermore, functional brain impairment, manifested as reduced brain volume, may be associated with deterioration in gait performance [31]. Accordingly, among the diagnostic indicators of sarcopenia, gait parameters may show a high correlation with the physical performance of participants.

However, these gait parameters were not associated with calf circumference or SMI, which are indicators of muscle mass. BIA, combined with the measurement of anthropometric metrics such as calf circumference, provides a practical approach for estimating muscle mass in nonclinical settings [11]. Individual factors, such as edema or diuretics may influence BIA [32], and muscle mass changes in older adults are not uniform [22], necessitating caution when interpretating their correlation with gait parameters. However, it is important to note that muscle mass is not a good predictor of adverse health–related outcomes such as mobility limitation, falls, and mortality in community-dwelling older adults [9]. Therefore, sarcopenia in older adults should be diagnosed holistically with muscle strength and physical performance taking precedence over muscle mass. Although this study did not identify a universal correlation between handgrip strength and gait parameters, quadriceps strength has been shown to influence gait speed, step length, and cadence [30]. This emphasizes the limitations of handgrip strength as an indicator of overall muscle strength in healthy older adults [33,34]. Further studies are required to determine the association between upper and lower limb muscle strength and gait parameters.

Among the gait parameters, gait speed, either alone or as part of a combined tool, such as the SPPB, has been confirmed to be associated with sarcopenia, which is the most comprehensive indicator of brain function and structural abnormalities [35]. Double support time correlates with balance in the cortico-subcortical white matter sensory and motor tracts [35]. Gait speed and double support time are direct indicators obtained while walking, whereas vertical oscillation and IVLR are known to be sensitive in reflecting changes in brain function as indirect data derived from direct indicators [36]. We confirmed the independent association between vertical oscillation and sarcopenia. Therefore, analyzing various gait parameters could improve the accuracy of sarcopenia diagnosis in older adults. Gait measurement using wearable sensors, which are easy to use and allow continuous and periodic monitoring of daily activities, offers significant promise for the early detection of sarcopenia in both the home and caregiving facilities. In general, older adults have difficulty with consistent use on consecutive days. Therefore, compliance with wearable sensors must be considered when evaluating the effectiveness of wearables in monitoring physical activity in older populations [37].

Limitations

However, several limitations of this study should be considered when applying gait measurements to older adults. First, the majority of participants were 65-74 years old, relatively healthy, and capable of independent daily living; hence, our results may not be applicable to all older populations. Additional research is required to apply this device to older adults living in nursing homes or other specialized care facilities. Second, we were unable to compare differences in wearable sensor techniques because our study used earbud-type devices rather than those more commonly placed on the foot or shank or used an insole pressure-sensing method [38]. We did not conduct repeated tests to assess test-retest reliability because it was difficult to revisit older participants. Third, because this study was conducted using a cross-sectional design, we could not establish a temporal relationship between the deterioration of gait parameters and sarcopenia. It is important to note that differences in the prevalence of sarcopenia based on the definition of muscle mass, strength, and physical performance may influence the correlation with gait parameters. Specifically, because our study did not compare gait parameters based on different rates of sarcopenia diagnosis, clinical interpretation should be cautious when predicting sarcopenia directly. Because race or lifestyle habits may have affected our results, it is important to consider the characteristics of the participants when applying the findings of this study. Finally, single and double support times, vertical oscillation, and IVLR still have limitations in practical applications, because diagnostic criteria have yet to be established. The algorithms for estimating these gait parameters are not currently disclosed; therefore, further studies are required to identify the clinical cutoff points to predict sarcopenia.

Conclusions

Using earbud-type wearable sensors, this study confirmed that gait parameters, particularly gait speed, double support time, and vertical oscillation, are correlated to muscle strength and physical performance in community-dwelling older adults. Given the limitations associated with the current diagnostic criteria and the potential unreliability of single measurements, further research is imperative to establish efficacy measures and broaden the applicability of diverse gait parameters in sarcopenia diagnosis.

XSL•FO
RenderX

Acknowledgments

We thank B-Lab and Beflex Inc for their cooperation. This research was funded by the Olive Healthcare (South Korea). The funders had no role in the design of the study, collection of data, analyses, interpretation of data, writing of the manuscript, or the decision to publish the results.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: JS, JC, and HJK
Formal analysis: JS and HJK
Funding acquisition: JS
Methodology: JS, JC, and HJK
Writing—original draft: JS
Writing—review and editing: JC and HJK

Conflicts of Interest

None declared.

References

1. Chen LK, Liu LK, Woo J, et al. Sarcopenia in Asia: consensus report of the Asian Working Group for Sarcopenia. J Am Med Dir Assoc 2014 Feb;15(2):95-101. [doi: 10.1016/j.jamda.2013.11.025] [Medline: 24461239]
2. Liang C, Shi L, Li B, He Z. The mediating role of sarcopenia in the association between physical activity and falls among Chinese older adults: a cross-sectional study. Healthcare (Basel) 2023 Dec 12;11(24):3146. [doi: 10.3390/healthcare11243146] [Medline: 38132036]
3. Chou MY, Nishita Y, Nakagawa T, et al. Role of gait speed and grip strength in predicting 10-year cognitive decline among community-dwelling older people. BMC Geriatr 2019 Jul 5;19(1):186. [doi: 10.1186/s12877-019-1199-7] [Medline: 31277579]
4. Wang DXM, Yao J, Zirek Y, Reijnierse EM, Maier AB. Muscle mass, strength, and physical performance predicting activities of daily living: a meta-analysis. J Cachexia Sarcopenia Muscle 2020 Feb;11(1):3-25. [doi: 10.1002/jcsm.12502] [Medline: 31788969]
5. Oh MH, Shin HE, Kim KS, Won CW, Kim M. Combinations of Sarcopenia Diagnostic Criteria by Asian Working Group of Sarcopenia (AWGS) 2019 guideline and incident adverse health outcomes in community-dwelling older adults. J Am Med Dir Assoc 2023 Aug;24(8):1185-1192. [doi: 10.1016/j.jamda.2023.04.005] [Medline: 37209713]
6. Dent E, Morley JE, Cruz-Jentoft AJ, et al. Physical frailty: ICFSR International Clinical Practice Guidelines for Identification and Management. J Nutr Health Aging 2019;23(9):771-787. [doi: 10.1007/s12603-019-1273-z] [Medline: 31641726]
7. Dent E, Morley JE, Cruz-Jentoft AJ, Arai H, Kritchevsky SB, Guralnik J, et al. International Clinical Practice Guidelines for Sarcopenia (ICFSR): screening, diagnosis and management. J Nutr Health Aging 2018;22(10):1148-1161. [doi: 10.1007/s12603-018-1139-9] [Medline: 30498820]
8. Voulgaridou G, Tyrovolas S, Detopoulou P, et al. Diagnostic criteria and measurement techniques of sarcopenia: a critical evaluation of the up-to-date evidence. Nutrients 2024 Feb 1;16(3):436. [doi: 10.3390/nu16030436] [Medline: 38337720]
9. Bhasin S, Travison TG, Manini TM, et al. Sarcopenia definition: the position statements of the Sarcopenia Definition and Outcomes Consortium. J Am Geriatr Soc 2020 Jul;68(7):1410-1418. [doi: 10.1111/jgs.16372] [Medline: 32150289]
10. Chen LK, Woo J, Assantachai P, et al. Asian Working Group for Sarcopenia: 2019 consensus update on sarcopenia diagnosis and treatment. J Am Med Dir Assoc 2020 Mar;21(3):300-307. [doi: 10.1016/j.jamda.2019.12.012] [Medline: 32033882]
11. Mijnarends DM, Meijers JMM, Halfens RJG, et al. Validity and reliability of tools to measure muscle mass, strength, and physical performance in community-dwelling older people: a systematic review. J Am Med Dir Assoc 2013 Mar;14(3):170-178. [doi: 10.1016/j.jamda.2012.10.009] [Medline: 23276432]
12. Abellan van Kan G, Rolland Y, Andrieu S, et al. Gait speed at usual pace as a predictor of adverse outcomes in community-dwelling older people an International Academy on Nutrition and Aging (IANA) Task Force. J Nutr Health Aging 2009 Dec;13(10):881-889. [doi: 10.1007/s12603-009-0246-z] [Medline: 19924348]
13. Lord S, Howe T, Greenland J, Simpson L, Rochester L. Gait variability in older adults: a structured review of testing protocol and clinimetric properties. Gait Posture 2011 Oct;34(4):443-450. [doi: 10.1016/j.gaitpost.2011.07.010] [Medline: 21920755]
14. Kobsar D, Charlton JM, Tse CTF, et al. Validity and reliability of wearable inertial sensors in healthy adult walking: a systematic review and meta-analysis. J Neuroeng Rehabil 2020 May 11;17(1):62. [doi: 10.1186/s12984-020-00685-3] [Medline: 32393301]

XSL•FO
RenderX

15. Contreras C, Stanley EC, Deschamps-Prescott C, Burnap S, Hopkins M, Browning B, et al. Evaluation of smartphone technology on spatiotemporal gait in older and diseased adult populations. Sensors (Basel) 2024 Sep 9;24(17):5839. [doi: 10.3390/s24175839] [Medline: 39275750]
16. Ho MH, Peng CY, Liao Y, Yen HY. Efficacy of a wearable activity tracker with step-by-step goal-setting on older adults' physical activity and sarcopenia indicators: clustered trial. J Med Internet Res 2024 Nov 1;26:e60183. [doi: 10.2196/60183] [Medline: 39486024]
17. Kang MG, Kang SJ, Roh HK, et al. Accuracy and diversity of wearable device-based gait speed measurement among older men: observational study. J Med Internet Res 2021 Oct 11;23(10):e29884. [doi: 10.2196/29884] [Medline: 34633293]
18. Biswas N, Chakrabarti S, Jones LD, Ashili S. Smart wearables addressing gait disorders: a review. Mater Today Commun 2023 Jun;35:106250. [doi: 10.1016/j.mtcomm.2023.106250]
19. Seifer AK, Kuderle A, Dorschky E, Moradi H, Hannemann R, Eskofier BM. Step length and gait speed estimation using a hearing aid integrated accelerometer: a comparison of different algorithms. IEEE J Biomed Health Inform 2024 Nov;28(11):6619-6628. [doi: 10.1109/JBHI.2024.3454824] [Medline: 39236137]
20. Seifer AK, Dorschky E, Küderle A, Moradi H, Hannemann R, Eskofier BM. EarGait: estimation of temporal gait parameters from hearing aid integrated inertial sensors. Sensors (Basel) 2023 Jul 20;23(14):6565. [doi: 10.3390/s23146565] [Medline: 37514858]
21. Shin J, Park E. Comparison between discrete multi-wavelength near-infrared spectroscopy and bioelectrical impedance analysis in the assessment of muscle mass for community-dwelling older people. J Clin Med 2024 Apr 18;13(8):2350. [doi: 10.3390/jcm13082350] [Medline: 38673621]
22. Hansen SS, Munk T, Knudsen AW, Beck AM. Concordance between changes in calf circumference and muscle mass exists: a narrative literature review. Clin Nutr ESPEN 2024 Feb;59:171-175. [doi: 10.1016/j.clnesp.2023.11.026] [Medline: 38220372]
23. do Amaral C, da Luz Goulart C, da Silva BM, et al. Low handgrip strength is associated with worse functional outcomes in long COVID. Sci Rep 2024 Jan 24;14(1):2049. [doi: 10.1038/s41598-024-52401-z] [Medline: 38267519]
24. Ryu JY, Kim M, Kim KS, Kim S, Won CW. Chair stand test as a proxy for physical performance and muscle strength in sarcopenia diagnosis: the Korean Frailty and Aging Cohort Study. Aging Clin Exp Res 2022 Oct;34(10):2449-2456. [doi: 10.1007/s40520-022-02172-2] [Medline: 35918606]
25. Guralnik JM, Simonsick EM, Ferrucci L, et al. A short physical performance battery assessing lower extremity function: association with self-reported disability and prediction of mortality and nursing home admission. J Gerontol 1994 Mar;49(2):M85-M94. [doi: 10.1093/geronj/49.2.m85] [Medline: 8126356]
26. Kim M, Won CW. Sarcopenia in Korean community-dwelling adults aged 70 years and older: application of screening and diagnostic tools from the Asian Working Group for Sarcopenia 2019 Update. J Am Med Dir Assoc 2020 Jun;21(6):752-758. [doi: 10.1016/j.jamda.2020.03.018] [Medline: 32386844]
27. Kim S, Kim M, Won CW. Validation of the Korean Version of the SARC-F questionnaire to assess sarcopenia: Korean Frailty and Aging Cohort Study. J Am Med Dir Assoc 2018 Jan;19(1):40-45. [doi: 10.1016/j.jamda.2017.07.006] [Medline: 28843526]
28. Jung CK, Kim J, Rhim HC. Validation of an ear-worn wearable gait analysis device. Sensors (Basel) 2023 Jan 21;23(3):1244. [doi: 10.3390/s23031244] [Medline: 36772282]
29. WHO Expert Consultation. Appropriate body-mass index for Asian populations and its implications for policy and intervention strategies. Lancet 2004 Jan 10;363(9403):157-163. [doi: 10.1016/S0140-6736(03)15268-3] [Medline: 14726171]
30. Callisaya ML, Blizzard L, Schmidt MD, McGinley JL, Lord SR, Srikanth VK. A population-based study of sensorimotor factors affecting gait in older people. Age Ageing 2009 May;38(3):290-295. [doi: 10.1093/ageing/afp017] [Medline: 19264860]
31. Rosano C, Aizenstein H, Brach J, Longenberger A, Studenski S, Newman AB. Special article: gait measures indicate underlying focal gray matter atrophy in the brain of older adults. J Gerontol A Biol Sci Med Sci 2008 Dec;63(12):1380-1388. [doi: 10.1093/gerona/63.12.1380] [Medline: 19126852]
32. Genton L, Karsegard VL, Kyle UG, Hans DB, Michel JP, Pichard C. Comparison of four bioelectrical impedance analysis formulas in healthy elderly subjects. Gerontology 2001;47(6):315-323. [doi: 10.1159/000052821] [Medline: 11721145]
33. Porto JM, Nakaishi APM, Cangussu-Oliveira LM, Freire Júnior RC, Spilla SB, Abreu DCC. Relationship between grip strength and global muscle strength in community-dwelling older people. Arch Gerontol Geriatr 2019;82:273-278. [doi: 10.1016/j.archger.2019.03.005] [Medline: 30889410]
34. Samuel D, Rowe P. An investigation of the association between grip strength and hip and knee joint moments in older adults. Arch Gerontol Geriatr 2012;54(2):357-360. [doi: 10.1016/j.archger.2011.03.009] [Medline: 21481952]
35. Rosano C, Brach J, Longstreth WT Jr, Newman AB. Quantitative measures of gait characteristics indicate prevalence of underlying subclinical structural brain abnormalities in high-functioning older adults. Neuroepidemiology 2006;26(1):52-60. [doi: 10.1159/000089240] [Medline: 16254454]
36. Wang Z, Ren K, Li D, et al. Assessment of brain function in patients with cognitive impairment based on fNIRS and gait analysis. Front Aging Neurosci 2022 May 24;14:799732. [doi: 10.3389/fnagi.2022.799732] [Medline: 35686022]

XSL•FO
RenderX

37. Ding H, Ho K, Searls E, et al. Assessment of wearable device adherence for monitoring physical activity in older adults: pilot cohort study. JMIR Aging 2024 Oct 25;7:e60209. [doi: 10.2196/60209] [Medline: 39454101]
38. Prasanth H, Caban M, Keller U, et al. Wearable sensor-based real-time gait detection: a systematic review. Sensors (Basel) 2021 Apr 13;21(8):2727. [doi: 10.3390/s21082727] [Medline: 33924403]

Abbreviations

AWGS: Asian Working Group for Sarcopenia
BIA: bioelectrical impedance analysis
IVLR: instantaneous vertical loading rate
SARC-F: strength, assistance with walking, rise from a chair, climb stairs and fall frequency
SMI: appendicular skeletal muscle mass index
SPPB: short physical performance battery

Edited by A Mavragani; submitted 03.07.24; peer-reviewed by G Welk, S Okita; revised version received 20.11.24; accepted 18.12.24; published 10.04.25.

Please cite as:
Shin J, Kweon HJ, Choi J
Assessment of Gait Parameters Using Wearable Sensors and Their Association With Muscle Mass, Strength, and Physical Performance in Korean Older Adults: Cross-Sectional Study
JMIR Form Res 2025;9:e63928
URL: https://formative.jmir.org/2025/1/e63928
doi:10.2196/63928

©Jinyoung Shin, Hyuk Jung Kweon, Jaekyung Choi. Originally published in JMIR Formative Research (https://formative.jmir.org), 10.4.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Centering Youth Voice in the Adaptation of an mHealth Intervention for Young Adults With HIV in South Texas, United States: Human-Centered Design Approach

Nhat Minh Ho[1], BS; Catherine Johnson[2], MSc; Autumn Chidester[3], MD, MPH; Ruby Viera Corral[1], BS; Jacundo Ramos[4], BS; Miguel Garcia[4], BFA; Rishi Gonuguntla[5], MD, MPH; Cyrena Cote[4], BS; Divya Chandramohan[4], MD; Hueylie Lin[1], MD, MPH; Anna Taranova[2], MD, MS; Ank E Nijhawan[6], MD, MPH, MSCS; Susan Kools[7], PhD; Karen Ingersoll[8], PhD; Rebecca Dillingham[8], MD, MPH; Barbara S Taylor[4], MD, MS

[1]Joe R. and Teresa Lozano Long School of Medicine, The University of Texas Health Science Center at San Antonio, San Antonio, TX, United States

[2]Department of Public Health, Innovation and Equity, University Health, San Antonio, TX, United States

[3]Department of Women's Health, Dell Medical School, The University of Texas at Austin, Austin, TX, United States

[4]Division of Infectious Diseases, Department of Medicine, The University of Texas Health Science Center at San Antonio, 7703 Floyd Curl Dr., MSC 7881, San Antonio, TX, United States

[5]Department of Psychiatry, Washington University School of Medicine in St. Louis, St Louis, MO, United States

[6]Division of Infectious Diseases, Departament of Medicine, The University of Texas Southwestern Medical Center, Dallas, TX, United States

[7]School of Nursing, University of Virginia, Charlottesville, VA, United States

[8]School of Medicine, University of Virginia, Charlottesville, VA, United States

Corresponding Author:
Barbara S Taylor, MD, MS
Division of Infectious Diseases, Department of Medicine, The University of Texas Health Science Center at San Antonio, 7703 Floyd Curl Dr., MSC 7881, San Antonio, TX, United States

Abstract

Background: Young adults living with HIV are less likely to engage in care and achieve viral suppression, compared to other age groups. Young adults living with HIV also have a high degree of self-efficacy and willingness to adopt novel care modalities, including mobile health (mHealth) interventions. Interventions to increase care engagement could aid young adults living with HIV in overcoming structural and social barriers and leveraging youth assets to improve their health outcomes.

Objective: The objective of the paper was to use an assets-based framework, positive youth development, and human-centered design principles to adapt an existing mHealth intervention, PositiveLinks (PL), to support care engagement for 18‐ to 29-year-olds with HIV.

Methods: We conducted a formative evaluation including semistructured interviews with 14 young adults with HIV and focus groups with 26 stakeholders (providers, nurses, case managers, and clinic staff). Interviews covered barriers to care, provider communication, and concerns or suggestions about mHealth interventions. The research team used thematic analysis to review interview transcripts. In the second phase, human-centered design processes informed adaptation of the existing PL platform using data from real-time use suggestions of 3 young adults with HIV. Throughout the formative evaluation and adaptation, a Youth Advisory Board (YAB) provided input.

Results: Young adults with HIV and stakeholders identified common elements of an mHealth intervention that would support care engagement including: the convenience of addressing needs through the app, online support groups to support interconnection, short videos or live chats with other young adults with HIV or providers, appointment and medication reminders, and medical information from a trustworthy source. Stakeholders also mentioned the need for youth empowerment. Concerns included worries about confidentiality, unintentional disclosures of status, urgent content in an unmoderated forum, and the impersonality of online platforms. Design suggestions from young adults with HIV included suggestions on appearance, new formatting for usability of the online support group, and prioritization of local content. Based on the feedback received, iterative changes were made to transform PL into Positive Links for Youth (PL4Y). Final votes on adaptations were made by the YAB. The overall appearance of the platform was changed, including logo, color, and font. The online support group was divided into 3 channels which support hashtags and content searches. The "Resources" and "Frequently Asked Questions" sections were condensed and revised to prioritize South Texas–specific content.

Conclusions: Our assets-based framework supported young adults with HIV and stakeholder input in the transformation of an mHealth intervention to meet the needs of 18- to 29-year-olds in South Texas. The human-centered design approach allowed

XSL•FO
RenderX

young adults with HIV to suggest specific changes to the intervention's design to support usability and acceptability. This adapted version, PL4Y, is now ready for pilot testing in the final phase of this implementation science project.

(JMIR Form Res 2025;9:e60531) doi:10.2196/60531

KEYWORDS

HIV; implementation science; youth; mHealth; adherence; young; mobile health; mobile health intervention; AIDS; US; adult; self-efficacy; willingness; health outcomes; mHealth intervention; interview; human-centered design; acceptability; usability; mobile phone

Introduction

Background

Young adults living with HIV face difficulty with care engagement [1]. While existing studies of young adults with HIV include varying age ranges, spanning 13 to 29 years, most agree that young adults with HIV across the age spectrum face unique challenges in care engagement. These include limited opportunities for care, higher rates of missed appointments, and medication nonadherence, all of which contribute to difficulties in achieving viral suppression [2-4]. Interventions to increase care engagement could help young adults with HIV overcome structural and social barriers to improve health outcomes. While young adults with HIV often seek mobile health technologies to support care engagement, few were developed or adapted with their input or needs in mind.

Most young adults with HIV have a high degree of self-efficacy and willingness to adopt novel care modalities, particularly technological innovations [5]. In 2018, a national survey revealed 45% of adolescents report being on the web almost constantly, and 95% of teens own a smartphone [6]. Mobile health (mHealth) interventions are well suited to these digital natives. Data show that mobile phone apps are an acceptable and often preferred way for adolescents to receive sexual health information and successfully engage those susceptible to HIV [7-10]. Considering the developmentally-appropriate desire for self-determination found among young adults with HIV, they may appreciate mHealth interventions offering an expanded range of services, particularly functions that support self-management of HIV and digital connections [11,12]. However, the majority of current mHealth interventions offer a limited number of functions, such as medication or appointment reminders [7,8,10,13].

mHealth Interventions and Care Engagement for Young Adults With HIV

A few mHealth interventions focusing on young adults with HIV have shown promising results [9,10,13]. A mobile intervention that included peer navigation and a webcomic series increased entry to care, decreased rates of loss to follow up, and supported sustained viral load suppression for young adults with HIV [9]. A review of mHealth intervention modalities for 18- to 34-year-old people living with HIV found advantages such as improved accessibility to information, enhanced self-efficacy, and reduced negative feelings about HIV diagnoses [13]. Other forms of mHealth interventions previously used in promoting the health of young adults with HIV include remote coaching, short message services, phone calls, and social media outreach [14,15]. mHealth intervention implementation faces the challenge of needing continuous updates based on input from stakeholders and app users [10].

Few studies in the existing literature describe how mHealth interventions can be tailored to the needs of young adults with HIV or adaptation processes that ensure they play an active role in intervention design. The Positive Youth Development (PYD) framework suggests that youth involvement and empowerment can lead to better health and reduced risk-taking behaviors [1,16,17]. Aligned with PYD, peer empowerment, knowledge seeking, and taking responsibility for health outcomes improve engagement in health-promoting processes and resilience among gay and bisexual youth with HIV [18]. These data suggest that the PYD framework may be used to incorporate youth input into intervention design and, ideally, lead to more impactful interventions.

Young adults with HIV in South Texas experience significant barriers to care engagement for HIV, making the region an important location to study innovative interventions [19]. In 2016, the US Centers for Disease Control and Prevention (CDC) began using phylogenetic analysis to identify geographic areas where HIV transmission and new diagnoses were more likely. The first iteration of this analysis identified several clusters of new diagnoses in Texas, the largest of which was in South Texas, where 78% of the individuals newly diagnosed were under 30 years old. The majority were also uninsured, without access to HIV prevention services, and Hispanic or Latino [20,21]. Other literature suggests that young men who have sex with men of color experience structural racism, which may alienate them from health care services [4,22]. Despite these challenges, South Texas also has a strong history of collaborative work between community clinics, the county health system, and the local health authority to end the HIV epidemic [23,24]. These factors make the region an ideal site to create and implement a new mHealth intervention targeted specifically toward young adults with HIV.

Investigation Goals and Significance

The aim of this investigation is to collaborate with young adults with HIV and create a space for their voices and input to adapt an existing mHealth intervention, PositiveLinks (PL), to improve care engagement in 18- to 29-year-old people living with HIV in South Texas [25]. This age range was selected as they experience higher rates of new diagnoses and a lower prevalence of sustained viral suppression [4]. PL is a clinic-based multimodal HIV care intervention that increases care engagement and viral suppression [25]. The platform includes daily medication adherence tracking, mood and stress

XSL•FO
RenderX

self-assessment, appointment reminders, secure messaging with clinic providers, and an anonymous online peer support network [25]. We used human-centered design thinking processes in this adaptation to ensure young adults with HIV input, as interventions are more effective when impacted communities are involved in their design and implementation [5].

We conducted a formative evaluation using data from semistructured interviews with young adults with HIV and stakeholders, as well as focus groups with stakeholders to understand acceptable and desired intervention characteristics [19]. Subsequently, we used design thinking processes to aid in our adaptation of PL [16]. This paper describes findings from each phase to enhance understanding of the components of mHealth interventions that are accepted and sought after by young adults with HIV.

Methods

Overview

Adaptation of the existing mHealth intervention, PL, into its novel iteration, Positive Links for Youth (PL4Y), was informed by three data sources: (1) components of the formative evaluation that addressed acceptable mHealth interventions, (2) a human-centered design thinking process, and (3) A Youth Advisory Board (YAB) of young adults with HIV that provided input throughout the formative evaluation and adaptation phases (Figure 1). There was no overlap in membership between the young adults with HIV participating in interviews and the YAB during the formative evaluation. A total of 3 young adults with HIV interviewed during the formative evaluation joined the YAB for subsequent phases of the project, the human-centered design thinking process, and the pilot intervention with the adapted PL4Y app, currently underway. The project used the Consolidated Framework for Implementation Research (CFIR), gathering knowledge of individuals involved and the inner and outer settings surrounding the intervention from a formative evaluation, and then iteratively adapting the intervention to fit the context [26]. A PYD framework also informed study activities, which incorporated input of young adults with HIV in the form of interviews, active feedback during the human-centered design thinking process, and the YAB [11,12]. All 3 components were overseen by the design working group, which included the principal investigator (PI), study coordinator, research assistants, and the scientists, clinic staff, and software designers on the original PL team.

XSL•FO
RenderX

Figure 1. Triangulation of data sources for adaptation of the mobile health (mHealth) intervention to the needs of young adults with HIV in South Texas. PL: PositiveLinks; YAB: Youth Advisory Board; FAQ: frequently asked questions.

XSL•FO
RenderX

Recruitment

Members of the YAB were recruited through meetings with community-based organizations, flyer distribution in HIV service organizations and clinics, and collaboration with advocates for young adults with HIV. For the formative evaluation phase and Think-Aloud interviews, young adults with HIV were recruited from 6 HIV treatment centers in South Texas using flyers and direct outreach to clinic staff and providers. A total of 14 participants were enrolled in the formative evaluation phase. Of the 3 participants in the Think-Aloud session, 2 had previously participated in the formative evaluation phase. Inclusion criteria required members to be 18- to 29-year-old and living with HIV in San Antonio, Texas. Stakeholders from 6 regional HIV treatment centers were also recruited for the formative evaluation phase.

Youth Advisory Board Role and Input

PL4Y began with the creation of a YAB consisting of 18 - to 29-year-old people living with HIV in South Texas [1,19]. As described previously by Chidester et al [1], the YAB's structure and format evolved in response to the COVID-19 pandemic and feedback from YAB members. YAB engagements occurred quarterly, with input collected through Qualtrics surveys to preserve YAB member anonymity.

Topics covered in the surveys included desired elements of an mHealth intervention, topics for the community message board, preferred mobile app security options (password, fingerprint, face ID, etc), and feedback on the design and look of the app. In the final stage of design iteration, responses regarding the graphics and the app interface from previous YAB surveys and the Think-Aloud sessions were reviewed by the design working group. The software developers incorporated this feedback into the design of a new look and logo for PL4Y. A Qualtrics survey was distributed to the YAB, asking respondents to rate and rank potential logos and designs (N=11). The top 3 logos and fonts (determined by average ranking and average rating) were implemented into a second Qualtrics survey (N=9). The highest average ranked combination of logo and font was selected as the official logo of PL4Y (Figure 2).

Figure 2. Overview of 4 human-centered design thinking driven adaptations to the existing intervention to respond to expressed needs of young adults with HIV during Think-Aloud sessions and the Youth Advisory Board (YAB). mHealth: mobile health; LGBTQIA+: lesbian, gay, bisexual, transgender, queer, intersex, asexual, and other minoritized gender and sexual identities.

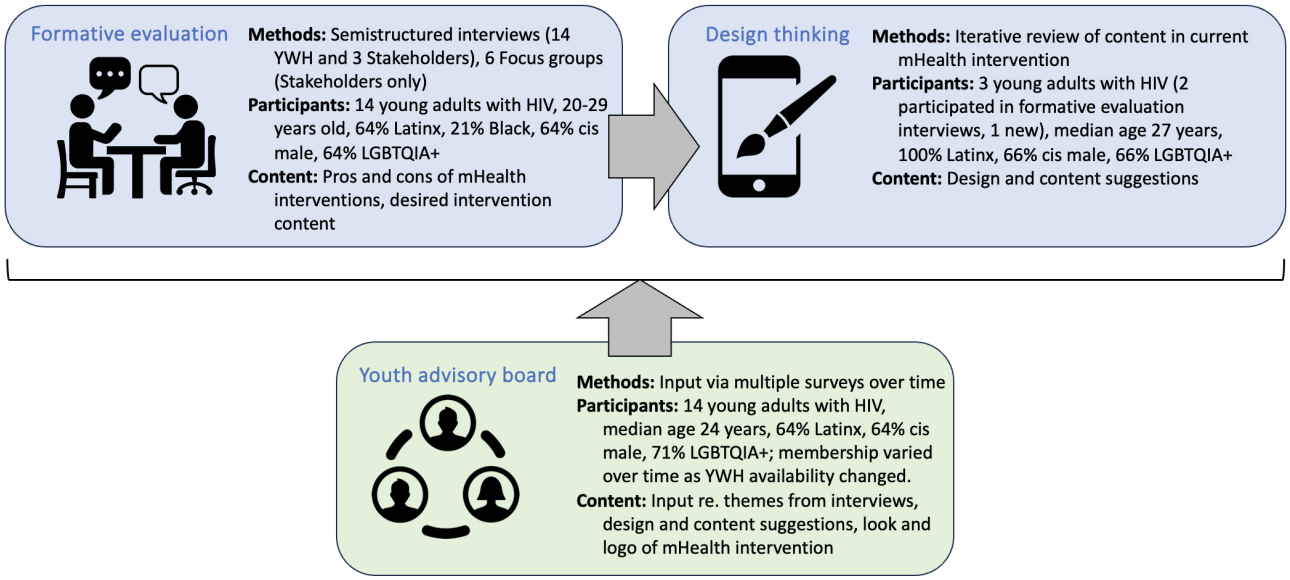

Formative Evaluation of Acceptable mHealth Intervention Components

The formative evaluation of care engagement for young adults with HIV in South Texas was conducted from October 2020 to December 2021. Eligibility criteria included being 18 - 29 years of age, residing in South Texas, fluency in either English or Spanish, and reporting current or past challenges with care engagement. Stakeholders included 26 office, nursing, and provider staff members from the regional HIV treatment centers, who provided input either through 6 focus groups or 3 individual interviews.

Interview and focus group guides were developed based on the theory of planned behavior (TPB) theoretical model, the PYD framework, and additional themes provided by the YAB. young adults with HIV were asked about first-hand experiences that affect their care engagement as young people residing in South Texas. Stakeholders were asked to describe their experiences and beliefs regarding serving young adults with HIV and to provide system-level insights on strategies for improving care engagement. To reduce bias and support generative reflection on desired components, neither young adults with HIV or stakeholders were provided with access to the original version of the PL app. A transcription service was used to transcribe individual interviews and focus groups. Transcripts were reviewed and edited by the interviewer (CJ) for accuracy. Transcripts were uploaded into NVivo 1.6.2 (Lumivero), which allows for tracking of responses by stakeholder or young adults with HIV characteristics and by clinic site.

The qualitative analysis team (QAT) created a master codebook of themes derived from TPB and PYD theoretical frameworks for thematic analysis [1,16,17]. The QAT included 4 research assistants with training in qualitative analysis (CJ, AC, HL, and

XSL•FO
RenderX

NMH), 5 faculty-level researchers with expertise in HIV care engagement (AT, AN, KI, RD, and BT), and 1 youth development expert (SK). A member of the primary coding team independently coded transcripts inductively (CJ, AC, and HL), with review by the QAT lead (BT). Secondary coders (CJ, AC, HL, and NMH) expanded on coding as needed. At this stage, the QAT also examined themes by participant type and noted differences between young adults with HIV and stakeholder perspectives. The QAT resolved differences in coding by consensus and confirmed revisions to codebooks. The full QAT participated in the final round of coding, where all themes were reviewed, representative quotes discussed, and themes were renamed, merged, or discarded. Overall findings from the formative evaluation regarding care engagement for young adults with HIV in this context were the topic of a separate analysis [19]. This analysis reviewed themes on advantages and disadvantages of mHealth interventions in general and themes regarding specific mHealth intervention components desired by young adults with HIV or stakeholders.

Human-Centered Design Adaptation of Positive Links by Young Adults with HIV

A total of 3 Think-Aloud interviews were conducted with young adults with HIV using human-centered design principles to ensure end-user input informed an iterative adaptation process [17,18]. A total of 2 Think-Aloud sessions were conducted in person at a local clinic, and 1 was conducted online through Zoom. In addition, 3 researchers (CJ, AC, and NMH) were present for interviews and a modified system usability scale guided conversation [27]. The Think-Aloud participants reviewed each aspect and function of a previous version of the app by walking through it on their phones in real time, providing suggestions for possible design changes and usability improvement. Adaptation recommendations were categorized by ease of implementation and frequency. The most common and feasible suggestions were conveyed to the University of Virginia app development team for final design and development.

Ethical Considerations

All young adults with HIV and stakeholders participating in the formative evaluation and human-centered design components reviewed an institutional review board (IRB)–approved information sheet about the study and provided verbal consent to participate. Participants in the focus groups and Think-Aloud sessions received a US $25 electronic gift card. Stakeholders did not receive monetary incentives for participation. YAB members also received a US $25 electronic gift card upon completion of each survey. This protocol was approved by the UT Health San Antonio and University Health IRBs (HSC20220752HU)

Results

Acceptability and Concerns Regarding mHealth Interventions From the Formative Evaluation

Qualitative data from young adults with HIV and stakeholders revealed common themes about benefits and challenges of an mHealth care engagement intervention. Overall, young adults with HIV and stakeholders felt ease of access to resources, labs, and clinic personnel would be beneficial, and that the anonymous online support group could help young adults with HIV feel less isolated. However, concerns were raised about confidentiality on an mHealth platform and potential pitfalls of the online support group, including the need for moderation and issues with web-based connections (Textbox 1).

Textbox 1. Summary of perceived benefits, challenges, and critiques of an mHealth care engagement intervention expressed by 18- to 29-year-olds living with HIV and stakeholders.

Benefits:

- Increased accessibility of care team.
- Increased access to accurate information about health issues and available resources.
- App-based medication and appointment reminders.
- Ability to track treatment progress through improvements in lab parameters.
- Improved access to a safe community through a online support group.
- Anonymity provided by online support group platform.

Challenges:

- Potential confidentiality concerns.
- Lack of streamlined features.
- Difficulty connecting to other members in the online setting.
- Need for moderation of online support group with potential for cyberbullying.
- Live chat could trigger negative or apprehensive feelings.

XSL•FO
RenderX

Perspectives on Benefits of the mHealth Intervention

Accessibility to Care and Knowledge

Youth and stakeholders believed the app would be highly effective in connecting patients with providers and case managers. Young adults with HIV noted that having an app on their smartphone could increase accessibility to their care team, making them more likely to participate in their care. They also emphasized the app could help them understand what programs they were eligible for and how to apply for support without visiting a clinical setting.

> *You'll be more likely to actually participate in that if it's, you know, easy. It's, like, virtual; so that's a big thing.* [Youth Voice YV 8]

> *If you have a phone or some type of smart device, you would be able to access it wherever you are.* [YV5]

Stakeholders further noted the online nature of the app provides a unique benefit, particularly in the COVID-19 era, by enabling patients to connect with their support network without meeting in person. They believed this benefit extended beyond the pandemic, as it also improves access for patients in rural settings.

> *We don't know how many people are at home alone.* [Stakeholder Voice SV 9]

Furthermore, young adults with HIV and stakeholders felt the app could centralize trustworthy information and serve as an assessable resource.

> *You may not want to ask the doctor, but just, like, you want a quick question answered.* [YV6]

Advantages of an Online Support Group

Youth recognized the benefits of an online support group, describing it as a safe community where people with similar life experiences could connect. They noted that such a group could reduce stigma by fostering a sense of camaraderie.

> *I think with a support group, it's, like, knowing that you're not alone and that, like, it's not a end-of-the-world thing; and you're not dying. You know, you have other people who are living with it and surviving; and it kind of decreases that stigma because you know that there are millions, you know, that are dealing with the same issue.* [YV6]

Both young adults with HIV and shareholders appreciated the anonymity provided by an app-based support group. Young adults with HIV emphasized the app's online platform offered privacy, as they would not need to visit a physical location to engage with a support group.

> *It's good because it's like I don't have to go and park there to go see somebody.* [YV13]

Stakeholders believed newly diagnosed patients would be able to ask about symptoms without feeling guilt. In addition, they highlighted that the anonymity of the app may enable more people to feel comfortable contributing to the community without fear of being judged.

> *Like if they want us to stay a little bit more anonymous but have some place where they can talk to some other people that are going through something similar, or to be like, hey, I'm having these weird symptoms, anybody else experienced that or whatever.* [SV25]

Reminders

Young adults with HIV noted app-based medication reminders would have been helpful when they were first diagnosed and had not already established a medication routine.

> *I think it would be good for other people that are barely starting, so they can remember to drink their meds.* [YV3]

Young adults with HIV and stakeholders reported they would welcome medication refill reminders and appointment reminders.

> *Oh, that would definitely be helpful—medication reminders for sure because I will go to my very last, like, tablet; and I will forget, and then it's a hassle.* [YV6]

Stakeholders identified that reminders could provide encouragement by tracking treatment progress and enabling users to share their milestones within a community. This, in turn, would empower young adults with HIV to continue engaging in care.

> *Hey, your viral load was this high and look how it is now. I think it would be helpful, you know, when they become undetectable or they can, you know, share with partners.* [SV18]

Young adults with HIV reported they wished to have "testimonials/videos of people living with HIV," and "facilitated conversations" on the app.

Disadvantages of an mHealth Intervention

Confidentiality and Anonymity

Although the app's presence on a smartphone increases accessibility, young adults with HIV and stakeholders expressed concerns about potential for unintentional disclosure if someone else sees the app on their device. Both groups identified notifications as a particular risk for exposure.

> *Somebody else gets access to your account, and I mean, that's biggest my fear… Maybe getting notifications when I didn't want to get a notification and it kind of exposing me.* [YV7]

Design Critiques

Some young adults with HIV and stakeholders mentioned having too many settings in the app may limit its usability. They wanted a streamlined design with access to educational content. In addition, youth expressed concerns about potential redundancy in notification, where the same message might appear in multiple locations on the smartphone.

> *It's just redundant when I'm using the same device over and over again.* [YV6]

XSL•FO
RenderX

Young adults living with HIV and stakeholders voiced concerns that online interactions could feel impersonal. A participant reported negative experiences with previous web-based communities of people living with HIV.

> *But I just never really, I never really identified—or like connected with any of them. They were … people from (other countries). But no Hispanic people.* [YV11]

Other young adults living with HIV reported that online resources did not feel genuine and preferred in-person interactions.

> *Going to an actual in-person group, having the actual in-person interaction with people for me would have more of an impact than just kind of talking to someone on my phone.* [YV5]

Despite the potential impersonal experience, stakeholders pointed out that people are more accustomed to virtual settings than in the past.

> *But a lot of things just nowadays is really virtual, just like what we're doing here today. You know, a lot of it is just really impersonal.* [SV3]

Concerns for Unmoderated Interactions

Young adults living with HIV were concerned that without moderation, there was the possibility of smaller communities developing within mHealth interventions and online communities, making the platform less approachable for new users. In addition, concerns were voiced about the lack of moderation allowing users to write irrelevant or offensive posts within the forum. A participant described their experience reading comments on a thread for a online discussion group:

> *Every time you scroll down and you think you're done, there's like 20 more in the bottom and they're just saying random stuff.* [YV9]

Young adults living with HIV and stakeholders had concerns regarding the potential for "trolls" or "cyberbullying." They were particularly concerned about those who would be able to download and join the app community following implementation.

> *Maybe if a certain type of group of people get ahold of it… They have easy access to basically come on and troll everybody.* [YV14]

Stakeholders additionally expressed concerns about delayed responses to urgent messages written in the live chats, given staff are not constantly monitoring the service. However, they believe this risk could be mitigated by educating patients that urgent messages should not be conveyed through the app and that messages will be reviewed within a designated time.

> *Patients may just send the chat like saying… I have fever and shortness of breath. And then Friday at 5:30 and nobody's there in the chat feature. So, I think that that's a little bit dangerous.* [SV18]

Young adults with HIV voiced apprehension about how video support groups or live chats would affect them. A user expressed they might experience a, "feeling of insignificance, just because maybe somebody is going through it worse than you are" (YV2). In addition, young adults living with HIV stated concerns about how listening to other people's testimonies may, "make it more real for them" (YV5).

and it is unpleasant if people minimize their experience on the app. They also speculated that users interested in building connections may migrate to other platforms, such as social media, for deeper conversations.

> *But I don't think that they're going to really utilize just this App to message back and forth. Because they'll be like, "Hey, you're cool." "Oh, you're cool, too. Here's my Instagram." And then they're going to use that.* [YV13]

Design Suggestions From Young Adults With HIV Derived From Think-Aloud Sessions

Young adults with HIV wanted to know who would have access to the information they put into the app, such as medication adherence, mood, or stress. App users can report daily medication adherence, stress levels, and mood "check ins" to log their emotions. Think-Aloud participants emphasized the importance of ensuring that no one views their data without consent.

Participants provided specific suggestions to improve the app's aesthetic appeal, requesting a more "modern look." They had strong opinions against the original color scheme of the app and the old logo, which featured two chain links. The original blue color scheme was deemed "too medical" by participants, who suggested either a "plain" or "vibrant" color scheme. Participants were quick to point out incorrect scaling on thumbnails and other visually unappealing color schemes that reduced the visual charm of the original PL app.

During Think-Aloud sessions, participants reported certain design aspects were nonintuitive, hindering the ease of use. Participants suggested reorganizing the community message board from a single thread to allow participants to self-organize threads, optimizing the search bar for resources and FAQs, having convenient "learn more" hyperlinks within lab results, and implementing "read receipts" to confirm whether messages and documents sent were seen by all users, including health care providers.

Participants expressed particular interest in finding local resources relevant to South Texas. They were eager to learn about the resources offered within the app but suggested making them more community specific. Participants also found the FAQ section useful for patients newly diagnosed with HIV and suggested including a glossary of acronyms to help understand frequently used terminology in HIV care. Participants also emphasized the need for customizability and improved integration within their devices. Specifically, they requested ways to customize the app's home screen icon and to modify notification settings to allow for privacy.

Several suggestions made by Think-Aloud participants were determined to be beyond the scope of the current adaptation, but important to consider in the future. These potential adaptations include saving posts, blocking specific users, and

XSL•FO
RenderX

prioritizing conversations within community message boards. Young adults living with HIV also requested the ability to add photos to posts, replies, and messages. Participants additionally suggested new features for access to care, such as online bus passes.

Adaptation Process

The design working group incorporated findings from the formative evaluation and Think-Aloud sessions into the app's redesign. The laboratory section was updated to be more understandable for young adults with HIV. The community message board was divided into 3 distinct channels, allowing members to "like" other posts, and implementing a search by word or hashtag function. The color scheme, font, and logo were updated based on feedback by the YAB, as described above (Figure 2). The "Resources and FAQ" section was condensed and revised to prioritize material relevant to South Texas. This adaptation process is ongoing as we continue to add requested content, such as videos explaining the steps to an HIV clinic visit.

Discussion

Principal Findings

We incorporated the voices of young adults with HIV throughout the process of adapting an mHealth care engagement intervention using an implementation science–informed approach. This process led to significant changes in the intervention. Young adults with HIV identified several advantages of mHealth interventions, including ease of access, community building, and medications and appointment reminders. However, they also voiced concerns about confidentiality, the impersonal nature of online interactions, and lack of content moderation. Based upon feedback from the formative evaluation, Think-Aloud sessions, and the YAB, the intervention was adapted to meet expressed needs. Modifications included changing the graphic design and logo, adding South Texas specific resources, updating visualization of laboratory data, and designing a more user-friendly and youth-appropriate community message board.

The medical literature suggests young adults with and those without HIV accept app-based mHealth interventions, which can be used to improve health behaviors for youth and emerging adults [9,10,13,28]. An app-based intervention was found to be successful in an adolescent population with asthma, as well as in an adolescent population in recovery following hematopoietic stem cell transplant [29,30]. There have been few studies investigating the usability of app-based interventions in patients with HIV. Aladin et al [9] enrolled 113 patients with HIV and found their app-based intervention to be associated with increased communication and engagement with care when combined with peer navigation and a comic series. However, the majority of mHealth interventions are text message–based between patient and clinic, and users lack the ability to form connections with each other [31]. Brooks et al [13] found that app-based interventions were most useful for young adults with HIV if they had medical information and reminders, leading to increased self-efficacy, supported feelings of connection to a community, and reduced stigma around HIV. These results indicate users of app-based interventions may benefit from "discussion board" components that allow users to communicate with one another and develop a community. Schnall et al [32] studied the functions of 15 publicly available app-based interventions targeting people living with HIV and found 4 enabled communication between peers and providers, 6 had medication reminders, 7 had medication logs, and 6 included a search function [32]. This analysis aligns with the existing literature as both users and stakeholders identified all of the above as critical functionalities of an app-based intervention, and these components were included in the final redesign.

Our findings add to this existing knowledge base by highlighting the importance young adults with HIV place on easy, asynchronous access to services. Young adults with HIV welcome the ability to message providers or other members of the care team, quick links to web-based resources for HIV care and social services, and automatic appointment and medication reminders. We discovered many young adults with HIV wished for connections to other young adults with HIV, something not currently featured in mHealth interventions for this demographic. They also appreciated the online, anonymous format of the community message board. Finally, young adults with HIV repeatedly emphasized that the graphic design and usability of the intervention should be influenced by what was most acceptable and appealing to them. These data influenced our own design adaptation of PL into PL4Y but are also relevant to framing other mHealth interventions for young adults with HIV or other youth communities.

Few studies investigate patient thoughts regarding current mHealth interventions or discuss the concerns of young adults with HIV with this intervention modality. A review of 7 mHealth interventions by Conserve et al [33] found cellular network restrictions were a significant barrier to the use of mobile technologies for young adults with HIV. Our findings echo this, as both young adults with HIV and the stakeholders anticipated access to a smartphone could be a barrier to use of our intervention.

We found additional young adults with HIV concerns with mHealth intervention that, to our knowledge, are not reflected in the existing literature. Many young adults with HIV expressed concerns about confidentiality and security of the data on their phone. Others worried that communicating by electronic messaging, rather than in person or by phone call, would feel impersonal. Young adults living with HIV were also concerned with exposure to harmful or disempowering narratives on the community message board. Though our intervention already has robust security and content monitoring protocols, these concerns will need to be addressed during implementation.

Limitations

Despite the meaningful findings of this investigation, there are limitations. Our adaptation process incorporated viewpoints of a limited number of young adults with HIV participants and YAB members. While their perspectives were valuable, we cannot assume they represent the perspectives of all young adults with HIV. Furthermore, our participants reflect the South Texas population, with the majority identifying as Latinx, which may limit the generalizability of our findings. Despite these

XSL•FO
RenderX

limitations, this study fills a crucial gap in the literature by identifying crucial barriers and facilitators to the adoption of an app-based mHealth intervention for young adults with HIV. It also increases our understanding of the understudied South Texas community's potential response to other mHealth interventions [34].

Conclusions

mHealth interventions targeting young adults with HIV can improve care engagement and health outcomes. However, few studies have identified the specific characteristics that promote or hinder patient adoption of app-based interventions [9,10,13,28]. Our adaptation process successfully incorporated input from a majority Hispanic or Latinx group of young adults with HIV in the design of a mHealth intervention. Our process demonstrated that young adults with HIV should play integral roles in the design of interventions for this population. We identified novel concerns not previously reflected in the literature including both barriers and facilitators to app-based mHealth interventions for young adults with HIV. We also describe a rigorous and iterative adaptation process that centers youth voice to ensure that the new intervention meets the needs of the community, successfully transforming PL into PL4Y. Currently, our team is testing the feasibility and acceptability of the adapted PL4Y mHealth intervention in a randomized control trial among young adults with HIV in South Texas. Our findings can also be used in the design of future app-based mHealth interventions to promote their adoption by young adults with HIV.

Acknowledgments

We are deeply grateful to all the HIV treatment centers and youth-focused community-based organizations in South Texas who partnered with us in this effort, and to all study participants and Youth Advisory Board members for their time and commitment to this process. The funding source for this project was a grant from the National Institute of Mental Health (NIH/NIMH R34 MH122332). The sponsor played no role in the development, review, or approval of this manuscript for publication.

Data Availability

The datasets generated and analyzed during this study are not publicly available due to the fact that individual participant stories of their experiences with HIV could be identifiable in the transcripts, but are available from the corresponding author on reasonable request.

Conflicts of Interest

RD and KI provide consulting services for Warm Health Technology, LLC, which distributes the PositiveLinks app outside Virginia. None declared by the other authors.

References

1. Chidester AB, Johnson CJ, Lin H, et al. Nothing about us without us: involving youth living with HIV in a virtual advisory board. J Adolesc Health 2023 Dec;73(6):1158-1161. [doi: 10.1016/j.jadohealth.2023.06.028] [Medline: 37665305]
2. Ryscavage PA, Anderson EJ, Sutton SH, Reddy S, Taiwo B. Clinical outcomes of adolescents and young adults in adult HIV care. J Acquir Immune Defic Syndr 2011;58(2):193-197. [doi: 10.1097/QAI.0B013E31822D7564]
3. Kapogiannis BG, Koenig LJ, Xu J, et al. The HIV continuum of care for adolescents and young adults attending 13 urban US HIV care centers of the NICHD-ATN-CDC-HRSA SMILE collaborative. J Acquir Immune Defic Syndr 2020 May 1;84(1):92-100. [doi: 10.1097/QAI.0000000000002308] [Medline: 32267659]
4. Behavioral and clinical characteristics of persons with diagnosed HIV infection-medical monitoring project, HIV surveillance special report united states. : Centers for Disease Control and Prevention; 2020.
5. Hightow-Weidman LB, Muessig KE, Bauermeister J, Zhang C, LeGrand S. Youth, technology, and HIV: recent advances and future directions. Curr HIV/AIDS Rep 2015 Dec;12(4):500-515. [doi: 10.1007/s11904-015-0280-x] [Medline: 26385582]
6. Anderson M. Teens, social media & technology 2018. Pew Research Center. 2018. URL: https://www.pewresearch.org/internet/2018/05/31/teens-social-media-technology-2018/ [accessed 2023-10-05]
7. Cooper V, Clatworthy J, Whetham J, Consortium E. mHealth interventions to support self-management In HIV: a systematic review. Open AIDS J 2017;11:119-132. [doi: 10.2174/1874613601711010119] [Medline: 29290888]
8. Hightow-Weidman LB, Muessig K, Rosenberg E, et al. University of North Carolina/Emory Center for Innovative Technology (iTech) for addressing the HIV epidemic among adolescents and young adults in the United States: protocol and rationale for center development. JMIR Res Protoc 2018 Aug 3;7(8):e10365. [doi: 10.2196/10365] [Medline: 30076126]
9. Aladin B, Thompson M, Addison D, et al. The YGetIt? program: a mobile application, PEEP, and digital comic intervention to improve HIV care outcomes for young adults. Health Promot Pract 2023 Jul;24(4):658-668. [doi: 10.1177/15248399221150789] [Medline: 36924286]
10. Paul ME, Castillo M, Emmanuel P, et al. Scale up mHealth HIV interventions: site and public health perspectives and lessons learned from P3. Mhealth 2021;7:38. [doi: 10.21037/mhealth-20-65] [Medline: 33898607]

XSL•FO
RenderX

11. Benson PL, Scales PC, Sesma Jr A. Positive youth development: theory, research, and applications. In: Handbook of Child Psychology: John Wiley & Sons; 2007. [doi: 10.1002/9780470147658]
12. Tamis-LeMonda LB, Catherine S, editors. Toward a New Vision and Vocabulary about Adolescence: Theoretical, Empirical, and Applied Bases of a “Positive Youth Development” Perspective: Child Psychol 2nd ed Psychology Press; 2005.
13. Brooks RA, Nieto O, Swendeman D, et al. Qualitative evaluation of social media and mobile technology interventions designed to improve HIV health outcomes for youth and young adults living with HIV: a HRSA SPNS initiative. Health Promot Pract 2020 Sep;21(5):693-704. [doi: 10.1177/1524839920938704] [Medline: 32757839]
14. Amico KR, Lindsey JC, Hudgens M, et al. Randomized Controlled Trial of a Remote Coaching mHealth Adherence Intervention in Youth Living with HIV: Springer; 2022, Vol. 26:3897-3913. [doi: 10.1007/s10461-022-03717-2]
15. Goldstein M, Archary M, Adong J, et al. Systematic review of mhealth interventions for adolescent and young adult HIV prevention and the adolescent HIV continuum of care in low to middle income countries. AIDS Behav 2023 May;27(Suppl 1):94-115. [doi: 10.1007/s10461-022-03840-0] [Medline: 36322217]
16. Harte R, Glynn L, Rodríguez-Molinero A, et al. A human-centered design methodology to enhance the usability, human factors, and user experience of connected health systems: A three-phase methodology. JMIR Hum Factors 2017 Mar 16;4(1):e8. [doi: 10.2196/humanfactors.5443] [Medline: 28302594]
17. Liedtka J, King A, Bennett K. Solving Problems with Design Thinking: Ten Stories of What Works: Columbia University Press; 2011.
18. Liedtka J, Ogilvie T, Brozenske R. Designing for Growth: A Design Thinking Tool Kit for Managers: Columbia University Press; 2011.
19. Johnson C, Chidester A, Chandramohan D, et al. A call for youth voice to support engagement in care for 18- to 29-year olds living with HIV in the US South. AIDS Patient Care STDS 2024 May;38(5):238-248. [doi: 10.1089/apc.2024.0006] [Medline: 38662471]
20. Monterosso A, Minnerly S, Goings S, et al. Identifying and investigating a rapidly growing HIV transmission cluster in Texas. 2017 Presented at: Conference on Retroviruses and Opportunistic Infections; Feb 19-22, 2023; Seattle, Washington.
21. Oster AM, France AM, Panneer N, et al. Identifying clusters of recent and rapid HIV transmission through analysis of molecular surveillance data. J Acquir Immune Defic Syndr 2018 Dec 15;79(5):543-550. [doi: 10.1097/QAI.0000000000001856] [Medline: 30222659]
22. Allan-Blitz LT, Mena LA, Mayer KH. The ongoing HIV epidemic in American youth: challenges and opportunities. Mhealth 2021;7:33. [doi: 10.21037/mhealth-20-42] [Medline: 33898602]
23. Taylor B, Sapien H, Courtois F, et al. San antonio’s end stigma end HIV alliance/fast-track cities initiative: community-wide collaboration and peer advocate driven strategies to end the HIV epidemic. : Fast Track Cities Conference; 2019 Presented at: .
24. Taylor BS, Sapién H. Determining the direction of human immunodeficiency virus transmission: benefits and potential harms of taking phylogenetic analysis one step further. Clin Infect Dis 2021 Jan 23;72(1):38-40. [doi: 10.1093/cid/ciz1248] [Medline: 31922539]
25. Dillingham R, Ingersoll K, Flickinger TE, et al. PositiveLinks: A mobile health intervention for retention in HIV care and clinical outcomes with 12-month follow-up. AIDS Patient Care STDS 2018 Jun;32(6):241-250. [doi: 10.1089/apc.2017.0303] [Medline: 29851504]
26. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. Implementation Sci 2009 Dec;4(1):50. [doi: 10.1186/1748-5908-4-50]
27. Bangor A, Kortum PT, Miller JT. An empirical evaluation of the system usability scale. Int J Hum Comput Interact 2008 Jul 29;24(6):574-594. [doi: 10.1080/10447310802205776]
28. Biello KB, Marrow E, Mimiaga MJ, Sullivan P, Hightow-Weidman L, Mayer KH. A mobile-based app (MyChoices) to increase uptake of HIV testing and pre-exposure prophylaxis by young men who have sex with men: protocol for a pilot randomized controlled trial. JMIR Res Protoc 2019 Jan 7;8(1):e10694. [doi: 10.2196/10694] [Medline: 30617042]
29. Rodgers CC, Krance R, Street RL, Hockenberry MJ. Symptom prevalence and physiologic biomarkers among adolescents using a mobile phone intervention following hematopoietic stem cell transplantation. Oncol Nurs Forum 2014 May;41(3):229-236. [doi: 10.1188/14.ONF.229-236] [Medline: 24769589]
30. Burbank AJ, Lewis SD, Hewes M, et al. Mobile-based asthma action plans for adolescents. J Asthma 2015;52(6):583-586. [doi: 10.3109/02770903.2014.995307] [Medline: 25494553]
31. Fedele DA, Cushing CC, Fritz A, Amaro CM, Ortega A. Mobile health interventions for improving health outcomes in youth: a meta-analysis. JAMA Pediatr 2017 May 1;171(5):461-469. [doi: 10.1001/jamapediatrics.2017.0042] [Medline: 28319239]
32. Schnall R, Mosley JP, Iribarren SJ, Bakken S, Carballo-Diéguez A, Brown Iii W. Comparison of a user-centered design, self-management app to existing mHealth apps for persons living with HIV. JMIR Mhealth Uhealth 2015 Sep 18;3(3):e91. [doi: 10.2196/mhealth.4882] [Medline: 26385783]

XSL•FO
RenderX

33. Conserve DF, Jennings L, Aguiar C, Shin G, Handler L, Maman S. Systematic review of mobile health behavioural interventions to improve uptake of HIV testing for vulnerable and key populations. J Telemed Telecare 2017 Feb;23(2):347-359. [doi: 10.1177/1357633X16639186] [Medline: 27056905]
34. Ramirez AG, Thompson IM, Vela L. South texas population characteristics. In: Ramirez AG, Thompson IM, Vela L, editors. South Texas Health Status Review: A Health Disparities Roadmap: Springer; 2013:7-9. [doi: 10.1007/978-3-319-00233-0_2]

Abbreviations

CDC: Centers for Disease Control and Prevention
CFIR: Consolidated Framework for Implementation Research
IRB: institutional review board
mHealth: mobile health
PI: principal investigator
PL: PositiveLinks
PL4Y: Positive Links for Youth
PYD: Positive Youth Development
QAT: qualitative analysis team
SV: Stakeholder Voice
TPB: theory of planned behavior
YAB: Youth Advisory Board
YV: Youth Voice

Edited by A Mavragani; submitted 14.05.24; peer-reviewed by D Cain, EDWARD NSUBUGA; revised version received 04.02.25; accepted 05.02.25; published 08.04.25.

Please cite as:
Ho NM, Johnson C, Chidester A, Viera Corral R, Ramos J, Garcia M, Gonuguntla R, Cote C, Chandramohan D, Lin H, Taranova A, Nijhawan AE, Kools S, Ingersoll K, Dillingham R, Taylor BS
Centering Youth Voice in the Adaptation of an mHealth Intervention for Young Adults With HIV in South Texas, United States: Human-Centered Design Approach
JMIR Form Res 2025;9:e60531
URL: https://formative.jmir.org/2025/1/e60531
doi:10.2196/60531

© Nhat Minh Ho, Catherine Johnson, Autumn Chidester, Ruby Viera Corral, Jacundo Ramos, Miguel Garcia, Rishi Gonuguntla, Cyrena Cote, Divya Chandramohan, Hueylie Lin, Anna Taranova, Ank E Nijhawan, Susan Kools, Karen Ingersoll, Rebecca Dillingham, Barbara S Taylor. Originally published in JMIR Formative Research (https://formative.jmir.org), 8.4.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Chatbot to Support the Mental Health Needs of Pregnant and Postpartum Women (Moment for Parents): Design and Pilot Study

Kelsey McAlister[1], PhD; Lara Baez[1], PhD; Jennifer Huberty[1], PhD; Marianna Kerppola[2], MS, MBA

[1]Fit Minded Inc, 14848 N 46th Place, Phoenix, AZ, United States

[2]Poisera Inc (dba Moment for Parents), Ann Arbor, MI, United States

Corresponding Author:
Kelsey McAlister, PhD
Fit Minded Inc, 14848 N 46th Place, Phoenix, AZ, United States

Abstract

Background: Maternal mental health disorders are prevalent, yet many individuals do not receive adequate support due to stigma, financial constraints, and limited access to care. Digital interventions, particularly chatbots, have the potential to provide scalable, low-cost support, but few are tailored specifically to the needs of perinatal individuals.

Objective: This study aimed to (1) design and develop Moment for Parents, a tailored chatbot for perinatal mental health education and support, and (2) assess usability through engagement, usage patterns, and user experience.

Methods: This study used a human-centered design to develop Moment for Parents, a rules-based chatbot to support pregnant and postpartum individuals. In phase 1, ethnographic interviews (n=43) explored user needs to inform chatbot development. In phase 2, a total of 108 pregnant and postpartum individuals were recruited to participate in a pilot test and had unrestricted access to the chatbot. Engagement was tracked over 8 months to assess usage patterns and re-engagement rates. After 1 month, participants completed a usability, relevance, and satisfaction survey, providing key insights for refining the chatbot.

Results: Key themes that came from the ethnographic interviews in phase 1 included the need for trusted resources, emotional support, and better mental health guidance. These insights informed chatbot content, including mood-based exercises and coping strategies. Re-engagement was high (69/108, 63.9%), meaning users who had stopped interacting for at least 1 week returned to the chatbot at least once. A large proportion (28/69, 40.6%) re-engaged 3 or more times. Overall, 28/30 (93.3%) found the chatbot relevant for them, though some noted repetitive content and limited response options.

Conclusions: The Moment for Parents chatbot successfully engaged pregnant and postpartum individuals with higher-than-typical retention and re-engagement patterns. The findings underscore the importance of flexible, mood-based digital support tailored to perinatal needs. Future research should examine how intermittent chatbot use influences mental health outcomes and refine content delivery to enhance long-term engagement and effectiveness.

(JMIR Form Res 2025;9:e72469) doi:10.2196/72469

KEYWORDS

perinatal support; human-centered design; digital health; maternal health; chatbot; digital tool

Introduction

Maternal mental health disorders are a critical public health concern due to their high and rising prevalence. The World Health Organization (WHO) estimates that 10% of pregnant women and 13% of postpartum women develop a mental health disorder [1]. In the United States, 1 in 5 women experience a mental health disorder during pregnancy or within the first year after birth [2]. While postpartum depression is the most prevalent, affecting approximately 1 in 8 women [3], other maternal mental health disorders affect women, including anxiety, psychosis, obsessive-compulsive disorder, bipolar disorder, posttraumatic stress disorder, and schizophrenia [4]. These challenges are even more pronounced among underrepresented populations, who experience higher rates of maternal mood disorders [5,6]. Although treatable, 75% of women with maternal mental health disorders go untreated [2], which poses serious health concerns for both the mother and the infant. Untreated maternal mental health conditions are linked to increased risks for preterm birth, low birth weight, developmental delays, and impaired maternal-infant bonding, and also raise the mother's risk of substance use, psychosis, and suicide [4,7]. The lack of ongoing education and support of mental health issues during and after pregnancy often leaves these women without timely support or interventions, exacerbating the problem.

There are several barriers to care contributing to untreated maternal mental health disorders, including stigma, lack of access to mental health care professionals, and high-cost challenges [8-11]. These challenges are especially prevalent in

XSL•FO
RenderX

underrepresented populations, such as Black women, who often face additional obstacles such as systemic racism, cultural insensitivity, and mistrust of health care providers [12-14]. Other barriers include reluctance to acknowledge symptoms and a lack of support from partners, family members, and health care professionals [11,15]. Social pressures and stigma often prevent women from discussing their feelings openly, leading to feelings of shame and fear of losing custody of their child [15]. Additionally, many mothers have limited knowledge of maternal mental health problems, hindering their ability to recognize symptoms and seek timely help [10,11]. In some cases, individuals may turn to online sources or personal networks for support [16,17], but these resources can be incomplete or inaccurate. Without adequate support, these barriers perpetuate high rates of untreated maternal mental health disorders, negatively affecting both mother and child health outcomes.

Digital tools are a promising solution to addressing barriers to maternal mental health care. These tools offer several advantages, including scalability, affordability, and accessibility, and provide real-time support without the stigma associated with in-person care. A wealth of digital tools for perinatal mental health have been tested in clinical trials, and a recent meta-analysis found that these tools are modestly effective in reducing symptoms of anxiety and depression [18]. However, challenges have emerged in engaging and retaining users of many digital tools due to the excess burden they place on users (eg, needing to download and learn to use a new mobile app). Chatbots have emerged recently as a way to engage perinatal users with digital interventions using a mode of communication that is ubiquitous and comfortable (eg, texting) [19]. Furthermore, chatbots simulate conversations with users, which may help them stay engaged [20]. Chatbots are promising because they offer cost-effective, real-time engagement [21,22]. Chatbots monitor mental health symptoms longitudinally, with each conversation functioning as an ecological momentary assessment, capturing real-time insights into emotional well-being and enabling timely interventions [23]. Importantly, chatbots promote honest disclosure for emotional well-being [24,25].

The success of these tools depends on adopting a human-centered approach, which ensures that the tool aligns with the specific needs and preferences of its users. A human-centered design (HCD) approach involves feedback from users during development, ensuring that interactions are empathetic, personalized, and accessible [26]. HCD integrates human factors, ergonomics, and usability techniques to enhance product design [26,27]. Additionally, it often involves ethnographic research and strategies that explore how individuals interact with systems, ensuring products are safer, easier to use, and aligned with the intended users' needs and abilities [26,27]. This approach helps create engaging experiences, fostering trust and consistent use, which are critical for maintaining long-term effectiveness in mental health support [26].

A growing body of research has demonstrated the utility of engaging in HCD products when developing tools for perinatal populations. Evidence shows that perinatal women widely use and value health-related apps, particularly for support, education, and self-monitoring. For example, many women find apps helpful for personalized health tracking, connecting with online communities, and addressing mental health concerns [28,29]. Furthermore, studies indicate that apps are most effective when designed with input from users and health care professionals, addressing gaps such as information reliability, security, and clinical validation [30,31]. HCD approaches have shown high feasibility and satisfaction, emphasizing the importance of tailoring interventions to the specific challenges and preferences of this population [32].

Despite the potential of digital tools, there remains a critical gap in solutions that offer both education and meaningful mental health support for pregnant and postpartum individuals. Many chatbots use evidence-based techniques such as cognitive behavioral therapy but are not designed for the unique needs of perinatal populations, often failing to address fluctuating emotions, stigma, and cultural sensitivities [33,34]. Additionally, these tools frequently lack longitudinal support and proactive interventions for perinatal mood and anxiety disorders [35]. Therefore, the purpose of this study was to develop and evaluate a chatbot called Moment for Parents. The goal of Moment for Parents is to address the current need for a digital mental health intervention to support the needs of pregnant and postpartum people. This study aimed to (1) design and develop a chatbot tailored specifically to educate and support pregnant and postpartum people on emotional well-being and mental health and (2) evaluate the usability of Moment for Parents in a pilot test by examining usage patterns, engagement levels, and user experience.

Methods

Design and Development

This study used a HCD approach to develop the Moment for Parents chatbot. HCD is an interactive system development approach that aims to create useful products by focusing on users, their needs, and their requirements [26,36]. HCD leads to inclusive health technology that improves mental health equity because the system engages users to improve adherence, adoption, and effectiveness [26,27,37]. By applying these principles, we aimed to develop a more inclusive and supportive chatbot for the mental health and well-being of pregnant and postpartum end users.

Figure 1 depicts the Moment for Parents study design. In phase 1, we engaged in ethnographic interviews with pregnant and postpartum end users to understand their needs and preferences. These findings informed the initial prototype of the Moment for Parents chatbot. In phase 2, the initial prototype was launched for pilot-testing to evaluate its usability, functionality, and relevance to users' needs. Throughout phase 2, the Moment for Parents chatbot experience was iteratively updated as participants provided feedback.

XSL•FO
RenderX

Figure 1. Overview of this study's design.

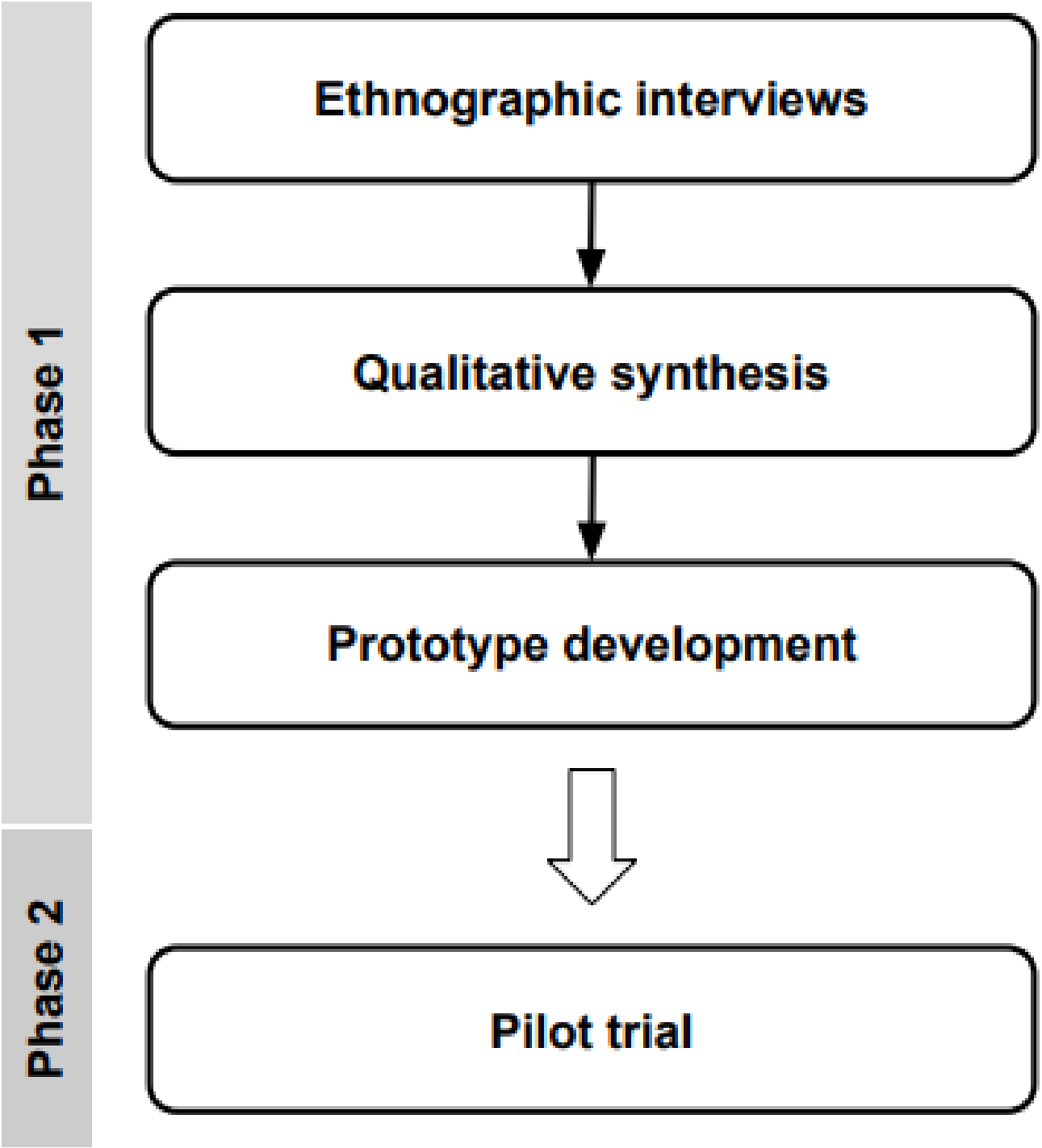

Ethical Considerations

This study is a retrospective analysis of secondary data, and therefore, it was exempt from additional consent requirements under human subjects regulations and received approval from the BRANY (Biomedical Research Alliance of New York) Institutional Review Board (study ID 24-338-1949). All study data were deidentified. Participants did not receive compensation for participation.

Phase 1 – Ethnographic Interviews

Participants

A total of 43 participants were recruited via snowball sampling, starting with existing networks and communities and expanding through participant referrals. Participants were eligible to participate in this study if they self-reported being pregnant or up to 1 year post partum.

Procedure

Data collection took place between March and November 2019. Ethnographic interviews were conducted via the videoconferencing service Zoom (Zoom Communications, Inc).

XSL•FO
RenderX

The same interviewer conducted all interviews. Interviews were recorded with participant permission, and the interviewer took reflexive notes throughout the interviews.

Measures

Of ethnographic interview, participants were asked to provide their gestational age if pregnant or the age of their youngest child if post partum. Participants were also asked how many additional children they had and their ages. No other demographic information was collected from participants. The ethnographic interviews were designed to explore participants' mental health needs, challenges during pregnancy or post partum, and expectations for digital support tools. Questions addressed a range of topics, including emotional well-being, resources they have found helpful, and desired features for a mental health chatbot. Sample questions included:

> *How has your pregnancy (if pregnant) been going so far? What have been your go-to resources to stay sane during pregnancy? What tools and/or resources were the most helpful?? After you had your baby (if postpartum), what tools or resources were the most helpful postpartum? Is there anything else you wish you had to help prepare for birth/motherhood? As you reflect on your pregnancy and postpartum experience, how did you think about self-care? What were the biggest hurdles/challenges in taking care of yourself?*

In addition to ethnographic interviews, additional input was collected from a multidisciplinary team of experts, including yoga and meditation instructors, midwives, doulas, public health professionals, and therapists specializing in maternal mental health. Insights for the Moment for Parents chatbot were also drawn from relevant maternal mental health literature, literature reviews on chatbots, and the team's lived experience.

Data Analysis

Audio recordings of interviews were transcribed, and reflexive interview notes were reviewed. Data were analyzed using a qualitative description approach [38], wherein the coder identified themes and subthemes relevant to the Moment for Parents prototype development.

Phase 2 – Pilot Test

Participants

A new sample of participants was recruited for the pilot test of the Moment for Parents prototype. Participants were recruited to the pilot test through personal networks and social media groups targeting pregnant and postpartum people who were interested in accessing mental health support. By recruiting these networks and online groups, we sought to reach a diverse range of users to ensure the prototype would be tested among people with different experiences in pregnancy and postpartum care.

Procedure

Data collection took place between November 2019 and June 2020. Upon enrollment, participants were provided with a no-cost, indefinite version of the Moment for Parents prototype, allowing participants to freely engage with the chatbot's features without time constraints. After 1 month of using Moment for Parents, participants were invited to complete a survey that asked about their perceptions of the Web app (ie, prototype feedback survey).

Measures

Chatbot Engagement

Participants' interactions with the Moment for Parents chatbot were tracked to assess usage patterns and calculate engagement metrics.

Prototype Feedback Survey

After 1 month of using Moment for Parents, participants completed a structured survey evaluating their experience with the chatbot. Participants received a text message inviting them to complete the survey, which included a link directing them to the questions in the Web app. Survey items assessed usability, perceived utility, relevance of content, and overall satisfaction. These responses allowed us to quantify participant satisfaction and identify trends in user experience.

Data Analysis

Descriptive statistics were used to summarize participant characteristics, engagement patterns, and feedback survey responses. Means and SDs were calculated for continuous variables, while categorical variables were reported as frequencies and percentages. Engagement over time was analyzed by calculating the proportion of participants who remained active each month. Re-engagement patterns were assessed by identifying users who returned to the chatbot after at least 1 week of inactivity. Stata (version 15; StataCorp LLC) was used for the analyses.

Results

Phase 1 – Ethnographic Interviews

Participants

A total of 43 individual interviews were conducted. Demographic characteristics of participants included 56% (n=24) who were first-time pregnant mothers, 23% (n=10) who were mothers expecting their second or third child, and 21% (n=9) who were post partum with at least 1 child under the age of 1 year.

Qualitative Description Results

A summary of the key themes and subthemes derived from the qualitative description can be found in Table 1. The three main themes identified were (1) resources, (2) mental health journey, and (3) gaps in support. When asked what types of support users were looking for during the perinatal period, they reported seeking guidance from friends, family, peers, mental health professionals, and medical providers. Users also described the emotional trajectory they experienced, highlighting unique challenges present during the various phases of pregnancy and postpartum. Finally, users identified a variety of gaps in support that they felt could be addressed with a chatbot. These included a lack of reliable evidence-based psychoeducation on maternal

XSL•FO
RenderX

mental health, social isolation from peers, and insufficient emotional support from medical professionals.

XSL•FO
RenderX

Table . Qualitative synthesis of findings from ethnographic interviews.

Theme		Description	Quotes
Resources[a]			
	Social support	Friends, family, and online communities eased anxiety through shared experiences and validation.	• "I feel like I leaned a lot on my friends." [P09[b], postpartum, aged 10 mo] • "It's good to just hear opinions of other pregnant people instead of like professional opinions or thoughts." [P22, pregnant, third trimester]
	Mental health tools	Apps and therapists provided emotional support, tracking, and coping strategies.	• "I've been in touch with my therapist since late last year... It's been really helpful just having an actual therapist to talk to and talk my anxieties out." [P41, pregnant, second trimester]
	Clinical guidance	Health care providers and doulas were essential for reassurance, though often limited in time for deeper support.	• "I listen to my doctor above all else. For my first birth, I didn't want a C-Section, and they let me go three days and respected my wishes. I had such a positive experience with my health care team, which I know not everyone has, so I trust my doctor above all else." [P07, pregnant, second trimester, aged 3 y]
Mental health journey[c]			
	Emotional trajectory	Anxiety, isolation, and the unpredictability of pregnancy and parenthood shaped mental health challenges. The pregnancy journey was marked by fluctuating emotions, anxiety, and adjustments in mental health.	• "Mentally a very isolating time anyways – feel like unless you know someone close to you who is pregnant around the same time, even in normal world standards, it can be very isolating." [P16, pregnant, third trimester] • "I'm not as happy as I was in the second trimester, I feel always tired. A little bit more emotionally trying, especially since I can't hang out with my friends." [P02, pregnant, third trimester]
	Early pregnancy stress	Intense physical symptoms and fear of complications caused heightened anxiety in the first trimester.	• "Trying to think that you know I don't want the baby to feel this. I want the baby to know that she's loved and don't want her to feel affected by the situation." [P40, postpartum, aged 15 mo]
	Late pregnancy anxiety	Worries about labor and delivery escalated in the third trimester, along with physical discomfort.	• "You're supposed to be happy about this, it shouldn't be terrible, not a walk in the park, but unknowns surrounding childbirth have been the biggest worry in back of my mind" [P30, pregnant, third trimester]

XSL•FO
RenderX

Theme		Description	Quotes
	Postpartum adjustment	Burnout, hormonal changes, and lack of support created emotional strain during the postpartum period.	• "Honestly, I was burned out. I was giving and giving, and it felt like I had nothing left for myself." [P15, postpartum, aged 10 mo] • "I think the hardest part for me postpartum was not taking care of the baby but how do I take care of myself." [P40, postpartum, aged 15 mo]
Gaps in support[d]			

XSL•FO
RenderX

Theme		Description	Quotes
	Maternal mental health	Overemphasis on baby care left mothers feeling unsupported in managing their own emotional well-being.	• "... would have been nice to have more information on raising kids developmentally along with doing things that help me cope with boiling of stress." [P07, pregnant, second trimester, aged 3 y] • "All the apps that I have are pretty much focused on the baby. Besides the workout one, I would say there's not an app that's more focused on the mother." [P11, postpartum, aged 5 mo]
	Community isolation	Young or first-time moms felt isolated and lacked connections to peers for emotional support.	• "I've never been around people that are pregnant and haven't spent much time around little babies so I feel ignorant to most things regarding pregnancy, giving birth and having a child. Starting from complete scratch…" [P01, pregnant, second trimester]
	Limited time and support from clinicians	Frustration with short, impersonal doctor visits that prioritized medical information over emotional support	• "[Appointments are so short] that I didn't want to bother my doctor all the time with little things I was feelings." [P16, pregnant, third trimester] • "Yeah, the first few months were hard and I felt like that was when I needed the most like reassurance that like 'This is temporary.'" [P26, postpartum, aged 1 y, aged 4 y]
	Lack of understandable and credible sources	Frequently encountered misinformation online and desired reliable, easy-to-understand sources.	• "Message boards mostly don't work for me -- joined some of the What to Expect forums, and 75% of the time they lead you down such a rabbit hole, people have such different opinions and its not helpful and maybe scares you more. 25% of the time I like them because knowing a bunch of people who are going through the same things is helpful." [P35, pregnant, second trimester, aged 3 y]

[a]Integration into prototype: Moment for Parents was designed to offer validation, reassurance, and coping strategies for mothers. Developed by emotional well-being experts—including a doula, perinatal therapist, and yoga and meditation instructor—it incorporated real-life examples from mothers to highlight shared experiences.

[b]P: participant.

[c]Integration into prototype: Moment for Parents included acceptance and commitment therapy exercises to boost self-acceptance and psychological flexibility, helping moms process emotions such as isolation and anxiety. It also used cognitive behavioral therapy techniques to address worries about physical symptoms and mindfulness practices for coping with discomfort. During each check-in, the chatbot encouraged users to create a self-care plan and suggested activities lasting 5, 15, 30, and 60 minutes.

[d]Integration into prototype: Moment for Parents focused entirely on mothers' emotional and mental well-being, reinforcing the importance of self-care in supporting their baby. The chatbot guided moms in building a support system through exercises on accepting and asking for help, challenging the "do it all" myth, and promoting a mindset of community care.

XSL•FO
RenderX

Prototype Development

The results from the analysis of the ethnographic interviews directly influenced the content in the first Moment for Parents prototype. First, the "resources" theme resulted in exercises aiming to provide validation, reassurance, and coping strategies for users. Given specific user needs around social support, Moment for Parents also included real examples from parents with mental health struggles to illustrate shared experiences. Second, results from the "mental health journey" theme inspired acceptance and commitment therapy based exercises around self-acceptance and psychological flexibility. The goal for these exercises was to help users process their negative emotions, including anxiety and isolation, that users reported experiencing throughout the perinatal period. The "mental health journey" theme also led to the development of exercises tailored to the different phases of the perinatal period. For example, cognitive behavioral therapy exercises to increase coping with worries about physical symptoms in early pregnancy, and exercises focused on increasing self-care exercises during the postpartum period. Finally, the "gaps in support" theme led to the inclusion of exercises designed to help users increase help-seeking and build their support systems. The multidisciplinary team of experts vetted and fleshed out the content based on the user interview themes. Other aspects of the prototype, such as the structure (eg, frequency of messaging), were also informed by insights from the team of experts.

The first prototype consisted of a rules-based chatbot called Moment for Parents embedded within a Web app. Upon logging in, participants completed a mood check-in. Based on their mood at the check-in, the chatbot offered an exercise tailored to their current mood. The Moment for Parents prototype had a library of 35 exercises that the chatbot would select based on the users' reported mood. A second version of the prototype was adopted after 4 weeks based on participant feedback. In this version, participants were offered the option to select their own journey through the app. Options included the following: (1) an exercise based on mood, (2) an exercise on a predetermined topic (ie, "topic of the day"), and (3) a self-care check-in. The following day, the chatbot followed up with participants to encourage accountability and engagement in the exercises provided by the chatbot.

Phase 2 – Pilot Test

Participant Characteristics

A total of 108 participants, 29 pregnant and 79 post partum, participated in the pilot test of the Moment for Parents prototype. Among pregnant participants, 69% (n=20) were experiencing their first pregnancy, while 31% (n=9) were pregnant for their second time or more. Additionally, 17% (n=5) of pregnant participants were in their first trimester, 31% (n=9) in their second trimester, and 52% (n=15) in their third trimester. For postpartum participants, 82% (n=65) had 1 child, and 18% (n=14) had multiple children. Additionally, 25% (n=20) participants were 0 - 13 weeks post partum, 25% (n=20) participants were 14 - 26 weeks postpartum, 17% (n=13) participants were 27 - 40 weeks postpartum, and 33% (n=26) participants were 41 - 52 weeks postpartum.

Overall Engagement

Participants completed a total of 1036 check-ins and exercises between November 2019 and June 2020, with an average of 9.6 (SD 14.5) check-ins per participant and a mean of 4.83 weeks (SD 5.17) of engagement. A total of 31.9% (n=331) of check-ins occurred during the first 2 weeks, and 46.1% (n=478) occurred within the first month. Over the second month, check-ins dropped to 17.5% (n=181), followed by 9.6% (n=99) in the third month. During the fourth and fifth months, 6.5% (n=67) of check-ins occurred each month, followed by a slight rise to 6.8% (n=68) in the sixth month. Check-ins then decreased to 2.9% (n=30) in the seventh month and 1.0% (n=10) in the eighth month.

Table 2 presents the number of participants who engaged with the Web app at least once within each given month for all participants and separately for pregnant and postpartum groups. At 1 month, nearly all participants (98.1%) remained active, with higher retention among postpartum users (100%) compared to pregnant users (93.1%). Retention remained high at 2 months, with almost half (47.2%) of users still engaged. While engagement declined over time, postpartum users showed sustained interest, with 15.2% active at 7 months and 7.6% at 8 months.

XSL•FO
RenderX

Table . Moment for parents monthly active users over the 8-month evaluation (n=108).[a]

	All (N=108), n (%)	Pregnant (n=29), n (%)	Postpartum (n=79), n (%)
1 month	106 (98.1)	27 (93.1)	79 (100)
2 months	51 (47.2)	9 (31.0)	42 (53.2)
3 months	22 (20.4)	3 (10.3)	19 (24.1)
4 months	17 (15.7)	3 (10.3)	14 (17.7)
5 months	22 (20.4)	4 (13.8)	18 (22.8)
6 months	14 (13)	1 (3.4)	13 (16.5)
7 months	12 (11.1)	—[b]	12 (15.2)
8 months	6 (5.6)	—	6 (7.6)

[a]Most users (n=90, 83.3%) were active during their first week of use; 47.2% (51/108; 9/29, 31% pregnant, 42/79, 53.2% postpartum) used the app during their second week.

[b]Not available.

Re-Engagement Patterns

A total of 25 (5 pregnant and 20 postpartum) participants used the app for a median of 1 week (mean 1.8, SD 3.33 wk) before stopping and never returning. In contrast, 69 (14 pregnant and 55 postpartum) participants eventually re-engaged with the app after at least one week of inactivity. Among these participants, 44.9% (31/69) returned only once (8/14, 57% of pregnant participants; 23/55, 42% of postpartum participants). Re-engagement was operationalized as the number of times participants returned to the app after stopping for at least 1 week of inactivity. Of participants who re-engaged, a large proportion (28/69, 41%) did so 3 or more times after at least 1 week of inactivity. Postpartum participants re-engaged more often and more quickly than pregnant participants. Most re-engagements took place between 1 and 2 weeks after initial re-engagement, but a significant percentage of re-engagements occurred after 4 weeks (58/402, 14%; Table 3).

Table . Re-engagement patterns among those who re-engaged with moment for parents after stopping (n=69 participants; n=402 check-ins).

		All	Pregnant	Postpartum
Time to re-engagement (week), mean (SD)[a]		2.8 (2.6)	3.2 (3.2)	2.6 (2.5)
Number of re-engagements, n (%)[a]				
	1 re-engagement	31 (44.9)	8 (57.1)	23 (41.8)
	2 re-engagements	10 (14.5)	2 (14.3)	8 (14.5)
	3+ re-engagements	28 (40.6)	4 (28.6)	24 (43.6)
Number of check-ins after re-engagement, n (%)[b]				
	After 1 week	160 (39.8)	28 (38.9)	132 (40)
	After 2 weeks	106 (26.4)	19 (26.4)	87 (26.4)
	After 3 weeks	78 (19.4)	15 (20.8)	63 (19.1)
	After 4+ weeks	58 (14.4)	10 (13.9)	48 (14.5)

[a]All (n=69), pregnant (n=14), and postpartum (n=55).

[b]All (n=402), pregnant (n=72), and postpartum (n=330).

Prototype Feedback

Feedback from users (n=30) who completed the 1-month survey is in Table 4. Most users signed up for Moment for Parents because a friend recommended it (n=15, 55%). Other top reasons for signing up included an interest in becoming more resilient (n=14, 47%), wanting support for a self-care routine (n=13, 43%), and looking for advice to take care of themselves (n=12, 40%). The vast majority (n=28, 93%) felt that Moment for Parents was created for people such as them. Most users thought that Moment for Parents was a good fit for them because it kept them accountable for self-care (n=21, 70%), it reminded them to reflect (n=19, 63%), and it helped them take a mental health break (n=17, 58%). A large proportion of users also felt that Moment for Parents validated what they were going through (n=13, 43%), gave helpful advice (n=13, 43%), and helped them navigate the ups and downs of motherhood (n=12, 40%). While users generally had a positive experience with Moment for Parents, some users reported several “pet peeves” or negative experiences. Half of the users felt that the check-in topics were repetitive (n=15), and 37% (n=11) of users reported a pet peeve when the check-ins had a glitch. A third of users felt that the check-ins did not have the response options they wanted (n=10). Less frequently reported pet peeves included the daily text

XSL•FO
RenderX

reminders (n=8, 27%), the lack of access via a mobile app (n=6, 20%), and feeling as if the check-in topic was not relevant (n=6, 20%). The vast majority of users spent less than 5 minutes completing their check-ins (n=28, 93%), and most (n=24, 80%) felt that the check-in length was “just right.”

XSL•FO
RenderX

Table . Feedback from users who completed the 1-month survey (n=30).

Feedback		Value, n (%)
Why did you sign up for Moment for Parents?		
	Friend recommended	15 (55.3)
	Interested in becoming more resilient	14 (46.7)
	Interested in trying a new product for moms	9 (30)
	Looking for advice to take care of baby	4 (13.3)
	Looking for advice to take care of myself	12 (40)
	Help coping with emotions	10 (33.3)
	Support for self-care routine	13 (43.3)
	Other	1 (3.3)
Was Moment for Parents created for moms like you?		
	Yes	28 (99.3)
Why is Moment for Parents a good fit for you?		
	Feels like a friend	11 (36.7)
	Gives me confidence as a mom	5 (16.7)
	Gives me helpful advice	13 (43.3)
	Helps me become resilient	7 (23.3)
	Helps me navigate ups and downs of motherhood	12 (40)
	Helps me take a mental health break	17 (56.7)
	I know it’s there in case I need it	1 (3.3)
	It keeps me accountable for self-care	21 (70)
	It offers helpful mindfulness exercises	9 (30)
	It reminds me to reflect	19 (63.3)
	It surprises me with new perspectives	6 (20)
	It understands what I’m going through	7 (23.3)
	It validates what I’m going through	13 (43.3)
What are your pet peeves about Moment for Parents?		
	Can’t access via an app	6 (20)
	Daily text reminders	8 (26.7)
	None–it’s great as-is	3 (10)
	Not enough support for sleep	4 (13.3)
	Tone is too empathetic	1 (3.3)
	Something else	5 (16.7)
	Too much emphasis on self-care	3 (10)
	When check-in has a glitch	11 (36.7)
	When check-in questions don’t have the response option I want	10 (33.3)
	When texts come too quickly	3 (10)
	When texts come too slowly	1 (3.3)
	When check-in topic doesn’t feel relevant to me	6 (20)
	When check-in topics are repetitive	15 (50)
	When Moment for Parents doesn’t respond to me appropriately	5 (16.7)
Time to complete check-in		

XSL•FO
RenderX

Feedback		Value, n (%)
	Less than 1 min	3 (10)
	1 - 3 min	20 (66.7)
	3 - 5 min	5 (16.7)
	More than 5 min	2 (6.7)
Check-in length		
	Just right	24 (80)
	Too long	5 (16.7)
	Too short	1 (3.3)

Discussion

Summary of Findings

The purpose of this study was to (1) design and develop a chatbot tailored specifically to educate and support pregnant and postpartum people on emotional well-being and mental health and (2) evaluate the usability of Moment for Parents in a pilot test by examining usage patterns, engagement levels, and user experience. In the design phase, we found that user needs included social support, coping strategies, more attention to mental health and self-care, and reliable medical and mental health information. Users also emphasized a need for support that is tailored to changes over the perinatal period. The end result of phase 1 was an initial prototype of the Moment for Parents chatbot that was used in the phase 2 pilot test. The results from the pilot test showed that participants engaged with the app early on and that sustained engagement decreased over time. However, many participants re-engaged with the chatbot after inactivity, suggesting that users may have turned to Moment for Parents when they needed support. The results from the feedback survey showed that participants were primarily interested in using Moment for Parents to build resilience, support their self-care routines, and seek emotional support during pregnancy and postpartum.

Themes from the qualitative analysis highlighted the key needs and preferences of perinatal users. Participants consistently emphasized the importance of evidence-based strategies to support their mental health, particularly as a complement to in-person care, where provider time for emotional support may be limited. Additionally, users described the fast-paced and unpredictable nature of the perinatal period, reinforcing the need for flexible and adaptive digital interventions that align with their evolving emotional and psychological needs. These findings underscore the importance of delivering personalized, flexible, and easily accessible digital support that aligns with the evolving needs of perinatal individuals. Such alignment may also play a critical role in sustaining user engagement over time.

In the pilot trial, user engagement was greatest in the first 2 months, then decreased over the next 6 months, both in terms of check-ins with the app and monthly active users. Importantly, 47% of users were still active after 2 months—this level of usage is much greater than metrics described in other publicly available mental health applications where users drop to only 4% of the initial sample after 2 weeks [39]. There are several possible explanations for this finding. First, Moment for Parents was developed using HCD, which is thought to enhance engagement with digital mental health interventions [40]. Feedback showed that most participants felt that the information and exercises in Moment for Parents were highly personalized and relevant, which may have led to more sustained engagement. Another possible explanation is that, given the time-consuming demands faced by perinatal individuals such as attending medical appointments, managing physical symptoms, breastfeeding, and childcare, engaging with Moment for Parents required relatively little time and energy from users, as most users reported spending less than 5 minutes with Moment for Parents daily. This is in contrast to other apps and chatbots that are more time and resource intensive [41,42].

While most participants disengaged after 2 months, many re-engaged with Moment for Parents at least once during the pilot. Among those who returned, a large proportion did so 3 or more times, suggesting they found the app useful or relevant to their needs. Re-engagement was highest in the first week (40%), but a notable portion (14%) continued to return after 4 weeks, indicating that users may have turned to the app specifically when they needed support or information. A recent review of 16 trials of digital mental health interventions for perinatal populations found that disengagement varied widely across trials (13% - 90% retention rates) and highlights the challenge of engaging perinatal users [43]. Moment for Parents was designed to deliver targeted support based on mood check-ins rather than continuous interaction, highlighting the importance of minimal contact and flexible mental health tools. Future research should prioritize understanding the timing and reasons behind user re-engagement, rather than solely aiming to increase engagement frequency, to enhance the mental health chatbots' ability to deliver timely, relevant support tailored to users' needs. This could include exploring the specific triggers for re-engagement, the impact of intermittent use on mental health outcomes, and how to refine content delivery to maximize effectiveness when users do return.

Interestingly, both engagement and re-engagement were greater among postpartum participants compared to pregnant participants. Postpartum users may have been experiencing greater mental health challenges than their pregnant counterparts in this sample and, therefore, turned to the app more frequently for exercises to improve their mood. Relatedly, postpartum users may have sought more information from the app as they learned to care for their infants, managed their postpartum

XSL•FO
RenderX

mental health, and adjusted to new parenthood. Additionally, the postpartum period is marked by significant hormonal changes, sleep deprivation, and increased caregiving demands, all of which contribute to low mood [44-46]. Given these challenges, postpartum individuals may be more open to digital mental health support as a convenient and accessible resource for coping strategies, reassurance, and parenting guidance [47]. In contrast, pregnant users may have relied more on in-person medical support or other existing resources, leading to lower engagement with the app. Alternatively, some users indicated in informal feedback that the content presented in Moment for Parents was geared toward postpartum more than pregnant user needs. Overall, the limited number of pregnant individuals in our sample prevents definitive conclusions, highlighting the need for further research on tailoring interventions to pregnant user needs.

The findings from the 1-month feedback survey reinforced the value of Moment for Parents in supporting perinatal mental health. The users reported that the tool provided helpful advice, supported self-care accountability, and facilitated emotional reflection, highlighting its role in bridging gaps in traditional care [48,49]. Many found the content relevant and beneficial, though some noted frustration with repetitive or irrelevant check-ins, emphasizing the need for customizable, phase-specific content. These findings underscore the potential of personalized digital interventions to enhance accessibility and engagement in perinatal mental health support.

Limitations

This study was not without limitations. First, the samples used in both phases of this study were relatively small and prone to bias because of the snowball sampling method used. Certain groups or participant characteristics may have been overrepresented, leading to skewed results. Furthermore, only 25% of the initial sample completed the 1-month survey, leading to potentially biased survey results if, for example, only users who enjoyed the app completed the survey. Relatedly, another limitation was that no demographic data was collected about study participants in either phase, limiting our ability to speak to the generalizability or biases in the samples studied. Second, the sample primarily consisted of postpartum users, limiting our ability to make inferences about the feasibility and utility of Moment for Parents across the perinatal period. Future work should focus on recruiting a larger sample of pregnant users at various stages of pregnancy, as well as postpartum users at various postpartum stages (eg, early postpartum vs late postpartum). Third, the way the feedback questions were asked may have led participants to provide overly positive feedback (eg, “why was Moment for Parents a good fit” vs “to what extent do you agree that Moment for Parents was a good fit?”). Future evaluations of Moment for Parents should include a more balanced survey to capture a variety of perspectives. Fourth, systematic data were not collected around the nature and timing of the iterative updates, which limits our ability to understand how these updates impacted engagement or satisfaction with Moment for Parents. Finally, data were not collected around user enactment of the Moment for Parents exercises outside of the check-ins. Future research should focus on understanding how users integrate the strategies they learn from Moment for Parents into their daily lives.

Conclusions

This study provides early evidence that Moment for Parents is a practical and scalable tool for perinatal mental health support. HCD informed the chatbot’s development, ensuring it aligned with the needs of perinatal users before being tested in a pilot trial. Findings showed that participants engaged with the chatbot far longer than is typical for digital interventions and often returned for support even weeks after initially stopping use—highlighting its usability and value. However, more research is needed to fully understand the chatbot’s efficacy in mitigating adverse perinatal mental health outcomes. These results contribute to the growing body of research on digital mental health interventions in the perinatal period, reinforcing the potential of chatbots to deliver accessible, on-demand support.

Data Availability

The dataset used for this retrospective analysis is not publicly available. However, aggregated and anonymized data may be shared with interested parties upon reasonable request.

Conflicts of Interest

MK is employed by Poisera, Inc (dba Moment for Parents), and Moment for Parents was used in this study. JH is a paid consultant of Moment for Parents, and KM and LB are employed by JH. However, authors’ employment status or salary are not dependent upon the results of their research.

References

1. Maternal mental health. World Health Organization. URL: https://www.who.int/teams/mental-health-and-substance-use/promotion-prevention/maternal-mental-health [accessed 2025-03-28]
2. Perinatal mental health and substance use disorders [white paper]. Americal Psychiatric Association. 2023. URL: https://www.psychiatry.org/getmedia/344c26e2-cdf5-47df-a5d7-a2d444fc1923/APA-CDC-Perinatal-Mental-and-Substance-Use-Disorders-Whitepaper.pdf [accessed 2025-03-28]
3. Reproductive health: symptoms of depression among women. CDC. URL: https://www.cdc.gov/reproductive-health/depression/index.html [accessed 2025-03-28]

XSL•FO
RenderX

4. Howard LM, Khalifeh H. Perinatal mental health: a review of progress and challenges. World Psychiatry 2020 Oct;19(3):313-327. [doi: 10.1002/wps.20769] [Medline: 32931106]
5. Petersen EE, Davis NL, Goodman D, et al. Racial/ethnic disparities in pregnancy-related deaths - United States, 2007-2016. MMWR Morb Mortal Wkly Rep 2019 Sep 6;68(35):762-765. [doi: 10.15585/mmwr.mm6835a3] [Medline: 31487273]
6. Ceballos M, Wallace G, Goodwin G. Postpartum depression among African-American and Latina mothers living in small cities, towns, and rural communities. J Racial Ethn Health Disparities 2016 Oct 19;4:916-927. [doi: 10.1007/s40615-016-0295-z] [Medline: 27761728]
7. Slomian J, Honvo G, Emonts P, Reginster JY, Bruyère O. Consequences of maternal postpartum depression: a systematic review of maternal and infant outcomes. Womens Health (Lond) 2019;15:1745506519844044. [doi: 10.1177/1745506519844044] [Medline: 31035856]
8. Biaggi A, Conroy S, Pawlby S, Pariante CM. Identifying the women at risk of antenatal anxiety and depression: a systematic review. J Affect Disord 2016 Feb;191:62-77. [doi: 10.1016/j.jad.2015.11.014] [Medline: 26650969]
9. Kingston D, Austin MP, Heaman M, et al. Barriers and facilitators of mental health screening in pregnancy. J Affect Disord 2015 Nov 1;186:350-357. [doi: 10.1016/j.jad.2015.06.029] [Medline: 26281038]
10. Ford E, Roomi H, Hugh H, van Marwijk H. Understanding barriers to women seeking and receiving help for perinatal mental health problems in UK general practice: development of a questionnaire. Prim Health Care Res Dev 2019 Dec 12;20:e156. [doi: 10.1017/S1463423619000902] [Medline: 31826794]
11. Button S, Thornton A, Lee S, Shakespeare J, Ayers S. Seeking help for perinatal psychological distress: a meta-synthesis of women's experiences. Br J Gen Pract 2017 Oct;67(663):e692-e699. [doi: 10.3399/bjgp17X692549] [Medline: 28847773]
12. Brailey C, Slatton BC. Centering Black women's voices: illuminating systemic racism in maternal healthcare experiences. Societies (Basel) 2024 May;14(5):70. [doi: 10.3390/soc14050070]
13. Estriplet T, Morgan I, Davis K, Crear Perry J, Matthews K. Black perinatal mental health: prioritizing maternal mental health to optimize infant health and wellness. Front Psychiatry 2022;13:807235. [doi: 10.3389/fpsyt.2022.807235] [Medline: 35573337]
14. Moore D, Ayers S, Drey N. A thematic analysis of stigma and disclosure for perinatal depression on an online forum. JMIR Ment Health 2016 May 19;3(2):e18. [doi: 10.2196/mental.5611] [Medline: 27197516]
15. Megnin-Viggars O, Symington I, Howard LM, Pilling S. Experience of care for mental health problems in the antenatal or postnatal period for women in the UK: a systematic review and meta-synthesis of qualitative research. Arch Womens Ment Health 2015 Dec;18(6):745-759. [doi: 10.1007/s00737-015-0548-6] [Medline: 26184835]
16. Peahl AF, Moniz MH, Heisler M, et al. Experiences with prenatal care delivery reported by black patients with low income and by health care workers in the US: a qualitative study. JAMA Netw Open 2022 Oct 3;5(10):e2238161. [doi: 10.1001/jamanetworkopen.2022.38161] [Medline: 36279136]
17. Song H, Cramer EM, McRoy S, May A. Information needs, seeking behaviors, and support among low-income expectant women. Women Health 2013;53(8):824-842. [doi: 10.1080/03630242.2013.831019] [Medline: 24215275]
18. Lewkowitz AK, Whelan AR, Ayala NK, et al. The effect of digital health interventions on postpartum depression or anxiety: a systematic review and meta-analysis of randomized controlled trials. Am J Obstet Gynecol 2024 Jan;230(1):12-43. [doi: 10.1016/j.ajog.2023.06.028]
19. Suharwardy S, Ramachandran M, Leonard SA, et al. Feasibility and impact of a mental health chatbot on postpartum mental health: a randomized controlled trial. AJOG Glob Rep 2023 Aug;3(3):100165. [doi: 10.1016/j.xagr.2023.100165] [Medline: 37560011]
20. Gaffney H, Mansell W, Tai S. Conversational agents in the treatment of mental health problems: mixed-method systematic review. JMIR Ment Health 2019 Oct 18;6(10):e14166. [doi: 10.2196/14166] [Medline: 31628789]
21. Car LT, Dhinagaran DA, Kyaw BM, et al. Conversational agents in health care: scoping review and conceptual analysis. J Med Internet Res 2020 Aug 7;22(8):e17158. [doi: 10.2196/17158] [Medline: 32763886]
22. Vaidyam AN, Wisniewski H, Halamka JD, Kashavan MS, Torous JB. Chatbots and conversational agents in mental health: a review of the psychiatric landscape. Can J Psychiatry 2019 Jul;64(7):456-464. [doi: 10.1177/0706743719828977] [Medline: 30897957]
23. Booth F, Potts C, Bond R, et al. A mental health and well-being chatbot: user event log analysis. JMIR mHealth uHealth 2023 Jul 6;11:e43052. [doi: 10.2196/43052] [Medline: 37410539]
24. Lucas GM, Gratch J, King A, Morency LP. It's only a computer: virtual humans increase willingness to disclose. Comput Human Behav 2014 Aug;37:94-100. [doi: 10.1016/j.chb.2014.04.043]
25. Lucas GM, Rizzo A, Gratch J, et al. Reporting mental health symptoms: breaking down barriers to care with virtual human interviewers. Front Robot AI 2017;4:51 [FREE Full text] [doi: 10.3389/frobt.2017.00051]
26. Stiles-Shields C, Cummings C, Montague E, Plevinsky JM, Psihogios AM, Williams KDA. A call to action: using and extending human-centered design methodologies to improve mental and behavioral health equity. Front Digit Health 2022;4:848052. [doi: 10.3389/fdgth.2022.848052] [Medline: 35547091]
27. Veinot TC, Mitchell H, Ancker JS. Good intentions are not enough: how informatics interventions can worsen inequality. J Am Med Inform Assoc 2018 Aug 1;25(8):1080-1088. [doi: 10.1093/jamia/ocy052] [Medline: 29788380]

XSL•FO
RenderX

28. Are pregnant and postpartum women interested in health-related apps? Implications for the prevention of perinatal depression. Consensus. URL: https://consensus.app/papers/are-pregnant-and-postpartum-women-interested-in-osma-barrera/43c0340c7c765fca9ee27233f11516f4/?utm_source=chatgpt [accessed 2025-03-28]
29. Connor K, Wambach K, Baird MB. Descriptive, qualitative study of women who use mobile health applications to obtain perinatal health information. J Obstet Gynecol Neonatal Nurs 2018 Nov;47(6):728-737. [doi: 10.1016/j.jogn.2018.04.138] [Medline: 29856963]
30. Tsai Z, Kiss A, Nadeem S, et al. Evaluating the effectiveness and quality of mobile applications for perinatal depression and anxiety: a systematic review and meta-analysis. J Affect Disord 2022 Jan 1;296:443-453. [doi: 10.1016/j.jad.2021.09.106] [Medline: 34624434]
31. Mackintosh N, Agarwal S, Adcock K, et al. Online resources and apps to aid self-diagnosis and help seeking in the perinatal period: a descriptive survey of women's experiences. Midwifery 2020 Nov;90:102803. [doi: 10.1016/j.midw.2020.102803] [Medline: 32717660]
32. Gordon M, Henderson R, Holmes JH, Wolters MK, Bennett IM, SPIRIT (Stress in Pregnancy: Improving Results with Interactive Technology) Group. Participatory design of ehealth solutions for women from vulnerable populations with perinatal depression. J Am Med Inform Assoc 2016 Jan;23(1):105-109. [doi: 10.1093/jamia/ocv109] [Medline: 26342219]
33. Rathnayaka P, Mills N, Burnett D, De Silva D, Alahakoon D, Gray R. A mental health chatbot with cognitive skills for personalised behavioural activation and remote health monitoring. Sensors (Basel) 2022 May 11;22(10):3653. [doi: 10.3390/s22103653] [Medline: 35632061]
34. Abd-Alrazaq AA, Alajlani M, Alalwan AA, Bewick BM, Gardner P, Househ M. An overview of the features of chatbots in mental health: a scoping review. Int J Med Inform 2019 Dec;132:103978. [doi: 10.1016/j.ijmedinf.2019.103978] [Medline: 31622850]
35. Mancinelli E, Magnolini S, Gabrielli S, Salcuni S. A chatbot (Juno) prototype to deploy a behavioral activation intervention to pregnant women: qualitative evaluation using a multiple case study. JMIR Form Res 2024 Aug 14;8:e58653. [doi: 10.2196/58653] [Medline: 39140593]
36. ISO 9241-210:2019. ISO. URL: https://www.iso.org/standard/77520.html [accessed 2025-03-28]
37. López L, Green AR, Tan-McGrory A, King R, Betancourt JR. Bridging the digital divide in health care: the role of health information technology in addressing racial and ethnic disparities. Jt Comm J Qual Patient Saf 2011 Oct;37(10):437-445. [doi: 10.1016/s1553-7250(11)37055-9] [Medline: 22013816]
38. Bradshaw C, Atkinson S, Doody O. Employing a qualitative description approach in health care research. Glob Qual Nurs Res 2017;4:2333393617742282. [doi: 10.1177/2333393617742282] [Medline: 29204457]
39. Baumel A, Muench F, Edan S, Kane JM. Objective user engagement with mental health apps: systematic search and panel-based usage analysis. J Med Internet Res 2019 Sep 25;21(9):e14567. [doi: 10.2196/14567] [Medline: 31573916]
40. Mohr DC, Riper H, Schueller SM. A solution-focused research approach to achieve an implementable revolution in digital mental health. JAMA Psychiatry 2018 Feb 1;75(2):113-114. [doi: 10.1001/jamapsychiatry.2017.3838] [Medline: 29238805]
41. Darcy A, Beaudette A, Chiauzzi E, et al. Anatomy of a Woebot® (WB001): agent guided CBT for women with postpartum depression. Expert Rev Med Devices 2023;20(12):1035-1049. [doi: 10.1080/17434440.2023.2280686] [Medline: 37938145]
42. Hussain-Shamsy N, Shah A, Vigod SN, Zaheer J, Seto E. Mobile health for perinatal depression and anxiety: scoping review. J Med Internet Res 2020 Apr 13;22(4):e17011. [doi: 10.2196/17011] [Medline: 32281939]
43. Davis JA, Ohan JL, Gibson LY, Prescott SL, Finlay-Jones AL. Understanding engagement in digital mental health and well-being programs for women in the perinatal period: systematic review without meta-analysis. J Med Internet Res 2022 Aug 9;24(8):e36620. [doi: 10.2196/36620] [Medline: 35943773]
44. Schiller CE, Meltzer-Brody S, Rubinow DR. The role of reproductive hormones in postpartum depression. CNS Spectr 2015 Feb;20(1):48-59. [doi: 10.1017/S1092852914000480] [Medline: 25263255]
45. Iranpour S, Kheirabadi GR, Esmaillzadeh A, Heidari-Beni M, Maracy MR. Association between sleep quality and postpartum depression. J Res Med Sci 2016;21(1):110. [doi: 10.4103/1735-1995.193500]
46. Saharoy R, Potdukhe A, Wanjari M, Taksande AB. Postpartum depression and maternal care: exploring the complex effects on mothers and infants. Cureus 2023 Jul;15(7):e41381. [doi: 10.7759/cureus.41381] [Medline: 37546054]
47. Feldman N, Perret S. Digital mental health for postpartum women: perils, pitfalls, and promise. NPJ Digit Med 2023 Jan 26;6(1):11. [doi: 10.1038/s41746-023-00756-4] [Medline: 36702915]
48. Zeng Z, Peng J, Liu L, Gong W. Translating research evidence into marketplace application: cohort study of internet-based intervention platforms for perinatal depression. J Med Internet Res 2023 Apr 17;25(1):e42777. [doi: 10.2196/42777] [Medline: 37067855]
49. Feldman N, Back D, Boland R, Torous J. A systematic review of mHealth application interventions for peripartum mood disorders: trends and evidence in academia and industry. Arch Womens Ment Health 2021 Dec;24(6):881-892. [doi: 10.1007/s00737-021-01138-z] [Medline: 33929636]

Abbreviations

BRANY: Biomedical Research Alliance of New York

XSL•FO
RenderX

HCD: human-centered design

Edited by A Mavragani; submitted 10.02.25; peer-reviewed by A Lewkowitz, R Joly; revised version received 10.03.25; accepted 11.03.25; published 08.04.25.

Please cite as:
McAlister K, Baez L, Huberty J, Kerppola M
Chatbot to Support the Mental Health Needs of Pregnant and Postpartum Women (Moment for Parents): Design and Pilot Study
JMIR Form Res 2025;9:e72469
URL: https://formative.jmir.org/2025/1/e72469
doi:10.2196/72469

© Kelsey McAlister, Lara Baez, Jennifer Huberty, Marianna Kerppola. Originally published in JMIR Formative Research (https://formative.jmir.org), 8.4.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Anticipated Acceptability of Blended Learning Among Lay Health Care Workers in Malawi: Qualitative Analysis Guided by the Technology Acceptance Model

Tiwonge E Mbeya-Munkhondya[1], PhD; Caroline J Meek[2], MPH; Mtisunge Mphande[3], BSc; Tapiwa A Tembo[3], MSc; Mike J Chitani[3], BA; Milenka Jean-Baptiste[4], MPH; Caroline Kumbuyo[3], BSoc; Dhrutika Vansia[3], MPH; Katherine R Simon[3,5], MD; Sarah E Rutstein[6], MD, PhD; Victor Mwapasa[7], PhD; Vivian Go[4], PhD; Maria H Kim[3,5], MD; Nora E Rosenberg[4], PhD

[1]School of Nursing, Kamuzu University of Health Sciences, Lilongwe, Malawi
[2]RTI International, Research Triangle Park, NC, United States
[3]Baylor College of Medicine Children's Foundation, Golden Peacock, 5 Floor, Lilongwe, Malawi
[4]Gillings School of Public Health, Chapel Hill, NC, United States
[5]Department of Pediatrics, Baylor College of Medicine, Houston, TX, United States
[6]Department of Medicine, University or North Carolina, Chapel Hill, NC, United States
[7]Department of Community and Environmental Health, Kamuzu University of Health Sciences, Blantyre, Malawi

Corresponding Author:
Dhrutika Vansia, MPH
Baylor College of Medicine Children's Foundation, Golden Peacock, 5 Floor, Lilongwe, Malawi

Abstract

Background: HIV index case testing (ICT) aims to identify people living with HIV and their contacts, engage them in HIV testing services, and link them to care. ICT implementation has faced challenges in Malawi due to limited counseling capacity among lay health care workers (HCWs). Enhancing capacity through centralized face-to-face training is logistically complex and expensive. A decentralized blended learning approach to HCW capacity-building, combining synchronous face-to-face and asynchronous digital modalities, may be an acceptable way to address this challenge.

Objective: The objective of this analysis is to describe factors influencing HCW anticipated acceptability of blended learning using the Technology Acceptance Model (TAM).

Methods: This formative qualitative study involved conducting 26 in-depth interviews with HCWs involved in the ICT program across 14 facilities in Machinga and Balaka, Malawi (November-December 2021). Results were analyzed thematically using TAM. Themes were grouped into factors affecting the 2 sets of TAM constructs: perceived usefulness and perceived ease of use.

Results: A total of 2 factors influenced perceived usefulness. First, HCWs found the idea of self-guided digital learning appealing, as they believed it would allow for reinforcement, which would facilitate competence. They also articulated the need for opportunities to practice and receive feedback through face-to-face interactions in order to apply the digital components. In total, 5 factors influenced perceived ease of use. First, HCWs expressed a need for orientation to the digital technology given limited digital literacy. Second, they requested accessibility of devices provided by their employer, as many lacked personal devices. Third, they wished for adequate communication surrounding their training schedules, especially if they were going to be asynchronous. Fourth, they wished for support for logistical arrangements to avoid work interruptions. Finally, they wanted monetary compensation to motivate learning, a practice comparable with offsite trainings.

Conclusions: A decentralized blended learning approach may be an acceptable method of enhancing ICT knowledge and skills among lay HCWs in Malawi, although a broad range of external factors need to be considered. Our next step is to integrate these findings into a blended learning package and examine perceived acceptability of the package in the context of a cluster randomized controlled trial.

International Registered Report Identifier (IRRID): RR2-10.1136/bmjopen-2023-077706

(JMIR Form Res 2025;9:e62741) doi:10.2196/62741

KEYWORDS

blended learning; technology acceptance model; index case testing; acceptability; partner notification; contact tracing

XSL•FO
RenderX

Introduction

While Malawi is making substantial progress in achieving the Joint United Nations Programme on HIV and AIDS (UNAIDS) 95-95-95 targets by 2030, the first target (HIV status awareness by persons living with HIV) has not been reached [1]. To find persons living with HIV unaware of their HIV status, the World Health Organization (WHO) promotes HIV index case testing (ICT), in which contacts of individuals living with HIV are identified, counselled, offered HIV testing, and linked to care [2]. Specifically, WHO recommends an option for voluntary assisted partner notification, in which health care workers help index clients invite their contacts for HIV testing, an approach that has been incorporated into Malawian standard of care [3-5].

Despite these recommendations, health care workers (HCWs) have limited counseling capacity that hinders full realization of voluntary assisted partner notification. For instance, formative work in Malawi identified limited counselling capacity among HCWs as a major challenge to enhancing the implementation of ICT [5]. ICT is offered by lay HCWs, cadres who lack professional training but are expected to perform select tasks within the health care system based on brief in-service training. Standard ICT training for lay HCWs is centralized, in-person, and delivered didactically. However, such approaches are logistically challenging and costly. A blended learning implementation package, which can be offered in a decentralized manner, asynchronously, and with nondidactic components is a promising alternative training approach for HCWs.

Blended learning, which combines face-to-face with digital modalities [6-8], is often more effective than digital or face-to-face training alone [9-11]. In recent years, the use of digital and blended educational approaches have increased, especially with COVID-19–driven pivots to remote learning [12-17]. Blended learning has promoted student motivation, skill-building, and achievement in several contexts [18-20]. However, blended learning has not been fully evaluated for implementation outcomes, especially among lay HCWs in low- and middle-income countries (LMICs) with limited technological literacy [21,22]. Thus, there is limited evidence as to whether lay cadres would find blended learning to be acceptable. We thus developed a cluster randomized controlled trial to assess the impact of a blended learning implementation package compared with the standard of care training on implementation, effectiveness, and cost-effectiveness outcomes [23]. Before developing this package, we conducted formative research to understand factors that may influence the acceptability of blended learning among lay HCWs in LMICs with limited technological literacy [24,25]. This formative work uses in-depth interviews (IDIs) to assess lay HCWs' perceptions of a blended learning approach to ICT training. Our analysis was guided by the Technology Acceptance Model (TAM).

Methods

Conceptual Framework

To explore the anticipated acceptability of a proposed blended learning implementation package, we used the TAM to guide data analysis. TAM is a theoretical model that examines constructs influencing the acceptance or rejection of technologies [26]. The model proposes that acceptability of any technology can be predicted by 2 factors: perceived usefulness and perceived ease of use (Figure 1). Perceived usefulness refers to the belief that a certain technology would increase an individual's job performance. Perceived ease of use is a belief that the use of the new technology would be effortless [27]. TAM proposes that a populations' perceptions of usefulness and ease of use of a technology impact the population's intention to use the technology, which in turn influences the population's actual usage of the technology. Furthermore, the key constructs of perceived usefulness and perceived ease of use are influenced by external factors, such as pedagogy, digital literacy, infrastructure, and the learning environment. Understanding these external factors is essential for designing a such a package. We used the TAM in this formative sub-study to guide the analysis of the relevant external factors influencing perceived usefulness and perceived ease of use surrounding blended learning in the context of ICT.

Figure 1. The Technology Acceptance Model.

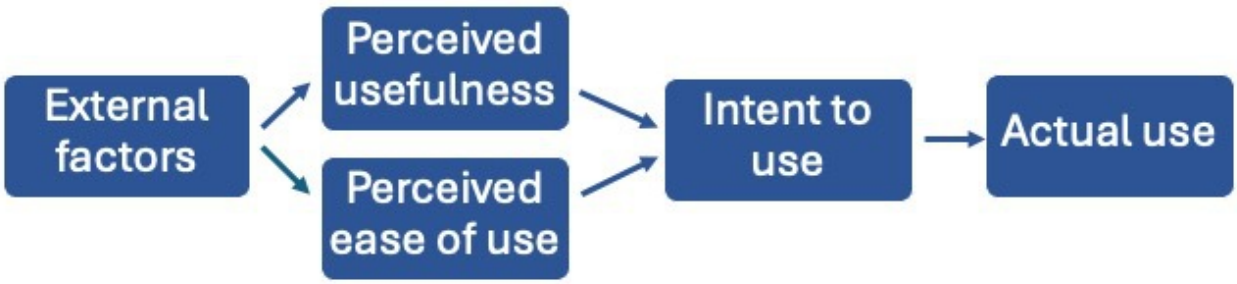

Setting

This study was conducted at health centers in Balaka and Machinga, two districts in Southeastern Malawi. Malawi is a country in Southeastern Africa with an adult HIV prevalence of 8.9% [1]. Although Malawi has met the second two UNAIDS 95-95-95 targets, it has not yet reached the first 95% HIV testing target, with only 88.3% of adults living with HIV aware of their HIV status at the time of the study [1]. Simultaneously, Malawi has one of the lowest rates of qualified HCWs globally with less than 1 medical doctor per 10,000 people [28], resulting in task shifting of HIV services to lay cadres with less formal training [29]. Particularly relevant to the proposed blended learning strategy, is the fact that Malawi has one of the lowest digital literacy levels (having the necessary skills and knowledge

XSL•FO
RenderX

to confidently and effectively use technology) globally ranking 131 out of 134 countries by Network Readiness Index [30].

Study Design and Population

This formative work was conducted to inform a cluster randomized controlled trial designed to examine the implementation, effectiveness, and cost-effectiveness of a blended learning ICT package [23]. Data for this qualitative substudy were generated through IDIs conducted with lay HCWs between November and December 2021. This formative study purposively selected facilities from both districts (Machinga and Balaka) and different levels of care (dispensaries, health centers, and hospitals) for purposes of generalizability. To be eligible for participation, HCWs had to be at least 18 years old, work full-time at one of the health facilities, and be involved in counseling index clients or tracing contact clients. With engagement from facility supervisors, the research team identified up to 2 eligible participants in each facility. These individuals were actively involved in ICT at the time of the study. A research team member visited all the eligible participants in the 14 facilities to give them more details about the study and to seek their consent. Contact details of all consenting participants were shared with one of the research team members who made arrangements with the participants for the date and time of the interviews.

Data Collection

IDIs were conducted by 3 trained Malawian qualitative interviewers in Chichewa, the local language. All interviews were conducted by 3 female external interviewers. Among them, 1 interviewer had a doctoral degree in social sciences and 2 had bachelor's degrees in social sciences. The interviewers were trained on techniques of normalization and probing to reduce the risk of socially desirable response bias and had previous training and experience in qualitative data collection and research positions within the organization (postdoctoral fellow, senior research officer, and research officer). Furthermore, interviews were conducted in a private setting at the facilities, with only the interviewers and participants present. Interviews were conducted using a semistructured guide developed for the study through a series of rigorous, iterative discussions among the research team. The questions for this analysis focused on feasibility and acceptability of a blended learning implementation package.

After consenting and before conducting the interview, the research staff introduced themselves and their role, explained the purpose of the interview, and reiterated that they could refuse to answer questions, that all data would be anonymized, and that responses would have no consequences on their employment. Interviewers first asked questions about participants' views of the standard training. Interviewers then provided participants with a brief description of the planned blended learning approach, explaining that it involved tablet-based asynchronous lessons modeling counseling approaches, as well as interactive synchronous in-person group sessions to practice the skills learned. Finally, they asked about perceived feasibility, probing on potential barriers. Before the start of the IDIs, the interview guides were translated, back translated, and pilot tested with iterative refinement. Interviews were conducted over the phone or face-to-face and lasted approximately an hour long. Repeat interviews were not carried out. Interviewers summarized each IDI immediately after it was conducted. Audio recordings were transcribed verbatim in both Chichewa and translated to English by professional transcribers. Transcriptions were not returned to the participants.

Data Analysis

The team took a thematic approach to coding [31]. First, the team read the summaries to identify key themes and observed that they were framed by TAM. The research team then developed a qualitative codebook based on a focused examination of the transcripts, debrief reports, and interview summaries. The codebook consisted of both deductive codes from the interview guides and inductive codes that mapped onto TAM constructs. In addition, 2 trained research analysts (TEM and CJM) applied codes to transcript text using the Dedoose (SocioCultural Research Consultants) web application. As a first step, each analyst independently coded the same transcript and compared code applications to ensure agreements in the coding process. The analysts discussed and resolved any coding disparities, updating the codebook as needed to facilitate a shared understanding of all codes. The analysts then met weekly throughout the coding process, which provided space to review code applications and address any challenges. After coding was completed, analysts reviewed the coded text to extract key themes through the identification of patterns, naming and defining themes and refining the identified themes. Analysis was organized by factors that influence the TAM's constructs: perceived usefulness (what participants perceived as gained or lost by the blended learning implementation package) and perceived ease of use (what participants perceived as facilitators or barriers to acceptance of the blended learning implementation package). The findings were not shared with the participants for their feedback.

Ethical Considerations

Ethical clearance was provided by the University of North Carolina at Chapel Hill institutional review board (#20 - 1810), the Malawi National Health Sciences Research Committee (#20/06/2566) and the Baylor College of Medicine institutional review board (H-48800). Interviewers obtained written informed consent from all participants before starting the IDIs, reminding participants that their participation was voluntary and could be withdrawn at any time. Participants were compensated the equivalent of US $10 per interview. All data were deidentified to protect participant confidentiality.

Results

Demographic Summary

A total of 14 facilities participated in the IDIs. Among those, 2 of the facilities were district hospitals, 9 were health centers, and 3 were dispensaries. Furthermore, 9 facilities were in Machinga and 5 were in Balaka. In total, 26 HCWs participated in the IDIs. A total of 7 were female and 19 were male. Furthermore, 16 participants were from Machinga and 10 were from Balaka. There were no participants that refused to participate. All the participants were involved in both screening

XSL•FO
RenderX

and counselling index clients and conducting contact tracing and had previous exposure to standard ICT training.

Themes

Our analysis identified two sets of factors influencing perceived usefulness: (1a) opportunity for self-guided learning through digital portions and (1b) opportunity for social exchange through face-to-face interactions. We identified 5 factors influencing perceived ease of use. In addition, 2 factors; (2a) need for orientation to the digital technology and (2b) accessibility of devices were related to perceived ease of use to the technology itself, while 3 factors (2c) communication around training schedules, (2d) support for logistical arrangements, and (2e) monetary compensation were related to perceived ease of use of the training approach (Figure 2). The following sections develop these themes, and Table 1 shows the illustrative quotes.

Figure 2. External factors that contribute to perceived usefulness and perceived ease of use.

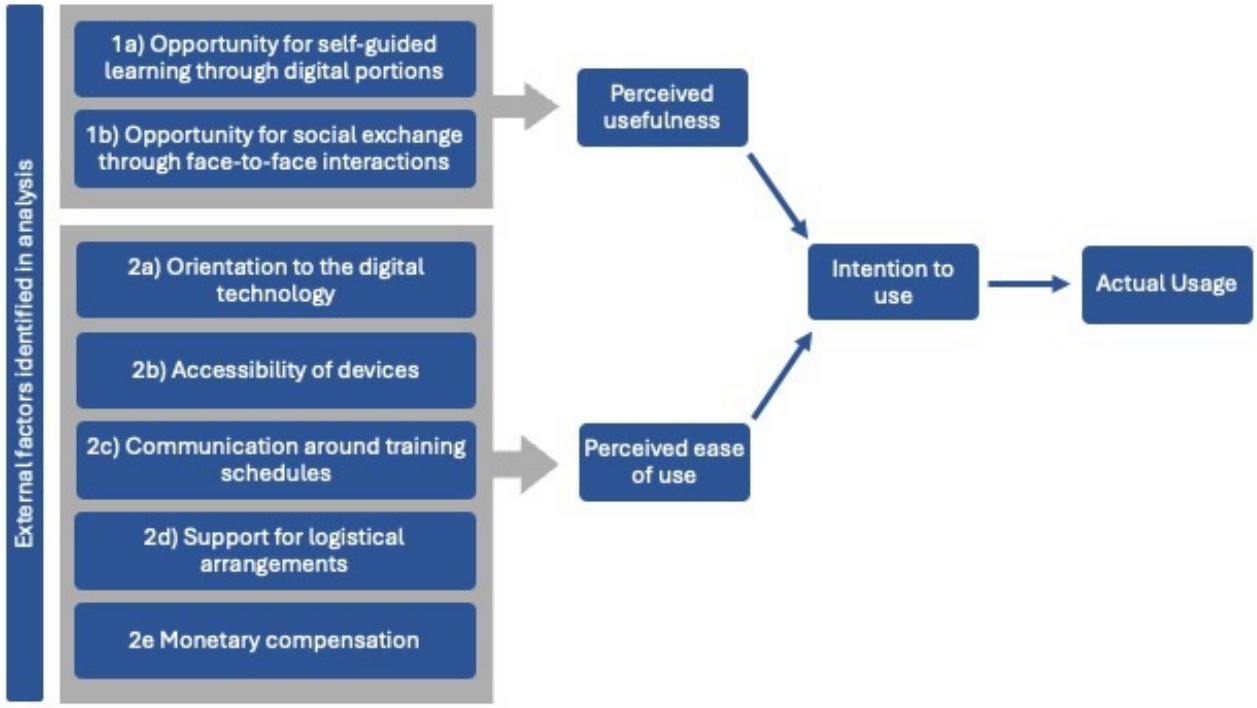

XSL•FO
RenderX

Table . Detailed explanations and illustrative quotations for each external factor.

External factors			Quote
1. Perceived usefulness			
	a. Opportunity for self-guided learning		
		Able to learn better.	• …. because I will be doing the things myself, I will be able to understand unlike having someone read to me or stand in front and say that this is this. [P 8021]
		Able to review the training content at their own pace	• You can repeat at your own time where you do not understand. [P 8007]
	b. Opportunity for interaction		
		Ability to ask questions and seek clarifications.	• ….the tablet cannot respond to questions that one may have as you do when you have a facilitator. [P 8003] • …. with the face-to-face portion if you have other questions maybe of which you failed to ask the tablet, … then you will ask the facilitator and he will. [P 8024]
		Ability to practice and receive feedback.	• … face to face, gives us … a chance to be practicing and see which areas we are doing well and areas in which we are lacking and help one another while we are together. [P 8013]
		Interacting with others	• … it means we will be able to meet people from different facilities and we will be able to share experiences. [P 8006]
2. Perceived ease of use			
	a. Need for orientation to the digital technology.		
		Need for orientation on use of tablet.	• There can be some people who have never used a tablet and before the training, they should be asked; “have you ever used a tablet?” If the answer is “no,” they should be taught how to use the tablet. [P 8011]
	b. Accessibility to the digital services		
		Questions and concerns regarding tablet availability	• ….if there would be a number of people with few tablets that will be a ‘challenge’ since you will be waiting for your friends to finish then wait for another one. [P 8023] • …if the networks are malfunctioning, it will definitely affect your work. [P 8012]

XSL•FO
RenderX

External factors			Quote
		Technical issues affecting usage and safety that could affect tablet access.	• …okay maybe at the time when you are using it [tablet] (…) maybe has run out of power or maybe that thing has a technical fault, [P 8007] • ….the tablet can get lost or maybe it fell down and it has broken. [P 8009]
	c. Communication around training schedules		
		Engaging in open communication about why certain individuals have been selected to attend tablet training.	• People can have thoughts if they can be seeing you working on a tablet and if the supervisor did not say that [name of participant] will be working on a tablet on this day. Your friends can think that you are just playing but if the supervisor can tell them they can know what is going on. [P 8004]
	d. Logistical arrangements		
		Location—Preference for the training location	• I would love if the training can be conducted in my catchment area and the other people would also want the training to be conducted in their catchment area. So, my opinion is that the training should be conducted at a convenient venue for most of the participants. [P 8006] • ….Just because' maybe 'you are far from' your home, 'you feel' that you are at a training… [P 8007]
		Time – perception of having the training over the weekend.	• So, if the training has happened on a weekend, there's no problem, it is your free time. [P 8012] • …we like using the weekend to…focus on our own things for our lives yeah so maybe many people would be affected if the training is done during the weekend. [P 8007]
	e. Monetary compensation		

XSL•FO
RenderX

External factors		Quote
	Need for allowance.	• On the allowances I can say that it would be good if we are given a little something so that we can be assisted. [P 8019] • …an allowance is an allowance. It needs to be there. [P 8022]
	Allowances motivate participants to concentrate.	• …this motivates people and even the concentration during the training is not compromised. [P 8006] • The benefit of receiving allowances is that the money helps you to care for yourself while you are doing the training, it can also help your family at the time when you are away from the family. It will help you to focus on the trainings and your home will not be disturbed due to lack of money. [P 8003]
	Need for equity when giving allowances.	• … So, the one working in NGO, because the training is hosted by the NGO, they only receive dinner allowance while those from government they receive dinner and daily subsistence allowance (DSA). [P 8005]
	Compensation is not the sole motivation	• I am not very concerned about the allowances because they just motivate us with the allowances…. what is paramount is the new knowledge that that we gain from the training. [P 8002]

Perceived Usefulness

Opportunity for Self-Guided Learning Through Digital Portions

Participants viewed the digital portion of blended learning as an opportunity for self-guided instruction. Some participants expected that a tablet-led training would allow them to learn ICT content better than facilitator-led training. This stemmed from a desire to review the training content at their own time and pace with no disturbances and to repeat difficult sections as needed. Participants expected they would keep the tablets at their facilities for future reference, an appealing feature.

Opportunity for Social Exchange Through Face-to-Face Interactions

Participants viewed opportunities to interact with facilitators and colleagues in-person as a central part of learning. Although participants were interested in individual learning, almost all also wished for social interaction during the learning process. Most participants were worried that working on the tablet would be like “talking to oneself” with no feedback mechanisms. They expected they would not get immediate feedback and clarifications. Participants viewed face-to-face sessions as essential for seeking clarification. In addition, participants felt face-to-face learning would allow them to practice and receive feedback. Practicing the skills with a facilitator and colleagues was considered useful because it would help them reinforce content and ultimately facilitate acquisition of knowledge and skills. It was also a time in which participants could receive support from their peers and facilitators on areas they needed to strengthen. In addition, participants considered the sharing of experiences with colleagues as integral.

Perceived Ease of Use

Need for Orientation to the Digital Technology

Orientation to the technology was highlighted as a key factor that would facilitate the ease of use of the blended learning implementation package. Conversely, some participants expressed fear about using a tablet for training without proper orientation due to not being well acquainted with the technology. To this end, participants suggested that they needed orientation on the use of the tablet as well as the applications that would

XSL•FO
RenderX

be used during the training since technological skills varied across participants.

Accessibility of Devices

Equipment access was a significant factor influencing perceived ease of use. Participants typically did not have personal tablets and assumed a tablet would be provided. Participants had several questions regarding the availability of the tablets within their facility, such as the duration when they would be available and whether they would be assigned to individuals or the facility. In addition, participants identified tablet storage, tablets running out of power, and poor network connectivity as key challenges that would affect tablet access. Participants feared that that the tablets were at risk of being lost at the facility, thus having appropriate tablet storage measures was suggested. Furthermore, participants were concerned that in seasons with frequent power blackouts, they would have challenges to charge the tablets if they ran out of power. Some participants stated that poor mobile network connectivity in some facilities could affect their ability to complete a tablet-based training.

Communication Around Training Schedules

Participants emphasized the need for open communication regarding training arrangements. They noted that training attendance can evoke envy among colleagues. To address this, participants suggested ensuring prior communication to all health care workers about the selection process and purpose of training, especially for tablet training, to minimize work disruption and misunderstandings.

Support for Logistical Arrangements

Time and location for both portions of the training were highlighted as key logistical considerations that would drive ease of use. Most participants considered having digital training at their place of work or a nearby location acceptable. They also were open to having portions of the training over the weekend or during workdays, with pros and cons for each of these different locations and times. Weekend training was considered desirable by some as it would not disrupt work schedules, but others noted it infringed on their personal commitments. Some participants explained that having the training on-site involved less effort for travel and would not disrupt the work schedule at the facility. However, to some individuals, having the training offsite was important because it created a mindset of higher excitement, motivation, and a sense of purpose for participants. They also noted that participants in offsite training would experience fewer distractions than in a work environment, allowing participants to fully focus on the training content.

Monetary Compensation

Provision of a training allowance (ie, monetary compensation) was noted as a contributor to ease of use of a blended training. The norm in the Malawian context is to receive an allowance to cover accommodation and meals when attending a centralized face-to-face training. Participants were asked to share their views regarding such allowances under a decentralized blended learning approach. The general perception was that they would still expect to receive an allowance. In fact, some participants indicated that they preferred offsite trainings, specifically because of the certainty of an allowance. They described how allowances act as a motivating factor that allows people to focus on training content and apply what they learn. In addition, allowances were perceived as important because they support the basic expenses for their household needs without detracting from their family's financial commitments while they were in the training. Participants preferred to receive a monetary allowance, rather than meals and accommodation, because it provided them with the possibility to save some of the money for other purposes. Participants voiced a desire for allowances to be disbursed equally across organization type, distance travelled, and rank. Although most participants noted allowances as important, many emphasized that this was not their sole ore even primary motivation for training.

Discussion

This study explored the anticipated acceptability of blended learning among lay HCWs in Malawi. Interest in self-guided learning was a key factor underlying perceived usefulness in digital learning. The desire for interaction was an important consideration for in-person learning. With respect to ease of use, blended learning could be facilitated through orientation, accessibility of digital devices, communication, and consideration of training logistics and compensation.

This study is among the first to examine perceived acceptability of a blended learning implementation package within the context of in-service training among lay HCWs in LMICs. Our study suggests that the use of technology could be acceptable among HCWs in this setting despite low digital literacy, as long as there is sufficient orientation and device access. The study revealed positive perspectives from HCWs on the use of the digital component as it would allow participants to learn the training content better at their own time, place, and pace. Consistently, studies that have used digital technology in tertiary education throughout health professions have reported positive perspectives [32-35], due largely to the ability to review content [16]. We posit that if blended learning proves acceptable in a country with one of the lowest digital literacy levels and many technology barriers, it can also be implemented in areas with higher digital literacy and fewer barriers.

Our findings support previous findings that suggest learners do not want to completely forego human interactions, which they perceive enhance learning [16,17,36-38]. While most participants welcomed the idea of digital ICT training, the lack of interaction was highlighted as a major hindrance to the use of technology alone. Hence, participants preferred to also have a face-to-face portion to resolve their unanswered questions and allow them to practice new skills. The desire for face-to-face encounters has also been highlighted in clinical training programs [32,39]. These findings around interest in both modalities reinforce the value of blended approaches [40] in which participants experience the flexibility of digital self-training and engagement and interactivity in face-to-face sessions [41]. We incorporated these findings into our package by having a set of digital modules focused on reviewing content, followed by face-to-face modules focused on practicing and receiving feedback.

XSL•FO
RenderX

Orientation was perceived as a key factor to facilitate the ease of use of digital technologies, such as tablets in this setting. Participants came from a setting with low technological literacy. According to the World Bank, mobile phone penetration in Malawi was at only at 60% in 2021, suggesting that digital skills are not as ubiquitous as they are in other settings. Therefore, for a blended learning implementation package to be successful, participants need basic navigation skills before initiation [14], as well as technical support to improve ease of use [42]. We incorporated these findings into our package training site supervisors on how to use the blended learning package. The site supervisors, in turn, were empowered to orient staff at their site and support those who experienced technology challenges.

Participants had different preferences regarding the time and location of the training, but they unanimously agreed on the desire for allowances, even with this new training modality. In addition, there was consensus around the importance of equity when distributing allowances. Allowances were felt to serve as motivating factors that allow participants to focus on training content. Such allowances are often provided to motivate participants [43-45]. However, an appeal of the blended learning approach is its potential to expand the skill set of HCWs in a more economical and less disruptive manner. By offering tablet-based training at or near their workplace, HCWs spend less time away from their other duties, and the overall costs could be reduced if no lodging or transportation expenses were incurred. Our study team took these findings under consideration in the development of a blended learning package that is now being tested in a randomized trial. In the trial, to address equity considerations, we offered the training to all lay health workers at each facility and provided a small allowance to recognize participants' time.

Our study has several limitations and next steps. The study was conducted before HCWs were exposed to blended learning and evaluated anticipated perceptions; actual perceptions may differ. This is something we will evaluate when we implement our trial. In addition, the study was conducted in a few facilities in 2 districts and perceptions may differ in other settings. However, we sampled from a broad range of districts and facility types and achieved saturation. These facilities are not meaningfully different from other facilities in these districts or the country overall, suggesting that our results may generalize broadly in Malawi.

In summary, our formative study showed that a blended learning implementation package may be an acceptable method of training HCWs in the Malawian setting if program developers attend to contextual factors that affect perceived usefulness and ease of use. Although this work offered immediate insights for our planned trial, we recommend additional research to assess acceptability of blended learning approaches in a range of LMIC settings among lay health workers.

Authors' Contributions

NER, TEM-M, TAT, MHK, VG, VM, SER, and KRS contributed to conceptualization. CK, TAT, MJC, and MM assisted with data collection. TEM-M, CJM, and MM assisted with the analysis. TEM-M, CJM, NER, TAT, MJC, MP, MJ-B, and DV contributed to the drafting of the paper and figures. All authors critically reviewed, edited, and approved the final paper.

Conflicts of Interest

None declared.

References

1. Malawi Ministry of Health. Malawi final report 2020-2021. 2022 Dec 1 URL: https://phia.icap.columbia.edu/malawi-final-report-2020-2021/ [accessed 2025-03-12]
2. World Health Organization. Guidelines on HIV Self-Testing and Partner Notification: Supplement to Consolidated Guidelines on HIV Testing Services: World Health Organization; 2016.
3. Rosenberg NE, Mtande TK, Saidi F, et al. Recruiting male partners for couple HIV testing and counselling in Malawi's option B+ programme: an unblinded randomised controlled trial. Lancet HIV 2015 Nov;2(11):e483-e491. [doi: 10.1016/S2352-3018(15)00182-4] [Medline: 26520928]
4. Kahabuka C, Plotkin M, Christensen A, et al. Addressing the First 90: a highly effective partner notification approach reaches previously undiagnosed sexual partners in Tanzania. AIDS Behav 2017 Aug;21(8):2551-2560. [doi: 10.1007/s10461-017-1750-5] [Medline: 28299518]
5. Tembo TA, Kim MH, Simon KR, et al. Enhancing an HIV index case testing passive referral model through a behavioural skills-building training for healthcare providers: a pre-/post-assessment in Mangochi District, Malawi. J Int AIDS Soc 2019 Jul;22 Suppl 3(Suppl Suppl 3):e25292. [doi: 10.1002/jia2.25292] [Medline: 31321917]
6. Picciano AG. Blended learning: implications for growth and access. OLJ 2006;10(3):95-102. [doi: 10.24059/olj.v10i3.1758]
7. Graham CR, Woodfield W, Harrison JB. A framework for institutional adoption and implementation of blended learning in higher education. Internet High Educ 2013 Jul;18:4-14. [doi: 10.1016/j.iheduc.2012.09.003]
8. Sakina R, Kulsum EM, Uyun AS. Integrating technologies in the new normal: a study of blended learning. ijqrm 2020;1(4):181-193 [FREE Full text] [doi: 10.46336/ijqrm.v1i4.81]
9. Means B, Toyama Y, Murphy R, Baki M. The effectiveness of online and blended learning: a meta-analysis of the empirical literature. Teachers College Record: The Voice of Scholarship in Education 2013 Mar;115(3):1-47. [doi: 10.1177/016146811311500307]

XSL•FO
RenderX

10. Karamizadeh Z, Zarifsanayei N, Faghihi AA, Mohammadi H, Habibi M. The study of effectiveness of blended learning approach for medical training courses. Iran Red Crescent Med J 2012 Jan;14(1):41-44. [Medline: 22737553]
11. Singh H. Building effective blended learning programs. In: Challenges and Opportunities for the Global Implementation of E-Learning Frameworks: IGI Global; 2021:15-23. [doi: 10.4018/978-1-7998-7607-6.ch002]
12. Skulmowski A, Rey GD. COVID-19 as an accelerator for digitalization at a German university: Establishing hybrid campuses in times of crisis. Hum Behav Emerg Technol 2020 Jul;2(3):212-216. [doi: 10.1002/hbe2.201] [Medline: 32838228]
13. Finlay MJ, Tinnion DJ, Simpson T. A virtual versus blended learning approach to higher education during the COVID-19 pandemic: The experiences of A sport and exercise science student cohort. J Hosp Leis Sport Tour Educ 2022 Jun;30:100363. [doi: 10.1016/j.jhlste.2021.100363] [Medline: 34867086]
14. Batubara HS, Riyanda AR, Rahmawati R, Ambiyar A, Samala AD. Implementasi model pembelajaran blended learning di masa pandemi Covid-19: meta-analisis [Article in Indonesian]. basicedu 2022;6(3):4629-4637 [FREE Full text] [doi: 10.31004/basicedu.v6i3.2816]
15. Mali D, Lim H. How do students perceive face-to-face/blended learning as a result of the Covid-19 pandemic? The International Journal of Management Education 2021 Nov;19(3):100552. [doi: 10.1016/j.ijme.2021.100552]
16. Baber H. Social interaction and effectiveness of the online learning – A moderating role of maintaining social distance during the pandemic COVID-19. AEDS 2022 Jan 3;11(1):159-171. [doi: 10.1108/AEDS-09-2020-0209]
17. Nacaroğlu O, Kızkapan O, Demir H. Middle school students' motivations and learning competencies in science: mediating role of digital literacy. Psychol Sch 2025. [doi: 10.1002/pits.23400]
18. Islam S, Baharun H, Muali C, et al. To boost students' motivation and achievement through blended learning. J Phys: Conf Ser 2018 Nov 1;1114:012046. [doi: 10.1088/1742-6596/1114/1/012046]
19. Ramachandran R, Kaur R, Roy A, Goel P, Deorari AK. Perceived usefulness of a blended learning approach for skills training among medical interns: a pilot study. BMC Med Educ 2024 Nov 25;24(1):1357. [doi: 10.1186/s12909-024-06379-x] [Medline: 39587590]
20. World Health Organization. WHO Guideline: Recommendations on Digital Interventions for Health System Strengthening: Web Supplement 2: Summary of Findings and GRADE Tables: World Health Organization; 2019.
21. Kenney J, Newcombe E. Adopting a blended learning approach: challenges encountered and lessons learned in an action research study. OLJ 2011;15(1):45-57. [doi: 10.24059/olj.v15i1.182]
22. Tembo TA, Mollan K, Simon K, et al. Does A blended learning implementation package enhance HIV index case testing in Malawi? A protocol for A cluster randomised controlled trial. BMJ Open 2024 Jan 22;14(1):e077706. [doi: 10.1136/bmjopen-2023-077706] [Medline: 38253452]
23. Hamilton AB, Finley EP. Qualitative methods in implementation research: An introduction. Psychiatry Res 2019 Oct;280:112516. [doi: 10.1016/j.psychres.2019.112516] [Medline: 31437661]
24. Elsey H, Khanal S, Manandhar S, et al. Understanding implementation and feasibility of tobacco cessation in routine primary care in Nepal: a mixed methods study. Implementation Sci 2015 Dec;11(1):1-12. [doi: 10.1186/s13012-016-0466-7]
25. Marangunić N, Granić A. Technology acceptance model: a literature review from 1986 to 2013. Univ Access Inf Soc 2015 Mar;14(1):81-95. [doi: 10.1007/s10209-014-0348-1]
26. Wallace LG, Sheetz SD. The adoption of software measures: A technology acceptance model (TAM) perspective. Information & Management 2014 Mar;51(2):249-259. [doi: 10.1016/j.im.2013.12.003]
27. Medical doctors (per 10,000).: World Health Organization; 2020. URL: https://www.who.int/data/gho/data/indicators/indicator-details/GHO/medical-doctors-(per-10-000-population) [accessed 2023-04-12]
28. Flick RJ, Simon KR, Nyirenda R, et al. The HIV diagnostic assistant: early findings from a novel HIV testing cadre in Malawi. AIDS 2019 Jun 1;33(7):1215-1224. [doi: 10.1097/QAD.0000000000002159] [Medline: 31045942]
29. Network readiness index 2023-malawi. : Portulans Institute; 2023.
30. Clarke V, Braun V, Hayfield N. Thematic analysis. In: Qualitative Psychology: A Practical Guide to Research Methods: Sage Publications; 2015, Vol. 3:222-248.
31. Kelly RF, Mihm-Carmichael M, Hammond JA. Students' engagement in and perceptions of blended learning in a clinical module in a veterinary degree program. J Vet Med Educ 2021 Apr;48(2):181-195. [doi: 10.3138/jvme.2019-0018] [Medline: 31873057]
32. Bhat S, Madi˙ M. Blended learning in undergraduate dental education. Cumhuriyet Dent J 2022;25(1):91-93. [doi: 10.7126/cumudj.968711]
33. Nagaraj C, Yadurappa SB, Anantharaman LT, Ravindranath Y, Shankar N. Effectiveness of blended learning in radiological anatomy for first year undergraduate medical students. Surg Radiol Anat 2021 Apr;43(4):489-496. [doi: 10.1007/s00276-020-02572-x] [Medline: 32964270]
34. Bouarar AC, Mouloudj S, Umar TP, Mouloudj K. Antecedents of physicians' intentions to engage in digital volunteering work: an extended technology acceptance model (TAM) approach. JICA 2023 Dec 6;31(4):285-299. [doi: 10.1108/JICA-03-2023-0017]
35. Jiang W. Interdependence of roles, role rotation, and sense of community in an online course. Distance Education 2017 Jan 2;38(1):84-105. [doi: 10.1080/01587919.2017.1299564]

XSL•FO
RenderX

36. Wu JH, Tennyson RD, Hsia TL. A study of student satisfaction in A blended e-learning system environment. Comput Educ 2010 Aug;55(1):155-164. [doi: 10.1016/j.compedu.2009.12.012]
37. Jusoff K, Khodabandelou R. Preliminary study on the role of social presence in blended learning environment in higher education. IES 2009;2(4):79-83. [doi: 10.5539/ies.v2n4p79]
38. Koch J, Ramjan LM, Everett B, Maceri A, Bell K, Salamonson Y. “Sage on the stage or guide on the side”-Undergraduate nursing students’ experiences and expectations of bioscience tutors in A blended learning curriculum: A qualitative study. J Clin Nurs 2020 Mar;29(5-6):863-871. [doi: 10.1111/jocn.15140] [Medline: 31855301]
39. Suwannaphisit S, Anusitviwat C, Tuntarattanapong P, Chuaychoosakoon C. Comparing the effectiveness of blended learning and traditional learning in an orthopedics course. Ann Med Surg (Lond) 2021 Dec;72:103037. [doi: 10.1016/j.amsu.2021.103037] [Medline: 34815858]
40. Kemp N, Grieve R. Face-to-face or face-to-screen? Undergraduates’ opinions and test performance in classroom vs. online learning. Front Psychol 2014;5:1278. [doi: 10.3389/fpsyg.2014.01278] [Medline: 25429276]
41. Kassably C, Ashaal A, Tahan S, Bouzakhem N, Youssef S, Ghosn F. Identification of factors influencing educators perception of blended learning and triggering adoption decisions. Migrat Lett 2024;21(4):126-144. [doi: 10.59670/ml.v21iS6.7217]
42. Nkamleu GB, Kamgnia BD. Uses and abuses of per-diems in africa: a political economy of travel allowances. : African Development Bank Group; 2014.
43. Vian T, Miller C, Themba Z, Bukuluki P. Perceptions of per diems in the health sector: evidence and implications. Health Policy Plan 2013 May;28(3):237-246. [doi: 10.1093/heapol/czs056] [Medline: 22684639]
44. Samb OM, Essombe C, Ridde V. Meeting the challenges posed by per diem in development projects in southern countries: a scoping review. Global Health 2020 May 28;16(1):48. [doi: 10.1186/s12992-020-00571-6] [Medline: 32466774]
45. Derntl M, Motschnig-Pitrik R. Patterns for blended, person-centered learning: strategy, concepts, experiences, and evaluation. Presented at: Proceedings of the 2004 ACM Symposium on Applied Computing; Mar 14-17, 2004; Nicosia, Cyprus. [doi: 10.1145/967900.968087]

Abbreviations

HCW: health care worker
ICT: index case testing
IDI: in-depth interview
LMIC: low- and middle-income country
TAM: Technology Acceptance Model
UNAIDS: Joint United Nations Programme on HIV and AIDS
WHO: World Health Organization

Edited by A Mavragani; submitted 30.05.24; peer-reviewed by K Mouloudj; revised version received 20.02.25; accepted 24.02.25; published 07.04.25.

Please cite as:
Mbeya-Munkhondya TE, Meek CJ, Mphande M, Tembo TA, Chitani MJ, Jean-Baptiste M, Kumbuyo C, Vansia D, Simon KR, Rutstein SE, Mwapasa V, Go V, Kim MH, Rosenberg NE
Anticipated Acceptability of Blended Learning Among Lay Health Care Workers in Malawi: Qualitative Analysis Guided by the Technology Acceptance Model
JMIR Form Res 2025;9:e62741
URL: https://formative.jmir.org/2025/1/e62741
doi:10.2196/62741

© Tiwonge E Mbeya-Munkhondya, Caroline J Meek, Mtisunge Mphande, Tapiwa A Tembo, Mike J Chitani, Milenka Jean-Baptiste, Caroline Kumbayo, Dhrutika Vansia, Katherine R Simon, Sarah E Rutstein, Victor Mwapasa, Vivian Go, Maria H Kim, Nora E Rosenberg. Originally published in JMIR Formative Research (https://formative.jmir.org), 7.4.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Impact of a Human Papillomavirus Facebook-Based Intervention (#HPVVaxTalks) Among Young Black (African American and Sub-Saharan African Immigrants) Adults: Pilot Pre- and Poststudy

Adebola Adegboyega[1], PhD; Amanda Wiggins[1], PhD; Abubakari Wuni[1], MS; Melinda Ickes[2], PhD

[1]College of Nursing, University of Kentucky, Lexington, KY, United States

[2]Department of Kinesiology and Health Promotion, College of Education, University of Kentucky, Lexington, KY, United States

Corresponding Author:
Adebola Adegboyega, PhD
College of Nursing, University of Kentucky, Lexington, KY, United States

Abstract

Background: Despite the availability of prophylactic human papillomavirus (HPV) vaccines, uptake remains suboptimal among young Black adults. Social media is a platform for the dissemination of health information and can be used to promote HPV vaccination among young Black adults.

Objective: This study aimed to assess the impact of a Facebook-based intervention (#HPVVaxTalks), which consisted of 40 posts over 8 weeks in improving cognitive outcomes, reducing vaccine hesitancy, and increasing vaccine intention, and uptake among young Black adults aged 18-26 years.

Methods: A pilot 1-group pre- and poststudy was conducted among 43 young Black adults who engaged in an 8-week Facebook intervention (#HPVVaxTalks). #HPVVaxTalks was developed in collaboration with a youth community advisory committee. Participants were actively recruited by research staff from community settings using flyers, and flyers were posted in public places in communities. Eligible participants were screened for eligibility and consented prior to study participation. Participants completed baseline surveys and were added to a Facebook page created for the study to receive intervention posts. Participants completed pre- and postdata on HPV knowledge, HPV vaccine knowledge, vaccine hesitancy, and vaccine uptake via REDCap (Research Electronic Data Capture) surveys distributed by email. Participants' satisfaction with the intervention was collected via individual interviews. Data were analyzed using 2-tailed paired *t* tests and repeated measures analysis

Results: Overall, 32 of the 43 (74%) participants completed the follow-up survey, and of the 23 participants who reported not having ever received the vaccine at baseline, 7 (30%) reported receiving the vaccine at follow-up. Participants demonstrated significant improvements in HPV knowledge and receiving the vaccine at follow-up. Participants demonstrated significant improvements in HPV knowledge (pre: mean 7.3, SD 4.2 and post: mean 11.1, SD 4.3; P=.004) and HPV vaccine knowledge (pre: mean 2.8, SD 2.5 and post: mean 4.7, SD 2.2; P=.003) and reduction in vaccine hesitancy (pre: mean 28.3, SD 4.2 and post: mean 29.9, SD 3.6; P=.007) after the intervention. However, there were no significant changes in other outcomes. Feedback from open-ended questions and qualitative interviews highlighted participants' satisfaction with the intervention and its role in increasing HPV and HPV vaccine awareness.

Conclusions: The findings from this study underscore the potential of social media platforms for health promotion among underrepresented populations and the importance of advocating for culturally appropriate interventions to improve HPV vaccination rates and reduce disparities.

(JMIR Form Res 2025;9:e69609) doi:10.2196/69609

KEYWORDS

human papillomavirus vaccination; HPV vaccination; Facebook-based intervention; young Black adults; social media

Introduction

Background

Human papillomavirus (HPV) vaccination provides primary prevention against HPV-related diseases [1,2]. However, HPV vaccine uptake remains low for female participants and lower for male participants. Within 12 years of HPV vaccine introduction, infections with the 4 HPV types prevented by Gardasil decreased to 88% among 14‑ to 19-year-old female participants and to 81% among 20‑ to 24-year-old female participants in the United States [3]. According to the Centers

XSL•FO
RenderX

for Disease Control and Prevention, only 21.5% of adults aged 18 - 26 years received the recommended number of doses of the HPV vaccine [4]. Completion rates for the HPV vaccine are low in Black female participants (44.7% vs 57.9% White female participants) and lower among Black male participants (29.4% vs 26.6% among White male participants) aged 18 - 26 years [5]. HPV vaccination is recommended for everyone through age 26 years if not adequately vaccinated when younger. Hence, Black young adults are a priority population to improve HPV vaccine uptake, given the low HPV vaccine uptake among this group [6,7].

In general, one of the most common barriers to HPV vaccination uptake is the lack of awareness of HPV and HPV infections. Research shows that sub-Saharan African immigrants have even less knowledge of HPV and HPV vaccination compared to White and Black American–born individuals [8-10]. The lack of knowledge results in misconceptions and the thinking that the HPV vaccine is unnecessary because individuals do not perceive themselves at risk of HPV infection [11]. In addition, the potential cost of the HPV vaccine series completion, perceptions of vaccine safety, and worry regarding anticipated side effects are barriers to the uptake of vaccine uptake for young adults [12]. In addition, system-level barriers such as limited access to vaccines, vaccine documentation requirements, and lack of strong provider recommendations may lead to missed vaccination opportunities [13-15]. Additionally, the African American population has been participants of historically unethical research, which may contribute to low vaccine uptake, medical mistrust, and misconceptions about vaccines [8,16,17].

Despite the high prevalence and complications that can occur from unresolved HPV infections, gaps in HPV-related cognition and HPV-related disparities exist for Black individuals. Catch-up HPV vaccination is recommended for young adults (up through age 26 years) who have not been previously vaccinated. However, few interventions have focused on Black young adults (18 - 26 years), much of the research has explored parental decision-making and system-level interventions for HPV vaccination among adolescents [18-22]. Transition to young adulthood is a critical period to institute health-promoting behaviors [23]. As young adults transition to adulthood, they begin to explore their sexuality, including initiating sexual activity for the first time or engaging in riskier sexual behaviors than in the past. These behaviors increase the risk of acquiring sexually transmitted infections including HPV, providing a strong justification for health-promoting behavior such as catch-up HPV vaccination [24]. As Black young adults assume health care decision-making responsibility and independence, they may benefit from interventions to promote cognitive outcomes (HPV and HPV vaccine-related knowledge, benefits, severity, self-efficacy, risk perception), vaccine intention, and catch-up HPV vaccination.

Social media is increasingly becoming the preferred platform for the dissemination of health information and can be used to raise awareness about HPV and HPV vaccinations, therefore increasing HPV vaccination rates [25]. Social media platforms are multifaceted, and their features can be used to serve multiple purposes and reach a wide audience. Social media platforms include video and photo posting, blogs, polls, and other functions [26,27]. In the United States, Facebook is the single most popular social network with over 2 billion daily users, and 73% of Black adults have ever used Facebook [28,29]. In addition, 83% of adult social media users report visiting Facebook in a typical week, and on average, young adults aged 18 to 24 years spend about 11 minutes every day on Facebook [30].

Recent studies have used Facebook-based interventions to promote HPV vaccination among young adults [31-33]. For example, a technology-mediated intervention including private Facebook posts improved knowledge among undergraduate college students in the HPV vaccination intervention relative to those in the control condition [32]. Similarly, a sponsored advertisement campaign on Facebook targeted at young female participants aged 18 - 26 years found that participants engaged with and were receptive to watching narrative-based health information videos on Facebook. Higher video engagement was associated with stronger intentions to talk with a health care professional, talk with friends or family, and vaccinate against HPV [31]. Despite the growing popularity of Facebook-based interventions, few have targeted young Black adults including sub-Saharan African immigrants. To fill this gap, we sought to leverage Facebook for an HPV vaccination intervention among young Black adults—a priority catch-up population with the highest burden of HPV. Thus, the objective of this study was to examine the impact of a Facebook-based intervention (#HPVVaxTalks) on HPV vaccination uptake and secondary outcomes among Black (African American and sub-Saharan African immigrants) young adults (18 - 26 years) of an 8-week Facebook-based intervention. We hypothesize that young Black adults who complete the intervention will have improvements in HPV knowledge and HPV vaccine knowledge and a reduction in vaccine hesitancy after the intervention.

Theoretical Framework

The theory of planned behavior states that the intention to perform a behavior is the strongest predictor of whether an individual engages in certain behaviors [34]. The theory of planned behavior has been used to understand and predict HPV vaccination rates by examining what constructs most correlate with people's intentions to receive HPV vaccine. For example, in a sample of college students, more positive attitudes and increased subjective norms were strongly correlated with increased intention to receive complete doses of the HPV vaccine [35]. Knowledge and health beliefs are antecedents of intention and subsequent behavior. Hence, the theory of planned behavior provides a framework to understand individual actions by identifying beliefs relevant to individuals and groups and to develop interventions to change beliefs [34].

Methods

Study Designs and Participants

This was a 1-group pre- and post-Facebook–based intervention pilot study conducted in a southeastern state. We used a mixed methods approach by administering a pre- and postintervention survey to the participants as well as conducting follow-up interviews with a subset of participants.

XSL•FO
RenderX

Participants for the study were young Black (African American and sub-Saharan African immigrants) adults between 18 and 26 years, given the age for HPV vaccination catch-up through age 26 years. We used the STROBE (Strengthening the Reporting of Observation Studies in Epidemiology) checklist to guide the reporting of the study (Checklist 1).

Intervention

A culturally appropriate Facebook-based intervention (#HPVVaxTalks) was developed in collaboration with a youth community advisory committee [36,37]. The research team created a library of evidenced-based Facebook posts, and a youth community advisory board reviewed the preliminary version of #HPVVaxTalks intervention materials and provided critiques and suggestions for refinement. The advisory board provided recommendations and input on different iterations of the posts for content preferences, priorities, user-friendliness, relevance, and cultural appropriateness [36]. To ensure posts were culturally appropriate, we followed strategies from Kreuter et al [38] including evidential, peripheral, linguistic, constituent-involving, and sociocultural strategies. Engagement with a youth community advisory board fostered the opportunity to discuss priorities reflecting the target group's interest, ensure the appeal and relevance of the intervention, and draw on the experience of the target populations. In addition, we conducted a content validation of the intervention by obtaining experts' feedback and participants' satisfaction on the components of the #HPVVaxTalks for content usability, acceptability, relevance, and visual appeal to inform further refinements [37]. The intervention addressed general information on HPV, the importance of HPV vaccination, vaccine hesitancy, self-efficacy, HPV vaccination as a cancer prevention strategy, risk perception, perceived behavioral control, and subjective norm. Intervention contents were mapped to constructs of the theory of planned behavior, and important concepts informed by the literature and youth advisory board. The Facebook posts contained relevant images, graphics, memes, short texts, and information on where to receive the HPV vaccine in the community (see Figures 1 and 2 for sample posts). We included embedded links to information from lead organizations (eg, American Cancer Society, American Sexual Health Association, and Centers for Disease Control and Prevention). The intervention consisted of 40 messages posted over 8 weeks, with 4 to 5 messages posted each week. One "secret" and "closed" Facebook group page solely for this study was created. Facebook was chosen as the platform for delivering HPV vaccination messages due to its features that allow for the creation of private groups. These groups enable only invited participants to see posts and engage in discussions, ensuring a level of privacy and confidentiality that can encourage open discussion [33]. Eligible participants were screened and consented, they received a link to complete the baseline survey and were invited to join #HPVVaxTalks Facebook group. Trained research staff posted intervention posts as scheduled (4 to 5 messages posted each week for 8 weeks) to promote engagement but not burden the participants with too much information. Enrolled participants were encouraged to turn on Facebook notification to know when posts are posted, view posts regularly, comment, like, and ask questions. The Facebook page was monitored throughout the intervention by the research team.

Figure 1. This is a sample of Facebook graphic posted as one of the messages for HPV vaccination promotion among young Black adults. Courtesy of National Foundation for Infectious Diseases [39].

XSL•FO
RenderX

Figure 2. This is a sample of Facebook graphic posted as one of the messages for HPV vaccination promotion among young Black adults [39]. HPV: human papillomavirus.

Recruitments

Black young adults aged 18 - 26 years were recruited throughout a southeastern state between April 2023 and June 2024 using approved study flyers, snowballing, and word of mouth. In addition, research staff carried out active recruitment throughout community organizations that have a high client base of young adults. Study flyers with research team contact information were posted in public areas and community settings throughout the state. Flyers described the project and asked interested participants to contact the study team to learn more about the study. Interested participants were provided with additional information about the study and screened for eligibility. Study eligibility included the following: self-identification as African American and sub-Saharan African immigrants, age 18 - 26 years, fluency in the English language, access to the internet, a Facebook account, and willingness to participate.

Ethical Considerations

Prior to the start of the study, all study procedures and materials were reviewed and approved by the University of Kentucky Institutional Review Board (protocol 76589). Participation was voluntary, and all study participants provided written informed consent prior to participation in any study activities. Confidentiality of participant's information was maintained at all times. Patient's data were coded with a unique subject identifier, and all data deidentified prior to analysis. Participants were compensated with US $30 gift card from a retail store for completing the baseline survey and US $20 for postintervention surveys.

Quantitative Data Collection

All research activities were completed remotely. Participants received REDCap (Research Electronic Data Capture; Vanderbilt University) links via emails to complete baseline and postintervention surveys. Quantitative data were collected using self-administered REDCap-based surveys at baseline and after the intervention (after 8 weeks). After the intervention, we repeated the items from the baseline survey and added open-ended questions to assess what participants liked about the intervention and ways for improvement. In addition, we assessed intervention satisfaction and asked if participants had received a dose of the HPV vaccination.

Measures

Overview

Participants completed a self-administered survey including sociodemographic characteristics, HPV vaccination history, HPV knowledge items, HPV vaccine knowledge items, vaccine hesitancy measures, risk perception items, HPV severity, HPV vaccination benefits, and HPV vaccination self-efficacy. Valid and reliable measures were adapted from the literature [40-43].

XSL•FO
RenderX

Sociodemographic Characteristics

Sociodemographic information collected via self-report included age, sex, country of origin, education level, income, and health insurance coverage.

HPV Vaccination Uptake

HPV vaccination uptake was assessed among those who reported not being vaccinated at baseline. In the follow-up survey, participants were asked if they had received at least 1 dose of the HPV vaccine since enrolling in the study (yes or no).

HPV Knowledge

HPV knowledge items included 16 items assessing knowledge of HPV (including transmission, consequences, and risk factors). Response options were true or false or do not know, with "do not know" scored as incorrect [41,42]. A summative score was calculated with a potential range of 0 - 16. The scale has high reliability (Cronbach α=0.84) among diverse populations [42], and Cronbach α for this study was 0.91.

HPV Vaccine Knowledge

HPV vaccine knowledge was assessed with an 8-item scale regarding the protection offered by HPV vaccines [41]. A summative score was calculated with a potential range of 0 - 8. Cronbach α for this sample was 0.84.

Vaccine Hesitancy

Vaccine hesitancy was measured with a 9-item scale that captures general attitudes to vaccination [43], sample items include "the HPV vaccine is effective" and "the HPV vaccine is important for my health." Responses options followed a 4-point Likert scale ranging from 1=strongly disagree to 4=strongly agree [43]. A summative score was calculated with a potential range of 9 - 36, with lower scores indicating greater hesitancy. Cronbach α for this sample was 0.80.

HPV-Related Risk Perception

HPV-related risk perception was assessed with the item, "If I don't get vaccinated for HPV, I think my chances of getting [HPV infection; HPV-related cancer; genital warts] sometime in the future would be," with responses of 1=almost zero, 2=small, 3=moderate, 4=large, and 5=almost certain [40].

HPV Severity, Benefits, and Self-Efficacy

These constructs were assessed with items adapted from the literature [40]. HPV severity was assessed with 4 items (eg, being infected with HPV would have major consequences on my life), HPV benefits were assessed with 3 items (eg, getting vaccinated for HPV will help protect me from HPV infection), and self-efficacy was assessed with 3 items (eg, I feel confident in my ability to get vaccinated for HPV, even if it is expensive). Responses options for each item followed a 5-point Likert scale ranging from 1=strongly disagree to 5=strongly agree. Summative scores were calculated for each construct, with potential ranges of 4 - 20, 3 - 15, and 3 - 15 for severity, benefits, and self-efficacy, respectively. Cronbach α for all scales exceeded 0.77.

HPV Vaccine Intention

Vaccine intention was assessed with an item, "I intend to get the HPV vaccine in the next 3 months?" Responses followed a 5-point Likert scale ranging from 1=strongly disagree to 5=strongly agree.

Intervention Satisfaction

Intervention satisfaction was assessed with items to measure usability (6 items; Cronbach α=0.93), acceptability (5 items; Cronbach α=0.92), and relevance of HPV posts or messages on Facebook (5 items; Cronbach α=0.92). Responses to each item followed a 5-point Likert scale ranging from 1=strongly disagree to 5=strongly agree, and summative scores were calculated for each construct separately [44,45].

Data Analysis: Quantitative Data Analysis

Sociodemographic characteristics among those who completed the follow-up survey (completers) and those who did not (noncompleters) were examined using the 2-tailed *t* test, chi-square test of association, or Mann-Whitney *U* test, as appropriate.

Descriptive statistics were used to analyze vaccination uptake. Among those who had not received the HPV vaccine at baseline, unadjusted changes in outcomes (HPV-related knowledge, HPV vaccine knowledge, vaccine hesitancy, risk perception, severity, vaccination benefits, self-efficacy, and intention and vaccination uptake) were evaluated using the paired 2-tailed *t* test. These changes were also evaluated using repeated measures analysis, adjusting for age, sex, and ethnicity. The repeated measures analysis is exploratory due to the small size. Intervention satisfaction, specifically usability, acceptability, and relevance of the messages, was summarized descriptively using means and SDs. All analysis was conducted using SAS (version 9.4; SAS Institute Inc), with an α level of .05 throughout.

Results

Quantitative Findings: Sample Characteristics

In total, 43 young Black adults enrolled in the study, 32 of the 43 (74%) participants completed the follow-up survey, and 9 participants were ineligible for HPV vaccination because of prior HPV vaccination or other reasons not identified at the time of recruitment. Flow participant diagram through pre- and postintervention is shown in Figure 3. In the baseline sample of 43 participants, the average age was 21.7 (SD 2.7 years; Table 1). The majority were female (32/43, 74%), and more than half were African-born (25/43, 58%). Participants were well-educated, with nearly half having a college degree or beyond (20/33, 46%), and over three-quarters of participants had health insurance coverage (33/43, 77%). In comparing the completers and noncompleters of the follow-up survey, a higher proportion of completers were African-born (*P*=.002), had higher educational attainment (*P*=.02), and had higher income levels (*P*=.04). There was no difference in age, sex, health insurance coverage, or receipt of the HPV vaccine (at baseline) between completers and noncompleters.

XSL•FO
RenderX

Figure 3. Flow participant diagram of the progress through the phases of a pre and post 1-group intervention (enrollment, baseline, completed intervention, follow-up, and data analysis). HPV: human papillomavirus.

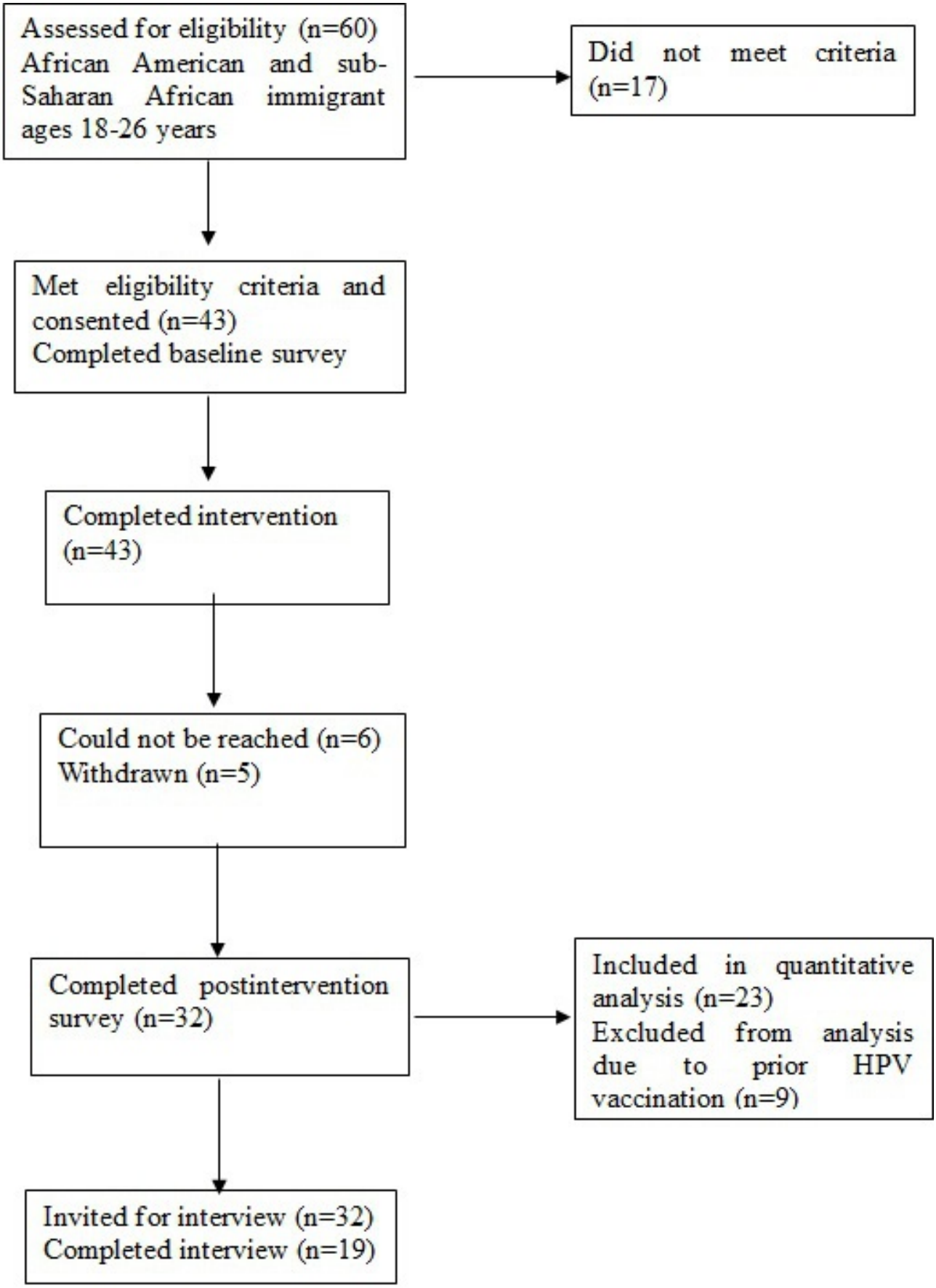

XSL•FO
RenderX

Table . Descriptive summary of sociodemographic characteristics of young adults (18 - 26 years).

	Full sample (N=43)	Completers (n=32)	Noncompleters (n=11)	*P* value
Age (years), mean (SD)	21.7 (2.7)	22.1 (2.9)	20.5 (1.6)	.09[a]
Sex, n (%)				.88[b]
Male	11 (26)	8 (25)	3 (27)	
Female	32 (74)	24 (75)	8 (73)	
Country of origin, n (%)				002[b]
US-born	18 (42)	9 (28)	9 (82)	
African-born	25 (58)	23 (72)	2 (18)	
Education, n (%)				.02[c]
High school or below	4 (9)	2 (6)	2 (18)	
Some college	19 (44)	12 (38)	7 (64)	
College graduate	13 (30)	11 (34)	2 (18)	
Master or higher	7 (16)	7 (22)	0 (0)	
Household income (US $), n (%)				.04[c]
Less than $25,000	14 (33)	11 (34)	3 (27)	
$25,000-$49,999	12 (28)	12 (38)	0 (0)	
$50,000-$79,999	7 (16)	5 (16)	2 (18)	
$80,000 or more	10 (23)	4 (13)	6 (55)	
Health insurance coverage, n (%)				.23[b]
Yes	33 (77)	26 (81)	7 (64)	
No	10 (23)	6 (18)	4 (36)	
Ever recieved the human papillomavirus vaccine, n (%)				.96[b]
Yes	12 (28)	9 (28)	3 (27)	
No or do not know	31 (72)	23 (72)	8 (73)	

[a]Two-tailed *t* test.

[b]Chi-square test of association.

[c]Mann-Whitney *U* test.

Among the 23 participants who completed the follow-up survey that reported not having ever received a dose of the HPV vaccine at baseline, 7 (30%) reported receipt of at least 1 dose of the HPV vaccine at follow-up. Among participants (n=23), there was a significant increase in HPV knowledge (*P*=.01) and HPV vaccine knowledge (*P*=.004; Table 2) in the unadjusted analysis. Based on a potential range of 0 - 16, average HPV knowledge scores increased from 7.3 (SD 4.2) to 11.1 (SD 4.3) from baseline to follow-up. For HPV vaccine knowledge, based on a potential range of 0 - 8, average scores increased from 2.8 (SD 2.5) on the baseline survey to 4.7 (SD 2.2) on the follow-up. Both of these changes remained statistically significant in the models adjusting for age, sex, and ethnicity (*P*=.004 and *P*=.003, respectively). There was also a significant change in HPV vaccine hesitancy. Scores on this scale increased from baseline (mean 28.3, SD 4.2) to follow-up (mean 29.9, SD 3.6; *P*=.02), which indicates a decrease in vaccine hesitancy. This change remained significant in the adjusted analysis (*P*=.007).

XSL•FO
RenderX

Table . Pre- and postchanges from baseline to follow-up among young adults (18 - 26 years) who have not received the human papillomavirus (HPV) vaccine at baseline (n=23).

	Potential range	Baseline, mean (SD)	Postintervention, mean (SD)	Unadjusted *P* value	Adjusted *P* value[a]
HPV knowledge	0 - 16	7.3 (4.2)	11.1 (4.3)	.01	.004
HPV vaccine knowledge	0 - 8	2.8 (2.5)	4.7 (2.2)	.004	.003
HPV vaccine hesitancy	9 - 36	28.3 (4.2)	29.9 (3.6)	.02	.007
HPV-related risk perception	1 - 5	2.4 (1.1)	2.7 (1.1)	.44	.58
HPV severity	4 - 20	16.5 (2.1)	16.5 (3.7)	.96	.93
HPV vaccination benefits	3 - 15	12.3 (1.8)	13.1 (2.2)	.19	.12
HPV vaccination self-efficacy	3 - 15	10.7 (2.1)	11.4 (2.6)	.34	.13
HPV vaccine intention	1 - 5	2.7 (0.8)	2.6 (1.1)	.88	.90

[a]Models adjusted for age, ethnicity, and sex.

There were no significant changes in HPV severity, vaccination benefits, vaccination self-efficacy, or intention among those who had not received the HPV vaccine at baseline. Among the completers of the intervention, baseline vaccine hesitancy and risk perception scores were near the midpoint of the scales, while HPV severity, vaccination benefits, and vaccination self-efficacy were relatively high.

For intervention satisfaction, participants rated the usability high (eg, information was important, credible, helpful, etc) with an average score of 24.8 (SD 3.4) of a potential maximum of 30. Similarly, based on potential ranges of 0 - 25, overall intervention acceptability (mean 20.0, SD 3.3; potential range 5 - 25) and relevance of the posts (mean 20.1, SD 2.7) were high. The usability, acceptability, and relevance of the posts or messages were supported further by one-on-one interviews (n=19) and open-ended questions (n=32).

Qualitative Data Collection

All participants who completed the postintervention assessments were invited for a follow-up interview via email. Follow-up one-on-one interviews were conducted with 19 participants to assess intervention satisfaction at 8 weeks (postintervention). The interviews were conducted on Zoom (Zoom Video Communications) by the first author (AA), a PhD-level qualitatively trained Black female researcher. Each interview began with an introduction, and the interview was guided by a semistructured interview guide developed by the research team. Each interview lasted 15 to 25 minutes. The interview questions were centered on how well the posts resonated with participants, how the intervention could be improved, and how the information learned from the intervention would be used. In addition, we asked if participants had received at least 1 dose of the HPV vaccination or scheduled an appointment. Participants were compensated with US $10 for interviews. Interviews were audio-recorded, transcribed verbatim, and deidentified for analysis.

Qualitative Data Analysis

Data from the one-on-one interviews and open-ended questions were analyzed by 2 qualitative trained research team members (Adegboyega Adebola and Wuni Abubakari). The researchers (AA and AW) read the transcripts and independently carried out a content analysis of the data. We did not use any qualitative software. Transcripts were coded line-by-line by the 2 researchers. Coding involved aggregating the data text into small categories of information and assigning a label to the code [46]. We used an iterative procedure to refine the codes and identify emergent themes until reaching saturation. Any discrepancies between results were discussed and resolved with the entire research team.

Qualitative Findings

In total, 19 participants completed the interviews, 12 (63%) were female participants. Two main themes emerged from the data: (1) feedback and recommendation on the intervention and (2) perceived impact of the intervention on vaccination decision.

Theme 1: Feedback and Recommendation on the Intervention

Participants expressed strong appreciation for the visual content used in the intervention, particularly memes and graphics. Participants found these visuals to be effective in delivering clear and engaging messages about HPV and vaccination. One participant stated, “the use of memes and graphics made the information clear and easy to understand.” Another participant noted, “visuals really caught my attention and made me more interested in learning about HPV.” The participants positively noted the notification system for new posts and the use of incentives to encourage engagement. One participant noted that “I like the fact that I get notified of when new information is posted.” Opportunities for improvements suggested by participants included the promotion of more participant engagement, the use of additional platforms, and shorter intervention duration. One participant suggested, “You should make it into a group that everybody is active, maybe like text

XSL•FO
RenderX

messaging, SMS, WhatsApp or something like a smaller group." Another participant noted, "I think I will like it shorter for half of the time it was done, like four weeks."

Theme 2: Perceived Impact of the Intervention on Vaccination Decision

Some participants who took the vaccine after the intervention reported that it significantly influenced their decision to vaccinate against HPV. Some participants indicated that the information provided through the intervention helped clarify misconceptions and increased their confidence in the vaccine's necessity. One participant noted, "After learning more from the intervention, I decided to get vaccinated." Another noted, "The intervention gave me the information I needed to make an informed decision about HPV vaccination."

Discussion

Principal Findings

This study was a pilot 1-group pre- and post-Facebook–based intervention (#HPVVaxTalks) study among 43 young Black adults in a southeastern state. Participants in the 8-week intervention showed significant improvements in HPV knowledge and HPV vaccine knowledge and a reduction in vaccine hesitancy after the intervention. Participants were satisfied with the intervention, and it was well-received. Participants did not show significant changes in HPV severity, vaccination benefits, vaccination self-efficacy, or intention among those who had not received the HPV vaccine at baseline.

The intervention led to HPV vaccination uptake, with 7 of 23 (30%) reporting that they have received at least 1 dose of the HPV vaccine at 8 weeks follow-up. This is a noteworthy result, given that the intervention targeted a high-risk group. This finding is in line with a study where 30% of Korean-American female participants aged 21 to 29 years who participated in a mobile health intervention received the first dose of the HPV vaccine after the intervention of a text messaging intervention to promote HPV vaccination [47]. Our finding highlights the impact of culturally appropriate intervention in promoting HPV vaccination among previously unvaccinated individuals. However, the low vaccination rates and no significant changes in HPV vaccine intention reported from our study show that this intervention had limited influence on HPV vaccination outcomes. According to the theory of planned behavior, the intention to perform a behavior is the strongest predictor of whether an individual engages in certain behaviors. The intention to engage in vaccination is determined by a person's beliefs, attitudes, and subjective norms within a specific context [34]. Hence, future interventions should use multicomponent interventions to address young adults' HPV vaccine–related beliefs, attitudes, and subjective norms. Multicomponent interventions should include effective patient-provider communication and HPV vaccine recommendations as a cancer prevention tool. In addition, there is a need for comprehensive social media campaigns to address myths and misinformation related to HPV targeted at young adults, given their wide reach and influence among young adults.

This study found a significant increase in HPV knowledge and HPV vaccine knowledge in the pre- and postintervention. The findings from this study are congruent with a social media intervention study where there was a statistically significant increase in knowledge after the intervention [48]. Furthermore, qualitative postintervention interviews with study participants supported the findings that our social media intervention increased their awareness and knowledge of HPV, particularly regarding transmission risks and preventive measures. This is in line with a similar study on mobile health interventions among unvaccinated youths, which showed improvement in HPV-related knowledge, vaccination intention, or vaccine initiation in a postintervention [49]. However, the findings from our study are inconsistent with a Twitter campaign that reported no significant changes in HPV knowledge after the campaign [50]. However, knowledge alone is insufficient for increasing HPV vaccination rates, as other factors beyond knowledge deficiency influence HPV vaccination decisions.

Our study found a significant reduction in HPV vaccine hesitancy. This finding supports the use of a Facebook-based intervention to decrease HPV vaccine hesitancy among young Black adults. A large-scale study across 67 countries showed that vaccine hesitancy can partly be attributed to a lack of confidence in vaccine safety, perceptions that vaccines do not work, distrust of information, perceived low risks of vaccine-preventable diseases, as well as a lack of trust in health care providers, authorities, and pharmaceutical companies [51]. Research has shown that interventions that support risk communication, community engagement, use of trained vaccine champions, tailored communication campaigns, and increased confidence in vaccines tend to reduce HPV vaccine hesitancy [52]. For this pilot study, we used a community-engaged approach by engaging with a youth advisory board for the development of #HPVVaxTalks intervention, which improved its cultural appropriateness for young Black adults. Future studies could use young adults as vaccine champions to deliver training and advocacy tailored to their communities to positively impact social norms [52].

Our study found no significant differences in HPV self-efficacy, intention to vaccinate, perceived severity, and perceived vaccination benefits. This may be because of the small sample size and relatively high baseline values for these variables. Future research with large sample size and longitudinal approach is warranted to test whether a Facebook-based intervention is effective in promoting significant change in HPV self-efficacy, intention to vaccinate, perceived severity, perceived vaccination benefits, and HPV vaccination uptake.

The intervention piloted in this study was well-received, with participants rating its usability, acceptability, and relevance highly. This study reiterates the importance of including the target population in formative research to promote usability, acceptability, and relevance to the target population. These findings are supported by research that reports that young adults are open to using social media to seek out health information and learn more about preventative health behaviors [32,48,50]. Using social media platforms could be an avenue and platform for delivering crucial information to reach young Black adults on a larger scale.

XSL•FO
RenderX

Participants reported that the intervention positively influenced their perception of the HPV vaccine's necessity and encouraged them to consider vaccination. These findings are supported by research that reports that young Black adults can benefit from social media interventions and seek out correct health information from social media platforms. Leader et al [31] showed that young adults were receptive to watching narrative-based health information videos on social media. Based on suggested opportunities for improvement, future research may explore shorter intervention duration to prevent attrition and use other social media avenues that may be relevant for health promotion and engagement for African American and sub-Saharan African immigrant Black young adults.

Implications for Practice or Research

The intervention's effect on HPV-related knowledge, vaccine knowledge, and vaccine hesitancy demonstrates its efficacy. Researchers can extend these findings by undertaking larger-scale research with varied groups and using other social media platforms to demonstrate the intervention's reproducibility and generalizability. Follow-up studies could determine whether knowledge increases, and behavioral improvements are sustainable over time. To determine if intervention effects persist beyond study completion and if participants complete the HPV vaccination series, longitudinal research is needed to monitor participants' HPV vaccination series completion. In addition, further research can determine which parts of the intervention (eg, specific message kinds and frequency of exposure) are most important to its effectiveness.

Tailored and community-driven social media interventions can be implemented to increase overall community uptake and support of the intervention and may improve health outcomes among minoritized and underserved populations. Health care professionals can use social media platforms to share evidence-based information regarding HPV and HPV vaccination. The intervention's success in increasing knowledge and vaccination shows that such initiatives could be beneficial on a larger scale. Messages can be tailored to young Black adults who are active on social media while taking into consideration cultural contexts and incorporating strategies to increase engagement and the effectiveness of future interventions. The usability, acceptability, and relevance of the intervention point to its usefulness in informing future large-scale Facebook-based studies.

Strengths and Limitations

Strengths of this study include the use of theory-driven intervention developed by a youth community advisory board, the use of a social media intervention, and a mixed methods approach. There are limitations to this study to consider. The sample is small and limited to those who had access to Facebook, and results are not generalizable to all young Black adults. The use of convenience sampling could have led to a homogenous participant pool and limited the diversity of perspectives. We collected self-reported measures, which could have social desirability bias. We also experienced a high attrition rate in the study. Future studies should oversample and send multiple routine reminders about follow-ups to minimize attrition. Another limitation of this study is the lack of a comparable control group. Similar to many pilot studies, the study was not powered, but provides a foundation for future studies, to be able to pilot-test the intervention and posts.

Conclusions

This study highlights significant improvements in HPV knowledge and HPV vaccine knowledge and a reduction in HPV vaccine hesitancy following a Facebook-based intervention among young Black adults. The findings from this study contribute to available research on the use of Facebook intervention among young Black adults. Intervention addressing knowledge gaps and improving HPV hesitancy among potential vaccine recipients is pivotal to promoting vaccine uptake. More research is needed to assess the type of messages and intervention dose that can most effectively facilitate changes in HPV perceived severity, vaccination benefits, self-efficacy, and intention to vaccinate to promote HPV vaccination behavior among young Black adults.

Acknowledgments

The authors thank all the volunteers who participated in the study. The study was supported by the American Cancer Society Institutional Grant (#IRG 19-140-31-IRG). The content is solely the responsibility of the authors and does not necessarily represent the official views of the sponsors.

Data Availability

The datasets analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Checklist 1
STROBE (STrengthening the Reporting of OBservational studies in Epidemiology) statement—checklist of items that should be included in reports of cohort studies.
[DOCX File, 33 KB - formative_v9i1e69609_app1.docx]

XSL•FO
RenderX

References

1. McNamara M, Batur P, Walsh JME, Johnson KM. HPV update: vaccination, screening, and associated disease. J Gen Intern Med 2016 Nov;31(11):1360-1366. [doi: 10.1007/s11606-016-3725-z] [Medline: 27184752]
2. Han JJ, Beltran TH, Song JW, Klaric J, Choi YS. Prevalence of genital human papillomavirus infection and human papillomavirus vaccination rates among US adult men: National Health and Nutrition Examination Survey (NHANES) 2013-2014. JAMA Oncol 2017 Jun 1;3(6):810-816. [doi: 10.1001/jamaoncol.2016.6192] [Medline: 28114440]
3. Rosenblum HG, Lewis RM, Gargano JW, Querec TD, Unger ER, Markowitz LE. Declines in prevalence of human papillomavirus vaccine-type infection among females after introduction of vaccine—United States, 2003–2018. MMWR Morb Mortal Wkly Rep 2021 Mar 26;70(12):415-420. [doi: 10.15585/mmwr.mm7012a2] [Medline: 33764964]
4. Chiou L, Tucker C. Social distancing, internet access and inequality. Public Goods & Regulation in Health Economics eJournal. 2020. URL: https://www.oxy.edu/sites/default/files/assets/Economics/Chiou/covid19_inequality_social_distancing.pdf [accessed 2020-03-10]
5. Boersma P, Black LI. Human papillomavirus vaccination among adults aged 18-26, 2013-2018. NCHS Data Brief 2020 Jan(354):1-8. [Medline: 32487295]
6. De P, Budhwani H. Human papillomavirus (HPV) vaccine initiation in minority Americans. Public Health (Fairfax) 2017 Mar;144:86-91. [doi: 10.1016/j.puhe.2016.11.005] [Medline: 28274389]
7. Pérez AE, Agénor M, Gamarel KE, Operario D. Nativity disparities in human papillomavirus vaccination among U.S. adults. Am J Prev Med 2018 Feb;54(2):248-258. [doi: 10.1016/j.amepre.2017.10.019] [Medline: 29241719]
8. Amboree TL, Darkoh C. Barriers to human papillomavirus vaccine uptake among racial/ethnic minorities: a systematic review. J Racial Ethn Health Disparities 2021 Oct;8(5):1192-1207. [doi: 10.1007/s40615-020-00877-6]
9. Adegboyega A, Adeyimika D, Omoadoni O, Mark D. HPV vaccination and cervical cancer screening promotion among Black individuals: social ecological perspectives from key informants interviews. Ethn Health 2023 Oct;28(7):1026-1040. [doi: 10.1080/13557858.2023.2193360] [Medline: 36973897]
10. Adegboyega A, Obielodan O, Wiggins AT, Dignan M, Williams LB. Beliefs and knowledge related to human papillomavirus (HPV) vaccine among African Americans and African immigrants young adults. Cancer Causes Control 2023 May;34(5):479-489. [doi: 10.1007/s10552-023-01678-y] [Medline: 36854988]
11. Zheng L, Wu J, Zheng M. Barriers to and facilitators of human papillomavirus vaccination among people aged 9 to 26 years: a systematic review. Sex Transm Dis 2021 Dec 1;48(12):e255-e262. [doi: 10.1097/OLQ.0000000000001407] [Medline: 33783412]
12. Kutz JM, Rausche P, Gheit T, Puradiredja DI, Fusco D. Barriers and facilitators of HPV vaccination in sub-Saharan Africa: a systematic review. BMC Public Health 2023 May 26;23(1):974. [doi: 10.1186/s12889-023-15842-1] [Medline: 37237329]
13. Holman DM, Benard V, Roland KB, Watson M, Liddon N, Stokley S. Barriers to human papillomavirus vaccination among US adolescents: a systematic review of the literature. JAMA Pediatr 2014 Jan;168(1):76-82. [doi: 10.1001/jamapediatrics.2013.2752] [Medline: 24276343]
14. Rosen BL, Shepard A, Kahn JA. US health care clinicians' knowledge, attitudes, and practices regarding human papillomavirus vaccination: a qualitative systematic review. Acad Pediatr 2018 Mar;18(2S):S53-S65. [doi: 10.1016/j.acap.2017.10.007] [Medline: 29502639]
15. Zhu X, Jacobson RM, MacLaughlin KL, Sauver JS, Griffin JM, Finney Rutten LJ. Parent-reported barriers and parental beliefs associated with intentions to obtain HPV vaccination for children in a primary care patient population in Minnesota, USA. J Community Health 2023 Aug;48(4):678-686. [doi: 10.1007/s10900-023-01205-9] [Medline: 36920709]
16. Galbraith-Gyan KV, Lechuga J, Jenerette CM, Palmer MH, Moore AD, Hamilton JB. HPV vaccine acceptance among African-American mothers and their daughters: an inquiry grounded in culture. Ethn Health 2019 Apr;24(3):323-340. [doi: 10.1080/13557858.2017.1332758] [Medline: 28553758]
17. Washington A, Chabaan J, Fakih A, et al. "Should I give it to my kids?": Factors that influence HPV vaccine hesitancy among African American parents. Cancer Epidemiol Biomarkers Prev 2022 Jul 1;31(7):1512-1512. [doi: 10.1158/1055-9965.EPI-22-0478]
18. Shapiro GK, Tatar O, Amsel R, et al. Using an integrated conceptual framework to investigate parents' HPV vaccine decision for their daughters and sons. Prev Med 2018 Nov;116:203-210. [doi: 10.1016/j.ypmed.2018.09.017]
19. Cox DS, Cox AD, Sturm L, Zimet G. Behavioral interventions to increase HPV vaccination acceptability among mothers of young girls. Health Psychol 2010 Jan;29(1):29-39. [doi: 10.1037/a0016942] [Medline: 20063933]
20. Perkins RB, Zisblatt L, Legler A, Trucks E, Hanchate A, Gorin SS. Effectiveness of a provider-focused intervention to improve HPV vaccination rates in boys and girls. Vaccine (Auckl) 2015 Feb;33(9):1223-1229. [doi: 10.1016/j.vaccine.2014.11.021]
21. McFadden SM, Ko LK, Shankar M, et al. Development and evaluation of an online continuing education course to increase healthcare provider self-efficacy to make strong HPV vaccine recommendations to East African immigrant families. Tumour Virus Res 2021 Jun;11:200214. [doi: 10.1016/j.tvr.2021.200214] [Medline: 33647533]
22. Chu H, Ko LK, Ibrahim A, et al. The impact of an educational forum intervention on East African mothers' HPV vaccine-related knowledge, attitudes, and intentions to vaccinate their adolescent children. Vaccine (Auckl) 2021 Jun 23;39(28):3767-3776. [doi: 10.1016/j.vaccine.2021.05.029] [Medline: 34053792]

XSL•FO
RenderX

23. Baranoski AS, Stier EA. Factors associated with time to colposcopy after abnormal Pap testing in HIV-infected women. J Womens Health (Larchmt) 2012 Apr;21(4):418-424. [doi: 10.1089/jwh.2011.3046]
24. Baranowski T, Cullen KW, Basen-Engquist K, et al. Transitions out of high school: time of increased cancer risk? Prev Med 1997;26(5 Pt 1):694-703. [doi: 10.1006/pmed.1997.0193] [Medline: 9327479]
25. Korda H, Itani Z. Harnessing social media for health promotion and behavior change. Health Promot Pract 2013 Jan;14(1):15-23. [doi: 10.1177/1524839911405850] [Medline: 21558472]
26. Ventola CL. Social media and health care professionals: benefits, risks, and best practices. P T 2014 Jul;39(7):491-520. [Medline: 25083128]
27. Mendoza-Herrera K, Valero-Morales I, Ocampo-Granados ME, Reyes-Morales H, Arce-Amaré F, Barquera S. An overview of social media use in the field of public health nutrition: benefits, scope, limitations, and a Latin American experience. Prev Chronic Dis 2020 Aug 6;17:E76. [doi: 10.5888/pcd17.200047] [Medline: 32762809]
28. Share of U.S. adults using social media, including Facebook, is mostly unchanged since 2018. Pew Research Center. 2019. URL: https://www.pewresearch.org/fact-tank/2019/04/10/share-of-u-s-adults-using-social-media-including-facebook-is-mostly-unchanged-since-2018 [accessed 2025-03-03]
29. Lebow S. Facebook is by far the most popular US social media platform. EMARKETER Inc. URL: https://www.emarketer.com/content/facebook-by-far-most-popular-us-social-media-platform [accessed 2025-03-10]
30. Average time spent on Facebook per day. Oberlo. 2024. URL: https://www.oberlo.com/statistics/average-time-spent-on-facebook [accessed 2025-03-03]
31. Leader AE, Miller-Day M, Rey RT, Selvan P, Pezalla AE, Hecht ML. The impact of HPV vaccine narratives on social media: testing narrative engagement theory with a diverse sample of young adults. Prev Med Rep 2022 Oct;29:101920. [doi: 10.1016/j.pmedr.2022.101920] [Medline: 36161120]
32. Brandt HM, Sundstrom B, Monroe CM, et al. Evaluating a technology-mediated HPV vaccination awareness intervention: a controlled, quasi-experimental, mixed methods study. Vaccines (Basel) 2020 Dec 10;8(4):749. [doi: 10.3390/vaccines8040749] [Medline: 33321975]
33. Wysota CN, Abroms LC, DeVarona H, Koban D, Napolitano M, Broniatowski DA. User experiences with a moderated Facebook group to promote vaccination. Am J Health Promot 2025 Jan;39(1):89-102. [doi: 10.1177/08901171241272061] [Medline: 39110567]
34. Montaño DE, Kasprzyk D. Theory of reasoned action, theory of planned behavior, and the integrated behavioral model. In: Glanz K, Rimer BK, Viswanath KV, editors. Health Behavior: Theory, Research and Practice, 5th edition: Jossey-Bass/Wiley; 2015:95-124.
35. Catalano H, Richards K, Hawkins K. Theory of planned behavior-based correlates of HPV vaccination intentions and series completion among university students in the Southeastern United States. Health Educ (Muncie) 2017;49(2):35-44 [FREE Full text]
36. Adegboyega A, Francis DB, Ebikwo C, Ntego T, Ickes M. Engagement with a youth community advisory board to develop and refine a Facebook HPV vaccination promotion intervention (#HPVvaxtalks) for young black adults (18-26 years old). Health Promot Pract 2023 Dec 5:15248399231216731. [doi: 10.1177/15248399231216731] [Medline: 38050905]
37. Adebola A, Charis E, Tristan N, Melinda I. Content validation of a Facebook HPV vaccination promotion intervention (#HPVvaxtalks) created for young Black adults (18–26 years old). Public Health Nurs 2024 Sep;41(5):1178-1187. [doi: 10.1111/phn.13356]
38. Kreuter MW, Lukwago SN, Bucholtz R, Clark EM, Sanders-Thompson V. Achieving cultural appropriateness in health promotion programs: targeted and tailored approaches. Health Educ Behav 2003 Apr;30(2):133-146. [doi: 10.1177/1090198102251021] [Medline: 12693519]
39. HPV social media graphics. National Foundation for Infectious Diseases. URL: https://www.nfid.org/resource/hpv-social-media-graphics/ [accessed 2025-03-10]
40. Gerend MA, Shepherd JE. Predicting human papillomavirus vaccine uptake in young adult women: comparing the health belief model and theory of planned behavior. Ann Behav Med 2012 Oct;44(2):171-180. [doi: 10.1007/s12160-012-9366-5] [Medline: 22547155]
41. Marlow LAV, Zimet GD, McCaffery KJ, Ostini R, Waller J. Knowledge of human papillomavirus (HPV) and HPV vaccination: an international comparison. Vaccine (Auckl) 2013 Jan;31(5):763-769. [doi: 10.1016/j.vaccine.2012.11.083]
42. Waller J, Ostini R, Marlow LAV, McCaffery K, Zimet G. Validation of a measure of knowledge about human papillomavirus (HPV) using item response theory and classical test theory. Prev Med 2013 Jan;56(1):35-40. [doi: 10.1016/j.ypmed.2012.10.028]
43. Luyten J, Bruyneel L, van Hoek AJ. Assessing vaccine hesitancy in the UK population using a generalized vaccine hesitancy survey instrument. Vaccine (Auckl) 2019 Apr;37(18):2494-2501. [doi: 10.1016/j.vaccine.2019.03.041]
44. Jones HJ, Kovacic MB, Bakas T. Establishing validity of the midlife Black women's stress-reduction wellness program materials using a mixed methods approach. J Cardiovasc Nurs 2022;37(5):446-455. [doi: 10.1097/JCN.0000000000000876] [Medline: 34935740]

XSL•FO
RenderX

45. Horvath KJ, Oakes JM, Rosser BRS, et al. Feasibility, acceptability and preliminary efficacy of an online peer-to-peer social support ART adherence intervention. AIDS Behav 2013 Jul;17(6):2031-2044. [doi: 10.1007/s10461-013-0469-1] [Medline: 23553347]
46. Creswell JW. Qualitative Inquiry and Research Design: Choosing Among Five Approaches: Sage Publications; 2012.
47. Lee HY, Koopmeiners JS, McHugh J, Raveis VH, Ahluwalia JS. mHealth pilot study: text messaging intervention to promote HPV vaccination. Am J Health Behav 2016 Jan;40(1):67-76. [doi: 10.5993/AJHB.40.1.8] [Medline: 26685815]
48. Reiter PL, Gower AL, Kiss DE, et al. Effects of a web-based HPV vaccination intervention on cognitive outcomes among young gay, bisexual, and other men who have sex with men. Hum Vaccin Immunother 2022 Nov 30;18(6):2114261. [doi: 10.1080/21645515.2022.2114261] [Medline: 36069662]
49. Ou L, Chen ACC, Amresh A. The effectiveness of mHealth interventions targeting parents and youth in human papillomavirus vaccination: systematic review. JMIR Pediatr Parent 2023 Nov 21;6:e47334. [doi: 10.2196/47334] [Medline: 37988155]
50. Allen JD, Hollander J, Gualtieri L, Alarcon Falconi TM, Savir S, Agénor M. Feasibility of a twitter campaign to promote HPV vaccine uptake among racially/ethnically diverse young adult women living in public housing. BMC Public Health 2020 Jun 1;20(1):830. [doi: 10.1186/s12889-020-08824-0] [Medline: 32487045]
51. Larson HJ, de Figueiredo A, Xiahong Z, et al. The state of vaccine confidence 2016: global insights through a 67-country survey. EBioMedicine 2016 Oct;12:295-301. [doi: 10.1016/j.ebiom.2016.08.042] [Medline: 27658738]
52. Tuckerman J, Kaufman J, Danchin M. Effective approaches to combat vaccine hesitancy. Pediatr Infect Dis J 2022 May 1;41(5):e243-e245. [doi: 10.1097/INF.0000000000003499] [Medline: 35213864]

Abbreviations

HPV: human papillomavirus
REDCap: Research Electronic Data Capture
STROBE: Strengthening the Reporting of Observation Studies in Epidemiology

Edited by A Mavragani; submitted 03.12.24; peer-reviewed by J Cunningham-Erves, W Tang; revised version received 07.02.25; accepted 08.02.25; published 02.04.25.

Please cite as:
Adegboyega A, Wiggins A, Wuni A, Ickes M
The Impact of a Human Papillomavirus Facebook-Based Intervention (#HPVVaxTalks) Among Young Black (African American and Sub-Saharan African Immigrants) Adults: Pilot Pre- and Poststudy
JMIR Form Res 2025;9:e69609
URL: https://formative.jmir.org/2025/1/e69609
doi:10.2196/69609

© Adebola Adegboyega, Amanda Wiggins, Abubakari Wuni, Melinda Ickes. Originally published in JMIR Formative Research (https://formative.jmir.org), 2.4.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Medical Device Based on a Virtual Reality–Based Upper Limb Rehabilitation Software: Usability Evaluation Through Cognitive Walkthrough

Seojin Hong[1], MS; Hyun Choi[2], PhD; Hyosun Kweon[1], PhD

[1]Department of Clinical Research Rehabilitation, National Rehabilitation Center, 58 Samgaksan-ro, Gangbuk-gu, Seoul, Republic of Korea

[2]Department of Healthcare and Public Health, National Rehabilitation Center, Seoul, Republic of Korea

Corresponding Author:
Hyosun Kweon, PhD
Department of Clinical Research Rehabilitation, National Rehabilitation Center, 58 Samgaksan-ro, Gangbuk-gu, Seoul, Republic of Korea

Abstract

Background: The use of virtual reality (VR) technology in rehabilitation therapy has been growing, leading to the development of VR-based upper-limb rehabilitation softwares. To ensure the effective use of such software, usability evaluations are critical to enhance user satisfaction and identify potential usability issues.

Objective: This study aims to evaluate the usability of a VR-based upper-limb rehabilitation software from the perspective of occupational therapists. Specifically, the study seeks to identify usability challenges and provide insights to improve user satisfaction.

Methods: The VR-based upper-limb rehabilitation software was tailored for therapists to operate while delivering therapy to patients. Usability testing was conducted with occupational therapists from the Korean National Rehabilitation Center using cognitive walkthroughs and surveys. Participants performed tasks that simulated real clinical scenarios, including turning the device on, assisting patients with wearing the device, and shutting it down. Observers recorded user reactions during task performance, and participants completed surveys to assess the ease of use of the user interface. This mixed-methods approach provided qualitative insights into user difficulties and their root causes.

Results: Usability evaluations were conducted with 6 participants. Cognitive walkthroughs revealed potential areas for improvement in the software, including (1) enhancements to the graphical user interface for ease of use, (2) refinements in the natural user interface, and (3) better user manuals for clearer product instructions. The ease-of-use score for the user interface averaged 1.58 on a 5-point scale (1=very easy to 5=very difficult).

Conclusions: This study provides valuable insights into improving user satisfaction by focusing on the needs of occupational therapists who operate a VR-based rehabilitation software. Future research should explore software refinement and clinical efficacy to maximize the therapeutic potential of such technologies.

(JMIR Form Res 2025;9:e68149) doi:10.2196/68149

KEYWORDS

usability; cognitive walkthrough; virtual reality-based upper limb rehabilitation software; upper limb; limb rehabilitation; rehabilitation; therapist; virtual reality; VR; medical device; formative evaluation; quantitative; qualitative; occupational therapy; user safety; usability testing; software; risk factor

Introduction

Background

Upper limb rehabilitation plays a critical role in restoring functional abilities in individuals with stroke or neurological injuries, enhancing their independence and autonomy [1,2]. Virtual reality (VR) technology has emerged as a promising tool in rehabilitation therapy, offering immersive and interactive features that can improve patient engagement and therapeutic outcomes [3-5]. Through tailored feedback and adjustable difficulty levels, VR-based rehabilitation programs can address diverse patient needs and therapeutic goals [6]. To achieve optimal clinical use, VR-based rehabilitation software must be designed to meet the needs of end users, including health care professionals such as occupational therapists, within their specific work environments. Usability evaluation is a critical step in this process, as it helps identify design flaws, reduce user errors, and enhance operational efficiency and satisfaction [7]. While existing studies on VR-based rehabilitation have largely focused on technical development and clinical efficacy [4], research on usability issues experienced by health care

XSL•FO
RenderX

professionals during actual software use remains limited [3] . Specifically, there is a lack of comprehensive analysis regarding the intuitive interaction, learning curve, and practical applicability of such software in real-world clinical settings [7].

Systematic usability evaluation targeting occupational therapists, as key stakeholders in VR-based rehabilitation, is essential to identify and address usability issues. Formative evaluation methods, such as cognitive walkthroughs, provide valuable insights during the early stages of software development, enabling iterative improvements and better user experience [3].

Usability Evaluation

Errors encountered during the use of medical devices are often linked to poorly designed interfaces rather than user mistakes. Although not all user errors result from design flaws, usability evaluations conducted before commercialization can effectively mitigate design-related issues [8,9]. Usability engineering processes play a vital role in identifying and addressing predictable user errors in clinical environments, ensuring the safety and efficiency of medical devices [10,11].

Usability evaluations can be categorized as formative or summative, depending on the timing and purpose of the evaluation [12]. Formative evaluations aim to identify usability issues and guide design improvements during the product development phase, while summative evaluations focus on validating outcomes at the final stages of development through objective data analysis [13]. Formative evaluations commonly use methods such as cognitive walkthroughs, task analysis, usability testing, and heuristic evaluation to explore and address usability challenges in depth [14].

Cognitive walkthroughs, in particular, are highly effective for evaluating software usability. They focus on assessing the system from the perspective of new users, enabling the identification of potential issues such as inconsistent interface behavior or a lack of clarity [15]. By providing insights into users' cognitive processes, cognitive walkthroughs empower developers to design more user-friendly, efficient, and satisfying interfaces.

Objective

This study aims to conduct a formative evaluation of a prototype VR-based upper limb rehabilitation software, focusing on its usability for occupational therapists. By identifying usability challenges and gathering feedback from intended users, this research seeks to provide actionable insights to guide iterative improvements in the software design. Ultimately, this study aims to contribute foundational data for user-centered design and optimize the software for effective clinical implementation.

Methods

Study Design

This study conducted a usability evaluation of a prototype VR-based upper limb rehabilitation software targeted for use by occupational therapists. The evaluation used cognitive walkthroughs and surveys, covering the entire user workflow, including activating the device, setting it up, and completing tasks. The software was assessed as an embedded system, including its hardware components.

Formative Evaluation Procedures

The formative evaluation was conducted in the clinical rehabilitation testbed of the Korean National Rehabilitation Center. Each session lasted approximately 90 minutes per participant (Table 1). Before the evaluation, the facilitator introduced the usability testing process, including its objectives, methods, and relevant product information.

XSL•FO
RenderX

Table . Formative evaluation procedures. The formative evaluation involved practitioners and evaluators, and was conducted for 90 minutes under the guidance of a facilitator.

Composition	Details	Time (minutes)
Orientation	• The facilitator explains the target product, purpose, overview, methodology, and other relevant aspects of the usability evaluation to the participant. • The facilitator introduces the concept of usability evaluation to participants who are unfamiliar with it.	10
Guidance on consent	• The facilitator provides the participant with detailed instructions regarding recording and transcription procedures. • The facilitator ensures that the participant fully comprehends the evaluation content and voluntarily provides written consent to participate in the usability evaluation.	10
Perform cognitive walkthrough	• The facilitator conducts a cognitive walkthrough following a predefined script and product usage instructions. • The participant operates the medical device in accordance with the facilitator's guidance. • The observer documents and analyzes the participant's responses during task performance.	50
Survey	• The facilitator administers a survey to the participant to assess the usability of the user interface.	20

Participants were asked about their previous experience with similar devices. Subsequently, they followed a structured script guiding them through a series of tasks using the software. Participants operated the software as instructed, while observers recorded their behaviors and responses throughout the process. Post-testing, participants were asked to evaluate their satisfaction with the system based on the VR-based upper limb rehabilitation software.

The usability testing was conducted in a simulated clinical environment resembling the intended use setting for the device (Figure 1). Environmental conditions, including lighting, temperature, and noise, were measured before the evaluation: lighting at 550 (SD 100) lx, temperature at 24°C (SD 2°C), relative humidity at 60% (SD 10%), and noise level at 50 (SD 5) dBA.

Figure 1. Test environment (left) and scene (right). The simulation environment is configured to resemble a rehabilitation treatment room where evaluators are used.

Cognitive Walkthrough

The cognitive walkthrough method evaluates how well users can navigate an interface without previous training or information about the device [16]. In this study, 6 occupational therapists conducted the cognitive walkthrough, assessing the ease of navigation for specific scenarios outlined in Table 2. Observers recorded and analyzed participants' reactions based on predefined observation sheets. Observation sheets are designed to record problems discovered during each task, user reactions and actions, and system feedback [17,18].

XSL•FO
RenderX

Table . Use scenarios for cognitive walkthrough. The facilitator verbally provided detailed subtask instructions to the occupational therapist, who performed the tasks in accordance with the usage scenario, while the simulated patient participated in the process under the occupational therapist's guidance.

Task	Subtask
1. Reviewing the user manual	• 1.1. Reviewing the user manual.
2. Checking components and power supply	• 2.1. Checking the HMD[a], controller, disposable face mask, PC set, and software in the PC. • 2.2. Checking the product's power supply.
3. Launching the program	• 3.1. Launching the "RehabwareVR" (Tech Village) icon and entering the password.
4. Creating patient information	• 4.1. Entering patient information. • 4.2. Providing instructions on patient data consent.
5. Donning the equipment	• 5.1. Instructing the patient to wear the disposable face mask, HMD, and controller.
6. Running the upper rehabilitation content	• 6.1. Selecting one of the upper limb rehabilitation exercises, explaining it to the patient, and running.
7. Running the content	• 7.1. Selecting the patient and configuring the catch ball activity. • 7.2. Saving these settings as User Setting 3, then starting the catch ball activity. • 7.3. After completing at least five repetitions, stopping the activity, modifying the environment settings, and rerunning the activity. • 7.4. Switching to the fruit transfer activity and running it. • 7.5. After completing the fruit transfer activity, proceeding with the bubble activity.
8. Updating patient information	• 8.1. Updating the patient information.
9. Running contents after updating patient information	• 9.1. Selecting the patient and starting the hammering activity. • 9.2. After completing the hammering activity, switching to the meteor avoidance activity. • 9.3. Ending the session.
10. Doffing the product	• 10.1. Removing the HMD, controllers, and disposable face mask from the patient.
11. Reviewing statistical analysis results	• 11.1. Reviewing the data graph of the patient. • 11.2. Reviewing the fruit transfer activity data for patient. • 11.3 Deleting the hammering activity data for patient.
12. Ending the program	• 12.1. Ending the program.

[a]HMD: head-mounted display.

Survey

Before the cognitive walkthrough, participants provided demographic information, clinical experience, and previous experience with similar devices. After completing the cognitive walkthrough, participants completed a survey to further assess the usability of the user interface based on the scenarios. The survey consisted of 23 items rated on a 5-point Likert scale (1=very easy to 5=very difficult). Results were analyzed using means and SDs.

VR-Based Upper Limb Rehabilitation Software

The prototype VR-based upper limb rehabilitation software used in this study was designed to improve upper limb function, including shoulder mobility, in patients with stroke or neurological injuries (Figure 2). Patients wear a head-mounted display (HMD; HTC VIVE Pro Full-Kit including VIBE base stations and controllers) and use handheld controllers to perform voluntary movements in a VR environment for rehabilitation (Figure 3). Occupational therapists, as intended users, set up the hardware, operate the software, and assist patients in improving their upper limb functions.

XSL•FO
RenderX

Figure 2. User interface used by occupational therapists in the virtual reality (VR)–based upper limb rehabilitation software: home screen (left), activity selection screen (middle), and activity options manipulation screen (right).

Figure 3. Hardware device of the virtual reality (VR)–based upper limb rehabilitation software. To use VR content, the patient wears the head-mounted device (HMD; left) on their head and holds the controller (right).

The software includes eight rehabilitation activities (Figure 4): (1) throwing a catching balls, (2) hammering nails, (3) popping bubbles, (4) moving fruits, (5) dodging meteors, (6) throwing balls, (7) playing the xylophone, and (8) stacking blocks.

Figure 4. Examples of virtual reality within the patient-worn head-mounted device (HMD), such as throwing a catching balls (left) and hammering nails (right).

The minimum computer requirements to run the software are Microsoft Windows 7 SP, Windows 8.1, or Windows 10, with a display resolution of 1280×72 0r higher. Recommended specifications include Microsoft Windows 10 Professional with the same resolution. The hardware used was the VIVE Pro Full-Kit.

Participants

The usability evaluation involved 3 usability practitioners (a facilitator and 2 observers), 6 occupational therapists as participants, and 3 mock patients. The facilitator conducted the evaluation, while observers recorded and analyzed participants' responses. The sample size of 6 participants was determined to balance cost and effectiveness, as this number is sufficient to identify most usability issues [19]. Mock patients were recruited to simulate the role of patients during testing, prioritizing safety by selecting healthy individuals. In the early stages of development, formative evaluations typically involve the use of simulated patients before including real patients. This approach ensures safety and control while being effective in

XSL•FO
RenderX

identifying functional issues in devices [20,21]. Furthermore, unlike real patients, simulated patients are trained to provide consistent conditions and responses, thereby maintaining the reliability and consistency of the evaluation process [22,23]. Accordingly, this study employed simulated patients instead of real patients to conduct the evaluation. All participants were informed of the study's purpose, methods, and procedures before testing and provided signed consent.

Recruitment Process

Recruitment was conducted through posted announcements at the Korean National Rehabilitation Center. The inclusion criteria encompassed possession of an occupational therapist license and previous experience with rehabilitation medicine devices, such as computerized cognitive rehabilitation programs, whereas the exclusion criteria comprised individuals who did not provide consent to participate in the study. Through eligibility screening, 6 occupational therapists were selected. The mock patients were 3 healthy adults aged 19 years and older and had no experience using the evaluation product. They were provided detailed explanations of the study and coordinated testing schedules. Written informed consent was obtained from all participants.

Ethical Considerations

This study was approved by the Institutional Review Board of the Korean National Rehabilitation Center (NRC-2023-06-041). All participants provided informed consent, and no financial compensation was offered. Data were anonymized to ensure participant privacy and confidentiality.

Results

Participant Characteristics

The clinical experience of the occupational therapists ranged from 6 months to 6 years, with an average of 3 years and 9 months. In addition, they had previous experience using similar medical devices, such as Comcog and RAPAEL (Neofect) (computerized cognitive rehabilitation programs), with their experience ranging from a minimum of 2 months to a maximum of 5 years (Table 3).

Table . General characteristics of participants. The participant in the cognitive walkthrough was an occupational therapist who had experience using similar medical devices.

Participant Number	Sex	Age (years)	Experience	Occupation	Experience using similar medical devices	Name of similar medical device
1	Female	28	6 years	Occupational therapist	6 months, twice a week	Comcog (Neofect), RAPAEL Smart Glove (Neofect), and RAPAEL Smart pegboard
2	Female	24	6 months	Occupational therapist	2 months, twice a week	RAPAEL
3	Female	27	5 years	Occupational therapist	5 years, twice a week	Comcog, RAPAEL, and Reha-Com (Hasomed)
4	Female	25	3 years	Occupational therapist	3 years, once a week	RAPAEL and Xbox (Microsoft)
5	Male	27	3 years, 9 months	Occupational therapist	2 years, once a day	Comcog
6	Female	29	5 years, 8 months	Occupational therapist	1 year, 3 times a month	Comcog and Reha-Com

Cognitive Walkthrough Results

A total of 12 usability errors were identified using cognitive walkthrough. In task 4, "Creating Patient Information," 4 errors were identified. In subtask 4.1, "Locating the Add Patient Icon," participants were confused as they could not find the icon leading to the screen for adding patient information. In subtask 4.2, "Registering Patient Information," participants hesitated when they had to enter such details as the diagnosis and dominant hand manually because moving to the next input field required a click on the mouse, as the tab key shortcut did not work. In subtask 4.3, "Guiding Patients' Consent for Using Personal Information," participants skipped the data consent step and almost automatically pressed the confirm button without explaining the process of consent for using personal information to the mock patient.

In task 7, "Performing the Activity," 5 usability issues were identified. In subtask 7.1, "Selecting Activity," participants attempted to select activities without first selecting the patient, which triggered a popup message indicating "No patient selected." In subtask 7.2, "Configuring the Activity Settings," participants committed multiple errors, including failure to estimate the perceived distance between the hand and the object within the virtual reality, and identifying which hand was being used for interaction. Furthermore, although the users clicked on the "User Settings" button, they failed to save the configurations. There were also challenges in adjusting precise decimal values using mouse clicks and dragging, and this time-consuming process delayed the therapeutic session. In subtask 7.4, "Changing Activities," participants switched activities without saving current data, which led to the loss of unsaved records. In addition, some therapists had difficulty

XSL•FO
RenderX

adjusting the settings after completing an activity, as they were unsure whether they needed to press the confirm button before proceeding to the next activity.

In task 11, “Reviewing Statistical Analysis Results,” 3 errors were identified. In subtask 11.1, “Reviewing Data Graphs,” no issues were observed. In subtask 11.2, “Reviewing Activity Data,” participants were unable to review the data because activity records were not saved, and there was inconsistency between the activity names on the playscreen and the statistical analysis screen, leading to further confusion. In subtask 11.3, “Deleting Data” participants clicked on the pie chart and searched for a delete button on the detailed graph screen.

Survey Results

The survey assessing the usability of user interface showed that most participants rated usability as a score of 2 or lower, indicating ease of use on a 5-point Likert scale (1=very easy to 5=very difficult; Table 4).

XSL•FO
RenderX

Table . Survey results of usability of user interface (n=5). Usability of user interface showed that most participants rated usability as a score of 2 (easy) or lower (1=very easy; 5=very difficult).

Usability of UI[a]	Mean (SD)
Reviewing user manual	
Identifying information	1.67 (SD 0.82)
Comprehensibility	1.83 (SD 0.98)
Checking the components	
Differentiating between components	1.50 (SD 0.55)
Verifying connection and power	1.17 (SD 0.41)
Launching the program	
Launching the software	1.00 (SD 0.00)
Entering the password	1.00 (SD 0.00)
Creating patient information	
Creating patient information	2.17 (SD 0.75)
Consenting to the disclosure of personal information	1.67 (SD 1.03)
Donning the product	
Donning the product	2.00 (SD 1.10)
Instructions regarding the product	1.83 (SD 0.98)
Running the upper rehabilitation content	
Engaging in the content	1.50 (SD 0.55)
Explaining the contents	1.50 (SD 0.55)
Performing the activity	
Pausing the content	1.67 (SD 1.03)
Configuring detailed settings	2.17 (SD 1.33)
Changing detailed settings	2.17 (SD 1.33)
Ending the content	1.67 (SD 1.03)
Updating patient information	
Updating patient information	1.33 (SD 0.82)
Running contents after updating patient information	
Changing the patient	1.67 (SD 1.03)
Doffing the product	
Doffing the product	1.50 (SD 0.84)
Reviewing statistical analysis results	
Checking the graph	1.33 (SD 0.82)
Checking data	1.33 (SD 0.82)
Deleting data	1.50 (SD 1.22)
Ending the program	
Ending the software	1.17 (SD 0.41)

[a]UI: user interface.

Discussion

Principal Findings

The following recommendations were suggested from the results of the cognitive walkthrough to enhance usability and user satisfaction with the interface of the VR-based upper limb rehabilitation software. First, enhancements to the graphical user interface (GUI) are essential. A GUI facilitates more intuitive and efficient information presentation using graphics instead of text [24]. This study identified usability issues stemming from the lack of clarity in the “Add Patient” icon,

XSL•FO
RenderX

underscoring the need for a more intuitive icon that clearly conveys its function. In addition, improvements in typography, such as optimizing font size and color contrast, are necessary to enhance the readability of activity-setting screens.

Furthermore, to streamline data entry, manual input for all patient information fields should be replaced with input controls such as pickers, which allow users to select from predefined options, or steppers, which enable users to increment or decrement numbers or values using "+" and "-" buttons. In addition, field navigation should be optimized by enabling users to switch between fields using the tab key, thereby facilitating user input. Finally, the functionality for controlling buttons within the software should be visibly displayed on the interface. When developing GUI-based medical software devices, the interface design should aim to integrate clear control mechanisms and simplify tasks, thereby minimizing users' cognitive load and memory demands [25].

Second, integrating natural user interface (NUI) elements could further enhance usability. NUI refers to an interface that relies on gestures rather than traditional input devices, such as a mouse or keyboard, to control content [26]. While the VR-based upper limb rehabilitation software uses a gesture-based interface using an HMD and controllers, the system does not differentiate between left and right sides. This limitation can hinder therapists from properly administering treatments for patients with unilateral paralysis. To address this issue, the software should include features to either automatically detect the controller's orientation or provide manual configuration options to distinguish the affected side. Furthermore, therapists reported difficulties in adjusting the perceived distance between the hand and objects displayed on their monitors, highlighting the need for improved distance control mechanisms in the VR environment. Therefore, a VR software interface should be designed to ensure ease of use for both therapists and patients.

Third, the user interface (UI) design requires improvements to enhance the interaction between the medical device and its users. In this study, we observed that therapists used inconsistent methods for adjusting settings and some failed to save their configurations before proceeding with the tasks. Although most medical device users are well-trained and experienced, errors can still occur when they rely heavily on their own previous knowledge. If the system can provide warning alerts when settings are unsaved or incorrect actions are detected, and provide confirmation prompts when actions are correct, unintended errors can be mitigated [27]. Therefore, warning alerts, confirmation popups, detailed instructions, and help guides within the software interface are recommended to prevent usage errors.

Formative evaluation is a type of usability test conducted during the medical device development phase to improve product design. In this study, we conducted a cognitive walkthrough and surveys with occupational therapists as participants to explore potential areas for improvement in the VR-based upper limb rehabilitation software.

Cognitive walkthrough is a usability test method used to systematically identify interactions between users and a system without previous knowledge of the system. It is particularly suitable for evaluating screen-based systems where users must actively navigate the interface [16]. Test tasks should be selected from the user's perspective, and the sequence of potential actions should first be outlined, followed by a step-by-step analysis of continuous actions to assess whether appropriate actions are taken at the correct moments. Successful interaction is determined by whether the system feedback helps users achieve their intended goals [28].

In this study, we identified usability issues that need to be addressed in the UI of the VR-based upper limb rehabilitation software as a medical device. However, there are some limitations to note. Although both health care providers delivering the treatment and patients receiving it were considered users, actual patients were not included in the participant group. In addition, the therapists involved in the study did not represent a diverse range of clinical experience levels. Despite these limitations, the rehabilitation specialists who participated in this study provided valuable insights, including perspectives from the patients' point of view. Future research should expand the scope of such work to include a broader range of users and conduct a more comprehensive analysis of the usability across various features and tasks.

Conclusion

VR-based rehabilitation represents a rapidly emerging field in health care. In this study, we conducted a usability evaluation of a prototype VR-based upper limb rehabilitation software. By ensuring a user-centered interface and addressing the identified UI issues and areas for improvement, enhancing the usability of the software as a medical device will contribute to the development of a solution that fully meets the needs of its users. These efforts will not only prioritize user satisfaction and ease of use but also accelerate the commercialization of the VR-based upper limb rehabilitation software. Furthermore, improving user-device interaction will enhance the usability and satisfaction associated with the device, ultimately contributing to the development of a medical device that delivers error-free performance, as intended by the system.

Acknowledgments

This study was supported by the National Rehabilitation Center Rehabilitation Research & Development Support Program, funded by the Ministry of Health and Welfare (NRCRSP-25TB01).

Data Availability

The datasets generated during and analyzed during this study are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Authors' Contributions

The research design and data analysis were primarily conducted by SH, while data collection and experimental execution were contributed by HC. The interpretation of results and drafting of the manuscript were carried out jointly. The final review and approval of the manuscript, as well as research direction adjustments, were overseen by HK.

Conflicts of Interest

None declared.

References

1. Carson RG, Hayward KS. Using mechanistic knowledge to appraise contemporary approaches to the rehabilitation of upper limb function following stroke. J Physiol 2025 Feb;603(3):635-650. [doi: 10.1113/JP285559] [Medline: 39129269]
2. Zhang T, Yao Z, Chen F, et al. Enhancing upper limb function and motor skills post-stroke through an upper limb rehabilitation robot. J Vis Exp 2024 Sep 6(211). [doi: 10.3791/66938] [Medline: 39311613]
3. Zhang T, Booth R, Jean-Louis R, et al. A primer on usability assessment approaches for health-related applications of virtual reality. JMIR Serious Games 2020 Oct 28;8(4):e18153. [doi: 10.2196/18153] [Medline: 33112243]
4. Seo NJ, Arun Kumar J, Hur P, Crocher V, Motawar B, Lakshminarayanan K. Usability evaluation of low-cost virtual reality hand and arm rehabilitation games. J Rehabil Res Dev 2016;53(3):321-334. [doi: 10.1682/JRRD.2015.03.0045] [Medline: 27271199]
5. Marques B, Oliveira S, Ferreira H, Amorim P, Dias P, Santos BS. The role of collaborative virtual reality engagement in stroke survivors' rehabilitation. In: Accessibility, Assistive Technology and Digital Environments: AHFE International; 2024, Vol. 121. URL: https://openaccess.cms-conferences.org/publications/book/978-1-958651-97-1/article/978-1-958651-97-1_3 [accessed 2025-01-26]
6. Oliveira S, Marques B, Amorim P, Dias P, Sousa Santos B. Stepping into recovery with an immersive virtual reality serious game for upper limb rehabilitation: a supermarket experience for stroke survivors. In: Chen JYC, Fragomeni G, editors. Virtual, Augmented and Mixed Reality Lecture Notes in Computer Science: Springer Nature Switzerland; 2024, Vol. 14708:142-153. [doi: 10.1007/978-3-031-61047-9_9]
7. Kamińska D, Zwoliński G, Laska-Leśniewicz A. Usability testing of virtual reality applications-the pilot study. Sensors (Basel) 2022 Feb 10;22(4):1342. [doi: 10.3390/s22041342] [Medline: 35214246]
8. Johnson TR, Tang X, Graham MJ, et al. Attitudes toward medical device use errors and the prevention of adverse events. Jt Comm J Qual Patient Saf 2007 Nov;33(11):689-694. [doi: 10.1016/s1553-7250(07)33079-1] [Medline: 18074717]
9. Koppel R, Metlay JP, Cohen A, et al. Role of computerized physician order entry systems in facilitating medication errors. JAMA 2005 Mar 9;293(10):1197-1203. [doi: 10.1001/jama.293.10.1197] [Medline: 15755942]
10. Bond RR, Finlay DD, Nugent CD, Moore G, Guldenring D. A usability evaluation of medical software at an expert conference setting. Comput Methods Programs Biomed 2014;113(1):383-395. [doi: 10.1016/j.cmpb.2013.10.006] [Medline: 24210868]
11. IEC I 62366-1: 2015 medical devices—part 1: application of usability engineering to medical devices. : International Electrotechnical Commission (IEC); 2015 URL: https://www.iso.org/standard/63179.html#:~:text=IEC%2062366%2D1%3A2015%20specifies,as%20it%20relates%20to%20SAFETY [accessed 2025-03-25]
12. Lewis JR. Usability: lessons learned … and yet to be learned. Int J Hum Comput Interact 2014 Sep 2;30(9):663-684. [doi: 10.1080/10447318.2014.930311]
13. Sauro J. A Practical Guide to Measuring Usability: 72 Answers to the Most Common Questions about Quantifying the Usability of Websites and Software: Measuring Usability LLC; 2010. URL: https://measuringu.com/wp-content/uploads/2017/05/QuantitativeUsabilityTestOnline.pdf [accessed 2025-03-25]
14. Choi CJ, Park IH, Kim SM, Kwon JY. A study on the usability evaluation for medical devices design using heuristic evaluation. Presented at: 2019 Spring Conference of ESK; May 31 to Jun 1, 2019; Yeosu, Korea.
15. Rieman J, Franzke M, Redmiles D. Usability evaluation with the cognitive walkthrough. Presented at: CHI '95: Conference Companion on Human Factors in Computing Systems; May 7-11, 1995; Denver, Colorado, United States. [doi: 10.1145/223355.223735]
16. Liu Y, Osvalder AL, Dahlman S. Exploring user background settings in cognitive walkthrough evaluation of medical prototype interfaces: a case study. Int J Ind Ergon 2005 Apr;35(4):379-390. [doi: 10.1016/j.ergon.2004.10.004]
17. Lavery D, Cockton G. Cognitive walkthrough: usability evaluation materials. Northumbria University Library. 1997. URL: https://nrl.northumbria.ac.uk/id/eprint/26027/ [accessed 2025-01-26]
18. Wharton C, Rieman J, Lewis C, Polson P. The cognitive walkthrough: a practitioner's guide. In: Usability Inspection Methods: John Wiley; 1994.
19. Nielsen J. Usability Engineering: Academic Press; 1993.
20. Clanchy K, Mitchell J, Mulholland K, et al. Towards co-design of rehabilitation technologies: a collaborative approach to prioritize usability issues. Front Rehabil Sci 2024;5:1302179. [doi: 10.3389/fresc.2024.1302179] [Medline: 38450206]

XSL•FO
RenderX

21. Alhasani R, George N, Radman D, Auger C, Ahmed S. Methodologies for evaluating the usability of rehabilitation technologies aimed at supporting shared decision-making: scoping review. JMIR Rehabil Assist Technol 2023 Aug 15;10:e41359. [doi: 10.2196/41359] [Medline: 37581911]
22. Park HK. Medical education using standardized patients. Hanyang Med Rev 2012;32(1):35. [doi: 10.7599/hmr.2012.32.1.35]
23. Gorski S, Prokop-Dorner A, Pers M, et al. The use of simulated patients is more effective than student role playing in fostering patient-centred attitudes during communication skills training: a mixed method study. Biomed Res Int 2022;2022:1498692. [doi: 10.1155/2022/1498692] [Medline: 36573197]
24. Study on the course of smart phones GUI UX and experiences design elements. journalofdigitaldesign 2014 Apr;14(2):545-554. [doi: 10.17280/jdd.2014.14.2.051]
25. Oh YH. A study on user interface design for improving usability of the medical device. In: KSDS Conference Proceeding: Korean Society of Design Science; 2009:126-127.
26. Kim B. A study on user experience of gesture-base interface in virtual reality [Master's thesis]. : Aju University; 2019.
27. Chae SZ, Kim AJ. A study on process of medical device development and guideline of ergonomics for safety improvement. ksdc 2017 Jun 30;23(2):701-710. [doi: 10.18208/ksdc.2017.23.2.701]
28. Hanington B, Martin B. The Pocket Universal Methods of Design: 100 Ways to Research Complex Problems, Develop Innovative Ideas and Design Effective Solutions: Rockport; 2017.

Abbreviations

GUI: graphical user interface
HMD: head-mounted display
NUI: natural user interface
UI: user interface
VR: virtual reality

Edited by A Mavragani; submitted 03.11.24; peer-reviewed by B Marques, G Guerrero-Contreras, J Fielder, JH Lee; revised version received 25.01.25; accepted 13.02.25; published 01.04.25.

Please cite as:
Hong S, Choi H, Kweon H
Medical Device Based on a Virtual Reality–Based Upper Limb Rehabilitation Software: Usability Evaluation Through Cognitive Walkthrough
JMIR Form Res 2025;9:e68149
URL: https://formative.jmir.org/2025/1/e68149
doi:10.2196/68149

© Seojin Hong, Hyun Choi, Hyosun Kweon. Originally published in JMIR Formative Research (https://formative.jmir.org), 1.4.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Development of a Comprehensive Decision Support Tool for Chemotherapy-Cycle Prescribing: Initial Usability Study

Sanna Iivanainen[1], PhD, MD; Reetta Arokoski[2], MSc; Santeri Mentu[2], MSc; Laura Lang[2], MSci; Jussi Ekström[2], PhD; Henri Virtanen[2], BDSc; Vesa Kataja[2], MD, PhD; Jussi Pekka Koivunen[1], MD, PhD

[1]Department of Oncology and Radiotherapy, Oulu University Hospital, Kajaanintie 50, Oulu, Finland

[2]Kaiku Health Oy, Helsinki, Finland

Corresponding Author:
Sanna Iivanainen, PhD, MD
Department of Oncology and Radiotherapy, Oulu University Hospital, Kajaanintie 50, Oulu, Finland

Abstract

Background: Chemotherapy cycle prescription is generally carried out through a multistep manual process that is prone to human error. Clinical decision support tools can provide patient-specific assessments that support clinical decisions, improve prescribing practices, and reduce medication errors.

Objective: We hypothesized that a knowledge-based, patient-derived, evidence-directed decision support tool consisting of multiple modules focusing on the core duties preceding chemotherapy-cycle prescription could result in a more cost-effective and error-free approach and streamline the workflow.

Methods: A 1-arm, multicenter, prospective clinical trial ("Follow-up of Cancer Patients Receiving Chemotherapy or Targeted Therapy by Electronic Patient Reported Outcomes-tool" [ECHO] 7/2019-1/2021; NCT04081558) was initiated to investigate the tool. The most important inclusion criteria were the presence of colorectal cancer (CRC) treated with oxaliplatin-based chemotherapy, age ≥18 years, Eastern Cooperative Oncology Group [ECOG] performance score of 0 to 2, and internet access. A decision support tool that included digital symptom monitoring, a laboratory value interface, and treatment schedule integration for semiautomated chemotherapy cycle prescribing was integrated into the care pathway. Performance was assessed by the percentage of chemotherapy cycles with sent and completed symptom questionnaires, while perceptions of health care professionals (HCPs) on the feasibility of the approach were collected through a 1-time semistructured interview.

Results: The ECHO trial included 43 patients with CRC treated with doublet or triplet chemotherapy in an adjuvant or metastatic setting. Altogether, 843 electronic patient-reported outcome (ePRO) symptom questionnaires were completed. Of the 15 recorded symptoms, fatigue (n=446, 52.9%) and peripheral neuropathy (n=429, 50.9%) were reported most often, while 137 grade 3 to 4 symptoms were recorded, of which diarrhea (n=5, 4%) and peripheral neuropathy (n=4, 3%) were the most common. During the study, 339 chemotherapy cycles were prescribed, and for the 77% (n=262) of new chemotherapy cycles, ePRO questionnaire data were available within preset limits (completed within 3 days prior to chemotherapy scheduling) while 65% of the cycles (n=221) had symptom questionnaire grading at ≤1%, and 67% of the cycles (n=228) had laboratory values in a preset range. The recommendations by the tool for a new chemotherapy cycle were tier 1 (green; meaning "go") in 145 (42.8%) of the cycles, tier 2 (yellow; "evaluate") in 83 (25%), and tier 3 (red; "hold") in 111 (32.7%). HCPs (n=3) were interviewed with a questionnaire (comprising 8 questions), revealing that they most valued the improved workflow, faster patient evaluation, and direct messaging option.

Conclusions: In this study, we investigated the feasibility of a decision support system for chemotherapy-cycle pre-evaluation and prescription that was developed for the prospective ECHO trial. The study showed that the functionalities of the investigated tool were feasible and that an automated approach to chemotherapy-cycle prescription was possible for nearly half of the cycles.

Trial Registration: ClinicalTrials.gov NCT04081558; https://clinicaltrials.gov/study/NCT04081558

(JMIR Form Res 2025;9:e62749) doi:10.2196/62749

KEYWORDS

cancer; chemotherapy; ePRO; electronic patient-reported outcome; decision support system

Introduction

In health care, digital approaches can potentially improve accessibility, increase the comprehensiveness of care, and streamline processes in a cost-effective manner. Digital tools are especially convenient in facilitating communication independently of time and place, as well as facilitating duties requiring repeated numerical comparison [1-3]. In oncology,

XSL•FO
RenderX

electronic patient-reported outcome (ePRO)–based monitoring has gained a great deal of interest in recent years. ePROs have been shown to improve quality of life (QoL) and survival and reduce the number of unscheduled visits among patients receiving chemotherapy for advanced cancer [4-6]. Reasons for the observed beneficial effects of ePROs are not well known but might be related to urgency algorithms, long-term symptom data, better management of side effects, symptom control, patient empowerment, and facilitated communication between the patient and care team [7-11].

For cytotoxic chemotherapy, the most limiting factors for dose intensity are generally bone marrow function, neuropathy, fatigue, and nausea. Dose intensity typically correlates to treatment outcomes [12], and methods to maintain the intensity level with a personalized approach could improve treatment. For monitoring of dose-intensity limiting factors, both laboratory values and symptom profiles are required.

In oncology units, prescription of a new chemotherapy cycle is based on general laboratory values such as neutrophil counts and transaminases being in the selected reference range and the absence of severe symptoms indicating treatment side effects or cancer progression. In most units, precycle assessment is done by phone-based symptom evaluation and manual comparison of laboratory values to a safety reference. ePRO-collected symptoms can be converted to numerical data, which enables automated comparison to a safety reference scale; the quality of automated comparison surpasses manual review.

Delivering evidence-based, personalized care that involves patients requires profound changes in the traditional, hospital organization–based structure, process, and organization of care, along with clinically validated incentives to support such changes. Health care providers must engage and apply a vast amount of scientific information to provide high-quality patient care. Decision support tools can provide patient-specific assessments that support clinical decisions, improve prescribing practices, and reduce medication errors, and these tools could be part of a solution to the increasing cognitive burden currently placed on clinicians [13].

Clinical decision support systems (CDSSs) have been classified and subdivided into various categories, frequently as knowledge based or non–knowledge based. In knowledge-based systems, outputs are based on rules (IF-THEN statements) relying on literature-based, practice-based, or patient-directed evidence. Similarly, CDSSs that are non–knowledge based require a data source, but these systems use artificial intelligence (AI) methodology in work-up rather than being programmed to follow expert medical knowledge. Non–knowledge-based CDSSs, although a rapidly growing use case for AI in medicine, come with challenges including problems understanding the logic that the AI uses in creating recommendations (the so-called black box issue) and problems with validated data availability [14,15].

Taking into consideration the present hurdles in the use of non–knowledge-based CDSSs in clinical practice, we chose to develop a practice-based, ePRO-directed multidimensional platform to personalize as well as automate the chemotherapy-prescribing process. We hypothesized that generation of a decision support tool consisting of multiple modules focusing on the core duties preceding a chemotherapy-cycle prescription could be a more cost-effective and error-free approach. The functionalities of the tool would include chemotherapy cycle–based scheduling for data collection and analysis, an ePRO symptom collection–linked urgency algorithm, personalized symptom management, laboratory value interface, and decision support.

Methods

Study Design

A 1-arm, multicenter, prospective clinical trial ("Follow-up of Cancer Patients Receiving Chemotherapy or Targeted Therapy by Electronic Patient Reported Outcomes-tool" [ECHO] 7/2019-1/2021; NCT04081558) investigated the use of a novel ePRO tool, Kaiku Health, in cancer care. The most important patient inclusion criteria were colorectal cancer (CRC), a plan to receive oxaliplatin-based chemotherapy as an adjuvant therapy or in the first- or second-line setting of advanced disease, age ≥18 years, an Eastern Cooperative Oncology Group [ECOG] performance score of 0 to 2, and internet access. A decision support tool consisting of a digital symptom-monitoring tool with an urgency algorithm, laboratory values, and integration of a treatment schedule to automate chemotherapy-cycle prescribing was created for the trial.

The ePRO Tool

The Kaiku Health ePRO follow-up module consists of a questionnaire that assess both the presence and severity of symptoms related to typical side effects of chemotherapy (blood in urine, dysuria, eye symptoms [decreased vision or other symptoms], peripheral sensory neuropathy, pain, constipation, cough, decreased appetite, diarrhea, fatigue, fever, mouth sores, nausea, rash or skin changes, shortness of breath, and vomiting). The tool also allows recording the presence of a symptom and a severity algorithm that grades the symptom according to National Cancer Institute Common Terminology Criteria for Adverse Events (NCI-CTCAE). The severity algorithm triggered an email alert to the study physician in the care unit within preset limits (the presence of a grade 3 or higher symptom or a rise in symptom severity from grade 0 to 2). The patients were informed that the ePRO follow-up was intended only for nonurgent communication, and, in critical matters, patients were advised to contact emergency care. In addition, patients received tailored, evidence-based, personalized self-care advice according to electronically reported symptoms and their grade. The tool also included a messaging option through which patients could communicate with their care team directly.

Laboratory Values

Prior to chemotherapy infusion, peripheral blood samples were taken to evaluate bone marrow, kidney, and liver function. The laboratory values include counts for red and white cells, thrombocytes, alanine and aspartate aminotransferases, bilirubin, and creatinine. Based on the summary of product characteristics of the chemotherapeutic agents used by the patient, prespecified laboratory values were set to assess the fitness of the patient for a new course of cancer medical therapy.

XSL•FO
RenderX

The Decision Support Tool

We developed and added a novel element to the Kaiku Health remote symptom-monitoring tool: in addition to collecting and grading ePROs, it automatically collected laboratory values via a novel interface (MyLab). The timing of the ePRO symptom questionnaires and laboratory value collection was automated according to chemotherapy cycles through another interface with a treatment-appointment reservation module (Oberon). The decision support tool analyzed the completed symptom questionnaires and preceding laboratory values and compared the results to a set of institutional reference values. The decision support tool provided a 3-tier recommendation to the care team on the initiation of the preplanned chemotherapy cycle. The 3-tier recommendations included green (meaning “go”), yellow (“evaluate and possibly postpone”), or red (“postpone”). A recommendation of go versus no-go and potential treatment visit details were sent to the patient through the Kaiku Health tool. With a no-go decision, an automated ePRO questionnaire and laboratory value collection for the rescheduled chemotherapy cycle were initiated based on the reservation module interface.

Ethical Considerations

The study was approved by the ethics committee of the Northern Ostrobothnia Hospital District (273/2017) and registered in an international clinical study registry (ClinicalTrials.gov NCT04081558). The study was carried out in accordance with the principles of the Declaration of Helsinki and Good Clinical Practice. All patients provided written informed consent. Study participants were not compensated for their participation in the trial.

Results

Development of the Decision Support Tool

Before a new chemotherapy cycle, a patient’s laboratory values and symptoms were evaluated, for example, for indications of treatable chemotherapy side effects, excess chemotherapy toxicity, and cancer progression. To streamline this work process, increase the comprehensiveness and repeatability of the precycle evaluation, and automate duties, we developed a decision support tool. The designed tool had 6 different elements that aimed to increase the efficiency and quality of cancer care (Figure 1; Multimedia Appendix 1).

XSL•FO
RenderX

Figure 1. Flowchart of the methodological process of the integration of electronic patient-reported outcomes (ePROs) and laboratory values. The tool included a scheduling module that synchronized the sending of symptom questionnaires to patients via Kaiku Health and laboratory value collection from the hospital's electronic health record (EHR) via an interface. The tool also included a symptom module with an ePRO-based questionnaire for 15 core symptoms. An evaluation module compared the symptoms and laboratory values to values in a set range and assigned a recommendation in 3 tiers (1=go; 2=hold; 3=postpone) for the chemotherapy cycle. HCP: health care professional.

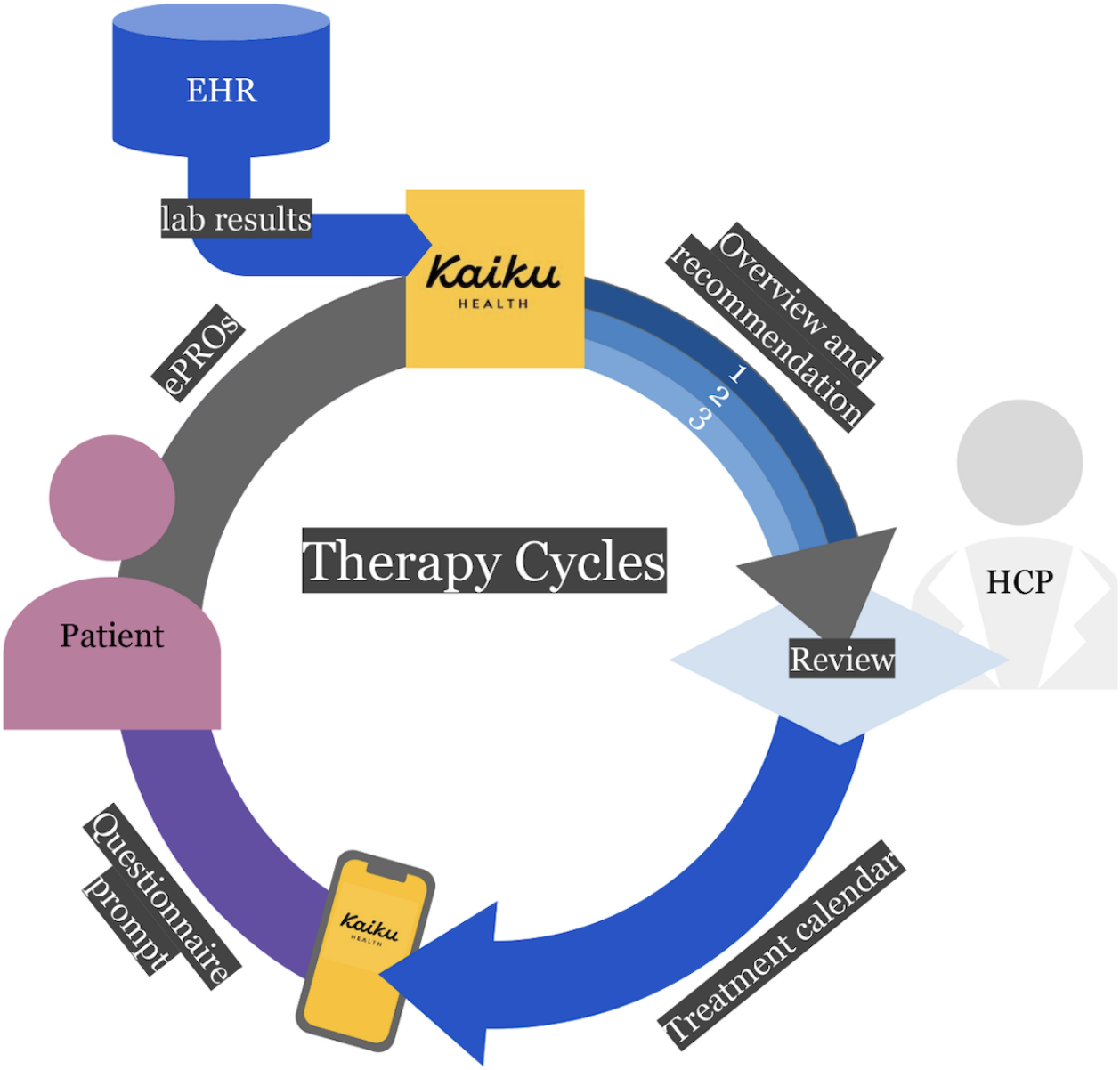

The first element was a scheduling module that synchronized the sending of symptom questionnaires to patients via Kaiku Health and laboratory value collection from hospital registries via an interface. The symptom module used ePRO-based (Kaiku Health) grading (NCI-CTCAE) for 15 core symptoms. The laboratory-value module collected specific central values via a software interface (Figure 2A). The evaluation module compared the symptoms and laboratory values to values in a set range and assigned a recommendation for the chemotherapy cycle in 1 of 3 tiers (Figure 2B). For the tier-1 (green/go) recommendation, laboratory values needed to be within a set range and all the symptoms needed to be at grade ≤1; thus, the nurse made the planned chemotherapy order to the hospital pharmacy, and an SMS text message with information on the timetable of drug infusion was sent through Kaiku Health. For the tier-2 (yellow/evaluate) recommendation, laboratory values had to be within a set range and ePROs had to include ≥1 symptom that was grade ≥2 (if they were not already at baseline); the nurse individually assessed the symptoms. If there was a need for cancer drug modification, typically dose reduction or supportive medications, the study physician was consulted as to whether the chemotherapy cycle could be initiated or modified, or if supportive medication should be initiated or modified. In the case of a "go" decision, confirmation of the treatment time was sent via Kaiku Health. In the tier-3 (red/hold) scenario, laboratory values were out of the set range, and the nurse postponed the chemotherapy cycle and sent the new infusion date through Kaiku Health. In the case of a postponed chemotherapy cycle, the scheduling module programmed a new

XSL•FO
RenderX

collection of symptoms and laboratory values according to the new infusion date.

Figure 2. Overview of the core functionalities of the decision support tool. (A) Electronic patient-reported outcome symptom questionnaire. (B) Laboratory integration module and its output. (C) Decision support tool recommendation.

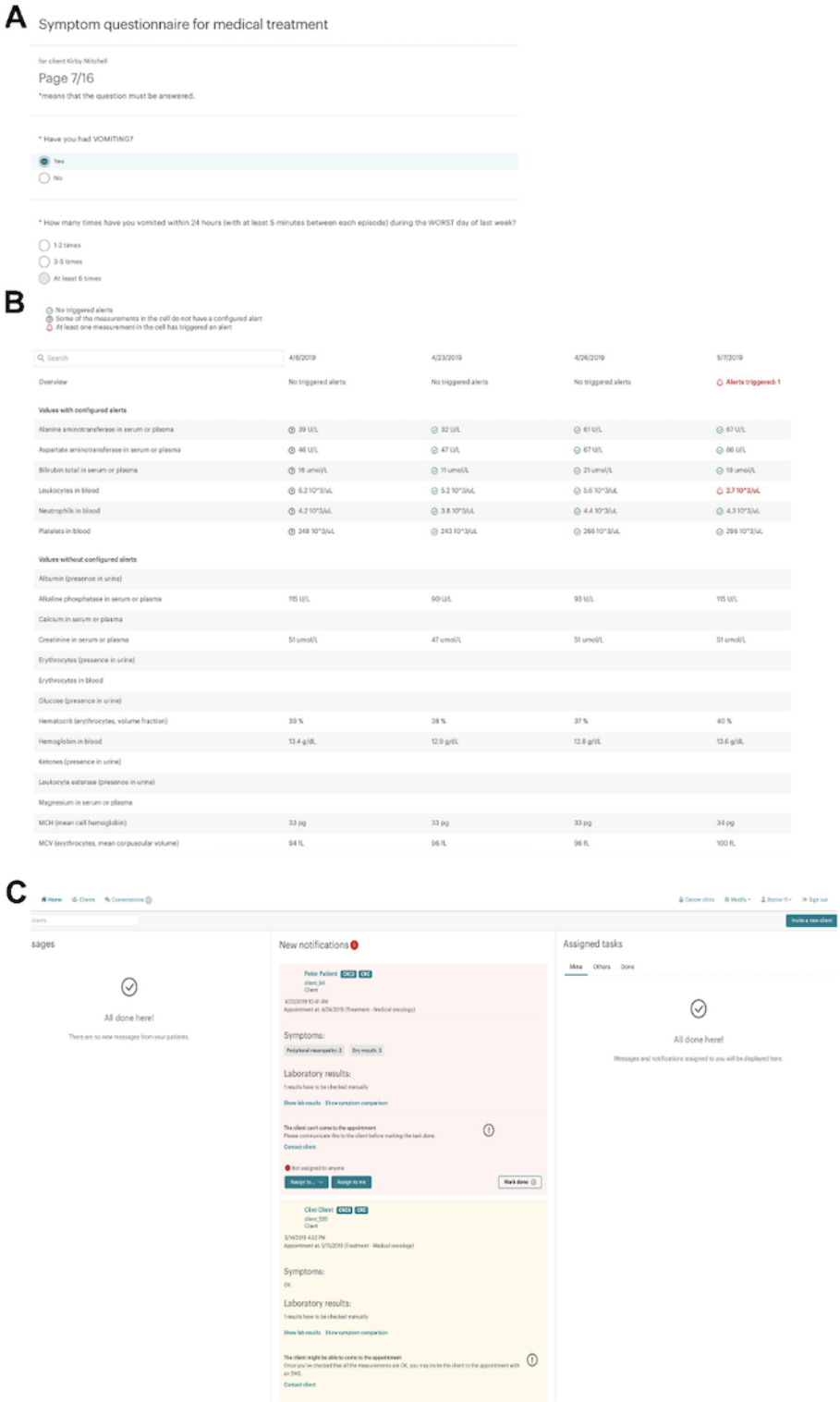

The care team could always override the recommendation of the decision support system. In the case of missing ePRO or laboratory data (or if the data were >3 days old) before the chemotherapy cycle, the decision support tool gave a tier-2 (yellow/evaluate) recommendation; thus, manual evaluation had to be done prior to prescribing the treatment. The SMS messaging module was used to inform the patient about the recommendation and potential treatment decision. Furthermore, the messaging module enabled secure messaging between the patient and care team. The symptom support module also provided the patient with personalized guidance for self-care based on the presence of specific symptoms and their grade (Figure 2C).

XSL•FO
RenderX

Performance of Decision Support Tool–Directed Chemotherapy Scheduling

Next, we analyzed the performance of the decision support tool based on data collected in a prospective, multicenter, single-arm clinical trial (ECHO; NCT04081558). Patient recruitment for the prospective cohort took place from July 2019 to January 2021 at the Oulu University Hospital and Vaasa Central Hospital oncology units. The study dataset included patients with CRC (n=43) in the ECHO trial who gave informed consent and were treated with doublet or triplet chemotherapy in an adjuvant or metastatic setting. During the study, a total of 339 chemotherapy cycles were prescribed. Altogether, 843 ePRO symptom questionnaires were completed during the ECHO study. Of the recorded 15 symptoms, fatigue (n=446, 52.9%) and peripheral neuropathy (n=429, 50.9%) were reported most often, while 137 grade 3 or 4 symptoms were recorded, of which diarrhea (n=5, 4%) and peripheral neuropathy (n=4, 3%) were the most common. Since the functioning of the Kaiku Health symptom tool has been evaluated in multiple previous trials [16-18], we focused here on the performance of the scheduling and evaluation modules.

The function of the scheduling module was assessed by evaluating the percentage of chemotherapy cycles with sent and completed symptom questionnaires. There were no false-positive or false-negative decision support tool recommendations. For over two-thirds (n=262, 77%) of the cycles, ePRO questionnaire data were available. Preceding a new chemotherapy cycle, 65% (n=221) of cycles had symptom questionnaires graded at ≤1, while 67% (n=228) of the cycles had laboratory values in the preset range. Evaluation module recommendations were tier 1 (green/go) in 145 (42.8%) of the cycles, tier 2 (yellow/evaluate) in 83 (25%), and tier 3 (red/hold) in 111 (32.7%), while 61% of the patients (n=27) needed a phone call before any of the planned chemotherapy cycles (Table 1).

Table . Performance metrics of the developed decision support tool for automated chemotherapy-cycle prescription. The feasibility of the scheduling module was assessed by evaluating the percentage of chemotherapy cycles with sent and completed symptom questionnaires and the availability of laboratory work. The 3-tier recommendation system operated as follows: tier 1 indicated "go" (the nurse proceeded with the initial treatment plan); tier 2 indicated "hold" (the laboratory values had to be within a set range, while the reported symptoms [at least one] were grade ≥2; the study physician was consulted); and tier 3 indicated "postpone" (electronic patient-reported outcomes were acceptable, but laboratory values were outside the preset limits; the nurse scheduled the laboratory work and postponed the infusion).

Criteria	Cycles (n=339), n (%)
Was an automated symptom questionnaire sent and completed prior to the chemotherapy cycle?	
Yes	262 (77.3)
No	77 (23)
Was the symptom questionnaire grading in the acceptable range prior to the planned chemotherapy cycle?	
Yes	221 (65.2)
No	118 (34.8)
Was the laboratory work in the acceptable range prior to the planned chemotherapy cycle?	
Yes	228 (67.3)
No	111 (32.7)
Evaluation module guidance (tier)	
1	145 (42.8)
2	83 (25)
3	111 (32.7)

The experience of the health care professionals (HCPs) with the decision support system integration into the care pathway was evaluated via interviews (n=4) that explored 5 topics: improved workflow, improved care, use of resources, challenges, and points for improvement. The HCPs appreciated the improved workflow, faster patient evaluation, and direct messaging option. The laboratory module originally collected data only daily, which was considered too infrequent by the HCPs, and data collection was reprogrammed to occur multiple times a day, which improved the usability of the tool. Some inconveniences were experienced with the scheduling module for patients with Baxter pump–based treatments, for whom 2 subsequent treatment reservations were made at once (treatment and pump removal); this led to incorrect scheduling of the symptom questionnaires. The HCPs felt that the messaging option should include read notifications, which would enable verification that the patient was aware of the chemotherapy schedule.

Discussion

Principal Findings

We present the results of a feasibility study investigating an integrated approach with an automated decision support system for chemotherapy-cycle prescription. We show that the investigated multidimensional integrated system is feasible and

XSL•FO
RenderX

can provide clinically meaningful decision support. The approach could decrease the HCP workload and increase the quality and safety of cancer care.

Comparison to Prior Work

In recent years, the use of ePRO symptom data has gained a great deal of interest in the oncology field. Previous studies have shown that ePRO monitoring can reduce emergency room visits, improve QoL, and, more importantly, increase survival [4-6]. Integration of an ePRO system into electronic patient records is considered essential for large-scale adoption of the approach. Timely integration of data from multiple electronic health care systems into ePROs could further enhance the clinical value of the captured data and facilitate decision support.

AI-based approaches can be superior for automatic detection, classification, and interpretation using big data. However, AIs can support improper decisions if they are presented with data beyond their original training set data. One could speculate that AI-based decision support systems could surpass our approach with a simple comparison to reference values. However, chemotherapy has a narrow therapeutic window and relying on prescription decisions made by AI alone may not satisfy ethical requirements without wide-scale training and validation [19-21].

A fast track for AI integration in clinical practice could lie in the use of large language models (LLMs). Methodological developments in LLMs such as in-context learning and retrieval-augmented generation could improve LLM performance and reduce hallucinations, consequently making the use of LLMs possible in clinical practice. It has been suggested that the recent development of advanced LLMs could improve patient care across several areas, such as clinical decision support or helping to answer patients' questions, and could even improve patient outcomes by identifying social determinants of health and reducing health disparities [22-24].

Strengths and Limitations

In this study, we investigated a next-generation approach aiming to streamline and automate the prescription of chemotherapy cycles. We identified symptom profiles and laboratory values as central decision mediators for safe chemotherapy prescription. We hypothesized that symptom profiles could be collected via ePROs in a more comprehensive manner and used as input for a decision support tool that would also collect laboratory data and schedule collections based on the expected cycle date in an automated fashion. Furthermore, our approach included an expanded ePRO module with an integrated urgency algorithm, symptom-based personalized feedback and self-care advice, and a safe messaging platform.

We were able to show that the investigated platform is technically feasible, with good user experiences among patients and HCPs. More importantly, we found that automated "go" decisions for new chemotherapy cycles were made in almost half of the chemotherapy cycles. It is likely that this proportion could be increased by focusing only on core symptoms, instead of the total of 17 symptoms that were recorded in this study. We speculate that the investigated tool would decrease the workload for HCPs and enable allocation of human resources to more value-creating duties. We strongly believe that the computerized comparison of data to a set reference is more precise and error-free than human assessment, even though our study did not provide any metrics on the specific matter.

Nevertheless, a few modifications to the platform were suggested by end users. First, in the testing phase, the frequency of laboratory data collection was already considered too low; thus, the laboratory module was reprogrammed to search for data multiple times a day, which greatly improved the usability of the subtool. Second, the HCPs felt that the messaging option regarding the schedule of the planned chemotherapy cycle should include read notifications, which would enable verification that the patient was aware of the coming cycle. Furthermore, due to the low number of study participants and the observational nature of the study, the results are to be interpreted with caution, and the generalizability of the feasibility of the studied platform is unclear.

Future Directions

Due to the feasibility design of the study, the results are generally hypothesis generating. Since the study investigated a complex technical tool, initial testing with a feasibility approach is therefore warranted before progressing to a larger clinical study. The integration of an LLM-based module for interactivity, such as a chat box, could enhance the supportive role of the developed tool by providing educational elements to empower patients as well as give guidance on milder-grade chemotherapy toxicity management on a more individual level. Furthermore, the rather large proportion of symptom questionnaires with grade 2 symptoms might be decreased by fine-tuning the symptom evaluation, possibly by using AI-based models to mimic diagnostic pattern recognition to identify red-flag symptoms in parallel with laboratory value assessment. To conclude, the results of our trial are promising and lay the groundwork for further studies as well as development of the platform. A randomized study design would be required to fully validate the clinical utility and cost-effectiveness of the system.

Acknowledgments

Generative artificial intelligence tools were not used in preparation of this manuscript. The study was funded by Roche Diagnostics. Kaiku Health employees were involved in data acquisition and analysis.

Data Availability

The datasets generated and/or analyzed during the current study are not publicly available but are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Authors' Contributions

SI, JPK, and RA are the guarantors of integrity of the entire study. SI, JPK, and RA contributed to study conceptualization and design. SI and JPK performed the literature search. SI, JPK, LL, JE, HV, and VK contributed to the clinical studies. SI, JPK, RA, SM, LL, and JE contributed to the experimental studies and data analysis. SI, JPK, SM, and JE performed the statistical analysis. SI, JPK, RA, SM, JE, LL, HV, and VK prepared the manuscript. SI, JPK, RA, SM, JE, LL, HV, and VK edited the manuscript. All authors read and approved the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Kaiku Bridge combines the Kaiku Health symptom data from the electronic patient-reported outcome platform and laboratory value data from the hospital records (Effica) according to the scheduled care events (Oberon). After evaluating both symptoms and laboratory values within predefined limits, the Kaiku interface provides guidance in 3 tiers on new chemotherapy-cycle prescription.
[PNG File, 89 KB - formative_v9i1e62749_app1.png]

References

1. Holch P, Warrington L, Bamforth LCA, et al. Development of an integrated electronic platform for patient self-report and management of adverse events during cancer treatment. Ann Oncol 2017 Sep 1;28(9):2305-2311. [doi: 10.1093/annonc/mdx317] [Medline: 28911065]
2. Kotronoulas G, Kearney N, Maguire R, et al. What is the value of the routine use of patient-reported outcome measures toward improvement of patient outcomes, processes of care, and health service outcomes in cancer care? A systematic review of controlled trials. J Clin Oncol 2014 May 10;32(14):1480-1501. [doi: 10.1200/JCO.2013.53.5948] [Medline: 24711559]
3. Lizée T, Basch E, Trémolières P, et al. Cost-effectiveness of web-based patient-reported outcome surveillance in patients with lung cancer. J Thorac Oncol 2019 Jun;14(6):1012-1020. [doi: 10.1016/j.jtho.2019.02.005] [Medline: 30776447]
4. Basch E, Deal AM, Dueck AC, et al. Overall survival results of a trial assessing patient-reported outcomes for symptom monitoring during routine cancer treatment. JAMA 2017 Jul 11;318(2):197-198. [doi: 10.1001/jama.2017.7156] [Medline: 28586821]
5. Denis F, Lethrosne C, Pourel N, et al. Randomized trial comparing a web-mediated follow-up with routine surveillance in lung cancer patients. J Natl Cancer Inst 2017 Sep 1;109(9):9. [doi: 10.1093/jnci/djx029] [Medline: 28423407]
6. Denis F, Yossi S, Septans AL, et al. Improving survival in patients treated for a lung cancer using self-evaluated symptoms reported through a web application. Am J Clin Oncol 2017 Oct;40(5):464-469. [doi: 10.1097/COC.0000000000000189] [Medline: 25811297]
7. Osborn J, Ajakaiye A, Cooksley T, Subbe CP. Do mHealth applications improve clinical outcomes of patients with cancer? A critical appraisal of the peer-reviewed literature. Support Care Cancer 2020 Mar;28(3):1469-1479. [doi: 10.1007/s00520-019-04945-4] [Medline: 31273501]
8. Moradian S, Voelker N, Brown C, Liu G, Howell D. Effectiveness of internet-based interventions in managing chemotherapy-related symptoms in patients with cancer: a systematic literature review. Support Care Cancer 2018 Feb;26(2):361-374. [doi: 10.1007/s00520-017-3900-8] [Medline: 28948360]
9. Trajkovic-Vidakovic M, de Graeff A, Voest EE, Teunissen S. Symptoms tell it all: a systematic review of the value of symptom assessment to predict survival in advanced cancer patients. Crit Rev Oncol Hematol 2012 Oct;84(1):130-148. [doi: 10.1016/j.critrevonc.2012.02.011] [Medline: 22465016]
10. Bennett AV, Jensen RE, Basch E. Electronic patient-reported outcome systems in oncology clinical practice. CA Cancer J Clin 2012;62(5):337-347. [doi: 10.3322/caac.21150] [Medline: 22811342]
11. Benze G, Nauck F, Alt-Epping B, et al. PROutine: a feasibility study assessing surveillance of electronic patient reported outcomes and adherence via smartphone app in advanced cancer. Ann Palliat Med 2019 Apr;8(2):104-111. [doi: 10.21037/apm.2017.07.05] [Medline: 29156896]
12. Lyman GH. Impact of chemotherapy dose intensity on cancer patient outcomes. J Natl Compr Canc Netw 2009 Jan;7(1):99-108. [doi: 10.6004/jnccn.2009.0009] [Medline: 19176210]
13. Walsh S, de Jong EEC, van Timmeren JE, et al. Decision support systems in oncology. JCO Clin Cancer Inform 2019 Feb;3(3):1-9. [doi: 10.1200/CCI.18.00001] [Medline: 30730766]
14. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med 2020;3:17. [doi: 10.1038/s41746-020-0221-y] [Medline: 32047862]
15. Nafees A, Khan M, Chow R, et al. Evaluation of clinical decision support systems in oncology: an updated systematic review. Crit Rev Oncol Hematol 2023 Dec;192:104143. [doi: 10.1016/j.critrevonc.2023.104143] [Medline: 37742884]

XSL•FO
RenderX

16. Iivanainen S, Alanko T, Peltola K, et al. ePROs in the follow-up of cancer patients treated with immune checkpoint inhibitors: a retrospective study. J Cancer Res Clin Oncol 2019 Mar;145(3):765-774. [doi: 10.1007/s00432-018-02835-6] [Medline: 30666409]
17. Iivanainen S, Alanko T, Vihinen P, et al. Follow-up of cancer patients receiving anti-PD-(L)1 therapy using an electronic patient-reported outcomes tool (KISS): prospective feasibility cohort study. JMIR Form Res 2020 Oct 28;4(10):e17898. [doi: 10.2196/17898] [Medline: 33112242]
18. Iivanainen SME, Ravi R, Jekunen A, et al. ePRO follow-up of colorectal cancer patients receiving oxaliplatin-based adjuvant chemotherapy is feasible and enhances the quality of patient care. Ann Oncol 2021 Sep;32:S1184-S1185. [doi: 10.1016/j.annonc.2021.08.1667] [Medline: 36809503]
19. Castaneda C, Nalley K, Mannion C, et al. Clinical decision support systems for improving diagnostic accuracy and achieving precision medicine. J Clin Bioinforma 2015;5(1):4. [doi: 10.1186/s13336-015-0019-3] [Medline: 25834725]
20. Shortliffe EH, Sepúlveda MJ. Clinical decision support in the era of artificial intelligence. JAMA 2018 Dec 4;320(21):2199-2200. [doi: 10.1001/jama.2018.17163] [Medline: 30398550]
21. Simon G, DiNardo CD, Takahashi K, et al. Applying artificial intelligence to address the knowledge gaps in cancer care. Oncologist 2019 Jun;24(6):772-782. [doi: 10.1634/theoncologist.2018-0257] [Medline: 30446581]
22. Iivanainen S, Lagus J, Viertolahti H, Sippola L, Koivunen J. Investigating large language model (LLM) performance using in-context learning (ICL) for interpretation of ESMO and NCCN guidelines for lung cancer. J Clin Oncol 2024 Jun 1;42(16_suppl):e13637-e13637. [doi: 10.1200/JCO.2024.42.16_suppl.e13637]
23. Chen D, Huang RS, Jomy J, et al. Performance of multimodal artificial intelligence chatbots evaluated on clinical oncology cases. JAMA Netw Open 2024 Oct 1;7(10):e2437711. [doi: 10.1001/jamanetworkopen.2024.37711] [Medline: 39441598]
24. Guevara M, Chen S, Thomas S, et al. Large language models to identify social determinants of health in electronic health records. NPJ Digit Med 2024 Jan 11;7(1):6. [doi: 10.1038/s41746-023-00970-0] [Medline: 38200151]

Abbreviations

AI: artificial intelligence
CDSS: clinical decision support system
CRC: colorectal cancer
ECHO: Follow-up of Cancer Patients Receiving Chemotherapy or Targeted Therapy by Electronic Patient Reported Outcomes-tool
ECOG: Eastern Cooperative Oncology Group
ePRO: electronic patient-reported outcome
HCP: health care professional
LLM: large language model
NCI-CTCAE: National Cancer Institute Common Terminology Criteria for Adverse Events
QoL: quality of life

Edited by A Mavragani; submitted 30.05.24; peer-reviewed by JCL Chow, SS Goswami; revised version received 12.02.25; accepted 13.02.25; published 31.03.25.

Please cite as:
Iivanainen S, Arokoski R, Mentu S, Lang L, Ekström J, Virtanen H, Kataja V, Koivunen JP
Development of a Comprehensive Decision Support Tool for Chemotherapy-Cycle Prescribing: Initial Usability Study
JMIR Form Res 2025;9:e62749
URL: https://formative.jmir.org/2025/1/e62749
doi:10.2196/62749

© Sanna Iivanainen, Reetta Arokoski, Santeri Mentu, Laura Lang, Jussi Ekström, Henri Virtanen, Vesa Kataja, Jussi Pekka Koivunen. Originally published in JMIR Formative Research (https://formative.jmir.org), 31.3.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Barriers and Determinants of Referral Adherence in AI-Enabled Diabetic Retinopathy Screening for Older Adults in Northern India During the COVID-19 Pandemic: Mixed Methods Pilot Study

Anshul Chauhan[1], PhD; Anju Goyal[2], PhD; Ritika Masih[1], MPH; Gagandeep Kaur[1], B Optometry; Lakshay Kumar[1], B Optometry; Neha[1], B Optometry; Harsh Rastogi[1], MSc; Sonam Kumar[1], MPH; Bidhi Lord Singh[1], PhD; Preeti Syal[2], MS; Vishali Gupta[1], MS; Luke Vale[3], PhD; Mona Duggal[1,4], MD

[1]Advanced Eye Centre, Post Graduate Institute of Medical Education and Research, Number 139, Kairon Block, Chandigarh, India

[2]Department of Statistics, Panjab University, Chandigarh, India

[3]Newcastle University, Newcastle, United Kingdom

[4]National Institute for Research in Digital Health and Data Sciences, Indian Council of Medical Research, New Delhi, India

Corresponding Author:
Mona Duggal, MD
Advanced Eye Centre, Post Graduate Institute of Medical Education and Research, Number 139, Kairon Block, Chandigarh, India

Abstract

Background: Diabetic retinopathy (DR) is a leading cause of blindness globally. DR has increasingly affected both individuals and health care systems as the population ages.

Objective: This study aims to explore factors and identify barriers associated with nonadherence to referral recommendations among older adult participants after DR screening (DRS) during the COVID-19 pandemic.

Method: This paper presents findings from a pilot study on artificial intelligence–enabled DRS conducted in two districts in Punjab, India (Moga and Mohali) during the COVID-19 pandemic. The screenings were conducted from March to June 2022 at community health center Badhani Kalan in Moga and from March to June 2021 in community settings (homes) in Block Boothgarh, Mohali. Participants were referred to the district hospital for an ophthalmological review based on artificial intelligence–enabled screening. After 1 month, the participants were contacted by telephone to assess adherence to the referral recommendations. Participants who did not adhere to the referral were then interviewed alongside health care providers to understand the barriers explaining their nonadherence.

Results: We aimed to recruit 346 and 600 older adult participants from 2 sites but enrolled 390. Key challenges included health facility closures due to COVID-19, low motivation among health personnel for recruitment, incomplete nonparticipation data, and high participant workloads. Approximately 45% of the participants were male and 55% female. Most participants (62.6%) were between 60 and 69 years old, while 37.4% were 70 or older, with a mean age of 67.2 (SD 6.2) years. In total, 159 participants (40.8%) were referred, while 231 participants (59.2%) were not. Only 23 (14.5%) of those referred followed through and visited a health facility for ophthalmological review, while 136 (85.5%) did not pursue further evaluation. Our analysis revealed no significant differences in the characteristics between adherent and nonadherent participants, suggesting that demographic and health factors alone do not predict adherence behavior in patients with DR. Interviews identified limited knowledge about DR, logistical challenges, financial constraints, and attitudinal barriers as the primary challenges.

Conclusions: This study, conducted during the COVID-19 pandemic, showed suboptimal adherence to referral recommendations among older adult patients due to knowledge gaps, logistical challenges, and health system issues. Quantifying and understanding adherence factors are crucial for targeted interventions addressing barriers to referral recommendations after DRS. Integrating teleophthalmology into and strengthening infrastructure for artificial intelligence–enabled diabetic retinopathy screening to enhance access and outcomes.

(JMIR Form Res 2025;9:e67047) doi:10.2196/67047

KEYWORDS

diabetic retinopathy; diabetes; gerontology; geriatric; old; aging; aged; artificial intelligence; retinopathy; retinal; referral; screening; optometry; ophthalmology; adherence; barriers

XSL•FO
RenderX

Introduction

Diabetic retinopathy (DR) is a prevalent and severe microvascular complication of diabetes mellitus (DM) [1]. In India, 10.9% (7.2% - 16.3%) of individuals aged 65 years and above with diabetes have DR, with 2.3% (1.2% - 4.4%) suffering from vision-threatening diabetic retinopathy (VTDR) [1], characterized by severe retinopathy or macular edema [2]. DR is typically asymptomatic in its early stages, and it can lead to visual impairment or blindness if left untreated [3]. The rate of blindness due to VTDR is expected to rise proportionately to the exponential increase in DM prevalence [4,5]. Visual impairment due to DR increases by 9.6% in individuals over 50 [6], emphasizing advancing age as a significant risk factor [7]. Individuals with VTDR face greater difficulties in vision-related tasks, such as reading, watching television, and driving [8]. Additionally, people with DR have a higher risk of falls, resulting in reduced mobility, social isolation, and increased physical dependence [9].

India is set to become home to the world's second-largest older adult population, with the number of people over 60 years of age projected to rise from 100 million in 2013 to 198 million by 2030 [10]. The National Programme for Health Care of the Elderly in India advocates for accessible, affordable, high-quality, long-term, comprehensive care services tailored to the needs of an aging population [10]. Significant scientific evidence shows that early screening and timely treatment referral can prevent most visual loss caused by DR [11]. Conventionally, DR screening (DRS) includes fundus (retina) examination by ophthalmologists or color fundus photography using conventional cameras (mydriatic or nonmydriatic) conducted by trained eye technicians or optometrists [12]. However, the sharp rise in diabetes cases, coupled with a shortage of trained retinal specialists and ophthalmologists [13], makes DRS services accessibility challenging, particularly as 80% of India's older adult population resides in rural areas [10]. The COVID-19 pandemic posed additional significant challenges to health care systems worldwide, inevitably leading to the curtailment of health services accessibility, including DRS [14,15].

Automated computer-based analysis of fundus images could ease the strain on health care systems by streamlining DRS and offering a more efficient management solution [13]. Artificial intelligence (AI) systems have reduced costs, enhanced diagnostic precision, and expanded patient accessibility to DRS services [16] and immediate AI-supported feedback on referral status was linked to higher referral adherence [17].

Previous studies have demonstrated telemedicine-based DRS programs' effectiveness, scalability, and sustainability in improving accessibility and reducing disparities [18]. The COVID-19 pandemic further emphasized the need for AI-enabled DRS and teleconsultations to standardize workflows and reduce human interactions to mitigate transmission risks [19,20].

These studies are particularly relevant for designing and implementing effective AI-enabled DRS in India, where challenges of health care accessibility and resource constraints exist [20]. Suboptimal adherence to follow-up recommendations impacts treatment outcomes and can undermine even the most effective treatment options [21]. A multicenter analysis showed a decline in laser procedures and vitrectomies for retinal detachment repair due to reduced clinic visits during the pandemic [22]. Patients faced barriers to follow-up eye exams, including transportation issues, high costs, and long wait times [23]. Programs such as a point-of-care DR examination program highlight the potential of telemedicine in reducing barriers to DRS, though adherence to follow-up recommendations after screening remains a major challenge [24].

However, some factors (individual, health) influencing adherence and nonadherence to referral services remain unknown [25]. During the COVID-19 pandemic, patients were prompted to limit hospital visits, minimize time spent there, and reduce direct contact with health care providers (HCPs). Limited research exists in India on older adults' adherence to DRS referral recommendations and the barriers influencing their decisions.

This study aims to (1) evaluate the referral adherence rates among older adult participants after AI-enabled DRS and identify the key factors affecting adherence during the COVID-19 pandemic and (2) identify and analyze the barriers to adherence to referral recommendations during the COVID-19 pandemic. The overall study design is described in Figure 1.

XSL•FO
RenderX

Figure 1. Overall study design. AI: artificial intelligence; DRS: diabetic retinopathy screening.

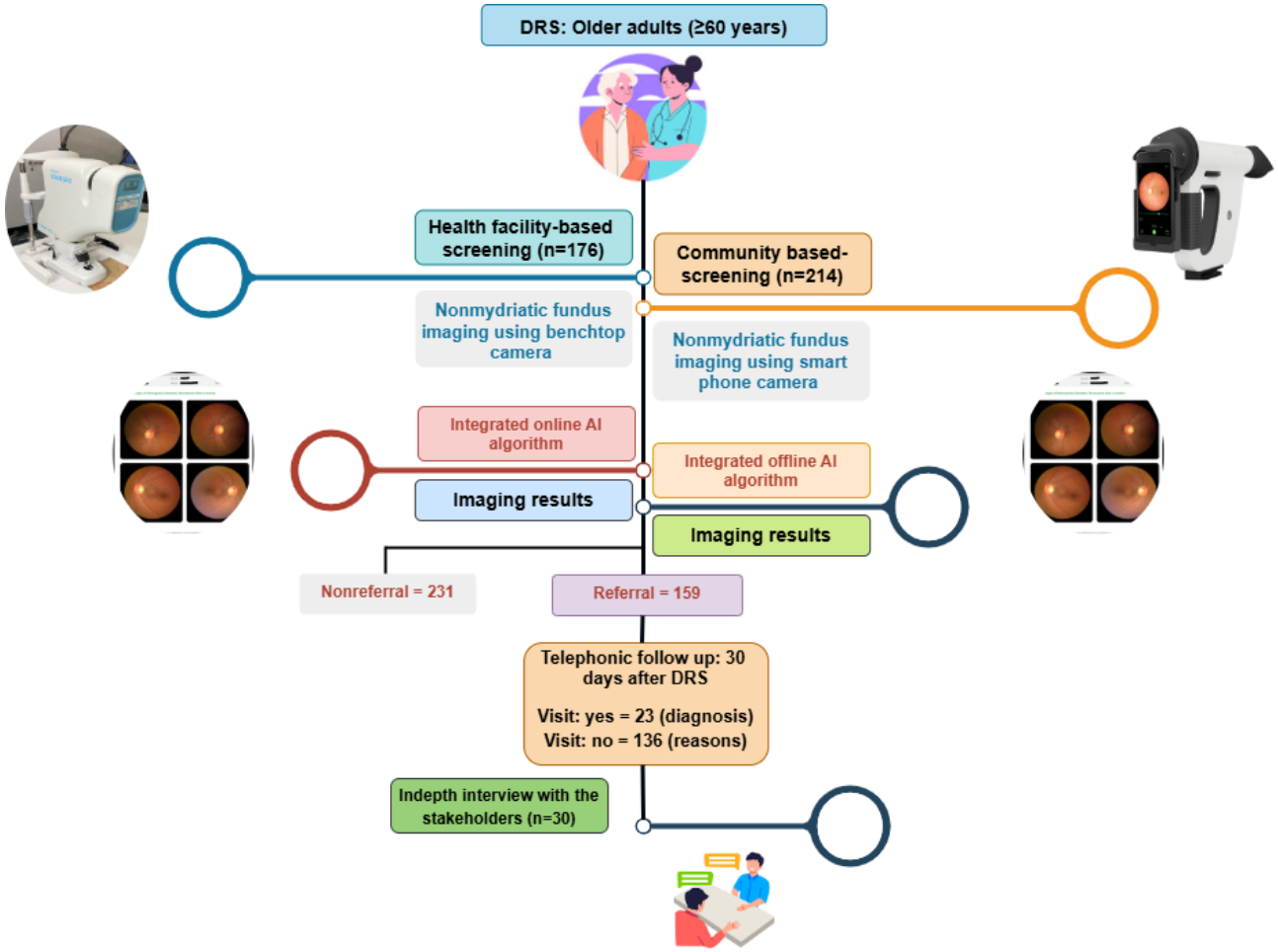

Methods

Study Setting and Participants

Quantitative

This paper presents findings from 2 districts in Punjab, India: Moga and Mohali. Screenings in Moga occurred at the community health center (CHC) Badhani Kalan from March to June 2022, while in Mohali, they were conducted in community settings (home) in Block Boothgarh from March to June 2021. The study included older adults [10] aged 60 years and above with a history of DM at both sites. Participants without a history of DM, those unwilling to provide informed consent, and those with a history of intraocular surgery, conjunctivitis, red eye, injury, or any eye inflammation were excluded.

A nurse at the medicine clinic at the CHC, Badhani Kalan, in district Moga, identified the elderlyolder adult participants with DM for DRS. In district Mohali, the list of diabeticparticipants with diabetes was obtained from thePa primary Hhealth Ccentere (PHC), in Boothgarh, and approached with the help of the village's Accredited Social and Health Activists (ASHA) worker. The participants agreeing to undergo DRS received an appointment card, and a reminder telephone call was placed a day before the screening to confirm their attendance at home.

The repeated lockdowns during the COVID-19 pandemic disrupted health care services, significantly increasing the burden on health systems. Consequently, recruitment efforts were hindered, making it difficult to reach the target sample sizes of 348 at CHC Badhani Kalan and 600 at Block Boothgarh. Since the populations shared similar health systems and implementation challenges, the data from both sites were pooled and presented. Consequently, data pooling enhanced sample size and statistical power and effectively addressed research questions [26].

Qualitative

The in-depth interviews were conducted with 2 distinct categories of participants: older adults and HCPs. Older adult participants who did not adhere to the recommended referral instructions were purposely sampled and approached for in-depth interviews to understand their barriers to referral adherence after DRS. The nurse at the ophthalmology clinic invited older adult participants to the CHC for in-depth interviews, while ASHA workers contacted older adult participants in the community to conduct interviews at homes.

The second group of participants, HCPs, were purposely selected according to their roles in diabetes and DR diagnosis, referral, and treatment (Table S1 in Multimedia Appendix 1). The inclusion criteria were having at least 3 years of experience working in public health care settings and a willingness to describe patient experiences and barriers to referral adherence (Table S2 in Multimedia Appendix 1). The HCPs included ASHAs from the community settings [27], Community Health

XSL•FO
RenderX

Officers (CHOs) from health and wellness centers [28], optometrists from a CHC [29] , ophthalmologists from the district hospital [30], and retina specialists from the retina unit of a tertiary eye care hospital.

Semistructured interviews were conducted to explore participants perspectives on the barriers to adhering to referral recommendations following AI-enabled DRS. The in-depth interview guides were developed in English, informed by existing literature on barriers to DR screening and referral [31-34], and later translated into Hindi and Punjabi (Table S3 in Multimedia Appendix 1). These guides, featuring open-ended questions and probes, explored participants' perspectives on referral adherence after DRS. Pilot interviews with 2 people with DM, an ASHA, and a CHO led to revisions for clarity, replacing ambiguous terms with simpler language for better participant understanding.

Ethical Considerations

Written informed consent was obtained from all participants, with both studies approved by the Institutional Ethics Committee of the Post Graduate Institute of Medical Education and Research, Chandigarh, India (PGI/IEC/2020/000741 and PGI/IEC/2020/001342). To ensure patient privacy and confidentiality, all images and associated metadata were de-identified prior to analysis. The studies adhered to the Declaration of Helsinki guidelines, and the study protocol is registered with the Clinical Trials Registry, India (CTRI 2022/10/046283 and 2022/10/046185).

Hardware and Software for DRS

Hardware

The data were collected from 2 locations across districts in Punjab, each employing a different DRS methodology. Health facility–based screening at CHCs requires desktop nonmydriatic fundus cameras, while community screenings (home) rely on portable cameras for enhanced accessibility.

Fundus Benchtop Camera

The screening at the CHC Badhani Kalan utilized the Forus 3 Nethra classic (benchtop) camera, supported by a laptop running MS Windows 10 or higher, featuring a 64-bit operating system and an i3 10th-generation processor [35,36].

Smartphone-Based Fundus Camera

The lightweight, portable Fundus on Phone Non Mydriatic 10 (Remidio) handheld device, with a power backup, was used for screening in community settings in Block Boothgarh [37].

Software

Both sites used validated AI screening algorithms for DRS. At the CHC, an online DRS algorithm (Revelo) was integrated into the benchtop fundus camera [38]. A camera-integrated offline AI algorithm (Medios) was used for community-based screening [37,39].

Data Collection

Fundus Photography Protocol

Training Protocol

Before patient recruitment, 2 optometrists underwent 15 days of 1-on-1 training at the Advanced Eye Centre of Post Graduate Institute of Medical Education and Research, Chandigarh, and a primary health care center in Block Boothgarh of district Mohali. The training covered recruitment and consenting procedures, history recording, retinal image acquisition, troubleshooting poor-quality images, automated grading with the AI system, and delivering grading reports to patients. Training and supervision continued until staff were comfortable independently collecting data and performing imaging.

Image Acquisition and Grading

Adequate ventilation was maintained for patient consultations and fundus photography. Optometrists wore N95 masks and followed strict hand hygiene. Patients without masks were required to wear one during interactions and in waiting areas. Optometrists obtained nonmydriatic macula and disc-centered image color fundus photographs at 45° field of view using low-cost cameras (Figure 1) for participants at both sites. The screening at the CHC was carried out in an ophthalmology clinic. On the day of screening, the optometrists were assisted by ASHA workers at the homes of the eligible participants. The AI algorithms provided results at both sites. Two human graders, a certified optometrist and an ophthalmologist, graded all the fundus images. A senior retina specialist resolved any disagreements between the graders. All the classifications of DR were based on the International Classification for Diabetic Retinopathy [40].

Referral and Telephonic Follow-Up

Participants received an AI-generated report detailing their DR status, severity, and referral recommendations. Patients diagnosed with moderate non-proliferative DR and above were referred to as referable DR [41] and, with ungradable images, were advised to visit the district hospital for an ophthalmological review. One month after the screening, participants were followed up with telephone calls to confirm adherence and obtain diagnosis details. Each participant was contacted thrice, on 3 consecutive days, before being classified as nonresponsive. A structured questionnaire was administered to participants via telephone. The questionnaire assessed factors influencing appointment attendance and treatment adherence, with responses recorded as yes, no, or do not know for each question [42,43]. The questionnaire assessed sociodemographic characteristics and perceived barriers to referral (knowledge, attitude, logistics) and included an open-ended question for unlisted reasons behind incomplete referrals.

In-Depth Interviews

Two research fellows (AC and HR) with master's degrees and over 5 years of qualitative research experience conducted the in-depth interviews in the participant's preferred languages (Hindi, Punjabi, and English). Participants with DM were interviewed in quiet hospital locations at the CHC or at home in a properly ventilated space while HCPs were interviewed

XSL•FO
RenderX

face-to-face or via Zoom, based on availability. Study procedures were explained, and informed consent was obtained, either written for in-person interviews or verbal for Zoom sessions. Each interview was audio recorded and lasted for 35 - 40 minutes. The final sample size for the qualitative study was determined based on the principle of data saturation [44,45].

Data Analysis

Using a mixed methods approach, the results from both the quantitative and qualitative components were compared, synthesized, and discussed.

Quantitative

All participant data were entered into Research Electronic Data Capture (REDCap) [46], exported in Excel format, cleaned, and subsequently analyzed using Stata/IC (version 14.2; StataCorp LLC). The analysis involved reporting the descriptive statistics. Histograms were used to assess the normality of continuous variables. Summary statistics include frequencies (percentages) for categorical variables and means (standard deviations) for normally distributed continuous variables; otherwise, medians and interquartile ranges were reported. A chi-square (X^2) test was conducted to compare the distribution of categorical variables (eg, gender, education, marital status) between the adherent and nonadherent participant groups.

Univariate logistic regression was used to estimate odds ratios (ORs) and 95% CIs for the association between each independent variable (eg, gender, age, education, occupation, marital status, income, health insurance, duration of diabetes, hypertension status, and previous eye facility visits) and the dependent variable (referral adherence status: adherent versus nonadherent). The multivariate logistic regression model included independent variables with a P value <.05 in univariate analysis to adjust for potential confounding factors and identify independent predictors of referral nonadherence, using a significance level of P<.05 for the final model.

Qualitative

The qualitative researchers proficient in Hindi and Punjabi transcribed the recordings verbatim, translated them into English, and removed all personal identifiers before analysis. We analyzed the interview transcripts using thematic analysis, following Braun and Clarke's 6-phase approach [47]. Although the phases progressed logically, the process was not linear but recursive and iterative, involving movement back and forth between phases as needed [47]. We systematically organized the data, coded it, and grouped similar codes into themes to identify barriers to referral adherence. The research team members (AC: research scholar; RM: research associate; HR: research scholar) performed independent open transcript coding. MD, an experienced public health specialist, reviewed and confirmed the initial coding framework. Coding and categorization under themes were done using Atlas.Ti 23 software.

We then began extracting relevant portions of text from each interview related to the categories, sorting and selecting quotes and placing them under the appropriate themes. Quotes were edited for readability to retain their original meaning; ellipses (...) show removed text. The participants are denoted in italics with a unique ID number and their group designator: P (people with DM), Opt (Ophthalmologists), ASHA (ASHA workers), CHO (Community Health Officer), Optom (Optometrists), and RS (Retina Specialist). Given that both study sites are in Punjab and share similar health system building blocks [48], the in-depth interviews were analyzed and presented collectively.

Results

Quantitative

Participant Characteristics

A total of 390 (41%) of 948 older adult participants were enrolled across both sites, with 176 (45%) males and 214 (55%) females, primarily due to COVID-19 restrictions and participant refusals. Participants aged 60 - 69 years comprised 62.6244% of the sample, while those aged ≥70 years constituted 37.4%, with a mean age of 67.2 (SD 6.2) years. Around 48.5% had no formal education, 45.9% had education up to the 10th standard, and 5.6% had education at the 12th standard or higher. Occupationally, 69.7% were unemployed or engaged in home duties, 13.8% were retired, and 16.5% were in other categories. Marital status indicated that 74% were married. Additionally, 69% had a monthly household income below 30,000 Indian rupees (INR; approximately US $350), 52.3% had a history of hypertension, and the mean DM duration was 8.03 (SD 6.9) years. Site-wise demographic details are available in Table S4 in Multimedia Appendix 1.

Referral Recommendations and Adherence Rates

Out of 390 screened participants, 231 (59.2%) were not referred, while 159 (40.8%) were referred for further review and management. Among those referred, only 23 (14.5%) followed the advice and visited a health facility (details in Table S5 in Multimedia Appendix 1), while 136 (85.5%) did not adhere to the referral recommendations or seek further medical attention.

Differences Between Referral Categories (Adherent and Nonadherent)

Table 1 shows the demographic and medical characteristic differences between adherent and nonadherent participants in the referral category. The comparison of 159 referral participants showed no statistically significant differences in demographics and health characteristics between adherent (n=23) and nonadherent (n=136) groups (all P>.05). This includes gender distribution (P=.23), age groups (P=.48), educational levels (P=.27), occupational status (P=.33), marital status (P=.97), household income (P=.71), health insurance coverage (P=.93), duration of DM (P=.51), presence of hypertension (P=.16), and previous eye facility visits (P=.45).

XSL•FO
RenderX

Table . Difference in demographics and health characteristics between adherent and nonadherent participants in the referral category.

Variable	Referral category (n=159)		
	Adherent (n=23)	Nonadherent (n=136)	*P* value
Gender, n (%)			.23
Male	12 (52.7)	53 (39)	
Female	11 (47.8)	83 (61)	
Age group (years), n (%)			.48
60 - 69	14 (60.8)	72 (53)	
≥70	9 (39.1)	64 (47)	
Education, n (%)			.27
No formal education	8 (34.8)	72 (53)	
Up to 10th standard	14 (60.8)	59 (43.4)	
12th standard and above	1 (4.4)	5 (3.6)	
Occupation, n (%)			.33
Retired from service	4 (17.4)	15 (11)	
Unemployed/home duties	14 (60.9)	103 (75.8)	
Other	5 (21.7)	18 (13.2)	.97
Marital status, n (%)			
Married	16 (69.5)	94 (69)	
Unmarried, divorced, separated, widow, widower	7 (30.4)	42 (31)	
Monthly household income (INR)[a], n (%)			.71
<30,000	15 (65.2)	94 (69)	
>30,000	8 (34.8)	42 (31)	
Health insurance, n (%)			.93
Yes	8 (34.8)	46 (33.8)	
No	15 (65.2)	90 (66.2)	
Duration of DM[b] (years), n (%)			.51
0 - 10	17 (73.9)	91 (67)	
≥10	6 (26.1)	45 (33)	
Hypertension (years), n (%)			.16
Yes	15 (65.2)	67 (49.2)	
No	8 (34.8)	69 (50.8)	
Previous visit to the eye facility, n (%)			.45
Once	19 (82.6)	120 (88.2)	
Never	4 (17.4)	16 (11.8)	

[a]INR: Indian rupee. 30,000 INR=US $350.

[b]DM: diabetes mellitus.

Analysis of Determinants for Nonadherence to Referral Recommendations

Table 2 compares determinant factors between 23 adherent and 136 nonadherent participants using univariate and multivariate logistic regression analyses. None of the variables showed statistically significant differences (all *P*>.05), indicating no evidence of associations with adherence status across the examined factors. Gender (female) showed nonsignificantly higher odds of nonadherence, for univariate (OR 1.7, 95% CI 0.7-4.2*; P*=.23) and multivariate (OR 1.07, 95% CI 0.31-3.72*; P*=.91). Age ≥70 years had univariate and multivariate ORs of 1.4, 95% CI 0.56-3.4 *(P*=.48) and 1.4, 95% CI 0.5-4.09 *(P*=.47), respectively, indicating no significant association with adherence. Similarly, no significant differences were found in occupation categories, marital status, household income, health

XSL•FO
RenderX

insurance status, duration of DM, hypertension, or previous eye facility visits (all *P*>.05).

Table . Univariate and multivariate analysis of the determinants for nonadherence to referral recommendations. Reference: Base category used for comparison.

Variable	Adherent (n=23)		Nonadherent (n=136)	
	Univariate OR[a] (95% CI)	*P* value	Multivariate OR (95% CI)	*P* value
Gender, n (%)				
Male	Reference		Reference	
Female	1.7 (0.7-4.2)	.23	1.07 (0.31-3.72)	.91
Age group (years)				
60 - 69	Reference		Reference	
≥70	1.4 (0.56-3.4)	.48	1.4 (0.5-4.09)	.47
Education, n (%)				
No formal education	Reference		Reference	
Up to 10th standard	0.47 (0.18-1.19)	.11	0.49 (0.16-1.45)	.20
12th and above	0.5 (0.06-5.36)	.61	0.57 (0.04-7.40)	.67
Occupation, n (%)				
Retired	Reference		Reference	
Unemployed/home	1.9 (0.56-6.7)	.28	1.98 (0.38-10.14)	.41
Other	0.9 (0.21-4.2)	.95	0.98 (0.17-5.6)	.98
Marital status, n (%)				
Married	Reference		Reference	
Unmarried, divorced, separated, widow, widower	1.02 (0.39-2.67)	.96	0.68 (0.23-1.97)	.48
Monthly household income[b], n (%)				
<30,000	Reference		Reference	
>30,000	0.84 (0.33-2.12)	.71	0.94 (0.32-2.77)	.92
Health insurance				
Yes	Reference		Reference	
No	1.3 (0.55-3.4)	.50	0.71 (0.24-2.03)	.52
Duration of diabetes mellitus (years)				
0 - 10	Reference		Reference	
≥10	1.4 (0.52-3.8)	.51	1.56 (0.52-4.68)	.42
Hypertension (years)				
Yes	Reference		Reference	
No	1.9 (0.76-4.8)	.10	2.09 (0.79-5.48)	.13
Previous visit to the eye facility				
Once	Reference		Reference	
Never	0.63 (0.19-2.09)	.46	0.54 (0.14-2.04)	.37

[a]OR: odds ratio.

[b]INR: Indian rupee. 30,000 INR=US $350.

Telephone Follow-Up: Assessing Referral Adherence Rates and Reasons

Participants with ungradable images or diagnosed with referable DR (n=136) were contacted by phone 1 month after screening. The most commonly reported barriers were categorized as other factors (n=70, 51.5%), logistical challenges (n=31, 22.8%), and attitudinal issues (n=27, 19.9%).

XSL•FO
RenderX

Regarding attitude, 17 (12.5%) participants believed their eyes were fine, while logistical obstacles included lack of family support (n=21, 15.4%) and family problems (n=8, 5.8%). Financial constraints, particularly high treatment costs, affected 4 (2.9%) participants. Additionally, 4 (2.9%) were unaware of the importance of DR treatment. Seasonal harvesting (n=14, 10.3%) and other health issues (n=9, 6.6%) were also contributing factors (Table 3).

Table . Reasons for nonadherence to referral recommendations.

Categories and barriers to referral adherence		Participants (n=136), n (%)
Attitude		
	Eyes are fine	17 (12.5)
	Not willing to treatment	3 (2.2)
	Lack of concern about diabetic retinopathy	3 (2.2)
	Not interested in treatment	4 (2.9)
Financial		
	High treatment cost	4 (2.9)
Awareness		
	Not aware that DR treatment is important	4 (2.9)
Logistical support		
	Lack of family support	21 (15.4)
	Travel distance to a health facility	2 (1.6)
	Family problems	8 (5.8)
Other		
	Harvesting seasons	14 (10.3)
	Work commitments	4 (2.9)
	Out of station	6 (4.4)
	Phone with family members	3 (2.2)
	Planning to visit a health facility soon	2 (2.2)
	Other health issues	9 (6.6)
	Weather conditions	1 (0.7)
	Technical error (wrong contact numbers, switch-offs, unavailability)	31 (22.8)

Qualitative

Characteristics of the Participants

In total, 28 in-depth interviews were conducted with 9 older adult people with DM and 19 HCPs. Of the 9 people with DM, 4 (44%) were male and 5 (56%) were female. Most (n=7, 78%) were aged 60 - 69 years, and 8 (89%) were married. Two-thirds (n=6) lacked formal education, and 6 (67%) were unemployed or engaged in home duties (Tables S6 and S7 in Multimedia Appendix 1). The HCP group included 1 retina specialist (5%), 4 ophthalmologists (21%), 2 optometrists (10%), 3 CHOs (16%), and 9 ASHA workers (47%). Among 19 health care providers, 13 (68%) were female and 6 (32%) were male. Most (n=12, 63%) were aged 31 - 40 years.

Barriers Associated With Low Referral Adherence

Overview

The data revealed three key themes that best explained the barriers to referral adherence: (1) awareness and knowledge-related obstacles, (2) logistical support challenges, and (3) health care system limitations. The results of the in-depth interviews with the stakeholders complemented the quantitative findings of nonadherent participants and shed further light on their findings (Figure 2). The thematic analysis suggested a range of factors that impact adherence to referral instructions and put low adherence or nonadherence into perspective. The quotes under the categories are available in Table S8 in Multimedia Appendix 1.

XSL•FO
RenderX

Figure 2. Barriers to referral nonadherence. DR: diabetic retinopathy; HF: health facility; VEGF: vascular endothelial growth factor.

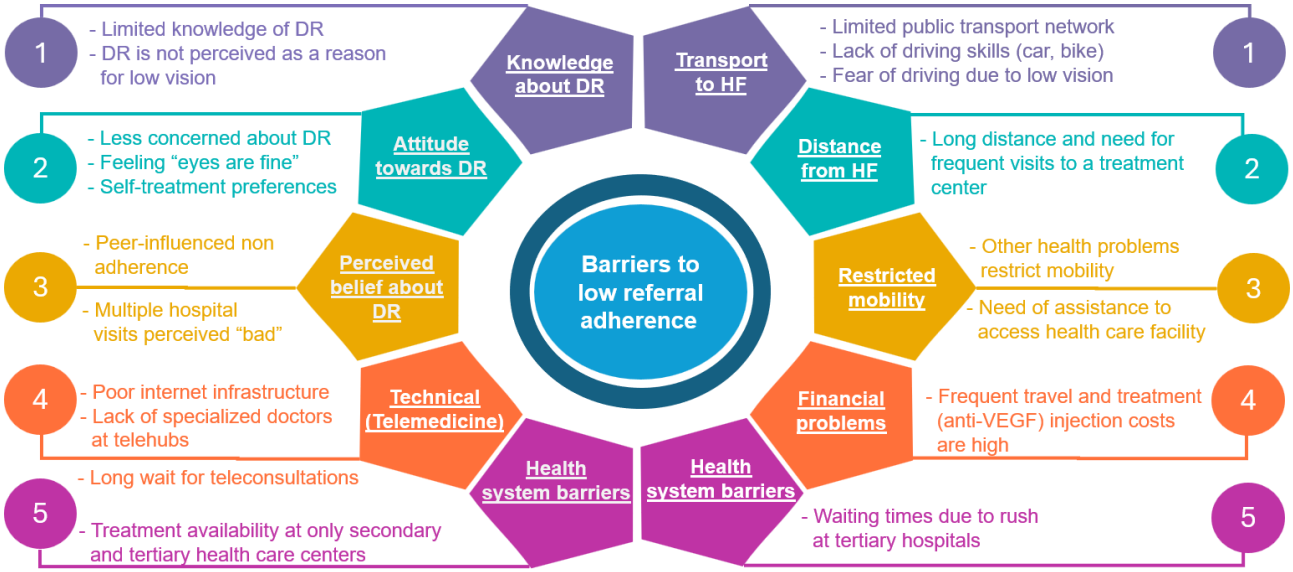

Awareness and Knowledge-Related Barriers

The HCPs observed that patients with DR often delayed treatment due to a lack of awareness about the condition and its complications, frequently mistaking vision changes for aging, cataracts, or simply needing new glasses. These misconceptions often lead patients to request remedies such as eye drops or new glasses, mistakenly believing these will resolve their vision problems. One patient recounted an interaction with fellow patients who had undergone extensive treatments and incurred substantial expenses without experiencing any improvement in vision. This perception highlights a common gap in patient knowledge. Individuals often rely excessively on peer anecdotal evidence, hindering adherence to evidence-based medical treatments. This issue is further complicated by a limited awareness of the crucial role regular follow-up appointments play in managing DR, resulting in poor compliance with treatment protocols. Additionally, some patients tend to avoid health care facilities under the false assumption that their ocular condition has stabilized while engaging in self-treatment, which introduces significant risks and compromises effective disease management.

Logistical Support Barriers

Mobility and Transportation

The HCPs said that the lack of public transport restricts participants from accessing local transport options for treatment, making it hard for patients to reach the facility for their appointments. One patient with a 2-wheeler reported difficulty driving alone because of their low vision. Longer distances to treatment facilities also decreased the frequency of visits, resulting in higher dropout rates despite initial engagement in treatment. Sometimes, multiple health issues can limit mobility and hinder access to eye care services, leading individuals to prioritize managing other health conditions over maintaining regular DR treatment appointments. This preference reflects a practical decision based on immediate health needs and logistical challenges posed by mobility restrictions rather than a lack of recognition of the importance of DR care.

Financial

Cost presents a formidable barrier, restricting many from affording regular travel to treatment centers. Initially, patients may adhere to treatment plans diligently, but as time progresses, dropout rates escalate. A patient articulated their predicament: "I was referred to another hospital for injections, which are financially out of reach due to the distance from home." The ophthalmologist and retina specialist underscored the substantial financial burden of DR treatment, specifically anti-VEGF (vascular endothelial growth factor) injections, which significantly impacts adherence. Financially disadvantaged patients frequently prioritize other expenses over necessary visits to eye care facilities, thereby exacerbating the challenge of managing DR effectively and consistently. This underscores the critical need for accessible and affordable eye care solutions to mitigate the economic barriers hindering patient care and treatment adherence.

Health Care System Barriers

Burdened Health System

The tertiary eye care hospital experiences high demand, resulting in delayed appointments that led many patients to discontinue treatment after 1-2 visits. Long queues also discourage patients from seeking care at larger eye care centers, underscoring the systemic challenges that affect access to DR treatment services. The large number of patients and extended waiting times pose significant challenges for individuals trying to attend appointments and receive timely treatment.

Technical Challenges

Telemedicine facilities are beneficial for extending eye care access [49], but they face implementation challenges. The limited availability of specialized doctors and unreliable internet in peripheral areas hinder teleconsultations. Village residents encountered prolonged wait times at telemedicine centres and expressed reluctance to wait for consultations with doctors at these facilities.

XSL•FO
RenderX

Discussion

Principal Findings

This study investigated barriers to referral adherence after AI-enabled DRS at primary (community) and secondary (CHC) sites among those who did not follow referrals to district hospitals for further management. We implemented AI-enabled DRS programs for older adults in public health settings in Punjab, India, during the COVID-19 pandemic. AI has been promoted as a transformative tool in health care to help attain health and health-related Sustainable Development Goals among older adults [50,51]. However, the impact of implementing AI on patient outcomes depends on many factors beyond diagnostic accuracy, including care access, referral adherence, and the implementation of treatment and management recommendations [52]. In our study, only 23 of 159 (14.5%) patients screened adhered to referral recommendations. Of these, only 9 visited the district hospital, while the others chose different facilities, indicating various factors influencing their decision-making regarding referral locations [53]. This finding contrasts with a study conducted in South India examining the factors influencing the utilization of referral services from secondary to tertiary eye care, which reported noncompliance rates of 31.65% and 24.2%, respectively [33].

Telephone interviews identified key barriers in our study, including low awareness about DR (n=4, 2.9%), attitudinal issues (n=27, 19.9%), and a lack of family support (n=21, 15.4%). Economic 27 (16.4%) and attitudinal barriers (44.2%) to referral service access were consistent with findings by Padhy et al [33] Our analysis showed no differences in the characteristics of adherent and nonadherent participants. However, nonstatistically significant differences between adherent and nonadherent groups do not imply that these factors are irrelevant or that there are no relationships. This indicates that we did not detect significant associations with the current sample size and variability. However, the wide confidence intervals suggest additional quantitative data are needed to improve estimate precision.

Nevertheless, we have no evidence that demographic and health variables alone are sufficient to understand or predict adherence behavior in patients with DR. Technical errors (Table 3) prevented us from contacting 31 of 136 (22.8%) participants, potentially impacting our quantitative analysis, including the univariate and multivariate analysis results. The unprecedented COVID-19 pandemic may have also negatively affected health-seeking behavior and treatment adherence [19,22].

We prioritized incorporating insights from key stakeholders through qualitative, in-depth interviews to guide informational power. Our study found that older adult individuals lack adequate awareness about DR complications, resulting in misconceptions about the causes of vision changes. Lack of awareness discourages service use, enabling unchecked harmful conditions and potentially leading to adverse outcomes due to delayed care-seeking. Further, the participants encountered difficulties affording transportation and treatment expenses and taking time off work for medical appointments, driven by concerns about their daily income [54]. Additionally, attitudes such as believing their eyes are healthy and logistical challenges like lack of family support impeded referral adherence [23]. Reliance on anecdotal evidence (do not go for eye treatment) from peers discouraged adherence to referral recommendations. Local and regional factors, such as the peak harvesting season (Table 3), hindered older adult adherence to referral recommendations (n=14, 10.3%). In-depth stakeholder interviews complemented the telephone follow-up findings, revealing barriers to referral recommendations. Similar programs, like the point-of-care DR examination program initiative, reported suboptimal follow-up rates despite targeted interventions, highlighting the need for more personalized strategies to enhance adherence [24]. Community-based programs have proven effective in enhancing access to eye care and improving patient adherence to continuity of care recommendations [24,55]. Bonilla-Escobar et al [24] showed higher adherence to referral recommendations when personalized approaches, such as phone calls, voicemails, and result letters with appointment details, were used. Transportation interventions, including bus passes, taxi vouchers, reimbursed travel, and tailored services, enhance health outcomes in chronic diseases, particularly for older adults and women [56].

Teleophthalmology, for instance, has accelerated health care access in rural and remote areas [57], overcoming significant travel constraints, time constraints, and economic barriers through community-based program initiatives [58,59]. However, the HCPs noted that technical challenges, such as underdeveloped internet infrastructure in rural areas, pose challenges to implementing teleophthalmology programs [60]. Hence, it is critical to identify methods for establishing DRS programs tailored to available resources and health care infrastructure [24,61]. In our study, older adult participants highlighted financial constraints as barriers to accessing referral services, a sentiment corroborated by the HCPs' in-depth interviews. The findings align with existing literature on economic challenges among individuals with visual impairment in developing countries [62]. The high costs of DR treatment often compel financially disadvantaged patients to prioritize other essential expenses over necessary eye care. Patients and HCPs should be informed about the comprehensive coverage offered by the Ayushman Bharat Pradhan Mantri Jan Arogya Yojana public health insurance program in India, including DRS and treatment. This ensures adequate access to necessary services through insurance benefits [63].

Our study findings revealed issues in the care process in rural areas, such as patients self-referring and bypassing the referral system hierarchy [64]. Systemic challenges such as delayed appointments and long queues discourage patients from continuing treatment beyond 1 or 2 visits due to access issues [23]. The World Report on Vision has recommended an integrated care model incorporating primary, secondary, and tertiary care, ensuring clear roles and appropriate referrals at each health care system level [65].

This study's strength lies in triangulating quantitative findings with qualitative insights to address research questions regarding referral adherence. Additionally, conducting the study across primary (community) and secondary (CHC) health care systems

XSL•FO
RenderX

provides comprehensive information on referral adherence within a network of public health systems.

A limitation of this study is its focus on a public health network in North India within one state, potentially limiting the generalizability of its findings to broader settings. Potential biases from self-reported data during telephone follow-ups may lead to underreporting or overreporting adherence based on perceived expectations and limited generalizability due to the study's regional focus. The study was conducted during the COVID-19 pandemic, marked by significant disruptions to health care systems and changes in patient behavior. These unique circumstances, including the impact of COVID-19 on decision-making regarding nonemergency medical care, may have introduced factors related to health care access that are specific to this time. Furthermore, nearly 22.8% of screened patients (Table 3) were not successfully contacted despite multiple attempts, which would increase the imprecision of estimates yielded by the regression analysis. For example, only 23 adherent participants contributed data to the multivariate model and therefore the analyses may be hindered due to issues related to statistical power and imbalanced group sizes [66]. This study did not extensively collect detailed information on health care system factors affecting adherence, such as hospital waiting times, hospital accessibility, and staff attitudes [33]. This highlights the necessity of exploring additional factors that influence adherence, including psychological, social, and systemic barriers [23,67].

In this pilot study, approximately 390 of 948 (41%) participants were recruited due to COVID-19 pandemic restrictions. Additionally, the limited sample size in the nonadherent group and high variability in the collected data resulted in insufficient statistical power to accurately identify determinants of nonadherence, leading to large confidence intervals and reduced precision in the estimates.

Conclusion and Policy Implications

This study, conducted during the COVID-19 pandemic, has demonstrated suboptimal adherence among older adult patients to referral recommendations. Participants considered barriers such as knowledge gaps, logistical challenges, and health system–related issues to have hindered referral adherence. Given the unique context of the pandemic, the results should be generalized with caution. Future studies should consider pragmatic trial study designs incorporating implementation science approaches using mixed methods to examine barriers to referral and adherence across diverse contexts. This could include exploring community beliefs and other social determinants influencing service uptake. Policy-level changes are essential to integrating teleophthalmology into public health programs and improving infrastructure for AI-enabled DRS, enhancing access and outcomes. An effective, well-coordinated referral system with defined protocols, streamlined communication, reliable transportation, trained personnel, integrated systems, and collaboration may address the challenges in the referral systems [68,69].

Acknowledgments

The authors sincerely thank all participants for their invaluable contributions to this research and the Department of Health and Family Welfare, Punjab, India, for granting permission to conduct the study. All authors declared that they had insufficient funding to support open access publication of this manuscript, including from affiliated organizations or institutions, funding agencies, or other organizations. JMIR Publications provided article processing fee (APF) support for the publication of this article.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary materials.
[PDF File, 248 KB - formative_v9i1e67047_app1.pdf]

References

1. Raman R, Vasconcelos JC, Rajalakshmi R, et al. Prevalence of diabetic retinopathy in India stratified by known and undiagnosed diabetes, urban-rural locations, and socioeconomic indices: results from the SMART India population-based cross-sectional screening study. Lancet Glob Health 2022 Dec;10(12):e1764-e1773. [doi: 10.1016/S2214-109X(22)00411-9] [Medline: 36327997]
2. Ting DSW, Cheung GCM, Wong TY. Diabetic retinopathy: global prevalence, major risk factors, screening practices and public health challenges: a review. Clin Exp Ophthalmol 2016 May;44(4):260-277. [doi: 10.1111/ceo.12696] [Medline: 26716602]
3. Vujosevic S, Aldington SJ, Silva P, et al. Screening for diabetic retinopathy: new perspectives and challenges. Lancet Diabetes Endocrinol 2020 Apr;8(4):337-347. [doi: 10.1016/S2213-8587(19)30411-5] [Medline: 32113513]
4. IDF Diabetes Atlas. URL: www.diabetesatlas.org [accessed 2025-03-26]
5. Teo ZL, Tham YC, Yu M, et al. Global prevalence of diabetic retinopathy and projection of burden through 2045: systematic review and meta-analysis. Ophthalmology 2021 Nov;128(11):1580-1591. [doi: 10.1016/j.ophtha.2021.04.027] [Medline: 33940045]

XSL•FO
RenderX

6. Flaxman SR, Bourne RRA, Resnikoff S, et al. Global causes of blindness and distance vision impairment 1990-2020: a systematic review and meta-analysis. Lancet Glob Health 2017 Dec;5(12):e1221-e1234. [doi: 10.1016/S2214-109X(17)30393-5] [Medline: 29032195]
7. Leley SP, Ciulla TA, Bhatwadekar AD. Diabetic retinopathy in the aging population: a perspective of pathogenesis and treatment. Clin Interv Aging 2021;16:1367-1378. [doi: 10.2147/CIA.S297494] [Medline: 34290499]
8. Lamoureux EL, Tai ES, Thumboo J, et al. Impact of diabetic retinopathy on vision-specific function. Ophthalmology 2010 Apr;117(4):757-765. [doi: 10.1016/j.ophtha.2009.09.035] [Medline: 20122736]
9. Fenwick E, Rees G, Pesudovs K, et al. Social and emotional impact of diabetic retinopathy: a review. Clin Exp Ophthalmol 2012;40(1):27-38. [doi: 10.1111/j.1442-9071.2011.02599.x] [Medline: 21575125]
10. National Programme for Health Care of the Elderly. Ministry of Health and Family Welfare. URL: https://mohfw.gov.in/?q=major-programmes/Non-Communicable-Diseases/Non-Communicable-Diseases-1 [accessed 2025-03-26]
11. Liew G, Michaelides M, Bunce C. A comparison of the causes of blindness certifications in England and Wales in working age adults (16-64 years), 1999-2000 with 2009-2010. BMJ Open 2014 Feb 12;4(2):e004015. [doi: 10.1136/bmjopen-2013-004015] [Medline: 24525390]
12. Ryan ME, Rajalakshmi R, Prathiba V, et al. Comparison Among Methods of Retinopathy Assessment (CAMRA) study: smartphone, nonmydriatic, and mydriatic photography. Ophthalmology 2015 Oct;122(10):2038-2043. [doi: 10.1016/j.ophtha.2015.06.011] [Medline: 26189190]
13. Padhy SK, Takkar B, Chawla R, Kumar A. Artificial intelligence in diabetic retinopathy: a natural step to the future. Indian J Ophthalmol 2019 Jul;67(7):1004-1009. [doi: 10.4103/ijo.IJO_1989_18] [Medline: 31238395]
14. Abbasi K. The scandals of covid-19. BMJ 2020 Apr(369):m1434. [doi: 10.1136/bmj.m1434]
15. Shih KC, Kwong ASK, Wang JHL, et al. Diabetic retinopathy screening during the coronavirus disease 2019 pandemic. Eye (Lond) 2020 Jul;34(7):1246-1247. [doi: 10.1038/s41433-020-0928-7] [Medline: 32367007]
16. Grzybowski A, Brona P, Lim G, et al. Artificial intelligence for diabetic retinopathy screening: a review. Eye (Lond) 2020 Mar;34(3):451-460. [doi: 10.1038/s41433-019-0566-0] [Medline: 31488886]
17. Mathenge W, Whitestone N, Nkurikiye J, et al. Impact of artificial intelligence assessment of diabetic retinopathy on referral service uptake in a low-resource setting: the RAIDERS randomized trial. Ophthalmol Sci 2022 Dec;2(4):100168. [doi: 10.1016/j.xops.2022.100168] [Medline: 36531575]
18. Bonilla-Escobar FJ, Ghobrial AI, Gallagher DS, Eller A, Waxman EL. Comprehensive insights into a decade-long journey: the evolution, impact, and human factors of an asynchronous telemedicine program for diabetic retinopathy screening in Pennsylvania, United States. PLoS One 2024;19(7):e0305586. [doi: 10.1371/journal.pone.0305586] [Medline: 38995899]
19. Ahmed I, Liu TYA. The impact of COVID-19 on diabetic retinopathy monitoring and treatment. Curr Diab Rep 2021 Sep 8;21(10):40. [doi: 10.1007/s11892-021-01411-6] [Medline: 34495377]
20. Kumar A. The transformation of the Indian healthcare system. Cureus 2023 May;15(5):e39079. [doi: 10.7759/cureus.39079] [Medline: 37378105]
21. Burkhart PV, Sabaté E. Adherence to long-term therapies: evidence for action. J Nurs Scholarsh 2003;35(3):207. [Medline: 14562485]
22. Breazzano MP, Nair AA, Arevalo JF, et al. Frequency of urgent or emergent vitreoretinal surgical procedures in the United States during the COVID-19 pandemic. JAMA Ophthalmol 2021 Apr 1;139(4):456-463. [doi: 10.1001/jamaophthalmol.2021.0036] [Medline: 33662093]
23. Piyasena M, Murthy GVS, Yip JLY, et al. Systematic review on barriers and enablers for access to diabetic retinopathy screening services in different income settings. PLoS One 2019;14(4):e0198979. [doi: 10.1371/journal.pone.0198979] [Medline: 31013274]
24. Bonilla-Escobar FJ, Eibel MR, Le L, Gallagher DS, Waxman EL. Follow-up in a point-of-care diabetic retinopathy program in Pittsburgh: a non-concurrent retrospective cohort study. BMC Ophthalmol 2024 Aug 20;24(1):356. [doi: 10.1186/s12886-024-03581-9] [Medline: 39164678]
25. Khanna RC, Kim S, Giridhar P, Mettla AL, Marmamula S, Rao GN. Barriers to uptake of referral services from secondary care to tertiary care and its associated factors in L V Prasad Eye Institute network in Southern India: a cross-sectional study. BMJ Open 2018 Jul 7;8(7):e020687. [doi: 10.1136/bmjopen-2017-020687] [Medline: 29982204]
26. Adhikari K, Patten SB, Patel AB, et al. Data harmonization and data pooling from cohort studies: a practical approach for data management. Int J Popul Data Sci 2021;6(1):1680. [doi: 10.23889/ijpds.v6i1.1680] [Medline: 34888420]
27. ASHA training modules. Ministry of Health and Family Welfare. URL: https://nhm.gov.in/index1.php?lang=1&level=3&sublinkid=184&lid=257 [accessed 2025-03-26]
28. Ayushman bharat comprehensive primary health care through health and wellness centres operational guidelines. Ministry of Health and Family Welfare. URL: http://www.ab-hwc.mohfw.gov.in/ [accessed 2023-10-13]
29. Indian public health standards. Community health centre 2022. National Health Systems Resource Centre. URL: https://nhsrcindia.org/sites/default/files/CHC%20IPHS%202022%20Guidelines%20pdf.pdf [accessed 2023-10-08]
30. Indian public health standards, sub district and district health hospital. Ministry of Health & Family Welfare. URL: https://nhsrcindia.org [accessed 2025-03-26]

XSL•FO
RenderX

31. Mukamel DB, Bresnick GH, Wang Q, Dickey CF. Barriers to compliance with screening guidelines for diabetic retinopathy. Ophthalmic Epidemiol 1999 Mar;6(1):61-72. [doi: 10.1076/opep.6.1.61.1563] [Medline: 10384685]
32. Watson MJG, McCluskey PJ, Grigg JR, Kanagasingam Y, Daire J, Estai M. Barriers and facilitators to diabetic retinopathy screening within Australian primary care. BMC Fam Pract 2021 Nov 30;22(1):239. [doi: 10.1186/s12875-021-01586-7] [Medline: 34847874]
33. Padhy D, Pyda G, Marmamula S, Khanna RC. Barriers to uptake of referral services from secondary eye care to tertiary eye care and its associated determinants in L V Prasad Eye Institute network in Southern India: a cross-sectional study-report II. PLoS ONE 2024;19(5):e0303401. [doi: 10.1371/journal.pone.0303401] [Medline: 38743737]
34. Kumar S, Kumar G, Velu S, et al. Patient and provider perspectives on barriers to screening for diabetic retinopathy: an exploratory study from southern India. BMJ Open 2020 Dec 10;10(12):e037277. [doi: 10.1136/bmjopen-2020-037277] [Medline: 33303431]
35. Shah P, Mishra DK, Shanmugam MP, Doshi B, Jayaraj H, Ramanjulu R. Validation of deep convolutional neural network-based algorithm for detection of diabetic retinopathy - artificial intelligence versus clinician for screening. Indian J Ophthalmol 2020 Feb;68(2):398-405. [doi: 10.4103/ijo.IJO_966_19] [Medline: 31957737]
36. Darwish DY, Patel SN, Gao Y, et al. Diagnostic accuracy and reliability of retinal pathology using the Forus 3nethra fundus camera compared to ultra wide-field imaging. Eye (Lond) 2019 May;33(5):856-857. [doi: 10.1038/s41433-019-0339-9] [Medline: 30679873]
37. Sosale B, Sosale AR, Murthy H, Sengupta S, Naveenam M. Medios- an offline, smartphone-based artificial intelligence algorithm for the diagnosis of diabetic retinopathy. Indian J Ophthalmol 2020 Feb;68(2):391-395. [doi: 10.4103/ijo.IJO_1203_19] [Medline: 31957735]
38. Duggal M, et al. Responsible adoption of cloud-based artificial intelligence in health care: A validation case study of multiple artificial intelligence algorithms for diabetic retinopathy screening in public health settings (forthcoming). .
39. Rajalakshmi R, Subashini R, Anjana RM, Mohan V. Automated diabetic retinopathy detection in smartphone-based fundus photography using artificial intelligence. Eye (Lond) 2018 Jun;32(6):1138-1144. [doi: 10.1038/s41433-018-0064-9] [Medline: 29520050]
40. Wong TY, Sun J, Kawasaki R, et al. Guidelines on diabetic eye care: the International Council of Ophthalmology recommendations for screening, follow-up, referral, and treatment based on resource settings. Ophthalmology 2018 Oct;125(10):1608-1622. [doi: 10.1016/j.ophtha.2018.04.007] [Medline: 29776671]
41. Raman R, Ramasamy K, Rajalakshmi R, Sivaprasad S, Natarajan S. Diabetic retinopathy screening guidelines in India: All India Ophthalmological Society diabetic retinopathy task force and Vitreoretinal Society of India Consensus Statement. Indian J Ophthalmol 2021 Mar;69(3):678-688. [doi: 10.4103/ijo.IJO_667_20] [Medline: 33269742]
42. Lee BW, Sathyan P, John RK, Singh K, Robin AL. Predictors of and barriers associated with poor follow-up in patients with glaucoma in South India. Arch Ophthalmol 2008 Oct;126(10):1448-1454. [doi: 10.1001/archopht.126.10.1448] [Medline: 18852425]
43. Vengadesan N, Ahmad M, Sindal MD, Sengupta S. Delayed follow-up in patients with diabetic retinopathy in South India: social factors and impact on disease progression. Indian J Ophthalmol 2017 May;65(5):376-384. [doi: 10.4103/ijo.IJO_620_16] [Medline: 28573993]
44. Mason M. Sample size and saturation in PhD studies using qualitative interviews. Forum Qual Soc Res 2010;11(3). [doi: 10.17169/fqs-11.3.1428]
45. Ritchie J, Lewis J. Qualitative Research Practice: A Guide for Social Science Students and Researchers: Sage Publications; 2003. URL: https://mthoyibi.files.wordpress.com/2011/10/qualitative-research-practice_a-guide-for-social-science-students-and-researchers_jane-ritchie-and-jane-lewis-eds_20031.pdf [accessed 2025-03-27]
46. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381. [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
47. Byrne D. A worked example of Braun and Clarke's approach to reflexive thematic analysis. Qual Quant 2022 Jun;56(3):1391-1412. [doi: 10.1007/s11135-021-01182-y]
48. Monitoring the building blocks of health systems: a handbook of indicators and their measurement strategies. World Health Organization. URL: https://apps.who.int/iris/bitstream/handle/10665/258734/9789241564052-eng.pdf [accessed 2025-03-26]
49. Kalogeropoulos D, Kalogeropoulos C, Stefaniotou M, Neofytou M. The role of tele-ophthalmology in diabetic retinopathy screening. J Optom 2020;13(4):262-268. [doi: 10.1016/j.optom.2019.12.004] [Medline: 31948924]
50. Sajwani RA. Artificial intelligence for sustainability development in healthcare. In: BUiD Doctoral Research Conference 2023 2024:264-272. [doi: 10.1007/978-3-031-56121-4_26]
51. Ageing, older persons and the 2030 agenda for sustainable development acknowledgements. : UNDP; 2017 URL: https://www.undp.org/publications/ageing-older-persons-and-2030-agenda-sustainable-development [accessed 2025-03-27]
52. Haynes B. Can it work? Does it work? Is it worth it? The testing of healthcareinterventions is evolving. BMJ 1999 Sep 11;319(7211):652-653. [doi: 10.1136/bmj.319.7211.652] [Medline: 10480802]

XSL•FO
RenderX

53. Piyasena M, Murthy GVS, Yip JLY, et al. Systematic review and meta-analysis of diagnostic accuracy of detection of any level of diabetic retinopathy using digital retinal imaging. Syst Rev 2018 Nov 7;7(1):182. [doi: 10.1186/s13643-018-0846-y] [Medline: 30404665]
54. Taber JM, Leyva B, Persoskie A. Why do people avoid medical care? A qualitative study using national data. J Gen Intern Med 2015 Mar;30(3):290-297. [doi: 10.1007/s11606-014-3089-1] [Medline: 25387439]
55. Solomon SD, Shoge RY, Ervin AM, et al. Improving access to eye care: a systematic review of the literature. Ophthalmology 2022 Oct;129(10):e114-e126. [doi: 10.1016/j.ophtha.2022.07.012] [Medline: 36058739]
56. Starbird LE, DiMaina C, Sun CA, Han HR. A systematic review of interventions to minimize transportation barriers among people with chronic diseases. J Community Health 2019 Apr;44(2):400-411. [doi: 10.1007/s10900-018-0572-3] [Medline: 30206755]
57. Chabouis A, Berdugo M, Meas T, et al. Benefits of Ophdiat, a telemedical network to screen for diabetic retinopathy: a retrospective study in five reference hospital centres. Diabetes Metab 2009 Jun;35(3):228-232. [doi: 10.1016/j.diabet.2008.12.001] [Medline: 19349202]
58. Hark LA, Katz LJ, Myers JS, et al. Philadelphia telemedicine glaucoma detection and follow-up study: methods and screening results. Am J Ophthalmol 2017 Sep;181:114-124. [doi: 10.1016/j.ajo.2017.06.024] [Medline: 28673747]
59. Wang YZ, He YG, Mitzel G, Zhang S, Bartlett M. Handheld shape discrimination hyperacuity test on a mobile device for remote monitoring of visual function in maculopathy. Invest Ophthalmol Vis Sci 2013 Aug 13;54(8):5497-5505. [doi: 10.1167/iovs.13-12037] [Medline: 23860761]
60. Yim D, Chandra S, Sondh R, Thottarath S, Sivaprasad S. Barriers in establishing systematic diabetic retinopathy screening through telemedicine in low- and middle-income countries. Indian J Ophthalmol 2021 Nov;69(11):2987-2992. [doi: 10.4103/ijo.IJO_1411_21] [Medline: 34708733]
61. Wong TY, Sabanayagam C. Strategies to tackle the global burden of diabetic retinopathy: from epidemiology to artificial intelligence. Ophthalmologica 2020;243(1):9-20. [doi: 10.1159/000502387] [Medline: 31408872]
62. Sengo DB, Marraca NA, Muaprato AM, García-Sanjuan S, Caballero P, López-Izquierdo I. Barriers to accessing eye health services in suburban communities in Nampula, Mozambique. Int J Environ Res Public Health 2022 Mar 25;19(7):7. [doi: 10.3390/ijerph19073916] [Medline: 35409600]
63. Government of india ministry of health and family welfare, AB-PMJAY health benefit package master 2022. National Health Authority. URL: https://pmjay.gov.in/sites/default/files/2022-04/HBP%202022%20..pdf [accessed 2024-07-14]
64. Eskandari M, Abbaszadeh A, Borhani F. Barriers of referral system to health care provision in rural societies in Iran. J Caring Sci 2013 Sep;2(3):229-236. [doi: 10.5681/jcs.2013.028] [Medline: 25276731]
65. Rao GN, Khanna RC, Athota SM, Rajshekar V, Rani PK. Integrated model of primary and secondary eye care for underserved rural areas: the L V Prasad Eye Institute experience. Indian J Ophthalmol 2012;60(5):396-400. [doi: 10.4103/0301-4738.100533] [Medline: 22944748]
66. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 1996 Dec;49(12):1373-1379. [doi: 10.1016/s0895-4356(96)00236-3] [Medline: 8970487]
67. Piyasena M, Murthy GVS, Yip JLY, et al. A qualitative study on barriers and enablers to uptake of diabetic retinopathy screening by people with diabetes in the Western Province of Sri Lanka. Trop Med Health 2019;47:34. [doi: 10.1186/s41182-019-0160-y] [Medline: 31139011]
68. Kruk ME, Gage AD, Arsenault C, et al. High-quality health systems in the Sustainable Development Goals era: time for a revolution. Lancet Glob Health 2018 Nov;6(11):e1196-e1252. [doi: 10.1016/S2214-109X(18)30386-3] [Medline: 30196093]
69. Kamau KJ, Osuga BO, Njuguna S. Challenges facing implementation of referral system for quality health care services in Kiambu County, Kenya. Health Syst Policy Res 2017;04(1):01. [doi: 10.21767/2254-9137.100067]

Abbreviations

ASHA: Accredited Social and Health Activists
CHC: community health center
CHO: community health officer
DM: diabetes mellitus
DR: diabetic retinopathy
DRS: DR screening
HCP: health care provider
VTDR: vision-threatening diabetic retinopathy

XSL•FO
RenderX

Edited by A Mavragani; submitted 30.09.24; peer-reviewed by F Bonilla-Escobar, T Intharah; revised version received 22.02.25; accepted 25.02.25; published 31.03.25.

Please cite as:
Chauhan A, Goyal A, Masih R, Kaur G, Kumar L, Neha , Rastogi H, Kumar S, Singh BL, Syal P, Gupta V, Vale L, Duggal M
Barriers and Determinants of Referral Adherence in AI-Enabled Diabetic Retinopathy Screening for Older Adults in Northern India During the COVID-19 Pandemic: Mixed Methods Pilot Study
JMIR Form Res 2025;9:e67047
URL: https://formative.jmir.org/2025/1/e67047
doi:10.2196/67047

© Anshul Chauhan, Anju Goyal, Ritika Masih, Gagandeep Kaur, Lakshay Kumar, Neha, Harsh Rastogi, Sonam Kumar, Bidhi Lord Singh, Preeti Syal, Vishali Gupta, Luke Vale, Mona Duggal. Originally published in JMIR Formative Research (https://formative.jmir.org), 31.3.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Personalized mHealth Intervention (StepAdd) for Increasing Physical Activity in Japanese Patients With Type 2 Diabetes: Secondary Analysis of Social Cognitive Theory Measurements of a Single-Arm Pilot Study

Kayo Waki[1,2,3], MPH, MD, PhD; Syunpei Enomoto[3], MAS; Toshimasa Yamauchi[2], MD, PhD; Masaomi Nangaku[4], MD, PhD; Kazuhiko Ohe[1,3], MD, PhD

[1]Department of Biomedical Informatics, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan

[2]Department of Diabetes and Metabolic Diseases, The University of Tokyo, Tokyo, Japan

[3]Department of Planning, Information and Management, University of Tokyo Hospital, Tokyo, Japan

[4]Division of Nephrology and Endocrinology, The University of Tokyo, Tokyo, Japan

Corresponding Author:
Kayo Waki, MPH, MD, PhD
Department of Biomedical Informatics, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan

Abstract

Background: A 12-week pilot of the StepAdd mobile health (mHealth) behavior change intervention based on social cognitive theory (SCT) saw an 86.7% increase in mean daily step counts among patients with type 2 diabetes. Due to the lack of exploration of theoretical implications in mHealth intervention studies, there is a need to understand the mechanism underlying the behavioral change to inform the future design of digital therapeutics.

Objective: This study aimed to examine the SCT drivers underlying the mean increase in exercise among Japanese patients with type 2 diabetes who participated in the StepAdd intervention.

Methods: This is a post hoc analysis of data collected in the single-arm pilot study of the 32 patients who completed the StepAdd intervention. The StepAdd app uses self-mastery and coping strategies to increase self-efficacy and thus increase walking. Self-mastery was measured by the goal completion (GC) rate, which is the percentage of days in which patients met these adapting goals. The use of coping strategies was measured by the strategy implementation (SI) rate, which is the percentage of days in which patients applied their selected coping strategies. We assessed correlations between GC, SI, and self-efficacy to increase walking via linear regression and analyzed relationships via structural equation modeling.

Results: We found statistically significant support for the SCT approach, including a correlation coefficient (ρ) of 0.649 between step increase and GC rate ($P<.001$); a ρ of 0.497 between the coping SI rate and self-efficacy increase ($P=.004$); a ρ of 0.446 between GC rate and self-mastery increase ($P=.01$); and a ρ of 0.355 between self-regulation increase and step increase ($P=.046$), giving us insight into why the behavior intervention succeeded. We also found significant correlations between self-efficacy for barriers and self-efficacy for task-specific behavior (ρ=0.358; $P=.04$), as well as self-regulation and self-efficacy for task-specific behavior (ρ=0.583; $P<.001$). However, a cross-lagged panel modeling analysis found no significant evidence that changes in self-efficacy preceded behavior changes in line with SCT.

Conclusions: Self-mastery and coping strategies contributed to the walking behavior change in StepAdd, supporting the SCT model of behavior change. Future research is needed to better understand the causal pathways proposed by SCT.

(JMIR Form Res 2025;9:e60221) doi:10.2196/60221

KEYWORDS

social cognitive theory; mHealth; mobile health; behavior change; diabetes; diabetic; type 2 diabetes; walk; step; exercise; physical activity; walking; digital therapeutics; secondary analysis; personalization; coping; self-efficacy

Introduction

Digital therapeutics (DTx), defined as evidence-based therapeutic interventions that are driven by high-quality software programs to treat, manage, or prevent a disease or disorder, can be effective in treating diabetes by inducing behavior changes [1]. However, studies of DTx generally do not explore the implications of the underlying behavior change theory. For example, we were unable to explain the mechanism behind behavior change with a randomized controlled trial (RCT) that

XSL•FO
RenderX

showed significant improvement in hemoglobin A_{1C} (HbA_{1c}) and albuminuria among patients with type 2 diabetes via DTx [2]. However, given the progress in DTx, it is not enough to simply show outcomes. Lifestyle change involves complex patient behavior comprising numerous decisions and significant effort. We need to explain the mechanism behind behavior change, to show the characteristics of the app in question that differ from those of other apps, and to explain the rationale for choosing the app to treat patients.

We analyzed results from a 12-week pilot study of 33 Japanese patients with diabetes who participated in the StepAdd mobile health (mHealth) behavior change intervention to explain the mechanism behind the behavior change [3]. This was a single-arm pilot study from August 2021 to December 2021 using pre-post evaluation to assess the feasibility and preliminary efficacy of a personalized mHealth intervention using the StepAdd app to improve physical activity among patients with type 2 diabetes, conducted in collaboration with Mitsui Memorial Hospital; Nihon Chouzai Co, Ltd (a community pharmacy); and Mitsui & Co, Ltd. Patients receiving type 2 diabetes treatment at Mitsui Memorial Hospital, with an HbA_{1c} of 7.5% or higher, were recruited by their physicians to participate in the pilot study evaluating the preliminary efficacy of StepAdd in increasing physical activity. StepAdd applies behavior change methods based on social cognitive theory (SCT) [4,5], and the pilot collected SCT-related measures throughout the intervention. The study included 6 visits spanning an initial step baseline phase and a 12-week continuous measurement phase. The transtheoretical model was used to categorize the participants who were currently at the contemplation stage (willing to change health behavior within the next 6 months), preparation stage (willing to change health behavior within the next month), or action stage (has made modifications to health behavior) to achieve the target goal of 10,000 steps a day. A community pharmacist briefed participants on this research study and on how to use the smartphone and the pedometer. The pilot saw an 86.7% increase in mean daily step counts, from 5436 (SD 2231) per day to 10,150 (SD 3908) per day, and a reduction of mean HbA_{1c} from 8.58% (SD 1.02%) to 7.79% (SD 1.11%) (−0.79%, SD 1.04%) [4]. A 24-week StepAdd RCT, with 80 participants per arm, is currently underway [6]. Analysis of RCTs has proven that increased walking causes clinically significant improvements to glycemic control among patients with diabetes [7,8]. The StepAdd pilot results show overall causality—the intervention caused the desired behavior and health improvement.

This secondary analysis seeks to use SCT measurements to understand the factors underlying the behavioral change, both to inform future analyses of the StepAdd RCT results and to improve the design of future interventions. We want to understand what features caused the change and whether the causal relationships defined by SCT can be proved.

SCT posits a central role to self-efficacy, the individual's belief that they can perform the behavior, with increases in self-efficacy causing increases in the behavior. StepAdd uses self-mastery and coping strategies to increase self-efficacy and thus increase walking. For self-mastery, StepAdd implements achievable but challenging goals, which are adapted weekly, with self-mastery being measured by the goal completion (GC) rate. For coping strategies, StepAdd has participants choose a barrier and an associated coping strategy to implement, increasing self-efficacy, with coping strategy use being measured by the strategy implementation (SI) rate. During the StepAdd pilot, we collected 3 measures of the underlying latent construct of self-efficacy at weeks 0, 4, 8, and 12: self-efficacy in achieving the targeted behavior, self-efficacy in dealing with barriers, and self-regulation (SR).

Methods

Overview

This is a post hoc analysis of data collected in the StepAdd pilot for the 32 patients who completed the intervention (the pilot analyzed 33 patients, but 1 dropped out after week 4 because he expected the study to be just 4 weeks). Data came from pedometers, the StepAdd app, and questionnaire instruments, with data collected at weeks 0, 4, 8, and 12 of the intervention.

The step data came from daily counts from wireless pedometers (OMRON HJA-405 T-W or Yamasa AW-001), averaged for the 2 weeks prior to the week 0, 4, 8, and 12 time points.

The intervention used 2 SCT-based methods. First, it targeted improving self-mastery by using a series of attainable yet growing goals. The GC rate measured the percentage of days in which patients met these adapting goals [3]. Second, it targeted enhancing SR and improving sociocultural factors in the SCT framework by allowing patients to evaluate possible barriers to achieving their step goals and choose to implement associated coping strategies. The SI rate measured the percentage of days in which patients applied their selected coping strategies [3].

The intervention is based in particular on increasing self-efficacy, the belief in one's own efficacy, applying it in this case to daily walking. We used 3 measures (self-efficacy for barriers [SE-B], self-efficacy for task-specific behavior [SE-T], and SR) of self-efficacy. SE-B is a self-efficacy scale to deal with barriers in achieving the targeted daily steps [9]. It is the sum of 4 questions, each rated 1 - 5, delivered via a questionnaire, amounting to a total score range of 4 - 20. SE-T is a self-efficacy scale for exercise in achieving the targeted daily steps [10]. It is the average of a participant's confidence (in percentages) in their ability to walk 4000; 6000; 8000; 10,000; or 12,000 steps daily, as measured via a questionnaire. SR was measured using the Japanese version of the 12-item Physical Activity Self-Regulation Scale (PASR-12) [11]. It is the sum of 12 questions, each rated 1 - 5, amounting to a total score range of 12 - 60.

We analyzed measurements using means and standard deviations. We assessed correlation via linear regression and analyzed relationships via structural equation modeling. All statistical analyses used SAS 9.4, with the significance level for all the statistical tests at $P<.05$ (2-tailed). We calculated 95% CIs where applicable.

XSL•FO
RenderX

Ethical Considerations

This study was approved by the Institutional Review Board of the University of Tokyo School of Medicine, under approval number 2021084NI-(2). This approval covers secondary analysis without additional consent. Study data were deidentified to ensure confidentiality and to protect all research participants.

Results

The behavior of interest, walking, increased steadily over the 12-week intervention (Figure 1). Both intervention method measurements (GC and SI) were consistently high (68% - 87%). The measures of self-efficacy (SE-B, SE-T, and SR) grew throughout the intervention.

Figure 1. Means and 95% CIs of change in 6 measured items over 12 weeks in the StepAdd intervention for Japanese patients with type 2 diabetes. GC: goal completion; SE-B: self-efficacy for barriers; SE-T: self-efficacy for task-specific behavior; SI: strategy implementation; SR: self-regulation.

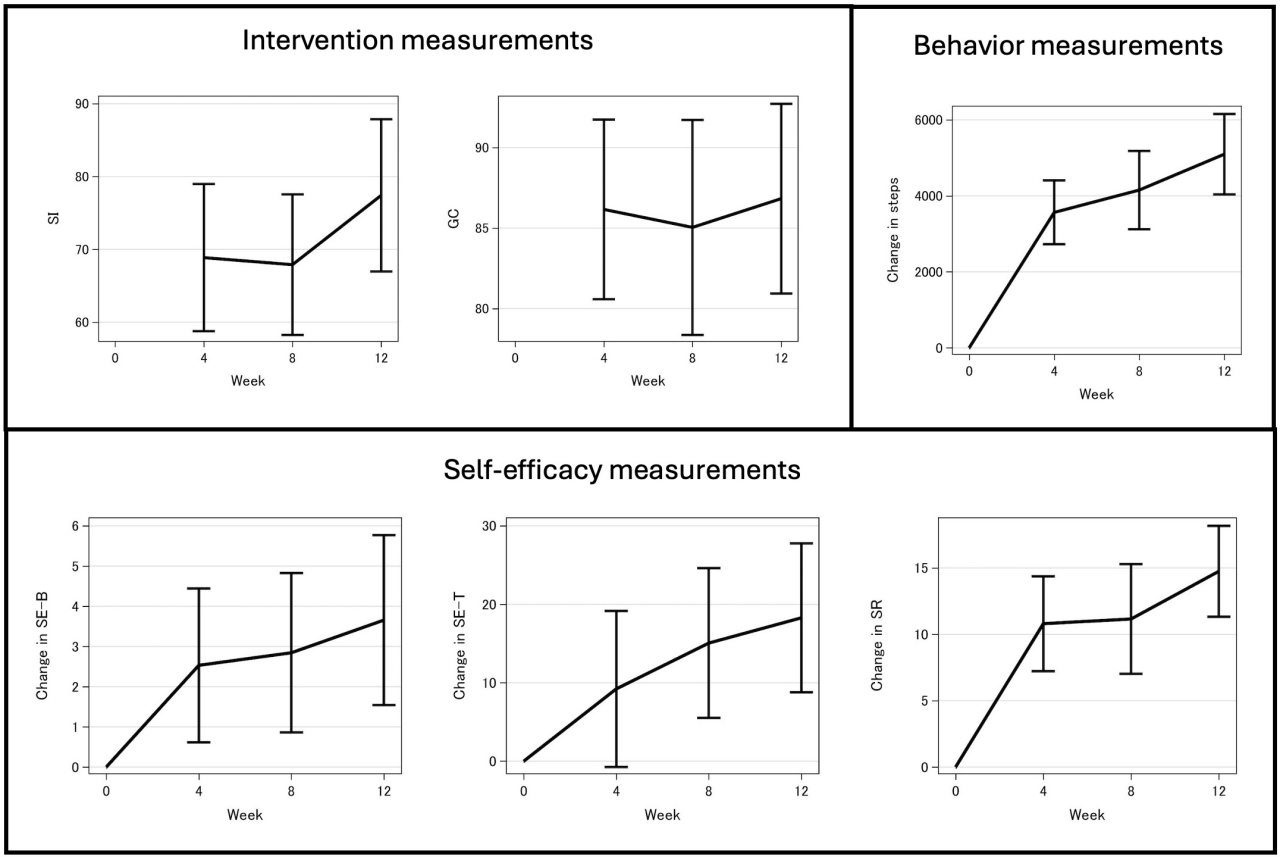

Analysis of the correlation of changes from baseline to week 12 among the 6 measurements revealed 6 relationships with statistical significance (Table 1 and Figure 2). When mapped to a simplified representation of SCT applied to this study, these relationships support but do not prove SCT's central mechanism of changes in self-efficacy causing changes in behavior. The correlation coefficient (ρ) of 0.497 between SI and change in SE-B suggests that the coping strategies target of the intervention was effective. Similarly, the ρ of 0.649 between change in steps and GC and the ρ of 0.446 between GC and change in SE-T suggest that the self-mastery target of the intervention was effective. The ρ of 0.355 between change in SR and change in steps supports the centrality of SR in the changes to exercise behavior [3].

XSL•FO
RenderX

Table . Correlations of social cognitive theory constructs at week 12 of the StepAdd intervention for Japanese patients with type 2 diabetes.

	SI[a]	GC[b]	ΔSteps	ΔSE-B[c]	ΔSE-T[d]	ΔSR[e]
SI						
ρ	1					
P value	__[f]					
GC						
ρ	0.267	1				
P value	.140	—				
ΔSteps						
ρ	0.028	0.649	1			
P value	.879	<.001	—			
ΔSE-B						
ρ	0.497	0.254	0.134	1		
P value	.004	.161	.465	—		
ΔSE-T						
ρ	0.328	0.446	0.279	0.358	1	
P value	.067	.01	.122	.04	—	
ΔSR						
ρ	−0.073	0.297	0.355	0.348	0.583	1
P value	.691	.098	.046	.051	<.001	—

[a]SI: strategy implementation.
[b]GC: goal completion.
[c]SE-B: self-efficacy for barriers.
[d]SE-T: self-efficacy for task-specific behavior.
[e]SR: self-regulation.
[f]Not applicable.

XSL•FO
RenderX

Figure 2. Social cognitive theory correlations at week 12 of the StepAdd intervention for Japanese patients with type 2 diabetes. (A) Intervention correlation between change in SE-B and SI. (B) Intervention correlation between change in steps and GC. (C) Behavior correlation between change in SR and change in steps. (D) Self-efficacy correlation between change in SE-B and change in SE-T. (E) Self-efficacy correlation between change in SE-T and GC. (F) Self-efficacy correlation between change in SE-T and SR. GC: goal completion; SE-B: self-efficacy for barriers; SE-T: self-efficacy for task-specific behavior; SI: strategy implementation; SR: self-regulation.

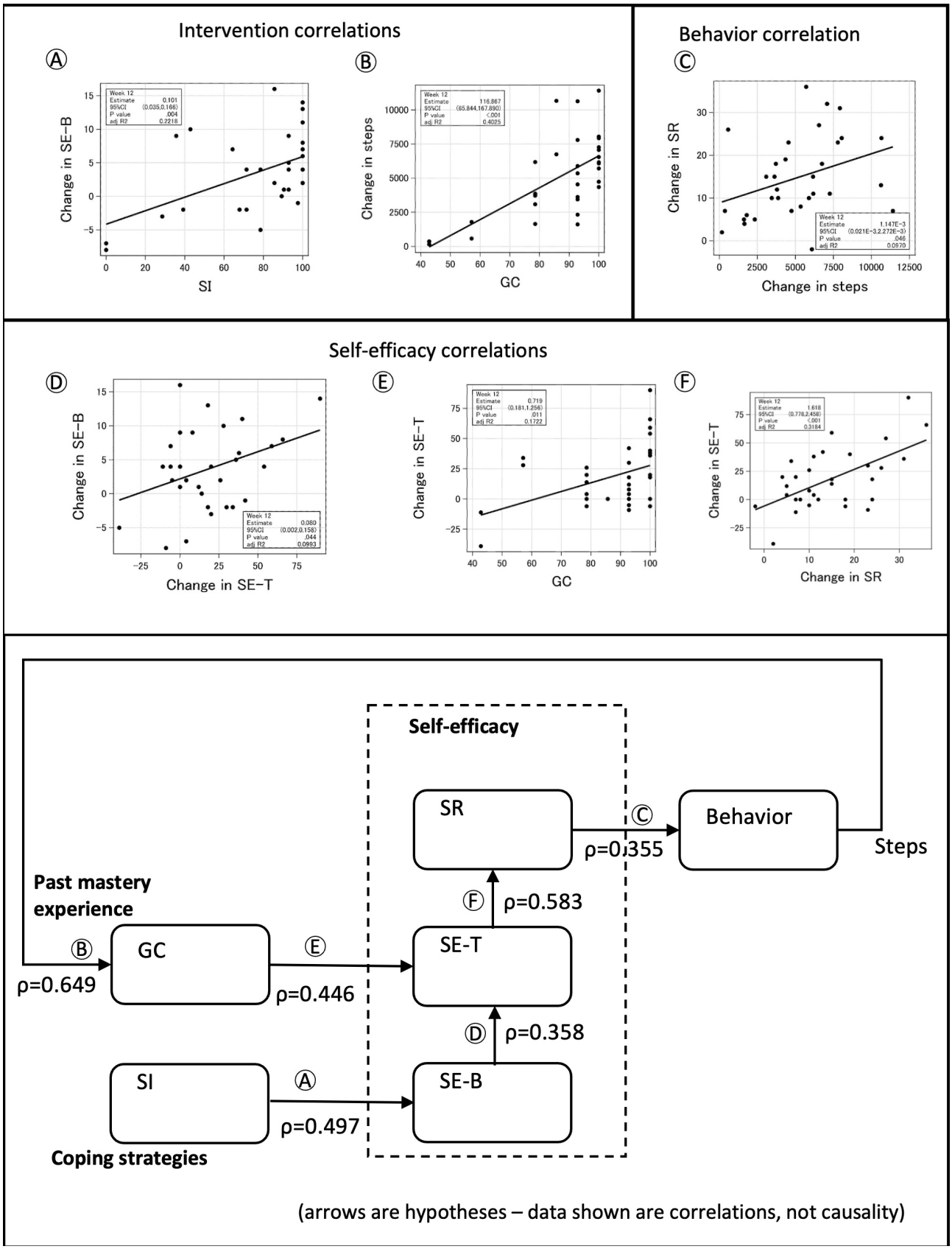

XSL•FO
RenderX

We found significant correlations among changes in the 3 measures of self-efficacy (SE-T to SE-B and SR to SE-T). We used structural equation modeling to conduct a confirmatory factor analysis to explore how the 3 self-efficacy–related measures relate to a latent variable of self-efficacy. The analysis with 3 variables and the 32-participant dataset did not support an assessment of goodness-of-fit. The StepAdd RCT currently underway [6] targets 160 patients, and this larger dataset may support a more effective confirmatory factor analysis.

We used cross-lagged panel modeling to explore whether changes in the self-efficacy measures preceded changes in the behavior, a necessary condition for the causality postulated in SCT. The analysis with our 32 participants and 4-week data spacing did not reveal any significant cross-lagged correlations.

Discussion

This study aimed to examine the SCT drivers underlying StepAdd's mean increase in exercise. Overall, our analyses found statistically significant support for the SCT approach in the StepAdd intervention. Significant correlation was found for the following parameters: step increase and GC rate; SR increase and step increase; GC rate and self-mastery increase; and coping SI rate and self-efficacy increase.

Our study findings are consistent with a study that reported that interventions that focus on removing obstacles to intended planned fitness activities are effective in increasing self-efficacy to engage in exercise [12]. Effective use of coping strategies such as problem-solving has been shown to enhance self-efficacy, which in turn promotes engagement in physical activity [13]. Self-mastery, which has been shown to be one of the most important predictors of self-efficacy for physical activity in community-dwelling older adults, is also reflected in our study findings [14]. Achieving goals allows individuals to experience a sense of mastery with authentic evidence of their capabilities, which in turn boosts self-efficacy to increase physical activity [15]. Further, SR strategies such as self-monitoring and goal setting have been associated with exercise adherence in older adults [16].

Despite the significant correlations, there was a lack of support for directional associations from the cross-lagged panel modeling results. A key limitation of this study is the difference between the timescale of behavioral decisions and the measurement intervals. Participants made physical activity decisions multiple times per day, often in response to real-time feedback from the intervention. However, the study's 4-week measurement intervals may not have fully captured these rapid behavior-feedback loops, making it difficult to assess causal relationships between intervention components and behavior change. Assessing causality is an area that requires further investigation. The pilot was relatively small, and further research is needed.

Our results provide insight into designing an effective DTx solution. Although some view DTx as centered on software implementation [17], we emphasize how DTx can remove time and space barriers, not just supporting patients but accompanying them 24/7 where they live, as a cost-effective low-labor augmentation to infrequent scheduled visits with health care professionals in a clinic.

Applying digital health to increase exercise levels is an active area of research [18], and more research is warranted. Interventions that are based on behavior change theory have been shown to be more effective than those not based on theory [19]. In our experience, building an intervention around a theory helps focus the effort on the key aspects, prioritizing them over less critical features. Our group has conducted numerous interventions not specifically based on a theory of human behavior change [2,20-23] that have produced results of interest, but we have found theory-based interventions [3,6,24] to be more effective. For example, the StepAdd study, which employs a theory-based approach, demonstrated a consistent increase in the average step count from week 1 to week 12 of the intervention [3]. In contrast, in the DialBetesPlus study, which is not theory-based, the average step count remained steady throughout the 12-month intervention period [2].

By understanding how the intervention relates to complex human behavior, researchers can put more emphasis on treatment elements that are effective over those that are not. Our results suggest that both of our intervention mechanisms, self-mastery and coping strategies, contributed to the change in behavior, supporting the SCT model of behavior change.

Acknowledgments

We thank Daniel Lane for his support in manuscript editing and scientific discussions. This work was supported by Mitsui & Co, Ltd. The funder played no role in the study design, data collection, analysis and interpretation of data, or writing of this manuscript.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author upon reasonable request for noncommercial purposes.

Authors' Contributions

KW drafted the manuscript, which was revised and approved by all authors, who also assume responsibility for its content. KW, TY, MN, and KO initiated the study. SE performed the analyses, and all authors discussed the analysis plan and results and provided input to the manuscript. All authors had access to the final study results. KW is the guarantor of this work and, as such, had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

XSL•FO
RenderX

Conflicts of Interest

A patent resulting from this research work was issued to Raxi Co, Ltd, and the University of Tokyo.

References

1. Digital Therapeutics Alliance. URL: https://dtxalliance.org/ [accessed 2025-03-04]
2. Waki K, Nara M, Enomoto S, et al. Effectiveness of DialBetesPlus, a self-management support system for diabetic kidney disease: randomized controlled trial. NPJ Digit Med 2024 Apr 27;7(1):104. [doi: 10.1038/s41746-024-01114-8] [Medline: 38678094]
3. Sze WT, Waki K, Enomoto S, et al. StepAdd: a personalized mHealth intervention based on social cognitive theory to increase physical activity among type 2 diabetes patients. J Biomed Inform 2023 Sep;145:104481. [doi: 10.1016/j.jbi.2023.104481] [Medline: 37648101]
4. Bandura A. Human agency in social cognitive theory. Am Psychol 1989 Sep;44(9):1175-1184. [doi: 10.1037/0003-066x.44.9.1175] [Medline: 2782727]
5. Beauchamp MR, Crawford KL, Jackson B. Social cognitive theory and physical activity: mechanisms of behavior change, critique, and legacy. Psychol Sport Exerc 2019 May;42:110-117. [doi: 10.1016/j.psychsport.2018.11.009]
6. Waki K, Tsurutani Y, Waki H, et al. Efficacy of StepAdd, a personalized mHealth intervention based on social cognitive theory to increase physical activity among patients with type 2 diabetes mellitus: protocol for a randomized controlled trial. JMIR Res Protoc 2024 Feb 23;13:e53514. [doi: 10.2196/53514] [Medline: 38393770]
7. Qiu S, Cai X, Schumann U, Velders M, Sun Z, Steinacker JM. Impact of walking on glycemic control and other cardiovascular risk factors in type 2 diabetes: a meta-analysis. PLoS One 2014;9(10):e109767. [doi: 10.1371/journal.pone.0109767] [Medline: 25329391]
8. Moghetti P, Balducci S, Guidetti L, et al. Walking for subjects with type 2 diabetes: a systematic review and joint AMD/SID/SISMES evidence-based practical guideline. Nutr Metab Cardiovasc Dis 2020 Oct 30;30(11):1882-1898. [doi: 10.1016/j.numecd.2020.08.021] [Medline: 32998820]
9. Yamawaki K, Harada K, Lee E, Oka K, Nakamura Y. Development of the stages of change and self-efficacy scales for walking behavior [Article in Japanese]. Nikken Kyogaku (Jpn J Health Educ Promot) 2009;17(2):87-96. [doi: 10.11260/kenkokyoiku.17.87]
10. Fukushima H. Development of a physical activity self-efficacy scale for young people [Article in Japanese]. Bulletin of the Faculty of Human Development Sciences 2011;6(1):99-106 [FREE Full text]
11. Watanabe K, Kawakami N, Adachi H, Inoue S, Meyer MRU. Internal consistency, convergent validity, and structural validity of the Japanese version of the Physical Activity Self-Regulation scale (PASR-12) among Japanese workers: a validation study. J Occup Health 2017 Jan 24;59(1):24-32. [doi: 10.1539/joh.16-0143-OA] [Medline: 27885244]
12. Artistico D, Pinto AM, Douek J, Black J, Pezzuti L. The value of removing daily obstacles via everyday problem-solving theory: developing an applied novel procedure to increase self-efficacy for exercise. Front Psychol 2013;4:20. [doi: 10.3389/fpsyg.2013.00020] [Medline: 23372560]
13. Pekmezi D, Jennings E, Marcus BH. Evaluating and enhancing self-efficacy for physical activity. ACSMs Health Fit J 2009;13(2):16-21. [doi: 10.1249/FIT.0b013e3181996571] [Medline: 29910597]
14. Warner LM, Schüz B, Wolff JK, Parschau L, Wurm S, Schwarzer R. Sources of self-efficacy for physical activity. Health Psychol 2014 Nov;33(11):1298-1308. [doi: 10.1037/hea0000085] [Medline: 24707842]
15. Bandura A. Self-Efficacy: The Exercise of Control: W H Freeman & Co; 1997.
16. McAuley E, Mullen SP, Szabo AN, et al. Self-regulatory processes and exercise adherence in older adults: executive function and self-efficacy effects. Am J Prev Med 2011 Sep;41(3):284-290. [doi: 10.1016/j.amepre.2011.04.014] [Medline: 21855742]
17. Wang C, Lee C, Shin H. Digital therapeutics from bench to bedside. NPJ Digit Med 2023 Mar 10;6(1):38. [doi: 10.1038/s41746-023-00777-z] [Medline: 36899073]
18. Mönninghoff A, Kramer JN, Hess AJ, et al. Long-term effectiveness of mHealth physical activity interventions: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2021 Apr 30;23(4):e26699. [doi: 10.2196/26699] [Medline: 33811021]
19. Glanz K, Bishop DB. The role of behavioral science theory in development and implementation of public health interventions. Annu Rev Public Health 2010;31:399-418. [doi: 10.1146/annurev.publhealth.012809.103604] [Medline: 20070207]
20. Waki K, Fujita H, Uchimura Y, et al. DialBetics: smartphone-based self-management for type 2 diabetes patients. J Diabetes Sci Technol 2012 Jul 1;6(4):983-985. [doi: 10.1177/193229681200600433] [Medline: 22920828]
21. Waki K, Fujita H, Uchimura Y, et al. DialBetics: a novel smartphone-based self-management support system for type 2 diabetes patients. J Diabetes Sci Technol 2014 Mar;8(2):209-215. [doi: 10.1177/1932296814526495] [Medline: 24876569]
22. Kondo M, Okitsu T, Waki K, Yamauchi T, Nangaku M, Ohe K. Effect of information and communication technology-based self-management system dialbeticslite on treating abdominal obesity in the specific health guidance in Japan: randomized controlled trial. JMIR Form Res 2022 Mar 24;6(3):e33852. [doi: 10.2196/33852] [Medline: 35323122]

XSL•FO
RenderX

23. Yamaguchi S, Waki K, Nannya Y, Nangaku M, Kadowaki T, Ohe K. Usage patterns of GlucoNote, a self-management smartphone app, based on ResearchKit for patients with type 2 diabetes and prediabetes. JMIR Mhealth Uhealth 2019 Apr 24;7(4):e13204. [doi: 10.2196/13204] [Medline: 31017586]
24. Shibuta T, Waki K, Miyake K, et al. Preliminary efficacy, feasibility, and perceived usefulness of a smartphone-based self-management system with personalized goal setting and feedback to increase step count among workers with high blood pressure: before-and-after study. JMIR Cardio 2023 Jul 21;7:e43940. [doi: 10.2196/43940] [Medline: 37477976]

Abbreviations

DTx: digital therapeutics
GC: goal completion
HbA_{1c}: hemoglobin A_{1c}
mHealth: mobile health
PASR-12: 12-item Physical Activity Self-Regulation Scale
RCT: randomized controlled trial
SCT: social cognitive theory
SE-B: self-efficacy for barriers
SE-T: self-efficacy for task-specific behavior
SI: strategy implementation
SR: self-regulation

Edited by A Mavragani; submitted 05.05.24; peer-reviewed by A Woodward; revised version received 15.02.25; accepted 16.02.25; published 28.03.25.

Please cite as:
Waki K, Enomoto S, Yamauchi T, Nangaku M, Ohe K
Personalized mHealth Intervention (StepAdd) for Increasing Physical Activity in Japanese Patients With Type 2 Diabetes: Secondary Analysis of Social Cognitive Theory Measurements of a Single-Arm Pilot Study
JMIR Form Res 2025;9:e60221
URL: https://formative.jmir.org/2025/1/e60221
doi:10.2196/60221

© Kayo Waki, Syunpei Enomoto, Toshimasa Yamauchi, Masaomi Nangaku, Kazuhiko Ohe. Originally published in JMIR Formative Research (https://formative.jmir.org), 28.3.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Accuracy of Smartphone-Mediated Snore Detection in a Simulated Real-World Setting: Algorithm Development and Validation

Jeffrey Brown, MD; Zachary Mitchell, BS; Yu Albert Jiang, MS; Ryan Archdeacon, BS

Bodymatter, Inc, 4343 Von Karman Ave, Suite 150J, Newport Beach, CA, United States

Corresponding Author:
Jeffrey Brown, MD
Bodymatter, Inc, 4343 Von Karman Ave, Suite 150J, Newport Beach, CA, United States

Abstract

Background: High-quality sleep is essential for both physical and mental well-being. Insufficient or poor-quality sleep is linked to numerous health issues, including cardiometabolic diseases, mental health disorders, and increased mortality. Snoring—a prevalent condition—can disrupt sleep and is associated with disease states, including coronary artery disease and obstructive sleep apnea.

Objective: The SleepWatch smartphone app (Bodymatter, Inc) aims to monitor and improve sleep quality and has snore detection capabilities that were built through a machine-learning process trained on over 60,000 acoustic events. This study evaluated the accuracy of the SleepWatch snore detection algorithm in a simulated real-world setting.

Methods: The snore detection algorithm was tested by using 36 simulated snoring audio files derived from 18 participants. Each file simulated a snoring index between 30 and 600 snores per hour. Additionally, 9 files with nonsnoring sounds were tested to evaluate the algorithm's capacity to avoid false positives. Sensitivity, specificity, and accuracy were calculated for each test, and results were compared by using Bland-Altman plots and Spearman correlation to assess the statistical association between detected and actual snores.

Results: The SleepWatch algorithm showed an average sensitivity of 86.3% (SD 16.6%), an average specificity of 99.5% (SD 10.8%), and an average accuracy of 95.2% (SD 5.6%) across the snoring tests. The positive predictive value and negative predictive value were 98.9% (SD 2.6%) and 93.8% (SD 14.4%) respectively. The algorithm performed exceptionally well in avoiding false positives, with a specificity of 97.1% (SD 3.5%) for nonsnoring files. Inclusive of all snoring and nonsnore tests, the aggregated accuracy for all trials in this bench study was 95.6% (SD 5.3%). The Bland-Altman analysis indicated a mean bias of −29.8 (SD 41.7) snores per hour, and the Spearman correlation analysis revealed a strong positive correlation (r_s=0.974; P<.001) between detected and actual snore rates.

Conclusions: The SleepWatch snore detection algorithm demonstrates high accuracy and compares favorably with other snore detection apps. Aside from its broader use in sleep monitoring, SleepWatch demonstrates potential as a tool for identifying individuals at risk for sleep-disordered breathing, including obstructive sleep apnea, on the basis of the snoring index.

(JMIR Form Res 2025;9:e67861) doi:10.2196/67861

KEYWORDS

snore detection; snore tracking; machine learning; SleepWatch; Bodymatter; neural net; mobile device; smartphone; smartphone application; mobile health; sleep monitoring; sleep tracking; sleep apnea

Introduction

The central importance of high-quality sleep in the maintenance of physical and mental health cannot be overstated, as well as the critical role of sleep in avoiding disease states and other physiologic impairments. Insufficient or poor-quality sleep has extensively documented associations with an array of negative consequences that affect both personal and public health. These issues include cardiometabolic disease states, such as hypertension, diabetes, and obesity, as well as increased frequency of heart attacks and strokes [1-3]. Sleep disruption profoundly impacts mental health and emotional well-being, increasing the frequency of depression, anxiety, inattentiveness, learning and memory impairment, and mood disorders [4-6]. From a public health perspective, poor-quality sleep correlates with workplace and vehicular accidents, and health issues arising from sleep impairment impart a massive economic burden to the health care system, as well as an increase in all-cause mortality [2,7-9].

Although the health risks of insufficient or poor-quality sleep are well known among the scientific community and society at large, most individuals struggle to obtain sleep of appropriate quality and quantity. The frequency of insufficient sleep varies by study, but as a conservative estimate, at least one-third of

XSL•FO
RenderX

the US adult population reports regularly sleeping less than the recommended 7 hours each night [10-12]. Moreover, there is evidence that sleep problems are worsening over time globally; the percentage of US adults reporting short sleep durations (defined as less than 6 h) has consistently increased over the last several decades [13,14]. The precise cause of this declining trend in sleep health is multifactorial but is influenced in large part by poor sleep hygiene related to the ongoing technological evolution of the modern world [15,16]. These concerns emphasize the need for interventions to address the profound personal and societal issue of sleep impairment.

Among other factors related to poor-quality sleep, a noteworthy and underappreciated cause of sleep impairment is the physiological phenomenon of snoring. Snoring is an auditory process that occurs during sleep as a result of air passing over relaxed tissues of the oropharynx, which produces vibrations and the sounds that are classically associated with snoring. Snoring is extremely common, especially among older adults, male individuals, and individuals with obesity; the reported frequency of habitual snoring among adults ranges from 20% to 60%, depending on risk factors [17-20]. Snoring is often benign but may be a source of personal embarrassment and a cause of relationship distress between sleep partners [19,21]. On a more concerning note, snoring may be associated with serious clinical conditions, including coronary artery disease, obstructive sleep apnea, and depressive disorders [17,22,23].

As with other areas of health optimization, smartphone-based apps have been developed to help address the need for improved sleep quality and sleep duration, and several such apps include snore detection capabilities. Because snoring occurs during sleep, individuals often lack insight into how frequently or how intensely they snore, and obtaining this information through formal polysomnography can be inconvenient, time-consuming, and cost-prohibitive. As such, there is an unmet need for novel, personalized methods of monitoring sleep quality, including evaluation of snoring and sleep-disordered breathing. There are several smartphone-based snoring detection apps in existence, and while these apps are promising in terms of their cost and scalability, most lack rigorous validation, and questions remain with regard to their real-world accuracy for individual users. This paper highlights the results of a bench study that evaluated a proprietary snore detection algorithm in a simulated real-world setting.

Methods

Study Design

The smartphone-based, acoustical snore detection algorithm used in the SleepWatch app was developed by the Bodymatter team in-house, using a deep neural net model trained on over 60,000 individually validated, real-world snore and nonsnore sounds. The SleepWatch snore detection capabilities were then tested against an array of snoring and sleep audio files in a controlled acoustic setting. Sleep audio files were derived from volunteers in an anonymized fashion, as outlined in the *Ethical Considerations* subsection. No raw or identifying human participant data were used in this study. Additionally, all training data were derived entirely from audio sources that were separate from the validation and test audio sets used in this bench study and prior trials.

In total, 36 trials were conducted based on test audio files that were 10 minutes in duration and were derived from 18 individual SleepWatch participants (male: n=9; female: n=9). Audio quality fidelity of the test files was ensured during their editing, exporting, and uploading. There were 2 audio files made from each participant, simulating a snoring index of 30, 60, 120, 240, 360, or 600 snores per hour. For each file with a snoring index of 120, 240, 360, or 600 snores per hour, 10 unique snores were repeated at regular intervals. For each file with a snoring index of 30 or 60 snores per hour, 5 unique snores were repeated at regular intervals. The duration of snore events was limited to a maximum of 5 seconds, in accordance with the scoring methodology described further in this paper. The intervening audio between inserted acoustic events was the ambient noise present at baseline during a given participant's sleep recording. Additionally, 9 trials were conducted based on test audio files that exclusively contained confounding, nonsnoring audio events (including percussive sounds, coughing, speech sounds, movement of bedding, etc). These confounding audio events reflected common ambient noise that was expected during typical sleeping conditions and had the potential to be erroneously detected as a snore. Again, no confounding acoustic event lasted longer than 5 seconds, in accordance with the scoring methodology. Of note, any audio samples that were determined to have corruption at any level of the editing or exporting process and resulted in distorted playback were discarded.

For each trial, a SleepWatch sleep recording session was started on an iPhone 13 smartphone (Apple Inc) that ran the SleepWatch app (v8.2.3.0-pkf33, public release v8.3.0) shortly before simultaneously starting the test audio playback. Test audio files were played from a dedicated JBL 305P MKII (Harman International Industries, Inc) studio speaker in close proximity to the detection smartphone to simulate product implementation in a real-world setting, where users would place their smartphone within 5 feet of their sleeping position. Trials were conducted under acoustically controlled ambient conditions. For each test, audio playback intensity was confirmed via an independent decibel meter and limited to a 125-millisecond (fast response) peak of 65 A-weighted decibels (dB[A]) and a floor of 50 dB(A). Trials were performed for 10 minutes, after which the SleepWatch app performance was recorded.

After each trial, the SleepWatch-detected snore count was compared against the known number of snoring events present in each audio file. Additionally, app performance was assessed by reporting sensitivity (true-positive rate), specificity (true-negative rate), and accuracy (ratio of true events to total events). As a scoring method, each audio file was discretized into 5-second intervals (N=120). An appropriately identified snoring event was considered a true-positive event for a 5-second interval, and conversely, a sound that was inaccurately reported as snoring was considered a false-positive event for the corresponding 5-second interval. False-negative events were missed snores, and true-negative events were 5-second regions of silence or ambient noise that were not reported as snoring. Sensitivity was calculated as "(true positive)/(true positive +

XSL•FO
RenderX

false negative)," and specificity was calculated as "(true negative)/(true negative + false positive)." Positive and negative predictive values were also determined, calculated as "(true positive)/(true positive + false positive)" and "(true negative)/(false negative + true negative)," respectively. Accuracy was determined as "(true positive + true negative)/all events" (all events: N=120). In the case of the 9 trials that used confounding nonsnore sounds, only specificity and accuracy could be described due to the absence of positive events, and this was factored into the aggregate accuracy. This scoring methodology is summarized in Table 1.

After the initial analysis, a Bland-Altman plot, which compared the average snoring index (snores/h) and difference in the snoring index between the app and test audio files, was produced as a visual method of analyzing bias associated with the app's performance. A Spearman correlation was performed between known and detected snoring index values, and the correlation coefficient (r_s) and statistical significance were determined. All statistical analyses were completed by using IBM SPSS Statistics (version 29.0.2.0; IBM Corp).

Table . Analytical validation reference table showing the schematic and equations[a,b,c,d,e] used to report snore detection performance of the SleepWatch app in this study.

		Constructed sleep audio (gold standard)	
		Snoring occurred	Snoring did not occur
Smartphone app			
	Snoring detected	True positive	False positive
	Snoring not detected	False negative	True negative

[a]Accuracy = (true positive + true negative)/(true positive + true negative + false positive + false negative).

[b]Sensitivity = true positive/(true positive + false negative).

[c]Specificity = true negative/(false positive + true negative).

[d]Positive predictive value = true positive/(true positive + false positive).

[e]Negative predictive value = true negative/(false negative + true negative).

Ethical Considerations

A principal aim of this study was to closely reproduce a real-world environment for the validation of Bodymatter software. As such, raw audio files were voluntarily obtained from individuals who were willing to participate in this study. Written consent was obtained for the use of their sleep audio for internal testing purposes. No compensation was provided to participants for their participation in this study. All audio files were anonymized at both the level of audio editing and the level of audio playback to ensure participant privacy. As no direct intervention was performed on study participants and this study posed no risk to participants, this study was granted an exemption from formal institutional review board approval for human subjects research.

Results

Across the 18 participants in this study, 9 were male and 9 were female. The age distribution skewed toward a younger adult demographic, with 28% (5/18) of participants aged 25 to 34 years, 44% (8/18) aged 35 to 44 years, 22% (4/18) aged 45 to 54 years, and 6% (1/18) aged 55 to 64 years. The average BMI was 31.21 (SD 5.03) kg/m^2 for male participants and 33.93 (SD 8.07) kg/m^2 for female participants. With respect to the presence of sleep partners, 56% (10/18) of participants reported sleeping alone, and 44% (8/18) reported having a consistent sleep partner.

The snore detection performance was evaluated for the 36 snoring tests and 9 confounding noise tests. During the development and internal validation of the SleepWatch snore detection algorithm, several trends were observed. The SleepWatch snore detection algorithm has a strong capacity to appropriately designate true-positive events (sensitivity) and an extremely strong capacity to appropriately define true-negative events (specificity); that is, the algorithm almost never reports a snore when one did not occur. Table 2 reports the sensitivity, specificity, and accuracy for both the snoring tests and the confounding audio (nonsnore) tests. The cumulative sensitivity for the snoring tests was 86.3% (SD 16.6%), the cumulative specificity was 99.5% (SD 10.8%), and the cumulative accuracy was 95.2% (SD 5.6%). The positive and negative predictive values were 98.9% (SD 2.6%) and 93.8% (SD 14.4%), respectively.

Consistent with the performance observed for the dedicated snore audio files, the SleepWatch algorithm continued to perform exceptionally well for the dedicated nonsnore audio files, with an overall specificity (ie, the true-negative capacity) of 97.1% (SD 3.5%). The sensitivity, positive predictive value, and negative predictive value cannot be reported, as there were no snoring or positive events in these audio files, but a false-positive rate of 2.9% was observed across all the noise audio files—defined as the ratio of erroneously detected snores to the total number of nonsnoring time intervals. Inclusive of the nonsnore audio files, the aggregated accuracy for all trials in this bench study was 95.6% (SD 5.3%).

XSL•FO
RenderX

Table . Detection performance of the SleepWatch app for cumulative, snoring, and nonsnore noise audio tests. The results were produced from 36 snoring trials and 9 nonsnoring noise trials.

Test condition	Sensitivity (%), mean (SD)	Specificity (%), mean (SD)	Positive predictive value (%), mean (SD)	Negative predictive value (%), mean (SD)	Accuracy (%), mean (SD)
Snore audio	86.3 (16.6)	99.5 (10.8)	98.9 (2.6)	93.8 (14.4)	95.2 (5.6)
Nonsnore audio	__[a]	97.1 (3.5)	—	—	97.1 (3.5)
Cumulative	86.3 (16.6)	98.0 (9.8)	98.9 (2.6)	93.8 (14.4)	95.6 (5.3)

[a]Not applicable.

Of note, the detected false-positive events were not evenly distributed across files. Algorithm performance was strong for coughing, percussive sounds, and loud speaking, for which no false-positive events were observed. Intermediate performance was observed for movement sounds, such as sounds resulting from pacing around a room, climbing stairs, or rustling pillows or bedding. Algorithm performance was poor for a specific file that included whispering and sleep talking, for which a false-positive rate of 10.8% was observed.

The results of the 36 snoring tests were compiled and displayed in a Bland-Altman plot (Figure 1). The average of the known versus detected snoring index (snores/h) was plotted on the x-axis, and the difference in the known versus detected snoring index was plotted on the y-axis. As expected, there were roughly 6 vertical data clusters around the x values 30, 60, 120, 240, 360, and 600 snores per hour, corresponding to the planned variations in snoring rate for each audio file. If the average of the known versus detected snores per hour aligned closely with these six x values, then excellent performance was interpreted. In contrast, if the difference in the known versus detected snores per hour was close to 0 on the y-axis across all tests, irrespective of the snoring rate, then ideal performance was also indicated. The mean bias of the SleepWatch app for this study was −29.8 (SD 41.7) snores per hour, indicating an average difference of approximately 30 snores per hour between the detected and expected snores. Most values were contained within the 95% limits of agreement (51.8 and −111.5 snores/h).

A Spearman correlation coefficient was calculated to statistically evaluate the association between the detected and known snoring index (snores/h) in each test audio file (Figure 2). The Spearman coefficient (r_s) for this cohort was 0.974, denoting an extremely strong, positive correlation between detected snores and expected snores (P<.001). These findings are highly encouraging and demonstrate a robust capacity for identifying true snores in a simulated real-world setting.

Figure 1. Bland-Altman plot of the difference in and average of the detected versus expected snores/h. The Bland-Altman plot demonstrates a mean bias of –29.8 (95% limits of agreement –111.5 to 51.8) snores/h for the average of the detected versus expected snores/h graphed against the difference in the detected versus expected snores/h. This demonstrates a relatively tight relationship between detected and expected snores.

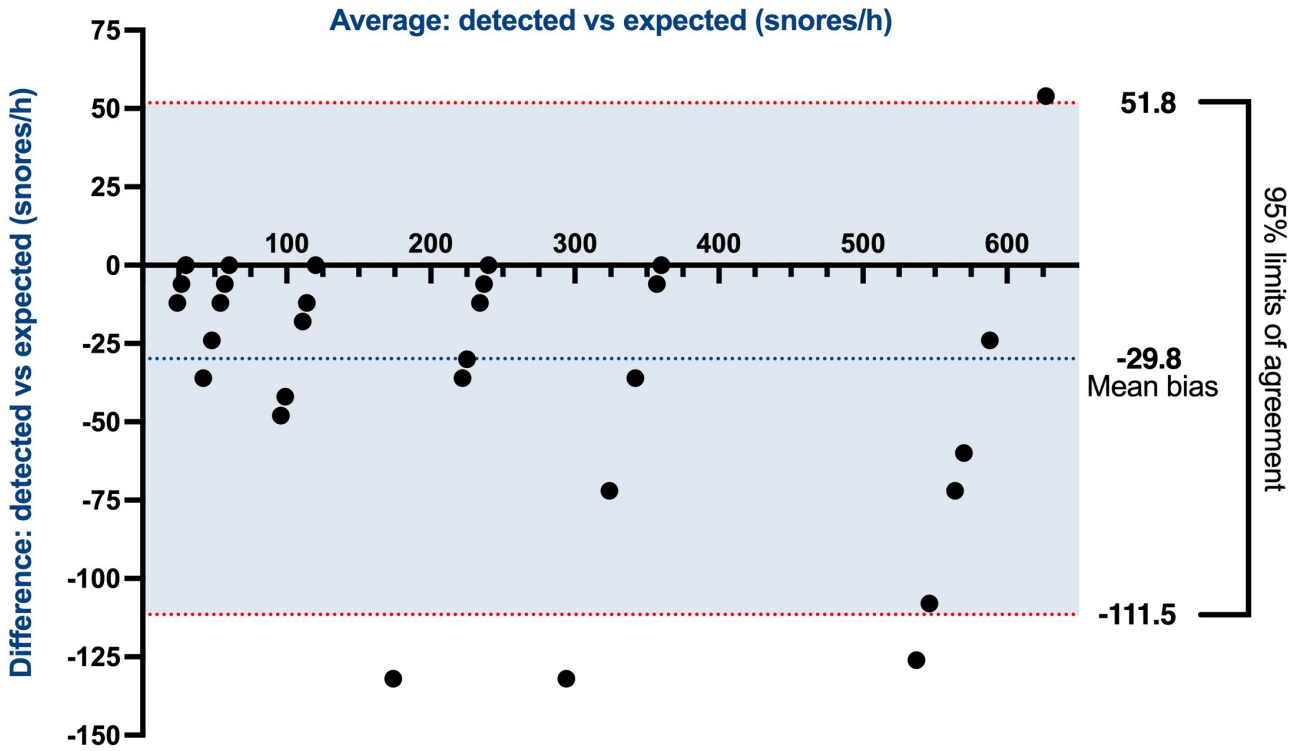

XSL•FO
RenderX

Figure 2. The expected snore index graphed against the snore index detected by the SleepWatch app. This demonstrates a tight grouping and a Spearman correlation coefficient of 0.974, indicating a strong, positive correlation between expected and detected snores.

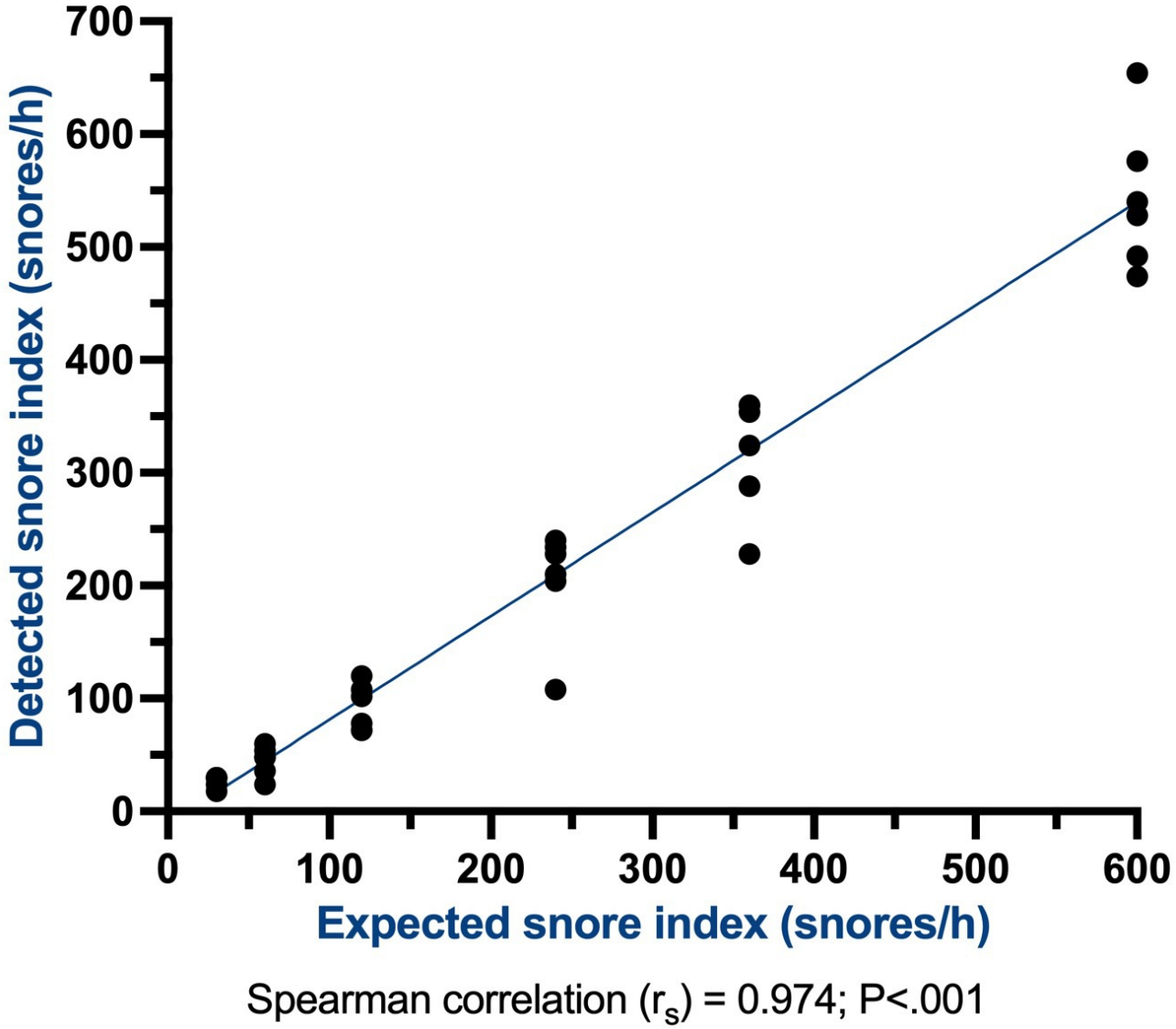

Discussion

SleepWatch Snore Detection Accuracy Compares Favorably to Similar Snore Detection Modalities

The SleepWatch app provides users with consistent, valuable insight into their sleep quality, and the snore detection feature constitutes an accuracy benchmark for acoustic snore detection software. This study reports the performance of the SleepWatch snore detection algorithm in detail; the algorithm showed an average sensitivity of 86.3% (SD 16.6%), an average specificity of 99.5% (SD 10.8%), and an average accuracy of 95.2% (SD 5.6%). Furthermore, the accuracy of the SleepWatch snore detection feature reflects performance that is superior to that of similar apps, and this feature is solely dependent on existing smartphone hardware; thus, it has powerful implications for improving sleep quality and screening for sleep-disordered breathing on an individualized basis.

There are several snore detection modalities in existence that have published results with respect to the accuracy of snore detection. The accuracy, sensitivity, specificity, positive predictive values, and negative predictive values of such modalities are summarized in Table 3, though these results must be compared with caution due to the differing methodologies between studies. Shin et al [24] integrated and analyzed a series of sleep and noise datasets for 10 individuals, attempting to distinguish verified snoring from ambient noise (playing music, running a fan, talking, etc). They reported a 95.1% overall accuracy in detecting snores and a sensitivity and specificity of 98.6% and 94.6%, respectively, as well as a positive and negative predictive value of 70.4% and 99.8%, respectively. Chiang et al [25] evaluated the sleep records of 11 patients using the Snore Clock app. The authors reported an overall mean snore detection accuracy of 95% and a sensitivity and specificity of 78.2% and 97%, respectively, along with a positive and negative predictive value of 65.3% and 97%, respectively. They also commented that the false-negative rate (type II error) was higher in patients with greater snoring frequencies. In a cohort of 19 patients, Klaus et al [26] compared the performance of the SnoreLab app with that of formal polysomnography, reporting an accuracy of 94.7% and a sensitivity and specificity of 100% and 94.1%, respectively, as well as a positive and

XSL•FO
RenderX

negative predictive value of 66.6% and 100%, respectively. These findings leave room for misinterpretation however, as these statistics were observed only in a group of participants that snored for >50% of the time recorded, constituting approximately 10% of the cohort. Inclusive of users who snored for >5% of the time recorded—a far less restrictive threshold comprising 42% of the participants—the accuracy, sensitivity, and specificity were 73.6%, 100%, and 54.5%, respectively, with a positive and negative predictive value of 61.5% and 100%. Klaus et al [26] also noted that the app had a strong tendency to overestimate the ratio of snoring time to quiet sleep time, relative to the comparison polysomnogram.

Table . Cross-study comparison of our study's snore detection app and other snore detection modalities reported in available literature. Our study's modality compares favorably against similar snore detection modalities relying on acoustic input.

Study	Accuracy, %	Sensitivity, %	Specificity, %	Positive predictive value, %	Negative predictive value, %
Our study	95.6	86.3	99.5	98.9	93.8
Shin et al [24][a]	95.1	98.6	94.6	70.4	99.8
Klaus et al [26][b]	73.6	100	54.5	61.5	100
Chiang et al [25][c]	95	78.2	97	65.3	97

[a]Shin et al [24] only used a smartphone for audio recording; snore detection was processed through a separate modality.

[b]Statistics from the Klaus et al [26] study are reported for participants with a snoring ratio of >5%, inclusive of the largest subset of participants evaluated.

[c]Chiang et al [25] selected participants from patients with known sleep apnea who were using an oral device that was intended for reducing snoring.

Collectively, the previously mentioned studies suggest that there is reasonable accuracy for smartphone-based snoring detection software, but there are several limitations with comparing our study to those snore detection studies. Notably, only 2 of those studies used working smartphone apps that recorded and processed snoring. Shin et al [24] only used a smartphone as a recording device and processed sleep audio and snore detection through a separate modality. Chiang et al [25] and Klaus et al [26] used smartphone apps with a methodology that was consistent with our study. However, Chiang et al [25] recruited participants with known sleep-disordered breathing or obstructive sleep apnea who had received a mandibular advancement device that was intended for reducing airway obstruction and snoring, which could potentially introduce a degree of selection bias and thereby limit cross-study comparison. Further, as previously stated, Klaus et al [26] reported results for an arbitrarily defined cohort that was considered to have spent >50% of the sleep session snoring, relative to quiet sleeping. App performance was far worse when the snoring threshold was >5%, with an accuracy of 73.6% in that subcohort. This discrepancy in methodology imparts limitations in the comparison of snore detection statistics between the study by Klaus et al [26] and our study, making cross-study comparison largely irrelevant.

The SleepWatch app by Bodymatter stands out for several reasons. Given that only 2 of the previously referenced studies involved a working smartphone app running snore detection software, many of the processes and protocols in the referenced studies do not seem practical to implement on an individualized, large-scale basis, whether due to concerns regarding privacy or concerns regarding convenience. Furthermore, the SleepWatch snore detection feature far outcompetes the referenced studies' modalities in positive predictivity and specificity, which, in practical language, denotes that (1) there is an extremely high likelihood that a true snore actually occurred when a snore was identified by the algorithm, and (2) the app has an extremely strong capacity to appropriately discriminate silence and nonsnoring noise events, with the latter representing one of the most significant challenges in snore detection. This reflects an engineering priority of avoiding false-positive reported snores in the development of the Bodymatter SleepWatch algorithm. Finally, this study carries the advantages of reporting performance on a snore-by-snore basis and including ambient noise and confounding sound events in the testing of the snore detection software, thereby evaluating performance in conditions that more closely reflect real-world settings and expected use.

Screening for Sleep-Disordered Breathing

As previously described, snoring is independently associated with an array of disease states, and snore detection smartphone apps provide individual users with insight into the presence and frequency of their snoring. Loud snoring is a cardinal symptom of obstructive sleep apnea—a disease characterized by airway collapse during sleep, which results in intermittent deoxygenation, apneic events, and arousal at night [27]. Furthermore, obstructive sleep apnea results in daytime sleepiness and an associated increase in accidents, as well as an increased risk of hypertension, coronary artery disease, depression, and death [28,29].

Diagnostic criteria for obstructive sleep apnea center around the apnea-hypopnea index (AHI), which describes the number of episodes where an individual stops breathing due to airway obstruction (apnea) or experiences decreased oxygenation due to inadequate breathing (hypopnea). A diagnosis of moderate-severity sleep apnea is made when >15 apneic or hypopneic events occur in 1 hour (AHI>15) [30]. Previous studies have described a positive correlation between the snoring index and AHI and, thereby, a positive correlation with the diagnosis of obstructive sleep apnea [31,32]. Moreover, it has been estimated that as many as 85% of individuals with obstructive sleep apnea are undiagnosed, highlighting the importance of novel, affordable, and scalable tools for disease screening [33].

XSL•FO
RenderX

These findings suggest that there is utility in exploring the SleepWatch app as a screening tool for stratifying users along a risk spectrum for sleep-disordered breathing, including obstructive sleep apnea. As previously mentioned, the snoring index alone is a useful metric for assessing user risk, aside from formal diagnostic criteria, such as apneic and hypopneic events. Ultimately, there is both significant need and potential for at-home technological innovations to assist with screening for underdiagnosed health conditions, including obstructive sleep apnea, and SleepWatch may have potential uses in a clinical context, aside from providing individual users with insights into their sleep health.

Limitations

Although the reported results of the SleepWatch snore detection feature are promising, this study must be evaluated with several limitations in mind. As mentioned previously, while false-positive snore events were infrequent, it is difficult to entirely eradicate such events. This is a priority for the future development of the snore detection algorithm. Further, this initial study involved a relatively small cohort, and a study with a larger sample size might more comprehensively evaluate the performance of the snore detection algorithm, due to a greater degree of acoustic heterogeneity. Additionally, there are intrinsic limitations in participant-derived audio collection, as individuals may place their phones at varying distances from the sleeping location, and the ambient acoustic environment naturally differs from person to person. From another perspective however, this acoustic diversity may represent a strength of this study, as it provides evidence that app performance is robust, irrespective of ambient sound conditions.

Conclusion

The SleepWatch app by Bodymatter has a strong track record of providing users with individualized insights into their sleep hygiene and represents a powerful and scalable modality for improving overall sleep quality. This study evaluated the performance of the deep neural net algorithm powering the SleepWatch snore detection feature. Over a series of trials involving snoring and nonsnoring sounds from 18 participants, a strong, positive correlation (r_s=0.974; P<.001) was found between detected and reference snoring indices that ranged from 30 to 600 snores per hour. The snore detection feature exhibited an overall accuracy (95.6%) that was competitive with those of similar modalities, as well as a remarkable ability to discriminate true snores from nonsnores, as reflected by the superior specificity and positive predictive value. This positions SleepWatch as a powerful tool for monitoring sleep quality and snoring. Although financial, geographic, and time constraints may prevent individuals from accessing traditional health and wellness resources, SleepWatch represents an accessible, scalable, and intuitive instrument for monitoring an individual's snoring over time and potentially identifying individuals at risk for sleep-disordered breathing, thus helping reduce the growing disease burden of impaired sleep.

Acknowledgments

No use of generative artificial intelligence was involved in the background research, writing, or editing of this study, and the writing in this paper represents original work and research by its contributors.

Data Availability

Consistent with the confidentiality agreement outlined for participants in this study, the raw and edited audio recordings used in this study are unavailable for public review. However, a summary table reporting the app performance for each participant is available in Multimedia Appendix 1. Readers may contact the corresponding author for further requests or information regarding access to data.

Conflicts of Interest

JB is a paid employee of Bodymatter, Inc. ZM and AJ are paid employees of and have stock incentives in Bodymatter, Inc. RA is a paid employee and board member of Bodymatter, Inc, and has stock incentives in Bodymatter, Inc.

Multimedia Appendix 1
Snore detection results broken down by participant, with app performance metrics for each individual.
[PDF File, 44 KB - formative_v9i1e67861_app1.pdf]

References

1. Institute of Medicine (US) Committee on Sleep Medicine and Research. Extent and health consequences of chronic sleep loss and sleep disorders. In: Colten HR, Altevogt BM, editors. Sleep Disorders and Sleep Deprivation: An Unmet Public Health Problem: National Academies Press (US); 2006. URL: https://www.ncbi.nlm.nih.gov/books/NBK19961/ [accessed 2024-10-07]
2. Medic G, Wille M, Hemels ME. Short- and long-term health consequences of sleep disruption. Nat Sci Sleep 2017 May 19;9:151-161. [doi: 10.2147/NSS.S134864] [Medline: 28579842]
3. Hanson JA, Huecker MR. Sleep Deprivation: StatPearls Publishing; 2024. URL: http://www.ncbi.nlm.nih.gov/books/NBK547676 [accessed 2024-10-07]

XSL•FO
RenderX

4. Chattu VK, Manzar MD, Kumary S, Burman D, Spence DW, Pandi-Perumal SR. The global problem of insufficient sleep and its serious public health implications. Healthcare (Basel) 2018 Dec 20;7(1):1. [doi: 10.3390/healthcare7010001] [Medline: 30577441]
5. Scott AJ, Webb TL, Martyn-St James M, Rowse G, Weich S. Improving sleep quality leads to better mental health: a meta-analysis of randomised controlled trials. Sleep Med Rev 2021 Dec;60:101556. [doi: 10.1016/j.smrv.2021.101556] [Medline: 34607184]
6. Telzer EH, Fuligni AJ, Lieberman MD, Galván A. The effects of poor quality sleep on brain function and risk taking in adolescence. Neuroimage 2013 May 1;71:275-283. [doi: 10.1016/j.neuroimage.2013.01.025] [Medline: 23376698]
7. Lim DC, Najafi A, Afifi L, et al. The need to promote sleep health in public health agendas across the globe. Lancet Public Health 2023 Oct;8(10):e820-e826. [doi: 10.1016/S2468-2667(23)00182-2] [Medline: 37777291]
8. Grandner MA. The cost of sleep lost: implications for health, performance, and the bottom line. Am J Health Promot 2018 Sep;32(7):1629-1634. [doi: 10.1177/0890117118790621a] [Medline: 30099900]
9. The Lancet Diabetes Endocrinology. Sleep: a neglected public health issue. Lancet Diabetes Endocrinol 2024 Jun;12(6):365. [doi: 10.1016/S2213-8587(24)00132-3] [Medline: 38740028]
10. Liu Y, Wheaton AG, Chapman DP, Cunningham TJ, Lu H, Croft JB. Prevalence of healthy sleep duration among adults--United States, 2014. MMWR Morb Mortal Wkly Rep 2016 Feb 19;65(6):137-141. [doi: 10.15585/mmwr.mm6506a1] [Medline: 26890214]
11. Centers for Disease Control and Prevention (CDC). Effect of short sleep duration on daily activities--United States, 2005-2008. MMWR Morb Mortal Wkly Rep 2011 Mar 4;60(8):239-242. [Medline: 21368739]
12. Grandner MA, Seixas A, Shetty S, Shenoy S. Sleep duration and diabetes risk: population trends and potential mechanisms. Curr Diab Rep 2016 Nov;16(11):106. [doi: 10.1007/s11892-016-0805-8] [Medline: 27664039]
13. Wang S, Rossheim ME, Nandy RR. Trends in prevalence of short sleep duration and trouble sleeping among US adults, 2005-2018. Sleep 2023 Jan 11;46(1):zsac231. [doi: 10.1093/sleep/zsac231] [Medline: 36152031]
14. Ford ES, Cunningham TJ, Croft JB. Trends in self-reported sleep duration among US Adults from 1985 to 2012. Sleep 2015 May 1;38(5):829-832. [doi: 10.5665/sleep.4684] [Medline: 25669182]
15. Fuller C, Lehman E, Hicks S, Novick MB. Bedtime use of technology and associated sleep problems in children. Glob Pediatr Health 2017 Oct 27;4:2333794X17736972. [doi: 10.1177/2333794X17736972] [Medline: 29119131]
16. Hysing M, Pallesen S, Stormark KM, Jakobsen R, Lundervold AJ, Sivertsen B. Sleep and use of electronic devices in adolescence: results from a large population-based study. BMJ Open 2015 Feb 2;5(1):e006748. [doi: 10.1136/bmjopen-2014-006748] [Medline: 25643702]
17. Bhattacharyya N. Sleep and health implications of snoring: a populational analysis. Laryngoscope 2015 Oct;125(10):2413-2416. [doi: 10.1002/lary.25346] [Medline: 25946644]
18. Pevernagie D, Aarts RM, De Meyer M. The acoustics of snoring. Sleep Med Rev 2010 Apr;14(2):131-144. [doi: 10.1016/j.smrv.2009.06.002] [Medline: 19665907]
19. Huang Z, Aarab G, Ravesloot MJL, de Vries N, Hilgevoord AAJ, Lobbezoo F. The annoying nature of snoring sounds is not only about intensity: a pilot study on exposed test subjects. J Clin Med 2023 Mar 31;12(7):2630. [doi: 10.3390/jcm12072630] [Medline: 37048713]
20. Knuiman M, James A, Divitini M, Bartholomew H. Longitudinal study of risk factors for habitual snoring in a general adult population: the Busselton Health Study. Chest 2006 Dec;130(6):1779-1783. [doi: 10.1378/chest.130.6.1779] [Medline: 17166996]
21. Virkkula P, Bachour A, Hytönen M, Malmberg H, Salmi T, Maasilta P. Patient- and bed partner-reported symptoms, smoking, and nasal resistance in sleep-disordered breathing. Chest 2005 Oct;128(4):2176-2182. [doi: 10.1378/chest.128.4.2176] [Medline: 16236871]
22. Hägg SA, Ilieva E, Ljunggren M, et al. The negative health effects of having a combination of snoring and insomnia. J Clin Sleep Med 2022 Apr 1;18(4):973-981. [doi: 10.5664/jcsm.9764] [Medline: 34753555]
23. Sowho M, Sgambati F, Guzman M, Schneider H, Schwartz A. Snoring: a source of noise pollution and sleep apnea predictor. Sleep 2020 Jun 15;43(6):zsz305. [doi: 10.1093/sleep/zsz305] [Medline: 31837267]
24. Shin H, Cho J. Unconstrained snoring detection using a smartphone during ordinary sleep. Biomed Eng Online 2014 Aug 15;13:116. [doi: 10.1186/1475-925X-13-116] [Medline: 25128409]
25. Chiang JK, Lin YC, Lin CW, Ting CS, Chiang YY, Kao YH. Validation of snoring detection using a smartphone app. Sleep Breath 2022 Mar;26(1):81-87. [doi: 10.1007/s11325-021-02359-3] [Medline: 33811634]
26. Klaus K, Stummer AL, Ruf S. Accuracy of a smartphone application measuring snoring in adults-how smart is it actually? Int J Environ Res Public Health 2021 Jul 8;18(14):7326. [doi: 10.3390/ijerph18147326] [Medline: 34299777]
27. Slowik JM, Sankari A, Collen JF. Obstructive Sleep Apnea: StatPearls Publishing; 2024. URL: http://www.ncbi.nlm.nih.gov/books/NBK459252 [accessed 2024-10-08]
28. Tregear S, Reston J, Schoelles K, Phillips B. Obstructive sleep apnea and risk of motor vehicle crash: systematic review and meta-analysis. J Clin Sleep Med 2009 Dec 15;5(6):573-581. [doi: 10.5664/jcsm.27662] [Medline: 20465027]
29. Harding SM. Complications and consequences of obstructive sleep apnea. Curr Opin Pulm Med 2000 Nov;6(6):485-489. [doi: 10.1097/00063198-200011000-00004] [Medline: 11100957]

XSL•FO
RenderX

30. Adult Obstructive Sleep Apnea Task Force of the American Academy of Sleep Medicine. Clinical guideline for the evaluation, management and long-term care of obstructive sleep apnea in adults. J Clin Sleep Med 2009 Jun 15;05(3):263-276. [doi: 10.5664/jcsm.27497]
31. Chiang JK, Lin YC, Lu CM, Kao YH. Snoring index and neck circumference as predictors of adult obstructive sleep apnea. Healthcare (Basel) 2022 Dec 15;10(12):2543. [doi: 10.3390/healthcare10122543] [Medline: 36554066]
32. Chiang JK, Lin YC, Lu CM, Kao YH. Correlation between snoring sounds and obstructive sleep apnea in adults: a meta-regression analysis. Sleep Sci 2022;15(4):463-470. [doi: 10.5935/1984-0063.20220068] [Medline: 36419807]
33. Motamedi KK, McClary AC, Amedee RG. Obstructive sleep apnea: a growing problem. Ochsner J 2009;9(3):149-153. [Medline: 21603432]

Abbreviations

AHI: apnea-hypopnea index
dB(A): A-weighted decibels

Edited by A Mavragani; submitted 22.10.24; peer-reviewed by M Agbede, X Gao; revised version received 24.02.25; accepted 25.02.25; published 28.03.25.

Please cite as:
Brown J, Mitchell Z, Jiang YA, Archdeacon R
Accuracy of Smartphone-Mediated Snore Detection in a Simulated Real-World Setting: Algorithm Development and Validation
JMIR Form Res 2025;9:e67861
URL: https://formative.jmir.org/2025/1/e67861
doi:10.2196/67861

© Jeffrey Brown, Zachary Mitchell, Yu Albert Jiang, Ryan Archdeacon. Originally published in JMIR Formative Research (https://formative.jmir.org), 28.3.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Mobile Health Tool to Capture Social Determinants of Health and Their Impact on HIV Treatment Outcomes Among People Who Use Drugs: Pilot Feasibility Study

Rachel E Gicquelais[1], PhD, MPH; Caitlin Conway[2], BSN; Olivia Vjorn[3], MS; Andrew Genz[4], MPH; Gregory Kirk[4,5], PhD, MD; Ryan Westergaard[1], MD, PhD, MPH

[1]School of Medicine and Public Health, University of Wisconsin-Madison, 603 WARF Office Building, 610 Walnut Street, Madison, WI, United States
[2]University of Wisconsin-Madison School of Nursing, Madison, WI, United States
[3]Center for Health Enhancement Systems Studies, University of Wisconsin-Madison College of Engineering, Madison, WI, United States
[4]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, United States
[5]Division of Infectious Diseases, Johns Hopkins University School of Medicine, Baltimore, MD, United States

Corresponding Author:
Rachel E Gicquelais, PhD, MPH
School of Medicine and Public Health, University of Wisconsin-Madison, 603 WARF Office Building, 610 Walnut Street, Madison, WI, United States

Abstract

Background: Active substance use, food or housing insecurity, and criminal legal system involvement can disrupt HIV care for people living with HIV and opioid use disorder (OUD). These social determinants of health are not routinely captured in clinical settings.

Objective: We evaluated whether real-time reports of social and behavioral factors using a smartphone app could predict viral nonsuppression and missed care visits to inform future mobile health interventions.

Methods: We enrolled 59 participants from the AIDS Linked to the Intravenous Experience (ALIVE) Study in Baltimore, Maryland, into a 12-month substudy between February 2017 and October 2018. Participants were eligible if they had OUD and had either a measured HIV RNA ≥1000 copies/mL or a ≥1-month lapse in antiretroviral therapy in the preceding 2 years. Participants received a smartphone and reported HIV medication adherence, drug use or injection, and several disruptive life events, including not having a place to sleep at night, skipping a meal due to lack of income, being stopped by police, being arrested, or experiencing violence on a weekly basis, through a survey on a mobile health app. We described weekly survey completion and investigated which factors were associated with viral nonsuppression (HIV RNA ≥200 copies/mL) or a missed care visit using logistic regression with generalized estimating equations adjusted for age, gender, smartphone comfort, and drug use.

Results: Participants were predominantly male (36/59, 61%), Black (53/59, 90%), and had a median of 53 years old. At baseline, 16% (6/38) were virally unsuppressed. Participants completed an average of 23.3 (SD 16.3) total surveys and reported missing a dose of antiretroviral therapy, using or injecting drugs, or experiencing any disruptive life events on an average of 13.1 (SD 9.8) weekly surveys over 1 year. Reporting use of any drugs (adjusted odds ratio [aOR] 2.3, 95% CI 1.4 - 3.7), injecting drugs (aOR 2.3, 95% CI 1.3 - 3.9), and noncompletion of all surveys (aOR 1.6, 95% CI 1.1 - 2.2) were associated with missing a scheduled care visit over the subsequent 30 days. Missing ≥2 antiretroviral medication doses within 1 week was associated with HIV viral nonsuppression (aOR 3.7, 95% CI: 1.2 - 11.1) in the subsequent 30 days.

Conclusions: Mobile health apps can capture risk factors that predict viral nonsuppression and missed care visits among people living with HIV who have OUD. Using mobile health tools to detect sociobehavioral factors that occur prior to treatment disengagement may facilitate early intervention by health care teams.

(JMIR Form Res 2025;9:e59953) doi:10.2196/59953

KEYWORDS

HIV; drug use; social determinants of health; mobile health; mHealth; smartphone

XSL•FO
RenderX

Introduction

In 2019, the US Department of Health and Human Services launched the Ending the HIV Epidemic initiative, establishing a goal to reduce new infections of HIV in the United States by 90% by the year 2030 [1]. Achieving this ambitious goal hinges on rapid linkage of newly diagnosed individuals to antiretroviral therapy (ART) and sustained HIV viral suppression to prevent transmission. Active injection drug use and use of other substances, such as alcohol, are associated with inconsistent adherence to ART and viral nonsuppression [2-5]. Accordingly, people who use drugs and are living with HIV are among the most vulnerable to disruptions in care that threaten sustained viral suppression.

The link between drug and alcohol use and viral nonsuppression often relates to unmet needs broadly defined as social determinants of health [6]. For example, housing instability among people who use drugs has been linked to an increased risk of both acquiring HIV and viral nonsuppression [6]. Injecting drugs raises the likelihood of unstable housing and hinders improvements in housing situations among people living with HIV [7,8], thereby threatening viral suppression. Further, inadequate access to food is highly prevalent among people who use drugs [9,10] and has been associated with viral nonsuppression [11], mortality [12], and engaging in behaviors that raise HIV transmission risk, such as receptive syringe sharing [13] and unprotected sexual intercourse [14]. The criminalization of drug use and social stigma of HIV further contribute to difficulties in engaging people who use drugs in both HIV and substance use disorder treatment [15] as well as to disruptions in care due to arrest or incarceration [16,17]. The link between social determinants of health and patients' capacity to engage and sustain in HIV care illustrates a need for innovative strategies that provide timely support to address the root causes of care disruption and viral nonsuppression.

Mobile health (mHealth) is defined broadly by the World Health Organization as "mobile and wireless technologies to support the achievement of health objectives" [18]. mHealth tools can provide a platform that facilitates real-time detection and the provision of timely assistance when events that could disrupt care engagement occur. Further, information captured through mHealth apps could detect early warning signs of care disengagement, such as missed ART doses or increases in drug use. These technologies may be particularly well-suited to augment existing patient-centered, comprehensive models of HIV care that are enabled by federal Ryan White funding in the United States [19], as these care models already feature medical case management, food pantries, support to obtain housing, and other services to address social determinants of health. In the present pilot feasibility study, we explored the potential of a smartphone app to detect disruptive life events and early warning signs of HIV care disengagement over a 1-year period among 59 people who use drugs living with HIV. Using a weekly survey within a smartphone app designed to support substance use disorder recovery, we assessed how often people who use drugs used the smartphone app and reported experiencing disruptive life events. Through subsequent linkage with electronic health record data, we assessed whether these disruptive life events were associated with 2 outcomes representing care disengagement: viral nonsuppression and missed care visits.

Methods

Participants, Recruitment, and Follow-Up

This manuscript reports on a pilot feasibility study of an mHealth app designed to detect disruptive events that could threaten sustained engagement in HIV care and examine the preliminary associations of these disruptive events with HIV care engagement. Participants for this study came from the AIDS Linked to the IntraVenous Experience (ALIVE) study, a prospective cohort study of people who inject drugs in Baltimore, Maryland, that has been described previously [20]. A total of 5506 individuals have been enrolled since the study began in 1988, and approximately 524 were living and had HIV at the time enrollment for the present substudy, which began in 2017 [21]. To enroll in the ALIVE study, participants must be aged 18 years or older and have a history of injection drug use. Participants are tested for HIV at enrollment. Follow-up is ongoing and involves twice-yearly visits to the ALIVE clinic where participants complete surveys about health history in the past 6 months and repeat HIV testing. They also provide consent for the researchers to access medical records to obtain health care utilization data.

The present analysis includes 59 ALIVE study participants living with HIV who were recruited through the HIV care clinic affiliated with Johns Hopkins Hospital in Baltimore, Maryland, United States. Medical records from ALIVE participants who had received HIV care at Johns Hopkins were reviewed to determine eligibility. If eligible, they were invited to meet with a study coordinator to complete a brief screening survey and enroll in the study. Recruitment flyers were also posted in the ALIVE study office and the Johns Hopkins clinic, inviting patients to self-refer for eligibility screening. The inclusion criteria were (1) ever had opioid use disorder (OUD), (2) used a nonprescribed opioid in the past year, (3) not treated with ART for ≥1 month in the past 2 years or had an HIV RNA viral load of ≥1000 copies/mL in the past 2 years, (4) a patient of the local HIV clinic from which participants were recruited, and (5) a participant of the ALIVE study.

All enrolled substudy participants received a smartphone with an unlimited voice and data plan and underwent a brief training with a study coordinator on using the study mHealth app. Participants completed an interviewer-administered baseline survey and follow-up surveys every 3 months for a 1-year period. The study team received HIV viral load testing results from the electronic health record at the local HIV clinic where recruitment was conducted and from routine semiannual testing performed as part of the ALIVE study follow-up. The date of all scheduled care visits at the local HIV clinic and whether visits were missed or attended was also collected. Enrollment and follow-up occurred during 2017 and 2018.

Ethical Considerations

This study was approved by the institutional review boards at the University of Wisconsin-Madison (protocol ID:

XSL•FO
RenderX

2016‐1190) and Johns Hopkins University (protocol ID: IRB00007523). Results were reported using the CONSORT (Consolidated Standards of Reporting Trials) guidelines for pilot studies. Written informed consent was obtained from all individual participants included in the study. All data collected were coded using a unique study identification number to ensure that individual participants could not be identified. Participants were compensated US $20 cash for completing a baseline survey and another US $20 for each completed follow-up survey. They also received a bonus of US $10 remuneration for each completed survey at the end of the 1-year follow-up period (up to US $50).

The A-CHESS mHealth App

The mHealth system used in this study, the Addiction-Comprehensive Health Enhancement Support System (A-CHESS), was designed to support recovery from alcohol use disorder and was found to reduce drinking in a prior study of people with alcohol use disorder [22]. A-CHESS has since been adapted for several other risk groups, including those living with OUD [23,24]. The theoretical framework supporting A-CHESS is the self-determination theory [25], which emphasizes that adaptive functioning to manage health conditions is improved by meeting 3 needs: competence (ie, the ability to cope with challenges), social relatedness (ie, having meaningful relationships and a feeling of connection), and autonomy (ie, a sense of control over behavior and goals). A-CHESS is therefore organized around these 3 pillars. Competence is supported by an extensive resource library containing podcasts, information, and other resources to support recovery. An in-app discussion board and private messaging function support relatedness by allowing app users to connect with one another and with case managers or providers who also have accounts on the app. Weekly self-monitoring surveys and a goal-setting feature within the app provide opportunities to build autonomy. Participants receive notifications to complete the self-monitoring survey on a weekly basis. For this study, the weekly self-monitoring survey was adapted to investigate potentially disruptive events that could threaten sustained engagement in HIV care, such as being arrested or not having a place to live, as well as events that may signal or raise risk of HIV care disengagement, such as missed ART doses and drug use (further described in Measures). Participants were invited to use A-CHESS on their study-provided cell phone for 12 months.

Measures

At the study enrollment visit, participants provided informed consent and completed a survey administered by study staff. They reported their HIV diagnosis date and whether they were currently using ART. Regarding drug use, participants reported whether they had used any illegal or street drugs (including marijuana) or used any medications not as prescribed in the past 30 days, what drugs they used, and whether they were currently taking any medications for OUD (including methadone, buprenorphine, or naltrexone). Participants also reported whether they had ever been diagnosed with any mental health conditions, including anxiety, depression, bipolar disorder, attention deficit hyperactivity disorder, obsessive-compulsive disorder, panic disorder, or post-traumatic stress disorder. Participants reported their comfort level using a smartphone on a numeric scale (1=not at all comfortable) to (5=very comfortable). Finally, participants reported sociodemographic characteristics, including age, gender, race, ethnicity, education, and employment.

The 2 primary outcomes of the study were abstracted through a review of electronic health records and ALIVE semiannual study visit data available for each participant during the 1-year follow-up period. These included viral nonsuppression, defined by the Centers for Disease Control and Prevention as an HIV viral load of ≥200 copies/mL (vs viral load of <200 copies/mL) [26] and missing a scheduled care visit, defined as missing any scheduled visit to the local HIV care clinic affiliated with the study (vs attending the visit). The HIV viral load at baseline was considered the closest HIV measurement to a participant's baseline survey within the period of 1 year before to 7 days after study enrollment. Electronic health records were also used to characterize baseline CD4+ count defined as the closest CD4 measurement to a participant's baseline survey within the 1 year before to 60 days after the baseline date. For the missed visit outcome, a total of 841 care visits were included in the analysis. These 841 visits include scheduled visits (regardless of visit attendance) with a medical provider, nurse, case manager, or mental health provider to reflect the breadth of care that patients living with HIV may receive through the United States Health Resources & Services Administration Ryan White HIV/AIDS Program. The primary outcome was whether a scheduled visit was missed versus attended. A total of 147 scheduled visits that were canceled before the appointment occurred were excluded from analyses.

The weekly self-monitoring survey included questions about the number of days ART was not taken, any drug use (defined by combining separate questions on use of heroin, any other opioids (including oxycodone, morphine, fentanyl, etc), cocaine, methamphetamine, or sedatives (including Valium, Ativan, or Xanax)), any injection drug use, and several potentially disruptive life events over the past 7 days. Potentially disruptive life events included not having a place to sleep, skipping a meal because of not having enough income, being stopped by police, being put in jail, and being robbed or beat up. Responses to questions on the weekly self-monitoring survey were used to form several binary indicator variables summarizing whether events of interest were reported on any weekly survey completed in the 30 days prior to having a viral load measurement and separately, in the 30 days prior to a scheduled care visit. Binary variables for noncompletion of all weekly surveys in the 30 days prior to each outcome and for reporting any significant event (ie, missed ART dose, any drug use, injecting drugs, or any disruptive life event) in the 30 days prior to each outcome were also created. These indicators were analyzed as the primary independent variables for the study. A-CHESS system use was summarized as the average number of days A-CHESS was used and the average number of weekly surveys completed per participant over the entire 1-year study period and per study month.

XSL•FO
RenderX

Analysis

Descriptive characteristics of the sample at baseline were calculated and compared between those who had at least 1 HIV viral load ≥200 copies/mL (vs those for whom all HIV test results across the 1-year study period indicated viral suppression). These characteristics were also compared for those who missed ≥50% of their care visits (vs <50%). Engagement with the A-CHESS smartphone app was described in the total sample and by the primary outcomes. To characterize whether reporting missed ART doses, any drug use, injecting drugs, or experiencing a disruptive life event were associated with viral nonsuppression or a missed care visit, logistic regression with generalized estimating equations accounting for repeated outcome measurements for each participant were used to estimate odds ratios with 95% CIs, as participants could have multiple HIV viral loads taken or care visits scheduled over the 1-year follow-up period. Unadjusted models and models adjusted for several confounders selected a priori were computed. Confounders included in adjusted models were sociodemographic characteristics (age, race, ethnicity, and gender), comfort using a smartphone (treated as a continuous variable with a range of 1-5), and whether the participant had used any drugs in the past 30 days, all of which were reported at study enrollment. In a sensitivity analysis, models were run with an alternate definition of HIV viral nonsuppression, ≥1000 copies/mL, as this matched one of the eligibility criteria for the study and was used to define virological failure in a prior, foundational study of HIV transmission [27].

Results

Participant Characteristics

A total of 59 participants enrolled in the 1-year study (Table 1). At baseline, participants were a mean age of 53 years and predominantly male (n=36, 61%), Black race (n=53, 90%), and non-Hispanic ethnicity (n=54, 92%). Approximately 52% (n=31) had less than a high school education, and only 5% (n=3) were employed at the time of the study. In the month before enrollment, 83% (n=49) had used any drugs, 71% (n=42) used opioids, 48% (n=28) used stimulants, and 27% (n=16) had injected drugs. While nearly all (n=55, 93%) reported currently taking ART, 16% (n=6) were virally nonsuppressed (defined as having a viral load of ≥1000 copies/mL at baseline), and 20% (n=10) had a CD4+ count <200 cells/mm^3. Participants had been living with HIV for a mean of 19 years prior to study enrollment, with 8% (n=5) having been diagnosed in the last 5 years.

XSL•FO
RenderX

Table . Descriptive characteristics at baseline among a sample of 59 people living with HIV and opioid use disorder enrolled in a pilot feasibility study of the A-CHESS (Addiction-Comprehensive Health Enhancement Support System) mobile health app during 2017 and 2018.

Characteristic	Total (n=59)	HIV viral suppression		Missed care visits	
		≥1 test ≥200 copies/mL (n=25)	All tests <200 copies/mL (n=27)	Missed ≥50% visits (n=23)	Missed <50% visits (n=33)
Age (years), median (IQR)	53 (11)	53 (9.0)	53 (12.5)	51 (10.0)	53 (11.1)
Female (vs male), n (%)	23 (39)	11 (44)	10 (37)	7 (30)	15 (46)
Black or African American race (vs another race), n (%)	53 (90)	25 (100)	24 (89)	21 (91)	29 (88)
Hispanic or Latinx (vs non-Hispanic), n (%)	5 (8)	0 (0)	4 (15)	2 (9)	3 (9)
Less than high school education (vs high school or more), n (%)	31 (52)	12 (48)	14 (52)	12 (52)	17 (52)
Employed (vs unemployed), n (%)	3 (5)	1 (4)	2 (7)	3 (13)	0 (0)
Comfort using a smartphone, mean (SD)	3.8 (1.3)	3.8 (1.4)	3.7 (1.3)	3.8 (1.3)	3.9 (1.3)
Any mental health concerns (vs none), n (%)	40 (68)	15 (60)	20 (74)	15 (65)	25 (76)
Used drugs in past 30 days (vs none), n (%)	49 (83)	19 (76)	25 (93)	20 (87)	27 (82)
Used opioids in past 30 days (vs none), n (%)	42 (71)	16 (64)	22 (82)	16 (70)	25 (76)
Used stimulants in past 30 days (vs none), n (%)	28 (48)	11 (44)	15 (56)	13 (56)	15 (46)
Injected drugs in past month (vs not), n (%)	16 (27)	6 (24)	7 (26)	7 (30)	8 (24)
Years living with HIV, mean (SD)	19.4 (8.7)	19.1 (7.5)	20.0 (9.2)	17.8 (7.2)	20.4 (9.1)
Currently taking antiretroviral therapy (vs not), n (%)	55 (93)	26 (96)	23 (92)	22 (96)	30 (91)
HIV RNA ≥1,000 copies/mL[a] (vs<1,000), n (%)	6 (16)	5 (29)	1 (6)	2 (14)	4 (19)
CD4+ count < 200cells/mm^3[b] (vs ≤200), n (%)	10 (20)	8 (36)	1 (4)	8 (40)	1 (4)

[a]37 participants had an available HIV viral load result at baseline.

[b]50 participants had an available CD4 count at baseline.

Of 52 participants who had at least 1 HIV viral load test during follow-up (comprising 237 total viral load tests), nearly half of the participants (n=25, 48%) had at least 1 detectable viral load of ≥200 copies/mL during the follow-up period. Regarding missed care visits, of 56 participants who had a visit scheduled during the follow-up period (comprising 841 total scheduled visits), 41% (n=23) of participants missed at least half of their scheduled care visits during the study period. The modest sample size limited our ability to test for differences in baseline characteristics between those who were virally nonsuppressed at least once over the follow-up period or who missed over half of their care visits, as shown in Table 1.

A-CHESS Engagement and Weekly Reports of Missed ART Doses, Drug Use, and Disruptive Life Events

Over the 1-year follow-up period, participants used A-CHESS an average of 119.4 (SD 82.8) total days and on an average of 11.5 (SD 4.4) days per month. The weekly survey was

XSL•FO
RenderX

completed an average of 23.3 (SD 16.3) times per participant across the follow-up period, leaving a total of 1353 weekly surveys available for analysis (Table 2). In total, 46.3% (627/1353) of weekly surveys included a report of missed ART doses, drug use, or a disruptive life event. Missing a dose of ART was reported on 25.2% (341/1353) of surveys, an average of 8.5 (SD 7.1) times per participant across the study period. Using any drugs was reported an average of 9.5 (SD 9.3) times per participant across the study period, and injecting drugs was reported 5.8 (SD 8.2) times per participant. Skipping a meal was reported an average of 6.6 (SD 6.5) times per participant. Not having a place to sleep was reported an average of 2.7 (SD 3.4) times per participant. Other disruptive life events (ie, being put in jail and being robbed or beaten up) were reported <2 times, on average, per participant.

Table . Missed antiretroviral therapy (ART) doses, drug use, and disruptive life events reported on weekly surveys completed by 59 participants living with HIV and opioid use disorder enrolled in a pilot feasibility study of the A-CHESS (Addiction-Comprehensive Health Enhancement Support System) mobile health app during 2017 and 2018.

Survey item	Number of weekly surveys, n (%)	Number of weekly surveys completed per participant, mean (SD)
Total number of weekly surveys completed	1353 (100)	23.3 (16.3)
Reported any missed ART, drug use, or disruptive life event	627 (46.3)	13.1 (9.9)
1 or more missed ART doses	341 (25.2)	8.5 (7.1)
2 or more missed ART doses	190 (14)	5.9 (6.1)
Any drug use	417 (30.8)	9.5 (9.3)
Injected drugs	116 (8.6)	5.8 (8.2)
Did not have a place to sleep at night	27 (2)	2.7 (3.4)
Skipped a meal because I didn't have enough money	131 (9.7)	6.6 (6.5)
Was stopped by police	22 (1.6)	1.8 (1.2)
Was put in jail	9 (0.7)	1.5 (0.6)
Was robbed or beat up	13 (1)	1.6 (0.9)

Table S1 in Multimedia Appendix 1 shows differences in A-CHESS utilization frequency and weekly survey completion during follow-up between participants who were virally nonsuppressed at least once over the follow-up period (vs suppressed throughout follow-up) as well as between those who missed 50% or more of their scheduled care visits (vs who missed fewer than 50% of their care visits). These comparisons indicate that more weekly surveys were completed by those who did not have any viral loads ≥200 copies/mL (P=.01) and who did not miss any care visits (P=.005). Those who did not miss any care visits during follow-up also tended to use A-CHESS more often (P=.03).

Association of Missed ART, Drug Use, and Disruptive Life Events Reported Through A-CHESS With Missed Care Visits and Viral Nonsuppression

In bivariable and adjusted analyses, injecting drugs, any drug use, and not completing any weekly surveys in the prior month were associated with missing a scheduled care visit in the subsequent 30 days (Table 3). Specifically, injecting drugs was associated with 2.3-fold higher adjusted odds (95% CI 1.3 - 3.9) and using any drugs was associated at a similar magnitude with missing a scheduled care visit after adjustment for age, comfort using a smartphone, gender, and any drug use at baseline. Not completing the weekly survey (vs completing at least 1 weekly survey) was associated with 57% higher odds of missing a care visit in the following 30 days (adjusted odds ratio [aOR] 1.6, 95% CI 1.1 - 2.2).

XSL•FO
RenderX

Table . Association of reporting missed antiretroviral therapy (ART) doses, drug use, and disruptive life events in the prior 30 days with missing a care visit among 59 participants living with HIV and opioid use disorder enrolled in a pilot feasibility study of the A-CHESS (Addiction-Comprehensive Health Enhancement Support System) mobile health app during 2017 and 2018.

Events reported in the month prior to scheduled care visit[a]	Unadjusted OR[b] (95% CI)	Adjusted OR (95% CI)[c]
Any missed ART, drug use, or disruptive life event	1.20 (0.80 - 1.81)	1.25 (0.81 - 1.93)
1 or more missed ART doses	0.91 (0.61 - 1.35)	0.89 (0.57 - 1.39)
2 or more missed ART doses	1.22 (0.70 - 2.15)	1.30 (0.69 - 2.43)
Any drug use	2.07 (1.28 - 3.35)	2.27 (1.41 - 3.66)
Injected drugs	2.11 (1.24 - 3.59)	2.26 (1.30 - 3.91)
Did not have a place to sleep at night	1.87 (0.38 - 9.27)	1.76 (0.28 - 11.14)
Skipped a meal due to not enough money	0.76 (0.36 - 1.60)	0.81 (0.37 - 1.74)
Stopped by police	1.07 (0.45 - 2.54)	1.11 (0.42 - 2.97)
No weekly surveys completed in past 30 days	1.42 (1.00 - 2.02)	1.57 (1.10 - 2.22)

[a]Odds ratios are reported relative to a referent group that completed at least 1 weekly survey in the 30 days prior to a scheduled care visit but did not indicate experiencing the disruptive life event.

[b]OR: odds ratio.

[c]Adjusted for age, gender, comfort using a smartphone (score: 1=not at all comfortable to 5=very comfortable), and any drug use in the past 30 days, all of which were reported at study enrollment.

In both unadjusted and adjusted analyses, missing 2 or more doses of ART was associated with having a nonsuppressed viral load in the subsequent 30 days (aOR 3.7, 95% CI 1.2 - 11.1; Table 4). These associations were similar in a sensitivity analysis in which the outcome was defined as a viral load of load ≥1000 (Table S2 in Multimedia Appendix 1). Notably, we were unable to adjust for race or ethnicity given the predominantly Black and non-Hispanic racial or ethnic makeup of the sample (90% Black and 8% Hispanic). Further, because being put in jail and robbed or beaten up were each reported on only a few surveys (<15 surveys each), we were unable to analyze their relationship with the primary outcomes of interest.

Table . Association of reporting missed antiretroviral therapy (ART) doses, drug use, and disruptive life events in the prior 30 days with HIV viral nonsuppression (viral load ≥200 copies/mL) among 59 participants living with HIV and opioid use disorder enrolled in a pilot feasibility study of the A-CHESS (Addiction-Comprehensive Health Enhancement Support System) mobile health app during 2017 and 2018.

Events reported in the month prior to viral load measurement[a]	Unadjusted OR[b] (95% CI)	Adjusted OR (95% CI)[c]
Any missed ART, drug use, or disruptive life event	1.30 (0.72 - 2.33)	1.28 (0.68 - 2.44)
1 or more missed ART doses	1.96 (0.88 - 4.35)	1.92 (0.76 - 4.76)
2 or more missed ART doses	4.00 (1.61 - 10.00)	3.70 (1.23 - 11.11)
Any drug use	0.93 (0.46 - 1.85)	0.92 (0.44 - 1.92)
Injected drugs	0.79 (0.22 - 2.86)	0.74 (0.19 - 2.78)
Did not have a place to sleep at night	2.38 (0.56 - 10.00)	2.44 (0.54 - 11.11)
Skipped a meal due to not enough money	0.64 (0.25 - 1.67)	0.6 (0.22 - 1.64)
Stopped by police	1.72 (0.34 - 8.33)	1.67 (0.32 - 9.09)
No weekly surveys completed in past 30 days	0.94 (0.53 - 1.69)	0.87 (0.46 - 1.64)

[a]Odds ratios are reported relative to a referent group that completed at least 1 weekly survey in the 30 days prior to a viral load test but did not indicate experiencing the disruptive life event.

[b]OR: odds ratio.

[c]Adjusted for age, gender, comfort using a smartphone (score: 1=not at all comfortable to 5=very comfortable), and any drug use in the past 30 days, all of which were reported at study enrollment.

Discussion

This study is a proof-of-concept demonstrating the utility and feasibility of using an mHealth system to capture time-varying threats to HIV viral suppression and care engagement among people living with HIV and OUD. Our system focused on detecting early warning signs of HIV care disengagement and disruptive events that fall within the realm of social determinants

XSL•FO
RenderX

of health known to impact the health of people who use drugs [6]. Over a 1-year period, we captured >600 reports of events that could be disruptive to HIV care and be early warning signs of care disengagement. Using or injecting drugs and reporting missed doses of ART were associated with an elevated likelihood to miss a care visit or be virally nonsuppressed during follow-up, respectively. Not completing any of the weekly check-in surveys was also associated with missing a care visit. While our study lacked the statistical power needed to assess whether disruptive life events reported less frequently, including housing insecurity, experiencing violence, and being arrested, were associated with poor outcomes, our results support the need to conduct larger studies of time-varying threats to care engagement in the future. An mHealth system like ours could be used in HIV care settings to assist case managers and social workers in proactively detecting and addressing these threats.

Our sample of people who use drugs answered an average of 23 of 52 total weekly surveys available to them during the 1-year follow-up period, a response rate of 44%. Despite that participants were not remunerated for completing weekly surveys, as they were operationalized as a self-assessment tool to enhance self-determined motivation as part of the A-CHESS app, this response rate is similar to several studies among people living with HIV [28]. Nonetheless, the utility of the weekly survey as an informative tool for clinical action may require a higher response rate. Thus, future studies may consider strategies such as incentivizing weekly surveys or sending surveys at a time most likely to get a response [29].

At the same time, our findings suggest that missing up to 4 consecutive surveys may be a helpful threshold to prompt a case manager or social worker to proactively reach out to a patient about any difficulties they may be experiencing and ensure any needs are met through appropriate referrals or direct provision of support. This conclusion is supported by considering multiple of our findings together, including our modest response rate, our finding that not answering at least 1 survey over a 4-week period was associated with missing a care visit, and our finding that those who did not experience viral nonsuppression or missed visits tended to engage with the app more frequently. A focus on monthly, rather than weekly, engagement with A-CHESS may also help to reduce alert fatigue, a phenomenon wherein people using technology become desensitized after receiving many alerts or notifications [30-33] and balance the extra staff time required to access A-CHESS and follow up with disengaged participants. Indeed, another ongoing study of A-CHESS in the substance use treatment setting suggests that staff time is a major barrier to using a clinician dashboard feature added to A-CHESS [34], which provides an easy-to-interpret interface to monitor patient survey results and app engagement. Prior studies of A-CHESS also suggest that providing dedicated support to implement A-CHESS within a health care organization using strategies such as coaching and process improvement models to ensure clinicians have dedicated time, technical assistance, and resources to integrate this new clinical tool into existing workflows [35,36]. Our finding that using and injecting drugs were associated with subsequently missing care visits further suggests that understanding the context of drug use in weekly surveys may assist with identifying supports available to app users at the time when drug use or craving occurs [37].

The need for novel and effective mHealth interventions is underscored by findings from several demonstration projects that were part of the Health Resources and Services Administration's Special Projects of National Significance focused on social media and mHealth for supporting retention in HIV care [38]. This work identified 3 promising foci for mHealth tools, including (1) managing HIV care; (2) fostering feelings of support and personal connectedness; and (3) helping alleviate negative feelings about status and mitigating HIV-related stigma [39]. A-CHESS can meet all 3 of these needs. Through its theoretical foundations in self-determination theory [25], A-CHESS facilitates patient-provider connection and support through messaging functions and a clinician dashboard that can highlight weekly self-monitoring survey results. Peer connections are augmented by the use of a discussion board, targeted resources, and other key features.

Limitations

The modest sample size of the study limits its internal and external validity and our ability to adjust for a broad set of confounders. Thus, we presented both unadjusted and adjusted results that could account for key confounders we identified a priori, including sociodemographic characteristics, comfort level with using a smartphone, and recent drug use at baseline. The overall conclusions drawn from the results were consistent in adjusted and unadjusted models. Additionally, several of the disruptive life events of interest (ie, housing instability, food insecurity, and law enforcement interaction) were reported infrequently over the follow-up period, which did not allow us to examine their association with missing a care visit or viral nonsuppression. As these are known contributors to disengagement with HIV care, they will be important to study in future, larger studies. Our results suggest that, when experienced, our smartphone app platform was able to capture at least some of these events on a weekly basis. Future research should also use qualitative methods to assess user experience, acceptability, and desired supports when disruptive life events occur. Finally, we were unable to include canceled care visits, and it was impossible to determine if or when the visit was rescheduled. Further analysis of patients who cancel a visit and do not return is warranted.

Conclusions

A-CHESS is a promising tool for capturing time-varying threats to engagement in HIV care among people who use drugs. Future research should examine whether receiving tailored support based on disruptive life events reported in A-CHESS is effective in reducing viral nonsuppression and HIV care disengagement and evaluate how to integrate A-CHESS within the HIV medical home setting.

XSL•FO
RenderX

Acknowledgments

The authors acknowledge the National Institute on Drug Abuse (NIDA) Avenir Award DP2DA042424 and NIDA U01DA036297.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to the small sample size but are available from the study principal investigator, RW, on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplemental tables.
[DOCX File, 358 KB - formative_v9i1e59953_app1.docx]

Checklist 1
CONSORT (Consolidated Standards of Reporting Trials) checklist.
[DOCX File, 3600 KB - formative_v9i1e59953_app2.docx]

References

1. Ending the HIV epidemic HIV.gov 2022. US Department of Health and Human Services. 2022. URL: https://www.hiv.gov/federal-response/ending-the-hiv-epidemic/overview [accessed 2023-04-26]
2. Jones TPW, Lampe FC, Arenas-Pinto A, et al. Alcohol, smoking, recreational drug use and association with virological outcomes among people living with HIV: cross-sectional and longitudinal analyses. HIV Med 2022 Mar;23(3):209-226. [doi: 10.1111/hiv.13156] [Medline: 34634176]
3. Ladak F, Socias E, Nolan S, et al. Substance use patterns and HIV-1 RNA viral load rebound among HIV-positive illicit drug users in a Canadian setting. Antivir Ther 2019;24(1):19-25. [doi: 10.3851/IMP3265] [Medline: 30230474]
4. Langebeek N, Gisolf EH, Reiss P, et al. Predictors and correlates of adherence to combination antiretroviral therapy (ART) for chronic HIV infection: a meta-analysis. BMC Med 2014 Aug 21;12:142. [doi: 10.1186/PREACCEPT-1453408941291432] [Medline: 25145556]
5. Westergaard RP, Hess T, Astemborski J, Mehta SH, Kirk GD. Longitudinal changes in engagement in care and viral suppression for HIV-infected injection drug users. AIDS 2013 Oct 23;27(16):2559-2566. [doi: 10.1097/QAD.0b013e328363bff2] [Medline: 23770493]
6. Galea S, Vlahov D. Social determinants and the health of drug users: socioeconomic status, homelessness, and incarceration. Public Health Rep 2002;117(Suppl 1):S135-S145. [Medline: 12435837]
7. Kidder DP, Wolitski RJ, Campsmith ML, Nakamura GV. Health status, health care use, medication use, and medication adherence among homeless and housed people living with HIV/AIDS. Am J Public Health 2007 Dec;97(12):2238-2245. [doi: 10.2105/AJPH.2006.090209] [Medline: 17971562]
8. Marcus R, de Groot A, Bachman S, et al. Longitudinal determinants of housing stability among people living with HIV/AIDS experiencing homelessness. Am J Public Health 2018 Dec;108(S7):S552-S560. [doi: 10.2105/AJPH.2018.304772] [Medline: 32941778]
9. Strike C, Rudzinski K, Patterson J, Millson M. Frequent food insecurity among injection drug users: correlates and concerns. BMC Public Health 2012 Dec 8;12(1058):1058. [doi: 10.1186/1471-2458-12-1058] [Medline: 23216869]
10. Schmitz J, Kral AH, Chu D, Wenger LD, Bluthenthal RN. Food insecurity among people who inject drugs in Los Angeles and San Francisco. Public Health Nutr 2016 Aug;19(12):2204-2212. [doi: 10.1017/S1368980016000306] [Medline: 26956477]
11. Singer AW, Weiser SD, McCoy SI. Does food insecurity undermine adherence to antiretroviral therapy? A systematic review. AIDS Behav 2015 Aug;19(8):1510-1526. [doi: 10.1007/s10461-014-0873-1] [Medline: 25096896]
12. Anema A, Chan K, Chen Y, Weiser S, Montaner JSG, Hogg RS. Relationship between food insecurity and mortality among HIV-positive injection drug users receiving antiretroviral therapy in British Columbia, Canada. PLoS ONE 2013;8(5):e61277. [doi: 10.1371/journal.pone.0061277] [Medline: 23723968]
13. Rouhani S, Allen ST, Whaley S, et al. Food access among people who inject drugs in West Virginia. Harm Reduct J 2021 Aug 21;18(1):90. [doi: 10.1186/s12954-021-00536-x] [Medline: 34419045]
14. Shannon K, Kerr T, Milloy MJ, et al. Severe food insecurity is associated with elevated unprotected sex among HIV-seropositive injection drug users independent of HAART use. AIDS 2011 Oct 23;25(16):2037-2042. [doi: 10.1097/QAD.0b013e32834b35c9] [Medline: 21811140]
15. Wood E, Kerr T, Tyndall MW, Montaner JSG. A review of barriers and facilitators of HIV treatment among injection drug users. AIDS 2008 Jul 11;22(11):1247-1256. [doi: 10.1097/QAD.0b013e3282fbd1ed] [Medline: 18580603]

XSL•FO
RenderX

16. Westergaard RP, Kirk GD, Richesson DR, Galai N, Mehta SH. Incarceration predicts virologic failure for HIV-infected injection drug users receiving antiretroviral therapy. Clin Infect Dis 2011 Oct;53(7):725-731. [doi: 10.1093/cid/cir491] [Medline: 21890777]
17. Westergaard RP, Hochstatter KR, Andrews PN, et al. Effect of patient navigation on transitions of HIV care after release from prison: a retrospective cohort study. AIDS Behav 2019 Sep;23(9):2549-2557. [doi: 10.1007/s10461-019-02437-4] [Medline: 30790170]
18. mHealth: New Horizons for Health through Mobile Technologies: World Health Organization; 2011.
19. Beane SN, Culyba RJ, DeMayo M, Armstrong W. Exploring the medical home in Ryan White HIV care settings: a pilot study. J Assoc Nurses AIDS Care 2014;25(3):191-202. [doi: 10.1016/j.jana.2013.10.007] [Medline: 24560357]
20. Vlahov D, Anthony JC, Muñoz A, et al. The ALIVE Study: a longitudinal study of HIV-1 infection in intravenous drug users: description of methods. J Drug Issues 1991 Oct;21(4):759-776. [doi: 10.1177/002204269102100406]
21. Sun J, Mehta SH, Astemborski J, et al. Mortality among people who inject drugs: a prospective cohort followed over three decades in Baltimore, MD, USA. Addiction 2022 Mar;117(3):646-655. [doi: 10.1111/add.15659] [Medline: 34338374]
22. Gustafson DH, McTavish FM, Chih MY, et al. A smartphone application to support recovery from alcoholism: A randomized clinical trial. JAMA Psychiatry 2014 May;71(5):566-572. [doi: 10.1001/jamapsychiatry.2013.4642] [Medline: 24671165]
23. Gustafson DH Sr, Landucci G, McTavish F, et al. The effect of bundling medication-assisted treatment for opioid addiction with mHealth: study protocol for a randomized clinical trial. Trials 2016 Dec 12;17(1):592. [doi: 10.1186/s13063-016-1726-1] [Medline: 27955689]
24. Gustafson DH, Landucci G, Vjorn OJ, et al. Effects of bundling medication for opioid use disorder with an mHealth intervention targeting addiction: a randomized clinical trial. Am J Psychiatry 2024 Feb 1;181(2):115-124. [doi: 10.1176/appi.ajp.20230055] [Medline: 37789744]
25. Deci EL, Ryan RM. Intrinsic Motivation and Self-Determination in Human Behavior: Plenum Press; 1985.
26. HIV treatment as prevention? HIV.gov. 2023. URL: https://www.hiv.gov/tasp [accessed 2025-03-18]
27. Cohen MS, Chen YQ, McCauley M, et al. Prevention of HIV-1 infection with early antiretroviral therapy. N Engl J Med 2011 Aug 11;365(6):493-505. [doi: 10.1056/NEJMoa1105243] [Medline: 21767103]
28. Smiley SL, Milburn NG, Nyhan K, Taggart T. A systematic review of recent methodological approaches for using ecological momentary assessment to examine outcomes in U.S. based HIV research. Curr HIV/AIDS Rep 2020 Aug;17(4):333-342. [doi: 10.1007/s11904-020-00507-0] [Medline: 32594365]
29. Aminikhanghahi S, Schmitter-Edgecombe M, Cook DJ. Context-aware delivery of ecological momentary assessment. IEEE J Biomed Health Inform 2020 Apr;24(4):1206-1214. [doi: 10.1109/JBHI.2019.2937116] [Medline: 31443058]
30. Ash JS, Sittig DF, Campbell EM, Guappone KP, Dykstra RH. Some unintended consequences of clinical decision support systems. AMIA Annu Symp Proc 2007 Oct 11;2007(26–30):26-30. [Medline: 18693791]
31. Baseman JG, Revere D, Painter I, Toyoji M, Thiede H, Duchin J. Public health communications and alert fatigue. BMC Health Serv Res 2013 Aug 5;13(1):295. [doi: 10.1186/1472-6963-13-295] [Medline: 23915324]
32. Gephart S, Carrington JM, Finley B. A systematic review of nurses' experiences with unintended consequences when using the electronic health record. Nurs Adm Q 2015;39(4):345-356. [doi: 10.1097/NAQ.0000000000000119] [Medline: 26340247]
33. Muench F, Baumel A. More than a text message: dismantling digital triggers to curate behavior change in patient-centered health interventions. J Med Internet Res 2017 May 26;19(5):e147. [doi: 10.2196/jmir.7463] [Medline: 28550001]
34. Fleddermann K, Molfenter T, Jacobson N, et al. Clinician perspectives on barriers and facilitators to implementing e-health technology in substance use disorder (SUD) treatment facilities. Subst Abuse 2021;15:11782218211053360. [doi: 10.1177/11782218211053360] [Medline: 34720585]
35. Gustafson D Sr, Horst J, Boss D, et al. Implementation of smartphone systems to improve quality of life for people with substance use disorder: interim report on a randomized controlled trial. JMIR Hum Factors 2022 Jul 14;9(3):e35125. [doi: 10.2196/35125] [Medline: 35834315]
36. Quanbeck A, Gustafson DH, Marsch LA, et al. Implementing a mobile health system to integrate the treatment of addiction into primary care: a hybrid implementation-effectiveness study. J Med Internet Res 2018 Jan 30;20(1):e37. [doi: 10.2196/jmir.8928] [Medline: 29382624]
37. Linas BS, Latkin C, Westergaard RP, et al. Capturing illicit drug use where and when it happens: an ecological momentary assessment of the social, physical and activity environment of using versus craving illicit drugs. Addiction 2015 Feb;110(2):315-325. [doi: 10.1111/add.12768] [Medline: 25311241]
38. Medich M, Swendeman DT, Comulada WS, et al. Promising approaches for engaging youth and young adults living with HIV in HIV primary care using social media and mobile technology interventions: protocol for the SPNS social media initiative. JMIR Res Protoc 2019 Jan 31;8(1):e10681. [doi: 10.2196/10681] [Medline: 30702434]
39. Brooks RA, Nieto O, Swendeman D, et al. Qualitative evaluation of social media and mobile technology interventions designed to improve HIV health outcomes for youth and young adults living with HIV: a HRSA SPNS initiative. Health Promot Pract 2020 Sep;21(5):693-704. [doi: 10.1177/1524839920938704] [Medline: 32757839]

XSL•FO
RenderX

Abbreviations

A-CHESS: Addiction Comprehensive Health Enhancement Support System
ALIVE: AIDS Linked to the Intravenous Experience
aOR: adjusted odds ratio
ART: antiretroviral therapy
CONSORT: Consolidated Standards of Reporting Trials
mHealth: mobile health
OUD: opioid use disorder

Edited by A Mavragani; submitted 26.04.24; peer-reviewed by S Ramsey, T Finlayson; revised version received 28.02.25; accepted 02.03.25; published 26.03.25.

Please cite as:
Gicquelais RE, Conway C, Vjorn O, Genz A, Kirk G, Westergaard R
Mobile Health Tool to Capture Social Determinants of Health and Their Impact on HIV Treatment Outcomes Among People Who Use Drugs: Pilot Feasibility Study
JMIR Form Res 2025;9:e59953
URL: https://formative.jmir.org/2025/1/e59953
doi:10.2196/59953

© Rachel E Gicquelais, Caitlin Conway, Olivia Vjorn, Andrew Genz, Gregory Kirk, Ryan Westergaard. Originally published in JMIR Formative Research (https://formative.jmir.org), 26.3.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Mobile App-Based Interactive Care Plan for Migraine: Survey Study of Usability and Improvement Opportunities

Nathan P Young[1], DO; Jennifer I Stern[1], MD; Stephanie J Steel[1], MD; Jon O Ebbert[2], MD

[1]Department of Neurology, Mayo Clinic, 200 First St. SW Eighth Floor, Rochester, MN, United States

[2]Department of Internal Medicine, Mayo Clinic, Rochester, MN, United States

Corresponding Author:
Nathan P Young, DO
Department of Neurology, Mayo Clinic, 200 First St. SW Eighth Floor, Rochester, MN, United States

Abstract

Background: We implemented a novel mobile app-based Migraine Interactive Care Plan (MICP) integrated with our electronic health records (EHRs). The MICP facilitates remote assessment of adult patients with migraine, educational content delivery, and care team communication. Feasibility of the MICP was demonstrated in a pilot implementation study.

Objective: We aimed to assess the preferences and satisfaction of patients with migraine users of a mobile app-based care plan integrated with the EHR.

Methods: An electronic survey was administered to a single cohort of MICP users between December 6, 2021, and December 30, 2021. The survey assessed patient preferences for which data to track, frequency of tracking, and satisfaction with the MICP. Survey responses were compared between subsets determined by patient-reported headache frequency and treatment with and without botulinum toxin and calcitonin gene-related peptide (CGRP) antagonist therapy. The Wilcoxon rank-sum test was used for continuous variables and the χ^2 test or Fisher exact test for categorical variables.

Results: The total sample size was 184 and the survey response rate was 30.4% (56/184). No significant differences in age (P=.26) or sex (P=.19) between respondents and nonrespondents were observed. Respondent median age was 42 (range 20 - 72) years and 94.6% (53/56) were female. Headache frequency was (1) 0 to 8 days (26/56, 46.4%), (2) 9 to 14 days (12/56, 21.4%), and (3) 15 or more days (18/56, 32.1%). No difference was observed in any survey responses based on headache frequency or treatment. The majority of respondents preferred to track headache days weekly (30/56, 53.6%) or daily (15/56, 26.8%) and preferred to change the frequency of headache tracking reminders (42/56, 75%). Respondents were somewhat or very interested in daily tracking personal observations in free text (41/52, 78.8%), medication treatment (43/52, 82.7%) and treatment response (39/56, 69.6%), class of medication treatment (36/52, 69.2%), severity of functional impairment (39/56, 69.6%), type of functional impairment (35/53, 66%), headache day (40/54, 74.1%), and headache pain level on a scale of 1 to 10 (38/53, 71.7%). Respondents agreed or strongly agreed that the education content was useful (31/51, 60.8%) but lacked personalization (25/51, 49%). Most respondents agreed or strongly agreed that they were satisfied with the MICP (38/50, 76%) and that it helped them communicate with their care team (38/53, 71.7%).

Conclusions: Most MICP users were motivated to track headache frequency, medication treatment with response, functional impairment, and pain intensity. Opportunities to improve the MICP include (1) allowing patients to change the frequency of assessments and notifications; (2) recording personal observations or comments through free text, which may include headache triggers; (3) assessment of headache severity using a 1 to 10 pain scale; and (4) tailoring headache education based on frequency and severity (episodic vs chronic migraine). These observations may be useful to improve the usability of the MICP and similar EHR-integrated migraine care platforms that others may develop.

(JMIR Form Res 2025;9:e66763) doi:10.2196/66763

KEYWORDS

migraine; remote monitoring; mobile app; mHealth; patient-reported outcomes; care plan; digital health; app; smartphone; eHealth; technology; survey study; headache; electronic health record; remote assessment; older adult; adult; electronic survey; pain; mobile phone; telehealth; telemedicine

XSL•FO
RenderX

Introduction

Background

Patients with migraine often face significant barriers in access to evidence-based care, with approximately one-quarter of those with episodic migraine and less than 5% of those with chronic migraine receiving appropriate treatment [1-4]. Access to care is especially challenging for women, racial and ethnic minority groups, and patients living in rural areas [5]. Novel health care delivery models may improve care access when resources are limited and projected to worsen [2,6]. Studies have demonstrated that telemedicine improves access to migraine care, and that it is associated with favorable outcomes including patient and provider satisfaction [7,8]. Internet or smartphone app-based remote assessment and monitoring of migraine [9,10] may increase the efficiency of care delivery and facilitate telemedicine, electronic [11], and face-to-face visits while delivering migraine educational content. Headache clinicians are comfortable treating patients through telemedicine and most have indicated a high level of interest in prescribing migraine apps [12] while citing the importance of integrating remote monitoring data into the electronic health record (EHR) [13]. Smartphone-based apps may also improve diagnosis [14], deliver educational content [15], guide behavioral treatment [16-18], and advance clinical trials [19]. There are improvement opportunities with respect to accessibility of apps for headache [20].

Previous Work

We implemented a novel smartphone app-based Migraine Interactive Care Plan (MICP) integrated with the EHR (Epic, Epic Systems) in our community neurology practice, and potentially available to other institutions using Epic EHR. Development and testing of the MICP has been previously published [15]. Feasibility of the MICP was demonstrated in a pilot implementation study, but usability did not reach a predetermined threshold of 75% of users completing at least one electronically assigned task 127/171 (74.3%) [15].

Access to the MICP was ordered by a provider at the time of a face-to-face neurology consult and integrated into the existing EHR app (ie, Epic MyChart platform). This integration allowed patient access to their medical record, care team messaging, and appointment management. The MICP was designed to facilitate remote monitoring of adult patients with migraine, deliver educational content, and streamline care delivery. Key features of the MICP included (1) weekly assessments of headache days, treatment days, reduced function, and absenteeism; (2) monthly assessments of patient satisfaction (Likert scale) with their migraine treatment plan; (3) Migraine Disability Assessment (MIDAS) survey [21] assessment every 3 months; (4) scheduled delivery of migraine education content; and (5) weekly assessment of medication compliance and concerns and electronic messaging with the care team if needed.

Goal of This Study

In this study, we aimed to identify opportunities to improve MICP usability by surveying MICP users to assess their satisfaction and preferences and whether they differ based on migraine frequency or type of preventive treatment.

Methods

Ethical Considerations

This quality improvement project was granted an exemption waiver for written consent from the Mayo Clinic (IRB 20 - 000606). All respondents electronically signed the Health Insurance Portability and Accountability Act (HIPAA) agreement before starting the survey. No compensation was offered to complete the survey. All survey responses were deidentified before analysis, and the data were on a secure server.

Overview

A novel electronic survey was developed and formatted (on Qualtrics). Between December 6, 2021, and December 30, 2021, all MICP users (n=184) who completed MICP enrollment and completed at least one task received an email invitation to a secure web-based survey with a personalized letter from a physician on the migraine care team (NPY) explaining the goals of the survey and encouragement to voluntarily participate. The survey cohort had all completed a neurological headache consultation within a general community neurology practice before enrollment in the MICP. Weekly email reminders were sent by a study coordinator to nonrespondents for a total of 4 weeks. Only 1 survey completion was allowed. Demographic variables, including age and sex, were retrieved from the EHR for respondents and nonrespondents and were compared using the *t* test. Categorical variables were summarized. The overall survey response rate was defined by the percentage of surveys returned. The analysis of the results from each individual question included only the completed response to the question and nonresponses to individual questions were excluded. Survey responses were compared between subsets determined by patient reported headache frequency (0 to 8 days, 9 to 14 days, or 15 or more days), survey self-reported treatment with and without botulinum toxin and calcitonin gene-related peptide (CGRP) antagonist therapy. The Wilcoxon rank-sum test was used for continuous variables and χ^2 test or Fisher exact test for categorical variables using a 2-tailed α level of 5%. Missing survey response data were excluded from analysis. Analysis was performed using SAS software (version 9.4; SAS Inc).

Results

The survey response rate was 30.4% (56/184). No significant differences in age (P=.26) or sex (P=.19) between respondents and nonrespondents were observed. Table 1 summarizes the survey responses of all respondents, and a comprehensive table including the subgroup analysis and P values are summarized in Multimedia Appendix 1 and Multimedia Appendix 2.

XSL•FO
RenderX

Table . Survey responses of migraine users of the Migraine Interactive Care Plan.

Survey question	Total respondents (N=56), n (%)
How often would you prefer to track your headache days on a smartphone app?	
Daily	15 (26.8)
Weekly	30 (53.6)
Monthly	8 (14.3)
Every 3 months	3 (5.4)
How often would you like to be reminded to record a headache day on a smartphone app?	
Daily	14 (25)
Weekly	33 (58.9)
Monthly	7 (12.5)
Every 3 months	2 (3.6)
Would you prefer to control and change the frequency of headache tracking reminders?	
Yes	42 (75)
No	14 (25)
In addition to headache days, please tell us what else you are interested in tracking with your headache days?	
Migraine/headache triggers	41 (73.2)
Stress level	38 (67.8)
Sleep	37 (66.1)
Step count	9 (16.1)
Heart rate	4 (7.1)
Diet	16 (28.6)
Exercise	15 (26.8)
Diet calories	9 (16.1)
Please rate your level of interest and motivation to track, on a daily basis, the following factors in the Mayo Clinic Migraine Care Plan	
Headache days (yes/no)	
Not at all interested	3 (5.6)
Not very interested	3 (5.6)
Neutral	8 (14.8)
Somewhat interested	15 (27.8)
Very Interested	25 (46.3)
If yes, then pain level on scale 1 - 10	
Not at all interested	3 (5.7)
Not very interested	1 (1.9)
Neutral	11 (20.8)
Somewhat interested	17 (32.1)
Very Interested	21 (39.6)
Functional impairment (mild, moderate, or severe)	
Not at all interested	3 (5.7)
Not very interested	2 (3.8)
Neutral	9 (17)

XSL•FO
RenderX

Survey question	Total respondents (N=56), n (%)
Somewhat interested	17 (32.1)
Very Interested	22 (41.5)
What type of functions was impaired (work, school, family, or personal)	
Not at all interested	3 (5.7)
Not very interested	4 (7.5)
Neutral	11 (20.8)
Somewhat interested	16 (30.2)
Very Interested	19 (35.8)
Did you take medication (yes or no)	
Not at all interested	3 (5.8)
Not very interested	0 (0)
Neutral	6 (11.5)
Somewhat interested	14 (26.9)
Very Interested	29 (55.8)
If yes, then which type of medication (pick from multiple in list)	
Not at all interested	5 (9.6)
Not very interested	1 (1.9)
Neutral	10 (19.2)
Somewhat interested	19 (36.5)
Very interested	17 (32.7)
Response to medication	
Not at all interested	3 (5.8)
Not very interested	1 (1.9)
Neutral	9 (17.3)
Somewhat interested	15 (28.8)
Very interested	24 (46.2)
Your own personal observations or comments (free text)	
Not at all interested	3 (5.8)
Not very interested	1 (1.9)
Neutral	7 (13.5)
Somewhat interested	20 (38.5)
Very interested	21 (40.4)
Please rate your level of agreement with the following statements about the Mayo Clinic Migraine Care Plan.	
I felt confident using the Mayo Clinic Care Plan	
Strongly disagree	1 (1.9)
Disagree	2 (3.8)
Neither agree nor disagree	8 (15.1)
Agree	24 (45.3)
Strongly agree	18 (34)
The Mayo Clinic Care Plan app was easy to use	
Strongly disagree	2 (3.8)
Disagree	4 (7.5)

XSL•FO
RenderX

Survey question	Total respondents (N=56), n (%)
Neither agree nor disagree	11 (20.8)
Agree	21 (39.6)
Strongly agree	15 (28.3)
The equipment helped in my care at home	
Strongly disagree	1 (1.9)
Disagree	8 (15.1)
Neither agree nor disagree	23 (43.4)
Agree	11 (20.8)
Strongly agree	10 (18.9)
I felt comfortable interacting with my care team through the Mayo Clinic Care Plan	
Strongly disagree	0 (0)
Disagree	3 (5.8)
Neither agree nor disagree	7 (13.5)
Agree	24 (46.2)
Strongly agree	18 (34.6)
It helped me better understand my condition	
Strongly disagree	1 (1.9)
Disagree	7 (13.2)
Neither agree nor disagree	23 (43.4)
Agree	12 (22.6)
Strongly agree	10 (18.9)
It helped me understand how to care for myself	
Strongly disagree	1 (1.9)
Disagree	9 (17)
Neither agree nor disagree	24 (45.3)
Agree	9 (17)
Strongly agree	10 (18.9)
It helped me understand what I should be tracking throughout my care	
Strongly disagree	1 (1.9)
Disagree	5 (9.4)
Neither agree nor disagree	24 (45.3)
Agree	14 (26.4)
Strongly agree	9 (17)
It helped me understand what steps I could take to improve my health	
Strongly disagree	1 (1.9)
Disagree	8 (15.1)
Neither agree nor disagree	25 (47.2)
Agree	11 (20.8)
Strongly agree	8 (15.1)
It helped me communicate with my care team	
Strongly disagree	2 (3.8)
Disagree	5 (9.4)

XSL•FO
RenderX

Survey question	Total respondents (N=56), n (%)
Neither agree nor disagree	8 (15.1)
Agree	27 (50.9)
Strongly agree	11 (20.8)
It helped to inform me when to contact my care team about concerning symptoms	
Strongly disagree	2 (3.8)
Disagree	9 (17)
Neither agree nor disagree	15 (28.3)
Agree	17 (32.1)
Strongly agree	10 (18.9)
Please rate your level of agreement with the following statements about the education provided while using the Mayo Clinic Care Plan.	
The educational materials were useful to me	
Strongly disagree	2 (3.9)
Disagree	6 (11.8)
Neither agree nor disagree	12 (23.5)
Agree	22 (43.1)
Strongly agree	9 (17.6)
The information was easy to understand	
Strongly disagree	1 (2)
Disagree	0 (0)
Neither agree nor disagree	11 (21.6)
Agree	26 (51)
Strongly agree	13 (25.5)
I was comfortable with how often I received educational materials	
Strongly disagree	2 (3.9)
Disagree	3 (5.9)
Neither agree nor disagree	11 (21.6)
Agree	26 (51)
Strongly agree	9 (17.6)
I was able to find the educational materials when I needed them	
Strongly disagree	3 (5.9)
Disagree	2 (3.9)
Neither agree nor disagree	13 (25.5)
Agree	25 (49)
Strongly agree	8 (15.7)
The educational materials matched my personal needs	
Strongly disagree	1 (2)
Disagree	5 (9.8)
Neither agree nor disagree	20 (39.2)
Agree	19 (37.3)
Strongly agree	6 (11.8)
The educational information from the Mayo Clinic Care Plan matched the information received from my Mayo Clinic Care Team	

XSL•FO
RenderX

Survey question	Total respondents (N=56), n (%)
Strongly disagree	1 (2)
Disagree	0 (0)
Neither agree nor disagree	18 (35.3)
Agree	21 (41.2)
Strongly agree	11 (21.6)
Please rate your level of agreement with the following statements.	
I would recommend the Mayo Clinic Care Plan to others with similar health condition(s)	
Strongly disagree	1 (2)
Disagree	2 (4)
Neither agree nor disagree	8 (16)
Agree	24 (48)
Strongly agree	15 (30)
Overall, I am satisfied with the Mayo Clinic Care Plan	
Strongly disagree	1 (2)
Disagree	4 (8)
Neither agree nor disagree	7 (14)
Agree	25 (50)
Strongly agree	13 (26)

The median age of respondents was 42 (range 20 - 72) years and 94.6% (53/56) were female. Respondents reported treatment with an injectable CGRP antagonist (17/56, 30.4%) or botulinum toxin (22/56, 39.3%). Patient-reported headache frequency at the time of the survey was (1) 0 to 8 days (26/56, 46.4%), (2) 9 to 14 days (12/56, 21.4%), and (3) 15 or more days (18/56, 32.1%). No differences in survey responses were observed comparing patients reporting CGRP antagonist treatment versus no CGRP treatment, botulinum toxin treatment versus no botulinum toxin treatment, or between headache frequency groups at the time of the survey.

In terms of frequency, respondents had the highest preference for tracking headache days weekly (30/56, 53.6%) followed by daily (15/56, 26.8%) and monthly (8/56, 14.3%), with a tracking frequency of every 3 months being the least preferred (3/56, 5.4%). Respondents indicated a similar preference for how frequently they would like to be reminded to record a headache day. Most respondents preferred to have control over the frequency of headache tracking reminders (42/56, 75%). In addition to tracking headache days, most respondents indicated a preference for tracking migraine and headache triggers (42/56, 73.2%), stress level (38/56, 67.8%), and sleep (37/56, 66.1%). Fewer participants preferred tracking diet (16/56, 28.6%), exercise (15/56, 26.8%), caloric intake (9/56, 16.1%), step count (9/56, 16.1%), and heart rate (4/56, 7.1%).

When asked about the level of interest and motivation to track on a daily basis, respondents were somewhat or very interested in tracking personal observations or comments in free text (41/52, 78.8%), medication treatment (43/52, 82.7%), response to medication (39/56, 69.6%), class of medication treatment chosen from a list (36/52, 69.2%), functional impairment graded mild, moderate, or severe (39/56, 69.6%), type of function impaired including work, school, family, personal (35/53, 66%), headache day (40/54, 74.1%), and headache pain level on scale 1 - 10 (38/53, 71.7%).

The majority of respondents agreed or strongly agreed with statements about the Migraine Care Plan that they “felt confident using” (42/53, 79.2%), “was easy to use” (36/53, 68%), “felt comfortable interacting with my care team” (42/52, 80.8%), “helped to the inform me when to contact my care team about concerning symptoms” (27/53, 50.9%), and “helped me communicate with my care team” (38/53, 71.7%). A minority of respondents agreed or strongly agreed with statements that “the equipment helped in my care at home” (21/53, 40%), “helped me better understand my condition” (22/53, 41.5%), “understand how to care for myself” (19/53, 35.8%), “understand what I should be tracking throughout my care” (23/53, 43.4%), and “understand steps I can take to improve my health” (19/53, 35.8%).

Respondents agreed or strongly agreed that the education materials were useful (31/51, 60.8%) and easy to understand (39/51, 76.5%), but fewer agreed that the education content “matched my personal needs” (25/51, 49%). Most respondents agreed or strongly agreed that they were satisfied with the Migraine Care Plan (38/50, 76%).

XSL•FO
RenderX

Discussion

Principal Findings

We assessed the satisfaction and preferences of MICP and did not observe differences based on migraine frequency or type of preventive treatment. However, analysis of all respondents highlight that the most notable MICP improvement opportunities include allowing patients to enter free text observational data and adjust the frequency of monitoring.

In addition, we observed that most respondents prefer to track headache outcomes that are of interest to clinicians engaged in migraine management, including headache days, treatment days and response, and functional impairment [22]. Most respondents in this study preferred weekly tracking as well as the option to change the frequency of tracking reminders along with the ability to enter free text data and observations and rate the severity of pain on a 1 to 10 pain scale, all of which were not included in the MICP. Most respondents were not interested in monitoring non-headache data such as diet, exercise, caloric intake, or heart rate. Most respondents agreed that the educational content delivered was useful but that the content lacked personalization. Most respondents were satisfied with the MICP and felt that it helped them communicate with their care team, supporting ongoing use in our practice.

Key MICP Improvement Opportunities and Comparison With Previous Work

Free Text Capability

We observed that 78.8% (41/52) of respondents were somewhat or very interested in recording their own free text observations. This capability was not included in the MICP because of concern that free text data might be difficult for clinicians to interpret or might not be clinically actionable or easy to summarize in the EHR. Minen et al [23] also reported a similar preference of a patient with migraine for recording free text data. Although such data may be difficult for clinicians to efficiently review within the EHR and contain clinical irrelevant observations [22], it may be important for patient engagement and improve usability. As we enter the era of artificial intelligence [24], such free text data [25] may be a powerful data set for artificial intelligence natural language processing tools that may be able to quickly summary such data or use in algorithm development to facilitate patient assessment and management [26-29].

Frequency of Monitoring

Only 15/56 (26.8%) respondents in our study indicated a preference for daily monitoring, and only 5/171 (2.9%) users in the pilot study of the MICP responded to nearly all assigned assessment tasks [15]. Raffaeli et al described a cohort of highly engaged users of a commercial app that engaged with the app daily for more than 7 - 13 months. This cohort consisted of only 1.8% of 85,000 active app users who used the app daily for 7 months [10]. These observations suggest that sustained engagement on a daily or weekly basis may not be a realistic goal for most patients. Whether giving patients the option to change the frequency of assessments would improve overall engagement is not known.

Our MICP allowed weekly tracking of the number of headache days, treatment days, and days with functional impairment. However, most of our patients indicated an interest in changing this tracking frequency (42/56, 75%). The range of different preferences in tracking frequency supports allowing patients to choose the frequency of monitoring even though the accuracy of the data may be less precise the longer a patient may be asked to recall. It is also possible that the preference for tracking frequency might depend on migraine frequency, for example, patients with less frequent episodic migraine may prefer less frequent monitoring than a patient with uncontrolled chronic migraine. We suspect that as patients are effectively treated with appropriate acute and preventative migraine therapies, headache frequency and disability decrease such that daily and even weekly monitoring would no longer be necessary. A study in a larger population of patients is needed to determine if flexibility would increase overall engagement.

Patient preferences for tracking frequency may not always align with clinician preference or insurance requirements. In addition, different tracking frequencies serve different purposes. For example, the diagnosis of episodic versus chronic migraine depends on 3 months of data, whereas a shorter period may be ideal for making changes in medications or other aspects of care.

The survey suggests that patients would prefer to control their tracking and tracking reminder frequency. Individualization of tracking and reminder frequency may enhance patient engagement and understanding, especially for those with a new diagnosis of migraine. Requiring monthly assessments of headache days, treatment days, function, and satisfaction will allow the clinician access to data that will inform best migraine care. For example, monthly tracking may quickly uncover medication overuse behavior. It may also allow the clinician early detection of the transition between episodic and chronic migraine, which in turn may allow new treatment options such as onabotulinumtoxin A. The MIDAS assessment every 3 months may not be needed except for research purposes to allow comparison with other studies or migraine care model intervention. Monthly data trends would be simple to summarize and easy to interpret for busy clinicians.

Educational Materials

Respondents agreed or strongly agreed that the education materials were useful (31/51, 60.8%) but the education did not always match personal needs (25/51, 49%). The migraine educational content was generally applicable and may be useful to a patient with a new migraine diagnosis. Most MICP users in this study had frequent migraine that led to neurology consultation, and as most were treated with botox or CGRP, it can be assumed majority failed to respond to multiple first-line preventive options. Education specific to chronic migraine was minimal. The MICP may be improved by delivering educational content that is tailored to the migraine diagnosis or current treatment. Delivering content that is already known or simply not of interest may disengage patients.

XSL•FO
RenderX

Strengths

We present survey findings from a unique group of patients that used a novel digital tool directly interfacing with the EHR within a community neurology practice. We assessed preferences for a broad range of features that may both inform migraine care and promote patient engagement. We compared respondents with nonrespondents and between subsets of patients. The low response rate was similar to other survey studies of the same population [6,30,31].

Limitations

Limitations include the small sample size, and the overall survey response rate was low, reducing the reproducibility and generalizability of our findings. In addition, the study was underpowered to detect differences in user preferences based on headache frequency or treatment type. Most patients in the survey likely had chronic migraine or frequent episodic migraine as is typical of our specialty practice. The majority of patients were treated with botulinum toxin injections or CGRP antagonist therapy at the time of the study, and the majority of patients reported less than 15 headache days per month, likely reflecting the effectiveness of their current treatment plan. We did not retrospectively review records to confirm the migraine diagnosis documented in the medication record. In addition, most of the population we assessed were white and female, with high migraine frequency and access to a neurology specialty clinic and for these reasons our observations are not generalizable to other patient populations or health care systems. Finally, we did not reassess the frequency or level of engagement of MICP users who completed the survey, increasing potential response bias.

Conclusions

Most MICP users were motivated to track clinically actionable data, including headache frequency, medication treatment with response, functional impairment, and pain intensity; features that should be maintained in the MICP and similar tools that others may develop. Opportunities to improve the MICP that may be useful for other similar remote monitoring tools include allowing patients to (1) change the frequency of assessments and notifications; (2) record personal observations or comments through free text, which may include headache triggers, accompanying symptoms, and prodrome and postdrome symptoms; (3) report headache severity using a 1 to 10 pain scale; and (4) tailor headache education based on frequency and severity (episodic vs chronic migraine). To balance the needs for patient engagement with the changes above, the MICP may be simplified by requiring a monthly assessment of headache days, treatment days, function, and satisfaction with treatment plan without MIDAS every 3 months. Remote assessment and monitoring of patients with migraine is feasible, and continued work to refine remote monitoring tools informed by patient preferences and reassessment of user preferences within diverse and larger populations, including both episodic and chronic patients with migraine, is needed.

Acknowledgments

We acknowledge Jay Mandrekar, PhD for his assistance with summarizing the survey data and statistical support. The study and manuscript were not created with the use of generative artificial intelligence. This study was funded by a grant from the Noaber Foundation and the Mayo Clinic Center for Digital Health.

Conflicts of Interest

Mayo Clinic and Epic are partners in the development and commercialization of Care Plans. No individuals involved in this study have a direct financial interest in the Migraine Interactive Care Plan. JOE has consulting agreements with K Health, Exact Sciences, Alnylam Pharmaceuticals, and MedinCell; and serves on the Scientific Advisory Board for Applied Aerosol Technologies. None declared by the other authors.

Multimedia Appendix 1
Survey responses in subsets based on headache frequency.
[DOCX File, 51 KB - formative_v9i1e66763_app1.docx]

Multimedia Appendix 2
Survey responses in subsets based on migraine treatment type.
[DOCX File, 40 KB - formative_v9i1e66763_app2.docx]

References

1. Ashina M, Katsarava Z, Do TP, et al. Migraine: epidemiology and systems of care. Lancet 2021 Apr 17;397(10283):1485-1495. [doi: 10.1016/S0140-6736(20)32160-7] [Medline: 33773613]
2. Raffaelli B, Rubio-Beltrán E, Cho SJ, et al. Health equity, care access and quality in headache - part 2. J Headache Pain 2023 Dec 13;24(1):167. [doi: 10.1186/s10194-023-01699-7] [Medline: 38087219]
3. Lipton RB, Silberstein SD. Episodic and chronic migraine headache: breaking down barriers to optimal treatment and prevention. Headache 2015 Mar;55 Suppl 2(S2):103-122. [doi: 10.1111/head.12505_2] [Medline: 25662743]

XSL•FO
RenderX

4. Dodick DW, Loder EW, Manack Adams A, et al. Assessing barriers to chronic migraine consultation, diagnosis, and treatment: results from the chronic migraine epidemiology and outcomes (CaMEO) study. Headache 2016 May;56(5):821-834. [doi: 10.1111/head.12774] [Medline: 27143127]
5. Buse DC, Armand CE, Charleston L 4th, et al. Barriers to care in episodic and chronic migraine: Results from the Chronic Migraine Epidemiology and Outcomes Study. Headache 2021 Apr;61(4):628-641. [doi: 10.1111/head.14103] [Medline: 33797078]
6. Young NP, Philpot LM, Vierkant RA, et al. Episodic and chronic migraine in primary care. Headache 2019 Jul;59(7):1042-1051. [doi: 10.1111/head.13543] [Medline: 31032913]
7. Robblee J. Telemedicine in headache medicine: a narrative review. Curr Pain Headache Rep 2023 Sep;27(9):371-377. [doi: 10.1007/s11916-023-01141-2] [Medline: 37515743]
8. Berk T, Silberstein S, McAllister P. A novel virtual-based comprehensive clinical approach to headache care. J Clin Med 2023 Aug 17;12(16):37629391. [doi: 10.3390/jcm12165349] [Medline: 37629391]
9. Mosadeghi-Nik M, Askari MS, Fatehi F. Mobile health (mHealth) for headache disorders: A review of the evidence base. J Telemed Telecare 2016 Dec;22(8):472-477. [doi: 10.1177/1357633X16673275] [Medline: 27799450]
10. Raffaelli B, Mecklenburg J, Overeem LH, et al. Determining the evolution of headache among regular users of a daily electronic diary via a smartphone app: observational study. JMIR Mhealth Uhealth 2021 Jul 7;9(7):e26401. [doi: 10.2196/26401] [Medline: 34255716]
11. Robblee J, Starling AJ. E-consultation in headache medicine: a quality improvement pilot study. Headache 2020 Nov;60(10):2192-2201. [doi: 10.1111/head.13981] [Medline: 33063855]
12. Minen MT, Szperka CL, Kaplan K, et al. Telehealth as a new care delivery model: The headache provider experience. Headache 2021 Jul;61(7):1123-1131. [doi: 10.1111/head.14150] [Medline: 34309828]
13. Minen MT, George A, Katara A, et al. Headache clinicians' perspectives on the remote monitoring of patients' electronic diary data: A qualitative study. Headache 2023;63(7):917-925. [doi: 10.1111/head.14519] [Medline: 37313636]
14. Roesch A, Dahlem MA, Neeb L, Kurth T. Validation of an algorithm for automated classification of migraine and tension-type headache attacks in an electronic headache diary. J Headache Pain 2020 Jun 12;21(1):75. [doi: 10.1186/s10194-020-01139-w] [Medline: 32532222]
15. Young NP, Ridgeway JL, Haddad TC, et al. Feasibility and usability of a mobile app-based interactive care plan for migraine in a community neurology practice: development and pilot implementation study. JMIR Form Res 2023 Oct 5;7:e48372. [doi: 10.2196/48372] [Medline: 37796560]
16. Minen MT, Adhikari S, Padikkala J, et al. Smartphone-delivered progressive muscle relaxation for the treatment of migraine in primary care: a randomized controlled trial. Headache 2020 Nov;60(10):2232-2246. [doi: 10.1111/head.14010] [Medline: 33200413]
17. Minen MT, Friedman BW, Adhikari S, et al. Introduction of a smartphone based behavioral intervention for migraine in the emergency department. Gen Hosp Psychiatry 2021;69(12-9):12-19. [doi: 10.1016/j.genhosppsych.2020.12.009] [Medline: 33485090]
18. Spreadbury JH, Young A, Kipps CM. A comprehensive literature search of digital health technology use in neurological conditions: review of digital tools to promote self-management and support. J Med Internet Res 2022 Jul 28;24(7):e31929. [doi: 10.2196/31929] [Medline: 35900822]
19. Santoro A, Delussi M, Leone M, et al. Effects of botulinum toxin on migraine attack features in chronic migraine: a six-month open-label observation study through electronic diary smartphone application. Toxins (Basel) 2019 Nov 15;11(11):668. [doi: 10.3390/toxins11110668] [Medline: 31731628]
20. Minen MT, George A, Camacho E, et al. Assessment of smartphone apps for common neurologic conditions (headache, insomnia, and pain): cross-sectional study. JMIR Mhealth Uhealth 2022 Jun 21;10(6):e36761. [doi: 10.2196/36761] [Medline: 35727625]
21. Lipton RB, Stewart WF, Sawyer J, Edmeads JG. Clinical utility of an instrument assessing migraine disability: the Migraine Disability Assessment (MIDAS) questionnaire. Headache 2001 Oct;41(9):854-861. [doi: 10.1111/j.1526-4610.2001.01156.x] [Medline: 11703471]
22. Dodick DW, Tepper SJ, Lipton RB, et al. Improving medical communication in migraine management: a modified delphi study to develop a digital migraine tracker. Headache 2018 Oct;58(9):1358-1372. [doi: 10.1111/head.13426] [Medline: 30362524]
23. Minen MT, Jaran J, Boyers T, Corner S. Understanding what people with migraine consider to be important features of migraine tracking: an analysis of the utilization of smartphone-based migraine tracking with a free-text feature. Headache 2020 Jul;60(7):1402-1414. [doi: 10.1111/head.13851] [Medline: 33300599]
24. Gutman B, Shmilovitch AH, Aran D, Shelly S. Twenty-five years of AI in neurology: the journey of predictive medicine and biological breakthroughs. JMIR Neurotech 2024;3:e59556. [doi: 10.2196/59556]
25. Lefkovitz I, Walsh S, Blank LJ, Jetté N, Kummer BR. Direct clinical applications of natural language processing in common neurological disorders: scoping review. JMIR Neurotech 2024;3:e51822. [doi: 10.2196/51822]

XSL•FO
RenderX

26. Chiang CC, Luo M, Dumkrieger G, et al. A large language model-based generative natural language processing framework finetuned on clinical notes accurately extracts headache frequency from electronic health records. medRxiv. 2023 Oct 3. [doi: 10.1101/2023.10.02.23296403] [Medline: 37873417]
27. Kreimeyer K, Foster M, Pandey A, et al. Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. J Biomed Inform 2017 Sep;73(14-29):14-29. [doi: 10.1016/j.jbi.2017.07.012] [Medline: 28729030]
28. Jensen K, Soguero-Ruiz C, Oyvind Mikalsen K, et al. Analysis of free text in electronic health records for identification of cancer patient trajectories. Sci Rep 2017 Apr 7;7(46226):46226. [doi: 10.1038/srep46226] [Medline: 28387314]
29. Messinger AI, Luo G, Deterding RR. The doctor will see you now: How machine learning and artificial intelligence can extend our understanding and treatment of asthma. J Allergy Clin Immunol 2020 Feb;145(2):476-478. [doi: 10.1016/j.jaci.2019.12.898] [Medline: 31883444]
30. Stern JI, Ebbert JO, Vierkant RA, Young NP. Migraine patient care model preferences in a community practice. Neurol Clin Pract 2023 Apr;13(2):e200131. [doi: 10.1212/CPJ.0000000000200131] [Medline: 37064592]
31. Jonker L, Fitzgerald L, Vanderpol J, Fisher S. Digital diary app use for migraine in primary care: prospective cohort study. Clin Neurol Neurosurg 2022 May;216(107225):107225. [doi: 10.1016/j.clineuro.2022.107225] [Medline: 35364371]

Abbreviations

CGRP: calcitonin gene-related peptide
EHR: electronic health record
MICP: Migraine Interactive Care Plan
MIDAS: Migraine Disability Assessment

Edited by A Mavragani; submitted 08.10.24; peer-reviewed by G Peral, M Straburzynski; revised version received 02.03.25; accepted 04.03.25; published 26.03.25.

Please cite as:
Young NP, Stern JI, Steel SJ, Ebbert JO
Mobile App-Based Interactive Care Plan for Migraine: Survey Study of Usability and Improvement Opportunities
JMIR Form Res 2025;9:e66763
URL: https://formative.jmir.org/2025/1/e66763
doi:10.2196/66763

© Nathan P Young, Jennifer I Stern, Stephanie J Steel, Jon O Ebbert. Originally published in JMIR Formative Research (https://formative.jmir.org), 26.3.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Changes in Health Education Literacy After Structured Web-Based Education Versus Self-Directed Online Information Seeking in Patients Undergoing Carpal Tunnel Release Surgery: Nonrandomized, Controlled Study

Mariella Seel[1], BA, MSc, Dr. sc. med.; Julian Alexander Mihalic[1,2], MD; Stefan Mathias Froschauer[1,2], MD, PhD; Bernhard Holzner[3], MD, PhD; Jens Meier[1,4], MD, PhD; Tobias Gotterbarm[1,2], MD, PhD; Matthias Holzbauer[1,2], MD, LLM

[1]Faculty of Medicine, Johannes Kepler University of Linz, Linz, Austria

[2]Department for Orthopedics and Traumatology, Kepler University Hospital GmbH, Krankenhausstrasse 9, Linz, Austria

[3]Department of Psychiatry, Psychotherapy, Psychosomatics and Medical Psychology, University Hospital of Psychiatry II, Innsbruck Medical University, Innsbruck, Austria

[4]Department of Anesthesiology and Intensive Care Medicine, Kepler University Hospital GmbH, Johannes Kepler University of Linz, Linz, Austria

Corresponding Author:
Matthias Holzbauer, MD, LLM
Faculty of Medicine, Johannes Kepler University of Linz, Linz, Austria

Abstract

Background: With advancements in anesthesiologic and surgical techniques, many surgeries are now performed as day-surgery procedures, requiring greater responsibilities for self-management from patients during the perioperative process. Online health information often lacks reliability and comprehensibility, posing risks for patients with low health literacy. Carpal tunnel release (CTR) surgery, a common day-surgery procedure, necessitates effective patient education for optimal recovery and self-management.

Objective: This study introduces the CTS Academy, a web-based education program designed for patients undergoing CTR day surgery. The study aimed to evaluate the CTS Academy's impact on patients' health education literacy (HEL) compared with self-directed online information seeking.

Methods: A scoping review on education programs focusing on the perioperative process of CTR was conducted before this study. In a nonrandomized controlled study, 60 patients scheduled for CTR were assigned to 2 groups based on the patients' preferences; the test group used the CTS Academy, while the control group performed self-directed online searches. HEL was assessed using the Health Education Literacy of Patients with chronic musculoskeletal diseases (HELP) questionnaire, focusing on patients's comprehension of medical information (COMPR), patients's ability to apply health-related information in an everyday life (APPLY), and patient's ability to communicate with health care professional (COMM). Secondary outcomes included content comprehensibility, patient preferences, platform usability, and clinical carpal tunnel syndrome (CTS)–related parameters.

Results: In the scoping review, 17 studies could be identified and included for full-text analysis. Eighteen patients each were included in the test group (13 women and 5 men) and in the control group (11 women and 7 men). The average time spent in the study was 167 and 176 days for the test and control groups, respectively. The test group showed significant improvements in APPLY (mean 28, SD 7.99 vs mean 24, SD 5.14; $P<.05$) and COMM (mean 30, SD 10.52 vs mean 25, SD 6.01; $P=.02$) after using the CTS Academy in a longitudinal analysis. No significant changes were observed in the control group. In a comparison between groups, the test group had significantly higher APPLY scores at follow-up (mean 24, SD 5.14 vs mean 33, SD 14.78; $P=.044$) and fewer comprehension issues at baseline (mean 38, SD 16.60 vs mean 50, SD 19.00; $P=.03$). The CTS-related knowledge assessment yielded 92% (66/72) versus 90% (65/72) correct answers in the test and control groups, respectively. The test group rated the CTS Academy highly in usability (6.22 of 7.00 points) and utility (6.13 of 7.00 points). Preferences leaned toward using CTS Academy alongside doctor consultations (16/18, 89%) and over self-directed searches (15/18, 84%). No significant differences were found in CTS-related symptoms between groups.

Conclusions: The CTS Academy effectively enhanced patients' HEL, especially in applying and communicating medical information. The platform's usability and utility were rated favorably, and patients preferred it over independent online information seeking. This suggests that structured, web-based education enhances patient self-management during the day surgery process.

(JMIR Form Res 2025;9:e65114) doi:10.2196/65114

XSL•FO
RenderX

KEYWORDS

carpal tunnel release surgery; patient education; structured web-based education; online information; health education literacy; web-based; health education; information seeking; carpal tunnel release; carpal tunnel surgery; non-randomized; controlled study; self-management; perioperative; online health information; health literacy; day surgery; online search; carpal tunnel; carpal

Introduction

Background

Due to continuous advancements in surgical and anesthesiologic techniques, an increasing number of surgical interventions are being performed in a day surgery setting. Consequently, patients spend less time in the hospital and are often overwhelmed by a large amount of information regarding their condition and the perioperative process provided during face-to-face interactions with medical professionals [1]. Thus, patients are required to assume greater responsibilities for self-management and self-care at home [1]. However, many patients lack the necessary knowledge and ability to effectively manage the day surgery situation, leading to an increased demand for comprehensive patient education [2]. Patients seeking online health information frequently encounter the risk of unreliable sources and incorrect information due to insufficient quality control or issues with comprehensibility [3]. Considering the low health literacy levels among some patients, this poses a substantial threat to effective health communication and patient safety [4].

Despite the low-risk profile of many ambulatory interventions, the recovery phase poses challenges, emphasizing the need for patient self-management and comprehensive education programs addressing both physical and psychological aspects [1,5] to empower patients. Patient empowerment can be defined as "a process through which people gain greater control over decisions and actions affecting their health" [6]. It can be achieved through patient education, a "process of assisting consumers of health care to learn how to incorporate health related behaviors (knowledge, skill, attitude) into everyday life with the purpose of achieving the goal of optimal health" [7]. This approach should provide patients with the health-related knowledge and skills to manage their medical conditions effectively [8]. Closely related is the concept of health education literacy (HEL), a patient's "cognitive and social skills […] to understand and use the information as conveyed in health education programmes" [9]. HEL is based on the concept of health literacy and consists of three dimensions, that are (1) the patient's ability to apply health-related information in an everyday life (APPLY), (2) the patient's ability to communicate with health care professionals (COMM), and (3) the patient's comprehension of medical information (COMPR) [9].

Goal of This Study

Open carpal tunnel release (CTR) surgery is a common day-surgery procedure for carpal tunnel syndrome (CTS) [10]. CTS affects approximately 3% of the general adult population [11] and is caused by compression of the median nerve and results in symptoms such as weakness, tingling, and pain [12]. In a recent meta-analysis, CTR as gold standard treatment results in reliable improvements in hand function and symptom severity on the long-term follow-up, while manual therapy only provides short term effects regarding pain relief [13]. To address time constraints and information overload experienced by patients during clinical visits and to enhance their comprehension of their medical condition and treatment, we aim to introduce the CTS Academy, a patient-centered web-based education program, in this study. Since the implementation of patient-centered web-based education programs focusing on the perioperative process of CTR is rare, a scoping review was conducted before this study to highlight recent gaps in existing research. Our program is tailored to the specific needs of patients undergoing day surgery for CTR and aims to provide valid and comprehensible information in a familiar environment, thereby facilitating patient self-management.

In a nonrandomized controlled study, we aimed to evaluate the CTS Academy for its effect on patients' HEL compared with self-directed online information seeking, 3 weeks after CTR (in comparison with the baseline survey).

We hypothesize that HEL is increased by participating in a web-based, structured educational programme (CTS Academy) for patients undergoing CTR. Furthermore, we examine the comprehensibility of educational content, patients' preferences in education, the usability of the CTS Academy platform, and clinical changes in CTS-related parameters. Our research addresses the scarcity of comparable implementations and aims to acquire further insights into the existing limited quantitative research in this domain.

Methods

Scoping Review of the Literature

Before the conduction of this study, we performed a scoping review of literature discussing the quality of online CTS information. From September to December 2022, we searched Epistemonikos, the Cochrane Library, PubMed, Connected Papers, and Google Scholar. Search prompts were for example "perioperative education," "web-based," "carpal tunnel release," and "structured patient education," and the complete search strategy can be requested from the authors. The search aimed to identify systematic reviews, original research, and related studies on web-based patient education and online information for patients undergoing open CTR surgery. No filters were applied for publication date or status, but only English and German full texts were analyzed. Publications with low evidence levels or lacking full study descriptions were excluded, while a broad search strategy ensured relevant materials not explicitly titled as web-based education were included. Reporting was performed based on the Joanna Briggs Institute Scoping Review Reporting Template [14] and the PRISMA-ScR (Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews) Checklist [15].

XSL•FO
RenderX

Design of the CTS Academy

We implemented the CTS Academy platform prototype using Microsoft Sway [16], a freely accessible web-based software for creating single-page websites.

The target group for this program consists of patients scheduled for open CTR day surgery who seek relevant information related to physical and mental preparation for CTR surgery.

The goal of CTS Academy is to assist patients preparing for surgery and managing the postoperative phase by providing textual information, multimedia content, self-check quizzes, downloadable or printable checklists for patients' self-directed online learning.

In the CTS Academy, we primarily focus on biophysiological, functional, experiential, and social aspects, excluding financial information due to the Austrian public health insurance system. The ethical and cognitive perspectives, emphasizing patient recognition, appreciation, and knowledge form the foundation of the CTS Academy. Patient education programs that prioritize patient-centeredness and engagement rely on evidence-based strategies, including plain language and multimodal information delivery [17]. Keulers et al [18] propose practical implications for interactive patient education, advocating for varied information presentation through text, graphics, and videos, with structured text using headings and subheadings. Summaries of lengthy texts determine the relevance of content, and the inclusion of graphs, pictures, and patient role models, such as animated characters, aids comprehension [18]. Incorporating entertaining elements, such as storytelling, enhances patients' self-efficacy, with a preference for visual materials like videos and images over plain text [19,20]. Modular approaches, including basic modules for essential knowledge and additional modules for in-depth information on living with a specific disease, are also recommended [20].

The CTS Academy is based on the information provided for informed surgery consent [21], but adds further structured and comprehensive content for various stages of day surgery. Furthermore, we included aspects of mental preparation for surgery [22-26]and instructions for the postoperative period, as well as training exercises [27]. Patients can use a printable checklist for perioperative preparation and access external links, for example, the Austrian pharmacy search to locate nearby pharmacies after hospital discharge. An animated character named Karla serves as a patient role model, guiding patients through the day-surgery process and sharing important advice. To improve accessibility, the spoken text is also displayed as written text throughout the entire video.

While the platform lacks structured patient assessments, it encourages self-assessment based on chapter summaries and offers knowledge checks. In Multimedia Appendix 1, we outline the content and media types presented on the CTS Academy platform. In Figure 1, screenshots of the CTS Academy platform are presented.

XSL•FO
RenderX

Figure 1. Screenshots of CTS academy including textual explanations of its content. It aims to support patients in preparing for surgery and managing postoperative recovery through self-directed online learning by offering textual information, multimedia content, self-check quizzes, and downloadable or printable checklists.

XSL•FO
RenderX

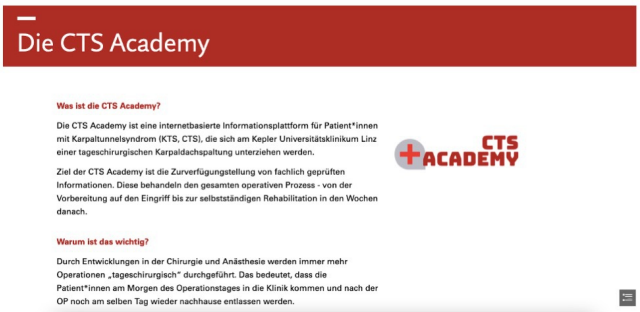

- Explanation of the project, the platform, project goals, relevance for the patients etc.
  - Why should I – as a patient – use this?

- Short summary at the beginning of a new chapter
  - Most important information as overview
  - Quick decision for continue reading or not
- Information on preoperative situation

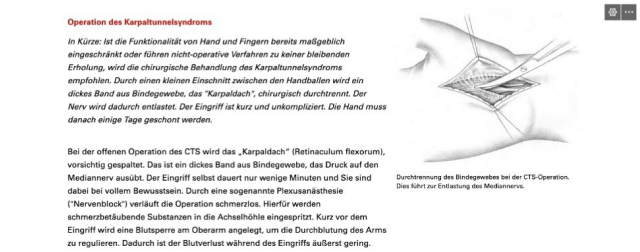

- Short summary
- Followed by longer and more detailed explanation
- Images for better comprehensibility

Unterwegs mit Karla (YouTube-Video)

- Video with spoken and written text
- Animated character "Karla" as a patient role model, sharing advice for undergoing CTS surgery
- Information on perioperative situation

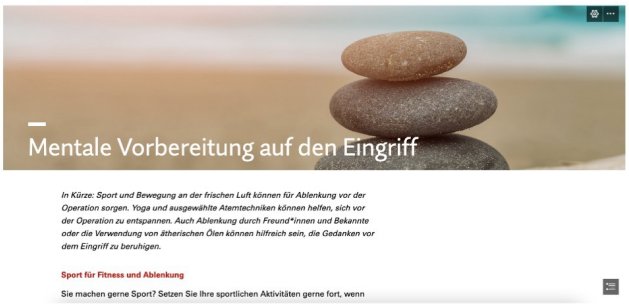

- Mental preparation for targeting stress and anxiety before surgery
  - Talks, exercise, music, breathing exercises, essential oils etc.

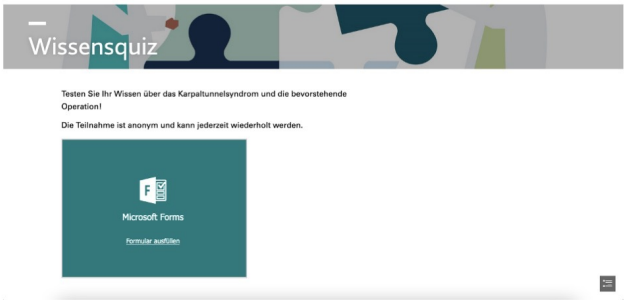

- Short quiz for checking understanding of important information
  - What is CTS? recurrence after surgery? etc.

XSL•FO
RenderX

Design of the Evaluation Study

Between April 2022 and April 2023, all patients presenting with CTS symptoms at the Department of Orthopedics and Traumatology and scheduled for CTR surgery after medical assessment were assessed for eligibility for inclusion in this nonrandomized, controlled, confirmatory study.

During consultations, patients received detailed information about their condition and perioperative procedures. After obtaining written informed consent for CTR, we introduced the CTS Academy study and explained the test or control groups. Allocation to the groups was based on the patient's preference—either participation in CTS Academy or self-directed online information seeking.

Inclusion criteria were a scheduled appointment for CTR surgery, voluntariness and age of majority (18 y of age or older), physical and cognitive aptitude to participate, sufficient language skills of written and spoken German, and access to web-connected devices.

Furthermore, patients needed sufficient skills for internet use and an active pursuit of online health information.

Patients in the test group were granted unlimited access to the CTS Academy (from informed consent to surgery until 3 w postoperatively), while the control group conducted independent online searches. Throughout the self-study period of both groups, reminders were sent via SMS text messages or email to encourage information access.

Participants completed an online questionnaire after inclusion for the study, and 3 weeks postoperatively (follow-up appointment). As the CTS Academy study was conducted with German-speaking patients, all questionnaires were used in validated German versions.

The primary outcome was assessed using the Health Education Literacy of Patients with chronic musculoskeletal diseases (HELP) questionnaire by Farin, Ullrich, and Nagl [9]. It comprises 3 subtopics, that are comprehension of medical information (COMPR), applying medical information (APPLY), and communicative competence in provider interactions (COMM). Patients rated their perceived difficulty in health-related situations on a 5-point scale, with lower values indicating fewer difficulties.

Secondary outcomes included assessing the comprehensibility of content provided in the CTS Academy using the Comprehensibility Of Health Education Programmes (COHEP) questionnaire by Farin-Glattacker et al [28] and Farin, Nagl, and Ullrich [29]. This questionnaire comprises 4 subscales: transferability scale, medical information scale, amount scale, and the trainer scale, while the latter was dismissed because we did not offer lessons by a trainer [29]. Patients had to rate the amount of information provided or the perceived usefulness of information on a 6-point ranging from 1 (strongly agree) to 6 (strongly disagree).

Knowledge checks involved answering 4 statements about CTS and its treatment [30] to check their knowledge at baseline after inclusion to the study, and 3 weeks postoperatively.

Patient preferences of information seeking in the test group were assessed using the following statements on a 5-point rating scale (1=strongly disagree, 2=disagree, 3=no preference, 4=agree, and 5=strongly agree) in the follow-up questionnaire:

- I prefer the additional use of the CTS Academy (in addition to talking with a doctor) over the sole standard patient information provided in the medical informed consent conversation.
- I prefer the use of the CTS Academy over independent online information-seeking.

Usability and utility were evaluated using selected items of meCUE 2.0 [31], rated in the follow-up questionnaire on a 7-point scale ranging from 1 (strongly disagree) to 7 (strongly agree).

Clinical CTS-related parameters were examined using the validated German-language version [32] of the Boston Carpal Tunnel Questionnaire (BCTQ) [33] using symptom severity scale, functional status scale, as well as the total BCTQ score at baseline and follow-up.

Before patient testing, we conducted a pre-, pilot-, and prototype testing of the CTS Academy platform in August 2021 to examine the usability as well as to detect functionality and usability issues. This testing involved 16 voluntary participants: 11 laypersons (some with previous CTS surgery experience) and 5 experts in hand surgery, psychology, sociology, e-learning, and online didactics. Laypersons were asked to imagine themselves in the role of CTS patients undergoing surgery. Experts reviewed the platform professionally. The feedback and suggestions were discussed with the other project team members and led to updates in pre- and postoperative recommendations, as well as revisions for clarity and gender-appropriate language (shortening of the sentence length, simplifying certain terms, and standardizing gender-inclusive language). Recommendations were added regarding preoperative and postoperative smoking abstinence, as well as the ergonomic use of computer mice and keyboards after surgery. A chapter about the author of the CTS Academy was added.

Statistical Analysis

Sample Size Estimation and Enrollment

We performed sample size estimation in G*Power (version 3.1.9.7) [34] using the following parameters: Wilcoxon-Mann-Whitney test (2 groups), 1-tailed hypothesis, Cohen *d* of 0.8, α error probability of 0.05, and power of 0.8 [35]. This yielded a required sample size of 21 patients per group, totaling 42 patients. Anticipating a 20% dropout rate, we aimed to recruit at least 50 patients. However, due to the COVID-19 pandemic's impact on surgical resourced and increased dropouts, we ultimately enrolled 60 patients. Between April 2022 and April 2023, 60 patients scheduled for open CTR at our hospital participated in the CTS Academy study. Patients presenting with CTS symptoms at the hand surgery outpatient clinic underwent diagnostic evaluation, and those requiring surgery were scheduled accordingly. During preoperative consultations, they received information about the procedure, risks, and postoperative care before providing written informed consent. Eligible patients—adults with sufficient German

XSL•FO
RenderX

language skills, cognitive and physical ability, regular internet use as well as willingness to seek online health information—were invited to join the study. No restrictions were placed on gender or education to ensure a representative sample. Interested patients could assign themselves to either the test group (structured web-based education) or control group (self-directed online information seeking) and received study details, including access to an online baseline questionnaire. Those in the test group were provided instructions for using the CTS Academy platform. Throughout the study, patients received reminders via email and SMS to complete study-related tasks, including the baseline and follow-up questionnaires. They were also notified before surgery and received postoperative messages encouraging engagement. Nonresponders were contacted within 2 days. Throughout the study, participants could reach the study team via phone, SMS, or email.

Data Exclusion and Dropouts

In our online questionnaires, all fields were mandatory except for the BCTQ functional scale, as some patients may not perform tasks specified in this BCTQ scale. Patients could review and change their answers before submitting the questionnaire. In the test group, 1% (2/144) of values were missing at baseline and at follow-up. In the control group, 6% (8/144) and 1% (2/144) were missing at baseline and follow-up, respectively. Following BCTQ guidelines, we excluded participants with more than 1 missing value, resulting in the exclusion of 1 participant from each group [36]. Since BCTQ activities are not directly comparable (eg, holding a book, doing chores, and getting dressed), missing values were not imputed.

Due to the per-protocol analysis of this study, 12 patients in each group dropped out of the study due to various reasons such as nonsubmission of questionnaires, withdrawal of consent, or canceled surgery appointments.

Data Preparation and Analysis

Microsoft Excel 16.72 for Mac, and IBM SPSS (version 29) were used for statistical testing.

Descriptive statistics included means and SD as well as the IQR to identify potential outliers. Normal distribution was tested using the Shapiro-Wilk-Test. The HELP questionnaire results were analyzed by summing each subscale (APPLY, COMM, and COMPR) for individual patients and transforming these sums to a 0‑100 scale. Delta values between pre- and postoperative scores (DIFF_APLLY, DIFF_COMM, and DIFF_COMPR) were calculated.

Depending on the parameters' normal distribution, either the exact Wilcoxon test or the *t* test for dependent variables and the Mann-Whitney *U* test or the *t* test for independent variables were used for comparisons. The total HELP score (a sum of the subtypes APPLY, COMM, and COMPR) was tested at baseline, follow-up, and as difference between the former using the multivariate analysis of variances (MANOVA).

Nominal data were tested using the chi-square test or the Fisher exact test if the requirements of the former test were not met. Effect sizes were determined using Psychometrica [37].

Demographic data were examined for group comparability, focusing on age and educational level. Statistical tests were primarily 1-sided, with significance set at $P<.05$. We converted results from default 2-sided tests to 1-sided outcomes and did not perform subgroup analyses.

The COHEP questionnaire results were analyzed by summing subscales and transforming them to a 0‑100 scale. Knowledge check results were reported as the frequency of correct answers within each group. Usability and utility of the CTS Academy platform were assessed using the meCue 2.0 data template [38]. For the assessment of CTS-related symptom changes, symptom severity, functional status, and total BCTQ, of which a delta value (DIFF_BCTQ) was calculated, were compared between cohorts as well as longitudinally.

Ethical Considerations

This study was approved and registered by the institutional review board (EK1212-2021). Written informed consent (Multimedia Appendix 2) based on the hospital's regulative for clinical studies was obtained from every patient willing to participate in the study. To maintain data security, patients were pseudonymized by assigning each one an alphanumeric ID at the time of study enrollment. This ID was used exclusively for all surveys. A list of IDs alongside patient names was kept in the study office and was accessible to the project management team, for example, to use the reminder system (eg, reminders to complete the questionnaire). During the analysis of the results, it was not possible at any time to link the identity of the test participants to the available data. No identification of individual participants in any of the figures or additional material is possible. Patients received no compensation for study participation.

The study was reported using the CHERRIES (Checklist for Reporting Results of Internet E-Surveys) checklist (Multimedia Appendix 3) [39].

Results

Scoping Review

In total, 17 reports were included for full-text analysis, finding that patients searching online for CTS information risk encounter incorrect or confusing content [3], which underlines the need for reliable education programs. An overview of the results is presented in Table 1.

XSL•FO
RenderX

Table . Outcomes of the 17 references included in the scoping review on patient-centered web-based education programs focusing on the perioperative process of carpal tunnel release conducted in Epistemonikos, the Cochrane Library, PubMed, Connected Papers, and Google Scholar up to December 2022, discussing the quality of online carpal tunnel syndrome information.

Reference	Topic	Relevant results
Cook et al [40]	Complexity of online information on hand surgery	The recommended 6th-grade reading level was not met in any of the analyzed CTS[a] material, which decreases understandability for many patients.
Goyal et al [41]	Analysis of misleading CTS information in YouTube videos	78% (47/60) of the analyzed videos contained at least 1 potentially misconception-reinforcing statement, more than 335% (21/60) contained more than 3 statements, mostly on CTS causes and treatment options.
Lutsky et al [42]	Re-evaluation of CTS information on the internet (updated study from 2000)	Significant improvement of online CTS information quality (P<.001), further improvement is needed.
Steimle et al [43]	Comparison of online CTS information with the Clinical Practice Guidelines of the American Academy of Orthopaedic Surgeons	Incorrect information might provoke misunderstandings and potential risks for patients.
Özbek et al [44]	Quality and reliability of CTS videos on YouTube	Significant differences in reliability and informational scores (P<.05) depending on the source of the video, and generally low reliability and educational quality.
Fang et al [45]	Content and design of online CTS education handouts	Handouts contained misleading information, lacked references or citations, and were complicated to understand.
Frueh et al [46]	Quality of online CTS information	Significant difference in information quality in older compared to newer CTS websites (P<.001). Significant difference in explaining the handling of complications (P=.03).
Frické et al [47]	Indicators of accuracy for online CTS information	Indicators like high Google ratings and inlinks were correlated with accuracy (P<.05).
Beredjiklian et al [48]	Source and content quality of online CTS information	Several websites had misleading or unconventional about CTS treatment. The mean information value score of all websites was low and differed based on authorship.
Kwak et al [49]	Quality and reliability of CTS YouTube videos	Low reliability and low educational quality of the analyzed videos.
Johnson et al [50]	Comprehensibility and applicability, cultural sensitivity, and readability of English and Spanish online CTS information material	Low comprehensibility, applicability, and readability scores, and no statistically significant differences in the overall educational quality between English and Spanish material.
Eberlin et al [51]	Readability of online CTS material	The recommended 6th grade reading level was not met in any of the analyzed CTS material, which decreases understandability for many patients.
Mert and Bozgeyik [52]	Quality and content of YouTube videos about CTS	Generally insufficient quality of content, only few videos reached good scores for global quality and accuracy.
Sproule et al [53]	Quality of online information about carpal tunnel release, Dupuytren's fasciectomy, and trigger finger release	The majority of material provided accurate but incomplete information, and a lack of information on prognosis and complications of surgical treatment.
Ozdemir et al [54]	Quality and reliability of CTS videos on YouTube	Overall low quality of videos concerning accuracy and transparency. Lack of information on treatment risks and benefits. Significantly better scores when authored by medical centers (P=.02).

XSL•FO
RenderX

Reference	Topic	Relevant results
Park et al [55]	Accuracy of online CTS images	Only 30% (6/20) of the analyzed images had incorrect anatomical information, for example, the position of the median nerve.
Akinleye et al [56]	Readability of online hand pathology material including CTS	The recommended 6th grade reading level was not met in any of the analyzed CTS material, which decreases understandability for many patients.

[a]CTS: carpal tunnel syndrome.

Patient Characteristics

Between April 2022 and April 2023, 30 patients were enrolled in each group. Figure 2 displays a flowchart of patients including the reasons for dropout.

Ultimately, both the test and the control group comprised of 18 patients each, whose characteristics are summarized in Table 2.

The average educational phase duration (CTS Academy platform or self-directed online information seeking) was 167 (range 43-267) days, from inclusion to 176 (range 43-253) days postoperatively.

Figure 2. Consort flowchart of patients' enrollment, follow-up, and analysis in the test group (CTS Academy) and control group (Self-directed online information seeking). CTR: carpal tunnel syndrome; CTS: carpal tunnel sydrome; HEL: health education literacy.

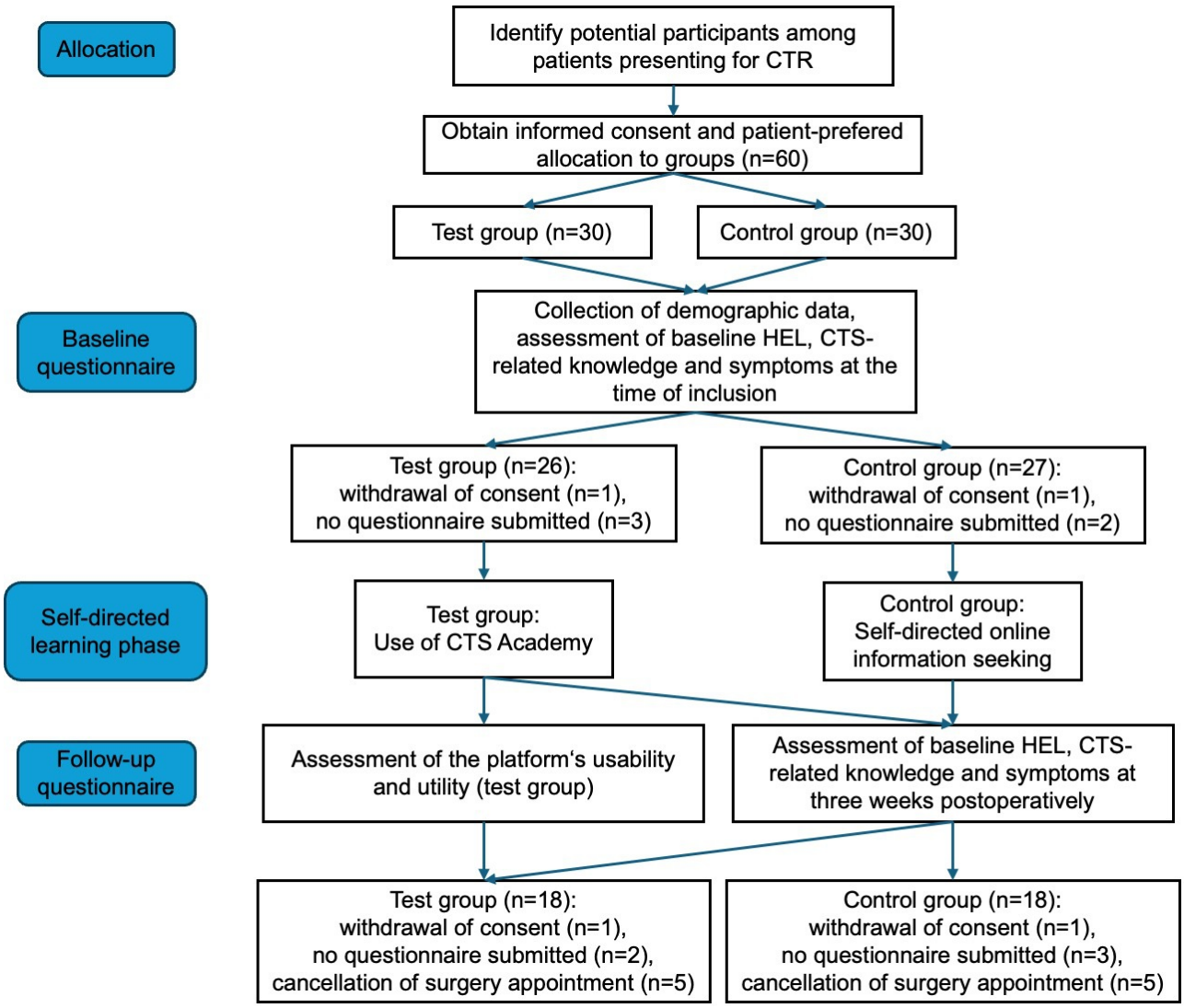

XSL•FO
RenderX

Table . Patient characteristics and demographics of the final test group (CTS Academy platform) and control group (self-directed online information seeking).

Categories and subcategories		Test group (n=18), n	Control group (n=18), n
Sex			
	Male	5	7
	Female	13	11
Age group			
	<30 years	0	0
	30 - 40 years	4	3
	41 - 50 years	3	3
	51 - 60 years	5	2
	61 - 70 years	3	6
	71 - 80 years	1	4
	>80 years	2	0
Highest education			
	Compulsory schooling or less	2	5
	High school (without A-levels) or apprenticeship	11	9
	High school with A-levels, diploma, apprenticeship on "Master" level	4	4
	Undergraduate degree (bachelor)	1	0
	Graduate degree (master, PhD, etc)	0	0
CTS[a] surgery			
	First time	14	11
	Previous CTS surgery (other hand)	4	6
	Previous CTS surgery (same hand)	0	0
	Previous CTS surgery (both hands)	0	1
Internet use			
	Several times a month	5	7
	Several times a week	5	3
	Daily	8	8
Average duration of CTS Academy use (days)		167 (range 43-267)	176 (range 43-253)

[a]CTS: carpal tunnel syndrome.

Primary Outcome

Outcomes of the primary parameter and its subtypes are presented in Table 3 and Table 4.

XSL•FO
RenderX

Table . Outcomes (pre-post intervention analysis) of health education literacy of patients with chronic musculoskeletal diseases questionnaire (health education literacy) via subitems comprehension of medical information, applying medical information, and communicative competence in provider interactions of the final test group (CTS Academy) and control group (self-directed online information seeking) including statistical results (all *P* values are reported as 1-sided *P* values).

	APPLY[a]			COMM[b]			COMPR[c]		
	Baseline, mean (SD)	Follow-up, mean (SD)	*P* value	Baseline, mean (SD)	Follow-up, mean (SD)	*P* value	Baseline, mean (SD)	Follow-up, mean (SD)	*P* value
Test group (n=18)	28 (7.99)	24 (5.14)	.006[d]	30 (10.52)	25 (6.01)	.017[d]2	38 (16.60)	35 (9.90)	.24
Control group (n=18)	33 (14.17)	33 (14.78)	.5	35 (17.12)	30 (15.19)	.0596	50 (19.00)	46 (23.75)	.18

[a]APPLY: applying medical information.

[b]COMM: communicative competence in provider interactions.

[c]COMPR: comprehension of medical information.

[d]Statistically significant results (α<.05).

Table . Outcomes (comparison between groups [*P* values]) of health education literacy of patients with chronic musculoskeletal diseases questionnaire (health education literacy) via subitems comprehension of medical information, applying medical information, and communicative competence in provider interactions of the final test group (CTS Academy) and control group (self-directed online information seeking) including statistical results (all *P* values are reported as 1-sided *P* values).

	APPLY[a]	COMM[b]	COMPR[c]	HEL[d]
Baseline	.2	.3	.03[e]	.13
Follow-up	.04[e]	.27	.06	.05[e]
Difference	.09	.38	.37	__[f]

[a]APPLY: applying medical information.

[b]COMM: communicative competence in provider interactions.

[c]COMPR: comprehension of medical information.

[d]HEL: health education literacy.

[e]Statistically significant results (α<.05).

[f]Not applicable.

In the test group, we found a statistically significant decrease in the group means for APPLY (*P*=.006) and COMM (*P*=.02) after using the CTS Academy, compared with baseline. However, we found no statistically significant difference regarding the COMPR subparameter. An overview of statistically significant postoperative changes on item-level is provided in Multimedia Appendix 4. In the control group analysis, we found no statistically significant difference.

In the comparison between the test and control group, significantly higher APPLY values were found in the test group during the follow-up assessment (*P*=.04). For COMPR, we found the test group had significantly fewer problems with comprehension of medical information at baseline (*P*=.03).

We found a statistically significant difference between the groups on the combined dependent variables at follow-up (*P*=.05). In the subsequent post hoc univariate ANOVA for each dependent variable (between-participant effects), we found a statistically significant difference between the groups for APPLY in the follow-up examination (*P*=.007).

Secondary Outcomes

As lower means indicate better results, we found the CTS Academy yielded good results on the transferability (mean 27, SD 9.74) and the medical information scale (mean 28, SD 10.12) and scored average results for the amount of medical information provided (mean 58, SD 24.61).

Both groups achieved high accuracy in describing CTS and the surgical procedure at baseline (63/72, 88% vs 62/72, 86% in test group and control group). At follow-up, the test group showed an average accuracy of 92%, with incorrect answers for only 1 statement. The control group yielded an average accuracy of 90% with incorrect answers for all 4 statements. No control group patient answered all statements correctly.

Regarding patients' preference for using the CTS Academy, 10 out of 18 (56%) patients totally agreed and 6 out of 18 (33%) agreed that they preferred the additional use of the CTS Academy (in addition to the consultation of a medical doctor). Furthermore, 2 patients had no preferences. When asked whether they preferred the CTS Academy over independent online information-seeking, 10 out of 18 (56%) patients totally agreed and 5 out of 18 (28%) agreed, while 3 patients had no preferences.

Patients rated usability with a mean 6.22 (SD 0.82) points (4.67 to 7.00 points), and utility with a mean 6.13 (SD 0.89) points (4.00 to 7.00 points).

XSL•FO
RenderX

The CTS-related symptoms assessed via the BCTQ are displayed in Table 5.

We found no statistically relevant differences between the groups (comparison of groups at baseline: P=.34 [2-sided], follow-up: P=.47 [2-sided], and for DIFF_BCTQ: P=.86 [2-sided]).

Table . Carpal tunnel syndrome–related symptoms of test group and control group at baseline and follow-up, assessed with the Boston carpal tunnel questionnaire.

	Symptom severity, mean (SD)		Functional status, mean (SD)	
	Baseline	Follow-up	Baseline	Follow-up
Test group (n=18)	3.03 (0.68)	1.99 (0.71)	2.24 (0.73)	2.24 (0.87)
Control group (n=18)	3.36 (0.92)	2.19 (0.98)	2.40 (1.04)	2.48 (1.04)

Discussion

Principal Results

This study introduced the CTS Academy to improve patients' HEL compared with self-directed online information seeking. Significant improvements were observed in the HELP subscales APPLY and COMM within the test group, indicating that the CTS Academy enhances patients' ability to apply medical information and communicate with health care professionals. Patients in the study group rated both utility (usefulness) and usability (ease of use) of the CTS Academy favorably.

CTS Academy content was evaluated favorably for transferability and medical information, with average ratings for the amount of medical information provided. Development relied on existing research in clinical patient education due to a lack of comprehensive guidelines, potentially affecting perceived content quality [17]. Future improvements could involve personalized pathways based on patient knowledge levels.

Both groups showed strong understanding of CTS etiology before and after the intervention. The test group's understanding of CTR surgery improved notably, with all participants (N=18) having all answers correct at follow-up. Misconceptions about complete symptom removal persisted in both groups, which might be due to a focus on postoperative recovery over potential complications.

Patients preferred the CTS Academy over self-directed online searches, indicating its value as a reliable information source. We conclude that integrating the CTS Academy into patient education could effectively enhance access to information and understanding of the perioperative situation, complementing discussions with medical professionals during informed consent.

High ratings for utility and usability suggest the platform is user-friendly and effective for CTR day surgery.

No significant differences were found between groups regarding perceived symptom severity or functional status, reflecting consistent surgical intervention.

The CTS Academy is the first web-based patient education program for CTS. No direct comparisons in medical literature are available, especially because no comparable programs have health education as primary outcome. Future updates to the CTS Academy will focus on restructuring content and enhancing readability and comprehensibility. The platform has the potential to evolve into an educational program with user accounts, enabling patients to follow structured, personalized learning paths. The integration of technologies such as artificial intelligence could further personalize and enhance the learning experience. Another future goal of this project is expanding the platform to include multiple languages relevant to clinical practice at our hospital.

Comparison With Previous Work

Patient education is crucial for day surgery, as it helps prevent adverse events and rehospitalization [57]. Pekonen et al [58] conducted a systematic review on tools for measuring patient empowerment after patient education interventions. While the Health Education Impact Questionnaire is well-established [59], the HELP questionnaire specifically addresses HEL. Because the latter not just assesses knowledge acquisition but also patients' ability to apply and understand educational interventions, it was deemed more appropriate for the CTS Academy study.

A publication by Farin-Glattacker et al [28] assessed HEL using the HELP questionnaire as an outcome measure in patients undergoing rehabilitation for chronic back pain or osteoarthritis [60]. The study group stated a high level of self-reported HEL, with the average patient experiencing slight difficulty on the APPLY and COMM scales, and slight to moderate difficulty on the COMPR scale [9]. It aligns with the findings from the CTS Academy that difficulty in understanding medical terminology appears to be a common challenge [61-64].

Regarding computer-based education on CTS, a randomized study compared knowledge levels and satisfaction scores between patients receiving the same information either using computer-based education or using face-to-face education with a medical doctor. Ultimately, the computer-based education group scored significantly higher on the knowledge questionnaire [35].

In total, 2 previous publications focused on the topic of structured web-based patient education versus self-directed online information seeking. Finkelstein and Bedra [65] conducted this comparison on hypertension knowledge in 60 patients with hypertension. The test group used a structured web-based patient education program on hypertension, while the control group searched the internet for information. The test group showed significantly higher hypertension knowledge scores compared with the baseline and compared with the

XSL•FO
RenderX

control group. Semistructured interviews indicated a better learning experience in the test group, particularly for those with limited computer and internet experience. Castro et al [66] found comparable results in a similar cohort study. Thus, both concluded that web-based patient education was effective for hypertension learning, even for participants with limited internet experience [65,66].

Although health literacy is often seen as empowering, it does not automatically lead to empowerment [67]. Health education is crucial for promoting health literacy [67], with health knowledge being a key dimension [68,69] according to the conceptual model of health literacy by Sørensen et al [70]. Low health literacy levels are associated with poorer health outcomes, increased morbidity and mortality [71], highlighting the importance of patient education. Several studies have proven that multimedia education programs can improve health knowledge, such as in patients with type 2 diabetes [72]. Despite existing literature on digital health interventions for improving health literacy, only few studies specifically measure health literacy outcomes, often focusing on knowledge-related outcomes instead [73]. This distinction underscores the importance of CTS Academy's focus on HEL.

Various publications on web-based patient education address the quality and readability of online educational materials, but few compare structured web-based patient education and self-directed online information seeking. Many examined online resources presenting orthopedic, trauma surgery, pediatrics, or cardiology content [74-76] exceed the recommended sixth-grade reading level [77], posing comprehension challenges. The quality and accuracy of online educational resources varies, with those authored by medical professionals only being slightly better than those authored by commercial organizations or laypersons [44,48,54]. Due to a lack of structured analyses of German educational online resources, quality and accuracy of online information for German-speaking CTS patients cannot be assessed. However, studies indicate that more than one-third of German-speaking orthopedic patients used the internet before their hospital appointment, and 80% (63/79) of these patients intended to gain additional knowledge of their condition and treatment options. The majority of patients performing self-directed online information seeking would recommend it to other orthopedic patients [78]. Hertling et al [79] reports that 66% (832/1262) of orthopedic patients used the internet as a source of disease-related information, and the majority found the information trustworthy and useful. However, it is unclear whether patients can assess the reliability and correctness of online information and identify incorrect information [79]. While both studies indicated widespread use and acceptance of online sources among German orthopedic patients, the quality of the information sources remains unexamined.

Comparing the comprehensibility of CTS Academy content with other programs using the COHEP questionnaire is challenging due to differing contexts. Existing studies by Farin-Glattacker et al [28], Nagl et al [60], and Finkelstein et al [65] focused on different patient populations and did not provide detailed comparative data.

Patients provided positive feedback on CTS Academy's usability and utility, appreciating its design and functionality, facilitating effective access and utilization of educational materials. Overall, these findings suggest that the CTS Academy is a valuable tool for educating and supporting patients undergoing day surgery for CTR.

In our study, we collected postoperative clinical parameters 3 weeks after CTR. Our clinical results align with the ones published by Mack et al [80] at 10 - 14 days postoperatively and Kim et al [81] and Aloi et al [82] at 14 days postoperatively, despite slight differences in patient characteristics and surgical approaches.

Limitations

This study also involves several limitations. The study was affected by COVID-19, especially by prolonged surgery waiting times leading patients to cancel their surgery appointments. This led to a smaller sample size than intended, potentially biasing our results. Furthermore, the patient-performed allocation process may have introduced selection bias. Moreover, group sizes were relatively small, potentially affecting the generalizability of the results. We did not ask patients to monitor the duration of information-seeking or the frequency of use of the CTS Academy. In addition, the quality of information accessed by the control group was not controlled, which could have influenced their outcomes and skewed comparisons with the intervention group. However, both groups completed identical baseline questions on CTS before accessing the platform content or other online resources, ensuring a basic understanding of the scope of information. These questions required basic knowledge, independent of the depth of research or exclusive platform content. This highlights a common challenge in independent online research, that is, patients struggle to identify essential information, and available content can be difficult to interpret. While the test group benefited from a patient-centered platform design, this advantage did not extend to the fundamental information itself. To preserve the natural research and learning process, patients were not required to document their sources, as this might disrupt their learning experience. For privacy reasons and to prevent artificially inflated research activities, browser histories were not collected to avoid potential identification or bias. In future research, randomization allocation to either the test or control group should be conducted to further investigate effects of the CTS Academy platform on HEL, ensuring that the results are not affected by a selection bias. In terms of usability, common usability testing guidelines report that 5 participants already detect almost all of the possible errors in a product [83]. Therefore, the usability conclusions can be seen as valid despite of the generally decreased study's power and generalizability.

Conclusions

Knowledge is commonly measured as a primary outcome in patient education programs for its ease of pre and post comparisons. However, the ability to apply knowledge in daily life is crucial for effective self-management. Health literacy models emphasize both knowledge and its practical application. Hence, we recommend a comprehensive evaluation for patient education programs, beyond knowledge assessments. Our

XSL•FO
RenderX

findings highlight the effectiveness of our educational intervention for patients undergoing CTR day-surgery. Despite challenges in comparing outcomes with other studies, statistical analysis favored the CTS Academy in improving overall HEL. Therefore, we rejected the null hypothesis, confirming the platform's efficacy over self-directed online information seeking. Future research could explore the quality of German-language online educational resources by having patients document and analyze their information sources. In addition, the effectiveness of the CTS Academy following its implementation at the Department of Orthopedics and Traumatology could be assessed. Investigating patient preferences and the utility of each specific chapter could help identify the most relevant educational needs and could be analyzed on item-level of each HELP or COHEP subscale for a granulated evaluation of patients' learning experience.

Acknowledgments

This study is supported by Johannes Kepler University Open Access Publishing Fund and the Federal State Upper Austria.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization, methodology, software, data curation, and writing—original draft preparation were contributed and handled by MS. Validation was contributed by TG, JM, and BH. Formal analysis and investigation was contributed by MS, JAM, and SMF. Resources was handled by TG. Writing—review and editing was contributed by MH. Visualization was contributed by MS and MH. Supervision was managed by BH, JM, and TG. Project administration was handled by SMF.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Overview of content and media types used in the CTS Academy (in order of appearance on the platform).
[DOCX File, 74 KB - formative_v9i1e65114_app1.docx]

Multimedia Appendix 2
Informed consent.
[PDF File, 239 KB - formative_v9i1e65114_app2.pdf]

Multimedia Appendix 3
Cherries (Checklist for Reporting Results of Internet E-Surveys) checklist.
[PDF File, 670 KB - formative_v9i1e65114_app3.pdf]

Multimedia Appendix 4
Overview of pre- and postoperative health education literacy on item-level for the test group.
[DOCX File, 24 KB - formative_v9i1e65114_app4.docx]

References

1. Hälleberg Nyman M, Nilsson U, Dahlberg K, Jaensson M. Association between functional health literacy and postoperative recovery, health care contacts, and health-related quality of life among patients undergoing day surgery: secondary analysis of a randomized clinical trial. JAMA Surg 2018 Aug 1;153(8):738-745. [doi: 10.1001/jamasurg.2018.0672] [Medline: 29710226]
2. Berg K, Arestedt K, Kjellgren K. Postoperative recovery from the perspective of day surgery patients: a phenomenographic study. Int J Nurs Stud 2013 Dec;50(12):1630-1638. [doi: 10.1016/j.ijnurstu.2013.05.002] [Medline: 23726224]
3. Tonsaker T, Bartlett G, Trpkov C. Health information on the internet: gold mine or minefield? Can Fam Physician 2014 May;60(5):407-408. [Medline: 24828994]
4. Connelly RA, Speer ME. Health literacy and health communication. In: Connelly RA, Turner T, editors. Health Literacy and Child Health Outcomes: Springer International Publishing; 2017:1-13. [doi: 10.1007/978-3-319-50799-6_1]
5. Charalambous A, Papastavrou E, Valkeapää K, et al. Content of orthopedic patient education provided by nurses in seven European countries. Clin Nurs Res 2018 Sep;27(7):770-789. [doi: 10.1177/1054773817713178] [Medline: 28691509]
6. World Health Organization. Empowerment. In: Health Promotion Glossary of Terms: World Health Organization; 2021. URL: https://www.who.int/publications/i/item/9789240038349 [accessed 2025-03-21]

XSL•FO
RenderX

7. Bastable SB, editor. Essentials of Patient Education, 2nd edition: Jones & Bartlett Learning; 2017.
8. Wittink H, Oosterhaven J. Patient education and health literacy. Musculoskelet Sci Pract 2018 Dec;38:120-127. [doi: 10.1016/j.msksp.2018.06.004] [Medline: 30017902]
9. Farin E, Ullrich A, Nagl M. Health education literacy in patients with chronic musculoskeletal diseases: development of a new questionnaire and sociodemographic predictors. Health Educ Res 2013 Dec;28(6):1080-1091. [doi: 10.1093/her/cyt095] [Medline: 24113084]
10. Towfigh H, Hierner R, Langer M, Friedel R, editors. Handchirurgie Bd: Springer; 2011, Vol. 2.
11. Atroshi I, Gummesson C, Johnsson R, Ornstein E, Ranstam J, Rosén I. Prevalence of carpal tunnel syndrome in a general population. JAMA 1999 Jul 14;282(2):153-158. [doi: 10.1001/jama.282.2.153] [Medline: 10411196]
12. American Academy of Orthopaedic Surgeons. Carpal tunnel syndrome. : OrthoInfo; 2016. URL: https://orthoinfo.aaos.org/en/diseases--conditions/carpal-tunnel-syndrome/ [accessed 2021-01-26]
13. Donati D, Boccolari P, Tedeschi R. Manual therapy vs. surgery: which is best for carpal tunnel syndrome relief? Life (Basel) 2024 Oct 11;14(10):1286. [doi: 10.3390/life14101286] [Medline: 39459587]
14. The Joanna Briggs Institute. Templates for scoping reviews. Downloadable templates in Word guide and assist reviewers in developing a scoping review protocol and scoping review. 2022. URL: https://jbi.global/scoping-review-network/resources [accessed 2022-12-30]
15. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
16. Microsoft Corporation. Microsoft sway. : 2021 URL: https://sway.office.com/ [accessed 2025-03-21]
17. Health Care Education Association. Patient education practice guidelines for health care professionals. 2021. URL: https://www.hcea-info.org/assets/hcea%20guidelines%20color%201-25-2021.pdf [accessed 2025-03-21]
18. Keulers BJ, Keulers MJ, Scheltinga MRM, Spauwen PHM. Evidence-based development of interactive patient education programs: a guide for achieving optimal computer-based patient education. Eur J Plast Surg 2006 Nov 13;29(4):169-175. [doi: 10.1007/s00238-006-0085-7]
19. Raunsbæk Knudsen L, Lomborg K, Ndosi M, Hauge EM, de Thurah A. The effectiveness of e-learning in patient education delivered to patients with rheumatoid arthritis: The WebRA study-protocol for a pragmatic randomised controlled trial. BMC Rheumatol 2021 Dec 20;5(1):57. [doi: 10.1186/s41927-021-00226-y] [Medline: 34924034]
20. Knudsen LR, Lomborg K, de Thurah A. Design and development of an e-learning patient education program for self-management support in patients with rheumatoid arthritis. PEC Innov 2022 Dec;1:100004. [doi: 10.1016/j.pecinn.2021.100004] [Medline: 37364010]
21. Sauerbier M, Allert S. Operation des karpaltunnelsyndroms (KTS) offen/endoskopisch [report in german]. 2017 URL: https://thieme-compliance.de/de/shop/Artikel/Aufkl%C3%A4rungsb%C3%B6gen/Karpaltunnelsyndrom-%28KTS%29%2C-Operation/p/DE61748501 [accessed 2025-03-25]
22. Norelli SK, Long A, Krepps JM. Relaxation techniques. In: StatPearls: StatPearls Publishing; 2021. URL: http://www.ncbi.nlm.nih.gov/books/NBK513238 [accessed 2023-07-19]
23. Anxiety and Depression Association of America (ADAA). Exercise for stress and anxiety. : Anxiety and Depression Association of America (ADAA); 2021. URL: https://adaa.org/living-with-anxiety/managing-anxiety/exercise-stress-and-anxiety [accessed 2021-05-06]
24. Pardede J, Simanjuntak GV, Manalu N. Effectiveness of deep breath relaxation and lavender aromatherapy against preoperative patent anxiety. Divers Equal Health Care 2020;17(4):168-173. [doi: 10.36648/206]
25. Uğraş GA, Yıldırım G, Yüksel S, Öztürkçü Y, Kuzdere M, Öztekin SD. The effect of different types of music on patients' preoperative anxiety: a randomized controlled trial. Complement Ther Clin Pract 2018 May;31:158-163. [doi: 10.1016/j.ctcp.2018.02.012] [Medline: 29705448]
26. Braden R, Reichow S, Halm MA. The use of the essential oil lavandin to reduce preoperative anxiety in surgical patients. J Perianesth Nurs 2009 Dec;24(6):348-355. [doi: 10.1016/j.jopan.2009.10.002] [Medline: 19962101]
27. American Academy of Orthopaedic Surgeons, OrthoInfo. Therapeutic exercise program for carpal tunnel syndrome. 2018. URL: https://orthoinfo.aaos.org/globalassets/pdfs/a00789_therapeutic-exercise-program-for-carpal-tunnel_final.pdf [accessed 2025-03-21]
28. Farin-Glattacker E, Schöpf A, Ullrich A, Nagl M. Abschlussbericht zum projekt „entwicklung und evaluation einer patientenschulung zur Förderung der gesundheitskompetenz Für chronisch kranke" (geko) [article in german[. Universitätsklinikum Freiburg. 2014. URL: https://docplayer.org/14332968-Abschlussbericht-zum-projekt.html [accessed 2025-03-21]
29. Farin E, Nagl M, Ullrich A. The comprehensibility of health education programs: questionnaire development and results in patients with chronic musculoskeletal diseases. Patient Educ Couns 2013 Feb;90(2):239-246. [doi: 10.1016/j.pec.2012.10.004] [Medline: 23127897]
30. Aung KK, Wu WK, Tokumi A, Kuo P, Day CS. Does a directive to an internet site enhance the doctor-patient interaction? A prospective randomized study for patients with carpal tunnel syndrome. J Bone Joint Surg Am 2015 Jul 1;97(13):1112-1118. [doi: 10.2106/JBJS.N.00741] [Medline: 26135078]

XSL•FO
RenderX

31. Minge M, Thüring M, Wagner I, Kuhr CV. The mecue questionnaire: A modular tool for measuring user experience. In: Soares M, Falcão C, Ahram TZ, editors. Advances in Ergonomics Modeling, Usability & Special Populations: Springer International Publishing; 2017, Vol. 486:115-128. [doi: 10.1007/978-3-319-41685-4_11]
32. Keilani M, Pernicka E, Paternostro-Sluga T, et al. Übersetzung und Validierung des „Boston Carpal Tunnel Syndrome Questionnaire" zum Einsatz bei deutschsprachigen Patienten. Phys Rehab Kur Med 2008 Jun;18(3):136-144. [doi: 10.1055/s-2007-1004550]
33. Levine DW, Simmons BP, Koris MJ, et al. A self-administered questionnaire for the assessment of severity of symptoms and functional status in carpal tunnel syndrome. J Bone Jt Surg 1993 Nov;75(11):1585-1592. [doi: 10.2106/00004623-199311000-00002]
34. Buchner A, Erdfelder E, Faul F, Lang AG. G*Power. URL: https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower [accessed 2025-03-21]
35. Keulers BJ, Welters CFM, Spauwen PHM, Houpt P. Can face-to-face patient education be replaced by computer-based patient education? A randomised trial. Patient Educ Couns 2007 Jul;67(1-2):176-182. [doi: 10.1016/j.pec.2007.03.012] [Medline: 17448621]
36. Multanen J, Ylinen J, Karjalainen T, Kautiainen H, Repo JP, Häkkinen A. Reliability and validity of the Finnish version of the Boston carpal tunnel questionnaire among surgically treated carpal tunnel syndrome patients. Scand J Surg 2020 Dec;109(4):343-350. [doi: 10.1177/1457496919851607] [Medline: 31132964]
37. Lenhard W, Lenhard A. Computation of effect sizes. 2016. [doi: 10.13140/RG.2.2.17823.92329]
38. Minge. meCue Data Template. 2013. URL: http://mecue.de/Homepage%20Content/english/meCUE_data_template.xlsx [accessed 2023-04-12]
39. Eysenbach G. Improving the quality of Web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34. [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
40. Cook JA, Sasor SE, Tholpady SS, Momeni A, Chu MW. Hand surgery resources exceed American health literacy. Hand (N Y) 2018 Sep;13(5):547-551. [doi: 10.1177/1558944717725384] [Medline: 28825342]
41. Goyal R, Mercado AE, Ring D, Crijns TJ. Most YouTube videos about carpal tunnel syndrome have the potential to reinforce misconceptions. Clin Orthop Relat Res 2021 Oct 1;479(10):2296-2302. [doi: 10.1097/CORR.0000000000001773] [Medline: 33847604]
42. Lutsky K, Bernstein J, Beredjiklian P. Quality of information on the internet about carpal tunnel syndrome: an update. Orthopedics 2013 Aug;36(8):e1038-e1041. [doi: 10.3928/01477447-20130724-20] [Medline: 23937750]
43. Steimle J, Gabriel S, Tarr R, Kohrs B, Johnston P, Martineau D. Comparing diagnostic and treatment recommendations of carpal tunnel syndrome available on the internet with AAOS clinical practice guidelines. Hand (N Y) 2020 Jul;15(4):514-520. [doi: 10.1177/1558944718821417] [Medline: 30654676]
44. Özbek EA, Armangil M, Karaca MO, Merter A, Dursun M, Kocaoğlu H. Evaluation of the reliability and quality of information in carpal tunnel syndrome shared on YouTube. J Wrist Surg 2022 Aug;11(4):295-301. [doi: 10.1055/s-0041-1735231] [Medline: 35971474]
45. Fang Y, Baker NA, Dole J, Roll SC. Quality of carpal tunnel syndrome patient education handouts available on the internet: a systematic analysis of content and design. Arch Phys Med Rehabil 2022 Feb;103(2):297-304. [doi: 10.1016/j.apmr.2021.08.010] [Medline: 34547274]
46. Frueh FS, Palma AF, Raptis DA, Graf CP, Giovanoli P, Calcagni M. Carpal tunnel syndrome: analysis of online patient information with the EQIP tool. Chir Main 2015 Jun;34(3):113-121. [doi: 10.1016/j.main.2015.04.003] [Medline: 26022522]
47. Frické M, Fallis D, Jones M, Luszko GM. Consumer health information on the internet about carpal tunnel syndrome: indicators of accuracy. Am J Med 2005 Feb;118(2):168-174. [doi: 10.1016/j.amjmed.2004.04.032] [Medline: 15694903]
48. Beredjiklian PK, Bozentka DJ, Steinberg DR, Bernstein J. Evaluating the source and content of orthopaedic information on the internet. The case of carpal tunnel syndrome. J Bone Joint Surg Am 2000 Nov;82(11):1540-1543. [doi: 10.2106/00004623-200011000-00004] [Medline: 11097441]
49. Kwak D, Park JW, Won Y, Kwon Y, Lee JI. Quality and reliability evaluation of online videos on carpal tunnel syndrome: a YouTube video-based study. BMJ Open 2022 Apr 15;12(4):e059239. [doi: 10.1136/bmjopen-2021-059239] [Medline: 35428646]
50. Johnson AR, Doval AF, Granoff MD, et al. A comparative multimetric assessment of English and Spanish carpal tunnel syndrome materials. J Surg Res 2019 Jun;238:64-71. [doi: 10.1016/j.jss.2019.01.032] [Medline: 30739070]
51. Eberlin KR, Vargas CR, Chuang DJ, Lee BT. Patient education for carpal tunnel syndrome: analysis of readability. Hand (N Y) 2015 Sep;10(3):374-380. [doi: 10.1007/s11552-014-9718-7] [Medline: 26330766]
52. Mert A, Bozgeyik B. Quality and content analysis of carpal tunnel videos on YouTube. Indian J Orthop 2022 Jan;56(1):73-78. [doi: 10.1007/s43465-021-00430-5] [Medline: 35070145]
53. Sproule JA, Tansey C, Burns B, Fenelon G. The web: friend or foe of the hand surgeon? Hand Surg 2003 Dec;8(2):181-185. [doi: 10.1142/s0218810403001716] [Medline: 15002095]
54. Ozdemir O, Diren F, Boyali O, Civelek E, Kabatas S. Metric evaluation of reliability and transparency of the videos about carpal tunnel syndrome surgery in the online platforms: assessment of YouTube videos' content. Neurospine 2021 Jun;18(2):363-368. [doi: 10.14245/ns.2142030.015] [Medline: 34218617]

XSL•FO
RenderX

55. Park KE, Allam O, Kim S, Prsic A. Accuracy of online pictorial resources for common hand surgery procedures. Hand (N Y) 2022 Mar;17(2):361-365. [doi: 10.1177/1558944720937360] [Medline: 32713204]
56. Akinleye SD, Garofolo-Gonzalez G, Montuori M, Culbertson MD, Hashem J, Edelstein DM. Readability of the most commonly accessed online patient education materials pertaining to pathology of the hand. Hand (N Y) 2018 Nov;13(6):705-714. [doi: 10.1177/1558944717726138] [Medline: 28836453]
57. Strouse S, Rosenberg S. Health literacy and discharge education. MSJ 2022;31(6):402 [FREE Full text] [doi: 10.62116/MSJ.2022.31.6.402]
58. Pekonen A, Eloranta S, Stolt M, Virolainen P, Leino-Kilpi H. Measuring patient empowerment - a systematic review. Patient Educ Couns 2020 Apr;103(4):777-787. [doi: 10.1016/j.pec.2019.10.019] [Medline: 31767243]
59. Osborne RH, Elsworth GR, Whitfield K. The Health Education Impact Questionnaire (heiQ): an outcomes and evaluation measure for patient education and self-management interventions for people with chronic conditions. Patient Educ Couns 2007 May;66(2):192-201. [doi: 10.1016/j.pec.2006.12.002] [Medline: 17320338]
60. Nagl M, Ullrich A, Farin E. Verständlichkeit von Patientenschulungen in der orthopädischen Rehabilitation: Qualitative Erhebung bei Rehabilitanden und Schulungsleitern. Rehabilitation (Bonn) 2012;52(1):34-39. [doi: 10.1055/s-0032-1312664]
61. Akbolat M, Amarat M, Ünal Ö, Şantaş G. A survey of health information seeking by cancer patients indicates some problems over medical explanations and terminology. Health Info Libraries J 2023 Mar;40(1):29-41. [doi: 10.1111/hir.12387]
62. Cosic F, Kimmel L, Edwards E. Patient comprehension of common orthopedic terminology. Health Lit Res Pract 2019 Jul;3(3):e187-e193. [doi: 10.3928/24748307-20190624-04] [Medline: 31428736]
63. Watermeyer J, Thwala Z, Beukes J. Medical terminology in intercultural health interactions. Health Commun 2021 Aug;36(9):1115-1124. [doi: 10.1080/10410236.2020.1735700] [Medline: 32202159]
64. Derevianchenko N, Lytovska O, Diurba D, Leshchyna I. Impact of medical terminology on patients' comprehension of healthcare. Georgian Med News 2018 Nov(284):159-163. [Medline: 30618411]
65. Finkelstein J, Bedra M. Is Internet search better than structured instruction for web-based health education? Stud Health Technol Inform 2013;190:65-67. [Medline: 23823377]
66. Castro H, Hise M, Finkelstein J. A comparison of two models of web-based education in older adults. AMIA Annu Symp Proc 2005;2005:914. [Medline: 16779201]
67. Crondahl K, Eklund Karlsson L. The nexus between health literacy and empowerment. SAGE Open 2016 Apr 1;6(2):215824401664641. [doi: 10.1177/2158244016646410]
68. Stars I. Health literacy as a challenge for health education. SHS Web Conf 2018;40:02004. [doi: 10.1051/shsconf/20184002004]
69. Gellert P, Tille F. What do we know so far? The role of health knowledge within theo-ries of health literacy. Eur Health Psychol 2015;16:266-274 [FREE Full text]
70. Sørensen K, Van den Broucke S, Fullam J, et al. Health literacy and public health: a systematic review and integration of definitions and models. BMC Public Health 2012 Jan 25;12(1):80. [doi: 10.1186/1471-2458-12-80] [Medline: 22276600]
71. Yim CK, Shumate L, Barnett SH, Leitman IM. Health literacy assessment and patient satisfaction in surgical practice. Ann Med Surg (Lond) 2018 Nov;35:25-28. [doi: 10.1016/j.amsu.2018.08.022] [Medline: 30263114]
72. Kandula NR, Nsiah-Kumi PA, Makoul G, et al. The relationship between health literacy and knowledge improvement after a multimedia type 2 diabetes education program. Patient Educ Couns 2009 Jun;75(3):321-327. [doi: 10.1016/j.pec.2009.04.001] [Medline: 19395223]
73. Davaris MT, Bunzli S, Trieu J, Dowsey MM, Choong PF. The role of digital health interventions to improve health literacy in surgical patients: a narrative review in arthroplasty. ANZ J Surg 2022 Oct;92(10):2474-2486. [doi: 10.1111/ans.17931] [Medline: 35924880]
74. Eltorai AEM, Ghanian S, Adams CA Jr, Born CT, Daniels AH. Readability of patient education materials on the american association for surgery of trauma website. Arch Trauma Res 2014 Jun;3(2):e18161. [doi: 10.5812/atr.18161] [Medline: 25147778]
75. Man A, van Ballegooie C. Assessment of the readability of web-based patient education material from major Canadian pediatric associations: cross-sectional study. JMIR Pediatr Parent 2022 Mar 16;5(1):e31820. [doi: 10.2196/31820] [Medline: 35293875]
76. Wasir AS, Volgman AS, Jolly M. Assessing readability and comprehension of web-based patient education materials by American Heart Association (AHA) and CardioSmart online platform by American College of Cardiology (ACC): How useful are these websites for patient understanding? Am Heart J Plus 2023 Aug;32:100308. [doi: 10.1016/j.ahjo.2023.100308] [Medline: 38510202]
77. Bühn S, Huppertz E, Weise A, et al. The effects of modifying elements of written informed consent forms for elective surgical or invasive procedures: a systematic review. Patient Educ Couns 2023 Feb;107:107576. [doi: 10.1016/j.pec.2022.107576] [Medline: 36455317]
78. Pennekamp P, Diedrich O, Schmitt O, Kraft C. Prävalenz und Stellenwert der Internetnutzung orthopädischer Patienten. Z Orthop Ihre Grenzgeb 2006 Jan;144(05):459-463. [doi: 10.1055/s-2006-942237]
79. Hertling S, Matziolis G, Graul I. Die Rolle des Internets als medizinische Informationsquelle für orthopädische Patienten [Article in German]. Orthopädie 2022;51(7):521-530. [doi: 10.1007/s00132-022-04238-5]

XSL•FO
RenderX

80. Mack EM, Callinan NJ, Reams M, Bohn DC, Chmielewski TL. Patient-reported outcomes after open carpal tunnel release using a standard protocol with 1 hand therapy visit. J Hand Ther 2017 Jan;30(1):58-64. [doi: 10.1016/j.jht.2016.03.007]
81. Kim JK, Koh YD, Kim JO, Choi SW. Changes in clinical symptoms, functions, and the median nerve cross-sectional area at the carpal tunnel inlet after open carpal tunnel release. Clin Orthop Surg 2016 Sep;8(3):298-302. [doi: 10.4055/cios.2016.8.3.298] [Medline: 27583113]
82. Aloi NF, Rahman H, Fowler JR. Changes in cross-sectional area of the median nerve and Boston carpal tunnel questionnaire scores after carpal tunnel release. Hand (N Y) 2024 Mar;19(2):212-216. [doi: 10.1177/15589447221127336] [Medline: 36189893]
83. Nielsen J. How Many Test Users in a Usability Study? NN/g Nielsen Norman Group. 2012. URL: https://www.nngroup.com/articles/how-many-test-users/ [accessed 2020-04-12]

Abbreviations

APPLY: patient's ability to apply health-related information in an everyday life
BCTQ: Boston carpal tunnel questionnaire
CHERRIES: Checklist for Reporting Results of Internet E-Surveys
COHEP: Comprehensibility Of Health Education Programmes
COMM: patient's ability to communicate with health care professionals
COMPR: patient's comprehension of medical information
CTR: carpal tunnel release
CTS: carpal tunnel syndrome
HEL: health education literacy
HELP: Health Education Literacy of Patients with chronic musculoskeletal diseases
MANOVA: multivariate analysis of variances
PRISMA-ScR: Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews

Edited by A Mavragani; submitted 05.08.24; peer-reviewed by R Tedeschi, X Liu; revised version received 03.02.25; accepted 04.02.25; published 25.03.25.

Please cite as:
Seel M, Mihalic JA, Froschauer SM, Holzner B, Meier J, Gotterbarm T, Holzbauer M
Changes in Health Education Literacy After Structured Web-Based Education Versus Self-Directed Online Information Seeking in Patients Undergoing Carpal Tunnel Release Surgery: Nonrandomized, Controlled Study
JMIR Form Res 2025;9:e65114
URL: https://formative.jmir.org/2025/1/e65114
doi:10.2196/65114

© Mariella Seel, Julian Alexander Mihalic, Stefan Mathias Froschauer, Bernhard Holzner, Jens Meier, Tobias Gotterbarm, Matthias Holzbauer. Originally published in JMIR Formative Research (https://formative.jmir.org), 25.3.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Impact of ChatGPT Exposure on User Interactions With a Motivational Interviewing Chatbot: Quasi-Experimental Study

Jiading Zhu[1], BASc; Alec Dong[2*]; Cindy Wang[1*]; Scott Veldhuizen[3,4,5], PhD; Mohamed Abdelwahab[1], MSc; Andrew Brown[1], MASc; Peter Selby[3,5,6], MBBS, MHSc; Jonathan Rose[1,3], PhD

[1]The Edward S. Rogers Sr. Department of Electrical & Computer Engineering, University of Toronto, Toronto, ON, Canada
[2]Department of Mechanical & Industrial Engineering, University of Toronto, Toronto, ON, Canada
[3]INTREPID Lab, Centre for Addiction and Mental Health, Toronto, ON, Canada
[4]Institute for Mental Health Policy Research, Centre for Addiction and Mental Health, Toronto, ON, Canada
[5]Department of Family and Community Medicine, University of Toronto, Toronto, ON, Canada
[6]Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada
[*]these authors contributed equally

Corresponding Author:
Jonathan Rose, PhD
The Edward S. Rogers Sr. Department of Electrical & Computer Engineering, University of Toronto, Toronto, ON, Canada

Abstract

Background: The worldwide introduction of ChatGPT in November 2022 may have changed how its users perceive and interact with other chatbots. This possibility may confound the comparison of responses to pre-ChatGPT and post-ChatGPT iterations of pre-existing chatbots, in turn affecting the direction of their evolution. Before the release of ChatGPT, we created a therapeutic chatbot, MIBot, whose goal is to use motivational interviewing to guide smokers toward making the decision to quit smoking. We were concerned that measurements going forward would not be comparable to those in the past, impacting the evaluation of future changes to the chatbot.

Objective: The aim of the study is to explore changes in how users interact with MIBot after the release of ChatGPT and examine the relationship between these changes and users' familiarity with ChatGPT.

Methods: We compared user interactions with MIBot prior to ChatGPT's release and 6 months after the release. Participants (N=143) were recruited through a web-based platform in November of 2022, prior to the release of ChatGPT, to converse with MIBot, in an experiment we refer to as MIBot (version 5.2). In May 2023, a set of (n=129) different participants were recruited to interact with the same version of MIBot and asked additional questions about their familiarity with ChatGPT, in the experiment called MIBot (version 5.2A). We used the Mann-Whitney *U* test to compare metrics between cohorts and Spearman rank correlation to assess relationships between familiarity with ChatGPT and other metrics within the MIBot (version 5.2A) cohort.

Results: In total, 83(64.3%) participants in the MIBot (version 5.2A) cohort had used ChatGPT, with 66 (51.2%) using it on a regular basis. Satisfaction with MIBot was significantly lower in the post-ChatGPT cohort (U=11,331.0; P=.001), driven by a decrease in perceived empathy as measured by the Average Consultation and Relational Empathy Measure (U=10,838.0; P=.01). Familiarity with ChatGPT was positively correlated with average response length (ρ=0.181; P=.04) and change in perceived importance of quitting smoking (ρ=0.296; P<.001).

Conclusions: The widespread reach of ChatGPT has changed how users interact with MIBot. Post-ChatGPT users are less satisfied with MIBot overall, particularly in terms of perceived empathy. However, users with greater familiarity with ChatGPT provide longer responses and demonstrated a greater increase in their perceived importance of quitting smoking after a session with MIBot. These findings suggest the need for chatbot developers to adapt to evolving user expectations in the era of advanced generative artificial intelligence.

(JMIR Form Res 2025;9:e56973) doi:10.2196/56973

KEYWORDS

chatbot; digital health; motivational interviewing; natural language processing; ChatGPT; large language models; artificial intelligence; experimental; smoking cessation; conversational agent

XSL•FO
RenderX

Introduction

Background

Generative chatbots are conversational systems that provide synthesized replies using deep learning techniques [1]. In recent years, generative chatbots based on large language models (LLMs) have made major advancements in their ability to engage in natural and human-like conversations [2]. ChatGPT, a popular LLM-based generative chatbot developed by OpenAI [3], has demonstrated significant potential to be applied in public health and medicine for a variety of purposes such as providing health information, supporting patient education, summarizing clinical notes, and assisting in administrative tasks [4,5]. However, these applications have not typically focused on therapeutic interactions, such as those using motivational interviewing (MI) [6] techniques.

Prior to ChatGPT's worldwide release in November 2022, our team developed MIBot [7], a therapeutic chatbot using MI techniques to guide smokers toward the decision to quit smoking. MIBot has a structured conversation flow, using scripted questions and LLM-generated MI-style reflections [6]. However, the popularity of ChatGPT and its advanced conversational capabilities raise questions about its influence on user interactions and expectations with simpler chatbots such as MIBot. These potential influences may in turn affect our future experiments with newer versions of MIBot, as the measurements used in future versions of MIBot may not be comparable to those in past versions, due to ChatGPT's potential influences on participants.

Related Work

Several recent studies have delved into the realm of generative chatbots, used in a medical or therapeutic context, and explored their effect on users. Perski et al [8] quantified the effect of the addition of a supportive chatbot to their smoking cessation application and found that it has increased user engagement and resulted in higher rates of quit success. Boucher et al [9] provided a comprehensive review of artificially intelligent chatbots in digital mental health interventions, discussing their acceptability, effects on user engagement and mental health outcomes, as well as their weaknesses and risks at the time, such as language proficiency and understanding. Moilanen et al [10] examined the effect of personality traits of a mental health chatbot on user engagement, finding that chatbots with a conscientious personality elicit the most user engagement and that users prefer their chatbot to be informative and confident rather than monotonic. Chow et al [11] framed ChatGPT as a disruptive technology and explored its impact on medical chatbots, claiming that it has the potential to improve access to health care services while expressing concerns over factors of risks such as reliability, transparency, privacy, and bias.

Together, these findings illustrate the growing influence of generative chatbots on user engagement and expectations in health-related contexts, particularly as advanced chatbots like ChatGPT redefine conversational norms. However, the impact of ChatGPT's widespread use on users' perceptions and behaviors toward other chatbots remains underexplored, raising questions about how exposure to such an advanced technology shifts user expectations and affects user behaviors. These findings served as important inspirations and motivations for this study, where we specifically investigate how users of one chatbot (MIBot) would change after being exposed to another chatbot (ChatGPT).

Study Objective

This study aims to investigate how user interactions with MIBot have changed following the release of ChatGPT and examines the relationship between familiarity with ChatGPT and these changes. By comparing dialogue sessions with MIBot conducted before and after ChatGPT's release, this study analyzes metrics such as average response length, satisfaction ratings, and effectiveness of the therapy. Additionally, we assess the relationship between participants' familiarity with ChatGPT and these metrics to better understand the impact of ChatGPT on user behaviors and expectations. Findings from this study are expected to provide insights into how user expectations have evolved since the release of ChatGPT and inform future development strategies for chatbots in health-related applications.

Methods

Experiments

Overview

An experiment was conducted on November 18, 2022, about 2 weeks before the worldwide introduction of ChatGPT by OpenAI [3]. We then repeated that experiment on May 16, 2023, about half a year after ChatGPT was released. The 2 experiments used the exact same version of MIBot [7] but on different groups of independently recruited participants. From this part on, we will refer to the first experiment as MIBot (version 5.2) and the second experiment as MIBot (version 5.2A).

Overall Experiment Flow

The overall experiment flow strictly followed what was detailed in the original MIBot paper [7]. Here is a brief overview:

1. Recruitment: Participants were recruited through the Prolific [12] paid web-based recruitment system after providing consent.
2. Preconversation surveys: Participants filled out surveys about their smoking habits on a custom website, including metrics used in the filtering process.
3. Conversation initiation: MIBot initiated with a text chat about smoking, continuing only with participant consent.
4. Core conversation: The main chatbot conversation about smoking cessation, which contained prescripted questions, responses from the user, and then generated reflections from the chatbot, described in more detail in the Design of MIBot section.
5. Postconversation: Participants completed another readiness-to-quit survey, the Consultation and Relational Empathy (CARE) Measure [13], and additional qualitative questions.
6. Reporting: Completion of tasks was recorded in the Prolific system [12].

XSL•FO
RenderX

7. One-week-later survey: A week later, participants answered a follow-up survey, also including readiness-to-quit, the completion of which, along with passing a manual data quality review, was required for the participants to receive their payments.

Design of MIBot

MIBot asked 2 types of questions. We call the first type "main questions" (Textbox 1) and the second type "yes or no questions" (Textbox 2). The main questions were open-ended questions that promoted self-reflection within users about their smoking habits, while yes or no questions were close-ended questions posed after each generated reflection to find out whether or not the generated reflection made sense, was on topic, and was used as a transition to the next main question.

Textbox 1. Main questions in MIBot conversations.

1. To start, what is the thing you like most about smoking?
2. What else do you like about smoking?
3. Now, what is the thing you like least about smoking?
4. What else do you dislike about smoking?
5. Now, what is one thing about your smoking that you would like to change?
6. What will it look like when you have made this change in your smoking addiction?
7. Finally, what are the steps you need to take to make this change?

Textbox 2. Yes or no questions in MIBot conversations.

1. Did that make sense?
2. Did what I said make sense to you?
3. Does this make sense to you?

The overall conversation structure was fixed as follows:

1. Introductory statement about MIBot and asking for permission to talk about the participant's smoking habits
2. Five to seven repetitions of
   a. Main question
   b. User response
   c. Generated reflection
   d. Yes or no question
   e. User response
3. Conclusion and thanking the participant for their time

Smoking Status

The context of our study requires us to keep track of the smoking status of the participants, measured by the following three metrics:

1. User status: A label that denotes if a participant would have passed the screening from their preconversation survey imported from our previous MIBot studies [7]. If they have passed the screening, they are denoted as low confidence or discordant; if not, they are instead denoted as high confidence and not discordant.
2. Heaviness of Smoking Index [14]: A validated survey metric calculated from cigarettes per day and time to the first cigarette of the day. Heaviness of Smoking Index is an integer, where a higher value indicates a heavier smoking habit.
3. Quit attempts made: A binary value that denotes if a participant has made at least 1 quit attempt in the week leading up to the MIBot conversation.

MIBot Surveys

There were 3 surveys delivered to the participants from the original MIBot experiment flow: preconversation survey, postconversation survey, and 1-week-later survey. These surveys were used to determine how effective MIBot was in motivating smokers to make the decision to quit smoking as well as how participants felt about the conversation with MIBot.

The relevant parts of the survey for this study are the following:

1. The CARE Measure [13] consists of questions regarding the participant's satisfaction with the conversation. Each question asks the participant to rate how well they think MIBot did on a scale from 1 to 5 or 0 if it does not apply. For example, participants were asked to evaluate MIBot's ability to "make you feel at ease" or "letting you tell your 'story'" back when they were interacting with MIBot. The CARE Measure is only included in the postconversation survey.
2. The Readiness Ruler [15] measures how ready the participant is to quit smoking. It consists of 3 ratings on a 0 - 10 scale: how confident they are about quitting smoking now, how important it is for them to quit smoking, and how ready they feel to quit smoking now. The Readiness Ruler is included in all 3 surveys.
3. Two additional feedback questions, included only in the postconversation survey:
   a. "What are 3 words that you would use to describe the chatbot?"
   b. "What would you change about the conversation?"

XSL•FO
RenderX

ChatGPT Survey

To determine the extent of exposure to ChatGPT, for each participant in MIBot (version 5.2A), we included an additional short survey in the 1-week-later survey referred to as the ChatGPT survey. It contained 8 new questions designed to evaluate the participant's knowledge and use of ChatGPT prior to engaging in MIBot (version 5.2A). The full ChatGPT survey can be found in Multimedia Appendix 1.

Recruitment and Data Inclusion

The steps taken for recruiting the participants for both MIBot (version 5.2) and MIBot (version 5.2A) were the exact same as detailed in the original MIBot paper [7]. Participants were recruited through the Prolific [12] web-based recruitment system, where they were informed by a recruitment description that they would engage in a text-based conversation with a chatbot designed to promote readiness to quit smoking, provide feedback on their experience, and complete the 1-week-later survey. The study was framed as an opportunity to contribute to research on chatbot-assisted smoking cessation, emphasized the confidentiality of participant data, and clearly stated that participation was voluntary, with the option to opt out if they did not agree to the terms. The entire study was delivered to the participants remotely, and they could participate in the study using their personal computers or mobile devices.

Notably, participants were screened using Prolific's filters based on the following inclusion criteria: participants could be located in any country, were at least 18 years of age, were fluent in English, had a smoking status defined as either a current smoker (smoking at least 5 cigarettes a day for at least 1 year) or a recent smoker (smoking at least 5 cigarettes a day for less than 1 year), and had a minimum approval rate of 90% on their prior Prolific studies. Additionally, Prolific was set to recruit an equal number of male and female participants. However, due to additional screening conducted after recruitment, the final sample was not perfectly balanced by sex.

The data collected were manually reviewed for data inclusion following the same criteria as listed in the MIBot paper [7] except for 1 major difference. In the original MIBot study, we only included participants with either of the following qualities in their Readiness Ruler scores:

- Low confidence: confidence level less than or equal to 5
- Discordant: importance level more than 5 points below the confidence level

As the focus of this study is instead about ChatGPT's influence on participants, we decided to include as many valid data entries as possible. Therefore, this filtering was not conducted in this study. As a result, the number of participants we included in this study increased from 100 to 143.

Evaluation Metrics

To evaluate MIBot conversations and user behaviors from various aspects, we designed several evaluation metrics, for which detailed definitions are provided below.

Response Length

In this study, the key aspect of user behavior we focused on was the length of their responses to questions prompted by MIBot. Response length is an important indicator of user engagement and willingness to interact with the chatbot. Longer responses suggest that participants are more actively reflecting on the conversation, which aligns with the therapeutic goals of MI sessions.

We defined response length by its word count, excluding any punctuation. We also categorized the questions MIBot asks into 2 categories and looked at the user responses separately to gain more insight. As a result, there are three metrics we applied to measure the lengths of participant responses:

1. Average response length: average length of responses to all questions.
2. Average response length (main): average length of responses to main questions, which are the scripted questions MIBot asks to provoke contemplation of the users' smoking habits.
3. Average response length (yes or no): average length of responses to yes or no questions, which are shown in Textbox 2.

Satisfaction With MIBot

Satisfaction with MIBot is a metric designed to measure user satisfaction after participating in a conversation with MIBot, a value between 0 and 1, calculated as the mean of the following three quantities:

1. Average CARE Measure (integers ranging from 1 to 5, in order of increasing satisfaction with MIBot), excluding answers where does not apply was selected. The mean is then rescaled to a number between 0 and 1 by subtracting 1 from the original value and dividing the result by 4.
2. Feedback sentiment score, based on answers to the first feedback question, "What are 3 words that you would use to describe the chatbot?" (integers ranging from 1 to 5, in order of increasing positivity). These scores were generated by automatically measuring the sentiment of the 3 words given by the user as feedback, described in the MIBot Survey section. The sentiment is computed using the neural network bert-base-multilingual-uncased-sentiment [16], which provides a score between 0 and 1.
3. MIBot improvement indicator, a binary value (0 or 1) if the participant indicated (in written feedback) that they would like MIBot improvement in response to the second feedback question, "What would you change about the conversation?" Responses are individually checked and manually labeled by a human collaborator. Since participants who suggested improvements are assumed to be less satisfied with MIBot, the value (1−MIBot improvement indicator) is used in the averaging calculation to get satisfaction with MIBot.

Effectiveness of MIBot

The goal of MIBot is to guide participants toward the decision to quit smoking by provoking contemplation. The more effective the conversation was, the more change we should observe in a participant. The effectiveness of MIBot is measured using 3

XSL•FO
RenderX

metrics related to the Readiness Ruler, as described in the MIBot Surveys section. The first of which, confidence change, is calculated as the difference between the confidence level before the conversation and the confidence level measured 1 week later, as recorded on the corresponding Readiness Rulers filled by the participants. Similarly, the importance change and readiness change were determined. As noted in the original MIBot paper [7], the primary outcome is the confidence change, as this is the most predictive of smoking cessation success [17-19].

Familiarity With ChatGPT

Familiarity with ChatGPT was given as a score between 0 and 126 (inclusive) to each participant based on their answers to the ChatGPT survey. Participants were first given a starting score of either 0 or 1: those who had not heard of ChatGPT were given a score of 0, and those who had were given a score of 1. Among those who had heard of ChatGPT, the score of those who had never used ChatGPT stayed at 1; for those who had used ChatGPT on top of knowing about it, they received an additional score according to the answers they had provided for 3 questions in the ChatGPT survey regarding their past use of ChatGPT. The three questions are as follows: (1) How long ago did you start using ChatGPT? (2) How frequently do you use ChatGPT? (3) On average, how long do your ChatGPT sessions last?

These 3 questions were presented as multiple-choice questions, the answers to which are presented in an ordered way and assigned a corresponding integer score from 1 to 5, as presented in Table 1. The final familiarity score they received was the product of the 3 integer scores, in addition to the 1 they already received. As a result, familiarity scores ranged from 0 to 126, with 0 being having never heard of ChatGPT, 1 being having heard of but never used ChatGPT, 2 being having used ChatGPT once a month or less for less than 5 minutes each session starting from less than a week ago, and 126 being having used ChatGPT multiple times per day with each session lasting longer than an hour for the past 5 months.

Table . Corresponding description for familiarity score assignment.

Corresponding score	Duration of use	Frequency of use	Length of each session
1	Less than a week	Once a month or less	Less than 5 minutes
2	Less than a month	Every 2 weeks	Less than 10 minutes
3	1 - 3 months	Every week	10 to 30 minutes
4	3 - 5 months	Every day	30 minutes to 1 hour
5	Over 5 months	Multiple times per day	More than 1 hour

Statistical Analysis

Comparison of Pre- and Post-ChatGPT Cohorts

To capture the differences in behavior between participants in the 2 cohorts—November 2022 (MI version 5.2) and May 2023 (MI version 5.2A)—we used the Mann-Whitney *U* test to compare various metrics, such as average response length, satisfaction with MIBot, and changes in Readiness Ruler scores. This nonparametric test provided test statistics and *P* values for each metric, which are used to determine significant differences between the 2 groups. The test was done using the Python programming language (Python Software Foundation) with the pandas [20] and SciPy [21] libraries.

Relationships Between Familiarity With ChatGPT and Other Variables

To examine the impact of familiarity with ChatGPT within the MIBot (version 5.2A) cohort, we performed Spearman rank correlation analysis. This test assessed the relationships between participants' familiarity scores (ranging from 0 to 126) and key metrics, such as average response length, satisfaction with MIBot, and changes in Readiness Ruler scores. Spearman rank correlation provided correlation coefficients and *P* values, indicating the strength and significance of each relationship. The test was also done using Python with the pandas [20] and SciPy [21] libraries.

Ethical Considerations

This research was approved by the University of Toronto Research Ethics Board (protocol # 35567), as amended on June 29, 2022. Participants voluntarily provided informed consent by agreeing to all the terms stated in a consent form presented digitally to all participants during recruitment, fully outlining the study's goals, procedures, potential risks, and privacy guarantees. It specified that no personally identifiable information would be collected, and any inadvertent identifiers would be removed prior to analysis or publication. Participants received a total of US $6.25 for MIBot (version 5.2) and US $6.18 for MIBot (version 5.2A) as compensation for completing all tasks in the study, and partial compensation was not provided to those who did not complete all components of the experiment. No identifiable participant information is included in any images or supplementary materials in the manuscript.

Results

Participant Demographic

Participant demographic data of both MIBot (version 5.2) and MIBot (version 5.2A) are listed in Tables 2 and 3, including categorical and continuous variables, respectively. Additional data of the demographics of the participants in both MIBot (version 5.2) and MIBot (version 5.2A) are listed in Multimedia Appendix 2.

XSL•FO
RenderX

Table . Categorical demographic data of participants in cohort MIBot (version 5.2; November 2022, pre-ChatGPT) and MIBot (version 5.2A; May 2023, post-ChatGPT).

Characteristic		MIBot (version 5.2) (n=143), n (%)	MIBot (version 5.2A) (n=129), n (%)
Sex			
	Male	72 (50.3)	63 (48.8)
	Female	71 (49.7)	66 (51.2)
User status			
	Low confidence or discordant	100 (69.9)	91 (70.5)
	High confidence and not discordant	43 (30.1)	38 (29.5)
Smoking status			
	I am a current smoker (smoke at least 5 cigarettes a day and have smoked this amount for at least 1 year)	121 (84.6)	107 (82.9)
	I am a recent smoker (smoke at least 5 cigarettes a day and have smoked this amount for less than 1 year)	22 (15.4)	22 (17.1)
Quit attempts made			
	Yes	69 (48.3)	46 (35.7)
	No	74 (51.7)	83 (64.3)

Table . Continuous demographic data of participants in cohort MIBot (version 5.2; November 2022, pre-ChatGPT) and MIBot (version 5.2A; May 2023, post-ChatGPT).

Characteristic	MIBot (version 5.2)		MIBot (version 5.2A)	
	Mean (SD)	Median (IQR)	Mean (SD)	Median (IQR)
Age (years)	29.22 (9.69)	26.00 (23.00 - 33.00)	32.75 (11.36)	29.00 (24.00 - 39.00)
HSI[a]	1.45 (1.38)	1.00 (0.00 - 2.00)	1.67 (1.45)	1.00 (0.00 - 3.00)
Readiness Rulers, preconversation				
Confidence	4.52 (2.75)	4.00 (2.00 - 7.00)	4.12 (2.60)	4.00 (2.00 - 6.00)
Importance	5.84 (2.84)	6.00 (3.00 - 8.00)	5.60 (2.65)	6.00 (3.00 - 8.00)
Readiness	5.36 (2.83)	5.00 (3.00 - 8.00)	4.79 (2.58)	5.00 (3.00 - 7.00)

[a]HSI: Heaviness of Smoking Index.

Reach of ChatGPT

The familiarity score, described in the Familiarity With ChatGPT section, provides an estimate of the reach of ChatGPT to the public. Table 4 provides a more detailed description of the distribution of the familiarity score, where we bin the participants based on familiarity score ranges and assign labels to each score range. “Unexposed” is the group of participants who have not heard of ChatGPT prior to this study. “Aware” denotes that the participants have heard of ChatGPT but did not use it in any capacity. People who have used ChatGPT are separated into 3 groups, which are “casual user,” “consistent user,” and “dedicated user,” based on their familiarity score.

Table . Distribution of ChatGPT familiarity levels among participants in the MIBot (version 5.2A) cohort.

ChatGPT familiarity level	Count, n (%)	Score
Unexposed	12 (9.3)	0
Aware	34 (26.4)	1
Casual user	17 (13.2)	2 to 8
Consistent user	37 (28.7)	9 to 27
Dedicated user	29 (22.5)	28 or higher

XSL•FO
RenderX

Comparison of Pre- and Post-ChatGPT Cohorts

Table 5 shows the Mann-Whitney *U* test results that compare between participants from the November 2022 (MI version 5.2) and May 2023 (MI version 5.2A) cohorts. In terms of average response length, there was no significant difference between the 2 cohorts, whether looking at all responses together or breaking them down into responses to main questions and responses to yes or no questions. Satisfaction with MIBot was significantly higher in MIBot (version 5.2) than in MIBot (version 5.2A; *U*=11,331.0; *P*=.001). Specifically, of the 3 metrics satisfaction with MIBot was aggregated of, Average CARE Measure was the only one showing any significant difference, with it being significantly higher in MIBot (version 5.2) compared to MIBot (version 5.2A; *U*=10,838; *P*=.01). As for changes in the Readiness Ruler, no significant differences were found between the cohorts in terms of confidence change, importance change, or readiness change.

Table . Results of the Mann-Whitney *U* test comparing key metrics between cohort MIBot (version 5.2; pre-ChatGPT) and MIBot (version 5.2A; post-ChatGPT).

Variable		MIBot (version 5.2)		MIBot (version 5.2A)		Test statistic (*U*)	*P* value
		Mean (SD)	Median (IQR)	Mean (SD)	Median (IQR)		
Response length							
	Average response length	4.98 (2.49)	4.46 (3.31 - 5.89)	5.36 (3.54)	4.29 (3.13 - 7.00)	9248.0	.97
	Average response length (main)	8.18 (4.00)	7.60 (5.50 - 9.69)	8.04 (4.71)	6.62 (4.86 - 10.43)	9936.5	.27
	Average response length (yes or no)	2.02 (1.91)	1.00 (1.00 - 2.07)	2.44 (3.04)	1.17 (1.00 - 2.80)	8301.5	.13
Satisfaction scores							
	Satisfaction with MIBot	0.74 (0.22)	0.85 (0.52 - 0.91)	0.66 (0.21)	0.61 (0.49 - 0.85)	11,331.0	.001
	Average CARE[a] Measure	0.77 (0.16)	0.78 (0.66 - 0.90)	0.72 (0.17)	0.74 (0.58 - 0.84)	10,838.0	.01
	MIBot improvement indicator	0.57 (1.01)	0.00 (0.00 - 1.00)	0.76 (1.12)	0.00 (0.00 - 1.00)	8192.5	.06
	Feedback sentiment score	0.79 (0.15)	0.80 (0.80 - 0.80)	0.76 (0.18)	0.80 (0.80 - 0.80)	10,028.0	.15
Readiness Ruler							
	Confidence change	0.86 (2.32)	1.00 (0.00 - 2.00)	0.60 (2.08)	0.00 (0.00 - 2.00)	9893.0	.29
	Importance change	0.69 (1.85)	0.00 (0.00 - 2.00)	0.40 (1.98)	0.00 (0.00 - 1.00)	10,003.5	.21
	Readiness change	0.35 (1.82)	0.00 (−1.00 to 1.00)	0.33 (2.07)	0.00 (−1.00 to 1.00)	9260.0	.95

[a]CARE: Consultation and Relational Empathy.

Relationships Between Familiarity With ChatGPT and Other Variables

Table 6 shows the Spearman rank correlation results evaluating relationships between familiarity with ChatGPT and various other variables within the MIBot (version 5.2A) cohort. In terms of average response length, there was a significant positive correlation between familiarity with ChatGPT and overall average response length (ρ=0.181; *P*=.04) as well as with average response length for main questions (ρ=0.180; *P*=.04) and average response length for yes or no questions (ρ=0.197; *P*=.03). Satisfaction with MIBot had a weak negative correlation with familiarity score, but it was not statistically significant (ρ=−0.171; *P*=.05). Of the 3 metrics aggregated into satisfaction with MIBot, only the MIBot improvement indicator showed a significant positive correlation with familiarity with ChatGPT (ρ=0.188; *P*=.03). Among the Readiness Ruler changes, a significant positive correlation was observed between familiarity with ChatGPT and importance change (ρ=0.296; *P*<.001) but not for confidence change or readiness change.

XSL•FO
RenderX

Table . Spearman rank correlation analysis for the MIBot (version 5.2A) cohort.

Variable		Familiarity with ChatGPT	
		ρ	*P* value
Response length			
	Average response length	0.181	.04
	Average response length (main)	0.180	.04
	Average response length (yes or no)	0.197	.03
Satisfaction scores			
	Satisfaction with MIBot	−0.171	.05
	Average CARE[a] Measure	−0.112	.21
	MIBot improvement indicator	0.188	.03
	Feedback sentiment score	−0.023	.80
Readiness Ruler			
	Confidence change	0.165	.06
	Importance change	0.296	<.001
	Readiness change	0.078	.38

[a]CARE: Consultation and Relational Empathy.

Discussion

Principal Findings

This study aimed to investigate how user interactions with MIBot changed following the release of ChatGPT as well as the relationship between ChatGPT familiarity and these changes. From the results, we found that users interacting with MIBot after the release of ChatGPT were less satisfied with the chatbot, particularly in terms of perceived empathy, but users more familiar with ChatGPT provided longer responses and showed greater increase in their perceived importance of quitting smoking.

The premise of this study is that people have been exposed to ChatGPT due to its high popularity, which is supported by the distribution of familiarity scores in the MIBot (version 5.2A) cohort. From Table 4, ChatGPT is observed to have a widespread reach, with only 12 (9.3%) participants of MIBot (version 5.2A) having never heard of it, and 34 (26.4%) having only learned about it but never used it. Collectively, 83 (64.3%) participants have used ChatGPT at least to some capacity, most of whom are consistent or even dedicated users, indicating that the public has received massive exposure to the popular chatbot.

Using the Mann-Whitney *U* test (Table 5) to compare participants from the MIBot (version 5.2) and MIBot (version 5.2A) cohorts, we observed that most variables did not show significant differences between the groups. However, satisfaction with MIBot was significantly lower in the MIBot (version 5.2A) cohort (U=11,331.0; P=.001), a finding primarily driven by differences in the Average CARE Measure. This specific component of the satisfaction score, which reflects perceived empathy in interactions, was significantly higher in the MIBot (version 5.2) cohort (U=10,838.0; P=.01). These findings suggest that participants in the post-ChatGPT cohort, many of whom were familiar with or at least aware of ChatGPT's advanced conversational capabilities, may have developed higher expectations for chatbot empathy, which MIBot may not fully meet.

Within the MIBot (version 5.2A) cohort, Spearman rank correlation analysis revealed that familiarity with ChatGPT was positively correlated with average response length across all response types, indicating that more experienced users were likely to engage in longer interactions. However, while users who are more familiar with ChatGPT are more likely to suggest improvements for MIBot (ρ=0.188; P=.03), satisfaction with MIBot as a whole had only a weak, nonsignificant negative correlation with familiarity (ρ=−0.171; P=.05), even though the U test results showed that satisfaction was significantly lower in the MIBot (version 5.2A) cohort. This discrepancy suggests that while ChatGPT exposure may influence satisfaction overall, familiarity alone does not predict satisfaction levels. It may be that other factors unique to the MIBot (version 5.2A) cohort, such as general exposure to more advanced conversational artificial intelligence or broader trends in technology expectations, are impacting satisfaction with MIBot in a way that the specific familiarity score we designed failed to capture. In other words, the drop in satisfaction with MIBot might reflect a general shift in expectations for chatbot performance rather than an effect specific to individual familiarity with ChatGPT.

Interestingly, although familiarity with ChatGPT appears to make users more critical of MIBot's capabilities, it is also associated with an increase in participants' perceived importance of quitting smoking (ρ=0.296; P<.001), a key measure of MIBot's effectiveness. This seemingly paradoxical finding could suggest that while exposure to advanced generative chatbots raises expectations for technical quality, it might also enhance users' receptiveness to the broader goals of MIBot, such as helping users resolve ambivalence and guiding them

XSL•FO
RenderX

toward change, and as a result, making a less advanced chatbot more effective.

Practical Implications

The findings from this study suggest several practical implications for future chatbot development, especially in health-related applications. The decrease in satisfaction with MIBot from post-ChatGPT users highlights a need for researchers and developers to adapt to higher expectations of chatbots, especially for empathy. Additionally, the tendency of users familiar with ChatGPT to provide longer, more detailed responses suggests that chatbots should be capable of handling more nuanced interactions. Finally, the positive correlation between ChatGPT familiarity and MIBot's effectiveness in increasing the perceived importance of quitting smoking implies that users more experienced with advanced generative chatbots may have more trust in the technology and be more receptive to interactions guiding themselves toward change.

Comparison With Prior Works

Unlike prior studies that examined and reviewed various aspects of chatbots that may affect user engagement and effectiveness in different contexts [8-10], this study specifically and uniquely investigates how exposure to and familiarity with a more advanced chatbot (ChatGPT) affect user interactions with a simpler chatbot (MIBot). Building on to our previous work, MIBot [7], our new experiments suggest that MIBot continues to increase users' readiness to quit even after the introduction of ChatGPT. Nevertheless, overall user satisfaction has decreased, possibly affected by higher user expectations among those exposed to ChatGPT. This could be a disruptive effect from ChatGPT as predicted by Chow et al [11]. Moreover, our findings also provide unique insights not found in related prior works, such as that familiarity with ChatGPT positively correlates with core metrics for user engagement with and effectiveness of MIBot, a simpler chatbot than ChatGPT. It also presented a measurement, at a specific point in time, of the exposure level to ChatGPT among our recruited participants, which demonstrates the extent of ChatGPT's reach at that time.

Limitations

There are several limitations regarding this study. Since we followed the same experiment flow as the original MIBot paper [7] and used mostly the same evaluation metrics, the limitations of the MIBot experiments documented in the original MIBot paper [7] also apply to this study. Moreover, there exists an additional type of sampling bias specific to this study. Since for both of our experiments, we recruited participants through Prolific [12], a web-based recruitment system, it is possible that these participants are more fluent with technology than the general public and therefore more likely to have been exposed to ChatGPT or have the effects of ChatGPT manifest in a way that does not fairly represent a more general population. Furthermore, our 2 experiments used different groups of people with some demographic differences, notably with a difference in their average age, which may result in different smoking behaviors and success rates of smoking cessation [22,23]. Finally, it is important to acknowledge the possibility of confounding factors not accounted for in our study, which may have influenced how participants interacted with MIBot or perceived its effectiveness, along with familiarity with ChatGPT.

Conclusions

Given MIBot, a generative chatbot for smoking cessation less advanced than ChatGPT, this study aimed to find any potential changes to how users interact with MIBot by comparing user interactions before and after ChatGPT's release. We found that post-ChatGPT users are less satisfied with MIBot overall, particularly in terms of perceived empathy. Moreover, as users gain more familiarity with ChatGPT, they provide longer responses and show a greater increase in their perceived importance of quitting smoking. These findings suggest the need for continuous innovation in chatbot technology, particularly in digital health, to meet the evolving expectations of users accustomed to more advanced chatbot interactions.

Acknowledgments

The authors acknowledge their long-standing collaboration with the Intrepid Lab of Dr Peter Selby at the Centre for Addiction and Mental Health in Toronto and Professor Matt Ratto in the Faculty of Information at the University of Toronto. This research was funded by a Natural Sciences and Engineering Research Council of Canada Discovery grant (RGPIN-2019-04395) and a New Frontiers in Research Fund grant (NFRFE-2021-00449).

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

JZ and JR conceptualized this research work and oversaw project administration in terms of managing, planning, and coordinating the experiments. JZ, AD, CW, MA, and AB contributed to data curation and to developing the chatbot and the scripts. JZ, AD, CW, and SV contributed to conducting all the formal analyses. PS and JR secured funding for the project and provided overall supervision of all the research activities. JZ, AD, CW, and MA did all the investigation work related to the study. JZ, SV, AB, and JR contributed to developing the methodology. JR provided the computing and experimental resources for the study. JZ, AD, and CW contributed to the validation and visualization of the data collected. JZ, AD, CW, and JR were the writers of the original

XSL•FO
RenderX

draft, while JZ and JR critically reviewed and edited the manuscript. All authors made significant contributions to this work and approved it for publication.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Questions related to ChatGPT on the week-later survey.
[DOCX File, 511 KB - formative_v9i1e56973_app1.docx]

Multimedia Appendix 2
Additional demographics data.
[DOCX File, 20 KB - formative_v9i1e56973_app2.docx]

References

1. Adamopoulou E, Moussiades L. Chatbots: history, technology, and applications. Mach Learn Appl 2020 Dec;2:100006. [doi: 10.1016/j.mlwa.2020.100006]
2. Shanahan M, McDonell K, Reynolds L. Role play with large language models. Nature New Biol 2023 Nov;623(7987):493-498. [doi: 10.1038/s41586-023-06647-8] [Medline: 37938776]
3. Introducing ChatGPT. OpenAI. 2022. URL: https://openai.com/blog/chatgpt/ [accessed 2023-12-19]
4. Biswas SS. Role of Chat GPT in public health. Ann Biomed Eng 2023 May;51(5):868-869. [doi: 10.1007/s10439-023-03172-7] [Medline: 36920578]
5. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023 Aug;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
6. Miller W, Rollnick S. Motivational Interviewing: Helping People Change and Grow: The Guilford Press; 2023.
7. Brown A, Kumar AT, Melamed O, et al. A motivational interviewing chatbot with generative reflections for increasing readiness to quit smoking: iterative development study. JMIR Ment Health 2023 Oct 17;10:e49132. [doi: 10.2196/49132] [Medline: 37847539]
8. Perski O, Crane D, Beard E, Brown J. Does the addition of a supportive chatbot promote user engagement with a smoking cessation app? An experimental study [published correction appears in Digit Health. 2020 May 22;6:2055207620930958]. Digit Health 2019 Jan;5:2055207619880676. [doi: 10.1177/2055207619880676]
9. Boucher EM, Harake NR, Ward HE, et al. Artificially intelligent chatbots in digital mental health interventions: a review. Expert Rev Med Devices 2021 Dec;18(sup1):37-49. [doi: 10.1080/17434440.2021.2013200] [Medline: 34872429]
10. Moilanen J, Visuri A, Suryanarayana SA, Alorwu A, Yatani K, Hosio S. Measuring the effect of mental health chatbot personality on user engagement. Presented at: MUM 2022: 21th International Conference on Mobile and Ubiquitous Multimedia; Nov 27-30, 2022; Lisbon, Portugal. [doi: 10.1145/3568444.3568464]
11. Chow JCL, Sanders L, Li K. Impact of ChatGPT on medical chatbots as a disruptive technology. Front Artif Intell 2023;6. [doi: 10.3389/frai.2023.1166014]
12. Palan S, Schitter C. Prolific.ac—a subject pool for online experiments. J Behav Exp Finance 2018 Mar;17:22-27. [doi: 10.1016/j.jbef.2017.12.004]
13. Mercer SW, Maxwell M, Heaney D, Watt GC. The consultation and relational empathy (CARE) measure: development and preliminary validation and reliability of an empathy-based consultation process measure. Fam Pract 2004 Dec;21(6):699-705. [doi: 10.1093/fampra/cmh621.cmh621]
14. Heatherton TF, Kozlowski LT, Frecker RC, Rickert W, Robinson J. Measuring the heaviness of smoking: using self-reported time to the first cigarette of the day and number of cigarettes smoked per day. Br J Addict 1989 Jul;84(7):791-799. [doi: 10.1111/j.1360-0443.1989.tb03059.x] [Medline: 2758152]
15. Readiness Ruler. Center for Evidence-Based Practices | Case Western Reserve University. 2021. URL: https://case.edu/socialwork/centerforebp/resources/readiness-ruler [accessed 2023-12-19]
16. Peirsman Y. bert-base-multilingual-uncased-sentiment. Hugging Face. 2020. URL: https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment [accessed 2020-12-19]
17. Smit ES, Hoving C, Schelleman-Offermans K, West R, de Vries H. Predictors of successful and unsuccessful quit attempts among smokers motivated to quit. Addict Behav 2014 Sep;39(9):1318-1324. [doi: 10.1016/j.addbeh.2014.04.017] [Medline: 24837754]
18. Gwaltney CJ, Metrik J, Kahler CW, Shiffman S. Self-efficacy and smoking cessation: a meta-analysis. Psychol Addict Behav 2009 Mar;23(1):56-66. [doi: 10.1037/a0013529] [Medline: 19290690]
19. Von Ah D, Ebert S, Ngamvitroj A, Park N, Kang DH. Factors related to cigarette smoking initiation and use among college students. Tob Induc Dis 2005 Dec 15;3(1):27-40. [doi: 10.1186/1617-9625-3-1-27] [Medline: 19570288]

XSL•FO
RenderX

20. pandas—Python Data Analysis Library homepage pandas—Python Data Analysis Library. URL: https://pandas.pydata.org/ [accessed 2024-11-17]
21. Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 2020 Mar 2;17(3):261-272. [doi: 10.1038/s41592-019-0686-2]
22. Townsend J, Roderick P, Cooper J. Cigarette smoking by socioeconomic group, sex, and age: effects of price, income, and health publicity. BMJ 1994 Oct 8;309(6959):923-927. [doi: 10.1136/bmj.309.6959.923] [Medline: 7950662]
23. Fidler J, Ferguson SG, Brown J, Stapleton J, West R. How does rate of smoking cessation vary by age, gender and social grade? Findings from a population survey in England. Addiction 2013 Sep;108(9):1680-1685. [doi: 10.1111/add.12241] [Medline: 23668684]

Abbreviations

CARE: Consultation and Relational Empathy
LLM: large language model
MI: motivational interviewing

Edited by A Mavragani; submitted 02.02.24; peer-reviewed by A Carcone, I Balan, J Wilson; revised version received 20.02.25; accepted 21.02.25; published 21.03.25.

Please cite as:
Zhu J, Dong A, Wang C, Veldhuizen S, Abdelwahab M, Brown A, Selby P, Rose J
The Impact of ChatGPT Exposure on User Interactions With a Motivational Interviewing Chatbot: Quasi-Experimental Study
JMIR Form Res 2025;9:e56973
URL: https://formative.jmir.org/2025/1/e56973
doi:10.2196/56973

© Jiading Zhu, Alec Dong, Cindy Wang, Scott Veldhuizen, Mohamed Abdelwahab, Andrew Brown, Peter Selby, Jonathan Rose. Originally published in JMIR Formative Research (https://formative.jmir.org), 21.3.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

A Brief Cognitive Behavioral Therapy–Based Digital Intervention for Reducing Hazardous Alcohol Use in South Korea: Development and Prospective Pilot Study

Manjae Kwon[1,2*], MD; Daa Un Moon[3*], MD; Minjae Kang[1,2], MD; Young-Chul Jung[1,2], MD, PhD

[1]Department of Psychiatry, Yonsei University College of Medicine, Seoul, Republic of Korea

[2]Institute of Behavioral Science in Medicine, Yonsei University College of Medicine, Seoul, Republic of Korea

[3]Department of Psychiatry and Neurosciences, Charité Campus Mitte, Charité Universitätsmedizin Berlin, Corporate Member of Freie Universität and Humboldt-Universität zu Berlin, Berlin, Germany

[*]these authors contributed equally

Corresponding Author:
Young-Chul Jung, MD, PhD
Department of Psychiatry, Yonsei University College of Medicine, Seoul, Republic of Korea

Abstract

Background: Alcohol consumption is a leading cause of death and disability worldwide, associated with numerous acute and chronic medical conditions. Digital health interventions offer a promising solution to overcome barriers associated with traditional treatment methods, providing accessible, scalable, and cost-effective means to support individuals in reducing hazardous drinking.

Objective: This pilot study aims to evaluate the feasibility, acceptability, and preliminary efficacy of the Sober smartphone app in individuals with hazardous alcohol use.

Methods: This single-group, pre- and postpilot study included 20 participants with risky alcohol use, identified using the Alcohol Use Disorder Identification Test. Participants used the Sober app for 4 weeks, incorporating cognitive behavioral therapy–based interventions. Feasibility was assessed by study and session completion rates, acceptability by participant satisfaction and perceived usefulness, and preliminary efficacy by changes in alcohol consumption and psychiatric symptoms. Semistructured interviews with participants and clinicians provided qualitative perspectives on the app's usability, efficacy, and areas for improvement.

Results: Of the 20 enrolled participants, 17 completed the study. The app demonstrated high feasibility with an 85% (17/20) study completion rate, and 59% (10/17) completed all cognitive behavioral therapy sessions. Participants reported positive acceptability, with average satisfaction and usefulness ratings of 3.8 and 3.7 of 5, respectively. Preliminary efficacy outcomes showed significant improvements: abstinence days increased from 67% to 85% (z=−3.17; P=.002), heavy drinking episodes decreased from 3.3 to 1.9 (t_{16}=−2.97; P=.003), and total alcohol consumption reduced from 456.8 to 195.9 mL (t_{16}=3.16; P=.002). Alcohol Use Disorder Identification Test scores dropped from 17.5 to 10.7 (t_{16}=4.51; P<.001). Additionally, depression (Patient Health Questionnaire-9) scores decreased from 5.8 to 4.4 (t_{16}=2.91; P=.01), and anxiety (Generalized Anxiety Disorder-7) scores from 3.4 to 2.1 (z=−2.80; P=.005). No adverse events were reported. Qualitative analysis found participants valued daily logging but noted usability issues, while clinicians called for tailored goals, enhanced communication features, and age-specific content.

Conclusions: The mobile app Sober shows promise as an effective tool for reducing hazardous alcohol consumption and improving related psychiatric symptoms. The study demonstrated high feasibility and positive acceptability, with significant preliminary efficacy in reducing alcohol use. Qualitative findings provided actionable evidence for refining the app's usability and clinical integration. Further research through a randomized controlled trial is warranted to confirm these findings and optimize the app's features and content.

Trial Registration: ClinicalTrials.gov NCT06502756; https://clinicaltrials.gov/study/NCT06502756

(JMIR Form Res 2025;9:e64459) doi:10.2196/64459

KEYWORDS

alcohol; hazardous alcohol use; digital intervention; cognitive behavioral therapy; mobile apps; prevention; therapy-based; cognitive behavioral; alcohol use; South Korea; prospective pilot study; pilot study; alcohol consumption; death; disability; chronic medical condition; digital health interventions; traditional treatment methods; Korean; hazardous drinking; acceptability; feasibility; smartphone app; alcohol use disorder; psychiatric symptoms; mobile phone

XSL•FO
RenderX

Introduction

Alcohol is the most widely used psychoactive substance worldwide and poses a significant public health problem with a massive socioeconomic burden [1,2]. It is a leading cause of death and disability globally, contributing to a wide range of acute and chronic medical conditions. Acute medical consequences of high alcohol intake include injuries, car accidents, and violence [3]. Chronic diseases associated with alcohol consumption include liver cirrhosis, cardiovascular diseases, and various cancers [4-8]. Additionally, alcohol use disorders are linked to numerous psychiatric comorbidities, such as mood disorders, anxiety disorders, personality disorders, and suicide [9,10].

Hazardous drinking, defined as a drinking behavior that significantly increases the risk of adverse alcohol-related outcomes [11], is prevalent among those who consume alcohol, with a global prevalence of 17% and 15.1% in South Korea [2,12]. Effective prevention typically includes screening and brief interventions to identify and address hazardous drinking early [13-15]. These interventions involve trained staff using a brief screening tool combined with motivational interviewing techniques to provide personalized feedback and clinical options tailored to the individual's level of risk, including a referral to treatment component, if necessary [11,16-18]. However, several barriers hinder the implementation of these interventions, including concerns about stigmatization, lack of resources, insufficient training, high clinician workload, and logistical and financial obstacles for patients seeking help [19-22].

Digital health interventions offer a promising solution to overcome this barrier, especially in countries like South Korea, which has the highest smartphone ownership rate in the world at approximately 95%, with minimal generational gaps in ownership [23]. With near-ubiquitous smartphone ownership, these interventions can improve health outcomes by providing inexpensive, easily accessible, and scalable methods to deliver education, support, and monitoring through personal devices [24]. Currently, many apps in commercial app stores aim to help users reduce their alcohol consumption, but few have been rigorously evaluated, and only a minority have published efficacy information [25]. While the overall evidence for alcohol reduction through digital interventions is promising, it remains inconclusive [25-31]. Moreover, reporting theoretical foundations in these apps is often limited and unclear, indicating a need for more rigorous evaluation and transparency in their development [31].

Hazardous alcohol use frequently co-occurs with psychological distress, including symptoms of depression and anxiety [32]. These symptoms often create a cyclical relationship with problematic alcohol use, where they exacerbate drinking patterns and complicate treatment outcomes [33,34]. Research shows that brief interventions targeting hazardous alcohol use can reduce consumption and alleviate psychological distress, depression, and anxiety, addressing both alcohol use and co-occurring mental health symptoms [35,36]. Studies also highlight the potential of digital interventions to achieve similar dual benefits, providing scalable solutions to this complex challenge [37,38].

This pilot study aims to estimate the potential of a digital intervention using the Sober smartphone app (WELT Corp Ltd) for secondary prevention in individuals with hazardous alcohol use. The primary objective is to evaluate the app's feasibility, acceptability, and preliminary efficacy in reducing hazardous drinking behaviors. In addition, the study examines secondary outcomes, including the impact of the intervention on symptoms such as depression, anxiety, and stress. The findings from this pilot study will be instrumental in refining the app's features and content and optimizing the study design for a planned randomized controlled trial (RCT).

Methods

Design and Setting

This single-group, pre- and postpilot study assessed the preliminary efficacy, feasibility, and acceptability of the mobile app Sober for reducing hazardous alcohol consumption. The single-group design was chosen to provide initial insights into the intervention's effects and to inform the design of future RCTs.

Ethical Considerations

The study was approved by the institutional review board of Yonsei Severance Hospital (approval 4-2023-0338) and adhered to the ethical standards of the Declaration of Helsinki. Prior to study participation, all participants provided written informed consent, including consent for the collection and analysis of their data. Participants were informed of their right to withdraw from the study at any time. No modifications to the content or methodology were made after the study commenced. We anonymized the participants using unique ID numbers. No identifying or potentially identifying information is included in this paper or Multimedia Appendices 1 and 2. All identifying or potentially identifying data are securely maintained by Yonsei Severance Hospital and accessible only to the authorized research team. The participants received approximately US $34 for each clinic visit for their time and travel expenses.

Recruitment

Participants were recruited through outpatient clinic bulletin boards, web-based advertisements (eg, university websites), flyers, and word of mouth. Recruitment targeted individuals who self-identified as having concerns about their drinking habits and sought mental health support. The study was conducted at Yonsei Severance Hospital, a tertiary care teaching hospital in Seoul, South Korea. Interested individuals received detailed information about the study and underwent an in-person screening process after providing written informed consent. Board-certified psychiatrists (MK, DUM, and YCJ) reviewed each participant's medical and medication history to determine eligibility following a psychiatric assessment [39], including a review of medical and medication history.

Eligibility Criteria

Participants aged 19 - 65 years who scored 8 or higher on the Alcohol Use Disorder Identification Test (AUDIT) [40,41] were

XSL•FO
RenderX

included, with risky alcohol use encompassing hazardous use, harmful use, and probable dependence. Additional inclusion criteria required participants to possess sufficient Korean language skills, own an Android smartphone, and have no difficulty using mobile apps independently. They also needed to understand the study information and provide written informed consent. Exclusion criteria included major psychiatric disorders such as schizophrenia spectrum disorders, bipolar disorder, and major depressive disorder. Participants with substance use disorders involving substances other than alcohol or nicotine (eg, illicit drugs or prescription medications) were excluded. Participants with active and progressive physical illnesses, unstable medical conditions, or a life expectancy of less than 6 months were also excluded. Pregnant individuals or those planning to become pregnant during the study period, as well as individuals who participated in another clinical trial within 4 weeks before screening, were not eligible.

Procedure

Enrollment and follow-up for the study occurred between May and August 2023 (Table 1). Participants who consented to join the pilot study downloaded the Sober app and were given usernames to sign in. All participants were encouraged to use the app for self-monitoring and to log in regularly over the 4 weeks. Additionally, participants were provided with a Samsung Galaxy Smart Watch 5 (model: SM-R900NZAAKOO) to collect biometric data throughout the study. Participants were required to visit the clinic 3 times: at baseline (visit 1), 2 weeks after the baseline visit (visit 2), and 4 weeks after the baseline visit (visit 3) for assessments. Each visit included a consultation with a board-certified psychiatrist who provided standard psychiatric consultation regarding alcohol consumption based on the data entered into the mobile app.

XSL•FO
RenderX

Table . Schedule of enrollment, interventions, and assessments for a single-group, pre-post pilot study evaluating the feasibility, acceptability, and preliminary efficacy of the Sober mobile app[a].

	Study period and time points			
	Enrollment	Visit 1	Visit 2	Visit 3 and closeout
	0 days	0 days	14±4 days	28±4 days
Enrollment				
Eligibility screen and informed consent	✓			
Intervention				
Mobile app Sober and consultation with a psychiatrist at each visit		✓	✓	✓
Assessments				
Sociodemographic and clinical characteristics	✓			
Alcohol consumption history	✓			
Primary objectives				
Feasibility of the Sober app			✓	✓
Acceptability of the Sober app				✓
Days abstinent in the past 2 weeks (%)		✓	✓	✓
Secondary objectives				
AUDIT[b]	✓		✓	✓
PHQ-9[c]		✓	✓	✓
GAD-7[d]		✓	✓	✓
PSS[e]		✓	✓	✓
CIWA-Ar[f]		✓	✓	✓
Amount of alcohol consumed (mL) in the past 2 weeks		✓	✓	✓
Adverse events		✓	✓	✓

[a]The study was conducted at Yonsei Severance Hospital, Seoul, South Korea, from May to August 2023.

[b]AUDIT: Alcohol Use Disorders Identification Test.

[c]PHQ-9: Patient Health Questionnaire-9.

[d]GAD-7: Generalized Anxiety Disorder-7.

[e]PSS: Perceived Stress Scale.

[f]CIWA-Ar: Clinical Institute Withdrawal Assessment of Alcohol Scale—Revised.

Sober App

The Sober app was developed as a digital intervention tool to reduce and prevent hazardous drinking (Figure 1). The development process involved a multidisciplinary team of experienced clinicians, software developers, product designers, and behavioral scientists to ensure clinical effectiveness and user-friendliness. Before developing the program, the team reviewed relevant materials, such as treatment manuals, intervention descriptions, guidelines, patient reports, and trial results. The program's development involved iterative evaluations of prototypes, with the findings used to refine and enhance the final app design. Designed for Android smartphones, the app supports clinicians, especially those with minimal addiction training, in providing structured, evidence-based interventions to hazardous drinkers in the general population. The cognitive behavioral therapy (CBT)–based app incorporates interventions that include psychoeducation, motivational interviewing, cognitive behavioral skills training (eg, problem-solving and behavior

XSL•FO
RenderX

substitution), and personalized normative feedback. It provides content on the effects of alcohol on the brain and behavior, risk awareness, and strategies for reducing risky drinking habits while encouraging users to consider alternatives to heavy drinking.

Figure 1. Screenshots of the Sober app's interface. Illustrative screenshots of the mobile app used in this study, showcasing features for daily logging, educational modules, and progress tracking. Original content in Korean has been translated into English for presentation.

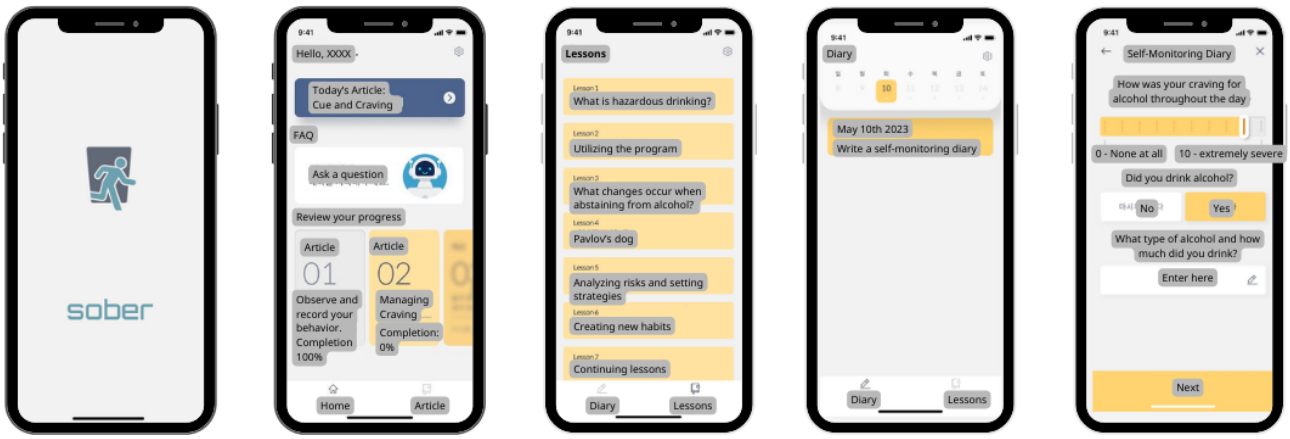

The modularly structured app features educational modules, self-monitoring tools, and interactive elements. The CBT-based modules are divided into 4 chapters, which include 8 sessions, each addressing different aspects of hazardous drinking and providing skills training and risk awareness (Table 2). The content is organized sequentially. The first chapter, "What is hazardous drinking?," introduces participants to the concept and helps them set initial goals, such as abstaining from alcohol for 4 weeks, entering self-monitoring reports, and engaging in the CBT program. Participants also receive personalized normative feedback on their drinking habits. The second chapter, "What changes occur when abstaining from alcohol?," covers psychoeducation about changes in sleep, appetite, and emotions resulting from abstinence, with participants recording these changes in their self-monitoring logs. The third chapter, "Pavlov's dog," explains the concepts of cues, cravings, behaviors, and outcomes, helping participants identify high-risk situations and understand how cravings lead to drinking behavior. This chapter guides participants in exploring and selecting alternative behaviors to replace drinking. The fourth chapter, "Forming a healthy drinking habit," reviews the previous 4 weeks and reinforces the skills learned. To facilitate self-monitoring and promote self-awareness, the app included a daily log function, where participants recorded their alcohol consumption, mood, appetite, and sleep quality.

XSL•FO
RenderX

Table . Description of cognitive behavioral therapy–based modules in the Sober mobile app intervention.

Week and chapter		Session	Key principles and contents
Week 1			
	What is hazardous drinking?	Do I belong to high-risk drinking group?	Introduction to the High-Risk Drinking Prevention Program: this includes a description of high-risk drinking behaviors and a self-assessment for participants to evaluate their own drinking habits. Normative feedback is provided for the participants' level of current drinking habit. The program also offers psychoeducation about the long-term risks associated with continuing these drinking habits.
	What is hazardous drinking?	Promise for 4 weeks abstinence	We set up the action plan with the initial goal of abstaining from alcohol for the next 4 weeks while participating in the program. Evaluate individuals' expectations regarding the outcomes of alcohol abstinence. Provide instructions for recording sleep, emotions, appetite, and alcohol consumption in the app.
Week 2			
	What changes occur when abstaining from alcohol?	Alcohol and sleep	Psychoeducation regarding the link between alcohol and sleep.
	What changes occur when abstaining from alcohol?	Alcohol and appetite	Psychoeducation regarding the link between alcohol and appetite.
	What changes occur when abstaining from alcohol?	Alcohol and emotion	Psychoeducation regarding the link between alcohol and emotion.
Week 3			
	Pavlov's dog	Cue	Description of the cue-craving-behavior-result cycle in alcohol consumption. This session focuses on identifying the cues that lead to this cycle.
	Pavlov's dog	Craving	Psychoeducation on the various ways "craving" may be identified in participants. Participants will set up alternative behaviors to resort to when they experience alcohol cravings.
Week 4			
	Forming a healthy drinking habit	Review and relapse prevention	Review of the intervention and establishment of relapse prevention strategies.

Assessments

At baseline (visit 1), participants provided sociodemographic information, lifestyle habits, past disease history, previous medication, and detailed information on their alcohol consumption habits. At each of the 3 visits, participants were evaluated using several standardized instruments: The AUDIT was used to assess hazardous and harmful patterns of alcohol consumption [40,41]. Symptoms of depression were measured using the Patient Health Questionnaire-9 (PHQ-9) [42,43], while symptoms of anxiety were evaluated using the Generalized Anxiety Disorder-7 (GAD-7) scale [44,45]. The Perceived Stress Scale (PSS) was used to assess perceived levels of stress [46,47], and the Clinical Institute Withdrawal Assessment of Alcohol Scale—Revised (CIWA-Ar) was used to evaluate alcohol withdrawal symptoms [48,49]. At the final visit (visit 3), the completion rate of the app-based sessions was assessed along with participants' satisfaction ratings, perceived usefulness of the app, and ease of mastering the app program, all assessed through a 5-point Likert scale. Adverse events were monitored and recorded at each visit to ensure participant safety.

XSL•FO
RenderX

Outcomes

The study's primary outcomes focused on feasibility, acceptability, and preliminary efficacy. Feasibility was assessed at visit 3 based on the study completion rate and the completion of all 8 CBT-based intervention sessions. Acceptability was evaluated at visit 3 using 3 measures: participant satisfaction, perceived usefulness of the app, and ease of mastering the app.

Preliminary efficacy was analyzed by comparing outcomes between baseline (visit 1) and postintervention (visit 3). The primary outcome was the change in the number of days participants abstained from alcohol in the past 14 days. Secondary outcomes included changes in total AUDIT scores, the total amount of alcohol consumed in the past 14 days, and psychiatric symptom scores (PHQ-9, GAD-7, PSS, and CIWA-Ar), all assessed at both baseline (visit 1) and postintervention (visit 3).

Statistical Analysis

All statistical analyses were conducted using SPSS (version 26 for Windows; IBM Corp). Frequencies and percentages described categorical variables, while means and SDs were used for continuous variables. Depending on the data distribution's normality (Shapiro-Wilk test), paired *t* tests (2-tailed) or Wilcoxon signed rank tests were used to compare pre- and postintervention outcomes. Statistical significance was set at $P<.05$. Results were reported with corresponding *P* values.

Participant and Clinician Interviews

Qualitative data were collected through semistructured, open-ended interviews conducted at the final visit (visit 3). All 17 participants who completed the study participated in individual interviews to explore their experiences with the app and the clinical trial. Additionally, 2 clinicians involved in the trial were interviewed separately following the study's conclusion to provide insights from a clinical perspective.

The interviews were conducted face-to-face (in private rooms at the study site) or via videoconferencing, depending on participant preferences, and lasted approximately 20‑30 minutes. An interview guide was used to ensure consistency across interviews. For participants, questions focused on app usability, content relevance, and perceived impact on drinking behaviors. For clinicians, the interviews explored the app's integration into clinical workflows, feasibility, and suggestions for tailoring the intervention to diverse patient needs. All interviews were conducted by members of the research team involved in the study. Interviews were audio-recorded with participant consent, transcribed verbatim, and anonymized to maintain confidentiality.

Thematic analysis was used to analyze the qualitative data, following Braun and Clarke's 6-phase framework [50]: 2 independent researchers (DUM and Yujin Lee) coded the transcripts to identify key patterns and themes. Codes were grouped into overarching themes and refined iteratively to ensure they accurately represented the data. Representative verbatim quotes were selected to illustrate key findings and provide depth to the analysis. Discrepancies in coding were resolved through discussion. The findings are reported following the COREQ (Consolidated Criteria for Reporting Qualitative Research) guidelines [51].

Results

Participant Eligibility and Baseline Characteristics

A total of 20 eligible participants with hazardous alcohol use, identified based on their AUDIT scores, were enrolled in the study. Of these, 17 completed the 4-week clinical trial. In total, 2 participants dropped out because they no longer had access to Android smartphones, and 1 withdrew voluntarily (Figure 2). All participants were male, with a mean age of 22.29 (SD 8.3) years. Most were students and single. Regarding smoking status, 3 participants were current smokers, 3 were ex-smokers, and 11 were never smokers. The mean age at first alcohol drinking was 18.9 (SD 1.2) years, and the mean age at first heavy drinking episode was 19.7 (SD 1.0) years. In total, 7 participants had a family history of alcohol-related problems, and 3 showed insights into their alcohol problems (Table 3).

XSL•FO
RenderX

Figure 2. Flowchart summarizing the participant recruitment, screening, enrollment, intervention, and follow-up processes. AUDIT: Alcohol Use Disorders Identification Test.

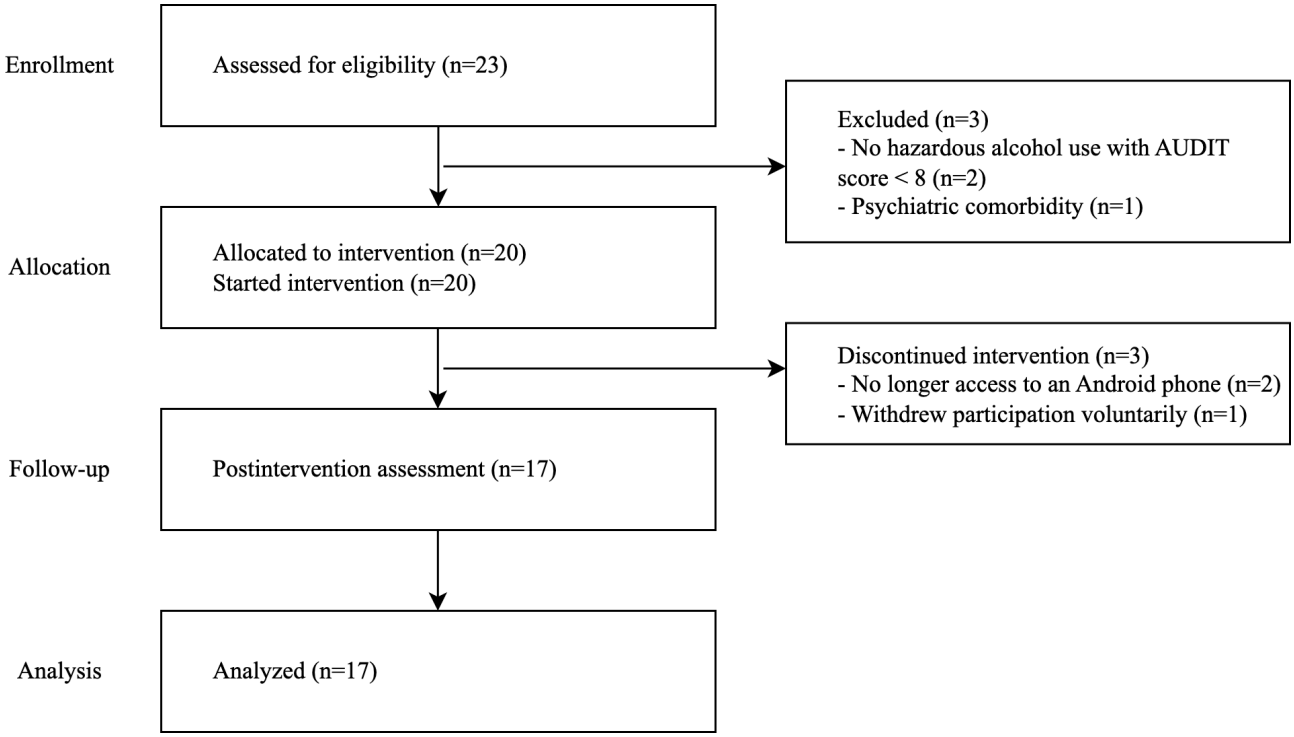

XSL•FO
RenderX

Table . Participants' baseline characteristics (n=17).

Characteristics	Values
Sex (male), n (%)	17 (100)
Age (years), mean (SD)	22.29 (8.3)
Height (cm), mean (SD)	174.9 (3.9)
Weight (kg), mean (SD)	74.4 (12.8)
Education, n (%)	
Middle school	1 (6)
High school	12 (71)
Bachelor's degree	3 (18)
Master's degree or postgraduate	1 (6)
Occupation, n (%)	
Student	14 (82)
Office worker	1 (6)
Labor worker	1 (6)
Unemployed	1 (6)
Religion, n (%)	
Christian	1 (6)
Others	0 (0)
None	16 (94)
Marital status, n (%)	
Single	16 (94)
Married	0 (0)
Divorced	1 (6)
Living arrangement, n (%)	
Living alone	8 (47)
Living with others	9 (53)
Smoking status, n (%)	
Current smokers	3 (18)
Ex-smoker	3 (18)
Never smoker	11 (65)
Age at first alcohol drinking (years), mean (SD)	18.9 (1.2)
Age at first binge drinking (years), mean (SD)	19.7 (1.0)
Presence of alcohol-related accident, n (%)	
No	10 (59)
Yes (once)	5 (29)
Yes (multiple)	2 (12)
Presence of alcohol-related legal problems, n (%)	
No	16 (94)
Yes (once)	0 (0)
Yes (multiple)	1 (56)
Presence of drunken driving, n (%)	
No	17 (100)
Yes	0 (0)

XSL•FO
RenderX

Characteristics	Values
Family history of alcohol-related problems, n (%)	
No	10 (59)
Yes	7 (41)
Insight of alcohol problems, n (%)	
No	14 (82)
Yes	3 (18)

Feasibility and Acceptability

The feasibility of the intervention was assessed based on the completion rates of the study and the CBT-based educational sessions within the app. Of the 20 participants initially enrolled, 17 (85%) completed the study. Among these 17 participants, 10 (59%) completed all 8 CBT-based sessions. Acceptability was evaluated in the 17 participants who completed the final visit (visit 3). Participants were asked to rate their satisfaction with the app on a scale from 0=not satisfied at all to 5=very satisfied. The average satisfaction score was 3.8 (SD 0.9). When asked to rate the usefulness of the app from 0=not useful at all to 5=very useful, the average response was 3.7 (SD 0.9). Additionally, participants rated their ease of mastering the app on a scale from 0=could not use without help to 5=used naturally without difficulty, with an average score of 3.5 (SD 0.6).

Preliminary Efficacy Outcomes

Preliminary efficacy outcomes showed significant improvements in several key measures (Table 4). The percentage of days participants remained abstinent increased from 67% to 85% (z=−3.17; P=.002). The frequency of heavy drinking episodes decreased from 3.3 to 1.9 (t_{16}=−2.97; P=.003). The amount of alcohol consumed reduced significantly from 456.8 to 195.9 mL (t_{16}=3.16; P=.002). Total AUDIT scores dropped from 17.5 to 10.7 (t_{16}=4.51; P<.001). Regarding psychological assessment outcomes, PHQ-9 scores decreased from 5.8 to 4.4 (t_{16}=2.91; P=.01), and GAD-7 scores from 3.4 to 2.1 (z=−2.80; P=.005). There were nonsignificant changes in PSS scores (17.2 to 16.1; P=.19) and CIWA-Ar scores (0.3 to 0.8; P=.74). No adverse events were reported by the study participants.

Table . Changes in alcohol consumption, psychological symptoms, and related measures before and after a 4-week intervention using the mobile app "Sober."

	Baseline (version 1), mean (SD)	Postintervention (version 3), mean (SD)	Difference, mean (SD)	t test (df=16)[a]	P value
Days abstinent in the past 2 weeks (%)	67 (25)	85 (16)	18.0 (4.2)	z=−3.17[b]	.002
Frequency of heavy drinking (AUDIT[c], question 3)	3.3 (0.7)	1.9 (1.4)	−1.4 (0.3)	−2.97	.003
Amount of alcohol consumed (mL)	456.8 (441.9)	195.9 (320.2)	−260.9 (82.6)	3.16	.002
AUDIT	17.5 (6.2)	10.7 (9.0)	−6.7 (1.5)	4.51	<.001
PHQ-9[d]	5.8 (5.4)	4.4 (5.6)	−1.4 (0.5)	2.91	.01
GAD-7[e]	3.4 (3.4)	2.1 (3.5)	−1.3 (0.5)	z=−2.80[b]	.005
PSS[f]	17.2 (2.9)	16.1 (2.6)	−1.1 (0.8)	1.38	.19
CIWA-Ar[g]	0.3 (0.6)	0.8 (2.7)	0.5 (0.7)	z=0.01[b]	.74

[a]Paired t test (2-tailed).

[b]Wilcoxon signed rank test.

[c]AUDIT: Alcohol Use Disorder Identification Test.

[d]PHQ-9: Patient Health Questionnaire-9.

[e]GAD-7: Generalized Anxiety Disorder-7.

[f]PSS: Perceived Stress Scale.

[g]CIWA-Ar: Clinical Institute Withdrawal Assessment of Alcohol Scale—Revised.

XSL•FO
RenderX

Participant and Clinician Perspectives on the Intervention

Qualitative feedback from participants and clinicians provided insights into the app's usability, effectiveness, and areas for improvement. For a detailed summary, see Multimedia Appendices 1 and 2.

Participant Feedback

Interviews with the 17 participants highlighted 3 primary themes: behavioral impact, usability challenges, and content suggestions.

Participants consistently reported that using the app increased their awareness of drinking behaviors and encouraged self-reflection, which helped them reduce alcohol consumption:

> *When I logged my drinking every day, it made me stop and think about how much I was drinking. It naturally helped me reduce my alcohol consumption.* [Participant 1]
>
> *Tracking how often and how much I drink made me more aware of my habits.* [Participant 3]

Participants also noted the app's role in helping them observe connections between their drinking and other aspects of their health, such as sleep and appetite:

> *It was helpful to see how my drinking affected things like my appetite and sleep. It made me think about how drinking impacts my health.* [Participant 13]

However, a participant reported unintended effects, such as increased cravings when focusing on abstinence:

> *Sometimes, thinking about not drinking actually made me want to drink more. It felt like a trigger at times.* [Participant 7]

Usability issues were commonly mentioned, including app crashes, lack of notifications, and vague criteria. Participants suggested improvements to enhance functionality:

> *The app would crash a lot when I tried to log something, which was really frustrating.* [Participant 6]
>
> *I wish the app had reminders, like notifications in the morning to remind me to log.* [Participant 4]
>
> *It felt like some of the criteria in the app were too vague. I wasn't always sure what I was supposed to do.* [Participant 3]
>
> *It would be nice to see trends over time, like a graph showing my progress.* [Participant 9]

Participants also recommended expanding the app's content with more educational materials, such as strategies for managing cravings and understanding the risks of alcohol consumption:

> *The app could include more about the long-term effects of drinking.* [Participant 17]
>
> *It would be great if I could log more details, like my emotions or cravings. Just using a simple scale didn't capture everything.* [Participant 12]

Clinician Feedback

Interviews with 2 clinicians involved in the trial provided additional perspectives on the app's clinical utility and suggestions for improvement.

Clinicians observed that participants who recorded their mood and sleep alongside abstinence became more aware of positive changes in their mental and physical health, reinforcing their motivation to reduce alcohol consumption:

> *By recording their mood and sleep daily while abstaining from alcohol, participants seemed to independently recognize the positive changes in their mind and body, which, in turn, reinforced their commitment to staying alcohol-free.* [Clinician 1]

They emphasized the importance of flexibility in accommodating participants' diverse goals, such as achieving abstinence or reducing binge episodes, suggesting features like weekly goal setting and tracking achievements:

> *Allowing for participants to set personalized goals every week and adding features that allow participants to record their achievement would be beneficial.* [Clinician 1]

To enhance clinical integration, clinicians recommended features that facilitate communication and monitoring, such as messaging options and tools to assess participants' motivation:

> *Features that connect clinicians and patients, like messaging or feedback options, would make follow-ups easier and more effective.* [Clinician 2]
>
> *If patients' motivation levels for change can be assessed through the app, it will help clinicians effectively tailor their approach to patient interventions in real-life settings.* [Clinician 2]

Additionally, clinicians underscored the importance of tailoring app content to different age groups and suggested incorporating features that allow participants to reflect on their motivations for change:

> *Younger and older users have very different needs and expectations. The app should include content that's tailored to different age groups.* [Clinician 2]

Discussion

Principal Findings

This pilot study aimed to assess the feasibility, acceptability, and preliminary efficacy of the Sober mobile app for reducing hazardous alcohol consumption. The results demonstrated significant improvements in key measures of alcohol use, including increased days of abstinence, reduction in heavy drinking episodes, and decreased overall alcohol consumption. Additionally, notable reductions in AUDIT scores and improvements in psychological assessments (PHQ-9 and GAD-7 scores) were observed.

Qualitative feedback from participants and clinicians provided valuable insights into the app's strengths and areas for improvement. Participants emphasized the app's ability to increase awareness of drinking behaviors and promote

XSL•FO
RenderX

self-reflection, which supported behavior change. However, usability challenges, such as technical issues and limited options for personalization, were highlighted. Clinicians stressed the importance of tailoring the app to diverse user goals and incorporating features to facilitate clinical integration. These findings will inform the refinement of the app and future iterations of the study design.

Comparison With Prior Work

Our findings align with previous research indicating the effectiveness of digital interventions in reducing alcohol consumption, particularly among the young male population [52], which mirrors the demographic of our study. Other studies have also shown positive effects in the general population, suggesting that such interventions could have broad applicability [29,31,53-55]. The promising use of CBT in digital formats is consistent with existing evidence that CBT delivered in clinical settings can reduce alcohol consumption [11,56,57]. Other digital interventions using CBT-based approaches have also demonstrated effectiveness in reducing alcohol consumption [58-60]. However, in the CBT framework, various interventions, including normative feedback, self-monitoring, psychoeducation, action planning, goal setting, problem-solving skills, and identifying or managing triggers and cravings, aim to strengthen "reflective" cognitive processes used to control behavior [25]. This approach diversity highlights the need for a more standardized framework for developing and evaluating digital interventions [31]. This study also supports findings that digital interventions can positively impact co-occurring psychological symptoms, such as depression and anxiety, further emphasizing their dual utility [61,62].

The app's high feasibility, demonstrated by an 85% (17/20) study completion rate, suggests that it is well-received and manageable for participants. However, the fact that only 59% (10/17) completed all 8 CBT-based educational sessions highlights the challenge of maintaining engagement with the full content. Strategies to enhance user engagement, such as incorporating gamification elements, personalized feedback, and real-time adjustments using predictive modeling [63-65], could address this gap.

Acceptability ratings were positive, with satisfaction and perceived usefulness ratings averaging 3.8 and 3.7 of 5, respectively. Feedback indicated that daily logging and self-monitoring features were valuable in promoting self-reflection and awareness of drinking habits. However, several usability improvements were suggested, such as showing averages of past drinking, fixing app crashes, adding notification settings, and integrating with other health apps. Clinician feedback emphasized the need for customizable goals, treatment periods, and content tailored to different age groups. These suggestions will be considered in future iterations of the app to ensure that it meets the diverse needs of its users. Furthermore, incorporating interactive elements and practical synthesis in workflows could enhance the overall intervention.

Strengths and Limitations

This study demonstrated the potential of the Sober app for reducing risky alcohol consumption; yet, several limitations must be acknowledged. The single-arm, pre- and postdesign limits causal inferences, and the study was not sufficiently powered. The short duration of the study and lack of long-term follow-up limit the ability to assess the sustained effects of the intervention. The inclusion of face-to-face sessions with clinicians may have influenced the outcomes, as the reinforcement and availability of clinicians could have impacted the results [66]. Furthermore, the homogeneous participant demographic, consisting solely of young male participants, limits the generalizability of the findings.

Despite these limitations, the study had several notable strengths. To our knowledge, this is the first study to evaluate a brief digital intervention for hazardous alcohol use in South Korea. The significant preliminary efficacy demonstrated suggests potential for real-life clinical implementation. The high overall completion rate and positive acceptability ratings indicate that the app is well-received and safe to use. Importantly, the inclusion of qualitative feedback provides a nuanced understanding of user experiences, directly informing app improvements and study design.

At a population level, addressing interventions for individuals who exhibit hazardous drinking can have the greatest impact on reducing alcohol-related problems [67]. By providing an easily accessible, scalable, and cost-effective intervention, the Sober app could help reduce the overall burden of hazardous drinking on health care systems. While numerous studies have shown that screening and brief intervention can lead to significant decreases in alcohol use in primary care populations [13,68], digital intervention lessens the burden on primary care providers, requires no extensive training, and mitigates the fear of stigma, as interventions can be anonymous [69]. This is particularly relevant in South Korea, where hazardous drinking rates are significant [12], and smartphone ownership is nearly ubiquitous [23].

Future Research

Future research should include an RCT to assess the app's efficacy compared to standard care without app use, with larger samples and follow-up assessments to determine long-term effects. Expanding the study to include a more diverse population will enhance the generalizability of the findings [70]. Optimizing the app's tailoring, structure, and content based on participant and clinician feedback is essential. Personalized content, interactive elements [71], gamification [72,73], and predictive modeling using data from daily logs and wearables could improve user engagement and intervention effectiveness by providing personalized feedback and real-time adjustments [63-65]. Integrating the app with other health monitoring tools and clinician workflow could give a more holistic approach to managing alcohol consumption [71,74].

Conclusions

The mobile app Sober shows promise as an effective tool for reducing hazardous alcohol consumption and improving related psychiatric symptoms. This pilot study demonstrated significant preliminary efficacy, high overall feasibility, and positive acceptability. The app's features, such as daily logging and CBT-based content, helped users develop greater self-awareness

XSL•FO
RenderX

and control over their drinking habits, showing potential for integration into clinical practice as a supportive tool. Qualitative feedback from participants and clinicians identified critical areas for improvement, such as usability, goal customization, and age-specific content, which will guide future app refinement. With further research and refinement, the Sober app has the potential to make an impact on public health by providing a scalable and accessible intervention for hazardous drinking.

Acknowledgments

This work was supported by a Seoul Business Agency grant funded by the Seoul Metropolitan Government (BT210029). The funding organization had no role in the study design, data collection, analysis, interpretation of results, or the decision to submit the manuscript for publication. The authors would like to acknowledge Yujin Lee (College of Pharmacy, Yonsei Institute of Pharmaceutical Sciences, Yonsei University, Incheon, Republic of Korea) for her contribution as one of the 2 independent researchers responsible for coding and analyzing the qualitative data. Generative artificial intelligence was not used in any portion of this manuscript.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

DUM was employed by WELT Corp Ltd. The other authors declare no competing interests.

Multimedia Appendix 1
Participant codes with acceptability measures.
[DOCX File, 20 KB - formative_v9i1e64459_app1.docx]

Multimedia Appendix 2
Thematic quotes from participant and clinician interviews.
[DOCX File, 21 KB - formative_v9i1e64459_app2.docx]

References

1. Health at a glance 2023: OECD indicators. : OECD Publishing; 2023. [doi: 10.1787/7a7afb35-en]
2. Global status report on alcohol and health and treatment of substance use disorders. World Health Organization. 2024. URL: https://www.who.int/publications/i/item/9789240096745 [accessed 2025-01-21]
3. Global status report on alcohol and health 2018. World Health Organization. 2018. URL: https://www.who.int/publications/i/item/9789241565639 [accessed 2025-01-21]
4. Rehm J, Gmel GE Sr, Gmel G, et al. The relationship between different dimensions of alcohol use and the burden of disease—an update. Addiction 2017 Jun;112(6):968-1001. [doi: 10.1111/add.13757] [Medline: 28220587]
5. Rehm J, Patra J, Brennan A, et al. The role of alcohol use in the aetiology and progression of liver disease: a narrative review and a quantification. Drug Alcohol Rev 2021 Nov;40(7):1377-1386. [doi: 10.1111/dar.13286] [Medline: 33783063]
6. Wild C, Weiderpass E, Stewart BW. World Cancer Report: Cancer Research for Cancer Prevention. : International Agency for Research on Cancer; 2020.
7. White IR, Altmann DR, Nanchahal K. Alcohol consumption and mortality: modelling risks for men and women at different ages. Bmj 2002 Jul 27;325(7357):191-191. [doi: 10.1136/bmj.325.7357.191]
8. Zahr NM, Kaufman KL, Harper CG. Clinical and pathological features of alcohol-related brain damage. Nat Rev Neurol 2011 May;7(5):284-294. [doi: 10.1038/nrneurol.2011.42] [Medline: 21487421]
9. Grant BF, Goldstein RB, Saha TD, et al. Epidemiology of DSM-5 alcohol use disorder: results from the national epidemiologic survey on alcohol and related conditions III. JAMA Psychiatry 2015 Aug;72(8):757-766. [doi: 10.1001/jamapsychiatry.2015.0584] [Medline: 26039070]
10. Puddephatt JA, Irizar P, Jones A, Gage SH, Goodwin L. Associations of common mental disorder with alcohol use in the adult general population: a systematic review and meta-analysis. Addiction 2022 Jun;117(6):1543-1572. [doi: 10.1111/add.15735] [Medline: 34729837]
11. MacKillop J, Agabio R, Feldstein Ewing SW, et al. Hazardous drinking and alcohol use disorders. Nat Rev Dis Primers 2022 Dec 22;8(1):80. [doi: 10.1038/s41572-022-00406-1] [Medline: 36550121]
12. Hong JW, Noh JH, Kim DJ. The prevalence of and factors associated with high-risk alcohol consumption in Korean adults: the 2009–2011 Korea National Health and Nutrition Examination Survey. PLoS ONE 2017;12(4):e0175299. [doi: 10.1371/journal.pone.0175299]

XSL•FO
RenderX

13. Kaner EF, Beyer FR, Muirhead C, et al. Effectiveness of brief alcohol interventions in primary care populations. Cochrane Database Syst Rev 2018 Feb 24;2(2):CD004148. [doi: 10.1002/14651858.CD004148.pub4] [Medline: 29476653]
14. Solberg LI, Maciosek MV, Edwards NM. Primary care intervention to reduce alcohol misuse ranking its health impact and cost effectiveness. Am J Prev Med 2008 Feb;34(2):143-152. [doi: 10.1016/j.amepre.2007.09.035] [Medline: 18201645]
15. The SAFER technical package: five areas of intervention at national and subnational levels. World Health Organization. 2019. URL: https://www.who.int/publications/i/item/the-safer-technical-package [accessed 2025-01-21]
16. US Preventive Services Task Force, Curry SJ, Krist AH, et al. Screening and behavioral counseling interventions to reduce unhealthy alcohol use in adolescents and adults: US Preventive Services Task Force recommendation statement. JAMA 2018 Nov 13;320(18):1899-1909. [doi: 10.1001/jama.2018.16789] [Medline: 30422199]
17. Miller WR, Rollnick S. Motivational Interviewing: Helping People Change: Guilford Press; 2012.
18. Vaca FE, Winn D. The basics of alcohol screening, brief intervention and referral to treatment in the emergency department. West J Emerg Med 2007 Aug;8(3):88-92. [Medline: 19561690]
19. Johnson M, Jackson R, Guillaume L, Meier P, Goyder E. Barriers and facilitators to implementing screening and brief intervention for alcohol misuse: a systematic review of qualitative evidence. J Public Health (Oxf) 2011 Sep 1;33(3):412-421. [doi: 10.1093/pubmed/fdq095]
20. Nygaard P, Aasland OG. Barriers to implementing screening and brief interventions in general practice: findings from a qualitative study in Norway. Alcohol Alcohol 2011 Jan 1;46(1):52-60. [doi: 10.1093/alcalc/agq073]
21. Wallhed Finn S, Bakshi AS, Andréasson S. Alcohol consumption, dependence, and treatment barriers: perceptions among nontreatment seekers with alcohol dependence. Subst Use Misuse 2014 May 12;49(6):762-769. [doi: 10.3109/10826084.2014.891616]
22. Kilian C, Manthey J, Carr S, et al. Stigmatization of people with alcohol use disorders: an updated systematic review of population studies. Alcoholism Clin Exp Res 2021 May;45(5):899-911. [doi: 10.1111/acer.14598]
23. Taylor K, Silver L. Smartphone ownership is growing rapidly around the world, but not always equally. Pew Research Center. 2019. URL: https://www.pewresearch.org/global/2019/02/05/smartphone-ownership-is-growing-rapidly-around-the-world-but-not-always-equally/ [accessed 2023-05-01]
24. Chong SOK, Pedron S, Abdelmalak N, Laxy M, Stephan AJ. An umbrella review of effectiveness and efficacy trials for app-based health interventions. NPJ Digit Med 2023 Dec 16;6(1):233. [doi: 10.1038/s41746-023-00981-x] [Medline: 38104213]
25. Colbert S, Thornton L, Richmond R. Smartphone apps for managing alcohol consumption: a literature review. Addict Sci Clin Pract 2020 Dec;15(1):17. [doi: 10.1186/s13722-020-00190-x]
26. Manning V, Piercy H, Garfield JBB, Clark SG, Andrabi MN, Lubman DI. A personalized approach bias modification smartphone app ("SWiPE") to reduce alcohol use: open-label feasibility, acceptability, and preliminary effectiveness study. JMIR Mhealth Uhealth 2021 Dec 10;9(12):e31353. [doi: 10.2196/31353] [Medline: 34890355]
27. Witkiewitz K, Desai SA, Bowen S, Leigh BC, Kirouac M, Larimer ME. Development and evaluation of a mobile intervention for heavy drinking and smoking among college students. Psychol Addict Behav 2014 Sep;28(3):639-650. [doi: 10.1037/a0034747] [Medline: 25000269]
28. Boyle SC, LaBrie JW, Trager BM, Costine LD. A gamified personalized normative feedback app to reduce drinking among sexual minority women: randomized controlled trial and feasibility study. J Med Internet Res 2022 May 13;24(5):e34853. [doi: 10.2196/34853] [Medline: 35559854]
29. Chiauzzi E, Green TC, Lord S, Thum C, Goldstein M. My student body: a high-risk drinking prevention web site for college students. J Am Coll Health 2005 May 1;53(6):263-274. [doi: 10.3200/JACH.53.6.263-274]
30. McDermott KT, Noake C, Wolff R, et al. Digital interventions to moderate alcohol consumption in young people: a Cancer Prevention Europe overview of systematic reviews. Front Digit Health 2023;5:1178407. [doi: 10.3389/fdgth.2023.1178407] [Medline: 37288171]
31. Kaner EF, Beyer FR, Garnett C, et al. Personalised digital interventions for reducing hazardous and harmful alcohol consumption in community-dwelling populations. Cochrane Database Syst Rev 2017 Sep 25;9(9):CD011479. [doi: 10.1002/14651858.CD011479.pub2] [Medline: 28944453]
32. Haynes JC, Farrell M, Singleton N, et al. Alcohol consumption as a risk factor for anxiety and depression: results from the longitudinal follow-up of the National Psychiatric Morbidity Survey. Br J Psychiatry 2005 Dec;187(6):544-551. [doi: 10.1192/bjp.187.6.544] [Medline: 16319407]
33. Strid C, Andersson C, Öjehagen A. The influence of hazardous drinking on psychological functioning, stress and sleep during and after treatment in patients with mental health problems: a secondary analysis of a randomised controlled intervention study. BMJ Open 2018 Mar 6;8(3):e019128. [doi: 10.1136/bmjopen-2017-019128] [Medline: 29511011]
34. Smith JP, Randall CL. Anxiety and alcohol use disorders: comorbidity and treatment considerations. Alcohol Res 2012;34(4):414-431. [Medline: 23584108]
35. Baker AL, Thornton LK, Hiles S, Hides L, Lubman DI. Psychological interventions for alcohol misuse among people with co-occurring depression or anxiety disorders: a systematic review. J Affect Disord 2012 Aug;139(3):217-229. [doi: 10.1016/j.jad.2011.08.004]

XSL•FO
RenderX

36. Magee KE, Connell AM. The role of substance use coping in linking depression and alcohol use from late adolescence through early adulthood. Exp Clin Psychopharmacol 2021 Dec;29(6):659-669. [doi: 10.1037/pha0000436] [Medline: 33539119]
37. Schouten MJE, Christ C, Dekker JJM, Riper H, Goudriaan AE, Blankers M. Digital interventions for people with co-occurring depression and problematic alcohol use: a systematic review and meta-analysis. Alcohol Alcohol 2022 Jan 8;57(1):113-124. [doi: 10.1093/alcalc/agaa147] [Medline: 33588432]
38. Stapinski LA, Prior K, Newton NC, et al. Are we making Inroads? A randomized controlled trial of a psychologist-supported, web-based, cognitive behavioral therapy intervention to reduce anxiety and hazardous alcohol use among emerging adults. eClinMed 2021 Sep;39:101048. [doi: 10.1016/j.eclinm.2021.101048] [Medline: 34622183]
39. Sheehan DV, Lecrubier Y, Sheehan KH, et al. The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. J Clin Psychiatry 1998;59 Suppl 20:22-33. [Medline: 9881538]
40. Babor TF, Higgins-Biddle JC, Saunders JB, Monteiro MG. AUDIT: the Alcohol Use Disorders Identification Test: guidelines for use in primary health care. World Health Organization. 2001. URL: https://www.who.int/publications/i/item/WHO-MSD-MSB-01.6a [accessed 2025-01-21]
41. Kim JS, Oh MK, Park BK, Lee MK, Kim GJ. Screening criteria of alcoholism by Alcohol Use Disorders Identification Test (AUDIT) in Korea. J Korean Acad Fam Med 1999;20(9):1152-1159.
42. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613. [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
43. Shin C, Lee SH, Han KM, Yoon HK, Han C. Comparison of the usefulness of the PHQ-8 and PHQ-9 for screening for major depressive disorder: analysis of psychiatric outpatient data. Psychiatry Investig 2019 Apr;16(4):300-305. [doi: 10.30773/pi.2019.02.01]
44. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
45. Ahn JK, Kim Y, Choi KH. The psychometric properties and clinical utility of the Korean version of GAD-7 and GAD-2. Front Psychiatry 2019;10:127. [doi: 10.3389/fpsyt.2019.00127]
46. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. J Health Soc Behav 1983 Dec;24(4):385-396. [Medline: 6668417]
47. Lee EH. Review of the psychometric evidence of the Perceived Stress Scale. Asian Nurs Res (Korean Soc Nurs Sci) 2012 Dec;6(4):121-127. [doi: 10.1016/j.anr.2012.08.004]
48. Choi HY, Seo JS, Lee SK. Assessment and treatment of alcohol withdrawal syndrome. Korean J Gastroenterol 2020 Aug 25;76(2):71-77. [doi: 10.4166/kjg.2020.76.2.71]
49. Sullivan JT, Sykora K, Schneiderman J, Naranjo CA, Sellers EM. Assessment of alcohol withdrawal: the revised Clinical Institute Withdrawal Assessment for Alcohol scale (CIWA-Ar). Br J Addict 1989 Nov;84(11):1353-1357. [doi: 10.1111/j.1360-0443.1989.tb00737.x] [Medline: 2597811]
50. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
51. Booth A, Hannes K, Harden A, Noyes J, Harris J, Tong A. COREQ (Consolidated Criteria for Reporting Qualitative Studies). In: Guidelines for Reporting Health Research: A User's Manual 2014:214-226. [doi: 10.1002/9781118715598]
52. Bertholet N, Cunningham JA, Faouzi M, et al. Internet - based brief intervention for young men with unhealthy alcohol use: a randomized controlled trial in a general population sample. Addiction 2015 Nov;110(11):1735-1743. [doi: 10.1111/add.13051]
53. Kraepelien M, Sundström C, Johansson M, Ivanova E. Digital psychological self-care for problematic alcohol use: feasibility of a new clinical concept. BJPsych Open 2023 May 24;9(3):e91. [doi: 10.1192/bjo.2023.73] [Medline: 37222099]
54. Bendtsen P, Bendtsen M, Karlsson N, White IR, McCambridge J. Online alcohol assessment and feedback for hazardous and harmful drinkers: findings from the AMADEUS-2 randomized controlled trial of routine practice in Swedish universities. J Med Internet Res 2015 Jul 9;17(7):e170. [doi: 10.2196/jmir.4020] [Medline: 26159179]
55. Blankers M, Koeter MWJ, Schippers GM. Internet therapy versus internet self-help versus no treatment for problematic alcohol use: a randomized controlled trial. J Consult Clin Psychol 2011;79(3):330-341. [doi: 10.1037/a0023498]
56. Magill M, Ray L, Kiluk B, et al. A meta-analysis of cognitive-behavioral therapy for alcohol or other drug use disorders: treatment efficacy by contrast condition. J Consult Clin Psychol 2019 Dec;87(12):1093-1105. [doi: 10.1037/ccp0000447] [Medline: 31599606]
57. Ray LA, Meredith LR, Kiluk BD, Walthers J, Carroll KM, Magill M. Combined pharmacotherapy and cognitive behavioral therapy for adults with alcohol or substance use disorders. JAMA Netw Open 2020 Jun 1;3(6):e208279. [doi: 10.1001/jamanetworkopen.2020.8279]
58. Augsburger M, Kaal E, Ülesoo T, et al. Effects of a minimal-guided on-line intervention for alcohol misuse in Estonia: a randomized controlled trial. Addiction 2022 Jan;117(1):108-117. [doi: 10.1111/add.15633] [Medline: 34184795]
59. Baumgartner C, Schaub MP, Wenger A, et al. "Take Care of You"—efficacy of integrated, minimal-guidance, internet-based self-help for reducing co-occurring alcohol misuse and depression symptoms in adults: results of a three-arm randomized

XSL•FO
RenderX

controlled trial. Drug Alcohol Depend 2021 Aug 1;225:108806. [doi: 10.1016/j.drugalcdep.2021.108806] [Medline: 34171823]
60. Kiluk BD, Ray LA, Walthers J, Bernstein M, Tonigan JS, Magill M. Technology-delivered cognitive-behavioral interventions for alcohol use: a meta-analysis. Alcohol Clin Exp Res 2019 Nov;43(11):2285-2295. [doi: 10.1111/acer.14189] [Medline: 31566787]
61. Schuckit MA, Hesselbrock V. Alcohol dependence and anxiety disorders: what is the relationship. Focus Tech Coop 2004;151(3):1723-1453. [doi: 10.1176/foc.2.3.440]
62. Boden JM, Fergusson DM. Alcohol and depression. Addiction 2011 May;106(5):906-914. [doi: 10.1111/j.1360-0443.2010.03351.x] [Medline: 21382111]
63. Chih MY, Patton T, McTavish FM, et al. Predictive modeling of addiction lapses in a mobile health application. J Subst Abuse Treat 2014 Jan;46(1):29-35. [doi: 10.1016/j.jsat.2013.08.004] [Medline: 24035143]
64. Bae S, Chung T, Ferreira D, Dey AK, Suffoletto B. Mobile phone sensors and supervised machine learning to identify alcohol use events in young adults: implications for just-in-time adaptive interventions. Addict Behav 2018 Aug;83:42-47. [doi: 10.1016/j.addbeh.2017.11.039] [Medline: 29217132]
65. Lindner P, Johansson M, Gajecki M, Berman AH. Using alcohol consumption diary data from an internet intervention for outcome and predictive modeling: a validation and machine learning study. BMC Med Res Methodol 2020 May 11;20(1):111. [doi: 10.1186/s12874-020-00995-z] [Medline: 32393245]
66. Kay-Lambkin F, Baker A, Lewin T, Carr V. Acceptability of a clinician-assisted computerized psychological intervention for comorbid mental health and substance use problems: treatment adherence data from a randomized controlled trial. J Med Internet Res 2011 Jan 27;13(1):e11. [doi: 10.2196/jmir.1522] [Medline: 21273184]
67. McGovern R, Kaner E. Alcohol prevention and treatment: interventions for hazardous, harmful, and dependent drinkers. In: The Handbook of Health Behavior Change, 4th edition: Springer Publishing Company; 2014:211-227.
68. Bertholet N, Daeppen JB, Wietlisbach V, Fleming M, Burnand B. Reduction of alcohol consumption by brief alcohol intervention in primary care. Arch Intern Med 2005 May 9;165(9):986. [doi: 10.1001/archinte.165.9.986]
69. Lapham GT, Hawkins EJ, Chavez LJ, et al. Feedback from recently returned veterans on an anonymous web-based brief alcohol intervention. Addict Sci Clin Pract 2012 Aug 28;7(1):17. [doi: 10.1186/1940-0640-7-17] [Medline: 23186354]
70. Saunders JB, Kypri K, Walters ST, Laforge RG, Larimer ME. Approaches to brief intervention for hazardous drinking in young people. Alcoholism Clin Exp Res 2004 Feb;28(2):322-329. [doi: 10.1097/01.ALC.0000113418.12889.08]
71. Fischer LC, Kölligan V, Wieland N, Klein M. Development and evaluation of a digital health intervention for substance use reduction in young refugees with problematic use of alcohol and/or cannabis-study protocol for a single-armed feasibility trial. Front Public Health 2021;9:557431. [doi: 10.3389/fpubh.2021.557431] [Medline: 33869122]
72. Boendermaker WJ, Boffo M, Wiers RW. Exploring elements of fun to motivate youth to do cognitive bias modification. Games Health J 2015 Dec;4(6):434-443. [doi: 10.1089/g4h.2015.0053] [Medline: 26421349]
73. Prins PJM, Brink ET, Dovis S, et al. “Braingame Brian”: toward an executive function training program with game elements for children with ADHD and cognitive control problems. Games Health J 2013 Feb;2(1):44-49. [doi: 10.1089/g4h.2013.0004]
74. Glass JE, Tiffany B, Matson TE, et al. Approaches for implementing digital interventions for alcohol use disorders in primary care: a qualitative, user-centered design study. Implement Res Pract 2022;3:26334895221135264. [doi: 10.1177/26334895221135264] [Medline: 37091101]

Abbreviations

AUDIT: Alcohol Use Disorders Identification Test
CBT: cognitive behavioral therapy
CIWA-Ar: Clinical Institute Withdrawal Assessment of Alcohol Scale—Revised
COREQ: Consolidated Criteria for Reporting Qualitative Research
GAD-7: Generalized Anxiety Disorder-7
PHQ-9: Patient Health Questionnaire-9
PSS: Perceived Stress Scale
RCT: randomized controlled trial

XSL•FO
RenderX

Edited by A Mavragani; submitted 22.07.24; peer-reviewed by S Nkosi; revised version received 02.12.24; accepted 02.12.24; published 19.03.25.

Please cite as:
Kwon M, Moon DU, Kang M, Jung YC
A Brief Cognitive Behavioral Therapy–Based Digital Intervention for Reducing Hazardous Alcohol Use in South Korea: Development and Prospective Pilot Study
JMIR Form Res 2025;9:e64459
URL: https://formative.jmir.org/2024/1/e64459
doi:10.2196/64459

© Manjae Kwon, Daa Un Moon, Minjae Kang, Young-Chul Jung. Originally published in JMIR Formative Research (https://formative.jmir.org), 19.3.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Exploring the Use of an Augmented Reality Device Learning Tool for Multidisciplinary Staff Training on Domestic Abuse and Sexual Violence: Postintervention Qualitative Evaluation

Dilroshini Karunaratne[1], MBChB; Jessica Whittock[2], BSc, MA; Amber Moore[1], MBBS, BSc; Krishna Dasigan[1], MBBS, MRCP; Jasmine Chevolleau[1], MBChB; Brent Bartholomew[3], MBChB, BSc, MRCP; Nikki Kelly[4]; Charlotte E Cohen[5], MBBS, BSc, MRCP

[1]Department of Medical Education, Chelsea and Westminster Hospital NHS Foundation Trust, London, United Kingdom

[2]Department of Safeguarding Adults and Children, Chelsea and Westminster NHS Foundation Trust, Standing Together Against Domestic Abuse, London, United Kingdom

[3]Department of Acute Medicine and Emergency Services, Faculty of Medicine, Chelsea and Westminster Hospital NHS Foundation Trust, London, United Kingdom

[4]The SafeLives Practice Team, SafeLives, London, United Kingdom

[5]Department of Genitourinary Medicine and HIV Services, Faculty of Medicine, Chelsea and Westminster Hospital NHS Foundation Trust, London, United Kingdom

Corresponding Author:
Dilroshini Karunaratne, MBChB
Department of Medical Education, Chelsea and Westminster Hospital NHS Foundation Trust, London, United Kingdom

Abstract

Background: Legislative policies published by National Health Service, England and the UK Government focus on prioritizing the creation of a stronger system. These frameworks emphasize on the improvement of health care staff's ability to identify and refer domestic abuse (DA) survivors as key areas for supporting workforce development. Health care staff are often the first professional contact of survivors of DA, and insufficient staff training is a key barrier to survivors being identified and directed to support. The Microsoft HoloLens2 is a mixed-reality headset that allows virtual objects (holograms) to be integrated into the real world. Mixed-reality headsets are being increasingly used within medical education and have the advantage of independent operation, reducing the staffing requirements for teaching. The HoloLens2 can be used to project HoloPatients (HPs), which resemble clinically unwell patients, into the classroom. Two of these HPs have been specifically designed to portray survivors of DA and sexual violence (SV).

Objective: This study explored potential uses of the HP in DA and SV training as a potential survivor-centered educational initiative that could be used as an adjunct to existing training for health care professionals and community sector workers.

Methods: Frontline staff and community stakeholders from the national health service, DA, and law enforcement sectors were invited on 3 separate occasions (n=14, 12, 22) to a HoloLens2 demonstration that displayed 9 HPs. The patient voice was to be outlined by personalized scripts, co-created alongside sector charities, ensuring survivor engagement and participation. Participants were given the opportunity to wear the headset and familiarize themselves with the technology during the sessions. A post-intervention evaluation research model was used to explore the feasibility and functionality of the HP as an educational tool.

Results: Thematic analysis described the HP as a "realistic," "adjustable" tool that "creates a safe learning environment." Participants suggested it could be useful in "pre-exposure preparation" by "improving communication" and allowing different approaches to be trialed in a safe environment. The use of survivor scripts was described as a useful tool to "bring the survivor into the learning space" in a safe way. Participants identified the HP as a suitable tool for workers inside and outside health care, including social sectors such as law enforcement (32%).

Conclusions: The HP acts as a low-risk, adaptable tool for trainees to develop skills in a safe environment. This study demonstrates that professionals perceived the HoloLens as an innovative means to amplify the lived experience voice. Further research will evaluate this additional impact on trainees' confidence and responses to survivors disclosing DA and SV within different disciplines to drive improved outcomes.

Trial Registration: Permission to conduct this study granted by the Research Innovation and Quality Improvement Team at Chelsea and Westminster NHS Foundation Trust

(JMIR Form Res 2025;9:e60075) doi:10.2196/60075

XSL•FO
RenderX

KEYWORDS

augmented reality; virtual reality; medical technology; domestic abuse; sexual violence; medical education; training; domestic violence; violence; assault; victim; survivor-centered; staff; community stakeholders; social care; innovation

Introduction

In the UK, it is estimated that nearly 16.6% of adults aged 16 years and older have experienced sexual assault [1]. Furthermore, it is estimated that 2.1 million people aged 16 years and older experienced domestic abuse (DA) in the year ending March 2023 [2]. These statistics are comparable to the number of patients currently living with type 2 diabetes mellitus [3]. However, unlike diabetes and other medical conditions, there is little evidence discussing the pathogenesis behind DA and sexual assault, thereby making prevention programs, both difficult to design and sustainably fund. With one of the highest rates of repeat victimization than any other crime [4], DA affects 1 in 5 adults during their lifetime [5]. In addition, the detrimental effects of DA are wide-ranging, encompassing physical, mental, behavioral, and reproductive consequences [6]. Survivors of DA seek health care more frequently than individuals who have not experienced abuse [7], varying from presentations related to the physical consequences of abuse to psychological health problems, including depression, anxiety, and post-traumatic stress disorder [8]. Therefore, health care providers have a key role in identifying and supporting survivors. A systematic review of barriers and facilitators to disclosure of abuse to health care professionals (HCPs) found that perception of the HCP's ability to respond to the disclosure was impactful. Any notion that the HCPs did not possess the skills or knowledge to appropriately respond deterred survivors from disclosing their experiences [8]. HCPs are often the first point of contact for survivors; hence, it is vital that they are equipped with the skills and tools to competently identify and support survivors [9].

According to the UK Government, all behaviors, inflicted on anyone perceived to be of a controlling, coercive, harmful, or sexual nature can be defined as DA or sexual violence (SV) [10]. Whilst the majority of survivors of DA are women [2], it is important that we do not neglect the experiences of those from the LGBTQ+ (lesbian, gay, bisexual, transgender, and queer) community, who are also more vulnerable to familial abuse and honor-based violence [11]. There are many barriers to LGBTQ+ individuals seeking support, including discriminative attitudes from the police leading to mistrust, as well as a lack of training for professionals from the criminal justice system and public sector on how to recognize and respond to DA in the LGBTQ+ community [12]. Although the data for England and Wales for the year ending March 2024 showed no statistically significant differences in the prevalence of domestic violence between ethnic groups [2], the literature shows that Black, Asian, and minority ethnic individuals experience multiple barriers in seeking help following abuse, including structural racism, cultural exclusion, and victim-blaming from both formal services and communities [13]. Furthermore, the 2024 statistics for England and Wales showed that a higher proportion of people with a disability experienced DA than those without [2]. These issues should not be considered as separate entities, but rather an intersectional approach must be undertaken, so we can understand how inequalities due to race, sex, disability, class, and sexual identity may compound each other, further preventing survivors from seeking help [11]. These statistics outline the need for services to adopt a survivor-centered approach. Strategic partnerships consulting survivors of all protected characteristics are needed to enable the development of pathways specifically designed to address their psychosocial needs. This is outlined in The Strategic Direction for Sexual Assault and Abuse Services [14], which is an initiative from National Health Service (NHS) England, with the goal of radically improving access to services and support for survivors of sexual abuse to help them heal and rebuild their lives. This requires collaboration between authorities, health services and strategic partners in the voluntary and community sectors as part of a multiagency care model [15].

Existing literature suggests a gap in the knowledge of HCPs on how to identify and respond to survivors of domestic violence and sexual assault, largely due to a lack of training [16,17]. A 2020 metasynthesis of 47 qualitative studies examined health practitioners' readiness to address domestic violence and abuse. Support from the health system was the largest theme, with HCPs suggesting further support was needed through upskilling staff in how to address domestic violence, as well as making enquiries into domestic violence routine and allocating additional time for sensitive conversations [16]. HCPs spoke about the importance of training in how to identify and respond to DA in their own specific clinical setting [16]. Evaluations of existing domestic violence support training programs for HCPs have demonstrated significant improvements in knowledge and confidence in how to identify and respond to survivors [17] and hence, we set out to design a training program adjunct, using a mixed reality device.

The emerging technologies of virtual, augmented, and mixed reality devices create an expansive new realm for education. Virtual reality is an immersive experience where a virtual world replaces the physical world, whereas augmented reality uses technology to superimpose virtual data (visual information) onto the physical world (real-life objects) [18]. Mixed reality is an in-between medium which allows virtual objects to be integrated and anchored onto the physical world [18]. The use of mixed reality headsets, such as the Microsoft Hololens2, is being increasingly used within medical education. There are many advantages to using these devices, including the option of repeated practice of learning, skills, and communication, without adverse effects on the patient [19]. These devices also can be used independently, reducing the time burden of teaching sessions on clinical educators. So far, mixed reality devices have been primarily used for teaching anatomy (through understanding of 3D object layers and components), procedures, and clinical skills [19-21]. We wanted to explore the benefits of using them for DA and SV training. In order to effectively teach HCPs about domestic violence and sexual assault, it is important to use examples based on real-life experiences of

XSL•FO
RenderX

survivors. To ask a survivor to facilitate a training session could be considered unethical, as it may be triggering to the survivor, with potential negative effects on their mental health and well-being; hence, the use of mixed reality devices creates an opportunity to recreate these learning opportunities without adverse effects on survivors.

The Domestic Abuse team at Chelsea & Westminster Hospital NHS Foundation Trust adopts a "whole health" approach by pioneering a longstanding partnership between health care and community stakeholders [22] widely accepted as a best response in acute health. Their model centers on 3 key features: a co-located Domestic Abuse Co-ordinator (DAC) responsible for training workers in DA; the co-location of Independent Domestic Violence Advisors (IDVAs) to provide holistic and specialist support to survivors that are identified; and a DA Lead. Data suggest that survivors 'felt safe' and were more likely to report to an on-site IDVA while in the hospital [23]. Early identification of potential survivors by frontline workers is essential and stresses the importance of high-quality in-house training on the identification, management and referral of survivors of DA and SV.

The aim of this study was to improve the quality of staff training on the identification and management of DA and SV survivors. In collaboration with the Undergraduate Education Team at Chelsea and Westminster Hospital NHS Foundation Trust, the DA Team proposed the use of the Microsoft HoloLens2 as a tool to channel the survivor voice. This evaluation explores both health care and community stakeholder views of the augmented reality headset as an additional learning tool to promote survivor-centered training in a safe setting. We hypothesized that the following experience would demonstrate the Microsoft HoloLens' potential as a teaching adjunct when educating staff on the identification and management of DA and SV survivors.

Methods

We conducted a formative evaluation of the Microsoft HoloLens2, specifically the GiGXR HoloPatient (HP) application, as a training tool within the DA and SV landscape.

Intervention

The Microsoft HoloLens2 is a mixed reality headset which has had a substantial impact in various industries since its release in 2019, including health care, education, and engineering [24]. Dissimilar to virtual reality, the headset is only part-immersive as it displays augmented content superimposed on the real surrounding environment. Organizations like GiGXR have created a series of augmented holograms, known as HPs, that resemble clinically unwell patients. The HP menu listed 17 holograms, each representing a clinical emergency, 2 of which were specifically designed to portray a survivor of DA (Lydia Johnson) and SV (Jenny Li). Each case consists of a series of video clips that depict the patient at different stages of their hospital journey (Figure 1). Aside from Jenny Li who vocalizes her experience via a prerecorded audio clip, all other holograms are nonverbal. As a result, the user has the flexibility to adapt the HP by overlaying a different patient script to suit their educational needs. This application, if used effectively, could expose students to unique training opportunities that are ethically challenging and difficult to recreate in a real clinical environment.

XSL•FO
RenderX

Figure 1. Lydia Johnson, a HoloPatient survivor of domestic abuse and sexual violence. Image A demonstrates Lydia on arrival to the emergency department. Image B displays her as a patient on a ward.

Recruitment

The evaluation was conducted in the form of live demonstrations. These demonstrations were held within a lecture theater at Chelsea and Westminster Hospital. Frontline staff and community stakeholders from the NHS, DA, social care, and law enforcement sectors within the North West London area were invited to 3 HP events, held between February and July 2023 (see spider diagram displaying the range of attendees; Figure 2). Participants from these sectors were invited due to their increased likelihood of encountering survivors in a professional setting and their shared use of North West London Safeguarding guidelines. All participants were invited by email, which described the aim and method of the study in detail (see recruitment email in Multimedia Appendix 1). Participants were reminded that participation is voluntary and that all data were to be anonymized on publication. Written consent was obtained at the stakeholder meeting after participants were given the opportunity to read further information on the HoloLens and its purpose in this study (see consent form in Multimedia Appendix 2).

XSL•FO
RenderX

Figure 2. Spider diagram displaying all the organizations within North West London which attended the stakeholder forums.

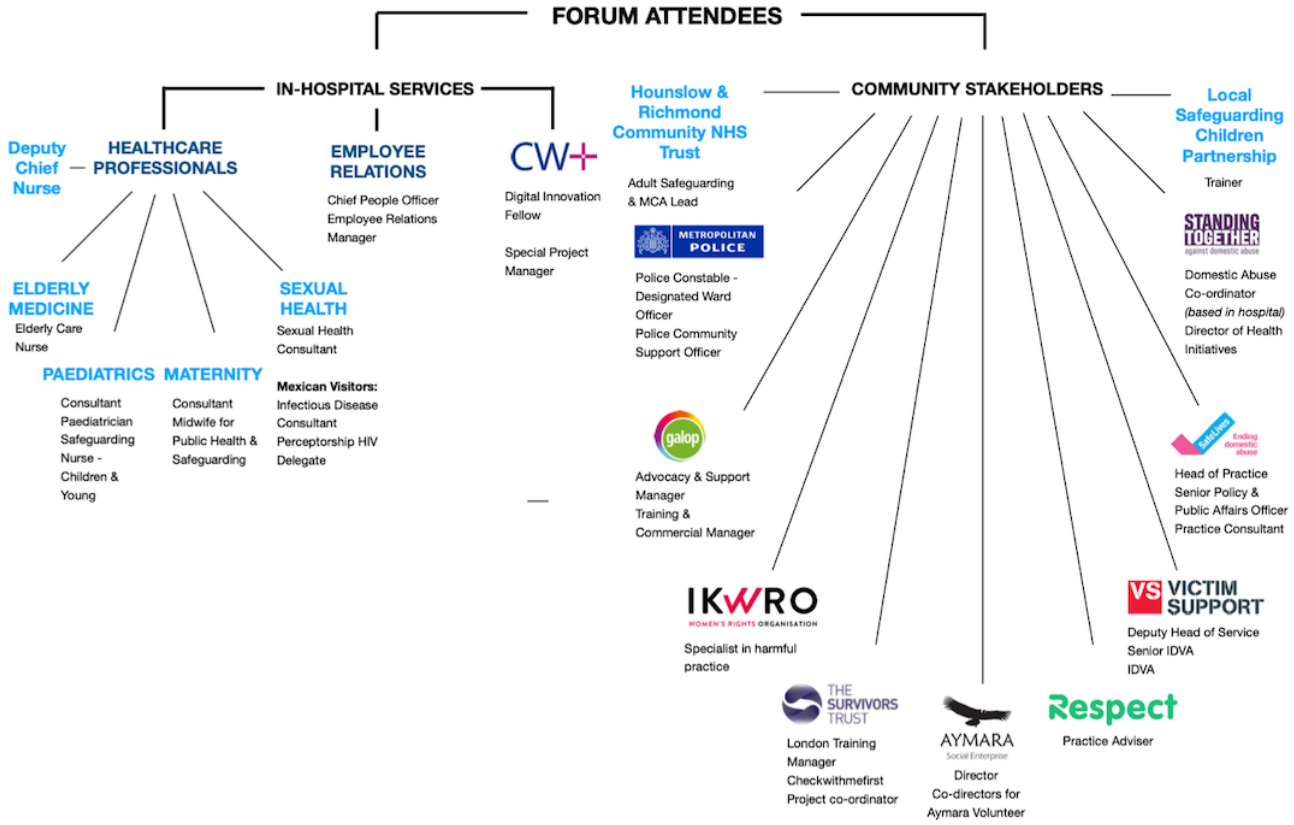

The Demonstrations

Following a brief introduction into augmented reality, attendees were shown data that evidenced the HP as an immersive tool when trialed in an undergraduate setting [25]. The HoloLens was ultimately proposed as a teaching adjunct to be used whilst training frontline workers on the management of DA and SV cases (see full presentation in Multimedia Appendix 3). It was envisaged to be a visual aid that amplified the lived experience, to be used after their initial training with the DAC. The patient voice was to be outlined by personalized scripts, co-created alongside sector charities, ensuring survivor engagement and participation. In total, 9 HPs, different in gender, age, and ethnicity, were projected into the room (Figure 3). Participants were given the opportunity to wear the headset and familiarize themselves with the technology during the sessions.

XSL•FO
RenderX

Figure 3. Nine HoloPatients selected to be displayed during the demonstrations. Their original clinical presentation demarcated in blue writing within each box.

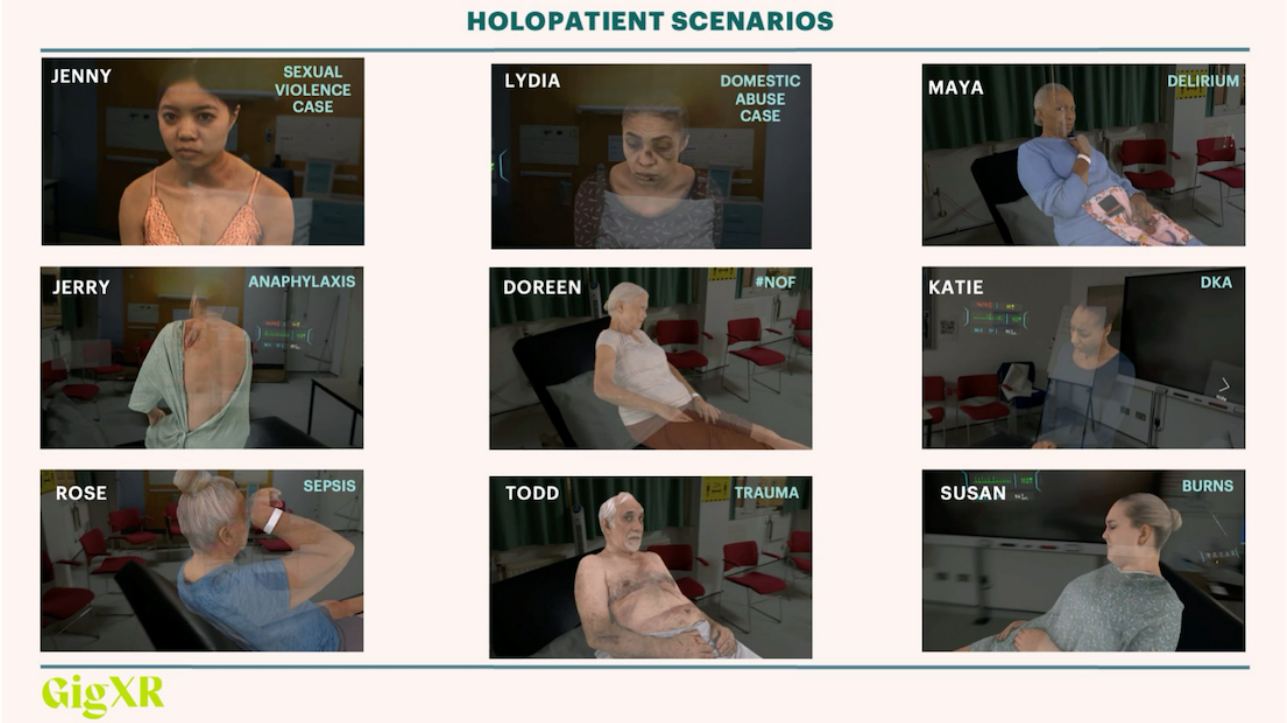

Data Collection and Analysis

A postintervention evaluation research model was used to explore the feasibility and functionality of the HP as an educational tool within DA and SV training. Furthermore, it evaluated the potential for a centralized multiagency augmented reality hub for professionals to access educational opportunities in the future. In order to achieve this, participants were prompted to answer 3 overarching questions: why should the HP be used, what should the HP be used for, and who should use the HP?

Data were collected in the form of a standardized questionnaire to improve reliability (see in Multimedia Appendix 4). It consisted of Likert scaling [26] and open-ended questions to collect quantitative and qualitative data, respectively. Likert scales were used for participants to rate how strongly they agreed or disagreed with the following statements: “HoloLens is an effective tool used in DA and SV,” and “Technological limitations prevent me from using this tool.” Open-ended questions gave participants the opportunity to provide feedback on how they felt wearing the device, if they thought it would be useful in health care or other fields, and if they felt there was value in having a centralized multiagency HoloLens training hub. A basic analysis conducted on the Likert scale questions displayed statistical means and modes. The qualitative data were thematically analyzed using Braun and Clark’s 6-step method [27] (see data sets in Multimedia Appendices 5, 6, and 7). One researcher (medically qualified) read each survey repeatedly to gain familiarity. Initial interpretations were discussed with the project supervisor, and together they extracted evidence from the surveys to refine each theme. Finally, a list of themes was presented with illustrative examples on multiple occasions to the wider team (see coding tree in Figure 4). As a final checking step, draft findings were discussed with a few participants (n=3) prior to publication. To ensure that this study was conducted in accordance with qualitative reporting guidelines, a COREQ checklist was completed (Checklist 1) [28].

XSL•FO
RenderX

Figure 4. Coding tree across all 3 stakeholder forums.

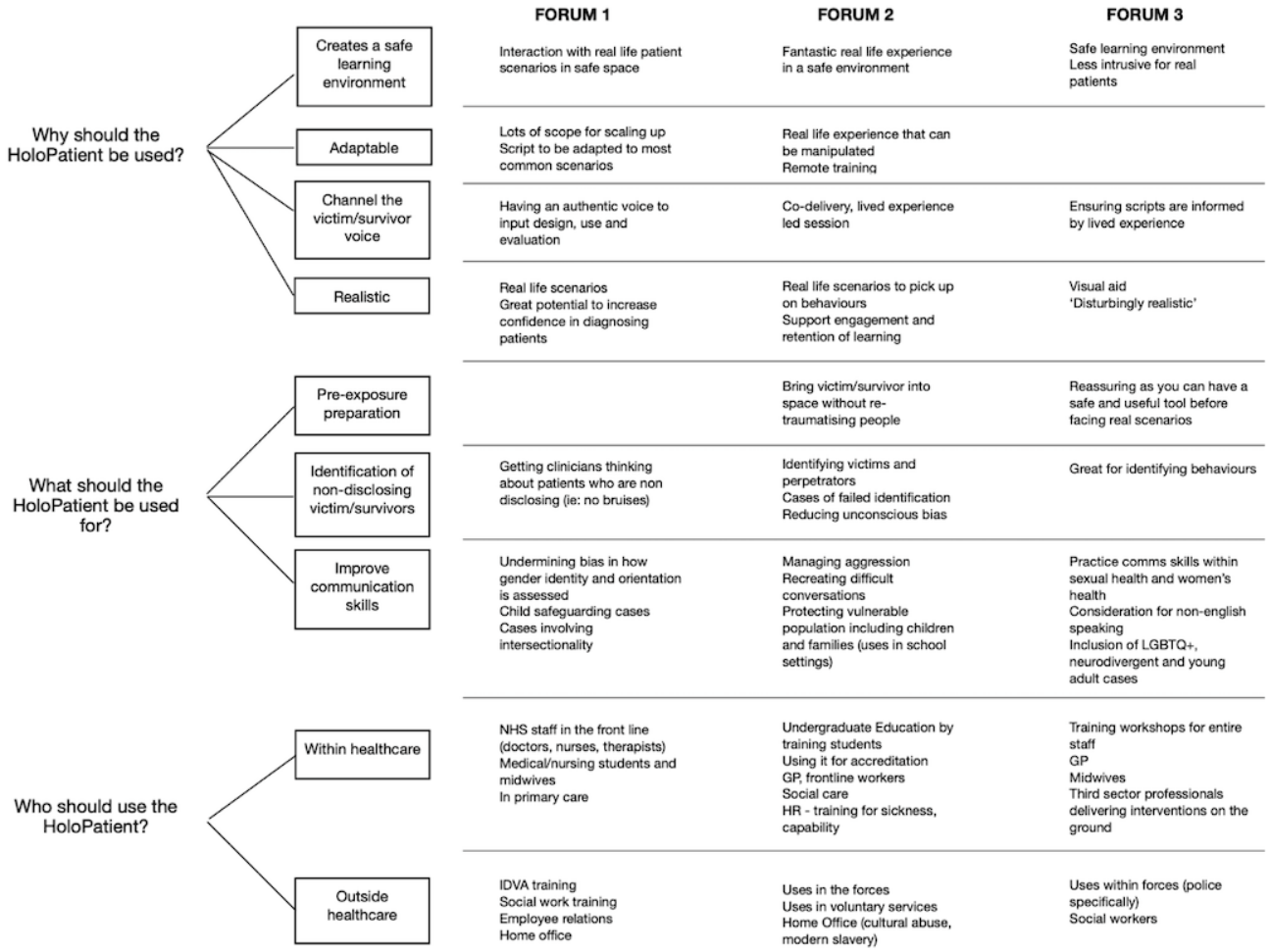

Ethical Considerations

This evaluation was approved by the Research Innovation and Quality Improvement board of Chelsea and Westminster Hospital NHS Foundation Trust under reference number 6038. Written informed consent was obtained from all participants. All study data are anonymous, and HP scripts, although informed by DA and sexual assault survivors, did not contain any identifiable information. Involvement in this study was optional and voluntary, and there was no financial compensation for participants or researchers.

Results

Thematic Analysis

A total of 48 questionnaires were collected across all 3 demonstrations (n=14, 12, 22). Approximately 29% (n=14) of the sample were local HCPs. The remainder consisted of representatives from community charities, local councils, and law enforcement (see spider diagram displaying range of attendees in Figure 2). No participants refused to participate or dropped out.

Analysis of participants' perspectives of the HP as a training tool in DA and SV sectors was divided into 3 overarching themes, based on our research questions: Why should the HP be used? What should the HP be used for? Who should use the HP?

A summary of this section is displayed in Table 1.

XSL•FO
RenderX

Table . Realistic table summarizing qualitative data following thematic analysis

Theme	Reason
Why should the HoloPatient be used?	
Realistic	• "disturbingly realistic" • "impactful" • "increased their sense of reality" • improving "retention of learning" • "more realistic than role-play"
Creates a safe learning environment	• brought "real-life learning experiences into a safe environment" • trainees can repeatedly interact, make errors, stumble • openly discuss case with peers without risk of re-traumatization
Adaptable	• allows "real-life experiences to be manipulated" • "scenarios adapted to each trainee's requirement" • "helpful for all professionals' • remote feature allows "scope to scale up"
Channel the survivor voice	Use of survivor script • "essential" feature that could "bring the survivor into the learning space" • to "incorporate language and cultural specific aspects" • "inclusion of survivors when creating the stories" • cover "common scenarios in different clinical settings and diverse cases"
What should the HoloPatient be used for?	
Pre-exposure preparation	• "safe and useful tool for practice before facing real-life scenarios" "important to understand how to deal with the situation before interacting with survivors" • law enforcement felt this could "support junior officers responding to their first call outs in a safer space."
Identification of nondisclosing survivors	• prepare trainees to identify survivors presenting without visible injury • "masking patients" • those "with a hidden agenda" • "great for identifying behaviors or visible indicators"
Improving communication skills with and when	
Vulnerable patients	• elderly • mental illness • neurodiversity • women • young people • NHS staff or other staff
LGBTQ + survivors	• "look at undermining bias in how people are assessed in regards to their their gender identity and sexual orientation" • support exploration of "intimate partner violence" and "hate crime" • have an "LGBTQ + lived experience voice" in the scenarios
Engaging with perpertrators	• "look at undermining bias in how people are assessed in regards to their their gender identity and sexual orientation"
Scenario could be adapted to	• where the survivor is "aggressive and protective of the perpetrator" • "on how to approach a survivor and respond appropriately with a perpetrator present"
Recreating difficult conversations	• "practicing mental capacity assessments" • "decision-making, end-of-life care planning" • "dealing with confrontation and bullying and harassment cases" • demonstrate "an example of an incorrect way to deal with a situation" • encourage "open discussion"

XSL•FO
RenderX

Theme	Reason
Who should use the HoloPatient?	
Within health care	
Undergraduate education	• "medical and nursing students" • training "first with the HoloPatient and then a certificate before the real practice"
Maternity	• "the risk of domestic abuse during pregnancy" • "trust building that happens between patient and midwife"
Other frontline staff	• 'all health professionals who may come into contact with a survivor' • 'no them and us' • targeting 'third sector professionals who intend to deliver interventions on the ground' • could be used in IDVA[a] training to 'significantly reduce the time taken to train'
Outside of health care	
Social care sector	• social workers • paid carers • other care home staff
Public sector	• within law enforcement • Home Office to recreate cases involving "minority ethnic survivors," "honour-based abuse" and "modern slavery.'

[a]IDVA: Independent Domestic Violence Advisor

Why Should the HoloPatient Be Used?

Realistic

Many attendees found the HP "disturbingly realistic" (Participant 1 (P1)) and were impressed with how "impactful" it was (P2, P3, and P4) (participant references taken from data sets—see Multimedia Appendices 5, 6, and 7). Attendees found that the HP "increased their sense of reality" (P5 and P6) hence improving engagement and "retention of learning" (P7). One participant found the HP "more realistic than role-play" (P2), as dynamics could be affected if a trainee is familiar with their actor or mentor.

Creates a Safe Learning Environment

Participants felt the HP brought "real-life learning experiences into a safe environment" (P8), where both trainees and survivors are better protected. In contrast to role-play, trainees can repeatedly interact, make errors, stumble, and openly discuss the HP's case with peers without risk of re-traumatization.

Adaptable

Attendees acknowledged that the HP allows "real-life experiences to be manipulated" (P9) by using "scenarios adapted to each trainee's requirement" (P1). As a result, a majority of participants saw it as a tool helpful for all professionals. HP demonstrations were live-streamed via Microsoft Teams to remote participants. One participant called this feature a "key benefit" (P10), which added "scope to scale up."

Channel the Survivor Voice

The concept of the survivor script was presented to all participants at the HoloLens demonstrations. The scripts received positive feedback, with stakeholders seeing it as an "essential" feature that could "bring the survivor into the learning space" (P11). The idea was to overlay the HP scenario with a voice-over, either prerecorded or delivered by a survivor facilitator on the day. Participants particularly encouraged the use of the facilitator's voice "to ensure that nuance isn't lost" (P9). The potential to create scripts that cover "common scenarios in different clinical settings (eg, Emergency Department, sexual health)" (P12) and "diverse cases" with different protected characteristics were discussed. The idea to "incorporate language and cultural specific aspects" (P7) into the scripts was strongly considered. Attendees agreed with the "inclusion of survivors when creating the stories" to ensure that the scripts are "informed by a lived experience" (P3), with agreement that "specialist domestic abuse" (P2) input was necessary to maintain the use of trauma-informed language. The creation of "an authentic voice advisory group" (P13) for their overall evaluation was suggested.

Overall, all stakeholders strongly agreed (n=45) or somewhat agreed (n=3) that the Microsoft HoloLens 2 could be an effective tool to deliver training on the identification and management of DA and SV survivors.

What Should the HoloPatient Be Used For?

Pre-Exposure Preparation

Stakeholders saw the HoloLens as a "safe and useful tool for practice before facing real-life scenarios" (P1). One participant described Jenny, the hologram of a SV survivor, as "incredibly uncomfortable to watch" (P14) explaining that it was "important to understand how to deal with the situation before interacting with survivors" (P14). Themes of realism and the HP's ability

XSL•FO
RenderX

to create a safe learning environment were commonly mentioned in conjunction with this theme. This is something that respondents cited clinicians often do not have access to, following any training. The HoloLens offers a training tool to embed the teaching from the DAC, before practicing in real life with patients. Law enforcement similarly felt this could support junior officers responding to their first call-outs in a safer space.

Identification of Nondisclosing Survivors

The HP's potential to prepare trainees to identify survivors presenting without visible injury became a reoccurring theme across all demonstrations. One participant described the device as "great for identifying behaviors or visible indicators" (P2).

Improve Communication Skills

Communicating With Vulnerable patients

Once all the HPs were displayed, participants were prompted to suggest potential scenarios that could aid the identification of vulnerable survivors of DA or SV. Public Health England defines vulnerability as "in need of special care, support, or protection because of age, disability, risk of abuse, or neglect" [29]. Categorized into these subgroups, the suggested scenarios are listed below in ascending order of prevalence based on the questionnaire data: Elderly patients (n=2): "delirium," "care-home scenarios"; women (n=4): "pregnant women," "female genital mutilation"; patients with mental illness (n=7): "suicidal tendencies," "eating disorders," and "psychosis"; patients with visible and nonvisible disability (n=8): "scenarios that include adults with learning disability," and "involving paid carers"; children and adolescents (n=16): "whole family scenarios," "involving teachers," "preparing families for court proceedings," "forced marriage," and "substance abuse".

Communicating With LGBTQ+ Survivors

A participant stated that "generalist staff have not always got the skills to deal with LGBTQ+ survivors" (P15). In response, they suggested the potential use of the HP to "look at undermining bias in how people are assessed in regards to their gender identity and sexual orientation" (P15). They saw how this device could support exploration of "intimate partner violence" and "hate crime" by editing the scenarios to have an "LGBTQ+ lived experience voice" (P15). A few participants highlighted the absence of any transgender HPs (P1, P15, P16, and P17). When asked to suggest designs for new HPs, 22% (n=11) of participants said LGBTQ+ HPs.

Dealing With Perpetrators

In the first stakeholder meeting, a scenario was adapted to include a perpetrator by getting a facilitator to interact with the HP by standing in the background. They are visible to the HoloLens user and all participants watching the live stream. It was suggested that the scenario could be adapted to "challenge the stereotype of a passive survivor" (P9) where the survivor is "aggressive and protective of the perpetrator" (P9). Stakeholders saw its training potential within the police force (n=19), "specifically on how to approach a survivor and respond appropriately with a perpetrator present" (P3). However, concerns with safety were highlighted. One participant stated that "we need to be clear on the expectation around engaging with a perpetrator and how to do it safely if at all" (P9). This was seen as a key advantage of the HoloLens tool, as this could be explored in a safe space.

Recreating Difficult Conversations

Stakeholders listed "practicing mental capacity assessments, decision-making, end-of-life care planning, dealing with confrontation, and bullying and harassment cases" (P18) as challenging scenarios that could be practiced with the HoloLens. One participant suggested to display "an example of an incorrect way to deal with a situation" (P8) as a teaching tool to encourage "open discussion" (P18).

Used in a Central Multiagency Training Hub

During the final demonstration, which took place in July 2023, attendees (n=22) were asked whether they supported the creation of a central multiagency training hub, potentially at Chelsea and Westminster NHS Foundation Trusts, which incorporated the use of augmented and virtual reality headsets. The intention of the hub is to provide community and health care services within the North West London region, shared access to the technology, and other training materials. This allows for a standardized training approach with greater opportunity for communication between teams. Overall creating more collaborative DA and SV multi-agency services in the future. A majority of participants either strongly (n=17) or somewhat agreed (n=2) in the creation of a central hub. It was stated that "sculpting multiagency responses into this (central hub) would be incredible and will change the health care landscape. The HoloLens can be used to continuously improve practice by dropping into an immersion center" (P4). One participant neither agreed nor disagreed with the hub, and 2 participants did not vote.

Who Should Use the HoloPatient?

Within Health Care

Undergraduate Education

When asked where this technology could be most useful, approximately 14% (n=7) of participants suggested its use in undergraduate education. In particular, "medical and nursing students" (P5, P18, P19, P20, P21, and P22). One attendee suggested using the HoloLens for accreditation by training "first with the HP and then a certificate before the real practice" (P20).

Maternity

At least one participant per demonstration (total n=3) acknowledged its need in midwifery due to "the risk of DA during pregnancy and the trust building that happens between patient and midwife. Additional attendees (n=5) saw potential for the HP to be used in "IDVA training" (P4), indicating it could significantly reduce the time taken to train if this were used in routine practice. IDVAs are trained specialists that work with clients "from the point of crisis to assess, discuss, and develop safety plans"[30].

Other Frontline Staff

Approximately 52% (n=25) of attendees felt as though the technology was suitable for "all health professionals who may come into contact with a survivor" (P9, P23). "No them and

XSL•FO
RenderX

us" (P24) was stated by one participant. Particular attention was given to primary care (n=6), emergency, and sexual health service staff. Other scoping included targeting "third sector professionals who intend to deliver interventions on the ground" (P7). One participant stated that training "could be in real-time with the holograms" (P5), providing clinicians with "authentic role-play" (P5) that was "less intrusive than real patients" (P5). This is a unique concept when compared to existing training programs, which mainly consist of digital modules at a beginner level [31].

Outside of Health Care

Social Care Sector

At all 3 demonstrations, over one-third of attendees (n=16) named social workers, paid carers, and other care home staff as members of the public who would benefit from DA and SV training with the HoloLens.

Public Sector

Similarly, around 39% (n=19) of participants described the HP's potential to train staff within law enforcement. Particularly those who "attend a domestic abuse emergency" (P12).

The use of the technology in the Home Office was discussed, especially to recreate cases involving "minority ethnic survivors" (P18), "honor-based abuse" (P2, P24, and P25), and "modern slavery" (P17 and P18).

Discussion

Study Findings and Comparison With Previous Works

The findings of this stakeholder evaluation indicate that despite minor technological limitations, all stakeholders saw the Microsoft HoloLens2 as a novel and effective tool to deliver training on DA and SV, within health, social care, law enforcement, and specialist settings. The data encourage the use of co-created survivor scripts alongside the HPs, to create a training program that amplifies their voice and lived experience. This enables the training to remain survivor-centered without needing a survivor present in the room as a facilitator. This minimizes the potential for adverse effects on the survivor.

A common theme across the groups was the ability to identify behaviors that signaled abuse in patients who did not have obvious physical signs of abuse such as bruising or other physical injury. The literature highlights how survivors often do not freely disclose abuse and often actively try to hide it [7], and hence being able to identify the more subtle signs is key to having good and comprehensive training on how to identify survivors. Positive feedback on the ease of use of the HoloLens device was received from all 3 groups, spanning health care workers, law enforcement, and the charity sector, demonstrating that this is a widely acceptable medium to deliver training.

Stakeholders suggested that the HPs were useful for pre-exposure preparation, which aligns with how we envisage it being used in the future. There was a strong theme of realism, with participants describing how the holograms had a profound emotional impact, which we hope will mirror some of the thoughts and feelings experienced when encountering survivors in real life. Enabling this first encounter to occur in a safe and simulated environment empowers participants to trial different approaches to managing the scenario without the risk of harm to a survivor so that they are better equipped to appropriately handle these situations in real life.

The existing literature clearly shows that one of the key barriers to recognizing and supporting survivors is a lack of training [12,16,17]. Ideally, this training needs to be embedded into the undergraduate curriculum of HCPs, so that doctors and nurses graduate with the skills to appropriately manage these encounters. We have demonstrated the potential of using the HoloLens2 in a professional setting and hope this could also be used as a tool for educating undergraduate health care students.

Limitations

We discovered numerous limitations within our project, both related to the Hololens2 device itself and to the application. Wi-Fi connectivity poses a significant issue. One attendee stated that the HoloLens is a "useful tool when in a Wi-Fi range." To open the HP application on the Microsoft HoloLens headset, the device must be connected to a high-speed wireless network. This is a key limitation, as internet connectivity is not reliable, especially in the community. To overcome this, the Undergraduate Education Team at Chelsea and Westminster Hospital has used the smartphone "Wi-Fi hotspot" feature to supply a mobile connection. If the HPs are to be rolled out as a training program more widely, then considerations will need to be made into the creation of a downloadable version that can be operated without an internet connection.

We also found the limited battery life of the headset was a barrier to producing longer or back-to-back training sessions. Microsoft estimate the HoloLens' battery life to have 2 - 3 hours of active use [32]. As both the Microsoft Teams and GiGXR HP applications were open during the demonstrations, it caused the device to malfunction at approximately 1.5 hours. This was highlighted as "an issue" by participants. One user felt that it hindered their ability to "maximize opportunity when using it in the training room." Microsoft has stated that the devices are fully functional while charging [32]. Therefore, the Undergraduate Education Team inserted a portable battery pack into the USB-C port to prolong battery life during the demonstrations. Despite these technological limitations, 64% of participants felt as though it did not affect their opinion of the HoloLens. Overall statistics demonstrated that 14% stated otherwise. while 23% remained impartial. Another limitation of the device is that as it uses more power, it heats significantly. This poses potential safety risks and makes the device less comfortable to wear. To combat this, training sessions will need to be limited to shorter periods of time. We would suggest not using the device for more than an hour at a time.

A further limitation of the HP application was the lack of diversity of the characters. Most participants (roughly 57%, n=27) stated that they would have liked a wider variety of "visible clues" to represent intersectional characters [33]. For example, "different ethnicities, cultural headwear, visible disabilities, walking aids, transgender patients, pregnant women, and children." Unfortunately, this would require the HP menu to be redesigned, at significant cost, which remains a limiting

XSL•FO
RenderX

factor when considering scalability. However, augmented reality allows the hologram's background to be adjusted to the user's preference therefore walking aids and other visual accessories can be added, in an alternative manner or with a real-life person in the scenario, either as a friend, a family member or a perpetrator.

We must also consider limitations related to the study design. Data were collected as postintervention questionnaires. Although this provided insights into how the HPs could be used, the questionnaire was fairly brief, consisting of 13 questions. A focus group of a random sample of participants may have provided more detailed insights into the utility and acceptability of the Hololens2, and the HPs as a training tool for DA and SV. Furthermore, a larger sample size would allow for more reliable results.

Conclusions

This technology has the potential to educate professionals on the use of trauma-informed language while challenging preconceived assumptions around protected characteristics in safe settings. Our evaluation has demonstrated that the Hololens2 is a widely acceptable tool for DA and SV training, that can be used as an adjunct to existing training programs. There is currently very little research on how virtual, mixed, and augmented reality devices can be used for training in DA and SV; hence, we hope the positive results from our evaluation will make way for further large-scale studies [34]. As the HoloLens2 can be operated independently, it gives scope for training sessions to be delivered with minimal staffing requirements.

Virtual reality companies like SimX have developed DA-specific virtual reality patient simulations. In the future we would want to evaluate the efficacy of these modalities as a teaching adjunct in the DA and SV space with the aim to create a core framework applicable to all future mixed and virtual reality technologies [35,36]. It could be of additional direct benefit to employee relation team training around sexual misconduct following the release of NHS England's sexual safety organizational charter [37]. The goal will be to roll out augmented reality Holosurvivors as a standard teaching tool, in addition to DACs, supporting the best multiagency team public health response to DA and SV in the UK.

Acknowledgments

We would like to thank GiGXR, who own the rights to the application and holograms, for granting us permission to evaluate their product throughout this study. We wish to acknowledge our hospital charity CW+, who enabled the purchasing of the technology to assist with teaching and training in the Trust, supported the stakeholder forums, and will be collaborating on future research with our domestic abuse team. Additionally, we would like to thank our stakeholders; Standing Together Against Domestic Abuse, Victim Support, Galop, SafeLives, Respect, IKWRO Women's Rights Organisation, Aymara, Metropolitan Police, Adult Community Safeguarding, and Children's Social Care for their attendance and much valued feedback during the stakeholder forums. Finally, we would like to praise our Undergraduate Education Team and our Domestic Abuse team at Chelsea and Westminster NHS Foundation Trust for their hard work and dedication on this project.

The authors attest that there was no use of generative artificial intelligence (AI) technology in the generation of text, figures, or other informational content of this manuscript.

Data Availability

The authors confirm that the data supporting the findings of this study are available within the accompanying multimedia appendix (see Multimedia Appendix 5-Multimedia Appendix 7).

Authors' Contributions

DK, JW, AM, KD, BB, NK, and CC conceptualized the study. DK, AM, KD, BB, and CC contributed to methodology; validation was done by DK, BB, and CC. Formal analysis was done by DK and CC. Investigation was done by DK, JW, AM, KD, BB, NK, and CC. DK, JW, AM, KD, BB, NK, and CC contributed to resources. DK and CC involved in writing – original draft. DK, JC, and CC involved in writing – review & editing. Visualization was done by DK and CC. CC involved in supervision. DK and CC involved in project administration.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Sample recruitment email sent to stakeholders.
[PDF File, 102 KB - formative_v9i1e60075_app1.pdf]

Multimedia Appendix 2
Sample participant information and consent form.

XSL•FO
RenderX

[PDF File, 104 KB - formative_v9i1e60075_app2.pdf]

Multimedia Appendix 3
Powerpoint presentation delivered by Undergraduate Education Team during HoloLens demonstrations.
[PDF File, 3454 KB - formative_v9i1e60075_app3.pdf]

Multimedia Appendix 4
Sample questionnaire given to participants during the HoloLens demonstrations.
[PDF File, 106 KB - formative_v9i1e60075_app4.pdf]

Multimedia Appendix 5
Questionnaire data from the first HoloLens demonstration (23/2/23).
[XLSX File, 12 KB - formative_v9i1e60075_app5.xlsx]

Multimedia Appendix 6
Questionnaire data from the second HoloLens demonstration (26/4/23).
[XLSX File, 13 KB - formative_v9i1e60075_app6.xlsx]

Multimedia Appendix 7
Questionnaire data from the third HoloLens demonstration (13/7/23).
[XLSX File, 17 KB - formative_v9i1e60075_app7.xlsx]

Checklist 1
COREQ checklist
[PDF File, 396 KB - formative_v9i1e60075_app8.pdf]

References

1. Office for National Statistics. Sexual offences in England and Wales overview - Office for National Statistics. 2023. URL: https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/bulletins/sexualoffencesinenglandandwalesoverview/march2022 [accessed 2025-03-03]
2. Office for National Statistics. Domestic Abuse in England and Wales Overview. URL: https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/bulletins/domesticabuseinenglandandwalesoverview/november2023 [accessed 2024-12-09]
3. Diabetes UK. Diabetes is Serious. 2022. URL: https://diabetes-resources-production.s3.eu-west-1.amazonaws.com/resources-s3/public/2022-04/Diabetes%20is%20Serious%20Report%20Digital_0.pdf [accessed 2025-03-03]
4. Government Equalities Office. Strengthening women's voices in government: a response to the public consultation. 2011. URL: https://assets.publishing.service.gov.uk/media/5a78ae7b40f0b62b22cbbf11/strengthening-womens-voices.pdf [accessed 2024-02-28]
5. Domestic abuse in england and wales overview: November 2022 (2022). Census 2021. URL: https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/bulletins/domesticabuseinenglandandwalesoverview/november2022 [accessed 2024-04-17]
6. World Health Organization. Understanding and addressing violence against women: Intimate partner violence: World Health Organization. 2012. URL: https://apps.who.int/iris/handle/10665/77432 [accessed 2025-03-03]
7. Tarzia L, Bohren MA, Cameron J, et al. Women's experiences and expectations after disclosure of intimate partner abuse to a healthcare provider: a qualitative meta-synthesis. BMJ Open 2020 Nov 27;10(11):e041339. [doi: 10.1136/bmjopen-2020-041339] [Medline: 33247027]
8. Heron RL, Eisma MC. Barriers and facilitators of disclosing domestic violence to the healthcare service: a systematic review of qualitative research. Health Soc Care Community 2021 May;29(3):612-630. [doi: 10.1111/hsc.13282] [Medline: 33440034]
9. Dheensa S, Halliwell G, Daw J, Jones SK, Feder G. "From taboo to routine": a qualitative evaluation of a hospital-based advocacy intervention for domestic violence and abuse. BMC Health Serv Res 2020 Feb 21;20(1):129. [doi: 10.1186/s12913-020-4924-1] [Medline: 32085771]
10. Department of Justice. Domestic and Sexual Violence and Abuse. 2016. URL: https://www.justice-ni.gov.uk/articles/domestic-and-sexual-violence-and-abuse [accessed 2025-03-03]
11. McGregor K. Domestic violence and abuse in LGBTQ+ communities. In: Ali P, Rogers MM, editors. Gender-Based Violence: A Comprehensive Guide: Springer; 2023. [doi: 10.1007/978-3-031-05640-6_34]

XSL•FO
RenderX

12. Wellock S, Tarpey E. LGBTQ + Community perceptions of domestic abuse, responsivity and inclusion within the criminal justice system. Sexual and Gender Diversity in Social Services 2024 Jul 2;36(3):354-378. [doi: 10.1080/10538720.2023.2267480]
13. Hulley J, Bailey L, Kirkman G, et al. Intimate partner violence and barriers to help-seeking among Black, Asian, minority ethnic and immigrant women: a qualitative metasynthesis of global research. Trauma Violence Abuse 2023 Apr;24(2):1001-1015. [doi: 10.1177/15248380211050590] [Medline: 35107333]
14. NHS England. Strategic direction for sexual assault and abuse services: lifelong care for victims and survivors: 2018-2023. 2018 URL: https://www.england.nhs.uk/wp-content/uploads/2018/04/strategic-direction-sexual-assault-and-abuse-services.pdf [accessed 2025-03-03]
15. Liu Y, Dong H, Zhang L, El Saddik A. Technical evaluation of HoloLens for multimedia: a first look. IEEE MultiMedia 2018;25(4):8-18. [doi: 10.1109/MMUL.2018.2873473]
16. Hegarty K, McKibbin G, Hameed M, et al. Health practitioners' readiness to address domestic violence and abuse: a qualitative meta-synthesis. PLoS ONE 2020;15(6):e0234067. [doi: 10.1371/journal.pone.0234067] [Medline: 32544160]
17. Cleak H, Hunt G, Hardy F, Davies B, Bell J. Health staff responses to domestic and family violence: the case for training to build confidence and skills. Australian Social Work 2021 Jan 2;74(1):42-54. [doi: 10.1080/0312407X.2020.1808029]
18. Lungu AJ, Swinkels W, Claesen L, Tu P, Egger J, Chen X. A review on the applications of virtual reality, augmented reality and mixed reality in surgical simulation: an extension to different kinds of surgery. Expert Rev Med Devices 2021 Jan;18(1):47-62. [doi: 10.1080/17434440.2021.1860750] [Medline: 33283563]
19. Barteit S, Lanfermann L, Bärnighausen T, Neuhann F, Beiersmann C. Augmented, mixed, and virtual reality-based head-mounted devices for medical education: systematic review. JMIR Serious Games 2021 Jul 8;9(3):e29080. [doi: 10.2196/29080] [Medline: 34255668]
20. Banjar A, Xu X, Iqbal MZ, Campbell A. A systematic review of the experimental studies on the effectiveness of mixed reality in higher education between 2017 and 2021. Computers & Education: X Reality 2023 Dec;3:100034. [doi: 10.1016/j.cexr.2023.100034]
21. Kolecki R, Pręgowska A, Dąbrowa J, et al. Assessment of the utility of mixed reality in medical education. Translational Research in Anatomy 2022 Sep;28:100214. [doi: 10.1016/j.tria.2022.100214]
22. Chelsea and Westminster Hospital NHS Foundation Trust. The Queen Consort visits our domestic abuse service. 2022. URL: https://www.chelwest.nhs.uk/about-us/news/news-archive/2022/the-queen-consort-visits-our-domestic-abuse-service [accessed 2024-02-28]
23. Halliwell G, Dheensa S, Fenu E, et al. Cry for health: a quantitative evaluation of a hospital-based advocacy intervention for domestic violence and abuse. BMC Health Serv Res 2019 Oct 21;19(1):718. [doi: 10.1186/s12913-019-4621-0] [Medline: 31638998]
24. Park S, Bokijonov S, Choi Y. Review of Microsoft HoloLens applications over the past five years. Appl Sci (Basel) 2021;11(16):7259. [doi: 10.3390/app11167259]
25. Karunaratne D, Moore A, Dasigan K, Bartholomew B. Exploring the use of the hololens as an innovative means to facilitate early clinical exposure in medical school. : MedEd Publish; 2023. [doi: 10.21955/mep.1115444.1]
26. Likert R. A technique for the measurement of attitudes. Arch Psychol 1932;22(140):55.
27. Braun V, Clarke V. Thematic analysis: APA handbooks in psychology®. In: Handbook of Research Methods in Psychology Ed: American Psychological Association; 2012:57-71. [doi: 10.1037/13620-004]
28. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
29. Office for Health Improvement & Disparities. Vulnerabilities: applying All Our Health. 2022. URL: https://www.gov.uk/government/publications/vulnerabilities-applying-all-our-health/vulnerabilities-applying-all-our-health [accessed 2025-03-03]
30. SafeLives. National definition of IDVA work. 2014. URL: https://safelives.org.uk/sites/default/files/resources/National%20definition%20of%20IDVA%20work%20FINAL.pdf [accessed 2025-03-03]
31. e-lfh. E-learning for Domestic Violence and Abuse. 2020. URL: https://www.e-lfh.org.uk/programmes/domestic-violence-and-abuse-elearning-for-health-visitors-and-nurses [accessed 2025-03-03]
32. Microsoft. About HoloLens 2. 2023. URL: https://learn.microsoft.com/en-us/hololens/hololens2-hardware [accessed 2025-03-03]
33. Evans E. Diversity matters: intersectionality and women's representation in the USA and UK: Table 1. Parliam Aff 2016 Jul;69(3):569-585. [doi: 10.1093/pa/gsv063]
34. Xue J, Hu R, Zhang W, et al. Virtual reality or augmented reality as a tool for studying bystander behaviors in interpersonal violence: scoping review. J Med Internet Res 2021 Feb 15;23(2):e25322 [FREE Full text] [doi: 10.2196/25322] [Medline: 33587044]
35. SimX. Domestic Violence Virtual Reality Medical Simulation. 2024 Sep 6. URL: https://www.simxvr.com/marketplace/domestic-violence-2/ [accessed 2024-12-17]
36. Hobson L. Virtual reality used to help domestic abuse in north east courts. BBC News. 2024 Sep 16 URL: https://www.bbc.co.uk/news/articles/cp8n67wxgjro [accessed 2025-03-03]

XSL•FO
RenderX

37. Russell S. Sexual safety in healthcare – organisational charter.: NHS England; 2023 Sep 4 URL: https://www.england.nhs.uk/publication/sexual-safety-in-healthcare-organisational-charter [accessed 2025-03-03]

Abbreviations

DA: domestic abuse
HCP: health care professional
HP: HoloPatient
IDVA: Independent Domestic Violence Advisor
NHS: National Health Service
SV: sexual violence

Edited by A Mavragani; submitted 30.04.24; peer-reviewed by P Vemavarapu, V Baker; revised version received 26.01.25; accepted 28.01.25; published 19.03.25.

Please cite as:
Karunaratne D, Whittock J, Moore A, Dasigan K, Chevolleau J, Bartholomew B, Kelly N, Cohen CE
Exploring the Use of an Augmented Reality Device Learning Tool for Multidisciplinary Staff Training on Domestic Abuse and Sexual Violence: Postintervention Qualitative Evaluation
JMIR Form Res 2025;9:e60075
URL: https://formative.jmir.org/2025/1/e60075
doi:10.2196/60075

© Dilroshini Karunaratne, Jessica Whittock, Amber Moore, Krishna Dasigan, Jasmine Chevolleau, Brent Bartholomew, Nikki Kelly, Charlotte E Cohen. Originally published in JMIR Formative Research (https://formative.jmir.org), 19.3.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Optimizing Testimonials for Behavior Change in a Digital Intervention for Binge Eating: Human-Centered Design Study

Isabel R Rooper[1,2], BA; Adrian Ortega[1,3], PhD; Thomas A Massion[1,2], MA; Tanvi Lakhtakia[1,3], BA; Macarena Kruger[1,3], BS; Leah M Parsons[1,2], BS; Lindsay D Lipman[1,2], BA; Chidiebere Azubuike[1,2], MS; Emily Tack[1,3], BA; Katrina T Obleada[4,5], PhD; Andrea K Graham[1,2], PhD

[1]Center for Behavioral Intervention Technologies, Feinberg School of Medicine, Northwestern University, 750 N Lake Shore Dr, 10th Floor, Chicago, IL, United States

[2]Medical Social Sciences, Feinberg School of Medicine, Northwestern University, Chicago, IL, United States

[3]Preventive Medicine, Feinberg School of Medicine, Northwestern University, Chicago, IL, United States

[4]Department of Psychiatry and Behavioral Sciences, Feinberg School of Medicine, Northwestern University, Chicago, IL, United States

[5]Potocsnak Family Division of Adolescent and Young Adult Medicine, Ann & Robert H. Lurie Children's Hospital of Chicago, Lurie Children's Hospital, Chicago, IL, United States

Corresponding Author:
Andrea K Graham, PhD
Center for Behavioral Intervention Technologies, Feinberg School of Medicine, Northwestern University, 750 N Lake Shore Dr, 10th Floor, Chicago, IL, United States

Abstract

Background: Testimonials from credible sources are an evidence-based strategy for behavior change. Behavioral health interventions have used testimonials to promote health behaviors (eg, physical activity and healthy eating). Integrating testimonials into eating disorder (ED) interventions poses a nuanced challenge because ED testimonials can promote ED behaviors. Testimonials in ED interventions must therefore be designed carefully. Some optimal design elements of testimonials are known, but questions remain about testimonial speakers, messaging, and delivery, especially for ED interventions.

Objective: We sought to learn how to design and deliver testimonials focused on positive behavior change strategies within our multisession digital binge eating intervention.

Methods: We applied human-centered design methods to learn users' preferences for testimonial speakers, messaging, and delivery (modalities, over time, and as "nudges" for selecting positive behavior change strategies they could practice). We recruited target users of our multisession intervention to complete design sessions. Adults (N=22, 64% self-identified as female; 32% as non-Hispanic Black, 41% as non-Hispanic White, and 27% as Hispanic) with recurrent binge eating and obesity completed individual interviews. Data were analyzed using methods from thematic analysis.

Results: Most participants preferred designs with testimonials (vs without) for their motivation and validation of the intervention's efficacy. A few distrusted testimonials for appearing too "commercial" or personally irrelevant. For speakers, participants preferred sociodemographically tailored testimonials and were willing to report personal data in the intervention to facilitate tailoring. For messaging, some preferred testimonials with "how-to" advice, whereas others preferred "big picture" success stories. For delivery interface, participants were interested in text, video, and multimedia testimonials. For delivery over time, participants preferred testimonials from new speakers to promote engagement. When the intervention allowed users to choose between actions (eg, behavioral strategies), participants preferred testimonials to be available across all actions but said that selectively delivering a testimonial with one action could "nudge" them to select it.

Conclusions: Results indicated that intervention users were interested in testimonials. While participants preferred sociodemographically tailored testimonials, they said different characteristics mattered to them, indicating that interventions should assess users' most pertinent identities and tailor testimonials accordingly. Likewise, users' divided preferences for testimonial messaging (ie, "big picture" vs "how-to") suggest that optimal messaging may differ by user. To improve the credibility of testimonials, which some participants distrusted, interventions could invite current users to submit testimonials for future integration in the intervention. Aligned with nudge theory, our findings indicate testimonials could be used as "nudges" within interventions—a ripe area for further inquiry—though future work should test if delivering a testimonial only with the nudged choice improves its uptake. Further research is needed to validate these design ideas in practice, including evaluating their impact on behavior change toward improving ED behaviors.

(JMIR Form Res 2025;9:e59691) doi:10.2196/59691

XSL•FO
RenderX

KEYWORDS

health behavior; health narratives; binge eating; user engagement; personalization; behavior change; digital health; intervention; human-centered design; behavioral health; preferences

Introduction

Testimonials are an evidence-based behavior change strategy with beneficial impacts on health behaviors. Often characterized by the first-person voice of a typical person [1], testimonials can explain an individual's experience with a health condition or intervention and can be integrated as part of an intervention's didactic approach. Testimonial speakers are often credible *sources* (eg, someone who has used an intervention), which is useful because highlighting credible sources is an evidence-based behavior change technique [2]. A meta-analysis by Xu [3] demonstrated that health narratives like testimonials have comparable effectiveness to statistical evidence for supporting behavior change. Specifically, testimonials produce effects on behavioral intentions and are persuasive for health beliefs and attitudes [4]. There is support for using testimonials to encourage health behavior change in a range of health-promoting interventions, including for physical activity [5,6], healthy eating [7], cancer screening [8], smoking cessation [9], and diabetes self-management [10].

In the literature about eating disorders (EDs), the behavioral impacts of testimonials are more nuanced. A small systematic review on mental health recovery narratives identified that in the context of anorexia nervosa, exposure to ED recovery testimonials can increase ED behaviors in testimonial viewers, indicating that particular care should be taken to avoid potentially triggering content in testimonials [11]. Specific to binge eating behaviors, in a qualitative study that exposed young women to video blogs depicting binge eating, participants reported that exposure to testimonials produced both positive (eg, reductions in their own binge eating) and negative effects (eg, desire to replicate depicted binge eating behaviors) [12]. Indeed, personal stories from credible peer sources on so-called "pro-Ana" (anorexia nervosa) and "pro-Mia" (bulimia nervosa) web-based forums can undermine ED recovery and validate ED behaviors [13]. Yet the popularity of these sites among people with EDs [14]—as well as the prominence of online communities [15] and social media influencers [16,17] focused on ED recovery—indicates that people often seek out and are exposed to ED-related testimonials online, such that they may be interested in testimonials from peer sources within interventions as well. To deliver testimonials to users of ED interventions, it is therefore critical to avoid potential safety risks by carefully designing testimonials.

Extant literature provides some guidance for designing safe, effective, and engaging testimonials. De Graaf et al [4] conducted a meta-review of 153 experimental studies of health narratives and identified features of effective testimonials. For testimonial messaging, they found that depicting a positive health behavior in a testimonial (eg, physical activity) is more effective for changing behavioral intentions than depicting the consequences of an unhealthy behavior (eg, health risks associated with a sedentary lifestyle), though this finding was mixed in physical activity studies [18,19]. The review also found that first-person testimonials are most effective, but that users do not find testimonials from speakers with similar characteristics to them (eg, shared gender, race, and age) more persuasive than testimonials from dissimilar speakers. However, others have found that testimonials from speakers with similar characteristics to users are more persuasive for changing health beliefs [20], improving engagement with the health message [21], and increasing behavioral intentions [22]. Indeed, a more recent meta-analysis by Chen et al [23] identified that sharing characteristics with a testimonial speaker improves testimonial persuasiveness. However, to our knowledge, the design of testimonials for ED interventions has not been explored, such that much remains unknown about designing testimonials for populations with EDs.

The purpose of this paper is to inform the design of testimonials for a digital binge eating intervention. This study was conducted as part of our ongoing efforts applying human-centered design methods to create a 16-week digital health intervention (FoodSteps) that targets binge eating and weight-related behaviors. We used human-centered design methodologies because they centralize the perspectives of users [24] and are useful for learning how to carefully design and package psychosocial interventions and their constituent parts [25]. Indeed, including users in human-centered design processes can yield engaging digital tools with high clinical impact [26]. We thus conducted virtual "design sessions" with users to learn their preferences related to testimonials.

We examined several specific components. First, given the behavioral benefits of testimonials and the popularity of ED testimonials in nonintervention contexts (eg, online ED discussion forums) [14,15], we sought to learn if users were interested in viewing testimonials as part of a digital ED intervention. Second, we investigated users' preferences for testimonial messaging, speakers, and delivery. Within testimonial delivery, we were interested in delivery modalities, as well as learning how to deliver testimonials (1) over time and (2) as "nudges" toward optimal choices within an intervention. Although many digital health interventions that use testimonials are longitudinal (ie, delivered over weeks or months) [27,28], to our knowledge, how to deliver testimonials over time has been unexplored and represents an important design question for motivating ongoing behavior change in a longitudinal intervention. Relatedly, we wanted to learn users' perspectives on using testimonials to "nudge" a health-oriented behavior, since nudge theory in behavioral economics posits that attaching a "nudge" to an optimal choice is a useful method to influence users' behavior [29]. Testimonials have been used for nudging in nonhealth contexts [30,31] and could be used in health interventions to nudge users toward a particular choice. We therefore sought to learn whether intervention users are open to testimonials as "nudges" and whether they perceive testimonials delivered as nudges would influence their choices. Collectively, results of this human-centered design study can inform the design of testimonials within health interventions,

XSL•FO
RenderX

with particular relevance for interventions that target binge eating and weight-related behaviors.

Methods

We conducted a qualitative study of FoodSteps target users' preferences for testimonial design and delivery. In reporting our data, we adhered to the Standards for Reporting Qualitative Research (see supplementary files) [32]. During codebook formation, the research team consisted of 3 bachelor's level researchers, 1 master's level researcher, and 3 clinical psychologist researchers.

Participants and Procedures

From October to December 2022, we recruited potential participants via flyers in Chicago, Illinois, and online via Craigslist, social media, and ResearchMatch. Participants were recruited to participate in design sessions to help design a mobile app for managing binge eating and weight-related behaviors. Through tailored recruitment materials, methods, and locations, we sought to recruit a sociodemographically diverse sample based on gender, race, ethnicity, age, and socioeconomic status in an effort to increase generalizability and relevance.

Interested individuals completed an online screener, and potential participants were invited to complete a baseline assessment to confirm eligibility. Participants were eligible if they were non-pregnant, English-speaking adults (≥18 years old) with regular access to a smartphone. Eligible participants also had obesity (BMI ≥30 kg/m^2 based on self-reported height and weight) and recurrent binge eating (≥12 objectively large binge eating episodes in the past 3 months based on the diagnostic portion of the Eating Disorder Examination semi-structured interview) [33]. To be eligible, participants had to endorse struggling with losing weight and be willing to use an app to reduce their binge eating. We established this inclusion criteria to obtain a sample of participants that mirrors target users of the FoodSteps intervention. We invited a representative group of 25 participants to our design sessions. Of these, 22 participants completed the interviews; 3 were lost to follow-up.

Ethical Considerations

This study was approved by Northwestern University's Institutional Review Board (STU00216998). All participants gave verbal informed consent. Participants received US$25 financial compensation for completing study procedures.

Design Sessions

The design sessions aimed to learn how to design and deliver testimonials in a digital binge eating intervention. The principal investigator (AKG) and a research assistant (TAM) iteratively developed a 1-hour interview guide for the design sessions. Individual virtual interviews, held via Zoom (Zoom Communications, Inc.), were led by TAM and frequently co-led by AKG.

During the semistructured design sessions, the research team member leading the interview began by briefly explaining the FoodSteps intervention to the participant. They were then shown nonfunctional mock-ups of intervention features, including designs that featured a testimonial (see Figure 1). Participants were asked which versions of the testimonial designs they preferred and their preferences for testimonial speakers (eg, sociodemographic characteristics), messaging (eg, how-to advice), delivery interface (eg, video), and delivery over time and as a "nudge" to a particular action. Specifically, because FoodSteps involves users choosing behavioral strategies each week, participants were shown designs in which a testimonial is used to nudge users to select a particular strategy.

In Table 1, we present the interview guide questions on testimonials. Because the interviews were semi-structured, we followed-up on participants' unique responses with additional questions based on the feedback that participants shared.

As saturation emerged (ie, when enough participants provided the same feedback such that the research team felt consensus had been reached), the designs and interview guide were iteratively updated, consistent with the iterative human-centered design process [26]. For example, as saturation developed regarding the set of identity characteristics participants wanted to share with the testimonial speaker, we asked more about participants' preferences for the longitudinal delivery of testimonials. As the study progressed, we also iteratively updated the visualizations we showed during design sessions based on participants' feedback. As such, participants were shown slightly different designs, and interviews were focused on different questions depending on the stage of the study at which they completed their design session.

XSL•FO
RenderX

Figure 1. Example slide of testimonial design options shown to participants. Designs iterations include (1) no testimonial, (2) Shanice's "big picture" testimonial, and (3) Shanice's "how-to" testimonial. In the latter options, a testimonial is used to nudge a user toward a specific action (in this case, the behavioral strategy "improve self-image").

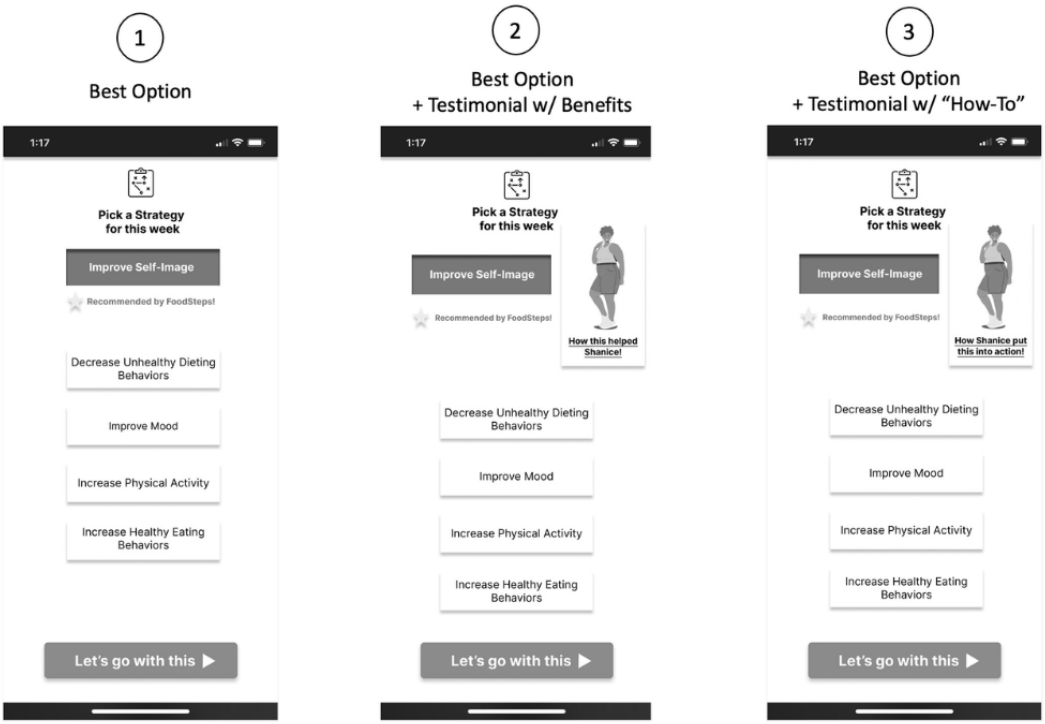

Table . Guiding questions from our interview guide to learn target users' preferences for testimonial design and delivery.

#1	What would you want to see in a testimonial?
#2	How would you want that testimonial presented to you (eg, video, article, pictures)?
#3	Is it important to you for the testimonials to be paired with your own identity? (eg, if you are a mother or father, seeing testimonials from other mothers or fathers)
#3a	If yes, ask: In order to make the testimonials related to you, would you be comfortable answering personal questions at the start to inform this personalization?
#4	Would you expect or want to see the same testimonial throughout the program, or different ones?
#5	Which of these designs do you prefer or like best, if any?
#5a	If there is one you prefer, why is it your favorite or what stands out to you about it?
#6	Which design(s) do you like least? What about them do you not like, find confusing, or feel is unhelpful?

Measures

We assessed participants' sociodemographic characteristics at screening (gender, race, ethnicity, age, and BMI) and through a questionnaire administered as part of the baseline eligibility assessment (household income and education). Objective binge eating was assessed at baseline by a trained assessor who administered the Eating Disorder Examination interview [33]; this assessor did not conduct the design sessions.

Analyses

We qualitatively analyzed interview text for themes using methods from thematic analysis [34,35]. Interviews were automatically transcribed by Zoom and edited for accuracy by a research assistant. A research team member (IRR) reviewed the interview data in full, then iteratively generated and applied an initial set of codes using Dedoose (SocioCultural Research Consultants), a qualitative data analysis software. After refining the codebook with another coder (AKG), IRR reviewed and recoded the data and began identifying themes. To strengthen the rigor of the coding process, the full study team reviewed the codebook and discussed the themes. IRR, AO, and AKG

XSL•FO
RenderX

refined and named the themes and met often to resolve coding ambiguities.

Results

Sample Characteristics

Of the 22 participants, 64% self-identified as female; 27% as Hispanic, 32% as non-Hispanic Black, and 41% as non-Hispanic White. On average, participants were 42 years (range 23-69 years; median 40 years; SD 14.10 years); had a BMI of 39.42 kg/m^2 (range 30.03-57.56 kg/m^2; median 37.48 kg/m^2; SD 7.15 kg/m^2); and experienced 49 binge episodes over the past 3 months (range 15-111; median 41; SD 30.59). Just over half (54%) had at least a bachelor's degree. Excluding 4 missing data points, participants' mean household income was US $79,616 (range US $10,000-$250,000; median US $53,000; SD US $64,991). See Table 2 for full sample characteristics.

Table . Sample characteristics of the target intervention users who participated in this study's design sessions.

ID	Gender	Race	Hispanic	Age (y)	Household income (USD)	Education	BMI (kg/m^2)	Number of binge eating episodes in the past 3 months
1	Female	Black	No	40	—	Master's degree	44.62	46
2	Female	White	No	36	$205,000	Master's degree	31.17	41
3	Female	White	No	43	$250,000	Master's degree	36.94	20
4	Female	White	No	56	$10,000	2-year college	38.16	36
5	Female	White	No	65	—	4-year college	30.99	15
6	Female	Black	No	29	$50,000	High school/GED	46.00	34
7	Female	Black	No	54	$56,000	Master's degree	47.42	111
8	Female	Black	No	28	—	Some college	57.56	22
9	Female	Black	No	40	$113,104	Professional degree	35.87	64
10	Male	White	No	69	$110,000	Professional degree	36.33	60
11	Female	White	Yes	28	—	4-year college	32.87	18
12	Female	More than one race	Yes	27	$30,000	Some college	35.66	111
13	Male	White	No	68	$60,000	Master's degree	41.55	84
14	Male	White	Yes	34	$65,000	Master's degree	34.72	48
15	Female	White	Yes	52	$99,000	Master's degree	40.85	108
16	Male	White	No	43	$150,000	4-year college	38.01	41
17	Male	White	No	48	$30,000	Some college	54.88	24
18	Female	Black	No	23	$50,000	Some college	38.84	67
19	Male	More than one race	Yes	30	$40,000	Some college	36.59	15
20	Male	Black	No	25	$50,000	Some college	36.18	32
21	Other	White	No	34	$25,000	Some college	42.06	32
22	Female	American Indian or Alaskan Native	Yes	45	$40,000	Some college	30.04	51

XSL•FO
RenderX

Preferences for Including Testimonials

Most users preferred designs featuring testimonials. They said testimonials personalized the intervention ("*that makes it a little more personalized for you*," P7) and validated its efficacy ("*look at all these people from various lots of life who have used this and are successful*," P2). Participants also found testimonials motivating ("*it will encourage someone*," P6) and engaging:

> *The testimonial person engages you. It gives you hope. It motivates you.* [P19]

Many said testimonials fostered connection ("*it's sort of a camaraderie in a way*," P13). They said binge eating is an illness for which social support is limited, so testimonials can fill this gap:

> *It's good to know that other people have the same issue.* [P11]
>
> *[It] feels more like you're doing it as a team effort, and that's […] important with such an isolating illness.* [P21]

While they generally preferred designs with testimonials, a subset distrusted testimonials for reasons of relevance, authenticity, and utility. A few recommended excluding testimonials altogether. Some participants found testimonials irrelevant because eating and weight are subjective topics:

> *The reason I don't like the testimonial option is because it's a very personal thing that you're doing.* [P3]
>
> *When it comes to my own weight loss journey, I wouldn't find how it helped somebody else [useful] […] that's just not something I would care to look into.* [P12]

Some found testimonials too "commercial" or distrusted the speaker:

> *I relate it mentally to too many, not scams, but like trying to sell a product [...] I might think it's just a paid actor. I don't trust it enough.* [P20]
>
> *It's like looking at an infomercial on TV […] Here's Suzy Cream Cheese, and she lost four hundred pounds in one week.* [P10]

A few participants also said that they would prefer advice come from the intervention, not a former user:

> *The 'how-to' is more important to me than the testimonial.* [P16]
>
> *Having some instruction as to how I can put it into action might be more beneficial for me personally than hearing how someone else put it into action.* [P9]

Participants also cited unique reasons for disliking testimonials. This participant said the speaker's success did not guarantee their own, so an unhelpful self-comparison could ensue:

> *If it doesn't help you and it did help someone else, then how are you going to internalize that? That's why I didn't like that one.* [P3]

Other critiques were of the illustration's larger body size, which raised concerns about the intervention's efficacy ("*if this helps Shanice, why does Shanice still look like this?*", P9) and prompted size comparisons ("*am I fatter than this person?*", P15). One participant worried the speaker's racial identity could alienate users ("*I would not use the image and the name, just because our country is so* [...] *ugly about race, and that could turn some people off*," P4).

Design Preferences

Participants described their preferences for testimonial speakers, interface, and messaging.

Speaker

Participants preferred speakers with shared identity characteristics such as age, race, gender, body size, lifestyle (ie, being busy), parenthood, and socioeconomic status.

> *I would immediately look at it and say I'm not a middle-aged African American woman.* [P10]
>
> *If it was someone smaller, I would be a bit more skeptical. But if it was someone that looks like me, I could say okay, she or he goes through the same struggle as I do.* [P18]
>
> *She's a busy lady like you are, and […] she still found time to squeeze it in.* [P1]
>
> *There's different challenges if you're a working mother compared to a stay-at-home mother, or someone who's not a mother.* [P2]
>
> *When Oprah lost her weight it was like, you have a personal chef, it's easy for you.* [P7]

Participants said sharing characteristics with speakers made those testimonials more comfortable ("*it's just something I feel comfortable with*," P22), relevant ("*I don't want to see how somebody was successful who I can't relate to*," P10), and inspiring ("*if it's someone I can relate to* […] *it makes it more real*," P7). They said these testimonials would make them more likely to listen:

> *If it shows some 6-foot tall, muscular dude […] I'm not going to listen to him.* [P8]

Yet, one user cautioned that sharing characteristics does not assure identification ("*just because this person is a female, and I'm a female, does not mean that I identify with this person*," P1).

Understanding which intersecting identities matter most to which users is important. We asked a subset of participants if they would be willing to report personal data in the intervention to facilitate identity-based matching and personalization; all of them said yes.

> *When you introduce that [the demographic questions] you should say: in order to really focus in on you, we're going to ask you these questions […] I want to know you're focusing in on me, because then it's worth my while to continue.* [P10]

These findings indicate that it will be important for interventions to deliver tailored testimonials to users, which can be facilitated by assessing their identity characteristics within the intervention.

XSL•FO
RenderX

Messaging

Generally, participants were shown 2 types of testimonials (see Figure 1). The first ("how this helped Shanice") conveyed the "big picture" impact of the behavioral strategy; the second ("how Shanice put this into action") conveyed "how-to" advice. Participants reported varying preferences between these options. Those who preferred "big picture" testimonials found the speaker's end results more motivating than the speaker's process ("*I'm more excited about the final results of losing weight, so this is more engaging*," P19). Others thought the big picture impact was obvious ("*it's not rocket science […] they know the benefits*," P8).

Some participants found "how-to" advice unhelpful ("*how one put it into action may not be how someone else puts it into action*," P9). Others favored "how-to" testimonials for their actionable insights ("*it's concrete. Easily implementable*," P14) that operationalize behavior change:

> *Sometimes when things are broken down […] it makes it a lot easier, especially if it's something that's new to you.* [P18]

These varying preferences indicate that both types of testimonial messaging ("big picture" and "how-to") are motivating to different users. To that end, this participant recommended assessing users' preferences within the intervention to discern the optimal messaging to motivate a user:

> *I could see there being an option in the setup for the app to choose […] Do you want to know testimonials? Do you want how-to? Which ones do you want?* [P17]

Understanding what testimonial messaging to deliver and to whom to best facilitate motivation is critical to optimize testimonials' behavioral impact.

Interface

Participants wanted testimonials with a simple interface and varying multimedia, such as videos, to encourage engagement and connection ("*it's more a personal connection when you have a video*," P2). A participant said they would prefer an image of the speaker, instead of an illustration ("*to the extent it could be a real person, that would be helpful*," P9). Others preferred testimonials delivered multimodally (eg, text and video) to give options for viewing the content:

> *Maybe having a video, and then the transcript below […] so I can kind of listen in my head where I may not have the audio, but I can still read and see them speaking.* [P2]
>
> *Every week, if there's something different; sometimes there might be audio of somebody talking about their experience, then there might be a write-up, and then there might be a video […] that might be motivating […] something that changes as weeks go by.* [P1]

One participant wanted multiple testimonials available so they could choose between speakers:

> *If you had a choice, like if you wanted to see Shanice, or Bob, or whomever. Like you could watch multiple testimonials from different people.* [P2]

Taken together, these preferences for multimodal content and choice between speakers can inform how intervention designers optimize testimonials' user interface.

Delivery Preferences

Testimonials could be a strategy to nudge users toward a particular action within the intervention when choice is involved (eg, in our intervention, users choose between behavioral strategies). We therefore asked participants about their preferences for seeing testimonials when there is an element of user choice. If multiple actions are available, this participant said a testimonial should only be delivered with the nudged action:

> *If I clicked on 'mood' this week, because 'mood' is not my recommended, there would be no testimonial […] I think I would be more motivated [to select the recommendation] if there was a testimonial.* [P2]

Other participants preferred to see testimonials across all available actions:

> *I thought it was going to show a visual next to [each behavioral strategy]. Like, on 'physical activity,' I thought it would have been somebody with weights, and on 'behaviors,' it would have somebody eating.* [P8)]
>
> *It would be great if even if it wasn't a recommendation, that they would have a testimonial there, too […] I still would like that option added to it.* [P7)]

Additionally, because most behavioral interventions are multi-session, we sought to understand users' preferences for the longitudinal delivery of testimonials. In our design sessions, we assessed this by asking users about their preferences after envisioning a week had passed in the program. Many participants wanted testimonial speakers to vary over time because new testimonials continue validating the intervention's efficacy:

> *It validates [it] if you can see real life people and every day it's a brand-new person.* [P19)]
>
> *I think different people each week. So that way you can see a variety of people, so you don't get into, like, oh my gosh, only one person's used this.* [P2)]

Some preferred testimonials from varied speakers each week to reduce potential boredom from hearing the same person's story repeatedly ("*especially if they're going through this week after week […] I really don't need to see that again*," P17). Delivering varied testimonials over time also increases the odds of identifying with the speaker:

> *If I click on it one day, and I see a Tom or something […] I don't really connect if it were Tom, but maybe next week I get somebody else.* [P1)]

However, one participant wanted to see the same speaker over time ("*I'd feel a lot more impressed if you showed me Bob all the time […] It's just like having a personal trainer*," P10). These findings indicate that when choice is involved, selectively delivering testimonials could nudge users toward the optimal action, though users preferred testimonials be available across

XSL•FO
RenderX

all actions. Understanding how to deliver testimonials over time will be crucial for intervention designers to facilitate motivation and engagement throughout a multi-session intervention.

Discussion

Principal Findings

We conducted design sessions to learn how to design and deliver testimonials within a multisession digital binge eating intervention. We found that participants generally preferred designs with testimonials because they promoted engagement, motivation, and connection. This finding indicates that target users of a digital intervention for binge eating and weight-related behaviors are interested in viewing testimonials as part of the intervention and aligns with other literature indicating interest in testimonials from peer sources among individuals with EDs [14,15]. We also described participants' preferences for testimonial design and delivery over time. From these results, we collated design recommendations, summarized in Table 3.

Table . Recommendations for optimizing the design of testimonials in digital health interventions.

General	Include testimonials from former users as part of digital interventions for eating and weight-related behaviors.
	Assess users' testimonial preferences in-app and incorporate choice around whether and how to receive testimonials (eg, speaker, messaging, or delivery).
Speaker	Design testimonials that match the user sociodemographically to improve relevance, understanding, and listening.
	Highlight the authenticity of the speaker to establish trust and avoid appearing too "commercial".
	Avoid showing users irrelevant testimonials that can disengage them.
	Facilitate sociodemographic matching of testimonials by leveraging users' willingness to input personal information in the app for matching purposes.
	Recognize that some users may prefer following the same person over time; as feasible, create and disseminate this content.
Messaging	Highlight actionable, nonprescriptive content that recognizes the personal nature of health journeys to transfer information without alienating users. eg, *"See how Shanice made this strategy work for her. Her unique approach might motivate you to find your own way to increase your daily activity."*
	Highlight the "big picture" impact of behavior changes to inspire users. eg, *"Here's how increasing her daily physical activity made Shanice feel stronger and helped her connect with her kids."*
	Include language to help users avoid fear of failure or unproductive self-comparison to the testimonial speaker. eg, *"The key is finding an approach that fit in my life. For me, finding what worked was a personal journey, and every small step has added up to big changes."*
Interface	Deliver testimonials through images, video, and audio to foster connection between the viewer and speaker, and capture motivating effects.
	Vary delivery method of testimonials to engage users.
Delivery	Offer testimonials from new speakers weekly (or on an intervention-appropriate cadence).
	When choice between actions is involved, consider leveraging testimonials to nudge users toward the optimal action.

Participants preferred testimonial speakers who shared their sociodemographic characteristics. This finding contrasts with de Graaf's [4] meta-analysis, but aligns with more recent research affirming the effectiveness of culturally tailored testimonials [36] and the persuasive influence of sharing race, age, and gender with a testimonial speaker [37]. When asked why tailored testimonials mattered to them, some participants said that nonrelatable testimonials (eg, from speakers with higher socioeconomic status or different racial identities) could alienate users, which underscores the importance of tailoring testimonials to users' personal characteristics. Notably, tailoring based on body size poses a nuanced challenge for weight-related interventions. While most participants preferred speakers with relatable body sizes, one said that this would make them think the intervention was ineffective. Indeed, because sharing certain characteristics was important to some participants, but not others, our results indicate that intervention designers should consider ways to assess users' preferences and characteristics

XSL•FO
RenderX

at the onset of the intervention to identify users' most pertinent identities (eg, is it more important for testimonials to match a user's gender or socioeconomic status?) and tailor testimonials accordingly. This can be facilitated by users' willingness to report personal information in the intervention, though continued design work is needed to understand user tolerance for testimonial personalization over time. Indeed, Robbins et al [38] observed that some Black participants were uncomfortable with a video testimonial intervention that featured overwhelmingly Black speakers. In multisession interventions, testimonial tailoring based on characteristics like race could feel increasingly targeted over time. It is thus necessary to learn users' preferences for tailored testimonials throughout an intervention during its actual delivery. Further research is also needed to empirically test if tailoring testimonials improves their effectiveness in the context of eating and weight-related interventions.

For testimonial messaging, we found users were relatively split between receiving "how-to" information versus "big picture" information, in that some preferred testimonials with actionable advice, whereas others preferred testimonials that highlighted the end result of the intervention. This divide also appears in the literature on testimonials for deterring maladaptive health behaviors. While Keer et al [39] found that testimonials describing the positive consequences of abstinence from binge drinking were most effective, Jawed and Hogan [40] found that highlighting actionable strategies for smoking cessation improved testimonial effectiveness and engagement. Our findings affirm that both forms of testimonial messaging may be effective for different subsets of users, depending on their preferred style of messaging. To this end, participants suggested incorporating choice around whether and how to receive testimonials, depending on the elements they find most motivating. User-driven personalization is highly feasible in digital interventions, given a sufficient bank of testimonial content, such that this is a promising area for further inquiry. Future work also would benefit from identifying factors that predict which design elements (eg, forms of messaging) are preferred by which subsets of users, toward effectively personalizing testimonials to users. Importantly, we did not explore users' preferences for testimonial messaging that included descriptions of binge eating or dangerous compensatory behaviors, given the behavioral consequences that can arise when viewers of ED testimonials are exposed to this kind of messaging [11,12]. Future work could explore opportunities to safely design testimonials that mention such behaviors, though our results indicate that potential users of our binge eating intervention were interested in and motivated by testimonials about positive health behaviors, such that testimonials describing ED behaviors could be avoided in interventions to reduce potential risks.

For delivery, we found that users were interested in written testimonials as well as video, audio, and multimedia testimonials. Because de Graaf [4] did not identify any modality as uniquely effective, and given participants' varied preferences, assessing users' preferred delivery modality at the onset of intervention is another promising opportunity for tailoring. For delivery over time, users preferred testimonials from varied speakers because they were more engaging and validated the intervention's efficacy. We did not learn how their messaging preferences may change over time, which is important because these preferences could meaningfully evolve (eg, a new user initially prefers "big picture" testimonials but later prefers "how tos"). Future work should thus examine methods to longitudinally assess users' testimonial preferences. When the intervention allowed users to choose between actions (in our case, behavioral strategies), users preferred testimonials to be delivered across all choices. While intervention designers could do this, our results indicate that selectively delivering a testimonial with the recommended action could encourage users to select it. This finding aligns with nudge theory in behavioral economics [29] and may be especially relevant for just-in-time adaptive interventions, which have deployed nudges as part of their behavior change strategies [41]. Our finding that users positively perceived testimonials indicates that testimonials are a promising design element to explore for deployment as nudges in digital health interventions, though future work should test if exclusively delivering a testimonial for the optimal action improves the action's uptake.

Despite the generally positive perceptions of testimonials, some users said testimonials would not motivate behavior change. This finding needs to be validated in practice through actual intervention delivery. Although much research supports the behavioral effects of testimonials [3], including for promoting abstinence from maladaptive health behaviors [9], Shen et al [42] found that cessation testimonials did not have significant effects, as opposed to prevention testimonials (eg, using condoms) and detection testimonials (eg, cancer screening), which did produce significant effects. This indicates that intervention designers—especially those who are designing interventions to dissuade maladaptive health behaviors—must continue identifying other elements to influence behavior and design to preempt testimonial pitfalls. Indeed, the subset of users who said testimonials would not influence their health behavior change universally critiqued the testimonials' trustworthiness. This objection can be understood through the lens of persuasive intent (ie, a message's implicit goal to persuade), which has been shown to undermine the credibility and persuasiveness of health communications [43,44]. Designing testimonials to highlight source credibility may help improve trust in the testimonial message [45]. Another approach to improve the credibility of testimonials could involve inviting users to submit their own testimonials for future integration into the intervention. This strategy could also help create a bank of testimonials from a diversity of users, which could then be deployed in the intervention.

Further research is needed to validate these findings with FoodSteps users in practice. While we believe these findings are relevant to interventions for other health topics, our study is limited by its focus on an intervention for binge eating and weight-related behaviors. Designers of other health interventions may wish to follow similar human-centered design methods to identify intervention-specific design preferences.

XSL•FO
RenderX

Conclusions

This study sought to learn how to optimize the design and delivery of testimonials in a multisession digital health intervention. Our findings confirm that target users of a digital intervention for binge eating and weight-related behaviors wanted testimonials delivered as part of the intervention. We generated novel insights into users' preferences for testimonial messaging, speakers, and delivery (through varied modalities, over time, and as a "nudge" toward a particular action). These results indicate opportunities for digital health intervention designers to tailor testimonials to users based on their preferences and identities. This study extends existing research by applying human-centered design methods to identify design preferences that can now be implemented and tested in practice. Our work offers a foundation for further inquiry into areas such as user preferences for sociodemographic tailoring over time and testimonials as "nudges" toward optimal actions in health interventions. Future research should continue applying human-centered design methods to address these and other questions, toward the goal of optimizing the behavioral impact of health testimonials.

Acknowledgments

This work was funded by the National Institute of Diabetes and Digestive and Kidney Diseases under grant number R01 DK133300 and the National Institute of Mental Health under grant number T32 MH115882.

Authors' Contributions

AKG and TAM designed the study. Data were analyzed by IRR, AO, and AKG. IRR, AO, and AKG drafted the manuscript. TL, TAM, MK, LMP, LDL, CA, ET, and KO critically reviewed and improved the manuscript with intellectual content. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Checklist 1
SRQR checklist.
[PDF File, 947 KB - formative_v9i1e59691_app1.pdf]

References

1. Bilandzic H, Busselle R. Narrative persuasion. In: The SAGE Handbook of Persuasion: Developments in Theory and Practice, 2nd edition: Sage Publications, Inc; 2013:200-219.
2. Michie S, Richardson M, Johnston M, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013 Aug;46(1):81-95. [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
3. Xu J. A meta-analysis comparing the effectiveness of narrative vs. statistical evidence: health vs. non-health contexts. Health Commun 2023 Dec;38(14):3113-3123. [doi: 10.1080/10410236.2022.2137750] [Medline: 36278821]
4. de Graaf A, Sanders J, Hoeken H. Characteristics of narrative interventions and health effects: a review of the content, form, and context of narratives in health-related narrative persuasion research. RCR 2016;4:88-131. [doi: 10.12840/issn.2255-4165.2016.04.01.011]
5. Perrier MJ, Martin Ginis KA. Changing health-promoting behaviours through narrative interventions: a systematic review. J Health Psychol 2018 Sep;23(11):1499-1517. [doi: 10.1177/1359105316656243] [Medline: 27387514]
6. Falzon C, Radel R, Cantor A, d'Arripe-Longueville F. Understanding narrative effects in physical activity promotion: the influence of breast cancer survivor testimony on exercise beliefs, self-efficacy, and intention in breast cancer patients. Support Care Cancer 2015 Mar;23(3):761-768. [doi: 10.1007/s00520-014-2422-x] [Medline: 25186211]
7. Oh HJ, Larose R. Tell me a story about healthy snacking and I will follow: comparing the effectiveness of self-generated versus message-aided implementation intentions on promoting healthy snacking habits among college students. Health Commun 2015;30(10):962-974. [doi: 10.1080/10410236.2014.910289] [Medline: 25256795]
8. Perrier MJ, Martin Ginis KA. Narrative interventions for health screening behaviours: a systematic review. J Health Psychol 2017 Mar;22(3):375-393. [doi: 10.1177/1359105315603463] [Medline: 26359288]
9. Igartua JJ, Rodríguez-Contreras L. Narrative voice matters! Improving smoking prevention with testimonial messages through identification and cognitive processes. Int J Environ Res Public Health 2020 Oct 5;17(19):7281. [doi: 10.3390/ijerph17197281] [Medline: 33028030]
10. Gardner L, Leshner G. The role of narrative and other-referencing in attenuating psychological reactance to diabetes self-care messages. Health Commun 2016;31(6):738-751. [doi: 10.1080/10410236.2014.993498] [Medline: 26528578]

XSL•FO
RenderX

11. Rennick-Egglestone S, Morgan K, Llewellyn-Beardsley J, et al. Mental health recovery narratives and their impact on recipients: systematic review and narrative synthesis. Can J Psychiatry 2019 Oct;64(10):669-679. [doi: 10.1177/0706743719846108] [Medline: 31046432]
12. Sung B, Lee JY. The influence of YouTube binge eating vlog viewing on perception and behavior among South Korean women, aged 20s to 30s, with binge eating tendencies: a qualitative analysis. Int J Ment Health 2024:1-27. [doi: 10.1080/00207411.2024.2377810]
13. Firkins A, Twist J, Solomons W, Keville S. Cutting ties with Pro-Ana: a narrative inquiry concerning the experiences of Pro-Ana disengagement from six former site users. Qual Health Res 2019 Aug;29(10):1461-1473. [doi: 10.1177/1049732319830425] [Medline: 30810096]
14. Sharpe H, Musiat P, Knapton O, Schmidt U. Pro‐eating disorder websites: facts, fictions and fixes. J Public Ment Health 2011 Mar 15;10(1):34-44. [doi: 10.1108/17465721111134538]
15. Nutley SK, Falise AM, Henderson R, Apostolou V, Mathews CA, Striley CW. Impact of the COVID-19 pandemic on disordered eating behavior: qualitative analysis of social media posts. JMIR Ment Health 2021 Jan 27;8(1):e26011. [doi: 10.2196/26011] [Medline: 33465035]
16. Nikolova I, LaMarre A. “If I Unfollow Them, It’s Not A Dig at Them”: A narrative analysis of Instagram use in eating disorder recovery. Psychol Women Q 2023 Sep;47(3):387-401. [doi: 10.1177/03616843231166378]
17. Herrick SSC, Hallward L, Duncan LR. “This is just how I cope”: an inductive thematic analysis of eating disorder recovery content created and shared on TikTok using #EDrecovery. Int J Eat Disord 2021 Apr;54(4):516-526. [doi: 10.1002/eat.23463] [Medline: 33382136]
18. Gray JB, Harrington NG. Narrative and framing: a test of an integrated message strategy in the exercise context. J Health Commun 2011 Mar;16(3):264-281. [doi: 10.1080/10810730.2010.529490] [Medline: 21058142]
19. Wirtz JG, Kulpavaropas S. The effects of narrative and message framing on engagement and eating intention among a sample of adult Hispanics. J Nutr Educ Behav 2014;46(5):396-400. [doi: 10.1016/j.jneb.2013.12.005] [Medline: 24517968]
20. de Graaf A. The effectiveness of adaptation of the protagonist in narrative impact: similarity influences health beliefs through self-referencing. Hum Commun Res 2014 Jan;40(1):73-90 [FREE Full text] [doi: 10.1111/hcre.12015]
21. Kim M, Shi R, Cappella JN. Effect of character-audience similarity on the perceived effectiveness of antismoking PSAs via engagement. Health Commun 2016 Oct;31(10):1193-1204. [doi: 10.1080/10410236.2015.1048421] [Medline: 26891148]
22. Lu AS. An experimental test of the persuasive effect of source similarity in narrative and nonnarrative health blogs. J Med Internet Res 2013 Jul 25;15(7):e142. [doi: 10.2196/jmir.2386] [Medline: 23887135]
23. Chen M, Dong Y, Wang J. A meta-analysis examining the role of character-recipient similarity in narrative persuasion. Communic Res 2024 Feb;51(1):56-82. [doi: 10.1177/00936502231204834]
24. Maguire M. Methods to support human-centred design. Int J Hum Comput Stud 2001 Oct;55(4):587-634. [doi: 10.1006/ijhc.2001.0503]
25. Lyon AR, Koerner K. User-centered design for psychosocial intervention development and implementation. Clin Psychol (New York) 2016 Jun;23(2):180-200. [doi: 10.1111/cpsp.12154] [Medline: 29456295]
26. Graham AK, Wildes JE, Reddy M, Munson SA, Barr Taylor C, Mohr DC. User-centered design for technology-enabled services for eating disorders. Int J Eat Disord 2019 Oct;52(10):1095-1107. [doi: 10.1002/eat.23130] [Medline: 31313370]
27. Brewer LC, Kumbamu A, Smith C, et al. A cardiovascular health and wellness mobile health intervention among Church-going African Americans: formative evaluation of the FAITH! app. JMIR Form Res 2020 Nov 17;4(11):e21450. [doi: 10.2196/21450] [Medline: 33200999]
28. Bruehlman-Senecal E, Hook CJ, Pfeifer JH, et al. Smartphone app to address loneliness among college students: pilot randomized controlled trial. JMIR Ment Health 2020 Oct 20;7(10):e21496. [doi: 10.2196/21496] [Medline: 33079071]
29. Patel MS, Volpp KG, Asch DA. Nudge units to improve the delivery of health care. N Engl J Med 2018 Jan 18;378(3):214-216. [doi: 10.1056/NEJMp1712984] [Medline: 29342387]
30. Gottschewski-Meyer PO, Lüddemann N, der LM, von VC. Exploring digital social norms nudges in E-grocery: typical consumer testimonials with a warm glow. 2023 Presented at: Internationale Tagung Wirtschaftsinformatik URL: https://aisel.aisnet.org/wi2023/59
31. Ouvrard B, Préget R, Reynaud A, Tuffery L. Nudging and subsidising farmers to foster smart water meter adoption. Eur Rev Agric Econ 2023 Jun 13;50(3):1178-1226. [doi: 10.1093/erae/jbad013]
32. O’Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med J Assoc Am Med Coll 2014;89(9):1245-1251. [doi: 10.1097/ACM.0000000000000388]
33. Fairburn CG, Cooper Z, O’Connor ME. The Eating Disorder Examination (EDE) 17.0D. 2014. URL: https://www.credo-oxford.com/7.2.html [accessed 2024-04-07]
34. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
35. Braun V, Clarke V. Can I use TA? Should I use TA? Should I not use TA? Comparing reflexive thematic analysis and other pattern‐based qualitative analytic approaches. Couns Psychother Res 2021 Mar;21(1):37-47. [doi: 10.1002/capr.12360]

XSL•FO
RenderX

36. Nugraha YR, Aditjondro E, Svenson D, et al. Harnessing the power of testimonial videos: a deep dive into the #1Minute2Quit campaign for spreading awareness on the health hazards of tobacco. JKMJC 2024;40(4):450-465 [FREE Full text] [doi: 10.17576/JKMJC-2024-4004-25]
37. Kim M. When similarity strikes back: conditional persuasive effects of character-audience similarity in anti-smoking campaign. Hum Commun Res 2019 Jan;45(1):52-77. [doi: 10.1093/hcr/hqy013] [Medline: 30631219]
38. Robbins R, Senathirajah Y, Williams NJ, et al. Developing a tailored website for promoting awareness about obstructive sleep apnea (OSA) among Blacks in community-based settings. Health Commun 2019 May;34(5):567-575. [doi: 10.1080/10410236.2018.1423865] [Medline: 29338353]
39. Keer M, van den Putte B, de Wit J, Neijens P. The effects of integrating instrumental and affective arguments in rhetorical and testimonial health messages. J Health Commun 2013;18(9):1148-1161. [doi: 10.1080/10810730.2013.768730] [Medline: 23659225]
40. Jawed A, Hogan A. Lessons learned from telltale testimonies: a descriptive study assessing coverage of the tips from former smokers campaign on YouTube. Am J Health Educ 2024 Mar 3;55(2):108-118. [doi: 10.1080/19325037.2023.2296950]
41. Hsu TCC, Whelan P, Gandrup J, Armitage CJ, Cordingley L, McBeth J. Personalized interventions for behaviour change: a scoping review of just-in-time adaptive interventions. Br J Health Psychol 2025 Feb;30(1):e12766. [doi: 10.1111/bjhp.12766] [Medline: 39542743]
42. Shen F, Sheer VC, Li R. Impact of narratives on persuasion in health communication: a meta-analysis. J Advert 2015 Apr 3;44(2):105-113. [doi: 10.1080/00913367.2015.1018467]
43. Lee EJ, Kim HS, Joo MH. Social media vs. mass media: mitigating the suspicion of ulterior motives in public health communication. Health Commun 2023 Oct;38(11):2450-2460. [doi: 10.1080/10410236.2022.2074781] [Medline: 35610554]
44. Wang W, Shen F. The effects of health narratives: examining the moderating role of persuasive intent. Health Mark Q 2019;36(2):120-135. [doi: 10.1080/07359683.2019.1575061] [Medline: 30907268]
45. Xu Y, Margolin D, Niederdeppe J. Testing strategies to increase source credibility through strategic message design in the context of vaccination and vaccine hesitancy. Health Commun 2021 Oct;36(11):1354-1367. [doi: 10.1080/10410236.2020.1751400] [Medline: 32308037]

Abbreviations

ED: eating disorder

Edited by A Mavragani; submitted 19.04.24; peer-reviewed by K Cooper; revised version received 15.01.25; accepted 10.02.25; published 21.03.25.

Please cite as:
Rooper IR, Ortega A, Massion TA, Lakhtakia T, Kruger M, Parsons LM, Lipman LD, Azubuike C, Tack E, Obleada KT, Graham AK Optimizing Testimonials for Behavior Change in a Digital Intervention for Binge Eating: Human-Centered Design Study JMIR Form Res 2025;9:e59691
URL: https://formative.jmir.org/2025/1/e59691
doi:10.2196/59691

© Isabel R Rooper, Adrian Ortega, Thomas A Massion, Tanvi Lakhtakia, Macarena Kruger, Leah M Parsons, Lindsay D Lipman, Chidiebere Azubuike, Emily Tack, Katrina T Obleada, Andrea K Graham. Originally published in JMIR Formative Research (https://formative.jmir.org), 21.3.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Interactive Care Coordination and Navigation mHealth Intervention for People Experiencing Homelessness: Cost Analysis, Exploratory Financial Cost-Benefit Analysis, and Budget Impact Analysis

Hannah P McCullough[1], PharmD, MPH; Leticia R Moczygemba[1], PharmD, PhD; Anton L V Avanceña[1,2], PhD; James O Baffoe[1], PharmD

[1]Health Outcomes Division, College of Pharmacy, The University of Texas at Austin, 2409 University Avenue, Austin, TX, United States

[2]Department of Internal Medicine, Dell Medical School, The University of Texas at Austin, Austin, TX, United States

Corresponding Author:
Leticia R Moczygemba, PharmD, PhD
Health Outcomes Division, College of Pharmacy, The University of Texas at Austin, 2409 University Avenue, Austin, TX, United States

Abstract

Background: The Interactive Care Coordination and Navigation (iCAN) mobile health intervention aims to improve care coordination and reduce hospital and emergency department visits among people experiencing homelessness.

Objective: This study aimed to conduct a three-part economic evaluation of iCAN, including a (1) cost analysis, (2) exploratory financial cost-benefit analysis, and (3) budget impact analysis (BIA).

Methods: We collected cost and expenditure data from a randomized controlled trial of iCAN to conduct a cost analysis and exploratory financial cost-benefit analysis. Costs were classified as startup and recurring costs for participants and the program. Startup costs included participant supplies for each participant and SMS implementation costs. Recurring costs included the cost of recurring services, SMS text messaging platform maintenance, health information access fees, and personnel salaries. Using the per participant per year (PPPY) costs of iCAN, the minimum savings reduction in the average health care costs among people experiencing homelessness that would lead to a benefit-cost ratio >1 for iCAN was calculated. This savings threshold was calculated by dividing the PPPY cost of iCAN by the average health care costs among people experiencing homelessness multiplied by 100%. The benefit-cost ratio of iCAN was calculated under different savings thresholds from 0% (no savings) to 50%. Costs were calculated PPPY under different scenarios, and the results were used as inputs in a BIA. A probabilistic sensitivity analysis was conducted to incorporate uncertainty around cost estimates. Costs are in 2022 US $.

Results: The total cost of iCAN was US $2865 PPPY, which was made up of US $265 in startup (9%) and US $2600 (91%) in recurring costs PPPY. The minimum savings threshold that would cause iCAN to have a positive return on investment is 7.8%. This means that if average health care costs (US $36,917) among people experiencing homelessness were reduced by more than 7.8% through iCAN, the financial benefits would outweigh the costs of the intervention. When health care costs are reduced by 25% ($9229/$36,917; equal to 56% [$9229/$16,609] of the average cost of an inpatient visit), the benefit-cost ratio is 3.22, which means that iCAN produces US $2.22 in health care savings per US $1 spent. The BIA estimated that implementing iCAN for 10,250 people experiencing homelessness over 5 years would have a financial cost of US $28.7 million, which could be reduced to US $2.2 million if at least 8% ($2880/$36,917) of average health care costs among people experiencing homelessness are reduced through the intervention.

Conclusions: If average costs of emergency department and hospital visits among people experiencing homelessness were reduced by more than 7.8% ($2880/$36,917) through iCAN, the financial benefits would outweigh the costs of the intervention. As the savings threshold increases, it results in a higher benefit-cost ratio.

Trial Registration: ClinicalTrials NCT05365867; https://clinicaltrials.gov/study/NCT05365867

(JMIR Form Res 2025;9:e64973) doi:10.2196/64973

KEYWORDS

people experiencing homelessness; budget impact analysis; financial cost-benefit analysis; mHealth; care coordination; care; mobile health; smartphones; homeless; hospitalization; cost analysis; health care cost; economic; emergency department; United States; cost-benefit; digital health

XSL•FO
RenderX

Introduction

People experiencing homelessness have higher rates of emergency department (ED) and hospital visits than housed groups. A systematic review reported that, on average, across 10 studies, the number of ED visits among people experiencing homelessness ranged from 0.7 to 5.8 visits annually. Furthermore, 5 of these studies compared ED use between people experiencing homelessness and housed groups and found that the rates were higher in people experiencing homelessness [1]. It has also been reported that people experiencing homelessness have 1.9 times the odds of hospital readmissions compared with housed groups [2]. Furthermore, a study of the hospital use patterns of people experiencing homelessness using electronic health data in an integrated health system in the United States found that people experiencing homelessness have higher rates of 30-day readmissions and longer lengths of stay in the hospital compared with those that are not homeless [3]. Some of the reasons for high ED and hospital utilization in people experiencing homelessness includes inadequate access to primary care and challenges taking medications due to medications being lost or stolen and competing priorities for basic needs such as shelter and food [4,5].

Mobile health (mHealth) is a type of digital technology that involves the use of mobile devices such as mobile phones to deliver health services [6]. The ubiquity and declining costs of mobile phones in the United States and other countries has generated interest in the use of mHealth to improve population and public health [7,8]. Studies have shown that mHealth can be effective in improving outcomes such as increasing patient adherence and healthy behaviors in chronic disease management [9], and reducing ED use [10-12]. According to 2 systematic reviews of cost effectiveness of digital interventions (eg, telephone, SMS text messaging, and mHealth apps), for those studies that included some type of economic evaluation, there is some evidence that these interventions may be cost-beneficial [9,13]. Even in the mHealth interventions where the outcomes were not statistically significant or cost-savings were not addressed, a reduction in preventable and nonemergent ED visits suggest likely significant cost-savings for health care payers in the community [13].

Given that mobile phones can reach traditionally hard to reach groups, including people experiencing homelessness [14], mHealth interventions should be considered when designing interventions to decrease ED and hospital visits in people experiencing homelessness. In a small pre-post study, people experiencing homelessness indicated that it was easy to use a mobile phone to participate in an mHealth intervention, and a qualitative study of people experiencing homelessness in the pre-post study found that mobile phones empowered people experiencing homelessness to self-manage their health and social needs [15,16]. A scoping review of 12 studies found that using technology on mobile phones such as SMS text messaging and apps could help improve access to mental health services for youth experiencing homelessness, with evidence of feasibility and acceptability by users [17]. The scalability of mHealth interventions for people experiencing homelessness will depend, in part, on whether its benefits outweigh the costs of the intervention. Thus, in this study, we will estimate the costs of an mHealth intervention, which will be described in the Methods section, that was implemented to decrease ED and hospital use among people experiencing homelessness and conduct an exploratory financial cost-benefit analysis and preliminary budget impact analysis to evaluate the potential cost-savings from this intervention for people experiencing homelessness.

Methods

Overview

Interactive Care Coordination and Navigation (iCAN) is a community-based mHealth intervention in the United States that was designed using findings from previous studies related to acceptability and benefits of mHealth interventions in people experiencing homelessness [15,16] and with feedback from homeless care providers and people with the lived experience of homelessness. The purpose of iCAN, which is implemented at community-based sites that provide homeless services, is to improve care coordination between health and social services for people experiencing homelessness with a goal of reducing hospitalizations and ED visits. It is a multicomponent intervention comprised of SMS text messages, preinstalled apps and resources, including a bus pass, and a case manager who conducted telephone encounters and used bidirectional SMS texting to communicate with participants. Participants received 3 - 5 messages daily regarding medication adherence and appointment reminders, general health messages, motivational messages such as encouragement to keep goals, and as needed messages for local information (eg, weather updates and resource fair information). Participants were also able to text the case manager or study staff. Within 72 hours of enrollment, the case manager conducted a telephone assessment to identify relevant health and social needs whereby they could connect participants to medical and social services in the community. Case managers received weekly reports twice about ED and hospital use from the local health information exchange (HIE) and contacted participants within 72 hours of an ED or hospital visit to help coordinate care needs for managing discharge instructions. People experiencing homelessness who were 18 years or older, owned a cell phone with service, had at least 2 chronic health conditions, had at least 2 hospitalizations or ED visits 6 months before enrollment, scored 4 or higher on the Rapid Estimate of Adult Literacy in Medicine-Short Form, and scored greater than 17 on the Mini-Mental State Exam were eligible to participate in iCAN. There was a total of 60 participants in the iCAN intervention. The mean age of participants in the iCAN intervention was 50.1 (SD 10.2) years. Most (42/60, 70%) were men; over half (34/60, 57%) reported their race as White and 32% (19/60) reported their race as Black. The randomized controlled trial (RCT) of iCAN completed recruitment in August 2023 and data analysis is ongoing.

Informed by previous studies of mHealth interventions [18-21], we conducted a three-part economic evaluation of iCAN, including a (1) cost analysis, (2) exploratory financial cost-benefit analysis, and (3) budget impact analysis. For the cost analysis, we used expenditure data from the iCAN RCT to determine the total cost and cost per participant per year (PPPY).

XSL•FO
RenderX

Using these costs, we conducted an exploratory cost-benefit analysis to understand the minimum health care savings that would cause iCAN to have a positive return on investment. Finally, we conducted a budget impact analysis to determine the financial requirements of scaling up the iCAN intervention in a metropolitan area. For the cost and cost-benefit analyses, we used scenario and probabilistic sensitivity analyses to determine the effect of different assumptions and uncertainty in existing data on our findings.

We followed the Consolidated Health Economic Evaluation Reporting Standards (CHEERS) [22], the International Training and Education Center for Health (I-TECH) Applied Economic Evaluation of Digital Health Interventions Report [23], the The International Society for Pharmacoeconomics and Outcomes Research (ISPOR) Budget Impact Analysis Report [24], and other guidance [25] in conducting this study. We used a public-payer perspective, assuming that the majority of people experiencing homelessness do not have private health insurance and cannot pay for out-of-pocket health care costs [26,27]. Thus, the costs of iCAN and health care services used by people experiencing homelessness are borne by the public sector through public health care payers (eg, Medicaid), health care systems, and local governments.

Data Collection

Intervention Costs

We collected expenditure records from the ongoing iCAN RCT to determine intervention costs. Intervention costs were separated into startup and recurring costs using a microcosting approach [28].

Startup or capital costs were separated into participant-related and program-related costs. Participant-related startup costs included a cell phone, phone case, armband, charging cable and block, rechargeable power bank, and drawstring backpack. Program-related startup costs included the initial payment for the SMS text messaging platform and printing costs for training materials.

Recurring costs were separated into participant-related, program-related, and personnel costs. Participant-related recurring costs included a monthly bus pass and a SIM card with a phone plan, which are given to each participant. Program-related recurring costs included monthly SMS text messaging platform maintenance fees and HIE data access fees. Access to HIE data are required to monitor ED and inpatient health care use of participants during the trial, however, this expense may be excluded during the scaling up of iCAN in the community. We explored this scenario in sensitivity analysis.

Personnel-related recurring costs included the salaries of a case manager and program coordinator. The case manager spent approximately 20 hours per week communicating with 24 enrolled participants; for 60 participants, this was equivalent to 1.25 case managers. The program coordinator, a role assumed by graduate research assistants during the iCAN trial, is responsible for several operational and administrative tasks, including setting up participants' cell phones (1 h per phone), training participants on the intervention (1 h per participant), and troubleshooting any technology-related issues (2 h per week). The case manager's annual salary was estimated based on the average hourly wage of social workers in Texas from the Bureau of Labor Statistics multiplied by the total work hours (2500) per year [29]. The salary of the program coordinator was based on the average hourly pay of graduate research assistants, which is available on the human resources website of The University of Texas at Austin [30]. The hours spent by the program coordinator on specific tasks were determined through interviews with trial staff and a review of staff records, and the total time spent was used to calculate the program coordinator's annual salary.

Health Care Costs

We used existing data to estimate the annual health care costs related to inpatient and ED care among people experiencing homelessness. We relied on previous studies on health care utilization among people experiencing homelessness in Texas to determine the average number and range of ED [31-34] and inpatient visits [31,35] (Multimedia Appendix 1). We took the weighted average and median of ED and inpatient visits across studies, which resulted in equivalent values and were comparable with the estimates from other jurisdictions and the United States [36-40]. We then multiplied these figures by the average cost of each service from the 2022 Medical Expenditure Panel Survey-Household Component (MEPS-HC) database [41]. With the average cost of inpatient (US $16,609, [SE $860]) and ED visits (US $1233 [SE $66]) from MEPS-HC, we used a total health care cost of US $36,916 PPPY. This average cost is comparable with what previous health care utilization studies among people experiencing homelessness who use health care services have reported in Texas and other locations [31-40].

Analysis

Cost Analysis

We calculated the total costs of iCAN by adding up all the startup and recurring intervention costs. We also calculated the PPPY costs by dividing the total costs by 60, which is the number of enrolled participants in the iCAN RCT. In scenario analyses, we varied assumptions around the cost inputs (Multimedia Appendix 2) and calculated total and PPPY costs under each scenario.

Cost-Benefit Analysis

Using the PPPY costs of iCAN, we calculated the minimum savings or reduction in the average health care costs among people experiencing homelessness that would lead to a benefit-cost ratio >1 for iCAN, where the financial benefits associated with the intervention outweigh its costs. We calculated this "savings threshold" by dividing the PPPY cost of iCAN by the average health care costs among people experiencing homelessness multiplied by 100%. For example, if the PPPY cost of iCAN is 10% of the average health care costs among people experiencing homelessness, a health care cost reduction of >10% (the savings threshold) would lead to a benefit-cost ratio >1 for iCAN. These savings can be accrued by reducing the number or severity of ED and inpatient visits. We then calculated the benefit-cost ratio of iCAN under different savings thresholds from 0% (no savings) to 50%.

XSL•FO
RenderX

To account for uncertainty in various cost inputs (ie, cost of the iCAN, number of ED, and inpatient visits), we conducted a probabilistic sensitivity analysis. A common approach in economic modeling, probabilistic sensitivity analysis involves generating probability distributions for each input using the best available parameter values and then drawing 10,000 random values of each input to produce a distribution of results. From these 10,000 estimates, we determined the mean, median, and 95% probable interval of the savings threshold for iCAN. We assigned gamma distributions for iCAN and health care costs and Poisson distributions for the annual number of inpatient and ED visits among people experiencing homelessness.

Budget Impact Analysis

We conducted a budget impact analysis to determine the financial requirements of implementing iCAN for people experiencing homelessness in a metropolitan area. We used a 5-year time horizon and selected Travis County, Texas, which includes the city of Austin where the iCAN trial was conducted, as the setting of the analysis. The number of unsheltered people experiencing homelessness in the first year was assumed to be 3990 based on estimates of the Homelessness Response System [42]. We assumed that the people experiencing homelessness population would grow annually by 3.4%, which is the estimated population growth rate for people experiencing homelessness by the US Department of Housing and Urban Development in 2023 [43]. We further assumed that only 60% of people experiencing homelessness would be eligible to participate in iCAN based on previous use of ED and inpatient services based on trial data and previously published studies. Finally, we assumed that only 80% of eligible people experiencing homelessness would participate in iCAN based on engagement data from the trial.

We included fixed and variable startup costs and variable recurring costs in the budget impact analysis. Fixed start-up costs, which were only paid in the first year of program implementation regardless of the number of participants, included the costs of the SMS text messaging platform and iCAN training materials. Variable startup costs included the 1-time cost of supplies (eg, cell phone) per participant. Recurring costs were the cost of services and salaries that varied based on the number of participants; these costs included SMS text messaging platform maintenance fees, HIE data access, personnel salaries, and annual bus passes and phone plans for each participant. We used a medical inflation rate of 2.3% in estimating future costs.

Ethical Considerations

The trial described in the study was approved by the University of Texas Institutional Board (STUDY00002666). The consent form indicated that research data would be kept for use in secondary analyses.

Results

Cost Analysis

The total cost of implementing iCAN to 60 participants for 1 year was US $170,921. The startup costs of iCAN (US $15,925) were estimated to be 9.3% ($15,925/$171,921) of the total cost (Table 1) and the remainder ($155,996/$171,921, 90.7%) were attributable to recurring costs (US $155,996), particularly personnel costs (US $84,908; Table 2). The startup and recurring costs PPPY were US $265 and US $2600, respectively, leading to a total PPPY cost of US $2865 PPPY (Table 3).

Table . Startup costs of Interactive Care Coordination and Navigation, a mobile health intervention for people experiencing homelessness in the United States, for a cost analysis. Total costs are calculated based on 60 participants. Costs are in 2022 US $.

Item		Unit cost per participant (US $)	Total cost (US $)
Participant-related			
	Cell phone	149.99	8999.40
	Phone case	12.99	779.40
	Arm band	8.99	539.40
	Charging block	3.50	210.00
	USB-C charging cable	3.50	210.00
	Power bank	14.99	899.40
	Drawstring backpack	1.34	80.40
	Subtotal	195.30	11,718.00
Program-related			
	Text messaging platform implementation	59.80	3588.00
	Printing costs for training binder supplies	10.31	618.59
	Subtotal	70.11	4206.59
Total		265.41	15,924.59

XSL•FO
RenderX

Table . Recurring costs of Interactive Care Coordination and Navigation, an mobile health intervention for people experiencing homelessness in the United States, for a cost analysis. Total costs were calculated for 60 participants. Costs are in 2022 US $.

Item		Unit cost (US $)	Units	Total annual cost (US $)
Participant-related				
	Bus pass	41.25 per month	12 months per participant	29,700
	Sim card or phone plan	210 per 6 months	12 months per participant	25,200
	Subtotal	—[a]	—	54,900
Program-related				
	SMS text messaging platform maintenance	99 per month	12 months for all participants	1188
	HIE[b] data access	200 per hour	750 hours for all participants	15,000
	Subtotal	—	—	16,188
Personnel-related				
	Program coordinator[c]	17 per hour	224 hours for all participants	3808
	Case worker	32.44 per hour	2500 hours for all participants	81,100
	Subtotal	—	—	84,908
Total		—	—	155,996

[a]Not applicable.

[b]HIE: health information exchange.

[c]Includes time spent on setting up (1 h/participant) and training (1 h/participant) participants and maintenance (2 h/wk).

Table . Per participant per year costs of Interactive Care Coordination and Navigation, a mobile health intervention for people experiencing homelessness in the United States, for a cost analysis. Costs are in 2022 US $.

Category		Cost PPPY[a] (US $)
Startup cost		
	Participant-related	195.30
	Program-related	70.11
	Subtotal	265.41
Recurring cost		
	Participant-related	915.00
	Program-related	269.80
	Personnel-related	1415.13
	Subtotal	2599.93
Total cost		2865.34

[a]PPPY: per participant per year.

The total PPPY cost of iCAN intervention varied slightly under different scenarios (Table 4). For example, when startup costs were excluded (Scenario A), recurring costs PPPY was US $2600. Excluding the cost of HIE data access (Scenario C) led to a similar PPPY cost of US $2615. The removal of bus passes and HIE data access (Scenario D) produced the lowest cost of US $2120 PPPY.

XSL•FO
RenderX

Table . Results of scenario analyses for a cost analysis of Interactive Care Coordination and Navigation, an mobile health intervention for people experiencing homelessness in the United States. All costs are in PPPY[a] 2022 US $.

Type of cost	Base case	Scenario A	Scenario B	Scenario C	Scenario D
Startup costs (US $)	265	0	265	265	265
Recurring costs (US $)	2600	2600	2105	2350	1855
Total iCAN[b] costs (US $)	2865	2600	2370	2615	2120
Savings threshold	7.8	7.0	6.4	7.1	5.7

[a]PPPY: per participant per year.

[b]iCAN: Interactive Care Coordination and Navigation.

In Table 4, the savings threshold refers to the percent (%) reduction in average health care costs for ED and inpatient care among people experiencing homelessness. The base case includes both startup and recurring costs. Scenario A includes recurring costs only. Scenario B includes startup costs and recurring costs without the bus pass. Scenario C include startup costs and recurring costs without HIE data access. Scenario D includes startup costs and recurring costs without the bus pass and HIE data access.

Cost-Benefit Analysis

When the cost of iCAN PPPY equals the base-case estimate (US $2865), the savings threshold is 7.8% (Table 2). This means that if average health care costs (US $36,917) among people experiencing homelessness were reduced by more than 7.8% through iCAN, the financial benefits would outweigh the costs of the intervention. A higher savings threshold is associated with a higher benefit-cost ratio for iCAN (Figure 1). When 10% ($3692/$N36,917) of average health care costs among people experiencing homelessness are averted (equal to the average cost of 3 ED visits), iCAN has a benefit-cost ratio of 1.29. When health care costs are reduced by 25% ($9229/N$36,917; equal to 56% [$9229/N$16,609] of the average cost of an inpatient visit), the benefit-cost ratio is 3.22, which means that iCAN produces US $2.22 in health care savings per US $1 spent.

Figure 1. Benefit-cost ratio of Interactive Care Coordination and Navigation, an mobile health intervention for people experiencing homelessness in the United States, under varying savings thresholds.

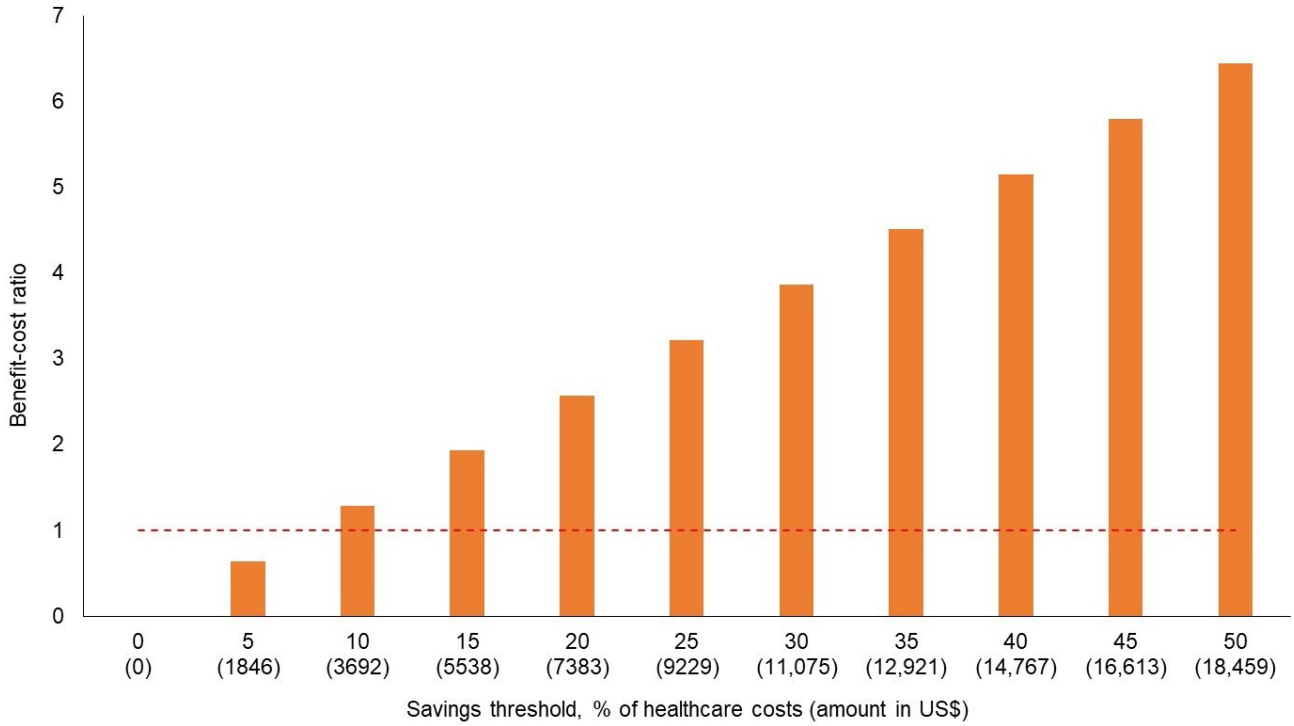

Figure 1 shows the benefit-cost ratio of iCAN under different savings thresholds, which is the percent (%) reduction in average health care costs for ED and inpatient care among people experiencing homelessness. Average health care costs were calculated using the weighted average of ED and inpatient visits among people experiencing homelessness from previous studies, whose value was equivalent to the median. The dotted red line marks a benefit-cost ratio of 1, which occurs when the benefits or financial savings of iCAN outweigh its costs. The probabilistic sensitivity analyses, which account for uncertainty in the costs of iCAN and the average health care costs of people experiencing homelessness, found a median savings threshold of 8.2% (95% probable interval 3.2 - 94.0). The average savings threshold was higher at 21%.

XSL•FO
RenderX

Budget Impact Analysis

The 5-year financial cost of implementing the iCAN intervention in Travis County for 10,250 people experiencing homelessness was around US $28.7 million (Table 5). If at least 5% of average health care costs among people experiencing homelessness are averted, the net budget impact is US $9.7 million. If average health care savings increased to 7%, the net budget impact of iCAN is further reduced to US $2.2 million.

Table . A 5-year budget impact analysis of Interactive Care Coordination and Navigation, an mobile health intervention for people experiencing homelessness in the United States (the 5-year total costs were calculated for the total number of people experiencing homelessness participating in Interactive Care Coordination and Navigation and health care savings were calculated based on the average health care costs for emergency department and inpatient care among people experiencing homelessness per year [US $36,917] among those who use these services).

	Year 1	Year 2	Year 3	Year 4	Year 5	5-year total
Total number of participants, n	1915	1980	2048	2117	2189	10,250
Intervention costs (US $)						
Fixed costs	4207	0	0	0	0	4207
Startup costs, PPPY[a]	195	195	195	195	195	2,001,764
Recurring costs, PPPY	2600	2600	2600	2600	2600	26,648,502
Total financial cost (US $)	5,357,637	5,535,448	5,723,653	5,918,257	6,119,478	28,654,472
Net budget impact (US $)						
5% health care savings	1,822,466	1,880,080	1,944,002	2,010,099	2,078,442	9,735,088
7% health care savings	408,397	417,933	432,142	446,835	462,028	2,167,335

[a]PPPY: per participant per year.

Discussion

Principal Results

In this study, we found, using a base-case estimate of US $2865, that a reduction in ED and hospital costs (estimated to be US $36,916 PPPY) of about 8%, which equals the cost of 2 ED visits PPPY, through iCAN would outweigh the costs of the intervention. Higher reductions in ED and hospital costs would result in a higher benefit-cost ratio for iCAN. While the economic value of mHealth interventions have been previously explored, none have focused on interventions for people experiencing homelessness. In total, 29 out of 39 studies evaluated in a systematic review conducted by Iribarren et al [44] in 2017 concluded that mHealth interventions show cost-effectiveness, economic benefit, or cost savings. Just over half of the mHealth interventions included SMS text messaging as the mHealth-related function in the intervention, with 77% (17/22) of these interventions reporting positive cost outcomes [44]. However, many limitations were identified in this systematic review. Of the 39 studies included in the review, only 12 were considered high-quality economic evaluations, which was defined as meeting at least 90% (21/24) of the items in the CHEERS checklist. In 2022, Gentili et al [13] conducted a systematic review of digital health interventions, which includes mHealth and other technologies such as telehealth, evaluating studies published after the time period evaluated by Iribarren et al [44]. Furthermore, 35 studies were included, and the findings were relatively consistent with that of Iribarren et al [44]. Only 12 of the 35 studies met at least 90% (21/24) of the items on the CHEERS checklist [13]. While the quality of economic evaluation was the focus of these systematic reviews, there is little evaluation of the budgetary impact and scalability of the included studies. Budget impact analyses were not conducted for these studies, and any evaluations on the economic scalability and sustainability of the interventions were preliminary, suggesting the need for formal budget impact analyses on mHealth interventions.

Another mHealth intervention used SMS to increase participation in their own care and decrease the number of no-show appointments in homeless veterans at a metropolitan Veterans Health Administration hospital. There was a significant decrease in the number of appointments cancelled by the patient and no-show appointments, and there was a significant reduction in ED visits. This study also showed that there were significant cost-savings when comparing the costs of implementing and maintaining SMS appointment reminders with the cost of unutilized services due to cancelled and no-show appointments [45]. A systematic review conducted in 2014 found that studies that evaluated cost-effectiveness showed cost savings associated with SMS services compared with standard of care while maintaining equal efficacy of interventions [46]. Even in the mHealth interventions where the outcomes were not statistically significant or cost-savings were not addressed, a reduction in preventable and nonemergent ED visits suggest likely significant cost-savings for health care payers in the community [13,44,47].

Our study adds to this literature by estimating the minimum health care savings needed for an mHealth intervention like iCAN to have positive net financial benefits. However, our analysis excludes other benefits that may be associated with iCAN, such as improvements in health outcomes and health-related quality of life. These health-related benefits

XSL•FO
RenderX

should be quantified, valued, and added to the cost-benefit analysis or cost-effectiveness analysis to gain a more comprehensive picture of iCAN's overall value.

Research has shown that people experiencing homelessness use more health care services than individuals with stable housing. In our analysis, we relied on previously published estimates of ED and inpatient health care use among people experiencing homelessness in Texas, which all found that use of these services is common (>1) among people experiencing homelessness who have a history of using these services. For example, single-center retrospective studies in Harris County, Texas, and Fort Worth, Texas found that people experiencing homelessness who used the ED had an average of 3.82 and 3.29 ED visits, respectively [33,34]. These numbers align with recent estimates from the National Hospital Ambulatory Medical Care Survey that found an ED visit rate of 310 per 100 persons per year in 2020 - 2021[37]. Estimates of inpatient visits among people experiencing homelessness in Texas are more limited and appear higher than numbers from other jurisdictions. A study of veterans in Pittsburgh, Pennsylvania, estimated the average number of inpatient visits at 0.7 (SD 1.6) among those who were hospitalized in 2012 - 2013, while single-center studies in Boston, Massachusetts, and Seattle, Washington State, found slightly higher numbers (>2) [40,48], which align with research showing a high risk of readmission among people experiencing homelessness [2,3]. Using these health care utilization numbers, we calculated an average health care cost comparable with previously published estimates of the cost of care among people experiencing homelessness [35,49,50]. Additional research is needed to understand the health care use and costs among people experiencing homelessness.

This study not only provides insight to the cost-benefit of the iCAN mHealth intervention, but also highlights the scalability and sustainability of an mHealth intervention like iCAN in a metropolitan area. Including a budget impact analysis of iCAN demonstrates that expanding the services provided by iCAN over a 5-year period is both possible and cost-saving. While existing literature on the economic implications of mHealth interventions exists, the majority of these studies did not conduct formal budget impact analyses to determine the sustainability of implementing an intervention on a larger scale [13,44]. Our economic evaluation and budget impact analysis balances both real-world data with necessary assumptions on a realistic scale, and many of the assumptions used in this study can be replaced with more specific real-world data after the iCAN RCT has concluded.

This economic evaluation and budget impact analysis provides insightful information in an emerging field where concrete data are lacking in several aspects. In published studies conducting budget impact analyses, the interventions being studied are typically medications or treatment protocols, not multicomponent behavioral health interventions [21,51-54]. While there are dozens of economic evaluations of mHealth interventions, there are very few formal budget impact analyses associated with these economic evaluations [13,44]. Our economic evaluation and budget impact analysis aims to fill this gap about mHealth interventions and the economic scalability and sustainability associated with expanding the services provided.

Limitations

The primary limitations of this study involve the assumptions made in the economic evaluation. The costs associated with the iCAN intervention were estimated from expenditure data from the RCT, which likely overestimate the actual costs of the intervention if implemented in real-world setting due to costs required to ensure accurate trial design and adherence to protocol [20]. We addressed this issue by adjusting the fixed costs included in our budget impact analysis, however, other economies of scale are potentially still excluded from our analysis. Health care costs related to ED and inpatient visits were estimated using national health care expenditure data (MEPS-HC) because such data were not available through the trial or for Travis County specifically. With possible variations in health care cost and utilization among people experiencing homelessness, our probabilistic sensitivity analyses explored the impact of uncertainty on our findings and indeed found a wide range of savings thresholds for iCAN. This suggests iCAN's economic value will be highly influenced by the change in health care costs and utilization of people experiencing homelessness who receive the intervention. Finally, we also had to make simplifying assumptions in the budget impact analysis, including the current size and growth of the people experiencing homelessness population in Travis County over the next 5 years. Our budget impact analysis can be easily updated with more precise inputs to understand the financial aspects of expanded access to iCAN.

Conclusions

This exploratory economic evaluation gives estimates of the cost of the iCAN mHealth intervention while exploring the cost-savings and budget impact of implementing the iCAN mHealth intervention on a larger scale. By using data from the iCAN RCT and other available resources, this economic evaluation determined that implementing the iCAN mHealth intervention for people experiencing homelessness in a metropolitan area provides financial cost-benefit if 1 hospitalization or 2 ED visits can be avoided. Future studies should explore the feasibility of implementing the iCAN mHealth intervention using data from the completed iCAN RCT to limit uncertainty in the cost of scaling up the intervention. Health care payers should examine the benefits of implementing mHealth interventions like this one to prevent unnecessary health care utilization and the associated costs among people experiencing homelessness.

Acknowledgments

Portions of the content (ie, the randomized controlled trial described in the methods) of this manuscript are funded under grant 1R18 HS027750-01 from the Agency for Healthcare Research and Quality (AHRQ), US Department of Health and Human

XSL•FO
RenderX

Services (HHS). The authors are solely responsible for the contents, findings, and conclusions, which do not necessarily represent the views of AHRQ. Readers should not interpret any statement in this report as an official position of AHRQ or of HHS. None of the authors have any affiliation or financial involvement that conflict with the material presented in this report. Portions of this manuscript were previously presented as abstracts and posters at the St. David's Center for Health Promotion and Disease Prevention Research in Underserved Populations Conference at the University of Texas School of Nursing in March 2023 and the AcademyHealth Annual Research Meeting in Seattle, WA, in June 2023.

Data Availability

The datasets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Annual emergency department and inpatient health care utilization among people experiencing homelessness.
[DOCX File, 16 KB - formative_v9i1e64973_app1.docx]

Multimedia Appendix 2
Scenarios explored in the cost analysis of Interactive Care Coordination and Navigation, an mHealthmobile health intervention for people experiencing homelessness in the United States.
[DOCX File, 15 KB - formative_v9i1e64973_app2.docx]

References

1. Vohra N, Paudyal V, Price MJ. Homelessness and the use of emergency department as a source of healthcare: a systematic review. Int J Emerg Med 2022 Jul 28;15(1):32. [doi: 10.1186/s12245-022-00435-3] [Medline: 35902803]
2. Miyawaki A, Hasegawa K, Figueroa JF, Tsugawa Y. Hospital readmission and emergency department revisits of homeless patients treated at homeless-serving hospitals in the USA: observational study. J Gen Intern Med 2020 Sep;35(9):2560-2568. [doi: 10.1007/s11606-020-06029-0] [Medline: 32666492]
3. Banerjee S, Yassin M, Dyer WT, Thomas TW, Rodriguez LA, Schmittdiel J. Hospital readmissions among patients experiencing homelessness: an electronic health record data study. Perm J 2024 Mar 15;28(1):55-61. [doi: 10.7812/TPP/23.075] [Medline: 38108331]
4. Liu M, Hwang SW. Health care for homeless people. Nat Rev Dis Primers 2021;7(1):5. [doi: 10.1038/s41572-020-00241-2] [Medline: 33446661]
5. Thurman W, Moczygemba LR, Baffoe JO. "Without my medication, I'm a wreck": photo-elicitation to explore medication use among people experiencing homelessness. Res Social Adm Pharm 2022 Jul;18(7):3149-3157. [doi: 10.1016/j.sapharm.2021.08.015] [Medline: 34479800]
6. What is digital health? US Food and Drug Administration. URL: https://www.fda.gov/medical-devices/digital-health-center-excellence/what-digital-health [accessed 2025-01-13]
7. Velmovitsky PE, Kirolos M, Alencar P, Leatherdale S, Cowan D, Morita PP. Leveraging mHealth technologies for public health. JMIR Public Health Surveill 2024 Sep 12;10:e49719. [doi: 10.2196/49719] [Medline: 39265164]
8. mHealth: use of appropriate digital technologies for public health: report by the director-general. : World Health Organization; 2017 URL: https://iris.who.int/bitstream/handle/10665/274134/B142_20-en.pdf?sequence=1&isAllowed=y [accessed 2025-03-06]
9. Kyaw TL, Ng N, Theocharaki M, Wennberg P, Sahlen KG. Cost-effectiveness of digital tools for behavior change interventions among people with chronic diseases: systematic review. Interact J Med Res 2023 Feb 16;12:e42396. [doi: 10.2196/42396] [Medline: 36795470]
10. Arora S, Peters AL, Burner E, Lam CN, Menchine M. Trial to examine text message-based mHealth in emergency department patients with diabetes (TExT-MED): a randomized controlled trial. Ann Emerg Med 2014;63(6):745-754. [doi: 10.1016/j.annemergmed.2013.10.012] [Medline: 24225332]
11. Eustache JH, Latimer EA, Liberman AS, et al. A mobile phone app improves patient-physician communication and reduces emergency department visits after colorectal surgery. Dis Colon Rectum 2023 Jan 1;66(1):130-137. [doi: 10.1097/DCR.0000000000002187] [Medline: 34933314]
12. Ladley A, Hieger AW, Arthur J, Broom M. Educational text messages decreased emergency department utilization among infant caregivers: a randomized trial. Acad Pediatr 2018 Aug;18(6):636-641. [doi: 10.1016/j.acap.2018.02.003] [Medline: 29432907]
13. Gentili A, Failla G, Melnyk A, et al. The cost-effectiveness of digital health interventions: a systematic review of the literature. Front Public Health 2022;10:787135. [doi: 10.3389/fpubh.2022.787135] [Medline: 36033812]

XSL•FO
RenderX

14. Moczygemba LR, Cox LS, Marks SA, Robinson MA, Goode JVR, Jafari N. Homeless patients' perceptions about using cell phones to manage medications and attend appointments. Int J Pharm Pract 2017 Jun;25(3):220-230. [doi: 10.1111/ijpp.12321] [Medline: 27896909]
15. Moczygemba LR, Thurman W, Tormey K, Hudzik A, Welton-Arndt L, Kim E. GPS mobile health intervention among people experiencing homelessness: pre-post study. JMIR Mhealth Uhealth 2021 Nov 3;9(11):e25553. [doi: 10.2196/25553] [Medline: 34730550]
16. Thurman W, Semwal M, Moczygemba LR, Hilbelink M. Smartphone technology to empower people experiencing homelessness: secondary analysis. J Med Internet Res 2021 Sep 29;23(9):e27787. [doi: 10.2196/27787] [Medline: 34586073]
17. Lal S, Elias S, Sieu V, Peredo R. The use of technology to provide mental health services to youth experiencing homelessness: scoping review. J Med Internet Res 2023 Jan 16;25:e41939. [doi: 10.2196/41939] [Medline: 36645703]
18. Prinja S, Gupta A, Bahuguna P, Nimesh R. Cost analysis of implementing mHealth intervention for maternal, newborn & child health care through community health workers: assessment of ReMIND program in Uttar Pradesh, India. BMC Pregnancy Childbirth 2018;18(1):390. [doi: 10.1186/s12884-018-2019-3] [Medline: 30285669]
19. Cheng Q, Dang T, Nguyen TA, et al. mHealth application for improving treatment outcomes for patients with multidrug-resistant tuberculosis in Vietnam: an economic evaluation protocol for the V-SMART trial. BMJ Open 2023;13(12):e076778. [doi: 10.1136/bmjopen-2023-076778] [Medline: 38081668]
20. Heapy AA, Driscoll MA, Buta E, et al. Co-operative pain education and self-management (copes) expanding treatment for real-world access (extra): pragmatic trial protocol. Pain Med 2020;21(Suppl 2):S21-S28. [doi: 10.1093/pm/pnaa365] [Medline: 33313733]
21. Avanceña ALV, Brody C, Chhoun P, Tuot S, Yi S. Connecting female entertainment workers in Cambodia to health care services using mHealth: economic evaluation of mobile link. JMIR Form Res 2024 Jul 25;8:e52734. [doi: 10.2196/52734] [Medline: 39052328]
22. Husereau D, Drummond M, Augustovski F, et al. Consolidated Health Economic Evaluation Reporting Standards 2022 (CHEERS 2022) statement: updated reporting guidance for health economic evaluations. BMJ 2022;376:e067975. [doi: 10.1136/bmj-2021-067975] [Medline: 35017145]
23. Babigumira JB, Dolan S, Shade S, et al. Applied economic evaluation of digital health interventions. I-TECH. URL: https://www.go2itech.org/wp-content/uploads/2021/02/I-TECH_HIS_Economic_Evaluation.pdf [accessed 2024-07-30]
24. Sullivan SD, Mauskopf JA, Augustovski F, et al. Budget impact analysis-principles of good practice: report of the ISPOR 2012 Budget Impact Analysis Good Practice II Task Force. Value Health 2014;17(1):5-14. [doi: 10.1016/j.jval.2013.08.2291] [Medline: 24438712]
25. Rascati K. Essentials of Pharmacoeconomics: Lippincott Williams & Wilkins; 2020.
26. Baggett TP, O'Connell JJ, Singer DE, Rigotti NA. The unmet health care needs of homeless adults: a national study. Am J Public Health 2010;100(7):1326-1333. [doi: 10.2105/AJPH.2009.180109] [Medline: 20466953]
27. Fryling LR, Mazanec P, Rodriguez RM. Barriers to homeless persons acquiring health insurance through the Affordable Care Act. J Emerg Med 2015 Nov;49(5):755-762. [doi: 10.1016/j.jemermed.2015.06.005] [Medline: 26281811]
28. Špacírová Z, Epstein D, García-Mochón L, Rovira J, Olry de Labry Lima A, Espín J. A general framework for classifying costing methods for economic evaluation of health care. Eur J Health Econ 2020;21(4):529-542. [doi: 10.1007/s10198-019-01157-9] [Medline: 31960181]
29. US Bureau of Labor Statistics. URL: https://data.bls.gov/timeseries/CUUR0000SAM?output_view=pct_12mths [accessed 2022-12-14]
30. Student employee compensation. The University of Texas at Austin Human Resources. URL: https://hr.utexas.edu/student/student-employee-compensation [accessed 2024-07-27]
31. Petrovich JC, Hunt JJ, North CS, Pollio DE, Roark Murphy E. Comparing unsheltered and sheltered homeless: demographics, health services use and predictors of health services use. Community Ment Health J 2020 Feb;56(2):271-279. [doi: 10.1007/s10597-019-00470-0] [Medline: 31552539]
32. Schiefelbein EL, Olson JA, Moxham JD. Patterns of health care utilization among vulnerable populations in Central Texas using data from a regional health information exchange. J Health Care Poor Underserved 2014;25(1):37-51. [doi: 10.1353/hpu.2014.0020] [Medline: 24509011]
33. Amato S, Nobay F, Amato DP, Abar B, Adler D. Sick and unsheltered: homelessness as a major risk factor for emergency care utilization. Am J Emerg Med 2019 Mar;37(3):415-420. [doi: 10.1016/j.ajem.2018.06.001] [Medline: 29891125]
34. Holmes CT, Holmes KA, MacDonald A, et al. Dedicated homeless clinics reduce inappropriate emergency department utilization. J Am Coll Emerg Physicians Open 2020;1(5):829-836. [doi: 10.1002/emp2.12054] [Medline: 33145527]
35. Buck DS, Brown CA, Mortensen K, Riggs JW, Franzini L. Comparing homeless and domiciled patients' utilization of the Harris County, Texas public hospital system. J Health Care Poor Underserved 2012;23(4):1660-1670. [doi: 10.1353/hpu.2012.0171] [Medline: 23698680]
36. Coe AB, Moczygemba LR, Harpe SE, Gatewood SBS. Homeless patients' use of urban emergency departments in the United States. J Ambul Care Manage 2015;38(1):48-58. [doi: 10.1097/JAC.0000000000000034] [Medline: 25469578]
37. Schappert SM, Santo L. Emergency department visits by homeless status and sex: United States, 2016-2021. Natl Health Stat Report 2024 Jun(204):1-21. [Medline: 38912857]

XSL•FO
RenderX

38. Ball MAZ, Sack DE, Druffner SA, et al. Characteristics and health care utilization of patients with housing insecurity in the ED. JAMA Netw Open 2024 Apr 1;7(4):e248565. [doi: 10.1001/jamanetworkopen.2024.8565] [Medline: 38669017]
39. Gabrielian S, Yuan AH, Andersen RM, Rubenstein LV, Gelberg L. VA health service utilization for homeless and low-income Veterans: a spotlight on the VA Supportive Housing (VASH) program in greater Los Angeles. Med Care 2014;52(5):454-461. [doi: 10.1097/MLR.0000000000000112] [Medline: 24714583]
40. Lin WC, Bharel M, Zhang J, O'Connell E, Clark RE. Frequent emergency department visits and hospitalizations among homeless people with Medicaid: implications for Medicaid expansion. Am J Public Health 2015 Nov;105 Suppl 5(Suppl 5):S716-S722. [doi: 10.2105/AJPH.2015.302693] [Medline: 26447915]
41. MEPS-HC data tools. AHRQ. URL: https://datatools.ahrq.gov/meps-hc [accessed 2024-11-08]
42. Austin count of homeless people returns this weekend. Here's what we know. Austin American-Statesman. URL: https://www.statesman.com/story/news/2023/01/27/austin-texas-annual-count-homeless-people-ending-community-homlessness-coalition/69839931007/ [accessed 2024-03-07]
43. Department of Housing and Urban Development (HUD). URL: https://www.hud.gov/ [accessed 2024-03-07]
44. Iribarren SJ, Cato K, Falzon L, Stone PW. What is the economic evidence for mHealth? A systematic review of economic evaluations of mHealth solutions. Plos one 2017;12(2):e0170581. [doi: 10.1371/journal.pone.0170581] [Medline: 28152012]
45. McInnes DK, Petrakis BA, Gifford AL, et al. Retaining homeless veterans in outpatient care: a pilot study of mobile phone text message appointment reminders. Am J Public Health 2014;104(Suppl 4):S588-S594. [doi: 10.2105/AJPH.2014.302061] [Medline: 25100425]
46. Beratarrechea A, Lee AG, Willner JM, Jahangir E, Ciapponi A, Rubinstein A. The impact of mobile health interventions on chronic disease outcomes in developing countries: a systematic review. Telemed J E Health 2014;20(1):75-82. [doi: 10.1089/tmj.2012.0328] [Medline: 24205809]
47. Katz R, Mesfin T, Barr K. Lessons from a community-based mHealth diabetes self-management program: it's not just about the cell phone. J Health Commun 2012;17(Suppl 1):67-72. [doi: 10.1080/10810730.2012.650613] [Medline: 22548601]
48. Srebnik D, Connor T, Sylla L. A pilot study of the impact of housing first-supported housing for intensive users of medical hospitalization and sobering services. Am J Public Health 2013;103(2):316-321. [doi: 10.2105/AJPH.2012.300867] [Medline: 23237150]
49. O'Toole TP, Johnson EE, Borgia M, et al. Population-tailored care for homeless veterans and acute care use, cost, and satisfaction: a prospective quasi-experimental trial. Prev Chronic Dis 2018 Feb 15;15:E23. [doi: 10.5888/pcd15.170311] [Medline: 29451116]
50. Mitchell MS, León CLK, Byrne TH, Lin WC, Bharel M. Cost of health care utilization among homeless frequent emergency department users. Psychol Serv 2017;14(2):193-202. [doi: 10.1037/ser0000113] [Medline: 28481604]
51. Faleiros DR, Álvares J, Almeida AM, et al. Budget impact analysis of medicines: updated systematic review and implications. Expert Rev Pharmacoecon Outcomes Res 2016;16(2):257-266. [doi: 10.1586/14737167.2016.1159958] [Medline: 26923561]
52. Jahn B, Todorovic J, Bundo M, et al. Budget impact analysis of cancer screening: a methodological review. Appl Health Econ Health Policy 2019 Aug;17(4):493-511. [doi: 10.1007/s40258-019-00475-6] [Medline: 31016686]
53. Mohammadnezhad G, Sattarpour M, Moradi N. Budget impact analysis of breast cancer medications: a systematic review. J Pharm Policy Pract 2022;15(1):105. [doi: 10.1186/s40545-022-00493-1] [Medline: 36581921]
54. Luo Z, Ruan Z, Yao D, Ung COL, Lai Y, Hu H. Budget impact analysis of diabetes drugs: a systematic literature review. Front Public Health 2021;9:765999. [doi: 10.3389/fpubh.2021.765999] [Medline: 34869180]

Abbreviations

CHEERS: Consolidated Health Economic Evaluation Reporting Standards
ED: emergency department
HIE: health information exchange
I-TECH: International Training and Education Center for Health
iCAN: Interactive Care Coordination and Navigation
ISPOR: The International Society for Pharmacoeconomics and Outcomes Research
MEPS-HC: Medical Expenditure Panel Survey-Household Component
mHealth: mobile health
PPPY: per participant per year
RCT: randomized controlled trial

XSL•FO
RenderX

Edited by A Mavragani; submitted 01.08.24; peer-reviewed by L Pirhonen, MP Montgomery; revised version received 13.01.25; accepted 16.01.25; published 18.03.25.

Please cite as:
McCullough HP, Moczygemba LR, Avanceña ALV, Baffoe JO
The Interactive Care Coordination and Navigation mHealth Intervention for People Experiencing Homelessness: Cost Analysis, Exploratory Financial Cost-Benefit Analysis, and Budget Impact Analysis
JMIR Form Res 2025;9:e64973
URL: https://formative.jmir.org/2025/1/e64973
doi:10.2196/64973

© Hannah P McCullough, Leticia R Moczygemba, Anton L V Avanceña, James O Baffoe. Originally published in JMIR Formative Research (https://formative.jmir.org), 18.3.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Digital Health Literacy and Attitudes Toward eHealth Technologies Among Patients With Cardiovascular Disease and Their Implications for Secondary Prevention: Survey Study

Greta Ullrich[1], Dr Med; Alexander Bäuerle[2,3], PhD; Hannah Vogt[1]; Amir Abbas Mahabadi[1], Prof Dr Med; Katrin Paldán[4], Dr Phil; Daniel Messiha[1], Dr Med; Lisa Maria Jahre[2,3], MSc; Christos Rammos[1], Prof Dr Med; Tienush Rassaf[1], Prof Dr Med; Julia Lortz[1], Prof Dr Med

[1]Department of Cardiology and Vascular Medicine, West-German Heart and Vascular Center Essen, University of Duisburg-Essen, Essen, Germany
[2]Clinic for Psychosomatic Medicine and Psychotherapy, LVR-University Hospital Essen, Essen, Germany
[3]Center for Translational Neuro- and Behavioral Sciences (C-TNBS), University of Duisburg-Essen, Essen, Germany
[4]HCT Research, Vorarlberg University of Applied Science, Dornbirn, Austria

Corresponding Author:
Julia Lortz, Prof Dr Med
Department of Cardiology and Vascular Medicine, West-German Heart and Vascular Center Essen, University of Duisburg-Essen, Essen, Germany

Abstract

Background: Cardiovascular disease is the major cause of death worldwide, leading to a significant socioeconomic burden. Although secondary prevention is a cornerstone in chronic disease management, adherence to guideline recommendations in this regard often falters, leading to suboptimal outcomes. While eHealth technologies are promising for improving treatment adherence, they also represent a new approach to secondary prevention. However, a common critique is that extensive digitalization may not adequately address the needs of older adults with chronic medical conditions.

Objective: This study aims to analyze eHealth literacy, digital use patterns, and general attitudes toward digital technologies in a collective of patients with cardiovascular disease to identify potential obstacles in implementing mobile health technologies in secondary preventive therapy.

Methods: This survey-based study was a part of the baseline examination of the PreventiPlaque trial. It involved 240 participants with known coronary artery disease. The assessment evaluated their current understanding of the general use of digital devices. The questionnaire covered aspects such as the duration of daily use, personal attitudes, and the perceived burden associated with digital media. eHealth literacy was assessed within the target population and general demographic data were gathered, focusing on cardiovascular comorbidities and risk factors.

Results: The analysis revealed an average age of 61.9 (SD 8.9) years, with 59.9% (n=144) of the participants being male. Overall, 37.3% (n=90) of the participants had previous knowledge of digital health interventions, while only 17.8% (n=41) had used them. Despite the generally low practical application within this study population, there was a high level of confidence in handling digital devices, with 61.9% (n=149) expressing themselves as either rather confident or very confident. Regarding the levels of eHealth literacy among the participants, 71.2% (n=170) claimed to be familiar with locating health information on the internet, and 64% (n=153) of participants felt capable of critically evaluating its quality. These levels of digital confidence were consistent across all age groups. Moreover, internet use rates remained high even among the older participants, with 80% (n=192) of those participants older than 75 years using the internet for 1-3 hours a day.

Conclusions: The study unveiled a notable confidence level among participants regarding the use of digital devices, coupled with a favorable attitude toward digital media evident across all age brackets. Remarkably, internet use rates remained high, even among older participants. The actual utilization of digital health interventions was relatively low, potentially stemming from challenges in locating reliable sources. These findings emphasize the prospect of future eHealth interventions customized to the distinct needs and preferences of patients in cardiovascular disease management. Recognizing the incongruity between confidence in device use and the restricted adoption of digital health tools can guide the development of focused interventions to narrow this divide.

Trial Registration: ClinicalTrials.gov NCT05096637; https://clinicaltrials.gov/study/NCT05096637

(JMIR Form Res 2025;9:e63057) doi:10.2196/63057

XSL•FO
RenderX

KEYWORDS

cardiovascular diseases; telemedicine; eHealth; patient-centered approach; digital health literacy; digital health; cardiovascular disease; mortality; artery disease; ischemic; heart disease; diabetes mellitus; obesity; patient education; eHealth literacy; mobile phone

Introduction

Cardiovascular disease remains the leading cause of global mortality, with coronary artery disease and ischemic heart disease alone contributing to 16% of total annual deaths worldwide [1]. Over the past 2 decades, ischemic heart disease has demonstrated the most rapid growth in overall death rates globally, resulting in nearly 9 million annual deaths [1]. Cardiovascular disease is intricately linked to various comorbidities and a severe mental health burden [2]. While advancements in the understanding and treatment of cardiovascular disease persist, there is an increasing emphasis on secondary prevention. Following guideline recommendations, secondary prevention aims to address modifiable risk factors such as hyperlipidemia, nicotine consumption, arterial hypertension, diabetes mellitus, obesity, chronic stress, and lack of physical activity [3]. Despite the preventive potential of lifestyle modifications supported by guidelines, their implementation remains inadequate. Primary obstacles include the challenge of ensuring adherence to long-term behavioral changes and the scarcity of medical resources and time needed for comprehensive patient education in chronic disease management [4,5]. Digital health interventions emerge as a promising avenue to educate and empower patients, encouraging an active role in disease management while optimizing the control and monitoring of modifiable risk factors [6]. They offer a potential solution to the current deficiencies in medical infrastructure, addressing the escalating demand associated with the increasing prevalence of cardiovascular disease [7]. However, a frequently cited challenge in the expanding digitization of health care is the consideration of special needs among older patients concerning the development and structure of digital health interventions [8]. Given that cardiovascular disease predominantly affects older individuals, it is crucial to not disadvantage them to fully use the modern treatment options.

The objective of this study was to delve into the digital use patterns of patients with cardiovascular disease expressing an interest in digital health interventions. The aim was to assess and identify potential obstacles and challenges in the implementation of mobile health (mHealth) or eHealth technologies for secondary prevention of cardiovascular disease. The study sought to evaluate patients' existing knowledge and opinions regarding smartphones, internet, and digital health interventions while focusing on uncovering possible age-dependent differences. The overarching goal was to gain insights that would inform the design of digital health interventions tailored to meet the specific needs and requirements of individuals managing cardiovascular conditions.

Methods

Study Design and Participants

We conducted a survey-based assessment at the University Hospital Essen, West German Heart and Vascular Center, Department of Cardiology and Vascular Medicine. The questionnaires used were part of the baseline examination of the PreventiPlaque trial. PreventiPlaque is a registered randomized clinical trial (NCT05096637) testing the effects of a smartphone app that included atherosclerotic plaque visualization on adherence to secondary preventive therapy [9]. The recruitment of participants for the PreventiPlaque trial occurred in 2022. Patients with atherosclerotic cardiovascular disease were eligible for participation in this trial. This included patients with documented ischemic heart disease, acute coronary syndrome, and patients with proven peripheral artery disease [10]. Another requirement was adult age (18 years and older). Participants had to own a smartphone that was suitable for potential app use. Finally, patients had to give written informed consent to comply with the study protocol and had to be willing to participate in the study. Patients with insufficient knowledge of the German language, or unwillingness to use a smartphone app, were excluded from the study.

Ethical Considerations

The conduct of the trial was approved by the ethics committee of the Medical Faculty of the University of Duisburg-Essen (20 - 9157-BO). Written informed consent was provided by each participant before inclusion in the study. The study was conducted in accordance to the Declaration of Helsinki. Data were collected and analyzed using a pseudonymous form. The participants did not receive any form of financial compensation for participating in this trial.

Measurements

Sociodemographic and Medical Data

Basic sociodemographic data and medical data were collected, using a questionnaire with 15 items, including the patient's marital status, level of education, and current profession. Moreover, patients were asked to self-assess their level of physical activity, as well as the quality of their diet. We extracted medical information from the hospital's electronic medical records regarding the presence of major cardiovascular risk factors such as diabetes mellitus, arterial hypertension, and hypercholesterolemia, as well as the prevalence of cardiovascular comorbidities including coronary artery disease, peripheral artery disease, stroke, aortic syndrome, or chronic heart failure.

eHealth Literacy

The revised German eHealth Literacy Scale (GR-eHEALS) was used to assess the participants' skills to find and critically evaluate health information on the internet [11]. It is a validated

XSL•FO
RenderX

tool to measure eHealth literacy in patients with cardiovascular disease and the modified version contains 8 items and 2 subscales [12]. Responses could be given on a 5-point Likert scale (eg, “I know how to use the internet to find answers to my health-related questions,” 1=does not apply to me, 5=does apply to me).

eHealth-Related Data

The self-generated eHealth data questionnaire started with 3 items examining the participant’s general confidence in handling digital media, digital devices, and web-based platforms on the internet. They could respond on a 5-point Likert scale (1=not confident at all, 5=very confident). Patients were asked to self-assess their daily internet utilization period for private or work-related purposes. Moreover, internet anxiety was measured by 6 items with possible responses on a 5-point Likert scale (eg “I have concerns about using the internet,” 1=does not apply to me, 5=does apply to me). Finally, it was determined whether the participants were already experienced in using digital health interventions, had heard about them, or knew where to find them. The self-generated eHealth items had already been used in previous studies [13-16] and have proven good reliability.

Statistical Analysis

We performed descriptive statistical analysis using SPSS (version 23; IBM Corp). Variables were presented as frequencies and percentages, as well as means and SDs. The participant’s BMI was calculated using the patient’s weight and height according to our hospital’s electronic medical records. Since it is often brought forward that older patients may not be a suitable target group for digital health interventions, we analyzed internet use frequency and digital confidence in relation to the participant’s age. To serve this purpose, we divided the participants into 4 age groups. To reach comparable sample sizes, we divided the age groups into <55, 55 - 64, 65 - 75, and >75-year-olds. Using these age groups, we performed an age-adjusted analysis of the variables “Internet-use,” as well as “confidence in handling internet platforms/digital media/mobile technologies.” This included an age-adjusted means comparison, using the Kruskal-Wallis test. We also conducted bivariate correlations between the participant’s age and the 4 items “Internet-use” and “confidence in handling Internet platforms/digital media/mobile technologies,” using Spearman’s Rho correlation for ordinally scaled variables. Finally, we analyzed these items mentioned above, describing the participants’ digital confidence adjusted to gender, and comparing the 2 formed subgroups “male” and “female.”

Results

Socioeconomic Characteristics

In total, 240 patients completed the assessment. With a mean age of 61.9 (SD 8.9) years and 17.4% (n=42) of participants being older than 70 years, the study showed an older people collective with a total of 59.9% (n=144) of all participants being male. With 60% (n=144) in total, most participants were married, 51.5% (n=124) of participants were retired, and 22.5% (n=54) of participants were working full-time.

Comorbidities and Cardiovascular Risk Factors

An analysis of the prevalence of major cardiovascular risk factors showed that 79.6% (n=191) and 80.4% (n=193) of the study participants had hypercholesterinemia and arterial hypertension, respectively. One-fourth of the participants were active smokers (n=60, 25%) and 17% (n=41) had type-2 diabetes while 36.5% (n=88) of participants were overweight with a BMI of 25 - 30 kg/m² and another 34.9% (n=84) of participants were obese with a BMI of >30 kg/m² (Table S1 in Multimedia Appendix 1). While all patients were diagnosed with coronary artery disease, a total of 36.7% (n=88) of patients had also been diagnosed with peripheral artery disease. Moreover, 26.3% (n=63) of participants had known congestive heart failure and 5.8% (n=14) of participants had once experienced a stroke (Table S1 in Multimedia Appendix 1).

eHealth Data

Overview

Analyzing the frequency and duration of internet use per day, only 5.9% (n=14) of the study population did not use the internet daily and 53.4% (n=128) of participants stated an internet use of 1 - 3 hours per day on average (Table 1). In general, levels of confidence when handling digital media, digital devices, and internet platforms were high in this study population (Table 2). Regarding internet-based programs to promote health or provide health information, only 34.4% (n=83) of participants knew how these programs worked and 47.5% (n=114) of participants knew where to find them (Table 3). When asked about having doubts about using the internet, a total of 75.7% (n=181) of the study population stated that this did not apply at all or rather did not apply. Furthermore, almost 80% (n=192) of the study population did not feel negatively affected by carrying a mobile phone with them (Table 4).

XSL•FO
RenderX

Table . Cardiovascular risk factors and comorbidities (N=240).

Cardiovascular risk factors	Prevalence, n (%)
BMI>30	84 (34.9)
BMI>25	88 (36.5)
Type 2 diabetes	41 (17)
Nicotine dependency	60 (25)
Hypercholesterolemia	191 (79.6)
Arterial hypertension	193 (80.4)
Congestive heart failure	63 (26.3)
Peripheral artery disease	88 (36.7)
Stroke	14 (5.8)

Table . Daily internet use (n=236).

	How long are you using the internet for private purposes per day?, n (%)
Not at all	14 (5.9)
Less than 1 hour	69 (29.2)
1 - 3 hours	126 (53.4)
3 - 5 hours	20 (8.5)
More than 5 hours	7 (3)

Table . Digital confidence in handling digital media (n=236).

	Very insecure, n (%)	Rather insecure, n (%)	Partly insecure, n (%)	Rather confident, n (%)	Very confident, n (%)
How confident are you when handling digital media?	16 (6.8)	17 (7.2)	56 (23.7)	92 (39)	55 (23.3)
How confident are you when handling internet platforms?	20 (8.5)	21 (8.9)	61 (25.8)	81 (34.3)	53 (22.5)
How confident are you when handling digital devices?	20 (8.5)	16 (6.8)	54 (22.9)	92 (39)	54 (22.9)

Table . Knowledge regarding internet-based programs to promote health (n=236).

	I do not agree at all, n (%)	I rather do not agree, n (%)	Neither, n (%)	I rather agree, n (%)	I fully agree, n (%)
I can imagine what that might be	27 (11.4)	45 (19.1)	35 (14.8)	97 (41.1)	32 (13.6)
I know how those programs work	36 (15.3)	66 (28)	53 (22.5)	67 (28.4)	14 (5.9)
I know where to find these programs	35 (14.8)	50 (21.2)	39 (16.5)	84 (35.6)	28 (11.9)

eHealth Literacy

As presented in Table S2 in Multimedia Appendix 1, the statistical analysis of the eHealth literacy questionnaire results showed that the majority (n=152, 63.5%) of the study participants rather or fully agreed to know how to find internet platforms with helpful information regarding their overall health. Moreover, a total of 71.2% (n=171) of participants stated to knew how to use the internet as a useful tool to get answers to their questions. When asked, whether the participants thought they were able to critically evaluate the credibility of health information on the internet, a total of 64% (n=152) of participants stated that they were able to do so. However, only 40% (n=96) of the participants stated to knew digital sources available for health information, and 47.9% (n=115) of

XSL•FO
RenderX

participants felt secure enough to use information from the internet in order to make health-related decisions (Table 5).

Table . eHealth literacy (n=236).

	I do not agree at all, n (%)	I rather not agree, n (%)	Neither, n (%)	I rather agree, n (%)	I fully agree, n (%)
I know how to find internet platforms with helpful health information (n=236)	17 (7.2)	30 (12.7)	39 (16.5)	89 (37.7)	61 (25.8)
I know how to use the internet to get answers to my questions (n=236)	16 (6.8)	28 (11.9)	24 (10.2)	108 (45.8)	60 (25.4)
I know which sources for health information are available (n=235)	18 (7.7)	45 (19.1)	45 (19.1)	90 (38.3)	37 (1.7)
I know where I can find helpful health information on the internet (n=236)	18 (7.6)	32 (19.1)	39 (16.5)	103 (43.6)	44 (18.6)
I know how to use health information from the internet to help me (n=235)	24 (10.2)	48 (20.4)	50 (21.3)	88 (37.4)	25 (10.6)
I can critically evaluate health information on the internet (n=236)	17 (7.2)	23 (9.7)	45 (19.1)	101 (42.8)	50 (21.2)
I can distinguish between questionable and trustworthy internet resources with health information (n=236)	19 (8.1)	34 (14.4)	41 (17.4)	105 (44.5)	37 (15.7)
I feel secure to use information from the internet to make decisions regarding my health (n=236)	27 (11.4)	48 (20.3)	48 (20.3)	78 (33.1)	35 (14.8)

Age-Related Differences in Internet Use and Digital Confidence

As Figure 1 shows, using the internet for more than one hour a day, particularly for 1 - 3 hours per day, was the most common answer when asked about the duration of daily internet use. This result can be seen throughout all age groups. A longer duration of daily internet use of at least 3 hours was more common in participants younger than 55 years than in older groups. When analyzing confidence in handling digital media, it becomes clear that with older age, starting from the age group of 65 - 74-year-olds, the share of “partly confident,” “rather not confident,” and “not confident” individuals slightly increases, while still more than 50% (n=120) of the population remains “rather confident” or “very confident” (Figure 2). The share of at least “partly confident” participants is even higher in terms of confidence in handling digital devices such as smartphones and computers (Figure 3). This is again apparent throughout all age groups, even the oldest group of aged older than 75 years, contains less than 20% (n=48) of participants who are “not confident at all” in handling digital devices (Figure 3). Finally, confidence in handling internet platforms was generally lower than confidence in handling digital devices, and the share of participants who were “very confident,” as well as “rather confident,” decreased with older age. In the oldest age group of aged older than 75 years, 42% (n=101) of participants were “very confident” or “rather confident” in using internet platforms (Figure 4). Using the Kruskal Wallis test to compare internet use rates and digital confidence between the 4 age groups, there was no statistically significant difference for any of the examined items (>.05). Testing for correlations between the participants age and digital confidence, as well as daily internet use, the Spearman’s-Rho correlation shows the only statistically significant result regarding the confidence in using internet platforms (ρ [rho]) of –.128; *P*=.048). In addition, there was no age-related statistically significant difference in the levels of confidence in using mobile technologies (ρ of –0.70; *P*=.28), digital media (ρ of –0.96; *P*=.14), or daily internet use (ρ of 0.36; *P*=.58).

XSL•FO
RenderX

Figure 1. Daily internet use in relation to age groups (indicated in years along the x-axis).

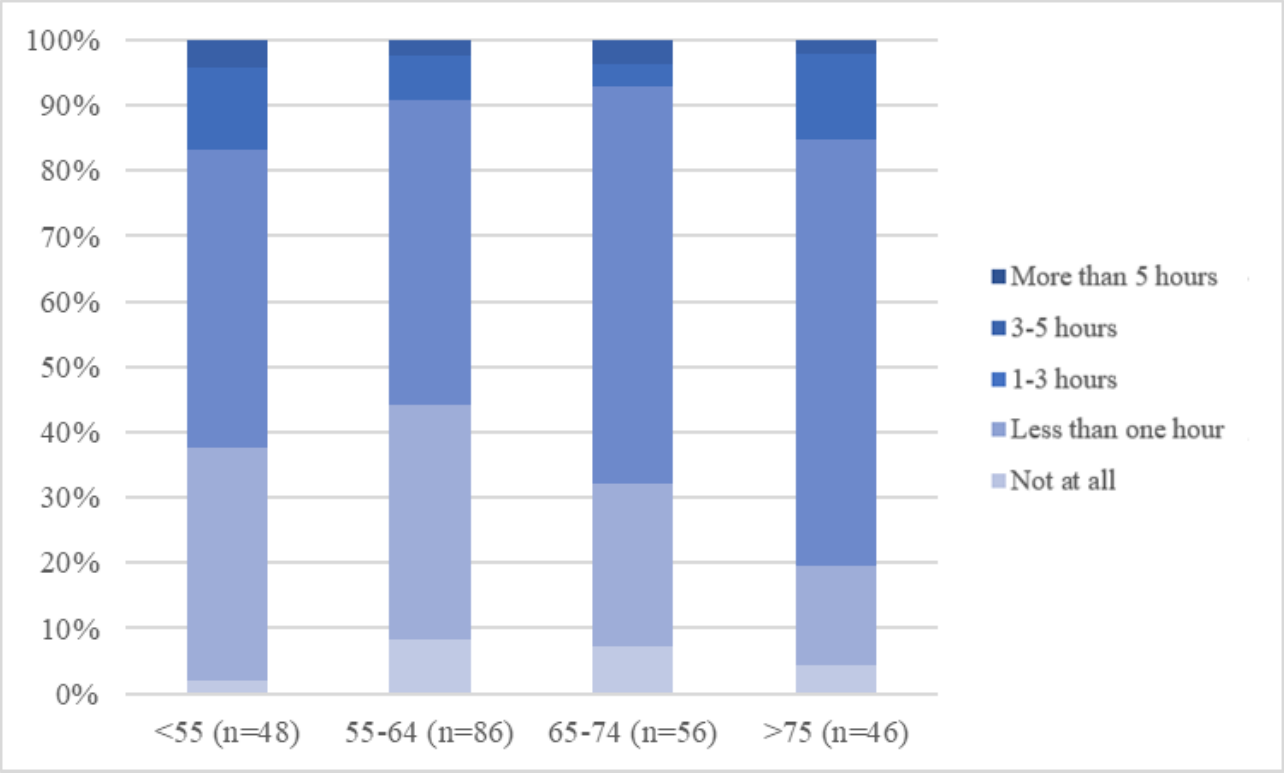

Figure 2. Confidence when handling digital media in relation to age groups (indicated in years along the x-axis).

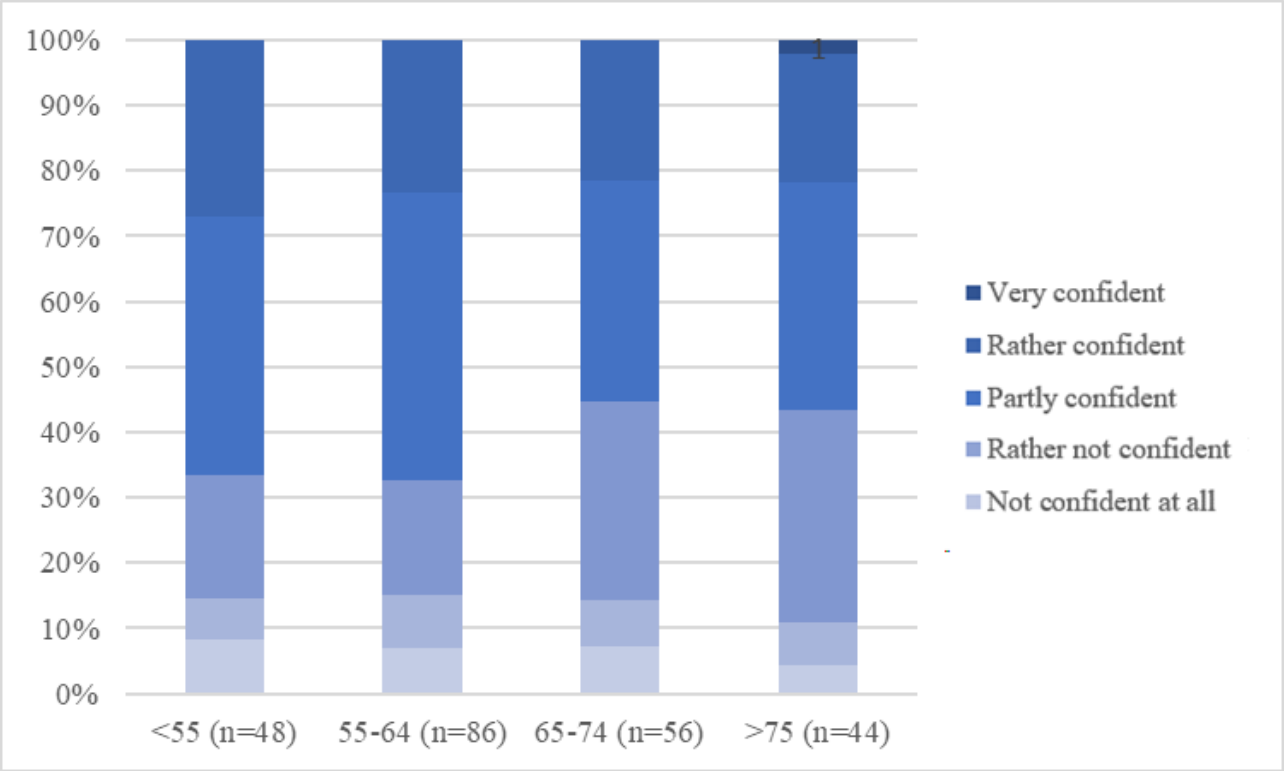

XSL•FO
RenderX

Figure 3. Confidence when handling digital devices in relation to age groups (indicated in years along the x-axis).

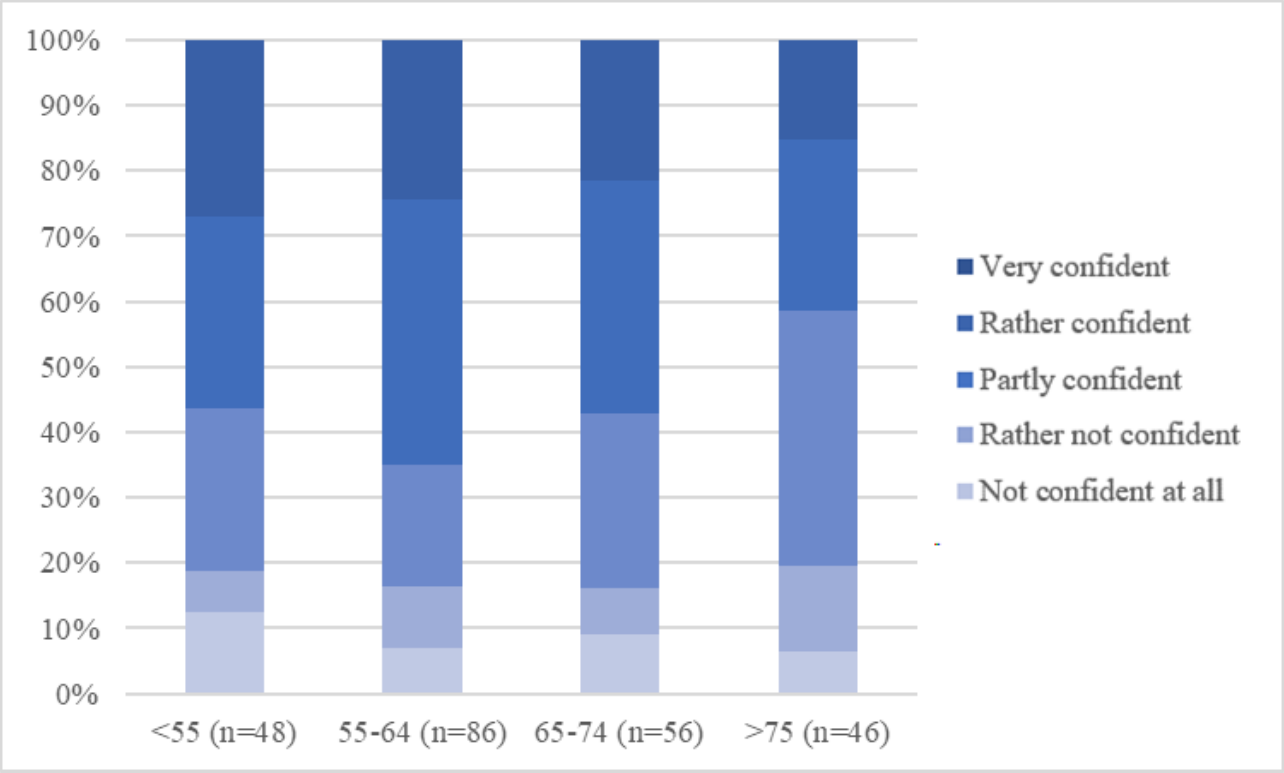

Figure 4. Confidence when handling internet platforms in relation to age groups (indicated in years along the x-axis).

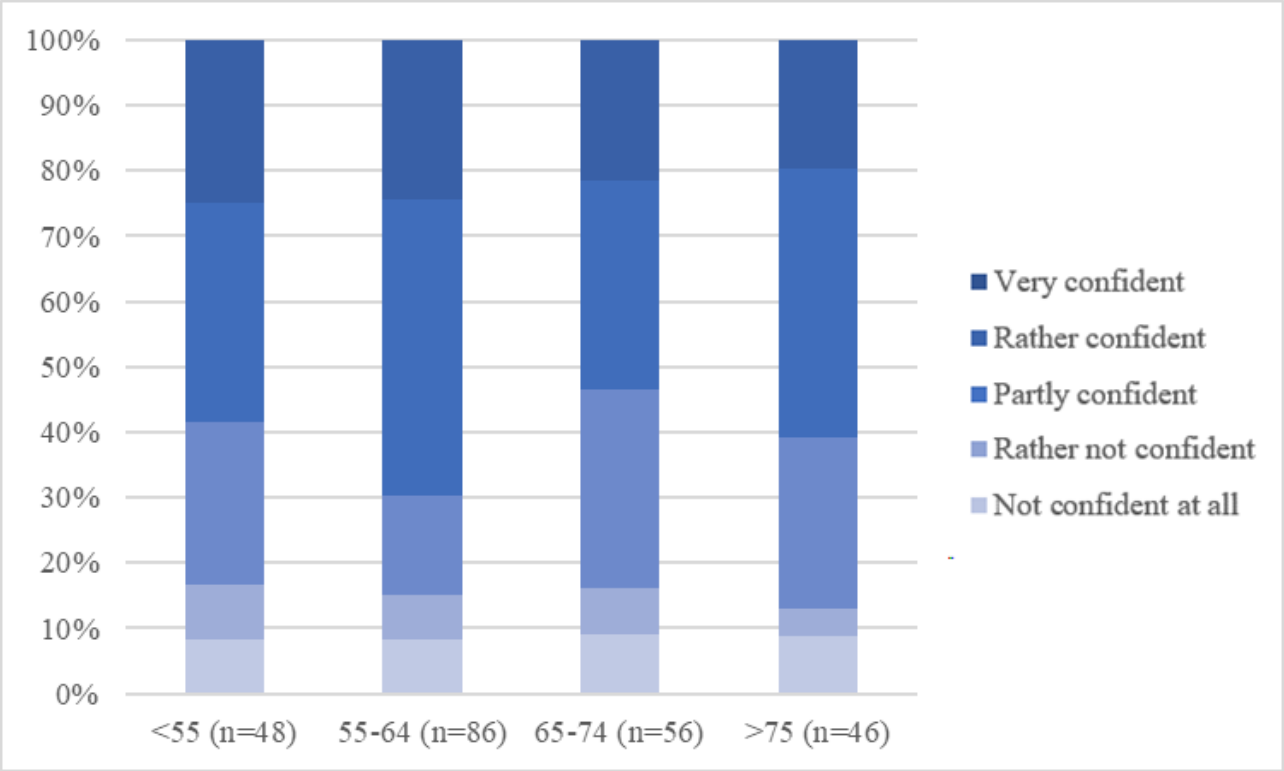

Gender-Specific Differences in the Level of Digital Confidence

As shown in Figure 5, we compared the differences in the levels of digital confidence in male and female participants. It is striking, that in all 3 categories, the levels of digital confidence seem to be at least slightly higher within the male subgroup. This difference becomes the most apparent comparing the levels of confidence when using internet platforms, with only about 10% (n=24) of the male participants feeling "rather not" or "not confident," while about 25% (n=60) of the female participants feel that way. Moreover, the share of participants who stated to be "very confident" in either of the 3 subcategories is relevantly higher in the male patients.

XSL•FO
RenderX

Figure 5. Gender-specific digital confidence.

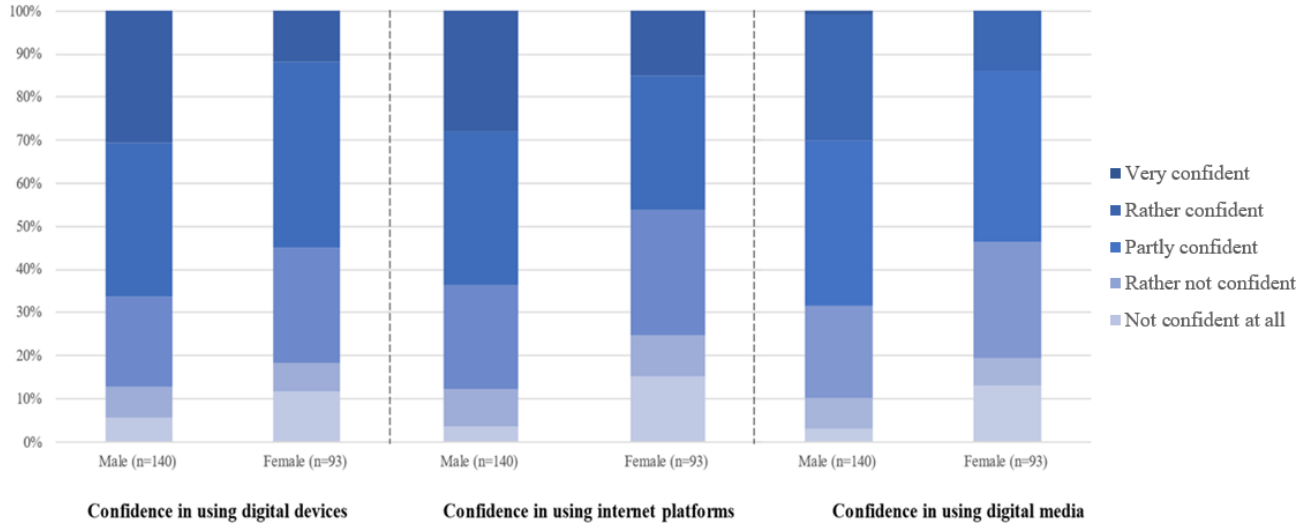

Discussion

Internet Use and Digital Knowledge

We conducted an analysis of current internet use rates among individuals affected by cardiovascular disease, examining their attitudes toward the internet and mobile technology, with a specific focus on digital health interventions. Additionally, we assessed participants' confidence in using digital technology and their ability to find credible health information on the internet.

Recent years have witnessed a notable increase in internet and smartphone use, particularly among the older population. From 2009 to 2019, the percentage of German citizens older than 65 years using the internet increased significantly, from approximately 30% to over 67% [17]. This trend was reflected in our trial, where only less than 6% (n=14) of the participants reported not using the internet for private purposes at all. Furthermore, the majority of the study collective (n=156, 64.9%) showed substantial daily internet use, spending at least one hour a day on the internet. Despite seemingly high internet use rates, knowledge about digital health interventions and internet-based health programs appeared to be limited. Only 34.3% (n=82) of participants indicated to understood how these programs worked, and merely 47.5% (n=114) of participants knew where to find them. This knowledge gap might be attributed to the potential lack of accessibility of digital interventions and a scarcity of programs that align with the specific needs of this patient demographic [18,19]. In the future, further research on patients' specific concerns and barriers regarding implementing digital health interventions will be necessary to gain a more in-depth understanding of the current low utilization rates.

Digital Confidence

Confidence levels in using digital media, internet platforms, and digital devices were notably high within this patient population, with over 60% (n=144) of participants expressing confidence in these 3 domains. Participants demonstrated particular assurance in handling digital media and devices. This heightened digital confidence is a crucial factor when contemplating the implementation of digital health interventions in this specific group, as a lack of confidence in digital technology is often cited as a significant barrier to integrating eHealth technologies into patients' chronic disease management [20,21]. Notably, these results deviate from prior research, which has frequently indicated that older patients, in particular, harbored doubts about eHealth technologies and felt insecure about using them [8]. Given the mean age of 61.9 (SD 8.9) years of the study participants, the observed high levels of digital confidence may signify the rapidly increasing rates of smartphone ownership and internet use among the older population [17]. Supporting this hypothesis, even in age-adjusted analyses, participants displayed sustained high levels of digital confidence, with most individuals feeling assured in handling digital media and devices. Only a minority expressed a lack of confidence in using these technologies. However, to ensure the inclusion of these individuals and provide them with an opportunity to benefit from digital health interventions, additional education on the matter will be imperative.

eHealth Literacy

To derive benefits from digital health solutions and effectively integrate them into daily life, possessing eHealth literacy is another crucial prerequisite [22]. Especially, in this study collective of patients with cardiovascular disease with a high prevalence of cardiovascular comorbidities and major risk factors, eHealth literacy is indispensable in order to use the rising possibilities that come with digitalization in chronic disease management.

eHealth literacy is defined as "the ability to seek, find, understand, and appraise health information from electronic sources and apply the knowledge gained to addressing or solving a health problem" [23]. It also involves the capability to discern between more or less credible sources of information. A common barrier to the adoption of digital health interventions is often the identification of trustworthy sources, leading to concerns about the security and privacy of these interventions [18,24]. In this trial, it was evident that most participants knew how to use the internet to find valuable health information and answers to health-related queries. However, it is noteworthy that only 40% (n=96) of the participants agreed that they knew about the various kinds of sources available for health

XSL•FO
RenderX

information on the internet, with only 1.7% (n=4) fully agreeing. This lack of awareness could be a contributing factor to the low rates of prior engagement with digital health interventions. Previous studies have highlighted that the lack of knowledge about eHealth interventions remains a pertinent barrier, especially among older patients who may not be adequately informed about the diverse options available to receive health support [25]. The trial's results indicated moderate levels of eHealth literacy. Most of the participants (n=154, 64%) felt capable of critically evaluating health information on the internet and distinguishing between questionable and trustworthy web-based resources. Nevertheless, only 47.9% (n=115) of the participants felt secure or reasonably secure in incorporating information from the internet into their health-related decision-making. Previous studies have indicated that even if health information is deemed credible, its implementation often falters due to the digital presentation, which is frequently described as "not user-friendly," "not meaningful," or generally challenging to comprehend [24]. One reason for this may be the sense of depersonalization of health information when presented digitally rather than in a face-to-face interaction, making it harder for individuals to connect with and apply relevant information [20].

Outlook—Challenges and Opportunities of Digital Health Interventions

To address patients' concerns regarding the security and reliability of health information on the internet, Germany took a significant step toward the digitalization of the health care system in 2019 by introducing digital health applications (DiGA) into standard care and supporting their use with statutory health insurance funding [26]. The approval of a mHealth or eHealth technology as a DiGA is strictly regulated and quality assurance for both prescribing physicians and patients using the intervention has to be ensured on a high level [27]. As highlighted in this study, the challenge of finding credible sources for health information on the internet remains a major concern. Having clear criteria to assess the quality and validity of health information that can be found on the internet is crucial to use the full potential of today's digital technology [28]. Confidence in using internet platforms was generally lower in this patient population compared to confidence in using digital devices (Figures 3 and 4). Therefore, officially approved digital health interventions with proven health care benefits that can be prescribed by treating physicians present an opportunity to make them available to more patients and ensure the high quality and credibility of the provided health information. Although this approach appears promising, the level of acceptance by both patients and physicians still needs to be thoroughly evaluated [29].

Limitations

It should be noted that the eHealth literacy score was self-assessed, and this study did not try to match the perceived eHealth literacy score to the actual abilities to use the internet to find health information and implement it in daily life. It has been shown that self-assessed eHealth literacy scores and actual eHealth literacy can differ when put into practice [30,31]. Digital use patterns and internet use were also assessed using self-reported data, which makes them subject to potential bias. For future research, it could be beneficial to include tracking individuals' daily use patterns in order to avoid potential bias and inaccuracies.

Moreover, since the study population was actively interested in participating in a trial revolving around mHealth technologies, a possible sampling bias should be considered when transferring the results to the general population. In addition to this, this trial particularly included patients diagnosed with coronary artery disease, which leads to a rather homogenous collective. This may limit the generalizability of these findings to a collective of patients with different cardiovascular diseases, such as heart failure, which was only present in one-fourth of the participants of this trial.

Conclusions

This study revealed that individuals with atherosclerotic cardiovascular disease were generally well-acquainted with using the internet and handling digital devices in all age groups. The internet was an integral part of their daily lives, leading to a high level of confidence in digital technologies. However, the adoption of digital health interventions remained relatively low, possibly reflecting reservations about incorporating internet-derived information into daily routines. This hesitancy might stem from uncertainties in identifying credible and trustworthy health sources on the internet, as well as the perceived lack of personalization in digital health interventions.

Despite their older age, the results suggest that patients with cardiovascular disease can be a suitable target group for digital health interventions, given their high internet use rates and digital confidence. The patient population appears representative in terms of cardiovascular comorbidities and the prevalence of cardiovascular risk factors. To address persistent skepticism toward digital technologies, it becomes crucial to incorporate personalized health information and adopt a patient-centered approach. Therefore, we need to delve into the specific needs and requirements of our patient collective regarding digital health interventions. This approach is essential to enhance the acceptance and effectiveness of digital health interventions within this patient population.

Acknowledgments

This work was financially supported by the Welcome-Back Grant of the University Duisburg-Essen Medical Faculty to Julia Lortz (grant D/401-00004). All authors declared that they had insufficient funding to support open access publication of this manuscript, including from affiliated organizations or institutions, funding agencies, or other organizations. JMIR Publications provided article processing fee (APF) support for the publication of this article.

XSL•FO
RenderX

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

AB conceived and designed the experiments, and analyzed and interpreted the data. AAM analyzed and interpreted the data. CR contributed reagents, materials, analysis tools, and data. DM performed the experiments. GU conceived and designed the experiments and performed the experiments. HV performed the experiments. JL conceived and designed the experiments, analyzed and interpreted the data, and contributed reagents, materials, analysis tools, or data. KP contributed reagents, materials, analysis tools, and data. LMJ analyzed and interpreted the data. TR conceived and designed the experiments and analyzed and interpreted the data. All authors wrote the paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary tables.
[DOCX File, 12 KB - formative_v9i1e63057_app1.docx]

References

1. The top 10 causes of death. WHO. 2020. URL: https://www.who.int/news-room/facts-sheets/details/the-top-10-causes-of-death [accessed 2024-02-16]
2. Hare DL, Toukhsati SR, Johansson P, Jaarsma T. Depression and cardiovascular disease: a clinical review. Eur Heart J 2014 Jun 1;35(21):1365-1372. [doi: 10.1093/eurheartj/eht462] [Medline: 24282187]
3. Visseren FLJ, Mach F, Smulders YM, et al. 2021 ESC Guidelines on cardiovascular disease prevention in clinical practice. Eur Heart J 2021 Sep 7;42(34):3227-3337. [doi: 10.1093/eurheartj/ehab484] [Medline: 34458905]
4. Chowdhury R, Khan H, Heydon E, et al. Adherence to cardiovascular therapy: a meta-analysis of prevalence and clinical consequences. Eur Heart J 2013 Oct;34(38):2940-2948. [doi: 10.1093/eurheartj/eht295] [Medline: 23907142]
5. Dall TM, Gallo PD, Chakrabarti R, West T, Semilla AP, Storm MV. An aging population and growing disease burden will require a large and specialized health care workforce by 2025. Health Aff (Millwood) 2013 Nov;32(11):2013-2020. [doi: 10.1377/hlthaff.2013.0714] [Medline: 24191094]
6. Gandapur Y, Kianoush S, Kelli HM, et al. The role of mHealth for improving medication adherence in patients with cardiovascular disease: a systematic review. Eur Heart J Qual Care Clin Outcomes 2016 Oct 1;2(4):237-244. [doi: 10.1093/ehjqcco/qcw018] [Medline: 29474713]
7. Schorr EN, Gepner AD, Dolansky MA, et al. Harnessing mobile health technology for secondary cardiovascular disease prevention in older adults: a scientific statement from the American Heart Association. Circ Cardiovasc Qual Outcomes 2021 May;14(5):e000103. [doi: 10.1161/HCQ.0000000000000103] [Medline: 33793309]
8. Nymberg VM, Bolmsjö BB, Wolff M, Calling S, Gerward S, Sandberg M. “Having to learn this so late in our lives…” Swedish elderly patients’ beliefs, experiences, attitudes and expectations of e-health in primary health care. Scand J Prim Health Care 2019 Mar;37(1):41-52. [doi: 10.1080/02813432.2019.1570612] [Medline: 30732519]
9. Ullrich G, Dönmez A, Mahabadi AA, et al. Effect of visual presentation of atherosclerotic carotid plaque on adherence to secondary preventive therapy using mHealth technologies (PreventiPlaque app): study protocol for a randomized controlled trial. Heliyon 2023 Mar;9(3):e14052. [doi: 10.1016/j.heliyon.2023.e14052] [Medline: 36915490]
10. Mach F, Baigent C, Catapano AL, et al. 2019 ESC/EAS guidelines for the management of dyslipidaemias: lipid modification to reduce cardiovascular risk: The Task Force for the management of dyslipidaemias of the European Society of Cardiology (ESC) and European Atherosclerosis Society (EAS). Eur Heart J 2019;41(1):111-188. [doi: 10.1093/eurheartj/ehz826]
11. Marsall M, Engelmann G, Skoda EM, Teufel M, Bäuerle A. Measuring electronic health literacy: development, validation, and test of measurement invariance of a revised German version of the eHealth literacy scale. J Med Internet Res 2022 Feb 2;24(2):e28252. [doi: 10.2196/28252] [Medline: 35107437]
12. Bäuerle A, Marsall M, Jahre LM, et al. Psychometric properties of the German revised version of the eHealth literacy scale in individuals with cardiac diseases: validation and test of measurement invariance. Digital Health 2023;9:20552076231194915. [doi: 10.1177/20552076231194915] [Medline: 37588160]
13. Bäuerle A, Mallien C, Rassaf T, et al. Determining the acceptance of digital cardiac rehabilitation and its influencing factors among patients affected by cardiac diseases. J Cardiovasc Dev Dis 2023 Apr 17;10(4):174. [doi: 10.3390/jcdd10040174] [Medline: 37103053]
14. Nurtsch A, Teufel M, Jahre LM, et al. Drivers and barriers of patients’ acceptance of video consultation in cancer care. Digital Health 2024;10:20552076231222108. [doi: 10.1177/20552076231222108] [Medline: 38188860]

XSL•FO
RenderX

15. Schröder J, Bäuerle A, Jahre LM, et al. Acceptance, drivers, and barriers to use eHealth interventions in patients with post-COVID-19 syndrome for management of post-COVID-19 symptoms: a cross-sectional study. Ther Adv Neurol Disord 2023;16:17562864231175730. [doi: 10.1177/17562864231175730] [Medline: 37255668]
16. Rentrop V, Damerau M, Schweda A, et al. Predicting acceptance of e-Mental health interventions in patients with obesity by using an extended unified theory of acceptance model: cross-sectional study. JMIR Form Res 2022 Mar 17;6(3):e31229. [doi: 10.2196/31229] [Medline: 35297769]
17. Ziebach M. Datenreport 2021—internetnutzung [Report in German]. 2021 URL: https://www.bpb.de/kurz-knapp/zahlen-und-fakten/sozialbericht-2024/553201/internetnutzung/ [accessed 2024-01-29]
18. Frederix I, Caiani EG, Dendale P, et al. ESC e-Cardiology Working Group Position Paper: overcoming challenges in digital health implementation in cardiovascular medicine. Eur J Prev Cardiol 2019 Jul;26(11):1166-1177. [doi: 10.1177/2047487319832394] [Medline: 30917695]
19. Torrent-Sellens J, Díaz-Chao Á, Soler-Ramos I, Saigí-Rubió F. Modelling and predicting eHealth usage in Europe: a multidimensional approach from an online survey of 13,000 European Union internet users. J Med Internet Res 2016 Jul 22;18(7):e188. [doi: 10.2196/jmir.5605] [Medline: 27450189]
20. Lie SS, Karlsen B, Oord ER, Graue M, Oftedal B. Dropout from an eHealth intervention for adults with type 2 diabetes: a qualitative study. J Med Internet Res 2017 May 30;19(5):e187. [doi: 10.2196/jmir.7479] [Medline: 28559223]
21. Herrera S, Salazar A, Nazar G. Barriers and supports in eHealth implementation among people with chronic cardiovascular ailments: integrative review. Int J Environ Res Public Health 2022 Jul 7;19(14):14. [doi: 10.3390/ijerph19148296] [Medline: 35886149]
22. Gee PM, Greenwood DA, Paterniti DA, Ward D, Miller LMS. The eHealth enhanced chronic care model: a theory derivation approach. J Med Internet Res 2015 Apr 1;17(4):e86. [doi: 10.2196/jmir.4067] [Medline: 25842005]
23. Norman CD, Skinner HA. eHealth literacy: essential skills for consumer health in a networked world. J Med Internet Res 2006 Jun 16;8(2):e9. [doi: 10.2196/jmir.8.2.e9] [Medline: 16867972]
24. Ware P, Bartlett SJ, Paré G, et al. Using eHealth technologies: interests, preferences, and concerns of older adults. Interact J Med Res 2017 Mar 23;6(1):e3. [doi: 10.2196/ijmr.4447] [Medline: 28336506]
25. Pywell J, Vijaykumar S, Dodd A, Coventry L. Barriers to older adults' uptake of mobile-based mental health interventions. Digital Health 2020;6:2055207620905422. [doi: 10.1177/2055207620905422] [Medline: 32110429]
26. Ludewig G, Klose C, Hunze L, Matenaar S. Digital health applications: statutory introduction of patient-centred digital innovations into healthcare [Article in German]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2021 Oct;64(10):1198-1206. [doi: 10.1007/s00103-021-03407-9] [Medline: 34529096]
27. Sauermann S, Herzberg J, Burkert S, Habetha S. DiGA—a chance for the German healthcare system. J Eur CME 2022;11(1):2014047. [doi: 10.1080/21614083.2021.2014047] [Medline: 34992948]
28. Duncker D, Billig H, Frey N, et al. Professioneller Umgang mit sozialen Medien in der Kardiologie [Article in German]. Kardiologie 2024 Mar;18(2):176-186. [doi: 10.1007/s12181-024-00665-0]
29. Uncovska M, Freitag B, Meister S, Fehring L. Patient acceptance of prescribed and fully reimbursed mHealth apps in Germany: an UTAUT2-based online survey study. J Med Syst 2023 Jan 27;47(1):14. [doi: 10.1007/s10916-023-01910-x] [Medline: 36705853]
30. van der Vaart R, van Deursen AJ, Drossaert CH, Taal E, van Dijk JA, van de Laar MA. Does the eHealth Literacy Scale (eHEALS) measure what it intends to measure? Validation of a Dutch version of the eHEALS in two adult populations. J Med Internet Res 2011 Nov 9;13(4):e86. [doi: 10.2196/jmir.1840] [Medline: 22071338]
31. Richtering SS, Morris R, Soh SE, et al. Examination of an eHealth literacy scale and a health literacy scale in a population with moderate to high cardiovascular risk: Rasch analyses. PLoS ONE 2017;12(4):e0175372. [doi: 10.1371/journal.pone.0175372] [Medline: 28448497]

Abbreviations

mHealth: mobile health

Edited by A Mavragani; submitted 08.06.24; peer-reviewed by J Kaswija, U Sinha; revised version received 21.10.24; accepted 07.11.24; published 19.03.25.

Please cite as:
Ullrich G, Bäuerle A, Vogt H, Mahabadi AA, Paldán K, Messiha D, Jahre LM, Rammos C, Rassaf T, Lortz J
Digital Health Literacy and Attitudes Toward eHealth Technologies Among Patients With Cardiovascular Disease and Their Implications for Secondary Prevention: Survey Study
JMIR Form Res 2025;9:e63057
URL: https://formative.jmir.org/2025/1/e63057
doi:10.2196/63057

XSL•FO
RenderX

© Greta Ullrich, Alexander Bäuerle, Hannah Vogt, Amir Abbas Mahabadi, Katrin Paldán, Daniel Messiha, Lisa Maria Jahre, Christos Rammos, Tienush Rassaf, Julia Lortz. Originally published in JMIR Formative Research (https://formative.jmir.org), 19.3.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Interprofessional Discussion for Knowledge Transfer in a Digital "Community of Practice" for Managing Pneumoconiosis: Mixed Methods Study

Varinn Avi Sood[1]; Heidi Rishel Brakey[2], MA; Orrin Myers[2], PhD; Xin Shore[2], PhD; Akshay Sood[2], MPH, MD

[1]Central New Mexico Community College, La Cueva High School, Albuquerque, NM, United States

[2]University of New Mexico Health Sciences Center, 1 University of New Mexico MSC 10 5550, Albuquerque, NM, United States

Corresponding Author:
Akshay Sood, MPH, MD
University of New Mexico Health Sciences Center, 1 University of New Mexico MSC 10 5550, Albuquerque, NM, United States

Abstract

Background: Pneumoconiosis prevalence is increasing in the United States, especially among coal miners. Contemporaneously with an increased need for specialized multidisciplinary care for miners, there is a shortage of experts to fulfill this need. Miners' Wellness ECHO (Extension for Community Health Outcomes) is a digital community of practice based on interprofessional discussion for knowledge transfer. The program has been demonstrated to increase participants' self-efficacy for clinical, medicolegal, and "soft" skills related to miners' health.

Objective: We aimed to examine characteristics associated with interprofessional discussions and suggest ways to strengthen knowledge transfer.

Methods: This mixed methods study used an exploratory sequential design. We video-recorded and transcribed ECHO sessions over 14 months from July 2018 to September 2019 and analyzed content to examine participant discussions. We focused on participants' statements of expertise followed by other participants' acceptance or eschewal of these statements (utterances). We conducted quantitative analyses to examine the associations of active participation in discussion (primary outcome variable, defined as any utterance). We analyzed the association of the outcome on the following predictors: (1) participant group status, (2) study time frame, (3) participant ECHO experience status, (4) concordance of participant group identity between presenter and participant, (5) video usage, and (6) attendance frequency. We used the generalized estimating equations approach for longitudinal data, logit link function for binary outcomes, and *LSMEANS* to examine least squares means of fixed effects.

Results: We studied 23 sessions with 158 unique participants and 539 total participants, averaging 23.4 (SD 5.6) participants per session. Clinical providers, the largest participant group, constituting 36.7% (n=58) of unique participants, were the most vocal group (mean 21.74, SD 2.11 average utterances per person-session). Benefits counselors were the least vocal group, with an average utterance rate of 0.57 (SD 0.29) per person-session and constituting 8.2% (n=13) of unique participants. Thus, various participant groups exhibited different utterance rates across sessions (P=.003). Experienced participants may have dominated active participation in discussion compared to those with less or intermediate experience, but this difference was not statistically significant (P=.11). When the didactic presenter and participant were from the same participant group, active participation by the silent group participants was greater than when both were from different groups. This association was not seen in vocal group participants (interaction P=.003). Compared to those participating by audio, those participating on video tended to have higher rates of active participation, but this difference was not statistically significant (P=.11).

Conclusions: Our findings provide insight into the mechanics of interprofessional discussion in a digital community of practice managing pneumoconiosis. Our results underscore the capacity of the novel ECHO model to leverage technology and workforce diversity to facilitate interprofessional discussions on the multidisciplinary care of miners. Future research will evaluate whether this translates into improved patient outcomes.

(JMIR Form Res 2025;9:e67999) doi:10.2196/67999

KEYWORDS

digital community of practice; knowledge transfer; pneumoconiosis; telementoring; rural health care; rural professionals; multidisciplinary management; interprofessional discussion; miner health; health equity; mixed methods; digital health; public health; digital community; self-efficacy; quantitative analyses; quantitative; technology; multidisciplinary care; patient outcome

XSL•FO
RenderX

Introduction

Mining dust-related diseases (eg, pneumoconiosis) are increasing in rural United States, especially among coal miners [1-4]. The 2017 prevalence of radiographic pneumoconiosis for coal miners with over 25 years of underground mining experience was greater than 10%, double the prevalence from the late 1990s [1]. Similarly, the 2014 rate of complicated pneumoconiosis (a particularly deadly form) among long-tenured underground coal miners was 1.1%, compared to 0.3% at its lowest point in the late 1990s [5].

Miners are medically vulnerable, underserved, and often underinsured. Many miners live in rural, remote, and mountainous locations in the Mountain West and Appalachia, constituting the "hot spot" regions of pneumoconiosis prevalence and mortality [6]. The prevalence of radiographic and complicated pneumoconiosis is the highest among all US miners in rural central Appalachia [7,8].

The re-emergence of pneumoconiosis presents unique challenges for rural communities. Compared with urban residents, those in rural areas have less access to outpatient pulmonary rehabilitation [9], primary care physicians [10], and pulmonologists [11]. Rural practitioners also face unique challenges, including professional isolation and complex patient profiles [12], and describe multiple barriers to knowledge acquisition, such as resources, personal costs, physical distance, and time [13].

Contemporaneously with an increase in the need for specialized multidisciplinary care for US miners, there is a shortage of experts to fulfill this need [14]. This critical gap is being addressed in New Mexico by innovative technology-based interventions such as mobile clinics equipped for telemedicine [15] and digital communities of practice for knowledge transfer. These interventions are recognized as rural and training innovations by the Rural Health Information Hub and the American Thoracic Society, respectively. The Miners' Wellness ECHO (Extension for Community Health Outcomes) Program is a digital community of practice, defined as a group of people who "share a concern or a passion for something they do and learn how to do better as they interact regularly" [16].

Project ECHO began in 2003 to connect primary care providers in rural New Mexico at "spoke" sites to University-based experts at a "hub" site so the former could effectively treat their patients with hepatitis C locally, instead of referring them to the University of New Mexico [17]. Since then, the ECHO movement has grown exponentially, spanned many health topics, and reached people around the globe. The Miners' Wellness ECHO Program provides structured longitudinal interprofessional telementoring to professionals caring for US miners in pneumoconiosis mortality hot spots, including respiratory therapists, home health professionals, benefits counselors, attorneys, clinical providers, and others.

Previous research of the Miners' Wellness ECHO established that participants in the pneumoconiosis mortality hot spots in the United States valued telementoring sessions for delivering relevant, evidence-based, balanced, and objective content [18]. Participants also showed a significant increase in self-efficacy scores overall and for domains related to clinical, medicolegal, and "soft" skills pertaining to miners' health, with "soft" skills including interpersonal and communication skills needed to navigate highly collaborative work in the care of miners [14]. The fact that existing participants showed greater improvement than fresh (newer) participants indicated that there might be a dose effect of telementoring on change in self-efficacy. In addition, greater improvement in outcomes among participants belonging to the clinical versus nonclinical professional groups might reflect the disproportionate emphasis on clinical didactics and case discussions in the program. Participants rated the collective efficacy of the virtual community of practice highly, stating feelings of being "closely knit," trustworthiness, and willingness to help each other [19].

The foundation of knowledge transfer in the program is the "all teach, all learn" model, which implies that every participant has unique knowledge to contribute and share during interprofessional discussions. Our previous research showed that most Miners' Wellness ECHO participants used knowledge ties from outside their participant group, emphasizing the interprofessional nature of knowledge transfer in miners' care [20]. The knowledge transfer was more efficient for participants with greater ties to rural-based miners than those with lesser ties [20]. Given the compensation challenges faced by US miners, benefits counselors and attorneys played an outsized role in knowledge transfer [18]. The findings from the preliminary evaluation establish the usefulness of this unique program and provide the study rationale for further examination of its learning processes, communication, and participation, an area of limited scholarship [21].

The objective of our study was to examine characteristics associated with interprofessional discussions and suggest ways to strengthen knowledge transfer. The long-term goal of this study is to reduce health inequity through greater investment in interprofessional telementoring efforts that promote collaborative health care in medically underserved mining communities by fusing technology with specialized multidisciplinary expertise. This approach may help rural communities counter the re-emergence of the pneumoconiosis epidemic by ameliorating their shortage of skilled expertise in mining-related diseases.

Methods

ECHO Model

The ECHO approach differs from traditional telemedicine, where clinical providers (eg, physicians, physician assistants, and advanced practice providers) assume short-term direct care of individual patients. Further, unlike webinars or conventional didactic lectures, the ECHO model provides a digital discussion of cases with expert panels in real time that is highly contextualized and thereby fulfills key learning theories, such as deliberate practice [22], social cognitive theory [23], and situated learning and communities of practice [24]. As detailed in a previous publication [18], the ECHO model is based on the following key principles: (1) use of internet-based technology for multipoint videoconferencing, which helps leverage scarce

XSL•FO
RenderX

resources; (2) use of an established evidence-based disease-management model that has been demonstrated to improve outcomes by reducing variation in processes of care and sharing best practices [17,25-27]; (3) use of case-based learning, based on discussion with experts and peers; (4) use of a digital community of practice, which emphasizes reciprocity in the sharing of skills and information; and (5) use of an internet-based database (ie, iECHO software [Hangzhou IECHO Science & Technology Co, Ltd]) to monitor outcomes.

Miners' Wellness ECHO Structure

Professionals involved in miners' health are invited to attend the program using an emailed web link connected to a REDCap (Research Electronic Data Capture; Vanderbilt University) survey [28]. As published previously [18], the ECHO sessions are scheduled at the same time twice a month for 75 minutes, adhering to a standard format. Following a facilitator's initial 10 minutes of introduction and announcements, an invited expert delivers a 15-minute didactic followed by a 20-minute facilitated question-answer session and a 30-minute digital facilitated case discussion. The medical director of the Miners' Wellness ECHO program is trained to facilitate group discussions during a 3-day immersion program and is periodically retrained [29]. The ECHO format, employing adult learning principles, focuses more on active learning through discussion than on didactic training. Upon completing a session-specific survey, participants can receive continuing education credits without charge. A multidisciplinary committee of experts follows a structured curriculum that is continually adapted based on the needs of the virtual community of practice, which are identified through a review of the continuing evaluation reports.

Attendance at ECHO sessions is open and voluntary. Through session attendance, participants can participate in the didactic and case discussions, provide information and insight from their own experience, and receive mentoring from their peers and an expert panel. Outside the program sessions, participants continue to have access to peers and experts for urgent consultation requests via telephone or email. Over time, with iterative practice and feedback, participants gain additional expertise, and become more confident in their skill sets related to caring for miners in their respective fields [14,18,20]. Recorded didactic sessions are made available through a web-based archive. The didactic and case discussion topics were not repeated during this study's time frame. All cases presented during this study's time frame were resolved using the collective expertise of the multidisciplinary community of practice, which included experts. The University of New Mexico has a sustainable business model for its ECHO Programs, with the cost of personnel shared between university budgets and grants and contracts.

Study Design and Recruitment

This 14-month exploratory sequential mixed methods study involved participants in the Miners' Wellness ECHO Program from July 2018 to September 2019. Participants are invited from across the nation by email, flyers, and word of mouth referrals to join the ECHO sessions. Their session participation is voluntary. At the start of the session, participants are informed that the session is video recorded. Together, the University of New Mexico School of Medicine and its rural partner Miners' Colfax Medical Center constitute the hub site of experts. The spoke partner sites are located across the pneumoconiosis mortality hot spot regions of the United States [18]. Although participants were allowed to attend multiple sessions, our analyses emphasize cross-sectional comparisons based on participant characteristics.

We asked the following 5 research questions on common ECHO processes based on the experience of this study's team, which are also related to the engagement theory model that focuses on participation, engagement, and learning within communities of practice [30]:

1. Do some participant groups make more utterances than others (eg, do clinical providers make more utterances than benefit counselors)?
2. Do participants who are more experienced with ECHO have greater rates of active participation in discussion than those with lower levels of experience?
3. Does the participant group identity of the didactic presenter predict active participation in discussion from the same participant group during that session (eg, if a didactic presenter is an attorney, are other attorneys more likely to participate in discussion in that session)?
4. Does video usage predict active participation in discussion, as compared to audio usage?
5. Does session attendance frequency predict active participation in discussion?

Video Recording and Transcription

Each ECHO session during this study time frame was video recorded and professionally transcribed. Participants were informed of the recording at the start of each session.

Summative Qualitative Content Analysis

Using an exploratory approach from a naturalistic point of view, we conducted summative qualitative content analysis to code interprofessional discussions among ECHO participants in each video. Specifically, we examined participants' verbal utterances, focusing on statements of expertise followed by acceptance or eschewal of the statements by other participants, using latent analysis for interpretation of content. We did not use a guiding theory, as this was exploratory. See Table 1 for the qualitative codebook with definitions of each. See Multimedia Appendix 1 for examples of conversations demonstrating statements of expertise and responses of acceptance and eschewal.

XSL•FO
RenderX

Table . Qualitative and quantitative content analysis codebook, from the mixed methods study of Miners' Wellness ECHO[a] participants, July 2018 to September 2019.

Variable, variable label, and variable code value				Definition or example of code value	Level of measurement
Qualitative codebook					
	Utterance type[b]				
		What type of utterance did the participant make?			
			1=Statement of expertise	Defining solutions to problems or answering questions. Anyone who speaks as an expert is an expert.	Nominal
			2=Acceptance of a statement of expertise	Others' acceptance of a statement of expertise. Includes adding to statement (expanding is acceptance) or verbalizations such as "That's great!" "Thank you so much for that." "That's really helpful!"	Nominal
			3=Eschewal of a statement of expertise	Others' eschewal in response to a statement of expertise. Eschewal can include providing new "expertise" that contradicts statements in disagreement with previously stated opinions, information, evidence, or statements. Can include partial disagreement or eschewal.	Nominal
			4=Neutral or no response to a statement of expertise	Neutral or no response. For example, if a speaker moves to next question or topic.	Nominal
Quantitative codebook					
	Number of utterances				
		How many utterances did the participant make during the session?		Total number of expertise, acceptance, eschewal, or neutral statements during a session.	Ordinal
	Participation				
		Did this person participate in a session?			
			0=No	At least 1 utterance during a session	Nominal
			1=Yes	At least 1 utterance during a session	Nominal
	Attendance				
		Did this person attend a session?			
			0=No	Logged into a session	Nominal
			1=Yes	Logged into a session	Nominal
	Session				

XSL•FO
RenderX

Variable, variable label, and variable code value			Definition or example of code value	Level of measurement
	What is the session number?		23 sessions from July 2018 to September 2019	Ordinal
Video				
	Did the participant use their video camera during the session?			
		1=Audio only	Use of video camera during a session	Nominal
		2=Video	Use of video camera during a session	Nominal
		3=Unknown	Use of video camera during a session	Nominal
Presenter				
	Was the participant a didactic presenter during the session?			
		0=No	__[c]	Nominal
		1=Yes	—	Nominal
Participant group				
	What participant group was the person a part of?			
		1=Attorney	Self-defined participant group status	Nominal
		2=Clinical provider	Self-defined participant group status	Nominal
		3=Home health professional	Self-defined participant group status	Nominal
		4=Respiratory therapist	Self-defined participant group status	Nominal
		5=Benefits counselor	Self-defined participant group status	Nominal
		6=Other	Self-defined participant group status	Nominal
Participant ECHO experience status				
	What joining group was the person a part of?			
		1=Experienced	Participant joined before the start of this study, May 9, 2018.	Nominal
		2=Intermediate	Participant joined at the start of this study, May 9, 2018.	Nominal
		3=Less experienced	Joined after September 12, 2018.	Nominal

[a]ECHO: Extension for Community Health Outcomes.

[b]Utterances include statements of expertise, acceptance, and eschewal. Additional details are provided in Tables S1 and S2 in Multimedia Appendix 1.

[c]Not applicable.

We realize that evaluating others' acceptance or eschewal can be subjective and difficult to determine by reading transcripts. Additionally, body language and tone of voice can convey meaning that written words cannot. To minimize these risks, our qualitative team consisted of 4 people with varied and complementary areas of expertise: (1) the principal investigator (AS) who is an expert pulmonologist and medical director/session facilitator for the Miners' Wellness Tele-ECHO Clinic; (2) an expert qualitative methodologist (HRB); and (3) two experts of ECHO methods and processes (director of program operations, Rachelle Rochelle, MPA, and senior program manager, Stephen Murillo, MBA). This team of 4 analysts met frequently during the project to interpret session content and discuss coding, potential biases, and reflexivity. We concurrently coded video recordings of participant behavior alongside transcripts using NVivo (version 12; QSR

XSL•FO
RenderX

International). The senior analyst created the preliminary codebook and independently coded the presence of each code across all ECHO sessions. She met with the principal investigator and 2 other team members to review the trial coding of the first 2 sessions as a group, modifying coding and definitions as needed. The principal investigator independently reviewed 2 additional sessions, and the other analysts independently reviewed 5 sessions; therefore, 7 sessions were trial-coded by 3 (5 sessions) or 4 (2 sessions) team members. The team regularly met to discuss coding until they were 100% in agreement, at which point the senior analyst coded the remaining 16 sessions. She continued to meet with the team periodically to discuss her coding generally and anything that was coded to be "unclear" (as defined in Table 1), at which point they jointly decided on whether it was better suited for another code. We removed all statements of expertise, acceptance, and eschewal of the session facilitator (AS), so we only evaluated participants' utterances. This approach minimized the bias resulting from the effect of the role and relationships of the session facilitator (AS) in the data analysis. There were no unintended consequences from the program participation or analysis. Since thematic saturation was not the goal of this study, the sampling was guided by the number of sessions rather than sampling saturation.

Quantitative Data Analysis

We worked with our senior statistician (OM) to convert our qualitative coding from NVivo into quantitative coding in Excel (Microsoft Corp). We coded additional variables to be analyzed with SAS (version 9.4; SAS Institute Inc). See Table 1 for the quantitative codebook.

Outcome and Predictor Variables

We examined the primary outcome variable, "active participation in discussion," defined by making any utterances (which included statements of expertise, acceptance, and eschewals). We analyzed the association of the outcome on the following predictor variables: (1) participant group status, (2) study time frame, (3) participant ECHO experience status, (4) concordance of participant group identity between presenter and participant, (5) video usage, and (6) attendance frequency.

Participant groups included attorneys, benefits counselors, clinical providers, home health professionals, respiratory therapists, and others. The participants' ECHO experience status was determined by the participant's date of joining the ECHO Program. Participants' ECHO experience status was defined by (1) the experienced group: those who joined the Miners' Wellness ECHO before the start of this study, May 9, 2018; (2) the intermediate group: those who joined at the start of this study, May 9, 2018; and (3) the less experienced group: those who joined after September 12, 2018. The participant groups were also classified as "silent" and "vocal" based on their preliminary analysis of their rates of utterances during the ECHO sessions. Participant group identity was considered concordant when the didactic presenter and participant belonged to the same participant group. Video status was determined if the video camera was turned on for any duration during a session by the participant.

Analytic Strategy

The number of utterances by type (statements of expertise, acceptance, and eschewal) and overall for each participant and ECHO session were collected. An average number of utterances by participant group, session, and type were computed. Binary variables for any active participation in discussion (1=yes for ≥1 utterance in the session, 0=no for 0 utterances in the session) were made for each utterance type and overall. Frequencies and percentages for active participation variables and other categorical variables were compiled by utterance type, participant group, and session.

We used the generalized estimating equations approach to extend the generalized linear model to handle longitudinal data, including predictor variables. The logit link function was applied for binary outcomes. Repeated statements with the subject option identified each session as a cluster. *LSMEANS* were used to compute and compare least squares means of fixed effects. A 2-tailed *P* value less than .05 was considered statistically significant.

Ethical Considerations

This study was an arm of a larger study. This arm of this study included participants who consented and enrolled in the larger study and those who were not enrolled in the larger study. This arm was approved as exempt by the institutional review board, University of New Mexico Human Research Protections Office (HRRC#18 - 386). Informed consent was waived by the institutional review board for this arm, separate from the larger study. The study data were deidentified for analysis. No compensation was offered for this arm, separate from the larger study.

Results

Descriptive Characteristics

We studied 23 ECHO sessions during this study's time frame, with 158 unique participants and 539 total participants. Further, 78 (49.4%) participants attended a single session, 28 (17.7%) attended 2 sessions, 11 (7%) attended 3 sessions, 13 (8.2%) attended 4 - 5 sessions, 13 (8.2%) attended 6 - 10 sessions, and 15 (9.5%) attended 11 - 20 sessions (overall mean 3.4, SD 4.1 sessions). Individual session attendance averaged 23.4 (SD 5.6; median 24, IQR 8; range=13 - 36) participants per session. During these sessions, 23 participants presented 23 patient cases, and 23 invited experts presented 23 didactics.

Do Some Participant Groups Make More Utterances Than Others?

Clinical providers, the largest participant group constituting 36.7% (58/158) of unique participants, were also the most vocal group (mean 21.74, SD 2.11, average utterances per person-session; Table 2). Attorneys, while the smallest group in number (11/158, 6.9%), were the second most vocal group (mean 6.73, SD 4.29, average utterances per person-session). Benefits counselors and home health professionals were the least vocal groups, with average utterance rates of 0.57 (SD 0.29) and 2.86 (SD 2.24) per person-session, respectively, constituting 8.2% (13/158) and 22.8% (36/158) of unique

XSL•FO
RenderX

participants, respectively. Thus, the various participant groups exhibited significantly different rates of utterances across sessions (*P*=.003).

Table . Summary of participant groups and their participation in discussion, with utterance rates per person-session across 23 Miners' Wellness ECHO[a] sessions during this study's time frame from July 2018 to September 2019.[b,c]

Participant group	Total PAS[d], n	Total PRAS[e], n/N (%)	Unique PAS, n	Unique PRAS, n/N (%)	Total statements of expertise, n	Average expertise per person-session	Total statements of acceptance, n	Average acceptance per person-session	Total statements of eschewal, n	Average eschew per person-session	Total utterances across sessions, n	Average utterances per person-session
Attorney	54	54/539 (10)	11	11/158 (6.9)	77	3.5	58	2.64	13	0.59	148	6.73
Benefits counselor	87	87/539 (16.1)	13	13/158 (8.2)	6	0.26	7	0.3	0	0	13	0.57
Clinical provider	186	186/539 (34.5)	58	58/158 (36.7)	280	12.17	182	7.91	38	1.65	500	21.74
Home health professional	89	89/539 (16.5)	36	36/158 (22.8)	36	1.71	21	1	3	0.14	60	2.86
Other or unknown	63	63/539 (11.7)	27	27/158 (17.1)	56	2.43	48	2.09	7	0.3	111	4.83
Respiratory therapist	60	60/539 (11.1)	13	13/158 (8.2)	53	2.3	31	1.35	12	0.52	96	4.17
Total	539	539/539 (100)	158	158/158 (100)	508	22.09	347	15.09	73	3.17	928	40.35

[a]ECHO: Extension for Community Health Outcome.

[b]Utterances include statements of expertise, acceptance, and eschewal. Additional detail of utterances is provided in Table S1 in Multimedia Appendix 1.

[c]Details of individual ECHO session content are shown in Table S2 in Multimedia Appendix 1 to provide more context to data shown in Figure 1.

[d]PAS: participants across sessions.

[e]PRAS: participant representation in all sessions.

Some topics generated a lot of discussion (ie, high rates of utterances), while others did not (Figure 1). Overall, utterances per person-session did not change over time (*P*=.50, Figure 1). It should be noted that in session 17, there were no utterances—this is the only session facilitated by a different person who was an experienced ECHO participant and expert clinical provider but not a trained facilitator. As shown in Table 2, most utterances are statements of expertise, and most statements are accepted. As shown in Figure 1, some topics generate a lot of controversy during discussion (ie, high rates of eschewal), while others do not.

XSL•FO
RenderX

Figure 1. Rates of utterances across the 23 Miners' Wellness ECHO sessions during this study's time frame, July 2018 to September 2019. This figure shows the average number of utterances per person-session (in gray line) and the percentage breakdown of each type of utterance (teal for eschewal, gray for acceptance, and red for statements of expertise). ECHO: Extension for Community Health Outcomes.

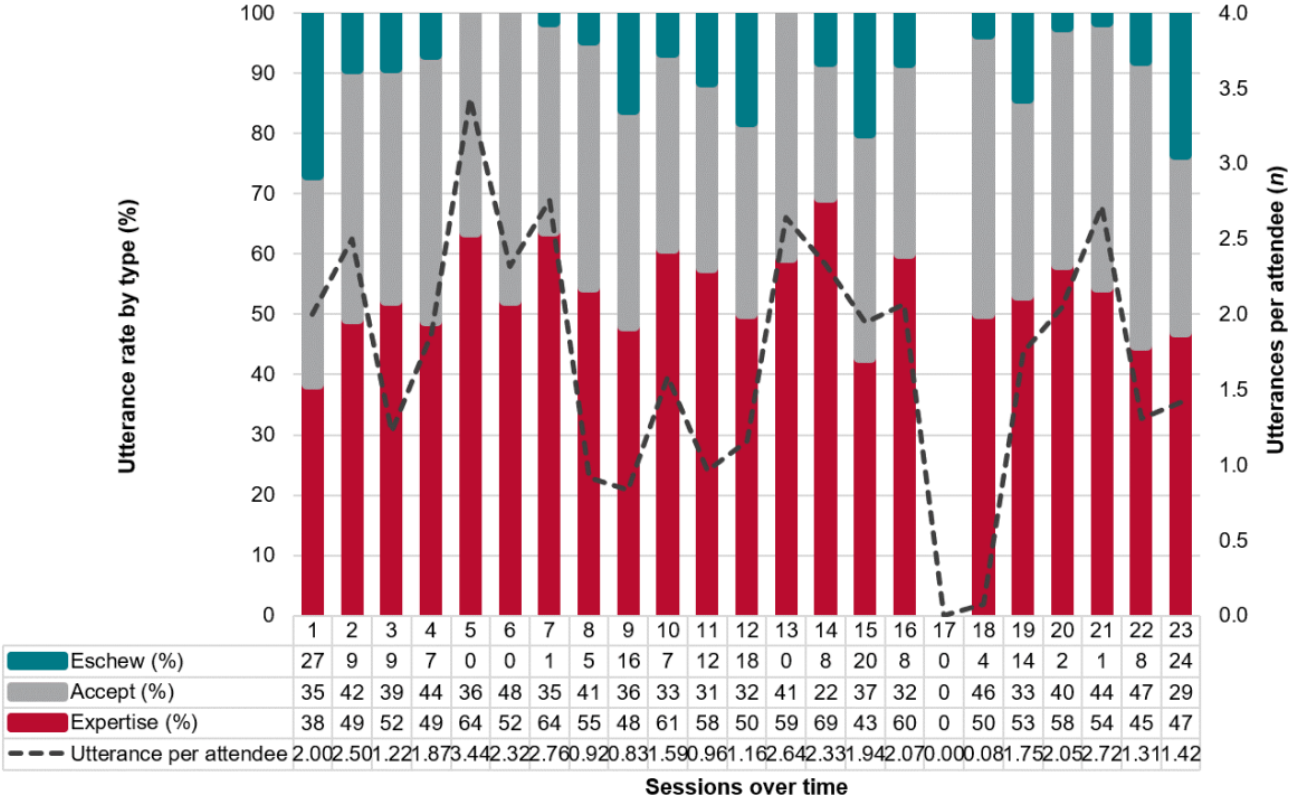

Do Participants Who Are More Experienced With ECHO Have Greater Rates of Active Participation in Discussion Than Those With Lower Levels of Experience?

Experienced and intermediate experienced groups accounted for 36.8% (n=104) and 26.9% (n=29) of active participation in discussion, respectively, while the active participation from the less experienced group accounted for 28.6% (n=40). Experienced participants may have, thus, tended to dominate active participation in discussion compared to participants with less or intermediate levels of experience, but this difference was not statistically significant (*P*=.11, Table 3).

Table . Association of participants' ECHO[a] experience with active participation in discussion, as measured by any utterance during an individual session, during this study's time frame from July 2018 to September 2019.

Group	Nonactive (n)	Active, n (%)	OR[b] (95% CI)	*P* value	Overall *P* value
Experienced	179	104 (36.8)	1.45 (0.88-2.39)	.14	.11
Intermediate experience	79	29 (26.9)	0.92 (0.52-1.64)	.77	Referent
Less experienced	100	40 (28.6)	Referent	__[c]	Referent

[a]ECHO: Extension for Community Health Outcomes.

[b]OR: odds ratio.

[c]Not applicable.

Does the Participant Group Identity of the Didactic Presenter Predict Active Participation in Discussion From the Same Participant Group During That Session?

Participant group identity was considered concordant when the didactic presenter and participant belonged to the same participant group. For this analysis, the participant groups were categorized as silent versus vocal, in which the benefits counselors and home health professionals constituted the former group and the clinical providers, attorneys and respiratory therapists in the latter group (with the other or unknown group excluded). When the didactic presenter and participant were from the same participant group, active participation in discussion (measured by any utterances within an individual ECHO session) by the silent group participants was 26.9% (n=7) compared to 9.5% (n=14) when the didactic presenter was from a different group (odds ratio 3.5, 95% CI 1.47-8.54, Table 4). Among vocal group participants, active participation was 46.4% (n=96) when presenters were in the same group as the

XSL•FO
RenderX

participant, compared to 46.2% (n=43) when presenters and participants were from different groups (odds ratio 1.01, 95% CI 0.58-1.75), with an interaction *P*=.003.

Table . Association of active participation in Miners' Wellness ECHO[a] session discussion by silent and vocal groups on concordance of participant group identity between didactic presenter and participant, during this study's time frame from July 2018 to September 2019.

Participant group concordance between presenter or participant[b]	Vocal group (n=300)			Silent group (n=173)			Interaction *P* value
	Nonactive (n)	Active, n (%)	OR[c] (95% CI)	Nonactive (n)	Active, n (%)	OR (95% CI)	
Concordant	111	96 (46.4)	1.01 (0.58-1.75)	19	7 (26.9)	3.5 (1.47-8.54)	.003
Discordant	50	43 (46.2)	Referent	133	14 (9.5)	Referent	Referent

[a]ECHO: Extension for Community Health Outcomes.

[b]For this analysis, the participant groups were categorized as silent versus vocal, in which the benefits counselors and home health professionals constituted the former group and the clinical providers, attorneys and respiratory therapists in the latter group (with the other or unknown group excluded).

[c]OR: odds ratio.

Does Video Usage Predict Active Participation in Discussion, as Compared to Audio Usage?

Compared to those joining a session by audio, those joining by video tended to have higher rates of active participation in discussion, as measured by any utterance during an individual session. However, this comparison did not reach statistical significance (Table 5, *P*=.11).

Table . Association of active participation in Miners' Wellness ECHO[a] session discussion on video camera use, during this study's time frame from July 2018 to September 2019.

Group	Nonactive participation (n)	Active participation, n (%)	OR[b] (95% CI)	*P* value
Audio[c]	110	54 (32.9)	Referent	.11
Video[c]	175	119 (40.5)	1.39 (0.93-2.07)	Referent

[a]ECHO: Extension for Community Health Outcomes.

[b]OR: odds ratio.

[c]Those whose audio and video status could not be determined (n=81) were excluded from the analysis.

Does Session Attendance Frequency Predict Active Participation in Discussion?

Unique participants' attendance ranged from 1 to 20 sessions and was categorized into 3 groups (Table 6). Overall, active participation in the discussion was significantly associated with attendance (*P*=.004); however, no dose response was noted.

Table . Association of active participation in Miners Wellness ECHO[a] session discussion on frequency of session attendance, during this study's time frame from July 2018 to September 2019.

Session attendance frequency	Nonactive participation (n)	Active participation, n (%)	OR[b] (95% CI)	*P* value	Overall *P* value
≤5 sessions	76	56 (42.4)	Referent	__[c]	.004
6 - 10 sessions	1	12 (92.3)	16.3 (2.1-128.9)	.008	Referent
11 - 20 sessions	4	11 (73.3)	3.7 (1.1-12.3)	.03	Referent

[a]ECHO: Extension for Community Health Outcomes.

[b]OR: odds ratio.

[c]Not applicable.

XSL•FO
RenderX

Discussion

Principal Findings

In this digital community of practice for managing pneumoconiosis with previously published efficacy outcomes [14,18,20], our study findings show that some participant groups are vocal while others are silent, based on utterance rates during interprofessional discussion. Some topics generate a lot of discussion (ie, relatively high rates of utterances), and some topics generate a lot of controversy during discussion (ie, relatively high rates of eschewal of statements of expertise). Inadequately trained facilitators may impede discussion during a session. Experienced participants may dominate the discussion and inhibit participation by less experienced participants, but this association did not reach statistical significance (P=.11, Table 3). When the didactic presenter and participant are from the same participant group, participation in discussion by the silent group participants rises significantly. Compared to those participating by audio, those participating by video camera may tend to have higher rates of participation in discussion, but this association did not reach statistical significance. Overall, active participation in the discussion was significantly associated with attendance frequency. Our study findings lend themselves to several best practice recommendations, as discussed below.

The primary objective of the Project ECHO movement is to decentralize knowledge for the care of patients through exchanging insights and information using the all teach, all learn principle. All participants have unique knowledge sets, and discussion inside and outside the ECHO session within and across participant groups facilitates the transfer of knowledge that would otherwise remain siloed within individual participant groups. Further, interprofessional discussion in the ECHO model may allow greater access to new and thought-provoking ideas and perspectives that foster learning and other growth-enhancing actions. Our study findings confirm the existence of silent participant groups that, by not fully engaging in discussions, may not be as effective in exchanging knowledge (including providing and acquiring knowledge) within the ECHO session. These groups need to be actively supported by a trained facilitator and by the judicious use of didactic presenters sharing participant group identity concordant with the silent group participants. However, our previous objective measurement of knowledge transfer using social network principles indicates that benefits counselors are among the groups most effective in knowledge transfer [18]. This prior finding would suggest that either benefits counselors convey relevant knowledge succinctly within the ECHO sessions or participate in knowledge transfer outside the strict bounds of the ECHO program via follow-up phone calls, emails, and virtual and physical meetings. This finding warrants further research as it is a possible indication that the informal “social network” system, perhaps provoked by or aided by the ECHO sessions, may be more important than and reinforce the formal sessions.

Project ECHO formally trains its facilitators on the best practices for managing discussions in a digital community of practice. Substituting a trained facilitator by another untrained expert can affect participation in the discussion. Experienced participants usually offer the greatest rates of statements of expertise, acceptance, and eschewal, reflecting their expert status in this complex field. However, the dominance of experienced groups can be avoided by the facilitator actively encouraging less experienced groups to make their minority opinions heard during discussion. Using video cameras during participation in this digital community of practice helps build community and promotes accountability and engagement at the cost of rural internet bandwidth. This practice should be encouraged, even though the findings did not reach statistical significance.

Our study has multiple strengths. To the best of our knowledge, our approach of studying interprofessional discussion in a digital “community of practice” for managing pneumoconiosis has never been used previously by another group of investigators. This study involves the innovative ECHO model intervention that addresses barriers to the care of miners by providing a multidisciplinary community of practice approach, which has been well studied in other diseases [17,25-27]. This study is topical and significant because it addresses a critical gap related to the emerging pneumoconiosis epidemic in rural United States. Since no one refused participation, there was likely no potential participation bias in this study. Since the ECHO model has been adopted nationwide and worldwide to care for patients with numerous chronic diseases, infrastructure already exists to rapidly scale the Miners’ Wellness ECHO Program nationwide and worldwide. Other strengths include the detailed contemporaneous interpretation of videos and transcripts of discussions; strict quality control in qualitative analysis; a mixed methods study design; and the simultaneous use of qualitative, quantitative, and subject content experts as coinvestigators.

Although our study focused on the ECHO program, which emphasizes chronic disease management with a strong focus on mentorship and collaborative learning, our findings have broader implications for other digital communities of practice that may be structured differently and operate differently. Examples include Elpha (Women in Tech Network), a community of practice for women in technology; Stack Overflow (Stack Exchange Inc) for programmers and developers; Digital Nomad World for nonprofessional digital nomads [31]; and various leadership communities of practice that emphasize succession planning and leadership skills development [32]. Such unrelated communities of practice can also benefit from our evidence-based findings and recommendations to reach out to silent groups, break down silos, provide facilitator training, and encourage active participation. On the other hand, the ECHO program could learn from other communities of practice on how to reward active participation and knowledge sharing (Stack Overflow) or create asynchronous platforms to supplement synchronous engagement (Digital Nomad World).

There are also limitations to this study. We cannot correlate interprofessional discussion for knowledge transfer to patient outcomes or changes in provider behavior. However, we have previously published a listing of qualitative changes our ECHO participants reported they would make in their practice, obtained as part of a continuing medical education survey requested at the end of each ECHO session [20]. Participants whose group identity was other or unknown were 11.7% (63/539) of the total

XSL•FO
RenderX

participant representation in all sessions. When analyzing participant group identity (as shown in Table 4), the other or unknown group was excluded. Although a small sample size raises the possibility of a type I error, individual professionals and teams of professionals trained in the ECHO model can reach a large number of miners, with the potential for creating exponential change. Miners impacted with pneumoconiosis do not participate in the ECHO model. This model characteristic is, however, not a limitation for the following reasons. The model ensures that knowledge is disseminated widely and effectively by focusing on provider education. By reaching out to more miner professionals, the ECHO model indirectly benefits more miners and is more scalable than traditional telemedicine models focusing on individual patients. Case-based learning, where providers present anonymized patient cases to a panel of experts, allows for in-depth discussion and learning without compromising patient confidentiality while maintaining the traditional provider-patient relationship [25,33].

Conclusions

Our findings provide insight into the mechanics of interprofessional discussion for knowledge transfer in a digital community of practice managing pneumoconiosis and potential recommendations to enhance the same. Our results underscore the capacity of the Project ECHO model to leverage technology and workforce diversity to facilitate interprofessional discussions on the multidisciplinary and complex care of miners and ultimately promote health equity among rural and medically underserved mining communities. Although this approach addresses a critical gap related to the emerging pneumoconiosis epidemic, future research will evaluate whether this translates into improved patient outcomes in mining communities, a priority need in rural United States.

Acknowledgments

Generative artificial intelligence was not used in any portion of this paper. This study was funded by the Alpha Foundation (grant AFC719-14 of US $250,000) and partly supported by the National Center for Advancing Translational Sciences, National Institutes of Health (UL1TR001449). The sponsors played no role in conducting, analyzing, and disseminating this study or reviewing and approving this paper for publication. We want to thank Rachelle Rochelle, MPA, and Stephen Murillo, MBA, who were integral to the collection of this data.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

All authors were involved with the following: (1) substantial contributions to the conception or design of this work (HRB, VAS, XS, OM, and AS) and the acquisition (HRB, XS, OM, and AS), analysis (HRB, XS, and OM), or interpretation of data for this work (HRB, VAS, XS, OM, and AS); (2) drafting this work or reviewing it critically for important intellectual content (HRB, VAS, XS, OM, and AS); (3) final approval of the version to be published (HRB, VAS, XS, OM, and AS); and (4) agreement to be accountable for all aspects of this work in ensuring that questions related to the accuracy or integrity of any part of this work are appropriately investigated and resolved (HRB, VAS, XS, OM, and AS).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Example conversations and details of individual ECHO (Extension for Community Health Outcomes) session content.
[DOCX File, 35 KB - formative_v9i1e67999_app1.docx]

References

1. Blackley DJ, Halldin CN, Laney AS. Continued increase in prevalence of coal workers' pneumoconiosis in the United States, 1970-2017. Am J Public Health 2018 Sep;108(9):1220-1222. [doi: 10.2105/AJPH.2018.304517] [Medline: 30024799]
2. Hall NB, Blackley DJ, Halldin CN, Laney AS. Continued increase in prevalence of r-type opacities among underground coal miners in the USA. Occup Environ Med 2019 Jul;76(7):479-481. [doi: 10.1136/oemed-2019-105691] [Medline: 31023786]
3. Laney AS, Attfield MD. Coal workers' pneumoconiosis and progressive massive fibrosis are increasingly more prevalent among workers in small underground coal mines in the United States. Occup Environ Med 2010 Jun;67(6):428-431. [doi: 10.1136/oem.2009.050757] [Medline: 20522823]
4. Laney AS, Blackley DJ, Halldin CN. Radiographic disease progression in contemporary US coal miners with progressive massive fibrosis. Occup Environ Med 2017 Jul;74(7):517-520. [doi: 10.1136/oemed-2016-104249] [Medline: 28408654]

XSL•FO
RenderX

5. Blackley DJ, Halldin CN, Laney AS. Resurgence of a debilitating and entirely preventable respiratory disease among working coal miners. Am J Respir Crit Care Med 2014 Sep 15;190(6):708-709. [doi: 10.1164/rccm.201407-1286LE] [Medline: 25221884]
6. Dwyer-Lindgren L, Bertozzi-Villa A, Stubbs RW, et al. Trends and patterns of differences in chronic respiratory disease mortality among US counties, 1980-2014. JAMA 2017 Sep 26;318(12):1136-1149. [doi: 10.1001/jama.2017.11747] [Medline: 28973621]
7. Almberg KS, Halldin CN, Blackley DJ, et al. Progressive massive fibrosis resurgence identified in U.S. coal miners filing for black lung benefits, 1970-2016. Ann Am Thorac Soc 2018 Dec;15(12):1420-1426. [doi: 10.1513/AnnalsATS.201804-261OC] [Medline: 30114941]
8. Arnold C. A scourge returns: black lung in Appalachia. Environ Health Perspect 2016 Jan;124(1):A13-A18. [doi: 10.1289/ehp.124-A13] [Medline: 26720594]
9. Fan VS, Giardino ND, Blough DK, Kaplan RM, Ramsey SD, Nett Research Group. Costs of pulmonary rehabilitation and predictors of adherence in the National Emphysema Treatment Trial. COPD 2008 Apr;5(2):105-116. [doi: 10.1080/15412550801941190] [Medline: 18415809]
10. Kirby JB, Yabroff KR. Rural-urban differences in access to primary care: beyond the usual source of care provider. Am J Prev Med 2020 Jan;58(1):89-96. [doi: 10.1016/j.amepre.2019.08.026] [Medline: 31862103]
11. Croft JB, Lu H, Zhang X, Holt JB. Geographic accessibility of pulmonologists for adults with COPD: United States, 2013. Chest 2016 Sep;150(3):544-553. [doi: 10.1016/j.chest.2016.05.014] [Medline: 27221645]
12. Miedema B, Hamilton R, Fortin P, Easley J, Tatemichi S. The challenges and rewards of rural family practice in New Brunswick, Canada: lessons for retention. Rural Remote Health 2009;9(2):1141. [Medline: 19496642]
13. Crockett LK, Leggett C, Curran JA, et al. Knowledge sharing between general and pediatric emergency departments: connections, barriers, and opportunities. CJEM 2018 Feb 22:1-9. [doi: 10.1017/cem.2018.7] [Medline: 29467040]
14. Sood A, Pollard C, Kalishman S, et al. Telementoring of healthcare teams in the care of miners. ATS Sch 2020 Dec 22;2(1):66-83. [doi: 10.34197/ats-scholar.2020-0073OC] [Medline: 33870324]
15. Evans K, Lerch S, Boyce TW, et al. An innovative approach to enhancing access to medical screening for miners using a mobile clinic with telemedicine capability. J Health Care Poor Underserved 2016;27(4A):62-72. [doi: 10.1353/hpu.2016.0182] [Medline: 27818414]
16. Wenger E. Communities of Practice: Learning, Meaning and Identity: Cambridge University Press; 1998.
17. Arora S, Thornton K, Murata G, et al. Outcomes of treatment for hepatitis C virus infection by primary care providers. N Engl J Med 2011 Jun 9;364(23):2199-2207. [doi: 10.1056/NEJMoa1009370] [Medline: 21631316]
18. Sood A, Assad N, Jarrell W, et al. A virtual community-of-practice approach by rural stakeholders in managing pneumoconiosis in the USA: A cross-sectional analysis. Rural Remote Health 2020 Aug;20(3):5784. [doi: 10.22605/RRH5784] [Medline: 32847365]
19. Files JA, Blair JE, Mayer AP, Ko MG. Facilitated peer mentorship: a pilot program for academic advancement of female medical faculty. J Womens Health (Larchmt) 2008;17(6):1009-1015. [doi: 10.1089/jwh.2007.0647] [Medline: 18681821]
20. Soller B, Myers O, Sood A. Transfer of knowledge on pneumoconiosis care among rural-based members of a digital community of practice: cross-sectional study. JMIR Form Res 2024 Jan 24;8:e52414. [doi: 10.2196/52414] [Medline: 38265861]
21. Ranmuthugala G, Plumb JJ, Cunningham FC, Georgiou A, Westbrook JI, Braithwaite J. How and why are communities of practice established in the healthcare sector? A systematic review of the literature. BMC Health Serv Res 2011 Oct 14;11:273. [doi: 10.1186/1472-6963-11-273] [Medline: 21999305]
22. Ericsson KA. Deliberate practice and acquisition of expert performance: a general overview. Acad Emerg Med 2008 Nov;15(11):988-994. [doi: 10.1111/j.1553-2712.2008.00227.x] [Medline: 18778378]
23. Bandura A. Social cognitive theory: an agentic perspective. Annu Rev Psychol 2001;52:1-26. [doi: 10.1146/annurev.psych.52.1.1] [Medline: 11148297]
24. Wenger E. How we learn. Communities of practice. The social fabric of a learning organization. Healthc Forum J 1996;39(4):20-26. [Medline: 10158755]
25. Arora S, Kalishman S, Thornton K, et al. Expanding access to hepatitis C virus treatment--Extension for Community Healthcare Outcomes (ECHO) project: disruptive innovation in specialty care. Hepatology 2010 Sep;52(3):1124-1133. [doi: 10.1002/hep.23802] [Medline: 20607688]
26. Bouchonville MF, Hager BW, Kirk JB, Qualls CR, Arora S. Endo echo improves primary care provider and community health worker self-efficacy in complex diabetes management in medically underserved communities. Endocr Pract 2018 Jan;24(1):40-46. [doi: 10.4158/EP-2017-0079] [Medline: 29368967]
27. Deming P, Arora S. Taribavirin in the treatment of hepatitis C. Expert Opin Investig Drugs 2011 Oct;20(10):1435-1443. [doi: 10.1517/13543784.2011.606214] [Medline: 21854301]
28. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research Electronic Data Capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381. [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]

XSL•FO
RenderX

29. Moss P, Hartley N, Russell T. Project ECHO®: a global cross-sectional examination of implementation success. BMC Health Serv Res 2024 May 3;24(1):583. [doi: 10.1186/s12913-024-10920-5] [Medline: 38702685]
30. Kearsley G, Shneiderman B. Engagement theory: a framework for technology-based teaching and learning. Educ Technol 1998;38(5):20-23.
31. Burton C. 6 community of practice examples (+ different types). Thinkific. 2022 Aug 16. URL: https://www.thinkific.com/blog/community-of-practice-examples/#non-professional [accessed 2025-02-22]
32. DePodesta M. The development of leadership communities of practice: utilizing the ADKAR framework. Nurs Adm Q 2024;48(3):225-233. [doi: 10.1097/NAQ.0000000000000648] [Medline: 38848484]
33. Zhou C, Crawford A, Serhal E, Kurdyak P, Sockalingam S. The impact of Project ECHO on participant and patient outcomes: a systematic review. Acad Med 2016 Oct;91(10):1439-1461. [doi: 10.1097/ACM.0000000000001328] [Medline: 27489018]

Abbreviations

ECHO: Extension for Community Health Outcomes
REDCap: Research Electronic Data Capture

Edited by A Mavragani; submitted 25.10.24; peer-reviewed by G Kernohan, W LaMendola; revised version received 23.01.25; accepted 24.01.25; published 13.03.25.

Please cite as:
Sood VA, Rishel Brakey H, Myers O, Shore X, Sood A
Interprofessional Discussion for Knowledge Transfer in a Digital "Community of Practice" for Managing Pneumoconiosis: Mixed Methods Study
JMIR Form Res 2025;9:e67999
URL: https://formative.jmir.org/2025/1/e67999
doi:10.2196/67999

© Varinn Avi Sood, Heidi Rishel Brakey, Orrin Myers, Xin Shore, Akshay Sood. Originally published in JMIR Formative Research (https://formative.jmir.org), 13.3.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Telemedicine Prescribing by US Mental Health Care Providers: National Cross-Sectional Survey

Mollie R Cummins[1,2], BSN, MSN, PhD; Julia Ivanova[2], PhD; Hiral Soni[2], MS, MBA, PhD; Zoe Robbins[1], DNP; Brian E Bunnell[2,3], PhD; Esteban López[2], MD, MBA; Brandon M Welch[2,4], MS, PhD

[1]College of Nursing, University of Utah, 10 S 2000 East, Salt Lake City, UT, United States

[2]Doxy.me, Inc, Charleston, SC, United States

[3]Department of Psychiatry and Behavioral Neurosciences, Morsani College of Medicine, University of South Florida, Tampa, FL, United States

[4]Biomedical Informatics Center, Public Health and Sciences, Medical University of South Carolina, Charleston, SC, United States

Corresponding Author:
Mollie R Cummins, BSN, MSN, PhD
College of Nursing, University of Utah, 10 S 2000 East, Salt Lake City, UT, United States

Abstract

Background: In the postpandemic era, telemedicine continues to enable mental health care access for many people, especially persons living in areas with mental health care provider shortages. However, as lawmakers consider long-term telemedicine policy decisions, some question the safety and appropriateness of prescribing via telemedicine, and whether there should be requirements for in-person evaluation, especially for controlled substances.

Objective: Our objective was to assess US telemental health care provider perceptions of comfort and perceived safety in prescribing medications, including controlled substances, via telemedicine.

Methods: We conducted a web-based, cross-sectional survey of US telemental health care providers who prescribe via telemedicine, using nonprobability, availability sampling of a national telehealth research panel from February 13 to April 28, 2024. We used descriptive statistics, visualization, and thematic analysis to analyze results. We assessed differences in response distribution by health care provider licensure type (physician vs nonphysician) and specialty (psychiatry vs nonpsychiatry) using the Mann-Whitney *U* test.

Results: A total of 115 screened and eligible panelists completed the survey. Overall, participants indicated high levels of comfort with prescribing via telemedicine, with 84% (102/115) of health care providers indicating they strongly agree with the statement indicating comfort in prescribing medications via telemedicine. However, participants indicated less comfort in prescribing if they have never seen a patient in person, or if the patient is located out-of-state. Most participants indicated they can safely prescribe controlled substances via telemedicine, without having previously provided care to a patient in person. However, 14.8% (17/115) to 19.1% (30/115) of health care providers (by schedule) felt that they could rarely or never safely prescribe controlled substances. There were some differences in perception of comfort and safety by licensure and specialty. Among controlled substance schedules, participants indicated the least perceived safety with schedule IV medications, and the most safety with schedule II and III medications.

Conclusions: These health care providers were highly comfortable prescribing both scheduled and unscheduled medications via telemedicine. Comfort and perceived safety with telemedicine prescribing varied somewhat by licensure type (physician vs nonphysician) and specialty (psychiatry vs nonpsychiatry). Perceived safety varied moderately for scheduled medications (controlled substances), especially for schedule IV and V medications. Participants indicated use of adaptive strategies to prescribe safely depending upon the clinical context. In ongoing efforts, we are analyzing additional survey results and conducting qualitative research related to telemedicine prescribing. A strong understanding of prescriber perspectives and experience with telemedicine prescribing is needed to support excellent clinical practice and effective policy making in the United States.

(JMIR Form Res 2025;9:e63251) doi:10.2196/63251

KEYWORDS

telemedicine; telehealth; telemental; provider; professional; experience; attitude; opinion; perception; perspective; prescribing; prescription; drug; pharmacology; pharmacotherapy; pharmaceutic; pharmaceutical; medication; mental health; digital health; informatics; buprenorphine; ketamine; cross sectional; survey; questionnaire

XSL•FO
RenderX

Introduction

During the COVID-19 public health emergency, the need for mental health services increased and telemedicine enabled critical access to care [1,2]. In the postpandemic era, telemedicine continues to enable mental health care access for many people, especially persons living in areas with mental health care provider shortages [3]. However, as lawmakers consider long-term policy decisions related to telemedicine, some have raised questions about the safety and appropriateness of prescribing via telemedicine, and whether there should be requirements for in-person evaluation, especially for controlled substances.

Before the US COVID-19 public health emergency declaration, most health care providers lacked telemedicine prescribing experience due to limited telemedicine adoption. Additionally, telemedicine-based prescribing of controlled substances was restricted during the prepandemic era, in compliance with the 2008 Ryan Haight Online Pharmacy Consumer Protection Act (Ryan Haight Act). Pursuant to the Ryan Haight Act, the US Drug Enforcement Agency (DEA) required that health care providers conduct an in-person evaluation of patients, before prescribing controlled substances via telemedicine. This measure was intended to prevent health care providers from prescribing potentially harmful medications with only minimal and inadequate online interaction with a patient. During the US public health emergency declaration, temporary policy flexibilities enabled controlled substance prescribing without the requirement of an in-person evaluation. In part, this enabled crucial access to buprenorphine for opioid use disorder treatment [4,5]. Now, consumers and stakeholders are calling for new policy that enables continued access to mental health care and substance use treatment, including medication-based treatment, via telemedicine [6]. The challenge is to develop evidence-based policy that supports safety while enabling critical health care access.

Current evidence is inconclusive but suggests that prescribing patterns can differ when care is provided via telemedicine versus in-person care. A 2023 study at a single institution showed that orthopedic patients are prescribed higher doses (milligram morphine equivalent) of opioids via video telemedicine than during in-person visits. McCabe et al [7] found that telemedicine is used less frequently than in-person visits for prescribing antibiotics. However, other studies have shown little or no difference in the prescribing patterns associated with telemedicine use [8,9]. These varied findings likely relate to differences in the clinical context for prescribing, including patient and health care provider characteristics, system-level factors, or the specific medications and their unique requirements for appropriate initiation and monitoring (and whether those requirements can be met via telemedicine). Evidence supporting safety and quality of telemedicine prescribing in mental health care is scarce, and also likely to be context dependent. However, a recent, large cohort study of health care claims data found no difference in safety or quality for telemedicine-based treatment versus in-person treatment of opioid use disorder, including medication-based opioid use disorder treatment [10].

Current health care provider perspectives and practices related to prescribing via telemedicine for mental or behavioral health care are not well-characterized. However, a better understanding of telemental health care provider's perspectives and experiences related to prescribing is needed to inform appropriate telemedicine program design, identify key research questions, pursue clinical practice guidelines, and develop curricula for professional educational programs. The perspectives and experience of practicing telemental health care providers should also inform policy decisions, such as the currently pending US DEA rules governing telemedicine prescribing of buprenorphine for opioid use disorder. The purpose of this study was to assess the perspectives of US telemental health care providers related to their comfort and perceived safety in prescribing medications, including controlled substances, via telemedicine.

Methods

Study Design

We conducted a web-based, cross-sectional survey of telemental health providers who prescribe via telemedicine, using nonprobability, availability sampling of a telehealth research panel from February 13 to April 28, 2024.

Survey

The web-based survey, administered using Qualtrics, was created by the research team and designed to elicit perspectives and practices of US mental or behavioral health care providers related to prescribing via telemedicine. The survey consisted of 8 sections: informed consent (1 question), verification of eligibility (5 questions), demographic and practice information (13 questions), comfort and safety (6 questions), health history (7 questions), physical assessment and diagnostic testing (7 questions), issuing a prescription (8 questions), and legal and regulatory environment (8 questions). A copy of the survey is provided in Multimedia Appendix 1.

This study is focused on a subset of the survey items, including Likert-scale items that measured agreement with statements indicating comfort with prescribing in varied scenarios, including in person (no telemedicine), via telemedicine, and via telemedicine with and without previous in-person care. We also analyzed Likert-scale items that measured agreement with statements of ability to safely prescribe medications, by US DEA scheduling. Recognizing that safe prescribing can depend upon individual patient characteristics and contextual factors aside from telemedicine use or pharmacotherapy, we designed the Likert scale to measure how often medications of a given schedule can be prescribed safely via telemedicine. Optional open-ended items prompted participants to elaborate upon their responses.

Survey Development

The interdisciplinary research team, which includes expertise in telemedicine, clinical pharmacology, and advanced practice nursing in primary care, developed the initial survey based on a literature review. A small group of telemedicine practitioners reviewed an early draft of the survey and provided input on its content, format, and relevance for clinical practice. We then refined the survey according to their input. Before initiating

XSL•FO
RenderX

data collection, we informally pretested and modified the survey within the research team, then formally pretested the survey with the target audience of 5 prescribing telemental health providers located in Utah and North Carolina who met the inclusion criteria. All 5 individuals who pretested the survey indicated that they strongly agree with the following statements: (1) the content and wording of the survey is appropriate; (2) the survey is easy to understand; (3) the survey is free of errors; (4) the time required to complete the survey is reasonable; and (5) the time required to complete the survey would be reasonable if the participants are compensated. Pretesters completed the survey in mean 11.9 (SD 4.24) minutes and recommended US $25 - $50 as compensation for completing the survey. The survey was configured to assign completion codes, so that the process of participant compensation, which requires collecting additional personal information, could be managed separately from data collection.

Sampling and Recruitment

We implemented nonprobability, availability sampling of a telehealth research panel. We calculated a target sample size of 382 for 95% CI. We currently lack precise data describing the number of US health care providers who (1) provide mental health care and (2) prescribe via telemedicine. Therefore, we based the target sample size Health Resources and Services Administration estimates of the total number of US mental health care providers in prescribing roles [11].

We recruited survey participants from the TelehealthEngage Research Panel, a panel of 7134 telemedicine users who have consented to be contacted about opportunities to participate in research. TelehealthEngage includes individuals from 49 states who are active users of Doxy.me, a commercial telemedicine platform. We invited all TelehealthEngage panelists who identified as a health professional in mental health, psychiatry, family practice, internal medicine, general practice, neurology, integrative medicine, or unknown fields or who identified as a physician, hospitalist, physician assistant (PA), nurse practitioner, or nurse. Recruitment was initiated on February 13, 2024, and closed on April 28, 2024. We divided our 2 recruitment waves over the course of 3 days each: the first invite was sent to 2343 unique panelists and the second to 3490. We sent a reminder email to any panelist who had not opened the email invitation (1491 in the first wave and 3200 in the second wave) after 3 weeks. The survey was closed out 2 weeks after the last invitation reminder was sent out. We required responses to all categorical or numeric survey items to complete the survey and be compensated for participation with a US $50 gift card.

Ethical Considerations

This study was reviewed and approved by the BRANY (Biomedical Research Alliance of New York) Institutional Review Board (IRB00010793). To protect the anonymity of participants, we configured Qualtrics to assign a random completion code to each completed survey. Upon survey completion, a code was displayed to participants along with instructions to visit a separate Qualtrics survey to submit the code and contact information for compensation. In this way, we separated participants' responses from their contact information. We stored the data in a secure environment with access limited to essential study personnel. Participants were compensated for participation with a US $50 gift card.

Analysis

We analyzed all completed surveys. We calculated descriptive statistics and frequencies, visualized response distributions, and aggregated free-text responses for qualitative analysis. Further, we assessed group differences in response distributions by licensure type (physician vs nonphysician) and specialty type (psychiatric vs nonpsychiatric) for items using Mann-Whitney *U* tests. We performed all statistical analysis using SPSS (version 29; IBM Corp) and performed a qualitative, thematic analysis of open-ended items using MAXQDA (MAXQDA - Distribution by VERBI GmbH). The results of the quantitative analyses informed the way we organized codes from the thematic analysis [12,13]. Further, 1 author coded the qualitative responses from surveys using the entire participant response as the unit of analysis. After 3 iterations of coding, a codebook was developed and honed with the help of all authors. We assigned each response one or more codes, and so the total number of codes exceeds the number of responses.

Results

Participation and Response

In total, 350 TelehealthEngage panelists accessed the survey. Of those, 336 (96%) completed all eligibility screening questions, and 121 (36%) of those screened were eligible. Further, 115/121 (95%) of screened and eligible panelists completed the survey. There was no missing data in completed surveys given the requirement that all items be completed, except optional free text fields. Only 4 surveys were initiated but not completed, so we did not analyze the data or individual items for nonresponse bias. See Table 1 illustrating study participation results.

Table . Participant count by stage.

Stage	Value, n
Accessed survey	350
Completed consent and screening items	336
Eligible	121
Initiated survey	119
Completed survey	115

XSL•FO
RenderX

Participant Characteristics

Demographic characteristics of the participants are summarized in Table 2 and Figure 1. The participants were largely White, non-Hispanic, and female, with a mean age of 51.2 (SD 12.4) years and primarily physicians and advanced practice nurses in smaller-sized practices. The most common specialties were psychiatry or psychiatry specialties. However, 28 (24.3%) indicated nonpsychiatry specialties. Geographically, the participants were distributed across 26 American states (Figure 1).

XSL•FO
RenderX

Table . Demographic and practice characteristics of participants (N=115).

Characteristic	Values, n (%)
Race	
American Indian or Alaska Native	1 (0.9)
Asian	9 (7.8)
Black or African American	8 (7)
Native Hawaiian or Pacific Islander	0 (0)
White	83 (72.2)
More than 1 race	5 (4.3)
Other race	2 (1.7)
Prefer not to answer	7 (6.1)
Unknown	0 (0)
Hispanic or Latino	
Hispanic or Latino	8 (7)
Not Hispanic or Latino	97 (84.3)
Prefer not to answer	6 (5.2)
Unknown	4 (3.5)
Gender	
Male	46 (40)
Female	66 (57.4)
Nonbinary or third gender	0 (0)
Prefer not to say	3 (2.6)
Licensure	
Doctor of medicine or doctor of osteopathic medicine	69 (60)
Advanced practice registered nurse or physician assistant	40 (34.8)
Doctor of philosophy clinical psychologist	4 (3.)
Pharmacist	0 (0)
Other	2 (1.74)
Percent of clients seen via telehealth	
None (0%)	0 (0)
Few (1% - 24%)	10 (8.7)
Some (25% - 49%)	23 (20)
Most (50% - 74%)	33 (28.7)
Almost all (75% - 99%)	29 (25.2)
All (100%)	20 (17.4)
Practice size	
Independent	53 (46.1)
Small group (2 - 5 health care providers)	29 (25.2)
Mid-size (6 - 15 health care providers)	20 (17.4)
Large group practice (16+ health care providers)	13 (11.3)
Setting[a]	
Academic	6 (5.2)
Community	47 (40.9)
Hospital	4 (3.5)

XSL•FO
RenderX

Characteristic	Values, n (%)
Clinic	58 (50.4)
School	0 (0)
Corrections	0 (0)
Federally qualified health center	4 (3.5)
Digital health care	17 (4.8)
Other	6 (5.2)
Specialty	
Psychiatry	63 (54.8)
Addiction medicine or psychiatry	6 (5.2)
Child and adolescent psychiatry	14 (12.2)
Geriatric psychiatry	4 (3.5)
Forensic psychiatry	0 (0)
Consultation liaison psychiatry	0 (0)
Family practice	11 (9.6)
Internal medicine	6 (5.2)
Pediatrics	3 (2.6)
Other	8 (7)
Experience in specialty (years)	
0 - 5	24 (20.9)
6 - 10	20 (17.4)
11 - 15	23 (20)
16+	48 (41.7)
Experience in telehealth (years)	
0 - 3.7 (since onset of COVID-19 pandemic)	70 (60.9)
3.7 - 10	39 (33.9)
11 - 15	3 (2.6)
16+	3 (2.6)

[a]Participants could select more than 1 descriptor.

XSL•FO
RenderX

Figure 1. Geographic distribution of participants (N=115). Created using Bing Maps [14] and published under the platform's limited license [15].

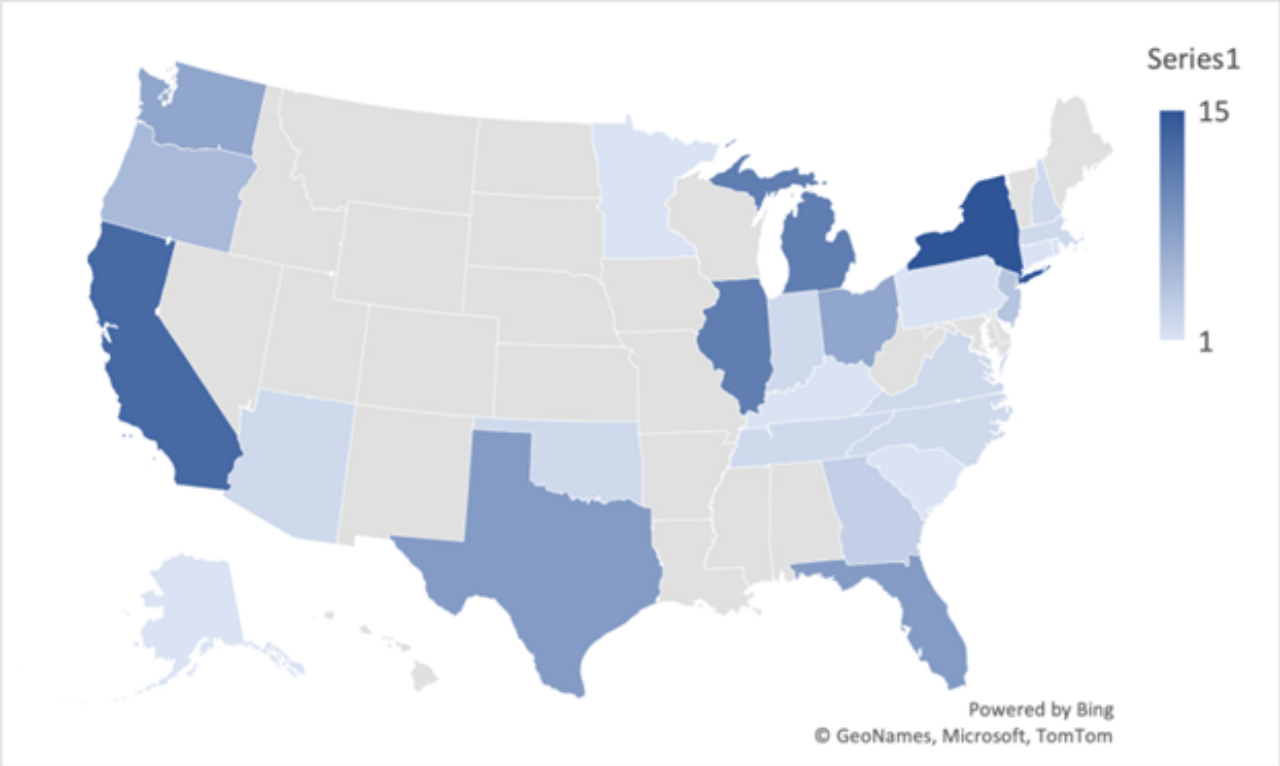

Main Findings

Overall, participant health care providers indicated high levels of comfort with prescribing via telemedicine, with 84% (102/115) of participants indicating they strongly agree with the statement indicating comfort in prescribing medications via telemedicine. Only 5 (4.3%) participants somewhat or strongly disagreed. However, participants indicated less comfort in prescribing if they have never seen a patient in person, or if the patient is located out-of-state. Figure 2 shows a visualization of the response distribution; detailed descriptive statistics are provided in Multimedia Appendix 2.

Figure 2. Participant agreement with statements of comfort in prescribing via telemedicine (N=115).

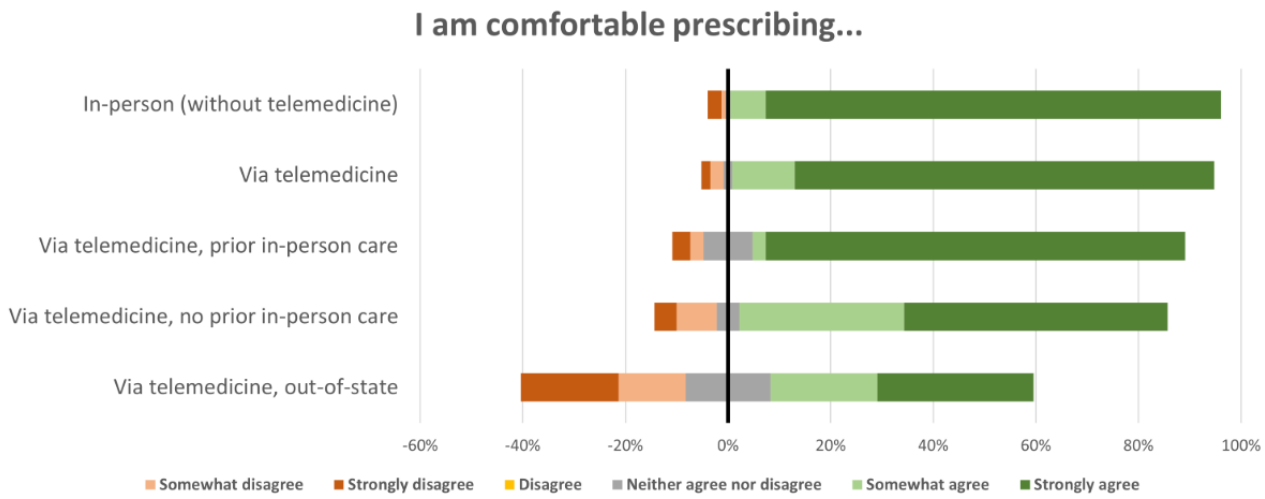

Further, 50% (58/115) of participants responded to the optional question "Please tell us more about the situations in which you feel comfortable or not comfortable prescribing." We found 3 general categories of comments related to comfort: comments related to feeling uncomfortable (1 instance), comfortable (17 instances), or more or less comfortable depending on the specific situation (conditionally comfortable, 36 instances).

The singular response that a participant health care provider unequivocally did not feel comfortable prescribing over telemedicine was in regard to controlled substances: "Will not prescribe classified drugs or anything for ADHD online." Health care providers who noted they unequivocally felt comfortable prescribing over telemedicine (17 instances) also included some explanation of prescribing situations regarding laboratories and assessments, laws and regulations, and types of medication:

XSL•FO
RenderX

"All but one of my current patients I've met in person, but my practice is now solely telemedicine. I do my best to assess the patient's condition, personality, and response. I think I am as comfortable prescribing online as in person." The majority (36 instances) of health care providers reported their comfort levels depended upon certain conditions. The approximate distribution of condition types is depicted in Figure 3. Health care providers were more (18 codes) or less comfortable (18 codes) due to reasons relating to the clinical scenario (including patient characteristics or behavior); laws and regulations; types of medications; and the care process (characteristics of the patient's care plan or care delivery including specific types of visits, assessments, and laboratories). Numerous health care providers noted how an initial in-person intake (9 codes) or working with established patients (9 codes) affected their experiences and comfort: "I felt comfortable after having an in person visits [sic] with the patient before telemedicine. I do a physical exam on all patients during the first visit snd sm [sic] not sure I feel comfortable prescribing without that initial physical encounter." Complete results of thematic analysis for this item are provided in Multimedia Appendix 3.

Figure 3. Condition types where health care providers feel conditionally more or less comfortable prescribing via telemedicine. *N refers to codes of more or less comfortable, depending upon conditions. n refers to code counts for the condition type. The total number of codes for condition type exceeds that of codes for comfort.

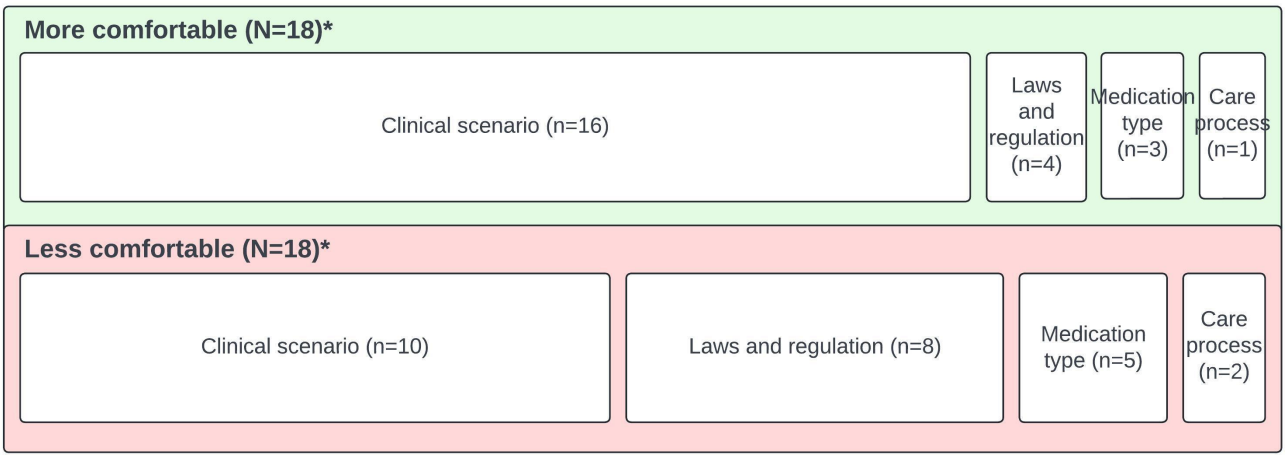

For telemedicine-based prescribing of controlled substances, the results are visualized in Figure 4 (detailed descriptive statistics in Multimedia Appendix 2). Most participants indicated that they can safely prescribe controlled substances via telemedicine, without having previously provided care to a patient in person. However, 14.8% - 19.1% of health care providers (by schedule) felt that they could rarely or never safely prescribe controlled substances. Among controlled substance schedules, participants indicated the least perceived safety with schedule IV medications, and the most safety with schedule II and III medications.

Figure 4. Perceived safety of telemedicine-based prescribing for medication types, by US Drug Enforcement Agency drug scheduling (N=115).

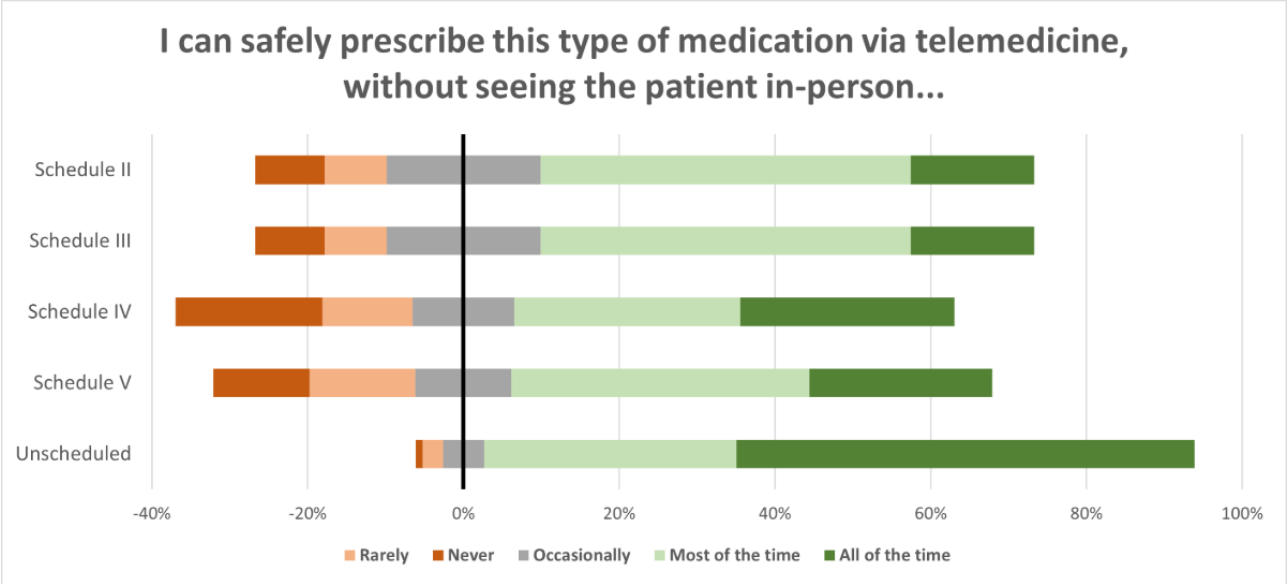

Further, 44% (51/115) of participants responded to the optional question "Tell us more about the types of medications you feel comfortable prescribing and/or monitoring via telemedicine, and the circumstances in which you feel an in-person assessment is appropriate…" We found 2 general categories of comments

XSL•FO
RenderX

related to comfort: comfortable (33 instances) and not comfortable (33 instances).

As depicted in (Figure 5), health care providers reported they felt comfortable (33 instances) prescribing certain types of medications (22 codes), prescribing with given certain care activities or processes (12 codes), in relation to laws and regulations (2 codes), in certain clinical scenarios (2 codes), and other (1 code). Further, 1 health care provider explained their requirements regarding medication types for prescribing over telemedicine: "No analgesics prescribed (no opiates, etc.) but I do prescribe stimulants for psychiatric reasons (ADHD, post-concussive pathology, as an adjunct for depression). My use of benzos is very conservative. Otherwise, I rarely prescribe anything but 'standard' psych meds–anti-depressants, mood stabilizers, non-benzo anxiolytics, at times anti-psychotics." Health care providers reported not feeling comfortable (31 instances) due to prescribing certain types of medications (19 codes), certain clinical scenarios (16 codes), and preferring in-person prescribing (1 code). Further, 1 health care provider noted how certain clinical scenarios and medication types affect their comfort: "I would only feel uncomfortable prescribing controlled substances virtually if I was [sic] concerned about other illicit substance use or elevated blood pressure/heart rate. I would want to see [the] patient in person for vital signs and UDS [urine drug screen]. The [state controlled substance database] allows us to monitor controlled prescriptions our patients have obtained." Complete results of thematic analyis for this item are provided in Multimedia Appendix 3.

Figure 5. Comments on comfort with specific types of medications and need for in-person appointments. *N refers to incidents and n refers to code counts. The total number of codes exceeds that of incidents.

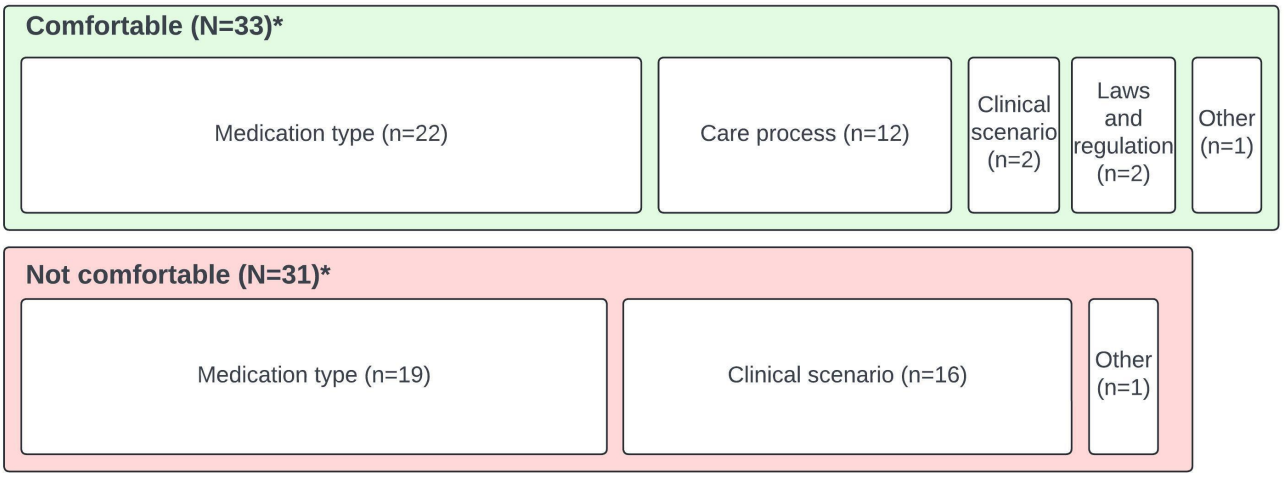

Findings by Licensure and Specialty

Table 3 summarizes the results of the Mann-Whitney *U* tests to determine whether response differed according to licensure type (physician vs nonphysician) or specialty (psychiatry vs nonpsychiatry). We found that comfort in prescribing differed by licensure type for in-person prescribing (no use of telemedicine, U=1857.5; P=.005; common language effect size [CL]=0.59), with physicians indicating stronger baseline agreement that they are comfortable with in-person prescribing (mean rank of 61.92 vs a mean rank of 52.12 for nonphysicians). We also found that physicians differed from nonphysicians in perception of their ability to safely prescribe schedule V medications (U=2018.5; P=.01; CL=0.64), with visualization showing that physicians indicated greater perceived safety (mean rank 64.25 vs 48.62). There were no other statistically significant differences in the distribution of responses to statements of comfort in prescribing by licensure type.

XSL•FO
RenderX

Table . Difference in distribution of responses by licensure type and specialty (N=115).

Item	Licensure (physician or nonphysician)			Specialty (psychiatric or nonpsychiatric)		
	U(SD)	*P* value	CL[a]	*U*	*P* value	CL[a]
I am comfortable prescribing medications…						
In person (no telemedicine)	1857.5	.005[a]	0.59	1216	.98	0.50
Telemedicine	1796.5	.08	0.57	1292.5	.47	0.53
Telemedicine, prior in person	1666	.50	0.52	1215.5	.98	0.50
Telemedicine, no prior in person	1435	.34	0.45	1571	.01	0.64
Telemedicine, out-of-state	1651.5	.71	0.52	1039	.23	0.43
I can safely prescribe this type of medication via telemedicine, without seeing the patient in person:						
Schedule II	1701.5	.50	0.54	1409.5	.19	0.58
Schedule III	1562	.88	0.49	880	.02	0.36
Schedule IV	1749	.34	0.55	1487.5	.07	0.61
Schedule V	2018.5	.01	0.64	1063	.30	0.44
Unscheduled medications	1774.5	.22	0.56	1419.5	.13	0.58

[a]CL: common language effect size.

By specialty, we found that the distribution of responses for prescribers from psychiatric specialties differed from that of nonpsychiatric specialty prescribers for comfort in prescribing via telemedicine when there has been no previous in-person care (U=1571; P=.01; CL=0.64), with prescribers from psychiatric specialties indicating more comfort (mean rank 62.06 vs 45.39). Prescribers from psychiatric specialties also differed for the perceived safety of prescribing schedule III medications (U=880; P=.02; CL=0.36), with visualization showing that prescribers from psychiatric specialties indicated lower perceived safety for this type of medication (mean rank 54.11 vs 70.07). Every participant with a specialty of addiction medicine or psychiatry (n=6) indicated that they can safely prescribe schedule III medications either “always” or “almost always.” However, the number of addiction medicine or psychiatry specialists was not sufficient to support testing of group differences.

Discussion

Principal Findings

In this survey of US telemental health care providers, we examined prescribing telemental health care provider perceptions of comfort and safety in prescribing medications via telemedicine. Largely White, non-Hispanic, female, and middle-aged, the participants’ demographic characteristics reflect the known demographics of the US mental health care workforce [16]. Licensure type and medical specialty varied. Participant practice settings were primarily individual or small group, and nonacademic. Most participants were physicians, advanced practice registered nurses, or PAs, and psychiatry was the most common specialty. Geographically, participants were fairly distributed within the United States. However, there was no representation of the mountain west.

Overall, we found that telemental health care providers were very comfortable prescribing via telemedicine. However, they were slightly less comfortable when prescribing to a patient not previously seen in person, and even less comfortable prescribing to an out-of-state patient. Psychiatry specialists expressed more comfort than nonpsychiatry specialists with providing care to patients not previously seen in person. The effect size was moderate for these differences. Qualitative analysis showed that individual prescriber comfort varied depending on the context, including the clinical scenario, types of medications being prescribed, laws and regulations, and care process. This finding is consistent with prior research indicating that a health care provider’s discomfort in prescribing over telemedicine appears to be tempered by access to additional information regarding the patient and situation such as access to laboratories, in-person follow-up, and coordination of care [17,18]

We further examined the participants’ perspectives on their ability to safely prescribe unscheduled medications as well as scheduled medications, which are controlled substances. Participants indicated that they are highly comfortable prescribing unscheduled medications. However, they varied in their comfort prescribing controlled substances. While participants in this study indicated they can usually prescribe schedule II-V medications safely, a minority indicated that they can rarely or never prescribe these medications safely. Variation in comfort by DEA scheduling may relate more to prescriber familiarity and experience with those medications, than with safety issues. For example, we found that physicians felt more comfortable than nonphysicians in prescribing schedule V

XSL•FO
RenderX

medications, with a moderate effect size. Examples of schedule V medications are pregabalin (Lyrica) and diphenoxylate or atropine (Lomotil). Given most psychiatric medications fall under schedules II-IV, this could reflect a broader medical practice of physicians in comparison to other licensure types. Psychiatry specialists were less comfortable than nonpsychiatry specialists in prescribing schedule III medications with no prior in-person care. Examples of schedule III medications include buprenorphine (Suboxone), testosterone, and ketamine. It is plausible that nonpsychiatry specialists, unlikely to prescribe buprenorphine or ketamine in their practices, may have based their response on other medications within schedule III, medications with which they are more familiar.

All addiction psychiatry or medicine specialists indicated a high level of comfort in prescribing schedule III medications. It is plausible that addiction specialists have cultivated practices that enhance safety and reduce harm, increasing their sense of safety with schedule III medications. A typical addiction specialist may enhance safety through a policy of checking controlled substance databases, requiring random drug toxicology screens and securing permission to speak with patients' families. Prescribers who routinely treat addiction also often have the advantage of intersecting with integrative partners such as substance use disorder treatment centers, court appointed case managers, etc, which may enhance their sense of comfort and safety in prescribing. Addiction specialists may be better prepared to learn that a patient has diverted medications, sought additional prescribers, or knowingly taken drugs or prescriptions outside of their prescribed regimen, and could be less likely to fear these circumstances in the context of treatment.

In open-ended responses, participants described substantial tailoring of their individual prescribing practices and decision-making according to the clinical context and their comfort level in prescribing. This finding is indicative of individual adaptation of practice to perceived risks and uncertainties. Currently, best practices for prescribing via telemedicine are general in nature, and health care providers rely on local practice settings' rules and processes for guidance [19]. Individual prescribers may differ in how they approach virtual and hybrid (virtual and in-person) care, but likely adapt their practice within their comfort level through measures that include coordination of care, in-person follow-up appointments, requirements for assessments and laboratories, and compliance with laws and regulations [20]. Ultimately, our findings show that telemental health care providers prescribe over telemedicine at their discretion, in the interest of their patients' health, and according to their individual sense of comfort. Our findings also initiate that prescribers implement care activities and processes that enhance safety.

Limitations

This study used a nonprobability sampling approach; we recruited a convenience sample from a research panel of telemental health care providers. The sample size and study design do not support the generalizability of findings to all US mental health care providers. However, these findings reflect the perspectives of a large, national sample of these health care providers. TelehealthEngage panelists are users of the Doxy.me telemedicine platform which is heavily used by solo and small clinic practices, and this is reflected in the participant demographics. Perspectives may differ for health care providers who are part of large enterprise settings or academic settings. Additionally, we may not have captured nuanced considerations of prescribing safety or the conditions under which prescribers feel comfortable prescribing via telemedicine. Medications within a single controlled substance schedule are used for different purposes and have different safety considerations. For example, the use of telemedicine to prescribe ketamine is currently highly controversial, as the mental health community continues to debate the appropriateness of using a medication that induces a dissociative state without the support of an in-person therapist or guide.

We also recognize that telemedicine implementation models vary and likely influence perspectives on prescribing comfort and safety. For example, a prescriber who provides telemental health care in the context of a hub-and-spoke model, where a registered nurse or medical assistant augments virtual care with an in-person assessment, may have perceptions of greater comfort and safety in prescribing than a prescriber practicing without such a model. In ongoing work, we are conducting qualitative research that entails interviewing prescribers to more fully elucidate this information. However, this survey study allowed us to first characterize the overarching perspectives of a larger number of prescribers.

Conclusion

We conducted a national, cross-sectional survey of US telemental health providers to assess their comfort and perceived safety in prescribing medications, including controlled substances, via telemedicine. The participants included physicians, advanced practice nurses, PAs, and clinical psychologists. These health care providers were highly comfortable prescribing both scheduled and unscheduled medications via telemedicine. Comfort and perceived safety with telemedicine prescribing varied somewhat by licensure type (physician vs nonphysician) and specialty (psychiatry vs nonpsychiatry). Perceived safety varied moderately for scheduled medications (controlled substances), especially for schedule IV and V medications. Participants indicated use of adaptive strategies to prescribe safely depending upon the clinical context. In ongoing efforts, we are analyzing additional survey results and conducting qualitative research related to telemedicine prescribing. A strong understanding of prescriber perspectives and experience with telemedicine prescribing is needed to support excellent clinical practice and effective policy making in the United States.

Acknowledgments

We gratefully acknowledge Beju Shah, PharmD, MBA, for his assistance in identifying individuals to pretest the survey.

XSL•FO
RenderX

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

MC and JI conceived and designed this study, obtained ethical approval, and conducted the statistical analysis. MC developed the survey instrument, conducted the statistical analysis, and coordinated pretesting. HS and JI implemented sampling, recruitment, and participant compensation. MC, JI, BEB, EL, ZR, and BMW contributed to the literature search and interpretation of findings. MC wrote the first draft of this paper. All authors reviewed and edited this paper and approved the final version of this paper.

Conflicts of Interest

BMW is a shareholder, and MC, JI, HS, EL and BEB are employees of Doxy.me Inc, a commercial telemedicine company. The authors declare no other competing interests.

Multimedia Appendix 1
Survey of telemental health providers and informed consent form.
[PDF File, 563 KB - formative_v9i1e63251_app1.pdf]

Multimedia Appendix 2
Descriptive statistics.
[DOCX File, 18 KB - formative_v9i1e63251_app2.docx]

Multimedia Appendix 3
Results of thematic analysis of open-ended questions.
[DOCX File, 19 KB - formative_v9i1e63251_app3.docx]

References

1. Penninx B, Benros ME, Klein RS, Vinkers CH. How COVID-19 shaped mental health: from infection to pandemic effects. Nat Med 2022 Oct;28(10):2027-2037. [doi: 10.1038/s41591-022-02028-2] [Medline: 36192553]
2. Baumgart JG, Kane H, El-Hage W, et al. The early impacts of the COVID-19 pandemic on mental health facilities and psychiatric professionals. Int J Environ Res Public Health 2021 Jul 29;18(15):8034. [doi: 10.3390/ijerph18158034] [Medline: 34360334]
3. Larson EH, Patterson DG, Garberson LA, Andrilla CHA. Supply and distribution of the behavioral health workforce in rural America. WWAMI Rural Health Research Center, University of Washington. 2016. URL: https://depts.washington.edu/fammed/rhrc/wp-content/uploads/sites/4/2016/09/RHRC_DB160_Larson.pdf [accessed 2025-01-24]
4. Huskamp HA, Riedel L, Uscher-Pines L, et al. Initiating opioid use disorder medication via telemedicine during COVID-19: implications for proposed reforms to the Ryan Haight Act. J Gen Intern Med 2022 Jan;37(1):162-167. [doi: 10.1007/s11606-021-07174-w] [Medline: 34713386]
5. Lynch MJ, Vargas D, Winger ME, et al. A telemedicine bridge clinic improves access and reduces cost for opioid use disorder care. Drug Alcohol Depend Rep 2024 Jun;11:100227. [doi: 10.1016/j.dadr.2024.100227] [Medline: 38550513]
6. Zionts A. The DEA relaxed online prescribing rules during covid. now it wants to rein them in. KFF Health News. URL: https://kffhealthnews.org/news/article/dea-online-prescribing-rules-telehealth-changes/ [accessed 2025-01-24]
7. McCabe BK, Linder JA, Doctor JN, et al. The protocol of improving safe antibiotic prescribing in telehealth: a randomized trial. Contemp Clin Trials 2022 Aug;119:106834. [doi: 10.1016/j.cct.2022.106834] [Medline: 35724841]
8. Fathy R, Briker S, Rodriguez O, Barbieri JS. Comparing antibiotic prescription rates between in-person and telemedicine visits. J Am Acad Dermatol 2022 Aug;87(2):438-440. [doi: 10.1016/j.jaad.2021.08.064] [Medline: 34499992]
9. Patel R, Irving J, Brinn A, et al. Impact of the COVID-19 pandemic on remote mental healthcare and prescribing in psychiatry: an electronic health record study. BMJ Open 2021 Mar;11(3):e046365. [doi: 10.1136/bmjopen-2020-046365]
10. Hailu R, Mehrotra A, Huskamp HA, Busch AB, Barnett ML. Telemedicine use and quality of opioid use disorder treatment in the US during the COVID-19 pandemic. JAMA Netw Open 2023 Jan 3;6(1):e2252381. [doi: 10.1001/jamanetworkopen.2022.52381] [Medline: 36692880]
11. Mapping supply of the US psychiatric workforce. Health Workforce Technical Assistance Center. 2018 Oct. URL: https://www.healthworkforceta.org/research-alerts/mapping-supply-of-the-us-psychiatric-workforce/ [accessed 2025-02-6]
12. Clarke V, Braun V. Thematic analysis. J Posit Psychol 2017 May 4;12(3):297-298. [doi: 10.1080/17439760.2016.1262613]
13. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
14. Bing Maps. 2025. URL: https://www.bingmapsportal.com/ [accessed 2025-05-05]

XSL•FO
RenderX

15. Microsoft bing maps platform apis terms of use. Microsoft. 2025. URL: https://www.microsoft.com/en-us/maps/bing-maps/product [accessed 2025-05-05]
16. Data tool: demographics of the U.S psychology workforce. American Psychological Association. 2024. URL: https://www.apa.org/workforce/data-tools/demographics [accessed 2025-01-24]
17. Uscher-Pines L, Sousa J, Raja P, Mehrotra A, Barnett M, Huskamp HA. Treatment of opioid use disorder during COVID-19: experiences of clinicians transitioning to telemedicine. J Subst Abuse Treat 2020 Nov;118:108124. [doi: 10.1016/j.jsat.2020.108124] [Medline: 32893047]
18. Uscher-Pines L, Parks AM, Sousa J, et al. Appropriateness of telemedicine versus in-person care: a qualitative exploration of psychiatrists' decision making. Psychiatr Serv 2022 Aug 1;73(8):849-855. [doi: 10.1176/appi.ps.202100519] [Medline: 35080419]
19. Smith K, Ostinelli E, Macdonald O, Cipriani A. COVID-19 and telepsychiatry: development of evidence-based guidance for clinicians. JMIR Ment Health 2020 Aug 28;7(8):e21108. [doi: 10.2196/21108] [Medline: 32658857]
20. Gough F, Budhrani S, Cohn E, et al. ATA practice guidelines for live, on-demand primary and urgent care. Telemed J E Health 2015 Mar;21(3):233-241. [doi: 10.1089/tmj.2015.0008] [Medline: 25658882]

ABBREVIATIONS

BRANY: Biomedical Research Alliance of New York
CL: common language effect size
DEA: Drug Enforcement Agency
PA: physician assistant

Edited by A Mavragani; submitted 14.06.24; peer-reviewed by MH Pourasad, V Chavarria; revised version received 20.12.24; accepted 21.12.24; published 11.03.25.

Please cite as:
Cummins MR, Ivanova J, Soni H, Robbins Z, Bunnell BE, López E, Welch BM
Telemedicine Prescribing by US Mental Health Care Providers: National Cross-Sectional Survey
JMIR Form Res 2025;9:e63251
URL: https://formative.jmir.org/2025/1/e63251
doi:10.2196/63251

© Mollie R Cummins, Julia Ivanova, Hiral Soni, Zoe Robbins, Brian E Bunnell, Esteban López, Brandon M Welch. Originally published in JMIR Formative Research (https://formative.jmir.org), 11.3.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Supplementing Consent for a Prospective Longitudinal Cohort Study of Infants With Antenatal Opioid Exposure: Development and Assessment of a Digital Tool

Jamie E Newman[1], MPH, PhD; Leslie Clarke[2], MS, MBA, RN; Pranav Athimuthu[1], BA; Megan Dhawan[3], MSN, CRNP; Sharon Owen[4], RN; Traci Beiersdorfer[5], RN; Lindsay M Parlberg[1], MPH; Ananta Bangdiwala[1], MS; Taya McMillan[1], MPH; Sara B DeMauro[3], MSCE, MD; Scott Lorch[3], MSCE, MD; Myriam Peralta-Carcelen[4], MPH, MD; Deanne Wilson-Costello[2], MD; Namasivayam Ambalavanan[4], MD; Stephanie L Merhar[5], MS, MD; Brenda Poindexter[6], MD; Catherine Limperopoulos[7], PhD; Jonathan M Davis[8], MD; Michele Walsh[9], MD; Carla M Bann[1], PhD

[1]RTI International, 3040 Cornwallis Road, Research Triangle Park, NC, United States
[2]Case Western Reserve University, Cleveland, OH, United States
[3]Children's Hospital of Philadelphia, Philadelphia, PA, United States
[4]University of Alabama at Birmingham, Birmingham, AL, United States
[5]University of Cincinnati College of Medicine, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, United States
[6]Emory University, Atlanta, GA, United States
[7]Children's National Hospital, Washington, DC, United States
[8]Tufts Medical Center, Boston, MA, United States
[9]National Institute of Child Health and Human Development, Rockville, MD, United States

Corresponding Author:
Jamie E Newman, MPH, PhD
RTI International, 3040 Cornwallis Road, Research Triangle Park, NC, United States

Abstract

Background: The Outcomes of Babies With Opioid Exposure (OBOE) study is an observational cohort study examining the impact of antenatal opioid exposure on outcomes from birth to 2 years of age. COVID-19 social distancing measures presented challenges to research coordinators discussing the study at length with potential participants during the birth hospitalization, which impacted recruitment, particularly among caregivers of unexposed (control) infants. In response, the OBOE study developed a digital tool (consenter video) to supplement the informed consent process, make it more engaging, and foster greater identification with the research procedures among potential participants.

Objective: We aim to examine knowledge of the study, experiences with the consent process, and perceptions of the consenter video among potential participants of the OBOE study.

Methods: Analyses included 129 caregivers who were given the option to view the consenter video as a supplement to the consent process. Participants selected from 3 racially and ethnically diverse avatars to guide them through the 11-minute video with recorded voice-overs. After viewing the consenter video, participants completed a short survey to assess their knowledge of the study, experiences with the consent process, and perceptions of the tool, regardless of their decision to enroll in the main study. Chi-square tests were used to assess differences between caregivers of opioid-exposed and unexposed infants in survey responses and whether caregivers who selected avatars consistent with their racial or ethnic background were more likely to enroll in the study than those who selected avatars that were not consistent with their background.

Results: Participants demonstrated good understanding of the information presented, with 95% (n=123) correctly identifying the study purpose and 88% (n=112) correctly indicating that their infant would not be exposed to radiation during the magnetic resonance imaging. Nearly all indicated they were provided “just the right amount of information” (n=123, 98%) and that they understood the consent information well enough to decide whether to enroll (n=125, 97%). Survey responses were similar between caregivers of opioid-exposed infants and unexposed infants on all items except the decision to enroll. Those in the opioid-exposed group were more likely to enroll in the main study compared to the unexposed group (n=49, 89% vs n=38, 51%; $P<.001$). Of 81 caregivers with known race or ethnicity, 35 (43%) chose avatars to guide them through the video that matched their background. Caregivers selecting avatars consistent with their racial or ethnic background were more likely to enroll in the main study (n=29, 83% vs n=43, 57%; $P=.01$).

XSL•FO
RenderX

Conclusions: This interactive digital tool was helpful in informing prospective participants about the study. The consenter tool enhanced the informed consent process, reinforced why caregivers of unexposed infants were being approached, and was particularly helpful as a resource for families to understand magnetic resonance imaging procedures.

Trial Registration: ClinicalTrials.gov NCT04149509; https://clinicaltrials.gov/study/NCT04149509

(JMIR Form Res 2025;9:e59954) doi:10.2196/59954

KEYWORDS

informed consent digital tool; avatars; video-assisted consent; MRI; antenatal opioid exposure; infant; antenatal; opioid exposure; caregiver; survey; magnetic resonance imaging; Outcomes of Babies With Opioid Exposure

Introduction

The importance of the informed consent process in today's clinical trials stems from ethical standards detailed in landmark documents such as the Nuremberg Code, the Declaration of Helsinki, and the Belmont Report. Recent research suggests that there is room for improvement in the consent processes to better inform prospective participants of study aims and potential risks. Numerous studies have suggested that minority populations, those with low levels of education, and those with intellectual disabilities are among the groups in which improvements could be made to heighten participants' understanding of the informed consent process (see Tam et al [1] for a review). Some studies have observed that fewer than half of the participants were able to correctly answer questions about the study's purpose after reviewing an informed consent form [2,3]. In the review by Tam et al [1], only 55% could name a risk associated with a study, with even fewer being able to define key methodological concepts such as the use of placebo or randomization.

Informed consent materials should be easily comprehended by the average research participant. Grade-level reading standards are the most common metric for measuring the readability of consent documents. It is generally agreed that material should target 6th- to 8th-grade level standards to be considered widely interpretable. Yet, a 2003 study exploring informed consent materials furnished by research institutions across the United States found that only 8% of their sample met their own institution's standards for comprehension, with the average readability score of a consent document exceeding stated standards by almost 3 grade levels [4]. If prospective participants cannot understand the structure and content of a consent form, they may decline to participate. If a study using English language materials involves nonnative English speakers, minority populations, or foreign immigrants, the reading level must be further simplified [5].

In response to criticisms of written consent forms, recent studies have harnessed technology to create consent procedures that simplify the information being conveyed while better engaging prospective participants. While some approaches have used electronic forms, others have developed videos, audio files, and multimedia tutorials to supplement the consent process.

Electronic consent tools found their earliest uses in medical procedures, and most evaluations of their effectiveness continue to reside in that domain [6,7]. Studies evaluating electronic consent have focused on comprehension [8], engagement [8-10], satisfaction and procedure-related anxiety reduction [11], recall of risks and benefits [4], and the potential to streamline the consent process by reducing time spent by staff explaining study procedures [7]. There are mixed findings on the effectiveness of video-assisted consent compared with written consent forms [9]. This may be due, in part, to technology advances over time. Videos created 20 years ago would have been limited in the graphics capacity of procedures they could display, whereas modern tools have the capacity to include 3D animation and guided walk-throughs. Multiple studies that have successfully demonstrated the effectiveness of video-assisted consent involve minority populations, study participants from rural areas, populations with English as a second language, older adults, and those with intellectual disabilities [12-15].

Certain capacities show particular promise at recruiting people traditionally underrepresented in medical research and vulnerable populations, warranting consideration in future studies. For example, Lawrence et al [16] describe an electronic consent framework using REDCap (Research Electronic Data Capture; Vanderbilt University) and customizable avatars. Videos provided participants with a choice of virtual assistants voiced by community members that guided them through the consent process. Participants had the ability to select their avatar from a variety of backgrounds, which were developed in consultation with targeted communities. Participants "emphasized the importance of representing minority populations [and] depicting supportive and helpful interactions with medical staff," as well as their strong preference for seeing procedures depicted using avatars [16]. Participants reported being inspired by the avatars and being able to see themselves in the avatars, which have translated into stronger intervention outcomes for the treatment group [17].

Digitally assisted consent also offers potential advantages in other research domains for informing participants of study procedures and potential risks such as neuroimaging methods (eg, magnetic resonance imaging [MRI]). Although MRI is largely a low-risk, noninvasive procedure that substitutes magnetic fields for high-energy radiation, misconceptions about its use abound, including false beliefs that MRI exposes participants to radiation [18]. To our knowledge, no studies of electronic consent tools have assessed the potential of video-assisted consent to help communicate risks associated with MRI procedures. The few studies involving neuroimaging have focused on computed tomography and other radiation methods that do not carry as many misconceptions [13]. Thus, our research fills a gap in this literature by assessing whether using a video supplement to visually depict MRI scanning

XSL•FO
RenderX

procedures and associated risks can adequately inform potential participants and dispel common MRI misperceptions in the Outcomes of Babies With Opioid Exposure (OBOE) study. It also builds on past literature on consent procedures, by including ethnically diverse avatars to describe MRI procedures.

Social distancing measures to mitigate the spread of COVID-19 presented new challenges to research coordinators talking at length with caregivers during the birth hospitalization about the OBOE study and building rapport through these face-to-face interactions. To mitigate these challenges, we developed the OBOE study consenter video to supplement the informed consent process and to make it more engaging for potential study participants, especially those from underrepresented populations. We developed this interactive digital tool to inform prospective participants about the study, reinforce why caregivers of unexposed (control) infants were being approached, provide a resource for families to understand the MRI procedures, and describe the longitudinal nature of the study while detailing visits across the 2-year study period. To enhance participant diversity, we developed 3 racially and ethnically diverse avatars, which participants could choose to guide them through the video. If potential participants were able to see themselves in the avatars, this may foster greater identification with the research procedures. This paper provides an overview of the digital tool and examines knowledge of the study, experiences with the consent process, and perceptions of the video among potential participants in the OBOE study.

Methods

OBOE Study Overview

The Advancing Clinical Trials in Neonatal Opioid Withdrawal OBOE study is a multisite prospective longitudinal cohort study of infants with antenatal opioid exposure and unexposed infants from birth to 2 years of age (ClinicalTrials.gov NCT04149509). The OBOE study protocol is described elsewhere in detail [19]. Briefly, the observational study aims to determine the impact of antenatal opioid exposure on brain development and neurodevelopmental outcomes over the first 2 years of life and to explore whether family, home, and community factors modify developmental trajectories. The study consists of 3 MRI visits at 0 - 1 month, 6 months, and 24 months; 1 home visit at 12 months; 1 telephone interview at 18 months; and developmental testing at 12 and 24 months. The OBOE study addresses a growing health concern—opioid use and misuse—among a particularly vulnerable population: pregnant individuals and their infants. Overdose has become one of the leading causes of maternal mortality [20], mirroring the rise in misuse among pregnant individuals and the greater population throughout the opioid epidemic.

Digital Tool Development: The Consenter Video

The first infant was enrolled into the OBOE study in August 2020, and the OBOE study began developing the consenter video a year later to assist with the slower-than-anticipated enrollment and to enhance participant diversity. The consenter tool is a patient-centered, interactive digital tool for informing prospective participants about clinical research studies and enhancing informed consent processes [21,22]. It transforms the informed consent process from the passive receipt of complex medical and scientific information to an active, engaging visual experience that can improve individuals' understanding of clinical research processes and result in a more informed, dedicated research participant [23]. Five main steps were involved in developing the OBOE study consenter tool. First, a detailed storyboard was developed, which included verbatim audio text and a written description of proposed visuals. The audio text was primarily pulled from the single institutional review board (IRB)–approved consent form. Next, the team created illustrations and visuals to describe the study and assessments. Third, a group of digital designers developed customized avatars. Fourth, the team selected 3 actors to record voice-overs that corresponded to the race and ethnicity of each avatar. Finally, the video production team combined the multimedia elements into a single video file. The development process was guided by a communication scientist to ensure that evidence-based communication practices were incorporated. Throughout each step in the process, staff at the clinical sites, including research coordinators, nurses, physicians, and MRI staff, reviewed the products, and their feedback was incorporated. All study team members attended a virtual consenter orientation held before implementation. The OBOE study launched the IRB-approved consenter tool in June 2022, at which time approximately half of the target sample had been enrolled.

Ethical Considerations

Human Subject Ethics Review Approvals

Through a single IRB at Cincinnati Children's Hospital Medical Center, all 4 clinical sites (Cincinnati Children's Hospital Medical Center, University of Alabama at Birmingham, Children's Hospital of Pennsylvania, and Case Western Reserve University), the Neuroimaging Core at Children's National Hospital, and the Data Coordinating Center at the RTI (Research Triangle Institute) received approval for human subjects research activities for the OBOE study, and informed consent was obtained for all participants. The STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) reporting guidelines for observational studies were followed [24].

Informed Consent

Caregivers of opioid-exposed infants and unexposed infants were asked to view the consenter video in addition to talking to coordinators about the study as well as reading and signing the informed consent form. Caregivers of infants born at or after 37 weeks gestation were recruited at the 4 clinical sites primarily during hospital stays following birth and, to a lesser extent, prenatal clinics. Caregivers of opioid-exposed infants were approached if their infant was exposed to opioids in the second or third trimester of pregnancy. Caregivers of unexposed (control) infants were approached if there was no antenatal drug exposure as determined by maternal history or maternal urine toxicology screen at delivery. OBOE study exclusion criteria for both groups were confirmed at delivery and are detailed elsewhere [19].

XSL•FO
RenderX

Privacy and Confidentiality

Caregivers approached for the OBOE study were assigned a participant ID number, and they chose from one of three racially and ethnically diverse avatars to guide them through the 11-minute video with recorded voiceovers and illustrations of study procedures (Figure 1). After viewing the video, caregivers were asked to complete a brief survey using their participant ID, regardless of their decision to participate in the main study. The survey assessed their knowledge of the study, experiences with the consent process, and perception of the video.

Figure 1. Consenter avatars that caregivers choose to guide them through the video as a supplement to the informed consent form.

Compensation Details

Participants were given a US $10 gift card for completing the survey.

Digital Tool Overview

To ensure ease of access to the consenter tool, participants were provided 3 options for viewing the video. These included: (1) an iPad with the video preloaded onto an app, which the participants used alongside a coordinator; (2) a coordinator lanyard card with a QR code that participants could scan to watch the consenter video on their personal mobile device; and (3) business cards with a QR code and web link directing to the video so that participants could access the video asynchronously away from the study site.

The interactive video described the study purpose, detailed the risks and benefits of joining the study, explained that participation is voluntary, and depicted MRI procedures via a verbal description and illustrations of a staff member preparing a swaddled, sleeping baby with protective earmuffs to enter the MRI scanner (Figure 2). The avatar explained that MRI scans are safe and do not expose the infant to any radiation, dispelling a common misperception of MRI.

Figure 2. Consenter video screenshots illustrating MRI procedures at the 0 - 1 month OBOE study visit. MRI: magnetic resonance imaging; OBOE: Outcomes of Babies With Opioid Exposure.

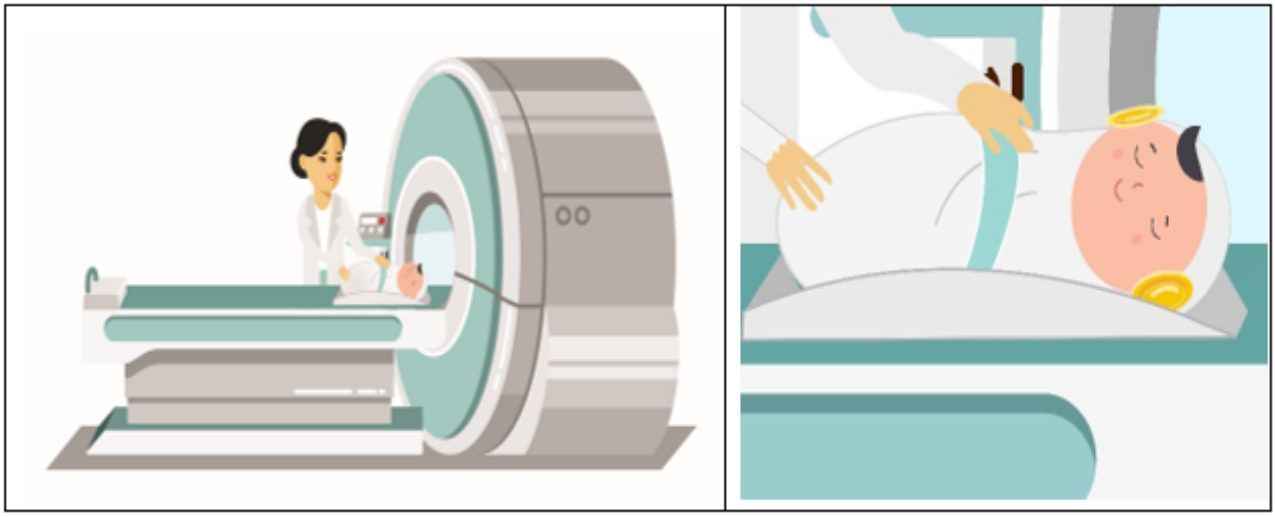

The consenter tool detailed the longitudinal nature of the study and described assessments at each visit time point. Figure 3 illustrates developmental testing at the 24-month visit.

XSL•FO
RenderX

Figure 3. Consenter video screenshot illustrating developmental testing at the 24-month OBOE study visit. OBOE: Outcomes of Babies With Opioid Exposure.

Statistical Analysis

All analyses were performed using SAS (version 9.4; SAS Institute). Chi-square tests were used to examine (1) whether there were differences between opioid-exposed and unexposed groups in survey responses and (2) whether participants who selected avatars consistent with their racial or ethnic background were more likely to report that they decided to enroll in the study than those who selected avatars that were not consistent with their background.

Results

Overview

Analyses included 129 caregivers (55 caregivers of opioid-exposed infants and 74 caregivers of unexposed infants) who were given the option to view the consenter video as a supplement to the informed consent process and take the survey.

Survey Responses

Participants demonstrated a good understanding of the information presented in the video (Table 1). Overall, 95% (n=123) correctly identified the purpose of the OBOE study. The majority also correctly indicated that the research team would collect the baby's umbilical cord (n=120, 93%), the infant would not be exposed to radiation during the MRI scan (n=112, 88%), and the study would not include a blood test (n=105, 83%). Nearly all indicated they were provided "just the right amount of information" (n=123, 98%) and understood the consent information well enough to decide whether to enroll (n=125, 97%).

XSL•FO
RenderX

Table . Survey responses of caregivers of opioid-exposed infants and unexposed infants after watching consenter video (N=129).

Item	All (n=129), n (%)	Opioid exposed (n=55), n (%)	Unexposed (n=74), n (%)
Knowledge of study			
1. Correctly identified the purpose of the OBOE[a] study	123 (95)	53 (96)	70 (95)
2. Correctly indicated, "If I enroll in the study, the research team will collect a piece of my baby's umbilical cord at delivery that would otherwise be thrown away."	120 (93)	50 (91)	70 (95)
3. Correctly indicated that infant would not be exposed to radiation during the MRI[b] scan	112 (88)	48 (87)	64 (88)
4. Which of the following is not one of the study activities:			
a. Three brain MRIs over the first 2 years of my baby's life	4 (3)	3 (5)	1 (1)
b. A 12-month home visit	9 (7)	6 (11)	3 (4)
c. A blood test (correct response)	105 (83)	44 (80)	61 (86)
d. Parent or caregiver surveys	8 (6)	2 (4)	6 (8)
Experience with consent process			
5. During the consent process, do you feel you were provided:			
a. Too much information	0 (0)	0 (0)	0 (0)
b. Just the right amount of information	123 (98)	53 (98)	70 (97)
c. Too little information	3 (2)	1 (2)	2 (3)
6. Felt you understood the consent information well enough to decide whether you and your baby should be in the study	125 (97)	55 (100)	70 (95)
Perceptions of consenter video			
7. Did you watch the video describing the study?			
a. Yes	124 (96)	51 (93)	73 (99)
b. No	5 (4)	4 (7)	1 (1)
7a. If yes, what did you like most about the video[c]?			
a. The explanation of why the study is important	62 (51)	27 (53)	35 (49)
b. The explanation of the study procedures	28 (23)	12 (24)	16 (23)
c. The explanation of the MRI	19 (16)	6 (12)	13 (18)
d. The explanation of how my personal information would be protected	7 (6)	4 (8)	3 (4)
e. I did not like watching the video	2 (2)	1 (2)	1 (1)
f. Other	4 (3)	1 (2)	3 (4)
7b. If yes, what did you not like about the video?			
a. It was too long	5 (4)	1 (2)	4 (6)
b. It was hard to understand	0 (0)	0 (0)	0 (0)

XSL•FO
RenderX

Item	All (n=129), n (%)	Opioid exposed (n=55), n (%)	Unexposed (n=74), n (%)
c. It did not cover everything I needed to know about the study	0 (0)	0 (0)	0 (0)
d. I had a hard time getting the video to play	10 (8)	3 (6)	7 (10)
e. I liked watching the video	93 (79)	41 (82)	52 (76)
f. Other	10 (8)	5 (10)	5 (7)
Study enrollment intentions			
8. Did you decide to enroll in the OBOE study[d]?			
a. Yes	87 (67)	49 (89)	38 (51)
b. No	25 (19)	1 (2)	24 (32)
c. Not yet decided	17 (13)	5 (9)	12 (16)
8a. If you decided not to enroll or have not yet decided, what was the primary reason?			
a. Too much time or effort to complete the study visits or surveys	3 (8)	1 (17)	2 (6)
b. I do not trust that my information will be kept safe	2 (5)	0 (0)	2 (6)
c. I do not want to disclose any information, even if my name is not attached to it	0 (0)	0 (0)	0 (0)
d. I am not interested in participating in any research	5 (13)	0 (0)	5 (15)
e. I live too far away from the hospital or follow-up clinic	5 (13)	0 (0)	5 (15)
f. Financial burden to come back for the visits	1 (3)	0 (0)	1 (3)
g. I do not want my baby to have a brain MRI	6 (15)	0 (0)	6 (18)
h. Other	18 (45)	5 (83)	13 (38)

[a]OBOE: Outcomes of Babies With Opioid Exposure.

[b]MRI: magnetic resonance imaging.

[c]Percentages do not add to 100% because 3 respondents selected multiple responses.

[d]$P<.001$.

Survey responses did not differ significantly between caregivers of opioid-exposed infants and those of unexposed infants on all items except decision to enroll. Those in the opioid-exposed group were significantly more likely to report that they decided to enroll in the study compared to the unexposed group (n=49, 89% vs n=38, 51%; $P<.001$).

Avatar Selection

Of 129 respondents who completed the consenter video survey, 1 (1%) was Asian, 1 (1%) was Black and Hispanic, 29 (22%) were Black and non-Hispanic, 69 (53%) White and non-Hispanic, and 29 (22%) were missing race. Among 81 respondents of known race who watched the video, 35 (43%) chose avatars to guide them through the video that matched their background. Participants who selected avatars consistent with their racial or ethnic background were more likely to report that they decided to enroll in the study (n= 29, 83% vs n=43, 57%; $P=.01$).

Other Findings

An unanticipated benefit of the consenter video was its usefulness in informing new caregivers about the study when there were custody changes. As the longitudinal study progressed, coordinators indicated that the consenter video was helpful in providing information to new caregivers (eg, foster parents) about the study. We only asked caregivers to take the consenter survey at their infant's enrollment into the study, so we do not have data to assess what new caregivers thought about the video in cases of custody changes.

Discussion

Principal Findings

Social distancing measures to mitigate the spread of COVID-19 placed limitations on face-to-face interactions between study staff and participants. The OBOE study addressed this challenge

XSL•FO
RenderX

by developing an avatar-based consenter tool to supplement the consent process by creating an interactive video to explain study procedures while engaging a diverse cohort of study participants, particularly those from underrepresented populations who are most likely to have difficulty with traditional consent materials. Of the 129 caregivers who viewed the consenter video and took the survey, almost all (n=123, 95%) were able to correctly identify the purpose of the study, and nearly all felt that the video provided "just the right amount" of information and understood the consent information well enough to decide whether to enroll.

Participants' most frequently cited strengths of the video were its ability to convey the importance of the study and its ability to explain the MRI and study procedures. This is an important finding given the lack of existing research assessing video-assisted consent for neuroimaging studies. In addition, when asked what they did not like about the video, 79% (n=93) of participants indicated they in fact liked watching the video, suggesting they were engaged in viewing the video-assisted consent. Compared with written consent forms, animation and visuals and such as those used in the consenter tool can better engage participants, potentially resulting in more informed research participants [22-24].

Participants who indicated that they did not like the consenter video cited technical issues (n=10, 8%) and length (n=5, 4%) as the primary reasons. Duration is a key element that affects engagement if the video length does not correlate with the complexity of the information conveyed. Researchers must try to avoid including excessive details, because doing so may result in informational overload thus minimizing the benefits of using a video [25,26]. Additionally, study staff must be adequately equipped to handle the technological needs associated with using digital tools. The OBOE study addressed the need for flexible viewing options by providing 3 ways for participants to view the video.

Given the potential for race-matched avatars to facilitate connections between participants and the study, we anticipated that participants would tend to choose the avatar that best matched their group identification. However, only 43% of participants whose race was known and who completed the consenter survey chose an avatar that aligned with their racial background. It is possible that participants chose avatars for reasons aside from self-representation, such as perceptions of avatars' attractiveness or how authentic they felt an avatar looked. Lee [27] found that Black and female participants of a virtual video game were more likely to choose Black and female avatars (respectively) when assured that they were entering a high-diversity virtual world with other minority participants, but without this assurance, many Black participants chose White avatars. Lee [27] interpreted this finding using a social identity theory framework [28], explaining that virtual avatars can be used to facilitate self-expression and feelings of belonging, or they can create "escapes" for minority participants by allowing them to camouflage themselves as a member of the majority group. When virtual worlds are framed as high in diversity, participants tend toward the former, projecting their identity onto their avatar and customizing it accordingly. However, without this framing, participants may default to using ingrained cultural lenses and may "pass" under White avatars. Zimmermann et al [29], using the theories of transformed social interaction [30] and impression management [31], explain this "passing" behavior as people's attempt to present an idealized representation of themselves in a given context.

Researchers must make active efforts to assure participants, through study framing and multimedia tool design, that the world they will enter as study participants is one that accepts them, includes them, and actively recognizes their autonomy to choose and consent, regardless of their background.

Limitations

Specific activities that supported the launch of the consenter tool included receiving robust feedback from a diverse set of clinical site staff on the video illustrations, avatar images, and voice-overs to ensure the video was culturally relevant and met the needs of the caregivers approached for the study. However, a limitation is that we did not explicitly obtain feedback during video development from those with lived experience, such as caregivers of opioid-exposed infants. For example, Naeim et al [32] describe the implementation of electronic video consent in their precision health research after the formation of a community advisory board consisting of respected individuals from the participants' home community who could speak of the communities' perspectives. They solicited regular feedback from the community advisory board, ensuring that the video language was suitable for a lay population and spoke effectively to the study's racially diverse sample.

Conclusions

This interactive digital tool was helpful in informing prospective participants about the study and reinforcing why caregivers of unexposed infants were being approached. The video was a helpful resource for families to understand the study's MRI procedures, because it provided a verbal description and included illustrations of a staff member preparing a swaddled, sleeping baby with protective earmuffs to enter the MRI scanner.

Acknowledgments

The authors acknowledge the caregivers who agreed to take part in this study; the medical and nursing colleagues at Case Western Reserve University, Cincinnati Children's Hospital Medical Center, University of Alabama at Birmingham, and Children's Hospital of Philadelphia; the Neuroimaging Core colleagues at Children's National Hospital; the data coordinating center staff at RTI (Research Triangle Institute) International; and National Institute of Child Health and Human Development colleagues Dr Andrew Bremer, MAS, MD, PhD, Dr Nahida Chakhtoura, MD, and Ms Stephanie Archer, MA. This research was supported by the National Institutes of Health (NIH) through the NIH Helping to End Addiction Long-Term (HEAL) Initiative (awards

XSL•FO
RenderX

1PL1HD101059-01, 3PL1HD101059-01S2). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or its NIH HEAL Initiative.

Data Availability

The dataset analyzed during this study is not currently publicly available because the OBOE study is ongoing. Following OBOE study completion, a public use dataset will be deposited in the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD) Data and Specimen Hub (DASH).

Conflicts of Interest

None declared.

References

1. Tam NT, Huy NT, Thoa LTB, et al. Participants' understanding of informed consent in clinical trials over three decades: systematic review and meta-analysis. Bull World Health Organ 2015 Mar 1;93(3):186-98H. [doi: 10.2471/BLT.14.141390] [Medline: 25883410]
2. Eiser C, Davies H, Jenney M, Glaser A. Mothers' attitudes to the randomized controlled trial (RCT): the case of acute lymphoblastic leukaemia (ALL) in children. Child Care Health Dev 2005 Sep;31(5):517-523. [doi: 10.1111/j.1365-2214.2005.00538.x] [Medline: 16101646]
3. Manafa O, Lindegger G, Ijsselmuiden C. Informed consent in an antiretroviral trial in Nigeria. Indian J Med Ethics 2007;4(1):26-30. [doi: 10.20529/IJME.2007.009] [Medline: 18630217]
4. Paasche-Orlow MK, Taylor HA, Brancati FL. Readability standards for informed-consent forms as compared with actual readability. N Engl J Med 2003 Feb 20;348(8):721-726. [doi: 10.1056/NEJMsa021212] [Medline: 12594317]
5. Palazzani L, Macioce F, Daverio M, Ferro V, Persampieri L. New strategies for increasing participation of patients from diverse cultural and religious backgrounds in clinical trials. BioLaw Journal - Rivista di BioDiritto 2019;1(15):101-122. [doi: 10.15168/2284-4503-403]
6. Rossi M, McClellan R, Chou L, Davis K. Informed consent for ankle fracture surgery: patient comprehension of verbal and videotaped information. Foot Ankle Int 2004 Oct;25(10):756-762. [doi: 10.1177/107110070402501011] [Medline: 15566709]
7. Shqaidef AJ, Saleh MYN, Ismail FK, Abu-Awad M, Khambay BS. A comparative assessment of information recall and comprehension between conventional leaflets and an animated video in adolescent patients undergoing fixed orthodontic treatment: a single-center, randomized controlled trial. Am J Orthod Dentofacial Orthop 2021 Jul;160(1):11-18. [doi: 10.1016/j.ajodo.2020.03.028] [Medline: 33902979]
8. Hall EW, Sanchez TH, Stein AD, et al. Use of videos improves informed consent comprehension in web-based surveys among internet-using men who have sex with men: a randomized controlled trial. J Med Internet Res 2017 Mar 6;19(3):e64. [doi: 10.2196/jmir.6710] [Medline: 28264794]
9. Glaser J, Nouri S, Fernandez A, et al. Interventions to improve patient comprehension in informed consent for medical and surgical procedures: an updated systematic review. Med Decis Making 2020 Feb;40(2):119-143. [doi: 10.1177/0272989X19896348] [Medline: 31948345]
10. Schenker Y, Fernandez A, Sudore R, Schillinger D. Interventions to improve patient comprehension in informed consent for medical and surgical procedures: a systematic review. Med Decis Making 2011;31(1):151-173. [doi: 10.1177/0272989X10364247] [Medline: 20357225]
11. Gesualdo F, Daverio M, Palazzani L, et al. Digital tools in the informed consent process: a systematic review. BMC Med Ethics 2021 Feb 27;22(1):18. [doi: 10.1186/s12910-021-00585-8] [Medline: 33639926]
12. Biesecker B, Raspa M, Rupert D, Moultrie R, Furberg R, McCormack LA. Making Clinical Trials More Patient-Centered Using Digital Interactive E-Consent Tools: RTI Press; 2019. [doi: 10.3768/rtipress.2019.op.0063.1910]
13. Cowan EA, Calderon Y, Gennis P, Macklin R, Ortiz C, Wall SP. Spanish and English video-assisted informed consent for intravenous contrast administration in the emergency department: a randomized controlled trial. Ann Emerg Med 2007 Feb;49(2):221-230. [doi: 10.1016/j.annemergmed.2006.07.934] [Medline: 17011074]
14. Fanaroff AC, Li S, Webb LE, et al. An observational study of the association of video- versus text-based informed consent with multicenter trial enrollment: lessons from the PALM study (patient and provider assessment of lipid management). Circ Cardiovasc Qual Outcomes 2018 Apr;11(4):e004675. [doi: 10.1161/CIRCOUTCOMES.118.004675] [Medline: 29625993]
15. McCormack LA, Wylie A, Moultrie R, et al. Supporting informed clinical trial decisions: results from a randomized controlled trial evaluating a digital decision support tool for those with intellectual disability. PLoS ONE 2019;14(10):e0223801. [doi: 10.1371/journal.pone.0223801] [Medline: 31644588]
16. Lawrence CE, Dunkel L, McEver M, et al. A REDCap-based model for electronic consent (eConsent): moving toward a more personalized consent. J Clin Transl Sci 2020 Apr 3;4(4):345-353. [doi: 10.1017/cts.2020.30] [Medline: 33244416]

XSL•FO
RenderX

17. Heffner JL, Watson NL, Serfozo E, et al. An avatar-led digital smoking cessation program for sexual and gender minority young adults: intervention development and results of a single-arm pilot trial. JMIR Form Res 2021 Jul 30;5(7):e30241. [doi: 10.2196/30241] [Medline: 34328430]
18. Hadskis M, Kenny N, Downie J, Schmidt M, D'Arcy R. The therapeutic misconception: a threat to valid parental consent for pediatric neuroimaging research. Account Res 2008;15(3):133-151. [doi: 10.1080/08989620801946917] [Medline: 18792535]
19. Bann CM, Newman JE, Poindexter B, et al. Outcomes of Babies with Opioid Exposure (OBOE): protocol of a prospective longitudinal cohort study. Pediatr Res 2023 May;93(6):1772-1779. [doi: 10.1038/s41390-022-02279-2] [Medline: 36042329]
20. Campbell J, Matoff-Stepp S, Velez ML, Cox HH, Laughon K. Pregnancy-associated deaths from homicide, suicide, and drug overdose: review of research and the intersection with intimate partner violence. J Womens Health (Larchmt) 2021 Feb;30(2):236-244. [doi: 10.1089/jwh.2020.8875] [Medline: 33295844]
21. Furberg RD, Ortiz AM, Moultrie RR, et al. A digital decision support tool to enhance decisional capacity for clinical trial consent: design and development. JMIR Res Protoc 2018 Jun 6;7(6):e10525. [doi: 10.2196/10525] [Medline: 29875084]
22. Furberg RD, Raspa M, Wheeler AC, McCormack LA, Bailey DB. A digital health app to assess decisional capacity to provide informed consent: protocol for a randomized controlled trial. JMIR Res Protoc 2018 Nov 19;7(11):e10360. [doi: 10.2196/10360] [Medline: 30455171]
23. Raspa M, Fitzgerald T, Furberg RD, et al. Mobile technology use and skills among individuals with fragile X syndrome: implications for healthcare decision making. J Intellect Disabil Res 2018 Oct;62(10):821-832. [doi: 10.1111/jir.12537] [Medline: 30105880]
24. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol 2008 Apr;61(4):344-349. [doi: 10.1016/j.jclinepi.2007.11.008]
25. Bester J, Cole CM, Kodish E. The limits of informed consent for an overwhelmed patient: clinicians' role in protecting patients and preventing overwhelm. AMA J Ethics 2016 Sep 1;18(9):869-886. [doi: 10.1001/journalofethics.2016.18.9.peer2-1609] [Medline: 27669132]
26. Mohammed-Ali AI, Gebremeskel EI, Yenshu E, et al. Informed consent in a tuberculosis genetic study in Cameroon: information overload, situational vulnerability and diagnostic misconception. Res Ethics 2022 Oct;18(4):265-280. [doi: 10.1177/17470161221106674]
27. Lee JER. Does virtual diversity matter?: Effects of avatar-based diversity representation on willingness to express offline racial identity and avatar customization. Comput Human Behav 2014 Jul;36:190-197. [doi: 10.1016/j.chb.2014.03.040]
28. Tajfel H, Turner JC. An integrative theory of intergroup conflict. In: Hogg MA, Abrams D, editors. Intergroup Relations: Essential Readings: Psychology Press; 1979:1-4 URL: https://alnap.org/help-library/resources/an-integrative-theory-of-intergroup-conflict/ [accessed 2025-02-19]
29. Zimmermann D, Wehler A, Kaspar K. Self-representation through avatars in digital environments. Curr Psychol 2023 Sep;42(25):21775-21789. [doi: 10.1007/s12144-022-03232-6]
30. Bailenson JN, Yee N, Blascovich J, Beall AC, Lundblad N, Jin M. The use of immersive virtual reality in the learning sciences: digital transformations of teachers, students, and social context. J Learn Sci 2008 Feb 15;17(1):102-141. [doi: 10.1080/10508400701793141]
31. Goffman E. The Presentation of Self in Everyday Life: Bantam; 1959.
32. Naeim A, Dry S, Elashoff D, et al. Electronic video consent to power precision health research: a pilot cohort study. JMIR Form Res 2021 Sep 8;5(9):e29123. [doi: 10.2196/29123] [Medline: 34313247]

Abbreviations

IRB: institutional review board
MRI: magnetic resonance imaging
OBOE: Outcomes of Babies With Opioid Exposure
REDCap: Research Electronic Data Capture
RTI: Research Triangle Institute
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

XSL•FO
RenderX

Edited by A Mavragani; submitted 26.04.24; peer-reviewed by L Narbey; revised version received 22.01.25; accepted 22.01.25; published 04.03.25.

Please cite as:
Newman JE, Clarke L, Athimuthu P, Dhawan M, Owen S, Beiersdorfer T, Parlberg LM, Bangdiwala A, McMillan T, DeMauro SB, Lorch S, Peralta-Carcelen M, Wilson-Costello D, Ambalavanan N, Merhar SL, Poindexter B, Limperopoulos C, Davis JM, Walsh M, Bann CM
Supplementing Consent for a Prospective Longitudinal Cohort Study of Infants With Antenatal Opioid Exposure: Development and Assessment of a Digital Tool
JMIR Form Res 2025;9:e59954
URL: https://formative.jmir.org/2025/1/e59954
doi:10.2196/59954

© Jamie E Newman, Leslie Clarke, Pranav Athimuthu, Megan Dhawan, Sharon Owen, Traci Beiersdorfer, Lindsay M Parlberg, Ananta Bangdiwala, Taya McMillan, Sara B DeMauro, Scott Lorch, Myriam Peralta-Carcelen, Deanne Wilson-Costello, Namasivayam Ambalavanan, Stephanie L Merhar, Brenda Poindexter, Catherine Limperopoulos, Jonathan M Davis, Michele Walsh, Carla M Bann. Originally published in JMIR Formative Research (https://formative.jmir.org), 4.3.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

A Digital Pornography Education Prototype Co-Designed With Young People: Formative Evaluation

Jake Turvey[1], MPH; Michelle Raggatt[1,2], BBiomedSci (Hons); Cassandra J C Wright[1,3], PhD; Angela C Davis[1,2], PhD; Meredith J Temple-Smith[4], DHSc; Megan S C Lim[1,2,5], PhD

[1]Young People's Health Research, Burnet Institute, 85 Commercial Rd, Melbourne, Australia
[2]School of Public Health and Preventive Medicine, Monash University, Melbourne, Australia
[3]Alcohol and other Drugs team, Menzies School of Health Research, Casuarina, Australia
[4]Department of General Practice and Primary Care, University of Melbourne, Melbourne, Australia
[5]Melbourne School of Population and Global Health, University of Melbourne, Parkville, Australia

Corresponding Author:
Megan S C Lim, PhD
Young People's Health Research, Burnet Institute, 85 Commercial Rd, Melbourne, Australia

Abstract

Background: Interventions to help young people make sense of sex and relationships in the context of widely available pornography are becoming increasingly supported in school settings. However, young people who experience disruptions to their education often have less access to such programs. Digital platforms may offer a more accessible method to deliver tailored sexual health and pornography literacy to young people who are disengaged from mainstream schooling, or who experience other types of structural disadvantage.

Objective: This study aimed to describe the formative evaluation of “The Gist” a co-designed online sexual health education and pornography literacy prototype designed to meet the sexual health information needs of structurally marginalized young people in Australia.

Methods: We conducted iterative workshops with 33 young people aged between 15 and 24 years recruited from an alternative education school in Melbourne, Australia. Through interactive activities, participants evaluated the overall prototype design, including its usability, desirability, inclusiveness, and potential for impact.

Results: Participants reported The Gist to be easy to use (17/20, 85%) and safe (19/23, 83%), with “hot” branding (25/30, 83%). However, perceived content relevance was dependent on the participants’ existing level of sexual health knowledge and experience, with only 31% (7/23) agreeing that “The Gist feels like it was made for me.” The interactive learning activities such as the debunked (myth-busting) and quiz features were among the most used and well-liked on The Gist platform. Low unprompted engagement with the prototype outside of facilitated workshop settings also confirmed previous researcher postulations that The Gist as a standalone digital platform is unlikely to meet the needs of this population group. Further design refinements are needed to improve user experience, including more interactive activities and visual information in place of heavily text-based features.

Conclusions: This study provides important insights into the design and sexual health information needs of structurally marginalized young people. Further research is needed to assess the overall efficacy of The Gist prototype, as well as its ability to positively influence young people’s sexual attitudes, beliefs, and behaviors. Future iterations should consider hybrid or face-to-face delivery models to better capture student engagement.

(JMIR Form Res 2025;9:e65859) doi:10.2196/65859

KEYWORDS

pornography; education; website; prototyping; evaluation; sexual health; sexual wellbeing; pornography literacy; young people; youth; adolescents; The Gist; sexual education; Australia; efficacy; digital health; co-design

Introduction

Pornography and Pornography Literacy

Pornography is easily accessible to, and commonly viewed by young people. In a 2017 cross-sectional survey of 941 Australians aged 15‑29 years, 84% of males and 23% of females viewed pornography at least weekly [1]. Some young people report that pornography is an important component of their sexual development [2], helping to define understandings of sexual pleasure, how to have sex, and diverse sexual experiences and sexualities [3]. However, pornographic content also commonly includes behavior which portrays problematic themes, including the promotion of heteronormative and

XSL•FO
RenderX

misogynistic gender norms, violence-supportive sexual objectification of women, lack of consent, and the poor representation of diverse people and body types [4]. Research has further shown that pornography can shape young people's sexual development in line with these problematic themes, and negatively influence their developing attitudes, beliefs, and sexual health behaviors [2,5,6].

Pornography literacy education is a technique which aims to reduce the adverse impacts of pornography use. Originating from media literacy theory, pornography literacy programs seek to increase young people's knowledge and understanding of healthy relationships and sexualities, whilst developing their ability to critique, evaluate, and challenge the various themes commonly present within pornographic content [7]. However, there remains a scarcity of formally evaluated pornography literacy interventions [8-10]. In a 2018 pilot study conducted in the United States, it was observed that students demonstrated improvements in various pornography-related knowledge items after participating in a college-based pornography literacy class [11]. Similarly, a longitudinal study of Dutch teenagers found that the relationship between pornography use and sexist attitudes became weaker the more a participant engaged in school-based pornography literacy classes [12]. These findings therefore suggest that there is potential for pornography literacy programs to shape the way in which young people perceive common underlying themes and messages within online pornographic content [12]. While pornography literacy programs show promise in mitigating the adverse impacts of pornography on young people, determining how best to integrate and deliver these programs within existing educational frameworks is crucial for their success.

Relationships and Sexuality Education

School-based sex education initiatives have traditionally been the preferred approach to improving sexual health outcomes amongst young people and mitigating risks of sexual health related harms such as sexually transmitted infections (STIs) and unplanned pregnancy [13]. In Australia, the delivery of relationships and sexuality education (RSE) has evolved in recent decades to include a broader range of contemporary sexual health topics and themes [14]. However, creating programs that adequately cater to the complex sexual health needs of diverse young people has posed an ongoing challenge for Australian educators and policymakers alike [15]. Furthermore, the actual education received by students is highly varied nationwide, as the development and implementation of RSE curricula remains the responsibility of each individual state or territory, while content delivery often differs between individual schools [14]. In a 2018 national survey of 6327 Australian high school students, more than 60% of respondents described their RSE experience as either irrelevant or marginally relevant to their sexual health needs, with many advocating for a greater focus on sexual communication, relationships, and pleasure [16]. The Australian government introduced mandatory consent education from 2023, covering areas of sexual consent, gender stereotypes, and power and relationship dynamics. However, there have been growing calls for the further addition other contemporary topics (such as pornography literacy) that take a more sex positive approach, and are more inclusive of alternative sexualities, gender identities, and other intersectionalities [17-19].

Furthermore, school-based education may not meet the needs of all young people. Young people have many different intersecting experiences that mean they have more or less opportunities for healthy sexual education and development. Young people who are disengaged from, or experience disruptions to their mainstream education or for example, have less access to school-based sex education programs [20]. In addition, school-based sex education is often not tailored to meet the specific needs of young people who are lesbian, gay, bisexual, transgender, queer, questioning, intersex, asexual (LGBTQIA+), young people from diverse cultural backgrounds, or young people with experiences of poverty [21-23]. International research also indicates that young people with experiences of trauma, family conflict, and other adverse childhood experiences have poorer access to sexual health information compared with other mainstream cohorts, and have an increased likelihood of engaging in risky sexual health behaviors [24]. Interventions to support young people's sexual development in the context of pornography therefore, need to account for this substantial variation in sexual health development among young people, and must consider the specific needs of individuals who have been excluded from current relationships and sexuality education.

Digital platforms have shown increasing promise as an acceptable way to reach and deliver tailored sexual health information to young people [25]. Recent systematic reviews have highlighted their effectiveness at shaping young people's knowledge of healthy sexual development, and reducing risky sexual health behaviors [25,26]. However, many of these interventions remain highly focused on singular or complementary behavioural outcomes in adolescents, such as STI prevention and contraception use [26]. As such, there is a distinct need for an online sexual health platform that is not only informed by, and tailored to the specific needs of structurally marginalized young people, but also provides them with a more comprehensive education on a wide range of sexual health topics, including pornography literacy.

Aims

The Gist is an online relationships- and sexuality-education platform, incorporating elements of pornography literacy, that was co-designed by young people, sexual health researchers, educators, and web developers [4]. Here, we describe a formative evaluation of The Gist prototype, an evaluation designed to pretest a program with the target group before implementation [27]. This evaluation aims to assess The Gist's usability, desirability, inclusiveness, and potential for impact, as well as identifying areas for future improvement.

Methods

The Intervention

Development of The Gist used a human-centred and co-design methodology [4]. Several months before this study, young people who identified as LGBTQIA+, were culturally and linguistically diverse, had experiences of family conflict or

XSL•FO
RenderX

educational disengagement participated in a series of iterative workshops to provide in-depth insights and design ideas for a sexual health education resource. Based on their insights and design recommendations, a minimally functioning digital prototype was developed. This is a working website that demonstrated most of the planned features of The Gist with some of the planned content. This study aimed to test and further refine the prototype design. To give context to this evaluation, the key design considerations and features of The Gist are briefly described below. More in-depth information about the methodology and findings from the design and development phase is available in a separate publication [4].

To give context to this evaluation, the key design considerations and features of The Gist are briefly described below. More in-depth information about the methodology and findings from the design and development phase is available in a separate publication [4].

The Gist prototype was designed as an online website application, with mobile-friendly capacity. Topics covered included pornography, body diversity, sex, consent, pleasure, healthy relationships, and sexual preferences. Language and imagery remained gender- and sexuality-neutral. Critical pornography literacy was built into this content by contrasting the healthy and ethical information about sex, relationships, and bodies with problematic messages often espoused in pornographic materials. The visual design of the digital prototype had an alternative, “retro 90s” and comic-book style as preferred by participants from the earlier co-design stage.

The website prototype presented short facts about various sexual health topics to engage, surprise and challenge the user. Users were then linked to more in-depth information about the topic, and encouraged to engage with interactive activities that further challenged their knowledge and built critical thinking skills around the respective topic. Key platform features included:

1. Get Facts: a series of rolling screens presented a short fact with an image, designed to reproduce the “stories” function used on social media apps such as Instagram and Snapchat.
2. Debunked*:* presented users with a short statement on a series of sexual health topics and prompted users to decide whether the statement was true or false. This feature was designed to challenge commonly held misconceptions and beliefs about certain sexual health topics.
3. Quizzes: brief interactive quizzes presented through multiple-choice questions designed to test the user’s knowledge of key sexual health themes.
4. Sexual Health Articles: provided in-depth written information on various sexual health topics broken down into digestible subsections.

Evaluation Questions

This paper reports on the testing and evaluation phase of the human-centered design process, which is part of an iterative cycle of prototyping, testing, and refining. The insights from this evaluation phase will be used to further refine and develop The Gist prototype for improved application. The key evaluation questions were: (1) is The Gist usable and accessible for young people?, (2) is The Gist desirable and relevant to young people?, (3) is The Gist inclusive and safe for young people?, and (4) do young people believe that The Gist has potential to positively impact other young people’s knowledge, attitudes, and confidence around respectful and healthy relationships, sexual health and wellbeing, and pornography literacy?

Recruitment and Setting

Evaluation participants were recruited from across 3 campuses of an alternative education school in Melbourne, Australia. Students at this school have experiences of fragmented school attendance or disengagement from mainstream school settings for a range of reasons including family conflict or breakdown, bullying, being gender and sexually diverse, living with a disability, or being culturally and linguistically diverse. Researchers first spoke to school staff, providing them with an overview of the study and demonstration of The Gist prototype so that teachers were informed of the content that would be shown to students, and could voice any concerns about certain topics that some students may find distressing. After feedback and strong support from staff was received, the research team then spoke to students, and provided a brief overview of the study, workshop activity and content themes, before distributing plain language participant information and consent forms to those interested.

Participants

A total of 33 young people participated in at least 1 of 7 workshops held across a 3-week period. Of the 33 participants; 15 identified as female, 13 identified as male and 1 identified as nonbinary (4 participants elected not to provide demographic information). Participants’ ages ranged from 15 - 24 years with a median age of 19 years. A total of 18 participants identified as heterosexual, 11 identified as LGBTQIA+, and 8 spoke a language other than English or in addition to English at home.

Workshop Methodology

In total, 3 groups of students across 2 campuses attended a series of 2 workshops each. Only 1 workshop was carried out at the third campus as these students had previously participated in the initial co-design research process for The Gist earlier in the year and were therefore already familiar with The Gist prototype.

All workshops were facilitated by 3 members of the research team, including 1 digital designer and 2 public health researchers. Two of these facilitators had been involved in the development of the prototype, all identified as young women (age<35 years old). School staff were present in 1 workshop (teacher supervision was required due to an incident earlier in the day unrelated to the research project). The duration of the workshops ranged from 70 - 135 minutes. An overview of the workshop aims is available in Multimedia Appendix 1. Workshop activities were specifically designed to be engaging, interactive, and trauma-informed, due to the unique learning and classroom needs of the participants. A total of 10 different activities were used across the various workshops, with a descriptive overview of each workshop activity available in Multimedia Appendix 2.

XSL•FO
RenderX

Data Collection

A range of qualitative and quantitative data was collected from across the 7 workshops.

Qualitative Data

Qualitative data included audio recordings of workshop activities and classroom discussions and physical material produced by students during workshop activities (ie, images of whiteboard brainstorming and completed worksheet activities). One researcher recorded written observations on classroom activities, levels of student engagement, and overheard classroom discussions. They used a semistructured note-taking guide developed by the lead researcher, to supplement audio recording with observations of nonverbal interactions. In addition, while students explored the prototype, 1 participant in each group (n=3) volunteered to be individually observed to collect more detailed information on website navigation and interactions. The facilitator recorded their actions in detail including which buttons were pressed and which features were accessed.

One researcher recorded written observations on classroom activities, levels of student engagement, and overheard classroom discussions. They used a semi-structured note-taking guide developed by the lead researcher, to supplement audio recording with observations of non-verbal interactions.

In addition, while students explored the prototype, one participant in each group (n=3) volunteered to be individually observed to collect more detailed information on website navigation and interactions. The facilitator recorded their actions in detail including which buttons were pressed and which features were accessed.

Quantitative Data

Quantitative data included surveys conducted during workshops explored students' receptiveness to The Gist prototype, including measures on usability and accessibility, desirability and relevance, safety and inclusivity, and potential to impact (Multimedia Appendix 3). A five-point Likert scale was used to assess survey responses, with response options ranging from 1= "Nah Shit," 2="Pfft," 3="Meeh," 4="It's OK," 5="Love it!!!." Students used stickers to indicate their responses on a large poster displaying the questions and website analytical data including unique page visits.

Website analytical data including unique page visits.

Data Analysis

This research was embedded in the constructivist paradigm. Following initial data cleaning and familiarisation, author [MR] developed a thematic framework guided by the evaluation questions, as well as emerging themes from initial data familiarisation processes. This framework was then applied to the data by authors [MTS and MR] by coding the observer notes, audio recordings, and images directly into NVivo 12 software (Lumivero). Researchers then summarized emerging themes into charts and tables to visualize the data as a whole, before mapping and synthesising differences across different participant groups and workshops. Throughout each phase, the thematic framework was refined as new themes or subthemes emerged.

Simultaneously, the quantitative data sourced from workshop activities and the workshop surveys were analysed by author [MR] using descriptive statistics. Then, a narrative integration approach was used to weave both qualitative and quantitative findings together on a theme-by-theme basis by authors [JT, MR, CW, and ML]. To achieve this, the quantitative findings were mapped to the thematic framework, and used to support or expand upon the emerging study themes, providing a stronger and more comprehensive understanding of the themes. Findings and key themes are arranged according to the evaluation questions. All supporting quotes are presented anonymously.

Ethical Considerations

Ethics approval for the research was granted by the Alfred Health Human Research Ethics Committee (373/18). All participants provided written informed consent. For participants under the aged of 18 years or younger, written or verbal parental consent confirmed via school staff was sought before workshop commencement. All data have been deidentified. Participants received a AUD20 (USD 13.90) voucher per workshop, plus a AUD10 (USD6.95) voucher if attending both workshops. Participants who were already familiar with the prototype and were only invited to a single workshop received a AUD25 (USD17.40) voucher.

Results

Usable and Accessible

Use and Usability

In general, participants were able to access and engage with The Gist website effectively, with 85% of workshop participants (17/20) responding with either "Love it!!!" or "It's OK" when asked if "The Gist is easy to use." However, website analytical data suggests that actual engagement with The Gist outside of the facilitated workshops was extremely low. Of the 1641 unique page visits recorded across the evaluation period, 1497 (91%) were recorded during workshop sessions (Figure 1), suggesting few participants independently engaged with the website.

XSL•FO
RenderX

Figure 1. Daily unique page visits on The Gist website during the evaluation period.

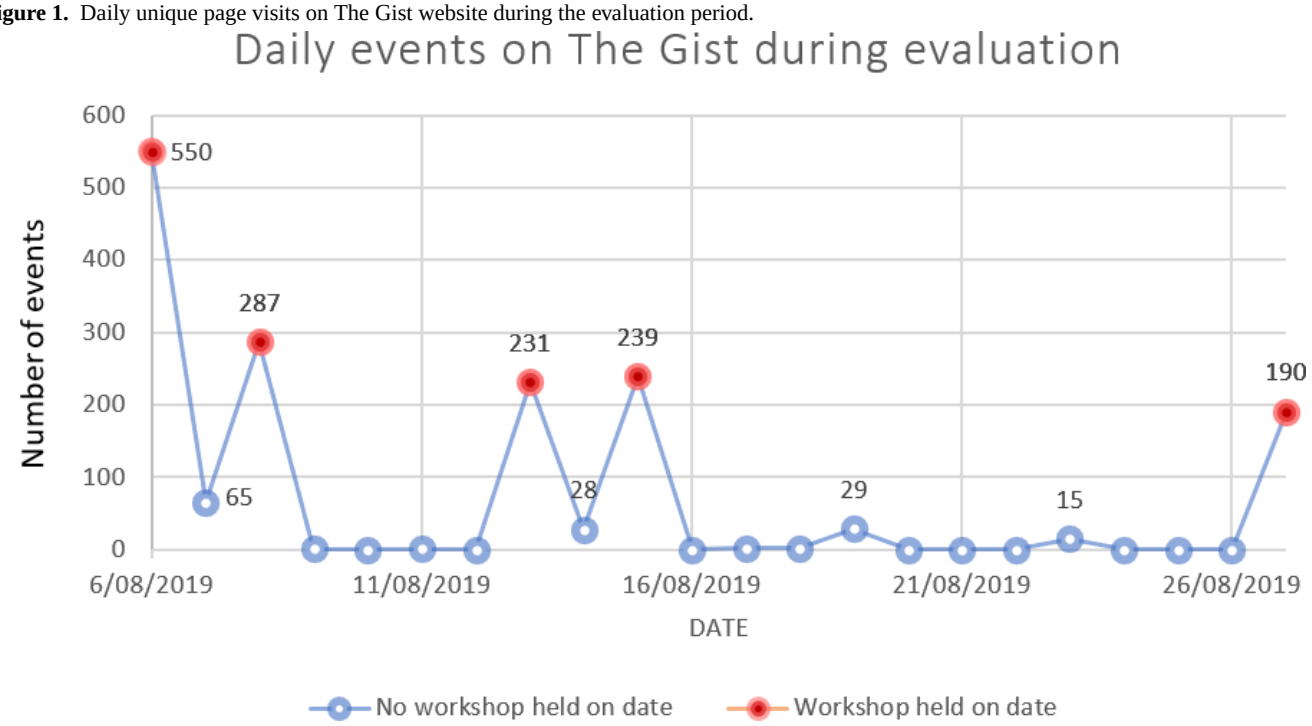

Researcher observations also noted that students primarily engaged with The Gist in small groups during workshop sessions, with many participants preferring to explore the website features alongside their peers, rather than as an individual endeavor. Students seemed to value the shared discussions which were stimulated by The Gist content and engaged each other in a collective learning process.

Non-Intuitive Design Features

Participants also reported design issues that meant several of The Gist features were not as intuitive or easy to use as participants would have liked. The Get Facts feature for example, was not clearly identifiable as content, with several participants initially attempting to use it as a navigation function. The ability to filter content within the search function, was not obvious to many participants who would either scroll past this option or requested clarification from facilitators.

> *It's all organised but I wouldn't really know how to go around using it such as what the app offers. But after a while at using it I would get used to it*

Further Information and Learning

Participants were receptive to the different interactive activities available. Over the evaluation period, the most popular and well used features were "Debunked" (302 events), "Quizzes" (127 events), and "Get Facts" (100 events). The home page received 240 visits and "Articles" were viewed 84 times.

> *I first checked "Get Facts" tab section. Informative, liked the flicking facts.*
>
> *When you click something and then you learn something, it's not complicated. The Get Facts [feature] works*

Articles were generally the least used learning feature, owing largely to their length and heavy use of text over interactive content and images, despite our attempts to prioritise brevity. Several participants commented that they did not enjoy reading, and quickly abandoned the feature or skipped over the articles altogether, preferring to access information through more interactive means. This is despite feedback from participants suggesting that the content did not go far enough with the information presented, with participants wanting the opportunity to learn more about specific topics that interested them. However, amongst the participants who reviewed the article feature, the majority (17/21, 81%) enjoyed (ie, "it's OK" or "Love it!!!") the format of the articles which used subheadings to break down relevant information, while 95% (20/21) thought the articles were easy to read. While a dot point summary of each article was provided, this was placed at the end of the article, with facilitators once again observing many participants either abandoning the article before reaching the summary, or simply scrolling past.

> *For me, because I don't like reading, I would say the articles weren't very important to me. But that's because I hate reading.*
>
> *I really love the [article] layout and the expandable boxes – the word I'd use is nifty.*
>
> *The articles were easy to read, [and] I did learn some cool things about consent.*

Desirable and Relevant

Content Relevance

Overall, participants held generally positive opinions of The Gist and its content, with 68% (15/22) of workshop participants agreeing with the statement, "The Gist is definitely something I would use." More than 73% (16/22) of participants also agreed that they "learnt some cool stuff from the Gist." Many participants reflected that The Gist was both helpful and informative, with the features of the website serving as effective

XSL•FO
RenderX

learning strategies to deliver new and engaging sexual health information.

> *Yes [relevant] because some of these things happen in my life*

Participants particularly enjoyed how the content was specifically catered toward young people, and sought to openly discuss topics that were frequently absent from within other mainstream sexual health education programs.

> *I really liked how The Gist is better than any sex education you get in school*
>
> *It's more appropriate than other sites because it's made for young people that don't understand sex ed at school or just want to learn.*
>
> *…get info about things that are not typically spoken about*
>
> *I like that it was open and uncensored*

However, while most participants felt that The Gist was relevant to young people in general, several participants reflected that the prototype was not overly relevant to them personally. Two opposing explanations were offered for this, with some participants suggesting that The Gist's content was more appropriate for a younger audience, with quizzes being "too easy," and that they already knew much of the information presented. In contrast, several other participants reflected that much of the content was too advanced for them, especially as they had little previous knowledge or sexual experience. They explained that there was not enough information for "beginners" who were not comfortable learning about sex-based topics. Similar apprehensions were expressed by a small number of students who identified as asexual, or those who had never engaged in watching pornographic material.

> *I probably won't use it that much because the info is stuff I mostly know.*
>
> *More information for younger/less experienced.*

Survey results supported this mixed response. When asked if "The Gist feels like it was made for me," 6 participants (26%) responded negatively ("Nah shit" or "Pfft"), 7 participants (31%) responded positively ("It's OK" or "Love it!!!"), and the largest proportion of students (10/22, 43%) remained neutral ("Meeh"). While many students did not think the content was personally relevant, most remained interested in participating and providing feedback on the prototype itself.

Aesthetically Appealing

Participants also gave feedback on the design and branding of The Gist, with the overall look and feel of the application appealing well to young people who took part in the evaluation. In total, 83% of participants (25/30) responded either "It's ok" or "I love it!!!" to the statement "The Gist branding is hot!" signifying significant approval of the website aesthetic and imagery.

Participants also indicated that they enjoyed the use of bright colours and specifically approved of the retro-90s styling which simultaneously appealed to a wide range of young people and created a nostalgic and safe feeling whilst using the site.

> *Comic book aesthetic is something that everyone can enjoy*
>
> *I really enjoy the colours and art style: it makes me feel safe and nostalgic somehow*

For some students, the visual depiction of genitalia, sex, and pornography (Figure 2) was seen as unnecessary and excessive at times, and may have worked to dissuade apprehensive students from engaging with the website.

> *Some of it is too much, this one that indicates cum in the Get Facts section, it's not necessary*

Figure 2. Features of The Gist, a digital prototype for pornography education. Panels show screenshots of the home page, Get Facts, article,and quiz.

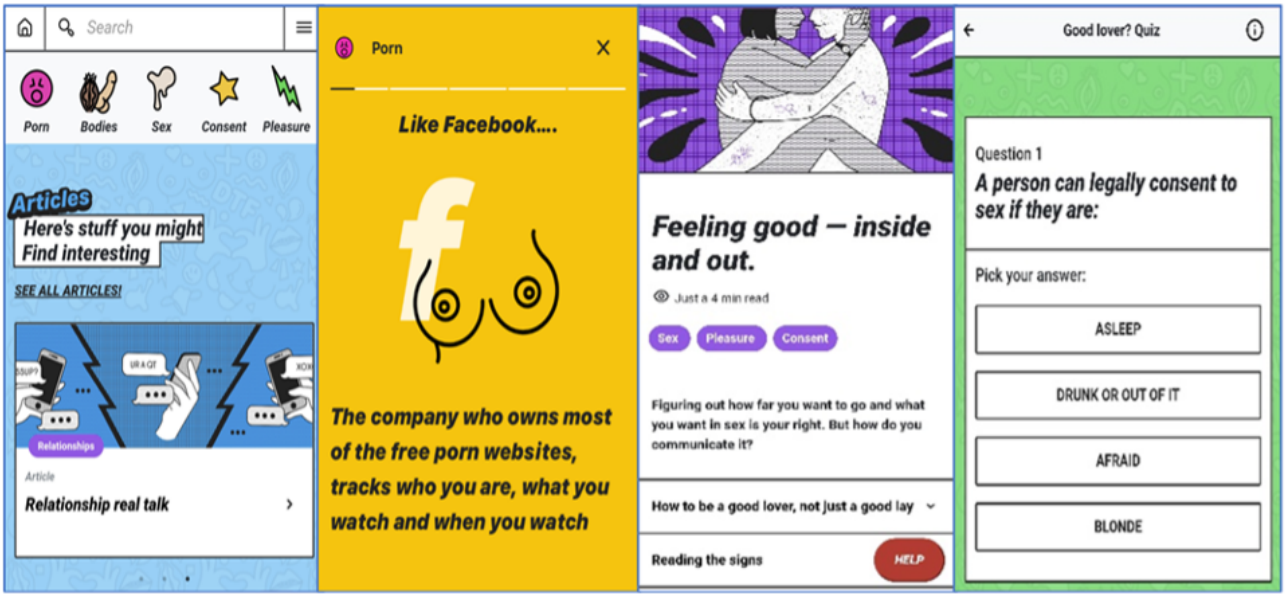

XSL•FO
RenderX

Safe and Inclusive

User Safety

When asked about how safe they felt using The Gist website, the majority (19/23, 83%) of participants responded either "It's ok" or "I love it!!!." Participants particularly enjoyed the ability to browse The Gist anonymously, especially due to the highly sensitive nature of The Gist's content and associated sexual health topics.

> *Safe place to get information that you might not want to ask someone about*
>
> *I like that its anonymous, you don't need an account to log in*

Participants were also asked if they would feel comfortable accessing The Gist website in public spaces such as on a train or tram, with a slight majority (12/22, 54%) responding with either "It's OK" or "I love it!!!." Those who felt less safe using The Gist expressed that these sentiments were largely driven by feelings of being unsafe on the internet more generally due to concerns of data monitoring and internet privacy rather than factors specifically related to The Gist website.

Inclusion of LGBTQIA+ Themes

In general, the content and language used throughout The Gist was seen as inclusive, particularly by those participants who identified as LGBTQIA+. Participants appreciated the nongendered wording used across the site, and the inclusive approach to sex and pornography use amongst gender and sexually diverse people.

> *It's not gender specific which is good*

Several participants, however, did express a desire to see greater inclusion of LGBTQIA+ themes, and for content to appear more obviously inclusive of sexual and gender diverse young people. This included having a dedicated LGBTQIA+ filter or tab section with relevant content rather than just having inclusive language and content spread throughout the website.

> *Make it more trans/non-binary (like a language filter). Have LGBTQIA+ filter e.g.: I click trans + all the content is trans-based*
>
> *fully inclusive – could be more obvious. Have LGBTQIA+ section*

Differences in Cultural Perspectives Toward Sex and Pornography

Discussions with participants who were recent migrants or from culturally and linguistically diverse backgrounds suggested that cultural understandings played a significant role in how they experienced The Gist website and its related content. Some young people indicated that in their culture or religion, acts of intimacy and sex were typically seen as only appropriate within marriage, and therefore influenced their view on the necessity of a sexual health education and pornographic literacy platform.

Several other participants were willing to discuss issues surrounding healthy relationships and consent, but were less comfortable discussing sexual behaviors and pornography. Even though some of these participants felt overwhelmed at times during the first workshop, all participants returned to the second workshop and appeared to be comfortable accessing The Gist and fully participating in the remaining activities. This highlights the importance of giving participants' time to develop rapport and to "warm up" to discussing sensitive topics during co-design processes.

> *Feel uncomfortable when quiz answers mention sex in relationships*
>
> *Confronting images*

Potential of The Gist to Positively Impact Other Young People

Participants were asked whether they believed The Gist had the potential to impact other young people's knowledge, attitudes, and confidence around healthy relationships, sexual health issues, and pornographic literacy. Most participants perceived that it would be an effective sexual health resource for other young students.

> *I like it and think it would be helpful and informative for teens and young adults.*
>
> *…it is a web site that helps young people no matter the gender/ethnicity to understand more about erotic matters.*

However, once again participants reflected that the ability of The Gist to impact other young people's sexual health and pornographic literacy was dependent on its relevance to the user, with greater efforts needed to provide appropriate information to different ages and knowledge levels.

> *It depends on what kinda age group that you want, I think more detail would be good if you want older people.*
>
> *I wish that it appealed to an older audience as well as a younger one*

When asked if "The Gist will make people want to watch porn," 63% of participants (14/22) disagreed with the statement, and only 1/22 agreed ("probs"). Participants stated that The Gist was more likely to reduce the stigma associated with watching pornography, and therefore makes young people feel less "scared" about watching pornographic material.

> *Using the website was easy to read & understand the gist will not make you want to watch porn it will instead give you an insight what porn really is!*
>
> *I think it might make people feel comfortable not scared to watch porn*

Discussion

Principal Findings

The Gist prototype was developed through a multistage co-design process, which includes this evaluation as part of the iteration and refinement stage. A key criticism made by participants was that much of the Gist's content and messaging was too text-heavy, particularly when providing users with in-depth or supplementary information on key sexual health topics. Feedback from participants emphasised this as a barrier to engagement, with many expressing their dislike of the articles

XSL•FO
RenderX

feature and other text-based elements of the website. This is despite this feature being intentionally designed based on previous research of young people's user experiences and online interaction patterns, as well as feedback from the initial prototype co-design stage, in which young people supported the progressive disclosure of supplementary text information [4]. This highlights the importance of a multistage co-design and prototype refinement process, with further work needed to optimize the Gist's articles for enhanced user engagement and acceptability. In contrast, the interactive learning activities such as the "Debunked" (myth-busting) and "quiz" features were amongst the most used and well-liked on The Gist platform. This affinity for highly interactive, image-based activities is unsurprising, given the well documented preference for kinaesthetic and visual learning styles amongst this cohort [28]. Additional efforts are therefore needed to further adapt key supplementary health information on The Gist website to more appropriate styles, including interactive activities, image-based content, and audio-visual materials. However, consideration must also be given to the resources associated with producing and maintaining such a large volume of supplementary content, including the capacity to update this information as young people's information needs change.

Despite enthusiastic engagement from young people during the evaluation workshops, website analytical data indicated that few participants accessed The Gist outside of the classroom setting. It was also noted that many participants preferred to explore The Gist collectively with fellow classmates, rather than as an individual learning exercise. This suggests that young people are potentially more likely to engage with The Gist as a sexual health and pornography literacy resource within a facilitated and supported learning environment. Consequently, this poses important questions for future iterations of The Gist, including how to strengthen independent engagement with the website, and whether The Gist as a standalone digital platform is a sufficient intervention to meet the needs of this population group. As noted by the research team in earlier co-design stages [4], previous international research has also shown that while young people appreciate the availability of sexual health materials online, they often seek out the guidance of trusted adults and authority figures to help them make sense of this information, and to determine its validity [29,30]. Thus, consideration should be given to the overall structure of The Gist as an educational intervention, and whether alternative delivery methods, (such as a potential hybrid model involving face-to-face sessions at schools, community sites, or services) are more effective approaches to supporting the sexual development and wellbeing of structurally marginalised young people.

This study adds to limited research into the development and evaluation of pornography education. A recent systematic review identified only 2 education programs that addressed pornography as a major focus [10]. Neither have described in depth the process of designing and developing the program.

Wider Implications

As so called "digital natives," there is a widespread perception of young people as highly proficient consumers and navigators of digital information. However, digital literacy is associated with socio-economic status, with online health information less accessible and interpretable to young Australians of lower socioeconomic standing [31,32]. The principal findings of this evaluation, therefore, provide valuable insights into how structurally marginalized young people perceive The Gist as a digital health resource, including key user interface and design preferences, information needs, and online learning requirements.

As pornography literacy programs emerge within mainstream education settings throughout Australia [7], it is important to recognize that young people are not homogenous, and their information needs are likely to differ despite similarities in demographic characteristics. Even amongst our small cohort of structurally marginalized young people, there was a wide variation in experience levels, previous knowledge, and attitudes toward sexual and pornographic themes. These differences reflect the fact that adolescent sexual development is a highly individualised process, shaped by a range of biological, psychological, and social factors [33]. For our participants, this also meant that the perceived relevance and desirability of the Gist prototype differed depending upon their own set of unique personal circumstances. Participants from culturally and linguistically diverse backgrounds also shared how their differing cultural perspectives and understandings of sex and pornography shaped their experiences with the prototype and receptiveness to the key health messages within. Future pornography literacy programs, therefore, need to ensure that the information and content presented is suited to, and tailorable for a wide range of experience and knowledge levels, even within similar demographic cohorts, and is offered in a way that is inclusive of all sexuality, gender, and culturally diverse people.

This evaluation has developed recommendations for further prototype refinements. It is vital that additional testing be carried out following these modifications to ensure The Gist remains highly relevant to structurally marginalized young people seeking accessible sexual health and pornography information. Future research is also needed to assess the impact of The Gist on young people's knowledge, attitudes, and behaviors, and to assess its effectiveness as a standalone sexual health and pornography literacy intervention.

Limitations

There are several limitations in which the findings described above should be considered. Study participants were recruited through a single service provider, and were therefore already familiar with other participants and supporting staff members. While this familiarity may have helped some participants to feel more comfortable during workshop sessions, it also introduces the potential for social and group biases. In the presence of their peers and teachers, participants may have sought to conform to the dominant group narrative by offering opinions that felt more socially acceptable. The sensitive nature of the research content may have further exacerbated this, with participants potentially hesitant to share their own personal experiences with sex and pornography. Similarly, while facilitators emphasized the need for critical and honest feedback,

XSL•FO
RenderX

participants may have felt inclined to provide positive assessments of the website as an unintended courtesy to the research team, particularly as 2 of the facilitators had been involved in development of the prototype. The design of the evaluation and the analysis of data was conducted by 3 researchers who were not directly involved in development of the prototype; however, they were close colleagues of those who did develop the prototype.

Young people were also free to engage with The Gist prototype as much or as little as they liked, however, not all participants attended both workshops, meaning that exposure varied. During the workshops, facilitators provided additional prompts and guidance only as needed, and there was no prerequisite that all elements of The Gist be explored and evaluated by participants. While this flexibility is an integral part of co-design research with structurally marginalized cohorts, it also means that again, participants may have had variable experiences with The Gist website, leading to differing assessments of the prototype. Similarly, as The Gist was an early prototype, access was only available through a password protected URL link, further limiting potential student engagement with the site. Finally, the evaluation did not use any validated instruments or apply techniques to enhance the validity or credibility of data analysis.

Conclusion

Amid growing concerns over the potential impact of pornography on the sexual development of young people, digital resources such as The Gist are emerging as potential alternatives to contemporary interventions seeking to combat these concerns. Research is still needed to determine the efficacy and impact of digital pornography literacy resources on young people's sexual attitudes, beliefs, and behaviors. The findings from this formative evaluation highlight the potential for creating engaging and highly pertinent online platforms for pornographic literacy, specifically tailored to the distinctive requirements of structurally marginalized young people. However, low engagement with the prototype outside of facilitated settings suggests that a hybrid delivery model that incorporates face-to-face elements may be preferable. This study emphasizes that intervention development is a long-term process requiring multiple iterations of prototype evaluation and refinement.

Acknowledgments

The project was funded by the Victorian Health Promotion Foundation (VicHealth), Melbourne, Australia. Angela Davis, Cassandra Wright, and Megan Lim were supported by National Health and Medical Research Council fellowships. Michelle Raggatt was supported by a Monash University Australian Postgraduate Award. The project was delivered in partnership with Hester Hornbrook Academy. The Gist prototype's content, illustration, design, and production were created with contributions from Mike Ebinum, Stacey Murphy, Jenny Walsh, Cat Wall, and Bonnie Eichelberger. The authors gratefully acknowledge the contribution to this work of the Victorian Operational Infrastructure Support Program received by Burnet Institute.

Data Availability

The data generated in this study are not publicly available due to participant confidentiality. Selected extracts of data could be made available upon reasonable request to the corresponding author, subject to approval from the relevant research ethics committee.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Workshop aims.
[DOCX File, 19 KB - formative_v9i1e65859_app1.docx]

Multimedia Appendix 2
Workshop activities.
[DOCX File, 19 KB - formative_v9i1e65859_app2.docx]

Multimedia Appendix 3
Workshop survey results.
[DOCX File, 19 KB - formative_v9i1e65859_app3.docx]

References

1. Lim MSC, Agius PA, Carrotte ER, Vella AM, Hellard ME. Young Australians' use of pornography and associations with sexual risk behaviours. Aust N Z J Public Health 2017 Aug;41(4):438-443. [doi: 10.1111/1753-6405.12678] [Medline: 28664609]

XSL•FO
RenderX

2. Quadara A, El-Murr A, Latham J. The Effects of Pornography on Children and Young People: An Evidence Scan: Australian Institute of Family Studies; 2017.
3. Waling A, Bellamy R, Ezer P, Kerr L, Lucke J, Fisher C. “It’s kinda bad, honestly”: Australian students’ experiences of relationships and sexuality education. Health Educ Res 2020 Dec 23;35(6):538-552. [doi: 10.1093/her/cyaa032] [Medline: 32929480]
4. Davis AC, Wright CJ, Murphy S, et al. A digital pornography literacy resource co-designed with vulnerable young people: development of “The Gist”. J Med Internet Res 2020 Jun 1;22(6):e15964. [doi: 10.2196/15964] [Medline: 32348268]
5. Marshall EA, Miller HA. The role of sexual scripts in the relationship between pornography use and sexual coercion. J Interpers Violence 2023 Apr;38(7-8):5519-5541. [doi: 10.1177/08862605221123291] [Medline: 36181240]
6. Massey K, Burns J, Franz A. Young people, sexuality and the age of pornography. Sex Cult 2021 Feb;25(1):318-336. [doi: 10.1007/s12119-020-09771-z]
7. Albury K. Porn and sex education, porn as sex education. Porn Studies 2014;1(1-2):172-181. [doi: 10.1080/23268743.2013.863654]
8. A presentation for parents and carersis now available!. It’s Time We Talked. URL: https://itstimewetalked.com [accessed 2025-02-14]
9. A teaching resource for years 7 to 10. The Practical Guide to Love, Sex and Relationships. URL: https://www.lovesexrelationships.edu.au [accessed 2025-02-12]
10. Aznar-Martínez B, Lorente-De-Sanz J, López-i-Martín X, Castillo-Garayoa JA. Pornography and gender-based violence: two neglected topics in sexuality education. A systematic review. Sex Educ :1-17. [doi: 10.1080/14681811.2024.2316154]
11. Rothman EF, Adhia A, Christensen TT, Paruk J, Alder J, Daley N. A pornography literacy class for youth: results of a feasibility and efficacy pilot study. Am J Sex Educ 2018 Jan 2;13(1):1-17. [doi: 10.1080/15546128.2018.1437100]
12. Vandenbosch L, van Oosten JMF. The relationship between online pornography and the sexual objectification of women: the attenuating role of porn literacy education. J Commun 2017 Dec;67(6):1015-1036 [FREE Full text] [doi: 10.1111/jcom.12341]
13. Crocker BCS, Pit SW, Hansen V, John-Leader F, Wright ML. A positive approach to adolescent sexual health promotion: a qualitative evaluation of key stakeholder perceptions of the Australian Positive Adolescent Sexual Health (PASH) Conference. BMC Public Health 2019 Jun 3;19(1):681. [doi: 10.1186/s12889-019-6993-9] [Medline: 31159767]
14. Pfitzner N, Ollis D, Stewart R, Allen KA, Fitz-Gibbon K, Flynn A. Respectful Relationships Education in Australian: National Stocktake and Gap Analysis of Respectful Relationships Education Material and Resources Final Report: Monash university Gender and Family Violence Prevention Centre; 2022.
15. Meehan C. “They’re much too young”: the entanglement of porn, pleasure and age in sex education. J Sex Res 2024 Apr 8:1-9. [doi: 10.1080/00224499.2024.2326960] [Medline: 38588605]
16. Fisher CM, Waling A, Kerr L, Bellamy R, Ezer P, Mikolajczak G, et al. 6th National Survey of Australian Secondary Students and Sexual Health 2018: Australian Research Centre in Sex, Health & Society, La Trobe University; 2019. [doi: 10.3389/fpubh.2019.00217]
17. Power J, Kauer S, Fisher C, Chapman-Bellamy R, Bourne A. The 7th National Survey of Australian Secondary Students and Sexual Health 2021: The Australian Research Centre in Sex, Health and Society, La Trobe University; 2022.
18. McKibbin G, Hamilton B, Humphreys C. Hold pornography to account – not education programs – for children’s harmful sexual behaviour. The Conversation 2016.
19. Crabbe M, Flood M. School-based education to address pornography’s influence on young people: a proposed practice framework. Am J Sex Educ 2021 Jan 2;16(1):1-37. [doi: 10.1080/15546128.2020.1856744]
20. Brown G, Sorenson A, Hildebrand J. How they got it and how they wanted it: marginalised young people’s perspective on their experiences of sexual health education. Sex Educ 2012 Nov;12(5):599-612. [doi: 10.1080/14681811.2011.634141]
21. O’Farrell M, Corcoran P, Davoren MP. Examining LGBTI+ inclusive sexual health education from the perspective of both youth and facilitators: a systematic review. BMJ Open 2021 Sep 2;11(9):e047856. [doi: 10.1136/bmjopen-2020-047856] [Medline: 34475162]
22. Lirios A, Mullens AB, Daken K, et al. Sexual and reproductive health literacy of culturally and linguistically diverse young people in Australia: a systematic review. Cult Health Sex 2024 Jun 2;26(6):790-807. [doi: 10.1080/13691058.2023.2256376]
23. Atkins R, Sulik MJ, Hart D, Ayres C, Read N. The effects of school poverty on adolescents’ sexual health knowledge. Res Nurs Health 2012 Jun;35(3):231-241. [doi: 10.1002/nur.21471] [Medline: 22431188]
24. Wood SK, Ford K, Madden HCE, Sharp CA, Hughes KE, Bellis MA. Adverse childhood experiences and their relationship with poor sexual health outcomes: results from four cross-sectional surveys. Int J Environ Res Public Health 2022 Jul 21;19(14):8869. [doi: 10.3390/ijerph19148869] [Medline: 35886718]
25. Wadham E, Green C, Debattista J, Somerset S, Sav A. New digital media interventions for sexual health promotion among young people: a systematic review. Sex Health 2019 Apr;16(2):101-123. [doi: 10.1071/SH18127] [Medline: 30819326]
26. Sewak A, Yousef M, Deshpande S, Seydel T, Hashemi N. The effectiveness of digital sexual health interventions for young adults: a systematic literature review (2010-2020). Health Promot Int 2023 Feb 1;38(1):1-14. [doi: 10.1093/heapro/daac104] [Medline: 36757346]

XSL•FO
RenderX

27. Bauman AE, Nutbeam D. In: North R, editor. Evaluation in A Nutshell: A Practical Guide to the Evaluation of Health Promotion Programs, 2nd edition: McGraw-Hill Education Medical; 2014.
28. Olivares-Cuhat G. Learner factors in a high-poverty urban middle school. Urban Educ 2011;9(1):1-11.
29. Albury K. 'Recognition of competition' versus will to app: rethinking digital engagement in Australian youth sexual health promotion policy and practice. Media International Australia 2019 May;171(1):38-50. [doi: 10.1177/1329878X19850224]
30. Lupton D. Young people's use of digital health technologies in the global north: narrative review. J Med Internet Res 2021 Jan 11;23(1):e18286. [doi: 10.2196/18286] [Medline: 33427684]
31. Thomson S, Bortoli L. Preparing Australian students for the digital world: results from the PISA 2009 digital reading literacy assessment. Australian Council for Educational Research 2012.
32. Neter E, Brainin E. eHealth literacy: extending the digital divide to the realm of health information. J Med Internet Res 2012 Jan 27;14(1):e19. [doi: 10.2196/jmir.1619] [Medline: 22357448]
33. Kar SK, Choudhury A, Singh AP. Understanding normal development of adolescent sexuality: a bumpy ride. J Hum Reprod Sci 2015;8(2):70-74. [doi: 10.4103/0974-1208.158594] [Medline: 26157296]

Abbreviations

LGBTQIA+: lesbian, gay, bisexual, transgender, queer, questioning, intersex, asexual
STI: sexually transmitted infection
VicHealth: Victorian Health Promotion Foundation

Edited by A Mavragani; submitted 27.08.24; peer-reviewed by H Pratomo, M Lindroth; revised version received 26.01.25; accepted 28.01.25; published 04.03.25.

Please cite as:
Turvey J, Raggatt M, Wright CJC, Davis AC, Temple-Smith MJ, Lim MSC
A Digital Pornography Education Prototype Co-Designed With Young People: Formative Evaluation
JMIR Form Res 2025;9:e65859
URL: https://formative.jmir.org/2025/1/e65859
doi:10.2196/65859

© Jake Turvey, Michelle Raggatt, Cassandra J C Wright, Angela C Davis, Meredith J Temple-Smith, Megan S C Lim. Originally published in JMIR Formative Research (https://formative.jmir.org), 4.3.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Purposes of Smartphone Health Care Apps and the Practicality of Their Functions in Disaster Situations: Qualitative Function Assessment Study

Sayuri Nonaka[1*], PhD; Susumu Fujii[1*], PhD; Kosuke Chris Yamada[1,2*], PhD

[1]Disaster Medical Informatics Lab, International Research Institute of Disaster Science, Tohoku University, Sendai, Japan

[2]Research Institute of Sport Medical Science, Tokai University, Hiratsuka, Japan

[*]all authors contributed equally

Corresponding Author:
Sayuri Nonaka, PhD
Disaster Medical Informatics Lab, International Research Institute of Disaster Science, Tohoku University, Sendai, Japan

Abstract

Background: Japan has experienced various natural disasters, including the Great East Japan Earthquake in 2011. It becomes crucial to focus on strengthening self-help measures through health care apps that are used in normal times to help people during disasters. However, little is known about what health care apps would be useful in times of disaster.

Objective: This study aimed to investigate the prevalent functionalities and purposes of using health care apps during normalcy, explore their potential utility, and propose strategies for disaster response through their utilization.

Methods: We focus on highly ranked health care apps (within the top 100 in the health care category for iPhones by Apple, Inc for a certain period) and reclassify their purpose of use, such as sleep, relaxation, and exercise, in detail. We also investigate the functions within each health care app (measurement, recording, advice, content provision, and guidance to actual services), based on which we determine their potential utilization during disasters and anticipate potential solutions to address disaster-related challenges. We also consider the ideal framework of health care apps in disaster response, exploring possibilities such as the necessity of new disaster-specific apps or the adaptation of existing health care apps for disaster scenarios.

Results: Among the 70 free apps, the predominant functions included "recording" (n=60 cases, 86%) and "measurement" (n=47 cases, 67%), primarily encompassing the mechanical functions of wearable devices and smartphones. A similar trend was seen in the 77 paid apps, but "content provision" (n=54 cases, 70%) was the most prevalent. Furthermore, the "content provision" function was particularly common in the "purposes of use" categories "sleep" (χ^2_{11}=29; P<.001), "relaxation" (χ^2_{11}=14.6; P<.001), and "exercise" (χ^2_{11}=9.3; P=.002). This suggested the possibility of using the content provision function in existing health care apps to support mental and physical health even during a disaster.

Conclusions: The widespread use of apps during normal times could minimize hesitation in adopting them during disasters. The findings emphasize the potential for augmenting disaster-specific content within existing apps rather than developing new ones. This approach aligns with the likelihood of preinstalled app use during emergencies, indicating a pragmatic strategy for enhancing disaster response content within prevalent apps.

(JMIR Form Res 2025;9:e56862) doi:10.2196/56862

KEYWORDS

health care; disaster medicine; mobile app; survey; disaster situations; self-reliance; Japan; disaster response; mobile phone

Introduction

Japan has experienced various natural disasters, including the Great East Japan Earthquake (GEJE) in 2011 [1-3], the Kumamoto earthquake in 2016, and wind and flood damage induced by Typhoons 15 and 19 in 2019 [4-6]. Furthermore, there is a 70% probability of a Nankai Trough earthquake and an earthquake directly under the Tokyo metropolitan area by 2050, necessitating urgent attention to disaster prevention, mitigation, and response [7].

The issues in drug prescribing that became apparent after the GEJE have been partially resolved using the Medication Record Book held by those affected by the disaster [8]. Electronic standards such as E-Kusu links have been established, and many drugstores and local governments are transitioning from paper-based to electronic medication handbooks [9]. Furthermore, various issues, such as managing chronic diseases, including diabetes and mental health during disasters, have been identified [10-17]. To reduce the burden on public and mutual aid support systems in times of disaster, we thought that familiar

XSL•FO
RenderX

digital technology, health care apps for smartphones, could be expected to strengthen self-help [18-21]. In this study, we focused on strengthening self-help measures during disasters using health care apps that are used in everyday life.

The best way to determine whether an app can be used effectively during a disaster is to investigate whether it can be used during a disaster. However, it is difficult to conduct such a survey during a disaster. In addition, it is time-consuming to apply the findings to the next disaster. Therefore, we assume that apps commonly used in the health care field related to human life will be immediately available at the time of a disaster and divide them into subgenres related to the issues at the time of a disaster. The potential for use in times of disaster was considered by surveying the implemented IT functions. This is because we thought that it would be difficult to use during a disaster if the function is mainly just measurement and record keeping. If the function can provide detailed methods such as content, there is a high possibility that it can be used during a disaster. In other words, rather than creating an app that provides content useful only for disasters, we thought it would be more effective to add only content for disaster apps within an app already installed and tied to the issues at the time of a disaster.

This study aimed to investigate the prevalent functionalities and purposes of using health care apps during normalcy, explore their potential utility, and propose strategies for disaster response through their utilization.

Specifically, we focus on highly ranked health care apps (within the top 100 in the health care category for iPhones by Apple, Inc for a certain period) and reclassify their purpose of use, such as sleep, relaxation, and exercise, in detail. We also investigate the functions within each health care app (measurement, recording, advice, content provision, and guidance to actual services), based on which we determine their potential utilization during disasters and anticipate potential solutions to address disaster-related challenges. We also consider the ideal framework of health care apps in disaster response, exploring possibilities such as the necessity of new disaster-specific apps or the adaptation of existing health care apps for disaster scenarios.

Methods

Survey Procedure

Overview

In this study, we conducted a comprehensive survey of current health care–related smartphone apps to determine their functionalities and intended purposes. Subsequently, we investigated the functions implemented for each specific purpose. Finally, based on the findings, we evaluated the potential utilization of each purpose during disaster scenarios, identifying their utility and challenges.

The procedure that we followed for the surveys is elaborated below.

Initial Selection (Survey 1)

For approximately 3 months, we targeted commonly used daily life health care or fitness apps by many Japanese people that were ranked within the top 100 on app download sites. The download site is Apple's App Store.

Functional Survey (Survey 2)

We focused on technical functionalities such as measurement (including wearable device integration), data recording, advice, and video provision. We classified functionalities into 5 categories: measurement function, recording function, advice provision function, content provision function, and actual service function.

Purpose Classification (Survey 3)

We surveyed and categorized the intended purposes for each app into eating habits, self-management, sleep, relaxation, walking, running, exercise, cycle management (menstruation or pregnancy), habit formation, childcare or parenting, teaching materials or practice, and actual service provision.

Analysis of Characteristics of Apps (Survey 4)

We examined the characteristics of functional implementation unique to each purpose identified in surveys 2 and 3. We also considered how to use them in the event of a disaster.

Survey Description

Overview

Using the cumulative results from surveys 1 to 4, we assessed the utility and challenges of each purpose during disaster situations. We evaluated whether the existing system could be directly used or required customization to address disaster-specific issues. Detailed descriptions of each survey method follow below.

Survey 1: Targeted Apps (Commonly Used Daily-Life Health Care Apps)

We included iPhone apps available on the App Store of Apple, Inc. There was no category for emergencies or disasters in the App Store, so the category for medical or health care–related options would be the best choice for dealing with disasters. According to the App Store category descriptions, the "Medical" category is described as "Apps that are focused on medical education, information management, or health reference for patients or health care professionals" [22]. The "Health & Fitness" category comprises "Apps related to healthy living, including stress management, fitness, and recreational activities" [22]. We targeted apps in the "Health & Fitness" category because they are accessible for general daily use rather than exclusively for health care professionals. The survey period was from October 2021 to April 2022, and the data were collected 4 times. The target apps were chargeable and free in the "Health & Fitness" category displayed exclusively on iPhones, excluding rankings on iPads and Macs. In addition, all apps had to be ranked within the top 100 during all 4 data collection periods. The App Store has a primary category for the app. If the category is inappropriate for an app, it violates the guidelines [22]. Whether an app is classified in the "Health & Fitness" category is determined by Apple, Inc.

XSL•FO
RenderX

Survey 2: Classification of "Implemented Function"

This survey aimed to identify 5 distinct functions (A: measurement function, B: recording function, C: advice provision function, D: content provision function, and E: actual service function) within the health care apps targeted in survey 1. Assessment of function implementation relied on scrutinizing app descriptions and screenshot images available in the apps. We did not verify these functions' actual performance or effectiveness or whether they precisely aligned with the description provided. As there is no existing classification system for these 5 functions, researchers (SF and KCY) who are an expert in app development and evaluation defined them, and another expert (SN) carried out the classification work. Multiple researchers discussed any doubt about the decision. The elucidation of each function is as follows.

- "A: measurement function" was examined to determine whether the function displays numerical values (measured values) by linking with smartphone sensors or external devices worn by individuals (eg, wearable devices and home-use measuring devices). Information on whether the values were stored was not sought. If the description mentioned support for the Apple health care app or Health Kit, it was deemed linked to a wearable device and was considered to possess a "measurement" function. If the value could only be entered manually, no "measurement function" was provided.
- "B: recording function" was defined as the ability to link, read, and save (record) values acquired from manual inputs and view the data. Information on the storage duration (history period) or storage frequency (history count) was sought. The recorded values could be manually entered, read from external data, or include sensor values obtained from "A: measurement function."
- "C: advice provision function" focused on investigating whether the app offered advice to users from various sources, including dedicated communities, experts, or artificial intelligence functions. The advice content was not limited to advice optimized based on personal data, advice from nonexperts in the community, or generalized advice. However, supportive advice, such as encouraging users, was not considered. Additionally, advice that only explained how to use the app or explained the provided content was not considered.
- "D: content provision function" was defined as providing specific health care methods (content), such as videos and recipes, within the app.
- "E: actual service function" focused on scrutinizing whether health care services were provided through human intervention. Services were assumed to be provided through digital services, such as scheduling appointments at physical stores, phone calls, and chats.

Figure 1 illustrates an app's function through a series of flow diagrams, depicted as a sequence from A to E.

Figure 1. Classification of functions. The function of an app through a series of flow diagrams, depicted as a sequence from A to E. Data measured by the user is recorded, advice is offered based on the recordings, content aligned with the advice is provided, and the actual service is offered in the form of human intervention. AI: artificial intelligence.

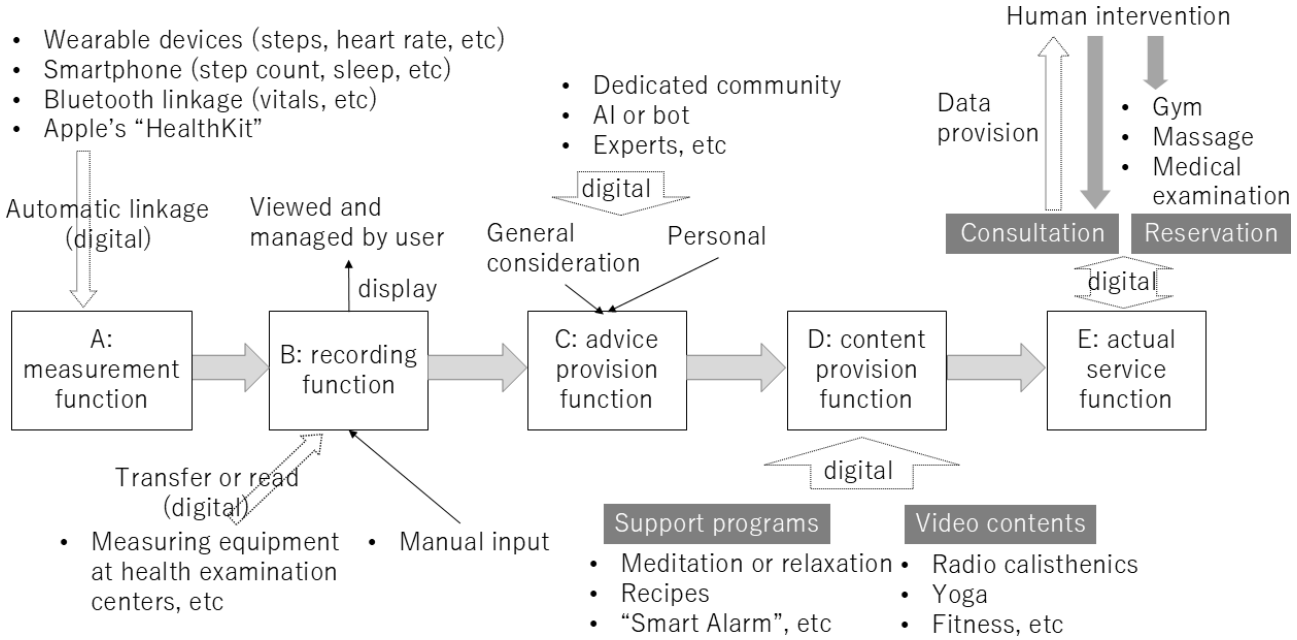

Survey 3: Categorization of "Purpose of Use"

Survey 1 results, categorized broadly under the health care classification by Apple, were reclassified into more detailed purposes of use because of the initial classification's broadness. The reclassification aimed to offer a more granular breakdown of the apps' intended purposes and was based on the service purpose outlined by the provider without considering any unintended or user-applied purposes of use. Determining the purpose of use relied on the information provided within the app description and screenshot images. Initially, the apps were classified into 15 categories (Multimedia Appendix 1), including "eating habits," "sleep," "relaxation," "walking," "running," "exercise," "cycle management (menstruation or pregnancy)," and "childcare or parenting." If an app did not fit any of these categories, they were further classified, and if an app served

XSL•FO
RenderX

multiple purposes, it was assigned multiple purposes of use. As there is no existing classification system for this category, researchers (SF and KCY) who are an expert in app development and evaluation defined them, and another expert (SN) carried out the classification work. Multiple researchers discussed any doubt about the decision.

Survey 4: Evaluation of Cross-Tabulation Results Between "Implemented Function" and "Purpose of Use"

Survey 4 involved a cross-tabulation evaluation between the purpose of use (survey 3) and the implemented functions (survey 2) classifications. The primary objective was to explore any notable characteristics in the implemented functions (A to E) concerning each purpose of use. For instance, the assessment aimed to discern commonalities, such as whether "A: measurement function" and "B: recording function" were prevalent in apps designed for sleep-related use or whether "A: measurement function" was rare in apps offering actual services. We also analyzed which "implemented function" was prominently implemented for each "purpose of use." Multiple response test was used for the analysis. A commercially available software program JMP (version 17.1.0; SAS Institute) was used for the analyses. This survey was compiled and analyzed by an expert health care app specialist (SN) and verified by some experts (SF and KCY) in disaster medicine. This evaluation, encompassing surveys 1‐4, facilitated discussions on the utility and challenges posed during disasters for each purpose of use. Additionally, it provided insights into precautions and suggestions for customization to enhance utility during disaster scenarios.

Ethical Considerations

This study does not involve the collection of data or interventions with human participants but is based on existing published app rankings and app descriptions. Therefore, we confirmed that this study did not meet the ethical review requirements at our institution. More specifically, this study satisfied the following conditions: it did not involve experiments, observations, or interventions in the human body. It did not collect or reuse personal information from participants and used only publicly available information, so the identification of individuals is impossible.

Results

Result 1: Targeted Survey of Commonly Used Daily-Life Health Care Apps

Apps in Apple's "Health & Fitness" category included those with "in-app purchases" (Apple Inc, 2023) for both free and paid downloads. "In-app purchases" included "consumable," "nonconsumable," "auto-renewable subscriptions," and "nonrenewing subscriptions" [23]. The presence of "in-app purchases" was not used to distinguish between free and paid download apps and those that required a fee to commence use were classified as paid download apps; Apple's classification at the time of the survey was used. The evaluation of free apps occurred on 4 occasions: October 22, 2021; October 29, 2021; November 22, 2021; and January 11, 2022. Paid apps were similarly scrutinized on 4 different dates: February 2, 2022; February 22, 2022; March 22, 2022; and April 12, 2022. In the second survey, the proportion of free apps was 86%, while that of paid apps was 84%. In the third survey, these values changed to 77% and 81%, respectively. By the fourth survey, the proportions further declined to 70% for free apps and 77% for paid apps. The results of survey 1 indicated that 147 (70 free and 77 paid) apps ranked in the top 100 on all 4 occasions.

Result 2: Classification of "Implemented Function"

Figure 2 provides an overview of the survey procedure. The results regarding the implemented functions within the apps are summarized in Table 1. It includes the 147 apps from Result 1, which ranked in the top 100 across all 4 survey points. Among the 70 free apps, the predominant function was "B: recording function" (n=60 cases, 86%), followed by "A: measurement function" (n=47 cases, 67%). These functionalities primarily leveraged the mechanical capabilities inherent in wearable devices and smartphones. The 77 paid apps also displayed a high implementation rate for "B: recording function" (n=48 cases, 62%) and "A: measurement function" (n=39 cases, 51%); however, the most notable difference lay in the implementation of "D: content provision function" (n=54 cases, 70 %) within paid apps, showcasing a higher rate than in free apps (n=28 cases, 40%). The app vendor specifically curated this content, marking a distinct trend from the mechanically driven performance witnessed in free apps.

XSL•FO
RenderX

Figure 2. Survey procedure overview. After survey 1 narrowed the target apps to 147, we examined the functions implemented in these apps (survey 2) and the purpose of using these apps (survey 3). Survey 4 analyzed the characteristics of apps based on their implemented functions and purpose.

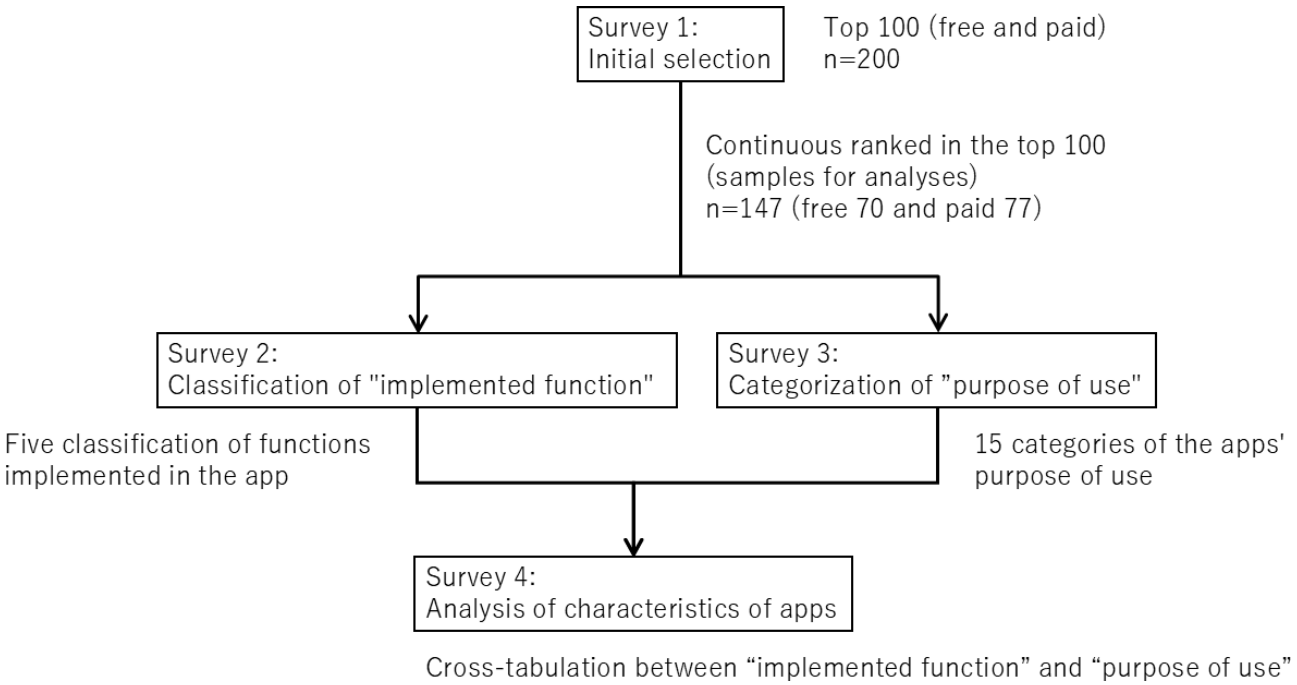

Table . Results of classification of functions implemented in the app.

	A: measurement	B: recording	C: advice provision	D: content provision	E: actual service
Free apps[a] (n=70), n (%)	47 (67)	60 (86)	32 (46)	28 (40)	10 (14)
Paid apps[b] (n=77), n (%)	39 (51)	48 (62)	18 (23)	54 (70)	0 (0)
Total downloaded apps (n=147), n (%)	86 (58.5)	108 (73.5)	50 (34)	82 (55.8)	10 (6.8)

[a]"Free" means free download apps.

[b]"Paid" means paid download apps.

Result 3: Categorization of "Purpose of Use"

The results of the "purpose of use" categorization are summarized in Table 2. Apple's categorization had a coarse granularity regarding the apps' "purpose of use"; therefore, the recategorization was performed. No additions other than the 15 "purpose of use" categorizations set during planning were performed. Among the 70 free apps, "walking" was the most common (n=16 cases, 23%), followed by "exercise" and "self-management" (n=15 cases each, 21%). Among the 77 paid apps, "sleep" (n=21, 27%) and "relaxation" (n=20, 26%) were the most common, followed by "exercise" (n=15, 20%). Overall (147 combined free and paid apps), the most popular activities were "exercise" and "sleep" (n=30 cases each, 20.4%), followed by "relaxation" (n=28, 19%) and "self-management" (n=25, 17%). Note that only paid apps were available for "childcare or parenting" and "teaching materials or practice," whereas only free apps were available for "actual service provision." The designation "not applicable" refers to apps lacking health care functionality. For the categories of "self-management," "habit formation," "teaching materials or practice," and "tools for Apple Health," a single purpose of use was assigned if it could not be determined to be the primary purpose of use, even if multiple purposes of use were plausible. However, if an app implied multiple purposes, such as "eating habits" and "exercise," the apps were tallied separately within each applicable category, leading to a cumulative count across various purposes.

XSL•FO
RenderX

Table . Results of the categorization of the apps' purpose of use.

Purpose of use	Free (n=70), n (%)	Paid (n=77), n (%)	Total (n=147), n (%)
Eating habits	10 (14)	3 (4)	13 (8.8)
Sleep	9 (13)	21 (27)	30 (20.4)
Relaxation	8 (11)	20 (26)	28 (19)
Walking	16 (23)	2 (3)	18 (12.2)
Running	5 (7)	2 (3)	7 (4.8)
Exercise	15 (21)	15 (20)	30 (20.4)
Cycle management (menstruation or pregnancy)	8 (11)	4 (5)	12 (8.2)
Childcare or parenting	0 (0)	3 (4)	3 (2)
Actual service provision	10 (14)	0 (0)	10 (6.8)
Self-management	15 (21)	10 (13)	25 (17)
Habit formation	2 (3)	4 (5)	6 (4.1)
Teaching materials or practice	0 (0)	4 (5)	4 (2.7)
Tools for Apple Health	0 (0)	3 (4)	3 (2)
Not applicable	9 (13)	0 (0)	9 (6.1)
Pack of apps	0 (0)	2 (3)	2 (1.4)

Result 4: Evaluation of Cross-Tabulation Results Between "Implemented Function" and "Purpose of Use"

The survey was conducted to determine the characteristics of the provided functions according to the purpose of use. "Tools for Apple Health," "pack of apps," and "not applicable" were excluded from survey 4 because they were not directly relevant to health-related purposes of use. The results of combining the "implemented function" categorized in survey 2 for each "purpose of use" classified in survey 3 are presented in Table 3.

Table . Cross-tabulation results of "implemented function" classifications by "purpose of use" of apps.

Purpose of use	A: measurement, n (%)	B: recording, n (%)	C: advice provision, n (%)	D: content provision, n (%)	E: actual service, n (%)
Eating habits (n=13)	12 (92)	13 (100)	10 (77)	11 (85)	1 (8)
Sleep (n=30)	17 (57)	16 (53)	11 (37)	30 (100)	1 (3)
Relaxation (n=28)	12 (43)	14 (50)	7 (25)	26 (93)	1 (4)
Walking (n=18)	18 (100)	18 (100)	8 (44)	8 (44)	2 (11)
Running (n=7)	7 (100)	7 (100)	5 (71)	6 (86)	0 (0)
Exercise (n=30)	21 (70)	27 (90)	13 (43)	26 (87)	2 (7)
Cycle management (menstruation or pregnancy; n=12)	8 (67)	12 (100)	11 (92)	7 (58)	3 (25)
Childcare or parenting (n=3)	1 (33)	1 (33)	1 (33)	3 (100)	0 (0)
Actual service provision (n=10)	4 (40)	9 (90)	6 (60)	6 (60)	10 (100)
Self-management (n=25)	22 (88)	25 (100)	9 (36)	3 (12)	0 (0)
Habit formation (n=6)	1 (17)	6 (100)	0 (0)	0 (0)	0 (0)
Teaching materials or practice (n=4)	0 (0)	0 (0)	0 (0)	4 (100)	0 (0)

Table 4 examines trends in the "purpose of use" for each "implemented function." For "A: measuring function," a significantly greater number of respondents reported "eating habits," "walking," "running," and "self-management," while

XSL•FO
RenderX

fewer respondents reported “relaxation,” “habit formation,” and “teaching materials or practice.” In the case of “B: recording function,” “eating habits,” “walking,” “cycle management,” and “self-management” were significantly more common, whereas “sleep,” “relaxation,” and “teaching materials or practice” were significantly less common. Regarding “C: advice provision function,” significantly more instances were found in “eating habits” and “cycle management” whereas “habit formation” reported significantly fewer occurrences. For “D: content provision function,” significantly more instances were observed in “sleep,” “relaxation,” “exercise,” and “teaching materials or practice,” whereas “self-management” and “habit formation” reported significantly fewer occurrences. In the case of “E: actual service function,” significantly more instances were noted in “actual service provision,” whereas “self-management” reported significantly fewer occurrences.

XSL•FO
RenderX

Table . Results of multiple response test: "purpose of use" categorized by "implemented function" of apps.

	χ^2 (*df*)	*P* value
A: measurement function		
Eating habits	5.7 (11)	.02
Sleep	0.7 (11)	.41
Relaxation	5.1 (11)	.02
Walking	16.1 (11)	<.001
Running	6.3 (11)	.01
Exercise	0.5 (11)	.48
Cycle management (menstruation or pregnancy)	0.0 (11)	.84
Childcare or parenting	1.2 (11)	.28
Actual service provision	2.4 (11)	.13
Self-management	7.5 (11)	.006
Habit formation	5.7 (11)	.02
Teaching materials or practice	8.2 (11)	.004
B: recording function		
Eating habits	5.4 (11)	.02
Sleep	12.0 (11)	<.001
Relaxation	13.8 (11)	<.001
Walking	7.5 (11)	.006
Running	2.9 (11)	.09
Exercise	1.8 (11)	.18
Cycle management (menstruation or pregnancy)	5.0 (11)	.03
Childcare or parenting	3.3 (11)	.07
Actual service provision	0.6 (11)	.44
Self-management	10.4 (11)	.001
Habit formation	2.5 (11)	.11
Teaching materials or practice	13.4 (11)	<.001
C: advice provision function		
Eating habits	8.4 (11)	.004
Sleep	0.0 (11)	.92
Relaxation	2.0 (11)	.16
Walking	0.4 (11)	.55
Running	3.3 (11)	.07
Exercise	0.4 (11)	.52
Cycle management (menstruation or pregnancy)	15.6 (11)	<.001
Childcare or parenting	0.0 (11)	.88
Actual service provision	2.1 (11)	.15
Self-management	0.0 (11)	.87
Habit formation	5.7 (11)	.02
Teaching materials or practice	3.8 (11)	.05
D: content provision function		
Eating habits	3.3 (11)	.07
Sleep	29.0 (11)	<.001

XSL•FO
RenderX

	χ^2 (*df*)	*P* value
Relaxation	14.6 (11)	<.001
Walking	2.2 (11)	.14
Running	2.0 (11)	.16
Exercise	9.3 (11)	.002
Cycle management (menstruation or pregnancy)	0.1 (11)	.81
Childcare or parenting	2.9 (11)	.09
Actual service provision	0.0 (11)	.91
Self-management	26.7 (11)	<.001
Habit formation	11.5 (11)	<.001
Teaching materials or practice	3.9 (11)	.049
E: actual service function		
Eating habits	0.0 (11)	.98
Sleep	0.9 (11)	.33
Relaxation	0.8 (11)	.38
Walking	0.3 (11)	.59
Running	1.1 (11)	.30
Exercise	0.0 (11)	.86
Cycle management (menstruation or pregnancy)	3.4 (11)	.06
Childcare or parenting	0.5 (11)	.49
Actual service provision	51.8 (11)	<.001
Self-management	3.9 (11)	.048
Habit formation	0.9 (11)	.33
Teaching materials or practice	0.6 (11)	.43

Discussion

Principal Results

The results from surveys 1‑4 provided insights into the potential usefulness of health care apps in the event of a disaster and highlighted specific concerns that need to be addressed for each purpose of use. We also discussed precautions and customization suggestions for use during times of disaster.

Survey 1 indicated that apps with enduring high rankings are more likely to be used during regular times. Considering the challenges of discovering and adopting new apps during disasters, using established apps beforehand could offer substantial advantages.

Survey 2 revealed that free apps primarily focus on measurement and recording, raising doubts about their direct applicability in times of disaster. By contrast, many paid apps offer content that can provide concrete advice, suggesting they could be used during disasters.

Survey 3 suggested a possibility of comprehensively addressing care demands during disasters by categorizing “exercise” as “physical” to maintain physical health and “sleep” and “relaxation” as “mental” (wellness). The survey also suggested using the app to provide comprehensive care during disasters.

Survey 4 explored the relationship between implemented functions and purpose of use, suggesting that health care apps require implementing functions such as recording and content provision. The implementation rate of content-providing functions was high for “physical” (eg, “exercise”) and “mental wellness” (eg, “sleep” and “relaxation”). In other words, apps frequently used during normal times may provide comprehensive (physical and mental) and specific content (response methods) during disasters. Using them during disaster scenarios is highly beneficial as they are already installed and configured.

In the following subsections, we discuss the anticipated effects of the apps’ specific “purpose of use” during disasters and suggest ways to improve their effectiveness in such scenarios, considering the apps’ “implemented functions.” The timing of using health care apps after a disaster is thought to vary from person to person. Therefore, the disaster phase we will examine here is the period after evacuation and after securing personal safety.

“Eating Habits” Apps

Eating habits during disasters are biased toward carbohydrates such as rice, bread, and noodles, rendering maintaining a nutritionally balanced diet difficult [24]. Moreover, this can lead to life-threatening problems for those with allergies or chronic illnesses because they cannot prepare their meals [25].

XSL•FO
RenderX

A study reported that, following the GEJE, patients with type 1 diabetes required 6 months for their HbA_{1c} values to return to predisaster levels [26].

The "eating habits" apps are characterized by their "C: advice provision function"; however, solely relying on this function in situations such as evacuation shelters can be challenging. In such times, obtaining suitable food items and meals customized to individual situations becomes crucial but can also be difficult to procure. For instance, devising a meal tailored to one's circumstances, especially considering evacuation rations, poses a significant challenge. Future endeavors should implement the "D: content provision function" to provide specific content. Moreover, the "C: advice provision function" includes a feature that allows information to be exchanged within a dedicated community. This feature enables the people impacted by the disaster to mutually exchange information and receive advice from medical specialists and nutrition managers.

"Sleep" and "Relaxation" Apps

Daily sleep is an indicator of mental health [27]. Based on sleep status, even in the event of a disaster, the early detection of mental health problems may be possible [28,29].

Many of the "sleep" and "relaxation" apps were equipped with "D: content provision function." Several users used paid apps, indicating that they have routine difficulties with "sleep" and "relaxation" and seek content that addresses these concerns. Although whether the content has a scientific basis is debatable, sleep and relaxation are particularly important for maintaining mental health even in times of disaster, and the possibility of providing a concrete method (content) can render the content useful in disaster situations. In the future, providing content tailored to the special environment of a disaster, such as living in an evacuation center, will be crucial.

"Walking," "Running," and "Exercise" Apps

In evacuation centers, space limitations restrict physical activity, leading to a sedentary lifestyle and potential health issues such as poor circulation and muscle weakness due to a lack of exercise. Physical exercises, particularly for older adults, play a crucial role in preventing frailty [30]. However, challenges were observed during the GEJE, where exercise acceptance within evacuation centers was limited [31]. To address these issues, strategies have been proposed to include exercise instructors in disaster medical assistance teams and form a disaster-exercise support network, emphasizing collaboration with other organizations and exercise initiatives that include local governments [32].

While "walking" and "running" apps showed no significant implemented functions, the "exercise" apps tended to have more "D: content provision function." Notably, "A: measurement function" was significantly implemented in "walking" and "running" apps. However, measuring the number of steps or distances might offer limited utility during disasters. Unless a specific means of "exercise" is provided, effectively using the system in the event of a disaster will be difficult. The challenge will be to provide an "exercise" content that can be implemented within limited spaces, including evacuation centers.

"Cycle Management (Menstruation or Pregnancy)" Apps

In the aftermath of the GEJE, female evacuees encountered challenges, including inadequate sanitary products, underwear, and essentials for infants and young children [33,34]. Women (especially pregnant women and caregivers) are at higher risk of facing mental health problems during disasters [35,36], highlighting the stressful environment for several women in such situations. Frameworks such as The Sendai Framework for Disaster Risk Reduction, which describes women's perspectives and participation in disaster reduction and building women's capacities [37], and guidelines from the Cabinet Office ("Women's Perspective for Strengthening Disaster Response Capabilities: Disaster Prevention and Recovery Guidelines from a Gender Equality Perspective") [38,39] focus on enhancing women's roles in disaster reduction and recovery.

Several women manage their menstrual and pregnancy cycles during normal times, suggesting that incorporating health care apps into their daily lives might be straightforward. For instance, as of 2022, the "LunaLuna" app, one of the "cycle management" apps, has been downloaded more than 19 million times [40].

By predicting these things, we can predict the demand for sanitary products, determine what is in short supply, and know when there is a problem with our bodies, such as when the cycle is different from normal. We can use this app effectively.

"Childcare or Parenting" Apps

Only 3 apps were identified, all of which were paid apps used by parents to support their children's development and growth. Therefore, their use during disasters would be challenging.

"Actual Service Provision" Apps

The "actual service" apps provided convenience to existing members, such as digital reservations and consultations. Currently, these apps are only intended for use during normal times. However, they could prove useful during evacuations if they offer real services related to supplies and medical care.

"Self-Management" Apps

Previous studies have demonstrated the potential benefits of using wearable devices to track activity, physiological functions, and lifestyle data. Having undergone rigorous testing for data accuracy, they can have a positive physical and mental impact [41-44].

The "self-management" apps tended to implement "A: measurement function" and "B: recording function," but lacked "C: advice provision function" and "D: content provision function." Currently, these apps only measure and record data from wearable devices, with limited use of the collected data.

However, ongoing research on wearable device technology suggests that it may become an essential tool in future health care apps by facilitating behavioral change and offering benefits to health care professionals and patients [45]. To date, most "self-management" apps focus on measurement and recording, rendering them less effective during a disaster scenario.

XSL•FO
RenderX

"Habit Formation" Apps

These apps primarily encouraged behavioral change through reminders and consistent recording. However, its utility during disaster situations seems improbable at present.

"Teaching Materials or Practice" Apps

These apps served only "D: content provision function." They could prove useful if tailored precisely to specific needs. Creating content based on accurate knowledge of the situation would be safer and more precise during disasters than a general internet search. They might serve as a straightforward contingency preparedness tool not typically used in normal circumstances.

Overall Perspective

Disaster situations require health care apps that not only measure and record day-to-day activities but also provide specific methods and actual services addressing unique disaster-related issues. These apps should encompass functionalities that can provide concrete methods specially tailored to disaster situations. We believe this is feasible using the content provision function in this study. For example, in terms of food, the blood sugar levels of diabetics will likely be difficult to control when eating emergency food in the event of a disaster. Emergency food generally consists mainly of carbohydrates, so there is a risk of sudden fluctuations in blood sugar levels and high blood sugar. In addition, there is a possibility that dehydration symptoms will worsen diabetes symptoms due to a lack of hydration due to concerns about using the toilet [46]. Therefore, if the app content could encourage people to drink more water and provide information on the fact that the rice balls and sweet buns provided immediately after a disaster contain sodium and the recommended amount of food to eat, it could be very useful. However, avoiding hypoglycemia and replenishing sugar levels would still be necessary.

Similarly, there is a trend for apps that deal with issues such as sleep, exercise, and stress management during disasters to have a significant implementation of the "D: content provision function," which also suggests the possibility of managing health in terms of both "mental" and "physical" aspects even during disasters. In other words, the key is not to develop new apps but to prepare content specific to disasters. Because these apps are widely used normally, they are highly reliable and have low initial implementation costs. Therefore, enhancing existing content is more reasonable than developing new apps as disaster countermeasures.

Limitations

There are some limitations to this study. First, this study focused only on Apple's App Store apps. Although we assumed that the top 100 apps were used among all the users of smartphones, they are not necessarily apps that are used frequently. It could be needed to expand the investigation to Android apps as well. Second, this study could not clearly define the "purpose of use" classification. The "purpose of use" classification assumed identical functionality for apps serving multiple purposes. For example, if an app served both "eating habits" and "exercise," the assumption was that identical functionality existed for both "purposes of use." However, while the "D: content provision function" might be implemented for "eating habits," it might not extend to "exercise." Despite this, the "D: content provision function" was also assumed to be implemented in the "exercise" category for discussion purposes. This was because of its existence in the app itself. In categories such as "self-management," "habit formation," and "teaching materials or practice," where the app providers did not specify distinct health care purposes, leaving them to the users, overlap between multiple "purposes of use" was assumed to be absent. Third, there are other aspects of the app's characteristics. In this study, we investigated the characteristics of apps that could be used during a disaster in ordinary time. It is also possible that apps from genres (categories) uncovered in this study could be adapted to the health field. Finally, we believe that the results of this study are not relevant in times of immediate disaster or response to disaster but are effective in the aftermath and the recovery phase.

Conclusions

In this study, we underscored the high utility of apps used in everyday contexts, highlighting their potential usefulness even in disaster situations owing to their reliability and reduced initial implementation.

The "D: content provision function" emerged as crucial, particularly during disasters when specific solutions are essential. Since the "D: content provision function" tended to be significantly implemented in the "sleep," "relaxation," and "exercise" categories, we observed the potential for health management in both the "mental" and "physical" realms even amid disasters. We believe that providing tailored content that can be used effectively during such crises is an important issue for the future.

In the future, apps in subgenres where the content provision function had a significantly high implementation rate will be studied individually to verify their effectiveness.

Acknowledgments

Part of this research was conducted as part of a joint research project with NTT Space Environment and Energy Laboratories entitled "Study of an intervention model for the construction of a next-generation evacuation system, the improvement of life after evacuation and the promotion of actions toward disaster readiness by local residents" and Research on "Digital Hospitals in Metaverse Spaces that Protect Local Lives" Realized through the Integration of Disaster Science and IOWN, and JST-OPERA: Joint Research Promotion Program for Industry-Academia Co-Creation Platform (Consortium for Co-Creation of Electric Power and Telecommunications Fusion Networks). Generative artificial intelligence was not used in any portion of the manuscript writing.

XSL•FO
RenderX

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

SN and SF conceived the study. SN compiled the data, developed the statistical analysis plan, and conducted statistical analyses. KCY contributed to the interpretation of the results. SN drafted the original manuscript. SF supervised the conduct of this study. All authors reviewed the manuscript draft and revised it critically on intellectual content. All authors approved the publication of the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Categorization of the app's purpose of use.
[DOCX File, 16 KB - formative_v9i1e56862_app1.docx]

References

1. 2011 Tohoku-Oacific Ocean earthquake (Great East Japan Earthquake) [Web page in Japanese]. Cabinet Office. 2022. URL: https://www.bousai.go.jp/2011daishinsai/pdf/torimatome20220308.pdf [accessed 2023-11-17]
2. Egawa S. Progress of disaster medicine during ten years after the 2011 Great East Japan Earthquake. Tohoku J Exp Med 2021 Mar;253(3):159-170. [doi: 10.1620/tjem.253.159] [Medline: 33762509]
3. Ishigaki A, Higashi H, Sakamoto T, Shibahara S. The Great East-Japan Earthquake and devastating tsunami: an update and lessons from the past Great Earthquakes in Japan since 1923. Tohoku J Exp Med 2013 Apr;229(4):287-299. [doi: 10.1620/tjem.229.287] [Medline: 23583960]
4. Damage caused by Typhoon No 15 in 2019 [Web page in Japanese]. Cabinet Office. 2019. URL: https://www.bousai.go.jp/updates/r1typhoon15/pdf/r1typhoon15_30.pdf [accessed 2023-11-17]
5. Damage from the 2016 Kumamoto Earthquake with the epicenter in Kumamoto Prefecture, Japan [Web page in Japanese]. Cabinet Office. 2019. URL: https://www.bousai.go.jp/updates/h280414jishin/pdf/h280414jishin_55.pdf [accessed 2023-11-17]
6. Damage caused by Typhoon no 19 in 2019 [Web page in Japanese]. Cabinet Office. 2020. URL: https://www.bousai.go.jp/updates/r1typhoon19/pdf/r1typhoon19_45.pdf [accessed 2023-11-17]
7. Summary of the white paper on land, infrastructure, transport and tourism in Japan. MLIT. 2021. URL: https://www.mlit.go.jp/statistics/content/001442571.pdf [accessed 2023-11-17]
8. Examples of the use of the medication handbook during the Great East Japan Earthquake [Web page in Japanese]. Japan Pharmaceutical Association. 2012. URL: https://www.nichiyaku.or.jp/files/co/activities/shinsai_techo.pdf [accessed 2025-02-26]
9. E-kusu link [Web page in Japanese]. Japan Pharmaceutical Association. 2016. URL: https://www.nichiyaku.or.jp/e_kusulink/ [accessed 2023-11-17]
10. Abe M, Akaishi T, Ishizawa K, et al. A massive natural disaster, the Great East Japan Earthquake, and the incidence of dialysis due to end-stage kidney disease. J Nephrol 2022 Apr;35(3):719-724. [doi: 10.1007/s40620-021-01140-9] [Medline: 34642850]
11. Hirai H, Okazaki K, Ohira T, et al. Suboptimal diabetic control and psychological burden after the triple disaster in Japan: the Fukushima Health Management Survey. BMJ Open Diabetes Res Care 2021 Sep;9(1):e002007. [doi: 10.1136/bmjdrc-2020-002007] [Medline: 34497045]
12. Ishikuro M, Obara T, Murakami K, et al. Relation of disaster exposure with maternal characteristics and obstetric outcomes: the Tohoku medical megabank project birth and three-generation cohort study. J Epidemiol 2023 Mar 5;33(3):127-135. [doi: 10.2188/jea.JE20210052] [Medline: 34219121]
13. Kishimoto M, Noda M. Diabetes care: after the Great East Japan Earthquake. J Diabetes Investig 2013 Jan 29;4(1):97-102. [doi: 10.1111/jdi.12025] [Medline: 24843638]
14. Kotozaki Y, Tanno K, Sakata K, et al. Association between social isolation and total mortality after the Great East Japan Earthquake in Iwate Prefecture: findings from the TMM CommCohort study. Int J Environ Res Public Health 2022 Apr 5;19(7):4343. [doi: 10.3390/ijerph19074343] [Medline: 35410023]
15. Nakaya N, Nakamura T, Tsuchiya N, Tsuji I, Hozawa A, Tomita H. The association between medical treatment of physical diseases and psychological distress after the Great East Japan Earthquake: the Shichigahama Health Promotion Project. Disaster Med Public Health Prep 2015 Aug;9(4):374-381. [doi: 10.1017/dmp.2015.52] [Medline: 25912962]
16. Ohira T, Nakano H, Nagai M, et al. Changes in cardiovascular risk factors after the Great East Japan Earthquake. Asia Pac J Public Health 2017 Mar;29(2_suppl):47S-55S. [doi: 10.1177/1010539517695436] [Medline: 28330394]

XSL•FO
RenderX

17. Orui M, Sato Y, Tazaki K, Kawamura I, Harada S, Hayashi M. Delayed increase in male suicide rates in tsunami disaster-stricken areas following the Great East Japan Earthquake: a three-year follow-up study in Miyagi Prefecture. Tohoku J Exp Med 2015 Mar;235(3):215-222. [doi: 10.1620/tjem.235.215] [Medline: 25765170]
18. Bucci S, Schwannauer M, Berry N. The digital revolution and its impact on mental health care. Psychol Psychother 2019 Jun;92(2):277-297. [doi: 10.1111/papt.12222] [Medline: 30924316]
19. Desveaux L, Shaw J, Saragosa M, et al. A mobile app to improve self-management of individuals with type 2 diabetes: qualitative realist evaluation. J Med Internet Res 2018 Mar 16;20(3):e81. [doi: 10.2196/jmir.8712] [Medline: 29549070]
20. Dugas M, Wang W, Crowley K, et al. Engagement and outcomes associated with contextual annotation features of a digital health solution. J Diabetes Sci Technol 2022 Jul;16(4):804-811. [doi: 10.1177/1932296820976409] [Medline: 33355003]
21. Fujii S, Nonaka S, Nakayama M. Use of medical information and digital services for self-empowerment before, during, and after a major disaster. Tohoku J Exp Med 2021 Nov;255(3):183-194. [doi: 10.1620/tjem.255.183] [Medline: 34853210]
22. Choosing a category. Apple Developer. 2023. URL: https://developer.apple.com/app-store/categories/ [accessed 2023-11-17]
23. In-app purchase. Apple Developer. 2023. URL: https://developer.apple.com/in-app-purchase/ [accessed 2023-11-17]
24. Tokumaru O, Fujita M, Nagai S, Minamikawa Y, Kumatani J. Medical problems and concerns with temporary evacuation shelters after great earthquake disasters in Japan: a systematic review. Disaster Med Public Health Prep 2022 Aug;16(4):1645-1652. [doi: 10.1017/dmp.2021.99] [Medline: 34103106]
25. Sudo N, Shimada I, Tsuboyama-Kasaoka N, Sato K. Revising “nutritional reference values for feeding at evacuation shelters” according to nutrition assistance by public health dietitians based on past major natural disasters in Japan: a qualitative study. Int J Environ Res Public Health 2021 Sep 24;18(19):10063. [doi: 10.3390/ijerph181910063] [Medline: 34639364]
26. Kamimura M, Hakoda A, Kanno J, Nishii A, Igarashi Y, Fujiwara I. Glycemic control in type 1 diabetic patients following the Great East Japan Earthquake and tsunami. J Jpn Diabetes Soc 2014;57:16-21. [doi: 10.11213/tonyobyo.57.16]
27. Recognize early signs of mental illness [Web page in Japanese]. National Center of Neurology and Psychiatry. 2023. URL: https://kokoro.ncnp.go.jp/health_understanding.php#02 [accessed 2023-11-17]
28. Li X, Buxton OM, Hikichi H, et al. Predictors of persistent sleep problems among older disaster survivors: a natural experiment from the 2011 Great East Japan earthquake and tsunami. Sleep 2018 Jul 1;41(7):zsy084. [doi: 10.1093/sleep/zsy084] [Medline: 29726979]
29. Nakamura Y, Suda T, Murakami A, et al. Sleep disturbance of evacuees in Minamisanriku Town after Great East Japan Earthquake: risk factors and treatment. Tohoku J Exp Med 2020 Jul;251(3):207-216. [doi: 10.1620/tjem.251.207] [Medline: 32669518]
30. Maeda K, Shamoto H, Furuya S. Feeding support team for frail, disabled, or elderly people during the early phase of a disaster. Tohoku J Exp Med 2017 Aug;242(4):259-261. [doi: 10.1620/tjem.242.259] [Medline: 28747593]
31. Yamashita M, Seino S, Nofuji Y, et al. The Kesennuma study in Miyagi, Japan: study design and baseline profiles of participants. J Epidemiol 2022 Dec 5;32(12):559-566 [FREE Full text] [doi: 10.2188/jea.JE20200599] [Medline: 33840651]
32. Survey and research project on physical and mental support and utilization of exercise and sports in the areas affected by the Great East Japan Earthquake [Web page in Japanese]. Japan Health Promotion & Fitness Foundation. 2012. URL: https://www.health-net.or.jp/tyousa/houkoku/pdf/h23_shinsaihoukokusyo.pdf [accessed 2023-11-17]
33. Disaster prevention and reconstruction from a gender equal society perspective—lessons from the Great East Japan Earthquake—(from the “White Paper on Gender Equality 2012”). Gender Equality Bureau Cabinet Office. 2023. URL: https://www.gender.go.jp/english_contents/about_danjo/whitepaper/pdf/2012-01.pdf [accessed 2025-02-26]
34. Survey on the status of response to the earthquake from the perspective of gender equality (July 2012) [Web page in Japanese]. Gender Equality Bureau Cabinet Office. 2012. URL: https://www.gender.go.jp/policy/saigai/pdf/hokokusyo_P17-P22.pdf [accessed 2023-11-17]
35. Orui M, Tanaka E, Maeda M, et al. Mental health and suicide prevention during disasters. Nihon Koshu Eisei Zasshi 2020;67(2):101-110. [doi: 10.11236/jph.67.2_101] [Medline: 32092726]
36. Research group for "Research on cooperation methods between disaster psychiatric assistance team (DPAT) and community mental health systems. Manual for Mental Health and Medical Welfare Activities in Disasters in Local Governments. 2021. URL: https://www.mhlw.go.jp/content/000772550.pdf [accessed 2023-11-17]
37. Sendai framework for disaster risk reduction 2015-2030. PreventionWeb. 2016. URL: https://www.preventionweb.net/files/43291_sendaiframeworkfordrren.pdf [accessed 2023-11-17]
38. Women’s perspective for strengthening disaster response capabilities—disaster prevention and recovery guidelines from a gender equality perspective [Web page in Japanese]. Gender Equality Bureau Cabinet Offic. 2020. URL: https://www.gender.go.jp/policy/saigai/fukkou/pdf/guidelene_01.pdf [accessed 2023-11-17]
39. Women’s perspective for strengthening disaster response capabilities. Gender Equality Bureau Cabinet Office. URL: https://www.gender.go.jp/policy/saigai/fukkou/pdf/pamphlet_en.pdf [accessed 2023-11-17]
40. LunaLuna. URL: https://sp.lnln.jp/brand/services/lnln_services [accessed 2023-11-17]
41. Bliudzius A, Svaikeviciene K, Puronaite R, Kasiulevicius V. Physical activity evaluation using activity trackers for type 2 diabetes prevention in patients with prediabetes. Int J Environ Res Public Health 2022 Jul 6;19(14):8251. [doi: 10.3390/ijerph19148251] [Medline: 35886100]

XSL•FO
RenderX

42. Lima FV, Kadiyala V, Huang A, et al. At the crossroads! time to start taking smartwatches seriously. Am J Cardiol 2022 Sep 15;179:96-101. [doi: 10.1016/j.amjcard.2022.06.020] [Medline: 35842279]
43. Magal N, Rab SL, Goldstein P, Simon L, Jiryis T, Admon R. Predicting chronic stress among healthy females using daily-life physiological and lifestyle features from wearable sensors. Chronic Stress (Thousand Oaks) 2022;6:24705470221100987. [doi: 10.1177/24705470221100987] [Medline: 35911618]
44. Schmidt LI, Jansen CP, Depenbusch J, Gabrian M, Sieverding M, Wahl HW. Using wearables to promote physical activity in old age: feasibility, benefits, and user friendliness. Z Gerontol Geriatr 2022 Aug;55(5):388-393. [doi: 10.1007/s00391-022-02083-x] [Medline: 35849158]
45. Kang HS, Exworthy M. Wearing the future-wearables to empower users to take greater responsibility for their health and care: scoping review. JMIR Mhealth Uhealth 2022 Jul 13;10(7):e35684. [doi: 10.2196/35684] [Medline: 35830222]
46. Hoshide S, Nishizawa M, Okawara Y, et al. Salt intake and risk of disaster hypertension among evacuees in a shelter after the Great East Japan Earthquake. Hypertension 2019 Sep;74(3):564-571. [doi: 10.1161/HYPERTENSIONAHA.119.12943] [Medline: 31280649]

Abbreviations

GEJE: Great East Japan Earthquake

Edited by A Mavragani; submitted 28.01.24; peer-reviewed by A Bodini, J Sands; revised version received 06.01.25; accepted 16.01.25; published 05.03.25.

Please cite as:
Nonaka S, Fujii S, Yamada KC
Purposes of Smartphone Health Care Apps and the Practicality of Their Functions in Disaster Situations: Qualitative Function Assessment Study
JMIR Form Res 2025;9:e56862
URL: https://formative.jmir.org/2025/1/e56862
doi:10.2196/56862

© Sayuri Nonaka, Susumu Fujii, Kosuke Chris Yamada. Originally published in JMIR Formative Research (https://formative.jmir.org), 5.3.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Assessment of the Sensitivity of a Smartphone App to Assist Patients in the Identification of Stroke and Myocardial Infarction: Cross-Sectional Study

Amar Dhand[1,2], MD, DPhil; Rama Mangipudi[3], BPharm, MSc; Anubodh S Varshney[4], MD; Jonathan R Crowe[5,6], MSc, MD, MPH; Andria L Ford[7], MD; Nancy K Sweitzer[8], MD, PhD; Min Shin[9], PhD; Samuel Tate[9], MS; Haissam Haddad[3], MD; Michael E Kelly[10], MD, PhD; James Muller[11], MD; Jay S Shavadia[3], MBChB

[1]Department of Neurology, Brigham and Women's Hospital, Harvard Medical School, 65 Landsdowne Street, Cambridge, MA, United States
[10]Division of Neurosurgery, Department of Surgery, University of Saskatchewan, Saskatoon, SK, Canada
[11]Division of Cardiovascular Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, United States
[2]Network Science Institute, Northeastern University, Boston, MA, United States
[3]Division of Cardiology, Department of Medicine, Unversity of Saskatchewan, Saskatoon, SK, Canada
[4]Division of Cardiovascular Medicine, Department of Medicine, Stanford University, Palo Alto, CA, United States
[5]Department of Neurology, University of Virgina, Charlottesville, VA, United States
[6]Department of Public Health Sciences, University of Virgina, Charlottesville, VA, United States
[7]Department of Neurology, Washington University School of Medicine, St. Louis, MO, United States
[8]Division of Cardiology, Department of Medicine, Washington University School of Medicine, St. Louis, MO, United States
[9]Department of Computer Science, University of North Carolina at Charlotte, Charlotte, NC, United States

Corresponding Author:
Amar Dhand, MD, DPhil
Department of Neurology, Brigham and Women's Hospital, Harvard Medical School, 65 Landsdowne Street, Cambridge, MA, United States

Abstract

Background: Most people do not recognize symptoms of neurological and cardiac emergencies in a timely manner. This leads to delays in hospital arrival and reduced access to therapies that can open arteries. We created a smartphone app to help patients and families evaluate if symptoms may be high risk for stroke or heart attack (myocardial infarction, MI). The ECHAS (Emergency Call for Heart Attack and Stroke) app guides users to assess their risk through evidence-based questions and a test of weakness in one arm by evaluating finger-tapping on the smartphone.

Objective: This study is an initial step in the accuracy evaluation of the app focused on sensitivity. We evaluated whether the app provides appropriate triage advice for patients with known stroke or MI symptoms in the Emergency Department. We designed this study to evaluate the sensitivity of the app, since the most dangerous output of the app would be failure to recognize the need for emergency evaluation. Specificity is also important, but the consequences of low specificity are less dangerous than those of low sensitivity.

Methods: In this single-center cross-sectional study, we enrolled patients presenting with symptoms of possible stroke or MI. The ECHAS app assessment consisted of a series of evidence-based questions regarding symptoms and a test of finger-tapping speed and accuracy on the phone's screen to detect unilateral arm weakness. The primary outcome was the sensitivity of the ECHAS app in detecting the need for ED evaluation. The secondary outcome was the sensitivity of the ECHAS app in detecting the need for hospital admission. Two independent and blinded board-certified physicians reviewed the medical record and adjudicated the appropriateness of the ED visit based on a 5-point score (ground truth). Finally, we asked patients semistructured questions about the app's ease of use, drawbacks, and benefits.

Results: We enrolled 202 patients (57 with stroke and 145 with MI). The ECHAS score was strongly correlated with the ground truth appropriateness score (Spearman correlation 0.41, $P<.001$). The ECHAS app had a sensitivity of 0.98 for identifying patients in whom ED evaluation was appropriate. The app had a sensitivity of 1.0 for identifying patients who were admitted to the hospital because of their ED evaluation. Patients completed an app session in an average of 111 (SD 60) seconds for the stroke pathway and 60 (SD 33) seconds for the MI pathway. Patients reported that the app was easy to use and valuable for personal emergency situations at home.

XSL•FO
RenderX

Conclusions: The ECHAS app demonstrated a high sensitivity for the detection of patients who required emergency evaluation for symptoms of stroke or MI. This study supports the need for a study of specificity of the app, and then a prospective trial of the app in patients at increased risk of MI and stroke.

(JMIR Form Res 2025;9:e60465) doi:10.2196/60465

KEYWORDS

time-to-treatment; digital health; patient-centered care; eHealth apps; smartphones; apps; cross-sectional study; stroke; myocardial infarction; heart attack; neurology; neurological assessment; accuracy; emergency department; finger-tapping test; hospital admission; physicians; medical records; mobile health; mhealth; symptom recognition; mobile phone

Introduction

Timely recognition and intervention in neurological and cardiac emergencies can prevent severe complications and improve survival rates. Prehospital care is emerging as a critical window for deploying innovative strategies to enhance early recognition, triage, and transport in these emergencies [1]. Prehospital care is particularly important in ischemic stroke and myocardial infarction (MI), where the time from symptom onset to treatment profoundly affects prognosis [2,3]. This is particularly true in marginalized populations, where prehospital delay is a key driver of disparities in outcomes [4,5]. Thus, addressing these delays offers broad benefits for public health, economic savings, and equity in health care [6].

The urgency in addressing delays in care for patients with stroke and MI is underscored by a wealth of literature emphasizing the “golden hour”—the initial period when medical intervention is most effective [7,8]. For stroke, thrombolytic or thrombectomy therapy within this critical window reduces the extent of brain damage and improves functional outcomes [9,10]. Similarly, for MI, early reperfusion therapy is crucial in reducing infarct size and mortality [11-13]. However, despite large-scale public health campaigns and education efforts, the majority of patients fail to recognize symptoms as life-threatening, leading to delayed presentations to the emergency department (ED) and missed opportunities for early intervention [2,14,15].

In addition to patient delay, patients often make mistakes in triage decisions. Some never seek the needed emergency evaluation, while others visit an ED for symptoms that do not require emergency evaluation [16]. The initial problem is one of sensitivity and the latter is one of specificity. For screening tests, it is important to first prioritize sensitivity before specificity. This is because not recognizing the need for emergency evaluation could lead to dangerous outcomes (eg, death or morbidity) which is more serious than overdiagnosis. We acknowledge the latter is also important for any decision support tool to be useful before final deployment [17].

The ECHAS (Emergency Call for Heart Attack and Stroke) smartphone app is an innovative approach to bridge this gap by assisting patients in identifying symptoms of neurological and cardiac emergencies and facilitating accurate triage and a quicker intervention (Figure 1). Leveraging the ubiquity of smartphones, ECHAS aims to be a scientifically validated and regulatory-approved digital medicine technology. The app is modeled on the “history and examination” of a neurologist or cardiologist by asking a series of evidence-based questions about the user’s medical history and symptoms, as well as a finger-tapping test designed to detect unilateral weakness. Based on the user or family responses, the app calculates a risk score and recommends one of the following three actions: (1) call 911, (2) call a medical hotline, or (3) call a primary care physician now.

In the first of a series of validation steps, this study evaluated the sensitivity and usability of ECHAS in appropriately triaging patients who have already sought care in the ED for symptoms of a possible stroke, MI, or related conditions.

XSL•FO
RenderX

Figure 1. ECHAS (Emergency Call for Heart Attack and Stroke) flowchart of "history and examination" functions for symptoms concerning cardiac or stroke emergencies.

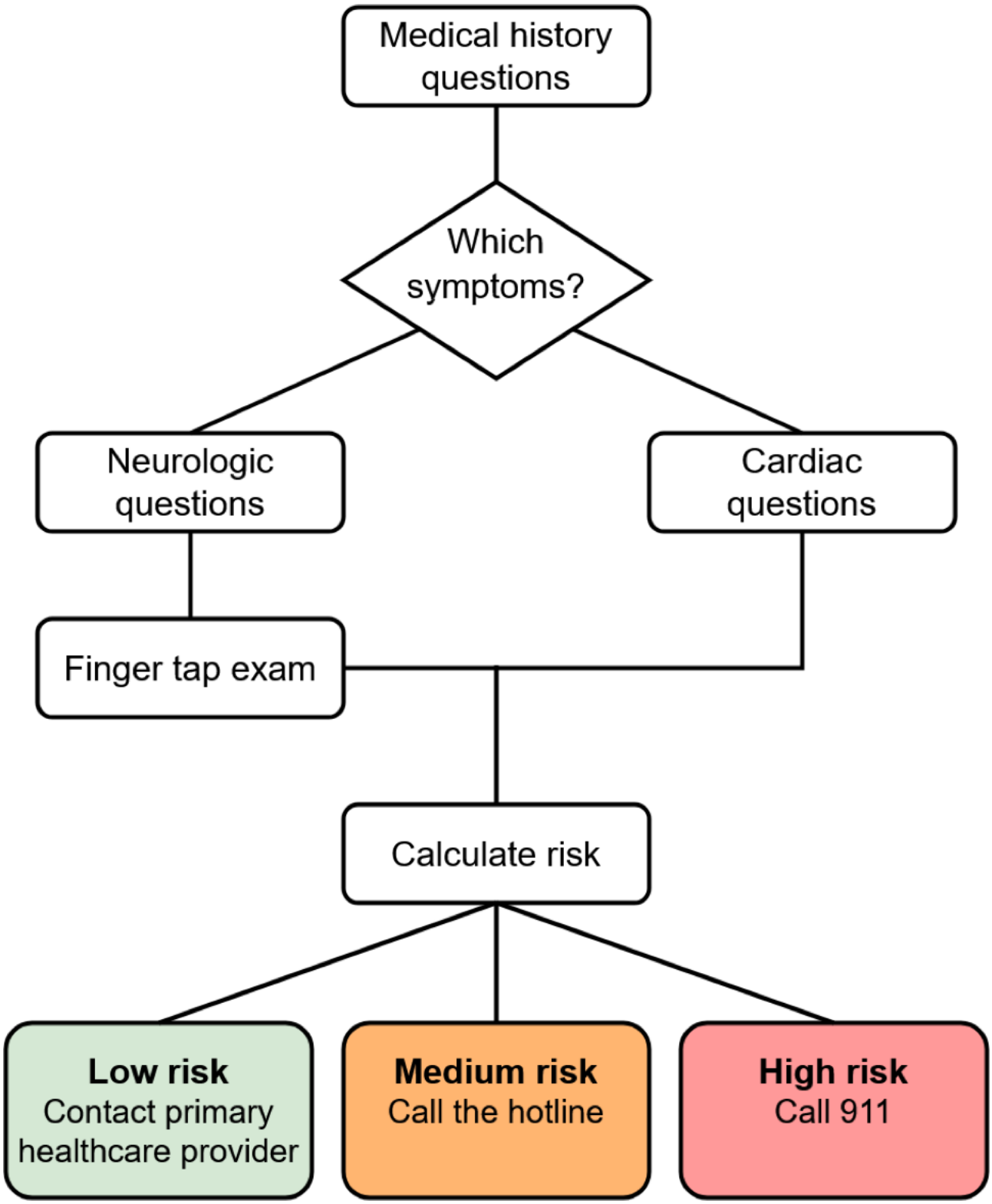

Methods

Study Design and Participants

This was a cross-sectional study at the Royal University Hospital in Saskatoon, Canada between June 2022 and July 2023. We recruited eligible patients in the ED once their initial evaluation was complete, and they were in stable condition.

The inclusion criteria for the study were patients who presented to the ED for evaluation of a possible stroke or MI. These patients were identified for exhibiting symptoms that warranted an "Stroke alert" or "MI alert" as judged by standard clinical protocol. The symptoms indicative of a stroke alert included slurred speech, asymmetric weakness or numbness, balance problems, vision changes, and headache. The symptoms that triggered an MI alert included chest discomfort, chest pressure

XSL•FO
RenderX

or pain, palpitations, shortness of breath, lightheadedness or presyncope, and syncope. In addition, eligible participants were required to be 18 years or older and capable of providing informed consent.

Patients were excluded from eligibility if they were non-English speaking, had a documented history of dementia, had severe visual impairment, were unable to use a smartphone (due to physical, cognitive, or other reasons), or were unable to provide consent. A single trained research staff member enrolled all patients.

Regarding sample size, we determined the sample size a priori based on statistical and practical considerations. We estimated that a sample size of 200 patients was sufficient to provide power to detect small to medium effect sizes, ensuring robust statistical analysis of sensitivity of the app. In addition, this number was deemed feasible for a single-center study within the constraints of the available budget and study period.

The sampling method was consecutive sampling. It involved including all patients who met inclusion criteria sequentially, as they presented within a defined period, until the desired sample size was reached. This minimized selection bias. For recruitment, the research team screened eligible participants through daily chart review. We then collaborated with nurse clinicians in the ED, cardiology, and neurology units to identify potential participants. We also displayed a poster in these nursing units to assist nurses and patients to understand the study. The research team requested the bedside nurse to gauge patient interest. If patient was interested, the research team approached them at the bedside. The research team member introduced the study, obtained informed written consent, and then began collecting data.

Ethical Considerations

We followed a protocol approved by the University of Saskatchewan Research Ethics Board (REB Bio #3380). All patients provided written informed consent (Multimedia Appendix 1) to the study. Regarding privacy, all data collected in the app was deidentified. Specifically, the responses to the questions and the finger-tapping amounts were not connected to the user. For compensation, participants did not receive any financial payment for their involvement in the study. We ensured that no individual participants may be identified in this paper.

ECHAS Description

All eligible patients completed a single session of ECHAS on a study-standardized Apple iPhone 13. Figure 2 displays screenshots of ECHAS. A single session included questions of symptoms and phone sensor–based examination of unilateral weakness for the stroke pathway. Participants were instructed to answer questions about symptoms based on what they experienced before coming to the ED. They answered yes or no questions about these symptoms and their previous history aligned with standard of clinical care (refer to Multimedia Appendix 2 for list of questions). Patients were asked about symptoms occurring during their decision to seek emergency evaluation that had occurred hours before engaging with the app. In many instances, the symptoms that had brought them to the ED were still present at the time of data entry in the app.

XSL•FO
RenderX

Figure 2. Screenshots of ECHAS (Emergency Call for Heart Attack and Stroke) including stroke question, myocardial infarction question, finger-tapping, and final suggestion. We obtained explicit consent from these individuals to use their images for publication.

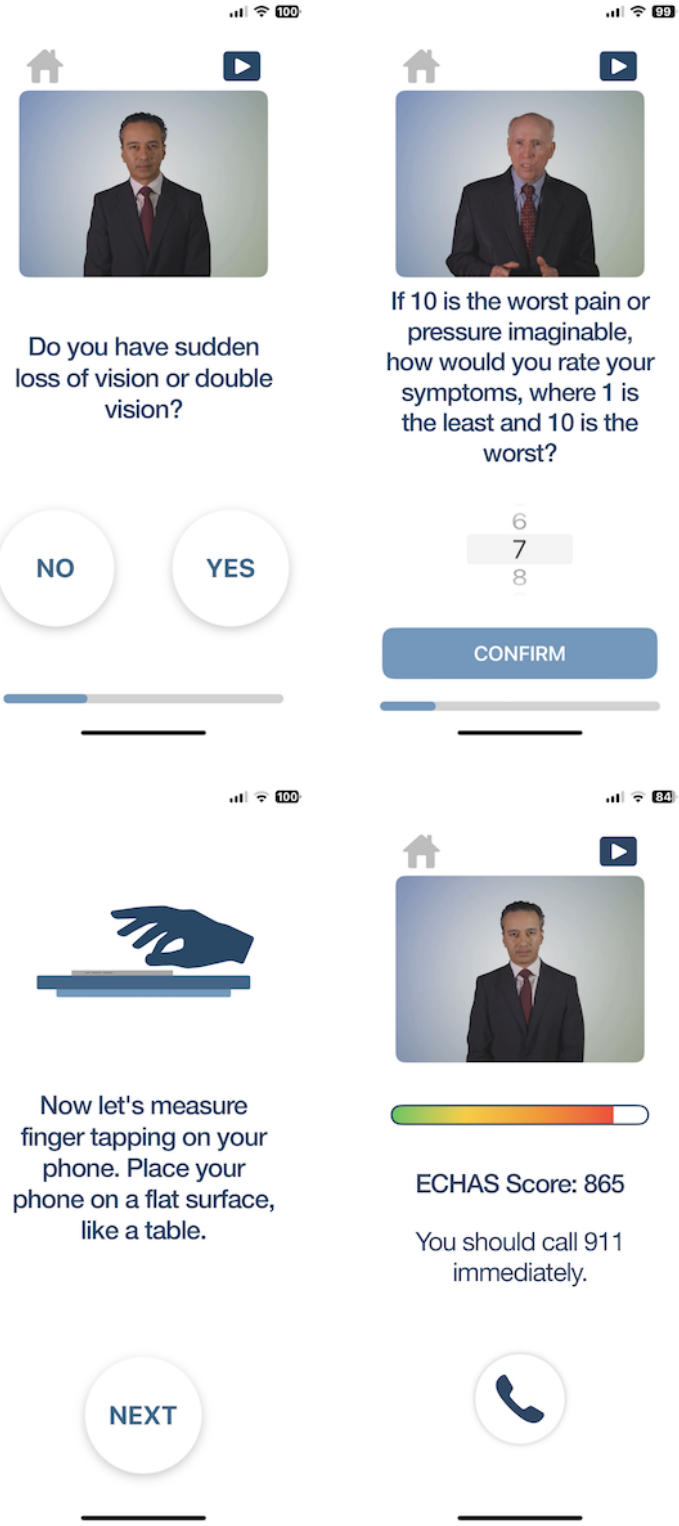

Patients in the stroke pathway additionally completed a phone sensor–based evaluation evaluating finger-tapping on the left and right hands. The app quantified the three following finger-tapping characteristics: (1) speed of finger-tapping for 10 seconds against literature-based standards, (2) comparison of finger-tapping speed on the left versus right hand, and (3) the accuracy of tapping in the middle of the target provided. Participants' responses to questions and their finger-tapping statistics were aggregated into an ECHAS score. The ECHAS score guided one of the following three triage decisions: (1) call 911, (2) call the hot line, or (3) call your primary care physician now.

Clinical End Points

The ground truth for the triage decision assessment was the appropriateness of the patient's visit to the ED as judged by a

XSL•FO
RenderX

retrospective review of the ED record performed by a board-certified neurologist and a board-certified cardiologist using pre-established criteria (Table 1). Each physician was blinded to the ECHAS score or triage decision and to the other adjudicator's decision. A positive (ie, considered reasonably appropriate) ED visit was defined as a patient whose Appropriateness Scale Score was ≥3. A negative (ie, considered inappropriate) ED visit was defined as a patient whose Appropriateness Scale Score was ≤2.

XSL•FO
RenderX

Table . The Appropriateness Scale for emergency department evaluation for potential neurological and cardiac emergency.

Appropriateness Scale score	Definition
Neurological	
1	No neurological imaging or neurology consultation
2	CT[a] head performed
3	Neurology consultation completed
4	Any of the following: • CTA[b] or CTV[c] head and neck performed • MRI[d] brain performed • MRA[e] or MRV[f] head and neck performed • Invasive angiogram performed • Lumbar puncture performed • Admitted to observation unit or hospital admission for noncerebrovascular diagnosis
5	Any of the following: • Diagnosed with stroke, TIA[g], ICH[h], epidural hematoma, SDH[i], SAH[j] • Admitted to hospital for a cerebrovascular diagnosis
Cardiac	
1	No ECG[k] taken nor troponin value measured
2	ECG taken or single troponin value measured
3	ECG taken, ≥2 troponin values measured
4	• ECG taken, ≥2 troponin values measured PLUS: • Coronary CTA performed • TTE[l] performed • Stress test performed • Invasive coronary angiography performed OR (no ECG or troponin required) • Arrhythmia requiring cardioversion • Admission to observation unit or hospital admission for a noncardiovascular diagnosis
5	Any of the following: • Diagnosed with ST-segment elevation MI[m] (STEMI), Non-ST elevation-acute coronary syndrome (NSTE-ACS), acute PE[n], acute aortic dissection, cardiac arrest, or cardiogenic shock • Admitted to hospital for a cardiovascular diagnosis

[a]CT: computed tomography.
[b]CTA: computed tomography angiography.
[c]CTV: computed tomography venography.
[d]MRI: magnetic resonance imaging.
[e]MRA: magnetic resonance angiography.
[f]MRV: magnetic resonance venography.
[g]TIA: transitory ischemic attack.
[h]ICH: intracerebral hemorrhage.
[i]SDH: subdural hematoma.
[j]SAH: subarachnoid hemorrhage.

XSL•FO
RenderX

[k]ECG: electrocardiogram.

[l]TTE: transthoracic echocardiography.

[m]MI: myocardial infarction.

[n]PE: pulmonary embolism.

The primary outcome was the sensitivity of the ECHAS app in detecting patients who needed ED evaluation. We compared the 2 highest acuity triage decisions of the app (call 911 or call the hotline) versus a positive classification (Appropriateness Scale Score ≥3) by the blinded physician adjudicators. We also calculated the sensitivity of ECHAS app to predict admission to the hospital for workup (Appropriateness Scale Score cutoff of 5 as positive and ≤4 as negative).

In the usability study, we determined the time to complete the ECHAS assessment and recorded the feedback of patients on the usability of the app. We recorded patient's feedback both in the form of 2-choice or 5-choice Likert scales to measure opinions and semistructured interview responses.

Statistical Analysis

We first completed descriptive analyses of the ECHAS score and the Appropriateness Scale Score. Then, we assessed their relationship using scatterplots, Spearman's correlation, and unadjusted linear regression. Finally, we created 2×2 tables of ECHAS versus the ground truth and calculated the sensitivity. The study was focused on sensitivity since failure to identify emergency cases is the most dangerous potential outcome of use of the app. The study population in this ED-based context had a high prevalence of serious neurological or cardiac conditions warranting ED evaluation. Specificity will be evaluated in a subsequent study in patients who did not seek ED evaluation.

For secondary outcomes, we calculated the mean and median time that patients took to complete the ECHAS session. We created bar plots of patients' responses on usability. Finally, we completed a qualitative analysis of the participants' interview responses. As the qualitative analysis was ancillary to the quantitative work, we followed a "small q qualitative research" approach [18]. Using single-participant interviews with fully structured, open-ended questions, we elicited brief responses, often single sentences. We coded for patterns to describe and summarize topics and categories [19]. Throughout the process, we engaged in reflexivity by discussing patterns as a group before, during, and after coding. These discussions ensured the confirmability, dependability, credibility, and transferability of our findings.

We used the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) Statement Guideline in reporting results (Checklist 1).

Results

Cohort Characteristics

The cohort examined in this study consisted of 202 enrolled participants, of which 200 had ECHAS scores available. In 2 patients, the scores were not recorded due to technical error. The distribution of patients by clinical presentation included 57 patients with stroke and 145 patients with MI. The overall sex distribution was 76 women and 126 men. The median age of the participants was 62 (IQR 51-67) years. Racial composition of the cohort comprised White (n=179), Asian (n=10), Black or African American (n=1), and Indigenous (n=12). In Table 2, we present the clinical characteristics stratified by illness including baseline medical, ED presentation, stroke, and MI data.

XSL•FO
RenderX

Table . Clinical characteristics of 202 study participants from Royal University Hospital emergency department in 2022 - 2023.

Variables		Presented to ED[a] for symptoms	
		Stroke (n=57)	MI[b] (n=145)
Baseline medical data			
Sex, n (%)			
	Male	25 (44)	102 (70)
	Female	32 (56)	44 (30)
Weight (kg), median (IQR)		86.0 (75.0-97.5)	90.0 (76.0-102.0)
Height (cm), median (IQR)		168.0 (163.0-175.0)	173.5 (165.0-180.3)
Race, n (%)			
	American Indian or Alaska Native	0 (0)	0 (0)
	Asian	0 (0)	10 (7)
	Black or African American	1 (1)	0 (0)
	Native Hawaiian or Other Pacific Islander	0 (0)	0 (0)
	White	53 (93)	127 (87)
	Other	3 (5)	9 (6)
Cigarette smoking, n (%)			
	Current (within last 30 days)	19 (33)	41 (28)
	Never	34 (60)	72 (50)
	Quit ≥1 month	4 (7)	32 (22)
Diabetes mellitus		9 (16)	30 (21)
Diabetes type, n (%)			
	Type 1	1 (11)	2 (7)
	Type 2	8 (89)	28 (93)
Medications taken to control blood sugars		9 (100)	23 (77)
Diabetes drug type, n (%)			
	Oral medication	7 (12)	20 (14)
	GLP-1[c] analogues	1 (2)	2 (1)
	Insulin	3 (5)	10 (7)
Dyslipidemia requiring treatment, n (%)		16 (28)	43 (29)
Hypertension requiring treatment, n (%)		19 (33)	65 (45)
Previous MIs, n (%)		2 (4)	26 (18)
Previous intervention within 30 days, n (%)		0 (0)	2 (1)
Type of intervention, n (%)			
	CABG[d]	0 (0)	0 (0)
	PCI[e] or stenting	0 (0)	2 (1)
	Hear ablation	0 (0)	0 (0)
	Heart valve surgery	0 (0)	0 (0)
Acute coronary syndrome, n (%)		3 (5)	28 (19)
Previous stroke, n (%)		6 (11)	7 (5)
Previous transient ischemic attack (TIA), n (%)		7 (12)	3 (2)

XSL•FO
RenderX

Variables			Presented to ED[a] for symptoms	
			Stroke (n=57)	MI[b] (n=145)
ED presentation data				
	Hours from symptom onset to ED arrival, n (%)			
		Less than 3 hours	24 (43)	58 (40)
		3 to 6 hours	5 (9)	18 (12)
		More than 6 hours	27 (48)	69 (46)
	Time from symptom onset to arrival, median (IQR)		340 (113-1140)	276 (85-2880)
	Mode of transportation to ED, n (%)			
		EMS[f] from home or residence	13 (23)	29 (20)
		EMS from rehab or other health care facility	0 (0)	1 (0.7)
		Personal auto	39 (68)	110 (76)
		Taxi	0 (0)	0 (0)
		Transfer from other hospital	4 (7)	5 (3)
		Other	1 (2)	0 (0)
Stroke data				
	Stroke severity by NIH[g] Stroke Scale, n (%)			
		Mild (0 - 4)	45 (79)	—[h]
		Moderate (5-14)	11 (19)	—
		Severe (15-42)	1 (2)	—
	Finger-tapping assessment, n (%)			
		Both sides are normal speed and equal	39 (75)	—
		Right side is slow compared with left	6 (12)	—
		Left side is slow compared with right	7 (14)	—
		Both sides are slow	0 (0)	—
	CT[i] results, n (%)			
		CT not done	10 (18)	—
		Normal	30 (53)	—
		Abnormal	17 (30)	—
	MRI[j] results, n (%)			
		MRI not done	36 (63)	—
		Normal	9 (16)	—
		Abnormal	12 (21)	—
	Source document to be submitted, n (%)			
		CT Report	46 (81)	—
		MRI Report	20 (35)	—
	TPA given, n (%)		2 (4)	—
	Mechanical thrombectomy completed, n (%)		1 (2)	—
MI data				

XSL•FO
RenderX

Variables		Presented to ED[a] for symptoms	
		Stroke (n=57)	MI[b] (n=145)
Cardiac biomarkers, n (%)			
	Troponin T	—	13 (9)
	Troponin I	—	17 (12)
	High sensitive troponin T	—	130 (89)
	Creatine kinase	—	121 (83)
	Not performed	—	11 (8)
Troponins test 1 result (ng/mL), median (IQR)		—	0.016 (0.007-0.057)
Troponins test 1 upper reference range, median (IQR)		—	0.014 (0.014-0.050)
Troponins test 2 result (ng/mL), median (IQR)		—	0.063 (0.016-0.711)
Troponins test 2 upper reference range, median (IQR)		—	0.014 (0.014-0.050)
Troponins test 3result (ng/mL), median (IQR)		—	0.083 (0.033-1.114)
Troponins test 3 upper reference range, median (IQR)		—	0.014 (0.014-0.050)
Creatine kinase result (U/L), median (IQR)		—	159.0 (94.0-1010.0)
ECG[k] changes, n (%)			
	Normal ECG during the event	—	49 (34)
	Abnormal	—	35 (24)
	ECG showed signs of new MI	—	34 (24)
	ECG showed signs of possible new MI	—	19 (13)
	ECG showed non-specific signs	—	7 (5)
Cardiac catheterization performed, n (%)		—	77 (53)
Coronary imaging performed, n (%)		—	12 (8)
Stress echo ordered, n (%)		—	3 (2)
New cardiac medications started, n (%)		—	80 (55)
Class of new cardiac medicines prescribed, n (%)			
	Antiplatelet	—	76 (52)
	Anticoagulation	—	74 (51)
	Lipid-lowering	—	30 (21)
	Antihypertensive	—	24 (16)

[a]ED: emergency department.
[b]MI: myocardial infarction.
[c]GLP-1: Glucagon-like peptide-1.
[d]CABG: coronary artery bypass graft.
[e]PCI: percutaneous coronary intervention.
[f]EMS: emergency medical services.
[g]NIH: National Institutes of Health.
[h]Not applicable.
[i]CT: computed tomography.
[j]MRI: magnetic resonance imaging.
[k]ECG: electrocardiogram.

XSL•FO
RenderX

ECHAS Score Versus Appropriateness Scale Scores

The distribution of scores for the ECHAS score and the Appropriateness Scale were similar, as illustrated in Figure 3. This similarity suggests alignment in the measurement of clinical outcomes by these 2 scales.

Figure 3. Distribution of the ECHAS (Emergency Call for Heart Attack and Stroke) score and appropriateness scale score (ground truth) with the median indicated by the blue line in 202 study participants.

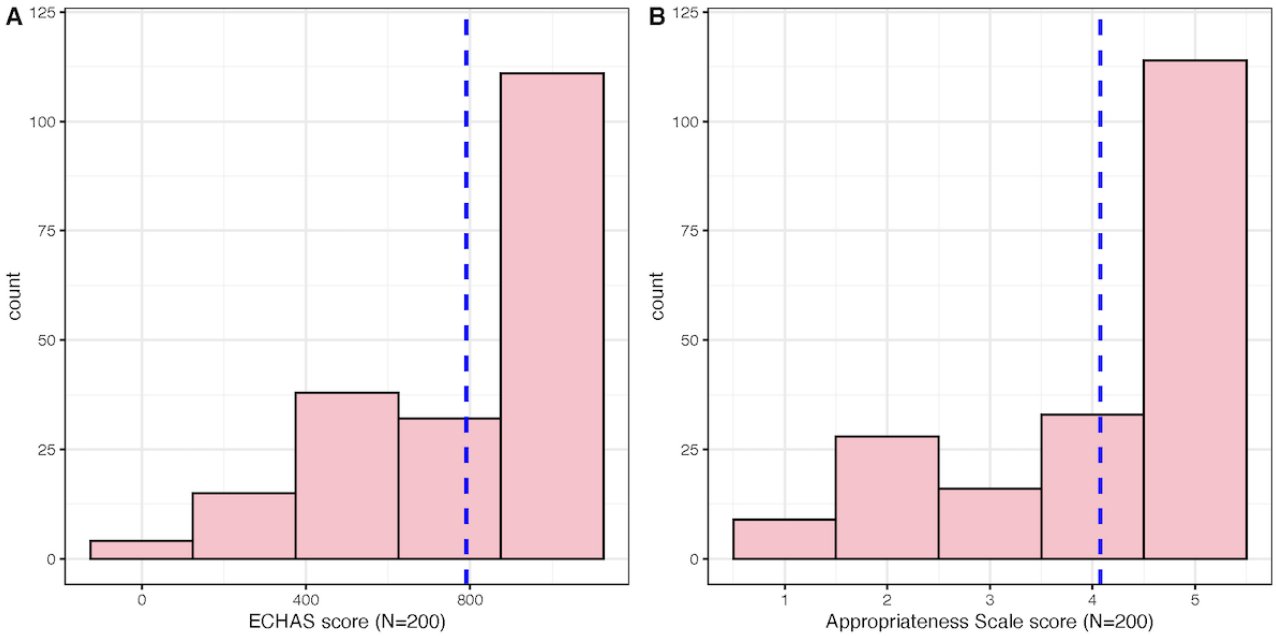

We observed a significant correlation between the ECHAS score and the Appropriateness Scale Score, with a Spearman correlation coefficient 0.41 (P<.001). This pattern of correlation was consistent when stratified for patients in both the MI and stroke pathways. This association was confirmed in a simple linear regression, where the ECHAS β-coefficient was determined to be 0.002 (SE=0.003, P<.01), with a constant of 2.557.

The agreement of ECHAS scores with the ground truth, as adjudicated by 2 independent physicians, is presented in Table 3. This 2×2 table shows that the physicians adjudicated the cases had 97% agreement, suggesting low interobserver variability.

Table . Two-by-two tables of the ECHAS (Emergency Call for Heart Attack and Stroke) app triage advice versus adjudicators' determination of appropriateness of emergency department visit.

ECHAS app		Positive	Negative	Sensitivity	Specificity
Adjudicator 1				0.98	0.08
	Positive	159	34		
	Negative	4	3		
Adjudicator 2				0.97	0.07
	Positive	153	40		
	Negative	4	3		

The average sensitivity of ECHAS for appropriate triage was 0.98. The sensitivity for identifying patients admitted to the hospital for possible stroke or MI was 1.0, indicating that ECHAS detected all patients who were highest acuity. The average specificity of ECHAS for appropriate triage was 0.08.

Secondary Outcomes

The average time to complete the assessment using ECHAS varied by 111 (SD 60) seconds for stroke and 60 (SD 33) seconds for MI.

Patients reported finding ECHAS usable, understandable, and worthwhile in acute emergencies at home, as shown in Figure 4. Specific comments from patients, which provide insights into their experiences and perceptions of the app, are detailed in Textbox 1.

XSL•FO
RenderX

Figure 4. Usability of the ECHAS (Emergency Call for Heart Attack and Stroke) app according to study participants.

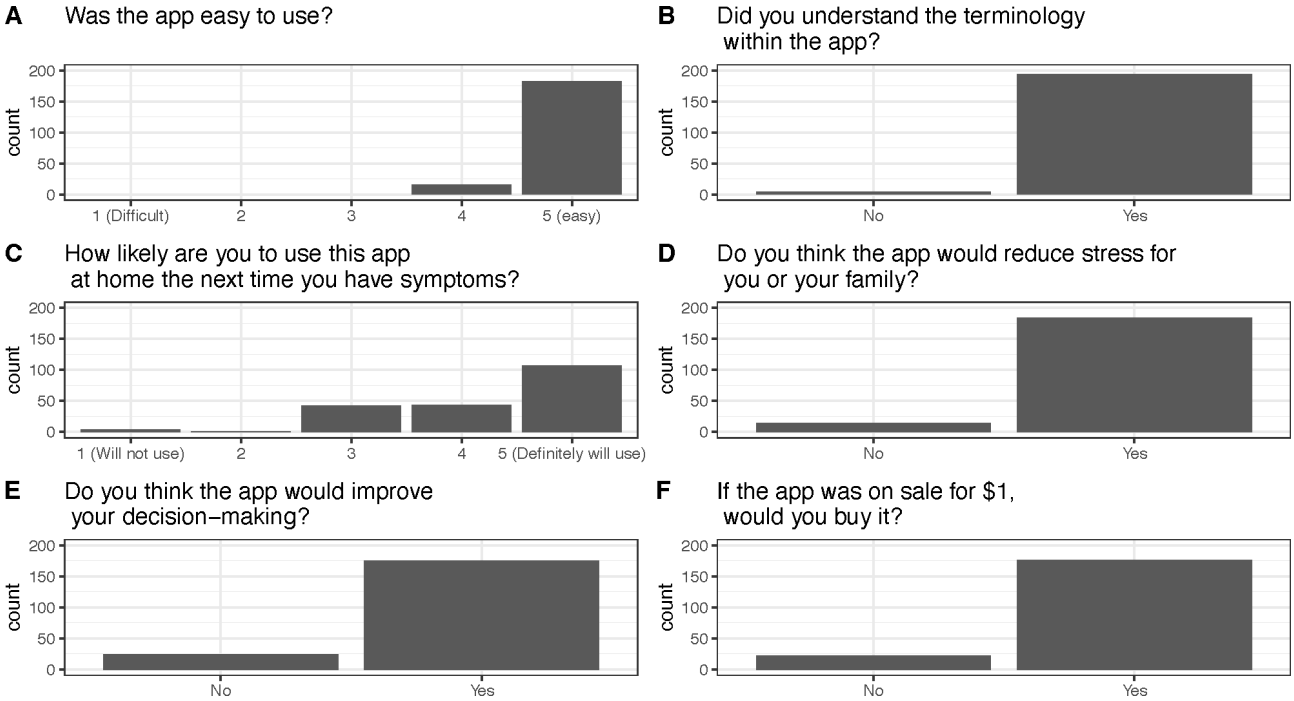

Textbox 1. Patients' experiences and perceptions of the ECHAS (Emergency Call for Heart Attack and Stroke) after use in the Royal University Hospital emergency department.

App is helpful, useful, and positive

- *Good app. Very useful for delayers. Could save lives.*
- *I wish I had this app in my mobile 2 days back.*
- *The next time I experience any symptoms that I don't understand, I would use the app to help me make the correct decision.*
- *I would inform my colleagues that this app is available. I would use it myself if I was having chest pain again.*
- *Consider if vitals measurement could be done by the smart phone.*
- *If I have symptoms again I would use again.*

App as important within family networks

- *I would definitely buy the app for my family members.*
- *If the app could get the summary of all the questions answered at the end (like a PDF copy)- may be beneficial for EMS/ER to look at the responses to reduce the time for reevaluation or asking same questions.*
- *I would use the app whenever I thinks my father is having a possible heart attack or stroke.* [The patient's daughter]

Suggestions for improvement:

- *Don't know what to select. I have both severe headache and chest pain. Thought could be either of stroke or heart attack.*
- *Chest pain felt like a prior heart attack- so called 911. I would call 911 irrespective of the app.*
- *Don't remember my blood pressure.*
- *Not sure if I am having heart attack or stroke which one to select to begin.*

Discussion

Principal Results

In this initial validation study, we evaluated the sensitivity and usability of the ECHAS smartphone app in 202 patients presenting to an ED for symptoms of a possible stroke or MI. We found a significant correlation between ECHAS scores and the Appropriateness Scale scores, indicating the app was successful in identifying patients who should seek ED evaluation. We found that ECHAS demonstrated high sensitivity of 0.98, especially in identifying high-acuity stroke or MI patients, with a unanimous detection rate for those patients admitted. In addition, the app proved to be efficient, requiring on average 111 (SD 60) seconds for stroke, and 60 (SD 33) seconds for MI. Patient feedback highlighted the potential

XSL•FO
RenderX

usability and effectiveness of ECHAS in emergency situations at home.

These findings support continued development of the app which requires study of specificity and a prospective study. The specificity in this study of patients who were already on the MI or Stroke Alert pathway in the emergency room was low (0.08). This means that ECHAS did not detect patients who did not need more comprehensive cardiac workup (ECG taken, ≥2 troponin values measured or more advanced studies) or neurological workup (neurology consultation and advanced imaging). The low specificity could mean that the app would over-recommend ED visits leading to false positives. However, it is unknown how ECHAS would perform in otherwise healthy persons at home. This needs further prospective study. In this study, we deliberately prioritized sensitivity because the consequences of low sensitivity are more dangerous than those of low specificity.

Comparison With Previous Work

The development of ECHAS is informed by a communication and health behavior theory suggesting that mobile health technologies address constraints of space and time that hinder more traditional forms of mass media communication [20]. Specifically, mobile health tools have shown promise to enhance self-management, adherence to treatment, and health literacy among patients with chronic conditions [21]. Furthermore, certain mobile health solutions provide an opportunity for real-time monitoring and assessment, which is particularly valuable in emergency scenarios where every second counts [22]. The ECHAS study contributes to these trends in the literature by demonstrating initial accuracy of a potentially ubiquitous tool that combines a physician-informed assessment and sensor-based monitoring to empower patients and families to determine actions during a worrisome moment.

The ECHAS study adds to the evolving landscape of digital medicine technologies aimed at enhancing acute symptom detection and emergency medical system alerting [23,24]. For example, Wasselius et al [25] showed the potential of wearable accelerometers to detect unilateral arm paresis. In the Apple Heart Study, Perez et al [26] showed that a wearable algorithm could correctly identify atrial fibrillation in users who were notified of irregular pulses. Garcia et al [27] introduced a mobile app that detects key symptoms of stroke, such as smile asymmetry, speech difficulties, and arm weakness, showing effective early detection capabilities. Similarly Stroke Cognitive Medical Assistant by Khriyenko et al [28] used IBM Watson tools to help users detect face asymmetry, speech difficulties, and arm weakness. These advancements collectively underscore a significant shift toward integrating technology in the fight against acute medical emergencies.

Study Strengths

This study's strengths included a large sample size of patients in acute settings, the use of 2 board-certified and blinded adjudicators for end point determination, and the evaluation of ECHAS against the appropriateness of ED visits rather than just confirmed stroke or MI diagnoses. In addition, this collaborative effort between neurologists and cardiologists addresses the needs of patients with vascular risk factors who are at risk for both neurological and cardiac emergencies. Finally, these positive results were achieved with a sensor that only detects finger-tapping weakness. Future versions of ECHAS, being tested at the time of this study, include sensors for facial droop detection and slurred or incomprehensible speech detection. This should improve specificity. Together with finger-tapping weakness, ECHAS future versions would allow objective detection of face, arm, and speech abnormalities (FAST) on a smartphone.

Limitations

As an initial investigation of ECHAS, this study had limitations. Its cross-sectional design may introduce recall bias, affecting the accuracy of reported outcomes. In addition, the performance of the app, tested in a controlled clinical environment on a standardized phone, might not mirror its usability or accuracy in real-world, at-home scenarios during worrisome moments. The study was conducted in a single center with limited diversity in terms of demographics and disabilities. The version of ECHAS evaluated in this study included sensors limited to detecting unilateral weakness through finger tapping. As stated above, ECHAS versions now being tested also include sensors for detecting facial asymmetry and speech difficulties. The study's cohort composition limited the ability to assess specificity, which is the app's ability to correctly identify those without the condition. This is an important metric for avoiding false positives that requires further study. We are planning future prospective studies involving at-home app deployment to address these limitations and provide a clearer understanding of ECHAS' accuracy, specificity, and usability in real-life settings with lower risk cohorts. We also aim to study ECHAS across diverse populations to test accessibility and usability in a wide range of users.

Conclusions

In this study of 202 patients, we found that ECHAS was highly sensitive and easily usable in guiding patients to identify medical emergencies without input by health care personnel. The findings support further research to determine specificity and a study of the app's prospective use in home settings. Overall, this "history and examination" app adds to the burgeoning array of digital medicine tools designed to improve prehospital care, highlighting its potential to empower patients and enhance early intervention strategies.

Acknowledgments

This study was funded by ECHAS, LLC. The cofounders of ECHAS, LLC played an active role in the design of the study, analysis, interpretation of data, and writing of the manuscript. We acknowledge the support of Seed Investors, particularly JanaLee Cherneski and Ian Desai, who enabled this study.

XSL•FO
RenderX

Data Availability

The data sets generated and/or analyzed during this study are not publicly available due to their sensitive medical nature protected by HIPAA, but are available from the corresponding author on reasonable request.

The underlying code for ECHAS is not publicly available for proprietary reasons. We understand that this restricts replication and transparency. Therefore, we will make the code for ECHAS used in this study available to qualified researcher who provide a reasonable request to the corresponding author.

Authors' Contributions

AD analyzed and interpreted the data and wrote the first draft of the manuscript. RM and JSS collected the data and were major contributors to the manuscript writing. ASV and JC adjudicated all data, adjudicated all cases to determine the Appropriateness Score, and were major contributors to the manuscript writing. ASV, JC, AF, NKS, MS, ST, HH, MEK, JM interpreted the data and provided substantial revisions to the data. ST, MS, and AD developed the smartphone app. All authors read and approved the final manuscript.

Conflicts of Interest

Authors AD and JM are co-founders of ECHAS, LLC. AD, JM, and MS hold shares in ECHAS, LLC and declare a financial interest. Authors ASV, JC, and ST have acted as paid consultants for ECHAS, LLC and declare a financial interest. All other authors declare no financial or non-financial competing interests.

Multimedia Appendix 1
Informed consent form.
[PDF File, 209 KB - formative_v9i1e60465_app1.pdf]

Multimedia Appendix 2
List of ECHAS questions.
[PDF File, 72 KB - formative_v9i1e60465_app2.pdf]

Checklist 1
STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) checklist.
[PDF File, 154 KB - formative_v9i1e60465_app3.pdf]

References

1. Fladt J, Ospel JM, Singh N, Saver JL, Fisher M, Goyal M. Optimizing patient-centered stroke care and research in the prehospital setting. Stroke 2023 Sep;54(9):2453-2460. [doi: 10.1161/STROKEAHA.123.044169] [Medline: 37548010]
2. Hoschar S, Albarqouni L, Ladwig KH. A systematic review of educational interventions aiming to reduce prehospital delay in patients with acute coronary syndrome. Open Heart 2020;7(1):e001175. [doi: 10.1136/openhrt-2019-001175] [Medline: 32201586]
3. Matsuo R, Yamaguchi Y, Matsushita T, et al. Association between onset-to-door time and clinical outcomes after ischemic stroke. Stroke 2017 Nov;48(11):3049-3056. [doi: 10.1161/STROKEAHA.117.018132] [Medline: 28974631]
4. Powell-Wiley TM, Baumer Y, Baah FO, et al. Social determinants of cardiovascular disease. Circ Res 2022 Mar 4;130(5):782-799. [doi: 10.1161/CIRCRESAHA.121.319811] [Medline: 35239404]
5. Cruz-Flores S, Rabinstein A, Biller J, et al. Racial-ethnic disparities in stroke care: the American experience: a statement for healthcare professionals from the American Heart Association/American Stroke Association. Stroke 2011 Jul;42(7):2091-2116. [doi: 10.1161/STR.0b013e3182213e24] [Medline: 21617147]
6. National Institutes of Health, National Institute of Neurological Disorders and Stroke. Inequities in Access and Delivery of Acute Stroke Care.: National Institutes of Health; 2022. URL: https://www.ninds.nih.gov/sites/default/files/documents/Inequities-in-Access-and-Delivery-of-Acute-Stroke-Care-A-Brain-Attack-Coalition-Report_508C.pdf [accessed 2025-01-23]
7. Hand M, Brown C, Horan M, Simons-Morton D. Access to timely and optimal care of patients with acute coronary syndromes — community planning considerations: a report by the National Heart Attack Alert Program. J Thromb Thrombolysis 1998;6(1):19-46. [doi: 10.1023/A:1008820104852]
8. Randhawa AS, Pariona-Vargas F, Starkman S, et al. Beyond the golden hour: treating acute stroke in the Platinum 30 minutes. Stroke 2022 Aug;53(8):2426-2434. [doi: 10.1161/STROKEAHA.121.036993] [Medline: 35545939]
9. Saver JL, Fonarow GC, Smith EE, et al. Time to treatment with intravenous tissue plasminogen activator and outcome from acute ischemic stroke. JAMA 2013 Jun 19;309(23):2480-2488. [doi: 10.1001/jama.2013.6959] [Medline: 23780461]
10. Saver JL, Goyal M, van der Lugt A, et al. Time to treatment with endovascular thrombectomy and outcomes from ischemic stroke: a meta-analysis. JAMA 2016 Sep 27;316(12):1279-1288. [doi: 10.1001/jama.2016.13647] [Medline: 27673305]

XSL•FO
RenderX

11. Moser DK, Kimble LP, Alberts MJ, et al. Reducing delay in seeking treatment by patients with acute coronary syndrome and stroke: a scientific statement from the American Heart Association Council on cardiovascular nursing and stroke council. Circulation 2006 Jul 11;114(2):168-182. [doi: 10.1161/CIRCULATIONAHA.106.176040] [Medline: 16801458]
12. Nallamothu BK, Normand SL, Wang Y, et al. Relation between door-to-balloon times and mortality after primary percutaneous coronary intervention over time: a retrospective study. Lancet 2015 Mar;385(9973):1114-1122. [doi: 10.1016/S0140-6736(14)61932-2]
13. Jollis JG, Granger CB, Zègre-Hemsey JK, et al. Treatment time and in-hospital mortality among patients with ST-segment elevation myocardial infarction, 2018-2021. JAMA 2022 Nov 22;328(20):2033-2040. [doi: 10.1001/jama.2022.20149] [Medline: 36335474]
14. Yoon CW, Oh H, Lee J, et al. Comparisons of prehospital delay and related factors between acute ischemic stroke and acute myocardial infarction. J Am Heart Assoc 2022 May 3;11(9):e023214. [doi: 10.1161/JAHA.121.023214] [Medline: 35491981]
15. Saczynski JS, Yarzebski J, Lessard D, et al. Trends in prehospital delay in patients with acute myocardial infarction (from the Worcester Heart Attack Study). Am J Cardiol 2008 Dec 15;102(12):1589-1594. [doi: 10.1016/j.amjcard.2008.07.056] [Medline: 19064010]
16. Newton EH. Addressing overuse in emergency medicine: evidence of a role for greater patient engagement. Clin Exp Emerg Med 2017 Dec;4(4):189-200. [doi: 10.15441/ceem.17.233] [Medline: 29306268]
17. Power M, Fell G, Wright M. Principles for high-quality, high-value testing. Evid Based Med 2013 Feb;18(1):5-10. [doi: 10.1136/eb-2012-100645] [Medline: 22740357]
18. Braun V, Clarke V. Toward good practice in thematic analysis: avoiding common problems and be(com)ing a knowing researcher. Int J Transgend Health 2023;24(1):1-6. [doi: 10.1080/26895269.2022.2129597] [Medline: 36713144]
19. Saldana J. The Coding Manual for Qualitative Researchers, 4th edition: SAGE Publications; 2021.
20. Glanz K, Rimer BK. Health Behavior: Theory, Research, and Practice, 5th edition: Jossey-Bass; 2015.
21. Hamine S, Gerth-Guyette E, Faulx D, Green BB, Ginsburg AS. Impact of mHealth chronic disease management on treatment adherence and patient outcomes: a systematic review. J Med Internet Res 2015 Feb 24;17(2):e52. [doi: 10.2196/jmir.3951] [Medline: 25803266]
22. Albahri OS, Albahri AS, Mohammed KI, et al. Systematic review of real-time remote health monitoring system in triage and priority-based sensor technology: taxonomy, open challenges, motivation and recommendations. J Med Syst 2018 Mar 22;42(5):80. [doi: 10.1007/s10916-018-0943-4] [Medline: 29564649]
23. Bat-Erdene BO, Saver JL. Automatic acute stroke symptom detection and emergency medical systems alerting by mobile health technologies: a review. J Stroke Cerebrovasc Dis 2021 Jul;30(7):105826. [doi: 10.1016/j.jstrokecerebrovasdis.2021.105826] [Medline: 33932749]
24. Varshney AS, Madias C, Kakkar R, Martin DT. Watching for disease: the changing paradigm of disease screening in the age of consumer health devices. J Gen Intern Med 2020 Jul;35(7):2173-2175. [doi: 10.1007/s11606-019-05626-y] [Medline: 31942673]
25. Wasselius J, Lyckegård Finn E, Persson E, et al. Detection of unilateral arm paresis after stroke by wearable accelerometers and machine learning. Sensors (Basel) 2021 Nov 23;21(23):7784. [doi: 10.3390/s21237784] [Medline: 34883800]
26. Perez MV, Mahaffey KW, Hedlin H, et al. Large-scale assessment of a smartwatch to identify atrial fibrillation. N Engl J Med 2019 Nov 14;381(20):1909-1917. [doi: 10.1056/NEJMoa1901183] [Medline: 31722151]
27. García L, Tomás J, Parra L, Lloret J. An m-health application for cerebral stroke detection and monitoring using cloud services. Int J Inf Manage 2019 Apr;45:319-327. [doi: 10.1016/j.ijinfomgt.2018.06.004]
28. Khriyenko O, Rönkkö K, Tsybulko V, Piik K. Stroke Cognitive Medical Assistant (StrokeCMA). GSTF J Comput 2018;6(1). [doi: 10.5176/2251-3043_6.1.112]

Abbreviations

ECHAS: Emergency Call for Heart Attack and Stroke
ED: emergency department
MI: myocardial infarction
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

XSL•FO
RenderX

Edited by A Mavragani; submitted 13.05.24; peer-reviewed by D Patel; revised version received 22.01.25; accepted 23.01.25; published 03.03.25.

Please cite as:
Dhand A, Mangipudi R, Varshney AS, Crowe JR, Ford AL, Sweitzer NK, Shin M, Tate S, Haddad H, Kelly ME, Muller J, Shavadia JS
Assessment of the Sensitivity of a Smartphone App to Assist Patients in the Identification of Stroke and Myocardial Infarction: Cross-Sectional Study
JMIR Form Res 2025;9:e60465
URL: https://formative.jmir.org/2025/1/e60465
doi:10.2196/60465

© Amar Dhand, Rama Mangipudi, Anubodh S Varshney, Jonathan R Crowe, Andria L Ford, Nancy K Sweitzer, Min Shin, Samuel Tate, Haissam Haddad, Michael E Kelly, James Muller, Jay S Shavadia. Originally published in JMIR Formative Research (https://formative.jmir.org), 3.3.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Text Messaging Interventions for Unhealthy Alcohol Use in Emergency Departments: Mixed Methods Assessment of Implementation Barriers and Facilitators

Megan O'Grady[1], PhD; Laura Harrison[2], MPH; Adekemi Suleiman[1], MBBS, MPH; Morica Hutchison[1], PhD, LMFT; Nancy Kwon[2,3], MD; Frederick Muench[4], PhD; Sandeep Kapoor[2,3], MD

[1]Department of Public Health Sciences, University of Connecticut School of Medicine, 263 Farmington Avenue, Farmington, CT, United States

[2]Northwell Health, New Hyde Park, NY, United States

[3]Zucker School of Medicine at Hofstra/Northwell, Hempstead, NY, United States

[4]Partnership to End Addiction, New York, NY, United States

Corresponding Author:
Megan O'Grady, PhD
Department of Public Health Sciences, University of Connecticut School of Medicine, 263 Farmington Avenue, Farmington, CT, United States

Abstract

Background: Many patients with unhealthy alcohol use (UAU) access health care in emergency departments (EDs). Scalable supports, such as SMS text messaging interventions, are acceptable and feasible to enhance care delivery for many health issues, including substance use. Further, SMS text messaging interventions have been shown to improve patient outcomes related to alcohol consumption (eg, reduced consumption compared to no intervention, basic health information, or drink tracking), but they are rarely offered in clinical settings.

Objective: This paper describes a mixed methods study using the Integrated Promoting Action on Research Implementation in Health Services (i-PARIHS) framework. The goal of this study was to use a stakeholder-engaged mixed methods design to assess barriers and facilitators to the implementation of SMS text messaging interventions for UAU in EDs with a focus on the recipient's characteristics, the innovation's degree of fit within the existing practice, and the unique nature of the inner and outer context.

Methods: This study was conducted in a large health system in the northeastern United States. We examined electronic health record data on alcohol screening in 17 EDs; surveyed 26 ED physician chairpersons on implementation feasibility, acceptability, and appropriateness; and interviewed 18 ED staff and 21 patients to understand barriers and facilitators to implementation. Interviews were analyzed according to the i-PARIHS framework to assess recipient characteristics, innovation degree of fit, and inner and outer context.

Results: Electronic health record data revealed high variability in alcohol screening completion (mean 73%, range 35% - 93%), indicating potential issues in identifying patients eligible to offer the intervention. The 26 ED chair surveys revealed a relatively high level of implementation confidence (mean 4, SD 0.81), acceptability (mean 4, SD 0.71), and appropriateness (mean 3.75, SD 0.69) regarding the UAU SMS text messaging intervention; feasibility (mean 3.5, SD 0.55) had the lowest mean, indicating concerns about integrating the text intervention in the busy ED workflow. Staff were concerned about staff buy-in and adding additional discussion points to already overwhelmed patients during their ED visit but saw the need for additional low-threshold services for UAU. Patients were interested in the intervention to address drinking and health-related goals.

Conclusions: ED visits involving UAU have increased in the United States. The results of this formative study on barriers and facilitators to the implementation of UAU SMS text messaging interventions in EDs indicate both promise and caution. In general, we found that staff viewed offering such interventions as appropriate and acceptable; however, there were concerns with feasibility (eg, low alcohol risk screening rates). Patients also generally viewed the SMS text messaging intervention positively, with limited drawbacks (eg, slight concerns about having time to read messages). The results provide information that can be used to develop implementation strategies that can be tested in future studies.

Trial Registration: ClinicalTrials.gov NCT05350878; http://clinicaltrials.gov/ct2/show/NCT05350878

(JMIR Form Res 2025;9:e65187) doi:10.2196/65187

KEYWORDS

unhealthy alcohol use; text messaging intervention; emergency department; barriers and facilitators; implementation; alcohol use; unhealthy; mixed methods assessment; mixed methods; assessment; facilitators; alcohol; health care; scalable supports; text

XSL•FO
RenderX

messaging; patient outcomes; health system; electronic health record; EHR; alcohol screening; acceptability; feasibility; survey; health-related goals

Introduction

Unhealthy alcohol use (UAU) costs the United States nearly US $250 billion per year [1], and alcohol-related deaths among those aged 16 years and older recently increased by 25% [2]. Emergency departments (EDs), where alcohol-related visits are rising [3-5], are sometimes the only health care touchpoint for patients with UAU, making it a promising point of intervention [6,7]. SMS text messages are one of the most ubiquitous and salient modes of mobile health interventions. They use brief, supportive messages to promote behavior change without significant user burden [8-10] and are acceptable and feasible to enhance care delivery for many health issues, ranging from high blood pressure to substance use [9,11-16]. A meta-analysis and study indicated that alcohol-targeted texting interventions reduced alcohol consumption compared with no intervention, basic health information, or drink tracking [17,18]. Notably, UAU text interventions for ED and trauma patients are acceptable and effective in reducing drinking [19-22]. However, there is limited uptake of text-messaging interventions in ED settings.

Reviews of substance use technology interventions have noted minimal focus on implementation strategies or outcomes; rather, most studies focus only on patient clinical outcomes [23,24]. Thus, prioritizing the examination of barriers and facilitators to understand ways to increase uptake is an important next step [25]. This paper describes a mixed methods study using the Integrated Promoting Action on Research Implementation in Health Services (i-PARIHS) framework, which posits optimal implementation occurs when interactive problem-solving promotes acceptance and use of a new innovation based on the recipient's characteristics (eg, motivation), the innovation's degree of fit within the existing practice (eg, usability), and the unique nature of the inner (eg, culture) and outer context (eg, mandates) [26,27]. The examination of barriers and facilitators to implementing UAU text interventions in EDs [26,28] was done by examining electronic health records (EHRs), surveys, and interview data from a large health system using i-PARIHS elements known to influence the uptake of new innovations.

Methods

Study Population

The study was conducted in 17 adult-serving EDs in the northeastern US urban and suburban locations; it is part of a larger project examining SMS text messaging interventions for UAU [29].

Ethical Considerations

This study was approved by the institutional review board at the Feinstein Institutes for Medical Research (approval number 21-0756). All participants consented to study procedures using verbal consent processes and consent information sheets. All qualitative data were deidentified prior to analysis; survey data were anonymous. Patients received a US $15 gift card incentive; staff participants were not compensated. Generative artificial intelligence was not used in any part of the manuscript writing.

Procedures and Materials

We used an explanatory, sequential mixed methods approach (quantitative → qualitative) [30]. Data were integrated at several points. Quantitative EHR and survey data classified sites on low or high implementation potential; classifications were used for data collection and site selection for qualitative interviews and analysis. Data were integrated during result interpretation. Reporting is in accordance with the COREQ (Consolidated Criteria for Reporting Qualitative Research) and CHERRIES (Checklist for Reporting Results of Internet e-Surveys) checklists for qualitative health research involving interviews and internet e-surveys.

EHR Data

The 3-item AUDIT-C (Alcohol Use Disorders Identification Test-Consumption) [31] detects UAU (≥3 for women and ≥4 for men) and was built into all participating ED EHRs; it provided information on each ED's capacity to address UAU (eg, low screening completion rates may indicate a lack of time). Data, extracted from 2021, included percentage of patients screened (number of AUDIT-C completions divided by the total patient census).

Staff Surveys

ED physician chairpersons from the 17 participating EDs received an email from the study team inviting them to complete a 13-item web-based survey of three validated 4-item scales assessing conceptually distinct, but interrelated constructs often used in formative research to examine implementation barriers: feasibility (eg, "A text messaging intervention for UAU seems possible in my ED"), acceptability (eg, "I would welcome our clinical team offering a text messaging intervention for UAU to patients during their ED visit"), and appropriateness (eg, "A text messaging intervention for UAU seems fitting for patients in my ED") [32]. An additional confidence item ("I am confident that our clinical ED team would offer a text messaging intervention for UAU to patients during their ED visit") was included. All items were rated on a 5-point scale (1=completely disagree to 5=completely agree). To complete site classification, we calculated each site's survey totals (range 4 - 65; higher values indicate positive implementation outcomes). All invited EDs participated; in cases where both primary and secondary chairs completed the survey, scores were averaged.

Implementation Potential Classification

We classified low versus high implementation potential based on AUDIT-C completion rates and survey scores (low: ≤median on both measures; high: >median on both measures), resulting in 5 low sites and 4 high. Sites with mixed classifications were not recruited. We considered several factors (eg, census and location) to select the 4 final EDs for interviews (2 low and 2 high) using stratified purposeful sampling [33,34].

XSL•FO
RenderX

Staff Interviews

In the 4 final EDs, chairpersons were interviewed and then each identified staff. They were invited via emails from study staff, resulting in 18 (82%) out of 22 invited interviewies. Interviews lasted ~30 minutes, were conducted over Zoom, recorded, and cleaned for analysis. Interview guides covered views on alcohol screening, SMS text messaging interventions and how to best offer them, barriers and facilitators to intervention implementation, and general practice change.

Patient Interviews

Patients were recruited from 6 EDs; sites were selected based on location and census to gain perspectives from different populations and ED sizes (eg, race or ethnicity and urban or suburban). Patients were approached during their ED visit if they screened positive on the AUDIT-C. Of the 30 patients invited, 21 (70%) were interviewed. Interviews lasted ~20 minutes and were recorded and transcribed. The interview guide was adapted from previous ED studies (eg, mobile device use and interest in receiving supportive messages on UAU reduction) [35,36].

Qualitative Analysis

We used rapid qualitative analysis [37] to analyze interview data following procedures previously outlined [37,38]. Two researchers (MO and AS) summarized 3 transcripts from patient and staff interviews to ensure consistent summaries. One researcher (AS) completed the remaining summaries with the second researcher (MO) reviewing for accuracy and alignment; discrepancies were resolved via discussion. For patients, we created comparison groups between AUDIT-C scores indicative of moderate to high risk (score 4 - 7) versus severe risk (score 8 - 12) to examine response differences; for staff, we created comparison groups by low versus high implementation potential sites. See Multimedia Appendix 1 for demographic information for chair surveys (n=26), staff (n=18), and patient interviews (n=21).

Results

EHR and Chair Survey

Average AUDIT-C completion rates (73%) were relatively high but ranged greatly (35% - 93%). Survey implementation scores indicated a relatively high level of confidence, acceptability, and appropriateness regarding the UAU texting intervention; feasibility had the lowest mean, indicating logistical implementation concerns. See Multimedia Appendix 2 for absolute values on AUDIT-C completion rates and implementation scores.

Staff Interviews

Table 1 illustrates implementation barriers and facilitators. In comparing staff in low (n=7) versus high (n=11) implementation sites, low sites more frequently indicated addressing alcohol use was not a high priority. All staff discussed challenges to include information about the intervention in discharge papers, noting the paperwork is bulky and patients rarely read it. Facilitators were similar across high and low sites. Low sites discussed the difficulty of making changes, and that change takes time and requires strong justification. When asked about recent practice changes, low sites had fewer examples of what worked well and more frequently mentioned low staff morale, high turnover, and burnout.

XSL•FO
RenderX

Table . ED[a] staff qualitative interview barrier and facilitator themes from rapid analysis in accordance with i-PARIHS[b] domains.

Overall theme		Subtheme	Exemplar quote	i-PARIHS domain
Implementation barriers				
	Screening views	There is not 100% buy-in from staff to do alcohol risk screening (eg, feeling that patients are not truthful on screenings, that the threshold is too low for "positive," that there is screening fatigue, or that patients may be surprised by being asked).	"...We try to have as good of a completion as possible...there isn't a one hundred percent buy in from the staff when it comes to it...a lack of an education, because the staff is in our population as well...when you work in an emergency room, and you see a lot of crazy things, and you deal with a lot of crazy health issues, somebody comes in for one thing, you're not necessarily too concerned about doing a screening kind of unrelated to the main diagnosis. But I think that is part is the main reason why completion rates aren't one hundred percent..."	Context/inner (culture/climate)
	Screening process	Physicians or other providers generally do not view alcohol screening results.	"It's kind of cumbersome, unless it's actually pertinent to the case...We're juggling so much. It's a little too much for us to handle."	Context/inner (culture/climate)
	Need for services	There was a focus mostly on the need for services for those with more severe problems who are frequent visitors to the ED.	"I think that it could be very helpful for a certain set of patients...the patients that keep coming back that we've tried to help...And yet, a week later they back."	Recipients (motivation/beliefs)
	ED as an appropriate setting for addressing alcohol use	The ED is viewed as a difficult environment to address problems (lack of privacy and acute health problems) and have conversations beyond screening, doing more falls to low mixed priority due to staff workload/burden unless the patient is there for an alcohol issue.	"...I think it really depends on if the patient is there, specifically for the alcohol issue...In addition to the medical aspect, we try to make it a priority. But if a patient is identified in triage,...needing an SBIRT consult and they're not there for then it's not a priority it just becomes another thing in the chart that you know we just don't have time to address, unfortunately."	Context/inner (culture/climate)
	SMS text messaging intervention usefulness for ED patients	The text intervention was viewed as potentially less useful for patients with severe alcohol problems, less motivated to change their drinking, or certain populations (eg, older adults and low mobile phone users).	"To be honest, I don't think it would be that effective to the majority, the patients that we see like the so-called frequent flyers, or the one-time binge drinker kind of deal. They're not going to be interested in that..."	Innovation (degree of fit)
	Reasons patients may or may not accept the text intervention	Some patients may not have mobile phones/texting plans or may have lower motivation to change, which may deter from accepting a text intervention.	"...Some of them don't have cell phones or some of them have like prepaid cell phones...I honestly don't think it would be helpful."	Innovation (degree of fit)

XSL•FO
RenderX

Overall theme		Subtheme	Exemplar quote	i-PARIHS domain
	Offering the text intervention to patients	Patients are overwhelmed during the ED visit and receive a lot of discharge instructions, so text intervention information could get lost.	"...whereas the doctor might put it on the discharge paper...who reads them? They're twelve pages. It's going in the garbage. No one is really you know paying attention to it."	Innovation (degree of fit)
	Barriers to implementing a text intervention in the ED	It was expressed that physicians and nurses do not have the time to discuss alcohol use with patients, and that there are too many screens already being done, and the EHR[c] is already full of things to click.	"...It would be tough to operationalize that in a busy...any ED but...to have that as another initiative for our PAs, attendings for cases that are not necessarily for that, may be difficult to operationalize. It's not impossible I just don't know if it would be feasible in every single case."	Recipients (time, resources, support)
	Staff training needs	Staff are already overwhelmed with training; it would be difficult to add more training modules for a new intervention.	"...We get so many emails and have so many required modules that we need to watch that I think that those would probably not be the most effective way..."	Recipients (time, resources, support)
	General ease of making practice changes in ED	For something new to be implemented, staff must see value for the patient, and it must fit into the workflow without adding a burden.	"...when the staff buy into it and understand why we're doing it for the patient. I think that helps. They need to understand why we're implementing change."	Context/inner (innovation/change)
	Reflections on recent practice changes	It was perceived that change takes a long time based on recent experiences in implementing new practices.	"...I think it could be challenging to make changes. Especially if it requires like extra work. I think for people, it takes time or some of them just will never do it that they're just not complying with the changes. But I think it does take time and people to get used to the idea."	Context/inner (innovation/change)
	Culture of innovation	Staff morale and burnout are still recovering from COVID and some long-time staff are not as open to change; this creates some ambivalence about making changes.	"...burnt out, probably from the most part, I think most staff is burnt still from Covid, and is still in a recovery phase in general, because volume is kind of coming back, everybody is a little deconditioned in terms of volume, because we had very low levels, and we had high levels. So I think there's a level of burnout out there that we're constantly battling that's probably the biggest challenge that we're having at the moment."	Context/inner outer (culture/climate)
Implementation facilitators				

XSL•FO
RenderX

Overall theme		Subtheme	Exemplar quote	i-PARIHS domain
	Views on alcohol screening	The ED is a good place for screening because alcohol-related health issues in the ED are common, and it is beneficial information for the clinical team because the problem may not be identified or visible otherwise; it helps to normalize and destigmatize the issue of substance use.	"...I think it's a pretty good place. Um, Ah to identify um people who have alcohol abuse because, like yeah, like, I said, we do have like a nice, decent amount of people that do come in..."	Context/inner (culture/climate)
	Views on the alcohol screening process	Most patients are being screened. AUDIT[d] screening tool is built into EHR, and most patients are initially screened by triage/primary nursing and followed up via social work or screening, brief intervention, and referral to treatment (SBIRT[e]) health coaches.	"AUDIT screening tool is built into EMR, and most patients are initially screened by triage/primary nursing and followed up via social work or SBIRT."	Context/inner (culture/climate)
	The need for services for people with alcohol problems	It was viewed that the prevalence of those with alcohol problems in the ED is high; they see people across the continuum of use.	"...I see people on two ends of the spectrum right. Either they're completely intoxicated, and they're over utilizing the system for a bed to sleep, or they're coming in from like a bar or a nightclub overly intoxicated. Or I see the other end of this spectrum, which is somebody that's in severe withdrawal, delirium tremens. I've seen people that need to be admitted to the ICU setting or a hospital monitored bed because that's the normal vital signs risk procedure...Then you also see the other complications of alcohol abuse, such as like pancreatitis, for example, That's...the majority of the cases that I interact with."	Recipients (motivation/beliefs)
	ED as an appropriate setting for addressing alcohol use	The ED is a good place to identify problems and get the conversation started. The ED is a key access point that sees a high volume of people.	"I think it's a good place to identify...I think there are resources around that we can pull from. But again, if there's not a coach or somebody here, we struggle with where to refer people to."	Context/inner (culture/climate)

XSL•FO
RenderX

Overall theme		Subtheme	Exemplar quote	i-PARIHS domain
	SMS text messaging intervention usefulness for ED patients	It was viewed that there was no downside to offering the text intervention, even with a potentially small/incremental benefit, and patients may be more inclined because it is something they can engage with after an ED visit and because it is automated.	“I think it would be helpful for a patient that would be willing and wanting to need services...it really just offers another resource. I think nowadays one of the many issues that are occurring is that patients don’t have access to resources. So I think any step that we can do to help provide other resources to patients that are in need, would be helpful. Something in real time or something where they don’t have to wait on the phone don’t have to dial number can just kind of use their phone and text would be definitely helpful. I think it would be good.”	Innovation (degree of fit)
	Reasons patients may or may not accept the text intervention	The text intervention could be introduced to patients in a way that describes how it could be beneficial while ensuring privacy/confidentiality.	“...just letting them know that it’s confidential, that it’s helpful for them. They might, maybe someone of a clinical nature could tell them that they think it’s important that they address this with someone, with a professional.”	Innovation (degree of fit)
	Offering the text intervention to patients	Ways to facilitate offering the text intervention could be pagers/flyers with strong marketing, including it in discharge instructions, and using a team approach to offering. Materials plus a human touch would be beneficial.	“...The other way to do this is a passive way in terms of flyers...somebody could just opt in reading material while they are there...then in terms of explaining it is probably just a script so you give them a flyer and you can train probably a lot of support staff to say I know this is not the reason for your visit today we are working on this initiative, We give this to everybody...”	Innovation (degree of fit)
	Barriers to implementing a text intervention in the ED	Tapping into already existing workforce members (eg, social work, SBIRT, and referral coordinators) could be useful for offering the text intervention.	“...I find that nurses have more time to talk to patients, but I think it would have to be a team approach coming from the attending and working with the nurses...I think it would have to really be a team effort on everybody’s part.”	Recipients (time, resources, support)

XSL•FO
RenderX

Overall theme		Subtheme	Exemplar quote	i-PARIHS domain
	Barriers to implementing a text intervention in the ED	It could be useful to plug into already operating workflows or offer passive ways of offering the text intervention (eg, QR code).	"...like QR Codes for people to open up YouTube videos as distraction videos...But I think that places like the bathroom walls and provide to our codes for them to scan phone numbers and then champions. And as for health coaches and our social workers, and our case managers are often utilized to help with, uh disposition of those patients, right?"	Context/inner (innovation/change)
	Staff training needs	Brief training using familiar modalities would help with educating team members about the text intervention (eg, learning management system modules, morning huddles, Microsoft PowerPoint slides with all relevant information about the text program, and morning brief).	"...PowerPoint that you could always like reference or something that they could go back on. Um. I know a lot of the times like providers kind of say like, Oh, social work could do this...Some type of education um would be good, I think, including maybe, examples of what the text messages are, and then examples of like, If they did reach out to the program, what are they able to actually help with..."	Recipients (time, resources, support)
	General ease of making practice changes in ED	The ED is a setting used to change and is very adaptable. It values flexibility and change and is nimble.	"...I hate to say it depends on the change right so, yeah, but I think I think to answer your question a little easier our department is open to change pretty regularly we change things, most emergency departments do but we're pretty aggressive about being flexible, we pride ourselves sort of on that..."	Context/inner (innovation/change)
	Reflections on recent practice changes	EDs have a lot of recent experiences with change and methods to disseminate information and gain feedback and to have consistency in communication.	"...the changes with COVID testing and requirements and admission process...but I think what went well, people appreciate consistency, I think in communication, so I think what I appreciate now is that everyone knows, where to look for updates and things like that..."	Context/inner (innovation/change)

XSL•FO
RenderX

Overall theme	Subtheme	Exemplar quote	i-PARIHS domain
Culture of innovation	EDs are overall welcoming of innovation and embrace a culture of change; leadership and a cohesive team are important in change/innovation.	"There has been a lot of change here. It's always changing. They've received multiple certifications. (redacted). Um, They've made new rooms. They've gone through a renovation. They've gone through new leadership..."	Context/inner outer (culture/climate)

[a]ED: emergency department.

[b]i-PARIHS: Integrated Promoting Action on Research Implementation in Health Services.

[c]EHR: electronic health record.

[d]AUDIT: Alcohol Use Disorders Identification Test.

[e]SBIRT: Screening, Brief Intervention, and Referral to Treatment.

Patient Interviews

Patient results indicated strong interest and potential benefit from a UAU text intervention with limited concerns (Table 2). When comparing patients who screened moderate to high (n=6) versus severe (n=15) on the AUDIT-C, only patients at moderate to high risk said they had a goal to use the intervention to prevent future heavy drinking, while both groups mentioned abstinence and moderation. Perceived benefits also varied. Patients at severe risk thought the program could help stop the progression of alcohol use disorder symptoms, change their lifestyle, and be a distraction from drinking; patients at moderate to high risk thought the program could improve their job roles and family relationships, and provide information about drinking and change their thoughts about drinking. Both groups thought the intervention could provide support and improve their health and had similar content preferences. Patients at severe risk additionally mentioned using the text intervention to track drinking, set and complete goals, engage family support, avoid people who trigger their use, and reduce heavy drinking during drinking episodes.

XSL•FO
RenderX

Table . ED[a] patient qualitative interview themes from rapid analysis in accordance with i-PARIHS[b] domains.

Overall theme	Subtheme	Exemplar quote	i-PARIHS domain
Patient phone use patterns	• Most patients had mobile phones and used them for a variety of purposes, including frequent texting • It was rare that patients used their mobile phones for health or to manage their alcohol use	N/A	Recipients • Time, resources, and support • Motivations, values, and beliefs
Interest in a text intervention for alcohol use	• Most patients were interested in receiving the text intervention • Patients were interested in different goals (eg, to prevent future heavy drinking, reduce current drinking, or stop drinking completely) • Patients were interested in a variety of reasons (eg, improve health and relationships and change drinking cognitions)	• "Yeah. Maybe for myself if, if, you know, in the future, I decide to drink – I start drinkin' more. When I see myself drinkin' more, I wanna know information. Maybe I could prevent myself from, you know, gettin' into alcoholism or somethin' like that." [Participant 27] • "Personally, maybe it might change my mind on why I'm drinking at that moment. Or if I'm feeling low, maybe it'd inspire me not to drink as much. Maybe those texts'll be useful." [Participant 17]	Innovation (degree of fit)
Content	• Patients were interested in a variety of content (eg, quotes, uplifting, inspirational, or humorous messages, memes, images, resources, videos, and links with information) • Patient preference on the desired frequency of texts ranged from 3 times per week to several times per day	• "I like resources. Resources as in like what you should do besides picking up a drink." [Participant 13]	Innovation (degree of fit)
Offer preferences	• Most patients had no preference of who (eg, nurse or doctor) should offer the intervention, but it was more important that it be a personal conversation showing concern and empathy • Most patients had no preference for whether they would prefer the intervention to be offered through the health system versus a third party	• "Um, maybe towards the end and I'm, um, gettin' discharged or somethin,' like, when I feel better." [Participant 27]	Innovation (degree of fit)
Concerns	• Patients did not generally cite concerns related to privacy or cost; concern was that the program does not send too many messages, that they would not be able to read messages immediately, or that messages would be too long	• "So if I can't get to the message right away, I won't. But usually when I open the messages up, I do take concern in what's being said or written." [Participant 17]	Innovation (degree of fit)

XSL•FO
RenderX

Overall theme	Subtheme	Exemplar quote	i-PARIHS domain
Motivation to change	• Most participants had at least some motivation to change their drinking (eg, quit or cut down) and mentioned several reasons (eg, family and job)	• "I'm at a 10 where I wanna change, Um, I have tried. I cut back, um, 'cause I used to drink like every day and I cut back a lot, so…looking to cut it back even further to, and just stop drinking alcohol." [Participant 25]	Recipients (motivations, values, beliefs)
Benefits to enrolling	• Patients saw a number of benefits to enrolling (eg, help with mental and physical health, receiving information about negative effects of alcohol, having help in limiting alcohol use, or prompting in the moment to think differently about drinking)	• "It just healthy to stop. Make your body feel better, more healthy." [Participant 26]	Innovation

[a]ED: emergency department.

[b]i-PARIHS: Integrated Promoting Action on Research Implementation in Health Services.

Discussion

The results of this formative study on barriers and facilitators to implementing UAU texting interventions in EDs indicate both promise and caution. Quantitative data suggested concerns with feasibility (eg, low alcohol risk screening rates and feasibility scores). Qualitative results confirmed and extended quantitative results; providers viewed offering the intervention as appropriate and acceptable, especially to address the high volume of alcohol-related visits and support patients longer term; patients viewed the intervention as a positive way to meet drinking goals and saw several health benefits; they saw limited drawbacks (eg, slight concerns about time to read messages).

Providers felt there was no downside to offering the intervention even if there was incremental patient benefit. However, differences were noted among low and high implementation potential sites: variable screening rates and lower ratings on feasibility (eg, competing priorities, staff burnout, and turnover), indicating potential additional support strategies needed to create staff buy-in and sustain intervention engagement in low-potential sites. Larger EDs with higher patient census' may have unique implementation barriers, as they tended to have lower alcohol risk screening completion rates that would impact their ability to identify those who would benefit from text interventions. Other studies examining provider acceptance, barriers, and facilitators to implementing technology-based interventions in their EDs (not focused on alcohol use, however) revealed similar findings: high ratings of usage and value, but lower rates of actual use due to insufficient training, lack of awareness of the program, and technical factors (eg, compatibility) [39,40]. Future studies should examine ED characteristics that may impact implementation and design and test specific strategies to address these characteristics.

Patients of varying alcohol use risk severity indicated potential differences in drinking goals: those with higher severity may be less likely to have moderation goals and more likely to have abstinence goals. Results may reflect differences in how drinking has differently impacted health, relationships, functioning, and experience of alcohol use disorder symptoms among those with lower or higher severity. Patients' reasons for using a text program for UAU in our study were similar to that of another study examining patients' reasons for attending a residential program for alcohol use, including health, lifestyle, and sobriety goals [41]. Future studies of SMS text messaging programs for UAU should take into consideration different patient goals and experiences and how this could impact implementation strategies to engage patients. Designing programs that can address both moderate and more severe patients' needs and uses is important; the ability of the program to tailor to the needs and preferences of individual patients should be inherent in the design.

This study had limitations, including being conducted in one health system, which may affect generalizability, and classifying implementation potential using the 5-point Likert scale, which was meaningful in identifying potential implementation issues, but may not indicate practically meaningful differences.

In conclusion, SMS text messaging interventions for alcohol use have the potential to offer ED staff another tool to assist patients with meeting alcohol-related health goals. Patients welcome these interventions as a way to receive support and information. However, EDs will have varying levels of motivation and readiness to implement these driven by barriers related to feasibility. Findings may be useful for the development and testing of implementation strategies in larger cluster-randomized trials, an important area for mobile behavioral health [42]. Automating enrollment (eg, QR codes and text invites), capitalizing on existing workflows, and engaging ED leadership may be fruitful implementation strategies to test in future ED implementation research.

XSL•FO
RenderX

Acknowledgments

We are thankful to the sites and staff for their participation in the study. This study is supported by the National Institute on Alcohol Abuse and Alcoholism of the National Institutes of Health under award number R21AA029734 (to MO). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Demographics for chair surveys (n=26), staff interviews (n=18), and patient interviews (n=21).
[DOCX File, 16 KB - formative_v9i1e65187_app1.docx]

Multimedia Appendix 2
ED EHR alcohol screening data and feasibility, appropriateness, acceptability, and confidence of implementing alcohol SMS text messaging intervention in the ED in 17 sites. ED: emergency department; EHR: electronic health record.
[DOCX File, 17 KB - formative_v9i1e65187_app2.docx]

Checklist 1
COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist.
[DOCX File, 17 KB - formative_v9i1e65187_app3.docx]

Checklist 2
CHERRIES (Checklist 2 for Reporting Results of Internet e-Surveys) checklist.
[DOCX File, 18 KB - formative_v9i1e65187_app4.docx]

References

1. Sacks JJ, Gonzales KR, Bouchery EE, Tomedi LE, Brewer RD. 2010 national and state costs of excessive alcohol consumption. Am J Prev Med 2015 Nov;49(5):e73-e79. [doi: 10.1016/j.amepre.2015.05.031] [Medline: 26477807]
2. White AM, Castle IJP, Powell PA, Hingson RW, Koob GF. Alcohol-related deaths during the COVID-19 pandemic. JAMA 2022 May 3;327(17):1704-1706. [doi: 10.1001/jama.2022.4308] [Medline: 35302593]
3. Theriault KM, Rosenheck RA, Rhee TG. Increasing emergency department visits for mental health conditions in the United States. J Clin Psychiatry 2020 Jul 28;81(5):20m13241. [doi: 10.4088/JCP.20m13241] [Medline: 32726001]
4. White AM, Slater ME, Ng G, Hingson R, Breslow R. Trends in alcohol-related emergency department visits in the United States: results from the nationwide emergency department sample, 2006 to 2014. Alcohol Clin Exp Res 2018 Feb;42(2):352-359. [doi: 10.1111/acer.13559] [Medline: 29293274]
5. Suen LW, Makam AN, Snyder HR, et al. National prevalence of alcohol and other substance use disorders among emergency department visits and hospitalizations: NHAMCS 2014-2018. J Gen Intern Med 2022 Aug;37(10):2420-2428. [doi: 10.1007/s11606-021-07069-w] [Medline: 34518978]
6. Strayer RJ, Friedman BW, Haroz R, et al. Emergency department management of patients with alcohol intoxication, alcohol withdrawal, and alcohol use disorder: a white paper prepared for the American Academy of Emergency Medicine. J Emerg Med 2023 Apr;64(4):517-540. [doi: 10.1016/j.jemermed.2023.01.010] [Medline: 36997435]
7. Hawk KF, D'Onofrio G. Time to treat alcohol use disorder in the emergency department. Ann Emerg Med 2023 Apr;81(4):450-452. [doi: 10.1016/j.annemergmed.2022.11.013] [Medline: 36775724]
8. Berman AH, Gajecki M, Sinadinovic K, Andersson C. Mobile interventions targeting risky drinking among university students: a review. Curr Addict Rep 2016;3(2):166-174. [doi: 10.1007/s40429-016-0099-6] [Medline: 27226948]
9. Tofighi B, Nicholson JM, McNeely J, Muench F, Lee JD. Mobile phone messaging for illicit drug and alcohol dependence: a systematic review of the literature. Drug Alcohol Rev 2017 Jul;36(4):477-491. [doi: 10.1111/dar.12535] [Medline: 28474374]
10. Ranney ML, Pisani AR, Chernick LS. The role of texting in addressing mental health. In: Moreno MA, Radovic A, editors. Technology and Adolescent Mental Health: Springer International Publishing; 2018:207-215. [doi: 10.1007/978-3-319-69638-6_15]
11. Keoleian V, Polcin D, Galloway GP. Text messaging for addiction: a review. J Psychoactive Drugs 2015;47(2):158-176. [doi: 10.1080/02791072.2015.1009200] [Medline: 25950596]

XSL•FO
RenderX

12. Schwebel FJ, Larimer ME. Text message reminders as an adjunct to a substance use intervention for adolescents and young adults: pilot feasibility and acceptability findings. Digit Health 2020;6. [doi: 10.1177/2055207620965052] [Medline: 33110614]
13. Hutton A, Prichard I, Whitehead D, et al. mHealth interventions to reduce alcohol use in young people: a systematic review of the literature. Compr Child Adolesc Nurs 2020 Sep;43(3):171-202. [doi: 10.1080/24694193.2019.1616008] [Medline: 31192698]
14. Mason M, Ola B, Zaharakis N, Zhang J. Text messaging interventions for adolescent and young adult substance use: a meta-analysis. Prev Sci 2015 Feb;16(2):181-188. [doi: 10.1007/s11121-014-0498-7] [Medline: 24930386]
15. Sibley AL, Noar SM, Muessig KE, et al. An automated text messaging intervention to reduce substance use self-stigma (Project RESTART): protocol for a feasibility and acceptability pilot study. JMIR Res Protoc 2024 Aug 9;13:e59224. [doi: 10.2196/59224] [Medline: 39121478]
16. Spieker AJ, Nelson LA, Rothman RL, et al. Feasibility and short-term effects of a multi-component emergency department blood pressure intervention: a pilot randomized trial. J Am Heart Assoc 2022 Mar;11(5):e024339. [doi: 10.1161/JAHA.121.024339] [Medline: 35195015]
17. Bendtsen M, McCambridge J, Åsberg K, Bendtsen P. Text messaging interventions for reducing alcohol consumption among risky drinkers: systematic review and meta-analysis. Addiction 2021 May;116(5):1021-1033. [doi: 10.1111/add.15294] [Medline: 33047865]
18. Muench F, Madden SP, Oommen S, et al. Automated, tailored adaptive mobile messaging to reduce alcohol consumption in help-seeking adults: a randomized controlled trial. Addiction 2024 Mar;119(3):530-543. [doi: 10.1111/add.16391] [Medline: 38009576]
19. Suffoletto B, Chung T, Muench F, Monti P, Clark DB. A text message intervention with adaptive goal support to reduce alcohol consumption among non-treatment-seeking young adults: non-randomized clinical trial with voluntary length of enrollment. JMIR Mhealth Uhealth 2018 Feb 16;6(2):e35. [doi: 10.2196/mhealth.8530] [Medline: 29453191]
20. Suffoletto B, Kristan J, Callaway C, et al. A text message alcohol intervention for young adult emergency department patients: a randomized clinical trial. Ann Emerg Med 2014 Dec;64(6):664-672.e4. [doi: 10.1016/j.annemergmed.2014.06.010] [Medline: 25017822]
21. Sharpe S, Kool B, Whittaker R, et al. Effect of a text message intervention to reduce hazardous drinking among injured patients discharged from a trauma ward: a randomized controlled trial. NPJ Digit Med 2018;1:13. [doi: 10.1038/s41746-018-0019-3] [Medline: 31304298]
22. Burner E, Zhang M, Terp S, et al. Feasibility and acceptability of a text message-based intervention to reduce overuse of alcohol in emergency department patients: controlled proof-of-concept trial. JMIR Mhealth Uhealth 2020 Jun 4;8(6):e17557. [doi: 10.2196/17557] [Medline: 32496203]
23. Nesvåg S, McKay JR. Feasibility and effects of digital interventions to support people in recovery from substance use disorders: systematic review. J Med Internet Res 2018 Aug 23;20(8):e255. [doi: 10.2196/jmir.9873] [Medline: 30139724]
24. Ramsey AT, Satterfield JM, Gerke DR, Proctor EK. Technology-based alcohol interventions in primary care: systematic review. J Med Internet Res 2019 Apr 8;21(4):e10859. [doi: 10.2196/10859] [Medline: 30958270]
25. Kmiec J, Suffoletto B. Implementations of a text-message intervention to increase linkage from the emergency department to outpatient treatment for substance use disorders. J Subst Abuse Treat 2019 May;100:39-44. [doi: 10.1016/j.jsat.2019.02.005] [Medline: 30898326]
26. Harvey G, Kitson A. PARIHS revisited: from heuristic to integrated framework for the successful implementation of knowledge into practice. Implement Sci 2016 Mar 10;11:33. [doi: 10.1186/s13012-016-0398-2] [Medline: 27013464]
27. Swindle T, Johnson SL, Whiteside-Mansell L, Curran GM. A mixed methods protocol for developing and testing implementation strategies for evidence-based obesity prevention in childcare: a cluster randomized hybrid type III trial. Implement Sci 2017 Jul 18;12(1):90. [doi: 10.1186/s13012-017-0624-6] [Medline: 28720140]
28. Kitson A, Harvey G. Facilitating an evidence-based innovation into practice: the novice facilitator's role. In: Implementing Evidence-Based Practice in Healthcare: Routledge; 2015:85-104. [doi: 10.4324/9780203557334-6]
29. O'Grady MA, Kapoor S, Harrison L, Kwon N, Suleiman AO, Muench FJ. Implementing a text-messaging intervention for unhealthy alcohol use in emergency departments: protocol for implementation strategy development and a pilot cluster randomized implementation trial. Implement Sci Commun 2022 Aug 6;3(1):86. [doi: 10.1186/s43058-022-00333-y] [Medline: 35933560]
30. Palinkas LA, Aarons GA, Horwitz S, Chamberlain P, Hurlburt M, Landsverk J. Mixed method designs in implementation research. Adm Policy Ment Health 2011 Jan;38(1):44-53. [doi: 10.1007/s10488-010-0314-z] [Medline: 20967495]
31. Bush K, Kivlahan DR, McDonell MB, Fihn SD, Bradley KA. The AUDIT alcohol consumption questions (AUDIT-C): an effective brief screening test for problem drinking. Arch Intern Med 1998;158(16):1789-1795. [doi: 10.1001/archinte.158.16.1789] [Medline: 9738608]
32. Weiner BJ, Lewis CC, Stanick C, et al. Psychometric assessment of three newly developed implementation outcome measures. Implement Sci 2017 Aug 29;12(1):108. [doi: 10.1186/s13012-017-0635-3] [Medline: 28851459]

XSL•FO
RenderX

33. Palinkas LA, Horwitz SM, Green CA, Wisdom JP, Duan N, Hoagwood K. Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. Adm Policy Ment Health 2015 Sep;42(5):533-544. [doi: 10.1007/s10488-013-0528-y] [Medline: 24193818]
34. Benoot C, Hannes K, Bilsen J. The use of purposeful sampling in a qualitative evidence synthesis: a worked example on sexual adjustment to a cancer trajectory. BMC Med Res Methodol 2016 Feb 18;16:21. [doi: 10.1186/s12874-016-0114-6] [Medline: 26891718]
35. Ranney ML, Choo EK, Wang Y, Baum A, Clark MA, Mello MJ. Emergency department patients' preferences for technology-based behavioral interventions. Ann Emerg Med 2012 Aug;60(2):218-227. [doi: 10.1016/j.annemergmed.2012.02.026] [Medline: 22542311]
36. Kool B, Smith E, Raerino K, Ameratunga S. Perceptions of adult trauma patients on the acceptability of text messaging as an aid to reduce harmful drinking behaviours. BMC Res Notes 2014 Jan 4;7:4. [doi: 10.1186/1756-0500-7-4] [Medline: 24387293]
37. Gale RC, Wu J, Erhardt T, et al. Comparison of rapid vs in-depth qualitative analytic methods from a process evaluation of academic detailing in the Veterans Health Administration. Implement Sci 2019 Feb 1;14(1):11. [doi: 10.1186/s13012-019-0853-y] [Medline: 30709368]
38. Hamilton AB. Qualitative methods in rapid turn-around health services research. Presented at: 2013 VA HSR&D Cyberseminar Spotlight on Women's Health; Dec 11, 2013.
39. Fujimori R, Liu K, Soeno S, et al. Acceptance, barriers, and facilitators to implementing artificial intelligence-based decision support systems in emergency departments: quantitative and qualitative evaluation. JMIR Form Res 2022 Jun 13;6(6):e36501. [doi: 10.2196/36501] [Medline: 35699995]
40. Jun S, Plint AC, Campbell SM, Curtis S, Sabir K, Newton AS. Point-of-care cognitive support technology in emergency departments: a scoping review of technology acceptance by clinicians. Acad Emerg Med 2018 May;25(5):494-507. [doi: 10.1111/acem.13325] [Medline: 28960689]
41. Marsh TN, Eshakakogan C, Spence M, et al. Patient experience of an abstinence-based indigenous residential treatment program in Northern Ontario: a descriptive qualitative study. Front Health Serv 2024;4:1387184. [doi: 10.3389/frhs.2024.1387184] [Medline: 39777229]
42. Graham AK, Lattie EG, Powell BJ, et al. Implementation strategies for digital mental health interventions in health care settings. Am Psychol 2020 Nov;75(8):1080-1092. [doi: 10.1037/amp0000686] [Medline: 33252946]

Abbreviations

AUDIT-C: Alcohol Use Disorders Identification Test-Consumption
CHERRIES: Checklist for Reporting Results of Internet e-Surveys
COREQ: Consolidated Criteria for Reporting Qualitative Research
ED: emergency department
EHR: electronic health record
i-PARIHS: Integrated Promoting Action on Research Implementation in Health Services
UAU: unhealthy alcohol use

Edited by A Mavragani; submitted 08.08.24; peer-reviewed by H Hutton; revised version received 30.01.25; accepted 30.01.25; published 03.03.25.

Please cite as:
O'Grady M, Harrison L, Suleiman A, Hutchison M, Kwon N, Muench F, Kapoor S
Text Messaging Interventions for Unhealthy Alcohol Use in Emergency Departments: Mixed Methods Assessment of Implementation Barriers and Facilitators
JMIR Form Res 2025;9:e65187
URL: https://formative.jmir.org/2025/1/e65187
doi:10.2196/65187

© Megan O'Grady, Laura Harrison, Adekemi Suleiman, Morica Hutchison, Nancy Kwon, Frederick Muench, Sandeep Kapoor. Originally published in JMIR Formative Research (https://formative.jmir.org), 3.3.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Patients' Experience With Evaluation by Both a Musculoskeletal Physician and Physical Therapist in the Same Digital Visit: Survey Study

Mary I O'Connor, MD; Carolyn Chudy, DO; Kaitlyn C Peters, NP; Megan Ribaudo; Carrie McCulloch, MD; Jared Aguilar, DPT; Trista Taylor, MPA; Ryan A Grant, MS, MBA, MD

Vori Health, Inc, Nashville, TN, United States

Corresponding Author:
Mary I O'Connor, MD
Vori Health, Inc, Nashville, TN, United States

Abstract

Background: Patients undergoing evaluation for musculoskeletal concerns are often seen by a physician and physical therapist in the in-person setting in a sequential manner. This process typically delays the onset of nonoperative care, inclusive of physical therapy, and creates the risk of inadequate clinical collaboration between physician and physical therapist. To address these issues, we designed a novel initial patient evaluation to a group visit in which both a specialty-trained musculoskeletal physician and physical therapist simultaneously evaluate a patient together in the digital encounter.

Objective: The aim of the study is to gain insights from patients on their experience with this innovative digital simultaneous musculoskeletal medical doctor and physical therapist (MD+PT) visit format for the initial evaluation of musculoskeletal concerns.

Methods: An electronic 7-question survey was sent to 750 patients who completed an MD+PT visit asking them to comment on prior musculoskeletal evaluations and their experience with the MD+PT format.

Results: In total, 195 (26%) patients responded to the survey with the frequent body regions of diagnosis being lumbar spine (n=65), knee (n=32), shoulder (n=21), cervical spine (n=20), hip (n=14), and hand (n=11). Most patients had prior musculoskeletal experience with a physician or nurse practitioner (171/195, 87.7%) or physical therapist (148/195, 75.9%) with nearly all such encounters in the in-person setting (161/171,94.2% for physician or nurse practitioner and 144/148, 97.3% for physical therapy). Only 3.1% (6/193) of patients reported seeing both a physician and physical therapist during the same in-person visit. Patients rated the simultaneous MD+PT visit very favorably: this type of digital evaluation saved them time (179/192, 93.2%) and permitted them to promptly start their treatment plan (174/192, 90.6%). Overall, 87.5% (168/192) rated the MD+PT visit as enjoyable, and 92.2% (177/192) responded that it increased their confidence with understanding their medical condition and how to start treating it.

Conclusions: Our early experience with the evaluation of patients with musculoskeletal conditions by both a specialty-trained musculoskeletal physician and physical therapist simultaneously in the same digital visit resulted in patients reporting a very positive experience with high satisfaction, engagement, and confidence in understanding their diagnosis and how to start treating it.

(JMIR Form Res 2025;9:e66744) doi:10.2196/66744

KEYWORDS

telemedicine; musculoskeletal care; patient satisfaction; multidisciplinary care; digital visit

Introduction

Musculoskeletal disorders are a leading cause of disability and health care expenditure worldwide [1]. Inappropriate care exacerbates the cost of musculoskeletal care as seen, for example, with high-cost spine [2] and total knee replacement [3] surgeries. Efficient and effective evidence-based evaluation of musculoskeletal disorders is critical to ensuring patient-centric care, delivering quality patient outcomes, and lowering health care costs. The type of care a patient with a musculoskeletal concern seeks and the way that care is delivered—especially at the onset of a care journey—can impact the risk of inappropriate and expensive imaging [4-6], procedures [7,8], and surgery [2,3,6]. In other words, what type of clinician a patient initially starts with influences the care path a patient will receive.

Most patients with musculoskeletal concerns first seek care from their primary care clinician [9]; however, many primary care clinicians lack fundamental competency in the musculoskeletal examination and do not feel comfortable evaluating, diagnosing, and treating musculoskeletal conditions [10]. In many states, patients may now directly see a physical

XSL•FO
RenderX

therapist without a physician referral, via direct access physical therapy, due to the expansion of the scope of practice for physical therapists [11]. From either primary care or physical therapy, many patients are subsequently referred to an orthopedic or spine surgeon, or in the minority of cases a specialty nonoperative musculoskeletal physician (eg, physical medicine and rehabilitation or sports medicine physician), collectively referred to as specialty musculoskeletal physician. Creation of a comprehensive treatment plan for these patients will be dependent upon the completion of a specialty musculoskeletal physician's clinical assessment, which often is delayed given the time required for separate physician and physical therapist appointments, and potentially by discordance in evaluations and advice by these 2 clinicians.

For many patients, musculoskeletal care can be effectively delivered through telemedicine. Digital delivery of medical services can be very convenient for patients, particularly for those with challenges to access in-person clinicians [12]. Physical examination of the patient with a musculoskeletal condition can be readily performed in the digital space [13]. The adequacy of that digital physical examination is supported by research showing good to very good concordance of the diagnosis of the patient with a musculoskeletal condition between the digital and in-person clinician [13]. Moreover, the concordance of treatment plans for patients with musculoskeletal conditions created in these 2 practice settings was "probably good to excellent" [14]. Finally, a meta-analysis of 35 systematic reviews assessing the use of telemedicine versus in-person care for musculoskeletal conditions showed telemedicine to be beneficial to or equal in patient-reported outcomes measures and objective measures, including physical function, and to have notably lower costs. Furthermore, in the small number of studies in the review reporting on patient experience measures, telemedicine was equivalent to in-person care [15].

We developed an initial clinical digital visit structure to address these delays in treatment created by the traditional sequential clinical visits between physician and physical therapist, streamline the patient experience, and advance evidence-based patient care. In this model, patients see a nonoperative musculoskeletal specialty physician (or a musculoskeletal nurse practitioner) and a physical therapist together—literally at the same time—during a simultaneous digital medical visit via our telehealth platform [16]. To our knowledge, such a simultaneous evaluation, which we will abbreviate as an "MD+PT" (musculoskeletal medical doctor and physical therapist) digital evaluation, has not been previously systematically implemented in either the in-person or digital care environment. To gain insights into the patient experience with this simultaneous patient-centric model, we created a descriptive survey to better understand the experience from the patients' point of view.

Methods

Ethical Considerations

Our study, including the consent language and questionnaire, was reviewed by The Institute for Evaluation and Research Institutional Review Board and approved as exempt. Patients who completed an MD+PT digital evaluation were sent an email asking them to voluntarily consent to the research study and complete a 7-question survey about their appointment experience with the MD+PT evaluation (Textbox 1). All patients were 18 years or older of age. The email informed patients that the purpose of the study was to examine the impact of enabling patients to see a medical doctor or nurse practitioner and physical therapist together in the same digital appointment. Patients were provided no compensation for completing the survey. Those who elected to participate were informed that they could withdraw from the study at any time and skip questions if they chose to do so. The consent informed patients that their survey results would be confidential, shared with our researchers and not their medical team. Participants could answer the questionnaire in an anonymous manner and complete it only once. The questions could not be answered until participants consented to participate in the study. Data integrity was protected with access only provided to the researchers conducting the data analysis.

XSL•FO
RenderX

Textbox 1. The 7-question survey used in this study of 195 patients who completed an evaluation with both a nonoperative musculoskeletal physician and physical therapist during the same digital encounter.

1. Before coming to Vori Health, have you ever seen a medical doctor (or nurse practitioner or physician assistant) for back, muscle, or joint pain?

- Yes, no

If yes, were any of these visits telehealth appointments? Yes/no

2. Before coming to Vori Health, have you ever seen a physical therapist for back, muscle, or joint pain?

- Yes, no

If yes, were any of these visits telehealth appointments? Yes/no

3. Before coming to Vori Health, have you ever been seen by both a medical doctor/nurse practitioner and physical therapist together in the same appointment for back, muscle, or joint pain?

- Yes, no

If yes, were any of these visits telehealth appointments? Yes/no

The next few questions will ask you about your experience at Vori Health seeing a medical doctor or nurse practitioner and physical therapist together in the same appointment.

4. Seeing a medical doctor/nurse practitioner and physical therapist together in the same appointment saved me time.

- Strongly agree, agree, neutral, disagree, strongly disagree

5. Seeing a medical doctor/nurse practitioner and physical therapist together in the same appointment increased my confidence in understanding my medical condition and how to start treating it.

- Strongly agree, agree, neutral, disagree, strongly disagree

6. Seeing a medical doctor/nurse practitioner and physical therapist together in the same appointment helped me to start my treatment plan faster compared to if I had to see the doctor and physical therapist in different appointments.

- Strongly agree, agree, neutral, disagree, strongly disagree

7. Seeing a medical doctor/nurse practitioner and physical therapist together in the same appointment made my care experience more enjoyable.

- Strongly agree, agree, neutral, disagree, strongly disagree

If disagree, or strongly disagree, show question below:

Please indicate why the appointment was not enjoyable for you (choose all that apply):

o The experience felt too complicated.

o I did not feel I had enough time with each clinician.

o I did not feel comfortable sharing confidential medical information with anyone other than my medical doctor.

o Other (free text):

Questionnaire

The questionnaire sought patients' prior experience with musculoskeletal care, both in the in-person and telemedicine settings. Patients were asked if they had ever been seen by both a medical doctor (or nurse practitioner or physician associate) and physical therapist in the same appointment. Additional questions were asked related to patients' experiences with their MD+PT digital evaluation. To recruit patients to the study, an email survey was sent to all patients who completed the MD+PT evaluation until we had 200 patients who consented to participate in the study. There were no other inclusion or exclusion criteria.

The MD+PT Clinical Evaluation and Care Plan

The MD+PT digital evaluation was performed by a specialty musculoskeletal physician (typically, physical medicine and rehabilitation), or for a smaller number of patients by an experienced musculoskeletal nurse practitioner, and a physical therapist who had a minimum of 5 years of musculoskeletal experience. Each integrated visit was 40 minutes in duration. The physician or nurse practitioner, physical therapist, and patient were each in separate physical locations and logged into the same videoconferencing link to be together in the digital space for the evaluation. The history and digital physical examination were performed by a combination of physician or nurse practitioner and physical therapist. Each clinician could ask additional questions to the patient to provide more details to the medical history. The physical examination was completely digital and based on published guidelines [13]. Both the physician or nurse practitioner and physical therapist could instruct the patient to perform various movements and tests to complete the physical examination to the satisfaction of both clinicians. Based on the available evidence, the physician or nurse practitioner then rendered a medical diagnosis, and the physical therapist provided a functional diagnosis. If outside records, inclusive of previous imaging, were not available to the clinicians at the time of this initial visit, then the physician or nurse practitioner would request the outside records and imaging for subsequent review.

XSL•FO
RenderX

At the completion of this visit, patients were immediately started on an evidence-based biopsychosocial personalized care plan that may include evidence-based prescriptions, imaging, laboratory studies, and a personalized home exercise program facilitated by computer vision motion-tracking technology. Patients were subsequently scheduled for follow-up digital physical therapy care and, if applicable, dedicated health coaching and registered dietitian care, given the overlap of modifiable risk factors with musculoskeletal conditions (eg, obesity). Follow-up physician or nurse practitioner visits were also scheduled as clinically appropriate.

Results

Overview

Surveys were sent to a total of 750 patients, with 200 patients consenting to the study. In total, 5 patients consented but did not answer any questions on the survey resulting in 195 responses, a 26% survey response rate. Surveys were fully answered by 192 patients and incompletely answered by 3 patients who each did not complete 1‑2 questions. Patients who responded to the survey had a variety of diagnoses involving the following body regions: lumbar spine (n=65), knee (n=32), shoulder (n=21), cervical spine (n=20), hip (n=14), hand (n=11), ankle (n=8), elbow or wrist (n=6), thoracic spine (n=5), and other and nonspecific body regions (n=13).

Prior Musculoskeletal Clinical Experience

Patients were asked about their prior experience with the evaluation of back, muscle, or joint pain. In total, 171 of 195 (87.7%) patients had previously seen a physician or nurse practitioner, with 161 of 171 (94.2%) of these encounters occurring in the in-person setting. A total of 148 of 195 (75.9%) patients had previously seen a physical therapist, with 144 of 148 (97.3%) of these encounters occurring in the in-person setting.

We asked patients about prior experience of being evaluated by both the physician (or nurse practitioner or physician assistant) and physical therapist in the same clinical visit for back, muscle, or joint pain. Only 3.1% (6/193) reported a prior visit that combined a physician and physical therapist with 1 patient commenting that this occurred during a hospital stay, with no patient reporting the encounter as digital. We concluded that our cohort had essentially no experience with simultaneous musculoskeletal clinical visits, particularly in the digital setting.

MD+PT Digital Experience

The remainder of our survey questions focused on the experience of the patient with the MD+PT digital evaluation. Table 1 shows that 93.2% (179/192) agreed that this appointment format saved them time (strongly agreed: 162/192, 84.4% and somewhat agreed: 17/192, 8.9%).

Table . Survey responses (n=192) to question asking patients whether they felt the MD+PT digital evaluation saved them time.

Response	Participant count, n
Strongly agree	162
Somewhat agree	17
Neither agree nor disagree	11
Somewhat disagree	0
Strongly disagree	2

When patients were asked if the MD+PT digital evaluation allowed them to start their treatment plan faster than if they had seen a physician and physical therapist in separate appointments, 90.6% (174/192) of patients responded affirmatively (strongly agreed: 155/192, 80.7% and somewhat agreed: 19/192, 9.9%; Table 2).

Table . Survey responses (n=192) to question asking patient whether the MD+PT digital evaluation helped them start a treatment plan faster compared to seeing a doctor and physical therapist in separate appointments.

Response	Participant count, n
Strongly agree	155
Somewhat agree	19
Neither agree nor disagree	16
Somewhat disagree	1
Strongly disagree	1

Given our focus on the patient experience, the 2 most significant survey questions asked about patient enjoyment of their visit experience and confidence in medical decision-making. Overall, Table 3 shows that 87.5% (168/192) rated the visit as enjoyable (strongly agree: 135/192, 70.3% and somewhat agree: 33/192, 17.2%).

XSL•FO
RenderX

Table . Survey responses (n=192) to question asking patients whether the MD + PT digital evaluation made their patient care experience more enjoyable.

Response	Participant count, n
Strongly agree	135
Somewhat agree	33
Neither agree nor disagree	21
Somewhat disagree	1
Strongly disagree	2

The patients who responded as "somewhat disagreed" (n=1) or "strongly disagreed" (n=2) each provided a comment: (1) the practitioner and therapist were not very engaging, and I could not connect with them; (2) digital appointment diminished the transfer of information between the patient and provider; and (3) in-person health care is so very important, especially the first visit.

When asked if MD+PT digital evaluation increased their confidence with understanding their medical condition and how to start treating it, Table 4 illustrates that 92.2% (177/192) positively responded (strongly agree: 137/192, 71.4% and somewhat agree: 40/192, 20.8%).

Table . Survey responses (n=192) to question asking patients whether the MD+PT digital evaluation increased their confidence in understanding their medical condition and how to start treating it.

Response	Participant count, n
Strongly agree	137
Somewhat agree	40
Neither agree nor disagree	12
Somewhat disagree	1
Strongly disagree	2

Discussion

Overview

With 1 in 2 US adults estimated to experience a musculoskeletal disorder [9], improving musculoskeletal care must be a national, if not global, priority. We innovated our digital initial clinical evaluation to bring both the physician or nurse practitioner and physical therapist together simultaneously with the patient for the duration of the telemedicine evaluation. More specifically, both clinicians were together with the patient for the entirety of the first digital visit. This allows the patient to provide 1 history and the clinicians to collaboratively perform 1 digital physical examination. The physician or nurse practitioner renders a medical diagnosis, and the physical therapist provides a functional diagnosis, with the opportunity for collaboration between both clinicians during the visit to drive clinician concordance. The resultant treatment plan is aligned with the patient's values and goals (ie, "What matters to me") [17], and the patient can ask clarifying questions to both clinicians. Our clinicians subjectively commented that this integrated, combined visit was a more satisfying and collaborative experience for them.

Principal Findings

Our survey results showed a very high level of positive responses from patients. Patients felt that this visit format saved them time (179/192, 93.2%), allowed them to start their treatment plan faster (174/192, 90.6%), and was enjoyable (168/192, 87.5%). Moreover, 92.2% (177/192) of patients responded that the visit increased their confidence with understanding their medical condition and how to start treating it. Improved patient understanding of their medical condition and associated treatment positively affects health outcomes [18]. Effective patient education supports an atmosphere of trust between patient and clinician, empowers patients to participate in their own health care, and supports informed decision-making [19].

We postulate that the high patient confidence in understanding their medical condition and how to start treating it was influenced by our MD+PT visit format. Our model was structured to create diagnostic and therapeutic agreements between clinicians. The model also provided patients with the benefit, real or perceived, of having both a musculoskeletal physician or nurse practitioner and physical therapist together with them in the clinical encounter. While direct access physical therapy has been shown to be a safe, less expensive, and effective model for triage of patients with musculoskeletal conditions compared to a required referral from a physician [20], physical therapists cannot order imaging, prescribe medications, or render a medical diagnosis in the United States. Patients may have a perception of a more complete evaluation when the physician (or nurse practitioner) is evaluating them, even if imaging or prescription medications are not ordered. Patients may also have a higher level of trust when a physician or nurse practitioner is creating the treatment plan with the physical therapist, particularly for patients with complex conditions or comorbidities. Future research comparing the experience of patients being evaluated by a physical therapist versus evaluation by both a physician or nurse practitioner and

XSL•FO
RenderX

physical therapist may provide data on the potential value (or not) of the MD+PT visit.

Research on concordance [21] in diagnosis between physicians and physical therapists shows varied results. In a systemic review of 19 studies involving 1745 patients and a total of 35 advanced practice physical therapists (APPTs), Lafrance et al [22] showed concordance between APPTs and physicians for musculoskeletal disorders to be "probably good to very good for diagnosis" (κ coefficient [21] 0.76) and "good to very good for surgical triage" (κ coefficient 0.71). All studies in this systemic review were conducted outside the United States, and in 5 studies, APPTs had the authority to order diagnostic imaging studies, including magnetic resonance imaging [22]. The ability of some APPTs in this study to order diagnostic imaging limits the application of these results to concordance between a physical therapist and a physician in the United States.

In other research, Madsen et al [23] reported on 69 patients independently evaluated twice on the same day by an orthopedic surgeon and an extended scope physical therapist. Agreement between the surgeon and physical therapist was 62% on the primary diagnosis and 79% on a combination of diagnoses [23]. In 2 small studies, the κ coefficient between APPTs and musculoskeletal surgeons for referral for surgical treatment was 0.46 (shoulder disorders) [24] and 0.69 (lumbar spine pain) [25], suggesting an opportunity for improvement. The experience of the physical therapist is a variable that may influence these outcomes: in a study of 105 patients with carpal tunnel syndrome, physical therapists with advanced experience had 100% agreement with surgeons, whereas early-career physical therapists only had fair agreement [26]. Depressive symptoms were also shown in one study of knee disorders to decrease diagnostic concordance between physical therapists and medical musculoskeletal physicians [27]. While all our clinicians have a least 5 years of musculoskeletal clinical experience, our MD+PT model may mitigate diagnostic discordance when the clinical experience of one of the clinicians is more limited.

We believe that concordance in the treatment plan in our MD+PT model was facilitated by direct and immediate communication between clinicians. In a survey of interprofessional communication between Canadian orthopedic surgeons and physical therapists, 65.6% of clinicians in stand-alone practices report a negative view of communication compared to 48.4% of those in collaborative practice. This same study also found that orthopedic surgeons and physical therapists did not agree on the clinical information that should be shared regarding postoperative rehabilitation [28]. In a study of 600 survey responses of physical therapists treating patients who underwent rotator cuff repair, 33% of physical therapists reported receiving copies of the operative report in <25% of patients, and 16% reported not receiving an operative note >50% of the time. With postoperative therapy protocols based on the size of the rotator cuff tear and the complexity of the repair, this lack of communication could negatively impact clinical care [29].

We also highlight the importance of patients being able to promptly begin their treatment plan. In the traditional model, the time to diagnosis and treatment of a patient often depends upon the sequencing of physician and physical therapist evaluations as well as the potential need for insurance authorization for care, with frequent delays in care. While in many states, patients can access physical therapy (direct PT) for initial treatment with no delay in beginning the physical therapy plan of care, one small study found that 25.6% of patients who began physical therapy and were subsequently referred to a physician had their physician visit at a range of 0.5 to 10 weeks following the initial physical therapy evaluation [30]. We were unable to find data on the length of time for a patient to see the physical therapist after referral from the physician, which is likely dependent on many factors including patient insurance and the need for prior authorization. Delaying the onset of treatment may lengthen the recovery time and negatively impact clinical outcomes [31,32]. Finally, 2 separate clinical visits can result in additional expenses to patients related to travel costs, time lost from work, and disruption of normal activities.

Finally, a very high percentage (168/192, 87.5%) of patients found the MD+PT visit enjoyable. We chose to ask this question as an attempt to gauge the emotional experience of the patient with 1 simple question. We believe that taking evidence-based care to the next level of effectiveness also requires delivering a better patient experience. We are not aware of others using this question to gain insights into the patient experience. The current Consumer Assessment of Healthcare Providers and Systems (CAPHS) survey, a standard instrument for measuring the patient experience, asks patients about various aspects of their experience including whether their visit started on time and if the clinician showed them respect, listened carefully, and gave them enough time. CAPHS asks patients to rate their visit on a scale of 0 (worse visit possible) to 10 (best visit possible) [33]. In our opinion, CAPHS is a very valuable tool but does not query the emotional response of the patient to the encounter. While we recognize "enjoyable" may never be a realistic experience for patients with very serious medical issues, we challenge the health care system to explore this, and similar, questions that assess the emotional response of the patient. We believe that the more enjoyable the clinical experience is for patients, the higher their engagement will be, which will translate into improved clinical outcomes.

Limitations

Our research has several limitations. The main limitation was our survey response rate of 26% (N=195) raising the potential for nonresponder selection bias. Consensus as to an acceptable survey response rate in medical research is lacking [34,35]. Studies show a range of patient response rates to email surveys. In a global overview paper of surgical patients, email studies gave an average response rate of 68%, although the United States was noted to have a lower overall response rate of 64.2% [33]. In a primary care practice, the response rate to a postvisit email survey was 60.9%. However, only 45.15% of patients who initially consented to completing a paper waiting room survey agreed to receive an email survey [36].

While these response rates are higher than ours, they may not be directly comparable. Patients who have undergone a surgical

XSL•FO
RenderX

procedure may be more willing to participate in a research study compared to those who had a telemedicine visit. In the primary care practice study cited in the preceding paragraph, the response rate to an email survey was 27.7% (60.9% of the 45.15% of patients who consented to an email survey), which is comparable to our response rate of 26% (N=195). In our study, we did not ask patients for consent to be surveyed, rather we sent the email survey to all eligible patients until 200 patients consented to the study.

In studies with methodology more aligned with ours, our survey response rate is favorable. A 2016 primary care study showed a 19.8% response rate to a 6-question email survey to patients on access to care and patient centeredness [37]. Another publication in 2018 reported 28% of patients opening an email containing a patient experience survey, with 80% completing the survey [38]. In a 2024 poster on patient satisfaction and outcomes, an email survey of patients having a genetic information session for kidney disease via telephone had a response rate of 12.1%, although the survey was much longer at 47 questions [39].

We did not conduct a corresponding survey of our clinicians using this model to potentially identify episodes of initial discordance in clinical assessment and treatment plan (and how these were resolved). All patients received a personalized, evidence-based, biopsychosocial treatment plan supported by the physician or nurse practitioner and physical therapist at the completion of the initial visit. Health coaches and registered dietitians were brought into the clinical team to support the patient as appropriate. We also did not survey our clinicians on their perception of whether communication between clinicians was facilitated with this model or whether the clinicians found this model to be enjoyable for them. We did not correlate the results of this survey with clinical outcomes or health care use; these are areas we plan to study in the future.

Patient's perception of the quality of care provided by each clinician, or the quality of care provided by the MD+PT visit, was not studied and is an area for potential future research. However, given that patients reported high confidence in their medical plan with the MD+PT visit, we postulate that this is a surrogate for patients' perceptions of a higher quality of care. In the literature, one study found that patients with osteoarthritis reported that physical therapists provided better information about how to take care of their condition, while surgeons provided patients with greater involvement in the decision-making process and facilitated a greater degree of expectations being met [40]. When compared to primary care clinicians, another study found that patient perception of physical therapy–led care was superior in all quality aspects evaluated [41].

Our model requires the resources of both a physician or nurse practitioner and physical therapist in the initial evaluation. Direct access physical therapy has been shown to be clinically effective and resource-sensitive, as many patients with musculoskeletal conditions will not require a physician or nurse practitioner assessment [11]. The scope of our study did not support any type of cost analysis of our model versus direct access physical therapy. We believe that patients with low-complexity musculoskeletal issues can be effectively evaluated in a physical therapy first model and those with high complexity are better evaluated in our MD+PT model.

Finally, our study did not directly address whether health literacy was improved in our patients. Health literacy is defined as "the degree to which individuals have the ability to find, understand, and use information and services to inform health-related decisions and actions for themselves and others" [42]. We speculate that patients having increased confidence in their understanding of their condition and how to start treating it may have greater health literacy. As annual health care costs of individuals with low health literacy are 4 times higher than those of higher health literacy individuals, and patients with higher health literacy are more likely to adhere to treatment recommendations [43], we will continue our research to identify the potential impact of our model on health literacy.

Conclusions

We found our innovative model of a musculoskeletal physician or nurse practitioner and physical therapist evaluating the patient with a musculoskeletal condition together (at the same time during the initial telemedicine visit) resulted in a high degree of positive patient feedback. Overwhelmingly, patients reported that this visit model saved them time, allowed them to start their treatment plan faster, was enjoyable, and increased their confidence in understanding their medical condition and how to start treating it. We will continue to seek additional ways to innovate patient-centric care that enhances the patient experience, clinical outcomes, and appropriate health care use.

Acknowledgments

There was no use of generative artificial intelligence in the generation of text, figures, or other informational content of this manuscript.

Data Availability

The datasets generated and analyzed during this study are not publicly available because the public sharing of respondent data was not included in the consent for study participation permission.

XSL•FO
RenderX

Authors' Contributions

MIO wrote and edited the manuscript and contributed to the study conception, methodology, and data analysis. MR curated the data, provided tables for the manuscript, and contributed to the study conception, methodology, and data analysis. CC and JA contributed to the study conception and methodology. KCP contributed to study conception, methodology, and data analysis. CM contributed to methodology and writing the manuscript. TT determined the software used in the study and supported data curation. RAG contributed with writing and editing the manuscript. All authors reviewed the final manuscript.

Conflicts of Interest

All authors are employees of Vori Health Medical Group, PLLC or Vori Health, Inc, and have a financial interest in the company.

References

1. Gill TK, Mittinty MM, March LM. Global, regional, and national burden of other musculoskeletal disorders, 1990-2020, and projections to 2050: a systematic analysis of the Global Burden of Disease Study 2021. Lancet Rheumatol 2023 Nov;5(11):e670-e682. [doi: 10.1016/S2665-9913(23)00232-1] [Medline: 37927903]
2. Unnecessary back surgery: older Americans put at risk while billions in Medicare funds wasted. Lown Institute. URL: https://lownhospitalsindex.org/unnecessary-back-surgery [accessed 2024-12-23]
3. Riddle DL, Jiranek WA, Hayes CW. Use of a validated algorithm to judge the appropriateness of total knee arthroplasty in the United States: a multicenter longitudinal cohort study. Arthritis Rheumatol 2014 Aug;66(8):2134-2143. [doi: 10.1002/art.38685] [Medline: 24974958]
4. Jacobs JC, Jarvik JG, Chou R, et al. Observational study of the downstream consequences of inappropriate MRI of the lumbar spine. J Gen Intern Med 2020 Dec;35(12):3605-3612. [doi: 10.1007/s11606-020-06181-7] [Medline: 32989711]
5. Khan MM, Pincher B, Pacheco R. Unnecessary magnetic resonance imaging of the knee: how much is it really costing the NHS? Ann Med Surg (Lond) 2021 Oct;70:102736. [doi: 10.1016/j.amsu.2021.102736]
6. Sogbein OA, Chen AG, McClure JA, et al. Unnecessary interventions for the management of hip osteoarthritis: a population-based cohort study. Can J Surg 2024;67(4):E300-E305. [doi: 10.1503/cjs.001624] [Medline: 39089818]
7. Shelton S, Yochum A DO, Walker E, Waters N. Effectiveness of epidural steroid injections for low back pain. Am Fam Physician 2022 Jul;106(1):89-90. [Medline: 35839356]
8. Hastings CS, Schilthuis AJ, Jackson CD. Physical therapy is better than glucocorticoid injections for osteoarthritis of the knee. J Gen Intern Med 2021 Oct;36(10):3301-3303. [doi: 10.1007/s11606-021-06870-x] [Medline: 34322835]
9. National Academies of Sciences, Engineering, and Medicine; Health and Medicine Division; Board on Health Care Services; Committee on Identifying Disabling Medical Conditions Likely to Improve with Treatment. Selected Health Conditions and Likelihood of Improvement with Treatment: National Academies Press; 2020.
10. Wu V, Goto K, Carek S, et al. Family medicine musculoskeletal medicine education: a CERA study. Fam Med 2022 May;54(5):369-375. [doi: 10.22454/FamMed.2022.975755]
11. APTA report: use of direct access among PTs is widespread, but barriers need to be addressed. American Physical Therapy Association. 2017. URL: https://www.apta.org/article/2017/10/20/apta-report-use-of-direct-access-among-pts-is-widespread-but-barriers-need-to-be-addressed [accessed 2025-02-12]
12. Haleem A, Javaid M, Singh RP, Suman R. Telemedicine for healthcare: capabilities, features, barriers, and applications. Sens Int 2021;2:100117. [doi: 10.1016/j.sintl.2021.100117] [Medline: 34806053]
13. Laskowski ER, Johnson SE, Shelerud RA, et al. The telemedicine musculoskeletal examination. Mayo Clin Proc 2020 Aug;95(8):1715-1731. [doi: 10.1016/j.mayocp.2020.05.026] [Medline: 32753146]
14. Vincent R, Charron M, Lafrance S, Cormier AA, Kairy D, Desmeules F. Investigating the use of telemedicine by health care providers to diagnose and manage patients with musculoskeletal disorders: systematic review and meta-analysis. J Med Internet Res 2024 Sep 23;26:e52964. [doi: 10.2196/52964] [Medline: 39312765]
15. Bargeri S, Castellini G, Vitale JA, et al. Effectiveness of telemedicine for musculoskeletal disorders: umbrella review. J Med Internet Res 2024 Feb 2;26:e50090. [doi: 10.2196/50090] [Medline: 38306156]
16. Vori Health. URL: https://www.vorihealth.com [accessed 2024-12-23]
17. Barry MJ, Edgman-Levitan S. Shared decision making—pinnacle of patient-centered care. N Engl J Med 2012 Mar 1;366(9):780-781. [doi: 10.1056/NEJMp1109283] [Medline: 22375967]
18. Simonsmeier BA, Flaig M, Simacek T, Schneider M. What sixty years of research says about the effectiveness of patient education on health: a second order meta-analysis. Health Psychol Rev 2022 Sep;16(3):450-474. [doi: 10.1080/17437199.2021.1967184] [Medline: 34384337]
19. AAFP core educational guidelines: recommended core educational guidelines for family practice residents. American Academy of Family Physicians. URL: https://www.aafp.org/pubs/afp/issues/2000/1001/p1712.html [accessed 2024-12-23]
20. Gallotti M, Campagnola B, Cocchieri A, Mourad F, Heick JD, Maselli F. Effectiveness and consequences of direct access in physiotherapy: a systematic review. J Clin Med 2023 Sep 7;12(18):5832. [doi: 10.3390/jcm12185832] [Medline: 37762773]
21. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22(3):276-282. [Medline: 23092060]

XSL•FO
RenderX

22. Lafrance S, Vincent R, Demont A, Charron M, Desmeules F. Advanced practice physiotherapists can diagnose and triage patients with musculoskeletal disorders while providing effective care: a systematic review. J Physiother 2023 Oct;69(4):220-231. [doi: 10.1016/j.jphys.2023.08.005] [Medline: 37714771]
23. Madsen MN, Kirkegaard ML, Klebe TM, et al. Inter-professional agreement and collaboration between extended scope physiotherapists and orthopaedic surgeons in an orthopaedic outpatient shoulder clinic—a mixed methods study. BMC Musculoskelet Disord 2021 Jan 4;22(1):4. [doi: 10.1186/s12891-020-03831-z] [Medline: 33397350]
24. Lowry V, Bass A, Lavigne P, et al. Physiotherapists' ability to diagnose and manage shoulder disorders in an outpatient orthopedic clinic: results from a concordance study. J Shoulder Elbow Surg 2020 Aug;29(8):1564-1572. [doi: 10.1016/j.jse.2019.11.030] [Medline: 32199757]
25. Robarts S, Stratford P, Kennedy D, Malcolm B, Finkelstein J. Evaluation of an advanced-practice physiotherapist in triaging patients with lumbar spine pain: surgeon-physiotherapist level of agreement and patient satisfaction. Can J Surg 2017 Aug;60(4):266-272. [doi: 10.1503/cjs.013416] [Medline: 28730987]
26. Lewis KJ, Coppieters MW, Vicenzino B, Hughes I, Ross L, Schmid AB. Occupational therapists, physiotherapists and orthopaedic surgeons agree on the decision for carpal tunnel surgery. Int J Health Policy Manag 2022 Jul 1;11(7):1001-1008. [doi: 10.34172/ijhpm.2020.227] [Medline: 33590739]
27. Lowry V, Bass A, Vukobrat T, et al. Higher psychological distress in patients seeking care for a knee disorder is associated with diagnostic discordance between health care providers: a secondary analysis of a diagnostic concordance study. BMC Musculoskelet Disord 2021 Jul 30;22(1):650. [doi: 10.1186/s12891-021-04534-9] [Medline: 34330250]
28. Longstaffe R, Slade Shantz J, Leiter J, Peeler J. Surgeon-therapist communication: do all members see eye-to-eye? Phys Sportsmed 2015 Nov;43(4):381-387. [doi: 10.1080/00913847.2015.1077096] [Medline: 26309155]
29. Schultzel M, Scheidt KB, McNeill B, Klein CM, Blout C, Itamura JM. Surgeon-therapist communication must be improved in rotator cuff repair rehabilitation: an electronic survey of physical therapists on postoperative rehabilitation protocols and communication with treating surgeons. Perm J 2021 May;25:20.088. [doi: 10.7812/TPP/20.088] [Medline: 33970072]
30. Boissonnault WG, Ross MD. Physical therapists referring patients to physicians: a review of case reports and series. J Orthop Sports Phys Ther 2012 May;42(5):446-454. [doi: 10.2519/jospt.2012.3890] [Medline: 22282166]
31. Ojha HA, Wyrsta NJ, Davenport TE, Egan WE, Gellhorn AC. Timing of physical therapy initiation for nonsurgical management of musculoskeletal disorders and effects on patient outcomes: a systematic review. J Orthop Sports Phys Ther 2016 Feb;46(2):56-70. [doi: 10.2519/jospt.2016.6138] [Medline: 26755406]
32. Deslauriers S, Déry J, Proulx K, et al. Effects of waiting for outpatient physiotherapy services in persons with musculoskeletal disorders: a systematic review. Disabil Rehabil 2021 Mar;43(5):611-620. [doi: 10.1080/09638288.2019.1639222] [Medline: 31304824]
33. CAHPS Clinician & Group Survey. URL: https://www.ahrq.gov/sites/default/files/wysiwyg/cahps/surveys-guidance/cg/adult-eng-cg40-3351a.pdf [accessed 2024-12-23]
34. Meyer VM, Benjamens S, Moumni ME, Lange JFM, Pol RA. Global overview of response rates in patient and health care professional surveys in surgery: a systematic review. Ann Surg 2022 Jan 1;275(1):e75-e81. [doi: 10.1097/SLA.0000000000004078] [Medline: 32649458]
35. Eysenbach G. Improving the quality of web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004;6(3):e34. [doi: 10.2196/jmir.6.3.e34]
36. Johnston S, Hogg W, Wong ST, Burge F, Peterson S. Differences in mode preferences, response rates, and mode effect between automated email and phone survey systems for patients of primary care practices: cross-sectional study. J Med Internet Res 2021 Jan 11;23(1):e21240. [doi: 10.2196/21240] [Medline: 33427675]
37. Slater M, Kiran T. Measuring the patient experience in primary care: comparing e-mail and waiting room survey delivery in a family health team. Can Fam Physician 2016 Dec;62(12):e740-e748. [Medline: 27965350]
38. Evaluation of the primary care patient experience survey tool. Te Tāhū Hauora Health Quality & Safety Commission. URL: https://www.hqsc.govt.nz/assets/Our-data/Publications-resources/MoH-PES-report-18April2018_2.pdf [accessed 2024-12-23]
39. Hager MM, Stein QP, Hendricks E, et al. Patient satisfaction and outcomes following a session with a genetic counselor for kidney diseases. J Am Soc Nephrol 2024;35(10S). [doi: 10.1681/ASN.2024w7s5gcnj]
40. Gustavsson L, Mohaddes M, Samsson K, Beischer S. No major difference in perceived quality of care in patients with hip or knee osteoarthritis assessed in a physical therapy-led triage compared with standard care: a randomized controlled trial. BMC Musculoskelet Disord 2023 Jun 29;24(1):530. [doi: 10.1186/s12891-023-06659-5] [Medline: 37386368]
41. Samsson KS, Bernhardsson S, Larsson MEH. Perceived quality of physiotherapist-led orthopaedic triage compared with standard practice in primary care: a randomised controlled trial. BMC Musculoskelet Disord 2016 Jun 10;17(17):257. [doi: 10.1186/s12891-016-1112-x] [Medline: 27286829]
42. Health literacy fact sheets. Center for Health Care Strategies. URL: https://www.chcs.org/resource/health-literacy-fact-sheets [accessed 2024-12-22]
43. Health literacy: statistics at-a-glance. AAACEUs. URL: https://www.aaaceus.com/courses/nl0610/article1.pdf [accessed 2024-12-22]

XSL•FO
RenderX

Abbreviations

APPT: advanced practice physical therapist
CAPHS: Consumer Assessment of Healthcare Providers and Systems
MD+PT: musculoskeletal medical doctor and physical therapist

Edited by A Mavragani; submitted 22.09.24; peer-reviewed by M Heng, R Vincent; revised version received 22.01.25; accepted 23.01.25; published 03.03.25.

Please cite as:
O'Connor MI, Chudy C, Peters KC, Ribaudo M, McCulloch C, Aguilar J, Taylor T, Grant RA
Patients' Experience With Evaluation by Both a Musculoskeletal Physician and Physical Therapist in the Same Digital Visit: Survey Study
JMIR Form Res 2025;9:e66744
URL: https://formative.jmir.org/2025/1/e66744
doi:10.2196/66744

© Mary I O'Connor, Carolyn Chudy, Kaitlyn C Peters, Megan Ribaudo, Carrie McCulloch, Jared Aguilar, Trista Taylor, Ryan A Grant. Originally published in JMIR Formative Research (https://formative.jmir.org), 3.3.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

App- and Wearable-Based Remote Monitoring for Patients With Myasthenia Gravis and Its Specialists: Feasibility and Usability Study

Maike Stein[1,2,3,4], MD; Regina Stegherr[5], PhD, Dr rer nat; Pushpa Narayanaswami[3], MD; David Legg[1], MSc; Meret Herdick[1,4], MD; Andreas Meisel[1,4,6], MD; Lea Gerischer[1,2,4*], MD; Sophie Lehnerer[1,2,4,6*], MD

[1]Department of Neurology with Experimental Neurology, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Charitéplatz 1, Berlin, Germany

[2]Digital Health Center, Berlin Institute of Health at Charité, Charité Universitätsmedizin Berlin, Berlin, Germany

[3]Department of Neurology, Beth Israel Deaconess Medical Center/Harvard Medical School, Boston, MA, United States

[4]Neuroscience Clinical Research Center, Department of Neurology with Experimental Neurology, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany

[5]Institute of Biometry and Clinical Epidemiology, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany

[6]Center for Stroke Research Berlin, Charité Universitätsmedizin Berlin, Berlin, Germany

[*]these authors contributed equally

Corresponding Author:
Maike Stein, MD
Department of Neurology with Experimental Neurology, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Charitéplatz 1, Berlin, Germany

Abstract

Background: Myasthenia gravis (MG) is rare, chronic autoimmune disorder of the neuromuscular junction that requires specialized care and often lifelong treatment, facing challenges due to its rarity and the limited availability of specialists. Telemedical solutions in specialized centers hold considerable promise in bridging this gap by increasing access to this care to a broader patient population in a timely manner. However, there is no research regarding interventional remote care solutions in the field of MG to date.

Objective: This study aimed to assess the feasibility and usability among patients with MG and specialists of a telemedicine platform, tailored to patients with MG and designed to facilitate remote monitoring, treated in a specialized center.

Methods: The telemedicine platform consisted of an app for patients and a web-based portal for physicians. Over a period of 3 months, 30 patients continuously monitored their vital parameters through external devices, including a digital spirometer and a wearable (activity tracker). Furthermore, patients completed 7 different patient-reported outcome measures (PROMs) through the app at predefined intervals. Specialists could review this monitoring data and adjust therapy as necessary. In addition, communication between patients and physicians was facilitated through a chat module. Feasibility was evaluated by total adherence rates for completing PROMs within the app, alongside the collection of spirometry and wearable data. Furthermore, user satisfaction was assessed among both patients with MG and physicians at the end of study.

Results: Total adherence rates ranged from 74.3% (1830/2464) to 97.9% (327/334) across all data types, with the highest adherence observed for PROMs (1139/1179, 96.6%), followed by spirometry (293/334, 87.7%) and wearables (1830/2261, 80.9%). Notably, patients wore the wearable longer than required by protocol and conducted a higher number of spirometry measurements during the study than required per protocol (median 20 h/d [IQR 15-24] vs 14 h/d and median 49 [IQR 15-59] measurements vs 11 measurements, respectively). Technical issues and discomfort with wearables were factors affecting lower adherence in some patients. The System Usability Scale yielded a median score of 85 indicating “excellent usability.” In addition, results from a more detailed user evaluation questionnaire showed high levels of user satisfaction among both patients and health care professionals across diverse categories, including their experience of the care program, communication, and evaluation of the program.

Conclusions: Remote monitoring of patients with MG through the telemedical platform demonstrated good feasibility and acceptability, as evidenced by above-average adherence rates and user satisfaction for both patients and physicians. The majority of patients wanted to continue using the app. These findings highlight the potential for user-friendly digital tools to enhance care for patients with MG, although addressing technical challenges and ensuring patient comfort with wearables are essential for

XSL•FO
RenderX

optimal implementation. Further research involving larger cohorts and longer study duration is warranted to validate these findings.

Trial Registration: Deutsches Register Klinischer Studien DRKS00029907; https://drks.de/search/en/trial/DRKS00029907

(JMIR Form Res 2025;9:e58266) doi:10.2196/58266

KEYWORDS

myasthenia gravis; myasthenia; remote monitoring; PROMs; digital platform; wearables; telemedicine; spirometry; app; usability; feasibility; autoimmune disorder; web-based portal; activity tracker; communication; wearable data; digital tool; mobile phone

Introduction

Myasthenia gravis (MG) is a rare autoimmune disease characterized by specific autoantibodies targeting the postsynaptic membrane of the neuromuscular junction, resulting in fluctuating fatigability and weakness of ocular, bulbar, and skeletal muscles, with the potential for life-threatening crises [1]. MG has a bimodal incidence pattern, with younger women and older men being most commonly affected [2]. Given its chronic nature, the majority of patients with MG require long-term and often lifelong, highly specialized care and therapy [3]. Furthermore, studies have shown that patients with MG experience a significantly reduced quality of life [4] and have a high demand for individualized counseling on various topics [5] and supportive interactions. In addition, characteristic symptom fluctuations are not captured effectively by conventional clinical assessments that rely on infrequent in-person assessments that are further challenged due to documentation overload for physicians, limited consultation time, and geographical barriers that limit access to specialists [6] . The characteristic symptom fluctuations in MG, but also the long intervals for therapies to effectively influence the disease including potential side effects, give rise to questions and challenges for timely therapy adjustments and individual counseling. However, in practice, these adjustments often do not occur in a timely manner or are conducted in ways that pose data privacy problems, are resource intensive and lack standardization. These challenges and the growing demand for personalized and on-demand available health care have spurred the development of telemedical and wearable devices for remote patient monitoring [7]. Wearable devices such as activity trackers facilitate continuous monitoring of various health parameters and have shown promise in reducing hospitalization costs, improving personalized health management [8-10] and health ‐ related quality of life [11]. Those digital health interventions hold potential for bridging the care gap by offering individualized support based on continuous health data derived from both subjective patient-reported outcome measures (PROMs) and objective wearable and sensor-based data. Although previous remote digital studies have shown the feasibility and usability of leveraging smartphones and wearable technology for assessing clinical changes in different diagnosis and naturalistic settings [12-14], there is to date no data available on telemedical interventions for patients with MG. Generally, successful integration of these technologies into health care relies on patients' active involvement and acceptance, underscoring the importance of digital literacy [15,16]. In the context of remote monitoring technologies, adherence serves as a crucial measure of patient compliance with study protocols (eg, the completion rate of providing data inputs at specified frequencies outlined in the protocol) and is dependent on their willingness and capacity to adopt new behaviors and practices. We hypothesize that the integration of digital tools into clinical care presents a viable approach for remotely monitoring the condition and symptoms of patients with MG. The aim of this study was to evaluate the feasibility and user satisfaction levels among patients with MG and health care professionals with a telemedicine platform specifically tailored to monitor patients with MG and enabling patient-physician interaction.

Methods

Telemedical Platform for Patients with MG

The telemedical platform for patients with MG, "MyaLink" used in the study was developed at Charité Universitätsmedizin Berlin together with software partner Qurasoft GmbH (based on the SaniQ platform). The platform is a Conformité Européenne (CE)-marked class-I medical device (Medical Device Directive) and is GDPR (General Data Protection Regulation) and BlnDSG (Berlin Data Protection Act) compliant. The platform consists of an app for patients (iOS and Android) and a web-based portal for physicians facilitating to review patient data, manage and set up individual monitoring plans and medication plans. It offers a secure interactive platform for exchange of information and health data between physicians and patients and informed treatment decisions (Figure 1). The MyaLink app contains the following modules: (1) transmission and integration of sensor-based data from wearables including heart rate, oxygen saturation, and step count (device: Garmin Vivosmart 5); (2) assessment of the forced vital capacity from a digital spirometer (device: MIR Spirobank Smart One); (3) assessment of questionnaires and PROMs (see details under Section 3.5); and a (4) communication module for individual requests and patient-physician interaction. Furthermore, the app contains a (5) medication plan with reminder function and (6) a module including data storage of important documents (eg, clinical reports), which can also be shared with the physician through the web-based portal. The web-based system for physicians enables real time access to this data and allows clinicians to obtain an accurate overview of the patients' condition and visualization of health data. If necessary, the patient can be contacted through the communication tool and adjustments to medication and the remote monitoring plans can be made.

XSL•FO
RenderX

Figure 1. Schematic overview of telemedical platform and its features (MyaLink): transmission and simultaneous integration of sensor-based data into the app is implemented by bluetooth from coupled external devices (wearable or digital spirometer). In addition, the platform facilitates the collection of patient-reported outcome measures (PROMs) through the app and a communication module enables interaction between patients and physicians. Health data can be shared with specialists who can review monitoring parameters through the web-based system and make therapy adjustments as needed. The data is transmitted between patient and physicians with double encryption (AES256, Advanced Encryption Standard). PROM: patient-reported outcome measure.

Subjects' Eligibility

Inclusion and exclusion criteria have been previously presented in detail in our methods paper [17]. Inclusion criteria included patients with a diagnosis of MG for at least 6 months before study enrollment, aged 18 years and older, and classified according to the Myasthenia Gravis Foundation of America (MGFA) classification status I-IV [18]. Exclusion criteria included patients with somatic, psychiatric, or neurological disorders who were undergoing treatments or were not stabilized under drug therapy that could adversely affect cognitive function. Patients were recruited through the specialized outpatient myasthenia center at Charité Universitätsmedizin Berlin.

Study Design

This analysis focused on data obtained from the telemedical intervention group from a randomized controlled trial (RCT) conducted at our specialized center at Charité Universitätsmedizin Berlin in Germany. The RCT included 45 patients with MG over a 12-week observation period between April and September 2023. Patients were randomized in a 2:1 ratio into the telemedical intervention group (n=30) or the control group (n=15), who received standard of care without telemedical intervention. Randomization was performed according to gender and clinical class stratified by MGFA score at baseline (using the randomizer tool [19]). The primary objective of the study was to investigate the feasibility of the remote monitoring platform in care of patients with MG. Feasibility was measured through adherence rates based on completion of PROMs in the app and data collection from external devices (including a digital spirometer and activity tracker). Furthermore, user satisfaction with the telemedical platform was assessed among patients and health care professionals comprised of the study team (n=5) at the end of the study. We, therefore, exclusively analyzed data from the 30 patients from the intervention group. Secondary objectives of the study (but not part of this analysis) included comparing standard of care with and without additional telemedical intervention. This comparison focused on evaluating clinical endpoints between the control and intervention group. Comprehensive baseline and end-of-study visits were conducted at the study center for both groups and included an in-person clinical assessment of medical history, medication, hospitalizations for MG, care-related questions (eg, accessibility of a specialist or areas of improvement regarding MG treatment), various PROMs and MG-specific clinical assessment through the quantitative myasthenia gravis score (QMGS) [20]. At baseline, the intervention group received a wearable (activity tracker) and a digital spirometer, along with the MyaLink app, which were set up and connected on the patients' smartphone. Instructions on app and external devices usage were provided by the study team. This enabled remote health monitoring throughout the study by active (PROMs and spirometry) and passive (wearable) continuous data collection. Patients were instructed to fill out PROMs and perform spirometry at predefined intervals. They were instructed to wear the wearable during waking hours (at least 14 h per day) throughout the study period and optionally at night. Telemedical check-ups occurred in weeks 4 and 8, comprising the reviewing of remote monitoring data by study physicians and contacting the patients if necessary (eg, changing medication). Patients could also initiate communication through the chat module at any time, with physicians monitoring messages on weekdays during working hours and responding within 24 hours. Patients completed usability questionnaires through the app a week before the end of the study, while health care professionals conducted their questionnaires after study completion. Furthermore, oral feedback from patients was collected and documented during the end-of-study visit. All data from the study visits and all data from the telemedical check-up was documented in REDCap (Research Electronic Data Capture) software (version 13.7.31; Vanderbilt University).

XSL•FO
RenderX

Patient-Reported Outcome Measures and Data From App

In total, 6 questionnaires and 1 test were assigned to patients through the app, administered weekly or monthly, depending on their specific type (Figure 2). Weekly questionnaires included the myasthenia gravis activities of daily living (MG-ADL) profile [21], myasthenia gravis quality of life, revised (MG-QoL15r) [22], and single simple question (SSQ) [23], which are commonly used in MG studies including large approval studies to assess functional status and quality of life. Patients were also instructed to perform a weekly breath count test [24]. Monthly questionnaires included Chalder Fatigue Scale (CFS 11) [25] and Hospital Anxiety and Depression Scale (HADS) [26] to assess common comorbidities in MG. In addition, Patient Acceptable Symptom State (PASS) was assessed monthly [27]. Furthermore, patients completed 2 user satisfaction questionnaires at the end of the study. For broad comparability, patients completed the System Usability Scale (SUS) [28], where patients ranked each of the 10 statements from 1 to 5 based on how much they agreed with the statement. The second questionnaire was adapted from Braun et al [29] and specifically tailored to the MyaLink telemedical platform. It features 17 items categorized in 3 sections and uses a 4-point Likert scale (strongly agree to strongly disagree). Originally formulated in German for publication purposes, it underwent consensus translation into English by 4 medical health care professionals, including 1 native speaker. After the study, the study team completed a modified version of the questionnaire adapted from Braun et al for health care professionals, respectively.

Figure 2. Monitoring plan and frequencies of assessed parameters: Active (PROMs, respiratory function assessment) and passive data (wearable data) were collected during the study period. PROMs data were collected weekly or monthly, depending on the questionnaire. Respiratory function assessment, assessed through breath count test and spirometry, was conducted weekly by the protocol. The collection of wearable data was continuous, with steps aggregated by the hour. PROMs: patient-reported outcome measures; MG-ADL: myasthenia gravis activities of daily living; MG-QoL15r: myasthenia gravis quality of life, revised; SSQ: single simple question; PASS: Patient Acceptable Symptom State; CFS: Chalder Fatigue Scale; HADS: Hospital Anxiety and Depression Scale.

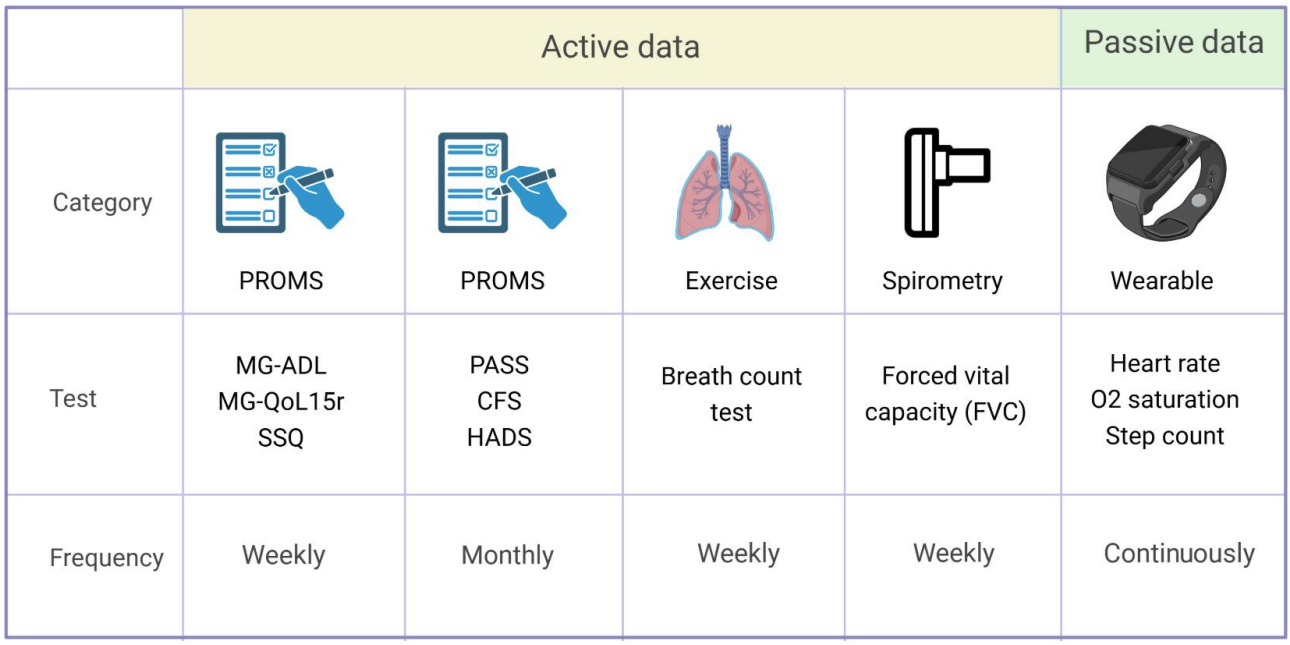

Data From External Device or Wearable

Patients were instructed to perform spirometry measurements weekly, with reminders sent through the app through push notifications. The wearable continuously collected oxygen saturation level (every minute), heart rate (every minute), and steps (hourly aggregated). Data from both the digital spirometer and the wearable were transmitted directly to the app by bluetooth. The study team regularly checked the web-based system to ensure receipt of data from external devices and questionnaires, sending reminder messages through the chat module when necessary.

Analysis

The statistical calculations were performed using R software (version 4.2.2; R Foundation for Statistical Computing) [30]. Descriptive statistics (median [IQR] and absolute and relative frequencies) are presented depending on the scale of the variables.

Feasibility was assessed by adherence rates. For PROMs, assignments of the questionnaire started a week or month after the baseline visit respectively, depending on the questionnaire. Questionnaires were assigned together with a push notification on the patient's phone and disappeared from the app upon completion or were "overwritten" upon receiving a new assignment. Questionnaires could only be submitted by patients through the app if they were fully completed. Each questionnaire was required to be completed within 7 days to be counted as "completed," including the monthly assigned questionnaires, with missing entries calculated accordingly. The final day for data acquisition from the last assigned weekly questionnaires ended the day before the end-of-study visit. Total adherence rates were calculated as the total number of completed questionnaires divided by the total number of assigned

XSL•FO
RenderX

questionnaires. Additional questionnaire assignments were allowed in cases of clinical deterioration but were excluded from adherence calculations. The study was considered successful if at least 60% of assigned PROMs through the app were answered by patients of the intervention group. Across the 30 patients of the intervention group, a mean adherence rate and a 2-sided 95% CI are determined. If the observed average adherence is at least 76%, with data from 30 patients, the 95% CI would range from 60.7% to 91.3%, thus exceeding the 60% threshold.

Spirometry adherence rates were calculated accordingly (ratio of the total number of actual measurements performed during the study vs total number of measurements required per the study protocol).

Adherence calculations for wearable data included measurements starting the day after the baseline visit and ending the day before the end-of-study visit. Adherence to wearable data from the activity tracker was evaluated by aggregating data over hourly windows, with 100% adherence defined as wearing the device for 14 hours per day per patient (based on the estimated waking hours). This was then extrapolated over the study duration. Each hour necessitated at least one heart rate measurement to validate wear time, considering heart rate a more reliable marker than oxygen saturation. Overall adherence was determined by the sum of the hours the patients had with at least one heart rate measurement divided by sum of the hours per day corresponding to 100% individual adherence.

The user satisfaction questionnaire was descriptively evaluated using the Likert scale. The SUS ranked each of the 10 statements from 1 to 5 based on user agreement, from which a SUS score was calculated: Each patient responses undergo conversion into a new numerical value (0 - 4), which is then aggregated and multiplied by 2.5 to transform the original scoring range of 0 - 40 into a 0 - 100 scale. These scores do not represent percentages but a percentile ranking [31]. Research suggests that a SUS score above 68 is considered acceptable, with scores below 68 classified as not acceptable. Therefore, a score above 68 on the SUS is interpreted as a positive result further supporting the platform's usability. A more detailed classification considers SUS scores between 73 and 84 as good, between 85 and 99 as excellent, and 100 as the best.

This analysis focuses on the primary study endpoints assessing feasibility and usability. Further exploratory analyses of secondary endpoints, eg, comparative clinical endpoint analyses (between control group and intervention group), evaluations of communication patterns from the MyaLink chat and telemedical interventions are currently prepared and results will be presented in further studies.

Ethical Considerations

This study received ethics approval by the ethics committee at Charité Universitätsmedizin Berlin (EA2/157/22). Written informed consent was obtained from the study participants before the start of the study. The study was conducted in accordance with the Declaration of Helsinki (2013) and its later amendments. Data storage and handling responsibilities were approved by the ethics committee of Charité Universitätsmedizin Berlin. The data collected through MyaLink is stored on certified data servers, which adhere to a management system according to ISO/IEC 27001:2013.

Results

Patient Characteristics

Patients were between 23 and 82 years (median 47.5 years; IQR 40.5-57.5) at time of enrollment. The majority of patients were female (21/30, 70%). Most patients in the study had a mild to moderate disease severity, with 36.7% (11/30) categorized in MGFA II (a or b) status and 43.3% (13/30) in MGFA III (a or b) status at the time of enrolment. Median scores for MG-ADL and QMG were 10 (IQR 5.25-11) and 15.5 points (IQR 12.25-20.75), respectively. On average, patients were diagnosed 5.5 years (IQR 2-9.75) before entry in the study. One patient withdrew from the study due to health reasons unrelated to MG. All results presented include data from the dropout, including measurements taken before the time of dropout (with the patient's consent). Detailed patient characteristics are presented in Table 1.

XSL•FO
RenderX

Table . Patient characteristics: age, clinical class, time since diagnosis, study duration, and operating system for intervention group as well as subgroups (female and male).

Characteristics	Overall (n=30)	Female (n=21)	Male (n=9)
Age, median (IQR)	47.5 (40.5 - 57.5)	47 (42 - 53)	56 (40 - 60)
MGFA[a] score, n (%)			
I	1 (3.3)	0 (0)	1 (11.1)
IIa	5 (16.7)	4 (19)	1 (11.1)
IIb	6 (20)	3 (14.3)	3 (33.3)
IIIa	3 (10)	2 (9.5)	1 (11.1)
IIIb	10 (33.3)	8 (38.1)	2 (22.2)
IVa	0 (0)	0 (0)	0 (0)
IVb	5 (16.7)	4 (19)	1 (11.1)
MG-ADL[b] score, median (IQR)	10 (5.25 - 11)	10 (5 - 11)	9 (6-12)
MG-QoL15r[c] score, median (IQR)	15.5 (12.25 - 20.75)	17 (12 - 21)	13 (13 - 19)
QMG[d] score, median (IQR)	15 (12.25 - 17)	15 (13 - 18)	14 (10 - 15)
Years since MG[e] diagnosis, median (IQR)	5.5 (2 - 9.75)	5 (2-10)	7 (4-9)
System, n (%)			
Android	19 (63.3)	15 (71.4)	4 (44.4)
iOS	11 (36.7)	6 (28.6)	5 (55.6)
Days between baseline and end-of-study visit, median (IQR)	84 (84 - 85)	84 (83 - 85)	84 (84 - 86)
Dropout, n (%)	1 (3.3)	1 (4.8)	0 (0)

[a]MGFA: Myasthenia Gravis Foundation of America.

[b]MG-ADL: myasthenia gravis activities of daily living.

[c]MG-QoL15r: myasthenia gravis quality of life, revised.

[d]QMG: quantitative myasthenia gravis.

[e]MG: myasthenia gravis.

Adherence

The total adherence rates for PROMs and external devices data ranged from 74.3%- 97.9% (1830/2464 - 327/334) across all data types and are detailed in Table 2. When examining adherence by data category, adherence to PROMs was the highest, followed by spirometry, and then wearable data from the activity tracker.

XSL•FO
RenderX

Table . Total adherence rates for different data groups.

Data group	Adherence (total completed/total assigned), n/N (%)
Overall PROMs[a]	1139/1179 (96.6)
MG-ADL[b]	321/334 (96.1)
MG-QoL15r[c]	325/334 (97.3)
SSQ[d]	327/334 (97.9)
CFS[e]	55/59 (93.2)
HADS[f]	5/59 (93.2)
PASS[g]	56/59 (94.9)
Breath count test	326/334 (97.6)
Spirometry	
FVC[h]	293/334 (87.7)
Wearables (at least 14 h per day)	
Pulse	1830/2464 (74.3)
Pulse (excluding days without any wearable measurement from the analysis)	1830/2261 (80.9)

[a]PROMS: patient-reported outcome measures.
[b]MG-ADL: myasthenia gravis activities of daily living.
[c]MG-QoL15r: myasthenia gravis quality of life, revised.
[d]SSQ: single simple question.
[e]CFS: Chalder Fatigue Scale.
[f]HADS: Hospital Anxiety and Depression Scale.
[g]PASS: Patient Acceptable Symptom State.
[h]FVC: forced vital capacity.

Adherence to PROMs ranged between 93.2% (55/59) and 97.9% (327/334), depending on the specific questionnaire. Patients were assigned weekly questionnaires and spirometry measurements between 11 and 13 times (drop out: 5 times) during the study, depending on individual study duration of a participant. Monthly questionnaires were assigned twice to all patients (drop out: 1). There was no difference in adherence regarding the questionnaire frequency. For the breath count test, adherence was 97.6% (326/334). Adherence to spirometry was slightly lower than PROMs (293/334, 87.7%), likely due to the higher task burden associated with these active measurements (Figure 3). Patients conducted a median of 49 spirometry measurements (IQR 15-59).

XSL•FO
RenderX

Figure 3. Adherence of spirometry measurements: The blue circles represent the end-of-study visit for each patient, while the red line indicates week 12, marking the duration of the study. Each black vertical line indicates when a spirometry measurement was assigned according to the study protocol (weekly intervals). Patients were given a 7-day window until the next weekly assignment for measurement to qualify as "completed." The black dots mark the times when the patient actually performed the spirometry measurement. Each measurement from the digital spirometer was timestamped and subsequently automatically transmitted to the app through bluetooth, with the values being integrated accordingly. FVC: forced vital capacity.

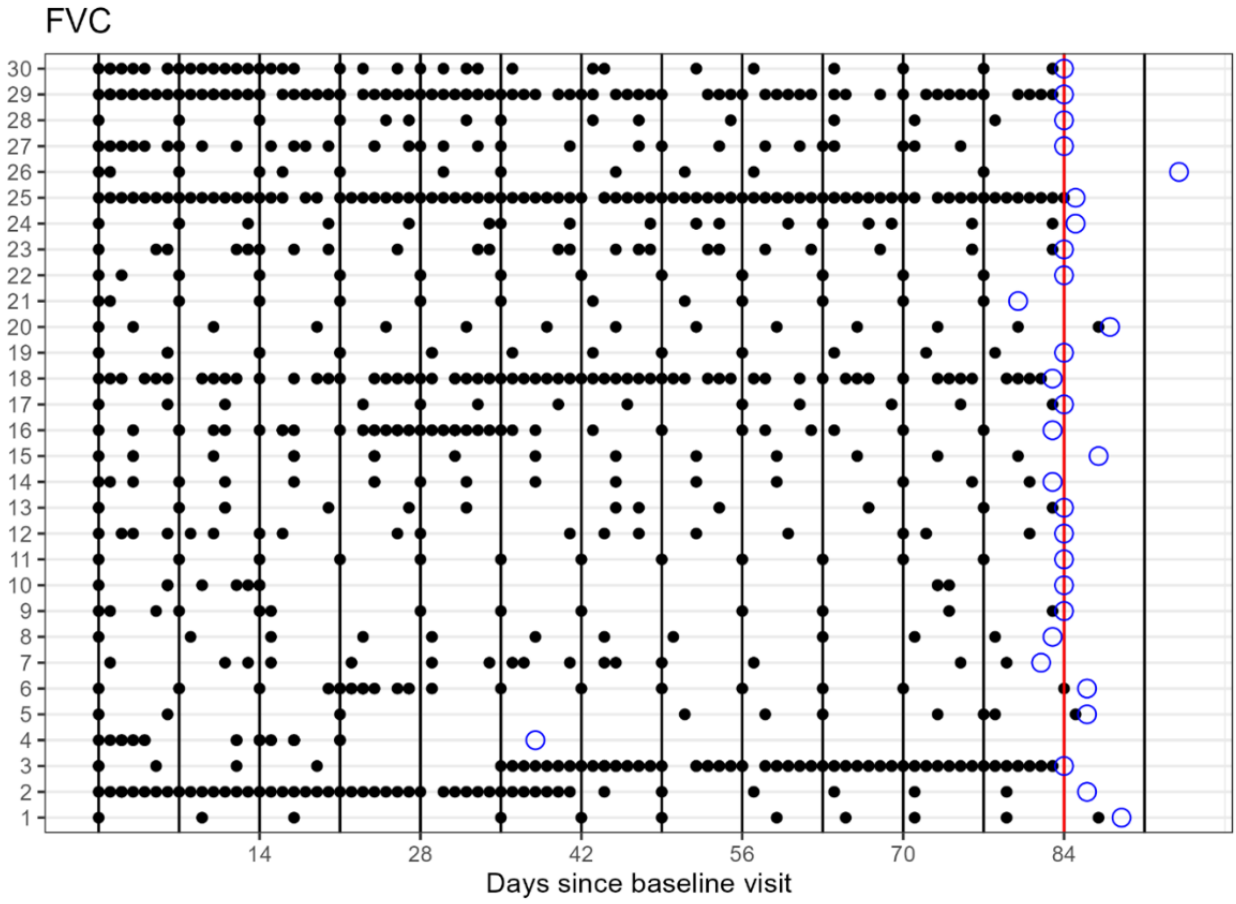

On average, patients wore the wearable for a median duration of 20 hours per day (IQR 15 - 24). Wearable adherence rates exhibited considerable variability among patients, ranging from 6.8% (4/59) to 100% (83/83) (Figure 4), despite an overall strong adherence of 80.9% (1830/2261) for all patients during the study when excluding days without any data input (no recorded heart rate measurements) from the calculation (Table 2). Including days without wearable data input to the calculation decreased adherence to 74.3% (1830/2464). Clinical insights from chat interactions, as evidenced by the timestamps within the chat messages, unveiled several factors impacting patients with decreased device use. For example, out of the 30 patients, 2 patients (6.7%) experienced a myasthenic crisis necessitating prolonged hospitalization and reported having forgotten the devices at home. Furthermore, 1 patient (3.3%) had difficulty with using the app (older patient who was overwhelmed by the technical operation and conducting measurements with the digital spirometer) with numerous interactions with the software provider's support team. Some patients experienced discomfort with the rubber strap of the activity tracker. Some patients had device-related issues such as (recurring) pairing problems and 1 out of the 30 patients (3.3%) received a defective activity tracker, necessitating a device replacement. However, even after replacement the patient reported ongoing connectivity issues between the wearable and his smartphone.

XSL•FO
RenderX

Figure 4. Adherence rates for wearable data from the activity tracker: Days since baseline visit on the X-axis, with patients represented on the Y-axis and the adherence rate per patient in the column on the right side. Days with wearable usage of at least 14 hours per day were classified as compliant (marked in green). A wearable hour was counted if at least one heart rate measurement value was recorded per hour. Days with less than 14 hours of wearable usage per day were marked as red. Days without any recorded heart rate measurements were marked in gray and excluded from the analysis.

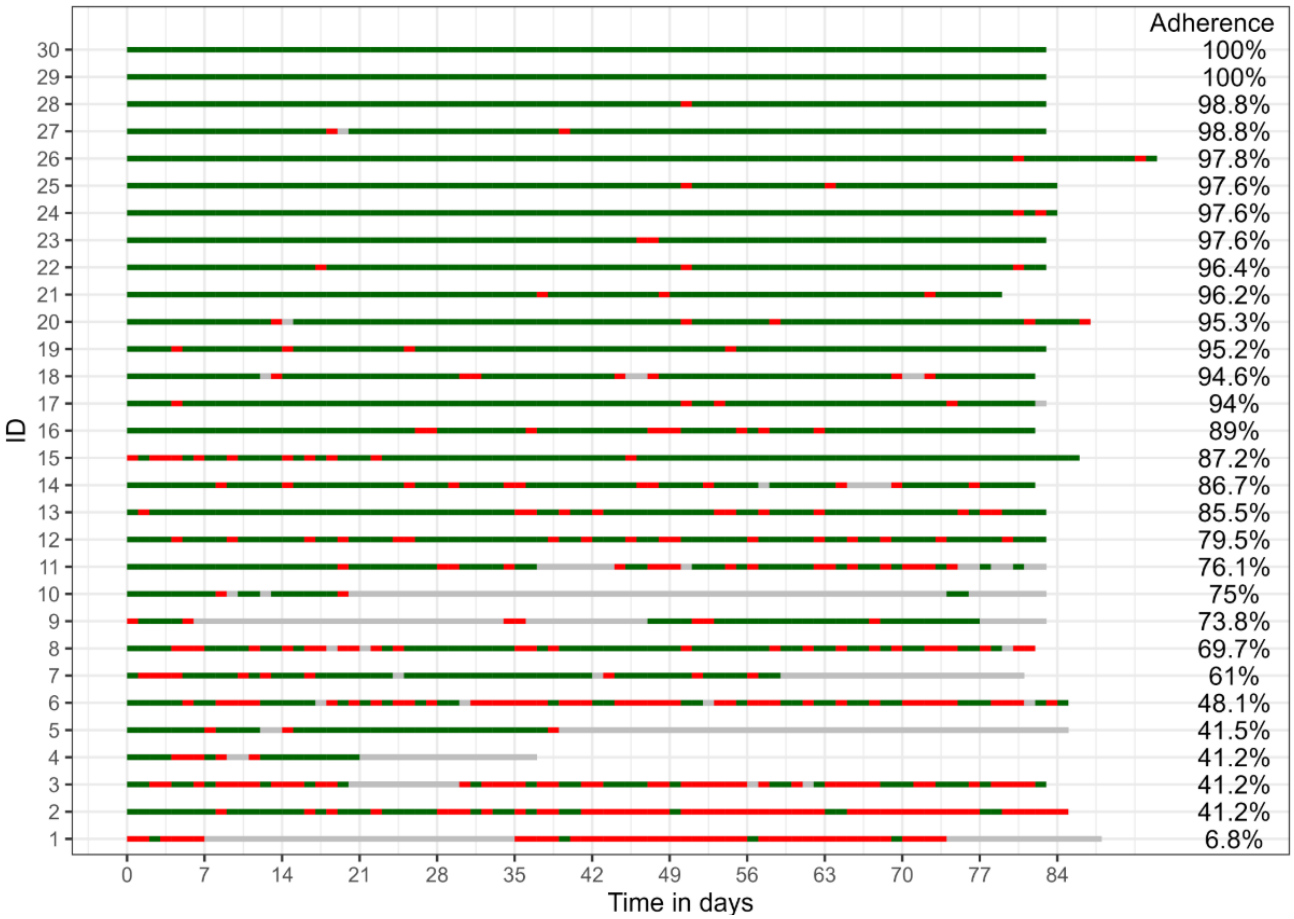

User Satisfaction

The highest levels of agreement were observed in the following order: experience with the care program, communication, and evaluation of the program. Regarding the care program evaluation, 89.7% (26/29) agreed that the app was easy to use while 93.1% (27/29) agreed that the time investment required for app usage was appropriate. In total, 89.6% (26/29) of patients agreed that describing their health status during follow-up appointments was easier. A total of 93.1% (27/29) indicated that treating physicians were interested in reviewing the collected monitoring parameters. In addition, 79.3% (23/29) of patients agreed that their quality of life had significantly improved since app usage. Futhermore, 96.6% (28/29) expressed willingness to recommend the app to other patients. In addition, 69% (20/29) agreed that they felt more confident in identifying MG-specific problems using the app. Views on the reduction of visits to neurologists or clinics were mixed; 55.2% (16/29) agreed to needing fewer visits to neurologists, while 51.7% (15/29) agreed to reduced visits to clinics, with 13.8% (4/29) strongly disagreeing with both categories (Figure 5). Out of the 29 patients, 14 patients (46.7%) responded that they did not agree that they required less frequent hospital visits; of those 14 patients, 8 patients (57.1%) were hospitalized during the study period due to MG (two of whom were in intensive care). In addition, 7 of these 14 patients (50%) experienced at least one exacerbation. Results of user evaluation from the health care professionals showed similar high acceptance rates regarding the telemedical platform, ease of use, and overall positive evaluation on the effect on patient care (Figure 6).

XSL•FO
RenderX

Figure 5. User evaluation for patients: The questionnaire was divided into 3 categories: "Experience with the care program," "Communication and Interaction" and "Evaluation of the program." All 17 questions could be answered using a 4-point Likert scale ranging from 1 (strongly agree) to 4 (strongly disagree).

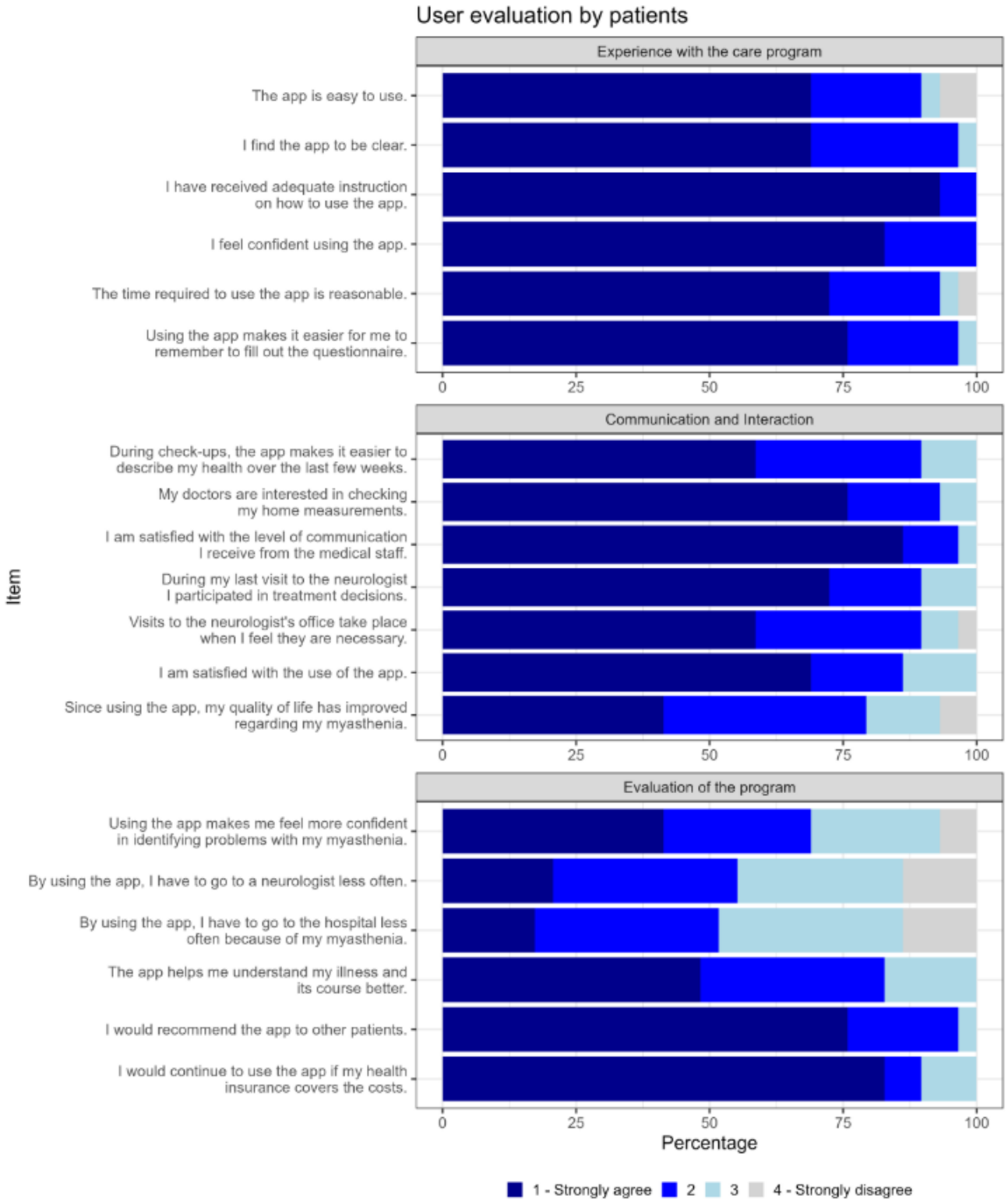

XSL•FO
RenderX

Figure 6. User evaluation for health care professionals: The questionnaire was divided into 3 categories: "Experience with the care program," "Communication and Interaction," and "Evaluation of the program." All 17 questions could be answered using a 4-point Likert scale ranging from 1 (strongly agree) to 4 (strongly disagree). Depending on the role within the study team (eg, physician and study nurse), only certain questions in the questionnaire could be answered.

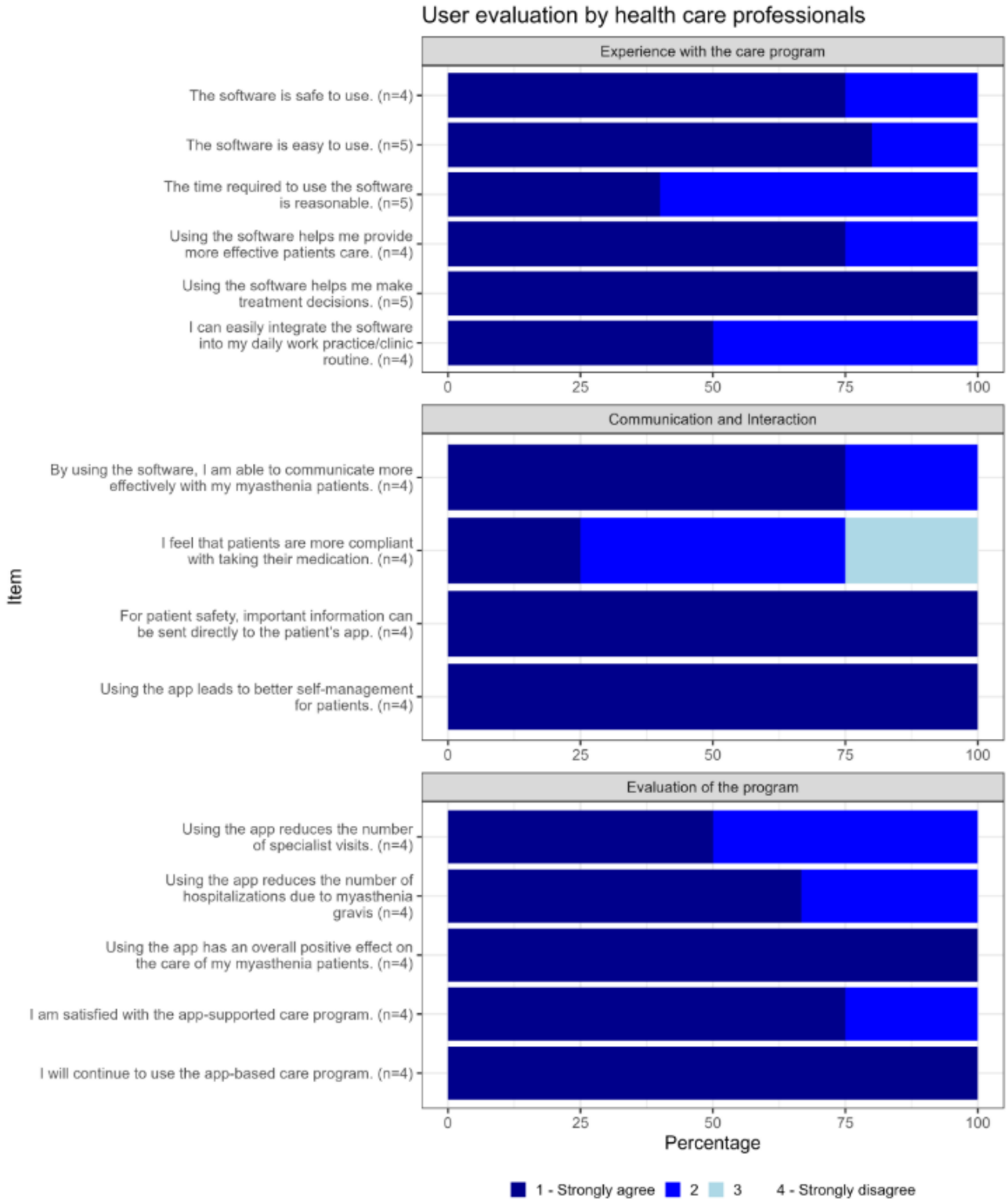

The overall median SUS score was 85 (IQR 75-90). Figure 7 shows that there was a high level of agreement regarding the system's ease of use (26/29, 90%), patients' confidence in operating it (25/29, 86%), and patients' belief that others would quickly learn to use the app (25/29, 86%). Notably, 21% (6/29) remained neutral regarding perceived inconsistencies in the app, although the majority (20/29, 69%) tended to disagree. Among patients (N=22) who responded at least once with a neutral (Likert-scale 3) or negative response to any question (agreeing or strongly agreeing with negatively formulated questions or disagreeing or strongly disagreeing with positively formulated questions), there were 14 females (63.6%) with a median age of 47.5 (IQR 39.25-61.25; mean 49.7, SD 15.70), and the MG-ADL score at baseline was median 9.5 (IQR 5-11; mean

XSL•FO
RenderX

8.5, SD 3.60). Of all 29 patients, 69% (20/29) expressed willingness to use the app regularly. Among the patients who responded negatively or neutral to this statement, the documented oral feedback collected at the end-of-study visit mainly encompassed technical feedback, such as suggestions for adjusting the size of the chat window, implementing filter options for monitoring parameters in the health diary, and presenting preferences for the parameter dashboard. Furthermore, patients also frequently noted necessary improvements for a more user-friendly medication plan. In addition, there were critiques regarding the omission of specific symptoms in the PROMs (eg, nonmotor symptoms, pain, and sleep) and requests from patients for a free-form diary function (eg, for documenting infections or daily fluctuations).

Figure 7. The System Usability Scale (SUS) uses a 10-item questionnaire with 5 response options for respondents, ranging from strongly agree to strongly disagree. Due to rounding, the percentage for Question 2 ("I feel confident using the app") is 99% instead of 100%.

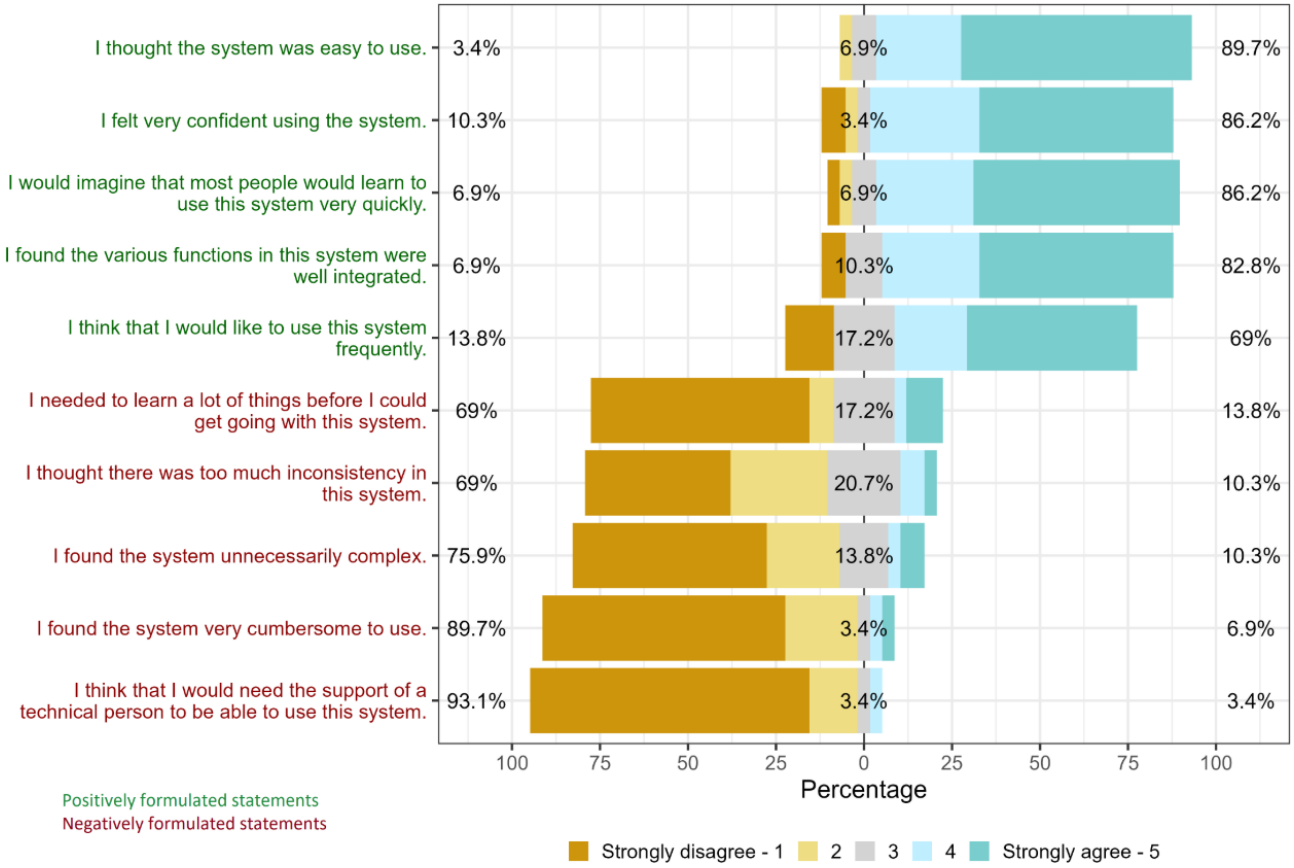

Discussion

Our study evaluated the feasibility and usability of a telemedicine platform for patient-physician interaction and monitoring for MG, incorporating wearable data collection. Feasibility, measured by adherence, ranged from 74.3%- 97.9% (1830/2464 - 327/334) across all data types, with the highest adherence observed for completion of PROMs (1139/1179, 96.6%) within the app, followed by adherence to spirometry measurements (293/334, 87.7%), and then wearable adherence (1830/2261, 80.9%). Next to adherence, user satisfaction among patients and health care professionals was assessed and generally rated high to very high across different categories (care program, communication, and evaluation of the program) and considered "excellent" according to the SUS score. The majority of patients stated their willingness to continue using the app.

Our results demonstrate that the introduction of such a digital tool in a clinical setting is feasible and facilitates clinical monitoring regardless of age or gender. These findings represent the first results in the field of MG for telemedical intervention. There was mutual acceptance from both health care professionals and patients, as evidenced by the high adherence rates and good to very good user satisfaction assessed at the end of the study.

Overall, patients demonstrated greater engagement with the app questionnaires compared with data collection through wearables and spirometry. Passive data collection by wearables displayed higher adherence rates than active spirometry, interestingly diverging from previous research [32]. Consistent with previous clinical studies [33-35], there was a variability particularly in wearable adherence. Our additional clinical data from the chat interactions revealed that commonly technical issues (eg, pairing problems and device malfunction) and discomfort (eg, wristband intolerance) were factors contributing to lower adherence rates for some patients. Only 1 patient consistently experienced difficulties (and ultimately discontinued wearable use). However, patients wore the device significantly longer than required by the study protocol, including during nighttime (median 20 h per day (IQR 15-24)). Similarly, spirometry was performed more frequently than required by protocol (49 measurements (median; IQR 15-59)). This indicates a high need and willingness of patients to monitor their own symptoms. Our

XSL•FO
RenderX

study results align with broader research on telemedicine for chronic diseases, demonstrating high feasibility and user acceptance, consistent with findings in other disorders. Similar to studies in palliative care in cancer patients [36], depression monitoring [37] and heart failure [38], our results confirm the feasibility of continuous monitoring through mobile technologies for chronic conditions. With an adherence rate of 80.9% (1830/2261) for wearables, our study falls at the higher end of reported ranges (53% - 94%) [36-43]. However, the higher adherence rates for active data (PROMs) over passive data (wearables) in our results contrasts with other studies finding [40,42], indicating potential disease-specific differences requiring further investigation. In addition, patients' sense of control during active data input or technical challenges with wearables could have affected adherence to passive monitoring. Still, comparison with other studies for different diseases remains difficult, as literature on wearable adherence expose inconsistency in reporting and study design.

Overall user satisfaction was high, particularly patients agreed on the ease of use, satisfaction, and improvement in understanding their symptoms. This finding aligns with a study in multiple sclerosis that reported sustained satisfaction with remote monitoring using smartphone and wearable technology over 6 months [40]. Further qualitative analyses, such as focus groups with patients, could aid in gaining a deeper understanding of the reasons behind these perceptions and to identify areas for improvements. First approaches have been made by the research team within the framework of additional small-scale pilot phases involving patients testing the platform use and workshops for feedback collection and development requests. Furthermore, mixed opinions were indicated by patients regarding whether the use of the telemedicine platform could reduce visits to neurologists or clinics, indicating that the platform may primarily serve as a complement to existing care structures rather than a replacement of personal interactions.

For successful implementation in clinical care, the efficient usage and integration of such data into clinical decision-making processes will be essential. This includes logistical efforts, but also necessitates a careful assessment of data quality and interpretation, and the "appropriate" monitoring intensity for different patient groups (eg, those with a short disease duration, highly active disease, and fast-acting treatment regimens may require more frequent assessments). Furthermore, addressing regulatory challenges, particularly regarding reimbursement structures for health care professionals will be crucial for long-term integration within health care systems.

Limitations

Our study encountered several challenges associated with remote monitoring technology: These included missing passive wearable data primarily due to technical issues and synchronization problems and patients' discomfort with the wristband from the wearable. Addressing these challenges and providing solutions, such as training sessions and available technical support, is crucial to ensure optimal patient experience and adherence in future studies using these technologies. Furthermore, the study cohort is relatively small, and the study duration is short. Patients with higher digital literacy may have been more likely to participate, potentially impacting the representativeness of the cohort. Regarding user evaluation of health care professionals, it should be noted that the study team size was small and partially comprised members of the development team for the platform, which could add to bias in answering the questionnaire.

Conclusion

In conclusion, our results demonstrate the feasibility of remote monitoring for collecting both subjective (active) and objective (passive) data from multiple sources among patients with MG, as evidenced by above-average adherence rates and acceptability. This underscores the potential for user-friendly digital tools for long-term monitoring of patients with MG and their willingness to track disease progression. Larger and prolonged studies are needed to benchmark our findings and to further explore the effective implementation of such tools and their clinical implications in practice.

Acknowledgments

We thank the Neuroscience Clinical Research Center at Charité Universitätsmedizin Berlin for their support in administration and throughout the formal processes essential for the successful execution of this study. The authors especially wish to thank Maja Olzewska, Josefina Kreusler, and Gabriela Jiron-Lozano for their support in establishing the REDCap database. Furthermore, special acknowledgments are extended to Jens Brestrich and Dike Remstedt for their support in conducting the study. In addition, the authors thank Qurasoft GmbH, namely Lale Zils, for the project assistance during the study and Sebastian Thunert, Lucas Keune, and Ragnar Hirschbrich for their technical support throughout the study. We wish to thank all the patients who participated in the study. Figure 2 was created with BioRender.com. Argenx, UCB, and Hormosan provided a grant to support the study but were not involved in the design of the study. The project was partly funded and supported by the Berlin Institute for Health/Charité Universitätsmedizin Berlin.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Authors' Contributions

MS wrote the paper. MS, SL, and LG developed the study design. AM critically reviewed the study. MS and SL conducted the study. DL helped with observation of the data. RS performed statistical analyses. MS, SL, and LG interpreted the data. RS, PN, DL, MH, AM, LG, and SL critically reviewed and edited the paper.

Conflicts of Interest

MS has received speaker's honoraria and honoraria for attendance at advisory boards from Argenx and Alexion. MH has received speaker´s honoraria from Argenx and speaker's honoraria and honoraria for attendance at advisory boards from Alexion. PN has received grant support from PCORI, Alexion, Dianthus, and Janssen. PN has received honoraria for consulting services for Alexion, Argenx, CVS Caremark, Dianthus, GSK, Immune Abs, Janssen, Novartis, and UCB. PN has served on DSMB for Sanofi. AM received speaker's honoraria from Alexion, Grifols, and Hormosan, and honoraria from Argenx, Alexion, MorphoSys, and UCB for consulting services, and financial research support from Alexion and Octapharma. He is chairperson of the medical advisory board of the DMG. SL has received speaker's honoraria for attendance at patient events and advisory boards from Argenx, Alexion, Biogen, Hormosan, Roche, and UCB. None declared by other authors.

References

1. Gilhus NE, Tzartos S, Evoli A, Palace J, Burns TM, Verschuuren J. Myasthenia gravis. Nat Rev Dis Primers 2019 May 2;5(1):1-19. [doi: 10.1038/s41572-019-0079-y]
2. Vinciguerra C, Iacono S, Bevilacqua L, et al. Sex differences in neuromuscular disorders. Mech Ageing Dev 2023 Apr;211:111793. [doi: 10.1016/j.mad.2023.111793] [Medline: 36806604]
3. Narayanaswami P, Sanders DB, Wolfe G, et al. International consensus guidance for management of myasthenia gravis: 2020 Update. Neurology (ECronicon) 2021 Jan 19;96(3):114-122. [doi: 10.1212/WNL.0000000000011124] [Medline: 33144515]
4. Lehnerer S, Jacobi J, Schilling R, et al. Burden of disease in myasthenia gravis: taking the patient's perspective. J Neurol 2022 Jun;269(6):3050-3063. [doi: 10.1007/s00415-021-10891-1] [Medline: 34800167]
5. Stein M, Hoffmann S, Gerischer L, et al. Myasthenia gravis - a retrospective analysis of e-mail inquiries made to a patient organisation and specialized center to uncover unmet needs from patients and caregivers. BMC Neurol 2022 Dec 7;22(1):455. [doi: 10.1186/s12883-022-02981-y] [Medline: 36476357]
6. Masanneck L, Räuber S, Schroeter CB, et al. Driving time-based identification of gaps in specialised care coverage: An example of neuroinflammatory diseases in Germany. Digit Health 2023;9:20552076231152989. [doi: 10.1177/20552076231152989] [Medline: 36762020]
7. Holko M, Litwin TR, Munoz F, et al. Wearable fitness tracker use in federally qualified health center patients: strategies to improve the health of all of us using digital health devices. NPJ Digit Med 2022 Apr 25;5(1):53. [doi: 10.1038/s41746-022-00593-x] [Medline: 35469045]
8. Ramezani R, Cao M, Earthperson A, Naeim A. Developing a smartwatch-based healthcare application: notes to consider. Sensors (Basel) 2023 Jul 25;23(15):6652. [doi: 10.3390/s23156652] [Medline: 37571436]
9. Dierckx R, Pellicori P, Cleland JGF, Clark AL. Telemonitoring in heart failure: big brother watching over you. Heart Fail Rev 2015 Jan;20(1):107-116. [doi: 10.1007/s10741-014-9449-4] [Medline: 24972644]
10. Koehler F, Koehler K, Deckwart O, et al. Efficacy of telemedical interventional management in patients with heart failure (TIM-HF2): a randomised, controlled, parallel-group, unmasked trial. Lancet 2018 Sep 22;392(10152):1047-1057. [doi: 10.1016/S0140-6736(18)31880-4] [Medline: 30153985]
11. Smolis-Bąk E, Dąbrowski R, Piotrowicz E, et al. Hospital-based and telemonitoring guided home-based training programs: Effects on exercise tolerance and quality of life in patients with heart failure (NYHA class III) and cardiac resynchronization therapy. A randomized, prospective observation. Int J Cardiol 2015 Nov;199:442-447. [doi: 10.1016/j.ijcard.2015.07.041]
12. Zhang Y, Pratap A, Folarin AA, et al. Long-term participant retention and engagement patterns in an app and wearable-based multinational remote digital depression study. NPJ Digit Med 2023 Feb 17;6(1):25. [doi: 10.1038/s41746-023-00749-3] [Medline: 36806317]
13. Druce KL, Dixon WG, McBeth J. Maximizing engagement in mobile health studies: lessons learned and future directions. Rheum Dis Clin North Am 2019 May;45(2):159-172. [doi: 10.1016/j.rdc.2019.01.004] [Medline: 30952390]
14. De Angel V, Lewis S, White K, et al. Digital health tools for the passive monitoring of depression: a systematic review of methods. NPJ Digit Med 2022 Jan 11;5(1):3. [doi: 10.1038/s41746-021-00548-8] [Medline: 35017634]
15. Svennberg E, Tjong F, Goette A, et al. How to use digital devices to detect and manage arrhythmias: an EHRA practical guide. Europace 2022 Jul 15;24(6):979-1005. [doi: 10.1093/europace/euac038] [Medline: 35368065]
16. Hillmann HAK, Soltani S, Mueller-Leisse J, Hohmann S, Duncker D. Cardiac rhythm monitoring using wearables for clinical guidance before and after catheter ablation. J Clin Med 2022 Apr 26;11(9):2428. [doi: 10.3390/jcm11092428] [Medline: 35566556]

XSL•FO
RenderX

17. Stein M, Meisel A, Moench M, et al. The telemedical platform myalink for remote monitoring in myasthenia gravis: rationale and protocol for a proof of concept study. MedRxiv. Preprint posted online on Jul 11, 2024. [doi: 10.1101/2024.07.10.24310229]
18. Jaretzki A 3rd, Barohn RJ, Ernstoff RM, et al. Myasthenia gravis: recommendations for clinical research standards. Task Force of the Medical Scientific Advisory Board of the Myasthenia Gravis Foundation of America. Neurology (ECronicon) 2000 Jul 12;55(1):16-23. [doi: 10.1212/wnl.55.1.16] [Medline: 10891897]
19. Randomization service for multicenter clinical trials. Randomizer. URL: https://www.randomizer.at/ [accessed 2025-02-27]
20. Barohn RJ, McIntire D, Herbelin L, Wolfe GI, Nations S, Bryan WW. Reliability testing of the quantitative myasthenia gravis score. Ann N Y Acad Sci 1998 May 13;841:769-772. [doi: 10.1111/j.1749-6632.1998.tb11015.x] [Medline: 9668327]
21. Wolfe GI, Herbelin L, Nations SP, Foster B, Bryan WW, Barohn RJ. Myasthenia gravis activities of daily living profile. Neurology (ECronicon) 1999 Apr 22;52(7):1487-1489. [doi: 10.1212/wnl.52.7.1487] [Medline: 10227640]
22. Burns TM, Sadjadi R, Utsugisawa K, et al. International clinimetric evaluation of the MG-QOL15, resulting in slight revision and subsequent validation of the MG-QOL15r. Muscle Nerve 2016 Dec;54(6):1015-1022. [doi: 10.1002/mus.25198] [Medline: 27220659]
23. Abraham A, Breiner A, Barnett C, Katzberg HD, Bril V. The utility of a single simple question in the evaluation of patients with myasthenia gravis. Muscle Nerve 2018 Feb;57(2):240-244. [doi: 10.1002/mus.25720] [Medline: 28590545]
24. Kukulka K, Gummi RR, Govindarajan R. A telephonic single breath count test for screening of exacerbations of myasthenia gravis: A pilot study. Muscle Nerve 2020 Aug;62(2):258-261. [doi: 10.1002/mus.26987] [Medline: 32447763]
25. Spitzer RL, Kroenke K, Williams JB. Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. Primary care evaluation of mental disorders. Patient health questionnaire. JAMA 1999 Nov 10;282(18):1737-1744. [doi: 10.1001/jama.282.18.1737] [Medline: 10568646]
26. Zigmond AS, Snaith RP. The hospital anxiety and depression scale. Acta Psychiatr Scand 1983 Jun;67(6):361-370. [doi: 10.1111/j.1600-0447.1983.tb09716.x] [Medline: 6880820]
27. Mendoza M, Tran C, Bril V, Katzberg HD, Barnett C. Patient-acceptable symptom states in myasthenia gravis. Neurology (ECronicon) 2020 Sep 22;95(12):e1617-e1628. [doi: 10.1212/WNL.0000000000010574] [Medline: 32759200]
28. Brooke J. SUS: A quick and dirty usability scale. In: Usability Evaluation in Industry: Taylor & Francis; 1995:189-194.
29. Braun M, Schmidt O, Schultz T, Woehrle H, Sundrup MG, Schöbel C. Erfahrungen mit der digitalen Versorgung von Patienten mit chronischen und akuten Lungenerkrankungen während der SARS-CoV-2-Pandemie [Article in German]. Internist 2022 Mar;63(3):255-265. [doi: 10.1007/s00108-022-01266-3]
30. R. The R Project for Statistical Computing. URL: https://www.r-project.org/ [accessed 2020-09-11]
31. Bangor A, Kortum PT, Miller JT. An empirical evaluation of the System Usability Scale. Int J Hum Comput Interact 2008 Jul 29;24(6):574-594. [doi: 10.1080/10447310802205776]
32. Wahle F, Kowatsch T, Fleisch E, Rufer M, Weidt S. Mobile sensing and support for people With depression: a pilot trial in the wild. JMIR Mhealth Uhealth 2016 Sep 21;4(3):e111. [doi: 10.2196/mhealth.5960] [Medline: 27655245]
33. Jacoby SF, Robinson AJ, Webster JL, Morrison CN, Richmond TS. The feasibility and acceptability of mobile health monitoring for real-time assessment of traumatic injury outcomes. Mhealth 2021;7:5. [doi: 10.21037/mhealth-19-200] [Medline: 33634188]
34. Lee S, Kim H, Park MJ, Jeon HJ. Current advances in wearable devices and their sensors in patients with depression. Front Psychiatry 2021;12:672347. [doi: 10.3389/fpsyt.2021.672347] [Medline: 34220580]
35. Chan YFY, Wang P, Rogers L, et al. The Asthma Mobile Health Study, a large-scale clinical observational study using ResearchKit. Nat Biotechnol 2017 Apr;35(4):354-362. [doi: 10.1038/nbt.3826] [Medline: 28288104]
36. Pavic M, Klaas V, Theile G, et al. Mobile health technologies for continuous monitoring of cancer patients in palliative care aiming to predict health status deterioration: a feasibility study. J Palliat Med 2020 May;23(5):678-685. [doi: 10.1089/jpm.2019.0342] [Medline: 31873052]
37. Pedrelli P, Fedor S, Ghandeharioun A, et al. Monitoring changes in depression severity using wearable and mobiles ensors. Front Psychiatry 2020;11:584711. [doi: 10.3389/fpsyt.2020.584711] [Medline: 33391050]
38. Werhahn SM, Dathe H, Rottmann T, et al. Designing meaningful outcome parameters using mobile technology: a new mobile application for telemonitoring of patients with heart failure. ESC Heart Fail 2019 Jun;6(3):516-525. [doi: 10.1002/ehf2.12425] [Medline: 30868756]
39. Wu JM, Ho TW, Chang YT, et al. Wearable-based mobile health app in gastric cancer patients for postoperative physical activity monitoring: focus group study. JMIR Mhealth Uhealth 2019 Apr 23;7(4):e11989. [doi: 10.2196/11989] [Medline: 31012858]
40. Midaglia L, Mulero P, Montalban X, et al. Adherence and satisfaction of smartphone- and smartwatch-based remote active testing and passive monitoring in people with multiple sclerosis: nonrandomized interventional feasibility study. J Med Internet Res 2019 Aug 30;21(8):e14863. [doi: 10.2196/14863] [Medline: 31471961]
41. Matcham F, Leightley D, Siddi S, et al. Remote Assessment of Disease and Relapse in Major Depressive Disorder (RADAR-MDD): recruitment, retention, and data availability in a longitudinal remote measurement study. BMC Psychiatry 2022 Feb 21;22(1):136. [doi: 10.1186/s12888-022-03753-1] [Medline: 35189842]

XSL•FO
RenderX

42. Sohn A, Speier W, Lan E, et al. Integrating remote monitoring into heart failure patients' care regimen: A pilot study. PLoS One 2020;15(11):e0242210. [doi: 10.1371/journal.pone.0242210] [Medline: 33211733]
43. Naslund JA, Aschbrenner KA, Bartels SJ. Wearable devices and smartphones for activity tracking among people with serious mental illness. Ment Health Phys Act 2016 Mar;10:10-17. [doi: 10.1016/j.mhpa.2016.02.001] [Medline: 27134654]

Abbreviations

BlnDSG: Berlin Data Protection Act
CFS: Chalder Fatigue Scale
GPDR: General Data Protection Regulation
HADS: Hospital Anxiety and Depression Scale
MG: myasthenia gravis
MG-ADL: myasthenia gravis activities of daily living
MG-QoL15r: myasthenia gravis quality of life, revised
MGFA: Myasthenia Gravis Foundation of America
PASS: Patient Acceptable Symptom State
PROMs: patient-reported outcome measures
QMGS: quantitative myasthenia gravis score
RCT: randomized controlled trial
REDCap: Research Electronic Data Capture
SSQ: single simple question
SUS: System Usability Scale

Edited by A Mavragani; submitted 03.04.24; peer-reviewed by C Vinciguerra, L Masanneck; revised version received 03.12.24; accepted 27.12.24; published 03.03.25.

Please cite as:
Stein M, Stegherr R, Narayanaswami P, Legg D, Herdick M, Meisel A, Gerischer L, Lehnerer S
App- and Wearable-Based Remote Monitoring for Patients With Myasthenia Gravis and Its Specialists: Feasibility and Usability Study
JMIR Form Res 2025;9:e58266
URL: https://formative.jmir.org/2025/1/e58266
doi:10.2196/58266

© Maike Stein, Regina Stegherr, Pushpa Narayanaswami, David Legg, Meret Herdick, Andreas Meisel, Lea Gerischer, Sophie Lehnerer. Originally published in JMIR Formative Research (https://formative.jmir.org), 3.3.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Enhancing Distress Tolerance Skills in Adolescents With Anorexia Nervosa Through the BALANCE Mobile App: Feasibility and Acceptability Study

Christina Miranda[1], BA; Brittany Matheson[2], PhD; Nandini Datta[2], PhD; Aileen Whyte[2], PhD; Hyun-Joon Yang[2], BS; Paul Schmiedmayer[3], PhD; Vishnu Ravi[3], MD; Oliver Aalami[3], MD; James Lock[2], MD, PhD

[1]School of Medicine, Stanford University, 291 Campus Dr, Stanford, CA, United States

[2]Department of Psychiatry and Behavioral Sciences, Stanford Medicine, Stanford, CA, United States

[3]Stanford Mussallem Center for Biodesign, Stanford University, Stanford, CA, United States

Corresponding Author:
Christina Miranda, BA
School of Medicine, Stanford University, 291 Campus Dr, Stanford, CA, United States

Abstract

Background: Anorexia nervosa is a severe psychiatric disorder with high morbidity and mortality, particularly among adolescents. Family-based treatment (FBT) is the leading evidence-based intervention for adolescent anorexia nervosa, involving parents in renourishment and behavior interruption. Despite its effectiveness, challenges in distress tolerance and emotion regulation during high-stress situations, such as mealtimes, contribute to suboptimal treatment outcomes, with only 35% to 50% of adolescents achieving full recovery. Enhancing distress tolerance skills during FBT may improve treatment responses and recovery rates. The BALANCE mobile app was developed to address this need, offering real-time, dialectical behavior therapy (DBT)–based distress tolerance skills to support adolescents and families during mealtimes.

Objective: Our aim was to explore the feasibility and acceptability of a mobile app designed to deliver distress tolerance skills to adolescents with and adolescents without anorexia nervosa. When fully programmed and optimized, we plan to use the mobile app to improve distress tolerance during mealtimes for adolescents with anorexia nervosa undergoing FBT.

Methods: BALANCE was developed collaboratively with Stanford University's Center for Biodesign, leveraging the expertise of clinical psychologists and using biodesign student input and the Stanford Spezi ecosystem. The app underwent an iterative development process, with feedback from adolescent users. The initial feasibility and acceptability of the app were assessed through self-reported questionnaires and structured interviews with 24 adolescents aged 12 to 18 years, including 4 diagnosed with anorexia nervosa and 20 healthy controls. Adolescents with anorexia nervosa specifically used the app during mealtimes, and healthy controls used it as needed. Participants assessed the app's usability, perceived effectiveness, and its impact on their distress tolerance.

Results: The app demonstrated high usability and acceptability. Of 24 participants, 83% (n=20) reported enjoying the app, 88% (n=21) would recommend it to peers, and 100% (n=24) found it user-friendly. Adolescents with anorexia nervosa reported that BALANCE helped them manage stressful mealtimes more effectively, highlighting features such as guided meditation, breathing exercises, and gamification elements as particularly effective. Healthy controls provided additional feedback, confirming the app's broad appeal to the target audience and potential scalability. Preliminary findings suggest that BALANCE may enhance distress tolerance in adolescents with and adolescents without anorexia nervosa.

Conclusions: BALANCE shows promise as an innovative mobile health intervention for enhancing distress tolerance in adolescents with anorexia nervosa. Its user-friendly design and tailored DBT-based skills make it a feasible tool for integration into FBT. Future research should explore its integration into clinical practice and its impact on treatment outcomes. As distress tolerance skills are relevant to a range of mental health conditions, future research may also expand BALANCE's application to broader adolescent populations.

(JMIR Form Res 2025;9:e70278) doi:10.2196/70278

KEYWORDS

mHealth; mobile health; mobile application; emotion regulation; eating disorders; family-based treatment; distress tolerance; mealtimes

XSL•FO
RenderX

Introduction

Anorexia nervosa is a severe psychiatric disorder that poses significant morbidity and mortality risks, particularly among adolescents. Family-based treatment (FBT) has been shown to be effective in addressing adolescent anorexia nervosa by involving parents to promote renourishment and disrupt eating disorder behaviors [1]. However, despite its success, only 35% to 50% of adolescents fully recover through FBT, highlighting the need for innovative strategies to enhance treatment outcomes [2,3].

Recent research highlights the role of distress tolerance in eating disorder recovery, with studies demonstrating that lower distress tolerance is associated with poorer treatment outcomes [4]. Despite the increasing adoption of mobile health (mHealth) interventions in psychiatric care, few apps specifically target distress tolerance during mealtimes for adolescents with anorexia nervosa. The BALANCE app was developed to address this gap, providing real-time, dialectical behavior therapy (DBT)–based distress tolerance strategies that can be integrated into FBT.

Several studies in the past 2 to 3 years have examined digital interventions for eating disorders, reinforcing the importance of personalized and interactive tools in improving treatment adherence [5-7]. Digital interventions have been found to enhance engagement by providing real-time coping strategies, facilitating adherence to treatment protocols, and supporting long-term behavior change [7]. Additionally, studies emphasize the importance of incorporating structured distress tolerance strategies into existing treatment models to optimize clinical outcomes [4]. By leveraging DBT principles, BALANCE aims to enhance distress tolerance skills, thereby ultimately improving treatment efficacy for adolescents with anorexia nervosa.

This study evaluated the feasibility and acceptability of the BALANCE mobile app, which was designed to improve distress tolerance in adolescents, with the aim of ultimately having the app be used during mealtimes for adolescents with anorexia nervosa undergoing FBT.

Methods

Recruitment Process and Procedures

To ensure a systematic and transparent recruitment process, we implemented the following step-by-step procedure: (1) identifying eligible participants, (2) screening for eligibility, (3) obtaining parental and participant consent, and (4) conducting an orientation session.

Identification of Eligible Participants

A total of 24 adolescents aged 12 to 18 years were recruited for this study. For the initial iteration of the app, we focused on acceptability and feasibility among the adolescent age group. In total, 20 healthy adolescent controls were recruited from the community through flyers on social media platforms. We also recruited 4 adolescents with anorexia nervosa from the Stanford University Comprehensive Care Program—a 15-bed inpatient unit within Lucile Packard Children's Hospital that provides multidisciplinary care for medically compromised adolescents with eating disorders—to preliminarily assess acceptability and feasibility with this clinical group.

Screening for Eligibility

Interested participants completed an initial phone call with study staff to determine eligibility based on diagnosis, age, and access to a smart device that could download the BALANCE app.

Parental and Participant Consent

For participants younger than 18 years, parental consent and adolescent assent were obtained. Participants aged 18 years provided informed consent independently. All participants received detailed study explanations, including explanations about potential risks and benefits.

Orientation Session

Eligible participants attended an orientation session where they learned how to use the app, including its features and intended usage during mealtimes.

Ethical Considerations

This study was approved by Stanford University's institutional review board, under the protocol titled "Refinement of Emotion Regulation App for Adolescents with Anorexia Nervosa at Mealtimes During Family-Based Treatment" (eProtocol #69550), adhering to ethical guidelines for human subject research. Informed consent was obtained from all participants aged 18 years, and for participants younger than 18 years, consent was secured from their parents or legal guardians alongside adolescent assent. To ensure privacy and confidentiality, all data were deidentified prior to analysis and securely stored on encrypted servers that were accessible only to the research team. Findings were reported in aggregate to prevent individual identification. Participants were not monetarily compensated but were provided early access to the BALANCE app and allowed continued use after this study's completion.

App Development and Collaboration

BALANCE was developed in collaboration with Stanford University's Eating Disorder Research Team and a Stanford biodesign course on mHealth app development [8]. The course enabled computer science undergraduate and graduate students to provide input on app features, interface design, and technical functionality, leveraging the Stanford Spezi open-source digital health framework [9] and student expertise in mHealth technology. Spezi was designed for the rapid development of modern, interoperable mHealth apps based on an ecosystem of modules that encapsulate common functionality and build upon international standards for health data. Prior research on app-based interventions for eating disorders demonstrated the importance of incorporating lived experiences of potential users into the development process [10]. Similarly, the BALANCE app development followed an iterative, user-centered design process, which incorporated 2 waves of feedback from adolescent users (N=10), ensuring the app's relevance and usability for the target adolescent age group. Licensed clinical psychologists with expertise in DBT-based distress tolerance skills guided the creation of the app's core content, while

XSL•FO
RenderX

students helped refine the user experience through feedback loops and usability testing [11].

BALANCE incorporates several features designed to improve distress tolerance in adolescents. Users can select a distress tolerance skill, such as distraction, relaxation, or emotion tracking, through the app's interface (Figure 1). Data on app use (eg, frequency, duration, and skill type) are collected. The app also includes a gamification feature, rewarding users with coins (ie, for customizing an avatar) based on their engagement. The app can be opened directly when prompted by the adolescents or their parents. Additionally, when the app is integrated with a user's personal smartwatch, notifications are sent to the user when their heart rate increases enough to indicate the possibility of emotional distress, prompting them to open the app. Therapists working with the adolescents in FBT have access to a dashboard that provides information about adolescents' app use in visual tables and figures. Therapists can integrate this information (eg, how often the app was used, what time of day app use occurred, and what app features were accessed) into FBT when working with the patients and their family.

Figure 1. BALANCE app interface and features. This figure showcases the design and core functionalities of the BALANCE app—a mobile health intervention developed to deliver distress tolerance skills to adolescents aged 12 to 18 years. The app incorporates guided meditation, emotion tracking, and gamification features, and it was developed through a collaborative process involving clinical psychologists and biodesign students. Screenshots depict the user interface and interactive elements.

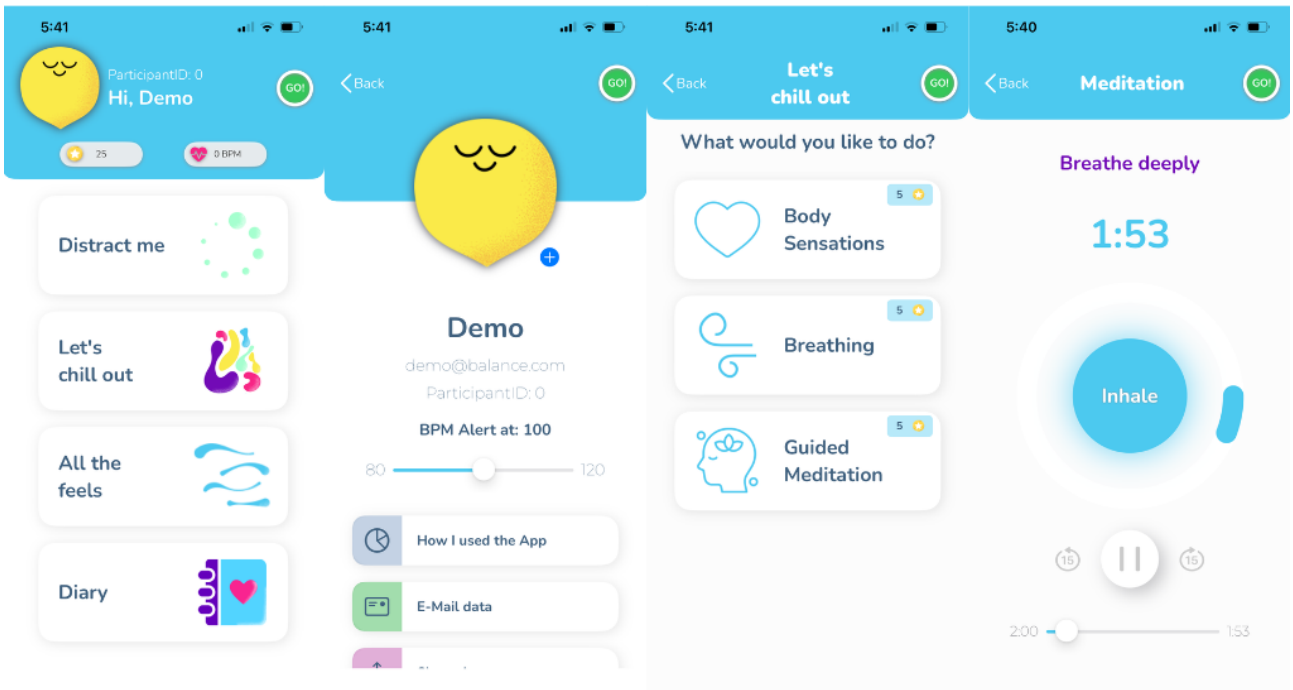

Study Design

This study assessed the BALANCE app's feasibility and acceptability, using self-reported questionnaires and interviews with adolescent users. A secondary aim was to explore app use among adolescents with anorexia nervosa. Healthy adolescents used the app for at least 30 minutes, while adolescents with anorexia nervosa used the app during mealtimes. Measures included usability, perceived effectiveness, and distress tolerance outcomes. Structured interviews allowed participants to elaborate on survey responses to suggest specific modifications to the app, comment on the design, alert us about any usability issues, and share any other feedback that was not captured by survey questions.

Measures

Participants rated their experience with the app by using Likert-scale surveys and open-ended questions about the app's functionality, design, and impact on emotion regulation.

Results

Participant Demographics

Participants consisted of healthy adolescents (n=20) and adolescents with anorexia nervosa (n=4). Participants were racially and ethnically diverse; of the 24 participants, 16 (67%) identified as female, 8 (33%) identified as male, 13 (54%) identified as White, 8 (33%) identified as Asian, and 3 (13%) identified as Hispanic or Latino.

Survey Results

The majority of participants found the app user-friendly (n=24, 100%) and helpful (n=22, 92%). Of the 24 participants, 20 (83%) reported enjoying the app, and 21 (88%) would recommend it to a friend. Notably, the guided meditation and breathing exercises were highlighted as particularly helpful during stressful situations. Participant ratings of the app were generally consistent between adolescents with and adolescents without anorexia nervosa (Table 1). Preliminary findings suggest that the BALANCE app improves adolescents' self-reported ability to manage emotional states. Adolescents with anorexia

XSL•FO
RenderX

nervosa reported a decrease in distress and increased ability to tolerate discomfort associated with eating (Tables 1-3).

Table . Survey responses from the 24 adolescents (4 with anorexia nervosa and 20 healthy controls) on the BALANCE app. Responses (collected from July 2023 to January 2024) highlight user satisfaction, ease of use, and distress tolerance skills.[a]

Survey statement	Strongly disagree, n (%)	Disagree, n (%)	Neutral, n (%)	Agree, n (%)	Strongly agree, n (%)
I enjoyed using this application	0 (0)	0 (0)	0 (0)	20 (83)	4 (17)
This app has functionality that is easy to understand	0 (0)	0 (0)	2 (8)	17 (71)	5 (21)
This app has a user-friendly interface design	0 (0)	0 (0)	0 (0)	14 (58)	10 (42)
This app is exhausting or difficult to use	11 (46)	10 (42)	2 (8)	1 (4)	0 (0)
I would recommend this app to a friend	0 (0)	0 (0)	3 (13)	21 (88)	0 (0)
This app helped me to regulate my emotions	0 (0)	1 (4)	8 (33)	15 (63)	0 (0)
I found this app to be triggering or unhelpful to my emotional state	16 (67)	6 (25)	0 (0)	2 (8)	0 (0)

[a]Percentages may not add up to 100% due to rounding.

Table . Survey responses from the 20 healthy adolescent controls on the BALANCE app. Responses (collected from July 2023 to January 2024) highlight user satisfaction, ease of use, and distress tolerance skills.[a]

Survey statement	Strongly disagree, n (%)	Disagree, n (%)	Neutral, n (%)	Agree, n (%)	Strongly agree, n (%)
I enjoyed using this application	0 (0)	0 (0)	0 (0)	16 (80)	4 (20)
This app has functionality that is easy to understand	0 (0)	0 (0)	1 (5)	14 (70)	5 (25)
This app has a user-friendly interface design	0 (0)	0 (0)	0 (0)	12 (60)	8 (40)
This app is exhausting or difficult to use	9 (45)	10 (50)	0 (0)	1 (5)	0 (0)
I would recommend this app to a friend	0 (0)	0 (0)	3 (15)	17 (85)	0 (0)
This app helped me to regulate my emotions	0 (0)	1 (5)	6 (30)	13 (65)	0 (0)
I found this app to be triggering or unhelpful to my emotional state	13 (65)	5 (25)	0 (0)	2 (10)	0 (0)

[a]Percentages may not add up to 100% due to rounding.

XSL•FO
RenderX

Table . Survey responses from the 4 adolescents with anorexia nervosa on the BALANCE app. Responses (collected from July 2023 to January 2024) highlight user satisfaction, ease of use, and distress tolerance skills.[a]

Survey statement	Strongly disagree, n (%)	Disagree, n (%)	Neutral, n (%)	Agree, n (%)	Strongly agree, n (%)
I enjoyed using this application	0 (0)	0 (0)	0 (0)	4 (100)	0 (0)
This app has functionality that is easy to understand	0 (0)	0 (0)	1 (25)	3 (75)	0 (0)
This app has a user-friendly interface design	0 (0)	0 (0)	0 (0)	2 (50)	2 (50)
This app is exhausting or difficult to use	2 (50)	0 (0)	2 (50)	0 (0)	0 (0)
I would recommend this app to a friend	0 (0)	0 (0)	0 (0)	4 (100)	0 (0)
This app helped me to regulate my emotions	0 (0)	0 (0)	2 (50)	2 (50)	0 (0)
I found this app to be triggering or unhelpful to my emotional state	3 (75)	1 (25)	0 (0)	0 (0)	0 (0)

[a]Percentages may not add up to 100% due to rounding.

Discussion

Summary of Findings

This study demonstrated the feasibility and acceptability of the BALANCE mobile app for delivering distress tolerance skills to adolescents, including those with anorexia nervosa. The app was well received, with 83% (20/24) of participants reporting enjoyment, 88% (21/24) reporting that they would recommend it to peers, and 100% (24/24) finding it user-friendly. Adolescents with anorexia nervosa reported improvements in their ability to tolerate mealtime distress, suggesting that the app may serve as a valuable adjunct to FBT. These findings align with this study's primary objective to assess the feasibility and acceptability of the app, paving the way for its integration into clinical practice. Additionally, this study provides initial evidence that BALANCE may help address distress-related barriers to effective meal completion—a crucial factor in anorexia nervosa recovery.

Interpretations

The BALANCE app provides a structured and scalable solution to a critical gap in FBT by delivering real-time distress tolerance skills for high-stress mealtimes [12,13]. Distress tolerance deficits contribute to poor treatment adherence and increased dropout rates among adolescents undergoing FBT [12,13]. The high usability ratings suggest that BALANCE may effectively supplement traditional therapeutic interventions, thereby improving mealtime coping mechanisms.

Compared to other mHealth apps for eating disorders, BALANCE is unique in its focus on real-time distress tolerance rather than self-monitoring or general emotional regulation [5,7]. Its DBT-based approach aligns with research showing the efficacy of distress tolerance training in improving treatment engagement and emotional resilience in adolescents with anorexia nervosa [14,15]. Gamification and interactive features further support engagement—a crucial factor in digital intervention success [5,6].

As digital tools become more integrated into clinical practice, BALANCE could be incorporated into standard FBT protocols to improve accessibility and engagement. By providing just-in-time intervention during mealtimes, it may help adolescents manage high-anxiety eating situations and support better treatment adherence and long-term recovery.

Limitations

This study's limitations include a small sample size, particularly the limited number of adolescents with anorexia nervosa. Additionally, the reliance on self-reported data introduces potential biases, and the short duration of app use prevents long-term effectiveness assessment. Future research should address these limitations by conducting larger-scale trials, exploring longitudinal impacts, and assessing BALANCE's integration into standard FBT protocols. Investigating potential barriers to app engagement in diverse clinical settings is also necessary for optimizing its implementation.

Conclusions and Broader Implications

The BALANCE app is a novel, acceptable, and feasible tool for delivering distress tolerance skills to adolescents with anorexia nervosa. It shows promise for use during high-stress mealtimes, addressing a critical barrier to the effectiveness of FBT. The pilot testing of the app's integration into FBT is underway, assessing its impact on treatment adherence and distress tolerance in real-world clinical settings. If successfully implemented, BALANCE could enhance treatment outcomes, reduce barriers to recovery, and improve FBT effectiveness. Future research should explore its broader applicability to other mental health conditions wherein distress tolerance plays a key

XSL•FO
RenderX

role, to ensure its scalability and long-term benefits for adolescent mental health interventions.

Acknowledgments

This project was funded by the Packard Foundation and Stanford University's undergraduate health application development course participants. The authors would like to thank the students and faculty from the Stanford Mussallem Center for Biodesign for their collaboration in developing the BALANCE app [8,9].

Data Availability

The datasets generated or analyzed during this study are not publicly available due to participant privacy and confidentiality considerations but are available from the corresponding author on reasonable request.

Conflicts of Interest

JL receives royalties from Guilford Press and Routledge. He is also the codirector of the Training Institute for Child and Adolescent Eating Disorders, LLC. ND receives royalties from Routledge. BM receives royalties from Routledge.

References

1. Lock J, Le Grange D, Agras WS, Moye A, Bryson SW, Jo B. Randomized clinical trial comparing family-based treatment with adolescent-focused individual therapy for adolescents with anorexia nervosa. Arch Gen Psychiatry 2010 Oct;67(10):1025-1032. [doi: 10.1001/archgenpsychiatry.2010.128] [Medline: 20921118]
2. Le Grange D, Lock J, Agras WS, Bryson SW, Jo B. Randomized clinical trial of family-based treatment and cognitive-behavioral therapy for adolescent bulimia nervosa. J Am Acad Child Adolesc Psychiatry 2015 Nov;54(11):886-894.e2. [doi: 10.1016/j.jaac.2015.08.008] [Medline: 26506579]
3. Lock JD, Le Grange D, Bohon C, Matheson B, Jo B. Who responds to an adaptive intervention for adolescents with anorexia nervosa being treated with family-based treatment? Outcomes from a randomized clinical trial. J Am Acad Child Adolesc Psychiatry 2024 Jun;63(6):605-614. [doi: 10.1016/j.jaac.2023.10.012] [Medline: 38142046]
4. Datta N, Matheson BE, Citron K, Van Wye EM, Lock JD. Evidence based update on psychosocial treatments for eating disorders in children and adolescents. J Clin Child Adolesc Psychol 2023;52(2):159-170. [doi: 10.1080/15374416.2022.2109650] [Medline: 35950931]
5. Wilks CR, Gurtovenko K, Rebmann K, Williamson J, Lovell J, Wasil AR. A systematic review of dialectical behavior therapy mobile apps for content and usability. Borderline Personal Disord Emot Dysregul 2021 Dec 3;8(1):29. [doi: 10.1186/s40479-021-00167-5] [Medline: 34857035]
6. Guala MM, Bul K, Skårderud F, Søgaard Nielsen A. A serious game for patients with eating disorders (Maze Out): pilot user experience and acceptance study. JMIR Form Res 2023 Jan 27;7:e40594. [doi: 10.2196/40594] [Medline: 36705956]
7. Lindgreen P, Lomborg K, Clausen L. Patient experiences using a self-monitoring app in eating disorder treatment: qualitative study. JMIR Mhealth Uhealth 2018 Jun 22;6(6):e10253. [doi: 10.2196/10253] [Medline: 29934285]
8. Stanford University. CS342 / MED253 - building for digital health: winter quarter – academic year 2023/2024. Notion. 2023. URL: https://stanfordbdhg.notion.site/CS342-MED253-Building-for-Digital-Health-115008f9653881e3b42ddf7dfbf5a156 [accessed 2025-02-18]
9. Schmiedmayer P, Bauer A, Zagar P, Ravi V, Aalami O. Stanford Spezi. Zenodo. 2024 Oct 29. URL: https://zenodo.org/records/14006855 [accessed 2025-02-18]
10. Jarman HK, McLean SA, Rodgers R, et al. Informing mHealth and web-based eating disorder interventions: combining lived experience perspectives with design thinking approaches. JMIR Form Res 2022 Oct 31;6(10):e38387. [doi: 10.2196/38387] [Medline: 36315225]
11. Safer DL, Robinson AH, Jo B. Outcome from a randomized controlled trial of group therapy for binge eating disorder: comparing dialectical behavior therapy adapted for binge eating to an active comparison group therapy. Behav Ther 2010 Mar;41(1):106-120. [doi: 10.1016/j.beth.2009.01.006] [Medline: 20171332]
12. Rowsell M, MacDonald DE, Carter JC. Emotion regulation difficulties in anorexia nervosa: associations with improvements in eating psychopathology. J Eat Disord 2016 May 18;4:17. [doi: 10.1186/s40337-016-0108-0] [Medline: 27195121]
13. Harrison A, Sullivan S, Tchanturia K, Treasure J. Emotional functioning in eating disorders: attentional bias, emotion recognition and emotion regulation. Psychol Med 2010 Nov;40(11):1887-1897. [doi: 10.1017/S0033291710000036] [Medline: 20102669]
14. Vogel EN, Singh S, Accurso EC. A systematic review of cognitive behavior therapy and dialectical behavior therapy for adolescent eating disorders. J Eat Disord 2021 Oct 18;9(1):131. [doi: 10.1186/s40337-021-00461-1] [Medline: 34663452]
15. Rozakou-Soumalia N, Dârvariu Ş, Sjögren JM. Dialectical behaviour therapy improves emotion dysregulation mainly in binge eating disorder and bulimia nervosa: a systematic review and meta-analysis. J Pers Med 2021 Sep 18;11(9):931. [doi: 10.3390/jpm11090931] [Medline: 34575707]

XSL•FO
RenderX

Abbreviations

DBT: dialectical behavior therapy
FBT: family-based treatment
mHealth: mobile health

Edited by A Mavragani; submitted 18.12.24; peer-reviewed by HJ Watson, NJ Goel; revised version received 03.02.25; accepted 04.02.25; published 28.02.25.

Please cite as:
Miranda C, Matheson B, Datta N, Whyte A, Yang HJ, Schmiedmayer P, Ravi V, Aalami O, Lock J
Enhancing Distress Tolerance Skills in Adolescents With Anorexia Nervosa Through the BALANCE Mobile App: Feasibility and Acceptability Study
JMIR Form Res 2025;9:e70278
URL: https://formative.jmir.org/2025/1/e70278
doi:10.2196/70278

© Christina Miranda, Brittany Matheson, Nandini Datta, Aileen Whyte, Hyun-Joon Yang, Paul Schmiedmayer, Vishnu Ravi, Oliver Aalami, James Lock. Originally published in JMIR Formative Research (https://formative.jmir.org), 28.2.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Probing Public Perceptions of Antidepressants on Social Media: Mixed Methods Study

Jianfeng Zhu[1], PhD; Xinyu Zhang[2], MA; Ruoming Jin[1], Prof Dr; Hailong Jiang[1], PhD; Deric R Kenne[3], PhD

[1]Department of Computer Science, Kent State University, 800 E Summit St, Kent, OH, United States

[2]Department of Communication, Michigan State University, East Lansing, MI, United States

[3]Center for Public Policy and Health, College of Public Health, Kent State University, Kent, OH, United States

Corresponding Author:
Jianfeng Zhu, PhD
Department of Computer Science, Kent State University, 800 E Summit St, Kent, OH, United States

Abstract

Background: Antidepressants are crucial for managing major depressive disorders; however, nonadherence remains a widespread challenge, driven by concerns over side effects, fear of dependency, and doubts about efficacy. Understanding patients' experiences is essential for improving patient-centered care and enhancing adherence**,** which prioritizes individual needs in treatment.

Objective: This study aims to gain a deeper understanding of patient experiences with antidepressants, providing insights that health care providers, families, and communities can develop into personalized treatment strategies. By integrating patient-centered care, these processes may improve satisfaction and adherence with antidepressants.

Methods: Data were collected from AskaPatient and Reddit, analyzed using natural language processing and large language models. Analytical techniques included sentiment analysis, emotion detection, personality profiling, and topic modeling. Furthermore, demographic variations in patient experiences were also examined to offer a comprehensive understanding of discussions around antidepressants.

Results: Sentiment and emotion analysis revealed that the majority of discussions (21,499/36,253, 59.3%) expressed neutral sentiments, with negative sentiments following closely (13,922/36,253, 38.4%). The most common emotions were fear (16,196/36,253, 44.66%) and sadness (12,507/36,253, 34.49%). The largest topic, "Mental Health and Relationships," accounted for 11.69% (3755/36,253) of the discussions, which indicated a significant focus on managing mental health conditions. Discussions around nonadherence were marked by fear, followed by sadness, while self-care discussions showed a notable trend of sadness.

Conclusions: These psychological insights into public perceptions of antidepressants provide a foundation for developing tailored, patient-centered treatment approaches that align with individual needs, enhancing both effectiveness and empathy of care.

(JMIR Form Res 2025;9:e62680) doi:10.2196/62680

KEYWORDS

antidepressant; AskaPatient; natural language processing; BERTopic; large language models; Reddit

Introduction

Background

Antidepressants are a cornerstone in the treatment of major depressive disorder worldwide, improving the quality of life for millions of individuals. However, nonadherence to antidepressant therapy remains a significant barrier to achieving successful outcomes. According to the National Center for Health Statistics, 13.2% of adults in the United States used antidepressants within the past 30 days, with higher rates among women (17.7%) than among men (8.4%) between 2015 and 2018 [1]. This disparity highlights potential differences in how men and women experience depression and respond to antidepressant therapy, which may influence their adherence behaviors [2].

The factors driving nonadherence are multifaceted, encompassing both patient- and clinician-related challenges. Patients may avoid medicines due to concerns about side effects, fear of dependency, or doubts about their effectiveness [3]. Furthermore, depression's symptoms, such as low motivation and emotion dysregulated, often prevent patients from filling prescriptions or attending follow-up appointments [2]. Clinician-related factors, such as insufficient patient education and inadequate follow-up, also contribute to poor adherence. These challenges underscore the need for a more patient-centered approach in managing depression.

Patient-centered care (PCC) is designed to optimize treatment adherence by ensuring that clinical interventions are closely aligned with each patient's preferences, needs, and values. This approach enhances communication between caregivers and

XSL•FO
RenderX

patients, supports collaborative decision-making, and increase overall patient satisfaction, all of which are essential to improving adherence rates [4,5]. However, long-term treatment with selective serotonin reuptake inhibitors raises concerns about emotion blunting, a condition in which patients report feeling emotionally numb, impacting both their positive and negative emotions, which potentially alter personality [6-8]. This highlights the need to examine the emotional and psychological dimensions of antidepressant use in real-world setting.

Traditional methods such as surveys and interviews are limited by recall bias and cannot capture the day-to-day emotional fluctuations that significantly influence medication adherence [9,10]. Real-time insights are crucial to understanding how emotions, which change rapidly in response treatment, impact medication adherence. Given that medication adherence is closely linked to patient emotions during treatment, consequently there is a need for innovative data collection that can offer information on patients' real-time experience.

Social media offers a unique opportunity to capture real-time, unfiltered opinions from users [11-13]. Unlike structured clinical trials or surveys, social media discussions reflect spontaneous, in the moment experiences, offering deeper insights into how patients perceive their medications and mental health [14-16]. Furthermore, social media provides large-scale, real-time data on patient experiences, which can be analyzed to better understand the emotional and psychological factors influencing antidepressant adherence.

Advancements in artificial intelligence, particularly with natural language processing (NLP) and large language models (LLMs), such as GPT-4, allow for more nuanced analysis of patient language [17-19]. These tools provide deeper insights by not only detecting sentimental analysis but also profiling personality traits, allowing researchers to explore psychological dimensions of antidepressant users. Through NLP and LLMs, it is possible to assess emotional states, sentiment, and personality traits from large volumes of patient-generated content.

This study aims to bridge the gap between clinical assessment and patient-reported experiences by using real-time social media data. This advanced approach offers a more comprehensive understanding of patient experiences, capturing emotional and psychological factors often missed by traditional methods. The findings support decision-making strategies for caregivers and enhance PCC strategies leading to increase treatment adherence.

Literature Review

Nonadherence to Antidepressants

Research consistently highlights significant variability in antidepressant adherence across different populations. A review of studies over the past decade reveals that approximately 50% of patients discontinue antidepressant therapy prematurely, which underscore the pervasive nature of nonadherence [3]. Pompili et al [20] identified several key predictors of nonadherence in patients with mood disorders, including those of younger age (younger than 40 years), comorbid substance use or personality disorders, negative beliefs, poor insight, treatment-related side effects, and a weak therapeutic alliance. The study on 332 patients by Kamaradova et al [21] found a negative correlation between self-stigma and adherence, with greater illness severity associated with lower adherence across all diagnostic groups. Similarly, De las Cuevas et al [22] identified that adherence was closely linked to positive attitudes toward treatment, while nonadherence was associated with severe depression and concerns about side effects. These findings underscore the multifactorial nature of antidepressant nonadherence, influenced by patient beliefs, illness severity, and the quality of patient-provider relationships [23].

Patient-Centered Care

PCC, which prioritizes the patient's preferences and values, plays a crucial role in addressing nonadherence. It improves patient satisfaction and outcomes by considering the psychological challenges, such as fear and uncertainty, faced by patients and their families [24]. Studies show that shared decision-making and effective communication between health care providers and patients foster trust, enabling patients to express their emotional concerns, ultimately strengthening the therapeutic alliance, and improving adherence [25,26].

Integration sentiment analysis into PCC allows health care providers to gain real-time insights into patient emotions and concerns, enabling timely interventions that enhance adherence and promote self-care strategies [27]. These tools, combined with personality profiling, offer a more nuanced understanding of patient experiences than traditional methods [28], allowing for personalized care that improves mental health outcomes.

Use of Social Media for Health Research

Social media provides real-time, unfiltered insights into patient experiences with medication and mental health challenges. Platforms such as X (Twitter) and Reddit are rich sources of data that help researchers understand patient interactions with their mental health conditions and the health care system. Social media enables real-time health monitoring and predictive analysis, allowing for early detection of changes in patient mood or attitudes toward medications, which can trigger timely interventions for nonadherence [29-31]. Sentiment analysis of social media posts has proven effective in identifying concerns about side effects, medication efficacy, and treatment challenges, complementing traditional health care systems by offering new avenues for improving adherence to antidepressants [30].

Integration of NLP and LLMs in Mental Health Research

Recent NLP models have shown promise in identifying and classifying even rare neuropsychiatric diseases. For instance, the language model RoBERTa has been used to identify clusters of individuals with primary progressive aphasia through speech analysis, providing health care providers with valuable insights into patient experiences [32,33]. LLMs, which increasingly power conversational agents, can simulate synthetic personalities, allowing for fine-tuned, personalized patient interactions. These models amplify mental health support by offering tailored perspectives that improve adherence to therapy and foster more effective patient care [34].

XSL•FO
RenderX

Emotion-based models, such as those using emotional transition fingerprints, have demonstrated greater generalizability than content-based models, helping researchers understand the social dynamics of antidepressant use [35]. Furthermore, BERT has been used to categorize antidepressant-related discussions on social media, such as mood, sleep, and somatic symptoms, showcasing NLP's ability to provide targeted analysis in specific domains [36]. These tools have also been applied to predict drug interactions and optimize treatment plans using techniques such as Adaptive Fuzzy Logic Neural Networks, improving accuracy in predicting side effects [37].

Despite substantial research on antidepressant nonadherence, existing studies often neglect the dynamic interplay of emotions, personality traits, and social media discourse in shaping patient behaviors. Traditional approaches have provided foundational insights, yet they frequently fail to capture the real-time and nuanced patient experiences accessible through modern social media platforms. Furthermore, while PCC is increasingly recognized for its potential to improve adherence, there is a critical gap on how to leverage advanced NLP and LLMs to investigate patients' present psychological reaction with antidepressants using social media datasets.

This research aims to bridge the gap, which provides a more granular understanding of the psychological factors by analyzing sentiment, emotion, and personality profiles related to antidepressant use across diverse demographics. Furthermore, we explore topics, especially topics associated with 2 themes: nonadherence and self-care across sentiment and personality profile, which enhances our understanding of antidepressant users' real-time psychological patterns. The following 4 research questions (RQs) align to the goal of this study.

RQ1: How do sentiment analyses reveal differences in antidepressant experiences across demographics?

RQ2: What predominant emotions (eg, fear, sadness, and joy) are associated with antidepressant use, and how do these vary by demographics?

RQ3: What personality traits (eg, extraversion and neuroticism) are prevalent among users discussing antidepressants, and how do these traits differ across demographics?

RQ4: What are the prevailing topics in web-based antidepressant discussions, and how do sentiment, emotion, and personality insights inform our understanding of nonadherence and self-care themes?

Through these questions, this study addresses key limitations identified in the existing literature and harnesses advanced NLP techniques on social media data. The findings offer deeper understanding of patients' experience, improving antidepressant treatments.

Methods

Data Collection and Preprocessing for Antidepressant Analysis

To explore public perceptions of antidepressants, data were collected from 2 major platforms: AskaPatient [38] and Reddit [39], both of which offer diverse, patient-reported experiences. AskaPatient provides structured reviews and ratings, while Reddit features spontaneous, unfiltered discussions, offering a broader range of insights.

From AskaPatient, data were collected for the top 10 antidepressants, based on 1571 reviews posted between 2015 and 2023. This dataset includes demographic information (eg, gender and age), satisfaction ratings, side effects, and free-text comments (see Table 1 for sample data). The analyzed antidepressants were Cymbalta, Effexor XR, Lexapro, Zoloft, Wellbutrin XL, Wellbutrin, Celexa, Paxil, Wellbutrin SR, and Prozac.

We filtered posts from Reddit between 2018 and 2023 using the same top 10 antidepressants analyzed in the AskaPatient dataset. Due to the absence of explicit demographic details in Reddit posts, we used regular expressions (regex) to deduce demographic information such as gender and age. For instance, in the post: "Hi, I am a 19 y/o male. 6'2" and approximately 210 lbs. I currently take venlafaxine (Effexor) to treat depression and generalized anxiety disorder. I do not smoke. For the last several months, I have been experiencing hiccups…," we inferred the demographic information as "19 y/o male." (For a sample dataset, refer to Multimedia Appendix 1.)

This process constructed 36,252 posts with both inferred gender and age information. Preprocessing steps included text standardization such as lowercasing and stopwords removal. This dataset included 72.33% (25,972/36,252) of females and 27.67% (9936/36,252) of males, with 78% (28,277/36,252) of those aged 18 - 35 years, reflecting a predominantly younger demographic.

XSL•FO
RenderX

Table . Sample AskaPatient dataset.

Name	Reason	Satisfaction	Side effects	Comments	Sex	Age (years)
Celexa	Anxiety and depression	4	Loss of appetite, headache, insomnia, and feeling numb	Definitely changing my life. I feel like a new person and ready to live every day like I never did. The side effects are not that bad, my body is slowly feeling normal again. Hoping not having any side effects soon!	Female	21
Celexa	Autism and anxiety	3	Increased aggression with higher doses (5 mg)	Took edge off my son's anxiety but didn't do much else. He was more aggressive when the dose was raised.	Male	9
Celexa	Fibromyalgia and depression	3	Weight gain and fatigue not helping with the pain from fibromyalgia	Quit taking feel like there is something that will work better. Having anger issues and uncontrollable outburst.	Female	56
Celexa	Anxiety	1	Hair loss, vision haze, heart palps, and laziness	This was the start of my downfall I lost all my hair from it and lost my life slowly avoid this crap as its led me down a path of destruction.	Male	24

Analytical Techniques

To gain insights into patient experiences with antidepressants, we applied several advanced analytical methods, offering multiple perspectives on sentiment, emotions, personality traits, and latent themes.

Sentimental Analysis

We used the *twitter-roberta-base-sentiment-latest* model to classify text into negative, neutral, or positive sentiment. This model was trained on 124 million tweets from 2018 to 2021 and fine-tuned using the TweetEval benchmark [40]. For example, the comment "The side effects are not that bad, my body is slowly feeling normal again" yielded probabilities of 0.67 for positive, 0.28 for neutral, and 0.04 for negative sentiment. A 0.7 threshold was used for overall sentiment classification, but the full probability distributions were also analyzed for more nuanced insights.

Emotional Analysis

We used the English DistilRoBERTa-base model [41], optimized for detecting 6 basic emotions—anger, disgust, fear, joy, sadness, and surprise—along with neutral. This model trained on diverse datasets and enabled the identification of emotional states influencing antidepressant use.

Personality Analysis

LLMs, such as GPT-4 [42], were used for personality predictions, focusing on the Big Five traits—openness, conscientiousness, extraversion, agreeableness, and neuroticism. Each comment was assigned personality scores (1-5). Role prompting [43], a powerful artificial intelligence technique, was applied to guide the LLMs' response in interpreting personality traits based on patient comments. (Detailed prompts can be found in Multimedia Appendix 1.)

Topic Modeling Insights

We used BERTopic [44], a topic modeling approach leveraging transformers and class-based term frequency-inverse document frequency to generate coherent topics. To enhance interpretability, we used GPT-4 to refine topic labels by analyzing keywords and representative documents. The following prompt was used (Textbox 1).

To further explore the themes of nonadherence and PCC within the generated topics, we selected the 2 predominate topics related to nonadherence (side effects, withdrawal symptoms, etc) and PCC (personalized treatment, communication with providers, etc). These topics were most populated, which allowed the study to focus on the exploration of how sentiment, emotion, and personality traits associated with the 2 themes.

XSL•FO
RenderX

Textbox 1. Large language model prompt for topic label generation.

```
Prompt = """
I have topic that contains the following documents: \n[DOCUMENTS]
The topic is described by the following keywords: [KEYWORDS]
Based on the above information, can you give a short label of the topic?
"""
```

Ethical Considerations

The authors confirm that this research adhered to ethical guidelines and legal requirements in the United States. For the data obtained from AskaPatient, a data use agreement was signed to ensure compliance with ethical standards. Reddit data were sourced from publicly available platforms [45] and did not contain any personally identifiable information. The dataset was fully anonymized, and no private identifiers were included.

The data used in this research were obtained from 2 publicly accessible sources: AskaPatient and Reddit. Access to AskaPatient data was granted under a signed terms of use agreement between Consumer Health Resource Group, LLC, and authors, which provides a nonexclusive, revocable, and limited license for academic purposes. This agreement ensures compliance with ethical and legal requirements, including the stipulation that no more than 10 individual records or postings are included in any publication, with proper attribution to the AskaPatient website. Data from Reddit were publicly available and did not include any private, sensitive, or identifiable information. This study was partially supported by the National Science Foundation project (IIS-2041065) and adhered to the ethical guidelines and legal requirements outlined in the institutional review board regulations of Kent State University (reference no. KSU IRB20-182) and was classified as “exempt” under 45 CFR §46.104 (d), the federal policy for the protection of human subjects in the United States.

Results

In our analysis, traditional statistical methods (eg, *t* tests) were used to validate the observed trends in sentiment and emotion across demographic groups, with detailed results provided in Multimedia Appendix 1.

Sentiment Analysis

Significant gender differences were found in sentiment scores (Figures S4-S6 in Multimedia Appendix 1). Negative sentiment was higher in males ($t_{35,905}$=−6.79; P<.001), while females exhibited higher neutral ($t_{35,905}$=4.47; P<.001) and positive sentiment ($t_{35,905}$=5.60; P<.001). The majority of posts were classified as neutral (21,499/36,253, 59.3%), followed by negative (13,922/36,253, 38.4%), and a smaller proportion of positive posts (798/36,253, 2.2%).

The gender difference in overall sentiment indicates that females exhibited 38.8% (10,181/26,245) negative, 59.1% (15,515/26,245) neutral, and 2.1% (551/26,245) positive sentiments, while males showed 37.6% (3764/10,008) negative, 59.8% (5986/10,008) neutral, and 2.6% (260/10,008) positive sentiments.

Across age groups, negative sentiment was notably higher in younger age groups (0 - 18 years), while neutral sentiment was lower in the 0 - 18 years age groups than in the older age groups, particularly 31 - 45 years age group ($t_{35,903}$=−15.77; P<.001). Positive sentiment was also significantly lower in younger age groups (0 - 18 years) than in older age groups.

Sentiment by age range (Figure S7 in Multimedia Appendix 1) indicated that the negative sentiment was highest in the 0 - 18 years age group (median score ~0.62) and gradually decreases with age, stabilizing around 0.52 for the 46 - 60 and 60+ years age groups. Neutral sentiment was relatively stable across age, with the highest scores in the 31 - 45 and 60+ years age groups (~0.37) and the lowest scores in the 0 - 18 years age group (~0.31). Positive sentiment remained consistently low across all ages, peaking in the 46 - 60 years age group (~0.14).

Emotional Analysis

Significant gender differences were found for anger, disgust, joy, neutral, and sadness (P<.05). No significant differences were observed for fear and surprise between genders (Figures S10-S12 in Multimedia Appendix 1). The emotion analysis shows that fear was the most prevalent emotion across the dataset, present in 44.66% (16,196/36,253) of posts, followed by sadness (12,507/36,253, 34.49%). Surprise (2455/36,253, 6.77%) and anger (2132/36,253, 5.88%) were less common, while positive emotions such as joy appeared in only 4.88% (1769/36,253) of posts. Neutral emotions accounted for 2.67% (968/36,253), with disgust being the least frequent emotion (232/36,253, 0.64%).

A comparison of the mean emotion scores between females (**F**) and males (**M**) showed that fear was dominant for both genders (~0.40), with sadness slightly higher in females and joy more expressed in males, although both emotions had low overall score. Anger, disgust, and surprise were reported at similarly low levels for both genders, with females showing slightly higher levels of sadness and anger.

Across different age groups (0 - 18, 19 - 30, 31 - 45, 46 - 60, and 60+ years), fear remained the most prevalent emotion, peaking in the younger users (0 - 18 years) and gradually declining with age. Sadness followed a similar pattern, with higher scores among younger users before decreasing. Joy and neutral emotions remain low across all age groups, with slight increases in users aged 46 years and older. Anger, disgust, and surprise showed no significant variation across age ranges.

XSL•FO
RenderX

Personality Analysis

A gender difference was observed in agreeableness ($t_{35,903}$=3.78; P<.001). No significant gender differences were found for openness, conscientiousness, extraversion, or neuroticism (Figures S1-S17 in Multimedia Appendix 1). The proportion of users who scored the highest level (5) across the 5 personality traits, separated by gender, showed that agreeableness was highest in males (429/10,008, 4.29%) compared with females (1034/26,24, 53.94%). Conscientiousness and extraversion showed similar proportions, with 4% (1039/25,972) of females and 4.06% (403/9936) of males scoring 5. Neuroticism exhibited minimal differences, with females at 3.97% (1031/25,972) and males at 4.05% (402/9936). Openness scores were also closely matched, with females scoring 4.07% (1056/25,972) and males scoring 3.93% (391/9,936).

Pairwise t tests across age groups for openness, conscientiousness, extraversion, agreeableness, and neuroticism showed no significant differences after adjustment (all P>.05). Personality trait scores varied across age group. Agreeableness showed little variation across age groups, peaking in the 60+ years age group (1511/36,253, 4.17%) and reaching its lowest in the younger than 30 years age group (1450/36,253, 4%). Conscientiousness remained stable, with slightly higher scores in the 19 - 30 and 46 - 60 years age groups (1462/36,253, 4.03%) and lower scores in the 60+ years age group (1355/36,253, 3.74%). Extraversion was highest in the 60+ years age group (1555/36,253, 4.29%) and lowest in the 0 - 18+ years age group (1348/36,253, 3.72%). Neuroticism peaked in the 60+ years age group (1627/36,253, 4.49%) and was lowest in the 46 - 60 years age group (1360/36,253, 3.75%). Openness followed a similar trend, with the 60+ years age group scoring the highest (1704/36,253, 4.70%) and the 46 - 60 years age group scoring the lowest (1414/36,253, 3.90%).

Topic Analysis

Table 2 shows the results of BERTopic analysis within dataset, identifying key topics related to antidepressants user experiences. The largest topic, “Mental Health and Relationships,” accounted for 11.69% (3755/32,121) of the discussions, indicating a significant focus on managing these mental health conditions. Other notable topics included “Lexapro Experiences and Alternatives” (2292/32,121, 7.14%) and “Zoloft Withdrawal and Anxiety” (2184/32,121, 6.80%), reflecting common concerns about antidepressant experience and desire to switch to alternative medicines to manage their anxiety and depression.

Several discussions centered around side effects, including “Medication and Sexual Dysfunction” (2146/32,121, 6.68%) , “Chronic Abdominal Pain Issues (1949/32,121, 6.07%), “Skin Rashes and Sensitivities” (1700/32,121, 5.70%), “Heart Symptoms and Anxiety” (1638/32,121, 5.10%), and “Chronic Pain and Health Issues” (1300/32,121, 4.05%), alongside concerns related to weight gain and sleep disorders, which highlight frequent side effects users encounter during treatment.

Two primary themes—nonadherence and self-care—were extracted from the dataset based on the highest documented counts from relevant topics identified by the BERTopic model. In the nonadherence theme, “Zoloft Withdrawal and Anxiety,” ” Effexor Experiences and Withdrawals,” “Cymbalta Withdrawal Experiences and Effects,” and “Psychiatric Drug Withdrawal Effects” were assigned. In the self-care theme, the topics “mental health and relationships,” “relationship struggles and mental health,” and “toxic relationships and mental health” were chosen. Furthermore, we investigate sentiment, emotion, and personality traits across the 2 themes.

Sentiment distribution varied between the themes of nonadherence and self-care (Figure S19 in Multimedia Appendix 1). Neutral sentiment was most prevalent in both, accounting for 57.70% (3679/6376) in nonadherence and 47.65% (2929/6147) in self-care. Negative sentiment followed closely, with 36.68% (2339/6376) in nonadherence and 50.69% (3116/6147) in self-care, while positive sentiment remained low (358/6376, 5.61% and 102/6147, 1.66%, respectively).

The emotion distribution has been compared across the themes of nonadherence and self-care (Figure S20 in Multimedia Appendix 1). Fear was dominant emotion in both themes, making up 55.83% (3560/6376) of emotions in nonadherence and 35.90% (2207/6147) in self-care. Sadness was more prominent in self-care (2733/6147, 44.46%) than in nonadherence (1566/6376, 24.56%). Anger was also more frequent in self-care (592/6147, 9.63%) compared with nonadherence (265/6376, 4.16%).

Personality traits (scoring 5) also differed across themes (Figure S21 in Multimedia Appendix 1). In the nonadherence theme, agreeableness was the most common trait (2023/6376, 31.73%), followed by extraversion (1983/6376, 31.10%) and neuroticism (1892/6376, 29.68%). In the self-care theme, neuroticism was highest (1891/6147, 30.76%), followed by openness (1887/6147, 30.69%), conscientiousness (1804/6147, 29.34%), and agreeableness (1789/6,147, 29.10%).

XSL•FO
RenderX

Table . BERTopic result.

Topic	OpenAI label	Document count	Proportion
0	Mental Health and Relationships	3755	11.69
1	Lexapro Experiences and Alternatives	2292	7.14
2	Zoloft Withdrawal and Anxiety	2184	6.80
3	Medication and Sexual Dysfunction	2146	6.68
4	ADHD[a] Medication Challenges and Insights	2090	6.51
5	Chronic Abdominal Pain Issues	1949	6.07
6	Skin Rashes and Sensitivities	1700	5.70
7	Chronic Pain and Fatigue Issues	1651	5.14
8	Heart Symptoms and Anxiety	1638	5.10
9	Abnormal Blood Test Concerns	1446	4.50
10	Chronic Pain and Health Issues	1300	4.05
11	Weight Gain and Medication Struggles	1272	3.96
12	Relationship Struggles and Mental Health	1260	3.92
13	Experiences With Prozac for Anxiety	1176	3.66
14	Irregular Periods and Birth Control	1146	3.57
15	Toxic Relationships and Mental Health	1132	3.53
16	Wellbutrin Experiences and Challenges	1006	3.13
17	Chronic Sleep Disorders and Fatigue	990	3.08
18	Dizziness and Anxiety Symptoms	788	2.45
19	Chronic Headache and Migraine	773	2.40
20	Psychiatric Drug Withdrawal Effects	766	2.38
21	Anxiety and Substance Reactions	734	2.28
22	Effexor Experiences and Withdrawals	681	2.12
23	Chronic UTI[b] Symptoms and Challenges	517	1.61
24	Cymbalta Withdrawal Experiences and Effects	453	1.41
25	Alcohol, Anxiety, and Treatment Concerns	402	1.25
27	Eye Health and Vision Concerns	385	1.20
28	Chronic Ear Infections and Tinnitus	274	0.85
29	Managing Anxiety and Depression Tips	2	0.01

[a]ADHD: attention-deficit/hyperactivity disorder.

[b]UTI: urinary tract infection.

XSL•FO
RenderX

Discussion

Principal Findings

The finding reveals that sentiment of most of the antidepressant users is neutral within the dataset, while there remains substantial negative sentiment, and few show positive. In the meantime, the emotional analysis exhibits that fear and sadness command patients' emotional response. Those findings may consider the experiences when the patients use those antidepressants during the daily life. The community and health care can provide support to reduce the fear of side effects or failure of antidepressants, and family can facilitate comfort toward the sadness and negative emotions. Consequently, antidepressant users can practice self-care and enhance their treatment adherence.

Personality has not indicated a significant difference within the dataset. Topic analysis expresses that discussions mainly emphasize managing anxiety and depression, the efficacy of medications, and their side effects. Furthermore, in the nonadherence theme, the neutral sentiment was more prominent, followed by the significant fear emotion. In the self-care theme, the negative sentiment was more distinct, with the sadness emotion most dominant. Those findings suggest that the most beneficial route for antidepressant users is to exercise focused self-care for their mental health issues, while continuing their antidepressants treatment.

RQ1: How Do Sentiment Analyses Reveal Differences in Antidepressant Experiences Across Demographics?

The sentiment analysis indicates that public perception of antidepressants is predominantly neutral, with a substantial portion reflecting negative sentiment and very few positive expressions. The neutral sentiment suggests ambivalence or cautious optimism toward antidepressants, while the limited positive sentiment points to restricted satisfaction with treatment outcomes. These trends reflect broader society uncertainty or skepticism about antidepressant effectiveness and side effects, consistent with the literature identifying nonadherence as a pervasive issue [46].

In terms of gender, the results show minor differences in sentiment between males and females, with females exhibiting slightly higher negative sentiment and lower positive sentiment than males. This may reflect the differences in antidepressant experiences, societal pressures, or varying expectations regarding treatment outcomes across genders. Addressing these gender-specific concerns may enhance treatment satisfaction and adherence of antidepressants.

Notably, younger individuals expressed the highest levels of negative sentiment, which gradually decreases with age. This could be attributed to heightened sensitivity to side effects or unaccomplished expectations from treatment among younger users [47]. Older age groups demonstrated more stable neutral sentiment and slightly higher positive sentiment, particularly in the 46 - 60 years age group. These findings underscore the importance of patient's sentiment during antidepressant treatment. Mental health may benefit from tailoring interventions based on these emotional responses. Therefore, addressing negative bias early may be a key mechanism of antidepressant drug action and a potentially useful predictor of therapeutic response [48].

RQ2: What Predominant Emotions (eg, Fear, Sadness, and Joy) Are Associated With Antidepressant Use, and How Do These Vary by Demographics?

The emotional analysis revealed that fear and sadness dominate patients' emotional responses, with fear accounting for nearly half (44.66%) and sadness is the second most prominent emotion (34.49%). These findings suggest that patients' experiences with antidepressants are characterized by deep concerns and dissatisfaction. The prominence of fear, especially among younger users, may indicate anxiety related to side effects, potential long-term dependence, or uncertainties regarding the treatment's efficacy. This aligns with research showing that emotional blunting, commonly associated with antidepressants use, is also a symptom of depression and is linked to poorer remission outcomes [49].

The prominence of fear across both genders highlights widespread concerns about side effects, dependence, and treatment efficacy. However, the slightly higher level of sadness and anger in females suggests gender-specific emotional responses to antidepressants. Tailored communication strategies that address these gender-specific emotional concerns could improve treatment satisfaction and adherence.

As patients age, fear tends to decrease, potentially due to greater familiarity with antidepressants or prolonged exposure to treatment. However, the emotional burden remains strong among younger users, emphasizing the need for enhanced communication regarding treatment plans and potential side effects. The high prevalence of sadness**,** particularly among younger users, may reflect frustration or disappointment with treatment outcomes, potentially linked to unmet expectations, a lack of education over antidepressant, or emotional blunting. The intent of providing time-phased educational materials to patients is to maximize the relevance of such information by synchronizing it with typical recovery processes and issues. Additional efforts at engaging patients earlier after the initiation of treatment might be of most benefit [50].

RQ3: What Personality Traits (eg, Extraversion and Neuroticism) Are Prevalent Among Users Discussing Antidepressants, and How Do These Traits Differ Across Demographics?

Personality traits exhibited minor yet meaningful variations across gender and age groups. Agreeableness was higher in males, suggesting that empathy and social cooperation may positively influence adherence to mental health treatment. Concerning the age difference, one study found that the lower the extroversion score, the better the response to antidepressant treatment for adolescents [51]. From our finding, extraversion was highest in the 60+ years age group (4.29%) and lowest in the 0 - 18 years age group (3.72%). Compared with the higher negative sentiment, fear, and sadness emotions among the 0 - 18 years age group, the health care providers may tailor treatment to maximize the outcome of antidepressant medication.

XSL•FO
RenderX

Agreeableness is dominant in the nonadherence theme; neuroticism is the largest personality in the self-care theme. These are not notable findings that align with existing literatures for personality impact to antidepressant adherence and self-care; the result may provide a new perspective in the future research.

RQ4: What Are the Prevailing Topics in Web-Based Antidepressant Discussions, and How Do Sentiment, Emotion, and Personality Insights Inform Our Understanding of Nonadherence and Self-Care Themes?

The topic analysis identified that discussions primarily focus on mental health and relationships, medication efficacy, and side effects such as sexual dysfunction and withdrawal symptoms. These findings highlight clear concerns among users regarding the mental health issues and overall effectiveness of antidepressants. The dominance of neutral and negative sentiment in these discussions, combined with the high presence of fear and sadness, indicates that many users remain uncertain about their treatment outcomes and dissatisfy their daily life regarding their experience with the antidepressants.

Discussions around nonadherence are marked by fear, followed by sadness emotion, with patients expressing concerns about withdrawal symptoms and medication efficacy. In contrast, self-care discussions show a notable trend at sadness, accompanied by an increase in expressions of anger. A study found that stigma and fear of adverse events inhibited initiation of therapy, while adverse events and ineffectiveness of antidepressants contributed to discontinuation. Patients with strong perceptions of the necessity and few concerns about antidepressants were more likely to adhere to treatment at all phases of adherence [52].

Limitations

This study has several limitations. First, it relies on self-reported data from AskaPatient and Reddit, which may introduce bias as individuals who share their experiences on the web may not represent the broader population of antidepressant users. These users may disproportionately report extreme experiences, potentially skewing the findings. Furthermore, the focus on social media users limits the generalizability of the results to the wider population.

Although sentiment and emotion analysis tools are sophisticated, they may not fully capture the complexity of human emotions, particularly mixed or nuanced sentiments. Personality traits inferred from text via LLMs might not always align with real-world behaviors, possibly leading to misinterpretations. Furthermore, the study provides a snapshot of discussions at a particular time, without tracking changes over time. Furthermore, while age and gender are considered in the analysis, other demographic factors, such as socioeconomic status, education, and geographic location, were not controlled for and may influence user experiences with antidepressants. These factors should be considered in future studies to provide a more comprehensive understanding of patients' antidepressant use.

Conclusions

This study offers valuable insights into public perceptions of antidepressants, revealing that neutral and negative sentiments dominate discussions, with fear and sadness as the most common emotions, particularly among younger users. Personality traits exhibited minor yet meaningful variations across gender and age groups. Agreeableness was higher in males, suggesting that empathy and social cooperation may positively influence adherence to mental health treatment.

The topic analysis revealed that discussions primarily revolve around mental health and relationships, medication efficacy, and side effects, such as sexual dysfunction and withdrawal symptoms. Discussions about nonadherence are characterized by fear as the dominant emotion, followed by sadness, with patients voicing concerns about withdrawal symptoms and the effectiveness of medications. In contrast, self-care discussions convey a predominant sense of sadness, accompanied by heightened expression of anger.

The findings underscore the importance of self-care that addresses psychological dimensions of antidepressant use. Tailored interventions based on individual negative sentiment, fear, and sadness emotional challenges could improve overall treatment outcomes. By leveraging real-time social media data and advanced analytical tools, this research contributes to a more patient-centered understanding of antidepressant use, offering pathways for improving patient care and enhancing their nonadherence for antidepressants.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed supplementary data on public perceptions of antidepressants discussed on social media, including topic analysis results, sentiment breakdowns, and associated statistical summaries.
[DOCX File, 1836 KB - formative_v9i1e62680_app1.docx]

References

1. Brody DJ, Gu Q. Antidepressant use among adults: United States, 2015-2018. NCHS Data Brief 2020 Sep(377):1-8. [Medline: 33054926]
2. Marasine NR, Sankhi S. Factors associated with antidepressant medication non-adherence. Turk J Pharm Sci 2021 Apr 20;18(2):242-249. [doi: 10.4274/tjps.galenos.2020.49799] [Medline: 33902271]

XSL•FO
RenderX

3. Sansone RA, Sansone LA. Antidepressant adherence: are patients taking their medications? Innov Clin Neurosci 2012 May;9(5-6):41-46. [Medline: 22808448]
4. Deen TL, Fortney JC, Pyne JM. Relationship between satisfaction, patient-centered care, adherence and outcomes among patients in a collaborative care trial for depression. Adm Policy Ment Health 2011 Sep;38(5):345-355. [doi: 10.1007/s10488-010-0322-z] [Medline: 20978932]
5. Rathert C, Wyrwich MD, Boren SA. Patient-centered care and outcomes: a systematic review of the literature. Med Care Res Rev 2013 Aug;70(4):351-379. [doi: 10.1177/1077558712465774] [Medline: 23169897]
6. Kramer PD. Listening to Prozac: The Landmark Book about Antidepressants and the Remaking of the Self: Penguin; 1997.
7. Opbroek A, Delgado PL, Laukes C, et al. Emotional blunting associated with SSRI-induced sexual dysfunction. Do SSRIs inhibit emotional responses? Int J Neuropsychopharmacol 2002 Jun;5(2):147-151. [doi: 10.1017/S1461145702002870] [Medline: 12135539]
8. Sansone RA, Sansone LA. SSRI-induced indifference. Psychiatry (Edgmont) 2010 Oct;7(10):14-18. [Medline: 21103140]
9. Stone AA, Shiffman S, Schwartz JE, Broderick JE, Hufford MR. Patient non-compliance with paper diaries. BMJ 2002 May 18;324(7347):1193-1194. [doi: 10.1136/bmj.324.7347.1193] [Medline: 12016186]
10. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol 2008;4(1):1-32. [doi: 10.1146/annurev.clinpsy.3.022806.091415] [Medline: 18509902]
11. Saha K, Sugar B, Torous J, Abrahao B, Kıcıman E, De Choudhury M. A social media study on the effects of psychiatric medication use. Proc Int AAAI Conf Weblogs Soc Media 2019;13:440-451. [doi: 10.1609/icwsm.v13i01.3242]
12. Saha K, Torous J, Kiciman E, De Choudhury M. Understanding side effects of antidepressants: large-scale longitudinal study on social media data. JMIR Ment Health 2021 Mar 19;8(3):e26589. [doi: 10.2196/26589] [Medline: 33739296]
13. Zhu J, Jin R, Kenne DR, Phan N, Ku WS. User dynamics and thematic exploration in r/depression during the COVID-19 pandemic: insights from overlapping r/SuicideWatch users. J Med Internet Res 2024 May 20;26:e53968. [doi: 10.2196/53968] [Medline: 38767953]
14. Carabot F, Fraile-Martínez O, Donat-Vargas C, et al. Understanding public perceptions and discussions on opioids through Twitter: cross-sectional infodemiology study. J Med Internet Res 2023 Oct 31;25:e50013. [doi: 10.2196/50013] [Medline: 37906234]
15. Park A, Conway M. Tracking health related discussions on Reddit for public health applications. AMIA Annu Symp Proc 2017;2017:1362-1371. [Medline: 29854205]
16. Zhu J, Yalamanchi N, Jin R, Kenne DR, Phan N. Investigating COVID-19's impact on mental health: trend and thematic analysis of reddit users' discourse. J Med Internet Res 2023 Jul 12;25:e46867. [doi: 10.2196/46867] [Medline: 37436793]
17. Stade EC, Stirman SW, Ungar LH, et al. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. NPJ Ment Health Res 2024 Apr 2;3(1):12. [doi: 10.1038/s44184-024-00056-z] [Medline: 38609507]
18. Yang K, Zhang T, Kuang Z, Xie Q, Huang J, Ananiadou S. MentaLLaMA: interpretable mental health analysis on social media with large language models. Presented at: WWW '24; May 13-17, 2024; Singapore, Singapore. [doi: 10.1145/3589334.3648137]
19. Chowdhury AK, Sujon SR, Shafi MSS, et al. Harnessing large language models over transformer models for detecting Bengali depressive social media text: a comprehensive study. Nat Lang Process J 2024 Jun;7:100075. [doi: 10.1016/j.nlp.2024.100075]
20. Pompili M, Venturini P, Palermo M, et al. Mood disorders medications: predictors of nonadherence—review of the current literature. Expert Rev Neurother 2013 Jul;13(7):809-825. [doi: 10.1586/14737175.2013.811976] [Medline: 23898852]
21. Kamaradova D, Latalova K, Prasko J, et al. Connection between self-stigma, adherence to treatment, and discontinuation of medication. Pat Prefer Adherence 2016;10:1289-1298. [doi: 10.2147/PPA.S99136] [Medline: 27524884]
22. De las Cuevas C, Peñate W, Sanz EJ. Risk factors for non-adherence to antidepressant treatment in patients with mood disorders. Eur J Clin Pharmacol 2014 Jan;70(1):89-98. [doi: 10.1007/s00228-013-1582-9] [Medline: 24013851]
23. Sawada N, Uchida H, Suzuki T, et al. Persistence and compliance to antidepressant treatment in patients with depression: a chart review. BMC Psychiatry 2009 Jun 16;9:38. [doi: 10.1186/1471-244X-9-38] [Medline: 19531229]
24. Epstein RM, Street RL. The values and value of patient-centered care. Ann Fam Med 2011;9(2):100-103. [doi: 10.1370/afm.1239] [Medline: 21403134]
25. 2023 webinars planning information. UNC Lineberger Cancer Network. URL: https://unclineberger.org/unclcn/wp-content/uploads/sites/871/2024/03/2023-Telehealth-Webinar-Planning-Guide-1.pdf [accessed 2024-11-24]
26. Levinson W. Patient-centred communication: a sophisticated procedure. BMJ Qual Saf 2011 Oct;20(10):823-825. [doi: 10.1136/bmjqs-2011-000323]
27. McPherson S, Wicks C, Tercelli I. Patient experiences of psychological therapy for depression: a qualitative metasynthesis. BMC Psychiatry 2020 Jun 18;20(1):313. [doi: 10.1186/s12888-020-02682-1] [Medline: 32552748]
28. Ell G. Encouraging self-care and positive lifestyle changes in patients with depression. Pharm J 2020 [FREE Full text]
29. Sidana S, Mishra S, Amer-Yahia S, Clausel M, Amini MR. Health monitoring on social media over time. Presented at: SIGIR '16; Jul 7-21, 2016; Pisa Italy. [doi: 10.1145/2911451.2914697]

XSL•FO
RenderX

30. Naslund JA, Bondre A, Torous J, Aschbrenner KA. Social media and mental health: benefits, risks, and opportunities for research and practice. J Technol Behav Sci 2020 Sep;5(3):245-257. [doi: 10.1007/s41347-020-00134-x] [Medline: 33415185]
31. Roeder A. Social Media Use Can Be Positive for Mental Health and Well-Being: Harvard TH Chan School of Public Health; 2020.
32. Rezaii N, Wolff P, Price BH. Natural language processing in psychiatry: the promises and perils of a transformative approach. Br J Psychiatry 2022 Jan 7;220(5):1-3. [doi: 10.1192/bjp.2021.188] [Medline: 35048814]
33. Stewart R, Velupillai S. Applied natural language processing in mental health big data. Neuropsychopharmacology 2021 Jan;46(1):252-253. [doi: 10.1038/s41386-020-00842-1] [Medline: 32895453]
34. Serapio-García G, Safdari M, Crepy C, et al. Personality traits in large language models. Research Square. Preprint posted online on Aug 8, 2023. [doi: 10.21203/rs.3.rs-3296728/v1]
35. Guo X, Sun Y, Vosoughi S. Emotion-based modeling of mental disorders on social media. Presented at: WI-IAT ’21; Dec 14-17, 2021; Melbourne, Australia. [doi: 10.1145/3486622.3493916]
36. de Anta L, Alvarez-Mon MA, Ortega MA, et al. Areas of interest and social consideration of antidepressants on English tweets: a natural language processing classification study. J Pers Med 2022 Jan 25;12(2):155. [doi: 10.3390/jpm12020155] [Medline: 35207644]
37. Duraisamy P, Natarajan Y, Preethaa KRS, Mouthami K. Sentiment analysis on drug reviews using diverse classification techniques. Presented at: 2022 3rd International Conference on Communication, Computing and Industry 40 (C2I4); Dec 15-16, 2022; Bangalore, India. [doi: 10.1109/C2I456876.2022.10051399]
38. AskaPatient. URL: https://www.askapatient.com/ [accessed 2025-02-13]
39. Reddit. URL: https://www.reddit.com/ [accessed 2025-02-13]
40. Barbieri F, Camacho-Collados J, Espinosa Anke L, Neves L. TweetEval: unified benchmark and comparative evaluation for tweet classification. Presented at: Findings of the Association for Computational Linguistics; Nov 16-20, 2020. [doi: 10.18653/v1/2020.findings-emnlp.148]
41. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv. Preprint posted online on Oct 2, 2019. [doi: 10.48550/arXiv.1910.01108]
42. OpenAI. GPT-4 technical report. 2023. URL: https://openai.com/research/gpt-4 [accessed 2025-02-13]
43. Leonard L. Role prompting: the secret to effective AI interactions. The Software Forum. URL: https://www.thesoftwareforum.com/role-prompting [accessed 2025-02-13]
44. Grootendorst M. BERTopic: neural topic modeling with a class-based TF-IDF procedure. arXiv. Preprint posted online on Mar 11, 2022. [doi: 10.48550/arXiv.2203.05794]
45. Reddit comments/submissions 2005-06 to 2023-09. Academic Torrents. URL: https://academictorrents.com/details/89d24ff9d5fbc1efcdaf9d7689d72b7548f699fc [accessed 2025-02-13]
46. Dell’Osso B, Albert U, Carrà G, et al. How to improve adherence to antidepressant treatments in patients with major depression: a psychoeducational consensus checklist. Ann Gen Psychiatry 2020;19:61. [doi: 10.1186/s12991-020-00306-2] [Medline: 33062034]
47. Luft MJ, Lamy M, DelBello MP, McNamara RK, Strawn JR. Antidepressant-induced activation in children and adolescents: risk, recognition and management. Curr Probl Pediatr Adolesc Health Care 2018 Feb;48(2):50-62. [doi: 10.1016/j.cppeds.2017.12.001] [Medline: 29358037]
48. Godlewska BR, Browning M, Norbury R, Cowen PJ, Harmer CJ. Early changes in emotional processing as a marker of clinical response to SSRI treatment in depression. Transl Psychiatry 2016 Nov 22;6(11):e957. [doi: 10.1038/tp.2016.130] [Medline: 27874847]
49. Goodwin GM, Price J, De Bodinat C, Laredo J. Emotional blunting with antidepressant treatments: a survey among depressed patients. J Affect Disord 2017 Oct 15;221:31-35. [doi: 10.1016/j.jad.2017.05.048] [Medline: 28628765]
50. Mundt JC, Clarke GN, Burroughs D, Brenneman DO, Griest JH. Effectiveness of antidepressant pharmacotherapy: the impact of medication compliance and patient education. Depress Anxiety 2001;13(1):1-10. [doi: 10.1002/1520-6394(2001)13:1<1::aid-da1>3.0.co;2-s] [Medline: 11233454]
51. Ran LY, Liu XY, Wang W, et al. Personality traits predict treatment outcome of an antidepressant in untreated adolescents with depression: an 8-week, open-label, flexible-dose study. J Affect Disord 2024 Apr 1;350:102-109. [doi: 10.1016/j.jad.2024.01.015] [Medline: 38199422]
52. Gupta R, Ariefdjohan M. Mental illness on Instagram: a mixed method study to characterize public content, sentiments, and trends of antidepressant use. J Ment Health 2021 Aug;30(4):518-525. [doi: 10.1080/09638237.2020.1755021] [Medline: 32325006]

Abbreviations

LLM: large language model
NLP: natural language processing
PCC: patient-centered care
RQ: research question

XSL•FO
RenderX

Edited by A Mavragani; submitted 28.05.24; peer-reviewed by E Chan, H Rafi, R Norel; revised version received 08.11.24; accepted 13.11.24; published 26.02.25.

Please cite as:
Zhu J, Zhang X, Jin R, Jiang H, Kenne DR
Probing Public Perceptions of Antidepressants on Social Media: Mixed Methods Study
JMIR Form Res 2025;9:e62680
URL: https://formative.jmir.org/2025/1/e62680
doi:10.2196/62680

© Jianfeng Zhu, Xinyu Zhang, Ruoming Jin, Hailong Jiang, Deric R Kenne. Originally published in JMIR Formative Research (https://formative.jmir.org), 26.2.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Novel Evaluation Metric and Quantified Performance of ChatGPT-4 Patient Management Simulations for Early Clinical Education: Experimental Study

Riley Scherr[1], BS; Aidin Spina[1], BS; Allen Dao[2], BS; Saman Andalib[1*], BS; Faris F Halaseh[1*], BS; Sarah Blair[3], BS; Warren Wiechmann[4], MD, MBA, MSED; Ronald Rivera[4], MD

[1]School of Medicine, University of California, Irvine School of Medicine, 1001 Health Sciences Road, Irvine, CA, United States

[2]Department of Medicine, Stanford Medicine, Stanford, CA, United States

[3]School of Medicine, Tufts University, Boston, MA, United States

[4]Department of Emergency Medicine, University of California Irvine, Orange, CA, United States

[*]these authors contributed equally

Corresponding Author:
Riley Scherr, BS
School of Medicine, University of California, Irvine School of Medicine, 1001 Health Sciences Road, Irvine, CA, United States

Abstract

Background: Case studies have shown ChatGPT can run clinical simulations at the medical student level. However, no data have assessed ChatGPT's reliability in meeting desired simulation criteria such as medical accuracy, simulation formatting, and robust feedback mechanisms.

Objective: This study aims to quantify ChatGPT's ability to consistently follow formatting instructions and create simulations for preclinical medical student learners according to principles of medical simulation and multimedia educational technology.

Methods: Using ChatGPT-4 and a prevalidated starting prompt, the authors ran 360 separate simulations of an acute asthma exacerbation. A total of 180 simulations were given correct answers and 180 simulations were given incorrect answers. ChatGPT was evaluated for its ability to adhere to basic simulation parameters (stepwise progression, free response, interactivity), advanced simulation parameters (autonomous conclusion, delayed feedback, comprehensive feedback), and medical accuracy (vignette, treatment updates, feedback). Significance was determined with χ^2 analyses using 95% CIs for odds ratios.

Results: In total, 100% (n=360) of simulations met basic simulation parameters and were medically accurate. For advanced parameters, 55% (200/360) of all simulations delayed feedback, while the Correct arm (157/180, 87%) delayed feedback was significantly more than the Incorrect arm (43/180, 24%; $P<.001$). A total of 79% (285/360) of simulations concluded autonomously, and there was no difference between the Correct and Incorrect arms in autonomous conclusion (146/180, 81% and 139/180, 77%; $P=.36$). Overall, 78% (282/360) of simulations gave comprehensive feedback, and there was no difference between the Correct and Incorrect arms in comprehensive feedback (137/180, 76% and 145/180, 81%; $P=.31$). ChatGPT-4 was not significantly more likely to conclude simulations autonomously ($P=.34$) and provide comprehensive feedback ($P=.27$) when feedback was delayed compared to when feedback was not delayed.

Conclusions: These simulations have the potential to be a reliable educational tool for simple simulations and can be evaluated by a novel 9-part metric. Per this metric, ChatGPT simulations performed perfectly on medical accuracy and basic simulation parameters. It performed well on comprehensive feedback and autonomous conclusion. Delayed feedback depended on the accuracy of user inputs. A simulation meeting one advanced parameter was not more likely to meet all advanced parameters. Further work must be done to ensure consistent performance across a broader range of simulation scenarios.

(JMIR Form Res 2025;9:e66478) doi:10.2196/66478

KEYWORDS

medical school simulations; AI in medical education; preclinical curriculum; ChatGPT; ChatGPT-4; medical simulation; simulation; multimedia; feedback; medical education; medical student; clinical education; pilot study; patient management

XSL•FO
RenderX

Introduction

Background

With the rise of generative artificial intelligence (AI) applications such as OpenAI's ChatGPT, research into its medical application has expanded. Early investigations assessed whether large language models (LLMs) could pass medical trainee licensing examinations [1-5]. Such studies indicated that LLMs can pass medical exams and possess solid foundations in medical reasoning [3,6-8].

An expanding body of literature has focused on LLMs in medical education, including the perspectives of both students and seasoned clinicians [9-13]. Medical students have shown interest in LLMs and frequently use or intend to use them educationally [14]. Studies have also suggested that ChatGPT can be an effective tool for students when entering the clinical wards [15]. Moving from theoretical, classroom-based instruction to hands-on patient care introduces new challenges [16]. The steep learning curve heightens the need for reliable training tools, and generative AI technologies may satisfy this need.

Current literature suggests a need for rigorous validation of student use of generative AI. Frameworks such as RISE (role, input, steps, expectations; ie, inputting the LLM's role, anticipated input, required steps, and desired expectations) exist for prompt engineering but have been infrequently applied to medical student LLM use [17]. Initial work established prompts for medical trainees to practice common clinical scenarios in platforms such as ChatGPT [9,18]. These studies established that, given the precise wording of prompts, LLMs can act as an effective simulator of basic clinical scenarios. Despite this promising result, these technologies' unknown accuracy and precision in displaying desirable parameters limits their broader applicability [19-21].

In this study, we reviewed evidence-based resources for medical simulation and multimedia educational design. Using The Society for Simulation in Healthcare guidelines along with Richard Mayer's multimedia design principles, we created an evaluation system with 3 main evaluation categories: basic simulation parameters, advanced simulation parameters, and medical accuracy parameters. Each category was further divided into 3 subparameters. Basic parameters were composed of stepwise progression, free response, and interactivity. Advanced parameters were composed of autonomous conclusion, delayed feedback, and comprehensive feedback. Medical accuracy parameters were composed of clinical vignette, updates based on treatment, and feedback. See Figure 1 for definitions.

XSL•FO
RenderX

Figure 1. Simulation parameters. Desirable simulation parameters of ChatGPT clinical simulations were based on multimedia educational technology principles and simulation in health care recommendations. Three major parameters were each divided into 3 corresponding subparameters with clear definitions. Subparameters constitute the total 9-part metric by which ChatGPT simulations were evaluated. "P" or green indicates a pass, and "F" or red indicates a fail for whether a simulation exhibits that characteristic.

Basic Parameters

Stepwise progression in simulation is supported by the segmenting principle of multimedia learning, meaning that learning is maximized when broken into smaller units [22]. At breaks between units, the learner provides the next steps in the simulated patient's management with reasoning. This is supported by Mayer's self-explanation principle, which shows strong learning when previous knowledge is integrated into current learning by explaining one's reasoning [23]. An user-responsive design further facilitates this and creates a generative learning activity poised to improve learning acquisition and retention [24]. These are both in-line with real medical practice where every step of a treatment or diagnosis must be chosen. The free-response format for responses similarly mimics real medical practice, as it allows space for treatment justifications. It is also supported by the active processing principle, which asserts that learning is improved when people actively organize information into cognitive models and integrate prior knowledge to address the new task, rather than passively absorbing teaching points. Just like in real life, where physicians create treatment plans de novo, rather than choosing from a selection of options [25]. Finally, interactivity promotes cognitive engagement and is beneficial to learning outcomes, especially when planned and augmented by peer-to-peer and peer-to-teacher learning [26,27].

Advanced Parameters

Advanced parameters were selected to improve authenticity and decrease the cognitive load devoted to simulation mechanics, allowing greater focus on learning engagement [28]. An autonomous conclusion is valuable because it provides a specific end point and prevents learners from following extraneous threads when learning goals have been achieved. Learners do not have to decide when they have learned enough, rather they must meet the simulation goals to finish. Delayed feedback

XSL•FO
RenderX

keeps novice learners from overrelying on immediate feedback cues to make decisions and promotes active learning in simulations [29]. However, Mayer suggests immediate feedback may be better for some tasks such as solo novice learning or guided reasoning with an expert, but the Society for Simulation in Healthcare guidelines suggest that delayed debrief is most appropriate, and we accordingly chose delayed feedback for our metric [30]. This also mimics the real-life practice of medicine, where the only indication of a treatment's accuracy is a change in patient stability. Delaying feedback also keeps these simulations in-line with the active-processing principle that active learning is more effective than passive learning from regular performance feedback cues [29]. Finally, comprehensive feedback falls in-line with Mayer's signaling principle. Signaling, or indicating what information is vital, highlights specific learning points rather than generalizing about a user's performance [31]. Comprehensive feedback thus ensures meaningful performance assessments with specific takeaways.

Medical Accuracy Parameters

Medical accuracy is perhaps the most important. This is a common-sense principle that undergirds all the other parameters' efficacy. Medical accuracy categories were defined whenever there were physiologic, pharmacologic, or pathologic assertions.

A 9-part model for evaluating AI medical simulations is therefore proposed. Using a previously engineered and verified starting prompt, this model helps quantify the consistency of an LLM (ChatGPT-4) in following prompt commands to create simulations exhibiting these 9 features of effective simulations. The provided results aim to better inform recommendations to students and educators who use such resources in their clinical education and establish a method for future evaluation of simulation reliability.

Methods

Study Design

The Society for Simulation in Healthcare recommends simulation sessions be high or low fidelity with low physical realism, making AI simulation sessions acceptable when high-fidelity sessions are unavailable [30]. This is also congruent with the authentic learning environments principle which suggests that learning can happen equivalently in any environment as long as the design adheres to effective learning principles [32]. The Society also recommends training should be spaced, frequent, short, and skill-oriented sessions. Therefore, we built our prompt around asthma exacerbation management, a short lesson with evidence-based best practices to be done in 5 to 10 minutes with specific learning takeaways. This is further supported by the Society's recommendation that each session end with focused feedback via structured debrief, so our prompt accordingly addressed this. The only criteria we were unable to meet were recommendations for in-situ practice and the use of interprofessional teams. However, the authentic learning environments principle again suggests that these simulations can have educational benefits even though they are not in situ, given our use of educational design principles. The following prompt, developed with features of RISE prompt engineering, was subsequently used [9]:

"Please create a clinical scenario on a patient presenting to the hospital with an acute asthma exacerbation and quiz me on what the proper next step of management is. Please make it free response and interactive, meaning you ask me what the next step is one question at a time, and then I write out what I would do, and then you ask me another question based on how my answer would affect the patient. Please update/change the patient's condition based on my actions, and do not tell me the right answers until the end of the scenario."

Simulations were run on 1 of 3 ChatGPT-4 accounts due to ChatGPT-4's token limits and split across Safari, Google Chrome, and Firefox browsers from March 29, 2024, to April 24, 2024. Each simulation was run on its own browser tab session of ChatGPT-4 to eliminate session-dependent memory. All cases simulated an acute asthma exacerbation and focused on treatment steps.

An appropriate treatment algorithm for acute asthma exacerbation was developed based on the review of recommendations from the American Thoracic Society (a leading pulmonology society) and the American Academy of Family Physicians, and confirmed by author RR, a board-certified emergency medicine physician at a high-volume academic medical center [33]. Movement to the next step of the algorithm depended on the simulated patient's clinical condition, and the tester could proceed to the final step (discharge) if the patient was appropriately stabilized in a stepwise manner. For example, if a patient was no longer in respiratory distress and had normal oxygen saturation after administration of a short-acting muscarinic antagonist, the tester would skip magnesium sulfate and endotracheal intubation and proceed to patient education and counseling. See Figure 2 for treatment workflow.

XSL•FO
RenderX

Figure 2. Acute asthma simulation treatment algorithms. Treatment algorithms for acute asthma exacerbation simulations were developed on societal guidelines and were used as a template for ChatGPT simulation inputs. The green pathway represents "correct treatment" and was followed for 180 simulations in the correct subgroup. The red path represents "incorrect treatment," a variation of the correct treatment algorithm, and was followed for 180 simulations in the incorrect subgroup.

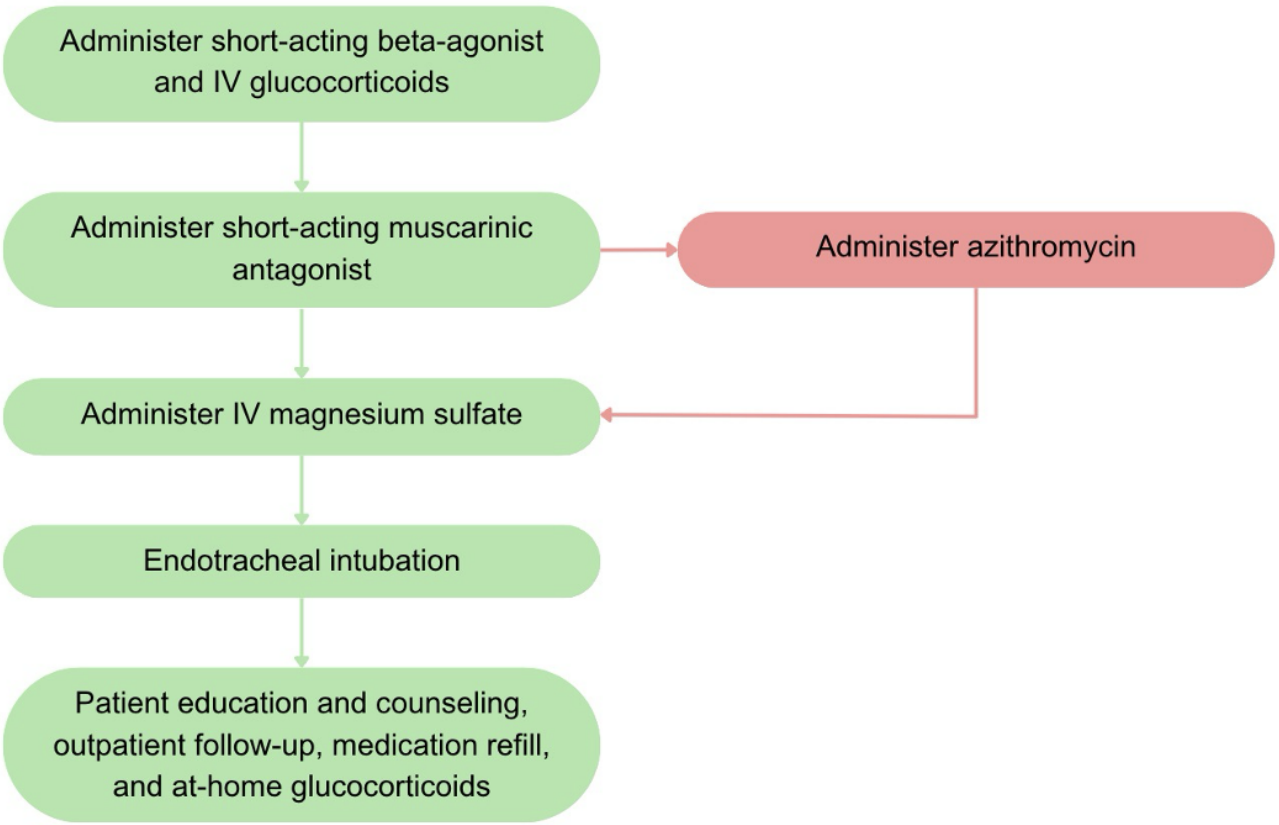

Simulations were divided into 2 subgroups based on adherence to the treatment workflow. The "Correct Treatment" subgroup followed the treatment workflow. The "Incorrect Treatment" subgroup added administration of a macrolide antibiotic (azithromycin) as the third step (Figure 2). Progression through the algorithm for each simulation depended on the patient's response to treatment, but all incorrect simulations included steps 1, 2, 3, and 6 at minimum. In total, 180 simulations each were run for the Correct and Incorrect treatment arms, leaving a total of 360 simulations for analysis. Simulations were evaluated according to our 9-part model using the definitions in Figure 1.

Author RS resolved unclear data points. Simulations were reviewed by all authors for medical accuracy and confirmed with author RR, a licensed, board-certified emergency medicine physician with 8 years of clinical experience.

Descriptive statistics were run on GraphPad Prism (version 10.2.3; Graphstats Technologies) and Excel (version 16.85; Microsoft Corp). Statistical significance was determined by χ^2 test, with an α level set at .05.

Ethical Considerations

Study data were collected by an author (AD) and did not involve personal information or observation of any person's private or public information. As such data were collected from ChatGPT, all study data were inherently anonymous. This study was reviewed under the University of California, Irvine's Institutional Review Board Protocol #3213 and deemed not human subjects research and not requiring institutional review board review.

Results

A summary of the results is listed in Table 1 and represented in Figure 3.

XSL•FO
RenderX

Table . Simulation characteristics[a].

	Combined (n=360), n (%)	Correct (n=180), n (%)	Incorrect (n=180), n (%)	χ^2 (*df*)	*P* value[b]
Basic parameters	360 (100)	180 (100)	180 (100)	0 (1)	≥.99
Medical accuracy	360 (100)	180 (100)	180 (100)	0 (1)	≥.99
Comprehensive feedback	282 (78)	137 (76)	145 (81)	1.05 (1)	.31
Autonomous conclusion	285 (79)	146 (81)	139 (77)	0.825 (1)	.36
Delayed feedback	200 (55)	157 (87)	43 (24)	146.2 (1)	<.001

[a]Simulations were given the correct or incorrect treatment algorithm inputs and evaluated on exhibition of characteristics against the 9-part metric. Successful exhibition of a parameter was calculated as a percent of all simulations in a subgroup. The table provides a summary of simulation outcomes.

[b]*P* values were calculated from χ^2 values

Figure 3. ChatGPT performance on simulation parameters by correct and incorrect inputs. Radar plot showing correct (blue) and incorrect (green) ChatGPT-4 performance on simulation parameters. Each point represents a simulation parameter. Lines indicate the percentage of simulations exhibiting that parameter, with the circumference of the shape indicating 100%. Simulations were given the correct or incorrect acute asthma exacerbation treatment algorithm inputs and evaluated by the 9-part metric developed from multimedia education and simulation principles. Successful exhibition of a parameter was calculated as a percent of all simulations in a subgroup and plotted on the radar plot. Correct and Incorrect subgroups performed similarly on all metrics except "delayed feedback."

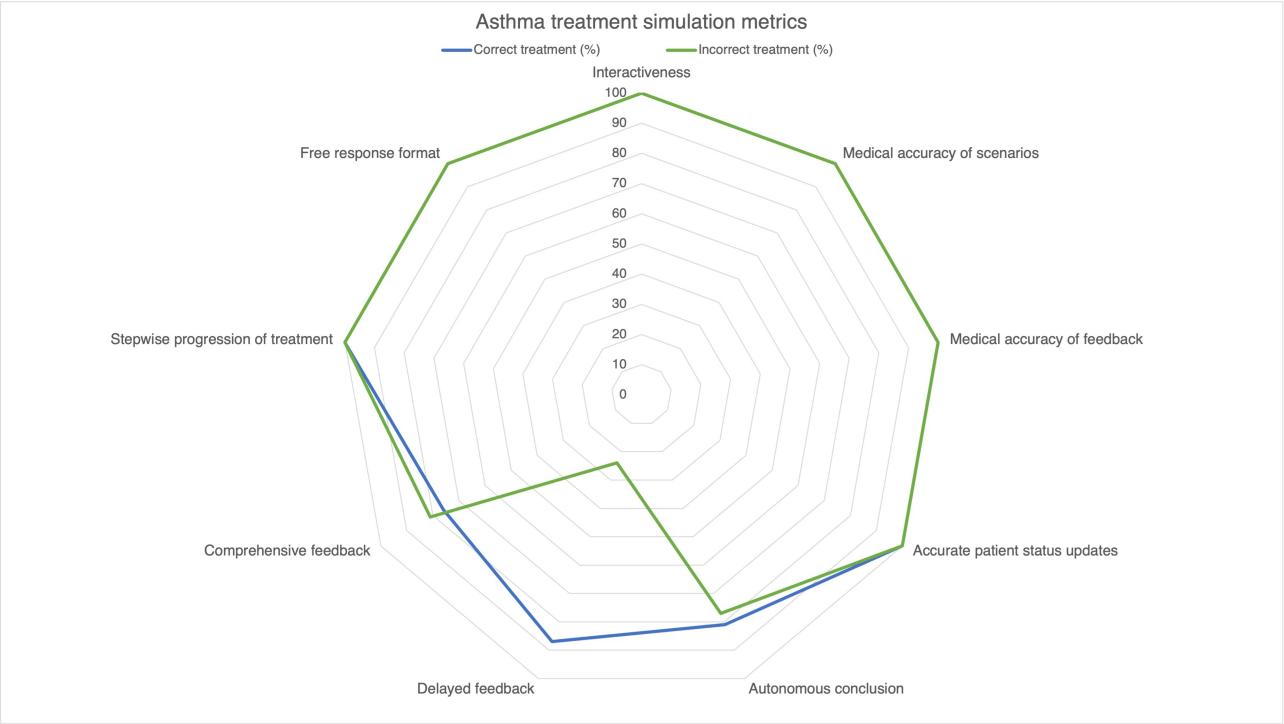

In 360 completed simulations, ChatGPT adhered to basic parameters of stepwise progression of treatment, appropriate patient status adjustments based on treatment inputs, and free response format in 100% (n=360) of simulations. It was also found to be 100% medically accurate in its vignettes, treatment updates, and feedback. There was no difference in medical accuracy between ChatGPT-4 outputs after correct versus incorrect tester treatment responses.

Results for advanced parameters were mixed. Among the correct treatment subgroup (n=180), 87% (157/180) of simulations delayed feedback until the end of the scenario. Only 24% (43/180) of the simulations with incorrect treatment (n=180) delayed feedback (*P*<.001). In all but 2 of these simulations with immediate (nondelayed) feedback, the feedback came after the administration of azithromycin. When combining correct and incorrect simulations (n=360), ChatGPT-4 demonstrated a 55% success rate in delaying feedback.

Simulations with correct treatments concluded the scenario autonomously 81% (146/180) of the time, while those with incorrect treatment concluded autonomously 77% (139/180; *P*=.36). Overall, 79% (285/360) of all simulations, regardless of treatment accuracy, concluded autonomously without tester input.

Correct simulations provided comprehensive feedback 76% (137/180) of the time while incorrect simulations provided comprehensive feedback 81% (145/180) of the time (*P*=.31).

XSL•FO
RenderX

Combined analysis showed ChatGPT-4 provided comprehensive feedback on 76% (137/180) of outputs.

Further analysis revealed that ChatGPT-4 was not more likely to conclude the simulation autonomously (P=.34) and provide comprehensive feedback (P=.27) when feedback was delayed compared to when feedback was not delayed (Table 2).

Table . Impact of delayed feedback on simulation autonomous conclusion and comprehensive feedback.

	Delayed feedback	Early feedback
Autonomous conclusion[a]	162	123
No autonomous conclusion	38	37
Comprehensive feedback[b]	161	121
No comprehensive feedback	39	39

[a]Number of all simulations regardless of subgroup that reached an autonomous conclusion versus those that did not based on delayed or early feedback timing.

[b]Number of all simulations regardless of subgroup that provided comprehensive feedback versus those that did not based on delayed or early feedback timing.

Discussion

Basic Parameters

ChatGPT-4 performed perfectly on basic simulation parameters, which is a requirement for it to be a viable educational tool. Its performance agrees with other research on prompt engineering showing proper prompt generation can produce high-quality results [34]. However, even given the use of proper prompt engineering techniques, perfect performance is surprising given ChatGPT's known inconsistencies [4,34,35]. This may reflect its algorithmic refinement and increased LLM training from widespread use. Given this known feature of LLMs, we should expect it to continue improving over time.

Medical Accuracy

ChatGPT was also medically accurate for both patient progression within the simulation and in its feedback; simulated patients had reasonable clinical presentations, responded appropriately to therapies, and feedback was in accordance with basic pharmacology, physiology, antibiotic stewardship, and clinical practice. This agrees with other findings on ChatGPT's clinical reasoning abilities, but the consistently excellent performance is again impressive. However, acute asthma exacerbation is not an exceedingly complex problem and has well-established management guidelines that likely exist in ChatGPT's training data, so its medical accuracy cannot be generalized to more complex clinical simulations like intensive care management [2,4]. Further research needs to be done into LLM performance differences between complex and simple simulations.

Advanced Parameters

ChatGPT's performance on feedback was variable. The observed significant difference in delaying feedback is noteworthy. Delayed feedback best mirrors real-life clinical scenarios where immediate feedback is not always available to redirect incorrect thought processes [36,37]. However, immediate feedback to incorrect answers mirrors the role a teacher might take with a student in simulation, where the teacher does not give a correct answer but instead explains why a learner's answer was incorrect in an attempt to give a better chance at the correct answer. It is also in-line with the expertise reversal effect that suggests novice learners will do better with more intense educational guidance, while advanced learners do better with less [38]. One possibility is that ChatGPT may be intentionally providing more structured guidance to learners who make errors, recognizing the need for additional support.

Delayed feedback best mirrors real-life clinical scenarios where immediate feedback is not always available to redirect incorrect thought processes. This may be particularly important for learners intent on practicing in smaller community settings where there are fewer opportunities for collaboration. However, this independence is a major developmental process in residency training, and thus, may be above the training level for which these simulations were targeted. Additionally, health care systems are interdisciplinary and collaborative. Expecting completely delayed feedback on patient care may be unrealistic considering shift hand-offs, pharmacists, nurses, medical record alerts, and other checks. Furthermore, the difference between delayed feedback for correct and incorrect treatments may offer different learning experiences for learners of different skills; it is unclear whether these different learning experiences would be inequitable or if they would offer new opportunities for learning (ie, learners who know the proper treatment are ready to practice independently, whereas learners still mastering treatment protocols are not forced to practice independently when they are not ready).

Another consideration is that different prompt engineering could resolve this variability. A prompt could include the instructions to "not give feedback until the end of the scenario, and instead of giving feedback on incorrect answers, make the patient more unstable as a sign that the choice was incorrect, then explain this incorrect choice in the summary feedback at the end of the simulation." Alternatively, because ChatGPT's algorithm learns from user inputs, learners running simulations in the same chat window could give ChatGPT feedback as they run simulations to hone in on their desired parameters, including desired feedback mechanisms. Future studies will need to evaluate differential prompt engineering, learner experiences, and learner preferences.

XSL•FO
RenderX

Of note, ChatGPT also gave hints to guide the tester after incorrect inputs. For example, ChatGPT could change from asking about the "next step in management" to asking about the "next step in controlling bronchoconstriction," thus nudging the tester toward administering a bronchodilator. This feature, as well as not delaying feedback after incorrect inputs, might be ChatGPT exhibiting flow, the educational and gaming principle that tasks should be made difficult enough to optimally challenge the learner, but not so difficult that the learner quits [39]. Again, further studies will need to evaluate learner experiences and preferences.

Comprehensive feedback was not significantly different between correct and incorrect simulations, with an overall rate of 78% (282/360). This high performance is encouraging, as comprehensive feedback offers more detailed learning opportunities and touchpoints for self-study. While this study did not explore the possibility of learners requesting expanded feedback, doing so could potentially increase the rate of comprehensive feedback to close to 100%, thereby enhancing the educational value of the simulations. ChatGPT concluded autonomously at similar rates to comprehensive feedback, which is also encouraging. Autonomous conclusion prevents learners from becoming "trapped" in the simulation, primed by the continuing simulation to believe there is more treatment needed even when the patient is stable and treatment should conclude. While autonomous conclusion at 100% would be best for efficiency and ease of use, forcing a learner to decide when to end the simulation can also mimic the testing style of oral examinations used by some medical specialties for board certification. This is a skill expected of advanced learners who have completed residency, but earlier learners could be easily prompted to end the simulation when they feel they have completed their treatments. Proper instructions and framing could address this problem from the prompt generator's perspective.

ChatGPT-4 was not significantly more likely to conclude simulations autonomously and provide comprehensive feedback when feedback was delayed. This suggests that ChatGPT-4 does not necessarily perform desirable actions in clusters, meaning that learners who make mistakes are not at risk of losing some of the simulation's desirable characteristics.

Future Work

This study focused on a single mistake in the treatment workflow, but future studies should look at performance with multiple mistakes. ChatGPT-4's ability to display desirable simulation parameters in the presence of multiple mistakes would be crucial for it to be a viable educational tool; real learners may make multiple mistakes of varying severity, and ChatGPT-4 would need to handle this. As is, the feasibility of using this tool given ChatGPT's varied performance on advanced parameters is still unclear. Medical educators may be hesitant to use a tool that is not completely standardized and risks inequitable learning outcomes, while others may welcome a tool that seemingly adapts to the learner's level. Quantified student learning gains after simulation use are thus needed for educators to make informed decisions in their classrooms and simulation centers and are the subject of ongoing work. However, with the aforementioned testing, we foresee the following potential implementation: educators would use the most advanced version of ChatGPT available and tailor the prompt provided here to their students' learning goals. Simulations would focus on simple diagnostic or therapeutic problems with straightforward algorithms. Simulations would be paired with robust pre- or postsimulation didactics to reinforce simulation concepts and not be the sole method of instruction. Ideally, simulations would be overseen by an experienced educator capable of monitoring simulations for accuracy.

Future work should focus on measuring student performance after using simulations to assess educational utility, trialing simulations on different LLMs to assess if the best LLM exists for this purpose, and trialing LLM performance on increasingly complex, nonalgorithmic simulations with multiple wrong answers. All of these are important for this tool's external validity.

Limitations

This study has several limitations which should be addressed. First, the simulations were limited to a single clinical scenario—acute asthma exacerbation—which may not be representative of ChatGPT-4's performance across a wider range of medical conditions. More complex scenarios reflecting the breadth of clinical practice and training level cannot be assumed to function as the simulations provided here. Our simulations were also run on only 1 prompt to limit confounders, but different prompts may produce varied results. Additionally, the study only compared correct treatment and 1 type of incorrect treatment, excluding a broader spectrum of clinical inaccuracies that may occur in real-world educational settings. Furthermore, simulations were run on 3 different ChatGPT-4 accounts and across different browsers, which, although necessary due to token limits, might introduce variability in the model's responses. Finally, the results of this study are contingent on the current application program interface settings established by ChatGPT and each new update of the LLM can potentially impact reproducibility. However, this does not detract from this study's conclusions given its intent to establish the methodology for evaluating simulations, rather than being a definitive recommendation for implementing LLM into preclinical medical education.

Conclusions

The 9-part model described here can be useful in evaluating ChatGPT simulations as a learning tool in accordance with established principles of medical simulation and multimedia design in educational technology. The use of this model can help standardize the evaluation of LLM simulation research. As an example use of this model, ChatGPT performed well in creating practice clinical scenarios that adhered to 3 simulation parameter categories. It performed excellently on 2 primary features, medical accuracy and basic simulation parameters, and reasonably well on advanced parameters. Variation in simulation characteristics based on the accuracy of learner inputs is a point of concern. This tool's impact on student learning is an important next step to explore, but these simulations

XSL•FO
RenderX

demonstrate ChatGPT's potential to be a reliable educational tool for the appropriate preclinical-to-clinical learning level.

Data Availability

All data generated or analyzed during this study are included in this published article and in Multimedia Appendix 1.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary files.
[XLSX File, 62 KB - formative_v9i1e66478_app1.xlsx]

References

1. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 8;9:e45312. [doi: 10.2196/45312] [Medline: 36753318]
2. Nakao T, Miki S, Nakamura Y, et al. Capability of GPT-4V(ision) in the Japanese National Medical Licensing Examination: evaluation study. JMIR Med Educ 2024 Mar 12;10:e54393. [doi: 10.2196/54393] [Medline: 38470459]
3. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digital Health 2023 Feb;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
4. Ali R, Tang OY, Connolly ID, et al. Performance of ChatGPT and GPT-4 on Neurosurgery written board examinations. Neurosurgery 2023 Dec 1;93(6):1353-1365. [doi: 10.1227/neu.0000000000002632] [Medline: 37581444]
5. Brin D, Sorin V, Vaid A, et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. Sci Rep 2023 Oct 1;13(1):16492. [doi: 10.1038/s41598-023-43436-9] [Medline: 37779171]
6. Mohammad B, Supti T, Alzubaidi M, et al. The pros and cons of using ChatGPT in medical education: a scoping review. Stud Health Technol Inform 2023 Jun 29;305:644-647. [doi: 10.3233/SHTI230580] [Medline: 37387114]
7. Boscardin CK, Gin B, Golde PB, Hauer KE. ChatGPT and generative artificial intelligence for medical education: potential impact and opportunity. Acad Med 2024 Jan 1;99(1):22-27. [doi: 10.1097/ACM.0000000000005439] [Medline: 37651677]
8. Khan RA, Jawaid M, Khan AR, Sajjad M. ChatGPT—reshaping medical education and clinical management. Pak J Med Sci 2023;39(2):605-607. [doi: 10.12669/pjms.39.2.7653] [Medline: 36950398]
9. Scherr R, Halaseh FF, Spina A, Andalib S, Rivera R. ChatGPT interactive medical simulations for early clinical education: case study. JMIR Med Educ 2023 Nov 10;9:e49877. [doi: 10.2196/49877] [Medline: 37948112]
10. Halaseh FF, Yang JS, Danza CN, Halaseh R, Spiegelman L. ChatGPT's role in improving education among patients seeking emergency medical treatment. West J Emerg Med 2024 Sep;25(5):845-855. [doi: 10.5811/westjem.18650] [Medline: 39319818]
11. Wu Y, Zheng Y, Feng B, Yang Y, Kang K, Zhao A. Embracing ChatGPT for medical education: exploring its impact on doctors and medical students. JMIR Med Educ 2024 Apr 10;10:e52483. [doi: 10.2196/52483] [Medline: 38598263]
12. Araujo SM, Cruz-Correia R. Incorporating ChatGPT in medical informatics education: mixed methods study on student perceptions and experiential integration proposals. JMIR Med Educ 2024 Mar 20;10:e51151. [doi: 10.2196/51151] [Medline: 38506920]
13. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023 Mar 19;11(6):887. [doi: 10.3390/healthcare11060887] [Medline: 36981544]
14. Alkhaaldi SMI, Kassab CH, Dimassi Z, et al. Medical student experiences and perceptions of chatgpt and artificial intelligence: cross-sectional study. JMIR Med Educ 2023 Dec 22;9:e51302. [doi: 10.2196/51302] [Medline: 38133911]
15. Skryd A, Lawrence K. ChatGPT as a tool for medical education and clinical decision-making on the wards: case study. JMIR Form Res 2024 May 8;8:e51346. [doi: 10.2196/51346] [Medline: 38717811]
16. Kim HW, Hong JW, Nam EJ, Kim KY, Kim JH, Kang JI. Medical students' perceived stress and perceptions regarding clinical clerkship during the COVID-19 pandemic. PLoS ONE 2022;17(10):e0277059. [doi: 10.1371/journal.pone.0277059] [Medline: 36315569]
17. Heston TF, Khun C. Prompt engineering in medical education. IME 2023;2(3):198-205. [doi: 10.3390/ime2030019]
18. Musallam E, Alhaj Ali A, Alkhafaji M. OpenAI's ChatGPT clinical simulation: an innovative teaching strategy for clinical nursing education. Nurse Educ 2024;49(6):E361-E362. [doi: 10.1097/NNE.0000000000001657] [Medline: 38728116]
19. Lucas HC, Upperman JS, Robinson JR. A systematic review of large language models and their implications in medical education. Med Educ 2024 Nov;58(11):1276-1285. [doi: 10.1111/medu.15402] [Medline: 38639098]

XSL•FO
RenderX

20. Jeyaraman M, K SP, Jeyaraman N, Nallakumarasamy A, Yadav S, Bondili SK. ChatGPT in medical education and research: a boon or a bane? Cureus 2023 Aug;15(8):e44316. [doi: 10.7759/cureus.44316] [Medline: 37779749]
21. van de Ridder JMM, Shoja MM, Rajput V. Finding the place of ChatGPT in medical education. Acad Med 2023 Aug 1;98(8):867. [doi: 10.1097/ACM.0000000000005254] [Medline: 37162206]
22. Clark RC. Multimedia learning in e-courses. In: Mayer RE, Fiorella L, editors. The Cambridge Handbook of Multimedia Learning, 2nd edition: Cambridge University Press; 2014:842-882. [doi: 10.1017/CBO9781139547369.040]
23. Wylie R, Chi MTH. The self-explanation principle in multimedia learning. In: Mayer RE, Fiorella L, editors. The Cambridge Handbook of Multimedia Learning, 2nd edition: Cambridge University Press; 2014:413-432. [doi: 10.1017/CBO9781139547369.021]
24. Luetner D, Schmeck A. The generative activity principle in multimedia learning. In: Mayer RE, Fiorella L, editors. The Cambridge Handbook of Multimedia Learning, 2nd edition: Cambridge University Press; 2014:433-448. [doi: 10.1017/CBO9781139547369.022]
25. Mayer RE. Cognitive theory of multimedia learning. In: Mayer RE, Fiorella L, editors. The Cambridge Handbook of Multimedia Learning, 2nd edition: Cambridge University Press; 2014:43-71. [doi: 10.1017/CBO9781139547369.005]
26. Low R, Sweller J. The modality principle in multimedia learning. In: Mayer RE, Fiorella L, editors. The Cambridge Handbook of Multimedia Learning, 2nd edition: Cambridge University Press; 2014:227-246. [doi: 10.1017/CBO9781139547369.012]
27. Plass JL, Schwartz RN. Multimedia learning with simulations. In: Mayer RE, Fiorella L, editors. The Cambridge Handbook of Multimedia Learning, 2nd edition: Cambridge University Press; 2014:729-761. [doi: 10.1017/CBO9781139547369.036]
28. Mayer RE, Fiorella L. Principles for reducing extraneous processing in multimedia learning. In: The Cambridge Handbook of Multimedia Learning, 2nd edition: Cambridge University Press; 2014:279-315. [doi: 10.1017/CBO9781139547369.015]
29. Johnson CI, Priest HA. The feedback principle in multimedia learning. In: Mayer RE, Fiorella L, editors. The Cambridge Handbook of Multimedia Learning, 2nd edition: Cambridge University Press; 2014:449-463. [doi: 10.1017/CBO9781139547369.023]
30. Stefanidis D, Cook D, Kalantar-Motamedi SM, et al. Society for simulation in healthcare guidelines for simulation training. Simul Healthc 2024 Jan 1;19(1S):S4-S22. [doi: 10.1097/SIH.0000000000000776] [Medline: 38240614]
31. van Gog T. The signaling (or cueing) principle in multimedia learning. In: Mayer RE, Fiorella L, editors. The Cambridge Handbook of Multimedia Learning, 2nd edition: Cambridge University Press; 2014:263-278. [doi: 10.1017/CBO9781139547369.014]
32. Feldon DF, Jeong S, Clark RE. Fifteen common but questionable principles of multimedia learning. In: Mayer RE, Fiorella L, editors. The Cambridge Handbook of Multimedia Learning, 3rd edition: Cambridge University Press; 2021:25-40. [doi: 10.1017/9781108894333.005]
33. Dabbs W, Bradley MH, Chamberlin SM. Acute asthma exacerbations: management strategies. Am Fam Physician 2024 Jan;109(1):43-50. [Medline: 38227870]
34. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023 Aug;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
35. Alkaissi H, McFarlane SI. Artificial hallucinations in ChatGPT: implications in scientific writing. Cureus 2023 Feb;15(2):e35179. [doi: 10.7759/cureus.35179] [Medline: 36811129]
36. Hattie J, Timperley H. The power of feedback. Rev Educ Res 2007 Mar;77(1):81-112. [doi: 10.3102/003465430298487]
37. Chamberland M, Setrakian J, St-Onge C, Bergeron L, Mamede S, Schmidt HG. Does providing the correct diagnosis as feedback after self-explanation improve medical students diagnostic performance? BMC Med Educ 2019 Jun 11;19(1):194. [doi: 10.1186/s12909-019-1638-3] [Medline: 31185971]
38. Kalyuga S. The expertise reversal principle in multimedia learning. In: Mayer RE, Fiorella L, editors. The Cambridge Handbook of Multimedia Learning, 2nd edition: Cambridge University Press; 2014:576-597. [doi: 10.1017/CBO9781139547369.028]
39. Csikszentmihalyi M, Abuhamdeh S, Nakamura J. In: Csikszentmihalyi M, editor. Flow and the Foundations of Positive Psychology: The Collected Works of Mihaly Csikszentmihaly: Springer; 2014:227-238. [doi: 10.1007/978-94-017-9088-8_15]

Abbreviations

AI: artificial intelligence
LLM: large language model
RISE: role, input, steps, expectations

XSL•FO
RenderX

Edited by A Mavragani; submitted 14.09.24; peer-reviewed by C Zickler, M Barakat, N Chien, S Patel, W Li; revised version received 30.01.25; accepted 31.01.25; published 27.02.25.

Please cite as:
Scherr R, Spina A, Dao A, Andalib S, Halaseh FF, Blair S, Wiechmann W, Rivera R
Novel Evaluation Metric and Quantified Performance of ChatGPT-4 Patient Management Simulations for Early Clinical Education: Experimental Study
JMIR Form Res 2025;9:e66478
URL: https://formative.jmir.org/2025/1/e66478
doi:10.2196/66478

© Riley Scherr, Aidin Spina, Allen Dao, Saman Andalib, Faris F Halaseh, Sarah Blair, Warren Wiechmann, Ronald Rivera. Originally published in JMIR Formative Research (https://formative.jmir.org), 27.2.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Capturing Everyday Parental Feeding Practices and Eating Behaviors of 3- to 5-Year-Old Children With Avid Eating Behavior: Ecological Momentary Assessment Feasibility and Acceptability Study

Abigail Pickard[1,2], PhD; Katie Edwards[1,3], PhD; Claire Farrow[1], PhD; Emma Haycraft[4], PhD; Jacqueline Blissett[1], PhD

[1]School of Psychology and Institute of Health and Neurodevelopment, Aston University, Birmingham, United Kingdom

[2]Department of Clinical Psychology, School of Health in Social Science, University of Edinburgh, Wilkie Building, Edinburgh, United Kingdom

[3]School of Psychology, University of Birmingham, Birmingham, United Kingdom

[4]School of Sport, Exercise and Health Sciences, Loughborough University, Loughborough, United Kingdom

Corresponding Author:
Abigail Pickard, PhD
School of Psychology and Institute of Health and Neurodevelopment, Aston University, Birmingham, United Kingdom

Abstract

Background: The wide use of smartphones offers large-scale opportunities for real-time data collection methods such as ecological momentary assessment (EMA) to assess how fluctuations in contextual and psychosocial factors influence parents' feeding practices and feeding goals, particularly when feeding children with high food approaches.

Objective: The main objectives of this study were to (1) assess parents/caregivers' compliance with EMA procedures administered through a smartphone app and (2) estimate the criterion validity of the EMA to capture children's eating occasions and parents' feeding practices. Participant adherence, technological challenges, and data quality were used to provide an overview of the real-time dynamics of parental mood, feeding goals, and contextual factors during eating occasions.

Methods: Parents in the United Kingdom with a child aged 3 to 5 years who exhibit avid eating behavior were invited to participate in a 10-day EMA study using a smartphone app. Of the 312 invited participants, 122 (39%) parents initiated the EMA study, of which 118 (96.7%) completed the full EMA period and the follow-up feasibility and acceptability survey.

Results: Of those parents who completed the EMA study, 104 (87.4%) parents provided at least 7 "full" days of data (2 signal surveys and 1 event survey), despite 51 parents (43.2%) experiencing technical difficulties. The parents received notifications for morning surveys (69.9% response rate), 3 daily mood surveys (78.7% response rate), and an end-of-day survey (84.6% response rate) on each of the 10 days. Over the EMA period, a total of 2524 child eating/food request surveys were self-initiated by the participants on their smartphones, an average of 2.1 times per day per parent (SD 0.18; min=1.7, max=2.3). The majority of parents felt that the surveys made them more aware of their feelings (105/118, 89%) and activities (93/118, 79%). The frequency of daily food requests estimated by parents at baseline was significantly correlated with the frequency of food requests reported daily during the EMA period (r=0.483, P<.001). However, the number of daily food requests per day estimated at baseline (mean 4.5, SD 1.5) was significantly higher than the number of food requests reported per day during the EMA period (mean 3.7, SD 1.1), (t_{116}=18.8, P<.001).

Conclusions: This paper demonstrates the feasibility of employing EMA to investigate the intricate interplay between parental mood, feeding goals, contextual factors, and feeding practices with children exhibiting an avid eating behavior profile. However, the use of EMA needs to be carefully developed and tested with parents' involvement to ensure successful data collection.

International Registered Report Identifier (IRRID): RR2-10.2196/55193

(JMIR Form Res 2025;9:e66807) doi:10.2196/66807

KEYWORDS

pediatric; paediatric; child; child eating; parent feeding; parent; ecological momentary assessment; mHealth; mobile health; mobile app; application; smartphone; digital; digital health; digital technology; digital intervention

XSL•FO
RenderX

Introduction

Children with an avid eating behavior profile pose unique challenges for parents, as this group of children has been shown to make greater requests for food, often in response to emotions and food cues in their environment rather than due to internal hunger signals [1,2]. Avid eating behavior in children is characterized by an intense and persistent interest in food, often leading to distinct challenges in managing feeding interactions [1,2]. Understanding the real-time dynamics of these interactions is crucial for developing targeted interventions aimed at promoting healthy eating behaviors in children.

However, despite evidence that both parental feeding practices and children's eating behaviors are dynamic, much research has investigated the association between the 2 constructs cross-sectionally, using questionnaires, interviews, and one-off mealtime observations [3]. Such research only allows between-subject investigation, neglecting the within-subject variation in both feeding practices and eating behavior. These limitations have spurred more recent research using ecological momentary assessment (EMA; also known as experience sampling methods).

EMA is a form of mobile health technology that allows the collection of comprehensive data on daily life, as it is directly experienced by an individual, by asking them to provide responses about their emotions, context, and behavior at several random time points each day over several days [4]. Commonly, EMA employs both signal-contingent surveys, which are scheduled by the researcher to be push-notified at specific times, and event-contingent surveys, which are to be initiated by the participant following a specific event or action [5]. Due to advancements in technology and the widespread use of smartphones, EMA methods have become more popular in eating behavior research [4,6]. Berge et al [7] conducted an EMA study examining feeding practices in a sample of parents (N=150) of a 5 - 7-year-old child with at least one other sibling between 2 and 12 years old. Their EMA study had very high compliance; 100% of participants (N=118) completed all 8 days of EMA data collection and met the minimum requirement per day (ie, 2 signal contingent, 1 event contingent, and 1 end-of-day contingent), with an average of 7.4 surveys per day [7].

A more recent study conducted by Loth et al [8] used EMA among a sample of 120 young adults with at least 1 child aged 2 - 5 years. Of the 118 participants who registered for their EMA study, 116 (98.3%) participants completed at least one eating occasion survey, 100 participants completed one "full" day (84.7%), and 80 (67.8%) participants completed "full" days for the complete 10-day EMA period (a "full" day was defined as providing data for 2 eating occasions, 2 random prompts, and 1 end of day prompt). Finally, a study examining parents' feeding practices as a predictor of children's physical activity and fruit and vegetable intake used an EMA design with 22 mothers, fathers, and 8 - 12-year-old-child triads completing a 7-day assessment period [9]. The response rates for the study period were high with 85%, 80%, and 85.9% responses to the survey notifications completed by the mothers, fathers, and children, respectively.

While these studies have provided important evidence for the use of EMA to investigate parental feeding interactions with children, there remains caution and concern within the field about the feasibility of using such methods with samples in which feeding and eating interactions may be more challenging. For example, it is unknown whether such methods are adequate to capture real-time information for parents of children with avid eating tendencies, who have previously been shown to have a higher risk of food insecurity [1] and where children are more likely to show an increased frequency of eating and food requests [2]. Therefore, this study assessed the feasibility, acceptability, and criterion validity of using EMA to explore the participation and everyday experiences of parenting children with avid eating behavior aged 3 to 5 years old across a 10-day sampling period.

Assessing the feasibility of such methods is not only important to guide future EMA research but also to ensure that the findings of this study are reliable. The study also aimed to determine the usefulness of using a commercially licensed software application on Android and iOS devices as a tool for EMA.

Methods

Overview

This EMA study was part of a larger program of research, the APPETItE project, which examines feeding and eating in preschool children with avid eating behavior to inform future intervention design and efficacy. An avid eating behavior profile in preschool children was identified and defined in an earlier APPETItE study using latent profile analysis (LPA) [1]. The LPA used the 8 subscales of the Children's Eating Behavior Questionnaire (CEBQ) [10] as index variables to create holistic yet distinct eating profiles for UK children aged 3 to 5 years.

Ethical Considerations

Recruitment, enrolment, and data collection for the EMA study took place between October 2023 and April 2024. Ethical approval was granted by the Aston University Health and Life Sciences Research Ethics Committee (HLS21079). All participants provided informed consent and were made aware that they had the right to opt out of the study at any time. For a comprehensive overview of the EMA study protocol and the measures included in each data collection wave, refer to Edwards et al [11]. The findings of the full study will be published at a later date.

Participants

Only parents of children aged 3 to 5 years who were assigned to the avid eating profile based on the screening questionnaire were eligible to participate in this EMA study. Eligibility criteria also included English-speaking primary caregivers from the United Kingdom who were responsible for feeding their child for more than half the time when their child was at home. Parents whose child was autistic, or had severe learning disabilities, or a chronic illness that directly influenced their dietary requirements and eating habits were not eligible to participate. A total of 312 parents met the eligibility criteria for the study and were invited to participate via email.

XSL•FO
RenderX

Recruitment

From the original LPA defining the avid eating profile, 101 parents with a child in the eligible age range (3 - 5 y) and assigned to the avid eating profile were invited to participate. The profile solution identified in the LPA study [1] was used to classify additional children whose parents expressed interest in the EMA study and submitted CEBQ [10] data through the recruitment platform Prolific. Using this profiling method, 168 children were identified as belonging to the avid eating profile. A further 43 parents who had completed the CEBQ screening questionnaire for a separate lab-based study as part of the APPETItE project were also identified as having eligible children and were invited to take part in the EMA study. If the parents completed the initial interest survey, they were then contacted via email with further instructions and a demo video [12] explaining how to complete the EMA study. If parents experienced any technical difficulties, had any questions, or wanted to go through the study procedures, they were able to contact the research team by email.

Data Collection Procedures

Participation in this study was remote; surveys were administered through a smartphone app, ExpiWell, which was downloaded directly to parents' smartphones. ExpiWell is a cloud-based platform designed for researchers to conduct experience sampling methods and EMA studies. It enables researchers to create, schedule, and distribute surveys to participants via free mobile applications available on iOS and Android devices. Parents who did not own a smartphone were informed that they could request one from the research team to use for the study period, but all parents used their own devices.

Parents completed a baseline questionnaire, 10 days of EMA, and an end-of-study questionnaire. The 10-day EMA period included both signal surveys (push notifications sent to the participant's smartphone) and event surveys (initiated by the participant following an eating occasion or request for food by the child). Parents were notified to complete 1 start-of-day survey, 3 signal-contingent examining mood and emotions, 1 end-of-day survey, and event surveys activated by the parent when their child ate or asked for food. Participants were able to schedule the morning and evening surveys to suit their typical routine. The full protocol and procedure are reported elsewhere [11].

Parents received a £100 (approximately US $126) shopping voucher if at least 8 "full" days of surveys were complete. Based on previous research, the criteria for one full day of EMA data included the completion of 2 signal-contingent surveys (morning, mood, or end-of-day) and one event-contingent (food) survey [8]. All parents who completed 10 days of EMA were entered into a prize draw to win an additional £100 (approximately US $126) shopping voucher. If participants withdrew prematurely from this study, their time was reimbursed on a pro-rated basis of £10 (approximately US $12.60) per complete day.

Measures

To evaluate the feasibility and criterion validity of using EMA to capture the eating and feeding experiences of parents with a child displaying avid eating, several questions were incorporated into the survey period. The baseline survey asked parents to report on average how many times their child requested food, and ate food, both on weekdays and weekends. In the end-of-day surveys, parents were also asked to report how many times the child had eaten or requested food on that specific day. These data were collected to determine how many of the average eating and food request occasions were being reported in the EMA period.

Feedback Questionnaire

A self-constructed feedback questionnaire, based on the items used by de Vries et al [13] and "The Open Handbook of Experience Sampling Methodology" [14], was used to gather participants' perspectives on the EMA procedure. Participants responded to 16 items about their experience of completing the study, the influence the EMA study had on their daily lives, the awareness the EMA study gave them of their actions and feelings, and the burden of completing the EMA study. Response options were based on a 5-point scale where 1="not at all," 2="a little," 3="moderately," 4="quite a bit," and 5="extremely." The final item asked about whether the participant had experienced any technical problems during the EMA period with response options of 1="yes," and 0="no." The full study, including a list of all the items used, is available on the Open Science Framework [15].

Data Analysis

Descriptives of the survey completion rates, forgotten reporting, withdrawal rates, nonresponse rates, survey duration, and any reported technical issues were examined to determine EMA feasibility. Criterion validity of EMA-reported frequency of eating occasions and food requests was examined by comparison with the initially reported frequencies and end-of-day frequencies using paired *t* tests and Pearson correlations. Study acceptability was assessed with descriptive statistics of the feedback responses at the end of the study.

Results

Demographics

Parents had a mean age of 34.97 years (SD 5.3; min=23.6, max=55.3) and children had a mean age of 52.57 months (SD 10.5; min=36.8, max=71.4). An overview of the parents' demographic background and child sex is provided in Table 1.

XSL•FO
RenderX

Table . Demographic details of the sample (N=118).

Demographics	Values, n (%)
Child sex	
Male	55 (46.6)
Female	63 (53.4)
Parent sex	
Male	26 (22)
Female	92 (78)
Parent ethnicity	
Asian	13 (11)
Black	10 (8.5)
White	93 (78.8)
Other	2 (1.7)
Education	
Degree	70 (59.3)
No degree	48 (40.7)
Parent or family structure	
Single parent	12 (10.2)
Dual parent	96 (81.4)
Cohabiting (not in relationship)	6 (5.1)
Extended family	4 (3.4)
Working status	
Unemployed	25 (21.2)
Working part-time (between 8 and 29 hours per week)	36 (30.5)
Working full-time (30 hours or more per week)	57 (48.3)
Adequacy of income	
Living comfortably	54 (45.8)
Managing	47 (39.8)
Finding it difficult	17 (14.4)
Household food security	
High or marginal food security	78 (66.1)
Low food security	20 (16.9)
Very low food security	20 (16.9)

EMA Feasibility

Of the 312 parents invited to participate, 122 (39.1%) parents downloaded the mobile application and completed the initial survey. Of those 122 participants, 118 (96.7%) participants completed the EMA period. Four parents withdrew from the EMA study due to either technical difficulties or time constraints. Of the remaining participants, 104 (87.4%) participants completed at least 7 days of the EMA period with an average of 58 completed surveys for the full 10 days (SD 18; min=9, max=103). Figure 1 provides an overview of the eligible sample sizes as well as completion and attrition rates at each phase of the EMA study.

XSL•FO
RenderX

Figure 1. Recruitment and retention figures for each EMA phase. EMA: ecological momentary assessment.

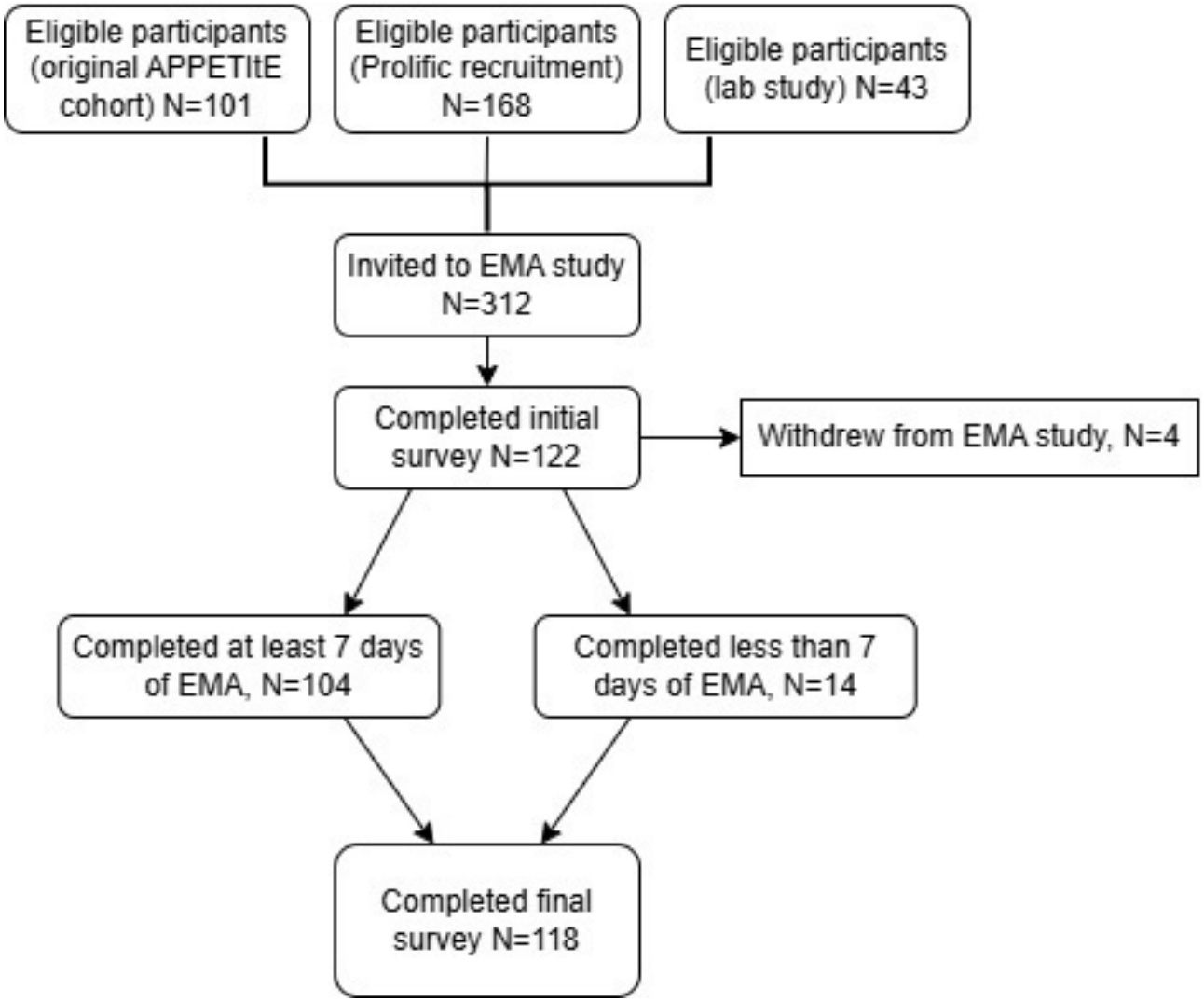

Over the 10-day EMA period, a total of 2524 child eating/food request surveys were self-initiated by the 118 participants on their smartphones, with 30 surveys partially completed. Parents self-initiated the request for food and eating surveys on average 2.1 times per day (SD 0.18; min=1.7, max=2.3). Of the 2524 eating and food request surveys completed at the moment, 2353 reported on an event in which the child ate food, and 171 reported on an instance in which the child requested food but did not eat.

At the end of all signal-contingent surveys and the food survey, parents were asked whether they had forgotten to report an eating occasion or a request for food from the child. On 870 occasions a parent retrospectively reported on an eating/food request that they had forgotten to report (Table 2 details the number of eating events that were reported on when the parent was reminded at the following survey).

XSL•FO
RenderX

Table . Number of retrospective eating occasions and food request reports.

Survey	Number of responses
Morning surveys	
Forgotten eating event	111
Forgotten food request event	4
Mood surveys	
Forgotten eating event	413
Forgotten food request event	28
End-of-day surveys	
Forgotten eating event	137
Forgotten food request event	13
Food surveys	
Forgotten eating event	141
Forgotten food request event	23
Total number of forgotten events reported	870

Response rates were calculated for the signal-contingent surveys, which included 1 survey in the morning, 3 surveys during the day, and 1 survey in the evening. From the first notification, parents had 60 minutes to complete the surveys before the link expired. A total of 1180 push notifications were received by all the participants (10 per participant) to complete the morning surveys, 825 morning surveys were initiated by participants following the notification, with 18 incomplete responses (response rate=69.9%). The parents also received notifications for 3 signal-contingent surveys (mood surveys) at semirandom times within one of three 120-minute blocks. A possible 3540 signal-contingent mood surveys were available to the 118 participants throughout the 10-day duration (30 per participant) with a total completion of 2778 surveys and 8 incomplete responses (78.7% response rate). The signal contingent end-of-day survey was sent to each participant every evening with a potential total of 1180 responses, 998 end-of-day surveys were initiated when prompted with 4 incomplete responses (84.6% response rate).

To test whether more participants responded to the 5 prompted surveys at the beginning (eg, 2-4 d) compared with the end (eg, 8-10 d) a paired *t* test was conducted on the daily compliance rates. Day 1 was treated as a practice day due to significantly lower compliance rates as participants became accustomed to the EMA protocol (Figure 2). There was no significant difference in compliance rate for the start of the study period compared with the end of the study period (t_{118}=0.127, P=.90).

XSL•FO
RenderX

Figure 2. Response rates for signal contingent surveys over EMA Period. EMA: ecological momentary assessment.

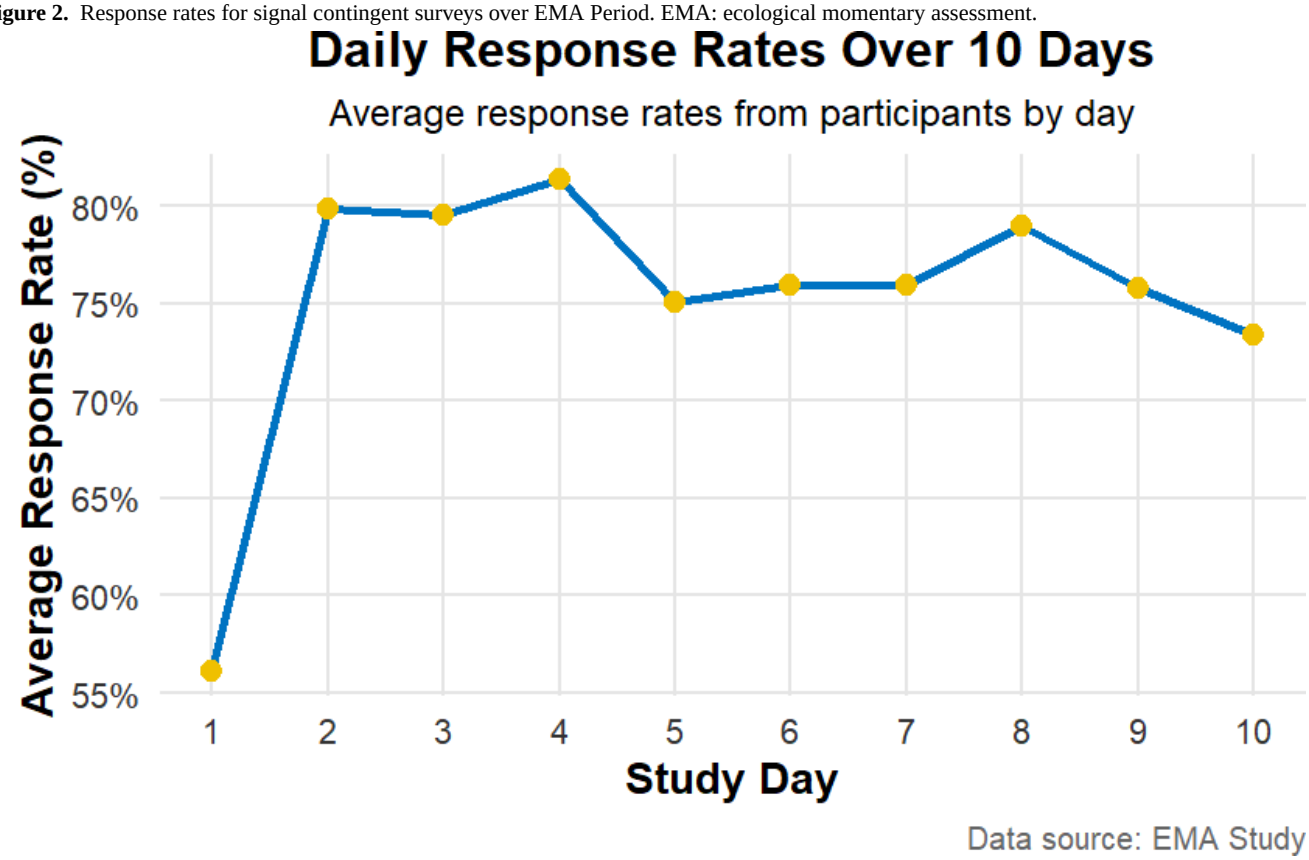

A 1-way ANOVA was conducted to determine whether the working status of the parent was associated with the number of surveys completed. There was a significant difference in the number of reported eating and food request occasions ($F_{2,115}$=12.41, P=.011), with parents working part-time reporting significantly more food occasions (mean=2.5 food surveys per day) than parents working full-time (mean=1.75 food surveys per day) but not significantly more than unemployed parents (mean=2.13). There were no significant differences in survey completion for all signal-contingent surveys between parents of differing working statuses.

Table 3 displays the descriptives of survey completion duration. The average duration to complete the morning surveys was 2.76 minutes (SD 6.83 minutes). The average time taken for the signaled mood surveys was 2.09 minutes (SD 6.16 minutes). The average duration to complete the food surveys was 3.13 minutes (SD 7.87 minutes) and 2524 food surveys were self-initiated and completed.

Table . Average survey completion duration by survey type.

	Duration (MM:SS)		
	Mean (SD)	Minimum	Maximum
Signaled surveys			
Morning surveys	02:29:00 (04:00:00)	00:38:00	58:32:00
Mood surveys	01:52:00 (02:21:00)	00:27:00	39:17:00
End-of-day surveys	02:20:00 (02:15:00)	00:47:00	31:44:00
Event surveys			
Food surveys	02:47:00 (02:29:00)	00:48:00	52:47:00

Almost half of the participants experienced technical issues with the application (51/118, 43.2%). Pearson chi-square test with Yates' continuity correction revealed that Android users were more likely to experience technical issues (30.16% of iOS users experienced technical issues compared with 62.26% of Android users).

Over half of the participants (71/118, 60%) reported that they did not make any mistakes when completing the surveys.

Criterion Validity

To determine the criterion validity of using EMA to accurately capture eating occasions and food requests of parents from children with avid eating, several additional measures of the frequency of children's eating and requests for food were taken.

Parents were asked at baseline to estimate how frequently their child ate on a typical day, during the EMA period they were also asked to report how many times their child had eaten that

XSL•FO
RenderX

day in the evening survey. A paired-samples *t* test and paired-samples correlations were then conducted to determine whether the scores were representative of what was reported by parents. Although the parents' baseline reports of eating frequency were correlated with daily reported eating frequency (r=0.365, P<.001), eating frequency was significantly higher in the baseline reports compared with the daily reports (t_{116}=5.78, P<.001). Parents reported greater eating frequency in the baseline survey (mean 4.5, SD 1.5) compared with the frequency that they reported in the end-of-day surveys (mean 3.7, SD 1.1).

Pairwise *t* testing showed that the approximate frequency of food requests per day reported at baseline was significantly correlated with the frequency of food requests reported daily during the EMA period (r=0.483, P<.001). However, pairwise *t* testing showed that the number of food requests per day at baseline was significantly higher (mean 5.6, SD 2.4) than the number of food requests reported per day during the EMA period (mean 1.9, SD 1.1; t_{116}=18.8; and P<.001). The number of daily food requests and eating occasions reported by parents in the baseline survey was not significantly correlated with the number of daily food requests and eating occasions reported by parents in the EMA period (r=0.121, P=.097). Furthermore, the mean number of daily eating/request events at baseline (mean 10.1, SD 3.5) was significantly higher than the number of daily eating/request events during the EMA period (mean 2.04, SD 1.2; t_{116}=24.9, and P=.001). The frequency of the daily eating/request events reported in the end-of-day survey was significantly correlated with the number of reports for in-the-moment eating/request events (r=0.244, P=.004). However, reports of average eating and food requests at the end of the day (mean 5.6, SD 1.8) were significantly higher than the average frequency of in-the-moment reporting of eating/food requests (mean 2.06, SD 1.2; t_{116}=20.2; and P<.001).

Study Acceptability

To determine whether the questions included in the EMA study were acceptable for measuring eating and feeding interactions with children with avid eating behavior, the parents were asked to rate the degree to which the questions included in the survey accurately reflected their experiences. The majority of the participants felt that the EMA surveys and questions accurately reflected their experience of a parent feeding a 3 - 5-year-old child with avid eating behavior, 20% (n=24) said extremely well, 44% (n=52) said quite a bit, and 30% (n=35) said moderately well.

Figure 3 presents participants' feedback on the EMA period. Nearly all participants reported that they were not burdened by the sound of the study notifications (116/118, 98%). Although parents were asked to keep their devices on loud, it was at their discretion whether they muted the notifications. Over half of the participants were not burdened by the length of the surveys (68/118, 58%). Over half of the participants considered the number of notifications at least a "little" burdensome (73/118, 62%), but 39% (n=46) did not consider the frequency of notifications a burden. Over half of parents (64/118, 54%) found that the surveys hindered their daily interactions a little to quite a bit. However, (76/118, 64%) reported that the surveys did not influence their activities at all and (51/118, 43%) stated that the surveys did not influence their mood.

Figure 3. Proportions of responses for items measuring study acceptability.

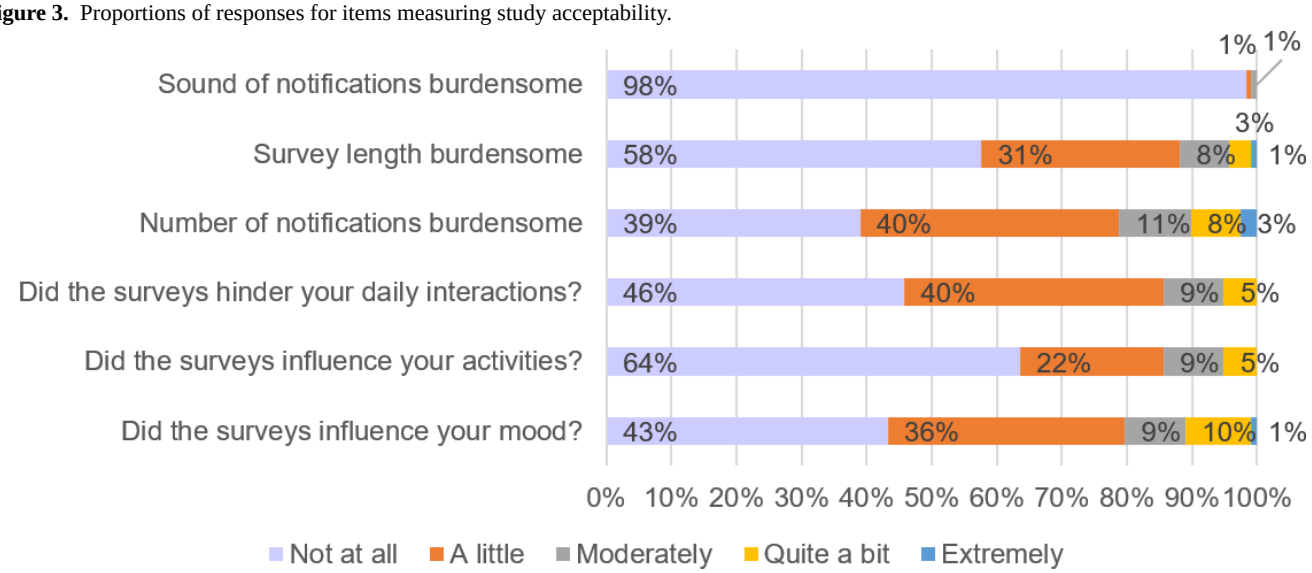

The majority of parents (105/118, 89%) reported that filling in the surveys made them more aware of their feelings and this was experienced as pleasant by 55% (58/105), neutral by 37% (39/105), and unpleasant by 8% (8/105) of parents. A total of 79% of parents (93/118) felt they were made more aware of their activities through completing the EMA study, of which 58 parents experienced this as pleasant, 34 parents experienced this as neutral, and 1 parent experienced this as unpleasant.

Parents were also asked in the signal-contingent end-of-day survey to rate how difficult it was to fill out the surveys that day, with 1="not at all" and 5="extremely". After the scores had been aggregated by participants the mean level of difficulty was 2 (equivalent to "a little") (SD 0.8; min=1, max=5).

XSL•FO
RenderX

Discussion

Principal Findings

This study aimed to assess the feasibility, acceptability, and criterion validity of using EMA to investigate parental feeding practices and their dynamic interactions with children exhibiting avid eating behaviors. The results demonstrate that EMA is a feasible and acceptable method for capturing real-time data on child eating occasions and food requests, but it also highlights some discrepancies in the data collected and usability issues specific to parents with young children.

Feasibility of, and Compliance With, EMA Methodology

Despite the monetary incentive for the study, the recruitment rate to the study was lower than expected with only 122 (39.1%) of contacted parents initiating the study. However, the study found a high completion rate of 96.7% (n=118) for the EMA period, suggesting that the method was generally feasible for the participating parents. Most parents (104/118, 87.4%) completed at least 7 days of the 10-day EMA period, with an average of 58 surveys completed per participant over the 10 days, which is above the minimum requirement of 2 signal-contingent surveys and 1 event-contingent survey per day. The overall response rates for the signal-contingent surveys were relatively high, ranging from 69.9% to 84.6% depending on the type of survey. These findings suggest that it was generally feasible for the participants to engage with the EMA protocol, despite the demands of responding to notifications to complete signal-contingent surveys as well as self-initiating surveys multiple times per day. However, working status played a role in the frequency of reported eating and food requests, with part-time working parents reporting more food occasions than full-time working parents. This finding suggests that the employment status of parents may influence the ability to engage with EMA protocols, which could have implications for future studies in families.

The inclusion of pro-rated remuneration appeared to be valuable for retaining participants and appropriately rewarding ongoing completion of study components. A small percentage of parents withdrew from the study due to technical difficulties or time constraints, and nearly half of the participants (51/118, 43.2%) experienced technical issues with the application, which could have affected data quality and study adherence. The provision of an email address through which participants could access rapid technical support from our study team ensured the retention of most participants in the study despite these challenges. Such support must be built into any future studies using EMA methods to ensure participant retention. Careful testing and consideration of the most appropriate software are crucial in future EMA research. The feedback also indicated that while the majority of parents did not find the sound of notifications burdensome, some participants perceived the frequency and timing of surveys as a hindrance to daily activities. These findings highlight the importance of carefully considering the design and user experience of the EMA app to minimize the burden on participants and reduce barriers to accurate data collection.

Criterion Validity of EMA for Capturing Feeding Interactions

While the EMA method showed promise in capturing in-the-moment data on eating occasions and food requests, discrepancies were found between the EMA reports and baseline recall data, suggesting that EMA should be validated with observational studies, given the challenges with any self-report methods. There were notable discrepancies between baseline reports of perceived frequency and EMA data of reported eating episodes: parents reported significantly fewer eating occasions and food requests during the EMA period compared with their baseline estimates. This pattern was consistent across different measures of eating frequency and food requests, suggesting a possible overestimation in baseline self-reports, or, perhaps more likely, underreporting of events during the EMA period. The finding that participants retrospectively reported missed eating or food request occasions in subsequent surveys supports the latter interpretation, suggesting some level of underreporting in the initial EMA responses.

There were 870 instances where parents, when prompted, reported missed eating or food request events that they had forgotten to log in the moment, highlighting a significant gap in real-time reporting accuracy. This retrospective reporting underscores the challenges inherent in capturing accurate eating behaviors in real-time, particularly with parents of young children. These results are consistent with previous studies indicating that retrospective prompts can improve data accuracy in EMAs [16]. The high incidence of missed reports emphasizes the importance of considering both real-time and reflective reporting methods in future research to enhance the validity and comprehensiveness of monitoring children's eating behavior [16]. Thus, whilst EMA may provide a good assessment of some eating interactions demonstrated by the wealth of data we collected; it is unlikely to capture all such interactions in the period of study.

The observed discrepancies between baseline reports and EMA data reflect the challenges associated with self-reporting in real-time contexts, particularly for parents of young children who display avid eating. If parents are unable to complete surveys at the moment due to competing demands or distractions, the more chaotic or stressful feeding interactions that we aim to capture are likely to be missed. Additionally, the high frequency of child eating behaviors and food requests, particularly among children with avid eating behaviors, may lead to reluctance in parents to consistently report these events [2]. Despite these limitations, the general correlation between EMA-reported data and baseline data indicates that EMA remains a valuable tool for capturing dynamic changes in feeding practices and children's eating behaviors. However, to mitigate these challenges, future EMA studies could incorporate design modifications, such as reducing the EMA period, providing more frequent reminders, or simplifying survey interfaces to encourage immediate and consistent reporting.

Acceptability of EMA

The study's findings indicate that our EMA protocol was generally acceptable to parents. The majority of participants reported that they were not burdened by the notification sounds

XSL•FO
RenderX

(116/118, 98%) and were not significantly affected by the length of the surveys (68/118, 58%). However, the frequency of notifications was considered burdensome by some participants (73/118, 62%), and over half (64/118, 54%) reported that the surveys hindered their daily interactions to some extent. Despite this, most parents reported that completing the EMA surveys increased their awareness of their feelings and activities, which was perceived as pleasant or neutral by the majority. These findings suggest that while EMA can be intrusive for some participants, it is generally well-tolerated and may even have positive side effects by enhancing self-awareness.

Implications for Future Research

The findings of this study have important implications for future research using EMA to investigate feeding practices and eating behaviors in children. First, the study underscores the need to further refine the EMA methodology to ensure that it is user-friendly and minimizes the burden on participants. For example, the user interface could be enhanced to reduce technical difficulties, and the frequency and timing of surveys could be further optimized based on participant feedback.

Second, the study provides evidence supporting the use of EMA to capture in-the-moment mood and context data related to parental feeding practices and children's eating behaviors. This is particularly important given the dynamic and context-dependent nature of these behaviors, which cannot be adequately captured through traditional cross-sectional methods. By employing EMA, researchers can gain a more nuanced understanding of how contextual factors, such as parental mood, feeding goals, and external pressures, influence feeding practices on a moment-to-moment basis.

Finally, this study contributes to the growing body of literature supporting the use of mobile technology for real-time data collection in behavioral research. The high compliance rates observed in this study suggest that parents are willing and able to use smartphone apps to report on their feeding practices and their children's eating behaviors, even over extended periods. This opens up new opportunities for large-scale studies that leverage mobile technology to explore complex behavioral dynamics in diverse populations.

Limitations

It is important to recognize that EMA methods have certain limitations. Since assessments are conducted repeatedly and in close relation to the behaviors, thoughts, and emotions being studied, there is a chance that the methods themselves could influence the variables being examined. This indeed is evidenced in the high percentage of participants who said they were made more aware of their emotions and activities because of the study. As such, there is the potential that throughout the EMA period, an individual's behaviors and self-reports begin to shift and this should be accounted for in subsequent analysis. Future studies should incorporate features to address participant reactivity, such as randomizing the order of questions.

It must also be noted that there are several general limitations with EMA methodologies. In particular, the surveys are self-reported so this may result in the same social desirability reporting that is seen in other self-report methods such as unrepeated questionnaires. Additionally, in an attempt to keep the surveys brief and accessible, very basic responses (ie, yes or no) were provided, which limits capturing the detail and nuance of eating occasions. Ideally, examples of child-parent interactions should be observationally recorded to allow for a richer interpretation of feeding interactions in young children with avid eating as well as determining the accuracy of parent reports. The results of this study also highlight that relying on parents to trigger event-based surveys often results in underreporting of eating occasions which impacts the accurate reporting of the frequency of eating occasions or food requests in children with avid eating. The development of wearable devices and camera recordings could be used in future research to capture more precise data on children's eating patterns.

Additionally, while EMA studies can be designed to include a diverse sample with various racial, ethnic, and socioeconomic backgrounds, it is essential to consider how likely different groups are to complete an EMA survey and how well the results can be generalized to the broader population. Although participants were offered the use of a smartphone for the study, no participant requested this, but it may still be the case that access to a smartphone was a barrier to participation. Equally, participants with higher completion rates than others may have greater time and fewer pressures around feeding interactions to be able to complete the surveys. Furthermore, there is also the possibility that parents who choose to participate and complete an EMA study may have a stronger interest in promoting healthy eating habits for their child than those who choose not to participate, which could introduce bias into the findings. Finally, although the adherence in our study aligns closely with previously reported short-term EMA studies [17], the 10-day duration may limit the ability to generalize to contexts requiring sustained monitoring over longer periods.

Conclusion

In conclusion, this study demonstrates the feasibility, acceptability, and potential of using EMA to investigate the intricate interplay between parental mood, feeding goals, contextual factors, and feeding practices in children with avid eating behaviors. While the results are promising, further refinement of the EMA methodology is needed to address technical challenges and reduce participant burden. Future research should continue to explore the use of EMA to capture the dynamic nature of feeding interactions and to inform the development of targeted interventions to promote healthy eating behaviors in children.

Acknowledgments

We would like to thank all the families for participating in the study. We are also grateful to Croker H, Farrow C, Haycraft E, Herle M, Llewellyn, and Blissett J for conceptualizing the research project and acquiring funding. This work was supported by

XSL•FO
RenderX

an Economic and Social Research Council (ESRC) research grant (ES/V014153/1). The funding organizations had no role in the design and conduct of the study; collection, management, analysis, and interpretation of data; and preparation, review, or approval of the manuscript.

Data Availability

The data sets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Each named author has substantially contributed to conducting the underlying research and drafting this manuscript. JB, CF, and EH have contributed to the conception and funding acquisition of the study. AP and KE have led the study design and data collection. AP has led the data analysis and write-up of the manuscript. JB, CF, EH, and KE have been involved in the editing and proofing of the manuscript.

Conflicts of Interest

None declared.

References

1. Pickard A, Croker H, Edwards K, et al. Identifying an avid eating profile in childhood: associations with temperament, feeding practices and food insecurity. Appetite 2023 Dec 1;191:107050. [doi: 10.1016/j.appet.2023.107050] [Medline: 37793473]
2. Edwards KL, Blissett J, Croker H, et al. Examining parents' experiences and challenges of feeding preschool children with avid eating behaviour. Appetite 2024 Jul 1;198:107372. [doi: 10.1016/j.appet.2024.107372] [Medline: 38657683]
3. Loth KA, Ji Z, Kohli N, Fisher JO, Fulkerson JA. Parents of preschoolers use multiple strategies to feed their children: findings from an observational video pilot study. Appetite 2023 Aug 1;187:106615. [doi: 10.1016/j.appet.2023.106615] [Medline: 37236362]
4. Hektner JM, Schmidt JA, Csikszentmihalyi M. Experience Sampling Method: Measuring the Quality of Everyday Life: Sage; 2007. [doi: 10.4135/9781412984201]
5. Napa SC, Prieto CK, Diener E. Experience sampling: promises and pitfalls, strength and weaknesses. In: Diener E, editor. Assessing Well-Being: Springer; 2009, Vol. 39:157-180. [doi: 10.1007/978-90-481-2354-4_8]
6. Runyan JD, Steinke EG. Virtues, ecological momentary assessment/intervention and smartphone technology. Front Psychol 2015;6:481. [doi: 10.3389/fpsyg.2015.00481] [Medline: 25999869]
7. Berge JM, Tate A, Trofholz A, et al. Examining variability in parent feeding practices within a low-income, racially/ethnically diverse, and immigrant population using ecological momentary assessment. Appetite 2018 Aug 1;127:110-118. [doi: 10.1016/j.appet.2018.04.006] [Medline: 29684414]
8. Loth KA, Ji Z, Wolfson J, Neumark-Sztainer D, Berge JM, Fisher JO. A descriptive assessment of A broad range of food-related parenting practices in A diverse cohort of parents of preschoolers using the novel real-time parent feeding practices survey. Int J Behav Nutr Phys Act 2022 Mar 2;19(1):22. [doi: 10.1186/s12966-022-01250-y] [Medline: 35236392]
9. Lopez NV, Lai MH, Yang CH, Dunton GF, Belcher BR. Associations of maternal and paternal parenting practices with children's fruit and vegetable intake and physical activity: preliminary findings from an ecological momentary study. JMIR Form Res 2022 Aug 10;6(8):e38326. [doi: 10.2196/38326] [Medline: 35947425]
10. Wardle J, Guthrie CA, Sanderson S, Rapoport L. Development of the children's eating behaviour questionnaire. J Child Psychol Psychiatry 2001 Oct;42(7):963-970. [doi: 10.1111/1469-7610.00792] [Medline: 11693591]
11. Edwards K, Croker H, Farrow C, et al. Examining parent mood, feeding context, and feeding goals as predictors of feeding practices used by parents of preschool children with avid eating behavior: protocol for an ecological momentary assessment study. JMIR Res Protoc 2024 Mar 19;13:e55193. [doi: 10.2196/55193] [Medline: 38502178]
12. APPETItE smartphone study: information video. Youtube. URL: https://www.youtube.com/watch?v=Gp5Ydnhk3K4 [accessed 2025-01-03]
13. de Vries LP, Baselmans BML, Bartels M. Smartphone-based ecological momentary assessment of well-being: a systematic review and recommendations for future studies. J Happiness Stud 2021;22(5):2361-2408. [doi: 10.1007/s10902-020-00324-7] [Medline: 34720691]
14. Myin-Germeys I, Kuppens P. The open handbook of experience sampling methodology. In: A Step-by-Step Guide to Designing, Conducting, and Analyzing ESM Studies: Leuven: Center for Research on Experience Sampling and Ambulatory Methods Leuven; 2022.
15. Pickard A, Blissett J, Farrow C, Edwards K. APPETItE - ecological momentary analysis. Published online. 2023. [doi: 10.17605/OSF.IO/N48YV]
16. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol 2008;4:1-32. [doi: 10.1146/annurev.clinpsy.3.022806.091415] [Medline: 18509902]

XSL•FO
RenderX

17. Berge JM, Trofholz A, Tate AD, et al. Examining unanswered questions about the home environment and childhood obesity disparities using an incremental, mixed-methods, longitudinal study design: the family matters study. Contemp Clin Trials 2017 Nov;62:61-76. [doi: 10.1016/j.cct.2017.08.002] [Medline: 28800894]

Abbreviations

CEBQ: Children's Eating Behavior Questionnaire
EMA: ecological momentary assessment
LPA: latent profile analysis

Edited by A Mavragani; submitted 23.09.24; peer-reviewed by M Stach, N Lopez; revised version received 07.01.25; accepted 07.01.25; published 27.02.25.

Please cite as:
Pickard A, Edwards K, Farrow C, Haycraft E, Blissett J
Capturing Everyday Parental Feeding Practices and Eating Behaviors of 3- to 5-Year-Old Children With Avid Eating Behavior: Ecological Momentary Assessment Feasibility and Acceptability Study
JMIR Form Res 2025;9:e66807
URL: https://formative.jmir.org/2025/1/e66807
doi:10.2196/66807

© Abigail Pickard, Katie Edwards, Claire Farrow, Emma Haycraft, Jacqueline Blissett. Originally published in JMIR Formative Research (https://formative.jmir.org), 27.2.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Mobilizing Health Promotion Through Canada's Student Mental Health Network: Concurrent, Mixed Methods Process Evaluation

Amy Ecclestone[1], BHSc; Brooke Linden[2], PhD; Jessica Rose[3], BSc; Kiran Kullar[3], BSc

[1]School of Public Health Sciences, University of Waterloo, Waterloo, ON, Canada

[2]Health Services and Policy Research Institute, Queen's University, 21 Arch Street, Kingston, ON, Canada

[3]Queen's University, Kingston, ON, Canada

Corresponding Author:
Brooke Linden, PhD
Health Services and Policy Research Institute, Queen's University, 21 Arch Street, Kingston, ON, Canada

Abstract

Background: Mental health issues among Canadian postsecondary students are prevalent. In tandem, an increased acknowledgment of the need for upstream mental health support has been highlighted. While the majority of institutions offer some form of mental health promotion, research suggests students are failing to access support due to barriers including lack of awareness, geographical and financial barriers, and lack of relevance in offerings. Canada's Student Mental Health Network is a web-based knowledge mobilization initiative designed to fill these gaps. With content created and curated "for-students, by-students" and reviewed by subject matter experts, the Network serves as a one-stop shop for evidence-based, mental health support for postsecondary students.

Objective: The goal of this research was to conduct the first component of a comprehensive program evaluation of the Network. This paper details a formative, process evaluation after approximately 1 year of operations, with the goal of assessing acceptability and feasibility.

Methods: Using a concurrent mixed methods study design, quantitative and qualitative data were simultaneously collected from students in order to evaluate the acceptability and feasibility of the Network as a mental health promotion resource. Quantitative data were automatically collected through Google Analytics via the website over the course of the first year of operations. Data collected included the number of users accessing the website, user engagement, and user "stickiness." Quantitative data were used to evaluate both accessibility and feasibility. Qualitative data were collected via individual, digital interviews conducted with a modest sample of students (n=8) across areas and levels of study. Qualitative data derived more detailed insights into user experience and website attributes, as well as feedback on content delivery, providing evidence used to evaluate feasibility.

Results: A total of 1200 users globally accessed the Network within the first year of operations, with Canadian users accounting for nearly 90% of total website traffic. An overall 66% engagement rate was observed, with the average user visiting 7 pages per session. Further support for the acceptability of the Network is demonstrated in the Canada-wide reach of the content development and review team. Evidence for the feasibility of the Network was observed through website use statistics indicating the most frequently viewed pages aligned with our goals: providing mental health education and increasing awareness of available resources. Qualitative feedback provided additional context surrounding the feasibility of the space, including positive feedback on the esthetics, relevance, usability, inclusion, and accessibility. Areas for content expansion and improvements to accessibility were also identified.

Conclusions: The results of this study provide evidence in support of the feasibility and acceptability of the Network as a web-based knowledge mobilization initiative in support of postsecondary students' mental health and well-being. Future research will pursue a summative, impact assessment to evaluate utility.

(JMIR Form Res 2025;9:e58992) doi:10.2196/58992

KEYWORDS

mental health; health promotion; program evaluation; postsecondary; knowledge translation

Introduction

Prevalence estimates for mental health–related problems, including above-average stress, psychological distress, and symptoms of mental illnesses, have increased significantly among Canadian postsecondary students over the past several years [1]. Indeed, data collected from Canadian postsecondary institutions via the 2022 National College Health Assessment III survey (n=11,322) revealed large proportions of students reporting moderate (47%) to high (37%) stress levels within the past 30-day period, with over half (52%) reporting their

XSL•FO
RenderX

stress levels to have impacted their academic performance [2]. Additionally, approximately one-third (33%) of respondents screened positive for serious psychological distress, while many reported having received a diagnosis for anxiety (23%) or depression (25%) [2].

Importantly, excessive stress has been linked to a number of negative academic (eg, decreased performance and decreased motivation) and health (eg, anxiety and depression) outcomes for postsecondary students [3,4]. In recent years, a more in-depth exploration of the sources of student stress has been conducted in the development and deployment of the Postsecondary Student Stressors Index. This work has revealed that stressors experienced by students within the postsecondary setting reach beyond academics, also spanning the learning environment, campus culture, personal, and interpersonal domains [5,6]. Importantly, this suggests a need for holistic mental health promotion support, including providing resources that extend beyond the academic realm. Further complicating things, postsecondary-aged Canadians were also the age group identified as the most likely to report dramatically declining mental health throughout the COVID-19 pandemic [7]. The long-term mental health fallout of the pandemic continues to be unknown, further reinforcing the need to provide well-being support to students as they navigate the stressors associated with the postsecondary setting.

Strategies to address and support postsecondary mental health have emerged at the national, provincial, and institutional levels [8]. At the postsecondary level, many institutions have attempted to improve student mental health by offering both upstream and downstream mental health services [8,9]. Upstream services aim to intervene prior to the development of symptoms of mental illness (eg, mental health promotion and mental illness prevention), while downstream interventions (eg, counseling and pharmacological therapy) aim to treat individuals who have already reached a clinical threshold. These service levels align with Keyes' seminal conceptual framework, the dual continuum of mental health and mental illness, which conceptualizes mental health and mental illness as existing on interrelated, but separate continuums [10]. A student who is experiencing a period of poor mental health, but who has not yet reached the clinical threshold for a mental illness, may benefit greatly from the use of upstream services. Unfortunately, the majority of students experiencing a decline in their mental health seek downstream treatments as an initial solution. As a result, the demand for downstream campus mental health services (eg, counseling) over the past decade has far outpaced institutions' capacities to deliver timely care [11]. Placing additional focus on upstream mental health promotion and mental illness prevention resources may contribute toward alleviating the current bottleneck observed at the treatment level by encouraging students to implement appropriate forms of help-seeking that align with their level of need. Bolstering upstream supports, including mental health promotion, may also improve students' resilience and education regarding adaptive and effective coping mechanisms, empowering students to better manage stressful situations.

While some form of mental health promotion is offered on most postsecondary campuses, a national review of mental health and well-being services on Canadian postsecondary campuses found that only 70% of respondents felt students were well-informed about available resources on campus, and about mental health issues in general [12]. Notably, many respondents (including campus service providers and institution administrators; 84%) also indicated they felt there was room for improvement in the mental health promotion efforts currently offered on their campus [12]. Further, mental health promotion activities were highly varied, with some offering programs aimed to inform students about available campus mental health services, reduce stigma, and educate students about mental illness. Other institutions focused on outreach, aiming to encourage students to seek help when needed. In many cases, the mental health supports offered were dependent on the resources (eg, financial, logistic, and staffing) available to the institution, with smaller institutions unable to provide as much support as their larger counterparts [12].

Together, these gaps and inconsistencies suggest a need for a comprehensive and holistic approach to mental health promotion, where resources are universally accessible to postsecondary students across Canada. With the overall aim of filling these gaps, Canada's Student Mental Health Network ("the Network") was created—a web-based, "one-stop shop" for mental health education and evidence-based resources. Content is tailored specifically to postsecondary students and is provided in a centralized, universally accessible location, focusing on three primary pillars of action: improving mental health literacy, encouraging the development of strong social support networks, and improving understanding of appropriate help-seeking and awareness of available resources.

A comprehensive evaluation of the Network, including formative and summative components, has been detailed and published elsewhere [13]. The purpose of this paper is to present the results of the formative, process evaluation of the Network after approximately 1 year of operation. Process evaluations are intended to assess 2 key components of program function: (1) whether the program is reaching the intended targeted population (acceptability), and (2) whether its delivery and function are consistent with program design intentions (feasibility) [14]. The key objectives of this formative evaluation were to determine the feasibility and acceptability of the Network to date and to inform the development of the protocol for the subsequent summative impact assessment.

Methods

Study Design

We used a concurrent mixed methods evaluation study design, combining both quantitative and qualitative data to gain more in-depth insights into the acceptability and feasibility of the Network. Best practices in evaluation dictate that the chosen evaluation design should align with the maturity of the initiative. Given the age of the Network at the time of this study, a formative process evaluation design was chosen, with a focus laid on process indicators selected to provide information about feasibility and acceptability, rather than focusing on outcomes evaluating impact. A future summative impact evaluation is planned for when the initiative has matured appropriately [13].

XSL•FO
RenderX

Figure 1 displays the logic model used to visually depict the program theory supporting the process evaluation.

Figure 1. Logic model depicting program theory for the Student Mental Health Network, with components relevant to the formative process evaluation highlighted.

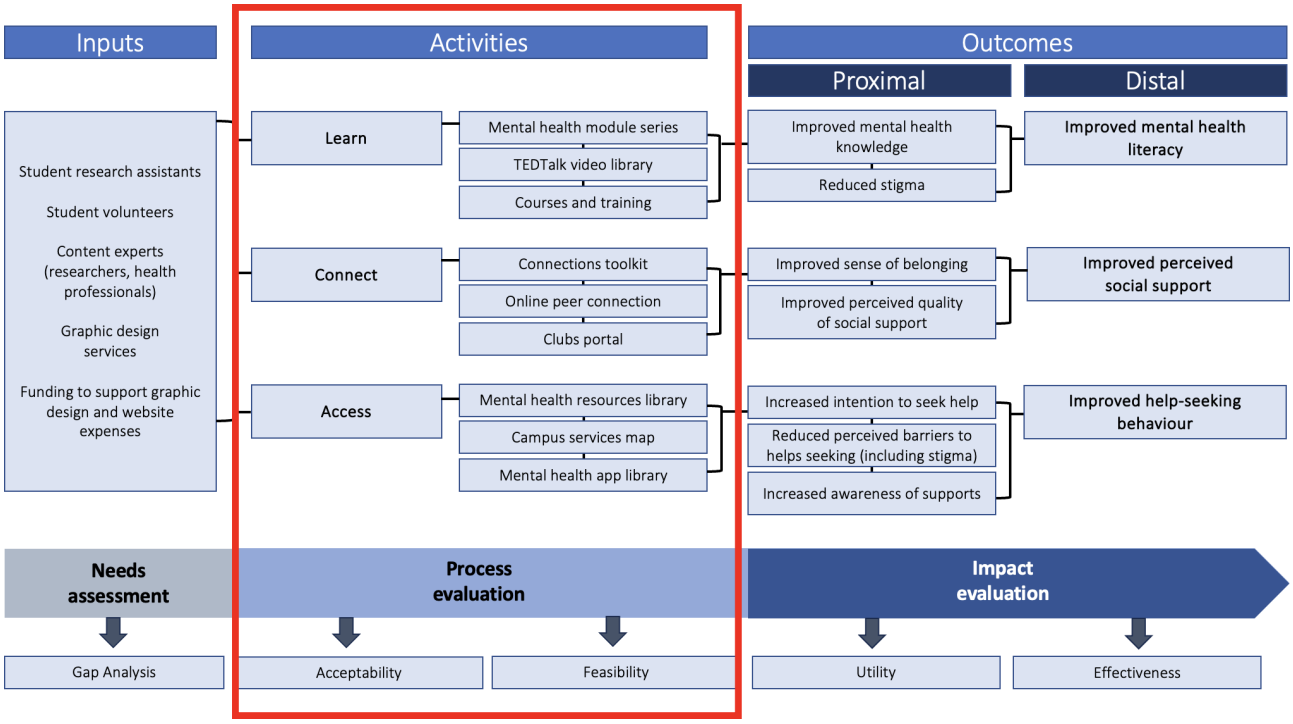

Initiative

The Student Mental Health Network is a website designed to be a web-based resource hub tailored toward postsecondary students, featuring mental health education and evidence-based tools and resources.

The target population for the Network itself is postsecondary students enrolled across Canada.

Centered around improving mental health literacy, encouraging the development of strong social support networks, and improving help-seeking behaviors, the content is intended to be relevant for all students, regardless of the type of institution attended (eg, university, college, and institute), years, areas, and levels of study. We have also aimed to develop content that is accessible, psychologically safe, and culturally competent for various subgroups of students who may be best supported by specific content areas, including medical students, members of the LGBTQ2S+ community, BIPOC students, Indigenous students, international students, and first-generation students. All content is created and curated by students and reviewed for validity by subject matter experts (eg, mental health researchers, clinicians, and medical professionals) using a collaborative, knowledge mobilization approach. Further information on the resources featured in each of the three content pillars (learn, connect, and access) has been published elsewhere [13].

Recruitment

To meet the objectives of this evaluation, data were collected using two methods: (1) quantitative website use statistics, collected through Google Analytics, and (2) qualitative data collected through individual, cognitive interviews.

To collect information regarding the reach and coverage of the Network in terms of website engagement and user experience, we used quantitative website metrics collected passively through Google Analytics. These metrics included: the total number of users, average number of users (weekly, monthly, and annually), average engagement time (time spent on each component of the website) and user "stickiness" (average length of visits), page views, event counts (clicks, file downloads), and user region. The definitions of these specific criteria per the Google Analytics glossary can be viewed in Multimedia Appendix 1.

Second, individual, cognitive interviews were conducted using a "think aloud" technique to further evaluate user experience and highlight gaps in the Network's delivery [15]. Participants were recruited using a convenience sampling method; a recruitment notice was posted on the Network website, as well as through the Network's social media channels. Participants had to be computer literate and currently enrolled at a postsecondary institution at the time of the interview. A total of 8 interviews were conducted before data saturation was reached (eg, the point at which additional interviews were not producing new information and we were beginning to hear repetitive answers to interview questions). Table 1 shows the demographic characteristics of the 8 interviewees. In terms of procedure, participants were asked to screen share over Zoom (Zoom Video Communications Inc) as they navigated the Network website and talked the interviewer through their exploration, identifying content-related issues or preferences, as well as any navigational challenges. The interviews included questions related to important website attributes (eg, appearance, content, interactivity, inclusivity or representation, and accessibility) and perceived acceptability and feasibility of the space. Interviews were audio recorded and automatically

XSL•FO
RenderX

transcribed, with field notes taken throughout the interview and immediately reviewed afterward.

Table . Descriptive statistics for demographics characteristics of cognitive interview sample.

Characteristic	Value (N=8), n (%)
Gender	
Woman	5 (60)
Man	3 (40)
Year of study	
Y1 undergraduate	1 (13)
Y2 undergraduate	2 (25)
Y3 undergraduate	2 (25)
Y4 undergraduate	2 (25)
Graduate	1 (13)
Area of study	
Health sciences	3 (38)
Science	3 (38)
Arts	1 (13)
Graduate studies	1 (13)
Institution	
Queen's University	5 (63)
Western University	1 (13)
University of Toronto	2 (25)

Data Analysis

Descriptive statistics provided through Google Analytics were displayed as frequency tables. Cognitive interview data were analyzed through the use of the website attribute codes displayed in Table 2. Field notes were taken electronically during the interviews and reviewed immediately afterward. Website attribute codes were applied to relevant transcript areas. In the event that a new issue arose that was not captured by existing problem codes, a new code was created.

Table . Website attribute codes used to categorize cognitive interview responses and appropriate corrective actions to be taken.

Website attribute	Description	Corrective action
Esthetics	Comments on the look and feel of the website, including whether the color scheme, images, and multimedia used are appropriate.	• Improve esthetics (additional graphics, multimedia, color)
Relevance	Comments on whether the content displayed on the website is not relevant to postsecondary student stress or mental health.	• Add additional relevant content • Remove irrelevant content
Accessibility	Comments on the accessibility of content is adequate (eg, ease of use for students with visual, audio, or other impairments).	• Improve accessibility features of the website
Inclusion	Comments on whether the content provided adequately caters to a wide variety of students and important subgroups.	• Change in wording • Change the order of items
Usability	Comments on ease of navigation of the website, whether the website renders on mobile and web browsers adequately, etc.	• Improve navigation • Improve optimization

Ethical Considerations

This research study was reviewed and approved for ethical compliance by the Queen's University Health Sciences and Affiliated Teaching Hospitals Research Ethics Board (TRAQ #6041988). Google Analytics data were collected passively from website users and were not connected to any identifying information. Therefore, there was no consent component associated with this part of the study. Interview participants provided verbal, informed consent to participate (including

XSL•FO
RenderX

audio and video recording of the interview) after reviewing the letter of information for the study. All qualitative data (eg, transcripts and extracted quotations) were deidentified for analyses. Interview recordings were kept confidential and stored as password-protected electronic files by the principal investigator (BL) as the data custodian. Interview participants were compensated with a CAD $20 (US $14.04) e-gift card to Amazon as thanks for their participation.

Results

Acceptability

Recall that acceptability refers to whether or not the program or initiative is reaching its intended targeted population [14]. Website use statistics were extracted from Google Analytics for the dates of May 1, 2022, to June 1, 2023, inclusive. A total of 1200 users globally accessed the Network within this time frame, where Canadian users accounted for nearly 90% of total website traffic (Multimedia Appendix 1). During this time frame, about 1500 "engaged sessions" were recorded, defined as the number of sessions (or visits) on the website that exceeded 10 seconds in length or had ≥2 page views. Users also spent some time navigating the Network, with the average user visiting approximately 7 pages per session. This resulted in an overall 66% engagement rate during the first year of operations. This data suggests that users are not only visiting the website but are spending time exploring its resources. Similarly, during this time frame, a total of 21,532 "events" (a specific interaction or occurrence on a website, such as a page view, click, or download) were recorded. Although we were not able to capture the age of these users, the continued and increased presence of users from across regions of Canada over time provides preliminary evidence in support of acceptability.

The acceptability of the Network is also reflected in the Canada-wide reach demonstrated through our content development team. Our student volunteers, who develop content and promote the Network, attend postsecondary institutions across the country and are enrolled across various programs, levels, and areas of study, including medicine, health science, political science, computer science, public health, education, and more. In addition to contributing students, we have a team of subject matter experts who serve as expert reviewers of website content to ensure its validity and reliability. These experts include faculty specializing in related research, Student Wellness Services staff and clinicians, and family physicians. Figure 2 shows the Canada-wide reach of the Network, including each of these contributing groups after 1 year of operations.

XSL•FO
RenderX

Figure 2. Map depicting the cross-Canada reach of the Student Mental Health Network after approximately 1 year of operations via the postsecondary institutions at which both student contributors and subject matter experts reside.

Feasibility

Feasibility is determined by whether a program's delivery and function are consistent with program design intentions [14]. We first gained insight into the feasibility of the Network by examining the website use statistics to determine which pages were viewed most frequently (Multimedia Appendix 1). In line with our goals, resources from the learn section of the Network were some of the most sought after the first year of operations (eg, module series, toolkits, and online courses). This was consistent with our initial needs assessment, which suggested that students were eager to improve their mental health education. Notably, several resources within the access pillar of the website were also within the most frequently visited pages (eg, campus services map, mental health resources bank, stress management resources, and mental health app library). This is consistent with our goals to encourage student awareness of available mental health resources. Finally, resources within the connect pillar of the website (eg, clubs portal and peer connection resources) were also among the most frequently visited pages, consistent with our goal to encourage the development of strong social support networks.

We also derived evidence for feasibility by conducting cognitive interviews with 8 students over the course of the evaluation period. Interviewees were asked to navigate through the Network website and comment on the appearance, relevance, accessibility, and usability. Overall, feedback was positive across interviews, with participants sharing that the site was visually appealing, easy to access, and relevant to them or their peers. Table 3 depicts the results of the interviews, as well as assigned website attribute codes.

XSL•FO
RenderX

Table . Responses to website attributes as indicated by interview participants (N=8).

Interviewee	Esthetics	Relevance	Accessibility	Inclusion	Usability
001	+[a]	+	+, x[b]	+	+
002	+	+	+	+	+
003	+, x	+	+	+	+, x
004	+	+	+	+	x
005	+	+	+	+	+
006	+	+	+, x	+	+, x
007	+, x	+	+	+	+
008	+	+	+	+	+

[a]+ indicates a positive response to this attribute.

[b]x indicates that suggestions for improvement were recommended by the participant.

Esthetics

When asked about the appearance of the website, participants indicated liking the color scheme and its consistency. The use of multimedia (eg, images, graphics, and videos) was reported to increase the visual appeal of the website. Participants also found that the content was displayed in an engaging way. In particular, they reported enjoying the format in which the toolkits and modules were presented, especially the integration of different types of media (eg, links to web pages, TED Talks, and worksheets). Participants indicated that the resources displayed as interactive multimedia were most esthetically pleasing (as well as easiest to navigate), suggesting that resources currently provided as downloadable documents be adapted to be displayed in a more esthetic, interactive alterative format.

Usability

With respect to usability, participants noted that the website was easy to navigate, with clear pathways to associated resources. Participants appreciated the list of included items at the bottom of each page, clearly indicating each subsection of website pillars. Participants were also asked which type of device they used to access the online space, and whether the website and its resources rendered properly on their device. Despite the type of device and operating system used, the website was rendered properly. Many respondents suggested that a mobile app or version be developed specifically with tablets and mobile devices in mind.

Relevance

When asked about the relevance of the website content to the student target population, all participants agreed that the content was relevant and valuable to a postsecondary audience. Participants also shared that it was clear content was intended for students from all schools and provinces and did not appear to be preferential to a single region. When asked if they would recommend the site or any of its resources to a friend or family member, every participant said yes, with some indicating that they had already done so.

Accessibility and Inclusivity

Interview participants were asked to share their impressions of the accessibility and inclusivity of the website. Respondents shared positive comments regarding the accessibility of the site, stating that content appeared to be directed toward a wide range of student backgrounds and not targeted to any specific group. Participants also indicated that they were able to find everything they were looking for with ease, suggesting the website was user-friendly in terms of accessibility. Regarding inclusivity, participants thought the website content was generally applicable to all students, the more specific resources provided (eg, graduate students' module, Indigenous well-being module, and intersectionality module) were especially inclusive as they targeted specific struggles experienced by students belonging to those groups. Respondents encouraged the development of additional resources (eg, modules and toolkits) targeting other specialty populations to enhance the inclusivity of the content. Overall, positive feedback was provided regarding both the accessibility and inclusivity of the website. Participants' recommendations regarding additional resources and content that could be improved or enhanced are displayed in Table 4.

XSL•FO
RenderX

Table . Recommendations provided by interview participants (N=8) for additional or adapted content for the Student Mental Health Network.

Area of website	Suggested addition or update
Learn	• A library of recommended self-help books fostering positive mental health, resilience, and well-being • Development of an international student module, LGBTQ+[a] module, and expansion of resources for students of color, LGBTQ+ students, international students, and Indigenous students
Connect	• A library of social media accounts that aim to promote mental health
Access	• Reformatting of the Mental Health Resources bank to be displayed similarly to the clubs bank
General	• A newsletter updating interested students on new resources added to the site, and on general student mental health news (eg, recent impactful research) • A Q&A[b] page for the site to answer any common questions students may have while navigating our site • Option to increase text size • Option to select "read aloud" or prerecorded audio for some sections and resources • French translation of website content • Development of a mobile-optimized version of the Network

[a]LGBTQ+: lesbian, gay, bisexual, transgender, queer, and other sexual and gender minority.

[b]Q&A: question and answer.

Discussion

Principal Findings

As prevalence estimates for mental health issues have increased over the past decade [1], increasing attention has been paid to postsecondary mental health. The development of institutional frameworks and guiding documents such as the Okanagan Charter [16] and Canada's National Standard for the Psychological Safety and Wellbeing of Postsecondary Students [17] have emerged, alongside new forms of data collection such as the Canadian Campus Wellbeing Survey [18]. Across the board, the recommendation for improvements in the targeting of mental health support for students is evident [8]. With the demand for downstream treatment-based services (eg, counseling and medication) continually surpassing the campus' capacities to provide care to all, there is a glaring need to work upstream. Bolstering and improving the targeting of these upstream mental health supports is imperative to reducing the burden on downstream campus mental health services by empowering students to improve their resiliency and develop a toolkit of resources to turn to during times of increased stress where lower-intensity support may result in marked improvements.

The Student Mental Health Network offers 1 such solution as a web-based mental health promotion resource specifically targeted toward postsecondary students and universally accessible across Canada. The Network provides unlimited access to mental health education and resources, addressing many of the common barriers cited by students attempting to access mental health supports, including financial restrictions, lengthy wait times, stigma, and logistical access (eg, time, location, or travel-based constraints) [19]. The findings from this formative process evaluation have provided preliminary evidence for both the acceptability and feasibility of the Network. Results showed the imitative is reaching its intended target population (acceptability), and that its delivery and function are consistent with program design intentions (feasibility). In addition to positive feedback from the target population, opportunities for growth were also identified.

Results from Google Analytics provided evidence for accessibility (that the initiative is reaching its intended targeted population), with web use statistics indicating that a total of 1200 users globally accessed the Network within the first year of operations, the vast majority of whom resided in Canada (90%). This data suggests that the Network is reaching its intended target population. Providing additional support to this conclusion is the Canada-wide reach of the Network, evidenced through the regional distribution of our content contributors and expert reviewers, including postsecondary students, mental health researchers, clinicians, and physicians from across the country. Users also spent time navigating the Network, with the average user visiting approximately 7 pages per session. This finding provides evidence that users are not simply popping in and out of the space, but rather, are taking the time to view multiple pages and absorb some of the information provided. Overall, the Network reported a high engagement rate (66%) within its first year of operations. Overall, this data suggests that users are not only visiting the website but are spending time exploring its resources, as we hoped they would. Collecting passive user data via Google Analytics allowed us to cast a much wider net with respect to determining usership and avoid issues common to survey-based research, such as low response rates and selection bias.

Feasibility was supported by Google Analytics data in addition to the data derived from cognitive interviews. Recall that the

XSL•FO
RenderX

goal of the Network was to provide Canadian postsecondary students with a web-based, “one-stop shop” for mental health education and evidence-based resources focusing on mental health education, social support, and improving awareness of available mental health resources. After the first year of operations, results suggest that the Network’s delivery and function have been consistent with our goals. Valid and reliable content has been created across all 3 areas of focus and is well-received by students. Data from Google Analytics showed that resources within all 3 sections fell within the top 20 most frequently visited pages of the Network, with resources in the learn and access sections being the most popular. Results from interviews further supported students’ endorsement of the Network and its resources, with positive reviews regarding website esthetics, relevance of content, accessibility, inclusion, and usability. Interviewees identified some areas for improvement, including additional content to fill some existing gaps, as well as minor esthetic or display changes to existing resources, but no resource was identified for removal or irrelevance. Here, we see the added value of the mixed methods design of the evaluation, where valuable contextual information was provided as a result of the cognitive interviews compared to what would have been gleaned from the collection of quantitative data alone. These interviews allowed for a detailed walkthrough of the website, facilitating a rich conversation about numerous elements (eg, relevance, accessibility, and esthetics). Overall, the design elements highlighted by study participants echo those identified by Garret et al [20] as key to thoughtful website design and driving user engagement.

Limitations

Despite its strengths, there are several limitations to this research that should also be acknowledged. While the Network is a novel initiative aiming to fill important gaps, its novelty also presents evaluation challenges given the absence of a standardized process for evaluating this type of initiative. Despite this limitation, we did our best to thoughtfully design a comprehensive program theory and evaluation plan, including both the process evaluation outlined here, as well as a future summative impact evaluation. In terms of data collection, only 8 interviews were conducted with all students attending postsecondary institutions in Ontario, resulting in limited geographic variability in our qualitative sample. However, we did see good variability in terms of other characteristics, such as gender, area, and level of study. Furthermore, while our qualitative sample was small, we did find that we reached data saturation in our interviews. Additionally, data collected through Google Analytics were able to provide us with information on the number of people visiting the website over time and from which regions, but not the age groups to which those visitors belonged. As a result, we can conclude that website traffic increased over time, but we do not have quantitative evidence to confirm that those visitors were postsecondary-aged individuals. Additionally, while the Network has demonstrated a cross-Canada reach after the first year of operations, there is an uneven distribution of users and content contributors with the majority situated in Ontario. This may be due to the fact that the Network was initially created by a group of students at a single, Eastern Ontario university, or due to the vast majority of postsecondary institutions in Canada being located in this province. Moving forward, we plan to continue working on extending the reach of the Network across the country. Finally, website content is currently only available in English, which poses an accessibility challenge for French-speaking postsecondary students. Going forward with Network development, we aim to translate content into French to provide equitable access to mental health resources to Canadian students in both official languages.

Conclusions

Bolstering upstream supports like mental health promotion may help alleviate the current bottleneck observed at the downstream service level on postsecondary campuses. Canada’s Student Mental Health Network aims to contribute toward alleviating this issue by delivering a novel, innovative method of providing universally accessible, evidence-based mental health promotion to postsecondary students across the country. The Network provides a centralized location for mental health education and resources, created for-students, by-students, and validated by subject matter experts, including physicians, clinicians, and mental health researchers. The space will contribute to breaking down common barriers to help-seeking, such as geographic restraints (eg, by providing universally accessible content that can serve online students, as well as those with travel or logistical obstacles), financial restraints (eg, by providing cost-free education and resources wherever possible), and lack of awareness (eg, by maintaining an active social media presence and partnering with Student Wellness Services at postsecondary institutions across Canada to bolster promotion). The participatory, for-students, by-students approach is a key strength of the Network and a characteristic that many students have highlighted as a unique and attractive quality of the space. The Network is a first-of-its-kind initiative that aims to fill gaps in current mental health promotion resources provided to students using an innovative, collaborative design. Through the provision of relevant and focused content and the facilitation of social connection, the Network will support and empower students to take ownership of their mental health and well-being by providing them with the toolbox they need to build and maintain resilience.

The formative process evaluation component for the Network has now been completed, with data collection for the impact evaluation currently underway. The next steps will include the ongoing creation, curation, and refreshing of the Network resources in order to meet the changing needs of the target population, as well as to fill any gaps identified throughout the evaluation process. Resources will continue to be updated to ensure a high level of accessibility and relevance. Preliminary results from this evaluation show the Network is meeting its preliminary objectives. The summative assessment of the Network will aim to further evaluate these objectives and provide additional evidence of overall impact.

XSL•FO
RenderX

Acknowledgments

The authors would like to acknowledge and thank all the student contributors involved in the development and curation of content for the Student Mental Health Network [21]. Dr. Brooke Linden's position is funded by the Bell Canada Mental Health and Anti-Stigma Research Chair.

Data Availability

The data generated or analyzed during this study are not publicly available due to Ethics Board restrictions but are available from the corresponding author for secondary analysis upon reasonable request and upon submission of an additional ethics request to the approving Ethics Board.

Authors' Contributions

AE and JR led data collection and analysis. AE, JR, and KK took lead roles in the drafting and writing of the manuscript. BL was the principal investigator on the research project and held a supervisory role throughout the data collection, analysis, reporting, writing, and editing of the manuscript. All authors approved the manuscript for final submission.

Conflicts of Interest

All authors have a nonfinancial competing interest as founding members of the Student Mental Health Network.

Multimedia Appendix 1
Google Analytics terms glossary and website usage statistics.
[DOCX File, 216 KB - formative_v9i1e58992_app1.docx]

References

1. Linden B, Boyes R, Stuart H. Cross-sectional trend analysis of the NCHA II survey data on Canadian post-secondary student mental health and wellbeing from 2013 to 2019. BMC Public Health 2021 Dec;21(1). [doi: 10.1186/s12889-021-10622-1]
2. Canadian reference group data report spring 2022. American College Health Association. 2022. URL: https://www.acha.org/wp-content/uploads/2024/07/NCHA-III_SPRING_2022_CANADIAN_REFERENCE_GROUP_DATA_REPORT.pdf [accessed 2025-02-07]
3. Pascoe MC, Hetrick SE, Parker AG. The impact of stress on students in secondary school and higher education. Int J Adolesc Youth 2020 Dec 31;25(1):104-112. [doi: 10.1080/02673843.2019.1596823]
4. Duffy A, Keown-Stoneman C, Goodday S, et al. Predictors of mental health and academic outcomes in first-year university students: identifying prevention and early-intervention targets. BJPsych Open 2020 May 8;6(3):e46. [doi: 10.1192/bjo.2020.24] [Medline: 32381150]
5. Linden B, Stuart H. Post-secondary stress and mental well-being: a scoping review of the academic literature. Can J Commun Ment Health 2020 May 1;39(1):1-32. [doi: 10.7870/cjcmh-2020-002]
6. Linden B, Stuart H, Ecclestone A. Trends in post-secondary student stress: a pan-Canadian study. Can J Psychiatry 2023 Jul;68(7):521-530. [doi: 10.1177/07067437221111365] [Medline: 35791667]
7. National polling initiative. Mental Health Research Canada. 2020. URL: https://www.mhrc.ca/national-polling-covid [accessed 2022-05-17]
8. Monaghan C, Linden B, Stuart H. Postsecondary mental health policy in Canada: a scoping review of the grey literature: Politique de santé mentale post-secondaire au Canada: un examen de la portée de la littérature grise. Can J Psychiatry 2021 Jul;66(7):603-615. [doi: 10.1177/0706743720961733] [Medline: 33016127]
9. DiPlacito-DeRango ML. Acknowledge the barriers to better the practices: support for student mental health in higher education. Can J Scholarsh Teach Learn 2016;7(2). [doi: 10.5206/cjsotl-rcacea.2016.2.2]
10. Keyes CLM. The mental health continuum: from languishing to flourishing in life. J Health Soc Behav 2002 Jun;43(2):207-222. [doi: 10.2307/3090197] [Medline: 12096700]
11. McGartland K. White paper: mental health on campus. Centre for Innovation in Campus Mental Health. 2018. URL: https://campusmentalhealth.ca/wp-content/uploads/2018/03/White-Paper-Mental-Health-on-Campus1.pdf [accessed 2022-05-17]
12. Jaworska N, De Somma E, Foneska B, Heck E, MacQueen G. Mental health services for students at postsecondary institutions: a national survey. Can J Psychiatry 2016;12(61). [doi: 10.1177/0706743716640752]
13. Ecclestone A, Linden B, Monaghan C, Zheng S. Canada's Student Mental Health Network: protocol for a comprehensive program evaluation. JMIR Res Protoc 2023 Jun 22;12:e41521. [doi: 10.2196/41521] [Medline: 37347511]
14. Rossi PH, Lipsey MW, Freeman HE. Evaluation: A Systematic Approach, 7th edition: SAGE Publications; 2003.
15. Wolcott MD, Lobczowski NG. Using cognitive interviews and think-aloud protocols to understand thought processes. Curr Pharm Teach Learn 2021 Feb;13(2):181-188. [doi: 10.1016/j.cptl.2020.09.005] [Medline: 33454077]

XSL•FO
RenderX

16. Okanagan charter: an international charter for health promoting universities and colleges. Best Practices in Canadian Higher Ed. 2015. URL: https://bp-net.ca/wp-content/uploads/2019/03/Okanagan-Charter.pdf [accessed 2025-01-20]
17. Mental health and well-being for post-secondary students (CSAZ 2003). : Canadian Standards Association; 2020.
18. Fagan MJ, Wunderlich KB, Faulkner G. Psychological distress among Canadian postsecondary students: a repeated cross-sectional analysis of the Canadian Campus Wellbeing Survey (CCWS) between spring 2020-2023. J Am Coll Health 2024 Jul 31;2024:1-6. [doi: 10.1080/07448481.2024.2382438] [Medline: 39083793]
19. Moghimi E, Stephenson C, Gutierrez G, et al. Mental health challenges, treatment experiences, and care needs of post-secondary students: a cross-sectional mixed-methods study. BMC Public Health 2023 Apr 6;23(1):655. [doi: 10.1186/s12889-023-15452-x] [Medline: 37020282]
20. Garett R, Chiu J, Zhang L, Young SD. A literature review: website design and user engagement. Online J Commun Media Technol 2016 Jul;6(3):1-14. [Medline: 27499833]
21. Student Mental Health Network. URL: www.studentmentalhealthnetwork.ca [accessed 2025-02-07]

Edited by A Mavragani; submitted 29.03.24; peer-reviewed by A AL-Asadi, KW Tay; revised version received 27.01.25; accepted 28.01.25; published 27.02.25.

Please cite as:
Ecclestone A, Linden B, Rose J, Kullar K
Mobilizing Health Promotion Through Canada's Student Mental Health Network: Concurrent, Mixed Methods Process Evaluation
JMIR Form Res 2025;9:e58992
URL: https://formative.jmir.org/2025/1/e58992
doi:10.2196/58992

© Amy Ecclestone, Brooke Linden, Jessica Rose, Kiran Kullar. Originally published in JMIR Formative Research (https://formative.jmir.org), 27.2.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Evaluation of Financial Support Workshops for Patients Under State Pension Age With Degenerative Cervical Myelopathy: Survey Study

Tanzil Rujeedawa[1], MB BChir; Zahabiya Karimi[2], BA; Helen Wood[2], MA; Irina Sangeorzan[2], MSc; Roy Smith[2], BSc; Iwan Sadler[2]; Esther Martin-Moore[2]; Adrian Gardner[3], PhD; Andreas K Demetriades[4], BChir, MB MPhil; Rohitashwa Sinha[5], BMBS, PhD; Gordan Grahovac[6], MD, PhD; Antony Bateman[7], MSc; Naomi Deakin[1], PhD; Benjamin Davies[1], BSc, MB ChB, MPhil, MRCS

[1]Division of Neurosurgery, Department of Clinical Neurosciences, University of Cambridge, The Old Schools, Trinity Ln, Cambridge, United Kingdom
[2]Myelopathy.org, Cambridge, United Kingdom
[3]Aston University, Birmingham, United Kingdom
[4]Edinburgh Spinal Surgery Outcome Studies Group, Department of Neurosurgery, Royal Infirmary Edinburgh, Edinburgh, United Kingdom
[5]Department of Neurosurgery, Leeds General Infirmary, Leeds, United Kingdom
[6]King's College Hospital, London, United Kingdom
[7]Royal Derby Spinal Centre, Derby, United Kingdom

Corresponding Author:
Benjamin Davies, BSc, MB ChB, MPhil, MRCS
Division of Neurosurgery, Department of Clinical Neurosciences, University of Cambridge, The Old Schools, Trinity Ln, Cambridge, United Kingdom

Abstract

Background: Degenerative cervical myelopathy (DCM), a form of slow-motion and progressive spinal cord injury caused by spinal cord compression secondary to degenerative pathology, leads to high levels of disability and dependence, and may reduce quality of life. Myelopathy.org is the first global scientific and clinical charity for DCM, providing an accessible platform freely disseminating information relevant to the DCM diagnosis and its treatment. Significant transient and long-term change to earnings do occur and can thrust individuals into poverty. People with DCM face many challenges accessing state financial assistance. This can have a cumulative negative financial effect due to the association between DCM and low socioeconomic index. Financial support available to patients under pension age include Universal Credit (UC), a payment that helps with living costs, and Personal Independence Payment (PIP), which helps with extra living costs if someone has both a long-term health condition or disability and difficulty doing certain everyday tasks.

Objective: This study aimed to assess if delivering workshops centered around access to financial support could assist people with DCM living in the United Kingdom.

Methods: A series of 2 internet-based workshops was targeted at accessing financial support for English patients under the state pension age, with an anonymized survey delivered to participants after each session. The first session was on UC and the second on PIP. The survey consisted of a mixture of Likert scales, free text and yes or no answers. Survey responses were analyzed using descriptive statistics and free text answers underwent inductive thematic analysis.

Results: The average rating on the use of UC was 9.00/10. Presession self-rated confidence levels were 5.11/10 rising to 8.00/10. The mean score of wanting further similar sessions was 8.67/10 with 56% (5/9) of participants wanting one-to-one sessions. For PIP, the average session use rating was 10/10. Presession self-rated confidence levels were 4.43/10 rising to 9.57/10. The mean score of wanting further similar sessions was 8.71/10, with 43% (3/7) of participants wanting one-to-one sessions . Following inductive thematic analysis, themes regarding the usefulness of such sessions and the challenges to accessing financial support emerged. One participant gave negative feedback, which included the length of the session and perceived problems around confidentiality and data protection.

Conclusions: The pilot series was largely perceived as a success, with participants finding them useful and increasing their self-rated confidence in navigating the UK financial support system. Given the small sample size, it is hard to predict the success of future sessions. Finally, given that the hurdles in accessing financial support extend beyond DCM, such workshops may be relevant to other organizations.

(JMIR Form Res 2025;9:e59032) doi:10.2196/59032

XSL•FO
RenderX

KEYWORDS

myelopathy; degenerative; spine; spinal; benefits; aid; financial; money; income; poverty; disability; disabled; finance; workshop; education; service; access; accessibility; navigate; confidence; government

Introduction

Degenerative cervical myelopathy (DCM) is a form of slow-motion and progressive spinal cord injury caused by spinal cord compression secondary to degenerative pathology. This includes disease processes such as cervical spondylosis, ossification of the posterior longitudinal ligament, ossification of the ligamentum flavum, and degenerative disc disease [1,2]. Globally, it is estimated to affect 2% of adults, although less than 10% are formally diagnosed at this time [3,4]. DCM can cause sensorimotor dysfunction in the upper or lower limbs, gait disturbance, bladder or bowel dysfunction, and pain [2,5-9]. Surgery is the mainstay of treatment for DCM and aims to decompress the spinal cord [2,10-12]. Whilst this typically stops further deterioration and may afford meaningful symptomatic improvement, recovery is generally incomplete. People with DCM experience high levels of disability and dependence, and are among those with the lowest quality of life scores for any chronic disease [13]. In fact, almost half of diagnosed individuals may be dependent on the support of others for activities of daily living (41.9% (326/778) in a recent study [14]).

This landscape, in the study mentioned above, results in great socioeconomic impact, with over a third of participants unable to return to work (35.7%; 278/778 of people with DCM [14]) due to DCM. Furthermore, 25.8% (201/778) were employed full time and 11.3%(88/778) were employed part-time [14]. In addition, 35.7% (278/778) were unable to work due to their disability [14]. Surgery can be helpful, with Lønne et al [15] showing that by 12 months after surgery, 65% (589/906) had returned to work, increasing to 75% (680/906) by 36 months. Clearly, therefore, DCM impairs patients' ability to contribute economically before retirement and can lead people into poverty [16]. Davies et al [17] found that, based on the average age of cases at presentation in the United Kingdom and an average age of retirement of 65 years, there is a potentially 15.1 years of affected productivity in people with DCM [17]. This equates to an estimated lifetime inflated loss of income of up to £347,112 (US $435,150) using UK Office for National Statistics figures or £474,719 (US $595,121) using Office for Economic Co-Operation figures [17]. In fact, there is an estimated annual loss of productivity of £362.6m (US $453.3m), in addition to disability financial support totaling £280.2m (US $350.3m) and an overall cost to society for this cohort of £681.6m (US $ 853.9m) in the United Kingdom [17]. This has a cumulative negative financial effect due to the association of DCM with a low socioeconomic index, a delayed diagnosis, and poor outcomes [18].

Whilst this provides an estimate of the current "demand" and potential future "societal cost" of DCM, the practicalities of ensuring that people with DCM can access support is seldom straightforward, since social support systems differ by both jurisdiction and patient circumstance. A study from Norway, where financial support is available to all citizens suffering ill-health, found that 20% of patients received medical income compensation financial support 1 year before surgery, rising to 100% at the time of surgery. However, for most countries, this "gold standard" access is complex. First, the term "DCM" is relatively new, having been introduced in 2015 [1]. Furthermore, being focused on undergoing treatment and lacking a true understanding of the long-term implications of the disease, people with DCM rarely recognize (or accept) the need for financial support until needs are heightened [19-21]. Professionals may be aware of these challenges, as patients often turn to their medical team for support, guidance, and written opinions. However, many health care professionals are not trained to advise patients on available options and are likely unequipped to best navigate the claim and assessment process.

Myelopathy.org is the first global scientific and clinical charity for DCM. The charity provides an accessible internet-based platform freely disseminating information relevant to the DCM diagnosis and its treatment. As an organization, it hosts peer-to-peer support forums, and fields inquiries from those affected by DCM. Across this media, contacts requesting information or assistance for accessing financial support is a frequent topic.

Several active organizations exist that offer advice and guidance to people applying for financial support. These include Hastings Advice and Representation Centre (HARC) [22], Citizens Advice [23], and Welfare Benefits Unit [24]. In fact, certain charities such as the British Heart Foundation have dedicated web pages specific to applying for financial support [25]. Some, such as Age Scotland, offer workshops to their beneficiaries about accessing benefits [26]. However, there is no literature assessing the utility of these interventions. We hypothesized that delivering workshops centered around access to financial support could assist people with DCM.

Methods

Overview

The authors, in collaboration with Myelopathy.org, piloted a series of 2 workshops in May 2023 targeted at accessing financial support for English patients under the state pension age.

Participants

People with DCM were recruited to the sessions virtually, using an advertisement within the Facebook (Meta) support group run by Myelopathy.org, which had 2065 members at the time of the study. Interested participants contacted the charity directly by email to book sessions. This was the only form of recruitment. Inclusion criteria was any person with DCM under state pension age from England. Where a session was oversubscribed, a random number generator was used to select 10 participants.

XSL•FO
RenderX

Procedures

The sessions were delivered in a paid partnership with HARC, a charitable organization, which provides welfare financial support advice and information to the people of East Sussex and the South Coast of England [22]. It aims to address inequalities, relieve poverty, and improve the quality of life for vulnerable and disadvantaged individuals and families [22]. Funding was from the National Lottery community fund, a UK specific fund [27].

The overall aim was to equip participants with the skills to navigate the financial support system independently. The workshop program consisted of 2 sessions delivered as a set. All sessions were held virtually on Zoom (Zoom Video Communications) and each lasted up to 2 hours. The workshop format was chosen in partnership with HARC. HARC provided expertise from generic sessions delivered previously, with the chosen workshop format encouraging interactivity and the ability to ask questions, as well as being cost-effective.

The sessions were targeted to patients currently under state pension age, allowing content to be tailored to age demographics. This is dependent on when someone was born [28]. The first session was on Universal Credit (UC), a payment that helps with living costs [28] and the second on Personal Independence Payment (PIP), which helps with extra living costs if someone has both a long-term health condition or disability and difficulty doing certain everyday tasks or getting [29]. The sessions will be termed UC and PIP, respectively. Each session consisted of a comprehensive overview of either UC or PIP and their application process with a question time at the end.

Measures

Participants were evaluated using an anonymized survey (Google Forms), sent by email to the participants by the charity after attending each workshop.

Analysis

Survey responses were analyzed using descriptive statistics (mean, median, mode, and range) using R (version 4.2.1; R Foundation for Statistical Computing). Free text answers underwent inductive thematic analysis to identify key themes of feedback by TR and BD separately. Themes were identified manually with TR an BD having a discussion afterwards to ensure validity. Given the small quantity of data, no codebook software was used.

Ethical Considerations

This study was conducted with ethical approval from the Myelopathy.org Board (M2023001). At the start of the workshop, participants were provided with an overview of the workshop and that feedback would be asked at the end that may be used for publication. By continuing into the workshop, participants were providing informed consent to participate. All data collected were anonymous. No incentives were offered for completion of the surveys.

Results

A total of 24 people with DCM registered for the workshops. There were 9 attendees in UC and 7 in PIP, with 2 individuals dropping out from the first session to the second. Demographic information was not recorded.

Quantitative Feedback

Participants were asked to rate the sessions across four questions, using a Likert scale (1 to 10) for each. Figure 1 summarizes the responses to the survey administered after each session, while Tables S1 and S2 (Multimedia Appendix 1) provide responses at the level of each participant.

The average rating on the usefulness of UC was 9.00/10 (SD 2.82). Presession self-rated confidence levels were 5.11/10 (SD 2.73) rising to 8.00/10 (SD 2.62) after the session. The mean score of wanting further similar sessions was 8.67/10 (SD 2.87), with 56% (5/9) of participants wanting one-to-one sessions.

For PIP, the average session usefulness rating was 10/10 (SD 0), with the desire for further sessions. Presession confidence levels were 4.43/10 (SD 1.76) rising to 9.57/10 (SD 0.49). The mean score of wanting further similar sessions was 8.71/10 (SD 1.75) with 43% (3/7) of participants wanting one-to-one sessions.

Figure 1. A. Box plots showing the response in the survey administered after the UC session. B. Box plots showing the response in the survey administered after the PIP session. UC: universal credit; PIP: personal independence payment.

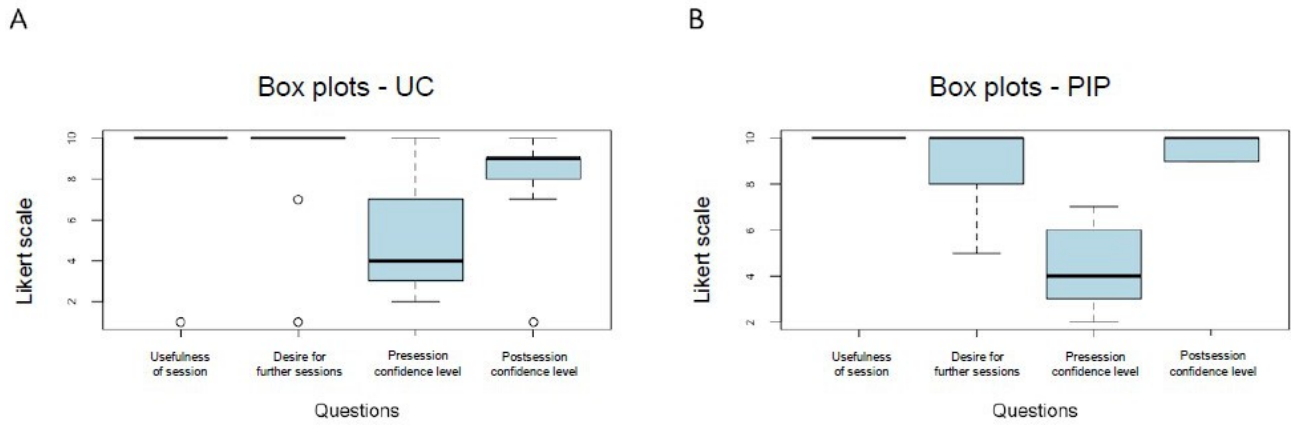

XSL•FO
RenderX

Qualitative Feedback

Several themes emerged during the thematic analysis, notably with regards to the usefulness of the sessions and the challenges to accessing financial support.

An important theme that emerged was the usefulness of the sessions. Participants noted that the sessions were "interactive," "amazing," "informative," and that they had "enjoyed this very much." Both sessions were generally viewed as very helpful, with several participants noting the information given being particularly useful with a participant noting that the sessions contained "specific information that was tailored to participants' needs, rather than just a bullet point description of financial support available." The information on what the questions on the forms actually meant and how to answer them was noted as especially helpful. In fact, most participants noted that they could not suggest any improvements for the sessions. The desire for further sessions was also noted, with several participants keen for more as demonstrated by one of the participants saying that they "would like more of these types of sessions as they are so so beneficial and also make you feel like you are less alone in the journey," and another saying "just more sessions would be amazing." One participant even suggested group discussions that include Myelopathy.org members who had made successful claims.

The challenges to accessing financial support were also highlighted. The confusing system was the main challenge noted. The lack of knowledge about eligibility, lack of understanding and knowledge from professionals and fear of judgment were also mentioned. One participant "felt alone and like I was being judged" and another "never applied before because I never felt like I was entitled to any, but I feel more confident now." In addition, participants found it hard to understand how detailed answers to the forms needed to be. Finally, physical disability making it hard to fill in forms was another issue.

Certain negative themes also came across from only 1 of the participants. They did not find the session helpful due to what was perceived as "unnecessary personal introductions, and waffle," and some technical issues that prevented the screen from being shared properly at the start. In addition, they felt that details of the session were not properly explained in advance. They also perceived that the facilitator had strayed "into areas of confidentiality and data protection." This was due to an incident whereby another one of the participants had mentioned to the facilitator that they were going to the toilet and the facilitator accidentally mentioned it to the whole group. The participant involved had no issues with this.

Discussion

With the support of a charitable partner, Myelopathy.org delivered a series of workshops that were effective in increasing awareness and confidence in obtaining financial support in the United Kingdom. No other similar study was found through a search on PubMed for comparison, even in a non-myelopathy context, thereby illustrating the novel nature of this initiative. Workshops were considered highly useful (9-10/10) and increased confidence by >40%.

Following inductive thematic analysis, themes regarding the usefulness of the sessions and the challenges to accessing financial support emerged. The interactivity of the session was viewed as an important positive. The usefulness of the sessions was also noted as well as the desire for further sessions. The challenges of applying for financial support were also explored, notably the confusing system, physical disability, the lack of knowledge about eligibility, lack of understanding and knowledge from professionals, and fear of judgment. One participant had negative feedback, which included the length of the session and perceived problems around confidentiality and data protection. The feedback was taken seriously and given to the facilitator. An apology was presented to the participant by ZK, the director of Myelopathy.org, and the misunderstanding was resolved. This participant did not attend PIP and could represent a subgroup of the people with DCM population for whom another mode of engagement would be more suitable such as anonymized access of online web pages or documents.

While the intervention was considered successful, it is important to consider areas for improvement, particularly to ensure the intervention can be sustained and exported to other regions or countries. Of note, participants were split about wanting further one-to-one sessions despite thinking further sessions would be helpful. This could be potentially explained by the participants preferring to be in a group setting rather than alone.

Experience from other settings is limited and is an avenue that should be explored. It should also be acknowledged that individuals who sign up for a group workshop may be more likely to like such settings. This initial evaluation also does not collect objective data relevant to actual or successful financial support applications. Other limitations include the fact that the sessions were focused solely on financial support available in England and particular economic situations. How these findings relate to outside England and situations such as self-employed status is therefore unclear. Furthermore, delivery of sessions virtually, available to only a limited number of participants, may provide only a narrow snapshot of people with DCM. This situation is compounded by the fact that Facebook-only recruitment may skew participant demographics. Another limitation is that no demographic data were collected that could have helped in contextualizing findings and identifying possible trends across specific demographic groups.

Despite the limitations outlined above, the participants largely deemed this initial pilot to be a welcome success. Given this success, Myelopathy.org aims to provide further similar sessions both in England and other countries. In addition, this can serve as a template for small charities focusing on relatively unknown conditions that cause a financial burden. The socioeconomic impact of DCM is a global research priority [30]. The chronic disability experienced by most, even if transiently during the drawn-out diagnostic process and aftermath from surgery, can dramatically alter household incomes and thrust people into poverty [16]. There is a cumulative negative effect due to the association of DCM with low socioeconomic conditions.

XSL•FO
RenderX

Conclusion

Despite the presented limitations and a small amount of negative feedback from one individual, the sessions were largely perceived favorably, with participants finding them useful and increasing their self-rated confidence in navigating the UK financial support system. The small sample size limits any further extrapolation; it is indeed hard to predict how future sessions will be received. Nonetheless, Myelopathy.org intends to co-host future sessions with the lessons learned from this pilot. Finally, difficulties in accessing financial support extend beyond DCM, and hence, the lessons learned in this evaluation may be helpful to other patient groups and support organizations.

Acknowledgments

BD is supported by an NIHR Clinical Doctoral Fellowship. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research, or the Department of Health.

Data Availability

The datasets generated during and/or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

TR contributed to conceptualization, manuscript drafting, and preparation. ZK, HW, I Sangeorzan, RS, I Sadler, and EMM contributed to conceptualization. ND contributed to manuscript review. BD contributed to conceptualization and manuscript review.

Conflicts of Interest

BD, HW, I Sadler, and RS are Trustees of Myelopathy.org. ZK is the director of Myelopathy.org.

Multimedia Appendix 1
Supplementary tables.
[DOCX File, 16 KB - formative_v9i1e59032_app1.docx]

References

1. Nouri A, Tetreault L, Singh A, Karadimas SK, Fehlings MG. Degenerative cervical myelopathy: epidemiology, genetics, and pathogenesis. Spine (Phila Pa 1976) 2015 Jun 15;40(12):E675-E693. [doi: 10.1097/BRS.0000000000000913] [Medline: 25839387]
2. Badhiwala JH, Ahuja CS, Akbar MA, et al. Degenerative cervical myelopathy - update and future directions. Nat Rev Neurol 2020 Feb;16(2):108-124. [doi: 10.1038/s41582-019-0303-0] [Medline: 31974455]
3. Grodzinski B, Stubbs DJ, Davies BM. Most degenerative cervical myelopathy remains undiagnosed, particularly amongst the elderly: modelling the prevalence of degenerative cervical myelopathy in the United Kingdom. J Neurol 2023 Jan;270(1):311-319. [doi: 10.1007/s00415-022-11349-8] [Medline: 36056204]
4. Smith SS, Stewart ME, Davies BM, Kotter MRN. The prevalence of asymptomatic and symptomatic spinal cord compression on magnetic resonance imaging: a systematic review and meta-analysis. Global Spine J 2021 May;11(4):597-607. [doi: 10.1177/2192568220934496] [Medline: 32677521]
5. Tracy JA, Bartleson JD. Cervical spondylotic myelopathy. Neurologist 2010 May;16(3):176-187. [doi: 10.1097/NRL.0b013e3181da3a29] [Medline: 20445427]
6. Kalsi-Ryan S, Karadimas SK, Fehlings MG. Cervical spondylotic myelopathy: the clinical phenomenon and the current pathobiology of an increasingly prevalent and devastating disorder. Neuroscientist 2013 Aug;19(4):409-421. [doi: 10.1177/1073858412467377] [Medline: 23204243]
7. Harrop JS, Naroji S, Maltenfort M, et al. Cervical myelopathy: a clinical and radiographic evaluation and correlation to cervical spondylotic myelopathy. Spine (Phila Pa 1976) 2010 Mar 15;35(6):620-624. [doi: 10.1097/BRS.0b013e3181b723af] [Medline: 20150835]
8. Tetreault L, Goldstein CL, Arnold P, et al. Degenerative cervical myelopathy: a spectrum of related disorders affecting the aging spine. Neurosurgery 2015 Oct;77 Suppl 4:S51-S67. [doi: 10.1227/NEU.0000000000000951] [Medline: 26378358]
9. Boerger T, Alsouhibani A, Mowforth O, et al. Moving beyond the neck and arm: the pain experience of people with degenerative cervical myelopathy who have pain. Global Spine J 2022 Sep;12(7):1434-1442. [doi: 10.1177/2192568220986143] [Medline: 33626937]
10. Fehlings MG, Ibrahim A, Tetreault L, et al. A global perspective on the outcomes of surgical decompression in patients with cervical spondylotic myelopathy: Results from the prospective multicenter aospine international study on 479 patients. Spine (Phila Pa 1976) 2015;40(17):1322-1328. [doi: 10.1097/BRS.0000000000000988]

XSL•FO
RenderX

11. Fehlings MG, Wilson JR, Kopjar B, et al. Efficacy and safety of surgical decompression in patients with cervical spondylotic myelopathy. JBJS 2013;95(18):1651-1658. [doi: 10.2106/JBJS.L.00589]
12. Wilson JR, Tetreault LA, Kim J, et al. State of the art in degenerative cervical myelopathy: an update on current clinical evidence. Neurosurgery 2017 Mar 1;80(3S):S33-S45. [doi: 10.1093/neuros/nyw083] [Medline: 28350949]
13. Oh T, Lafage R, Lafage V, et al. Comparing quality of life in cervical spondylotic myelopathy with other chronic debilitating diseases using the short form survey 36-health survey. World Neurosurg 2017 Oct;106:699-706. [doi: 10.1016/j.wneu.2016.12.124] [Medline: 28065875]
14. Pope DH, Mowforth OD, Davies BM, Kotter MRN. Diagnostic delays lead to greater disability in degenerative cervical myelopathy and represent a health inequality. Spine (Phila Pa 1976) 2020 Mar 15;45(6):368-377. [doi: 10.1097/BRS.0000000000003305] [Medline: 31658234]
15. Lønne VV, Hara S, Gulati S, et al. Return to work after surgery for degenerative cervical myelopathy: a nationwide registry-based observational study. Acta Neurochir (Wien) 2023 Mar;165(3):779-787. [doi: 10.1007/s00701-023-05521-w] [Medline: 36795223]
16. Davies B, Matters SIM. Series 2 ep 1 - poverty & myelopathy. : Mylopathy matters URL: https://soundcloud.com/myelopathymatters/s2e1-poverty-myelopathy [accessed 2025-02-04]
17. Davies BM, Phillips R, Clarke D, et al. Establishing the socio-economic impact of degenerative cervical myelopathy Is fundamental to improving outcomes [AO Spine RECODE-DCM Research Priority Number 8]. Global Spine J 2022 Feb;12:122S-129S. [doi: 10.1177/21925682211039835] [Medline: 35174730]
18. Rethorn ZD, Cook CE, Park C, et al. Social risk factors predicting outcomes of cervical myelopathy surgery. J Neurosurg Spine 2022 Jul 1;37(1):41-48. [doi: 10.3171/2021.12.SPINE21874] [Medline: 35090132]
19. Sangeorzan I, Andriopoulou P, Davies BM, McNair A. The information needs of people with degenerative cervical myelopathy: A qualitative study to inform patient education in clinical practice. PLoS ONE 2023;18(5):e0285334. [doi: 10.1371/journal.pone.0285334] [Medline: 37205664]
20. Davies BM, Khan DZ, Barzangi K, et al. We choose to call it “Degenerative Cervical Myelopathy”: findings of AO Spine RECODE-DCM, an international and multi-stakeholder partnership to agree a standard unifying term and definition for a disease. Global Spine J 2024 Mar;14(2):503-512. [doi: 10.1177/21925682221111780] [Medline: 35769029]
21. Davies BM, Mowforth O, Wood H, et al. Improving awareness could transform outcomes in degenerative cervical myelopathy [AO Spine RECODE-DCM Research Priority Number 1]. Global Spine J 2022 Feb;12(1_suppl):28S-38S. [doi: 10.1177/21925682211050927] [Medline: 35174734]
22. HARC Hastings Advice and Representation Centre. URL: https://harcuk.com [accessed 2025-02-04]
23. Citizens advice. 2023. URL: https://www.citizensadvice.org.uk/benefits [accessed 2025-02-04]
24. Welfare benefits unit. 2023. URL: http://www.welfare-benefits-unit.org.uk [accessed 2025-02-04]
25. Can I get financial support or benefits if I have a heart condition? British Heart Foundation. URL: https://www.bhf.org.uk/informationsupport/support/practical-support/can-i-get-financial-support-or-benefits-if-i-have-a-heart-condition [accessed 2025-02-06]
26. Benefits workshops. Age Scotland. URL: https://www.agescotland.org.uk/information-advice/information-workshops/benefits-workshops [accessed 2025-02-06]
27. The national lottery community fund. URL: https://www.tnlcommunityfund.org.uk/ [accessed 2025-02-06]
28. Line A, insurance N, pension S, pension AS state pension. How and when you can claim. URL: https://www.gov.uk/universal-credit [accessed 2025-02-11]
29. Personal independence payment. Gov.uk. URL: https://www.gov.uk/pip [accessed 2025-02-04]
30. Davies BM, Phillips R, Clarke D, et al. Establishing the socio-economic impact of degenerative cervical myelopathy Is fundamental to improving outcomes [AO Spine RECODE-DCM Research Priority Number 8]. Global Spine J 2022 Feb;12(1_suppl):122S-129S. [doi: 10.1177/21925682211039835] [Medline: 35174730]

Abbreviations

DCM: degenerative cervical myelopathy
HARC: Hastings Advice and Representation Centre
PIP: personal independence payment
UC: universal credit

XSL•FO
RenderX

Edited by A Mavragani; submitted 31.03.24; peer-reviewed by F Khan, G Goodman; revised version received 26.11.24; accepted 15.12.24; published 24.02.25.

Please cite as:
Rujeedawa T, Karimi Z, Wood H, Sangeorzan I, Smith R, Sadler I, Martin-Moore E, Gardner A, K Demetriades A, Sinha R, Grahovac G, Bateman A, Deakin N, Davies B
Evaluation of Financial Support Workshops for Patients Under State Pension Age With Degenerative Cervical Myelopathy: Survey Study
JMIR Form Res 2025;9:e59032
URL: https://formative.jmir.org/2025/1/e59032
doi:10.2196/59032

©Tanzil Rujeedawa, Zahabiya Karimi, Helen Wood, Irina Sangeorzan, Roy Smith, Iwan Sadler, Esther Martin-Moore, Adrian Gardner, Andreas K Demetriades, Rohitashwa Sinha, Gordan Grahovac, Antony Bateman, Naomi Deakin, Benjamin Davies. Originally published in JMIR Formative Research (https://formative.jmir.org), 24.2.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Snowball Group Usability Testing for Rapid and Iterative Multisite Tool Development: Method Development Study

Katherine L Dauber-Decker[1,2], PhD; David Feldstein[3], MD; Rachel Hess[4], MS, MD; Devin Mann[5,6], MD, MS; Eun Ji Kim[1,2], MS, MSc, MD; Pranisha Gautam-Goyal[7], MD; Jeffrey Solomon[1,2], BFA; Sundas Khan[8,9], MD; Fatima Malik[1,2], MHA; Lynn Xu[5], MPH; Ainsley Huffman[10], MS; Paul D Smith[3], MD; Wendy Halm[11], DNP, FNP-BC, APNP; Alice Yuroff[3], PhD; Safiya Richardson[5,6,12], MPH, MD

[1]Feinstein Institutes for Medical Research, Northwell Health, 600 Community Drive, Suite 403, Manhasset, NY, United States
[10]Utah Clinical and Translational Science Institute, University of Utah, Salt Lake City, UT, United States
[11]University of Wisconsin School of Nursing, Madison, WI, United States
[12]Institute for Excellence in Health Equity, NYU Langone Health, New York, NY, United States
[2]Northwell Health, 2000 Marcus Avenue, Suite 300, New Hyde Park, NY, United States
[3]University of Wisconsin School of Medicine and Public Health, Madison, WI, United States
[4]University of Utah Health, Salt Lake City, UT, United States
[5]Department of Population Health, NYU Grossman School of Medicine, New York, NY, United States
[6]Department of Health Informatics, Medical Center Information Technology, NYU Langone Health, New York, NY, United States
[7]Donald and Barbara Zucker School of Medicine at Hofstra/Northwell, Manhasset, NY, United States
[8]Department of Medicine, Baylor College of Medicine, Houston, TX, United States
[9]Center for Innovations in Quality, Effectiveness and Safety, Michael E. DeBakey VA Medical Center, Houston, TX, United States

Corresponding Author:
Katherine L Dauber-Decker, PhD
Feinstein Institutes for Medical Research, Northwell Health, 600 Community Drive, Suite 403, Manhasset, NY, United States

Abstract

Background: Usability testing is valuable for assessing a new tool or system's usefulness and ease-of-use. Several established methods of usability testing exist, including think-aloud testing. Although usability testing has been shown to be crucial for successful clinical decision support (CDS) tool development, it is often difficult to conduct across multisite development projects due to its time- and labor-intensiveness, cost, and the skills required to conduct the testing.

Objective: Our objective was to develop a new method of usability testing that would enable efficient acquisition and dissemination of results among multiple sites. We sought to address the existing barriers to successfully completing usability testing during CDS tool development.

Methods: We combined individual think-aloud testing and focus groups into one session and performed sessions serially across 4 sites (snowball group usability testing) to assess the usability of two CDS tools designed for use by nurses in primary and urgent care settings. We recorded each session and took notes in a standardized format. Each site shared feedback from their individual sessions with the other sites in the study so that they could incorporate that feedback into their tools prior to their own testing sessions.

Results: The group testing and snowballing components of our new usability testing method proved to be highly beneficial. We identified 3 main benefits of snowball group usability testing. First, by interviewing several participants in a single session rather than individuals over the course of weeks, each site was able to quickly obtain their usability feedback. Second, combining the individualized think-aloud component with a focus group component in the same session helped study teams to more easily notice similarities in feedback among participants and to discuss and act upon suggestions efficiently. Third, conducting usability testing in series across sites allowed study teams to incorporate feedback based on previous sites' sessions prior to conducting their own testing.

Conclusions: Snowball group usability testing provides an efficient method of obtaining multisite feedback on newly developed tools and systems, while addressing barriers typically associated with traditional usability testing methods. This method can be applied to test a wide variety of tools, including CDS tools, prior to launch so that they can be efficiently optimized.

Trial Registration: Clinicaltrials.gov NCT04255303; https://clinicaltrials.gov/study/NCT04255303

(JMIR Form Res 2025;9:e55316) doi:10.2196/55316

XSL•FO
RenderX

KEYWORDS

clinical decision support; CDS; decision aid; clinical aid; cough; sore throat; strep pharyngitis; snowball group usability testing; snowball group; usability testing

Introduction

Usability testing assesses a new tool or system's usefulness and ease-of-use. Usability testing is particularly important in clinical decision support (CDS) development. CDS tools are evidence-based tools that help clinicians make decisions to improve patient care. Usability testing during CDS development increases the likelihood of tool adoption and impact by creating a tool that is workflow-integrated, useful, and easy to use [1-5]. The current standard for usability testing of CDS tools is think-aloud testing in which approximately 3 - 5 end users are asked to think out loud while interacting with a prototype of the tool during individual sessions.

Although the benefits of usability testing are well established, it is often not conducted during typical CDS development due to tight project timelines. Think-aloud testing with 3 - 5 users may add weeks to a project timeline. This becomes a larger challenge in multisite development projects. Typically, usability testing is done on the same tool at multiple sites in parallel, where end-users at each site are likely to report similar issues, adding weeks to the project timeline at each site with minimal added value of testing at each site. An efficient, effective method for usability testing CDS across multiple sites is needed.

In addition to think-aloud testing done in individual sessions, focus groups are also used to obtain feedback on a tool. These two types of testing can provide different types of insight from future end-users [6,7]. Focus groups include interactions between different participants, which can lead to the development of consolidated ideas through conversation. However, a single vocal participant can influence the direction of the discussion and may not represent all the participants' views. One-on-one interviews, in contrast, enable the research team to discuss individual-level feedback with each participant [7]. Including both types of usability testing allows for the highest yield of testing [1,4].

We sought to develop a new method of usability testing, termed snowball group usability testing, that would enable efficient, high-yield multisite testing. We applied our method to a project aimed at decreasing antibiotic overprescribing in the outpatient setting. In the United States, antibiotics are frequently prescribed inappropriately, contributing to antibiotic resistance [8-13]. The goals of our testing were to (1) determine the feasibility and practicality of snowball group usability testing and (2) facilitate the development of two useful and usable CDS tools. Here we describe our method for snowball group usability testing.

Methods

Ethical Considerations

All study activities were approved by the New York University Langone Institutional Review Board (i19-01222). Verbal informed consent was obtained from all usability testing participants. Any identifiable data collected as part of the study were stored on secure drives and only accessible to members of the study team. Participants were not compensated for partaking in this research study. This manuscript does not include any identifiable study data.

Setting and Recruitment

We conducted this study at 4 large academic sites in New York, Wisconsin, and Utah. The study teams at each site were experienced in CDS design, development, and evaluation, with a focus on using human-centered design strategies. Nurses who were eligible to participate in usability testing were identified based on their roles at study clinics and eligibility criteria. The study was presented to nurses by the study team. Eligible nurses were contacted by email for recruitment. Nurses who were interested in participating provided verbal consent and were given a key information sheet that outlined their research participation.

Description of CDS Tools

For this study, we used the Heckerling rule for cough [14] and the Centor score for sore throat [15-17] to develop CDS tools for use by nurses in an outpatient setting. We developed each of these tools for use in two different electronic health records (EHRs), Epic and Allscripts TouchWorks, due to site differences. Each tool included an in-person nurse visit section and a triage section that could be applied either over the phone or in-person. During triage, the nurse or medical assistant recorded the patient's symptoms using a tool that helped determine the risk and severity level for the patient. This was used to determine if the patient needed to be seen in person for those who were assessed over the phone. Patients who needed to be seen or were already in the office but did not have severe symptoms or other significant illnesses were deemed eligible for an in-person nurse visit. Alternatively, the patient could be sent to an emergency department or physician visit, or for those who were assessed over the phone, they could be advised that they do not need a visit and could self-treat with supportive care at home (Figure 1). The assessment questions were developed based on current site standards for patients with cough and sore throat [18].

XSL•FO
RenderX

Figure 1. Cough and sore throat clinical decision support workflow diagram. The workflow for both tools included 4 steps: (1) triage, (2) evaluation, (3) risk calculation, and (4) testing and treatment based on risk.

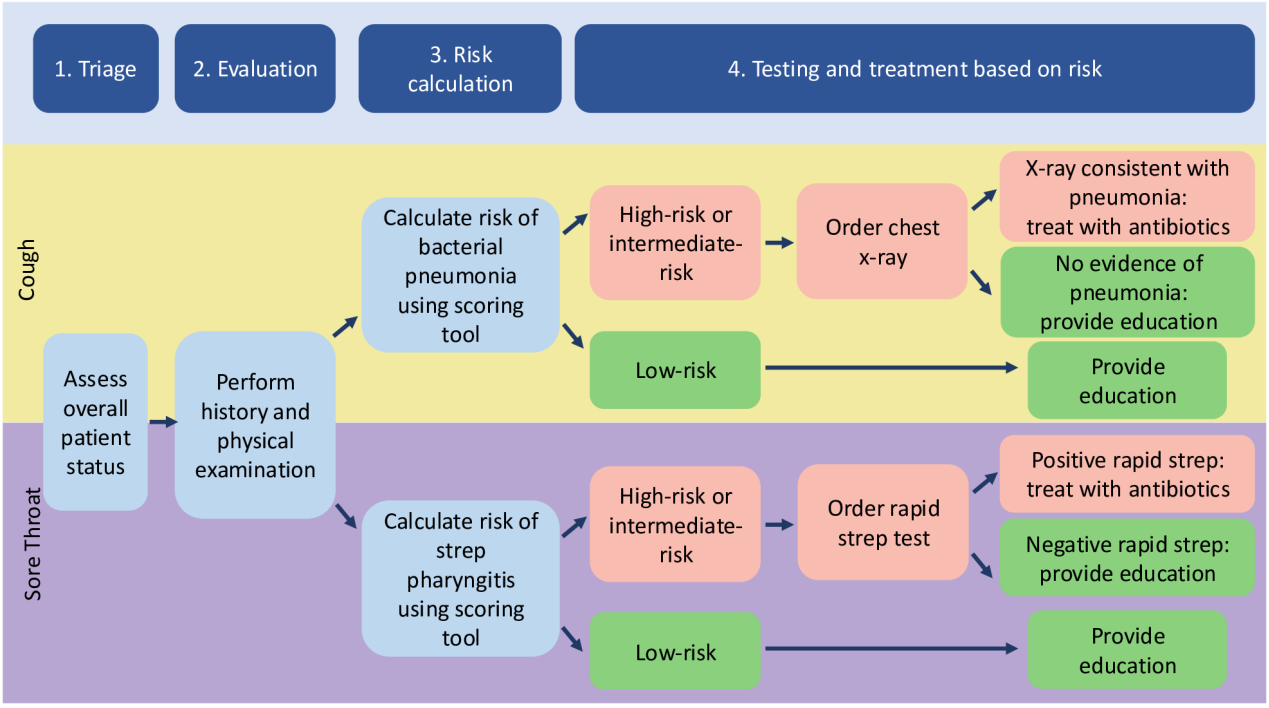

If a nurse visit was deemed appropriate for the patient, the nurse completed the visit section during an in-person encounter. The visit tools included a note for documenting history and the physical exam, a risk calculator to determine the risk of bacterial infection, and order sets based on the results of the risk calculator. Orders could be placed or pended by nurses to be signed by providers. The risk calculators were based on the Heckerling rule or Centor score for cough or sore throat, respectively, and recommendations were given based on the overall score (Table 1). Based on the number of criteria with which the patient presented, rule users could stratify patients into low-, intermediate-, or high-risk categories, indicating different levels of care.

Table . Heckerling rule and Centor score criteria.

Rule	Criteria	Low-risk	Intermediate- or high-risk
Heckerling rule (pneumonia)	• Temperature of >100°F (37.8 °C) • Heart rate of >100 beats/min • Crackles (rales) • Decreased breath sounds • No history of asthma	0 - 1: supportive care indicated	2 - 5: chest x-ray and antibiotics if x-ray was positive
Centor score (strep pharyngitis)	• History of fever: feels feverish, sudden onset of cold with shivering sweats, or oral temperature of ≥100.4°F (38.0 °C) • Absence of cough: new onset, frequent and/or persistent, different from baseline • Tonsillar exudates • Tender anterior cervical nodes	0 - 1: no rapid strep test indicated	2 - 3: rapid strep test indicated and antibiotics if the test is positive

Snowball Group Usability Testing Sessions

We combined individual think-aloud sessions and focus groups into one session. We conducted these sessions at each site serially, with the hope that each subsequent organization would reach saturation more quickly by benefitting from the previous site's insights. This form of testing will hereafter be referred to as snowball group usability testing, in which "snowballing" refers to conducting usability testing serially at each site, transferring feedback regarding the CDS tools from site to site as testing progresses, and "group" refers to conducting testing with multiple participants at once. To our knowledge, snowballing has not previously been used in the context of usability testing. We applied snowballing with the goal of increasing the efficiency at which results were obtained and passed along between study institutions. Each site performed a

XSL•FO
RenderX

minimum of one group think-aloud session for the cough tool and another for the sore throat tool with one exception. Site 4 did not complete a cough CDS testing session. Sessions were conducted with the EHR encoded tool within a sandbox environment (for the Epic EHR sites) or a Research Electronic Data Capture (REDCap) mock-up of the tool (for the Allscripts TouchWorks EHR site). Each session was structured to include an individual think-aloud component with 1 to 4 nurse participants (and in one session, a medical assistant) simultaneously, followed by a focus group component with all participants that included debriefing questions about the tool. Results were documented in a pretemplated format for rapid dissemination to the other study sites to enable the efficient integration of feedback prior to the next site's testing session. Sessions were completed remotely on Microsoft Teams, Cisco Webex, or Zoom depending on site requirements.

Study team moderators were different for each site due to Institutional Review Board regulations (ie, study team members at each site could only conduct testing with their own site's participants). The main moderator presented an overview of the session and the tools. This included instructions on how to moderate think-aloud sessions. To obtain individualized feedback, each nurse worked with an additional session moderator in an individual web-based breakout room. Here, the nurse used the think-aloud method while navigating through the tool or a prototype of the tool. Each nurse used the tool or prototype on mock clinical cases. The main moderator circulated among the breakout rooms throughout the session to answer any questions, resolve any issues, and ensure that the session ran smoothly.

Following the individual breakout room think-aloud sessions, all nurse participants were brought back together into the main meeting room for a focus group to answer debriefing questions about their opinions of the tool. All participants were encouraged to speak, and individual participants were asked for their thoughts if they were not frequently volunteering their opinions. Sites shared their results with each other between sessions so that any necessary tool modifications could be made prior to the next testing session, enabling rapid iteration of the tools. Sites requiring more feedback following their initial sessions completed an additional session.

Snowballing Approach to Method Development

Prior to initiating our usability testing sessions, we assembled a reporting form template in PowerPoint that we could quickly fill in with structured feedback from our sessions. As each site ran their testing session, and moderators filled out the template with session results and shared it with the other participating sites. This enabled the other sites to make changes as needed before conducting their own usability testing sessions. In addition, the PowerPoint template included space for sites to share information about how to iteratively adjust the methods to efficiently run the sessions based on their experiences. Each site highlighted any challenges they had with completing the sessions and suggestions for how to avoid such challenges in future sessions, allowing sites to implement changes as needed. With the goal of rapidly iterating on our tools among sites, we did not use traditional in-depth thematic analysis in favor of quick feedback summaries. Our informal thematic analysis enabled us to quickly resolve the most easily addressable issues between sessions.

Results

A summary of our session durations and participant breakdowns is provided in Table 2.

XSL•FO
RenderX

Table . Time and participant breakdown of snowball group usability testing sessions.

Variable	Average value, mean	Site 1	Site 2		Site 3	Site 4[a]
		Session 1	Session 1	Session 2[b]	Session 1	Session 1
Cough clinical decision support testing						
Participants, n	3.3	3	4	—	3	—
Facilitators, n	4.7	4	4	—	6	—
Think-aloud time (min)	35.8	30	30 - 45[c]	—	40	—
Focus group time (min)	18	21	__[d]	—	15	—
Sore throat clinical decision support testing						
Participants, n	3	4	2	1	3	5
Facilitators, n	4	4	4	3	6	3
Think-aloud time (min)	36.4	27	40	45	30	40
Focus group time (min)	14.1	24	__[d]	5	17.5	10

[a]Site 4 did not complete usability testing sessions for the cough tool.

[b]Site 2 completed 2 sessions to test the sore throat tool.

[c]This range represents differences in the amount of time participants were able to take part in the session.

[d]This session did not include a focus group component.

Cough and Sore Throat CDS Tool Feedback

Snowball group usability testing enabled us to efficiently obtain and incorporate iterative feedback across our study sites (Figure 2). Overall, participants felt that the CDS tools were useful but identified some areas for potential improvement. Overarching themes from the sessions included suggestions for changes in the tool's wording and formatting, as well as changes to the overall workflow. For example, participants identified the wording of "recent travel" as unclear, as they were unsure what timeframe would constitute recent. Similarly, "disease exposure" contained ambiguity as to which diseases the exposure could encompass. In terms of workflow, participants at one site pointed out that it would not make sense for the tool to ask about COVID results during the visit, as this would have been a prerequisite to the patient coming in for a visit at the time this study was conducted. Each of these suggestions provided valuable feedback for modifications to be made prior to the next testing session.

XSL•FO
RenderX

Figure 2. Snowball group usability testing enabled efficient acquisition of iterative feedback across study sites. (**A**) Snowballing enabled iterative integration of feedback between testing sessions. Circles indicate feedback obtained at during each site's testing sessions. Snowballing is indicated by the increasing size of the circles from site to site. (**B**) Each session included both think-aloud testing and focus group components except where indicated in Table 2.

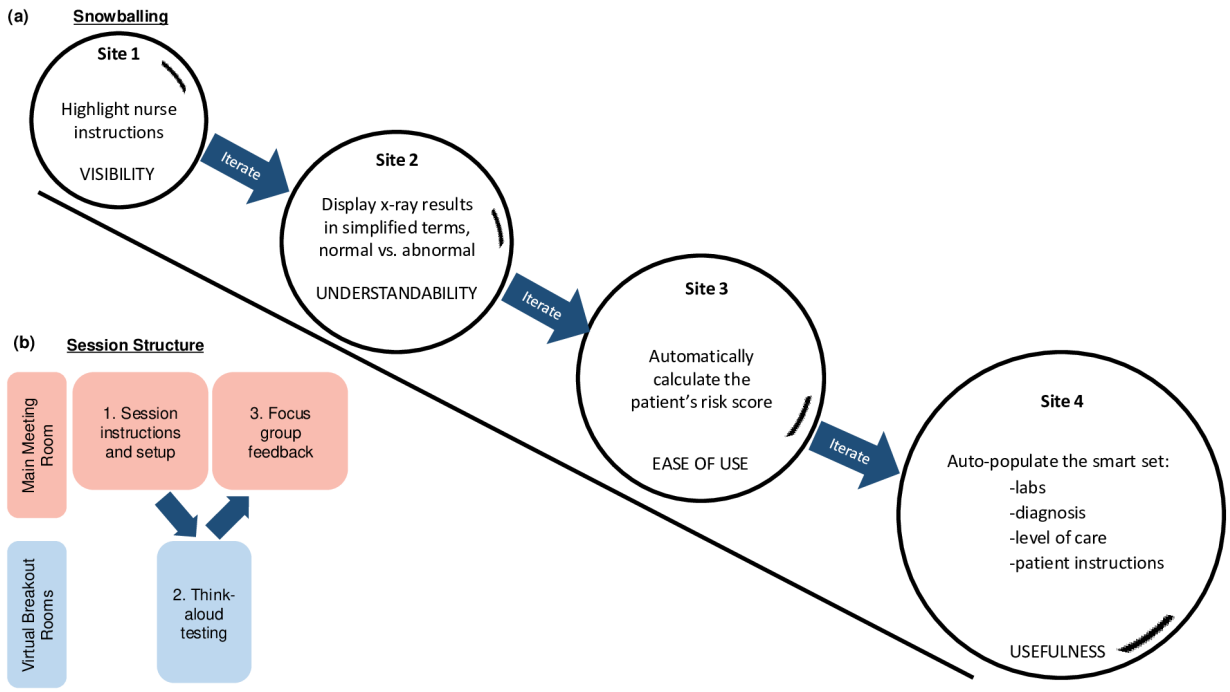

Focus Group Testing Component

The focus group testing component of our new method proved to be highly beneficial. We asked debriefing questions regarding participants' opinions of the tool in a focus group setting. One key feature was that, by collecting feedback from multiple participants at once rather than over a series of multiple separate sessions, we were more easily able to notice similarities in feedback among participants. In addition, each site's study team was able to meet soon after the single group session, rather than once after several sessions had been completed. This made it easier to discuss results, find similarities in participant feedback, and generalize, due to the recency of the information being obtained.

Snowballing Component

The snowballing component of our method also proved to be beneficial. A key benefit of snowballing was that study teams were able to incorporate changes based on previous sites' sessions. For example, nurses at one site did not properly follow instructions for the order of how to choose antibiotics based on patient allergies. The study team made changes to the cough and sore throat CDS tools at this site to highlight these instructions. These changes were also incorporated into the CDS tools under development at the 2 other sites that used Epic EHRs in the study. In addition, during the first site's testing session, it was determined that the location of the patient comorbidities, used to determine if a nurse visit was appropriate during triage, was not optimal. Comorbidity information was moved to below the nurse visit symptoms to improve ease of use. This feedback was incorporated into the second site's tool prior to their usability testing session.

While snowballing was useful for identifying and implementing certain changes to the CDS tools among sites, differences in the 2 EHRs and their respective limitations precluded some tool components from being transferable between sites. For example, while the Epic EHR versions of the tools were able to automatically calculate results, the TouchWorks EHR versions lacked this capability due to system limitations. In addition, a key feature of the Epic CDS tool was that it could link to order sets, enabling the user to seamlessly proceed from the CDS to the appropriate orders for a patient. TouchWorks did not support connection to order sets, and therefore, the transition from the CDS to ordering in the TouchWorks versions of the CDS tools was more manual. This, unfortunately, could not be changed. Finally, differences in primary care and urgent care workflows precluded uniformity in our approaches in these two environments.

Discussion

Principal Findings

Snowball group usability testing is a novel method that can be used to rapidly and iteratively test the usability of a new CDS tool or workflow. To our knowledge, this is the first time usability testing has been performed in this manner. We identified 3 main benefits of snowball group usability testing. First, by interviewing several participants in a single session rather than individuals over the course of weeks, each site was able to rapidly obtain a large amount of data at once. With traditional testing approaches, in which one participant is interviewed at a time and each must be scheduled separately, it may take weeks or even months to obtain the same amount of data. Importantly, snowball group usability testing helped to

XSL•FO
RenderX

expedite the overall usability testing timeline. By using our method sequentially (or in limited cases, concurrently) and at different sites, we were able to rapidly make changes, then retest our CDS tools with the next group of participants.

Second, combining the individualized think-aloud component with a focus group component in the same session helped us obtain a variety of feedback. In a focus group, one participant's feedback can become amplified, as a more talkative participant may dominate the conversations, and others may simply agree with their thoughts. Similar to any other focus group, session moderators encouraged all nurses in the group to contribute to the conversation, calling out individuals if they had not provided feedback after a while. By combining methodologies, we first obtained in-depth individual-level feedback with the think-aloud sessions and then elicited group-based feedback with the focus groups when participants built upon each other's opinions. Both types of feedback were useful; in the think-aloud sessions, the study team mostly observed, whereas in the focus group, the study team was more involved in asking for feedback on how to improve the tool. Therefore, combining approaches enabled us to get different types of feedback from the same set of participants all at once, which was invaluable when considering tight research and development timelines.

Third, our method enabled us to identify new and varied feedback from each site because we were able to edit key issues with our tools between sessions. Rather than hearing the same feedback in each session and making all of the changes at the end, our method resulted in a greater variety of feedback and a greater ability to optimize the tools.

Expediting the timeline and maximizing the amount of feedback gained are particularly valuable when developing EHR tools due to the inherent time constraints in the development process. Working with multiple sites and internal teams, such as those in charge of coding and implementing changes to the EHR, poses its own timing challenges as a result of the need to coordinate schedules and workloads. Thus, usability testing is often not considered a top priority; however, it remains an essential component of tool and system development. The use of snowball group usability testing effectively addresses this problem by reducing the time it takes to obtain and share feedback among participating sites. We also performed our testing remotely, which further addressed the issue of tight timelines and schedules, as participants did not have to travel from their clinics to be in the same location as other participants and the study team for the testing sessions. This enabled participants from different clinics to participate at the same time, without taking extra time out of their busy schedules for travel. Of note, remote testing can be used to mitigate scheduling issues with other forms of usability testing, including single-participant think-aloud sessions, interviews, and focus groups.

Limitations

There were challenges and limitations with snowball group usability testing. We experienced some incompatibility between different EHR systems, limiting transferability. Although the content of the CDS tools was transferable, the overall structure of the note and workflow were quite different. In addition, there were some components of the tools (such as the automatic calculator and linking to order sets) that were possible in one system but not the other, requiring site-specific modifications and testing.

In addition, timing, staffing, and availability issues presented a challenge. There were variations in each site's experience with usability testing. Finding 4 - 5 experienced usability testing moderators for a given session was challenging for some sites. As a result, staff needed to be trained in the methodology, but identifying times when staff were available was challenging. Another site experienced research staff turnover at the beginning of the usability testing phase of our study, leading to a need for extra training and reshuffling of resources. Additionally, nurses participating in the sessions had limited time available to complete their sessions and the study teams were unable to get through all of their planned cases and questions. For example, in one session, only think-aloud testing was completed due to time constraints, as the nurses had competing clinical demands. In the same session, one of the nurse participants was frequently interrupted by medical staff and phone calls. Two sites cut the think-aloud session short, setting a time limit to ensure that the focus group component was included. Additional challenges with time included accounting for the different speeds at which participants gave feedback and a lack of opportunity for the study team to iteratively improve the timing issues from session to session, as most sites only conducted one. However, learning from earlier sessions, we were able to change future sessions so that each nurse started off with a different mock case during their respective think-aloud sessions. This ensured that each case was tested, even though each nurse did not complete every case themselves.

Finally, an additional limitation was the fact that, due to the pragmatic nature of our study, we did not perform highly rigorous qualitative or quantitative evaluations of our usability testing results. Future testing of this method should include validated usability measures, such as the System Usability Scale [19], in all rounds of testing, as well as qualitative thematic analysis to rigorously assess the impact of snowball group usability testing on the usability of the CDS tools being developed. However, based on our collective years of experience performing usability testing, the suggestions and improvements derived of from snowball group usability testing were on par with what we would expect to obtain from established usability testing techniques, and the quality of our findings was higher than what we have seen using previous testing methods.

Conclusion

Snowball group usability testing provides a novel, efficient method of obtaining feedback on newly developed tools and systems, while addressing barriers typically associated with traditional usability testing methods. We successfully developed and used this method to test two CDS tools and rapidly iterated the process among 4 sites. Snowball group usability testing can be applied to test a wide variety of tools and workflows, including CDS tools, prior to launch so that they can be efficiently optimized, ultimately leading to higher adoption and end-user satisfaction.

XSL•FO
RenderX

Acknowledgments

This work was supported by grants 2R01AI108680-07A1 from the National Institute of Allergy and Infectious Diseases and K23HL145114 from the National Heart, Lung, and Blood Institute. Generative artificial intelligence was not used in any portion of the manuscript writing or for any part of this study.

Data Availability

Data for this study are limited to session recordings and notes. Deidentified usability testing data will be provided upon request.

Conflicts of Interest

KLDD's spouse works at the NYU IRB. The remaining authors declare no conflict of interest.

References

1. Kushniruk AW, Patel VL. Cognitive and usability engineering methods for the evaluation of clinical information systems. J Biomed Inform 2004 Feb;37(1):56-76. [doi: 10.1016/j.jbi.2004.01.003] [Medline: 15016386]
2. Richardson S, Mishuris R, O'Connell A, et al. "Think aloud" and "near live" usability testing of two complex clinical decision support tools. Int J Med Inform 2017 Oct;106:1-8. [doi: 10.1016/j.ijmedinf.2017.06.003] [Medline: 28870378]
3. Press A, McCullagh L, Khan S, Schachter A, Pardo S, McGinn T. Usability testing of a complex clinical decision support tool in the emergency department: lessons learned. JMIR Hum Factors 2015 Sep 10;2(2):e14. [doi: 10.2196/humanfactors.4537] [Medline: 27025540]
4. Li AC, Kannry JL, Kushniruk A, et al. Integrating usability testing and think-aloud protocol analysis with "near-live" clinical simulations in evaluating clinical decision support. Int J Med Inform 2012 Nov;81(11):761-772. [doi: 10.1016/j.ijmedinf.2012.02.009] [Medline: 22456088]
5. Dauber-Decker KL, Basile M, King D, et al. Developing a decision aid to facilitate informed decision making about invasive mechanical ventilation and lung transplantation among adults with cystic fibrosis: usability testing. JMIR Hum Factors 2021 Apr 14;8(2):e21270. [doi: 10.2196/21270] [Medline: 33851921]
6. Gill P, Stewart K, Treasure E, Chadwick B. Methods of data collection in qualitative research: interviews and focus groups. Br Dent J 2008 Mar 22;204(6):291-295. [doi: 10.1038/bdj.2008.192] [Medline: 18356873]
7. Rosenthal M. Qualitative research methods: why, when, and how to conduct interviews and focus groups in pharmacy research. Curr Pharm Teach Learn 2016 Jul;8(4):509-516. [doi: 10.1016/j.cptl.2016.03.021]
8. Antibiotic resistance threats in the United States, 2019. : Centers for Disease Control and Prevention; 2019 Nov. [doi: 10.15620/cdc:82532]
9. Dadgostar P. Antimicrobial resistance: implications and costs. Infect Drug Resist 2019;12:3903-3910. [doi: 10.2147/IDR.S234610] [Medline: 31908502]
10. Bush K, Courvalin P, Dantas G, et al. Tackling antibiotic resistance. Nat Rev Microbiol 2011 Nov 2;9(12):894-896. [doi: 10.1038/nrmicro2693] [Medline: 22048738]
11. Nadimpalli ML, Chan CW, Doron S. Antibiotic resistance: a call to action to prevent the next epidemic of inequality. N Med 2021 Feb;27(2):187-188. [doi: 10.1038/s41591-020-01201-9] [Medline: 33462445]
12. Fleming-Dutra KE, Hersh AL, Shapiro DJ, et al. Prevalence of inappropriate antibiotic prescriptions among US ambulatory care visits, 2010-2011. JAMA 2016 May 3;315(17):1864-1873. [doi: 10.1001/jama.2016.4151] [Medline: 27139059]
13. Havers F, Thaker S, Clippard JR, et al. Use of influenza antiviral agents by ambulatory care clinicians during the 2012-2013 influenza season. Clin Infect Dis 2014 Sep 15;59(6):774-782. [doi: 10.1093/cid/ciu422] [Medline: 25034419]
14. Heckerling PS, Tape TG, Wigton RS, et al. Clinical prediction rule for pulmonary infiltrates. Ann Intern Med 1990 Nov 1;113(9):664-670. [doi: 10.7326/0003-4819-113-9-664] [Medline: 2221647]
15. Fine AM, Nizet V, Mandl KD. Large-scale validation of the Centor and McIsaac scores to predict group A streptococcal pharyngitis. Arch Intern Med 2012 Jun 11;172(11):847-852. [doi: 10.1001/archinternmed.2012.950] [Medline: 22566485]
16. McIsaac WJ, White D, Tannenbaum D, Low DE. A clinical score to reduce unnecessary antibiotic use in patients with sore throat. CMAJ 1998 Jan 13;158(1):75-83. [Medline: 9475915]
17. Centor RM, Witherspoon JM, Dalton HP, Brody CE, Link K. The diagnosis of strep throat in adults in the emergency room. Med Decis Making 1981;1(3):239-246. [doi: 10.1177/0272989X8100100304] [Medline: 6763125]
18. Briggs J. In: Magee SW, McAvey MM, Conville T, editors. Telephone Triage Protocols for Nursing, 5th edition: Wolters Kluwer Health; 2015.
19. Brooke J. SUS: A "quick and dirty" usability scale. In: Jordan PW, Thomas B, Weerdmeester BA, McClelland IL, editors. Usability Evaluation in Industry: Taylor & Francis; 1996.

XSL•FO
RenderX

Abbreviations

CDS: clinical decision support
EHR: electronic health record
REDCap: Research Electronic Data Capture

Edited by A Mavragani; submitted 11.12.23; peer-reviewed by D Chrimes, J Brooke; revised version received 29.10.24; accepted 30.10.24; published 18.02.25.

Please cite as:
Dauber-Decker KL, Feldstein D, Hess R, Mann D, Kim EJ, Gautam-Goyal P, Solomon J, Khan S, Malik F, Xu L, Huffman A, Smith PD, Halm W, Yuroff A, Richardson S
Snowball Group Usability Testing for Rapid and Iterative Multisite Tool Development: Method Development Study
JMIR Form Res 2025;9:e55316
URL: https://formative.jmir.org/2025/1/e55316
doi:10.2196/55316

© Katherine L Dauber-Decker, David Feldstein, Rachel Hess, Devin Mann, Eun Ji Kim, Pranisha Gautam-Goyal, Jeffrey Solomon, Sundas Khan, Fatima Malik, Lynn Xu, Ainsley Huffman, Paul D Smith, Wendy Halm, Alice Yuroff, Safiya Richardson. Originally published in JMIR Formative Research (https://formative.jmir.org), 18.2.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Exploring the Effect of an 8-Week AI-Composed Exercise Program on Pain Intensity and Well-Being in Patients With Spinal Pain: Retrospective Cohort Analysis

Annika Griefahn[1,2,3], MSc; Florian Avermann[3], MSc; Christoff Zalpour[2], Prof Dr Med; Robert Percy Marshall[4,5], MD; Inés Cordon Morillas, PhD; Kerstin Luedtke[1], Prof Dr

[1]Department of Physiotherapy, Institute of Health Sciences, University of Lübeck, Lübeck, Germany
[2]Faculty of Business Management and Social Sciences, Hochschule Osnabrück, Albrechtstraße 30, Osnabrück, Germany
[3]Evidence and Evaluation Department, medicalmotion GmbH, München, Germany
[4]Medical Department, RasenBallsport Leipzig GmbH, Leipzig, Germany
[5]Department of Orthopedic and Trauma Surgery, Martin Luther University Halle-Wittenberg, Halle, Germany

Corresponding Author:
Annika Griefahn, MSc
Department of Physiotherapy, Institute of Health Sciences, University of Lübeck, Lübeck, Germany

Abstract

Background: Spinal pain, one of the most common musculoskeletal disorders (MSDs), significantly impacts the quality of life due to chronic pain and disability. Physical activity has shown promise in managing spinal pain, although optimizing adherence to exercise remains a challenge. The digital development of artificial intelligence (AI)-driven applications offers a possibility for guiding and supporting patients with MSDs in their daily lives.

Objective: The trial aimed to investigate the effect of an 8-week AI-composed exercise program on pain intensity and well-being in patients with spinal pain. It also examined the relationship between exercise frequency, pain intensity, and well-being. In addition, app usage frequency was examined as a proxy for app engagement.

Methods: Data from users who met the inclusion criteria were collected retrospectively from the medicalmotion app between January 1, 2020, and June 30, 2023. The intervention involved the use of the medicalmotion app, which provides 3 - 5 personalized exercises for each session based on individual user data. The primary outcomes assessed pain intensity and well-being using the numeric rating scale (NRS) and the Likert scale. Data were collected at baseline (t0), 4 weeks (t1), and 8 weeks (t2). The correlation between exercise frequency, pain intensity, and well-being was analyzed as a secondary outcome. In addition, average session length and frequency were measured to determine app engagement. Statistical analysis included ANOVA and Spearman correlation analysis.

Results: The study included 379 participants with a mean age of 50.96 (SD 12.22) years. At t2, there was a significant reduction of 1.78 points on the NRS ($P<.001$). The score on the Likert scale for well-being improved by 3.11 points after 8 weeks. Pain intensity showed a negative correlation with the number of daily exercises performed at t1 and t2. Well-being had a small negative correlation with the average number of exercises performed per day. The average number of exercises performed per day was 3.58. The average session length was approximately 10 minutes, and the average interaction with the app was 49.2% (n=27.6 days) of the 56 available days.

Conclusions: Overall, the study demonstrates that an app-based intervention program can substantially reduce pain intensity and increase well-being in patients with spinal pain. This retrospective study showed that an app that digitizes multidisciplinary rehabilitation for the self-management of spinal pain significantly reduced user-reported pain intensity in a preselected population of app users.

Trial Registration: OSF Registries osf.io/kjhef; https://osf.io/kjhef

(JMIR Form Res 2025;9:e57826) doi:10.2196/57826

KEYWORDS

exercise; mHealth; app engagement; spinal pain; artificial intelligence; AI; intensity; well-being; mobile health; apps; applications; retrospective analysis; physical activity; adults; questionnaire

XSL•FO
RenderX

Introduction

Musculoskeletal disorders (MSDs) can affect muscles, tendons, cartilage, ligaments, and nerves [1]. MSDs are a major cause of chronic pain, physical disability, and loss of quality of life worldwide. They include more than 150 conditions and syndromes [2]. For example, one of the most common MSDs is spinal pain, and more specifically, low back pain (LBP) [2]. The lifetime prevalence of LBP is high, ranging from 75% to 85% [3]. Nonspecific LBP, characterized by the absence of a clear anatomical cause, accounts for approximately 90% of LBP cases [3]. Numerous risk factors for LBP have been identified, including age, obesity, sedentary lifestyle, poor posture, smoking, and psychosocial factors such as stress and depression [4]. These factors contribute to LBP's complex and multifactorial nature, including physical, psychological, occupational, and lifestyle influences [4].

Several studies have shown that regular physical activity is effective in reducing chronic spinal pain and is consistently recommended in clinical guidelines for the management of nonspecific LBP [5-8]. A systematic review by Jordan et al [9] assessed the effects of interventions that may improve adherence to exercise and physical activity. A total of 42 trials were included, mainly focusing on patients with knee osteoarthritis and spinal pain with relatively short follow-up measurements. Promising strategies to improve exercise adherence included supervised exercise, individualized exercise, refresher or follow-up sessions, the provision of supplementary materials such as audio- or videotapes with graded exercises, self-management programs, and cognitive behavioral techniques [9].

The Clinical Practice Guideline on Intervention for the Management of Acute and Chronic LBP suggests that specific complaints and causes should be considered when selecting exercises [10]. For example, for acute complaints involving leg pain, specific exercises to activate the trunk muscles and improve the strength as well as the endurance of the back muscles should be considered. For chronic spinal pain, exercises designed to improve movement control should be included [10].

The increasing digitalization of health care and therapeutic services allows patients to use various mobile and web applications to manage their conditions [11]. Digital solutions provide a scalable and widely accessible approach, enabling the management of spinal pain in rural areas where the availability of physiotherapists is limited [12]. Digital therapeutic care applications mostly include video-based treatment programs and educational materials [13]. Not surprisingly, the number of studies investigating digital interventions has increased in recent years, including studies on the effectiveness of digital exercises in patients with spinal pain [14-18].

Another aspect of digitalization is integrating artificial intelligence (AI) methods into digital therapy applications. AI can be used to better address the heterogeneity of patients with MSDs and to personalize therapy. The use of AI in health care, particularly to support diagnosis, has increased in recent years [19]. Few studies have investigated the effects of AI-based exercises [20-22]. In addition, maintaining an exercise routine is an important factor in reducing the risk of recurrence [7,23,24]. Nearly 70% of people who recover from spinal pain experience another episode within 12 months [25]. Most maintenance strategies involve supervised group exercise, with or without equipment. However, this can be costly for the individual and the health care system [7]. Therefore, an approach using digital health applications and AI offers a new perspective to increase individual adherence to exercise at a lower cost, which could reduce the risk of recurrent pain episodes in the future [26].

To increase knowledge and provide additional evidence on the use of digital applications, the following four research questions were defined: (1) What are the effects of an 8-week app-based AI-composed exercise program in patients with spinal pain regarding pain intensity? (2) What are the effects of an 8-week app-based AI-composed exercise program on patients with spinal pain regarding their well-being? (3) Is there a correlation between the frequency of exercise and pain intensity and well-being? (4) What are the parameters of app engagement, in terms of average session duration and frequency, for patients with spinal pain in an 8-week app-based AI-composed exercise program?

Methods

Ethical Considerations

The study was conducted in accordance with the ethical principles of the Declaration of Helsinki. All procedures and materials used in this study were approved by the ethics committee of the University of Applied Sciences Osnabrueck (HSOS/2022/1/2). The data provided by medicalmotion GmbH were anonymized and stored on a password-protected server at the University of Applied Sciences Osnabrueck. All users consented to the collection of data by agreeing to the terms and conditions of the use of medicalmotion.

Trial Design

The study is reported following the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guideline (Multimedia Appendix 1) [27]. The study was registered a priori via the Open Science Framework [28]. The study is a retrospective analysis of user data. Users were recruited to the medicalmotion app between January 1, 2020, and June 30, 2023, through various recruitment opportunities, such as health insurance companies or physiotherapists.

Participants

Inclusion criteria were as follows: (1) men and women who reported spinal pain (including upper and lower back) in the initial therapeutic questionnaire, (2) individuals aged between 18 and 65 years with access to the medicalmotion app, and (3) participants who had at least 8 interactions with the medicalmotion app during the 8-week evaluation period. Exclusion criteria were as follows: (1) persons diagnosed with a neurological or mental illness, (2) persons currently suffering from an infection or systemic disease, or (3) persons regularly taking medication for mental illness. The sample for this study consisted of all enrolled participants who met the inclusion criteria.

XSL•FO
RenderX

Intervention

The medicalmotion GmbH offers a comprehensive mobile and web application for various pain conditions. The mobile app not only offers AI-based exercises but also includes a wide range of additional features, including relaxation exercises, podcasts, a chat function, and a health cockpit, which tracks the user's condition throughout use. The AI-chosen exercises consider medical history data, including pain characteristics, lifestyle information, and well-being. In addition, user feedback from previous exercise sessions is incorporated into the composition process to ensure a personalized approach.

The system dynamically selects the most appropriate exercises based on real-time data such as the user's pain location and intensity, comfort level, and feedback from previous exercises. It creates an exercise needs profile and matches it with exercises from an extensive database. This adaptive approach allows adjustments for acute pain or changing circumstances. The user has the option to choose the number of daily exercises (3 to 5), all of which are delivered as real-time audio-based exercise videos. Each exercise has an average length of 167 seconds. The system also identifies potential causes of pain and adapts the exercise composition accordingly, ensuring a personalized exercise program with transparent traceability. Its training repertoire consists of a selection of 300 exercises. These exercises are categorized into 3 groups: release, mobility, and strength. Release exercises aim to induce muscle and tissue relaxation. Mobility exercises involve the full spectrum of body movement with complex movement patterns. Strength exercises strengthen muscles, tendons, and supporting tissues. An example of each type of exercise is included in Multimedia Appendix 2.

Outcomes

The primary outcome measure was pain intensity. An 11-point numeric rating scale (NRS) assessed participants' pain intensity. A change of 1 - 2 points was considered clinically relevant [29]. The secondary outcome was well-being, measured using an 11-point Likert scale. A score of zero indicates poor well-being, while a score of 10 indicates perfect well-being. Data were collected at baseline (t0), 4 weeks (t1), and 8 weeks (t2). In addition, the following parameters were collected at t0 via a therapeutic questionnaire as part of the onboarding within the medicalmotion app: age, work mode, sports frequency, and sex. The initial pain areas' anatomical location and duration were recorded using a body map. Pain duration was categorized into chronic, subacute, and acute. The anatomical location of pain was recorded for the buttocks, lower back, upper back, and neck areas. Multiple responses were allowed for anatomical location. Data collection included the weekly training volume, the number of training sessions with more than one exercise, the number of training sessions in which all exercises were completed, and the total number of exercises completed. The application evaluated the average length and frequency of sessions to assess engagement with the app. Data were only collected within the medicalmotion app.

Statistical Analysis

For the statistical analysis, an ANOVA with repeated measurement was conducted to examine the effects of the 8-week app usage on pain intensity and well-being. If the assumption of sphericity was not met, the Greenhouse-Geisser correction was used. For post hoc analysis, the Bonferroni correction was used.

In addition, Pearson correlation (r) between pain intensity, well-being, average number of exercises per day, total number of exercises, and skipped exercises was examined. If there was no linear correlation, Spearman correlation (ρ) was performed. The interpretation of the correlation coefficients was based on Cohen thresholds [30].

The average session length was calculated by multiplying the average number of exercises performed by the average exercise duration of 167 seconds. Session frequency was calculated by dividing the total number of active days by the total number of available days (56 days). Session frequency was expressed as a percentage. Session length was given in seconds and minutes.

The significance level was set at P=.05. The effect size was described by partial η^2 [30]. Results are reported as mean, SD, minimum, maximum, CIs, F value, and df. Statistical analysis of the data was performed using IBM SPSS Statistics version 29.

Results

Descriptive Results of the Participants

The total sample included 379 participants, of whom 138 (36.4%) were men and 241 (63.6%) were women. The mean age was 50.96 (SD 12.22) years. Data on mode of work, frequency of sport, initial pain areas, and completed exercises per day are shown in Table 1.

XSL•FO
RenderX

Table . Descriptive results of the cohort.

Variables			Values
Categorical variables			
	Participants, n (%)		
		Female	241 (63.6)
		Male	138 (36.4)
	Work mode, n (%)		
		Sitting and standing	165 (43.5)
		Sitting	177 (46.7)
		Hard work	17 (4.5)
		Standing	20 (5.3)
	Sport frequency, n (%)		
		Never	59 (15.6)
		1 - 3 times per week	248 (65.4)
		>3 times per week	72 (19)
	Initial pain areas (anatomical location), n (%)		
		Lower back	229 (60.4)
		Neck	201 (53)
		Buttocks	109 (28.8)
		Upper back	184 (48.5)
Continuous Variables			
	Age (years), mean (SD)		50.96 (12.22)
	Number of initial areas of pain based on the duration of the pain, mean (SD)		
		Total	4.86 (4.02)
		Chronic pain areas (>6 months)	2.43 (3)
		Subacute pain areas (>1 week and <6 months)	1.52 (2.18)
		Acute pain areas (<1 week)	0.91 (2.33)

Pain Intensity

At t0, the intervention group showed a mean pain intensity of 6.08 (SD 2.16) on the 0 - 10 NRS (Table 2). After 8 weeks (t2), the intervention group showed a mean reduction of 1.78 (SD 2.05) points on the 0 - 10 NRS (Figure 1). The factor "time" was significant ($F_{1.88,710.96}$=179.861, P<.001) and showed a large effect with η^2=0.32. Post hoc analysis with Bonferroni correction showed a significant difference at each time point with P<.001 (Table 2).

XSL•FO
RenderX

Table . Statistical analysis of pain intensity and well-being.

Measurement points	Mean (SD)	Mean difference (SD)		Percentage difference, %		95% CI		Min	Max	*P* value	Partial η^2[a]
		t0[b]	t1[c]	t0	t1	Lower	Upper				
Pain intensity (NRS[d])										<.001	0.32
t0	6.08 (2.16)	—[e]	—	—	—	5.85	6.3	1	10	—	—
t1	5.54 (2.23)	0.53 (1.9)	—	8.3	—	5.32	5.77	0	10	—	—
t2[f]	4.3 (2.28)	1.78 (2)	1.25 (1.6)	29.3	22.4	4.07	4.52	0	10	—	—
Well-being (Likert scale)										<.001	0.64
t0	4.95 (1.65)	—	—	—	—	4.8	5.1	2	10	—	—
t1	7.07 (1.43)	2.1 (1.8)	—	42.8	—	6.92	7.22	4	10	—	—
t2	8.06 (1.38)	3.1 (1.7)	1 (1.5)	62.8	14	7.9	8.2	4	10	—	—

[a]η^2 = effect size.

[b]t0: baseline.

[c]t1: after 4 weeks.

[d]NRS: numeric rating scale.

[e]Not applicable.

[f]t2: after 8 weeks.

Figure 1. Changes in pain intensity over time. An asterisk (*) denotes a significant difference. NRS: numeric rating scale (from 0 to 10); lower whisker: minimum; upper whisker: maximum.

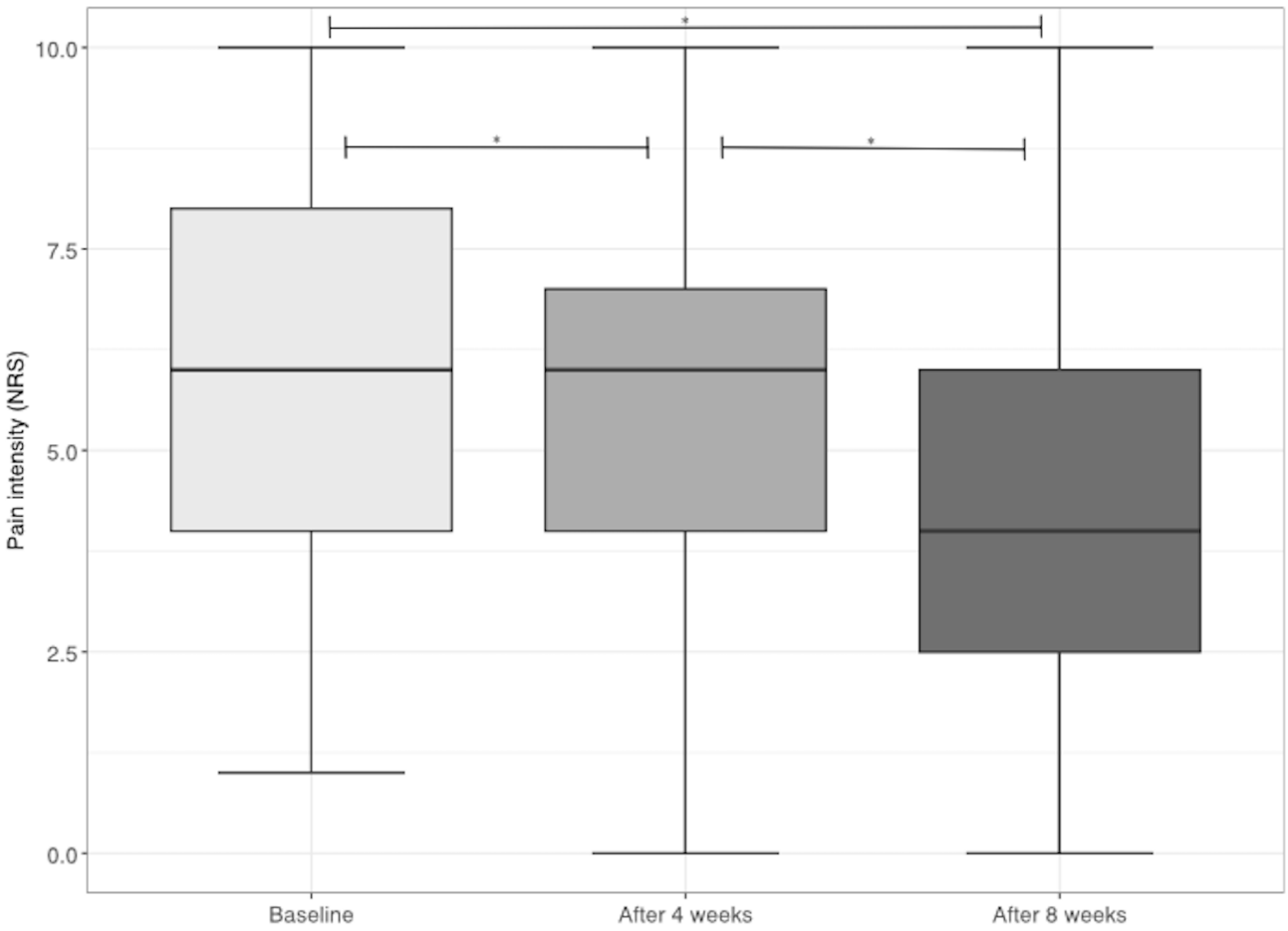

XSL•FO
RenderX

Well-Being

At t0, mean well-being was 4.95 (SD 1.65) on the 11-point Likert scale (Table 2). After 8 weeks (t2), the well-being improved by 3.11 points (Figure 2). The factor "time" was significant ($F_{1.94,732.57}$=671.97, P<.001) and showed a large effect with η^2=0.64. Post hoc analysis with Bonferroni correction showed a significant difference at each time point with P<.001 (Table 2).

Figure 2. Changes in well-being over time are shown with mean and SD. An asterisk (*) denotes a significant difference.

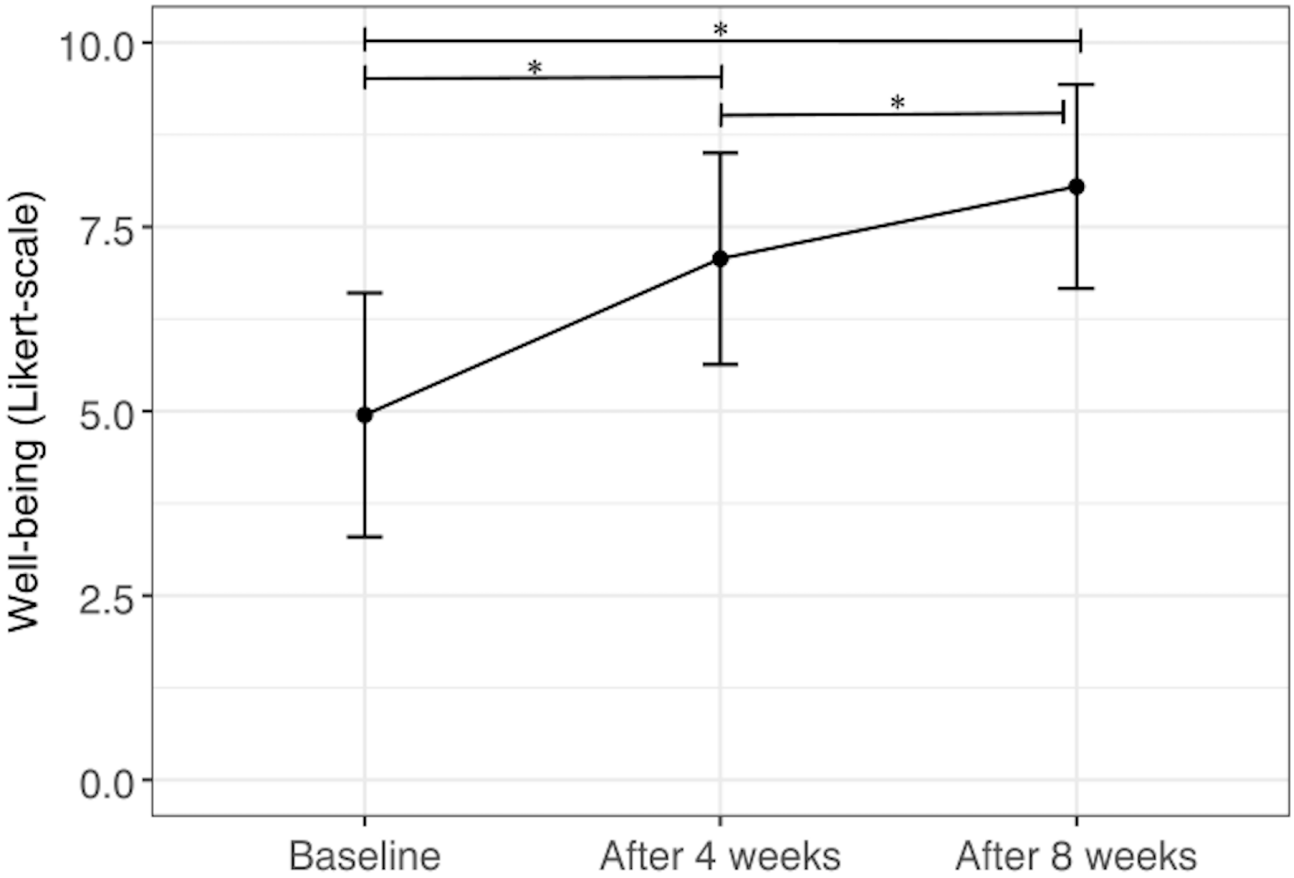

App Engagement

The average session length was 597.86 seconds (~10 minutes). The medicalmotion app was used on an average of 49.2% (27.6 days) of the 56 available days. Table 3 shows the variables: active days, total number, and average number of exercises performed.

Table . Overview of the average activity within the medicalmotion app.

Variable	Mean (SD)	Min	Max
Active days within 8 weeks trial (days)	27.54 (10.98)	16	57
Finished exercises (number)	99.34 (49.9)	1	279
Skipped exercises (number)	1.36 (4.87)	0	67
finished exercises per day (number)	3.58 (0.9)	1	5

Correlation Between Pain Intensity and Exercise Completion

Pain intensity levels showed a statistically significant negative correlation with the average number of exercises performed per day at t1 (P<.001) and t2 (P=.004). This suggests that as pain intensity increased, the number of exercises performed per day tended to decrease. Conversely, as pain intensity decreased, the number of exercises performed per day tended to increase. A similar, although slightly weaker, negative correlation was observed between pain intensity at t2 and the total number of exercises completed (P=.04). Notably, no statistically significant correlation was found between missed exercises and pain intensity. Details of the correlation coefficients and P values can be found in Multimedia Appendix 3. Figure 3 shows the relationship between the pain intensity at t1 and t2, the difference between t2 and t0, and the average number of exercises per day using a heat map.

XSL•FO
RenderX

Figure 3. Heatmap of the Spearman correlation between pain intensity and average finished exercises per day. The x-axis represents the average number of exercises per day (0 - 5), and the y-axis represents pain intensity (0 - 10). The color scale from green to yellow to red represents the percentage distribution. Green means a low percentage, yellow a medium percentage, and red a high percentage.

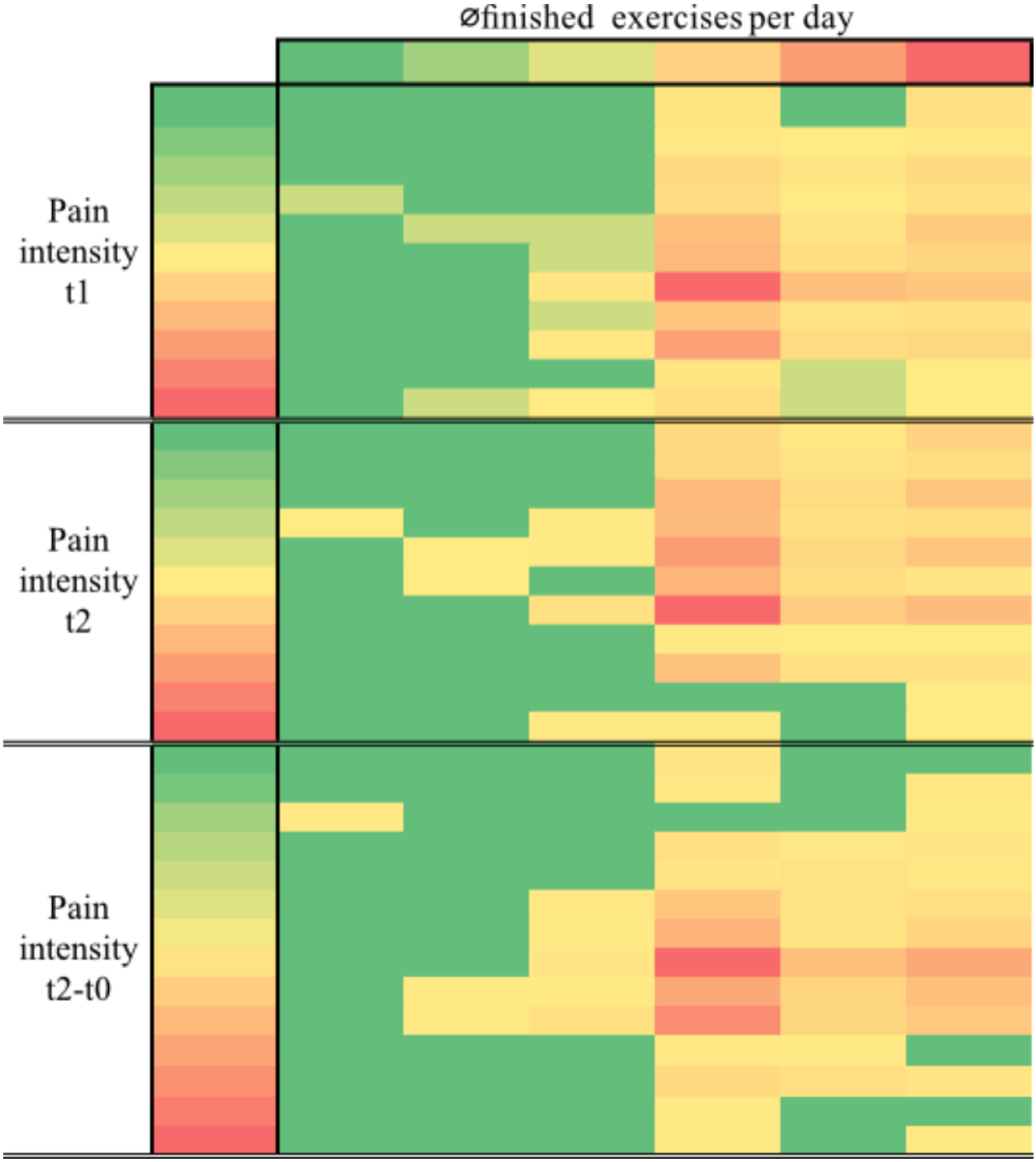

Correlation Between Well-Being and Exercise Completion

The average daily number of exercises completed showed a statistically significant, albeit small, negative correlation with the change in well-being over 8 weeks (P=.01). This means that as well-being increased, the number of exercises completed tended to decrease. Conversely, as well-being decreased, the average number of exercises completed tended to increase.

Figure 4 shows the relationship between the well-being at t1 and t2, the difference between t0 and t2, and the average number of exercises per day using a heat map.

XSL•FO
RenderX

Figure 4. Heatmap of the Spearman correlation between well-being and finished exercises per day. The x-axis represents the average number of exercises per day (0 - 5), and the y-axis represents well-being (0 - 10). The color scale from green to yellow to red represents the percentage distribution. Green means a low percentage, yellow a medium percentage, and red a high percentage.

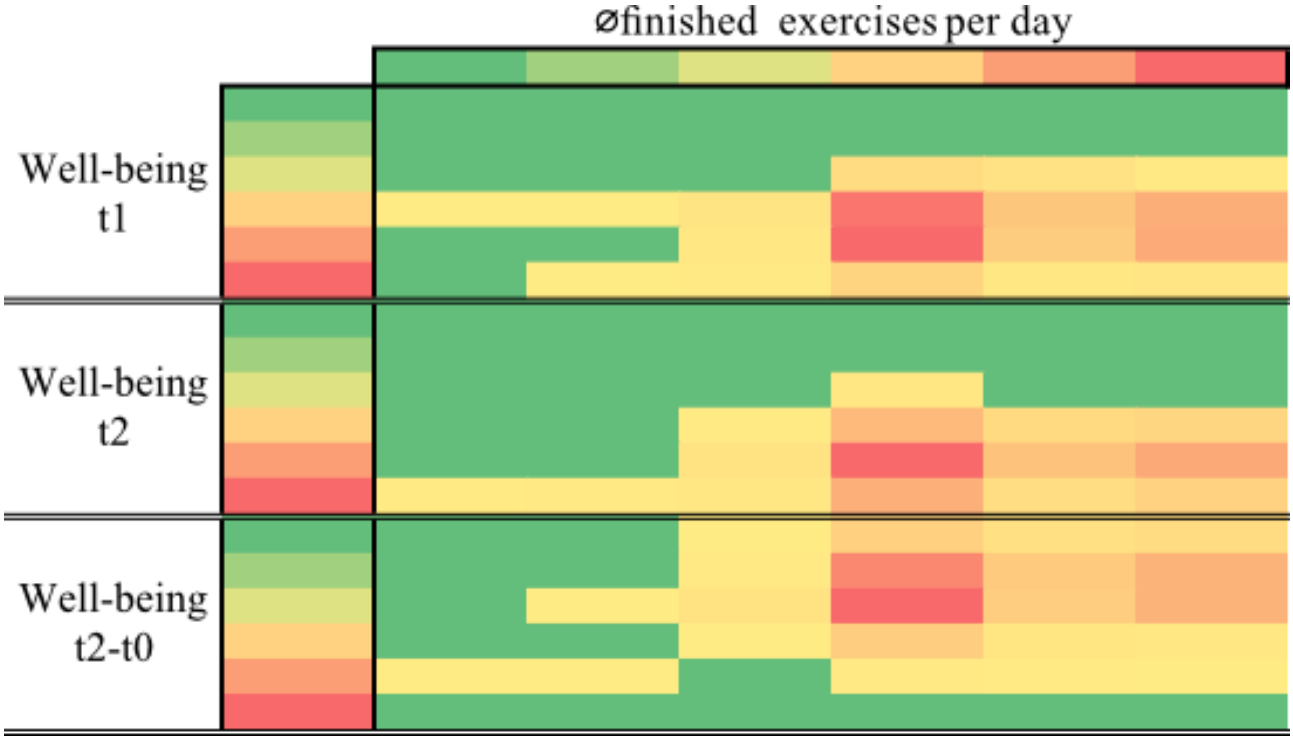

Discussion

Principal Results

This study aimed to evaluate the effect of an app-based, AI-composed exercise program on pain and well-being in patients with spinal pain. A significant effect size (η^2=0.32) was achieved with a clinically relevant reduction in pain intensity of 1.78 points. A similar picture emerges with regard to well-being. The observed change in well-being of 3.11 points represents a substantial improvement. In addition, the large effect size (η^2=0.64) indicates that a significant proportion of the variation in well-being is attributable to the intervention. The study showed significant negative correlations between pain intensity levels at t1 (P<.001) and t2 (P=.004) and the average number of daily exercises, meaning that a higher pain intensity was associated with reduced daily exercise. A weaker negative correlation was observed between pain intensity at t2 and the total number of exercises performed (P=.04). The graph shows that both too much and too little exercise did not result in a significant change in pain intensity.

In addition, the average number of exercises completed per day showed a small but significant negative correlation with the change in well-being over 8 weeks (P=.01), suggesting that increased well-being was associated with a tendency to exercise less per day. In terms of session frequency, the medicalmotion app was used on 49.2% (27.6 days) of the 56 available days. With an average exercise duration of 167 seconds, this cohort's average use time was approximately 10 minutes per 8-week session.

Limitations

While this study provides valuable insights, its limitations require careful consideration. One limitation of the AI-based app is that it categorizes the exercises into strength, mobility, and release. However, it needs to be made clear whether these categories or the individualized plan contributed significantly to the observed effects. Future studies should focus on a more detailed evaluation of the specific effects of AI on outcomes.

Pain intensity (NRS) and well-being (Likert scale) were treated as metrically scaled in the statistical analysis, allowing for the calculation of repeated measures ANOVA. Therefore, the results are presented with the mean and SD. Based on the study by Nair and Diwan [31], parametric tests can be used for ordinal scaled data (pain intensity) if a normal distribution is present. This fact should be taken into account when interpreting the results.

Because of the broad inclusion criteria used, there is variability in the participants. This variability results from the wide age range and the broad categorization of spinal pain as an inclusion criterion. As a result, the population may have included individuals with a medical diagnosis of spinal pain as well as those with no diagnosis or an inconclusive diagnosis. This diversity in the study population could affect the influence of the app on the participants. In order to increase the specificity of the results and improve the assessment of effectiveness, it may be beneficial to introduce a differentiation based on diagnostic categories.

Another limitation is the data collection period, from 2020 to 2023. The COVID-19 pandemic that was ongoing during this period, with its peak years in 2020 and 2021, may have confounded the results. During this time, the amount of physical

XSL•FO
RenderX

activity and opportunities for public exercise in gyms were limited. In addition, many people were experiencing psycho-emotional stress, which may have affected their perception of pain. Finally, the long-term effects of COVID-19 infection can cause heterogeneous symptom complexes, which may further influence the physical complaints of the participants.

A further limitation of this study is the short duration of the intervention, which was limited to 8 weeks. Such a short period of time limits the conclusions that can be drawn about the long-term effects of the intervention and the sustainability of the results. In particular, aspects such as long-term compliance and the associated changes in exercise frequency could not be adequately assessed during this short observation period. Future studies with longer follow-ups are needed to make reliable statements about the long-term benefits and stability of the intervention.

In the context of these limitations, it is important to acknowledge that this study has inherent drawbacks, including potential selection bias, lack of a control group, and reliance on self-reported data. These limitations may introduce bias and affect the generalizability of the findings. The retrospective cohort study design also has its own set of drawbacks, including potential data quality issues and variability in user recruitment. In particular, the lack of control for concomitant pain-relieving interventions, such as physiotherapy, adds complexity and requires caution in drawing definitive conclusions. Consequently, the results of the study should be used primarily for hypothesis generation, with validation sought through prospective study designs to address these limitations and provide a more comprehensive understanding of the issue.

Comparison to Prior Work

These results on pain intensity reduction can be compared with home-based exercise therapy, as shown in a review by Quentin et al [32], who demonstrated a reduction in pain intensity in nonspecific LBP and a reduction in functional impairment with home-based exercise therapy and also reported a high effect size. A study by Weise et al [33] compared the effects of digital exercise therapy with physiotherapy. This study also showed a reduction in pain intensity of 2.92 points after 6 weeks in the intervention group. The results of this study showed similar effects. However, the added value of the AI-composed exercises cannot be deduced from these results. Therefore, it cannot be concluded whether the individualized exercises provided by the mobile app have added value in reducing pain or whether simply exercising induced the effect. However, a study by Cimarras-Otal et al [34] showed that exercises tailored to the needs of workers had a better outcome than general exercises.

The interaction between well-being and chronic pain has been reported previously [35]. For example, improvements in well-being have been shown to lead to reductions in pain catastrophizing and depression [35,36]. Based on the association between exercise and increased subjective well-being, the exercises provided in this study also influenced the psychological aspects of chronic pain.

Although a correlation only indicates an association and does not imply causation, the results of the correlation analysis can be interpreted in several ways. On the one hand, a reduction in pain could be the reason why users exercised more. In other words, more exercise or movement became possible with less pain. On the other hand, increased practice could also cause pain reduction; that is, users could reduce their pain through practice. Further studies are therefore needed to establish a causal relationship and the ideal frequency of exercise. The relationship between exercise frequency and changes in pain intensity was demonstrated in a large-scale study by Marshall et al [37]. Interestingly, the frequency of exercise was very similar between participants with significant improvement and those with deterioration. This fact may be due to the inclusion of different types of exercise, making it difficult to assess the effect of specific types of exercise. Therefore, an individualized training program for each person may be necessary to improve pain intensity successfully [37].

Another factor that may influence exercise frequency may be motivation. Increased pain and decreased well-being could both increase and decrease motivation to exercise. On the one hand, if you are in pain, you need to do something about it, or conversely, if you are in pain, there may be a fear or concern that exercise will only worsen the condition. Both scenarios are plausible and require further investigation in future studies. In this context, motivation to exercise should be assessed to identify possible relationships between pain intensity and motivation.

A 1985 review by Dishman et al [38] identified several factors that influence exercise frequency. Self-motivation, anticipated personal health benefits, perceived well-being, and enjoyment of physical activity positively contributed to the likelihood of participating in a supervised exercise program. Conversely, factors such as mood disturbance, health concerns, and knowledge about exercise or health negatively influenced the likelihood. This observation underscores the importance of motivation to exercise and highlights its significant influence on the frequency of engagement in physical activity.

A study by Meyer et al [39] showed a positive association between the frequency of physical activity and self-rated health. This finding contrasts somewhat with the current results, as a negative association was found between an average number of daily exercises and a change in self-rated health over 8 weeks. One possible explanation for this discrepancy could be the distinction between well-being and self-rated health. Another explanation could be that the lower the number of complaints, the lower the compliance to continue exercising. Users may have felt better and consequently exercised less because they no longer felt the need to exercise. One possible explanation may be the influence of self-motivation on training frequency, which is already a known factor for training frequency [38]. The correlation cannot answer these considerations; hence, further research is needed to establish a causal relationship [38]. In addition, the method of assessing the frequency of physical activity may also contribute to the difference. In the study by Meyer et al [39], participants provided a self-assessment of physical activity using a questionnaire, in contrast to the objective values obtained from the app in this study. Another factor to consider for the discrepancy between Meyer et al [39] and this study's results is the study population. The

XSL•FO
RenderX

aforementioned study examined healthy participants and not a pain-specific population [39].

To date, few studies have reported on the frequency of use of eHealth products. In a study by Labinsky et al [40], 9 out of 39 participants reported having only used the applications once. A study by Tian et al [41] also found that the average usage time of common digital products was between 2 and 10 minutes. No reliable values could be found specifically for eHealth or mobile health products.

Conclusions

In summary, the study investigated the multiple effects of an 8-week, app-based, AI-composed exercise program on patients with spinal pain. It addressed critical dimensions, including pain intensity and well-being, while exploring the potential relationship between exercise frequency and these outcomes. The results highlight the importance of exercise frequency in influencing pain intensity and well-being, revealing a compelling interplay between engagement in prescribed exercises and observed improvements in health outcomes. This finding strengthens the rationale for personalized exercise programs and highlights the importance of adherence and regularity in achieving optimal outcomes.

Hypotheses that were derived from this retrospective analysis for a randomized controlled trial: (1) participants in the intervention group following an 8-week AI-composed exercise program will experience a statistically significant reduction in pain intensity and a significant improvement in well-being compared with the control group (alternative hypothesis) and (2) the frequency of engagement in the prescribed exercises will be positively correlated with decreased pain intensity and increased well-being among participants with spinal pain (correlation hypothesis).

Acknowledgments

The authors would like to thank Philipp Schlüter for his assistance. We would also like to thank medicalmotion GmbH for providing the patient examples.

Author Ines Cordon Morillas is not affiliated with any institution but is an independent scholar/freelancer who mainly supported data analysis for this study.

No funding, grants, or other support has been received.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: AG, FA, KL
Formal analysis: AG, FA, ICM
Investigation: AG, FA, KL
Methodology: AG, CZ, KL
Project administration: KL, AG
Writing – original draft: AG, FA
Writing – review & editing: AG, FA, CZ, RPM, ICM, KL

Conflicts of Interest

AG and FA are employees of medicalmotion GmbH. They report personal fees and nonfinancial support provided by medicalmotion GmbH. The remaining authors declare no competing interests.

Multimedia Appendix 1
STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) statement.
[PDF File, 123 KB - formative_v9i1e57826_app1.pdf]

Multimedia Appendix 2
Exercise examples.
[PDF File, 68 KB - formative_v9i1e57826_app2.pdf]

Multimedia Appendix 3
Exercise-induced correlation of pain intensity and well-being.
[PDF File, 111 KB - formative_v9i1e57826_app3.pdf]

References

XSL•FO
RenderX

1. Musculoskeletal health. World Health Organization. 2022 Jul 14. URL: https://www.who.int/news-room/fact-sheets/detail/musculoskeletal-conditions [accessed 2023-03-24]
2. GBD 2017 Disease and Injury Incidence and Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet 2018;392:1789-1858. [doi: 10.1016/S0140-6736(18)32279-7]
3. Wu A, March L, Zheng X, et al. Global low back pain prevalence and years lived with disability from 1990 to 2017: estimates from the Global Burden of Disease Study 2017. Ann Transl Med 2020 Mar;8(6):299. [doi: 10.21037/atm.2020.02.175] [Medline: 32355743]
4. Knezevic NN, Candido KD, Vlaeyen JWS, Van Zundert J, Cohen SP. Low back pain. Lancet 2021 Jul 3;398(10294):78-92. [doi: 10.1016/S0140-6736(21)00733-9] [Medline: 34115979]
5. Hayden JA, van Tulder MW, Malmivaara A, Koes BW. Exercise therapy for treatment of non-specific low back pain. Cochrane Database Syst Rev 2005 Jul 20;2005(3):CD000335. [doi: 10.1002/14651858.CD000335.pub2] [Medline: 16034851]
6. Geneen LJ, Moore RA, Clarke C, Martin D, Colvin LA, Smith BH. Physical activity and exercise for chronic pain in adults: an overview of Cochrane Reviews. Cochrane Database Syst Rev 2017 Jan 14;1(1):CD011279. [doi: 10.1002/14651858.CD011279.pub2] [Medline: 28087891]
7. Steffens D, Maher CG, Pereira LSM, et al. Prevention of low back pain: a systematic review and meta-analysis. JAMA Intern Med 2016 Feb;176(2):199-208. [doi: 10.1001/jamainternmed.2015.7431] [Medline: 26752509]
8. Oliveira CB, Maher CG, Pinto RZ, et al. Clinical practice guidelines for the management of non-specific low back pain in primary care: an updated overview. Eur Spine J 2018 Nov;27(11):2791-2803. [doi: 10.1007/s00586-018-5673-2] [Medline: 29971708]
9. Jordan JL, Holden MA, Mason EE, Foster NE. Interventions to improve adherence to exercise for chronic musculoskeletal pain in adults. Cochrane Database Syst Rev 2010 Jan 20;2010(1):CD005956. [doi: 10.1002/14651858.CD005956.pub2] [Medline: 20091582]
10. George SZ, Fritz JM, Silfies SP, et al. Interventions for the management of acute and chronic low back pain: revision 2021. J Orthop Sports Phys Ther 2021 Nov;51(11):CPG1-CPG60. [doi: 10.2519/jospt.2021.0304] [Medline: 34719942]
11. Hewitt S, Sephton R, Yeowell G. The effectiveness of digital health interventions in the management of musculoskeletal conditions: systematic literature review. J Med Internet Res 2020 Jun 5;22(6):e15617. [doi: 10.2196/15617] [Medline: 32501277]
12. Hou J, Yang R, Yang Y, et al. The effectiveness and safety of utilizing mobile phone-based programs for rehabilitation after lumbar spinal surgery: multicenter, prospective randomized controlled trial. JMIR Mhealth Uhealth 2019 Feb 20;7(2):e10201. [doi: 10.2196/10201] [Medline: 30785406]
13. Lewkowicz D, Slosarek T, Wernicke S, Winne A, Wohlbrandt AM, Bottinger E. Digital therapeutic care and decision support interventions for people with low back pain: systematic review. JMIR Rehabil Assist Technol 2021 Nov 19;8(4):e26612. [doi: 10.2196/26612] [Medline: 34807837]
14. Machado GC, Pinheiro MB, Lee H, et al. Smartphone apps for the self-management of low back pain: a systematic review. Best Pract Res Clin Rheumatol 2016 Dec;30(6):1098-1109. [doi: 10.1016/j.berh.2017.04.002] [Medline: 29103552]
15. Priebe JA, Haas KK, Moreno Sanchez LF, et al. Digital treatment of back pain versus standard of care: the cluster-randomized controlled trial, Rise-uP. J Pain Res 2020;13:1823-1838. [doi: 10.2147/JPR.S260761] [Medline: 32765057]
16. Bailey JF, Agarwal V, Zheng P, et al. Digital care for chronic musculoskeletal pain: 10,000 participant longitudinal cohort study. J Med Internet Res 2020 May 11;22(5):e18250. [doi: 10.2196/18250] [Medline: 32208358]
17. Lara-Palomo IC, Antequera-Soler E, Matarán-Peñarrocha GA, et al. Comparison of the effectiveness of an e-health program versus a home rehabilitation program in patients with chronic low back pain: a double blind randomized controlled trial. Digit Health 2022;8:20552076221074482. [doi: 10.1177/20552076221074482] [Medline: 35111332]
18. Özden F, Sarı Z, Karaman Ö, Aydoğmuş H. The effect of video exercise-based telerehabilitation on clinical outcomes, expectation, satisfaction, and motivation in patients with chronic low back pain. Ir J Med Sci 2022 Jun;191(3):1229-1239. [doi: 10.1007/s11845-021-02727-8] [Medline: 34357527]
19. Alowais SA, Alghamdi SS, Alsuhebany N, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ 2023 Sep 22;23(1):689. [doi: 10.1186/s12909-023-04698-z] [Medline: 37740191]
20. Rasmussen CDN, Sandal LF, Holtermann A, Stochkendahl MJ, Mork PJ, Søgaard K. Effect of a smartphone self-management digital support system for low-back pain (selfBACK) among workers with high physical work demands: secondary analysis of a randomized controlled trial. Scand J Work Environ Health 2024 Dec 1;50(8):613-621. [doi: 10.5271/sjweh.4186] [Medline: 39264347]
21. Marcuzzi A, Klevanger NE, Aasdahl L, et al. An artificial intelligence-based app for self-management of low back and neck pain in specialist care: process evaluation from a randomized clinical trial. JMIR Hum Factors 2024 Jul 9;11:e55716. [doi: 10.2196/55716] [Medline: 38980710]
22. Hartmann R, Avermann F, Zalpour C, Griefahn A. Impact of an AI app-based exercise program for people with low back pain compared to standard care: a longitudinal cohort-study. Health Sci Rep 2023 Jan;6(1):e1060. [doi: 10.1002/hsr2.1060] [Medline: 36660258]

XSL•FO
RenderX

23. de Campos TF, Maher CG, Fuller JT, Steffens D, Attwell S, Hancock MJ. Prevention strategies to reduce future impact of low back pain: a systematic review and meta-analysis. Br J Sports Med 2021 May;55(9):468-476. [doi: 10.1136/bjsports-2019-101436] [Medline: 32646887]
24. Huang R, Ning J, Chuter VH, et al. Exercise alone and exercise combined with education both prevent episodes of low back pain and related absenteeism: systematic review and network meta-analysis of randomised controlled trials (RCTs) aimed at preventing back pain. Br J Sports Med 2020 Jul;54(13):766-770. [doi: 10.1136/bjsports-2018-100035] [Medline: 31672696]
25. da Silva T, Mills K, Brown BT, et al. Recurrence of low back pain is common: a prospective inception cohort study. J Physiother 2019 Jul;65(3):159-165. [doi: 10.1016/j.jphys.2019.04.010] [Medline: 31208917]
26. Pocovi NC, Lin CWC, French SD, et al. Effectiveness and cost-effectiveness of an individualised, progressive walking and education intervention for the prevention of low back pain recurrence in Australia (WalkBack): a randomised controlled trial. Lancet 2024 Jul 13;404(10448):134-144. [doi: 10.1016/S0140-6736(24)00755-4] [Medline: 38908392]
27. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol 2008 Apr;61(4):344-349. [doi: 10.1016/j.jclinepi.2007.11.008] [Medline: 18313558]
28. AI-based exercise recommendations for people with back pain: a retrospective cohort analysis. OSF Registries. URL: https://osf.io/kjhef [accessed 2025-02-03]
29. Salaffi F, Stancati A, Silvestri CA, Ciapetti A, Grassi W. Minimal clinically important changes in chronic musculoskeletal pain intensity measured on a numerical rating scale. Eur J Pain 2004 Aug;8(4):283-291. [doi: 10.1016/j.ejpain.2003.09.004] [Medline: 15207508]
30. Cohen J. Statistical Power Analysis for the Behavioral Sciences, 2nd edition: Routledge; 2013.
31. Nair AS, Diwan S. Pain scores and statistical analysis—the conundrum. Ain-Shams J Anesthesiol 2020 Dec;12(1). [doi: 10.1186/s42077-020-00085-8]
32. Quentin C, Bagheri R, Ugbolue UC, et al. Effect of home exercise training in patients with nonspecific low-back pain: a systematic review and meta-analysis. Int J Environ Res Public Health 2021 Aug 10;18(16):8430. [doi: 10.3390/ijerph18168430] [Medline: 34444189]
33. Weise H, Zenner B, Schmiedchen B, et al. The effect of an app-based home exercise program on self-reported pain intensity in unspecific and degenerative back pain: pragmatic open-label randomized controlled trial. J Med Internet Res 2022 Oct 28;24(10):e41899. [doi: 10.2196/41899] [Medline: 36215327]
34. Cimarras-Otal C, Marcen-Cinca N, Rabal-Pelay J, et al. Adapted exercises versus general exercise recommendations on chronic low back pain in industrial workers: a randomized control pilot study. Work 2020;67(3):733-740. [doi: 10.3233/WOR-203322] [Medline: 33164978]
35. Yildizeli Topcu S. Relations among pain, pain beliefs, and psychological well-being in patients with chronic pain. Pain Manag Nurs 2018 Dec;19(6):637-644. [doi: 10.1016/j.pmn.2018.07.007] [Medline: 30181033]
36. Marshall PWM, Schabrun S, Knox MF. Physical activity and the mediating effect of fear, depression, anxiety, and catastrophizing on pain related disability in people with chronic low back pain. PLoS ONE 2017;12(7):e0180788. [doi: 10.1371/journal.pone.0180788] [Medline: 28686644]
37. Marshall PW, Morrison NMV, Gibbs M, Schabrun SM. The effect of exercise engagement on low back disability at 12-months is mediated by pain and catastrophizing in a community sample of people with chronic low back pain. Behav Res Ther 2022 Dec;159:104205. [doi: 10.1016/j.brat.2022.104205] [Medline: 36215852]
38. Dishman RK, Sallis JF, Orenstein DR. The determinants of physical activity and exercise. Public Health Rep 1985;100(2):158-171. [Medline: 3920714]
39. Meyer OL, Castro-Schilo L, Aguilar-Gaxiola S. Determinants of mental health and self-rated health: a model of socioeconomic status, neighborhood safety, and physical activity. Am J Public Health 2014 Sep;104(9):1734-1741. [doi: 10.2105/AJPH.2014.302003] [Medline: 25033151]
40. Labinsky H, Gupta L, Raimondo MG, Schett G, Knitza J. Real-world usage of digital health applications (DiGA) in rheumatology: results from a German patient survey. Rheumatol Int 2023 Apr;43(4):713-719. [doi: 10.1007/s00296-022-05261-7] [Medline: 36543961]
41. Tian Y, Zhou K, Pelleg D. What and how long: prediction of mobile app engagement. ACM Trans Inf Syst 2022 Jan 31;40(1):1-38. [doi: 10.1145/3464301]

Abbreviations

AI: artificial intelligence
LBP: low back pain
MSD: musculoskeletal disorder
NRS: numeric rating scale
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

XSL•FO
RenderX

Edited by A Mavragani; submitted 28.02.24; peer-reviewed by JB Lechauve, Z Rahemi; revised version received 11.12.24; accepted 12.12.24; published 18.02.25.

Please cite as:
Griefahn A, Avermann F, Zalpour C, Marshall RP, Cordon Morillas I, Luedtke K
Exploring the Effect of an 8-Week AI-Composed Exercise Program on Pain Intensity and Well-Being in Patients With Spinal Pain: Retrospective Cohort Analysis
JMIR Form Res 2025;9:e57826
URL: https://formative.jmir.org/2025/1/e57826
doi:10.2196/57826

© Annika Griefahn, Florian Avermann, Christoff Zalpour, Robert Percy Marshall, Inés Cordon Morillas, Kerstin Luedtke. Originally published in JMIR Formative Research (https://formative.jmir.org), 18.2.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Involving End Users in the Development and Usability Testing of a Smartphone App Designed for Individuals With Prediabetes: Mixed-Methods Focus Group Study

Natalie Grieve[1], BA, MSc; Kyra Braaten[1], BSc, MSc; Megan MacPherson[1], BA, MOT, PhD; Sam Liu[2], BPHE, MSc, PhD; Mary E Jung[1], HBKin, MSc, PhD

[1]Faculty of Health and Social Development, University of British Columbia, 1238 Discovery Way, Kelowna, BC, Canada
[2]Exercise Science, Physical and Health Education, University of Victoria, Victoria, BC, Canada

Corresponding Author:
Mary E Jung, HBKin, MSc, PhD
Faculty of Health and Social Development, University of British Columbia, 1238 Discovery Way, Kelowna, BC, Canada

Abstract

Background: Technology is more likely to be used when it is designed to meet the needs of end users. To supplement the Small Steps for Big Changes diabetes prevention program, a smartphone app was developed in partnership with past Small Steps for Big Changes clientele. Usability testing is critical for the ongoing use and adoption of mobile health apps by providing insight on where appropriate adjustments and improvements need to be made to ensure user satisfaction.

Objective: A focus group with 7 participants was conducted to examine the app's usability and collect feedback for future iterations.

Methods: Past Small Steps for Big Changes clientele participated in a cognitive walkthrough of 8 novel tasks and completed the System Usability Scale survey. Participants were then given the option to use the app for 3 weeks before completing the User-Mobile Application Rating Scale.

Results: Analysis of the cognitive walkthrough identified 26 usability problems; each was coded using a heuristic evaluation to describe usability errors. The most frequently coded errors included inappropriate progress feedback, information appearing in an illogical order, counterintuitive design, and issues with app aesthetics. A mean summary score of 66.8% (SD 18.91) was reported for the System Usability Scale, representing a marginal acceptability score and indicating that design issues needed to be resolved. A User-Mobile Application Rating Scale mean score of 3.59 (SD 0.33) was reported, implying an average acceptability rating.

Conclusions: These findings identified necessary improvements in the app, ranging from minor aesthetic problems to major functionality problems. Involving end users allows the app to be tailored to the client's preferences and increases the likelihood of usage. This app aligns with Small Steps for Big Changes' program components and behavior change techniques that can improve health outcomes for future clients and allow them to self-monitor their exercise, diet, and goals.

(JMIR Form Res 2025;9:e59386) doi:10.2196/59386

KEYWORDS

usability evaluation; mHealth; usability testing; app; end-user; focus group; participant; survey; diabetes; user-centered; cognitive walkthrough; cognitive walkthroughs; questionnaire; mobile phone; digital health; prediabetes

Introduction

Background

Involving end users in mobile health (mHealth) technology development increases the value and practical use of the product [1,2]. Recognizing a client's expertise through collaboration allows end users to effectively express their needs and expectations during the development phases of mHealth apps [3]. mHealth technologies have been used to increase accessibility to health care resources [4] and increase patient autonomy [5]. The uptake of mHealth technology has been shown to improve chronic disease management including improvements in weight management and hypertension, and decreased hospitalizations [6]. Within diabetes research, mHealth tools have assisted in self-management efforts, improved glycemic control [6-8], and are acceptable by users [9]. mHealth apps allow users to play an active role in their health, subsequently improving health outcomes. To capitalize on these advantages, an mHealth app was developed to supplement a diabetes prevention program run in British Columbia, Canada.

XSL•FO
RenderX

Small Steps for Big Changes is a community-based diabetes prevention program designed for individuals with prediabetes [10]. Small Steps for Big Changes is delivered in 6 sessions over 4 weeks using a client-centered counseling style known as motivational interviewing with follow-ups over 12 months. Clients work one-on-one with their coach to cover several topics including goal setting, healthy food choices, and exercise adherence. During the program's first session, clients are introduced to self-monitoring techniques, including tracking diet and exercise. Various forms of tracking are discussed, including commercial mHealth apps, journals, paper tracking, and exercise logs. While the most prominent method of tracking used in the program is commercially available mHealth apps, several challenges have been reported by Small Steps for Big Changes clients and coaches with such tools [11]. This program-specific mHealth app was developed to supplement the Small Steps for Big Changes program and to fill the gap in mHealth apps previously identified by Small Steps for Big Changes participants. Involving end users during development supported initiatives to ensure the app was relevant and useful. This mHealth app was developed to support the Small Steps for Big Changes program by integrating the frequently used behavior change techniques throughout the 6 sessions, emphasizing goals, planning, feedback, and monitoring [12].

Development of the Small Steps for Big Changes mHealth app was guided by the FASTER (Framework for Accelerated and Systematic Technology-Based Intervention Development and Evaluation Research), a 3-phase framework that recognizes the complexities of conducting research on mHealth technologies in a timely fashion while also consistently engaging end users throughout [13]. Phase 1 included a needs assessment, literature review, and preliminary evaluation; the feedback and results from phase 1 helped inform the app prototype used in this usability study [14]. This study focused on the second phase of FASTER, examining the progressive usability of the app [13]. In this phase, small-scale usability testing helped uncover usability errors, levels of acceptability, and improvements needed before large-scale implementation (phase 3 of FASTER). Usability testing is critical for sustained use and uptake of mHealth apps by identifying where changes can be made [13]. An app that is difficult to navigate with an interface that is nonintuitive limits usage. mHealth technologies provide opportunities to self-monitor and manage health behaviors for individuals with chronic disesase [15]. Successful self-monitoring of behaviors involves consistency; if carried out on an app, this needs an interface that is easy to use and will encourage the likelihood that goals will be recorded and revisited. Understanding where usability errors exist will inform appropriate changes needed to improve app features and increase user satisfaction [13].

Objective

The aim of this study was the answer the following questions: where do existing usability errors exist in the Small Steps for Big Changes mHealth app and is the app deemed usable by past Small Steps for Big Changes participants? To answer this question, this study used a mixed-methods approach involving a cognitive walkthrough (CW) focus group, take-home testing period, and self-reported usability surveys to engage end users in the examination of usability of a Small Steps for Big Changes mHealth app.

Methods

Small Steps for Big Changes App Development

Pathverse, the app-building company used for this project that specializes in no-code app building for research and learning [16], worked with our research team to create a platform specific to the needs of the Small Steps for Big Changes program. Several meetings took place with Pathverse and the research team to organize content, create select features, and tailor the content to the Small Steps for Big Changes program. This app included basic self-monitoring features, including tracking for steps, minutes of daily exercise, weight, waist circumference, and hemoglobin A_{1c} (HbA_{1c}). Additional capabilities included logging exercise sessions, recording goals, diary entry opportunities, and access to additional frequently asked questions and resources including nutritional cooking videos and program content. Additional details on layout and design can be seen in the attached screenshots available in Multimedia Appendix 1.

Participants

A total of 7 participants were recruited to participate in this study. A sample size between 6 and 10 participants was determined a priori based on past literature exploring user-centered CW methods in mHealth chronic disease self-management [17,18]. Secondary measures used to collect additional usability data included the System Usability Scale (SUS) [19,20] and User-Mobile Application Rating Scale (uMARS) [21]. Past Small Steps for Big Changes clients who had completed the 4-week training phase of the program were invited to take part in this study. Eligible participants included past Small Steps for Big Changes clients who were adults aged 18 years or older, able to read and write in English, assessed for prediabetes by one of the following means: (1) physician-diagnosed prediabetes, (2) HbA_{1c} values between 5.7% and 6.4%, and (3) an American Diabetes Association risk questionnaire score indicating increased risk of diabetes (>5). Participants also needed to have access to a mobile device with internet access (such as a tablet or smartphone). In an attempt to represent the current demographic of the Small Steps for Big Changes program (75% of individuals identified as female and 25% as male), participants were purposively sampled. Participants were asked to bring an internet-enabled mobile device during their laboratory visit with capabilities to connect and access the device over the 3-week take-home period.

Ethical Considerations

Informed consent was obtained from all individual participants and all procedures performed were in accordance with the ethical standards of the University of British Columbia Behavioural Research Ethics Board (H22-01399).

CW: Phase 1

Participants were invited to engage in a CW focus group at the University of British Columbia Okanagan Campus. CWs involve a task-based method to explore the usability of digital health

XSL•FO
RenderX

technologies and when used in conjunction with questionnaires are part of a multistep process to evaluate usability [17,18,22]. As participants arrived, research assistants helped to provide assistance to connect to the internet, download the app, and create a user profile. A 10-minute demonstration of the Small Steps for Big Changes app on a large projector screen was presented before the CW began to account for the app's novelty. During this demonstration a researcher lead a familiarization walkthrough of each of the function landing pages (home page, tracking page, and resources page) as well as some of the button functions (go-back, save, and delete). Participants followed along on their devices and were given time to play with navigating the app before starting the novel tasks. After this period, each participant was provided with a task booklet outlining 8 novel tasks to be completed during the CW focus group. Each task had 3 corresponding questions used to identify usability problems about the task and assist with the facilitation of group discussion. The workbook was designed for additional observations to be recorded during each task. During the CW, participants were asked to rotate through leading the completion of a task using their instincts by informing the facilitator how to perform the task on a large screen for the group to follow. After each task, all participants were asked to record their observations and share their thoughts about any usability problems they encountered. Each participant lead at least 1 task during the CW. Once all tasks were completed in the CW, participants were granted access to the app for an additional 3 weeks to use it as they pleased, simulating a real-world setting. During this time, participants were encouraged to make note of any additional usability errors, feedback for future modifications, and general reactions or evaluations. This information was provided along with compensation following completion of this study.

Mobile Device Proficiency Questionnaire and Demographics Survey

The Mobile Device Proficiency Questionnaire (MDPQ-16) [23] was used to assess participants' mobile device proficiency, and demographics were collected for descriptive reporting. Participants self-reported their age, sex, gender, whether they were born in Canada, whether they identify as Indigenous or a visible minority, their highest level of education completed, occupational status, household income, and marital status. During the initial laboratory visit, participants were asked to complete a demographics survey and the MDPQ-16 on iPads (Apple Inc) before beginning the CW. The MDPQ-16 includes 16 questions organized into 8 subscales such as mobile device basics, communication, internet, troubleshooting, and are answered on a 5-point scale (1=never tried and 5=tried very easily). Each subscale score is averaged and then summed, with a maximum score of 40 (a higher score indicates greater proficiency [23]).

About SUS

After completing the CW tasks, participants were asked to complete the SUS [19]. The SUS was used to assess system usability. This brief 10-question survey asks general usability questions about ease of use and complexity on a 5 point-scale (1=strongly disagree and 5=strongly agree). The SUS is a widely used scale, a recent meta-analysis concluded that the SUS was a suitable measure for evaluating the usability of digital health apps [20]. Final SUS scores range from 0 - 100, with higher scores indicating better usability [19] with a benchmark acceptability score of 68 [20]. This survey was delivered immediately after the CW to each participant on an iPad.

User Version of the Mobile App Rating Scale: Phase 2

After completing an optional take home period, the uMARS [21] was used to assess the quality and usability of the mHealth app. The uMARS items are answered on a 5-point scale, representing the following 1-inadequate, 2-poor, 3-acceptable, 4-good, and 5-excellent [21]. The uMARS is a reliable tool to measure app quality and has been used to assess mHealth apps including self-management behaviors associated with type 2 diabetes [24]. This survey offered additional detail and depth compared to the SUS and was delivered to the participants after an optional 3-week at-home phase. The uMARS consists of 4 subscales focused on engagement, functionality, aesthetics and information. A mean score for each subscale is calculated, and a total app quality score is calculated by averaging the subscale means. The survey also includes additional questions about app subjective quality, perceived impact, and an option to leave further written comments. Participants were given information about the take-home phase of this study and were asked to use the app freely for 3 weeks. After this take-home period, participants were sent a survey link to the uMARS.

Analysis

All survey data from the demographics, MDPQ-16, SUS, and uMARS were descriptively analyzed using SPSS (version 27; IBM Corp) to examine mean scores, SD, and range. The CW was analyzed through an audio recording and transcript. The usability problems identified in the CW task workbook and transcript were coded using Nielsen's heuristic evaluation [25,26]. The 10 usability heuristics outline general principles for user interface design; if an identified usability problem violates any of the 10 heuristics, it is coded accordingly. Furthermore, 2 coders completed the qualitative analysis and assigned severity scores to each of the usability errors identified.

Results

Demographics and MDPQ-16

A total of 5 of the 7 participants were aged older than 70 years, and 2 were between 55 and 65 years of age. No individuals identified as Indigenous or a visible minority. Further, 2 participants identified as having a disability. There was variance in education level and family income; 5 participants were retired. Table 1 reports on more detailed demographic results. MDPQ-16 scores were reported to assess levels of mobile device proficiency. The mean (SD) MDPQ-16 total score was 36.36 and SD of 3.42, individual MDPQ-16 scores are reported in Multimedia Appendix 1. These MDPQ-16 scores represented high levels of mobile device proficiency for our participants compared to those identified in previous literature [23].

XSL•FO
RenderX

Table . Demographic characteristics of participants.

Sample characteristic	Participants who selected this response
Age (years), mean (SD)	70 (7.47)
Sex, n (%)	
Female	4 (57)
Male	3 (43)
Gender, n (%)	
Woman	4 (57)
Man	3 (43)
Born in Canada, n (%)	
Yes	4 (57)
No	3 (43)
No	7 (100)
No	7 (100)
Identifies as a person with a disability, n (%)	
Yes	2 (29)
No	4 (57)
Prefer not to answer	1 (14)
Level of education, n (%)	
High school	1 (14)
University certificate or diploma below the bachelor level	3 (43)
Postgraduate degree	2 (29)
College, CEGEP[a], or other nonuniversity certificate or diploma	1 (14)
Working status, n (%)	
Working full time	2 (29)
Retired	5 (71)
Annual income level (US $)	
$41,688 to $62,531	1 (14)
$62,532 to $82,681	1 (14)
$104,220 or more	3 (43)
Prefer not to answer	2 (29)
Marital status, n (%)	
Married	6 (86)
Widowed	1 (14)

[a]CEGEP, Collège d'enseignement general et professionnel, which means general and professional teaching college

About CW

The CW lasted 60 minutes in duration. Participants were asked to complete 8 novel tasks while engaging with the Small Steps for Big Changes app. A total of 26 usability problems were identified and assigned severity scores in the recording transcript and client workbooks. The highest number of usability problems identified for a specific task was 7 for task 1, and the lowest was 0 for task 8. All severity scores and usability problems identified for each task and corresponding Nielsen heuristics are reported in Table 2.

XSL•FO
RenderX

Table . Usability problems identified in associated workbook task completion.

Task	Number of usability problems identified	Mean severity score (SD)	Taps expected to complete task (n)	Actual taps to complete task (n)	Nielsen heuristics identified (*f*)
1-Record a weight value	7	2.1 (0.69)	5	8	1 (3), 2 (4), 3 (1), 4 (2), 6 (4), 8 (2)
2-Navigate to the FAQ[a] page	2	0.4 (0.53)	4	4	3 (1), 6 (1), 8 (1)
3-Log an exercise	3	1 (1.15)	17	18	1 (1), 2 (1), 4 (1), 6 (2), 8 (2)
4-Create a diary entry	1	0.5 (0.58)	7	7	10 (1)
5-Record a goal	5	1.9 (1.38)	12	15	1 (1), 2 (2), 3 (1), 4 (1), 5 (1), 6 (2), 9 (1)
6-Navigate to a module	4	1.6 (0.53)	6	9	2 (3), 3 (1), 8 (1), 10 (2)
7-Update the hemoglobin A_{1c} value within the tracking page	4	0 (0)	10	10	2 (1), 4 (2), 6 (2), 8 (1)
8-Navigate to a video resource	0	0 (0)	5	5	0

[a]FAQ: frequently asked questions.

The usability problems identified in the CW included issues with icon and text sizes, scroll sensitivity, system preferences, and logical location of features and information. Common system preference issues were challenges saving and editing goals and the exclusion of a 24-hour clock. Participants also had difficulty scrolling and reported favorability for a text box for selecting values. Usability problems were analyzed by coding each into 10 categories. The 26 usability problems highlighted errors associated with information appearing in an illogical order and lack of visibility of information. For example, the usability error named “tracker alignment” indicates that the weight tracking feature’s dates did not match what was seen on the graph displayed in the participant’s progress summary. This was coded as 1 (incorrect feedback provided by the system) and 2 (information appearing in an illogical order). All usability problems were coded accordingly; some were assigned only 1 Nielsen heuristic, while others were assigned multiple.

NG and KB double-coded the 26 identified usability errors according to heuristic evaluation. Nielsen heuristics of 2 (match between system and the real world) and 6 (recognition rather than recall) were the most frequently coded (11 times each), while heuristics of 5 (error prevention), 7 (flexibility and efficiency of use), and 9 (help users recognize, diagnose, and recover from errors) were rarely coded. Table 3 outlines each heuristic score, their coded frequency, and the corresponding percentage of the 49 total codes used.

Table . Nielsen heuristic frequency for usability problem codes from a total of 26 usability problems and 49 total codes.

Nielsen heuristic	Frequency (%[a])
Visibility of system status	5 (10)
Match between system and the real world	11 (22)
User control and freedom	4 (8)
Consistency and standards	6 (12)
Error prevention	1 (2)
Recognition rather than recall	11 (22)
Flexibility and efficiency of use	0 (0)
Aesthetic and minimalist design	7 (14)
Help users recognize, diagnose, and recover from errors	1 (2)
Help and documentation	3 (6)

[a]%: percentage of the total frequency of codes.

XSL•FO
RenderX

About SUS

The mean SUS score was 66.8% for the sample, with an SD of 18.91 and a large range from 50 to 97.5. Further, 4 participants rated the usability moderately low, between 50% and 60%, while the remaining 3 rated the usability highly, resulting in scores above 85%.

About uMARS

In total, 5 of the 7 clients who participated in the CW completed the full uMARS. The app quality total score was calculated to be a mean score of 3.59/5 with an SD of 0.33, with the highest-rated subcategory being aesthetics at a mean score of 3.8 (SD 0.51) and the lowest being engagement at 3.12 (SD 0.33). The highest perceived impact scores were behavior change and awareness, both receiving a mean score of 3.33 (SD 1.03), with knowledge being the lowest at a mean of 2.8 (SD 0.84). Additional feedback for improvements included adding analytics for the progress made on the app, adding synchronizing capabilities with smartwatches to avoid manually recording physical activity and including tracking standards for specific age groups and sex. A preference to the Fitbit app was also reported by 2 participants during discussion of additional feedback after the CW. Positive comments resulting from completion of the uMARs included comments about the potential for the app to be a very useful tool, the value of recording anthropometrics such as weight and HbA_{1c}, the emphasis on diabetes, and the use of the graphs as a visual representation of the client's progress. Survey scores for usability data are reported in Table 4, and additional results from the usability surveys are included in Multimedia Appendix 1.

Table . Usability scores for quantitative surveys (N=18).

Survey	Mean (SD)	Range
MDPQ-16[a]	36.36/40 (3.42)	31 - 40
SUS[b]	66.8/100 (18.91)	50 - 97.5
uMARS[c]	3.59/5 (0.33)	3.03 - 3.83

[a]MDPQ-16: Mobile Device Proficiency Questionnaire.

[b]SUS: System Usability Scale.

[c]uMARS: User-Mobile Application Rating Scale.

Follow-Up With Pathverse

Once the research team noted the usability problems and feedback, 2 follow-up meetings were conducted with Pathverse to address the usability concerns and make appropriate improvements. Updates made by the Pathverse team included correcting minor bugs with diary notification scheduling, fixing the scaling of images and icons, and bringing consistency across devices (Samsung vs Apple). Additions made to the app included 24-hour time when inputting goals and tracking, opportunity to input manual steps, and updates for some of the language and icons used (the menu symbol, save button, and trash icon).

Discussion

Findings

The 26 usability problems identified in the CW identified relevant usability errors seen during task-based discussion within the group of participants. Further, 9 of the 10 Nielsen heuristics were coded, indicating usability errors across multiple domains of the app; however, most usability problems were addressed for preference changes and minor updates in the most recent iteration of the app. Other suggestions, including a hover function to show a description of functions within the app, and the ability to link the app with a Fitbit have not yet been addressed due to cost, time, and practicality implications.

The usability score of 66.8 for the SUS represented a marginal usability score [19,20]. The variance could be due to a range in expectations of the app and its functions; a few users expressed they had been using different health monitoring apps, which they were quite satisfied with, which may have influenced their perceptions of the Small Steps for Big Changes app. After 3 weeks of optional use, a uMARS mean score of 3.59 was reported, describing an acceptable score [22]. MDPQ-16 scores were relatively high in our sample; we can have confidence that the feedback provided was independent of a lack of skills or ability to use mobile apps. Within our sample, moderate levels of usability may have been due to a low level of willingness to track in general. We did not purposively sample those who favored smartphone tracking over other methods such as tracking in a journal or workbook,; and some users may have had little interest in tracking at all.

The literature has identified perceived usability as an existing barrier to mHealth usage [27]; this aligns with the moderate levels of usability reported in this study. However, high levels of usability have been reported by select mHealth technologies designed for individuals with type 2 diabetes [9,28], which counter the moderate levels of usability seen in this study. Although the levels of usability have demonstrated mixed results, when comparing user feedback, various themes did align with our findings. Users agreed on noting similar improvements, including adding more knowledge and information available on the mHealth apps [28], suggesting that mHealth app content may be oversimplified. Interestingly, this study reported high levels of mobile device proficiency from adults with a mean age of 70 years, which may contradict previous research noting older age as a barrier to mHealth usage [27].

This study identified usability problems and necessary improvements for the Small Steps for Big Changes program

XSL•FO
RenderX

app. With only moderate levels of usability reported, these results emphasize the need for user feedback during the development process. This highlights the importance of the end user's involvement in app development and prototype stages. Several iterations of mHealth technology should be tested before large-scale implementation to ensure the end product is relevant and highly useful to users. This study demonstrates that early usability evaluations can help identify and address issues at the beginning of the development cycle, saving time and resources in the long run. The iterative nature of this study's approach, where usability problems were identified and subsequently addressed, underscores the importance of continuous improvement. Following the steps of the FASTER framework helped ensure that necessary steps were not overlooked; during this process, essential findings for decision-making can save on cost and the long-term success of an mHealth app [13]. This study highlights the importance of tailoring mobile apps, especially health-related ones, to meet the specific needs and preferences of older adults. Developers should prioritize user-centered design principles, considering factors such as font size, icon clarity, and ease of navigation to enhance usability for this demographic.

Past research has highlighted the importance of integrating multiple methods to assess usability in digital health evaluation [22],; this study incorporated various modes of feedback through a CW, real-world use for 3 weeks, validated questionnaires, and qualitative feedback. Diverse feedback sources provide a more comprehensive understanding of usability and user experiences. Each feedback method uncovers different aspects of usability. Combining various feedback sources allows for a holistic evaluation of the mobile app. This approach ensures that usability problems are not only identified but also validated across different dimensions. It helps prioritize issues that have the most significant impact on users. The real-world use over 3 weeks provides insights into how usability evolves over time. It helps identify whether initial issues persist or if users adapt to the app. The combined feedback informs tailored iterations. Specific issues can be addressed when identified through different sources in a targeted manner. For example, usability problems identified in CW may lead to design changes, while questionnaire data can guide overall satisfaction improvements.

Limitations

Limited diversity within our sample was seen, specifically for those identifying as Indigenous or a visible minority. While this is representative of the current demographics served by Small Steps for Big Changes, as Small Steps for Big Changes expands to more communities, the potential to target more diverse opinions should be highlighted, especially for those who may benefit the most from increased accessibility (eg, individuals living in rural communities). This study did not explore long-term user compliance and security issues; further research and considerations should be explored to identify solutions around these limiting factors [29]. Considerations including (1) a small sample size when reporting uMARS results and (2) an unrealistic simulation of app implementation (providing assistance during installation and a familiarization of the app) should be taken into account when interpreting the results.

Future Directions for Research

Future research may look toward phase 3 of FASTER through large-scale implementation and evaluation of the app as a supplement to the Small Steps for Big Changes diabetes prevention program to examine the app's effectiveness in improving health behaviors and clinical outcomes for clients. mHealth prompts such as texts and push notifications have been well-received by users in diabetes prevention [30]. Looking toward these capabilities within the Small Steps for Big Changes app is an additional function that may benefit client health outcomes by allowing them to self-monitor their exercise, diet, and goals. Future iterations of the app will call for additional cycles of usability testing and evaluation to ensure a satisfactory user experience is reached. Usability is an important aspect of user experience; however, additional factors including incentivizing, convenience, and practicality of mHealth apps may also influence decisions to integrate their use in practice. For example, the results of this study report that the lowest category in the uMARS scale was engagement. In the uMARS, user engagement refers to whether the app is entertaining, interesting, customizable and interactive; if users do not find the app engaging they may be less incentivized to use it.

Conclusions

Diabetes self-management efforts can be improved using mHealth technologies; it can help establish meaningful routines, allow clients to better understand their health conditions and status, and provide additional outlets for interactions with health care professionals and resources [31]. Providing tools to make behavior change efforts easier for clients will lead to better outcomes; however, these tools need to be well-accepted by users to increase the likelihood of usage. This study successfully assessed the usability of a self-monitoring app for individuals with prediabetes by identifying 26 usability errors and recognizing where improvements needed to be made. These activities are necessary for generating research evidence for technology-based interventions [13] to ensure usage is not a limiting factor. Moderate usability levels were reported by users, prompting a new iteration of the app, emphasizing the importance of user involvement. Involving end users in this process allows the mHealth app usage experts to make decisions during development. Overlooking usability testing during applicaton development can result in premature large-scale implementation before critical changes and errors can be addressed. Embracing multiple feedback modes reflects a user-centric approach to design and evaluation. It demonstrates a commitment to understanding the user experience from various angles, which is essential for creating user-friendly apps.

Acknowledgments

We respectfully acknowledge that the work and research conducted for this study was on the traditional, ancestral, and unceded territory of the Syilx Nation.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional usability data and screenshots of app platform.
[DOCX File, 2565 KB - formative_v9i1e59386_app1.docx]

References

1. de Beurs D, van Bruinessen I, Noordman J, Friele R, van Dulmen S. Active involvement of end users when developing web-based mental health interventions. Front Psychiatry 2017;8:72. [doi: 10.3389/fpsyt.2017.00072] [Medline: 28515699]
2. Timmers T, van der Weegen W, Janssen L, Kremer J, Kool RB. Importance of patient involvement in creating content for ehealth interventions: qualitative case report in orthopedics. JMIR Form Res 2022 Nov 3;6(11):e39637. [doi: 10.2196/39637] [Medline: 36326799]
3. Song T, Yu P, Bliokas V, et al. A clinician-led, experience-based co-design approach for developing mHealth services to support the patient self-management of chronic conditions: development study and design case. JMIR mHealth uHealth 2021 Jul 20;9(7):e20650. [doi: 10.2196/20650] [Medline: 34283030]
4. Price M, Yuen EK, Goetter EM, et al. mHealth: a mechanism to deliver more accessible, more effective mental health care. Clin Psychol Psychother 2014;21(5):427-436. [doi: 10.1002/cpp.1855] [Medline: 23918764]
5. Sahin C. Rules of engagement in mobile health: what does mobile health bring to research and theory? Contemp Nurse 2018;54(4-5):374-387. [doi: 10.1080/10376178.2018.1448290] [Medline: 29502472]
6. Marcolino MS, Oliveira JAQ, D'Agostino M, Ribeiro AL, Alkmim MBM, Novillo-Ortiz D. The impact of mHealth interventions: systematic review of systematic reviews. JMIR mHealth uHealth 2018 Jan 17;6(1):e23. [doi: 10.2196/mhealth.8873] [Medline: 29343463]
7. Whitehead L, Seaton P. The effectiveness of self-management mobile phone and tablet apps in long-term condition management: a systematic review. J Med Internet Res 2016 May 16;18(5):e97. [doi: 10.2196/jmir.4883] [Medline: 27185295]
8. Stevens S, Gallagher S, Andrews T, Ashall-Payne L, Humphreys L, Leigh S. The effectiveness of digital health technologies for patients with diabetes mellitus: a systematic review. Front Clin Diabetes Healthc 2022;3:936752. [doi: 10.3389/fcdhc.2022.936752] [Medline: 36992773]
9. Alanzi T, Alanazi NR, Istepanian R, Philip N. Evaluation of the effectiveness of mobile diabetes management system with social networking and cognitive behavioural therapy (CBT) for T2D. mHealth 2018;4:35. [doi: 10.21037/mhealth.2018.06.05] [Medline: 30221168]
10. Bean C, Dineen T, Locke SR, Bouvier B, Jung ME. An evaluation of the reach and effectiveness of a diabetes prevention behaviour change program situated in a community site. Can J Diabetes 2021 Jun;45(4):360-368. [doi: 10.1016/j.jcjd.2020.10.006] [Medline: 33323314]
11. Dineen TE, Bean C, Jung ME. Successes and challenges from a motivational interviewing-informed diabetes prevention program situated in the community. Health Promot Pract 2024 Mar;25(2):274-284. [doi: 10.1177/15248399221115066] [Medline: 36047459]
12. MacPherson MM, Dineen TE, Cranston KD, Jung ME. Identifying behaviour change techniques and motivational interviewing techniques in small steps for big changes: a community-based program for adults at risk for type 2 diabetes. Can J Diabetes 2020 Dec;44(8):719-726. [doi: 10.1016/j.jcjd.2020.09.011] [Medline: 33279096]
13. Wang RH, Kenyon LK, McGilton KS, et al. The time is now: a FASTER approach to generate research evidence for technology-based interventions in the field of disability and rehabilitation. Arch Phys Med Rehabil 2021 Sep;102(9):1848-1859. [doi: 10.1016/j.apmr.2021.04.009] [Medline: 33992634]
14. MacPherson M, Grieve N, Gala B, Braaten K, Jung M. Development of a small steps for big changes diabetes prevention app: application of the development phase of FASTER. J Technol Behav Sci 2024;9(4):681-702. [doi: 10.1007/s41347-023-00373-8]
15. Yi JY, Kim Y, Cho YM, Kim H. Self-management of chronic conditions using mhealth interventions in Korea: a systematic review. Healthc Inform Res 2018 Jul;24(3):187-197. [doi: 10.4258/hir.2018.24.3.187] [Medline: 30109152]
16. Liu S, La H, Willms A, Rhodes RE. A "no-code" app design platform for mobile health research: development and usability study. JMIR Form Res 2022 Aug 18;6(8):e38737. [doi: 10.2196/38737] [Medline: 35980740]
17. Georgsson M, Staggers N, Årsand E, Kushniruk A. Employing a user-centered cognitive walkthrough to evaluate a mHealth diabetes self-management application: a case study and beginning method validation. J Biomed Inform 2019 Mar;91:103110. [doi: 10.1016/j.jbi.2019.103110] [Medline: 30721757]
18. Beauchemin M, Gradilla M, Baik D, Cho H, Schnall R. A multi-step usability evaluation of a self-management app to support medication adherence in persons living with HIV. Int J Med Inform 2019 Feb;122:37-44. [doi: 10.1016/j.ijmedinf.2018.11.012] [Medline: 30623782]
19. Bangor A, Kortum PT, Miller JT. An empirical evaluation of the system usability scale. Int J Hum Comput Interact 2008 Jul 29;24(6):574-594. [doi: 10.1080/10447310802205776]

XSL•FO
RenderX

20. Hyzy M, Bond R, Mulvenna M, et al. System usability scale benchmarking for digital health apps: meta-analysis. JMIR mHealth uHealth 2022 Aug 18;10(8):e37290. [doi: 10.2196/37290] [Medline: 35980732]
21. Stoyanov SR, Hides L, Kavanagh DJ, Wilson H. Development validation of the user version of the mobile application rating scale and (uMARS). JMIR mHealth uHealth 2016 Jun 10;4(2):e72. [doi: 10.2196/mhealth.5849] [Medline: 27287964]
22. Davis R, Gardner J, Schnall R. A review of usability evaluation methods and their use for testing ehealth hiv interventions. Curr HIV/AIDS Rep 2020 Jun;17(3):203-218. [doi: 10.1007/s11904-020-00493-3] [Medline: 32390078]
23. Roque NA, Boot WR. A new tool for assessing mobile device proficiency in older adults: the mobile device proficiency questionnaire. J Appl Gerontol 2018 Feb;37(2):131-156. [doi: 10.1177/0733464816642582] [Medline: 27255686]
24. Tripathi D, Vikram NK, Chaturvedi S, Bhatia N. Development of "DiabetesSutra" a mobile application for lifestyle management of type 2 diabetes in India. J Diabetes Metab Disord 2024 Jun;23(1):709-720. [doi: 10.1007/s40200-023-01339-6] [Medline: 38932794]
25. Nielsen J. Usability inspection methods. Presented at: Conference Companion on Human Factors in Computing Systems - CHI 94; Apr 24-28, 1994; Boston, Massachusetts, United States p. 413-414 URL: http://portal.acm.org/citation.cfm?doid=259963 [accessed 2024-01-31] [doi: 10.1145/259963.260531]
26. Kushniruk AW, Monkman H, Tuden D, Bellwood P, Borycki EM. Integrating heuristic evaluation with cognitive walkthrough: development of a hybrid usability inspection method. Stud Health Technol Inform 2015;208:221-225. [doi: 10.3233/978-1-61499-488-6-221]
27. Byambasuren O, Beller E, Hoffmann T, Glasziou P. Barriers to and facilitators of the prescription of mhealth apps in australian general practice: qualitative study. JMIR mHealth uHealth 2020 Jul 30;8(7):e17447. [doi: 10.2196/17447] [Medline: 32729839]
28. Goswami R, Garner SL, George CE. Usability of an mHealth application in India for type 2 diabetes mellitus. Health Educ J 2023 Feb;82(1):108-116. [doi: 10.1177/00178969221138483]
29. Istepanian RSH, Al-Anzi TM. m-Health interventions for diabetes remote monitoring and self management: clinical and compliance issues. mHealth 2018;4:4. [doi: 10.21037/mhealth.2018.01.02] [Medline: 29552566]
30. MacPherson MM, Merry KJ, Locke SR, Jung ME. mHealth prompts within diabetes prevention programs: a scoping review. mHealth 2022;8:20. [doi: 10.21037/mhealth-21-22] [Medline: 35449504]
31. Torbjørnsen A, Ribu L, Rønnevig M, Grøttland A, Helseth S. Users' acceptability of a mobile application for persons with type 2 diabetes: a qualitative study. BMC Health Serv Res 2019 Sep 6;19(1):641. [doi: 10.1186/s12913-019-4486-2] [Medline: 31492176]

Abbreviations

CW: cognitive walkthrough
FASTER: Framework for Accelerated and Systematic Technology-Based Intervention Development and Evaluation Research
HbA_{1c}: hemoglobin A_{1c}
MDPQ-16: Mobile Device Proficiency Questionnaire
mHealth: mobile health
SUS: System Usability Scale
uMARS: User-Mobile Application Rating Scale

Edited by A Bucher; submitted 10.04.24; peer-reviewed by A Gandhi, K Rawlings, M Braun, Y Kim; revised version received 08.11.24; accepted 09.11.24; published 11.02.25.

Please cite as:
Grieve N, Braaten K, MacPherson M, Liu S, Jung ME
Involving End Users in the Development and Usability Testing of a Smartphone App Designed for Individuals With Prediabetes: Mixed-Methods Focus Group Study
JMIR Form Res 2025;9:e59386
URL: https://formative.jmir.org/2025/1/e59386
doi:10.2196/59386

© Natalie Grieve, Kyra Braaten, Megan MacPherson, Sam Liu, Mary E Jung. Originally published in JMIR Formative Research (https://formative.jmir.org), 11.2.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete

XSL•FO
RenderX

bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Activity Trackers in Physical Therapy for People With Chronic Obstructive Pulmonary Disease in the Netherlands: Cross-Sectional Study on Current Use and Implementation Determinants

Darcy Ummels[1,2], PhD; Esther Bols[1], PhD; Roel Johannes Anna Frantzen[1], BSc; Tim Frantzen[1], BSc; Levy Robeerts[1], BSc; Emmylou Beekman[1,3], PhD

[1]Research Centre for Autonomy and Participation of Persons with a Chronic Illness, Academy for Speech and Language Therapy, Zuyd University of Applied Sciences, Heerlen, Netherlands

[2]Department of Rehabilitation Medicine, Care and Public Health Research Institute, Faculty of Health, Medicine & Life Sciences, Maastricht University, Maastricht, Netherlands

[3]Department of Family Medicine, Care and Public Health Research Institute, Faculty of Health, Medicine & Life Sciences, Maastricht University, Maastricht, Netherlands

Corresponding Author:
Darcy Ummels, PhD
Research Centre for Autonomy and Participation of Persons with a Chronic Illness, Academy for Speech and Language Therapy, Zuyd University of Applied Sciences, Heerlen, Netherlands

Abstract

Background: In the Netherlands, 545,900 people experienced chronic obstructive pulmonary disease (COPD) in 2022. Physical therapy following the Royal Dutch Society for Physiotherapy (Koninklijk Nederlands Genootschap voor Fysiotherapie) guidelines for COPD treatment is a proven effective treatment for people with COPD. The revised guidelines include a new recommendation: a patient's physical activity level should be assessed with an activity tracker (AT). Literature shows that the implementation of eHealth in clinical practice, in this case, ATs, is challenging.

Objective: This study aims (1) to assess how and why ATs are currently used in physical therapy in patients with COPD and (2) to determine which barriers and facilitators are of relevance for optimal implementation of ATs during the clinical reasoning process of physical therapists in patients with COPD.

Methods: A cross-sectional study was used to evaluate the implementation of ATs in physical therapy. Included participants were physical therapists who were affiliated with Chronisch ZorgNet and had a specialization in COPD treatment. The survey content was based on the Consolidated Framework for Implementation Research, the theory of planned behavior, the framework "experiences of patients with commercially available ATs," and the Koninklijk Nederlands Genootschap voor Fysiotherapie guidelines for COPD. Physical therapists were questioned via a digital survey.

Results: In total, 211 completed surveys were analyzed. Of the 211 participating physical therapists, 108 (51.2%) used ATs, whereas most of them (n=82, 75.9%) already used ATs before it was advised in the guidelines. Physical therapists indicated that the most important reason to use ATs is that they experience it as an added health care value. Both users and nonusers indicated that the most important reason why they do not use ATs is because their patients do not want to use an AT. The second reason was a lack of knowledge in the nonuser group. Moreover, both users and nonusers indicated that the implementation of ATs was not prepared and planned for within their center.

Conclusions: Overall, these results show that ATs are not yet fully implemented in the Dutch general physical therapy practice in patients with COPD, as recommended by current evidence-based guidelines. Physical therapists need guidance for the successful implementation of ATs. This could be accomplished by providing training for physical therapists, integrating ATs into the education of (future) physical therapists, and providing support during the implementation process of ATs for both the physical therapists and management.

(JMIR Form Res 2025;9:e59533) doi:10.2196/59533

KEYWORDS

physical therapy; activity tracker; implementation; technology; COPD; chronic obstructive pulmonary disease; eHealth; meaningful use; health measurement; active lifestyle; Netherlands; physical activity; barriers; therapy

XSL•FO
RenderX

Introduction

Chronic obstructive pulmonary disease (COPD) is a heterogeneous lung condition characterized by chronic respiratory symptoms (dyspnea, cough, sputum production, and/or exacerbations) due to abnormalities of the airways (bronchitis and bronchiolitis) and/or alveoli (emphysema) that cause persistent, often progressive, airflow obstruction [1,2]. In the Netherlands, 545,900 people experienced COPD in 2022. It is the fourth leading disease in relation to disease burden and the sixth leading disease that causes mortality in the Dutch population [3]. The health care expenses for people with COPD were estimated at US $775 million in 2019, totaling 1% of the total Dutch health care expenses and 23% of the total expenditure for all respiratory diseases [4]. Physical therapy according to the Royal Dutch Society for Physiotherapy (Koninklijk Nederlands Genootschap voor Fysiotherapie [KNGF]) guidelines for COPD treatment [5] is a proven effective treatment for people with COPD. In general, the treatment is a major component of pulmonary rehabilitation and aims to increase a patient's physical activity level and physical capacity and is accomplished with home exercises and education [5]. The KNGF guidelines describe the clinical reasoning process for physical therapists based on the latest evidence. In the revised guidelines, the new recommendation is that a patient's physical activity level (eg, number of steps) is measured with an activity tracker (AT). The patient's physical activity level is an important starting point, as it codetermines whether physical therapy is indicated or a lifestyle coach is sufficient. Therefore, if ATs are not or incorrectly used, the clinical reasoning process is compromised. This could result in misclassification of patients and hence the wrong allocation to physical therapy or lifestyle coaching.

Apart from assigning patients to treatment, ATs can also be used to monitor and evaluate the physical activity level during and after the treatment [6]. Using ATs as an intervention can also enhance the physical activity level of patients with COPD when combined with the guidance of a health care professional [7]. Moreover, they can improve patients' self-management and self-efficacy [8-10].

Literature shows that the implementation of eHealth in clinical practice, in this case, ATs, is challenging [6,11,12]. Barriers in the implementation of eHealth are among others, the lack of knowledge among health care professionals concerning eHealth possibilities, unclear benefits of eHealth, resistance to change, and clinicians' perception of impersonal care [13-16]. The degree of actual use of ATs and areas of application by physical therapists in patients with COPD are currently unclear. In addition, more insight into potential barriers and facilitators for successful future implementation of ATs is necessary to optimize the treatment of patients with COPD.

A theoretical framework to assess potential barriers and facilitators for implementing innovations is the Consolidated Framework for Implementation Research (CFIR) [17]. The CFIR consists of 5 domains associated with effective implementation: intervention characteristics, inner setting, characteristics of individuals, outer setting, and process. Besides assessing potential barriers and facilitators for successful implementation, physical therapists need to change their current behavior. The theory of planned behavior (TPB) supports predicting an individual's intention to engage in a (new) behavior. The TPB consists of 4 aspects: attitude, subjective norms, perceived behavioral control, and intention [18].

This study combines the aspects of the CFIR and the TPB as a background (1) to assess how and why ATs are currently used in physical therapy in patients with COPD and (2) to determine which barriers and facilitators are of relevance for optimal implementation of ATs during the clinical reasoning process of Dutch physical therapists in patients with COPD.

Methods

A cross-sectional study was performed, in which Dutch physical therapists were questioned via a voluntary digital closed survey.

Population

Included participants were physical therapists who were affiliated with Chronisch ZorgNet and had a specialization in COPD treatment. Chronisch ZorgNet is a nationwide network that provides training for treating patients with chronic diseases according to the latest evidence (guidelines). The network consists of 3500 physical therapists working in primary, secondary, and tertiary care, of which 1200 physical therapists are specialized in lung physical therapy [19]. This network provided access to this target population. The survey was distributed by emails from Chronisch ZorgNet. All specialized lung physical therapists who were Chronisch ZorgNet members were eligible to participate in this survey. Physical therapists were excluded from this study if they treated less than 1 unique patient with COPD per week.

Survey

The survey (Multimedia Appendix 1) was drafted from documents based on international scientific literature: (1) the checklist "the use of measurement instruments in clinical practice" [20] based on the CFIR [17] (Q1), (2) the questionnaire "meaningful use of patient-specific goal setting within the clinical reasoning of physical therapists" [21] based on the TPB [18] (Q2), (3) the framework "experiences of patients with commercially available ATs" [22] (Q3), and (4) the KNGF guidelines for COPD [5]. Questions of the checklist Q1 and the questionnaire Q2 were adapted for the application of ATs instead of general measurement instruments. To gain insight into relevant barriers and facilitators, the topics of the framework Q3 were used. The survey was supplemented with questions based on the recommendations for the use of ATs in the COPD guidelines [5]. In total, 9 questions determined the eligibility of the participants (eg, sex, age, and work experience). Based on the question "I use ATs in people with COPD" (yes or no), participants answered either 59 questions if they used ATs or 33 questions if they did not.

Of the 59 questions for the AT users, 26 were based on the checklist "the use of measurement instruments in clinical practice"' Q1 [20], 5 were based on questionnaire Q2 [21], 20 were based on the framework Q3 [22], and 8 were based on the COPD guidelines [5]. Of the 33 questions for the nonusers, 19

XSL•FO
RenderX

were based on Q1 [20], 5 were based on Q2 [21], and 9 were based on Q3 [22]. The questions were either answered on a 7-point Likert scale (0=totally disagree and 7=totally agree), on a multiple-answer scale, or were open comment questions. Only 4 questions were necessarily answered by yes or no (eg, ATs are provided to our patients: yes or no). Participants were able to review and change their answers before completing the survey.

To ensure the content validity of the combined questionnaires, the draft survey was sent to 5 experts for feedback on the usability and face validity of the survey: 2 experts in the field of measurement tools in health care, 1 expert in the field of ATs in health care, and 2 specialized physical therapists. The feedback required no changes to the survey and the final survey was constructed (Multimedia Appendix 1).

Data collection was performed in May 2021. The survey was completed via clickable link, and QR code in the email leading to the survey in QuestBack Essentials (2021). For each question, a digital page emerged; in total, there were between 33 and 59 pages excluding the start- and the end page and 7 questions regarding population characteristics. All questions were mandatory questions and needed to be completed before the survey automatically continued. One reminder was sent after 1 week. Participants were asked to answer the questions as accurate as possible, and it was explicitly addressed that the results were analyzed anonymously. Data were stored in a secured data repository of Zuyd University of Applied Sciences for 15 years, and only the participating researchers had access to these data.

Data Analyses

Data analyses were performed using SPSS Statistics (version 23.0; IBM Crop). Only fully completed surveys were analyzed. Descriptive statistics of the participants' characteristics were presented as a number (percentage) for the categorical variables sex, workplace setting, and whether they completed a recent COPD course. For the continuous variables, age, work experiences, and the number of treated unique patients with COPD per week were presented as a mean (SD). Answers to questions with a 7-point Likert scale were presented as mean (SD). Responses to open comment questions were treated as qualitative data and categorized into themes by using directed content analyses via inductive coding [23] and presented as a number (percentage) per theme. Open answers that did not fit any code were labeled as other. Responses were separately analyzed for users and nonusers of ATs. In case of any missing data, listwise deletion was applied.

Ethical Considerations

No ethics approval was required since the research was not subject to the Medical Research Involving Human Subjects Act (Wet medisch-wetenschappelijk onderzoek met mensen). The research was part of continuous quality monitoring and improvement of daily COPD care within a national quality network (Chronisch ZorgNet). This study did not require written informed consent. Participating physical therapists were informed about the purpose of this survey prior to commencing the survey. Physical therapists voluntarily participated in this study. Study results were processed anonymously. Physical therapists were always able to end the survey without any consequences. There was no compensation for the physical therapists. The CHERRIES (Checklist for Reporting Results of Internet E-Surveys) was used (Checklist 1).

Results

Population Characteristics

In total, 216 physical therapists completed the survey. Five physical therapists were excluded because they did not meet the inclusion criteria (treating less than 1 unique patient with COPD per week), leaving 211 responses for analysis. There were no missing data. The characteristics of the physical therapists are displayed in Table 1.

Table . Characteristics of the participating physical therapist (N=211).

Characteristic	Participants
Sex (male), n (%)	74 (35.1)
Age (years), mean (SD)	46.5 (11.7)
Work experience (years), mean (SD)	22.8 (11.1)
Work setting, n (%)	
Primary care	197 (93.5)
Secondary care	5 (2.5)
Primary and secondary care	8 (3.8)
Tertiary care	1 (0.5)
Number of unique patients with COPD[a] per week, mean (SD)	8.2 (6.5)
Completed a recent COPD course, n (%)	208 (98)

[a]COPD: chronic obstructive pulmonary disease.

Results of the survey are presented per research questions.

How ATs Are Used

Of the 211 participating physical therapists, 108 (51.2%) used an AT, whereas most of them (n=82, 75.9%) already used an

XSL•FO
RenderX

AT before it was advised in the guidelines. Most physical therapists measured the number of steps (n=64, 55.2%) and used a standard app on a smartphone (n=97, 89.8%). Other parameters named were, for example, activities of daily living, distance, physical capacity, and heart rate. The 2 most common goals for the use of an AT were to inventory the physical activity level (n=94, 87%) and to stimulate the physical activity level of the patient (n=88, 81.5%). All results on how ATs were used in physical therapy in patients with COPD are presented in Table 2.

XSL•FO
RenderX

Table . Results on how activity trackers (ATs) are currently used in physical therapy in patients with chronic obstructive pulmonary disease (COPD; N=211; based on digital survey).

Question	User (n=108)	Nonuser (n=103)
I use AT in people with COPD, n (%)	• Yes: 108 (100)[a]	• No: 103 (100)[a]
Before the release of the COPD guidelines, I already used AT in people with COPD, n (%)	• Yes: 82 (75.9) • No: 26 (24.1)	—[b]
It is clear what I want to measure with an AT[c], mean (SD)	• 5.8 (1.3)	—
Which parameter or concept do you measure with an AT? n (%)	• Number of steps: 64 (55.2) • Physical activity: 43 (37.1) • Other[d]: 9 (7.7)	—
It is clear to me with which AT I can measure these parameters[c], mean (SD)	• 5.3 (1.5)	—
There are agreements within our center about the choice of the used type of AT[c], mean (SD)	• 3.2 (1.9)	—
Within our center, we use the following type of AT, n (%)	• App on a smartphone: 97 (89.8) • Commercially available pedometer: 57 (52.7) • Commercially available accelerometer: 14 (12.9) • Already owned by patient: 2 (1.8)	—
Within our center, we use the following brand AT, n (%)	• Standard app on a smartphone: 26 (24.1) • Unknown brand: 23 (21.3) • Already owned by patient: 14 (12.9) • Fitbit: 12 (11.1) • Omron: 8 (7.4) • Other[d]: 7 (6.4) • Yamax: 3 (2.7) • McRobberts: 2 (1.8)	—
The AT is worn by people with COPD on the following locations, n (%)	• Trouser pocket: 88 (81.4) • Wrist: 52 (48.1) • Hip: 13 (12) • Bag: 13 (12) • Other[d]: 5 (4.6) • Chest pocket: 4 (3.7) • Lower back: 2 (1.8) • Ankle: 2 (1.8)	—
Which brand of AT would you recommend to your patients? n (%)	• Do not know: 49 (45.3) • Apps: 23 (21.2) • Fitbit: 21 (19.2) • I do not recommend a brand: 6 (5.5) • Garmin: 5 (4.6) • Other: 5 (4.6) • Apple: 4 (3.7)	—
I stimulate my patients to purchase an AT[c], mean (SD)	• 4.9 (1.4)	—
At this moment, in how many unique patients with COPD per week are you using an AT? mean (SD)	• 5.5 (4.2)	—
I use AT for the following goals in patients with COPD, n (%)	• Inventory: 94 (87) • Stimulation: 88 (81.5) • Allocating: 64 (59.2) • Evaluation: 60 (55.5)	—

XSL•FO
RenderX

Question	User (n=108)	Nonuser (n=103)
My patients with COPD wear an AT for the recommended number of days (7 days)[c], mean (SD)	• 4.4 (1.5)	—

[a]Of the 211 respondents, 108 (51.2%) answered “yes” and 103 (38.8) answered “no.”

[b]Not available.

[c]Likert scale with a range of 0 - 7 (0=totally disagree and 7=totally agree).

[d]Multiple open answers which were categorized as other.

Why ATs Are Used

Physical therapists who used ATs because they believed that it is an added value to health care (n=46, 42.5%) and because the guidelines recommend to use ATs (n=44, 40.7%). Physical therapists considered the cutoff value of 5000 steps per day for patients with COPD to be a realistic measure of sufficient activity (mean 4.7, SD 1.5). The primary reason physical therapists did not use ATs both among users (n=58, 53.7%) and nonusers (n=35, 33.9%) was that patients themselves indicated that they did not want to use them. In the nonusers group, lack of knowledge (n=31, 30.9%) was the second important reason for nonuse. Other reasons mentioned were, for example, bad experiences, difficulty implementing ATs, and concerns about the clinometric properties of the ATs.

In most centers, there was no consensus on why they were going to use ATs (mean 3.4, SD 1.8). Users scored high on questions regarding whether ATs contributed to the assessment of physical activity levels (mean 5.8, SD 1.1), allocation of patients into profiles (mean 5.8, SD 1.1), evaluation of physical activity levels (mean 5.9, SD 0.9), and stimulation of physical activity levels (mean 5.9, SD 1.0). Results on why ATs were used in physical therapy in patients with COPD are presented in Table 3.

XSL•FO
RenderX

Table . Results on why activity trackers (ATs) are currently used in physical therapy in patients with chronic obstructive pulmonary disease (COPD; N=211; based on digital survey).

Question	User (n=108)	Nonuser (n=103)
Within our center, we have agreed on why we (are going to) use AT[a], mean (SD)	• 3.4 (2.0)	• 3.2 (1.9)
The choice for the concerned AT was made based on, n (%)	• Availability: 73 (67.6) • Costs: 49 (45.3) • Feasibility: 43 (39.8) • Own experience: 19 (17.5) • Validity: 15 (13.8) • Already owned by the center: 5 (4.6) • Other[b]: 6 (5.5%)	—[c]
I want to measure the following specific activities, n (%)	• Walking: 105 (97.2) • Bicycling: 52 (48.1) • Running: 11 (10.1) • Activities of daily living: 8 (7.4) • Swimming: 7 (6.4) • Other[b]: 5 (4.6) • Walking stairs: 2 (1.8)	—
I want to measure the following parameters, n (%)	• Number of steps: 104 (96.2) • Active minutes: 73 (67.5) • Walked distance: 53 (49) • Heart rate: 35 (32.4) • Passive minutes: 34 (31.4) • Calories: 13 (11.1) • Other[b]: 3 (2.7)	—
I think the cutoff value of 5000 steps per day to be sufficiently active is a realistic cutoff value[a], mean (SD)	• 4.7 (1.5)	—
The most important reason I am using an AT in patients with COPD is, n (%)	• Because it is an added health care value: 46 (42.5) • Because of the guidelines: 44 (40.7) • Own interest: 9 (8.3) • Other[b]: 5 (4.6) • Because it is mandatory in my network: 4 (3.7)	—
When I use an AT, it contributes to the inventory of the physical activity level and the physiotherapeutic diagnosis of a patient[a], mean (SD)	• 5.8 (1.1)	—
When I use an AT, it contributes to allocating patients into the profiles of the guidelines[a], mean (SD)	• 5.8 (1.1)	—
When I use an AT, it contributes to evaluating the physical activity during the treatment process in patients with COPD[a], mean (SD)	• 5.9 (0.9)	—
When I use an AT, it contributes to stimulating physical activity as part of my intervention[a], mean (SD)	• 5.9 (1.0)	—
I think AT should be used in patients with COPD[a], mean (SD)	• 5.2 (1.7)	• 4.3 (1.7)
My patients with COPD think it is important to (start) using AT[a], mean (SD)	• 4.1 (1.2)	• 3.9 (2.0)
My patients with COPD are more aware of their functioning by using an AT[a], mean (SD)	• 5.7 (1.1)	• 5.8 (1.4)

XSL•FO
RenderX

Question	User (n=108)	Nonuser (n=103)
Health care centers, comparable to my center are using AT[a], mean (SD)	• 4.2 (0.9)	• 3.3 (1.2)
External organizations are obligating the use of AT[a], mean (SD)	• 3.4 (1.8)	• 2.6 (1.6)
The most important reason I do not use AT in some patients with COPD is, n (%)	• Patients do not want to: 58 (53.7) • No added value: 13 (12) • No knowledge: 9 (8.3) • No AT available: 8 (7.4) • Costs: 7 (6.4) • Other[b]: 6 (5.5)	• Patients do not want to: 35 (33.9) • No knowledge: 31 (30.9) • No AT available: 9 (8.7) • Bad experience: 8 (7.7) • Other[b]: 8 (7.7) • No added value: 7 (6.7) • Costs: 7 (6.7)

[a]Likert scale with a range of 0 - 7 (0=totally disagree and 7=totally agree).

[b]Multiple open answers which were categorized as other.

[c]Not available.

Barriers and Facilitators for Implementation of ATs

No remarkable differences exist between users and nonusers regarding the experienced barriers and facilitators. For both groups, a relevant barrier to using ATs was the costs (users: n=75, 69.4% vs nonusers: n=77, 74.7%) and the cognitive or communication skills of the patient (users: n=75, 69.4% vs nonusers: n=50, 48.5%). Other reasons that were mentioned were prioritizing the patient over scientific pursuits, privacy concerns, and AT fails to measure parameters that therapists find meaningful. Important facilitators in both groups for using ATs were the motivation of the patient (users: n=90, 83.3% vs nonusers: n=70, 67.9%) and the user-friendliness of the AT (users: n=64, 59.2% vs nonusers: n=57, 55.3%). Both users (mean 3.6, SD 1.8) and nonusers (mean 2.5, SD 1.6) indicated that they lack the knowledge to provide a recommendation of commercially available ATs to their patients. Likewise, both groups indicated insufficient education or coaching possibilities for the health care professionals on ATs within their center (users: mean 3.7, SD 1.9 and nonusers: mean 2.7, SD 1.8). Further barriers and facilitators are presented in Table 4.

XSL•FO
RenderX

Table . Results on relevant barriers and facilitators for optimal implementation of activity trackers (ATs) during the clinical reasoning process of Dutch physical therapists in patients with chronic obstructive pulmonary disease (COPD; N=211; based on digital survey).

Question	User (n=108)	Nonuser (n=103)
I have sufficient knowledge to recommend commercially available AT to my patients regarding price, functions, and quality[a], mean (SD)	• 3.6 (1.8)	• 2.5 (1.6)
I am capable to draft personalized physical activity goals together with my patient[a], mean (SD)	• 5.8 (1.0)	• 5.8 (1.3)
I am capable to teach my patients about the functions, use, and interpretations of AT[a], mean (SD)	• 5.4 (1.3)	• 4.9 (1.8)
ATs are provided to our patients, n (%)	• No: 71 (65.7) • Yes: 37 (34.3)	• No: 99 (96.1) • Yes: 4 (3.9)
The following factors stimulate the use of AT during the clinical reasoning in patients with COPD, n (%)	• Motivation for the patient: 90 (83.3) • User-friendliness: 64 (59.2) • The added value: 59 (54.6) • Reliability and validity: 38 (35.1) • Costs: 28 (25.9) • Time investment: 25 (23.1)	• Motivation for the patient: 70 (67.9) • User-friendliness: 57 (55.3) • The added value: 54 (52.4) • Reliability and validity: 40 (38.8) • Costs: 25 (24.2) • Time investment: 18 (17.4) • Other[b]: 5 (4.8)
The following factors obstruct the use of AT during clinical reasoning in patients with COPD, n (%)	• Costs: 75 (69.4) • Cognitive or communication skills of the patient: 75 (69.4) • Time investment for the therapist: 37 (34.2) • User-friendliness: 36 (33.3) • Time investment for the patient: 34 (31.4) • Reliability and validity: 26 (24) • Other[b]: 3 (2.7)	• Costs: 77 (74.7) • Cognitive or communication skills of the patient: 50 (48.5) • Time investment for the therapist: 46 (44.6) • User-friendliness: 45 (43.6) • Reliability and validity: 40 (38.8) • Time investment for the patient: 40 (38.8) • Other[b]: 7 (6.7)
I know how I can measure physical activity with an AT[a], mean (SD)	• 5.5 (1.4)	—[c]
I know how I can interpret the results of the AT[a], mean (SD)	• 5.4 (1.2)	—
I know how I can discuss the results of the AT with my patients[a], mean (SD)	• 5.8 (1.0)	—
The use of AT is part of our mission and vision and is incorporated into our policy[a], mean (SD)	• 5.2 (2.0)	• 3.8 (2.3)
The use of AT fits in our care process[a], mean (SD)	• 5.6 (1.2)	• 4.6 (1.9)
There is sufficient support and involvement from the management when using AT[a], mean (SD)	• 5.2 (1.9)	• 4.1 (1.9)
In our center, there is someone who guides the implementation process of AT[a], mean (SD)	• 2.7 (2.0)	• 2.2 (1.7)
In our center, there is sufficient education or coaching for the health care professionals about AT[a], mean (SD)	• 3.7 (1.9)	• 2.7 (1.8)
In our center, there is sufficient time to try to use AT[a], mean (SD)	• 4.0 (1.9)	• 3.4 (1.7)
In our center, the use of AT is incorporated in our electronic patients' files[a], mean (SD)	• 4.5 (2.2)	• 3.4 (2.3)
In our center, there are sufficient AT available, n (%)	• No: 63 (58.3) • Yes: 24 (22.2) • Not applicable: 20 (18.5)	• No: 67 (65) • Yes: 5 (4.8) • Not applicable: 21 (20.3)

XSL•FO
RenderX

Question	User (n=108)	Nonuser (n=103)
I am open to (start to) using AT in patients with COPD[a], mean (SD)	• 6.5 (0.8)	• 5.8 (1.4)
I am planning to continue or start using AT in patients with COPD[a], mean (SD)	• 6.4 (0.8)	• 5.2 (1.5)
I have sufficient knowledge to (start to) use AT[a], mean (SD)	• 5.5 (1.3)	• 4.3 (1.7)
I trust I can use AT in patients with COPD[a], mean (SD)	• 5.5 (1.2)	• 4.8 (1.7)
My colleagues are open to (start) using an AT[a], mean (SD)	• 5.5 (1.5)	• 5.2 (2.0)
My colleagues have sufficient skills and knowledge to (start) using AT[a], mean (SD)	• 5.1 (1.7)	• 4.7 (2.2)
My patients with COPD are open to (start) using AT[a], mean (SD)	• 4.6 (1.1)	• 4.0 (1.8)
My patients with COPD have sufficient skills and knowledge to (start) using an AT[a], mean (SD)	• 4.0 (1.4)	• 3.6 (2.0)
My patients with COPD experience commercially available AT as technical complex[a], mean (SD)	• 5.1 (1.3)	• 5.7 (1.8)
In our center, the implementation of AT was prepared and planned[a], mean (SD)	• 3.5 (2.3)	• 2.7 (2.4)
In our center, we started using AT on a trial basis and started to adapt our care process[a], mean (SD)	• 4.2 (1.9)	• 2.2 (1.5)
In our center, relevant persons were involved in the implementation process: health care professionals, patients (representatives), and supporting professionals (policy)[a], mean (SD)	• 3.4 (2.0)	• 2.9 (2.5)
In our center, the use of AT are frequently evaluated and improved[a], mean (SD)	• 3.0 (1.9)	• 2.9 (2.8)

[a]Likert scale with a range of 0 - 7 (0=totally disagree and 7=totally agree).
[b]Multiple open answers which were categorized as other.
[c]Not available.

Discussion

Principal Findings

This research aimed to assess how and why ATs are currently used in physical therapy in patients with COPD and to determine which barriers and facilitators are of relevance for optimal implementation of ATs during the clinical reasoning process of Dutch physical therapists.

It is known that the implementation of an innovation is often challenging [11]. However, of the 211 participating physical therapists, 108 (51.2%) used an AT, and most of them (n=82, 75.9%) already used an AT before it was advised in the guidelines. This relatively high percentage could be explained by the fact that physical therapists indicated that the most important reason for using ATs is the experienced added health care value. Another explanation could be that the included physical therapists were affiliated with Chronisch ZorgNet and might therefore be more sensitized to provide evidence-based care. Furthermore, it was notable that regardless whether a physical therapist was a user or nonuser, the barriers and facilitators for using ATs were roughly the same. Both users and nonusers indicated that the most important reason for not using ATs is because their patients do not want to use an AT. However, it is debatable whether this is an assumption of the physical therapist or the actual opinion of the patient, as a recent review describes that patients with COPD in general enjoyed using ATs and found using the ATs easy [24]. The second (nonusers) and third (users) reason for physical therapists to not use ATs was a general lack of knowledge. Physical therapists indicated that they did not know which brand of commercially available AT they could recommend to their patients, which is related to the experienced lack of knowledge regarding price, functions, and quality. In contradiction, knowledge regarding how to measure with ATs, how to interpret the data of the AT, and how to discuss the data of the AT with a patient was scored

XSL•FO
RenderX

high by the physical therapists. A possible explanation could be that therapists lack different knowledge: the knowledge on how to measure with ATs, how to interpret the data of the AT, and how to discuss the data of the AT.

Comparison to Prior Work

The general adaption rate of eHealth in physical therapy seems to be similar to the adaption rate of 52% found in this study [25,26]. However, the use of the ATs in this study was self-reported; therefore, the presence of biases in the reported use and perception of ATs might be present. Most physical therapists used an app on a smartphone as an AT, which links to the fact that availability and costs were important factors when selecting an AT. Most smartphones automatically measure step count, and apps are mostly free of use. However, the validity of apps regarding measuring physical activity is low, especially in older adults during low walking speed, as is the case in activities of daily living [27]. Most physical therapists considered the cutoff value of 5000 steps per day to be sufficiently active, used in the new national guidelines, to be realistic. This is valuable information, considering that this cutoff value in the revised guidelines is based on previous studies. Several studies mention a minimum of 7500 - 10,000 steps per day in order to have a healthy active lifestyle [28-30]. However, it seems that a healthy person reaches only 5500 - 6000 steps per day on average, and people with a chronic disease 3500 - 5500 steps per day [29-31]. Consequently, in the national guidelines, a cutoff value of 5000 steps per day is used for people with COPD to be sufficiently active [32,33]. This study suggests that 5000 steps per day might be a minimal clinical important threshold in daily physical therapy practice.

Both users and nonusers indicated that the implementation of ATs was not prepared and planned for within their center. To successfully implement eHealth, internal and external facilitation, audit and feedback, management support, and training of clinicians are essential according to a recent systematic review [34] and are in line with the CFIR [17]. Wilde et al [35] similarly showed the physical therapists' need for guidance and information to support the integration of ATs within the practice. A solution could be to provide education and training to management and physical therapists. Management should be trained to guide an implementation trajectory; however, the use of implementation support practitioners could also be an option [36]. Implementation support practitioners are professionals who support organizations, leaders, and staff in their implementation of evidence-informed practices and policies [37]. Current physical therapists and physical therapy students should be educated and trained. To optimize the adoption of ATs in clinical settings, educational interventions should be developed based on the best available evidence, clinical expertise, and patient values and circumstances [38].

However, it is important to train physical therapists' and students' general eHealth competencies in order to use eHealth in a broader spectrum than ATs only. It is suggested that education consists of 5 competency themes: information communication technology attitudes and skills, interpretation and analysis of eHealth data, support and guidance, communication skills, and privacy and confidentiality [39]. Current health care professionals are insufficiently trained in all of these competencies and therefore struggle to perform these skills in their daily clinical practice [40]. A better understanding and more digital competency could also lead to better experiences and better outcomes. On the other hand, a previous study showed that patients with chronic diseases (including COPD) experience ATs as technically complicated and wish to receive support by their physical therapist during the use within a diagnostic and/or therapeutic process [22]. Currently, several (digital) tools and websites are available to support health care professionals in the goal-oriented selection of eHealth [41,42].

Strengths and Limitations

Several limitations of this study need to be mentioned. First, this study addressed technology acceptance on a limited basis. The level of technology acceptance might influence whether ATs are used or not. Studies showed that if health care professionals are already experienced eHealth users, they report fewer implementation barriers and experience more advantages (eg, a more positive attitude toward eHealth) [43]. Second, participating physical therapists were affiliated with Chronisch ZorgNet, an organization in which physical therapists treating patients with chronic diseases are trained to work evidence-based. This might have biased the results since affiliated physical therapists might be more prone to adhere to recent guidelines and be motivated to implement innovations. On the other hand, all participating physical therapists were members of Chronisch ZorgNet, which includes almost all physical therapists who are specialized in lung physical therapy. Therefore, the study sample constitutes a representative sample for primary care. A strength of this study is that the survey was based on relevant international scientific literature and models in relation to implementation and behavioral change (CFIR and TPB), which enables generalization beyond the Dutch context [5,13-18,22,44].

Conclusions

Overall, these results show that ATs are not yet fully implemented in the Dutch general physical therapy practice in patients with COPD, as recommended by current evidence-based guidelines. Both users and nonusers indicated that the most important reason for not using ATs is because their patients do not want to or at least the health care professionals experienced it that way. However, it is debatable whether this is an assumption of the physical therapist or the actual opinion of the patient. Physical therapists need guidance for the successful implementation of ATs. This could be accomplished by providing training for current physical therapists, integrating ATs into the education of (future) physical therapists, and providing support during the implementation process of ATs for both the physical therapists and management.

XSL•FO
RenderX

Acknowledgments

The authors would like to thank Chronisch ZorgNet for their assistance and coordination in data collection and all participating physical therapists for their responses. The authors thank Anita Stevens for her expertise on adapting questionnaire Q2. The authors declare that no generative artificial intelligence was used in writing this manuscript.

Data Availability

The datasets generated and analyzed during this study are not publicly available due to the restrictions of the quality network Chronisch ZorgNet but are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Questionnaire for both users and nonusers.
[DOCX File, 35 KB - formative_v9i1e59533_app1.docx]

Checklist 1
CHERRIES (Checklist for Reporting Results of Internet E-Surveys).
[DOCX File, 32 KB - formative_v9i1e59533_app2.docx]

References

1. Global strategy for the diagnosis management, and prevention of chronic obstructive pulmonary disease. Global Initiative for Chronic Obstructive Lung Disease. 2023. URL: https://goldcopd.org/2023-gold-report-2/ [accessed 2025-02-04]
2. Celli B, Fabbri L, Criner G, et al. Definition and nomenclature of chronic obstructive pulmonary disease: time for its revision. Am J Respir Crit Care Med 2022 Dec 1;206(11):1317-1325. [doi: 10.1164/rccm.202204-0671PP] [Medline: 35914087]
3. COPD [Web page in Dutch]. VZinfo. 2022. URL: https://www.vzinfo.nl/copd [accessed 2023-01-09]
4. COPD health care costs [Web page in Dutch]. VZinfo. 2019. URL: https://www.vzinfo.nl/copd/zorgkosten#:~:text=Zorgkosten%20voor%20COPD%2075%20miljoen%20euro%20in%202018&text=De%20meeste%20%C3%B3%20uit%20zorgkosten%20voor%20de%20dwangbehandelingen [accessed 2024-03-27]
5. Vreeken HL, Beekman E, Doornmaal M, Post MHL, Meerhoff GA, Spruit MA. COPD [Web page in Dutch]. KNGF. 2020. URL: https://www.kennisplatformfysiotherapie.nl/richtlijnen/copd/a-algemene-informatie/ [accessed 2025-02-04]
6. Ummels D, Beekman E, Braun SM, Beurskens AJ. Using an activity tracker in healthcare: experiences of healthcare professionals and patients. Int J Environ Res Public Health 2021 May 12;18(10):34066296. [doi: 10.3390/ijerph18105147] [Medline: 34066296]
7. Qiu S, Cai X, Wang X, et al. Using step counters to promote physical activity and exercise capacity in patients with chronic obstructive pulmonary disease: a meta-analysis. Ther Adv Respir Dis 2018;12:1753466618787386. [doi: 10.1177/1753466618787386] [Medline: 29993339]
8. Chiauzzi E, Rodarte C, DasMahapatra P. Patient-centered activity monitoring in the self-management of chronic health conditions. BMC Med 2015 Apr 9;13:77. [doi: 10.1186/s12916-015-0319-2] [Medline: 25889598]
9. Karapanos E, Gouveia R, Hassenzahl M, Forlizzi J. Wellbeing in the making: peoples' experiences with wearable activity trackers. Psychol Well Being 2016;6:4. [doi: 10.1186/s13612-016-0042-6] [Medline: 27376017]
10. Richard AA, Shea K. Delineation of self-care and associated concepts. J Nurs Scholarsh 2011 Sep;43(3):255-264. [doi: 10.1111/j.1547-5069.2011.01404.x] [Medline: 21884371]
11. Wouters M, Huygens M, Voogdt H, et al. E-health monitor: samen aan zet [Web page in Dutch]. Nivel. 2019. URL: https://www.nivel.nl/nl/publicatie/samen-aan-zet-ehealth-monitor-2019 [accessed 2025-02-04]
12. Taylor J, Coates E, Wessels B, Mountain G, Hawley MS. Implementing solutions to improve and expand telehealth adoption: participatory action research in four community healthcare settings. BMC Health Serv Res 2015 Dec 1;15:529. [doi: 10.1186/s12913-015-1195-3] [Medline: 26626564]
13. Schreiweis B, Pobiruchin M, Strotbaum V, Suleder J, Wiesner M, Bergh B. Barriers and facilitators to the implementation of eHealth services: systematic literature analysis. J Med Internet Res 2019 Nov 22;21(11):e14197. [doi: 10.2196/14197] [Medline: 31755869]
14. Ross J, Stevenson F, Lau R, Murray E. Factors that influence the implementation of e-health: a systematic review of systematic reviews (an update). Implement Sci 2016 Oct 26;11(1):146. [doi: 10.1186/s13012-016-0510-7] [Medline: 27782832]
15. Scott Kruse C, Karem P, Shifflett K, Vegi L, Ravi K, Brooks M. Evaluating barriers to adopting telemedicine worldwide: a systematic review. J Telemed Telecare 2018 Jan;24(1):4-12. [doi: 10.1177/1357633X16674087]

XSL•FO
RenderX

16. Granja C, Janssen W, Johansen MA. Factors determining the success and failure of eHealth interventions: systematic review of the literature. J Med Internet Res 2018 May 1;20(5):e10235. [doi: 10.2196/10235] [Medline: 29716883]
17. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. Implementation Sci 2009 Dec;4(1):50. [doi: 10.1186/1748-5908-4-50]
18. Ajzen I. The theory of planned behavior. Organ Behav Hum Decis Process 1991 Dec;50(2):179-211. [doi: 10.1016/0749-5978(91)90020-T]
19. Chronish ZorgNet [Website in Dutch]. URL: https://www.chronischzorgnet.nl/ [accessed 2025-02-04]
20. Beurskens A, Swinkels R, Wittink H, Peppen R, Vermeulen H. Meten in de Praktijk: Bohn Stafleu van Loghum; 2020. [doi: 10.1007/978-90-368-2459-0]
21. Stevens JGA. Ready for Goal Setting?: Maastricht University; 2017.
22. Ummels D, Beekman E, Moser A, Braun SM, Beurskens AJ. Patients' experiences with commercially available activity trackers embedded in physiotherapy treatment: a qualitative study. Disabil Rehabil 2020 Nov;42(23):3284-3292. [doi: 10.1080/09638288.2019.1590470] [Medline: 30973026]
23. Elo S, Kyngäs H. The qualitative content analysis process. J Adv Nurs 2008 Apr;62(1):107-115. [doi: 10.1111/j.1365-2648.2007.04569.x] [Medline: 18352969]
24. Wilde LJ, Sewell L, Percy C, Ward G, Clark C. What are the experiences of people with COPD using activity monitors?: A qualitative scoping review. COPD 2022;19(1):88-98. [doi: 10.1080/15412555.2022.2033192] [Medline: 35132933]
25. Rausch AK, Baur H, Reicherzer L, et al. Physiotherapists' use and perceptions of digital remote physiotherapy during COVID-19 lockdown in Switzerland: an online cross-sectional survey. Arch Physiother 2021 Jul 7;11(1):18. [doi: 10.1186/s40945-021-00112-3] [Medline: 34233763]
26. Miller MJ, Pak SS, Keller DR, Gustavson AM, Barnes DE. Physical therapist telehealth delivery at 1 year into COVID-19. Phys Ther 2022 Nov 6;102(11):36124699. [doi: 10.1093/ptj/pzac121] [Medline: 36124699]
27. Ummels D, Beekman E, Theunissen K, Braun S, Beurskens AJ. Counting steps in activities of daily living in people with a chronic disease using nine commercially available fitness trackers: cross-sectional validity study. JMIR Mhealth Uhealth 2018 Apr 2;6(4):e70. [doi: 10.2196/mhealth.8524] [Medline: 29610110]
28. Hancock C. The benefits of regular walking for health, well - being and the environment. : C3 Collaborating for Health; 2012 URL: https://image.guardian.co.uk/sys-files/Guardian/documents/2012/11/02/C3-report-on-walking.pdf [accessed 2025-02-05]
29. Lee IM, Shiroma EJ, Kamada M, Bassett DR, Matthews CE, Buring JE. Association of step volume and intensity with all-cause mortality in older women. JAMA Intern Med 2019 Aug 1;179(8):1105-1112. [doi: 10.1001/jamainternmed.2019.0899] [Medline: 31141585]
30. Tudor-Locke C, Bassett DR. How many steps/day are enough? Sports Med 2004;34(1):1-8. [doi: 10.2165/00007256-200434010-00001]
31. Tudor-Locke CE, Myers AM. Methodological considerations for researchers and practitioners using pedometers to measure physical (ambulatory) activity. Res Q Exerc Sport 2001 Mar;72(1):1-12. [doi: 10.1080/02701367.2001.10608926] [Medline: 11253314]
32. Depew ZS, Novotny PJ, Benzo RP. How many steps are enough to avoid severe physical inactivity in patients with chronic obstructive pulmonary disease? Respirology 2012 Aug;17(6):1026-1027. [doi: 10.1111/j.1440-1843.2012.02207.x] [Medline: 22672739]
33. Tudor-Locke C, Craig CL, Thyfault JP, Spence JC. A step-defined sedentary lifestyle index: <5000 steps/day. Appl Physiol Nutr Metab 2013 Feb;38(2):100-114. [doi: 10.1139/apnm-2012-0235] [Medline: 23438219]
34. Varsi C, Solberg Nes L, Kristjansdottir OB, et al. Implementation strategies to enhance the implementation of eHealth programs for patients with chronic illnesses: realist systematic review. J Med Internet Res 2019 Sep 27;21(9):e14255. [doi: 10.2196/14255] [Medline: 31573934]
35. Wilde LJ, Percy C, Clark C, Ward G, Wark PA, Sewell L. Views and experiences of healthcare practitioners supporting people with COPD who have used activity monitors: "More than just steps". Respir Med 2023 Nov;218:107395. [doi: 10.1016/j.rmed.2023.107395] [Medline: 37633422]
36. Albers B, Metz A, Burke K. Implementation support practitioners—a proposal for consolidating a diverse evidence base. BMC Health Serv Res 2020 May 1;20(1):368. [doi: 10.1186/s12913-020-05145-1] [Medline: 32357877]
37. Burke K, Albers B, Ward C. Implementation support practitioner profile guiding principles and core competencies for implementation practice. : University of North Carolina at Chapel Hill; 2020.
38. Components of evidence-based practice. American Physical Therapy Association. 2020 Mar 23. URL: https://www.apta.org/patient-care/evidence-based-practice-resources/components-of-evidence-based-practice#:~:text=Evidence-based%20practice%20includes%20the,three%20elements%20are%20equally%20important [accessed 2024-10-16]
39. Barakat A, Woolrych RD, Sixsmith A, Kearns WD, Kort HS. eHealth technology competencies for health professionals working in home care to support older adults to age in place: outcomes of a two-day collaborative workshop. Med 20 2013;2(2):e10. [doi: 10.2196/med20.2711]

XSL•FO
RenderX

40. Wentink MM, Siemonsma PC, van Bodegom-Vos L, et al. Teachers' and students' perceptions on barriers and facilitators for eHealth education in the curriculum of functional exercise and physical therapy: a focus groups study. BMC Med Educ 2019 Sep 6;19(1):343. [doi: 10.1186/s12909-019-1778-5] [Medline: 31492129]
41. Meten met ehealth [Website in Dutch]. 2024. URL: https://metenmetehealth.nl [accessed 2025-02-05]
42. Beweegtech [Website in Dutch]. 2024. URL: https://beweegtech.nl/ [accessed 2025-02-05]
43. Ariens LF, Schussler-Raymakers FM, Frima C, et al. Barriers and facilitators to eHealth use in daily practice: perspectives of patients and professionals in dermatology. J Med Internet Res 2017 Sep 5;19(9):e300. [doi: 10.2196/jmir.7512] [Medline: 28874336]
44. Spruit MA, Van't Hul A, Vreeken HL, et al. Profiling of patients with COPD for adequate referral to exercise-based care: the Dutch model. Sports Med 2020 Aug;50(8):1421-1429. [doi: 10.1007/s40279-020-01286-9] [Medline: 32333261]

Abbreviations

AT: activity tracker
CFIR: Consolidated Framework for Implementation Research
CHERRIES: Checklist for Reporting Results of Internet E-Surveys
COPD: chronic obstructive pulmonary disease
KNGF: Koninklijk Nederlands Genootschap voor Fysiotherapie
TPB: theory of planned behavior

Edited by A Mavragani; submitted 15.04.24; peer-reviewed by MV Ittersum, Y Chu; revised version received 19.12.24; accepted 19.12.24; published 12.02.25.

Please cite as:
Ummels D, Bols E, Frantzen RJA, Frantzen T, Robeerts L, Beekman E
Activity Trackers in Physical Therapy for People With Chronic Obstructive Pulmonary Disease in the Netherlands: Cross-Sectional Study on Current Use and Implementation Determinants
JMIR Form Res 2025;9:e59533
URL: https://formative.jmir.org/2025/1/e59533
doi:10.2196/59533

© Darcy Ummels, Esther Bols, Roel Johannes Anna Frantzen, Tim Frantzen, Levy Robeerts, Emmylou Beekman. Originally published in JMIR Formative Research (https://formative.jmir.org), 12.2.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Understanding Providers' Attitude Toward AI in India's Informal Health Care Sector: Survey Study

Sumeet Kumar, PhD; Snehil Rayal, BTech; Raghuram Bommaraju, PhD; Navya Pratyusha Varasala, MA; Sirisha Papineni, MD; Sarang Deo, PhD

Indian School of Business, Gachibowli, ISB Road, Hyderabad, India

Corresponding Author:
Sumeet Kumar, PhD
Indian School of Business, Gachibowli, ISB Road, Hyderabad, India

Abstract

Background: Tuberculosis (TB) is a major global health concern, causing 1.5 million deaths in 2020. Diagnostic tests for TB are often inaccurate, expensive, and inaccessible, making chest x-rays augmented with artificial intelligence (AI) a promising solution. However, whether providers are willing to adopt AI is not apparent.

Objective: The study seeks to understand the attitude of Ayurveda, Yoga and Naturopathy, Unani, Siddha, and Homoeopathy (AYUSH) and informal health care providers, who we jointly call AIPs, toward adopting AI for TB diagnosis. We chose to study these providers as they are the first point of contact for a majority of TB patients in India.

Methods: We conducted a cross-sectional survey of 406 AIPs across the states of Jharkhand (162 participants) and Gujarat (244 participants) in India. We designed the survey questionnaire to assess the AIPs' confidence in treating presumptive TB patients, their trust in local radiologists' reading of the chest x-ray images, their beliefs regarding the diagnostic capabilities of AI, and their willingness to adopt AI for TB diagnosis.

Results: We found that 93.7% (270/288) of AIPs believed that AI could improve the accuracy of TB diagnosis, and for those who believed in AI, 71.9% (194/270) were willing to try AI. Among all AIPs, 69.4% (200/288) were willing to try AI. However, we found significant differences in AIPs' willingness to try AI across the 2 states. Specifically, in Gujarat, a state with better and more accessible health care infrastructure, 73.4% (155/211) were willing to try AI, and in Jharkhand, 58.4% (45/77) were willing to try AI. Moreover, AIPs in Gujarat who showed higher trust in the local radiologists were less likely to try AI (odds ratio [OR] 0.15, 95% CI 0.03 - 0.69; P=.02). In contrast, in Jharkhand, those who showed higher trust in the local radiologists were more likely to try AI (OR 2.11, 95% CI 0.9 - 4.93; P=.09).

Conclusions: While most AIPs believed in the potential benefits of AI-based TB diagnoses, many did not intend to try AI, indicating that the expected benefits of AI measured in terms of technological superiority may not directly translate to impact on the ground. Improving beliefs among AIPs with poor access to radiology services or those who are less confident of diagnosing TB is likely to result in a greater impact of AI on the ground. Additionally, tailored interventions addressing regional and infrastructural differences may facilitate AI adoption in India's informal health care sector.

(JMIR Form Res 2025;9:e54156) doi:10.2196/54156

KEYWORDS

artificial intelligence; tuberculosis; health care providers; cross-sectional studies; trust; x-rays; India

Introduction

Tuberculosis (TB) remains a significant global health challenge, with over 80% of reported cases and deaths originating from low- and middle-income countries (LMICs) worldwide [1]. Among these countries, India shoulders a substantial burden, accounting for a quarter of all TB cases and resulting in approximately 89,000 deaths in the year 2019 alone [2]. The COVID-19 pandemic further worsened these global inequalities, particularly by disrupting TB diagnostic and treatment services [3,4]. Data from the World Health Organization (WHO) reveals a concerning trend in the incidence rate of TB. After experiencing a decline of around 2% per year over the past 2 decades, there has been a 3.6% increase in new TB cases per 100,000 population between 2020 and 2021, as indicated in the Global Tuberculosis Report [5]. Additionally, due to factors like underreporting, underdiagnosis, and limited access to health care for TB patients, a significant gap exists between the estimated and reported number of cases [6]. In India, 29% of TB cases were either undiagnosed or missed in 2021 [7,8]. This resurgence of TB cases [9] is particularly alarming for LMICs like India, which bear a disproportionate disease burden [10] and face numerous challenges [6] in their fragile health care supply chains.

XSL•FO
RenderX

To lessen this burden, recent advancements in diagnosis based on molecular testing and chest x-ray (CXR) scans offer promising avenues. However, the highly accurate molecular diagnostic tests for TB, such as Xpert MTB/RIF (*Mycobacterium tuberculosis*/rifampin), remain expensive, difficult to access, and challenging to maintain in LMICs [11]. Additionally, their roll-out has been slow [12]. Although WHO recommends molecular diagnostics as the preferred first-line testing method, only 38% of individuals diagnosed with TB in 2021 were tested with a WHO-recommended rapid molecular diagnostic at the initial diagnosis [12]. The other option of clinical diagnosis based on CXRs continues to be the mainstay among health care providers in India [13]. CXR is recommended by WHO for screening, triaging, and assisting in the diagnosis of TB. It is also featured in the national TB diagnostic guidelines of 19 out of 22 countries with the highest TB burden [14]. However, there is a shortage of adequately trained radiologists proficient in analyzing CXRs and producing high-quality reports [15]. This shortage has important implications, as the majority of TB care providers are not formally trained to interpret CXRs [16].

In the absence of qualified radiologists, the presence of automated artificial intelligence (AI) systems for interpreting CXRs could prove to be highly advantageous [17,18]. Employing AI-assisted interpretation of CXRs can improve the speed [19] and enhance the accuracy of TB diagnosis [20]. Addressing the delays in TB diagnosis through earlier TB detection is a key element of the WHO's TB control and elimination strategy [21]. Moreover, the more recent deep learning–based AI systems for CXR interpretation have demonstrated on par or higher specificity than radiologists [22-24], and their usage can significantly reduce the cost of TB screening [25]. These systems are also portable and can additionally be accessed remotely via the internet, making them a good choice for providers who often serve in remote locations.

However, the successful adoption of AI for the accurate diagnosis of TB depends on many factors [26]. In particular, adoption by the private sector is needed, as it plays a significant role in treating TB cases in India, diagnosing twice as many cases as compared to the public sector, amounting to an estimated 2.2 million TB cases [27]. The private sector is fragmented, consisting of a variety of health care providers, clinics, and practitioners that lack integration, especially in small towns and villages [28]. A study conducted [29] revealed that the average Indian village has 3.2 primary health care providers, with 68% lacking formal medical training and referred to as informal health care providers (IPs). In addition, India has Ayurveda, Yoga and Naturopathy, Unani, Siddha, and Homoeopathy (AYUSH) providers formally trained in one of the 6 traditional Indian systems of medicine, constituting 22.8% of formally trained medical practitioners in India [30]. From this point forward, we will use the abbreviation "AIPs" to collectively denote both informal health care providers and AYUSH practitioners. For faster and more accurate TB diagnosis, AIPs need to adopt AI systems. However, the adoption will only happen if AIPs perceive value in AI, which has not been examined in prior research.

In this paper, we conducted a survey study to understand the attitudes and perspectives of AIPs toward adopting AI for TB diagnosis. AIPs are the first point of contact in patient pathways for the majority of TB cases in India, so their willingness to incorporate AI-enhanced CXR systems into the decision-making is a key factor in driving AI adoption. Adopting AI could enhance AIPs' capability to interpret CXR scans and detect TB cases earlier and more accurately. In order to assess AIPs' beliefs in the diagnostic capabilities of AI and their willingness to adopt AI for TB diagnosis, we conducted a cross-sectional survey of 406 AIPs across the states of Jharkhand (162 participants) and Gujarat (244 participants).

Methods

Overview

We conducted a cross-sectional survey of AIPs in 2 Indian states (Gujarat and Jharkhand) to understand their current practices and potential barriers in adopting AI in their practice. The guidelines in the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) were followed (Checklist 1).

Jharkhand is one of the poorer states in India with lower technology penetration and lower access to radiologic facilities as well as radiologists, and overall a less developed private sector. In contrast, Gujarat is a more developed state with an active and accessible private health care infrastructure.

Participants Recruitment and Enrollment

We collaborated with World Health Partner (WHP), a non-profit organization that has partnered with x-ray labs and AIPs in Gujarat and Jharkhand. In Gujarat, WHP worked with 73 x-ray labs and 757 AIPs, while in Jharkhand, they collaborated with 113 x-ray labs and 406 AIPs.

Multiple methods were used to identify the AIPs. Field investigators from WHP mapped the AIPs in the districts of Ranchi and East Singhbhum in Jharkhand, as well as Surat and Gandhinagar in Gujarat. Investigators queried TB patient cohorts (those in their first year of treatment) about the health care providers they visited during their treatment. Allopathic providers were interviewed to identify nearby AIPs who frequently referred cases to them. Additionally, government officials and partner organizations at the community level helped identify local health care provider networks. Radiology units offering CXR services were also consulted to locate AIPs in their vicinity. The investigators used the mobile Commcare application to map these providers, capturing details such as contact information, qualifications, outpatient department load, referral linkages, and more, which were then used to generate a unique ID for each provider (form included in Multimedia Appendix 1). Each AIP clinic functioned as an independent study or clinical setting.

For our study, we asked all identified AIPs working with WHP for their participation and enrolled 406 AIPs based on their willingness to participate in the survey. In return, we provided a mobile recharge of 300 (US $3.47) for completing the survey. These participants were from 18 blocks across 2 districts in Gujarat and 22 blocks spanning 2 districts in Jharkhand. The survey was conducted between February 2022 and September 2022.

XSL•FO
RenderX

Data

We developed the initial questionnaire by leveraging questions from established scales and modifying them to suit our study's context. This process included adapting sections on AI in TB diagnosis from existing literature [31,32] and customizing questions to assess AIPs' comfort with technology and demographics [33]. Additionally, we adapted from the literature [34] the questions regarding AIPs' trust in radiologists, ensuring they were pertinent to our specific research objectives.

The questionnaire underwent 7 iterative revisions in collaboration with our field partner, WHP, known for its extensive experience with TB providers and patients. Subsequently, Outline India, the data collection agency, conducted pretests that led to further refinements based on feedback from pilot studies. This phase involved translating the questionnaire into Hindi and Gujarati and making 8 additional revisions to improve clarity and relevance. Our research team oversaw the translation process to ensure its accuracy. The final, refined version of the questionnaire, which incorporated comprehensive feedback and amendments, received validation by the strategy lead at WHP.

We structured the questionnaire into multiple subsections to evaluate the following aspects: AIPs' understanding of TB diagnosis; confidence in diagnosing TB patients; trust levels in local radiologists; beliefs and intentions regarding the adoption of AI in TB diagnosis; comfort levels with technology; and demographic details such as age, gender, state of residence, and years of experience. We conducted a survey of the registered AIPs by partnering with Outline India. This agency deployed field investigators proficient in the local languages (Hindi and Gujarati) and used specialized software for digital data collection.

We gathered data from 406 participating AIPs. The survey was conducted at the participants' locations and lasted 10-20 minutes. Out of 406 AIPs, 39.9% (162/406) participants were from Jharkhand and 60.1% (244/406) participants were from Gujarat. In total, 85.2% (346/406) of the respondents were men, while women comprised 14.8% (60/406). The age of respondents ranged from 22 to 86 (mean 41.8, SD 11.4) years.

Responses to most of the questions were on a 5-point Likert scale as shown in Table 1. To illustrate the use of this scale, a score of "4" indicates that the participant somewhat agrees with the statement under consideration, whereas a score of "5" indicates strong agreement with the statement. For a few questions, respondents' opinions were recorded as 1=Yes and 0=No using a dichotomous scale. The detailed questionnaire can be referenced in Multimedia Appendix 2.

Table . Likert Scale of 0 - 5 capturing responses from "strongly disagree" to "strongly agree".

Scale	Respondent's opinion to question
5	Strongly agree
4	Somewhat agree
3	Neither agree nor disagree
2	Somewhat disagree
1	Strongly disagree

We encountered certain absent (not available) data points, stemming from incomplete surveys. These data points were eliminated before employing the data for regression, resulting in the removal of 92 observations. Out of these, 17 were from Gujarat, and 75 were from Jharkhand. The main reason for the reduction in the number of observations was the lack of responses for questions on "Trust in local radiologists" because many survey participants did not have access to radiologists.

In addition, the outliers were identified and removed using boxplot and Cook distance analysis to eliminate any highly influential data points that could distort the results. This process resulted in a final data set consisting of 288 observations, with 211 observations from Gujarat and 77 observations from Jharkhand. We show the entire workflow for data cleaning in Figure 1. Multimedia Appendix 3 contains further information on outlier analysis and on descriptor variables for both the initial sample collected and the final sample used in the regression analysis.

XSL•FO
RenderX

Figure 1. Data cleaning workflow reflecting multiple steps to arrive at the final dataset.

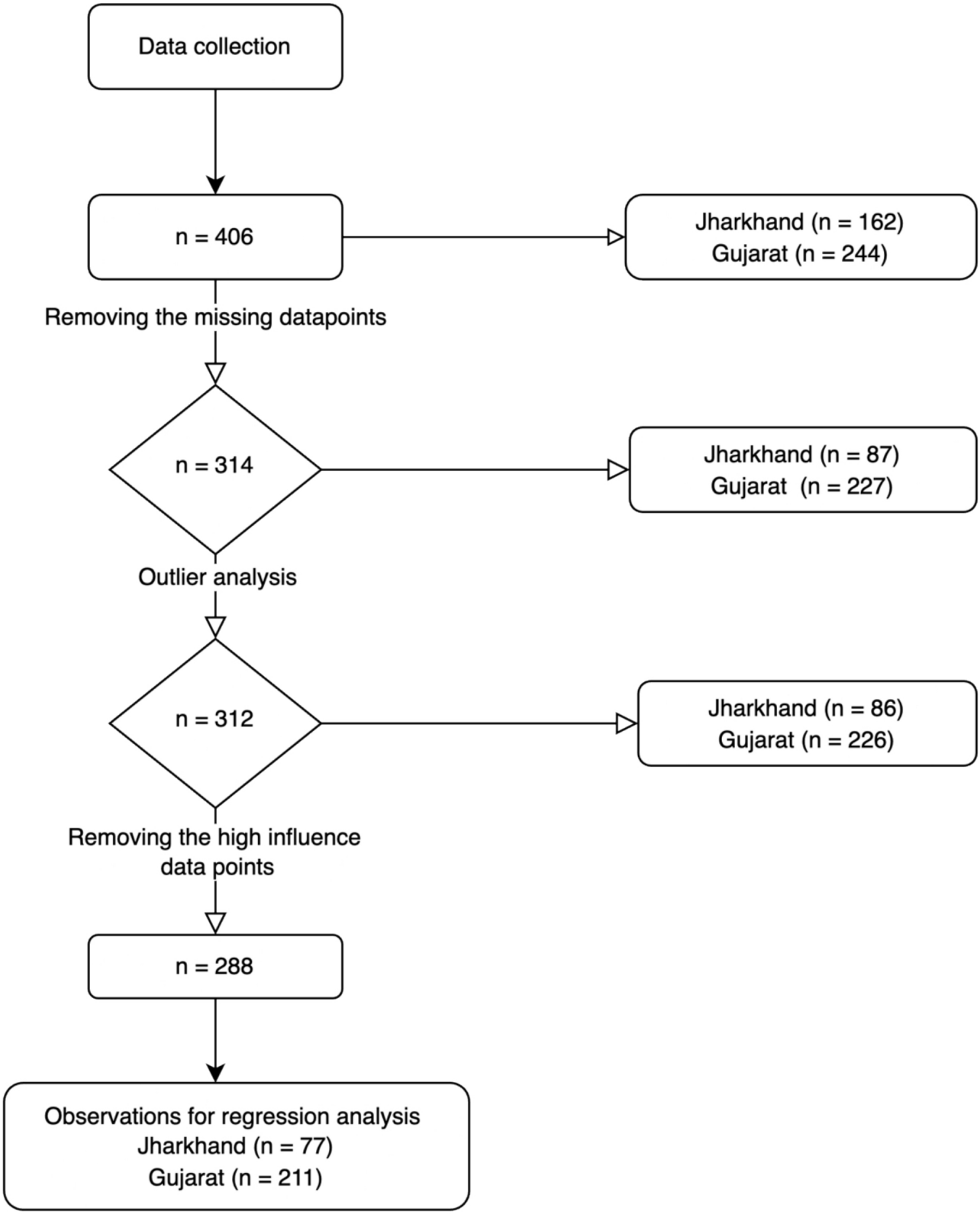

Variable Description and Analysis

In this study, we investigated two main constructs: (1) AI Intent and (2) AI Belief. Previous research on technology adoption has demonstrated that the actual utilization of a technology is preceded by factors such as "Intent to use," "Attitude toward technology," and considerations of "Ease of use" among others [35,36]. The ease of use of a product often evolves with the incorporation of new features and regular enhancements to the user experience, a common occurrence in the realm of internet technology products. Hence, we specifically focused on understanding individuals' "Intent to use" AI products, which we refer to as "AI Intent"; and their "Attitude toward AI," which we refer to as "AI Belief."

XSL•FO
RenderX

We coded the response to the survey question, “Are you willing to try an AI software that could generate an x-ray report from an uploaded x-ray?” as a categorical variable “AI Intent” with 2 options: “Yes” or “No.” We used logistic regression to examine its association with the independent variables.

We created the “AI Belief” variable by combining the responses to 2 questions: “I believe AI algorithms can reliably detect imaging findings that suggest TB” and “I believe the use of AI could improve the accuracy of diagnosis of tuberculosis.” We calculated the mean of the responses to these questions to construct the “AI Belief” variable. It represents participants’ overall belief in the effectiveness of AI in improving TB diagnosis.

For independent variables in the regression analysis, we included “Confidence in diagnosing TB” and “Trust in local radiologists,” and demographic variables such as age, gender, and state of residence.

We coded “Confidence in diagnosing TB” as the mean of the responses to 5 questions from the survey questionnaire: “I feel confident in my ability to read an x-rays report to detect a TB case,” “I feel confident in my ability to interpret an x-rays film without an x-rays report to detect a TB case,” “I feel confident in my ability to deal with TB patients (if allowed to treat or medicate),” “I feel confident in my ability to use relevant diagnostic tests to detect a TB case,” and “I feel confident in my ability to collect relevant health-related information from patients.”

Similarly, we coded “Trust in local radiologists” as the mean of the responses to 3 questions: “The language and style of radiology reports are mostly clear,” “The radiologist who reads the x-rays and writes the report is competent,” and “I am generally satisfied with x-rays reports I receive from radiologists.” We calculated the Cronbach α using all the survey responses to assess the consistency of the scale. Gender and state of residence were treated as categorical.

We used Python (version 3.9.13) for data cleaning and data visualization, while data analysis was carried out using StataCorpSTATA SE (version 17.0) and Microsoft Excel.

Ethical Considerations

The study protocols underwent review and approval by the Institutional Review Board at the Indian School of Business (application number ISB-IRB 2022 - 04) and were approved on January 28, 2022. Informed consent was obtained from the participants before recording their responses. Information provided to participants included the purpose of the study, what participants will be asked to do if they participate, how much of the subjects’ time participation may take, and what type of information they might be asked to provide. It was also made clear that participation is voluntary, and subjects may withdraw from the study at any time. The data were deidentified to protect participants’ privacy. The consent form can be referenced in Multimedia Appendix 4.

Results

Descriptive Statistics

We found that over 90% of AIPs believed in the benefits of AI, with 93.8% (270/288) responding positively to “AI Improves Accuracy of TB Diagnosis,” and 90.9% (262/288) responding positively to “AI Can Detect Image Findings.” We also found that 69.4% (200/288) of AIPs were willing to try AI. For those who believed in AI, 71.9% (194/270) were willing to try AI.

Of the final 288 observations used for regression analysis, 73.3% (211/288) were from Gujarat, and 27.7% (77/288) were from Jharkhand. In the sample, 11.1% (32/288) were women, and 88.9% (256/288) were men with ages between 22 and 68 (mean 40.7, SD 9.7) years. See Table 2 for the summary statistics of the data.

Table . Summary statistics for each variable in regression analysis, for Jharkhand and Gujarat, February-September 2022.

	Jharkhand (n=77)[a]				Gujarat (n=211)[b]			
Variable	Mean	SD	Min	Max	Mean	SD	Min	Max
Age (years)	40.87	10.92	24	67	40.60	9.47	22	68
AI[c] Intent	0.58	0.50	0	1	0.73	0.44	0	1
AI Belief	4.39	0.71	2.5	5	4.59	0.55	3	5
Confidence in diagnosing TB[d]	4.41	0.46	3	5	4.50	0.50	2.4	5
Trust in local radiologists	4.56	0.62	1.33	5	4.85	0.34	3.33	5

[a]Of the 77 respondents, 61 (79.2%) were male and 16 (20.8) were female.

[b]Of the 211 respondents, 195 (92.4%) were male and 16 (7.6%) were female.

[c]AI: artificial intelligence.

[d]TB: tuberculosis.

The mean score for “Confidence in diagnosing TB” was similar in both states, 4.50 (SD 0.50) for Gujarat and 4.41 (SD 0.46)

XSL•FO
RenderX

for Jharkhand. The mean score for "Trust in local radiologists" was higher among AIPs in Gujarat (4.85, SD 0.34) than in Jharkhand (4.56, SD 0.62).

The Cronbach α value for the questions measuring "Confidence in diagnosing TB" was 0.65, while it was 0.82 for "Trust in the local radiologist" and 0.79 for "AI Belief."

AI Intent

The results of the logistic regression analysis of "AI Intent" on predictor variables are shown in Table 3. We found that "Confidence in diagnosing TB" was positively associated for both states but statistically significant only for Gujarat (odds ratio [OR] 1.943, 95% CI 1.038-3.637; *P*=.04) and not for Jharkhand (OR 2.186, 95% CI 0.695-6.875; *P*=.18).

Table . Regression results for dependent variable "AI Intent", for Jharkhand and Gujarat, February-September 2022.

	Jharkhand (n=77)				Gujarat (n=211)			
Predictor variables	Odds ratio	SE	*P* value	95% CI	Odds ratio	SE	*P* value	95% CI
Age (years)	0.964	0.025	.15	0.916-1.014	0.990	0.016	.52	0.959-1.021
Gender	2.573	1.874	.20	0.617-10.728	2.073	1.252	.23	0.635-6.771
Confidence in diagnosing TB[a]	2.186	1.278	.18	0.695-6.875	1.943	0.622	.04	1.038-3.637
Trust in local radiologists	2.107	0.914	.09	0.900-4.932	0.149	0.117	.02	0.032-0.691

[a]TB: tuberculosis.

"Trust in local radiologists" was positively associated with "AI Intent" for AIPs in Jharkhand (OR 2.107, 95% CI 0.9-4.932; *P*=.09) but negatively associated for AIPs in Gujarat (OR 0.149, 95% CI 0.032 - 0.691; *P*=.02).

AI Belief

The results of the multivariate linear regression (ordinary least squares) analysis are shown in Table 4. First, we observed that "Confidence in diagnosing TB" was positively associated with "AI Belief" in both Jharkhand (β=0.72, *P*<.001) and Gujarat (β=0.177, *P*=.008). Further, "Trust in local radiologists" was also positively associated with "AI Belief" in both Jharkhand (β=0.215, *P*=.02) and Gujarat (β=0.66, *P*<.001).

Table . Regression results for dependent variable "AI Belief", for Jharkhand and Gujarat, February-September 2022.

	Jharkhand (n=77)				Gujarat (n=211)			
Predictor variables	β	SE	*P* value	95% CI	β	SE	*P* value	95% CI
Age (years)	−0.007	0.006	.24	−0.020 to 0.005	0.002	0.004	.64	−0.005 to 0.009
Gender	0.166	0.196	.40	−0.225 to 0.557	0.021	0.105	.84	−0.186 to 0.227
Confidence in diagnosing TB[a]	0.720	0.137	<.001	0.447 to 0.993	0.177	0.066	.008	0.047 to 0.308
Trust in local radiologists	0.215	0.089	.02	0.037 to 0.392	0.660	0.105	<.001	0.453 to 0.866

[a]TB: tuberculosis.

We found that none of the demographic factors were significantly associated with "AI Belief."

Discussion

Principal Findings

The integration of AI with CXRs holds promise as a cost-effective and accurate solution for TB diagnosis. However, despite this potential, the adoption of AI tools among health care providers for TB diagnosis remains limited. In this study, we conducted a cross-sectional survey of AIPs in 2 Indian states (Gujarat and Jharkhand) to understand and assess their belief in AI; their willingness to adopt AI for improved diagnosis; and the association of these with the AIPs' confidence in diagnosing TB as well as trust in the local radiologists. We found that AIPs are mostly aware of AI and believe in AI, but the intent to adopt AI is lower and depends on AIPs' confidence in diagnosing TB and their trust in local radiologists. Previous studies that examined the perceptions of health care professionals about AI in the medical field have yielded mixed results [37,38]. For instance, Oh et al [37] found that physicians in Korea had a favorable attitude toward AI, considering it as a complementary

XSL•FO
RenderX

tool rather than a replacement for their roles. In contrast, Abdullah et al [38] discovered that health care employees in Saudi Arabia needed to be made aware of the benefits of AI and expressed concerns about being replaced by AI in their jobs.

While previous studies primarily focused on formal health care professionals [37-39], students [40], and the general public [41], our study contributes to understanding perceptions of AI among other medical professionals, specifically AIPs in India. Our findings indicate that most AIPs (over 90%) believed in AI's capabilities; however, a significant proportion (88/288, 30.6%) were unwilling to try AI. These findings align with previous research among physicians [42], where the intention to use AI was observed only when there was a belief in the role of AI. However, unlike in prior research [42], where the participants were medical college students who received AI education as part of their curriculum, AIPs are unlikely to have similar formal educational opportunities for exposure to AI. Therefore, AIPs would require additional support, for example, in terms of training programs to know about AI and subsequently try AI. Additionally, AI systems should strive to be more explainable to make it easier to believe in AI, for example, by providing underlying reasoning that leads to their output, which is also known as explainable AI. An explainable AI approach would assist AIPs in developing trust in the system's output and bolster their belief in AI.

Our study has uncovered a noteworthy association between AIPs' confidence in diagnosing TB and their willingness to adopt AI, which is somewhat surprising when considering prior research has indicated that highly confident clinicians are less likely to change their decision [43]. However, AIPs differ from general clinicians in a crucial aspect, as they lack formal training in reading x-rays films. Therefore, we posit that the confidence of AIPs is likely influenced by the availability of high-quality diagnostic inputs, such as access to microbiological tests and well-prepared radiology reports. While the contextual factors may be at play in our findings, this suggests that the overall impact of AI adoption on health care outcomes might be more restrained than earlier studies have suggested. This distinction arises from the fact that those choosing to adopt AI may already have access to a high-quality diagnostic infrastructure, and thus, their adoption of AI might not significantly boost health outcome metrics like case detection rates. In light of this insight, it is advisable for companies and health care organizations to not rely exclusively on mean performance differences between AI systems and radiologists when estimating the potential enhancements in health outcomes.

Moreover, we identified a substantial difference in the intention to adopt AI between the states of Gujarat and Jharkhand. A higher willingness to try AI was observed among respondents from Gujarat compared to their counterparts in Jharkhand. The observed disparity between the 2 states could be attributed to Gujarat being a more developed state in India, with higher levels of education and better health care infrastructure. The presence of a well-established, technology-enabled infrastructure in Gujarat may have provided first-hand experience of the benefits of technology in health care, leading to a stronger intent to try technological products like AI. This implies that, in the context of AIPs, states with better technology-based infrastructure are the ones likely to have higher adoption when AI-based systems are implemented on the ground.

Another factor that might have affected willingness to try AI was the unavailability of qualified radiologists in less developed states like Jharkhand. One would expect AIPs to be more willing to try AI systems in the absence of qualified radiologists. However, our findings indicate that AIPs in Jharkhand were less likely to try AI. The reasons for this attitude are not immediately apparent, but experiencing lower-quality radiology reports may bias AIPs into believing that CXR technology is unreliable. To overcome such barriers to adoption, implementing agencies will need to provide additional evidence of the benefits of AI and related technologies like CXR scans.

By integrating AI into the primary health care centers, which are already actively engaged in TB service provision, one can extend AI-powered TB diagnostics to grassroots levels [15]. While AIPs acknowledge the potential of AI to enhance TB diagnosis, concerns regarding the financial impact of such technology are significant. AI's integration into TB diagnosis through CXR screenings entails an additional cost, estimated at a modest 100 (approximately $1.3) per screening [44]. This cost, while relatively small, would be borne by AIPs and possibly transferred to patients. Considering India's per capita net income was 98,374 (US $1138.71) in 2022 - 23 [45], this expense may be manageable for a large population segment. However, affordability may pose a barrier in economically disadvantaged regions, potentially necessitating government intervention to facilitate access. The urgency of integrating AI in diagnostics must be contextualized within India's acute radiologist shortfall, with the current ratio standing at approximately one radiologist per 100,000 people, starkly below international norms [15]. This shortage underscores the broader necessity of AI deployment in augmenting health care capacity.

Our study has unveiled several important findings that are instrumental in understanding the landscape of AI in India's TB-related health care services. First and foremost, a significant fraction of AIPs have a positive attitude toward AI, with many expressing a readiness to integrate AI into their practice. However, we have identified a discernible gap between those who acknowledge AI's benefits and those who are prepared to embrace the technology. This discrepancy is further nuanced by a variation in attitudes based on providers' self-assessed diagnostic competencies, providers with greater confidence in their abilities tend to be more receptive to AI. Additionally, there is noticeable heterogeneity in the willingness to adopt AI across different geographic regions, which points to the influence of local health care infrastructure and socioeconomic factors on technology acceptance.

The implications of these findings are multifaceted and suggest several actionable strategies for enhancing AI adoption in health care settings. To translate the positive attitudes of AIPs into actual use of AI, tailored initiatives that address barriers to implementation are necessary. This might include providing more robust evidence of AI's effectiveness, integrating AI education into continuous professional development programs, and offering financial incentives. As diagnostic pathways in TB are majorly driven by the private sector [27], steps should be

XSL•FO
RenderX

taken to make the integration of AI in the TB care cascade profitable for providers to ensure its sustainability. Moreover, there is a compelling need to target AIPs with suboptimal diagnostic capabilities, as enhancements in this group's diagnostic services could positively impact overall health care outcomes.

To ensure these strategies are effective across India's diverse health care landscape, future research should replicate this study in multiple geographies. Such research should also examine the intricacies of trust and financial relationships between AIPs and radiologists to uncover deeper insights into the factors that influence AI adoption and to devise informed interventions that can bolster the technology's uptake.

Limitations

Like other survey-based studies, this study too has a few limitations. First, we had to remove many observations for the state of Jharkhand because many AIPs in Jharkhand did not respond to questions on "Trust in local radiologists." The lower response rate to these questions was because of AIPs' limited access to radiologists. To ascertain if the removal of such AIPs from the data impacted the findings, we compared the demographic variables of the dataset before and after the removal and did not find any statistically significant difference (more information in Multimedia Appendix 5).

Second, the responses in the questionnaire that measured opinions about AI may have been subjected to social desirability bias among AIPs. To mitigate this effect, we informed the participants about the anonymity of the survey and the separation of personal information from their responses.

Third, the AIPs' perception of AI in TB diagnosis may have been influenced by their overall opinion of AI and the extent of their knowledge of AI. To allow more informed responses, participants were apprised of the AI approach to recognize TB cases.

Conclusions

Our study provides insights into the attitudes of AYUSH and informal health care providers toward AI in TB diagnosis. It contributes to the understanding that adopting AI is not solely determined by a belief in AI but also relates to factors like confidence in diagnosing TB and other local factors like the quality of radiology services. Furthermore, regional variations between Gujarat and Jharkhand underscore the importance of tailored interventions to effectively integrate AI into diagnostic practices, with implications beyond TB diagnosis.

Acknowledgments

This research has been supported in part by the Max Institute of Healthcare Management (MIHM) at the Indian School of Business. Access to Ayurveda, Yoga and Naturopathy, Unani, Siddha, and Homoeopathy (AYUSH) practitioners and informal health care providers was facilitated by World Health Partners. We acknowledge a grant of 820,000 (US $9491.78) for the project titled "Effectiveness of AI for clinical diagnosis – Case of TB diagnosis in India." We also want to acknowledge the support of Dr Nisarg Desai for validating the survey questionnaire. We used the generative artificial intelligence tool ChatGPT by OpenAI in some parts of the paper for paraphrasing and checking grammar.

Data Availability

The datasets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

SD, SP, SK, and RB designed the study. NPV and SP coordinated with the survey agency to conduct the study and collect the data. SR and SK conducted the analysis and prepared figures and tables. SK, SR, RB, and SD wrote the paper. All authors have reviewed the final manuscript and provided critical feedback.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Application form to map Ayurveda, Yoga and Naturopathy, Unani, Siddha, and Homoeopathy practitioners and informal health care providers.
[PDF File, 29 KB - formative_v9i1e54156_app1.pdf]

Multimedia Appendix 2
Survey Questionnaire.
[PDF File, 650 KB - formative_v9i1e54156_app2.pdf]

Multimedia Appendix 3
Outlier analysis.

XSL•FO
RenderX

[PDF File, 292 KB - formative_v9i1e54156_app3.pdf]

Multimedia Appendix 4
Institutional review board application and approval documents.
[PDF File, 2160 KB - formative_v9i1e54156_app4.pdf]

Multimedia Appendix 5
Descriptor variables.
[PDF File, 262 KB - formative_v9i1e54156_app5.pdf]

Checklist 1
STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) checklist.
[PDF File, 117 KB - formative_v9i1e54156_app6.pdf]

References

1. Tuberculosis. WHO. 2023. URL: https://www.who.int/news-room/fact-sheets/detail/tuberculosis [accessed 2025-01-27]
2. Varshney K, Patel H, Kamal S. Trends in tuberculosis mortality across India: improvements despite the COVID-19 pandemic. Cureus 2023 Apr;15(4):e38313. [doi: 10.7759/cureus.38313] [Medline: 37261163]
3. McQuaid CF, Vassall A, Cohen T, Fiekert K, Covid/Tb Modelling Working Group, White RG. The impact of COVID-19 on TB: a review of the data. Int J Tuberc Lung Dis 2021 Jun 1;25(6):436-446. [doi: 10.5588/ijtld.21.0148] [Medline: 34049605]
4. Migliori GB, Thong PM, Alffenaar JW, et al. Gauging the impact of the COVID-19 pandemic on tuberculosis services: a global study. Eur Respir J 2021 Nov;58(5):2101786. [doi: 10.1183/13993003.01786-2021] [Medline: 34446465]
5. Global tuberculosis report. WHO. 2022. URL: https://iris.who.int/bitstream/handle/10665/363752/9789240061729-eng.pdf?sequence=1 [accessed 2025-01-27]
6. Chakaya J, Khan M, Ntoumi F, et al. Global Tuberculosis Report 2020 - Reflections on the Global TB burden, treatment and prevention efforts. Int J Infect Dis 2021 Dec;113 Suppl 1(Suppl 1):S7-S12. [doi: 10.1016/j.ijid.2021.02.107] [Medline: 33716195]
7. Ricks S, Singh A, Sodhi R, Pal A, Arinaminpathy N. Operational priorities for engaging with India's private healthcare sector for the control of tuberculosis: a modelling study. BMJ Open 2024 Mar 19;14(3):e069304. [doi: 10.1136/bmjopen-2022-069304] [Medline: 38508628]
8. India TB report 2022, coming together to end TB altogether. : Ministry of Health and Family Welfare URL: https://tbcindia.mohfw.gov.in/wp-content/uploads/2023/05/TBAnnaulReport2022.pdf [accessed 2024-10-29]
9. Natarajan A, Beena PM, Devnikar AV, Mali S. A systemic review on tuberculosis. Indian J Tuberc 2020 Jul;67(3):295-311. [doi: 10.1016/j.ijtb.2020.02.005] [Medline: 32825856]
10. Suttels V, Guedes Da Costa S, Garcia E, et al. Barriers and facilitators to implementation of point-of-care lung ultrasonography in a tertiary centre in Benin: a qualitative study among general physicians and pneumologists. BMJ Open 2023 Jun 26;13(6):e070765. [doi: 10.1136/bmjopen-2022-070765] [Medline: 37369423]
11. Pai M, Dewan PK, Swaminathan S. Transforming tuberculosis diagnosis. Nat Microbiol 2023 May;8(5):756-759. [doi: 10.1038/s41564-023-01365-3] [Medline: 37127703]
12. Reid M, Agbassi YJP, Arinaminpathy N, et al. Scientific advances and the end of tuberculosis: a report from the Lancet Commission on Tuberculosis. Lancet 2023 Oct;402(10411):1473-1498. [doi: 10.1016/S0140-6736(23)01379-X]
13. Chest radiography in tuberculosis detection. WHO. 2016. URL: https://iris.who.int/bitstream/handle/10665/252424/9789241511506-eng.pdf [accessed 2025-01-27]
14. Pande T, Pai M, Khan FA, Denkinger CM. Use of chest radiography in the 22 highest tuberculosis burden countries. Eur Respir J 2015 Dec;46(6):1816-1819. [doi: 10.1183/13993003.01064-2015]
15. Thakur A. Shortage of radiologists a concern, fostering research in AI can augment capabilities of radiologists: experts. The Indian Express. 2023. URL: https://indianexpress.com/article/cities/bangalore/shortage-radiologists-research-ai-capabilities-experts-8884446/ [accessed 2025-01-27]
16. Guidance document on ntep_ayush collaborative activities to end TB. TBC India. URL: https://tbcindia.gov.in/showfile.php?lid=3631 [accessed 2023-09-19]
17. Hansun S, Argha A, Liaw ST, Celler BG, Marks GB. Machine and deep learning for tuberculosis detection on chest x-rays: systematic literature review. J Med Internet Res 2023 Jul 3;25:e43154. [doi: 10.2196/43154] [Medline: 37399055]
18. Kotei E, Thirunavukarasu R. A comprehensive review on advancement in deep learning techniques for automatic detection of tuberculosis from chest x-ray images. Arch Computat Methods Eng 2024 Jan;31(1):455-474. [doi: 10.1007/s11831-023-09987-w]
19. Lakhani P, Sundaram B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology 2017 Aug;284(2):574-582. [doi: 10.1148/radiol.2017162326] [Medline: 28436741]

XSL•FO
RenderX

20. Rajpurkar P, O'Connell C, Schechter A, et al. CheXaid: deep learning assistance for physician diagnosis of tuberculosis using chest x-rays in patients with HIV. NPJ Digit Med 2020;3(1):115. [doi: 10.1038/s41746-020-00322-2] [Medline: 32964138]
21. Bhargava A, Bhargava M, Juneja A. Social determinants of tuberculosis: context, framework, and the way forward to ending TB in India. Expert Rev Respir Med 2021 Jul;15(7):867-883. [doi: 10.1080/17476348.2021.1832469] [Medline: 33016808]
22. Qin ZZ, Sander MS, Rai B, et al. Using artificial intelligence to read chest radiographs for tuberculosis detection: a multi-site evaluation of the diagnostic accuracy of three deep learning systems. Sci Rep 2019 Oct 18;9(1):15000. [doi: 10.1038/s41598-019-51503-3] [Medline: 31628424]
23. Codlin AJ, Dao TP, Vo LNQ, et al. Independent evaluation of 12 artificial intelligence solutions for the detection of tuberculosis. Sci Rep 2021 Dec 13;11(1):23895. [doi: 10.1038/s41598-021-03265-0] [Medline: 34903808]
24. Engle E, Gabrielian A, Long A, Hurt DE, Rosenthal A. Performance of Qure.ai automatic classifiers against a large annotated database of patients with diverse forms of tuberculosis. PLOS One 2020;15(1):e0224445. [doi: 10.1371/journal.pone.0224445] [Medline: 31978149]
25. Tzelios C, Nathavitharana RR. Can AI technologies close the diagnostic gap in tuberculosis? Lancet Dig Health 2021 Sep;3(9):e535-e536. [doi: 10.1016/S2589-7500(21)00142-4]
26. Kulkarni V, Gawali M, Kharat A. Key technology considerations in developing and deploying machine learning models in clinical radiology practice. JMIR Med Inform 2021 Sep 9;9(9):e28776. [doi: 10.2196/28776] [Medline: 34499049]
27. Arinaminpathy N, Batra D, Khaparde S, et al. The number of privately treated tuberculosis cases in India: an estimation from drug sales data. Lancet Infect Dis 2016 Nov;16(11):1255-1260. [doi: 10.1016/S1473-3099(16)30259-6]
28. Yellappa V, Lefèvre P, Battaglioli T, Devadasan N, Van der Stuyft P. Patients pathways to tuberculosis diagnosis and treatment in a fragmented health system: a qualitative study from a south Indian district. BMC Public Health 2017 Aug 4;17(1):635. [doi: 10.1186/s12889-017-4627-7] [Medline: 28778192]
29. Das J, Daniels B, Ashok M, Shim EY, Muralidharan K. Two Indias: the structure of primary health care markets in rural Indian villages with implications for policy. Soc Sci Med 2022 May;301:112799. [doi: 10.1016/j.socscimed.2020.112799] [Medline: 32553441]
30. The health workforce in india. WHO. 2016. URL: https://iris.who.int/bitstream/handle/10665/250369/9789241510523-eng.pdf?sequence=1 [accessed 2025-01-27]
31. Coppola F, Faggioni L, Regge D, et al. Artificial intelligence: radiologists' expectations and opinions gleaned from a nationwide online survey. Radiol Med 2021 Jan;126(1):63-71. [doi: 10.1007/s11547-020-01205-y]
32. Abuzaid MM, Elshami W, Tekin H, Issa B. Assessment of the willingness of radiologists and radiographers to accept the integration of artificial intelligence into radiology practice. Acad Radiol 2022 Jan;29(1):87-94. [doi: 10.1016/j.acra.2020.09.014]
33. Kelamane S, Satyanarayana S, Nagaraja SB, et al. Engaging informal private health care providers for TB case detection: experiences from RIPEND project in India. Tuberc Res Treat 2021;2021:9579167. [doi: 10.1155/2021/9579167] [Medline: 34239728]
34. Ignácio FDC, de Souza L, D'Ippolito G, Garcia MM. Radiology report: what is the opinion of the referring physician? Radiol Bras 2018;51(5):308-312. [doi: 10.1590/0100-3984.2017.0115] [Medline: 30369658]
35. Kemp A, Palmer E, Strelan P. A taxonomy of factors affecting attitudes towards educational technologies for use with technology acceptance models. Br J Educ Tech 2019 Sep;50(5):2394-2413. [doi: 10.1111/bjet.12833]
36. Rauniar R, Rawski G, Yang J, Johnson B. Technology acceptance model (TAM) and social media usage: an empirical study on Facebook. J Enterp Inf Manag 2014 Feb 4;27(1):6-30. [doi: 10.1108/JEIM-04-2012-0011]
37. Oh S, Kim JH, Choi SW, Lee HJ, Hong J, Kwon SH. Physician confidence in artificial intelligence: an online mobile survey. J Med Internet Res 2019 Mar 25;21(3):e12422. [doi: 10.2196/12422]
38. Abdullah R, Fakieh B. Health care employees' perceptions of the use of artificial intelligence applications: survey study. J Med Internet Res 2020 May 14;22(5):e17620. [doi: 10.2196/17620]
39. Nitiéma P. Artificial intelligence in medicine: text mining of health care workers' opinions. J Med Internet Res 2023 Jan 27;25:e41138. [doi: 10.2196/41138] [Medline: 36584303]
40. Teng M, Singla R, Yau O, et al. Health care students' perspectives on artificial intelligence: countrywide survey in Canada. JMIR Med Educ 2022 Jan 31;8(1):e33390. [doi: 10.2196/33390] [Medline: 35099397]
41. Beets B, Newman TP, Howell EL, Bao L, Yang S. Surveying public perceptions of artificial intelligence in health care in the United States: systematic review. J Med Internet Res 2023 Apr 4;25:e40337. [doi: 10.2196/40337] [Medline: 37014676]
42. Wagner G, Raymond L, Paré G. Understanding prospective physicians' intention to use artificial intelligence in their future medical practice: configurational analysis. JMIR Med Educ 2023 Mar 22;9:e45631. [doi: 10.2196/45631] [Medline: 36947121]
43. Dreiseitl S, Binder M. Do physicians value decision support? A look at the effect of decision support systems on physician opinion. Artif Intell Med 2005 Jan;33(1):25-30. [doi: 10.1016/j.artmed.2004.07.007]
44. qXR. India Health Fund. URL: https://www.indiahealthfund.org/our-portfolio/qxr/#:~:text=It%20is%20an%20affordable%20screening,100%20(~%241.3)%20per%20test [accessed 2025-01-27]

XSL•FO
RenderX

45. Income in india. Wikipedia. URL: https://en.wikipedia.org/wiki/Income_in_India [accessed 2025-01-27]

Abbreviations

AI: artificial intelligence
AIP: Ayurveda, yoga and naturopathy, Unani, Siddha, and homoeopathy practitioner and informal health care provider
AYUSH: Ayurveda, Yoga and Naturopathy, Unani, Siddha, and Homoeopathy
CXR: chest x-ray
LMIC: low- and middle-income country
MTB/RIF: *Mycobacterium tuberculosis*/rifampin
OR: odds ratio
TB: tuberculosis
WHO: World Health Organization
WHP: World Health Partner

Edited by A Mavragani; submitted 31.10.23; peer-reviewed by AK Madhukeshwar, K Varshney; revised version received 18.11.24; accepted 18.11.24; published 10.02.25.

Please cite as:
Kumar S, Rayal S, Bommaraju R, Varasala NP, Papineni S, Deo S
Understanding Providers' Attitude Toward AI in India's Informal Health Care Sector: Survey Study
JMIR Form Res 2025;9:e54156
URL: https://formative.jmir.org/2025/1/e54156
doi:10.2196/54156

© Sumeet Kumar, Snehil Rayal, Raghuram Bommaraju, Navya Pratyusha Varasala, Sirisha Papineni, Sarang Deo. Originally published in JMIR Formative Research (https://formative.jmir.org), 10.2.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Developing an ICD-10 Coding Assistant: Pilot Study Using RoBERTa and GPT-4 for Term Extraction and Description-Based Code Selection

Sander Puts[1], PhD; Catharina M L Zegers[1], PhD; Andre Dekker[1], PhD; Iñigo Bermejo[1,2], PhD

[1]Department of Radiation Oncology (Maastro), GROW Research Institute for Oncology and Reproduction, Maastricht University Medical Centre+, P.O. Box 616, Maastricht, Netherlands

[2]Data Science Institute (DSI), Hasselt University, Belgium

Corresponding Author:
Sander Puts, PhD
Department of Radiation Oncology (Maastro), GROW Research Institute for Oncology and Reproduction, Maastricht University Medical Centre+, P.O. Box 616, Maastricht, Netherlands

Abstract

Background: The International Classification of Diseases (ICD), developed by the World Health Organization, standardizes health condition coding to support health care policy, research, and billing, but artificial intelligence automation, while promising, still underperforms compared with human accuracy and lacks the explainability needed for adoption in medical settings.

Objective: The potential of large language models for assisting medical coders in the ICD-10 coding was explored through the development of a computer-assisted coding system. This study aimed to augment human coding by initially identifying lead terms and using retrieval-augmented generation (RAG)–based methods for computer-assisted coding enhancement.

Methods: The explainability dataset from the CodiEsp challenge (CodiEsp-X) was used, featuring 1000 Spanish clinical cases annotated with ICD-10 codes. A new dataset, CodiEsp-X-lead, was generated using GPT-4 to replace full-textual evidence annotations with lead term annotations. A Robustly Optimized BERT (Bidirectional Encoder Representations from Transformers) Pretraining Approach transformer model was fine-tuned for named entity recognition to extract lead terms. GPT-4 was subsequently employed to generate code descriptions from the extracted textual evidence. Using a RAG approach, ICD codes were assigned to the lead terms by querying a vector database of ICD code descriptions with OpenAI's text-embedding-ada-002 model.

Results: The fine-tuned Robustly Optimized BERT Pretraining Approach achieved an overall F_1-score of 0.80 for ICD lead term extraction on the new CodiEsp-X-lead dataset. GPT-4-generated code descriptions reduced retrieval failures in the RAG approach by approximately 5% for both diagnoses and procedures. However, the overall explainability F_1-score for the CodiEsp-X task was limited to 0.305, significantly lower than the state-of-the-art F_1-score of 0.633. The diminished performance was partly due to the reliance on code descriptions, as some ICD codes lacked descriptions, and the approach did not fully align with the medical coder's workflow.

Conclusions: While lead term extraction showed promising results, the subsequent RAG-based code assignment using GPT-4 and code descriptions was less effective. Future research should focus on refining the approach to more closely mimic the medical coder's workflow, potentially integrating the alphabetic index and official coding guidelines, rather than relying solely on code descriptions. This alignment may enhance system accuracy and better support medical coders in practice.

(JMIR Form Res 2025;9:e60095) doi:10.2196/60095

KEYWORDS

International Classification of Diseases; ICD-10; computer-assisted-coding; GPT-4; coding; term extraction; code analysis; computer assisted coding; transformer model; artificial intelligence; AI automation; retrieval-augmented generation; RAG; large language model; LLM; Bidirectional Encoder Representations from Transformers; Robustly Optimized BERT Pretraining Approach; RoBERTa; named entity recognition; NER

Introduction

Background and Significance

The International Classification of Diseases (ICD), developed by the World Health Organization (WHO), serves as a universal standard for coding health-related conditions [1]. The ICD system steers health care policymaking, assists in billing and reimbursement, underpins research, ensures quality monitoring, and promotes the standardization of medical information exchange. ICD-10 is divided into diagnosis and procedure codes,

XSL•FO
RenderX

encompassing over 100,000 codes. The system is hierarchically organized and includes detailed coding guidelines.

Medical coding entails the allocation of unique codes to medical records and is a standard procedure in most hospitals. Hospitals collect a list of ICD codes relevant to each patient's hospital admission. Medical coding is a time-consuming and error-prone task, which has led to interest in automating it, and in turn, to the emergence of the medical coding subfield within medical natural language processing (NLP) [1,2]. However, the task of effectively computerizing ICD coding remains a challenge. While the accuracy of codes assigned by medical coders may be debated [2-4], current state-of-the-art (SOTA) automated systems for ICD-10 coding still fall short of human performance [5-7]. In addition, many artificial intelligence (AI) solutions are "black boxes," not offering explanations for their predictions. However, model transparency is required to gain trust, which is crucial for its adoption in medical settings [8].

ICD Coding

Diagnosis codes in the ICD system are organized into an alphanumeric format across 21 chapters, capturing a range of categories including specific diseases, symptoms, and external causes, among others.

The hierarchical structure of ICD codes begins with a letter that corresponds to a chapter, together with the 2 subsequent digits that form the category of the code. Following the category code, the subclassification uses up to 5 additional alphanumeric characters. These detail specific aspects of the disease or condition, like laterality or severity, as illustrated in Figure 1. The procedure codes, separate from diagnostic codes, consist of 7 characters termed as "axis," with each axis representing a specific aspect of the medical procedure performed [9]. ICD-10 provides a comprehensive 3-volume coding manual that includes an alphabetical index. This index helps in finding medical codes using "lead terms," which are crucial keywords associated with medical conditions, listed in the alphabetical index. Medical coders first identify a diagnosis's lead term in the text (Figure 2). In this example, the lead term "cirrhosis" from the textual evidence corresponds to an alphabetic index entry. Under "Cirrhosis," the relevant subentry is "liver," "alcoholic," leading to the back-reference code K70.30. The next coding step involves consulting the tabular index for K70.30 for detailed instructions.

Figure 1. Structure and components of an ICD-10 diagnosis code. ICD-10: International Classification of Diseases, 10th version.

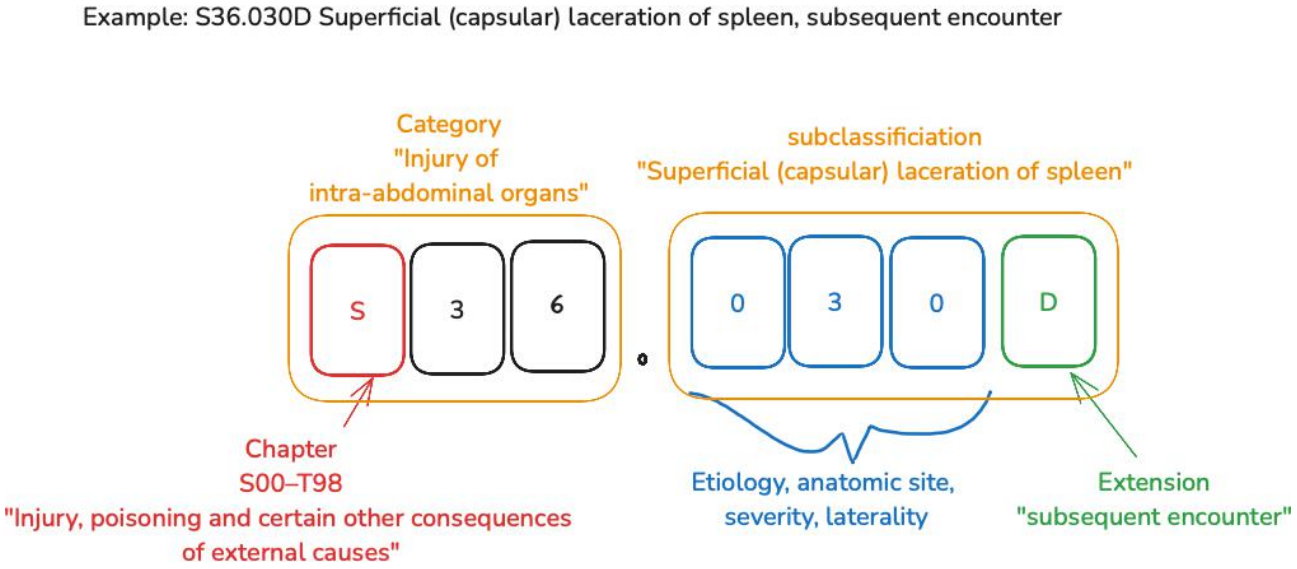

XSL•FO
RenderX

Figure 2. Illustration of concepts "code description," "textual evidence (green)," "lead terms (red)," and "back references" in the "alphabetical index". ICD: International Classification of Diseases.

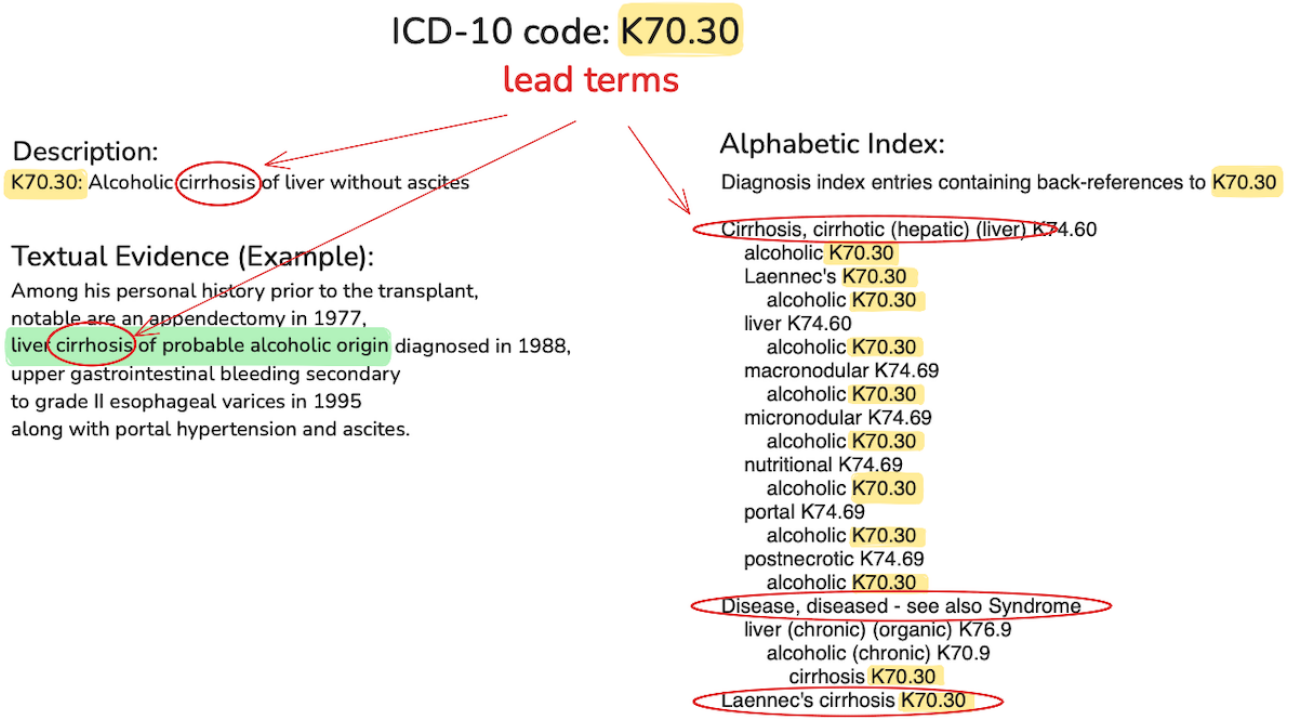

For each diagnosis and procedure that needs an ICD code, the initial step is to identify the lead term, which is comparable to the named entity recognition (NER) task in NLP. NER involves extracting specific categories like names, locations, medications, and diseases from unstructured text. Only a few studies proposing lead term-based approaches for ICD coding have been published [5], possibly due to the absence of public datasets annotated with lead terms. Previous NER approaches focused on extracting full textual evidence, rather than lead terms [5].

Explainability

Recent studies have proposed adversarial robustness training strategies and feature attribution, such as the use of attention mechanisms, for the explainability of NLP [10-12]. A limitation of these approaches is the absence of efforts to validate whether the identified "explainable" terms align with official ICD guidelines. For instance, while "insulin" may be a predictive term for the disease "diabetes", it is not suitable as evidence for coding according to the WHO guidelines and educational textbooks [13] as insulin can be prescribed for various medical conditions other than diabetes.

Transformers

In the original transformer architecture, which encompasses both encoder and decoder components [14], it is crucial to recognize the distinct roles of these elements. Encoder-only models, such as the BERT (Bidirectional Encoder Representations from Transformers) architecture and its variant, RoBERTa (Robustly Optimized BERT Pretraining Approach) [15], excel at capturing relevant information from the input data and constructing meaningful representations [16-18] and can be found in current SOTA models in ICD coding [5,7]. Conversely, decoder models, exemplified by GPT, are specialized in generating coherent and contextually relevant text [19]. With the evolution of the GPT architecture, remarkable advancements have been achieved in NLP [20,21]. GPT has already been applied across various health care disciplines and use cases [22]. However, early studies indicate that while GPT-3.5, GPT-4, and ChatGPT show potential, their performance in ICD classification has been mixed, with out-of-the-box negative results [23]. Generative approaches may be especially useful for ICD coding of rare diseases [24]. OpenAI's GPT was, for a time, the leading large language model (LLM) in generative AI [20], but other models, such as Mixtral, Gemini, Grok, Llama, and Claude [25], are available and quickly catching up [26,27].

Objective

Our objective was to develop a GPT-4-based computer-assisted coding system to aid medical coders in ICD-10 coding. Our approach enhances the identification of diagnoses and procedures in texts by initially focusing on lead terms rather than extracting the full textual evidence for a code at once. The reason for initially extracting lead terms is that it aligns perfectly with the first step followed by medical coders in their procedure. This approach would therefore be a good fit for assisting with the process. For code normalization, our approach exclusively employs code descriptions, moving away from reliance on model-based training. Specifically, the ICD code for identified lead terms was assigned using GPT-4 and ICD descriptions, employing a retrieval-augmented generation (RAG) approach. This reliance on an external knowledge source is strategically chosen to improve generalizability, enhancing the system's ability to accurately classify codes unseen during training. Furthermore, this method offers increased adaptability to changes in the coding system. The research also investigates the optimal method for querying a database to retrieve ICD codes by description. It compares the effectiveness of 2

XSL•FO
RenderX

approaches: using direct textual evidence as the query and employing code descriptions generated (invented) by GPT-4 from that textual evidence. Additionally, the study aimed to assess GPT-4's ability to accurately select the best code by matching code descriptions with textual diagnoses. This objective is rooted in the understanding that ICD coding guidelines and descriptions are designed for humans, and it seeks to determine if GPT-4's capabilities in following instructions translate to robust performance in this specific context.

Methods

GPT-4

This study used GPT-4 (gpt-4 1106-preview) via OpenAI's application programming interface and the Langchain Python library [11]. To ensure reproducibility, the model's temperature was set to 0. GPT-4's parametric knowledge, or out-of-the-box performance, was evaluated in 2 ways. First, it was prompted to generate descriptions for 100 random ICD codes from the CodiEsp dataset. These descriptions were then compared with the ground truth by prompting GPT-4 in a new chat with, "Do both descriptions refer to the same ICD code?" Second, GPT-4 was tasked with matching the correct ICD-10 code to each of the 100 official descriptions.

Dataset CodiEsp

CodiEsp is a unique corpus of 1000 clinical cases in Spanish, annotated by professionals for explainability, offering token-level annotations that provide textual evidence for each assigned code [28]. The annotation agreement rates were 88.6% for diagnosis, 88.9% for procedures, and 80.5% for textual evidence [28]. The test set contains 1514 unique ICD codes (1143 diagnoses and 371 procedures), some of which are not present in the training set, specifically, 439 diagnosis codes (38.3%) and 231 procedure codes (62.3%). Among the procedure codes, 143 (38.5%) are 4-axis ICD-10 procedure codes, which officially should have 7-axis ICD-10 codes. These incomplete codes lack descriptions and account for 602 of the total of 1112 procedure codes (54%). Overall, 828 out of 4777 (17.3%) test annotations do not have an exact match in the training set. In the current SOTA on CodiEsp-X, the entity extraction and normalization are both transformer-based [5].

Dataset CodiEsp-X-Lead

This approach does not use NER to extract complete textual evidence; instead, it mimics human coders by first identifying lead terms. An annotated dataset was required to train and evaluate a NER model for this task. However, manual annotation of 18,435 items was deemed infeasible due to the time and effort involved. Instead, GPT-4 few-shot prompting with 20 manually curated examples was used to extract lead terms from CodiEsp-X textual evidence. The author (SP) reviewed and corrected the extraction. The resulting CodiEsp-X-lead dataset is identical to CodiEsp-X but includes additional columns for the GPT-extracted lead terms.

Approach

For this study, we followed the process described in 2 main phases and 5 steps in Figure 3, mimicking the process human medical coders follow.

Figure 3. Schematic overview of the process followed for ICD coding. ICD: International Classification of Diseases; RAG: retrieval-augmented generation; RoBERTa: Robustly Optimized BERT Pretraining Approach

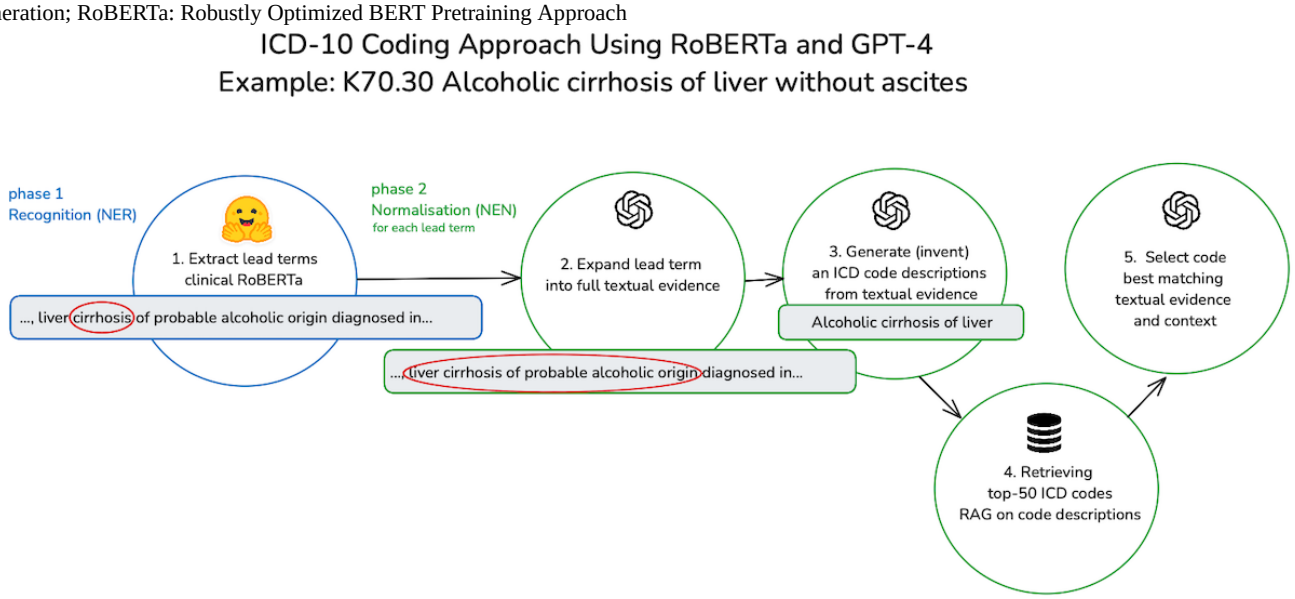

Step 1: Extract Lead Terms

A RoBERTa-based biomedical-clinical language model for Spanish was fine-tuned for NER of lead terms using the CodiEsp-X-lead dataset and Hugging Face's transformer library (version 4.26.0) [19,20], using the official dataset splits. The fine-tuned model enabled precise identification of lead terms within medical documents. To evaluate its performance, the same approach was applied to the CodiEsp-X dataset to extract full-textual evidence, which is more commonly used in AI methods for code assignment.

Step 2: Expand Lead Terms Into Full Textual Evidence

Instead of using NER to extract textual evidence, GPT-4 was instructed to complete a template for each lead term which included a field to capture the full textual evidence for the code.

XSL•FO
RenderX

Alongside the textual evidence, the template includes a field designed to capture contextual information, which is described as the text snippet containing all the necessary information for coding the lead term. In the prompt, the entire text was supplied, using HTML tags to highlight the lead terms within the original text.

Step 3: Generate (Invent) an ICD Code Description From Textual Evidence

In this part, we tested the theory that GPT-4 can improve text embeddings-based database search accuracy by aligning queries with official ICD code descriptions. This assumption arose from the observation that terms in official ICD code descriptions often markedly differ from those used in clinical texts. GPT-4 was tasked to generate (invent) ICD code descriptions from extracted textual evidence, aiming to reconcile these terminological discrepancies. This process, known as query alignment, refines a query to mirror the language and structure of database entries more closely, with the goal of enhancing search accuracy. To verify this assumption, the study compared different types of input queries; raw clinical text versus GPT-4 refined descriptions, to assess their effectiveness in achieving precise database lookups.

Step 4: Retrieving ICD Codes Using Descriptions and RAG

The complete lists of code descriptions for diagnoses and procedures contain around 2.5 million tokens each, far exceeding the context window size of current generative models. While including an entire knowledge base in the prompt is feasible for smaller datasets, RAG is necessary here to retrieve relevant context efficiently. To address this, we converted the official ICD code descriptions into vectors using OpenAI's "text-embedding-ada-002" model via the Langchain library [21], and stored them in the Chroma vector database. This enables efficient searching by embedding-based similarity rather than exact matches. By retrieving only the most relevant information, RAG ensures that the results fit within the model's context window, allowing it to process the required details effectively.

Step 5: Select the Code Best Matching Textual Evidence

From the 50 results retrieved from the vector store query, GPT-4 was prompted to choose the best-matching ICD code. For the decision, GPT was provided with the lead term, textual evidence, and additional context extracted in Step 2.

Evaluation

The performance of the ICD coding pipeline and its components was evaluated using precision, recall, and F_1-score metrics calculated with the CodiEsp scripts [18,22]. To assess performance loss at each step, predictions were replaced with actual values in subsequent steps only if the predictions at the current step were accurate. Specifically, in lead term extraction, a prediction was deemed correct if the predicted text span overlapped with the textual evidence. For textual evidence extraction, if there was an overlap, all values except the span were replaced. In the code retrieval step, the predicted code was replaced with the actual code if the predicted code was among the top-k results. In the code assignment, the predictions were not altered. Additionally, the accuracy of lead term extraction was also measured using F_1-scores calculated for the actual spans, independently of the CodiEsp script.

Ethical Considerations

This study used only publicly available data from the CodiEsp challenge, a synthetic corpus of 1000 clinical case studies, that does not contain real patient information, ensuring anonymity and eliminating the need for ethical approval.

Results

GPT-4

GPT-4 generated descriptions for 100 randomly selected ICD diagnosis codes and in a new chat compared these with official ICD-10 descriptions. It found that only 52% matched the official descriptions, as detailed in Multimedia Appendix 1. When GPT-4 was tasked with associating the correct ICD-10 code to each of the 100 official descriptions, only 47% of the ICD codes were correct.

Dataset CodiEsp-X-Lead

A key outcome of this study is the creation of the CodiEsp-X-lead dataset using GPT-4. This dataset consists of 7748 unique textual evidence strings, with 6532 of these containing multiple terms. As instructed, GPT-4 typically extracted a single lead term from these multiterm strings. However, in 549 cases (8%), multiple lead terms were extracted. The researchers manually resolved 325 of these cases to a single lead term, leaving the remaining cases unmapped due to ambiguity in determining the appropriate lead term.

Step 1: Extracting Lead Terms

After fine-tuning the Biomedical-clinical language model for Spanish for NER to extract lead terms [29], it achieved an F_1-score of 0.80 (precision of 0.78 and a recall) of 0.83 on the CodiEsp-X-lead test set (Table 1). When this model was solely fine-tuned and tested for diagnoses, it exhibited a slightly better performance, with an F_1-score of 0.82 (Table 1).

XSL•FO
RenderX

Table . Named entity recognition (NER) performance on extracting lead terms and textual evidence for diagnoses and or procedures on CodiEsp-X and CodiEsp-X-lead.

NER	F_1-score
Textual evidence diagnoses and procedures combined (ours-baseline) [29]	0.65
Textual evidence diagnoses (SOTA[a]) [5]	0.71
Textual evidence procedures (SOTA) [5]	0.61
Main-terms diagnoses (ours) [29]	0.82
Main-terms diagnoses and procedures combined (ours) [29]	0.80

[a]SOTA: state-of-the-art.

Step 2: Expand Lead Terms Into Full Textual Evidence

Using GPT-4, full textual evidence was extracted based on preannotated lead terms, which were identified in Step 1. GPT-4 completed a template that incorporated the full textual evidence for each provided lead term. This approach achieved an F_1-score of 0.509, as shown in Figure 4.

Figure 4. F1-Score performance loss for each individual step in the approach.

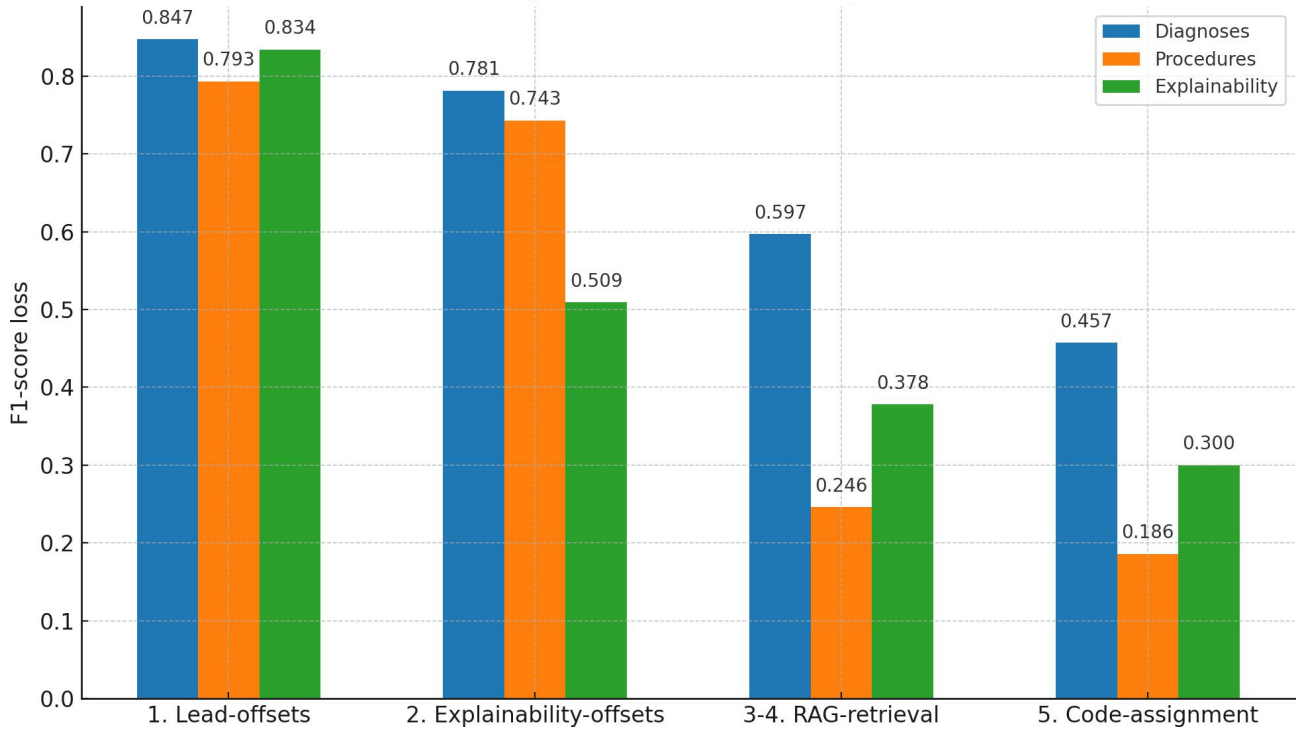

The base model from Step 1 was fine-tuned to extract textual evidence solely for benchmarking, achieving an F_1-score of 0.65 on the CodiEsp-X test set (Table 1). This model was not used in the actual approach, which aims to mimic medical coders by prioritizing the extraction of lead terms. Note that results from Table 1, assessing exact offsets, differ from those in Figure 5 where the official CodiEsp scripts allow for an error margin.

XSL•FO
RenderX

Figure 5. Comparison of failed retrievals @50 in a vector database containing all official ICD diagnosis and procedure codes. The chart compares results for 25 documents (507 codes), contrasting searching by textual evidence versus GPT-4 zero-shot generated code descriptions based on the textual evidence.

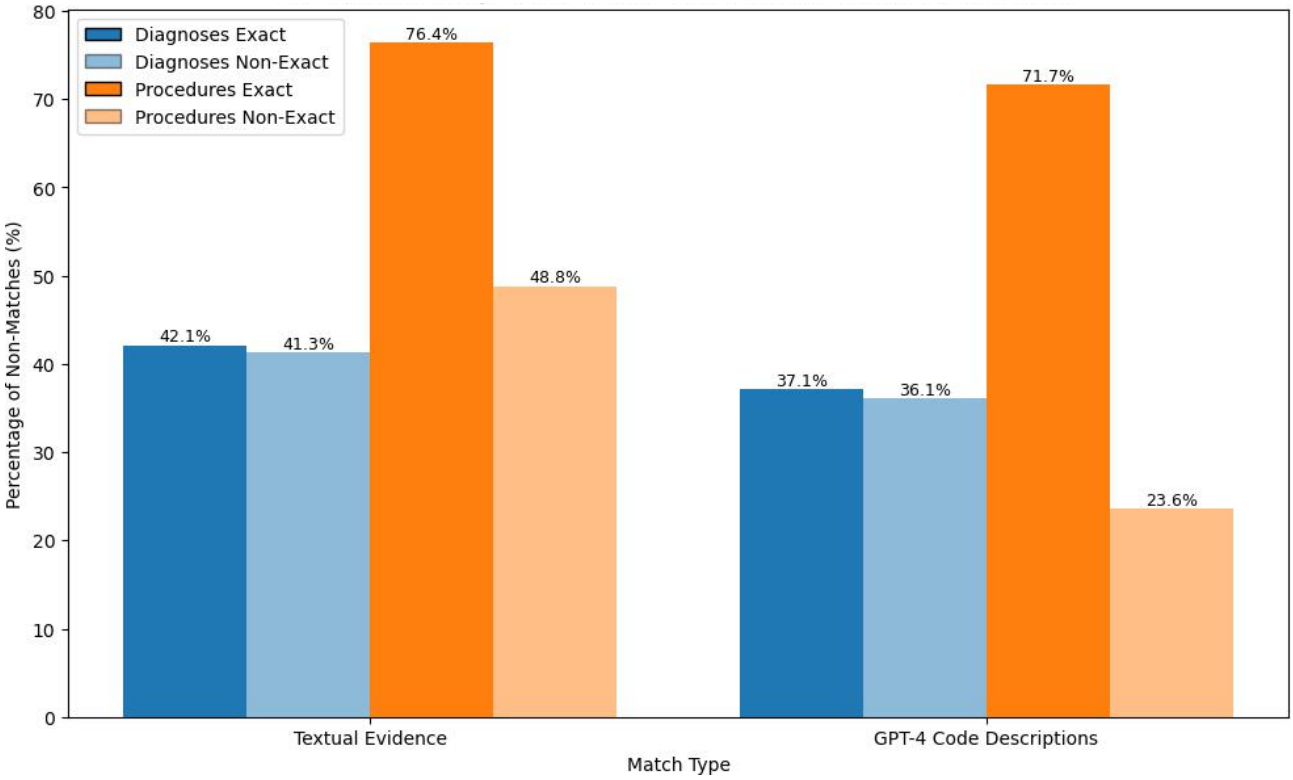

Step 3: Generate (Invent) an ICD Code Description From Textual Evidence

GPT-4 was used to generate ICD code descriptions from textual evidence, evaluated on 25 documents containing 507 codes in a RAG setting. When comparing GPT-4-generated descriptions to direct retrieval using textual evidence (Figure 5), retrieval failures decreased from 160 to 141 (42.1% to 37.1%) for the 380 diagnoses codes and from 97 to 91 (76.4% to 71.7%) for the 127 procedure codes.

For procedure codes, many were assigned as 4-axis codes, lacking the detail of the official 7-axis codes and without available descriptions in the resources. As a comparison, a nonexact matching criterion was applied, where a match was considered valid if the assigned code was a substring of the retrieved code. Under this nonexact approach, retrieval failures for procedures dropped significantly, from 62 to 30 out of 127 (48.8% to 23.6%). For diagnoses, which were mostly assigned in full detail, the decrease in failures was smaller, from 157 to 137 out of 380 (41.3% to 36.1%).

Step 4: Retrieving ICD Codes Using Descriptions and RAG

When RAG retrieval was evaluated within the full pipeline, a significant drop in explainability F_1-score was observed, falling from 0.509 to 0.378, as shown in steps 3 and 4 in Figure 4. Procedures had the largest contribution to this drop, with their F_1-score decreasing from 0.743 to 0.246. Diagnoses also experienced a decline, with their F_1-score dropping from 0.781 to 0.597.

Step 5: Select the Code Best Matching Textual Evidence

In the final step, GPT-4 had to select the correct ICD code from the top 50 retrieved results based on the previously predicted textual evidence. This selection process caused a further decline in performance, with the explainability F_1-score dropping from 0.378 to 0.3 (Table 2 and Figure 4).

XSL•FO
RenderX

Table . Comparison of performance of different models in the CodiEsp challenge.

Team name	Diagnoses (F_1-score)	Procedures (F_1-score)	Explainability (F_1-score)
FLE	0.679	0.514	0.611
IAM	0.687	0.522	0.611
BETO-Galén+mBERTGalén+XLM-R-Galén	—[a]	—	0.633
Ours	0.457	0.204	0.305

[a]Not available.

Discussion

Summary of Main Findings

In this study, we proposed methods to assist ICD-10 coding using LLMs. We successfully created a dataset of lead terms, "CodiEsp-X-lead," using few-shot prompting on GPT-4. Using this dataset, we effectively fine-tuned a transformer model (F_1-score of 0.80) that can assist medical coders in the initial step of medical coding: identifying ICD lead terms. To enhance RAG retrieval, we employed GPT-4 to generate code descriptions from textual evidence, aiming to improve ICD code identification within a code description database. Although the initial step, focusing on extracting lead terms, closely mimicked medical coders and performed well, subsequent RAG steps based on textual descriptions were less effective, yielding a poor F_1-score of 0.305 on the CodiEsp-X task. We will discuss the shortcomings of the RAG implementation and analyze each step's performance, highlighting areas of significant improvement.

Motivation of Choices

RoBERTa was selected over GPT-4 for NER of lead terms because encoder-based models are better suited for token-level tasks like NER, providing detailed context understanding. RoBERTa is lighter and more efficient to fine-tune, potentially outperforming generative, decoder-only models like GPT-4 [17,30]. Additionally, as of 2023, fine-tuning GPT-4 was not possible. Previous research indicates that GPT models perform poorly in traditional, example-rich NER tasks [31].

In contrast, GPT-4 excels in few-shot NER tasks and was chosen for other aspects of the study due to its SOTA performance in NLP. Medical coding involves following specific guidelines, and a well-configured generative model like GPT-4, trained for instruction-following, is likely to perform well in such roles. Therefore, GPT-4 was used to create the dataset by extracting lead terms through few-shot NER, as manually compiling the dataset was not feasible. Its natural language generation capabilities and instruction-following skills were also leveraged to generate code descriptions, align queries, and identify the correct ICD-10 codes from a given list.

Using code descriptions exclusively for normalizing lead terms to ICD codes offers independence from training data-based fine-tuning. Given the vastness of ICD-10, with over 100,000 codes, many are rarely encountered in practice. The analysis of the CodiEsp dataset revealed that 17.3% of specific test set codes were not included in the training data. If the proposed method, which primarily relies on official code descriptions, were to be successful, it likely would be very generalizable across different ICD datasets.

GPT-4 Parameteric Knowledge

The preliminary results of GPT-4's performance were disappointing. However, it is crucial to approach negative outcomes based on prompts with caution, as the prompts themselves may be suboptimal. Nevertheless, we can definitively state that GPT-4's out-of-the-box performance on 2 basic ICD-10 coding tasks—coding official descriptions and generating code descriptions for diagnoses—was mediocre. The approach of using GPT-4 generated descriptions for procedure codes was found to be completely ineffective in the study. These findings and those from the first available publications on GPT-3.5 and GPT-4's abilities with ICD billing codes underscore their subpar performance for downstream applications [23].

Creating CodiEsp-X-Lead With GPT-4

In this study, we introduced "CodiEsp-X-lead", a curated selection of lead terms from the CodiEsp-X dataset. Using GPT-4 as an annotator and employing few-shot prompting, we extracted these terms from the dataset's textual evidence. This method showcases its utility for other use cases where manual annotation is unfeasible, but reviewing the generated annotations is feasible. While the authors manually reviewed and made minor adjustments to this extraction, there is potential for further refinement. For improved lead term identification, lead terms derived from the codes assigned to that textual evidence could be suggested by using an inverse lookup on the official alphabetical index. This approach can benefit both the GPT-4 prompts given and our manual review process. Last, to ensure the highest accuracy, domain experts, such as medical coders, could annotate or validate a portion of the lead terms and can help improve the quality of the examples used in the (few-shot) prompt. Another suggestion to enhance the dataset's quality is to allow multiple lead terms for each assigned code. Limiting to a single lead term when several are applicable results in inconsistencies and reduces the quality of the dataset and the model's performance.

Step 1: Extracting Lead Terms

The results demonstrate that using NER for lead term identification holds distinct advantages over extracting the full textual evidence all at once.

XSL•FO
RenderX

First, since the task is more straightforward, significantly better performance can be obtained when extracting lead terms compared with extracting the full textual evidence. For instance, the SOTA approach for the CodiEsp-X task yielded an F_1-score of 0.71 for NER of full-textual evidence for diagnoses compared with our score of 0.82 for NER of lead terms (Table 1) [2]. High performance in extracting the correct lead term is crucial, as it sets the performance ceiling for the entire coding approach.

Second, starting with lead term extraction mirrors the initial step of human coders, facilitating seamless integration into their workflow and providing standalone value. A coding assistant that accurately identifies lead terms with high recall and precision would likely be favored over a system that offers full ICD assignments but has a higher error rate. The latter could compel human coders to frequently reread the text and make extensive corrections, frequently starting from scratch. In contrast, a fully autonomous approach that meets performance requirements without needing human intervention would likely favor extracting full-textual evidence at once, due to its potential for higher efficiency. However, in such a scenario, the relevance of explainability becomes questionable, particularly when the system operates satisfactorily with no human oversight.

The proposed NER model for extracting lead terms performs well, but further improvements can be achieved by enhancing the quality of the CodiEsp-X-lead dataset. It may be beneficial to fine-tune the model to prioritize recall over precision, as medical coders might favor removing incorrect lead term suggestions instead of revisiting a text to verify that all diagnoses are captured. Additionally, optimizing parameters might enhance the performance of the model.

Step 2: Expand Lead Terms Into Full Textual Evidence

From Step 2, our method deviated from ICD-10 coding guidelines by adopting a RAG approach, which requires a database lookup rather than following the traditional coding steps.

Using a zero-shot GPT-4 prompt to extract full textual evidence resulted in mediocre outcomes, with the F_1-score for explainability dropping from 0.824 to 0.509 (Figure 4). This highlights a significant limitation of this approach. In contrast, skipping lead term extraction and directly employing a fine-tuned BERT model yielded better performance, achieving F_1-scores of 0.71 and 0.61 for diagnoses and procedures respectively, which are SOTA, and a combined score of 0.65 for our model [5]. For future work, we propose to align more closely with medical coders' practices by continuing to use extracted lead terms and leveraging the alphabetic index and official guidelines for code lookups, allowing textual evidence to emerge naturally. This differs from the current RAG approach, which relies on a single lookup using predefined semantic queries.

Step 3: Generate (Invent) an ICD Code Description From Textual Evidence

The hypothesis we examined suggested that GPT-generated code descriptions, based on textual evidence and context, would semantically more closely align with a database of ICD code descriptions than the original textual evidence. This expected alignment is intended to enhance RAG retrieval, suggesting that rephrasing textual evidence into descriptions using AI could improve the accuracy of searching for ICD codes by their descriptions. From the experiments on a subset of the data (Figure 5), we confirm the hypothesis, as all cases show a reduction in failed retrievals. Although the improvements are subtle in most scenarios, with only a few percentage points fewer failed retrievals, there is a marked improvement for nonexact matching procedure codes. Specifically, failed retrievals dropped from 48.8% to 23.6%. This notable performance in nonexact matching is likely due to the majority of procedure codes lacking descriptions (59%), which is why nonexact matching was evaluated and resulted in a significant reduction in the number of classes.

Step 4: Retrieving ICD Codes Using Descriptions and RAG

In step 3, we tested nonexact matching for procedure codes, but the official Codiesp evaluation uses exact matching. Our approach, relying on NER extraction and descriptions for code lookup, fails when many assigned codes lack descriptions and are invalid.

In the RAG retrieval step, F_1-score performance for procedure codes drops sharply from 0.743 to 0.246. Diagnosis code performance is also mediocre at 0.597, but this would still be acceptable if GPT-4 could select the correct code in the next step without causing another drop.

Step 5: Select the Code Best Matching Textual Evidence

In step 5, GPT-4 selects the final answer from the retrieved codes, but another drop in performance is observed, resulting in a final F_1-score of 0.3 on CodiEsp-x, compared with the SOTA score of 0.633 (Table 2) [5]. It is likely that the proposed textual evidence (span) is incorrect in some cases, making it valuable to evaluate this step in isolation to assess its impact on overall performance.

Method Recap and Evaluation

While the initial NER step for extracting lead terms is promising and can directly assist medical coders, subsequent steps perform inadequately. Although using RAG with code descriptions seems intuitive for ICD-10 coding, the implementation underperforms. While lead term extraction works well, the following steps should better mimic the workflow of medical coders. If this is not done, step 2 could be replaced by extracting textual evidence via NER, which achieves a higher F_1-score (0.65 in Table 1) compared with the current method (0.509 in Figure 5).

Instead of relying on descriptions to find the correct code, it would be more effective to use the alphabetic index of lead terms. Textual evidence could then be derived from this lookup using iterative steps, rather than the current approach, which requires knowing the exact spans of evidence in advance before performing the database lookup, as we need to know exactly what we are searching for.

XSL•FO
RenderX

Additionally, terms like NEC (not elsewhere classified) and NOS (not otherwise specified), which are present in some code descriptions, could be useful for an agent-based approach, allowing an agent to interpret and follow these instructions in coding scenarios. This would enhance performance, as an agent can adaptively handle these cases, which are often challenging for a simple database lookup.

Incorporating more coding guidelines into the process would further align with an agent-based model, closely replicating the real-world workflow of medical coders.

Limitations

Our hypothesis was that GPT-4 can achieve effective medical ICD-10 code normalization (assignment) by solely using the official WHO-supplied ICD-10 code descriptions. The findings suggest that the combination of GPT-4 with solely code descriptions might not be sufficient.

Relying solely on the WHO descriptions for medical coding can be problematic. A prime example is the challenge in coding “lower limb weakness” when it must fit into a broad category description like “Other malaise”, highlighting issues with semantic differences. Although the embedding methods used address some similarity issues, using an alphabetic index or leveraging training data might be more effective and accurate, but care must be taken to maintain generalizability.

A major challenge in using GPT-4 for research lies in its cost, processing speed, and privacy concerns. For each individual medical text, processing with GPT-4 using the proposed prompts typically takes several minutes at a minimum. Furthermore, when evaluating 50 code candidates, each accompanied by descriptions for the lead term extracted, the token use becomes significant, resulting in substantial costs. Additionally, GPT-4’s closed-source nature presents further complications in research applications.

Another limitation is the exclusive use of the CodiEsp dataset, which restricts the generalizability of our findings. While the MIMIC dataset is a commonly used public resource for ICD coding, it lacks annotations for explainability. No other publicly available ICD dataset is currently coded for explainability.

Future Research

For NER of lead terms and their normalization to ICD codes, we suggest continuing with the NER approach for extracting lead terms, given its promising performance and its seamless integration into the workflow of medical coders. For future research, we propose implementing an agent-based LLM approach that mimics all steps of the medical coders’ workflow in detail.

Another point for future research is to explore the use of locally runnable LLMs, ensuring that data can remain in-house for privacy and security purposes.

Conclusions

A dataset of ICD-10 lead terms, “CodiEsp-X-lead”, was created using few-shot prompting on GPT-4. This method of model-assisted labeling showcases its utility for other use cases where manual annotation is unfeasible, but reviewing the generated annotations is feasible. As GPT models are not the SOTA in NER [31], a RoBERTa-based model was fine-tuned, achieving an F_1-score of 0.80. This model can aid medical coders in the primary step of medical coding by identifying ICD lead terms. GPT-4 was employed to generate code descriptions based on textual evidence, improving RAG retrieval for code lookups in a code description database. However, exclusive reliance on code descriptions, coupled with GPT-4 prompting for ICD-10 coding, led to only mediocre outcomes. Continuing with the NER approach for lead term extraction is advisable. Relying solely on code descriptions, along with GPT-4 prompts for ICD-10 coding, led to mediocre outcomes, partly because some codes in the dataset lacked descriptions. Additionally, this approach is likely less effective than mimicking the complete workflow of medical coders, which could yield more accurate results. A key limitation of using RAG for medical coding is its requirement for full-textual evidence upfront, making it less flexible than methods that extract evidence during the lookup process using official coding resources and guidelines.

Acknowledgments

The authors acknowledge the use of ChatGPT (GPT-4 and GPT-4o, OpenAI, 2023/2024) to assist in improving the language and readability of this manuscript. ChatGPT was specifically employed to rephrase sentences for clarity and enhance the overall articulation of the text. The authors carefully reviewed and refined the rephrased content to ensure that the final text accurately reflected their intended meaning and original ideas.

Data Availability

The datasets generated or analyzed during this study are available in the "Enhancing ICD coding with LLMs" repository [32].

The official CodiEsp dataset is openly available at [33].

Conflicts of Interest

None declared.

Multimedia Appendix 1
GPT-4 parametric ICD-10 knowledge.

XSL•FO
RenderX

[DOCX File, 17 KB - formative_v9i1e60095_app1.docx]

References

1. Alharbi MA, Isouard G, Tolchard B. Historical development of the statistical classification of causes of death and diseases. Cog Med 2021 Jan 1;8(1):1893422. [doi: 10.1080/2331205X.2021.1893422]
2. Kim BH, Ganapathi V. Read, attend, and code: pushing the limits of medical codes prediction from clinical notes by machines. 2021 Presented at: Proceedings of the 6th Machine Learning for Healthcare Conference; Aug 6-7, 2021; Proceedings of Machine Learning Research p. 196-208.
3. Joos C, Lawrence K, Jones AE, Johnson SA, Witt DM. Accuracy of ICD-10 codes for identifying hospitalizations for acute anticoagulation therapy-related bleeding events. Thromb Res 2019 Sep;181:71-76. [doi: 10.1016/j.thromres.2019.07.021] [Medline: 31357146]
4. Burns EM, Rigby E, Mamidanna R, Bottle A, Aylin P, Ziprin P, Faiz OD A, Aylin P, Ziprin P, et al. Systematic review of discharge coding accuracy. J Public Health (Oxf) 2012 Mar;34(1):138-148. [doi: 10.1093/pubmed/fdr054] [Medline: 21795302]
5. López-García G, Jerez JM, Ribelles N, Alba E, Veredas FJ. Explainable clinical coding with in-domain adapted transformers. J Biomed Inform 2023 Mar;139:104323. [doi: 10.1016/j.jbi.2023.104323] [Medline: 36813154]
6. Barros J, Rojas M, Dunstan J, Abeliuk A. Divide and conquer: an extreme multi-label classification approach for coding diseases and procedures in spanish. : Association for Computational Linguistics; 2022 Presented at: Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI); Abu Dhabi, United Arab Emirates (Hybrid). [doi: 10.18653/v1/2022.louhi-1.16]
7. Huang CW, Tsai SC, Chen YN. PLM-ICD: automatic ICD coding with pretrained language models. arXiv. Preprint posted online on 2022 URL: http://arxiv.org/abs/2207.05289 [accessed 2023-02-22]
8. Hulsen T. Explainable artificial intelligence (XAI): concepts and challenges in healthcare. AI 2023;4(3):652-666. [doi: 10.3390/ai4030034]
9. Centers for Medicare & Medicaid Services. ICD-10-PCS official guidelines for coding and reporting 2023. URL: https://www.cms.gov/files/document/2023-official-icd-10-pcs-coding-guidelines.pdf [accessed 2024-11-20]
10. Xie X, Xiong Y, Yu PS, Zhu Y. EHR coding with multi-scale feature attention and structured knowledge graph propagation. : Association for Computing Machinery; 2019 Nov 3 Presented at: Proceedings of the 28th ACM International Conference on Information and Knowledge Management; Nov 3-7, 2019; Beijing China p. 649-658. [doi: 10.1145/3357384.3357897]
11. Amjad H, Ashraf M, Sherazi S, et al. Attention-based explainability approaches in healthcare natural language processing. : SCITEPRESS; 2023 Presented at: Proceedings of the 16th International Joint Conference on Biomedical Engineering Systems and Technologies; Lisbon, Portugal p. 689-696. [doi: 10.5220/0011927300003414]
12. Edin J, Maistro M, Maaløe L, Borgholt L, Havtorn JD, Ruotsalo T. An unsupervised approach to achieve supervised-level explainability in healthcare records. arXiv. Preprint posted online on 2024 URL: http://arxiv.org/abs/2406.08958 [accessed 2024-10-09]
13. Safian SC, Johnson MA. Let's Code It! 2019-2020 Code Edition, 2nd edition: McGraw-Hill Education; 2020.
14. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. arXiv. URL: http://arxiv.org/abs/1706.03762
15. Liu Y, Ott M, Goyal N, Du J J, Joshi M, Chen D, et al. RoBERTa: a robustly optimized BERT pretraining approach. arXiv. Preprint posted online on 2019 URL: http://arxiv.org/abs/1907.11692
16. Devlin J, Chang MW, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding. arXiv. Preprint posted online on 2018 URL: http://arxiv.org/abs/1810.04805
17. Lu Q, Li R, Wen A, Wang J, Wang L, Liu H. Large language models struggle in token-level clinical named entity recognition. arXiv. Preprint posted online on 2024 URL: http://arxiv.org/abs/2407.00731
18. Clark K, Khandelwal U, Levy O, At M. What does BERT look at? an analysis of BERT's attention. arXiv. Preprint posted online on 2019 URL: http://arxiv.org/abs/1906.04341
19. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. OpenAI 2018 [FREE Full text]
20. OpenAI GPT-4 technical report. arXiv. Preprint posted online on 2023 URL: http://arxiv.org/abs/2303.08774
21. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. arXiv. Preprint posted online on 2020 URL: http://arxiv.org/abs/2005.14165 [accessed 2023-10-12]
22. Li J, Dada A, Puladi B, Kleesiek J, Egger J. ChatGPT in healthcare: a taxonomy and systematic review. Comput Methods Programs Biomed 2024 Mar;245:108013. [doi: 10.1016/j.cmpb.2024.108013] [Medline: 38262126]
23. Soroush A, Glicksberg BS, Zimlichman E, et al. Assessing GPT-3.5 and GPT-4 in generating international classification of diseases billing codes. medRxiv. Preprint posted online on Jul 7, 2023. [doi: 10.1101/2023.07.07.23292391]
24. Yang Z, Kwon S, Yao Z, Yu H. Multi-label few-shot ICD coding as autoregressive generation with prompt. Proc AAAI Conf Artif Intell 2023 Jun 26;37(4):5366-5374. [doi: 10.1609/aaai.v37i4.25668] [Medline: 37635946]
25. Safavi-Naini SAA, Ali S, Shahab O, Shahhoseini Z, Savage T, Rafiee ST, et al. Vision-language and large language model performance in gastroenterology. arXiv. Preprint posted online on 2024 URL: http://arxiv.org/abs/2409.00084

XSL•FO
RenderX

26. Gema AP, Leang JOJ, Hong G, Devoto A, Mancino ACM, Saxena R, et al. Are we done with MMLU. arXiv. Preprint posted online on 2024 URL: http://arxiv.org/abs/2406.04127
27. Chiang WL, Zheng L, Sheng Y, Angelopoulos AN, Li T, Li D, et al. Chatbot arena: an open platform for evaluating llms by human preference. arXiv. Preprint posted online on 2024 URL: http://arxiv.org/abs/2403.04132
28. Miranda-Escalada A, Gonzalez-Agirre A, Armengol-Estapé J, Krallinger M. Overview of automatic clinical coding: annotations, guidelines, and solutions for non-English clinical cases at codiesp track of CLEF eHealth 2020. 2020 Presented at: Working Notes of Conference and Labs of the Evaluation (CLEF) Forum CEUR Workshop Proceedings; Sep 22-25, 2020; Thessaloniki, Greece.
29. Carrino CP, Llop J, Pàmies M, et al. Pretrained biomedical language models for clinical NLP in spanish. : Association for Computational Linguistics; 2022 Presented at: Proceedings of the 21st Workshop on Biomedical Language Processing; Dublin, Ireland p. 193-199. [doi: 10.18653/v1/2022.bionlp-1.19]
30. Hu Y, Chen Q, Du J, Peng X, Keloth VK, Zuo X, et al. Improving large language models for clinical named entity recognition via prompt engineering. J Am Med Inform Assoc 2024 Sep 1;31(9):1812-1820. [doi: 10.1093/jamia/ocad259] [Medline: 38281112]
31. Wang S, Sun X, Li X, Ouyang R, Wu F, Zhang T, et al. GPT-NER: named entity recognition via large language models. arXiv. Preprint posted online on 2023 URL: http://arxiv.org/abs/2304.10428 [accessed 2023-07-20]
32. Enhancing ICD coding with LLMs. GitHub. 2024. URL: https://github.com/putssander/icdllmeval [accessed 2025-01-20]
33. Miranda EA, Gonzalez AA, Krallinger M. CodiEsp corpus: gold standard Spanish clinical cases coded in ICD10 (CIE10) - eHealth CLEF2020. Z 2020. [doi: 10.5281/zenodo.3625746]

Abbreviations

BERT: Bidirectional Encoder Representations from Transformer
ICD: International Classification of Diseases
LLM: large language model
NEC: not elsewhere classified
NER: named entity recognition
NLP: natural language processing
NOS: not otherwise specified
RAG: retrieval-augmented generation
RoBERTa: Robustly Optimized BERT Pretraining Approach
SOTA: state-of-the-art
WHO: World Health Organization

Edited by A Mavragani; submitted 01.05.24; peer-reviewed by SS Sohail, T Karen; revised version received 20.11.24; accepted 24.11.24; published 11.02.25.

Please cite as:
Puts S, Zegers CML, Dekker A, Bermejo I
Developing an ICD-10 Coding Assistant: Pilot Study Using RoBERTa and GPT-4 for Term Extraction and Description-Based Code Selection
JMIR Form Res 2025;9:e60095
URL: https://formative.jmir.org/2025/1/e60095
doi:10.2196/60095

© Sander Puts, Catharina M L Zegers, Andre Dekker, Iñigo Bermejo. Originally published in JMIR Formative Research (https://formative.jmir.org), 11.2.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Web-Based, Algorithm-Guided Insulin Titration in Insulin-Treated Type 2 Diabetes: Pre-Post Intervention Study

Nishanth Thiagarajan, MBBS; Hong Chang Tan, MBBS; Suresh Rama Chandran, MBBS; Phong Ching Lee, MBChB; Yun Ann Chin, MBBS; Wanling Zeng, MD; Emily Tse Lin Ho, MBBS; David Carmody, MBChB; Su-Yen Goh, MBBS; Yong Mong Bee, MBBS

Department of Endocrinology, Singapore General Hospital, 20 College Road, Academia, Singapore, Singapore

Corresponding Author:
Nishanth Thiagarajan, MBBS
Department of Endocrinology, Singapore General Hospital, 20 College Road, Academia, Singapore, Singapore

Abstract

Background: Self-monitoring of blood glucose (SMBG) using web-based diabetes management platforms has demonstrated promise in managing type 2 diabetes (T2D). However, the effectiveness of such systems incorporating algorithm-guided insulin titration has not been extensively studied in Asian populations.

Objective: This study evaluates the efficacy and safety of the ALRT telehealth solution—a US Food and Drug Administration–cleared, web-based platform that integrates SMBG with algorithm-driven insulin dose adjustments—in improving glycemia in insulin-treated T2D.

Methods: This 24-week, pre-post intervention study enrolled 25 adults with T2D (mean age 58.9, SD 7.0 y; n=14, 56% male) on twice-daily premixed insulin. Inclusion criteria included a baseline hemoglobin A_{1c} (HbA_{1c}) level between 7.5% to 9.9% (58 - 86 mmol/mol), a BMI ≤40 kg/m², and experience with SMBG. Participants uploaded twice-daily SMBG data weekly via a mobile app, which generated insulin titration recommendations based on a predefined algorithm. Physicians reviewed and approved the recommendations, which were then communicated back to participants via the app. The primary outcome was the change in HbA_{1c} level from baseline to 24 weeks. Secondary outcomes included changes in fasting plasma glucose, insulin dose, hypoglycemia incidence, and SMBG adherence.

Results: Participants achieved significant reductions in HbA_{1c} level from 8.6% (70 mmol/mol) at baseline to 7.4% (57 mmol/mol) at 24 weeks (P<.001), with reductions of 0.8% and 0.4% in the first and second 12 weeks, respectively. Fasting plasma glucose decreased from 8.7 (SD 2.0) mmol/L to 7.1 (SD 1.4) mmol/L (P<.001). Mean total daily insulin dose increased modestly from 0.73 (SD 0.31) units/kg/day to 0.79 (SD 0.34) units/kg/day (P=.007). Participants demonstrated high adherence, completing 97.3% (327/336) of prescribed SMBG measurements. During the study, 48% (12/25) of participants experienced at least 1 hypoglycemia episode, predominantly mild hypoglycemia (85/96, 88.5%; glucose 3.0 - 3.9 mmol/L). Hypoglycemia episodes increased from 24 during weeks 0 - 12 to 72 during weeks 13 - 24. There were no episodes of severe hypoglycemia requiring external assistance. BMI increased slightly from 29.0 (SD 3.6) kg/m² to 29.5 (SD 3.6) kg/m² (P=.03), reflecting a modest weight gain associated with improved glycemia.

Conclusions: In conclusion, patients with insulin-treated T2D initiated on a web-based glucose monitoring system with algorithm-guided dosing recommendations showed significant improvement in glycemic control compared to baseline. High adherence rates underscore the feasibility of integrating algorithm-guided insulin titration into routine care. While hypoglycemia incidence rose slightly, episodes were predominantly mild, and no severe events occurred. This intervention shows promise for broader adoption in T2D management, particularly in resource-constrained settings.

(JMIR Form Res 2025;9:e68914) doi:10.2196/68914

KEYWORDS

diabetes; insulin; monitoring; technology; mobile; app; intervention

Introduction

The global prevalence of diabetes mellitus has risen dramatically, with projections estimating that over 1.3 billion people will be affected by 2050 [1]. This chronic disease profoundly impacts both quality of life [2] and health care expenditure [3], and has driven significant efforts toward early detection and treatment.

Effective glycemic control remains the corner stone of diabetes management, correlating with improved long-term outcomes [4-6]. Despite rapid advances in pharmacotherapy, up to

XSL•FO
RenderX

two-thirds of patients fail to achieve their glycemic targets [7]. Self-monitoring of blood glucose (SMBG) is a pivotal component of diabetes management [8], enhancing patient engagement and glycemic control [9]. The benefits of SMBG are most pronounced [10-12] when a structured approach is taken. Ideally, blood glucose readings should be collected at specified intervals and returned to health care providers for interpretation. In this regard, a safe and consistent dose adjustment regimen will be needed if any treatment modification is required. Finally, patients need to be promptly informed of the necessary changes. Thus, structured SMBG's effectiveness is contingent on regular, frequent physician reviews and adjustments of therapy [13,14].

However, having frequent in-person clinic visits to review SMBG and adjust insulin doses is impractical, given the massive patient load and long waiting time of many diabetes clinics. Furthermore, a vast amount of SMBG data needs to be reviewed and interpreted by the physician during the limited allocated clinic time. This process is laborious and time-consuming, yet decision-making for insulin dose adjustments needs to be rapid. Providers with less clinical experience will find this especially challenging. Consequently, the SMBG data are often not fully capitalized to improve a patient's diabetes management.

Several web-based programs and mobile apps have been developed to overcome some of the abovementioned challenges. Instead of manual recording of SMBG and then providing feedback to their provider in person, these programs allow patients to record and transmit SMBG data directly via the app, reducing the frequency of face-to-face visits [15-17]. In an earlier study, we demonstrated the ability of a web-based blood glucose monitoring system to optimize insulin dosages for patients receiving a basal-plus or basal-bolus insulin regimen. Importantly, glycemic control improved without increasing hypoglycemia rates [18]. However, although this approach reduced the need for face-to-face visits for insulin titration, physicians still need to review a large amount of SMBG data before recommending any insulin dose adjustment. In addition, insulin doses are adjusted based on the physician's discretion and experience. Such an approach requires extensive workforce resources and is unlikely to solve the real-world problem of delays in treatment optimization. One potential solution is to use an automated insulin dose adjustment system that could interpret SMBG data and provide dose adjustment recommendations based on a preprogrammed algorithm. Studies have demonstrated the superiority of such systems in improving glycemic control compared to physician-dependent adjustments alone [19]. However, such programs have not been studied in an Asian cohort.

We thus designed the Glucose Monitoring and Intervention in Insulin-treated Type 2 Diabetes Patients (GEMINI-T2D) study to examine the efficacy and safety of a web-based platform that incorporates monitoring and insulin titration recommendations based on a built-in algorithm in patients treated with insulin.

Methods

Study Design

The GEMINI-T2D study was a prospective, single-site, single-group, pre-post interventional study.

Study Participants

All participants were recruited from the Singapore General Hospital's Diabetes & Metabolism Centre from September 2020 to May 2022. Those deemed potentially eligible were contacted by phone, by email, or during scheduled visits to the Diabetes & Metabolism Centre. Those who expressed interest in participating were given the informed consent form and invited to participate. Final eligibility of those who consented to participate was based on satisfactory completion of the screening process, demonstrating full eligibility based on inclusion and exclusion criteria.

Inclusion Criteria

We recruited adults (aged ≥21 y) with type 2 diabetes (T2D) of more than 6-month duration on twice-daily premixed insulin regimen for ≥3 months. The total daily dose (TDD) of insulin must be <1 unit/kg, hemoglobin A_{1c} (HbA_{1c}) level must be between 7.5% (58 mmol/mol) to 9.9% (86 mmol/mol), and BMI must be ≤40 kg/m². We chose to study patients on premixed insulin as it represents a common and preferred insulin formulation in Asia [20,21]. Participants were required to have experience performing SMBG and own a compatible smartphone capable of weekly data uploads via the study app.

Exclusion Criteria

We excluded individuals with hypoglycemia unawareness (Gold score ≥4) to limit the risk of severe hypoglycemia during treatment. Pregnant or breastfeeding women, as well as those with severe renal impairment (estimated glomerular filtration rate <30 mL/min/1.73 m²), hemoglobinopathies, systemic corticosteroid use, and medical disease with a life expectancy of less than 1 year, were excluded.

Intervention

Eligible participants were given a glucometer and instructed to perform SMBG twice daily. They were trained to use the ALRT system—a US Food and Drug Administration–cleared diabetes management system—and asked to upload their SMBG data to the ALRT app weekly for the next 24 weeks. Participants were also educated on hypoglycemia recognition and management, and instructed to perform SMBG when experiencing symptoms suggestive of hypoglycemia.

Insulin Dose Adjustment

The ALRT system analyzed uploaded SMBG data every 7 days. When capillary blood glucose readings consistently fell outside the predefined range of 4 - 8 mmol/L, a predefined algorithm adjusted the prebreakfast and/or predinner insulin doses, implementing an increment or decrement of up to 15% to 20% as necessary (details of the algorithm are provided in Multimedia Appendix 1). The treating physician was notified via a web-based user interface with the recommended insulin dose adjustments and was required to accept or decline the

XSL•FO
RenderX

recommendation. Subsequently, the adjusted insulin doses were communicated to the participants via the ALRT app every 7 days, or more frequently in the event of hypoglycemia. Participants were required to acknowledge the recommended dose adjustments via the app.

Baseline and Follow-Up Assessments

Medical history, vital signs, body measurements, and laboratory measurements were collected at baseline. HbA_{1c} level and fasting plasma glucose (FPG) were measured at baseline, week 12, and week 24.

Outcome Measures

The primary outcome measure was the change in HbA_{1c} level at the end of 24 weeks relative to baseline. Secondary outcome measures included the incidence of hypoglycemia, defined as the number of episodes of hypoglycemia per participant, and the adherence to the prescribed SMBG regimen. Episodes of hypoglycemia were classified as follows:

- Level 1: Glucose value of 3.0 - 3.9 mmol/L
- Level 2: Glucose value of <3.0 mmol/L
- Level 3: Severe hypoglycemia causing altered mentation requiring external assistance for recovery

Adherence to the SMBG regimen was defined as the proportion of the prescribed twice-daily SMBG that the participant successfully completed and uploaded to the ALRT platform. The incidence of hypoglycemia was defined as the number of episodes of hypoglycemia that occurred during each 12-week period of the study.

Statistical Analysis

Continuous data were presented as mean (SD) and categorical data were presented as numbers (percentages). We compared the variables at baseline, week 12, and week 24, using repeated-measures ANOVA for normally distributed variables with post hoc Tukey method, and Friedman ANOVA for nonnormally distributed variables with post hoc Wilcoxon signed-rank tests with Bonferroni correction. Statistical analyses were performed using R (version 4.1.1; R Foundation for Statistical Computing).

Ethical Considerations

The SingHealth Institutional Review Board reviewed and approved the study (reference 2019/2874). Informed consent was obtained from all participants, and this was inclusive of secondary analysis. The study data were deidentified. Participants were compensated with a nominal transport allowance of SGD $80 (US $59.14) for each of the clinic visits at weeks 0, 12, and 24.

Results

Participant Characteristics

A total of 25 participants were recruited and followed up over 24 weeks. The baseline demographic and clinical characteristics are summarized in Table 1. The study cohort had a mean age of 58.9 (SD 7.0) years, was 56% (n=14) male, and had a mean BMI of 29.0 (SD 3.6) kg/m^2.

XSL•FO
RenderX

Table . Baseline demographic and clinical characteristics of 25 adults with type 2 diabetes enrolled in the GEMINI-T2D[a] study, a 24-week pre-post intervention study conducted at the Singapore General Hospital from September 2020 to May 2022. All participants were on twice-daily premixed insulin and met inclusion criteria for HbA_{1c}[b] level (7.5% - 9.9%), BMI (≤40 kg/m²), and experience with self-monitoring of blood glucose.

Characteristics	Value (N=25)
Age (years), mean (SD)	58.9 (7.0)
Sex, n (%)	
Male	14 (56)
Female	11 (44)
Ethnicity, n (%)	
Chinese	13 (52)
Malay	4 (16)
Indian	6 (24)
Others	2 (8)
BMI (kg/m^2), mean (SD)	29.0 (3.6)
HbA_{1c} level (%), mean (SD)	8.6 (0.7)
HbA_{1c} level (mmol/mol), mean (SD)	70.4 (7.7)
Fasting plasma glucose (mmol/L), mean (SD)	8.7 (2.0)
Duration of diabetes (years), mean (SD)	18.9 (6.7)
Duration of insulin use (years), mean (SD)	6.3 (3.5)
Total daily dose of insulin (units/day), mean (SD)	57.3 (24.4)
Total daily dose of insulin (units/kg/day), mean (SD)	0.73 (0.31)
Hypertension, n (%)	22 (88)
Hyperlipidemia, n (%)	23 (92)
Ischemic heart disease, n (%)	3 (12)
Stroke, n (%)	1 (4)
Retinopathy, n (%)	12 (48)
Neuropathy, n (%)	4 (16)
Nephropathy, n (%)	6 (24)
Peripheral vascular disease, n (%)	2 (8)

[a]GEMINI-T2D: Glucose Monitoring and Intervention in Insulin-treated Type 2 Diabetes Patients.
[b]HbA_{1c}: hemoglobin A_{1c}.

HbA_{1c} Level and Other Glycemic Variables

Mean baseline HbA_{1c} level and FPG were 8.6% (SD 0.7%) and 8.7 (SD 2.0) mmol/L, respectively (Table 2). The mean baseline TDD of insulin was 57.3 (SD 24.4) units/day or 0.73 (SD 0.31) units/kg/day. All 25 participants experienced a reduction in HbA_{1c} level of ≥0.4% at 24 weeks. Mean HbA_{1c} level decreased by 1.2% over 24 weeks, from 8.6% (SD 0.7%) to 7.4% (SD 0.6%; $P<.001$; Figure 1). From weeks 0 - 12, the mean HbA_{1c} level decreased by 0.8% ($P<.001$), and from weeks 13 - 24, the mean HbA_{1c} level decreased by 0.4% ($P<.001$). Individuals with a higher baseline HbA_{1c} level experienced a greater absolute decrease in HbA_{1c} level over 24 weeks ($P<.001$; Figure 1).

Mean FPG also decreased by 1.6 mmol/L over 4 weeks, from 8.7 (SD 2.0) mmol/L to 7.1 (SD 1.4) mmol/L ($P<.001$; Figure 1). From weeks 0 - 12, mean FPG decreased by 1.3 mmol/L ($P=.02$), and from weeks 13 - 24, mean FPG decreased by 0.3 mmol/L ($P>.99$).

The mean TDD of insulin increased from 0.73 (SD 0.31) units/kg/day to 0.79 (SD 0.34) units/kg/day ($P=.01$; Figure 1). From weeks 0 - 12, the mean TDD of insulin increased by 0.05 units/kg/day ($P=.02$), and from weeks 13 - 24, the mean TDD of insulin increased by 0.01 units/kg/day ($P>.99$). BMI rose from 29.0 kg/m^2 to 29.5 kg/m^2 over 24 weeks ($P=.03$).

XSL•FO
RenderX

Table . Changes in key glycemic outcomes, insulin doses, and BMI from baseline to weeks 12 and 24 in the GEMINI-T2D[a] study, which evaluated the efficacy of a web-based, algorithm-guided insulin titration system in adults with insulin-treated type 2 diabetes at the Singapore General Hospital from September 2020 to May 2022. *P* values indicate trends over the study period.

Variable	Baseline, mean (SD)	Week 12, mean (SD)	Week 24, mean (SD)	*P* value for trend
HbA_{1c}[b] level (%)	8.6 (0.7)	7.8 (0.6)	7.4 (0.6)	<.001
HbA_{1c} level (mmol/mol)	70.4 (7.7)	61.4 (7.0)	57.3 (6.7)	<.001
Fasting plasma glucose (mmol/L)	8.7 (2.0)	7.4 (1.9)	7.1 (1.4)	<.001
Total daily dose of insulin (units/day)	57.3 (24.4)	61.3 (25.7)	62.1 (25.4)	.01
Total daily dose of insulin (units/kg/day)	0.73 (0.31)	0.78 (0.33)	0.79 (0.34)	.007
BMI (kg/m^2)	29.0 (3.6)	29.3 (3.6)	29.5 (3.6)	.04

[a]GEMINI-T2D: Glucose Monitoring and Intervention in Insulin-treated Type 2 Diabetes Patients.

[b]HbA_{1c}: hemoglobin A_{1c}.

XSL•FO
RenderX

Figure 1. Trends in (A) hemoglobin A_{1c} (HbA_{1c}) level, (B) mean change in HbA_{1c} level stratified by baseline HbA_{1c} level, (C) fasting plasma glucose (FPG), and (D) total daily dose (TDD) of insulin (units/kg/day) during the 24-week GEMINI-T2D study. The study assessed the impact of algorithm-guided insulin titration delivered via a web-based telehealth platform in 25 adults with insulin-treated type 2 diabetes at the Singapore General Hospital from September 2020 to May 2022. GEMINI-T2D: Glucose Monitoring and Intervention in Insulin-treated Type 2 Diabetes Patients.

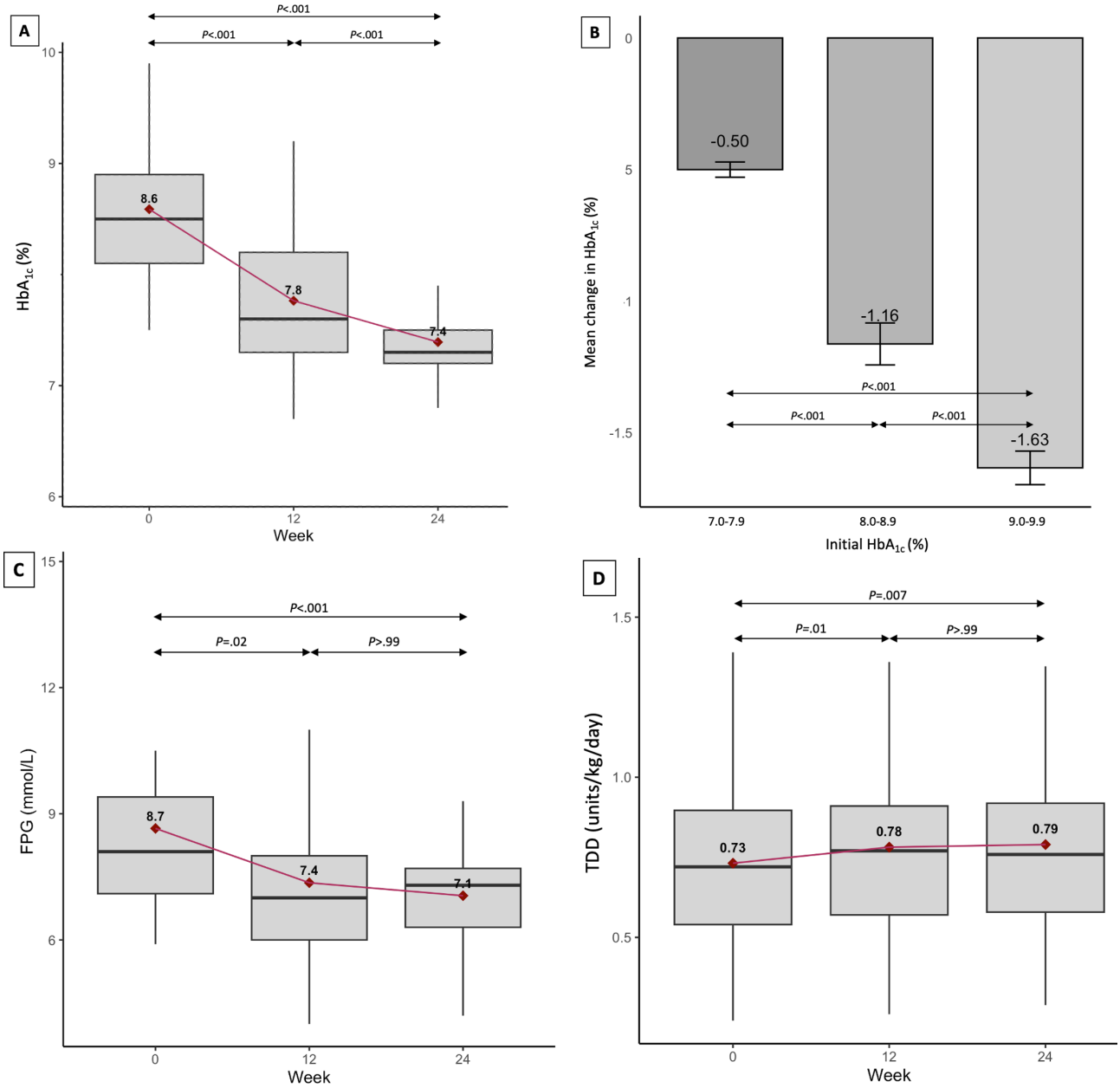

Hypoglycemia and Adherence to SMBG

At baseline, 44% (11/25) of participants reported at least 1 episode of hypoglycemia in the preceding month. During our study, 48% (12/25) of participants experienced at least 1 episode of hypoglycemia. In total, 24 episodes of hypoglycemia occurred in weeks 0 - 12, while 72 episodes occurred in weeks 13 - 24. The episodes of hypoglycemia were predominantly confined to level 1 (mild; blood glucose 3.0 - 3.9 mmol/L), with 85 (89%) out of 96 episodes falling in this category (Table 3). There were no level-3 (severe) episodes of hypoglycemia reported during the study. Adherence to the prescribed SMBG frequency was 98.2% (165/168) for weeks 0 - 12 and 96.5% (162/168) for weeks 12 - 24.

XSL•FO
RenderX

Table . Incidence and severity of hypoglycemia episodes per participant during the 24-week GEMINI-T2D[a] study. Episodes were categorized by level: mild (3.0 - 3.9 mmol/L), moderate (<3.0 mmol/L), and severe (requiring external assistance). Data highlight an increase in mild hypoglycemia episodes from weeks 0 - 12 to weeks 13 - 24. The study was conducted in 25 adults with insulin-treated type 2 diabetes using a web-based telehealth platform at the Singapore General Hospital from September 2020 to May 2022.

Level of hypoglycemia	Incidence of hypoglycemia	
	Weeks 1 - 12	Weeks 13 - 24
1 (mild)	0.88	2.52
2 (moderate)	0.08	0.36
3 (severe)	0	0

[a]GEMINI-T2D: Glucose Monitoring and Intervention in Insulin-treated Type 2 Diabetes Patients.

Discussion

The GEMINI-T2D study demonstrated that remote monitoring of SMBG data via a glucose management platform, coupled with algorithm-guided insulin titration, may improve glycemic control in patients with insulin-treated T2D. To the best of our knowledge, this is the first study of an integrated, mobile phone–based solution in an Asian cohort with insulin-treated diabetes. The degree of HbA_{1c} improvement in our study was comparable to studies on smartphone-based diabetes management platforms [15,18,22] and a larger trial on an automated insulin titration system [19].

Several key factors likely contributed to the improvement in glycemic control in our cohort. The ALRT system empowered individuals to participate in structured SMBG, through a platform that allowed for ease of upload of blood glucose readings directly from a glucometer. It also provided a convenient way to store, visualize, and communicate the data with their physicians, receiving timely feedback. The weekly review of data followed by insulin dose adjustments also tightened the feedback loop between patients and their physicians, enabling more prompt insulin dose adjustments than usually possible in a clinic setting. Physicians are also likely to have been empowered by the algorithm-guided insulin dose recommendations, which provide a framework to base their therapeutic decisions upon.

In addition, we observed that the improvement in HbA_{1c} level during the first half of the study persisted into the second half, demonstrating sustained benefits from the intervention. Interestingly, although the average increase in TDD of insulin was only 0.05 units/kg/day or 3 units/day for a 60-kg person, HbA_{1c} level decreased by 1.7%. This modest increase in insulin dose alone is unlikely to account for such a substantial improvement in glycemia. We speculate that the “outsized” glycemic improvement may be attributed to a more optimal redistribution of insulin doses facilitated by structured SMBG.

Another contributing factor to the improved glycemic control could be enhanced patient empowerment, which may have led to positive changes in overall health, diet, and lifestyle. This is supported by the high adherence to the prescribed SMBG regimen observed in our cohort, exceeding rates typically reported in the literature [23]. Enhanced patient engagement, consistently recognized as a critical factor in managing chronic diseases such as diabetes [24-26], was a central focus of our intervention. This engagement was fostered through patient education, twice-daily SMBG, and frequent review with feedback from physicians.

With regard to the safety of our intervention, while an increase in level-1 (mild) episodes of hypoglycemia occurred with tighter glycemic control, we found a low rate of level-2 (moderate) hypoglycemia, with no episodes of level-3 (severe) hypoglycemia. This safety signal was likely contributed by a few measures, such as patient education on hypoglycemia recognition and management at the time of enrollment, along with automated alerts with algorithm-guided insulin dosing suggestions to the treating physician via the ALRT app in event of hypoglycemia, prompting timely adjustment of insulin doses ahead of the scheduled weekly SMBG reviews. The safety profile aligns with a US-based randomized controlled trial [19] that showed similar hypoglycemia rates with automated insulin titration guidance compared to standard care. Nevertheless, dynamic adjustments to the algorithm to minimize hypoglycemia rates will be warranted in future implementation of this intervention in other populations.

A web-based platform with algorithm-guided insulin dose titration holds promise for closing the titration gap for people on insulin therapy, offering dynamic and frequent adjustments based on data, while assisting busy physicians in delivering necessary care safely and conveniently. Moving forward, it is also fathomable that artificial intelligence may allow for yet larger scales of SMBG data management and analysis, and provide more personalized insulin dosing recommendations that take into account factors beyond glycemic trend such as diet, lifestyle, and general health.

The primary limitation of our study is a small sample size, which reflects its design as a pilot study of an intervention. While the pre-post intervention design of our study suggests the effectiveness of our intervention compared to prior standard care, a longer study with a larger sample size would be warranted to evaluate the long-term effectiveness and scalability of such a web-based platform in diabetes care. Incorporating a comparison arm would further clarify the intervention’s performance against standard care and help identify the factors contributing to the improved glycemic control observed in our cohort. Additionally, our strict inclusion criteria may limit the generalizability of our findings to a broader population with diabetes, warranting further exploration. Finally, understanding physicians’ experiences with this intervention would provide

XSL•FO
RenderX

valuable insights into the practical challenges and opportunities in its integration into routine clinical practice.

In conclusion, patients with insulin-treated T2D initiated on a web-based glucose monitoring system with algorithm-guided dosing recommendations showed significant improvement in glycemic control compared to baseline. High adherence rates underscore the feasibility of integrating algorithm-guided insulin titration into routine care. While hypoglycemia incidence rose slightly, episodes were predominantly mild, and no severe events occurred. This intervention shows promise for broader adoption in T2D management, particularly in resource-constrained settings.

Acknowledgments

No generative artificial intelligence (AI) or AI-assisted technologies were used in this study. This research was supported by A*STAR, Singapore, under its Industry Alignment Pre-Positioning Fund (grant H19/01/a0/023 – Diabetes Clinic of the Future), and ALR Technologies.

Data Availability

The datasets generated during and/or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Insulin dose titration algorithm.
[DOCX File, 15 KB - formative_v9i1e68914_app1.docx]

References

1. GBD 2021 Diabetes Collaborators. Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: a systematic analysis for the Global Burden of Disease Study 2021. Lancet 2023 Jul 15;402(10397):203-234. [doi: 10.1016/S0140-6736(23)01301-6] [Medline: 37356446]
2. Trikkalinou A, Papazafiropoulou AK, Melidonis A. Type 2 diabetes and quality of life. World J Diabetes 2017 Apr 15;8(4):120-129. [doi: 10.4239/wjd.v8.i4.120] [Medline: 28465788]
3. Magliano DJ, Boyko EJ, IDF Diabetes Atlas 10th Edition Scientific Committee. IDF Diabetes Atlas, 10th Edition: International Diabetes Federation; 2021. URL: https://www.ncbi.nlm.nih.gov/books/NBK581934/ [accessed 2025-01-20]
4. Diabetes Control and Complications Trial Research Group, Nathan DM, Genuth S, et al. The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. N Engl J Med 1993 Sep 30;329(14):977-986. [doi: 10.1056/NEJM199309303291401] [Medline: 8366922]
5. Ohkubo Y, Kishikawa H, Araki E, et al. Intensive insulin therapy prevents the progression of diabetic microvascular complications in Japanese patients with non-insulin-dependent diabetes mellitus: a randomized prospective 6-year study. Diabetes Res Clin Pract 1995 May;28(2):103-117. [doi: 10.1016/0168-8227(95)01064-k] [Medline: 7587918]
6. Holman RR, Paul SK, Bethel MA, Matthews DR, Neil HAW. 10-year follow-up of intensive glucose control in type 2 diabetes. N Engl J Med 2008 Oct 9;359(15):1577-1589. [doi: 10.1056/NEJMoa0806470] [Medline: 18784090]
7. Mannucci E, Monami M, Dicembrini I, Piselli A, Porta M. Achieving HbA1c targets in clinical trials and in the real world: a systematic review and meta-analysis. J Endocrinol Invest 2014 May;37(5):477-495. [doi: 10.1007/s40618-014-0069-6] [Medline: 24699860]
8. Benjamin EM. Self-monitoring of blood glucose: the basics. Clin Diabetes 2002 Jan 1;20(1):45-47. [doi: 10.2337/diaclin.20.1.45]
9. Karter AJ, Ackerson LM, Darbinian JA, et al. Self-monitoring of blood glucose levels and glycemic control: the Northern California Kaiser Permanente Diabetes registry. Am J Med 2001 Jul;111(1):1-9. [doi: 10.1016/s0002-9343(01)00742-2] [Medline: 11448654]
10. Polonsky WH, Fisher L, Schikman CH, et al. Structured self-monitoring of blood glucose significantly reduces A1C levels in poorly controlled, noninsulin-treated type 2 diabetes: results from the Structured Testing Program study. Diabetes Care 2011 Feb;34(2):262-267. [doi: 10.2337/dc10-1732] [Medline: 21270183]
11. Parkin CG, Buskirk A, Hinnen DA, Axel-Schweitzer M. Results that matter: structured vs. unstructured self-monitoring of blood glucose in type 2 diabetes. Diabetes Res Clin Pract 2012 Jul;97(1):6-15. [doi: 10.1016/j.diabres.2012.03.002] [Medline: 22464874]
12. Lalić NM, Lalić K, Jotić A, et al. The impact of structured self-monitoring of blood glucose combined with intensive education on HbA1c levels, hospitalizations, and quality-of-life parameters in insulin-treated patients with diabetes at

XSL•FO
RenderX

primary care in Serbia: the multicenter SPA-EDU study. J Diabetes Sci Technol 2017 Jul;11(4):746-752. [doi: 10.1177/1932296816681323] [Medline: 28625073]
13. Chow N, Shearer D, Aydin Plaa J, et al. Blood glucose self-monitoring and internet diabetes management on A1c outcomes in patients with type 2 diabetes. BMJ Open Diabetes Res Care 2016 Apr 29;4(1):e000134. [doi: 10.1136/bmjdrc-2015-000134] [Medline: 27158516]
14. Chircop J, Sheffield D, Kotera Y. Systematic review of self-monitoring of blood glucose in patients with type 2 diabetes. Nurs Res 2021;70(6):487-497. [doi: 10.1097/NNR.0000000000000542] [Medline: 34292228]
15. Yu Y, Yan Q, Li H, et al. Effects of mobile phone application combined with or without self-monitoring of blood glucose on glycemic control in patients with diabetes: a randomized controlled trial. J Diabetes Investig 2019 Sep;10(5):1365-1371. [doi: 10.1111/jdi.13031] [Medline: 30815973]
16. Cui M, Wu X, Mao J, Wang X, Nie M. T2DM self-management via smartphone applications: a systematic review and meta-analysis. PLoS One 2016 Nov 18;11(11):e0166718. [doi: 10.1371/journal.pone.0166718] [Medline: 27861583]
17. Katalenich B, Shi L, Liu S, et al. Evaluation of a remote monitoring system for diabetes control. Clin Ther 2015 Jun 1;37(6):1216-1225. [doi: 10.1016/j.clinthera.2015.03.022] [Medline: 25869625]
18. Rama Chandran S, Tan HC, Chen Q, et al. Telemonitoring with a connected glucose meter improves glycemia among people with insulin-treated type 2 diabetes. J Diabetes Sci Technol 2023 Jul;17(4):909-915. [doi: 10.1177/19322968231157387] [Medline: 36825611]
19. Bergenstal RM, Johnson M, Passi R, et al. Automated insulin dosing guidance to optimise insulin management in patients with type 2 diabetes: a multicentre, randomised controlled trial. Lancet 2019 Mar 16;393(10176):1138-1148. [doi: 10.1016/S0140-6736(19)30368-X] [Medline: 30808512]
20. Kalra S, Balhara YPS, Sahay BK, Ganapathy B, Das AK. Why is premixed insulin the preferred insulin? novel answers to a decade-old question. J Assoc Physicians India 2013 Jan;61(1 Suppl):9-11. [Medline: 24482980]
21. Chinese Elderly Type 2 Diabetes Prevention and Treatment of Clinical Guidelines Writing Group, Geriatric Endocrinology and Metabolism Branch of Chinese Geriatric Society, Geriatric Endocrinology and Metabolism Branch of Chinese Geriatric Health Care Society, Geriatric Professional Committee of Beijing Medical Award Foundation, National Clinical Medical Research Center for Geriatric Diseases (PLA General Hospital). Clinical guidelines for prevention and treatment of type 2 diabetes mellitus in the elderly in China [Article in Chinese]. Zhonghua Nei Ke Za Zhi 2022 Jan 1;61:12-50. [doi: 10.3760/cma.j.cn112138-20211027-00751] [Medline: 34979769]
22. Zhou W, Chen M, Yuan J, Sun Y. Welltang - a smart phone-based diabetes management application - Improves blood glucose control in Chinese people with diabetes. Diabetes Res Clin Pract 2016 Jun;116:105-110. [doi: 10.1016/j.diabres.2016.03.018] [Medline: 27321324]
23. Hu ZD, Zhang KP, Huang Y, Zhu S. Compliance to self-monitoring of blood glucose among patients with type 2 diabetes mellitus and its influential factors: a real-world cross-sectional study based on the Tencent TDF-I blood glucose monitoring platform. Mhealth 2017 Jun 27;3:25. [doi: 10.21037/mhealth.2017.06.01] [Medline: 28736734]
24. Barello S, Graffigna G, Vegni E. Patient engagement as an emerging challenge for healthcare services: mapping the literature. Nurs Res Pract 2012;2012:905934. [doi: 10.1155/2012/905934] [Medline: 23213497]
25. Forbat L, Cayless S, Knighting K, Cornwell J, Kearney N. Engaging patients in health care: an empirical study of the role of engagement on attitudes and action. Patient Educ Couns 2009 Jan;74(1):84-90. [doi: 10.1016/j.pec.2008.07.055] [Medline: 18790594]
26. Chang YT, Tu YZ, Chiou HY, Lai K, Yu NC. Real-world benefits of diabetes management app use and self-monitoring of blood glucose on glycemic control: retrospective analyses. JMIR Mhealth Uhealth 2022 Jun 15;10(6):e31764. [doi: 10.2196/31764] [Medline: 35704371]

Abbreviations

FPG: fasting plasma glucose
GEMINI-T2D: Glucose Monitoring and Intervention in Insulin-treated Type 2 Diabetes Patients
HbA_{1c}: hemoglobin A_{1c}
SMBG: self-monitoring of blood glucose
T2D: type 2 diabetes
TDD: total daily dose

XSL•FO
RenderX

Edited by A Mavragani; submitted 17.11.24; peer-reviewed by A Andersen, X Gao; revised version received 28.12.24; accepted 29.12.24; published 07.02.25.

Please cite as:
Thiagarajan N, Tan HC, Rama Chandran S, Lee PC, Chin YA, Zeng W, Ho ETL, Carmody D, Goh SY, Bee YM
Web-Based, Algorithm-Guided Insulin Titration in Insulin-Treated Type 2 Diabetes: Pre-Post Intervention Study
JMIR Form Res 2025;9:e68914
URL: https://formative.jmir.org/2025/1/e68914
doi:10.2196/68914

© Nishanth Thiagarajan, Hong Chang Tan, Suresh Rama Chandran, Phong Ching Lee, Yun Ann Chin, Wanling Zeng, Emily Tse Lin Ho, David Carmody, Su-Yen Goh, Yong Mong Bee. Originally published in JMIR Formative Research (https://formative.jmir.org), 7.2.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Assessment of Digital Capabilities by 9 Countries in the Alliance for Healthy Cities Using AI: Cross-Sectional Analysis

Hocheol Lee, PhD

Department of Health Administration, College of Software and Digital Healthcare Convergence, Yonsei University, PO Box 26493, Wonju, Repubic of Korea

Corresponding Author:
Hocheol Lee, PhD
Department of Health Administration, College of Software and Digital Healthcare Convergence, Yonsei University, PO Box 26493, Wonju, Repubic of Korea

Abstract

Background: The Alma-Ata Declaration of 1978 initiated a global focus on universal health, supported by the World Health Organization (WHO) through healthy cities policies. The concept emerged at the 1984 Toronto "Beyond Health Care" conference, leading to WHO's first pilot project in Lisbon in 1986. The WHO continues to support regional healthy city networks, emphasizing digital transformation and data-driven health management in the digital era.

Objective: This study explored the capabilities of digital healthy cities within the framework of digital transformation, focusing on member countries of the Asian Forum of Healthy Cities. It examined the cities' preparedness and policy needs for transitioning to digital health.

Methods: A cross-sectional survey was conducted of 9 countries—Australia, Cambodia, China, Japan, South Korea, Malaysia, Mongolia, the Philippines, and Vietnam—from August 1 to September 21, 2023. The 6-section SPIRIT (setting approach and sustainability; political commitment, policy, and community participation; information and innovation; resources and research; infrastructure and intersectoral; and training) checklist was modified to assess healthy cities' digital capabilities. With input from 3 healthy city experts, the checklist was revised for digital capabilities, renaming "healthy city" to "digital healthy city." The revised tool comprises 8 sections with 33 items. The survey leveraged ChatGPT (version 4.0; OpenAI, Microsoft), accessed via Python (Python Software Foundation) application programming interface. The *openai* library was installed, and an application programming interface key was entered to use ChatGPT (version 4.0). The "GPT-4 Turbo" model command was applied. A qualitative analysis of the collected data was conducted by 5 healthy city experts through group deep-discussions.

Results: The results indicate that these countries should establish networks and committees for sustainable digital healthy cities. Cambodia showed the lowest access to electricity (70%) and significant digital infrastructure disparities. Efforts to sustain digital health initiatives varied, with countries such as Korea focusing on telemedicine, while China aimed to build a comprehensive digital health database, highlighting the need for tailored strategies in promoting digital healthy cities. Life expectancy was the highest in the Republic of Korea and Japan (both 84 y). Access to electricity was the lowest in Cambodia (70%) with the remaining countries having had 95% or higher access. The internet use rate was the highest in Malaysia (97.4%), followed by the Republic of Korea (97.2%), Australia (96.2%), and Japan (82.9%).

Conclusions: This study highlights the importance of big data-driven policies and personal information protection systems. Collaborative efforts across sectors for effective implementation of digital healthy cities. The findings suggest that the effectiveness of digital healthy cities is diminished without adequate digital literacy among managers and users, suggesting the need for policies to improve digital literacy.

(JMIR Form Res 2025;9:e62935) doi:10.2196/62935

KEYWORDS

digital capabilities; digital health cities; digital transformation; Asian Forum of Healthy Cities; assessment; digital health; artificial intelligence; AI; World Health Organization; WHO; healthy city; data; health management; digital era; qualitative analysis; cross-sectional survey; database; digital health database; effectiveness; digital literacy

Introduction

The Alma-Ata Declaration of 1978 marked a global shift toward universal health, a movement endorsed by the World Health Organization (WHO) through its advocacy for healthy cities as a strategic policy to enhance global health [1]. The idea of a healthy city was first conceptualized at the 1984 Toronto "Beyond Health Care" conference, and the first pilot project was spearheaded by the WHO European Office in Lisbon in April 1986 [2]. Subsequently, the International Health Cities Committee was established, and over 3000 cities now have healthy cities policies. The WHO remains a steadfast supporter

XSL•FO
RenderX

of regional healthy city networks. In Europe, the WHO European Healthy Cities Network, founded in 1988, continuously promotes networking through assemblies every 5 years, involving over 1400 municipalities [3]. In North America, a network was established with Canada as the central player in 1984, which now includes over 200 North American cities. In the Western Pacific region, the Alliance for Healthy Cities (AFHC) was established in 2003. It certifies and manages healthy cities as alliance members, facilitating dynamic networking through biennial general assemblies [4]. In the Western Pacific region, 9 countries—Australia, Cambodia, China, Japan, the Republic of Korea, Malaysia, Mongolia, the Philippines, and Vietnam—are actively participating as full members of the AFHC. The AFHC is an international network focused on protecting and promoting the health of urban residents through collaborative efforts.

Healthy cities are dedicated to improving physical and social environments where all community members collaborate to enhance the health and quality of life of citizens [5]. The evaluation of healthy cities is categorized into distinct areas: personal, community, socioeconomic infrastructure, environment, and evaluation management. These domains all adopt a top-down approach, underscoring the pivotal role of policy makers in formulating and executing policies and initiatives for healthy cities [6]. Research has shown that the proactive stance of the city's leaders has a positive impact on community residents' health [7]. Additionally, healthy city projects must be developed considering the specific contemporary and environmental demands of the country. Worldwide, healthy cities are adopting unique characteristics, transitioning to models such as senior-friendly cities, smart healthy cities, digital healthy cities, and green healthy cities [8,9]. Following the coronavirus disease 2019 pandemic, there was an acceleration in digital transformation, a trend pivotal in the context of healthy cities. The WHO highlighted that healthy cities in the digital era should focus on data-driven monitoring and the implementation of digital platforms for managing the public's health [10]. With the advancement of digital technology, it has become crucial for healthy cities to prioritize the development of digital competencies.

The WHO's "Global Strategy on Digital Health 2020 - 2025" emphasizes the role of healthy cities in the digital age, including citizen-led health management; policy making based on big data; and the expansion of digital platforms, including telemedicine [11]. Particularly, the strategy underscores the importance of health management through information technology devices and the significance of digital health at the national level [11]. Additionally, the AFHC stresses the need to integrate aging, smart technology, and digital solutions in healthy city strategies for the Western Pacific region. It prioritizes digital-focused strategies for advancing healthy cities [12].

The world is preparing for the digital era through evident forms of digital transformation such as digital healthy cities and smart healthy cities. Previous studies have stressed the importance of equipping nations and cities with digital competency readiness to prepare for this transformation [13]. However, according to United Nations digital indicators, there are disparities in digital capabilities among countries in the Western Pacific region and the AFHC member countries [14].

Several methods are used to measure the capabilities of healthy cities, such as the WHO European Healthy City Evaluation Model and the AFHC SPIRIT (setting approach and sustainability political commitment, policy, and community participation; information and innovation; resources and research; infrastructure and intersectoral; and training) checklist [15]. These 2 models, conceptualized and evaluated by healthy city experts, provide a structured, logical framework for objectively assessing various aspects of a city, including its health status, capabilities, lifestyle, vision, strategy, and network. Healthy cities worldwide use these models, and they have recently been used to understand the capabilities of healthy cities in the digital era.

In the context of digital transformation, this study aims to investigate the capabilities of digital healthy cities in all AFHC member countries, exploring policy implications for areas that require further development and readiness to effectively promote the transition to digital healthy cities.

Methods

Study Design

This cross-sectional study aims to inspect the digital healthy city capabilities among AFHC member countries. Nine countries were included: Australia, Cambodia, China, Japan, the Republic of Korea, Malaysia, Mongolia, the Philippines, and Vietnam.

Instruments

We used a modified version of the SPIRIT checklist developed and used by AFHC to assess the capabilities of healthy cities [16,17]. The checklist consists of 6 sections: setting approach and sustainability, political commitment, policy, and community participation, information and innovation, resources and research, infrastructure and intersectoral, and training. We modified the tool with input from 3 healthy city experts to more specifically evaluate digital capabilities, particularly changing the term "healthy city" used in the original tool to "digital healthy city." The modified tool contained 8 sections with 33 items.

Data Collection

Data were collected using a database powered by a large-scale artificial intelligence (AI) application. ChatGPT (version 4.0), developed by OpenAI, was accessed via an application programming interface in Python and used to conduct our survey. We installed the *openai* library in Python and entered an application programming interface key issued for the use of ChatGPT (version 4.0). In the *openai* library, the "GPT-4 Turbo" model command was used; GPT-4 Turbo is a large AI language model that has acquired knowledge up to March 2023.

The query command was set up to allow the input of questions. Additionally, the "message" command was used to define the conversation's history for generation. The system option facilitated the setup of a chatbot helper for conversation options, and messages were relayed directly to the helper via the assistant. The user option was used to save previous

XSL•FO
RenderX

conversations and maintain dialogue continuity. The results were made available through the answer command.

The questions on the SPIRIT checklist were revised to reflect names of the 9 surveyed AFHC member countries. The survey was conducted from August 1 to September 21, 2023, and the collected data were stored in a report. No human participants were involved, and thus institutional review board (IRB) approval was not obtained.

Data Analysis

Five healthy city experts performed a qualitative analysis on the collected data through a group deep discussion. The experts were recruited from October 1, 2023, to November 3, 2023. They had experience conducting activities within the AFHC International Network, held PhDs, and were affiliated with healthy city networks in their respective countries. Specifically, the group comprised 3 public health PhDs from Yonsei University's Healthy City Research Center and two professors who are academic members of the Healthy City Society. A consultation meeting was held on November 13, 2023. Prior to the qualitative analysis meeting, written informed consent was obtained from the 5 experts; the consent covered this study's purpose and plan and the objectives of the qualitative analysis meeting. Through this process, we obtained the experts' advice on the appropriateness of the contents of the SPIRIT checklist.

First, the collected reports were summarized and organized based on key areas on the checklist. Second, a relational analysis was conducted of keywords from the summarized reports to visualize the results. Third, 3 digital healthy city experts conducted a final review and discussed the digital capabilities and directions for each country. The indicators for healthy cities in the 9 surveyed countries (Table 1) were analyzed using the official AFHC data, World Bank's DataBank, and International Telecommunication Union Data Hub [14,18,19].

XSL•FO
RenderX

Table . AFHC[a] member cities.

	AUS[b]	KHM[c]	CHN[d]	JPN[e]	KOR[f]	MYS[g]	MNG[h]	PHL[i]	VNM[j]
Number of healthy cities, n[k]	5	1	34	31	109	1	5	12	1
Population, million (year)	25.9 (2021)	16.7 (2022)	1410 (2022)	125.1 (2022)	51.6 (2022)	33.9 (2022)	3.3 (2022)	115.5 (2022)	98.1 (2022)
GDP[l] per capita, US $ (year)	64,491 (2022)	1786 (2022)	12,720 (2022)	33,815 (2022)	32,254 (2022)	11,971 (2022)	4946 (2022)	3498 (2022)	4163 (2022)
GDP growth, annual % (year)	3.6 (2022)	1.1 (2022)	3 (2022)	−0.4 (2022)	2.6 (2022)	8.7 (2022)	4.8 (2022)	7.6 (2022)	8 (2022)
Poverty headcount ratio, % (year)	0.5 (2018)	N/A[m]	0.1 (2019)	0.7 (2013)	0.2 (2016)	0 (2018)	0.7 (2018)	3 (2018)	0.7 (2018)
Life expectancy at birth, total years (year)	83 (2021)	70 (2021)	78 (2021)	84 (2021)	84 (2021)	75 (2021)	71 (2021)	69 (2021)	71 (2021)
Access to electricity, %	100	82	100	100	100	100	100	97	100
Safely managed sanitation services, %	74	N/A	70	81	100	77	56	61	N/A
Internet use rate, %	96.2	60.2	75.6	82.9	97.2	97.4	81.6	52.7	78.6

[a]AFHC: Alliance for Healthy Cities.
[b]AUS: Australia.
[c]KHM: Cambodia.
[d]CHN: China.
[e]JPN: Japan.
[f]KOR: the Republic of Korea.
[g]MYS: Malaysia.
[h]MNG: Mongolia.
[i]PHL: the Philippines.
[j]VNM: Vietnam.
[k]The number of cities enrolled in the Healthy Cities initiative in the respective country.
[l]GDP: gross domestic product.
[m]N/A: not available.

Ethical Considerations

All components of this survey were approved by the IRB of Yonsei University in Korea (IRB 1041849-202401-SB-018-01). This study does not involve human subjects research; therefore, no compensation procedures were required, and there are no issues related to privacy or confidentiality. These aspects were thoroughly reviewed and confirmed by the Yonsei University IRB.

Results

AFHC Member Countries

Of the 9 AFHC member countries included in this study, the Republic of Korea had the largest number of healthy cities (109), followed by China (34) and Japan (31). China had the largest population (1410 million), followed by Japan (125.1 million) and the Philippines (115.5 million). Gross domestic product per capita was the highest in Australia (64,491), followed by Japan (33,815) and Korea (32,254). Life expectancy was the highest in the Republic of Korea and Japan (both 84 y). Access to

XSL•FO
RenderX

electricity was the lowest in Cambodia (70%) with the remaining countries having 95% or higher access. The internet use rate was the highest in Malaysia (97.4%), followed by the Republic of Korea (97.2%), Australia (96.2%), and Japan (82.9%).

SPIRIT Approach for the Digital Era

Setting and Sustainability (S)

Digital healthy cities were evaluated in terms of setting and sustainability. The findings indicated that Australia, Cambodia, and the Republic of Korea must ensure sustainability through telemedicine, while China should establish a digital healthy city database by integrating the public's electronic medical records. Malaysia, Mongolia, and the Philippines need to sustain their digital healthy cities by establishing digital platform settings for communities (Table 2).

Table . Setting and sustainability (S) of the SPIRIT[a] checklist by country.

Country	Content
Australia	• Sustainability through telemedicine and digital health networking • Sustainable digital healthy city through digital medical infrastructure
Cambodia	• Provide high-quality medical service by addressing the digital medical disparity • Expand sustainability by providing stable high-speed internet
China	• Digital health database that integrates the public's electronic medical records • Health application, wearable, and digital health education
Japan	• Sustainable urban planning through exemplary cases of digital healthy cities • Policy decisions based on integration, education, and collaboration-focused data
The Republic of Korea	• Establish a telemedicine system by digitalizing the existing health care facilities • Establish systems to protect data and prevent leakage of personal information
Malaysia	• Digital healthy city approach through digital technology and public health collaboration • Secure data by strategically placing smart kiosks in the communities
Mongolia	• Establish a digital healthy city hub through a digital platform • Expand sustainability through digital health experts and digital coordinators
The Philippines	• Digital healthy city with community participation • Ensure sustainability by providing stable high-speed internet
Vietnam	• Establish a sustainable digital healthy city by installing data protection measures • Evidence-based digitally healthy city powered by research

[a]SPIRIT: setting approach and sustainability; political commitment, policy, and community participation; information and innovation; resources and research; infrastructure and intersectoral; and training.

Political Commitment, Policy, and Community Participation (P)

We examined political commitment, policy, and community participation in the context of digital healthy cities across 9 countries. The findings emphasized interdepartmental collaboration involving private organizations, schools, and civil institutions. Cambodia, China, and the Philippines referred to building digital healthy cities through community participation via digital platforms. Additionally, the Republic of Korea is implementing digital healthy cities by disclosing health-related public data (Table 3).

XSL•FO
RenderX

Table . Political commitment, policy, and participation (P) of the SPIRIT[a] checklist by country.

Country	Content
Australia	• Health in all policies (health care, environment, and education) • Interdepartmental collaboration • Competency skills training • Data protection • Monitoring
Cambodia	• Digital leadership • Interdepartmental collaboration • Strengthen competency for digital health • Community involvement (digital platform) • Partnership with the technology industry
China	• Community involvement (digital platform) • Intensive training of promising talents in each department • Centralized and comprehensive policy approach through digital healthy cities • Comprehensive and encompassing digital healthy city framework
Japan	• Health in All Policy approach • Digital healthy city policies centered on transportation, education, housing, and digital infrastructure • All departments demand to receive a health impact evaluation • Health in All Policy approach by forming an interdepartmental committee
The Republic of Korea	• Dedicated task force team for Health in All Policy approach • Open data policy to allow access to health-related data • Digital healthy city education for all personnel in charge of health-related matters in all departments • Accelerate the transition to a healthy city by installing digital devices in communities
Malaysia	• Establish digital health laboratories to make effective policies • Share policies among various departments and set common health goals • Devise digital healthy city policies to preserve the environment
Mongolia	• Integration of various areas in digital healthy city policies • Bolster responsibility of digitally healthy city policy makers • Build public-private partnership
The Philippines	• Propose digital healthy city policies through interspecialty collaboration, including health, transportation, education, environment, and information technology • Digital healthy city through digital health platforms
Vietnam	• Digital healthy city policy decisions based on data • Feedback from digital healthy city experts every 2 years

[a]SPIRIT: setting approach and sustainability; political commitment, policy, and community participation; information and innovation; resources and research; infrastructure and intersectoral; and training.

Information and Innovation (I)

We analyzed digital healthy cities in terms of information and innovation. Australia has approached this from a data-centralization perspective. Cambodia has provided telemedicine services through digital health kiosks. China is using digital mental health platforms such as lifestyle fitness apps, green technology, and AI diagnostic tools. Japan is enhancing the digital-use competencies of older adults in their aged society and is planning innovative digital healthy cities based on their needs, such as promoting digital detox weekend campaigns. The Republic of Korea is realizing digital healthy cities with AI platforms for mental health counseling and integrated electronic medical record systems. Malaysia is installing digital kiosks where community residents can obtain instant health information, training community health professionals for this purpose. Mongolia has recognized the need to build digital healthy cities for nomadic communities by facilitating mobile health and telemedicine participation. The Philippines provides telemedicine consultations through the digital health access drive and aims to develop digital healthy

XSL•FO
RenderX

cities based on health care big data. Vietnam is developing digital healthy cities based on data and the digital city profile (Table 4).

Table . Information and innovation (I) of the SPIRIT[a] checklist by country.

Country	Content
Australia	• Address urban noncommunicable diseases and mental health deterioration • Develop public parks and bike lanes • Require collaboration among various departments • Digital healthy city through centralization (eg, data centralization)
Cambodia	• Telemedicine service by implementing digital health kiosks • Require increasing awareness for a healthy lifestyle • Promote active participation of community members and analyze feedback • Improve urban living environment through public-private collaboration
China	• Digital mental health platform • Lifestyle fitness app • Green technology spaces • Data-based AI[b] diagnostic tool
Japan	• Strengthen older adults' digital competency • Innovative digital healthy city planning using digital technology for older adults • Digital detox weekends campaign
The Republic of Korea	• Digital platform for AI mental health counseling • Establish a real-time digital monitoring system for respiratory disorders • Comprehensive electronic medical record system • Digital program for walking through collaboration between Ministry of Health and Welfare and the Ministry of Land, Infrastructure, and Transport
Malaysia	• Immediately acquire health information by installing digital kiosks in public areas • Digital health education through community health care professionals • Expand digital health by integrating nonhealth sectors, such as education and entertainment
Mongolia	• Develop a digital healthy city system for nomadic communities • Reduce health disparity by promoting mobile health and telemedicine for nomads • Collaboration between the green healthy city and digital healthy city • Develop mobile health care facilities
The Philippines	• Provide telemedicine counseling and health education through the digital health access drive • Develop and analyze the digital city profile • Develop a digital healthy city based on health care big data
Vietnam	• Inspect walking facilities through Geographic Information System • Digital healthy city approach • Digital healthy city based on the digital city profile

[a]SPIRIT: setting approach and sustainability; political commitment, policy, and community participation; information and innovation; resources and research; infrastructure and intersectoral; and training.

[b]AI: artificial intelligence.

Resources and Research (R)

We examined the funding and research for digitally healthy cities. Australia is analyzing the digital health gap and planning research on digital healthy cities in cooperation with the Australian Research Council. Cambodia evaluated its digital infrastructure and defined key performance indicators by mapping health priorities. China monitors health burden

XSL•FO
RenderX

assessments and conducts research to enhance digital literacy while also studying the environment, policy, and regulations for digital healthy cities. Japan approaches digital healthy cities by analyzing digital health accessibility and identifying gaps. Additionally, they are conducting research on telemedicine platforms. The Republic of Korea aims to expand its digital healthy cities through the Korea Healthy City Partnership. The partnership seeks to develop key performance indicators and establish shared objectives for digital healthy cities. Mongolia is developing digital healthy cities with privacy-protected digital health regulations (Table 5).

Table . Resources and research (R) of the SPIRIT[a] checklist by country.

Country	Content
Australia	• Establish priority goals for a digital healthy city • Analyze digital divide • Health outcome metrics • Joint resources investment by public and private sectors • Research on digital healthy city through collaboration with the Australian Research Council
Cambodia	• Networks of major stakeholders (eg, leaders, health care professionals, urban planning, and companies) • Evaluate digital infrastructure (eg, internet, digital device use, and platform) • Health priority mapping • Define key performance indicators for a digitally healthy city
China	• Health burden evaluation monitoring • Research to improve digital literacy • Foster a digital healthy city environment through sociocultural evaluation • Research on policies and regulations for a digital healthy city
Japan	• Develop a framework to evaluate the need for a digital healthy city • Analyze digital health access among various demographic groups • Analyze health disparities for a digital healthy city • Research on artificial intelligence diagnosis and telemedicine platforms
The Republic of Korea	• Expand digital healthy city through the Korea Healthy Cities Partnership • Establish and share a digital healthy city database among cities • Establish plans based on the latest literature and information on digital healthy city
Malaysia	• Develop health initiatives to develop regions with poor digital infrastructure • Analyze digital health disparities and implement a digital healthy city focused on vulnerable regions and groups • Develop key performance indicators and establish common digital healthy city goals
Mongolia	• Understand Mongolia's legal digital health regulations and develop a digital healthy city that protects personal information
The Philippines	• Foster experts through focus group interviews with various stakeholders • Understand the digital health policy environment and innovate regulations • Develop key performance indicators and establish common digital healthy city goals
Vietnam	• Appropriate allocation of resources and funds through financial analysis, technical analysis, and risk assessment of digital healthy cities

[a]SPIRIT: setting approach and sustainability; political commitment, policy, and community participation; information and innovation; resources and research; infrastructure and intersectoral; and training.

XSL•FO
RenderX

Infrastructure and Intersectoral (I)

We examined the infrastructure and intersectoral measures for promoting digital healthy cities. Australia is strengthening the functions and capabilities of departments associated with digital healthy cities, requiring regular workshops for knowledge sharing. Cambodia and China are focusing on enhancing infrastructure through digital healthy city operation committees and collaboration with the Ministry of Health and Ministry of Urban Planning as well as other city administration departments. Japan is establishing a digital healthy city committee and uses local media and digital platforms for communication. The Republic of Korea is emphasizing building infrastructure for digital health companies and networking through digital healthy city conferences. The Philippines is developing infrastructure involving urban information and communication technology developers and coordinators to build networks of environmental experts (Table 6).

Table . Infrastructure and intersectoral (I) of the SPIRIT[a] checklist by country.

Country	Content
Australia	• Need to strengthen the functions and competencies of departments linked to digital healthy cities • Sharing of knowledge and skills through periodic workshops • Campaigning and promotions through online platforms and the internet
Cambodia	• Bolster infrastructure through the activities of the digital healthy city operation committee • Cambodia digital healthy city initiative operation committee needs to collaborate with the Ministry of Health, Ministry of Urban Planning, and city administrative bodies
China	• Bolster infrastructure through the activities of the digital healthy city operation committee • Collaboration among public agencies (Ministry of Health), private sector (information technology companies), hospital personnel (physicians and traditional Chinese medicine practitioners), academia (college faculty), and civil society (health nongovernmental organization)
Japan	• Establish a digital healthy city committee with the Ministry of Health and local government leaders • Communication through local media and digital platforms
The Republic of Korea	• Establish various digital health company infrastructure around artificial intelligence conglomerates • Expand the network around the Korea Healthy Cities Partnership
Malaysia	• Establish a department dedicated to technological advancements for digital healthy cities • Establish a digital healthy city network by hosting various digital healthy city conferences
Mongolia	• Launch Mongolia digital healthy city partnerships • Consider the expertise of the digital healthy city policy strategy team, operation, and research teams
The Philippines	• Develop infrastructure involving information and communication technology developers and development coordinators • Establish a network involving environmental experts to develop digital healthy cities
Vietnam	• Establish and execute sustainable digital healthy city plans by developing professional project teams for a digital healthy city

[a]SPIRIT: setting approach and sustainability; political commitment, policy, and community participation; information and innovation; resources and research; infrastructure and intersectoral; and training.

Training (T)

We analyzed training initiatives for realizing digital healthy cities in each country. Australia regularly engages in education and research through exchanges with academic institutions. Cambodia requires settings, such as a digital healthy city initiative conference and the Global Health Summit (Table 7).

XSL•FO
RenderX

Table . Training (T) of the SPIRIT[a] checklist by country.

Country	Content
Australia	• Educate digital healthy city-related personnel through periodic workshops • Education and research through exchanges with academic institutions
Cambodia	• Cambodia digital healthy city initiative conference • Need digital healthy city Asia conference • Global health summit setting
China	• Healthy China 2030 conference • Host a digital healthy city forum • Smart and healthy city exhibition
Japan	• Host conference with the involvement of Japan Smart Community Alliance • Education using digital kiosks • Education using community fitness apps
The Republic of Korea	• Provide public health promotion workshops focused on digital health • Expand digital healthy city through the Korea digital health association • Organize digital health master class
Malaysia	• Host Digital Health Malaysia to invite healthy cities in nearby countries for a joint education on digital healthy city planning and directions • Host smart city conference
Mongolia	• Education and training on digital innovation in health care • Organize Mongolian Smart City Association
The Philippines	• Host an international digital healthy city forum through the Digital Health Philippines Summit • Need collaboration through digital healthy city sister city agreements
Vietnam	• Actively promote digital healthy cities through social media • Research on health technologies and collaboration

[a]SPIRIT: setting approach and sustainability, political commitment, policy, and community participation, information and innovation, resources and research, infrastructure and intersectoral, and training.

Discussion

Principal Results

This study analyzed the digital capabilities of AFHC member countries amid the ongoing transition toward a digital era using the AFHC SPIRIT checklist with the *openai* library in Python and the ChatGPT (version 4.0) large language model. The following implications were obtained.

First, it is necessary to prepare for digital healthy cities by forming digital healthy city networks and committees to promote sustainability and infrastructure. Specifically, an interdepartmental committee should be established to facilitate collaboration among health city-related departments. Further, a holistic "Healthy in All Policy during digital era" approach should be adopted. The AFHC should enhance intercity networking in the digital era during its biennial general meetings. Additionally, it is important to create a framework based on an in-depth understanding of social dynamics, policy structures, governance networks, and urban infrastructure [20].

Second, as digital healthy cities increasingly rely on data-driven policies, the importance of data protection and personal information security has been highlighted. With the volume of daily generated data expanding worldwide, the era of big data is imminent. In such an era, personal information protection is a significant social issue [21]. This study found that data-based policy decisions are crucial for digital healthy cities and requires reviewing and analyzing various data types. The increasing use of community residents' data necessitates diverse protective measures, such as legal regulations and training for responsible personnel. For instance, some healthy cities have adopted AI care robots to provide home-based care services for older adults. As all information about the client is collected through these care robots, heightened precautions must be taken to protect personal information [22].

Third, there is a need for cooperation between various organizations and sectors, such as public agencies, industries, schools, hospitals, and private organizations, to build a digital healthy city network and implement intersectoral healthy city policies. Such cooperation is crucial to promote digital healthy

XSL•FO
RenderX

cities. It is particularly important to establish an infrastructure for community health management by linking digital information technology and AI-related industries with school research institutions and including public institutions, hospitals, and private organizations. Developing a shared platform across these sectors is essential to systematically implement digital healthy cities initiatives [7].

Fourth, there is a need for digital healthy city policies tailored to each country's unique digital capabilities, characteristics, policies, and cultures. The WHO reviewed the focus group analyses of healthy cities worldwide and emphasized the importance of developing appropriate policies for each country and city, defining this as the "glocal approach" [23]. With advancements in digital technology, there is a growing focus on personalized medicine and public health. Accordingly, digital healthy city policies must be customized, considering the diverse characteristics of each country.

Finally, this study highlights the importance of education for key groups in digital healthy cities—policy makers, civil servants, health care providers, and residents. As digital transformation gains traction worldwide, AFHC member countries are also witnessing accelerated digitalization. However, the digital divide between cities, genders, and socioeconomic classes in these countries is widening [24]. The WHO cautions that, without adequate digital literacy among managers and users, the efficacy of digital technologies remains limited even after commercialization [10]. Telemedicine, digital health kiosks, and data management systems are key components of digital healthy cities. To use these services as intended, competency education is essential. Most importantly, digital literacy education for older adults is crucial, given the rapid pace of population aging [25,26].

Limitations

This study has a few limitations. First, as ChatGPT (version 4.0) is trained only until April 2023, the latest developments were not reflected in our survey. Second, the number of healthy cities varied across countries. Third, the SPIRIT checklist was interpreted by only 3 experts. While objective assessments were performed based on their expertise, the number of experts may not have been sufficient for subjective results. Thus, future studies should involve at least 5 experts. Fourth, owing to the proprietary nature of OpenAI's ChatGPT large language model, its full workings and application in this study could not be completely explained. Additionally, the differing levels of political commitment and resources among countries complicate comparisons. Wealthier nations such as South Korea and Japan invest heavily in digital health, while countries such as Cambodia and Mongolia focus on basic health care needs. Tailored strategies are needed to account for these economic and health care disparities, and future studies should consider long-term progress in countries with less developed infrastructures.

Conclusions

This study analyzed the digital health city capabilities of 9 AFHC member countries in the digital era. The findings suggested that these countries must prepare to attain sustainable digital healthy cities by establishing networks and committees. Additionally, digital healthy cities develop big data-driven policies, and this requires systems to ensure personal information protection. This study also emphasized the need for collaborative initiatives across various sectors to systematically implement digital healthy cities. Finally, digital healthy cities are expected to be less effective even with good policies and environments if users have poor digital literacy; thus, policies for improving digital literacy among both managers and users will be required.

Acknowledgments

We would like to thank the stakeholders of Alliance for Healthy Cities countries for their contributions to digital healthy cities. This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2021R1C1C2005464).

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

References

1. World Health Organization. Revised list of indicators and procedure for monitoring progress towards health for all in the European region (1987–1988). : WHO European Bureau; 1989 URL: https://iris.who.int/bitstream/handle/10665/355931/39wd04e-HFATgtsUpdat-890007.pdf?sequence=1&isAllowed=y [accessed 2025-01-28]
2. Ashton J, Grey P, Barnard K. Healthy cities — WHO's new Public Health initiative. Health Promot Int 1986;1(3):319-324. [doi: 10.1093/heapro/1.3.319]
3. WHO European healthy cities network. World Health Organization. 2023. URL: https://www.who.int/europe/groups/who-european-healthy-cities-network [accessed 2024-05-03]

XSL•FO
RenderX

4. Nakamura K. A network of healthy cities in asia and the pacific: the alliance for healthy cities. In: Muto T, Nakahara T, Nam EW, editors. Asian Perspectives and Evidence on Health Promotion and Education: Springer Japan; 2011:155-161. [doi: 10.1007/978-4-431-53889-9_15]
5. Raea SM, Alsulami ANM, Aljuaid TH, et al. Designing healthy cities: linking infrastructure to quality of life. Int J of Health Sci 2020;4(S1):509-529. [doi: 10.53730/ijhs.v4nS1.15432]
6. Duhl LJ. The healthy city: its function and its future. Health Promot Int 1986;1(1):55-60. [doi: 10.1093/heapro/1.1.55]
7. de Leeuw E. Evidence for healthy cities: reflections on practice, method and theory. Health Promot Int 2009 Nov;24 Suppl 1:i19-i36. [doi: 10.1093/heapro/dap052] [Medline: 19914985]
8. Acuto M, Morissette M, Tsouros A. City diplomacy: towards more strategic networking? Learning with WHO Healthy Cities. Global Policy 2017 Feb;8(1):14-22 [FREE Full text] [doi: 10.1111/1758-5899.12382]
9. Cocchia A. Smart and digital city: a systematic literature review. In: Dameri RP, Rosenthal-Sabroux C, editors. Smart City: Springer International Publishing; 2014:13-43. [doi: 10.1007/978-3-319-06160-3_2]
10. Rudnicka E, Napierała P, Podfigurna A, Męczekalski B, Smolarczyk R, Grymowicz M. The World Health Organization (WHO) approach to healthy ageing. Maturitas 2020 Sep;139:6-11. [doi: 10.1016/j.maturitas.2020.05.018] [Medline: 32747042]
11. Lapão LV, Correia JC, Jevtic M. Public health framework for smart cities within the comprehensive approach to sustainability in Europe: case study of diabetes. Sustainability 2023;15(5):4269. [doi: 10.3390/su15054269]
12. Is the Eastern Mediterranean region ready for digitalizing health? implications from the global strategy on digital health. : World Health Organization; 2021 URL: https://www.slideshare.net/WHO-EMRO/is-the-eastern-mediterranean-region-ready-for-digitalizing-health [accessed 2025-01-28]
13. Lin Y, Shen Z. Evaluation of urban health efficiency in japanese cities using network data envelopment analysis. J Urban Plann Dev 2023 Jun;149(2):04023011. [doi: 10.1061/JUPDDM.UPENG-4069]
14. Mouton M, Ducey A, Green J, et al. Towards “smart cities” as “healthy cities”: health equity in a digital age. Can J Public Health 2019 Jun;110(3):331-334. [doi: 10.17269/s41997-019-00177-5] [Medline: 30701413]
15. World. ITU. URL: https://datahub.itu.int/data/ [accessed 2024-05-04]
16. Eriksson C. Learning and knowledge-production for public health: a review of approaches to evidence-based public health. Scand J Public Health 2000 Dec;28(4):298-308. [Medline: 11228118]
17. Nam EW, Moon JY, Lee A. Evaluation of Healthy City Project using SPIRIT checklist: Wonju city case. Korean J Health Educ Promot 2010;27:15-25 [FREE Full text] [doi: G704-000885.2010.27.5.001]
18. 2021 SPIRIT template. Alliance for Healthy Cities. 2021. URL: https://www.alliance-healthycities.com/docs/AFHCAward2021/AFHC%20Awards%202021_SPIRIT_TEMPLATE.docx [accessed 2024-05-12]
19. Members. Alliance for Healthy Cities. URL: https://www.alliance-healthycities.com/htmls/members/index_members.html [accessed 2024-03-03]
20. Data bank. World Bank Group. URL: https://databank.worldbank.org/ [accessed 2025-01-14]
21. Ramaswami A, Russell AG, Culligan PJ, Sharma KR, Kumar E. Meta-principles for developing smart, sustainable, and healthy cities. Science 2016 May 20;352(6288):940-943. [doi: 10.1126/science.aaf7160] [Medline: 27199418]
22. Zou H. Protection of personal information security in the age of big data. In: 12th International Conference on Computational Intelligence and Security (CIS): IEEE Publications; 2016:586-589. [doi: 10.1109/CIS.2016.0142]
23. Lee H, Chung MA, Kim H, Nam EW. The effect of cognitive function health care using artificial intelligence robots for older adults: systematic review and meta-analysis. JMIR Aging 2022 Jun 28;5(2):e38896. [doi: 10.2196/38896] [Medline: 35672268]
24. de Leeuw E. Global and local (glocal) health: the WHO Healthy Cities programme. Global Change and Human Health 2001;2(1):34-45. [doi: 10.1023/A:1011991014805]
25. Longoria IAI, Bustamante-Bello R, Ramírez-Montoya MS, Molina A. Systematic mapping of digital gap and gender, age, ethnicity, or disability. Sustainability 2022;14(3):1297. [doi: 10.3390/su14031297]
26. Lee H, Lim JA, Nam HK. Effect of a digital literacy program on older adults’ digital social behavior: a quasi-experimental study. Int J Environ Res Public Health 2022 Sep 29;19(19):12404. [doi: 10.3390/ijerph191912404] [Medline: 36231707]

Abbreviations

AFHC: Alliance for Healthy Cities
AI: artificial intelligence
IRB: institutional review board
SPIRIT: setting approach and sustainability; political commitment, policy, and community participation; information and innovation; resources and research; infrastructure and intersectoral; and training
WHO: World Health Organization

XSL•FO
RenderX

Edited by A Mavragani; submitted 05.06.24; this is a non–peer-reviewed article;accepted 05.12.24; published 07.02.25.

Please cite as:
Lee
Assessment of Digital Capabilities by 9 Countries in the Alliance for Healthy Cities Using AI: Cross-Sectional Analysis
JMIR Form Res 2025;9:e62935
URL: https://formative.jmir.org/2025/1/e62935
doi:10.2196/62935

© Hocheol Lee Originally published in JMIR Formative Research (https://formative.jmir.org), 7.2.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

A Web-Based Resource Informed by Cognitive Behavioral Therapy and Positive Psychology to Address Stress, Negative Affect, and Problematic Alcohol Use: A Usability and Descriptive Study

Ingrid Serck-Hanssen[1], MS, MBA; Marit Solheim-Witt[2], MA; Justin J Anker[3], PhD; Dawn E Sugarman[4,5], PhD

[1]Department of Clinical Counseling, Alpha Element Institute, LLC, 6160 Summit Dr N, Suite 200, Minneapolis, MN, United States

[2]Department of Holistic Health, Alpha Element Institute, LLC, Minneapolis, MN, United States

[3]Psychiatry and Behavioral Sciences, University of Minnesota, Minneapolis, MN, United States

[4]Division of Alcohol, Drugs, and Addiction, McLean Hospital, Belmont, MA, United States

[5]Department of Psychiatry, Harvard Medical School, Boston, MA, United States

Corresponding Author:
Ingrid Serck-Hanssen, MS, MBA
Department of Clinical Counseling, Alpha Element Institute, LLC, 6160 Summit Dr N, Suite 200, Minneapolis, MN, United States

Abstract

Background: Research documents that drinking to cope behavior can be disrupted by enhancing emotion regulation and coping skills related to the experience of stress and negative affect. The Alpha Element Self-Coaching Plan incorporates principles of positive psychology and cognitive behavioral therapy to redirect negative thinking and emotions and, therefore, has the potential to benefit individuals who use alcohol to cope with stress.

Objective: This study aimed to evaluate satisfaction and usability of the web-based Alpha Element Self-Coaching Plan in order to inform the development of an expanded digital platform based on the Alpha Element framework.

Methods: Participants enrolled in the web-based program as part of their clinical care were eligible to participate. A total of 20 individuals (14 women and 6 men) between ages 30 and 79 (mean 54.5, SD 14.14) years completed web-based questionnaires to assess product performance in areas such as ease of technology use, quality of videos and handouts, and the value of the activities. Participants also completed the System Usability Scale (SUS) and provided background and demographic information, including alcohol use.

Results: Only 1 participant reported no alcohol use in the past year; 55% (11/20) of participants drank alcohol 2‑4 times per month or less and 45% (9/20) reported drinking alcohol 2‑3 times per week or more. The average SUS score of 76.38 (SD 17.85) was well above the commonly accepted threshold of 68, indicating high system usability. A majority of the sample (16/19, 84%) agreed or strongly agreed that the activities in the program inspired behavioral changes; and most agreed or strongly agreed that the program was engaging (16/20, 80%), well-organized (18/20, 90%), and easy to follow (17/20, 85%). Only 2 participants endorsed experiencing difficulty using the program on a smartphone. Suggestions for program improvements included expanding the platform, updating the web format, adding user interactivity, and enhancing navigation.

Conclusions: These data suggest that participants were generally satisfied with the web-based Alpha Element Self-Coaching Plan, and rated usability of the program as favorable. Importantly, a significant portion of participants reported that the program inspired behavioral changes. More research is needed with a larger sample to obtain specific data about alcohol consumption and investigate associations between alcohol use and program components, as well as examine gender differences. Data collected from this study will be used to expand the platform and improve user experience.

(JMIR Form Res 2025;9:e63819) doi:10.2196/63819

KEYWORDS

alcohol misuse; stress; drinking to cope; DTC; negative affect; positive psychology intervention; PPI; cognitive behavior therapy; CBT; alcohol use; drinking; usability; descriptive study; behavior; emotion; coping skill; positive psychology; psychology; online resource; mobile phone

XSL•FO
RenderX

Introduction

Stress is a well-documented risk factor associated with the initiation, maintenance and relapse of alcohol use disorder [1]. The self-medication model suggests some individuals drink to cope (DTC) with stress due to the belief that drinking will alleviate stress and negative affect [2,3]. However, although alcohol may provide temporary relief from emotional discomfort, in the long-term, drinking to cope with stress exacerbates negative affect and puts individuals at risk for alcohol-related consequences [4].

Positive psychology interventions (PPIs) may be beneficial in disrupting the DTC cycle of alcohol use, as emerging data suggest that the processes underlying PPI may facilitate more adaptive behavioral choices about substance use [5]. PPIs are strength-based approaches that apply individual traits along with positive subjective experiences to promote well-being [5,6]. PPIs aim to foster optimism, psychological well-being, and resilience through practices like cognitive flexibility, positive reframing, and the cultivation of strengths, representing a shift away from focusing solely on reducing symptoms and instead building emotional assets. Furthermore, research indicates that positivity can be learned and lead to increased well-being and decreased alcohol consumption [7]. A wide range of PPIs exist that vary in type, delivery mode, and intensity. PPIs have mostly been studied in nonclinical samples; however, preliminary evidence exists for their efficacy in clinical settings [8]. Research suggests that cognitive changes may explain the positive effect of PPIs on emotions [6]. Therefore, cognitive behavioral therapy (CBT) can be personalized with PPI components that target motivational and emotional factors. With a focus on increasing positive emotions, behaviors, and cognitions, PPI approaches align well with CBT principles and both can be leveraged with the objective to reduce stress with negative affect aimed at mitigating cravings for alcohol to self-medicate.

Critical to the success of CBT-based strategies in the clinical setting is engaging clients in homework exercises [9,10]. Technology-based tools such as smartphone apps and web-based self-management programs provide scalable solutions for individuals to engage with self-directed exercises. Emerging research indicates that web-based interventions can serve a protective function for mental health by mitigating the effects of stress [11]. Ongoing measurement of engagement with web-based homework exercises along with strategic approaches to tailoring intervention content delivery to individuals may help increase sustained practice [12]. The purpose of this quality improvement (QI) study was to evaluate usability and satisfaction of the web-based Alpha Element Self-Coaching Plan (hereafter referred to as the Self-Coaching Plan [13]) applied in a community clinical counseling setting. The Self-Coaching Plan incorporates a combination of CBT and PPI approaches and features content based on a novel positive psychological framework of Vitality Type patterns in behavior and cognition [14-17]. The goal of the Self-Coaching Plan is to empower individuals with tools to discover a new and positive perspective on themselves, along with psychoeducation and personalized skills to proactively manage stress and negative affect, as well as address problematic drinking behavior. The aim of this QI study was to examine satisfaction and usability of the web-based Self-Coaching Plan in order to facilitate ongoing improvement of this tool.

Methods

Study Setting

In this QI study the Self-Coaching Plan was offered as adjunct homework to clients in the clinical setting. Participation in the project was optional and did not affect the therapeutic relationship with their clinical provider. All participants in this study were receiving mental health counseling through telehealth in a community counseling clinic in Minneapolis, Minnesota.

Study Design

Data were gathered for administrative purposes within the context of routine efforts to examine and improve clinical care services and operations. This QI study was conceptualized as a formative evaluation to assess product performance of the Self-Coaching Plan. The evaluation surveys focused on system usability, technology ease of use, quality of videos, and downloadable handouts, as well as the value of the program content and activities.

Participants

Participants were recruited through a self-selection process. We applied an all-comers approach, meaning individuals enrolled in the Self-Coaching Plan as part of their standard care were invited to participate. Of the 35 clients invited to participate, 20 (14 women and 6 men) agreed. To recruit participants, a University of Minnesota REDCap (Research Electronic Data Capture; Vanderbilt University) survey link was included in weekly touch-base emails.

Intervention

A web-based tool for self-directed learning, the Self-Coaching Plan, was applied in this study as a homework resource in the clinical setting. The Self-Coaching Plan is based on the premise that introducing individuals to a framework of 4 Vitality Type patterns as a strategy to disrupt negative self-schemas can help provide new perspectives on existing behavior patterns as well as cognitive resources to reframe distorted thinking that is impairing the regulation of negative affect (Table 1).

XSL•FO
RenderX

Table . The premise of the Self-Coaching Plan is to apply 4 Vitality Type categories to self-identify empowered and disempowered patterns in behavior and cognition. Throughout the Self-Coaching Plan, individuals are reminded that words are cognitive tools that frame our perceptions and shape our reality.

Vitality Type patterns	Catalyst	Torchbearer	Pathfinder	Pragmatist
Empowered behavior patterns	• Discovery • Open-minded • Uplifting • Playful • Social • "Let's try it now."	• Action • Direct • Fast • Results-driven • Bold • "What can I do?"	• Development • Investigative • Process-oriented • Diplomatic • Consensus-building • "I'll think about it."	• Completion • Scrutinizing • Focused • Discerning • Structured • "It has to be done right."
Disempowered behavior patterns	• Self-indulgence • Erratic	• Irritability • Restless	• Avoidance • Overwhelmed	• Perfectionism • Judgmental

The 4 categories, Catalyst Vitality, Torchbearer Vitality, Pathfinder Vitality, and Pragmatist Vitality, represent distinct patterns in outlook and behavior (Table 1). The essence of a catalyst is to spark activity. Positive words associated with the Catalyst Vitality pattern include discovery, open-minded, uplifting, playful, and social. The essence of a torch bearer is to challenge status quo and pave the way to take action. Positive words associated with the Torchbearer Vitality pattern are action, direct, fast, results-driven, and bold. The essence of a pathfinder is to be methodical. Positive words associated with the Pathfinder Vitality pattern include development, investigative, process-oriented, diplomatic, and consensus-building. The essence of a pragmatist is to bring clarity and stability to their environment. Positive words associated with the Pragmatist Vitality pattern are completion, scrutinizing, focused, discerning, and structured. As part of the Self-Coaching Plan, participants study all 4 Vitality Type categories by exploring the words linked to each. Through a series of exercises, they apply these words to personalize their self-coaching strategies. By organizing positive words into categories (Table 1) the Vitality Type patterns combine elements from both CBT and PPI.

Exploring how to break unhealthy habits and cognitive distortions and replacing them with positive self-talk can create shifts in outlook that can lead to changes in behavior. The Vitality Type categories represent a range of possibilities in patterns of cognition and behavior which in the Self-Coaching Plan are applied as cognitive tools to help individuals self-identify and articulate aspects of their empowered and disempowered patterns. Exploring words linked to the Vitality Type patterns and applying these words as cognitive tools to change behavior and outlook, help individuals take back their power instead of leaning on alcohol and other negative coping strategies. It is posited that the framework of 4 Vitality Type categories offers resources to promote positive self-perception to inspire new self-awareness, self-discovery, and development of personalized self-management strategies. Vitality Type patterns can also be applied as a cognitive strategy for considering differences in relationships and interpersonal communication. Managing relationships with self and others are key aspects of managing stress and problematic drinking behavior.

It is proposed that learning about these patterns (schemas) can help individuals orient and learn new terminology to reframe negative schemas they have about themselves. By applying the constructs of 4 Vitality Types as cognitive resources to identify and articulate empowered and disempowered patterns, individuals are introduced to a fresh perspective on themselves along with terminology that can provide a bridge to describe potential movement away from pathology toward hope and well-being. This self-reflective process can lead to a deeper understanding of how behaviors and thoughts impact outlook and overall health. By using the Vitality Type categories as a framework of schemas and tools to foster cognitive flexibility, the Self-Coaching Plan adapts CBT elements and PPI principles as a therapeutic strategy to create synergy and inspire self-investigation. Therefore, the terminology of the 4 Vitality Type categories are applied to all activities to provide a map for organizing and personalizing CBT-PPI strategies.

In this QI study, the objective of the Self-Coaching Plan was for clients, between sessions, to explore and process patterns of disempowerment and negative outlook and learn skills to redirect and self-regulate. This web-based program offers psychoeducation and activities through videos and visual graphics; topics include how to apply the Vitality Type patterns to foster self-compassion, growth mindset, and preparing for positive change as well as facts about gender-specific vulnerabilities and substance use. During the Self-Coaching Plan, individuals are encouraged to use the Vitality Type patterns as tools to help achieve their goals and focus on what they are for—as a shift away from focusing on what they are against and working to change. Self-Coaching Plan activities are organized into sections with the objective of completing an action plan to stay motivated on their journey toward positive change (Table 2). It is recommended that the proposed order of activities is followed, however, individuals are free to choose the order they prefer. The goal is for individuals to apply the Vitality Type framework to gain a new perspective on predominant go-to behavior patterns and understand the importance of developing their less dominant patterns to self-advocate, set boundaries, and feel more empowered. The Self-Coaching Plan is designed to help individuals get prepared and stay motivated to achieve their goals over an 8-week period. They learn to apply words as cognitive tools to change their

XSL•FO
RenderX

behavior and take their power back instead of leaning on alcohol and other negative coping strategies. The Self-Coaching Plan can be applied as both a guided and unguided protocol.

XSL•FO
RenderX

Table . A web-based tool for self-directed learning, the Self-Coaching Plan, was applied in this study as a homework resource in the clinical setting. The goal was to empower individuals through a personalized process of activities and self-reflection to foster healthy habits, learn to set boundaries, manage emotions, and change unhealthy habits such as problematic drinking behavior.

Module name	Description	Components
Growth Mindset	Psychoeducation about how our abilities and frame of mind can be developed through dedication and hard work. A growth mindset is knowing that it is never too late to make a change.	• Welcome video. • Course intro. • Psychoeducation about growth mindset and self-compassion.
Get Ready for Positive Change	Psychoeducation to get clarity about priorities and steps to achieve positive lifestyle changes.	• Course objectives. • Gender specific vulnerabilities for substance use. • Substance use and drug facts.
Vitality Type Indicator	The objective of this self-assessment is to gain awareness of empowered and disempowered patterns in communication and relationships.	• Self-assessment objective. • Connection between stress management, outlook, and drinking behavior.
Vitality Type Patterns	Psychoeducation about how managing the relationship with self and others is key to managing the relationship with alcohol.	• Introducing the 4 Vitality Types as tools to self-manage and lay the foundation for developing a growth mindset. • Vitality Type matrix of empowered and disempowered patterns.
Top Priority	Identify the most important area of desired change. Individuals are encouraged to look at themselves as a priority.	• Exploring boundaries required for positive change. • Self-care as a means to manage stress. • Connection between mind, body and soul in healing.
Self-Coaching Guide	The Self-Coaching Guide is the Vitality Type of the highest score on the Vitality Type Indicator self-assessment.	• The approach of the Self-Coaching Guide is key to achieving our Top Priority. • Self-Coaching exercise applies the highest scoring Vitality Type pattern as a technique and practice to reflect on goals, challenges, and personal growth. • Exercise to create positive “I am” statements using words from the word-bank associated with their highest self-assessed Vitality Type score.
Hidden Resource	The Hidden Resource is the Vitality Type of the lowest score on the Vitality Type Indicator self-assessment.	• The approach of the Hidden Resource indicates untapped potential that can help on the path to sobriety. • Exercise to redirect thoughts by applying the lowest scoring Vitality Type as a strategy to address challenges and difficult situations in a new way. • Reflection exercise to discover how to leverage aspects of the Hidden Resource approach as support during the healing journey.
Shadow Aspects	Shadow Aspects represent disempowered behaviors taken on during times of stress or fear.	• Exercise to self-identify behaviors that are barriers in communication and relationships. • Reflection activity about self-compassion and growth mindset as strategies to help turn around a disempowered state.

XSL•FO
RenderX

Module name	Description	Components
Power Words	Power Words are positive characteristics and behaviors linked to each Vitality Type.	• Activities to choose Power Words, which are then applied to personalize strategies to help disrupt Shadow Aspects. • Exercise to identify 1 Power Word linked to the Self-Coaching Guide and another word linked to the Hidden Resource as strategy to stay proactive rather than reactive in stressful situations. • Reflection activity to become more aware of how to deal with joy and conflict can be key to understanding the relationship with alcohol.
Self-Coaching Plan Commitment Sheet	Course summary: maintaining sobriety and a sober-minded lifestyle is much more manageable when the focus is on empowerment, growth mindset and self-compassion.	• Resources to explore how to continue using the Vitality Types to stay committed to healing, health and wellness.

As part of the Self-Coaching Plan, individuals take a self-assessment (Vitality Type Indicator, Table 2) that sorts responses according to the categories of the 4 Vitality Type categories. The results of the self-assessment provide a map for individuals to discover a range of Vitality Type patterns in their behavior, cognition, and communication. Individuals' highest Vitality Type score indicates a predominant pattern and is identified as their Self-Coaching Guide. An individual's lowest Vitality Type score indicates untapped potential and is identified as their Hidden Resource. Attributes linked to an individual's Hidden Resource represent capabilities outside of the person's typical patterns of behavior. The concepts of Self-Coaching Guide and Hidden Resource are incorporated into the program as cognitive tools and a positive approach to develop personalized strategies for stress-management and self-regulation. Individuals are prompted to practice their Hidden Resource capabilities in order to prepare themselves to proactively handle challenges and difficult situations with a new perspective. This information is applied in the Self-Coaching Plan as CBT-PPI informed reflective learning activities and homework assignments.

Through a series of activities of using words as cognitive resources to make positive shifts, individuals practice using specific words that represent empowered behavior to formulate their desired change and develop healthy habits. A word bank is linked to each of the 4 Vitality Type categories as a tool to discover a range of possibilities. Individuals are asked to reflect on which words and aspects of the 4 patterns resonate with their view of themselves, and which do not. In a self-coaching module, individuals explore how to use their Self-Coaching Guide and their Hidden Resource (Table 2) to manage stress and move toward sobriety and a healthier, more balanced life. Table 1 presents an overview of the 4 categories and illustrates how each proactive (empowered) Vitality Type pattern has a reactive (disempowered) opposite. In the Self-Coaching Plan (Table 2), the disempowered opposite is referred to as a Shadow Aspect. Shadow Aspects represent disempowered behaviors taken on during times of stress or fear. To dig deeper into understanding their disempowered reactions, participants are asked to self-identify 2 primary Shadow Aspects. To redirect disempowerment, individuals then choose 2 words from the word bank they consider to be empowering. These words are referred to as Power Words. Power Words are positive characteristics and behaviors linked to each Vitality Type which are applied to develop their personalized strategies to redirect. During this CBT-PPI process of applying Vitality Type categories in self-discovery and reflective learning, individuals gain a newfound perspective on their go-to behavior patterns along with personalized strategies to use their less dominant patterns as a resource to set boundaries, self-advocate, and feel more empowered.

Procedures

A survey was offered through a REDCap link as part of touch-base emails frequently sent out to all registered users of the Self-Coaching Plan. We informed all users that the goal of the project was quality improvement and their answers would be used to improve user experience. Users were not able to proceed to the survey unless they had checked the box indicating they agreed to participate. Participants completed a custom questionnaire, the System Usability Scale (SUS), as well as background and demographic information, including alcohol use. The custom questionnaire was designed to evaluate (1) the relevance and ease-of use of the Self-Coaching Plan material and (2) satisfaction with the Self-Coaching Plan.

Measures

The SUS is a 10-item scale designed to provide a global view of subjective assessments of usability. The scale covers a variety of aspects of usability, including the need for support and training while using the system as well as its level of complexity. The SUS evaluation instrument yields a single number that represents a composite measure of the overall usability of the system being studied with current literature suggesting that a score of 68 as a useful benchmark (mean SUS score) [18].

The custom questionnaire was developed to gather feedback about program engagement along with dimensions of perceived interactivity and ease of use. The questions were developed by creators of the Self-Coaching Plan and other subject matter experts. The data were collected to establish a baseline along key dimensions against which change over time can be measured. Key dimensions measured included strengths and

XSL•FO
RenderX

limitations of the technology, videos, and exercises; need for tech support, notifications, and reminders; program navigation and logistics features; time commitment; aesthetics and visual elements; as well as suggestions for improvement. Specific questions were asked about engagement, self-perception and behavior change as well as knowledge about key terms and concepts, such as Vitality Types, Self-Coaching Guide, and Hidden Resource. Each item was rated on a 5-point Likert scale ranging from strongly disagree to strongly agree with a midpoint of neither agree nor disagree. Recommendations for enhancing program improvement were obtained through optional write-in responses to capture comments and suggestions in the language of participants. Alcohol consumption was measured with 1 question about frequency of alcohol use in the past year with the following scale: Never; Monthly or less; 2 - 4 times a month; 2 - 3 times a week; 4 or more times.

Data Collection

Participants completed the surveys through REDCap. REDCap is an encrypted, electronic database that is HIPAA (Health Insurance Portability and Accountability Act) compliant. It is approved by University of Minnesota Institutional Review Board for the administration and storage of human subject information.

Statistical Analysis

Analyses for this study focused on evaluating the Self-Coaching Plan using descriptive statistics and were performed using R statistical software (R Foundation for Statistical Computing). Mean and standard deviation were used for continuous variables such as age and SUS scores. For categorical variables, including demographic information (eg, sex, education level, and race) and survey responses (eg, engagement with self-care routine and frequency of alcohol consumption), frequencies were calculated.

Ethical Considerations

The project was determined not to qualify as human subject research by the University of Minnesota Institutional Review Board (STUDY00020268). Confidentiality was strictly protected. No identifying information was collected.

Results

Participant Characteristics

Participants (N=20) were between 30 and 79 years of age (mean 54.5, SD 14.14), predominantly female (n=14, 70%) White (n=15, 75%) and well educated with 35% (n=7) college graduates and 35% (n=7) reporting as graduate level or above. Only 1 participant reported no alcohol use in the past year; 55% (n=11) of the sample drank alcohol 2 - 4 times per month or less and 45% (n=9) reported drinking alcohol 2 - 3 times per week or more.

System Usability

Average SUS score from this study was 76.38 (SD 17.85). This result is well above the commonly accepted threshold of 68. The score indicates high usability of the system and its interfaces.

Custom Questionnaire

As shown in Table 3, a majority (16/19, 84%) of participants agreed or strongly agreed that the activities in the program inspired positive behavioral changes while 90% (18/20) agreed or strongly agreed that participation in the program positively influenced their self-perception. There was strong endorsement of knowledge and understanding about concepts unique to the program such as Vitality Types, Self-Coaching Guide and Hidden Resource. Of note, all participants (n=20, 100%) agreed or strongly agreed that the Vitality Type Indicator self-assessment was a helpful tool in their self-development; 90% (n=18) endorsed the purpose of the Self-Coaching Guide as clear and understandable while 85% (n=17) reported they could relate to their Self-Coaching Guide; 80% (n=16) of participants strongly endorsed understanding the concept of a Hidden Resource. A majority of participants (n=16, 80%) endorsed the program as engaging and 80% (n=16) found the videos were effective in motivating them to complete the homework activities and most participants agreed or strongly agreed that the visual design of the platform was appealing. With regards to technical interface, 80% (n=16) of participants did not experience significant difficulties using the Self-Coaching Plan on smartphones or computers. Generally, participants (n=18, 90%) found the program well-organized.

XSL•FO
RenderX

Table . Results from the quality improvement study (N=20).

Quality improvement survey questions	Strongly agreed or agreed, n (%)	Neither agreed or disagreed, n (%)	Disagreed or strongly disagreed, n (%)
The activities in this program inspired me to change my behavior.[a]	16 (84)	—[b]	3 (16)
Participating in this course positively impacted my self-perception.	18 (90)	2 (10)	—
The order of the activities in the program was logical and helped in my understanding.	18 (90)	1 (5)	1 (5)
The videos were effective in motivating me to complete the activities.	16 (80)	3 (15)	1 (5)
The Vitality Type categories were understandable.	18 (90)	2 (10)	—
The Vitality Type Indicator self-assessment was helpful in my self-development.	20 (100)	—	—
The purpose of the Self-Coaching Guide was clearly explained and understandable.	18 (90)	2 (10)	—
I related to my Self-Coaching Guide.	17 (85)	3 (15)	—
The concept of a Hidden Resource was clear to me.	16 (80)	4 (20)	—
I found the program to be engaging.	16 (80)	2 (10)	2 (10)
The content within the program was well organized.	18 (90)	2 (10)	—
The time required to complete this course felt appropriate to me.	17 (85)	2 (10)	1 (5)

[a]One participant missed this item.

[b]Not endorsed by any participants.

As part of the custom questionnaire, participants were asked to respond to 3 open-ended questions. The first question asked participants to share their thoughts about which program components they considered to be the most effective. Predominant themes that emerged from responses to this question were (1) empowerment, (2) personalization, and (3) engagement. The following comments illustrate a sense of empowerment: “Understanding better how I react to things” and “A clear view of my strengths and weaknesses really helped me choose where I wanted to focus my energy.” The theme personalization was illustrated by the following quotes: “I love that I get to use a personal Self-Coaching Guide based on my behaviors specifically designed for me!” and “I enjoyed exploring the Vitality Types that align with my personality and observing how they influence my daily behavior. After completing the test (Vitality Type self-assessment), I identified which trait was most prominent for me.” Third, the theme engagement, was illustrated by these comments: “I found the homework was very helpful for me to have. It encouraged me to continue my healing in between sessions and then to have a follow-up conversation with my therapist about the homework really helped tie everything together for me.” “The activities were most effective because they made you actually pause and reflect.”

The second open-ended question asked participants to share their views about program components they considered the least effective. Here the predominant theme was navigation, as illustrated by the following comments, “I did find it a little confusing in the beginning to find the correct files at the end of the program,” and “It was hard to find the Vitality Types breakdowns through the library.” The third open-ended question asked for recommendations to enhance the effectiveness of the program. Here feedback again focused on navigation, as the following comments illustrate. “I think the navigation could be cleaned up, but overall, it was an effective program.” “The online interface needs to be updated.” Other themes for program improvements included expanding the platform and adding user interactivity as illustrated by the following comments: “I’d be excited to witness user interaction on the platform. In addition, incorporating a gamification element, perhaps tracking our progress through the program, would be an engaging feature.” Another participant stated, “I hope for a platform where I can engage with others regarding Vitality Types, perhaps in the form of a community for individuals who have undergone this course.”

XSL•FO
RenderX

Discussion

Principal Findings

Findings from this study document high usability and satisfaction with the web-based Self-Coaching Plan in the clinical setting. A significant portion of participants reported that the program inspired behavioral changes as well as positive self-perception and found the program to be well-organized and easy to follow. Most participants (18/20, 90%) reported the Vitality Type categories were understandable, with 85% (17/20) stating they could relate to their Self-Coaching Guide, and 84% (16/19) endorsed that they were inspired by program activities to change their behavior. However, more research is needed to examine the relationship between these constructs, participant self-perception, and behavior change regarding alcohol use. Several participants expressed a desire for improved navigation.

Comparison With Previous Work

An increasing amount of research points to the double jeopardy of stress and negative affect as a detrimental set of vulnerabilities that lead to problematic drinking behavior and DTC [2,19]. This trend is of increasing concern and spotlights the need for research and development of innovative self-management approaches. Data from this QI study suggest that participants were generally satisfied with the Self-Coaching Plan, and rated usability of the program as favorable. The implications from this study support the application of the Vitality Type categories as a CBT-PPI approach for individuals to gain a fresh and more positive self-perception as well as personalized skills to self-regulate. All participants endorsed the Vitality Type Indicator as a helpful tool in their self-development. Through reflection activities, individuals were encouraged to discover a new view of themselves and apply the Vitality Type patterns as cognitive tools to disrupt barriers, reframe, and manage stress.

Findings from this QI study indicate the 4 Vitality Types patterns can be adapted to a web-based program for individuals to orient and reflect on these 4 categories of attributes and explore the possibility of using the terminology to describe aspects of themselves and develop strategies to redirect. Responses to the custom questionnaire document that 90% (18/20) of participants endorsed the purpose of the Self-Coaching Guide as clear and understandable. Furthermore, 85% (17/20) reported they could relate to their Self-Coaching Guide. The majority of participants 80% (16/20) strongly endorsed understanding the concept of a Hidden Resource. It is noteworthy that 90% (18/20) of participants thought the Self-Coaching Plan positively impacted their self-perception.

A focus of CBT approaches is to support clients in their efforts to examine their thoughts and behaviors and become scientific investigators of their own lives [9]. Results from the open-ended questions in the custom questionnaire document that participants applied the self-directed exercises in the Self-Coaching Plan as tools to self-reflect and personalize the content. This finding is underscored by the comment, “The activities were most effective because they made you actually pause and reflect.” Individuals who engage with homework activities are taking a hopeful step by demonstrating the belief that their own efforts will help them overcome their clinical problems [10]. It has been proposed that integrating PPIs into existing treatment approaches may be beneficial because they produce an immediate reward that is nonsubstance-related [20].

Historically, PPI research has focused on individual positive qualities and attributes such as gratitude, optimism, and kindness [21] and not development of novel conceptual frameworks of patterns in behavior and cognition to address the larger context of disempowerment created by negative self-schemas. The Vitality Type categories offer a map of words for individuals to reflect on a range of factors impacting their behavior and communication without focusing on pathology. The Self-Coaching Plan is distinguished from other web-based psychosocial programs by providing a novel conceptual framework of 4 Vitality Type categories as proposed schemas in behavior and cognition. The proposed schematic conceptualizations of distinct patterns of emotion, behavior and cognitions as 4 Vitality Type patterns represent a model that can be applied to expand upon the traditional positive psychology approaches. As a psychosocial approach to self-development, the Self-Coaching Plan represents a new context for self-reflection and an innovative way of looking at differences in relationships and interpersonal communication. The current lack of standardized PPI protocols may limit the implementation of PPIs in the clinical setting and it may be beneficial to develop new, positive-focused interventions for substance use and other conditions instead of simply including traditional PPIs into existing treatments [22]. One study found that addiction counselors were using positive psychology exercises in their sessions over half of the time without classifying their methods as explicitly PPI [23]. To our knowledge, current web-based PPI informed programs do not apply a map of 4 empowered and disempowered patterns in behavior and cognition as schemas to navigate stress and proactively manage negative affect. As documented by the findings in this study, 90% (18/20) of participants endorsed the Self-Coaching Plan as positively impacting their self-perception. These results indicate that awareness of the Vitality Type categories and terminology can be leveraged to help individuals gain a more optimistic view of themselves as well as develop a personalized self-coaching plan to work toward health and well-being.

Data collected from this study will be applied to improve and expand the web-based Self-Coaching Plan. More research is needed with a larger sample to obtain specific data about alcohol consumption and investigate associations between alcohol use and program components. Furthermore, web-based programs that incorporate gender-specific components have been shown to enhance outcomes for women with substance use disorders [24]. Therefore, it is important for future research to examine gender differences, as well as feedback on gender-specific elements of the program.

Limitations

The small number of participants is a limitation in this study. Participants were recruited through a self-selection process which could be a limitation if participants who chose to partake in the survey were more positively biased than nonparticipants.

XSL•FO
RenderX

Another limitation is that drinking to cope was not a criterion for eligibility, thus the intervention was not specifically tested on individuals who drink to cope. Individuals who drink to cope might have provided different data regarding acceptability and usefulness. Hence, future research should focus on testing the intervention in a target population with DTC. In addition, the sample was predominantly White, female, and college-educated. While this demographic distribution is representative of the community where the research was conducted, it limits the wider generalizability of the findings. Future research should seek to gain feedback on usability and satisfaction from a more diverse sample. Of note, recent research suggests alcohol use is increasing more among middle-aged and older adults than among younger drinkers [25], therefore the wide age range in the present sample could be an indication of the program's appeal to this age group.

Conclusions

In this QI study, we examined usability and satisfaction with a web-based Self-Coaching Plan that applies CBT and PPI principles featuring a novel framework of Vitality Type categories to engage clients in a structured self-development process to increase skills to manage stress and negative affect. Findings from this quality improvement study indicate that the Self-Coaching Plan was engaging as well as acceptable to participants, positively influencing their self-perception and behavior. Data from the custom questionnaire document that applying the framework of 4 Vitality Type patterns is an effective approach to empower individuals with information, strategies, and a plan to self-regulate. Taken together, results from this QI study warrant investigating if the Self-Coaching Plan's approach to personalize assets-based self-management strategies targeting stress and negative affect, could help empower individuals with healthy habits to cope with stress instead of turning to alcohol.

Acknowledgments

The study was internally funded by Alpha Element Institute, LLC. Effort for this paper was supported by the National Institute on Drug Abuse K23DA050780 (DES) and UG3DA059414 (JJA).

Data Availability

The datasets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

ISH and MSW cowrote the paper, assisted in study conception, development, and qualitative data analysis. JJA led data analysis and assisted in study conception and development. DES conceived of the study, oversaw the project, and edited the paper. All authors reviewed and approved the final paper.

Conflicts of Interest

ISH and MSW are cofounders and co-owners of Alpha Element Institute, LLC, and have equity in the company. DES and JJA report no conflicts of interest.

References

1. Peltier MR, Verplaetse TL, Mineur YS, et al. Sex differences in stress-related alcohol use. Neurobiol Stress 2019 Feb;10:100149. [doi: 10.1016/j.ynstr.2019.100149] [Medline: 30949562]
2. Frone MR. Work stress and alcohol use: developing and testing a biphasic self-medication model. Work Stress 2016;30(4):374-394. [doi: 10.1080/02678373.2016.1252971] [Medline: 28090129]
3. Hawn SE, Cusack SE, Amstadter AB. A systematic review of the self-medication hypothesis in the context of posttraumatic stress disorder and comorbid problematic alcohol use. J Trauma Stress 2020 Oct;33(5):699-708. [doi: 10.1002/jts.22521] [Medline: 32516487]
4. Armeli S, Sullivan TP, Tennen H. Drinking to cope motivation as a prospective predictor of negative affect. J Stud Alcohol Drugs 2015 Jul;76(4):578-584. [doi: 10.15288/jsad.2015.76.578] [Medline: 26098033]
5. Trinh CD, Girard R, Schick MR, Spillane NS. Positive psychological interventions on alcohol use and consequences: pilot randomized trial in a young adult cannabis-using sample. J Subst Use Addict Treat 2024 Mar;158:209241. [doi: 10.1016/j.josat.2023.209241] [Medline: 38056630]
6. Stone BM, Schmidt K. Positive psychological interventions and cognition. J Posit Psychol 2020 Sep 2;15(5):629-632. [doi: 10.1080/17439760.2020.1789714]
7. Akhtar M, Boniwell I. Applying positive psychology to alcohol-misusing adolescents. GPWK 2010;20(3):6-31. [doi: 10.1921/gpwk.v20i3.707]
8. Chakhssi F, Kraiss JT, Sommers-Spijkerman M, Bohlmeijer ET. The effect of positive psychology interventions on well-being and distress in clinical samples with psychiatric or somatic disorders: a systematic review and meta-analysis. BMC Psychiatry 2018 Jun 27;18(1):211. [doi: 10.1186/s12888-018-1739-2] [Medline: 29945603]

XSL•FO
RenderX

9. Padesky CA. Collaboration and guided discovery. Cogn Behav Pract 2022 Aug;29(3):545-548. [doi: 10.1016/j.cbpra.2022.02.003]
10. Strunk DR. Homework. Cogn Behav Pract 2022 Aug;29(3):560-563. [doi: 10.1016/j.cbpra.2022.02.009]
11. Enrique Roig A, Mooney O, Salamanca-Sanabria A, Lee CT, Farrell S, Richards D. Assessing the efficacy and acceptability of a web-based intervention for resilience among college students: pilot randomized controlled trial. JMIR Form Res 2020 Nov 11;4(11):e20167. [doi: 10.2196/20167] [Medline: 33174530]
12. Goldin PR, Lindholm R, Ranta K, Hilgert O, Helteenvuori T, Raevuori A. Feasibility of a therapist-supported, mobile phone-delivered online intervention for depression: longitudinal observational study. JMIR Form Res 2019 Jan 22;3(1):e11509. [doi: 10.2196/11509] [Medline: 30682726]
13. Transform your life with the Alpha Element Self-Coaching Plan. Alpha Element. 2023. URL: https://alphaelement.podia.com/alpha-element-academy [accessed 2025-01-28]
14. Anker JJ, Serck-Hanssen I, Solheim-Witt M, Sugarman DE. Quality improvement of an online cognitive behavioral therapy and positive psychology informed resource to manage alcohol use, stress, and negative affect. In: Alcohol, Clinical and Experimental Research: Volume 48, Issue S1 Supplement: Speaker and Poster Abstracts of the Research Society on Alcohol, 47th Annual Scientific Meeting, June, 2024: Wiley-Blackwell; 2024. URL: https://onlinelibrary.wiley.com/toc/29937175/2024/48/S1 [accessed 2025-01-28]
15. Solheim-Witt M, Serck-Hanssen I. Self-development and interpersonal communication. Presented at: Interprofessional Education Summit: High-Performance Innovation Teams: An Interprofessional Approach to Health; Apr 11, 2017; St Paul, MN, United States.
16. Solheim-Witt M, Serck-Hanssen I. Communication as a bridge to proactive behavior, resilience and well-being. Presented at: Interprofessional Education Summit: What if? The Impact of Innovation and Interprofessionalism on Healthcare; Apr 9, 2018; St Paul, MN, United States.
17. Solheim-Witt M, Serck-Hanssen I. Resilience, resourcefulness and self-leadership: evaluation of a self-management program and novel conceptual framework of vitality type patterns in behavior and communication. Presented at: National Institute on Drug Abuse Genetics and Epigenetics Cross Cutting Research Team Meeting; May 16-17, 2023; Bethesda, MD, United States.
18. Hyzy M, Bond R, Mulvenna M, et al. System usability scale benchmarking for digital health apps: meta-analysis. JMIR mHealth uHealth 2022 Aug 18;10(8):e37290. [doi: 10.2196/37290] [Medline: 35980732]
19. Anker JJ, Kushner MG, Thuras P, Menk J, Unruh AS. Drinking to cope with negative emotions moderates alcohol use disorder treatment response in patients with co-occurring anxiety disorder. Drug Alcohol Depend 2016 Feb 1;159:93-100. [doi: 10.1016/j.drugalcdep.2015.11.031] [Medline: 26718394]
20. Stone BM. Positive psychology for substance use disorders: a rationale and call to action. J Stud Alcohol Drugs 2022 Nov;83(6):959-961. [doi: 10.15288/jsad.22-00259] [Medline: 36484594]
21. Mahipalan M, S S. Workplace spirituality and subjective happiness among high school teachers: gratitude as a moderator. Explore (NY) 2019;15(2):107-114. [doi: 10.1016/j.explore.2018.07.002] [Medline: 30150147]
22. Stone BM. Integrating positive psychology into substance use treatments. Encyc 2023;3(3):1133-1144. [doi: 10.3390/encyclopedia3030082]
23. Krentzman AR, Barker SL. Counselors' perspectives of positive psychology for the treatment of addiction: a mixed methods pilot study. Alcohol Treat Q 2016;34(4):370-385. [doi: 10.1080/07347324.2016.1217705] [Medline: 27980355]
24. Sugarman DE, Meyer LE, Reilly ME, Greenfield SF. Feasibility and acceptability of a web-based, gender-specific intervention for women with substance use disorders. J Womens Health (Larchmt) 2020 May;29(5):636-646. [doi: 10.1089/jwh.2018.7519] [Medline: 31584313]
25. White AM, Orosz A, Powell PA, Koob GF. Alcohol and aging - an area of increasing concern. Alcohol 2023 Mar;107:19-27. [doi: 10.1016/j.alcohol.2022.07.005] [Medline: 35940508]

Abbreviations

CBT: cognitive behavioral therapy
DTC: drink to cope
HIPAA: Health Insurance Portability and Accountability Act
PPI: positive psychology intervention
QI: quality improvement
REDCap: Research Electronic Data Capture
SUS: System Usability Scale

XSL•FO
RenderX

Edited by A Mavragani; submitted 30.06.24; peer-reviewed by E Grekin; revised version received 05.12.24; accepted 06.12.24; published 10.02.25.

Please cite as:
Serck-Hanssen I, Solheim-Witt M, Anker JJ, Sugarman DE
A Web-Based Resource Informed by Cognitive Behavioral Therapy and Positive Psychology to Address Stress, Negative Affect, and Problematic Alcohol Use: A Usability and Descriptive Study
JMIR Form Res 2025;9:e63819
URL: https://formative.jmir.org/2025/1/e63819
doi:10.2196/63819

© Ingrid Serck-Hanssen, Marit Solheim-Witt, Justin J Anker, Dawn E Sugarman. Originally published in JMIR Formative Research (https://formative.jmir.org), 10.2.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Step Count Accuracy of the Life Plus Connected Watch at Different Localizations and Speeds in Healthy Adults, Patients With Cardiovascular Disease, and Patients With Peripheral Artery Disease: Step Count Validation Study in Laboratory Settings

Anne-Noëlle Heizmann[1,2], PhD; Edouard Ollier[1,3], MCU-PH; Pierre Labeix[4], PhD; Ivan Goujon[4], MSc; Frédéric Roche[1,2,4], PUPH; Claire Le Hello[1,2,5], PUPH

[1]INSERM, SAINBIOSE U1059, Saint-Priest-en-Jarez, France

[2]University Jean Monnet, Saint-Etienne, France

[3]Innovation and Pharmacology Clinical Research Unit, University Hospital of Saint-Etienne, Saint-Etienne, France

[4]Department of Physiology, University Hospital of Saint-Etienne, Saint-Etienne, France

[5]Department of Therapeutics and Vascular Medicine, University Hospital of Saint-Etienne, Saint-Etienne, France

Corresponding Author:
Anne-Noëlle Heizmann, PhD
INSERM, SAINBIOSE U1059, Saint-Priest-en-Jarez, France

Abstract

Background: Smartwatches are increasingly used to monitor and motivate physical activity. Patients with cardiovascular disease (CVD) and peripheral artery disease (PAD) often do not meet national physical activity recommendations. They may, thus, benefit from a physical activity program using smartwatches. The Life Plus smartwatch is designed to facilitate activity monitoring by counting steps, but its validity needs to be determined, particularly in patients who may not have a normal gait, such as those with cardiovascular pathology.

Objective: This study evaluates the accuracy of the Life Plus smartwatch (versions 2 and 3) in healthy adults, patients with CVD, and patients with PAD at different walking speeds (1.8, 2.5, 3.2, and 4 km/h) and different localizations (wrists, hips, and ankles) to determine best accuracy.

Methods: In total, 34 participants, comprising healthy individuals (n=10), patients with CVD (n=14), and patients with PAD (n=10), wore 6 Life Plus watches simultaneously (3 of version 2 and 3 of version 3), located on wrists, hips, and ankles. Participants walked on a treadmill for 3-minute sessions at speeds of 1.8, 2.5, 3.2, and 4 km/h; they then performed a 10-minute free walking on the ground and again walked for 3-minute sessions on a treadmill at the same speeds. Actual step counts were recorded through video footage.

Results: When worn at the wrist, no significant difference between the actual number of steps and step count by version 2 watches was found in each group independently (healthy group: P=.25; CVD group: P=.50; and PAD group: P=.37). Significant differences were found with the version 2 watches at the wrist in the healthy group at 3.2 (−5.26%; P=.01) and 4 km/h (−6.13%; P=.008) and in the CVD group at 2.5 (−5.94%; P=.008), 3.2(−13.1%; P=.008), and 4 km/h (−13.96%; P=.004). When worn at the wrist, no significant difference between actual number of steps and step count by version 3 watches was found in the healthy group (P=.02) and the PAD group (P=.90). Significant differences were found with the version 3 watches at the wrist in the CVD group at 3.2 (−6.43%; P=.001) and 4 km/h (−7.3%; P=.01) and in the PAD group at 4 km/h (−5.77%; P=.04).

Conclusions: For optimized counting when worn at the wrist, healthy individuals should prefer version 2 watches for slow walking (1.8 and 2.5 km/h) and version 3 for higher speeds (3.2 and 4 km/h). Patients (CVD and PAD) should prefer version 2 watches at 1.8 km/h and version 3 watches for higher speeds (2.5, 3.2, and 4 km/h).

(JMIR Form Res 2025;9:e58964) doi:10.2196/58964

KEYWORDS

physical activity assessment; accuracy; validation; steps; smartwatches; wearables; cardiovascular disease; cardiovascular; cardiology; heart; artery; peripheral artery disease; wearable physical activity monitoring device; physical activity; exercise; walking; monitoring; mobile phone

XSL•FO
RenderX

Introduction

The benefits of regular physical activity on global health are well documented [1-4]. The key role of physical activity in cardiovascular diseases (CVDs) including peripheral artery disease (PAD) has been well reported within the scientific literature and has even been recognized as an effective nonmedical therapy and core component of cardiac rehabilitation [5,6]. Walking is the most natural physical activity that confers many benefits with minimal adverse effects [7-9]. In this context, daily step count seems to be an interesting target for achieving physical activity recommendations and setting physical activity goals [10]. Daily step count can easily be assessed and monitored by the patients themselves through wearable physical activity monitoring devices like a smartwatch. In healthy populations, wearable devices such as smartwatches are frequently used as motivational and self-monitoring tools for physical activity [11-14]. These wearable physical activity tracker devices offer considerable advantages in health care and personalized physical activity management in populations with chronic disease [15]. The gait of these patients can be different than those of healthy persons and may therefore influence the wearable device's accuracy [16-19]. Patients with PAD are known to have an altered gait cycle and a reduction in step lengths [20-23].

The Life Plus connected watch is a wearable physical activity monitoring device that provides information about daily steps and daily walking activity. It also offers personalized and progressive goals, feedback, encouragements, and personalized innovative challenges, taking into account the users' personal physical activity level and capacity. This watch was designed to be easy to use and not requiring a smartphone app, synchronization, or an internet connection. This watch seems to be an interesting tool to motivate patients to walk more. As this connected watch has not been evaluated, we wanted to estimate its accuracy in a healthy adult population as well as in populations with CVD or PAD. The aim of this study was to determine the accuracy of the Life Plus connected watch (versions 2 and 3) in counting steps under controlled treadmill walking conditions at different speeds in healthy adults and patients with CVD or PAD, with each participant wearing 6 watches at different locations (wrists, hips, and ankles). In view of the current literature, we hypothesize that the Life Plus watch is reliable for step counting with variations according to speeds, localizations, and the different groups (healthy adults, patients with CVD, and patients with PAD).

Methods

Recruitment

The inclusion criteria were adult healthy volunteers, adult persons with CVD or PAD (asymptomatic stage or stress ischemia stage) able to give written informed consent, to understand walking instructions to walk on a treadmill alone without technical help for several periods of 3 minutes at different speeds (1.8 to 4 km/h). Patients with CVD were recruited in the "Viellissement, Système Nerveux Autonome et Sommeil" (aging, autonomic nervous system, and sleep) center of the University Hospital of Saint-Etienne during their cardiac rehabilitation program. Patients with PAD were recruited in the Vascular Medicine Department of the University Hospital of Saint-Etienne during a scheduled ultrasound test for PAD. Investigators (CLH, ANH, FR, PL, and IG) explained the purpose and described the procedure to the patients who then decided whether or not they wanted to take part in the study. Healthy adults were volunteers recruited from medical and paramedical professionals at Saint-Etienne University Hospital, who gave their oral consent after a detailed presentation of the study. Exclusion criteria were volunteers and patients unable to walk independently, patients with PAD at chronic limb–threatening ischemia stage, and patients with walking limitations due to other pathologies. The ability to walk on a treadmill for 3 minutes at different paces was reported by all the patients recruited. None of the patients or healthy adults underwent any surgical procedures in the 3 months preceding the treadmill walking sessions.

Ethical Considerations

This study was approved by the ethics committee of the University Hospital of Saint-Etienne in June 2022 (reference CE 2022 - 36). The purpose of the study and the description of the procedure were explained to the participants before they signed an informed consent. Retrieved study data were deidentified and anonymized. No compensation was provided for participation.

Material and Data Collection

Each participant wore simultaneously 6 Life Plus connected watches: 3 version 2 and 3 version 3, version 2 positioned on one side with one at the wrist, one at the hip, and one at the ankle; and version 3 on the other side in the same localizations. This disposition was determined to evaluate versions 2 and 3 Life Plus connected watches and to identify the best placement for better accuracy since the position may influence its accuracy [24]. Treadmill walking provided a controlled and easily replicable physical activity for evaluating the accuracy of step counting at different speeds by watches [25].

The Life Plus connected watch uses a triaxial accelerometer. Versions 2 and 3 are embedded with different accelerometers and step counting algorithms. Life Plus watches integrate a microelectromechanical system technology that is a spring-levered step counter. Step count recording was collected directly through the Life Plus watch screen. Videos of the participant's lower limbs were recorded for a visual count of steps taken during all the walking sessions. Actual numbers of steps were counted a posteriori through visual inspection by the investigators with a camera of a Samsung Galaxy TAB S6 Lite under Android, which has a number of effective pixels of 12.8 megapixels (per million) with a full high-definition resolution of 1920×1080 (30 frames per screen).

The following data were retrieved: step counts by the watches, actual numbers of steps (videos), watch version (versions 2 and 3), group (healthy adults, patients with CVD, and patients with PAD), watch position (wrist, hip, and ankle), walking speed (1.8, 2.5, 3.2, and 4 km/h), help from the treadmill's bars (yes

XSL•FO
RenderX

and no), patients characteristics (age, BMI, and ankle-brachial index for patients with PAD).

Study Design

After giving written informed consent, participants were invited to walk 4 sessions of 3 minutes on the treadmill at different speeds (1.8, 2.5, 3.2, and 4 km/h). As far as possible, participants were asked to let go of the treadmill bars as soon as they started their walking movement so that their arms could swing normally. Between each session, participants had to sit still on a chair for at least 2 minutes to allow step count data to be refreshed on the watch screen. After the first 4 sessions, the participants had a 10-minute free walking session at their own pace in the hospital corridor to simulate a fatigue pattern. Participants then again walked 4 sessions of 3 minutes on the treadmill at the same speeds as the first time. The treadmill walking sessions were filmed so that 2 investigators could count the actual number of steps taken during all the treadmill walking sessions (criterion measure). The study design is illustrated in Figure 1.

Figure 1. Study design.

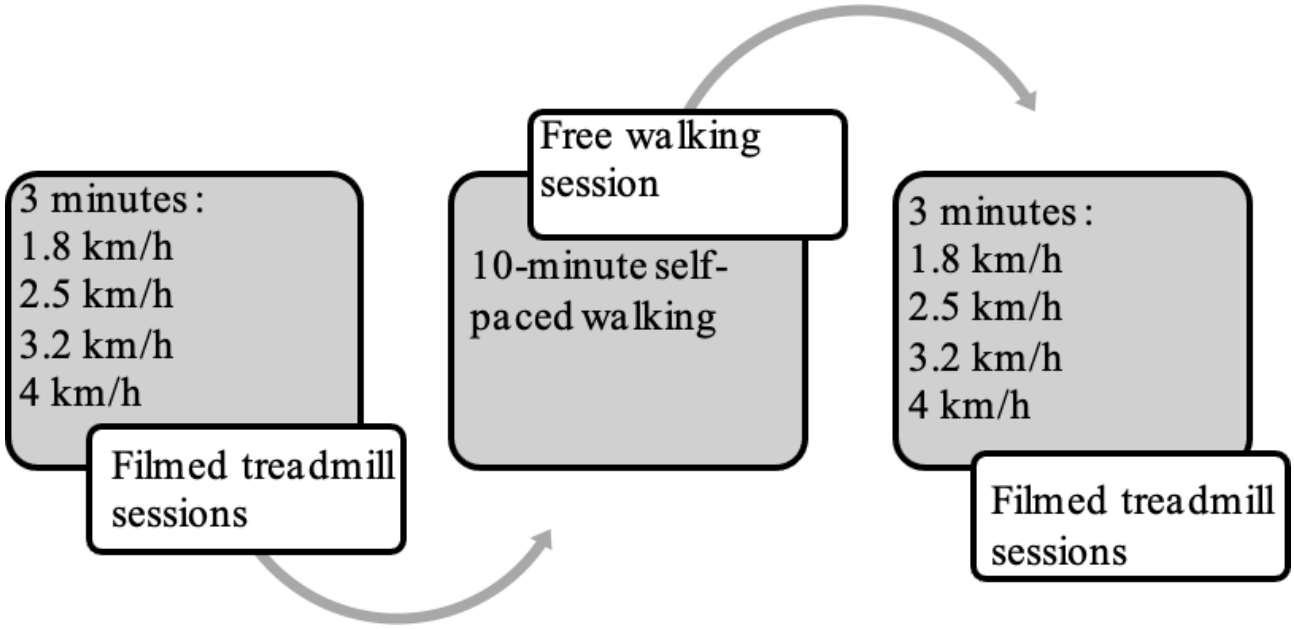

Statistical Analysis

Each participant was assessed in 8 walking sessions at 1.8, 2.5, 3.2, and 4 km/h. In all the conditions, the difference between the step count recorded by the Life Plus connected watches and the actual number of steps counted by the investigators was computed. We calculated the quantiles Q1 and Q3 for the data on actual number of steps, then the IQR=Q3−Q1; then we calculated the lower and upper thresholds for the outliers with lower threshold=Q1−1.5*IQR and upper threshold=Q3+1.5*IQR. Outliers were removed from the data collected by the watches. For all analyses, the α significance threshold used was 5%. Concordance between the step count recorded by the connected watches and the actual number of steps was quantified by calculating the difference between the actual number of steps and the step count measured by the watches. These differences were analyzed using the Wilcoxon test as the Shapiro-Wilk test was performed and showed deviance from normality for actual number of steps and step count by watches.

ANOVA was computed to compare differences in age and BMI between groups, and the Tukey test was used to determine which differences were significant. Finally, Bland-Altman charts were constructed to evaluate the mean bias and the limits of agreement between step count by watches and actual numbers of steps. All analyses were conducted using the R Statistical language (version 4.0.4; R Foundation for Statistical Computing).

Results

Demographics

In total, 34 participants were included in the study: 10 in the healthy group, 14 in the CVD group, and 10 in the PAD group. Characteristics of the 3 groups are summarized in Table 1.

XSL•FO
RenderX

Table . Demographic characteristics of the 3 groups included in the Life Plus watch validation study (healthy adults, patients with cardiovascular disease [CVD], and patients with peripheral artery disease [PAD]).

	Healthy group (n=10)	CVD group (n=14)	PAD group (n=10)
Age (years), mean (SD)	32.5 (11)	59.7 (14.1)	65.1 (9.8)
Sex, n (%)			
Female	5 (50)	2 (14)	2 (20)
Male	5 (50)	12 (86)	8 (80)
BMI (kg/m^2), mean (SD)	22.18 (2.98)	24.22 (1.89)	26.85 (4.03)
Ankle-brachial index, mean (SD)			
Right	—[a]	—	0.84 (0.28)
Left	—	—	0.88 (0.18)
Toe pressure (mm Hg), mean (SD)			
Right	—	—	78.6 (20.1)
Left	—	—	77.6 (26.6)

[a]Not available.

No significant age difference between the CVD and PAD groups (*P*=.53) was found, but significant differences between the healthy group and PAD group and the healthy group and CVD group (*P*<.001) were found. For BMI, analyses showed no difference between the CVD and PAD groups (*P*=.14) and between the healthy group and CVD group (*P*=.29) but significant differences between the healthy group and PAD group (*P*=.005).

Evaluation Outcomes

Global statistical parameters are summarized in Table 2.

Table . Global statistical parameters for the number of steps counted by the investigators (number of actual steps) and counted by the Life Plus watches (versions 2 and 3).

	Number of actual steps	Number of steps measured by the watch (outliers removed)
Range	185-399	176-412
Median (IQR)	296.5 (264-324)	310 (277-343)
Mean (SD)	292.6 (43.73)	307.3 (49.71)

Independently from any condition, the Life Plus step watch overestimates the actual number of steps by mean+14.7 steps (*P*=.001; *r*=0.08; 95% CI 2-8.5).

Life Plus Watch Version 2

Independently from all conditions, the version 2 watch did not correctly trigger automatic step counting (no step counted or less than the lower threshold of outliers) in 34% (n=81) walking sessions of the healthy group, 39% (n=116) in the CVD group, and 40% (n=95) in the PAD group. When looking closer, this technical issue occurred more frequently at slow speeds (1.8 and 2.5 km/h), when participants were holding the treadmill bars, and when the watch was worn at the hip or ankle. In the healthy group at 3.2 and 4 km/h when worn at the wrist, the version 2 watch triggered step counting in each walking session. In the CVD and PAD groups, at 3.2 and 4 km/h speeds, the version 2 watch triggered step counting more efficiently. Detailed statistics are summarized in Table S1 in Multimedia Appendix 1.

When worn at the wrist, no significant difference between actual number of steps and step count by watches was found (*P*=.89; 95% CI −8.0 to 9.0). The same result was found in each group independently (healthy group: *P*=.25; 95% CI −5.99 to −21; CVD group: *P*=.50; 95% CI −19.9 to 8.9; and PAD group: *P*=.37; 95% CI −8.0 to 22).

Detailed mean differences and mean percentage differences between the actual number of steps and step count by the watches are summarized in Table 3.

In the healthy group, when the watches were worn at the wrist, without holding treadmill bars and with speeds of 3.2 and 4 km/h, they underestimated the actual number of steps by mean 5.4% (*P*=.008 at 3.2 and 4 km/h). Differences between actual number of steps and step count by the watches were the lowest and nonsignificant when the watch was worn at the wrist at 1.8 km/h (*P*=.50; 95% CI 8-22) and 2.5 km/h (*P*=.21; 95% CI −22 to 3.5). In the CVD group, the mean percentage difference between the actual number of steps and the step count by the watches ranged from −13.96% (n=) at 4 km/h (*P*=.004; 95% CI −15.5 to −2.5) to −0.83% at 1.8 km/h (*P*=.13; 95% CI −126 to 24) when the watch was worn at the wrist. In the PAD group, the mean percentage difference between actual number of steps and the number of steps counted by the watches ranged from −12.44% at 4 km/h (*P*=.65; 95% CI −46 to 13.5) to +0.70% at

XSL•FO
RenderX

1.8 km/h (*P*=.81; 95% CI −49.5 to 35) when the watch was worn at the wrist.

Table . Mean step differences and mean percentage differences between actual number of steps and number of steps counted by version 2 watch in the 3 groups (healthy adults, patients with cardiovascular disease [CVD], and patients with peripheral artery disease [PAD]) at different localizations (wrist, hip, and ankle) and different speeds (1.8, 2.5, 3.2, and 4 km/h).

Speed and position		Groups		
		Healthy, mean step difference (%)	CVD, mean step difference (%)	PAD, mean step difference (%)
1.8 km/h				
	Wrist	9.37 (7.01)	−11.93 (−0.83)	−6.64 (0.70)
	Hip	17.94 (10.17)	−3.37 (2.33)	1.92 (3.86)
	Ankle	55.87 (23.85)	34.57 (16.02)	39.86 (17.54)
2.5 km/h				
	Wrist	2.24 (1.89)	−19.06 (−5.94)	−13.78 (−4.42)
	Hip	10.80 (5.06)	−10.50 (2.78)	5.21 (−1.26)
	Ankle	48.73 (18.74)	27.43 (10.90)	32.72 (12.43)
3.2 km/h				
	Wrist	−16.92 (−5.26)	−38.00 (−13.10)	−32.93 (−11.57)
	Hip	−8.36 (−2.10)	−29.66 (−9.94)	24.37 (−8.41)
	Ankle	29.58 (11.58)	8.28 (3.75)	13.56 (5.27)
4 km/h				
	Wrist	−18.76 (−6.13)	−40.06 (−13.96)	−34.78 (−12.44)
	Hip	10.20 (−2.96)	−31.50 (−10.80)	−26.21 (−9.27)
	Ankle	27.73 (10.72)	6.43 (2.88)	11.72 (4.41)

The Bland-Altman plots revealed no systematic differences between the actual number of steps and the number of steps counted by the watches for the 3 groups when the watch was worn at the wrist (Figure S1 in Multimedia Appendix 1).

Life Plus Watch Version 3

Independently from all conditions, the version 3 watch did not correctly trigger automatic step counting (no step counted or less than the lower threshold of outliers) in 35% (n=83) of walking sessions in the healthy group, 43% (n=129) in the CVD group, and 43% (n=103) in the PAD group. When looking closer, this technical issue occurred more frequently at 1.8 and 2.5 km/h, when participants were holding the treadmill bars, and when the watch was worn at the hip or ankle. At 3.2 and 4 km/h and with the watch worn at the ankle, the watch version 3 triggered step counting more frequently. Detailed statistics are summarized in Table S2 in Multimedia Appendix 1.

When worn at the wrist, no significant difference between actual number of steps and number of steps counted by the watches (*P*=.52; 95% CI −15 to 7.0). The same result was found in the healthy group (*P*=.02; 95% CI 2.99-41.99) and the PAD group (*P*=.90; 95% CI −19 to +20). A significant difference was found in the CVD group (*P=.006; P*=.01; 95% CI −43 to +9) at higher speeds (3.2 and 4 km/h) but not at lower speeds (1.8 and 2.5 km/h). Mean differences and mean percentage differences between the actual number of steps and the number of steps counted by the watches are summarized in Table 4.

XSL•FO
RenderX

Table . Mean step differences and mean percentage differences between actual number of steps and number of steps counted by version 3 watch in the 3 groups (healthy adults, patients with cardiovascular disease [CVD], and patients with peripheral artery disease [PAD]) at different positions (wrist, hip, and ankle) and different speeds (1.8, 2.5, 3.2, and 4 km/h).

Speed and position		Groups		
		Healthy, mean step difference (%)	CVD, mean step difference (%)	PAD, mean step difference (%)
1.8 km/h				
	Wrist	28.28 (13.67)	6.98 (5.84)	12.27 (7.36)
	Hip	36.84 (16.83)	15.54 (8.99)	20.83 (10.52)
	Ankle	74.78 (30.51)	53.48 (22.68)	58.77 (24.20)
2.5 km/h				
	Wrist	21.15 (8.56)	−0.16 (0.72)	5.13 (2.25)
	Hip	29.71 (11.72)	8.41 (3.88)	13.70 (5.41)
	Ankle	67.64 (25.40)	46.34 (17.57)	51.63 (9.09)
3.2 km/h				
	Wrist	1.99 (1.40)	−14.03 (−6.43)	−19.31 (−4.91)
	Hip	10.55 (4.56)	−10.75 (−3.27)	−5.46 (−1.75)
	Ankle	48.49 (11.58)	27.18 (10.41)	32.47 (11.94)
4 km/h				
	Wrist	0.15 (0.54)	−21.16 (−7.30)	−15.87 (−5.77)
	Hip	8.71 (3.70)	−12.59 (−4.14)	−7.31 (−2.61)
	Ankle	46.64 (17.38)	25.34 (9.55)	30.63 (11.07)

Results showed that in the healthy group, differences between actual number of steps and steps counted by the watches were lower and nonsignificant when the watch was worn at the wrist at all speeds. In the healthy group, the version 3 watch tended to globally overestimate step counts in all conditions. In the CVD group, the mean percentage difference between actual number of steps and number of steps counted by the watch ranged from −7.29% (P=.01; 95% CI −32.5 to 108.5) at 4 km/h to +5.84% (P=.25; 95% CI −91 to 22) at 1.8 km/h when the watch was worn at the wrist. In the PAD group, the mean percentage of difference between the actual number of steps and the number of steps counted by the watches ranged from −5.77% (P=.04; 95% CI −59 to −4) at 4 km/h to +7.36% (P=1.0; 95% CI −156 to 81) at 1.8 km/h when the watch was worn at the wrist.

The Bland-Altman plots revealed no systematic differences between the actual number of steps and the step count obtained by the watches when the watch was worn at the wrist (Figure S2 in Multimedia Appendix 1).

Discussion

Principal Findings

Activity trackers were designed to help people to monitor their daily physical activity. A smartwatch able to accurately count the number of steps and provide motivating features, such as the Life Plus watch, would benefit patients' physical activity. Indeed, it is important to have a reliable activity tracker [26] to support and coach patients in home-based physical activity rehabilitation. Inaccurate information on physical activity may lead to frustration for the patient if his activity is underestimated or to false positive feedback if overestimated. As these tools often are tested in healthy populations, we wanted to test the accuracy of the Life Plus watch for counting steps in healthy volunteers, patients with CVD, and patients with PAD.

On the basis of our investigation, the Life Plus smartwatches seemed to be accurate for counting steps in healthy persons when walking on a treadmill without holding the lateral bars. The version 3 version was better in counting steps.

Our results indicated that in the healthy group, when worn at the wrist, the version 2 watches were more accurate at 1.8 and 2.5 km/h; whereas at 3.2 and 4 km/h, the version 3 watches were more accurate. Healthy adults should wear the watch, version 2 or 3, at the wrist for better accuracy. In the CVD and PAD groups, versions 2 and 3 watches have similar accuracy at 1.8 and 2.5 km/h when worn at the wrist or hips; at 3.2 and 4 km/h, the version 3 watches are more accurate than the version 2 when worn at the wrist or hips. In practice, the version 3 watch seems to be the better option in all groups with better accuracy at all speeds when worn at the wrist or hips.

Differences between the number of steps counted by watches according to their version (versions 2 and 3) are explained by the different accelerometers and corresponding step counting algorithm embedded in versions 2 and 3.

Comparison With Prior Work and Limitations

Globally, our investigation showed that step count accuracy decreased as walking speed decreased, which is consistent with

XSL•FO
RenderX

other validation studies [27,28]. In our study design, speeds were not counterbalanced; therefore, we could not ensure that there was no order effect on step counts.

The localization of the watch on the body also influences the step count accuracy, with overestimation when worn at the ankles, and underestimation when worn at the hips at higher speeds. This result follows other validation studies [29].

The watch had slightly lower accuracy in the CVD and PAD groups. This could be explained by the fact that the patients were more often holding the treadmill lateral bars to maintain stability, and wrist-worn wearable devices typically undercount the number of steps when hands are fixed on lateral hand bars [30]. Moreover, patients may have had some impaired arm swing and gait pattern as shown in several studies [17,18,20-23]. Furthermore, age, sex, and BMI also may influence the validity of wearable devices [31]. These could also have affected our results as the mean age of the healthy participant's group was much lower than that of the 2 disease groups. This makes the comparison between groups difficult. Results could also be influenced by the fact that the study design did not propose a familiarization session on the treadmill. Walking on a treadmill is not natural and requires a lot of attention at first. Several studies showed indeed higher cadences, shorter step lengths, shorter swing phases, longer stance phases, and higher walking variability when walking on a treadmill [32,33]. A 10-minute session to familiarize with walking on a treadmill could enable to be in conditions nearer to natural walking [34].

Participants only walked for 3 minutes for each speed. The accelerometer has a wake-up threshold, meaning it only starts measuring movement when a certain level of acceleration is detected. At slower speeds, the arm swing may not be sufficiently pronounced to reach the accelerometer's step count threshold. The algorithm has set specific thresholds that need to be met for a movement to be recognized as a step. If the arm swing is too subtle, it may not reach these thresholds, resulting in steps not being counted. This could explain the percentage of errors in step count.

Our research evaluated the accuracy of the Life Plus watch, versions 2 and 3, in laboratory conditions. Future research should investigate the accuracy of the watch during free-living conditions.

Conclusions

We investigated the accuracy of versions 2 and 3 of the Life Plus watch in healthy adults, patients with CVD, and patients with PAD. Accuracy of the version 3 was better for all walking speeds when worn at the wrist. Mean percentage differences between actual number of steps (videos) and number of steps counted by watches were less than 10% when version 3 was worn at the wrist by each group at all speeds except 1.8 km/h. Globally, independently from the watch version, accuracy slightly decreased when worn by patients with CVD or PAD walking at slower speeds. In the CVD and PAD groups, the accuracy of version 3 was better when worn at the wrist for slower walking speeds (1.8 and 2.5 km/h) and when worn at the hips for higher walking speeds. In practice, when worn at the wrist as it is intended, healthy individuals should prefer version 2 watches for slow walking (1.8 and 2.5 km/h) and version 3 for higher speeds (3.2 and 4 km/h). Patients (CVD and PAD) should prefer version 2 watches at 1.8 km/h and version 3 watches for higher speeds (2.5, 3.2, and 4 km/h).

Acknowledgments

The researchers wish to thank all the volunteers for their time and the staff of the Viellissement, Système Nerveux Autonome et Sommeil rehabilitation center of Saint Etienne University Hospital for the treadmill use and patient recruitment help.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

ANH and CLH conducted the experiment, EO helped with the statistical analyses, IG and PL helped in patients' recruitment, and FR allowed access to cardiac rehabilitation. All coauthors reviewed and approved the manuscript for publication.

Conflicts of Interest

ANH is a PhD student working in collaboration with the Life Plus company as part of her Convention Industrielle de Formation par la Recherche (industrial agreements for training through research) thesis (optimization and use of the Life Plus watch in rehabilitation of patients with cardiovascular disease). All other authors have no known conflict of interest to disclose.

Multimedia Appendix 1
Outliers table and Bland-Altman plots.
[DOCX File, 1096 KB - formative_v9i1e58964_app1.docx]

References

XSL•FO
RenderX

1. Ekelund U, Tarp J, Steene-Johannessen J, et al. Dose-response associations between accelerometry measured physical activity and sedentary time and all cause mortality: systematic review and harmonised meta-analysis. BMJ 2019 Aug 21;366:l4570. [doi: 10.1136/bmj.l4570] [Medline: 31434697]
2. Li J, Siegrist J. Physical activity and risk of cardiovascular disease—a meta-analysis of prospective cohort studies. Int J Environ Res Public Health 2012 Feb;9(2):391-407. [doi: 10.3390/ijerph9020391] [Medline: 22470299]
3. Warburton DER. Health benefits of physical activity: the evidence. Can Med Assoc J 2006 Mar 14;174(6):801-809. [doi: 10.1503/cmaj.051351]
4. Cardiovascular diseases (CVDs). World Health Organization. 2021. URL: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) [accessed 2025-01-09]
5. Gerhard-Herman MD, Gornik HL, Barrett C, et al. 2016 AHA/ACC guideline on the management of patients with lower extremity peripheral artery disease: executive summary: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. Circulation 2017 Mar 21;135(12):e686-e725. [doi: 10.1161/CIR.0000000000000470] [Medline: 27840332]
6. Casillas JM, Troisgros O, Hannequin A, et al. Rehabilitation in patients with peripheral arterial disease. Ann Phys Rehabil Med 2011 Oct;54(7):443-461. [doi: 10.1016/j.rehab.2011.07.001] [Medline: 21917541]
7. Hall KS, Hyde ET, Bassett DR, et al. Systematic review of the prospective association of daily step counts with risk of mortality, cardiovascular disease, and dysglycemia. Int J Behav Nutr Phys Act 2020 Jun 20;17(1):78. [doi: 10.1186/s12966-020-00978-9] [Medline: 32563261]
8. Morris JN, Hardman AE. Walking to health. Sports Med 1997 May;23(5):306-332. [doi: 10.2165/00007256-199723050-00004] [Medline: 9181668]
9. Oja P, Kelly P, Murtagh EM, Murphy MH, Foster C, Titze S. Effects of frequency, intensity, duration and volume of walking interventions on CVD risk factors: a systematic review and meta-regression analysis of randomised controlled trials among inactive healthy adults. Br J Sports Med 2018 Jun;52(12):769-775. [doi: 10.1136/bjsports-2017-098558] [Medline: 29858464]
10. Kraus WE, Janz KF, Powell KE, et al. Daily step counts for measuring physical activity exposure and its relation to health. Med Sci Sports Exerc 2019;51(6):1206-1212. [doi: 10.1249/MSS.0000000000001932]
11. Bravata DM, Smith-Spangler C, Sundaram V, et al. Using pedometers to increase physical activity and improve health: a systematic review. JAMA 2007 Nov 21;298(19):2296-2304. [doi: 10.1001/jama.298.19.2296] [Medline: 18029834]
12. Sanders JP, Loveday A, Pearson N, et al. Devices for self-monitoring sedentary time or physical activity: a scoping review. J Med Internet Res 2016 May 4;18(5):e90. [doi: 10.2196/jmir.5373] [Medline: 27145905]
13. Vetrovsky T, Siranec M, Marencakova J, et al. Validity of six consumer-level activity monitors for measuring steps in patients with chronic heart failure. PLOS ONE 2019;14(9):e0222569. [doi: 10.1371/journal.pone.0222569] [Medline: 31518367]
14. Brickwood KJ, Watson G, O'Brien J, Williams AD. Consumer-based wearable activity trackers increase physical activity participation: systematic review and meta-analysis. JMIR Mhealth Uhealth 2019 Apr 12;7(4):e11819. [doi: 10.2196/11819] [Medline: 30977740]
15. Park LG, Elnaggar A, Lee SJ, et al. Mobile health intervention promoting physical activity in adults post cardiac rehabilitation: pilot randomized controlled trial. JMIR Form Res 2021 Apr 16;5(4):e20468. [doi: 10.2196/20468] [Medline: 33861204]
16. Ata R, Gandhi N, Rasmussen H, et al. Clinical validation of smartphone-based activity tracking in peripheral artery disease patients. NPJ Digit Med 2018;1(1):66. [doi: 10.1038/s41746-018-0073-x] [Medline: 31304343]
17. Ozcan EB, Saglam M, Vardar-Yagli N, et al. Impaired balance and gait characteristics in patients with chronic heart failure. Heart Lung Circ 2022 Jun;31(6):832-840. [doi: 10.1016/j.hlc.2021.10.015] [Medline: 35177316]
18. Panizzolo FA, Maiorana AJ, Naylor LH, et al. Gait analysis in chronic heart failure: the calf as a locus of impaired walking capacity. J Biomech 2014 Nov 28;47(15):3719-3725. [doi: 10.1016/j.jbiomech.2014.09.015] [Medline: 25307437]
19. Thorup CB, Andreasen JJ, Sørensen EE, Grønkjær M, Dinesen BI, Hansen J. Accuracy of a step counter during treadmill and daily life walking by healthy adults and patients with cardiac disease. BMJ Open 2017 Mar 31;7(3):e011742. [doi: 10.1136/bmjopen-2016-011742] [Medline: 28363918]
20. Gardner AW, Forrester L, Smith GV. Altered gait profile in subjects with peripheral arterial disease. Vasc Med 2001;6(1):31-34. [doi: 10.1177/1358836X0100600106] [Medline: 11358158]
21. McDermott MM, Ohlmiller SM, Liu K, et al. Gait alterations associated with walking impairment in people with peripheral arterial disease with and without intermittent claudication. J Am Geriatr Soc 2001 Jun;49(6):747-754. [doi: 10.1046/j.1532-5415.2001.49151.x] [Medline: 11454113]
22. Myers SA, Johanning JM, Stergiou N, Celis RI, Robinson L, Pipinos II. Gait variability is altered in patients with peripheral arterial disease. J Vasc Surg 2009 Apr;49(4):924-931. [doi: 10.1016/j.jvs.2008.11.020] [Medline: 19217749]
23. Szymczak M, Krupa P, Oszkinis G, Majchrzycki M. Gait pattern in patients with peripheral artery disease. BMC Geriatr 2018 Feb 20;18(1):52. [doi: 10.1186/s12877-018-0727-1] [Medline: 29458330]
24. Tudor-Locke C, Barreira TV, Schuna JM. Comparison of step outputs for waist and wrist accelerometer attachment sites. Med Sci Sports Exerc 2015 Apr;47(4):839-842. [doi: 10.1249/MSS.0000000000000476] [Medline: 25121517]

XSL•FO
RenderX

25. Moore CC, McCullough AK, Aguiar EJ, Ducharme SW, Tudor-Locke C. Toward harmonized treadmill-based validation of step-counting wearable technologies: a scoping review. J Phys Act Health 2020 Aug 1;17(8):840-852. [doi: 10.1123/jpah.2019-0205] [Medline: 32652514]
26. Kang HS, Exworthy M. Wearing the future-wearables to empower users to take greater responsibility for their health and care: scoping review. JMIR Mhealth Uhealth 2022 Jul 13;10(7):e35684. [doi: 10.2196/35684] [Medline: 35830222]
27. Bassett DR, Toth LP, LaMunion SR, Crouter SE. Step counting: a review of measurement considerations and health-related applications. Sports Med 2017 Jul;47(7):1303-1315. [doi: 10.1007/s40279-016-0663-1]
28. Gould ZR, Mora-Gonzalez J, Aguiar EJ, et al. A catalog of validity indices for step counting wearable technologies during treadmill walking: the CADENCE-Kids study. Int J Behav Nutr Phys Act 2021 Dec;18(1):97. [doi: 10.1186/s12966-021-01167-y] [Medline: 36076265]
29. Kotler D, Wu C, Thompson K, et al. The accuracy of fitness trackers varies with their location on the body (P16-019-19). Curr Dev Nutr 2019 Jun;3:nzz050.P16-019-19. [doi: 10.1093/cdn/nzz050.P16-019-19]
30. Lynn R, Pfitzer R, Rogers RR, Ballmann CG, Williams TD, Marshall MR. Step-counting validity of wrist-worn activity monitors during activities with fixed upper extremities. J Meas Phys Behav 2020;3(3):197-203. [doi: 10.1123/jmpb.2019-0055]
31. Musa SB, Ellis R, Chafe B, et al. Wearable device validity in measuring steps, energy expenditure, and heart rate across age, gender, and body mass index: data analysis from a systematic review. J Phys Act Health 2023 Feb 1;20(2):100-105. [doi: 10.1123/jpah.2022-0160] [Medline: 36535270]
32. Murray MP, Spurr GB, Sepic SB, Gardner GM, Mollinger LA. Treadmill vs. floor walking: kinematics, electromyogram, and heart rate. J Appl Physiol (1985) 1985 Jul;59(1):87-91. [doi: 10.1152/jappl.1985.59.1.87] [Medline: 4030579]
33. Warabi T, Kato M, Kiriyama K, Yoshida T, Kobayashi N. Treadmill walking and overground walking of human subjects compared by recording sole-floor reaction force. Neurosci Res 2005 Nov;53(3):343-348. [doi: 10.1016/j.neures.2005.08.005] [Medline: 16182398]
34. Van de Putte M, Hagemeister N, St-Onge N, Parent G, de Guise JA. Habituation to treadmill walking. Biomed Mater Eng 2006;16(1):43-52. [Medline: 16410643]

Abbreviations

CVD: cardiovascular disease
PAD: peripheral artery disease

Edited by A Mavragani; submitted 29.03.24; peer-reviewed by G Piva, RJ Mays; revised version received 28.07.24; accepted 02.09.24; published 10.02.25.

Please cite as:
Heizmann AN, Ollier E, Labeix P, Goujon I, Roche F, Le Hello C
Step Count Accuracy of the Life Plus Connected Watch at Different Localizations and Speeds in Healthy Adults, Patients With Cardiovascular Disease, and Patients With Peripheral Artery Disease: Step Count Validation Study in Laboratory Settings
JMIR Form Res 2025;9:e58964
URL: https://formative.jmir.org/2025/1/e58964
doi:10.2196/58964

© Anne-Noëlle Heizmann, Edouard Ollier, Pierre Labeix, Ivan Goujon, Frédéric Roche, Claire Le Hello. Originally published in JMIR Formative Research (https://formative.jmir.org), 10.2.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Mediating Effect of Social Capital on the Association Between Digital Literacy and Life Satisfaction Among Older Adults in South Korea: Cross-Sectional Study

Hyein Jung[1], MA; Hocheol Lee[2], PhD; Eun Woo Nam[1,2], MPH, PhD

[1]Department of Health Administration, Yonsei University Graduate School, Wonju, Republic of Korea

[2]Division of Health Administration, College of Software and Digital Healthcare Convergence, Yonsei University, Unit 412, Chang-jo gwan 1 Yonseidae-gil, Wonju, Republic of Korea

Corresponding Author:
Eun Woo Nam, MPH, PhD
Department of Health Administration, Yonsei University Graduate School, Wonju, Republic of Korea

Abstract

Background: As Korea rapidly transforms into a super-aged society, research indicates that digital literacy among older adults enhances their life satisfaction. Digital literacy refers to the ability to efficiently use digital technologies, encompassing access, competency, and utilization. It reflects the capacity to navigate and benefit from digital environments effectively. Furthermore, social capital positively influences the quality of life, and digital literacy facilitates social capital formation. However, since most studies have only focused on the direct relationship between digital literacy and life satisfaction, research on the mediating role of social capital remains limited.

Objective: To analyze the effect of digital literacy on the life satisfaction of older adults in South Korea and to verify whether social capital acts as a mediating factor in this process.

Methods: This descriptive cross-sectional study used data from the 2023 Report on the Digital Divide—an annual survey conducted by the Korean Ministry of Science and Information and Communications Technology. The study targeted individuals aged 65 years or older. Descriptive statistics, the Pearson correlation analyses, and the 3-step multiple regression analysis proposed by Baron and Kenny were performed. The bootstrap method was employed, and all analyses were conducted using R, version 4.4.1.

Results: The study included 869 participants. Digital literacy had a significant positive effect on their life satisfaction (β=0.103; *P*=.008). Social capital was also positively associated with life satisfaction (β=0.337; *P*<.001). Mediation analysis showed that digital literacy influenced life satisfaction both directly (β=0.103; *P*=.006) and indirectly through social capital (β=0.037; *P*=.03). Bootstrapping confirmed the significance of the indirect effect (β=0.037, 95% CI 0.005 - 0.070; *P*=.03). The total effect of digital literacy on life satisfaction was also significant (β=0.140, 95% CI 0.058 - 0.230; *P*=.002).

Conclusions: This study analyzed the association between digital literacy, social capital, and life satisfaction among older adults in Korea. We identified that social capital mediates the association between digital literacy and life satisfaction among older adults. These findings indicate that tailored digital literacy programs and support policies that promote social capital formation could help bridge the digital divide and foster social inclusion. These measures would enable older adults to access essential services, reduce social isolation, and enhance health and well-being, ultimately improving the overall quality of life.

(JMIR Form Res 2025;9:e68163) doi:10.2196/68163

KEYWORDS

digital literacy; digital access; digital competency; digital utilization; life satisfaction; older adults; social capital; mediating effect; aging

Introduction

South Korea is rapidly approaching a super-aged society, where over 20% of the population is 65 years and older. As of 2024, approximately 9.9 million, or 19.2%, of the total population fall with in this age group. By 2025, this is projected to exceed 20%, officially marking the country's transition into a super-aged society [1]. This increase in the older adult population has led to a decrease in the proportion of the economically active population. Therefore, the financial and psychological burdens at the individual, family, and national levels are expected to increase due to aging-related diseases and care issues. In 2019, the medical expenses of older adults in Korea accounted for over 40% of the nation's total medical expenses, and the country recorded the highest elderly poverty rate among the Organisation for Economic Co-operation and Development (OECD) countries

XSL•FO
RenderX

[2]. Moreover, the life satisfaction of older adults aged 65 years and over decreased by 2.4% compared to the preceding year, reaching 31.9%, which is lower than the overall average. Life satisfaction also tends to decrease with age [1].

Enhancing the digital technology proficiency of older adults has been proposed as a potential solution to address these challenges [3,4]. Although internet use has grown significantly in recent years, disparities remain, especially among older adults. In South Korea, older adults aged 55 years and above are categorized as one of the 4 major digitally vulnerable groups, alongside people with disabilities, low-income households, farmers, and fishermen. Among these groups, older adults have the lowest level of digital literacy, highlighting a significant digital divide [5]. This digital gap limits access to essential resources and opportunities, hindering efforts to improve their quality of life.

As social systems become increasingly digitized, older adults experience considerable difficulties in managing various daily tasks, such as internet-based purchases, bill payments, and financial transactions. Furthermore, the rapid expansion of telemedicine and digital health care services following the COVID-19 pandemic has significantly increased reliance on digital platforms for health management. This shift has exacerbated challenges for older adults in effectively utilizing these technologies [6,7]. These difficulties are further intensified by resistance to new technologies, psychological barriers, and negative self-perceptions related to aging [4,8,9]. Consequently, these challenges isolate older adults socially and economically and hinder their ability to lead an independent and fulfilling life.

Digital literacy is the ability to efficiently use digital technologies, grounded in fundamental literacy skills such as reading, writing, and arithmetic. It encompasses digital access, digital competency, and digital utilization, enabling individuals to search, evaluate, utilize, share, and create content using digital tools such as the internet and smartphones [10,11].

Research has shown that digital literacy positively impacts the life satisfaction of older adults by enhancing personal development, social connections, and self-esteem [12]. Activities such as searching for financial and health information and engaging in internet-based communication boost self-efficacy, foster positive perceptions of aging, alleviate loneliness and isolation, and improve overall quality of life [13-15]. Domestic research has also demonstrated that the digital literacy level of older adults positively impacts their life satisfaction [16-18]. Moreover, it plays a crucial role in maintaining and expanding relationships with family, friends, and the community, thereby positively influencing the formation of social capital [19,20].

Social capital, which refers to the social networks, norms, and trust that arise from interactions among individuals and groups, serves as a resource that promotes a sense of community and enables individuals to lead fulfilling lives [21,22]. Previous studies have identified social capital as a vital resource that helps alleviate loneliness and enhances life satisfaction, thereby positively influencing the mental well-being of older adults [23-27]. When older adults effectively utilize digital technologies, their social capital is further strengthened, helping them overcome social isolation and enhance their quality of life [19]. Specifically, the enhancement of digital literacy contributes to the growth of social capital, thereby improving emotional stability and life satisfaction [15]. Digital literacy allows older adults to connect with broader social networks, which elevates their overall quality of life.

While several studies have explored the relationship between digital literacy and life satisfaction among older adults and the connection between social capital and quality of life, most have focused only on their direct relationships. Thus, research on how social capital mediates this connection remains limited. This highlights the need to better understand how digital literacy contributes to life satisfaction, considering the potentially mediating role of social capital. Therefore, this study aims to analyze the impact of digital literacy on the life satisfaction of Korean older adults and to verify whether social capital acts as a mediating factor in this process.

Methods

Study Design and Aim

This cross-sectional study is aimed at analyzing the mediating effect of social capital on the relationship between digital literacy and life satisfaction among older adults in Korea.

Data Collection

This study used raw data from the 2023 Report on the Digital Divide—an annual survey conducted by the Korean Ministry of Science and Information and Communications Technology (MSIT). Considering the OECD's definition of older adults and the age criteria commonly used for pension and social security benefits, this study defines older adults as those aged 65 years and above. Accordingly, response data from this demographic were extracted, processed, and analyzed. The survey employed a structured questionnaire administered through face-to-face interviews. Measures were taken to minimize response bias by standardizing the interview process and providing training for surveyors. The survey period lasted from October to December 2023. For analysis, the data were deidentified in compliance with the Personal Information Protection Act. The sample was selected using a proportional and stratified probability sampling method, accounting for inclusion probability based on region, gender, and age, and estimates of standard error for each stratum. Among the 1239 participants aged 65 years and older, data from 869 individuals were analyzed after excluding responses with missing values (n=370).

Study Variables

Life Satisfaction

The dependent variable in this study, life satisfaction, was measured across 8 domains of daily life: leisure and cultural activities, financial status, social activities, interpersonal relationships, family relationships, work, physical and mental health, and political engagement. These were assessed using a 4-point Likert scale, and the scores were converted to a 100-point scale. A Cronbach α of 0.806 was calculated to assess the internal consistency of the measurement items.

XSL•FO
RenderX

Digital Literacy

The independent variable in this study, digital literacy, was evaluated by integrating 3 components: digital access, digital competency, and digital utilization. Digital access was assessed using 4 items, examining whether participants owned digital devices such as desktops, laptops, and mobile phones and whether they had constant internet access (1 for yes, 0 for no). Digital competency was measured using 7 items on personal computer competency and 7 items on mobile device competency, each assessed using a 4-point Likert scale ranging from “Strongly disagree” (1) to “Strongly agree” (4). Digital utilization was assessed based on 50 items categorized by wired or mobile internet usage, the diversity of internet services used, and advanced internet utilization. The diversity of internet service usage included activities such as information and news searches, email, media content, educational content, social networking services, messengers, blogs, internet-based communities, cloud services, daily information, e-commerce, financial transactions, and public service utilization. Advanced internet utilization measured the extent of information production and sharing, networking, social participation, and economic activities. All items were rated on a 4-point Likert scale ranging from “Do not use at all” (1) to “Use frequently” (4). Scores for digital access, digital competency, and digital utilization were calculated on a weighted 100-point scale, as outlined in the 2023 Report on the Digital Divide. A Cronbach α of 0.967 confirmed the internal consistency of these items.

Social Capital

The mediating variable, social capital, was measured by assessing the presence of social support networks, such as family and friends, in the internet-based environment for older adults. The survey used a shortened version of the Internet Social Capital Scale (ISCS) developed by Williams [28], modified and adapted by the National Information Society Agency of Korea to evaluate social capital levels. The original ISCS consists of 20 items, with 10 items each for bonding social capital and bridging social capital. Bonding social capital assesses aspects such as emotional support, access to limited resources, and the ability to mobilize solidarity. Bridging social capital evaluates aspects such as an outward-looking perspective, contact with a wide range of people, a sense of belonging to a larger group, and interactions with broader communities. In this study, the shortened scale with 5 items each for bonding and bridging social capital, totaling 10 items, was used. All items were rated on a 4-point Likert scale from “Strongly disagree” (1) to “Strongly agree” (4) and converted to a 100-point scale, with a Cronbach α of 0.863 confirming reliability.

Control Variables

The control variables included sociodemographic characteristics such as gender, age, education level, disability status, household type, and monthly income. Age was categorized into 65 - 69, 70 - 74, 75 - 79, and 80 years or older. Education level was classified as lower than elementary school, middle school graduate, high school graduate, and college graduate or higher. Disability status was binary (“yes” or “no”), and household type was classified as either single-person households or households with 2 or more people. Monthly income was divided into 5 quintiles based on income levels.

Data Analysis

The data were analyzed using the R statistical program, version 4.4.1, and the specific analysis methods are as follows. First, descriptive statistical analyses, including frequency analysis, percentages, means, and SDs, were conducted to understand the participants’ sociodemographic characteristics, digital literacy level, life satisfaction, and social capital. Second, the Pearson correlation analyses were performed to identify the associations between the main variables of the study model: digital literacy, life satisfaction, and social capital. Third, to verify the mediating effect of social capital in the association between digital literacy and life satisfaction, a 3-step multiple regression analysis based on Baron and Kenny’s procedure [29] was conducted. The bootstrap method was applied to test the statistical significance of mediation effects, using a 95% CI.

Ethical Considerations

This study was conducted in accordance with the Declaration of Helsinki and received review and approval from the Institutional Research Ethics Committee of the Yonsei University Institutional Review Board in Wonju, South Korea (Approval No.1041849 - 202410-SB-213-01). All procedures followed the relevant institutional guidelines and regulations, and informed consent was obtained from all participants or their legal guardians before participation. To ensure privacy and confidentiality, the data used in this study underwent anonymization and deidentification processes.

Results

Participants Characteristics

A total of 869 participants were included in this study (Table 1). There were marginally more women (n=450, 51.8%) than men (n=419, 48.2%). By age, the largest group included those aged 65 - 69 years old (n=367, 42.2%), followed by those aged 70 - 74 years old (n=296, 34.1%), 75 - 79 years old (n=164, 18.9%), and 80 years and older (n=42, 4.8%). Regarding education level, high school graduates accounted for the largest proportion (n=345, 39.7%), followed by middle school graduates (n=289, 33.3%), those with elementary school education or lower (n=182, 20.9%), and college graduates or higher (n=53, 6.1%). In terms of disability status, most respondents (n=855, 98.4%) did not have a disability, while 1.6% (n=14) reported having a disability. Household type showed that 18.2% (n=158) lived in single-person households, while 81.8% (n=711) lived in households with 2 or more people. Regarding monthly income, the largest group earned ₩2 - ₩2.99 million (US $1392.30-$2081.49) (n=233, 26.8%), followed by ₩1 - ₩1.99 million (US $696.15-$1385.34) (n=216, 24.9%), ₩4 million or more (US $2784.61) (n=188, 21.6%), ₩3 - ₩3.99 million (US $2088.45-$2777.64) (n=152, 17.5%), and less than ₩1 million (US $696.15) (n=80, 9.2%).

XSL•FO
RenderX

Table . Respondents' characteristics and degree of study variables.

Characteristic	Participants (n=869)
Gender, n (%)	
Male	419 (48.2)
Female	450 (51.8)
Age (years), n (%)	
65 - 69	367 (42.2)
70 - 74	296 (34.1)
75 - 79	164 (18.9)
≥80	42 (4.8)
Education level, n (%)	
Elementary school or lower	182 (20.9)
Middle school	289 (33.3)
High school	345 (39.7)
College or higher	53 (6.1)
Disability, n (%)	
No	855 (98.4)
Yes	14 (1.6)
Live with someone, n (%)	
No	158 (18.2)
Yes	711 (81.8)
Income, n (%)[a]	
<1 million	80 (9.2)
1 - 1.99 million	216 (24.9)
2 - 2.99 million	233 (26.8)
3 - 3.99 million	152 (17.5)
≥4 million	188 (21.6)
Digital literacy[b], mean (SD)[c]	56.2 (10.5)
Digital access	62.4 (16.6)
Digital competency	47.2 (16.5)
Digital utilization	62.2 (6.2)
Life satisfaction, mean (SD)[c]	64.9 (11.0)
Social capital, mean (SD)[c]	67.9 (12.0)

[a]Income reported in Korean won (₩1=US $0.00070) and categorized into ranges.

[b]Digital literacy: the sum of access, competency, and utilization, with respective weights of 0.2, 0.4, and 0.4.

[c]All scores presented on a 100-point scale.

The mean digital literacy score was 56.2 (SD 10.5), with the subcomponents showing the following mean scores: 62.4 (SD 16.6) for digital access; 47.2 (SD 16.5) for digital competency; and 62.2 (SD 6.2) for digital utilization. The mean life satisfaction score was 64.9 (SD 11.0), and the mean social capital score was 67.9 (SD 12.0).

Correlation Analysis Between Life Satisfaction, Digital Literacy, and Social Capital

The results of the Pearson correlation analyses, conducted to examine the correlation between digital literacy, social capital, and life satisfaction, are presented in Table 2. Life satisfaction showed a statistically significant positive correlation with both digital literacy (r=0.276, P<.001) and social capital (r=0.438, P<.001), with a stronger correlation observed between life satisfaction and social capital. This indicates that higher levels

XSL•FO
RenderX

of digital literacy and social capital are associated with greater life satisfaction. Additionally, digital literacy showed a significant positive correlation with social capital (r=0.216, P<.001), indicating that as digital literacy increases, social capital tends to increase as well.

Table . Correlation analysis of life satisfaction, digital literacy, and social capital.

Variables	Life satisfaction	Digital literacy	Social capital
Life satisfaction			
r	—[a]	0.276	0.438
P value	—	*P*<.001	*P*<.001
Digital literacy			
r	0.276	—	0.216
P value	*P*<.001	—	*P*<.001
Social capital			
r	0.438	0.216	—
P value	*P*<.001	*P*<.001	—

[a]Not applicable.

The Mediating Effect of Social Capital on the Association Between Digital Literacy and Life Satisfaction

Following the procedure proposed by Baron and Kenny [29], a 3-step multiple regression analysis was conducted to examine the effects of digital literacy on life satisfaction and the mediating role of social capital (Table 3), controlling for gender, age, education level, disability status, and income level.

XSL•FO
RenderX

Table . The mediating effect of social capital on the association between digital literacy and life satisfaction.[a]

Variables	Model 1[b]		Model 2[c]		Model 3[d]	
	β (SE)	*t* (*P*[e])	β (SE)	*t* (*P*[e])	β (SE)	*t* (*P*[e])
Gender						
Female (reference: male)	1.253 (0.818)	1.531 (.13)	−0.276 (0.736)	−0.375 (.71)	−0.699 (0.684)	−1.022 (.31)
Age (years)						
70 - 74 (reference: 65 - 69)	−1.324 (0.934)	−1.417 (.16)	−1.342 (0.840)	−1.597 (.11)	−0.896 (0.780)	1.148 (.25)
75 - 79 (reference: 65 - 69)	0.172 (1.171)	0.147 (.88)	0.138 (1.053)	0.131 (.90)	0.080 (0.977)	0.082 (.93)
≥80 (reference: 65 - 69)	−0.477 (1.940)	−0.246 (.81)	−0.287 (1.745)	−0.164 (.87)	−0.1255 (1.619)	−0.078 (.94)
Education level						
Middle school (reference: ≤ elementary school)	3.213 (1.173)	2.738 (.006)	0.745 (1.055)	0.706 (.48)	−0.339 (0.983)	−0.345 (.73)
High school (reference: ≤ elementary school)	4.244 (1.285)	3.302 (.001)	1.686 (1.156)	1.459 (.15)	0.255 (1.079)	0.236 (.81)
≥ College (reference: ≤ elementary school)	5.206 (2.084)	2.497 (.01)	5.399 (1.875)	2.879 (.004)	3.643 (1.745)	2.087 (.04)
Disability						
Yes (reference: no)	1.038 (3.182)	0.326 (.74)	3.403 (2.862)	1.189 (.24)	3.053 (2.655)	1.150 (.25)
Live with someone						
Yes (reference: no)	0.059 (1.224)	0.048 (.96)	−0.363 (1.101)	−0.330 (.74)	−0.383 (1.021)	−0.375 (.71)
Income[f]						
1 - 1.99 million (reference:<1 million)	1.643 (1.655)	0.992 (.32)	2.579 (1.489)	1.732 (.08)	2.025 (1.382)	1.465 (.14)
2 - 2.99 million (reference:<1 million)	3.676 (1.775)	2.071 (.04)	5.642 (1.596)	3.534 (<.001)	4.402 (1.484)	2.965 (.003)
3 - 3.99 million (reference:<1 million)	4.765 (1.939)	2.458 (.01)	6.524 (1.744)	3.741 (<.001)	4.916 (1.623)	3.029 (.003)
≥4 million (reference:<1 million)	5.502 (1.962)	2.805 (.005)	7.371 (1.765)	4.177 (<.001)	5.515 (1.644)	3.354 (<.001)
Digital literacy	0.110 (0.046)	2.364 (.02)	0.140 (0.042)	3.350 (<.001)	0.103 (0.039)	2.647 (.008)
Social capital	—[g]	—	—	—	0.337 (0.029)	11.81 (<.001)

[a]All models were adjusted for control variables, including gender, age, education level, disability status, household type, and income level.

[b]Model 1: effect of digital literacy on social capital (constant=54.969; R^2=0.096; adjusted R^2=0.082; F_{852}=6.51; P<.001).

[c]Model 2: effect of digital literacy on life satisfaction (constant=51.755; R^2=0.133; adjusted R^2=0.119; F_{852}=9.38; P<.001).

[d]Model 3: effect of digital literacy and social capital on life satisfaction (constant=33.214; R^2=0.255; adjusted R^2=0.242; F_{851}=19.48; P<.001).

[e]P values <.05 are considered significant.

[f]Income reported in Korean won (₩1=US $0.00070) and categorized into ranges.

[g]Not applicable.

XSL•FO
RenderX

Model 1 analyzed the effect of digital literacy on social capital. Digital literacy showed a statistically significant positive effect on social capital (β=0.110; *P*=.02), indicating that higher levels of digital literacy tend to increase social capital. Moreover, higher education and income levels were positively associated with social capital.

Model 2 examined the effect of digital literacy on life satisfaction. Digital literacy had a statistically significant positive effect on life satisfaction (β=0.140; *P*<.001). Higher life satisfaction was observed among those with a college degree or higher (β=5.399; *P*=.004) and those with a monthly income of ₩2 - ₩2.99 million (β=5.642; *P*<.001), ₩3 - ₩3.99 million (β=6.524; *P*<.001), and ₩4 million or more (β=7.371; *P*<.001).

Model 3 analyzed the effects of digital literacy and social capital on life satisfaction. Both digital literacy (β=0.103; *P*=.008) and social capital (β=0.337; *P*<.001) had statistically significant positive effects on life satisfaction, indicating that both factors contribute to enhancing life satisfaction. Additionally, higher life satisfaction was observed among those with a college degree or higher (β=3.643; *P*=.04) and those with a monthly income of ₩2 - ₩2.99 million (β=4.402; *P*=.003), ₩3 - ₩3.99 million (β=4.916; *P*=.003), and ₩4 million or more (β=5.515; *P*<.001).

The *F* and *P* values for each model (models 1, 2, and 3) were statistically significant (*P*<.001), indicating that the models fit the data well. Notably, model 3 demonstrated the highest explanatory power, with an adjusted R^2 of 0.242, when social capital was included.

The regression coefficient of the mediating variable, social capital, was statistically significant, and the direct effect of the independent variable, digital literacy, on the dependent variable, life satisfaction (β=0.140), decreased when social capital was included as a mediator (β=0.103). This confirms that social capital had a partial mediating effect on the relationship between digital literacy and life satisfaction (Figure 1). Thus, social capital played a key role in enhancing life satisfaction, and digital literacy positively influenced life satisfaction through social capital.

Figure 1. Association between digital literacy, social capital, and life satisfaction.

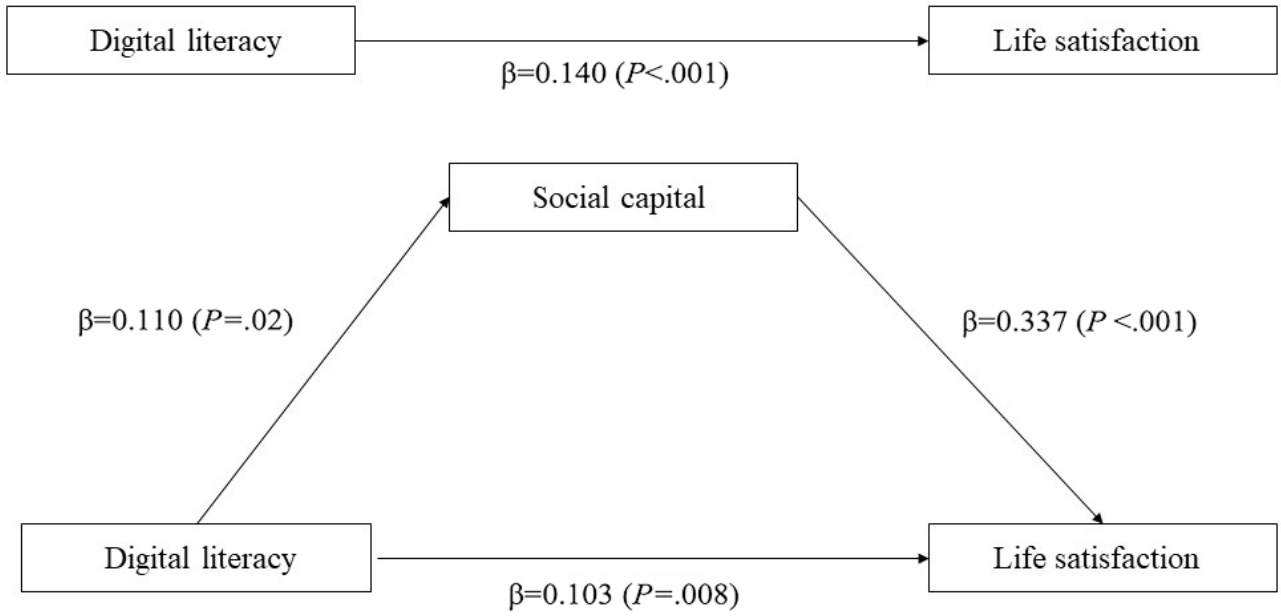

Verification of the Significance of the Mediating Effect of Social Capital

Table 4 presents the results of the bootstrapping analysis performed to verify the statistical significance of the mediating effect of social capital. First, the direct effect of digital literacy on life satisfaction, without mediation by social capital, was 0.103 (*P*=.006), confirming that digital literacy had a significant direct effect on life satisfaction. The indirect effect, representing the effect of digital literacy on life satisfaction through social capital, was 0.037 (*P*=.03). The total effect, which combines the direct and indirect effects, was 0.140 (*P*=.002), indicating that the overall impact of digital literacy on life satisfaction was statistically significant. Thus, digital literacy was an important factor that positively influenced life satisfaction, both directly and indirectly, through the mediating role of social capital.

XSL•FO
RenderX

Table . Verification of the significance of the mediating effect of social capital.[a]

Mediation pathway	β (95% CI)	*P* value[b]
Indirect effect		
Digital literacy → Social capital → Life satisfaction	0.037 (0.005-0.070)	.03
Direct effect		
Digital literacy → Life satisfaction	0.103 (0.030-0.180)	.006
Total effect	0.140 (0.058-0.230)	.002

[a]Results are based on bootstrapping simulations (n=5000).

[b]*P* values <.05 are considered significant.

Discussion

Principal Findings

This study analyzed the effect of digital literacy on life satisfaction among older adults in Korea and investigated the mediating role of social capital in this relationship. The results showed that digital literacy has a positive effect on life satisfaction. These findings align with prior research indicating that higher levels of digital access, competency, and utilization are associated with greater life satisfaction [16,18,30]. Studies on older adults have highlighted the internet's role as a valuable tool in daily life, supporting interpersonal communication, information searches, task completion, and leisure activities, all of which contribute positively to life satisfaction [31,32]. These findings underscore the internet as a crucial resource that empowers older adults to lead independent and active lives. This study reinforces these findings by demonstrating that a higher level of digital literacy positively influences life satisfaction among older adults in Korea.

The COVID-19 pandemic accelerated the digitalization of public services in Korea, significantly increasing reliance on digital platforms for essential services, including health care, administration, and finance. Although Korea is known for its advanced digital infrastructure, older adults face new challenges in accessing these essential services. This often leads to delays in medical care and limited access to health information, increasing the risk of adverse health outcomes [33]. Moreover, the rapid expansion of financial digitalization has exposed older adults with low digital literacy and limited decision-making abilities to economic risks, such as financial fraud [34]. According to a report comparing OECD countries, Korea's younger population (aged 25 - 34 years) demonstrates digital proficiency levels higher than the OECD average, whereas the older population aged 55 - 65 years exhibits proficiency levels below the OECD average. Consequently, Korea is identified as having one of the largest generational digital proficiency gaps among OECD countries [35].

Public and private organizations in Korea have provided digital literacy education for older adults, including training in computer and smartphone usage. However, these programs have predominantly taken the form of one-to-many classroom-style lectures focused on basic skills, which may not sufficiently address the diverse needs of older adults [36]. To better meet these needs, it is recommended to first assess digital literacy levels to determine older adults' specific digital proficiency levels and then provide tailored training based on these levels. For beginners, training on basic smartphone usage and sending text messages would be appropriate, while intermediate and advanced courses could cover digital health management, internet-based financial education, and the creation of digital photo albums.

Such tailored education can also be observed in international examples. The United Kingdom's Mi WiFi project (2017 - 2018) provided smart devices and digital education to vulnerable groups, including older adults and individuals with disabilities, helping them acquire basic digital skills. Similarly, the United States' Cyber Seniors project involved young mentors from volunteer organizations visiting older adults in their homes or facilities to provide one-to-one digital education, significantly enhancing their digital competency [37]. The older adult population differs widely in terms of income, education level, and health status. Therefore, digital education programs should be designed to reflect the diverse characteristics and needs of these groups.

In this study, social capital was found to have a positive impact on older adults' life satisfaction, consistent with previous studies' findings [38-41]. As social networks shrink and participation opportunities diminish with age, maintaining or acquiring social capital becomes increasingly challenging [42]. Social capital is often divided into 2 dimensions: bonding social capital and bridging social capital [43]. Bonding social capital refers to emotional support, access to resources, and a sense of solidarity formed within close relationships, such as family, friends, and neighbors. When older adults have networks from which they can obtain resources or receive emotional support, their life satisfaction significantly improves. Meanwhile, bridging social capital is formed through interactions with a wider community and diverse individuals, which helps reduce social isolation and contributes to higher life satisfaction among older adults.

These dimensions of social capital are shaped by components such as networks, social participation, norms, and social trust. Among these, social trust in family, friends, neighbors, and institutions was identified as the most important factor affecting life satisfaction in older adults [44]. By strengthening relationships with family and the local community, social trust contributes to psychological stability and enhances the quality of life for older adults. This highlights the need for

XSL•FO
RenderX

community-based support environments that consider the unique characteristics of older adults, enabling them to build and maintain social capital. Communities should provide programs and spaces that allow older adults to share their thoughts and experiences, fostering meaningful relationships. Such environments are essential to reinforcing social capital and improving the life satisfaction of older adults.

In this study, social capital showed a partial mediating effect on the relationship between digital literacy and life satisfaction among older adults. This implies that digital literacy not only directly enhances life satisfaction but also indirectly influences it through social capital. These findings are consistent with those of previous research showing that social capital mediates the relationship between digital utilization and life satisfaction among individuals aged 55 years and older [45]. Older adults with higher levels of digital literacy tend to build social relationships and participate in voluntary social activities through internet-based platforms, leading to improvements in their life satisfaction [46-48]. For example, Yoon et al [49] demonstrated that during the COVID-19 pandemic, a pilot program using digital platforms for social prescriptions effectively expanded the social networks of Korean older adults and improved their life satisfaction. South Korean society continues to emphasize family-centered values, with older adults particularly prioritizing their relationships with family members. Maintaining connections with family and friends via social networking services or remote communication positively impacts psychological well-being and overall quality of life [50,51].

Recently, some local governments in Korea have proposed tailored support measures for older adults, categorized based on their digital competency levels and specific needs. For instance, support programs have been planned for groups requiring digital tools for daily life, those seeking social interactions, and those pursuing self-development and cultural enrichment [52]. These programs aim to enable older adults to effectively utilize digital tools, foster social interactions, and mitigate social isolation. Ultimately, it is essential to implement practical support measures that expand social capital in a digital environment and enable older adults to lead more independent and fulfilling lives.

Limitations and Future Directions

This study has the following limitations. First, as a cross-sectional study, it cannot track changes in causal relationships between variables over time. Future studies should utilize longitudinal data to better understand how relationships between variables change over time. Second, this study focuses on older adults in Korea, which limits the generalizability of the results to other countries or groups with different cultural backgrounds. The relationship between digital literacy and social capital formation may differ across countries, highlighting the need for international comparative studies. Third, most of the study participants (855/869, 98.4%) were not disabled and belonged to the upper-middle-income group, which may restrict the generalizability of the findings. Therefore, the results should be interpreted with caution.

Nevertheless, this study is significant as it used nationwide raw data from the Korean MSIT, analyzing data representative of the older adult population in Korea. It emphasizes the importance of improving digital literacy and expanding social capital to enhance the quality of life of older adults in Korea.

Additionally, future research should explore the distinct impacts of the components of digital literacy—digital access, competency, and utilization—in greater detail. Particularly, examining specific aspects of digital literacy, such as media utilization skills, online safety awareness, and information-seeking behaviors, could help identify their relative effects on life satisfaction among older adults. This analysis would provide critical insights for effectively reducing the digital divide among older adults and designing tailored digital education and support policies to meet their specific needs.

Conclusion

This study analyzed the associations between digital literacy, social capital, and life satisfaction among older adults in Korea. The results showed significant correlations among digital literacy, social capital, and life satisfaction, with social capital mediating the relationship between digital literacy and life satisfaction. These findings underscore that digital literacy, when combined with social capital, contributes to enhancing the life satisfaction of older adults. Therefore, improving digital literacy and expanding social capital are crucial for promoting the psychological well-being and quality of life of older adults. Furthermore, strengthening digital literacy in the rapidly aging society of Korea can not only improve the quality of life for older adults but also help bridge generational gaps and foster greater social cohesion.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Ministry of Education (2024S1A5A2A03038024).

Data Availability

The datasets generated and analyzed during this study were collected by the Ministry of Science and Information and Communication Technology (MSIT) and the Korean National Information Society Agency and are available on the agency's website [5].

XSL•FO
RenderX

Conflicts of Interest

None declared.

References

1. 2024 elderly statistics press release [Web page in Korean]. Statistics Korea. 2024. URL: https://kostat.go.kr/board.es?mid=a10301010000&bid=10820&act=view&list_no=432917 [accessed 2025-01-22]
2. The 5th National Health Plan (Health Plan 2030, 2021-2030) [Web page in Korean]. Ministry of Health and Welfare. 2021. URL: https://www.mohw.go.kr/board.es?mid=a10401000000&bid=0008&tag=&act=view&list_no=365340 [accessed 2025-01-22]
3. Ageing in a digital world – from vulnerable to valuable. : ITU Publications; 2021 URL: https://www.unapcict.org/sites/default/files/2021-05/Ageing%20in%20a%20digital%20world%20%20from%20vulnerable%20to%20valuable.pdf [accessed 2025-01-22]
4. Hou F, Cao L, Wang Q. Call for the ageing-inclusive digital economy and related standards: based on a global survey. Presented at: ICEME 2022; Jul 16-18, 2022; Beijing, China. [doi: 10.1145/3556089.3556175]
5. The 2023 Report on the Digital Divide [Report in Korean]. : The Ministry of Science and ICT; 2024 URL: https://www.nia.or.kr/site/nia_kor/ex/bbs/List.do?cbIdx=81623 [accessed 2025-01-27]
6. Haimi M, Sergienko R. Adoption and use of telemedicine and digital health services among older adults in light of the COVID-19 pandemic: repeated cross-sectional analysis. JMIR Aging 2024 Apr 24;7:e52317. [doi: 10.2196/52317] [Medline: 38656768]
7. Mace RA, Mattos MK, Vranceanu AM. Older adults can use technology: why healthcare professionals must overcome ageism in digital health. Transl Behav Med 2022 Dec 30;12(12):1102-1105. [doi: 10.1093/tbm/ibac070] [Medline: 36073770]
8. Choi EY, Kim Y, Chipalo E, Lee HY. Does perceived ageism widen the digital divide? And does it vary by gender? Gerontologist 2020 Sep 15;60(7):1213-1223. [doi: 10.1093/geront/gnaa066] [Medline: 32539099]
9. Mubarak F, Suomi R. Elderly forgotten? Digital exclusion in the information age and the rising grey digital divide. Inquiry 2022;59:469580221096272. [doi: 10.1177/00469580221096272] [Medline: 35471138]
10. Bawden D. Origins and concepts of digital literacy. In: Lankshear C, Knobel M, editors. Digital Literacies: Concepts, Policies and Practices: Peter Lang Publishing; 2008:17-32.
11. van Dijk JAGM. The Deepening Divide: Inequality in the Information Society: Sage; 2005. [doi: 10.4135/9781452229812]
12. Hasan H, Linger H. Enhancing the wellbeing of the elderly: social use of digital technologies in aged care. Educ Gerontol 2016 Oct;42(11):749-757. [doi: 10.1080/03601277.2016.1205425]
13. Zheng R, Spears J, Luptak M, Wilby F. Understanding older adults' perceptions of internet use: an exploratory factor analysis. Educ Gerontol 2015 Mar 9;41(7):504-518. [doi: 10.1080/03601277.2014.1003495]
14. Fowler JG, Gentry JW, Reisenwitz TH. Analyzing Chinese older people's quality of life through their use of the internet. Int J Consumer Studies 2015 Apr;39(4):324-334. [doi: 10.1111/ijcs.12194]
15. Welch V, Ghogomu ET, Barbeau VI, et al. Digital interventions to reduce social isolation and loneliness in older adults: an evidence and gap map. Campbell Syst Rev 2023 Dec;19(4):e1369. [doi: 10.1002/cl2.1369] [Medline: 38024780]
16. Kim SK, Shin HR, Kim YS. The effect of mobile device capability of middle and older aged adults on life satisfaction: focusing on the mediating effect of mobile social participation [Article in Korean]. J Digit Convergence 2020;18(3):23-34. [doi: 10.14400/JDC.2020.18.3.023]
17. Lee SH, Han M, Lee J, Kim S. The effect of digital literacy in the elderly on life satisfaction: serial multiple mediating effects of social participation activities and depression [Article in Korean]. KAIS 2023;24(6):406-416. [doi: 10.5762/KAIS.2023.24.6.406]
18. Lee H. Analysis of the impact of digital literacy on life satisfaction (2019-2022) for older adults in South Korea: a national community-based panel study. Sci Rep 2024 Sep 2;14(1):20399. [doi: 10.1038/s41598-024-71397-0] [Medline: 39223233]
19. Morris ME, Adair B, Ozanne E, et al. Smart technologies to enhance social connectedness in older people who live at home. Australas J Ageing 2014 Sep;33(3):142-152. [doi: 10.1111/ajag.12154] [Medline: 24730370]
20. Khosravi P, Rezvani A, Wiewiora A. The impact of technology on older adults' social isolation. Comput Human Behav 2016 Oct;63:594-603. [doi: 10.1016/j.chb.2016.05.092]
21. Coleman JS. Social capital in the creation of human capital. American Journal of Sociology 1988 Jan;94:S95-S120. [doi: 10.1086/228943]
22. Putnam RD. The prosperous community: social capital and public life. Am Prospect 1993;13:35-42 [FREE Full text]
23. Pénard T, Poussing N. Internet use and social capital: the strength of virtual ties. J Econ Issues 2010 Sep;44(3):569-595. [doi: 10.2753/JEI0021-3624440301]
24. Mason M, Sinclair D, Berry C. Nudge or Compel? Can Behavioural Economics Tackle the Digital Exclusion of Older People?: International Longevity Centre; 2012.
25. Nyqvist F, Forsman AK, Giuntoli G, Cattan M. Social capital as a resource for mental well-being in older people: a systematic review. Aging Ment Health 2013;17(4):394-410. [doi: 10.1080/13607863.2012.742490] [Medline: 23186534]

XSL•FO
RenderX

26. Lu N, Wu B. Perceived neighborhood environment, social capital and life satisfaction among older adults in Shanghai, China. Sci Rep 2022 Apr 23;12(1):6686. [doi: 10.1038/s41598-022-10742-7] [Medline: 35461348]
27. Larnyo E, Tettegah S, Griffin B, et al. Effect of social capital, social support and social network formation on the quality of life of American adults during COVID-19. Sci Rep 2024 Feb 1;14(1):2647. [doi: 10.1038/s41598-024-52820-y] [Medline: 38302613]
28. Williams D. On and off the 'net: scales for social capital in an online era. J Comp Mediated Comm 2006 Jan;11(2):593-628. [doi: 10.1111/j.1083-6101.2006.00029.x]
29. Baron RM, Kenny DA. The moderator–mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. J Pers Soc Psychol 1986;51(6):1173-1182. [doi: 10.1037/0022-3514.51.6.1173]
30. Hwang HJ, Hwang YS. Gaps and reasons of digital divide within the elderly in Korea: focusing on household composition [Article in Korean]. Journal of Social Sciences 2017 Sep 30;24:359-386. [doi: 10.46415/jss.2017.09.24.3.359]
31. Seifert A, Schelling HR. Seniors online: attitudes toward the internet and coping with everyday life. J Appl Gerontol 2018 Jan;37(1):99-109. [doi: 10.1177/0733464816669805] [Medline: 27655237]
32. Lifshitz R, Nimrod G, Bachner YG. Internet use and well-being in later life: a functional approach. Aging Ment Health 2018 Jan;22(1):85-91. [doi: 10.1080/13607863.2016.1232370] [Medline: 27657190]
33. Han MH. Effects of digital information competency on life satisfaction among the elderly in the era of COVID-19 [Article in Korean]. KAIS 2023;24(2):559-567. [doi: 10.5762/KAIS.2023.24.2.559]
34. Kim J. Digital divide among older consumers and measures to enhance their digital literacy [Article in Korean]. J Consum Policy Stud 2022;53:27-54. [doi: 10.15723/jcps.53.3.202212.27]
35. Skills Matter: Additional Results from the Survey of Adult Skills: OECD Publishing; 2019. [doi: 10.1787/1f029d8f-en]
36. Kim SW, Lee SS. A study of the current status of older adults' digital literacy programs in public libraries and improvement plans [Article in Korean]. Korean Soc Inform Manag 2022;39(4):49-74. [doi: 10.3743/KOSIM.2022.39.4.049]
37. A study on the human rights situation of the elderly due to the digital divide [Report in Korean]. : National Human Rights Commission of Korea; 2022 URL: https://www.humanrights.go.kr/base/board/read?boardManagementNo=17&boardNo=7608724&menuLevel=3&menuNo=115 [accessed 2025-01-27]
38. Chen Y, Li W, Zhang X, Cheng H, Tian Y, Yang H. Association between social capital and quality of life in older adults with subjective cognitive decline: a cross-sectional study. Appl Nurs Res 2024 Feb;75:151773. [doi: 10.1016/j.apnr.2024.151773] [Medline: 38490797]
39. Kim BJ, Linton KF, Lum W. Social capital and life satisfaction among Chinese and Korean elderly immigrants. J Soc Work (Lond) 2015 Jan;15(1):87-100. [doi: 10.1177/1468017313504699]
40. Kim J, Kim SE. Social capital and life satisfaction of elderly who lives alone: focusing on regional difference in depression and public pension [Article in Korean]. Journal of Social Science 2018 Jul 30;29(3):267-284. [doi: 10.16881/jss.2018.07.29.3.267]
41. Kim MY, Jeon HJ. The effects of smartphone use on life satisfaction in older adults: the mediating role of participation in social activities [Article in Korean]. Korean J Gerontol Soc Welfare 2017;72(3):343-370. [doi: 10.21194/kjgsw.72.3.201709.343]
42. Jeon SN. The effects of elderly's depression and social capital on successful aging [Article in Korean]. JKSSCHE 2019 Dec 31;20(3):29-42. [doi: 10.35133/kssche.20191231.03]
43. Chung S, Sung MH. Relationship between social capital and life satisfaction: a comparison of three different age groups [Article in Korean]. Health Soc Welf Rev 2012 Dec;32(4):249-272. [doi: 10.15709/hswr.2012.32.4.249]
44. Putnam RD. Bowling alone: the collapse and revival of American community. Presented at: CSCW00: Computer Supported Cooperative Work; Dec 2-6, 2000; Philadelphia, PA, United States. [doi: 10.1145/358916.361990]
45. Yoon HJ, Shin HR, Kim YS. A study on the relationship between level of digital information usage and life satisfaction: the mediating effect of the social capital [Article in Korean]. Info Policy 2020;27(4):85-100. [doi: 10.22693/NIAIP.2020.27.4.085]
46. Lee BJ, Myung SH. The impact of perception of informatization and internet use on the formation of social networks among the elderly [Article in Korean]. J Korean Assoc Reg Inf Soc 2010;13(4):151-179. [doi: 10.22896/karis.2010.13.4.007]
47. Delello JA, McWhorter RR. Reducing the digital divide: connecting older adults to iPad technology. J Appl Gerontol 2017 Jan;36(1):3-28. [doi: 10.1177/0733464815589985] [Medline: 26084479]
48. Francis J, Kadylak T, Cotten SR, Rikard RV. When it comes to depression, ICT use matters: a longitudinal analysis of the effect of ICT use and mattering on depression among older adults. In: Stephanidis C, editor. HCI International 2016 – Posters' Extended Abstracts: Springer; 2016:301-306. [doi: 10.1007/978-3-319-40542-1_49]
49. Yoon HJ, Kim HK, Lim JA, et al. Evaluation of the effectiveness of a social prescribing pilot project for the elderly using a digital platform in the COVID-19 pandemic [Article in Korean]. Korean Public Health Res 2022;48(2):1-17. [doi: 10.22900/kphr.2022.48.2.001]
50. Kim MY. The effects of smartphone use on life satisfaction, depression, social activity, and social support of older adults [Article in Korean]. J Korea Acad Ind Coop Soc 2018;19(11):264-277. [doi: 10.5762/KAIS.2018.19.11.264]
51. Kwon MO, Lee KT. The effects of self-disclosure on online social capital and subjective well-being in online network service [Article in Korean]. J Korean Manag Assoc 2015;28(5):1485-1503 [FREE Full text]

XSL•FO
RenderX

52. Kim CN, Jo JH, Lee HS, Lee HS, Yoon MH, Lee MY. A study on strategies to enhance digital literacy among the elderly in Gyeonggi province [Report in Korean]. : Gyonggi Welfare Foundation; 2022 2022-04 URL: https://ggwf.gg.go.kr [accessed 2025-01-22]

Abbreviations

OECD: Organisation for Economic Co-operation and Development
MSIT: Ministry of Science and Information and Communications Technology
ISCS: Internet Social Capital Scale

Edited by A Mavragani; submitted 02.11.24; peer-reviewed by C Nnajiofor, M Gasmi; revised version received 23.12.24; accepted 24.12.24; published 06.02.25.

Please cite as:
Jung H, Lee H, Nam EW
Mediating Effect of Social Capital on the Association Between Digital Literacy and Life Satisfaction Among Older Adults in South Korea: Cross-Sectional Study
JMIR Form Res 2025;9:e68163
URL: https://formative.jmir.org/2025/1/e68163
doi:10.2196/68163

© Hyein Jung, Hocheol Lee, Eun Woo Nam. Originally published in JMIR Formative Research (https://formative.jmir.org), 6.2.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Addressing the "Black Hole" of Low Back Pain Care With Clinical Decision Support: User-Centered Design and Initial Usability Study

Robert S Rudin[1], PhD; Patricia M Herman[2], ND, PhD; Robert Vining[3], DC, DHSc

[1]RAND, 20 Park Plaza, Suite 910, Boston, MA, United States

[2]RAND, Santa Monica, CA, United States

[3]Palmer Center for Chiropractic Research, Palmer College of Chiropractic, Davenport, IA, United States

Corresponding Author:
Robert S Rudin, PhD
RAND, 20 Park Plaza, Suite 910, Boston, MA, United States

Abstract

Background: Low back pain (LBP) is a highly prevalent problem causing substantial personal and societal burden. Although there are specific types of LBP, each with evidence-based treatment recommendations, most patients receive a nonspecific diagnosis that does not facilitate evidence-based and individualized care.

Objectives: We designed, developed, and initially tested the usability of a LBP diagnosis and treatment decision support tool based on the available evidence for use by clinicians who treat LBP, with an initial focus on chiropractic care.

Methods: Our 3-step user-centered design approach consisted of identifying clinical requirements through the analysis of evidence reviews, iteratively identifying task-based user requirements and developing a working web-based prototype, and evaluating usability through scenario-based interviews and the System Usability Scale.

Results: The 5 participating users had an average of 18.5 years of practicing chiropractic medicine. Clinical requirements included 44 patient interview and examination items. Of these, 13 interview items were enabled for all patients and 13 were enabled conditional on other input items. One examination item was enabled for all patients and 16 were enabled conditional on other items. One item was a synthesis of interview and examination items. These items provided evidence of 12 possible working diagnoses of which 3 were macrodiagnoses and 9 were microdiagnoses. Each diagnosis had relevant treatment recommendations and corresponding patient educational materials. User requirements focused on tasks related to inputting data, and reviewing and selecting working diagnoses, treatments, and patient education. User input led to key refinements in the design, such as organizing the input questions by microdiagnosis, adding a patient summary screen that persists during data input and when reviewing output, adding more information buttons and graphics to input questions, and providing traceability by highlighting the input items used by the clinical logic to suggest a working diagnosis. Users believed that it would be important to have the tool accessible from within an electronic health record for adoption within their workflows. The System Usability Scale score for the prototype was 84.75 (range: 67.5 - 95), considered as the top 10th percentile. Users believed that the tool was easy to use although it would require training and practice on the clinical content to use it effectively. With such training and practice, users believed that it would improve care and shed light on the "black hole" of LBP diagnosis and treatment.

Conclusions: Our systematic process of defining clinical requirements and eliciting user requirements to inform a clinician-facing decision support tool produced a prototype application that was viewed positively and with enthusiasm by clinical users. With further planned development, this tool has the potential to guide clinical evaluation, inform more specific diagnosis, and encourage patient education and individualized treatment planning for patients with LBP through the application of evidence at the point of care.

(JMIR Form Res 2025;9:e66666) doi:10.2196/66666

KEYWORDS

low back pain; clinical decision support; user-centered design; usability; back pain; low back pain care; pain; clinical decision; societal burden; substantial; burden; evidence-based; treatment; diagnosis; support tool; clinicians; chiropractic; chiropractor; reviews; scenario-based interviews; interviews

Introduction

Low back pain (LBP) is highly prevalent globally, causing personal and societal burden, chronic disability, and substantial health care expenditures, ranging from US $62 to US $124 billion annually in the United States [1-6]. Most LBP is caused by 1 or more benign conditions that are not associated with dangerous pathology [7]. Instead, LBP is influenced by injury, inflammation, and dysfunction of myofascial tissues, joints, and

XSL•FO
RenderX

nerve roots or peripheral nerves, with symptoms also influenced by nonphysical factors (eg, psychological, social, and environmental) and central sensitization processes [8]. However, despite extensive evidence in the literature informing diagnoses and treatments for specific types of LBP, nonpathological LBP is typically referred to as nonspecific [7]. The nonspecific LBP label lacks details to offer a working understanding of a patient's condition or inform personalized treatment plans. Furthermore, some diagnoses for common "specific" and neurologically related conditions, such as stenosis, neurogenic claudication, radiculopathy, and sciatica, are ill-defined, overlapping, and often used incorrectly, contributing to the challenge of addressing LBP with existing clinical evidence [9]. The uncertainty that pervades LBP diagnosis and management also causes patients substantial distress [10].

To help distill the evidence for LBP and make it more widely available in clinical practice, one of the authors (RV) conducted prior work that defines an evidence-based examination for LBP in a chiropractic research setting [11]; systematically reviewed the literature of LBP diagnostic studies, which recommends standardizing diagnostic terminology for common conditions contributing to LBP [9]; and developed a revised office-based clinical examination [12] and paper-based clinical decision aid that coordinates working diagnoses with evidence-based nonpharmacological treatments offered by doctors of chiropractic [12,13].

To further the goal of supporting clinicians in applying the latest evidence for LBP in busy clinical settings, we leveraged this prior work to develop an electronic clinical decision support (CDS) tool. Our goal was to produce a web-based application to support evidence-based in-office examination, diagnosis, treatment decisions, and patient education for nonpathological conditions contributing to LBP. We used an iterative, user-centered design process to identify clinical requirements from published evidence, elicit user requirements from iterative interviews, and develop and evaluate the usability of an electronic CDS tool that facilitates and encourages evidence-based diagnosis and treatment.

Methods

Overview

We conducted a three-step user-centered design process [14] to develop the CDS tool for LBP informed by best practices [15] in developing such tools: (1) identify clinical requirements based on our prior published reviews of the evidence for LBP diagnosis and treatment; (2) iteratively identify user requirements, create wireframes, and develop a working prototype; and (3) test usability as part of a preliminary assessment (Figure 1). Steps (1) and (2) occurred in June-August 2023, and Step (3) occurred in April and May 2024. We present the results of each step. User input was received during the second and third steps from 5 chiropractors serving clinical teaching faculty roles with Palmer College of Chiropractic. Participants were recruited from the College's main campus in Davenport, Iowa (n=4) and branch campus in Port Orange, Florida (n=1). Each campus maintains teaching clinics staffed by licensed chiropractic doctors who provide care and oversee senior student clerkships. Faculty were recruited via an institutional review board (IRB)-approved email that offered information about the study, including participation requirements, and a request for a reply if interested. Those interested were sent additional information and scheduled for an interview.

Participants were targeted for recruitment as part of a purposive sampling strategy with the intention of including variation in terms of users by sex and experience. We intentionally targeted clinicians with more extensive experience in clinical practice and teaching, rather than novice users, to gain expert insights into usability and workflow issues in real-world settings and educational contexts. This approach helps ensure that the tool meets the demands and complexities of their working environments and builds credibility and acceptance among the clinical community. Users were recruited via email invitation; all invited users agreed to participate. Users verbally consented before each session. All user sessions were audio- and video-recorded.

Figure 1. User-centered design process for low back pain decision support tool. After identifying relevant clinical findings from prior systematic literature reviews, we followed a standard, iterative process to identify and refine user requirements and design solutions before conducting a test of the usability using fictitious scenarios.

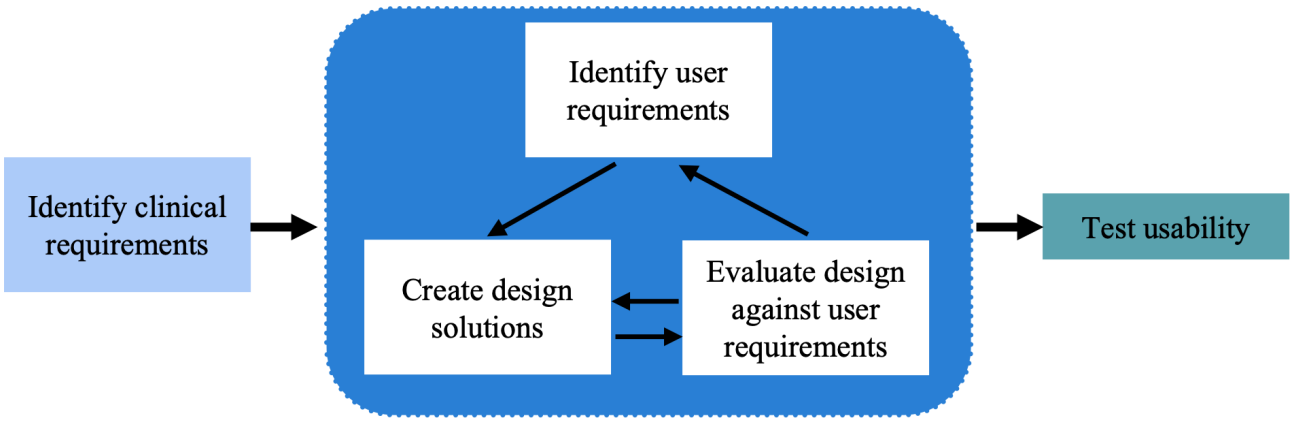

XSL•FO
RenderX

Clinical Requirements Identification

As described in the introduction, in prior work, we developed an evidence-based classification system for LBP, conducted a systematic review of LBP diagnostic studies, made recommendations for standardizing diagnostic terminology for common conditions contributing to LBP, and developed recommendations for LBP evaluation and treatment that were distilled in a series of publications including a paper-based clinical decision aid [9,11-13]. To identify clinical requirements for the CDS tool in this work, we reviewed our prior work to derive (1) clinical interview and examination inputs with adaptive characteristics that disable or enable items based on prior responses, (2) working diagnoses based on selected input items, and (3) evidence-based treatment recommendations corresponding to each working diagnosis. Furthermore, the clinician on the research team (RV) drafted patient education materials corresponding to each working diagnosis for user testing.

User Requirement Elicitation and CDS Tool Design and Development

We iteratively elicited user requirements and developed and refined our prototype through wireframing and then software programming. We developed a discussion guide covering topics that included current approaches used for diagnosing and treating LBP, workflows related to LBP, pain points, ideas for tool design, and reactions to the wireframes showing the input items and example diagnoses, treatment recommendations, and patient educational information. Two research team members with expertise in informatics and design (RSR) and LBP clinical practice and research (RV) interviewed each user. Prior to each first design session, users were sent prior published literature related to LBP diagnosis and treatment. Users were asked to review in advance as a reminder of current evidence related to LBP evaluation, diagnosis, and treatment, and to gain familiarity with key sources underlying the CDS tool. Users were encouraged to “think aloud” during the design sessions.

After each design session, 1 research team member (RSR) reviewed the recording and summarized key findings using conventional content analysis [16] and the framework method [17]. Categories of findings were related to user requirements, changes were needed to make the wireframes and prototypes address the emerging requirements, and new design ideas emerged during the sessions organized by user task. Between design sessions, we brainstormed or ideated new design ideas for the tool and worked with an experienced user interface designer to create additional wireframes. We then developed a fully functional prototype of the tool as a web-based application. A second research team member (RV) reviewed the recordings and confirmed or amended key findings. Results were finalized through consensus among research team members (RSR and RV).

Usability Test

To assess usability, 1 clinician-author (RV) developed 4 fictional clinical scenarios to help assess how the users would use the CDS tool in realistic situations. These involved a condition-specific history and examination findings for conditions matching criteria for acute nociceptive pain, chronic nociceptive pain, chronic nociceptive pain combined with symptoms of neurogenic claudication, and radicular pain. Each user was tested on 1 or 2 of these scenarios. At the start of each usability testing session, the user was given a brief overview of the tool and then asked to imagine that he or she was treating a patient with LBP. The user was then read a brief summary of a fictional patient’s LBP symptoms and was instructed to ask questions, as they would during a patient interview, to complete the input questions. The user was encouraged the “think aloud” as they used the CDS tool. After completing at least 1 scenario, the user was asked to complete the System Usability Scale (SUS), a 10-item measure that is commonly used to assess the usability [18]. SUS scores range from 0 to 100, with higher scores indicating greater usability. SUS scores >68 are considered above average, and those above 80 are in the top 10th percentile [19]. Users were encouraged to explain reasons for their answers and results were summarized [16].

Ethical Considerations

This study was reviewed and approved by the RAND Human Subjects Protection Committee (HSPC ID: 2023-N0175). The IRB at Palmer College of Chiropractic was the relying IRB through a formal reliance agreement. All written and audiovisual data (which were not anonymized) were stored on secure servers accessible only to the research team. Because the interviews did not address protected or sensitive topics, but instead focused on perceptions of a potential website and clinical workflows, no written consent was obtained. Instead, verbal consent was obtained prior to conducting interviews. Participants were compensated US $150 per interview, with a maximum of 2 interviews.

Results

User Characteristics

All users practiced chiropractic in teaching clinics at Palmer College of Chiropractic at either an Iowa or Florida campus location. See Table 1 for user characteristics. All encountered LBP on a daily basis when they treated patients. Users varied on how often they attempted to arrive at a specific diagnosis from rarely to every visit. One user recently switched to an administrative position and was no longer engaging in direct patient care.

XSL•FO
RenderX

Table . Characteristics of chiropractors at Palmer College of Chiropractic (N=5) who participated in user design sessions to assess the low back pain decision support tool.

Characteristic	Statistical values
Sex, n (%)	
Male	2 (40)
Female	3 (60)
Years practicing chiropractic, mean (range)	18.5 (7-34)
Location, n (%)	
Iowa	4 (80)
Florida	1 (20)
Patient visits per week, n	~100
Years using an EHR[a], mean (range)	10 (7 - 12)

[a]EHR: electronic health record.

Clinical Requirements

We identified 14 possible working diagnoses divided into 3 macrocategories and 9 microcategories (Figure 2). For the input items and logic rules to enable them, we identified a total of 44 items: 13 interview items are enabled for all patients, 13 interview items are enabled conditional on other input items, 1 examination item is enabled for all patients, 16 examination items are enabled conditional on other input items, and 1 item is a synthesis of interview and examination findings and enabled for all patients. Items that were enabled for all patients helped identify macrodiagnoses; items that were conditionally enabled were related to microdiagnoses. For example, if the user selected "Sensory changes in a nerve root distribution," which offers some evidence for the macrodiagnostic category of neuropathic pain, 8 interview items and 6 examination items are enabled to clarify 1 or more potential microdiagnoses.

To specify the clinical logic rules for each working diagnosis, we used Boolean logic. For example, piriformis syndrome was included as a possible diagnosis based on the following logic combining the relevant input items: "radiating pain into an ipsilateral leg" OR "tenderness of the greater sciatic notch" OR "buttock pain" OR "positive straight leg raise test" OR "increased pain with prolonged sitting." The inputs and logic represent diagnostic information or criteria consistent with piriformis syndrome. If indicated by the literature, we included estimates of the strength of the evidence based on the quantity of inputs corresponding to a working diagnosis (eg, 1-5). Anchor terms of less likely (eg, 1) and more likely (eg, 5) are used because working diagnoses cannot be completely definitive.

Treatment recommendations were assigned to each working diagnosis based on published evidence. In many cases, 1 treatment recommendation applied to multiple diagnoses. For example, graded exposure and activity training are recommended for both neurogenic claudication and nociplastic pain. Some treatment recommendations apply to all working diagnoses (eg, education about condition and general exercise when safe and tolerated).

Figure 2. Types of low back pain. The literature identifies 3 macrodiagnoses and 9 microdiagnoses. These categories served as outputs of the decision support tool for low back pain.

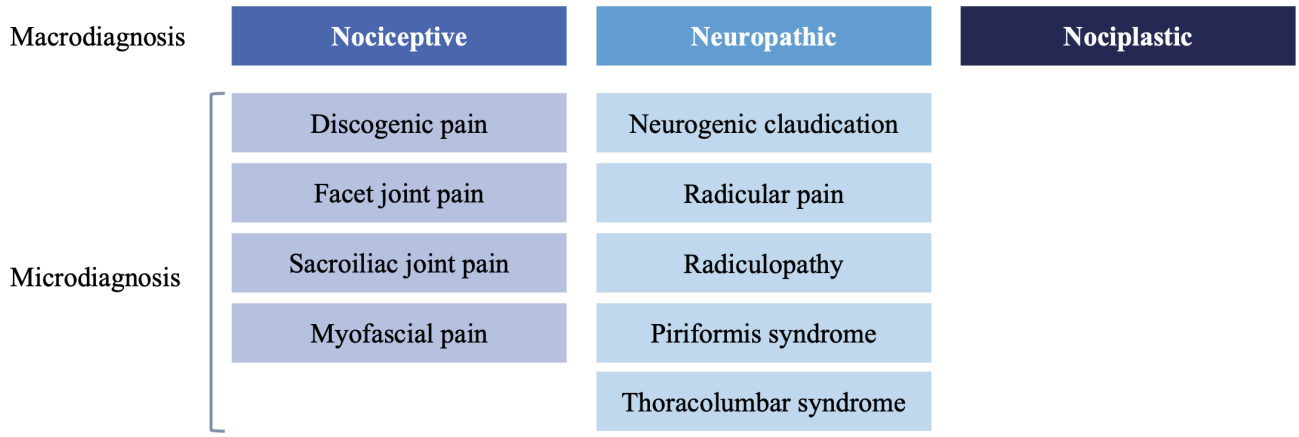

User Requirements and CDS Tool Design

Example wireframes used in design sessions are shown in Figure S1 in Multimedia Appendix 1. Users were positive about the potential value of the CDS tool: "This tells you exactly where to focus your attention on. I like it" (User 3). Users said that the primary use case for using the tool would be when a new patient presents with LBP. Some might also use it if treatment was not working for reassessing their diagnoses and treatment plan and informing the potential for a referral. All users

XSL•FO
RenderX

suggested that it would be useful both for practicing chiropractors and as an education tool.

User requirements and design features are shown in Table 2 organized by user task. User input led to key refinements in the design, such as organizing the input questions by microdiagnosis, adding a patient summary screen that persists during data input and when reviewing the output, adding more information buttons and graphics to input questions, and providing traceability by highlighting the input items used by the clinical logic to suggest a working diagnosis.

The prototype CDS tool was built as a web-based application using the React javascript library and responsive design. For this prototype, all of the data entered by the user are stored within the user's web browser—no database or server-side storage of data were used (Figure 3). Screenshots of the tool are shown in Figure 4 and Figures S2 and S3 in Multimedia Appendix 1.

XSL•FO
RenderX

Table . User requirements and design features by task for the low back pain decision support tool.

Task	User requirement	Design feature
Input interview and examination findings	Make it easy to input the relevant data.	Order interviews items before examination items to parallel clinical workflows; disable input items for microdiagnoses until they are indicated based on items for more general diagnoses; and cluster input items based on microdiagnosis.
	Allow the user to learn more about data that need to be inputted if desired.	Information buttons and graphics.
	Show a summary of the interview/examination findings as user enters them.	Patient summary section.
Review potential working diagnoses	Allow the user to see all potential working diagnoses, which reflects the multifactorial nature of LBP[a], with relative strength of evidence based on interview/examination findings.	List diagnoses with likeliness scales to enable rapid visual confirmation of the extent of evidence supporting each condition, allowing users to confirm data accuracy and determine whether the interview and examination data align with the overall clinical presentation.
	For any specific working diagnosis, allow the user to view the specific interview/examination findings that constitute the evidence for it.	When user hovers over a diagnosis, highlight the related interview/examination findings in patient summary.
	Allow the user to go back and change/alter the interview/examination findings at any time to alter the potential working diagnoses.	Navigation arrow and button to return to the input screen.
Select potential working diagnoses	Allow the user to select/deselect from the list of working diagnoses based on the user's clinical assessment.	Toggles for each diagnosis.
	Allow the user to document his or her selected working diagnoses in his or her notes.	Allow copy-paste; automated input into EHR[b,c].
Save data for next visit	Allow the user to save his or her selected working diagnoses and interview/examination findings for review in future.	Save data tied to visit in EHR or database[b].
Review treatment recommendations	Allow the user to see all evidence-based treatment recommendations for selected working diagnoses.	List treatments based on evidence-based guidelines for diagnoses.
	For any specific treatment recommendation, allow the user to see which diagnoses he or she corresponds to.	List diagnoses below each treatment recommendation.
Determine treatment recommendations	Allow the user to select/deselect and customizing treatment recommendations based on the user's clinical assessment and experience with the various options.	Toggles for each treatment recommendation[b].
Select relevant patient educational information	Allow the user to easily print the patient educational information.	Make educational information available in PDF or other printable format.
	Allow the user to select/deselect and adapt patient educational information.	Select buttons to turn on/off specific contents[b].
	Allow the user to tailor the patient educational information based on patient and user preferences.	Let user edit content for each patient and configuration option for default content[b].
Global	Allow the user to access the tool within the EHR.	Implement CDS[d] tool within patient chart in EHR[b].

[a]LBP: low back pain.

[b]Not yet implemented in prototype CDS tool.

[c]EHR: electronic health record.

[d]CDS: clinical decision support.

XSL•FO
RenderX

Figure 3. Technical architecture of web-based application for low back pain decision support. The sequence 1 - 4 shows each step that occurs as the user interacts with the tool.

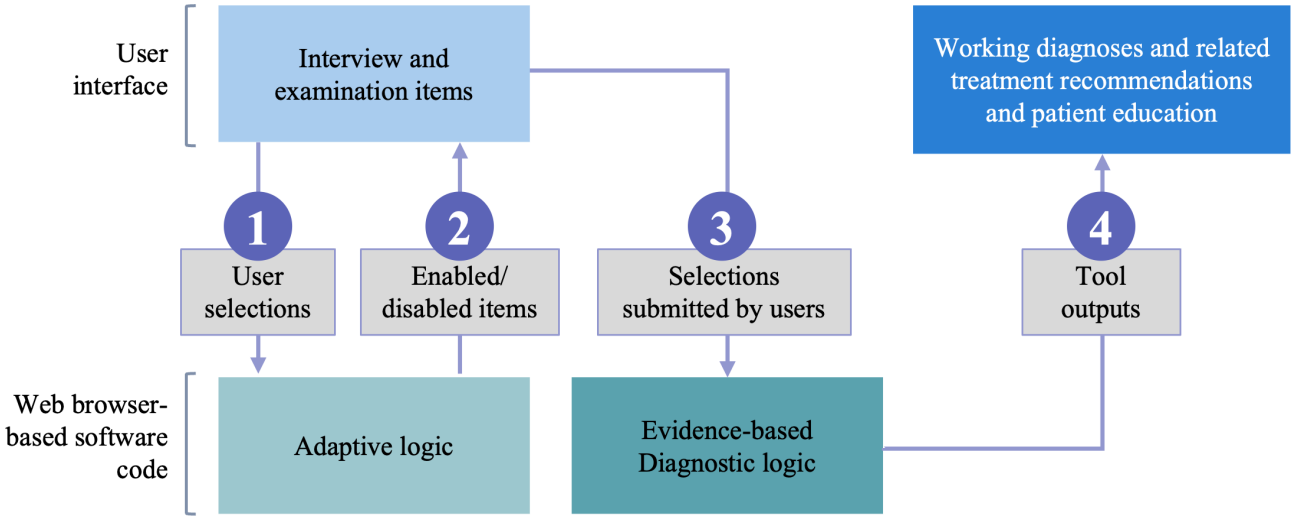

XSL•FO
RenderX

Figure 4. Complex multifactorial diagnosis with evidence for 3 macrodiagnoses of low back pain (nociceptive, neuropathic, and nociplastic pain). Screenshots of the tool show (1) user selections based on the interview and examination, with (2) items disabled according to the adaptive logic in the model. The greatest evidence suggests neurogenic claudication with overlying nociplastic pain. Piriformis syndrome was rejected by the model because symptoms overlap with neurogenic claudication and because historical and examination evidence is stronger for neurogenic claudication. The (3) summary panel to the right, shown only in the top right screenshot, shows the selections as submitted by the user. The bottom screenshots show the tool outputs (4), with the result of the evidence-based diagnostic logic displaying the working diagnoses (left) and recommendations for treatments and patient education (right).

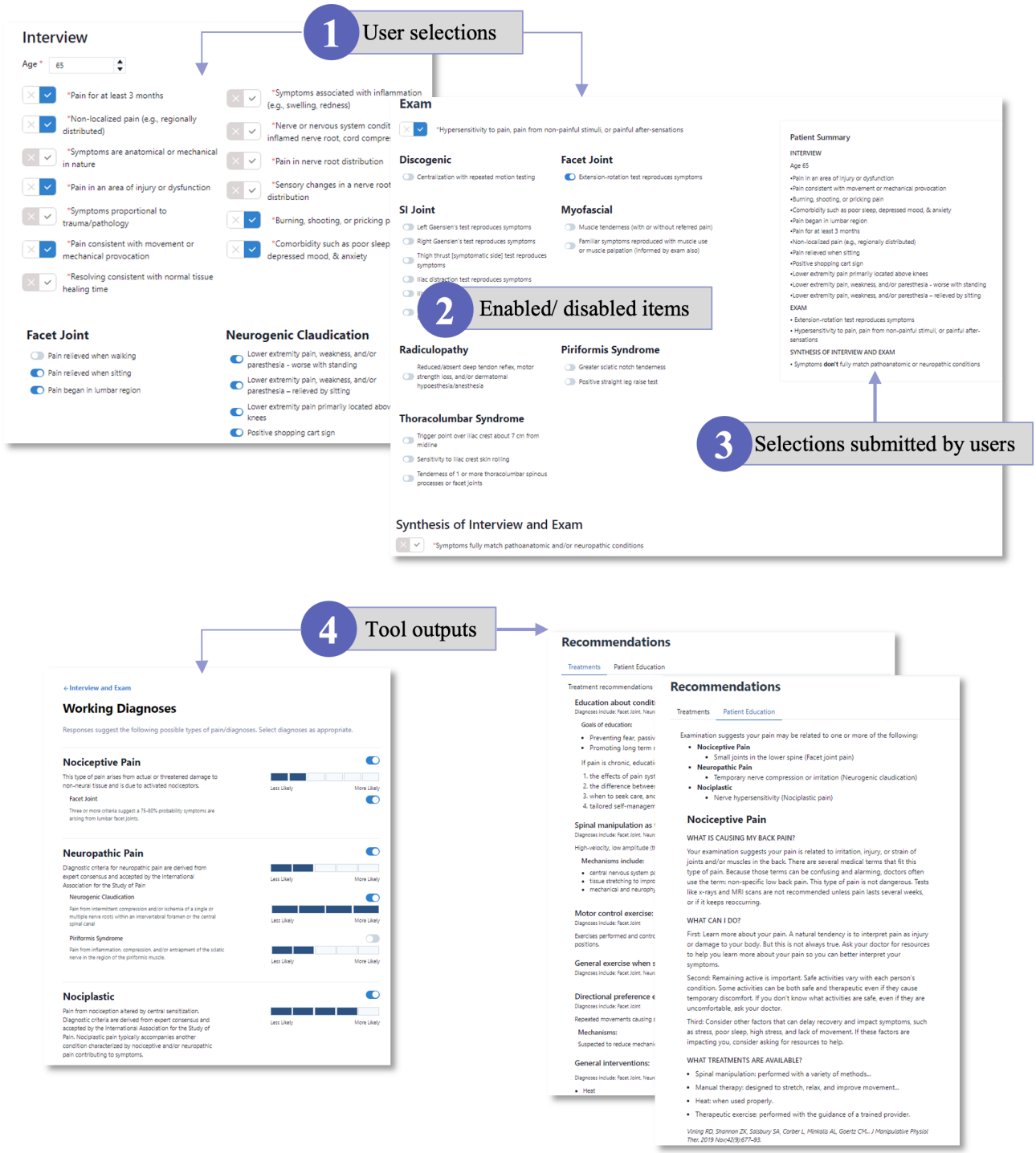

Usability Test

The mean score of the SUS was 84.75 (range 67.5 - 95). Qualitative results explained these findings (Table 3). Overall, users were enthusiastic about using the CDS tool because they perceived that it would help them focus on the most relevant information, especially for more complex patients. All users thought that the tool was easy to use from a technical perspective but required training and practice on the clinical content to use it effectively and efficiently. While using the tool during a study session, one user quickly transitioned from an initial impression that substantial effort would be needed to use it to expressing enthusiasm for the tool's potential use for improving back pain diagnosis and treatment: "This (tool) is like walking into a storm, but it wouldn't be like that the second time and certainly by the

XSL•FO
RenderX

50th time it would be fabulous" (User 4). One user said that the training needed to use the tool effectively was a good thing because that training would improve care for LBP—they strongly agreed with the statement from the SUS that said, "I needed to learn a lot of things before I could get going with this system" (User 3), but added that they meant that "in a positive connotation way."

Users also reacted to the usability of specific features. When using the input page, user reactions included the following: enabling examination items only when indicated could help eliminate tests that might be done routinely but are not clinically relevant; the patient summary was useful for guiding the input process; the wording or clinical intention behind some items was difficult to understand (eg, "symptoms fully match pathoanatomic and/or neuropathic condition;" and "resolving consistent with normal tissue healing time"); and it might be difficult to have time to go through all items during a visit.

When using the list of working diagnoses, reactions from users included positive views of the feature that highlighted relevant input items when hovering over a diagnosis to show the inputs (ie, evidence) supporting it; clarity of the scale used to indicate strength of evidence; and helpfulness of having a list of diagnoses to prevent anchoring on one and also for explaining the clinical reasoning process to patients. Reactions to the treatment recommendations were generally positive and believed to be helpful for decision-making. Two users noted that they had preferences for some treatment options more than others based on what they were familiar with.

For the patient education materials, 3 users emphasized the importance of being able to print out copies. Two believed that the materials would be acceptable to patients while 3 found them too complex especially for those with low health literacy levels. Two users preferred handwriting education information to make it more personal. One preferred to print out the materials and highlight key information.

Users also described how they would likely use this CDS tool in their workflows. All users believed that it was important to make the tool available within the EHR to allow for easy access to it and to minimize the need to switch between applications or devices. All users would use the CDS tool while with the patients but 2 users mentioned that they might leave the room to use the tool separately. One suggested that a freestanding web-based version of the tool maybe useful as well in some scenarios. Two users wanted to modify their patient intake questions to match the CDS tool's input items to help the clinician prepare for the visit and facilitate tool usage.

Users suggested several ideas to improve use and usability including use of colors on the input screen to break up the text more, autopopulating the CDS tool based on patient intake questions, providing a crosswalk between generated working diagnoses and *International Classification of Diseases, Tenth Revision* (*ICD-10*) codes to facilitate billing, indicating in the input screen if the combination input items selected did not make clinical sense and should be revisited, further emphasizing which treatments should be of higher or lower priority based on evidence (if it exists), and providing more visuals for patient education materials.

XSL•FO
RenderX

Table . Summary of qualitative usability evaluation results of the low back pain decision support tool.

Category	Selected finding	Illustrative quote
Overall impact of the CDS[a] tool on clinical care	Users believed that the tool would help users focus on the relevant information, improve diagnostic accuracy, and use of evidence-based treatments.	"I'm very likely to use this… you know why? It makes me better. It makes me smarter… This is an elegant way to approach the black hole of low back pain." (User 4)
		"It helps you focus your thoughts" (User 3)
Training required to use the CDS tool	Although the tool is simple to use technically, clinical knowledge is needed to use it effectively and will require training.	"It doesn't seem complex; it just seems like I need to familiarize myself with it." (User 1)
Input items	Enabling items adaptively can help guide information gathering.	"It is nice that it does close off those other options that maybe get in our way… I think this helps cater the treatment to the patient's diagnoses rather than allowing the provider to cater the treatment to their own preferences." (User 5)
	It might be challenging for users to go through all items during visits with a patient.	"I think it could be challenging even with familiarity… to answer it real-time [as] I'm having a conversation with somebody." (User 1)
Working diagnoses	Degree of evidence is helpful and clearly shown in bar format.	"Very visually clear." (User 4)
Treatment recommendations	The recommendations can help inform treatment for both clinician and student.	"Beyond nice, this is very educational… I love this, this is brilliant." (User 3)
Patient education	The content may be helpful to some patients but too complex for others.	"I think this is very useful especially if someone starts catastrophizing." (User 4)
		"I think it would be hugely beneficial at the report of findings to explain to the patient what the pain generator is and how that is affecting their presentation." (User 5)
		"[For] 90% of [my patient] this would be far too complex… it is far beyond a lot of our patients understanding." (User 3)

[a]CDS: clinical decision support.

Discussion

Principal Results

This study reports the initial development and evaluation of a CDS tool facilitating evidence-based diagnosis, clinical management, and patient education for LBP. The tool is web-based and can be embedded into the EHR for use during the clinical encounter as a support for interview, examination, care planning, and patient education. It can also be used independent of the patient encounter. The SUS score for the prototype was 84.75 (range: 67.5 - 95), which is in the top 10th percentile. Users believed that the CDS tool was easy to use, although training in clinical content and practice was considered necessary for effective and efficient use. With such training and practice, users believed that it would improve care and shed light on the "black hole" of LBP diagnoses and treatment.

The CDS tool inputs included 44 patient interview and examination items that provided evidence of 12 possible working diagnoses. Input items were ordered to begin with interview items followed by examination items—conditionally displayed if relevant—so as to parallel clinical workflows. The output user interface allows multiple potential working diagnoses, reflecting the multifactorial nature of LBP. Likeliness scales enable rapid visual assessment of the extent of evidence supporting each condition. This feature enables clinicians to confirm data accuracy and determine whether the interview and examination data align with the overall clinical presentation. The ability to remove or reject working diagnoses for which there is little evidence in favor of others supported by stronger evidence can assist in prioritizing care. The patient education materials written at an eighth-grade level were designed to help further explain LBP, offer reassurance, list possible treatment and self-care options, describe other factors influencing pain severity (eg, stress, poor sleep, and lack of movement), and encourage patients to communicate with clinicians about these topics. Some interviews suggested that these materials had potential to be useful, although some revision may be needed. Review by, and feedback from, patients with LBP would best inform further development of this feature.

Contribution and Comparison With Other Work

To our knowledge, this is the first provider-facing tool developed to assist clinicians in the diagnosis and treatment of LBP. A prior study describes the development a CDS tool to assist Australian pharmacists in advising patients about watchful waiting for acute LBP and recommending medications and provider types to seek for persistent symptoms [20]. Our work is complementary to that work in that it focuses on offering providers a CDS tool once patients seek care. The CDS tool offers a systematic process for a clinical evaluation which, in

XSL•FO
RenderX

turn, generates working diagnoses and links these outputs with guideline-based nonpharmacological treatments offered by chiropractors and other practitioners providing similar treatments. Our tool also shares some design similarities with currently available web-based risk-screening tools, such as the Lung Cancer Screening Risk Calculator and the Fracture Risk Assessment Tool (FRAX) [21,22]. These calculators are clinician-facing applications with input variables (eg, age, sex, smoking status, and bone mineral density) for estimating the risk for future lung cancer and osteoporotic fracture.

In addition to distilling the evidence for LBP diagnosis and treatment and making it available to clinicians at the point of care, this tool has the potential to help clinicians address patient needs identified in qualitative research for LBP. Costa et al [10] describe difficulties encountered by people with LBP due to uncertainty stemming from diagnoses and treatment. Our CDS tool has the potential to support clinicians in explaining the nature of the problem, treatment plan, and what patients can do on their own to address the problem [10]. Cox et al [23] make specific recommendations to clinicians about communicating LBP diagnoses, including sharing thought processes and diagnostic reasoning and providing written information about diagnosis. Our tool addresses these recommendations [23].

The CDS tool we developed may enable clinicians to avoid common errors in clinical reasoning leading to misdiagnosis, such as anchoring and premature closure [24]. Anchoring is an attachment to a diagnosis despite contradictory evidence; premature closure occurs when a diagnosis is accepted before verification. By offering objective, evidence-based, and repeatable provisional working diagnoses based on history and examination data, with visual displays of the strength of evidence supporting working diagnoses, this tool may help clinicians more systematically and consciously consider each potential working diagnosis in the context of the strength of evidence.

Limitations and Future Work

User input was gathered from 5 practicing chiropractic clinicians who participated in both the development and usability testing phases. This sample size is sufficient for formative research and for identifying major usability issues [25-27]. While there is a potential for bias due to their involvement in the design process, the likelihood of such bias is reduced, as the design sessions took place 7 months prior to the usability testing, limiting recall of specific design details. Informed by our findings and to address limitations in our scope, we plan to build on this work in 6 areas. First, we will implement the tool within a health system's EHR to enable easy access within the clinical workflow and enhance certain features such as allowing users to save information for each patient. Second, we will iteratively refine the patient-facing content with patients, clinicians, and with consideration of evidence in literature. Third, we will expand to other clinical specialties that treat LBP. Our initial focus was on chiropractic, a profession focused on nonpharmacological diagnosis and treatment of neuro-musculoskeletal conditions and consisting of approximately 77,000 licensed practitioners in the United States [28]. There is potential to further develop the tool for physical therapy, primary care, and other specialties. Fourth, the tool may be expanded beyond the biological perspective to include other factors (eg, psychological, spiritual, environmental, and social) relevant to LBP diagnosis and management [29]. Fifth, we will examine opportunities to incorporate artificial intelligence (AI). AI-based approaches are gaining traction in diagnosis [30,31] and other areas. However, AI models are limited by the data quality used to create them. By standardizing the definition of interview and examination inputs, our tool has the potential to improve the quality of data used by AI. Finally, the CDS tool developed in this study does not screen for signs or symptoms of pathology (eg, fracture, infection, and malignancy). There is no current consensus on such screening, and findings can be interpreted differently with varied clinical histories and among patients in different age groups. However, inputs for major red flags (eg, history of cancer and pulsatile pain with abdominal bruit suggesting aortic aneurysm) can be added to prompt and facilitate the clinical screening process.

Conclusions

The LBP tool we have developed has potential to help standardize the process of clinical evaluation leading to systematically derived working diagnoses that inform individually tailored treatment decisions. Our process of defining clinical requirements based on rigorous evidence reviews and applying user-centered design principles through eliciting user requirements produced a clinician-facing decision support tool that was viewed positively and with enthusiasm by clinical users. If implemented, this tool may shed light on the "black hole" of LBP care, facilitate dissemination of LBP evidence as it is produced, and catalyze innovation in LBP care.

Acknowledgments

Research reported in this publication was supported by the RAND Technology Incubator and the National Center for Complementary and Integrative Health of the National Institutes of Health under award number U24AT012549. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors thanks Andrew Sloan for writing the software code for the tool and Alyson Youngblood for help designing the user interface.

Data Availability

The datasets generated during and/or analyzed during this study are not publicly available to protect the privacy of the research subjects.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional examples of wireframes and screenshots.
[DOCX File, 2515 KB - formative_v9i1e66666_app1.docx]

References

1. Hartvigsen J, Hancock MJ, Kongsted A, et al. What low back pain is and why we need to pay attention. Lancet 2018 Jun 9;391(10137):2356-2367. [doi: 10.1016/S0140-6736(18)30480-X] [Medline: 29573870]
2. Pitcher MH, Von Korff M, Bushnell MC, Porter L. Prevalence and profile of high-impact chronic pain in the United States. J Pain 2019 Feb;20(2):146-160. [doi: 10.1016/j.jpain.2018.07.006] [Medline: 30096445]
3. Dutmer AL, Schiphorst Preuper HR, Soer R, et al. Personal and societal impact of low back pain: the Groningen spine cohort. Spine (Phila Pa 1976) 2019 Dec 15;44(24):E1443-E1451. [doi: 10.1097/BRS.0000000000003174] [Medline: 31369481]
4. Wu A, March L, Zheng X, et al. Global low back pain prevalence and years lived with disability from 1990 to 2017: estimates from the Global Burden of Disease Study 2017. Ann Transl Med 2020 Mar;8(6):299. [doi: 10.21037/atm.2020.02.175]
5. Lo J, Chan L, Flynn S. A systematic review of the incidence, prevalence, costs, and activity and work limitations of amputation, osteoarthritis, rheumatoid arthritis, back pain, multiple sclerosis, spinal cord injury, stroke, and traumatic brain injury in the United States: a 2019 update. Arch Phys Med Rehabil 2021 Jan;102(1):115-131. [doi: 10.1016/j.apmr.2020.04.001]
6. Dieleman JL, Cao J, Chapin A, et al. US health care spending by payer and health condition, 1996-2016. JAMA 2020 Mar 3;323(9):863-884. [doi: 10.1001/jama.2020.0734] [Medline: 32125402]
7. Maher C, Underwood M, Buchbinder R. Non-specific low back pain. Lancet 2017 Feb;389(10070):736-747. [doi: 10.1016/S0140-6736(16)30970-9]
8. Gräper PJ, Scafoglieri A, Clark JR, Hallegraeff JM. Sensory profiles predict symptoms of central sensitization in low back pain: a predictive model research study. J Clin Med 2024 Aug 1;13(16):4677. [doi: 10.3390/JCM13164677/S1]
9. Vining RD, Shannon ZK, Minkalis AL, Twist EJ. Current evidence for diagnosis of common conditions causing low back pain: systematic review and standardized terminology recommendations. J Manipulative Physiol Ther 2019 Nov;42(9):651-664. [doi: 10.1016/j.jmpt.2019.08.002]
10. Costa N, Butler P, Dillon M, et al. “I felt uncertain about my whole future”-a qualitative investigation of people’s experiences of navigating uncertainty when seeking care for their low back pain. Pain 2023 Dec 1;164(12):2749-2758. [doi: 10.1097/j.pain.0000000000002975] [Medline: 37478013]
11. Vining R, Potocki E, Seidman M, Morgenthal AP. An evidence-based diagnostic classification system for low back pain. J Can Chiropr Assoc 2013 Sep;57(3):189-204. [Medline: 23997245]
12. Vining RD, Minkalis AL, Shannon ZK, Twist EJ. Development of an evidence-based practical diagnostic checklist and corresponding clinical exam for low back pain. J Manipulative Physiol Ther 2019 Nov;42(9):665-676. [doi: 10.1016/j.jmpt.2019.08.003]
13. Vining RD, Shannon ZK, Salsbury SA, Corber L, Minkalis AL, Goertz CM. Development of a clinical decision aid for chiropractic management of common conditions causing low back pain in veterans: results of a consensus process. J Manipulative Physiol Ther 2019 Nov;42(9):677-693. [doi: 10.1016/j.jmpt.2019.03.009] [Medline: 31864769]
14. ISO. ISO 9241-210:2019 —Ergonomics of human-system interaction — part 210: human-centred design for interactive systems. 2019. URL: https://www.iso.org/standard/77520.html [accessed 2024-06-25]
15. Kawamoto K, Houlihan CA, Balas EA, Lobach DF. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. BMJ 2005 Apr 2;330(7494):765. [doi: 10.1136/bmj.38398.500764.8F] [Medline: 15767266]
16. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005 Nov;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
17. Gale NK, Heath G, Cameron E, Rashid S, Redwood S. Using the framework method for the analysis of qualitative data in multi-disciplinary health research. BMC Med Res Methodol 2013;13(1):117. [doi: 10.1186/1471-2288-13-117]
18. Bangor A, Kortum PT, Miller JT. An empirical evaluation of the System Usability Scale. Int J Hum Comput Interact 2008 Jul 29;24(6):574-594. [doi: 10.1080/10447310802205776]
19. Measuring U. Measuring usability with the System Usability Scale (SUS). 2011. URL: https://measuringu.com/sus/ [accessed 2024-08-13]
20. Downie AS, Hancock M, Abdel Shaheed C, et al. An electronic clinical decision support system for the management of low back pain in community pharmacy: development and mixed methods feasibility study. JMIR Med Inform 2020 May 11;8(5):e17203. [doi: 10.2196/17203] [Medline: 32390593]

XSL•FO
RenderX

21. ScreenLC. Lung cancer screening risk calculator. URL: https://screenlc.com/dpp-vue/index.html [accessed 2024-09-03]
22. FRAX. Fracture risk assessment tool. URL: https://frax.shef.ac.uk/FRAX/tool.aspx [accessed 2024-09-03]
23. Cox C, Hatfield T, Willars J, Fritz Z. Identifying facilitators and inhibitors of shared understanding: an ethnography of diagnosis communication in acute medical settings. Health Expect 2024 Aug;27(4):e14180. [doi: 10.1111/hex.14180] [Medline: 39180375]
24. Rylander M, Guerrasio J. Heuristic errors in clinical reasoning. Clin Teach 2016 Aug;13(4):287-290. [doi: 10.1111/tct.12444] [Medline: 26395611]
25. Nielsen Norman Group. Why you only need to test with 5 users. 2000 Mar 18. URL: https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/ [accessed 2024-11-19]
26. Virzi RA. Refining the test phase of usability evaluation: how many subjects is enough? Hum Factors 1992 Aug;34(4):457-468. [doi: 10.1177/001872089203400407]
27. Shuldiner J, Kiran T, Agarwal P, et al. Developing an audit and feedback dashboard for family physicians: user-centered design process. JMIR Hum Factors 2023;10(1):e47718. [doi: 10.2196/47718]
28. NBCE. Practice analysis of chiropractic 2020—NBCE survey analysis. 2020. URL: https://www.nbce.org/practice-analysis-of-chiropractic-2020/ [accessed 2024-09-03]
29. King D. Diagnosis: Oxford University Press; 2015. [doi: 10.1093/ACREFORE/9780199381135.013.2138]
30. Fu Y, Zhou J, Li J. Diagnostic performance of ultrasound-based artificial intelligence for predicting key molecular markers in breast cancer: a systematic review and meta-analysis. PLoS ONE 2024;19(5):e0303669. [doi: 10.1371/journal.pone.0303669] [Medline: 38820391]
31. Nowroozi A, Salehi MA, Shobeiri P, et al. Artificial intelligence diagnostic accuracy in fracture detection from plain radiographs and comparing it with clinicians: a systematic review and meta-analysis. Clin Radiol 2024 Aug;79(8):579-588. [doi: 10.1016/j.crad.2024.04.009]

Abbreviations

AI: artificial intelligence
CDS: clinical decision support
FRAX: Fracture Risk Assessment Tool
LBP: low back pain
SUS: System Usability Scale

Edited by A Mavragani; submitted 19.09.24; peer-reviewed by S Kandaswamy; revised version received 02.12.24; accepted 03.12.24; published 04.02.25.

Please cite as:
Rudin RS, Herman PM, Vining R
Addressing the "Black Hole" of Low Back Pain Care With Clinical Decision Support: User-Centered Design and Initial Usability Study
JMIR Form Res 2025;9:e66666
URL: https://formative.jmir.org/2025/1/e66666
doi:10.2196/66666

© Robert S Rudin, Patricia M Herman, Robert Vining. Originally published in JMIR Formative Research (https://formative.jmir.org), 4.2.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Summer Research Internship Curriculum to Promote Self-Efficacy, Researcher Identity, and Peer-to-Peer Learning: Retrospective Cohort Study

Yulia A Levites Strekalova[1], MBA, PhD; Rachel Liu-Galvin[1], MBChB; Samuel Border[1], MS; Sara Midence[1], BS; Mishal Khan[1], MHA; Maya VanZanten[2], BS; John Tomaszewski[3], MD; Sanjay Jain[4], MD, PhD; Pinaki Sarder[1], PhD

[1]Department of Health Services Research, Management and Policy, College of Public Health and Health Professions, University of Florida, Gainesville, FL, United States

[2]Ann & Robert H Lurie Children's Hospital of Chicago, Chicago, IL, United States

[3]Department of Pathology and Anatomical Sciences, University at Buffalo, Buffalo, NY, United States

[4]John T Milliken Department of Medicine, Washington University, St. Louis, MO, United States

Corresponding Author:
Yulia A Levites Strekalova, MBA, PhD
Department of Health Services Research, Management and Policy, College of Public Health and Health Professions, University of Florida, Gainesville, FL, United States

Abstract

Background: Common barriers to students' persistence in research include experiencing feelings of exclusion and a lack of belonging, difficulties developing a robust researcher identity, perceptions of racial and social stigma directed toward them, and perceived gaps in research skills, which are particularly pronounced among trainees from groups traditionally underrepresented in research. To address these known barriers, summer research programs have been shown to increase the participation and retention of undergraduate students in research. However, previous programs have focused predominantly on technical knowledge and skills, without integrating an academic enrichment curriculum that promotes professional development by improving students' academic and research communication skills.

Objective: This retrospective pre-then-post study aimed to evaluate changes in self-reported ratings of research abilities among a cohort of undergraduate students who participated in a summer research program.

Methods: The Human BioMolecular Atlas Program (HuBMAP) piloted the implementation of a web-based academic enrichment curriculum for the Summer 2023 Research Internship cohort, which was comprised of students from groups underrepresented in biomedical artificial intelligence research. HuBMAP, a 400-member research consortium funded by the Common Fund at the National Institutes of Health, offered a 10-week summer research internship that included an academic enrichment curriculum delivered synchronously via the web to all students across multiple sites. The curriculum is intended to support intern self-efficacy, researcher identity development, and peer-to-peer learning. At the end of the internship, students were invited to participate in a web-based survey in which they were asked to rate their academic and research abilities before the internship and as a result of the internship using a modified Entering Research Learning Assessment instrument. Wilcoxon matched-pairs signed rank test was performed to assess the difference in the mean scores per respondent before and after participating in the internship.

Results: A total of 14 of the 22 undergraduate students who participated in the internship responded to the survey. The results of the retrospective pre-then-post survey indicated that there was a significant increase in students' self-rated research abilities, evidenced by a significant improvement in the mean scores of the respondents when comparing reported skills self-assessment before and after the internship (improvement: median 1.09, IQR 0.88-1.65; W=52.5, $P<.001$). After participating in the HuBMAP web-based academic enrichment curriculum, students' self-reported research abilities, including their confidence, their communication and collaboration skills, their self-efficacy in research, and their abilities to set research career goals, increased.

Conclusions: Summer internship programs can incorporate an academic enrichment curriculum with small-group peer learning in addition to a laboratory-based experience to facilitate increased student engagement, self-efficacy, and a sense of belonging in the research community. Future research should investigate the impact of academic enrichment curricula and peer mentoring on the long-term retention of students in biomedical research careers, particularly retention of students underrepresented in biomedical fields.

(JMIR Form Res 2025;9:e54167) doi:10.2196/54167

XSL•FO
RenderX

KEYWORDS

artificial intelligence; biomedical research; curriculum; training programs; workforce

Introduction

Several population groups continue to be underrepresented among trainees pursuing biomedical research disciplines [1-5]. According to 2023 census data, Hispanic, Black, and American Indian and Alaska Native individuals comprised 34.5% of the US population (19.5%, 13.7%, and 1.3%, respectively) [6]. Yet, only 16.0% of students awarded science and engineering doctorate degrees in 2020 were from these groups [7]. These disparities carry over into the workforce, with Hispanic individuals comprising only 15.0% of the STEM (science, technology, engineering, and math) workforce in 2021, 9.0% of Black individuals, and <1.0% of American Indians and Alaska Native individuals [7]. These figures illustrate a disparity between the racial and ethnic diversity of the US population and of individuals working in biomedical and STEM fields.

The underrepresentation of trainees from diverse backgrounds is particularly pronounced in the fields of biomedical artificial intelligence (AI) and machine learning [8,9]. Biomedical research has seen significant growth in AI applications with the development of tools for diagnosing, monitoring, and predicting prognoses [10,11]. With AI's increasing role in society, AI researchers' diversity must mirror the diverse populations they serve. A skilled research workforce in domain-specific AI applications is needed to maintain sustained progress in AI research. According to a 2019 report from the AI Now Institute, women comprised only 18% of authors at AI conferences, while 80% of AI professors were male [8]. A lack of diverse perspectives can have significant consequences on resulting research products, as lack of fairness (the ability of AI to be free from bias) and explainability (the ability to understand why an AI model makes certain predictions, which is crucial for identifying biases) have plagued the AI field for years [10,11].

Common barriers to students' participation and persistence in research include experiencing feelings of exclusion and a lack of belonging, difficulties developing a strong research identity, perceptions of racial and social stigma directed toward them, and perceived gaps in research skills [1-3,12,13]. One study, for example, observed that reasons for dropout from computer science degrees included unclear expectations about the chosen field, feeling as if they did not belong, and social isolation [12]. These challenges are particularly pronounced among trainees from groups traditionally underrepresented in research, as articulated by the National Institutes of Health (NIH) [14]. The NIH reports that several racial and ethnical groups are underrepresented in research: Black or African American individuals, Hispanic or Latinx individuals, American Indians or Alaska Natives, and Native Hawaiians and other Pacific Islanders, with women carrying a higher burden of underrepresentation [14]. To address these known barriers, research exposure programs for undergraduate students can include a curriculum of academic skill development components that promote student self-efficacy, identity, and a sense of belonging in research. Specifically, evidence shows that students encouraged to develop their "science identity" are more likely to integrate into the wider culture of scientific research [3]. Students can develop their self-confidence, self-efficacy, and identity as scientists through science support experiences and connections with peers and mentors, thus benefiting from exposure to diverse perspectives and experiences [5,15,16]. Additionally, experiences that foster a sense of belonging to the research community have been shown to lead to greater persistence in STEM fields, especially for students from groups underrepresented in research [16-19]. Furthermore, STEM undergraduates who engage in small-group and active learning demonstrate increased self-efficacy, achieve better grades, and are more likely to remain in STEM fields [16,20,21].

There is a critical need for nationwide collaborative training programs to fund, develop, and mentor students from underrepresented groups to ensure AI representation in biomedical research [22]. Summer research programs delivered via both in-person and web-based formats have been shown to increase the participation and retention of undergraduate students in research [23]. Evidence shows that involvement in both face-to-face and remote summer research experiences equips undergraduate students with knowledge and research skills and promotes enrollment in postgraduate training and entry into scientific- or health care–related careers [24,25]. In addition to being exposed to research and learning research skills, students can also gain confidence as scientists and develop a researcher identity and intention to pursue a career in STEM [19,26,27]. However, previous programs have primarily focused on equipping students with technical knowledge and skills, with varying degrees of emphasis on professional development, without integrating a universal academic enrichment curriculum that is delivered synchronously via the web to students across multiple sites and that focuses on active learning and improving students' academic and research skills [23-27]. The effective development of students' research skills and familiarity with scientific methods calls for supplementing laboratory research experience with academic enrichment sessions in order to equip students with the necessary academic and professional skills to thrive as researchers building their identities in the research space. These sessions should allow students to actively engage in discussions about their identities as researchers while offering a supportive environment that fosters a sense of belonging and facilitates peer-to-peer information exchange. This study aimed to report the findings of delivering academic enrichment sessions to undergraduate students from groups underrepresented in AI research and to test the hypothesis that a guided academic enrichment curriculum, as part of a summer research internship that also included a laboratory-based experience, promotes self-perceived research skills, researcher identity, and professional skills.

Methods

Program Background

The Human BioMolecular Atlas Program (HuBMAP) is a consortium of diverse research teams with expertise in biology,

XSL•FO
RenderX

imaging, computation, data management, and visualization. The program is funded by the NIH Common Fund and aims to create a multiscale spatial atlas of the human body at single-cell resolution by applying advanced technologies and disseminating resources to the community [28]. Extensive infrastructure has been established through a coordinated effort among various HuBMAP integration, visualization, and engagement teams; tissue-mapping centers; technology and tools development groups; and rapid technology implementation teams and working groups [29]. To ensure the pipeline of emerging research talent familiar with HuBMAP's tools, data, and processes, the program began offering summer research internships for undergraduate students in 2021. The internship included placement of the student within an active research laboratory for 8 - 10 weeks, during which time they were mentored by an established investigator, and at the end of which they were encouraged to deliver a presentation of their research project. The 2023 HuBMAP internship program hosted students at 19 different research laboratories across 12 different institutions. Participants were recruited nationally by the NIH coordinating center; they submitted applications that included a research interest and career goal statement and were selected for a paid internship based on matching their research interests to those of participating mentors. A complete list of participating laboratories and institutions is included in Multimedia Appendix 1.

In 2023, the HuBMAP internship program was enhanced with an academic enrichment curriculum. The curriculum was informed by the evidence-based Entering Research curriculum, developed by a faculty trained as an Entering Research facilitator [30], and designed to address the needs and challenges of a diverse group of undergraduate students from groups underrepresented in biomedical and AI research. The academic enrichment curriculum included a series of interactive workshops, peer discussions, and hands-on writing assignments aimed at supporting students to develop their research identities and enhance their research communication skills. Each session lasted 2 hours and was delivered synchronously via Zoom, enabling students participating in internships at 19 different research laboratories across 12 different institutions across the United States to attend. A complete list of the sessions constituting the academic enrichment curriculum can be found in Multimedia Appendix 2. Sessions were structured into 1 hour of didactics, small-group activities, and discussions and 1 hour of guided work on academic products. Examples of academic products included a research statement, an individual development plan, a web-based ORCID profile, and a research poster, among others. An institutional Google Drive account was used to capture interns' work systematically and securely, promote peer-to-peer learning, and facilitate continuous access to training materials and assignments.

Data Collection and Instruments

In the summer of 2023, a total of 22 students participated in the HuBMAP summer internship, and their self-reported demographic data were collected through the initial application form. Demographic variables collected during the application process included sex, race and ethnicity, class standing, degree major, and home institution type (categorized as per Carnegie classification).

At the end of the internship, all 22 students were invited to participate in a web-based survey in which they were asked to rate their research skills self-efficacy (operationalized on the survey instrument as academic and research abilities) before the internship (referred to as the before questions) and as a result of the internship (referred to as the after questions). It is important to note that the after questions specifically asked the students to rate their abilities as a result of the internship, and not as a result of the academic enrichment curriculum. The survey employed a 17-item scale and validated questions adapted from the Entering Research Learning Assessment instrument [31]. The Entering Research Learning Assessment instrument is designed to measure learning gains of undergraduate or graduate research trainees across 7 areas: research comprehension and communication skills, practical research skills, research ethics, researcher identity, equity and inclusion awareness and skills, and professional and career development skills. The survey was piloted with undergraduate students and presented at several meetings to internship leaders at the NIH and participating sites to gain expert feedback and ensure the content and external validity of the items. Examples of questions in the survey included those that asked about students' self-reported ability to communicate the context, methods, and results of their research, to tailor research communications to different audiences, to work comfortably in the research environment, to explore possible research career pathways, to set research career goals, and to develop a plan to pursue a research career. For each item, students were asked to indicate their abilities on a scale of 1 - 5, with 1 representing "no ability," 2 representing "a little ability," 3 representing "moderate ability," 4 representing "good ability," and 5 representing "great ability." The complete list of survey questions can be found in (Multimedia Appendix 3).

The retrospective pre-then-post survey was administered anonymously, and the participants were asked to rereport their sex, race and ethnicity, and class standing. Since the retrospective pre-post survey was anonymous, we were unable to tell which students had or had not responded to the survey; thus, we compared the demographic characteristics of the respondents (n=14) to that of the whole cohort (N=22).

Statistical Analyses

The Cronbach α was applied to measure the internal consistency of the items included to assess the participants' research skills and self-efficacy. Median, minimum, and maximum scores for each of the before and after questions were calculated. Mean scores per respondent for the before questions compared with the after questions were also calculated, and a Wilcoxon matched-pairs signed rank test was performed to assess the median difference in the mean scores per respondent for the before questions compared with the after questions. Although the survey results were anonymous, we were able to match responses to each before and after question to individual survey respondents based on anonymous identification numbers.

XSL•FO
RenderX

All statistical analyses were performed using SAS (version 9.4; SAS Institute Inc). All P values were 2-sided, with $P<.05$ considered statistically significant.

Ethical Considerations

The study was submitted for review to the University of Florida Institutional Review Board and determined to be exempt (#ET00041828). Informed consent was obtained from all survey respondents. No compensation was provided for participants in this study, and no identification of individual participants in any manuscript or supplementary material is possible.

Results

Participant Characteristics

Participants in the 2023 HuBMAP summer internship and curriculum included 22 students from undergraduate institutions across the United States. Students had declared majors in STEM fields such as biology, bioinformatics, biochemistry, and engineering. While class standing varied among the participants, half of the students were at a sophomore level before beginning their summer internship. Related to their incoming career goals, participants often reported interest in pursuing MD, PhD, or MD/PhD degrees and sought to use the internship to explore the various career opportunities available in biomedicine. Almost half of the participants (9/22, 40.9%) were from high (R2) and very high (R1) research activity academic institutions, with the rest of the participants (13/22, 59.1%) from master, baccalaureate, and associate colleges. Demographic information of the 22 students is presented in Table 1.

XSL•FO
RenderX

Table . Characteristics of Human BioMolecular Atlas Program summer 2023 interns (N=22) and survey respondents (N=14).

Demographic	Interns, n (%)	Survey respondents, n (%)
Sex		
Male	7 (31.8)	3 (21.4)
Female	14 (63.6)	10 (71.4)
Other	1 (4.5)	1 (7.1)
Race		
Black or African American	6 (27.3)	5 (35.7)
Asian or Pacific Islander	5 (22.7)	3 (21.4)
White	4 (18.2)	1 (7.1)
Other race	0 (0)	1 (7.1)
More than one race	2 (9.1)	1 (7.1)
N/A[a] or prefer not to answer	5 (22.7)	3 (21.4)
Ethnicity		
Hispanic or Latino	7 (31.8)	3 (21.4)
Not Hispanic or Latino	15 (68.2)	11 (78.6)
Class standing		
Freshman	5 (22.7)	2 (14.3)
Sophomore	11 (50.0)	7 (50.0)
Junior	3 (13.6)	2 (14.3)
Senior	3 (13.6)	3 (21.4)
Major		
Biology	8 (36.4)	—[b]
Biochemistry	5 (22.7)	—
Bioinformatics	2 (9.1)	—
Chemistry	1 (4.5)	—
Biomedical/electrical engineering	2 (9.1)	—
Genetics	1 (4.5)	—
Information technology	1 (4.5)	—
Nanomedicine	1 (4.5)	—
Neuroscience	1 (4.5)	—
Home institution type per Carnegie classification		
Doctoral universities: very high research activity (R1)	5 (22.7)	—
Doctoral universities: high research activity (R2)	4 (18.2)	—
Master colleges and universities: larger programs (M1)	4 (18.2)	—
Master colleges and universities: smaller programs (M3)	1 (4.5)	—
Baccalaureate colleges	7 (31.8)	—
Associate colleges	1 (4.5)	—

[a]N/A: not applicable.

[b]Not available.

XSL•FO
RenderX

Invitations to complete a web-based survey were sent out to all 22 students, and 14 survey responses were received and included in data analysis. The makeup of the survey respondents aligned with the overall sample, but the sample size was too small for statistical analyses. The demographic makeup of all interns who participated in the program and the subset who responded to the survey are reported in Table 1. The postprogram survey did not collect data on degree major or institutional affiliation; thus, this information is not reported for the Survey Respondents column in Table 1.

Results of Retrospective Pre-Post Survey

Cronbach α was calculated to measure the level of internal consistency among the survey questions, and its value of 0.96 for the preintervention questions and 0.92 for the postintervention questions indicated an excellent level of internal consistency, that is, all questions in the survey were assessing the primary underlying concepts being assessed: researcher skills, researcher identity, and self-efficacy in research.

Table 2 reports the median, minimum, and maximum values of the responses to each of the before questions compared with each of the after questions. Given the ordinal nature of the data, where the median was a decimal number due to the averaging of 2 middle values, the median was rounded up to the nearest whole number. Of all survey responses, 3 responses were missing (one to each of After Question 3, After Question 5, and After Question 9); missing responses therefore comprised <1% of all data.

XSL•FO
RenderX

Table . Median and range of responses to self-reported research skills assessment before versus after questions.

Skill assessed	Before, median (IQR; min, max)	After, median (IQR; min, max)
Research comprehension and communication skills		
Communicate the context, methods, and results of your research.	3 (2-4.25; 1, 5)	4 (4-5; 3, 5)
Tailor your research communications for different audiences (eg, general public and disciplinary conference)	3 (3-4; 1, 5)	5 (4-5; 4, 5)
Work in the research environment comfortably	3 (2-4.25; 1, 5)	5 (4-5; 4, 5)
Ask questions to clarify your understanding of your research project	4 (3-4; 1, 5)	5 (4-5; 4, 5)
Practice regular and open communication with your research mentor	3 (2.75-4; 1, 5)	5 (4-5; 4, 5)
Practice regular and open communication with your research team members	3 (2.75-4; 1, 5)	5 (4-5; 4, 5)
Practical research skills		
Use the tools, materials, and equipment needed to conduct research	3 (2-3.25; 1, 5)	5 (4-5; 4, 5)
Make a case for your research question based on the literature	3 (2-3; 1, 5)	4 (4-5; 3, 5)
Research ethics		
Identify forms of unethical practices or research misconduct	3 (2.75-4; 1, 5)	4 (4-5; 3, 5)
Researcher identity		
Think of yourself as a scientist/researcher	3 (2-3; 1, 5)	4 (3.75-5; 2, 5)
Feel like you belong in research	3 (2-3.25; 1, 5)	4 (3-5; 2, 5)
Researcher confidence and independence		
Work independently on your research project	2 (1.75-3.25; 1, 4)	5 (4-5; 2, 5)
Determine the next steps in your research project	2 (1.75-3; 1, 5)	4 (3.75-5; 3, 5)
Equity and inclusion awareness and skills		
Understand how others might experience research differently based on their identity (eg, race, socioeconomic status, first-generation status)	3 (1-4; 1, 5)	5 (4-5; 1, 5)
Professional and career development skills		
Explore possible research career pathways	3 (2-4; 1, 5)	5 (3.75-5; 3, 5)
Set research career goals	3 (2-3; 1, 5)	5 (4-5; 3, 5)
Develop a plan to pursue a research career (determine the next step in your training)	3 (2-3.25; 1, 4)	5 (4-5; 3, 5)

The results from the Wilcoxon signed rank matched-pairs test indicated that there was a significant improvement in the mean scores of the respondents when comparing the before questions to the after questions (improvement: median 1.09, IQR 0.88-1.65; W=52.5, $P<.001$). The mean scores of each respondent for the before questions compared with the after questions are presented in Multimedia Appendix 4. A review of individual items revealed an increase in all items, indicating that overall, students felt their research abilities had improved following the HuBMAP Summer Research Internship.

Discussion

This paper reports on the pilot implementation of an academic enrichment curriculum delivered as part of the summer, AI-focused, undergraduate research internship administered through the HuBMAP program. Following participation in the internship, students reported increases in self-assessed research skills, including their abilities to set research career goals, to think of themselves as researchers, and to work comfortably in the research environment, and increases in their feelings of belonging in research. Our curricular intervention included

XSL•FO
RenderX

academic enrichment content supplemented by peer learning and hands-on activities.

Earlier research has demonstrated that perceived gaps in research skills and feelings of isolation have detrimental effects on students' persistence in computer science and AI-focused training [12]. While most summer research internships focus on individual training and technical aspects of research with varying degrees of emphasis on professional development [23-27], the format and content of the supplemental academic enrichment curriculum we piloted showed promise in creating a nurturing, active, and collaborative learning environment for student interns across multiple institutions, with demonstrated self-reported improvements in abilities that went beyond technical research skills.

Based on our study's findings, we have formulated propositions and hypotheses about the mechanisms through which academic enrichment curricula promote research skills self-efficacy and researcher identity and result in sustained engagement in research training. First, delivering an academic enrichment curriculum enhances and amplifies the learning experiences of students participating in summer research programs and promotes the long-term retention of these students in biomedical research careers. Second, academic enrichment sessions help to build a research community and promote a sense of belonging among students. Third, providing students with opportunities to learn from one another encourages peer learning and enhances students' critical thinking skills, resourcefulness, and resilience within academia. Additional research questions that we would like to answer through future studies and longitudinal assessments include how students can learn from one another and whether peer learning is superior to learning from mentors. Future studies should also explore the effect of individual versus cohort-based research experiences. The academic enrichment curriculum we delivered to participants of the HuBMAP summer internship program would be considered a mix of cohort (delivered through the academic enrichment sessions) and hybrid (since students were based at separate laboratories and institutions with individual mentors) research experiences. In addition, further research is needed to determine the optimal structure of the academic enrichment curriculum, including the frequency of sessions (once weekly vs every other week, for example).

Our experience suggests that the integration of an academic enrichment curriculum into a laboratory-based AI-focused summer research internship can help students increase their confidence, communication skills, self-efficacy, and sense of belonging in the research community. In line with our findings, a growing body of evidence shows that a structured curriculum supports undergraduate students from traditionally underrepresented groups in developing their identity as scientists, and in their overall self-efficacy to use scientific thinking, through science-support experiences and connections formed with peers [5,15,16]. Thus, we recommend that future programs should incorporate small group peer learning, active learning strategies, and laboratory-based research experience as evidence-supported methods to facilitate increased self-efficacy and a sense of belonging in the research community, which may also have the potential to promote the retention of undergraduate students in STEM and biomedical disciplines [16,20,21].

Despite these encouraging results, it is important to recognize that the generalizability of the results of this study is limited by its relatively small sample size of 22 students, of which 14 students (63.6%) completed the pre- and postprogram surveys. As a result, this study's findings may not fully represent the broader experiences of other students across other programs. It is also important to note that it was not possible to separate the effect of our academic enrichment curriculum from other activities that students may have been engaged in during their HuBMAP Summer Research Internship. Multiple individual laboratories across different institutions hosted students participating in the 2023 HuBMAP summer internship program. Therefore, while we delivered this academic enrichment curriculum to all students via the web, some students may have been provided additional learning opportunities (remote or in-person) at the laboratories or institutions where they were based. The nature and quantity of such other learning opportunities may have varied across the different sites and should be explored in future studies. It is possible that a portion of the gains in students' self-reported research abilities were attributable to these supplementary activities and not because of the academic enrichment curriculum. Future research should therefore investigate the specific impact of academic enrichment curricula on the long-term retention of students in biomedical research careers, particularly retention of students underrepresented in biomedical fields. Furthermore, each student received one-to-one and group mentoring from lead investigators and senior trainees of the host labs. The effects of mentoring and the role of peer-to-peer interaction as contributing factors in student academic skill development can also be quantified and measured in future evaluations of summer research experiences. Additionally, this study used a pre-then-post evaluation design, which combined the questions in which students were asked to rate their research abilities "before" and "as a result of" the HuBMAP Summer Research Internship into a single survey administered post program. This methodological design allowed the students to assess knowledge gained from a perspective informed by a 10-week experience. However, the effects of recall and social desirability biases can be evaluated with sufficiently powered studies that use different evaluation formats.

Following their participation in the HuBMAP 2023 summer internship program, which included an academic enrichment curriculum delivered remotely, students' self-reported research abilities, including their confidence and communication skills, their self-efficacy in research, and their abilities to set research career goals, increased. Students also reported an increased ability to think of themselves as researchers and to work comfortably in the research environment, as well as increased feelings of belonging in research after participating in the program. We recommend that summer internship programs incorporate an academic enrichment curriculum with small-group peer learning in addition to a laboratory-based experience to facilitate increased student engagement, self-efficacy, and a sense of belonging in the research community. Strengthening the pipeline of skilled and diverse

XSL•FO
RenderX

professionals entering these fields will help in addressing both current and future workforce needs, reducing bias in AI fields, and cultivating the next generation of scientists and leaders in public health, with the goal of improving health outcomes for all populations.

Acknowledgments

This work was supported in part by the National Institutes of Health Common Fund (award number 3OT2OD033753).

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

YALS contributed to the writing of the original draft, conceptualization, data curation, funding acquisition, investigation, methodology, resources, supervision, and validation. RLG was responsible for writing the original draft and conducting formal analysis. MV contributed to the writing of the original draft, data curation, and resources. SB reviewed and edited the manuscript and provided resources. SM managed project administration and provided resources. MK contributed to project administration, provided resources, and reviewed and edited the manuscript. JT reviewed and edited the manuscript and provided resources. SJ reviewed and edited the manuscript and contributed to funding acquisition. PS contributed to the review and editing of the manuscript, funding acquisition, resources, and supervision.

Conflicts of Interest

PS is an advisor of digipath.ai. The authors have no further interests to declare.

Multimedia Appendix 1
List of research laboratories/institutions that hosted students for the 2023 Human BioMolecular Atlas Program summer internship program.
[DOCX File, 15 KB - formative_v9i1e54167_app1.docx]

Multimedia Appendix 2
Schedule and objectives of academic enrichment sessions offered to summer 2023 Human BioMolecular Atlas Program interns.
[DOCX File, 17 KB - formative_v9i1e54167_app2.docx]

Multimedia Appendix 3
Academic enrichment evaluation instrument.
[DOCX File, 17 KB - formative_v9i1e54167_app3.docx]

Multimedia Appendix 4
Mean scores per respondent for before questions compared to after questions.
[DOCX File, 15 KB - formative_v9i1e54167_app4.docx]

References

1. Jantzer J, Kirkman T, Furniss KL. Understanding differences in underrepresented minorities and first-generation student perceptions in the introductory biology classroom. J Microbiol Biol Educ 2021 Dec;22(3):e00176-21. [doi: 10.1128/jmbe.00176-21] [Medline: 34804325]
2. Fisher AJ, Mendoza-Denton R, Patt C, et al. Structure and belonging: pathways to success for underrepresented minority and women PhD students in STEM fields. PLoS One 2019;14(1):e0209279. [doi: 10.1371/journal.pone.0209279] [Medline: 30625173]
3. Hurtado S, Cabrera NL, Lin MH, Arellano L, Espinosa LL. Diversifying science: underrepresented student experiences in structured research programs. Res High Educ 2009 Mar;50(2):189-214. [doi: 10.1007/s11162-008-9114-7] [Medline: 23503690]
4. Estrada M, Burnett M, Campbell AG, et al. Improving underrepresented minority student persistence in STEM. CBE Life Sci Educ 2016;15(3):es5. [doi: 10.1187/cbe.16-01-0038] [Medline: 27543633]
5. McGee R Jr, Saran S, Krulwich TA. Diversity in the biomedical research workforce: developing talent. Mt Sinai J Med 2012;79(3):397-411. [doi: 10.1002/msj.21310] [Medline: 22678863]
6. Population estimates, July 1, 2023, (V2023) quickfacts US. US Census Bureau. 2024. URL: https://www.census.gov/quickfacts/fact/table/US [accessed 2024-09-17]

XSL•FO
RenderX

7. Diversity and STEM: women, minorities, and persons with disabilities. National Science Foundation. 2023. URL: https://ncses.nsf.gov/pubs/nsf23315/report/ [accessed 2023-12-18]
8. West SM. Discriminating systems: gender, race, and power in AI - report. AI Now Institute. 2019. URL: https://ainowinstitute.org/publication/discriminating-systems-gender-race-and-power-in-ai-2 [accessed 2024-09-17]
9. Zhang D, Maslej N, Brynjolfsson E, et al. The AI index 2022 annual report. arXiv. Preprint posted online on 2022. [doi: 10.48550/arXiv.2205.03468]
10. Border SP, Sarder P. From what to why, the growing need for a focus shift toward explainability of AI in digital pathology. Front Physiol 2021;12:821217 [FREE Full text] [doi: 10.3389/fphys.2021.821217] [Medline: 35087427]
11. Wong C. AI "fairness" research held back by lack of diversity. Nature New Biol 2023 Mar 30. [doi: 10.1038/d41586-023-00935-z] [Medline: 36997714]
12. Schefer-Wenzl S, Miladinovic I, Bachinger-Raithofer S, Muckenhumer C. A study on reasons for student dropouts in a computer science bachelor's degree program. In: Auer ME, Cukierman UR, Vendrell Vidal E, Tovar Caro E, editors. Towards a Hybrid, Flexible and Socially Engaged Higher Education: Proceedings of the 26th International Conference on Interactive Collaborative Learning (ICL2023), Volume 4: Springer Nature; 2024:391-400. [doi: 10.1007/978-3-031-53382-2_38]
13. Estrada M, Woodcock A, Hernandez PR, Schultz PW. Toward a model of social influence that explains minority student integration into the scientific community. J Educ Psychol 2011;103(1):206-222. [doi: 10.1037/a0020743]
14. NOT-OD-20-031: notice of NIH's interest in diversity NIH.gov. National Institutes of Health. 2019. URL: https://grants.nih.gov/grants/guide/notice-files/NOT-OD-20-031.html [accessed 2024-09-17]
15. Chemers MM, Zurbriggen EL, Syed M, Goza BK, Bearman S. The role of efficacy and identity in science career commitment among underrepresented minority students: efficacy and identity in science career commitment. J Soc Issues 2011 Sep;67(3):469-491. [doi: 10.1111/j.1540-4560.2011.01710.x]
16. Connors PK, Lanier HC, Erb LP, et al. Connected while distant: networking CUREs across classrooms to create community and empower students. Integr Comp Biol 2021 Oct 4;61(3):934-943. [doi: 10.1093/icb/icab146] [Medline: 34190987]
17. Ostrove JM, Long SM. Social class and belonging: implications for college adjustment. Rev High Educ 2007 Jun;30(4):363-389. [doi: 10.1353/rhe.2007.0028]
18. Rosenthal L, London B, Levy SR, Lobel M. The roles of perceived identity compatibility and social support for women in a single-sex STEM program at a co-educational university. Sex Roles 2011 Nov;65(9-10):725-736. [doi: 10.1007/s11199-011-9945-0]
19. Bruthers CB, Matyas ML. Undergraduates from underrepresented groups gain research skills and career aspirations through summer research fellowship. Adv Physiol Educ 2020 Dec 1;44(4):525-539. [doi: 10.1152/advan.00014.2020]
20. Springer L, Stanne ME, Donovan SS. Effects of small-group learning on undergraduates in science, mathematics, engineering, and technology: a meta-analysis. Rev Educ Res 1999 Mar;69(1):21-51. [doi: 10.3102/00346543069001021]
21. Freeman S, Eddy SL, McDonough M, et al. Active learning increases student performance in science, engineering, and mathematics. Proc Natl Acad Sci U S A 2014 Jun 10;111(23):8410-8415. [doi: 10.1073/pnas.1319030111] [Medline: 24821756]
22. Stathoulopoulos K, Mateos-Garcia JC. Gender diversity in AI research. SSRN J 2019. [doi: 10.2139/ssrn.3428240]
23. Ajayi TB, Mueller AL, Okwuosa IS, et al. Innovations in undergraduate research training through multisite collaborative programming: American Heart Association summer undergraduate research experience syndicate. J Am Heart Assoc 2022 Apr 19;11(8):e022380. [doi: 10.1161/JAHA.121.022380] [Medline: 35388707]
24. Wierzchowski A, Wink DJ, Zhang H, Kambanis K, Robles JOR, Rosenhouse-Dantsker A. CoLab: a workshop-based undergraduate research experience for entering college students. J Chem Educ 2022 Dec 13;99(12):4085-4093. [doi: 10.1021/acs.jchemed.1c01290] [Medline: 37519308]
25. Heggland SJ, Hovde CJ, Minnich SA, Liou LE, Daniels RL. Building a biomedical pipeline: the impact of the Idaho IDeA INBRE summer research experience at a primarily undergraduate institution. Adv Physiol Educ 2021 Dec 1;45(4):786-796. [doi: 10.1152/advan.00232.2020] [Medline: 34529538]
26. Johnson MDL, Knox CJ. National Summer Undergraduate Research Project (NSURP): a virtual research experience to deliver REAL (Retention, Equity, Access, and Life-Changing) outcomes for underrepresented minorities in STEM. J Microbiol Biol Educ 2022 Apr;23(1):e00335-21. [doi: 10.1128/jmbe.00335-21] [Medline: 35496677]
27. Hudson JK, Hudson BG. Aspirnaut: a rural high school pipeline to increase diversity in STEM. Nat Rev Nephrol 2019 Oct;15(10):591-592. [doi: 10.1038/s41581-019-0190-z]
28. The Human BioMolecular Atlas Program (HuBMAP). National Institutes of Health. 2024. URL: https://commonfund.nih.gov/HuBMAP [accessed 2024-09-17]
29. Jain S, Pei L, Spraggins JM, et al. Advances and prospects for the Human BioMolecular Atlas Program (HuBMAP). Nat Cell Biol 2023 Aug;25(8):1089-1100. [doi: 10.1038/s41556-023-01194-w] [Medline: 37468756]
30. Balster N, Pfund C, Rediske R, Branchaw J. Entering research: a course that creates community and structure for beginning undergraduate researchers in the STEM disciplines. CBE Life Sci Educ 2010;9(2):108-118. [doi: 10.1187/cbe.09-10-0073] [Medline: 20516356]

XSL•FO
RenderX

31. Butz AR, Branchaw JL. Entering Research Learning Assessment (ERLA): validity evidence for an instrument to measure undergraduate and graduate research trainee development. CBE Life Sci Educ 2020 Jun;19(2):ar18. [doi: 10.1187/cbe.19-07-0146] [Medline: 32412837]

Abbreviations

AI: artificial intelligence
HuBMAP: Human BioMolecular Atlas Program
NIH: National Institutes of Health
STEM: science, technology, engineering, and math

Edited by A Mavragani; submitted 14.03.24; peer-reviewed by CM Ambrosino, G White; revised version received 14.10.24; accepted 15.10.24; published 03.02.25.

Please cite as:
Levites Strekalova YA, Liu-Galvin R, Border S, Midence S, Khan M, VanZanten M, Tomaszewski J, Jain S, Sarder P
Summer Research Internship Curriculum to Promote Self-Efficacy, Researcher Identity, and Peer-to-Peer Learning: Retrospective Cohort Study
JMIR Form Res 2025;9:e54167
URL: https://formative.jmir.org/2025/1/e54167
doi:10.2196/54167

© Yulia A Levites Strekalova, Rachel Liu-Galvin, Samuel Border, Sara Midence, Mishal Khan, Maya VanZanten, John Tomaszewski, Sanjay Jain, Pinaki Sarder. Originally published in JMIR Formative Research (https://formative.jmir.org), 3.2.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Web-Based Exercise and Nutrition Intervention to Improve Leg Muscle Strength and Physical Functioning in Older Adults: Pre-Post Pilot Study

Berber Gijsbertha Dorhout[1,2], PhD; Nick Wezenbeek[3], MSc; Lisette C P G M de Groot[1], Prof Dr; Pol Grootswagers[1], PhD

[1]Division of Human Nutrition and Health, Wageningen University & Research, Stippeneng 4, PO Box 17, Wageningen, Netherlands

[2]Research Group Innovation of Human Movement Care, HU University of Applied Sciences Utrecht, Utrecht, Netherlands

[3]CleverMove, Wageningen, Netherlands

Corresponding Author:
Berber Gijsbertha Dorhout, PhD
Division of Human Nutrition and Health, Wageningen University & Research, Stippeneng 4, PO Box 17, Wageningen, Netherlands

Abstract

Background: The lifestyle intervention ProMuscle, which combines resistance exercise and an increased protein intake, was effective in improving muscle strength, muscle mass, and physical functioning in older adults. However, due to a growing shortage of health care professionals, the rapidly growing aging population cannot be personally guided in the future. Therefore, Uni2Move, a scalable web-based variant of ProMuscle, was designed to reach larger groups of older adults without putting additional burden on health care professionals.

Objective: The current study investigated the effects of a web-based lifestyle intervention on muscle strength, protein intake, and physical functioning in healthy older adults. In addition, we conducted a qualitative study to gather key insights of the participants involved, as little is known about older adults' perceptions of web-based lifestyle interventions.

Methods: A pre-post pilot study was conducted in the Netherlands. In the 24-week intervention, 19 healthy adults aged >55 years were included. They performed resistance training at home twice a week for 24 weeks via web-based workout videos. Videos (45 - 60 minutes) were recorded by the fitness trainer and mainly focused on training leg muscles. In addition, older adults were advised on increasing protein intake via two web-based consultations by a dietitian in the first 12 weeks and via an e-learning course in the second 12 weeks. Intervention adherence and acceptance was measured in week 25. The 1 repetition maximum knee extension strength, repeated chair rise test, and protein intake were measured at baseline, week 13, and week 25. Linear mixed models were used to test differences over time. Semistructured interviews were used to gather experiences of participants. Atlas.ti version 22 was used to analyze the interviews.

Results: The mean age of participants (n=19) at baseline was 69 (SD 7) years. The 1 repetition maximum knee extension strength and repeated chair rise test improved significantly during the 24-week intervention with a mean difference of 7.0 kg (95% CI 4.8-9.3; $P<.001$) and −1.2 seconds (95% CI −1.7 to −0.6; $P<.001$), respectively. Total protein intake per day did not change, whereas protein intake during breakfast had increased significantly after 13 weeks with a mean difference of 6.9 g (95% CI 1.1-12.7; $P=.01$). Qualitative research revealed that advantages of the program included no need to travel and exercising in their own environment. Disadvantages were the lack of physical interaction and no corrections by the trainer.

Conclusions: The results of the web-based exercise and nutrition intervention Uni2Move indicate potential improvements of muscle strength and physical functioning in healthy middle-aged and older adults. Providing such lifestyle interventions on the internet could reach an increased number of older adults, providing the opportunity to contribute to the health and independence of the rapidly growing aging population.

(JMIR Form Res 2025;9:e54392) doi:10.2196/54392

KEYWORDS

web-based lifestyle intervention; resistance exercise; protein intake; muscle strength; muscle mass; older adults; web-based exercise; nutrition; exercise; resistance training; sarcopenia

XSL•FO
RenderX

Introduction

The aging population is growing and life expectancy is increasing [1]. With advancing age, many people are experiencing sarcopenia. Sarcopenia can be defined as the age-related decline in skeletal muscle mass and function, highlighting the importance of advancing age as a factor of sarcopenia [2]. Sarcopenia is a natural part of the aging process, but its progression can be hastened by several factors, such as inflammation associated with aging (inflammaging), chronic diseases, lack of physical activity, inadequate nutrition, unintentional weight loss, and disuse events [3]. As the global population of older adults increases, the prevalence of sarcopenia is anticipated to grow accordingly [2,4].

As a consequence, the demand for care and support is increasing, whereas at the same time, the shortage of health care personnel is growing [5]. Governments worldwide have called for novel public health approaches concerning healthy aging, in response to the demographic shift and limited availability of health care professionals in the future [3,6]. Therefore, it is important to focus on the prevention of sarcopenia, with the goal of improving the health of the aging population and adding life to years [7].

Fortunately, lifestyle and behavioral interventions may reduce the prevalence of sarcopenia. The recommended lifestyle advice for treatment of sarcopenia consists of an increased dietary protein intake [8,9] and progressive resistance exercise [10,11]. The ProMuscle intervention, which combines resistance exercise with an increased protein intake, has been shown to be effective in improving muscle health of older adults in various settings [12-14]. In view of the rapidly growing aging population and the potential of digitalization in today's world, a web-based version of ProMuscle seems promising. For that purpose, ProMuscle was transformed to a web-based program: Uni2Move.

Several studies showed that web-based exercise training (ie, participants train at home) resulted in an improved physical performance, physical activity, balance, mobility, and muscle strength in older adults [15,16]. In addition, a recent study showed that a blended home-based exercise and dietary protein intervention resulted in improved gait speed, physical activity level, muscle mass, strength, and dietary protein intake in older adults [17]. However, it is unknown whether an intervention combining resistance exercise and an increased protein intake that is offered completely on the internet is effective in improving muscle health. In addition, little is known about the perceptions and preferences of middle-aged and older adults regarding web-based interventions. Therefore, we investigated the effects of the web-based lifestyle intervention Uni2Move on muscle strength, physical functioning, and protein intake in healthy middle-aged and older adults and conducted qualitative research to gather the key insights and experiences from participants involved.

Methods

Uni2Move Program

The Uni2Move program was a 24-week web-based lifestyle intervention, based on the ProMuscle program.

Web-based resistance exercise training

Participants followed full body progressive resistance exercise training at home twice a week via web-based videos. Workouts lasted 45 to 60 minutes. The main focus was training the leg muscles, as these muscles are important in performing daily activities. At least 50% of each training session consisted of leg exercises, while the main upper body muscle groups were alternately trained once per week (shoulders, chest, back, and abdominal muscles).

The resistance training sessions comprised of a warm-up (5 - 7 minutes), strength training (40 - 50 minutes) and a cooldown (3 - 5 minutes). The warm-up consisted of aerobic exercises, mobility exercises, and light strength exercises like air squats and push-ups against the wall. During the cooldown, light aerobic and dynamic stretching exercises were performed. The strength training session consisted of 4 - 5 blocks. Each block consisted of 2 - 4 unique exercises that were performed for 3 - 4 sets of 40 - 45 seconds. Each set of exercises was followed by a 20 second break. In total, 10 - 12 unique exercises were performed per training session. An example of a block is as follows: set 1 squats, 20 second break, set 2 push-ups, 20 second break, set 3 squats, 20 second break, set 4 push-ups, etc. Effective muscle resting time was 80 seconds. This set-up was chosen as we expected long resting breaks to be detrimental for motivation and adherence in a home-based situation with only digital interaction possibilities.

Participants used their body weight, resistance bands, and weights (eg, dumbbells or bottles filled with sand). Exercises performed with bodyweight or weights were multiple variations of squats, box squats, lunges, wall-sits, bent-over rows, hip bridges, calf raises, planks, shoulder presses, lateral shoulder raises, reverse flies, push-ups, pullovers, crunches, and dips. Exercises performed with resistance bands were leg curls, leg extensions, leg abductions, glute kickbacks, knee lifts, horizontal rows, and shoulder external rotations. The exercise regimen was designed in a progressive manner by increasing the difficulty of the exercises during the 24-week period and encouraging participants to use higher weights, stronger bands, or to choose a more difficult variation of an exercise, which was provided by the trainer. In addition, more unilateral exercise variations were included from week 7 onward.

Although the main focus was on training around 12 - 15 reps (2 - 3 seconds for the eccentric movement and 1 - 2 seconds for the concentric movement), a small number of exercises were performed isometrically (ie, wall-sit and plank) or with a focus on eccentric load (downward movement of a one-leg box squat). From week 13 onward, we introduced power training for the leg muscles in 50% of the trainings (ie, squats, lunges, and kettlebell swings) performed at high speed to stimulate improvements in physical functioning.

XSL•FO
RenderX

Training sessions were prerecorded by a fitness trainer and broadcasted at set times. A fitness trainer was present on the web at the start and end of the broadcast to welcome participants and answer questions. Participants could also perform a missed training at a time of their preference.

Increased protein intake

Before the start of the program, participants recorded their nutritional intake for 3 days in the Traqq app [18]. Based on their nutritional intake, participants received personal advice from the dietitian about increasing their protein intake up to 25 grams during breakfast, lunch, and dinner. A protein intake of 25 grams for each main meal was chosen because it has been shown to be the requirement for appropriate muscle protein synthesis [19]. A lower protein intake may attenuate the skeletal muscle protein synthesis response in older adults [20]. In addition, guidelines of the effective ProMuscle program, in which a dietary protein intake of 25 grams per main meal was advised, were used in the Uni2Move program [13]. Consultations took place via a web-based video call or by phone. During the evaluation consultation in week 6, the advice was adjusted if necessary. From week 13 onward, participants enrolled in a web-based nutrition module. The nutrition module consisted of 12 submodules, with a new submodule released every week. Various themes related to proteins were discussed in the form of videos, short texts, assignments, quizzes, and recipes. Each participant accessed the module with a personal account and performed it at their own pace.

The pilot study did not include a comparative control group of participants because of the previous ProMuscle studies performed. ProMuscle is an in-person program and has been proven to be effective in controlled and practice settings. This pilot study was an exploration to study the effects of ProMuscle when offered completely on the internet. In case of positive results, a larger study could be conducted including a larger study population and a control group.

Participants

Adults aged 55 years and older living in and around Wageningen (The Netherlands) were recruited through flyers and news items in local newspapers. The participant's eligibility for the web-based program was assessed through a physical screening; non-frail adults were allowed to participate. Participants' ability to use the internet was assessed using two questions: (1) are you able to use email? (yes/no); (2) are you able to use internet explorer? (yes/no). Their ability to adjust the performance of exercises based on instructions was assessed by visual inspection of the fitness trainer to ensure safe participation. Participants were excluded during screening if participation might not be safe, in cases of medical limitations (such as severe heart failure or recovering from hip or knee arthroplasty), or in cases of an uncontrolled disease (ie, diabetes type 1 or 2, chronic obstructive pulmonary disease, hypertension, and cancer).

In total, 21 middle-aged and older adults were eligible for inclusion. After screening, 2 participants were excluded. Reasons for exclusion were (1) computer technical problems, resulting in them being unable to follow the web-based intervention and (2) not specified by the participant.

Measurements and Analyses

Quantitative Measures

Questionnaires were used to collect baseline characteristics including age, sex, ethnicity, education level, care use, and sports. Body weight was measured to the nearest 0.1 kg using a digital scale and height was measured to the nearest 0.1 cm using a stadiometer. Measurements were performed at week 0, 13, and 25. Uni2Move was based on the effective ProMuscle program, which has a duration of 25 weeks, and effectiveness was measured at 0, 13, and 25 weeks [12]. In order to be able to compare Uni2Move (web-based program) to ProMuscle (in-person program) measurements of Uni2Move were also performed at the beginning, half way, and end of the program (0, 13, and 25 weeks). In addition, strength training needed to be performed for at least 6 months to assure maintenance of muscle strength and mass [21]. Muscle strength was measured by a fitness trainer with a 1-repetition maximum knee extension test using the leg extension machine. Physical performance was measured by the repeated chair rise test. The chair rise test was measured as the time to complete 5 chair rises with hands folded against the chest. The same chair was used during the measurements in week 0, 13, and 25. The measurements were performed according to a standardized protocol. In addition, fall frequency was measured using a questionnaire. Nutritional intake was measured for 3 days with the Traqq app in weeks 0, 13, and 25. The average protein intake over 3 days was calculated in gram per day, g/kg body weight/d and per meal moment (breakfast, lunch, or dinner).

Adherence

Training sessions were broadcasted using Webinargeek. The number of participants that watched the training session was tracked via Webinargeek. The percent adherence was calculated as the number of training sessions watched, divided by the number of training sessions that were provided, multiplied by 100. Adherence of the dietary consultations was registered by the dietitian for each participant separately. Engagement in the web-based nutrition module was registered for each submodule by the researchers by checking the percentage of participants that completed the submodule.

Acceptance

During the measurement in week 25, acceptance of the program was checked by asking the following question: "What grade do you give the Uni2Move program on a scale of 1 - 10?" The question was asked for the total program and separately for the training sessions, as well as for the consultations with the dietitian and the nutrition module.

Qualitative Measures

A trained research assistant conducted semistructured interviews with participants to gain insight into experiences with the web-based program. Interview questions were based on a pretested interview guide from the ProMuscle in Practice Study and the guide was supplemented with questions specifically focusing on the web-based program [22]. The goal of the interview was to answer the following research questions:

XSL•FO
RenderX

1. How did participants experience the web-based training sessions?
2. How did participants experience the web-based dietary consultations?
3. How did participants experience the web-based nutrition module?

Statistical Analyses

Univariate procedures were used to check for normal distribution of the data. Baseline data were expressed as means with SD or as percentages. Effects on muscle strength and physical performance were analyzed using linear mixed models in SPSS version 29 (IBM Corp). Fixed factors were age, sex, and measuring time. Statistical significance was indicated with a *P* value of <.05. Qualitative data were analyzed in Atlas.ti version 22 (Lumivero). Interviews were taped and transcribed verbatim, and transcripts from the interviews were coded using an inductive approach. First, the interviews were read to familiarize with the data. After that, parts of the transcript were identified and classified as open codes. Finally, repeating patterns were identified and sorted in themes.

Ethical Considerations

The research team of the study officially consulted the medical ethical committee, with the question if medical ethical approval was necessary. No medical ethical approval was needed as this research did not fall within the remit of the Dutch Medical Research Involving Human Subjects Act (in Dutch: WMO; File number METC East-Netherlands: 2021 - 13365). All steps in this study have been conducted ethically. Participants provided informed consent before participation [23]. Privacy was guaranteed by working via SmartPIA procedures. Participants were covered (for incidents related to the intervention) by the insurance of the organization. Study data were deidentified as every participant was assigned a participant number. Participants were not identifiable from participant numbers. Participants did not receive a compensation for the research. Since Uni2Move is a program that is supposed to be offered regularly, we asked participants to pay a set price of 157 euros (US $161.58) in total. We aimed to include older adults with a low socioeconomic position (SEP), by letting them participate for a reduced price of 30 euros (US $30.88) in total.

Results

Baseline Characteristics

A total of 21 community-dwelling older adults were eligible to participate, and 19 participants could be included. The mean age at baseline was 69 (SD 7) years and the majority were female (n=14, 74%; Table 1).

XSL•FO
RenderX

Table . Baseline characteristics of the intervention group (n=19).

Characteristic	Value
Age (years), mean (SD)	69 (7)
Sex, n (%)	
Female	14 (74)
Male	5 (26)
Height (cm), mean (SD)	168.4 (9.4)
Weight (kg), mean (SD)	71.4 (14.6)
Native Dutch ethnicity, n (%)	19 (100)
Education level, n (%)	
Low[a]	0 (0)
Intermediate[b]	2 (10)
High[c]	17 (90)
Receiving care[d], n (%)	
Yes	1[e] (5)
No	18 (95)
Doing sports weekly, n (%)	
Yes	15 (79)
No	4 (21)
Type of sports, n (%)	
Hiking	11 (58)
Cycling	9 (47)
Nordic walking	2 (10)
Running	1 (5)
Aqua gym	1 (5)
Swimming	1 (5)
Yoga	1 (5)
Tai chi	1 (5)
Dancing	1 (5)
Canoeing	1 (5)
Golf	1 (5)
Total exercise hours per week, mean (SD)	4 (2)

[a]Low: primary education, prevocational secondary education, secondary vocational education level 1, undergraduate senior general secondary education, or undergraduate preuniversity education.

[b]Intermediate: secondary vocational education level 2, 3, or 4; senior general secondary education; or preuniversity education.

[c]High: higher vocational education or university.

[d]Care was defined as informal care, housekeeping, personal care, nursing care, individual counseling, or group counseling.

[e]Care received was housekeeping, once a week.

Quantitative Study

Results are presented in Table 2 and Figures 1 and 2. Participants significantly improved knee extension strength after 13 weeks by 5.1 kg (95% CI 2.8-7.3; $P<.001$) and by 7.0 kg (95% CI 4.8-9.3; $P<.001$) after 25 weeks, compared to baseline. Performance on the repeated chair test improved significantly after 13 weeks (−0.9 seconds, 95% CI −1.5 to −0.3; $P=.001$) and 25 weeks (−1.2 seconds, 95% CI −1.7 to −0.6; $P<.001$). Frequency of falls in the past 6 months decreased from 3 out of 17 (18%) at baseline to 1 out of 17 (6%) in weeks 13 and 25. Total protein intake did not change significantly between baseline and week 13 or week 25. Protein intake during breakfast increased significantly by 6.9 g (95% CI 1.1-12.7; $P=.01$) between week 0 and 13. Mean differences of protein intake during other meal moments did not change significantly.

XSL•FO
RenderX

Figure 1. Effects of muscle strength in kilogram at week 0, 13, and 25. Error bars indicate SD. *Indicates a significant increase compared to baseline (*P*<.05).

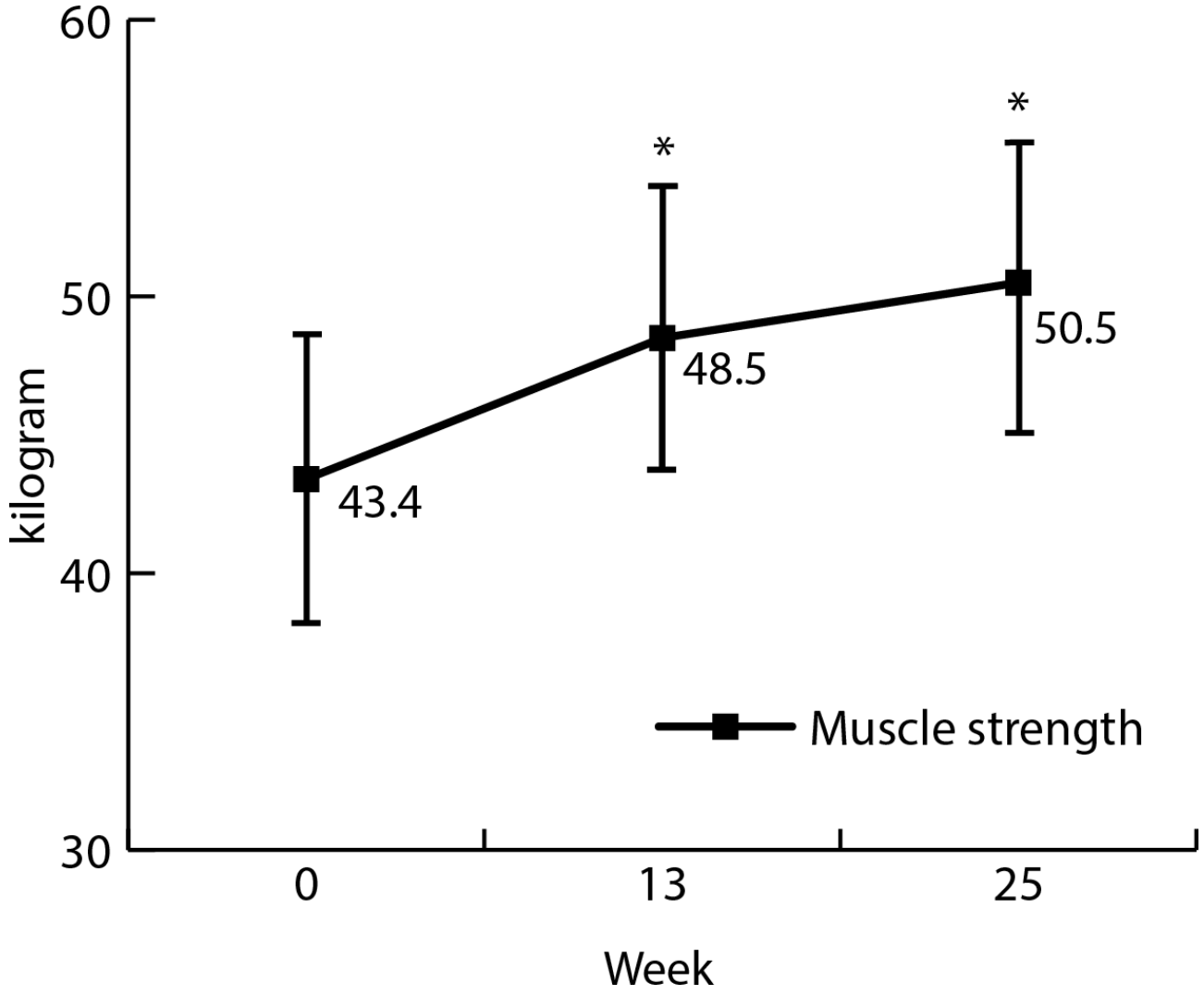

XSL•FO
RenderX

Figure 2. Effects of repeated chair rise test in seconds at week 0, 13, and 25. Error bars indicate SD. *Indicates a significant increase compared to baseline (*P*<.05).

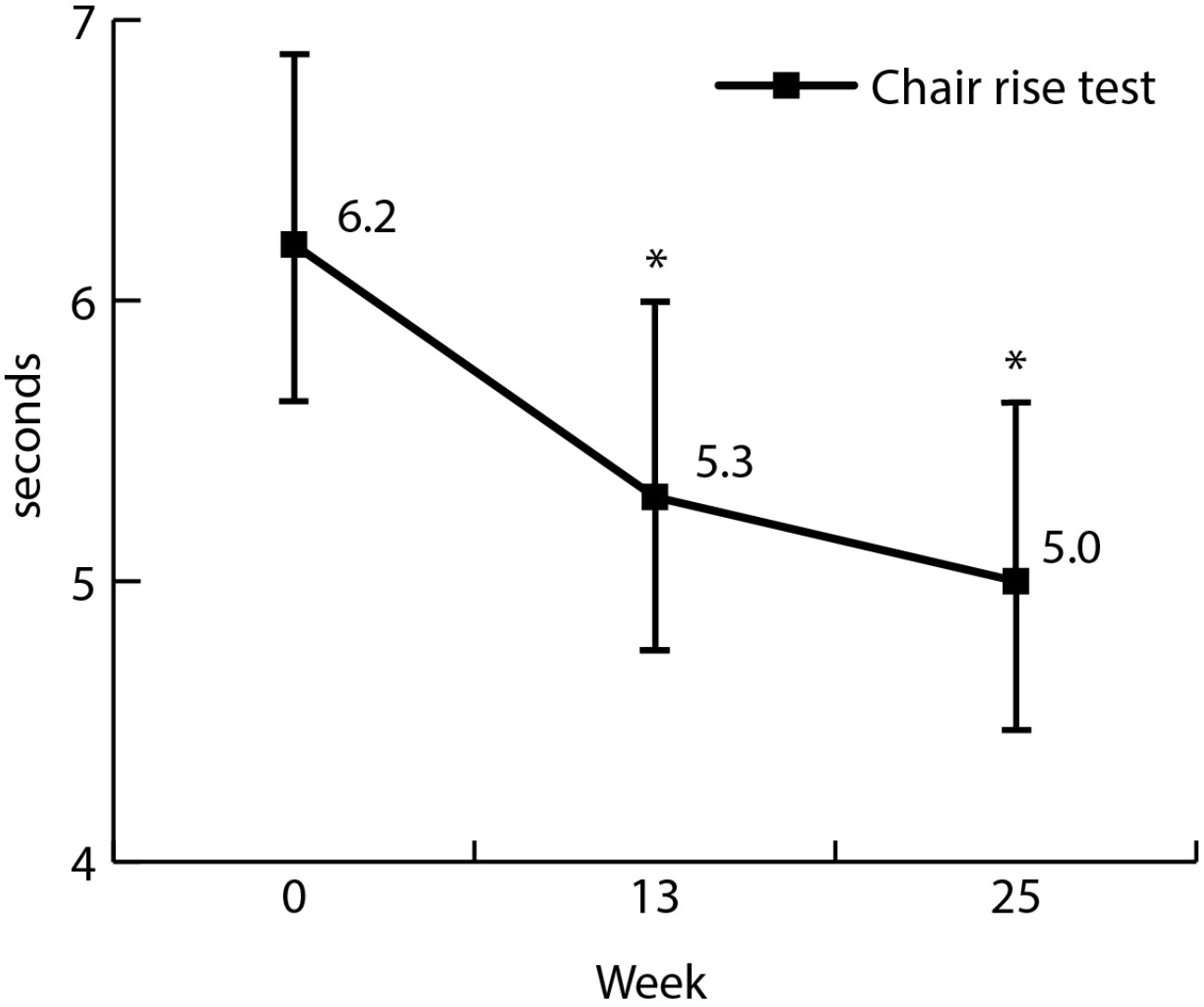

XSL•FO
RenderX

Table . Study outcomes at week 0, 13, and 25 and the time effects of weeks 0 - 25, 0 - 13, and 13 - 25.

Study outcome	Week 0, mean (95% CI)	Week 13, mean (95% CI)	Week 25, mean (95% CI)	Week 0 - 25		Week 0 - 13		Week 13 - 25	
				Mean difference (95% CI)	*P* value	Mean difference (95% CI)	*P* value	Mean difference (95% CI)	*P* value
Chair rise test (s)	6.2 (5.6-6.9)	5.3 (4.7-6.0)	5.0 (4.4-5.7)	−1.2 (−1.7 to −0.6)	<.001	−0.9 (−1.5 to −0.3)	.001	−0.3 (−0.9 to 0.3)	.65
1RM[a] knee extension (kg)	43.4 (38.5-48.3)	48.5 (43.6-53.4)	50.5 (45.5-55.4)	7.0 (4.8-9.3)	<.001	5.1 (2.8-7.3)	<.001	2.0 (−0.3 to 4.2)	.10
Protein intake (g/day)**	79.4 (65.6-93.2)	90.2 (75.9-104.4)	85.2 (70.3-100.2)	5.9 (−12.6 to 24.3)	>.99	10.8 (−6.9 to 28.5)	.39	−4.9 (−23.5 to 13.7)	>.99
Protein intake (g/kg body weight/day)	1.2 (0.9-1.5)	1.3 (1.0-1.6)	1.2 (0.9-1.5)	0.1 (−0.2 to 0.4)	>.99	0.2 (−0.1 to 0.4)	.54	−0.1 (−0.4 to 0.2)	>.99
Protein intake breakfast (g/day)	16.1 (10.9-21.3)	23.0 (17.6-28.4)	20.3 (14.6-26.1)	4.2 (−2.1 to 10.5)	.32	6.9 (1.1-2.7)	.01	−2.7 (−9.2 to 3.8)	.94
Protein intake lunch (g/day)	22.0 (17.1-27.0)	25.5 (20.0-31.0)	28.6 (22.9-34.4)	6.6 (−0.4 to 13.6)	.07	3.5 (−3.3 to 10.3)	.65	3.1 (−4.4 to 10.6)	.95
Protein intake dinner (g/day)	33.6 (25.9-41.4)	37.6 (29.2-45.9)	35.1 (26.3-44.0)	1.5 (−8.6 to 11.6)	>.99	3.9 (−5.6 to 13.5)	.96	−2.5 (−13.2 to 8.3)	>.99

[a]1RM: 1 repetition maximum.

Adherence

Participants followed on average 72% of the training sessions that were offered in the 24-week program. In the first 12 weeks, participants followed 77% of the training sessions, as compared to 67% in the second 12 weeks. The percentage of participants that participated in the web-based nutrition module, consisting of 12 modules, differed per week. On average, 80% of the participants performed the first 6 modules, and on average, 60% of the participants conducted the last 6 modules.

Qualitative Study

Participants' experiences were retrieved with interviews (n=7). The average length of the interviews was 27 minutes. For the interviewees, 6 out of 7 (86%) were woman with an average age of 66 years. The focus of the interviews was gaining insights in the experiences with the web-based training sessions, consultations, and nutrition module. Participants rated the total Uni2Move program 8.5 on average (scale 1 - 10).

Experiences With Web-Based Training Sessions

The training sessions were rated 8.6 on average (scale 1 - 10). Most of the participants did not follow the web-based training sessions live but caught up at a time of their preference. They indicated several advantages of a web-based program, such as no need to travel, exercising in their own environment, and the possibility to train at a self-determined time. Participants appreciated that the fitness trainer provided variations to exercises, for example receiving an alternative exercise if they had a prosthesis.

> *The benefit is that you have no travel time of course. All I have to do here is walk from one room to another.* [Participant, female]

Disadvantages mentioned by participants were the lack of physical interaction, social control, training in a group, and corrections by the trainer. Some participants also indicated that more discipline was needed.

> *You need a little more discipline to think: tomorrow morning I am going to do it.* [Participant, female]

> *A drawback is that there is a little bit less correction. You need to think by yourself; How do I place my back and how do I stand? There is nobody who says; Hello, a bit to the front, a bit to the back.* [Participant, female]

Experiences With Web-Based Dietary Consultations

Participants rated the consultations 7.3 on average (scale 1 - 10). Participants appreciated the fact that the consultations with the dietitian could be planned at any time of day. In addition, the form of the consultation was flexible and participants mentioned that the dietitian always checked whether the information they provided was sufficient.

> *She always ended the consultation with asking whether the advise she provided was clear and*

XSL•FO
RenderX

sufficient, and for me it was fulfilling. [Participant, female]

For several participants, the web-based consultations made the contact with the dietitian less personal, leaving limited room for discussion. In addition, the consultations were a challenge for a minority of the participants. Due to limited digital skills, video calling was challenging for them. As an alternative, these participants conducted phone consultations.

Experiences With the Web-Based Nutrition Module

The web-based nutrition module was rated 7.1 on average (scale 1 - 10). Participants mentioned that the level of difficulty for the modules was suitable. However, some participants thought it should be more challenging. Participants appreciated the schedule of the nutrition module, with a new module being presented on the web every week. In addition, the different elements of the nutrition module (knowledge clips, text, quizzes, and recipes) were mentioned as assets.

Participants missed the connection between the exercise and nutrition program, as they were offered separately and integration between the two was not present. According to the participants, a feature that should be added to the nutrition module is the possibility to have an web-based conversation in order to gather experiences from other users. On the content level, participants mentioned the need for information on skipping a meal, as well as vegetarian and vegan diets.

Discussion

Principal Findings

To our knowledge this is the first study evaluating the effects of an intervention combining resistance exercise and an increased protein intake that is completely web-based. Participants significantly improved knee extension strength after 13 weeks by 5.1 kg (95% CI 2.8-7.3; $P<.001$) and by 7.0 kg (95% CI 4.8-9.3; $P<.001$) after 25 weeks, compared to baseline. Performance on the repeated chair test improved significantly after 13 weeks by −0.9 seconds (95% CI −1.5 to −0.3; $P=.001$) and after 25 weeks by −1.2 seconds (95% CI −1.7 to −0.6; $P<.001$). Protein intake during breakfast increased significantly between week 0 and 13. Total protein intake per day did not change significantly. Participants positively valued Uni2Move, rating the total program an 8.5 (scale 1 - 10). Advantages of the web-based training sessions and dietary consultations were the flexibility to conduct the program in their own environment and at a self-determined time. Challenges were the lack of personal contact and corrections by the trainer.

Web-Based vs Offline Program

Results of the web-based Uni2Move program are in line with the offline program ProMuscle in Practice for the chair rise test after 13 weeks, which were −0.9 seconds (15%) versus −0.8 seconds (6%), respectively, and knee extension strength improved by 5.1 kg (12%) versus 9.6 kg (14%), respectively [13]. These findings show that the offline program is a promising strategy in improving muscle health of middle-aged and older adults. However, several challenges of the web-based program in comparison to the face-to-face program should be addressed. The average protein intake decreased after week 13, and the adherence dropped slightly during the course of the web-based program. Several participants indicated in the interviews that discipline was needed to proceed with the adjusted diet, and personal contact was lacking in the nutrition module. In line with this, Aalbers et al [10] concluded in their systematic review that tailored web-based interventions were rated as more interesting and helped middle-aged participants more, compared to generic web-based interventions [24]. This emphasizes the importance of personalizing web-based interventions and corresponding to the needs of participants. As suggested by participants, the web-based nutrition module can be improved by adding an option to chat with other participants. Social support is known to be an effective technique for the maintenance of a behavior change [25]. In addition, the nutrition module can be extended with advanced modules for extra challenge and additional information.

Strengths and Limitations

First, a major strength of this study is the combination of quantitative data and qualitative data. In this way, insights on the effects of the program are combined with experiences of participants with the practicalities and feasibility of the program. Second, the limited costs of the intervention are a strength. Costs of preventive interventions are always challenging for professionals as well as participants. A web-based intervention limits costs by reducing travel motions, guiding large groups of participants during web-based training, reusing training videos, and reusing the nutrition module. This decreases a possible cost-related barrier for participation.

Several limitations should be mentioned. First, as participants perform the workout in their own environment, the performance of the exercises is less regulated. As a consequence, the intensity of the performed exercises could not be measured. On the other hand, the results show improvements for the chair rise test as well as the knee extension strength, and adherence levels and rating of the training sessions are satisfactory. Based on that, we may assume that participants performed the exercises with a sufficient intensity.

Second, the lack of physical attendance during the intervention limits the possibility of measuring adherence to the program in an accurate way. Attendance to training sessions was based on whether participants watched the video of the training session. In a face-to-face setting, the fitness trainer would be able to check who is present and actively participates in the training sessions. In addition, fitness trainers are not able to correct participants during at home training sessions. During the interviews, several participants also indicated the lack of guidance and corrections on the performance of exercises, whereas others did not miss the guidance. For older adults who live independently at home and without complex disorders, web-based guidance can be suitable. Physical guidance on site may be appropriate for frail older people who need more assistance. Further research is warranted to carefully assess eligibility of the program in different groups.

Third, the small study population is a limitation of this pilot study. In addition, this study population was highly educated for the most part. In general, adults with a higher education

XSL•FO
RenderX

level have better health compared to their less-educated peers [26]. The latter group should also be included in future research to study the effects of Uni2Move on their muscle health and to contribute to reducing health inequalities. During the recruitment phase, several strategies that were shown to be successful strategies for reaching adults with a low SEP were executed in order to include participants with a lower SEP [24]. Multilayered recruitment strategies were conducted to increase engagement. First, we adjusted the recruitment material by adopting the language level and a personal approach by recruiting at low budget shops with flyers suitable for adults with low literacy levels. Second, we deployed existing networks and community networks by recruiting at the local food bank and neighborhoods where many adults with a low SEP live. Third, we tackled the common barrier of having limited resources to join the program by lowering the price of participation from 157 euros (US $161.58) to 30 euros (US $30.88) for participants with a low SEP. These strategies led to the inclusion of 2 participants with an intermediate education level. SEP was not measured because of participants' privacy. Recently, Collombon et al [4] published a paper on recruitment strategies for reaching adults aged 50 years and older with low SEP for participation in web-based physical activity interventions [27]. Results show that recruitment via invitation letters administered by municipality was most promising; however, the costs were relatively high (>2000 euros; US $2058.34). This strategy could be deployed in future research to include a diverse group of participants.

Future Directions

The main challenges of the web-based intervention include recruiting participants with a low SEP and adherence to the program. Recruitment of more participants with a low SEP should be attempted in follow-up studies by sending invitations via municipalities, as well as by focusing on recruitment on social media and at gyms [27]. Adherence to the program can be increased in several ways. Promising strategies appear to be increasing personal contact and tailoring the intervention to the needs of participants. This can be executed by conducting a hybrid intervention, alternating web-based and physical training sessions. In this way, fitness trainers can tailor the level of the exercises to the needs of the participants and can correct participants if needed. Personal contact evolves during training sessions, not only between participants but also between participants and fitness trainers. The advantages of a hybrid intervention are lower participation costs, flexibility of training sessions at home, less health care professionals needed, increased safety due to face-to-face training sessions, and expectedly higher adherence on the long term due to tailoring and personal contact.

Conclusion

In summary, the web-based lifestyle intervention Uni2Move is a pilot study that resulted in an improvement of muscle strength and performance on the repeated chair rise test, comparable to the previous offline ProMuscle studies. The results should be interpreted with caution, as the study population is small. Advantages of the current intervention are flexibility for participants as well as professionals, as well as a lack of travel time and costs. Further research is needed to evaluate the intervention's effects in larger groups, and should include participants with a low SEP, be in a hybrid form, and focus on adherence in the long term.

Acknowledgments

Within Uni2Move, Wageningen University collaborated with CleverMove and dietitians from Zoeetje. CleverMove is specialized in guiding people over 55 with (web-based) strength training. Zoeetje is specialized in advising on nutrition in health and disease. We would like to thank the fitness trainer of CleverMove, the dietitians of Zoeetje, the participants of Uni2Move, Susanne Wildemast, and the students, Eline Jonkers and Carlijn de Vries, involved in conducting the Uni2Move program, for their contribution. The research described was funded by ZonMw (05460402110005). The development of the web-based nutrition module was financed by the Rijksinstituut voor Volksgezondheiden Milieu (RIVM). These organizations had no role in the design, analysis, or writing of this article. The study was funded by ZonMw.

Data Availability

The data sets generated during and/or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: BGD, NW, LCPGMG, PG
Data curation: BGD, PG
Formal analysis: BGD
Funding acquisition: BGD, NW, LCPGMG, PG
Investigation: BGD
Methodology: BGD, NW, LCPGMG, PG
Project administration: BGD, NW, PG
Resources: BGD, NW, PG
Software: BGD
Supervision: BGD, NW, LG, PG

XSL•FO
RenderX

Validation: BGD, NW, PG
Visualization: BGD
Writing – original draft: BGD
Writing – review & editing: NW, LG, PG

Conflicts of Interest

NW is owner of CleverMove, which is a commercial organization. NW conducted the training session of Uni2Move. NW had no role in the analyses or interpretation of the data.

References

1. Shen Y, Liu D, Li S, et al. Effects of exercise on patients important outcomes in older people with sarcopenia: an umbrella review of meta-analyses of randomized controlled trials. Front Med (Lausanne) 2022;9:811746. [doi: 10.3389/fmed.2022.811746] [Medline: 35186999]
2. Tieland M, Dirks ML, van der Zwaluw N, et al. Protein supplementation increases muscle mass gain during prolonged resistance-type exercise training in frail elderly people: a randomized, double-blind, placebo-controlled trial. J Am Med Dir Assoc 2012 Oct;13(8):713-719. [doi: 10.1016/j.jamda.2012.05.020] [Medline: 22770932]
3. Wet medisch-wetenschappelijk onderzoek met mensen. Overheid.nl. 2024. URL: https://wetten.overheid.nl/BWBR0009408/2025-01-01 [accessed 2025-01-23]
4. Collombon E, Bolman CAW, de Bruijn GJ, Peels DA, Lechner L. Recruitment strategies for reaching adults aged 50 years and older with low socioeconomic status for participation in online physical activity interventions. Front Digit Health 2024;6:1335713. [doi: 10.3389/fdgth.2024.1335713] [Medline: 38800097]
5. van Dongen EJI, Haveman-Nies A, Doets EL, Dorhout BG, de Groot L. Effectiveness of a diet and resistance exercise intervention on muscle health in older adults: ProMuscle in Practice. J Am Med Dir Assoc 2020 Aug;21(8):1065-1072. [doi: 10.1016/j.jamda.2019.11.026] [Medline: 31948853]
6. Duurzame zorg en preventie (SPR 2019-2022). : Rijksinstituut voor Volksgezondheiden Milieu; 2024 URL: https://www.rivm.nl/over-het-rivm/strategisch-programma-rivm/duurzame-zorg-en-preventie [accessed 2025-01-23]
7. Kwasnicka D, Dombrowski SU, White M, Sniehotta F. Theoretical explanations for maintenance of behaviour change: a systematic review of behaviour theories. Health Psychol Rev 2016 Sep;10(3):277-296. [doi: 10.1080/17437199.2016.1151372] [Medline: 26854092]
8. Izquierdo M, Merchant RA, Morley JE, et al. International exercise recommendations in older adults (ICFSR): expert consensus guidelines. J Nutr Health Aging 2021;25(7):824-853. [doi: 10.1007/s12603-021-1665-8] [Medline: 34409961]
9. van Dongen EJI, Doets EL, de Groot L, Dorhout BG, Haveman-Nies A. Process evaluation of a combined lifestyle intervention for community-dwelling older adults: ProMuscle in Practice. Gerontologist 2020 Nov 23;60(8):1538-1554. [doi: 10.1093/geront/gnaa027] [Medline: 32259835]
10. Aalbers T, Baars MAE, Rikkert M. Characteristics of effective Internet-mediated interventions to change lifestyle in people aged 50 and older: a systematic review. Ageing Res Rev 2011 Sep;10(4):487-497. [doi: 10.1016/j.arr.2011.05.001] [Medline: 21628005]
11. Ouderen. Statistiek CB voor de. 2024. URL: https://www.cbs.nl/nl-nl/visualisaties/dashboard-bevolking/leeftijd/ouderen [accessed 2025-01-17]
12. Coletta G, Phillips SM. An elusive consensus definition of sarcopenia impedes research and clinical treatment: a narrative review. Ageing Res Rev 2023 Apr;86:101883. [doi: 10.1016/j.arr.2023.101883] [Medline: 36792012]
13. Paddon-Jones D, Rasmussen BB. Dietary protein recommendations and the prevention of sarcopenia. Curr Opin Clin Nutr Metab Care 2009 Jan;12(1):86-90. [doi: 10.1097/MCO.0b013e32831cef8b] [Medline: 19057193]
14. Lucassen DA, Brouwer-Brolsma EM, van de Wiel AM, Siebelink E, Feskens EJM. Iterative development of an innovative smartphone-based dietary assessment tool: Traqq. J Vis Exp 2021 Mar 19(169). [doi: 10.3791/62032] [Medline: 33818566]
15. Cruz-Jentoft AJ, Bahat G, Bauer J, et al. Sarcopenia: revised European consensus on definition and diagnosis. Age Ageing 2019 Jul 1;48(4):601. [doi: 10.1093/ageing/afz046] [Medline: 31081853]
16. UN decade of healthy ageing plan of action 2021-2030. : World Health Organization; 2020 URL: https://cdn.who.int/media/docs/default-source/decade-of-healthy-ageing/decade-proposal-final-apr2020-en.pdf [accessed 2025-01-17]
17. Hill KD, Hunter SW, Batchelor FA, Cavalheri V, Burton E. Individualized home-based exercise programs for older people to reduce falls and improve physical performance: a systematic review and meta-analysis. Maturitas 2015 Sep;82(1):72-84. [doi: 10.1016/j.maturitas.2015.04.005] [Medline: 25989701]
18. Katsanos CS, Kobayashi H, Sheffield-Moore M, Aarsland A, Wolfe RR. Aging is associated with diminished accretion of muscle proteins after the ingestion of a small bolus of essential amino acids. Am J Clin Nutr 2005 Nov;82(5):1065-1073. [doi: 10.1093/ajcn/82.5.1065] [Medline: 16280440]
19. Global strategy on human resources for health: workforce. : World Health Organization; 2016 URL: https://www.who.int/publications/i/item/9789241511131 [accessed 2025-01-17]

XSL•FO
RenderX

20. Dorhout BG, de Groot L, van Dongen EJI, Doets EL, Haveman-Nies A. Effects and contextual factors of a diet and resistance exercise intervention vary across settings: an overview of three successive ProMuscle interventions. BMC Geriatr 2022 Mar 9;22(1):189. [doi: 10.1186/s12877-021-02733-6] [Medline: 35264105]
21. Bautmans I, Van Puyvelde K, Mets T. Sarcopenia and functional decline: pathophysiology, prevention and therapy. Acta Clin Belg 2009;64(4):303-316. [doi: 10.1179/acb.2009.048] [Medline: 19810417]
22. Raghupathi V, Raghupathi W. The influence of education on health: an empirical assessment of OECD countries for the period 1995-2015. Arch Public Health 2020;78:20. [doi: 10.1186/s13690-020-00402-5] [Medline: 32280462]
23. Coelho-Júnior HJ, Milano-Teixeira L, Rodrigues B, Bacurau R, Marzetti E, Uchida M. Relative protein intake and physical function in older adults: a systematic review and meta-analysis of observational studies. Nutrients 2018 Sep 19;10(9):1330. [doi: 10.3390/nu10091330] [Medline: 30235845]
24. Nelson ME, Layne JE, Bernstein MJ, et al. The effects of multidimensional home-based exercise on functional performance in elderly people. J Gerontol A Biol Sci Med Sci 2004 Feb;59(2):154-160. [doi: 10.1093/gerona/59.2.m154] [Medline: 14999030]
25. Rosenberg IH. Sarcopenia: origins and clinical relevance. J Nutr 1997 May;127(5 Suppl):990S-991S. [doi: 10.1093/jn/127.5.990S] [Medline: 9164280]
26. Johnson NR, Kotarsky CJ, Mahoney SJ, et al. Evenness of dietary protein intake is positively associated with lean mass and strength in healthy women. Nutr Metab Insights 2022;15:11786388221101829. [doi: 10.1177/11786388221101829] [Medline: 35734029]
27. van den Helder J, Mehra S, van Dronkelaar C, et al. Blended home-based exercise and dietary protein in community-dwelling older adults: a cluster randomized controlled trial. J Cachexia Sarcopenia Muscle 2020 Dec;11(6):1590-1602. [doi: 10.1002/jcsm.12634] [Medline: 33103379]

Abbreviations

SEP: socioeconomic position

Edited by A Mavragani; submitted 14.11.23; peer-reviewed by B Egan, M Stein; revised version received 15.11.24; accepted 02.12.24; published 30.01.25.

Please cite as:
Dorhout BG, Wezenbeek N, de Groot LCPGM, Grootswagers P
Web-Based Exercise and Nutrition Intervention to Improve Leg Muscle Strength and Physical Functioning in Older Adults: Pre-Post Pilot Study
JMIR Form Res 2025;9:e54392
URL: https://formative.jmir.org/2025/1/e54392
doi:10.2196/54392

© Berber Gijsbertha Dorhout, Nick Wezenbeek, Lisette C P G M de Groot, Pol Grootswagers. Originally published in JMIR Formative Research (https://formative.jmir.org), 30.1.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Identifying Digital Markers of Attention-Deficit/Hyperactivity Disorder (ADHD) in a Remote Monitoring Setting: Prospective Observational Study

Heet Sankesara[1], BTech; Hayley Denyer[2], PhD; Shaoxiong Sun[1,3], PhD; Qigang Deng[2], MSc; Yatharth Ranjan[1], MSc; Pauline Conde[1], BSc; Zulqarnain Rashid[1], PhD; Philip Asherson[2], MBBS, PhD; Andrea Bilbow[4]; Madeleine J Groom[5,6], PhD; Chris Hollis[6], PhD FRCPsych; Richard J B Dobson[1,7,8,9,10], PhD; Amos Folarin[1,7,8,9,10], PhD; Jonna Kuntsi[2], PhD

[1]Department of Biostatistics & Health Informatics, Institute of Psychiatry, Psychology and Neuroscience, King's College London, 16 De Crespigny Park, London, United Kingdom

[10]NIHR Biomedical Research Centre, University College London Hospitals NHS Foundation Trust, London, United Kingdom

[2]Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom

[3]Department of Computer Science, University of Sheffield, Sheffield, United Kingdom

[4]Attention Deficit Disorder Information and Support Service (ADDISS), Edgware, Middlesex, United Kingdom

[5]School of Medicine, Mental Health & Clinical Neurosciences, Institute of Mental Health, University of Nottingham, Nottingham, United Kingdom

[6]NIHR MindTech Healthcare Technology Co-operative, Institute of Mental Health, University of Nottingham, Nottingham, United Kingdom

[7]Institute of Health Informatics, University College London, London, United Kingdom

[8]NIHR Biomedical Research Centre, South London and Maudsley NHS Foundation Trust and King’s College London, London, United Kingdom

[9]Health Data Research UK London, University College London, London, United Kingdom

Corresponding Author:
Heet Sankesara, BTech
Department of Biostatistics & Health Informatics, Institute of Psychiatry, Psychology and Neuroscience, King's College London, 16 De Crespigny Park, London, United Kingdom

Abstract

Background: The symptoms and associated characteristics of attention-deficit/hyperactivity disorder (ADHD) are typically assessed in person at a clinic or in a research lab. Mobile health offers a new approach to obtaining additional passively and continuously measured real-world behavioral data. Using our new ADHD remote technology (ART) system, based on the Remote Assessment of Disease and Relapses (RADAR)–base platform, we explore novel digital markers for their potential to identify behavioral patterns associated with ADHD. The RADAR-base Passive App and wearable device collect sensor data in the background, while the Active App involves participants completing clinical symptom questionnaires.

Objective: The main aim of this study was to investigate whether adults and adolescents with ADHD differ from individuals without ADHD on 10 digital signals that we hypothesize capture lapses in attention, restlessness, or impulsive behaviors.

Methods: We collected data over 10 weeks from 20 individuals with ADHD and 20 comparison participants without ADHD between the ages of 16 and 39 years. We focus on features derived from (1) Active App (mean and SD of questionnaire notification response latency and of the time interval between questionnaires), (2) Passive App (daily mean and SD of response time to social and communication app notifications, the SD in ambient light during phone use, total phone use time, and total number of new apps added), and (3) a wearable device (Fitbit) (daily steps taken while active on the phone). Linear mixed models and *t* tests were employed to assess the group differences for repeatedly measured and time-aggregated variables, respectively. Effect sizes (*d*) convey the magnitude of differences.

Results: Group differences were significant for 5 of the 10 variables. The participants with ADHD were (1) slower (P=.047, d=1.05) and more variable (P=.01, d=0.84) in their speed of responding to the notifications to complete the questionnaires, (2) had a higher SD in the time interval between questionnaires (P=.04, d=1.13), (3) had higher daily mean response time to social and communication app notifications (P=.03, d=0.7), and (4) had a greater change in ambient (background) light when they were actively using the smartphone (P=.008, d=0.86). Moderate to high effect sizes with nonsignificant P values were additionally observed for the mean of time intervals between questionnaires (P=.06, d=0.82), daily SD in responding to social and communication app notifications (P=.05, d=0.64), and steps taken while active on the phone (P=.09, d=0.61). The groups did not differ in the total phone use time (P=.11, d=0.54) and the number of new apps downloaded (P=.24, d=0.18).

XSL•FO
RenderX

Conclusions: In a novel exploration of digital markers of ADHD, we identified candidate digital signals of restlessness, inconsistent attention, and difficulties completing tasks. Larger future studies are needed to replicate these findings and to assess the potential of such objective digital signals for tracking ADHD severity or predicting outcomes.

(JMIR Form Res 2025;9:e54531) doi:10.2196/54531

KEYWORDS

ADHD; smartphones; wearable devices; mobile health; mHealth; remote monitoring; surveillance; digital markers; attention-deficit/hyperactivity disorder; behavioral data; real world; adult; adolescent; participants; digital signals; restlessness; severity; predicting outcomes

Introduction

The core clinical symptoms of attention-deficit/hyperactivity disorder (ADHD), which affects 5.9% of children and 2.5% of adults, are inattention, impulsivity, and hyperactivity [1]. The diagnosis relies on a semistructured interview designed to evaluate the diagnostic criteria for ADHD symptoms and associated functional impairments. Cognitive tasks, which are often used in research, also reveal behavioral responses that suggest lapses in attention or impulsive responses in people with ADHD [2-4]. For example, ADHD is strongly associated with increased reaction time variability: people with ADHD are highly inconsistent in their speed of responding to cognitive tasks, which is linked to difficulties with the regulation of attention and arousal [5,6].

Beyond the conventional in-person clinic- or lab-based assessments on clinical symptoms and cognitive functioning, the field of mobile health (mHealth) offers a new approach to obtaining additional, more detailed, and longer-term behavioral data in the real world. Passive monitoring using smartphone sensors enables ongoing, objective, and unobtrusive data collection while the participant continues with their everyday activities [7-10]. The real-world behaviors captured using such passive monitoring may, for example, provide digital signals of ADHD severity or predict outcomes. Emerging evidence from research on other disorders, such as depression and schizophrenia, indicates the potential for passive smartphone data to identify markers related to symptom severity or warning signs of relapse [11,12].

We recently developed a new remote measurement technology system for adults and adolescents with ADHD (older than 16 years)—the ADHD remote technology (ART) system—which uses the Remote Assessment of Disease and Relapses (RADAR)–base mHealth platform [13]. RADAR-base is an open-source platform to leverage data from wearable devices and mobile technologies. It provides scalable and customizable capabilities for remote real-time data collection from a wide range of sources and apps, providing a unified system for researchers to store, manage, and analyze the collected data. The RADAR-base Passive App collects background data from several sensors on modern smartphones, such as ambient noise, ambient light, phone use information (eg, which apps have been used and for how long), passive audio, GPS (relative) location, local Bluetooth device connectivity, battery life, gyroscope, steps, and acceleration from smartphone sensors. The Active App collects data requiring conscious effort (eg, questionnaires and other tasks), which are customizable according to study requirements. The platform sends notifications via the Active App to remind study participants to fill out the questionnaires and other prompts; related timestamps are recorded when the notification is delivered when participants start filling out the questionnaires and finish them.

Data previously collected using the Passive App in the RADAR-CNS (Remote Assessment of Disease and Relapse – Central Nervous System) long-term remote monitoring project [14] on adults with major depressive disorder revealed digital biomarkers that have the potential to predict depression severity [15]. For example, analyses focusing on a median follow-up period of 4 months showed that nearby Bluetooth device count—a proxy for social isolation—was associated with depression questionnaire scores [16]. The nearby Bluetooth device count can reflect the participants' social connections and interactions with family, friends, co-workers, and strangers. Therefore, the data can also reflect participants' time at home, mobility, social isolation, working status, and the number of other Bluetooth devices in the house and working environment. Further analyses of geolocation data over 24 months showed that an increase in depression symptom severity was preceded by an increase in time spent at home and a reduction in the average time spent spread across different locations. These findings from the RADAR-major depressive disorder study illustrate how data from the Passive App can be used to identify digital markers of the severity of psychiatric symptoms, supporting its application to other disorders, such as ADHD.

We aimed to quantify and investigate whether adolescents and adults with ADHD differ from people without ADHD on a range of measures hypothesized to capture lapses in attention or restless or impulsive behaviors. As shown in Figure 1, we focus on analyzing passively collected data from the wearable device and a dedicated Passive App. Similar to the behavioral data gathered by the Passive App, the analysis of data from the Active App explores participants' behavior during questionnaire completion. While the content of the questionnaires and cognitive data themselves are valuable for assessing ADHD and associated symptoms, they fall outside the scope of this particular paper. We aimed to investigate digital traces passively collected from participants' smartphones and wearable devices during their daily lives.

XSL•FO
RenderX

Figure 1. Graphical abstract showing study methodology and feature generation. A total of 20 participants with and 20 without ADHD were recruited in the United Kingdom between August and November 2020 and matched on age and gender. Participants completed the clinical symptom questionnaires via the Active App, while the smartphone sensor data were collected from the Passive App and the wearable sensor data from Fitbit. From the Active App data, the variables questionnaire notification response latency and time intervals between questionnaires are derived. Social and communication apps notification response latency, ambient light, number of new apps added, and active session duration are derived from the Passive App, whereas the Fitbit steps during the active sessions are derived from the wearable device. ART: attention-deficit/hyperactivity disorder remote technology; ADHD: attention-deficit/hyperactivity disorder; RADAR: Remote Assessment of Disease and Relapses.

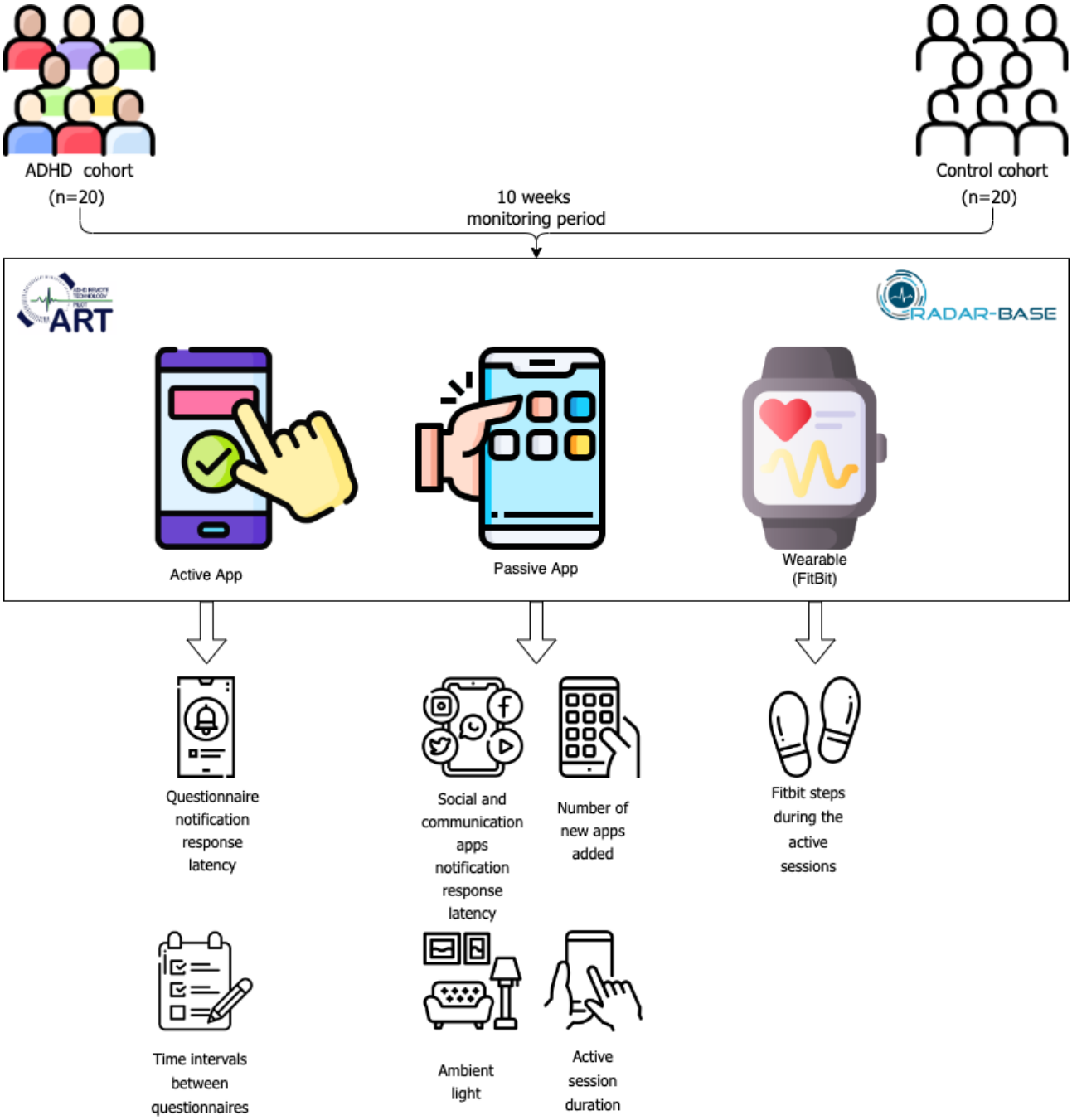

Methods

Participants

Between August and November 2020, we recruited 20 individuals with ADHD and 20 comparison individuals without ADHD between the ages of 16 and 39 years into the study (Table 1). Participants were recruited from previous studies (where they had indicated that they were willing to be contacted regarding future research studies) via the Attention Deficit Disorder Information and Support Service using an advert placed on their website and emailed to members, social media (Facebook and Twitter), and placing an advert on the King's College London Volunteer circular and on the "Call for Participants" website. Exclusion criteria were as follows: (1) having psychosis, major depression, mania, drug dependency, or a major neurological disorder; (2) any other major medical condition that might impact upon the individual's ability to participate in normal daily activity; (3) pregnancy; and (4) IQ of <70. Participants in the ADHD group additionally met the diagnostic criteria for ADHD, and participants in the comparison

XSL•FO
RenderX

group could not meet the diagnostic criteria for ADHD based on the self-report on the Barkley Adult ADHD Rating Scale (BAARS) on current symptoms (BAARS-IV) and Barkley ADHD functional impairment questionnaire.

Table . Demographics for the groups of participants with and without attention-deficit/hyperactivity disorder (ADHD).

Group	Female, n (%)	Age (years), mean (SD)	WASI-II[a] vocabulary subscale, mean (SD)
ADHD (n=20)	15 (75)	27.49 (6.04)	57.85 (7.53)
Without ADHD (n=20)	15 (75)	27.79 (6.17)	56.80 (8.35)

[a]WASI-II: Wechsler Abbreviated Scale of Intelligence.

Ethical Considerations

The study was approved by the North East – Tyne and Wear South Research Ethics Committee (REC reference: 20/NE/0034). Informed consent was obtained from participants before the assessments started. To maintain participant anonymity, their data were pseudonymized, and participants were identified in the study data with a code. Data collected from the apps, wearable devices, and interviews were only associated with this code and stored separately from any personally identifiable information. Participants were compensated £30 (US $37) after completion of the baseline sessions, £20 (US $25) after the first remote active monitoring follow-up (end of week 5), and a further £50 (US $62) at the study end point (end of week 10). The participants did not receive additional compensation for completing the debrief interviews.

Procedure

ART is an observational, nonrandomized, noninterventional study using commercially available wearable technology and smartphone sensors, representing no change to the usual care or treatments of participants due to participation.

Participants attended 2 remote baseline sessions with a research worker using Microsoft Teams. The first remote baseline session with the participants with ADHD included the administration of the following assessments: (1) the Diagnostic Interview for ADHD in adults [17] to confirm ADHD diagnosis, (2) vocabulary and digit span subscales from the Wechsler Abbreviated Scale of Intelligence, and Wechsler Adult Intelligence Scale, respectively, and (3) web-based Research Electronic Data Capture (REDCap) [18] baseline questionnaires. The baseline REDCap questionnaires for the participants with ADHD included the About You (demographic) questionnaire, AQ-10 (autism), and the COVID-19 baseline questionnaire. The same was true for participants without ADHD however, in addition, they were required to complete the BAARS/Barkley Functional Impairment Scale (BFIS) (ADHD) to ensure they did not reach the research threshold for ADHD symptoms.

The second session was administered once participants had received their wearable device and smartphone by post, approximately a week after the first session. The second session included: (1) the administration of 2 cognitive tasks (combined cued continuous performance test and Go/NoGo task, and the Fast task) and (2) a training session on the use of the wearable device (Fitbit) and the smartphone Active and Passive Apps. The participant also received a leaflet summarizing key information (Participant Technology User Guide) and researcher contact details for future reference. Each session lasted for approximately 1.5 hours. Comparison participants were assessed in the same way, except that instead of the full ADHD diagnostic interview, they completed the ADHD symptom and impairment questionnaire [19].

During the baseline session, the participants with ADHD were asked to identify a partner, parent, or close friend who could complete informant-report versions of the active monitoring questionnaires on ADHD symptoms, impairment, and irritability using the web-based REDCap. The research worker contacted the informant and invited him/her to complete the questionnaires at each of the 3 remote active monitoring time points.

Each participant was, therefore, in the study for 10 weeks. The content of questionnaires and cognitive data are beyond the scope of this analysis. We describe below measures for variables used in this analyses.

Measures

Active App: Clinical Questionnaires

Active monitoring involved the participant completing clinical symptom questionnaires on the RADAR-base smartphone Active App and 2 cognitive tasks on their home PC or laptop, which took place 3 times: at 2 weeks (the first remote self-administered assessment), 6 weeks (the second remote self-administered assessment) and 10 weeks (the third remote self-administered assessment) after the baseline remote researcher-led session.

Participants were requested to fill out the following questionnaires every 2 weeks: Patient Health Questionnaire depression scale[19], Affective Reactivity Index questionnaire [20], BAARS-IV [21], Generalized Anxiety Disorder-7 [22], and Reactive-Proactive Aggression Questionnaire [23]. Participants with ADHD also completed a medication use questionnaire; however, we excluded this questionnaire from our analyses, as the comparison group was not asked to complete it. Participants were informed that the questionnaires would be available for 3 days; however, they were told to complete them as soon as it was safe to do so. As the aim of this paper focuses on the passive measures of ADHD classification, the responses to the questionnaires are beyond the scope of this paper.

Passive Monitoring Measures

Passive monitoring, where data are collected without conscious effort from the participants using the smartphone and the wearable device (Fitbit Charge 3), started from the second

XSL•FO
RenderX

baseline session and continued for 10 weeks for each participant. Passive data were collected continuously on a 24/7 basis. Participants with an existing Android phone had the option of either using their existing phone (1 participant with ADHD and 4 participants without ADHD) or upgrading it to a study Android phone (Motorola G7 Play or G7 Power; 3 participants with ADHD and 6 without ADHD). Participants who had an iPhone were asked to replace the iPhone with a study phone for the duration of the study (16 participants with ADHD and 10 without ADHD). Before enrolling in the study, participants were told they could only use an Android phone during the study period due to the availability of sensors and use data on Android devices. It was explained to them that due to our study design, they must only use one phone. During the study, the research team monitored phone battery status and phone app use to ensure compliance. Furthermore, we analyzed the battery discharging rate and found no significant differences in the discharging rate throughout the study period, confirming that the study phone was being used as the primary phone. Participants were asked to wear the wearable device on their nondominant hands.

The Passive App collected data on ambient noise, ambient light, phone use information (eg, which apps were used and how long, when the phone was unlocked), GPS location, Bluetooth connectivity, battery life, gyroscope, steps, and acceleration. Several passive measures were used to derive features to assess the smartphone use behavior of individuals with ADHD.

One such measure is phone status data. It logs the state of a smartphone at any given moment. The state can be “STANDBY” (device on standby), “UNLOCKED” (the device is unlocked), “SHUTDOWN” (the device is shut down), or “BOOTED” (the device is booting or starting). During the period between the device being unlocked and standby again, it is assumed that the participant is actively using the smartphone. We named this period an active phone use session.

The active phone use session can help us understand how long and when a participant is active on their smartphone. Another Passive App measure is the app event data. An “event” can represent any activity in the phone app. The app event data capture background (events that happen without the help of any active user interference), interactive (events that are initiated by the user using the app’s user interface), and foreground (events that are shown in the app notification section) events. The app event data can be used to understand the user’s smartphone activity (ie, what apps they use and how long they use them).

Combining this with the aforementioned active phone use data can be used to understand the smartphone behavior of individuals with ADHD. We calculated how long a participant took to open the phone after receiving a notification from any social and communication apps. We defined social and communication as applications that are labeled as “social” (Instagram, Facebook, etc) or “communication” (WhatsApp, Facebook Messenger, SMS, etc) in the Google Play store.

The smartphone also uses its light sensor to measure ambient light exposure level (ie, illuminance). We used the light sensors to assess the ambient information of participants. The data collection and management were handled by the open-source mHealth platform RADAR-base.

Feature Generation

Using the measures described, we derived the following features that we hypothesize capture ADHD tendencies (Table 2).

XSL•FO
RenderX

Table . List of features for the analyses and their descriptions.

Number	Feature name	Feature description
1	Questionnaire notification response latency	The time taken by the participant to start completing the questionnaire after receiving the notification.
2	SD in questionnaire response latency	SD in the time taken by the participant to start completing the questionnaire after receiving the notification.
3	Mean interval between questionnaires	The mean time interval between finishing one questionnaire and starting the next one.
4	SD in intervals between questionnaires	SD in the time interval between finishing one questionnaire and starting the next one.
5	Daily mean of social and communication apps' notification response latency	Daily mean of the response time to notifications from social and communication apps.
6	Daily SD in social and communication apps' notification response latency	SD in the response time to notifications from the social and communication apps.
7	SD in ambient light	SD in ambient light while actively using the smartphone.
8	Fitbit steps during the active sessions	Fitbit step count during each active phone use session.
9	Active session duration	The time participant spends actively using their smartphone device in each active phone use session.
10	Number of new apps added	The total number of new apps added daily.

Questionnaire Response Time Features

The variable capturing the response latency of the questionnaire notification enables us to assess whether individuals with ADHD are slower to start answering the questionnaires following the notification. The variable mean intervals between questionnaires allow us to explore whether individuals with ADHD have a longer time interval between finishing one questionnaire and starting the next one. The SD in questionnaire response latency and intervals between questionnaires assess the variability in the response time and questionnaire intervals.

Passive Phone Use Features

We further created variables that allow us to assess whether participants with ADHD are more variable and differ overall in their speed of responding to notifications from social and communication apps. The notification data obtained using the Passive app is used to determine how individuals with ADHD respond to notifications from social and communication apps compared with the participants without ADHD.

When a notification from a social or communication app occurs, the phone will record an OTHER event type followed by a FOREGROUND or INTERACTION event if the participant has interacted with the notification. We have excluded all the notifications from the analyses that appeared when the participants were, indicated by their Fitbit data, to be sleeping to remove instances as it is highly unlikely that they could have seen the notifications. We analyzed the time it took for participants to unlock their phones after receiving a notification from a social or communication app.

Specifically, we computed the daily mean and SD of the notification data to derive the participants' daily aggregate of smartphone behaviors.

Ambient Light Features

The variables on the SD in ambient light and steps while active on their smartphones were hypothesized to be related to restlessness by allowing us to assess whether individuals with ADHD move around more than participants without ADHD when they are actively using their devices.

Phone Activity Features

We also created variables measuring the length of each active phone use session and the number of new apps downloaded to assess whether shorter active phone use sessions and a greater number of new apps downloaded are observed in the participants with ADHD.

Statistical Methods

We used linear mixed models (LMMs) to examine differences between the groups for repeatedly measured variables over the study period. These variables include questionnaire notification response latency, mean intervals between questionnaires, the SD in intervals between questionnaires, daily mean of social and communication apps' notification response latency, the daily SD in social and communication apps' notification response, the SD in ambient light, Fitbit steps during the active sessions, active session duration and the number of new apps added.

We chose this approach as it takes the nonindependence of data into account. Therefore, we can use repeated measures from the same participant. LMMs are also highly robust and do not need

XSL•FO
RenderX

to satisfy the prerequirements necessary for ANOVA models [24]. Moreover, we investigated the interaction effect between the days since the recruitment and the control group with respect to questionnaire notification response latency and intervals between questionnaires to evaluate how being in the study for a longer duration impacts these behavioral variables.

For nontemporal aggregated variables (SD in questionnaire notification response latency), we used the standard *t* test to compare the group differences. The *P* values are noted by *P* in the text, and variables are defined as statistically significant if $P<.05$. Along with *P* values, we have also reported the Standard Score (Z value) and effect size (*d*). For repeatedly measured variables where we used LMM, we have computed the effect size in Formula 1 [25]. The effect size of nontemporal aggregated variables is computed using Cohen *d* [26].

Formula 1, measuring effect sizes on the results of LMM [25], is as follows:

$$d=\frac{\text{difference between means}}{\sqrt{\text{varinterceptpart}+\text{varinterceptitem}+\text{varslopepart}+\text{varslopeitem}+\text{varresidual}}}$$

where

d=effect size

varinterceptpart=the intercept coefficient per participant

varinterceptitem=the intercept coefficient per item

varslopepart=the slope coefficient per participant

varslopeitem=the slope coefficient per item

varresidual=the residual variance coefficient

The effect size $d>0.6$ has been employed to determine nonsignificant trends [27]. The cut-off was used to correspond to moderate to high effect sizes. As this is primarily an exploratory study and to inform future larger-scale studies, multiple testing corrections were not undertaken to reduce the chance of type-two errors (ie, false negative results) and to avoid over-correction for multiple comparisons involving multiple correlated variables.

We identified and removed outliers by using the IQR proximity rule, a well-established method for outlier detection in skewed distributions [28]. We calculated the upper limit (UL) and the lower limit (LL), a statistical threshold computed from data points according to Formula 2. Any values in the data above UL and below LL are deemed outliers and removed from the analyses. The derived limits serve as a reference point to identify and manage outliers or unusual data points in a data set. To assess the influence of potential outliers, we conducted a sensitivity analysis. This involved reanalyzing the data after excluding the 2 participants with the highest SD in ambient light values.

Formula 2, the IQR proximity rule [28], is as follows:

data:List of data points

Q1=Percentile(data,25)

Q3=Percentile(data,75)

IQR=Q3−Q1

LL=Q1−1.5×ILR;UL=Q3+1.5×ILR

Results

Group differences were significant for 5 out of the 10 variables: questionnaire notification response latency, the SD in questionnaire notification response latency, the SD in intervals between questionnaires, daily mean of social and communication apps' notification response latency, and SD in ambient light (Table 3). While not significant, the mean intervals between questionnaires, the daily SD in social and communication apps' notification response latency and Fitbit steps during the active session show results from moderate to high effect size, that is, $d>0.60$.

XSL•FO
RenderX

Table . Summary of results for all the 10 derived features.

Number	Feature name	ADHD[a], mean (SD)	Without ADHD, mean (SD)	*P* value	Standard score (Z)	Effect size (*d*)
1	Questionnaire notification response latency (in hours)[b]	19.14 (12.98)	11.69 (10.31)	.047	−1.99	1.05
2	SD in questionnaire notification response latency (in hours)[b]	15.88 (9.27)	8.97 (6.83)	.01	2.68	0.84
3	Mean intervals between questionnaires (in hours)	1.12 (2.17)	0.18 (0.64)	.06	−1.18	0.82
4	SD in intervals between questionnaires (in hours)[b]	2.01 (5.04)	0.35 (2.23)	.04	−2.03	1.13
5	Daily mean of social and communication apps' notification response latency (in seconds)[b]	2304.27 (1406.62)	1516.45 (785.78)	.03	−2.17	0.7
6	Daily SD in social and communication apps' notification response latency (in seconds)	2736.38 (1091.17)	2125.15 (840.86)	.05	−1.96	0.64
7	SD in ambient light (in lux)[b]	83.96 (155.57)	43.63 (73.57)	.008	−2.67	0.86
8	Fitbit steps during the active sessions (in steps)	313.14 (263.88)	179.20 (177.26)	.09	−1.67	−0.61
9	Active session duration (in seconds)	1778.07 (1653.87)	1047.68 (1060.45)	.11	−1.60	0.54
10	Number of new apps added	1.18 (0.40)	1.43 (0.81)	.24	1.18	0.18

[a]ADHD: attention-deficit/hyperactivity disorder.

[b]Statistically significant result.

For questionnaire notification response latency, mean intervals between questionnaires and SD in intervals between questionnaires, we observed the interaction of these features with time and found that the features increased slightly, though non significantly, with each day. We did not find any significant interaction between the control group and the day for the features. The negative coefficient observed corresponding to the interaction effect for the features implies that the increase in features per day is smaller in the comparison group compared with the ADHD group.

Figures 2-6 illustrate the results for questionnaire notification response latency, SD in questionnaire notification response latency, SD in intervals between questionnaires, daily mean of social and communication apps' notification response latency, and SD in ambient light, respectively. Figure 5 shows the SD in ambient light while participants are actively using their smartphones. This reveals that participant IDs 5 and 36, who have ADHD, are potential outliers. The sensitivity analysis confirmed that the group difference remained statistically significant after these exclusions. Therefore, we have chosen to include all participants in the final analysis for transparency and robustness.

XSL•FO
RenderX

Figure 2. The figure presents the differences in the time taken by the participant to start completing the questionnaire after receiving the notification (questionnaire notification response latency) in participants with and without ADHD. ADHD: attention-deficit/hyperactivity disorder.

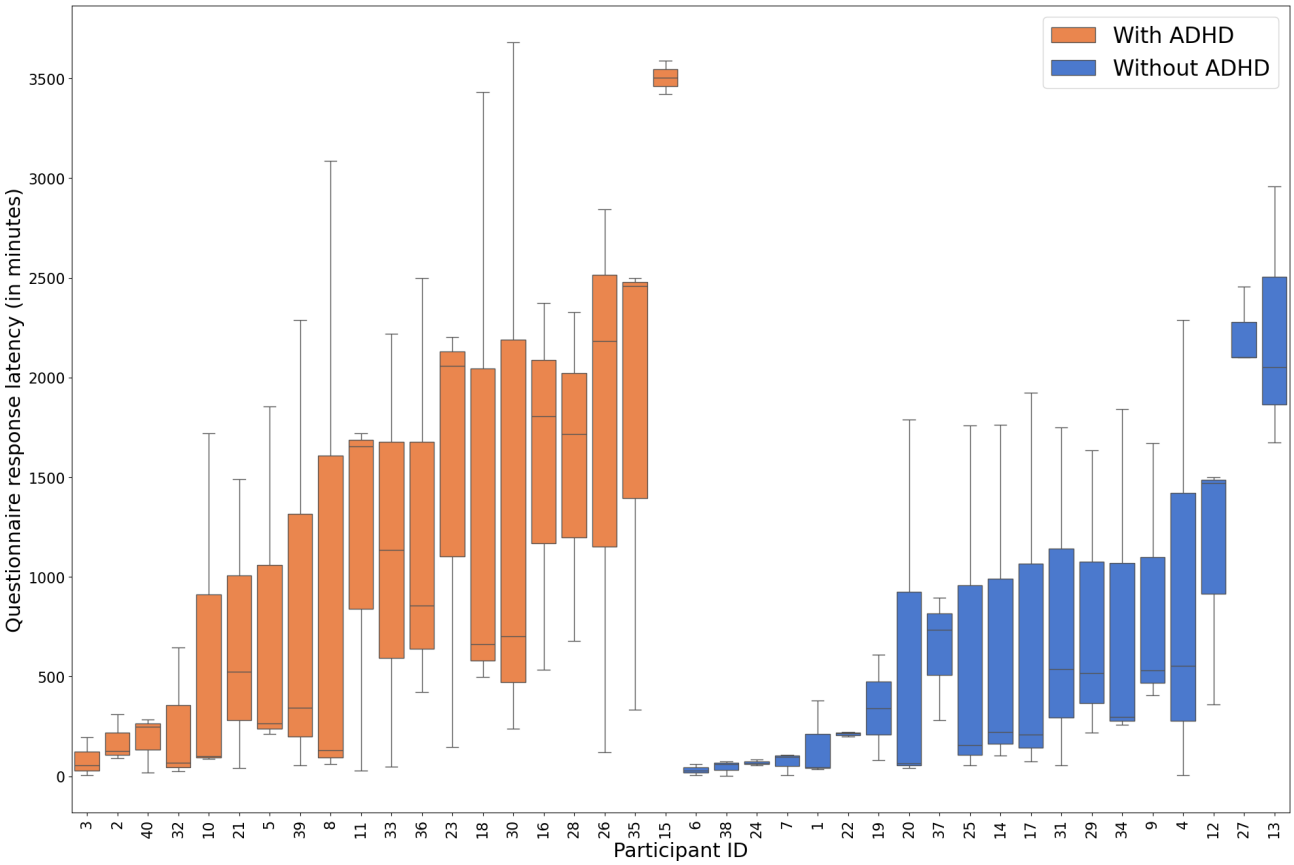

Figure 3. SD in the time taken by the participant to start completing the questionnaire after receiving the notification (SD in questionnaire notification response latency) in individuals with and without ADHD. The boxplot represents the median and IQR of SD in questionnaire notification response latency for both groups. ADHD: attention-deficit/hyperactivity disorder.

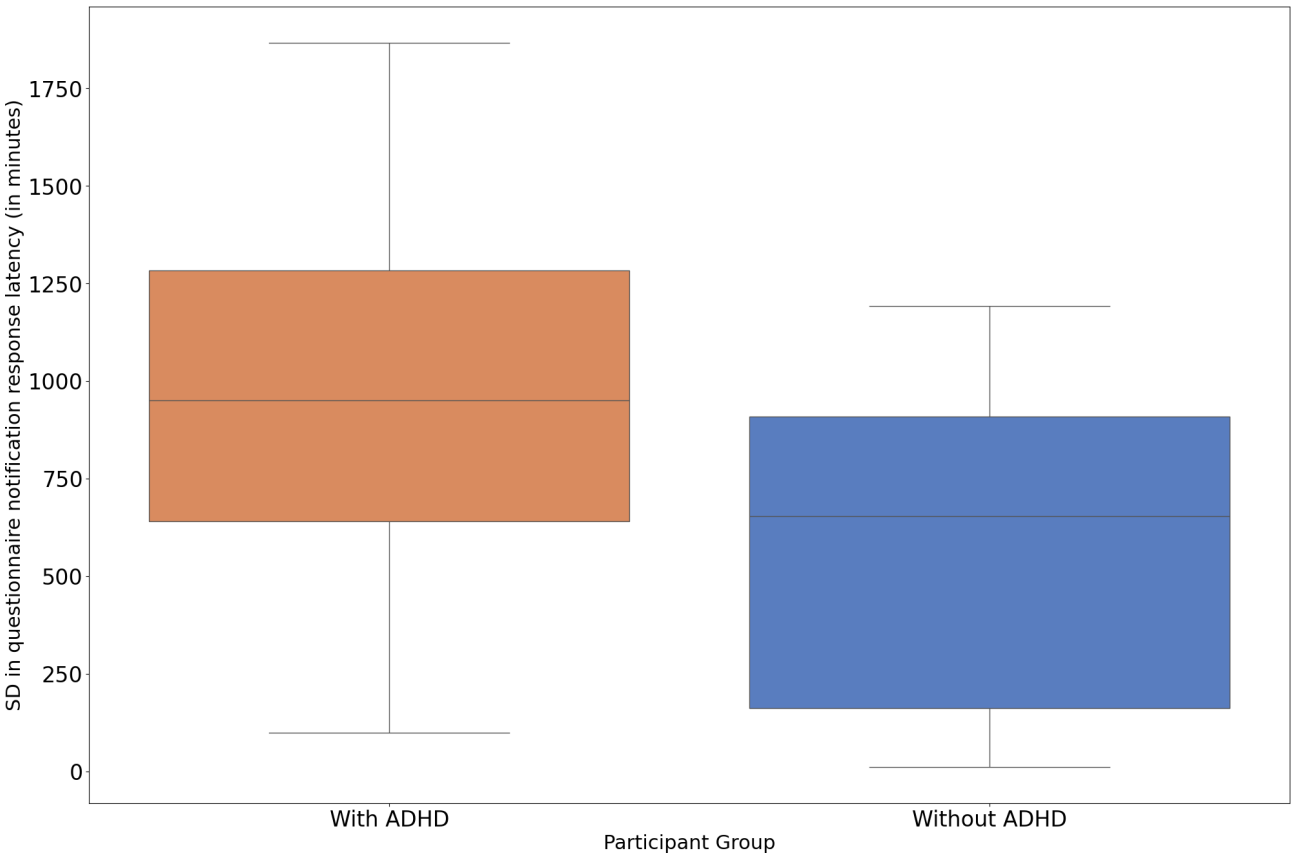

XSL•FO
RenderX

Figure 4. Log of SD in the interval between finishing one questionnaire and starting the next one (SD in intervals between questionnaires) in participants with and without ADHD. ADHD: attention-deficit/hyperactivity disorder.

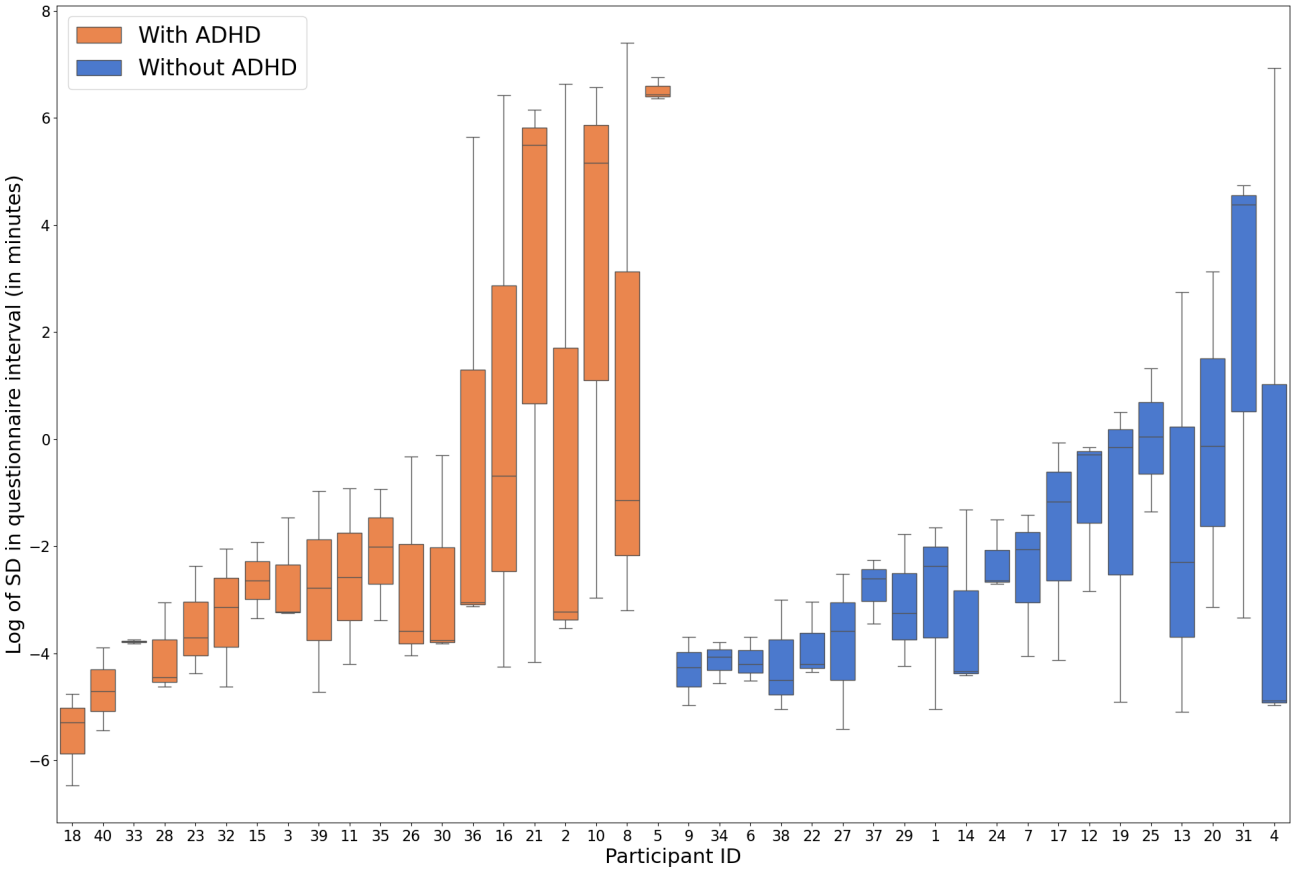

Figure 5. Daily mean of response time to notifications from social and communication apps (daily mean of social and communication apps' notification response latency) in participants with and without ADHD. ADHD: attention-deficit/hyperactivity disorder.

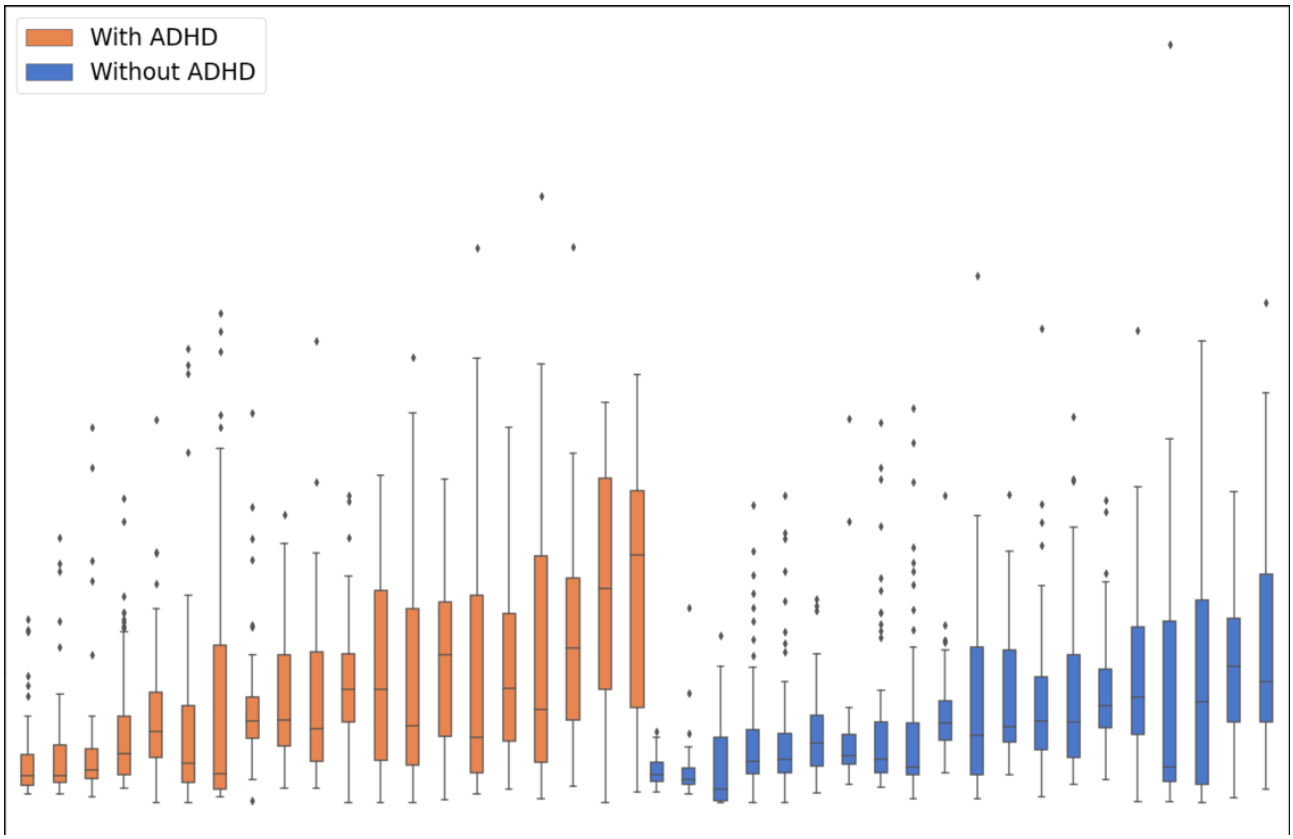

XSL•FO
RenderX

Figure 6. SD in ambient light while actively using the smartphone in participants with and without ADHD. ADHD: attention-deficit/hyperactivity disorder.

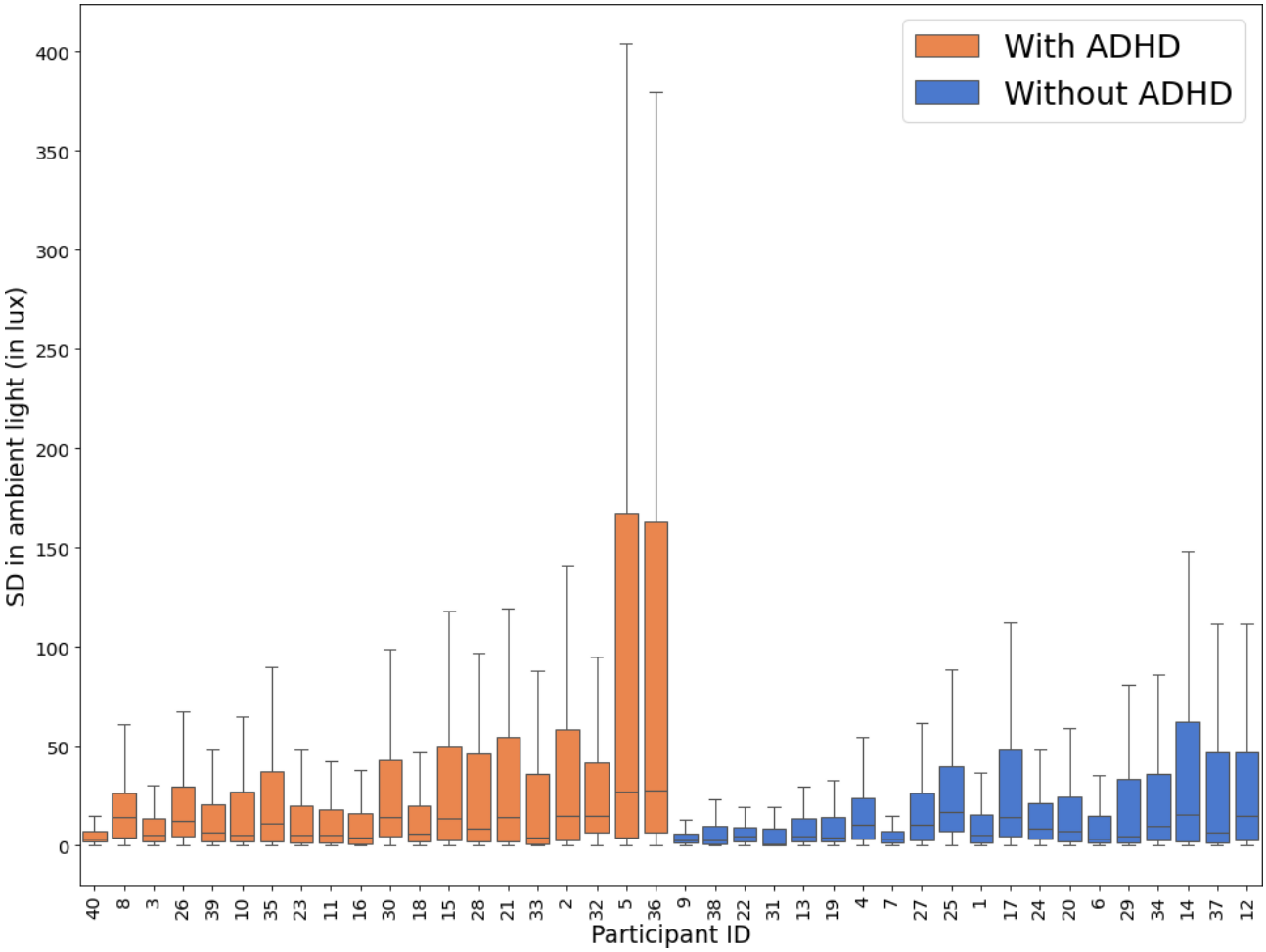

Discussion

Principal Findings

In a novel exploration of digital markers of ADHD, we identified candidate digital signals of restlessness, inconsistent attentional focusing, and difficulties completing tasks as part of a 10-week remote monitoring pilot study on our ART system.

Analyses of the data from the smartphone Active App, where participants were requested to complete a series of questionnaires at 4-week intervals, revealed that individuals with ADHD were slower and more variable in starting to complete questionnaires after receiving a notification, and more variable in the time interval between finishing one questionnaire and starting the next one, than the participants without ADHD. The Passive App data further showed that the participants with ADHD were slower to open their smartphones following a notification from a social or communication app compared with the group without ADHD. These novel findings—as well as further potential signals that emerged at trend-level only (greater mean in the ADHD group in the intervals between questionnaires and variability in the time taken to open the phone after receiving a social or communication app notification)—await replication in larger future studies.

The above findings are in line with the frequently reported difficulties in completing tasks among people with ADHD [3] and may further reflect characteristic inconsistent attentional focusing. Future research is needed to explore whether the slower and more variable speed of responding captured on the smartphone data reflects the same difficulties with the regulation of attention and arousal reported in lab-based cognitive studies in individuals with ADHD. Such cognitive studies strongly link ADHD with inconsistency in the speed of responding to cognitive tasks, which has been linked in neurophysiological studies of the regulation of attention and arousal [5,6]. Whether passive remote data collection using smartphones could offer a real-world passive alternative or augmentation to the more intensive cognitive task administration to measure behavioral characteristics associated with ADHD requires careful further study. Given the general availability of smartphones in the general population, the possibility of using this approach for symptom monitoring remains intriguing.

An additional group difference from our analyses of the Passive App data was the SD in ambient (background) light concurrent with the use of their mobile phone, which was significantly higher among the participants with ADHD. Further investigation would be useful to understand this better, but we presently hypothesize it may relate to participants with ADHD moving between differentially lit environments while using their smartphones. A related group comparison on Fitbit step count while active on their smartphone was not significant, but the high effect size observed again emphasizes the need for future

XSL•FO
RenderX

research with increased statistical power. We did not observe a difference between participants with and without ADHD in the number of new apps downloaded or in the length of active smartphone use sessions, which suggests that these markers do not capture impulsivity, restlessness, or other symptoms of ADHD.

Limitations

Since this was a pilot study, the sample size was small (n=40), and the findings require replication in future studies with larger samples. While this sample size provided sufficient power to explore initial relationships between ADHD and phone-based behavioral markers, it limits the generalizability of the findings to the broader ADHD population. Smaller sample sizes can increase the likelihood of chance findings and decrease the ability to detect subtle but meaningful effects. Therefore, replicating these findings in larger and more diverse samples is necessary to confirm their validity and generalizability. A strength of our design was, however, the large amount of data collected per participant: 10 weeks' worth of multimeasure remote monitoring data, which, in the case of passive monitoring data, was collected 24 hours a day. During the 10 weeks of the study period, we collected an average of 99,180 Fitbit step data points per participant. We collected an average of 149,821 different app activities per participant from 1855 unique apps during the same period.

Another limitation is that the data were collected during the COVID pandemic when restrictions were in place that may have affected daily activities (eg, physical activity, sleep, and smartphone activity) and well-being more generally. Importantly for our analyses, however, we note that such restrictions were concurrent for both groups.

Another limitation to consider is that 75% of our participants are women, and hence, there is a gender imbalance in our study. However, the groups were matched on gender. Future larger studies should aim to recruit samples with more representative gender distributions.

Furthermore, we provided several participants (n=35) who owned an Apple smartphone or a noncompatible Android smartphone with a new compatible Android smartphone. Despite instructing participants to use the new Android phone as their primary phone during their remote monitoring period, as participants had access to their Apple smartphone and noncompatible smartphone, it is difficult to assess how much they may have used their nonstudy phones.

Future Research

We are currently conducting a large remote monitoring study on adults with ADHD—the ADHD remote technology study of cardiometabolic risk factors and medication adherence (ART-CARMA) [29]. In ART-CARMA, we obtain remote monitoring data from the participants with ADHD both before and after initiation of ADHD medication treatment, which will allow us to examine the effects of medication on the digital markers of ADHD that we identified here. In total, we will obtain remote monitoring data from each participant for 12 months, enabling detailed over-time analyses of the multiple measures. Another direction for future research would be to explore potential age- and gender-related differences in digital markers, from childhood through adolescence and young adulthood, to older adulthood, in both men and women.

Conclusions

Using the ART system in a 10-week remote monitoring study of participants with and without ADHD, we identified novel candidate digital markers of ADHD: digital signals that we propose to capture ADHD symptoms of restlessness, inconsistent attentional focusing, and difficulties completing tasks. Future research can assess the potential for the candidate digital markers of ADHD, as identified here, to track ADHD severity or predict outcomes.

Acknowledgments

We thank all who made this research possible: all participants who contributed their time to the study and the funders. The

attention-deficit/hyperactivity disorder (ADHD) remote technology (ART) pilot study was funded by a King's Together Strategic Grant. This research was funded in part by the National Institute for Health Research (NIHR), University College London Hospitals Biomedical Research Centre, Maudsley Biomedical Research Centre at South London and Maudsley NHS (National Health Service) Foundation Trust, and King's College London. HD is supported by the UK Medical Research Council (MR/N013700/1) and King's College London member of the MRC Doctoral Training Partnership in Biomedical Sciences. QD is supported by the China Scholarship Council (201908440362). Generative artificial intelligence tools like ChatGPT were not used in writing this manuscript. However, Grammarly Pro was used for editing, grammar, and formatting purposes.

Data Availability

For open access, the author has applied a Creative Commons Attribution (CC BY) license to any author-accepted manuscript version. The data supporting this study are not openly available but may be shared on request following a formal review process by the chief investigator and coinvestigators.

Conflicts of Interest

The authors disclosed the following potential conflicts of interest related to the research, authorship, and publication of this article: JK has delivered talks at educational events sponsored by Medice, with all funds directed to King's College London to support

XSL•FO
RenderX

attention-deficit/hyperactivity disorder (ADHD) research. PA has received research funding from Vifor Pharma, delivered sponsored talks, and served as an advisor for Shire, Janssen-Cilag, Eli Lilly, Flynn Pharma, and Pfizer on the diagnosis and treatment of ADHD; all funds were received by King's College London and allocated to ADHD studies. The remaining authors have reported no conflicts of interest.

References

1. Faraone SV, Banaschewski T, Coghill D, et al. The World Federation of ADHD international consensus statement: 208 evidence-based conclusions about the disorder. Neurosci Biobehav Rev 2021 Sep;128:789-818. [doi: 10.1016/j.neubiorev.2021.01.022] [Medline: 33549739]
2. Kuntsi J, Wood AC, Rijsdijk F, et al. Separation of cognitive impairments in attention-deficit/hyperactivity disorder into 2 familial factors. Arch Gen Psychiatry 2010 Nov;67(11):1159-1167. [doi: 10.1001/archgenpsychiatry.2010.139] [Medline: 21041617]
3. Franke B, Michelini G, Asherson P, et al. Live fast, die young? A review on the developmental trajectories of ADHD across the lifespan. Eur Neuropsychopharmacol 2018 Oct;28(10):1059-1088. [doi: 10.1016/j.euroneuro.2018.08.001]
4. Sun S, Denyer H, Sankesara H, et al. Remote administration of ADHD-sensitive cognitive tasks: a pilot study. J Atten Disord 2023 Jul;27(9):1040-1050. [doi: 10.1177/10870547231172763] [Medline: 37269091]
5. James SN, Cheung CHM, Rijsdijk F, Asherson P, Kuntsi J. Modifiable arousal in attention-deficit/hyperactivity disorder and its etiological association with fluctuating reaction times. Biol Psychiatry Cogn Neurosci Neuroimaging 2016 Nov;1(6):539-547. [doi: 10.1016/j.bpsc.2016.06.003] [Medline: 27840854]
6. Cheung CHM, McLoughlin G, Brandeis D, Banaschewski T, Asherson P, Kuntsi J. Neurophysiological correlates of attentional fluctuation in attention-deficit/hyperactivity disorder. Brain Topogr 2017 May;30(3):320-332. [doi: 10.1007/s10548-017-0554-2] [Medline: 28289850]
7. Girolamo G, Barattieri di San Pietro C, Bulgari V, et al. The acceptability of real‐time health monitoring among community participants with depression: a systematic review and meta‐analysis of the literature. Depress Anx 2020 Sep;37(9):885-897. [doi: 10.1002/da.23023]
8. Ranjan Y, Rashid Z, Stewart C, et al. RADAR-base: open source mobile health platform for collecting, monitoring, and analyzing data using sensors, wearables, and mobile devices. JMIR Mhealth Uhealth 2018;7(8):e11734. [doi: 10.2196/11734] [Medline: 29674305]
9. Pratap A, Atkins DC, Renn BN, et al. The accuracy of passive phone sensors in predicting daily mood. Depress Anxiety 2019 Jan;36(1):72-81. [doi: 10.1002/da.22822] [Medline: 30129691]
10. Dalla Costa G, Leocani L, Montalban X, et al. Real-time assessment of COVID-19 prevalence among multiple sclerosis patients: a multicenter European study. Neurol Sci 2020 Jul;41(7):1647-1650. [doi: 10.1007/s10072-020-04519-x] [Medline: 32617741]
11. Barnett I, Torous J, Staples P, Sandoval L, Keshavan M, Onnela JP. Relapse prediction in schizophrenia through digital phenotyping: a pilot study. Neuropsychopharmacology 2018 Jul;43(8):1660-1666. [doi: 10.1038/s41386-018-0030-z] [Medline: 29511333]
12. Saeb S, Zhang M, Karr CJ, et al. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. J Med Internet Res 2015 Jul 15;17(7):e175. [doi: 10.2196/jmir.4273] [Medline: 26180009]
13. Ranjan Y, Rashid Z, Stewart C, et al. RADAR-base: open source mobile health platform for collecting, monitoring, and analyzing data using sensors, wearables, and mobile devices. JMIR Mhealth Uhealth 2019 Aug 1;7(8):e11734. [doi: 10.2196/11734] [Medline: 31373275]
14. Matcham F, Leightley D, Siddi S, et al. Remote assessment of disease and relapse in major depressive disorder (RADAR-MDD): recruitment, retention, and data availability in a longitudinal remote measurement study. Eur Psychiatr 2022 Jun;65(S1):S112-S112. [doi: 10.1192/j.eurpsy.2022.315]
15. Zhang Y, Folarin AA, Sun S, et al. Relationship between major depression symptom severity and sleep collected using a wristband wearable device: multicenter longitudinal observational study. JMIR Mhealth Uhealth 2021 Apr 12;9(4):e24604. [doi: 10.2196/24604] [Medline: 33843591]
16. Zhang Y, Folarin AA, Sun S, et al. Predicting depressive symptom severity through individuals' nearby bluetooth device count data collected by mobile phones: preliminary longitudinal study. JMIR Mhealth Uhealth 2021 Jul 30;9(7):e29840. [doi: 10.2196/29840] [Medline: 34328441]
17. Kooij J, Francken MH. DIVA 2.0 diagnostic interview for ADHD in adults. : DIVA Foundation; 2010 URL: www.advancedassessments.co.uk/resources/ADHD-Screening-Test-Adult.pdf
18. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381. [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
19. Kroenke K, Strine TW, Spitzer RL, Williams JBW, Berry JT, Mokdad AH. The PHQ-8 as a measure of current depression in the general population. J Affect Disord 2009 Apr;114(1-3):163-173. [doi: 10.1016/j.jad.2008.06.026] [Medline: 18752852]

XSL•FO
RenderX

20. Stringaris A, Goodman R, Ferdinando S, et al. The affective reactivity index: a concise irritability scale for clinical and research settings. J Child Psychol Psychiatry 2012 Nov;53(11):1109-1117. [doi: 10.1111/j.1469-7610.2012.02561.x] [Medline: 22574736]
21. Barkley RA, Murphy KR. Attention-Deficit Hyperactivity Disorder: A Clinical Workbook: Guilford Press; 2006.
22. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder. Arch Intern Med 2006 May 22;166(10):1092. [doi: 10.1001/archinte.166.10.1092]
23. Raine A, Dodge K, Loeber R, et al. The reactive-proactive aggression questionnaire: differential correlates of reactive and proactive aggression in adolescent boys. Aggress Behav 2006 Apr 1;32(2):159-171. [doi: 10.1002/ab.20115] [Medline: 20798781]
24. Schielzeth H, Dingemanse NJ, Nakagawa S, et al. Robustness of linear mixed ‐ effects models to violations of distributional assumptions. Methods Ecol Evol 2020 Sep;11(9):1141-1152. [doi: 10.1111/2041-210X.13434]
25. Brysbaert M, Stevens M. Power analysis and effect size in mixed effects models: a tutorial. J Cogn 2018 Jan 12;1(1):9. [doi: 10.5334/joc.10] [Medline: 31517183]
26. Lakens D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Front Psychol 2013 Nov 26;4:863. [doi: 10.3389/fpsyg.2013.00863] [Medline: 24324449]
27. Nead KT, Wehner MR, Mitra N. The use of “Trend” statements to describe statistically nonsignificant results in the oncology literature. JAMA Oncol 2018 Dec 1;4(12):1778-1779. [doi: 10.1001/jamaoncol.2018.4524] [Medline: 30347042]
28. Vinutha HP, Poornima B, Sagar BM. Detection of Outliers Using Interquartile Range Technique from Intrusion Dataset Information and Decision Sciences: Springer; 2018:511-518. [doi: 10.1007/978-981-10-7563-6_53]
29. Denyer H, Ramos-Quiroga JA, Folarin A, et al. ADHD remote technology study of cardiometabolic risk factors and medication adherence (ART-CARMA): a multi-centre prospective cohort study protocol. BMC Psychiatry 2022 Dec 20;22(1):813. [doi: 10.1186/s12888-022-04429-6] [Medline: 36539756]

Abbreviations

ADHD: attention-deficit/hyperactivity disorder
ART: attention-deficit/hyperactivity disorder remote technology
ART-CARMA: attention-deficit/hyperactivity disorder remote technology study of cardiometabolic risk factors and medication adherence
BAARS: Barkley Adult Attention-Deficit/Hyperactivity Disorder Rating Scale
BFIS: Barkley Functional Impairment Scale
LL: lower limit
LMM: linear mixed model
mHealth: mobile health
RADAR: Remote Assessment of Disease and Relapses
REDCap: Research Electronic Data Capture
UL: upper limit

Edited by A Mavragani; submitted 14.11.23; peer-reviewed by CI Campbell, HA Golkashani, TA Reddy Sure; revised version received 24.10.24; accepted 29.10.24; published 29.01.25.

Please cite as:
Sankesara H, Denyer H, Sun S, Deng Q, Ranjan Y, Conde P, Rashid Z, Asherson P, Bilbow A, Groom MJ, Hollis C, Dobson RJB, Folarin A, Kuntsi J
Identifying Digital Markers of Attention-Deficit/Hyperactivity Disorder (ADHD) in a Remote Monitoring Setting: Prospective Observational Study
JMIR Form Res 2025;9:e54531
URL: https://formative.jmir.org/2025/1/e54531
doi:10.2196/54531

© Heet Sankesara, Hayley Denyer, Shaoxiong Sun, Qigang Deng, Yatharth Ranjan, Pauline Conde, Zulqarnain Rashid, Philip Asherson, Andrea Bilbow, Madeleine J Groom, Chris Hollis, Richard J B Dobson, Amos Folarin, Jonna Kuntsi. Originally published in JMIR Formative Research (https://formative.jmir.org), 29.1.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Development of an Instagram Reel-Based Bystander Intervention Message Among College Students: Formative Survey and Mixed Methods Pilot Study

Leticia Couto, PhD

College of Communication, DePaul University, Daley Bldg, 14 E Jackson Blvd, Chicago, IL, United States

Corresponding Author:
Leticia Couto, PhD
College of Communication, DePaul University, Daley Bldg, 14 E Jackson Blvd, Chicago, IL, United States

Abstract

Background: Bystander intervention is a common method to address the ubiquitous issue that is sexual violence across college campuses. Short messages that incentivize bystander intervention behavior can be another tool to fight sexual violence.

Objective: This study aimed to conduct formative research surrounding social norms and bystander barriers to pilot and develop Instagram (Meta) reel-based messages addressing bystander intervention among college students.

Methods: The first step was to conduct a formative survey to identify peer norms and actual behavior of the intended population. Once that data were collected, a mixed methods message pilot was conducted by a survey where participants randomly saw 5 of the 12 messages developed, assessing them for credibility, perceived message effect, and intended audience.

Results: The formative survey was conducted among 195 college students from the same institution, and the pilot test was conducted among 107 college students. The formative survey indicated a discrepancy between perceived peer behavior and actual behavior of the participants in all 3 measures, allowing for the development of normative messaging. The pilot testing indicated the credibility was acceptable (eg, mean 3.94, SD 1.15 on a 5-point scale) as well as the perceived message effect (eg, mean 4.26, SD 0.94 on a 5-point scale). Intended audiences were also identified and reached. Qualitative results indicated that the messages may have lacked credibility, although the quantitative results suggest otherwise.

Conclusions: Participants understood the messages concerned bystander intervention, and perceived message effects results indicated the messages to be effective in assisting bystander intervention engagement by normative messaging. Messages were considered credible and reached the intended audience. The qualitative results provided further insights on how the messages can be adapted before being tested for effects. Future research should focus on further adapting the messages and testing their effects among the studied population.

(JMIR Form Res 2025;9:e66769) doi:10.2196/66769

KEYWORDS

bystander intervention; message development; sexual health; college; student; sexual violence; bystander; reel-based; Instagram; social media; short message; formative research; mixed methods; social norms; perceived behavior; qualitative; behavioral health; digital health

Introduction

Background

Bystander intervention trainings and messages are two of the main ways universities in the United States attempt to address the sexual violence crisis in college campuses [1,2]. Although it is common for bystander intervention messaging (here, *bystander intervention* is defined as the ability to identify situations in which a third party may assist someone in a risky and act to intervene [3]) to use bystander models as a theoretical framework [4,5], the social norms approach is also used to address sexual violence interventions [6]. For the social norms approach to be successful, it is necessary that the population of the message experiences pluralistic ignorance, where there is a difference between their own behavior and their estimate of peer behavior [6]. The social norms approach suggests that people who believe that their peers are engaging in a behavior are more likely to engage in that behavior themselves [6], whereas bystander models, in particular bystander barriers, suggest that the belief that others will take responsibility removes their own responsibility to engage in bystander intervention [7]. Many trainings and programs exist that focus on stimulating bystander intervention behavior [8,9], but less is known about the role of supporting media messages. In addition, it is necessary to investigate how these 2 opposing perspectives may be perceived as messaging mechanisms for college students.

XSL•FO
RenderX

Sexual Violence and Bystander Intervention

There is a sexual assault crisis on college campuses. Estimates indicate that approximately 1 in 4 college women are sexually assaulted during their time in college [1], which is high compared with the rate of 1 in 5 among the general population of women in the United States [2]. In addition, certain populations seem to be more at risk, such as first-year students [10] and LGBTQ+ (lesbian, gay, bisexual, transgender, queer+) students [1]. Universities across the country have tried to curb the problem by using a series of prevention efforts against sexual violence, with one of the most popular ones being bystander intervention [11].

Bystander intervention training and campaigns, including in-person and online training, are effective ways of increasing intervention intentions among college students in multiple scenarios [5,12]. An example of these campaigns is Green Dot, which features a focus on motivational speech and in-person training [12]. Over 500 universities in the United States have adopted Green Dot [13]. It is worth noting, however, that the bystander effect has been brought into question [14,15]. This can be in part due to its questionable origins in the Kitty Genovese issue [16]. Even still, this event led to an eventual refinement of what it means to intervene as a bystander, and the trainings developed from there proved to be successful [4,5,12,17,18].

Instagram as a Channel for Interventions

Using Instagram as a platform for public health concerns has been studied mostly in an exploratory manner [19]. However, some studies suggest that the issue is not necessarily reach but the ability of the entities governing these accounts to communicate with their publics [20]. In addition, such messaging developed for Instagram can be easily adapted for other channels, facilitating eventual scalability efforts.

Instagram reels, in particular, are more popular than regular Instagram posts and stories because they have better reach and can be found more easily than regular posts [21]. They also have more editing features that can be used to make messages engaging for the audience [21]. Instagram is also highly used among college-aged populations, more so than other platforms [22], such as Twitter (now X), Facebook, and Snapchat. Thus, the Instagram reel-inspired format can be an optimal channel for short and targeted messaging.

This Study

In this study, a formative survey was conducted to assess if pluralistic ignorance exists in a college population in relation to bystander intervention behavior in sexual assault situations. This step needed to be established first to ensure a social norms–based message would be feasible. In the presence of pluralistic ignorance, 3 theoretically based Instagram reel-based video messages with short scripts messages were developed and piloted in terms of credibility, intended audience, perceived message effectiveness, and a manipulation check of the conditions. One condition on bystander behavior barriers, focusing on showing participants only the actual behavioral data reported in the survey, was called the message readjustment condition in this study; another condition was based on the social norms approach, which was called the norm reinforcement condition in this study, as it focused on showing participants only the normative information from the data collected in the survey; and 1 combined both theoretical frameworks, showing both the normative information and the behavioral information with the intent of highlighting the discrepancy between the 2. The goal of this study is to assess the perceived message effectiveness, credibility, and intended audience reach of these messages pertaining to bystander intervention.

Methods

Formative Survey

In the first semester of 2023 in a large, public, American university, data were collected from 195 undergraduate students aged 18 to 29 years (mean 20.52, SD 2.22 years), most of whom identified as women (118/195, 60.5%), cisgender (184/195, 94.4%), heterosexual (123/195, 63.1%), White (139/195, 71.6%), not Hispanic (110/195, 56.7%), and middle class (124/195, 63.6%). Participants were recruited from a randomized list of 2000 emails from the university's registrar's office, representative of the university as whole.

Formative Research Measurements

Bystander Intervention Behavior

Three items were adapted from Hust et al [23]. Participants were asked to indicate how much they agree (5) or disagree (1) with statements related to perceived norms. Specifically, these items concerned the participants own behavior. These items included "I have made sure my friends are ok if I see him/her in an uncomfortable sexual situation at a party" and "I have discouraged a friend who said they planned to get someone drunk to have sex." The Cronbach α was assessed for the entire sample, for a value of .83.

Bystander Intervention Perceived Norms

Three items were adapted from Hust et al [23]. Participants were asked to indicate how much they agree (5) or disagree (1) with statements related to perceived norms. Specifically, these items refer to how they perceived their peers would act in these situations. Items include the stem "most of my friends would," followed by items such as "make sure their friend is ok if they see him/her in an uncomfortable sexual situation at a party" and "discourage a friend who said they planned to get someone drunk to have sex." The Cronbach α was assessed for the entire sample, for a value of .8.

Formative Research Results

Table 1 demonstrates that there was a discrepancy between peer perception and actual behavior of participants. The means were higher among the normative statements (bottom 3) as compared with their actual behavior items counterparts, indicating that participants said they more strongly agreed with the normative statements. This indicates that there was an opportunity for a social norms message to be developed, as pluralistic ignorance was identified.

XSL•FO
RenderX

Table . Means and SDs of items used for message development.

Items	Mean (SD)	n/N[a] (%)
I have made sure my friends are ok if I see him/her in an uncomfortable sexual situation at a party.	4.43 (1.08)	163/195 (83.6)
I have warned someone if I saw a drug being slipped into their drink.	3.63 (1.41)	93/192 (47.7)
I have discouraged a friend who said they planned to get someone drunk to have sex.	3.73 (1.43)	107/193 (55.5)
Most of my friends would make sure their friend is ok if they see him/her in an uncomfortable sexual situation at a party.	4.86 (0.41)	191/195 (98)
Most of my friends would warn someone if they saw a drug being slipped into their drink.	4.89 (0.38)	191/195 (98)
Most of my friends would discourage a friend who said they planned to get someone drunk to have sex.	4.82 (0.47)	188/195 (96.4)

[a]Percentage that indicated either “agree” (5) or “somewhat agree” (4) for each of these statements.

Pilot Testing

Messages were developed based on the items tested in the formative research survey step. A total of 12 messages were developed, 3 per condition. The experimental conditions and the theoretical basis for each can be found on Table 2.

Table . Experiment conditions based on theoretical frameworks.

Condition	Theoretical framework	Outcomes
Norm readjustment	Bystander barriers [7]	Intention to take responsibility
Norm reinforcement	Social norms approach [6]	Perceived peer norm
Norm readjustment and reinforcement	Social norms approach and bystander barriers [6,7]	Intentions
Control	__[a]	—

[a]Not applicable.

Data were collected in October of 2023. This data collection was exempt from institutional review board (IRB) review for similar reasons as the formative survey, where the university’s IRB reviewed the proposal and indicated that a full-board review was not necessary considering the minimal risks. To be eligible to participate, participants had to be currently in the United States, be an undergraduate student at the university being studied, and be between the ages of 18 and 29 years. Participants were recruited through a pool of students taking communication courses and were compensated with extra credit. Other recruitment methods were attempted, such as social media recruitment, but they yielded very few participants. These participants were compensated with Starbucks gift cards. These 3 conditions were added as screeners to ensure only participants who conformed to them participated in the pilot. This type of pilot testing strengthens the validity and reliability of the intervention, and it has been used in similar efforts [24,25].

Participants saw 5 of 12 Instagram reel-based short video messages developed in random order (3 per condition: norm readjustment, norm reinforcement, combined condition, and control). The following instructions were provided to the participants before they started the pilot: “Next, you will see 5 short videos. You will answer some questions about those videos and then some overall questions. Please make sure you can hear the videos as you watch them. Please click on the button on the bottom right to continue.” Participants were also provided with the following description of bystander intervention: “Definition of bystander intervention: Bystander Intervention is recognizing a potentially risky situation or interaction and choosing to respond in a way that could positively affect the outcome. For the purposes of this study, we are mostly referring to intervening in situations that could escalate to sexual violence.” As these messages had to be highly tailored, the messages referred to the students with the name of the university’s mascot (eg, if the university’s mascot was a tiger, the participants were called tigers). These references were removed from this paper to ensure the anonymity of the participants and an unbiased peer review process.

A total of 107 individuals participated in the pilot, but 81 participants responded to demographics items. Participants’ ages ranged from 18 to 24 years (mean 20.16, SD 1.33 years). Most identified as women (52/81, 64.2%) and cisgender (80/81, 98.8%). Most identified as heterosexual (60/81, 74.1%), but a considerable portion of the sample either chose a sexual minority (15/81, 18.4%) or chose not to answer (6/81, 7.4%). Most participants were non-Hispanic/Latine (44/81, 54.3%) and were

XSL•FO
RenderX

Caucasian/White (61/81, 75.3%). Most participants also self-identified as middle class (58, 71.6%).

Pilot Measures

Perceived Message Effectiveness

An adaptation of the 3 items from the University of North Carolina's Perceived Message Effectiveness Scale (UNC PMES) was used to measure perceived message effectiveness [26]. Example items include "This message makes me concerned about the lack of bystander intervention engagement on campus" and "This message discourages me from not engaging in bystander intervention." Participants were asked to indicate how much they agree (5) or disagree (1) with these statements. The Cronbach α for the messages varied between .68-.92. The average α was .87.

Credibility

To measure credibility, participants were asked how much they agreed (5) or disagreed (1) that the message was trustworthy, credible, believable, and realistic. The 4 items were combined to form a credibility scale. This scale was self-developed based on previous literature exploring credibility of messages [27]. The Cronbach α range for the credibility composite scores was .89-.95 with an average of .92.

Intended Audience for the Message

Participants were asked to indicate how much they disagree (1) or agree (5) that the following groups are the intended audience of the message: college students, people between 18 - 29 years old, and [university] students. The goal of this item was to assess whether the message resonates with its intended audience (ie, college students, [university] students, people aged between 18 - 29 years). These items were self-developed to address the specific intended audiences of these messages.

Manipulation Checks

A total of 6 manipulation checks were developed. The participants were asked to indicate how much they agreed (5) or disagreed (1) with the statements that had the stem "the message I saw," followed by "encouraged me to be an active bystander," "encouraged me to be part of the group that helps" (items that distinguished control messages from the other messages), "suggests very few [university mascot] help others at parties" (refers to readjustment condition messages and combination condition messages), "suggests a lot of [university mascot] help others at parties" (refers to reinforcement condition), "suggests [university mascot] trust their friends" (refers to reinforcement condition and combination condition messages), and "contains tips on how to party safely" (refers to the control messages). These were evaluated based on means and how responses compared to one another.

Open-Ended Items

Participants were asked to indicate what they liked and disliked about the messages, as well as to identify what they believed to be the purpose of the message.

Initial Message Design

The messages were designed on Canva, using available music and visuals. The script was created based on the bystander items described in this paper. The scripts are approximately 10 - 15 words from each other in terms of length. Please refer to Textbox 1 for a complete breakdown of each initial message, the condition that it belongs to, and the number by which the message is referred to in the *Results* section. Figure 1 shows an example of the message.

XSL•FO
RenderX

Textbox 1. Initial messages and the conditions they belong to (message numbers are those referred to in the Results section).

Norm readjustment

Message 1:

BE AN ACTIVE BYSTANDER

Only 48% of [UNIVERSITY MASCOT] have told someone that their drink had been drugged

IT'S ON YOU!

Be part of the group that helps!

Message 2:

BE AN ACTIVE BYSTANDER

Only 84% of [UNIVERSITY MASCOT] have made sure their friend was ok if they seem to be in an uncomfortable situation at a party

IT'S ON YOU!

Be part of the group that helps!

Message 3:

BE AN ACTIVE BYSTANDER

Only 55% of [UNIVERSITY MASCOT] have discouraged a friend from getting someone drunk to have sex with them

IT'S ON YOU!

Be part of the group that helps!

Norm reinforcement

Message 4:

BE AN ACTIVE BYSTANDER

More than 97% of [UNIVERSITY MASCOT] believe their friends would tell someone that their drink had been drugged

IT'S ON YOU!

Be part of the group that helps!

Message 5:

BE AN ACTIVE BYSTANDER

More than 97% of [UNIVERSITY MASCOT] believe that their friends would make sure another friend was ok if they seem to be in an uncomfortable situation at a party

IT'S ON YOU!

Be part of the group that helps!

Message 6:

BE AN ACTIVE BYSTANDER

More than 95% of [UNIVERSITY MASCOT] believe that their friends would discourage a friend from getting someone drunk to have sex with them

IT'S ON YOU!

Be part of the group that helps!

Norm readjustment and reinforcement

Message 7:

BE AN ACTIVE BYSTANDER

More than 97% of [UNIVERSITY MASCOT] believe their friends would tell someone that their drink had been drugged, but only 48% of [UNIVERSITY MASCOT] have done so

IT'S ON YOU!

Be part of the group that helps!

Message 8:

BE AN ACTIVE BYSTANDER

More than 97% of [UNIVERSITY MASCOT] believe that their friends would make sure another friend is ok if they seem to be in an uncomfortable situation at a party, but only 84% of [UNIVERSITY MASCOT] have done so

XSL•FO
RenderX

IT'S ON YOU!

Be part of the group that helps!

Message 9:

BE AN ACTIVE BYSTANDER

More than 95% of [UNIVERSITY MASCOT] believe that their friends would discourage a friend from getting someone drunk to have sex with them, but only 55% of [UNIVERSITY MASCOT] have done so

IT'S ON YOU!

Be part of the group that helps!

Control

Message 10:

BE SAFE AT YOUR NEXT PARTY!

At parties, make sure you know where the nearest exit is in case of an emergency. You never know when you might have to bounce!

PLAN AHEAD!

Safe partying is more fun!

Message 11:

BE SAFE AT YOUR NEXT PARTY!

At parties, make sure you know where you place important belongings like wallet and keys. You never know when you may need to bounce!

PLAN AHEAD!

Safe partying is more fun!

Message 12:

BE SAFE AT YOUR NEXT PARTY!

At parties, make sure you know how you are getting back home. You never know when you may need to bounce!

PLAN AHEAD!

Safe partying is more fun!

XSL•FO
RenderX

Figure 1. Piloted message example.

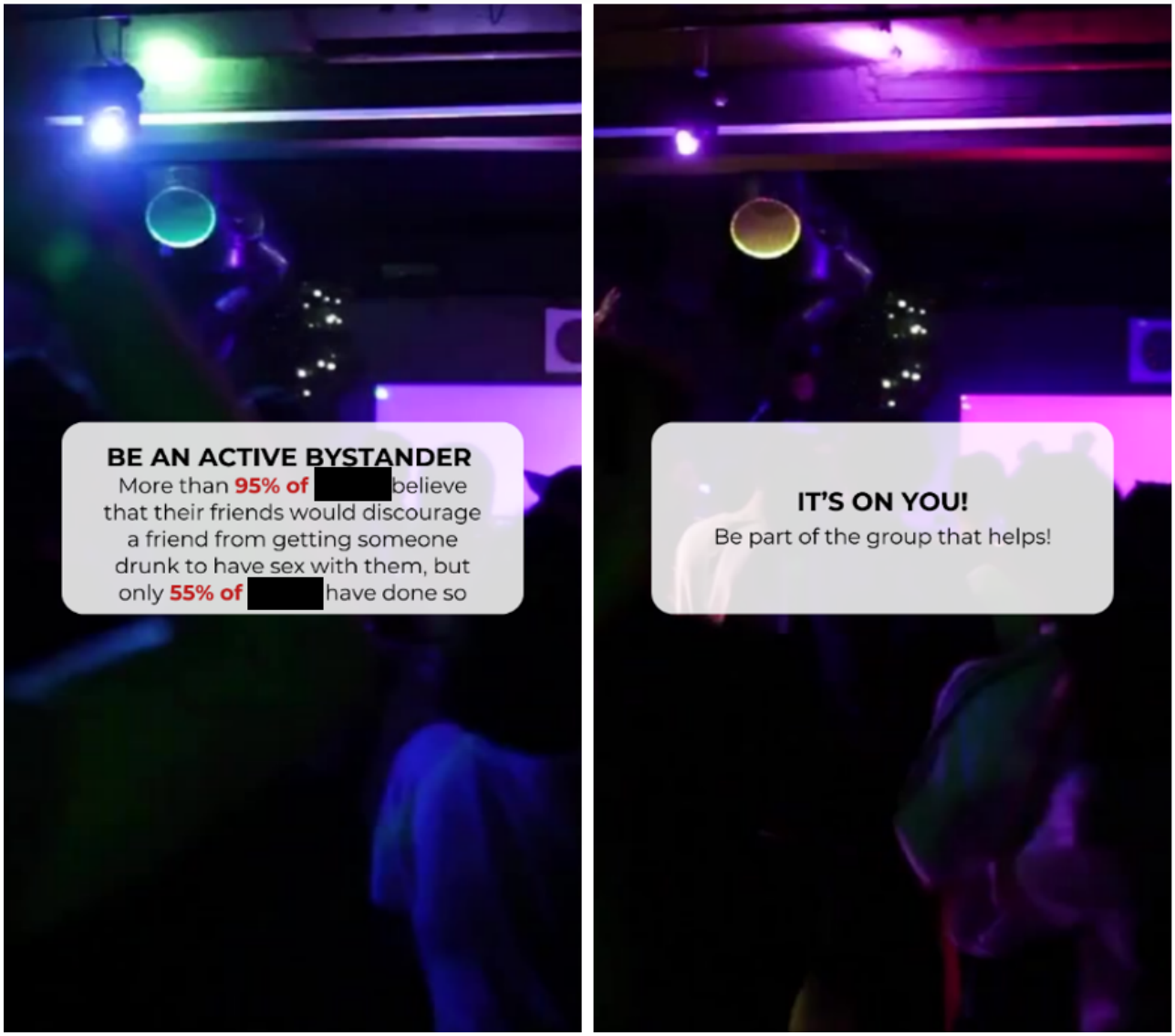

Qualitative Analysis

To evaluate the qualitative open-ended questions, thematic analysis was used [28,29]. Thematic analysis was chosen instead of other potential qualitative methods to allow for a more organic identification of themes and discussion among coders. The coders were the author and a colleague in the same discipline as the author. The first step in Braun and Clarke's [28,29] thematic analysis is to become familiarized with the data. The coders read all the responses before starting to code them. After that, the responses were prepped in a spreadsheet format. The coders highlighted cells based on how they may represent a theme. If a segment may be referring to 2 or more themes, coders were to add a column next to the column where the segments were and highlight the cell next to the original segment. This procedure repeated itself for each question (likes, dislikes, and purpose) and per message (9 messages total). Although 12 messages were tested, the qualitative coding only happened for the messages that were related to bystander intervention and not the control messages (messages 10 - 12). Thus, there is a total of 24 units to analyze, where each has between 28 - 37 segments as responses. A total of 930 segments were coded. Then, we used 1 set of messages to train, where we coded the qualitative responses related to message 1 while we were together, but the coding happened separately. During the training, we established the scope of the themes, and we established that for the priming question, we were coding for 2 themes: participants understood the message, and participants did not understand the message. After coding the segments associated with message 1, we found that we agreed in most instances on the development of themes and how we approached the priming segments. We used the segments associated with messages 1 - 4 to establish the most common themes and used those themes to code the rest of the segments in messages 5 - 9. Thus, approximately 44% of the data were used to establish the themes, and those themes were reinforced using the rest of the data. We then moved on to code the segments associated with messages 2 - 9 by ourselves, and compared the coding of the coder and my coding to assess if there was a need for us to meet again and discuss any major disagreements. Minor disagreements were addressed by email, phone calls, and text messages.

XSL•FO
RenderX

Statistical Analysis

The pilot was assessed using descriptive statistics only, as the goal of this stage of the messages is to check for credibility, perceived audience, manipulation checks, and perceived effects. Means and frequencies were used to make these assessments.

Ethical Considerations

Participants were compensated with the opportunity to enter a drawing to receive a gift card by Tango Cards in the value of US$100 (three available cards), US$50 (six available cards), or US$25 (eight available cards) in the formative survey and with extra credit in a communication course of their choice for their participation in the pilot for most participants, with very few receiving Starbucks gift cards instead. Data collection happened through an online survey and was collected anonymously. The drawing information was completely unlinked from the survey responses. This study was exempt from review by the author's institution. Exempt review means the protocol was reviewed by a member of the university's IRB, but a full-board review was not deemed necessary, as risks were considered minimal.

Results

Qualitative Results

Eye-Catching Statistics and Information

In general, participants indicated that they appreciated how informative the messages were and that statistics were provided. One participant wrote that they liked that "[The message] gives a statistic that catches your attention," and another participant mentioned, "I like the percentages displayed in the video." This was a pattern across all test messages where these were things participants appreciated in the messages.

Production Passed the Vibe Check

What is conceptualized here as production value is both aspects, such as the video chosen and the music in the background, as well as descriptions of "vibes" and entertainment value. Participants indicated that the message was "upbeat" and that they liked the background music often. They also appreciated how short the message was while still providing valuable information. One participant said that "[The message] was simple and effective," and a different participant said that "the background video and music was engaging."

No Source to Back It Up? Participants Disliked the Lack of Source for the Information Provided

The most common theme of dislike toward the message was the lack of source for the statistics provided. Multiple participants expressed disdain or skepticism toward the statistics, which could come from a lack of source but also a reluctance to accept that they were true in relation to fellow college students in their institution. One participant said that they disliked that there was "no source to back it up," while another one explained that they disliked that the message "did not show how they got the statistic."

Maybe a Bit Boring

This theme was developed from participants mentioning that the messaging is "boring" or "not very engaging." This theme is hard to conciliate with the presence of participants who did express they appreciated that the message was short and kept them interested throughout. One participant said that "[The message] is boring and unengaging," and another indicated that "[The message] was boring to watch."

Message Understood!

In general, participants understood what the messages were trying to achieve with their call to action and information. One participant explained that their understanding of the message was that "this message is telling people that [UNIVERSITY MASCOT] intervene when it comes to getting people drunk for sex, and so should they," while another participant indicated that "Be an active bystander" was the main message. This is important information to be taken together with the results from the quantitative analysis described in the following paragraphs.

Quantitative Results

Perceived Message Effectiveness

To assess the quantitative results, the mean of the scales was used in reference to the midpoint of the scale, that is, above the midpoint represented acceptance and below midpoint represented rejection. All bystander-related messages (1-9) were above the midpoint (3 on a 1 - 5 scale, where 1=disagree and 5=agree) in terms of perceived message effectiveness. The bottom 2 means in perceived message effectiveness belonged to 2 control messages: message 10 (mean 2.77, SD 1.36) and message 12 (mean 2.85, SD 1.12). Although all condition messages (1-9) were above the midpoint, the lowest mean among them belonged to message 5 (mean 3.19, SD 1.33), and that was lower than the mean for message 11 (mean 3.26, SD 1.27), which is a control message. The highest means belonged to message 6 (mean 3.94, SD 1.15) and message 9 (mean 3.85, SD 1.18). The results can be seen in their completion in Table 3.

XSL•FO
RenderX

Table . Perceived message effectiveness results.

Condition and message		Number of participants per condition	Mean (SD)
Readjustment			
	1	37	3.46 (1.2)
	2	38	3.56 (1.24)
	3	38	3.76 (1.12)
Reinforcement			
	4	34	3.79 (1.13)
	5	36	3.19 (1.33)
	6	36	3.94 (1.15)
Combination			
	7	36	3.67 (1.29)
	8	37	3.69 (1.24)
	9	35	3.85 (1.18)
Control			
	10	40	2.77 (1.36)
	11	35	3.26 (1.27)
	12	36	2.85 (1.12)

Credibility

When assessing the means for credibility for each message, we see that generally, all messages were above the midpoint (3 on a 1 - 5 scale). The highest was achieved by message 6, at mean 4.26 (SD 0.94), and the lowest were obtained for messages 5 and 11, at mean 3.90 (SD_5 0.96, SD_{11} 0.98). The mean (SD) values are presented in Table 4.

Table . Message credibility results.

Condition and message		Number, n	Mean (SD)
Readjustment			
	1	36	4.20 (0.74)
	2	38	3.95 (0.87)
	3	38	3.82 (1.03)
Reinforcement			
	4	34	4.30 (0.87)
	5	36	3.90 (0.96)
	6	36	4.26 (0.94)
Combination			
	7	36	4.22 (0.94)
	8	37	4.18 (0.79)
	9	35	4.10 (0.97)
Control			
	10	40	4.06 (0.75)
	11	34	3.90 (0.98)
	12	36	4.11 (1)

XSL•FO
RenderX

Message Audiences

Participants were asked to assess what audiences they believed the messages they saw would be most suited to. That was done using a Likert-type scale ranging from 1 (disagree) to 5 (agree). All major intended audiences for the message, college students, [university] students, and people who are 18‐29 years old, had high means in terms of participants perceiving the messages as being appropriate for them as an audience. Among the condition messages, the highest one was Message 7 among college students (mean 4.86, SD 0.42), and the lowest one was Message 3 among people between ages 18‐29 (mean 4.37, SD 1.08). Table 5 shows a full breakdown of results.

XSL•FO
RenderX

Table . Audience perception means and SDs.

Audience, condition, and message			Number, n	Mean (SD)
College students				
	Readjustment			
		1	36	4.78 (0.54)
		2	38	4.58 (0.79)
		3	38	4.68 (0.7)
	Reinforcement			
		4	34	4.79 (0.54)
		5	36	4.67 (0.63)
		6	35	4.77 (0.55)
	Combination			
		7	36	4.86 (0.42)
		8	37	4.73 (0.65)
		9	35	4.77 (0.49)
	Control			
		10	40	4.67 (0.66)
		11	34	4.62 (0.74)
		12	36	4.78 (0.59)
People aged between 18‑29 years				
	Readjustment			
		1	36	4.47 (1.06)
		2	38	4.53 (0.73)
		3	38	4.37 (1.08)
	Reinforcement			
		4	34	4.56 (0.93)
		5	36	4.44 (0.91)
		6	35	4.6 (0.74)
	Combination			
		7	36	4.61 (0.87)
		8	37	4.57 (0.77)
		9	35	4.57 (0.95)
	Control			
		10	40	4.6 (0.78)
		11	34	4.41 (0.86)
		12	36	4.81 (0.4)
[University name] students				
	Readjustment			
		1	36	4.67 (0.76)
		2	38	4.61 (0.76)
		3	38	4.68 (0.7)
	Reinforcement			
		4	34	4.74 (0.71)
		5	36	4.81 (0.47)

XSL•FO
RenderX

Audience, condition, and message		Number, n	Mean (SD)
	6	35	4.74 (0.7)
Combination			
	7	36	4.81 (0.53)
	8	37	4.73 (0.65)
	9	35	4.74 (0.61)
Control			
	10	40	4.65 (0.62)
	11	34	4.38 (0.99)
	12	36	4.67 (0.72)

Manipulation Checks

The goal of the manipulation checks was to ascertain if the message conditions were properly different from each other in what they were trying to achieve. A total of 6 statements were developed to assess different aspects of the messages and check that manipulation efforts are working. Results indicated that, indeed, all message conditions were properly differentiated from each other. For example, the control condition message 11, which had the highest mean among the control messages, had a lower mean in the manipulation check "the message I saw encouraged me to be an active bystander" (mean 3.15, SD 1.46) in a 5-point scale than the lowest noncontrol message (message 3: mean 3.74, SD 1.35). Full results can be found in the Table 6.

XSL•FO
RenderX

Table . Manipulation check descriptive statistics.

Manipulation check, condition, and message		Number, n	Mean (SD)
The message I saw encouraged me to be an active bystander			
Readjustment			
	1	36	4 (1.24)
	2	38	3.82 (1.18)
	3	38	3.74 (1.35)
Reinforcement			
	4	34	4.18 (1.24)
	5	35	4.03 (1.2)
	6	35	4.06 (1.35)
Combination			
	7	36	4.17 (1.23)
	8	37	4.16 (1.07)
	9	34	4.09 (1.26)
Control			
	10	39	2.97 (1.58)
	11	34	3.15 (1.46)
	12	36	2.86 (1.48)
The message I saw encouraged me to be part of the group that helps			
Readjustment			
	1	36	4.5 (0.74)
	2	38	4.08 (1)
	3	38	4.13 (0.99)
Reinforcement			
	4	34	4.41 (0.96)
	5	35	4.37 (0.81)
	6	35	4.63 (0.81)
Combination			
	7	36	4.47 (0.94)
	8	37	4.49 (0.77)
	9	34	4.47 (0.75)
Control			
	10	39	3.05 (1.62)
	11	34	3.65 (1.28)
	12	36	3.17 (1.52)
The message I saw suggests very few [university mascot] help others at parties			
Readjustment			
	1	36	3.94 (1.04)
	2	38	2.95 (1.16)
	3	38	3.84 (1.24)
Reinforcement			
	4	34	3.06 (1.67)
	5	35	3.03 (1.67)

XSL•FO
RenderX

Manipulation check, condition, and message		Number, n	Mean (SD)
	6	35	2.8 (1.71)
Combination			
	7	36	3.97 (1.21)
	8	37	3.38 (1.34)
	9	34	3.53 (1.24)
Control			
	10	39	2.49 (1.3)
	11	34	2.68 (1.36)
	12	36	2.5 (1.4)
The message I saw suggests a lot of [university mascot] help others at parties			
Readjustment			
	1	36	3 (1.24)
	2	38	3.87 (0.84)
	3	38	2.55 (1.13)
Reinforcement			
	4	34	3.97 (1.31)
	5	35	4.23 (1.14)
	6	35	4.49 (0.7)
Combination			
	7	36	2.89 (1.47)
	8	37	3.73 (1.17)
	9	34	3.06 (1.25)
Control			
	10	39	2.59 (1.23)
	11	34	2.97 (1.34)
	12	36	2.67 (1.29)
The message I saw suggests [university mascot] trust their friends			
Readjustment			
	1	36	3.11 (1.17)
	2	38	3.74 (0.86)
	3	38	2.89 (1.23)
Reinforcement			
	4	34	4.32 (0.95)
	5	35	4.37 (0.94)
	6	35	4.43 (0.85)
Combination			
	7	36	3.44 (1.3)
	8	37	4.11 (0.84)
	9	34	3.74 (1.11)
Control			
	10	39	2.69 (1.22)
	11	34	3.41 (1.33)
	12	36	2.83 (1.28)

XSL•FO
RenderX

Manipulation check, condition, and message		Number, n	Mean (SD)
The message I saw contains tips on how to party safely			
Readjustment			
	1	36	3.22 (1.36)
	2	38	3.11 (1.43)
	3	38	2.61 (1.37)
Reinforcement			
	4	34	3.12 (1.45)
	5	35	3.2 (1.28)
	6	35	3.83 (1.29)
Combination			
	7	36	3.06 (1.55)
	8	37	3.19 (1.43)
	9	34	3.15 (1.44)
Control			
	10	39	4.31 (1.22)
	11	34	4.35 (0.98)
	12	36	4.44 (0.84

Discussion

Principal Findings

In this study, the feasibility of a social norms message was assessed by the means of a formative survey study, demonstrating that there was indeed a discrepancy in terms of pluralistic ignorance. The presence of pluralistic ignorance allows for the development of social normative messaging. After establishing the presence of pluralistic ignorance, a message was piloted, and results indicated that the messages reached the intended audience, had the intended perceived effect, were perceived as credible, and the manipulation checks indicated the conditions were properly differentiated from each other while representing their purpose. The feasibility of messages focusing on bystander barriers and combining bystander barriers and the social norms approach was also explored. These messages were also highly accepted among the participants, as well as highly understood. Further research is necessary to explore which theoretical approach is ideal to reach this population in terms of the topic of bystander intervention in sexual assault situations. This was the first study to consider combining opposing approaches to bystander intervention messaging—one where it is suggested that people should consider the behavior of their peers [6] and one where they should take more responsibility for their own actions [7].

Although all 3 approaches showed appropriate levels of perceived message effect and credibility, it is necessary to underscore that the reinforcement condition and the combination condition were particularly successful. This suggests that the bystander barrier associated with taking responsibility may not be as strong of a factor in bystander intervention messaging as it is an attitude to be changed by messaging. The fact that the common denominator among the more successful messages is the presence of normative information is in line with the current literature surrounding sexual violence prevention in the form of media messages [30] as well as theoretical frameworks used to address sexual violence prevention [6,7], further building on it by adding a component that goes beyond norms only. However, more must be explored in terms of combined message by proper experimental designs.

The discrepancy between the qualitative and quantitative results of the pilot in terms of credibility highlighted the need to conduct mixed methods pilots to holistically assess messages that concern public health. Although the quantitative results indicated that credibility was appropriate as is, the qualitative data indicated that there were ways to improve credibility even more. The message should be adjusted before further evaluation of its effects. Possible reasons for this discrepancy could be that, although participants find the message credible overall, as in, the message had face validity to them, they would expect proper sourcing of this information. When asked quantitatively if the messages were credible, they may tend to say it is credible, but when they have the space to expand on that, they explain how this credibility could become more formal.

In terms of practical implications, this study shows the possibility for universities and other organizations that conduct bystander intervention training to provide their publics with messages that can underscore these trainings. It is unlikely that an Instagram reel-based message alone can change one's behavior. However, media messages can serve to reinforce adopted norms [30-32]. This can, in turn, maintain a change in behavior from a more complex intervention that goes beyond media exposure.

XSL•FO
RenderX

Limitations

This study, although novel, is not without limitations. The messages were highly tailored for the students of the university where the study was conducted. Although this is ideal for social normative messaging, it can interfere with the scalability of the study. Furthermore, more assessments using inferential statistics are necessary to establish the potential effects of these messages. These are the next steps in this line of research. In addition, the messages lacked proper sourcing, which participants identified particularly in the qualitative results as a flaw. Future developments of these messages should provide proper sourcing to enhance the credibility of this intervention.

Conclusion

Overall, this study explores how different theories can contribute in different ways to the same public health concern. The messages assessed here have proven to be adequate for the next stage of evaluation with an experimental design. However, this next step would not be as rigorous if the procedures discussed here had not been conducted first.

Data Availability

Data can be made available upon request at the discretion of the author.

Conflicts of Interest

None declared.

Checklist 1
CHERRIES (Checklist for Reporting Results of Internet E-Surveys) checklist of items that apply to this study.
[PDF File, 27 KB - formative_v9i1e66769_app1.pdf]

References

1. Cantor D, Fisher B, Chibnall S, et al. Report on the AAU campus climate survey on sexual assault and sexual misconduct. Westat. 2019. URL: https://cpb-us-w2.wpmucdn.com/voices.uchicago.edu/dist/6/294/files/2019/10/2019-Campus-Climate-Survey-Results.pdf [accessed 2024-12-25]
2. Smith SG, Zhang X, Basile KC, et al. The National Intimate Partner and Sexual Violence Survey: 2015 data brief — updated release. Centers for Disease Control and Prevention. 2018. URL: https://stacks.cdc.gov/view/cdc/60893 [accessed 2025-01-21]
3. Your role in preventing sexual assault. RAINN. URL: https://www.rainn.org/articles/your-role-preventing-sexual-assault [accessed 2023-08-12]
4. Coker AL, Bush HM, Fisher BS, et al. Multi-college bystander intervention evaluation for violence prevention. Am J Prev Med 2016 Mar;50(3):295-302. [doi: 10.1016/j.amepre.2015.08.034] [Medline: 26541099]
5. Kleinsasser A, Jouriles EN, McDonald R, Rosenfield D. An online bystander intervention program for the prevention of sexual violence. Psychol Violence 2015 Jul;5(3):227-235. [doi: 10.1037/a0037393] [Medline: 26240776]
6. Berkowitz AD. An overview of the social norms approach. In: Changing the Culture of College Drinking: A Socially Situated Health Communication Campaign: Hampton Press; 2005:193-214.
7. Burn SM. A situational model of sexual assault prevention through bystander intervention. Sex Roles 2009 Jun;60(11-12):779-792. [doi: 10.1007/s11199-008-9581-5]
8. Orchowski LM, Merrill JE, Oesterle DW, et al. Integrated alcohol use and sexual assault prevention program for college men who engage in heavy drinking: randomized pilot study. JMIR Form Res 2023 Nov 23;7:e47354. [doi: 10.2196/47354] [Medline: 37995129]
9. Pant I, Kang BA, Rimal R. Improving bystander self-efficacy to prevent violence against women through interpersonal communication using mobile phone entertainment education: randomized controlled trial. JMIR Form Res 2023 Apr 14;7:e38688. [doi: 10.2196/38688] [Medline: 37058330]
10. Kimble M, Neacsiu AD, Flack WF, Horner J. Risk of unwanted sex for college women: evidence for a red zone. J Am Coll Health 2008;57(3):331-338. [doi: 10.3200/JACH.57.3.331-338] [Medline: 18980890]
11. McMahon S, Banyard VL. When can I help? a conceptual framework for the prevention of sexual violence through bystander intervention. Trauma Violence Abuse 2012 Jan;13(1):3-14. [doi: 10.1177/1524838011426015] [Medline: 22096017]
12. Coker AL, Cook-Craig PG, Williams CM, et al. Evaluation of green dot: an active bystander intervention to reduce sexual violence on college campuses. Violence Against Women 2011 Jun;17(6):777-796. [doi: 10.1177/1077801211410264] [Medline: 21642269]
13. Green Dot. Culture of Respect. 2014. URL: https://cultureofrespect.org/program/green-dot/ [accessed 2023-08-11]
14. Fischer P, Krueger JI, Greitemeyer T, et al. The bystander-effect: a meta-analytic review on bystander intervention in dangerous and non-dangerous emergencies. Psychol Bull 2011 Jul;137(4):517-537. [doi: 10.1037/a0023304] [Medline: 21534650]
15. Ganti N, Baek S. Why people stand by: a comprehensive study about the bystander effect. J Stud Res 2021;10(1). [doi: 10.47611/jsrhs.v10i1.1390]

XSL•FO
RenderX

16. Lurigio AJ. Crime narratives, dramatizations, and the legacy of the Kitty Genovese murder. Crim Just Behav 2015 Jul;42(7):782-789. [doi: 10.1177/0093854814562954]
17. Coker AL, Bush HM, Cook-Craig PG, et al. RCT testing bystander effectiveness to reduce violence. Am J Prev Med 2017 May;52(5):566-578. [doi: 10.1016/j.amepre.2017.01.020] [Medline: 28279546]
18. Salazar LF, Schipani-McLaughlin AM, Sebeh Y, Nizam Z, Hayat M. A web-based sexual violence, alcohol misuse, and bystander intervention program for college women (RealConsent): randomized controlled trial. J Med Internet Res 2023 Jun 21;25:e43740. [doi: 10.2196/43740] [Medline: 37342080]
19. Fung ICH, Blankenship EB, Ahweyevu JO, et al. Public health implications of image-based social media: a systematic review of Instagram, Pinterest, Tumblr, and Flickr. Perm J 2020;24:18.307. [doi: 10.7812/TPP/18.307] [Medline: 31852039]
20. Pinto PA, Antunes MJL, Almeida AMP. Public Health on Instagram: an analysis of health promotion strategies of Portugal and Brazil. Procedia Comput Sci 2021;181:231-238. [doi: 10.1016/j.procs.2021.01.142]
21. Gandola K. Instagram reels VS instagram stories. Net Influencer. 2022. URL: https://www.netinfluencer.com/instagram-reels-vs-instagram-stories/ [accessed 2023-08-11]
22. Social media fact sheet. Pew Research Center. 2021. URL: https://www.pewresearch.org/internet/fact-sheet/social-media/ [accessed 2023-08-12]
23. Hust SJT, Marett EG, Lei M, et al. Health promotion messages in entertainment media: crime drama viewership and intentions to intervene in a sexual assault situation. J Health Commun 2013;18(1):105-123. [doi: 10.1080/10810730.2012.688241] [Medline: 23020668]
24. Couto L, Willoughby JF. #LoveYourBody: an experimental test of the effects of objectification and body appreciation content on Instagram fitness and health posts among young women. Health Commun 2024;39(11):2298-2306. [doi: 10.1080/10410236.2023.2265647]
25. Willoughby JF, Hust SJT, Li J, Couto L. Social media, marijuana and sex: an exploratory study of adolescents' intentions to use and college students' use of marijuana. J Sex Res 2022 Jan;59(1):85-97. [doi: 10.1080/00224499.2020.1827217] [Medline: 33034543]
26. Baig SA, Noar SM, Gottfredson NC, Boynton MH, Ribisl KM, Brewer NT. UNC perceived message effectiveness: validation of a brief scale. Ann Behav Med 2019 Jul 17;53(8):732-742. [doi: 10.1093/abm/kay080] [Medline: 30321252]
27. Appelman A, Sundar SS. Measuring message credibility: construction and validation of an exclusive scale. J Mass Commun Q 2016;93(1):59-79. [doi: 10.1177/1077699015606057]
28. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
29. Braun V, Clarke V. Conceptual and design thinking for thematic analysis. Qual Psychol 2022;9(1):3-26. [doi: 10.1037/qup0000196]
30. Hust SJT, Kang S, Couto L. Using narrative-based social norms messaging to affect group norms. In: Orchowski LM, DeJong W, editors. Health Behavior and Social Norms: Theory, Research, and Strategies for Change (Forthcoming): American Psychological Association.
31. Hust SJT, Kang S, Couto L, Li J, Rodgers KB. Explaining college men's rape myth acceptance: the role of sports media, masculine norms & fraternity membership. J Health Commun 2023;28(8):477-486. [doi: 10.1080/10810730.2023.2222077]
32. Hust SJT, Marett EG, Lei M, Ren C, Ran W. Law & Order, CSI, and NCIS: the association between exposure to crime drama franchises, rape myth acceptance, and sexual consent negotiation among college students. J Health Commun 2015;20(12):1369-1381. [doi: 10.1080/10810730.2015.1018615] [Medline: 26418170]

Abbreviations

IRB: institutional review board
LGBTQ+: lesbian, gay, bisexual, transgender, queer+
UNC PMES: University of North Carolina's Perceived Message Effectiveness Scale

Edited by A Mavragani; submitted 22.09.24; peer-reviewed by K Fortson, T Vertommen; revised version received 25.11.24; accepted 26.11.24; published 27.01.25.

Please cite as:
Couto L
The Development of an Instagram Reel-Based Bystander Intervention Message Among College Students: Formative Survey and Mixed Methods Pilot Study
JMIR Form Res 2025;9:e66769
URL: https://formative.jmir.org/2025/1/e66769
doi:10.2196/66769

XSL•FO
RenderX

© Leticia Couto. Originally published in JMIR Formative Research (https://formative.jmir.org), 27.1.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Evaluating Older Adults' Engagement and Usability With AI-Driven Interventions: Randomized Pilot Study

Marcia Shade[1*], BS, PhD; Changmin Yan[2*], PhD; Valerie K Jones[2*], PhD; Julie Boron[3*], PhD

[1]College of Nursing, University of Nebraska Medical Center, 985330 Nebraska Medical Center, Omaha, NE, United States

[2]College of Journalism and Mass Communications, University of Nebraska Lincoln, Lincoln, NE, United States

[3]Department of Gerontology, University of Nebraska at Omaha, Omaha, NE, United States

[*]all authors contributed equally

Corresponding Author:
Marcia Shade, BS, PhD
College of Nursing, University of Nebraska Medical Center, 985330 Nebraska Medical Center, Omaha, NE, United States

Abstract

Background: Technologies that serve as assistants are growing more popular for entertainment and aiding in daily tasks. Artificial intelligence (AI) in these technologies could also be helpful to deliver interventions that assist older adults with symptoms or self-management. Personality traits may play a role in how older adults engage with AI technologies. To ensure the best intervention delivery, we must understand older adults' engagement with and usability of AI-driven technologies.

Objective: This study aimed to describe how older adults engaged with routines facilitated by a conversational AI assistant.

Methods: A randomized pilot trial was conducted for 12-weeks in adults aged 60 years or older, self-reported living alone, and having chronic musculoskeletal pain. Participants (N=50) were randomly assigned to 1 of 2 intervention groups (standard vs enhanced) to engage with routines delivered by the AI assistant Alexa (Amazon). Participants were encouraged to interact with prescribed routines twice daily (morning and evening) and as needed. Data were collected and analyzed on routine engagement characteristics and perceived usability of the AI assistant. An analysis of the participants' personality traits was conducted to describe how personality may impact engagement and usability of AI technologies as interventions.

Results: The participants had a mean age of 79 years, with moderate to high levels of comfort and trust in technology, and were predominately White (48/50, 96%) and women (44/50, 88%). In both intervention groups, morning routines (n=62, 74%) were initiated more frequently than evening routines (n=52, 62%; $z=-2.81$, $P=.005$). Older adult participants in the enhanced group self-reported routine usability as good (mean 74.50, SD 11.90), and those in the standard group reported lower but acceptable usability scores (mean 66.29, SD 6.94). Higher extraversion personality trait scores predicted higher rates of routine initiation throughout the whole day and morning in both groups (standard day: B=0.47, $P=.004$; enhanced day: B=0.44, $P=.045$; standard morning: B=0.50, $P=.03$; enhanced morning: B=0.53, $P=.02$). Higher agreeableness (standard: B=0.50, $P=.02$; enhanced B=0.46, $P=.002$) and higher conscientiousness (standard: B=0.33, $P=.04$; enhanced: B=0.38, $P=.006$) personality trait scores predicted better usability scores in both groups.

Conclusions: he prescribed interactive routines delivered by an AI assistant were feasible to use as interventions with older adults. Engagement and usability by older adults may be influenced by personality traits such as extraversion, agreeableness, and conscientiousness. While integrating AI-driven interventions into health care, it is important to consider these factors to promote positive outcomes.

Trial Registration: ClinicalTrials.gov NCT05387447; https://clinicaltrials.gov/study/NCT05387447

(JMIR Form Res 2025;9:e64763) doi:10.2196/64763

KEYWORDS

voice assistant; interventions; usability; engagement; personality; older adults; aging; technology; artificial intelligence; AI; self-management; pilot trial; chronic; musculoskeletal pain; AI assistant; Alexa; user experience; digital health; digital intervention; mobile phone

Introduction

Digital technologies may be beneficial for helping older adults stay connected and directing them to health resources [1]. Newer, technology-based interventions may provide a way to assist with the self-management of chronic illnesses and improve social connections for older adults [2]. Traditional technologies such as computers and mobile tablets have been adopted by aging adults but can be abandoned depending on acceptance and ease of use [3]. Newer technologies such as conversational

XSL•FO
RenderX

voice assistants are a type of artificial intelligence (AI) that have the capacity to assist and have basic interactions with individuals. Older adults can engage with conversational AI through voice, which offers an advantage over spending extensive time learning a new program or manually operating small sized device features. Among Americans aged 18 years and older, 62% use an AI conversational voice assistant on devices such as smart speakers, smartphones, TV remotes, in-car systems, computers, and tablets among others [4]. With the increasing population of older adults, conversational voice assistants may be an optimal modality to deliver interventions.

The use of conversational voice assistants has been studied as an intervention for chronic conditions such as heart failure, lung disorders, and mental health [5]. Their use has been explored for promoting behaviors such as physical activity and for health care support [6,7]. Studies are emerging in relation to conversational voice assistants independently influencing loneliness, social isolation, and pain [8-13]. These studies have primarily focused on acceptability, barriers, and the design of conversational voice assistants for older adult populations. In addition, a recent review of the literature only focused on the use and acceptability of conversational voice assistants to reduce loneliness; none of the conversational voice assistant-based interventions were personalized [13,14]. Kocabelli et al [15] found several conversational voice assistants delivered interventions that were personalized, but lacked theoretical framework or evidence-based approach for implementation. Further exploration is needed to understand older adults' engagement with conversational voice assistants to ensure optimal outcomes when used as a modality for intervention delivery.

An individual's personality can influence acceptance and needs to be considered as a part of usability and the engagement with technology-driven interventions [16]. In younger populations, openness to experience and agreeableness, have been shown to be positively related to usability of technology [17].

Personality traits are characteristics of individuals that typically remain stable over the adult life course. Adults from their mid-to-late 20s and older were found to score similarly across the lifespan on the Big Five traits (ie, openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism [18]. The OCEAN acronym is used to describe these traits. The "O" personality trait of being open to experience is the appreciation of a variety of experiences, "C" conscientiousness is the exhibition of self-discipline, and "E" extraversion is the amount of engagement one has with the external world. The trait of "A" is, agreeableness, which is a concern for social harmony and the "N" is neuroticism trait, which is the tendency of a person to exhibit negative emotions. An individual's personality traits play a role in health care, particularly when implementing patient-centered care and encouraging intervention adherence [19]. Loneliness, which is prevalent in older adults, may influence the presentation of personality traits. A meta-analysis documented negative associations between loneliness and extraversion, agreeableness, conscientiousness, and openness, and a positive association between loneliness and neuroticism [20]. Another study found that loneliness was associated with varying levels of extraversion, neuroticism, and agreeableness [21]. When using technology to deliver interventions in older adults, engagement may differ in those that self-report loneliness; however, this warrants further exploration.

An individual's personality may influence the engagement and user experience with AI technology. Within the realm of traditional technologies, the strongest correlates of older adults' perceived usefulness and ease of use of computers were the personality traits of agreeableness and openness [22,23]. Openness to experience has been associated with increased probability of internet use [16]. Furthermore, higher agreeableness and lower neuroticism in older adults have been associated with perceived usability with an automated vehicle [24].

Since the AI boom, the relationship between the personality traits of older adults and use of conversational voice assistants has not been extensively explored and may impact engagement with AI driven interventions. A commercially available AI voice assistant, that is, Alexa, was used to deliver prescribed interactive routines as nonpharmacological interventions for pain and loneliness. In this analysis of the data, the following research questions were explored:

1. When do older adults initiate the various prescribed routines?
2. What are the older adults' perceived usability scores with different intervention routines delivered by AI through conversational voice assistants?
3. What is the relationship between older adults' personality and their engagement with the different intervention routines?
4. What is the relationship between older adults' personality and the perceived usability of different intervention routines delivered by AI through conversational voice assistants?

Methods

Overview

This study was a 12-week pilot randomized trial conducted with a convenience sample of participants who were 60 years of age or older, self-reported living alone, spoke English, and experienced pain symptoms. A sample size calculation was performed a priori for this pilot study. A sample of 59 participants was calculated to account for 10% study attrition, 80% power, 0.35 effect size, and 0.05 significance. The Mini Cog was used to screen individuals to ensure cognitive ability of participants [25]. Individuals with memory loss or with previous or current use of conversational voice assistants were excluded. Informed consent was obtained prior to initiating study related activities. Once participants were enrolled, they were given a study identification number, and a computerized sorter randomly assigned them to one of the two intervention groups. Participants were compensated for their time during this study. One $25 dollar gift card was given to participants after completing baseline questionnaires, one $25 gift card after completing 12-week questionnaires, and were allowed to keep the Amazon Echo Dot device.

XSL•FO
RenderX

Intervention Groups

Prescribed routines were used for the intervention, such as the frequency of prescription medications (eg, daily, twice daily, as needed). The routines were created based on nonpharmacologic strategies [26] and selected from evidence-based interventions that have a positive influence on pain and loneliness [2,27,28]. After screening and enrollment, participants were randomly assigned to intervention groups.

The 2 intervention groups included the conversational voice assistant standard (CVA-S) routine and the conversational voice assistant enhanced (CVA-E) routine. Standard routines were more generalized and static, while the enhanced routines were tailored, as it has been well documented that individualized or tailored interventions increase engagement and adherence [29-31]. The AI routines in the CVA-S group included a generalized (without personalization) Alexa greeting such as stating, "good morning" or "good evening." The routines consisted of a combination of a generalized (1) meditation, (2) game, (3) music, and (4) jokes. The CVA-S group's "as needed" routine was a general greeting and interactive activity. The AI routines in the CVA-E group included more sociable greetings that were personalized to each participant's name preference such as stating, "hi there, Jim" or "good evening, Betty." The CVA-E routine consisted of a combination of personalized (1) pain location and pain type meditation (2) interactive game preferences, (3) genre or musical artist, and (4) choice of preferred joke style, devotionals, interactive stretching and or other interactive activities.

Routine Programming

The conversational voice assistant "Alexa" was used for this study because the 65+ population prefers Alexa among commercially available voice assistants [32]. Routines were programmed within the Alexa app using available actions and skills created by software app developers. Actions are standard preprogrammed interactions such as playing music or telling a joke, and skills are voice-activated apps such as health fitness, education, games, podcasts, and meditations that can be used with Alexa and the smart speaker. Before skill selection for this study, research personnel reviewed ratings and tested the skills to make sure they worked and would be appropriate for older populations. Routines were initiated by participants speaking to an Amazon Echo Dot device and stating "Alexa, start my routine," which then cued the sequence of selected actions and skills.

Measures and Instruments

Data collected from participants was de-identified and included characteristics such as demographics, health history, and comfort with and trust of technology, assessed on a 4-point scale ranging from 1 (none) to 4 (high). The Big Five Inventory (BFI) was used to measure personality traits of the participants. The BFI contains 44 questions that represent the different traits of openness (Cronbach α=0.77), conscientiousness (Cronbach α=0.72), extraversion (Cronbach α=0.79), agreeableness (Cronbach α=0.78), and neuroticism (Cronbach α=0.70) [18,33].

Feasibility outcomes included capturing engagement with the AI assistant routines and self-reported usability. Engagement with routines was measured at the end of the study. Participants' voice assistant profiles were securely accessed to track the date and time stamps on the incidence of routine initiation with the standard and enhanced loneliness routines. Usability was measured using the System Usability Scale (SUS) administered at the end of the study. A 10-item system usability scale (ranging from 0 to 100) was created (Cronbach α=0.75) [34].

Data Analysis

The data analysis strategy for this study was multifaceted, aimed at understanding the impact of different intervention conditions on routine initiation rates, the influence of personality traits on engagement with Alexa-based routines, and perceptions of system usability. Given the nature of the data and the research questions, the analysis used both descriptive statistics and inferential statistical tests to explore these relationships and test hypotheses. This comprehensive data analysis strategy, incorporating both nonparametric tests for hypothesis testing and regression analysis for exploring relationships between personality traits and key outcomes, facilitated a nuanced understanding of the factors influencing routine initiation and system usability in the context of Alexa-based interventions.

Descriptive statistics in addition to normality tests were conducted. Initial analysis involved descriptive statistics (mean, median, SD, skewness, and kurtosis) to summarize the central tendency and dispersion of routine initiation rates and system usability scores. The Shapiro-Wilk test was used to assess the normality of distributions for these variables. Given the significant results from the Shapiro-Wilk test indicating nonnormal distributions for both routine initiation rates and system usability scores, nonparametric tests were chosen for hypothesis testing.

Nonparametric tests were applied to test the differences in routine initiation rates and SUS scores across conditions (morning vs evening and standard vs enhanced). The Wilcoxon Signed-Rank test was used to compare morning and evening routine initiation rates within participants. Mann-Whitney *U* test was used to test differences between the standard and enhanced conditions in terms of routine initiation rates and SUS scores. This test was chosen due to its suitability for comparing 2 independent samples without assuming normal distribution.

Personality Traits Analysis

Multiple regression analyses were conducted to explore the relationship between participants' personality traits and their engagement with Alexa routines, as well as the perceived usability of the system. These analyses were performed for the total sample and separately for each condition (standard vs enhanced) to determine if personality traits differentially influenced outcomes based on the intervention type. The choice of multiple regression was guided by the interest in assessing the predictive power of multiple independent variables (personality traits) on a continuous dependent variable (number of routine initiations or system usability scores).

Ethical Considerations

The University of Nebraska Medical Center Institutional Review Board approved the study protocol (IRB #0177 - 21-EP).

XSL•FO
RenderX

Results

Demographic findings

The study included 50 participants with mean age 79.24 (SD 7.85; median 78, range 65 - 98). The majority (44/50, 88%) were female (male: 6/50, 12%) and most (48/50, 96%) participants identified as White (African American: 2/50, 44%). Most of the older adults were either widowed (22/50, 44%) or legally divorced or separated (21/50, 42%), while the remaining participants were either living apart from their partner due to health reasons (5/50, 10%) or single (2/50, 4%). Most ((40/50, 80%) participants had completed a college degree or higher, while 16% (8/50) had completed high school, and (2/50, 4%) had attended trade school. The majority (32/50, 64%) reported incomes between US $10,000-$50,000, while 4% (2/50) earned below US $10,000, 14% (7/50) earned between US $50,001-$90,000, and 18% (9/50) earned above US $90,000. Participants reported moderate to high levels of comfort (mean 2.75, SD 0.67) and trust in technology (mean 2.94, SD 0.91) on a scale of 1-5.

Engagement and Perceived Usability of Prescribed Routines

Participants were encouraged to engage with and initiate the routines twice daily for 12 weeks, which equated to 168 interactions in total. As shown in Table 1, the rates of routine initiations were significantly higher among participants in the CVA-E group than the CVA-S group. While observing the prescribed time of day for engagement, the routine initiation rates were higher in the morning.

Table . Comparisons of daily routine initiations of participants in the CVA-S[a] and CVA-E[b] groups in this pilot study of older adults' engagement with 12-week artificial intelligence (AI)-driven interventions.

Condition	Routine initiation (Time of day)	Routine initiation frequency (actual number of interactions/total number of possible interactions, %)	Number of initiations, mean (SD)	*z* score	*P* value
Full sample (n=50)	Morning	62/84[c], 74	62.13 (32.85)	−2.81	.005
Full sample (n=50)	Evening	52/84, 62	51.94 (32.13)		
CVA-E (n=26)	Entire day	126/168[d], 75	126.53 (59.9)	−1.98	.048
CVA-S (n=24)	Entire day	101/168, 60	100.56 (54.88)		
CVA-E (n=26)	Morning	62/84, 74	62.24 (31.86)	−3.16	.002
CVA-S (n=24)	Morning	49/84, 58	49.01 (29.17)		
CVA-E (n=26)	Evening	44/84, 52	52.29 (32.3)	−0.13	.90
CVA-S (n=24)	Evening	44/84, 52	51.55 (32.63)		

[a]CVA-S: conversational voice assistant standard.

[b]CVA-E: conversational voice assistant enhanced.

[c]Number of interactions during either morning and evening.

[d]Total number of interactions during the day.

Given the variances in rates of routine initiations, and a notable preference for morning over evening initiations across conditions, group comparisons were conducted using Wilcoxon analyses to examine the interaction between intervention timing (morning vs evening) and intervention groups (CVA-S vs CVA-E using the Mann-Whitney *U* test). As presented in Table 1, the CVA-E group initiated "morning routines" significantly more often than those in the CVA-S group. In the evening, the rates of routine initiation were not significantly different between the groups.

The perceived usability of the AI assistant showed favorable results. The SUS scores of the overall sample of participants were "good" (mean 70.56, SD 10.58). When conditions were analyzed separately, the CVA-E group's SUS scores were good (mean 74.50, SD 11.90), while the CVA-S group scores were "OK" (mean 66.29, SD 6.94).

Personality and Engagement With Prescribed Routines

As displayed in Figures 1 and 2, the overall model including all Big Five personality traits was not a good fit for predicting of the daily number of routine initiations ($F_{5,44}$=1.45, P=.23, R^2=.14), morning routine initiation ($F_{5,44}$=2.19, P=.07, R^2=.20), or evening routine initiation ($F_{5,44}$=0.72, P=.61, R^2=.08) at α=.05.

However, the extraversion personality trait was significantly associated with routine initiations. It was a significant predictor of overall daily routine initiations (B=0.42, t=2.65, P=.01) and initiation of morning routines (B=0.47, t=3.07, P=.004), but not the evening routines (B=0.28, t=1.73, P=.09). The other 4 personality traits did not predict routine initiations.

XSL•FO
RenderX

Figure 1. Personality trait predictors of routine initiations throughout the day in this pilot study of older adults' engagement with 12-week artificial intelligence (AI)-driven interventions.

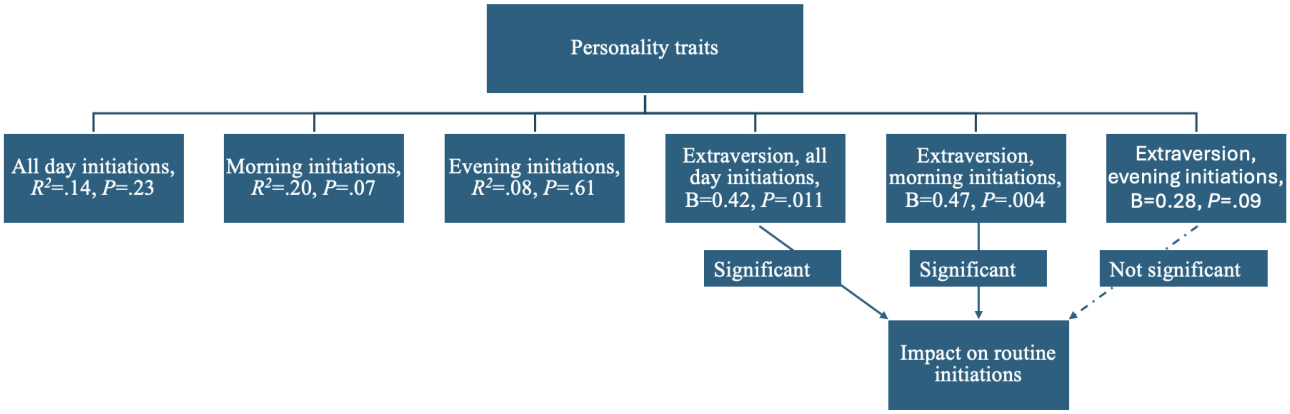

Figure 2. Personality trait predictors of daily routine initiations of participants in the CVA-S (conversational voice assistant standard) and CVA-E (conversational voice assistant enhanced) groups in this pilot study of older adults' engagement with 12-week artificial intelligence (AI)-driven interventions.

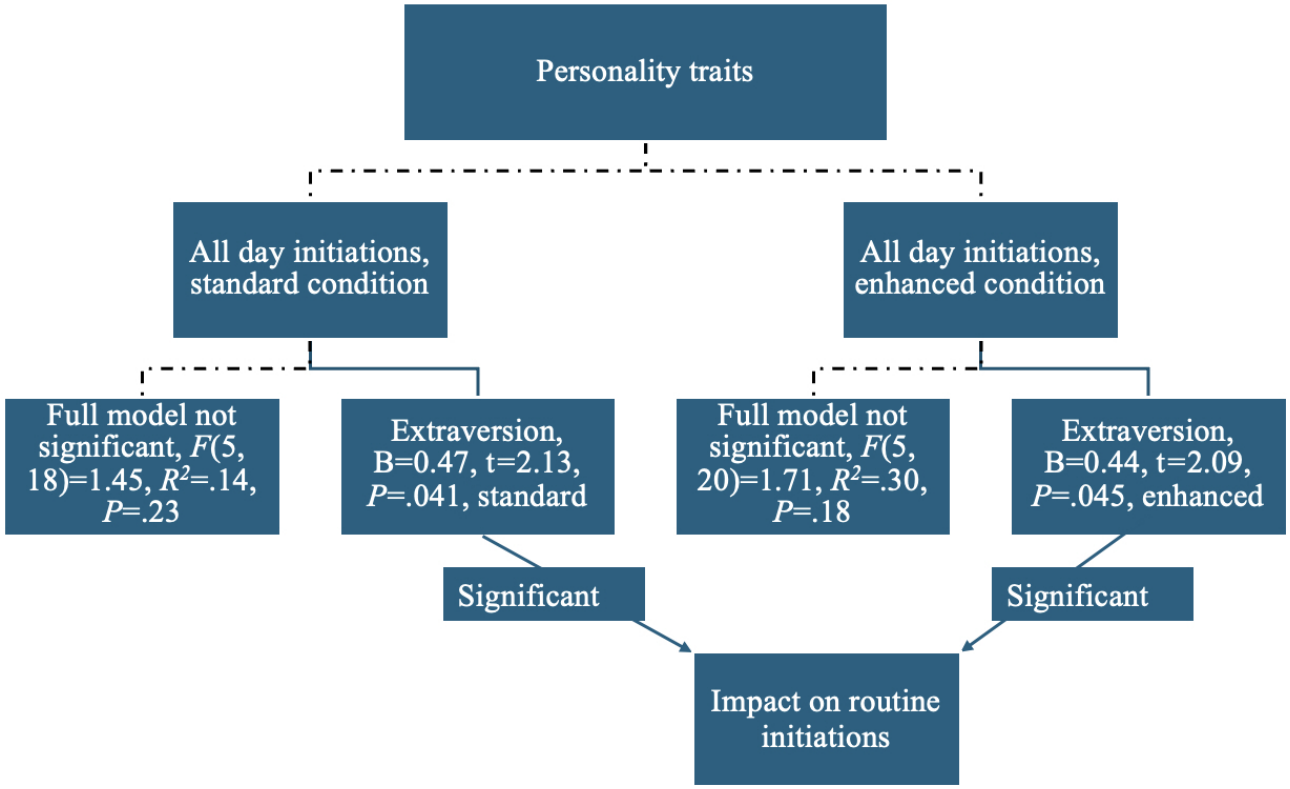

Similar patterns were observed when comparing the CVA-S and CVA-E groups. Extraversion significantly predicted all routine initiations in the CVA-S group, (B=0.47, t=2.13, P=.041 and the CVA-E group, B=0.44, t=2.09, P=.045). As displayed in Figure 3, extraversion was a significant predictor of morning routine initiations in the CVA-S group, B=0.50, t=2.35, P=.03, and the CVA-E group, B=0.53, t=2.54, P=.02. None of the Big Five personality traits significantly predicted evening routine initiations in either group.

XSL•FO
RenderX

Figure 3. Personality trait predictors of morning routine initiations of participants in the CVA-S (conversational voice assistant standard) and CVA-E (conversational voice assistant enhanced) groups in this pilot study of older adults' engagement with 12-week artificial intelligence (AI)-driven interventions.

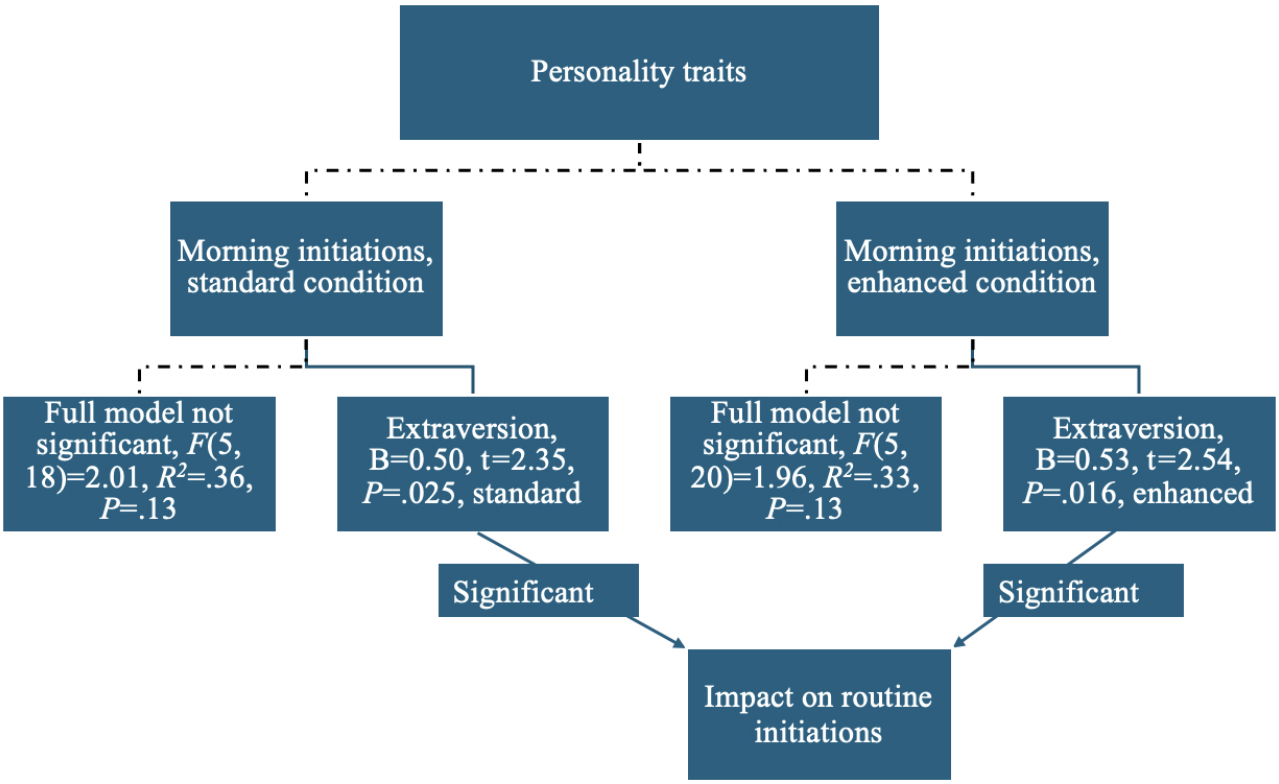

Personality and Perceived Usability of Prescribed Routines

As depicted in Figure 4, among the Big Five personality traits, only agreeableness (B=0.43, t=4.67, P<.001) and conscientiousness (B=0.35, t=3.99, P<.001) positively predicted SUS scores in the overall sample ($F_{5,44}$=29.60, P<.001, R^2=0.77).

As shown in Figure 5, when analyzed separately, both the CVA-S (B=0.50, t=2.49, P=.02) and CVA-E (B=0.46, t=3.31, P=.002) groups reported agreeableness respectively and conscientiousness (B=0.33, t=2.15, P=.04; B=0.38, t=2.97, P=.006 respectively) were positively associated with SUS scores.

XSL•FO
RenderX

Figure 4. Personality trait predictors of participant self-reported usability in this pilot study of older adults' engagement with 12-week artificial intelligence (AI)-driven interventions.

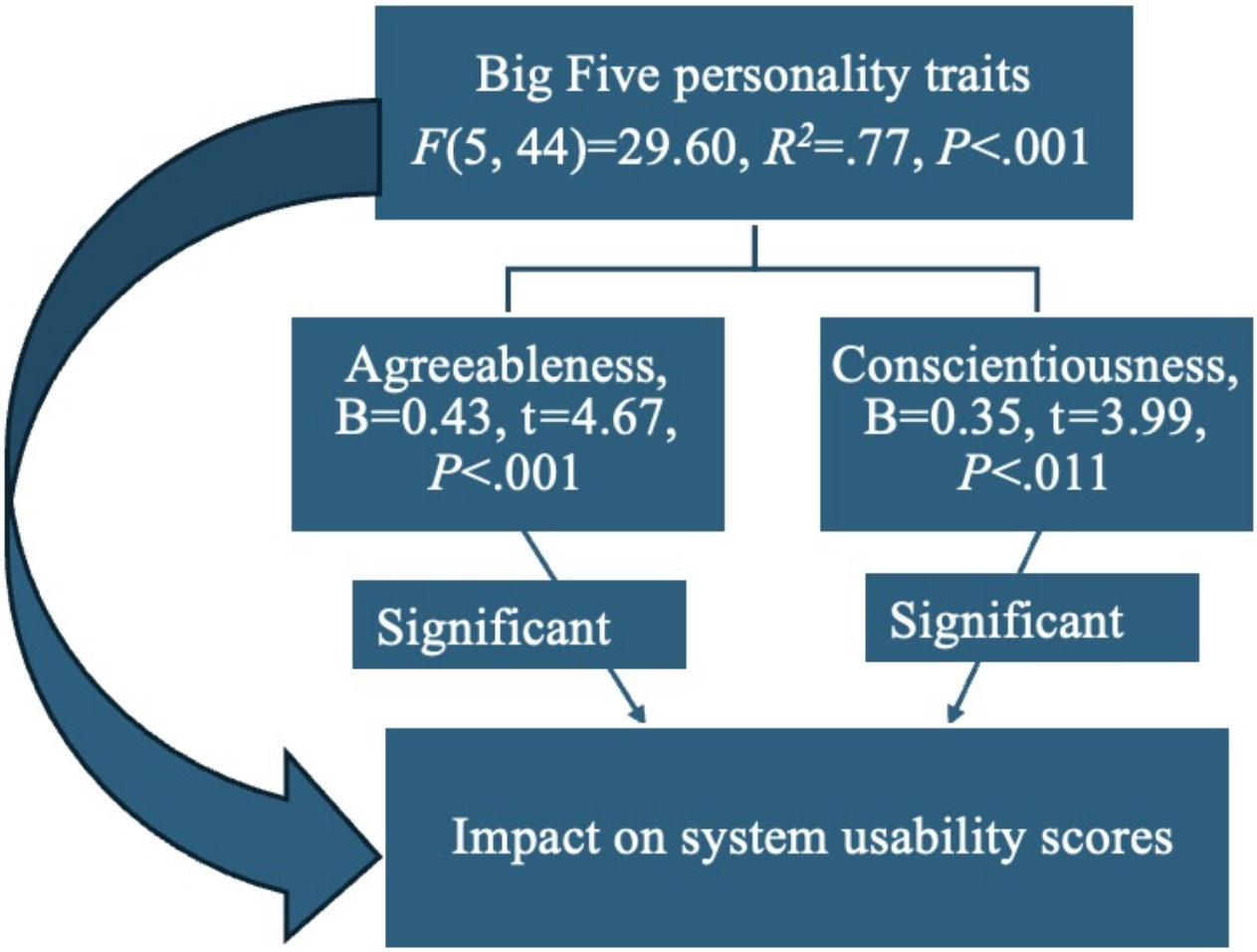

Figure 5. Personality trait predictors of participants self-reported usability in the CVA-S (conversational voice assistant standard) and CVA-E (conversational voice assistant enhanced) groups in this pilot study of older adults' engagement with 12-week artificial intelligence (AI)-driven interventions.

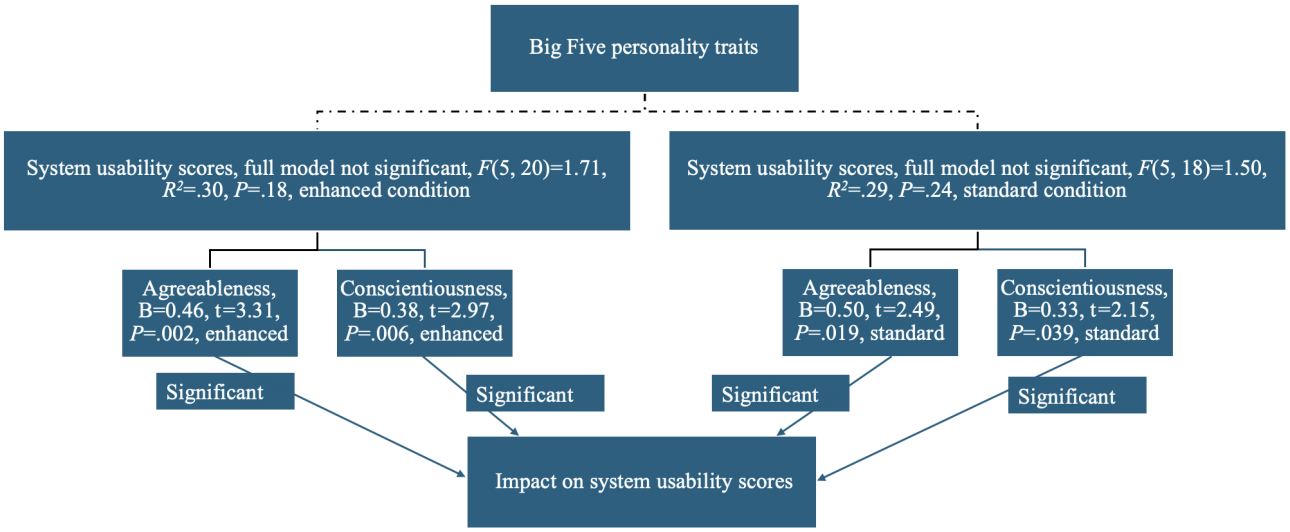

Discussion

Principal Findings

Technology is becoming a popular vehicle for delivering interventions for health care and self-management. In this analysis, we examined older adults' engagement with and usability of AI-delivered routine interventions. Overall, the older participants engaged with the assistant throughout the 12 weeks of the study. Perceived usability was acceptable, with routine initiation rates being higher in the morning. Personality

XSL•FO
RenderX

traits of the participants in this study may have impacted engagement with the prescribed routines. The following discussion provides additional insights into the findings.

Older Adults' Engagement With Prescribed Routines and Perceived Usability

Our study revealed significant differences in routine initiation rates between morning and evening sessions, with participants demonstrating a greater propensity to engage in prescribed routines in the morning. A potential explanation is behavioral habit. Engagement and adherence may be greater in the morning than in the evening because of internal behavioral cues [35]. Another potential explanation for this finding lies within circadian rhythm science. Circadian rhythms are internal processes that regulate the sleep-wake cycle, repeating roughly every 24 hours. They influence physical, mental, and behavioral changes in living organisms. Circadian rhythms are crucial for determining human sleep patterns and have significant implications for designing technology that aligns with these natural processes [36]. Research suggests that morning hours are associated with higher alertness and cognitive functioning, which could facilitate the initiation of health-related routines [37]. The enhanced intervention components may have amplified these effects, indicating that tailored interventions can effectively increase engagement with health-promoting behaviors. The integration of technology with human circadian rhythms or habits may enhance intervention engagement and adherence.

Another possible explanation for the higher initiation rates in the morning could be arthritic pain. Inclusion criteria was self-reported pain symptoms and most of the participants reported a diagnosis of arthritis. Commonly, individuals experience stiffness in the morning that may impact pain symptoms and need for management [38]. In older adults with musculoskeletal pain, technology-based routines or interventions may be best encouraged in the morning.

Older Adults' Personality and Engagement

The results of this study highlight the influence of personality traits on engagement and usability. Extraversion was a significant predictor of routine initiation, particularly in the morning, which is consistent with previous research indicating that extraverted individuals are more likely to engage in health-promoting activities [39]. Extraverts, characterized by their energy, enthusiasm, and sociability, might have the most inherent motivation for social interaction and stimulation. When they perceived the routines as opportunities for engaging experiences or personal growth, it may have made the process of initiating routines more appealing [40,41]. The finding that this effect was more pronounced in the morning suggests that, consistent with their circadian rhythm, extraverts may be prone to engage with technology during early hours, setting a positive tone for the rest of their day.

Consistent with Kortum and Oswald [17], agreeableness was one personality trait that aligned with usability in our sample of older adults. Agreeable and conscientious personality predicted higher perceived usability, suggesting that these traits may influence individuals' perceptions and interactions with technology-based health interventions. It is worth noting the older participants in this study reported moderate to high levels of comfort and trust in technology. Agreeable individuals, known for their cooperative and trusting nature, may be more predisposed to view the Alexa-based intervention in a positive light, attributing any potential usability issues to their own learning curve rather than the system's design flaws. This optimistic viewpoint may lead to higher perceived usability. Conscientious individuals, with their focus on diligence, organization, and reliability, may appreciate systems that are consistent and efficient. The participants' positive evaluation of the system's usability might stem from a recognition of its potential to support their goal-oriented behavior and routine adherence. The structured nature of the preset routines could align well with their preference for orderly environments, further enhancing their perception of the system's usability.

The study reveals a nuanced differentiation between participants' engagement with prescribed routines and their perceptions of system usability, emphasizing the distinct roles of personality traits. Extraversion was found to drive the frequency of routine initiations, particularly in the morning, suggesting that an action-oriented engagement is influenced by a desire for stimulation and interaction. Older adults with extraverted personality traits may be more likely to seek that interaction and treat AI with anthropomorphic properties [42]. Anthropomorphism is when a human attributes human traits to entities or devices that are not human [43]. Researchers have found that people interact with conversational voice assistants in anthropomorphic ways, greeting them, asking them personal questions, exhibiting polite behaviors such as saying "please" and "thank you," and reacting to what the voice assistant did or did not say [12]. Thus, there is a possibility that anthropomorphic interactions with the conversational voice assistants may have impacted engagement and usability of conversational voice assistants to deliver interventions.

Conversely, agreeableness and conscientiousness were associated with more favorable perceptions of system usability, indicating a cognitive evaluation process that values trust, adaptability, and reliability. This distinction underscores the importance of differentiating the motivational drivers behind the use of technology-based health interventions and the cognitive evaluations that shape user satisfaction. By acknowledging the interplay between action-oriented engagement and cognitive evaluations, designers can create more effective, user-friendly interventions tailored to diverse user preferences and personality profiles.

Extraverts, who are stimulated by social interaction and are more action-oriented, particularly in the morning, may benefit from interventions scheduled during this time. Health platforms can incorporate adaptive scheduling features that customize the timing of prompts based on the user's personality profile, thus optimizing the likelihood of engagement and adherence which could impact health outcomes.

Furthermore, understanding the role of personality traits in the acceptance and usability of health technologies can inform the development of more personalized, effective, and user-friendly interventions. The significant role of agreeableness and

XSL•FO
RenderX

conscientiousness in shaping perceptions of system usability suggests that these traits should be considered when designing user interfaces. Users high in agreeableness and conscientiousness value trust, adaptability, and reliability. Therefore, health interventions should prioritize these design elements to cater to their cognitive evaluation processes. For example, features that provide consistent feedback, allow for personalization, and are easy to navigate could be more appealing to these users. Researchers have found that conversational voice assistants exhibit "personality type" traits. The Google and Alexa AI assistants have personality traits of functional intelligence, sincerity, sociability, and creativity [44]. The personality and voice of the assistant has shown to enhance engagement [45]. Future interventions could be tailored for optimal engagement when matching AI assistants and human dominant personalities.

Although the results of this pilot study are informative there are limitations worth noting. There were an even number of participants in each intervention group, however, our sample was relatively small and lacked diversity. The findings are not generalizable to all older adults who report loneliness and pain. Our participants self-reported loneliness and pain, but our study may not have enrolled individuals who were experiencing more intense loneliness and pain. Primary care providers could help us identify older adults with more severe pain and or loneliness. Those without regular internet access in their home due to location (eg, rural areas) or lack of broadband access (due to personal choice or financial restraints) would not be able to interact with conversational voice assistants. Some of these sources of variation could also interact with personality differences. For example, those that choose not to have internet in their homes for personal reasons, may differ from those that do. Furthermore, exploring how to engage those with differing personality characteristics is an important area for future research, so that routines are individualized enough to produce an impact on daily life. Finally, our study was 12-weeks in duration; a longer study would be beneficial to explore the length of time needed to integrate routines that impact well-being and to sustain long term impact.

Conclusions and Implications

Commercially available conversational voice assistants have the potential to deliver nonpharmacologic interventions that are individualized for older adults. Prescribed interactive routines were feasible to use in older adults who self-reported loneliness and pain. Routine engagement may be influenced by the older adults' pain characteristics and personalities. Tailoring these routines to align with individual pain profiles and personality traits may improve the efficacy of interventions and user satisfaction. A personalized approach can enhance adherence and overall outcomes by addressing specific needs and preferences of older adults, making technology-assisted interventions more relevant and efficient in managing their conditions.

Acknowledgments

We acknowledge the role of Victoria Dube, Hyeon Jung Heselton, Taylor Bacon, and our wonderful study participants. This work was supported by Great Plains IDeA-CTR–NIH (NCT05387447).

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Checklist 1
CONSORT guidelines for randomized pilot study.
[PDF File, 94 KB - formative_v9i1e64763_app1.pdf]

References

1. Sen K, Prybutok G, Prybutok V. The use of digital technology for social wellbeing reduces social isolation in older adults: a systematic review. SSM Popul Health 2022 Mar;17:101020. [doi: 10.1016/j.ssmph.2021.101020] [Medline: 35024424]
2. Fakoya OA, McCorry NK, Donnelly M. Loneliness and social isolation interventions for older adults: a scoping review of reviews. BMC Public Health 2020 Feb 14;20(1):129. [doi: 10.1186/s12889-020-8251-6] [Medline: 32054474]
3. Bhattarai P, Newton-John TRO, Phillips JL. Quality and usability of arthritic pain self-management apps for older adults: a systematic review. Pain Med 2018 Mar 1;19(3):471-484. [doi: 10.1093/pm/pnx090] [Medline: 28541464]
4. Serrano KJ, Yu M, Riley WT, et al. Willingness to exchange health information via mobile devices: findings from a population-based survey. Ann Fam Med 2016;14(1):34-40. [doi: 10.1370/afm.1888]
5. Bérubé C, Schachner T, Keller R, et al. Voice-based conversational agents for the prevention and management of chronic and mental health conditions: systematic literature review. J Med Internet Res 2021 Mar 29;23(3):e25933. [doi: 10.2196/25933] [Medline: 33658174]

XSL•FO
RenderX

6. Luo TC, Aguilera A, Lyles CR, Figueroa CA. Promoting physical activity through conversational agents: mixed methods systematic review. J Med Internet Res 2021 Sep 14;23(9):e25486. [doi: 10.2196/25486] [Medline: 34519653]
7. Tudor Car L, Dhinagaran DA, Kyaw BM, et al. Conversational agents in health care: scoping review and conceptual analysis. J Med Internet Res 2020 Aug 7;22(8):e17158. [doi: 10.2196/17158] [Medline: 32763886]
8. Kim S, Choudhury A. Exploring older adults' perception and use of smart speaker-based voice assistants: a longitudinal study. Comput Human Behav 2021 Nov;124:106914. [doi: 10.1016/j.chb.2021.106914]
9. Shade MY, Rector K, Soumana R, Kupzyk K. Voice assistant reminders for pain self-management tasks in aging adults. J Gerontol Nurs 2020 Oct 1;46(10):27-33. [doi: 10.3928/00989134-20200820-03] [Medline: 32852047]
10. Shade MY, Hama RS, Eisenhauer C, Khazanchi D, Pozehl B. "Ask, 'when you do this, how much pain are you in?'": content preferences for a conversational pain self-management software application. J Gerontol Nurs 2023 Jan;49(1):11-17. [doi: 0.3928/00989134-20221205-04]
11. Jones VK, Yan C, Shade MY, et al. Reducing loneliness and improving social support among older adults through different modalities of personal voice assistants. Geriatrics (Basel) 2024 Feb 22;9(2):22. [doi: 10.3390/geriatrics9020022] [Medline: 38525739]
12. Jones VK, Hanus M, Yan C, Shade MY, Blaskewicz Boron J, Maschieri Bicudo R. Reducing loneliness among aging adults: the roles of personal voice assistants and anthropomorphic interactions. Front Public Health 2021;9:750736. [doi: 10.3389/fpubh.2021.750736] [Medline: 34957013]
13. Corbett CF, Wright PJ, Jones K, Parmer M. Voice-activated virtual home assistant use and social isolation and loneliness among older adults: mini review. Front Public Health 2021;9:742012. [doi: 10.3389/fpubh.2021.742012] [Medline: 34708017]
14. Marziali RA, Franceschetti C, Dinculescu A, et al. Reducing loneliness and social isolation of older adults through voice assistants. Lit Rev and Bibl Anal J Med Internet Res 2024;26:e50534. [doi: 10.2196/50534]
15. Kocaballi AB, Berkovsky S, Quiroz JC, et al. The personalization of conversational agents in health care: systematic review. J Med Internet Res 2019 Nov 7;21(11):e15360. [doi: 10.2196/15360] [Medline: 31697237]
16. Özbek V, Alnıaçık Ü, Koc F, Akkılıç ME, Kaş E. The impact of personality on technology acceptance: a study on smart phone users. Proc Soc Behav Sci 2014 Sep;150:541-551. [doi: 10.1016/j.sbspro.2014.09.073]
17. Kortum P, Oswald FL. The impact of personality on the subjective assessment of usability. Int J Hum-Comput Interact 2018 Feb;34(2):177-186. [doi: 10.1080/10447318.2017.1336317]
18. John OP, Srivastava S. The big-five trait taxonomy: history, measurement, and theoretical perspectives. 1999 URL: https://www.academia.edu/13361041/The_Big_Five_trait_taxonomy_History_measurement_and_theoretical_perspectives
19. Redelmeier DA, Najeeb U, Etchells EE. Understanding patient personality in medical care: five-factor model. J Gen Intern Med 2021 Jul;36(7):2111-2114. [doi: 10.1007/s11606-021-06598-8] [Medline: 33506393]
20. Barańczuk U. The Five Factor Model of personality and social support: A meta-analysis. J Res Pers 2019 Aug;81:38-46. [doi: 10.1016/j.jrp.2019.05.002]
21. Schutter N, Koorevaar L, Holwerda TJ, Stek ML, Dekker J, Comijs HC. "Big Five" personality characteristics are associated with loneliness but not with social network size in older adults, irrespective of depression. Int Psychogeriatr 2020 Jan;32(1):53-63. [doi: 10.1017/S1041610219000231] [Medline: 30968789]
22. Mitzner TL, Rogers WA, Fisk AD, et al. Predicting older adults' perceptions about a computer system designed for seniors. Univers Access Inf Soc 2016 Jun;15(2):271-280. [doi: 10.1007/s10209-014-0383-y] [Medline: 31186624]
23. Mitzner TL, Boron JB, Fausset CB, et al. Older adults talk technology: technology usage and attitudes. Comput Human Behav 2010 Nov 1;26(6):1710-1721. [doi: 10.1016/j.chb.2010.06.020] [Medline: 20967133]
24. Voinescu A, Morgan PL, Alford C, Caleb-Solly P. The utility of psychological measures in evaluating perceived usability of automated vehicle interfaces – a study with older adults. Transp Res Part F: Traffic Psychol Behav 2020 Jul;72:244-263. [doi: 10.1016/j.trf.2020.05.003]
25. Borson S, Scanlan JM, Chen P, Ganguli M. The Mini-Cog as a screen for dementia: validation in a population-based sample. J Am Geriatr Soc 2003 Oct;51(10):1451-1454. [doi: 10.1046/j.1532-5415.2003.51465.x] [Medline: 14511167]
26. Ambrose KR, Golightly YM. Physical exercise as non-pharmacological treatment of chronic pain: why and when. Best Pract Res Clin Rheumatol 2015 Feb;29(1):120-130. [doi: 10.1016/j.berh.2015.04.022] [Medline: 26267006]
27. Welsh TP, Yang AE, Makris UE. Musculoskeletal pain in older adults: a clinical review. Med Clin North Am 2020 Sep;104(5):855-872. [doi: 10.1016/j.mcna.2020.05.002] [Medline: 32773050]
28. Arnstein P, Shade M, Herr KA, Young HM, Fishman SM. Managing older adults' chronic pain: lower-risk interventions. AJN Am J Nurs 2023;123(2):46-52. [doi: 10.1097/01.NAJ.0000919740.00088.93]
29. Schubart JR, Stuckey HL, Ganeshamoorthy A, Sciamanna CN. Chronic health conditions and internet behavioral interventions: a review of factors to enhance user engagement. Comput Inform Nurs 2011 Feb;29(2):81-92. [doi: 10.1097/NCN.0b013e3182065eed] [Medline: 21164337]
30. Dworschak C, Heim E, Kuhn N, Schwager J, Tröster A, Maercker A. User-centered development of an internet-based CBT intervention for the treatment of loneliness in older individuals. Internet Interv 2024 Mar;35:100720. [doi: 10.1016/j.invent.2024.100720] [Medline: 38328277]

XSL•FO
RenderX

31. Martorella G, Boitor M, Berube M, Fredericks S, Le May S, Gélinas C. Tailored web-based interventions for pain: systematic review and meta-analysis. J Med Internet Res 2017 Nov 10;19(11):e385. [doi: 10.2196/jmir.8826] [Medline: 29127076]
32. EMARKETER. Siri, Alexa, and Google Assistant Popularity Varies Across US Age Groups. Secondary siri, alexa, and google assistant popularity varies across US age groups 2024. 2021. URL: https://www.emarketer.com/content/siri-alexa-google-assistant-popularity-varies-across-us-age-groups
33. Goldberg LR. The structure of phenotypic personality traits. Am Psychol 1993 Jan;48(1):26-34. [doi: 10.1037//0003-066x.48.1.26] [Medline: 8427480]
34. Brooke J. System usability scale (SUS): a quick-and-dirty method of system evaluation user information reading, UK: digital equipment co ltd U10 - exported. 1986 URL: https://www.researchgate.net/publication/228593520_SUS_A_quick_and_dirty_usability_scale
35. Phillips LA, Burns E, Leventhal H. Time-of-day differences in treatment-related habit strength and adherence. Ann Behav Med 2021 Mar 20;55(3):280-285. [doi: 10.1093/abm/kaaa042] [Medline: 32542355]
36. Roenneberg T, Kuehnle T, Juda M, et al. Epidemiology of the human circadian clock. Sleep Med Rev 2007 Dec;11(6):429-438. [doi: 10.1016/j.smrv.2007.07.005] [Medline: 17936039]
37. Valdez P. Circadian rhythms in attention. Yale J Biol Med 2019 Mar;92(1):81-92 [FREE Full text] [Medline: 30923475]
38. COBB S, MERCHANT WR, RUBIN T. The relation of symptoms to osteoarthritis. J Chronic Dis 1957 Feb;5(2):197-204. [doi: 10.1016/0021-9681(57)90135-2] [Medline: 13405980]
39. Rizzuto D, Mossello E, Fratiglioni L, Santoni G, Wang HX. Personality and survival in older age: the role of lifestyle behaviors and health status. Am J Geriatr Psychiatry 2017 Dec;25(12):1363-1372. [doi: 10.1016/j.jagp.2017.06.008] [Medline: 28711464]
40. Chapman BP, Duberstein PR, Sörensen S, Lyness JM. Personality and perceived health in older adults: the five factor model in primary care. J Gerontol B Psychol Sci Soc Sci 2006 Nov;61(6):362-365. [doi: 10.1093/geronb/61.6.p362] [Medline: 17114306]
41. Duberstein PR, Sörensen S, Lyness JM, et al. Personality is associated with perceived health and functional status in older primary care patients. Psychol Aging 2003 Mar;18(1):25-37. [doi: 10.1037/0882-7974.18.1.25] [Medline: 12641310]
42. Kaplan AD, Sanders T, Hancock PA. The relationship between extroversion and the tendency to anthropomorphize robots: A Bayesian analysis. Front Robot AI 2018;5:135. [doi: 10.3389/frobt.2018.00135] [Medline: 33501013]
43. Epley N, Waytz A, Cacioppo JT. On seeing human: a three-factor theory of anthropomorphism. Psychol Rev 2007 Oct;114(4):864-886. [doi: 10.1037/0033-295X.114.4.864] [Medline: 17907867]
44. Poushneh A. Humanizing voice assistant: The impact of voice assistant personality on consumers' attitudes and behaviors. J Retail Consum Serv 2021 Jan;58:102283. [doi: 10.1016/j.jretconser.2020.102283]
45. Zhong R, Ma M. Effects of communication style, anthropomorphic setting and individual differences on older adults using voice assistants in a health context. BMC Geriatr 2022 Sep 15;22(1):751. [doi: 10.1186/s12877-022-03428-2] [Medline: 36104775]

Abbreviations

AI: artificial intelligence
BFI: Big Five Inventory
CVA-E: conversational voice assistant enhanced
CVA-S: conversational voice assistant standard
SUS: System Usability Scale

Edited by A Mavragani; submitted 25.07.24; peer-reviewed by E Ross, N Hird; revised version received 20.11.24; accepted 26.11.24; published 24.01.25.

Please cite as:
Shade M, Yan C, Jones VK, Boron J
Evaluating Older Adults' Engagement and Usability With AI-Driven Interventions: Randomized Pilot Study
JMIR Form Res 2025;9:e64763
URL: https://formative.jmir.org/2025/1/e64763
doi:10.2196/64763

© Marcia Shade, Changmin Yan, Valerie K Jones, Julie Boron. Originally published in JMIR Formative Research (https://formative.jmir.org), 24.1.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete

XSL•FO
RenderX

bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Characterizing Telehealth Barriers and Preferences to Promote Acceptable Implementation Strategies in Central Uganda: Multilevel Formative Evaluation

Michael Kizito[1*], BSCS, MSCS, PhD; Erina Nabunjo Mugabi[2*], BEd, MBA; Sabrina Ford[3*], BScPT, PhD; Bree Holtz[4*], BA, MSIS, PhD; Kelly Hirko[5], BScB, MPH, PhD

[1]Department of Computer Science, Makerere University Kampala, Kampala, Uganda
[2]Department of Health Policy Planning and Management, School of Public Health, Makerere University Kampala, Kampala, Uganda
[3]College of Human Medicine, Michigan State University, East Lansing, MI, United States
[4]Department of Advertising & PR, College of Communication Arts & Sciences, Michigan State University, East Lansing, MI, United States
[5]Department of Epidemiology and Biostatistics, College of Human Medicine, Michigan State University, East Lansing, MI, United States
[*]these authors contributed equally

Corresponding Author:
Michael Kizito, BSCS, MSCS, PhD
Department of Computer Science, Makerere University Kampala, Kampala, Uganda

Abstract

Background: Telehealth approaches can address health care access barriers and improve care delivery in resource-limited settings around the globe. Yet, telehealth adoption in Africa has been limited, due in part to an insufficient understanding of effective strategies for implementation.

Objective: This study aimed to conduct a multi-level formative evaluation identifying barriers and facilitators for implementing telehealth among health service providers and patients in Central Uganda.

Methods: We collected surveys characterizing telehealth perceptions, barriers, and preferences from health care providers and patients seeking primary care in the Central Region of Uganda from January 2022 to July 2022. Survey development was informed by the technology acceptance model and evaluated predictors of technology acceptance (ie, perceived usefulness, ease of use, and attitudes). We used descriptive statistics to characterize telehealth perceptions and examined differences according to provider and patient characteristics using Student *t* tests.

Results: Nearly 79% (n=48) of 61 providers surveyed had used telehealth, and perceptions were generally favorable. While 93.4% (n=57) reported that telehealth adds value to clinical practice, less than half (n=30, 49.2%) felt telehealth was more efficient than in-person visits. Provider-reported barriers to telehealth included technology challenges for the patient (34/132, 26%), low patient engagement (25/132, 19%), and lack of implementation support (24/132, 18%). Telehealth use was lower among the 91 surveyed patients, with only 19.8% (n=18) having used telehealth. Although 89% (n=81) of patients reported saving time with telehealth approaches, 33.3% (n=30) of patients reported that telehealth made them feel uncomfortable, and 43.8% (n=39) reported concerns about confidentiality. Over 72% (n=66) of patients who had used telehealth previously reported satisfaction with the telehealth services they received. Several differences in perceptions of telehealth according to patient's self-reported health status were observed.

Conclusions: Perceptions of telehealth were generally favorable, although higher among providers than patients. Barriers impeding telehealth use include technology challenges and the lack of infrastructure and implementation support. Findings from this study can inform the implementation of acceptable telehealth approaches to address disparities propagated by health care access barriers in Sub-Saharan Africa.

(JMIR Form Res 2025;9:e60843) doi:10.2196/60843

KEYWORDS

telehealth; telemedicine; health care; disparities; technology; barriers; resource-limited; preferences; Uganda; Africa; barrier; formative evaluation; health service provider; primary care; satisfaction; Sub-Saharan Africa; survey; utility

XSL•FO
RenderX

Introduction

Access to quality health care services is critical to achieving optimal health outcomes and reducing health disparities [1]. Individuals residing in resource-limited settings often encounter various economic, social, cultural, and geographic barriers to health care, including provider shortages, inadequate health insurance coverage, transportation challenges, and limited health literacy [2-4]. Health care access barriers are especially stark in many countries in Sub-Saharan Africa, which have 0.23 doctors for every 10,000 people, compared to the highest ratio of 84.2 doctors in the most high-income countries [5]. Increasing access to quality health care and preventive services via telehealth is a promising strategy to reduce health disparities in low- and middle-income countries.

Telehealth refers broadly to the remote delivery of health care services using electronic and telecommunications technology, including voice, chat, and video-based platforms. There are numerous types of telehealth applications, such as live video conferencing for the provision of clinical services to patients (ie, telemedicine), provider-to-provider health care consultation, and remote delivery of nonclinical services, including the use of mobile health (mHealth) tools to support health promotion and wellbeing efforts [6]. Telehealth approaches offer a practical and often cost-effective method to expand the reach of health care services and address health care access barriers in resource-limited settings [7]. Despite the widespread adoption of mobile phones and other digital tools that can support telehealth applications, and the acknowledged potential of using these technologies to improve health care service delivery [8], telehealth adoption in clinical settings remains low. This is particularly true in health care professional shortage areas, which arguably stand to benefit most from these technologies. Indeed, results from a recent review demonstrate the promise of telemedicine approaches in revolutionizing health care delivery in Africa by improving access, efficiency, and patient outcomes [9]. Yet, significant challenges related to the effective implementation and adoption of telehealth remain.

The COVID-19 pandemic highlighted the need for innovative, cost-effective, technology-enabled systems for health care delivery and resulted in a rapid increase in telehealth adoption worldwide. Indeed, the rise in innovative telehealth start-up companies is beginning to transform health care access across Africa, despite concerns about data privacy, inadequate infrastructure, and a lack of policy frameworks [9,10]. For example, Rocket Health based in Kampala, Uganda [11], offers telehealth services as a licensed and registered clinic, laboratory, and pharmacy, including doctor consultations through phone calls and SMS text messages, as well as laboratory sample pickups and medicine deliveries. However, these services can be costly and may not be accessible to those residing in rural areas of Africa with limited internet access. Thus, continued efforts are urgently needed to ensure the broad and equitable reach of telehealth applications in Africa.

An insufficient understanding of effective strategies for implementation has limited telehealth adoption in under-resourced settings. Indeed, very little is known about specific challenges and preferences for telehealth program implementation in African countries, particularly in Uganda [12]. This dearth of knowledge severely limits the ability to implement effective and appropriate telehealth solutions to improve health care access in Africa equitably. Thus, the purpose of this study was to conduct a multi-level formative evaluation identifying barriers and facilitators for implementing telehealth among health service providers and patients in Central Uganda.

Methods

Study Population

The study population included health service providers (ie, doctors, nurses, midwives, and lab technicians) and patients seeking health care services at regularly scheduled clinic visits between January and July 2022 at the following health care facilities in the Central Region of Uganda: Namulonge Health Centre III, Wattuba Health Centre III, Kasangati Health Centre IV, and Buwambo Health Centre IV. Uganda's health care system works on a referral basis and includes small village health teams, health centers (II, III, and IV), district general hospitals, and regional and national referral hospitals, with more intensive services provided at larger centers or hospitals [13]. A health center III facility has approximately 18 staff, including a senior clinical officer, and provides laboratory services and a general outpatient clinic and maternity ward. Health centers IV are small hospitals, with all of the services available at health center III, along with additional medical providers, the ability to admit patients, and a theater for surgical procedures. Although phone and internet access is widespread in Uganda's health center IV facilities, telemedicine integration is minimal, and technology has been used largely in referral hospital settings to initiate referrals and consultations and for knowledge sharing between providers [14]. Of note, the adoption of telecommunication technologies in Uganda is relatively low, with mobile phone coverage at approximately 60% and mobile internet subscriptions at around 9% in 2016, and primarily among urban residents [15]. Although marked variations in internet coverage across different regions in Uganda are evident, the vast majority of the population in Central Uganda has access to at least 2G networks [16].

Study Design

We used a cross-sectional study design engaging providers and patients through surveys using a convenience sampling approach. Inclusion criteria required participants to be providers at the study sites or patients receiving health care at the study sites at the time of study recruitment. Study participants were not selected based on prior telehealth use. Recruitment was facilitated through partnership with the Wakiso District Health Officer (DHO). The research team met with the DHO to describe the study objectives, and the DHO drafted a letter of Introduction to the Health centers for the research team to initiate contact with health centers in the study region. The research team then met with the medical superintendents at health centers in the Central Region of Uganda (Kasangati HCIV, Namulonge Health Centre III, Wattuba Health Centre III, and Buwambo Centre HCIV) to provide the introductory letter from the DHO and to

XSL•FO
RenderX

schedule site visits with the designated nurse at each site to determine the best approaches for distributing surveys at each site. Research team members distributed paper copies of the patient and provider surveys at each site visit with the designated nurse. Surveys were distributed to all patients during health care encounters at the health centers during the research team's site visit. The designated nurse at each site helped distribute and collect the surveys from providers. Through our existing contacts, digital surveys were also emailed to service providers from academia, health IT professionals, and the Infectious Diseases Institute (IDI) from January through June 2022. Given the descriptive nature of this cross-sectional study, we did not conduct power analyses but aimed to have a minimum sample of 50 participants in each group to ensure sufficient data for descriptive analyses.

Ethical Considerations

This study was reviewed for human subject research ethics and was approved as exempt by the Institutional Review Board at Michigan State University (Study ID 00006715). An informed consent waiver, including a description of the study purpose and contact information for the study investigators, was provided on the study surveys. The consent waiver on the surveys specified that participation was voluntary and that responses were confidential. Surveys were anonymous and did not collect any personally identifiable information. Participants were not compensated for survey completion.

Data Collection

Study surveys were developed for patients and providers. The patient survey questions were based on the validated Service User Technology Acceptability Questionnaire, which measures user perceptions about the acceptability of telehealth in domains of enhanced care, increased acceptability, privacy and discomfort, care personnel concerns, telehealth as a substitution, and satisfaction [17]. The provider survey was based on a questionnaire developed by the American Medical Association to assess telehealth preferences and impact on clinical practice, including perceptions of clinical value and efficiency of telehealth services [18]. This modified questionnaire was deployed in our previous study assessing telehealth satisfaction among providers in rural Northwest Michigan, United States of America [19]. Informed by the Technology Acceptance Model [20], the surveys also assessed the following factors influencing an individual's intention to use new technology: perceived usefulness and perceived ease of use. Survey respondents provided perceptions of telehealth using a 5-point Likert scale, where 1 indicated "strongly agree" and 5 indicated "strongly disagree." Additionally, the survey ascertained demographic information, including age range (<30, 30 - 49, 50 - 65, and 65+ years), gender, length of time in practice (providers only), and self-reported health status (patients only). The survey also asked whether respondents had heard about telemedicine and used telemedicine prior to completing the survey. The provider survey also ascertained information on barriers to telehealth using the question: "What are your biggest challenges, if any, that you face regarding telehealth visits? Select all that apply." The following options were available for selection, lack of reimbursement, licensure, technology challenges for the patients, technology challenges for the health care provider/practice, low patient engagement, lack of implementation support, no challenges, or other. This list was based on a prior review of barriers to telehealth in Sub-Saharan Africa [21], with reimbursement issues reflecting barriers due to differential financial return for telehealth visits and licensure indicating legal and regulatory barriers to telehealth implementation. Survey data were entered into Qualtrics, either directly by participants completing the online survey or by study staff for paper-based surveys.

Statistical Analysis

The survey data were downloaded from Qualtrics and collated in an Excel spreadsheet for analysis. We conducted descriptive statistics of the survey items, describing participant characteristics, perceptions of telehealth, and barriers to using telehealth among providers. Descriptive statistics were calculated to describe provider-reported barriers to telehealth. We also assessed whether telehealth perceptions varied according to participant age, and previous telehealth use, length of time practicing medicine (providers only), and health status (patients only) using Student *t* tests. In secondary analyses, we assessed patient perceptions among the subset of individuals who had used telehealth previously. Missing data were excluded from analysis. All statistical analyses were performed using SAS (version 9.4; SAS Institute) and statistical significance was defined at $P<.05$.

Results

Surveys from 61 providers and 91 patients were completed and included in the analysis. For providers, 41 surveys were completed on paper and 20 were completed online. For patients, all 91 surveys were completed on paper. As shown in Table 1, most providers were between the ages of 30 - 49 years (n=43, 71.1%) and reported female gender (n=38, 63.3%). Providers varied in the length of time in practice, with 20% (n=12) of respondents practicing less than 5 years, 43.3% (n=25) practicing between 5 and 10 years, and 36.7% (n=22) practicing more than 10 years. Most of the surveyed providers had heard about telemedicine before (n=42, 68.9%) and used telemedicine previously (n=48, 78.7%). Of note, several providers (n=6) noted that they had not heard of telemedicine previously but reported using telemedicine after reading the description that telemedicine included telephone-based visits. Patient survey respondents were younger than providers, with 57.8% (n=52) of patients under the age of 30 years. Most patient participants were female (n=67, 74.4%) and reported health status as "Excellent" or "Very good" (n=26, 28.9% and n=23, 25.6%, respectively). Compared to providers, a lower percentage of patients had heard of telemedicine (n=32, 35.2%) or used telemedicine previously (n=18, 19.8%).

XSL•FO
RenderX

Table . Characteristics of provider and patient survey respondents. Missing data on age, gender, length of time in practice and self-reported health status for 1 patient survey.

	Providers (n=61), n (%)	Patients (n=91), n (%)
Age, years		
<30	10 (16.7)	52 (57.8)
30 - 49	43 (71.7)	22 (24.4)
50 - 65	6 (10)	10 (11.1)
>65	1 (1.7)	6 (6.7)
Gender		
Female	38 (63.3)	67 (74.4)
Male	22 (36.7)	23 (25.6)
Length of time in practice, years		
<5	12 (20)	—[a]
5 - 10	25 (43.3)	—
11 - 20	16 (26.7)	—
>20	6 (10.0)	—
Self-reported health status		
Excellent	—	26 (28.9)
Very good	—	23 (25.6)
Good	—	17 (18.9)
Fair	—	20 (22.2)
Poor	—	4 (4.4)
Heard about telemedicine previously		
Yes	42 (68.9)	32 (35.2)
No	19 (31.2)	59 (64.8)
Used telemedicine previously		
Yes	48 (78.7)	18 (19.8)
No	13 (21.3)	73 (80.2)

[a]Not applicable

As shown in Figure 1, provider perceptions of telehealth were generally positive, with 96.7% (n=59) agreeing or strongly agreeing that telehealth could help them monitor patients more rapidly and 98.4% (n=60) reporting that they could easily learn how to use telehealth. While 93.4% (n=57) reported that telehealth adds value to clinical practice, less than half (n=30, 49.2%) felt that telehealth was more efficient than in-person visits. Most providers (n=56, 91.8%) reported feeling comfortable with information and communication technologies, but only 49.2% (n=30) thought their health center had the necessary infrastructure to support telehealth use. Indeed, 83.3% (n=51) of providers reported that telehealth use would imply major changes in their clinical practice. Finally, all 48 providers who had used telehealth would recommend telehealth services to others, and all 13 providers who had not used telehealth would be interested in using telehealth services in the future.

XSL•FO
RenderX

Figure 1. Provider perceptions of telehealth from a cross-sectional survey conducted in Central Uganda, January 2022-July 2022 (n=61).

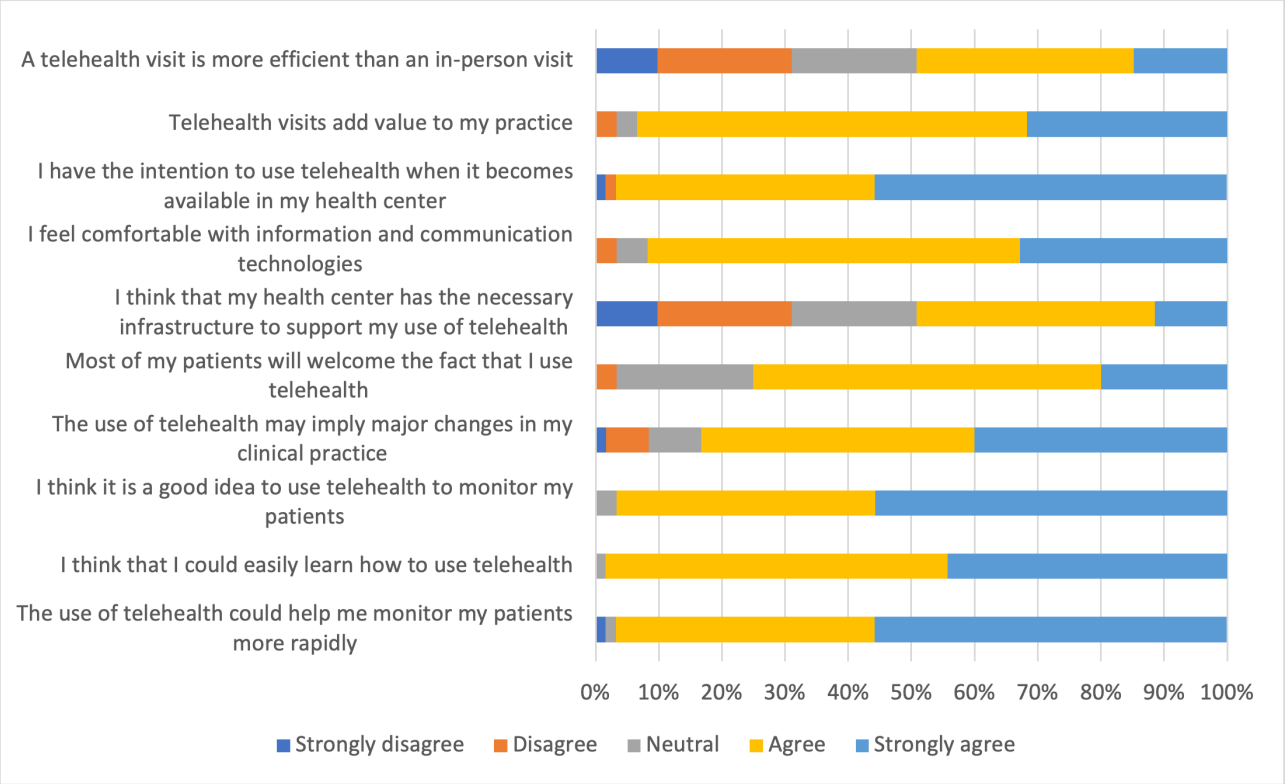

Perceptions of telehealth were less favorable among patients compared to providers (Figure 2). Nearly 78% (n=70) agreed or strongly agreed that telehealth services should be recommended to people in a similar condition, and 76% (n=68) agreed or strongly agreed that telehealth services would allow their caretakers to better monitor their health condition. Although approximately 33% (n=30) of patient survey respondents agreed or strongly agreed that telehealth services have made them feel uncomfortable, and 43.8% (n=39) agreed or strongly agreed that telehealth services make them worried about the confidentiality of private information. Roughly half of the respondents (n=46, 51.1%) did not agree that telehealth could replace regular care, and 43.4% (n=39) agreed or strongly agreed that telehealth services are not as suitable as regular face-to-face consultations. In secondary analyses among the subset of patients who had used telehealth previously (n=18), favorable perceptions were more pronounced, with nearly 90% (n=82) reporting that telehealth services increased their access to care, 89% (n=81) reporting that telehealth saved time, and 88% (n=80) reporting that telehealth services have helped them improve their health. Over 72% (n=66) of patients who had used telehealth previously reported satisfaction with the telehealth services they received. All 18 patients who had used telehealth would recommend telehealth services to others. However, only 27.8% (n=25) of those who had used telehealth agreed or strongly agreed that telehealth services allowed them to be less concerned about their health status (data not shown). Of the 73 patients who had not used telehealth, 60 (82.2%) would be interested in using telehealth services in the future.

Provider-reported barriers to telehealth, shown in Table 2, included technology challenges for the patient (34/132, 26%), low patient engagement (25/132, 19%), and lack of implementation support (24/132, 18%). Lack of reimbursement (19/132, 14%) and technology challenges for the health care provider or practice (19/132, 14%) were also noted.

XSL•FO
RenderX

Figure 2. Patient perceptions of telehealth from a cross-sectional survey conducted in Central Uganda, January 2022-July 2022 (n=91).

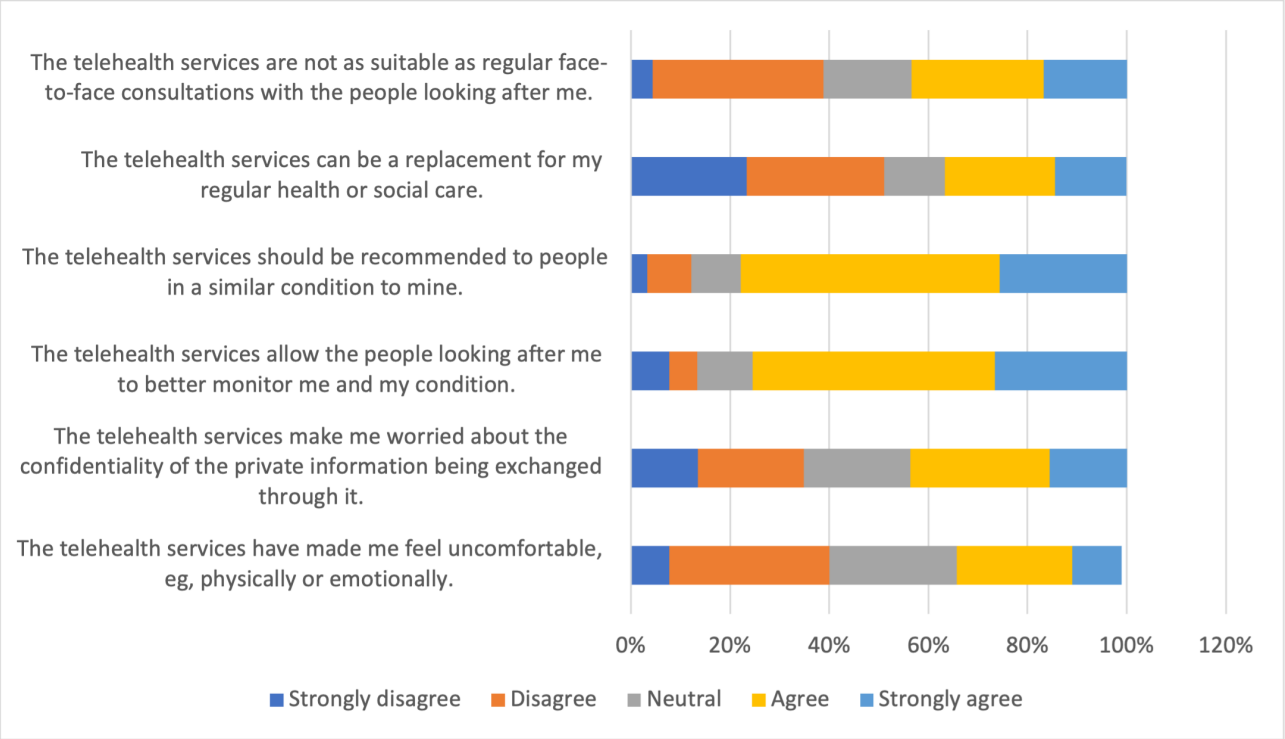

Table . Provider-reported barriers to telehealth from a cross-section survey conducted in Central Uganda, January 2022-July 2022.

	Number of provider-reported barriers (n=132), n (%)
Lack of reimbursement	19 (14)
Licensure	5 (4)
Technology challenges for the patient	34 (26)
Technology challenges for the health care provider or practice	19 (14)
Low patient engagement	25 (19)
Lack of implementation support	24 (18)
No challenges	2 (2)
Other	4 (3)

Overall, provider perceptions of telehealth did not differ according to age, gender, length of time practicing medicine, or based on prior telehealth use (Table 3). However, intentions to use telehealth when it becomes available were slightly higher among male versus female providers (4.73 vs 4.34 on Likert 1-5 scale; *P*=.03). Conversely, several differences in perceptions of telehealth according to patient characteristics were observed (Table 4). For example, concerns about confidentiality with telehealth were slightly higher among patients who did not use telehealth compared to users (3.29 vs 2.35; *P*=.01). Additionally, patients with better self-reported health status were more likely to agree that telehealth should be recommended to people in a similar condition to theirs compared to those with poorer health status (4.14 vs 3.53; *P*=.003). Although not statistically significant, males were slightly more likely than females to report feeling uncomfortable with using telehealth (3.30 vs 2.80; *P*=.06). Interest in participating in telehealth visits in the future was higher among younger (<50 y) versus older participants (3.22 vs 2.56; *P*=.07), among females versus males (4.08 vs 3.43; *P*=.05) and among those with better versus worse self-reported health status (4.12 vs 3.64; *P*=.04).

XSL•FO
RenderX

Table . Provider perceptions overall and according to provider characteristics.

	Overall (n=61), mean (SD)	Age			Sex			Length of time practicing			Telehealth use		
		<50 years (n=53), mean (SD)	≥50 years (n=7), mean (SD)	*P* value	Male (n=22), mean (SD)	Female (n=38), mean (SD)	*P* value	≤10 years (n=37), mean (SD)	>10 years (n=22), mean (SD)	*P* value	Used telehealth (n=48), mean (SD)	Did not use telehealth (n=13), mean (SD)	*P* value
The use of telehealth could help me monitor my patients more rapidly	4.49 (0.7)	4.47 (0.7)	4.57 (0.5)	.73	4.41 (0.6)	4.53 (0.8)	.54	4.58 (0.6)	4.32 (0.9)	.22	4.52 (0.7)	4.38 (0.5)	.54
I think that I could easily learn how to use telehealth	4.43 (0.5)	4.42 (0.5)	4.57 (0.5)	.47	4.45 (0.5)	4.42 (0.6)	.82	4.45 (0.6)	4.41 (0.5)	.79	4.38 (0.5)	4.62 (0.51)	.15
I think it is a good idea to use telehealth to monitor my patients	4.52 (0.6)	4.49 (0.6)	4.71 (0.5)	.33	4.55 (0.5)	4.50 (0.6)	.77	4.47 (0.6)	4.59 (0.5)	.45	4.48 (0.6)	4.69 (0.5)	.23
The use of telehealth may imply major changes in my clinical practice	4.13 (0.9)	4.15 (0.9)	4.43 (0.5)	.43	4.14 (0.9)	4.22 (0.8)	.73	4.24 (0.8)	4.09 (0.9)	.52	4.09 (1.0)	4.31 (0.9)	.65
Most of my patients will welcome the fact that I use telehealth	3.92 (0.7)	3.94 (0.8)	3.71 (0.8)	.45	3.82 (0.7)	3.97 (0.8)	.45	3.84 (0.8)	4.05 (0.7)	.32	4.00 (0.8)	3.62 (0.5)	.10

XSL•FO
RenderX

	Overall (n=61), mean (SD)	Age			Sex			Length of time practicing			Telehealth use		
		<50 years (n=53), mean (SD)	≥50 years (n=7), mean (SD)	*P* value	Male (n=22), mean (SD)	Female (n=38), mean (SD)	*P* value	≤10 years (n=37), mean (SD)	>10 years (n=22), mean (SD)	*P* value	Used telehealth (n=48), mean (SD)	Did not use telehealth (n=13), mean (SD)	*P* value
I think that my health center has the necessary infrastructure to support my use of telehealth	3.20 (1.2)	3.19 (1.2)	3.14 (1.1)	.93	3.18 (1.2)	3.18 (1.2)	.99	3.13 (1.3)	3.27 (1.1)	.66	3.27 (1.2)	2.92 (1.2)	.36
I feel comfortable with information and communication technologies	4.21 (0.7)	4.21 (0.7)	4.29 (0.8)	.78	4.18 (0.7)	4.24 (0.7)	.77	4.29 (0.7)	4.09 (0.7)	.29	4.23 (0.7)	4.15 (0.6)	0.73
I have the intention to use telehealth when it becomes available in my health center	4.48 (0.7)	4.43 (0.8)	4.86 (0.4)	.16	4.73 (0.5)	4.34 (0.8)	.03	4.53 (0.6)	4.41 (0.9)	.56	4.46 (0.7)	4.54 (0.9)	.73
Telehealth visits add value to my practice	4.22 (0.7)	4.21 (0.7)	4.29 (0.5)	.79	4.18 (0.7)	4.24 (0.6)	.74	4.13 (0.7)	4.38 (0.6)	.17	4.21 (0.7)	4.23 (0.8)	.93
A telehealth visit is more efficient than an in-person visit	3.23 (1.2)	3.28 (1.2)	3.00 (1.4)	.57	2.95 (1.2)	3.42 (1.2)	.16	3.21 (1.2)	3.32 (1.3)	.75	3.23 (1.2)	3.23 (1.4)	.99

XSL•FO
RenderX

Table . Patient perceptions overall and according to patient characteristics.

	Overall (n=91), mean (SD)	Age			Sex			Health status			Telehealth use		
		<50 years (n=74), mean (SD)	≥50 years (n=16), mean (SD)	*P* value	Male (n=23), mean (SD)	Female (n=67), mean (SD)	*P* value	<Very good (n=41), mean (SD)	≥Very good (n=49), mean (SD)	*P* value	Used telehealth (n=18), mean (SD)	Did not use telehealth (n=73), mean (SD)	*P* value
The telehealth services have made me feel uncomfortable, eg, physically, or emotionally.	2.96 (1.1)	3.74 (1.1)	3.75 (1.0)	.97	3.30 (1.1)	2.80 (1.1)	.06	2.98 (1.0)	2.90 (1.2)	.75	2.56 (1.1)	3.06 (1.1)	.09
The telehealth services make me worried about the confidentiality of the private information being exchanged through it.	3.11 (1.3)	3.94 (0.9)	3.50 (0.9)	.08	3.09 (1.0)	3.11 (1.4)	.96	2.90 (1.4)	3.27 (1.2)	.18	2.35 (1.1)	3.29 (1.3)	.01
The telehealth services allow the people looking after me to better monitor me and my condition.	3.81 (1.1)	3.88 (1.0)	3.56 (0.8)	.24	4.00 (1.1)	3.76 (1.1)	.38	3.70 (1.2)	3.92 (1.1)	.37	4.06 (1.2)	3.75 (1.1)	.31
The telehealth services should be recommended to people in a similar condition to mine.	3.88 (1.0)	2.86 (1.1)	3.25 (0.9)	.21	3.65 (1.1)	3.94 (1.0)	.24	3.53 (1.1)	4.14 (0.9)	.003	4.11 (0.7)	3.82 (1.1)	.16

XSL•FO
RenderX

	Overall (n=91), mean (SD)	Age			Sex			Health status			Telehealth use		
		<50 years (n=74), mean (SD)	≥50 years (n=16), mean (SD)	*P* value	Male (n=23), mean (SD)	Female (n=67), mean (SD)	*P* value	<Very good (n=41), mean (SD)	≥Very good (n=49), mean (SD)	*P* value	Used telehealth (n=18), mean (SD)	Did not use telehealth (n=73), mean (SD)	*P* value
The telehealth services can be a replacement for my regular health or social care.	2.77 (1.4)	3.14 (1.3)	3.38 (0.9)	.48	2.96 (1.4)	2.68 (1.4)	.42	2.75 (1.4)	2.76 (1.4)	.99	3.11 (1.6)	2.68 (1.4)	.25
The telehealth services are not as suitable as regular face to face consultations with the people looking after me.	3.17 (1.2)	3.88 (0.9)	3.50 (0.9)	.15	3.26 (1.3)	3.14 (1.2)	.67	3.28 (1.1)	3.08 (1.3)	.46	2.89 (1.3)	3.24 (1.2)	.28
I would be interested in continuing to participate in telehealth visits in the future.	3.92 (1.1)	3.22 (1.3)	2.56 (1.3)	.07	3.43 (1.1)	4.08 (0.9)	.05	3.64 (1.1)	4.12 (1.0)	.04	4.11 (1.0)	3.87 (1.1)	.42

Discussion

Findings from this study demonstrate generally favorable perceptions of telehealth among providers and patients in Central Uganda. Providers were more likely than patients to have used telehealth, and most noted the clinical value of telehealth and reported feeling comfortable using the technology. Importantly, this study also identified provider barriers impeding telehealth use, including technology challenges for the patient and the lack of infrastructure and implementation support. Patient perceptions of telehealth were largely neutral to positive, and concerns around confidentiality of information and comfort level using telehealth services were reported. Overall, our findings suggest the strong potential of telehealth approaches to improve health care service delivery in Central Uganda. Our results demonstrate the need for future telehealth implementation efforts that address technology and infrastructure-related challenges.

Overall, the favorable perceptions of telehealth among providers in this study align with results from another cross-sectional study in Uganda, where attitudes toward telemedicine were overwhelmingly positive, yet gaps in technology readiness were pronounced [22]. Results from the providers in this study demonstrate evidence for perceived usefulness of telehealth in monitoring patient care and providing clinical value to patients. These findings align with results from prior studies [23-25], including our prior study of rural providers in the United States

XSL•FO
RenderX

[19]. Additionally, most providers in this study reported ease of use in telehealth technology platforms. However, concerns around the efficiency of telehealth visits compared to in-person clinic visits were also commonly reported. Concerns around the efficiency of telehealth approaches were also noted in our team's prior study of rural providers in the United States [19]. These findings differ, however, from results in other study settings where improvements in efficiency with telehealth approaches were observed [23,24]. The lack of supportive infrastructure for telehealth visits, which was noted as an important barrier to telehealth among survey respondents, may contribute to the discrepancies in findings across studies. Indeed, most providers in this study reported that telehealth approaches would require major changes in current clinical practice and workflows. Such changes may help improve overall efficiencies with the telehealth system and overcome this reported barrier.

In this study, patient perceptions of telehealth were less favorable compared to providers. This may be due in part to lower levels of telehealth utilization and comfort in using the technology among patients. Indeed, while most providers in this study had used telehealth, both awareness and telehealth use among patients were much lower. This suggests the need to better understand patient-level barriers to telehealth use and to target patient education efforts to increase overall knowledge about telehealth applications from the patient perspective. Importantly, our findings also demonstrate satisfaction among patients who had previously used telehealth, and most previous users also recognized the clinical utility of telehealth for care monitoring. However, there were concerns among patients related to the comfort of using telehealth and the confidentiality of the private information shared through the technology. These findings emphasize the need for future efforts to increase awareness of telehealth services, offer educational trainings on digital literacy, and address concerns related to the confidentiality of information.

Perceptions of telehealth in this study did not differ substantially by most provider and patient characteristics. Interestingly, we did not observe any significant differences in telehealth perceptions according to provider or patient age. This differs from findings in U.S. study populations, where telehealth perceptions were more favorable among younger patients and providers [19,26,27]. In these prior studies, age-related differences in telehealth perceptions were largely attributed to lower technology access and digital literacy among older adults [28]. Participants in this study were younger overall, and this may have contributed to the lack of observed differences in telehealth perceptions by age in our study. Intentions to use telehealth in clinical practice were modestly higher among male compared to female providers in our study. This finding differs from those assessing mental health providers in the United States, where no gender differences in intention to continue using telehealth after the COVID-19 pandemic were observed [29]. It is important to note that the intention to use telehealth was high among female and male providers in our study. Thus, future studies are needed to assess whether this gender difference persists in other study settings. Interestingly, patients with better self-reported health status were more favorable about recommending telehealth to people in similar conditions and were also more interested in continuing to participate in telehealth visits in the future. These findings highlight the growing recognition that telehealth may not be appropriate or feasible for all patients or all clinical care situations [30].

The strengths of this study include the focus on multiple perspectives, including those of both providers and patients across multiple clinical sites in this understudied region. Moreover, our study used a community-engaged approach, working in partnership with community members and incorporating input on study processes and outcomes from clinical and community partners. This approach can help guide implementation efforts toward those with the most community relevance and potential impact. We also held dissemination workshops in the community to discuss the findings and inform next steps to ensure that research efforts address the relevant needs and preferences of the community members in this region of Uganda. This research provides a replicable and scalable model that can help inform telehealth implementation efforts in other under-resourced settings. This study did have several limitations. First, using a convenience sampling strategy, we were unable to calculate a response rate given resource limitations for tracking survey distribution across multiple study sites. While the survey development was informed by the technology acceptance model [20] and modeled on surveys used in prior studies [31,32], the survey questions were not specifically validated. Additionally, we were unable to assess whether perceptions of telehealth differed according to the type of health care provider, as this information was not collected on the provider survey. This should be assessed in future studies, particularly given that readiness to integrate telemedicine services varied across health facility types and job roles in a prior study conducted in Uganda [14]. Finally, this study focuses on a single region in Uganda, and results may not be broadly generalizable across other populations and settings.

Overall, our findings demonstrate both satisfaction with telehealth among users and willingness to use it among nonusers for both providers and patients. Indeed, providers and patients who previously used telehealth unanimously reported recommending the service to others. All surveyed providers and most patients who had not used telehealth were interested in using this approach in the future. These findings suggest positive attitudes and behavioral intentions to use telehealth, which are key components to technology acceptance according to the technology acceptance model [20]. Taken together, our results suggest that the perceived usefulness, ease of use, and attitudes toward telehealth are supportive of future implementation efforts to increase the reach of telehealth services in Uganda and in other low- and middle-income countries.

Findings from this study can inform the implementation of acceptable telehealth approaches to address disparities propagated by health care access barriers in Sub-Saharan Africa.

XSL•FO
RenderX

Acknowledgments

Funding for this study was received from Michigan State University, Alliance for African Partnerships: Partnerships for Innovative Research in Africa, Early Career Investigators Award (MK and KH).

Data Availability

The datasets generated during this study are not publicly available due to participants not consenting to the sharing of data beyond the study team.

Conflicts of Interest

None declared.

References

1. Institute of Medicine (US) Committee on Understanding and Eliminating Racial and Ethnic Disparities in Health Care. In: Smedley BD, Stith AY, Nelson AR, editors. Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care: National Academies Press; 2003.
2. Douthit N, Kiv S, Dwolatzky T, Biswas S. Exposing some important barriers to health care access in the rural USA. Pub Health (Fairfax) 2015 Jun;129(6):611-620. [doi: 10.1016/j.puhe.2015.04.001]
3. Syed ST, Gerber BS, Sharp LK. Traveling towards disease: transportation barriers to health care access. J Community Health 2013 Oct;38(5):976-993. [doi: 10.1007/s10900-013-9681-1]
4. Nutbeam D, Lloyd JE. Understanding and responding to health literacy as a social determinant of health. Annu Rev Public Health 2021 Apr 1;42:159-173. [doi: 10.1146/annurev-publhealth-090419-102529] [Medline: 33035427]
5. Hsiao A, Vogt V, Quentin W. Effect of corruption on perceived difficulties in healthcare access in sub-Saharan Africa. PLoS ONE 2019;14(8):e0220583. [doi: 10.1371/journal.pone.0220583] [Medline: 31433821]
6. Kamulegey L, Ssebwana J, Abigaba W, Bwanika J, Musinguzi D. Mobile health in Uganda: a case study of the Medical Concierge Group. EASci 2019 Mar 25;1(1):9-14. [doi: 10.24248/easci.v1i1.12]
7. LeRouge C, Garfield MJ. Crossing the telemedicine chasm: have the U.S. barriers to widespread adoption of telemedicine been significantly reduced? Int J Environ Res Public Health 2013 Nov 28;10(12):6472-6484. [doi: 10.3390/ijerph10126472] [Medline: 24287864]
8. Hoque MR, Bao Y, Sorwar G. Investigating factors influencing the adoption of e-Health in developing countries: a patient's perspective. Inf Health Soc Care 2017 Jan 2;42(1):1-17. [doi: 10.3109/17538157.2015.1075541]
9. Mahmoud Z, Sanusi M, Nartey C, Adedinsewo D. Using technology to deliver cardiovascular care in African countries. Curr Cardiol Rep 2023 Dec;25(12):1823-1830. [doi: 10.1007/s11886-023-01988-2] [Medline: 37486571]
10. Ogodo O. Dozens of telehealth companies are innovating in Africa. N Med 2023 Feb;29(2):281-282. [doi: 10.1038/s41591-022-02089-3] [Medline: 36759675]
11. RocketHealth. URL: https://rockethealth.africa/ [accessed 2022-08-08]
12. Isabalija D, Mayoka KG, Rwashana DA, Mbarika PVW. Factors affecting adoption, implementation and sustainability of telemedicine information systems in Uganda. J Health Inform Dev Ctries 2011;5(2).
13. Medard T, Yawe BL, Bosco OJ. Health care delivery system in Uganda: a review. Tanzan J Health Res 2023;24(2):57. [doi: 10.4314.thrb.v24i2.5]
14. Kiberu VM, Scott RE, Mars M. Assessment of health provider readiness for telemedicine services in Uganda. Health Inf Manag 2019 Jan;48(1):33-41. [doi: 10.1177/1833358317749369] [Medline: 29359588]
15. Paul KJ, Munuulo J, Nagujja S. Empowerment digital divide: case of internet adoption in Uganda's rural-urban areas. Adv Soc Sci Res J 2015 Nov;2(11):11. [doi: 10.14738/assrj.211.1567]
16. Uganda: driving inclusive socio-economic progress through mobile-enabled digital transformation. Global System for Mobile Communications Association. 2019. URL: https://www.gsma.com/solutions-and-impact/connectivity-for-good/mobile-for-development/wp-content/uploads/2019/03/GSMA_Connected_Society_Uganda_Overview.pdf [accessed 2025-01-17]
17. Hirani SP, Rixon L, Beynon M, et al. Quantifying beliefs regarding telehealth: development of the whole systems demonstrator service user technology acceptability questionnaire. J Telemed Telecare 2017 May;23(4):460-469. [doi: 10.1177/1357633X16649531] [Medline: 27224997]
18. AMA digital health research: physicians' motivations and requirements for adopting digital health. American Medical Association. 2020 Feb. URL: https://www.ama-assn.org/system/files/2020-02/ama-digital-health-study.pdf [accessed 2025-01-20]
19. Klee D, Pyne D, Kroll J, James W, Hirko KA. Rural patient and provider perceptions of telehealth implemented during the COVID-19 pandemic. BMC Health Serv Res 2023 Sep 12;23(1):981. [doi: 10.1186/s12913-023-09994-4] [Medline: 37700286]

XSL•FO
RenderX

20. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989 Sep;13(3):319. [doi: 10.2307/249008]
21. Dodoo JE, Al-Samarraie H, Alzahrani AI. Telemedicine use in Sub-Saharan Africa: barriers and policy recommendations for Covid-19 and beyond. Int J Med Inform 2021 Jul;151:104467. [doi: 10.1016/j.ijmedinf.2021.104467]
22. Kiberu VM, Scott RE, Mars M. Assessing core, e-learning, clinical and technology readiness to integrate telemedicine at public health facilities in Uganda: a health facility - based survey. BMC Health Serv Res 2019 Apr 29;19(1):266. [doi: 10.1186/s12913-019-4057-6] [Medline: 31035976]
23. Beiser M, Lu V, Paul S, et al. Electronic health record usage patterns: assessing telemedicine's impact on the provider experience during the COVID-19 pandemic. Telemed E-Health 2021 Aug 1;27(8):934-938. [doi: 10.1089/tmj.2020.0490]
24. Saiyed S, Nguyen A, Singh R. Physician perspective and key satisfaction indicators with rapid telehealth adoption during the coronavirus disease 2019 pandemic. Telemed J E Health 2021 Nov;27(11):1225-1234. [doi: 10.1089/tmj.2020.0492] [Medline: 33513045]
25. Gentry MT, Puspitasari AJ, McKean AJ, et al. Clinician satisfaction with rapid adoption and implementation of telehealth services during the COVID-19 pandemic. Telemed J E Health 2021 Dec;27(12):1385-1392. [doi: 10.1089/tmj.2020.0575] [Medline: 33606560]
26. Ladin K, Porteny T, Perugini JM, et al. Perceptions of telehealth vs in-person visits among older adults with advanced kidney disease, care partners, and clinicians. JAMA Netw Open 2021 Dec 1;4(12):e2137193. [doi: 10.1001/jamanetworkopen.2021.37193] [Medline: 34870680]
27. Fischer SH, Ray KN, Mehrotra A, Bloom EL, Uscher-Pines L. Prevalence and characteristics of telehealth utilization in the United States. JAMA Netw Open 2020 Oct 1;3(10):e2022302. [doi: 10.1001/jamanetworkopen.2020.22302] [Medline: 33104208]
28. Gordon NP, Crouch E. Digital information technology use and patient preferences for internet-based health education modalities: cross-sectional survey study of middle-aged and older adults with chronic health conditions. JMIR Aging 2019 Apr 4;2(1):e12243. [doi: 10.2196/12243] [Medline: 31518291]
29. Wilczewski H, Paige SR, Ong T, et al. Providers' perspectives on telemental health usage after the COVID-19 pandemic: retrospective analysis. JMIR Form Res 2022 Nov 11;6(11):e39634. [doi: 10.2196/39634] [Medline: 36322787]
30. Chaet D, Clearfield R, Sabin JE, Skimming K, Council on Ethical and Judicial Affairs American Medical Association. Ethical practice in telehealth and telemedicine. J Gen Intern Med 2017 Oct;32(10):1136-1140. [doi: 10.1007/s11606-017-4082-2] [Medline: 28653233]
31. Holtz B, Mitchell K, Hirko K, Ford S. Using the technology acceptance model to characterize barriers and opportunities of telemedicine in rural populations: survey and interview study. JMIR Form Res 2022 Apr 15;6(4):e35130. [doi: 10.2196/35130] [Medline: 35436207]
32. Holtz BE. Patients perceptions of telemedicine visits before and after the coronavirus disease 2019 pandemic. Telemed J E Health 2021 Jan;27(1):107-112. [doi: 10.1089/tmj.2020.0168] [Medline: 32614689]

Abbreviations

DHO: District Health Officer

Edited by A Mavragani; submitted 23.05.24; peer-reviewed by M Mars; revised version received 19.11.24; accepted 20.11.24; published 23.01.25.

Please cite as:
Kizito M, Mugabi EN, Ford S, Holtz B, Hirko K
Characterizing Telehealth Barriers and Preferences to Promote Acceptable Implementation Strategies in Central Uganda: Multilevel Formative Evaluation
JMIR Form Res 2025;9:e60843
URL: https://formative.jmir.org/2025/1/e60843
doi:10.2196/60843

© Michael Kizito, Erina Nabunjo Mugabi, Sabrina Ford, Bree Holtz, Kelly Hirko. Originally published in JMIR Formative Research (https://formative.jmir.org), 23.1.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Testing an Electronic Patient-Reported Outcome Platform in the Context of Traumatic Brain Injury: PRiORiTy Usability Study

Christel McMullan[1], PhD; Grace Turner[2], PhD; Ameeta Retzer[1], PhD; Antonio Belli[3], MD, FRCS, FRCS (SN); Elin Haf Davies[4], PhD; Laura Nice[5], PhD; Luke Flavell[1]; Jackie Flavell[1]; Melanie Calvert[1], PhD

[1]Centre for Patient Reported Outcomes Research, Institute of Applied Health Research, University of Birmingham, Birmingham, United Kingdom

[2]Sport, Exercise and Rehabilitation Sciences, University of Birmingham, Birmingham, United Kingdom

[3]Institute of Inflammation and Ageing, University of Birmingham, Birmingham, United Kingdom

[4]Aparito Ltd, Wrexham, United Kingdom

[5]Centre for Conflict Wound Research, University of Birmingham, Birmingham, United Kingdom

Corresponding Author:
Christel McMullan, PhD
Centre for Patient Reported Outcomes Research, Institute of Applied Health Research, University of Birmingham, Birmingham, United Kingdom

Abstract

Background: Traumatic brain injury (TBI) is a significant public health issue and a leading cause of death and disability globally. Advances in clinical care have improved survival rates, leading to a growing population living with long-term effects of TBI, which can impact physical, cognitive, and emotional health. These effects often require continuous management and individualized care. Traditional paper-based assessments can be cumbersome, potentially impeding regular monitoring of patient-reported outcomes (PROs). Electronic PROs (ePROs) offer a promising alternative by enabling real-time symptom tracking, which can facilitate early identification of issues, support shared decision-making, and improve outcomes for patients with TBI.

Objective: This study evaluates the usability of an ePRO platform—Atom5—for individuals with TBI. By analyzing how patients use the system to report their symptoms, the study aims to identify usability issues, assess user satisfaction, and determine the potential of Atom5 to support ongoing patient-centered care.

Methods: Atom5 was customized to enable individuals with TBI to report their symptoms. Usability testing was conducted through one-on-one sessions with participants recruited from Headway UK—an organization supporting brain injury survivors. Each participant took part in cognitive interviews using with the "Think Aloud" method, encouraging them to verbalize their thoughts and experiences while using the platform. This approach provided qualitative insights into areas of difficulty, usability strengths, and accessibility barriers. User satisfaction was quantitatively assessed with a brief 4-item questionnaire based on the System Usability Scale. Usability outcomes were analyzed for critical and noncritical errors, focusing on user experience and overall satisfaction.

Results: In total, 9 participants completed a single usability testing session using Atom5, including 4 men, 4 women, and 1 nonbinary individual; 4 participants were under 55 years old, and 6 had their TBI <10 years ago. Finally, 8 participants used an Android device. The platform included measures for anxiety (Generalized Anxiety Disorder-2 item), depression (Patient Health Questionnaire-2), posttraumatic stress disorder (Posttraumatic Stress Disorder checklist 2), and TBI-specific quality of life (Traumatic Brain Injury – Quality of Life Short form) and a total of 26 questions. Overall, all participants were satisfied with the system, noting that it was easy to navigate and accessible despite difficulties in understanding some questions. Further, 6 participants encountered no errors, while 1 participant reported one critical error and 2 others reported one noncritical error each. The participants rated their overall satisfaction with the platform at an average score of 3.9 (SD 0.49) out of 5.

Conclusions: This usability study suggests that individuals living with TBI can effectively report symptoms using the Atom5 ePRO platform, with generally high satisfaction and few usability issues, thereby enabling continuous monitoring and proactive symptom management. Future ePRO development should focus on inclusivity and adaptability to address the diverse needs of patients with TBI, ensuring these tools can effectively support a wide range of users.

(JMIR Form Res 2025;9:e58128) doi:10.2196/58128

KEYWORDS

usability study; usability; patient reported outcome; PRO; electronic patient reported outcome; ePRO; traumatic brain injury; TBI; think aloud; cognitive interviews; early warning; early detection; mobile phone

XSL•FO
RenderX

Introduction

Background

A traumatic brain injury (TBI) is defined as "a traumatically induced structural injury and/or physiological disruption of brain function as a result of an external force" [1]. Each year, it is estimated that over 50 million TBIs occur worldwide [2]. In the United Kingdom, around 900,000 people attend accident and emergency departments (immediate critical care) with a head injury every year, with approximately 1.3 million people living with the effects of a TBI [3].

Although most TBIs result in mild symptoms and do not require hospitalization, many individuals experience disabling long-term symptoms [4-7]. The impact of a TBI falls into several categories: physical (headaches, dizziness, and blurred vision), cognitive (trouble with memory, concentration and attention, and impaired reasoning), emotional (personality changes, mood swings, depression, and anxiety), and behavioral (fatigue and anger) [8-11]. This means that patients with mild TBI have different needs, and health care must be tailored to each patient [12].

In addition, advances in management and guidelines [13] mean that people now live longer with the effects caused by their TBI. It is therefore important for clinicians to monitor these effects over time to improve the quality of life of people living with the effects of a TBI.

As a result, TBI is being increasingly regarded as a chronic condition, as it impacts on multiple health domains and functions [4,14]. Research has shown that TBI is a risk factor for dementia [14], psychiatric, cardiovascular, genitourinary and neurodegenerative conditions [15-17].

One way to monitor the effects of TBI on patients' symptoms and quality of life over time is by using patient-reported outcomes measures (PROMs). PROMs are defined as "questionnaires completed by patients to assess the effects of disease or treatment (or both) on symptoms, functioning, and health related quality of life from their perspective" [18]. PROMs have traditionally been administered in a paper format. Research has shown that there are several drawbacks to paper PROMs. These include lengthy completion time, risk of poor data quality, and being prone to human error [19]. The growth of eHealth technologies and the omnipresence of smartphones and tablets [20] has allowed patients to play a more active role in their health care [21-23].

Digital health is a rapidly evolving field and provides a cost-effective way to remotely monitor patients [12,23]. Often referred to as mobile health or "mhealth," recent research has shown that the use of these new technologies in health care has been received positively both by patients [12,24] and clinicians [24-26]. In addition to this, new technologies have given clinicians and patients the opportunity to use electronic PROMs (ePROMs) [20]. Unlike paper PROMs, ePROMs can facilitate real-time data, remote monitoring of symptoms and send or receive feedback [27], and can improve acceptance rate [28]. Examples of successful ePROs systems include AmbuFlex in Denmark with patients with renal failure [29] and the eRAPID system in the United Kingdom to monitor side effects of chemotherapy [30]. However, disadvantages of ePROMs include concerns around privacy issues, technical difficulties, a large initial financial investment, and the potential digital divide [19].

Like any new product being developed, it is vital to test the usability of a new electronic patient-reported outcome (ePRO) system in order to improve and optimize the final product [31,32]. Usability testing is defined by the International Organisation for Standardization (ISO) as "the extent to which a product can be used by specified user to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use" [33].

Research Context: The PRiORiTy Study

This usability study is the second stage of the wider National Institute for Health and Care Research Surgical Reconstruction and Microbiology Research Centre (NIHR SRMRC)–funded PRiORiTy (Patient-Reported Outcomes Research in Trauma) study. The PRiORiTy study aimed to develop and test the usability and feasibility of an ePRO system with people with a TBI. Prior to this usability study, we conducted a qualitative study on the views and perspectives of patients living with the effects of a TBI, their carers or family members, and health care professionals on using PROMs or ePROMs. The findings of the qualitative study were reported in an article published elsewhere [27].

Aim

The aim of this study was to test the usability of using an ePRO system (Atom5) with individuals who had a TBI.

Methods

Study Participants and Settings

People who had a TBI were recruited in July 2021 from Headway UK, a UK-based charity aiming to promote understanding of all aspects of brain injury and provide information, support and services to survivors, their families and carers. The sample size was informed by previous usability studies, which were in favor of the 10±2 rules, which advocated that around 10 participants could identify 80% of the issues around the system being tested [34-36]. Other studies showed that as low as 5 people are needed if conducting a series of testing sessions [31,37]. However, since we conducted only one testing session, we decided to increase our sample size and it was clear that we reached data saturation by the time we conducted our last testing session. Inclusion criteria were being 18 years of age and older, having the ability to converse in English, and to give informed consent. We excluded anyone who was not able to give informed consent.

Recruitment Process

A lead contact within Headway UK identified people who were deemed eligible (purposive sampling), approached them in person, and introduced them to the study. The Headway UK staff member briefly explained the study to eligible people and gave them the information sheet. They then sent the researcher (CM), via a secure email address, the contact details of the people who had verbally agreed to take part in the usability

XSL•FO
RenderX

study. CM phoned these people to arrange a mutually suitable date and time to conduct the testing session. Written informed consent was obtained from all the participants and data were anonymized.

ePRO Platform (Atom5)

The ePRO platform used in this study was Atom5, which was developed by Aparito Ltd [38], a UK-based medical technology company. Atom5 consists of two interfaces (Multimedia Appendix 1):

1. A clinician dashboard accessed via a web browser that can be accessed by specific members of the research team with appropriate permission.
2. A patient-facing interface accessed via an app on Android or iOS devices onto which study participants input their patient-reported outcome data.

The Atom5 app for the PRiORiTy study was designed in conjunction with our Patient and Public Involvement and Engagement group, the research team (including a consultant neurosurgeon) and was informed by the findings from the project's qualitative study [27].

Aparito Ltd are accredited for ISO 13485:2016 for medical device quality management system, ISO 27001:2001 for information security management systems, and Cyber Essential Plus recertification.

PROMs Selection

The selection of PROMs was informed by the findings of our previous qualitative research [27], literature searches, and input from our patient partners and clinical collaborators. Our qualitative research showed that people who had a TBI often experienced anxiety, depression, and posttraumatic stress disorder, and this was further evidenced by the existing literature [8,9,27]. Therefore, the PROMs included in this study were the Generalized Anxiety Disorder-2 item (GAD-2), Patient Health Questionnaire-9 (PHQ-9), Posttraumatic Stress Disorder checklist 2 (PCL-2) and Traumatic Brain Injury – Quality of Life Short form (TBI QOL SF; ability to participate in social roles and satisfaction with social roles and activities). Quality of Life after Brain Injury (QOLIBRI) was considered as a potential TBI-specific measure. However, because it includes 37 items, the Patient and Public Involvement and Engagement and clinical collaborators advised against it, as they felt participants with TBI would not be able to complete it due to their inability to stay focused for long periods of time. Instead they suggested using 2 submeasures of the TBI QOL SF (ability to participate in social roles and satisfaction with social roles and activities) for a total of 20 questions. In addition, clinical collaborators had already been using TBI QOL questionnaire in their clinical practice.

The details of the four PROMs selected are mentioned in Table 1.

XSL•FO
RenderX

Table . Patient-reported outcome measures included in the Patient-Reported Outcomes Research in Trauma usability study.

PROM[a] category and PROMs		Description
General population		
	GAD-2[b] (Anxiety)	• Screening tool for generalized anxiety disorder [39] derived from GAD-7[c] • Two items • Recall period of 2 weeks • Validated measure which has retained the same psychometrics properties of the GAD-7 (86% sensitivity and 83% specificity) [40]
	PHQ-2[d] (Depression)	• Screening tool for depression derived from the PHQ-9[e] • Two items • Recall period of 2 weeks • Validated measure with sensitivity of 83% and specificity of 92% [41]
	PCL-2[f] (PTSD[g])	• Abbreviated version of the PTSD Checklist – Civilian version (PCL-C) • Two items • Screening tool for PTSD • Recall period of 1 month • Validated measure with sensitivity of .97 and specificity of .58 [42]
Disease-specific		
	TBI QOL SF[h] (ability to participate in social roles; satisfaction with social roles and activities)	• Health-related Quality of Life • 20 items • Recall period of 7 days • Validated measures with excellent psychometric characteristics, although more research is needed to establish evidence of construct validity and evaluate the measures' sensitivity to change [43]. In addition, the measures need to be validated when using via mobile health.

[a]PROM: patient-reported outcome measures.
[b]GAD-2: Generalized Anxiety Disorder-2 item.
[c]GAD-7: Generalized Anxiety Disorder-7 item.
[d]PHQ-2: Patient Health Questionnaire-2.
[e]PHQ-9: Patient Health Questionnaire-9.
[f]PCL-2: Posttraumatic Stress Disorder checklist 2.
[g]PTSD: posttraumatic stress disorder.
[h]TBI QOL SF: Traumatic Brain Injury – Quality of Life Short form.

Data Collection and Testing Procedure

Testing Sessions

The usability testing aimed at assessing the effectiveness and efficiency of the ePRO platform. This consisted of one-to-one sessions conducted either online on a University of Birmingham secure Zoom account or face-to-face on Headway premises in the West Midlands. The testing sessions and interviews were conducted by CM, PhD, who is a female qualitative research fellow with extensive experience in qualitative data collection, including conducting interviews. She did not know the participants prior to the start of the research. The participants were fully informed of the reasons for doing the research prior to the data collection.

At the start of the session, participants were given a patient leaflet that outlined the main stages involved in navigating through the app and completing the questionnaires. They were then asked to download the app and perform a series of tasks on the app while sharing their thoughts about the platform (Think Aloud technique) [44]. Finally, the participants took part in a short interview (10 - 15 min) on their experience of using the electronic platform. A carer or family member was allowed to attend the testing session. A topic guide was developed and pilot-tested with a patient with TBI. Field notes were kept during the testing sessions.

The tasks performed by the participants are detailed in Textbox 1.

XSL•FO
RenderX

Textbox 1. Tasks performed by participants during the testing sessions.

- Download the Atom5 app
- Onboard
- Consent
- Navigate through homepage
- Navigate through Questionnaires module
- Complete and submit Generalized Anxiety Disorder-2
- Complete and submit Patient Health Questionnaire-2
- Complete and submit Posttraumatic Stress Disorder checklist 2
- Complete and submit Traumatic Brain Injury – Quality of Life Short form
- Navigate through frequently asked questions module

Satisfaction Questionnaire

Finally, the participants were asked to answer a brief 4-question satisfaction questionnaire to rate their satisfaction with the usability of the electronic platform (5-point scale: 1–poor or never to 5–excellent or yes) [36,45].

Data Saturation

There has been an ongoing debate about sample sizes for usability testing. Our sample size has been been informed by previous research, which shows that 8 - 10 participants are required to detect over 80% of issues.

Data Analysis

Data were analyzed by two independent researchers. The participant characteristics were summarized as frequency and percentage. The testing sessions were audio-recorded. The participants' comments during the testing sessions and the researcher's observations were collated into a table. Errors were recorded and divided into critical (requiring the researcher's input to be able to continue) and noncritical (not requiring the researcher's input to be able to continue). Suggestions to improve the electronic platform were also recorded. Participant ratings for the 4 satisfaction questions were used to calculate a mean score per question. Overall mean score and SD was also calculated.

Data collected from the ePROMs was used for the purpose of the usability study only and was not downloaded.

The COREQ (Consolidated Criteria for Reporting Qualitative Research) reporting checklist was used to report findings (Checklist 1).

Patient and Public Involvement and Engagement

One person who was living with the effects of a TBI and a family member were involved throughout the study. We met regularly (on average every 3 mo) and they reviewed patient-facing documents (patient information sheet, consent form, and instructions leaflet), assisted with recruitment and provided advice on the design and development of the Atom5 platform. In addition to meeting regularly, we were in regular email contact in order to quickly address any urgent issues.

Ethical Considerations

The study was approved the University of Birmingham Ethics committee (reference ERN_17 - 1253). Informed consent was obtained from all participants. The participants were not compensated for taking part in the study. All data were deidentified to ensure participant privacy and confidentiality.

Results

Participants' Characteristics

A total of 9 people who had a TBI took part in a testing session. This included 4 men, 4 women, and 1 person identifying as nonbinary. A total of 4 out of the 9 participants (44%) were under 55 years old, and 5 were over 55 years old. The majority of the participants (6/9 participants, 66%) had their TBI less than 10 years ago. Finally, 8 out of 9 participants (89%) used an Android device, and 1 used an iOS device. The participants' characteristics are summarized in Table 2.

In total, 5 testing sessions took place remotely by video using Zoom and the other 4 took place face-to-face on Headway UK premises. No one else was present in the interview room, although Headway UK staff were in the building. The testing sessions lasted between 42 and 63 minutes.

Among the 4 male participants, 3 of them were visually impaired and needed some help to read the questions. The researcher (CM) read the questions for them verbatim and helped them to input their answers onto the electronic device. These 3 male participants did not own an electronic device and had to borrow one from friends.

XSL•FO
RenderX

Table . Participants' characteristics.

Variables		Values, n (%)
Age (years)		
	<55	4 (44)
	>55	5 (56)
Sex		
	Male	4 (44)
	Female	4 (44)
	Nonbinary	1 (12)
Ethnicity		
	White	9 (100)
Number of years since traumatic brain injury		
	<10 years ago	6 (67)
	10 - 20 years ago	2 (22)
	>20 years ago	1 (11)
Device used		
	iOS	1 (11)
	Android	8 (89)

Effectiveness and Efficiency of the Atom5 App

One critical error and 2 noncritical errors were recorded and all the participants completed the testing sessions. The critical error was due the participant not knowing what to do after inputting the onboarding code (please note that after downloading the Atom5 platform, users need to either scan a QR code or enter a numerical code to be able to access the questionnaires. The QR code or numerical code was given to them by the researcher (CM). The two noncritical errors were due to downloading the app on an iPad not compatible with the platform and not being able to answer a question because the question was missing. Table 3 shows how each participant addressed the critical and noncritical errors.

Table . Descriptions of critical and noncritical errors encountered by participants and how they were addressed during testing sessions.No differences were found between iOS and Android users.

Error (error type)	How participants addressed noncritical error
Not knowing what to do after inputting the onboarding code (critical error)	Researcher advised the participant
Downloading the app on an iPad not compatible with the platform (non-critical error)	Participant used a newer device
Not being able to answer a question because the question was missing (noncritical error)	Participant had to go back to previous page to read the question

It was difficult to record how long it took participants to complete the PROMs on Atom5 as they discussed their experience with the researcher while carrying out the tasks required. However, the researcher observed that while the majority had very little difficulty completing GAD-2, PHQ-2, and PCL-2 and completed them in under 30 seconds, most of the participants struggled with the TBI QOL SF questionnaire and took longer to complete it. One of the reasons for this is that it is a longer questionnaire (20 questions for TBI QOL SF vs 6 questions for GAD-2, PHQ-2, and PCL-2 combined). However, another reason for struggling with it was the difficulty understanding some of the questions, such as the following questions, and not knowing how to answer them:

- "I am able to do all the community activities that I want to do"
- "I can do everything for work that I want to do (including work at home)."
- "I am satisfied with the amount of time I spend doing work (including work at home)."

Issues and Suggestions for Improvement

The participants identified a number of issues and made some suggestions to improve their experience of using the Atom5 platform (Table 4). These issues and suggestions related to font size, lack of clarity, names of questionnaires, questions numbers, clarity of some the text, what to do after onboarding, communication with clinical team, length of questionnaires, questions wording, colors, and name of the app.

XSL•FO
RenderX

Table . Summary of issues and suggestions for improvement made by participants during the testing sessions.

Issue	Details	Suggestions made by participants
Font size too small	Some items were difficult to read: • Questions in GAD-2[a] were difficult to read • FAQs[b]	• Increase font
Lack of clarity	Difficulty reading and understanding: • Consent text • FAQs • Thank you message after questionnaires • Unsure what the FAQs are about	• Reformat using bullet points or add paragraphs • Include information in FAQs about study and what participants have to do
Lack of information about study	• What participants need to do after onboarding was not always obvious to all participants	• Add instructions on app
Communication with clinical team	• Need for some participants to leave comments	• Add free text boxes at the end of each questionnaire
Length of questionnaires	• TBI QOL SF[c] was lengthy	• Add the question number at the start of TBI QOL SF question (eg, 1/20; 2/20, etc)
Questions wording	• Difficulty understanding some of the TBI QOL SF questions	• Reword the questions[d] • Add emojis/pictures
Orange and white colors	• Most participants did not understand the significance of the orange and white colors • The white color was difficult to see by one participant	• Add explanation in FAQs • Increase font size • Bold words that are on white background
Name of the app Atom5	• Name of app was not related to TBI[e]	• Change the name to make it more relevant to TBI

[a]GAD-2: Generalized Anxiety Disorder-2 item.

[b]FAQ: frequently asked questions.

[c]TBI QOL SF: Traumatic Brain Injury – Quality of Life Short form.

[d]This may not be possible, as it may impact on validity of measures and would require work with the developers.

[e]TBI: traumatic brain injury.

Participants' Overall Comments and Researchers' Observations

Overall, participants were positive about using Atom5 to report their symptoms. They found the app easy to download and use. Most of them liked the white and orange colors of the app, even though they did not understand the significance of it. The main issues mentioned by the participants was regarding the TBI QOL SF questions, which they had difficulty understanding and therefore were not sure how to answer them. One participant reported they would be unlikely to use the platform again because it would be too difficult to use on their own as they would forget to use it (Textbox 2).

XSL•FO
RenderX

Textbox 2. Participants' overall comments on using the Atom5 platform and researchers' observations.

Participants' comments

- Very easy to use.
- App is quite good and looks smart.
- Easier than a piece of paper.
- Things are laid out very well in the app.
- Unlikely to use it again because it would be difficult to use on my own.
- I wouldn't remember to use it.
- Downloading the app might be an issue for some people with a traumatic brain injury.
- Made sense but had to think about the questions.

Researchers' observations

- Several participants had to read the Traumatic Brain Injury – Quality of Life Short form (TBI QOL SF) questions several times, difficulty understanding what the questions mean.
- Two participants were getting tired halfway through completing the TBI QOL SF questionnaire.

Participants' Satisfaction

The summary scores for participants' satisfaction and rating of usability of Atom5 are shown in Table 5. The overall mean score was 3.9 (SD 0.33) and the mean score for individual questions was between 3.4 (SD 0.52) and 4.6 (SD 0.72).

Table . Participants' usability and satisfaction scores of using the Atom5 electronic platform (range: 1 - 5).

Question	Average score (SD)
Q1. How easy was the system to use and navigate?	3.9 (0.33)
Q2. How satisfied are you with the content?	4.6 (0.72)
Q3. How satisfied are you with the display?	3.4 (0.52)
Q4. How likely are you to use it again or recommend it to others?	3.9 (1.16)
Average usability and satisfaction score	3.9 (0.49)

Discussion

Summary of Findings

This article reports the results of the usability testing of the Atom5 platform used by people living with the effects of a TBI. Our ePRO platform was informed by existing literature, patients, and a clinician. A total of 4 PRO measures were selected and programmed in Atom5. The results of the interviews show that most of the people living with the effects of a TBI who took part in the usability study were able to report their symptoms using the Atom5 platform. The number of errors were low (1 critical and 2 noncritical errors) and were due to incompatibility of a device, not knowing what to do after onboarding, or a missing question. Overall, the platform was found to be effective. Suggestions to improve the electronic platform included increasing the font, reformatting the text and layout to make it clearer, and adding instructions or study information and free text boxes.

Findings in Relation to Existing Literature

The results of our usability study reflect results from previous usability studies to some extent [24]. Although our participants were positive about their overall experience of using Atom5, their satisfaction scores were slightly lower than in other studies [36,45,46], reminding us that barriers to the use of ePRO platforms still exist. Participants' lower satisfaction scores could be attributed to several factors, including visual difficulty, fatigue, cognitive impairment, and some questions difficult to understand and answer.

As we were aware that a TBI, even a mild TBI, can cause fatigue and attention deficit, among many other symptoms [47,48], we deliberately kept the number of questions as low as possible while being useful enough for the clinical team. However, a few participants started to feel tired a few minutes into the testing sessions. This is sometimes known as "respondent fatigue," and it refers to the way respondents lose concentration as they complete questionnaires, especially in the latter part of the questionnaires [49]. The fact that all the participants managed to complete the questionnaires and the tasks required is encouraging and suggests that they believed in the potential of the ePRO platform enough to complete the testing sessions [50].

In addition, several participants had to read some of the questions several times, especially questions from the TBI QOL SF questionnaire. Even after reading the questions several times, they did not always understand them. While this could be due to a lack of concentration, it is also worth wondering if the

XSL•FO
RenderX

wording of the questionnaires is suitable for individuals living with the effects of a TBI and whether the measure is suitable for use in routine care. As this issue might not directly relate to the ePRO platform, it would be important to explore this further in a future study.

One of the main risks of using ePRO platforms is excluding certain groups of patients [19,51]. This concerned 2 specific groups in our study. First, 3 participants who were visually impaired after their TBI had to be helped with reading the questions. Visual impairment is a common consequence of TBI [52], and therefore it is essential to include individuals who have visual impairment after a TBI. The second group of people who risks of being excluded is people who do not own an electronic device or are not familiar with technology [19]. A total of 3 of our participants did not have a smartphone or tablet and borrowed someone else's device in order to take part in the usability study. This could, however, raise issues around data privacy and security.

Finally, 1 participant felt they would not remember to use the electronic platform unless they were sent a reminder and suggested being sent a reminder, which has been shown to increase response rate [53].

Implications for ePRO Developers, Health Care Professionals, and Researchers

The fact that some of our participants were visually impaired should not exclude them from using ePROMs, as demonstrated in our study. In the case of our participants with visual impairment, the researcher read the questions. Therefore, ePRO developers, health care providers, and researchers should ensure electronic platforms are accessible for all, to ensure inclusivity and maximize participation [46] by allowing family members or carers to complete proxy measures [54] or by providing a "Read aloud" option for patients.

Similarly, although this was not the case with our participants, the suggestion that some people who had a TBI might struggle with downloading the app suggests issues around digital literacy. ePRO developers and researchers should ensure participants receive adequate training or support. The fact that 3 of our participants did not have access to a mobile device indicates that electronic devices should be provided to patients to allow them to complete ePROMs. This would also remove privacy and security issues if using their own devices or borrowing someone else's. Maximizing participation could also be increased if ePRO developers sent reminders to participants.

Several participants mentioned they would have liked to see their previous answers to the questionnaires in order to view their progress over time. Therefore, adding visual representations of users' previous answers would help them to do so.

Finally, it is recommended for health care professionals and researchers to use validated PROMs.

Strengths and Limitations

The main strength of our study is the Think Aloud technique used for data collection; allowing participants to share their views while carrying out the requested tasks provides real-time feedback [44,55], in particular, about the questions from the TBI QOL that they had difficulty understanding and responding to.

There are several limitations to our study. The first one is the lack of ethnic diversity among participants, meaning that study findings might not be representative of the wider population. The research team recognizes the need to promote inclusive data collection. In addition, the lack of text-to-speech or voice commands in the app made it more difficult for participants with vision impairment to use the app.

The time it took participants to complete the tasks was not representative of the time it would take to complete the ePROs, as they shared their views while carrying out the tasks. This made it difficult to assess efficiency of the electronic platform.

The fact that, despite excellent psychometric characteristics and a reading age of fifth-grade level (10/11 years old), some of the participants struggled to understand and answer some of the questions, implying that their answers did not reflect how they felt.

Finally, the thinking aloud approach used in this usability study means that some of the participants could become a little tired during the testing sessions because of the cognitive issues some of them faced. Researchers should ask participants to reflect on their experience after each task rather than at the end of the session.

Conclusion

Usability testing is a vital stage of any new product development. In our case, it showed that although individuals living with the effects of a TBI can report their symptoms on an electronic platform, ePRO developers, health care providers, and researchers should ensure that electronic platforms are inclusive and can be adapted to people's needs.

Future research should include testing the Atom5 platform in TBI clinic settings and with ethnic minorities.

Acknowledgments

The authors wish thank Headway UK for their contribution to recruitment.

Data Availability

The datasets generated during and/or analyzed during this study are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Authors' Contributions

CM, GT, AR, TB, LN, JF, LF, and MC designed the study. CM completed the data collection and data analysis. CM drafted the manuscript. EHD/Aparito developed the Atom5 platform and provided technical support. All authors reviewed and provided feedback on the manuscript.

Conflicts of Interest

CM receives funding from the National Institute for Health Research (NIHR) Surgical Reconstruction and Microbiology Research Centre, the NIHR Blood and Transplant Research Unit (BTRU) in Precision Transplant and Cellular Therapeutics, Innovate UK. GT receives funding from the NIHR. AR receives funding from the NIHR Birmingham Biomedical Research Centre, University of Birmingham, and the NIHR ARC West Midlands, University of Birmingham. MC is an NIHR senior investigator and receives funding from the NIHR Birmingham Biomedical Research Centre, the NIHR Surgical Reconstruction and Microbiology Research Centre, and NIHR ARC West Midlands at the University of Birmingham and University Hospitals Birmingham NHS Foundation Trust, Health Data Research UK, UKRI, Macmillan Cancer Support, UCB Pharma, Gilead, Janssen, and GSK. MC has received personal fees from Astellas, Aparito Ltd, CIS Oncology, Takeda, Merck, Daiichi Sankyo, Glaukos, GSK, and the Patient-Centred Outcomes Research Institute (PCORI) outside the submitted work. DK reports grants from Macmillan Cancer Support, the NIHR, NIHR Birmingham Biomedical Research Centre, and NIHR SRMRC at the University of Birmingham and University Hospitals Birmingham NHS Foundation Trust, and personal fees from Merck and GSK outside the submitted work.

Multimedia Appendix 1
Screenshots of Atom5 app and clinical dashboard.
[DOCX File, 395 KB - formative_v9i1e58128_app1.docx]

Checklist 1
COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist.
[PDF File, 599 KB - formative_v9i1e58128_app2.pdf]

References

1. Head injury. Quality standards and indicators: briefing paper. : National Institute for Health and Care Excellence; 2014 URL: https://www.nice.org.uk/guidance/qs74/resources/head-injury-briefing-paper2 [accessed 2024-11-29]
2. Feigin VL, Theadom A, Barker-Collo S, et al. Incidence of traumatic brain injury in New Zealand: a population-based study. Lancet Neurol 2013 Jan;12(1):53-64. [doi: 10.1016/S1474-4422(12)70262-4] [Medline: 23177532]
3. Traumatic brain injury across the life course: priorities, challenges, and opportunities. : UK Research and Innovation/Medical Research Council; 2022 URL: https://www.ukri.org/wp-content/uploads/2022/12/MRC-07122022-MRC-Traumatic-Brain-Injury-Workshop-report_June-2022.pdf [accessed 2024-11-29]
4. Maas AIR, Menon DK, Manley GT, et al. Traumatic brain injury: progress and challenges in prevention, clinical care, and research. Lancet Neurol 2022 Nov;21(11):1004-1060. [doi: 10.1016/S1474-4422(22)00309-X] [Medline: 36183712]
5. Levin HS, Diaz-Arrastia RR. Diagnosis, prognosis, and clinical management of mild traumatic brain injury. Lancet Neurol 2015 May;14(5):506-517. [doi: 10.1016/S1474-4422(15)00002-2] [Medline: 25801547]
6. Varner C, Thompson C, de Wit K, Borgundvaag B, Houston R, McLeod S. Predictors of persistent concussion symptoms in adults with acute mild traumatic brain injury presenting to the emergency department. CJEM 2021 May;23(3):365-373. [doi: 10.1007/s43678-020-00076-6] [Medline: 33555602]
7. Masel BE, DeWitt DS. Traumatic brain injury: a disease process, not an event. J Neurotrauma 2010 Aug;27(8):1529-1540. [doi: 10.1089/neu.2010.1358] [Medline: 20504161]
8. Ryan LM, Warden DL. Post concussion syndrome. Int Rev Psychiatry 2003 Nov;15(4):310-316. [doi: 10.1080/09540260310001606692] [Medline: 15276952]
9. Traumatic brain injury (TBI). Headway. 2022. URL: https://www.headway.org.uk/about-brain-injury/individuals/types-of-brain-injury/traumatic-brain-injury [accessed 2024-12-04]
10. Tran S, Kenny B, Power E, et al. Cognitive-communication and psychosocial functioning 12 months after severe traumatic brain injury. Brain Inj 2018;32(13-14):1700-1711. [doi: 10.1080/02699052.2018.1537006] [Medline: 30346807]
11. Voormolen DC, Polinder S, von Steinbuechel N, Vos PE, Cnossen MC, Haagsma JA. The association between post-concussion symptoms and health-related quality of life in patients with mild traumatic brain injury. Injury 2019 May;50(5):1068-1074. [doi: 10.1016/j.injury.2018.12.002] [Medline: 30554897]
12. Karvandi E, Barrett L, Newcombe V, Hutchinson P, Helmy A. Digital health interventions for remote follow-up after mild traumatic brain injury. Br J Neurosurg 2024 May 6;6(1-7):1-7. [doi: 10.1080/02688697.2024.2346564] [Medline: 38711206]
13. Lawrence T, Helmy A, Bouamra O, Woodford M, Lecky F, Hutchinson PJ. Traumatic brain injury in England and Wales: prospective audit of epidemiology, complications and standardised mortality. BMJ Open 2016 Nov 24;6(11):e012197. [doi: 10.1136/bmjopen-2016-012197] [Medline: 27884843]

XSL•FO
RenderX

14. Dams-O'Connor K, Juengst SB, Bogner J, et al. Traumatic brain injury as a chronic disease: insights from the United States Traumatic Brain Injury Model Systems Research Program. Lancet Neurol 2023 Jun;22(6):517-528. [doi: 10.1016/S1474-4422(23)00065-0] [Medline: 37086742]
15. Izzy S, Chen PM, Tahir Z, et al. Association of traumatic brain injury with the risk of developing chronic cardiovascular, endocrine, neurological, and psychiatric disorders. JAMA Netw Open 2022 Apr 1;5(4):e229478. [doi: 10.1001/jamanetworkopen.2022.9478] [Medline: 35482306]
16. Brett BL, Gardner RC, Godbout J, Dams-O'Connor K, Keene CD. Traumatic brain injury and risk of neurodegenerative disorder. Biol Psychiatry 2022 Mar 1;91(5):498-507. [doi: 10.1016/j.biopsych.2021.05.025] [Medline: 34364650]
17. Livingston G, Huntley J, Liu KY, et al. Dementia prevention, intervention, and care: 2024 report of the Lancet standing Commission. Lancet 2024 Aug 10;404(10452):572-628. [doi: 10.1016/S0140-6736(24)01296-0] [Medline: 39096926]
18. Calvert M, Kyte D, Price G, Valderas JM, Hjollund NH. Maximising the impact of patient reported outcome assessment for patients and society. BMJ 2019 Jan 24;364:k5267. [doi: 10.1136/bmj.k5267] [Medline: 30679170]
19. Meirte J, Hellemans N, Anthonissen M, et al. Benefits and disadvantages of electronic patient-reported outcome measures: systematic review. JMIR Perioper Med 2020 Apr 3;3(1):e15588. [doi: 10.2196/15588] [Medline: 33393920]
20. Glenwright BG, Simmich J, Cottrell M, et al. Facilitators and barriers to implementing electronic patient-reported outcome and experience measures in a health care setting: a systematic review. J Patient Rep Outcomes 2023 Feb 14;7(1):13. [doi: 10.1186/s41687-023-00554-2] [Medline: 36786914]
21. Shaw T, McGregor D, Brunner M, Keep M, Janssen A, Barnet S. What is eHealth (6)? Development of a conceptual model for eHealth: qualitative study with key informants. J Med Internet Res 2017 Oct 24;19(10):e324. [doi: 10.2196/jmir.8106] [Medline: 29066429]
22. Cao J, Lim Y, Sengoku S, Guo X, Kodama K. Exploring the shift in International trends in mobile health research from 2000 to 2020: bibliometric analysis. JMIR Mhealth Uhealth 2021 Sep 8;9(9):e31097. [doi: 10.2196/31097] [Medline: 34494968]
23. Guo X, Chen S, Zhang X, Ju X, Wang X. Exploring patients' intentions for continuous usage of mHealth services: elaboration-likelihood perspective study. JMIR Mhealth Uhealth 2020 Apr 6;8(4):e17258. [doi: 10.2196/17258] [Medline: 32250277]
24. Bonn MM, Graham LJ, Marrocco S, Jeske S, Moran B, Wolfe DL. Usability evaluation of a self-management mobile application for individuals with a mild traumatic brain injury. Dig Health 2023;9:20552076231183555. [doi: 10.1177/20552076231183555] [Medline: 37426589]
25. Ganefianty A, Songwathana P, Damkliang J, Imron A, Latour JM. A mobile health transitional care intervention delivered by nurses improves postdischarge outcomes of caregivers of patients with traumatic brain injury: a randomized controlled trial. World Neurosurg 2024 Apr;184:191-201. [doi: 10.1016/j.wneu.2024.01.078] [Medline: 38244683]
26. Sánchez-Rodríguez MT, Pinzón-Bernal MY, Jiménez-Antona C, et al. Designing an informative app for neurorehabilitation: a feasibility and satisfaction study by physiotherapists. Healthcare (Basel) 2023 Sep 14;11(18):2549. [doi: 10.3390/healthcare11182549] [Medline: 37761746]
27. McMullan C, Retzer A, Slade A, et al. Care providers' and patients' attitudes toward using electronic-patient reported outcomes to support patients with traumatic brain injury: a qualitative study (PRiORiTy). Brain Inj 2020 May 11;34(6):723-731. [doi: 10.1080/02699052.2020.1740944] [Medline: 32223338]
28. Velikova G, Wright EP, Smith AB, et al. Automated collection of quality-of-life data: a comparison of paper and computer touch-screen questionnaires. J Clin Oncol 1999 Mar;17(3):998-1007. [doi: 10.1200/JCO.1999.17.3.998] [Medline: 10071295]
29. Schougaard LMV, Larsen LP, Jessen A, et al. AmbuFlex: tele-patient-reported outcomes (telePRO) as the basis for follow-up in chronic and malignant diseases. Qual Life Res 2016 Mar;25(3):525-534. [doi: 10.1007/s11136-015-1207-0] [Medline: 26790427]
30. Holch P, Warrington L, Bamforth LCA, et al. Development of an integrated electronic platform for patient self-report and management of adverse events during cancer treatment. Ann Oncol 2017 Sep 1;28(9):2305-2311. [doi: 10.1093/annonc/mdx317] [Medline: 28911065]
31. Aiyegbusi OL. Key methodological considerations for usability testing of electronic patient-reported outcome (ePRO) systems. Qual Life Res 2020 Feb;29(2):325-333. [doi: 10.1007/s11136-019-02329-z] [Medline: 31691202]
32. Jaspers MWM. A comparison of usability methods for testing interactive health technologies: methodological aspects and empirical evidence. Int J Med Inform 2009 May;78(5):340-353. [doi: 10.1016/j.ijmedinf.2008.10.002] [Medline: 19046928]
33. 20282-2:2013(en) IT. Usability of consumer products and products for public use — part 2: summative test method. : International Organization for Standardization; 2013 URL: https://www.iso.org/standard/62733.html [accessed 2024-11-29]
34. Faulkner L. Beyond the five-user assumption: benefits of increased sample sizes in usability testing. Behav Res Methods Instrum Comput 2003 Aug;35(3):379-383. [doi: 10.3758/bf03195514] [Medline: 14587545]
35. Hwang W, Salvendy G. Number of people required for usability evaluation. Commun ACM 2010 May;53(5):130-133. [doi: 10.1145/1735223.1735255]
36. McMullan C, Hughes SE, Aiyegbusi OL, Calvert M. Usability testing of an electronic patient-reported outcome system linked to an electronic chemotherapy prescribing and patient management system for patients with cancer. Heliyon 2023 Jun;9(6):e16453. [doi: 10.1016/j.heliyon.2023.e16453] [Medline: 37260889]

XSL•FO
RenderX

37. Macefield R. How to specify the participant group size for usability studies: a practitioner's guide. J Usability Stud 2009;5(1):34-45.
38. Aparito. URL: https://www.aparito.com [accessed 2024-12-24]
39. Sapra A, Bhandari P, Sharma S, Chanpura T, Lopp L. Using Generalized Anxiety Disorder-2 (GAD-2) and GAD-7 in a primary care setting. Cureus 2020 May 21;12(5):e8224. [doi: 10.7759/cureus.8224] [Medline: 32582485]
40. Plummer F, Manea L, Trepel D, McMillan D. Screening for anxiety disorders with the GAD-7 and GAD-2: a systematic review and diagnostic metaanalysis. Gen Hosp Psychiatry 2016;39:24-31. [doi: 10.1016/j.genhosppsych.2015.11.005] [Medline: 26719105]
41. Kroenke K, Spitzer RL, Williams JBW. The Patient Health Questionnaire-2: validity of a two-item depression screener. Med Care 2003 Nov;41(11):1284-1292. [doi: 10.1097/01.MLR.0000093487.78664.3C] [Medline: 14583691]
42. Lang AJ, Stein MB. An abbreviated PTSD checklist for use as a screening instrument in primary care. Behav Res Ther 2005 May;43(5):585-594. [doi: 10.1016/j.brat.2004.04.005] [Medline: 15865914]
43. Heinemann AW, Kisala PA, Boulton AJ, et al. Development and calibration of the TBI-QOL ability to participate in social roles and activities and TBI-QOL satisfaction with social roles and activities item banks and short forms. Arch Phys Med Rehabil 2020 Jan;101(1):20-32. [doi: 10.1016/j.apmr.2019.07.015] [Medline: 31473208]
44. Eccles DW, Arsal G. The think aloud method: what is it and how do I use it? Qual Res Sport Exerc Health 2017 Aug 8;9(4):514-531. [doi: 10.1080/2159676X.2017.1331501]
45. Aiyegbusi OL, Kyte D, Cockwell P, et al. Development and usability testing of an electronic patient-reported outcome measure (ePROM) system for patients with advanced chronic kidney disease. Comput Biol Med 2018 Oct 1;101:120-127. [doi: 10.1016/j.compbiomed.2018.08.012] [Medline: 30130638]
46. McMullan C, Retzer A, Hughes SE, et al. Development and usability testing of an electronic patient-reported outcome (ePRO) solution for patients with inflammatory diseases in an Advanced Therapy Medicinal Product (ATMP) basket trial. J Patient Rep Outcomes 2023 Oct 9;7(1):98. [doi: 10.1186/s41687-023-00634-3] [Medline: 37812323]
47. Maas AIR, Menon DK, Adelson PD, et al. Traumatic brain injury: integrated approaches to improve prevention, clinical care, and research. Lancet Neurol 2017 Dec;16(12):987-1048. [doi: 10.1016/S1474-4422(17)30371-X] [Medline: 29122524]
48. Nelson LD, Temkin NR, Dikmen S, et al. Recovery after mild traumatic brain injury in patients presenting to US level I trauma centers: a Transforming Research and Clinical Knowledge in Traumatic Brain Injury (TRACK-TBI) study. JAMA Neurol 2019 Sep 1;76(9):1049-1059. [doi: 10.1001/jamaneurol.2019.1313] [Medline: 31157856]
49. Hochheimer CJ, Sabo RT, Krist AH, Day T, Cyrus J, Woolf SH. Methods for evaluating respondent attrition in web-based surveys. J Med Internet Res 2016 Nov 22;18(11):e301. [doi: 10.2196/jmir.6342] [Medline: 27876687]
50. O'Reilly-Shah VN. Factors influencing healthcare provider respondent fatigue answering a globally administered in-app survey. PeerJ 2017;5:e3785. [doi: 10.7717/peerj.3785] [Medline: 28924502]
51. Calvert MJ, Cruz Rivera S, Retzer A, et al. Patient reported outcome assessment must be inclusive and equitable. Nat Med 2022 Jun;28(6):1120-1124. [doi: 10.1038/s41591-022-01781-8] [Medline: 35513530]
52. Armstrong RA. Visual problems associated with traumatic brain injury. Clin Exp Optom 2018 Nov;101(6):716-726. [doi: 10.1111/cxo.12670] [Medline: 29488253]
53. Triplet JJ, Momoh E, Kurowicki J, Villarroel LD, Law TY, Levy JC. E-mail reminders improve completion rates of patient-reported outcome measures. JSES Open Access 2017 Mar;1(1):25-28. [doi: 10.1016/j.jses.2017.03.002] [Medline: 30675535]
54. Holden RJ, Valdez RS, Schubert CC, Thompson MJ, Hundt AS. Macroergonomic factors in the patient work system: examining the context of patients with chronic illness. Ergonomics 2017 Jan;60(1):26-43. [doi: 10.1080/00140139.2016.1168529] [Medline: 27164171]
55. Usability. Digital.gov. 2013. URL: https://www.usability.gov/get-involved/blog/2013/04/moderating-usability-tests.html [accessed 2024-12-04]

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Research
ePRO: electronic patient-reported outcomes
ePROM: electronic patient-reported outcome measures
GAD-2: Generalized Anxiety Disorder-2 measure
ISO: International Organisation for Standardization
mTBI: mild traumatic brain injury
NIHR SRMRC: National Institute for Health and Care Research Surgical Reconstruction and Microbiology Research Centre
PCL-2: Posttraumatic Stress Disorder Checklist 2
PHQ-2: Patient Health Questionnaire-2
PRiORiTy : Patient-Reported Outcomes Research in Trauma
PRO: patient-reported outcomes

XSL•FO
RenderX

PROM: patient-reported outcome measure
QOLIBRI: Quality of Life after Brain Injury
TBI: traumatic brain injury
TBI QOL SF: Traumatic Brain Injury – Quality of Life Short form

Edited by A Mavragani; submitted 12.03.24; peer-reviewed by H Mühlan, S Juengst; revised version received 08.11.24; accepted 10.11.24; published 23.01.25.

Please cite as:
McMullan C, Turner G, Retzer A, Belli A, Davies EH, Nice L, Flavell L, Flavell J, Calvert M
Testing an Electronic Patient-Reported Outcome Platform in the Context of Traumatic Brain Injury: PRiORiTy Usability Study
JMIR Form Res 2025;9:e58128
URL: https://formative.jmir.org/2025/1/e58128
doi:10.2196/58128

© Christel McMullan, Grace Turner, Ameeta Retzer, Antonio Belli, Elin Haf Davies, Laura Nice, Luke Flavell, Jackie Flavell, Melanie Calvert. Originally published in JMIR Formative Research (https://formative.jmir.org), 23.1.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Association of Psychological Factors With Willingness to Share Health-Related Data From Technological Devices: Cross-Sectional Questionnaire Study

Marijn Eversdijk, MSc; Emma Rixt Douma, MSc; Mirela Habibovic, PhD; Willem Johan Kop, PhD

Department of Medical and Clinical Psychology, Center of Research on Psychological Disorders and Somatic Diseases (CoRPS), Tilburg University, Tilburg, the Netherlands

Corresponding Author:
Marijn Eversdijk, MSc
Department of Medical and Clinical Psychology, Center of Research on Psychological Disorders and Somatic Diseases (CoRPS), Tilburg University, Tilburg, the Netherlands

Abstract

Background: Health-related data from technological devices are increasingly obtained through smartphone apps and wearable devices. These data could enable physicians and other care providers to monitor patients outside the clinic or assist individuals in improving lifestyle factors. However, the use of health technology data might be hampered by the reluctance of patients to share personal health technology data because of the privacy sensitivity of this information.

Objective: This study investigates to what extent psychological factors play a role in people's willingness to share personal health technology data.

Methods: Data for this cross-sectional study were obtained by quota sampling based on age and sex in a community-based sample (N=1013; mean age 48.6, SD 16.6 years; 522/1013, 51.5% women). Willingness to share personal health technology data and related privacy concerns were assessed using an 8-item questionnaire with good psychometric properties (Cronbach's α=0.82). Psychological variables were assessed using validated questionnaires for optimism (Life Orientation Test—Revised), psychological flexibility (Psychological Flexibility Questionnaire), negative affectivity (Type D Scale-14—Negative Affectivity), social inhibition (Type D Scale-14—Social Inhibition), generalized anxiety (Generalized Anxiety Disorder-7), and depressive symptoms (Patient Health Questionnaire-9). Data were analyzed using multiple linear regression analyses, and network analysis was used to visualize the associations between the item scores.

Results: Higher levels of optimism (β=.093; *P*=.004) and psychological flexibility (β=.127; *P*<.001) and lower levels of social inhibition (β=−.096; *P*=.002) were significantly associated with higher levels of willingness to share health technology data when adjusting for age, sex, and education level in separate regression models. Other associations with psychological variables were not statistically significant. Network analysis revealed that psychological flexibility clustered more with items that focused on the benefits of sharing data, while optimism was negatively associated with privacy concerns.

Conclusions: The current results suggest that people with higher levels of optimism and psychological flexibility and those with lower social inhibition levels are more likely to share health technology data. The magnitude of the effect sizes was low, and future studies with additional psychological measures are needed to establish which factors identify people who are reluctant to share their data such that optimal use of devices in health care can be facilitated.

(JMIR Form Res 2025;9:e64244) doi:10.2196/64244

KEYWORDS

health data sharing; privacy concerns; wearable health technology; personality; psychological flexibility; optimism; social inhibition; psychological factors; willingness; health-related data; mobile phone

Introduction

Recent advances in health technology, including wearable devices and smartphone apps, have facilitated remote patient health monitoring [1-3]. Sharing data from these devices with care providers, researchers, or device developers has multiple advantages for the patient, including early detection of health-related abnormalities, risk stratification, individualized health interventions, and overall better disease management [4,5]. The health care system might also benefit by applying machine learning algorithms to the large amounts of collected data to improve future health care and increase cost-efficiency [6].

Despite the potential health benefits, uptake of these health technologies is often impeded by patient and other user concerns related to the privacy and confidentiality of sharing sensitive

XSL•FO
RenderX

personal data [7-9]. These privacy-related concerns are relevant because they influence the willingness to share their health technology data [10,11]. While respect for personal autonomy in the decision to share data is an important cornerstone in medical ethics, privacy-related objections could be overcome by adjusting the technological features of a device based on specific desires or patient characteristics [12]. Assessing individual differences in willingness to share technology data could help develop a personalized approach toward sharing health data to facilitate lower privacy concerns, higher user satisfaction, higher satisfaction with care, and better health outcomes.

Willingness to share personal health technology data can be defined as an individual tendency to allow others to have access to personal biobehavioral (eg, heart rate) or other personal data (eg, geographic location) collected using health technology (eg, wearable devices and smartphone apps) [11,13,14]. In general, the extent to which a person is willing to share or protect personal data is determined by an interaction between the sensitivity of the information (eg, medical or socioeconomic information, purchase history, and location), the context of data sharing (eg, health care or commercial setting), and personal dispositions such as privacy-related attitudes [15]. These personal dispositions are most relevant for the question of who is willing to share personal health technology data. Research has identified not only digital literacy, older age, and privacy knowledge as factors associated with willingness to share technology data but also psychological factors such as personality traits and generalized trust [13,16,17]. However, there is a knowledge gap regarding the magnitude of associations between psychological factors and willingness to share personal health technology data.

Current evidence regarding positive psychological factors, although scarce, indicates that patients who are overall more optimistic are more likely to expect positive outcomes from sharing their data (eg, better health monitoring) or are more likely to believe that they are less at risk of privacy breaches than others [18,19]. Optimism is also associated with higher psychological flexibility, which is the ability to distance from current mindsets and consider other possible mindsets [20,21]. People with high levels of psychological flexibility might be better able to act upon goals that involve new technologies to facilitate better personalized care that might also help other patients by sharing data even when this comes at the cost of less privacy for the user.

In addition to positive psychological traits, negative psychological factors are also expected to be associated with the willingness to share health-related data. Experiencing distress (eg, anxiety and depressive symptoms) is linked to threat-avoidant behaviors and reduced engagement in reward-seeking behaviors [22]. These tendencies might translate into overvaluing the risks and undervaluing the potential rewards of sharing data, leading to a lower willingness to share health technology data. Prior research has shown that anxiety and depressive symptoms are more likely to occur in patients with negative personality traits such as negative affectivity (NA) and social inhibition (SI) [23]. The relative importance of positive psychological characteristics versus negative characteristics with the willingness to share health technology data has yet to be researched.

This study explores associations between psychological factors and the willingness to share personal health technology data. It was expected that positive psychological characteristics (ie, optimism and psychological flexibility) would be associated with a higher willingness to share health technology data, whereas negative psychological characteristics (ie, anxiety, depressive symptoms, and measures of distress-related personality traits: NA and SI) would be associated with less willingness to share health technology data. The role of sociodemographic background factors (age, sex, and education level) and whether or not participants have previously used health technology apps was also investigated using multivariable models. Exploratory network analyses on the questionnaire items were conducted to reveal potential psychological mechanisms between the participant responses on the psychological constructs and items that reflect the willingness to share data.

Methods

Participants and Procedure

This cross-sectional study was based on data obtained from a Dutch community-based sample (N=1013) as part of an annually recurring survey study (for additional details, see the study by Kop et al [24]). Research assistants were instructed to apply quota sampling within their personal network by approaching a minimum number of participants from each sex and age group to ensure equal representation (eg, to ensure the number of men aged between 31 and 40 years in the total sample is equal to the number of women aged between 31 and 40 years in the total sample). Inclusion criteria were age above 18 years and being fluent in Dutch. Although there were no geographical restrictions, most participants lived in the south of the Netherlands, as a result of the sampling procedure. Data for this study were collected between February 2022 and April 2022. Data collection took place online using Qualtrics.

Ethical Considerations

Ethical approval was obtained from the Tilburg School of Social and Behavioral Sciences ethical review board (protocol RP55). All participants gave informed consent before taking part in this study, and their data were deidentified prior to data analysis. Participants were not compensated for their participation.

Measures

Data collection was conducted using a web-based survey after participants provided informed consent for their participation. All data were obtained using validated questionnaires or purpose-designed questions (see Multimedia Appendix 1 for an overview of the used questionnaires).

Health Technology Data-Sharing Inventory

The willingness to share health technology data was assessed using an 8-item questionnaire designed for the purposes of this study. Six questions focused on the potential benefits of sharing data (eg, “I would allow my personal data to be shared if it would help me”). Two items were derived from the Service

XSL•FO
RenderX

User Technology Acceptability Questionnaire and focused on the feeling that health technology violates confidentiality or people's privacy [25]. Participants were asked to indicate their level of agreement on a 5-point Likert scale (1=strongly disagree, 5=strongly agree). After reverse-coding the 2 negatively phrased items, a continuous total score was calculated by adding up the items (range 8-40). A higher score indicates a higher willingness to share health technology data. The Cronbach's α coefficient for this scale was 0.82 in this study, which indicates a good internal consistency of this inventory.

Optimism

Optimism was measured using the 10-item version of the Life Orientation Test—Revised (LOT-R) [26]. The LOT-R includes 3 items about optimism, 3 items about pessimism, and 4 filler items. Examples of questions are "I'm always optimistic about my future" (ie, optimism) and "I hardly ever expect things to go my way" (ie, pessimism). Participants were asked to indicate their level of agreement on a 5-point Likert scale (0=strongly disagree, 4=strongly agree). A continuous total score for optimism (range 0 - 24) was calculated by adding the scores on optimism to the reversed scores on pessimism. A higher score indicates higher levels of optimism. The published internal consistency of the LOT-R total score is acceptable (Cronbach's α=0.78) [26], which was similar and acceptable in the present sample (Cronbach's α=0.74).

Psychological Flexibility

The Psychological Flexibility Questionnaire (PFQ) [27] was used to measure psychological flexibility. Participants were asked to indicate their degree of agreement on a 5-point Likert scale (1=strongly disagree, 5=strongly agree), enabling participants to enter a neutral response (3=neither agree nor disagree). Examples of items are "I feel open to changes" and "Concepts may possess different meanings when perceived in different contexts." The answers to all 20 questions were added up to a continuous total score (range 20-100). A higher score indicates higher psychological flexibility. Internal consistency of the questionnaire was good in the present sample (Cronbach's α=0.89).

Negative Affectivity and Social Inhibition

NA and SI were measured using the DS-14 (Type D Scale-14) [28]. These scales have been developed to assess a construct identifying individuals with a distressed personality trait (ie, type D), but the subscales are used as separate dimensions in this paper, as the focus was not on the interaction of these 2 subscales, which is needed when investigating the type D construct [29]. The questionnaire contains a 7-item subscale assessing NA and a 7-item scale assessing SI. An example of a question for NA is "I often feel unhappy," and for SI, an example is "I am a closed kind of person." Two questions of the SI subscale are reverse-phrased and recoded before calculating total scores. Participants rated their agreement on a 5-point Likert scale (0=strongly disagree, 4=strongly agree). A continuous total score was calculated by adding the scores on each subscale (range 0 - 28). A higher score indicates a higher presence of the personality trait. The Cronbach's α value for NA in this sample was 0.88 and for SI 0.88, indicating good internal consistency.

Generalized Anxiety and Depressive Symptoms

The Generalized Anxiety Disorder-7 (GAD-7) questionnaire [30] was used to measure the presence of anxiety-related complaints over the last 2 weeks (example items are "Feeling nervous, anxious or on edge" and "Not being able to stop or control worrying"). Participants were asked to indicate the presence of these complaints on a 4-point Likert scale (0=not at all, 3=nearly every day). A continuous total score was calculated by adding the scores of the 7 items (range 0 - 21). A higher score indicates a higher presence of anxiety-related complaints. The Cronbach's α value in this study was 0.87, indicating good internal consistency.

Depressive symptoms were assessed using the Patient Health Questionnaire-9 (PHQ-9) [31], asking about the presence of depressive symptoms such as "Little interest or pleasure in doing things" or "Feeling down, depressed, or hopeless." Questions focused on the presence of complaints over the past 2 weeks and had to be filled in on a 4-point Likert scale (0=not at all, 3=nearly every day). A continuous total score was calculated by adding the scores on all 9 items (range 0-27). A higher score indicates a higher presence of depressive symptoms. Internal consistency was considered good, with a Cronbach's α value of 0.85 in the present sample.

Background Variables and Covariates

Demographic measures (age, self-reported sex, and education level) were obtained from the self-report questionnaire. The dataset contained only a measurement for sex and not for gender. Although gender would have better matched the psychological and behavioral variables in the dataset, the authors decided to report sex since this measure better represents the construct measured in the questionnaire. Level of education was divided into low (completed secondary school or less), middle (completed vocational education or high school only), or high (completed college or higher). Participants also reported their health status by indicating the presence of a lifetime diagnosed medical condition (eg, cardiovascular, cancer, lung disease, gastrointestinal disorder, and diabetes mellitus). In addition, an assessment was made concerning participants' use of smartphone-based health apps or related wearable technology.

Statistical Analysis

Descriptive statistics were obtained for all demographic and psychological variables by calculating the mean and SD for continuous variables and the number of occurrences and percentages for all categorical variables. Pearson and nonparametric Spearman correlation analyses were used to obtain an overview of the bivariate associations between the psychological measures and willingness to share health technology data. For illustrative purposes, odds ratios (ORs) with 95% CIs were also calculated using tertile-based dichotomized data of the psychological measures and the willingness to share health data (a value higher than 1 indicates a higher chance of willingness to share data based on the upper tertile of the psychological measure).

XSL•FO
RenderX

Multiple linear regression analyses were performed to assess which variables were independently associated with the willingness to share health technology data. Assumptions for regression analysis, including linearity, normality of residuals, homoscedasticity, and the presence of influential cases, were evaluated using the plot function in R (version 4.2.2; R Foundation for Statistical Computing) (findings are shown in Table S1 and Figure S1 in Multimedia Appendix 2). Data are presented in terms of the overall explained variance of the model (R^2) and the standardized regression weights per variable (β). Analyses first focused on the results of the full sample (N=1013). Subsequent sensitivity analyses were conducted for participants who had used or currently used health technology apps (N=244) and those not using such apps (N=769).

Total scores were coded as missing values if participants answered <60% of all items of that specific scale. If at least 60% of the items of a questionnaire were completed, then missing items of that scale were imputed by the mean of the answered questions of that scale. The percentage of missing values among all the psychological item scores was <0.1%, and there appeared to be no patterns in missingness. For the multiple regression models, listwise deletion was used to deal with missing values on the total scores of each predictor. The potential of biased parameters estimated as a result of mean imputation was limited, as it was applied only within variables, the percentage of missing values was extremely low, and listwise deletion was used [32].

Exploratory network analyses were performed to further investigate the associations between the item scores of all the psychological scales. Network analyses were visualized with the "qgraph" package (version 1.9.3) in R [33]. Each node represents an individual item of the questionnaire, while edges represent the associations between 2 nodes (ie, questionnaire items).

Results

Participants and Demographics

Table 1 displays the participant characteristics. The mean age was 48.6 (SD 16.6) years, and 51.5% (522/1013) of the participants were women. Levels of education were relatively high (52.3% higher education). A current or history of a medical disorder was reported by 36.9% (374/1013) of the participants, and average anxiety and depressive symptom levels were relatively low.

Table . Characteristics of all included participants.

	Total sample (N=1013)	Technology users (N=244)	Not using health technology (N=769)	*P* value
Age (years), mean (SD)	48.6 (16.6)	46.3 (15.0)	49.4 (17.1)	.01
Sex (female), n (%)	522 (51.5)	127 (52.0)	395 (51.4)	.91
Education, n (%)				<.001
Low	129 (12.7)	26 (10.7)	103 (13.4)	
Middle	354 (34.9)	65 (26.6)	289 (37.6)	
High	530 (52.3)	153 (62.7)	377 (49.0)	
Lifetime medical disorder, n (%)	374 (36.9)	94 (38.5)	280 (36.4)	.60
Variables (range), mean (SD)				
Optimism (0-24)	15.5 (3.5)	16.0 (3.5)	15.4 (3.4)	.02
Psychological flexibility (20-100)	75.1 (8.6)	76.6 (9.1)	74.6 (8.4)	.002
Negative affectivity (0-28)	7.7 (5.2)	7.2 (4.8)	7.9 (5.4)	.09
Social inhibition (0-28)	8.7 (5.3)	8.7 (5.1)	8.7 (5.4)	.87
Generalized anxiety (0-21)	3.0 (3.4)	3.0 (3.3)	2.9 (3.4)	.77
Depressive symptoms (0-27)	3.5 (3.9)	3.4 (3.2)	3.6 (4.1)	.42
Health data sharing (8-40)	23.8 (6.0)	25.0 (5.6)	23.4 (6.0)	<.001

The variable willingness to share health technology was approximately normally distributed, with scores ranging from 8 to 40 (mean 23.8, SD 6.0, median 24.0) (Figure S2 in Multimedia Appendix 2). Furthermore, 24.1% (244/1013) of the participants were currently using digital health apps or other forms of health technology (ie, wearable devices or smartphone apps aimed at improving physical activity, sleep, and medication adherence, or at reducing stress, alcohol use, and smoking). Participants who used technological devices were younger, were more likely to have a high education level, and had higher psychological flexibility levels than participants who did not use such devices, whereas no differences were found on the other variables listed in Table 1.

Associations of Psychological Factors With Willingness to Share Health Technology Data

As shown in Table 2, high levels of optimism (r=0.07; P=.03) and psychological flexibility (r=0.13; P<.001) were significantly

XSL•FO
RenderX

correlated with higher levels of willingness to share health technology data. In addition, higher levels of SI were associated with a lower willingness to share health technology data (r=−0.08; P=.01). Other bivariate correlations between psychological factors and willingness to share data were nonsignificant. When repeating these analyses with dichotomized scores, it was found that participants scoring in the highest tertile of psychological flexibility were 63% more likely to score high (upper tertile) on willingness to share health technology data (OR 1.63, 95% CI 1.25-2.12). Participants scoring in the lowest tertile of SI were also more likely to be willing to share their data (OR 1.53, 95% CI 1.17-2.00), whereas none of the other dichotomized psychological measures revealed significant ORs with a high level of willingness to share data.

Table . Correlation and regression coefficients of a linear regression model with all psychological measures and willingness to share health technology data.

Variable	Spearman ρ	*P* value	Pearson *r*	*P* value	*B* (95% CI)	β	*P* value
Optimism (LOT-R[a])	0.07	.03	0.07	.03	0.094 (−0.043 to 0.231)	.054	.18
Psychological flexibility (PFQ[b])	0.13	<.001	0.13	<.001	0.067 (0.017 to 0.117)	.097	.008
Negative affectivity (DS-14: NA[c])	−0.04	.20	−0.04	.26	0.050 (−0.067 to 0.166)	.043	.41
Social inhibition (DS-14: SI[d])	−0.10	.002	−0.08	.01	−0.068 (−0.149 to 0.013)	−.061	.10
Generalized anxiety (GAD-7[e])	0.01	.80	0.01	.84	0.115 (−0.072 to 0.301)	.065	.23
Depressive symptoms (PHQ-9[f])	−0.01	.71	−0.02	.62	−0.097 (−0.257 to 0.063)	−.063	.23
Age	−0.08	.01	−0.09	.004	−0.031 (−0.055 to −0.007)	−.086	.01
Sex[g]	0.07	.04	0.07	.03	0.873 (0.124 to 1.622)	.073	.02
Education level[h]	−0.06	.05	−0.06	.06	−0.629 (−1.160 to −0.096)	−.074	.02

[a]LOT-R: Life Orientation Test—Revised.

[b]PFQ: Psychological Flexibility Questionnaire.

[c]DS-14 NA: Type D Scale-14 Negative Affectivity

[d]DS-14 SI: Type D Scale-14 Social Inhibition

[e]GAD-7: Generalized Anxiety Disorder-7.

[f]PHQ-9: Patient Health Questionnaire-9.

[g]Sex coded as 0 (female) and 1 (male; a positive value indicates that men have higher willingness to share health technology data than women).

[h]Education level coded as 0 (low), 1 (middle), and 2 (high; a positive value indicates that a higher education is associated with a higher willingness to share health technology data).

Of the background factors, male participants (Cohen d=−0.14; P=.03) and technology users (Cohen d=−0.25; P<.001) scored significantly higher on willingness to share technology data (Table S2 in Multimedia Appendix 2). When adjusting for age, sex, and education level in separate linear regression models, optimism (β=.093; P=.004), psychological flexibility (β=.127; P<.001), and SI (β=−.096; P=.002) remained significantly associated with willingness to share health technology data.

To evaluate which psychological factor was associated with willingness to share health technology data, independent of other psychological factors, a multiple linear regression analysis was used entering all psychological factors simultaneously while also adjusting for age, sex, and education level (adjusted R^2=0.03; P<.001). Higher psychological flexibility was the psychological variable that was independently associated with more willingness to share data (β=.097; P=.008) while adjusting for covariates. Other correlates of willingness to participate in health technology data sharing were younger age (β=−.086; P=.01), male sex (β=.073; P=.02), and lower education level (β=−.074; P=.02) (Table 2).

When repeating the analyses after stratifying the sample by the use of health technology, a similar pattern of results was found for both groups. Among health technology users (N=244), higher levels of psychological flexibility were independently associated with more willingness to share data (β=.204; P=.006). Other associations with the willingness to share data among technology

XSL•FO
RenderX

users were nonsignificant (adjusted R^2 for the full model=0.04; P=.03). In the subset of participants who did not use health technology (N=769), lower SI was associated with a higher willingness to share data (β=−.092; P=.03), while male sex (β=.076; P=.04) and lower education level (β=−.089; P=.02) were also independently associated with willingness to share data (Table S3 in Multimedia Appendix 2).

Network Analysis of the Interrelationship Between Psychological Factors and Willingness to Share Health Technology Data

Figure 1 shows the results of the network analysis based on the correlational structure of all questionnaire items. Every edge represents a correlation larger than 0.1. Each underlying construct is represented by a different color. The network visualizes the extent to which items cluster together, indicating the measurement of the same underlying construct. The network visualizes that the items on data sharing do not cluster together with most of the psychological predictors. Within the health technology data-sharing items, it appeared that items 7 (“I am worried about the confidentiality of the private information being exchanged by the health technology”) and 8 (“Health technology will violate my privacy”) were distinct from the other items (eg, “I would allow my personal data to be shared with my physician if it would improve my treatment”), based on their relative spatial position. The item-level comparisons show that items 7 and 8 cluster more toward items on optimism, while the other data-sharing items cluster more toward psychological flexibility. A regularized network analysis can be found in Figure S3 in Multimedia Appendix 2.

Figure 1. Network correlation structure of all psychological measurements and willingness to share health technology data.

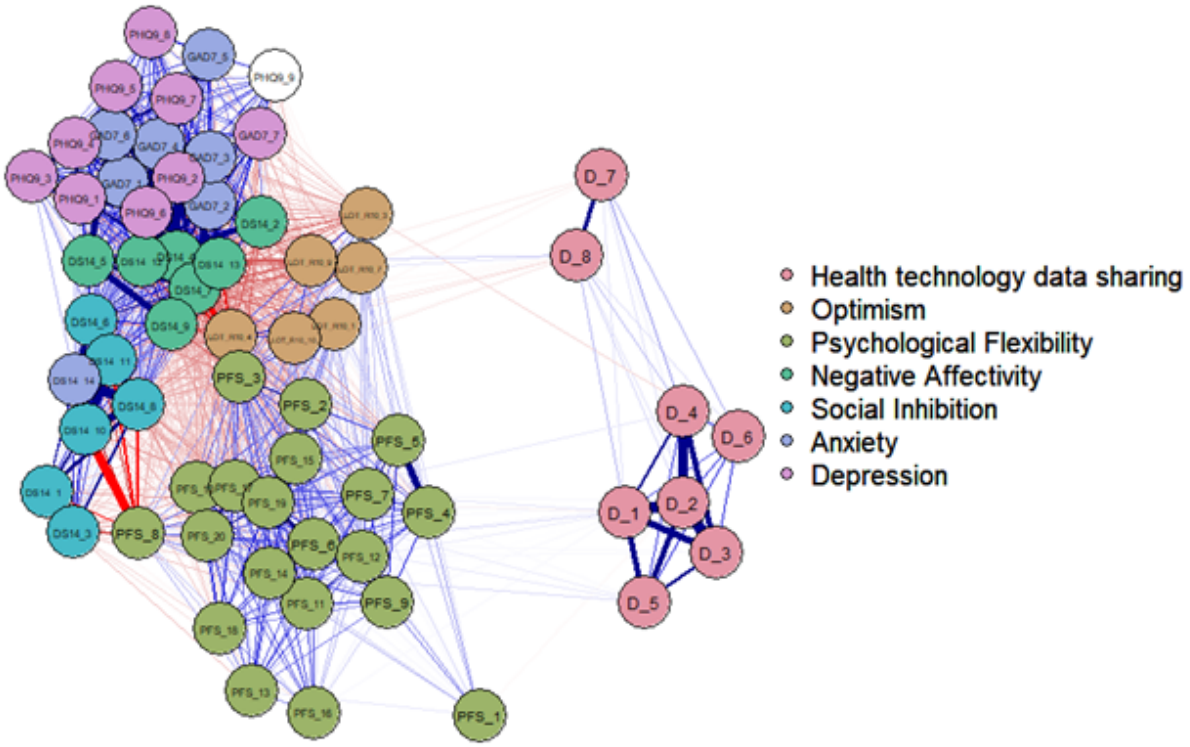

Discussion

Principal Findings

This study shows that higher levels of optimism and psychological flexibility and lower levels of SI are associated with higher levels of willingness to share health technology data. Multivariable analyses further indicated that psychological flexibility was the strongest and independent psychological factor associated with the willingness to share health technology data. Younger age, male sex, and lower education level were also independently associated with a higher willingness to share health technology data. Item-level network analysis revealed that optimism was associated with less concerns about confidentiality and privacy, whereas psychological flexibility was associated with a higher willingness to share health technology data if it could benefit the user or others.

Integration of the Present Findings With the Existing Literature

The current findings are consistent with previous research indicating that positive psychological traits, such as optimism and trust, are associated with a higher willingness to share health data [17,19]. The association of psychological flexibility with willingness to share health technology data has not yet been described in the scientific literature. People high on psychological flexibility are able to adapt to situational demands and balance competing desires, needs, and life domains [20].

XSL•FO
RenderX

Furthermore, they are more open toward new forms of behavior or shifting behavioral repertoires to achieve goals related to their personal values, which could play a central role in the decision to share personal health technology data with others. When balancing between giving up a part of their privacy compared and helping themselves or others by sharing their data, their prosocial values can take precedence over their protective intuitions, resulting into a higher willingness to share data. By shifting behaviors toward sharing data, people accept potential privacy concerns and remain true to personal values such as helping others, pursuing optimal health outcomes, or other values related to the benefits of data sharing [34].

Limitations and Strengths

The findings of this study need to be interpreted in the context of several limitations. The questionnaire to evaluate willingness to share health technology data has not been validated in prior settings, although the psychometric properties of the questionnaire were good. There was a lack of diversity in the sample with regard to education level and ethnic background. The levels of generalized anxiety and depressive symptoms were also low in this community-based sample, potentially resulting in a floor effect for their relationship with willingness to share health data. The cross-sectional design precludes causal inferences and a more in-depth qualitative research approach, such as conducting individual interviews or focus groups, might provide additional information about the underlying values and arguments in the decision to share health technology data. However, even with additional knowledge on these motivations, there is a gap between data-sharing intention and action [35]. This observation indicates the need for future studies measuring actual privacy-related behaviors instead of evaluating intentions to engage in such behaviors. A strength of this study is the large sample covering a wide range of ages and an equal number of women and men, in which the relative importance of multiple psychological constructs could be studied. However, the set of psychological constructs was limited in the already existing dataset, and the models indicated very small explained variance, which pleads for an extension of other relevant psychological factors, such as technological self-efficacy, performance expectancy, trust, or other personality traits [14,17,36-39].

Conclusions and Future Directions

This study documents a significant association of optimism, psychological flexibility, and low SI with a higher likelihood of individuals being willing to share health technology data. Of these variables, psychological flexibility appeared to be the strongest factor, particularly among individuals who use health technology. In addition, younger age, male sex, and low education levels were associated with a higher willingness to share health technology data, particularly among people who currently do not use health technology. Examining the underlying network structure of a large set of psychological variables revealed that psychological flexibility is mainly associated with the benefits of sharing health data, while people low on optimism show more concerns about confidentiality and privacy. Still, the magnitude of effect sizes was low, and additional work is needed on the underlying psychological mechanisms and reasons behind the disclosure and sharing of health-related data in general and health technology data in particular. Future studies that investigate psychological factors with reasons for sharing health data could identify which specific reasons are important for whom on an individual level instead of a population level. Additional studies with a broader set of psychological variables should be conducted to identify those variables that contribute the most to the willingness to share health data before specific clinical implications can be formulated. To increase the explained variance of the regression models, a broader set of psychological measures should be considered in studying their relationship with the willingness to share health technology data. Findings from these studies would provide important building blocks for the design of future health technology and promote individually tailored functionalities. This would enable customization of the data-sharing settings of personal health technology data sharing in accordance with individual preferences, rather than a binary decision to share all data or no data at all. The possibility to retract or modify prior consent to health technology data sharing with minimal effort could also increase people's initial participation in data sharing [40]. This personalized approach might result in less privacy concerns, higher user satisfaction with the technology, higher satisfaction with health care, and potentially also better health outcomes.

Data Availability

The datasets generated during and/or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Overview of questionnaire items used for analysis.
[DOCX File, 26 KB - formative_v9i1e64244_app1.docx]

Multimedia Appendix 2
Supplemental tables and figures with further data on continuous psychological variables, dichotomous background factors, willingness to share data, and the regularized correlation structure of all scales.
[DOCX File, 147 KB - formative_v9i1e64244_app2.docx]

XSL•FO
RenderX

References

1. Afshin A, Babalola D, Mclean M, et al. Information technology and lifestyle: a systematic evaluation of internet and mobile interventions for improving diet, physical activity, obesity, tobacco, and alcohol use. JAHA 2016 Sep;5(9). [doi: 10.1161/JAHA.115.003058]
2. Chaudhry B, Wang J, Wu S, et al. Systematic review: impact of health information technology on quality, efficiency, and costs of medical care. Ann Intern Med 2006 May 16;144(10):742-752. [doi: 10.7326/0003-4819-144-10-200605160-00125] [Medline: 16702590]
3. Habibović M, Rollman B. Technological innovations in biobehavioral and psychosomatic medicine. Psychosom Med 2023 Sep 1;85(7):565-567. [doi: 10.1097/PSY.0000000000001240] [Medline: 37506296]
4. Banerjee S, Hemphill T, Longstreet P. Wearable devices and healthcare: data sharing and privacy. Inf Soc 2018 Jan 1;34(1):49-57. [doi: 10.1080/01972243.2017.1391912]
5. Bayoumy K, Gaber M, Elshafeey A, et al. Smart wearable devices in cardiovascular care: where we are and how to move forward. Nat Rev Cardiol 2021 Aug;18(8):581-599. [doi: 10.1038/s41569-021-00522-7] [Medline: 33664502]
6. Bauer M, Glenn T, Monteith S, Bauer R, Whybrow PC, Geddes J. Ethical perspectives on recommending digital technology for patients with mental illness. Int J Bipolar Disord 2017 Dec;5(1):6. [doi: 10.1186/s40345-017-0073-9] [Medline: 28155206]
7. Martinez-Martin N, Luo Z, Kaushal A, et al. Ethical issues in using ambient intelligence in health-care settings. Lancet Digit Health 2021 Feb;3(2):e115-e123. [doi: 10.1016/S2589-7500(20)30275-2] [Medline: 33358138]
8. Segura Anaya LH, Alsadoon A, Costadopoulos N, Prasad PWC. Ethical implications of user perceptions of wearable devices. Sci Eng Ethics 2018 Feb;24(1):1-28. [doi: 10.1007/s11948-017-9872-8] [Medline: 28155094]
9. Baines R, Stevens S, Austin D, et al. Patient and public willingness to share personal health data for third-party or secondary uses: systematic review. J Med Internet Res 2024 Mar 5;26:e50421. [doi: 10.2196/50421] [Medline: 38441944]
10. Wetzels M, Broers E, Peters P, Feijs L, Widdershoven J, Habibovic M. Patient perspectives on health data privacy and management: “Where Is my data and whose is it?”. Int J Telemed Appl 2018;2018:3838747. [doi: 10.1155/2018/3838747] [Medline: 30631347]
11. Kim TK, Choi M. Older adults’ willingness to share their personal and health information when adopting healthcare technology and services. Int J Med Inform 2019 Jun;126:86-94. [doi: 10.1016/j.ijmedinf.2019.03.010] [Medline: 31029268]
12. Rubeis G, Schochow M, Steger F. Patient autonomy and quality of care in telehealthcare. Sci Eng Ethics 2018 Feb;24(1):93-107. [doi: 10.1007/s11948-017-9885-3]
13. Karampela M, Ouhbi S, Isomursu M. Connected health user willingness to share personal health data: questionnaire study. J Med Internet Res 2019;21(11):e14537. [doi: 10.2196/14537]
14. Rising CJ, Gaysynsky A, Blake KD, Jensen RE, Oh A. Willingness to share data from wearable health and activity trackers: analysis of the 2019 Health Information National Trends Survey Data. JMIR Mhealth Uhealth 2021 Dec 13;9(12):e29190. [doi: 10.2196/29190] [Medline: 34898448]
15. Brough AR, Martin KD. Critical roles of knowledge and motivation in privacy research. Curr Opin Psychol 2020 Feb;31:11-15. [doi: 10.1016/j.copsyc.2019.06.021] [Medline: 31376581]
16. Baruh L, Secinti E, Cemalcilar Z. Online privacy concerns and privacy management: a meta-analytical review. J Commun 2017 Feb;67(1):26-53. [doi: 10.1111/jcom.12276]
17. Bansal G, Zahedi FM, Gefen D. Do context and personality matter? Trust and privacy concerns in disclosing private information online. Inf Manag 2016 Jan;53(1):1-21. [doi: 10.1016/j.im.2015.08.001]
18. Kokolakis S. Privacy attitudes and privacy behaviour: a review of current research on the privacy paradox phenomenon. Comput Secur 2017 Jan;64:122-134. [doi: 10.1016/j.cose.2015.07.002]
19. Cho H, Lee JS, Chung S. Optimistic bias about online privacy risks: testing the moderating effects of perceived controllability and prior experience. Comput Human Behav 2010 Sep;26(5):987-995. [doi: 10.1016/j.chb.2010.02.012]
20. Kashdan TB, Rottenberg J. Psychological flexibility as a fundamental aspect of health. Clin Psychol Rev 2010 Nov;30(7):865-878. [doi: 10.1016/j.cpr.2010.03.001] [Medline: 21151705]
21. Demirtaş AS. Optimism and happiness in undergraduate students: cognitive flexibility and adjustment to university life as mediators. Anales de Psicología 2020;36(2):320-329. [doi: 10.6018/analesps.381181]
22. Bishop SJ, Gagne C. Anxiety, depression, and decision making: a computational perspective. Annu Rev Neurosci 2018 Jul 8;41:371-388. [doi: 10.1146/annurev-neuro-080317-062007] [Medline: 29709209]
23. Mols F, Denollet J. Type D personality among noncardiovascular patient populations: a systematic review. Gen Hosp Psychiatry 2010 Jan;32(1):66-72. [doi: 10.1016/j.genhosppsych.2009.09.010]
24. Kop WJ, Toussaint A, Mols F, Löwe B. Somatic symptom disorder in the general population: associations with medical status and health care utilization using the SSD-12. Gen Hosp Psychiatry 2019;56:36-41. [doi: 10.1016/j.genhosppsych.2018.10.004] [Medline: 30578985]
25. Hirani SP, Rixon L, Beynon M, et al. Quantifying beliefs regarding telehealth: Development of the Whole Systems Demonstrator Service User Technology Acceptability Questionnaire. J Telemed Telecare 2017 May;23(4):460-469. [doi: 10.1177/1357633X16649531]

XSL•FO
RenderX

26. Scheier MF, Carver CS, Bridges MW. Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem): a reevaluation of the Life Orientation Test. J Pers Soc Psychol 1994 Dec;67(6):1063-1078. [doi: 10.1037//0022-3514.67.6.1063] [Medline: 7815302]
27. Ben-Itzhak S, Bluvstein I, Maor M. The Psychological Flexibility Questionnaire (PFQ): development, reliability and validity. WebmedCentral Psychology 2014;5(4):WMC004606.
28. Denollet J. DS14: standard assessment of negative affectivity, social inhibition, and type D personality. Psychosom Med 2005;67(1):89-97. [doi: 10.1097/01.psy.0000149256.81953.49] [Medline: 15673629]
29. Lodder P. Modeling synergy: how to assess a Type D personality effect. J Psychosom Res 2020 May;132:109990. [doi: 10.1016/j.jpsychores.2020.109990] [Medline: 32151819]
30. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
31. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613. [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
32. Lodder P. To impute or not impute: that's the question. In: Mellenbergh GJ, Adler HJ, editors. Advising on Research Methods: Selected Topics 2013: Johannes van Kessel Publishing; 2014.
33. Epskamp S, Cramer AOJ, Waldorp LJ, Schmittmann VD, Borsboom D. qgraph: network visualizations of relationships in psychometric data. J Stat Softw 2012;48(4). [doi: 10.18637/jss.v048.i04]
34. Kalkman S, van Delden J, Banerjee A, Tyl B, Mostert M, van Thiel G. Patients' and public views and attitudes towards the sharing of health data for research: a narrative review of the empirical evidence. J Med Ethics 2022 Jan;48(1):3-13. [doi: 10.1136/medethics-2019-105651] [Medline: 31719155]
35. Kim J, Im E, Kim H. From intention to action: the factors affecting health data sharing intention and action. Int J Med Inform 2023 Jul;175:105071. [doi: 10.1016/j.ijmedinf.2023.105071]
36. Chandrasekaran R, Katthula V, Moustakas E. Too old for technology? Use of wearable healthcare devices by older adults and their willingness to share health data with providers. Health Informatics J 2021;27(4):14604582211058073. [doi: 10.1177/14604582211058073] [Medline: 34802315]
37. Tolentino DA, Costa DK, Jiang Y. Determinants of American adults' use of digital health and willingness to share health data to providers, family, and social media: a cross-sectional study. Comput Inform Nurs 2023 Nov 1;41(11):892-902. [doi: 10.1097/CIN.0000000000001025] [Medline: 37310724]
38. Cherif E, Bezaz N, Mzoughi M. Do personal health concerns and trust in healthcare providers mitigate privacy concerns? Effects on patients' intention to share personal health data on electronic health records. Soc Sci Med 2021 Aug;283:114146. [doi: 10.1016/j.socscimed.2021.114146] [Medline: 34242891]
39. Bansal G, Zahedi FM, Gefen D. The impact of personal dispositions on information sensitivity, privacy concern and trust in disclosing health information online. Decis Support Syst 2010 May;49(2):138-150. [doi: 10.1016/j.dss.2010.01.010]
40. Budin-Ljøsne I, Teare HJA, Kaye J, et al. Dynamic consent: a potential solution to some of the challenges of modern biomedical research. BMC Med Ethics 2017 Jan 25;18(1):4. [doi: 10.1186/s12910-016-0162-9] [Medline: 28122615]

Abbreviations

DS-14: Type D Scale-14
GAD-7: Generalized Anxiety Disorder-7
LOT-R: Life Orientation Test—Revised
NA: negative affectivity
OR: odds ratio
PFQ: Psychological Flexibility Questionnaire
PHQ-9: Patient Health Questionnaire-9
SI: social inhibition

Edited by A Mavragani; submitted 12.07.24; peer-reviewed by J Odumegwu; revised version received 02.12.24; accepted 02.12.24; published 23.01.25.

Please cite as:
Eversdijk M, Douma ER, Habibovic M, Kop WJ
The Association of Psychological Factors With Willingness to Share Health-Related Data From Technological Devices: Cross-Sectional Questionnaire Study
JMIR Form Res 2025;9:e64244
URL: https://formative.jmir.org/2025/1/e64244
doi:10.2196/64244

XSL•FO
RenderX

© Marijn Eversdijk, Emma Rixt Douma, Mirela Habibovic, Willem Johan Kop. Originally published in JMIR Formative Research (https://formative.jmir.org), 23.1.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

AI Machine Learning–Based Diabetes Prediction in Older Adults in South Korea: Cross-Sectional Analysis

Hocheol Lee, PhD; Myung-Bae Park, PhD; Young-Joo Won, PhD

Department of Health Administration, College of Software and Digital Healthcare Convergence, Yonsei University, Changjogwan, Yonseidae-gil 1, Wonju, Republic of Korea

Corresponding Author:
Young-Joo Won, PhD
Department of Health Administration, College of Software and Digital Healthcare Convergence, Yonsei University, Changjogwan, Yonseidae-gil 1, Wonju, Republic of Korea

Abstract

Background: Diabetes is prevalent in older adults, and machine learning algorithms could help predict diabetes in this population.

Objective: This study determined diabetes risk factors among older adults aged ≥60 years using machine learning algorithms and selected an optimized prediction model.

Methods: This cross-sectional study was conducted on 3084 older adults aged ≥60 years in Seoul from January to November 2023. Data were collected using a mobile app (Gosufit) that measured depression, stress, anxiety, basal metabolic rate, oxygen saturation, heart rate, and average daily step count. Health coordinators recorded data on diabetes, hypertension, hyperlipidemia, chronic obstructive pulmonary disease, percent body fat, and percent muscle. The presence of diabetes was the target variable, with various health indicators as predictors. Machine learning algorithms, including random forest, gradient boosting model, light gradient boosting model, extreme gradient boosting model, and k-nearest neighbors, were employed for analysis. The dataset was split into 70% training and 30% testing sets. Model performance was evaluated using accuracy, precision, recall, F1 score, and area under the curve (AUC). Shapley additive explanations (SHAPs) were used for model interpretability.

Results: Significant predictors of diabetes included hypertension (χ^2_1=197.294; P<.001), hyperlipidemia (χ^2_1=47.671; P<.001), age (mean: diabetes group 72.66 years vs nondiabetes group 71.81 years), stress (mean: diabetes group 42.68 vs nondiabetes group 41.47; t_{3082}=−2.858; P=.004), and heart rate (mean: diabetes group 75.05 beats/min vs nondiabetes group 73.14 beats/min; t_{3082}=−7.948; P<.001). The extreme gradient boosting model (XGBM) demonstrated the best performance, with an accuracy of 84.88%, precision of 77.92%, recall of 66.91%, F1 score of 72.00, and AUC of 0.7957. The SHAP analysis of the top-performing XGBM revealed key predictors for diabetes: hypertension, age, percent body fat, heart rate, hyperlipidemia, basal metabolic rate, stress, and oxygen saturation. Hypertension strongly increased diabetes risk, while advanced age and elevated stress levels also showed significant associations. Hyperlipidemia and higher heart rates further heightened diabetes probability. These results highlight the importance and directional impact of specific features in predicting diabetes, providing valuable insights for risk stratification and targeted interventions.

Conclusions: This study focused on modifiable risk factors, providing crucial data for establishing a system for the automated collection of health information and lifelog data from older adults using digital devices at service facilities.

(JMIR Form Res 2025;9:e57874) doi:10.2196/57874

KEYWORDS

diabetes; prediction model; super-aging population; extreme gradient boosting model; geriatrics; older adults; aging; artificial intelligence; machine learning

Introduction

With advancements in medical technology contributing to longer life expectancies, the world is witnessing a rapid acceleration in population aging. The United Nations projects that the global older adult population will increase from 10% in 2022 to 16% by 2025 [1]. This demographic shift significantly drives the increased prevalence of noncommunicable diseases such as diabetes, hypertension, hyperlipidemia, renal failure, arthritis, and Alzheimer disease, which burden the primary health care system [2]. Specifically, the aging population presents challenges in geriatric care within primary health care systems, including shortages of caregiving personnel, financial constraints, and psychological stresses associated with family caregiving.

Among the Organisation for Economic Co-operation and Development (OECD) countries, South Korea is expected to become the first super-aged society—a society where the older

XSL•FO
RenderX

adult population accounts for more than 20% of the total population [3]. The rising older adult population and concurrent increase in chronic diseases represent a critical public health issue in Korea. Notably, diabetes is a serious issue, with 39.2% of older adults experiencing diabetes [4]. Diabetes management is crucial, as inadequate control can lead to severe complications, including hypertension and hyperlipidemia [5]. However, the diabetes management rate among the Korean older adult population stands at a mere 30.3%, significantly lower than the awareness rates (84%) and treatment rates (74.8%) [6]. Older adults are particularly vulnerable to diabetes due to aging and consequent physiological changes, as well as lifestyle modifications resulting from physical decline and other medical conditions [7].

The early prevention of diabetes is imperative in older adults. Previous studies have shown that early intervention in older adults can reduce the risk of complications such as cardiovascular diseases, renal failure, and vision impairment [8]. Yet, clinical interventions such as fast blood glucose tests are often needed for diabetes management, and these may not be feasible due to older adults' physical frailty. Thus, health education and behavioral interventions, not clinical treatments and diagnostics, are vital for preventing diabetes and other chronic conditions in this population [9]. Research indicates that the factors affecting diabetes in older adults differ from those in the general population [10]. While obesity is a risk factor in individuals in their 40s, being underweight elevates the risk for various chronic conditions, including obesity in older adults [11]. The differences in diabetes risk factors for older adults highlight the importance of specific research on the older adult population.

With the global advancement of computer technology, machine learning and deep learning have been used across various fields. Machine learning algorithms can analyze complex and large datasets and identify patterns and risk factors that might not be apparent through traditional statistical methods. In medicine, the potential of machine learning has been demonstrated through an increase in research on disease prediction, personalized medicine, and personalized public health services in clinical and public health care. Particularly, machine learning algorithms can incorporate a broader array of factors and use a wide range of data types to produce generalizable results compared to conventional statistical approaches [12].

However, research using machine learning algorithms to predict diabetes in the Korean older adult population remains sparse. Studies targeting older adults are particularly lacking, partly due to challenges in accessing physical and mental health data, as well as daily lifelog data, due to the mobility constraints of the older population. Thus, a system for the automated collection of health information and lifelog data from older adults must be established using digital devices at service facilities they frequent [10]. This study determines diabetes risk factors among older adults aged ≥60 years by using machine learning algorithms and selects an optimized model. We hypothesize that older adults may be affected by different and identical risk factors compared to younger generations and provide evidence to facilitate policy-making. The findings could serve as a model for other countries with similar demographic changes and health care challenges.

Methods

Data Collection

This cross-sectional study was conducted on older adults aged ≥60 years, and the survey was conducted from January to November 2023. Participants aged ≥60 years living in Seoul, South Korea, were recruited among individuals who had participated in the Mind Care Provider Project. Participants in the project voluntarily collected data through public institution promotions aimed at users interested in using health measurement services. To collect data from the participants, a mobile app called Gosufit was developed, which was installed on participants' digital devices [13].

First, the app measured indicators such as depression, stress, anxiety, basal metabolic rate (BMR), oxygen saturation, heart rate, and average daily step count, and the data were stored on a server. Second, a health coordinator (registered nurse) measured and recorded data on diabetes, hypertension, hyperlipidemia, chronic obstructive pulmonary disease (COPD), percent body fat, and percent muscle. In total, 3674 older adults aged ≥60 years participated in the survey from January 1 to November 30, 2023. After excluding 590 participants due to nonresponse, dropout, or missing data, 3084 participants were included in the final analysis. Nonresponse and dropout occurred when participants did not wish to continue or when data transmission was interrupted due to issues with the app during the survey.

Instruments

The target variable was the presence of diabetes. Those diagnosed with diabetes by a physician were coded as 1, and those who have not been diagnosed were coded as 0. The predictor variables included hypertension, posttraumatic stress disorder (PTSD), stress, anxiety, depression, BMR, oxygen saturation, average daily step count, hyperlipidemia, COPD, percent body fat, and percent muscle. Hypertension, hyperlipidemia, and COPD were coded as 1 for a physician diagnosis or 0 otherwise. PTSD, stress, and anxiety and depression were assessed using the 5-item PTSD checklist (PCL-5), the short version of the Geriatric Depression Scale (SGDS), and the Hospital Anxiety and Depression Scale (HADS), respectively, with 100-point–based scoring. Oxygen saturation was determined using a digital oximeter per 100% saturation. The average daily step count was measured via the app.

Statistical Analysis

The Boruta-based feature selection method (FSM) was used for feature selection, which is a wrapper-based FSM that uses the random forest classification algorithm [14]. The entire dataset was randomly divided into a 70% training set and a 30% testing set using a stratified sampling procedure [15].

Machine Learning Algorithm

The random forest algorithm is an ensemble learning method used for classification and regression. It constructed multiple

XSL•FO
RenderX

decision trees during training and output the class or mean prediction (for regression) of individual trees. Random forests corrected overfitting to the training set that is often seen in decision trees. The gradient boosting model (GBM) is a machine learning technique used for regression and classification problems. It generated a predictive model as an ensemble of weak prediction models (typically decision trees). Like other boosting methods, it built the model in a stage-wise fashion and optimized any differentiable loss function to generalize the model. The light gradient boosting model (LGBM) is a gradient-boosting framework that uses a tree-based learning algorithm. It was designed for distributed and efficient operation with a faster training speed, higher efficiency, lower memory use, and better accuracy. The LGBM was capable of processing large-scale data with numerous features and data points while maintaining performance.

The extreme gradient boosting model (XGBM) is an algorithm for tree-based ensemble learning that addresses the slow performance speed and overfitting regularization issues of gradient boosting models. It featured built-in cross-validation and the automatic handling of missing values. The k-nearest neighbors (KNN) model is a simple, versatile, and easy-to-implement supervised machine learning algorithm used for classification and regression. It classified data points based on how their neighbors were classified, stored all available cases, and classified new cases based on a majority vote of its k neighbors. The case assigned to the class was the most common class among the k-nearest neighbors.

Performance Evaluation Criteria

The performance of the machine learning models was evaluated based on accuracy, precision, recall, F1 score, and area under the curve (AUC). The confusion matrix evaluated classification models by dividing them into positive or negative categories based on the match between actual and predicted classes.

Accuracy represented the proportion of correctly classified data among the total predictions for the hypertension risk group. It measured how accurate the predictions for the hypertension risk group were. The mathematical equation for accuracy is as follows:

$$\text{Accuracy}\%=tp+tntp+tn+fp+fn\times100$$

Precision represented the proportion of actual positive samples among the cases that were predicted as positive by the machine learning model. In other words, it indicated the ratio of samples positive for hypertension to those predicted to have hypertension by the model. The mathematical equation for precision is as follows:

$$\text{Precision}\%=tptp+fp\times100$$

Recall measured the proportion of those predicted by the machine learning model to have hypertension within the actual hypertension group. It provided the percentage of cases predicted to be at risk for hypertension from the entire hypertension risk group. The mathematical equation is as follows:

$$\text{Recall}\%=tptp+fn\times100$$

The F1 score was the harmonic mean of precision and recall. The mathematical equation is as follows:

$$\text{F1score}\%=2tp2tp+fp+fn\times100$$

The AUC referred to the area under the receiver operating characteristic curve, which was used to evaluate the performance of binary classification models, and an AUC close to 1 indicated better model performance. The AUC equation is as follows:

$$\text{AUC}=\int X=01TPRFPR-1xdx$$

Model Interpretability

Shapley additive explanations (SHAPs) interpret the prediction outcomes of machine learning models. SHAPs were introduced by Lundberg and Lee [16] in 2017 and were designed based on a game theory concept known as Shapley values. These values supported prediction interpretation, enabling the assessment of the relative importance among features. Additionally, they helped understand the characteristics with the greatest influence on the model's predictions, assisting in model improvement or decision-making processes. The mathematical equation is as follows:

$$\varnothing k(v)=1M!\sum S\subseteq M/\{k\}|S|!(M-|S|-1)![v(S\cup\{k\})-v(S)]$$

Ethical Considerations

Ethical approval for this study was obtained from Yonsei University Mirae Institutional Review Board (No. 1041849 - 202401-SB-021-01), including a supplementary application for expanded data collection. All procedures and data management were conducted following the General Data Protection Regulation and ethical principles outlined in the Helsinki Declaration. Informed consent was obtained from all study participants regarding data collection and the analysis of the data. The questionnaires were submitted entirely anonymously. No form of compensation was provided to the participants.

Results

Participant Information

In total, 3084 individuals participated in this study (Table 1). The study population comprised of 895 (29%) individuals with diabetes and 2189 (71%) individuals without. In addition, 1730 (56.1%) had hypertension, and diabetes prevalence significantly differed according to hypertension (χ^2_1=197.294; P<.001). In total, 1803 (58.5%) participants had hyperlipidemia, and diabetes prevalence significantly differed according to hyperlipidemia (χ^2_1=47.671; P<.001). The mean age was 72.66 years in the diabetes group and 71.81 years in the nondiabetes group. The mean stress score was 42.68 in the diabetes group and 41.47 in the nondiabetes group and significantly differed between the two groups (t_{3082}=−2.858; P=.004). The mean heart rate was 75.05 beats/min in the diabetes group and 73.14 beats/min in the nondiabetes group and significantly differed between the two groups (t_{3082}=−7.948; P<.001).

XSL•FO
RenderX

Table . Respondent characteristics.

Risk factor	Diabetes (n=895)	Nondiabetes (n=2189)	Chi-square or *t* test (*df*)	*P* value
Hypertension, n (%)			197.294[a] (1)	<.001
No	209 (23.4)	1115 (50.9)		
Yes	686 (76.6)	1074 (49.1)		
Hyperlipidemia, n (%)			47.671[a] (1)	<.001
No	286 (32)	995 (45.5)		
Yes	609 (68)	1194 (54.5)		
COPD[c], n (%)			7.764[b] (3082)	.005
No	856 (95.6)	2135 (97.5)		
Yes	39 (4.4)	54 (2.5)		
Sex, n (%)			3.472[b] (3082)	.06
Male	60 (6.7)	191 (8.7)		
Female	835 (93.3)	1998 (91.3)		
Age (years), mean (SD)	72.66 (6.31)	71.81 (6.32)	−3.395[b] (3082)	.001
PTSD[d], mean (SD)	14.45 (10.02)	14.21 (9.87)	−0.614[b] (3082)	.5
Stress, mean (SD)	42.68 (9.83)	41.47 (10.97)	−2.858[b] (3082)	.004
Anxiety, mean (SD)	15.58 (12.09)	15.67 (12.38)	0.179[b] (3082)	.86
Depression, mean (SD)	19.39 (12.48)	19.42 (12.92)	0.063[b] (3082)	.95
Percent body fat, mean (SD)	31.00 (8.74)	30.75 (8.66)	–0.727[b] (3082)	.47
Percent muscle, mean (SD)	35.73 (5.30)	35.95 (5.35)	–0.465[b] (3082)	.31
BMR[e], mean (SD)	1192.54 (159.05)	1187.71 (272.54)	–0.496[b] (3082)	.62
Oxygen saturation (%), mean (SD)	97.12 (3.48)	97.28 (2.44)	1.455[b] (3082)	.15
Heart rate (beats/min), mean (SD)	75.05 (6.29)	73.14 (5.97)	–7.948[b] (3082)	<.001
Daily step count, mean (SD)	11789.99 (18061.89)	13710.84 (34980.68)	1.560[b] (3082)	.12

[a]Chi-square.

[b]1-tailed *t* test.

[c]COPD: chronic obstructive pulmonary disease.

[d]PTSD: posttraumatic stress disorder.

[e]BMR: basal metabolic rate.

Risk Factor Selection Using Boruta

The importance of features was measured using the Boruta-based FSM (Multimedia Appendix 1). Hypertension, age, percent body fat, heart rate, hyperlipidemia, BMR, stress, and oxygen saturation were identified as important features. These features were included in the machine learning model for diabetes prediction.

Performance Comparison of Machine Learning Models

The performances of the 5 machine learning models used in the study were compared based on accuracy, precision, recall, F1 score, and AUC. Model performance was the highest for the XGBM, followed by the LGBM, random forest model, GBM, and KNN model (Table 2, Figure 1). The XGBM had an accuracy of 84.88%, precision of 77.92%, recall of 66.91%, F1 score of 72.00, and AUC of 0.7957, showing a high performance for its prediction.

XSL•FO
RenderX

Table . Performance of ml methods.

Ranking	Models	Accuracy (%)	Precision (%)	Recall (%)	F1 score	AUC[a]
1	Extreme gradient boosting model	84.88	77.92	66.91	72.00	0.7957
2	Light gradient boosting model	84.77	78.57	65.42	71.39	0.7906
3	Random forest model	81.53	78.82	49.81	61.04	0.7216
4	Gradient boosting model	77.32	77.57	30.85	44.14	0.6360
5	K-nearest neighbors model	66.95	42.06	36.43	39.04	0.5794

[a]AUC: area under the curve.

Figure 1. Receiver operating characteristic curves and precision-recall curves of 4 predictive models. AUC: area under the curve; LBGM: light gradient boosting model; ROC: receiver operating characteristic; XBGM: extreme gradient boosting model.

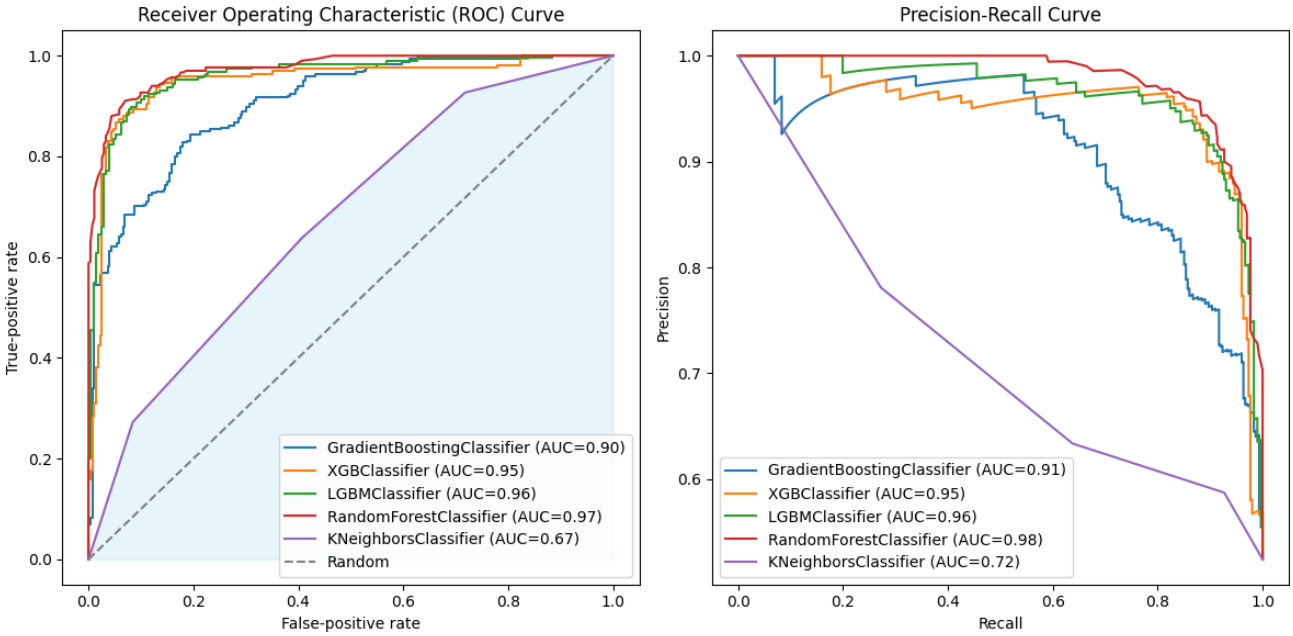

Interpretable Risk Factors of Hypertension

SHAP analysis was performed for the XGBM—the best-performing diabetes prediction model. The features selected through the FSM (hypertension, age, percent body fat, heart rate, hyperlipidemia, BMR, stress, and oxygen saturation) were included in the SHAP analysis. Red SHAP values indicated a higher impact on diabetes prediction, and blue SHAP values suggested a greater influence on nondiabetes outcomes (Figure 2). The findings indicated positive SHAP values for hypertension, indicating a stronger prediction of having diabetes, while not having hypertension was strongly linked to not having diabetes. Age presented mixed SHAP values, but notably, higher ages correlated with increased diabetes prediction. Similarly, higher heart rates and the presence of hyperlipidemia were associated with increased diabetes risk. Additionally, elevated stress levels were linked to a higher probability of diabetes. The SHAP analysis revealed a mixed impact of age on diabetes prediction. Particularly, as age increased, the likelihood of a diabetes prediction also increased. Additionally, a higher heart rate and the presence of hyperlipidemia were also linked to a higher diabetes probability. Moreover, an increase in stress levels elevated the risk of developing diabetes.

XSL•FO
RenderX

Figure 2. Importance of risk factors based on Shapley additive explanation values. SHAP: Shapley additive explanation.

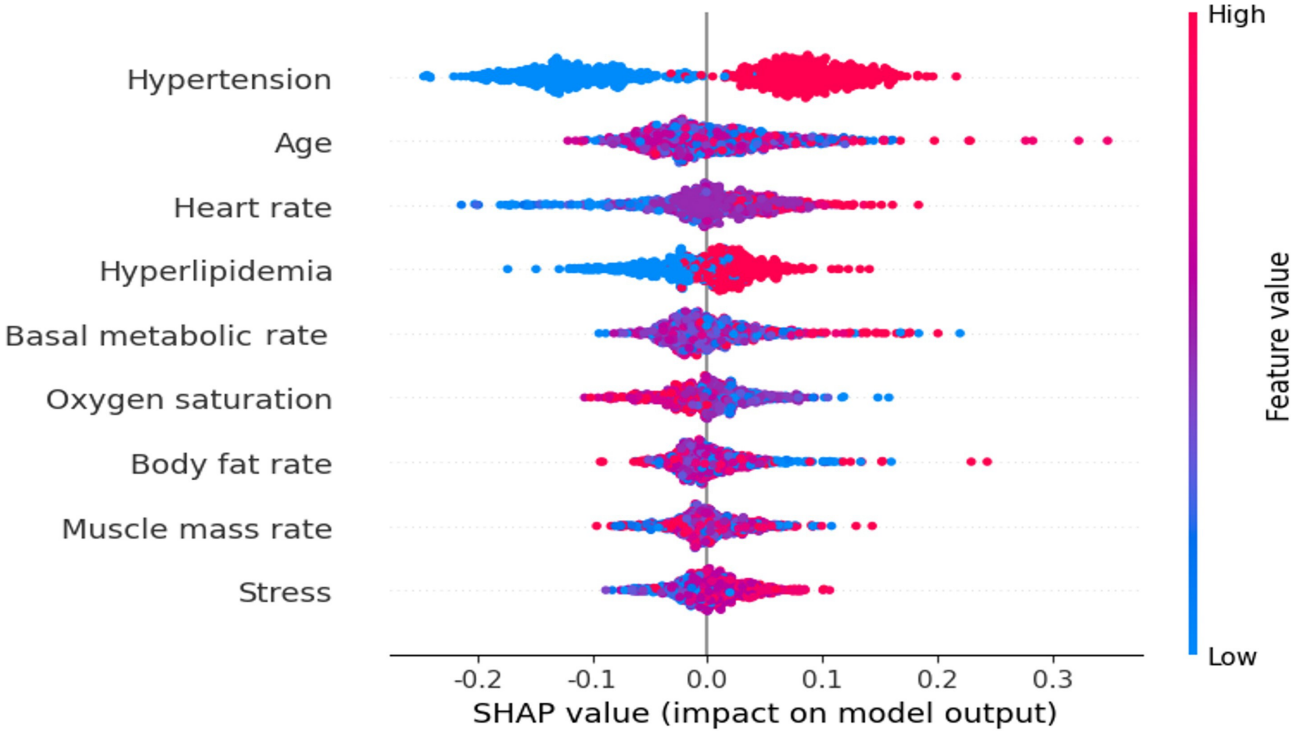

Discussion

Principal Results

This study analyzed the performance of 5 machine learning–based algorithms in predicting diabetes risk in older Korean adults. The XGBM performed the best, supporting previous research that shows XGBM as the superior model for diabetes prediction with an AUC of 84% [17], which was also similar to this study. In contrast, other studies have assessed the predictive accuracy of decision trees, naive Bayes algorithms, and random forest algorithms for diabetes risk factors, finding the random forest algorithm to be the top performer with 94% accuracy and precision [18], which was identical to our predictive algorithm. Systematic reviews of machine learning approaches also mention that support vector machines, artificial neural networks, and decision trees are frequently used prediction classification models [19]. However, these were excluded due to the anticipated challenges in managing nonlinear patterns with linear approaches [20].

In this study, hypertension, hyperlipidemia, old age, heart rate, and stress were identified as diabetes risk factors, which was similar to previous reports. Existing studies establish a close link between hypertension and diabetes, with a stronger correlation observed in older adults [21], and this was also seen in our results. A study analyzing diabetes risk factors using large-scale health care data from Kuwait demonstrated that the logistic regression had the highest accuracy at 80.7% and that hypertension, obesity, and sex were also strongly associated with diabetes risk [22]. Further, an increase in heart rate was predicted to elevate diabetes risk. A study of 30,000 participants reported that diabetes risk increased with every 10 beats per minute rise in pulse rate [23], which was consistent with the findings of this study. Hyperlipidemia was also identified as a diabetes risk factor. Hyperlipidemia, hypertension, and diabetes are 3 major chronic conditions that are interrelated as mutual risk factors, and this study confirmed that hyperlipidemia is a diabetes risk factor [24].

This study predicted diabetes in older adults, and certain factors differed from previously known diabetes risk factors. In the existing literature, diabetes risk is elevated with increasing obesity [25], but in this study, the SHAP value for percent body fat was not clear, with no significant differences in diabetes risk according to percent body fat in the independent *t* test. Thus, the severity of obesity was not a significant risk factor for diabetes. On the contrary, diabetes risk increased with decreasing body weight. An array of study results has been reported regarding this issue. One such report suggested that weight loss positively impacts diabetes in older adults but also complicates the treatment for optimal blood glucose regulation in patients with type 2 diabetes [26]. Here, weight loss in older adults entails loss of muscles and bone density, which may have a detrimental impact on diabetes in the long term [27].

Certain predictor variables identified as significant diabetes risk factors in previous studies were not included in this study. Particularly, family history is a known risk factor for diabetes [28]. One important factor of this study was that it focused on modifiable risk factors, rather than family history, which is an unmodifiable risk factor. Particularly, we examined other risk factors besides unmodifiable risk factors, such as family history, age, and sex, as being able to provide interventions for these risk factors is essential. The Korean government should emphasize preventive health education in primary health care facilities, senior welfare centers, community centers, and public health centers to address modifiable diabetes risk factors in the older adult population, such as hypertension, heart rate, hyperlipidemia, and stress. Additionally, while education for older adults is crucial, prevention is more important. Thus,

XSL•FO
RenderX

health education should be provided for those in their 40s to prepare for healthy older adulthood.

Limitations

This study had a few limitations. First, the study population comprised older adults aged ≥60 years who reside in Seoul, Korea. Thus, it did not represent the entire 60-and-over population in the country. Further, while existing studies set the age criterion for older adults as 65 years and older, we set the age to 60 years and older because the retirement age in Korea is 60 years. Second, we could not obtain personally identifiable information such as personal income and education level due to legal regulations. Thus, even though these factors may predict diabetes, this study was limited to other characteristics. Lastly, this study randomly split the dataset into training and testing subsets (70% and 30%). This split was not stratified, which means the distribution of target variables in the subsets may not perfectly match the original dataset's distribution. This could potentially impact the generalizability of the model's performance to other datasets. Future studies could consider employing a stratified splitting method to ensure a balanced representation of target variables across subsets.

Conclusions

This study analyzed machine learning algorithms for diabetes prediction in older adults in Korea. Hypertension, hyperlipidemia, and stress were identified as modifiable diabetes risk factors. Additionally, body fat percentage did not significantly predict diabetes in older adults, presumably because body fat loss is closely linked to muscle strength and bone mass loss. The findings suggest that targeted interventions focusing on managing hypertension, hyperlipidemia, and stress can significantly reduce diabetes risk in this population.

Furthermore, these diabetes predictors in older adults could be mitigated by promoting healthier lifestyle choices and behaviors, such as regular physical activity, balanced nutrition, and stress management techniques. The government should implement comprehensive health education programs across various facilities, including primary health care facilities and welfare centers, to raise awareness about these modifiable risk factors.

Moreover, educational interventions should be initiated at a younger age—particularly for individuals in their 40s—to foster proactive health management and prevent the onset of diabetes in older adulthood. By adopting a preventive approach and addressing modifiable risk factors early, we can enhance the overall health and quality of life for the aging population in Korea. It is necessary to develop personalized modeling that predicts major chronic diseases such as diabetes, hypertension, hyperlipidemia, and obesity through the advancement of prediction algorithms. This will provide a foundation for creating personalized health promotion education and programs. Future research should focus on refining machine learning–based models by incorporating diverse datasets and longitudinal data to improve generalizability and predictive performance. Additionally, exploring the integration of behavioral and environmental factors into machine learning algorithms may further enhance the accuracy and applicability of these models in real-world settings.

Acknowledgments

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea funded by the Ministry of Education (NRF-2021R1C1C2005464).

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Risk factors selection using the Boruta-based feature selection method.
[JPEG File, 127 KB - formative_v9i1e57874_app1.jpeg]

References

1. World population prospects 2022: summary of results. United Nations. 2022. URL: https://www.un.org/development/desa/pd/sites/www.un.org.development.desa.pd/files/wpp2022_summary_of_results.pdf [accessed 2024-12-31]
2. Chen C, Ding S, Wang J. Digital health for aging populations. N Med 2023 Jul;29(7):1623-1630. [doi: 10.1038/s41591-023-02391-8] [Medline: 37464029]
3. Kim KW, Kim OS. Super aging in South Korea unstoppable but mitigatable: a sub-national scale population projection for best policy planning. Spat Demogr 2020 Jul;8(2):155-173. [doi: 10.1007/s40980-020-00061-8] [Medline: 34222615]
4. Kwon HS. Prevalence and treatment status of diabetes mellitus in Korea. J Korean Med Assoc 2023 Jul;66(7):404-407. [doi: 10.5124/jkma.2023.66.7.404]
5. de Boer IH, Bangalore S, Benetos A, et al. Diabetes and hypertension: a position statement by the American Diabetes Association. Diabetes Care 2017 Sep;40(9):1273-1284. [doi: 10.2337/dci17-0026] [Medline: 28830958]

XSL•FO
RenderX

6. Ha KH, Lee KA, Han KD, Moon MK, Kim DJ. Diabetes screening in South Korea: a new estimate of the number needed to screen to detect diabetes. Korean J Intern Med 2023 Jan;38(1):93-100. [doi: 10.3904/kjim.2022.283] [Medline: 36420563]
7. Flack KD, Davy KP, Hulver MW, Winett RA, Frisard MI, Davy BM. Aging, resistance training, and diabetes prevention. J Aging Res 2010 Dec 15;2011:127315. [doi: 10.4061/2011/127315] [Medline: 21197110]
8. Rooney MR, Rawlings AM, Pankow JS, et al. Risk of progression to diabetes among older adults with prediabetes. JAMA Intern Med 2021 Apr 1;181(4):511-519. [doi: 10.1001/jamainternmed.2020.8774] [Medline: 33555311]
9. Kramer MK, Vanderwood KK, Arena VC, et al. Evaluation of a diabetes prevention program lifestyle intervention in older adults: a randomized controlled study in three senior/community centers of varying socioeconomic status. Diabetes Educ 2018 Apr;44(2):118-129. [doi: 10.1177/0145721718759982] [Medline: 29514568]
10. Caprani N, Gurrin C, O'Connor N. I like to log: a questionnaire study towards accessible lifelogging for older users. Presented at: 12th International ACM SIGACCESS Conference on Computers and Accessibility; Oct 25-27, 2010; Orlando, FL p. 263-264. [doi: 10.1145/1878803.1878862]
11. Wilson PWF, Kannel WB. Obesity, diabetes, and risk of cardiovascular disease in the elderly. Am J Geriatr Cardiol 2002;11(2):119-123. [doi: 10.1111/j.1076-7460.2002.00998.x] [Medline: 11872970]
12. Henglin M, Stein G, Hushcha PV, Snoek J, Wiltschko AB, Cheng S. Machine learning approaches in cardiovascular imaging. Circ Cardiovasc Imaging 2017 Oct;10(10):e005614. [doi: 10.1161/CIRCIMAGING.117.005614] [Medline: 28956772]
13. GoSuFit. Google Play. URL: https://play.google.com/store/apps/details?id=com.healthbridge.hfact20&pcampaignid=web_share&pli=1 [accessed 2024-12-31]
14. Pudjihartono N, Fadason T, Kempa-Liehr AW, O'Sullivan JM. A review of feature selection methods for machine learning-based disease risk prediction. Front Bioinform 2022;2:927312. [doi: 10.3389/fbinf.2022.927312] [Medline: 36304293]
15. May RJ, Maier HR, Dandy GC. Data splitting for artificial neural networks using SOM-based stratified sampling. Neural Netw 2010 Mar;23(2):283-294. [doi: 10.1016/j.neunet.2009.11.009] [Medline: 19959327]
16. Lundberg SM, Su-In L. A unified approach to interpreting model predictions. Presented at: 31st International Conference on Neural Information Processing Systems; Dec 4-9, 2017; Long Beach, CA p. 4768-4777 URL: https://dl.acm.org/doi/10.5555/3295222.3295230 [accessed 2024-12-31]
17. Lai H, Huang H, Keshavjee K, Guergachi A, Gao X. Predictive models for diabetes mellitus using machine learning techniques. BMC Endocr Disord 2019 Oct 15;19(1):101. [doi: 10.1186/s12902-019-0436-6] [Medline: 31615566]
18. Yuvaraj N, SriPreethaa KR. Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster. Cluster Comput 2019 Jan;22(S1):1-9. [doi: 10.1007/s10586-017-1532-x]
19. Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine learning and data mining methods in diabetes research. Comput Struct Biotechnol J 2017;15:104-116. [doi: 10.1016/j.csbj.2016.12.005] [Medline: 28138367]
20. Tang Y. Deep learning using linear support vector machines. arXiv. Preprint posted online on Feb 21, 2015. [doi: arXiv:1306.0239v4]
21. Ye C, Fu T, Hao S, et al. Prediction of incident hypertension within the next year: prospective study using statewide electronic health records and machine learning. J Med Internet Res 2018 Jan 30;20(1):e22. [doi: 10.2196/jmir.9268] [Medline: 29382633]
22. Farran B, Channanath AM, Behbehani K, Thanaraj TA. Predictive models to assess risk of type 2 diabetes, hypertension and comorbidity: machine-learning algorithms and validation using national health data from Kuwait--a cohort study. BMJ Open 2013 May 14;3(5):e002457. [doi: 10.1136/bmjopen-2012-002457] [Medline: 23676796]
23. Lee DH, de Rezende LFM, Hu FB, Jeon JY, Giovannucci EL. Resting heart rate and risk of type 2 diabetes: a prospective cohort study and meta-analysis. Diabetes Metab Res Rev 2019 Feb;35(2):e3095. [doi: 10.1002/dmrr.3095] [Medline: 30378246]
24. Sullivan PW, Ghushchyan VH, Ben-Joseph R. The impact of obesity on diabetes, hyperlipidemia and hypertension in the United States. Qual Life Res 2008 Oct;17(8):1063-1071. [doi: 10.1007/s11136-008-9385-7] [Medline: 18777200]
25. Du J, Yang S, Zeng Y, Ye C, Chang X, Wu S. Visualization obesity risk prediction system based on machine learning. Sci Rep 2024 Sep 28;14(1):22424. [doi: 10.1038/s41598-024-73826-6] [Medline: 39342032]
26. Kyrou I, Tsigos C. Obesity in the elderly diabetic patient: is weight loss beneficial? No. Diabetes Care 2009 Nov;32(Suppl 2):S403-S409. [doi: 10.2337/dc09-S348] [Medline: 19875589]
27. Newman AB, Lee JS, Visser M, et al. Weight change and the conservation of lean mass in old age: the Health, Aging and Body Composition Study. Am J Clin Nutr 2005 Oct;82(4):872-878. [doi: 10.1093/ajcn/82.4.872] [Medline: 16210719]
28. Harrison TA, Hindorff LA, Kim H, et al. Family history of diabetes as a potential public health tool. Am J Prev Med 2003 Feb;24(2):152-159. [doi: 10.1016/s0749-3797(02)00588-3] [Medline: 12568821]

Abbreviations

AUC: area under the curve
BMR: basal metabolic rate
COPD: chronic obstructive pulmonary disease

XSL•FO
RenderX

FSM: feature selection method
GBM: gradient boosting model
HADS: Hospital Anxiety and Depression Scale
KNN: k-nearest neighbors
LGBM: light gradient boosting model
OECD: Organisation for Economic Co-operation and Development
PCL-5: 5-item PTSD checklist
SGDS: short version of the Geriatric Depression Scale
SHAP: Shapley additive explanation
XGBM: extreme gradient boosting model

Edited by A Mavragani; submitted 28.02.24; peer-reviewed by BM Kunar, HB Drissi, J Nalepa; revised version received 09.12.24; accepted 09.12.24; published 21.01.25.

Please cite as:
Lee H, Park MB, Won YJ
AI Machine Learning–Based Diabetes Prediction in Older Adults in South Korea: Cross-Sectional Analysis
JMIR Form Res 2025;9:e57874
URL: https://formative.jmir.org/2025/1/e57874
doi:10.2196/57874

©Hocheol Lee, Myung-Bae Park, Young-Joo Won. Originally published in JMIR Formative Research (https://formative.jmir.org), 21.1.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Insights From Diverse Perspectives on Social Media Messages to Inform Young Adults With Cancer About Clinical Trials: Focus Group Study

Melissa P Beauchemin[1], PhD, RN, CPNP, CPON; Desiree Walker[2], BA; Allison Rosen[3], MS; Maria Frazer[4], BSE; Meital Eisenberger[5], BS; Rhea K Khurana[1], BSc; Edward Bentlyewski[5], MSN, APN, NP-C, AOCNP, CCRP; Victoria Fedorko[6], BA; Corey H Basch[7], EdD, MPH; Grace C Hillyer[6], EdD, MPH

[1]Columbia University School of Nursing, New York, NY, United States
[2]Afiya, New York, NY, United States
[3]Advocates for Collaborative Education, Santa Cruz, CA, United States
[4]Pediatrics, Hematology/Oncology, University of Iowa Health Care, Iowa City, IA, United States
[5]Cancer Clinical Protocol and Data Management, Columbia University Herbert Irving Comprehensive Cancer Center, New York, NY, United States
[6]Department of Epidemiology, Columbia University Mailman School of Public Health, New York, NY, United States
[7]Department of Public Health, William Paterson University, Wayne, NJ, United States

Corresponding Author:
Melissa P Beauchemin, PhD, RN, CPNP, CPON
Columbia University School of Nursing, New York, NY, United States

Abstract

Background: Low rates of adolescent and young adult (YA; aged 15-39 y) clinical trial enrollment (CTE), particularly among underserved groups, have resulted in a lack of standardized cancer treatments and follow-up guidelines for this group that may limit improvement in cancer treatments and survival outcomes for YAs.

Objective: To understand and address unique barriers to CTE, we conducted focus groups to learn about informational, financial, and psychosocial needs of YAs surrounding CTE and identify strategies to address these barriers.

Methods: We conducted 5 focus groups in 2023 among a diverse sample of YA patients from across the United States. An interview guide was developed collaboratively with YA advocates. Specifically, informational needs, financial concerns, and psychosocial issues were explored, and participants were probed to suggest strategies, especially those that leverage technology, to address these barriers. Sessions were audio recorded, transcribed, and coded using direct content analysis. Findings were synthesized through consensus discussions.

Results: We confirmed the previously proposed thematic barriers regarding YA CTE and identified 9 subthemes: awareness, lack of clear and accessible CTE information, fear of the unknown, assumptions about costs, insurance coverage, navigating financial responsibilities, clinical trial discussions, clinical trial misconceptions, and desire for a support network. Throughout, YAs mentioned needs that might be addressed through informational outreach leveraging digital technology, the internet, and social media.

Conclusions: This study expands knowledge of YA perceived barriers to CTE. These findings suggest that leveraging digital technology to disseminate reliable information to address needs may be an effective strategy to improve clinical trial participation in the YA population.

(JMIR Form Res 2025;9:e64265) doi:10.2196/64265

KEYWORDS

young adults; adolescent and young adult cancer; cancer; cancer treatment; clinical trials; clinical trial awareness; clinical trial enrollment; clinical trial knowledge; clinical trial attitudes; clinical trial enrollment barriers; social media; social media messages; psychosocial; United States

Introduction

Among adolescents and young adults (AYAs; aged 15 - 39 y) with cancer [1], clinical trial enrollment (CTE) rates range from 3% to 14%, a number disproportionately below the observed annual cancer incidence in this age group [2]. Previous research shows the disparity in CTE among AYAs is stark when compared to the cancer incidence rates across other age groups. For instance, pediatric and older adult patients demonstrate higher CTE rates compared to AYAs, although AYAs

XSL•FO
RenderX

experience higher cancer incidence [3-5]. Despite cancer incidence being 3 times greater and with lower survival rates across all major tumor types for AYAs [3-8], there remains insufficient progress in discovering and developing new therapies, investigating molecular differences, and CTE for this population [9,10]. Inferior survival outcomes are further compounded among AYAs from marginalized groups (eg, non-White, publicly insured, low-income, rural residents, and non-English speakers [9,11]).

Young adults (YAs; aged 18-39 y) with cancer are a distinct group with needs that differ greatly from pediatric and adult populations. A cancer diagnosis as a YA can derail developmental transitions and impact emotional, cognitive, and social development [12,13]. As digital natives, a term used to describe individuals who have grown up with digital technology as an integral part of their daily lives, YAs approach health information-seeking differently than their younger and older counterparts [14]. Their proficiency with digital technology makes them especially adept at accessing medical information online using a wider variety of digital resources than older adults [15,16]. Interventions to increase CTE have demonstrated feasibility and acceptability among YAs with cancer [17], however, interventions developed thus far incorporate little input from YA patients.

Abrahão et al [18] explored CTE barriers from the perspectives of physicians and AYA patients using individual interviews and found that poor understanding of clinical trials, financial issues, and psychosocial problems were the greatest obstacles to CTE for AYAs. As the first phase of a larger study (TALK TO ME: Tailored YA Trial Messaging, HF CU22-2020), we conducted focus groups (FGs) to capitalize on interpersonal communication among AYAs with cancer and inform the development of tailored, relevant, and salient social media messages to address barriers to CTE among YAs with cancer.

Methods

Study Design

To build on prior work understanding barriers and facilitators to CTE among YAs, we used a qualitative descriptive approach to conduct FGs among YAs with cancer, as this methodology is ideal for providing a rich description of an experience by participants rather than interpreting or theorizing [19,20]. As the goal of this phase of this study was to develop social media messages for YAs regarding CTE, this methodologic approach was ideal to stay close to the participant experiences and perceptions. We used FGs as the mode of data collection to tap into the communication, interactions, and processes that are more readily observed among groups than in individual interviews [21].

Study Population and Recruitment

FG participants were recruited by leveraging the networks and social media platforms (eg, Facebook support groups and Instagram) of this study team's YA advocates via social media posts or by direct outreach. Outreach messages explained that we were recruiting individuals who were diagnosed with cancer between 18 and 39 years of age to participate in a FG to explore YA perceptions about clinical trials with the goal of informing the development of social media messages that aim to improve understanding and address concerns about clinical trials for YAs with cancer. Interested respondents were asked to provide basic sociodemographic (current age, gender identity, race or ethnicity, and city or state of residence, rural vs urban), whether they were active social media users, clinical information, including age at diagnosis and cancer type and availability (weekday vs weekend and daytime vs evening) for group participation, via a Google Forms survey link. As we sought to understand reasons for both participation and nonparticipation in cancer clinical trials, participation in a clinical trial was not an eligibility criterion.

This study team's YA advocates contacted potential participants to explain this study's procedures, confirm eligibility (diagnosed with cancer between 18 and 39 years of age, active social media user, and English speaker), and invite participation. Potential participants were then purposively sampled to represent a range of experiences (1 group all female and any race or ethnicity; 1 group all male and any race or ethnicity; 3 groups mixed gender identity, race or ethnicity, and any cancer type).

FG Data Collection

An FG guide was developed during weekly study team meetings guided by the Abrahaō framework from individual interviews among AYA participants [18]. We applied an iterative process to the guide development; pilot testing was carried out by this study's team members, including patient advocate partners [22]. To ensure study participants were oriented to the topic, a brief, 2‑ to 3-minute overview was prepared by this study team to introduce and define a clinical trial [23].

FGs were conducted virtually over Zoom between May and October 2023 and were 1 to 1½ hours in duration. Moderators included female YA advocates and, for the all-male group, a male oncology nurse practitioner. An information sheet was distributed via email to all participants before the FG.

The sessions began with a brief explanation of clinical trials and important definitions. Participants shared their experiences with cancer clinical trials including experiences of family members or friends and were asked about reasons for participating or not, and the support needed for trial involvement specifically as related to information, finances, and their psychosocial needs [18]. They also discussed their trust in social media information, preferred platforms, where they often seek health information, and suggested ways to use it to share clinical trial information with YAs.

Data Analysis

All FG sessions were audio-recorded and transcribed verbatim using NVivo Transcription, Lumivero, LLC. Participants were assigned a speaker ID, and transcripts were deidentified and cleaned using a denaturalized approach [24] to facilitate analysis of the participants' words rather than a more interpretive approach. We used a directed content analysis [25] to enrich and expand our understanding of barriers and facilitators to refine existing themes rather than identify new ones. Study team members trained (MPB, ME, RKK, and GCH) in qualitative analysis coded the transcripts first individually; then consensus

XSL•FO
RenderX

discussions were held to discuss coding discrepancies, synthesis of findings, and to ensure intercoder agreement. Saturation was anticipated given the deductive nature of the analytic approach but was tracked during consensus discussions [26].

Ethical Considerations

Verbal consent, including consent to be audio recorded, was obtained at the start of each FG. Participants were informed of their right to withdraw from this study at any time without penalty. For their time and effort, participants were compensated with a gift card valued at US $50. Regulatory approval was obtained from Columbia University Institutional Review Board (ID# AAAU4337).

Results

Overview

A total of 33 individuals participated in the FGs (maximum 8, minimum 4). The mean age was 32.2 (SD 5.7) years. Males and females were equally represented; 42.4% (n=14) were White and 42.4% (n=14) were Black or African American. Most (n=24, 72.7%) resided or sought treatment in urban areas, and 27.3% (n=9) were diagnosed with breast cancer (Table 1). Findings are grouped by themes (informational, financial, and psychosocial) and subtheme. Exemplary quotes further supporting subthemes are displayed in Table 2.

Table . Characteristics of focus group participants (N=33).

Demographic	Value
Age (years)	
Mean (SD)	32.2 (5.7)
Range	24-42
Gender, n (%)	
Female	17 (51.5)
Male	16 (48.5)
Race, n (%)	
White	14 (42.4)
Black or African American	14 (42.4)
Asian	1 (3)
Mixed race	3 (9.1)
Did not specify	1 (3)
Rural versus urban, n (%)	
Rural	9 (27.3)
Urban	24 (72.7)
Cancer type, n (%)	
Skin cancer	2 (6.1)
Breast cancer	9 (27.3)
Lymphoma	2 (6.1)
Other	20 (60.6)

XSL•FO
RenderX

Table . Interview themes and categories with additional illustrative quotes.

Theme and categories		Representative quote
Theme 1: Information or knowledge		
	Awareness	"I was given two options as far as radiation and chemo. And basically, they were like, these are the two options you have, which one do you want? Clinical trials never even brought up." [FG1[a]]
	Lack of standardized or central system	"One of the biggest issues, I think upon diagnosis as a patient and as a as a caregiver is information overload." [FG1]
	Fear of the unknown	"Being reminded that...my health is more important than just a number or an experiment or a statistic. And then in terms of. Like in a form of social media that I would go to, I think like visual media tends to be more compelling to me like to be able to see your face and and a name and to know you know who that person is giving me information." [FG4]
Theme 2: Psychosocial		
	Clinical trial discussions	"You know, being upfront about the risks obviously is important, but the potential benefits too, I think just prevent presenting balanced information about the nature of the trial is probably all you can do, really." [FG5]
	Clinical trial misconceptions	"I feel like clinical trials have been sort of this last-ditch effort after thought" [FG1]
	Desire for support network	"Social media is the connection to community" [FG2]
Theme 3: Financial		
	Assumptions or knowledge about clinical trials	"And another one said, In order to find out if you're eligible, you need to pay $750 out of pocket to do this consultation. So that just makes people I mean, it just it's just not fair." [FG1]
	Young adult specific concerns	"The trial sponsor may be able to offer some form of compensation or some sort of resource, but there's also the sort of unintended financial impact that can happen, especially to this age group taking time off from work, finding childcare, meeting the needs for other bills, that kind of stuff." [FG1]
	Communication	"And I think that the visits they go by so quickly anyways, there's so much to cover. There probably wouldn't even be enough time for a provider to sit down and have a full conversation about what is a clinical trial." [FG5]

[a]FG: focus group.

Theme 1: Informational

Subtheme 1: Awareness

Across all FGs, YAs cited varying levels of exposure to information and lack of awareness of cancer generally, their specific cancer, and treatment options that may have included clinical trials. Many reported facing the challenges of gathering important information about their diagnosis and treatment options on their own and conveying their informational needs to others. One participant reflected saying, "when I was diagnosed, I was still on my parent's health insurance and my dad handled all of the insurance...I didn't know anything like, I feel like if I hadn't had that support, like I can't even imagine."

Participants also noted that they wanted to understand their disease and treatment options to better participate in their own care, however finding relatable and accessible information was difficult. One participant remarked "I thought I was really educated until I got diagnosed with cancer and then I just felt like an idiot." Some participants felt that they were tasked with processing information about their cancer diagnosis and

XSL•FO
RenderX

treatment that was not easily digestible or so complex that "no normal person is just going to know."

The process of deciding on treatment, whether the standard of care or a potential clinical trial, was described as "draining." In some cases, participants said that they were completely unaware that a clinical trial may have been an option. As one person recalled, "I was given two options as far as radiation and chemo. And basically, they were like, these are the two options you have, which one do you want? Clinical trials [were] never even brought up."

More specifically, YAs wanted to be better informed about their possible treatment options and how they might impact them in the future. They expressed the need to comprehend the risks and benefits associated with each treatment option to make an informed decision. One participant summarized this as "there's so many different things that you have to consider when making these decisions and side effects are no joke for cancer treatment…there's just it's stuff that's very lasting and those are other things you have to consider when making the decision." Most importantly, participants conveyed that they wanted their providers to consider them as a whole person when treating their cancer, specifically considering their reproductive health concerns and fertility preservation. One participant told us "I had like seven days to decide if I wanted to freeze my eggs or not." They believed that the provider should make these YA-unique issues a top priority when creating cancer treatment plans and ensure that YAs fully understand these implications at the start.

Subtheme 2: Lack of Standardized or Central System

Participants repeatedly spoke of the difficulty in finding centralized databases with information on their cancer or resources relevant to their experiences. When trying to understand the concept of a clinical trial, participants found the information to be complex and filled with medical terminology they did not understand only added to the stress of their illness. Many felt intimidated, overwhelmed, and frightened by the information they found, especially for YAs with rare cancers. Among participants who searched these databases, the consensus was there needs to be "some sort of database that like pulls that information [together]...clinical trial wise…it would be nice to know that there are options for people that have like, not so ideal cancer circumstances." Additionally, participants felt this information should be engaging, compassionate, and written in plain language.

When researching clinical trials, YAs said that they would like information to come from trusted sources such as their medical team or the organization implementing the clinical trial. It helps, they said when "someone is maybe [a] medical personnel…or a governmental organization[s] that you can then be able to trust that information." One participant who participated in a clinical trial explained that "the biggest differentiator for me was my oncologist explaining to me all of the options for clinical…and really went into detail…he explained to me in a way that I could understand...and I trusted him. I trusted him implicitly."

Participants addressed the role of caregivers seeking treatment information for their YAs. They emphasized the need for resources to comfort both YAs and their caregivers throughout treatment.

> *The caregiver is often the person doing this research or supporting this. And so, I think it's really important to consider them as your equal audience.*

They also warned that "One of the biggest issues, I think upon diagnosis as a patient and as a caregiver is information overload."

Subtheme 3: Fear of the Unknown

A common fear cited by participants regarding clinical trials was adverse side effects of the experimental treatment. They also wanted to connect with others like themselves to hear firsthand about their experience with cancer and the treatment options they were considering to put their minds at ease. They stated that when they could not access the support and information they needed through their care team, they sought it on social media using Instagram hashtags and Facebook groups.

> *I think the biggest thing for me, where I found the most helpful information was searching hashtags on Instagram. I searched my chemo regimen hashtags. I searched my diagnosis. I searched by stage. I searched AYA cancer...And I would find people there that would have a community on their pages and resources on their pages.*

YAs that used social media to supplement their understanding and find support expressed concern over the lack of reliable sources and the spread of misinformation on some platforms. Participants suggested that it is difficult to differentiate between helpful information and potentially harmful content on social media. Another issue related to the use of social media reported by participants was "doom scrolling," which can cause increased conflict within the patient. One participant explained that "because of the information overload, I've really had to limit my information intake from social media."

Theme 2: Psychosocial Issues

Subtheme 1: Clinical Trial Discussions

YAs cited that care teams should remember that patients usually have limited knowledge of clinical trials, including the risks and benefits. As one participant explained, "I actually refused [experimental] treatment because I thought my only option was chemotherapy and radiation...if I knew that I had other options, I probably would have been more proactive, no matter what the cost was like." Participants stated that initiating simple dialogue that normalizes CTE in the context of cancer treatment would motivate patients to consider clinical trials as a treatment option.

Subtheme 2: Clinical Trial Misconceptions

Participants explained that common misconceptions about clinical trials were that they were reserved for patients with advanced disease and considered a "last ditch effort" for only those with terminal disease. These misconceptions add to the fear surrounding clinical trials and can present as a key barrier to CTE and psychological distress. Participants explained that

XSL•FO
RenderX

providers should anticipate and address misconceptions about clinical trials during treatment discussions.

Subtheme 3: Desire for a Support Network

Participants reported that the effects of their illness and treatment took a significant toll on their mental health and that additional check-ins with the medical team would have been beneficial. For those who participated in a clinical trial, the additional contact and attention by the clinical trial care team eased the burden. One participant who participated in a clinical trial explained that "the additional support for me was one of the best things that came out of the clinical trial because I was at such a large hospital…This connection is especially important during trial enrollment as this is an uncertain time for patient[s]."

Participants explained that they would have appreciated a peer support network while navigating life and treatment simultaneously. As one participant explained, "cancer is the worst part-time job on the entire planet." They use social media as a "connection to community." Participants expressed interest in a social media account or online group that connects patients more easily since it can be valuable for communication with those who have had similar experiences.

Some participants mentioned that discomfort while considering a clinical trial can be further exacerbated by family members' lack of understanding or distrust of medical research. This may lead family members to caution patients away from participating. One family responded to the idea of a clinical trial their son was considering by saying "No, don't do it. You know, they're going to do x y z. It's not safe. You know, you're not that sick." The participant explained that their family saw a clinical trial as "a last-line option and not a first-line option." The importance of providing caregivers and others surrounding the patient with accurate information to support the YAs treatment decisions was stressed.

Theme 3: Financial Concerns

Subtheme 1: Assumptions About Clinical Trials Costs

Financial hardship was mentioned by many participants as a key deterrent to CTE. Several reported that the fear of financial problems prevented them from inquiring about clinical trials because they assumed resources would not be provided or that the resources provided would not be sufficient to cover additional costs. Patients applauded treatment centers that provide financial navigation and resources to patients considering clinical trials. They believed this could ease barriers such as travel expenses and limited insurance coverage and potentially increase CTE.

Subtheme 2: Insurance Coverage

Some participants indicated that the insurance framework is difficult to navigate on its own and can be further complicated by seeking nonstandard of care treatment. One participant explained that they "wanted to get enrolled in a clinical trial that was at another health institution" but they did not proceed because "my insurance wouldn't cover it because it was out of network." Alongside this, the struggle to balance their schooling, family, and work responsibilities impacts the state of their finances.

Subtheme 3: Navigating Financial Responsibilities

Some of the participants assumed that the patient was responsible for additional costs which deterred them from inquiring about clinical trials as a treatment option. Many participants were unaware that most health insurers cover many, if not all, costs associated with clinical trials, that costs related to increasingly more clinical trials are covered by the sponsor, or that there are occasions when sponsors offer financial assistance to address costs not covered by insurance or the clinical trial. One participant remarked, "I think a lot of us don't even realize that our hospital might have financial navigators or patient navigators who exist to help answer those questions." YAs believed providers and any online sources should educate patients about their financial options while addressing common misperceptions.

Discussion

In this study, we explored categorical individual-level barriers [18] to CTE among a diverse group of individuals diagnosed with cancer as YAs to extend our knowledge and inform future strategies to address these barriers. Through FGs representative of gender identity, race, and geographic region in the United States, we found that lack of knowledge and imperfect communication were the greatest barriers to CTE in this population. Participants indicated that limited awareness about clinical trials, commonly held misperceptions, misinformation on websites and social media, and a lack of clear and accessible information about clinical trial risks and benefits likely contribute to underrepresentation among YAs. Proactively providing anticipatory guidance may circumvent distress if the topic is raised and reduce patient hesitancy in participation.

Consistent with other literature, our study supports that YAs often seek information on the web or social media [27,28] and participants favored an intervention using social media to address CTE barriers. They also provided clear guidance on how information might be tailored to YAs and what specific information may be most helpful in addressing concerns and barriers to CTE.

In many ways, the financial concerns and psychosocial issues reflected a subset of the informational gaps reported by participants. Finances were a significant concern, and several participants were unaware of insurance coverage requirements or that resources such as financial navigators exist to help with the costs of CTE [29]. Having financial information readily available during the cancer treatment decision-making process can support discussions around a clinical trial option. A digital support network may also provide a connection with others sharing a similar lived experience during times of loneliness and fear.

We followed rigorous qualitative methodology to conduct our study, however, there are limitations to acknowledge. Potential selection bias may be present, as participants were recruited through our YA patient advocates on the internet and, thus, may have been more technologically proficient and familiar with social media and clinical trials than the general YA patients with cancer community. However, the vast majority of YAs

XSL•FO
RenderX

have an online presence and use social media and the internet regularly. Further, as with other qualitative research, the transferability (ie, generalizability) of these data may be limited although great effort was expended to ensure that our sample was diverse in terms of sociodemographic characteristics to yield FGs representative of the YA population. More than half of our FG participants were non-White which provides a deeper view of experiences from underrepresented YAs regarding engagement with and perceptions of CTE. Although we based our discussions a priori on themes identified by Abrahão et al [18], moderators followed the discussions and probed further when any topic arose organically to ensure all potential barriers were explored and thus, resulted in the expansion of subthemes described.

Our research enhances our understanding of the obstacles that YAs face when making decisions about cancer treatment, including the option to participate in a clinical trial. It also provides further evidence that using a combination of approaches at different levels may be most effective in engaging YAs in discussions about CTE. While barriers to CTE at the provider, institutional, and system levels have been extensively studied, the YAs in our study highlight individual-level challenges that need attention. Our findings indicate that a social media intervention focused on sharing reliable information about CTE could help YAs process complex information and make informed treatment decisions, thus improving CTE in this population. Engaging with YAs through social media allows researchers and medical professionals to connect with them in a way that is familiar and comfortable. Further research should address the gaps in information identified in our study to develop resources and educational materials that can enhance understanding of CTE among YAs, their families, and caregivers. When developing treatment plans, it is crucial to prioritize patients' informational needs and provide all YAs with the tools and resources to have meaningful discussions with their health care providers about CTE.

Acknowledgments

This study was funded by a grant to GCH from the Hope Foundation (HF CU22-2020).

Conflicts of Interest

None declared.

References

1. Unger JM, Beauchemin M, Hershman DL. Adolescent and young adult enrollment to a National Cancer Institute-sponsored National Clinical Trials Network Research Group over 25 years. Cancer 2021 Dec 15;127(24):4574-4584. [doi: 10.1002/cncr.33855] [Medline: 34351619]
2. Jacob SA, Shaw PH. No improvement in clinical trial enrollment for adolescents and young adults with cancer at a children's hospital. Pediatr Blood Cancer 2017 Dec;64(12). [doi: 10.1002/pbc.26638] [Medline: 28509440]
3. Keegan THM, Ries LAG, Barr RD, et al. Comparison of cancer survival trends in the United States of adolescents and young adults with those in children and older adults. Cancer 2016 Apr 1;122(7):1009-1016. [doi: 10.1002/cncr.29869] [Medline: 26848927]
4. Bleyer A, Budd T, Montello M. Adolescents and young adults with cancer: the scope of the problem and criticality of clinical trials. Cancer 2006 Oct 1;107(7 Suppl):1645-1655. [doi: 10.1002/cncr.22102] [Medline: 16906507]
5. Bleyer A, Viny A, Barr R. Cancer in 15- to 29-year-olds by primary site. Oncologist 2006 Jun;11(6):590-601. [doi: 10.1634/theoncologist.11-6-590] [Medline: 16794238]
6. Bleyer A OLM, Barr R, Ries LAG. Cancer Epidemiology in Older Adolescents and Young Adults 15 to 29 Years of Age, Including SEER Incidence and Survival: 1975-2000: National Cancer Institute (NCI); 2006.
7. Berkman AM, Livingston JA, Merriman K, et al. Long-term survival among 5-year survivors of adolescent and young adult cancer. Cancer 2020 Aug 15;126(16):3708-3718. [doi: 10.1002/cncr.33003] [Medline: 32484922]
8. Kahn JM, Keegan THM, Tao L, Abrahão R, Bleyer A, Viny AD. Racial disparities in the survival of American children, adolescents, and young adults with acute lymphoblastic leukemia, acute myelogenous leukemia, and Hodgkin lymphoma. Cancer 2016 Sep 1;122(17):2723-2730. [doi: 10.1002/cncr.30089] [Medline: 27286322]
9. Collins CL, Malvar J, Hamilton AS, Deapen DM, Freyer DR. Case-linked analysis of clinical trial enrollment among adolescents and young adults at a National Cancer Institute-designated comprehensive cancer center. Cancer 2015 Dec 15;121(24):4398-4406. [doi: 10.1002/cncr.29669] [Medline: 26393950]
10. de Rojas T, Neven A, Terada M, et al. Access to clinical trials for adolescents and young adults with cancer: a meta-research analysis. JNCI Cancer Spectr 2019 Dec;3(4):kz057. [doi: 10.1093/jncics/pkz057] [Medline: 32337483]
11. Roth M, Beauchemin M, Kahn JM, Bleyer A. Patterns of National Cancer Institute-sponsored clinical trial enrollment in Black adolescents and young adults. Cancer Med 2021 Nov;10(21):7620-7628. [doi: 10.1002/cam4.4292] [Medline: 34592782]
12. Rueda A, MacIntyre J. Challenges faced by adolescent and young adult patients with cancer. Oncology Nursing News. 2022. URL: https://www.oncnursingnews.com/view/challenges-faced-by-adolescent-and-young-adult-patients-with-cancer [accessed 2024-05-27]

XSL•FO
RenderX

13. Silin C. Help AYA survivors and patients with cancer navigate infertility. ONS Voice. 2020. URL: https://voice.ons.org/stories/help-aya-survivors-and-patients-with-cancer-navigate-infertility [accessed 2024-05-27]
14. Catley J. Here's why teens and young adults are healthcare digital natives. Health Works Collective. 2018. URL: https://www.healthworkscollective.com/heres-why-teens-and-young-adults-are-healthcare-digital-natives/ [accessed 2025-01-10]
15. Digital natives. American Library Association. 2015. URL: https://www.ala.org/future/trends/digitalnatives [accessed 2024-10-31]
16. Corzo C. Tech savvy plays a role on two generations' quest for helath information, FIU research finds. FIU College of Business. 2024. URL: https://business.fiu.edu/news/2024/tech-savvy-plays-a-role-on-two-generations-quest-for-health-information.html [accessed 2024-10-31]
17. Avutu V, Monga V, Mittal N, et al. Use of communication technology to improve clinical trial participation in adolescents and young adults with cancer: consensus statement from the children's oncology group adolescent and young adult responsible investigator network. JCO Oncol Pract 2022 Mar;18(3):224-231. [doi: 10.1200/OP.21.00554] [Medline: 34905405]
18. Abrahão R, Alvarez EM, Waters AR, et al. A qualitative study of barriers and facilitators to adolescents and young adults' participation in cancer clinical trials: oncologist and patient perspectives. Pediatr Blood Cancer 2022 Apr;69(4):e29479. [doi: 10.1002/pbc.29479] [Medline: 34913583]
19. Sandelowski M. What's in a name? Qualitative description revisited. Res Nurs Health 2010 Feb;33(1):77-84. [doi: 10.1002/nur.20362] [Medline: 20014004]
20. Sandelowski M. Whatever happened to qualitative description? Res Nurs Health 2000 Aug;23(4):334-340. [doi: 10.1002/1098-240x(200008)23:4<334::aid-nur9>3.0.co;2-g] [Medline: 10940958]
21. Kitzinger J. The methodology of focus groups: the importance of interaction between research participants. Soc Health & Illn 1994 Jan;16(1):103-121 [FREE Full text] [doi: 10.1111/1467-9566.ep11347023]
22. Kallio H, Pietilä AM, Johnson M, Kangasniemi M. Systematic methodological review: developing a framework for a qualitative semi-structured interview guide. J Adv Nurs 2016 Dec;72(12):2954-2965. [doi: 10.1111/jan.13031] [Medline: 27221824]
23. What are clinical trials? National Cancer Institute. 2024. URL: https://www.cancer.gov/research/participate/clinical-trials/what-are-clinical-trials [accessed 2024-10-01]
24. Sandelowski M. Notes on transcription. Res Nurs Health 1994 Aug;17(4):311-314. [doi: 10.1002/nur.4770170410] [Medline: 8036279]
25. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005 Nov;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
26. Morse JM. "Data were saturated...". Qual Health Res 2015 May;25(5):587-588. [doi: 10.1177/1049732315576699] [Medline: 25829508]
27. Psihogios AM, Lane-Fall MB, Graham AK. Adolescents are still waiting on a digital health revolution: accelerating research-to-practice translation through design for implementation. JAMA Pediatr 2022 Jun 1;176(6):545-546. [doi: 10.1001/jamapediatrics.2022.0500] [Medline: 35404381]
28. Rideout V, Fox S, Trust WB. Digital health practices, social media use, and mental well-being among teens and young adults in the US. Providence Digital Commons. 2018. URL: https://digitalcommons.providence.org/publications/1093/ [accessed 2024-06-16]
29. Doherty MJ, Thom B, Gany F. Evidence of the feasibility and preliminary efficacy of oncology financial navigation: a scoping review. Cancer Epidemiol Biomarkers Prev 2021 Oct;30(10):1778-1784. [doi: 10.1158/1055-9965.EPI-20-1853] [Medline: 34341051]

Abbreviations

AYA: adolescent and young adult
CTE: clinical trial enrollment
FG: focus group
YA: young adult

Edited by A Mavragani; submitted 12.07.24; peer-reviewed by JG Barnard, L Super, S Chantziara; revised version received 21.11.24; accepted 24.11.24; published 20.01.25.

Please cite as:
Beauchemin MP, Walker D, Rosen A, Frazer M, Eisenberger M, Khurana RK, Bentlyewski E, Fedorko V, Basch CH, Hillyer GC Insights From Diverse Perspectives on Social Media Messages to Inform Young Adults With Cancer About Clinical Trials: Focus Group Study
JMIR Form Res 2025;9:e64265
URL: https://formative.jmir.org/2025/1/e64265
doi:10.2196/64265

XSL•FO
RenderX

© Melissa P Beauchemin, Desiree Walker, Allison Rosen, Maria Frazer, Meital Eisenberger, Rhea K Khurana, Edward Bentlyewski, Victoria Fedorko, Corey H Basch, Grace C Hillyer. Originally published in JMIR Formative Research (https://formative.jmir.org), 20.1.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Sentiments of Individuals with Interstitial Cystitis/Bladder Pain Syndrome Toward Pentosan Polysulfate Sodium: Infodemiology Study

Yulin Hswen[1], MPH, ScD; Qiuyuan Qin[2], MS, PhD; Pressley Smith[3], MS; Alison Swierczynski[3], BS; Stuart Bauer[4,5], MD; Erika Ladson[6], MS, PhD; Amanda Leigh Garrett[6], PhD; Catherine A Brownstein[3,7], MPH, PhD

[1]Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, CA, United States
[2]Department of Public Health and Sciences, University of Rochester, Rochester, NY, United States
[3]Division of Genetics and Genomics, Boston Children's Hospital, Boston, MA, United States
[4]Department of Urology, Boston Children's Hospital, Boston, MA, United States
[5]Department of Surgery, Harvard Medical School, Boston, MA, United States
[6]Inspire, Arlington, VA, United States
[7]Department of Pediatrics, Harvard Medical School, Boston, MA, United States

Corresponding Author:
Catherine A Brownstein, MPH, PhD
Division of Genetics and Genomics, Boston Children's Hospital, Boston, MA, United States

Abstract

Background: Interstitial cystitis/bladder pain syndrome (IC/BPS) is a multifactorial, chronic syndrome involving urinary frequency, urgency, and bladder discomfort. These IC/BPS symptoms can significantly impact individuals' quality of life, affecting their mental, physical, sexual, and financial well-being. Individuals sometimes rely on peer-to-peer support to understand the disease and find methods of alleviating symptoms. The only US Food and Drug Administration–approved medication to treat IC/BPS is pentosan polysulfate sodium (PPS). However, ocular pigmentary maculopathy has been described in some individuals, with greater severity associated with prolonged PPS exposure.

Objective: While prior research has separately assessed the benefits and side effects of PPS, this study sought to identify (1) sentiments of individuals with IC/BPS toward PPS and (2) topics discussed by individuals with IC/BPS in conjunction with PPS through use of an internet peer-to-peer forum.

Methods: Data were collected from Inspire—an anonymous web-based health community where individuals gather by condition to find support and information. Sentiment analysis and percentages of negative, positive, and neutral sentiment for PPS discussions encompassing each topic was conducted using VADER (Valence Aware Dictionary for Sentiment Reasoning). Topic modeling was conducted using latent Dirichlet allocation. Words with the highest probability were ranked to categorize each topic, and authors manually investigated and labeled discussions.

Results: There were 354 forum posts related to PPS. Topic modeling with latent Dirichlet allocation revealed 5 topic categories: "ineffectiveness or discontinued use," "alternative treatments," "personal treatment suggestions based on experience," "severe side effects," and "risk of long-term use." Topics related to "severe side effects" and "risk of long-term use" garnered less discussion, with the former also having the lowest positive sentiment (4.28, 14.29%). The topic "ineffectiveness or discontinued use" was most frequently discussed. This topic also had the highest percentage of negative posts (52/152, 34.21%). However, the average compound score was within the neutral compound score range (−0.094, SD 0.625). In addition, forum data highlighted individuals' acknowledgment of the efficacy of PPS in improving their quality of life, with statements such as "saved my sanity" being representative. The overall compound individuals' sentiment toward PPS was −0.083, split across 32.49% (115/354) negative, 22.03% (78/354) positive, and 45.48% (161/354) neutral sentiment categories.

Conclusions: The overall authentic sentiment toward PPS is broad but balances to neutral. This neutral sentiment suggests that while some individuals express concerns about the side effects and long-term risks associated with PPS, others appreciate its positive impact on their quality of life. This research confirms that individuals with IC/BPS actively engage with health forums like Inspire to seek information, share their experiences, and explore different treatment options. As IC/BPS remains a complex syndrome, this study highlights the value of patient-led discussions in informing treatment decisions. Furthermore, these findings suggest that health care providers might benefit from considering the insights shared on peer-to-peer forums to better understand individual preferences, concerns, and expectations.

XSL•FO
RenderX

(JMIR Form Res 2025;9:e54209) doi:10.2196/54209

KEYWORDS

interstitial cystitis; IC; painful bladder syndrome; bladder pain syndrome; BPS; social media; social network; pain; treatment; chronic condition; chronic disease; chronic illness; Elmiron; pentosan polysulfate sodium; PPS; internet forum

Introduction

Interstitial Cystitis/Bladder Pain Syndrome and Current Treatment

Interstitial cystitis/bladder pain syndrome (IC/BPS) is a multifactorial chronic syndrome that affects the urinary bladder [1]. The condition is characterized by urinary frequency, urgency, and bladder pain [1], involving both women and men [2], that impact their emotional, psychological, and social well-being [1]. The etiology of IC/BPS remains elusive; the disease is diagnosed not with a single test but a combination of multiple investigations to rule out other diagnoses [3]. Many individuals with IC/BPS present heterogeneous syndromes from several causes [4] suggesting it should be classified into different subtypes [5]. Recent investigations of genetics and biomarkers have begun to uncover the molecular mechanisms of IC/BPS [4,6]. First-degree relatives of women with IC/BPS have a greater prevalence of the disease than women without affected relatives, indicating a genetic component [7,8]. Twin studies have shown a greater concordance of IC/BPS among monozygotic than dizygotic twins [7,9]. The Golgi-localized adenosine triphosphatase gene *ATP2C1* has also been identified as having increased rare variant burden in an IC/BPS cohort (OR [odds ratio]=6.76, 95% CI 1.1-75.9) [6] and is a candidate gene.

Pentosan Polysulfate Sodium–Associated Pigmentary Maculopathy

Currently, the only US Food and Drug Administration (FDA)–approved treatment is oral pentosan polysulfate sodium (PPS). Use of PPS has been demonstrated to improve symptoms [10], although recent studies link PPS to a unique presentation of pigmentary maculopathy [11]. Pigmentary maculopathy can produce moderate visual impairment, macular hyperpigmented spots, yellow-orange deposits, and patchy retinal pigment epithelium atrophy [12]. The first study detailed 6 individuals over 2 years who developed new visual symptoms after exposure to PPS [11]. Several additional studies have replicated the association between PPS use and pigmentary maculopathy [13-16].

Support Groups and the Inspire Platform

Patients and caregivers often join web-based health communities to find support for incurable or rare diseases. Through interpersonal exchanges, patients and caregivers can develop an understanding of their illnesses. Social Comparison Theory postulates that social behaviors could be predicted on the assumption that individuals seek the maintenance of normalcy and accuracy in their world by using opinions of others to evaluate their thinking and feeling [17]. The benefits of support groups can be explained by social support proximal factors, for example, attachment, group membership, and collective identity [18]. Secure adult attachment is described as having comfort, intimacy, and trust such that another person can provide support, empathy, and care. Thus, individuals with chronic illnesses seek comfort from others who can provide this type of attachment [18]. There is also evidence that attachment security benefits individuals with health-related distress, as it was found to be negatively associated with depression among individuals with chronic pain. Therefore, illness support groups can provide security and a positive effect on individuals [18]. Many forums allow users to maintain anonymity, which can encourage more honest and candid discussions, especially about sensitive topics, while being easily accessible to anyone with internet access.

Group membership also describes the evolution of humans having an automatic tendency to categorize themselves and others as in-group and out-group based on similar characteristics, for example, having the same chronic illness [18]. Individuals with chronic illnesses sometimes deviate from social norms, that may impact their physical appearance and actions and can lead to stigmatization, discrimination, and social exclusion, threatening a sense of belonging. Due to these effects, individuals look toward in-group membership, leading to beneficial cohesion [18].

The perceived benefits of participating in support communities include meeting and befriending others with the same condition and similar experiences, learning about the disease and related treatments, giving and receiving emotional support to group members, and accountability for reaching health-related goals [19-21]. Digital communities often serve as spaces where users collaborate to find solutions to common challenges, drawing on the collective wisdom of the group [22].

The Inspire platform is the world's largest and fastest-growing health community designed as an environment for individuals to feel comfortable sharing personal experiences and sensitive health information on an anonymous forum. More than 1o million annual users on the Inspire platform have joined hundreds of disease-specific community groups with patients and caregivers with similar diseases [23]. The Inspire platform maintains anonymity while allowing individuals to obtain information about, and participate in, clinical trials and research efforts [23]. The Interstitial Cystitis Association community on Inspire has over 40,000 members and is the most active IC/BPS forum [23].

Web-based communities provide a rich source of information for health research [24]. For example, monitoring web-based social networks has been an effective tool for identifying adverse events [25] and barriers to accessing health care [26]. The use of web-based social networks for disease surveillance is well documented [27-30]. Notably, data from web-based networks has also been used to obtain a deeper understanding of patient sentiment [31]. Therefore, this study sought to identify (1) sentiments of individuals with IC/BPS toward PPS and (2) topics

XSL•FO
RenderX

discussed by individuals with IC/BPS in conjunction with PPS through use of an internet peer-to-peer forum.

Methods

Study Design and Data Sources

To examine both the sentiment and the topics discussed concerning PPS, the Inspire forum for individuals with IC/BPS was used to derive quantifiable data points. We used the IC/BPS discussion forum posts and replies from the United States from January 12, 2019, to November 30, 2022. In our analytical sample, 3232 patient forum posts were collected from the Inspire forum, and 354 discussed PPS. The data corpus was created to gain insight into the triggers or factors flaring or worsening IC/BPS symptomology. As such, specific text terms and razor tags were used to bring these topics to the surface: flare-, flair-, worse-, increase-, more, trigger-, caus-, inducing, induce-, producing, produce-, factor, factors, provok-, activat-, exacerbate-, irritat-, hypersensitive-, sensiti-, react-, risky, acute, #flare, #riskfactor, #environmentalfactor, #cause, #causality, #hypersensitivity, and #sensitization. Texts were analyzed for specific discussions regarding PPS and other concerns of individuals with IC/BPS. Conversations and discussion posts were limited to sentences with the term PPS and the trade name for PPS. These sentences were used for further analysis alongside the previous sentence relative to the usage of “PPS” and the following sentence relative to the usage of “PPS”.

All analyses used this definition of PPS-related discussions and were conducted using Python version 3.6 (Python Software Foundation). In addition, a combination of topic modeling and sentiment analysis was used in the retroactive investigation.

Topic Modeling

Topic modeling is statistical modeling for discovering abstract “topics” in a text collection. Latent Dirichlet allocation (LDA) is a statistical topic model used to discover and classify text to a particular collection. LDA builds a topic per document model and words per topic model, creating Dirichlet distributions. Each topic has a probability of generating different words in the text. Topics are surfaced based on the likelihood of word co-occurrence and the probability of words making up each topic.

Before using LDA for topic modeling, raw data were preprocessed (lower cased, removing URL links, nonletter characters, and abbreviations) and tokenized (divided into sets) by returning a list of strings for each post to prepare for analysis using the GenSim package in Python [32]. Furthermore, bigrams were used on the tokenized data to combine frequently occurring words to analyze topics. Bigrams were also implemented on the tokenized data to combine the words that occurred together frequently. Common stop words such as “the,” “a,” “an,” “in,” and so on, were filtered out using the Natural Language Toolkit in Python, which contains a list of stop words stored in the English language [33]. The spacy package was used to lemmatize words (reducing words to their root word, eg, “running” and “ran” can be reduced to “run”) and ensure the thorough removal of stop words. The words “PPS” (and the associated trade name) were also excluded from the LDA analysis because the data were filtered explicitly for discussions about “PPS.” For the intertopic distance map, multidimension scaling is used for the visualization of the topics in a 2D space with the 2 principal components, prinicipal component 1 and principal component 2. We used coherence score to select the potential numbers of topics for LDA model. Coherence score measures the consistency, clarity, and relevance of each topic in LDA modeling [34,35]. We then selected the maximum number of topics that could be produced without overlapping the intertopical distance map. We also used perplexity score to examine the model performance. Perplexity score measures the predication performance of a LDA model on held-out documents [36,37]. The size of the topic represented by circles, is proportional to the proportion of posts about PPS by individuals on Inspire. Topics having more words in common are represented closer together, thus the distance between the topics is related to how different the topics are in the posts.

After preprocessing and tokenizing the data, each forum post was categorized to the LDA topic with the highest proportion to identify hidden topics within LDA modeling. Words with the highest probably were ranked for each topic. Finally, topics were manually labeled to determine the overarching theme for each topic (AS, EL, CAB, and ALG).

Sentiment Analysis

Sentiment analysis was used to computationally identify and categorize opinions expressed on the IC/BPS Inspire forum to find the emotional tone behind PPS-associated text. Sentiment analysis was conducted using the Valence Aware Dictionary for Sentiment Reasoning (VADER). VADER is a lexicon and rule-based sentiment analysis tool to identify the sentiment attributed to a topic of interest. It is also considered the standard for analyzing sentences to determine the sentiment of a piece of literature.

The results generated by VADER include a dictionary of 4 percentage scores classified as negative, neutral, positive, and compound. The negative, neutral, and positive scores are the relative percentages of negative, neutral, and positive sentiments surrounding the topic discussion. The compound score is a continuous score ranging from −1 (most negative) to +1 (most positive). Definitions most commonly used for compound scores are: −1.0 to −0.5 (a negative score), −0.5 to +0.5 (a neutral score), and +0.5 to +1.0 (a positive score). Sentiment analysis was conducted on all PPS-related discussions and on all topics determined by LDA topic.

Ethical Considerations

Inspire community members may anonymously engage with others through public discussion posts in the forum. The participants agree to have their posts or discussion comments used for research purposes upon registering for Inspire. Deidentified community posts, with all personally identifiable information removed, are available for secondary data analysis. Individuals were not compensated for participating in the forum. This study was approved by the Boston Children’s Hospital Institutional Review Board (IRB-P00040767).

XSL•FO
RenderX

Results

Overview

Results are based on the 354 posts that related to PPS. Examples of forum posts with sentiment categorization by VADER are listed in Table 1.

Table . Select interstitial cystitis/bladder pain syndrome forum posts describing their experiences with pentosan polysulfate sodium. Posts with positive, negative, and neutral sentiment are shown.

Example notable quotes	Sentiment categorization by VADER[a]
"IC is not just an inflamed bladder. It is a disease with no known cause or cure. It is a lifetime of pain and misery. I know I have had it for over 20 years now. IC causes the lining of your bladder to shed each time you urinate. It is so painful it feels like lava coming through your thighs."	Overall neutral sentiment toward topic 1 (ineffectiveness or discontinuing use)
"Interstitial Cystitis is insidious, terribly painful, and has been a direct causality of what I can and could not do throughout the entire duration of my lifetime."	Overall negative sentiment toward topic 3 (personal treatment suggestions based on the experience)
"You must be patient and not give up hope, be willing to try different remedies as not one thing works for everyone. Relief can be fleeting and not every day since the condition is chronic."	Overall positive sentiment toward topic 3 (personal treatment suggestions based on the experience)
"…with all the class action suits, I got scared and stopped taking [...] I have had several flare-ups since I stopped and no quality of life, as someone else said. I think eyesight is a big side effect to consider, but the alternative is pain [...] still deciding if I should go back on it?"	Overall neutral sentiment toward topic 1 (ineffectiveness or discontinuing use)
"[...] actually would give me a reasonable baseline of normal, but I have chosen to go off of it due to the new vision concerns. I am praying that they will determine the vision issue to be very low risk, and I will consider going back on because it is really what works best for me."	Overall positive sentiment toward topic 2 (alternative treatment)
"I foolishly went off [...] in January because of the possible retinal eye damage and of course the high cost. Really big mistake. I'm back on [...] and hoping that it doesn't take 3 - 4 months for it to kick in again."	Overall negative sentiment toward topic 3 (personal treatment suggestions based on the experience)
"a lifesaver because [they] also [were] in severe pain, burning, etc, and running to the bathroom. After [...], [they] had no pain, [they] could eat anything [they] wanted."	Overall negative sentiment toward topic 5 (risk of long-term use)
"I am so glad to find someone besides myself that isn't stopping [...] I was suicidal before [...] went to work on me. The fear of returning to the hot mess I was before taking [...] terrifies me. I am comforted to find out I am not the only one sticking with it."	Overall negative sentiment toward topic 5 (risk of long-term use)
"[...] is garbage and has terrible side effects."	Overall negative sentiment toward topic 1 (ineffectiveness or discontinuing use)
"antihistamines, amitriptyline, [...], and an ultra-strict diet seem to help for a few days, and then I am right back to where I started."	Overall negative sentiment toward topic 1 (ineffectiveness or discontinuing use)

[a]VADER: Valence Aware Dictionary for Sentiment Reasoning.

In Table 1, all posts were analyzed using VADER for sentiment categorization. To determine sentiment, VADER begins by splitting the text into individual words. It then assigns each word a score to indicate whether it is positive or negative. Using these predefined scores, VADER computes an overall sentiment score for the entire text. VADER also accounts for sentiment intensity (eg, capitalization and punctuation) and looks for modifiers that might alter the meaning of nearby words.

LDA Modeling Output

LDA topic modeling revealed 5 topics, the minimum number without overlap. Figure 1 displays the intertopic distance map, the area of the topic circles proportional to the number of words that belong to each topic. The distance between topics represents the degree of difference between each. Topics having more words in common are closer together. The greatest proportion of discussions was topic 1 (152/354), then topic 2 (92/354), topic 3 (56/354), topic 4 (28/354), and topic 5 (26/354). In Figure 1, topics 1, 2, and 5 were more closely clustered than topics 3 and 4, indicating they were most dissimilar to topics 1, 2, and 5. The highest probability of co-occurring words making up each LDA topic is shown in Table 2. Using the co-occurence of words and manual investigation, Topic 1 was identified as "ineffectiveness or discontinued use," topic 2 as "alternative treatments," topic 3 as "personal treatment suggestions based on experience," topic 4 as "severe side effects," and topic 5 as "risk of long-term use."

XSL•FO
RenderX

Figure 1. Display of intertopic distance and marginal distribution of the 5 topics. Instead of representing the divergence of topics as a phylogentic tree in 1D, this map allows visualization of how topics relate in 2D space–based principal components. Intertopic distance map measures the distance between topics. Similar topics are closer to each other, while very different topics are farther apart. The size of the circle represents the number of words in each topic; larger circles mean more words belong to that topic.

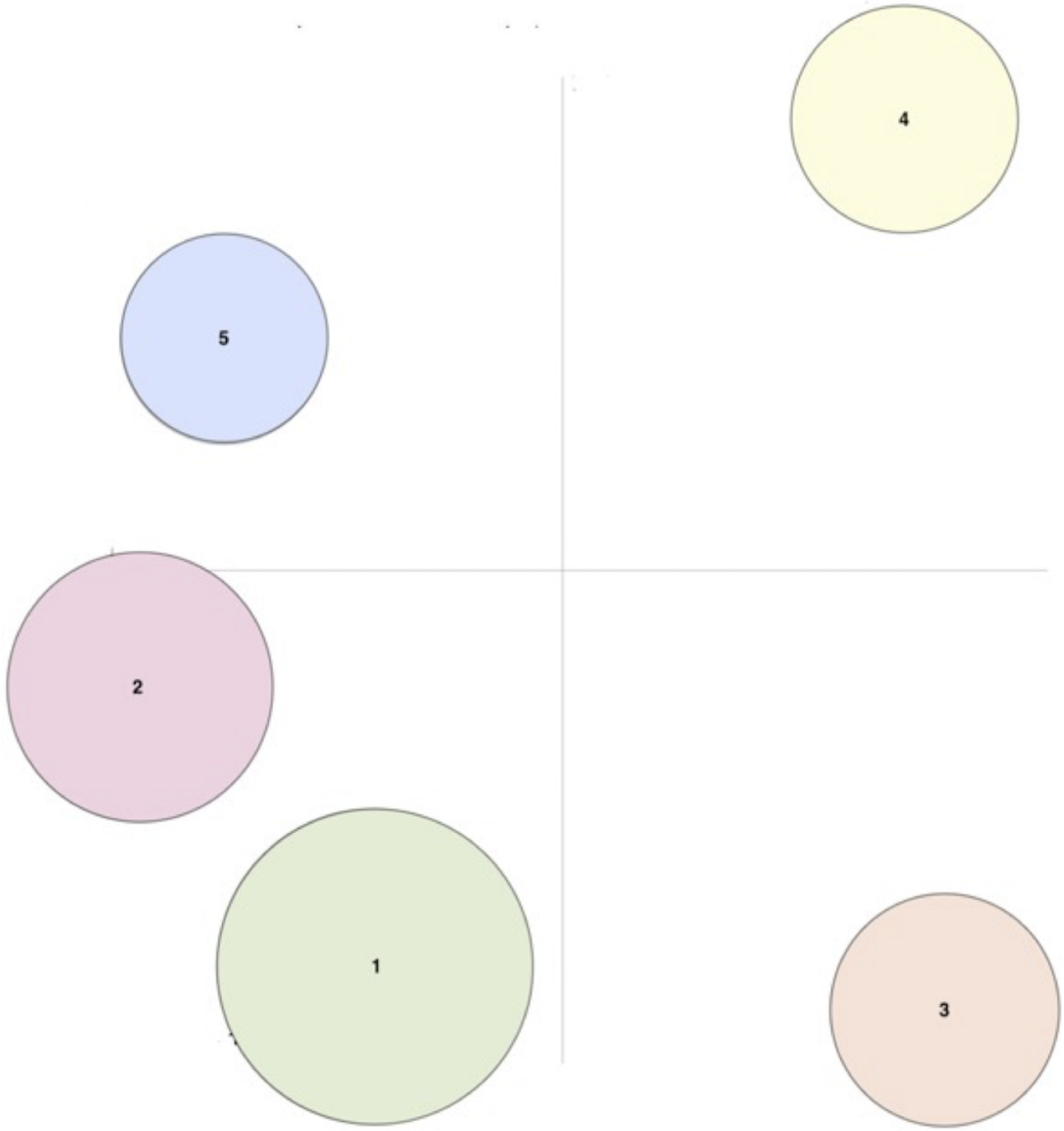

XSL•FO
RenderX

Table . Latent Dirichlet allocation word occurrence probability for pentosan polysulfate sodium–related discussions in the interstitial cystitis/bladder pain syndrome forum.

Topic 1: ineffectiveness or discontinued use (152/354, 42.9%)		Topic 2: alternative treatments (92/354, 26%)		Topic 3: personal treatment suggestions based on the experience (56/354, 15.8%)		Topic 4: severe side effects (28/354, 7.9%)		Topic 5: risk of long-term use (26/354, 7.3%)	
Probability	Word	Probability	Word	Probability	Word	Probability	Word	Probability	Word
.08	Take	.08	Go	.12	Eye	.05	Try	.05	Night
.08	Year	.08	Also	.09	Damage	.04	Bladder	.04	Problem
.04	Stop	.07	Pain	.07	Cause	.04	Start	.04	Really
.04	Flare	.04	Help	.06	Retinal	.04	Have Been	.04	Think
.03	IC	.04	Month	.05	See	.03	Back	.03	Pigmentary
.03	Time	.03	Day	.04	Issue	.03	Continue	.03	Vision
.03	Work	.02	Capsule	.04	Hear	.03	Make	.03	Maculopathy
.03	Find	.02	Get	.03	Vision	.03	Still	.03	Macular degeneration
.02	Thing	.02	Med	.03	Retina	.03	Medication	.02	End
.02	Know	.02	Diet	.03	Take	.03	Burn	.02	Cost
.02	*Help*	.02	Bladder	.03	Ask	.02	Due	.02	Symptom
.02	People	.02	Give	.03	Drug	.02	Want	.02	Many
.02	Try	.02	Back	.03	Condition	.02	Give	.02	Much
.02	Dr	.01	Week	.02	Doctor	.02	Ever	.02	Risk
.02	Hope	.01	Mg	.02	Long	.02	Instillation	.02	Course
.02	Lot	.01	Seem	.02	Hair	.02	Fall	.02	Patient
.02	Life	.01	Urologist	.02	Light	.02	Free	.02	Hour
.02	Month	.01	Time	.01	Specific	.02	Develop	.02	Concern
.01	Possible	.01	Food	.01	Term	.02	Little	.02	Hydroxyzine
.01	Day	.01	Eat	.01	Stop	.02	Pill	.02	Helpful

In Table 2, the LDA modeling method is used to reveal the main topics and how they are distributed across a dataset (ie, the IC/BPS forum). Five main topics were identified. The proportion, number of posts discussing each topic, and manual topic labels for each LDA topic are shown.

Sentiment and Topic Analysis

The overall sentiment scores for each topic are presented in Table 3. Overall, the VADER compound sentiment of discussions about PPS was −0.083, (negative sentiment 32.49% [115/354], positive 22.03% [78/354], and neutral 45.48% [161/354]).

Table . Sentiment scores for discussions surrounding pentosan polysulfate sodium (n=354). Posts were stratified by the identified topics from the latent Dirichlet allocation analysis.

Topic number and label	Average compound score (SD)	Negative, n (%)	Positive, n (%)	Neutral, n (%)
Total (n=354)	−0.083 (0.603)	115 (32.49)	78 (22.03)	161 (45.48)
Topic 1: ineffectiveness or discontinuing use (n=152)	−0.094 (0.625)	52 (34.21)	35 (23.03)	65 (42.76)
Topic 2: alternative treatments (n=92)	−0.082 (0.620)	31 (33.7)	23 (25)	38 (41.3)
Topic 3: personal treatment suggestions based on the experience (n=56)	−0.004 (0.573)	15 (26.79)	12 (21.43)	29 (51.79)
Topic 4: severe side effects (n=28)	−0.162 (0.523)	9 (32.14)	4 (14.29)	15 (53.57)
Topic 5: risk of long-term use (n=26)	−0.119 (0.581)	8 (30.77)	4 (15.38)	14 (53.85)

XSL•FO
RenderX

In Table 3, each topic number represents a specific theme that emerged during the analysis of the text corpus. The average compound score indicates whether the overall tone of discussions for each topic is positive, negative, or neutral. The number and percentage of scored negative, postitive, and neutral posts is shown. By comparing sentiment across these topic numbers, we can gain insights into how different aspects of PPS are perceived within the discussions.

Topic 1 identified by the LDA topic modeling was manually labeled "ineffectiveness" based on the total posts about PPS on the IC/BPS Inspire forum using its relative size in Figure 1 and number of posts in Table 3. In Topic 1, the highest probability of co-occurring and relevant words compromising this topic were: take (.08), year (.08), stop (.04), and flare (.04). Its distance in relationship to topics 2 (alternative treatments) and 5 (risk of long-term use), seen in Figure 1, indicates a higher degree of similarity based on the types of words they have in common. Topics 4 (severe side effects) and 3 (personal treatment suggestions based on experience) were furthest from topic 1. shows the overall compound scores, and percent negative, positive, and neutral for the entire dataset and for each topic of PPS posts by individuals from Inspire. Topic 1 had the highest percentage of negative posts with 34.21% (52/152), 23.03% (35/152) positive sentiment posts, while the percent of neutral sentiment was 42.76% (65/152). However, despite having the most percent negative posts compared with any other topic, the average compound score was −0.094 which falls within the neutral compound score range of −0.50 to +0.50.

Topic 2 was identified by LDA modeling and manually labeled as "alternative treatments." This topic had the most overlap of words in common with topic 1 Figure 1. Topic 2 was the second largest topic with 26% (92/354) of the total posts about PPS on the IC/BPS Inspire forum. The highest probability of co-occurring words most relevant terms to the topic was: pain (.07), help (.04), diet (.02), urologist (.01), time (.01). Moreover, it had the second-largest proportion of discussions on the IC/BPS Inspire forum and the second most posts associated with the topic; note its relative size (Figure 1). Compared with the other 4 topics, topic 2 had the second-highest percentage of negative posts at 33.7% (31/92). However, it also had the highest percentage positive sentiment score (23/92, 25%) and percentage of neutral (38/92, 41.3%). The overall compound score in topic 2 was −0.082, the second most negative, falling within the neutral compound score range of −0.50 to +0.50.

Topic 3 was identified by LDA modeling and manually labeled as "personal treatment suggestions based on experience." It was the third largest topic with 15.8% (56/354) about PPS on the IC/BPS Inspire forum. The highest probability of co-occurring words most relevant to the topic was: eye (.12), damage (.09, retinal (.06), vision (.03), retina (.03), and doctor (.02). Compared with the other 4 topics, topic 3 had the least negative compound sentiment (−0.004) and the least percent of negative (15/26, 26.79%), positive (12/56, 21.43%), and (29/56, 51.79%) neutral posts. This is reflective of the distance of topic 3 in the intertopic distance map whereby it is further from to topic 2.

Topic 4 was identified by the LDA modeling and manually labeled as "severe side effects." It was the second smallest topic with 7.9% (28/354) of the total posts about PPS on the IC/BPS Inspire forum. The highest probability of co-occurring words most relevant to the topic was bladder (.04), medication (.03), burn (.02), and instillation (.02). It had the highest intertopic distance compared with the topics 2, 1, and 3, making it the least similar topic based on word commonality. Topic 4 had the most negative compound score, −0.162 with 32.14% (9/28) negative, 14.29% (4/28) positive, and 53.57% (15/28) neutral posts.

Topic 5 was identified by the LDA modeling and manually labeled as "risk of long-term use." It was the smallest topic with 7.3% (26/354) about PPS on the IC/BPS Inspire forum. The highest probability of co-occurring words most relevant to the topic was: night (.05), problem (.04), pigmentary (.03), vision (.03), maculopathy (.03), and macular degeneration (.03). The average compound score for was −.119 (SD 0.581) which falls within the neutral compound range of −0.50 to +0.50. It had the highest percentage of neutral sentiment posts (14/26, 53.85%), compared with the other topics.

Discussion

Principal Findings

PPS is widely prescribed to relieve IC/BPS symptoms [1]; it is the only FDA-approved oral medication available for symptom alleviation. However, PPS has been linked to a unique presentation of pigmentary maculopathy [16]. This potential side-effect has created a need for individuals exposed to PPS to be consistently monitored by an optometrist for early detection [11,13].

Although previous studies have assessed the benefits and side effects of PPS, they have done so separately without capturing individuals' thoughts on if or why they are willing to assume the side effects for the relief of pain. This study provides information as to the sentiment of the IC/BPS community toward PPS, which was determined to be neutral, with 5 topics discussed in the forum all producing a neutral compound score.

Studying sentiment toward medications offers several important benefits, beginning with improved patient-centered care that aligns with individual preferences, concerns, and quality of life. An understanding of patient sentiment may assist health care professionals in assessing whether the perceived benefits of a medication outweigh the risks from the patient's perspective. As the data here indicate that patient sentiment is neither strongly for nor against PPS, patients may be less influenced by emotional or external biases, making it easier to assess the medication's effects based on objective outcomes rather than preconceived notions. This can inform shared decision-making and guide personalized treatment options. It is arguable that understanding patient's sentiment toward their medication could reduce anxiety and improve trust, both of which are related to treatment adherence [38].

As a whole, the forum data describe a willingness among some individuals to risk the potential side effects associated with PPS due to its effectiveness at relieving pain. Multiple posts indicate PPS, "saved [their] sanity" or, "changed [their] life" (Table 1). However, the proportion of negative sentiment is close to

XSL•FO
RenderX

one-third (115/354, 32.49%), with positive sentiment taking up a proportion of just over one-fifth (78/354, 22.03%) and neutral just under one half (161/354, 45.48%). The mean compound sentiment score in the forum is −0.083 (SD 0.603), indicating an overall neutral sentiment.

Topics 4 (severe side effects) and 5 (risk of long-term use), were the least discussed. While topic 4 did not surface as often in the LDA model and had a lower discussion frequency, it did have the lowest positive sentiment. Topic 5 had the highest neutral sentiment of the topics.

Limitations

Some drawbacks of using web-based health communities such as Inspire include difficulty controlling the quality of information shared on such sites. Uncertain and unknowable generalizability of these results is another limitation. Other potential drawbacks include privacy risks to patients and caregivers utilizing the platform, as "data intruders" could lead to discrimination by employers, insurance companies, family, and friends. Although there are policies, such as the Genetic Information Nondiscrimination Act, which prohibits genetic information discrimination in employment, privacy is still a risk that must be considered [20] and individuals may be unwilling to speak freely.

VADER uses a lexicon-based approach system. In such an approach, while broader term coverage is made available, words have a finite assignment to sentiment orientation, which does not consider figures of speech or negations, as mentioned above [39]. LDA topic modeling limitations include the number of topics that can be generated and the inability to depict correlations [40,41]. Topics are not developed over time, and the modeling assumes that words are exchangeable and that sentence structure is not modeled.

Confounding factors, such as current symptoms and overall IC/BPS management, should be included in future studies to further the understanding of the underlying determinants of sentiment toward PPS. Comparitive attitudes toward different IC/BPS treatments should also be explored to assess varying disease management strategies.

Conclusions

Individuals use health forums such as Inspire to learn more about their disease, medications, and treatments. Therefore, this study aimed to use forum posts from Inspire to investigate individuals with IC/BPS' overall sentiment toward PPS. Using VADER, sentiments regarding PPS were found to be neutral. These results also suggest that patients are open to considering both the benefits and risks of PPS, which can lead to informed and balanced decision-making regarding their IC/BPS treatment. Both the breadth of discussion topics and sentiment demonstrate the complexity inherent in individual consideration of a medication with publicized side effects. Future studies of sentiment analyses toward other IC/BPS treatments and their comparison to PPS may provide additional context and direction in the treatment of this painful condition.

Acknowledgments

We thank the patients, families, and caregivers for their participation in this study. This publication was supported by the Centers for Disease Control and Prevention of the U.S. Department of Health and Human Services (HHS) as part of a financial assistance award totaling $2,400,000 with 100 percent funding from the CDC/HHS (U01DP006634-01). The contents are those of the author(s) and do not necessarily represent the official views of, nor are they endorsed by, the CDC/HHS or the U.S. Government.

Data Availability

Aggregated and anonymized data are available upon reasonable request to Inspire [42].

Conflicts of Interest

None declared.

References

1. Lim Y, O'Rourke S. Interstitial Cystitis: StatPearls Publishing; 2022.
2. Messing EM, Stamey TA. Interstitial cystitis: early diagnosis, pathology, and treatment. Urology 1978 Oct;12(4):381-392. [doi: 10.1016/0090-4295(78)90286-8] [Medline: 213864]
3. Hanno P, Nordling J, van Ophoven A. What is new in bladder pain syndrome/interstitial cystitis? Curr Opin Urol 2008 Jul;18(4):353-358. [doi: 10.1097/MOU.0b013e3282fcea88] [Medline: 18520754]
4. Liu HT, Kuo HC. Biomarkers for patients with interstitial cystitis/bladder pain syndrome. Urol Sci 2015 Dec;26(4):225-229. [doi: 10.1016/j.urols.2015.02.002]
5. Akiyama Y, Hanno P. Phenotyping of interstitial cystitis/bladder pain syndrome. Int J Urol 2019 Jun;26 Suppl 1:17-19. [doi: 10.1111/iju.13969] [Medline: 31144756]
6. Estrella E, Rockowitz S, Thorne M, et al. Mendelian disorders in an interstitial cystitis/bladder pain syndrome cohort. Adv Genet (Hoboken) 2023 Mar;4(1):2200013. [doi: 10.1002/ggn2.202200013] [Medline: 36910591]
7. Warren JW, Jackson TL, Langenberg P, Meyers DJ, Xu J. Prevalence of interstitial cystitis in first-degree relatives of patients with interstitial cystitis. Urology 2004 Jan;63(1):17-21. [doi: 10.1016/j.urology.2003.08.027] [Medline: 14751339]

XSL•FO
RenderX

8. Dimitrakov JD. A case of familial clustering of interstitial cystitis and chronic pelvic pain syndrome. Urology 2001 Aug;58(2):281. [doi: 10.1016/s0090-4295(01)01138-4] [Medline: 11489726]
9. Warren JW, Keay SK, Meyers D, Xu J. Concordance of interstitial cystitis in monozygotic and dizygotic twin pairs. Urology 2001 Jun;57(6 Suppl 1):22-25. [doi: 10.1016/s0090-4295(01)01120-7] [Medline: 11378045]
10. Parsons CL, Schmidt JD, Pollen JJ. Successful treatment of interstitial cystitis with sodium pentosanpolysulfate. J Urol 1983 Jul;130(1):51-53. [doi: 10.1016/s0022-5347(17)50948-9] [Medline: 6191049]
11. Pearce WA, Chen R, Jain N. Pigmentary maculopathy associated with chronic exposure to pentosan polysulfate sodium. Ophthalmology 2018 Nov;125(11):1793-1802. [doi: 10.1016/j.ophtha.2018.04.026] [Medline: 29801663]
12. Parkinson R, Thurau S, Maag KP. Pigmentary maculopathy in interstitial cystitis/bladder pain syndrome treated with oral pentosan polysulfate: a review. EMJ Urol 2022 Aug 15:74-81. [doi: 10.33590/emjurol/10054331]
13. Kailavasan M, Goddard JC. Association of Elmiron (pentosanpolysulphate sodium) with pigmented maculopathy: an update for urologists and patients. J Clin Urol 2023 Sep;16(5):458-462. [doi: 10.1177/20514158211053699]
14. Doiron RC, Bona M, Nickel JC. Possible drug-induced, vision-threatening maculopathy secondary to chronic pentosan polysulfate sodium (Elmiron®) exposure. Can Urol Assoc J 2020 Feb;14(2):10-11. [doi: 10.5489/cuaj.6401] [Medline: 31999543]
15. Higgins K, Welch RJ, Bacorn C, et al. Identification of patients with pentosan polysulfate sodium-associated maculopathy through screening of the electronic medical record at an academic center. J Ophthalmol 2020;2020:8866961. [doi: 10.1155/2020/8866961] [Medline: 33489347]
16. Philip AM, Wannamaker KW, Miller DM. Prevalence and dose dependency analysis of pentosan polysulfate sodium maculopathy. Ophthalmic Epidemiol 2023 Feb;30(1):82-87. [doi: 10.1080/09286586.2022.2031227] [Medline: 35081852]
17. Davison KP, Pennebaker JW, Dickerson SS. Who talks? The social psychology of illness support groups. Am Psychol 2000 Feb;55(2):205-217. [Medline: 10717968]
18. Lehardy EN, Fowers BJ. Ultimate (evolutionary) explanations for the attraction and benefits of chronic illness support groups: attachment, belonging, and collective identity. Curr Psychol 2020 Aug;39(4):1405-1415. [doi: 10.1007/s12144-018-9841-7]
19. Delisle VC, Gumuchian ST, Rice DB, et al. Perceived benefits and factors that influence the ability to establish and maintain patient support groups in rare diseases: a scoping review. Patient 2017 Jun;10(3):283-293. [doi: 10.1007/s40271-016-0213-9] [Medline: 28004275]
20. Solberg LB. The benefits of online health communities. Virtual Mentor 2014 Apr 1;16(4):270-274. [doi: 10.1001/virtualmentor.2014.16.04.stas1-1404] [Medline: 24735575]
21. Wicks P, Keininger DL, Massagli MP, et al. Perceived benefits of sharing health data between people with epilepsy on an online platform. Epilepsy Behav 2012 Jan;23(1):16-23. [doi: 10.1016/j.yebeh.2011.09.026] [Medline: 22099528]
22. Wicks P, Massagli M, Frost J, et al. Sharing health data for better outcomes on PatientsLikeMe. J Med Internet Res 2010 Jun 14;12(2):e19. [doi: 10.2196/jmir.1549] [Medline: 20542858]
23. Inspire. URL: https://www.inspire.com [accessed 2025-01-02]
24. Brownstein JS, Freifeld CC, Madoff LC. Digital disease detection--harnessing the web for public health surveillance. N Engl J Med 2009 May 21;360(21):2153-2155. [doi: 10.1056/NEJMp0900702] [Medline: 19423867]
25. Lee JY, Lee YS, Kim DH, Lee HS, Yang BR, Kim MG. The use of social media in detecting drug safety-related new black box warnings, labeling changes, or withdrawals: scoping review. JMIR Public Health Surveill 2021 Jun 28;7(6):e30137. [doi: 10.2196/30137] [Medline: 34185021]
26. Kline E, Garrett AL, Brownstein C, et al. Using social media listening to understand barriers to genomic medicine for those living with Ehlers-Danlos syndromes and hypermobility spectrum disorders. Health Expect 2023 Aug;26(4):1524-1535. [doi: 10.1111/hex.13755] [Medline: 37062887]
27. Bisanzio D, Kraemer MUG, Brewer T, Brownstein JS, Reithinger R. Geolocated Twitter social media data to describe the geographic spread of SARS-CoV-2. J Travel Med 2020 Aug 20;27(5):taaa120. [doi: 10.1093/jtm/taaa120] [Medline: 32701135]
28. Harris JK, Hawkins JB, Nguyen L, et al. Using Twitter to identify and respond to food poisoning: The Food Safety STL Project. J Public Health Manag Pract 2017;23(6):577-580. [doi: 10.1097/PHH.0000000000000516] [Medline: 28166175]
29. Young SD, Rivers C, Lewis B. Methods of using real-time social media technologies for detection and remote monitoring of HIV outcomes. Prev Med 2014 Jun;63:112-115. [doi: 10.1016/j.ypmed.2014.01.024] [Medline: 24513169]
30. Clemente L, Lu F, Santillana M. Improved real-time influenza surveillance: using internet search data in eight Latin American countries. JMIR Public Health Surveill 2019 Apr 4;5(2):e12214. [doi: 10.2196/12214] [Medline: 30946017]
31. Sewalk KC, Tuli G, Hswen Y, Brownstein JS, Hawkins JB. Using Twitter to examine web-based patient experience sentiments in the United States: longitudinal study. J Med Internet Res 2018 Oct 12;20(10):e10043. [doi: 10.2196/10043] [Medline: 30314959]
32. Haider MM, Hossin M, Mahi HR, Arif H. Automatic text summarization using gensim word2vec and k-means clustering algorithm. Presented at: 2020 IEEE Region 10 Symposium (TENSYMP); Jun 5-7, 2020; Dhaka, Bangladesh. [doi: 10.1109/TENSYMP50017.2020.9230670]

XSL•FO
RenderX

33. Bird S, Klein E, Loper E. Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit: O'Reilly Media, Inc; 2009, Vol. ISBN.
34. O'Callaghan D, Greene D, Carthy J, Cunningham P. An analysis of the coherence of descriptors in topic modeling. Expert Syst Appl 2015 Aug;42(13):5645-5657. [doi: 10.1016/j.eswa.2015.02.055]
35. Syed S, Spruit M. Full-text or abstract? Examining topic coherence scores using latent Dirichlet allocation. Presented at: 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA); Oct 19-21, 2017; Tokyo, Japan. [doi: 10.1109/DSAA.2017.61]
36. Zhao W, Chen JJ, Perkins R, et al. A heuristic approach to determine an appropriate number of topics in topic modeling. BMC Bioinformatics 2015;16:S8. [doi: 10.1186/1471-2105-16-S13-S8] [Medline: 26424364]
37. Rüdiger M, Antons D, Joshi AM, Salge TO. Topic modeling revisited: new evidence on algorithm performance and quality metrics. PLoS ONE 2022;17(4):e0266325. [doi: 10.1371/journal.pone.0266325] [Medline: 35482786]
38. Costello K, Kennedy P, Scanzillo J. Recognizing nonadherence in patients with multiple sclerosis and maintaining treatment adherence in the long term. Medscape J Med 2008;10(9):225. [Medline: 19008986]
39. D'Andrea A, Ferri F, Grifoni P, Guzzo T. Approaches, tools and applications for sentiment analysis implementation. IJCA 2015;125(3):26-33. [doi: 10.5120/ijca2015905866]
40. Barde BV, Bainwad AM. An overview of topic modeling methods and tools. Presented at: 2017 International Conference on Intelligent Computing and Control Systems (ICICCS); Jun 15-16, 2017; Madurai, India. [doi: 10.1109/ICCONS.2017.8250563]
41. Lafferty J, Blei D, editors. Correlated Topic Models: MIT Press; 2005.
42. Inspire research. Inspire. URL: https://inspireresearch.com/ [accessed 2025-01-07]

Abbreviations

FDA: Food and Drug Administration
IC/BPS: interstitial cystitis/bladder pain syndrome
LDA: latent Dirichlet allocation
OR: odds ratio
PPS: pentosan polysulfate sodium
VADER: Valence Aware Dictionary for Sentiment Reasoning

Edited by A Mavragani; submitted 03.11.23; peer-reviewed by E Said-Hung, R Taneja; revised version received 04.10.24; accepted 05.10.24; published 17.01.25.

Please cite as:
Hswen Y, Qin Q, Smith P, Swierczynski A, Bauer S, Ladson E, Garrett AL, Brownstein CA
Sentiments of Individuals with Interstitial Cystitis/Bladder Pain Syndrome Toward Pentosan Polysulfate Sodium: Infodemiology Study
JMIR Form Res 2025;9:e54209
URL: https://formative.jmir.org/2025/1/e54209
doi:10.2196/54209

© Yulin Hswen, Qiuyuan Qin, Pressley Smith, Alison Swierczynski, Stuart Bauer, Erika Ladson, Amanda Leigh Garrett, Catherine A Brownstein. Originally published in JMIR Formative Research (https://formative.jmir.org), 17.1.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Using Health Information Resources for People With Cognitive Impairment (digiDEM Bayern): Registry-Based Cohort Study

Florian Weidinger[1], MSc; Nikolas Dietzel[1], PhD; Elmar Graessel[2], MD, PhD; Hans-Ulrich Prokosch[3], PhD; Peter Kolominsky-Rabas[1], MBA, MD, PhD

[1]Interdisciplinary Centre for Health Technology Assessment and Public Health, Friedrich-Alexander-Universität Erlangen-Nürnberg, Schwabachanlage 6, Erlangen, Germany

[2]Department of Psychiatry and Psychotherapy, Center for Health Services Research in Medicine, Universitätsklinikum Erlangen, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

[3]Department of Medical Informatics, Biometrics and Epidemiology, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

Corresponding Author:
Florian Weidinger, MSc
Interdisciplinary Centre for Health Technology Assessment and Public Health, Friedrich-Alexander-Universität Erlangen-Nürnberg, Schwabachanlage 6, Erlangen, Germany

Abstract

Background: Dementia is a growing global health challenge with significant economic and social implications. Underdiagnosis of dementia is prevalent due to a lack of knowledge and understanding among the general population. Enhancing dementia literacy through improved health information–seeking behavior is crucial for the self-determined management of the disease by those affected. Understanding the relationship between dementia literacy, health information–seeking behavior, and the use of various information sources among individuals with cognitive impairment is of high importance in this context.

Objective: The aim of this study was to analyze the relevance of different sources of health information from the perspective of people with cognitive impairment, while also evaluating differences based on age, gender, and disease progression.

Methods: This study is part of the ongoing project "Digital Dementia Registry Bavaria – digiDEM Bayern." The Digital Dementia Registry Bavaria is a multicenter, prospective, longitudinal register study in Bavaria, Germany. People with cognitive impairment rated several information sources by using Likert scales with the values unimportant (1) to very important (5). Data were analyzed descriptively, and multiple 2-sample, 2-tailed *t* tests were used to evaluate differences by cognitive status and gender and using multiple one-way ANOVA to evaluate differences by age group.

Results: Data of 924 people with cognitive impairment (531 with dementia, 393 with mild cognitive impairment) were evaluated. The most relevant health information sources were "Personal visit to a medical professional" (mean 3.9, SD 1.1) and "Family / Friends" (mean 3.9, SD 1.2). "Internet" was 1 of the 2 lowest-rated information sources by people with cognitive impairment (mean 1.6, SD 1.1), with nearly three-quarters (684/924, 74%) of the participants rating the source as unimportant. The age-specific analyses showed significant differences for the sources "Internet" ($F_{2,921}$=61.23; P<.001), "Courses / Lectures" ($F_{2,921}$=18.88; P<.001), and "Family / Friends" ($F_{2,921}$=6.27; P=.002) for the 3 defined age groups. There were several significant differences between people with mild cognitive impairment and dementia whereby the first group evaluated most sources higher, such as "Internet" (mean difference=0.6; t_{640}=7.52; P<.001). The only sources rated higher by the dementia group were "TV / Radio" and "Family / Friends," with none of them showing significant differences. Gender-specific analyses showed women with cognitive impairment valuing every evaluated source higher than men apart from "Internet" (mean difference=0.4; t_{685}=4.97; P<.001).

Conclusions: To enhance health and dementia literacy, the best way to communicate health information to people with cognitive impairment is through interpersonal contact with medical professionals and their friends and family. Slight changes in valuation should be considered as the medical condition progresses, along with variations by age and gender. In particular, the evaluation and use of the internet are dependent on these factors. Further research is needed to capture potential changes in the valuation of the internet as a health information source.

International Registered Report Identifier (IRRID): RR2-10.1136/bmjopen-2020-043473

(JMIR Form Res 2025;9:e54460) doi:10.2196/54460

KEYWORDS

dementia; mild cognitive impairment; cognitive impairment; information sources; health information; health information–seeking behavior; Digital Dementia Registry Bavaria; digiDEM

XSL•FO
RenderX

Introduction

Dementia is one of the most significant public health challenges of our time, and it will continue to be so in the future. To date, there are an estimated 57.4 million people with dementia worldwide, with estimates of 78 million people affected by 2030 and up to 152.8 million cases by 2050 [1,2]. In financial terms, these figures translate into an estimated global cost of dementia of US $1313.4 billion, which equates to an annual cost of US $23,796 per person with dementia [3]. Estimations also show that 75% of people with dementia worldwide live without a confirmed diagnosis. One of the reasons for this underdiagnosis is a lack of knowledge and understanding of dementia and its symptoms [1].

One way to address this problem of underdiagnosis of dementia is to improve health literacy and, therefore, dementia literacy [4]. However, levels of health literacy are low across Europe, particularly in Germany. According to the European Health Literacy Population Survey 2019 - 2021, 46% of the European population and 72% of the German population had limited health literacy in 2021. These figures were even more pronounced for vulnerable subpopulations such as people of advanced age or people with chronic diseases [5].

Nutbeam [6] defines health literacy as "the ability of individuals to gain access to, understand, and use information to promote and maintain good health." Dementia literacy is a subcategory and describes the knowledge for the recognition, management, and prevention of dementia [7]. There are multiple conceptual frameworks that describe the model of health literacy with most of them recognizing the process of using information as an essential element of health literacy. One example is the integrated model of health literacy by Sørensen et al [8], which is based on the 4 competencies of access, understanding, appraisal, and application. The first competency includes the ability to seek, find, and obtain health information. This implies that health information–seeking behavior also plays an important role in the process of how health literacy affects health outcomes [9].

Zimmerman et al [10] describe health information–seeking behavior as the purposeful behavior by an individual to find health information. The strategies for finding health information can take different forms, for example, through active seeking or more passive measures. An important aspect of health information–seeking behavior is which sources people use to obtain health information [10]. This implies a relationship between health literacy and the use of health information sources [9].

According to a recent survey conducted in 2021, 55% of Europeans aged 16 - 74 years have sought health-related information via the web [11]. The US Health Information National Trends Survey indicates a similar situation for the American general population, with 72.7% using the internet first for their most recent search for health information in 2019 [12]. However, health information–seeking behavior and preferences for information sources depend heavily on various individual aspects, such as gender and age, as well as situational aspects, such as having a chronic disease like dementia [13,14]. Women are generally more interested in health information than men and also use health information sources more frequently [15]. For chronic conditions, Oh and Cho [16] found that chronic patients were generally more likely to search for health information than healthy individuals. To date, no studies have investigated which sources of health information people with cognitive impairment use and how they value them.

Therefore, the following research questions (RQs) were asked: RQ1 (How do people with cognitive impairment evaluate different sources of health information?), RQ2 (Are there differences in the evaluation based on the age of the person affected?), RQ3 (Are there differences in the evaluation based on the gender of the person affected?), and RQ4 (Are there differences in the evaluation based on the progress of the medical condition of the person affected?).

Methods

Study Design

This study is part of the ongoing project "Digital Dementia Registry Bavaria – digiDEM Bayern." The Digital Dementia Registry Bavaria is a multicenter, prospective, longitudinal register study conducted in all administrative regions of Bavaria. The detailed methodology of the project is described elsewhere [17,18].

Study Population

Participants are people with mild cognitive impairment (MCI) and people with mild or moderate dementia living in Bavaria, who are in the following collectively referred to as people with cognitive impairment. MCI is a memory impairment that results in below the expected performance for the patient's age and level of education. People with MCI are not demented and generally maintain their independence in functional abilities of daily life [19,20]. To identify eligible participants, people have to undergo a screening based on the Mini-Mental State Examination (MMSE) and Montreal Cognitive Assessment (MoCA) prior to inclusion [21,22]. If available, a family caregiver is included with the person with cognitive impairment [17,18].

Recruitment and Data Collection

Participants are recruited by specially trained research partners in all of the 7 administrative regions in Bavaria, beginning in August 2020. Research partners are institutions that are specialized and have experience in the management and care of people with cognitive impairment and their family caregivers. Data collection is conducted through standardized face-to-face interviews using a web-based data entry system [17,18].

Ethical Considerations

The Digital Dementia Registry Bavaria is performed following the ethical standards of the Declaration of Helsinki, and it obtained ethical approval from the ethics committee of the Medical Faculty of the Friedrich-Alexander-Universität Erlangen-Nürnberg (application number: 253_20 B). Informed consent from the participants or their authorized representative is acquired before screening and study inclusion. The project collects and stores all personal data separately from the registry

XSL•FO
RenderX

data on different stand-alone systems to ensure data protection. All participants are pseudonymized. The data protection concept was approved by the local data protection supervisor of the Friedrich-Alexander-Universität Erlangen-Nürnberg and authorized by the Bavarian Data Protection Commissioner. Participation is voluntary, and participants are not compensated [17].

Measures

Sociodemographic data of the people with cognitive impairment were collected, such as age, gender, educational degree, family status, score in cognitive assessment, and whether they had a family caregiver.

The relevance of 8 different sources of health information was rated by people with cognitive impairment using Likert scales ranging from unimportant (1) to very important (5). Based on previous studies, the rated health information sources were "Internet," "TV / Radio," "Books / Brochures," "Courses / Lectures," "Newspaper / Journals," "Family / Friends," "Pharmacy," and "Personal visit to a medical professional" [23,24]. In order to further evaluate the target group's use of the internet, the participants were asked about the last time they had used the internet. Participants who had last used the internet 3 months ago or had never used it were also asked about their reasons for not using it.

Statistical Analysis

Subgroup analyses were performed for the cognitive status and gender of the people with cognitive impairment using multiple 2-sample, 2-tailed *t* tests and for the age of the participants using multiple one-way ANOVA to evaluate potential differences depending on the progress of the medical condition and gender- and age-specific differences. Based on the age distribution in the sample, the participants were divided into 3 age groups (<70 y, 70 - 79 y, and 80 y or older) for the analysis of age-specific differences. In order to minimize or avoid the probability of a type 1 error, we applied the Bonferroni correction for multiple testing. Accordingly, the adjusted significance level of $P<.002$ was set for all subgroup analyses (nonadjusted significance level: $P<.05$) [25]. Effect sizes (Cohen *d* or η^2) were determined [26]. Statistical analysis was performed using IBM SPSS Statistics (version 28; IBM Corp).

Results

Study Population

In the period between August 2020 and July 2023, in total, 924 people with cognitive impairment were included in the Digital Dementia Registry Bavaria. The following presented analyses refer to the baseline survey.

Table 1 shows the characteristics of the participants in total and for the 2 subgroups "MCI" and "dementia." The age of the study population ranged from 54 to 102 years and was 80.5 (SD 7.7) years on average. Around 59.4% (549/924) of the participants were female, with a mean MMSE score of 22.8 (SD 3.8) points and a mean MoCA result of 19.9 (SD 2.7) points. With 58.2% (538/924), most participants had a middle education, and 50.3% (465/924) were in a partnership. Around 64.1% (592/924) of the people with cognitive impairment regularly received help from a family caregiver at least once per week. Based on their MMSE score, 42.5% (393/924) of the study population was assigned to the MCI group and 57.5% (531/924) of the study population was assigned to the dementia group. There were significant differences between the 2 subgroups for their age ($P<.001$), their MMSE result ($P<.001$), their education ($P<.001$), their family status ($P=.003$), and whether they regularly received support from a family caregiver ($P<.001$).

XSL•FO
RenderX

Table . Description of this study population with the classification into the groups "MCI" (mild cognitive impairment) and "dementia" in a registry-based cohort study in Germany from 2020 to 2023.

		Total (N=924)	MCI (n=393; 42.5%)	Dementia (n=531; 57.5%)	*P* value
Age (years), mean (SD)		80.5 (7.7)	78.3 (7.8)	82.1 (7.2)	<.001
Gender, n (%)					.46
	Female	549 (59.4)	228 (58.0)	321 (60.5)	
	Male	375 (40.6)	165 (42.0)	210 (39.5)	
Family caregiver, n (%)					<.001
	No family caregiver	332 (35.9)	201 (51.1)	131 (24.7)	
	Has family caregiver	592 (64.1)	192 (48.9)	400 (75.3)	
Cognitive assessment, mean (SD)					
	MMSE[a]	22.8 (3.8)	26.5 (1.7)	20.1 (2.5)	<.001
	MoCA[b]	19.9 (2.7)	19.9 (2.7)	—[c]	—
Education, n (%)					<.001
	Lower education	197 (21.3)	62 (15.8)	135 (25.4)	
	Middle education	538 (58.2)	234 (59.5)	304 (57.3)	
	Higher education	189 (20.5)	97 (24.7)	92 (17.3)	
Family status, n (%)					.003
	No partnership	459 (49.7)	173 (44.0)	286 (53.9)	
	In partnership	465 (50.3)	220 (56.0)	245 (46.1)	

[a]MMSE: Mini-Mental State Examination.

[b]MoCA: Montreal Cognitive Assessment.

[c]Not applicable.

General Valuation of Health Information Sources

The most relevant health information sources for people with cognitive impairment were "Family / Friends" as well as "Personal visit to a medical professional" (eg, doctors or physiotherapists) (Table 2). With an average score of 3.9 (SD 1.1), 33.0% (305/924) of the people with cognitive impairment rated "Personal visit to a medical professional" as a very important source, and around three-quarters (694/924, 75.1%) rated it as at least "important." A similar result was found for "Family / Friends," with an average score of 3.9 (SD 1.2), and 39.6% (366/924) of the people with cognitive impairment rated them as a very important source of health information. Going by the highest average score, these 2 sources were followed by "TV / Radio," "Newspaper / Journals," "Pharmacy," and "Books / Brochures." The 2 lowest-rated information sources by people with cognitive impairment were "Internet" and "Courses / Lectures," with an average score of 1.6 (SD 1.1) and 1.6 (SD 1.0) on the Likert scale. Nearly three-quarters (684/924, 74.0%) of the participants rated "Internet" as an unimportant source, with a similar amount (680/924, 73.6%) stating that they had not used the internet in the last 3 months; the majority of those participants (569/680, 83.7%) stated that they had never used the internet before. The most frequently cited reasons for not using the internet were "generally no interest in the internet" (446/680, 65.6%) and "too complicated" (386/680, 56.8%) (Multimedia Appendix 1).

XSL•FO
RenderX

Table . General evaluation of health information sources by people with cognitive impairment using Likert scales ranging from unimportant (1) to very important (5) in a registry-based cohort study in Germany from 2020 to 2023.

	Likert scale ratings (% of participants who selected each rating)					Likert scale score, mean (SD)
	Unimportant	Less important	Partially	Important	Very important	
Internet	74.0	6.8	8.0	7.4	3.8	1.6 (1.1)
TV/Radio	13.7	9.3	20.9	38.5	17.5	3.4 (1.3)
Books/Brochures	41.0	19.9	18.1	14.0	7.0	2.3 (1.3)
Courses/Lectures	69.0	13.6	9.4	6.5	1.4	1.6 (1.0)
Newspaper/Journals	18.8	11.9	19.4	33.7	16.2	3.2 (1.4)
Family/Friends	8.0	7.3	13.2	31.9	39.6	3.9 (1.2)
Pharmacy	26.8	12.3	19.7	28.6	12.6	2.9 (1.4)
Personal visit to a medical professional	6.6	5.7	12.6	42.1	33.0	3.9 (1.1)

Valuation of Health Information Sources by Age

The age-specific analysis shows heterogeneous results for the 8 health information sources (Multimedia Appendix 2). The sources "Internet" ($F_{2,921}$=61.23; P<.001), "Courses / Lectures" ($F_{2,921}$=18.88; P<.001), and "Family / Friends" ($F_{2,921}$=6.27; P=.002) showed significant differences for the 3 defined age groups. People with cognitive impairment valued "Internet" and "Courses / Lectures" lower as age increased, from people under the age of 70 years ("Internet": mean 2.7, SD 1.5; "Courses / Lectures": mean 2.1, SD 1.3), to people aged 70 to 79 years ("Internet": mean 1.8, SD 1.3; "Courses / Lectures": mean 1.7, SD 1.0), and to people aged 80 years and older ("Internet": mean 1.3, SD 0.9; "Courses / Lectures": mean 1.4, SD 0.9). In comparison, the valuation of the information source "Family / Friends" grew with increasing age, from people younger than 70 years (mean 3.7, SD 1.2) and people between the ages of 70 and 79 years (mean 3.7, SD 1.3), to people aged 80 years and older (mean 4.0, SD 1.2). While age-specific differences in the information sources "Family / Friends" (η^2=0.01) and "Courses / Lectures" (η^2=0.04) showed small effect sizes, the effect size for the information source "Internet" (η^2=0.12) was medium.

Valuation of Health Information Sources by Gender

Concerning the gender-specific evaluation of information sources, the results show that apart from "Internet" and "Personal visit to a medical professional," women with cognitive impairment valued every explored source more than men with cognitive impairment on average (Multimedia Appendix 3). Two of the information sources showed significant gender-specific differences. As already described, men with cognitive impairment rated "Internet" 0.4 points higher on average than women with cognitive impairment (t_{685}=4.97; P<.001), with a small effect size (Cohen d=0.35). Female participants valued "Pharmacy" higher than male participants by 0.4 points on average (t_{772}=3.98; P<.001), with a small effect size (Cohen d=0.27). Another information source that was valued higher by women with cognitive impairment was the source "Family / Friends," with a mean difference of 0.2 (t_{756}=2.49; P=.013) and a small effect size (Cohen d=0.17).

Valuation of Health Information Sources by Cognition

Regarding the relevance of health information sources categorized by cognitive status, there were several significant differences between the MCI and dementia groups (Multimedia Appendix 4). People with MCI rated the sources "Internet," "Books / Brochures," "Courses / Lectures," "Newspapers / Journals," "Pharmacy," and "Personal visit to a medical professional" higher. Except for the sources "Books / Brochures" and "Newspapers / Journals," all of those group differences showed significant results. The source "Internet" showed the most considerable mean difference between both groups, with 0.6 points (t_{640}=7.52; P<.001) and a medium effect size (Cohen d=0.53). The only sources rated higher by the dementia group were "TV / Radio" and "Family / Friends," with none of them showing significant differences.

Discussion

Principal Results

To the best of our knowledge, this is the first study that describes the relevance of different sources of health information from the perspective of people with cognitive impairment.

Our analyses show that interpersonal contact with family and friends and that with medical professionals are the most important sources of health information for people with cognitive impairment. This interpersonal communication is followed closely by traditional media formats such as TV, radio, newspapers, and journals. People with cognitive impairment rated the internet and attending courses or lectures the lowest, with each being valued as less important or unimportant.

General Valuation of Health Information Sources

There are a number of possible reasons for the low relevance of the internet for people with cognitive impairment. The participants of our study mainly reported that they were generally not interested in the internet as a medium and that the

XSL•FO
RenderX

internet was too complicated for them. Other frequently cited reasons were that relatives take care of matters on the internet, no advantage is seen in using it, or traditional media are sufficient. These findings are partially in line with the literature, although it should be noted that the majority of the participants in our study had never used the internet before and therefore had no experience with the internet. Dixon et al [27] defined 4 main barriers that prevented people with cognitive impairment from seeking the internet for health information. These are the scarcity of relevant information and the inaccessibility, inaccuracy, and distrust of the information that was found [27]. More importantly, Dixon et al [28] see a chance of the internet becoming a more relevant source for people with cognitive impairment in the future, especially if their close relatives and health professionals support them in validating internet-based health information.

Our findings are in line with the large-scale survey by Weber et al [29] who assessed adults older than 60 years and reported high use of interpersonal communication with medical professionals, family, and friends as well as traditional mass media or brochures and infrequent use of the internet for seeking health information.

Stehr et al [30] came to similar results and showed that interpersonal contact and traditional mass media are the most used health information sources for people aged 65 years or older while they also confirmed that the internet is used little or not at all. Further, they noticed that the use and valuation of communication with medical professionals increases for people with chronic diseases, which include cognitive impairment [30].

The above results can be explained by, among other things, the trust that older adults or people with cognitive impairment place in the various sources. Baumann et al [31] reported that medical professionals, in this case doctors, are the most trusted health information source followed by family and friends, while the internet is one of the least trusted sources. These findings implicate that the trustiness of the various sources is a significant factor in the valuation by people with cognitive impairment.

Pertl et al [32] also stated that people with MCI show problems with understanding numerical health information. These problems are likely to increase for people with dementia, as they found a statistically significant correlation with global cognitive status [32]. These findings suggest that people with MCI and dementia might require explanation through communicating with medical professionals and their friends or family to understand their health issues.

Valuation of Health Information Sources by Age

The age-specific results show that the relevance of the information sources “Internet” and “Courses / Lectures” decreased significantly with age. Contrary to this, the evaluation of the source “Family / Friends” enlarged with increasing age. No significant differences were found for the 5 other sources based on participants’ age. The difference in the information source “Internet” is in particular noteworthy, with a medium effect size and high differences in the mean values between the 3 age groups.

Our age-specific findings regarding the internet align with the results of the “D80+ report” by Reissmann et al [33], which found low internet use among people older than 80 years in Germany, at 37%. Weber et al [29] also found that age is a decisive factor influencing internet use as the tendency of infrequent use of the internet for seeking health information became more pronounced with increasing age.

Reasons for these observations may include the often limited digital health literacy or generally limited digital media literacy and internet skills of the older population [34]. These limitations become even more pronounced with increasing age, especially for people with cognitive impairment [34,35]. The high average age of 80.5 years in this sample could therefore also explain the generally low rating of the internet as a source of health information.

Valuation of Health Information Sources by Gender

Our gender-specific analyses show that women and men with cognitive impairment differ in assessing the relevance of health information sources. Except for the internet, all our evaluated sources were more important to women with cognitive impairment, though these differences between men and women were small. The sources “Pharmacy” and “Family / Friends” showed the most significant differences.

These gender-specific findings are in line with current scientific research. For the general population, Manierre [15] discovered that women seek health information more often than men. In the group of adults aged 65 years and older, Stehr et al [30] found higher use of most health information sources by women. They also noted the most remarkable differences between the sources “Pharmacy” and “Internet,” with the former being much more used by women and the latter by men [30]. The same applies to the results of the KomPaS study by Horch et al [36], who found similar differences in the use of these 2 sources.

Valuation of Health Information Sources by Cognition

When categorized by cognitive status, our results show that people with MCI rate most of the health information sources evaluated more highly than people with dementia. People with dementia only value contact with family and friends as well as watching TV or listening to the radio more.

It is noticeable that the mean values for health information sources that require some activity to obtain information, for example, leaving home to see a doctor, are all higher for the group of people with MCI. The same applies to the mean for sources of information that can be used passively, such as watching television or receiving advice from a friend, which is higher for people with dementia. This could be explained by the progressive social withdrawal of people with cognitive impairment as the medical condition progresses [37].

Dixon et al [28] also found similar results with their research on the changing strategies of people with cognitive impairment seeking health information. They reported that most people with cognitive impairment apply an active search technique after receiving their diagnosis and then transition to a monitoring strategy for ongoing seeking of health information. With the progression of their medical condition, they transition to a proxy

XSL•FO
RenderX

information search, as they have problems accessing the health information needed when just actively searching by themselves. With further progression, the authors even describe a strategy of information avoidance and selective exposure to health information by people with cognitive impairment [28]. These changes in information behavior could explain the distribution of different sources in the 2 cognitive groups. While people with MCI are likely to use one of the first, more active search strategies described, people with dementia are more likely to use one of the latter strategies due to their disease progression.

Limitations

This study has limitations. First, our sample was a convenience sample with a specific target group of people with cognitive impairment in Bavaria. We cannot exclude that we may have a selection bias as a result of the different referral patterns in the different regions of Bavaria. Although our research partners were institutions specialized in managing and caring for people with cognitive impairment in all of the 7 administrative regions in Bavaria, this study is not representative of the total population. This limits the comparability and transferability of the results to the entirety of people with cognitive impairment.

Furthermore, the results of the group differences mostly show small effect sizes, except for the group differences between people with MCI and dementia and the 3 age groups for the information source “Internet.” This limits the power and generalizability of the results while highlighting the need for further research with even larger case numbers.

Conclusions

To enhance health and dementia literacy in people with cognitive impairment, our results suggest that interpersonal contact with medical professionals and their friends and family is the best way to communicate health information to people with cognitive impairment. Another way is to disseminate information through traditional mass media. As the disease progresses, the importance of family and friends as sources of information increases, while the relevance of other sources decreases. Women with cognitive impairment are generally easier to reach and more interested in health information. At the same time, men and younger people with cognitive impairment also use the internet to some extent and are open to receiving information via this medium.

While the internet was of little importance as a source of health information to most of our study population of people with cognitive impairment older than 70 years, this may change in the future, as recent surveys from the European Union and the United States indicate [11,12]. Due to the continuing improvement in digital literacy in older adults, the internet might become more valuable as a source of health information. This implicates a need for further research in the future to capture these potential changes.

Acknowledgments

The project "Digital Dementia Registry Bavaria –digiDEM Bayern" is funded by the Bavarian State Ministry of Health, Care and Prevention as part of the funding initiative "BAYERN DIGITAL II" (grant number: G42d-G8300-2017/1606-83). The present work was performed by FW in (partial) fulfillment of the requirements for obtaining the degree “Dr. rer. biol. hum.” at the Medical Faculty of the Friedrich-Alexander-Universität Erlangen-Nürnberg. No generative artificial intelligence tools were used in the preparation and writing of the manuscript.

Data Availability

The datasets used and analyzed during this study are not publicly available since this is an ongoing research project and data will be used in the course of future analysis. Deidentified or anonymized data are available from the Bavarian State Ministry of Health, Care and Prevention upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Frequency of reasons given for not using the internet by people with cognitive impairment in a registry-based cohort study in Germany from 2020 to 2023.
[PNG File, 35 KB - formative_v9i1e54460_app1.png]

Multimedia Appendix 2
Evaluation of health information sources by people with cognitive impairment, with differentiation by age, using Likert scales ranging from unimportant (1) to very important (5) in a registry-based cohort study in Germany from 2020 to 2023.
[PNG File, 83 KB - formative_v9i1e54460_app2.png]

Multimedia Appendix 3
Evaluation of health information sources by people with cognitive impairment, with differentiation by gender, using Likert scales ranging from unimportant (1) to very important (5) in a registry-based cohort study in Germany from 2020 to 2023.

XSL•FO
RenderX

[PNG File, 74 KB - formative_v9i1e54460_app3.png]

Multimedia Appendix 4
Evaluation of health information sources by people with cognitive impairment, with differentiation by their cognitive status, using Likert scales ranging from unimportant (1) to very important (5) in a registry-based cohort study in Germany from 2020 to 2023.
[PNG File, 75 KB - formative_v9i1e54460_app4.png]

References

1. Gauthier S, Rosa-Neto P, Morais JA, Webster C. World Alzheimer Report 2021: Journey Through the Diagnosis of Dementia: Alzheimer's Disease International; 2021.
2. Nichols E, Steinmetz JD, Vollset SE, et al. Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the global burden of disease study 2019. Lancet Public Health 2022 Feb;7(2):e105-e125. [doi: 10.1016/S2468-2667(21)00249-8]
3. Wimo A, Seeher K, Cataldi R, et al. The worldwide costs of dementia in 2019. Alzheimers Dement 2023 Jul;19(7):2865-2873. [doi: 10.1002/alz.12901] [Medline: 36617519]
4. Aihara Y, Maeda K. Dementia literacy and willingness to dementia screening. Int J Environ Res Public Health 2020 Nov 4;17(21):33158014. [doi: 10.3390/ijerph17218134] [Medline: 33158014]
5. The HLS19 Consortium of the WHO Action Network M-POHL. International report on the methodology, results, and recommendations of the european health literacy population survey 2019-2021 (HLS19) of M-POHL. : Austrian National Public Health Institute: Vienna; 2021.
6. Nutbeam D. Health literacy as a public health goal: a challenge for contemporary health education and communication strategies into the 21st century. Health Promot Int 2000 Sep 1;15(3):259-267. [doi: 10.1093/heapro/15.3.259]
7. Low LF, Anstey KJ. Dementia literacy: recognition and beliefs on dementia of the Australian public. Alzheimers Dement 2009 Jan;5(1):43-49. [doi: 10.1016/j.jalz.2008.03.011] [Medline: 19118808]
8. Sørensen K, Van den Broucke S, Fullam J, et al. Health literacy and public health: a systematic review and integration of definitions and models. BMC Public Health 2012 Jan 25;12:80. [doi: 10.1186/1471-2458-12-80] [Medline: 22276600]
9. Chen X, Hay JL, Waters EA, et al. Health literacy and use and trust in health information. J Health Commun 2018;23(8):724-734. [doi: 10.1080/10810730.2018.1511658] [Medline: 30160641]
10. Zimmerman MS, Shaw G Jr. Health information seeking behaviour: a concept analysis. Health Info Libr J 2020 Sep;37(3):173-191. [doi: 10.1111/hir.12287] [Medline: 32052549]
11. EU citizens: over half seek health information online. Eurostat. 2022. URL: https://ec.europa.eu/eurostat/en/web/products-eurostat-news/-/edn-20220406-1 [accessed 2023-08-22]
12. The most recent time you looked for information about health or medical topics, where did you go first?- HINTS 5 cycle 3. National Cancer Institute Health Information National Trends Survey. 2019. URL: https://hints.cancer.gov/view-questions/question-detail.aspx?PK_Cycle=12&qid=688 [accessed 2023-08-22]
13. Rossmann C, Lampert C, Stehr P, Grimm M. Nutzung Und Verbreitung von Gesundheitsinformationen: Ein Literaturüberblick Zu Theoretischen Ansätzen Und Empirischen Befunden. Use and Dissemination of Health Information: A Literature Review of Theoretical Approaches and Empirical Findings: Bertelsmann Stiftung; 2018. [doi: 10.11586/2017051]
14. Chasiotis A, Wedderhoff O, Rosman T, Mayer AK. Why do we want health information? the goals associated with health information seeking (GAINS) questionnaire. Psychol Health 2020 Mar;35(3):255-274. [doi: 10.1080/08870446.2019.1644336] [Medline: 31340675]
15. Manierre MJ. Gaps in knowledge: tracking and explaining gender differences in health information seeking. Soc Sci Med 2015 Mar;128(151-8):151-158. [doi: 10.1016/j.socscimed.2015.01.028] [Medline: 25618604]
16. Oh YS, Cho Y. Examining the relationships between resources and online health information seeking among patients with chronic diseases and healthy people. Soc Work Health Care 2015;54(2):83-100. [doi: 10.1080/00981389.2014.987940] [Medline: 25674723]
17. Dietzel N, Kürten L, Karrer L, et al. Digital Dementia Registry Bavaria-digiDEM Bayern: study protocol for a multicentre, prospective, longitudinal register study. BMJ Open 2021 Feb 8;11(2):e043473. [doi: 10.1136/bmjopen-2020-043473] [Medline: 33558357]
18. Reichold M, Dietzel N, Chmelirsch C, Kolominsky-Rabas PL, Graessel E, Prokosch HU. Designing and implementing an IT architecture for a digital multicenter dementia registry: digiDEM Bayern. Appl Clin Inform 2021 May;12(3):551-563. [doi: 10.1055/s-0041-1731286] [Medline: 34134149]
19. Albert MS, DeKosky ST, Dickson D, et al. The diagnosis of mild cognitive impairment due to alzheimer's disease: recommendations from the National Institute on Aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease. Alzheimers Dement 2011 May;7(3):270-279. [doi: 10.1016/j.jalz.2011.03.008] [Medline: 21514249]
20. Petersen RC, Smith GE, Waring SC, Ivnik RJ, Tangalos EG, Kokmen E. Mild cognitive impairment: clinical characterization and outcome. Arch Neurol 1999 Mar;56(3):303-308. [doi: 10.1001/archneur.56.3.303] [Medline: 10190820]

XSL•FO
RenderX

21. Folstein MF, Folstein SE, McHugh PR. “Mini-mental state”. a practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res 1975 Nov;12(3):189-198. [doi: 10.1016/0022-3956(75)90026-6] [Medline: 1202204]
22. Nasreddine ZS, Phillips NA, Bédirian V, et al. The Montreal cognitive assessment, MoCA: a brief screening tool for mild cognitive impairment. J Am Geriatr Soc 2005 Apr;53(4):695-699. [doi: 10.1111/j.1532-5415.2005.53221.x] [Medline: 15817019]
23. Gräßel E, Bleich S, Meyer-Wegener K, Schmid U, Kornhuber J, Prokosch HU. Das Internet als Informationsquelle für pflegende Angehörige eines Demenzpatienten. The internet as an information source for informal caregivers of dementia patients. Psychiat Prax 2009 Apr 3;36:115-118. [doi: 10.1055/s-2008-1067550]
24. Lausen B, Potapov S, Prokosch HU. Health-related use of the internet in Germany 2007. GMS Med Inform Biometr Epidemiol 2008 Sep;4(2).
25. Armstrong RA. When to use the Bonferroni correction. Ophthalmic Physiol Opt 2014 Sep;34(5):502-508. [doi: 10.1111/opo.12131] [Medline: 24697967]
26. Cohen J. Statistical power analysis. Curr Dir Psychol Sci 1992 Jun;1(3):98-101. [doi: 10.1111/1467-8721.ep10768783]
27. Dixon E, Anderson J, Blackwelder D, Radnofsky M, Lazar A. Barriers to online dementia information and mitigation. Proc SIGCHI Conf Hum Factor Comput Syst 2022 Apr;2022:513. [doi: 10.1145/3491102.3517554] [Medline: 35789135]
28. Dixon E, Anderson J, Blackwelder DC, Radnofsky ML, Lazar A. The human need for equilibrium: qualitative study on the ingenuity, technical competency, and changing strategies of people with dementia seeking health information. J Med Internet Res 2022 Aug 11;24(8):e35072. [doi: 10.2196/35072] [Medline: 35969426]
29. Weber W, Reinhardt A, Rossmann C. Lifestyle segmentation to explain the online health information-seeking behavior of older adults: representative telephone survey. J Med Internet Res 2020 Jun 12;22(6):e15099. [doi: 10.2196/15099] [Medline: 32530433]
30. Stehr P, Weber W, Rossmann C. Gesundheitsinformationsverhalten 65+: Erreichbarkeit älterer Zielgruppen. Health seeking behavior 65+: reaching older target groups. Bundesgesundheitsbl 2021 Jan;64(1):37-44. [doi: 10.1007/s00103-020-03253-1]
31. Baumann E, Czerwinski F, Rosset M, Seelig M, Suhr R. Wie informieren sich die Menschen in Deutschland zum Thema Gesundheit? erkenntnisse aus der ersten Welle von HINTS Germany. How do people in Germany seek health information? insights from the first wave of HINTS Germany. Bundesgesundheitsbl 2020 Sep;63(9):1151-1160. [doi: 10.1007/s00103-020-03192-x]
32. Pertl MT, Benke T, Zamarian L, et al. Do patients with mild cognitive impairment understand numerical health information? J Alzheimers Dis 2014;40(3):531-540. [doi: 10.3233/JAD-131895] [Medline: 24473188]
33. Reissmann M, Oswald V, Zank S, Tesch-Römer C, Cologne Center for Ethics, Rights, Economics, and Social Sciences of Health (ceres). Digitale Teilhabe in der Hochaltrigkeit (D80+ Kurzberichte, 6). Digital participation in very old age (D80+ abridged reports, 6). Social Science Open Access Repository. 2022. URL: https://nbn-resolving.org/urn:nbn:de:0168-ssoar-78429-7 [accessed 2024-10-07]
34. Hargittai E, Piper AM, Morris MR. From internet access to internet skills: digital inequality among older adults. Univ Access Inf Soc 2019 Nov;18(4):881-890. [doi: 10.1007/s10209-018-0617-5]
35. Talbot CV, Briggs P. The use of digital technologies by people with mild-to-moderate dementia during the COVID-19 pandemic: a positive technology perspective. Dementia (London) 2022 May;21(4):1363-1380. [doi: 10.1177/14713012221079477] [Medline: 35333111]
36. Horch K, Jordan S. Sachbericht zur studie "Kommunikation und Information im Gesundheitswesen aus Sicht der Bevölkerung. patientensicherheit und informierte Entscheidung" (KomPaS). Report on the study "Communication and information in the healthcare system from the perspective of the population. patient safety and informed decision making" (KomPaS). Bundesministerium für Gesundheit. 2019. URL: https://www.bundesgesundheitsministerium.de/fileadmin/Dateien/5_Publikationen/Praevention/abschlussbericht/2019-08-01_Sachbericht_KomPaS.pdf [accessed 2023-08-22]
37. Porcelli S, Van Der Wee N, van der Werff S, et al. Social brain, social dysfunction and social withdrawal. Neurosci Biobehav Rev 2019 Feb;97:10-33. [doi: 10.1016/j.neubiorev.2018.09.012] [Medline: 30244163]

Abbreviations

MCI: mild cognitive impairment
MMSE: Mini-Mental State Examination
MoCA: Montreal Cognitive Assessment
RQ: research question

XSL•FO
RenderX

Edited by A Mavragani; submitted 10.11.23; peer-reviewed by A Reinhardt, N Döring, T Brkovic; revised version received 07.10.24; accepted 15.10.24; published 15.01.25.

Please cite as:
Weidinger F, Dietzel N, Graessel E, Prokosch HU, Kolominsky-Rabas P
Using Health Information Resources for People With Cognitive Impairment (digiDEM Bayern): Registry-Based Cohort Study
JMIR Form Res 2025;9:e54460
URL: https://formative.jmir.org/2025/1/e54460
doi:10.2196/54460

© Florian Weidinger, Nikolas Dietzel, Elmar Graessel, Hans-Ulrich Prokosch, Peter Kolominsky-Rabas. Originally published in JMIR Formative Research (https://formative.jmir.org), 15.1.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Mobile Phone App to Promote Lifestyle Change in People at Risk of Type 2 Diabetes: Feasibility 3-Arm Randomized Controlled Trial

Gyri Skoglund[1], MSc; Gunvor Hilde[1], PhD; Pernille Lunde[1], PhD; Venessa Vera Cruz Naceno[1], PT; Cecilie Fromholt Olsen[2], PhD; Birgitta Blakstad Nilsson[3], PhD

[1]Department of Rehabilitation Science and Health Technology, Faculty of Health Sciences, Oslo Metropolitan University, PB 4, St. Olavs plass, Oslo, Norway

[2]Section for Allied Health Professionals, Department of Neurology, Oslo University Hospital, Oslo, Norway

[3]Section for Physiotherapy, Division of Medicine, Oslo University Hospital, Oslo, Norway

Corresponding Author:
Gyri Skoglund, MSc
Department of Rehabilitation Science and Health Technology, Faculty of Health Sciences, Oslo Metropolitan University, PB 4, St. Olavs plass, Oslo, Norway

Abstract

Background: The use of mobile health interventions, such as apps, are proposed to meet the challenges faced by preventive health care services due to the increasing prevalence of type 2 diabetes (T2D). Thus, we developed and conducted initial feasibility testing of the Plunde app for promoting and monitoring individual goals related to lifestyle change for people at risk of T2D.

Objective: The primary aim of this study was to assess the feasibility of an app for promoting lifestyle change in people at risk of T2D. The secondary aim was to assess recruitment rate, resource requirements, and change in potential outcomes for a full scale randomized controlled trial (RCT) study .

Methods: A 3-arm feasibility RCT lasting 12 weeks was designed. Participants were recruited from 9 general practitioners in Norway. Eligible participants were randomized to either (1) app follow-up; (2) app follow-up and referral to care as usual in Healthy Life Centers; or (3) referral to care as usual in a Healthy Life Center, only. The primary outcome was feasibility and was measured by app adherence (actual usage of the app), the System Usability Scale, and app motivation score gained from a questionnaire designed for this study. Criteria for success were preset based on these measures. Secondary outcomes included recruitment rate, resource requirements, and potential primary outcomes of a full-scale RCT. This included change in body weight, waist circumference, and self-evaluated functional health status, assessed with the Dartmouth Primary Care Cooperative Research Network/World Organization of Family Doctors (COOP/WONCA) functional health assessment chart.

Results: Within 8 months, 9 general practitioners recruited a total of 54 participants, of which 45 were eligble for participation in the study. Mean age was 61 (SD 13) years and 53% (n=24) were female. App adherence was 86%, the mean System Usability Scale score was 87.3 (SD 11.9), and the mean app motivation score was 74.8 (SD 30.3). Throughout the intervention period, health care professionals spent on average 3.0 (SD 1.0) minutes per participant per week providing follow-up. Statistically significant reduction in body weight and waist circumference was shown in group 1 and 3.

Conclusions: Based on the preset criteria for success, the Plunde app is feasible in providing support for lifestyle change. The Plunde app had excellent user satisfaction. The amount of time spent on monitoring and promoting lifestyle change through the app was low; however, the recruitment was slow. Results from this study will guide the development of further research within this field.

Trial Registration: ClinicalTrials.gov NCT06117098; https://clinicaltrials.gov/study/NCT06117098

(JMIR Form Res 2025;9:e63737) doi:10.2196/63737

KEYWORDS

prevention; risk of type 2 diabetes; lifestyle change; feasibility; app adherence; mobile health; mHealth; mobile phone

Introduction

There is an urgent need to prevent and combat the global epidemic of type 2 diabetes (T2D). The number of adults with diabetes is expected to reach 67 million by 2030 and 69 million by 2045 [1]. Prediabetes, an intermediate stage of glucose dysregulation that may precede T2D, affected approximately 720 million individuals worldwide in 2021 and is estimated to

XSL•FO
RenderX

affect estimated 1 billion people by 2045 [1,2]. Prediabetes is diagnosed when blood glucose levels are above the normal range but not high enough to be classified as diabetes [2].

Healthy diet, regular physical activity, and maintaining a normal body weight are established cornerstones of prevention or delayed onset of T2D [3-5]. Hence, lifestyle modification is the preferred initial approach when having confirmed prediabetes or being at high risk of T2D [2]. Siginficant risk reduction in delaying or preventing progression into T2D has been found in people receiving lifestyle interventions [3,6]. A post-hoc analysis of randomized, controlled, multicentre Prediabetes Lifestyle Intervention Study showed that a weight loss of 5 percent or more, could decrease the risk of T2D with 72%, and that full remission of T2D should be the target aim [7]. A larger waist circumference (WC), independent of overall adiposity, has been found to be strongly and linearly associated with a higher risk of T2D [8]. Consequently, both body weight and WC may be important predictors of the development for T2D.

Healthy Life Centers (HLCs) is a part of the public health care service in the municipalities in Norway, providing in person lifestyle modification programs for people with, or in high risk of disease [9]. Considering the burden of noncommunicable diseases (NCDs), together with an aging population, the health care system faces fundamental challenges in delivering optimal care [10]. Thus, the primary health care services are at a threshold limit.

The use of mobile health (mHealth) technologies has been proposed to meet challenges related to delivery of sustainable prevention strategies [11]. The beneficial effects of mHealth interventions regarding lifestyle factors have been demonstrated in patients with NCDs in several systematic reviews [12-14]. However, few studies have examined the effect of using an app to deliver support and guidance for risk reduction in patients at risk of T2D.

In 2022, we investigated barriers and facilitators for lifestyle change in people at risk of T2D in a qualitative meta-synthesis [15]. We found that central barriers were time, cost, and availability [15]. Additionally individual tailoring of lifestyle interventions was highlighted to facilitate the process of making healthy changes [15]. The advantages of digital health solutions lie, in its potential to overcome the barriers for lifestyle change and at the same time enabling personalized, empowering, and person-centered health interventions [16]. Therefore, based on former research, user orientation- and involvement, we developed the Plunde app (People Living UNDEr change) to monitor and guide individuals in initiating and maintaining lifestyle change and healthy behavior [17].

Randomized controlled trials (RCTs) are needed to assess potential effect of lifestyle interventions delivered by an app. Before commencing such a trial, investigating the feasibility of using an app to promote lifestyle change is warranted. Therefore, the primary aim of this study was to assess if the Plunde app was feasible to use for promoting lifestyle change in people at risk of T2D. Preset criteria for success was based on app adherence (actual usage of the app), the System Usability Scale (SUS) and app motivation score which was gained from a questionnaire designed for this feasibility study. Secondary, we aimed to assess recruitment rate, resource requirements and change in body weight, WC, and self-evaluated functional health status.

Methods

Study Design

This 3-arm feasibility RCT, had an intervention period lasting for 12 weeks. The participants were randomized to receive: lifestyle intervention delivered by the Plunde app (group 1) or referral to group-based lifestyle intervention at the local HLCs in addition to follow-up with the Plunde app (group 2) or to usual care which was referral to group-based lifestyle intervention at the local HLCs (group 3). A computer-generated block randomization scheme was used to allocate the participants to 1 of the 3 groups via concealed envelopes after baseline assessment. The block sizes varied by 3 and 6.

Reporting follows the Consilidated Standard of Reporting Trials (CONSORT) 2010 statement: extension to randomized pilot and feasibility trials [18]. Additionally, principles from the Consilidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth (CONSORT-eHEALTH) statement [19] were applied (Checklist 1).

Setting and Recruitment

Participants were recruited from general practitioner (GP) clinics in the eastern part of Norway from May 2023 and throughout December 2023. Initially the patients were recruited from 2 GP clinics. From September 2023 we included additionally 3 GPs clinics. Prior to recruitment, information meetings at the clinic sites were held. Eligible participants were asked by their GP regarding willingness to participate in the study, and with their consent they were contacted by a researcher (BBN or VVCN). The baseline assessments and follow-ups were conducted at 3 collaborating HLCs by 2 of the project members (BBN and VVCN).

Participants

Eligible participants were women and men aged 18 years or older, who were assessed by their GP to be at high risk of developing T2D. The criteria for being at high risk was HbA_{1c} within the threshold level of 42-48 mL/mmol. In addition, participants had to own and use an Android or iOS smartphone and be able to read and understand Norwegian or English.

By September 2023 only 25 participants were included; therefore, we invited 3 more GP clinics to recruit participants to the project. To increase the recruitment rate, we also widened the criteria for being at high risk of T2D. Hypertension, hypercholesteremia, and being obese were set as criteria in addition to, or instead of, HbA_{1c} levels between 42 mL/mmol and 48 mL/mmol.

Baseline Assessment

At baseline we collected demographic and descriptive data within different lifestyle domains from the standardized “HLC Startup” questionnaire developed by the Norwegian Health Directorate [20]. The demographic data collected included

XSL•FO
RenderX

gender, age, educational level, and occupational status. The descriptive data included physical activity level, dietary habits, tobacco, alcohol, sleep habit, and social support. This questionnaire was primarily used to guide and structure the plan for lifestyle change and setting of goals and tasks. Anthropometric data (body weight, WC, and height) were collected at baseline and at follow-up by the same researcher.

The Plunde App

Participants randomized to the app intervention groups, group 1 and 2, received access to Plunde and guiding in how to use it after baseline assessment. Plunde was developed to promote initiation of behavioral change and adherence to healthy behavior [17]. The app is highly individualized as it is based on the patient's own goals to initiate or maintain a lifestyle change [17]. Plunde permits the user to set personal goals (Figure 1B) with tasks and accompanying reminders. Goals and tasks decided during the baseline consultation were added to the app. The goals set by the participants were related to the different areas of lifestyle they wanted to work on during the intervention period. The main lifestyle domains presented (through the questionnaire) were dietary habits, physical activity level and exercise, sleep, mental well-being, alcohol, and smoking habits. Each participant set different tasks to their goals, and 1 goal could have 1 or more related tasks. To each of the tasks, the participant set a reminder and decided when and how often the reminders of their tasks should appear. Additionally, they could choose the wording of their reminders. Plunde provided push notifications related to the individual tasks, and the participants replied "completed" or "not completed."

Figure 1. Screenshots showing the user interface.

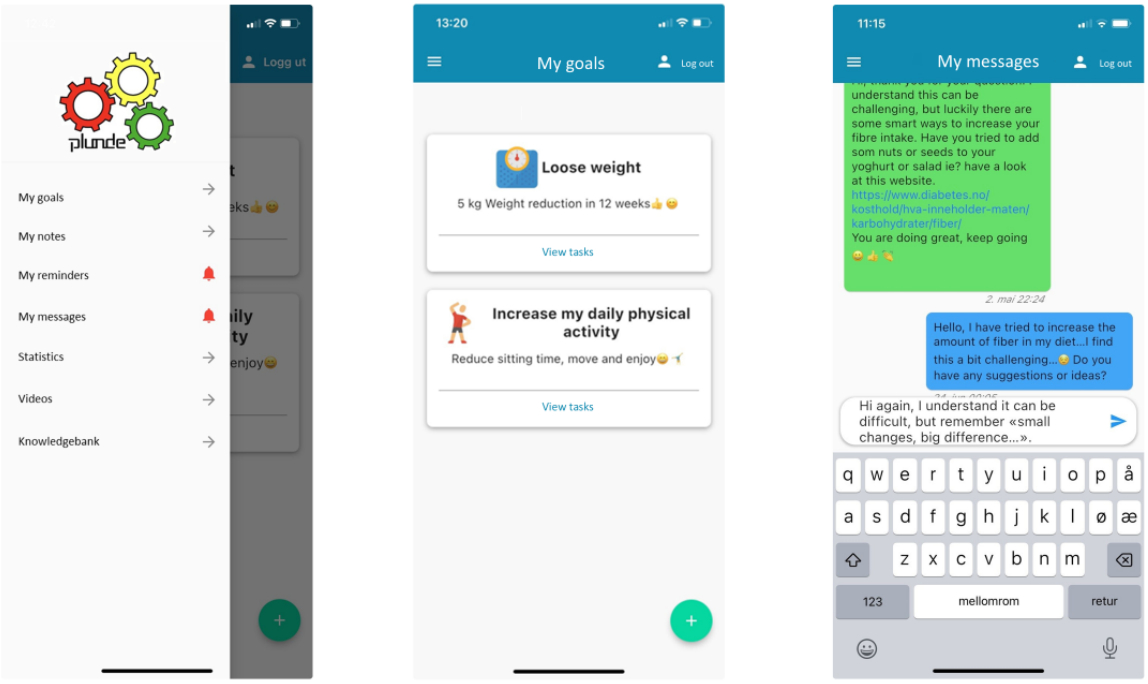

Other features and functions in Plunde include a note function, messages, videos, and a knowledge bank (Figure 1A). The note function allowed the participant to write relevant notes or comments on their goal achievement process, or for example writing an exercise diary. The message function allowed the participants to chat with their supervisor during the follow-up period (Figure 1C). Relevant videos based on the participants' goals were added under the video function. In the knowledge bank, the supervisor could add written educational content based on the participants' goals and requests.

A supervisor had access to an administrator interface and monitored the goals, tasks, and notes of each participant throughout the intervention period. During the follow-up period the supervisor could send weekly encouraging messages to support their process of lifestyle change. If the participants asked any questions in the message function, these were answered within 2 working days. The 2 supervisors in this study were both physiotherapists, 1 being a researcher with comprehensive experience in cardiac rehabilitation and cardiac research and the other with experience from HLCs and community health services. The same supervisor who conducted the baseline consultation monitored and provided feedback throughout the study period.

Referral to Healthy Lifestyle Center

Participants randomized to group 2 and 3 were referred to the HLC by the respective supervisor immediately after the baseline consultation, with an expected enrollment to the HLC program within 2 weeks. The lifestyle modification program at the HLC included individual and group-based counseling and courses for increased physical activity, healthy diet, improved sleep, and tobacco cessation. The participants were encouraged to attain at least 1 or 2 exercise sessions per week at the HLC. The participants in all the 3 study arms, received counseling based on elements from motivational interviewing, which is a client-centered method of intervention focused on enhancing

XSL•FO
RenderX

intrinsic motivation and behavioral change [21]. Goal setting and planning were crucial parts of this conversation, based on the needs, values, and perspectives of the participants.

Primary Outcome

The primary outcome was feasibility measured as app adherence, user satisfaction, and app motivation. Data related to feasibility of Plunde were collected from patients in groups 1 and 2.

Adherence to the app was registered in terms of actual use, which we defined to be the percentage of reminders answered per week throughout the 12-week intervention period. These data were collected from the administrator interface.

The user satisfaction with the app was assessed with the SUS. The SUS is a technology independent, 10-item questionnaire with a score between 0 and 100, where 0 represents low usability and 100 represent high usability [22]. The elements of the SUS scale cover aspects of usability such as ease of use, integration of functions, learning curve, satisfaction, user confidence, and complexity [22,23].

App motivation was measured with a questionnaire designed for this study. This questionnaire consisted of 17 questions—13 questions with Likert scale (0 to 100) options, 3 multiple-choice questions, and 1 open-ended question (Multimedia Appendix 1). The main questions defining app motivation score was based on 2 of the questions; to what extent Plunde was experienced as motivating and if the individual feedback provided from the supervisor was experienced as motivating. In addition, we asked for how long the participants would have liked to continue using the app and, in this period, how often they would like to receive feedback from their supervisors. The criteria for feasibility success of the app that were preset were (1) at least 80% of the participants reminders were answered per week (app adherence), (2) mean SUS score ≥65, and (3) app motivation score ≥75.

Secondary Outcomes

To inform a potential full-scale RCT, data on recruitment rate and resource requirements were collected throughout the study period. We calculated the recruitment rate as the number of participants successfully recruited for the study divided by the total amount of time required for recruitment. Resource requirements in groups having the app intervention was assessed by the supervisors who logged all time spent on monitoring and providing feedback to participants. Additionally, potential primary outcomes in a future scale RCT were evaluated. This included body weight, WC, and self-evaluated functional health status. These outcomes were assessed to get experience with the measurements and to examine a potential change.

Body weight was measured in kilogram using a digital scale (Body Composition Analyzer, BC-418, ADE MeWa GMbh, Schwerin and Beurer GS400 Signature Line weight). The evaluation was carried out in the “standard mode,” introducing the participant’s age, sex, and height. We strived to use the same equipment for measuring body weight at baseline and at follow-up. WC was measured with a measuring tape (in cm; Seca 201) at the end of several consecutive natural breaths, midpoint being between the top of the iliac crest and the lower margin of the last palpable rib in the mid axillary line [24]. The cut off points for abdominal obesity and increased risk of T2D is WC ≥80 in women and WC ≥90 in men [25,26].

Self-evaluated functional health status was assessed with the Dartmouth Primary Care Cooperative Research Network/World Organization of Family Doctors (COOP/WONCA) chart [27]. The questionnaire is a generic questionnaire comprising 6 domains—physical fitness, feelings (mental well-being), daily activities, social activities, change in health, and overall health. Each dimension is illustrated pictorially, numerically, and in writing inquiring about the patient’s status during the past 2 weeks. The response categories were scored from 1 to 5, where higher score indicate worse health status.

Statistical Analysis

Data were analysed using SPSS Statistics for Windows (version 29.0.0; IBM Corp). Descriptive statistics were used to assess feasibility measures. Descriptive statistics are reported in mean and SD or median and range for continuous variables and in number and percentages for categorical variables. To examine a potential change throughout the intervention period for body weight, WC, and self-evaluated functional health status, we performed within-group analyses on mean change from baseline to 12-weeks follow-up. If the data qualified for a normal distribution, the paired sample *t* test was used. If not normally distributed, the paired Wilcoxon signed rank test was used. All statistical tests were 2-sided and *P* values <.05 were considered statistically significant. To ensure a robust sample for the purpose of this feasibility study, we aimed to include a total of 60 participants [18].

Ethical Considerations

The Regional Committee for Medical Research Ethics in Southeast Norway did not find approval to be required. The study is approved by The Norwegian Data Protection Service for Research and the study has been registered on ClinicalTrials.gov (NCT06117098). The participants provided written informed consent before participation.

Results

Recruitment

From May 2023 and throughout December 2023, 9 GPs recruited a total of 54 participants of which 45 (82%) were eligible and included in this study. As shown in Figure 2, three participants dropped out during the intervention period. No reasons for drop out were provided.

XSL•FO
RenderX

Figure 2. Recruitment and participant flow.

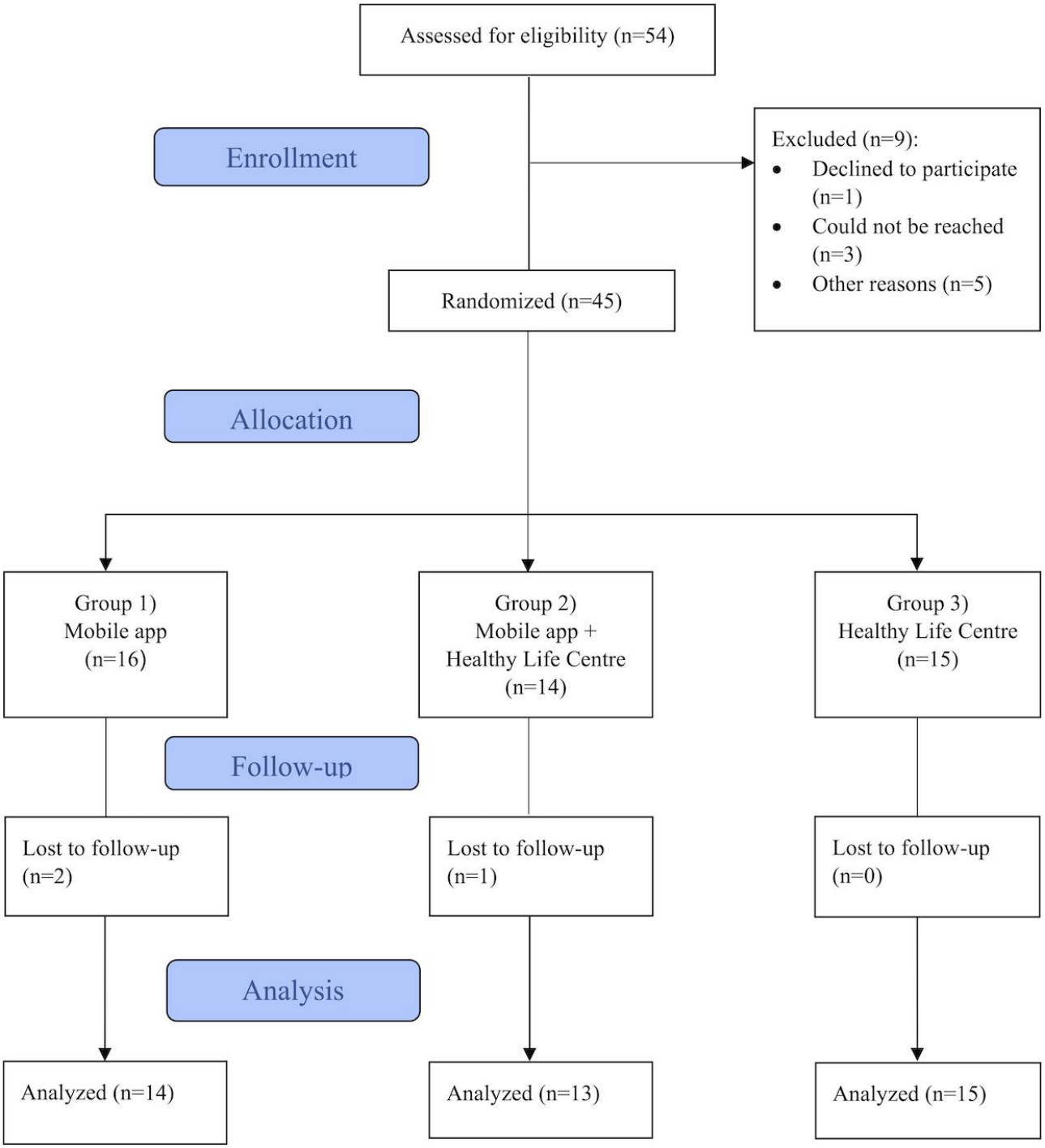

Participant Characteristics

Baseline characteristics are presented in Table 1. In total, 53% (n=24) were female and the mean age was 61 (SD 13) years with a range from 31 to 83 years. A total of 49% (n=22) of the participants had BMI ≥30, 88% (n=39) of the women had WC ≥80 cm, and 95% (n=43) of the men had WC ≥90 cm. The goals of the participants pertained to the 5 lifestyle domains of weight reduction, reduced HbA_{1c}, healthy diet, physical activity, and sleep.

XSL•FO
RenderX

Table . Baseline characteristics.

Characteristics	Study sample (N=45)	App (group 1; n=16)	App + Healthy Life Center (group 2; n=14)	Healthy Life Center (group 3; n=15)
Female, n (%)	24 (53)	6 (38)	9 (38)	9 (38)
Age (years), mean (SD)	60.8 (12.6)	54.5 (12.4)	59.1(12.7)	68.1 (9.1)
Body weight (kg), mean (SD)	91.0 (22.3)	96.6 (19.6)	84.6 (24.3)	89.7 (22.8)
BMI, mean (SD)	31.0 (6.6)	32.3 (5.4)	30.0 (7.0)	30.0 (7.0)
Waist circumference (cm), mean (SD)	105.2 (14.9)	108.5 (12.1)	99.0 (16.0)	106.7 (15.4)
Educational level, n (%)				
Elementary or high school	20 (46)	7 (44)	7 (50)	7 (47)
University or college >3 years	9 (20)	3 (19)	2 (14)	4 (27)
University or college ≤3 years	14 (33)	5 (31)	5 (36)	4 (27)
Employment, n (%)				
Full-time work or part-time work	19 (43)	9 (56)	8 (57)	4 (27)
Disability benefits	8 (18)	5 (31)	0 (0)	3 (20)
Retired	17 (39)	2 (13)	6 (43)	8 (53)
COOP/WONCA[a] score, mean (SD)				
Physical fitness	2.3 (1.0)	2.0 (0.9)	2.2 (1.1)	2.8 (1.0)
Feelings	1.8 (1.2)	1.8 (1.1)	1.8 (1.1)	1.8 (1.3)
Daily activities	1.5 (0.9)	1.4 (0.7)	1.5 (0.9)	1.7 (1.1)
Social activities	1.4 (0.8)	1.4 (0.8)	1.1 (0.3)	1.7 (1.0)
Change in health	2.6 (0.7)	2.5 (0.6)	2.6 (0.6)	2.8 (0.7)
Overall health	2.4 (0.8)	2.5 (0.7)	2.2 (0.8)	2.5 (0.7)
Total score, mean (SD)	11.8 (3.2)	11.6 (2.8)	10.6 (3.3)	13.1 (3.0)

[a]COOP/WONCA: Dartmouth Primary Care Cooperative Research Network/World Organization of Family Doctors functional health assessment chart.

Participants randomized to Plunde (groups 1 and 2; n=27) had a mean of 1.5 (SD 0.7) goals, 1 - 3 tasks related to each goal and a mean of 7.7 (SD 7.1) weekly reminders of their goal related tasks. The most frequent goal was related to weight reduction and reduced HbA_{1c}. Further, 17 (61%) participants randomized to Plunde had iOS smartphone and 11 (39%) had Android smartphone. For 2 (7%) participants, data regarding the operative system were not collected. No technical problems or bugs was reported during the intervention period.

Primary Outcome

At the 12 weeks follow-up, user log data could be retrieved from 27 participants. The overall app adherence score was 86% (n=27) for the whole intervention period. The mean SUS score was 86.6 (SD 12.0), indicating a grade A and excellent usability (n=22).

The mean app motivation score was 74.8 (SD 30.3). The motivational messaging from the supervisor was reported (n=26) to be highly motivating, with a mean score of 93.5 (SD 19.8). When asked for how long the participants would prefer to be followed up through the Plunde app, the most frequent answer was 12 months. On the question on how often the participants would find it useful to get feedback from their supervisor, every other week was the most frequent answer.

Secondary Outcomes

From May 2023 and throughout December 2023, 54 participants were recruited from 9 GP clinics within 8 months, of these 45 participants were eligible for the study. The recruitment rate was 1.4 participants per week.

The first consultation with the participants in group 1 and 2 (n=27) lasted in average for 91.3 (SD 26.5) minutes. Monitoring and follow-up through the app during the study period, took an average of 3.0 (SD 1.0) minutes per participant per week.

As presented in Table 2, preliminary pre-post interventions findings showed statistically significant within-group changes in 1 or more outcomes in all 3 groups. A significant reduction in body weight and WC was shown in group 1 and 3, and

XSL•FO
RenderX

improved changes in self-evaluated health was shown in all 3 groups. In the data analysis related to weight reduction, there was 1 data point missing from group 1. For WC, 1 data point was missing from both group 1 and 2. Meanwhile, for the COOP/WONCA analysis, group 3 had 1 missing data point.

Table . Changes in body weight, waist circumference, and self-evaluated functional health status at baseline (N=45) and after 12 weeks follow-up (n=42).

Outcomes and measures	Baseline, mean (SD)	12 weeks, mean (SD)	Mean difference (95% CI)	*P* value
App (group 1; n=14)				
Body weight (kg)	96.7 (20.1)	93.8 (18.5)	2.8 (0.8 to 4.9)	.011
Waist circumference (cm)	108.9 (12.4)	104.3 (11.3)	4.6 (2.0 to 7.1)	.002
COOP/WONCA[a] total score	11.9 (2.8)	9.4 (1.9)	2.5 (1.0 to 4.0)	.003
Physical fitness	2.1 (0.9)	1.8 (0.7)	0.4 (−0.2 to 0.9)	.174
Feelings	1.9 (1.1)	1.5 (0.9)	0.4 (−0.2 to 1.0)	.139
Daily activities	1.4 (0.8)	1.3 (0.7)	0.1 (−0.4 to 0.6)	.547
Social activities	1.4 (0.9)	1.0 (0.0)	0.4 (−0.6 to 0.9)	.082
Change in health	2.5 (0.7)	1.6 (0.6)	0.9 (0.5 to 1.2)	<.001
Overall health	2.6 (0.8)	2.1 (0.8)	0.4 (−0.1 to 0.9)	.082
App + Healthy Life Center (group 2; n=13)				
Body weight	81.5 (22.3)	80.2 (21.5)	1.3 (−0.5 to 3.1)	.136
Waist circumference (cm)	98.1 (13.8)	95.5 (15.3)	2.6 (−0.7 to 6.0)	.114
COOP/WONCA total score	10.3 (3.2)	9.4 (2.3)	0.9 (−1.5 to 3.3)	.423
Physical fitness	2.1 (1.0)	1.8 (0.8)	0.3 (−0.1 to 0.7)	.104
Feelings	1.8 (1.2)	1.8 (1.1)	0.0 (−0.7 to 0.7)	>.99
Daily activities	1.5 (1.0)	1.1 (0.3)	0.4 (−0.2 to 1.0)	.209
Social activities	1.1 (0.3)	1.1 (0.3)	__[b]	—
Change in health	2.5 (0.7)	2.2 (0.9)	0.9 (−0.3 to 1.5)	.004
Overall health	2.2 (0.8)	1.6 (0.7)	0.5 (0.1 to 0.9)	.012
Healthy Life Center (group 3; n=15)				
Body weight (kg)	89.7 (22.8)	86.9 (21.4)	2.7 (0.9 to 4.6)	.007
Waist circumference (cm)	106.7 (15.4)	103.2 (13.9)	3.5 (1.1 to 6.0)	.009
COOP/WONCA total score	13.1 (3.1)	11.4 (3.3)	1.8 (-0.2 to 3.7)	.070
Physical fitness	2.8 (1.0)	2.4 (1.1)	0.4 (−0.3 to 1.1)	.233
Feelings	1.8 (1.3)	2.1 (1.5)	−0.3 (−0.9 to 0.3)	.265
Daily activities	1.8 (1.1)	1.4 (0.9)	0.4 (−0.4 to 1.1)	.336
Social activities	1.7 (1.1)	1.1 (0.4)	0 (−0.02 to 1.2)	.055
Change in health	2.8 (0.7)	2.1 (0.8)	0.6 (−0.02 to 1.3)	.057
Overall health	2.5 (0.8)	2.0 (0.7)	0.5 (0.1 to 0.9)	.029

[a]COOP/WONCA: Dartmouth Primary Care Cooperative Research Network/World Organization of Family Doctors functional health assessment chart.
[b]The paired sample effect size could not be produced.

Discussion

Principal Findings

According to our preset criteria for feasibility, we found the Plunde app to be feasible to use for promoting and monitoring lifestyle change for people at risk of T2D. App adherence was high, and Plunde was rated with excellent user satisfaction. Furthermore, the app was evaluated as motivating, and feedback from the supervisors was scored as the most motivating factor. The amount of time spent on monitoring and promoting lifestyle change through the app was low. However, the recruitment was time consuming.

XSL•FO
RenderX

To our knowledge this is the first study assessing the feasibility of a mobile phone app to promote and monitor lifestyle change in people at risk of T2D. In our study, the adherence to the app was high. This is an important finding as the effectiveness of mHealth interventions have been proven to be closely related to app adherence and engagement [28]. Recently, Jakob et al [28] found that 5 intervention-related factors indicated positive effects on adherence to mHealth apps for prevention or management of NCDs. This included tailoring and personalization to the individual needs of the user, reminders in the format of individualized push notifications, being user-friendly, a technically stable app design, and personal support that complemented the digital intervention [28]. In line with this, 2 systematic reviews by Dugas et al [29] and Asbjørnsen et al [30], showed that the most frequent behavior change techniques used in mHealth interventions were goal-setting and planning, personalization, feedback and monitoring as well as prompts and cues. All these factors were integrated in the Plunde app and could, therefore, have contributed to the high adherence to the app. The integration of digital features and behavior change techniques may optimize T2D prevention interventions to achieve clinically significant weight loss [31] and improve self-management [32].

The app motivation scores in our study were high. Particularly, participants had high scores on the motivational feedback from the supervisor. Having “a person behind the app” can also be aligned with a so-called blended care model, characterized by a combination of digital interventions and therapist-guided interventions [33-35]. Blended care intervention has shown promising results on promoting lifestyle change in terms of weight loss and physical activity [34-36]. In previous research, the communication with the supervisor was highlighted as crucial, because of the possibility to communicate and receive individualized feedback by a real person [37]. On the contrary, the lack of direct contact and involvement with health care professionals has in previous study been pointed out as a main barrier for feasibility of an app-based intervention for cardiovascular disease and diabetes risk awareness and prevention [38]. An important component of this form of blended care model is the “onboarding” phase of the app intervention [39]. This is likely to be another factor having contributed to the high app adherence in this study. The initial face-to-face consultation enabled further tailored communication through the app, as it focused on the individual’s goals and tasks, and the associated barriers and facilitators to reach them. The freedom of choice and flexibility in a tailored program allows participants to set personalized and meaningful goals [15]. A systematic review and meta-analysis by Joiner et al [12] showed that technology combined with online health coaching resulted in greater weight loss compared to fully automated electronic health interventions. Thus, it is reasonable to assume that a crucial factor for adherence to the app is the individualized follow-up provided by a supervisor. Consequently, to succeed in a full-scale RCT, each participant should have a supervisor which can serve as the person behind the app and provide personalized and tailored feedback.

Weight reduction is the most important factor in preventing T2D, delaying the onset of T2D, and even remission of prediabetes and micro- and macrovascular risk [6,7]. In this feasibility study we assessed body weight, WC, and self-evaluated health at baseline and at 12 weeks follow-up. This was done to get experience with the measurements and to examine a potential change in these outcomes for a potential full-scale trial. In such a trial, one could consider to include the outcome waist-to-height ratio, as this variable has been shown to have an even stronger association with T2D risk than BMI, WC, and the waist-to-hip ratio [25,40,41]. Importantly, this study was designed as a feasibility RCT and did not aim to detect possible differences between groups, thus between-group analysis was not performed. We got valuable experiences with conducting the measurements and collecting data on body weight and WC.

Strengths and Limitations

A strength of this study is the heterogeneity of our sample in terms of age, gender, educational level, and sociodemographic status [42]. This increases the reliability and generalizability of our results, particularly regarding app adherence, app motivation, and user satisfaction. However, participation bias can be considered as a limitation in this study, as the people who agreed to participate are more likely to have a higher motivation for lifestyle change and being more aware of their risk of T2D. It is also important to consider that there is a potential risk in failing to recruit people at the highest risk [43]. However, considering the baseline characteristics of our study sample we seem to have reached the targeted population for risk reduction of T2D.

An important limitation to this study is that we changed our inclusion criteria during the recruitment process, jeopardizing a correct picture of the recruitment process. However, widening the inclusion criteria was a pragmatic choice, to achieve an adequate sample for being able to evaluate the feasibility of Plunde. If we had used the widened criteria from the beginning, the recruitment rate could have been faster. Thus, this has given us valuable considerations to the recruitment process in a potential full-scale RCT.

According to the recommendations for complex interventions and assessing feasibility [44] mixed methods is recommended [45]. Thus, not having qualitative interviews at the follow-up after ended intervention may present a limitation. Qualitative interviews could have provided us with valuable and more extensive information to better understand the complexity of adherence and engagement in use of the app. However, researchers from our group have previously investigated participants experiences with app-based follow-up over a period of 1 year [37]. Qualitative data and results from their work prepared the ground when designing the Plunde app in terms of functionality and likewise the role of the app supervisor to promote adherence and engagement to the app [17].

From a Feasibility Study to a Full-Scale RCT Study

Despite the overall feasibility of the Plunde, our study revealed some adjustments to be made prior commencing a full-scale RCT. To address the challenge of the slow recruitment rate, involvement of a substantial number of GP clinics seems necessary. However, it is important to consider that people at

XSL•FO
RenderX

risk of T2D, do not necessarily visit their GP very often, being in a state often without clinical symptoms [2].

Several recruitment strategies could be considered to enhance the recruitment rate and to increase the number of people available for recruitment. This could be done by developing a campaign for recruitment of participant into a full RCT, mediated or passive recruitment by using a wide variety of social media platforms. In person recruitment of people at risk of T2D has been conducted in venues like public spaces, libraries, churches, and different meeting places to reach people assessable for eligibility [46]. Rapid risk assessments and point-of-care testing of HbA_{1c} levels in community settings, for example GP offices, has been suggested to increase access to testing, facilitating early diagnosis, glycemic awareness, and lifestyle change [47], and could potentially increase the recruitment rate. Furthermore, using electronic health record-supported recruitment approaches has been found to be feasible and promising, and have the potential to increase the reach of eligible participants [48]. This could also be an important approach to facilitate the engagement of GPs and health care professionals in the research [48].

In terms of conducting a potential future 3-arm full-scale RCT, this feasibility study has provided us with valuable insights to the final decision of the design. We got experience with the administration of 3 intervention groups, including acceptability of the randomization procedure, and importantly, we gained important experience on the overall recruitment rate. In terms of conducting a potential future 3-arm full-scale RCT, we consider it as a strength to have evaluated feasibility of the app when used as “stand alone” intervention (group 1) and when used in combination with face-to-face group intervention (group 2). However, the complexity and resource requirements required to conduct a 3-arm design in a full-scale RCT must be taken under consideration. Regarding the usual care group (group 3), the variation in modalities offered by the HLCs and the risk of prolonged inclusion time to the HLC program, due to capacity and resource challenges, adds to the complexity and underscores the importance of an integrated stakeholder involvement in the planning and implementation of a full-scale RCT.

An advantage of conducting a full scale RCT with a 3-arm design, would be the potential to investigate whether an app intervention alone is as good as and not inferior to usual care (HLC). In addition, an evaluation whether a lifestyle intervention delivered as combination of app and HLC would be superior to usual care alone. Follow-up through an app instead of participating in HLC could be the preferred alternative for people because of the flexibility and accessibility an app provides. As a significant proportion of the targeted population for risk reduction of T2D are people committed to work and family responsibilities, and in terms of finding optimal prevention strategies, it is a challenge that the HLCs for the most offer their program during daytime and main working hours.

Despite the growing evidence on the effectiveness of mHealth interventions and lifestyle change for risk reduction of T2D, more evidence is needed regarding long-term adherence and effectiveness [28,49,50]. Considering our positive findings on the low amount of time spent on monitoring and providing follow-up, an assessment of long-term adherence and effectiveness implemented in a full-scale RCT would be feasible and sustainable in terms of resource requirements. When we asked the participants in this study how long they would have continued to use the app, most of the participants answered 12 months. In terms of individualized feedback from the supervisor, most of the app participants answered that they found feedback every month appropriate, suggesting that long-term follow does not necessarily need to be time-consuming from a health care professionals’ perspective.

The high speed of technology, innovative app development and the potential of artificial intelligence chatbots as supervisors provides a huge potential in providing scalable mHealth interventions for people at risk of T2D. The need for more research to understand which interventions and components have the greatest reach, promotes engagement and adherence and being most effective in the long term is consistently elucidated [50]. However, this feasibility study leads us to suggest that the potential of digital interventions or mHealth is also to be found beyond the technology itself. The blended care model design of the Plunde app contributes to the research on finding the optimal interaction model between health care professionals and people at risk of T2D.

Conclusions

Based on our preset criteria for success the Plunde app is feasible to monitor and promote lifestyle change in patients at risk of T2D. The amount of time spent on monitoring and promoting lifestyle change through the app was low; however, recruitment was slow. This feasibility study has provided important information and guidance for planning and executing a potential randomized controlled trial.

Acknowledgments

The authors would like to express their gratitude to the people who participated and made this study possible. The authors confirm that all personal identifiers have been removed or disguised so that those described are not identifiable and cannot be identified. They would like to thank the collaborating general practitioners and their secretaries for recruiting participants to our study and the Healthy Life Centers for including the participants to their service and following up in their program. They would like to thank associate professor Way Kiat Bong and MSc student Duy Lee at the Faculty of Technology, Art and Design (TKD) for providing standby IT technical support and resource. This project was internally funded by Oslo Metropolitan University.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Questions related to the experience with the setup and use of the smartphone application Plunde.
[PDF File, 106 KB - formative_v9i1e63737_app1.pdf]

Checklist 1
CONSORT-eHEALTH checklist (V1.6.1).
[PDF File, 8365 KB - formative_v9i1e63737_app2.pdf]

References

1. Sun H, Saeedi P, Karuranga S, et al. IDF Diabetes Atlas: global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. Diabetes Res Clin Pract 2022 Jan;183:109119. [doi: 10.1016/j.diabres.2021.109119] [Medline: 34879977]
2. Echouffo-Tcheugui JB, Perreault L, Ji L, Dagogo-Jack S. Diagnosis and management of prediabetes: a review. JAMA 2023 Apr 11;329(14):1206-1216. [doi: 10.1001/jama.2023.4063] [Medline: 37039787]
3. Magkos F, Hjorth MF, Astrup A. Diet and exercise in the prevention and treatment of type 2 diabetes mellitus. Nat Rev Endocrinol 2020 Oct;16(10):545-555. [doi: 10.1038/s41574-020-0381-5] [Medline: 32690918]
4. Kim HJ, Kwon O. Nutrition and exercise: cornerstones of health with emphasis on obesity and type 2 diabetes management-a narrative review. Obes Rev 2024 Aug;25(8):e13762. [doi: 10.1111/obr.13762] [Medline: 38715378]
5. Uusitupa M, Khan TA, Viguiliouk E, et al. Prevention of type 2 diabetes by lifestyle changes: a systematic review and meta-analysis. Nutrients 2019;11(11):2611. [doi: 10.3390/nu11112611]
6. Glechner A, Keuchel L, Affengruber L, et al. Effects of lifestyle changes on adults with prediabetes: a systematic review and meta-analysis. Prim Care Diabetes 2018 Oct;12(5):393-408. [doi: 10.1016/j.pcd.2018.07.003] [Medline: 30076075]
7. Sandforth A, von Schwartzenberg RJ, Arreola EV, et al. Mechanisms of weight loss-induced remission in people with prediabetes: a post-hoc analysis of the randomised, controlled, multicentre Prediabetes Lifestyle Intervention Study (PLIS). Lancet Diabetes Endocrinol 2023 Nov;11(11):798-810. [doi: 10.1016/S2213-8587(23)00235-8]
8. Jayedi A, Soltani S, Motlagh SZT, et al. Anthropometric and adiposity indicators and risk of type 2 diabetes: systematic review and dose-response meta-analysis of cohort studies. BMJ 2022 Jan 18;376:e067516. [doi: 10.1136/bmj-2021-067516] [Medline: 35042741]
9. Healthy Life Centres in Norway. Helsedirektoratet. 2017. URL: https://www.helsedirektoratet.no/english/healthy-life-centres [accessed 2024-12-27]
10. Tromp J, Jindal D, Redfern J, et al. World Heart Federation roadmap for digital health in cardiology. Glob Heart 2022;17(1):61. [doi: 10.5334/gh.1141] [Medline: 36051317]
11. mHealth: New Horizons for Health through Mobile Technologies: World Health Organization; 2011.
12. Joiner KL, Nam S, Whittemore R. Lifestyle interventions based on the diabetes prevention program delivered via eHealth: a systematic review and meta-analysis. Prev Med 2017 Jul;100:194-207. [doi: 10.1016/j.ypmed.2017.04.033]
13. Lunde P, Nilsson BB, Bergland A, Kværner KJ, Bye A. The effectiveness of smartphone apps for lifestyle improvement in noncommunicable diseases: systematic review and meta-analyses. J Med Internet Res 2018 May 4;20(5):e162. [doi: 10.2196/jmir.9751] [Medline: 29728346]
14. Barengo NC, Diaz Valencia PA, Apolina LM, et al. Mobile health technology in the primary prevention of type 2 diabetes: a systematic review. Curr Diab Rep 2022 Jan;22(1):1-10. [doi: 10.1007/s11892-021-01445-w] [Medline: 35113334]
15. Skoglund G, Nilsson BB, Olsen CF, Bergland A, Hilde G. Facilitators and barriers for lifestyle change in people with prediabetes: a meta-synthesis of qualitative studies. BMC Public Health 2022 Mar 21;22(1):553. [doi: 10.1186/s12889-022-12885-8] [Medline: 35313859]
16. Zangger G, Bricca A, Liaghat B, et al. Benefits and harms of digital health interventions promoting physical activity in people with chronic conditions: systematic review and meta-analysis. J Med Internet Res 2023;25:e46439. [doi: 10.2196/46439]
17. Lunde P, Skoglund G, Olsen CF, Hilde G, Bong WK, Nilsson BB. Think aloud testing of a smartphone app for lifestyle change among persons at risk of type 2 diabetes: usability study. JMIR Hum Factors 2023 Nov 15;10(1):e48950. [doi: 10.2196/48950] [Medline: 37966894]
18. Eldridge SM, Chan CL, Campbell MJ, et al. CONSORT 2010 statement: extension to randomised pilot and feasibility trials. BMJ 2016 Oct 24;355:i5239. [doi: 10.1136/bmj.i5239] [Medline: 27777223]
19. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res 2011;13(4):e126. [doi: 10.2196/jmir.1923]
20. Tilbud ved frisklivssentraler og veilederkurs [Article in Norwegian]. Helsedirektoratet. 2023. URL: https://www.helsedirektoratet.no/tema/frisklivssentraler/tilbud-ved-frisklivssentraler-og-veilederkurs [accessed 2024-12-27]

XSL•FO
RenderX

21. Armstrong MJ, Mottershead TA, Ronksley PE, Sigal RJ, Campbell TS, Hemmelgarn BR. Motivational interviewing to improve weight loss in overweight and/or obese patients: a systematic review and meta-analysis of randomized controlled trials. Obes Rev 2011 Sep;12(9):709-723. [doi: 10.1111/j.1467-789X.2011.00892.x] [Medline: 21692966]
22. Brooke J. SUS: a quick and dirty usability scale. In: Jordan PW, McClelland IL, Weerdmeester B, editors. Usability Evaluation In Industry: Taylor & Francis; 1996:189-194.
23. Hyzy M, Bond R, Mulvenna M, et al. System Usability Scale benchmarking for digital health apps: meta-analysis. JMIR Mhealth Uhealth 2022 Aug 18;10(8):e37290. [doi: 10.2196/37290] [Medline: 35980732]
24. Waist circumference and waist-hip ratio: report of a WHO expert consultation, Geneva, 8-11 December 2008. World Health Organization Institutional Repository for Information Sharing. 2008. URL: iris.who.int/bitstream/handle/10665/44583/?sequence=1 [accessed 2024-12-27]
25. Ke JF, Wang JW, Lu JX, Zhang ZH, Liu Y, Li LX. Waist-to-height ratio has a stronger association with cardiovascular risks than waist circumference, waist-hip ratio and body mass index in type 2 diabetes. Diabetes Res Clin Pract 2022 Jan;183:109151. [doi: 10.1016/j.diabres.2021.109151] [Medline: 34863718]
26. Siren R, Eriksson JG, Vanhanen H. Waist circumference a good indicator of future risk for type 2 diabetes and cardiovascular disease. BMC Public Health 2012 Aug 9;12(1-6):631. [doi: 10.1186/1471-2458-12-631] [Medline: 22877354]
27. Kinnersley P, Peters T, Stott N. Measuring functional health status in primary care using the COOP-WONCA charts: acceptability, range of scores, construct validity, reliability and sensitivity to change. Br J Gen Pract 1994 Dec;44(389):545-549. [Medline: 7748664]
28. Jakob R, Harperink S, Rudolf AM, et al. Factors influencing adherence to mHealth apps for prevention or management of noncommunicable diseases: systematic review. J Med Internet Res 2022 May 25;24(5):e35371. [doi: 10.2196/35371] [Medline: 35612886]
29. Dugas M, Gao GG, Agarwal R. Unpacking mHealth interventions: a systematic review of behavior change techniques used in randomized controlled trials assessing mHealth effectiveness. Dig Health 2020;6:2055207620905411. [doi: 10.1177/2055207620905411] [Medline: 32128233]
30. Asbjørnsen RA, Smedsrød ML, Solberg Nes L, et al. Persuasive system design principles and behavior change techniques to stimulate motivation and adherence in electronic health interventions to support weight loss maintenance: scoping review. J Med Internet Res 2019 Jun 21;21(6):e14265. [doi: 10.2196/14265] [Medline: 31228174]
31. Van Rhoon L, Byrne M, Morrissey E, Murphy J, McSharry J. A systematic review of the behaviour change techniques and digital features in technology-driven type 2 diabetes prevention interventions. D Health 2020;6:2055207620914427. [doi: 10.1177/2055207620914427] [Medline: 32269830]
32. El-Gayar O, Ofori M, Nawar N. On the efficacy of behavior change techniques in mHealth for self-management of diabetes: a meta-analysis. J Biomed Inform 2021 Jul;119:103839. [doi: 10.1016/j.jbi.2021.103839] [Medline: 34139330]
33. Obro LF, Heiselberg K, Krogh PG, et al. Combining mHealth and health-coaching for improving self-management in chronic care. A scoping review. Patient Educ Couns 2021 Apr;104(4):680-688. [doi: 10.1016/j.pec.2020.10.026] [Medline: 33143907]
34. Berry MP, Sala M, Abber SR, Forman EM. Incorporating automated digital interventions into coach-delivered weight loss treatment: a meta-analysis. Health Psychol 2021;40(8):534-545. [doi: 10.1037/hea0001106]
35. Hohberg V, Fuchs R, Gerber M, Künzler D, Paganini S, Faude O. Blended care interventions to promote physical activity: a systematic review of randomized controlled trials. Sports Med Open 2022 Jul 30;8(1):100. [doi: 10.1186/s40798-022-00489-w] [Medline: 35907158]
36. Kouwenhoven-Pasmooij TA, Robroek SJW, Kraaijenhagen RA, et al. Effectiveness of the blended-care lifestyle intervention “PerfectFit”: a cluster randomised trial in employees at risk for cardiovascular diseases. BMC Public Health 2018 Jun 19;18(1):766. [doi: 10.1186/s12889-018-5633-0] [Medline: 29921255]
37. Lunde P, Bye A, Bruusgaard KA, Hellem E, Nilsson BB. Patients’ experiences of using a smartphone app after cardiac rehabilitation: qualitative study. JMIR Hum Factors 2022 Mar 23;9(1):e34294. [doi: 10.2196/34294] [Medline: 35319476]
38. Buss VH, Varnfield M, Harris M, Barr M. Remotely conducted app-based intervention for cardiovascular disease and diabetes risk awareness and prevention: single-group feasibility trial. JMIR Hum Factors 2022 Jul 1;9(3):e38469. [doi: 10.2196/38469] [Medline: 35776504]
39. Bostrøm K, Børøsund E, Varsi C, et al. Digital self-management in support of patients living with chronic pain: feasibility pilot study. JMIR Form Res 2020 Oct 23;4(10):e23893. [doi: 10.2196/23893] [Medline: 33094734]
40. Moosaie F, Fatemi Abhari SM, Deravi N, et al. Waist-to-height ratio is a more accurate tool for predicting hypertension than waist-to-hip circumference and BMI in patients with type 2 diabetes: a prospective study. Front Public Health 2021;9:726288. [doi: 10.3389/fpubh.2021.726288] [Medline: 34692623]
41. Chen N, Hu LK, Sun Y, et al. Associations of waist-to-height ratio with the incidence of type 2 diabetes and mediation analysis: two independent cohort studies. Obes Res Clin Pract 2023 Jan;17(1):9-15. [doi: 10.1016/j.orcp.2022.12.005]
42. Kontochristopoulou AM, Karatzi K, Karaglani E, et al. Sociodemographic, anthropometric, and lifestyle correlates of prediabetes and type 2 diabetes in europe: the Feel4Diabetes study. Nutr Metab Cardiovasc Dis 2022 Aug;32(8):1851-1862. [doi: 10.1016/j.numecd.2022.03.021] [Medline: 35610084]

XSL•FO
RenderX

43. Bayley A, Stahl D, Ashworth M, et al. Response bias to a randomised controlled trial of a lifestyle intervention in people at high risk of cardiovascular disease: a cross-sectional analysis. BMC Public Health 2018 Sep 4;18(1):1092. [doi: 10.1186/s12889-018-5939-y] [Medline: 30180833]
44. Craig P, Dieppe P, Macintyre S, et al. Developing and evaluating complex interventions: the new Medical Research Council guidance. BMJ 2008 Sep 29;337:a1655. [doi: 10.1136/bmj.a1655] [Medline: 18824488]
45. Skivington K, Matthews L, Simpson SA, et al. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. BMJ 2021 Sep 30;374:2061. [doi: 10.1136/bmj.n2061] [Medline: 34593508]
46. Paglialunga S, Bond R, Jaycox SH. Evaluation of HbA1c screening during outreach events for prediabetes subject recruitment for clinical research. Trials 2018 Jan 22;19(1):60. [doi: 10.1186/s13063-018-2459-0] [Medline: 29357915]
47. Gourlay A, Sutherland C, Radley A. Point-of-care testing of HbA1c levels in community settings for people with established diabetes or people at risk of developing diabetes: a systematic review and meta-analysis. Prim Care Diabetes 2024 Feb;18(1):7-16. [doi: 10.1016/j.pcd.2023.10.011] [Medline: 37925311]
48. Aroda VR, Sheehan PR, Vickery EM, et al. Establishing an electronic health record-supported approach for outreach to and recruitment of persons at high risk of type 2 diabetes in clinical trials: the vitamin D and type 2 diabetes (D2d) study experience. Clin Trials 2019 Jun;16(3):306-315. [doi: 10.1177/1740774519839062] [Medline: 31007049]
49. Mönninghoff A, Kramer JN, Hess AJ, et al. Long-term effectiveness of mHealth physical activity interventions: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2021 Apr 30;23(4):e26699. [doi: 10.2196/26699] [Medline: 33811021]
50. Stowell M, Dobson R, Garner K, Baig M, Nehren N, Whittaker R. Digital interventions for self-management of prediabetes: a scoping review. PLoS One 2024;19(5):e0303074. [doi: 10.1371/journal.pone.0303074] [Medline: 38728296]

Abbreviations

CONSORT: Consilidated Standard of Reporting Trials
CONSORT-eHEALTH: Consilidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
GP: general practitioner
HLC: Healthy Life Center
mHealth: mobile health
NCD: noncommunicable disease
RCT: randomized controlled trial
SUS: System Usability Scale
T2D: type 2 diabetes
WC: waist circumference

Edited by A Mavragani; submitted 28.06.24; peer-reviewed by S Bonn; revised version received 29.10.24; accepted 30.10.24; published 15.01.25.

Please cite as:
Skoglund G, Hilde G, Lunde P, Cruz Naceno VV, Fromholt Olsen C, Blakstad Nilsson B
Mobile Phone App to Promote Lifestyle Change in People at Risk of Type 2 Diabetes: Feasibility 3-Arm Randomized Controlled Trial
JMIR Form Res 2025;9:e63737
URL: https://formative.jmir.org/2025/1/e63737
doi:10.2196/63737

© Gyri Skoglund, Gunvor Hilde, Pernille Lunde, Venessa Vera Cruz Naceno, Cecilie Fromholt Olsen, Birgitta Blakstad Nilsson. Originally published in JMIR Formative Research (https://formative.jmir.org), 15.1.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Considering Comorbidities and Individual Differences in Testing a Gaming Behavioral Activation App for Perinatal Depression and Anxiety: Open Trial Pilot Intervention Study

Gabriella E Hamlett[1,2], MA; Chloe Schrader[1,2], BA; Craig Ferguson[3], MS; Lauren A Kobylski[1,4], MPH; Rosalind Picard[3], ScD; Joseph J Locascio[5,6], PhD; Richard J McNally[2], PhD; Lee S Cohen[1,7], MD; Rachel Vanderkruik[1,7], MSc, PhD

[1]Center for Women's Mental Health, Massachusetts General Hospital, Boston, MA, United States
[2]Department of Psychology, Harvard University, Cambridge, MA, United States
[3]MIT Media Lab, Massachusetts Institute of Technology, Cambridge, MA, United States
[4]Department of Psychological & Brain Sciences, George Washington University, Washington, DC, United States
[5]Harvard Catalyst Biostatistical Consulting Group and Department of Neurology, Massachusetts General Hospital, Boston, MA, United States
[6]Department of Neurology, Harvard Medical School, Boston, MA, United States
[7]Department of Psychiatry, Harvard Medical School, Boston, MA, United States

Corresponding Author:
Rachel Vanderkruik, MSc, PhD
Center for Women's Mental Health, Massachusetts General Hospital, Boston, MA, United States

Abstract

Background: There is increasing interest in the development of scalable digital mental health interventions for perinatal populations to increase accessibility. Mobile behavioral activation (BA) is efficacious for the treatment of perinatal depression; however, the effect of comorbid anxiety and depression (CAD) on symptom trajectories remains underexplored. This is important given that at least 10% of women in the perinatal period experience CAD.

Objective: We assessed whether there were differences in symptom trajectories in pregnant participants with CAD as compared to those with depression only (ie, major depressive disorder [MDD]) during intervention with a BA mobile gaming app.

Methods: Pregnant adults with either CAD (n=10) or MDD (n=7) used a BA app for 10 weeks and completed biweekly symptom severity questionnaires for depression and anxiety. We assessed whether baseline diagnoses were associated with differential symptom trajectories across the study with mixed effects longitudinal models.

Results: When controlling for baseline symptoms, results revealed a significant interaction between baseline diagnosis and the quadratic component of study week on anxiety (β=.18, SE 0.07; t_{62}=2.61; P=.01), revealing a tendency for anxiety in the CAD group to increase initially and then decrease at an accelerated rate, whereas MDD symptoms were relatively stable across time. There was a significant effect of linear time on depression (β=−.39, SE 0.11; t_{68}=−3.51; P=.001), showing that depression declined steadily across time for both groups. There was a significant effect of baseline diagnosis on depression (β=−8.53, SE 3.93; t_{13}=−2.17; P=.05), suggesting that those with MDD had higher follow-up depression compared to those with CAD when holding other predictors constant.

Conclusions: The app was beneficial in reducing depression symptoms in perinatal individuals with different comorbidity profiles. With respect to anxiety symptom trajectories, however, there was more variability. The app may be especially effective for the treatment of anxiety symptoms among individuals with CAD, as it encourages in-the-moment ecologically relevant exposure to anxiety-provoking stimuli. Despite no significant group difference in baseline anxiety symptoms, the MDD group did not have a significant reduction in their anxiety symptoms across the study period, and some individuals had an increase in anxiety. Findings may point to opportunities for the augmentation of BA gaming apps for those with MDD to more effectively target anxiety symptoms. Overall, findings suggest there may be value in considering comorbidities and individual variations in participants when developing scalable mobile interventions for perinatal populations.

(JMIR Form Res 2025;9:e59154) doi:10.2196/59154

KEYWORDS

perinatal anxiety; perinatal depression; behavioral activation; digital mental health; mobile phone

XSL•FO
RenderX

Introduction

During the perinatal period, 19% - 40% of women experience anxiety or depression, and at least 10% experience comorbid anxiety and depression (CAD [1,2]). Despite its prevalence, CAD in perinatal populations is relatively underexplored [3]. CAD is associated with greater symptom severity, decreased treatment response, and longer episode duration [2-5]. Therefore, more research is needed to assess treatment considerations for perinatal individuals with CAD.

Behavioral activation (BA) is a behavioral intervention that aims to boost engagement in pleasurable activities, emphasizing value-driven behaviors and avoidance reduction, and is efficacious for the treatment of perinatal depression and anxiety [6]. While there is a strong need for the treatment of anxiety and depression in perinatal populations, access to treatment remains challenging and thus further highlights the need for scalable treatments that are accessible for use in ecologically valid settings [7]. Due to accessibility difficulties and maternity health care deserts [8], there is increasing interest in the development of scalable digital mental health interventions for perinatal populations [9-13].

Vanderkruik et al's [14] pilot study on the feasibility and acceptability of a BA gaming app for pregnant women suggests its potential to decrease depression symptoms during pregnancy. However, the effect of comorbidity on symptom trajectories remains unexplored, and few studies focus on individual differences in mobile BA treatment outcomes [14,15]. In this pilot study, we take a hypothesis-generating approach to assess differences in anxiety and depression symptom trajectories in participants with either CAD or major depressive disorder (MDD).

Methods

Recruitment

Participants (pregnant adults, native English speakers, smartphone users, and at least moderately depressed [14]) were recruited throughout the United States between 2021 and 2022 via clinician referrals, the Massachusetts General Hospital Center for Women's Mental Health website [16], and social media advertising. Ineligibility criteria included imminent risk of self-harm, current substance abuse, psychotic disorder, or active mania. All participants (n=18) met the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* (*DSM-5*) criteria for a current major depressive episode, 10 of whom additionally met criteria for generalized anxiety disorder (GAD) as assessed by the Mini International Neuropsychiatric Interview (MINI [17]). One participant dropped from the study due to active substance use, resulting in a final sample of 17.

Participants were instructed to use the app for the 10-week study period and invited to complete daily "Adventures" (eg, exercising and cleaning) in their real life to earn points and progress in the game. Participants completed surveys for depression and anxiety symptoms biweekly over the 10-week study period [14].

Ethical Considerations

This study is a secondary analysis of deidentified data from the parent study. All original study procedures were approved by the Mass General Brigham institutional review board (2021P001400). All participants provided informed consent in the parent study, which stipulated that deidentified data may be analyzed in secondary analyses without additional consent. Participants were not provided compensation for their participation.

Measures

Eligibility was assessed at baseline with the MINI [17], a semistructured interview conducted by a trained research assistant [14]. Participant grouping (CAD vs MDD) was determined at baseline using MINI diagnostic criteria. The CAD group included individuals who met *DSM-5* criteria for GAD (excessive, difficult to control anxiety and worry for most days for at least 6 months and 3 of the following symptoms: restlessness, being easily fatigued, difficulty concentrating, irritability, muscle tension, and sleep disturbance [18]) and MDD (a depressed mood for most of the day or anhedonia for at least 2-weeks, as well as 5 of the following symptoms: significant unintentional weight or appetite change, insomnia or hypersomnia, psychomotor agitation or retardation, fatigue, feelings of worthlessness or inappropriate or excessive guilt, decreased ability to concentrate or make decisions, and recurrent thoughts of death or suicidal ideation [19]). Those in the MDD group only met the criteria for MDD. Self-reported anxiety and depression symptoms within the past two weeks were measured with the Generalized Anxiety Disorder-7 (GAD-7 [18]) and Patient Health Questionnaire-9 (PHQ-9 [14]), respectively. Study completers were defined as participants who completed the final assessment at the end of the 10-week study period. As published previously [14], app engagement was measured by assessing days using the app (ie, the number of days a participant logged on) and activities completed (ie, the total number of activities a participant completed in the game).

Statistical Analysis

Given the scope of this pilot study, we chose a hypothesis-generating approach with the intent of informing future research, though we expected to see less favorable depression and anxiety symptom trajectories (according to PHQ-9 and GAD-7, respectively) for CAD relative to the MDD group. The analyses were conducted in R (version 4.3.0; R Core Team). Demographic variables, group differences in reductions of anxiety and depression symptoms, app engagement, study completion, and individual symptom trajectories by group were analyzed with descriptive statistics. Independent group *t* tests (2-tailed) explored differences in baseline measures of depression and anxiety by baseline diagnosis.

To assess individual variation in depression and anxiety symptom decreases over time within each group, we plotted individual participant data to visualize trajectories of change for anxiety and depression symptoms across time. To determine whether treatment outcomes changed differentially across time by baseline diagnosis while controlling for baseline depression and anxiety symptom levels, we used mixed effects longitudinal

XSL•FO
RenderX

models. In separate models, PHQ-9 and GAD-7 (hereafter referred to as depression and anxiety symptoms, respectively) were the dependent variables. We controlled for baseline depression and anxiety to ensure that trajectories were due to the app and diagnostic group rather than differences in baseline symptom levels. Fixed predictors were baseline diagnosis group (CAD vs MDD), time (linear and quadratic components of the week in study), baseline depression or anxiety symptoms for the analysis of their respective dependent variables, and all 2 and 3-way interactions of group, time, and baseline depression and anxiety symptoms. Participant-level covariates of baseline age and baseline week gestation were also included to ensure effects were due to the app and not differences in age or weeks gestation. The random effect was participant intercept. We progressively removed nonsignificant covariates and higher-order terms and reran the model. We checked the model residuals to ensure conformance to model assumptions of normality and calculated the proportion of variance in the dependent variable attributable to the fixed effects.

Results

Out of the 17 participants, 59% (n=10) had CAD and 41% (n=7) had MDD at baseline. Participants were primarily White, heterosexual, married, employed, and educated, with private health insurance, as seen in Table 1.

XSL•FO
RenderX

Table . Demographics and participant characteristics.

		Comorbid anxiety and depression (n=10)	Major depressive disorder (MDD; n=7)	Total (N=17)
Age (years), mean (SD)		33.4 (3.0)	35.8 (2.7)	34.4(3.1)
Weeks gestation, mean (SD)		18.1 (8.4)	15.1 (5.9)	16.8 (7.3)
Race or ethnicity, n (%)				
	White	10 (100)	2 (29)	12 (71)
	Black or African American	0 (0)	3 (43)	3 (18)
	Asian	0 (0)	2 (29)	2 (12)
	Non-Hispanic or Latina	9 (90)	6 (86)	15 (88)
	Hispanic or Latina	1 (10)	1 (14)	2 (12)
Sexual orientation, n (%)				
	Heterosexual	7 (70)	7 (100)	14 (82)
	Bisexual	2 (20)	0 (0)	2 (12)
	Queer	1 (10)	0 (0)	1 (6)
Marital status, n (%)				
	Married	9 (90)	3 (43)	12 (71)
	Divorced	0 (0)	1 (14)	1 (6)
	Never married	1 (10)	3 (43)	4 (24)
Employment status, n (%)				
	Employed	7 (70)	7 (100)	14 (82)
	Student	3 (30)	0 (0)	3 (18)
	Disabled or unable to work	0 (0)	0 (0)	0 (0)
Insurance status, n (%)				
	Private health insurance	10 (100)	6 (86)	16 (94)
	Medicaid	0 (0)	1 (14)	1 (6)
Education level, n (%)				
	Postgraduate training	7 (70)	5 (71)	12 (71)
	Bachelor's degree	3 (30)	1 (14)	4 (24)
	Some college	0 (0)	0 (0)	0 (0)
	High school diploma	0 (0)	0 (0)	0 (0)
	Some high school	0 (0)	1 (14)	1 (6)
Treatment, n (%)				
	Psychiatric medication in the last 2 months	6 (60)	2 (29)	8 (47)
	Psychosocial treatment in the last 2 months	5 (50)	3 (43)	8 (47)

Table 1 shows the demographics and participant characteristics of our participants in a pilot study conducted between 2021 and 2022 in the United States testing whether baseline symptom and diagnostic differences influence the symptom trajectory of pregnant individuals using a BA gaming app for perinatal depression and anxiety.

There were no significant group differences in age ($mean_{CAD}$ 33.4, SD_{CAD} 3.0 years and $mean_{MDD}$ 35.8, SD_{MDD} 2.7 years), baseline weeks gestation ($mean_{CAD}$ 18.1, SD_{CAD} 8.4 and $mean_{MDD}$ 15.1, SD_{MDD} 5.9), or baseline depression symptoms ($mean_{CAD}$ 12.8, SD_{CAD} 3.3 and $mean_{MDD}$ 11.7, SD_{MDD} 3.5). There was no significant difference in baseline past 2-week anxiety symptoms ($mean_{CAD}$ 10.7, SD_{CAD} 2.7 and $mean_{MDD}$ 7.0, SD_{MDD} 2.7; Table 2). That is, while both diagnostic groups started with similar past 2-week levels of anxiety, only the CAD group met full *DSM-5* criteria for GAD, which requires a duration of at least 6 months of symptoms. The CAD group was 3.6 times more likely to be using psychotropic medications (6

XSL•FO
RenderX

out of 10, 60% of participants with CAD, 2 out of 7, 29% of participants with MDD) though with this sample size, it did not reach statistical significance (*P*=.32). The CAD group was 1.2 times more likely to be in therapy (5 out of 10, 50% participants with CAD and 3 out of 7, 43% participants with MDD); findings did not reach significance (*P*=.63).

Table . Mean changes from baseline to follow-up in anxiety and depression for completers.

		PHQ-9[a] (mean, SD)	GAD-7[b] (mean, SD)
Comorbid anxiety and depression			
	Baseline (n=10)	12.8 (3.7)	10.7 (2.7)
	Baseline completers (n=7)	12.7 (3.9)	10.0 (2.9)
	Follow-up (n=7)	7.0 (3.4)	6.0 (2.6)
	Completers mean change score	−5.7 (4.3)	−4.0 (4.2)
Major depressive disorder			
	Baseline (n=7)	11.7 (3.5)	7.0 (2.7)
	Baseline completers (n=3)	13.7 (4.5)	7.3 (0.6)
	Follow-up (n=3)	7.7 (4.9)	9.7 (6.5)
	Completers mean change score	−6.0 (1.7)	2.3 (6.5)

[a]PHQ-9: Patient Health Questionnaire-9.

[b]GAD-7: Generalized Anxiety Disorder-7.

Of the 10 participants with CAD, 7 (70%) completed the last study assessment and of the original 7 participants with MDD, 3 (43%) completed the last study assessment. Baseline diagnosis (CAD vs MDD) was not a significant predictor of whether participants completed the last assessment (β=−1.36, SE 1.00; *z* score=−1.35; *P*=.18). There was also no significant difference in the number of days using the app between the CAD group and the MDD group ($mean_{CAD}$ 18.0, SD_{CAD} 16.20; $mean_{MDD}$ 11.6, SD_{MDD} 19.79; $t_{7.04}$=0.62; *P*=.56; 95% CI −18.09 to 30.89). There was no significant difference in the number of tasks completed between the CAD group and the MDD group ($mean_{CAD}$ 13.22, SD_{CAD} 14.07; $mean_{MDD}$ 6.0, SD_{MDD} 12.86; $t_{9.08}$=0.97; *P*=.36; 95% CI −9.55 to 23.99).

There was no significant difference between the depression symptom mean change score for completers from baseline to follow-up for the CAD group compared to the MDD group ($mean_{CAD}$ −5.7, SD_{CAD} 4.3; $mean_{MDD}$ −6.0, SD_{MDD} 1.7; t_8=0.11; *P*=.92; 95% CI −5.8 to −6.4). There was also no significant difference in the anxiety symptom mean change score for completers from baseline to follow-up for the CAD group compared to the MDD group ($mean_{CAD}$ −4.0, SD_{CAD} 4.2; $mean_{MDD}$ 2.3, SD_{MDD} 6.5; t_8=−1.87; *P*=.09; 95% CI −14.1 to 1.5).

When assessing individual-level trajectories of depression and anxiety symptom scores across study week by participant, we see within-group variation both for depression and anxiety symptom scores (Figure 1). In the MDD group, most individuals' anxiety symptoms decreased over time, though 1 individual demonstrated an increase in anxiety. The anxiety of some individuals with CAD increased before decreasing over time. While overall depression symptoms decreased over time, a few individuals in both groups displayed increases in symptoms.

XSL•FO
RenderX

Figure 1. Individual symptom trajectories of anxiety (GAD-7) and depression (PHQ-9) by baseline diagnosis. Figure 1 shows raw score symptom trajectories for each participant on anxiety (GAD-7) and depression (PHQ-9) symptoms by baseline diagnosis in a pilot study conducted between 2021 and 2022 in the United States testing whether baseline symptom and diagnostic differences influence symptom trajectory in pregnant individuals using a behavioral activation gaming app for perinatal depression and anxiety. Red triangles denote the mean score at each study week, with the red line representing a least squares regression line fit to these means. CAD: comorbid anxiety and depression; GAD-7: Generalized Anxiety Disorder-7; MDD: major depressive disorder; PHQ-9: Patient Health Questionnaire-9.

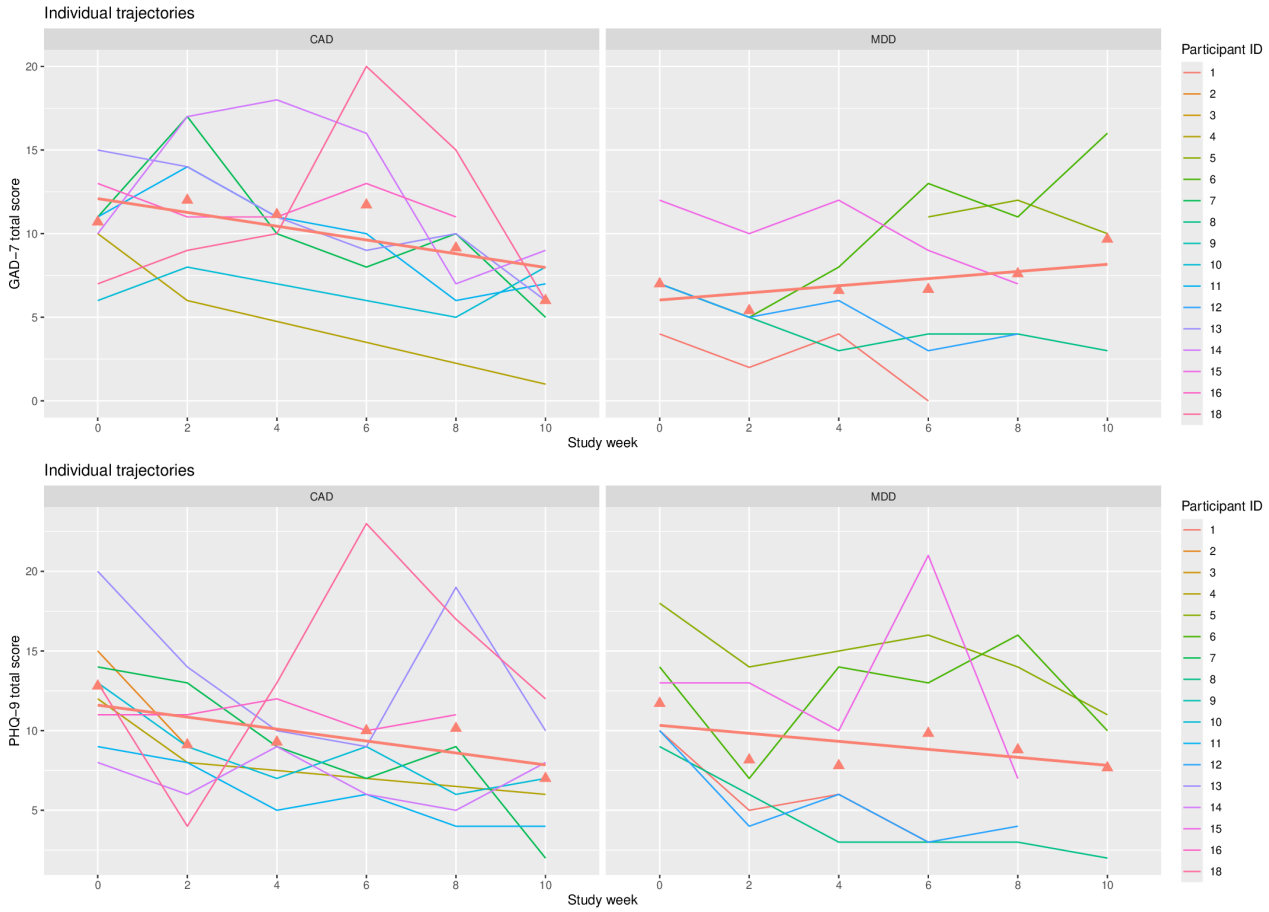

For mixed effects models, nonsignificant higher-order terms and covariates were removed, and models rerun. Relevant to the assessment of any association of baseline diagnosis to change in anxiety across time controlling for baseline anxiety, there was a significant interaction between baseline diagnosis and quadratic study week (β=.18, SE 0.07; t_{62}=2.61; *P*=.01). This interaction reflects a tendency for the CAD group anxiety symptoms to increase initially and then decrease at an accelerated rate, whereas the MDD groups' anxiety symptoms were relatively stable across time or even showed a slight "U" shape pattern across time (Figure 2). Overall, the fixed effects accounted for 33.9% of the total variance in the anxiety symptom scores. All other fixed effects of interest were nonsignificant (Table 3).

For the assessment of baseline diagnosis of depression across time controlling for baseline depression, there was a significant effect of linear time on depression symptom scores (β=−.39, SE 0.11; t_{68}=−3.51; *P*<.001), showing that depression symptoms decline steadily across time for both groups. We also found a significant effect of baseline diagnosis on depression symptoms at follow-up (β=−8.53, SE 3.93; t_{13}=−2.17; *P*=.05), suggesting that being in the MDD group was associated with slightly higher follow-up depression symptom scores compared to the CAD groups' follow-up depression symptom scores across time (Figure 2), holding other predictors (eg, baseline depression symptoms), constant. There was also a marginally significant interaction of baseline diagnosis with baseline depression symptoms (β=.65, SE 0.31; t_{13}=2.12; *P*=.054) suggesting that the difference in follow-up depression scores by baseline diagnosis difference was slightly stronger among those with lower baseline depression levels than at higher levels.

XSL•FO
RenderX

Figure 2. Model fixed effect predicted values of GAD-7 and PHQ-9 over time. Figure 2 shows effects plots from the 2 mixed effects models predicting anxiety (GAD-7; top panel) and depression (PHQ-9; bottom panel) across time in a pilot study conducted between 2021 and 2022 in the United States testing whether baseline symptom and diagnostic differences influence symptom trajectory in pregnant individuals using a behavioral activation gaming app for perinatal depression and anxiety. Predicted values from model fixed effects are shown with 95% confidence bands around the lines, controlling for baseline symptoms. CAD: comorbid anxiety and depression; GAD-7: Generalized Anxiety Disorder-7; MDD: major depressive disorder; PHQ-9: Patient Health Questionnaire-9.

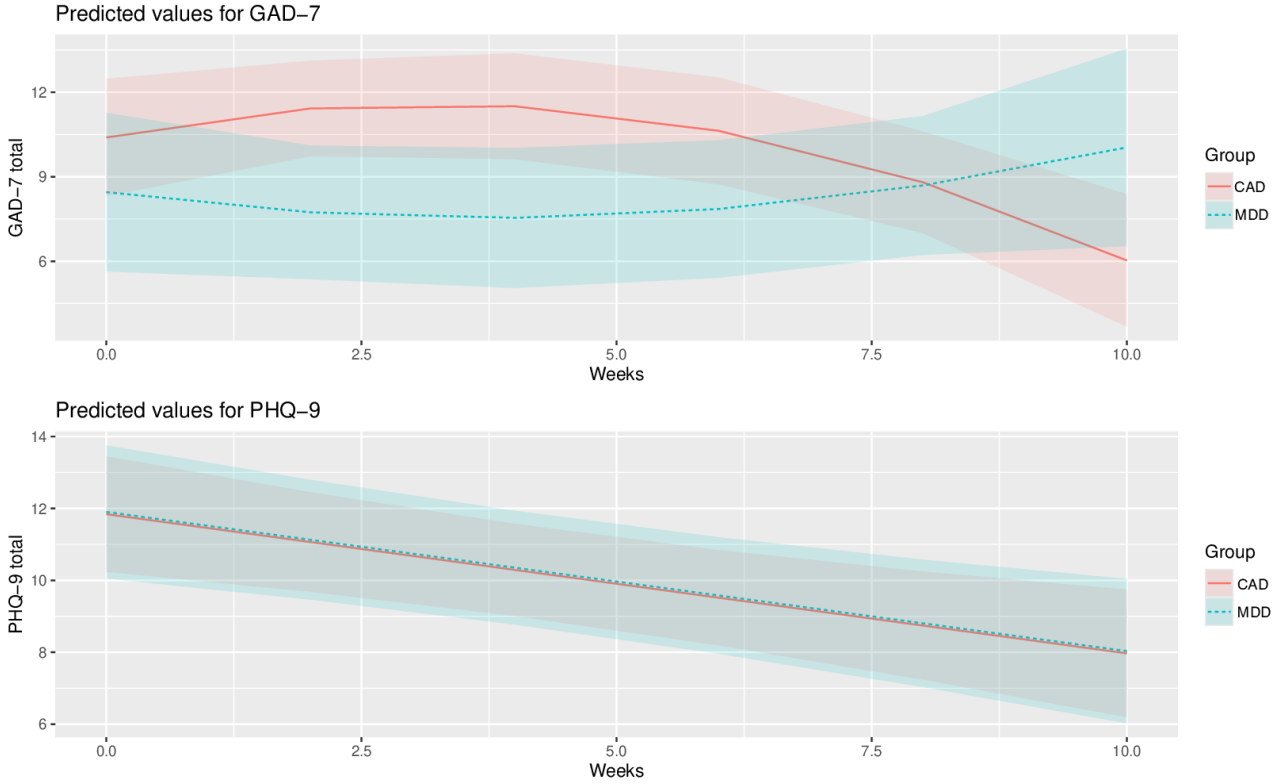

XSL•FO
RenderX

Table . Mixed models fixed effects results. Table 3 shows fixed effects results from both mixed effects models in a pilot study conducted between 2021 and 2022 in the US testing whether baseline symptom and diagnostic differences influence symptom trajectory in pregnant individuals using a behavioral activation gaming app for perinatal depression and anxiety.

		β (SE)	*t* test (*df*)	*P* value
GAD-7 [a] model				
	Intercept	5.06 (2.62)	1.93 (15)	.07
	Baseline diagnosis	−1.94 (1.81)	−1.07 (32)	.29
	Time[b]	.75 (0.44)	1.72 (63)	.09
	Time^2[c]	−.12 (0.04)	−2.76 (62)	.007
	Baseline GAD-7	.53 (0.23)	2.35 (13)	.04
	Baseline diagnosis time[b]	−1.24 (0.69)	−1.81 (62)	.08
	Baseline diagnosis time^2[c]	.18 (0.07)	2.61 (62)	.01
PHQ-9 [d] model				
	Intercept	4.37 (2.52)	1.73 (12)	.11
	Baseline diagnosis	−8.53 (3.93)	−2.17 (13)	.049
	Baseline PHQ-9	.57 (0.18)	3.07 (11)	.01
	Time	−.39 (0.11)	−3.51 (68)	.001
	Baseline diagnosis baseline PHQ-9	.65 (0.31)	2.12 (13)	.054

[a]GAD-7: Generalized Anxiety Disorder-7.

[b]Time: linear time.

[c]Time^2: quadratic time.

[d]PHQ-9: Patient Health Questionnaire-9.

Discussion

Principal Results

In this pilot study, our objective was to take a hypothesis-generating exploratory approach to assess differences in anxiety and depression symptom trajectories in participants with either CAD or MDD. While we expected to see less favorable depression and anxiety symptom trajectories (according to PHQ-9 and GAD-7, respectively) for CAD relative to the MDD group, our first main finding was that anxiety in the CAD group increased initially and then decreased at an accelerated rate, whereas in the MDD group, anxiety symptoms remained relatively stable across time despite no difference in baseline anxiety symptoms. Our other main finding was that depression symptoms declined steadily across time for both groups; however, those with MDD had higher follow-up depression scores compared to those with CAD, despite no differences in baseline depression symptoms. Given that there were no group differences in app engagement, these findings may not be explained by adherence or engagement. Our findings provide preliminary support for considering individual differences in baseline diagnostic characteristics when developing personalized mobile interventions, particularly in the context of treating perinatal individuals with psychological comorbidities.

With respect to the anxiety symptom trajectories, our findings are in line with the literature showing that compared to active controls, BA has a large effect on the reduction of depressive symptoms and a small effect on the reduction of anxiety symptoms when anxiety is secondary to depression [20]. In our study, it appears that the BA gaming app may be associated with a substantial reduction of anxiety symptoms for individuals with CAD. For these individuals, it may be that the BA components of the app target avoidance behaviors, which could result in a temporary increase in anxiety followed by reductions in anxiety potentially due to increased reward, habituation, and violation of expectations [6,21]. These findings suggest that the app may be especially effective for individuals with CAD, as it encourages in-the-moment ecologically relevant exposure to anxiety-provoking stimuli [7]. Particularly for individuals with CAD, for whom daily external stimuli may be especially provoking, the app may be especially useful and serve as an exposure.

Despite no significant group difference in baseline anxiety, the MDD group did not have a significant reduction in their anxiety symptoms across the study period, and some individuals had an increase in anxiety. Future studies should explore whether the anxiety experienced by the MDD group is maintained by other cognitive mechanisms to improve targeting anxiety symptoms when anxiety symptoms are secondary to depression symptoms. Findings may point to opportunities for the

XSL•FO
RenderX

augmentation of BA gaming apps for those with MDD to more effectively target anxiety symptoms. Future treatment personalization studies should also assess whether notifying patients of common treatment trajectories based on baseline presentations may improve treatment adherence. Findings of individual variability in anxiety and depression over time point to an opportunity for future more highly powered studies investigating the risk for increasing anxiety and depression across treatment during pregnancy.

Regarding our hypothesis that the CAD group may have less favorable depression symptom trajectories, we instead found that depression symptoms decreased linearly across time for both groups and that the MDD group had higher depression symptom scores at follow-up compared to the CAD groups' follow-up symptoms, suggesting that the app may be beneficial in reducing depression in perinatal individuals with different comorbidity profiles. These findings are in line with results from traditional and mobile BA interventions that BA is efficacious for the reduction of depressive symptoms (eg, [6,20,22]) even for individuals with different comorbidity profiles.

Limitations

This pilot study has several limitations (eg, modest sample size and no control group; discussed further in Vanderkruik et al [14]). Future fully powered studies should consider app engagement when assessing differences in symptom trajectories by baseline diagnosis. Due to the small sample size and lack of a control group, the results should be interpreted with caution and be considered as hypothesis generating.

Conclusions

The findings provide preliminary support for considering individual differences and comorbidities when developing scalable mobile interventions for perinatal populations. This pilot study has potentially important broad implications. Demand for treatment, particularly in perinatal populations, is high and yet can be hard to access for a variety of reasons, including the fact that many traditional interventions require presence in a formal and controlled therapy environment. This study represents an initial step toward developing scalable interventions and highlights the need to consider baseline differences in future studies that may personalize the app experience to increase adherence and effectiveness. Further research involving an adequately powered randomized controlled trial may further illuminate individual differences affecting the impact of mobile interventions.

Acknowledgments

The authors would like to thank all study participants for their participation in this study. This study was funded by the Harvard Medical School Kaplen Fellowship on Depression to RV at the MGH Center for Women's Mental Health. Generative artificial intelligence was not used for this paper.

Authors' Contributions

LSC, RV, and GEH contributed to the study conception and design. Material preparation and data collection were performed by RV, LAK, and CF. Analyses were performed by GEH and CS. JJL provided consultation. The first draft of the paper was written by GEH with contributions from RV and CS, and all authors commented on previous versions of the study. All authors read and approved the final paper. Funding was acquired by RV, and supervision was provided by JJL, RP, RJM, LSC, and RV.

Conflicts of Interest

LSC, as of February 26, 2024, receives research support as an employee of Massachusetts General Hospital (MGH) and works with the MGH National Pregnancy Registry. The current sponsors of MGH National Pregnancy Registry are Alkermes, Inc. (2016 to present); Dr. Reddy's Laboratories, Inc. (2023 to present); Eisai Inc. (2022 to present); Otsuka America Pharmaceutical, Inc. (2008 to present); Supernus Pharmaceuticals (2021 to present); and Teva Pharmaceutical Industries Ltd. (2018 to present). As an employee of MGH, LSC works with the MGH
Clinical Trials Network and Institute, which has had research funding from multiple pharmaceutical companies and the National Institute of Mental Health. LSC reports other research support from National Institutes of Health and SAGE Therapeutics but no support from advisory or consulting roles, no speaking fees or honoraria, no royalties or patents, and no other income. RV, as of March 14, 2024, receives research support as an employee of MGH and works on projects funded by The J Willard and Alice S Marriott Foundation and the National Eating Disorder Association. RV consults for the World Health Organization and Soulside, Inc.

References

1. Dennis CL, Falah-Hassani K, Shiri R. Prevalence of antenatal and postnatal anxiety: systematic review and meta-analysis. Br J Psychiatry 2017 May;210(5):315-323. [doi: 10.1192/bjp.bp.116.187179] [Medline: 28302701]
2. Falah-Hassani K, Shiri R, Dennis CL. The prevalence of antenatal and postnatal co-morbid anxiety and depression: a meta-analysis. Psychol Med 2017 Sep;47(12):2041-2053. [doi: 10.1017/S0033291717000617] [Medline: 28414017]

XSL•FO
RenderX

3. Goodman JH, Tyer-Viola L. Detection, treatment, and referral of perinatal depression and anxiety by obstetrical providers. J Womens Health (Larchmt) 2010 Mar;19(3):477-490. [doi: 10.1089/jwh.2008.1352]
4. Cuijpers P, Miguel C, Ciharova M, et al. Psychological treatment of depression with other comorbid mental disorders: systematic review and meta-analysis. Cogn Behav Ther 2023 May;52(3):246-268. [doi: 10.1080/16506073.2023.2166578] [Medline: 36718645]
5. Hopwood M. Anxiety symptoms in patients with major depressive disorder: commentary on prevalence and clinical implications. Neurol Ther 2023 Apr;12(Suppl 1):5-12. [doi: 10.1007/s40120-023-00469-6] [Medline: 37115459]
6. Dimidjian S, Goodman SH, Sherwood NE, et al. A pragmatic randomized clinical trial of behavioral activation for depressed pregnant women. J Consult Clin Psychol 2017;85(1):26-36. [doi: 10.1037/ccp0000151]
7. Ebrahimi OV, Asmundson GJG. Scaling up psychological interventions into the daily lives of patients with anxiety and related disorders. J Anxiety Disord 2024 Aug;106:102916. [doi: 10.1016/j.janxdis.2024.102916] [Medline: 39178811]
8. DiPietro Mager NA, Zollinger TW, Turman JE Jr, Zhang J, Dixon BE. Routine healthcare utilization among reproductive-age women residing in a rural maternity care desert. J Community Health 2021 Feb;46(1):108-116. [doi: 10.1007/s10900-020-00852-6] [Medline: 32488525]
9. Hadfield H, Wittkowski A. Women's experiences of seeking and receiving psychological and psychosocial interventions for postpartum depression: a systematic review and thematic synthesis of the qualitative literature. J Midwife Womens Health 2017 Nov;62(6):723-736. [doi: 10.1111/jmwh.12669]
10. Chen C, Wang X, Xu H, Li Y. Effectiveness of digital psychological interventions in reducing perinatal depression: a systematic review of meta-analyses. Arch Womens Ment Health 2023 Aug;26(4):423-439. [doi: 10.1007/s00737-023-01327-y] [Medline: 37231116]
11. Lara-Cinisomo S, Ramirez Olarte A, Rosales M, Barrera AZ. A systematic review of technology-based prevention and treatment interventions for perinatal depression and anxiety in Latina and African American women. Matern Child Health J 2021 Feb;25(2):268-281. [doi: 10.1007/s10995-020-03028-9] [Medline: 33389589]
12. Lewkowitz AK, Whelan AR, Ayala NK, et al. The effect of digital health interventions on postpartum depression or anxiety: a systematic review and meta-analysis of randomized controlled trials. Am J Obstet Gynecol 2024 Jan;230(1):12-43. [doi: 10.1016/j.ajog.2023.06.028]
13. Franco P, Olhaberry M, Muzard A, Harismendy Á, Kelders S. Developing a guided web app for postpartum depression symptoms: user-centered design approach. JMIR Form Res 2024 Aug 19;8(1):e56319. [doi: 10.2196/56319] [Medline: 39159447]
14. Vanderkruik RC, Ferguson C, Kobylski LA, et al. Testing a behavioral activation gaming app for depression during pregnancy: multimethod pilot study. JMIR Form Res 2024 Jan 26;8:e44029. [doi: 10.2196/44029] [Medline: 38277191]
15. Mancinelli E, Dell'Arciprete G, Pattarozzi D, Gabrielli S, Salcuni S. Digital behavioral activation interventions during the perinatal period: scoping review. JMIR Pediatr Parent 2023 Feb 28;6:e40937. [doi: 10.2196/40937] [Medline: 36853756]
16. Women's mental health across the life cycle. Massachusetts General Hospital Center for Women's Mental Health. URL: https://womensmentalhealth.org/ [accessed 2024-10-01]
17. Lecrubier Y, Sheehan D, Weiller E, et al. The MINI International Neuropsychiatric Interview (MINI). A short diagnostic structured interview: reliability and validity according to the CIDI. Eur psychiatr 1997;12(5):224-231. [doi: 10.1016/S0924-9338(97)83296-8]
18. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
19. Löwe B, Unützer J, Callahan CM, Perkins AJ, Kroenke K. Monitoring depression treatment outcomes with the patient health questionnaire-9. Med Care 2004 Dec;42(12):1194-1201. [doi: 10.1097/00005650-200412000-00006] [Medline: 15550799]
20. Stein AT, Carl E, Cuijpers P, Karyotaki E, Smits JAJ. Looking beyond depression: a meta-analysis of the effect of behavioral activation on depression, anxiety, and activation. Psychol Med 2021 Jul;51(9):1491-1504. [doi: 10.1017/S0033291720000239] [Medline: 32138802]
21. Hamlett GE, Foa EB, Brown LA. Exposure therapy and its mechanisms. In: Milad MR, Norrholm SD, editors. Fear Extinction: From Basic Neuroscience to Clinical Implications: Springer International Publishing; 2023:273-288. [doi: 10.1007/78542023428]
22. Dahne J, Wahlquist AE, Kustanowitz J, et al. Behavioral activation-based digital smoking cessation intervention for individuals with depressive symptoms: randomized clinical trial. J Med Internet Res 2023 Nov 1;25:e49809. [doi: 10.2196/49809] [Medline: 37910157]

Abbreviations

BA: behavioral activation
CAD: comorbid anxiety and depression
***DSM-5*:** *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition*
GAD: generalized anxiety disorder

XSL•FO
RenderX

GAD-7: Generalized Anxiety Disorder-7
MDD: major depressive disorder
MINI: Mini International Neuropsychiatric Interview
PHQ-9: Patient Health Questionnaire-9

Edited by A Mavragani; submitted 29.04.24; peer-reviewed by M Zimmermann; revised version received 01.11.24; accepted 05.11.24; published 14.01.25.

Please cite as:
Hamlett GE, Schrader C, Ferguson C, Kobylski LA, Picard R, Locascio JJ, McNally RJ, Cohen LS, Vanderkruik R
Considering Comorbidities and Individual Differences in Testing a Gaming Behavioral Activation App for Perinatal Depression and Anxiety: Open Trial Pilot Intervention Study
JMIR Form Res 2025;9:e59154
URL: https://formative.jmir.org/2025/1/e59154
doi:10.2196/59154

© Gabriella E Hamlett, Chloe Schrader, Craig Ferguson, Lauren A Kobylski, Rosalind Picard, Joseph J Locascio, Richard J McNally, Lee S Cohen, Rachel Vanderkruik. Originally published in JMIR Formative Research (https://formative.jmir.org), 14.1.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Verification of the Reliability of an Automated Urine Test Strip Colorimetric Program Using Colorimetric Analysis: Survey Study

Keigo Inagaki[1*], DPhil; Daisuke Tsuriya[2*], DPhil; Takuya Hashimoto[2], DPhil; Katsumasa Nakamura[2], Prof Dr Med

[1]Clinical Nursing, Gerontological Nursing, Faculty of Nursing, Hamamatsu University School of Medicine, 1-20-1 Handayama, Chuo-ku, Hamamatsu city, Shizuoka, Japan

[2]Hamamatsu University School of Medicine, Hamamatsu City, Chuo-ku, Japan

[*]these authors contributed equally

Corresponding Author:
Keigo Inagaki, DPhil
Clinical Nursing, Gerontological Nursing, Faculty of Nursing, Hamamatsu University School of Medicine, 1-20-1 Handayama, Chuo-ku, Hamamatsu city, Shizuoka, Japan

Abstract

Background: One method for noninvasive and simple urinary microalbumin testing is urine test strips. However, when visually assessing urine test strips, accurate assessment may be difficult due to environmental influences—such as lighting color and intensity—and the physical and psychological influences of the assessor. These complicate the formation of an objective assessment.

Objectives: This study developed an "automated urine test strip colorimetric program" (hereinafter referred to as "this program") to objectively assess urine test strips. Using this program may allow urine tests to be conducted at home. In this study, urine samples from hospitalized or outpatient patients were randomly obtained, and the reliability of this program was verified by comparing the agreement rate between this program and an automatic urine analyzer (US-3500 [Eiken Chemical Co, Ltd] and LABOSPECT 006 [Hitachi High-Tech Co, Ltd]). Furthermore, the sensitivity and specificity of the urine albumin test were investigated, and its applicability to screening for microalbuminuria was verified.

Methods: A urine test strip was placed in a photography box with constant light intensity and color temperature conditions. The image taken with a smartphone camera on top of the photography box was judged by this program. This program used Accelerated KAZE to perform image-matching processing to reduce the effect of misalignment during photography. It also calculated and judged the item with the smallest color difference between the color chart and the urine test strip using the CIEDE2000 color difference formula. The agreement rate of the results of this program was investigated using the results of an automatic urine analyzer as the gold standard.

Results: Compared with the judgments of an automatic urine analyzer, the average agreement rate for 12 items (protein, glucose, urobilinogen, bilirubin, ketone bodies, specific gravity, occult blood, pH, white blood cells, nitrite, creatinine, and albumin) was 78.6%. Furthermore, the average agreement rate of the 12 items within ±1 rank was 95.4%. The results showed a sensitivity of 100% and a specificity of 58.6% in determining albumin in urine, which is important for determining the stage of diabetic nephropathy. Finally, the area under the curve (0.907) derived from the receiver operating characteristic curve was satisfactory.

Conclusions: The program developed by the authors can determine urine test strips without requiring calibration in a certain shooting environment. If this program can be used at home to perform urinary microalbumin tests, the early detection and treatment of diabetic nephropathy may prevent the condition from becoming severe.

(JMIR Form Res 2025;9:e62772) doi:10.2196/62772

KEYWORDS

urine test strip; reliability; automatic urine analyzer; quasi-experimental study; colorimetric analysis; colorimetric; urinalysis; urinary; urine; evaluation; mobile phone

Introduction

Urinalysis testing displays vast physiological information and may be an effective tool for diagnosing various human diseases, including metabolic dysfunctions (eg, diabetes, urinary tract infections, and malignancies) and renal, cardiovascular, or hepatic disease [1-3]. Urinalysis test strips are an inexpensive, easy-to-use method that has long been used in clinical settings [4,5].

Urinalysis test strips change color in response to the concentration of the analyte present in the urine sample. Urinalysis test strips that react with urine analytes are checked against a reference color chart of several color blocks, each representing a specific analyte concentration.

XSL•FO
RenderX

The interpretation of colorimetric changes can be influenced by the health and psychological state of the observer [6,7]. Due to the lack of standardization and reading methods, it is not routinely used at home. The concept of self-screening at home has been progressively explored from multiple perspectives, and various types of devices have been developed depending on the application. For example, a Food and Drug Administration–approved product (Healthy.io) that uses a smartphone app to test items on urinalysis test strips for the early detection of chronic kidney disease and urinary tract infections is on the market. Various technologies exist for objectively judging urine test strips using smartphone apps such as “Vivoo” and “Uchek.”

This study developed an automated urine test strip colorimetric program to automatically judge urinalysis test strips from smartphone images [8]. This program automatically identified the item with the most negligible color difference between the reference color chart and urinalysis test strip and displayed the urinalysis results. Since the assessment was unaffected by the measurer’s physical condition or psychological state, the proposed method provided an objective rather than a subjective visual evaluation.

This study verified the reliability of an automated urine test strip colorimetric program, investigated the sensitivity and specificity of urinary albumin testing, and verified its applicability for microalbuminuria screening. Using real patient urine specimens, the concordance rate between creatinine and albumin and the 10 previously validated items was assessed. The concordance rate for each judgment item and the sensitivity and specificity of albumin were further investigated.

Methods

Overview

Figure 1 depicts this study’s experimental environment. A photography box (PU5025B; PULUZ) was used to maintain a constant lighting environment. An Xperia 5II (XQ-AS42, SONY) smartphone was used. The smartphone had a standard triple-lens camera (ZEISS lens) with a 1/1.7-inch large format sensor, approximately 12.2 million pixels/24-mm focal length, and a 1.7/dual photodiode/hybrid image stabilization (optical+electronic) *F* value. A urinalysis test strip (Uropaper III Eiken [E-UR97]; Eiken Chemical Co, Ltd) capable of testing 12 items was used as the test strip. The reference color chart, commonly used in clinical practice, was used as a comparison target for the urinalysis test strips (Figure 1). From June 2022 to January 2023, 1184 urine specimens were randomly extracted from specimens handled by the laboratory department of a hospital in Hamamatsu City. The inclusion criteria were urine samples from people aged 20-90 years who were evaluated at the laboratory department of a hospital in Hamamatsu City. The exclusion criteria were urine samples from people admitted to the pediatric ward and people with cloudy urine or gross hematuria. Basic information on the specimens is shown in Table 1. The US-3500 (Eiken Chemical Co, Ltd) and LABOSPECT 006 (Hitachi High-Tech Co, Ltd) were used as automatic urine analyzers for comparison with the results of the proposed method. The US-3500 was used to compare urobilinogen, occult blood, protein, glucose, ketone bodies, bilirubin, nitrite, specific gravity, white blood cells, and pH, and the LABOSPECT 006 was used to compare creatine and albumin levels. All 1184 specimens of urobilinogen, occult blood, protein, glucose, ketone bodies, bilirubin, nitrite, specific gravity, and pH were obtained using an automatic analyzer. Automated analyzer data were obtained for 1182 leukocyte, 214 creatine, and 42 albumin samples. The automatic urine test strip colorimetric program developed in this study was created using a desktop computer (HP EliteDesk 800 G4 TWR, HP), with Python as the development language and OpenCV library for image processing.

Figure 1. Reference color chart and shooting environment used for color comparison of urine test strips.

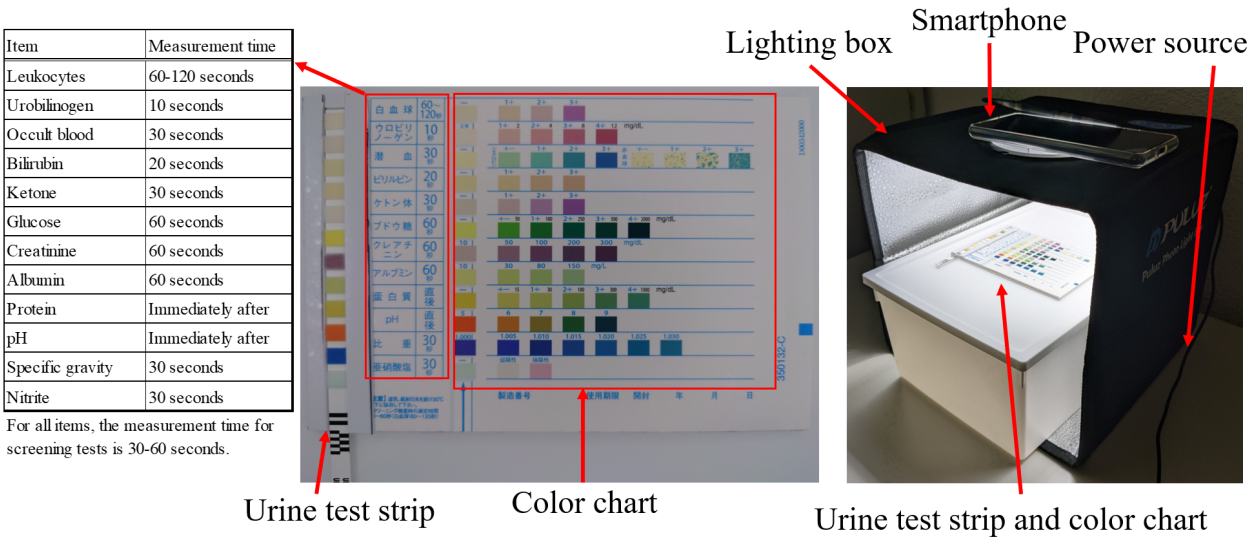

Item	Measurement time
Leukocytes	60-120 seconds
Urobilinogen	10 seconds
Occult blood	30 seconds
Bilirubin	20 seconds
Ketone	30 seconds
Glucose	60 seconds
Creatinine	60 seconds
Albumin	60 seconds
Protein	Immediately after
pH	Immediately after
Specific gravity	30 seconds
Nitrite	30 seconds

For all items, the measurement time for screening tests is 30-60 seconds.

XSL•FO
RenderX

Table . Basic information on 1184 urine samples randomly extracted from specimens handled by the testing department of a hospital in Hamamatsu City from June 2022 to January 2023.

Content	Individuals (N=1184)
Sex, n (%)	
Female	648 (54.7)
Male	536 (45.3)
Sex, Average age (SD)	
Female	48.16 (21.4)
Male	58.00 (22.8)
Clinical department, n (%)	
Department of rehabilitation	1 (0.1)
General surgery	12 (1)
Lower gastrointestinal surgery	14 (1.2)
Liver internal medicine	36 (3)
Hepato-biliary-pancreatic surgery	3 (0.3)
Ophthalmology	57 (4.8)
Hematology	55 (4.6)
Vascular surgery	1 (0.1)
Respiratory surgery	8 (0.7)
Respiratory medicine	31 (2.6)
Obstetrics and gynecology	263 (22.2)
Oral surgery	5 (0.4)
Otorhinolaryngology	6 (0.5)
Cardiology	24 (2)
Pediatrics	86 (7.3)
Gastroenterology	23 (1.9)
Upper gastrointestinal surgery	53 (4.5)
Cardiovascular surgery	18 (1.5)
Nephrology	73 (6.2)
Orthopedics	66 (5.6)
Psychiatry	4 (0.3)
Endocrinology and metabolism	131 (11.1)
Breast surgery	5 (0.4)
Neurosurgery	10 (0.8)
Urology	169 (14.3)
Dermatology	14 (1.2)
Immunology/rheumatology	15 (1.3)
Clinical pharmacology	1 (0.1)

XSL•FO
RenderX

Experimental Methods

Judgment of the Automatic Urine Dipstick Colorimetric Program

Automated Urine Dipstick Colorimetric Program

Figure 2 shows the processing flow of the automatic urine dipstick colorimetric program. The colors of the test paper for the urine test and color chart were compared according to the following procedure:

1. Specify the image file of the urinalysis strip.
2. Adjust the image position using Accelerated KAZE (AKAZE).
3. Acquire RGB (red, green, and blue; from the RGB color model) values from the central part (10×10 pixels) of the color chart and urinalysis test strip.
4. Convert to the L*a*b* color space.
5. Calculate the color difference between the color chart and the urine test strip using the CIEDE2000 color difference formula (color difference calculation with each color patch of the same selected item).
6. Display the judgment results with the smallest color difference.

Figure 2. Process flow for comparing the color of urine test strips and color charts. RGB: red, green, and blue.

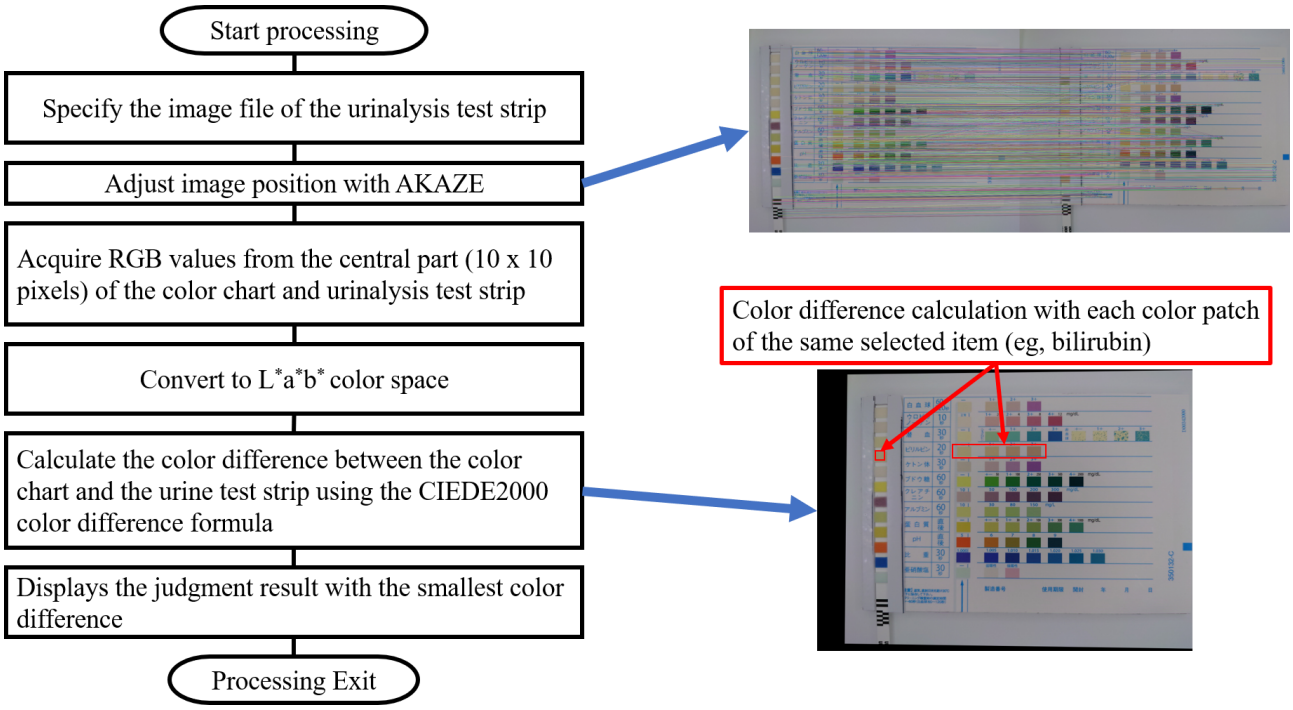

Filming Environment and Procedures

The color was evaluated using the following procedure in the same environment using a photography box (PU5025B):

1. Turn on the shooting box. Leave for 30 minutes to stabilize the lighting.
2. Start shooting 30 minutes after turning on the power.
3. Install the smartphone with the camera activated on top of the shooting box. Set the camera color temperature to sunlight, which is the same as that of the lighting of the shooting box.
4. Soak the urine test strip completely in the urine (approximately 1 or 2 seconds).
5. Lightly touch the side of the test paper with the tissue paper to remove excess urine on the test paper.
6. Place the test paper in a shooting box.
7. Take a picture 50 seconds after removing it from the urine sample (ie, when the paper is removed from the Spitz).
8. Judge the photographed images using an automatic urine dipstick colorimetric program.
9. Acquire age, sex, clinical department, and urinalysis results from electronic medical records.
10. Match the results with those of the automatic urine analyzer.

Accelerated KAZE

In this study, AKAZE, which is available in the OpenCV library, was used in image matching to reduce the effects of misalignment during shooting. AKAZE is an accelerated version of the algorithm used in KAZE [9], which is a 2D feature detection and description algorithm in nonlinear scale space developed by Alcantarilla et al [10]. This feature extraction algorithm is robust to changes in scaling, rotation, and lighting, and is similar to scale-invariant feature transform (SIFT). The KAZE algorithm is an improved version of SIFT and uses a nonlinear diffusion filter instead of a Gaussian filter to extract features while preserving the original image features and edges. SIFT uses a Gaussian filter to detect features. The center pixel of interest to be processed has the largest weight, and the outer pixels have smaller weights. This allows for images to be blurred while retaining more information around the pixels.

Judging Method for Urine Test Strips

The following methods are used for judging:

1. Approximate selection method: select the color on the color chart that is closest to the color of the test strip.

XSL•FO
RenderX

2. Rounding down method: if the color of the test strip does not reach the color on the color chart, judge it as a color with a lower density.
3. Rounding up method: if the color of the test strip is even slightly darker than that of the color chart, judge it as a color with a higher density.

Currently, there is no standard judgment method, but the Urinary Protein Measurement Committee has reported that "The approximate selection method seems to be theoretically correct, but it is prone to differences due to lighting and differences in each person's vision" [11]. Therefore, if the effects of lighting and individual differences can be reduced using this program, the "approximate selection method" may be suitable as an evaluation method. In this study, the "approximate selection method" was used.

Statistical Analysis

The rate of agreement was calculated using the results of the automatic urine analyzer. A concordance rate of ±1 was calculated as the concordance rate within 1 rank before and after the matching item. IBM SPSS Statistics 28th edition was used for the receiver operating characteristic (ROC) curve and area under the curve (AUC).

Ethical Considerations

This research was reviewed and approved by the ethics committee of the Hamamatsu University School of Medicine (approval number: 21 - 239). This study used existing samples in addition to existing information. However, as the samples were submitted to outpatient and other clinics, it was difficult to obtain direct consent. As such, information about this study, including the purpose of use of the samples and information, was made public on the Hamamatsu University School of Medicine's website and posted on notices within the hospital. Research subjects were given an opt-out form via information disclosure document and were guaranteed the opportunity to refuse participation in the study.

The samples and information provided, including personal information, were managed responsibly by the principal investigator (KI) and were anonymized such that specific individuals could be identified using a correspondence table or other means. This study collected only existing samples and information accumulated in the course of normal medical treatment and care, and there were no adverse events or malfunctions caused by the implementation of the study. This study targeted only the collection of existing samples and information accumulated during routine medical treatment and care, and there was no possibility of health damage occurring as a result of the study, so compensation for health damage and insurance enrollment did not apply.

Results

Agreement Rates and ±1 Rank Agreement Between Automated Urine Analyzers and Automated Urine Test Paper Colorimetric Programs

In this study, the US-3500, which is capable of qualitative evaluation using urine test strips, and the LABOSPECT 006, which is capable of quantitative evaluation of creatinine and albumin, were used. The results of the automated urine analyzers (US-3500 and LABOSPECT 006) were referred to the values recorded in the electronic medical records of the cooperating hospitals in the study. Table 2 shows the concordance rate of the automated urine analyzer and automated urine test strip colorimetric program, a concordance rate of ±1 rank, and the number of specimens. Urobilinogen, occult blood, protein, glucose, ketone bodies, bilirubin, nitrite, specific gravity, and pH were evaluated for concordance in 1184 samples. For white blood cells, 1182 specimens were evaluated for concordance. The amount of data obtained for creatinine and albumin is small due to the small number of daily test requests at the hospitals cooperating in the study. Therefore, the concordance rate was evaluated using 214 specimens for creatine and 42 specimens for albumin, which was in line with the data obtained from the automatic urine analyzer. The concordance rate for each item was more than 60%, except for specific gravity. The mean concordance for the 12 items was 78.6%. The matching rate of the ±1 rank for each item exceeded 90% for items other than specific gravity. The mean ±1 rank agreement for the 12 items was 95.4%. Multimedia Appendix 1 shows the concordance rates for each inspection item. The results indicate that an automated urine test paper colorimetric program could be applied to screening within ±1 rank.

XSL•FO
RenderX

Table . The concordance and ±1 rank match rate of the output results of the automated urine analyzer and the automated urine test strip colorimetric program and the number of samples obtained[a].

Item	Concordance rate, %	±1 rank match rate, %	Number of samples, n
Protein	83.1	100	1184
Glucose	94.8	100	1184
Urobilinogen	86	99.9	1184
Bilirubin	72.4	99.4	1184
Ketone	98	100	1184
Specific gravity	14.4	49.7	1184
Occult blood	87.8	97.8	1184
pH	80.7	100	1184
Leukocytes	86.6	99.1	1182
Nitrite	98.5	100	1184
Creatinine	64	98.6	214
Albumin	76.2	100	42
Average of 12 items	78.6	95.4	N/A[b]

[a]The samples were randomly selected from samples handled in the testing department of a hospital in Hamamatsu City from June 2022 to January 2023.
[b]N/A: not applicable.

Sensitivity and Specificity for Albumin

Trace albumin is a useful biomarker for early detection of nephropathy in patients with diabetes. Therefore, if trace albumin, which is important in determining the second stage of nephropathy, can be detected, it may lead to early detection of the progression of nephropathy. To confirm the applicability of our automated urine test paper colorimetric program for screening trace albuminuria, we calculated the sensitivity and specificity of the program for albumin. Since this program is a qualitative evaluation using urine test strips, the albumin value cannot be corrected for creatinine. Therefore, only the albumin value in urine was used for validation. Table 3 shows a cross-table for calculating the sensitivity and specificity of albumin. A diagnostic factor for diabetic nephropathy is microalbuminuria, which ranges from 30 to 299 mg/L. The cross-table was created assuming that patients with albumin levels of ≥30 and <30 were classified as having and not having the disease. Program judgment results show the program's output; microalbumin (urine automatic analysis data) shows the results recorded by LABOSPECT 006 in the electronic medical records of the collaborating hospitals. Cross-tab calculations yielded a sensitivity and specificity of 100% and 58.6%, respectively. Figure 3 shows the ROC curves. The AUC was 0.907, which is very favorable. Therefore, the results indicate that this program could be applied to screening for stage 2 nephropathy in diabetic nephropathy.

Table . Albumin sensitivity and specificity[a].

	Microalbumin (urine automatic analysis data)		Total
	With disease (albumin ≥30 mg/L)	With no disease (albumin <30 mg/L)	
Program judgment result			
Positive (albumin ≥30 mg/L)	13	12	25
Negative (albumin <30 mg/L)	0	17	17
Total	13	29	

[a]Patients with albumin levels of ≥30 and <30 were classified as those with diabetes and with no diabetes, respectively. Sensitivity and specificity were calculated by creating a cross-tabulation using the output of the automated urine analyzer and the automated urine test strip colorimetric program. From June 2022 to January 2023, 42 urine specimens were extracted from specimens handled by the laboratory department of a hospital in Hamamatsu City.

XSL•FO
RenderX

Figure 3. Receiver operating characteristic curve showing sensitivity and specificity of albumin. Patients with albumin levels of ≥30 were classified as those with diabetes, and those with albumin levels <30 were classified as patients with no diabetes. Sensitivity and specificity were calculated by creating a cross-table. From June 2022 to January 2023, 42 urine specimens were extracted from specimens handled by the laboratory department of a hospital in Hamamatsu City.

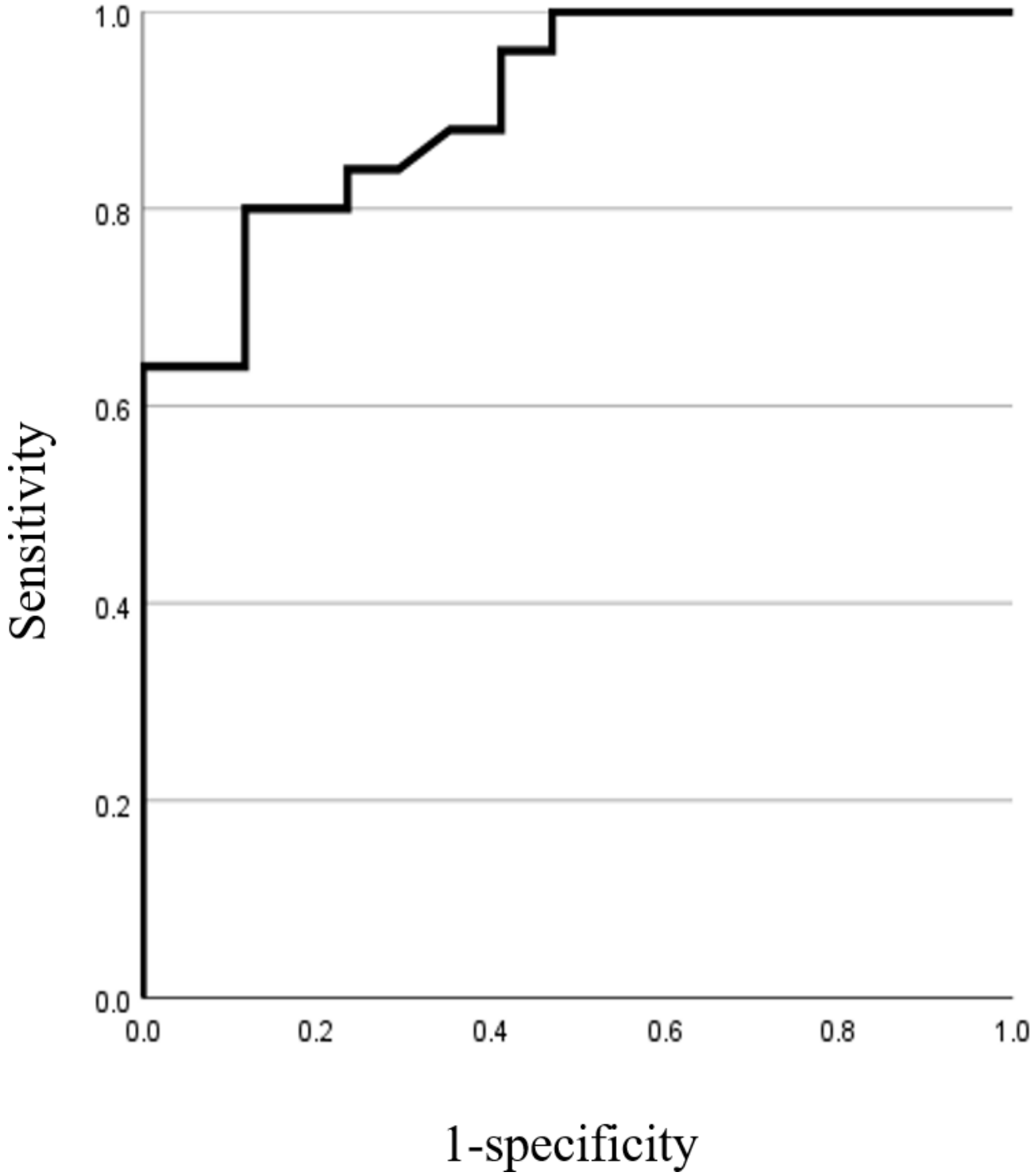

Discussion

Principal Results

This study used real patient urine specimens to investigate the agreement between an automated urine test strip colorimetric program and an automated urine analyzer. The concordance rate for each judgment item and the sensitivity and specificity of albumin were also evaluated. The concordance rate of the ±1 rank was 90% or more; thus, the concordance rate could be improved by devising the judgment threshold and judgment algorithm of the program.

The concordance rate of specific gravity obtained with this program was very low and was consistent with previous studies [12]. In the visual judgment, the colors of the reference color chart and test paper for urinalysis differed. Determining specific gravity with a low matching rate by simple color comparison

XSL•FO
RenderX

was challenging; therefore, a determination algorithm should be considered.

The detailed results for each item showed many negative and few positive judgments. However, despite the small number of positive tests, many items matched, supporting the effectiveness of the proposed method.

The proposed method determined the matching rate using a program that identifies the color closest to the reference color chart. Since the proposed method used the "approximate selection method," some items may present a low matching rate. Therefore, other evaluation methods should be considered in the future.

Previous reports have included many studies evaluating the reliability of programs using real urine samples, artificial urine, and reagents [13-17]. In a report using real urine samples, urine testing was performed using a smartphone app developed by Healthy.io, and its acceptability and feasibility were evaluated [16,17]. The satisfaction rate was very high among patients and caregivers, and the app could greatly enhance patient-centered care [16]. The potential of remote digital urine testing has also been reported [17].

Rahmat et al [18] validated the scanner color comparison using real urine specimens and reported an accuracy of 95.45% using a high-resolution scanner. The low-resolution scanner reported 83%, comparable with our automated urine test strip colorimetric program. Africa and Velasco [19] also verified the accuracy of color comparison using 45 samples from an accredited urinalysis laboratory and reported an accuracy of 96.519%. However, besides evaluating 10 items (ie, protein, glucose, ketone, specific gravity, occult blood, leukocytes, nitrite, urobilinogen, bilirubin, and pH), these 2 studies did not provide the breakdown of the concordance rate (ie, negative and positive breakdown and item-by-item classification). Urinary albumin and creatinine levels, necessary for determining diabetic nephropathy, have not yet been evaluated in the literature. Determining albumin and creatinine may help screen for microalbuminuria, enabling intervention before diabetic nephropathy becomes severe.

The sensitivity and specificity of this method for albumin detection were 100% and 58.6%, respectively. These suggest potentially high and low rates of false positives and negatives, respectively, allowing for people with diseases to be identified. Furthermore, the AUC derived from the ROC curve was 0.907, indicating satisfactory results. Therefore, the proposed method may be used as initial screening.

The proposed method may also support point-of-care (POC) testing at home to prevent the progression of diabetic nephropathy. Thakur et al [15] reported a method for albumin screening using smartphones and urine test strips. The study reported a 92% accuracy rate, validated only with reagents. In contrast, this study used actual urine specimens, and excellent screening results were obtained with a concordance rate of 76.2%, sensitivity of 100%, and specificity of 58.6%, relative to the results of an automated urine analyzer. Due to the COVID-19 pandemic, many people refrained from seeing doctors. As such, the proposed method is expected to be particularly useful for POC tests for home health management. Therefore, we are currently developing an application that uses the proposed system.

The proposed automatic urine test strip colorimetric program focuses only on color differences but does not require a large amount of data from machine learning. In addition, if standard color chart data from each manufacturer become available, urinalysis test strips from various manufacturers worldwide may be used without requiring calibration. In the future, the accuracy of determination may be improved by using data obtained from specimens, adjusting the threshold for determination, and examining determination methods.

Limitations

Despite its contributions, this study has limitations. The results do not apply to all smartphones because verification was performed using only 1 smartphone model. In the future, the program should be tested on other operating systems, smartphones, and cameras. In addition, owing to lighting limitations due to the shooting box, development did not progress sufficiently enough that judgment could be formulated with just a smartphone. This system is not suitable as a POC test; therefore, future investigations should verify whether similar results can be obtained using the flash function of a smartphone. In the future, the differences between multiple models and methods that do not use shooting boxes should be considered.

In addition, this study used urine test strips to evaluate the agreement between the white blood cell count obtained by an automated urine analyzer and the program developed in this study. Normally, the agreement with sediment analysis is verified. Future studies should compare the results of the formulated program with those from sediment analysis to compare the accuracy.

When urine samples were randomly collected, many samples did not have albumin analysis in the doctor's orders. Therefore, the albumin results from the automated urine analyzer were not recorded in the electronic medical record, and the match rate could not be calculated for many samples. The study protocol should be adjusted so that an adequate sample of albumin tests can be collected.

Normally, creatinine should be corrected in urinalysis results. However, given that the coincidence rate of creatinine was not high, only albumin was highlighted. In the future, accuracy can be improved by evaluating the albumin-creatinine ratio.

The analyzers used in this study (US-3500 and LABOSPECT 006) can evaluate reflectance, allowing for a more accurate evaluation of the consistency rate of urine test strips. However, this study used data stored in electronic medical records, and reflectance data could not be obtained. In the future, reflectance data should also be analyzed.

In this study, the agreement rate by qualitative evaluation was verified using urine test strips. However, sensitivity is low when using urine test strips. Therefore, in the future, reliability should be verified by comparing data that can be quantitatively measured, such as immunoturbidimetric methods. Finally, to

XSL•FO
RenderX

avoid over- or underestimating clinical scenarios, sample collection periods should span several years.

Conclusions

This study used real patient urine specimens to investigate the agreement between an automated urine test strip colorimetric program and an automated urine analyzer. It also investigated the concordance rate for each judgment item and the sensitivity and specificity of the program in evaluating albumin. The average concordance rate and rate of ±1 rank for the mean of the 12 inspected items were 78.6% and 95.4%, respectively. These results clarified that colorimetric analysis in a fixed imaging environment could be used to determine urine test strips without calibration.

Urinary albumin determination, essential for determining the stage of diabetic nephropathy, resulted in a sensitivity of 100% and a specificity of 58.6%. In addition, the AUC derived from the ROC curve was 0.907, indicating good results. The proposed method may be used for the screening of microalbuminuria to help prevent the exacerbation of diabetic nephropathy.

Acknowledgments

The authors thank Dr Masato Maekawa, Mr Keita Yamashita, Ms Misaki Hidaka, and the laboratory staff of Hamamatsu University Hospital for providing urine specimens and coordinating schedules in the midst of their busy work schedules. Professor Katsumasa Nakamura provided oversight and leadership responsibility for the research activity planning and execution, including mentorship external to the core team. Dr Daisuke Tsuriya and Dr Takuya Hashimoto were responsible for the management and coordination of the planning and execution of research activities. This work was supported by JSPS KAKENHI (grant 24K15836) and Terumo Life Science Foundation (21-Ⅲ7009), Hamamatsu University School of Medicine grant-in-aid. The funder played no role in the study design, data collection, data analysis, and interpretation, or manuscript writing.

Disclaimer

No artificial intelligence, language modeling, machine learning, or similar technologies have been used to create or assist in the creation or editing of any content (eg, text, tables, figures, and video) included in this submission.

Data Availability

The datasets generated during and analyzed during this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary Table
[DOCX File, 43 KB - formative_v9i1e62772_app1.docx]

References

1. Coppens A, Speeckaert M, Delanghe J. The pre-analytical challenges of routine urinalysis. Acta Clin Belg 2010;65(3):182-189. [doi: 10.1179/acb.2010.038] [Medline: 20669786]
2. Devillé WLJM, Yzermans JC, van Duijn NP, Bezemer PD, van der Windt DAWM, Bouter LM. The urine dipstick test useful to rule out infections. A meta-analysis of the accuracy. BMC Urol 2004 Jun 2;4:1-14. [doi: 10.1186/1471-2490-4-4] [Medline: 15175113]
3. Simerville JA, Maxted WC, Pahira JJ. Urinalysis: a comprehensive review. Am Fam Physician 2005 Mar 15;71(6):1153-1162. [Medline: 15791892]
4. Gong MM, Sinton D. Turning the page: advancing paper-based microfluidics for broad diagnostic application. Chem Rev 2017 Jun 28;117(12):8447-8480. [doi: 10.1021/acs.chemrev.7b00024] [Medline: 28627178]
5. Reboud J, Xu G, Garrett A, et al. Paper-based microfluidics for DNA diagnostics of malaria in low resource underserved rural communities. Proc Natl Acad Sci U S A 2019 Mar 12;116(11):4834-4842. [doi: 10.1073/pnas.1812296116] [Medline: 30782834]
6. Hong JI, Chang BY. Development of the smartphone-based colorimetry for multi-analyte sensing arrays. Lab Chip 2014 May 21;14(10):1725-1732. [doi: 10.1039/c3lc51451j] [Medline: 24671456]
7. Memişoğulları R, Yüksel H, Yıldırım HA, Yavuz Ö. Performance characteristics of dipstick and microscopic urinalysis for diagnosis of urinary tract infection. Eur J Gen Med 2010;7(2):174-178. [doi: 10.29333/ejgm/82846]
8. Inagaki K. Comparison of judgment results between the color comparison method of urine analysis test papers using the CIEDE2000 color-difference formula and the automatic urine analyzer. J Nurs Sci Eng 2021;8:170-176. [doi: 10.24462/jnse.8.0_170]

XSL•FO
RenderX

9. Alcantarilla PF, Nuevo J, Bartoli A. Fast explicit diffusion for accelerated features in nonlinear scale spaces. Presented at: British Machine Vision Conference 2013; Sep 9-13, 2013; Bristol, United Kingdom. [doi: 10.5244/C.27.13]
10. Alcantarilla PF, Bartoli A, Davison AJ. KAZE features. In: Fitzgibbon A, Lazebnik S, Perona P, Sato Y, Schmid C, editors. Computer Vision - ECCV Lecture Notes in Computer Science: Springer; 2012, Vol. 7577:214-227. [doi: 10.1007/978-3-642-33783-3_16]
11. Urinary Protein Measurement Committee. Report of the committee on renal function (GFR) and urinary protein measurement of the Japanese Society of Nephrology [Article in Japanese]. J Jpn Soc Nephrol 2001;43(1):1-19. [doi: 10.14842/jpnjnephrol1959.43.1]
12. Inagaki K, Nakamura K. Verification of individual differences in visual judgment of urine test strips and proposal of objective evaluation method. Lab Med Int 2024;3(2):42-49. [doi: 10.51041/lmi.3.2_42]
13. Kim SC, Cho YS. Predictive system implementation to improve the accuracy of urine self-diagnosis with smartphones: application of a confusion matrix-based learning model through RGB semiquantitative analysis. Sensors (Basel) 2022 Jul 21;22(14):5445. [doi: 10.3390/s22145445] [Medline: 35891125]
14. Ra M, Muhammad MS, Lim C, Han S, Jung C, Kim WY. Smartphone-based point-of-care urinalysis under variable illumination. IEEE J Transl Eng Health Med 2018;6:2800111. [doi: 10.1109/JTEHM.2017.2765631] [Medline: 29333352]
15. Thakur R, Maheshwari P, Datta SK, Dubey SK, Shakher C. Machine learning-based rapid diagnostic-test reader for albuminuria using smartphone. IEEE Sensors J 2021;21(13):14011-14026. [doi: 10.1109/JSEN.2020.3034904]
16. Erez DL, Derwick H, Furth S, et al. Dipping at home: is it better, easier, and more convenient? A feasibility and acceptability study of a novel home urinalysis using a smartphone application. Pediatr Nephrol 2023 Jan;38(1):139-143. [doi: 10.1007/s00467-022-05556-8] [Medline: 35445981]
17. Stauss M, Dhaygude A, Ponnusamy A, Myers M, Woywodt A. Remote digital urinalysis with smartphone technology as part of remote management of glomerular disease during the SARS-CoV-2 virus pandemic: single-centre experience in 25 patients. Clin Kidney J 2022 May;15(5):903-911. [doi: 10.1093/ckj/sfab286] [Medline: 35498887]
18. Rahmat RF, Muchtar MA, Adnan S, et al. Automated color classification of urine dipstick image in urine examination. J Phys Conf Ser 2018 Mar 1;978(1):012008. [doi: 10.1088/1742-6596/978/1/012008]
19. Africa AD, Velasco J. Development of a urine strip analyzer using artificial neural network using an android phone. ARPN J Eng Appl Sci 2017;12(6):1706-1712.

Abbreviations

AKAZE: Accelerated KAZE
AUC: area under the curve
POC: point-of-care
RGB: red, green, and blue
ROC: receiver operating characteristic
SIFT: scale-invariant feature transform

Edited by A Mavragani; submitted 10.06.24; peer-reviewed by AM Matos, J Favresse; revised version received 12.11.24; accepted 13.11.24; published 14.01.25.

Please cite as:
Inagaki K, Tsuriya D, Hashimoto T, Nakamura K
Verification of the Reliability of an Automated Urine Test Strip Colorimetric Program Using Colorimetric Analysis: Survey Study
JMIR Form Res 2025;9:e62772
URL: https://formative.jmir.org/2025/1/e62772
doi:10.2196/62772

© Keigo Inagaki, Daisuke Tsuriya, Takuya Hashimoto, Katsumasa Nakamura. Originally published in JMIR Formative Research (https://formative.jmir.org), 14.1.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Determinants of Dropping Out of Remote Patient-Reported Outcome–Based Follow-Up Among Patients With Epilepsy: Prospective Cohort Study

Sofie Bech Vestergaard[1,2*], MHSc; Mette Roost[3*], MHSc; David Høyrup Christiansen[2,4,5], MHSc, PhD; Liv Marit Valen Schougaard[2,3], MHSc, PhD

[1]Department of Rheumatology, Aarhus University Hospital, Aarhus N, Denmark
[2]Department of Clinical Medicine, Aarhus University, Aarhus N, Denmark
[3]AmbuFlex, Centre for Patient-Reported Outcomes, Gødstrup Hospital, Herning, Denmark
[4]Centre for Research in Health and Nursing, Research, Regional Hospital Central Jutland, Viborg, Denmark
[5]University Clinic, Centre of Elective Surgery, Regional Hospital Silkeborg, Silkeborg, Denmark
[*]these authors contributed equally

Corresponding Author:
Sofie Bech Vestergaard, MHSc
Department of Rheumatology, Aarhus University Hospital, Aarhus N, Denmark

Abstract

Background: The use of patient-reported outcome (PRO) measures is an emerging field in health care. In the Central Denmark Region, epilepsy outpatients can participate in remote PRO-based follow-up by completing a questionnaire at home instead of attending a traditional outpatient appointment. This approach aims to encourage patient engagement and is used in approximately half of all epilepsy outpatient consultations. However, dropout in this type of follow-up is a challenging issue.

Objective: This study aimed to examine the association between potential self-reported determinants and dropout in remote PRO-based follow-up for patients with epilepsy.

Methods: This prospective cohort study (n=2282) explored the association between dropout in remote PRO-based follow-up for patients with epilepsy and 9 potential determinants covering 3 domains: health-related self-management, general and mental health status, and patient satisfaction. The associations were examined using multiple logistic regression analyses with adjustment for sex, age, education, and cohabitation.

Results: A total of 770 patients (33.7%) dropped out of remote PRO-based follow-up over 5 years. Statistically significant associations were identified between all potential determinants and dropouts in PRO-based follow-up. Patients with low social support had an odds ratio of 2.20 (95% CI 1.38-3.50) for dropout. Patients with poor health ratings had an odds ratio of 2.17 (95% CI 1.65-2.85) for dropout. Similar estimates were identified for the remaining determinants in question.

Conclusions: Patients with reduced self-management, poor health status, and low patient satisfaction had higher odds of dropout in remote PRO-based follow-up. However, further research is needed to determine the reasons for dropout.

(JMIR Form Res 2025;9:e58258) doi:10.2196/58258

KEYWORDS

patient-reported outcome measures; dropouts; digital solutions; outpatient care; epilepsy; seizure disorder; neurological condition; cohort study; health care; Denmark; self-reported; self-management; mental health; patient satisfaction; logistic regression; social support

Introduction

Chronic conditions are a global burden that is increasing in most countries. In 2019, more than 65% of the population in Denmark had one or more chronic conditions [1]. To manage the worldwide development of these chronic conditions, it is crucial that health care resources are being optimally used. The application of telehealth initiatives could be a cost-effective solution for providing high-quality health care [2]. Over the past 50 years, the availability and use of telehealth have expanded and include various tools to manage health care [3]. The World Health Organization (WHO) defines telehealth as *"the delivery of health care services, where distance is a critical factor, by all health care professionals using information and communication technologies for the exchange of valid information for diagnosis, treatment, and prevention of disease and injuries, research and evaluation, and for the continuing education of health care providers, all in the interests of*

XSL•FO
RenderX

advancing the health of individuals and their communities" [4]. After the WHO declared COVID-19 as a pandemic in March 2020, the use of telemedicine has rapidly increased [5,6]. The pandemic forced clinicians to rethink the course of treatment for their patients in various clinical settings to be effectively managed from a distance [6]. One way of using telehealth could be using patient-reported outcomes (PROs) in remote symptom monitoring. This could transform the delivery of the traditional treatment for patients with chronic conditions with prescheduled appointments by optimizing the use of resources and ensuring patient-centered care [7-10].

The use of PROs was originally developed for utilization in research but has, over time, evolved to be utilized at a patient level in clinical practice [11,12]. The use of PROs in clinical practice has been documented to improve patient survival, provide individualized care, and ensure efficient usage of health care resources [13]. In Denmark, a generic clinical PRO system for data collection in clinical practice, called AmbuFlex, offers tailor-made PRO solutions [8]. The overall purpose is to improve patient involvement and quality of treatment while optimizing resource allocation [8,9]. As of September 2023, AmbuFlex has developed 65 PRO solutions for various chronic and malignant conditions in remote outpatient care [8]. The use of PROs in remote outpatient follow-up provides real-time symptom monitoring, early detection of health issues, and rapid clinical intervention [11]. This type of follow-up requires the patient to be willing and capable of taking more responsibility as well as playing an active role in monitoring and identifying symptoms of disease [14]. Patient experiences with this course of treatment are diverse. For patients with epilepsy, it has been documented to promote health-related self-management [15]. Research on patients with rheumatoid arthritis has shown that some patients find remote PRO-based follow-up to be a flexible, time- and resource-saving solution. Others worry about the absence of face-to-face communication with a clinician and find the responsibility to cause insecurity [14].

Dropout and nonuse of telehealth solutions is a common reported challenge [16,17]. A scoping review highlights that the reasons for the nonuse of digital PRO solutions are diverse and, among others, cover the ability to use PRO, engagement, emotional issues, lack of time, and technical barriers [18]. Several contributing factors such as sex [18-21], age [19,22,23], education [18,22], and cohabitation [24-27] have been found to influence the dropout rate in digital solutions. Previous research on PRO-based follow-up for patients with epilepsy has shown that socioeconomically advantaged patients with a high level of health literacy, self-efficacy, patient activation, good well-being, and general health are more likely to be referred to PRO-based follow-up than socioeconomically disadvantaged patients [24]. Whether the same determinants apply to dropout has not been investigated. With the limited amount of literature covering dropout in remote PRO-based follow-up, further research is needed. Thus, this study aimed to investigate potential determinants for dropout in remote PRO-based follow-up among patients with epilepsy. We examined the association between dropout in PRO-based follow-up and determinants from the following three domains: health-related self-management, general and mental health status, and patient satisfaction. We hypothesized that reduced health-related self-management, poor general and mental health status, and low patient satisfaction were associated with an increased risk of dropout in PRO-based follow-up.

Methods

Remote PRO-Based Follow-Up for Patients With Epilepsy

PRO-based follow-up was implemented for patients with epilepsy at 3 neurological departments in the Central Denmark Region in 2012. In PRO-based follow-up, patients complete a questionnaire at home instead of having prescheduled appointments at the outpatient clinic. To attend PRO-based follow-up, patients are individually referred by a health professional. Individual referral requires that the patient must be 15 years of age or older, have no cognitive impairments, and be capable of reading and writing in Danish [28]. The questionnaire is sent at a predefined time interval of 3, 6, 12, or 24 months either electronically through a secure electronic platform or as a paper version [28].

The epilepsy questionnaire includes 47 items and deals with the frequency and intensity of seizures, medical treatment, and the patient's well-being. The AmbuFlex system, based on an automated decision algorithm, can assess whether the patient needs clinical attention [8]. To identify patients in need of clinical attention, all response options in the questionnaire are categorized as red, yellow, or green. Red indicates that the patient needs or wants personal contact with the clinic. Yellow indicates that the patient might need personal contact. If all the patients' answers are categorized as green, the system automatically sends a new questionnaire within the predefined period. All answers that indicate a need for clinical attention are reviewed by a nurse within 14 days. The patients' questionnaire responses were displayed in a graphical overview available in the patients' electronic health records. Where clinical attention is needed, the questionnaire data are used to support patient-clinician communication and clinical decision-making [28]. If a patient wants to discontinue or a clinician decides to end a PRO-based follow-up treatment, the clinician must actively register the patient as "deregistered" in the AmbuFlex system. Deceased patients are automatically deregistered in the system. PRO-based follow-ups are offered to approximately 50% of the entire patient population at the 3 neurological departments and have been used since 2012 [24]. In January 2016, approximately 3000 epilepsy outpatients attended the PRO-based follow-up program at the 3 departments.

Study Population and Setting

We conducted a prospective cohort study among patients with epilepsy who attended PRO-based follow-ups at 3 neurology departments in the Central Denmark Region. During the period of January 1st to December 31st, 2016, patients were invited to answer a research questionnaire in addition to their usual scheduled epilepsy questionnaire from the outpatient clinic. The research questionnaire contained information about aspects related to self-management, general and mental health status, and patient satisfaction. Patients could respond to either a

XSL•FO
RenderX

web-based or paper version of the questionnaire. Nonresponders received up to 3 reminders [29]. Clinicians assessed the usual epilepsy questionnaire according to their normal routine, but they were blinded to the research questionnaire.

Outcome

The outcome was the dropout status in remote PRO-based follow-up in February 2021. This information was retrieved from the AmbuFlex system. As a clinician actively has to register the patient as "deregistered" in the AmbuFlex system, we defined dropouts in this study as "patients who, after using PRO-based follow-up, were deregistered by a clinician." The reason for the dropout was not registered. If the patients were deregistered due to death, this information was shown in the AmbuFlex system. Patients who died during follow-up were excluded from the study.

Potential Determinants

Self-Management

We assessed patient's health-related self-management using 3 different constructs: health literacy, self-efficacy, and patient activation. Health literacy was measured using part of the multi-dimensional Health Literacy Questionnaire (HLQ) [30]. The questionnaire includes 9 subscales covering different areas of health literacy to assess strengths and challenges. The HLQ has well-documented psychometric properties considering the original questionnaire and the Danish translation, with strong content and construct validity as well as high reliability [30,31]. In this study, the following 3 subscales were used: subscale #4, "Social support for health"; subscale #6, "Ability to actively engage with health care providers," and subscale #9, "Understanding health information well enough to know what to do." Subscale #4 includes 5 items with a 4-point ordinal response option ranging from, 1, "strongly disagree," 2, "disagree," 3, "agree," to 4, "strongly agree." Subscales #6 and #9 both include 5 items as well a 5-point ordinal response option ranging from, 1, "cannot do," 2, "very difficult," 3, "quite difficult," 4, "quite easy," to 5, "very easy." For all 3 subscales, a lower score indicates lower health literacy. An average score was calculated for each of the 3 subscales based on a guide provided when acquiring the license. If 1 item was missing, the mean score of the existing items was used to estimate the score. If 2 or more items were missing, the average score was not calculated. As the variables did not exhibit a log-linear relationship, the score was dichotomized into high and low health literacy. For subscale #4, high health literacy was a score >2, and low health literacy was a score ≤2. For subscales #6 and #9, high health literacy was a score >3, and low health literacy was a score ≤3.

Self-efficacy was measured using the General Self-efficacy Scale (GSES) [32]. The psychometric properties have been assessed across many nations and populations, showing a strong construct validity and a high reliability [32,33]. The GSES includes 10 questions regarding one's belief in the ability to cope with stress and challenges. Each question has a 4-point ordinal response option ranging from 1, "not at all true," 2, "hardly true," 3, "moderately true," to 4, "exactly true." The total GSES score ranges from 10 to 40. A lower score indicates a lower degree of self-efficacy [32]. If 1 or more items were missing, the total score was not calculated. The score was included as a continuous variable.

Patient activation was assessed using part of the Patient Activation Measure (PAM) [34]. The original questionnaire includes 13 items that assessed activation in illness or health conditions [35]. In this study, 2 single items were used in a modified form: "I am confident that I can tell when I need to get outpatient care" and "I am confident I can figure out solutions when new situations or problems arise with my health condition." Both items had 4 response options "strongly disagree," "disagree," "agree," and "strongly agree." Both items were dichotomized into strongly disagree/disagree and agree/strongly agree.

General and Mental Health Status

Mental well-being was measured using the WHO-5 Well-being Index (WHO-5) [36]. The WHO-5 includes 5 positively worded items focusing on mental well-being over the previous 2 weeks [36]. The questionnaire has well-documented psychometric properties for various diseases, including epilepsy [28,36,37]. Each item has a 6-point ordinal response option ranging from 5, "All the time," 4, "most of the time," 3, "more than half of the time," 2, "less than half of the time," 1, "some of the time," to 0, "at no time." To calculate the total score, the point for each item was added and afterward multiplied by 4, resulting in a total score between 0 and 100. A score of 100 indicates the best imaginable well-being [36]. If 1 or more items were missing, the total score was not calculated. The score was included as a continuous variable.

The patients' general health status was assessed using a single item from Short Form-36 (SF-36) [38]. The SF-36 covers 8 different domains related to health-related quality of life [39]. It has been documented that a single item from the questionnaire can predict general health [40]. In this study, the single item "In general, would you say your health is: excellent, very good, good, fair, or poor?" was therefore used. The 5 response options were categorized into "Excellent/very good," "good," and "fair/poor."

Patient Satisfaction

The domain of patient satisfaction included patient involvement, confidence, comfort, and satisfaction. Information regarding these determinants was collected using questions inspired by the Danish Cancer Society's Barometer Survey [41] and modified to fit the epilepsy outpatient setting. Information about patient involvement, confidence, and safety was obtained via the following questions: "Do you feel sufficiently involved in your course of treatment?," "Are you confident that the epilepsy outpatient clinic contributes in the best possible way in your course of treatment?," and "Are you comfortable with the follow-up you are receiving?." These questions had five response options: "Yes, to a great extent," "Yes, to some extent," "To a minor extent," "No, not at all," or "Do not know." Information about satisfaction was obtained through: "Overall, how would you assess your treatment in the epilepsy outpatient clinic?" with the following 5 response options: "Very good," "Mostly good," "Mostly bad," "Very bad," or "Do not know,"

XSL•FO
RenderX

For all questions, the category "Do not know" was excluded in the analyses. Patient involvement, confidence, and safety were dichotomized into "Yes, to a great extent/Yes, to some extent" and "To a minor extent/not at all." Satisfaction was dichotomized into "Very good/Mostly good" and "Mostly bad/Very bad."

Statistical Analyses

A Wilcoxon rank-sum test was utilized to compare age and sex between responders and nonresponders. Characteristics for the study population were presented by outcome. Continuous data were all nonnormally distributed and presented as the median and IQR. Categorical and dichotomized data were presented as numbers (n) and percentage distribution (%). The association between the potential determinants and dropouts in PRO-based follow-up was analyzed through multiple logistic regression analysis. All the odds ratio (OR) estimates were tested at a 5% significance level and reported with 95% CIs. Variables were continuously included if the linearity requirements were met. The assumption on linearity was deemed fulfilled for self-efficacy and well-being. Sex, age, education, and cohabitation were included in the adjusted analyses. The confounder variables were defined a priori based on a systematic literature review [18-27]. For the patient satisfaction domain, most of the determinants were controlled only for sex and cohabitation, as some categories had a limited number of observations. Each potential determinant was examined separately as the determinants were assessed as interdependent. All analyses were performed using the available observations giving varying observations for crude and adjusted estimates. An examination of all missing data was conducted prior to this. Upon inspection, no systematic or significant missing data were identified. For the continuous variables, postestimation analyses were performed to nuance the significance of different scores and dropouts in PRO-based follow-up. The statistical analyses were performed using Stata/IC 16.1 (StataCorp LLC).

Ethical Considerations

The study was approved by the Danish Data Protection Agency (number 1-16-02-691-14). The Ethics Committee of the Central Denmark Region was consulted and decided that the study did not require their approval. According to Danish law, only studies involving human biological material require committee approval. Correspondence with the ethics committee is available from the authors upon request. Informed consent for participation in the study was obtained in accordance with guidelines from the Danish Data Protection Agency. Information was provided to patients in a letter along with the questionnaire. Patients were informed that responding to the questionnaire constituted active consent for their participation and the use of their data in the research project. They were also informed that they could withdraw their consent at any time. All data were stored and handled with confidentiality.

Results

Patient Characteristics

From January 1st to December 31st, 2016, a total of 2975 patients received a paper or web-based questionnaire. Of these, 2464 patients answered the questionnaire, resulting in a response rate of 83%. During the follow-up period, 182 patients died. The final study population included 2282 patients (Figure 1). The median age in the study population was 49 (IQR 29.7) years, 1125 (49.3%) were male, 982 (43%) had a low education, and 1645 (72.1%) were not living alone (Table 1). A total of 770 patients (33.7%) had dropped out of PRO-based follow-up in February 2021. No statistical differences were found between responders and nonresponders in sex distribution (P=.34) or age distribution (P=.70). Missing data ranged from 1.5% (general health) to 7.8% (general self-efficacy) (Table 1).

XSL•FO
RenderX

Figure 1. Flowchart of patients included in the study.

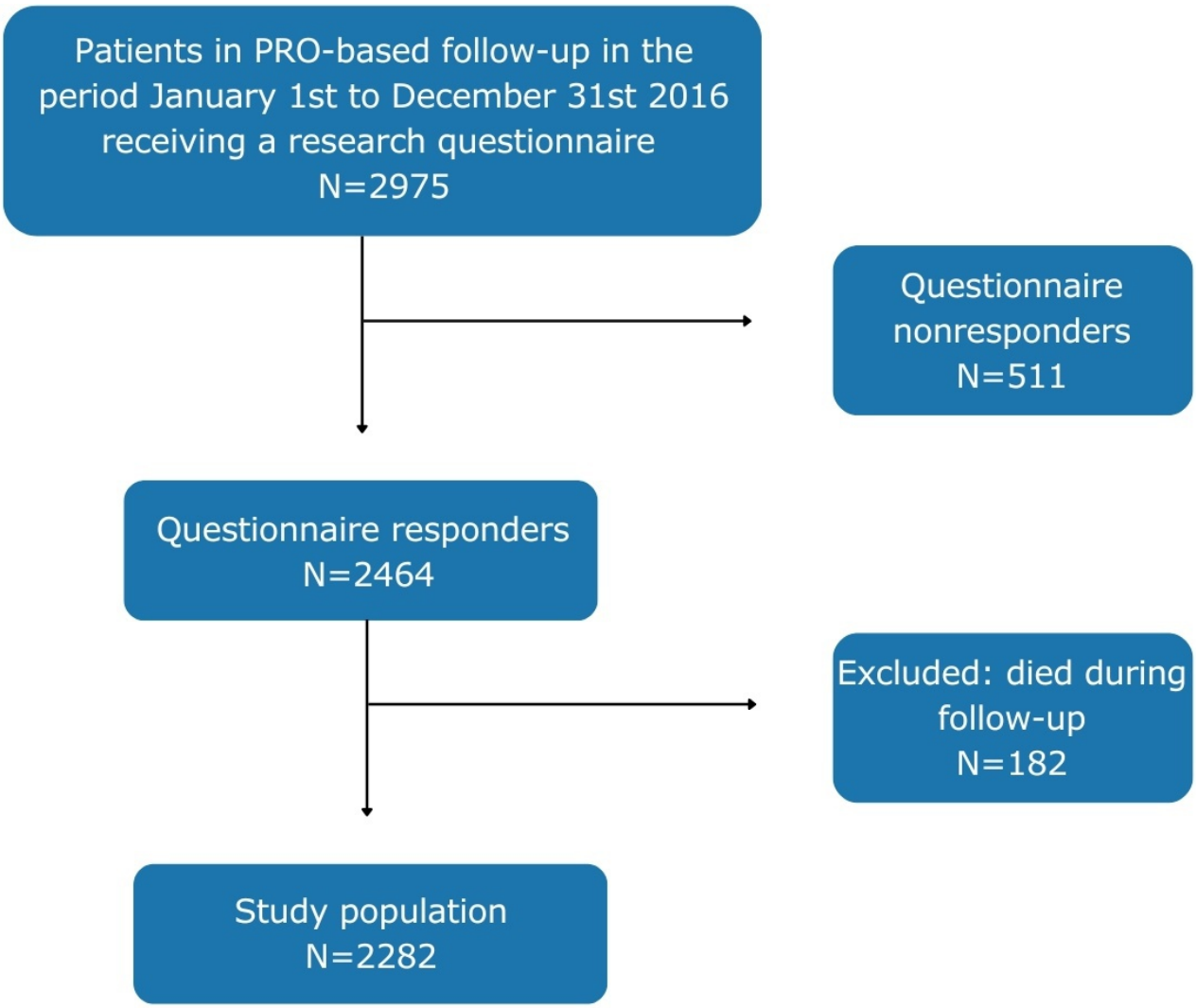

XSL•FO
RenderX

Table . Baseline characteristics and patient-reported potential determinants by status in patient-reported outcome (PRO)–based follow-up among patients with epilepsy in February 2021.

Characteristics	Total, n=2282	Active, n=1512	Dropout, n=770
Sex, n (%)			
Male	1125 (49.3)	757 (50.1)	368 (47.8)
Female	1157 (50.7)	755 (49.9)	402 (52.2)
Age, years, n (%)			
15 - 24	245 (10.7)	128 (8.5)	117 (15.2)
25 - 39	538 (23.6)	340 (22.5)	198 (25.7)
40 - 54	586 (25.7)	407 (26.9)	179 (23.3)
55 - 69	579 (25.4)	421 (27.8)	158 (20.5)
≥70	334 (14.6)	216 (14.3)	118 (15.3)
Education, n (%)			
None	512 (22.4)	317 (21.0)	195 (25.3)
Low	982 (43.0)	674 (44.6)	308 (40.0)
Medium/high	622 (27.3)	426 (28.2)	196 (25.5)
Missing	166 (7.3)	95 (6.3)	71 (9.2)
Cohabitation, n (%)			
Living alone	526 (23.1)	327 (21.6)	199 (25.8)
Not living alone	1645 (72.1)	1126 (74.5)	519 (67.4)
Missing	111 (4.9)	59 (3.9)	52 (6.8)
Health literacy #4[a], n (%)			
Low ≤2	77 (3.4)	37 (2.5)	40 (5.2)
High >2	2049 (89.8)	1395 (92.3)	654 (84.9)
Missing	156 (6.8)	80 (5.3)	76 (9.9)
Health literacy #6[b], n (%)			
Low ≤3	398 (17.4)	222 (14.7)	176 (22.9)
High >3	1728 (75.7)	1209 (80.0)	519 (67.4)
Missing	156 (6.8)	81 (5.4)	75 (9.7)
Health literacy #9[c], n (%)			
Low ≤3	323 (14.2)	172 (11.4)	151 (19.6)
High >3	1796 (78.7)	1253 (82.9)	543 (70.5)
Missing	163 (7.1)	87 (5.8)	76 (9.9)
Self-efficacy			
General Self-efficacy Scale score, median (IQR)	30 (25 - 33)	30 (26 - 34)	29 (24 - 32)
Missing, n (%)	177 (7.8)	91 (6.0)	86 (11.2)
PAM #5[d], n (%)			
Strongly disagree/disagree	221 (9.7)	118 (7.8)	103 (13.4)
Agree/strongly agree	1926 (84.4)	1323 (87.5)	603 (78.3)
Missing	135 (5.9)	71 (4.7)	64 (8.3)
PAM #12[e], n (%)			
Strongly disagree/disagree	342 (15.0)	193 (12.8)	149 (19.4)

XSL•FO
RenderX

Characteristics	Total, n=2282	Active, n=1512	Dropout, n=770
Agree/strongly agree	1797 (78.8)	1246 (82.4)	551 (71.6)
Missing	143 (6.3)	73 (4.8)	70 (9.1)
Mental well-being			
WHO-5 Well-being Index score, median (IQR)	76 (60 - 80)	76 (64 - 80)	72 (52 - 80)
Missing, n (%)	64 (2.8)	35 (2.3)	29 (3.8)
General health[f], n (%)			
Excellent/very good	1030 (45.1)	719 (47.6)	311 (40.4)
Good	886 (38.8)	597 (39.5)	289 (37.5)
Fair/poor	333 (14.6)	181 (12.0)	152 (19.7)
Missing	33 (1.5)	15 (1.0)	18 (2.3)
Patient involvement, n (%)			
A great extent/some extent	1820 (79.8)	1259 (83.3)	561 (72.9)
A lesser extent/not at all	189 (8.3)	102 (6.8)	87 (11.3)
Do not know	140 (6.1)	82 (5.2)	58 (7.5)
Missing	133 (5.8)	69 (4.6)	64 (8.3)
Confidence, n (%)			
A great extent/some extent	1945 (85.2)	1324 (87.6)	621 (80.7)
A minor extent/not at all	99 (4.3)	56 (3.7)	43 (5.6)
Do not know	116 (5.1)	68 (4.5)	48 (6.2)
Missing	122 (5.4)	64 (4.2)	58 (7.5)
Comfort, n (%)			
A great extent/some extent	1896 (83.1)	1307 (86.4)	589 (76.5)
A minor extent/not at all	108 (4.7)	56 (3.7)	52 (6.8)
Do not know	148 (6.5)	80 (5.3)	68 (8.8)
Missing	130 (5.7)	69 (4.6)	61 (7.9)
Satisfaction, n (%)			
Very good/mostly good	1863 (81.6)	1288 (85.2)	575 (74.7)
Mostly bad/very bad	93 (4.1)	49 (3.2)	44 (5.7)
Do not know	197 (8.6)	106 (7.0)	91 (11.8)
Missing	129 (5.7)	69 (4.6)	60 (7.8)

[a]Health Literacy Questionnaire subscale #4 “Social support for health.”

[b]Health Literacy Questionnaire subscale #6 “Ability to actively engage with health care providers.”

[c]Health Literacy Questionnaire subscale #9 “Understanding health information well enough to know what to do.”

[d]Patient Activation Measure item #5: “I am confident that I can tell when I need to get outpatient care.”

[e]Patient Activation Measure item #12: “I am confident I can figure out solutions when new situations or problems arise with my health condition.”

[f]Assessed using a single item from Short Form-36.

Associations Between Self-Management and Dropout in PRO-Based Follow-Up

The ORs for the self-management determinants and dropout in PRO-based follow-up are presented in Table 2. Statistically significant associations were found between all potential self-management determinants and dropout. A lower degree of self-management was associated with dropout. Patients reporting having low social support for health (HLQ4) had higher odds of dropout (OR 2,02, 95% CI 1.38-3.50). Similar associations were identified for the remaining determinants regarding health literacy (Table 2). Patients who had lower confidence in knowing when to get outpatient care (PAM5) had higher odds of dropout (OR 1.94, 95% CI 1.44-2.62). A high level of

XSL•FO
RenderX

self-efficacy had decreased odds for dropout per point (OR 0.96, 95% CI 0.95-0.98). When comparing a patient with the lowest GSES score (score of 10) to a patient with the highest possible GSES score (score of 40), the patient with the low self-efficacy had an OR of 3.21 (95% CI 2.05-5.03) for dropout (Table 3).

Table . Crude and adjusted odds ratio (OR) for potential determinants of self-management and drop-out in patient-reported outcome (PRO)–based follow-up among patients with epilepsy.

Potential determinants	Crude		Adjusted[a]	
	OR (95% CI)	*P* value	OR (95% CI)	*P* value
Social support for health (HLQ 4)	n=2126		n=2099	
High >2	ref		ref	
Low ≤2	2.31(1.46-3.64)	<.001	2.20 (1.38-3.50)[b]	<.001
Ability to actively engage with health care providers (HLQ 6)	n=2126		n=2050	
High >3	ref		ref	
Low ≤3	1.85 (1.48-2.31)	<.001	1.82 (1.44-2.31)	<.001
Low ability to understand health information well enough to know what to do (HLQ9)	n=2119		n=2046	
High >3	ref		ref	
Low ≤3	2.03 (1.59-2.58)	<.001	2.15 (1.65-2.80)	<.001
Self-efficacy	n=2105		n=2032	
General Self-efficacy Scale score [c]	0.96 (0.95-0.98)[d]	<.001	0.96 (0.95-0.98)[d]	<.001
PAM #5	n=2147		n=2066	
Strongly agree/agree	ref		ref	
Disagree/strongly disagree	1.92 (1.45-2.54)	<.001	1.94 (1.44-2.62)	<.001
PAM #12	n=2139		n=2060	
Strongly agree/agree	ref		ref	
Disagree/strongly disagree	1.75 (1.38-2.21)	<.001	1.84 (1.43-2.36)	<.001

[a]Adjusted for sex, age, education, and cohabitation.

[b]Adjusted for sex and cohabitation due to low observations in some categories.

[c]General Self-Efficacy Scale score in the interval 10 - 40.

[d]Per point increase in the General Self-Efficacy Scale score. The score 0 is reference.

Table . Adjusted odds ratio (OR) for dropout in patient-reported outcome (PRO)–based follow-up among patients with epilepsy at different scores relative to a patient with the highest possible score (40) on the General Self-Efficacy Scale.

General Self-efficacy scale score	OR (95% CI)[a]	*P* value
40 points	ref	
30 points	1.47 (1.27-1.71)	<.001
20 points	2.18 (1.61-2.94)	<.001
10 points	3.21 (2.05-5.03)	<.001

[a]OR for dropout relative to a reference with a score on 40 adjusted for sex, age, education, and cohabitation

XSL•FO
RenderX

Associations Between General and Mental Health Status and Dropout in PRO-Based Follow-Up

We found that patients with good general health status had statistically significantly higher odds for dropout (OR 1.28, 95% CI 1.65-2.85) compared to patients who reported excellent/very good health status. Similarly, patients with fair/poor general heath status had statistically significantly higher odds for dropout (OR 2.17, 95% CI 1.65-2.85) compared to patients who reported excellent/very good health status (Table 4). Moreover, patients with lower mental health status had higher odds of dropout. A patient with the lowest possible WHO-5 score (0) had an OR of 3.16 (95% CI 1.94-5.15) for dropout relative to a patient with the highest score (100) (Table 5).

Table . Crude and adjusted Odds ratio (OR) for potential determinants of general and mental health status and drop-out in PRO-based follow-up among patients with epilepsy.

Potential determinants	Crude		Adjusted[a]	
	OR (95% CI)	*P* value	OR (95% CI)	*P* value
Mental well-being	n=2218		n=2063	
WHO-5 Well-being Index score[b]	0.99 (0.98-0.99)[c]	<.001	0.99 (0.98-0.99)[c]	<.001
General health (SF-36)[d]	n=2249		n=2088	
Excellent/very good	ref		ref	
Good	1.12 (0.92-1.36)	.25	1.28 (1.04-1.57)	.02
Fair/poor	1.94 (1.51-2.50)	<.001	2.17 (1.65-2.85)	<.001

[a]Adjusted for sex, age, education, and cohabitation.

[b]WHO-5 Well-being Index score in the interval 0 - 100.

[c]Per point increase in the WHO-5 Well-being Index score. The score 0 is reference.

[d]Using a single item from the Short Form-36.

Table . Adjusted odds ratio (OR) for dropout in patient-reported outcome (PRO)-based follow-up among patients with epilepsy at different scores relative to a patient with the highest possible score (100) on the WHO-5 Well-being Index score.

WHO-5 Well-being Index score	OR (95% CI)[a]	*P* value
0 points	3.16 (1.94-5.15)	<.001
25 points	2.37 (1.64-3.42)	<.001
50 points	1.78 (1.39-2.27)	<.001
75 points	1.33 (1.18-1.51)	<.001

[a]OR for dropout relative to a reference with a score on 100 adjusted for sex, age, education, and cohabitation.

Associations Between Patient Satisfaction and Dropout in PRO-Based Follow-Up

The associations between patient satisfaction and dropout in PRO-based follow-up are presented in Table 6. Patients who, to a minor extent/not at all, felt sufficiently involved in the course of treatment had an OR of 1.93 (95% CI 1.42-2.63) of dropping out compared to patients who felt sufficiently involved. Similar associations were found for the determinants of confidence and comfort. Patients who were not satisfied and found their overall treatment to be mostly bad/very bad had an OR of 2.00 (95% CI: 1.31;3.06) for dropout compared to satisfied patients.

XSL•FO
RenderX

Table . Crude and adjusted odds ratio (OR) for potential determinants for patient satisfaction and drop-out in patient-reported outcome (PRO)–based follow-up among patients with epilepsy.

Potential determinants	Crude		Adjusted[a]	
	OR (95% CI)	*P* value	OR (95% CI)	*P* value
Patient involvement	n=2009		n=1935	
A great extent/some extent	ref		ref	
A minor extent/not at all	1.91 (1.41-2.59)	<.001	1.93(1.42-2.63)[b]	<.001
Confidence	n=2044		n=2021	
A great extent/some extent	ref		ref	
A minor extent/not at all	1.64 (1.09-2.46)	.02	1.62 (1.07-2.44)	.02
Comfort	n=2004		n=1981	
A great extent/some extent	ref		ref	
A minor extent/not at all	2.06 (1.40-3.04)	<.001	2.04 (1.38-3.02)	<.001
Satisfaction	n=1956		n=1934	
Very good/mostly good	ref		ref	
Mostly bad/very bad	2.01(1.32-3.06)	<.001	2.00 (1.31-3.06)	<.001

[a]Adjusted for sex and cohabitation.

[b]Adjusted for sex, education, and cohabitation.

Discussion

Principal Findings and Comparison With Previous Works

The majority of the patients reported a high level of self-management, good health status, and a high degree of patient satisfaction. A strong association was found between all 9 self-reported potential determinants and dropout in remote PRO-based follow-up for patients with epilepsy in the Central Denmark Region. Patients with reduced self-management, poor health status, and low patient satisfaction had higher odds of dropout in PRO-based follow-up. The association appeared independent, as the correction for sex, age, education, and cohabitation showed only a minor impact on the relation between the potential determinants and dropout.

To the best of our knowledge, previous research has not primarily focused on investigating dropout rates; instead, dropout has typically been included as a secondary finding [18,22]. Furthermore, past research has not given exclusive attention to PRO-based follow-up but has focused on telehealth more broadly, encompassing app-based interventions [18,22]. Using a meta-analysis, Meyerowitz-Katz et al determined that app-based chronic disease management interventions exhibit a dropout rate of 49% (95% CI 27%-70%) [22]. A scoping review of 31 studies on patient reasons for not using digital PRO concepts identified a dropout range of 2.4% to 72.3% [18]. Both studies reported higher dropout rates compared to the dropout rate of 33.7% in this study. The studies included follow-up periods with varying response frequencies, ranging from daily to every 3 months [18,22]. This variability in follow-up schedules may have contributed to the higher dropout rates observed in the comparative studies, along with the contextual differences between dropout rates in digital interventions and dropout rates in continuous chronic disease follow-up. Adherence to technology is another concept that can indicate how patients use digital solutions. A recently published systematic review investigated factors associated with adherence to tele-monitoring using electronic PRO measures in patients with chronic diseases, including heart failure, rheumatoid arthritis, and chronic pain [42]. This review found adherence rates to range between 61% and 96% in the included studies [42], which aligns with the dropout rate of approximately 30% found in our study.

Patient characteristics associated with dropout have been investigated mainly in relation to sociodemographic factors, such as sex, age, education, and cohabitation status in other studies [18,22,42]. Self-reported aspects, including self-management and health literacy, have to our knowledge not been reported in other studies regarding the dropout rate in digital interventions. Thus, our study contributes novel knowledge in this field and underlines the importance of considering different individual needs and competencies when implementing digital solutions in clinical care. Nielsen et al also investigated reasons for not using digital PRO concepts in clinical practice among patients with long-term conditions. They found several reasons for digital nonuse of PRO; among others, the themes covered the ability to use PRO data, engagement, emotional distress, and technical barriers [18]. From the patient's point of view, these reasons are important to consider before implementing a digital PRO solution in a clinical setting. Reasons for dropout in PRO-based follow-up will be further

XSL•FO
RenderX

investigated in future research by interviewing patients as well as clinicians.

The clinicians' perspective is important to investigate because dropout could be related not only to digital skills or engagement by patients. From the health care provider's point of view, there may also be concerns that arise during the treatment of patients included in a digital solution. Qualitative research regarding the clinicians' perspective of using PRO-based follow-up in patients with epilepsy has shown reluctance among clinicians [43]. For example, some of the clinicians found the lack of interpersonal contact as a negative consequence of PRO-based follow-up and felt unsure about some of the patients' capabilities to participate even though the patients were already included in PRO-based follow-up [43]. There is no standardized guideline regarding referral to PRO-based follow-up in patients with epilepsy in the Central Denmark Region; thus, the personal preferences amongst clinicians may play a role when deciding to refer or exclude patients from PRO-based follow-up. Prior research investigating factors associated with referral to PRO-based follow-up among patients with epilepsy and patients with rheumatoid arthritis supports this statement since patients with a higher socioeconomic background more often were referred to PRO-based follow-up compared to patients with a lower socioeconomic background [24,44]. Similar to our finding, the study in patients with epilepsy also found a positive association between a high level of health literacy and referral to PRO-based follow-up [24]. These findings in addition to our current study support the importance on focusing on the involvement of vulnerable patient groups when designing digital PRO solutions in clinical practice.

Strengths and Limitations

We conducted a large Danish prospective cohort study among a representative epilepsy population enrolled in remote PRO-based follow-up. Self-reported data were collected before the outcome of interest; thus, potential misclassification would be nondifferential and not bias the results. The outcome (dropout) was register-based, and missing data were acceptable for both self-reported and register-based data. Thus, the internal and external validity of this study are considered sufficient. However, some limitations should be noted. First, we possessed only baseline information about the potential determinants. It is possible that the patients' perceptions of self-management, health status, and patient satisfaction changed during the follow-up period.

Second, the initial response rate in this study was high (83%); however, the only data available regarding the nonrespondents were sex and age. No differences were found between responders and nonresponders, but we cannot rule out that nonresponders did differ from the respondents considering other characteristics. In general, we found high levels of self-management, health status, and satisfaction amongst the responders. Nondifferential selection bias could occur if a lower level of these aspects may have been reported by the nonresponders. Third, an issue occurred with the interpretation of the question regarding the confounder variable education. This was evident for some younger patients who reported having a higher educational level than plausible for their age. It was not possible to determine if this issue was present throughout the study, and there is a possibility that this misclassification may have contributed to insufficient correction for education. Furthermore, despite adjusting for the most common sociodemographic confounders, a risk of residual confounding cannot be completely ruled out. Finally, differentiating the reasons for dropout would have enhanced the strength of this study. However, this information was not available and will be explored in future research.

Conclusions

In this study, we found that reduced self-management, poor health status, and patient dissatisfaction were associated with an increased odds of dropout in remote PRO-based follow-up. Thus, findings from this study support that a remote approach in outpatient follow-up may not be suitable to the whole patient population, but a selected group of patients. However, further research is necessary to explore the reasons for patient dropout. The findings underline the importance of involving patients with different needs and competencies when designing and implementing digital PRO solutions in the health care system.

Acknowledgments

This manuscript has been written without artificial intelligence technology.

Data Availability

The datasets generated and analyzed during this study are available in anonymous form from the corresponding author on request or contact the AmbuFlex-Center for Patient-reported Outcome [45].

Authors' Contributions

Data Analysis: MR and SBV
Supervison: DHC and LMVS
Writing – original draft: MR and SBV
Writing – review & editing: MR, SBV, DHC, and LMVS

References

XSL•FO
RenderX

1. Hvidberg MF, Johnsen SP, Davidsen M, Ehlers L. A nationwide study of prevalence rates and characteristics of 199 chronic conditions in Denmark. Pharmacoecon Open 2020 Jun;4(2):361-380. [doi: 10.1007/s41669-019-0167-7] [Medline: 31342402]
2. Ryu S. Telemedicine: opportunities and developments in member states: report on the second global survey on eHealth 2009 (Global Observatory for eHealth Series, Volume 2). Healthc Inform Res 2012;18(2):153. [doi: 10.4258/hir.2012.18.2.153] [Medline: 22509475]
3. Weinstein RS, Krupinski EA, Doarn CR. Clinical examination component of telemedicine, telehealth, mHealth, and connected health medical practices. Med Clin North Am 2018 May;102(3):533-544. [doi: 10.1016/j.mcna.2018.01.002] [Medline: 29650074]
4. Scalvini S, Vitacca M, Paletta L, Giordano A, Balbi B. Telemedicine: a new frontier for effective healthcare services. Monaldi Arch Chest Dis 2004;PMID(4):226-233. [doi: 10.4081/monaldi.2004.686]
5. Scott A, Bai T, Zhang Y. Association between telehealth use and general practitioner characteristics during COVID-19: findings from a nationally representative survey of Australian doctors. BMJ Open 2021 Mar 24;11(3):e046857. [doi: 10.1136/bmjopen-2020-046857] [Medline: 33762248]
6. Bashshur R, Doarn CR, Frenk JM, Kvedar JC, Woolliscroft JO. Telemedicine and the COVID-19 pandemic, lessons for the Future. Telemed J E Health 2020 May;26(5):571-573. [doi: 10.1089/tmj.2020.29040.rb] [Medline: 32275485]
7. Gandrup J, Ali SM, McBeth J, van der Veer SN, Dixon WG. Remote symptom monitoring integrated into electronic health records: a systematic review. J Am Med Inform Assoc 2020 Nov 1;27(11):1752-1763. [doi: 10.1093/jamia/ocaa177] [Medline: 32968785]
8. Schougaard LMV, Larsen LP, Jessen A, et al. AmbuFlex: tele-patient-reported outcomes (telePRO) as the basis for follow-up in chronic and malignant diseases. Qual Life Res 2016 Mar;25(3):525-534. [doi: 10.1007/s11136-015-1207-0] [Medline: 26790427]
9. Hjollund NHI, Larsen LP, de Thurah AL, et al. Patient-reported outcome (PRO) measurements in chronic and malignant diseases: ten years' experience with PRO-algorithm-based patient-clinician interaction (telePRO) in AmbuFlex. Qual Life Res 2023 Apr;32(4):1053-1067. [doi: 10.1007/s11136-022-03322-9] [Medline: 36639598]
10. Aiyegbusi OL, Hughes SE, Peipert JD, Schougaard LMV, Wilson R, Calvert MJ. Reducing the pressures of outpatient care: the potential role of patient-reported outcomes. J R Soc Med 2023 Feb;116(2):44-64. [doi: 10.1177/01410768231152222] [Medline: 36758615]
11. Calvert M, Kyte D, Price G, Valderas JM, Hjollund NH. Maximising the impact of patient reported outcome assessment for patients and society. BMJ 2019 Jan 24;364:k5267. [doi: 10.1136/bmj.k5267] [Medline: 30679170]
12. Black N. Patient reported outcome measures could help transform healthcare. BMJ 2013 Jan 28;346:f167. [doi: 10.1136/bmj.f167] [Medline: 23358487]
13. Aiyegbusi OL, Nair D, Peipert JD, Schick-Makaroff K, Mucsi I. A narrative review of current evidence supporting the implementation of electronic patient-reported outcome measures in the management of chronic diseases. Ther Adv Chronic Dis 2021;12:20406223211015958. [doi: 10.1177/20406223211015958] [Medline: 34104376]
14. Knudsen LR, de Thurah A, Lomborg K. Experiences with telehealth followup in patients with rheumatoid arthritis: a qualitative interview study. Arthritis Care Res (Hoboken) 2018 Sep;70(9):1366-1372. [doi: 10.1002/acr.23483] [Medline: 29195011]
15. Mejdahl CT, Schougaard LMV, Hjollund NH, Riiskjær E, Thorne S, Lomborg K. PRO-based follow-up as a means of self-management support - an interpretive description of the patient perspective. J Patient Rep Outcomes 2017;2:38. [doi: 10.1186/s41687-018-0067-0] [Medline: 30238083]
16. Eysenbach G. The law of attrition. J Med Internet Res 2005 Mar 31;7(1):e11. [doi: 10.2196/jmir.7.1.e11] [Medline: 15829473]
17. Donkin L, Christensen H, Naismith SL, Neal B, Hickie IB, Glozier N. A systematic review of the impact of adherence on the effectiveness of e-therapies. J Med Internet Res 2011 Aug 5;13(3):e52. [doi: 10.2196/jmir.1772] [Medline: 21821503]
18. Nielsen AS, Kidholm K, Kayser L. Patients' reasons for non-use of digital patient-reported outcome concepts: a scoping review. Health Informatics J 2020 Dec;26(4):2811-2833. [doi: 10.1177/1460458220942649] [Medline: 32731773]
19. Pedersen DH, Mansourvar M, Sortsø C, Schmidt T. Predicting dropouts from an electronic health platform for lifestyle interventions: analysis of methods and predictors. J Med Internet Res 2019 Sep 4;21(9):e13617. [doi: 10.2196/13617] [Medline: 31486409]
20. Multone E, Vader JP, Mottet C, et al. Characteristics of non-responders to self-reported questionnaires in a large inflammatory bowel disease cohort study. Scand J Gastroenterol 2015;50(11):1348-1356. [doi: 10.3109/00365521.2015.1041150] [Medline: 25921677]
21. Højmark K, Støttrup C, Carreon L, Andersen MO. Patient-reported outcome measures unbiased by loss of follow-up. Single-center study based on DaneSpine, the Danish spine surgery registry. Eur Spine J 2016 Jan;25(1):282-286. [doi: 10.1007/s00586-015-4127-3] [Medline: 26208938]
22. Meyerowitz-Katz G, Ravi S, Arnolda L, Feng X, Maberly G, Astell-Burt T. Rates of attrition and dropout in app-based interventions for chronic disease. Syst Rev Meta-Analysis J Med Internet Res 2020;22(9):e20283. [doi: 10.2196/20283] [Medline: 32990635]

XSL•FO
RenderX

23. Portz JD, LaMendola WF. Participation, retention, and utilization of a web-based chronic disease self-management intervention among older adults. Telemed J E Health 2019 Feb;25(2):126-131. [doi: 10.1089/tmj.2017.0208] [Medline: 29782228]
24. Schougaard LMV, de Thurah A, Christensen J, et al. Sociodemographic, personal, and disease-related determinants of referral to patient-reported outcome-based follow-up of remote outpatients: a prospective cohort study. Qual Life Res 2020 May;29(5):1335-1347. [doi: 10.1007/s11136-019-02407-2] [Medline: 31900763]
25. Christensen AI, Ekholm O, Glümer C, et al. The Danish National Health Survey 2010. Study design and respondent characteristics. Scand J Public Health 2012 Jun;40(4):391-397. [doi: 10.1177/1403494812451412] [Medline: 22786925]
26. Christensen AI, Ekholm O, Gray L, Glümer C, Juel K. What is wrong with non-respondents? Alcohol-, drug- and smoking-related mortality and morbidity in a 12-year follow-up study of respondents and non-respondents in the Danish Health and Morbidity Survey. Addiction 2015 Sep;110(9):1505-1512. [doi: 10.1111/add.12939] [Medline: 25845815]
27. Parai C, Hägg O, Willers C, Lind B, Brisby H. Characteristics and predicted outcome of patients lost to follow-up after degenerative lumbar spine surgery. Eur Spine J 2020 Dec;29(12):3063-3073. [doi: 10.1007/s00586-020-06528-y] [Medline: 32666273]
28. Schougaard LMV, de Thurah A, Christiansen DH, Sidenius P, Hjollund NH. Patient-reported outcome (PRO) measure-based algorithm for clinical decision support in epilepsy outpatient follow-up: a test-retest reliability study. BMJ Open 2018 Jul 25;8(7):e021337. [doi: 10.1136/bmjopen-2017-021337] [Medline: 30049693]
29. Schougaard LMV, Mejdahl CT, Christensen J, et al. Patient-initiated versus fixed-interval patient-reported outcome-based follow-up in outpatients with epilepsy: a pragmatic randomized controlled trial. J Patient Rep Outcomes 2019 Sep 13;3(1):61. [doi: 10.1186/s41687-019-0151-0] [Medline: 31520247]
30. Osborne RH, Batterham RW, Elsworth GR, Hawkins M, Buchbinder R. The grounded psychometric development and initial validation of the Health Literacy Questionnaire (HLQ). BMC Public Health 2013 Jul 16;13:658. [doi: 10.1186/1471-2458-13-658] [Medline: 23855504]
31. Maindal HT, Kayser L, Norgaard O, Bo A, Elsworth GR, Osborne RH. Cultural adaptation and validation of the Health Literacy Questionnaire (HLQ): robust nine-dimension Danish language confirmatory factor model. Springerplus 2016;5(1):1232. [doi: 10.1186/s40064-016-2887-9] [Medline: 27536516]
32. Luszczynska A, Scholz U, Schwarzer R. The general self-efficacy scale: multicultural validation studies. J Psychol 2005 Sep;139(5):439-457. [doi: 10.3200/JRLP.139.5.439-457] [Medline: 16285214]
33. Scholz U, Doña BG, Sud S, Schwarzer R. Is general self-efficacy a universal construct? Psychometric findings from 25 countries. Eur J Psychol Assess 2002;18(3):242-251. [doi: 10.1027//1015-5759.18.3.242]
34. Maindal HT, Sokolowski I, Vedsted P. Translation, adaptation and validation of the American short form Patient Activation Measure (PAM13) in a Danish version. BMC Public Health 2009 Jun 29;9:209. [doi: 10.1186/1471-2458-9-209] [Medline: 19563630]
35. Maindal HT, Vedsted P, Mikkelsen EM. Oversættelse og kulturel tilpasning af den danske version af "The Patient Activation Measure". KS 2011 Mar 2;25(1):7-17. [doi: 10.18261/ISSN1903-2285-2011-01-03]
36. Topp CW, Østergaard SD, Søndergaard S, Bech P. The WHO-5 Well-Being Index: a systematic review of the literature. Psychother Psychosom 2015;84(3):167-176. [doi: 10.1159/000376585] [Medline: 25831962]
37. Heun R, Burkart M, Maier W, Bech P. Internal and external validity of the WHO Well-Being Scale in the elderly general population. Acta Psychiatr Scand 1999 Mar;99(3):171-178. [doi: 10.1111/j.1600-0447.1999.tb00973.x] [Medline: 10100911]
38. Bjorner JB, Thunedborg K, Kristensen TS, Modvig J, Bech P. The Danish SF-36 Health Survey: translation and preliminary validity studies. J Clin Epidemiol 1998 Nov;51(11):991-999. [doi: 10.1016/s0895-4356(98)00091-2] [Medline: 9817117]
39. Bjorner JB, Damsgaard MT, Watt T, Groenvold M. Tests of data quality, scaling assumptions, and reliability of the Danish SF-36. J Clin Epidemiol 1998 Nov;51(11):1001-1011. [doi: 10.1016/s0895-4356(98)00092-4] [Medline: 9817118]
40. Macias C, Gold PB, Öngür D, Cohen BM, Panch T. Are single-item global ratings useful for assessing health status? J Clin Psychol Med Settings 2015 Oct 22. [doi: 10.1007/s10880-015-9436-5] [Medline: 26492891]
41. Bekæmpelse K. Kræftpatienters behov og oplevelser med sundhedsvæsenet under udredning og behandling -kræftens bekæmpelses barometerundersøgelse (in Danish). : Danish Cancer Society; 2017 URL: https://www.cancer.dk/om-os/udgivelser-og-rapporter/2017kraeftpatienters-behov-og-oplevelser-med-sundhedsvaesenet-under-udredning-og-behandling-2017/ [accessed 2024-12-19]
42. Wiegel J, Seppen B, van der Leeden M, van der Esch M, de Vries R, Bos W. Adherence to telemonitoring by electronic patient-reported outcome measures in patients with chronic diseases: a systematic review. Int J Environ Res Public Health 2021 Sep 27;18(19):10161. [doi: 10.3390/ijerph181910161] [Medline: 34639463]
43. Mejdahl CT, Schougaard LMV, Hjollund NH, Riiskjær E, Lomborg K. Exploring organisational mechanisms in PRO-based follow-up in routine outpatient care - an interpretive description of the clinician perspective. BMC Health Serv Res 2018 Jul 13;18(1):546. [doi: 10.1186/s12913-018-3352-y] [Medline: 30001729]
44. Schougaard LMV, Knudsen LR, Grove BE, et al. Socioeconomic, disease-related, and personal factors associated with participation in remote follow-up in rheumatoid arthritis: a cross-sectional study. Arthritis Care Res (Hoboken) 2023 Oct;75(10):2044-2053. [doi: 10.1002/acr.25105] [Medline: 36785998]

XSL•FO
RenderX

45. AmbuFlex, center for PRO. Health in the Central Denmark Region. URL: www.ambuflex.dk [accessed 2025-01-13]

Abbreviations

CI: confidence interval
GSES: General Self-Efficacy Scale
HLQ: Health Literacy Questionnaire
IQR: Interquartile range
OR: odds ratio
PAM: patient activation measure
PRO: patient-reported outcome
SF-36: Short Form-36
WHO: World Health Organization
WHO-5: WHO-5 Well-being Index

Edited by A Mavragani, C Hoving; submitted 11.03.24; peer-reviewed by A Hassan, J Mistry; revised version received 04.11.24; accepted 06.11.24; published 14.01.25.

Please cite as:
Vestergaard SB, Roost M, Christiansen DH, Schougaard LMV
Determinants of Dropping Out of Remote Patient-Reported Outcome–Based Follow-Up Among Patients With Epilepsy: Prospective Cohort Study
JMIR Form Res 2025;9:e58258
URL: https://formative.jmir.org/2025/1/e58258
doi:10.2196/58258

© Sofie Bech Vestergaard, Mette Roost, David Høyrup Christiansen, Liv Marit Valen Schougaard. Originally published in JMIR Formative Research (https://formative.jmir.org), 14.1.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Impact of a Virtual Care Navigation Service on Member-Reported Outcomes Among Lesbian, Gay, Bisexual, Transgender, and Queer Populations: Case Study

Seul Ki Choi[1], MPH, PhD; Jaclyn Marshall[2], MS; Patrina Sexton Topper[3], RN, MS, PhD; Andrew Pregnall[4], MSc, MPhil; José Bauermeister[1], MPH, PhD

[1]School of Nursing, University of Pennsylvania, 418 Curie Blvd, Philadelphia, PA, United States
[2]Included Health, San Francisco, CA, United States
[3]College of Nursing, Villanova University, Villanova, PA, United States
[4]Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States

Corresponding Author:
Seul Ki Choi, MPH, PhD
School of Nursing, University of Pennsylvania, 418 Curie Blvd, Philadelphia, PA, United States

Abstract

Background: While the significance of care navigation in facilitating access to health care within the lesbian, gay, bisexual, transgender, queer, and other (LGBTQ+) communities has been acknowledged, there is limited research examining how care navigation influences an individual's ability to understand and access the care they need in real-world settings. By analyzing private sector data, we can bridge the gap between theoretical research findings and practical applications, ultimately informing both business strategies and public policy with evidence grounded in real-world efficacy.

Objective: The objective of this study was to evaluate the impact of specialized virtual care navigation services on LGBTQ+ individuals' ability to comprehend and access necessary care within a national cohort of commercially insured members.

Methods: This case study is based on the experience of commercially insured members, aged 18 or older, who used the LGBTQ+ Health Care Navigation (LGBTQ+ Navigation) service by Included Health between January 26 and July 31, 2023. Care coordinators assisted members by connecting them with vetted identity-affirming in-network providers, helping them navigate and understand their LGBTQ+ health benefits, and providing education and advocacy for clinical and nonclinical needs. We examined the impact of navigation on 5 member-reported outcomes. In addition to reporting the proportion who agreed or strongly agreed, we calculated an impact score that averaged assigned numerical values to all 5 question responses (1=strongly disagree to 5=strongly agree) for each respondent. We used ANOVA with Tukey post hoc tests and *t* tests to explore the relationships between the impact score and member characteristics, including optional self-reported demographics.

Results: Out of 4703 LGBTQ+ Navigation cases, 7.53% (n=354) had member-reported outcomes. A large majority of LGBTQ+ members agreed or strongly agreed that care navigation resulted in less stress (315/354, 89%), less care avoidance (305/354, 86.2%), higher confidence in finding an identity-affirming provider (327/354, 92.4%), improved ability to comprehend health care information (312/354, 88.1%), and improved ability to engage with providers (308/354, 87%). The average impact score was 4.44 (SD 0.69), with statistically significant differences by gender identity (P=.003), race (P=.01), ethnicity (P=.008), and pronouns (P=.02). The scores were highest for members with multiple gender identities (mean 4.56, SD 0.37), and members who did not provide their race, ethnicity, or their pronouns (mean 4.55, SD 0.64). Impact scores were lowest for transgender members (mean 4.11, SD 0.95).

Conclusions: The LGBTQ+ Navigation service, by enhancing members' comprehension and use of necessary care, demonstrates potential public health utility and value. Continuous evaluation of navigation services can serve as a supplementary tool for employers seeking to promote health equity and improve belonging among employees. This is particularly important as discrimination and stigma against LGBTQ+ communities persist in the United States. Therefore, scalable and system-level changes that use navigation services are essential to reach a larger proportion of the LGBTQ+ population.

(JMIR Form Res 2025;9:e64137) doi:10.2196/64137

KEYWORDS

health care navigation; LGBTQ+; lesbian; gay; bisexual; transgender; queer; access to care; care avoidance; identity-affirming care; minority; stress; stigma; health outcomes

XSL•FO
RenderX

Introduction

Compelling evidence indicates that communities comprised of lesbian, gay, bisexual, transgender, queer, and a broader spectrum of sexual orientations and/or gender identities (LGBTQ+) disproportionately endure a range of health issues [1-3]. According to minority stress theory, the stigma associated with sexual and/or gender minority identities acts as a stressor, potentially serving as a foundational contributor to health inequities among LGBTQ+ people [4]. This inequity includes adverse health outcomes, such as various forms of cancer, mental health disorders including depression and anxiety, HIV, and higher rates of substance use compared to their heterosexual and cisgender counterparts [5-10].

Research has identified impeded access to health care services as a key driver of adverse health outcomes within the LGBTQ+ communities, further amplifying health inequities [11-13]. Drawing upon the comprehensive framework for health care access of Levesque et al [14], it is imperative to examine health care accessibility from a multifaceted perspective. This perspective can involve accounting for factors like lack of inclusive policies, limited availability of health insurance, financial challenges that make health care unaffordable, insufficient awareness of available health care services, a sense of mistrust toward health care providers, and poor quality health care related to discrimination based on sexual orientation and gender identity, as well as a lack of health care providers' understanding of LGBTQ+ health care needs [15-24].

Policies and interventions that take this multidimensional view of health care access [14] are essential for improving access to health care, as well as improving access to the evidence-based medical information, all of which play a significant role in facilitating health care access [25]. One such intervention is care navigation, acknowledged as a successful model for increasing access to quality health care, particularly for marginalized populations [26]. Recently, small studies have demonstrated the ability of patient navigation to connect and engage individuals within diverse LGBTQ+ communities with HIV care [27] and facilitate gender-affirming care for transgender populations [28, 29].

While the significance of care navigation in facilitating access to care within the LGBTQ+ communities has been acknowledged [27-29], research in this area is limited. One gap that remains is understanding how care navigation influences an individual's ability to understand and access the care they need in real-world settings. Evaluating these services that are delivered by the private sector offers unique value that complements traditional research studies by providing real-world, actionable insights. Unlike controlled environments in research, private sector evaluations reflect the complexities of actual consumer behaviors, operational challenges, and market conditions. This allows for the observation of how services perform in diverse, dynamic settings where human behavior is less predictable and influenced by a range of contextual factors. Moreover, private sector data often encompasses larger, more varied populations, offering a broader scope of generalizability. By analyzing this data, we can bridge the gap between theoretical research findings and practical applications, ultimately informing both business strategies and public policy with evidence grounded in real-world efficacy. Therefore, the objective of the study was to evaluate the use of an LGBTQ+-specialized virtual care navigation service among Included Health's members and assess the impact of these services on members' abilities to understand and access the care they need among a national, commercially-insured cohort.

Methods

This study follows STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) reporting guidelines for cohort studies.

LGBTQ+ Virtual Health Care Navigation Service

This study is based on the experience of commercially insured members, aged 18 or older, who used the LGBTQ+ Health Care Navigation (LGBTQ+ Navigation) service by Included Health. LGBTQ+ Navigation is Included Health's phone- and chat-based health care navigation service for LGBTQ+ communities. It is offered by self-insured employers as a health care benefit for employees and their dependents and by health plans for their insured members. The LGBTQ+ Navigation service is consistent across all members who are eligible for the service. Employers share information about the navigation service with their employees and their family members similarly to other health benefits, which may include mailings, webinars, and their intranet. Members then proactively initiate a case by completing an intake form on Included Health's website or within the digital health app. A team of care coordinators, 90% of whom identify as LGBTQ+, provide concierge advocacy and guidance for members across 3 areas: connections to vetted identity-affirming in-network providers, benefits navigation focused on LGBTQ+ health, and education and advocacy for clinical and nonclinical needs.

These providers, ranging from primary care physicians to specialists, have been vetted by Included Health to ensure familiarity with LGBTQ+ health needs and inclusive practices. Coordinators, trained in investigative interviewing, gather information by speaking directly to providers and their staff on the provider's experience with population-level health needs, the treatment of LGBTQ+ members, and confirmation of inclusive practices, such as using the correct name and pronouns for patients regardless of their gender assigned at birth. If there are not multiple provider options for the member in the directory, coordinators conduct additional research and vetting to find suitable providers.

Included Health uses a rigorous vetting process to ensure that providers recommended for LGBTQ+ care offer a high standard of cultural competence and inclusivity. Rather than relying on binary yes or no questions or allowing providers to self-identify as inclusive, the care coordinators—who bring specialized training and lived experience—conduct in-depth conversations with providers. Care coordinators are trained over a 4 to 5 week program where they are prepared to address various topics including benefits plans, social determinants of health, family building, and common systemic health conditions faced by

XSL•FO
RenderX

LGBTQ+ individuals. In addition, they receive training in patient advocacy and engage in intuitive listening practices to round out their deep practical knowledge with the skills to uncover needs and empower members. This approach allows coordinators to evaluate providers' practical experience with LGBTQ+ communities and their commitment to delivering culturally affirming, high-quality care. To further ensure consistency in care standards, Included Health actively solicits feedback from LGBTQ+ members who visit these recommended providers. Providers who do not meet member expectations for culturally competent care are reviewed and, if necessary, removed from the directory. This continuous feedback loop, guided by both professional assessment and member experiences, helps to maintain a directory of providers who align with Included Health's standards for LGBTQ+ inclusive care.

Care coordinators also assist members with understanding their benefits related to LGBTQ+ health, such as family expansion and gender-affirming care. For example, members interested in family expansion may have access to adoption support and reproductive endocrinology interventions. In addition, care coordinators assist the member in completing all the steps along gender-affirming care pathways, which may include any necessary documentation such as World Professional Association for Transgender Health (WPATH) letters and access to travel benefits if needed. They are trained to uncover additional needs, allowing them to connect members to benefits they might not know they have. In addition, the care coordinators support members through education and advocacy for clinical and nonclinical needs. This support may include helping the member navigate a claim denial, interpret the workplace policy guidance, identify community resources (eg, parents of transgender children support group), complete required paperwork prior to receiving care, and inform the member on how to change their legal name.

Data Source

Our study sample consisted of all completed service requests (closed cases) of LGBTQ+ Navigation from January 26 to July 31, 2023. We merged data from member intake forms and the postcase member surveys. Members initiated one or more LGBTQ+ Navigation cases by completing an LGBTQ+ Navigation member request form. The form requested information to confirm LGBTQ+ Navigation eligibility, how the member could be supported (eg, find a provider or resource request), and geographic location, as well as optional fields for race, sexual orientation, pronouns, and gender identity. At case closure, the LGBTQ+-specialized care coordinator emailed an optional survey, without an incentive offered for completing the survey, to the member specific to the LGBTQ+ Navigation offering.

Ethical Considerations

The University of Pennsylvania Institutional Review Board (IRB) reviewed the study's protocol and concluded it was exempt as it did not meet the regulatory definition of human subjects research. The committee stated that "the IRB determined that this project qualifies as quality/program improvement meaning that the project does not meet the regulatory definition of human participants research and therefore does not require formal IRB review." When submitting the form to request navigation services, the member agreed to the terms of service which included the Included Health privacy policy, which stated anonymous information may be used for research. Data were deidentified which ensured that the authors did not have access to any identifiable information about members. Compensation was not provided for study participation.

Measures

Member and Case Characteristics

The member's geographic region was defined as Midwest, Northeast, South, or West based on the US census regions. We also included the type of case request as a binary variable (a provider request or seeking resource/support) and whether the member had more than one case. Race, ethnicity, gender identity, sexual orientation, and pronouns were optional fields within the member intake form. The race and ethnicity question allowed respondents to select all that apply, aligning with US Census Bureau recommendations [30]. Due to limited sample size, we derived 2 variables: one indicating whether they identified as White and the other indicating whether they identified as Hispanic. For pronouns, individuals were presented with the option to select from he/him/his, she/her/hers, they/them/theirs, and ze/zir/zirs using a choose-all-that-apply format. Subsequently, we derived a variable indicating whether they chose a single set of pronouns or multiple pronouns.

In the original survey, participants could select all applicable options for gender identity and sexual orientation (see Supplement 2 in Multimedia Appendix 1) [31]. To create mutually exclusive variables for gender identity and sexual orientation, we compiled all the selected options and categorized them into distinct sexual orientations and gender identities. If respondents chose a single identity for either sexual orientation or gender identity, they were merged into a new variable informed by the National Academies' Consensus Report [32]. For both gender identity and sexual orientation, some distinct identities had low sample sizes. In those instances, we combined the respondents into one category. If they selected multiple options for sexual orientation or gender identity, they were placed in a category for multiple options. Consequently, we established mutually exclusive categories for gender identity (cisgender; transgender; other gender identities including agender, gender fluid, intersex, or nonbinary; and multiple gender identities) and sexual orientation (bisexual; gay/lesbian; other sexual orientations including aromantic, asexual, pansexual, queer, or heterosexual; and multiple sexual orientations). For cases in which members did not complete the optional form fields, their member characteristics were presented as not available to indicate missing information. Details on the original responses for gender identity and sexual orientation, as well as how this data was merged into new mutually exclusive categories, can be found in the Supplement 2 in Multimedia Appendix 1.

XSL•FO
RenderX

Member-Reported Outcomes

Our study outcome examined members' evaluation of LGBTQ+ Navigation using a 5-point scale (1=strongly disagree to 5=strongly agree). The questions began, "Since working with Included Health, I am…" and the 5 fragments were (1) less stressed about using health care services, (2) less likely to avoid care, (3) better able to find, understand, and use the information I need to inform my health care choices, (4) better able to find health care providers who understand my needs, and (5) more prepared to engage with health care providers.

In addition to the individual responses, we calculated a composite score for each member to denote the LGBTQ+ Navigation impact. Responses were averaged across the 5 questions (Cronbach α=0.94). A score of 4 or higher was considered an overall positive impact of LGBTQ+ Navigation, as it meant the respondent selected agree (4 points) or strongly agree (5 points) on average, across the 5 separate questions. A score less than 4 was considered a neutral to no impact of LGBTQ+ Navigation, since it indicated that the member was more likely to select neither agree or disagree, disagree, or strongly disagree across the questions.

Statistical Analyses

We used descriptive statistics, including means and distributions, to report the study sample characteristics and the member-reported outcomes. To understand the differences between those with and without member-reported outcomes, we used Fisher exact tests or χ^2 tests for categorical variables and 2-tailed *t* tests for continuous variables. Among the subset of individuals who completed a follow-up survey, we explored the relationships between the LGBTQ+ Navigation impact score and member characteristics. We used ANOVA with Tukey post hoc tests and *t* tests. The α level was defined as .05. All analyses were conducted in SAS version 9.4 (SAS Institute).

Results

There were 4703 LGBTQ+ Navigation cases in the study sample, of which 354 (7.53%) cases were linked to postcase surveys with member-reported outcomes (Table 1). LGBTQ+ Navigation cases with member-reported outcomes were more likely to be for members with multiple cases (202/354, 57.1% vs 1761/4349, 40.49%; *P*<.001) and less likely to be from members who live in the South (98/354, 27.8% vs 1635/4349, 37.97%; *P*<.003) compared to cases without member-reported outcomes. Moreover, members who provided member-reported outcomes were less likely to provide their demographic characteristics, including race and ethnicity (178/354, 50.3% vs 1484/4349, 34.12%), pronouns (178/354, 50.3% vs 1484/4349, 34.12%), gender identity (205/354, 57.9% vs 1869/4349, 42.98%), and sexual orientation (190/354, 53.7% vs 1663/4349, 38.24%), in comparison to those who did not submit member-reported outcomes (*P*<.001). There was not a statistically significant difference in the proportion of cases that requested assistance in finding an identity-affirming provider (350/354, 98.9% vs 4229/4349, 97.24%; *P*=.07).

Among cases linked to member-reported outcomes, half of the respondents chose not to provide their race (178/354, 50.3%) and ethnicity (178/354, 50.3%), pronouns (178/354, 50.3%), gender identity (205/354, 57.9%), or sexual orientation (190/354, 53.7%). For the cases where race, ethnicity, and pronouns were provided (176/354, 49.7%), a majority identified as White (113/176, 64.2%) and non-Hispanic (141/176, 80.1%), and they used singular pronouns (149/176, 84.7%). For the cases where gender identities were provided (149/354, 42.1%), half of the members identified as cisgender (75/149, 50.3%), followed by transgender (33/149, 22.2%); other gender identities including agender, gender fluid, intersex, or nonbinary (30/149, 20.1%); or multiple gender identities (11/149, 7.4%). For the cases where sexual orientations were provided (164/354, 46.3%), 46.3% (76/164) identified as gay or lesbian, followed by multiple sexual orientations (40/164, 24.4%); other sexual orientations including aromantic, asexual, pansexual, queer, or heterosexual (30/164, 18.3%); or bisexual (18/164, 11%).

We found that a large majority of respondents agreed or strongly agreed that LGBTQ+ Navigation had a positive impact on their ability to understand and use the care they needed. Roughly 89% (315/354) felt less stressed about accessing health care services, and 86.2% (305/354) reported they were less likely to avoid care (Table 2). Furthermore, 88.1% (312/354) of respondents agreed or strongly agreed that they improved their ability to find, comprehend, and use the information necessary for making informed health care decisions. Over 92% (327/354) agreed or strongly agreed they had increased confidence in finding health care providers, and 87% (308/354) felt more prepared to engage with these providers.

The average composite LGBTQ+ Navigation impact score was 4.44 (SD=0.69). Approximately 44.6% (158/354) of members had a composite score of 5, indicating they responded "strongly agree" to each of the 5 statements. In addition, 36.2% (128/354) of scores were 4 or more but less than 5, while 19.2% (68/354) were less than 4 (Table 3).

The LGBTQ+ Navigation impact scores varied by case and member characteristics (Table 4). The mean impact scores were higher for members in the South and Midwest (South: 4.69; Midwest: 4.59; Northeast: 4.39; West: 4.20; *P*<.001); members with multiple gender identities (multiple gender identities: 4.56; not available: 4.55; cisgender: 4.36; other gender identities: 4.25; transgender: 4.11; *P*=.003); and members who did not provide their race (not available: 4.55; non-White: 4.40; White: 4.30; *P*=.01), their ethnicity (not available: 4.55; Hispanic: 4.46; non-Hispanic: 4.30; *P*=.008), or their pronouns (not available: 4.55; multiple pronouns 4.36; single set of pronouns: 4.33; *P*=.02). The impact scores did not differ significantly by sexual orientation.

XSL•FO
RenderX

Table . Associations between case and member characteristics and the completion status of member-reported outcomes (n=4703).

Variable	Total (n=4703), n (%)	Member-reported outcomes, n (%)		*P* value
		No (n=4349)	Yes (n=354)	
Region				.003
Midwest	694 (14.9)	635 (14.75)	59 (16.8)	
Northeast	809 (17.37)	738 (17.14)	71 (20.2)	
South	1733 (37.2)	1635 (37.97)	98 (27.8)	
West	1422 (30.53)	1298 (30.14)	124 (35.2)	
Requester service				.07
Provider request	4579 (97.36)	4229 (97.24)	350 (98.9)	
Resources/support	124 (2.64)	120 (2.76)	4 (1.1)	
Member had more than one case				<.001
No	2740 (58.26)	2588 (59.51)	152 (42.9)	
Yes	1963 (41.74)	1761 (40.49)	202 (57.1)	
Race				<.001
Non-White	1302 (27.68)	1239 (28.49)	63 (17.8)	
White	1739 (36.98)	1626 (37.39)	113 (31.9)	
N/A[a]	1662 (35.34)	1484 (34.12)	178 (50.3)	
Ethnicity				<.001
Non-Hispanic	2484 (52.82)	2343 (53.87)	141 (39.8)	
Hispanic	558 (11.86)	523 (12.03)	35 (9.9)	
N/A	1661 (35.52)	1483 (34.1)	178 (50.3)	
Pronouns				<.001
Single pronouns	2725 (57.94)	2576 (59.23)	149 (42.1)	
Multiple pronouns	316 (6.72)	289 (6.65)	27 (7.6)	
N/A	1662 (35.64)	1484 (34.12)	178 (50.3)	
Gender identity				<.001
N/A	2074 (44.1)	1869 (42.98)	205 (57.9)	
Cisgender	1677 (35.66)	1602 (36.84)	75 (21.2)	
Multiple gender identities	86 (1.83)	75 (1.72)	11 (3.1)	
Other[b]	422 (8.97)	392 (9.01)	30 (8.5)	
Transgender	444 (9.44)	411 (9.45)	33 (9.3)	
Sexual orientation				<.001
N/A	1853 (39.4)	1663 (38.24)	190 (53.7)	
Bisexual	276 (5.87)	258 (5.93)	18 (5.1)	
Gay/lesbian	1402 (29.81)	1326 (30.49)	76 (21.5)	
Multiple sexual orientations	497 (10.57)	457 (10.51)	40 (11.3)	
Other[c]	675 (14.35)	645 (14.83)	30 (8.5)	

[a]N/A: not available.

[b]Other gender identities included agender, gender fluid, intersex, or nonbinary.

[c]Other sexual orientations included aromantic, asexual, pansexual, queer, or heterosexual.

XSL•FO
RenderX

Table . Unadjusted average and distribution of responses by member-reported outcomes (n=354).

Response item	Response score, mean (SD)	Responses, n (%)				
		Strongly disagree	Disagree	Neutral	Agree	Strongly agree
Less stress	4.44 (0.78)	4 (1.1)	3 (0.8)	32 (9)	109 (30.8)	206 (58.2)
Less likely to avoid care	4.37 (0.92)	9 (2.5)	7 (2)	33 (9.3)	100 (28.2)	205 (57.9)
Better able to find and use health care information	4.45 (0.76)	2 (0.6)	3 (0.8)	37 (10.4)	102 (28.8)	210 (59.3)
Better able to find provider	4.51 (0.74)	4 (1.1)	3 (0.8)	20 (5.6)	108 (30.5)	219 (61.9)
More prepared to engage with physician	4.43 (0.78)	3 (0.8)	2 (0.6)	41 (11.6)	102 (28.8)	206 (58.2)

Table . Unadjusted composite LGBTQ+[a] Health Care Navigation impact score (n=354).

Composite scores[b]	LGBTQ+ Navigation impact values
Average score, mean (SD)	4.44 (0.69)
Score of less than 4, n (%)	68 (19.2)
Score of 4 or more but less than 5, n (%)	128 (36.2)
Score or 5, n (%)	158 (44.63)

[a]LGBTQ+: lesbian, gay, bisexual, transgender, queer, and others.

[b]Composite scores are the unadjusted average of the 5 individual member-reported outcomes for each member.

XSL•FO
RenderX

Table . Association between LGBTQ+ Health Care Navigation impact scores and case and member characteristics (n=354).

	LGBTQ+ Health care navigation impact score, mean (SD)	*P* value
Region		<.001[b]
Midwest	4.59 (0.52)	
Northeast	4.39 (0.72)	
South	4.69 (0.51)	
West	4.20 (0.79)	
Requester service		.41[c]
Provider request	4.45 (0.67)	
Resources/support	3.6 (1.77)	
Returning user		.054[c]
No	4.36 (0.67)	
Yes	4.50 (0.71)	
Race		.01[b]
Non-White	4.40 (0.74)	
White	4.30 (0.73)	
N/A[d]	4.55 (0.64)	
Ethnicity, n (%)		.008[b]
Non-Hispanic	4.30 (0.74)	
Hispanic	4.46 (0.69)	
N/A	4.55 (0.64)	
Pronouns, n (%)		.02[b]
Single pronouns	4.33 (0.76)	
Multiple pronouns	4.36 (0.51)	
N/A	4.55 (0.64)	
Gender identity, n (%)		.003[b]
N/A	4.55 (0.64)	
Cisgender	4.36 (0.66)	
Multiple gender identities	4.56 (0.37)	
Other[e]	4.25 (0.77)	
Transgender	4.11 (0.95)	
Sexual orientation, n (%)		.05[b]
N/A	4.54 (0.65)	
Bisexual	4.48 (0.59)	
Gay/lesbian	4.31 (0.69)	
Multiple sexual orientations	4.34 (0.68)	
Other[f]	4.27 (0.97)	

[a]LGBTQ+: lesbian, gay, bisexual, transgender, queer, and others.

[b]Determined using ANOVA with Tukey post hoc tests.

[c]Determined using 2-tailed *t* tests.

[d]N/A: not available.

[e]Other gender identities included agender, gender fluid, intersex, or nonbinary.

XSL•FO
RenderX

[f]Other sexual orientations included aromantic, asexual, pansexual, queer, or heterosexual.

Discussion

Principal Findings

To our knowledge, this is the first study to empirically measure the impact of a national, virtual LGBTQ+-specialized care navigation service on member-reported outcomes among commercially insured adults. We found the navigation service, which was considered a comprehensive framework to health care access [14], had a meaningful impact on multiple process outcomes that were associated with the use of preventive services, improved health, and lower long-term costs.

While LGBTQ+ individuals are more likely to receive preventive screenings and have better management of mental health conditions when under the care of an identity-affirming provider [33], the challenge lies in finding LGBTQ+ affirming providers [33,34]. Our study highlighted that members who used LGBTQ+ Navigation reported improved confidence in finding identity-affirming providers and reported being prepared to interact with them. Care avoidance is a significant health care concern as well, leading to the reduced use of preventive services, compromised health outcomes, and elevated long-term health care expenses [35,36]. This is particularly pertinent to the LGBTQ+ communities, which are more prone than other populations to avoid seeking care due to barriers and discrimination [37].

Our study found that 86.2% (305/354) of members reported they were less likely to avoid care after using the LGBTQ+ Navigation service. Moreover, LGBTQ+ Navigation increased members' ability to find, understand, and use health care information, which may lead to potential improvements in medication adherence, increased use of preventive care, reduced emergency room visits, and enhanced management of chronic care [38]. These results echo existing literature discussing the positive impact of LGBTQ+ navigation services on health care access and health care outcomes [28,39,40]. On the other hand, a fifth of members had an LGBTQ+ Navigation impact score that was less than 4, which means they were more likely to respond with neutral, disagree, or strongly disagree to the survey questions. This may indicate the presence of systemic barriers, such as laws, legislation, and insurance policies, that LGBTQ+ Navigation services are unable to address. Further analysis of members who did not respond affirmatively is needed to enhance services and better understand their unmet needs.

We found that care navigation positively impacted all member subgroups, with scores above 4 on the LGBTQ+ Navigation impact, though there were statistically significant differences between the groups. Notably, we found statistically significant differences by gender identity, with the lowest score for transgender individuals. In the context of historical stigma and continual discrimination toward transgender individuals compared to cisgender individuals, the lower scores suggest the need for additional gains toward more inclusive care environments. We also found the highest impact scores were among members who did not provide their demographic information on intake forms. This makes it difficult to interpret differences and highlights an area for future research when more data is available.

Research to date on the impact of care navigation for LGBTQ+ communities has been limited to small samples and focused on subsets of the communities, specifically those receiving care for chronic diseases (eg, HIV) or gender-affirming care [27,28]. This formative evaluation's results are encouraging, and similar tailored care navigation services have the potential to broaden the reach and scalability of health care access, particularly benefiting LGBTQ+ populations who often face challenges in navigating their health care needs. While promising, we recognize that the observed differences between respondents and nonrespondents could introduce biases into our findings, particularly concerning regional disparities and engagement levels. For instance, even though we observed no differences between respondents and nonrespondents in the use of care navigation services, the participants included in our analyses were less likely to be from the South. Moreover, members with multiple cases were more likely to provide feedback. These patterns suggest that certain groups may face barriers to participation, which could skew the overall results. To address this, we recommend targeted outreach efforts aimed at increasing survey participation among underrepresented groups, particularly those in the South and individuals with fewer cases. By implementing these strategies, we aim to create a more balanced dataset in future research, which would ultimately improve the generalizability and inclusivity of the study's conclusions.

Overall, continuous evaluation of these services can serve as a supplementary tool for employers seeking to promote health equity and improve belonging among employees. This is particularly important as discrimination and stigma against these communities persist in the United States. Such services may help reduce barriers by lowering out-of-pocket costs, allowing employees to access care without financial strain, while also addressing both insurance policy limitations and state-level policy constraints [28]. Furthermore, they show promise for enhanced health literacy among employees [41], which may empower them to navigate the health care system and make informed decisions about their health. Therefore, scalable and system-level changes that use navigation services are essential to reaching a larger proportion of the LGBTQ+ population.

Limitations

This study is not without limitations. First, the results may lack generalizability. Included Health was used as a case study, and the members who completed member-reported outcomes differed from those who did not. Second, there may be survey-response bias or a ceiling effect in member-reported outcomes. We recognize that a subset of members provided follow-up feedback through the surveys. Future research examining strategies to encourage increased survey completion rates among members may offer greater insights into patient-reported outcomes related to the Included Health LGBTQ+ Navigation service. Third, while members were asked to report the impact since working with Included Health, member-reported outcomes were solicited at case completion

XSL•FO
RenderX

only and do not reflect changes from a presurvey to postsurvey. Fourth, some members had access to additional Included Health services, such as virtual care, which may have influenced the responses to the survey. However, less than 7% of respondents used virtual care services and the survey came directly from their LGBTQ+ care coordinator after their case was closed, limiting the likelihood that the survey captured their experience with other Included Health services. Fifth, we were unable to differentiate the impact of LGBTQ+ Navigation by the multiple race and ethnicity, sexual orientation, and gender identities captured in the intake form due to small sample sizes and the need for mutually exclusive categories to support statistical testing. However, we felt it was important to examine if any differences were present by subgroup definitions with the data available. By consolidating members into broader race, sexual orientation, and gender identity categories, we were able to maximize data availability and better understand the differences between groups. Finally, most members did not provide optional demographic characteristics, which limited our ability to draw conclusions from the subgroup analyses. Moreover, we were unable to distinguish whether the examined case was for the policy member or for their family members since the data was not initially collected for research purposes. Despite these limitations, the strengths of the study include a national sample of members from the LGBTQ+ communities and an initial assessment of the impact of navigation on key process outcomes that can lead to improved health and reduced costs.

Conclusion

Health care navigation services can offer advocacy and guidance to members by facilitating connections to verified in-network providers, focusing on benefits navigation tailored to LGBTQ+ health needs and providing education and advocacy for both clinical and nonclinical needs. The Included Health LGBTQ+ Navigation service demonstrates potential public health utility and value. Further research is needed to examine the use of care navigation within segments of the LGBTQ+ communities. Moreover, investigating how care navigators addressed specific barriers will inform future navigation opportunities and policy development.

Acknowledgments

The authors are grateful to all the members and care coordinators of Included Health. We also recognize the contributions of staff in Included Health and the University of Pennsylvania. In particular, the authors would like to thank Fiona Weeks, Cass Wahl, and Zachary Taylor at Included Health and James Wolfe and Xandro Xu at Eidos LGBTQ+ Health Initiative, University of Pennsylvania.

Data Availability

The data sets generated and analyzed for this study are not publicly available due to them containing proprietary information of Included Health. The data is confidential and was used under the data use agreement between University of Pennsylvania and Included Health for the current study. Data access can be granted to authorized individuals subject to the company's approval and a nondisclosure agreement.

Authors' Contributions

SKC performed the analysis and led the writing of the manuscript. JM gathered the data and contributed to the conceptualization and drafting of this manuscript. PST contributed to the conceptualization and drafting of this manuscript. AP cleaned and analyzed the data and contributed to the drafting of this manuscript. JB contributed to the conceptualization and drafting of this manuscript. All authors provided feedback and helped shape the research, analysis, and manuscript.

Conflicts of Interest

JM is employed by Included Health. All other authors declare that they do not have competing interests.

Multimedia Appendix 1
Original member responses to optional gender identity and sexual orientation fields within member request form (n=176).
[DOCX File, 20 KB - formative_v9i1e64137_app1.docx]

Checklist 1
STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) checklist of items that should be included in reports of cohort studies.
[DOCX File, 34 KB - formative_v9i1e64137_app2.docx]

References

1. Zeeman L, Sherriff N, Browne K, et al. A review of lesbian, gay, bisexual, trans and intersex (LGBTI) health and healthcare inequalities. Eur J Public Health 2019 Oct 1;29(5):974-980. [doi: 10.1093/eurpub/cky226] [Medline: 30380045]

XSL•FO
RenderX

2. Bariola E, Lyons A, Leonard W, Pitts M, Badcock P, Couch M. Demographic and psychosocial factors associated with psychological distress and resilience among transgender individuals. Am J Public Health 2015 Oct;105(10):2108-2116. [doi: 10.2105/AJPH.2015.302763] [Medline: 26270284]
3. Fredriksen-Goldsen KI, Kim HJ, Barkan SE, Muraco A, Hoy-Ellis CP. Health disparities among lesbian, gay, and bisexual older adults: results from a population-based study. Am J Public Health 2013 Oct;103(10):1802-1809. [doi: 10.2105/AJPH.2012.301110] [Medline: 23763391]
4. Meyer IH. Prejudice, social stress, and mental health in lesbian, gay, and bisexual populations: conceptual issues and research evidence. Psychol Bull 2003 Sep;129(5):674-697. [doi: 10.1037/0033-2909.129.5.674] [Medline: 12956539]
5. Ferlatte O, Salway T, Samji H, et al. An application of syndemic theory to identify drivers of the syphilis epidemic among gay, bisexual, and other men who have sex with men. Sex Transm Dis 2018 Mar;45(3):163-168. [doi: 10.1097/OLQ.0000000000000713] [Medline: 29420444]
6. Kamen C, Palesh O, Gerry AA, et al. Disparities in health risk behavior and psychological distress among gay versus heterosexual male cancer survivors. LGBT Health 2014 Jun;1(2):86-92. [doi: 10.1089/lgbt.2013.0022] [Medline: 26789618]
7. Kamen CS, Smith-Stoner M, Heckler CE, Flannery M, Margolies L. Social support, self-rated health, and lesbian, gay, bisexual, and transgender identity disclosure to cancer care providers. Oncol Nurs Forum 2015 Jan;42(1):44-51. [doi: 10.1188/15.ONF.44-51] [Medline: 25542320]
8. Chan ASW, Leung LM, Wong FKC, et al. Needs and experiences of cancer care in patients' perspectives among the lesbian, gay, bisexual, transgender and queer community: a systematic review. Soc Work Health Care 2023;62(8-9):263-279. [doi: 10.1080/00981389.2023.2226182] [Medline: 37338823]
9. Hsieh N, Shuster SM. Health and health care of sexual and gender minorities. J Health Soc Behav 2021 Sep;62(3):318-333. [doi: 10.1177/00221465211016436] [Medline: 34528481]
10. National Academies of Sciences, Engineering, and Medicine. Understanding the Well-Being of LGBTQI+ Populations: National Academies Press; 2020.
11. Dodge B, Herbenick D, Friedman MR, et al. Attitudes toward bisexual men and women among a nationally representative probability sample of adults in the United States. PLoS ONE 2016;11(10):e0164430. [doi: 10.1371/journal.pone.0164430] [Medline: 27783644]
12. Grant J, Mottet L, Tanis J. National transgender discrimination survey report on health and health care. National LGBT Cancer Network. 2010. URL: https://cancer-network.org/wp-content/uploads/2017/02/National_Transgender_Discrimination_Survey_Report_on_health_and_health_care.pdf [accessed 2025-01-02]
13. Lund EM, Burgess CM. Sexual and gender minority health care disparities: barriers to care and strategies to bridge the gap. Prim Care 2021 Jun;48(2):179-189. [doi: 10.1016/j.pop.2021.02.007] [Medline: 33985698]
14. Levesque JF, Harris MF, Russell G. Patient-centred access to health care: conceptualising access at the interface of health systems and populations. Int J Equity Health 2013 Mar 11;12:18. [doi: 10.1186/1475-9276-12-18] [Medline: 23496984]
15. Cruz TM. Assessing access to care for transgender and gender nonconforming people: a consideration of diversity in combating discrimination. Soc Sci Med 2014 Jun;110:65-73. [doi: 10.1016/j.socscimed.2014.03.032] [Medline: 24727533]
16. Garbers S, Heck CJ, Gold MA, Santelli JS, Bersamin M. Providing culturally competent care for LGBTQ youth in school-based health centers: a needs assessment to guide quality of care improvements. J Sch Nurs 2018 Dec;34(6):424-429. [doi: 10.1177/1059840517727335] [Medline: 28830333]
17. Conron KJ, Mimiaga MJ, Landers SJ. A population-based study of sexual orientation identity and gender differences in adult health. Am J Public Health 2010 Oct;100(10):1953-1960. [doi: 10.2105/AJPH.2009.174169] [Medline: 20516373]
18. Durso LE, Meyer IH. Patterns and predictors of disclosure of sexual orientation to healthcare providers among lesbians, gay men, and bisexuals. Sex Res Social Policy 2013 Mar 1;10(1):35-42. [doi: 10.1007/s13178-012-0105-2] [Medline: 23463442]
19. Holland D, White LCJ, Pantelic M, Llewellyn C. The experiences of transgender and nonbinary adults in primary care: a systematic review. Eur J Gen Pract 2024 Dec;30(1):2296571. [doi: 10.1080/13814788.2023.2296571] [Medline: 38197305]
20. Chong LSH, Kerklaan J, Clarke S, et al. Experiences and perspectives of transgender youths in accessing health care: a systematic review. JAMA Pediatr 2021 Nov 1;175(11):1159-1173. [doi: 10.1001/jamapediatrics.2021.2061] [Medline: 34279538]
21. Wittgens C, Fischer MM, Buspavanich P, Theobald S, Schweizer K, Trautmann S. Mental health in people with minority sexual orientations: a meta-analysis of population-based studies. Acta Psychiatr Scand 2022 Apr;145(4):357-372. [doi: 10.1111/acps.13405] [Medline: 35090051]
22. Yu H, Flores DD, Bonett S, Bauermeister JA. LGBTQ + cultural competency training for health professionals: a systematic review. BMC Med Educ 2023 Aug 9;23(1):558. [doi: 10.1186/s12909-023-04373-3] [Medline: 37559033]
23. Holt NR, Eldridge-Smith ED, Griffin JA, Stepleman LM, Drescher CF, Casanova T. Differences in health care access, utilization, and experiences among LGBTQ+ subgroups in the Southern United States. Fam Community Health 2023;46(1):58-68. [doi: 10.1097/FCH.0000000000000340] [Medline: 35943219]
24. Hatzenbuehler ML, Lattanner MR, McKetta S, Pachankis JE. Structural stigma and LGBTQ+ health: a narrative review of quantitative studies. Lancet Public Health 2024 Feb;9(2):e109-e127. [doi: 10.1016/S2468-2667(23)00312-2] [Medline: 38307678]

XSL•FO
RenderX

25. Butkus R, Rapp K, Cooney TG, Engel LS, Health and Public Policy Committee of the American College of Physicians. Envisioning a better U.S. health care system for all: reducing barriers to care and addressing social determinants of health. Ann Intern Med 2020 Jan 21;172(2 Suppl):S50-S59. [doi: 10.7326/M19-2410] [Medline: 31958803]
26. Johnson CV, Mimiaga MJ, Bradford J. Health care issues among lesbian, gay, bisexual, transgender and intersex (LGBTI) populations in the United States: introduction. J Homosex 2008;54(3):213-224. [doi: 10.1080/00918360801982025] [Medline: 18825859]
27. Roland KB, Higa DH, Leighton CA, Mizuno Y, DeLuca JB, Koenig LJ. Client perspectives and experiences with HIV patient navigation in the United States: a qualitative meta-synthesis. Health Promot Pract 2020 Jan;21(1):25-36. [doi: 10.1177/1524839919875727] [Medline: 31597497]
28. Dozier R. How navigators influence insurance coverage for gender-affirming surgeries: a qualitative study. Perm J 2023 Mar 15;27(1):72-76. [doi: 10.7812/TPP/22.115] [Medline: 36916175]
29. Zamudio-Haas S, Koester K, Venegas L, et al. “Entre Nosotras:” a qualitative study of a peer-led PrEP project for transgender latinas. BMC Health Serv Res 2023 Sep 20;23(1):1013. [doi: 10.1186/s12913-023-09707-x] [Medline: 37730598]
30. Matthews K, Phelan J, Jones M, et al. 2015 national content test race and ethnicity analysis report. US Census Bureau.: U.S. Department of Commerce; 2017. URL: https://www.documentcloud.org/documents/4316468-2015nct-Race-Ethnicity-Analysis [accessed 2023-12-01]
31. Bauermeister J, Choi SK, Bruehlman-Senecal E, et al. An identity-affirming web application to help sexual and gender minority youth cope with minority stress: pilot randomized controlled trial. J Med Internet Res 2022 Aug 1;24(8):e39094. [doi: 10.2196/39094] [Medline: 35916700]
32. National Academies of Sciences, Engineering, and Medicine. Measuring Sex, Gender Identity, and Sexual Orientation: National Academies Press; 2022. [doi: 10.17226/26424]
33. McKay T, Tran NM, Barbee H, Min JK. Association of affirming care with chronic disease and preventive care outcomes among lesbian, gay, bisexual, transgender, and queer older adults. Am J Prev Med 2023 Mar;64(3):305-314. [doi: 10.1016/j.amepre.2022.09.025] [Medline: 36460525]
34. Khalili J, Leung LB, Diamant AL. Finding the perfect doctor: identifying lesbian, gay, bisexual, and transgender–competent physicians. Am J Public Health 2015 Jun;105(6):1114-1119. [doi: 10.2105/AJPH.2014.302448] [Medline: 25880937]
35. Hostetter J, Schwarz N, Klug M, Wynne J, Basson MD. Primary care visits increase utilization of evidence-based preventative health measures. BMC Fam Pract 2020 Jul 28;21(1):151. [doi: 10.1186/s12875-020-01216-8] [Medline: 32718313]
36. Gao J, Moran E, Grimm R, Toporek A, Ruser C. The effect of primary care visits on total patient care cost: evidence from the Veterans Health Administration. J Prim Care Community Health 2022;13:21501319221141792. [doi: 10.1177/21501319221141792] [Medline: 36564889]
37. Finneran P, Toribio MP, Natarajan P, Honigberg MC. Delays in accessing healthcare across the gender spectrum in the All of Us research program. J Gen Intern Med 2024 May;39(7):1156-1163. [doi: 10.1007/s11606-023-08548-y] [Medline: 38049655]
38. U.S. Department of Health and Human Services, Office of Disease Prevention and Health Promotion. National action plan to improve health literacy. US Department of Health and Human Services. 2010. URL: https://health.gov/sites/default/files/2019-09/Health_Literacy_Action_Plan.pdf [accessed 2023-12-14]
39. Doblecki-Lewis S, Butts S, Botero V, Klose K, Cardenas G, Feaster D. A randomized study of passive versus active PrEP patient navigation for a heterogeneous population at risk for HIV in South Florida. J Int Assoc Provid AIDS Care 2019;18:2325958219848848. [doi: 10.1177/2325958219848848] [Medline: 31131679]
40. Hostetter CR, Call J, Gerke DR, Holloway BT, Walls NE, Greenfield JC. “We are doing the absolute most that we can, and no one is listening”: barriers and facilitators to health literacy within transgender and nonbinary communities. Int J Environ Res Public Health 2022 Jan 22;19(3):1229. [doi: 10.3390/ijerph19031229] [Medline: 35162254]
41. Loignon C, Dupéré S, Fortin M, Ramsden VR, Truchon K. Health literacy - engaging the community in the co-creation of meaningful health navigation services: a study protocol. BMC Health Serv Res 2018 Jun 28;18(1):505. [doi: 10.1186/s12913-018-3315-3] [Medline: 29954407]

Abbreviations

IRB: Institutional Review Board
LGBTQ+: lesbian, gay, bisexual, transgender, queer, and others
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
WPATH: World Professional Association for Transgender Health

XSL•FO
RenderX

Edited by A Mavragani; submitted 09.07.24; peer-reviewed by DZ Nowaskie, M Cabell Jonas; revised version received 02.12.24; accepted 02.12.24; published 09.01.25.

Please cite as:
Choi SK, Marshall J, Sexton Topper P, Pregnall A, Bauermeister J
Impact of a Virtual Care Navigation Service on Member-Reported Outcomes Among Lesbian, Gay, Bisexual, Transgender, and Queer Populations: Case Study
JMIR Form Res 2025;9:e64137
URL: https://formative.jmir.org/2025/1/e64137
doi:10.2196/64137

© Seul Ki Choi, Jaclyn Marshall, Patrina Sexton Topper, Andrew Pregnall, José Bauermeister. Originally published in JMIR Formative Research (https://formative.jmir.org), 9.1.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

A Mobile App for Promoting Breastfeeding-Friendly Communities in Hong Kong: Design and Development Study

Heidi Sze Lok Fan[1], BNurs, MPhil, PhD; Emily Tsz Yan Leung[2], BA, BEd; Ka Wing Lau[2], BSc; Janet Yuen Ha Wong[3], BNurs, MNurs, PhD; Edmond Pui Hang Choi[2], BNurs, MPhil, PhD; Christine Lam[4], RM, BHS, MMid, IBCLC; Marie Tarrant[1], BN, MN, MPH, PhD; Hextan Yuen Sheung Ngan[5], MBBS, MD; Patrick Ip[6], MBBS, DCH, Dip Med; Chia Chin Lin[2], BSN, MSN, EMBA, PhD; Kris Yuet Wan Lok[2], BSc, MSc, PhD

[1]School of Nursing, Faculty of Health and Social Development, The University of British Columbia, Kelowna, BC, Canada

[2]School of Nursing, Li Ka Shing Faculty of Medicine, University of Hong Kong, 5/F, Academic Building, Pokfulam, Hong Kong, China (Hong Kong)

[3]School of Nursing and Health Sciences, Hong Kong Metropolitan University, Hong Kong, China (Hong Kong)

[4]Department of Obstetrics and Gynaecology, Queen Elizabeth Hospital, Kowloon, Hong Kong, China (Hong Kong)

[5]Department of Obstetrics and Gynaecology, Li Ka Shing Faculty of Medicine, University of Hong Kong, Hong Kong, China (Hong Kong)

[6]Li Ka Shing Faculty of Medicine, Department of Paediatrics and Adolescent Medicine, University of Hong Kong, Hong Kong, China (Hong Kong)

Corresponding Author:
Kris Yuet Wan Lok, BSc, MSc, PhD
School of Nursing, Li Ka Shing Faculty of Medicine, University of Hong Kong, 5/F, Academic Building, Pokfulam, Hong Kong, China (Hong Kong)

Abstract

Background: Breastfeeding is vital for the health and well-being of both mothers and infants, and it is crucial to create supportive environments that promote and maintain breastfeeding practices.

Objective: The objective of this paper was to describe the development of a breastfeeding-friendly app called “bfGPS” (HKU TALIC), which provides comprehensive territory-wide information on breastfeeding facilities in Hong Kong, with the goal of fostering a breastfeeding-friendly community.

Methods: The development of bfGPS can be categorized into three phases, which are (1) planning, prototype development, and preimplementation evaluation; (2) implementation and updates; and (3) usability evaluation. In phase 1, a meeting was held with experts, including maternal and child health researchers, app developers, breastfeeding individuals, and health professionals, to discuss the focus and functionality of the breastfeeding app. A prototype was developed, and breastfeeding facilities in various public venues in Hong Kong were assessed using a structured checklist. For the preimplementation evaluation, 10 focus groups and 19 one-on-one interviews were conducted between May 2019 and October 2020 with staff working in public premises (n=29) and breastfeeding individuals (n=29). For phase 2, bfGPS was published on iOS (Apple Inc) and Android (Google) platforms in September 2020. App updates were launched in September 2021 and May 2022 based on the suggestions provided by the participants in the preimplementation evaluation. For the usability evaluation, semistructured, in-depth, one-to-one interviews were conducted with breastfeeding individuals (n=30) to understand their experiences of using bfGPS. Content analysis was used to analyze the data.

Results: bfGPS is a mobile app that was developed to assist breastfeeding individuals in locating breastfeeding facilities in public venues in Hong Kong. In the preimplementation evaluation, the participants gave comments on the layout and interface of bfGPS, and suggestions were given on incorporating new functions into the app. Based on the suggestions of the participants in the preimplementation evaluation, a few additional functions were added into bfGPS, including allowing the users to rate and upload recent information about breastfeeding facilities and an infant tracker function that encourages users to record infant development. In the usability evaluation, 3 main themes emerged—bfGPS improves the community experience for breastfeeding individuals, facilitates tracking the infant’s growth, and provides suggestions for further development.

Conclusions: The bfGPS app is the first user-friendly tool designed to assist users in locating breastfeeding facilities within the community. It stands as a guide for similar health care app developments, emphasizing the importance of accurate, current data to ensure user adoption and long-term use. The app’s potential lies in the support and reinforcement of breastfeeding practices coupled with self-management strategies.

(JMIR Form Res 2025;9:e64191) doi:10.2196/64191

XSL•FO
RenderX

KEYWORDS

Baby-Friendly Community Initiative; Baby-Friendly Hospital Initiative; breastfeeding; community; stakeholders; mobile app; friendly communities; baby-friendly; well-being; mother; infant; application; mHealth; qualitative; user-friendly; self-management

Introduction

Background

Breastfeeding provides numerous benefits to both parents and infants [1]. It is considered one of the most effective health interventions for preventing infant and child mortality [2]. The World Health Organization (WHO) recommends that infants be exclusively breastfed for the first 6 months of life. After 6 months of age, complementary feeding should be provided with continuing breastfeeding up to 2 years of age or beyond [3]. However, breastfeeding individuals face numerous challenges that can lead to early weaning. Less than one-half of infants under 6 months old were exclusively breastfed globally, and only 45% of infants can be breastfed up to 2 years of age [4]. The determinants of breastfeeding continuation are multifactorial, with various demographic [5,6], psychological [7], and social factors [8].

Breastfeeding support is crucial for long-term breastfeeding. Technology-based interventions effectively improve breastfeeding outcomes [9,10]. However, a meta-analysis of randomized controlled trials found that mobile app usage was not significantly associated with prolonged breastfeeding duration. However, only a few studies were included in the meta-analysis, and there was high heterogeneity among the interventions [11]. Recently published randomized controlled trials showed that the use of mobile apps can significantly increase [12] exclusive breastfeeding duration [13].

The Baby-Friendly Hospital Initiative (BFHI) was launched by the WHO and the United Nations International Children's Emergency Fund in 1991 [14]. A large body of evidence shows that the BFHI can significantly improve breastfeeding outcomes [15,16]. The BFHI comprises 10 steps that aim to promote, protect, and support breastfeeding in a hospital setting [14]. Step 10 of the BFHI is about fostering the establishment of breastfeeding support groups in the community and referring breastfeeding individuals to them on discharge from the hospital or clinic [14]. Community support is essential for extending breastfeeding duration [17]. The Baby-Friendly Community Initiative (BFCI) is an extension of the BFHI [18] and is focused on providing community-based support for breastfeeding individuals [19]. However, the BFCI is implemented in only a few countries and resources from the government are needed to support the BFCI [19].

Globally, breastfeeding in public is not yet normalized. In Western countries, the proportion of individuals supporting breastfeeding in public ranges from 50% to 75% [20-23]. In China, although 65% of the community agrees that breastfeeding in public is acceptable, 30% believe that women should breastfeed only in private spaces [24]. Furthermore, breastfeeding individuals face numerous challenges when breastfeeding in public. Embarrassment has been identified as a key barrier to public breastfeeding [25] and unwanted attention and unsuitable structural environments are common challenges faced outside of the home [26]. Studies show that some public areas are unwelcoming to breastfeeding individuals [27], highlighting the lack of comfortable breastfeeding spaces [28]. A study has shown that breastfeeding in public is associated with a longer duration of breastfeeding [29]. It is crucial for breastfeeding individuals to find breastfeeding-friendly places when they visit public venues. Several mobile apps are available for breastfeeding individuals to locate breastfeeding facilities in their communities, such as "FeedFinder (GarbTech)," primarily used in the United Kingdom [30] and "Moommae (Moon Active)" in Thailand [31]. These apps can increase the confidence of breastfeeding individuals to go out and breastfeed [30] while supporting breastfeeding on demand [32]. However, users have emphasized the need for more breastfeeding facilities to be included in the app [30,32]. Therefore, we developed a social media mobile app with comprehensive territory-wide information on breastfeeding facilities in Hong Kong to promote breastfeeding-friendly attitudes within the community.

Breastfeeding Practices and Public Breastfeeding in Hong Kong

In Hong Kong, approximately 84.7% of postpartum individuals initiate breastfeeding [33]. However, only 23.9% of parents are able to continue to breastfeed for 12 months and only 22.2% exclusively breastfeed until 6 months [34]. Breastfeeding individuals face various barriers in public places. Frequently reported unpleasant experiences included being stared at by others and being asked to breastfeed in unsanitary or unsuitable places, such as public restrooms [35]. A significant number of women stated that they reduced going out with their children and resorted to feeding them with infant formula to avoid public breastfeeding [35]. However, given that supplementing with infant formula is strongly associated with breastfeeding cessation [36,37], this is not an optimal solution. One study demonstrated breastfeeding in public was associated with longer breastfeeding duration [29]. Therefore, it is crucial to offer community support.

The objective of this paper was to describe the development of a breastfeeding-friendly app called "bfGPS (HKU TALIC)," which provides comprehensive territory-wide information on breastfeeding facilities in Hong Kong, with the goal of fostering a breastfeeding-friendly community. We also gain insights into how breastfeeding individuals and public venue staff perceive bfGPS through qualitative interviews.

Methods

Design

This study was conducted as part of the BFCI, which aims to promote breastfeeding-friendly attitudes in the community. The BFCI program encompasses the following activities: (1) providing training to the staff and management of public venues to support breastfeeding individuals, (2) offering breastfeeding support to pregnant and breastfeeding individuals, and (3)

XSL•FO
RenderX

developing the bfGPS mobile app that aims to assist breastfeeding individuals in locating breastfeeding facilities in public venues in Hong Kong. The details of the BFCI program were described elsewhere [38].

In this study, we report the development of bfGPS, which is divided into three phases: (1) planning, prototype development, and preimplementation evaluation; (2) implementation and updates; and (3) usability evaluation (Figure 1). Phase 1 involves facilities assessment, the development of the prototype of bfGPS, and a preimplementation evaluation. The preimplementation evaluation involves focus groups and individual interviews to collect feedback on the app. For phase 2 of the development, we published bfGPS on iOS (Apple Inc) and Android (Google) platforms, and we updated the apps to add additional functions, which are rating of breastfeeding facilities and infant tracker. Phase 3 is a usability evaluation. It includes qualitative interviews with breastfeeding individuals to understand their experiences of using bfGPS.

Figure 1. Flow diagram of the development of bfGPS.

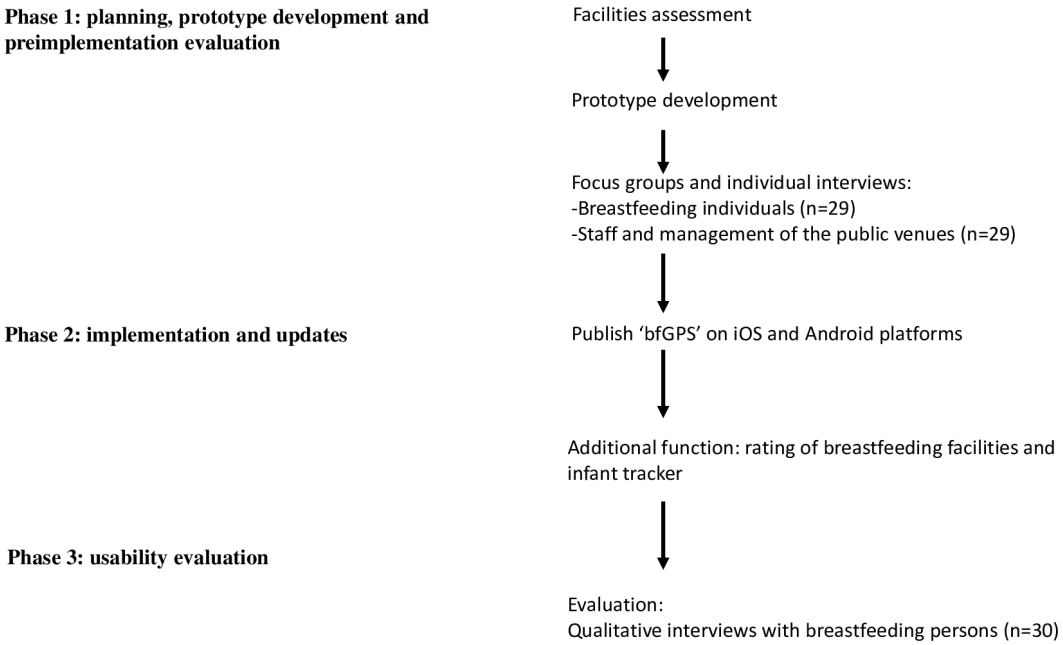

Phase 1: Planning, Prototyping, and Preimplementation Evaluation

Setting and Relevant Context

A meeting was held with experts, including maternal and child health researchers, app developers, breastfeeding individuals, and health professionals, to discuss the focus and functionality of the breastfeeding app. Based on the results of the Department of Health (2016) survey [35], the team decided to design an app that could offer user-friendly guidance on breastfeeding to the general public.

The content of the app was outlined, and a prototype of the bfGPS app was developed (Figure 2). bfGPS is available in English, Traditional Chinese, and Simplified Chinese. Between June 2019 and June 2020, undergraduate nursing students assessed breastfeeding facilities in various public venues in Hong Kong, such as government premises, shopping malls, and metro stations, using a structured checklist.

XSL•FO
RenderX

Figure 2. The interface of bfGPS for locating breastfeeding facilities is available in (**A**) English, (**B**) traditional Chinese, and (**C**) simplified Chinese. bfGPS provides a search bar that enables users to search breastfeeding facilities across Hong Kong. Users can view nearby facilities and the distance from their current location by enabling location access. The details of nearby facilities are shown with icons and descriptions. In addition, bfGPS allows for customized searches to enable users to filter results based on their needs, such as private rooms or handwashing facilities.

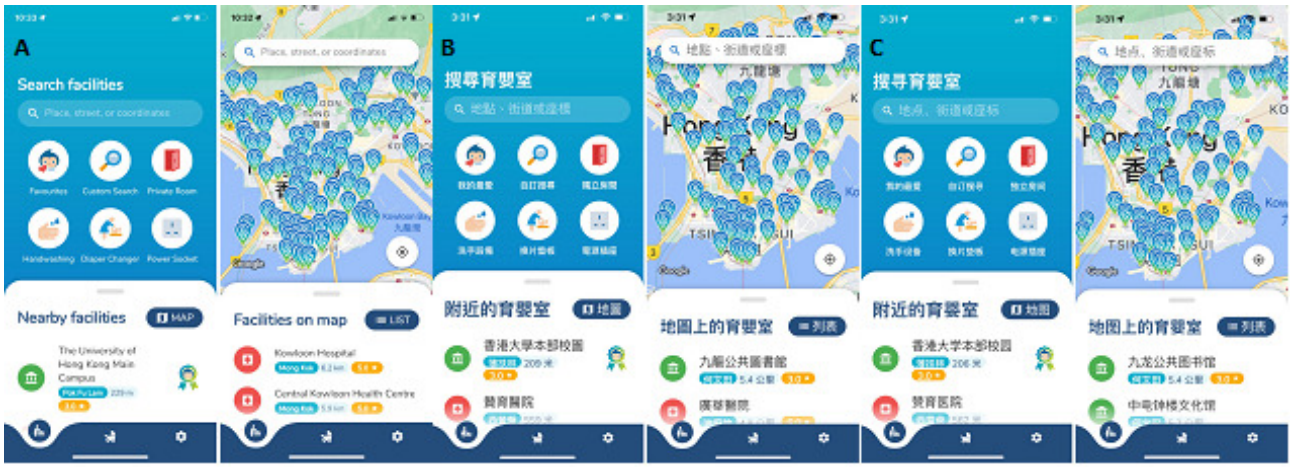

After developing the bfGPS prototype, breastfeeding individuals and staff from public venues were invited to participate in preimplementation evaluation. This preimplementation evaluation included focus groups and one-on-one interviews to gather feedback on the app's functionality, design, performance, and suggestions for further development of bfGPS. Both public venue staff and breastfeeding individuals were recruited, as it was important to confirm the clarity of the content related to the public premises.

Recruitment and Data Collection

A total of 29 public facility staff members and 29 breastfeeding individuals were invited to participate in the focus groups and one-on-one interviews between May 2019 and October 2020. Convenience sampling was used to recruit participants. Public venue staff were recruited through social media advertisements or by participating in staff training within the BFCI program. Staff training for the BFCI program was described elsewhere [38]. In brief, a convenience sample of shopping malls and premises was identified throughout Hong Kong using the internet search. Invitation letters were sent to senior managers and customer service of the identified shopping malls and premises. The emails and letters of invitation include a description of the BFCI program, which included exposure to a breastfeeding mobile app, bfGPS. The shopping malls and premises employees were invited to attend a 2-hour training course based on the WHO breastfeeding counseling course. The employees who attended the BFCI training were invited to participate in the preimplementation evaluation through SMS text messages or emails.

The staff members had diverse occupations and job roles, ranging from frontline staff to managers. The inclusion criteria for staff members were individuals who were 18 years of age or older, worked in public venues, and interacted with the public as part of their jobs. Breastfeeding individuals were recruited if they had breastfed their infants in public or had used breastfeeding facilities in public venues. The recruitment was conducted through the researchers' network. Participants received an HK $100 (approximately US $12.80) gift card as compensation.

Phase 2: Implementation and Updates (Setting and Relevant Context)

Beta testing was conducted to provide early versions of the app to experienced app testers who examined it for errors, crashes, layout issues, software bugs, or other problems. The research team and app developers performed beta testing. bfGPS was published on iOS and Android platforms in September 2020. An app update was launched in September 2021 which allowed users to rate and evaluate breastfeeding facilities and included an online function for the public to upload recent information about breastfeeding facilities.

Based on the participants' suggestions in the focus groups and one-on-one interviews, the ratings and evaluation stage were introduced in September 2021, along with an online function for the public to upload recent information about breastfeeding facilities (Figure 3). The infant tracker was a new add-on feature launched in May 2022 to encourage users to record infant development, including the amount of breast milk, breastfeeding times, sleep patterns, and reminders for feeding and vaccinations (Figure 4).

XSL•FO
RenderX

Figure 3. Interactive function for the users to upload information related to breastfeeding facilities in bfGPS. Users can submit information about breastfeeding facilities, such as name, type of premise, address, and photos of the premises. The study team then reviews the information to ensure accuracy before updating the app. This approach helps maintain up-to-date and reliable data for bfGPS users.

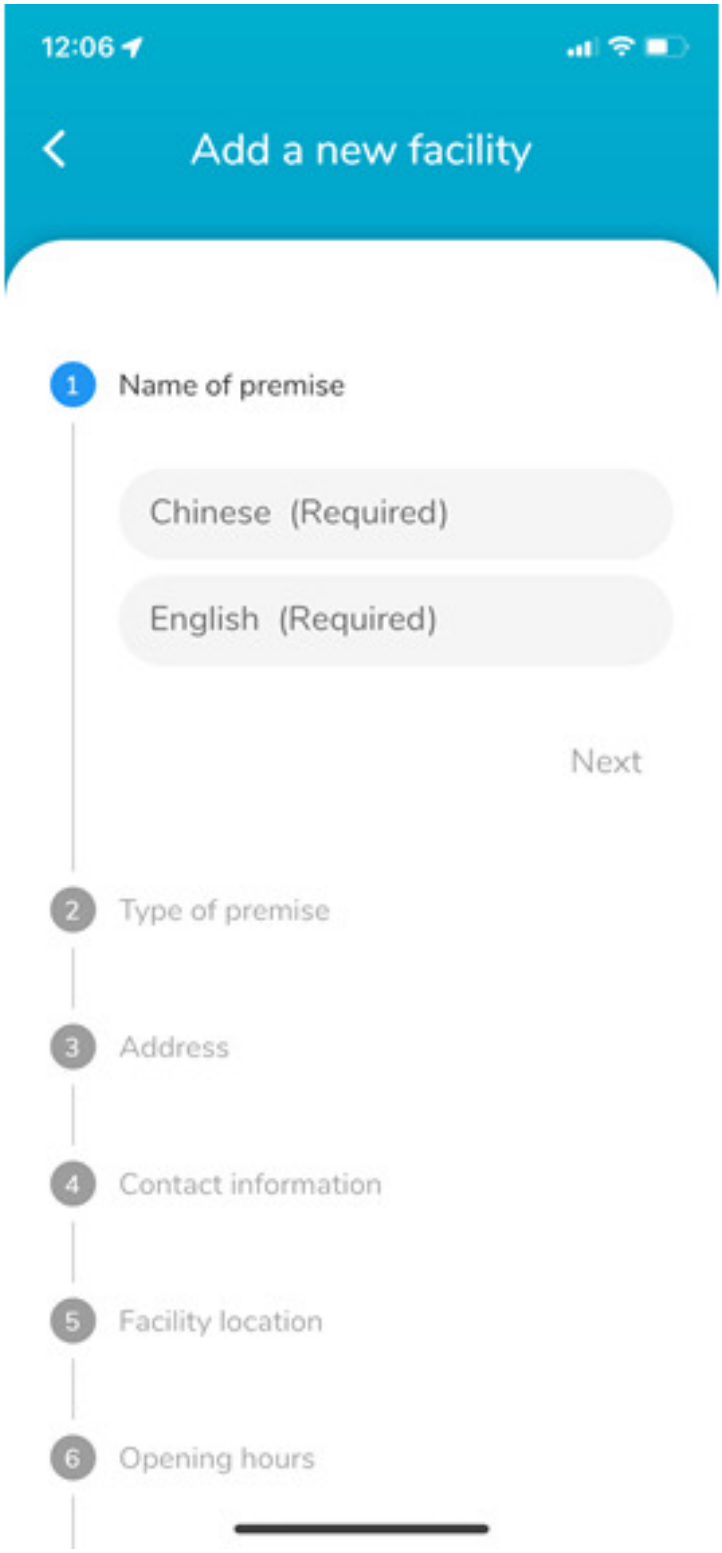

XSL•FO
RenderX

Figure 4. The infant tracker function of bfGPS. The infant tracker enables users to create multiple infant profiles. Key functions include recording infant feeding, human milk expression, diaper changes, sleep patterns, growth metrics, and immunization schedules. The growth patterns are plotted on growth charts using data from the Hong Kong Growth Study 2020 [39]. In addition, the infant tracker calculates upcoming vaccination dates based on the infant's birth date, following the schedule of the Hong Kong Childhood Immunization Programme [40]. The infant record can be exported in a PDF file.

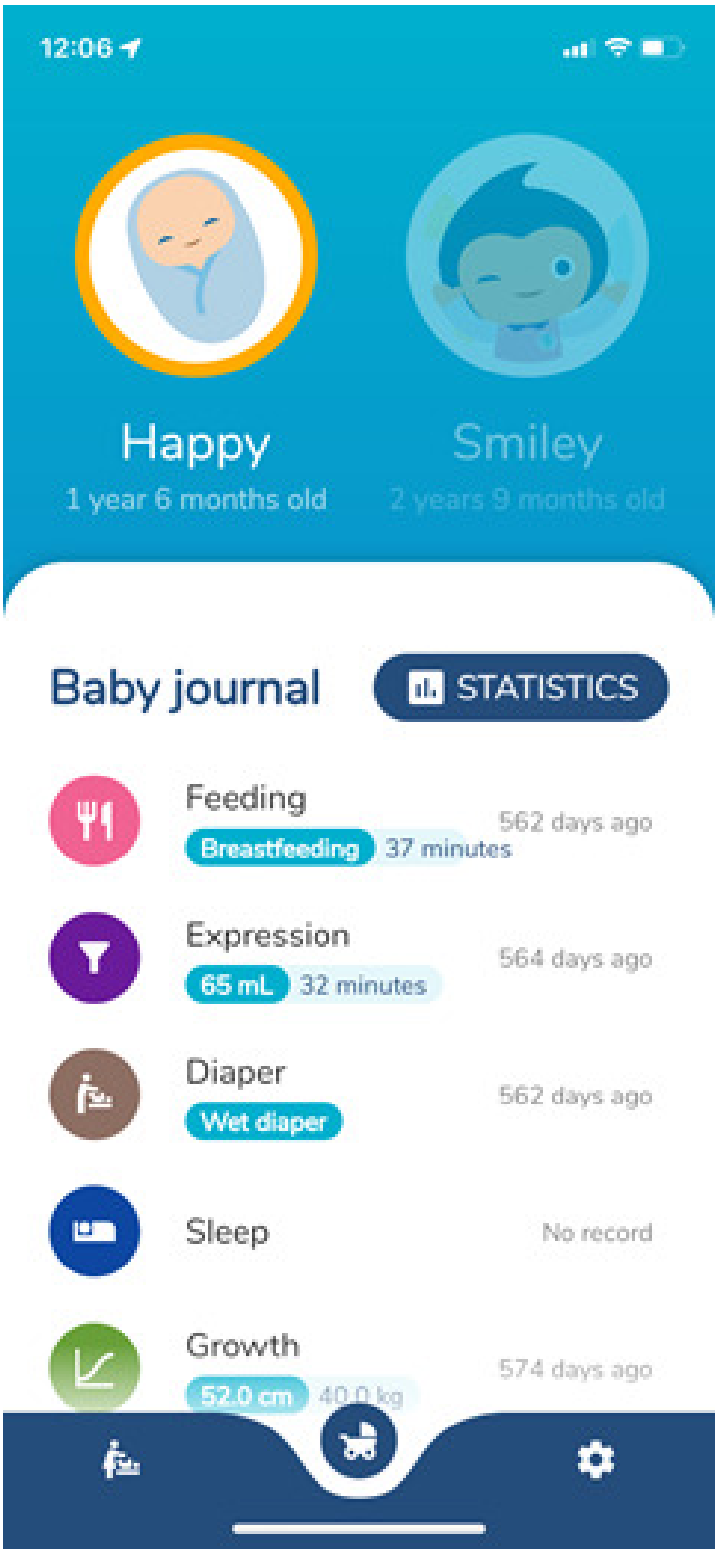

Phase 3: Usability Evaluation (Setting and Relevant Context, Recruitment, and Data Collection)

Following the introduction of the infant tracker function, qualitative interviews were conducted with 30 breastfeeding individuals who used location mapping and the infant tracker function from August to September 2022. Breastfeeding individuals were recruited through Facebook (Meta) and Instagram (Meta) posts containing brief descriptions of the study and a link for study registration. The following criteria were used: (1) 18 years of age or older, (2) Cantonese speaking, (3) pregnant or recently gave birth to a live newborn, (4) Hong

XSL•FO
RenderX

Kong residents, and (5) had ever used bfGPS. In the study registration form, participants were asked to indicate which functions did they used in bfGPS. Data saturation was achieved, as no new information emerged from the interviews [41]. Participants received HK $100 (approximately US $12.80) in coupons for a local grocery store as an honorarium for participating in the interviews.

Data Collection

Semistructured interview guides with open-ended questions were used to collect the data. The participants' experiences of using bfGPS, their insights into the strengths and limitations of bfGPS, recommendations for future development, and significant features for improving the user experience were comprehensively examined. The interview guides were refined as new insights, and perspectives were obtained throughout the study. Focus groups and one-on-one interviews were scheduled based on participants' preferences and conducted at various locations. The interviews were conducted by the authors, who are experienced in conducting qualitative interviews (HSLF and KWL). For the preimplementation evaluation (phase 1), 7 focus groups and 1 staff one-on-one interview were conducted in public commercial rooms and at the university campus. The 3 focus groups and 18 one-on-one interviews with breastfeeding individuals took place in public commercial rooms. The usability evaluation interviews (phase 3) were conducted at public venues convenient for the interviewees (n=30). Each focus group and the one-on-one interview lasted approximately 30 minutes and were audio-recorded with the participants' written permission.

Data Analysis

Descriptive statistics were used to summarize the participants' characteristics. Trained transcriptionists transcribed the audio recordings verbatim into Chinese. The transcripts were reviewed multiple times to increase the researcher's familiarity with the participants' experiences of using bfGPS. The transcripts were then denaturalized and translated into English. Content analysis was used to analyze the transcripts, as it allows for the identification of core consistencies and meanings in qualitative data [42]. The research team reviewed each transcribed interview and developed an open code list derived directly from the data. All textual data were coded and grouped into conceptually similar and overarching themes [43]. To enhance familiarity with the data, a manual data management approach was used, aided by using Microsoft Word [44].

Ethical Considerations

Ethical approval was obtained from the Institutional Review Board of the University of Hong Kong and Hospital Authority Hong Kong West Cluster (UW18-489). The study was conducted in accordance with the principles of the Helsinki Declaration. Participants were informed that they could withdraw from the study at any time without any repercussions. They received HK $100 (approximately US $12.80) in grocery store coupons as an honorarium for interview participation. Written informed consent was obtained from all participants. No specific identifiers were included in the data used for analysis, and participants' confidentiality was maintained at all times.

Results

Breastfeeding Facilities Assessment and Function of bfGPS

A total of 658 breastfeeding facilities in 502 sites were assessed, and the prototype of bfGPS was developed. bfGPS uses the GPS function to locate the nearest breastfeeding facilities. Users can use the advanced search function to narrow their search based on the type of premises and specific equipment provided by the breastfeeding facilities. In addition, photographs of the listed breastfeeding facilities were made available, along with equipment details.

Preimplementation Evaluation of bfGPS

Tables 1 and 2 present the characteristics of the staff members and breastfeeding individuals who participated in the focus groups and one-on-one interviews. Approximately 40% of the staff members were 35 years of age or above (Table 1). The majority of them were female (24/29, 83%), worked full-time (28/29, 97%), and were employed in the service industry (17/29, 59%). Table 2 provides the characteristics of the breastfeeding individuals, revealing that the majority were aged 35 years or older (15/27, 56%), primiparous (16/27, 59%), and returned to work within 6 months post partum (14/27, 52%).

XSL•FO
RenderX

Table . Characteristics of the public facility staff who participated in the preimplementation evaluation.

Demographic variable		Total (N=29), n (%)
Age (years)		
	18 - 29	10 (37)
	30 - 34	6 (22)
	≥35	11 (41)
	Not reported	2 (7)
Sex		
	Female	24 (83)
	Male	5 (17)
Employment status		
	Work full-time	28 (97)
	Work part-time	1 (4)
Occupation		
	Service industry	17 (59)
	Catering industry	4 (14)
	Property management	6 (21)
	Others	2 (7)

Table . Characteristics of the breastfeeding individuals who participated in the preimplementation evaluation.

Demographic variable		Total (N=27[a]), n (%)
Maternal age (years)		
	18 - 29	2 (7)
	30 - 34	10 (37)
	≥35	15 (56)
Parity		
	Primiparous	16 (59)
	Multiparous	11 (41)
Age of the child during the interview		
	<3 months	5 (19)
	3 to <6 months	5 (19)
	6 to <12 months	4 (15)
	≥12 months	13 (48)
Return to work within 6 months post partum		
	No	13 (48)
	Yes	14 (52)

[a]Of total, 2 participants had missing sociodemographic characteristics.

Participants provided feedback on various aspects of bfGPS, including layout, app performance, functionality, and suggestions for further development. Overall, the participants expressed satisfaction with the functions of bfGPS but preferred a more visually appealing layout. They also provided suggestions to enhance user experiences, such as implementing a rating function for the breastfeeding facilities and an infant tracker.

Layout and App Performance

Most participants stated that the layout and interface of bfGPS were user-friendly and easy to use. The app effectively helped users locate breastfeeding facilities. While the participants found the facility photos useful, some expressed concerns about the loading speed and the app's ability to quickly provide the information they needed, such as the location, name, and

XSL•FO
RenderX

equipment details. Furthermore, participants recommended enhancing the visual appeal by improving the layout of bfGPS.

> *With just one click to enter the page, you can choose locations and facilities, which are actually quite convenient. Moreover, opening hours were provided in the app. Parents might find it inconvenient to locate breastfeeding rooms, especially at night, which can be very helpful for them.* [Staff #3, Focus group #1]

> *When using the app for the first time, the premises' name and address should load first…. Waiting for photos to load before showing the premises' name and location can be time-consuming. We want to be fast.* [Breastfeeding individual #7]

> *bfGPS would attract me easily if it has an appealing appearance.* [Breastfeeding individual #6]

Areas of Improvement

Participants were asked to suggest ways to maximize the usage of bfGPS. In addition to the facilities locator function, participants recommended incorporating the following functions into the app: an infant tracker, more up-to-date information, increased uniqueness, and an interface for commenting on and reading other users' feedback.

All participants expressed concerns about the accuracy and timeliness of the information provided by the app. They emphasized that having inaccurate or outdated information would cause inconvenience to them and discourage them from using the app. To address this, participants suggested adding a channel for app users to report updates on breastfeeding facilities. The system administrators can then verify and approve the updates, ensuring the app provides accurate and up-to-date information.

> *I suggest the option for users to upload new baby-care facilities. Alternatively, companies can provide information directly to you (the research team) if they have opened new nursing rooms. You can check the information and upload it to the app after approval. If you need to do all the work by yourself, you may be unable to have a timely update…. This could significantly affect me if the information is outdated. If I go to the baby-care facility and discover it has been demolished, it will waste my time and mess up my plan.* [Breastfeeding individual #8]

To prolong the usage of the app, participants suggested allowing them to enter feeding or baby information, such as feeding time or body weight of the infants. This feature would incentivize them to continue using the app and make them less likely to delete it.

> *Honestly, my phone always runs out of storage space, making it easy to uninstall and reinstall bfGPS. However, if bfGPS provides more personalized functions, e.g. body weight record, feeding schedule, or which shopping mall I go to. I won't delete the app. If I do not go out with the baby or only go to the shopping malls I am familiar with, I don't see the need to keep the app as it uses up my storage space.* [Breastfeeding individual #19]

> *If the app includes a function for entering the baby's height, weight, and head circumference, we are more likely to use it and keep it for an extended period.* [Breastfeeding individual #1]

In addition to the standard information uploaded by the developers, participants expressed their desire to see the photos uploaded by other users and read their comments about the baby-care facilities.

> *I would like to view photos taken by users.* [Staff #4, Focus group #2]

Based on participants' feedback from the preimplementation evaluation, we integrated an infant tracking function, incorporated public ratings, and added a function for uploading recent breastfeeding facility information in bfGPS. These enhancements were introduced in phase 2 of the study, as outlined in the methods section.

Usability Evaluation

The characteristics of the participants are shown in Table 3. Approximately two-thirds were working full-time, and a majority had a bachelor's degree or higher (17/30, 57%), were 30‑34 years of age (16/30, 53%), and lived in Hong Kong since birth (27/30, 90%).

XSL•FO
RenderX

Table . Characteristics of participants in the end-user evaluations.

Demographic variable		Total (N=30), n (%)
Maternal age (years)		
	18 - 29	3 (10)
	30 - 34	16 (53)
	≥35	11 (37)
Employment status		
	Full-time	20 (67)
	Part-time	2 (7)
	Stay-at-home parent	8 (27)
Maternal education		
	Lower than university degree	13 (43)
	University degree or above	17 (57)
Length of residence in Hong Kong		
	15 years or above	3 (10)
	Since birth	27 (90)

bfGPS Improves Community Experience

The majority of participants reported that bfGPS is very convenient to use, providing prompt and detailed information about nearby breastfeeding facilities. bfGPS enables users to assess the breastfeeding facilities and plan ahead before going out with their babies. It offers flexibility to parents in their community experiences, as they can quickly locate the breastfeeding rooms.

> *Before using bfGPS, I wouldn't stay outside for too long. I would return home to feed within a few hours. With bfGPS, I know the locations of nursing rooms, allowing me to plan ahead for where I can go later. This enables me to explore farther places and extend the duration of my time going out.* [Participant #14]

In addition, participants appreciated the rating function and the inclusion of photos of breastfeeding facilities, as it allowed them to gain insights into the settings of these facilities.

> *The best feature of bfGPS is the provided photos, which clearly show the condition of the nursing rooms to me.… I think the 'ratings' function is beneficial because only those who have used the rooms would rate. The rating is important to me…I clicked on the page to see the issues of the breastfeeding facilities with 1-star ratings.* [Participant #32]

bfGPS Helps Track the Infants' Growth

bfGPS allows people to track the growth of infants and helps them retrieve infants' feeding records easily, especially during health checkups.

> *bfGPS allows me to input my baby's growth data, and it can generate a growth chart similar to the one in the maternal and child health center's booklet. I find it beneficial, so I continue to record (the growth data) here…. The breastfeeding timer is also useful and convenient. It enables me to track how much milk or the duration that my baby was fed. I can observe the amount is increasing with the graph.* [Participant #1]

> *If my family member is willing to record in bfGPS, I can easily access information about my baby's sleeping times. This allows me to plan her sleep schedule, especially on weekends, as I am the primary caregiver. My family member handwrites it, and I find it less convenient as I have to check the notebook each time.* [Participant #18]

Suggestions on Further Development

Similar to the findings in the preimplementation evaluation, the majority of participants emphasized the importance of providing written comments. In addition, participants highlighted the significance of sharing data from the infant tracker with family members. They suggested giving information on infant feeding and knowledge about infants' growth and parenting into bfGPS.

> *Users would benefit more from leaving comments after using the nursing rooms instead of solely relying on rating with the number of stars. For example, if the room is clean but the music is too loud, I can describe it clearly with words and leave other users to choose whether to use it. When you (the research team) receive the comments, you may communicate them to the person in charge and let them improve it.* [Participant #19]

> *It is important to have the sharing function, especially when other family members or helpers take care of your baby. This allows me to stay informed about how my baby is doing…. It is good to provide parenting information for parents. You may consider collaborating with the maternal and child health centers as they have a webpage on this.* [Participant #9]

XSL•FO
RenderX

> *(I would like to have information related to) the optimal time for introducing solid food or links to articles on the benefits of breastfeeding or other relevant breastfeeding.… It will be too monotonous if bfGPS is solely for record-keeping. It is better to have additional reference material. It will be challenging to continue using the app without information.* [Participant #7]

Evaluation of bfGPS

bfGPS was ranked number 3 in the Hong Kong App Store's "Health and Fitness" free app category in September 2020. The total downloads in the iOS and Android platforms were 11,542 in December 2023. As of December 2023, the app usage data demonstrated impact and behavior change, with more than 11,000 unique users having used the app and performing more than 484,000 facilities searches. On average, each user conducted 50 (SD 7.25) searches, with an average of 4 (SD 2.47) searches per app usage session. All users viewed the facility details over 150,000 (SD 39,827.06) times in total. Each user viewed the detailed information of 19 (SD 3.43) facilities on average, and each session of app usage involved 1.5 (SD 0.77) views on average. The app attained average user ratings of 4.8 and 4.111 (out of 5.0) on the App Store and Google Play, respectively. We collected 2287 ratings of breastfeeding facilities and 217 uploads of breastfeeding facility information.

Discussion

Principal Results

This study provides an overview of the app development process of bfGPS, which primarily focuses on locating breastfeeding facilities in the community. We have refined the app based on the feedback provided by participants in the preimplementation evaluation, incorporating features such as ratings for breastfeeding facilities and the ability to upload recent breastfeeding facility information.

The infant tracker has the potential to increase the participants' engagement. A study conducted in the United States showed that approximately 57% of participants used a mobile app to track infant feeding [45]. An infant tracker provides objective data to breastfeeding individuals regarding infant feeding methods and other aspects of baby care [46]. However, some breastfeeding individuals may express concerns about relying too heavily on the infant tracker feature [46]. Therefore, a large-scale quantitative study may be necessary to understand the effect of infant tracking on the well-being of breastfeeding individuals and its impact on breastfeeding outcomes.

For the future development of bfGPS and further research, we aim to transform it into a comprehensive platform that empowers users with access to breastfeeding knowledge and information, ultimately enhancing their self-efficacy and improving breastfeeding outcomes. We plan to collaborate with lactation consultants, registered nurses, midwives, breastfeeding individuals, and maternal and child health researchers to develop videos and informational leaflets. These resources will provide valuable information to help breastfeeding individuals overcome challenges. Furthermore, we intend to evaluate the effectiveness of bfGPS in supporting breastfeeding individuals within the community. The option to provide written comments on breastfeeding facilities is currently unavailable due to resource constraints, as close monitoring of the comments and additional staff is required. We will explore methods to enable a more comprehensive rating or comment system in the future.

The participants suggested the inclusion of breastfeeding information in the app. Studies have shown that incorporating breastfeeding information in the app can significantly improve breastfeeding outcomes [47,48]. While providing breastfeeding knowledge, it is essential to ensure the accuracy of the information. A study indicates that a high proportion of infant feeding apps provide incomplete or incorrect information [49]. In addition, most of these apps are provided by commercial sectors, which may result in biased information regarding infant feeding methods [49]. Beyond the app's content, gamification can be an important element to incorporate, as it is associated with hedonic well-being [50] and can encourage long-term app usage [51]. By maintaining user engagement, gamification is an effective component for promoting behavior change [52].

The development of bfGPS involved cross-disciplinary and multidisciplinary collaboration among public, nongovernmental, governmental, and commercial organizations. The fundamental goal is to have an impact on health in Hong Kong and beyond, influencing public, nongovernmental, governmental, and commercial organizations, as well as policies to develop an infrastructure that transforms breastfeeding culture. To date, only a limited number of studies have been conducted to understand the effects of the BFCI on breastfeeding outcomes [19]. Further research is needed to understand the effective elements in creating a breastfeeding-friendly community.

The engagement of community partners and breastfeeding individuals during the app development phase is vital. This approach allows us to understand the community's needs and ensures a user-centered design. Providing accurate information about breastfeeding facilities is important in encouraging breastfeeding individuals to use the app. Including commercial sectors in the BFCI program is important in gaining insights into the perspectives of the community members. This involvement enhances community engagement, encouraging community partners to offer breastfeeding-friendly services, and maintain up-to-date information on breastfeeding facilities. Ultimately, this collaborative effort contributes to fostering a breastfeeding-friendly attitude within the community.

Limitations

Our study has several limitations. First, the end users who participated in the usability evaluation were recruited from the BFCI project's social media page, which may have resulted in a selection bias, as participants were potentially more favorable toward breastfeeding, limiting the generalizability of the study findings. Second, conducting focus groups with public premises staff might have constrained them from freely expressing their individual opinions, as fellow participants were their colleagues or supervisors. Third, the objective of the usability evaluation was to understand the experiences of using the GPS feature. However, some participants did not use either the breastfeeding facilities locations or infant tracker functions, limiting their

XSL•FO
RenderX

ability to provide feedback on these specific features. In addition, some focus groups and one-on-one interviews were conducted before the COVID-19 pandemic, while others were conducted during the pandemic. Participants' concerns, needs, and expectations regarding breastfeeding facilities may have changed. This may limit the transferability of the findings.

Conclusion

This paper presents the user-centered process used in developing the bfGPS app, which aims to improve the breastfeeding experiences of families in the community. In addition, the app includes an infant tracker feature as an add-on, which can enhance various aspects of infant care. Our objective is for this work to serve as a model or guide for applying a user-centered methodology when developing new health care apps. Ensuring the app provides accurate and up-to-date information on breastfeeding facilities is crucial to promoting its continued and widespread use among breastfeeding individuals. This approach holds promise as an effective method for enhancing breastfeeding experiences in public and the community. The app has the potential to both reinforce and support the breastfeeding information provided to parents while also offering strategies for self-management to users. The next step involves testing the app as a therapeutic tool through a randomized clinical trial to verify its effectiveness in sustaining breastfeeding duration and addressing immediate breastfeeding issues. This ultimate goal serves as the motivation behind the creation and development of the app.

Acknowledgments

We would like to thank the participants in this study for sharing their experiences. During the preparation of this work, ChatGPT was used to enhance readability and improve the language. Following the use of this tool, HSLF and KYWL thoroughly reviewed and edited the content as necessary and take full responsibility for the accuracy and integrity of the final publication. This research is funded by the Health and Medical Research Fund (HKU-BF).

Data Availability

The datasets generated during and/or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

References

1. Victora CG, Bahl R, Barros AJD, et al. Breastfeeding in the 21st century: epidemiology, mechanisms, and lifelong effect. Lancet 2016 Jan 30;387(10017):475-490. [doi: 10.1016/S0140-6736(15)01024-7] [Medline: 26869575]
2. Jones G, Steketee RW, Black RE, Bhutta ZA, Morris SS, Bellagio Child Survival Study Group. How many child deaths can we prevent this year? Lancet 2003 Jul 5;362(9377):65-71. [doi: 10.1016/S0140-6736(03)13811-1] [Medline: 12853204]
3. World Health Organization, UNICEF. Global strategy for infant and young child feeding. : World Health Organization; 2003.
4. World Health Organization. Global breastfeeding scorecard 2023: rates of breastfeeding increase around the world through improved protection and support. : World Health Organization; 2023.
5. Holowko N, Jones M, Koupil I, Tooth L, Mishra G. High education and increased parity are associated with breast-feeding initiation and duration among Australian women. Public Health Nutr 2016 Oct;19(14):2551-2561. [doi: 10.1017/S1368980016000367] [Medline: 26996672]
6. Tarrant M, Fong DYT, Wu KM, et al. Breastfeeding and weaning practices among Hong Kong mothers: a prospective study. BMC Pregnancy Childbirth 2010 May 29;10:27. [doi: 10.1186/1471-2393-10-27] [Medline: 20509959]
7. Brown A, Rance J, Bennett P. Understanding the relationship between breastfeeding and postnatal depression: the role of pain and physical difficulties. J Adv Nurs 2016 Feb;72(2):273-282. [doi: 10.1111/jan.12832] [Medline: 26494433]
8. Bai DL, Fong DYT, Lok KYW, Tarrant M. Relationship between the infant feeding preferences of Chinese mothers' immediate social network and early breastfeeding cessation. J Hum Lact 2016 May;32(2):301-308. [doi: 10.1177/0890334416630537] [Medline: 26887843]
9. Almohanna AA, Win KT, Meedya S. Effectiveness of internet-based electronic technology interventions on breastfeeding outcomes: systematic review. J Med Internet Res 2020 May 29;22(5):e17361. [doi: 10.2196/17361] [Medline: 32469315]
10. Lau Y, Htun TP, Tam WSW, Klainin-Yobas P. Efficacy of e-technologies in improving breastfeeding outcomes among perinatal women: a meta-analysis. Matern Child Nutr 2016 Jul;12(3):381-401. [doi: 10.1111/mcn.12202] [Medline: 26194599]
11. Ziebart M, Kammermeier M, Koletzko B, Patro-Golab B. Mobile applications for promoting and supporting breastfeeding: systematic review and meta-analysis. Matern Child Nutr 2024 Oct 11:e13733. [doi: 10.1111/mcn.13733] [Medline: 39390945]
12. Vila-Candel R, Mena-Tudela D, Franco-Antonio C, Quesada JA, Soriano-Vidal FJ. Effects of a mobile application on breastfeeding maintenance in the first 6 months after birth: randomised controlled trial (COMLACT study). Midwifery 2024 Jan;128:103874. [doi: 10.1016/j.midw.2023.103874] [Medline: 37979550]

XSL•FO
RenderX

13. Koçak V, Ege E, İyisoy MS. The development of the postpartum mobile support application and the effect of the application on mothers' anxiety and depression symptoms. Arch Psychiatr Nurs 2021 Oct;35(5):441-449. [doi: 10.1016/j.apnu.2021.06.009] [Medline: 34561057]
14. World Health Organization, UNICEF. Baby-friendly hospital initiative: revised, updated and expanded for integrated care. : World Health Organization; 2009.
15. Guajardo-Villar A, Pelat C, Blondel B, et al. The impact of the Baby-Friendly Hospital Initiative on breastfeeding rates at maternity units in France. Int J Epidemiol 2024 Apr 11;53(3):dyae080. [doi: 10.1093/ije/dyae080] [Medline: 38857529]
16. Pérez-Escamilla R, Martinez JL, Segura-Pérez S. Impact of the Baby-friendly Hospital Initiative on breastfeeding and child health outcomes: a systematic review. Matern Child Nutr 2016 Jul;12(3):402-417. [doi: 10.1111/mcn.12294] [Medline: 26924775]
17. Fair FJ, Morrison A, Soltani H. The impact of Baby Friendly Initiative accreditation: an overview of systematic reviews. Matern Child Nutr 2021 Oct;17(4):e13216. [doi: 10.1111/mcn.13216] [Medline: 34076344]
18. Kavle JA, Ahoya B, Kiige L, et al. Baby-Friendly Community Initiative-from national guidelines to implementation: a multisectoral platform for improving infant and young child feeding practices and integrated health services. Matern Child Nutr 2019 Jan;15 Suppl 1(Suppl 1):e12747. [doi: 10.1111/mcn.12747] [Medline: 30748118]
19. Walsh A, Pieterse P, Mishra N, et al. Improving breastfeeding support through the implementation of the Baby-Friendly Hospital and Community Initiatives: a scoping review. Int Breastfeed J 2023 Apr 15;18(1):22. [doi: 10.1186/s13006-023-00556-2] [Medline: 37061737]
20. Huang Y, Osborne D, Sibley CG. Public support and sociodemographic correlates of public breastfeeding support in New Zealand. N Z Med J 2018 Dec 14;131(1487):44-53. [Medline: 30543611]
21. Meng X, Daly A, Pollard CM, Binns CW. Community attitudes toward breastfeeding in public places among Western Australia Adults, 1995-2009. J Hum Lact 2013 May;29(2):183-189. [doi: 10.1177/0890334413478835] [Medline: 23504475]
22. Mulready-Ward C, Hackett M. Perception and attitudes: breastfeeding in public in New York City. J Hum Lact 2014 May;30(2):195-200. [doi: 10.1177/0890334414524988] [Medline: 24614263]
23. Russell K, Ali A. Public attitudes toward breastfeeding in public places in Ottawa, Canada. J Hum Lact 2017 May;33(2):401-408. [doi: 10.1177/0890334417695203] [Medline: 28418805]
24. Zhao Y, Ouyang YQ, Redding SR. Attitudes of Chinese adults to breastfeeding in public: a web-based survey. Breastfeed Med 2017 Jun;12:316-321. [doi: 10.1089/bfm.2017.0013] [Medline: 28440676]
25. US Department of Health Human Services. Executive summary: the surgeon general's call to action to support breastfeeding. Breastfeed Med 2011 Feb;6(1):3-5. [doi: 10.1089/bfm.2011.9996]
26. Hauck YL, Kuliukas L, Gallagher L, Brady V, Dykes C, Rubertsson C. Helpful and challenging aspects of breastfeeding in public for women living in Australia, Ireland and Sweden: a cross-sectional study. Int Breastfeed J 2020 May 12;15(1):38. [doi: 10.1186/s13006-020-00281-0] [Medline: 32398087]
27. Grant A. Breasts and the city: an urban ethnography of infant feeding in public spaces within Cardiff, United Kingdom. Int Breastfeed J 2021 Apr 29;16(1):1-9. [doi: 10.1186/s13006-021-00384-2] [Medline: 33926506]
28. Grant A, Pell B, Copeland L, et al. Views and experience of breastfeeding in public: a qualitative systematic review. Matern Child Nutr 2022 Oct;18(4):e13407. [doi: 10.1111/mcn.13407] [Medline: 35914544]
29. Scott JA, Kwok YY, Synnott K, et al. A comparison of maternal attitudes to breastfeeding in public and the association with breastfeeding duration in four European countries: results of a cohort study. Birth 2015 Mar;42(1):78-85. [doi: 10.1111/birt.12138] [Medline: 25339304]
30. Balaam M, Comber R, Jenkins E, Sutton S, Garbett A, editors. FeedFinder: a location-mapping mobile application for breastfeeding women. Presented at: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems; Apr 18-23, 2015; New York, NY. [doi: 10.1145/2702123.2702328]
31. Chaovalit P, Pongnumkul S, editors. Moommae: A location-aware mobile information system for breastfeeding mothers in thailand (invited paper). Presented at: 2017 10th International Conference on Ubi-media Computing and Workshops (Ubi-Media); Aug 1-4, 2017; Pattaya, Thailand. [doi: 10.1109/UMEDIA.2017.8074104]
32. Wang CJ, Chaovalit P, Pongnumkul S. A breastfeed-promoting mobile app intervention: usability and usefulness study. JMIR Mhealth Uhealth 2018 Jan 26;6(1):e27. [doi: 10.2196/mhealth.8337] [Medline: 29374000]
33. World breastfeeding week 2023: annual survey summary. : Baby Friendly Hospital Initiative Hong Kong Association; 2023 URL: https://www.babyfriendly.org.hk/wp-content/uploads/2023/07/2023-WBW-Annual-Survey_E.pdf [accessed 2025-01-06]
34. Breastfeeding survey 2021. Department of Health. 2021. URL: https://www.fhs.gov.hk/english/reports/files/BF_survey_2021.pdf [accessed 2025-01-06]
35. Survey on mothers' breastfeeding experience in public places. Department of Health. 2016. URL: https://www.fhs.gov.hk/english/archive/files/reports/Survey_on_Mothers_views_on_BF_exp_in_pubplace_2016_final.pdf [accessed 2025-01-06]
36. Howel D, Ball H. Association between length of exclusive breastfeeding and subsequent breastfeeding continuation. J Hum Lact 2013 Nov;29(4):579-585. [doi: 10.1177/0890334413492908] [Medline: 23821654]
37. Scott J, Ahwong E, Devenish G, Ha D, Do L. Determinants of continued breastfeeding at 12 and 24 months: results of an Australian cohort study. Int J Environ Res Public Health 2019 Oct 18;16(20):3980. [doi: 10.3390/ijerph16203980] [Medline: 31635280]

XSL•FO
RenderX

38. Lok KYW, Fan HSL, Ip HL, et al. Building a baby-friendly community: development, strategy, implementation and evaluation of a BFCI program. J Adv Nurs 2024 Jun 14. [doi: 10.1111/jan.16267] [Medline: 38877670]
39. Hong Kong 2020 growth references 2024. Department of Health HKSAR. URL: https://www.dh.gov.hk/english/useful/useful_HP_Growth_Chart/files/growth_charts.pdf [accessed 2025-01-06]
40. Schedule of Hong Kong Childhood Immunisation Programme. Family Health Service Department of Health HKSAR. 2022. URL: https://www.fhs.gov.hk/english/main_ser/child_health/child_health_recommend.html [accessed 2024-12-12]
41. Fusch PI, Ness LR. Are we there yet? Data saturation in qualitative research. TQR 2015;20(9):1408. [doi: 10.46743/2160-3715/2015.2281]
42. Patton MQ. Qualitative Research & Evaluation Methods, 3rd edition: Sage Publications; 2002. URL: https://aulasvirtuales.wordpress.com/wp-content/uploads/2014/02/qualitative-research-evaluation-methods-by-michael-patton.pdf [accessed 2025-01-06]
43. Bengtsson M. How to plan and perform a qualitative study using content analysis. NursingPlus Open 2016;2:8-14. [doi: 10.1016/j.npls.2016.01.001]
44. Webb C. Analysing qualitative data: computerized and other approaches. J Adv Nurs 1999 Feb;29(2):323-330. [doi: 10.1046/j.1365-2648.1999.00892.x] [Medline: 10197931]
45. Dinour LM. Infant feeding tracker applications: cross-sectional analysis of use. J Nutr Educ Behav 2022 Sep;54(9):835-843. [doi: 10.1016/j.jneb.2022.03.012] [Medline: 35643748]
46. Dienelt K, Moores CJ, Miller J, Mehta K. An investigation into the use of infant feeding tracker apps by breastfeeding mothers. Health Informatics J 2020 Sep;26(3):1672-1683. [doi: 10.1177/1460458219888402] [Medline: 31793799]
47. Doan TTD, Pham NM, Binns C, et al. Effect of a mobile application on breastfeeding rates among mothers who have cesarean deliveries: a randomized controlled trial. Breastfeed Med 2022 Oct;17(10):832-840. [doi: 10.1089/bfm.2022.0088] [Medline: 36083249]
48. Seyyedi N, Rahmatnezhad L, Mesgarzadeh M, Khalkhali H, Seyyedi N, Rahimi B. Effectiveness of a smartphone-based educational intervention to improve breastfeeding. Int Breastfeed J 2021 Sep 20;16(1):70. [doi: 10.1186/s13006-021-00417-w] [Medline: 34544429]
49. Cheng H, Tutt A, Llewellyn C, et al. Content and quality of infant feeding smartphone apps: five-year update on a systematic search and evaluation. JMIR Mhealth Uhealth 2020 May 27;8(5):e17300. [doi: 10.2196/17300] [Medline: 32459187]
50. Lin Y, Tudor-Sfetea C, Siddiqui S, Sherwani Y, Ahmed M, Eisingerich AB. Effective behavioral changes through a digital mHealth app: exploring the impact of hedonic well-being, psychological empowerment and inspiration. JMIR Mhealth Uhealth 2018 Jun 15;6(6):e10024. [doi: 10.2196/10024] [Medline: 29907557]
51. White B, Giglia RC, White JA, Dhaliwal S, Burns SK, Scott JA. Gamifying breastfeeding for fathers: process evaluation of the milk man mobile app. JMIR Pediatr Parent 2019 Jun 20;2(1):e12157. [doi: 10.2196/12157] [Medline: 31518324]
52. El-Hilly AA, Iqbal SS, Ahmed M, et al. Game on? Smoking cessation through the gamification of mHealth: a longitudinal qualitative study. JMIR Serious Games 2016 Oct 24;4(2):e18. [doi: 10.2196/games.5678] [Medline: 27777216]

Abbreviations

BFCI: Baby-Friendly Community Initiative
BFHI: Baby-Friendly Hospital Initiative
WHO: World Health Organization

Edited by A Mavragani; submitted 17.07.24; peer-reviewed by A Eisingerich, T Ma; revised version received 05.11.24; accepted 06.11.24; published 10.01.25.

Please cite as:
Fan HSL, Leung ETY, Lau KW, Wong JYH, Choi EPH, Lam C, Tarrant M, Ngan HYS, Ip P, Lin CC, Lok KYW
A Mobile App for Promoting Breastfeeding-Friendly Communities in Hong Kong: Design and Development Study
JMIR Form Res 2025;9:e64191
URL: https://formative.jmir.org/2025/1/e64191
doi:10.2196/64191

© Heidi Sze Lok Fan, Emily Tsz Yan Leung, Ka Wing Lau, Janet Yuen Ha Wong, Edmond Pui Hang Choi, Christine Lam, Marie Tarrant, Hextan Yuen Sheung Ngan, Patrick Ip, Chia Chin Lin, Kris Yuet Wan Lok. Originally published in JMIR Formative Research (https://formative.jmir.org), 10.1.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete

XSL•FO
RenderX

bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Predicting Age and Visual-Motor Integration Using Origami Photographs: Deep Learning Study

Chien-Yu Huang[1,2], PhD; Yen-Ting Yu[1,3], PhD; Kuan-Lin Chen[3,4,5], PhD; Jenn-Jier Lien[6], PhD; Gong-Hong Lin[7], PhD; Ching-Lin Hsieh[1,2,8], PhD

[1]School of Occupational Therapy, College of Medicine, National Taiwan University, Taipei, Taiwan

[2]Department of Physical Medicine and Rehabilitation, National Taiwan University Hospital, Taipei, Taiwan

[3]Department of Occupational Therapy, College of Medicine, National Cheng Kung University, Tainan, Taiwan

[4]Institute of Allied Health Sciences, College of Medicine, National Cheng Kung University, Tainan, Taiwan

[5]Department of Physical Medicine and Rehabilitation, National Cheng Kung University Hospital, Tainan, Taiwan

[6]Department of Computer Science and Information Engineering, College of Electrical Engineering and Computer Science, National Cheng Kung University, Tainan, Taiwan

[7]International PhD Program in Gerontology and Long-Term Care, School of Nursing, Taipei Medical University, No. 250 Wu-Xing Street, Taipei, Taiwan

[8]Department of Occupational Therapy, College of Medical and Health Sciences, Asia University, Taipei, Taiwan

Corresponding Author:
Gong-Hong Lin, PhD
International PhD Program in Gerontology and Long-Term Care, School of Nursing, Taipei Medical University, No. 250 Wu-Xing Street, Taipei, Taiwan

Abstract

Background: Origami is a popular activity among preschool children and can be used by therapists as an evaluation tool to assess children's development in clinical settings. It is easy to implement, appealing to children, and time-efficient, requiring only simple materials—pieces of paper. Furthermore, the products of origami may reflect children's ages and their visual-motor integration (VMI) development. However, therapists typically evaluate children's origami creations based primarily on their personal background knowledge and clinical experience, leading to subjective and descriptive feedback. Consequently, the effectiveness of using origami products to determine children's age and VMI development lacks empirical support.

Objective: This study had two main aims. First, we sought to apply artificial intelligence (AI) techniques to origami products to predict children's ages and VMI development, including VMI level (standardized scores) and VMI developmental status (typical, borderline, or delayed). Second, we explored the performance of the AI models using all combinations of photographs taken from different angles.

Methods: A total of 515 children aged 2-6 years were recruited and divided into training and testing groups at a 4:1 ratio. Children created origami dogs, which were photographed from 8 different angles. The Beery–Buktenica Developmental Test of Visual-Motor Integration, 6th Edition, was used to assess the children's VMI levels and developmental status. Three AI models—ResNet-50, XGBoost, and a multilayer perceptron—were combined sequentially to predict age *z* scores and VMI *z* scores using the training group. The trained models were then tested using the testing group, and the accuracy of the predicted VMI developmental status was also calculated.

Results: The R^2 of the age and the VMI trained models ranged from 0.50 to 0.73 and from 0.50 to 0.66, respectively. The AI models that obtained an R^2>0.70 for the age model and an R^2>0.60 for the VMI model were selected for model testing. Those models were further examined for the accuracy of the VMI developmental status, the correlations, and the mean absolute error (MAE) of both the age and the VMI models. The accuracy of the VMI developmental status was about 71%-76%. The correlations between the final predicted age *z* score and the real age *z* score ranged from 0.84 to 0.85, and the correlations of the final predicted VMI *z* scores to the real *z* scores ranged from 0.77 to 0.81. The MAE of the age models ranged from 0.42 to 0.46 and those of the VMI models ranged from 0.43 to 0.48.

Conclusion: Our findings indicate that AI techniques have a significant potential for predicting children's development. The insights provided by AI may assist therapists in better interpreting children's performance in activities.

(JMIR Form Res 2025;9:e58421) doi:10.2196/58421

XSL•FO
RenderX

KEYWORDS

artificial intelligence; origami; child development screening; child development; visual motor integration; children; developmental status; activity performance; deep learning

Introduction

Origami is a common activity for preschool children [1]. From an occupational therapist's perspective, the quality of origami products may reflect a child's developmental age and visual-motor integration (VMI) [2]. For example, a 2-year-old child with typical development is able to roughly fold a sheet of paper in half, and a 3-year-old child is able to fold one diagonally and precisely. Therefore, origami activities and their products could help therapists identify children's development levels, including developmental age and VMI development, through the activity's process and the quality of the origami products.

Origami can be used by therapists as an evaluation tool to identify children's development in clinics because it is easy to implement, attractive to children, requires little time, and uses a simple material (only a piece of paper). However, therapists typically judge children's origami products based mainly on their own background knowledge and clinical experience, providing subjective and descriptive comments on the products (descriptive information). A few developmental assessments have used origami to assess children's fine motor skills. For example, the Peabody Developmental Motor Scales, Second Edition [3]; the Comprehensive Developmental Inventory for Infants and Toddlers [4]; the Stanford-Binet Intelligence Scale, Fifth Edition [5]; and the Bruininks-Oseretsky Test of Motor Proficiency, Second Edition [6] all contain one or more items involving paper folding to assess VMI, visual-spatial skills, or fine motor skills. However, these items are relatively simple and cannot solely be used to predict a child's developmental status. A study by Travers et al [7] investigated how children fold paper across early childhood, and they found that paper folding emerges as early as 27 months and becomes more accurate with age. Moreover, at least 50% of children between 4 and 5.54 years of age completed folds. However, their study only demonstrated children's ability to fold paper at various ages and provided no information on other developmental statuses. As a result, whether origami products can be used to predict children's development remains unsupported by evidence.

Artificial intelligence (AI) can be defined as a computer that acts or thinks like a human. Two AI techniques widely used in the medical and rehabilitation fields are computer vision and data prediction [8-10]. These two techniques may provide objective and quantitative data on origami products. Computer vision can transform image features into numerical data for data prediction, whereas the AI technique of data prediction uses data mining and probability to predict or estimate specific outcomes. Compared to traditional statistical analyses, AI data prediction has more flexibility to deal with categorical, ordinal, and numerical data. By combining the techniques of computer vision and data prediction, the important features of origami images can be extracted and further used to predict outcomes of interest for occupational therapists (eg, the children's age and scores of the standardized VMI assessment in our study).

Some studies have shown that AI techniques can analyze images and can identify children's behaviors and diagnoses with high accuracy [10-12]. In our preliminary study, we employed a simplified AI model (a single AI model) with a small sample size (n=119) to use photographs of origami products to predict children's ages [13]. The AI models showed promising performances (R^2=0.10 - 0.69). To extend upon our previous works, there were two purposes of our study. First, we refined the AI models, using a combination of 2 types of models, with a larger sample of origami products to predict children's ages (*z* scores), VMI levels (*z* scores), and VMI developmental status (typical, borderline, or delayed). Second, we explored the performances of the AI models using all combinations of photographs taken from different angles.

Methods

Participants and Procedure

The data included in our study were collected as part of an ongoing AI project (An Artificial Intelligence System for Assessing Gross Motor and Fine Motor, Handwriting, Attention and Emotion in Children with Developmental Delay), and some of the data were used in the preliminary study [13]. Preschool children enrolled in kindergarten were recruited to the study. The inclusion criteria were as follows: aged 2-6 years, ability to complete group administrations, and informed consent given by caregivers. The exclusion criterion was any obvious physical or mental disability.

The researchers first explained research purposes and procedures to the kindergarten principals. After the kindergarten principals agreed to participate in this study, researchers then contacted the children's parents to invite them to participate in the study. The researchers visited the children's classroom at the kindergarten and led them in two group activities (ie, origami and the administration of the VMI subtest). During the group activities, one researcher was a leader, and the other researchers (2-3 people) were coleaders who provided assistance to the children. Each coleader helped 3 - 6 children, depending on the number of children in the class. After leading the group activities, the researchers took photographs of the products. The researchers took 8 photographs of each origami work, including 4 angles (0, 60, 180, and 240 degrees) from the front side and the same 4 angles from the back side (Figure 1; the photographs are numbered 1 to 8 for each child depending on the shooting angle).

XSL•FO
RenderX

Figure 1. Origami photographs captured from different angles.

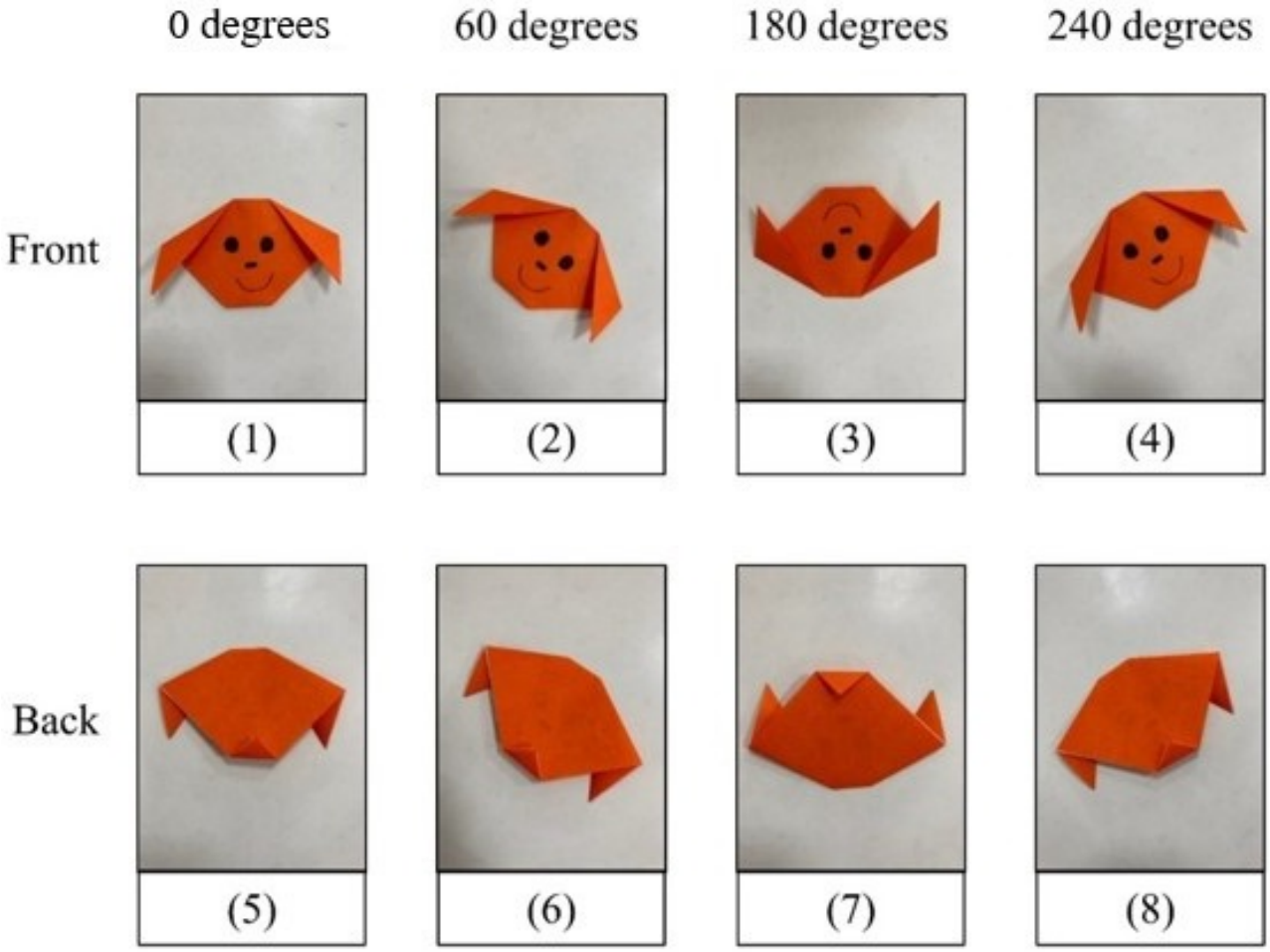

Measures

Origami

An origami dog was made by each of the children (see example in Figure 1). The dog was constructed in three steps: (1) fold a square paper into a triangle shape; (2) from the 2 corners of the long side of the triangle, diagonally fold 2 small triangles as 2 ears; and (3) turn the paper over, and from the other corner, fold a small triangle as the dog's chin. After the dog's face was made, the origami was turned to the front side. Facial features were then drawn on the dog by the child.

BEERY VMI Beery-Buktenica Visual-Motor Integration, Sixth Edition

The BEERY VMI Beery-Buktenica Visual-Motor Integration, Sixth Edition (BEERY VMI-6) was used to assess the children's VMI development [14]. The BEERY VMI-6 consists of three subtests: VMI, visual perception, and motor coordination. Only the VMI subtest was used in this study. The VMI subtest contains 30 pictures. A child needed to copy each picture, and the copy was rated as 0 (incorrect) or 1 (correct). The total score of the VMI test ranged from 0 to 30. The total score of the VMI test was transformed into standard scores based on the VMI norm, and the standard scores were classified into 3 developmental statuses: typical, borderline, or delayed. The total VMI scores (raw VMI scores were referred to as the VMI level in this study) and the developmental status were used for the analyses. The BEERY VMI-6 has good psychometric properties [14].

Statistical Analysis

The participants' products were randomly divided into a training group and a testing group at a 4:1 ratio for the model training and testing. For the training group, 4 steps were conducted to build the AI models that would use origami photographs to respectively predict the children's age *z* scores and VMI *z* scores (Figure 2). First, we used a deep residual network of 50 layers, which is called ResNet-50 (developed by He et al [15]), to extract origami features of various photos at different angles [16]. Second, we trained XGBoost 1.0.0 models (developed by Chen and Guestrin [17]) to use a child's photo features to respectively predict the child's age and raw VMI score. ResNet-50 extracted 2048 features from the photos and the XGBoost model predicted age and raw VMI scores. All combinations of the extracted photo features (from 1 photo alone to 8 photographs together) were respectively fitted to the XGBoost models. The parameters of the XGBoost model were set as follows: n_estimators=100, random_state=1, max_depth=3, and subsample=0.8. The parameters of the multilayer perceptron model were set as follows: random_state=1, max_iter=300, tol=$1e^{-3}$, and n_iter_no_change=5.

XSL•FO
RenderX

Figure 2. An artificial intelligence model example of the deep learning procedure. VMI: visual-motor integration.

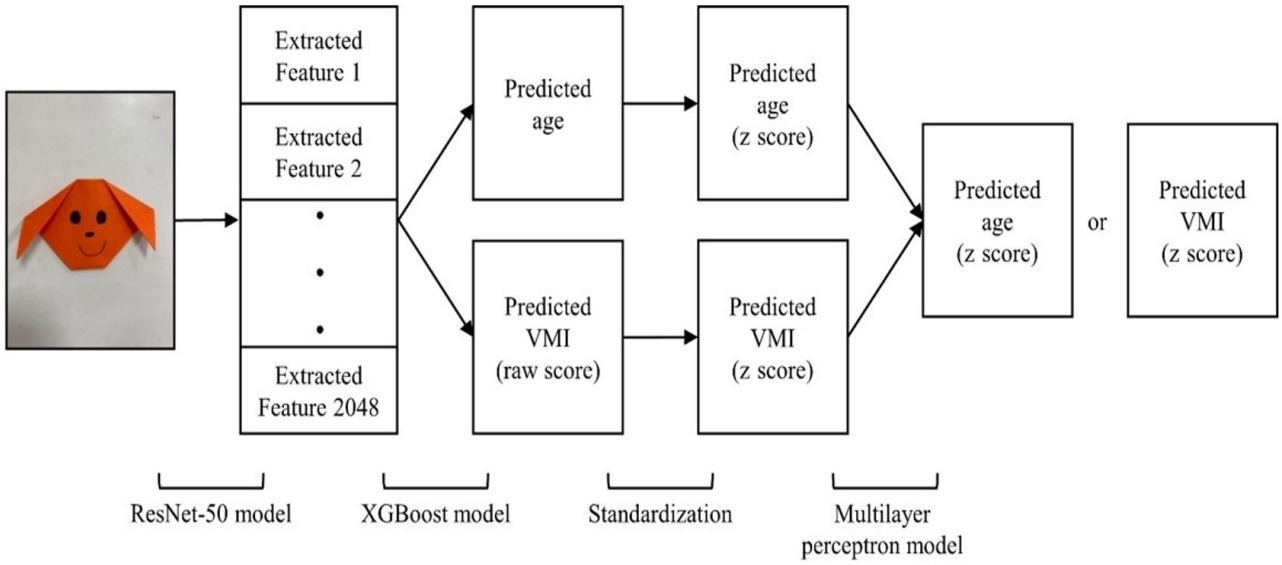

Third, the ages and the raw VMI scores predicted by the XGBoost models were respectively transformed into *z* scores (ie, a predicted age *z* score and a predicted VMI *z* score). We adopted the mean and SD of the training group's age and raw VMI scores for the transformation of *z* scores. We did so because the variables of age and raw VMI score were rated on different scales, and the standardization step ensured that the scales of the age and VMI score would be consistent (both between –1 and 1), which facilitated the subsequent model of data prediction [18,19]. Last, we trained multilayer perceptron models using both the predicted age *z* score and the predicted VMI *z* score to predict the real age *z* score and the real VMI *z* score, respectively.

After the two models (the XGBoost model and the multilayer perceptron model) were trained, the testing group was applied to the two trained models to examine the prediction performance of the models. The R^2, the correlation coefficient, and the mean absolute error (MAE) were used to identify the fitness of the models. The MAE represented the absolute score differences (ie, errors) between the real scores and the predicted scores. It was calculated as follows:

$$\sum i=1nyi-xi\vee n$$

y_i = predicted scores (eg, predicted age and predicted raw VMI scores);

x_i = real scores (eg, real age and real VMI scores).

Moreover, we intended to identify if the predicted VMI scores could reflect children's VMI developmental status. The predicted VMI *z* scores were further used for the analyses. First, the final predicted VMI *z* scores were transformed back to the predicted raw VMI scores. The predicted raw VMI scores were than transformed into predicted standard scores according to the norm of the VMI. Based on the predicted standard scores, we identified children's VMI developmental status as typical, borderline, or delayed. To examine the accuracy of the AI models in predicting the VMI developmental status, we compared the predicted developmental status to the child's real developmental status. Python 3.9 (Python Software) was used to perform the model training and the model testing.

Ethical Considerations

The ethics committee of I-Da hospital approved this study (REC number: EMRP46107N). This study was also approved by the Institutional Review Board of National Taiwan University Hospital (202111093RINB). All methods were conducted in accordance with relevant guidelines and regulations. Informed consent was obtained from the caregivers of the participating children. The study data were deidentified and securely stored in a private cloud space. Participants received a 200 New Taiwan dollar ($6.14 US) coupon after completing the evaluation. No identifying information of individual participants was present in any images included in the manuscript. Participation in this study is entirely voluntary. Participants had the right to withdraw at any time without providing a reason and without facing any negative consequences.

Results

Sample Descriptions

A total of 515 children were included in our study. The data for 412 children were used for the training group, and data for the other 102 children were used for the testing group. The sample characteristics and the VMI scores of the children are presented in Table 1. There were no significant differences in age and sex between the training and testing groups.

XSL•FO
RenderX

Table . Characteristics of the preschool children recruited from kindergartens in the cities of Kaohsiung and Tainan (n=515).

Variables	Training group (n=412)	Testing group (n=103)	*P* value
Age (years), mean (SD)	57.14 (14.46)	57.49 (12.03)	.79
Sex, n (%)			.63
Male	211 (51.2)	50 (48.5)	
Female	201(48.8)	53 (51.5)	
Raw VMI[a] scores, mean (SD)	11.22 (3.74)	11.34 (4.13)	.78
VMI age-equivalent scores, mean (SD)	58.02 (15.04)	58.70 (16.57)	.99
VMI standard scores, mean (SD)	99.06 (10.32)	99.04 (10.08)	.97
VMI percentiles, mean (SD)	47.51 (22.38)	47.61 (22.77)	.69
VMI developmental status, n (%)			.29
Typical	338 (82)	74 (71.8)	
Borderline	63 (15.3)	27 (26.2)	
Delayed	11 (2.7)	2 (2)	

[a]VMI: visual-motor integration.

AI Model Testing

All combinations of photo features were fitted to the XGBoost models, and then the predicted age *z* scores and the predicted VMI *z* scores were fitted to the multilayer perceptron models to produce the final predicted age *z* score and the VMI *z* score. We first inspected the R^2 values of the multilayer perceptron models for predicting the age *z* score and the VMI *z* score to select the better-performing models for further analyses. The R^2 of the age and the VMI models ranged from 0.50 to 0.73 and from 0.50 to 0.66, respectively. We selected AI models that had an R^2>0.70 for the age model and R^2>0.60 for the VMI model, which indicated better performance.

Those models were further examined for their accuracy in predicting the VMI developmental status, the correlations, and the MAE of both the age and the VMI models. Table 2 presents the final performances of the selected models. The accuracy of VMI developmental status was about 71%-76%. The correlations between the final predicted age *z* score and the real age *z* score ranged from 0.84 to 0.86, and the correlations of the final predicted VMI *z* scores to the real VMI *z* scores ranged from 0.7 to 0.81. The MAEs of the age models were between 0.42-0.46, and the MAEs of the VMI models were from 0.43 to 0.48.

Table . Comparison of origami prediction models using photos with different origami angles.

Photo numbers[a]	R^2 (age)	R^2 (VMI[b])	Accuracy (VMI)	Correlation (age)	Correlation (VMI)	MAE[c] (age)	MAE (VMI)
2, 5, 6	0.70	0.61	0.76	0.85	0.79	0.44	0.46
2, 5, 8	0.72	0.66	0.75	0.86	0.81	0.42	0.43
2, 7, 8	0.71	0.61	0.74	0.84	0.78	0.42	0.47
2, 3, 5, 6, 7	0.70	0.63	0.74	0.85	0.79	0.44	0.46
1, 3, 5, 7, 8	0.71	0.60	0.72	0.85	0.78	0.46	0.48
1, 3, 4, 5, 6, 7	0.73	0.61	0.71	0.86	0.78	0.44	0.47
2, 3, 4, 5, 6, 8	0.70	0.62	0.72	0.84	0.79	0.44	0.46
1, 2, 3, 5, 7, 8	0.70	0.62	0.73	0.84	0.79	0.45	0.46
2, 3, 4, 5, 7, 8	0.71	0.63	0.71	0.85	0.79	0.44	0.45
1, 2, 4, 5, 6, 7, 8	0.71	0.60	0.74	0.85	0.78	0.44	0.47

[a]Numbers correspond to the origami photographs shot from different sides and angles. Please see Figure 1 for the side and angle of each photo.

[b]VMI: visual-motor integration.

[c]MAE: mean absolute error.

Among all the combinations of photo features, the combination of photographs 2 (front 60 degrees), 5 (back 0 degrees), and 6

XSL•FO
RenderX

(back 60 degrees) and the combination of photographs 2, 5, and 8 (back 240 degrees) had better performances for predicting children's age *z* scores and VMI *z* scores.

Discussion

Principal Findings

This is the first study to apply AI techniques to origami images for predicting age, VMI level, and VMI developmental status in children. The results showed that AI models could be trained using photographs of children's origami creations to predict their ages (R^2 ranging from 0.70 to 0.73) and VMI levels (R^2 ranging from 0.60 to 0.66). Furthermore, the accuracy of predicting VMI developmental status was approximately 71%-76%. These findings suggest that AI techniques hold significant potential for analyzing origami images to further assess children's VMI development.

Analyzing origami photographs using AI techniques to predict children's age, as done in this study, presents two advantages. First, age is an interval indicator, and an interval indicator is an appropriate outcome measure for testing a prediction model. Second, children's age, to some extent, reflects their developmental level. As mentioned in the introduction, a 2-year-old child with typical development is supposed to be able to roughly fold a paper in half, and a 3-year-old child should be able to fold a paper diagonally and precisely. Therefore, predicting children's age through origami may help caregivers and therapists identify children's developmental status. For example, if an AI model prediction reports that a 5-year-old boy has the origami skills of a 3-year-old, the boy may have a developmental delay in some aspects (eg, fine motor manipulation, visual perception, or VMI). Thus, the boy can be referred for further evaluation and possible intervention. In summary, predicting children's ages could not only benefit researchers testing the AI model but also provide information about a child's development for therapists and caregivers.

Comparing all photo combinations for the final models showed that the combinations of photographs 2, 5, and 6 and of photographs 2, 5, and 8 performed better in predicting children's development. These results indicate that those photo combinations provide more useful information (or important features) about the origami products. This may be because the images of the origami dogs taken from those angles fully capture the dog's facial features, including the 2 ears, the chin, and the hand-drawn facial features, which could reflect the quality of the origami products. Using a particular and limited number of photographs can promote the efficiency and clinical utility of AI applications.

It is noted that the final predictions of the AI model used the age *z* score and the VMI *z* score as the input and output values. The reason was that the age and the raw VMI scores were two variables based on different scales, and normalization of the two variables could facilitate the model training process [18,19]. Moreover, the predicted age *z* score and the predicted VMI *z* score could be transformed back to the predicted age and predicted raw VMI scores based on the calculation formula. Therefore, future users could obtain the predicted age and the predicted VMI score from the AI model for use in the interpretation of origami performances.

Based on our findings, an evaluation tool for VMI development could be developed and used in both community and clinical settings. This tool would be an app designed to implement an AI model. By allowing users to upload photographs of a child's origami dogs, the app would immediately calculate and report the child's predicted age and raw VMI score. This app could provide 4 key benefits. First, the app would offer quantitative data, such as the predicted age, VMI level, and VMI developmental status, derived from a child's performance in folding origami dogs. Therapists could combine this quantitative information with personal observations to better interpret children's origami performance. Second, caregivers and teachers, who are not medical professionals, could use the app to better recognize children's development. The app's reports could help identify children with delayed VMI development or a predicted age significantly behind their chronological age, prompting caregivers and teachers to refer those children to medical facilities for further evaluation. Third, the app's internet connectivity would ensure easy accessibility, particularly benefiting children in remote regions or those who cannot undergo standardized evaluations. Fourth, as more people use the app, the continuous input of new data would help refine the AI model, thereby increasing the accuracy of the predicted age, VMI level, and VMI developmental status. This iterative process would enhance the app's effectiveness over time. Given these advantages, the app has significant potential for application in education and rehabilitation fields. Future research could explore incorporating other origami products into the model prediction to enhance its robustness. Examples include origami airplanes, frogs, and flowers. Allowing children to choose products based on their motivation and preferences could increase the model's utility and make its predictions of VMI development more robust.

Limitations

Several limitations were noted in this study. First, a small sample size of children aged 24 to 29 months (n=10) was used. This limited dataset may have resulted in insufficient data for AI learning, potentially increasing prediction errors for this age group. Second, most children recruited in the study exhibited typical development, and those with developmental conditions had mixed diagnoses. Consequently, the results may predominantly reflect the performance of typically developing children. Moreover, the model may tend to predict no developmental issues, increasing the likelihood of false negative results. Third, the images used for the model prediction were restricted to dog origami products, as only these were used to train the model. This limitation restricts the model's applicability and generalizability to other origami forms. Fourth, due to constraints in evaluation time and research personnel, only the standardized VMI score was used for model prediction. However, executing origami activities also requires manual dexterity, cognition, and visual perception, which were not included in the assessment. Fifth, since the predictions were primarily based on an AI model and conducted via the internet, there are concerns about children's privacy and data security. Furthermore, it is recommended that clinicians ensure

XSL•FO
RenderX

appropriate use and understanding when interpreting the results. Future studies should consider using standardized assessments of other developmental dimensions as model prediction outcomes to provide a more comprehensive understanding of children's development through origami activities. Additionally, increasing the sample size of young children and children with development conditions and incorporating different origami products in model training would enhance the model's utility and robustness.

Conclusion

Overall, it was previously known that origami activities and their products could help therapists identify children's development levels, including developmental age and VMI development, and photographs of origami could predict children's age and VMI. However, we have shown that AI techniques can provide quantitative information about activity products, and photographs of origami dogs can be used as an development evaluation tool for children. In conclusion, our study applied AI techniques to images of origami products to predict children's ages and VMI development. The results demonstrated that children's ages and VMI development could be predicted using photographs of origami dogs. Given that origami served as a successful example, AI techniques appear to have the significant potential for application to various products of children's activities, which could provide more quantitative insights.

Acknowledgments

We gratefully appreciate the children and caregivers who participated in our study. We also thank the principals and teachers from the Nian-Tze Preschool, the affiliated Da-She Elementary School Preschool, the affiliated Yan-Chao Elementary School Preschool, Mary Help of Christians Preschool, the affiliated Dong-Guang Elementary School Preschool, the affiliated Da-Tong Elementary School Preschool, and the New Generation Preschool. This work was supported by the Ministry of Science and Technology, Republic of China (110-2628-B-038-013, 109-2636-B-214-001, and 110-2636-B-214-001).

Data Availability

The datasets generated or analyzed during this study are available from the first author, CYH, upon reasonable request.

Authors' Contributions

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by GHL, CLH and CYH. The first draft of the manuscript was written by CYH and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript. CYH is the guarantor of this paper.

Conflicts of Interest

None declared.

References

1. Arief YS, Binarwati D, Ismiwiranti R. Playing origami enhance the creativity of school aged children. J Ners 2008;3(1):42-48. [doi: 10.20473/jn.v3i1.4979]
2. Imaroonrak S, Phunwutikorn P, Phattharayuttawat S, Ngamthipwatthana T, Sumalrot T, Auampradit N. The effects of origami training on creativity and visual-motor integration in preschool children. J Med Assoc Thailand 2018;101(1):S85-S89. [doi: 10.6594/JTOTA.202112_39(2).0005]
3. Folio MR, Fewell RR. Peabody Developmental Motor Scales, 2nd edition: Pearson; 2000.
4. Hwang AW, Weng LJ, Liao HF. Construct validity of the comprehensive developmental inventory for infants and toddlers. Pediatr Int 2010 Aug;52(4):598-606. [doi: 10.1111/j.1442-200X.2010.03102.x] [Medline: 20202156]
5. Roid GH, Pomplun M. The Stanford-Binet Intelligence Scales, Fifth Edition. In: Contemporary Intellectual Assessment: Theories, Tests, and Issues: The Guilford Press; 2012.
6. Bruininks RH, Bruininks BD. BOT2: Bruininks-Oseretsky Test of Motor Proficiency, 2nd edition: Pearson; 2005.
7. Travers BG, Kirkorian HL, Jiang MJ, et al. Knowing how to fold 'em: paper folding across early childhood. J Mot Learn Dev 2018 Jun;6(1):147-166. [doi: 10.1123/jmld.2016-0044] [Medline: 33981774]
8. Li X, Shi Y. Computer vision imaging based on artificial intelligence. Presented at: 2018 International Conference on Virtual Reality and Intelligent Systems (ICVRIS); Aug 10-11, 2018; Changsha, Hunan, China p. 22-25. [doi: 10.1109/ICVRIS.2018.00014]
9. Hamet P, Tremblay J. Artificial intelligence in medicine. Metab Clin Exp 2017 Apr;69S:S36-S40. [doi: 10.1016/j.metabol.2017.01.011] [Medline: 28126242]
10. Agrawal A, Gans JS, Goldfarb A. Exploring the impact of artificial intelligence: prediction versus judgment. Inf Econ Policy 2019 Jun;47:1-6. [doi: 10.1016/j.infoecopol.2019.05.001]

XSL•FO
RenderX

11. Zhang Y, Ma Y. Application of supervised machine learning algorithms in the classification of sagittal gait patterns of cerebral palsy children with spastic diplegia. Comput Biol Med 2019 Mar;106:33-39. [doi: 10.1016/j.compbiomed.2019.01.009] [Medline: 30665140]
12. Ihlen EAF, Støen R, Boswell L, et al. Machine learning of infant spontaneous movements for the early prediction of cerebral palsy: a multi-site cohort study. J Clin Med 2019 Dec 18;9(1):5. [doi: 10.3390/jcm9010005] [Medline: 31861380]
13. Yu YT, Lin GH, Huang CY, Hsieh CL. Using deep learning to analyze pictures of origami to pedict the age of three-to-six-year-old children. TOTA 2021;39(2):211-231. [doi: 10.6594/jtota.202112_39(2).0005]
14. Beery KE, Buktenica NA, Beery NA. Beery VMI: The Beery-Buktenica Developmental Test of Visual-Motor Integration: Pearson; 2004.
15. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Presented at: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Jun 27-30, 2016; Las Vegas, NV p. 770-778. [doi: 10.1109/CVPR.2016.90]
16. Hershey S, Chaudhuri S, Ellis DPW, et al. CNN architectures for large-scale audio classification. Presented at: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); Mar 5-9, 2017; New Orleans, LA p. 131-135. [doi: 10.1109/ICASSP.2017.7952132]
17. Chen T, Guestrin C. Xgboost: a scalable tree boosting system. Presented at: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Aug 13-17, 2016; San Francisco, CA p. 785-794. [doi: 10.1145/2939672.2939785]
18. Salimans T, Kingma DP. Weight normalization: a simple reparameterization to accelerate training of deep neural networks. Presented at: The 30th Conference on Neural Information Processing Systems; Dec 5-10, 2016; Barcelona, Spain URL: https://dl.acm.org/doi/10.5555/3157096.3157197 [accessed 2024-12-11]
19. Sola J, Sevilla J. Importance of input data normalization for the application of neural networks to complex industrial problems. IEEE Trans Nucl Sci 1997;44(3):1464-1468. [doi: 10.1109/23.589532]

Abbreviations

AI: artificial intelligence
BEERY VMI-6: BEERY VMI Beery-Buktenica Visual-Motor Integration, Sixth Edition
MAE: mean absolute error
VMI: visual-motor integration

Edited by A Mavragani; submitted 15.03.24; peer-reviewed by D Kohen; revised version received 07.10.24; accepted 10.10.24; published 10.01.25.

Please cite as:
Huang CY, Yu YT, Chen KL, Lien JJ, Lin GH, Hsieh CL
Predicting Age and Visual-Motor Integration Using Origami Photographs: Deep Learning Study
JMIR Form Res 2025;9:e58421
URL: https://formative.jmir.org/2025/1/e58421
doi:10.2196/58421

© Chien-Yu Huang, Yen-Ting Yu, Kuan-Lin Chen, Jenn-Jier Lien, Gong-Hong Lin, Ching-Lin Hsieh. Originally published in JMIR Formative Research (https://formative.jmir.org), 10.1.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Comparative Evaluation of Consumer Wearable Devices for Atrial Fibrillation Detection: Validation Study

Femke Wouters[1,2*], MSc; Henri Gruwez[1,2,3,4*], MD, PhD; Christophe Smeets[1,2], MSc, PhD; Anessa Pijalovic[2], MSc; Wouter Wilms[2], MSc; Julie Vranken[1,2], MSc; Zoë Pieters[5], MSc, PhD; Hugo Van Herendael[3], MD; Dieter Nuyens[3], MD, PhD, Prof Dr; Maximo Rivero-Ayerza[3], MD, PhD; Pieter Vandervoort[1,2,3], MD, PhD, Prof Dr; Peter Haemers[4], MD, PhD, Prof Dr; Laurent Pison[1,3], MD, PhD, Prof Dr

[1]Limburg Clinical Research Center/Mobile Health Unit, Faculty of Medicine and Life Sciences, Hasselt University, Hasselt, Belgium

[2]Department Future Health, Ziekenhuis Oost-Limburg, Genk, Belgium

[3]Department of Cardiology, Ziekenhuis Oost-Limburg, Genk, Belgium

[4]Department of Cardiovascular Sciences, KU Leuven, Leuven, Belgium

[5]Data Science Institute, Hasselt University, Hasselt, Belgium

[*]these authors contributed equally

Corresponding Author:
Laurent Pison, MD, PhD, Prof Dr
Limburg Clinical Research Center/Mobile Health Unit, Faculty of Medicine and Life Sciences, Hasselt University, Hasselt, Belgium

Abstract

Background: Consumer-oriented wearable devices (CWDs) such as smartphones and smartwatches have gained prominence for their ability to detect atrial fibrillation (AF) through proprietary algorithms using electrocardiography or photoplethysmography (PPG)–based digital recordings. Despite numerous individual validation studies, a direct comparison of interdevice performance is lacking.

Objective: This study aimed to evaluate and compare the ability of CWDs to distinguish between sinus rhythm and AF.

Methods: Patients exhibiting sinus rhythm or AF were enrolled through a cardiology outpatient clinic. The participants were instructed to perform heart rhythm measurements using a handheld 6-lead electrocardiogram (ECG) device (KardiaMobile 6L), a smartwatch-derived single-lead ECG (Apple Watch), and two PPG-based smartphone apps (FibriCheck and Preventicus) in a random sequence, with simultaneous 12-lead reference ECG as the gold standard.

Results: A total of 122 participants were included in the study: median age 69 (IQR 61-77) years, 63.9% (n=78) men, 25% (n=30) with AF, 9.8% (n=12) without prior smartphone experience, and 73% (n=89) without experience in using a smartwatch. The sensitivity to detect AF was 100% for all devices. The specificity to detect sinus rhythm was 96.4% (95% CI 89.5%-98.8%) for KardiaMobile 6L, 97.8% (95% CI 91.6% - 99.5%) for Apple Watch, 98.9% (95% CI 92.5% - 99.8%) for FibriCheck, and 97.8% (95% CI 91.5% - 99.4%) for Preventicus (*P*=.50). Insufficient quality measurements were observed in 10.7% (95% CI 6.3%-17.5%) of cases for both KardiaMobile 6L and Apple Watch, 7.4% (95% CI 3.9% - 13.6%) for FibriCheck, and 14.8% (95% CI 9.5% - 22.2%) for Preventicus (*P*=.21). Participants preferred Apple Watch over the other devices to monitor their heart rhythm.

Conclusions: In this study population, the discrimination between sinus rhythm and AF using CWDs based on ECG or PPG was highly accurate, with no significant variations in performance across the examined devices.

Trial Registration: ClinicalTrials.gov NCT06023290; https://clinicaltrials.gov/study/NCT06023290

(JMIR Form Res 2025;9:e65139) doi:10.2196/65139

KEYWORDS

atrial fibrillation; AF; mobile health; photoplethysmography; electrocardiography; smartphone; consumer wearable device; wearable devices; detection; electrocardiogram; ECG; mHealth

Introduction

Atrial fibrillation (AF) is the most prevalent cardiac arrhythmia among adults and is associated with an increased risk of stroke and mortality [1-4]. Since AF is often asymptomatic and paroxysmal, its diagnosis can be challenging. Asymptomatic AF frequently remains undetected until a thromboembolic event occurs. The diagnosis of AF requires an electrocardiogram

XSL•FO
RenderX

(ECG) showing irregular R-R intervals (when atrioventricular conduction is not impaired) and the absence of distinct repeating P waves [4].

Historically, ECG diagnostics were confined to medical settings, using devices administered or prescribed by health care professionals. However, the 21st century has witnessed a rapid surge in the availability and use of consumer-oriented wearable devices (CWDs) capable of monitoring heart rhythm [5-8]. Although this trend represents a significant democratization of access to heart rhythm monitoring, it also presents unique challenges for practicing clinicians to diagnose AF based on CWD data, due to the heterogeneous technologies and evolving algorithms used by these devices [9].

The underlying CWD technology is mainly based on electrocardiography or photoplethysmography (PPG). While an ECG measures electrical signals from the heart using surface electrodes on the skin, PPG analyzes the heart rhythm through an optical technique that measures the peripheral pulse [10]. This PPG technology has been adopted by smartwatches and smartphones, using light-emitting diodes and sensors or smartphone flashlights and cameras to generate PPG waveforms. The CWDs analyze ECG or PPG measurements using proprietary algorithms to notify the users of potential AF episodes [9,11,12]. However, only a few CWDs have received medical certification from the US Food and Drug Administration or the Conformité Européenne mark [13].

The emergence of CWDs equipped with automatic heart rate analysis presents a significant opportunity for the early detection of AF. These devices empower patients to monitor their heart rhythms independently, beyond traditional health care settings or without the need for physicians' prescriptions [9,14-16]. This shift enables more frequent and timely heart rhythm assessments, potentially leading to the earlier identification of AF episodes. However, this democratization of health monitoring also implies that physicians have limited involvement in selecting devices used by patients for heart rhythm analysis. Consequently, it becomes imperative for health care providers to adeptly interpret data from CWDs and assess the reliability of automated rhythm classifications of these devices. These skills are crucial to counsel patients who receive AF detections from CWDs effectively, especially as patients increasingly seek guidance on the accuracy of CWD measurements and on selecting the most suitable device.

Despite numerous validation studies of individual CWDs, a comprehensive comparison of their methodologies and outcomes is challenging, which leaves a gap in our understanding of their comparative effectiveness. This study aims to bridge this knowledge gap by evaluating and comparing the ability of digital CWDs to classify heart rhythms, specifically focusing on their accuracy in distinguishing between sinus rhythm and AF compared to the gold standard 12-lead ECG.

Methods

Study Design

Patients were recruited from July to November 2023. The VALIDATION (Validation of Digital Heart Rhythm Devices in the Detection of Atrial Fibrillation) study was a monocentric, prospective, randomized, interventional study that validated the use of PPG- and ECG-based CWDs to detect AF compared to a simultaneous 12-lead ECG in ambulatory cardiology patients in a supervised hospital environment.

Ethical Considerations

The study protocol is in accordance with the Declaration of Helsinki and the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use – Good Clinical Practice guidelines and was approved by 2 medical ethics committees (ie, Ziekenhuis Oost-Limburg, Genk, Belgium, Z-2023023; and Hasselt University, Hasselt, Belgium, B3712023000006). The study was registered at ClinicalTrials.gov (NCT06023290) and meets all STARD 2015 criteria. Informed consent was obtained in writing from all participants. Moreover, they were informed about their right to opt out of the study at any time without consequences. Data collected during the study was pseudonymized to protect participant confidentiality. Unique identifiers were used, and direct personal information was stored separately, accessible only to authorized personnel. Participants were not compensated for their involvement in the study as participation did not require an additional hospital visit and involved minimal additional time or effort.

Study Population

The study employed two distinct methods for participant inclusion, targeting adult patients from the cardiology outpatient clinic (Multimedia Appendix 1). First, all patients visiting the outpatient cardiology clinic were screened by chart review for eligibility on selected days when study personnel were available. Second, to enrich the study population with a higher prevalence of AF, all patients scheduled for electrical cardioversion at the cardiology outpatient clinic were also identified for screening by chart review.

Patients who met the following inclusion and exclusion criteria were invited to participate in the study: (1) age 18 years or older and (2) exhibit sinus rhythm or AF. Exclusion criteria were (1) presence of a pacemaker, (2) participation in another clinical trial that might interfere with the study protocol, and (3) physical or cognitive restrictions, including language barriers that hindered study measurements.

Study Measurements

After the participants provided informed consent, they were instructed by study personnel to perform heart rhythm measurements using 4 CWDs in a random sequence consisting of 2 ECG- and 2 PPG-based devices, according to the manufacturer's instructions. During all measurements, the patients were also instructed to stay still to avoid motion artifacts. All CWDs were provided by the study team to prevent device-induced bias.

KardiaMobile 6L (AliveCor Inc, Mountain View, CA) and a smartwatch, Apple Watch (Apple Inc, Cupertino, CA), were used to derive a single-lead ECG of 30 seconds. A lead-1 ECG was recorded using KardiaMobile 6L by positioning the device on the left knee with both thumbs on the device's front

XSL•FO
RenderX

electrodes [17]. To capture a lead-1 ECG using the Apple Watch, the participant wore the watch on the left wrist, ensuring skin contact with the electrode on the watch's underside. Measurement was initiated by the participant by touching a second electrode on the watch's crown with the right hand's index finger.

The PPG measurements were performed with two smartphone apps: (1) FibriCheck (Qompium NV, Hasselt, Belgium) and (2) Preventicus Heartbeats App (Preventicus GmbH, Jena, Germany). While initiating a PPG measurement in either smartphone app, the camera's flashlight was activated. Participants were instructed by study personnel to place an index finger over the smartphone camera. Using the camera, the app then captured skin color variations caused by heartbeat-induced blood flow changes. These variations were analyzed by the app to generate the PPG waveform.

The final rhythm diagnosis was made by the device's proprietary algorithm for both ECG and PPG measurements. Although 6 leads were measured using the KardiaMobile 6L device, its algorithm only took lead 1 into account. To minimize the occurrence of missing data, measurements were repeated up to 3 attempts until a good-quality measurement was obtained with each device. If good-quality measurements could not be obtained with a specific device, the patient's results were not excluded from the analyses. Good quality was defined as sufficient quality for analysis by the device algorithm. Rhythm diagnosis from CWDs were provided by the proprietary device algorithms and categorized as sinus rhythm or AF (Textbox 1). Concurrently with the CWD measurements, a continuous 12-lead ECG was performed for reference monitoring (MAC5000, GE Healthcare, Chicago, IL). Participants were excluded from the analyses if the heart rhythm changed on the reference ECG during the study conduct. The reference diagnosis on the 12-lead ECG was provided by two cardiologists blinded to the CWD classifications and the participant's clinical information.

XSL•FO
RenderX

Textbox 1. Categorization of consumer-oriented wearable device outcome labels.

Categorized as sinus rhythm

- KardiaMobile 6L
- Normal
- Bradycardia
- Tachycardia
- Apple Watch
- Sinus rhythm
- High heart rate - no atrial fibrillation
- Low or high heart rate
- FibriCheck
- Normal
- Mild irregularities
- Preventicus
- Normal
- Mild irregularities

Categorized as atrial fibrillation

- KardiaMobile 6L
- Possible atrial fibrillation
- Apple Watch
- Atrial fibrillation
- Atrial fibrillation - high heart rate
- FibriCheck
- Possible atrial fibrillation
- Preventicus
- Possible atrial fibrillation

Categorized as insufficient quality

- KardiaMobile 6L
- No analysis
- Unclassified
- Unreadable
- Apple Watch
- Inconclusive
- Poor reading
- FibriCheck
- Insufficient quality

XSL•FO
RenderX

- Preventicus
- Insufficient quality

Study Questionnaire

Following the completion of study measurements, participants were provided a 5-point Likert scale questionnaire designed to assess their experiences with and preferences regarding the devices used in the study. Additionally, the questionnaire aimed to evaluate the participants' digital literacy levels prior to their involvement in the study.

Data Analysis

Data analysis was performed using SPSS software (version 29.0; IBM Corp) and SAS (version 9.4; SAS Institute). $P<.05$ was considered statistically significant. The Shapiro-Wilk statistic was used to assess the normality of continuous data. Continuous variables are presented as median and IQR as appropriate and were compared using the Mann-Whitney U test. Discrete variables are presented as absolute numbers and percentages (%) and compared using Pearson χ^2 test. The performance of the CWDs were evaluated by calculating their sensitivity, specificity, and accuracy compared to the 12-lead reference ECG. These metrics were estimated and compared between CWDs using a generalized estimation equation model, accounting for the correlated nature of the measurements of the same patient. Measurements of insufficient quality were not included in these analyses.

Results

Overview

On designated study days, 200 patients (145 attending outpatient consultations and 55 undergoing elective cardioversions) were invited to participate in the study. Among these, 123 subjects were included, of which 122 completed the study protocol (see Multimedia Appendix 1). The remaining 77 patients either declined participation due to a lack of interest or were unwilling to commit the additional time required for study procedures. One participant with paroxysmal AF, who exhibited both sinus rhythm and AF on the reference ECG during the study, was excluded from the analysis to avoid confounding results. This decision was made because the specific rhythm present during the CWD recordings could not be reliably determined.

Study Population Characteristics

The median age of the study population was 69 (IQR 61-77) years, with 44 of the 122 (36.1%) participants being women. All participants were White. Detailed demographic and clinical characteristics are given in Table 1, and information on drug therapy is provided in Multimedia Appendix 2. There were no adverse events related to the use of the CWDs. Based on the 12-lead reference ECG, AF was present in 30 (24.6%) patients. The majority of the study population demonstrated a fair level of digital literacy, with 91 (74.6%) participants reporting daily internet usage and 96 (78.7%) using smartphones daily, respectively, as illustrated in Figure 1. However, 16 (13.1%) participants were identified as potentially digitally illiterate, having no prior experience with the internet or a smartphone. Furthermore, 89 (73%) participants reported no previous experience with CWDs, except smartphones.

XSL•FO
RenderX

Table . Demographic and clinical characteristics of the study population.

Participant characteristics	Participants			*P* value[a]
	Overall (n=122)	SR[b] (n=92)	AF[c] (n=30)	
Age (years), median (IQR)	69 (61-77)	68 (60.25 - 77.0)	72.5 (63.5 - 78.5)	.17
Gender (women), n (%)	44 (36.1)	34 (37)	10 (33.3)	.72
BMI (kg/m^2), median (IQR)	26.9 (24.1 - 30.8)	26.9 (23.5 - 30.3)	28.2 (25.7 - 33.6)	.02
Hypertension, n (%)	69 (56.6)	51 (55.4)	18 (60)	.66
Diabetes, n (%)	21 (17.2)	17 (18.5)	4 (13.3)	.52
Hypercholesterolemia, n (%)	83 (68)	65 (70.7)	18 (60)	.28
Heart failure, n (%)	21 (17.2)	12 (13)	9 (30)	.03
Vascular disease history, n (%)	26 (21.3)	22 (23.9)	4 (13.3)	.22
Stroke, n (%)	19 (15.6)	15 (16.3)	4 (13.3)	≥.99
CHA_2DS_2-VASc scores[d], n (%)				.50
0	12 (9.8)	9 (9.8)	3 (10)	
1	22 (18)	16 (17.4)	6 (20)	
2	21 (17.2)	18 (19.6)	3 (10)	
3	33 (27)	23 (25)	10 (33.3)	
4	21 (17.2)	18 (19.6)	3 (10)	
5	6 (4.9)	3 (3.3)	3 (10)	
6	7 (5.7)	5 (5.4)	2 (6.7)	

[a]Chi-square test, Fisher exact test, and Mann-Whitney *U* test were used in the analyses.

[b]SR: sinus rhythm.

[c]AF: atrial fibrillation.

[d]CHA_2DS_2-VASc is a points-based system used to stratify the risk of stroke in patients with AF. CHA_2DS_2-VASc is calculated as congestive heart failure, hypertension, age ≥75 years (doubled), diabetes, stroke (doubled), vascular disease, aged 65-74 years, and sex category (female) [18].

XSL•FO
RenderX

Figure 1. Experience with digital technology. Questions: I use the internet; I use a computer and/or laptop; I use a smartphone and/or tablet; I use a wearable device (eg, fitness tracker, smartwatch).

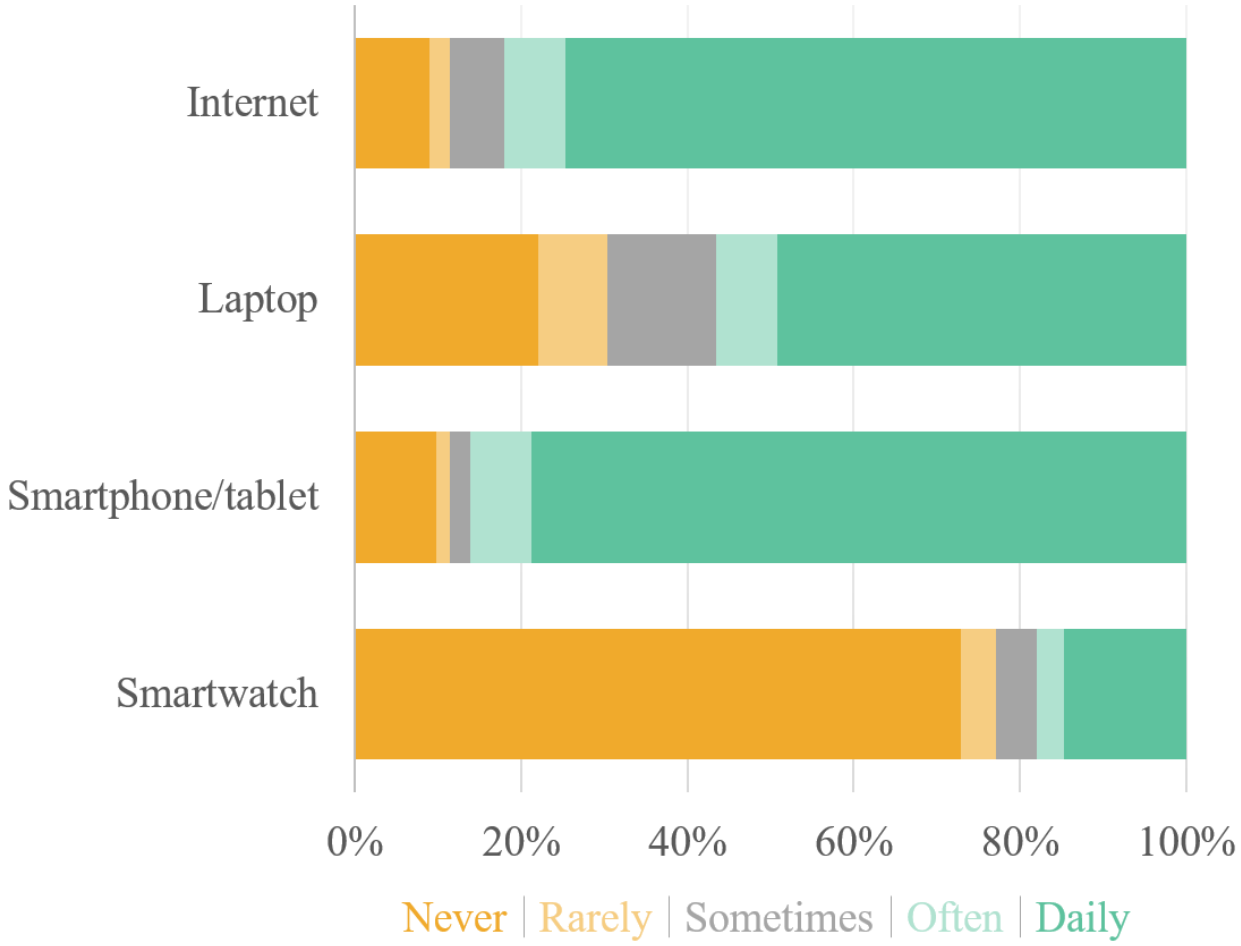

Signal Quality

Measurements of insufficient quality were repeated until the CWD algorithm provided a rhythm diagnosis. The number of additional measurements required for diagnosis is illustrated in Figure 2. Good-quality measurements could not be obtained after three attempts in 9 (7.4%) patients with KardiaMobile 6L, 4 (3.3%) with Apple Watch, 2 (1.6%) with FibriCheck, and 4 (3.3%) with Preventicus. Achieving good-quality measurements required 5 additional attempts for KardiaMobile 6L, 11 for Apple Watch, 7 for FibriCheck, and 18 for Preventicus. Consequently, the prevalence of insufficient quality measurements was 10.7% (95% CI 6.3%-17.5%) for both KardiaMobile 6L and Apple Watch, 7.4% (95% CI 3.9%-13.6%) for FibriCheck, and 14.8% (95% CI 9.5%-22.2%) for Preventicus, with no significant differences between the devices (P=.21). Furthermore, the distribution of insufficient quality measurements was comparable between patients in sinus rhythm and AF (11.7% vs 8.3%, P=.31).

XSL•FO
RenderX

Figure 2. Signal quality for each digital device. The flowchart describes the number of patients in which no good-quality measurement could be obtained after three attempts and the number of attempts necessary to obtain a good-quality measurement for each device.

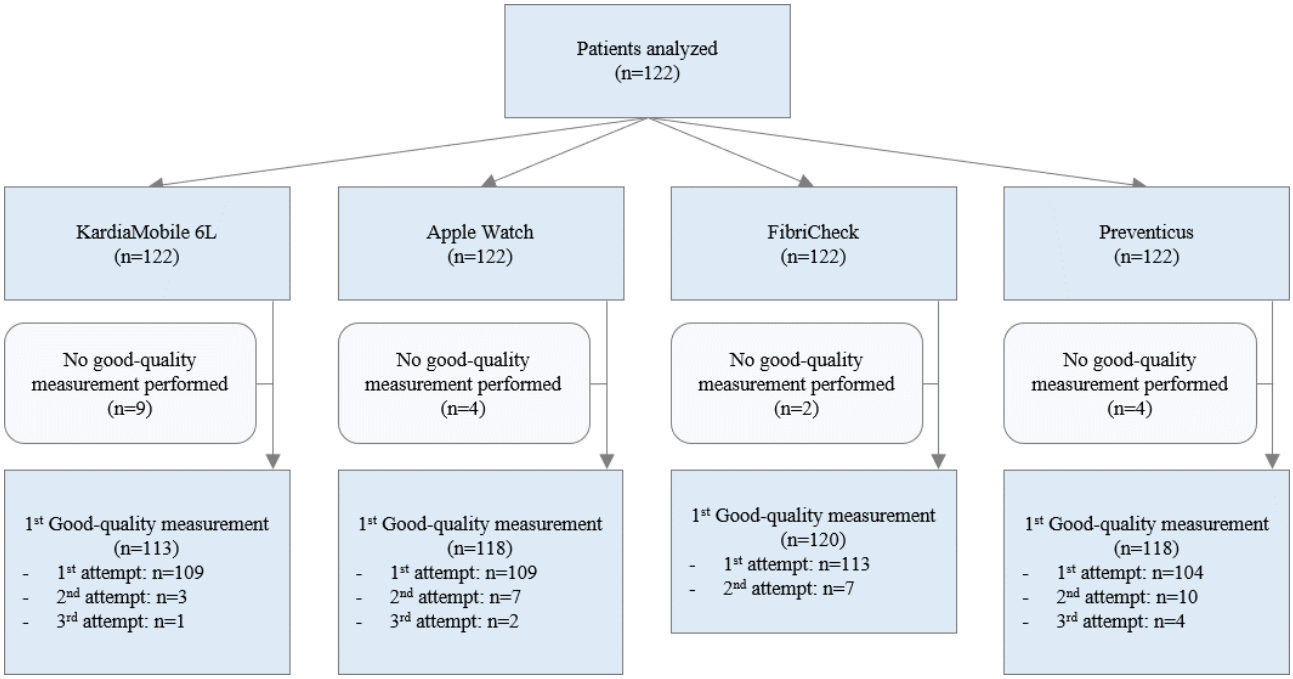

Patients with insufficient quality recordings from both ECG and PPG devices were significantly older, with a median age of 83 (IQR 77-86) years compared to 68 (IQR 60-76) years (*P*<.001). However, there was no significant age difference between patients with insufficient quality recordings from either ECG or PPG devices and those with only good-quality recordings with a median age of 71 (IQR 62-80) years and 68 (IQR 60-75) years, respectively (*P*=.14). Similarly, no significant age difference was observed among patients with insufficient quality recordings (median age 74, IQR 62-80 years) versus those with only good-quality recordings (median age 68, IQR 60-75 years) from PPG devices specifically (*P*=.14). For ECG devices, measurements of insufficient quality were more often performed by older patients with a median age of 74 (IQR 64-82) years versus 68 (IQR 60-75) years (*P*=.02). Despite reduced digital literacy in older patients (*P*<.001), there was no significant association between digital literacy and the occurrence of insufficient data quality from either ECG- or PPG-based CWDs (*P*=.14).

Heart Rhythm Classification

Based on the first good-quality measurement, all patients with AF were correctly identified by all CWDs (Table 2). Therefore, the sensitivity for all CWDs to detect AF was 100%. The specificity rates for identifying sinus rhythm were 96.4% for KardiaMobile 6L, 97.8% for Apple Watch, 98.9% for FibriCheck, and 97.8% for Preventicus. The differences in specificity between devices were not statistically significant (*P*=.50). The diagnostic performance of all study devices is detailed in Table 3.

XSL•FO
RenderX

Table . Cross-tabulation of the consumer-oriented wearable device results compared to the 12-lead reference electrocardiogram (ECG).

	12-lead ECG rhythm (n=113)	
	SR[a]	AF[b]
KardiaMobile 6L, n/N[c] (%)		
SR	81/84 (96.4)	0/29 (0)
AF	3/84 (3.6)	29/29 (100)
Apple Watch, n/N (%)		
SR	89/91 (97.8)	0/27 (0)
AF	2/91 (2.2)	27/27 (100)
FibriCheck, n/N (%)		
SR	89/90 (98.9)	0/30 (0)
AF	1/90 (1.1)	30/30 (100)
Preventicus, n/N (%)		
SR	87/89 (97.8)	0/29 (0)
AF	2/89 (2.2)	29/29 (100)

[a]SR: sinus rhythm.

[b]AF: atrial fibrillation.

[c]N represents the number of detections by the 12-lead reference ECG for the SR and AF columns.

Table . Comparison of performance of the consumer-oriented wearable devices compared to the 12–lead reference electrocardiogram.[a]

Performance metrics	KardiaMobile 6L[b]	Apple Watch[b]	FibriCheck[c]	Preventicus[c]	*P* value
Good-quality tracing, % (95% CI)	89.3 (82.5-93.7)	89.3 (82.5-93.7)	92.6 (86.4-96.1)	85.3 (77.8-90.5)	.21
Sensitivity, % (95% CI)	100	100	100	100	__[d]
Specificity, % (95% CI)	96.4 (89.5-98.8)	97.8 (91.6-99.5)	98.9 (92.5-99.8)	97.8 (91.5-99.4)	.50
Accuracy, % (95% CI)	97.4 (92.1-99.1)	98.3 (93.5-99.6)	99.2 (94.3-99.9)	98.3 (93.5-99.6)	.48

[a]Generalized estimation equation model was used for the analysis.

[b]Single-lead electrocardiogram (lead 1).

[c]Smartphone photoplethysmography (finger over the camera).

[d]Not applicable.

Study Questionnaire

After performing at least two measurements with each CWD, a significant majority of participants (over 80%) rated the study devices favorably across various criteria, including user-friendliness, likelihood of sustained use, design and layout, interesting content, overall satisfaction, and likelihood of recommending the devices to others (Figure 3). In these evaluations, the Apple Watch scored higher on the questionnaire, followed by FibriCheck, Preventicus, and KardiaMobile 6L. Despite these positive assessments, cost emerged as the principal barrier to broader adoption and implementation of these devices.

XSL•FO
RenderX

Figure 3. Study questionnaire responses. Questions: This device / app was easy to use; I would use this device / app again; I like the design and layout of this device / app; the content of this device / app is interesting to me; in general, I am satisfied with this device / app; I would recommend this device / app to people who would benefit from it; I would pay for this device / app (no price disclosed).

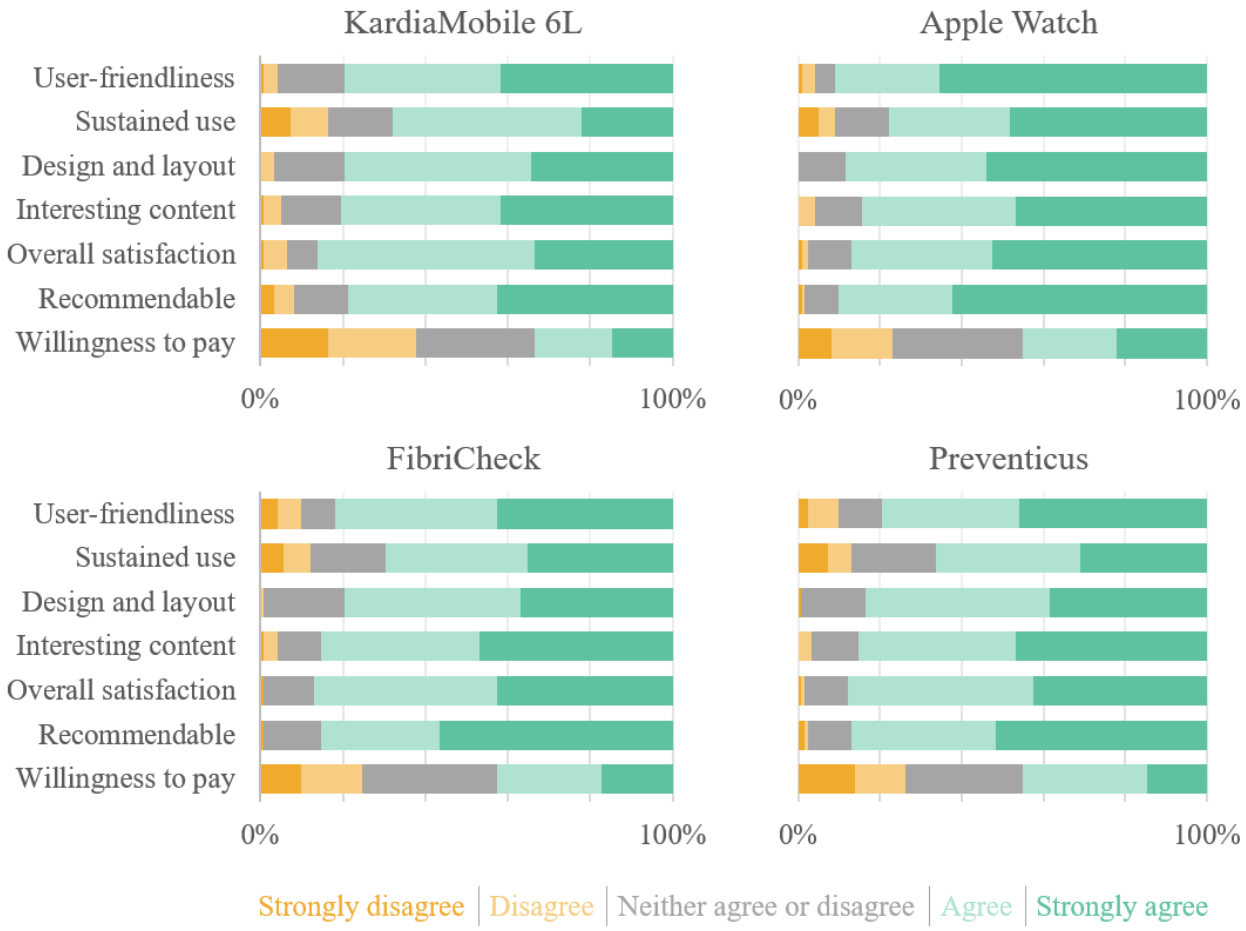

Discussion

Principal Findings

This study validated the ability of four CWDs to differentiate between sinus rhythm and AF using ECG or PPG technology and their proprietary algorithms. The devices were compared against a concurrently obtained 12-lead reference ECG, demonstrating very high sensitivity and specificity, thus confirming a near-perfect association with the diagnosis made using the 12-lead reference ECG.

To our knowledge, this is the first study to directly compare the accuracy of automated AF detection by ECG- and PPG-based CWDs against simultaneous 12-lead reference ECG monitoring. These findings align with prior studies that separately validated PPG and single-lead ECG CWDs [17,19-21]. However, the exclusion of patients and handling of uninterpretable results can significantly impact outcomes and hamper comparisons between studies [17]. A meta-analysis of 28 validation studies of PPG-based CWDs revealed significant biases, mainly concerning participant selection and insufficient information on study flow, design, and timing [19].

Addressing the limitations of previous studies, this trial adopted an inclusive approach by including all patients with a scheduled visit on designated study days, conducting simultaneous reference and index tests, and randomizing the sequence of index tests to minimize biases. However, a large number of the patients could not be screened or did not meet inclusion criteria, limiting findings to a specific demographic of adult White patients without a pacemaker in AF or sinus rhythm without frequent extrasystoles, who were capable of performing CWD measurements. Unlike previous studies, which have shown a greater incidence of uninterpretable measurements in up to one-third of cases, this study limited exclusions to 7.4%-14.8% by iterating measurements until a diagnosis was obtained. These measurements were excluded from this validation study, leading to an overestimation of the algorithm's accuracy [17,21,22]. This approach reflects real-world use, where users typically repeat measurements until a diagnosis is obtained from the device. Older patients, expectedly, were more likely to produce recordings of insufficient quality, frequently due to tremors.

Future validation studies are needed to validate these results in non-White populations and in individuals with rhythms other than sinus rhythm or AF, or with frequent extrasystoles that may challenge the algorithm's specificity. In our study, false-positive results from both ECG- and PPG-based CWDs were more likely when the 12-lead reference ECG showed premature atrial or ventricular contractions. These findings suggest that frequent extrasystoles can lead to rhythm misclassifications, highlighting the need for further research into how these irregularities affect device performance,

XSL•FO
RenderX

particularly in populations with diverse arrhythmias or more frequent extrasystoles. The study population included patients with both high and low digital literacy, which did not affect their performance in a supervised hospital environment. However, CWDs are mostly used in unsupervised settings. Recent real-world validation studies in unsupervised environments have exhibited similar results [23,24].

This study has implications for clinical practice as it provides physicians with information about the accuracy of CWDs. Physicians are likely to encounter CWDs, either by the patient's choice or owing to these devices being marketed and sold directly to the consumer. This comparative validation study of several devices with heterogeneous technologies and algorithms enables physicians to counsel patients on measurement results or the selection of medically certified devices. Similar studies should be repeated over time as algorithms are frequently updated, potentially improving diagnostic accuracy.

Given the high prevalence of uninterpretable device ECGs, mass utilization by the general population may lead to an increased and unnecessary health care burden [25]. To support clinical decisions based on CWD heart rhythm measurements, physicians should not rely solely on algorithm results, instead they should interpret traces manually when required. Although manual interpretation of CWDs was not within the scope of this study, it was investigated in the INTERPRET-AF study which reported similar accuracy for PPG- and ECG-based CWDs upon interpretation by physicians, with only modest reductions compared to the 12-lead ECG, even without prior PPG training [26].

The potential of CWDs in detecting new AF cases in high-risk populations (eg, secondary prevention in cryptogenic stroke patients) or monitoring AF recurrence in patients with an established diagnosis highlights the evolving landscape of cardiac care [23,27]. Despite increased AF detection and cost-effectiveness of these implementations, the stroke risk and therapeutic implications for AF detected by screening remain debatable [28]. Patients' willingness to adopt these CWDs was reported in over 80% of the questionnaire responses, signaling an increasing likelihood of implementation in the future.

Limitations

This study has several limitations. First, tracings deemed to be of insufficient quality were excluded from the analyses, potentially overestimating the performance and influencing the generalizability of the results. Second, patients with frequent extrasystoles or rhythms other than sinus rhythm or AF, as documented in their medical records, were excluded from this study, which could have reduced the specificity. Nevertheless, extrasystoles that occurred during the study procedures were included in the analyses, leading to false positive findings. The results of the CWD's algorithms were categorized into only two categories (ie, sinus rhythm and AF), without accounting for a broader range of differential diagnoses. Third, participants received instructions on device usage and were corrected when necessary to optimize signal quality. Fourth, this study could not evaluate the effect of skin color on the diagnostic accuracy of these devices. Fifth, the algorithm of the CWD with multiple ECG leads (ie, KardiaMobile 6L), which may have increased the specificity, was evaluated using 1 lead [17].

Conclusions

In this selected population, the distinction between sinus rhythm and AF using CWDs based on ECG or PPG was highly accurate, with no discernible variations observed across the examined devices.

Acknowledgments

FW is supported as a predoctoral student by the Bijzonder Onderzoeksfonds of Hasselt University (BOF20DOC17). HG is supported as a predoctoral strategic basic research fellow by the Fund for Scientific Research Flanders (FWO 1S83221N). This study is part of Limburg Clinical Research Center, supported by the foundation Limburg Sterk Merk, province of Limburg, Flemish government, Hasselt University, Ziekenhuis Oost-Limburg, and Jessa Hospital. The consumer-oriented wearable devices were provided by Qompium NV. The authors would like to thank Daimy Roebroek for her contribution to this work.

Authors' Contributions

Methodology: FW, HG, CS
Investigation: WW, AP, HG, FW
Statistical analysis: ZP, FW, HG
Writing – original draft: FW, HG
Writing – review and editing: FW, HG, CS, JV, HVH, DN, MRA, PV, PH, LP

Conflicts of Interest

None declared.

Multimedia Appendix 1
Flowchart of study screening and inclusion. AF: atrial fibrillation; SR: sinus rhythm; SVES: supraventricular extrasystoles; VES: ventricular extrasystoles.
[PNG File, 116 KB - formative_v9i1e65139_app1.png]

XSL•FO
RenderX

Multimedia Appendix 2
Medication use of the study population. Chi-square test was used in the analyses. ACE: angiotensin-converting enzyme; AF: atrial fibrillation; ARB, angiotensin receptor blockers; CCB: calcium channel blockers; OAC: oral anticoagulation; SR: sinus rhythm.
[PNG File, 134 KB - formative_v9i1e65139_app2.png]

References

1. Wolf PA, Abbott RD, Kannel WB. Atrial fibrillation as an independent risk factor for stroke: the Framingham Study. Stroke 1991 Aug;22(8):983-988. [doi: 10.1161/01.str.22.8.983] [Medline: 1866765]
2. McAllen PM, Marshall J. Cardiac dysrhythmia and transient cerebral ischaemic attacks. Lancet 1973 Jun 2;1(7814):1212-1214. [doi: 10.1016/s0140-6736(73)90527-8] [Medline: 4122563]
3. Chugh SS, Blackshear JL, Shen WK, Hammill SC, Gersh BJ. Epidemiology and natural history of atrial fibrillation: clinical implications. J Am Coll Cardiol 2001 Feb;37(2):371-378. [doi: 10.1016/s0735-1097(00)01107-4] [Medline: 11216949]
4. Hindricks G, Potpara T, Dagres N, et al. Corrigendum to: 2020 ESC Guidelines for the diagnosis and management of atrial fibrillation developed in collaboration with the European Association for Cardio-Thoracic Surgery (EACTS): The Task Force for the diagnosis and management of atrial fibrillation of the European Society of Cardiology (ESC) developed with the special contribution of the European Heart Rhythm Association (EHRA) of the ESC. Eur Heart J 2021 Oct 21;42(40):4194-4194. [doi: 10.1093/eurheartj/ehab648]
5. Dagher L, Shi H, Zhao Y, Marrouche NF. Wearables in cardiology: here to stay. Heart Rhythm 2020 May;17(5 Pt B):889-895. [doi: 10.1016/j.hrthm.2020.02.023] [Medline: 32354455]
6. Seneviratne MG, Connolly SB, Martin SS, Parakh K. Grains of sand to clinical pearls: realizing the potential of wearable data. Am J Med 2023 Feb;136(2):136-142. [doi: 10.1016/j.amjmed.2022.10.006] [Medline: 36351523]
7. Hong W. Advances and opportunities of mobile health in the postpandemic era: smartphonization of wearable devices and wearable deviceization of smartphones. JMIR Mhealth Uhealth 2024 Jan 22;12:e48803. [doi: 10.2196/48803] [Medline: 38252596]
8. Nazarian S, Lam K, Darzi A, Ashrafian H. Diagnostic accuracy of smartwatches for the detection of cardiac arrhythmia: systematic review and meta-analysis. J Med Internet Res 2021 Aug 27;23(8):e28974. [doi: 10.2196/28974] [Medline: 34448706]
9. Svennberg E, Tjong F, Goette A, et al. How to use digital devices to detect and manage arrhythmias: an EHRA practical guide. Europace 2022 Jul 15;24(6):979-1005. [doi: 10.1093/europace/euac038] [Medline: 35368065]
10. Allen J. Photoplethysmography and its application in clinical physiological measurement. Physiol Meas 2007 Mar;28(3):R1-39. [doi: 10.1088/0967-3334/28/3/R01] [Medline: 17322588]
11. Varma N, Cygankiewicz I, Turakhia MP, et al. 2021 ISHNE/HRS/EHRA/APHRS collaborative statement on mHealth in arrhythmia management: digital medical tools for heart rhythm professionals: from the International Society for Holter and Noninvasive Electrocardiology/Heart Rhythm Society/European Heart Rhythm Association/Asia Pacific Heart Rhythm Society. Cardiovasc Digit Health J 2021 Feb;2(1):4-54. [doi: 10.1016/j.cvdhj.2020.11.004] [Medline: 35265889]
12. Narayan SM, Wan EY, Andrade JG, et al. Visions for digital integrated cardiovascular care: HRS Digital Health Committee perspectives. Cardiovasc Digit Health J 2024 Apr;5(2):37-49. [doi: 10.1016/j.cvdhj.2024.02.003] [Medline: 38765620]
13. Baman JR, Mathew DT, Jiang M, Passman RS. Mobile health for arrhythmia diagnosis and management. J Gen Intern Med 2022 Jan;37(1):188-197. [doi: 10.1007/s11606-021-07007-w] [Medline: 34282532]
14. Lopez Perales CR, Van Spall HGC, Maeda S, et al. Mobile health applications for the detection of atrial fibrillation: a systematic review. EP Europace 2021 Jan 27;23(1):11-28. [doi: 10.1093/europace/euaa139]
15. Bonini N, Vitolo M, Imberti JF, et al. Mobile health technology in atrial fibrillation. Expert Rev Med Devices 2022 Apr;19(4):327-340. [doi: 10.1080/17434440.2022.2070005] [Medline: 35451347]
16. Klier K, Koch L, Graf L, Schinköthe T, Schmidt A. Diagnostic accuracy of single-lead electrocardiograms using the Kardia Mobile App and the Apple Watch 4: validation study. JMIR Cardio 2023 Nov 23;7:e50701. [doi: 10.2196/50701] [Medline: 37995111]
17. Scholten J, Jansen WPJ, Horsthuis T, et al. Six-lead device superior to single-lead smartwatch ECG in atrial fibrillation detection. Am Heart J 2022 Nov;253(53-8):53-58. [doi: 10.1016/j.ahj.2022.06.010] [Medline: 35850242]
18. Lip GYH, Nieuwlaat R, Pisters R, Lane DA, Crijns H. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the Euro Heart Survey on Atrial Fibrillation. Chest 2010 Feb;137(2):263-272. [doi: 10.1378/chest.09-1584] [Medline: 19762550]
19. Gill S, Bunting KV, Sartini C, et al. Smartphone detection of atrial fibrillation using photoplethysmography: a systematic review and meta-analysis. Heart 2022 Sep 26;108(20):1600-1607. [doi: 10.1136/heartjnl-2021-320417] [Medline: 35277454]
20. O'Sullivan JW, Grigg S, Crawford W, et al. Accuracy of smartphone camera applications for detecting atrial fibrillation: a systematic review and meta-analysis. JAMA Netw Open 2020 Apr 1;3(4):e202064. [doi: 10.1001/jamanetworkopen.2020.2064] [Medline: 32242908]

XSL•FO
RenderX

21. Mannhart D, Lischer M, Knecht S, et al. Clinical validation of 5 direct-to-consumer wearable smart devices to detect atrial fibrillation: BASEL wearable study. JACC Clin Electrophysiol 2023 Feb;9(2):232-242. [doi: 10.1016/j.jacep.2022.09.011] [Medline: 36858690]
22. Seshadri DR, Bittel B, Browsky D, et al. Accuracy of Apple Watch for detection of atrial fibrillation. Circulation 2020 Feb 25;141(8):702-703. [doi: 10.1161/CIRCULATIONAHA.119.044126] [Medline: 32091929]
23. Gruwez H, Ezzat D, Van Puyvelde T, et al. Real-world validation of smartphone-based photoplethysmography for rate and rhythm monitoring in atrial fibrillation. Europace 2024 Mar 30;26(4):38630867. [doi: 10.1093/europace/euae065]
24. Fernstad J, Svennberg E, Åberg P, Kemp Gudmundsdottir K, Jansson A, Engdahl J. Validation of a novel smartphone-based photoplethysmographic method for ambulatory heart rhythm diagnostics: the SMARTBEATS study. Europace 2024 Mar 30;26(4):euae079. [doi: 10.1093/europace/euae079] [Medline: 38533836]
25. Perez MV, Mahaffey KW, Hedlin H, et al. Large-scale assessment of a smartwatch to identify atrial fibrillation. N Engl J Med 2019 Nov 14;381(20):1909-1917. [doi: 10.1056/NEJMoa1901183] [Medline: 31722151]
26. Gruwez H, Evens S, Proesmans T, et al. Accuracy of physicians interpreting photoplethysmography and electrocardiography tracings to detect atrial fibrillation: INTERPRET-AF. Front Cardiovasc Med 2021;8(734737):734737. [doi: 10.3389/fcvm.2021.734737] [Medline: 34616786]
27. Wouters F, Gruwez H, Vranken J, et al. The potential and limitations of mobile health and insertable cardiac monitors in the detection of atrial fibrillation in cryptogenic stroke patients: preliminary results from the REMOTE trial. Front Cardiovasc Med 2022;9:848914. [doi: 10.3389/fcvm.2022.848914] [Medline: 35498000]
28. Lowres N, Neubeck L, Salkeld G, et al. Feasibility and cost-effectiveness of stroke prevention through community screening for atrial fibrillation using iPhone ECG in pharmacies. The SEARCH-AF study. Thromb Haemost 2014 Jun;111(6):1167-1176. [doi: 10.1160/TH14-03-0231] [Medline: 24687081]

Abbreviations

AF: atrial fibrillation
CWD: consumer-oriented wearable device
ECG: electrocardiogram
PPG: photoplethysmography
VALIDATION: Validation of Digital Heart Rhythm Devices in the Detection of Atrial Fibrillation

Edited by A Mavragani; submitted 06.08.24; peer-reviewed by D Sokas; revised version received 05.11.24; accepted 19.11.24; published 09.01.25.

Please cite as:
Wouters F, Gruwez H, Smeets C, Pijalovic A, Wilms W, Vranken J, Pieters Z, Van Herendael H, Nuyens D, Rivero-Ayerza M, Vandervoort P, Haemers P, Pison L
Comparative Evaluation of Consumer Wearable Devices for Atrial Fibrillation Detection: Validation Study
JMIR Form Res 2025;9:e65139
URL: https://formative.jmir.org/2025/1/e65139
doi:10.2196/65139

©Femke Wouters, Henri Gruwez, Christophe Smeets, Anessa Pijalovic, Wouter Wilms, Julie Vranken, Zoë Pieters, Hugo Van Herendael, Dieter Nuyens, Maximo Rivero-Ayerza, Pieter Vandervoort, Peter Haemers, Laurent Pison. Originally published in JMIR Formative Research (https://formative.jmir.org), 9.1.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Real-Time Analytics and AI for Managing No-Show Appointments in Primary Health Care in the United Arab Emirates: Before-and-After Study

Yousif Mohamed AlSerkal[1*], BMed, MSc, MBBS, BAO; Naseem Mohamed Ibrahim[1*], MBBCH, Arab Board Speciality; Aisha Suhail Alsereidi[1*]; Mubaraka Ibrahim[1*]; Sudheer Kurakula[2*]; Sadaf Ahsan Naqvi[2*], MBBS, SQIL; Yasir Khan[2*]; Neema Preman Oottumadathil[2*]

[1]Emirates Health Services, Dubai, United Arab Emirates

[2]Oracle, The Edge Building, Al Falak Street, Dubai Internet City, Dubai, United Arab Emirates

[*]all authors contributed equally

Corresponding Author:
Sadaf Ahsan Naqvi, MBBS, SQIL
Oracle, The Edge Building, Al Falak Street, Dubai Internet City, Dubai, United Arab Emirates

Abstract

Background: Primary health care (PHC) services face operational challenges due to high patient volumes, leading to complex management needs. Patients access services through booked appointments and walk-in visits, with walk-in visits often facing longer waiting times. No-show appointments are significant contributors to inefficiency in PHC operations, which can lead to an estimated 3%-14% revenue loss, disrupt resource allocation, and negatively impact health care quality. Emirates Health Services (EHS) PHC centers handle over 140,000 visits monthly. Baseline data indicate a 21% no-show rate and an average patient wait time exceeding 16 minutes, necessitating an advanced scheduling and resource management system to enhance patient experiences and operational efficiency.

Objective: The objective of this study was to evaluate the impact of an artificial intelligence (AI)-driven solution that was integrated with an interactive real-time data dashboard on reducing no-show appointments and improving patient waiting times at the EHS PHCs.

Methods: This study introduced an innovative AI-based data application to enhance PHC efficiency. Leveraging our electronic health record system, we deployed an AI model with an 86% accuracy rate to predict no-shows by analyzing historical data and categorizing appointments based on no-show risk. The model was integrated with a real-time dashboard to monitor patient journeys and wait times. Clinic coordinators used the dashboard to proactively manage high-risk appointments and optimize resource allocation. The intervention was assessed through a before-and-after comparison of PHC appointment dynamics and wait times, analyzing data from 135,393 appointments (67,429 before implementation and 67,964 after implementation).

Results: Implementation of the AI-powered no-show prediction model resulted in a significant 50.7% reduction in no-show rates ($P<.001$). The odds ratio for no-shows after implementation was 0.43 (95% CI 0.42-0.45; $P<.001$), indicating a 57% reduction in the likelihood of no-shows. Additionally, patient wait times decreased by an average of 5.7 minutes overall ($P<.001$), with some PHCs achieving up to a 50% reduction in wait times.

Conclusions: This project demonstrates that integrating AI with a data analytics platform and an electronic health record systems can significantly improve operational efficiency and patient satisfaction in PHC settings. The AI model enabled daily assessments of wait times and allowed for real-time adjustments, such as reallocating patients to different clinicians, thus reducing wait times and optimizing resource use. These findings illustrate the transformative potential of AI and real-time data analytics in health care delivery.

(JMIR Form Res 2025;9:e64936) doi:10.2196/64936

KEYWORDS

electronic health record; EHR; artificial intelligence; AI; no-show appointments; real-time data; primary health care; risk prediction; clinic waiting time; operational efficiency

XSL•FO
RenderX

Introduction

Primary health care (PHC) centers are the initial contact point for health care and are considered the gateway to the health care system for a large proportion of the population [1,2]. PHC centers experience a high daily patient influx, making PHC operations complex and requiring efficient systems. While many patients book appointments in advance, others rely on walk-ins, often leading to longer wait times for older and pediatric patients who require quicker attention. However, a significant issue is no-show appointments, where patients fail to attend their scheduled appointments without prior cancellation.

No-show appointments create a significant burden by unnecessarily blocking health care resources. They can negatively impact the financial and operational planning of health care organizations, leading to an estimated revenue loss of 3% - 14% [3,4]. Various interventions have been proposed to address this issue, focusing on understanding patient behaviors and reducing no-show rates [5]. Many organizations use data analytics to predict no-show probabilities and devise interventions to improve clinical efficiencies [6-8]. Literature suggests that while machine learning models have been explored to predict no-show likelihoods [9], a gap exists in operationalizing these insights at a large scale, especially in settings like the Emirates Health Services (EHS), which manages a vast network of health care facilities.

Another key indicator of patient satisfaction is waiting time, which directly affects the quality of care and patient experience [10-12]. Numerous studies have demonstrated that reducing patient waiting times can significantly enhance patient satisfaction and overall service quality. These studies highlight that multiple interventions, such as resource optimization and system redesign, are necessary to improve operational efficiency. Identifying bottlenecks and implementing practical interventions can significantly reduce waiting times [13-17]. With the evolution of artificial intelligence (AI), organizations can now leverage real-time data analytics to develop predictive models and suggest proactive interventions.

While there are examples from various health care settings demonstrating the successful use of AI models and patient reminders to reduce no-shows [18,19], there is limited research on how such systems can be optimized specifically for large health care networks like the EHS. Several studies highlight the importance of real-time analytics in improving health care outcomes, allowing organizations to respond promptly to emerging challenges and optimize resource allocation [20-22]. Hence, this study aims to address this gap by providing an electronic health record (EHR)-driven, real-time analytics solution tailored to the operational needs of PHC centers.

The EHS is a vast health care network across 6 emirates in the United Arab Emirates, managing hospitals, clinics, specialized centers, and supporting services [23]. The PHC centers within the EHS handle over 140,000 visits each month, resulting in a high patient turnover. While this often leads to longer-than-expected wait times negatively impacting patient experience, some appointment slots remain underused due to an average no-shows rate of 21%. Baseline data show the average physician waiting time exceeds 16 minutes, and while resources are distributed based on center needs, occasional shortages contribute to operational delays.

Considering these challenges, our study aims to evaluate the impact of a real-time dashboard that projects operational metrics and integrates an AI model to predict no-show appointments in PHC centers at the EHS. By using this AI-enriched dashboard, clinic administrators can identify high-risk patients (risk of no-show) and proactively contact them to emphasize the importance of their scheduled visits. The primary goals of this evaluation were to assess if the dashboard improved operational efficiency, reduced no-show appointments, optimized resource allocation, and decreased patient waiting times.

Methods

Study Design

This project was conceived as an innovative data-driven intervention using our data intelligence platform seamlessly integrated with the EHR (Multimedia Appendix 1). Using real-time PHC data, we created a dashboard with an AI model to guide the operational workflow and to reduce clinic wait times and no-show appointments.

This was a before-and-after study that evaluated the impact of a real-time dashboard with an AI-powered no-show prediction model at EHS PHCs 3 months before and after implementation. By leveraging this dashboard, clinic coordinators could proactively manage PHC traffic flow, enhancing operational efficiency and reducing patient waiting times. To capture key metrics accurately, several data definitions were established to track patient flow within the PHC setting. Clinic waiting time was defined as the intervals between various points in a patient's journey, which were all recorded as time stamps in the EHR system. Specifically, this included:

- Nurse waiting time: time from registration to the nurse station.
- Physician waiting time: time from being ready to be seen by the physician to the actual consultation.
- Check-in to checkout time: total time from check-in to final checkout by the physician.
- Appointment no-show: defined as a patient missing a scheduled appointment without prior notice to the PHC.

For comprehensive analysis, all visit and attendance data were sourced from the EHR and the human resources attendance system, both of which were updated in real time to ensure data accuracy and relevance.

Research Road Map Intervention and Dashboard Implementation

Our research followed a structured path from data acquisition to intervention and evaluation. We first integrated patient and clinic staff data from the EHR into a real-time dashboard that incorporated an AI model trained on historical EHR data to predict no-show appointments. This model, along with other key PHC metrics—such as footfall, resource availability, and wait times—were displayed on the dashboard to support clinic coordinators in managing appointments proactively. The AI

XSL•FO
RenderX

model projected daily no-show risk, allowing coordinators to reach out to high-risk patients for confirmation or rescheduling. Additionally, real-time resource constraints were visible on the dashboard, enabling operational adjustments as needed. All users received training on the new dashboard and workflow, and continuous monitoring and evaluation processes were established to assess ongoing performance. Finally, we analyzed the results to measure the dashboard's impact on clinic operations (Multimedia Appendices 2 and 3).

AI Model

Our AI model was developed to predict no-show appointments using historical EHR data from EHS PHCs, incorporating 16 distinct features: 4 demographic, 3 on patient history, and 9 appointment-specific factors (Multimedia Appendix 4). We used a random forest classification technique to assess the model's effectiveness in handling high-dimensional data, achieving an accuracy rate of 86%. The model categorized appointments into a high, medium, or low risk of a no-show, based on probability thresholds of ≥90%, 89% - 80%, and 79% - 70%, respectively. To validate the AI model, we used bootstrapping to ensure robustness and to reduce overfitting.

Statistical Analysis and Experimental Validation

Data were analyzed using SPSS version 17 (IBM Corp) and SAS version 03.05 software (SAS Institute). Descriptive statistics were computed for all variables. Independent 2-tailed sample *t* tests were conducted to compare mean waiting times before and after the intervention. To analyze changes in appointment no-show rates, *z* tests were performed. Logistic regression was used to estimate odds ratios (OR) and determine the likelihood of a no-show after implementation. Statistical significance was set at an α level of .05, with 95% CI computed for all estimates. The statistical validation involved performing hypothesis tests to determine the significance of changes in waiting times and no-show rates.

Outcome Measures

The outcome measures were the mean difference in waiting times, which was calculated using the waiting times before and after implementation and the percent reduction in no-shows, which was calculated using the no-show rates before and after implementation.

Ethical Considerations

Ethical approval was obtained from the Dubai Research Ethics Committee at the Ministry of Health and Prevention, United Arab Emirates. (Approval Reference No. MOHAP/DXB-REC/S.O.O/No.136/2023). The study involved deidentified, retrospective data negating the need for individual patient consent. Patient compensation was not applicable for this study. Bias was minimized by using deidentified data and implementing objective data quality control measures.

Results

We implemented our project across all PHCs in the EHS in October 2022. To compare the outcomes of this tool, we analyzed the data 3 months before (July to September 2022) and 3 months after (November 2022, to January 2023) implementation, and the outcomes were grouped and labeled as before or after, respectively. Descriptive analytics revealed a total of 67,429 booked visits before implementation compared to 67,964 visits after implementation with a similar sex and nationality distribution between both periods (Multimedia Appendix 5). Our analysis showed that most of the visits were through walk-ins instead of booked appointments. There was a slight decrease in walk-in appointments between the two periods, but the percentage of these visits remained relatively stable (80.9% to 80.7%). There was a notable change in the number of no-show appointments, drastically reducing from 20.82% to 10.25% (Table 1).

Table . Visit frequencies and dynamics before and after program implementation in primary health care.

	Before implementation, n/N (%)	After implementation, n/N (%)
Walk-ins	285,700/353,129 (80.91)	284,015/351,979 (80.69)
Appointments	67,429/353,129 (19.09)	67,964/351,979 (19.31)
No-shows	14,038/67,429 (20.82)	6966/67,964 (10.25)

Using a *z* test, our analysis showed that the implementation of the no-show AI model led to a significant 50.7% reduction in no-show rates across the PHC department (*P*<.001). Binary logistic regression further assessed the dashboard's impact on reducing no-shows, yielding an OR of 0.43 (95% CI 0.42 - 0.45; *P*<.001). This OR being less than 1 indicated that the intervention effectively decreased no-show occurrences, reducing the likelihood of no-shows by 57% after implementation of the AI model (Table 2).

Table . Mean difference and impact analysis of appointment no-show rates before and after program implementation in primary health care centers.

Measurements	Values	*P* value
Difference in no-show rate after implementation, n/N (%)	50.7	<.001
Likelihood of no-shows after implementation, odds ratio (95% CI)	0.43 (0.42-0.45)	<.001

Table 3 reflects the different waiting times for the nurse assessment, physician assessment, and overall consultation time (Multimedia Appendix 6). We observed reduced average waiting times at all stations across the PHC journey. We found a

XSL•FO
RenderX

decrease in patient wait times by an average of 5.7 minutes (from 54 to 49 minutes; *P*<.001). Based on the average time saved per visit (5.7 minutes), a total of 387,394.8 minutes were saved in the 3 months after implementation. This is equivalent to approximately 6,456.6 hours saved during this period. The reduction in waiting time from check-in to checkout has been consistent since implementing this solution.

Table . Comparison of patient waiting times at different stations during patient visits before and after program implementation at primary health care centers.

Waiting times	Before implementation (min), mean (SE)	After implementation (min), mean (SE)	Difference following implementation (min)	*P* value
Nurse waiting time	8.4 (0.02)	7.2 (0.02)	1.2	<.001
Physician waiting time	16.3 (0.03)	15.0 (0.03)	1.3	<.001
Physician consultation to checkout time	29.3 (0.06)	26.2 (0.06)	3.1	<.001
Check-in to checkout time	54.2 (0.08)	48.5 (0.07)	5.7	<.001

Discussion

Principal Results and Comparison With Prior Work

The implementation of an AI-powered no-show prediction model, integrated with a real-time operational dashboard at PHC centers within the EHS, has demonstrated significant improvements in operational efficiency and patient care. Our main findings revealed a significant 50.7% reduction in no-show appointments (*P*<.001), and a 5.7-minute reduction in overall patient waiting times, highlighting the direct impact of our intervention on streamlining health care delivery. These results underlined the value of leveraging predictive analytics and real-time data in optimizing resource allocation, reducing manual processes, and enhancing the patient experience. Unlike other studies, where the emphasis has been to explore the factors and behaviors contributing to patient no-shows using machine learning and other techniques [24], we successfully operationalized our machine learning model and connected it to a closed loop workflow which helped reduce absenteeism.

A key achievement of this intervention was the substantial reduction in no-show appointments. By proactively contacting high-risk patients identified by the AI model, clinic administrators were able to significantly reduce missed appointments. This was a notable success, as it not only improved the efficiency of the appointments but also allowed walk-in patients to fill the available slots, further maximizing clinic capacity. Additionally, the use of real-time dashboards provided clinic managers with immediate insights into patient volumes, waiting times, and staff availability, enabling swift adjustments in resource allocation during peak hours. These operational efficiencies, particularly the reduction in waiting times at multiple stages of the patient journey (nurse assessment, physician consultation, and overall clinic visit), underscore the importance of real-time analytics in improving the flow of care. Our robust analysis reassured that the findings of this before and after implementation were significant and could be attributed to our intervention.

The impact of these factors has been previously studied in isolation. There are successful examples of AI models to predict no-show appointments, which have helped reduce the burden of these instances [25-27]. Our observations were in alignment with these studies. Like other studies, we have seen that connecting with patients before their scheduled appointment reduces their chances of missing it [28]. Several strategies, like text messages, patient portal reminders, and phone calls, have been examined, and it has been found that calling patients before their scheduled visit is effective for reducing no-shows [29].

Operationally, real-time dashboards and decision support tools for clinic managers are helpful for identifying operational bottlenecks. Numerous studies have examined the impact of real-time dashboards on reducing patient waiting times in health care settings. These studies suggest that real-time dashboards are crucial for improving patient flow and reducing waiting times across various health care settings, emphasizing their potential as valuable tools in optimizing health care delivery [30-33].

As recommended by other studies, using data to calculate and predict interday scheduling precisely is undeniably beneficial for large health care organizations [30,34]. Trend analysis helps estimate the expected number of unbooked versus booked visits and helps allocate resources accordingly. Many PHC centers reserve a proportion of their scheduling slots for walk-in visits. Since the EHS has many walk-in visits, we should follow a similar strategy and use predictive analysis to help improve our resource allocation. We must carefully design this strategy to reduce our patient waiting times. In the future, we could further enhance the outcomes achieved in this study. By scaling up the AI-model to other services and departments, we could increase the benefits of this program for the EHS. The flexibility and the scalability of our real-time data intelligence platform makes it a valuable tool to resolve operational challenges in our health care system.

Strengths

This study presents a novel approach by integrating an AI-driven predictive model with a real-time dashboard to manage no-show appointments and reduce waiting times in a PHC setting. Unlike previous studies, which primarily focus on predictive modeling without an operational framework, our approach goes a step further by embedding predictive insights into an actionable, real-time dashboard that enables proactive intervention. By operationalizing predictive analytics within a large-scale health care network, this study contributes a scalable and practical

XSL•FO
RenderX

solution to the current state of digital health interventions aimed at improving health care efficiency and patient outcomes.

The major strengths of our study include big data analysis, the use of case mix, the contextualization to the local population, crossfunctional analytics for the clinician-to-patient ratio, real-time analysis for immediate identification of prolonged waiting times, and AI-based modeling for no-show appointment prediction that is coupled with patient contact information. Our holistic approach to use an AI model and real-time insights for a guided operational management workflow is unique in its nature and it can be used by organizations with similar resources.

Limitations

Certain limitations of this study must be acknowledged. First, the mitigation workflow is outside of EHRs, so we cannot quantify the efforts accurately. It is an established fact that a reduction in no-show appointments result in improved financial outcomes [35,36], and we acknowledge that a detailed analysis of the financial implications of reducing no-shows was beyond the scope of this study. Future research could address this aspect to provide a more comprehensive understanding of the intervention's economic impact.

Conclusion

In conclusion, this project represents the first instance of the EHS using a real-time, AI-driven analytics platform to improve operational efficiency, reduce no-show appointments, and enhance patient experience. The success of our solution in reducing no-shows and waiting times highlights the importance of integrating advanced analytics into daily health care operations. By providing clinic administrators with real-time actionable data, this platform enables better decision-making, optimized resource use, and improved patient flow. The scalability of this application across other health care settings further emphasizes its potential to drive widespread improvements in health care delivery, making it a valuable tool for health care organizations looking to modernize their operations through data-driven solutions. As we continue to refine this system, we expect further reductions in waiting times and improvements in patient satisfaction.

Acknowledgments

We would like to acknowledge Benjamin Vinodson. The study was fully funded by the Emirates Health Services, with no external sources involved.

Data Availability

The data sets generated during and/or analyzed during this study are available from the corresponding author on reasonable request

Authors' Contributions

NMI was the principal investigator of this project. All authors contributed to the design, analysis, and manuscript creation according to the sequence of authorship assigned.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Data architecture diagram representing the layers of data transition from source of information to the dashboard.
[PPTX File, 718 KB - formative_v9i1e64936_app1.pptx]

Multimedia Appendix 2
Program workflow for using the real-time artificial intelligence–powered dashboard to manage no-shows and reduce waiting times in primary health care.
[PPTX File, 45 KB - formative_v9i1e64936_app2.pptx]

Multimedia Appendix 3
A screenshot of the data dashboard showing real-time analytics for primary health care visits in a day.
[PPTX File, 1025 KB - formative_v9i1e64936_app3.pptx]

Multimedia Appendix 4
Major predictors contributing to the no-show prediction artificial intelligence model from primary health care visit data.
[PPTX File, 139 KB - formative_v9i1e64936_app4.pptx]

Multimedia Appendix 5

XSL•FO
RenderX

Frequency distribution of booked visits before and after the artificial intelligence program implementation in primary health care centers (July 2022, to January 2023).
[PDF File, 84 KB - formative_v9i1e64936_app5.pdf]

Multimedia Appendix 6
Reduction in patient waiting times at different stages of care following artificial intelligence program implementation in primary health care centers.
[PPTX File, 48 KB - formative_v9i1e64936_app6.pptx]

References

1. Primary health care. World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/primary-health-care [accessed 2024-10-09]
2. Leslie LK, Mehus CJ, Hawkins JD, et al. Primary health care: potential home for family-focused preventive interventions. Am J Prev Med 2016 Oct;51(4 Suppl 2):S106-S118. [doi: 10.1016/j.amepre.2016.05.014] [Medline: 27498167]
3. Guzek LM, Gentry SD, Golomb MR. The estimated cost of “no-shows” in an academic pediatric neurology clinic. Pediatr Neurol 2015 Feb;52(2):198-201. [doi: 10.1016/j.pediatrneurol.2014.10.020] [Medline: 25497123]
4. Moore CG, Wilson-Witherspoon P, Probst JC. Time and money: effects of no-shows at a family practice residency clinic. Fam Med 2001;33(7):522-527. [Medline: 11456244]
5. Harris SL, Samorani M. On selecting a probabilistic classifier for appointment no-show prediction. Decis Support Syst 2021 Mar;142:113472. [doi: 10.1016/j.dss.2020.113472]
6. Lin Q, Betancourt B, Goldstein BA, Steorts RC. Prediction of appointment no-shows using electronic health records. J Appl Stat 2020;47(7):1220-1234. [doi: 10.1080/02664763.2019.1672631] [Medline: 35707022]
7. Denney J, Coyne S, Rafiqi S. Machine learning predictions of no-show appointments in a primary care setting. SMU Data Sci Rev 2019;2(1) [FREE Full text]
8. Harvey HB, Liu C, Ai J, et al. Predicting no-shows in radiology using regression modeling of data available in the electronic medical record. J Am Coll Radiol 2017 Oct;14(10):1303-1309. [doi: 10.1016/j.jacr.2017.05.007] [Medline: 28673777]
9. Yang Y, Madanian S, Parry D. Enhancing health equity by predicting missed appointments in health care: machine learning study. JMIR Med Inform 2024 Jan 12;12:e48273. [doi: 10.2196/48273] [Medline: 38214974]
10. Dansky KH, Miles J. Patient satisfaction with ambulatory healthcare services: waiting time and filling time. Hosp Health Serv Adm 1997;42(2):165-177. [Medline: 10167452]
11. Huang XM. Patient attitude towards waiting in an outpatient clinic and its applications. Health Serv Manage Res 1994 Feb;7(1):2-8. [doi: 10.1177/095148489400700101] [Medline: 10133292]
12. Leddy KM, Kaldenberg DO, Becker BW. Timeliness in ambulatory care treatment. An examination of patient satisfaction and wait times in medical practices and outpatient test and treatment facilities. J Ambul Care Manage 2003;26(2):138-149. [doi: 10.1097/00004479-200304000-00006] [Medline: 12698928]
13. Michael M, Schaffer SD, Egan PL, Little BB, Pritchard PS. Improving wait times and patient satisfaction in primary care. J Healthc Qual 2013;35(2):50-59. [doi: 10.1111/jhq.12004] [Medline: 23480405]
14. Vinitsky A, David B, Michalik L, et al. Reducing wait time in a high-volume pediatric neuro-oncology clinic by optimizing process flow: a quality improvement project. Pediatr Qual Saf 2022;7(3):e557. [doi: 10.1097/pq9.0000000000000557] [Medline: 35720861]
15. Santibáñez P, Chow VS, French J, Puterman ML, Tyldesley S. Reducing patient wait times and improving resource utilization at British Columbia Cancer Agency’s ambulatory care unit through simulation. Health Care Manag Sci 2009 Dec;12(4):392-407. [doi: 10.1007/s10729-009-9103-1] [Medline: 20058528]
16. Le V, Wagar EA, Phipps RA, Del Guidice RE, Le H, Middleton LP. Improving patient experience of wait times and courtesy through electronic sign-in and notification in the phlebotomy clinic. Arch Pathol Lab Med 2020 Jun;144(6):769-775. [doi: 10.5858/arpa.2019-0139-OA] [Medline: 31755777]
17. Almomani I, AlSarheed A. Enhancing outpatient clinics management software by reducing patients’ waiting time. J Infect Public Health 2016;9(6):734-743. [doi: 10.1016/j.jiph.2016.09.005] [Medline: 27663518]
18. Saeed S, Somani N, Sharif F, Kazi AM. Evaluating the effectiveness of text messaging and phone call reminders to minimize no show at pediatric outpatient clinics in Pakistan: protocol for a mixed-methods study. JMIR Res Protoc 2018 Apr 10;7(4):e91. [doi: 10.2196/resprot.9294] [Medline: 29636321]
19. Horvath M, Levy J, L’Engle P, Carlson B, Ahmad A, Ferranti J. Impact of health portal enrollment with email reminders on adherence to clinic appointments: a pilot study. J Med Internet Res 2011 May 26;13(2):e41. [doi: 10.2196/jmir.1702] [Medline: 21616784]
20. Milosevic Z, Chen W, Berry A, Rabhi F. Real-time analytics. In: Big Data 2016:39-61. [doi: 10.1016/B978-0-12-805394-2.00002-7]
21. Wills MJ. Decisions through data: analytics in healthcare. J Healthc Manag 2014;59(4):254-262. [doi: 10.1097/00115514-201407000-00005] [Medline: 25154123]

XSL•FO
RenderX

22. Kaur J, Kulwinder Singh Mann D. AI based healthcare platform for real time, predictive and prescriptive analytics using reactive programming. J Phys Conf Ser 2018 Jan 3;933(1):012010. [doi: 10.1088/1742-6596/933/1/012010]
23. Emirates Health Services - UAE. URL: https://www.ehs.gov.ae/en/home [accessed 2024-12-20]
24. Salazar LHA, Parreira WD, Fernandes ADR, Leithardt VRQ. No-show in medical appointments with machine learning techniques: a systematic literature review. Inf 2022;13(11):507. [doi: 10.3390/info13110507]
25. AlMuhaideb S, Alswailem O, Alsubaie N, Ferwana I, Alnajem A. Prediction of hospital no-show appointments through artificial intelligence algorithms. Ann Saudi Med 2019;39(6):373-381. [doi: 10.5144/0256-4947.2019.373] [Medline: 31804138]
26. Goffman RM, Harris SL, May JH, et al. Modeling patient no-show history and predicting future outpatient appointment behavior in the veterans health administration. Mil Med 2017 May;182(5):e1708-e1714. [doi: 10.7205/MILMED-D-16-00345] [Medline: 29087915]
27. Dunstan J, Villena F, Hoyos JP, et al. Predicting no-show appointments in a pediatric hospital in Chile using machine learning. Health Care Manag Sci 2023 Jun;26(2):313-329. [doi: 10.1007/s10729-022-09626-z] [Medline: 36707485]
28. Liu C, Harvey HB, Jaworsky C, Shore MT, Guerrier CE, Pianykh O. Text message reminders reduce outpatient radiology no-shows but do not improve arrival punctuality. J Am Coll Radiol 2017 Aug;14(8):1049-1054. [doi: 10.1016/j.jacr.2017.04.016] [Medline: 28583321]
29. Childers RE, Laird A, Newman L, Keyashian K. The role of a nurse telephone call to prevent no-shows in endoscopy. Gastrointest Endosc 2016 Dec;84(6):1010-1017. [doi: 10.1016/j.gie.2016.05.052] [Medline: 27327847]
30. Cayirli T, Yang KK, Quek SA. A universal appointment rule in the presence of no‐shows and walk‐ins. Prod Oper Manag 2012 Jul;21(4):682-697. [doi: 10.1111/j.1937-5956.2011.01297.x]
31. Franklin A, Gantela S, Shifarraw S, et al. Dashboard visualizations: supporting real-time throughput decision-making. J Biomed Inform 2017 Jul;71:211-221. [doi: 10.1016/j.jbi.2017.05.024] [Medline: 28579532]
32. Marmor YN, Wasserkrug S, Zeltyn S, et al. Toward simulation-based real-time decision-support systems for emergency departments. Presented at: The 2009 Winter Simulation Conference (WSC); Dec 13-16, 2009; Austin, TX p. 2042-2053. [doi: 10.1109/WSC.2009.5429416]
33. Alaeddini A, Murty KG. DSS (decision support system) for allocating appointment times to calling patients at a medical facility. In: Murty KG, editor. Case Studies in Operations Research: Springer; 2015:83-109. [doi: 10.1007/978-1-4939-1007-6_5]
34. Wiesche L, Schacht M, Werners B. Strategies for interday appointment scheduling in primary care. Health Care Manag Sci 2017 Sep;20(3):403-418. [doi: 10.1007/s10729-016-9361-7] [Medline: 27000079]
35. Berg BP, Murr M, Chermak D, et al. Estimating the cost of no-shows and evaluating the effects of mitigation strategies. Med Decis Making 2013 Nov;33(8):976-985. [doi: 10.1177/0272989X13478194] [Medline: 23515215]
36. Kheirkhah P, Feng Q, Travis LM, Tavakoli-Tabasi S, Sharafkhaneh A. Prevalence, predictors and economic consequences of no-shows. BMC Health Serv Res 2016 Jan 14;16:13. [doi: 10.1186/s12913-015-1243-z] [Medline: 26769153]

Abbreviations

AI: artificial intelligence
EHR: electronic health record
EHS: Emirates Health Services
OR: odds ratio
PHC: primary health care

Edited by A Mavragani; submitted 12.08.24; peer-reviewed by A Ala, F Yang; revised version received 14.11.24; accepted 14.11.24; published 06.01.25.

Please cite as:
AlSerkal YM, Ibrahim NM, Alsereidi AS, Ibrahim M, Kurakula S, Naqvi SA, Khan Y, Oottumadathil NP
Real-Time Analytics and AI for Managing No-Show Appointments in Primary Health Care in the United Arab Emirates: Before-and-After Study
JMIR Form Res 2025;9:e64936
URL: https://formative.jmir.org/2025/1/e64936
doi:10.2196/64936

© Yousif Mohamed AlSerkal, Naseem Mohamed Ibrahim, Aisha Suhail Alsereidi , Mubaraka Ibrahim, Sudheer Kurakula, Sadaf Ahsan Naqvi, Yasir Khan, Neema Preman Oottumadathil. Originally published in JMIR Formative Research (https://formative.jmir.org), 6.1.2025. This is an open-access article distributed under the terms of the Creative Commons

XSL•FO
RenderX

Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Effects of Smart Goggles Used at Bedtime on Objectively Measured Sleep and Self-Reported Anxiety, Stress, and Relaxation: Pre-Post Pilot Study

Sharon Danoff-Burg, PhD; Elie Gottlieb, PhD; Morgan A Weaver, BA; Kiara C Carmon, BA; Duvia Lara Ledesma, MA; Holly M Rus, PhD

Scientific Research, SleepScore Labs, 2175 Salk Avenue, Suite 150, Carlsbad, CA, United States

Corresponding Author:
Sharon Danoff-Burg, PhD
Scientific Research, SleepScore Labs, 2175 Salk Avenue, Suite 150, Carlsbad, CA, United States

Abstract

Background: Insufficient sleep is a problem affecting millions. Poor sleep can trigger or worsen anxiety; conversely, anxiety can lead to or exacerbate poor sleep. Advances in innovative consumer products designed to promote relaxation and support healthy sleep are emerging, and their effectiveness can be evaluated accurately using sleep measurement technologies in the home environment.

Objective: This pilot study examined the effects of smart goggles used at bedtime to deliver gentle, slow vibration to the eyes and temples. The study hypothesized that objective sleep, perceived sleep, self-reported stress, anxiety, relaxation, and sleepiness would improve after using the smart goggles.

Methods: A within-participants, pre-post study design was implemented. Healthy adults with subclinical threshold sleep problems (N=20) tracked their sleep nightly using a polysomnography-validated noncontact biomotion device and completed daily questionnaires over two phases: a 3-week baseline period and a 3-week intervention period. During the baseline period, participants followed their usual sleep routines at home. During the intervention period, participants used Therabody SmartGoggles in “Sleep” mode at bedtime. This mode, designed for relaxation, delivers a gentle eye and temple massage through the inflation of internal compartments to create a kneading sensation combined with vibrating motors. Each night, the participants completed questionnaires assessing relaxation, stress, anxiety, and sleepiness immediately before and after using the goggles. Daily morning questionnaires assessed perceived sleep, complementing the objective sleep data measured every night.

Results: Multilevel regression analysis of 676 nights of objective sleep parameters showed improvements during nights when the goggles were used compared to the baseline period. Key findings include sleep duration (increased by 12 minutes, P=.01); duration of deep sleep (increased by 6 minutes, P=.002); proportion of deep sleep (7% relative increase, P=.02); BodyScore, an age- and gender-normalized measure of deep sleep (4% increase, P=.002); number of nighttime awakenings (7% decrease, P=.02); total time awake after sleep onset (reduced by 6 minutes, P=.047); and SleepScore, a measure of overall sleep quality (3% increase, P=.02). Questionnaire responses showed that compared to baseline, participants felt they had better sleep quality (P<.001) and woke feeling more well-rested (P<.001). Additionally, participants reported feeling sleepier, less stressed, less anxious, and more relaxed (all P values <.05) immediately after using the goggles each night, compared to immediately before use. A standardized inventory administered before and after the 3-week intervention period indicated reduced anxiety (P=.03), confirming the nightly analysis.

Conclusions: The use of smart goggles at bedtime significantly improved objectively measured sleep metrics and perceived sleep quality. Further, participants reported increased feelings of relaxation along with reduced stress and anxiety. Future research expanding on this pilot study is warranted to confirm and expand on the preliminary evidence presented in this brief report.

(JMIR Form Res 2025;9:e58461) doi:10.2196/58461

KEYWORDS

relaxation; stress; anxiety; sleep; health technology; intervention

Introduction

Insufficient sleep affects approximately one-third of the population [1] and is associated with adverse health outcomes and impaired performance [2]. Poor sleep can trigger or worsen anxiety, while conversely, anxiety can lead to or exacerbate poor sleep [3]. Technological advances in unobtrusive sleep measurement enable intervention studies to be conducted in the comfort of research participants’ own bedrooms, providing ecologically valid results while capturing accurate objective

XSL•FO
RenderX

data [4,5]. Concomitantly, the development of innovative consumer products designed to promote relaxation and support healthy sleep is emerging. Their effectiveness can be evaluated in field studies using ambulatory measurement technologies [6].

The aim of this research was to examine the effects of smart goggles used at bedtime to deliver gentle vibration to the eyes and temples, on sleep as well as on perceived stress, anxiety, and relaxation. Although a variety of evidence-based relaxation techniques [7] and sleep-enhancing products [8] already exist, devices such as smart goggles may appeal to individuals who wish to use such technological tools as an option within their repertoire of strategies for winding down at bedtime. Previous research suggests that vibration can increase relaxation [9], induce drowsiness [10], and may be a useful nonpharmacological intervention for poor sleep [11-13]. For example, a preliminary study documented the use of vibration in improving objectively measured sleep outcomes in people with mild to moderate symptoms of insomnia [11]. In this pilot study on adults with subclinical threshold sleep problems, we hypothesized that objective and perceived sleep outcomes as well as self-reported stress, anxiety, relaxation, and sleepiness would improve after using smart goggles delivering slow vibrations at bedtime.

Methods

Participants

Invitations to complete an eligibility questionnaire for a study testing smart goggles were emailed to registrants in a large database of individuals interested in sleep research and using SleepScore technology. Eligible respondents were invited to enroll based on the following selection criteria: difficulty falling or staying asleep, no history of sleep disorders, absence of other medical issues affecting sleep, substance use that could affect sleep, and no lifestyle issues such as shift work that might influence their sleep. The study included adults (N=20) with subclinical threshold sleep problems who were willing to track their sleep and use the smart goggles as instructed.

Design and Procedures

A within-participants, pre-post study design was implemented. The participants were aware that the intervention had the potential to affect their sleep. Following a 3-week baseline period during which participants measured their sleep data at home without any intervention, they used the Therabody SmartGoggles (Therabody Inc, Los Angeles, CA) at bedtime for 15 minutes (within 30 minutes of their intended sleep time) over a 3-week intervention period. The participants were instructed to use “Sleep” mode, designed for relaxation and inducing sleepiness. This mode delivers slow and gentle massage to the eye and temple areas by the inflation of internal compartments to create a kneading sensation and vibrating motors. Additionally, two other modes (SmartRelax and Focus) are available; however, since these provide different experiences, they were not used in this study.

During the entire 6-week study, nightly objective sleep measurements were collected and participants completed online questionnaires each morning and evening. Data collection was synchronized across all participants to account for weekday or weekend variation.

Measurement

Objective sleep data were collected with SleepScore Max (Consumer Sleep Solutions LLC, Carlsbad, CA), a noncontact monitoring device that uses respiratory signal and motion sensing to detect sleep. The device is placed next to the bed and controlled using a companion app. It uses ultra-low power radiofrequency waves to monitor body movement and respiration patterns when in bed; the measurement is unaffected by bedding or nightwear. If a partner is present, only the sleep of the individual closest to the device is measured. The device captures high-resolution magnitude and duration data of gross movements, micromovements, and full breathing cycles, which are transformed into 30-second epoch sleep stage data (wake, light, deep, rapid eye movement [REM]) using proprietary algorithms. Studies have shown good agreement with gold-standard polysomnography [14,15], exceeding the accuracy typically reported for actigraphy-based devices [16].

Using the 30-second epoch data, standard sleep metrics were calculated. In addition, 3 SleepScore technology proprietary sleep metrics reflecting sleep quality, all ranging from 0 to 100 and normalized for age and gender using reference values from the meta-analysis of quantitative sleep parameters by Ohayon and colleagues [17], were calculated:

- *SleepScore* is an overall sleep quality metric that includes objectively measured total sleep time, sleep onset latency, and sleep stage durations.
- *BodyScore* reflects the age- and gender-normalized amount of deep (non–rapid eye movement stage 3 [NREM-3]) sleep.
- *MindScore* reflects the age- and gender-normalized amount of REM sleep.

Self-reported data were collected daily, across the entire study via 100-point visual analog scales. Morning assessments measured perceived sleep quality and feeling well-rested upon waking. At night, the 100-point visual analog scales assessed relaxation, stress, anxiety, and sleepiness before and after goggle use. The construct validity and discriminate sensitivity of visual analog scales to assess perceived stress and related constructs have been documented [18]. Participants completed these scales immediately before using the goggles at bedtime and after using the goggles for 15 minutes, and then went to sleep. A 6-item version of the state scale of the Spielberger State-Trait Anxiety Inventory (STAI) [19] was administered once before and once after the intervention period.

Statistical Analyses

Nightly objective sleep data and daily self-reported data were analyzed using multilevel regression with random intercept, accounting for nights nested within participants, comparing nights during the baseline period to nights during the intervention period for each outcome. The regression model was $Sleepmeasure_{ij} = \beta_0 + \beta_1 * TestPeriod_{ij} + u_{0j} + e_{ij}$; TestPeriod, coded as 0 for observations during baseline and 1 for nights during the intervention period. Similarly, analysis of the nightly self-reported data used the same model to compare pre- and

XSL•FO
RenderX

post-goggle use. A 2-tailed paired-samples *t* test was used to analyze changes in the 6-item STAI scores.

Discrepancies in sample sizes (N=20 for objective and self-reported sleep, n=17 for 6-item STAI) were due to incomplete data sources. Participants tracked their sleep at home and, at times, were not fully compliant with the use of measurement tools or the completion of online surveys. This is common in field research collecting longitudinal and daily assessment data. All reported results reflect the largest available sample for each set of analyses.

Ethical Considerations

The study was approved by Sterling Institutional Review Board (ID 11012), and all procedures were conducted in accordance with the ethical standards of the Declaration of Helsinki. All participants provided informed consent using a secure online platform to review, electronically sign, and return a copy of the document to the research coordinator. They were informed that the study was voluntary, and of their right to withdraw at any time. Both objective and self-reported study data were deidentified prior to analysis and accessible only to members of the research team. As compensation, participants kept the smart goggles and sleep measurement device used during the study.

Results

Of the 20 participants, 40% (n=8) were women, and the age range was 26-75 (mean 50.41, SD 13.12) years. Further demographic details are provided in Table 1.

XSL•FO
RenderX

Table . Demographic information of participants at baseline (N=20).

Demographics	Values
Age (years)	
Mean (SD)	50.41 (13.12)
Range	26 - 75
Gender, n (%)	
Men	8 (40)
Women	12 (60)
Other identities	0 (0)
Race/Ethnicity, n (%)	
American Indian or Alaska Native	0 (0)
Asian	1 (5)
Black or African American	0 (0)
Hispanic/Latino	4 (20)
Native Hawaiian or other Pacific Islander	1 (5)
White	14 (70)
Household composition, n (%)	
I live alone	7 (35)
I live with my partner - (currently) no children in the house	3 (15)
I live with my children - (currently) no partner in the house	1 (5)
I live with my partner and child(ren)	6 (30)
I live with a family member(s)	3 (15)
I live with a roommate(s)	0 (0)
Other	0 (0)
Employment, n (%)	
Working full-time	14 (70)
Working part-time	2 (10)
Homemaker	2 (10)
Full-time student	1 (5)
Retired	1 (5)
Education, n (%)	
High school degree or equivalent (eg, GED[a])	1 (5)
Some college, no degree	5 (25)
Associate degree	1 (5)
Bachelor's degree	8 (40)
Master's degree	3 (15)
Doctoral degree	2 (10)
Annual household income (US$), n (%)	
<$25,000	0 (0)
$25,000-$49,999	3 (15)
$50,000-$74,999	4 (20)
$75,000-$99,999	3 (15)
$100,000-$124,999	2 (10)
$125,000-$149,999	2 (10)

XSL•FO
RenderX

$150,000-$174,999	0 (0)
$175,000-$199,999	0 (0)
≥$200,000	5 (25)
Would rather not answer this question	1 (5)

[a]GED: General Educational Development.

Objective Sleep

Nightly objective measurement of sleep (n=676 nights nested within 20 participants) revealed multiple improvements when participants used the goggles at bedtime. Key findings included increased sleep duration (+12 min, *P*=.01); increased deep sleep, reflected both in duration (+6 min, *P*=.002) and proportion of the night (7% relative increase, *P*=.02); enhanced BodyScore (+4%, *P*=.002); fewer nighttime awakenings (−7%, *P*=.02); reduced total wake time at night after sleep onset (−6 min, *P*=.047); and improved SleepScore, indicating overall sleep quality (+3%, *P*=.02). Detailed objective sleep metrics are displayed in Table 2.

Table . Multilevel regression results for objective sleep (n=676 nights nested within 20 participants), comparing the baseline period (sleep at home prior to intervention) to the intervention period (sleep at home using smart goggles at bedtime).

Outcomes	Objective measurement of sleep, observed mean (SD)[a]		Estimated marginal means[b]		
	Baseline period	Intervetion period	Intercept (SE)	β[c]	*P* value
Total sleep time in minutes	331.55 (57.49)	343.48 (62.69)	331.52 (4.87)	12.04	.01
Sleep onset latency in minutes	23.34 (18.33)	23.20 (18.17)	23.28 (1.64)	−0.13	.92
Number of awakenings	5.51 (2.10)	5.14 (2.12)	5.51 (0.17)	−0.38	.02
Time spent awake after sleep onset in minutes	60.50 (31.30)	54.32 (31.38)	60.17 (2.72)	−5.39	.047
Light sleep in minutes	215.36 (49.64)	219.86 (50.57)	215.47 (3.99)	4.54	.26
Deep sleep in minutes	57.52 (19.92)	63.19 (24.27)	57.66 (1.79)	5.58	.002
REM[d] sleep in minutes	58.67 (23.54)	60.44 (25.55)	58.52 (1.93)	1.94	.32
Percentage of time spent awake after sleep onset	16 (7)	14 (7)	15.36 (0.58)	−1.02	.08
Percentage of time in light sleep	54 (8)	55 (8)	54.50 (0.60)	0.02	.97
Percentage of time in deep sleep	15 (6)	16 (7)	15.10 (0.49)	1.14	.02
Percentage of time in REM sleep	15 (6)	15 (6)	14.92 (0.46)	−0.12	.80
SleepScore[e]	69.57 (10.89)	71.65 (10.87)	69.61 (0.87)	2.02	.02
BodyScore[e]	73.14 (10.38)	76.13 (11.15)	73.42 (0.86)	2.64	.002
MindScore[e]	69.30 (13.76)	69.68 (14.48)	69.18 (1.13)	0.55	.62

[a]For the baseline and intervention periods, each mean was calculated by averaging nights across participants, then averaging those participants' averages to a single simple average.

[b]These are the outcomes of separate multilevel regression analyses. Each row shows results from a different single-predictor, single-outcome model.

[c]The beta values are unstandardized and can therefore be interpreted on the same scale as the original data.

[d]REM: rapid eye movement.

[e]These scores range from 0 to 100. SleepScore is an age- and gender-normalized measure of overall sleep quality, BodyScore is an age- and gender-normalized measure of deep sleep, and MindScore is an age- and gender-normalized measure of REM sleep.

Self-Reported Sleep Quality and Well-Restedness

Multilevel regression analyses of daily self-reported sleep data (N=723 nights nested within 20 participants) showed that participants perceived better sleep quality (β=12.37, *P*<.001) and felt more rested in the morning (β=12.13, *P*<.001) when

XSL•FO
RenderX

using the goggles at bedtime compared to baseline, as assessed on a scale from 0 to 100.

Sleepiness, Anxiety, Stress, and Relaxation

Across 334 nights of the intervention period, multilevel regression analyses comparing responses immediately after using the goggles to those reported immediately before showed that participants felt sleepier (β=9.98, $P<.001$), less stressed (β=−10.38, $P<.001$), less anxious (β=−12.87, $P<.001$), and more relaxed (β=11.76, $P<.001$), all rated on a scale from 0 to 100.

At the end of the intervention period, compared to the end of the baseline period, participants' scores on the 6-item STAI showed reduced anxiety (t_{16}=2.31, P=.03), reflecting a 10% decrease and confirming the nightly analyses.

Discussion

Nonpharmacological techniques for promoting relaxation and improving sleep have the potential to help millions of individuals who experience suboptimal sleep [20]. This study evaluated the effectiveness of smart goggles designed to induce relaxation and support healthy sleep. Outcomes were measured using both self-reported data and a polysomnography-validated, noncontact biomotion device. The study population included a nonclinical sample of adults reporting poor sleep in the absence of diagnosed sleep disorders.

Compared to baseline, using the goggles at bedtime led to objective improvements in both sleep quality and duration. Although total sleep time remained less than 6 hours per night on average, objective improvements were seen in several parameters, including:

- sleep duration;
- increased deep sleep, both in duration and as a proportion of the night;
- reduced number of nighttime awakenings;
- decreased time spent awake at night after initially falling asleep; and
- enhanced sleep quality.

Aligning with these objective results, questionnaire data showed that participants perceived improvement in their sleep quality and felt more well-rested in the morning. In addition, immediately after using the goggles, participants felt sleepier, less stressed, less anxious, and more relaxed, compared to their experience immediately before using the goggles. A standardized inventory administered before and after the 3-week intervention period also indicated reduced anxiety, confirming the nightly analysis.

The observed improvements in the objective and self-reported sleep data may be attributed to increased relaxation resulting from the use of the goggles at bedtime. Vibration has been previously shown to be able to induce physiological relaxation [9,10] and support sleep [11-13]. This interpretation is supported by the changes in perceived relaxation, stress, and anxiety. However, objective parameters of physiological relaxation prior to sleep were not assessed, presenting an interesting avenue for future research.

Further studies could explore the intervention through a controlled trial, including comparison to goggles without vibration or to other relaxation techniques. This study assessed changes in parameters from the baseline to the intervention period while using the product at home, resembling how it is used outside of a research setting. The within-participants study design, which included long-term product use following a baseline period without the intervention, provides confidence that significant effects are due to the intervention itself. Although this study design has limitations, it reflects the real-world experience of introducing an intervention into the home environment, while also accounting for night-to-night variations in sleep patterns.

While longitudinal data provides an advantage by enabling the detection of nightly within-person differences, this pilot study is limited by its small sample size. Although we compared participants' improvements in sleep and related outcomes to their own baseline, the small sample size could have introduced a potential bias, and therefore these results should be interpreted with caution. Future research should investigate the effects of this intervention using larger sample sizes. Another limitation is that contactless technology is not ambulatory and therefore does not capture objective measures of daytime activity. However, the noncontact sleep measurement system used in this study offers the advantage of significantly higher overall accuracy, particularly in specificity, compared to traditional research-based actigraphy [16].

To conclude, this study found that using vibration therapy administered via smart goggles before going to sleep was associated with improvements in both objective sleep measures and self-reported outcomes. While these findings must be interpreted with caution, our data suggest that a device delivering gentle vibrations to the eye and temple area may have the potential to promote relaxation, decrease anxiety, and support healthy sleep when used at bedtime by adults with suboptimal sleep. Not only did sleep quality and sleep duration increase relative to baseline, but there was a decrease in the number of awakenings and the duration of time spent awake during the night. Further, improvement was seen in multiple metrics of deep sleep, which is vital for brain health and physical recovery. Future research expanding on this pilot study is warranted to confirm the preliminary evidence presented in this study.

Acknowledgments

Therabody provided the SmartGoggles and funding support to SleepScore Labs to design, conduct, and report the study.

XSL•FO
RenderX

Data Availability

The data analyzed in this study may be available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors are employed by SleepScore Labs.

References

1. Grandner MA. Epidemiology of insufficient sleep and poor sleep quality. In: Grandner MA, editor. Sleep and Health: Sleep Health Acad Press; 2019:11-20. [doi: 10.1016/B978-0-12-815373-4.00002-2]
2. Consensus Conference Panel, Watson NF, Badr MS, et al. Recommended amount of sleep for a healthy adult: a joint consensus statement of the American Academy of Sleep Medicine and Sleep Research Society. J Clin Sleep Med 2015 Jun 15;11(6):591-592. [doi: 10.5664/jcsm.4758] [Medline: 25979105]
3. Chellappa SL, Aeschbach D. Sleep and anxiety: from mechanisms to interventions. Sleep Med Rev 2022 Feb;61:101583. [doi: 10.1016/j.smrv.2021.101583] [Medline: 34979437]
4. Guillodo E, Lemey C, Simonnet M, et al. Clinical applications of mobile health wearable-based sleep monitoring: systematic review. JMIR Mhealth Uhealth 2020 Apr 1;8(4):e10733. [doi: 10.2196/10733] [Medline: 32234707]
5. Ko PRT, Kientz JA, Choe EK, Kay M, Landis CA, Watson NF. Consumer sleep technologies: a review of the landscape. J Clin Sleep Med 2015 Dec 15;11(12):1455-1461. [doi: 10.5664/jcsm.5288] [Medline: 26156958]
6. Watson NF, Lawlor C, Raymann RJEM. Will consumer sleep technologies change the way we practice sleep medicine? J Clin Sleep Med 2019 Jan 15;15(1):159-161. [doi: 10.5664/jcsm.7596] [Medline: 30621844]
7. Bertisch SM, Wells RE, Smith MT, McCarthy EP. Use of relaxation techniques and complementary and alternative medicine by American adults with insomnia symptoms: results from a national survey. J Clin Sleep Med 2012 Dec 15;8(6):681-691. [doi: 10.5664/jcsm.2264] [Medline: 23243402]
8. Maust DT, Solway E, Clark SJ, Kirch M, Singer DC, Malani P. Prescription and nonprescription sleep product use among older adults in the United States. Am J Geriatr Psychiatry 2019 Jan;27(1):32-41. [doi: 10.1016/j.jagp.2018.09.004] [Medline: 30409547]
9. Zhou Y, Murata A, Watanabe J. The calming effect of heartbeat vibration. Presented at: 2020 IEEE HAPTICS Symposium (HAPTICS); Mar 28-31, 2020; Crystal City, VA. [doi: 10.1109/HAPTICS45997.2020.ras.HAP20.157.5a2e1551]
10. Zhang N, Fard M, Bhuiyan MHU, Verhagen D, Azari MF, Robinson SR. The effects of physical vibration on heart rate variability as a measure of drowsiness. Ergonomics 2018 Sep;61(9):1259-1272. [doi: 10.1080/00140139.2018.1482373] [Medline: 29871584]
11. Himes BJ, Blotter JD, Kay DB, et al. The effect of beat frequency vibration on sleep latency and neural complexity: a pilot study. IEEE Trans Neural Syst Rehabil Eng 2021;29:872-883. [doi: 10.1109/TNSRE.2021.3076983] [Medline: 33950842]
12. Choi SH, Yoon H, Jin HW, et al. Effect of closed-loop vibration stimulation on heart rhythm during naps. Sensors (Basel) 2019 Sep 24;19(19):4136. [doi: 10.3390/s19194136] [Medline: 31554268]
13. Choi SH, Kwon HB, Jin HW, et al. Weak closed-loop vibrational stimulation improves the depth of slow-wave sleep and declarative memory consolidation. Sleep 2021 Jun 11;44(6):zsaa285. [doi: 10.1093/sleep/zsaa285] [Medline: 33367712]
14. O'Hare E, Flanagan D, Penzel T, Garcia C, Frohberg D, Heneghan C. A comparison of radio-frequency biomotion sensors and actigraphy versus polysomnography for the assessment of sleep in normal subjects. Sleep Breath 2015 Mar;19(1):91-98. [doi: 10.1007/s11325-014-0967-z] [Medline: 24614968]
15. Schade MM, Bauer CE, Murray BR, et al. Sleep validity of a non-contact bedside movement and respiration-sensing device. J Clin Sleep Med 2019 Jul 15;15(7):1051-1061. [doi: 10.5664/jcsm.7892] [Medline: 31383243]
16. Chinoy ED, Cuellar JA, Huwa KE, et al. Performance of seven consumer sleep-tracking devices compared with polysomnography. Sleep 2021 May 14;44(5):zsaa291. [doi: 10.1093/sleep/zsaa291] [Medline: 33378539]
17. Ohayon MM, Carskadon MA, Guilleminault C, Vitiello MV. Meta-analysis of quantitative sleep parameters from childhood to old age in healthy individuals: developing normative sleep values across the human lifespan. Sleep 2004;27:1255-1273. [doi: 10.1093/sleep.27.7.1255] [Medline: 15586779]
18. Lesage FX, Berjot S, Deschamps F. Clinical stress assessment using a visual analogue scale. Occup Med (Lond) 2012 Dec;62(8):600-605. [doi: 10.1093/occmed/kqs140] [Medline: 22965867]
19. Marteau TM, Bekker H. The development of a six-item short-form of the state scale of the Spielberger State-Trait Anxiety Inventory (STAI). Br J Clin Psychol 1992 Sep;31(3):301-306. [doi: 10.1111/j.2044-8260.1992.tb00997.x] [Medline: 1393159]
20. Maness DL, Khan M. Nonpharmacologic management of chronic insomnia. Am Fam Physician 2015 Dec 15;92(12):1058-1064. [Medline: 26760592]

Abbreviations

NREM-3: non–rapid eye movement stage 3

XSL•FO
RenderX

REM: rapid eye movement
STAI: State-Trait Anxiety Inventory

Edited by A Mavragani; submitted 15.03.24; peer-reviewed by GR Elkins, R Chamorro; revised version received 21.10.24; accepted 29.10.24; published 03.01.25.

Please cite as:
Danoff-Burg S, Gottlieb E, Weaver MA, Carmon KC, Lara Ledesma D, Rus HM
Effects of Smart Goggles Used at Bedtime on Objectively Measured Sleep and Self-Reported Anxiety, Stress, and Relaxation: Pre-Post Pilot Study
JMIR Form Res 2025;9:e58461
URL: https://formative.jmir.org/2025/1/e58461
doi:10.2196/58461

© Sharon Danoff-Burg, Elie Gottlieb, Morgan A Weaver, Kiara C Carmon, Duvia Lara Ledesma, Holly M Rus. Originally published in JMIR Formative Research (https://formative.jmir.org), 3.1.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Use of Social Media on Enhancing Dental Care and Practice Among Dental Professionals: Cross-Sectional Survey Study

Joseph Macadaeg Acosta[1,2], MDE, DMD; Palinee Detsomboonrat[3], MSc, DDS, PhD; Pagaporn Pantuwadee Pisarnturakit[3], MSc, DDS, PhD; Nipaporn Urwannachotima[3], MEd, DDS, PhD

[1]International Graduate Program in Dental Public Health, Department of Community Dentistry, Faculty of Dentistry, Chulalongkorn University, Bangkok, Thailand

[2]School of Dentistry, University of Baguio, Baguio City, Philippines

[3]Faculty of Dentistry, Department of Community Dentistry, Chulalongkorn University, 34 Henri Dunant Road, Pathumwan, Bangkok, Thailand

Corresponding Author:
Nipaporn Urwannachotima, MEd, DDS, PhD
Faculty of Dentistry, Department of Community Dentistry, Chulalongkorn University, 34 Henri Dunant Road, Pathumwan, Bangkok, Thailand

Abstract

Background: As digitalization continues to advance globally, the health care sector, including dental practice, increasingly recognizes social media as a vital tool for health care promotion, patient recruitment, marketing, and communication strategies.

Objective: This study aimed to investigate the use of social media and assess its impact on enhancing dental care and practice among dental professionals in the Philippines.

Methods: A cross-sectional survey was conducted among dental practitioners in the Philippines. The study used a 23-item questionnaire, which included 5 questions on dentists' background and demographic information and 18 questions regarding the use, frequency, and purpose of social media in patient advising and quality of care improvement. Data were analyzed using SPSS software, with frequency distributions and χ^2 tests used to assess the association between social media use and demographic variables and the impact on dental practice.

Results: The 265 dental practitioners in this study were predominantly female (n=204, 77%) and aged between 20 - 30 years (n=145, 54.7%). Most of the participants were general practitioners (n=260, 98.1%) working in a private practice (n=240, 90.6%), with 58.5% (n=155) having 0 - 5 years of clinical experience. Social media use was significantly higher among younger practitioners (20 - 30 years old) compared to older age groups (P<.001), though factors such as sex, dental specialty, and years of clinical practice did not significantly influence use. The majority (n=179, 67.5%) reported using social media in their practice, primarily for oral health promotion and education (n=191, 72.1%), connecting with patients and colleagues (n=165, 62.3%), and marketing (n=150, 56.6%). Facebook (n=179, 67.5%) and YouTube (n=163, 61.5%) were the most frequented platforms for clinical information, with Twitter (subsequently rebranded X) being the least used (n=4, 1.5%). Despite widespread social media engagement, only 8.7% (n=23) trusted the credibility of web-based information, and 63.4% (n=168) perceived a potential impact on the patient-dentist relationship due to patients seeking information on the internet. Social media was also perceived to enhance practice quality, with users reporting significant improvements in patient care (P=.001).

Conclusions: The findings highlight that social media is widely used among younger dental practitioners, primarily for education, communication, and marketing purposes. While social media use is associated with perceived improvements in practice quality and patient care, trust in information on social media remains low, and concerns remain regarding its effect on patient relationships. It is recommended to establish enhanced guidelines and provide reliable web-based resources to help dental practitioners use social media effectively and responsibly.

(JMIR Form Res 2025;9:e66121) doi:10.2196/66121

KEYWORDS

social media; oral health promotion; oral health education; dentists; dental practice; dental professionals; dental practitioners

Introduction

Traditionally, dental practices advertised their services primarily through local yellow pages, such as telephone directories listing local businesses or by placing promotional notices in the windows of dental offices [1]. Word-of-mouth was also a common method, where information about the quality of service was shared from one person to another [2]. However, with the advent of technological advancements, there has been a significant shift toward the use of social media. Social media

XSL•FO
RenderX

platforms offer the potential to reach diverse audiences, including those seeking information on oral health management, thereby promoting oral health [3].

Social media refers to a group of web-based applications built on the foundations of Web 2.0 [4,5]. According to the Merriam-Webster Dictionary, social media is defined as "any form of electronic communication through which users create web-based communities to share information, personal messages, ideas, and other content such as photos and videos" [6]. Popular platforms include social networking sites like Facebook and Twitter (subsequently rebranded X), media-sharing sites like YouTube and Instagram, blogs, and microblogging sites [7]. These platforms are widely used for searching, sharing, and communicating information, including health-related content [8,9]. Facebook fosters connections through its friend-based network, creating a web-based community where users can stay in touch and share content [10]. Twitter's microblogging feature allows users to share ideas and opinions in real time [11,12]. YouTube is a leading video-sharing platform [13-15], while Instagram, a photo and video social networking site, facilitates real-time communication through comments and direct messages and supports professional engagement and knowledge sharing [16-18]. TikTok, known for its short video content, has also gained popularity [19,20].

Social media serves as a dynamic tool for exchanging information and communicating with colleagues, patients, and the public on health-related issues. Public health professionals and organizations use these platforms for a variety of purposes, including health education, telemedicine, social marketing, scientific research, recruitment, career development, and professional networking [9]. Dental practitioners are increasingly leveraging social media to educate and inform the public by sharing detailed content on various dental procedures and educational activities [8,21-25], providing oral hygiene instructions and promoting oral health [26,27], enhancing communication [28-30], and engaging in marketing efforts [1,25,31,32]. Given the widespread use of social media, using these platforms responsibly and effectively in health care is crucial to ensure that they contribute positively to patient health outcomes.

As of April 2024, there were 5.44 billion internet users worldwide, representing 67.1% of the global population. Of this total, 5.07 billion, or 62.6% of the world's population, were social media users [33]. Based on the number of monthly active users, the most popular social networks globally are Facebook, YouTube, Instagram, WhatsApp, and TikTok [34]. In the Philippines, approximately 68.72 million people are social media users, with an average daily use of over 3 hours, the highest across the Asia-Pacific region. Facebook, Instagram, TikTok, and X (formerly known as Twitter) are currently some of the most popular platforms, providing opportunities for connecting with family and friends, sharing content via digital platforms, and expanding the reach of promotional marketing through web-based banners [35]. Social media in the Philippines has gained significant traction for entertainment, communication, marketing, and professional education. In dentistry, practitioners use these platforms to share content related to oral health, modern treatment options, trends in dental materials, and treatment costs.

Although several reports have highlighted the use of social media among dentists and its influence on dental practice [3,8,9,22,30,32,36,37], as well as its impact on dental education among faculty and students [7,21,23,38,39], relatively few studies have examined the use and influence of social media on dental practice in Southeast Asia, particularly in the Philippines where these platforms are widely used. This study aims to investigate the use of social media and assess its impact on enhancing dental care and practice among dental professionals in the Philippines.

Methods

Design and Study Population

This study used a cross-sectional survey design with purposive sampling, targeting all eligible licensed dental professionals, including both general dentists and specialists working in private or public clinical settings and who were involved in patient care, education, and practice management. Participants were also required to be active members of the Philippine Dental Association-Baguio City Chapter for the 2023 - 2024 term. The study population consisted of 265 dental practitioners. Practitioners who were not active members of the chapter were excluded from the study.

The survey questionnaire was adapted from previously published literature related to social media use [22,36,38]. For this study, social media included popular platforms such as Facebook, YouTube, Twitter, Instagram, TikTok, and Viber, which dental practitioners may have used for professional purposes like patient education, practice marketing, peer communication, and access to clinical information. A panel of experts validated the questionnaire items and the questionnaire underwent pretesting to ensure applicability and reliability. Revisions and minor adjustments were made based on the pilot study results and expert feedback. The 23-item questionnaire was divided into two sections. The first section gathered background and demographic information, including age, sex, dental specialty, work sector, and years of clinical practice. The second section comprised 18 questions designed to gather information on social media use; frequency of use; and purpose in dental practice, particularly its role in advising patients and improving the quality of care. This study was conducted within dental practice settings, with participants recruited through web-based methods. Recruitment emphasized voluntary participation, with all respondents providing informed consent before inclusion in the study. The questionnaire was administered via Google Forms, with the survey link disseminated through email, a Facebook page, and the Baguio City Chapter group chat. Respondents were encouraged to provide suggestions, assist in recruitment, and further share the questionnaire with other dental practitioners to reach more participants. They were given one day to complete the form.

Data Analysis

Data were analyzed using SPSS software version 28.0 (IBM Corp). The analysis included frequency distributions that were

XSL•FO
RenderX

expressed in numbers and percentages, as well as χ^2 tests to assess the association between social media use and demographic or social variables and the impact on dental practice. Statistical significance was set at an α level of .05. This study adhered to the reporting guidelines for cross-sectional studies following the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement.

Ethical Considerations

Ethical approval was obtained from the Human Research Ethics Committee of the Faculty of Dentistry, Chulalongkorn University (Study Code: HREC-DCU 2023 - 108). Informed consent was obtained from all participants through a written consent form, which provided details about the study's objectives and potential benefits. Participation in the study was entirely voluntary, with participants having the right to withdraw or discontinue at any time without any penalties or loss of benefits. All information was kept strictly confidential and anonymized to ensure participants could not be identified. No financial rewards or incentives were provided for participation.

Results

The background and demographic details of the 265 study participants are summarized in Table 1. The majority of participants were female (n=204, 77%), with over half (n=145, 54.7%) aged between 20 to 30 years. This was followed by participants aged 31 - 40 years (n=52, 19.6%), and both groups of participants aged 41 - 50 years and over 50 years represented 12.8% (n=34) of participants. Most participants (n=260, 98.1%) were general practitioners, and 90.6% (n=240) worked as private dentists. Regarding clinical experience, 58.5% (n=155) had 0 to 5 years of practice, while 29.8% (n=79) had over 10 years, and 11.7% (n=31) had 6 to 10 years of practice.

Table . Demographic variables and their association with the use of social media among dental practitioners.

Variable	Use of social media, n (%)[a]			Total (n=265), n (%)	*P* value[b]
	Yes	Sometimes	No		
Age (years)					<.001
20 - 30	102 (70.3)	35 (24.1)	8 (5.5)	145 (54.7)	
31 - 40	31 (60)	21 (40)	0 (0)	52 (19.6)	
41 - 50	27 (79)	2 (6)	5 (15)	34 (12.8)	
>50	19 (56)	15 (44)	0 (0)	34 (12.8)	
Sex					.66
Male	40 (66)	19 (31)	2 (3)	61 (23)	
Female	139 (68.1)	54 (26.5)	11 (5.4)	204 (77)	
Dental specialty					.25
General practitioners	177 (68.1)	70 (26.9)	13 (5)	260 (98.1)	
Specialists	2 (40)	3 (60)	0 (0)	5 (1.9)	
Working sector					.30
Government	10 (71)	2 (14)	2 (14)	14 (5.3)	
Private	160 (66.6)	69 (28.8)	11 (4.6)	240 (90.6)	
Both (government and private)	9 (82)	2 (18)	0 (0)	11 (4.2)	
Number of years of clinical practice					.19
0 to 5	109 (70.3)	38 (24.5)	8 (5.2)	155 (58.5)	
6 to 10	24 (77)	7 (23)	0 (0)	31 (11.7)	
>10	46 (58)	28 (35)	5 (6)	79 (29.8)	

[a]Percentages were calculated based on the number of participants corresponding to each variable.

[b]Chi-square test; α=.05.

There was a significant increase in social media use among the younger age group (20 - 30 years) compared to older age groups (31 - 40 years, 41 - 50 years, and >50 years), with a *P* value of <.001. However, sex, dental specialty, work sector, and years of clinical practice did not significantly influence social media use.

Table 2 illustrates the frequency distribution of social media use among dental practitioners. Of the 265 participants, most reported moderate use of social media with 37.4% (n=99) using it 1 - 2 days per week. However, nearly two-thirds (n=164, 61.9%) reported using social media at least 3 days weekly. Facebook (n=179, 67.5%) and YouTube (n=163, 61.5%) were the most commonly visited platforms for obtaining clinical information, followed by Instagram (n=115, 43.4%). Twitter

XSL•FO
RenderX

(n=4, 1.5%) was the least popular platform. Social media was mainly used for oral health education and promotion (n=191, 72.1%); communication with patients, friends, and family (n=165, 62.3%); and marketing and advertising (n=150, 56.6%).

Table . Frequency, preferred platforms, and purpose of social media use.

Variables	Frequency (n=265), n (%)
Social media use	
Never	2 (0.8)
Seldom (average of 1 - 2 d/wk)	99 (37.4)
Occasionally (average of 3 - 5 d/wk)	96 (36.2)
Frequently (average of 6 - 7 d/wk)	68 (25.7)
Social media website used[a]	
Twitter	4 (1.5)
Facebook	179 (67.5)
Instagram	115 (43.4)
YouTube	163 (61.5)
TikTok	38 (14.3)
Viber	23 (8.7)
Others	75 (28.3)
Purpose of social media use[a]	
Oral health promotion and education	191 (72.1)
Connect and communicate with patients, friends, and family	165 (62.3)
Exchange opinions and views regarding cases with colleagues	113 (42.6)
Marketing and advertising	150 (56.6)

[a]Given that study participants could indicate multiple responses, the sum of the percentages is not 100.

The responses of dental practitioners to questions regarding social media use are presented in Figure 1. Of the 265 participants, 67.5% (n=179) indicated that they use social media in their dental practice. Despite this, 83% (n=220) of dentists did not have their own web-based practice forums or websites. However, the majority (n=227, 85.7%) reported using social media to communicate with other dental professionals, and 41.5% (n=110) used it for marketing purposes. Although 41.1% (n=109) of dentists referred to articles or research information from social media for clinical practice, nearly half (n=127, 47.9%) were reluctant to broadcast treatment outcomes on the internet to attract patients, and 41.9% (n=111) were open to providing web-based consultations. Only 8.7% (n=23) of dentists trusted the credibility of information on social media, while 63.4% (n=168) believed that social media can affect the patient-dentist relationship when patients seek information on the internet. Additionally, most participants (n=154, 58.1%) did not allow patients to access their information through a website, but 64.2% (n=170) would recommend a trusted website to patients.

Table 3 explores the relationship between social media use and its impact on dental practice. Participants who used social media in their practice believed that it enhanced the quality of care provided to patients (P=.001) and significantly reported that it improved their dental practice (P=.02).

XSL•FO
RenderX

Figure 1. Percentage of dental practitioners' responses to questions on social media use (n=265).

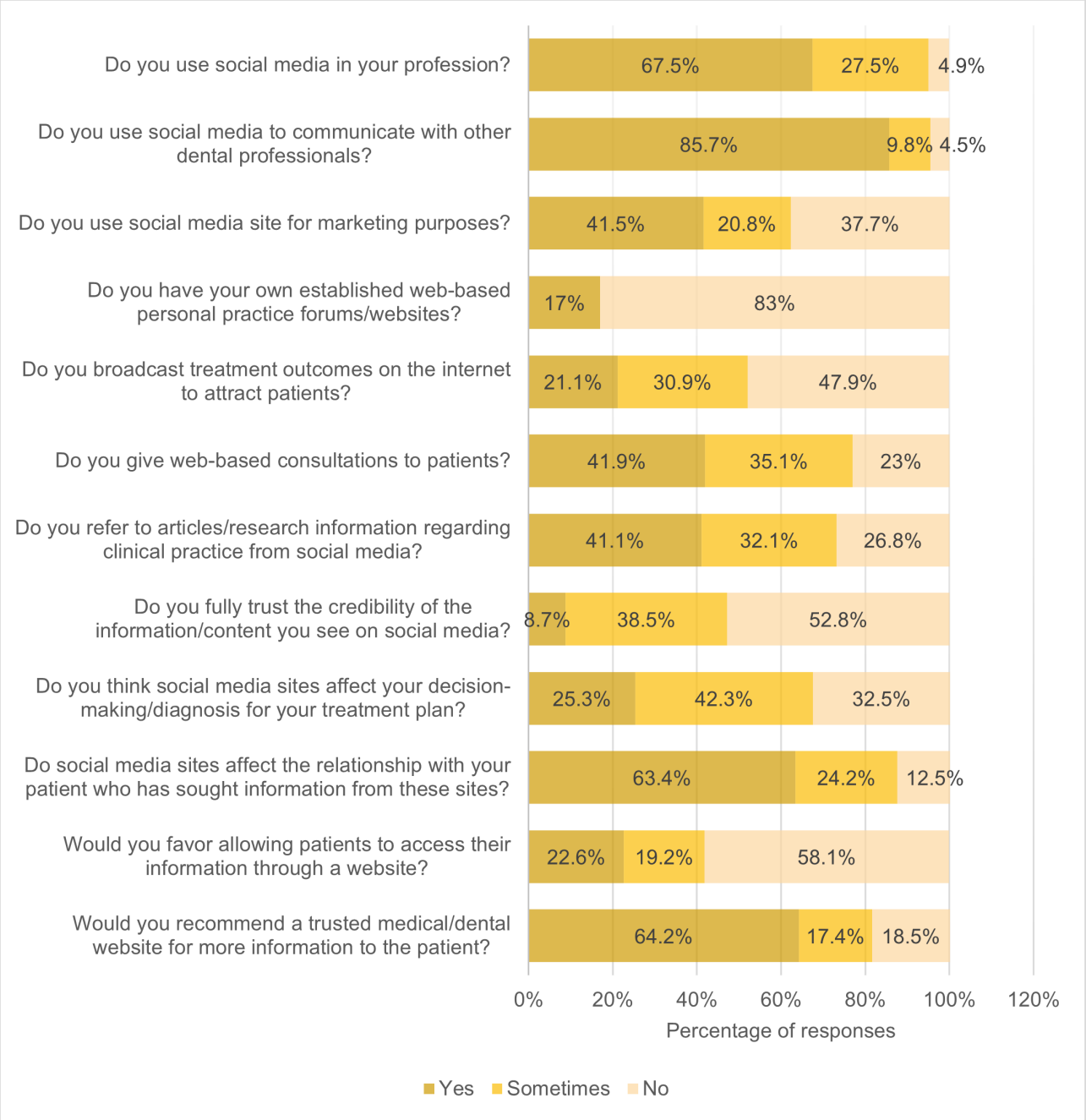

XSL•FO
RenderX

Table . Use of social media and respondents' perceived impact on enhancing patient care quality and dental practice.

Variable	Use of social media, n (%)[a]			Total (n=265), n (%)	*P* value[b]
	Yes	Sometimes	No		
Social media improves the quality of care delivered to patients					.001
Yes	126 (74.1)	40 (23.5)	4 (2.4)	170 (64.2)	
Maybe	42 (56)	24 (32)	9 (12)	75 (28.3)	
No	11 (55)	9 (45)	0 (0)	20 (7.5)	
Social media improves dental practice					.02
Yes	106 (74.7)	30 (21.1)	6 (4.2)	142 (53.6)	
Maybe	48 (57)	29 (35)	7 (8)	84 (31.7)	
No	25 (64)	14 (36)	0 (0)	39 (14.7)	

[a]Percentages were calculated based on the number of participants corresponding to each variable.

[b]Chi-square test was used with an α level of .05.

Discussion

Principal Findings

This study aimed to investigate the use of social media and assess its impact on enhancing dental care and practice among dental professionals. The findings revealed that over half of the participants aged 20 - 30 years are actively using social media in their dental profession, aligning with other studies that indicate younger dentists or health care professionals are more likely to use social media compared to their older counterparts [36,38,40-42]. Variables such as sex, dental specialty, work sector, and years of clinical practice did not significantly influence social media use.

Most respondents reported using social media 1 - 2 days per week to engage with the public and other dental professionals. This is in contrast to other studies where social media was used at least once daily for obtaining information [32,42]. Among social media users in the Philippines, Facebook and YouTube were identified as the primary platforms for accessing clinical information, which is consistent with The Digital Report 2024, highlighting that Facebook has 86.75 million active users in the Philippines [43]. These findings are similar to those reported among dentists in the United Kingdom and the United States, where Facebook is the most commonly used platform in dental practice [1,31]. This was also true as among dental students [7]. In comparison, dentists in Saudi Arabia predominantly use Twitter [22,25] and Instagram [32], while WhatsApp is the preferred platform among Indian dentists [37]. Similarly, South African dentists primarily use Google Plus and Facebook for personal rather than professional purposes [42].

Previous studies have shown that marketing is a common reason for using social media [1,25,31,32]. In our study, 72.1% (191/265) of participants indicated that their primary reason for using social media was for oral health promotion and education. This finding is consistent with earlier research, which found that a high proportion of respondents use social media for dental education and prefer to share clinical work [21]. Other studies also support the notion that education and sourcing oral health information are primary purposes for social media use among dental professionals [22-25]. Social media interventions have positively impacted various aspects of oral health, particularly among adolescents, where platforms like YouTube, WhatsApp, Instagram, and Telegram have been effectively used to promote oral health [27]. For instance, a study demonstrated that oral health education via Telegram significantly improved oral health outcomes among secondary school students [44] and adolescents [45] in Iran. Additionally, Instagram has been shown to support orthodontists in motivating young patients to maintain proper oral hygiene [46], and WhatsApp has been integrated successfully into oral hygiene protocols to enhance patient compliance and oral health during orthodontic treatment [47]. Presenting audiovisual information through YouTube has also significantly improved patient knowledge [48].

Another key purpose of social media use is communication and the exchange of opinions with colleagues. According to the survey, 62.3% (165/265) of participants used social media to connect with patients, friends, and family, while 85.7% (227/265) used it to communicate with other dental professionals. This is supported by studies showing that social media communication, such as Instagram, can reduce patient anxiety prior to dental procedures [28], enhance dentists' communication skills with other professionals [38], and improve patient interactions [30]. Another study found that a significant proportion of the younger population engages with social media to discuss dentistry. This highlights the need for careful management to ensure the dissemination of accurate dental health information [25].

The respondents of the survey also indicated that social media had the potential to enhance dental practice and the quality of care provided to patients. This finding is in agreement with study that shows that dentists who use social media in their practice not only recommend it to their peers but also believe it enhances their practice [22]. Social media can improve various aspects of dental practice, including service provision, advertising, counseling, and oral health education, and it can also be a tool for professional development [49]. Dental organizations and educators can use social media to disseminate information and updates [36,50]. Furthermore, multidimensional health care approaches, including social media, have been

XSL•FO
RenderX

proven highly successful in improving patient care, increasing public knowledge, facilitating research, connecting health care providers, improving medical education, and addressing public health crises [29].

This study has several limitations that should be considered when interpreting the results. First, the study population was limited to dental practitioners in Baguio City, Philippines, which may restrict the generalizability of the findings to the broader dentist population in the country. The uneven distribution of survey participation across various demographic variables, such as age, sex, dental specialty, working sector, and years of clinical practice, may further impact the ability to extrapolate the results to the entire dental community. Additionally, the use of an electronic survey as the sole data collection method presented inherent challenges. A significant portion of the target population may not respond to or use electronic surveys, potentially introducing nonresponse bias and affecting the representativeness of the sample. This bias could be particularly relevant in the context of social media use, as those who are less inclined to participate in electronic surveys may also be less likely to engage with social media platforms. Furthermore, the cross-sectional nature of the study limits the ability to establish causal relationships between social media use and its impact on dental practice and patient care. Longitudinal studies would be necessary to better understand the long-term effects of social media integration in dental settings. Despite these limitations, this study provides valuable insights into the current use of social media among dental practitioners in the Philippines and its perceived impact on dental care and practice. Future research should aim to include larger, more representative samples across different regions and settings to enhance the generalizability of the findings. Additionally, mixed-methods approaches combining quantitative and qualitative data collection techniques could provide a more comprehensive understanding of the topic.

Conclusions

Social media is widely used among younger dental practitioners, primarily for education, communication, and marketing, with a notable impact on enhancing practice quality and patient care. Social media platforms, especially Facebook and YouTube, were commonly used by practitioners to promote oral health, engage with patients and colleagues, and support marketing efforts.

Social media offers a convenient space for professional growth, patient education, and community interaction, which many dentists perceive as beneficial to their practice. However, a low trust in social media information and concerns about its influence on patient-dentist relationships indicate the need for clear guidelines and quality web-based resources. Enhanced support in these areas can help dental professionals maximize the positive impacts of social media on dental care and practice.

Future studies should compare social media with traditional marketing methods in the dental practice such as print ads, community outreach, and patient referrals. These traditional methods remain valuable for building trust, local credibility, and reaching offline patients. Comparing these strategies can help find the best way to combine digital and traditional approaches for better patient engagement and practice success.

Acknowledgments

The authors would like to thank all participants who participated in this study. We also appreciated the funding support from the Faculty of Dentistry at Chulalongkorn University.

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

References

1. Parmar N, Dong L, Eisingerich AB. Connecting with your dentist on Facebook: patients' and dentists' attitudes towards social media usage in dentistry. J Med Internet Res 2018 Jun 29;20(6):e10109. [doi: 10.2196/10109] [Medline: 29959108]
2. Jung YS, Yang HY, Choi YH, et al. Factors affecting use of word-of-mouth by dental patients. Int Dent J 2018 Oct;68(5):314-319. [doi: 10.1111/idj.12387] [Medline: 29572927]
3. Mary AV, Kesavan R, V V, Pi J, G D, S S. Impact of social media on dental treatment needs: a cross sectional pilot study. Int J Appl Dent Sci 2023 Jan 1;9(4):313-317. [doi: 10.22271/oral.2023.v9.i4e.1887]
4. Afful-Dadzie E, Afful-Dadzie A, Egala SB. Social media in health communication: a literature review of information quality. Hlth Inf Manag 2023 Jan;52(1):3-17. [doi: 10.1177/1833358321992683] [Medline: 33818176]
5. Kaplan AM, Haenlein M. Users of the world, unite! The challenges and opportunities of social media. Bus Horiz 2010 Jan;53(1):59-68. [doi: 10.1016/j.bushor.2009.09.003]
6. Social media. Merriam-Webster Dictionary. URL: https://www.merriam-webster.com/dictionary/social%20media [accessed 2024-12-17]

XSL•FO
RenderX

7. Kenny P, Johnson IG. Social media use, attitudes, behaviours and perceptions of online professionalism amongst dental students. Br Dent J 2016 Nov 18;221(10):651-655. [doi: 10.1038/sj.bdj.2016.864] [Medline: 27857111]
8. Sivaramakrishnan G, AbdulAmeer F, Faisal F, et al. Use of social media to view and post dentistry-related information in Bahrain: a cross-sectional study. Healthc Inform Res 2023 Jan;29(1):31-39. [doi: 10.4258/hir.2023.29.1.31] [Medline: 36792099]
9. Freire Y, Sánchez MG, Suárez A, Joves G, Nowak M, Díaz-Flores García V. Influence of the use of social media on patients changing dental practice: a web-based questionnaire study. BMC Oral Health 2023 Jun 6;23(1):365. [doi: 10.1186/s12903-023-03078-9] [Medline: 37277765]
10. MacDonald J, Sohn S, Ellis P. Privacy, professionalism and Facebook: a dilemma for young doctors. Med Educ 2010 Aug;44(8):805-813. [doi: 10.1111/j.1365-2923.2010.03720.x] [Medline: 20633220]
11. Malik A, Heyman-Schrum C, Johri A. Use of Twitter across educational settings: a review of the literature. Int J Educ Technol High Educ 2019 Dec;16(1):1-22. [doi: 10.1186/s41239-019-0166-x]
12. Forgie SE, Duff JP, Ross S. Twelve tips for using Twitter as a learning tool in medical education. Med Teach 2013;35(1):8-14. [doi: 10.3109/0142159X.2012.746448] [Medline: 23259608]
13. Madathil KC, Rivera-Rodriguez AJ, Greenstein JS, Gramopadhye AK. Healthcare information on YouTube: a systematic review. Health Informatics J 2015 Sep;21(3):173-194. [doi: 10.1177/1460458213512220] [Medline: 24670899]
14. Simsek H, Buyuk SK, Cetinkaya E. YouTube as a source of information on oral habits. J Indian Soc Pedod Prev Dent 2020;38(2):115-118. [doi: 10.4103/JISPPD.JISPPD_357_19] [Medline: 32611855]
15. Guo J, Yan X, Li S, Van der Walt J, Guan G, Mei L. Quantitative and qualitative analyses of orthodontic-related videos on YouTube. Angle Orthod 2020 May 1;90(3):411-418. [doi: 10.2319/082019-542.1] [Medline: 33378439]
16. Ooi HL, Kelleher MGD. Instagram dentistry. Prim Dent J 2021 Mar;10(1):13-19. [doi: 10.1177/2050168420980964] [Medline: 33722130]
17. Patil P, Pawar AM, Agustin Wahjuningrum D, Maniangat Luke A, Reda R, Testarelli L. Connectivity and integration of Instagram use in the lives of dental students and professionals: a country-wide cross-sectional study using the InstaAA Questionnaire. J Multidiscip Healthc 2023;16:2963-2976. [doi: 10.2147/JMDH.S430973] [Medline: 37818291]
18. Buyuk SK, Alpaydın MT, Imamoglu T. Social media and orthodontics: an analysis of orthodontic-related posts on Instagram. Aust Orthod J 2020 Jan 1;36(2):153-159. [doi: 10.21307/aoj-2020-017]
19. Fraticelli L, Smentek C, Tardivo D, et al. Characterizing the content related to oral health education on TikTok. Int J Environ Res Public Health 2021 Dec 16;18(24):13260. [doi: 10.3390/ijerph182413260] [Medline: 34948869]
20. Kılınç DD. Is the information about orthodontics on Youtube and TikTok reliable for the oral health of the public? A cross sectional comparative study. J Stomatol Oral Maxillofac Surg 2022 Oct;123(5):e349-e354. [doi: 10.1016/j.jormas.2022.04.009] [Medline: 35470122]
21. Kurian N, Varghese IA, Cherian JM, et al. Influence of social media platforms in dental education and clinical practice: a cross-sectional survey among dental trainees and professionals. J Dent Educ 2022 Aug;86(8):1036-1042. [doi: 10.1002/jdd.12914] [Medline: 35257384]
22. Hamasha AAH, Alghofaili N, Obaid A, et al. Social media utilization among dental practitioner in Riyadh, Saudi Arabia. Open Dent J 2019 Feb 28;13(1):101-106. [doi: 10.2174/1874210601913010101]
23. Aboalshamat K, Alkiyadi S, Alsaleh S, et al. Attitudes toward social media among practicing dentists and dental students in clinical years in Saudi Arabia. Open Dent J 2019 Mar 28;13(1):143-149. [doi: 10.2174/1874210601913010143]
24. Maharani DA, El Tantawi M, Yoseph MG, Rahardjo A. The use of internet platforms for oral health information and associated factors among adolescents from Jakarta: a cross sectional study. BMC Oral Health 2021 Jan 7;21(1):22. [doi: 10.1186/s12903-020-01387-x] [Medline: 33413315]
25. Althunayan A, Alsalhi R, Elmoazen R. Role of social media in dental health promotion and behavior change in Qassim province, Saudi Arabia. Int J Med Health Res 2018;4(2):98-103 [FREE Full text]
26. El Tantawi M, Bakhurji E, Al-Ansari A, AlSubaie A, Al Subaie HA, AlAli A. Indicators of adolescents' preference to receive oral health information using social media. Acta Odontol Scand 2019 Apr;77(3):213-218. [doi: 10.1080/00016357.2018.1536803] [Medline: 30632864]
27. Farrokhi F, Ghorbani Z, Farrokhi F, Namdari M, Salavatian S. Social media as a tool for oral health promotion: a systematic review. PLoS ONE 2023;18(12):e0296102. [doi: 10.1371/journal.pone.0296102] [Medline: 38113216]
28. Sivrikaya EC, Yilmaz O, Sivrikaya P. Dentist-patient communication on dental anxiety using the social media: a randomized controlled trial. Scand J Psychol 2021 Dec;62(6):780-786. [doi: 10.1111/sjop.12769] [Medline: 34333783]
29. Farsi D. Social media and health care, part I: literature review of social media use by health care providers. J Med Internet Res 2021 Apr 5;23(4):e23205. [doi: 10.2196/23205] [Medline: 33664014]
30. Ghandhi D, Bodani N, Lal A, et al. Evaluation of social media usage by dental practitioners of Pakistan for professional purposes - a cross-sectional study. Clin Cosmet Investig Dent 2022;14:245-252. [doi: 10.2147/CCIDE.S374111] [Medline: 35975279]
31. Henry RK, Molnar A, Henry JC. A survey of US dental practices' use of social media. J Contemp Dent Pract 2012 Mar 1;13(2):137-141. [doi: 10.5005/jp-journals-10024-1109] [Medline: 22665737]

XSL•FO
RenderX

32. Bahabri RH, Zaidan AB. The impact of social media on dental practice promotion and professionalism amongst general dental practitioners and specialists in KSA. J Taibah Univ Med Sci 2021 Jun;16(3):456-460. [doi: 10.1016/j.jtumed.2020.12.017] [Medline: 34140874]
33. Number of internet and social media users worldwide as of July 2024 (in billions). Statista. URL: https://www.statista.com/statistics/617136/digital-population-worldwide/ [accessed 2024-12-17]
34. Most popular social networks worldwide as of April 2024, by number of monthly active users (in millions). Statista. URL: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/ [accessed 2024-12-17]
35. Social media in the Philippines - statistics & facts. Statista. URL: https://www.statista.com/topics/6759/social-media-usage-in-the-philippines/#topicOverview [accessed 2024-12-17]
36. Al-Khalifa KS, Al-Swuailem AS, AlSheikh R, et al. The use of social media for professional purposes among dentists in Saudi Arabia. BMC Oral Health 2021 Jan 12;21(1):26. [doi: 10.1186/s12903-021-01390-w] [Medline: 33435911]
37. Jamkhande A, Pangarikar AB, Khairnar MR, Gavali N, Patil Y, Wadgave U. Social media use in dental practice by dental practitioners of India: a cross-sectional survey. J Clin Diagn Res 2022;16(6). [doi: 10.7860/JCDR/2022/56404.16533]
38. Makkar N, Lawrence D, Veeresh D, Banik A, Kumar A, Mehar AS. Assessment of social media usage and its influence among dental faculty members in Davangere city – a cross sectional survey. J Indian Assoc Public Health Dent 2018;16(3):246. [doi: 10.4103/jiaphd.jiaphd_143_17]
39. Kamarudin Y, Mohd Nor NA, Libamin AC, et al. Social media use, professional behaviors online, and perceptions toward e-professionalism among dental students. J Dent Educ 2022 Aug;86(8):958-967. [doi: 10.1002/jdd.12912] [Medline: 35247218]
40. Almaiman S, Bahkali S, Al Farhan A, Bamuhair S, Househ M, Alsurimi K. The prevalence of using social media among healthcare professionals in Saudi Arabia: a pilot study. Stud Health Technol Inform 2015;213:263-266. [Medline: 26153010]
41. Alsobayel H. Use of social media for professional development by health care professionals: a cross-sectional web-based survey. JMIR Med Educ 2016 Sep 12;2(2):e15. [doi: 10.2196/mededu.6232] [Medline: 27731855]
42. Snyman L, Visser JH. The adoption of social media and social media marketing by dentists in South Africa. SADJ 2014 Jul;69(6):258. [Medline: 26548200]
43. Kemp S. Digital 2024: the Philippines. Datareportal. 2024 Feb 21. URL: https://datareportal.com/reports/digital-2024-philippines [accessed 2024-12-17]
44. Zareban I, Rahmani A, Allahqoli L, et al. Effect of education based on trans-theoretical model in social media on students with gingivitis; a randomized controlled trial. Health Educ Health Prom 2022;10(1):1-7 [FREE Full text]
45. Scheerman JFM, Hamilton K, Sharif MO, Lindmark U, Pakpour AH. A theory-based intervention delivered by an online social media platform to promote oral health among Iranian adolescents: a cluster randomized controlled trial. Psychol Health 2020 Apr;35(4):449-466. [doi: 10.1080/08870446.2019.1673895] [Medline: 31621423]
46. Scribante A, Gallo S, Bertino K, Meles S, Gandini P, Sfondrini MF. The effect of chairside verbal instructions matched with Instagram social media on oral hygiene of young orthodontic patients: a randomized clinical trial. Appl Sci (Basel) 2021;11(2):706. [doi: 10.3390/app11020706]
47. Zotti F, Dalessandri D, Salgarello S, et al. Usefulness of an app in improving oral hygiene compliance in adolescent orthodontic patients. Angle Orthod 2016 Jan;86(1):101-107. [doi: 10.2319/010915-19.1] [Medline: 25799001]
48. Al-Silwadi FM, Gill DS, Petrie A, Cunningham SJ. Effect of social media in improving knowledge among patients having fixed appliance orthodontic treatment: a single-center randomized controlled trial. Am J Orthod Dentofacial Orthop 2015 Aug;148(2):231-237. [doi: 10.1016/j.ajodo.2015.03.029] [Medline: 26232831]
49. Al-Nahdi S, Alanzi T, Alhodaib H. Use of social media by healthcare providers in saudi arabia: reasons and usefulness for their professional development. Presented at: 19th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES); Oct 16-19, 2020; Xuzhou, China. [doi: 10.1109/DCABES50732.2020.00091]
50. Terrasse M, Gorin M, Sisti D. Social media, e-health, and medical ethics. Hast Cent Rep 2019 Jan;49(1):24-33. [doi: 10.1002/hast.975] [Medline: 30790306]

Abbreviations

STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

XSL•FO
RenderX

Edited by A Mavragani; submitted 05.09.24; peer-reviewed by J Hunsrisakhun, K Sadasiva; revised version received 19.11.24; accepted 20.11.24; published 03.01.25.

Please cite as:
Acosta JM, Detsomboonrat P, Pisarnturakit PP, Urwannachotima N
The Use of Social Media on Enhancing Dental Care and Practice Among Dental Professionals: Cross-Sectional Survey Study
JMIR Form Res 2025;9:e66121
URL: https://formative.jmir.org/2025/1/e66121
doi:10.2196/66121

© Joseph Macadaeg Acosta, Palinee Detsomboonrat, Pagaporn Pantuwadee Pisarnturakit, Nipaporn Urwannachotima. Originally published in JMIR Formative Research (https://formative.jmir.org), 3.1.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Artificial Intelligence–Powered Training Database for Clinical Thinking: App Development Study

Heng Wang[1*], MEd; Danni Zheng[1*], PhD; Mengying Wang[2], MEng; Hong Ji[2], MEng; Jiangli Han[3], MD; Yan Wang[4], MD; Ning Shen[5], MD; Jie Qiao[6], MD

[1]Education Department, Peking University Third Hospital, Beijing, China
[2]Information Management and Big Data Center, Peking University Third Hospital, Beijing, China
[3]Department of Cardiology, Peking University Third Hospital, Beijing, China
[4]Department of Obstetrics and Gynecology, Peking University Third Hospital, Beijing, China
[5]Department of Pulmonary and Critical Care Medicine, Peking University Third Hospital, Beijing, China
[6]Centre for Reproductive Medicine, Department of Obstetrics and Gynecology, Peking University Third Hospital, Beijing, China
[*]these authors contributed equally

Corresponding Author:
Jie Qiao, MD
Centre for Reproductive Medicine, Department of Obstetrics and Gynecology, Peking University Third Hospital, Beijing, China

Abstract

Background: With the development of artificial intelligence (AI), medicine has entered the era of *intelligent medicine*, and various aspects, such as medical education and talent cultivation, are also being redefined. The cultivation of clinical thinking abilities poses a formidable challenge even for seasoned clinical educators, as offline training modalities often fall short in bridging the divide between current practice and the desired ideal. Consequently, there arises an imperative need for the expeditious development of a web-based database, tailored to empower physicians in their quest to learn and hone their clinical reasoning skills.

Objective: This study aimed to introduce an app named "XueYiKu," which includes consultations, physical examinations, auxiliary examinations, and diagnosis, incorporating AI and actual complete hospital medical records to build an online-learning platform using human-computer interaction.

Methods: The "XueYiKu" app was designed as a contactless, self-service, trial-and-error system application based on actual complete hospital medical records and natural language processing technology to comprehensively assess the "clinical competence" of residents at different stages. Case extraction was performed at a hospital's case data center, and the best-matching cases were differentiated through natural language processing, word segmentation, synonym conversion, and sorting. More than 400 teaching cases covering 65 kinds of diseases were released for students to learn, and the subjects covered internal medicine, surgery, gynecology and obstetrics, and pediatrics. The difficulty of learning cases was divided into four levels in ascending order. Moreover, the learning and teaching effects were evaluated using 6 dimensions covering systematicness, agility, logic, knowledge expansion, multidimensional evaluation indicators, and preciseness.

Results: From the app's first launch on the Android platform in May 2019 to the last version updated in May 2023, the total number of teacher and student users was 6209 and 1180, respectively. The top 3 subjects most frequently learned were respirology (n=606, 24.1%), general surgery (n=506, 20.1%), and urinary surgery (n=390, 15.5%). For diseases, pneumonia was the most frequently learned, followed by cholecystolithiasis (n=216, 14.1%), benign prostate hyperplasia (n=196, 12.8%), and bladder tumor (n=193, 12.6%). Among 479 students, roughly a third (n=168, 35.1%) scored in the 60 to 80 range, and half of them scored over 80 points (n=238, 49.7%). The app enabled medical students' learning to become more active and self-motivated, with a variety of formats, and provided real-time feedback through assessments on the platform. The learning effect was satisfactory overall and provided important precedence for establishing scientific models and methods for assessing clinical thinking skills in the future.

Conclusions: The integration of AI and medical education will undoubtedly assist in the restructuring of education processes; promote the evolution of the education ecosystem; and provide new convenient ways for independent learning, interactive communication, and educational resource sharing.

(JMIR Form Res 2025;9:e58426) doi:10.2196/58426

KEYWORDS

artificial intelligence; clinical thinking ability; virtual medical records; distance education; medical education; online learning

XSL•FO
RenderX

Introduction

Clinical thinking refers to the ability of doctors to understand, diagnose, and treat diseases through analysis, synthesis, judgment, reasoning, and other thinking activities by using their medical knowledge, skills, and experience [1]. Clinical reasoning skills include a series of strategies such as correlating information, conducting a comprehensive analysis to form one or more diagnostic hypotheses, weighing the risks and benefits of diagnostic tests and treatments, as well as formulating a reasonable diagnosis and treatment plan [2]. This detailed process includes collecting and evaluating clinical information, selecting diagnostic tests, evaluating test results, developing diagnostic hypotheses, and weighing treatment options and risks [3].

Teaching clinical thinking ability is a challenging task even for experienced clinical teachers. As clinical knowledge is the basis of clinical thinking, the cultivation of clinical thinking needs to go through the links of guidance, deliberate practice, and feedback, and it takes a long time to cultivate efficient clinical thinking ability. The level of clinical thinking of residents is directly related to the safety and quality of medical treatment [4]. Therefore, we should constantly improve the understanding of the clinical thinking training and improve the training quality of residents.

Therefore, the cultivation of clinical thinking ability is a necessary condition for the transition from medical students to clinicians and is currently the key point and difficulty among resident training programs. Among the training processes, the most effective method is based on real or virtual case training, combined with feedback and reflection to continuously improve. Previous studies have shown that clinical thinking ability is scarce among physicians, while offline training cannot meet the gap between the present reality and the ideal requirements [5]. Therefore, there is an urgent need for the development of a web-based database to help physicians learn and exercise clinical thinking ability.

Artificial intelligence (AI) is a novel technology that can be utilized to research and develop the theories, methods, technologies, and applications used to simulate, extend, and expand human intelligence. AI belongs to the interdisciplinary field of natural science, social science, and technological science [6]. AI has a history of more than 70 years, and its application in the medical field mainly focuses on five aspects: medical imaging, auxiliary diagnosis, drug research and development, health management, and disease prediction.

There have been several reports and studies on the use of AI in medical teaching and training of clinical thinking ability. For example, Diagnostic Reasoning (DxR) Clinician (DxR Development Group, Inc.) was developed in 1994 by Myers and Dorsey in the United States to train clinical thinking among students [7]. In 2006, the BP network neural algorithm in AI was applied to teach quality monitoring instead of manually teaching daily monitoring [8]. In 2011, Knewton, an adaptive education platform, offered personalized learning services by extracting learning data from students. In 2017, Tsinghua University launched a study system combined with AI and full quantitative reality virtual technology. In this system, patients' image data can be converted into a holographic three-dimensional anatomical structure of the human body and mapped in the virtual space [9]. In 2020, a competency-based learning and assessment system (CBLAS) for residency training was designed to provide resident physicians with clinical assessments and learning in order to enhance the learning of trainees and reduce the burden of assessments [10]. However, in the era of AI, there are still issues that exist in medical teaching, such as the lack of general education, the lack of classroom interaction, and the lag in curriculum construction [11].

On this basis, the present study introduces an app named "XueYiKu," which includes consultations, physical examinations, auxiliary examinations, and diagnosis, by incorporating AI and actual complete hospital medical records to build an online-learning platform using human-computer interaction. This application could cultivate students' ability to solve clinical problems as well as improve teaching quality and efficiency, thereby achieving the sharing of high-quality teaching resources.

Methods

Design and Study Population

The "XueYiKu" app was designed as a contactless, self-service, trial-and-error system application based on actual complete hospital medical records and natural language processing technology to build an online-learning platform using human-computer interaction. Training participants included medical undergraduates, medical graduate students, residents, chief residents, and attending doctors. The purpose of the app was to improve clinical thinking ability through independent study accompanied by real-time feedback. The XueYiKu app was launched on the Android platform on May 14, 2019, and on the IOS platform on August 3, 2019. After 27 updates, the last version of the XueYiKu app was updated on January 12, 2022.

Data Collection

The general database of the XueYiKu app was based on a collection of real medical records and data from general teaching, including a standard clinical question database, virtual medical record database, and student learning record database. Using information extraction technology in natural language processing from real medical records, the XueYiKu app could poststructure the content from real electronic medical records, generate questions, and match answers online through machine learning. In the teaching cases, the automatically extracted answers matched 90% of the original medical records. The structured cases in the electronic medical record system of the hospital were then screened by a teaching team from Peking University Third Hospital, and appropriate medical records were selected and reviewed by the teaching team. Once the teaching case was determined, relevant inspection results were also extracted, including inspections, auxiliary inspections, and ultrasound, imaging, and pathology findings.

XSL•FO
RenderX

Outcome Measures

Clinical thinking ability involves extremely high requirements for the objectivity and standardization of evaluation. Hence, a multidimensional automatic evaluation system is required. With reference to the residents' standardized training report, the whole process of student learning was recorded systematically, and clinical thinking was evaluated in multiple dimensions according to the learning situation of the student, including preciseness, logic, systematicness, agility, knowledge expansion, and multidimensional and comprehensive application (Table 1). Based on the scores of students in various dimensions of learning, a radar chart of clinical thinking ability is drawn to assess the compliance of medical students' clinical thinking ability.

Table . Seven elements for clinical thinking evaluation in the XueYiKu app.

Elements	Contents
Preciseness	Comprehensive evaluation of diagnosis and differential diagnosis
Logic	The logical sequence evaluation of inquiry, physical examination, auxiliary examination, and diagnosis
Systematicness	Comprehensive evaluation of the completion rate of inquiry, physical examination, auxiliary examination, diagnosis, and differential diagnosis
Agility	Comprehensive evaluation of the time of inquiry, physical examination, auxiliary examination, diagnosis, and differential diagnosis
Knowledge expansion	Accuracy and logical evaluation of diagnosis and differential diagnosis
Multidimensional evaluation indicators	Comprehensive evaluation of diagnosis and treatment programs
Comprehensive application	Comprehensive application ability of basic and clinical knowledge (for evaluation of level 4 individually)

Evaluation and Feedback

We recruited all students in Peking University Hospital to evaluate the teaching effect of the XueYiKu app, including medical undergraduates, medical graduate students, residents, chief residents, and attending doctors. Both the questionnaire method and interview method were utilized. Web-based surveys, developed by the study team, were conducted upon completion of the case study so that students could fill in and evaluate their learning effect. Through interviews with relevant teachers and experts, the common clinical path extracted using AI technology was determined. Meanwhile, the range of clinical thinking ability evaluation was confirmed according to the requirements of different training stages. The questionnaire method was used to investigate the following two factors: (1) the current status of clinical thinking training regarding the ability of medical students and residents to clarify the problems in clinical thinking training and improve the content of the database; (2) the experience of medical students and residents regarding the use of the case database and self-evaluation of learning effects to complete the effect evaluation of the XueYiKu app.

Data Processing

Details of the virtual case system and AI teaching development based on actual complete hospital medical records and natural language processing technology have been described previously [12]. A flow chart of the program design from data collection to interaction between students and teachers is shown in Figure 1. Case extraction was performed at a hospital's case data center, and the best-matching cases were extracted through natural language processing, word segmentation, synonym conversion, and sorting. A standard clinical questioning data module, virtual case data module, and student learning difficulty module were established to achieve simulation. Students can view the objective examination and inspection data of actual cases, including details of the consultation and physical examination, and automatically provide their learning response via a multidimensional evaluation system. In order to assess the changes in students' clinical thinking after using the XueYiKu app, 15 medical graduate students were subjected to two simulation tests before and after learning through the virtual case system. The tests, which included the full-process case examination of cases having the same difficulty level, examined core clinical thinking test points such as consultation, physical examination, and disposal, and generated multidimensional evaluation indicators (rigor, logic, system, agility, and knowledge expansion). Thus, a complete and credible evaluation system was developed.

For statistical analysis, standard descriptive statistics were used to describe the study population: continuous variables were presented as the mean and SD, and categorical variables were presented as percentages.

XSL•FO
RenderX

Figure 1. Flow chart of the program design from data collection to interaction between students and teachers in the XueYiKu app.

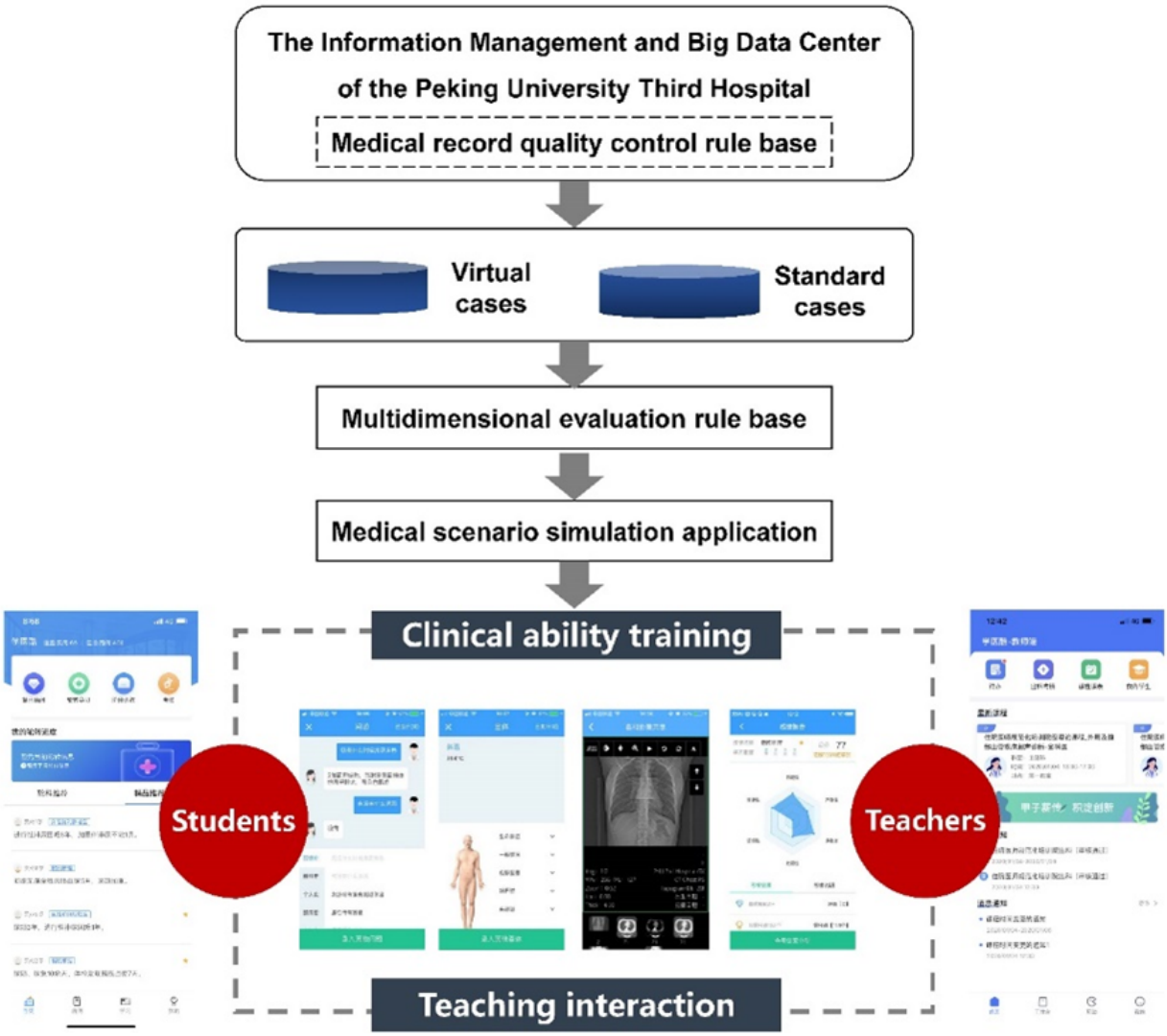

Ethical Considerations

This study was reviewed and approved by the Peking University Third Hospital Medical Science Research Ethics Committee (No. IRB00006761-M2022063). Invitations were dispatched to prospective participants via the XueYiKu app, accompanied by a comprehensive informed consent form outlining the nature of the investigation. Participants were clearly notified that the interviews would be audio-recorded solely for analytical purposes and would remain exclusively accessible to the research team. Furthermore, they were assured of their right to withdraw their consent at any juncture during the interview process. All information gathered about the participants was kept private and confidential. All study data were deidentified.

Results

Overview

After concealment processing, more than 400 teaching cases covering 65 kinds of diseases were released for students to learn from, and the subjects covered internal medicine, surgery, gynecology and obstetrics, and pediatrics (Table 2). The difficulty in learning cases was divided into 4 levels in ascending order: level 1 provided all score items for students to choose from; except for items from level 1, level 2 provided additional interference items; compared with level 2, level 3 displayed some score items and all interference items to enhance difficulty; for level 4, there was neither a score item nor an interference item, and the system would match the results from the students' voice or manual input of the questions.

XSL•FO
RenderX

Table . Categories of subjects and diseases for learning and evaluation in the database of the XueYiKu app.

Subjects	Category	Diseases
Internal Medicine	20	Coronary heart disease, heart failure, arrhythmia, hypertension, pneumonia, chronic airway inflammatory disease, pleural effusion, pulmonary embolism, upper gastrointestinal bleeding, pancreatitis, inflammatory bowel disease, ascites, nephrotic syndrome, acute kidney injury, diabetes, hyperthyroidism, leukemia, anemia, systemic lupus erythematosus, rheumatoid arthritis
Surgery	20	Inguinal hernia, appendicitis, gallbladder stones, bile duct stones, stomach cancer, colon cancer, rectal cancer, liver cancer, pancreatic cancer, lung cancer, pneumothorax, pituitary tumor, coronary heart disease, benign prostatic hyperplasia, bladder tumor, acute respiratory distress syndrome, cervical myelopathy disease, lumbar disc herniation, knee osteoarthritis, ankle fracture
Gynecology and obstetrics	15	Ectopic pregnancy, uterine fibroids, ovarian tumors, dysfunctional uterine bleeding, endometrial cancer, cervical cancer, ovarian cancer, hypertension during pregnancy (including gestational hypertension, preeclampsia, eclampsia, chronic hypertension, and chronic hypertension combined with preeclampsia), gestational diabetes or diabetes (impaired glucose tolerance) combined with pregnancy, prenatal hemorrhage (including placenta previa and abruption of the placenta)
Pediatrics	10	Neonatal hyperbilirubinemia, neonatal respiratory distress syndrome, neonatal sepsis, neonatal hypoxic-ischemic encephalopathy, pneumonia, diarrhea, nutritional diseases, nervous system disease, Kawasaki disease, and kidney disease

App Use Status

From the first launch on the Android platform on May 14, 2019, to the last version of the XueYiKu app updated in May 2023, the total number of teacher and student users was 6209 and 1180, respectively. Figure 2 represents the use status of the XueYiKu app during the test period and operation period. In total, there were 3224 person-times in the last 3 years, with an average of 771 person-times per year. From 2019 to 2021, the person-times logged in the application were the highest in 2019, at 1032 (125.4 on average per month), and increased from 646 in 2020 (53.8 on average per month) to 994 in 2021 (76.6 on average per month).

Figure 2. Changes in the number of student users of the XueYiKu app from May 2019 to February 2023.

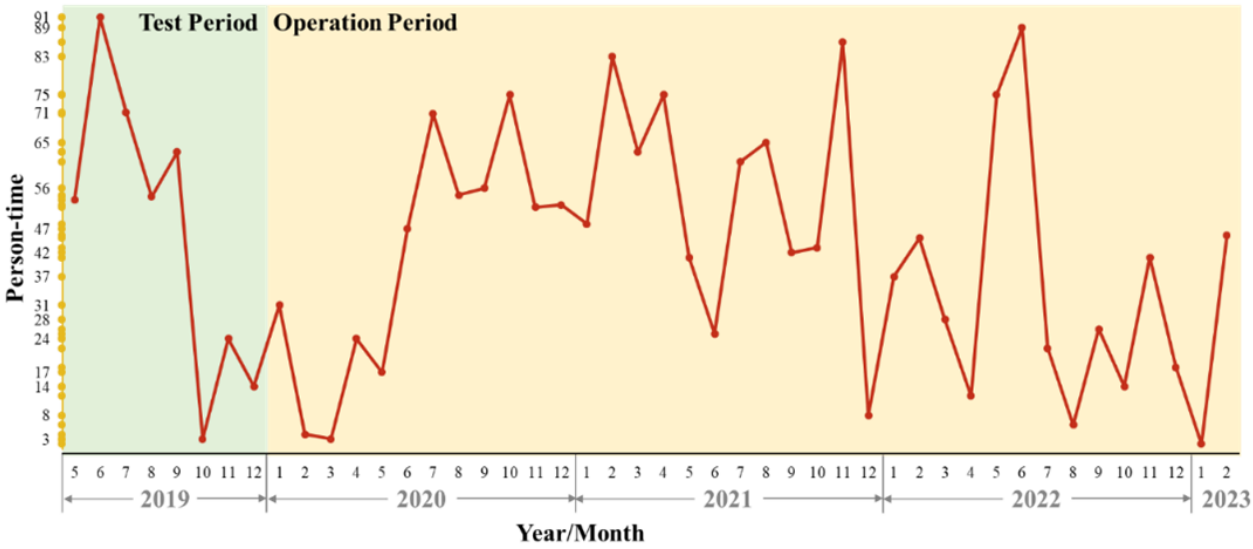

XSL•FO
RenderX

We calculated the frequency at which subjects and related diseases were learned by students, and the top 10 most frequently learned subjects and diseases are listed in Table 3 in descending order. For the second discipline, the most frequently learned subjects were respirology (n=606, 62.2% in internal medicine), general surgery (n=506, 48.8% in surgery), and urinary surgery (n=390, 37.6% in surgery). For diseases, pneumonia was the most frequently learned, with 452 person-times (58.4% in internal medicine), followed by cholecystolithiasis (n=216, 28.4% in surgery), benign prostate hyperplasia (n=196, 25.8% in surgery), and bladder tumor (n=193, 25.4% in surgery).

Table . The top 10 most frequently learned subjects and diseases in the XueYiKu app.

	Person-time	Percentage (%)
Subjects[a]		
Internal Medicine	975	38.8
Respirology	606	62.2
Cardiology	201	20.6
Gastroenterology	88	9
Endocrinology	80	8.2
Surgery	1036	41.2
General surgery	506	48.8
Urinary surgery	390	37.6
Thoracic surgery	140	13.5
Gynecology and obstetrics	338	13.5
Gynecology	175	51.8
Obstetrics	163	48.2
Pediatrics	163	6.5
Diseases[a]		
Internal Medicine	774	50.4
Pneumonia	452	58.4
Coronary heart disease	99	12.8
Lung cancer	79	10.2
Arrhythmia	76	9.8
Chronic airway inflammation	68	8.8
Surgery	761	49.6
Cholecystolithiasis	216	28.4
Benign prostate hyperplasia	196	25.8
Bladder tumor	193	25.4
Appendicitis	79	10.4
Inguinal hernia	77	10.1

[a]The top 10 subjects and diseases were listed in descending order due to quantitative limitations.

Evaluation of the App and its Learning Effects

To evaluate the learning effect of the XueYiKu app, the scores of 479 students were analyzed. On the basis of a full score of 100, 15% (n=73) of students scored less than a passing grade of 60 points, roughly a third (n=168, 35.1%) had a score in the 60 to 80 range, and half of them had a score of over 80 points (n=238, 49.7%). A radar chart (Figure 3) with specific scores (Table 4) was presented considering multiple dimensions to evaluate the learning situation of students: students who learned cases with difficulty levels 3 - 4 achieved excellent learning outcomes regarding preciseness (83.4 for level 3), multiple dimensions (87.4 for level 4), and knowledge expansion (86.5 for level 3). However, students who learned cases with a difficulty of level 4 did not score as high as the other students in systematicness (34.1 for level 4), agility (30.1 for level 4), and logic (21.1 for level 4). In contrast, students who learned from cases with level 3 difficulty had positive effects in many areas (86.5 in knowledge expansion and 83.4 in preciseness). In fact, there was a high level of satisfaction with knowledge expansion and preciseness in all learners, but it is worth noting

XSL•FO
RenderX

that all levels were completely negative about systematicness, agility, and logic.

Figure 3. Radar chart with specific scores of clinical thinking evaluation for the XueYiKu app by difficulty in learning cases.

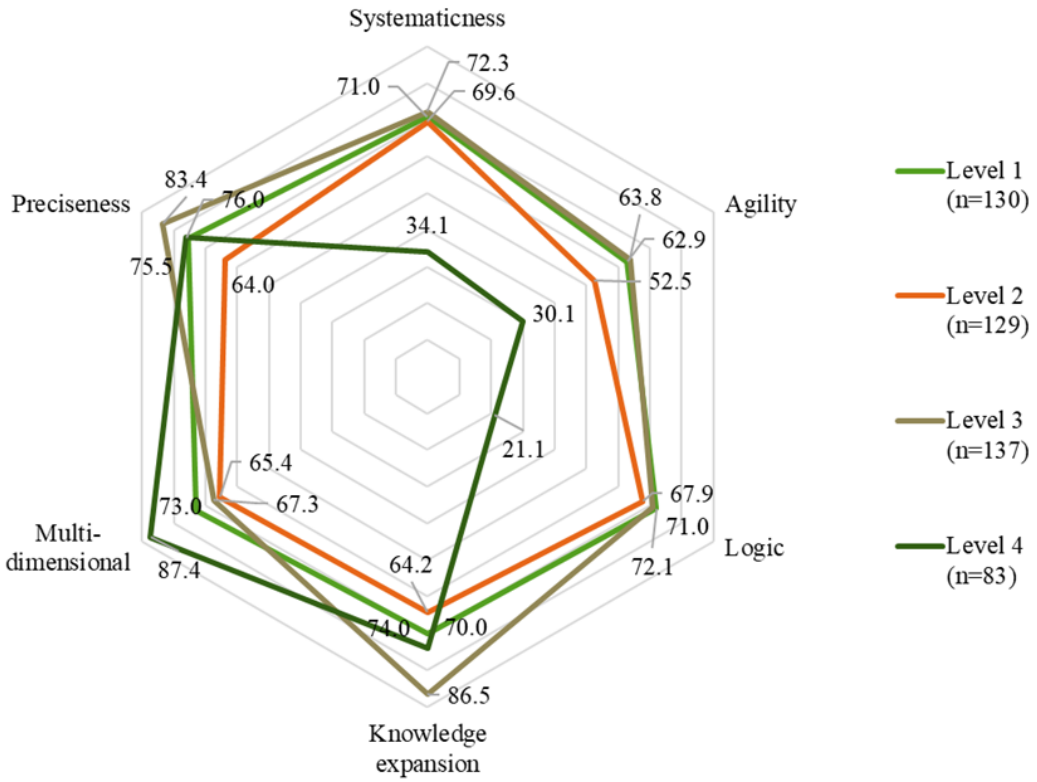

Table . Scores of clinical thinking for the XueYiKu app by the difficulty in learning cases and 6 dimensions according to the learning situation of the student.

	Level 1 (n=130)	Level 2 (n=129)	Level 3 (n=137)	Level 4 (n=83)
Systematicness	71.0	69.6	72.3	34.1
Agility	62.9	52.5	63.8	30.1
Logic	72.1	67.9	71.0	21.1
Knowledge expansion	70.0	64.2	86.5	74.0
Multidimensional evaluation indicators	73.0	65.4	67.3	87.4
Preciseness	75.5	64.0	83.4	76.0

Evaluation of the App and Its Learning Effects

This section provides information regarding the 70 students' satisfaction of the XueYiKu app and their self-assessment of its learning effect (Figure 4). Overall, most students were satisfied with the app. The degree of satisfaction kept increasing considering the app function, the personal learning effect, and the case quality. For the self-assessment of learning outcomes, all sections received a rating of more than satisfactory, but the method of examination was rated the least satisfactory. In other sections, the technique of questioning received the highest satisfaction rating, followed by the interpretation of examination items, with both categories receiving more than 80% positive feedback. In contrast, the diagnostic accuracy and the order and interpretation of treatment ranked next in order (~70%), and the other factors were at approximately 60%. Most of the dissatisfaction was rated for the order of treatment and the comprehensiveness of the diagnosis, which were responsible for more than 20% of the negative ratings. Except for the questioning technique, the proportion of negative ratings was more pronounced in all sections. Additionally, for the method of examination, about 30% rated "medium," which differed greatly from other sections' ratings.

XSL•FO
RenderX

Figure 4. Self-assessment and satisfaction of 70 students towards the learning outcomes for the XueYiKu app.

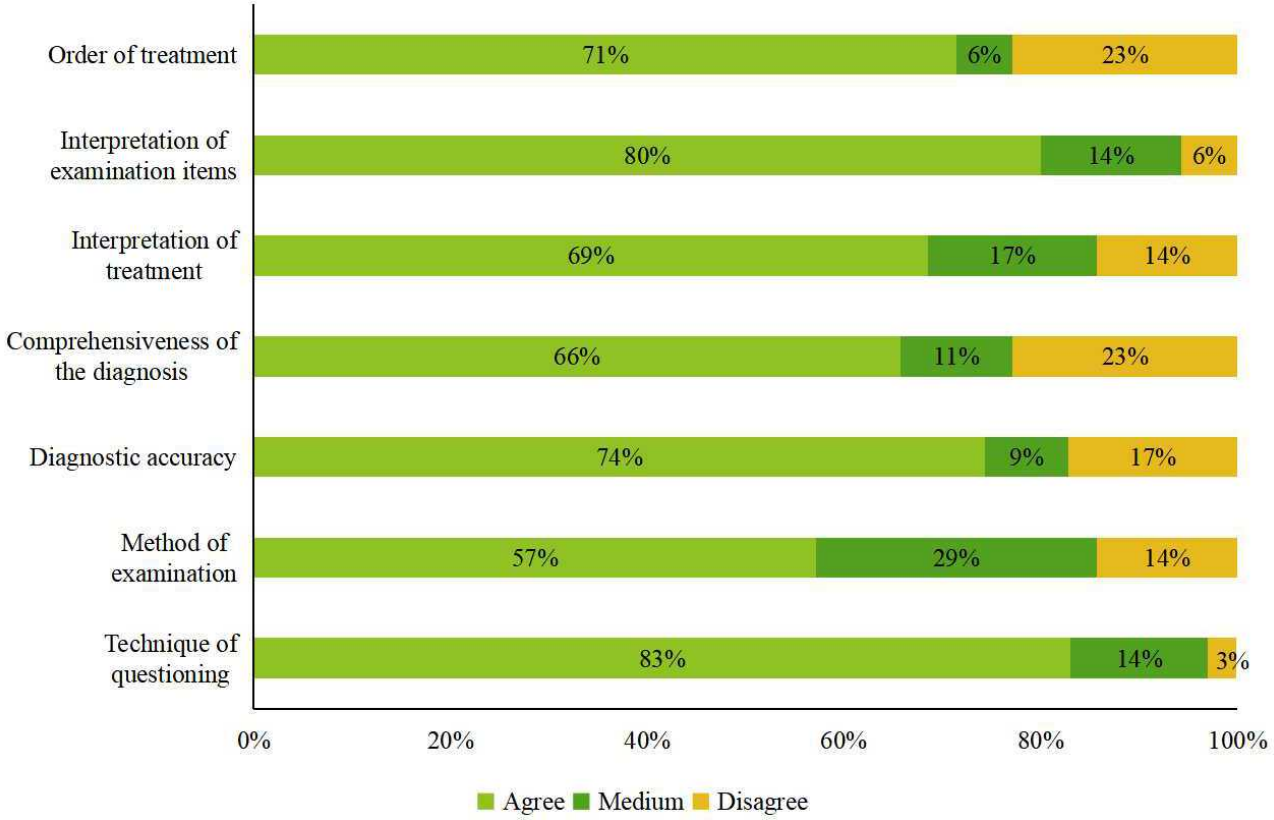

Feedback on the App and Its Teaching Effects

According to the feedback, there were some issues associated with the app that are worthy of attention. First, the original login method should be simplified, as well as that for the online case communication and sharing section. Additionally, discussion and comment functions should be added. Regarding data processing, the AI answers could optimize the fuzzy search, and the system could be better connected to the residency system so that students can check and manage the patient's detailed medical record information. Regarding the case corpus, there is still no section in the app to read images. In addition, not only do the users of the app want to obtain clearer guidance or feedback in future updates, but it would also be beneficial if special departments fulfilled the auxiliary examination and clinical guidelines. Finally, some of the cases included in the app are so simple that they are only suited for undergraduate medical students; it would be beneficial to add cases that are rare, complex, or even life-threatening, and to include the postoperative complications, judgment, treatment, and other information relevant to the case.

Discussion

Principal Findings

Clinical thinking training for medical students is very important; however, few design cases rely on IT for self-learning. The current study uses AI and big data platforms to establish a virtual learning platform based on real clinical cases. The XueYiKu app can assist in the development of medical students' clinical thinking skills by allowing them to perform consultation, physical examination, auxiliary examination, diagnosis, and treatment on virtual patients; the app can also design real cases with different difficulties to meet the needs of residents' clinical thinking training at different stages. Through the evaluation of students' experiences with the app, this study concludes that medical students' learning can become more active and self-motivated, using a variety of formats, with real-time feedback through the platform's assessment. The learning effect is also satisfactory overall and provides important methods and ideas for establishing scientific models and methods for assessing clinical thinking skills in the future.

Comparison to Prior Work

Analysis of the data shows that the number of case studies was positively correlated with the number of residents in the specialty. The top 4 diseases most frequently learned included pneumonia, cholecystolithiasis, benign prostate hyperplasia, and bladder tumor, all of which are common, require adequate diagnosis and differential diagnosis, and meet the basic goal of clinical thinking training at the residency stage. The platform contains 65 common diseases and 400 study cases covering internal medicine, surgery, obstetrics and gynecology, and pediatrics. The design is reasonable and can meet the basic learning needs of residents in various specialties. In the 1990s, DxR Clinician was developed by DxR Development Group, Inc. [7] in the United States, and it was later introduced and translated into Chinese by the China Union Medical University Press in 2009 [13]. However, the use of this software in China is still in its infancy, with few relevant literature reports. A previous study showed that the cases included in the DxR software were virtual cases and were limited by the purchase [7]. Students needed to complete the study on the computer, and the average time for students to become familiar with the

XSL•FO
RenderX

software was as long as 2 hours. In contrast, the types of diseases and the number of cases covered in the XueYiKu app, as well as the ease of use, are advantages. In addition, the framework of this platform is reasonable, as new case types of diseases can be added according to students' needs, and the user interface is simple and fast. Therefore, this app is more suitable for residents' clinical thinking training in the global internet and postepidemic era.

There is a decline in the training and formation of clinical thinking in residents. Based on this, the real case base used in this app was designed to present 4 levels of difficulty for the same case. Evaluation and feedback in the process of clinical thinking training formation and refinement are also important. As a result, students were evaluated in multiple dimensions according to the learning situation. Analysis of the results showed that the learning process meets the clinical diagnosis and treatment reality. For first- to second-year residents, level 2 - 3 difficulty was the most appropriate, as illustrated by their better performance in knowledge expansion and logic, among other factors. The clinical learning process is in line with clinical practice and meets the learning needs of residents. Specific to the learning process at different levels of difficulty, the primary difficulty allowed for a complete diagnosis based on the completed clinical information, and as the difficulty level increased, there was an advancement of the learning model; the results of the 4 levels of difficulty were in line with the development of the resident's clinical thinking, realizing the intended purpose of the app. One similar app currently in use is Brianly, a medical tutoring and social networking platform that provides personalized medical learning tutoring through AI algorithms [14]. In Tsinghua University, "artificial intelligence + virtual reality" within Brianly is used for clinical training, which allows direct viewing of the anatomical details of a patient's real body structure in a virtual space, as well as virtual surgery. However, there is currently no learning software for medical students that is designed to match the growth and learning process, and that provides a difficulty gradient based on the clinical thinking ability. In this context, our current app will continue to develop, improve, and provide feedback related to the 4 levels of difficulty to be better used for the clinical thinking training of senior residents.

Based on current usage, the highest number of uses per capita was usually found during the admission and start of residency, that is, from September to February each year. The survey results show that students had a high level of satisfaction with the consultation aspect of the app. The case study on the app allows for both a generic history-taking process and a structured questioning for a particular disease, which is inextricably linked to the teaching case base from clinical consultations of real cases. However, there are some differences between the examination and the actual situation, and the contents of the specialist examination and the physical examination for a particular disease should be further optimized to be more concise and focused.

Strengths and Limitations

One of the greatest advantages of our app is the inclusion of real cases, which could provide a new model of clinical thinking training by building a system to comprehensively assess the "clinical competence" of residents at different stages. For students, the strongest advantage is that they can customize clinical competency training according to their needs without the limitations of time and space. For app designers, the XueYiKu app is inherently user-friendly and sustainable, which allows the addition of cases in bulk based on the existing framework.

In addition to the benefits of the app, there are several limitations that could be further improved. First, current cases and diseases could be further expanded to cover more subjects, which is the next step of development. Second, in terms of use, students currently use the app independently, while the frequency of use is mainly based on students' interests combined with required course study and clinical rotation. Third, current evaluation results are almost subjective evaluations by humans. Thus, the evaluation results may be biased.

Conclusion

The emergence of virtual network learning systems such as the XueYiKu app can not only rapidly cultivate students' ability to solve clinical problems but also save teaching costs, improve teaching quality and efficiency, and achieve the sharing of high-quality teaching resources. In addition to enriching the content of clinical teaching, students with diverse experience and innovative thinking are trained, which promotes the basic clinical knowledge of medical students and effectively cultivates their clinical thinking ability.

Acknowledgments

The authors would like to thank all participants for willingly sharing their experiences. This study was supported by grant #22-458 Development and implementation of an IT-based competency assessment system for Internal Medicine and Obstetrics & Gynecology resident training programs in China. The study funders had no rule in the study design, implementation, analysis, manuscript, preparation, or decision to submit this article for publication.

Data Availability

The datasets generated during and/or analyzed during this study are available from the corresponding author upon reasonable request.

XSL•FO
RenderX

Authors' Contributions

YW, NS, and JQ contributed equally to this study. YW and HW conceptualized the study and collected the data. HW and DZ jointly analyzed the data and drafted the article, which was later refined by YW, who also drafted and revised the paper. MW and HJ contributed to data analysis and offered comments. JH and NS provided critical comments and reviewed the important intellectual content. JQ contributed to conception and design as well as the final approval of the version to be published as the corresponding author.

Conflicts of Interest

None declared.

References

1. Richards JB, Hayes MM, Schwartzstein RM. Teaching clinical reasoning and critical thinking: from cognitive theory to practical application. Chest 2020 Oct;158(4):1617-1628. [doi: 10.1016/j.chest.2020.05.525] [Medline: 32450242]
2. Reinus JF. Patient-data management and clinical reasoning: neglected essential skills. Am J Med 2022 Dec;135(12):1393-1394. [doi: 10.1016/j.amjmed.2022.08.022] [Medline: 36063858]
3. Karunaratne D, Sibbald M, Chandratilake M. Understanding cultural dynamics shaping clinical reasoning skills: A dialogical exploration. Med Educ 2024 Aug 1. [doi: 10.1111/medu.15479] [Medline: 39089689]
4. Locke R, Mason A, Coles C, Lusznat RM, Masding MG. The development of clinical thinking in trainee physicians: the educator perspective. BMC Med Educ 2020 Jul 16;20(1):226. [doi: 10.1186/s12909-020-02138-w] [Medline: 32678045]
5. Lee DS, Abdullah KL, Subramanian P, Bachmann RT, Ong SL. An integrated review of the correlation between critical thinking ability and clinical decision-making in nursing. J Clin Nurs 2017 Dec;26(23-24):4065-4079. [doi: 10.1111/jocn.13901] [Medline: 28557238]
6. Carin L. On artificial intelligence and deep learning within medical education. Acad Med 2020 Nov;95(11S):S10-S11. [doi: 10.1097/ACM.0000000000003630] [Medline: 32769462]
7. Myers JH, Dorsey JK. Using diagnostic reasoning (DxR) to teach and evaluate clinical reasoning skills. Acad Med 1994 May;69(5):428-429. [doi: 10.1097/00001888-199405000-00052] [Medline: 8086079]
8. Xu X, Liu F. Optimization of online education and teaching evaluation system based on GA-BP neural network. Comput Intell Neurosci 2021;2021:8785127. [doi: 10.1155/2021/8785127] [Medline: 34422036]
9. Nagi F, Salih R, Alzubaidi M, et al. Applications of artificial intelligence (AI) in medical education: a scoping review. Stud Health Technol Inform 2023 Jun 29;305:648-651. [doi: 10.3233/SHTI230581] [Medline: 37387115]
10. Hsiao CT, Chou FC, Hsieh CC, Chang LC, Hsu CM. Developing a competency-based learning and assessment system for residency training: analysis study of user requirements and acceptance. J Med Internet Res 2020 Apr 14;22(4):e15655. [doi: 10.2196/15655] [Medline: 32286233]
11. Webster CS. Artificial intelligence and the adoption of new technology in medical education. Med Educ 2021 Jan;55(1):6-7. [doi: 10.1111/medu.14409] [Medline: 33155316]
12. Wang M, Sun Z, Jia M, et al. Intelligent virtual case learning system based on real medical records and natural language processing. BMC Med Inform Decis Mak 2022 Mar 4;22(1):60. [doi: 10.1186/s12911-022-01797-7] [Medline: 35246134]
13. Meng Q, Zhaoxia T. Promoting effect of information products on case-based clinical teaching model. J Med Inform 2012(8):89-92. [doi: 10.3969/j.issn.1673-6036.2012.08.024]
14. Han ER, Yeo S, Kim MJ, Lee YH, Park KH, Roh H. Medical education trends for future physicians in the era of advanced technology and artificial intelligence: an integrative review. BMC Med Educ 2019 Dec 11;19(1):460. [doi: 10.1186/s12909-019-1891-5] [Medline: 31829208]

Abbreviations

AI: artificial intelligence
CBLAS: competency-based learning and assessment system
DxR: diagnostic reasoning

Edited by A Mavragani; submitted 26.04.24; peer-reviewed by D Hu, Q Xia; revised version received 23.09.24; accepted 24.09.24; published 03.01.25.

Please cite as:
Wang H, Zheng D, Wang M, Ji H, Han J, Wang Y, Shen N, Qiao J
Artificial Intelligence–Powered Training Database for Clinical Thinking: App Development Study
JMIR Form Res 2025;9:e58426
URL: https://formative.jmir.org/2025/1/e58426
doi:10.2196/58426

XSL•FO
RenderX

© Heng Wang, Danni Zheng, Mengying Wang, Hong Ji, Jiangli Han, Yan Wang, Ning Shen, Jie Qiao. Originally published in JMIR Formative Research (https://formative.jmir.org), 3.1.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Physical Activity at Work (PAW) Program in Thai Office Workers: Mixed Methods Process Evaluation Study

Katika Akksilp[1,2], MD, PhD; Thomas Rouyard[3], PhD; Wanrudee Isaranuwatchai[2], PhD; Ryota Nakamura[4], PhD; Falk Müller-Riemenschneider[1,5,6], PhD; Yot Teerawattananon[1,2], PhD, MD; Cynthia Chen[1,5], PhD

[1]Health Intervention and Technology Assessment Programme, Ministry of Public Health, Nonthaburi, Thailand
[2]Saw Swee Hock School of Public Health, National University of Singapore and National University Health System, Singapore, Singapore
[3]Graduate School of Public Health & Health Policy, City University of New York, New York, NY, United States
[4]Hitotsubashi Institute for Advanced Study, Hitotsubashi University, Tokyo, Japan
[5]Yong Loo Lin School of Medicine, National University of Singapore and National University Health System, Singapore, Singapore
[6]Digital Health Center, Berlin Institute of Health, Charité-Universitätsmedizin Berlin, Berlin, Germany

Corresponding Author:
Katika Akksilp, MD, PhD
Health Intervention and Technology Assessment Programme
Ministry of Public Health
6th Floor, 6th Building, Department of Health
Tiwanon Rd., Muang
Nonthaburi, 11000
Thailand
Phone: 66 2 590 4549
Fax: 66 2 590 4369
Email: gochi.akk@gmail.com

Abstract

Background: An increasing number of multicomponent workplace interventions are being developed to reduce sedentary time and promote physical activity among office workers. The Physical Activity at Work (PAW) trial was one of these interventions, but it yielded an inconclusive effect on sedentary time after 6 months, with a low uptake of movement breaks, the main intervention component.

Objective: This study investigates the factors contributing to the outcomes of the PAW cluster randomized trial.

Methods: Following the Medical Research Council's guidance for process evaluation of complex interventions, we used a mixed methods study design to evaluate the PAW study's recruitment and context (how job nature and cluster recruitment affected movement break participation), implementation (dose and fidelity), and mechanisms of impact (assessing how intervention components affected movement break participation and identifying the facilitators and barriers to participation in the movement breaks). Data from accelerometers, pedometers, questionnaires, on-site monitoring, and focus group discussions were used for the evaluation. Linear mixed effects models were used to analyze the effects of different intervention components on the movement breaks. Subsequently, qualitative analysis of the focus group discussions provided additional insights into the relationship between the intervention components.

Results: The participation in movement breaks declined after the third week, averaging 12.7 sessions (SD 4.94) per participant per week for the first 3 weeks, and continuing to decrease throughout the intervention. On-site monitoring confirmed high implementation fidelity. Analysis of Fitbit data revealed that each additional movement break was associated with a reduction of 6.20 (95% CI 6.99-5.41) minutes in sedentary time and an increase of 245 (95% CI 222-267) steps. Regarding the mechanisms of impact, clusters with higher baseline sedentary time demonstrated greater participation in movement breaks, while those with frequent out-of-office duties showed minimal engagement. Moreover, clusters with enthusiastic and encouraging movement break leaders were associated with a 24.1% (95% CI 8.88%-39.4%) increase in participation. Environmental and organizational support components using posters and leaders' messages were ineffective, showing no significant change in percentage participation in movement breaks (4.49%, 95% CI –0.49% to 9.47% and 1.82%, 95% CI –2.25% to 5.9%, respectively). Barriers such as high workloads and meetings further hindered participation, while the facilitators included participants' motivation to feel active and the perceived health benefits from movement breaks.

XSL•FO
RenderX

Conclusions: Despite high fidelity, the PAW trial did not significantly reduce sedentary time, with limited uptake of movement breaks due to context-related challenges, ineffective environmental support, and high workloads during the COVID-19 pandemic.

(JMIR Form Res 2025;9:e57604) doi:10.2196/57604

KEYWORDS

process evaluation; sedentary behavior; physical activity; workplace; movement breaks

Introduction

Background

Extensive research has examined workplace interventions to reduce sedentary behavior and promote physical activity globally [1,2]. Numerous interventions have yielded noteworthy results, showcasing their effectiveness, whereas some have encountered challenges in achieving desired outcomes [2-5]. While early meta-analyses, such as one from 1998, suggested minimal impact of workplace interventions on physical activity [6], subsequent reviews, including one from 2003, began to recognize their potential benefits [7]. However, more recent evidence has provided a clearer picture, indicating that more successful interventions tended to use pedometers, apply internet-based approaches, and include activities at social and environmental levels [5]. One systematic review found that physical environment interventions, whether implemented at the workstation, building, or neighborhood level, were effective in promoting physical activity in the workplace [8]. Finally, complex or multicomponent interventions emerge as the most effective workplace strategy for reducing sitting time [4], highlighting the intricate process leading to successful behavioral changes [9].

However, multicomponent intervention trial reports often fail to elucidate the rationale and mechanisms behind their results, leaving readers with questions regarding the efficacy of the interventions. Many scholars have criticized these trials as being akin to a “black box” because the underlying reasons for their success or failure remain unknown [10]. As a result, process evaluations, which assess how interventions were implemented in practice, are essential for deciphering the implications of the results from multicomponent intervention trials. These evaluations highlight aspects of the intervention that may require improvement to increase the probability of success [11,12].

Several frameworks advocate for process evaluations to explore the context, implementation, and impact mechanisms of intervention programs, encompassing aspects such as recruitment, reach, dose, fidelity, and challenges [10-13]. Process evaluations have become particularly relevant because they can help determine whether success or failure lies in the implementation, the intervention itself, or a combination of both factors. Notably, the Medical Research Council’s guidance on process evaluation of complex interventions is particularly valuable for assessing multicomponent interventions [11]. By contrast, frameworks like the one proposed by Steckler and Linnan [13] focus more on specific elements such as dose and fidelity but may not provide the broader perspective necessary to understand the interactions between context, mechanisms, and implementation. The emphasis of the Medical Research Council’s guidance on context and mechanisms was pivotal in capturing the socioecological dynamics within our intervention.

In addition, there is increasing acknowledgment that incorporating both qualitative and quantitative data, as well as using theoretical frameworks within process evaluations, plays a crucial role in promoting evidence-based practice [14]; for example, the Older People’s Exercise Intervention in Residential and Nursing Accommodation trial [15] used a mixed methods process evaluation to assess whether the multicomponent intervention changed residents’ home culture to increase physical activity and whether residents engaged in exercise activities. The study found no notable cultural shifts or sufficient engagement, thus explaining the trial’s null results [15,16]. Another process evaluation explored the implementation and impact mechanisms of a park prescription intervention trial, revealing key mediators of the intervention effects, such as park physical activity levels, as well as barriers that may have weakened intervention effectiveness [17].

Objectives

In Thailand, it has been reported that the majority of adults who engage in sufficient physical activity reside in rural areas, typically due to their work in the agricultural sector, whereas physical activity levels tend to be lower among office workers [18]. Numerous nonresearch initiatives have been launched to promote physical activity and reduce sedentary behavior among office workers in Thai companies and organizations. The Physical Activity at Work (PAW) trial [19] marked the first comprehensive cluster randomized trial aimed at promoting physical activity and reducing sedentary behavior among Thai office workers. This trial incorporated a multicomponent intervention and used accelerometer-measured data, laying a solid foundation for physical activity research in Thailand [20]. However, the trial produced inconclusive findings, leaving critical questions regarding the factors contributing to its outcomes unanswered. We observed a suboptimal uptake of movement breaks, the primary intervention component, and inferred that this, coupled with a low recruitment rate, may explain the absence of statistically significant outcomes [20]. Understanding these underlying reasons is essential for making necessary improvements in both research and policy. To address this, we conducted a process evaluation to investigate the factors contributing to the trial’s outcome.

Following the Medical Research Council’s guidance for process evaluation of complex interventions [11], we used a mixed methods study design to evaluate the PAW study’s recruitment and context, implementation, and impact mechanisms. Specifically, we examined how job nature and recruitment processes within each cluster influenced participation in movement breaks (context), the overall dose and fidelity of the

XSL•FO
RenderX

intervention (implementation), and the effects of intervention components on participation in movement breaks as well as the facilitators and barriers to movement break participation (impact mechanisms).

Methods

Description of the Cluster Randomized Trial

The PAW study was conducted in a free-living setting at the Ministry of Public Health offices in Nonthaburi, Thailand. Detailed methodological information regarding the cluster randomized trial (Thai Clinical Trials Registry: TCTR20200604007) was published in the protocol [19] and main results [20] manuscripts. Briefly, between July and September 2021, we recruited 282 office workers (age: mean 38.6, SD 10.4, years; female: n=228, 80.9%) from 18 separate offices. After baseline data collection, the participants were randomized into 9 intervention offices (140/282, 49.6%) and 9 control offices (142/282, 50.4%). The 6-month intervention, which took place between September 2021 and March 2022, consisted of 6 components across 4 levels, as outlined in Textbox 1.

Textbox 1. Key elements of the Physical Activity at Work cluster randomized trial.

Individual-level components

- Providing a wearable device with real-time activity feedback (Fitbit Inspire HR; Google LLC)
- Using a Fitbit smartphone app
- Offering individual weekly lottery-based financial incentives, where 1 intervention group participant who participated in at least 70% of the movement breaks within the previous week was randomly selected to receive a reward of THB 500 (US $16)

Social-level components

- Team movement breaks of light to moderate intensity, lasting at least 4 minutes and occurring 4 times a day (alarm clocks with speakers were provided to movement break leaders to initiate sessions), which served as the primary intervention component
- Team-based incentives of an additional weekly lottery reward of THB 500 (US $16) given to the winner if at least 70% of the participants in the cluster also attended at least 70% of the movement breaks within the previous week; 4 alarm reminders were set (9:30 AM, 10:30 AM, 2:30 PM, and 3:30 PM), and trained movement break leaders managed the starting times, songs, and movements; participants working from home were encouraged to join the sessions via web conferencing

Environmental-level component

- Three types of posters providing information on health risks associated with high sedentary time and the benefits of physical activity, as well as examples of stretching exercises

Organizational-level component

- Leadership support in the form of messages sent by office directors twice a week via Line (LY Corporation) to encourage participants to reduce sedentary time using movement breaks and increase physical activity, as well as to announce reward winners with congratulatory messages and photographs of the reward ceremony

The multicomponent intervention was developed using the socioecological model [9,21]. The intervention components were strategically designed to complement one another, with a particular focus on the primary component: the movement breaks. In addition, department directors actively encouraged participants to attend more movement break sessions and announced the weekly lottery reward, which was contingent upon the frequency of participation in the movement breaks. Two movement break leaders per office were trained to oversee these sessions. Moreover, posters were displayed to inform participants about the adverse effects of prolonged sedentary behavior and to provide examples of strategies to break up sedentary time in the office. All intervention participants were equipped with a Fitbit Inspire HR smartwatch (Google LLC), preset to remind them to interrupt prolonged sitting every 30 minutes. Consequently, the intervention operated as illustrated in Figure 1.

We formulated our research questions to encompass the context, implementation, and impact mechanisms, with a specific emphasis on the primary intervention component: the movement breaks. This emphasis stemmed from two key considerations: (1) as previously noted, our intervention design places the movement break as the central intervention component, supported by other components; and (2) we observed a low participation rate (median 31.5%, IQR 20.4%-42.7%) in the movement break sessions. Moreover, 77.5% (107/138) of the intervention participants attended fewer than half of the breaks (240/480, 50%) throughout the intervention period. We hypothesized that the absence of a significant intervention effect could be attributed to low attendance in the movement break sessions [20]. Hence, our focus is to investigate the underlying reasons behind this issue. Table 1 describes the process evaluation components, specific questions, and the data sources.

XSL•FO
RenderX

Figure 1. The mechanism of associations between the Physical Activity at Work cluster randomized trial intervention components and outcomes.

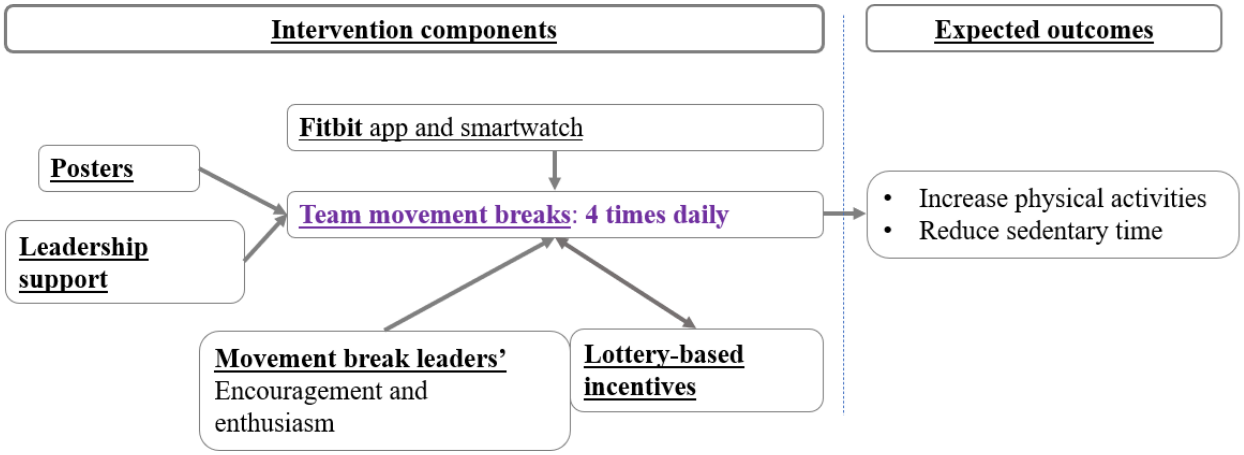

Table 1. Descriptions of the Physical Activity at Work process evaluation components, specific questions, and the data sources.

Process evaluation components	Specific process evaluation questions	Quantitative data	Qualitative data
Recruitment and context			
Recruitment	How does recruitment influence participation in movement breaks?	Recruitment rate	Recruitment fashion
Context	What impact do job descriptions and baseline characteristics of each cluster have on movement break participation?	Baseline summary statistics of each cluster	Job description of each cluster
Implementation			
Dose	What was the extent of intervention delivery?	Web monitoring: Fitbit wear time and conducts of movement breaks	On-site monitoring
Fidelity	Was the intervention delivered as intended?	Associations between participation in movement breaks and daily Fitbit sedentary time data as well as step counts during the intervention period	On-site monitoring
Mechanisms of impact			
Testing the assumptions behind the intervention design	How did supporting components influence participation in movement breaks?	Effects of intervention components on the percentage of weekly movement break participation	Focus group discussions
Participants' attitude	What were the facilitators and barriers to participation in movement breaks?	Additional questionnaire regarding facilitators and barriers to participation in movement breaks	Focus group discussions

Ethical Considerations

Ethics approval for both the quantitative and qualitative components of this study was granted by the Ethical Review Committee for Research in Human Subjects, Ministry of Public Health, Thailand (IRB00001629), in accordance with relevant ethical guidelines for human research. Any modifications to the approved protocol will be submitted to the ethics committee for review. All participants provided written informed consent before participation, which included consent for the use of deidentified data in secondary analyses and for the publication of results. To protect participants' privacy and confidentiality, all data were deidentified, and no personally identifiable information was included in the final dataset, with all published data presented in aggregate form. Participants' data (eg, case record forms, laboratory test results, information sheets, and consent forms) are stored in a locked cabinet in a researcher's office. All data will be destroyed by researchers within 5 years of publication. During the study, only deidentified data were used, and the data were only accessible to the research team. Participants received compensation of THB 250 (approximately US $7.5) for each data collection session in which they participated.

Dissemination Policy and Authorship Guidelines

In addition to disseminating our research findings to the funder of this study, the Ministry of Public Health, we will disseminate our findings to other countries, the study participants, and the research community. We have followed the authorship guidelines of the International Committee of Medical Journal Editing.

Data Collection

Baseline and Follow-Up Data Collection (Questionnaire)

An interviewer-administered questionnaire, based on the Thai National Statistical Office's health survey and capturing

XSL•FO
RenderX

sociodemographic data such as age and education [22], was used to collect participant data at both baseline and follow-up. In addition, intervention participants were asked supplementary questions regarding the implementation of the intervention during the follow-up data collection session (Multimedia Appendix 1).

Data Monitoring (Movement Break Schedule)

A data monitoring team, consisting of the trial implementer and an administrator, conducted weekly random field visits to 4 offices during the scheduled movement break sessions, adhering to the initial timetable. Monitored data included (1) the occurrence of sessions, (2) session quality, (3) participant attendance, and (4) any issues encountered during the session.

Movement break leaders were responsible for submitting weekly schedules via Microsoft Excel Online, indicating session timings and dates. Updates were required if a session did not fall within the 30-minute alarm windows (9:30 AM, 10:30 AM, 2:30 PM, and 3:30 PM), indicating absence or starting times outside the specified windows.

Fitbit Data and Movement Break Participation

During the 6-month intervention period, we collected Fitbit data, which included wear time, sedentary time, and step counts. Daily activity data were anonymized and used to compile weekly lists of potential reward recipients and to assess compliance with movement breaks. We determined participation in each movement break session by (1) analyzing a quadratic function of steps, (2) detecting quadratic functions within the 1-hour time frame of movement breaks, and (3) ensuring that participants had accumulated at least 100 steps.

Quantitative Data Analysis

Linear mixed models were used to examine the associations between intervention components and activity outcomes, as well as between the intervention components themselves. These models incorporated random intercepts from both individual and cluster levels, while also accounting for random slope by including intervention weeks. Weekly individual data were derived from daily Fitbit records, follow-up questionnaire responses, and monitoring data (Table 2). Data were analyzed using RStudio (version 4.0.3; Posit Software, PBC) and Stata software (version 14.2; StataCorp LLC), with significance level set at 5%.

XSL•FO
RenderX

Table 2. Variables used in the Physical Activity at Work process evaluation.

Variable	Explanation	Scale
Associations between movement break participation and daily Fitbit sedentary time and step count data during the intervention period		
Outcome variables		
Sedentary time	Daily sedentary time data from Fitbit	Continuous (min)
Steps	Daily step count data from Fitbit	Continuous (step count)
Exposure variables		
Movement break participation	Daily movement break participation	Ordinal; 1-4 (sessions)
Effects of intervention components on weekly movement break participation percentage		
Outcome variable		
Movement break participation percentage	The frequency of an individual's participation in movement break sessions within a specific week compared to the maximum number of sessions available during that week	Continuous (percentage)
Exposure variables		
Individual reward	We designated the weeks following individual reward wins as {1}, while all other weeks were coded as {0}, using the web monitoring data	1=won a reward last week; 0=others
Cluster reward	We assigned {1} to the weeks following office colleagues' reward wins and {0} to all other weeks, using the web monitoring data	1=a colleague in the same office cluster won a reward last week; 0=others
Fitbit wear time	We categorized the average weekly Fitbit wear time for each individual into a binary variable using the median	1=wore Fitbit at or above the median wear time; 0=wore Fitbit below the median wear time
Fitbit wear time (alternative)	How often Fitbit led individuals to engage in physical activity in the office in the last 2 weeks of the intervention	1=at least 4 d/wk; 0=<4 d/wk
Leadership support	How often the directors' support led individuals to engage in physical activity in the office in the last 2 weeks of the intervention	1=at least 4 d/wk; 0=<4 d/wk
Movement break leaders	How the movement break leaders' (1) encouragement and (2) enthusiasm contributed to movement break participation (We asked the movement break leaders and used the data as cluster-representative data)	3=both factors contributed *somewhat* or *a lot*; 2=only the encouragement contributed *somewhat* or *a lot*; 1=only the enthusiasm contributed *somewhat* or *a lot*; 0=both factors contributed *not at all*, *very little*, or *a little*
Posters	How often the posters led individuals to engage in physical activity in the office in the last 2 weeks of the intervention	1=at least 4 d/wk; 0=<4 d/wk
Posters (alternative)	How accurate individuals were in reporting the number of different poster designs in their office	1=3-4 designs; 0=0-2 designs

Focus Group Discussions

The qualitative aspect of the study involved focus group discussions aimed at exploring participants' perspectives on facilitators and barriers to engagement in movement breaks, as well as their attitudes toward intervention components. Intervention clusters were initially ranked by mean percentage participation in movement breaks. Subsequently, using purposive sampling, we invited up to 6 participants with the highest percentage participation from the top 2 clusters and up to 6 participants with the lowest percentage participation from the bottom 2 clusters for each of the 4 focus group discussions.

The 4 focus group discussions took place via 45- to 75-minute Zoom (Zoom Video Communications, Inc) meetings, with participants joining from their respective offices. The focus group discussions were recorded with participants' consent (both video and audio were captured), while notes were taken during the sessions. Before each focus group discussion, participants were briefed about its purpose, format, and estimated duration, with the option to interrupt as needed.

Verbatim transcriptions were manually completed by hired research assistants. The transcriptions were then subjected to deductive thematic analysis, with facilitator and barrier serving as overarching themes, and the socioecological model used as subtheme [9]. Two analysts (KA and WI) independently coded each transcript using the pre-established framework. References under themes and subthemes were compared and discussed to ensure consistency and accuracy. Intercoder reliability was not analyzed in this study.

An interview guide (Table S1 in Multimedia Appendix 2) was developed based on the research question "What are the facilitators and barriers to movement break participation?" The socioecological model informed this guide. A total of 4 interviewers (female: n=2, 50%) conducted the focus group discussions. They comprised the PAW trial implementer (first author; doctor of medicine), 2 study administrators (bachelor

XSL•FO
RenderX

of communication arts and master of political science), and a research assistant (bachelor of clinical pharmacology) who was not a trial staff member. All interviewers underwent a comprehensive 2-day training program tailored to the qualitative data collection and analysis requirements of the study. This training was led by a Thai senior researcher (PhD in anthropology). The focus group discussion guide served as a reference tool, ensuring that critical topics were not overlooked. During the focus group discussions, the researcher flexibly adjusted the sequence, content, and style based on individual responses. Emotions and nonverbal cues expressed by participants were carefully documented.

Code names were used in place of real names in the recorded data to protect participant privacy. Data collection and analysis proceeded concurrently throughout the study. After coding the transcripts from the 4 focus group discussions, the researchers and supervisors deliberated on data saturation. Subsequently, 2 additional focus group discussions were conducted with participants from clusters ranked in the top and bottom thirds, following the same procedures. No new themes emerged from these final 2 focus group discussions, indicating that data saturation had been achieved.

Results

Recruitment and Context

The PAW study recruited participants from all clusters within the Department of Medical Services building and the International Health Policy Program, Ministry of Public Health, Thailand, between July and September 2020. The ministry's governance structure is typically bureaucratic, with each department led by a single director general and each office headed by an office director. In total, 18 clusters were successfully recruited, with 15 (83%) falling under the Department of Medical Services purview; the remaining 3 (17%) clusters were under the jurisdiction of the Office of the Permanent Secretary (n=2, 67% were part of the International Health Policy Program and located in a distinct building complex; Figure S1 in Multimedia Appendix 2).

Recruitment procedures at the cluster level proceeded seamlessly. The directors of each office endorsed the active involvement of their respective teams in the trial, resulting in a cluster-level recruitment rate of 100% (18/18). For individual-level recruitment, we initiated the process by organizing office-specific group meetings to thoroughly explain the trial details. After these meetings, office workers interested in participating could engage with the trial staff to complete the individual informed consent process. To accommodate those unable to attend the scheduled meetings due to prior commitments, additional sessions were arranged.

Despite the bureaucratic governance style, the individual recruitment rate reached 62.9% (282/449). Job descriptions varied and included research-related roles and academic, finance, law, digital, and other administrative positions. At baseline, 3 (17%) of the 18 clusters exhibited a mean daily sedentary time exceeding 9 hours (mean sedentary time of 546, SD 90.3 min). Moreover, mean daily moderate to vigorous physical activity time was generally higher than the current physical activity guideline, reflecting that participants were relatively active. No significant associations were observed among job descriptions, baseline time spent in sedentary or moderate to vigorous physical activity, and monitoring data. However, clusters with higher baseline sedentary time demonstrated increased participation in movement breaks (Figure S2 in Multimedia Appendix 2). Furthermore, mean age and cluster size seemed to influence participation in movement breaks within the intervention group. Notably, individuals in the cluster with the youngest and smallest number of participants showed minimal participation in movement breaks (Table 3).

Table 3. Recruitment and context of intervention participants and the implementation of the Physical Activity at Work intervention (n=140).

	Baseline (n=140)			Data monitoring (6-mo intervention period) (n=138)			
Job description	Cluster size, n (%)	Recruitment rate n/N (%)	Sedentary time (min/d), mean (SD)	Lottery-incentive wins, n (%)	Movement break participation (%), mean (SD)	Maximum work from home, n/N (%)	Notes
Research related	23 (16.4	23/25 (92)	541 (90.5)	15 (63)	59.7 (14.7)	2.5/5 (50)	__[a]
Research related	8 (5.7)	8/44 (18)	483 (99.4)	0 (0)	25.8 (13.2)	5/5 (100)	Different building and department
Nursing	13 (9.3)	13/37 (35)	463 (117)	4 (17)	38.8 (16.6)	0/5 (0)	Different department
Finance	10 (7.1)	10/11 (90)	460 (95.9)	0 (0)	32.9 (5.49)	2/5 (40)	Broadcast from another cluster
Finance	14 (10.0)	14/17 (82)	481 (94.1)	1 (4.17)	29 (16.4)	2/5 (40)	—
Human resource	34 (24.3)	34/35 (97)	473 (113)	2 (8)	31.5 (11)	2.5/5 (50)	—
Human resource	18 (12.9)	18/31 (58)	394 (141)	1 (4)	25 (7.99)	2.5/5 (50)	Broadcast from another cluster
Digital	15 (10.7)	15/25 (60)	447 (116)	1 (4)	19.7 (9.83)	2/5 (40)	—
Inspection	5 (3.6)	5/18 (27)	350 (121)	0 (0)	3.88 (4.54)	2/5 (40)	n=3 (50%) participants at follow-up

[a]Not applicable.

XSL•FO
RenderX

Implementation

Fidelity: Was the Intervention Delivered as Intended?

On-Site Monitoring

The on-site monitoring team conducted 65 field visits to the intervention group participants' offices during their scheduled movement break sessions. During these visits, 35 movement breaks were observed, with 28 (80%) sessions featuring the Department of Medical Services theme song alongside another preferred Thai song. Importantly, there were no instances of cheating, such as merely shaking the Fitbit device without engaging in actual physical activity. However, 1 (8%) of the 13 participants from cluster 13 chose to take a 10-minute walk outside the office instead of participating in the team movement breaks, even when the sessions were not prompted. This participant managed to secure the weekly reward 3 times during the 24-week intervention period. Finally, no participants from the control group attended the sessions.

All alarm clocks were operational; nevertheless, leaders opted to broadcast the movement break songs using alternative devices, including the built-in broadcast system and office speakers. Notably, 2 (22%) of the 9 clusters did not independently initiate their movement break sessions; instead, they waited for another cluster to start the activity because they shared the same broadcast systems.

The movement break schedule reports submitted by the leaders were accurate; nonleaders who independently initiated sessions assisted with the reporting process. This collaboration ensured that the recorded data accurately reflected actual session initiation and participation. This also supports the finding that in some of the clusters (3/9, 33%), participants took the initiative to lead sessions without leader involvement, demonstrating shared responsibility in maintaining intervention fidelity.

Weekly rewards were distributed to winners by department directors. Participants received the team-based incentive on only 2 (8%) out of 24 occasions. A photograph capturing the reward ceremony, featuring face-to-face reward distribution, was shared with all intervention participants through Line OpenChat.

We received notifications regarding Fitbit issues, including syncing, freezing, and charging problems, both during and outside of our field visits. Fortunately, some participants with technological expertise were able to provide assistance, alleviating the need for the implementation team to respond to every notification. During our observations, we noticed that some of the participants (9/138, 6.5%) occasionally left their Fitbit devices at home when attending the movement break sessions. Despite understanding that their participation would not be recorded without the Fitbit, they still chose to engage in the sessions. In addition, 1 (0.7%) of the 138 participants was observed wearing the Fitbit only during the movement breaks.

All posters remained undamaged and visible, without any changes to their placement.

Due to the COVID-19 pandemic, a work-from-home policy was implemented. However, fewer than half of the participants (49/138, 35.5%) worked from home, with only 1 (11%) of the 9 clusters having a 100% work-from-home arrangement. This cluster used Zoom for meetings to conduct their movement breaks. Upon visiting the Zoom link twice, we found only 1 (0.7%) of the 138 participants in the Zoom meeting, with no movement breaks initiated. By contrast, another cluster coordinated their participation through a Line group chat when working from home, ensuring simultaneous session participation without requiring an online meeting.

Online Monitoring

We examined the correlations between engaging in movement breaks and daily Fitbit sedentary time and step count data. Each additional movement break was associated with a reduction of 6.20 (95% CI 6.99-5.41) minutes in sedentary time and an increase of 245 (95% CI 222-267) steps, adjusting for Fitbit wear time, number of public holidays in that week, and participants' age, sex, and education, with cluster and ID as random intercepts and intervention week number as a random slope.

This analysis, based on pedometer-measured outcomes, supported the fidelity of the movement break implementation because our prescribed minimum requirement for the duration of a single session was 4 minutes. Furthermore, the cumulative steps surpassed the eligibility criterion of 100 steps. The analysis also suggests successful data synthesis regarding movement break participation (Table 4).

Table 4. Associations between movement break participation on daily Fitbit sedentary time and step counts during the intervention period in the Physical Activity at Work cluster randomized trial.

	Model A		Model B	
	β[a] (95% CI)	*P* value	β[b] (95% CI)	*P* value
Sedentary time (min)	–6.32 (–7.06 to –5.57)	<.001	–6.20 (–6.99 to –5.41)	<.001
Steps (count)	263 (242 to 283)	<.001	245 (222 to 267)	<.001

[a]Linear mixed-effect model adjusted for Fitbit wear time, with cluster and ID as random intercepts and intervention week number as the random slope.
[b]Further adjusted for the number of public holidays in that week, age, gender, and education of the participants.

XSL•FO
RenderX

Dose: What Was the Extent of Intervention Delivery? (Online Monitoring)

According to the movement break leaders' weekly reports, many movement break sessions were never initiated by the leaders in each cluster. By the third week of the intervention, approximately 40% (72/180) of the scheduled movement break sessions had not been initiated (Figure S3 in Multimedia Appendix 2). Similarly, the overall average participation in movement breaks declined after the third week of the intervention, reducing to an average of 8 sessions per week per participant. Subsequently, the participation rate continued to decrease (Figure S4 in Multimedia Appendix 2).

We plotted the average daily Fitbit wear time to assess adherence during the intervention period (Figure S5 in Multimedia Appendix 2). We found that participants typically wore the devices for 10 to 15 hours per day at the start, with a slight decrease to 8 to 14 hours occurring 2 to 3 months into the intervention.

Mechanisms of Impact

Overview

Building on Figure 1, we present the effects of each component derived from a mixed methods analysis aimed at exploring the associations between each intervention component and movement break participation. We combine results from both quantitative and qualitative analyses to comprehensively address the question.

Quantitative Analysis

A total of 3200 participant-week data points were extracted by combining participants' demographics, Fitbit wear time, sedentary time and step count data, answers from the attitude-toward-intervention-components questionnaire, and online monitoring of rewards (Table 2).

Qualitative Analysis

Overview

There were 28 participants (n=3, 11% to n=6, 21% in each session) across 6 focus group discussions. The primary data analysts for the main results included 4 interviewers: the head of the data collection team (male), a program administrator (female), a researcher (male), and a staff member from the Health Intervention and Technology Assessment Program (a semiautonomous research unit under the Ministry of Public Health; female). Demographic details of all participants from the clusters (coded aby C12, C13, C5, C8, C16, and C17) involved in the focus group discussions are presented in Table S2 in Multimedia Appendix 2. Generally, the mean age of participants from the 3 top-ranked clusters was higher than that of participants from the bottom-ranked clusters. The majority of the participants (≥80%; 23/28, 82%) were female. Notably, the cluster with the lowest participation rate of 2.1% (10/480) comprised only 3 (2.2%) of the 138 participants at the 6-month follow-up (Table S2 in Multimedia Appendix 2).

How Did the Supporting Components Influence Participation in Movement Breaks?

Lottery-Based Incentives

Individuals who won the weekly lottery rewards the previous week demonstrated an 8.64% (95% CI 0.985%-16.3%) increase in movement break participation compared to other data points (Table 5). However, it is crucial to note the possibility of an overestimation when comparing data from winners against those from nonwinners. To address this, we conducted another subgroup analysis exclusively including winners, revealing a 5.1% (95% CI –3.44% to –13.6%) increase in movement break participation. However, this increase lacked statistical significance due to the small sample size (Table S3 in Multimedia Appendix 2).

Table 5. Effects of the Physical Activity at Work intervention components on weekly movement break participation percentage.

	β[a] (%; 95% CI)	Standardized β (%; 95% CI)	*P* value
Individual reward	8.64 (.985 to 16.3)	.0257 (.00293 to .0484)	.03
Cluster reward	–.325 (–2.64 to 1.99)	–.004 (–.0325 to .0245)	.78
Fitbit wear time	3.96 (2.28 to 5.65)	.0696 (.04 to .0993)	.001
Leadership support	1.82 (–2.25 to 5.9)	.0309 (–.0381 to .1)	.38
Enthusiastic	2 (–17.2 to 21.2)	0.077 (–.663 to .817)	.84
Encouraging	1.49 (–18.8 to 21.8)	0.0576 (–.724 to .84)	.89
Both enthusiastic and encouraging	24.1 (8.96 to 39.2)	0.929 (.346 to 1.51)	.002
Posters	4.49 (–.493 to 9.47)	0.0695 (–.00764 to .147)	.08

[a]Linear mixed effects model adjusted for the previous-week movement break participation percentage, number of public holidays in that week, age, sex, and education of the participants, with cluster and ID as random intercepts and intervention week number as the random slope.

During the focus group discussions, some of the participants suggested that rather than a single winner receiving a substantial cash prize, there should be multiple winners receiving more affordable rewards:

> *Rather than cash, it should be acknowledgments, such as showing who reach this many steps so we can compete with each other.* [C13]

> *[C]ould be something cheaper, don't have to be cash. Cheap shirts will do!* [C10]

XSL•FO
RenderX

> *[I]t was indeed a motivating factor, but the conditions for obtaining it should be somewhat more lenient. That is, to make it accessible to a broader audience, but the current conditions seem to be quite high. If it were reduced a bit, say to 100, I believe that providing rewards would be motivating enough.* [C8]

A participant in the worst-performing cluster did not remember the details of the financial incentive:

> *Sorry but how much was the reward?* [C17]

Team-Based Incentives

We observed no difference in movement break participation percentage (–0.325%, 95% CI –2.64% to 1.99%) in the weeks after a colleague from the same office won lottery rewards compared to other weeks. Contrary to our assumptions during the intervention development phase, peer pressure seems to have had no discernible effects on motivating movement break participation, as indicated by the quantitative analysis model (Table 5).

Different ideas emerged during the focus group discussions, as outlined in the following paragraphs.

Peer support was mentioned among members in the best-performing cluster as an important motivator:

> *Because it helped the team. If we dance, someone gets 1000THB, if we don't, it's 500THB only.* [C12]

However, others expressed discouragement:

> *When I see others got it and I never won for weeks, I was disheartened.* [C12]

Some of the participants viewed rewards as not motivating:

> *I joined the sessions because I want to be healthier. Rewards are meaningless for me.* [C13]

Fitbit Wear Time

There was a 3.96% (95% CI 2.28%-5.65%) increase in movement break participation among individuals with higher Fitbit wear time compared to those with lower wear time, using the median weekly wear time as the cutoff (Table 5). However, this increase may be attributed to motivation triggered by Fitbit notifications (eg, reminders to break prolonged sitting) or potentially to an information bias, wherein participants who wore their Fitbit devices for longer periods were detected more frequently in movement break participation.

To address potential bias, we compared the result with another model where Fitbit exposure was based on self-report, using the question "How often did you look at the Fitbit tracker during the last 2 weeks of the intervention period?" The analysis revealed that looking at the tracker more frequently (≥4 times/wk) was associated with a 1.97% (95% CI –1.42% to 5.36%) higher movement break participation percentage, without statistical significance (Table S4 in Multimedia Appendix 2).

Qualitative analysis indicated that participants perceived Fitbit as beneficial not necessarily for encouraging additional movement breaks but rather for motivating increased exercise and providing real-time feedback:

> *I joined short breaks a lot at first, but after a while, I forgot...However, I always wear the watch; look at the daily data.* [C13]
>
> *I set my goal to walk 10,000 steps a day after I got the watch, and I succeeded!* [C16]
>
> *[S]aying Fitbit encourages more movement breaks is wrong for me, but I feel the urge to exercise more from wearing the watch, like running after work.* [C17]

Leadership Support

Participants perceived the encouragement from directors as ineffective in motivating them to participate in more movement breaks.

Although office directors joined very few movement breaks, participants believed that their presence helped motivate everyone in the cluster:

> *He joined, I saw. But mostly he's busy.* [C16]
>
> *It made us stand up and dance...if we didn't it'd be awkward.* [C5]
>
> *We all danced every time our director was present (laugh).* [C8]

With regard to the movement break leaders' encouragement and enthusiasm, clusters with leaders who self-evaluated as more encouraging and enthusiastic had a significant increase of 24.1% (95% CI 8.88%-39.4%) in movement break participation compared to clusters with less encouraging and less enthusiastic leaders. However, because these exposure variables are self-reported by cluster leaders, the detected difference may be influenced by other cluster-specific factors.

Participants believed that movement break leaders' enthusiasm and encouragement helped them join more break sessions:

> *He was very active and always encourages everyone to stand up and dance.* [C8]
>
> *We were aware of the scheduled time, but if we were occupied, the leader would notify us.* [C13]

Nevertheless, some of the participants mentioned that they did not rely on leaders:

> *No matter how many people in the office, we danced. no leaders, no problem at all.* [C12]
>
> *I'm not sure...leaders always initiated the activity, but we did not always join.* [C16]
>
> *No one was available.* [C17]

Posters

Participants who reported that posters motivated them to engage in more movement breaks during the last 2 weeks of the intervention exhibited approximately a 5% higher participation rate, although this increase lacked statistical significance (Table 5). Notably, this increase may indicate participants' attitudes toward the intervention component rather than the direct effects of the posters themselves. Therefore, we compared the result by converting the variable to whether participants accurately identified the number of different styles of posters in their offices. The analysis revealed no significant difference in

XSL•FO
RenderX

movement break adherence between those who answered correctly and those who did not (standardized β=1.94, 95% CI –1.34 to 5.23; Table S5 in Multimedia Appendix 2).

Qualitative analysis showed that while the posters initially captured interest, over time, they failed to sustain attention:

> *At first I read them. I thought it was helpful and tried to follow some moves. After a while I just ignored them.* [C2]
>
> *I didn't really read it that much...just walk past.* [C3]

The standardized β coefficients of all exposures indicate that the self-evaluated enthusiasm and encouragement of movement break leaders were the most important variables among those included, followed by Fitbit wear time and the individual reward (Table 5).

Facilitators and Barriers to Movement Break Participation

Figure 2 illustrates participants' attitudes regarding the facilitators and barriers to movement break participation. The top 2 facilitators were the positive feelings associated with being physically active and the perceived health benefits, followed by encouragement from leaders, directors, and colleagues. The appeal of engaging in enjoyable exercises also motivated their participation. Interestingly, weekly rewards ranked lowest among the facilitators.

Figure 2. Facilitators and barriers to movement break participation in the Physical Activity at Work cluster randomized trial.

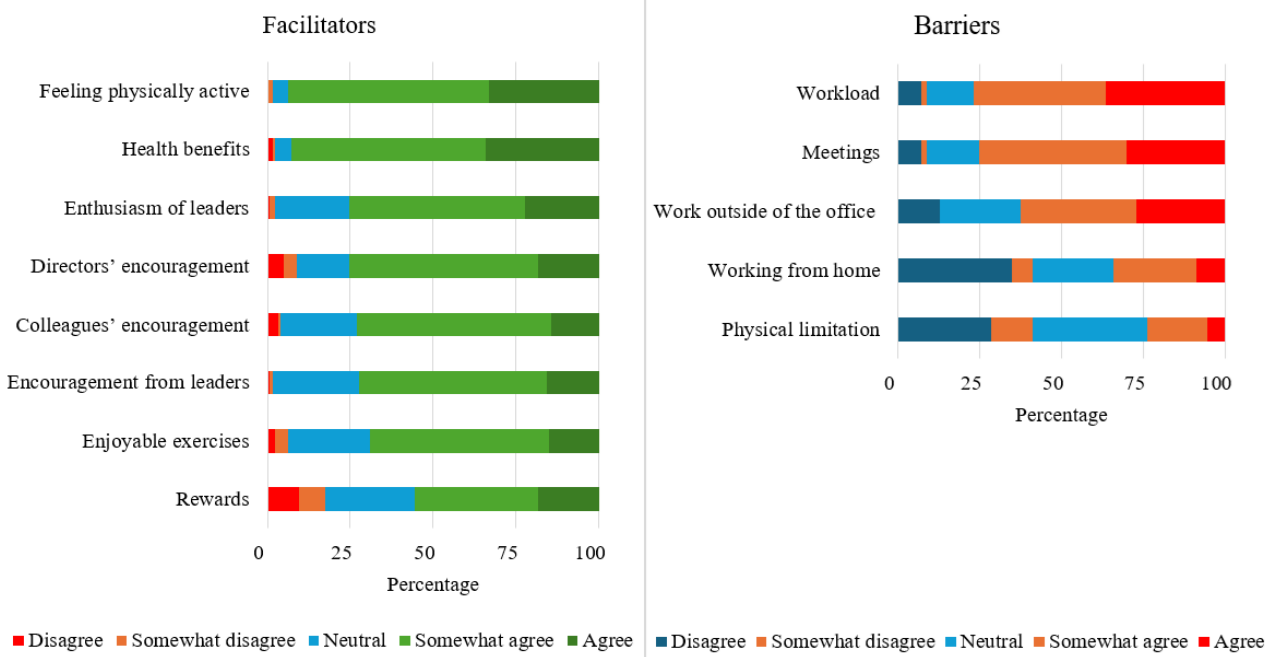

Conversely, the primary barrier perceived by participants was their workload, followed by meetings and working outside of the office. Surprisingly, many participants did not view working from home as a barrier to their participation.

Thematic analysis revealed positive attitudes toward intervention components:

> *I think having Fitbit is a really good motivation to move.* [C17]
>
> *Previously when someone invited me to run, I wanted to but it was hard. This [movement breaks] is easier and fun.* [C12]

Participants also noted barriers to the intervention, such as the lottery reward design:

> *It [individual reward] somehow motivated me, but I guess I just could not do as good as others.* [C16]

Another important barrier was the monotonous design of the movement breaks:

> *Dancing with the same moves gets boring after a while.* [C8]
>
> *I think we can change the songs to make it more interesting.* [C16]

Table 6 presents thematic analysis results with references under each theme and subtheme from all 6 focus group discussions.

XSL•FO
RenderX

Table 6. Thematic analysis of focus group discussions on the facilitators and barriers to movement break participation in the Physical Activity at Work cluster randomized trial.

Themes and subthemes	Quotes
Facilitators	
Psychological capability	• "Previously when someone invited me to run, I wanted to, but it was hard. This is easier and fun." (C12)
Social opportunity	• "Both songs and moves...it was like mass hysteria; others enjoyed the sessions and I wanted to join" (C12) • "Our director joined, and we danced happily" (C8) • "When my boss dance and I just sat there working, it felt strange, so I joined" (C5)
Physical opportunity	• "I'm happy that there is a new opportunity for me to exercise...in workplace. Normally I don't have time." (C12) • "Our office is spacious. There are empty spaces for dancing." (C5)
Reflective motivation	• "My personal goal was actually health, not rewards" (C13) • "I set my goal to get the reward and lose weight and cholesterol level because the programme measured those" (C5)
Automatic Motivation	• "It was relaxing, both mind and body...especially the mind." (C17) • "Just that we got to dance...when the session started, it relieved the stress quite a lot" (C5)
Barriers	
Psychological capability	• "10 min. dancing is too short. 30 min. working out is better for your health for the day." (C13) • "My programming work never allows me to lose focus" (C16)
Social opportunity	• "Meetings definitely prevented us to join movement sessions" (C8) • "We were not really motivated by directors and colleagues" (C17)
Physical opportunity	• "Sometimes workspace is not wide enough...also I was afraid I'll annoy others who were not in the project" (C13) • "During high workload, we cannot join" (C5) • "The high workload never allows us to do anything else" (C17) • "Ever since COVID situation got worse, I've been sitting all the time at my desk. It's the workload, can't do anything else." (C16)
Reflective motivation	• "Rewards could be anything, like shoes, incentives don't need to be money, like shirts..." (C5)
Automatic motivation	• "Each programme shouldn't last long. I mean, we should always change the stimulant to avoid boredom..." (C13) • "Dancing with the same moves gets boring after a while" (C8) • "I think we can change the songs to make it more interesting." (C16)

Discussion

The PAW cluster randomized trial of a multicomponent intervention, developed based on the socioecological framework [9], showed a decrease in participants' waking sedentary time and an increase in moderate to vigorous physical activity, although without statistical significance [20]. We conducted a mixed methods process evaluation, following the Medical Research Council's guidance [11], to comprehensively describe the PAW trial's (1) recruitment and context, (2) implementation, and (3) impact mechanisms.

Recruitment and Context

Our team anticipated a high recruitment rate once the office directors approved the inclusion of their offices in the trial, given the bureaucratic nature of the ministry organization in Thailand. However, recruitment rates were low in many clusters, weakening the trial's statistical power and risking reduced intervention adherence, particularly in the social-level component.

Cluster 17 serves as a notable example, demonstrating minimal engagement in movement breaks (Table 3; Figure S4 in Multimedia Appendix 2). This could be attributed to the recruitment of too few, relatively young participants alongside a high number of nonparticipants (Tables 3 and 4). Despite its potential for easy qualification for the team-based incentive with just 3 participants (at follow-up), cluster 17 might have faced challenges due to the negative influence of an environment where breaks are discouraged. This aligns with a previous study reporting reduced break taking in disapproving work environments compared to more supportive ones [23]. Cluster 13 also faced challenges with a low recruitment rate; yet, the engagement of the participants in this cluster in movement breaks was high. Two reasons might explain this: (1) the participant who chose to walk outside the office and secured 3 weekly rewards increased the cluster's average participation

XSL•FO
RenderX

rate, and (2) recruiting at least 13 participants could create an environment conducive to making movement breaks feel enjoyable and secure to participate in.

Job characteristics might also play a role in movement break participation. Cluster 17, “the inspection office,” frequently required its members to leave the office for inspections, which was reflected in their low sedentary time at baseline (Table 3) and echoed in the focus group interviews. In addition, cluster 16 represented a digital office where most participants worked on their laptop computers all day. While this setting might seem ideal for implementing the intervention, participants encountered challenges with breaks because they needed continual focus to code. This aligns with the suggestion that identifying the optimal fit between the organizational context and the intervention is crucial [24]. However, challenges remain when implementing such interventions across organizations with diverse offices, each characterized by unique job contexts and work styles.

Implementation

We implemented the trial with a comprehensive dose delivery, ensuring intervention fidelity by closely adhering to the original plan for each component. Moreover, the findings in Table 3 affirm the application of precise criteria for enumerating movement breaks, underscoring their correlations with the expected reductions in sedentary time and increases in step counts. Nevertheless, we encountered challenges related to the frequency of movement break sessions conducted by participants, which we refer to as “dose received” [13]. This inconsistency may be attributed to the performance of the movement break leaders, as illustrated in Figure S3 in Multimedia Appendix 2, which shows that many sessions were never initiated. However, several clusters relied on automatic timing mechanisms to initiate breaks, ensuring that sessions commenced even without leaders present. We discuss this further in the next subsection.

Mechanisms of Impact

Testing the Designed Intervention Components

Healthy behaviors are maximized when environments and policies support healthful choices, and individuals are motivated and educated to make them [25]. As one of the main principles of the ecological model, the interaction of influences means that variables within the system work together synergistically [26]. Hence, our study examined the impact of the supporting components on the movement break participation using multivariate linear mixed models (Table 4).

The enthusiasm and encouragement of the movement break leaders seem to have contributed to higher participation percentages. However, the cluster that exhibited the best performance challenged this observation, asserting that the members initiated sessions independently, even without leaders. Unlike other cluster randomized trials where exercise sessions were led by nonparticipants such as physiotherapists [16], the movement break leaders in our study were participants and could be replaced if absent. Hence, without the leaders, other members could initiate sessions automatically. Nevertheless, the influence of the leaders’ encouragement and enthusiasm likely played a role in motivating the rest of the team. Prior research has also found that workplace team leaders play a significant role in facilitating the implementation of workplace interventions [27,28]. Moreover, a comprehensive review emphasizes that effective team performance underscores the fulfillment of leadership styles, supportive team behaviors, communication, and performance feedback [29].

The impact of the team-based incentive on movement break participation was evidently negligible. Regarding individual lottery-based incentives, their influence remains somewhat unclear. Systematic reviews suggest that financial rewards for physical activity have positive short-term effects, surpassing unconditional incentives [30,31]. Moreover, another study indicates that increasing reward values may lead to improved results [32]. Nevertheless, insights from online monitoring data and focus group discussions revealed that only 1 (11%) of the 9 clusters actively pursued rewards and engaged in friendly competition, while the others (8/9, 89%) were indifferent. Awarding only 1 winner per week might have demotivated participants who faithfully adhered to the intervention but never won. This phenomenon can be explained using goal-setting theory, which suggests that setting clear and appropriately challenging goals is crucial [33]. By contrast, setting unachievable goals may induce stress, anxiety, and perceived pressure [34]. Therefore, overly difficult goals, such as securing the weekly lottery reward, may have discouraged participants.

Fitbit was perceived as a helpful tool for real-time data monitoring and might help motivate leisure physical activity. However, current evidence indicates unfavorable outcomes regarding the effectiveness of pedometers in increasing physical activity within the workplace or in motivating sedentary breaks [35,36]. The component may have primarily served as a data collection tool rather than actively supporting movement breaks. By contrast, the posters, which served as the environmental component, proved ineffective and should be replaced with physical environment interventions, which have been shown to be more effective [4,8].

Finally, leadership support was considered helpful at the start of the intervention but provided no lasting effect over time. The component could be perceived as a nudge, which has been widely used to prompt behavioral change among participants by providing alternative options to sedentary habits in the workplace [37]. Leadership support was expected to facilitate these behavioral shifts. Management support has also been found to be an essential enabler for workplace intervention participation to reduce sitting time [27,38,39]. However, the sustainability of the effects depends on the meticulous design of the intervention component. Future studies evaluating long-term effects must enhance the intervention component to ensure sustainability.

Facilitators and Barriers

Automatic motivation, driven by the desire to feel active and relaxed, was ranked as the top facilitator of movement break participation, followed by reflective motivation, including perceived health benefits. The findings are in line with a recent systematic review [39]. Another review also supports the idea that microbreaks at work can improve well-being by boosting vigor and reducing fatigue [40]. In addition, participants

XSL•FO
RenderX

generally grasped the concept of health benefits associated with movement breaks. However, some of the individuals compared movement breaks to aerobic exercise, expressing doubt regarding their health benefits.

Workload and meetings were perceived as the main barriers to movement breaks, which aligns with a previous study's prediction that high workloads, although positively related to the desire to detach from work, would also deter employees from actually taking breaks [23]. In addition, recent findings from a systematic review supported the notion that microbreaks increase well-being but do not necessarily improve work performance, especially in tasks that require high cognitive engagement. Moreover, the review suggested that breaks of >10 minutes may be necessary to enhance work performance [40].

By contrast, meetings presented clear barriers to our intervention design. Future research should be dedicated to advocating feasible and context-specific active meetings within an active workplace; for example, an exploratory study found that standing meetings were feasible, well-received by employees, and may reduce sitting time among the population [41]. However, widespread adoption faces obstacles due to prevailing sedentary work cultures and concerns about self-consciousness in front of senior staff, highlighting the need for broader social behavior change initiatives [42].

We hypothesized that movement break participation might be low due to the work arrangement during the peaks of the COVID-19 pandemic [20]. However, participants did not consider working from home to be the main barrier to movement break participation. Instead, they thought that the COVID-19 pandemic did not significantly hinder their adherence to the intervention due to the work-from-home policy. Instead, they attributed the difficulty to increased workload. Nevertheless, working from home remained 1 of the barriers to movement break participation. This is in line with another workplace cluster randomized trial, which reported that engaging in physical exercise with colleagues during working hours was more effective than home-based exercise in enhancing vitality and managing pain-related concerns among health care workers [43].

Strengths and Limitations

We conducted a comprehensive mixed methods process evaluation of the PAW multicomponent intervention, covering recruitment and context, implementation, and impact mechanisms. Rigorous analyses were applied to both quantitative and qualitative data. The data are prospective, spanning 6 months of follow-up. Although the overall results show no significant impact of the intervention, this process evaluation offers insights that could be crucial for the future development and evaluation of intervention packages aimed at reducing sedentary time while improving physical activity levels.

Nevertheless, our study had some limitations. First, we could not test mediators to understand the underlying mechanisms through which one variable influences another due to the trial's lack of efficacy [20]. Second, constructing intervention theories will involve reconstructing a logic model [11,44], a task we plan to undertake in future studies. Third, we could not complete the management and analysis of process evaluation data before the conclusion of the PAW trial. As a result, the process evaluation was conducted post hoc to elucidate the trial outcomes. Furthermore, our evaluation only incorporated participants' data and perspectives, overlooking input from other stakeholders such as organizational directors and nonparticipants. Finally, the team members responsible for process evaluation were also involved in the outcome evaluation. While our team members possess the most comprehensive understanding of the trial details, potential bias in interpretations must be acknowledged [11].

Conclusions

The PAW trial did not significantly reduce sedentary time among Thai office workers. Although the trial implementation was satisfactory regarding dose delivery and fidelity, there was limited uptake of the movement breaks, the key intervention component. This limited uptake could be attributed to (1) context-related challenges (including jobs requiring high cognitive engagement or frequent out-of-office work and meetings), (2) the absence of goal-setting aspects in the detailed design of individual and social components, (3) the lack of effective and sustainable supporting components at the environmental and organizational levels, and (4) elevated workloads in specific clusters (exacerbated during peak periods of the COVID-19 pandemic) serving as a significant barrier.

Acknowledgments

The authors thank Buppa Pardang, Pimonphan Thongoun, Saudamini Dabak, Kewalin Chomrenoo, Nachawish Kittibovorndit, and Budsadee Soboon for their help with the focus group discussions, further helpful discussions, and administrative support that contributed to this work. The trial was funded by revenue from taxes on tobacco and alcohol via the Thai Health Promotion Foundation. This study was supported by the Singapore Ministry of Health's National Medical Research Council under its Population Health Research Grant (PHRGOC22Jul-0020, ID: MOH-001315). The Health Intervention and Technology Assessment Program's International Unit is supported by the International Decision Support Initiative to provide technical assistance on health intervention and technology assessment to governments in low- and middle-income countries. RN and TR are supported by the Japan Society for the Promotion of Science Core-to-Core Program (JPJSCCB20200002). The funders had no role in the study design, data collection or analysis, the preparation of the manuscript, or the decision to publish. The views and opinions expressed in this paper are those of the authors and do not necessarily reflect those of the funders or the institutions to which the authors belong.

XSL•FO
RenderX

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

All authors contributed to the study design or delivery of the trial. KA was the principal investigator of the trial. FM-R and CC provided statistical expertise and data management. TR, RN, YT, WI, and FM-R provided expertise on conceptualizing the process evaluation aspects and the narrative of the manuscript. RN and TR provided expertise on behavioral economics, such as the impact of lottery-based and team-based incentives. TR and FM-R provided expertise on physical activity promotion theories and frameworks. KA, YT, and CC drafted the manuscript. All authors reviewed the manuscript draft and read and approved the final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional questions for intervention participants.
[DOCX File , 740 KB - formative_v9i1e57604_app1.docx]

Multimedia Appendix 2
Supplementary figures and tables.
[DOCX File , 15 KB - formative_v9i1e57604_app2.docx]

References

1. Buckingham SA, Williams AJ, Morrissey K, Price L, Harrison J. Mobile health interventions to promote physical activity and reduce sedentary behaviour in the workplace: a systematic review. Digit Health 2019 Mar 27;5:2055207619839883 [FREE Full text] [doi: 10.1177/2055207619839883] [Medline: 30944728]
2. Zhou L, Deng X, Xu M, Wu Y, Shang X, E F, et al. The effects of active workstations on reducing work-specific sedentary time in office workers: a network meta-analysis of 23 randomized controlled trials. Int J Behav Nutr Phys Act 2023 Jul 27;20(1):92 [FREE Full text] [doi: 10.1186/s12966-023-01467-5] [Medline: 37501138]
3. Mulchandani R, Chandrasekaran AM, Shivashankar R, Kondal D, Agrawal A, Panniyammakal J, et al. Effect of workplace physical activity interventions on the cardio-metabolic health of working adults: systematic review and meta-analysis. Int J Behav Nutr Phys Act 2019 Dec 19;16(1):134 [FREE Full text] [doi: 10.1186/s12966-019-0896-0] [Medline: 31856826]
4. Chu AH, Ng SH, Tan CS, Win AM, Koh D, Müller-Riemenschneider F. A systematic review and meta-analysis of workplace intervention strategies to reduce sedentary time in white-collar workers. Obes Rev 2016 May;17(5):467-481. [doi: 10.1111/obr.12388] [Medline: 26990220]
5. To QG, Chen TT, Magnussen CG, To KG. Workplace physical activity interventions: a systematic review. Am J Health Promot 2013;27(6):e113-e123. [doi: 10.4278/ajhp.120425-LIT-222] [Medline: 23631453]
6. Dishman RK, Oldenburg B, O'Neal H, Shephard RJ. Worksite physical activity interventions. Am J Prev Med 1998 Nov;15(4):344-361. [doi: 10.1016/s0749-3797(98)00077-4] [Medline: 9838977]
7. Proper KI, Koning M, van der Beek AJ, Hildebrandt VH, Bosscher RJ, van Mechelen W. The effectiveness of worksite physical activity programs on physical activity, physical fitness, and health. Clin J Sport Med 2003 Mar;13(2):106-117. [doi: 10.1097/00042752-200303000-00008] [Medline: 12629429]
8. Zhu X, Yoshikawa A, Qiu L, Lu Z, Lee C, Ory M. Healthy workplaces, active employees: a systematic literature review on impacts of workplace environments on employees' physical activity and sedentary behavior. Built Environ 2020 Jan;168:106455. [doi: 10.1016/j.buildenv.2019.106455]
9. Sallis J, Owen N, Fisher EB. Ecological models of health behavior. In: Glanz BK, K V, editors. Health Behavior and Health Education: Theory, Research, and Practice. Hoboken, NJ: Jossey-Bass; 2008.
10. Grant A, Treweek S, Dreischulte T, Foy R, Guthrie B. Process evaluations for cluster-randomised trials of complex interventions: a proposed framework for design and reporting. Trials 2013 Jan 12;14:15 [FREE Full text] [doi: 10.1186/1745-6215-14-15] [Medline: 23311722]
11. Moore GF, Audrey S, Barker M, Bond L, Bonell C, Hardeman W, et al. Process evaluation of complex interventions: medical research council guidance. BMJ 2015 Mar 19;350:h1258 [FREE Full text] [doi: 10.1136/bmj.h1258] [Medline: 25791983]
12. Oakley A, Strange V, Bonell C, Allen E, Stephenson J, RIPPLE Study Team. Process evaluation in randomised controlled trials of complex interventions. BMJ 2006 Feb 18;332(7538):413-416 [FREE Full text] [doi: 10.1136/bmj.332.7538.413] [Medline: 16484270]

XSL•FO
RenderX

13. Steckler A, Linnan L. Process Evaluation for Public Health Interventions and Research. San Francisco, CA: Jossey-Bass; 2022.
14. Liu H, Andersson G, Manchaiah V. Editorial: the process evaluation of clinical trials. Front Med (Lausanne) 2022 Jul 11;9:950637 [FREE Full text] [doi: 10.3389/fmed.2022.950637] [Medline: 35899205]
15. Ellard DR, Taylor SJ, Parsons S, Thorogood M. The OPERA trial: a protocol for the process evaluation of a randomised trial of an exercise intervention for older people in residential and nursing accommodation. Trials 2011 Feb 02;12:28 [FREE Full text] [doi: 10.1186/1745-6215-12-28] [Medline: 21288341]
16. Ellard DR, Thorogood M, Underwood M, Seale C, Taylor SJ. Whole home exercise intervention for depression in older care home residents (the OPERA study): a process evaluation. BMC Med 2014 Jan 03;12:1 [FREE Full text] [doi: 10.1186/1741-7015-12-1] [Medline: 24386900]
17. Petrunoff N, Yao J, Sia A, Ng A, Ramiah A, Wong M, et al. Activity in nature mediates a park prescription intervention's effects on physical activity, park use and quality of life: a mixed-methods process evaluation. BMC Public Health 2021 Jan 22;21(1):204 [FREE Full text] [doi: 10.1186/s12889-021-10177-1] [Medline: 33482787]
18. Tuangratananon T. Differences in physical activity levels between urban and rural adults in Thailand: findings from the 2015 National Health and Welfare Survey. Health Syst Res 2018;12(2018):27-41 [FREE Full text]
19. Chen C, Dieterich AV, Koh JJ, Akksilp K, Tong EH, Budtarad N, et al. The physical activity at work (PAW) study protocol: a cluster randomised trial of a multicomponent short-break intervention to reduce sitting time and increase physical activity among office workers in Thailand. BMC Public Health 2020 Sep 01;20(1):1332 [FREE Full text] [doi: 10.1186/s12889-020-09427-5] [Medline: 32873258]
20. Akksilp K, Koh JJ, Tan V, Tong EH, Budtarad N, Xueying G, et al. The physical activity at work (PAW) study: a cluster randomised trial of a multicomponent short-break intervention to reduce sitting time and increase physical activity among office workers in Thailand. Lancet Reg Health Southeast Asia 2022 Oct 19;8:100086 [FREE Full text] [doi: 10.1016/j.lansea.2022.100086] [Medline: 37384135]
21. Owen N, Sugiyama T, Eakin EE, Gardiner PA, Tremblay MS, Sallis JF. Adults' sedentary behavior determinants and interventions. Am J Prev Med 2011 Aug;41(2):189-196. [doi: 10.1016/j.amepre.2011.05.013] [Medline: 21767727]
22. The Health and Welfare Survey 2015. National Statistical Office of Thailand. 2023. URL: https://www.nso.go.th/nsoweb/storage/survey_detail/2023/20230505165728_20473.pdf [accessed 2024-12-17]
23. Phan V, Beck JW. Why do people (not) take breaks? An investigation of individuals' reasons for taking and for not taking breaks at work. J Bus Psychol 2023;38(2):259-282 [FREE Full text] [doi: 10.1007/s10869-022-09866-4] [Medline: 36593864]
24. Nielsen K, Taris TW, Cox T. The future of organizational interventions: addressing the challenges of today's organizations. Work Stress 2010 Oct 20;24(3):219-233. [doi: 10.1080/02678373.2010.519176]
25. Ottawa charter for health promotion. World Health Organization. 2012 Jun 16. URL: https://www.who.int/publications/i/item/WH-1987 [accessed 2024-10-31]
26. Glanz K, Viswanath K, Rimer BK. Health Behavior and Health Education: Theory, Research, and Practice. Hoboken, NJ: John Wiley & Sons; 2008.
27. Goode AD, Hadgraft NT, Neuhaus M, Healy GN. Perceptions of an online 'train-the-champion' approach to increase workplace movement. Health Promot Int 2019 Dec 01;34(6):1179-1190. [doi: 10.1093/heapro/day092] [Medline: 30452649]
28. Hopkins JM, Glenn BA, Cole BL, McCarthy W, Yancey A. Implementing organizational physical activity and healthy eating strategies on paid time: process evaluation of the UCLA WORKING pilot study. Health Educ Res 2012 Jun;27(3):385-398 [FREE Full text] [doi: 10.1093/her/cys010] [Medline: 22323279]
29. Salcinovic B, Drew M, Dijkstra P, Waddington G, Serpell BG. Factors influencing team performance: what can support teams in high-performance sport learn from other industries? A systematic scoping review. Sports Med Open 2022 Feb 22;8(1):25 [FREE Full text] [doi: 10.1186/s40798-021-00406-7] [Medline: 35192078]
30. Barte JC, Wendel-Vos GC. A systematic review of financial incentives for physical activity: the effects on physical activity and related outcomes. Behav Med 2017;43(2):79-90. [doi: 10.1080/08964289.2015.1074880] [Medline: 26431076]
31. Mitchell MS, Goodman JM, Alter DA, John LK, Oh PI, Pakosh MT, et al. Financial incentives for exercise adherence in adults: systematic review and meta-analysis. Am J Prev Med 2013 Nov;45(5):658-667. [doi: 10.1016/j.amepre.2013.06.017] [Medline: 24139781]
32. Finkelstein EA, Linnan LA, Tate DF, Birken BE. A pilot study testing the effect of different levels of financial incentives on weight loss among overweight employees. J Occup Environ Med 2007 Sep;49(9):981-989. [doi: 10.1097/JOM.0b013e31813c6dcb] [Medline: 17848854]
33. Latham GP, Locke EA. Self-regulation through goal setting. Organ Behav Hum Decis Process 1991 Dec;50(2):212-247. [doi: 10.1016/0749-5978(91)90021-k]
34. Latham GP, Locke EA. Enhancing the benefits and overcoming the pitfalls of goal setting. Organ Dyn 2006;35(4):332-340. [doi: 10.1016/j.orgdyn.2006.08.008]
35. Harrison J, O'Donoghue A, Hill J. The use of pedometers in the workplace to increase physical activity. Br J Card Nurs 2021 Aug 02;16(8):1-4 [FREE Full text] [doi: 10.12968/bjca.2021.0066] [Medline: 38125904]

XSL•FO
RenderX

36. Freak-Poli R, Cumpston M, Albarqouni L, Clemes SA, Peeters A. Workplace pedometer interventions for increasing physical activity. Cochrane Database Syst Rev 2020 Jul 21;7(7):CD009209. [doi: 10.1002/14651858.CD009209.pub3] [Medline: 32700325]
37. Forberger S, Wichmann F, Comito CN. Nudges used to promote physical activity and to reduce sedentary behaviour in the workplace: results of a scoping review. Prev Med 2022 Feb;155:106922. [doi: 10.1016/j.ypmed.2021.106922] [Medline: 34933021]
38. Mackenzie K, Such E, Norman P, Goyder E. Sitting less at work: a qualitative study of barriers and enablers in organisations of different size and sector. BMC Public Health 2019 Jul 04;19(1):884 [FREE Full text] [doi: 10.1186/s12889-019-7148-8] [Medline: 31272484]
39. Mackenzie K, Such E, Norman P, Goyder E. The development, implementation and evaluation of interventions to reduce workplace sitting: a qualitative systematic review and evidence-based operational framework. BMC Public Health 2018 Jul 04;18(1):833 [FREE Full text] [doi: 10.1186/s12889-018-5768-z] [Medline: 29973188]
40. Albulescu P, Macsinga I, Rusu A, Sulea C, Bodnaru A, Tulbure BT. "Give me a break!" A systematic review and meta-analysis on the efficacy of micro-breaks for increasing well-being and performance. PLoS One 2022 Aug 31;17(8):e0272460 [FREE Full text] [doi: 10.1371/journal.pone.0272460] [Medline: 36044424]
41. Danquah IH, Tolstrup JS. Standing meetings are feasible and effective in reducing sitting time among office workers-walking meetings are not: mixed-methods results on the feasibility and effectiveness of active meetings based on data from the "take a stand!" study. Int J Environ Res Public Health 2020 Mar 05;17(5):1713 [FREE Full text] [doi: 10.3390/ijerph17051713] [Medline: 32151044]
42. Biddle SJ, O'Connell SE, Davies MJ, Dunstan D, Edwardson CL, Esliger DW, et al. Reducing sitting at work: process evaluation of the SMArT Work (Stand More At Work) intervention. Trials 2020 May 13;21(1):403 [FREE Full text] [doi: 10.1186/s13063-020-04300-7] [Medline: 32404181]
43. Jakobsen MD, Sundstrup E, Brandt M, Andersen LL. Psychosocial benefits of workplace physical exercise: cluster randomized controlled trial. BMC Public Health 2017 Oct 10;17(1):798 [FREE Full text] [doi: 10.1186/s12889-017-4728-3] [Medline: 29017479]
44. Rush B, Ogborne A. Program logic models: expanding their role and structure for program planning and evaluation. Can J Program Eval 1991;6(2):95-106. [doi: 10.3138/cjpe.6.005]

Edited by A Mavragani; submitted 23.02.24; peer-reviewed by E Breckin; comments to author 09.09.24; revised version received 29.09.24; accepted 10.10.24; published 02.01.25.

Please cite as:
Akksilp K, Rouyard T, Isaranuwatchai W, Nakamura R, Müller-Riemenschneider F, Teerawattananon Y, Chen C
The Physical Activity at Work (PAW) Program in Thai Office Workers: Mixed Methods Process Evaluation Study
JMIR Form Res 2025;9:e57604
URL: https://formative.jmir.org/2025/1/e57604
doi:10.2196/57604
PMID:

©Katika Akksilp, Thomas Rouyard, Wanrudee Isaranuwatchai, Ryota Nakamura, Falk Müller-Riemenschneider, Yot Teerawattananon, Cynthia Chen. Originally published in JMIR Formative Research (https://formative.jmir.org), 02.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

ChatGPT's Attitude, Knowledge, and Clinical Application in Geriatrics Practice and Education: Exploratory Observational Study

Huai Yong Cheng[1], MPH, MS, MD

Minneapolis VA Health Care System, Minneapolis, MN, United States

Corresponding Author:
Huai Yong Cheng, MPH, MS, MD
Minneapolis VA Health Care System
1 Veterans Dr.
Minneapolis, MN, 55417
United States
Phone: 1 6124672051
Fax: 1 6127252084
Email: wchengwcheng@gmail.com

Abstract

Background: The increasing use of ChatGPT in clinical practice and medical education necessitates the evaluation of its reliability, particularly in geriatrics.

Objective: This study aimed to evaluate ChatGPT's trustworthiness in geriatrics through 3 distinct approaches: evaluating ChatGPT's geriatrics attitude, knowledge, and clinical application with 2 vignettes of geriatric syndromes (polypharmacy and falls).

Methods: We used the validated University of California, Los Angeles, geriatrics attitude and knowledge instruments to evaluate ChatGPT's geriatrics attitude and knowledge and compare its performance with that of medical students, residents, and geriatrics fellows from reported results in the literature. We also evaluated ChatGPT's application to 2 vignettes of geriatric syndromes (polypharmacy and falls).

Results: The mean total score on geriatrics attitude of ChatGPT was significantly lower than that of trainees (medical students, internal medicine residents, and geriatric medicine fellows; 2.7 vs 3.7 on a scale from 1-5; 1=*strongly disagree*; 5=*strongly agree*). The mean subscore on positive geriatrics attitude of ChatGPT was higher than that of the trainees (medical students, internal medicine residents, and neurologists; 4.1 vs 3.7 on a scale from 1 to 5 where a higher score means a more positive attitude toward older adults). The mean subscore on negative geriatrics attitude of ChatGPT was lower than that of the trainees and neurologists (1.8 vs 2.8 on a scale from 1 to 5 where a lower subscore means a less negative attitude toward aging). On the University of California, Los Angeles geriatrics knowledge test, ChatGPT outperformed all medical students, internal medicine residents, and geriatric medicine fellows from validated studies (14.7 vs 11.3 with a score range of –18 to +18 where +18 means that all questions were answered correctly). Regarding the polypharmacy vignette, ChatGPT not only demonstrated solid knowledge of potentially inappropriate medications but also accurately identified 7 common potentially inappropriate medications and 5 drug-drug and 3 drug-disease interactions. However, ChatGPT missed 5 drug-disease and 1 drug-drug interaction and produced 2 hallucinations. Regarding the fall vignette, ChatGPT answered 3 of 5 pretests correctly and 2 of 5 pretests partially correctly, identified 6 categories of fall risks, followed fall guidelines correctly, listed 6 key physical examinations, and recommended 6 categories of fall prevention methods.

Conclusions: This study suggests that ChatGPT can be a valuable supplemental tool in geriatrics, offering reliable information with less age bias, robust geriatrics knowledge, and comprehensive recommendations for managing 2 common geriatric syndromes (polypharmacy and falls) that are consistent with evidence from guidelines, systematic reviews, and other types of studies. ChatGPT's potential as an educational and clinical resource could significantly benefit trainees, health care providers, and laypeople. Further research using GPT-4o, larger geriatrics question sets, and more geriatric syndromes is needed to expand and confirm these findings before adopting ChatGPT widely for geriatrics education and practice.

(JMIR Form Res 2025;9:e63494) doi:10.2196/63494

XSL•FO
RenderX

KEYWORDS

ChatGPT; geriatrics attitude; ageism; geriatrics competence; geriatric syndromes; polypharmacy; falls; aging, older adults

Introduction

Background

ChatGPT stands for Chat Generative Pretrained Transformer and was developed by an artificial intelligence (AI) research company, OpenAI. It is an AI chatbot technology that can process our natural human language and generate a response. ChatGPT was released to the public by OpenAI on November 30, 2022, and surpassed 1 million users in just 5 days. This performance set a record for the second fastest-growing user base after that of Threads. ChatGPT currently has >180 million users (as of October 26, 2024) [1]. ChatGPT users have increased exponentially, and it has been increasingly used in medicine, resulting in 4625 publications on PubMed (as of October 26, 2024) [2]. Notably, ChatGPT was not programmed originally to be used for medical education and clinical application [1] despite reports of its wide application for medical education [3-7] and clinical practice [5-12]. ChatGPT has been studied in multiple specialties and subspecialties, such as psychiatry [13], radiology [14], and cardiology [15]. Its rapid progress in such a brief time has raised challenges, concerns, and limitations, including privacy, ethnic and racial bias, and legal risks [16-21].

Any application of ChatGPT to medical education and clinical practice could be promising but needs to be well designed and investigated. In the context of a growing aging population, geriatrics education, geriatrics workforces, and geriatrics practice have been facing significant challenges for the last 4 decades [22-26]. ChatGPT could offer an opportunity to improve geriatrics education and clinical practice [27-30]. Preliminary studies have shown promise [27-30]. For example, older adults often have polypharmacy issues [31], and providers review and deprescribe medications [32]. A specifically trained large language model may provide useful clinical support in polypharmacy management for primary care physicians [29]. ChatGPT has scored better on general and theoretical questions in geriatrics than on complex decisions and end-of-life situations, with the lowest scores related to diagnosis and performing complex tests [30]. Overall, the application of ChatGPT to geriatrics education and clinical practice is sporadic and needs to be expanded and further investigated. Therefore, this study was designed to determine whether we can trust ChatGPT to be used in geriatrics education and clinical practice using 3 distinct approaches.

The first approach was to evaluate any age-biased outputs generated by ChatGPT. The aging narrative spans >210 years [33]. The term ageism, coined by Robert Butler in 1969, refers to age-based stereotyping, prejudice, and discrimination [34]. Self-perception of aging is another emergent concept related to ageism [35]. Ageism is global and prevalent in daily life [34,36,37], health care systems [34,38], and the news media [39,40]. It is present in various medicine specialties such as oncology [41] and cardiology [42] and was persistent during the COVID-19 pandemic [43,44]. Ageism is significantly associated with poor health outcomes and quality of life [34,35]. Interventions to reduce ageism have been widely studied [45,46]. ChatGPT has been found to exhibit privacy, ethnicity, gender, and racial bias and entail legal risks [16-21]. For example, ChatGPT recommends fewer female than male ophthalmologists [47]. However, whether ChatGPT generates age-biased outputs has not been reported. This study aimed to demonstrate whether ChatGPT exhibits ageism by testing its geriatrics attitude using a validated University of California, Los Angeles (UCLA) geriatrics attitude instrument [48,49] and comparing its results with those obtained by medical students residents, and geriatrics medicine fellows from published articles

The second approach was to evaluate the geriatrics knowledge of ChatGPT. ChatGPT has passed the United States Medical Licensing Examination (USMLE) [50,51] and many other medicine exams [52-56], demonstrating its competence in medical knowledge. However, it has been less studied in geriatrics. For example, in one Geriatrics and AI in Spain project in Spain [30], ChatGPT was prompted with 10 questions about geriatric medicine. Compared to 130 physicians who answered the questionnaire, ChatGPT scored better on general and theoretical questions than on complex decisions and end-of-life situations, with the lowest scores for diagnosis and performing complex tests [30]. AI is likely to be incorporated into some areas of geriatric medicine, but it still presents significant limitations, mainly in complex medical decision-making [30]. This study was designed to demonstrate the geriatrics competency of ChatGPT by examining its performance on the validated UCLA geriatrics knowledge test [48,49,57] and comparing it with that of medical students, internal medicine residents, and geriatric medicine fellows from previously published articles in the literature.

The third approach was to evaluate ChatGPT's knowledge of 2 geriatric syndromes and its clinical application to them (polypharmacy and falls). Multiple previous studies have shown that ChatGPT performs well on clinical vignette questions [58-64]. For example, ChatGPT achieved 71.7% accuracy overall on 36 clinical vignettes from the Merck Sharpe and Dohme clinical manuals and impressive accuracy in clinical decision-making [58]. In another study, GPT-4o was queried for diagnoses and management plans with 20 physician-written clinical vignettes in otolaryngology, showing high agreement with physicians [59]. In addition, ChatGPT demonstrated appreciable knowledge and interpretation skills in psychiatry through 100 clinical case vignettes [60]. In another study, 33 physicians across 17 specialties generated 284 medical questions. ChatGPT generated largely accurate information in response to diverse medical queries as judged by academic physician specialists, with improvement over time [61]. ChatGPT also offered accurate recommendations on managing hypertension based on clinical practice guidelines [62]. ChatGPT provided accurate recommendations for 8 out of 10 clinical scenarios in cardiology [63]. In another study, GPT-4o generated a good rehabilitation plan for a stroke case from a textbook [64]. Finally, several studies have reported the application of

XSL•FO
RenderX

ChatGPT to predict common drug-drug interactions [65], managing polypharmacy [29], clinical pharmacy [66-68], and medical pharmacology as a self-learning tool [69]. Taken together, all previous studies suggest that ChatGPT could potentially be applied to complex geriatric syndrome vignettes with comprehensive clinical questions. This study was designed to demonstrate whether ChatGPT has solid geriatrics knowledge and can apply it to 2 common complex geriatric syndrome vignettes (polypharmacy and falls) by responding to comprehensive questions.

Study Objectives

We designed these 3 distinct approaches to provide evidence of whether ChatGPT can be trusted to be potentially applied to geriatrics education and clinical practice as an assistive tool.

Methods

This observational study used the validated UCLA geriatrics attitude instrument [48,49] and geriatrics knowledge test [57].

The Geriatrics Attitude Instrument

The geriatrics attitude instrument comprises 16 questions (Multimedia Appendix 1) [48,49]. In total, 6 of these statements reflect positive attitudes toward aging, such as "Most old people are pleasant to be with" (question 1). A total of 10 statements reflect negative attitudes toward aging, such as "Treatment of chronically ill old patients is hopeless" (question 11). Responses are graded on a Likert scale from 1 (*strongly disagree*) to 5 (*strongly agree*). The mean total score of ChatGPT was calculated and compared to scores from medical students, internal medicine residents, and geriatric medicine fellows based on previously published studies in the literature. Subtotals for positive and negative attitudes toward aging were also calculated to compare them to those in previously published studies. For statements reflecting a positive attitude toward aging, higher scores indicated a more positive attitude toward aging. For statements reflecting a negative attitude toward aging, lower scores indicated a less biased attitude toward aging.

The Geriatrics Knowledge Test

The validated UCLA geriatrics knowledge test contains 18 questions (Multimedia Appendix 2) [57]. In total, 8 questions are true-or-false statements, such as "Most older people are living in nursing homes" (question 4; false). A total of 10 questions are short clinical vignettes, such as "A 78-year-old nursing home resident with mild dementia associated with Alzheimer's disease is best determined by their ability to understand treatment options" (question 9; correct answer: B).

To minimize unfounded guessing, the following scoring system was used: +1 for a correct answer, –1 for a wrong answer, and 0 for "don't know." Therefore, the total score ranged from −18 to +18 [57]. Correct answers were determined by the author based on the literature.

ChatGPT Inputs and Outputs

GPT-3.5 was prompted to undertake the validated UCLA geriatrics attitude instrument and geriatrics knowledge test. Prompts were derived from original published papers [48,49,57]. Each prompt was repeated 4 times to evaluate the consistency and obtain an average response. In addition, ChatGPT was prompted to respond to 2 common geriatric syndrome vignettes (polypharmacy and falls). These 2 vignettes with questions were used in the author's previously published geriatrics curricula [70,71] and are described fully in the following sections.

The Polypharmacy Vignette

This vignette was used from a previously published workshop of prescribing and deprescribing [70]:

> *Ms. Smith, an 85-year-old white woman, has HTN, CAD, HFpEF, A-Fib, DM (II), HLD, hypothyroidism, gout, OA. On Warfarin 2.5 mg/d, furosemide 20 mg/d, Glyburide 5 mg/d, Spironolactone 50 mg/d, Amiodarone 300mg/d, Carvedilol 6.25 mg twice a day, Lisinopril 40 mg/d, Lipitor 80 mg/d, Levothyroxine 250 mg/d, Allopurinol 100 mg three times a day, acetaminophen 500 mg two tabs q6h for joint pain, and several supplements. Her pulse is 62 beat/min. Her BP is 120/70 (sit) and 98/64 (upright), weight 60 kg. No JVD, HR regularly, no murmur, few crackles on lung bases. Abdomen exam is benign, + pitting edema of ankles. Serum Cr is 1.2 mg/dL and Bun is 15 mg/dL. INR is 3.5. LDL is 50. albumin 4, Hb1ac is 6.1. TSH is 2. EF is 60%.*

ChatGPT was prompted with the polypharmacy vignette following 2 steps. First, ChatGPT was prompted to answer 4 general questions on the appropriate use of medications to examine its geriatric pharmacology knowledge: (1) "Which of the commonly used 8 drugs are potentially inappropriate medication (PIM) and should be avoided in older adults?" (2) "What is PIM in older adults?" (3) "How to identify PIM in older adults?" (4) "What to do with PIM?" Second, ChatGPT was prompted to answer 4 additional specific questions related to the polypharmacy vignette to evaluate its application of geriatric pharmacology to a clinical vignette: (1) "Any drug-drug interactions for this patient?" (2) "Any drug is ineffective for this patient?" (3) "This patient was taking so many medications. Any way to reduce the number of her medications?" (4) "Any drug-disease interaction?"

The Fall Vignette

This vignette was used from a previously published workshop of prescribing and deprescribing [71]:

> *An 82-year-old community-dwelling woman with PMH of multiple falls, right hip fracture, Dx of hypertension, coronary heart disease, knee arthritis, came to my clinic after she fell after her meal at home this morning. She felt pain of R head. She had a cane and walker at home but didn't use that often. She thought she didn't need it. She has taken care of her husband for several years. Her medication included Doxazosin (Cardura) and diltiazem (Cardizem), and acetaminophen as needed. I asked many questions related to her fall. Her vital signs were normal without orthostatic hypotension. Her cardiovascular and neurological examinations were benign. Her mini-mental status examination (MMSE) was 30/30.*

XSL•FO
RenderX

She had independent activities of daily living (ADL). Her geriatric depression scale (GDS) was 0/15. She passed timed up and go test. Because of her right black eye and pain of right head, I sent her to UVA to rule out orbital fracture. Please ask the following questions: List all potential fall risk factors for this patient What are three fall screening questions you need to ask for elderly patients in an outpatient setting based on 2010 fall prevention guideline you want to follow for this patient? Which one of three pathways in figure 1 from 2010 fall prevention guideline you want to follow for this patient? List all physical examinations that you think are important to perform for this patient. List your recommendations as many as you think that could help her to prevent her falling.

ChatGPT was prompted with the fall vignette following 3 steps. First, ChatGPT was prompted to answer 1 general question on fall risks in older adults. Second, ChatGPT was prompted to answer five pretest questions on fall prevention in older adults: (1) "When an older person falls, should you ask the following questions (one or more or none)? the circumstances of fall, mental status, medication changes, vision change, drinking"; (2) "Which of the following illnesses is a fall risk (one or more or none)? Pneumonia, Parkinson's disease, Diabetes mellitus, Dementia, Orthostatic hypotension"; (3) "Which of the following medications are not associated with an increased fall risk (one or more)? Levothyroxine, Diazepam, Oxycodone, Amitriptyline"; (4) "Which ones of the following are true (one or more or none)? An elderly person may become so fearful of falling that they restrict his or her mobility Most falls are associated with significant injuries Falls is a leading cause of accidental death in older population About 5% community-dwelling elderly person falls each year Survivors of fall-related hip fracture are rarely institutionalized Exercise improves function and reduces fall risk and injurious falls"; and (5) "Which of the following is NOT an environmental fall risk? (only pick one answer) Throw rug, freshly waxed kitchen floor, Grab bars, Electrical cord lying on the floor." Finally, ChatGPT was prompted to answer 5 additional questions to examine the application of fall prevention knowledge to the fall vignette: (1) "List all potential fall risk factors for this patient"; (2) "What are three fall screening questions you need to ask for elderly patients in an outpatient setting based on 2010 fall prevention guideline you want to follow for this patient?"; (3) "Which one of three pathways in figure 1 from 2010 fall prevention guideline you want to follow for this patient?"; (4) "List all physical examinations that you think they are important to perform for this patient"; and (5) "List your recommendations as many as you think that could help her to prevent her falling."

Evaluation of ChatGPT Outputs

ChatGPT outputs were collected and analyzed by the author. The author judged the correctness of the ChatGPT outputs on the polypharmacy and fall vignette questions based on evidence from clinical practice guidelines, systematic reviews, and other types of evidence in the literature. The descriptive analysis was performed using SPSS (version 28; IBM Corp).

In summary, we designed these 3 distinct approaches to provide evidence on whether we can trust ChatGPT to be potentially applied to geriatrics education and clinical practice as an assisting tool.

Ethical Considerations

This study was not human participant related, including the secondary data analysis. No informed consent was needed.

Results

Geriatrics Attitude

The total geriatrics attitude score of ChatGPT was significantly lower than that of trainees (medical students, internal medicine residents, and geriatric medicine fellows) from validation [49,57] and follow-up studies [48] (Multimedia Appendix 3). However, the mean subscore on positive geriatrics attitude of ChatGPT was higher than that of the trainees (medical students, internal medicine residents, and neurologists; Multimedia Appendix 4), where a higher score is better. The higher subscore on positive geriatrics attitude indicates a better attitude toward aging. Conversely, the mean subscore on negative geriatrics attitude of ChatGPT was lower than that of the trainees and neurologists, where lower subscores are better (Multimedia Appendix 4). The lower subscore on negative geriatrics attitude indicates a less age-biased attitude toward aging. Individual responses to 14 geriatrics attitude statements by ChatGPT, trainees, and neurologists are shown in Multimedia Appendix 4. The subscores of the comparison group in Multimedia Appendix 4 were based on previously published studies [72-74]. Notably, responses to statements 15 and 16 were rarely reported in previously published studies (Multimedia Appendix 4).

Of the 14 geriatrics attitude statements (Multimedia Appendix 4), 5 (36%) were positive, and 9 (64%) were negative. Regarding the positive geriatrics attitude statements, ChatGPT had 1 lower positive geriatric attitude response (indicating a less positive geriatrics attitude) and 5 similar or better positive geriatric attitude responses compared to medical students and residents (indicating a similar or better positive geriatrics attitude). Regarding the negative geriatrics attitude statements, ChatGPT had 1 similar negative geriatric attitude response and 8 better negative geriatrics attitude responses than medical students and residents, indicating a less negative geriatrics attitude.

The response from ChatGPT to the first prompt of the geriatrics attitude questions did not follow the Likert-scale format (1-5; 1=*strongly disagree*; 5=*strongly agree*; Multimedia Appendices 3 and 4) as ChatGPT provided only comments. The response to the third prompt included both the Likert-scale format (1-5; 1=*strongly disagree*; 5=*strongly agree*) and comments (Multimedia Appendix 4).

Geriatrics Knowledge

The total score of ChatGPT (mean 14.25) was significantly higher than the scores from all trainees (medical students, residents, and internal medicine fellows) in the validation studies [57] (mean 11.3; Multimedia Appendix 3) and was also significantly higher than that of first-year (mean 9.9) and second-year (mean 9.5) medical students (Multimedia Appendix

XSL•FO
RenderX

3) in the follow-up studies [48]. It was slightly higher than the scores of third-year medical students (mean 13.6) and internal medicine interns (mean 13.2), slightly lower than the scores of second- and third-year internal medicine residents (mean 14.7 and 14.9, respectively), and significantly lower than the scores of geriatric medicine fellows (mean 17.5) in the follow-up studies [48]. ChatGPT's performance on geriatrics knowledge in the first prompt was lower than that on the following 3 repeat prompts. For the first prompt, ChatGPT provided rationales without selecting options (Multimedia Appendix 4). For the third prompt, ChatGPT selected options and provided rationales for its choices (Multimedia Appendix 4). For the second and fourth prompts, ChatGPT select options without providing rationales (Multimedia Appendix 4).

To make it easy for the reader to replicate the results of this study, Multimedia Appendix 5 shows a few examples of prompts on geriatrics attitude and knowledge test questions and outputs by ChatGPT (all original outputs of ChatGPT are available upon request).

The Polypharmacy Vignette

Overall, ChatGPT performed well on the polypharmacy vignette involving a woman aged 85 years (Multimedia Appendix 6). ChatGPT provided appropriate responses to 4 general drug therapy questions and moderate responses to 4 specific drug therapy questions based on the vignette. ChatGPT correctly identified 5 drug-drug interactions and suggested deprescribing. However, it missed an ineffective medication (a supplement), 1 drug-drug interaction, and 3 drug-disease interactions (Multimedia Appendix 6). Despite the patient not having lung disease, ChatGPT provided 2 irrelevant drug-disease interactions specific to the vignette (ChatGPT hallucination; Multimedia Appendix 6).

The Fall Vignette

ChatGPT performed well on the fall vignette involving a woman aged 82 years (Multimedia Appendix 7). ChatGPT correctly summarized 10 common fall risks in older adults and correctly answered 3 out of 5 pretest questions. In the remaining 2 pretest questions, ChatGPT missed a few correct responses. In the fall vignette, ChatGPT provided appropriate responses to all 4 prompts (Multimedia Appendix 7). ChatGPT recognized most fall risk factors and responded perfectly to fall screening questions following the latest Centers for Disease Control and Prevention (CDC) fall guidelines even though the prompt specified the 2010 fall prevention guidelines. ChatGPT accurately followed the 3 pathways from the 2010 fall guidelines. It missed a few physical examinations. ChatGPT provided almost all recommended fall prevention strategies, with only a few omissions.

Discussion

Principal Findings

With the increasing use of ChatGPT in medical education [3-7] and clinical practice [5-12], the aim of this study was to evaluate whether we can trust ChatGPT to be used in geriatrics. The major findings of this study demonstrated that ChatGPT had less age-biased output than the trainees, outperformed the trainees on the validated UCLA geriatrics knowledge test, and reasonably applied geriatrics knowledge to 2 common geriatric syndrome vignettes (polypharmacy and falls). The preliminary findings of this study are promising and demonstrate that ChatGPT could be trusted and used as an assistant tool in geriatrics practice and education via 3 approaches, which are fully discussed in the following paragraphs.

The first approach was to evaluate ChatGPT's performance on a validated UCLA geriatrics attitude instrument [48,49]. To the best of our knowledge, this is the first study to demonstrate that ChatGPT is less likely to generate age-biased outputs, which contrasts with a few ethical and other concerns regarding ChatGPT [16-21]. This finding is significant because ageism is prevalent and associated with multiple poor outcomes, including health outcomes [34,35]. Various instruments exist to measure geriatrics attitude to assess ageism [34,35], with the UCLA geriatrics attitude instrument [48,49] being one of these validated tools [75], which we used in this study [48,49]. The UCLA geriatrics attitude instrument includes 16 items [48,49], 6 of which refer to a positive geriatrics attitude (eg, "most old people are pleasant to be with") where a higher score is better, and 10 of which refer to a negative geriatrics attitude where a lower score is better. Previous studies using the UCLA geriatrics attitude instrument have primarily used the total score on the geriatrics attitude scales without calculating and reporting the mean subscores on positive and negative geriatrics attitudes [48,49], which makes it difficult to interpret the significance of geriatrics attitude. However, a few studies have reported individual responses to the UCLA geriatrics attitude instrument, including positive and negative geriatrics attitude [48,49], for which we were able to calculate the subscores for comparison. We recommend that the subscores of the UCLA geriatrics attitude instrument should be reported to assess positive and negative geriatrics attitude. Overall, ChatGPT's individual responses to the UCLA geriatrics attitude statements were better than those of medical students and residents, with higher scores for positive attitude and lower scores for negative attitude (Multimedia Appendix 4), suggesting that ChatGPT has less age-biased outputs and could be trusted from an ethical perspective. This study suggests that ChatGPT is better than the news media and humans in providing less age-biased information [39,40]. However, a reliable and valid instrument with which to quantify modern medical student attitudes toward older people has not yet been developed. An adaptation of the Aging Semantic Differential scale for contemporary use has been recommended [76]. This study can be repeated using the Aging Semantic Differential scale to see whether ChatGPT still generates less age-biased outputs in the future.

The second approach was to evaluate ChatGPT's geriatrics knowledge. This study is the first to assess ChatGPT's performance on a validated UCLA geriatrics knowledge test [49,57]. ChatGPT has a substantial knowledge base and has been reported to pass the USMLE [50,51] and perform well on many other examinations [52-56]. Previous studies have suggested that ChatGPT could significantly impact medical education [3-7] and clinical practice, including clinical decision-making [5-12]. For example, a recent systematic review and meta-analysis on the performance of ChatGPT in medical

XSL•FO
RenderX

examinations showed that ChatGPT has been evaluated in multiple specialties, including plastic surgery, ophthalmology, neurosurgery, orthopedics, diabetes, gastroenterology, radiology, cardiology, dermatology, and anesthesia [61]. It can be reasonably believed that ChatGPT has good geriatrics knowledge. This study supported this belief and demonstrated that ChatGPT performed better than the average trainee in the validation studies (mean score 14.7 vs 11.3) and follow-up studies (mean score 14 vs 13.3; Multimedia Appendix 3). ChatGPT's performance was better than that of medical students and first-year internal medicine residents (PGY 1 in Multimedia Appendix 3), comparable to that of second- and third-year residents (PGY 2 and PGY 3 in Multimedia Appendix 3), but lower than that of geriatric medicine fellows. This suggests that ChatGPT could be a valuable supplemental resource for health professional trainees, health care providers, and laypeople. Given that GPT-4o has more parameters, it is reasonable to believe that GPT-4o could perform better on the geriatrics knowledge test. The UCLA geriatrics knowledge test only has 18 questions [57] and much less questions than the USMLE tested in previous studies [50,51]. Future studies should evaluate GPT-4o's performance on more multiple-choice questions from geriatrics board examinations to confirm the findings of this study before adopting ChatGPT for geriatrics education and clinical practice.

The third approach was to evaluate ChatGPT's responses to 2 common geriatric syndrome vignettes—polypharmacy and falls in older adults—from the author's previously published geriatrics curricula [70,71]. This study demonstrated that ChatGPT performed well on these 2 geriatric syndrome vignettes (Multimedia Appendices 6 and 7), which was promising and will be fully discussed in the following paragraphs.

For the polypharmacy vignette, ChatGPT was able to provide systematic recommendations for polypharmacy in older adults, which has not been investigated before. In this study, ChatGPT provided appropriate general principles for prescribing and deprescribing; identified PIMs; and suggested resources and criteria such as the Beers criteria [77], the Screening Tool of Older Persons' Prescriptions [75], the Drug Burden Index [78], and the Medication Appropriateness Index [79] (Multimedia Appendix 6). In addition, ChatGPT was able to provide 5 appropriate ways to manage PIMs in older adults (Multimedia Appendix 6) that are consistent with the evidence from geriatrics practice guidelines and prescribing principles [75,77-80]. In particular, this study demonstrated that ChatGPT can identify all 8 exemplary PIMs (Multimedia Appendix 6). These medications should be avoided in older adults [75,77-80]. ChatGPT was able to identify 4 drug-disease interactions but missed 3 drug-disease interactions in the polypharmacy vignette (Multimedia Appendix 6). This suggests room for improvement in training ChatGPT further. ChatGPT also identified drug-drug interactions accurately in the polypharmacy vignette (Multimedia Appendix 6), which was different from a previous study in which ChatGPT identified 39 out of 40 drug-drug interactions when it was given 40 pairs of drugs [65]. Therefore, the question is what the implications are of what ChatGPT can do for polypharmacy. Previous studies have shown that within the last 10 years medical students, postgraduate residents, primary care providers, and pharmacists were unaware of PIMs and the standard guidelines for older adults, such as the Beers criteria [81-84]. This study has shown that ChatGPT is more knowledgeable than humans. Both geriatricians and general internists still prescribed 7.2% to 8.7% of PIMs, respectively [84].

The persistent prevalence of PIMs and polypharmacy in older adults was well-reported [31,32]. What ChatGPT demonstrated its good knowledge and ability to do with polypharmacy in this study suggests that ChatGPT has great potential to be used by trainees and providers as an assistant tool. One suggested example will be to include an older patient's medical history and a list of all their medications in the ChatGPT prompt used in this study for medication review and to generate ChatGPT outputs which can be reviewed by providers at Clinic or other clinical settings. This will help nurses, other providers, and trainees at the point of care. It could help laypeople self-check PIMs. It could also help health professional trainees in the self-study of geriatric pharmacology.

Falling is another significant US and global problem [85,86] and one of the common geriatric syndromes in older adults [87]. ChatGPT has not been tested to provide systematic fall prevention recommendations for an older adult. This study demonstrated for the first time that ChatGPT has good knowledge of fall prevention as it performed well on pretests and provided a comprehensive summary of 10 common fall risks in older adults and specific recommendations in the fall vignette (Multimedia Appendix 7).

The 2024 US Preventive Services Task Force fall prevention guidelines listed several fall risks, including age; history of falls; cognitive and sensory deficits; presence of acute and chronic medical conditions; certain medications; environmental or occupational hazards; home or neighborhood features and alcohol or drug use; and impairments in mobility, gait, and balance [87]. The 2023 CDC Stopping Elderly Accidents, Deaths, and Injuries also has a long list of fall risks [88]. In addition, a systematic review and meta-analysis showed evidence-based fall risk factors among the aging population [89]. This study showed that ChatGPT identified most of fall risks for the fall vignette (Multimedia Appendix 7) consistent with the aforementioned fall prevention guidelines [87,88] and systematic review [89] except for missing risky behaviors such as standing on a chair instead of a step stool and loss of sensation in the feet.

This study also used 5 pretests to further evaluate ChatGPT's knowledge of fall prevention (Multimedia Appendix 7). ChatGPT generated responses consistent with the guidelines [88,89] and a systematic review [90]. In pretest 1, ChatGPT provided 5 good recommendations with reasoning for taking the medical history of an older adult. In pretest 2, ChatGPT correctly identified several medical conditions as fall risks, including Parkinson disease, dementia, and orthostatic hypotension but missing pneumonia and diabetes mellitus. This indicates that ChatGPT has some limitations. In pretest 3, ChatGPT correctly identified diazepam, oxycodone, and amitriptyline as fall risks [75,77-79] and levothyroxine as not a fall risk. In pretest 4, ChatGPT underestimated the prevalence

XSL•FO
RenderX

of falls in community-dwelling older adults (5%) but correctly answered most other true and false questions about fall risks and consequences. In pretest 5, ChatGPT correctly answered all questions about environmental fall risks and identified grab bars as not a risk. Overall, ChatGPT performed very well on 5 pretests with a few mistakes, indicating good knowledge of fall prevention for older adults.

To further test ChatGPT's clinical application to individual patients, the fall vignette was used to prompt ChatGPT. ChatGPT correctly identified 6 common fall risk factors except for potential home environment factors, consistent with the evidence from fall prevention guidelines [88,89] and a systematic review [90]. ChatGPT impressively recognized older caregiver burden, hypertension, and coronary heart diseases as fall risks, also consistent with the evidence from fall prevention guidelines [88,89] and a systematic review [90]. In response to fall screening questions, ChatGPT was correct if the 2023 CDC guidelines were used [89] but incorrect if the 2010 American Geriatrics Society guidelines were used [91]. This study indicates that ChatGPT can use the latest guidelines, although specific prompts might be unclear to ChatGPT. Regarding physical examinations relevant to the fall vignette, ChatGPT missed several but recommended multiple appropriate physical examinations [89,91], indicating a reasonably good performance. In the final prompt on fall prevention recommendations for the fall vignette, ChatGPT provided 6 appropriate fall prevention recommendations, including home environment–related prevention despite previously missing the home environment as a fall risk [88-91]. It was notable that the home environment was not on the list of fall risks in ChatGPT's response to the first prompt. However, ChatGPT correctly answered pretest 5 (home environment fall risk) and recommended home environment–related prevention for the patient in the fall vignette (Multimedia Appendix 7). This suggests some discordance between the fall assessment and action plans by ChatGPT, indicating a possibly different thinking and decision process within ChatGPT.

The crucial question is whether ChatGPT can be helpful and necessary in screening for fall risks and providing fall prevention recommendations to older adults in daily practice. Providers need knowledge and willingness to screen for fall risks in older adults, but multiple studies have shown that providers often lack fall prevention knowledge and are unwilling to screen for fall risks [92-96]. For example, one study showed that emergency providers lacked knowledge on which patients to be screened and were unwilling to spend more than a few minutes on screening for fall intervention [95]. Another study revealed that only 14% of providers at accountable care organizations were aware of the CDC's fall risk assessment algorithm (Stopping Elderly Accidents, Deaths, and Injuries) [93]. Furthermore, 43% of primary care providers did not agree that they had the expertise or time to perform fall risk assessments, and only a small percentage billed for fall risk screening despite being aware of Medicare reimbursement [96]. In addition, general practitioners were often unaware of their frail patients' fall history or fear of falling, with most patients not receiving fall prevention care. Less than 40% of providers asked most or all their older patients if they had fallen in the previous 12 months, and less than a quarter referred their patients to physical therapists for balance or gait training [92]. A recent CDC report showed some improvement, but gaps remain. Only 20% of providers were aware of any injury prevention resources, and a higher percentage screened for fall risk when older adults presented with specific concerns [94].

Given the high prevalence of fall among older adults and poor knowledge of fall prevention among providers, this study demonstrated that ChatGPT has better knowledge and ability to apply the fall prevention guidelines than many providers, suggesting that it can assist providers in assessing fall risks and preventing falls in older adults. For instance, entering an older patient's history into ChatGPT and reviewing the outputs generated can help providers identify patients with fall risk at the point of care. ChatGPT can also be used as an assistant tool to help health professional trainees with self-study.

Limitations and Quality Assessment of This Study

There are serval limitations to this study, which will be discussed in this section. ChatGPT's responses are based on the data that it was trained on, which may not include the most current geriatrics research, best practices, or region-specific guidelines, potentially limiting the trustworthiness of its advice. This study did not account for the quality of the information sources that ChatGPT draws from as ChatGPT cannot differentiate between authoritative and nonauthoritative content in real time like all other ChatGPT-based research. Despite GPT-4o having more data for training without releasing its data sources, this will not solve this basic problem. This study indicates that applying ChatGPT to geriatrics practice and education should be cautious and it should be used as an assistant tool only. ChatGPT will never or is unlikely to replace clinicians. As this study evaluated ChatGPT based on geriatric syndrome vignettes, it does not fully account for the complexities and nuances of real-world clinical decision-making and patient interactions, where additional context, history, and human judgment play critical roles. In addition, ChatGPT's responses might be overly generalized as it lacks the ability to perform individualized assessments or make nuanced clinical decisions based on patient-specific data. This is similar to vignette-based simulation education. This indicates that ChatGPT will not make a decision for someone but could assist with the decision-making. The effectiveness of ChatGPT's responses varies significantly depending on how well the user phrases the questions or provides necessary information. The input provided in this study was based on the author's geriatrics practice experience for >20 years, which might not mimic the way in which other clinicians would interact with the tool. Therefore, the results of this study may not be generalizable. Prompt engineering is developing to help improve the quality of the input. However, a standardized prompts among different providers can be very challenging. The author suggests practicing prompts to improve human-ChatGPT interaction and obtain reliable outputs from ChatGPT.

ChatGPT lacks clinical experience and the ability to engage in human clinical judgment, whereas it can process vast amounts of medical information. This may limit its application in critical decision-making scenarios or nuanced treatment planning for

XSL•FO
RenderX

complex geriatrics cases. Providers must use their clinical judgment integrated with clinical circumstances and not fully depend on the outputs of ChatGPT.

ChatGPT might face ethical and legal consequences. ChatGPT or AI hallucinations could cause ethical and legal risks. After the release of GPT-4o, ChatGPT often says, "I am AI and not clinician and cannot make decisions for you." It seems that ChatGPT developers recognize the potential legal and ethical consequences and protect ChatGPT.

Human-AI interaction is complex. Providers should be cautious in interpreting ChatGPT outputs. ChatGPT outputs should be interpreting with evidence-based clinical guidelines, systematic reviews, and randomized controlled trials shown in the Comments from the author columns in Multimedia Appendices 6 and 7. Whether the ChatGPT outputs are appropriate in this study should be judged and verified by a group of geriatricians

In addition, this study has several other limitations. This study used GPT-3.5. GPT-4o might produce better results. Comparisons between ChatGPT and human performance on geriatrics attitude and knowledge were based on previously reported results in the literature many years ago. A new study assessing geriatrics attitude and knowledge among current medical students, residents, and geriatric medicine fellows is needed to confirm the findings of this study. ChatGPT produced some errors, including ChatGPT hallucinations such as unrelated drug-disease interactions in Multimedia Appendix 6, consistent with previous reports [97,98]. We should be aware that ChatGPT can make mistakes, just like humans. ChatGPT should not be used alone in geriatrics practice and education at this point. ChatGPT needs to be pretrained specifically for medical education and clinical practice.

Like any other research, systematic appraisal of the current ChatGPT research is needed. A preliminary checklist, Model, Evaluation, Timing, Range and Randomization, Individual Factors, Counts, and Specificity of Prompts and Language (METRICS), to standardize the design and reporting of studies on generative AI–based models in health care education and practice has been recently developed [99]. The current study was started and completed before its development. However, it is worthwhile to use METRICS to evaluate the quality of this study. The following seven domains of METRICS will be used:

1. Model—ChatGPT was described and used.
2. Evaluation—both subjective and objective evaluation were described and used.
3. Timing—the study date was documented, but the duration was not documented; transparency—the geriatrics attitude instrument and knowledge test, the vignettes, and prompts were well described, and the outputs were exactly presented.
4. Range—the topics were well described, including geriatrics attitude and knowledge; randomization was not used.
5. Individual—there was subjective involvement with ChatGPT, such as judging the correctness of its outputs. This study was conducted by a single investigator.
6. Count—a total of 4 prompts were used. This study did not have a sample size.
7. Specificity of the prompts or language—for geriatrics attitude and knowledge, the prompts were from previously published studies. It is unclear whether they are appropriate for ChatGPT. For the 2 vignettes on polypharmacy and fall, the prompts were from previously published curricula. It is also unclear whether they are appropriate for ChatGPT.

Overall, the quality of this study is reasonably good based on METRICS. METRICS should be used in the reporting of any AI-based research in all journals.

In summary, this study took 3 distinct approaches to demonstrate the trustworthiness of ChatGPT to be used as an assistant tool in geriatrics practice and education. The implications of this preliminary study for geriatrics practice and education are further discussed in the following section.

Implications for Geriatrics Practice and Education

This study showed that ChatGPT outputs are not age-biased using the validated geriatrics attitude test and that ChatGPT performed well on the validated geriatrics knowledge test and on applied geriatrics knowledge tests (2 cases of common geriatric syndromes). These findings suggest that ChatGPT could potentially help geriatrics practice and education as an assistant tool in numerous ways. It is very important to know that ChatGPT is algorithm based. Clinical practice often uses algorithm- and pathway. Therefore, the underlying rationale for ChatGPT and clinical practice is similar. However, ChatGPT is not intended to replace physicians and other providers.

ChatGPT can be used as an assistant geriatrics tool and tutor to support self-learning and immediate feedback and self-assessment. For example, when trainees are studying geriatric syndromes, they can obtain responses on geriatric syndromes from ChatGPT. Therefore, ChatGPT could be incorporated to supplement the existing program in geriatrics education and support continuing medical education. ChatGPT can be potentially to be integrated to clinical reasoning and decision in a timely fashion and potentially improve the point of care in all clinical settings. For example, ChatGPT can help identify individual older patients who might have fall risks and drug-drug interactions.

ChatGPT can be used as an assistant or autonomous clinical tool to alleviate the geriatrics workforce crisis. For example, a nongeriatrician health provider could look for answers to clinical geriatrics questions, such as recommendations for preventing falls and reducing polypharmacy in older patients, from ChatGPT and help answer their patients' questions in a timely fashion. This is particularly helpful for rural practice where geriatrics consults are rare or nonexistent. This will improve physicians' and other providers' efficiency and accuracy in the care of older patients and provider-patient interactions.

It is expected that ChatGPT can help with post-visit summaries by providing patient education materials, such as home safety and environment modification for fall prevention. ChatGPT can also draft Clinic letters such as providing lab results and their interpretation to the patient in the Clinic letter.

ChatGPT can help older patients with self-screening or initial self-assessment when they are sitting in a clinic waiting room. For example, they can input their medical history into ChatGPT

XSL•FO
RenderX

such as age, physical function medications, and chronic conditions to ask whether they are at higher risk of fall.

Importantly, GPT-3.5 is free to use and can be cost-saving for providers and the health care system.

Finally, with more data training, ChatGPT will be more reliable and helpful in geriatrics practice and education.

Comparison With Previous Work

To the best of our knowledge, health professional trainees and providers but not ChatGPT have been assessed for geriatrics attitude [48,49]. ChatGPT has been assessed for geriatrics knowledge [28-30]. For example, ChatGPT was asked to answer 10 geriatrics questions [30]. The correctness of the outputs by ChatGPT were rated by a group of geriatricians [30]. In contrast, in this study, the correctness of the outputs by ChatGPT was determined by the author only, which could lead to errors. In another study, ChatGPT was prompted with 1 simple question [28]. The questions in the aforementioned studies were not validated. However, this study used validated UCLA geriatrics knowledge tests [57,72-74]. Several previous studies have demonstrated that ChatGPT has geriatric pharmacology knowledge [29,65-69]. For example, ChatGPT could identify drug-drug interactions from 40 drug-drug pairs [65]. In another study, ChatGPT could manage polypharmacy and make decisions on deprescribing 2 to 3 inappropriate medications based on a vignette [29]. Another study compared the answers from ChatGPT and clinical pharmacists to real clinical cases and clinical competency assessments [68]. However, these studies are different from this study, which used a systemic approach to polypharmacy (Multimedia Appendix 6). To the best of our knowledge, ChatGPT has not approached falls in older adults.

Future Directions

This study was a pilot with both promising findings and limitations. With the fast-growing use of ChatGPT and the new version of GPT-4o being released, the author suggests the following research directions.

A larger study using ChatGPT is needed to extend and confirm the findings of this study before adopting ChatGPT for geriatrics education and clinical practice. For example, more geriatrics knowledge questions and vignettes should be examined using GPT-4o. For example, the performance of ChatGPT on geriatrics certification exams should be tested.

Coinvestigators are needed to reduce biases. In particular, the correctness of the ChatGPT outputs should be evaluated and judged by a group of experts in geriatrics and integrated with the latest evidence, such as clinical practice guidelines, systematic reviews, and randomized controlled trials. For example, the ChatGPT outputs to a prompt should be rated from *strongly agree* to *strongly disagree* by a group of geriatrics experts in addition to determining their consistency with evidence-based clinical practice guidelines

Those who conduct ChatGPT research should receive training on prompt engineering. The author is doing this and feels that it is beneficial. This could reduce the variation in prompts.

The reliability of ChatGPT performance needs to be tested using different prompts to improve generalizability. The outcomes of the application of ChatGPT to geriatrics practice and education need to be further investigated.

Conclusions

This study suggests that ChatGPT could be a valuable assistant tool in geriatrics education and practice, helping health professional trainees and providers combat ageism, supplement geriatrics knowledge, and address common geriatric syndromes such as polypharmacy and falls. This study demonstrated that we could trust ChatGPT to be used in geriatrics practice and education by using 3 distinct approaches. One strength of this study is that Multimedia Appendices 5-7 provide the details of prompts to ChatGPT and the outputs generated by ChatGPT, which allows readers to apply a similar approach to their geriatrics education and practice studies. The findings of this study are promising but need more investigation before ChatGPT is widely adopted in geriatrics practice and education.

Data Availability

All data generated or analyzed during this study including all ChatGPT prompts and outputs are included in this published article and its supplementary information files on the journal website.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Geriatrics attitudes tests.
[DOCX File , 23 KB - formative_v9i1e63494_app1.docx]

Multimedia Appendix 2
University of California, Los Angeles, geriatrics knowledge test (questions 1-18).
[DOCX File , 25 KB - formative_v9i1e63494_app2.docx]

Multimedia Appendix 3

XSL•FO
RenderX

Comparison of geriatrics attitude and knowledge test performance between ChatGPT and trainees.
[DOCX File , 25 KB - formative_v9i1e63494_app3.docx]

Multimedia Appendix 4
Comparison of performance on the geriatrics attitude subscales by ChatGPT, trainees, and neurologists.
[DOCX File , 17 KB - formative_v9i1e63494_app4.docx]

Multimedia Appendix 5
Examples of ChatGPT outputs on geriatrics attitude and knowledge questions.
[DOCX File , 15 KB - formative_v9i1e63494_app5.docx]

Multimedia Appendix 6
ChatGPT output on the polypharmacy vignette about a woman aged 85 years.
[DOCX File , 17 KB - formative_v9i1e63494_app6.docx]

Multimedia Appendix 7
Fall risk identification and management in a woman aged 82 years.
[DOCX File , 20 KB - formative_v9i1e63494_app7.docx]

References

1. Number of ChatGPT users (Dec 2024). Exploding Topics. 2024 Dec 3. URL: https://explodingtopics.com/blog/chatgpt-users [accessed 2024-10-26]
2. PubMed homepage. PubMed. URL: https://pubmed.ncbi.nlm.nih.gov/ [accessed 2024-05-15]
3. Boscardin CK, Gin B, Golde PB, Hauer KE. ChatGPT and generative artificial intelligence for medical education: potential impact and opportunity. Acad Med 2023 Aug 31;99(1):22-27. [doi: 10.1097/acm.0000000000005439]
4. Preiksaitis C, Rose C. Opportunities, challenges, and future directions of generative artificial intelligence in medical education: scoping review. JMIR Med Educ 2023 Oct 20;9:e48785 [FREE Full text] [doi: 10.2196/48785] [Medline: 37862079]
5. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023 Mar 19;11(6):887 [FREE Full text] [doi: 10.3390/healthcare11060887] [Medline: 36981544]
6. Thirunavukarasu AJ, Ting DS, Elangovan K, Gutierrez L, Tan TF, Ting DS. Large language models in medicine. Nat Med 2023 Aug 17;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
7. Meng X, Yan X, Zhang K, Liu D, Cui X, Yang Y, et al. The application of large language models in medicine: a scoping review. iScience 2024 May 17;27(5):109713 [FREE Full text] [doi: 10.1016/j.isci.2024.109713] [Medline: 38746668]
8. Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. J Med Internet Res 2023 Jun 28;25:e48568 [FREE Full text] [doi: 10.2196/48568] [Medline: 37379067]
9. Andrew A. Potential applications and implications of large language models in primary care. Fam Med Community Health 2024 Jan 30;12(Suppl 1):e002602 [FREE Full text] [doi: 10.1136/fmch-2023-002602] [Medline: 38290759]
10. Li J, Dada A, Puladi B, Kleesiek J, Egger J. ChatGPT in healthcare: a taxonomy and systematic review. Comput Methods Programs Biomed 2024 Mar;245:108013 [FREE Full text] [doi: 10.1016/j.cmpb.2024.108013] [Medline: 38262126]
11. Rao SJ, Isath A, Krishnan P, Tangsrivimol JA, Virk HU, Wang Z, et al. ChatGPT: a conceptual review of applications and utility in the field of medicine. J Med Syst 2024 Jun 05;48(1):59. [doi: 10.1007/s10916-024-02075-x] [Medline: 38836893]
12. Momenaei B, Mansour HA, Kuriyan AE, Xu D, Sridhar J, Ting DS, et al. ChatGPT enters the room: what it means for patient counseling, physician education, academics, and disease management. Curr Opin Ophthalmol 2024 May 01;35(3):205-209. [doi: 10.1097/ICU.0000000000001036] [Medline: 38334288]
13. Cheng SW, Chang CW, Chang WJ, Wang HW, Liang CS, Kishimoto T, et al. The now and future of ChatGPT and GPT in psychiatry. Psychiatry Clin Neurosci 2023 Nov 11;77(11):592-596 [FREE Full text] [doi: 10.1111/pcn.13588] [Medline: 37612880]
14. Temperley HC, O'Sullivan NJ, Mac Curtain BM, Corr A, Meaney JF, Kelly ME, et al. Current applications and future potential of ChatGPT in radiology: a systematic review. J Med Imaging Radiat Oncol 2024 Apr;68(3):257-264. [doi: 10.1111/1754-9485.13621] [Medline: 38243605]
15. Lautrup AD, Hyrup T, Schneider-Kamp A, Dahl M, Lindholt JS, Schneider-Kamp P. Heart-to-heart with ChatGPT: the impact of patients consulting AI for cardiovascular health advice. Open Heart 2023 Nov;10(2):e002455 [FREE Full text] [doi: 10.1136/openhrt-2023-002455] [Medline: 37945282]
16. Mohammad B, Supti T, Alzubaidi M, Shah H, Alam T, Shah Z, et al. The pros and cons of using ChatGPT in medical education: a scoping review. Stud Health Technol Inform 2023 Jun 29;305:644-647. [doi: 10.3233/SHTI230580] [Medline: 37387114]

XSL•FO
RenderX

17. Duffourc M, Gerke S. Generative AI in health care and liability risks for physicians and safety concerns for patients. JAMA 2023 Jul 25;330(4):313-314. [doi: 10.1001/jama.2023.9630] [Medline: 37410497]
18. Kanter GP, Packel EA. Health care privacy risks of AI chatbots. JAMA 2023 Jul 25;330(4):311-312. [doi: 10.1001/jama.2023.9618] [Medline: 37410449]
19. Hanna JJ, Wakene AD, Lehmann CU, Medford RJ. Assessing racial and ethnic bias in text generation for healthcare-related tasks by ChatGPT. medRxiv. Preprint posted online on August 28, 2023 [FREE Full text] [doi: 10.1101/2023.08.28.23294730] [Medline: 37693388]
20. Amin KS, Forman HP, Davis MA. Even with ChatGPT, race matters. Clin Imaging 2024 May;109:110113. [doi: 10.1016/j.clinimag.2024.110113] [Medline: 38552383]
21. Erren TC, Lewis P, Shaw DM. Brave (in a) new world: an ethical perspective on chatbots for medical advice. Front Public Health 2023 Aug 17;11:1254334 [FREE Full text] [doi: 10.3389/fpubh.2023.1254334] [Medline: 37663854]
22. Korcok M. Medical schools face challenge of preparing physicians to care for fast-growing elderly population. JAMA 1985 Mar 01;253(9):1225-7, 1231. [Medline: 3968847]
23. Besdine R, Boult C, Brangman S, Coleman EA, Fried LP, Gerety M, et al. Caring for older Americans: the future of geriatric medicine. J Am Geriatr Soc 2005 Jun;53(6 Suppl):S245-S256. [doi: 10.1111/j.1532-5415.2005.53350.x] [Medline: 15963180]
24. Institute of Medicine, Committee on the Future Health Care Workforce for Older Americans, Board on Health Care Services. Retooling for an Aging America: Building the Health Care Workforce. Washington, DC: National Academies Press; 2008.
25. Dawson CM, Abiola AO, Sullivan AM, Schwartz AW, Members of the GERI Team Research Group. You can't be what you can't see: a systematic website review of Geriatrics Online-Visibility at US medical schools. J Am Geriatr Soc 2022 Oct;70(10):2996-3005 [FREE Full text] [doi: 10.1111/jgs.17997] [Medline: 36082829]
26. Gurwitz JH. The paradoxical decline of geriatric medicine as a profession. JAMA 2023 Aug 22;330(8):693-694. [doi: 10.1001/jama.2023.11110] [Medline: 37540519]
27. Fear K, Gleber C. Shaping the future of older adult care: ChatGPT, advanced AI, and the transformation of clinical practice. JMIR Aging 2023 Sep 13;6:e51776 [FREE Full text] [doi: 10.2196/51776] [Medline: 37703085]
28. Haque N. Artificial intelligence and geriatric medicine: new possibilities and consequences. J Am Geriatr Soc 2023 Jun 17;71(6):2028-2031. [doi: 10.1111/jgs.18334] [Medline: 36930059]
29. Rao A, Kim J, Lie W, Pang M, Fuh L, Dreyer KJ, et al. Proactive polypharmacy management using large language models: opportunities to enhance geriatric care. J Med Syst 2024 Apr 18;48(1):41. [doi: 10.1007/s10916-024-02058-y] [Medline: 38632172]
30. Rosselló-Jiménez D, Docampo S, Collado Y, Cuadra-Llopart L, Riba F, Llonch-Masriera M. Geriatrics and artificial intelligence in Spain (Ger-IA project): talking to ChatGPT, a nationwide survey. Eur Geriatr Med 2024 Aug;15(4):1129-1136. [doi: 10.1007/s41999-024-00970-7] [Medline: 38615289]
31. Nicholson K, Liu W, Fitzpatrick D, Hardacre KA, Roberts S, Salerno J, et al. Prevalence of multimorbidity and polypharmacy among adults and older adults: a systematic review. Lancet Healthy Longev 2024 Apr;5(4):e287-e296 [FREE Full text] [doi: 10.1016/S2666-7568(24)00007-2] [Medline: 38452787]
32. Hung A, Kim YH, Pavon JM. Deprescribing in older adults with polypharmacy. BMJ 2024 May 07;385:e074892. [doi: 10.1136/bmj-2023-074892] [Medline: 38719530]
33. Ng R, Chow TY. Aging narratives over 210 years (1810-2019). J Gerontol B Psychol Sci Soc Sci 2021 Oct 30;76(9):1799-1807 [FREE Full text] [doi: 10.1093/geronb/gbaa222] [Medline: 33300996]
34. Global report on ageism. World Health Organization. 2021 Mar 18. URL: https://www.who.int/teams/social-determinants-of-health/demographic-change-and-healthy-ageing/combatting-ageism/global-report-on-ageism [accessed 2024-12-12]
35. Hu RX, Luo M, Zhang A, Li LW. Associations of ageism and health: a systematic review of quantitative observational studies. Res Aging 2021 Aug;43(7-8):311-322. [doi: 10.1177/0164027520980130] [Medline: 33317402]
36. Allen JO, Solway E, Kirch M, Singer D, Kullgren JT, Moïse V, et al. Experiences of everyday ageism and the health of older US adults. JAMA Netw Open 2022 Jun 01;5(6):e2217240 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.17240] [Medline: 35704314]
37. Chang ES, Kannoth S, Levy S, Wang SY, Lee JE, Levy BR. Global reach of ageism on older persons' health: a systematic review. PLoS One 2020;15(1):e0220857 [FREE Full text] [doi: 10.1371/journal.pone.0220857] [Medline: 31940338]
38. Araújo PO, Soares IM, Vale PR, Sousa AR, Aparicio EC, Carvalho ES. Ageism directed to older adults in health services: a scoping review. Rev Lat Am Enfermagem 2023;31:e4019 [FREE Full text] [doi: 10.1590/1518-8345.6727.4019] [Medline: 37820219]
39. Amundsen D. A critical gerontological framing analysis of persistent ageism in NZ online news media: don't call us "elderly"!. J Aging Stud 2022 Jun;61:101009 [FREE Full text] [doi: 10.1016/j.jaging.2022.101009] [Medline: 35654544]
40. Murphy E, Fallon A, Dukelow T, O'Neill D. Don't call me elderly: a review of medical journals' use of ageist literature. Eur Geriatr Med 2022 Aug;13(4):1007-1009. [doi: 10.1007/s41999-022-00650-4] [Medline: 35581468]
41. Haase KR, Sattar S, Pilleron S, Lambrechts Y, Hannan M, Navarrete E, et al. A scoping review of ageism towards older adults in cancer care. J Geriatr Oncol 2023 Jan;14(1):101385. [doi: 10.1016/j.jgo.2022.09.014] [Medline: 36244925]

XSL•FO
RenderX

42. Bowling A. Ageism in cardiology. BMJ 1999 Nov 20;319(7221):1353-1355 [FREE Full text] [doi: 10.1136/bmj.319.7221.1353] [Medline: 10567146]
43. Silva MF, Silva DS, Bacurau AG, Francisco PM, Assumpção D, Neri AL, et al. Ageism against older adults in the context of the COVID-19 pandemic: an integrative review. Rev Saude Publica 2021 Apr 05;55:4 [FREE Full text] [doi: 10.11606/s1518-8787.2021055003082] [Medline: 33886953]
44. Altın Z, Buran F. Attitudes of health professionals toward elderly patients during the COVID-19 pandemic. Aging Clin Exp Res 2022 Oct;34(10):2567-2576 [FREE Full text] [doi: 10.1007/s40520-022-02209-6] [Medline: 35986878]
45. Martínez-Arnau FM, López-Hernández L, Castellano-Rioja E, Botella-Navas M, Pérez-Ros P. Interventions to improve attitudes toward older people in undergraduate health and social sciences students. A systematic review and meta-analysis. Nurse Educ Today 2022 Mar;110:105269 [FREE Full text] [doi: 10.1016/j.nedt.2022.105269] [Medline: 35063781]
46. Apriceno M, Levy SR. Systematic review and meta-analyses of effective programs for reducing ageism toward older adults. J Appl Gerontol 2023 Jun 17;42(6):1356-1375 [FREE Full text] [doi: 10.1177/07334648231165266] [Medline: 37069824]
47. Oca M, Meller L, Wilson K, Parikh AO, McCoy A, Chang J, et al. Bias and inaccuracy in AI chatbot ophthalmologist recommendations. Cureus 2023 Sep;15(9):e45911 [FREE Full text] [doi: 10.7759/cureus.45911] [Medline: 37885556]
48. Kishimoto M, Nagoshi M, Williams S, Masaki KH, Blanchette PL. Knowledge and attitudes about geriatrics of medical students, internal medicine residents, and geriatric medicine fellows. J Am Geriatr Soc 2005 Jan 22;53(1):99-102. [doi: 10.1111/j.1532-5415.2005.53018.x] [Medline: 15667384]
49. Reuben DB, Lee M, Davis JWJ, Eslami MS, Osterweil DG, Melchiore S, et al. Development and validation of a geriatrics attitudes scale for primary care residents. J Am Geriatr Soc 1998 Nov 27;46(11):1425-1430. [doi: 10.1111/j.1532-5415.1998.tb06012.x] [Medline: 9809767]
50. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the united states medical licensing examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 08;9:e45312 [FREE Full text] [doi: 10.2196/45312] [Medline: 36753318]
51. Yaneva V, Baldwin P, Jurich DP, Swygert K, Clauser BE. Examining ChatGPT performance on USMLE sample items and implications for assessment. Acad Med 2024 Feb 01;99(2):192-197. [doi: 10.1097/ACM.0000000000005549] [Medline: 37934828]
52. Levin G, Horesh N, Brezinov Y, Meyer R. Performance of ChatGPT in medical examinations: a systematic review and a meta-analysis. BJOG 2024 Feb;131(3):378-380. [doi: 10.1111/1471-0528.17641] [Medline: 37604703]
53. Kung JE, Marshall C, Gauthier C, Gonzalez TA, Jackson JB3. Evaluating ChatGPT performance on the orthopaedic in-training examination. JB JS Open Access 2023;8(3):e23.00056 [FREE Full text] [doi: 10.2106/JBJS.OA.23.00056] [Medline: 37693092]
54. Passby L, Jenko N, Wernham A. Performance of ChatGPT on specialty certificate examination in dermatology multiple-choice questions. Clin Exp Dermatol 2024 Jun 25;49(7):722-727. [doi: 10.1093/ced/llad197] [Medline: 37264670]
55. Ali R, Tang OY, Connolly ID, Zadnik Sullivan PL, Shin JH, Fridley JS, et al. Performance of ChatGPT and GPT-4 on neurosurgery written board examinations. Neurosurgery 2023 Dec 01;93(6):1353-1365. [doi: 10.1227/neu.0000000000002632] [Medline: 37581444]
56. Long C, Lowe K, Zhang J, Santos AD, Alanazi A, O'Brien D, et al. A novel evaluation model for assessing ChatGPT on otolaryngology-head and neck surgery certification examinations: performance study. JMIR Med Educ 2024 Jan 16;10:e49970 [FREE Full text] [doi: 10.2196/49970] [Medline: 38227351]
57. Reuben DB, Lee M, Davis JW, Eslami MS, Osterweil DG, Melchiore S, et al. Development and validation of a geriatrics knowledge test for primary care residents. J Gen Intern Med 1997 Jul;12(7):450-452 [FREE Full text] [doi: 10.1046/j.1525-1497.1997.00078.x] [Medline: 9229284]
58. Rao A, Pang M, Kim J, Kamineni M, Lie W, Prasad AK, et al. Assessing the utility of ChatGPT throughout the entire clinical workflow. medRxiv. Preprint posted online on February 26, 2023 [FREE Full text] [doi: 10.1101/2023.02.21.23285886] [Medline: 36865204]
59. Qu RW, Qureshi U, Petersen G, Lee SC. Diagnostic and management applications of ChatGPT in structured otolaryngology clinical scenarios. OTO Open 2023;7(3):e67 [FREE Full text] [doi: 10.1002/oto2.67] [Medline: 37614494]
60. Franco D'Souza R, Amanullah S, Mathew M, Surapaneni KM. Appraising the performance of ChatGPT in psychiatry using 100 clinical case vignettes. Asian J Psychiatr 2023 Nov;89:103770. [doi: 10.1016/j.ajp.2023.103770] [Medline: 37812998]
61. Goodman RS, Patrinely JR, Stone CAJ, Zimmerman E, Donald RR, Chang SS, et al. Accuracy and reliability of chatbot responses to physician questions. JAMA Netw Open 2023 Oct 02;6(10):e2336483 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.36483] [Medline: 37782499]
62. Kassab J, El Dahdah J, Chedid El Helou M, Layoun H, Sarraju A, Laffin LJ, et al. Assessing the accuracy of an online chat-based artificial intelligence model in providing recommendations on hypertension management in accordance with the 2017 American College of Cardiology/American Heart Association and 2018 European Society of Cardiology/European Society of Hypertension Guidelines. Hypertension 2023 Jul;80(7). [doi: 10.1161/hypertensionaha.123.21183]
63. Rizwan A, Sadiq T. The use of AI in diagnosing diseases and providing management plans: a consultation on cardiovascular disorders with ChatGPT. Cureus 2023 Aug;15(8):e43106 [FREE Full text] [doi: 10.7759/cureus.43106] [Medline: 37692649]

XSL•FO
RenderX

64. Zhang L, Tashiro S, Mukaino M, Yamada S. Use of artificial intelligence large language models as a clinical tool in rehabilitation medicine: a comparative test case. J Rehabil Med 2023 Sep 11;55:jrm13373 [FREE Full text] [doi: 10.2340/jrm.v55.13373] [Medline: 37691497]
65. Juhi A, Pipil N, Santra S, Mondal S, Behera JK, Mondal H. The capability of ChatGPT in predicting and explaining common drug-drug interactions. Cureus 2023 Mar;15(3):e36272 [FREE Full text] [doi: 10.7759/cureus.36272] [Medline: 37073184]
66. Fournier A, Fallet C, Sadeghipour F, Perrottet N. Assessing the applicability and appropriateness of ChatGPT in answering clinical pharmacy questions. Ann Pharm Fr 2024 May;82(3):507-513. [doi: 10.1016/j.pharma.2023.11.001] [Medline: 37992892]
67. Al-Dujaili Z, Omari S, Pillai J, Al Faraj A. Assessing the accuracy and consistency of ChatGPT in clinical pharmacy management: a preliminary analysis with clinical pharmacy experts worldwide. Res Social Adm Pharm 2023 Dec;19(12):1590-1594. [doi: 10.1016/j.sapharm.2023.08.012] [Medline: 37696742]
68. Huang X, Estau D, Liu X, Yu Y, Qin J, Li Z. Evaluating the performance of ChatGPT in clinical pharmacy: a comparative study of ChatGPT and clinical pharmacists. Br J Clin Pharmacol 2024 Jan;90(1):232-238. [doi: 10.1111/bcp.15896] [Medline: 37626010]
69. Choi W. Assessment of the capacity of ChatGPT as a self-learning tool in medical pharmacology: a study using MCQs. BMC Med Educ 2023 Nov 13;23(1):864 [FREE Full text] [doi: 10.1186/s12909-023-04832-x] [Medline: 37957666]
70. Cheng HY, Bradley E. Improving identification of potentially inappropriate medications among older patients: a preliminary report of a single-blind, controlled trial teaching fourth year medical students value- and preference-based prescribing/de-prescribing. Med Sci Educ 2016 Jun 28;26:431-435. [doi: 10.1007/s40670-016-0283-2]
71. Cheng H. Falls and fall prevention. MedEdPORTAL 2014 Sep 26:9910. [doi: 10.15766/mep_2374-8265.9910]
72. Alqahtani R, Almuhaidib S, Jradi H. A cross-sectional study: exploring knowledge and attitude of medical and nursing students to Care for Elders in the future. BMC Geriatr 2022 Nov 14;22(1):856 [FREE Full text] [doi: 10.1186/s12877-022-03551-0] [Medline: 36376782]
73. Hogan TM, Chan SB, Hansoti B. Multidimensional attitudes of emergency medicine residents toward older adults. West J Emerg Med 2014 Jul;15(4):511-517 [FREE Full text] [doi: 10.5811/westjem.2014.2.19937] [Medline: 25035760]
74. Seferoğlu M, Yıldız D, Pekel NB, Güneş A, Yıldız A, Tufan F. Attitudes of neurology specialists toward older adults. Aging Clin Exp Res 2017 Aug;29(4):787-792. [doi: 10.1007/s40520-016-0606-6] [Medline: 27456680]
75. O'Mahony D, Cherubini A, Guiteras AR, Denkinger M, Beuscart JB, Onder G, et al. STOPP/START criteria for potentially inappropriate prescribing in older people: version 3. Eur Geriatr Med 2023 Aug;14(4):625-632 [FREE Full text] [doi: 10.1007/s41999-023-00777-y] [Medline: 37256475]
76. Wilson MA, Kurrle S, Wilson I. Medical student attitudes towards older people: a critical review of quantitative measures. BMC Res Notes 2018 Jan 24;11(1):71 [FREE Full text] [doi: 10.1186/s13104-018-3186-z] [Medline: 29361969]
77. By the 2023 American Geriatrics Society Beers Criteria® Update Expert Panel. American Geriatrics Society 2023 updated AGS Beers Criteria® for potentially inappropriate medication use in older adults. J Am Geriatr Soc 2023 Jul 04;71(7):2052-2081. [doi: 10.1111/jgs.18372] [Medline: 37139824]
78. Hilmer SN, Mager DE, Simonsick EM, Cao Y, Ling SM, Windham BG, et al. A drug burden index to define the functional burden of medications in older people. Arch Intern Med 2007 Apr 23;167(8):781-787. [doi: 10.1001/archinte.167.8.781] [Medline: 17452540]
79. Samsa GP, Hanlon JT, Schmader KE, Weinberger M, Clipp EC, Uttech KM, et al. A summated score for the medication appropriateness index: development and assessment of clinimetric properties including content validity. J Clin Epidemiol 1994 Aug;47(8):891-896. [doi: 10.1016/0895-4356(94)90192-9] [Medline: 7730892]
80. Pravodelov V. Thoughtful prescribing and deprescribing. Med Clin North Am 2020 Sep;104(5):751-765. [doi: 10.1016/j.mcna.2020.06.001] [Medline: 32773043]
81. Ramaswamy R, Maio V, Diamond JJ, Talati AR, Hartmann CW, Arenson C, et al. Potentially inappropriate prescribing in elderly: assessing doctor knowledge, confidence and barriers. J Eval Clin Pract 2011 Dec;17(6):1153-1159. [doi: 10.1111/j.1365-2753.2010.01494.x] [Medline: 20630004]
82. Abd Wahab MS. The relevance of educating doctors, pharmacists and older patients about potentially inappropriate medications. Int J Clin Pharm 2015 Dec;37(6):971-974. [doi: 10.1007/s11096-015-0203-3] [Medline: 26415737]
83. Rathore A, Sharma R, Bansal P, Chhabra M, Arora M. Knowledge, attitude, and practice of medical interns and postgraduate residents on American Geriatric Society updated Beers criteria. J Educ Health Promot 2023;12:1 [FREE Full text] [doi: 10.4103/jehp.jehp_769_22] [Medline: 37034871]
84. Vandergrift JL, Weng W, Leff B, Gray BM. Geriatricians, general internists, and potentially inappropriate medications for a national sample of older adults. J Am Geriatr Soc 2024 Jan;72(1):37-47. [doi: 10.1111/jgs.18489] [Medline: 37350649]
85. Moreland B, Kakara R, Henry A. Trends in nonfatal falls and fall-related injuries among adults aged ≥65 years - United States, 2012-2018. MMWR Morb Mortal Wkly Rep 2020 Jul 10;69(27):875-881 [FREE Full text] [doi: 10.15585/mmwr.mm6927a5] [Medline: 32644982]
86. Xiong W, Wang D, Ren W, Liu X, Wen R, Luo Y. The global prevalence of and risk factors for fear of falling among older adults: a systematic review and meta-analysis. BMC Geriatr 2024 Apr 05;24(1):321 [FREE Full text] [doi: 10.1186/s12877-024-04882-w] [Medline: 38580924]

XSL•FO
RenderX

87. Rausch C, van Zon SK, Liang Y, Laflamme L, Möller J, de Rooij SE, et al. Geriatric syndromes and incident chronic health conditions among 9094 older community-dwellers: findings from the lifelines cohort study. J Am Med Dir Assoc 2022 Jan;23(1):54-9.e2 [FREE Full text] [doi: 10.1016/j.jamda.2021.02.030] [Medline: 33798484]
88. US Preventive Services Task Force, Nicholson WK, Silverstein M, Wong JB, Barry MJ, Chelmow D, et al. Interventions to prevent falls in community-dwelling older adults: US Preventive Services Task Force recommendation statement. JAMA 2024 Jul 02;332(1):51-57. [doi: 10.1001/jama.2024.8481] [Medline: 38833246]
89. Clinical resources. Centers for Disease Control and Prevention STEADI. 2024 Apr 22. URL: https://www.cdc.gov/steadi/hcp/clinical-resources/index.html [accessed 2024-06-25]
90. Xu Q, Ou X, Li J. The risk of falls among the aging population: a systematic review and meta-analysis. Front Public Health 2022;10:902599 [FREE Full text] [doi: 10.3389/fpubh.2022.902599] [Medline: 36324472]
91. Panel on Prevention of Falls in Older Persons, American Geriatrics Society and British Geriatrics Society. Summary of the Updated American Geriatrics Society/British Geriatrics Society clinical practice guideline for prevention of falls in older persons. J Am Geriatr Soc 2011 Jan;59(1):148-157. [doi: 10.1111/j.1532-5415.2010.03234.x] [Medline: 21226685]
92. Smith ML, Stevens JA, Ehrenreich H, Wilson AD, Schuster RJ, Cherry CO, et al. Healthcare providers' perceptions and self-reported fall prevention practices: findings from a large New York health system. Front Public Health 2015;3:17 [FREE Full text] [doi: 10.3389/fpubh.2015.00017] [Medline: 25964942]
93. Howland J, Hackman H, Taylor A, O'Hara K, Liu J, Brusch J. Older adult fall prevention practices among primary care providers at accountable care organizations: a pilot study. PLoS One 2018;13(10):e0205279 [FREE Full text] [doi: 10.1371/journal.pone.0205279] [Medline: 30307974]
94. Ortmann N, Haddad YK, Beck L. Special report from the CDC: provider knowledge and practices around driving safety and fall prevention screening and recommendations for their older adult patients, DocStyles 2019. J Safety Res 2023 Sep;86:401-408. [doi: 10.1016/j.jsr.2023.06.011] [Medline: 37718068]
95. Davenport K, Cameron A, Samson M, Sri-On J, Liu SW. Fall prevention knowledge, attitudes, and behaviors: a survey of emergency providers. West J Emerg Med 2020 Jul 10;21(4):826-830 [FREE Full text] [doi: 10.5811/westjem.2020.4.43387] [Medline: 32726252]
96. Meekes WM, Leemrijse CJ, Weesie YM, van de Goor IA, Donker GA, Korevaar JC. Falls prevention at GP practices: a description of daily practice. BMC Fam Pract 2021 Sep 21;22(1):190 [FREE Full text] [doi: 10.1186/s12875-021-01540-7] [Medline: 34548022]
97. Siontis KC, Attia ZI, Asirvatham SJ, Friedman PA. ChatGPT hallucinating: can it get any more humanlike? Eur Heart J 2024 Feb 01;45(5):321-323. [doi: 10.1093/eurheartj/ehad766] [Medline: 38088452]
98. Hatem R, Simmons B, Thornton JE. A call to address AI "hallucinations" and how healthcare professionals can mitigate their risks. Cureus 2023 Sep;15(9):e44720 [FREE Full text] [doi: 10.7759/cureus.44720] [Medline: 37809168]
99. Sallam M, Barakat M, Sallam M. A preliminary checklist (METRICS) to standardize the design and reporting of studies on generative artificial intelligence-based models in health care education and practice: development study involving a literature review. Interact J Med Res 2024 Feb 15;13:e54704 [FREE Full text] [doi: 10.2196/54704] [Medline: 38276872]

Abbreviations

AI: artificial intelligence
CDC: Centers for Disease Control and Prevention
METRICS: Model, Evaluation, Timing, Range and Randomization, Individual Factors, Counts, and Specificity of Prompts and Language
PIM: potentially inappropriate medication
UCLA: University of California, Los Angeles
USMLE: United States Medical Licensing Examination

Edited by A Mavragani; submitted 30.06.24; peer-reviewed by J Mira, P Heyn; comments to author 17.10.24; revised version received 26.10.24; accepted 17.11.24; published 03.01.25.

Please cite as:
Cheng HY
ChatGPT's Attitude, Knowledge, and Clinical Application in Geriatrics Practice and Education: Exploratory Observational Study
JMIR Form Res 2025;9:e63494
URL: https://formative.jmir.org/2025/1/e63494
doi:10.2196/63494
PMID:39752214

XSL•FO
RenderX

©Huai Yong Cheng. Originally published in JMIR Formative Research (https://formative.jmir.org), 03.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Digital Parenting Intervention With Intimate Partner Violence Prevention Content: Quantitative Pre-Post Pilot Study

Moa Schafer[1], DPhil; Jamie Lachman[1,2], DPhil; Paula Zinser[1], MSc; Francisco Antonio Calderón Alfaro[1], MSc; Qing Han[1], PhD; Chiara Facciola[3], PhD; Lily Clements[3], PhD; Frances Gardner[1], DPhil; Genevieve Haupt Ronnie[2], MA; Ross Sheil[4], BSc

[1]Centre for Evidence Based Intervention, Department of Social Policy and Intervention, University of Oxford, Oxford, United Kingdom

[2]Centre for Social Science Research, University of Cape Town, Cape Town, South Africa

[3]Innovations in Development, Education, and the Mathematical Sciences (IDEMS) International, Reading, United Kingdom

[4]UNICEF Jamaica, Kingston, Jamaica

Corresponding Author:
Moa Schafer, DPhil
Centre for Evidence Based Intervention
Department of Social Policy and Intervention
University of Oxford
Barnett House
32-37 Wellington Square
Oxford, OX1 2ER
United Kingdom
Phone: 44 01865 2703
Email: moa.schafer@spi.ox.ac.uk

Abstract

Background: Intimate partner violence (IPV) and violence against children are global issues with severe consequences. Intersections shared by the 2 forms of violence have led to calls for joint programming efforts to prevent both IPV and violence against children. Parenting programs have been identified as a key entry point for addressing multiple forms of family violence. Building on the IPV prevention material that has been integrated into the parenting program ParentText, a digital parenting chatbot, this pilot study seeks to explore parents' engagement with the IPV prevention content in ParentText and explore preliminary changes in IPV.

Objective: This study aimed to assess parents' and caregivers' level of engagement with the IPV prevention material in the ParentText chatbot and explore preliminary changes in experiences and perpetration of IPV, attitudes toward IPV, and gender-equitable behaviors following the intervention.

Methods: Caregivers of children aged between 0 and 18 years were recruited through convenience sampling by research assistants in Cape Town, South Africa, and by UNICEF (United Nations Children's Fund) Jamaica staff in 3 parishes of Jamaica. Quantitative data from women in Jamaica (n=28) and South Africa (n=19) and men in South Africa (n=21) were collected electronically via weblinks sent to caregivers' phones using Open Data Kit. The primary outcome was IPV experience (women) and perpetration (men), with secondary outcomes including gender-equitable behaviors and attitudes toward IPV. Descriptive statistics were used to report sociodemographic characteristics and engagement outcomes. Chi-square tests and 2-tailed paired dependent-sample *t* tests were used to investigate potential changes in IPV outcomes between pretest and posttest.

Results: The average daily interaction rate with the program was 0.57 and 0.59 interactions per day for women and men in South Africa, and 0.21 for women in Jamaica. The rate of completion of at least 1 IPV prevention topic was 25% (5/20) for women and 5% (1/20) for men in South Africa, and 21% (6/28) for women in Jamaica. Exploratory analyses indicated significant pre-post reductions in overall IPV experience among women in South Africa (P=.01) and Jamaica (P=.01) and in men's overall harmful IPV attitudes (P=.01) and increases in men's overall gender-equitable behaviors (P=.02) in South Africa.

Conclusions: To the best of our knowledge, this is the first pilot study to investigate user engagement with and indicative outcomes of a digital parenting intervention with integrated IPV prevention content. Study findings provide valuable insights into user interactions with the chatbot and shed light on challenges related to low levels of chatbot engagement. Indicative results

XSL•FO
RenderX

suggest promising yet modest reductions in IPV and improvements in attitudes after the program. Further research using a randomized controlled trial is warranted to establish causality.

(JMIR Form Res 2025;9:e58611) doi:10.2196/58611

KEYWORDS

intimate partner violence; SMS text messaging; chatbot; user engagement; parenting; violence; mobile phone

Introduction

Background

Violence against women (VAW) and violence against children (VAC) are global issues with severe, long-lasting consequences, which affect individuals and communities worldwide. Global reports have revealed alarmingly high rates of both forms of violence [1,2]. Prevalence estimates of intimate partner violence (IPV), which is the most common form of VAW, have found that >27% of ever-partnered women aged >15 years have experienced physical IPV, sexual IPV, or both at least once in their life [1]. Reports of VAC are also of great concern, with systematic review findings suggesting that >50% of all children worldwide, that is, >1 billion children globally, have experienced past-year violence [3,4]. Research findings also indicate that prevalence rates are higher for both forms of violence in low- and middle-income countries [5,6].

A growing body of evidence shows that IPV and VAC frequently cooccur in the same families [7]. In recent years, the intersections between IPV and VAC have received increasing attention, with a growing demand for prevention efforts that target both forms of violence concurrently [8]. Intersections shared by both IPV and VAC include common risk factors, such as acceptability of family violence and parental history of physical abuse [9,10], similar short- and long-term consequences, including mental and physical health problems [11], and various intergenerational effects [12]. In the past, despite their overlaps, the fields of IPV and VAC prevention research have predominantly been separated [8]. Therefore, initiatives that seek to address these common risk factors may help shift violence prevention efforts from a siloed approach to one that targets multiple types of violence simultaneously. By targeting multiple and intergenerational forms of violence, this strategy is likely to be more sustainable long-term, helping to reduce costs and increasing opportunities for scale-up and impact [13].

In the field of IPV prevention, researchers and practitioners are increasingly adopting what is known as gender-transformative approaches in programs that aim to address multiple forms of violence, that is, both VAW and VAC [14]. More specifically, gender-transformative approaches can be defined as those that aim to address harmful social and gender norms and are "designed specifically to encourage men and boys to adopt and enact gender-equitable, nonviolent attitudes and behaviors" [15]. Many programs that adopt this approach include activities and modules that focus on topics such as improving interpersonal skills, redefining responsibilities in the home, and changing perceptions around gender roles and violence [14]. Emerging evidence suggests that gender-transformative approaches offer promising results in preventing IPV perpetration, shifting restrictive gender norms, and promoting gender-equitable behaviors and attitudes [16-18]. Gender-transformative approaches are being adopted in family and sexual health interventions, further highlighting the shift toward a more holistic approach toward violence prevention and recognition of the intergenerational elements of family violence and shared risk factors [19,20].

Alongside gender-transformative interventions, in the field of VAC, parenting programs have been identified as a major strategy for reducing and preventing violence [21]. Findings from a recent systematic review of parenting programs for reducing child maltreatment in low- and middle-income countries identified several promising results across child, parent, and family outcomes, with meta-analyses results revealing reductions in child maltreatment and harsh parenting [22]. One program that has integrated both VAC and IPV prevention content includes the gender-transformative Bandebereho program in Rwanda, which engaged men and partners with the aim of encouraging healthier couple relationships and enhancing caregiving skills [23]. The program used fatherhood as an avenue to promote gender equality and foster positive shifts in men's relations with both their children and partners [18]. A randomized controlled trial (RCT) of the program found promising short- and long-term effects, with reductions in IPV and physical punishment of children at both 21 months posttest and at a 6-year follow-up [18].

Despite the promising potential of parenting programs to tackle multiple forms of family violence, scaling up parenting interventions remains a challenge [24]. To address barriers to scale, digital parenting interventions have been gaining momentum as a means to increase reach, engagement, and accessibility [25]. One such program, developed by Parenting for Lifelong Health [26,27], is a chatbot intervention called ParentText that provides caregivers with social learning–based parenting content that seeks to improve parent-child interaction and prevent child maltreatment [28]. Given the emerging evidence based on the potential to integrate parenting and gender-transformative approaches to prevent multiple forms of family violence, this study seeks to explore how users engage with a parenting chatbot that contains integrated IPV prevention material. Building on the development of IPV prevention material designed and integrated into the ParentText intervention in 2021 [28], the purpose of this pilot study was to examine user engagement with the IPV prevention content in the chatbot as well as explore preliminary changes in IPV.

Objectives

The overarching aim of this pilot study is 2-fold. First, it seeks to assess parents' and caregivers' level of engagement with the integrated IPV prevention material in the ParentText chatbot.

XSL•FO
RenderX

Second, it aims to conduct exploratory analyses of preliminary pre-post changes in experiences and perpetration of IPV and related outcomes, including attitudes toward IPV and gender-equitable behaviors following the intervention.

Consequently, this study seeks to answer the following research questions:

1. What is the level of user engagement with the IPV prevention content integrated into the parenting chatbot, ParentText, among parents and caregivers with low income in South Africa and Jamaica?
2. What, if any, are the preliminary changes in experiences and perpetration of IPV (primary outcomes) and related risk factors, including attitudes toward IPV and gender-equitable behaviors (secondary outcomes) among caregivers with low income in South Africa and Jamaica, after using the ParentText chatbot?

Methods

Study Setting

The study was conducted in 2 countries where IPV is particularly widespread: South Africa and Jamaica. Despite differing in their cultural contexts and geographical regions, South Africa and Jamaica share similar challenges surrounding high rates of IPV, making them comparable settings for exploring the integration of IPV prevention strategies in a parenting chatbot. South Africa has one of the highest rates of IPV globally, with studies revealing prevalence rates of 20% to 50% in which women reported having experienced IPV sometime during their life [29-31]. In Jamaica, rates of VAW are also alarming, with 24% of women having experienced physical violence, sexual violence, or both, from an intimate partner sometime in their life [32]. These high rates of violence underscore that urgent prevention efforts that seek to prevent and reduce violence are crucial. Moreover, research findings also indicate a high prevalence of patriarchal social norms and restrictive beliefs around women's roles in both Jamaica [33] and South Africa [34], further highlighting the harmful sociocultural norms prevalent in both contexts and the overlapping challenges that both countries face. In Jamaica, this study was conducted in the parishes: Kingston, St Catherine, and St Elizabeth; and in South Africa, this study was conducted in Cape Town.

Participants and Recruitment

In Cape Town, South Africa, participants were recruited through convenience sampling by local research assistants through the organization Clowns Without Borders South Africa. Recruitment mainly took place in neighborhoods, townships, and communities in urban Cape Town. In Jamaica, recruitment also took place through convenience sampling by members of UNICEF (United Nations Children's Fund) Jamaica staff members in partnership with the National Parenting Support Commission, an agency of the Jamaican Ministry of Education and Youth. Participants were recruited primarily at schools and at women's centers for adolescent mothers in urban and rural communities. While the aim was to recruit both men and women, due to resource constraints, in the end, only women were recruited in Jamaica. Furthermore, while originally, the study sought to only include parents aged >18 years, following an interest expressed by UNICEF Jamaica and the Jamaican National Parenting Support Commission to address challenges faced by adolescent parents in the country, additional procedures were put in place (refer to the Ethical Considerations section for more details) to include adolescent parents aged >16 years in Jamaica. Unfortunately, due to staff and resource limitations in Cape Town, it was only possible to recruit parents aged >18 years in South Africa.

Consequently, the inclusion criteria for participants in this study were as follows: (1) age >18 years in South Africa and >16 years in Jamaica (where adolescent parents aged >16 years were also recruited), (2) currently caring for a child aged between 0 and 17 years, (3) being in a relationship (defined as either having a partner or being married—refer to Multimedia Appendix 1 for further details), (4) having access to a phone that could receive messages on WhatsApp (Meta Platforms) or Telegram (Telegram Messenger Inc) and that could connect to 3G or the internet, (5) provided consent to participate in the study, and (6) being able to speak either English or isiXhosa (South Africa only). Exclusion criteria were participants who were not currently in a partnered relationship and participants who were not caregivers, parents, or currently caring for a child. As an initial pre-post evaluation, this study was not designed to detect significant intervention effects but rather to assess user engagement and conduct exploratory analyses of intervention outcomes [35]. Due to funding limitations, the sample size of the study was restricted to 86 participants, with some lost due to drop out (Figure 1). Nevertheless, the study used a G*Power 3 calculator with a sensitivity power analysis to calculate the effect size (Cohen *d*) needed to obtain a significant intervention effect. Input parameters included the use of 2-tailed paired *t* tests based on the study's primary outcomes. Assuming a type I error of $P<.05$, 80% power, this sample size was sufficiently powered to detect a moderately significant intervention effect of Cohen $d=0.67$.

XSL•FO
RenderX

Figure 1. Theory of change for the ParentText intimate partner violence (IPV) prevention content, including risk factors, program components, immediate-term outcomes, and long-term outcomes.

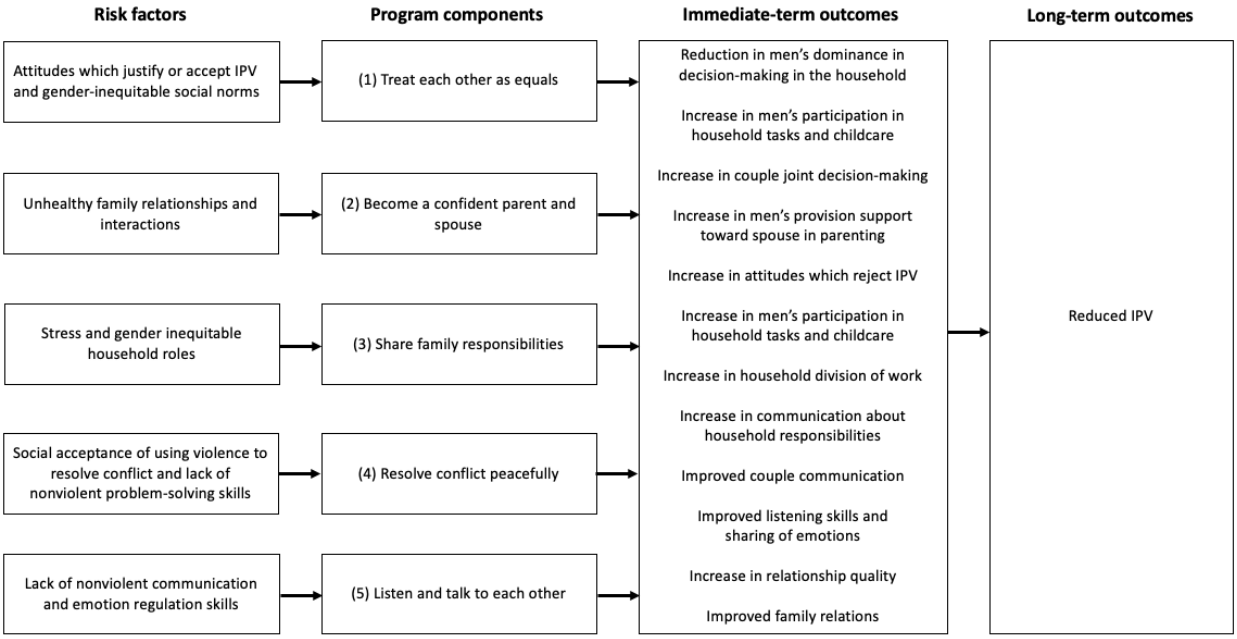

Intervention

ParentText is a chatbot intervention for parents and caregivers of children aged between 0 and 17 years [28]. The technical architecture is open source and available for developers through GitHub. The digital parenting intervention sends automated messages to users through social messaging platforms, such as WhatsApp, Telegram, and Facebook messenger (Meta Platforms), and is also available via SMS text messaging for individuals without smartphone access (for an example of program components and example messages, refer to Figures 1-3 and Multimedia Appendix 2). The parenting content of ParentText is derived from the in-person Parenting for Lifelong Health programs [26,27]. Enrolled users receive 23 days of daily ParentText messages personalized for their chosen child's developmental stage (0-23 months, 2-9 years, and 10-17 years). Content is provided using a range of formats, including text, image, video, and audio. Structured and sequential parenting content is derived from common elements of social learning theory-based programs along two main themes as follows: (1) positive relationship building and (2) limit setting and nonviolent discipline [36]. In addition, content specifically targeting positive partner relationships and IPV prevention for users in partnered relationships was integrated into the chatbot by drawing on gender-transformative material from a range of interventions found to reduce IPV and improve gender-equitable attitudes [23,37,38]. It is the engagement with this content that is the primary focus of this study. The IPV prevention content in ParentText is divided into 5 topics (refer to program components in Figure 1). The proposed causal pathway between each IPV prevention topic, identified risk factor, and behavior change domain is illustrated in the theory of change in Figure 1. Schafer et al [28] and Multimedia Appendix 3 give further details on the development of the IPV prevention content and theory of change. Additional information and examples on how parents can interact with the chatbot, along with examples of the sequencing of the ParentText messages, are available in Figures 2 and 3.

XSL•FO
RenderX

Figure 2. Flow of the intimate partner violence (IPV) prevention content text messages in ParentText, which illustrates the flow of text messages that users in ParentText receive, including introduction text messages and text messages providing options to view further IPV prevention content.

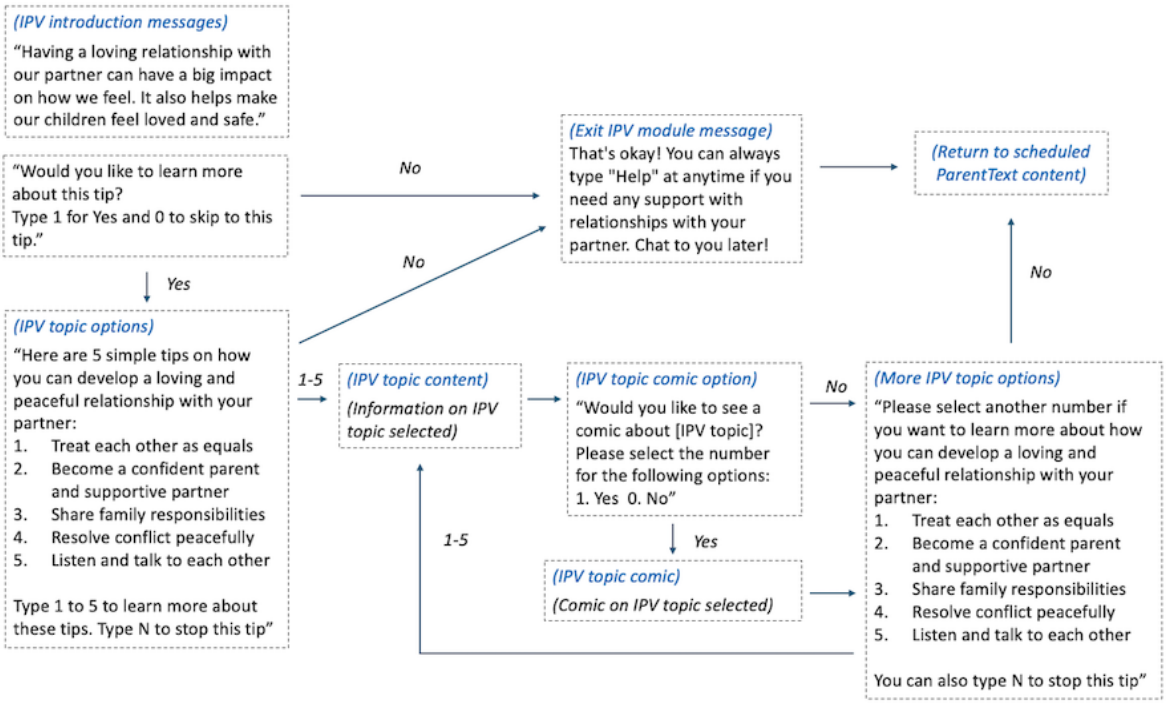

Figure 3. Flow of the check-in message on the intimate partner violence (IPV) topic 2 "joint decision-making," which illustrates the flow of the check-in text message that users in ParentText receive to provide an opportunity to reflect on the topic of "joint decision-making.".

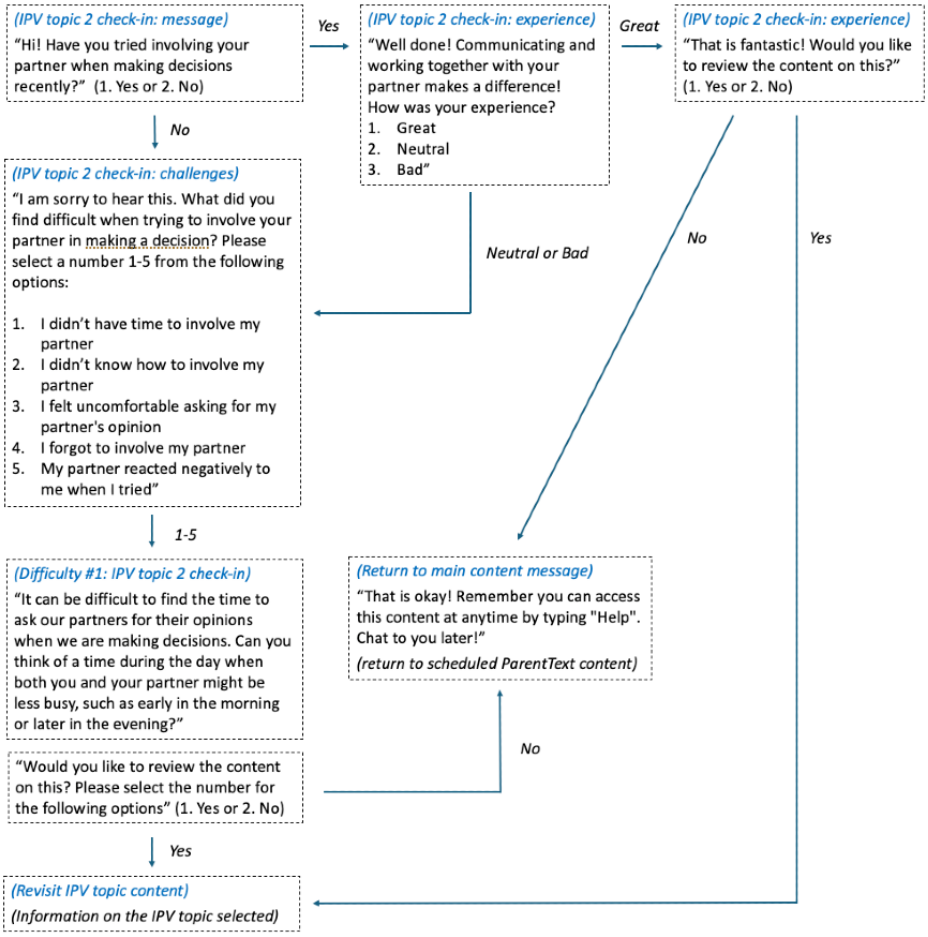

XSL•FO
RenderX

Design and Procedures

This study used a repeated-measures, single-arm design. Pre and posttest assessments were conducted at baseline (preintervention) and at 6 weeks after baseline (ie, 3 weeks after the end of the intervention) using the application Open Data Kit (ODK; refer to the Measures section and Multimedia Appendix 1 for more details). The posttest was planned as a 1-month posttest (1 month after the end of intervention); however, due to fieldwork challenges that arose and time constraints, it needed to be changed to 6 weeks after baseline (Figure 4).

Figure 4. Overview of the ParentText chatbot timeline, illustrating what content is delivered on what days of the program from day 1 to day 23, and demonstrating when the 3-week posttest assessment is delivered. IPV: intimate partner violence.

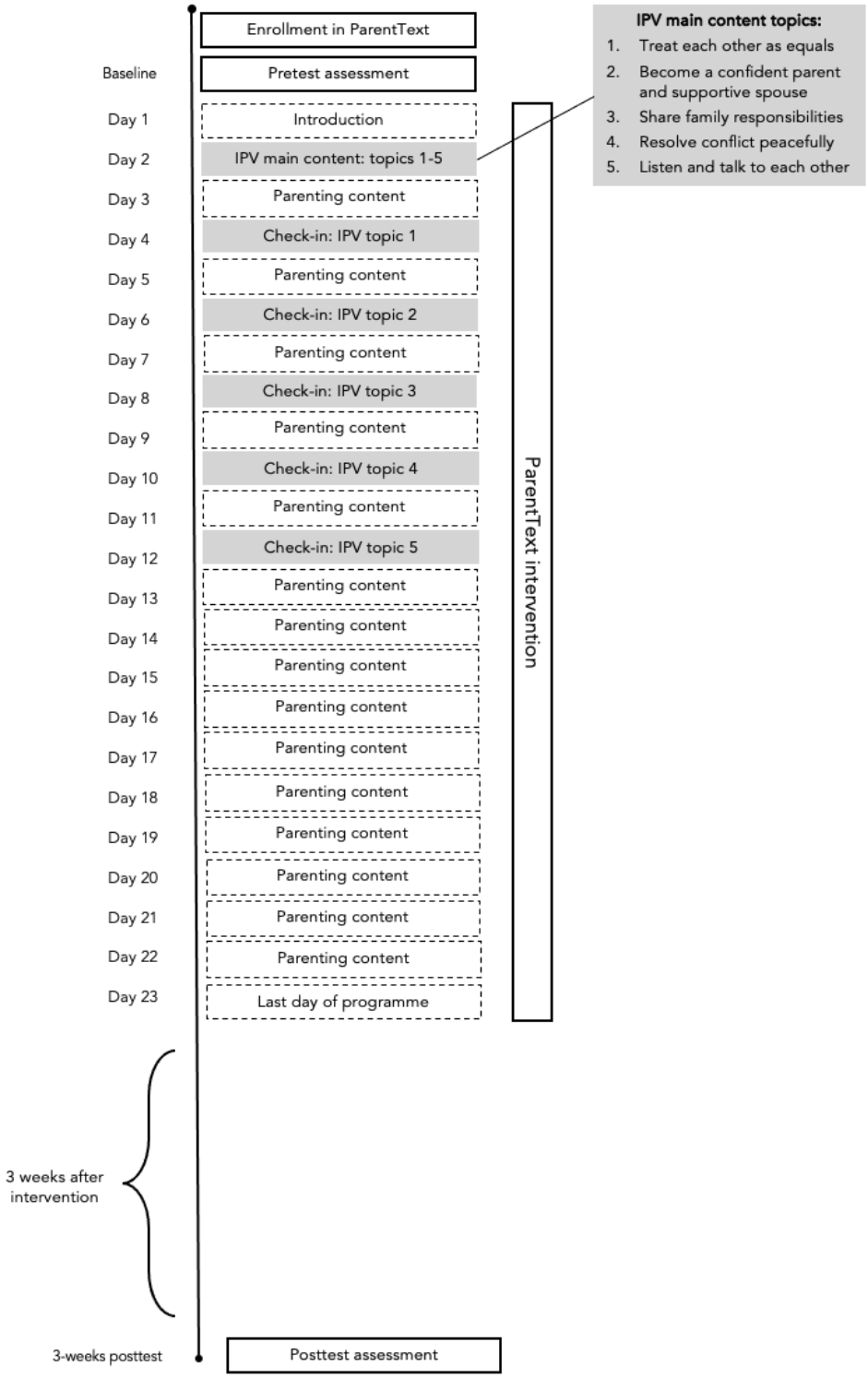

Reporting follows the CONSORT (Consolidated Standards of Reporting Trials) 2010 guidelines extension to randomized pilot and feasibility trials [39] (refer to Multimedia Appendix 4 for the CONSORT checklist). Recruitment, delivery, and

XSL•FO
RenderX

assessments in South Africa and Jamaica were conducted on a rolling basis from September 2022 to March 2023. Participants who enrolled in the intervention received data credit using a local service provider to ensure they would have access to 3G during the intervention. During recruitment, participants received a message that included a link to the program embedded within a WhatsApp or Telegram business account. After clicking the link, participants were directed to the ParentText chatbot account and asked to select their preferred language for the program. In keeping with recommendations from the formative research [28], the language was set to English in Jamaica, and the options were English and isiXhosa in South Africa. Participants were asked if they consented to participate in the intervention and evaluation and were provided with a link to a more detailed information sheet. Participants who did not consent exited the intervention. In alignment with guidelines for conducting statistical surveys on VAW [40], the survey included skip patterns, which determine the eligibility of specific questions that respondents are asked [40,41]. For example, in this study, men who responded that their partner was taking part in the intervention were not asked the IPV items in the IPV assessment and were instead only asked the questions on attitudes toward IPV and gender roles and on gender-equitable behaviors. Further information about this is provided in the Ethical Considerations section.

At baseline, participants were invited to complete the IPV assessment, which was optional, on an external server via Oxford's Linux virtual machine using ODK to ensure responses were not saved on their mobile devices. Participants provided their phone number when completing the IPV survey for their responses to be recorded and linked with a unique user ID. These data were then deidentified, exported as CSV files, and uploaded to a secure server in Oxford, which was only accessible by the research team and password protected. All personal identifying data were deleted once endline data collection was completed. Participants were invited to respond to the IPV assessment at 3 weeks postintervention via the phone number they provided when completing the IPV assessment at baseline.

Measures

The user engagement, demographic, primary outcomes, and secondary outcomes are provided in Table 1.

XSL•FO
RenderX

Table 1. Outcome measurements in the study, including user engagement, demographics, background variables, and primary and secondary outcomes, with descriptions of the study measurement used and the number of items used for each outcome.

Outcomes	Study measurement	Items, n
User engagement outcomes		
Number of days in the ParentText program	Measured by the final day the participant used the chatbot subtracted by the first day participant used the chatbot (including inactive days)	1
Number of active days the user actively interacted with the chatbot	Number of days participants actively interacted with ParentText as measured by chatbot interactions (excluding inactive days)	1
Overall chatbot interaction rate	The number of interactions logged per day divided by the total number of days participants actively interacted with ParentText	1
Participant engagement with each IPV[a] topic	The number of participants who viewed each IPV topic, in line with existing digital parenting intervention research where module viewing is frequently used as a proxy measure for the level of participant engagement [42,43]	5
Completion rate of IPV topics	Number of IPV topics completed out of the total 5 topics (measured by the number of IPV topics participants viewed that were recorded via RapidPro, UNICEF's open source framework developed to send and receive data using mobile phones [44])	5
Overall completion rate of ParentText parenting modules	Number of parenting modules completed (measured by the number of modules viewed that were recorded via the RapidPro (UNICEF) software)	1
Primary behavioral outcomes		
Women's experiences of IPV	Items adapted from the WHO[b] VAWI[c] and the core questionnaire in the WHO-MCS[d] [45]: physical (1 item), psychological (1 item), sexual (1 item), coercion (1 item), and economic abuse (1 item). Participants asked to give a frequency score on a scale of 0 to 8 or more times. The original VAWI tool measures past-year experiences of IPV [45]; however, a 1-year reporting time frame surpassed the duration of the intervention. The reporting timeframe and response codes were therefore adapted from the past-year (once, a few times, and many times) to the number of experiences of IPV in the past month (0, 1, 2, 3, 4, 5, 6, 7, and >8) to enhance the sensitivity of measuring changes in experiences of IPV. These adjustments are similar to those made to measure child maltreatment with the ISPCAN[e] Child Abuse Screening Tool (ICAST)-Trial measurement by Meinck et al [46]. Past-month time frame when measuring IPV has been used in other similar pieces of research, such as a study by Ebert and Steinert [47], which examined experiences of IPV during COVID-19 in Germany. Responses were dichotomously recoded, where 1 indicated any experience of each subsequent type of IPV and 0 indicated no experience of each type of IPV in the past month [48].	5
Men's perpetration of IPV	Items adapted from the WHO VAWI and the core questionnaire in the WHO-MCS [45]: physical (1 item), psychological (1 item), sexual (1 item), coercion (1 item), and economic abuse (1 item). Participants were asked to give a frequency score on a scale of 0 to >8 times, with responses dichotomized into any perpetration of IPV in the past month.	5
Secondary behavioral outcomes		
Men's experiences of IPV	Two items used and adapted from Abramsky et al [49], namely: "In the past month, during any potential times that you may have used violence against your partner, did they ever fight back physically to defend themselves?" and "In the past month, have you ever been hit or physically mistreated by your partner when you were not hitting or physically mistreating them?" Respondents were asked to give a frequency score on a scale of 0 to 8 or more times, with responses dichotomized.	2
Women's perpetration of IPV	Two items were adapted from a questionnaire developed by researchers at the London School of Hygiene and Tropical Medicine for a randomized controlled trial of a violence prevention intervention in Tanzania [49]. Items included: "During any potential times that you were hit in the past month, did you ever fight back physically to defend yourself?" and "In the past month, have you ever hit or physically mistreated your partner when they were not hitting or physically mistreating you?" Respondents were asked to give a frequency score on a scale of 0 to 8 or more times, with responses dichotomized.	2
Women's positive partner interactions	Item adapted from Abramsky et al [49] asking: "How many times in the past month did your partner show you they cared and respected your feelings even though they disagreed with you?" Participants were asked to give a frequency score on a scale of 0 to 8 or more times.	1
Men's positive partner interactions	One item adapted from Abramsky et al [49], that asks, "How many times in the past month did you show your partner you cared and respected their feelings even though you disagreed?." Participants asked to give a frequency score on a scale of 0 to 8 or more times.	1

XSL•FO
RenderX

Outcomes	Study measurement	Items, n
Gender-equitable behaviors	Measured using items adapted from questionnaires used in previous violence prevention interventions [23,49]. In total, 3 items were adapted from Abramsky et al [49], which included questions on couple communication, joint decision-making, and partner conflict resolution, with questions such as: "In the past week, how many times did you and your partner talk about your worries and feelings?" and "In the past week, how many times did you and your partner make a decision together?" The fourth item was adapted from a questionnaire developed by researchers of a randomized controlled trial of a gender-transformative violence prevention intervention in Rwanda [23], which asks: "In the past week, how many times did you and your partner share housework and caregiving tasks equally?" Respondents were asked to report whether in the past week the specific behavior occurred on a scale of 0 (never) to 3 (many times). The scores were summed for all items for the overall gender-equitable behavior score. The possible overall score ranged from 0 to 12.	4
Attitudes toward gender roles and IPV	Measured using 6 items in the "Attitudes toward gender roles" section of the WHO-MCS [45], such as "A woman should obey her husband's wishes even if she disagrees." Respondents were asked to indicate whether they agree or disagree with the statements using a Likert scale ranging from 1 (strongly agree) to 5 (strongly disagree). Negative items were reverse-coded so that higher scores denoted more gender-equitable attitudes. The scores were summed for all items for the overall attitude score. The possible overall score ranged from 6 to 30.	6
Background variables		
Parent gender, age, relationship status, living with partner, education level, employment status, partner employment status, and number of children	Sociodemographic information	8 (1 item each)
Women's past-year IPV experience (measured at baseline)	Measured using 1 item adapted from the WHO VAWI and the core questionnaire in the WHO-MCS [45], which asks, "How many times, in the past 12 months, did your partner do one of the following: insult or shout at you, hit or shove you, or force you to have sex?" Respondents were asked to give a frequency score on a scale of 0 to 8 or more times. Due to the skewed response patterns in the data for IPV, responses for IPV were dichotomously recoded, where 1 indicated any experience of IPV and 0 indicated no experience of IPV in the past year [48].	1

[a]IPV: intimate partner violence.

[b]WHO: World Health Organization.

[c]VAWI: Violence Against Women Instrument.

[d]MCS: Multi-Country Study on Domestic Violence.

[e]ISPCAN: International Society for Prevention of Child Abuse and Neglect.

Statistical Analysis

Data analyses were conducted using R (R Core Team). Descriptive statistics and frequencies were used to describe the sociodemographic characteristics of participants and to report engagement outcomes, summarized using means (SDs), medians (range), and N (%). To investigate changes in outcomes between pre and posttest, McNemar chi-square tests (for dichotomous variables) and paired dependent-sample *t* tests (for continuous variables) were conducted, with corresponding 95% CIs and *P* values [50,51]. Assumptions were examined to confirm no violations existed. Data were analyzed using complete case analysis.

Ethical Considerations

Ethical procedures were followed to ensure the safety and welfare of study participants. At the start of the IPV assessment, screening procedures ensured that men and women from the same household were not interviewed about IPV. This is based on the World Health Organization VAW research guidelines [52] and ethical and safety recommendations for intervention research on VAW [45] to prevent putting women at risk. Consequently, men who responded that their partner was taking part in the intervention were not asked the IPV items in the IPV assessment and were only asked instead the questions on attitudes toward IPV and gender roles and on gender-equitable behaviors. In addition, efforts were also in place to ensure responses to the IPV questions could not be viewed by others by delivering the IPV assessment using ODK. This procedure was implemented to protect participants by ensuring no responses were saved on their mobile devices. Participants who disclosed experiences of IPV were automatically provided referrals to local services, customized to their regional context that supported individuals experiencing violence. Local referral information could also be accessed by writing "Help Me" and choosing "Other Support" in the ParentText free text field. In line with UNICEF risk communication and community engagement guidelines [53], ParentText is also formatted to detect high-risk keywords to identify potential disclosure of dangerous situations via the free text field. Following detection, ParentText automatically offers the participant an empathetic and empowering reply with referral details that are customized to the country, which supports parent and child safety (eg, hotlines, police, and ambulance).

XSL•FO
RenderX

Further ethical considerations that were taken within the research project include recognitions surrounding power and limitations of positionality [54]. For instance, it is worth noting that while collaborators and members of the research team came from several countries, the primary author of this paper is from and is part of a research institution in the global north. Efforts to address limitations associated with this positionality were put in place, including integrating reflexivity across the stages of the research process [55]. For example, during the research design and data collection planning phase, members of local grass-roots organizations, communities, and stakeholders in Jamaica and South Africa were continuously consulted to review and revise the IPV prevention content and assessment materials [28]. Integrating these embedded consultations and opportunities for research amendments with members of the local community was a pivotal part of the research project, particularly given the first author's positionality as an "outsider" [55].

Ethics approval was obtained by the Department Research Ethics Committee at the University of Oxford (Central University Research Ethics Committee 2 Ref No: R69569/RE009). All methods were carried out in accordance with relevant guidelines and regulations. Informed consent to participate in the study was obtained from all participants before enrolling in the study and the intervention.

Results

Participants Characteristics

Participant characteristics are provided in Table 2 by country and with an overall summary. A study flow diagram illustrating the flow of participants through the trial is provided in Figure 5. The average age of participants across both countries was 34.4 (SD 8.74; range 16-57) years. The average age of women in South Africa was 34.9 (SD 6.77; range 25-49) years and men 28.8 (SD 2.68; range 25-36) years. In Jamaica, where only women were recruited into the study, the average age of women was 39.9 (SD 11.52; range 16-57) years. Most participants, across both countries, were partnered but not married (48/68, 71%). On average, participants had 2.23 (SD 1.19) children. Approximately half of the participants across both countries had a secondary education (36/68, 53%), and 7% (5/68) had only primary education or less. In South Africa, among women participants, 42% (8/19) were working and 58% (11/19) were unemployed or looking for work. Among men, slightly more were working (13/21, 62%), with 29% (6/21) unemployed or looking for work. In Jamaica, among women participants, this was similar, with 43% (12/28) working and 54% (15/28) unemployed or looking for work. Across both countries, the employment of partners was high (50/68, 74%), with only 18% (12/68) with partners who were unemployed or looking for work. Experience of IPV (women-report only due to safety reasons) at baseline was high: 21% (4/19) of women participants in South Africa and 29% (8/28) of women in Jamaica had experienced some form of IPV in the past year, respectively.

XSL•FO
RenderX

Table 2. Demographic characteristics of study participants in South Africa and Jamaica.

	South Africa		Jamaica	Overall
	Women (n=19)	Men (n=21)	Women (n=28)	Total (N=68)
Gender, n (%)				
Women	19 (48)	N/A[a]	28 (100)	47 (69)
Men	N/A	21 (53)	__[b]	21 (31)
Age (y), mean (SD)	34.9 (6.77)	28.8 (2.68)	39.9 (11.52)	34.4 (8.74)
Relationship, n (%)				
Married	6 (32)	2 (10)	12 (43)	20 (29)
Partnered but not married	13 (68)	19 (90)	16 (57)	48 (71)
Number of children, mean (SD)	2.53 (1.17)	1.19 (0.68)	2.61 (1.25)	2.23 (1.19)
Education, n (%)				
Primary	1 (5)	0 (0)	4 (14)	5 (7)
Secondary	9 (47)	10 (48)	17 (61)	36 (53)
Higher	9 (47)	11 (52)	7 (25)	27 (40)
Employment, n (%)				
Working	8 (42)	13 (62)	12 (43)	33 (49)
Unemployed or looking for work	11 (58)	6 (29)	15 (54)	32 (47)
Student	0 (0)	2 (10)	0 (0)	2 (3)
Retired	0 (0)	0 (0)	1 (4)	1 (2)
Partner's employment, n (%)				
Working	11 (58)	17 (81)	22 (79)	50 (74)
Unemployed or looking for work	7 (37)	2 (10)	3 (11)	12 (18)
Student	0 (0)	2 (10)	0 (0)	2 (3)
Retired	0 (0)	0 (0)	2 (7)	2 (3)
Refuse to answer	1 (5)	0 (0)	1 (4)	2 (3)
Experience of IPV[c], n (%)				
Experienced IPV in the past year[d]	4 (21)	X[e]	8 (29)	12 (26)[f]

[a]N/A: not applicable.
[b]No men were recruited in Jamaica.
[c]IPV: intimate partner violence.
[d]For safety reasons, men were not asked this question at baseline.
[e]Men were not asked this question at baseline due to safety reasons.
[f]Overall IPV here only includes women.

XSL•FO
RenderX

Figure 5. Study flow diagram illustrating the flow of participants through the trial. CONSORT: Consolidated Standards of Reporting Trials.

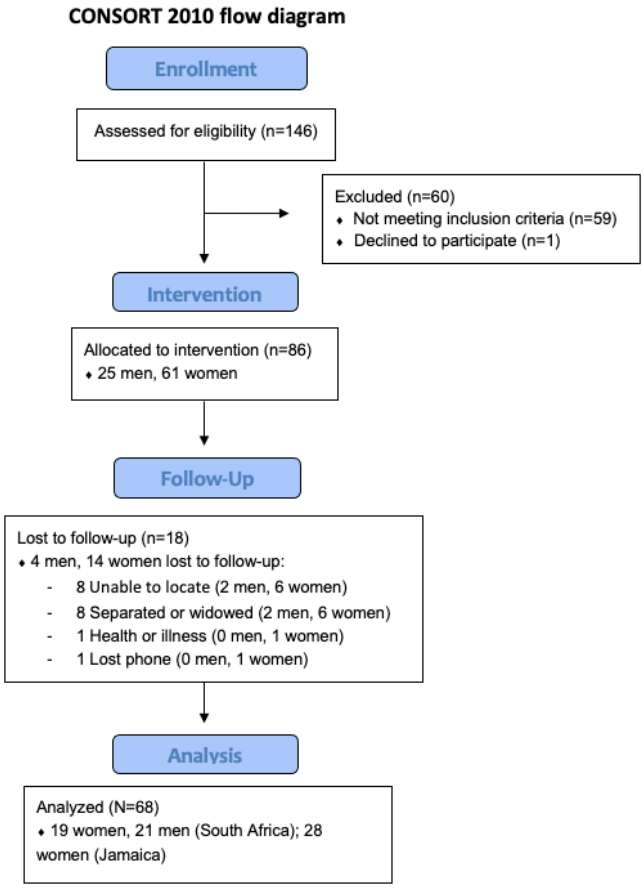

User Engagement Outcomes

Overall engagement outcomes are provided in Table 3. The ParentText program lasted 23 days, with the main IPV prevention content delivered on day 2 and check-in messages delivered within the first 12 days (Multimedia Appendix 2 provides an overview of the program content delivered each day). In terms of engagement, in South Africa, the average number of days in the program was 13 (SD 23.90) and 3 (SD 3.78) days for women and men, respectively. The average number of days users interacted with the chatbot was 3 (SD 1.58) for women and 2.6 (SD 1.52) for men, each with an average interaction rate of 0.57 (SD 0.26) interactions and 0.59 (SD 0.17) interactions per day, respectively. In Jamaica, the average number of days in the program was 27 (SD 23.10) days, and the average total number of days users interacted with the chatbot was 5.2 (SD 5.45) days. The average interaction rate in Jamaica was 0.21 (SD 0.20) interactions per day. In terms of engagement across the 5 relationship topics, there was a greater engagement with topics in Jamaica where engagement with each topic ranged from 11% (3/28) to 29% (8/28) compared to South Africa, where it ranged from 0% to 11% (2/19). In terms of completion rates, in South Africa, 25% (5/20) and 5% (1/20) of women and men completed 1 out of the 5 IPV topics, and 5% (1/20) and 0% completed all 5 IPV topics, respectively. In comparison to the completion rate of overall ParentText modules, 16% (3/19) and 9% (2/21) of women and men completed all the parenting (ie, the non-IPV) ParentText modules in South Africa. In contrast, in Jamaica, 21% (6/28) completed 1 out of the 5 IPV topics, 11% (3/28) completed all 5 IPV topics, and 18% (5/28) completed all the parenting ParentText modules.

XSL•FO
RenderX

Table 3. ParentText user engagement outcomes of the 23-day program in South Africa and Jamaica, including overall retention, engagement per intimate partner violence (IPV) topic, and IPV topic completion rate.

Outcomes	South Africa		Jamaica
	Women (N=19)	Men (N=21)	Women (N=28)
Overall retention, mean (SD)			
Number of days in program[a]	13 (23.90)	3 (3.78)	27 (23.10)
Number of active days the user interacted with the chatbot[b]	3 (1.58)	2.6 (1.52)	5.2 (5.45)
Number of interactions per day[c]	0.57 (0.26)	0.59 (0.17)	0.21 (0.20)
Engagement per IPV topic[d]			
Participants who viewed each IPV topic, n (%)			
Relationship topic 1: “Treat each other as equals”	2 (11)	2 (10)	7 (25)
Relationship topic 2: “Become a confident parent and supportive spouse”	2 (11)	0 (0)	4 (14)
Relationship topic 3: “Share family responsibilities”	2 (11)	1 (5)	5 (18)
Relationship topic 4: “Resolve conflict peacefully”	2 (11)	1 (5)	3 (11)
Relationship topic 5: “Listen and talk to each other”	2 (11)	1 (5)	8 (29)
All relationship topics (viewed all relationship topics)	1 (5)	0 (0)	3 (11)
Completion rate of IPV topics completed, n (%)[e]			
1	5 (25)	1 (5)	6 (21)
2	0 (0)	0 (0)	3 (11)
3	0 (0)	0 (0)	1 (4)
4	0 (0)	1 (5)	0 (0)
5	1 (5)	0 (0)	3 (11)
Completion rate of ParentText parenting modules, mean (SD)	16 (0.19)	9 (0.06)	18 (0.20)

[a]Measured by the final day participant used the chatbot, subtracted by the first day the participant used the chatbot (note: this measure includes inactive days).

[b]Number of days participants actively interacted with ParentText (note: this measure excludes inactive days).

[c]Measured by the number of interactions logged per day divided by the total number of days the participant actively interacted with ParentText.

[d]Measured based on the number of participants who viewed each IPV topic.

[e]Number of IPV topics completed out of the total 5 topics.

Behavioral Outcomes

Experiences and Perpetration of IPV

Table 4 presents the descriptive results of the IPV outcomes, summarizing experiences of IPV among women and perpetration of IPV among men using past-month frequencies at pre and posttest and *P* values for each outcome. At posttest, 73% (11/15) of women in South Africa and 72% (13/18) of women in Jamaica reported experiencing some form of IPV, compared to 84% (16/19) and 96% (23/24) at baseline, reductions that were significant in South Africa (*P*=.01) and in Jamaica (*P*=.01). Experiences of all subtypes of IPV for women in both South Africa and Jamaica also showed reductions at posttest, and so did acts of self-defense and perpetration (Multimedia Appendix 5). For perpetration of IPV among men, at posttest, 75% (9/12) of men in South Africa perpetrated some form of IPV, compared to 85% (11/13) at baseline, which was nonsignificant for overall perpetration of IPV (*P*=.06). Analyses of subtypes of IPV, however, found significant reductions in psychological (*P*=.04), physical (*P*<.001), and sexual (*P*<.001) violence among men in South Africa (Multimedia Appendix 5). Detailed tables of the primary and secondary outcomes are provided in Multimedia Appendix 5.

XSL•FO
RenderX

Table 4. Primary behavioral outcomes for women's past-month experience of overall intimate partner violence (IPV) in South Africa and Jamaica and men's past-month perpetration of overall IPV in South Africa.

	Baseline overall IPV, n (%)[a]	Posttest overall IPV, n (%)	*P* value[b]
Women's experience of overall IPV (past month)			
South Africa	16/19 (84)	11/15 (73)	.01
Jamaica	23/24 (96)	13/18 (72)	.01
Men's perpetration of overall IPV (past month)			
South Africa	11/13 (85)	9/12 (75)	.06

[a]% used is proportional to the baseline and posttest samples, respectively.

[b]McNemar chi-squared *P* value.

Positive Partner Interactions

Table 5 presents outcomes on positive partner interactions measured through respectful behaviors, summarized using means and SDs for pre and posttest along with *P* values. There were no significant effects detected for change in positive partner interactions reported by men or women.

Table 5. Secondary behavioral outcomes for positive partner interactions in the past month, overall attitude toward gender roles and intimate partner violence (IPV), and overall gender-equitable behaviors in the past week for women and men in South Africa and women in Jamaica.

	Baseline, mean (SD)	Posttest, mean (SD)	*P* value (*t* test)
Positive partner interactions			
South Africa			
Women (n=15)	4.74 (2.84)	4.14 (2.98)	.57
Men (n=12)	3.92 (2.50)	5.67 (2.46)	.10
Jamaica			
Women (n=15)	5.08 (2.60)	3.89 (2.54)	.14
Overall attitude[a] toward gender roles and IPV			
South Africa			
Women (n=14)	10.11 (2.71)	10.13 (3.89)	.98
Men (n=16)	11.90 (3.82)	8.88 (2.63)	.01
Jamaica			
Women (n=15)	11.39 (3.45)	11.89 (2.35)	.56
Overall gender-equitable behaviors[b]			
South Africa			
Women (n=15)	7.37 (2.95)	7.93 (3.53)	.63
Men (n=16)	7.81 (1.94)	9.75 (1.61)	.02
Jamaica			
Women (n=15)	7.14 (3.49)	7.39 (2.85)	.80

[a]A lower score indicates less harmful attitudes toward IPV and gender roles.

[b]A higher score indicates more gender-equitable behaviors in the past week.

Attitudes Toward (Harmful) Gender Roles and IPV

Table 5 summarizes means and SDs for the attitude outcomes at pre and posttest, with *P* values for each attitude outcome also provided. Analyses indicate a significant reduction in harmful attitudes toward IPV and gender roles reported by men in South Africa (*P*=.01); however, no effects were reported among women in South Africa (*P*=.98) or in Jamaica (*P*=.56).

Gender-Equitable Behaviors

Analyses found a significant increase in gender-equitable behaviors for men (*P*=.02; Table 5). No significant effects were detected for women in South Africa (*P*=.63) or in Jamaica (*P*=.80).

XSL•FO
RenderX

Discussion

Principal Findings

To the best of our knowledge, this pre-post pilot study is the first that examines user engagement and indicative outcomes in a digital parenting intervention with integrated IPV prevention content [22]. The study offers valuable, preliminary insights across 2 countries on user engagement with a digital parenting intervention with integrated IPV prevention, as well as provides exploratory, indicative results on IPV experience and perpetration, attitudes toward IPV, and gender-equitable behavior outcomes. First, we will discuss the key findings on levels of user engagement. While the average daily interaction rate was higher in South Africa (0.57 and 0.59 interactions per day for women and men, respectively) compared to Jamaica (0.21 interactions per day), the average number of days users stayed in the program in Jamaica was greater (27 days), compared to South Africa (13 and 3 days for women and men, respectively). Due to WhatsApp messaging restrictions, "push" messaging was paused after 24 hours of user inactivity, until user activity resumed. Given research findings that highlight the importance of messaging reminders in digital interventions [56], this restriction may have contributed to some of the low engagement levels observed.

The completion rate of engaging with at least 1 IPV topic in the program was lower than expected across the intervention, with completion rates of 5% (1/20) for men and 25% (5/20) for women in South Africa, and 21% (6/28) for women in Jamaica. The completion rates of all 5 IPV topics were much lower, with 0% of men and 5% (1/20) of women in South Africa, and 11% (3/28) of women in Jamaica completing all 5 IPV prevention topics. Even though user engagement and retention rates have been noted as a challenge in digital interventions [57], completion rates in this study were lower than anticipated in comparison to findings in the field. For example, a review of digital interventions for parents with young children found that attrition rates in digital parenting programs ranged from 30% to 50% [58]. Accordingly, the low completion rates of the IPV topics observed in ParentText raise questions about whether user engagement in the program was sufficient to change complex behaviors and deeply rooted attitudes surrounding violence and gender roles.

There are various possible reasons for the low levels of engagement with the IPV content observed in ParentText. First is the format of the program schedule. On day 2 of the ParentText program, participants were provided with a list of all 5 IPV prevention topics, which they could choose to explore. After selecting a topic to view, users were given the option to view one of the other IPV topics. However, it is possible that viewing >1 IPV prevention topic per day may have been overwhelming for users. Therefore, this may have inhibited users from engaging with multiple IPV topics in the program. Another element that may have affected engagement was the timing of the program "check-in" reminders. Participants received "check-in" messages in the latter part of the program (days 4-12; Multimedia Appendix 2), where they were encouraged to revisit and explore the IPV topics they had not yet engaged with. However, given the already low average number of days for which the participants remained in the program, it is likely that many participants did not receive these messages. These findings raise important questions on how to increase engagement with the IPV modules integrated into the ParentText. For example, amending the timing of content delivery might be one way to increase engagement. For example, a study assessing a text-messaging intervention focusing on promoting health relationships found that the timing of message delivery played a key role in program engagement and that if delivered at the wrong time, users might be less likely to engage with the text messages [59].

Taken together, these study results give valuable insight into how users engage with IPV content integrated into a chatbot and suggest that rather than interacting with the chatbot content daily, users appear to interact with the chatbot at their own pace. Consequently, providing more options for user-led selection of program content would be important for future iterations of the chatbot to ensure users can access and revisit the information they need more readily. The importance of user-led content delivery has also been underscored by other digital parenting intervention research. For instance, findings from a study on the digital fatherhood program "SMS4dads" revealed that the length of time between users receiving a text message with a link to program content to that of users clicking the link was, on average, 2.11 (SD 3.94) days [60]. This highlights how engagement with digital intervention content is often not instantaneous, but rather asynchronous and driven by users' schedules [60]. Accordingly, in future versions of ParentText, combining "push" notification reminders with user-led selection of program content might help increase user engagement. Notably, since the present trial of ParentText in South Africa and Jamaica, newer versions of the WhatsApp settings have been released, which allow future iterations of the program to override the 24-hour limit inhibiting "push" messages after user inactivity.

More broadly, the relatively low levels of user engagement found in this pilot study raise questions as to whether a digital modality is sufficient when delivering intervention content related to complex gender-equitable beliefs and behaviors. It may be the case that a hybrid approach, combining both digital content delivery and in-person sessions, might be necessary. For instance, a study on enhancing father engagement in parenting programs revealed that in terms of intervention delivery, fathers prefer less intensive and low-dose programs, including internet - based interventions [61]. However, a recent systematic review of interventions promoting gender equality noted that in various studies it was found that dialog was critical for driving change in gender norms [62]. The importance of dialog is further underscored by research examining men's engagement in family and health interventions, revealing that men value and often seek the opportunity to connect with other men to discuss parenting and fatherhood, even though cultural norms sometimes hinder this. Together, these findings suggest that a combination of digitally delivered content and in-person discussion opportunities might be needed. Various studies also suggest that using face-to-face recruitment and participatory, peer-engaged methods is an effective approach for engaging

XSL•FO
RenderX

parents, particularly fathers [63,64]. Research findings also suggest that providing opportunities to critically discuss and engage in dialog with others, especially on difficult topics, plays a valuable role not only in terms of engagement but also in terms of creating change and shifting harmful gender attitudes [17,62,64]. Taken together, it may be possible that low-intensity, 1-way engagement with a chatbot is insufficient to create change in more deeply rooted behaviors and attitudes among users and that more interactive engagement is needed to shift behaviors and beliefs. While these hybrid options and amendments may improve engagement, it should be noted that they will also increase intervention costs and limit scalability, which are some of the benefits of digital-only interventions [24].

It is also possible that the low level of user engagement with the IPV prevention topic was due to other factors, such as the format of the program. The IPV content in ParentText was only delivered on 1 specific day, after which only check-in messages were sent (refer to Figure 4 for an overview of the content delivery schedule). Receiving the option to view all the IPV content on the same day may potentially have been overwhelming for participants. This may have reduced user engagement rates due to intervention fatigue [65], which is cognitive or emotional weariness resulting from competing demands of intervention engagement and other burdens in daily life [66] and has been linked to intervention adherence. Considerations for future iterations of ParentText include (1) delivering all the content automatically (since optional and supplemental content in digital interventions is often associated with low engagement [67]) and (2) adjusting the timing of content and spreading out the delivery of the IPV content to allow mental rests in between new material, strategies that have been suggested to reduce intervention fatigue [65]. Notably, this is already an amendment that future iterations of ParentText are incorporating and that is currently being implemented in a trial of an updated version of ParentText in Malaysia [68]. This integrated approach of implementing IPV prevention content beyond a single module is also a strategy that is favored and advocated by gender-transformative approaches and is one that is being advocated more broadly in the parenting field [69].

The exploratory analyses of the behavioral outcomes, while only preliminary, are also important to discuss. In terms of IPV, women in Jamaica reported a significant reduction in overall experiences of IPV at posttest. These tentative findings are in line with results from other parenting and gender-transformative programs that have found reductions in both IPV experience at posttest and even at follow-up [18,23,37]. For example, findings from an RCT of a gender-transformative intervention in Rwanda found that, compared to the control, women in the intervention group reported reductions in both physical and sexual IPV following the intervention [23].

In contrast to the significant reductions in IPV experience among women, no significant effects for overall IPV perpetration were detected among men in South Africa. However, analyses of IPV subtypes found significant reductions in psychological violence, physical violence, and sexual violence among men in South Africa (Multimedia Appendix 5). The lack of change in overall IPV perpetration might be related to the low level of male engagement. Notably, the user engagement and the IPV completion rate of men in South Africa were 50% less than those of women. Hence, the limited interaction men had with the chatbot may not have been sufficient to change these behaviors.

It may also be the case that more interactive engagement is needed to shift these complex behaviors. For instance, emerging findings suggest that content modality can impact digital intervention outcomes [42], as highlighted in a recent study on ChattyCuz, a chatbot intervention that aims to support young women in navigating intimate relationships in South Africa [70]. Interestingly, an RCT of ChattyCuz found that while the gamified version of the chatbot (treatment 1) led to modest but significant reductions in IPV experience compared to those without treatment, in the narrative treatment, there was no effect on IPV experience [70]. The authors hypothesize that the lack of measurable effect on IPV experience in the narrative treatment arm of the study might be related to the lack of user interaction, that greater engagement with intervention content is necessary for behavior change, and that an information-only approach is insufficient [70]. These findings highlight how low levels of user engagement in ParentText might play a role in the lack of behavior changes detected and underscore that, amendments to the program modality might be needed to make the content more interactive to shift behaviors. Recent studies have also demonstrated the strong impact that interpersonal relationships have on gender attitudes, underscoring the need for interventions to not only target individuals but also social networks, such as peers and family members through, for example, critical reflection and dialog among social groups [71]. Research suggests that programs that have shown evidence of change in gender norms involve the engagement of stakeholders at different levels of the social-ecological model, include activities that allow for active participation, and promote critical awareness [14]. Consequently, it is possible that greater interactive engagement is needed to see greater changes in IPV experience and perpetration.

This study also revealed a significant overall reduction in harmful attitudes toward IPV as well as a significant increase in overall gender-equitable behaviors among men in South Africa. These indicative effects are promising, and while caution is warranted given the small sample, the findings align with other interventions that have observed reductions in harmful attitudes among men at posttest and detected increases in gender-equitable behaviors following gender-transformative interventions [72,73].

Strengths and Limitations

The study has various strengths worth noting. As the first study of a digital parenting intervention with integrated IPV prevention content, the present research makes a valuable contribution to the field of violence prevention and targeted efforts seeking to address VAC and IPV concurrently. With a growing emphasis in policy and research on preventing multiple forms of violence simultaneously [74], this study sheds light on preliminary behavioral outcomes and how users engage in IPV prevention content that has been integrated into a digital parenting intervention. Another strength of the study is the use of data from 2 countries. By delivering the program in 2 different

XSL•FO
RenderX

regions and cultures, this study offers an insight into important considerations to bear in mind before scaling up the intervention. For instance, the challenges with recruiting men in Jamacia compared to South Africa suggest that alternative recruitment strategies might be needed when scaling up the intervention in the Caribbean region.

The study also has various limitations. The first relates to the lack of a control group, without which causality cannot be established from the pre-post data. The study also had a small sample size, which means that the research findings need to be interpreted with caution given that there is limited statistical power in detecting pre-post intervention effects. Consequently, the next step is a fully powered RCT to establish effectiveness. Future research would also benefit from incorporating a factorial experiment design, which can be used to examine which specific intervention components have an effect on outcomes and which components interact with each other [75]. Knowing which components should be kept to sustain impact when taking an intervention to scale or delivering it in a new setting is critical, especially when considering program length and implementation costs [76]. Including follow-up assessments to examine whether intervention effects remain, become reduced, or are delayed over time [77] would also be beneficial in future studies as this study was limited to a 6-week posttest due to resource and time constraints. In addition, the study used a convenience sampling approach, which may limit the generalizability of the study findings. For example, it is possible that caregivers who opted to take part in the parenting program were more willing to change, which may bias the results. The outcomes in the study were also self-reported, and consequently, they may be subject to disclosure bias, recall bias, and social-desirability bias, leading to under reporting of unfavorable behaviors and over reporting of what might be perceived as desirable answers. However, given that the self-reports were carried out on participants' phones and participants were told the reports were fully anonymized, this will likely have reduced desirability bias in the quantitative surveys.

Another important limitation is that while in this study, participants were provided with mobile credit, this may not always be realistic in a real-world setting. While access to the internet and mobile data are increasing globally [78], it is important to explore potential options of being able to use the program without mobile data, especially when planning strategies for scale-up and delivery. One digital intervention that has explored this is the parenting program ParentApp for Teens, which is a mobile app designed for offline use that targets parents of teenagers [79,80]. There are also assessment limitations related to the self-reported measurements used in this study. Given the sensitive nature surrounding IPV and attitudes toward violence, it is possible that participants were hesitant to disclose violence due to stigma or fear of being endangered, leading to under reporting and consequently leading to undetected changes in behaviors [81]. Similarly, social-desirability bias may have also impacted the measurement of behavioral outcomes, particularly the measurement of IPV perpetration, which has been found to be more prone to social-desirability bias than reports of IPV experience [82].

An additional limitation worth noting is that men were underrepresented in the study. Recruitment of male caregivers is a known difficulty with parenting interventions [83], and even though it was possible to recruit both male and female caregivers in South Africa, it was only possible to recruit female caregivers in Jamaica. The lack of male caregivers from Jamaica means that certain experiences may have been underestimated due to the low number of fathers included. This challenge highlights the need for future studies to ensure greater efforts are taken to adopt recruitment strategies that ensure men are also included. In future studies, it would also be valuable to include people aged between 16 and 17 years in the South African sample to better understand how adolescent parents in South Africa engage with the chatbot.

Conclusions

This research presents tentative findings on the first-ever attempt to research the integration of IPV and parenting content into a digital intervention. Some of the preliminary results of the intervention are promising, with indications of reductions in overall IPV for women and a potential decrease in harmful IPV attitudes among men. However, the high rates of IPV experiences and perpetration reported at the posttest are still concerning and suggest that greater intervention efforts and programmatic amendments are likely necessary to tackle deep-rooted attitudes and harmful behaviors. In addition, the unexpectedly low levels of engagement in the intervention raise questions about whether interactions with the chatbot were sufficient to shift behaviors and attitudes and suggest that a hybrid approach to program delivery might be necessary. Given the urgent need for scalable prevention efforts that can tackle multiple forms of violence concurrently [84], more research is necessary to further explore how to increase chatbot user engagement and achieve greater reductions in IPV and harmful attitudes toward gender roles and IPV.

Acknowledgments

This research was funded in whole, or in part, by the United Kingdom Research and Innovation (UKRI; grant ES/P000649/1). For the purpose of open access, the author has applied for a CC BY public copyright license to any author accepted manuscript version arising from this submission. This research was also funded by UNICEF (United Nations Children's Fund) Jamaica and the Economic and Social Research Council Grand Union Doctoral Training Partnership in collaboration with the University of Oxford Saven European Scholarship and the Balliol College Dervorguilla Scholarship in Social Sciences. This study is part of the Global Parenting Initiative, which is funded by the LEGO Foundation, Oak Foundation, the World Childhood Foundation (grant 16191), the Human Safety Net, and the UK Research and Innovation Global Challenges Research Fund (grant ES/S008101/1). Aside from UNICEF's involvement in guiding the selection of the research site and study time frame, none of the funders were

XSL•FO
RenderX

involved in study design, data collection, data analysis and interpretation, or the writing of the manuscript. The authors would like to thank those who made this research possible: the respondents who participated in this study, UNICEF Jamaica, the research assistants, the study funders, the researchers and stakeholders who provided input during intervention development and delivery, and the Department of Social Policy and Intervention at the University of Oxford for administrative support. Generative artificial intelligence was not used in any portion of the manuscript writing.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

MS, JL, and FG participated in the design of the study and oversaw data collection and cleaning. MS managed the data collection process, with support from PZ, FACA, GHR, and RS. MS designed the statistical analysis, with JL, FG, QH, FACA, CF, and LC providing guidance on the analysis approach. MS drafted the paper, with JL and FG contributing. All authors approved the final manuscript.

Conflicts of Interest

MS is a researcher at the University of Oxford, and JL and FG are employed as staff at the university. In addition, JL is the chief executive officer of Parenting for Lifelong Health, a charitable organization based in the United Kingdom that developed the ParentText intervention. MS, JL, PZ, QH, FACA, GHR, and FG have participated and are participating in several research studies involving the program as investigators at the universities of Oxford, Cape Town, and Glasgow. CF and LC are employees of IDEMS International, a technology development company that codeveloped the ParentText intervention and provided support on data analysis of user engagement. RS is an employee at UNICEF and provided support on data collection. Conflict is avoided by declaring these potential conflicts of interest and by conducting and disseminating rigorous, transparent, and impartial evaluation research on both this and other similar parenting programs. No profits will be made from this program.

Multimedia Appendix 1
Intimate partner violence assessment.
[DOCX File , 28 KB - formative_v9i1e58611_app1.docx]

Multimedia Appendix 2
Structure of the ParentText intervention.
[DOCX File , 24 KB - formative_v9i1e58611_app2.docx]

Multimedia Appendix 3
Overview of ParentText intimate partner violence prevention content.
[DOCX File , 20 KB - formative_v9i1e58611_app3.docx]

Multimedia Appendix 4
CONSORT (Consolidated Standards of Reporting Trials) checklist.
[DOCX File , 374 KB - formative_v9i1e58611_app4.docx]

Multimedia Appendix 5
Detailed primary and secondary outcome tables.
[DOCX File , 46 KB - formative_v9i1e58611_app5.docx]

References

1. Violence against women prevalence estimates, 2018. World Health Organization. 2021 Mar 9. URL: https://www.who.int/publications/i/item/9789240022256 [accessed 2024-12-10]
2. Whitten T, Tzoumakis S, Green MJ, Dean K. Global prevalence of childhood exposure to physical violence within domestic and family relationships in the general population: a systematic review and proportional meta-analysis. Trauma Violence Abuse 2024 Apr 10;25(2):1411-1430 [FREE Full text] [doi: 10.1177/15248380231179133] [Medline: 37300288]
3. Hillis S, Mercy J, Amobi A, Kress H. Global prevalence of past-year violence against children: a systematic review and minimum estimates. Pediatrics 2016 Mar;137(3):e20154079 [FREE Full text] [doi: 10.1542/peds.2015-4079] [Medline: 26810785]
4. Global status report on preventing violence against children 2020. World Health Organization. URL: https://www.unicef.org/media/70731/file/Global-status-report-on-preventing-violence-against-children-2020.pdf [accessed 2024-12-10]

XSL•FO
RenderX

5. Global and regional estimates of violence against women: prevalence and health effects of intimate partner violence and non-partner sexual violence. World Health Organization. 2013 Oct 20. URL: https://www.who.int/publications/i/item/9789241564625 [accessed 2024-12-10]
6. Cerna-Turoff I, Fang Z, Meierkord A, Wu Z, Yanguela J, Bangirana CA, et al. Factors associated with violence against children in low- and middle-income countries: a systematic review and meta-regression of nationally representative data. Trauma Violence Abuse 2021 Apr;22(2):219-232 [FREE Full text] [doi: 10.1177/1524838020985532] [Medline: 33461441]
7. Pearson I, Page S, Zimmerman C, Meinck F, Gennari F, Guedes A, et al. The co-occurrence of intimate partner violence and violence against children: a systematic review on associated factors in low- and middle-income countries. Trauma Violence Abuse 2023 Oct;24(4):2097-2114 [FREE Full text] [doi: 10.1177/15248380221082943] [Medline: 35481390]
8. Guedes A, Bott S, Garcia-Moreno C, Colombini M. Bridging the gaps: a global review of intersections of violence against women and violence against children. Glob Health Action 2016 Jun 20;9:31516 [FREE Full text] [doi: 10.3402/gha.v9.31516] [Medline: 27329936]
9. Alhusen JL, Ho GW, Smith KF, Campbell JC. Addressing intimate partner violence and child maltreatment: challenges and opportunities. In: Korbin J, Krugman R, editors. Handbook of Child Maltreatment. Dordrecht, The Netherlands: Springer; 2014.
10. Gracia E, Rodriguez CM, Martín-Fernández M, Lila M. Acceptability of family violence: underlying ties between intimate partner violence and child abuse. J Interpers Violence 2020 Sep;35(17-18):3217-3236. [doi: 10.1177/0886260517707310] [Medline: 29294751]
11. Chisholm CA, Bullock L, Ferguson JE. Intimate partner violence and pregnancy: epidemiology and impact. Am J Obstet Gynecol 2017 Aug;217(2):141-144. [doi: 10.1016/j.ajog.2017.05.042] [Medline: 28551446]
12. Guedes A, Mikton C. Examining the intersections between child maltreatment and intimate partner violence. West J Emerg Med 2013 Aug;14(4):377-379 [FREE Full text] [doi: 10.5811/westjem.2013.2.16249] [Medline: 23997846]
13. Edwards KM, Kumar M, Waterman EA, Mullet N, Madeghe B, Musindo O. Programs to prevent violence against children in sub-Saharan Africa: a systematic review. Trauma Violence Abuse 2024 Jan;25(1):593-612 [FREE Full text] [doi: 10.1177/15248380231160742] [Medline: 36964686]
14. Levy JK, Darmstadt GL, Ashby C, Quandt M, Halsey E, Nagar A, et al. Characteristics of successful programmes targeting gender inequality and restrictive gender norms for the health and wellbeing of children, adolescents, and young adults: a systematic review. Lancet Glob Health 2020 Feb;8(2):e225-e236 [FREE Full text] [doi: 10.1016/S2214-109X(19)30495-4] [Medline: 31879212]
15. Brush LD, Miller E. Trouble in paradigm: "gender transformative" programming in violence prevention. Violence Against Women 2019 Nov;25(14):1635-1656. [doi: 10.1177/1077801219872551] [Medline: 31640536]
16. Marcos-Marcos J, Nardini K, Briones-Vozmediano E, Vives-Cases C. Listening to stakeholders in the prevention of gender-based violence among young people in Spain: a qualitative study from the positivMasc project. BMC Womens Health 2023 Jul 26;23(1):393 [FREE Full text] [doi: 10.1186/s12905-023-02545-3] [Medline: 37496067]
17. Casey E, Carlson J, Two Bulls S, Yager A. Gender transformative approaches to engaging men in gender-based violence prevention: a review and conceptual model. Trauma Violence Abuse 2018 Apr;19(2):231-246. [doi: 10.1177/1524838016650191] [Medline: 27197533]
18. Doyle K, Levtov RG, Karamage E, Rakshit D, Kazimbaya S, Sayinzoga F, et al. Long-term impacts of the Bandebereho programme on violence against women and children, maternal health-seeking, and couple relations in Rwanda: a six-year follow-up of a randomised controlled trial. EClinicalMedicine 2023 Sep 26;64:102233 [FREE Full text] [doi: 10.1016/j.eclinm.2023.102233] [Medline: 37781160]
19. van den Berg W, Hendricks L, Hatcher A, Peacock D, Godana P, Dworkin S. 'One man can': shifts in fatherhood beliefs and parenting practices following a gender-transformative programme in Eastern Cape, South Africa. Gend Dev 2013 Feb 01;21(1):111-125 [FREE Full text] [doi: 10.1080/13552074.2013.769775] [Medline: 24000276]
20. Ruane-McAteer E, Gillespie K, Amin A, Aventin Á, Robinson M, Hanratty J, et al. Gender-transformative programming with men and boys to improve sexual and reproductive health and rights: a systematic review of intervention studies. BMJ Glob Health 2020 Oct;5(10):e002997 [FREE Full text] [doi: 10.1136/bmjgh-2020-002997] [Medline: 33051283]
21. WHO guidelines on parenting interventions to prevent maltreatment and enhance parent–child relationships with children aged 0–17 years. World Health Organization. 2023 Feb 10. URL: https://www.who.int/publications/i/item/9789240065505 [accessed 2023-03-09]
22. Backhaus S, Gardner F, Schafer M, Melendez-Torres GJ, Knerr W, Lachman JM. World Health Organization guidelines on parenting interventions to prevent maltreatment and enhance parent–child relationships with children aged 0-17 Years: report of the systematic reviews of evidence. Center for Evidence-Based Intervention, University of Oxford. 2023. URL: https://pure.uva.nl/ws/files/192859093/systematic_reviews-for-the-who-parenting-guideline-jan-27th-2023.pdf [accessed 2024-12-10]
23. Doyle K, Levtov RG, Barker G, Bastian GG, Bingenheimer JB, Kazimbaya S, et al. Gender-transformative Bandebereho couples' intervention to promote male engagement in reproductive and maternal health and violence prevention in Rwanda: findings from a randomized controlled trial. PLoS One 2018 Apr 4;13(4):e0192756 [FREE Full text] [doi: 10.1371/journal.pone.0192756] [Medline: 29617375]

XSL•FO
RenderX

24. Sanders MR, Divan G, Singhal M, Turner KM, Velleman R, Michelson D, et al. Scaling up parenting interventions is critical for attaining the sustainable development goals. Child Psychiatry Hum Dev 2022 Oct;53(5):941-952 [FREE Full text] [doi: 10.1007/s10578-021-01171-0] [Medline: 33948778]
25. Harris M, Andrews K, Gonzalez A, Prime H, Atkinson L. Technology-assisted parenting interventions for families experiencing social disadvantage: a meta-analysis. Prev Sci 2020 Jul;21(5):714-727. [doi: 10.1007/s11121-020-01128-0] [Medline: 32415543]
26. Cluver LD, Meinck F, Steinert JI, Shenderovich Y, Doubt J, Herrero Romero R, et al. Parenting for lifelong health: a pragmatic cluster randomised controlled trial of a non-commercialised parenting programme for adolescents and their families in South Africa. BMJ Glob Health 2018 Jan 31;3(1):e000539 [FREE Full text] [doi: 10.1136/bmjgh-2017-000539] [Medline: 29564157]
27. Ward CL, Wessels IM, Lachman JM, Hutchings J, Cluver LD, Kassanjee R, et al. Parenting for lifelong health for young children: a randomized controlled trial of a parenting program in South Africa to prevent harsh parenting and child conduct problems. J Child Psychol Psychiatry 2020 Apr;61(4):503-512 [FREE Full text] [doi: 10.1111/jcpp.13129] [Medline: 31535371]
28. Schafer M, Lachman JM, Gardner F, Zinser P, Calderon F, Han Q, et al. Integrating intimate partner violence prevention content into a digital parenting chatbot intervention during COVID-19: intervention development and remote data collection. BMC Public Health 2023 Sep 04;23(1):1708 [FREE Full text] [doi: 10.1186/s12889-023-16649-w] [Medline: 37667352]
29. Shai NJ, Sikweyiya Y. Programmes for change: addressing sexual and intimate partner violence in South Africa. S Afr Crime Q 2015 Mar;51:31-41. [doi: 10.4314/sacq.v51i1.4]
30. Sere Y, Roman NV, Ruiter RA. Coping with the experiences of intimate partner violence among South African women: systematic review and meta-synthesis. Front Psychiatry 2021 May 26;12:655130 [FREE Full text] [doi: 10.3389/fpsyt.2021.655130] [Medline: 34122178]
31. Gordon C. Intimate partner violence is everyone's problem, but how should we approach it in a clinical setting? S Afr Med J 2016 Sep 10;106(10):962-965. [doi: 10.7196/samj.2016.v106i10.11408]
32. Violence against women. Organisation for Economic Co-operation and Development. URL: https://www.oecd.org/en/data/indicators/violence-against-women.html [accessed 2024-12-10]
33. Jamaica gender landscape. World Bank Group. URL: https://documents1.worldbank.org/curated/en/099106107012217020/pdf/IDU1101a2e811893114b471a77e16578ca7db7a4.pdf [accessed 2024-12-10]
34. Phaswana ED. Women, gender and race in post-apartheid South Africa. In: Yacob-Haliso O, Falola T, editors. The Palgrave Handbook of African Women's Studies. Cham, Switzerland: Palgrave Macmillan; Feb 09, 2021.
35. Thabane L, Cambon L, Potvin L, Pommier J, Kivits J, Minary L, et al. Population health intervention research: what is the place for pilot studies? Trials 2019 May 30;20(1):309 [FREE Full text] [doi: 10.1186/s13063-019-3422-4] [Medline: 31146768]
36. Kaehler LA, Jacobs M, Jones DJ. Distilling common history and practice elements to inform dissemination: Hanf-model BPT programs as an example. Clin Child Fam Psychol Rev 2016 Sep;19(3):236-258 [FREE Full text] [doi: 10.1007/s10567-016-0210-5] [Medline: 27389606]
37. Ashburn K, Kerner B, Ojamuge D, Lundgren R. Evaluation of the responsible, engaged, and loving (REAL) fathers initiative on physical child punishment and intimate partner violence in northern Uganda. Prev Sci 2017 Oct;18(7):854-864 [FREE Full text] [doi: 10.1007/s11121-016-0713-9] [Medline: 27738782]
38. Kapiga S, Harvey S, Mshana G, Hansen CH, Mtolela GJ, Madaha F, et al. A social empowerment intervention to prevent intimate partner violence against women in a microfinance scheme in Tanzania: findings from the MAISHA cluster randomised controlled trial. Lancet Glob Health 2019 Oct;7(10):e1423-e1434 [FREE Full text] [doi: 10.1016/S2214-109X(19)30316-X] [Medline: 31537372]
39. Eldridge SM, Chan CL, Campbell MJ, Bond CM, Hopewell S, Thabane L, et al. CONSORT 2010 statement: extension to randomised pilot and feasibility trials. BMJ 2016 Oct 24;355:i5239 [FREE Full text] [doi: 10.1136/bmj.i5239] [Medline: 27777223]
40. Guidelines for producing statistics on violence against women. United Nations. 2014. URL: https://unstats.un.org/unsd/gender/docs/guidelines_statistics_vaw.pdf [accessed 2024-12-10]
41. Gugerty MK, Karlan D. The Goldilocks Challenge: Right-Fit Evidence for the Social Sector. Oxford, UK: Oxford University Press; May 24, 2018.
42. Breitenstein SM, Gross D, Christophersen R. Digital delivery methods of parenting training interventions: a systematic review. Worldviews Evid Based Nurs 2014 Jun;11(3):168-176. [doi: 10.1111/wvn.12040] [Medline: 24842341]
43. Collins DA, Tully LA, Piotrowska PJ, Hawes DJ, Dadds MR. Perspectives on ParentWorks: learnings from the development and national roll-out of a self-directed online parenting intervention. Internet Interv 2019 Mar;15:52-59 [FREE Full text] [doi: 10.1016/j.invent.2018.12.002] [Medline: 30656140]
44. UNICEF's open source approach to innovation. UNICEF Innovation. 2020. URL: https://www.unicef.org./innovation/stories/unicefs-open-source-approach-innovation [accessed 2024-12-08]
45. García-Moreno C, Jansen HA, Watts C, Ellsberg M, Heise L. WHO multi-country study on women's health and domestic violence against women: summary report of initial results on prevalence, health outcomes and women's responses. World

XSL•FO
RenderX

Health Organization. 2005. URL: https://iris.who.int/bitstream/handle/10665/43310/9241593512_eng.pdf?sequence=1 [accessed 2024-12-10]
46. Meinck F, Boyes ME, Cluver L, Ward CL, Schmidt P, DeStone S, et al. Adaptation and psychometric properties of the ISPCAN Child Abuse Screening Tool for use in trials (ICAST-Trial) among South African adolescents and their primary caregivers. Child Abuse Negl 2018 Aug;82:45-58 [FREE Full text] [doi: 10.1016/j.chiabu.2018.05.022] [Medline: 29860107]
47. Ebert C, Steinert JI. Prevalence and risk factors of violence against women and children during COVID-19, Germany. Bull World Health Organ 2021 Jun 01;99(6):429-438 [FREE Full text] [doi: 10.2471/BLT.20.270983] [Medline: 34108753]
48. McNaughton Reyes HL, Maman S, Chen MS, Groves AK, Moodley D. Patterns of intimate partner violence victimization among South African women and their relation to emotional distress during pregnancy and postpartum. J Interpers Violence 2021 Apr 10;36(7-8):NP4230-NP4249 [FREE Full text] [doi: 10.1177/0886260518786738] [Medline: 29991312]
49. Abramsky T, Kapinga I, Mshana G, Lees S, Hansen CH, Hashim R, et al. Couples data from north-western Tanzania: insights from a survey of male partners of women enrolled in the MAISHA cluster randomized trial of an intimate partner violence prevention intervention. PLoS One 2020 Oct 2;15(10):e0240112 [FREE Full text] [doi: 10.1371/journal.pone.0240112] [Medline: 33006997]
50. Lachenbruch PA. McNemar test. In: Wiley StatsRef: Statistics Reference Online. Hoboken, NJ: John Wiley & Sons; 2014.
51. Trajman A, Luiz RR. McNemar chi2 test revisited: comparing sensitivity and specificity of diagnostic examinations. Scand J Clin Lab Invest 2008;68(1):77-80. [doi: 10.1080/00365510701666031] [Medline: 18224558]
52. Ethical and safety recommendations for intervention research on violence against women. World Health Organization. 2016 Nov 29. URL: https://www.who.int/publications/i/item/9789241510189 [accessed 2024-12-10]
53. COVID-19 Risk Communication and Community Engagement (RCCE) guidance note on quantitative data collection. United Nations International Children's Fund. 2020. URL: https://www.corecommitments.unicef.org/kp/rcce-covid19---quantitative-data-collection-guidance-note.pdf [accessed 2024-12-01]
54. Leung L, Miedema S, Warner X, Homan S, Fulu E. Making feminism count: integrating feminist research principles in large-scale quantitative research on violence against women and girls. Gend Dev 2019 Oct 29;27(3):427-447. [doi: 10.1080/13552074.2019.1668142]
55. Jamieson MK, Govaart GH, Pownall M. Reflexivity in quantitative research: a rationale and beginner's guide. Soc Personal Psychol Compass 2023 Feb;17(4):e12735. [doi: 10.1111/spc3.12735]
56. Richardson BP, van der Linde J, Pillay B, Swanepoel DW. Do text messages about health and development in young children affect caregiver behaviour and child outcomes? A systematic review. Health Educ J 2020 Oct 21;80(2):225-237. [doi: 10.1177/0017896920965992]
57. Dadds MR, Sicouri G, Piotrowska PJ, Collins DA, Hawes DJ, Moul C, et al. Keeping parents involved: predicting attrition in a self-directed, online program for childhood conduct problems. J Clin Child Adolesc Psychol 2019 Aug 01;48(6):881-893. [doi: 10.1080/15374416.2018.1485109] [Medline: 30067388]
58. Hall CM, Bierman KL. Technology-assisted interventions for parents of young children: emerging practices, current research, and future directions. Early Child Res Q 2015;33:21-32 [FREE Full text] [doi: 10.1016/j.ecresq.2015.05.003] [Medline: 27773964]
59. Guillot-Wright SP, Lu Y, Torres ED, Le VD, Hall HR, Temple JR. Design and feasibility of a school-based text message campaign to promote healthy relationships. School Ment Health 2018 Feb 9;10(4):428-436. [doi: 10.1007/s12310-018-9255-6]
60. Fletcher R, Kay-Lambkin F, May C, Oldmeadow C, Attia J, Leigh L. Supporting men through their transition to fatherhood with messages delivered to their smartphones: a feasibility study of SMS4dads. BMC Public Health 2017 Dec 13;17(1):953 [FREE Full text] [doi: 10.1186/s12889-017-4978-0] [Medline: 29237439]
61. Lechowicz ME, Jiang Y, Tully LA, Burn MT, Collins DA, Hawes DJ, et al. Enhancing father engagement in parenting programs: translating research into practice recommendations. Aus Psychol 2020 Nov 12;54(2):83-89. [doi: 10.1111/ap.12361]
62. Guthridge M, Kirkman M, Penovic T, Giummarra MJ. Promoting gender equality: a systematic review of interventions. Soc Just Res 2022 Sep 01;35(3):318-343. [doi: 10.1007/s11211-022-00398-z]
63. Yaremych HE, Persky S. Recruiting fathers for parenting research: an evaluation of eight recruitment methods and an exploration of fathers' motivations for participation. Parent Sci Pract 2023 Feb 25;23(1):1-32 [FREE Full text] [doi: 10.1080/15295192.2022.2036940] [Medline: 37346458]
64. Jeong J, Sullivan EF, McCann JK, McCoy DC, Yousafzai AK. Implementation characteristics of father-inclusive interventions in low- and middle-income countries: a systematic review. Ann N Y Acad Sci 2023 Feb;1520(1):34-52 [FREE Full text] [doi: 10.1111/nyas.14941] [Medline: 36482863]
65. Nahum-Shani I, Smith SN, Spring BJ, Collins LM, Witkiewitz K, Tewari A, et al. Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. Ann Behav Med 2018 May 18;52(6):446-462 [FREE Full text] [doi: 10.1007/s12160-016-9830-8] [Medline: 27663578]
66. Heckman BW, Mathew AR, Carpenter MJ. Treatment burden and treatment fatigue as barriers to health. Curr Opin Psychol 2015 Oct 01;5:31-36 [FREE Full text] [doi: 10.1016/j.copsyc.2015.03.004] [Medline: 26086031]
67. Hanno EC. Nudging early educators' knowledge, beliefs, and practices: an embedded randomized controlled trial of text message supports. Early Educ Dev 2022 Mar 20;34(2):530-550. [doi: 10.1080/10409289.2022.2053476]

XSL•FO
RenderX

68. Cooper H, Nadzri FZ, Vyas S, Juhari R, Ismail N, Arshat Z, et al. A hybrid digital parenting program delivered within the Malaysian preschool system: protocol for a feasibility study of a small-scale factorial cluster randomized trial. JMIR Res Protoc 2024 Apr 26;13:e55491 [FREE Full text] [doi: 10.2196/55491] [Medline: 38669679]
69. UNICEF Innocenti - Global Office of Research and Foresight, Prevention Collaborative, Equimundo. Parenting programmes to reduce violence against children and women. Why it is important. Brief 1.: UNICEF Innocenti; 2023. URL: https://www.unicef.org/innocenti/media/3456/file/UNICEF-Innocenti-Parenting-Programmes-Reduce-Violence-Against-Children-Women-Brief-1-2024-EN.pdf
70. De Filippo A, Bellatin P, Tietz N, Grant E, Whitefield A, Nkopane P, et al. Effects of digital chatbot on gender attitudes and exposure to intimate partner violence among young women in South Africa. PLOS Digit Health 2023 Oct 16;2(10):e0000358 [FREE Full text] [doi: 10.1371/journal.pdig.0000358] [Medline: 37844088]
71. Kågesten A, Gibbs S, Blum RW, Moreau C, Chandra-Mouli V, Herbert A, et al. Understanding factors that shape gender attitudes in early adolescence globally: a mixed-methods systematic review. PLoS One 2016 Jun 24;11(6):e0157805 [FREE Full text] [doi: 10.1371/journal.pone.0157805] [Medline: 27341206]
72. Falb KL, Asghar K, Blackwell A, Baseme S, Nyanguba M, Roth D, et al. Improving family functioning and reducing violence in the home in North Kivu, Democratic Republic of Congo: a pilot cluster-randomised controlled trial of Safe at Home. BMJ Open 2023 Mar 06;13(3):e065759 [FREE Full text] [doi: 10.1136/bmjopen-2022-065759] [Medline: 36878658]
73. Halim N, Steven Mzilangwe E, Reich N, Badi L, Simmons E, Servidone M, et al. Together to end violence against women in Tanzania: results of a pilot cluster randomized controlled trial to evaluate preliminary effectiveness of interpersonal and community level interventions to address intimate partner violence. Glob Public Health 2019 Dec 13;14(12):1653-1668 [FREE Full text] [doi: 10.1080/17441692.2019.1609062] [Medline: 31084264]
74. Gender action plan, 2022-2025: a vision for lasting, transformative change. United Nations Children's Fund. URL: https://www.unicef.org/gender-equality/gender-action-plan-2022-2025
75. Collins LM. Optimization of Behavioral, Biobehavioral, and Biomedical Interventions: The Multiphase Optimization Strategy (MOST). Cham, Switzerland: Springer; 2018.
76. Starmann E, Collumbien M, Kyegombe N, Devries K, Michau L, Musuya T, et al. Exploring couples' processes of change in the context of SASA!, a violence against women and HIV prevention intervention in Uganda. Prev Sci 2017 Feb 29;18(2):233-244 [FREE Full text] [doi: 10.1007/s11121-016-0716-6] [Medline: 27682273]
77. van Aar J, Leijten P, Orobio de Castro B, Overbeek G. Sustained, fade-out or sleeper effects? A systematic review and meta-analysis of parenting interventions for disruptive child behavior. Clin Psychol Rev 2017 Feb;51:153-163. [doi: 10.1016/j.cpr.2016.11.006] [Medline: 27930935]
78. Broadband statistics. Organisation for Economic Co-operation and Development. URL: https://www.oecd.org/en/topics/sub-issues/broadband-statistics.html [accessed 2023-10-31]
79. Awah I, Green O, Baerecke L, Janowski R, Klapwijk J, Chetty AN, et al. 'It provides practical tips, practical solutions!': acceptability, usability, and satisfaction of a digital parenting intervention across African countries. Psychol Health Med 2022 Aug 18;27(sup1):107-123. [doi: 10.1080/13548506.2022.2113106] [Medline: 35980251]
80. Janowski R, Green O, Shenderovich Y, Stern D, Clements L, Wamoyi J, et al. Optimising engagement in a digital parenting intervention to prevent violence against adolescents in Tanzania: protocol for a cluster randomised factorial trial. BMC Public Health 2023 Jun 23;23(1):1224 [FREE Full text] [doi: 10.1186/s12889-023-15989-x] [Medline: 37353844]
81. Cullen C. Method matters: the underreporting of intimate partner violence. World Bank Econ Rev 2023 Feb;37(1):49-73. [doi: 10.1093/wber/lhac022]
82. Visschers J, Jaspaert E, Vervaeke G. Social desirability in intimate partner violence and relationship satisfaction reports: an exploratory analysis. J Interpers Violence 2017 May 09;32(9):1401-1420. [doi: 10.1177/0886260515588922] [Medline: 26058978]
83. McGirr S, Torres J, Heany J, Brandon H, Tarry C, Robinson C. Lessons learned on recruiting and retaining young fathers in a parenting and repeat pregnancy prevention program. Matern Child Health J 2020 Sep 20;24(Suppl 2):183-190 [FREE Full text] [doi: 10.1007/s10995-020-02956-w] [Medline: 32564249]
84. Bott S, Guedes A, Subrahmanian R, Chowdhury S, Rumble L, Daban RA. Gender dimensions of violence against children and adolescents. United Nations Children's Fund (UNICEF). 2020 Apr. URL: https://www.unicef.org/media/93986/file/Child-Protection-Gender-Dimensions-of-VACAG-2021.pdf [accessed 2024-12-10]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
IPV: intimate partner violence
ODK: Open Data Kit
RCT: randomized controlled trial
UNICEF: United Nations Children’s Fund
VAC: violence against children
VAW: violence against women

XSL•FO
RenderX

Edited by A Mavragani; submitted 20.03.24; peer-reviewed by M Wang, S Willan, RG Levtov; comments to author 25.06.24; revised version received 13.10.24; accepted 30.10.24; published 03.01.25.

Please cite as:
Schafer M, Lachman J, Zinser P, Calderón Alfaro FA, Han Q, Facciola C, Clements L, Gardner F, Haupt Ronnie G, Sheil R
A Digital Parenting Intervention With Intimate Partner Violence Prevention Content: Quantitative Pre-Post Pilot Study
JMIR Form Res 2025;9:e58611
URL: https://formative.jmir.org/2025/1/e58611
doi:10.2196/58611
PMID:

©Moa Schafer, Jamie Lachman, Paula Zinser, Francisco Antonio Calderón Alfaro, Qing Han, Chiara Facciola, Lily Clements, Frances Gardner, Genevieve Haupt Ronnie, Ross Sheil. Originally published in JMIR Formative Research (https://formative.jmir.org), 03.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Short Paper

Validation of Sleep Measurements of an Actigraphy Watch: Instrument Validation Study

Mari Waki[1*]; Ryohei Nakada[1*]; Kayo Waki[1]; Yuki Ban[1]; Ryo Suzuki[2]; Toshimasa Yamauchi[3]; Masaomi Nangaku[4]; Kazuhiko Ohe[1]

[1]Department of Biomedical Informatics, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan
[2]Department of Diabetes, Metabolism and Endocrinology, Tokyo Medical University, Tokyo, Japan
[3]Department of Diabetes and Metabolic Diseases, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan
[4]Division of Nephrology and Endocrinology, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan
[*]these authors contributed equally

Corresponding Author:
Kayo Waki
Department of Biomedical Informatics
Graduate School of Medicine
The University of Tokyo
7-chōme-3-1 Hongō
Bunkyo City
Tokyo, 113-8654
Japan
Phone: 81 03 5800 9129
Fax: 81 03 5800 9173
Email: kwaki-tky@m.u-tokyo.ac.jp

Abstract

Background: The iAide2 (Tokai) physical activity monitoring system includes diverse measurements and wireless features useful to researchers. The iAide2's sleep measurement capabilities have not been compared to validated sleep measurement standards in any published work.

Objective: We aimed to assess the iAide2's sleep duration and total sleep time (TST) measurement performance and perform calibration if needed.

Methods: We performed free-living sleep monitoring in 6 convenience-sampled participants without known sleep disorders recruited from within the Waki DTx Laboratory at the Graduate School of Medicine, University of Tokyo. To assess free-living sleep, we validated the iAide2 against a second actigraph that was previously validated against polysomnography, the MotionWatch 8 (MW8; CamNtech Ltd). The participants wore both devices on the nondominant arm, with the MW8 closest to the hand, all day except when bathing. The MW8 and iAide2 assessments both used the MW8 EVENT-marker button to record bedtime and risetime. For the MW8, MotionWare Software (version 1.4.20; CamNtech Ltd) provided TST, and we calculated sleep duration from the sleep onset and sleep offset provided by the software. We used a similar process with the iAide2, using iAide2 software (version 7.0). We analyzed 64 nights and evaluated the agreement between the iAide2 and the MW8 for sleep duration and TST based on intraclass correlation coefficients (ICCs).

Results: The absolute ICCs (2-way mixed effects, absolute agreement, single measurement) for sleep duration (0.69, 95% CI –0.07 to 0.91) and TST (0.56, 95% CI –0.07 to 0.82) were moderate. The consistency ICC (2-way mixed effects, consistency, single measurement) was excellent for sleep duration (0.91, 95% CI 0.86-0.95) and moderate for TST (0.78, 95% CI 0.67-0.86). We determined a simple calibration approach. After calibration, the ICCs improved to 0.96 (95% CI 0.94-0.98) for sleep duration and 0.82 (95% CI 0.71-0.88) for TST. The results were not sensitive to the specific participants included, with an ICC range of 0.96-0.97 for sleep duration and 0.79-0.87 for TST when applying our calibration equation to data removing one participant at a time and 0.96-0.97 for sleep duration and 0.79-0.86 for TST when recalibrating while removing one participant at a time.

Conclusions: The measurement errors of the uncalibrated iAide2 for both sleep duration and TST seem too large for them to be useful as absolute measurements, though they could be useful as relative measurements. The measurement errors after calibration

XSL•FO
RenderX

are low, and the calibration approach is general and robust, validating the use of iAide2's sleep measurement functions alongside its other features in physical activity research.

(JMIR Form Res 2025;9:e63529) doi:10.2196/63529

KEYWORDS

actigraphy; sleep; Motion Watch 8; iAide2; total sleep time

Introduction

Background

The iAide2 (Tokai) is a physical activity monitoring system that measures intensity levels, step counts, temperature, and pulse rate [1,2] to assess activity and sleep, with minute-by-minute assessment of wake/sleep [3,4]. It is based on TDK's Silmee W20/W22 actigraphs [4-6] used in previous research [5,7-9]. The iAide2 supports long trials (via rechargeability and off-watch data storage) and timely behavior feedback (via a wireless EVENT button and near–real-time access to data). Its sleep measurements are not well validated; Kimura et al [7] compared its predecessor, the W20, to video-monitoring sleep assessment (which is itself unvalidated), and reported 5 sleep duration values, with detailed methodology not given [7].

Actigraphs are commonly validated against polysomnography (PSG) or another actigraph. PSG, though imperfect, is the gold standard of sleep measurement [10]. PSG's complexity requires sleep in a laboratory [11], which is costly and unrepresentative of free-living sleep. Watch-based actigraphy can efficiently collect multiple nights' data in a natural environment [12,13].

Objective

We assessed iAide2-measured sleep duration and total sleep time (TST) in participants without sleep disorders and added calibration when needed.

Methods

Approach

To assess free-living sleep, we validated iAide2 against a second actigraph, the MotionWatch 8 (MW8; CamNtech Ltd) [14], using methods similar to other actigraph-to-actigraph comparisons [11,15-17]. The MW8 is validated against PSG [18-20], with high correlations (r=0.920) and no significant bias (7.0 minutes) for TST [19]. Researchers have used the MW8 to assess sleep in people without sleep disorders [21-23]. We used validated MW8 settings (mode 1, 30-second epochs, and a threshold of 20 [24,25]).

Data Collection

We collected data from a convenience sample of members of the Waki DTx Laboratory at the Graduate School of Medicine, University of Tokyo. The laboratory is investigating lifestyle interventions to improve sleep and exercise and identified the iAide2 as a candidate measurement device. This validation study assessed the accuracy of iAide2 sleep monitoring in 6 individuals without known sleep disorders. They wore both devices on the nondominant arm, with the MW8 closest to the hand, all day except when bathing.

The participants used the MW8 EVENT-marker button to record bedtime (when they began trying to sleep) and risetime (when they stopped trying to sleep). The duration between these times is time in bed (TIB). Our MW8 and iAide2 assessments both used these nonactigraphy measurements.

MotionWare (version 1.4.20; CamNtech Ltd) provided sleep onset ("Fell Asleep" in MotionWare, the earliest time after bedtime when the participant was asleep), sleep offset ("Woke Up," the latest time prior to risetime when the participant transitioned from asleep to awake), and TST (the sum of all minutes of sleep between sleep onset and offset) [25]. From these measurements, we derived sleep onset latency (the duration between bedtime and sleep onset), sleep offset latency (the duration between sleep offset and risetime), and sleep duration (the duration between sleep onset and sleep offset).

We used a similar process with the iAide2, using iAide2 software (version 7.0). We defined sleep onset as the time of the first sleep after bedtime. Similarly, we defined sleep offset as 1 minute after the time of final sleep prior to risetime. We defined TST as the sum of minutes asleep within the sleep duration period.

Statistical Analysis

We compared measurements using intraclass correlation coefficients (ICCs; 2-way mixed effects, absolute agreement, single measurement, unless otherwise specified) [26,27]. We used the *irr* package 0.84.1 in R for point estimation of ICC and calculating 95% CIs, and R (version 4.2.1) for calibration.

Ethical Considerations

The University of Tokyo School of Medicine institutional review board (2024137NI) approved this study, which abided by the Declaration of Helsinki. Participants provided verbal consent for data collection upon recruitment. All participants provided written consent for the use of their data. Participants' data were anonymized. No compensation was provided.

Results

Participants

We recruited 6 participants (aged 24-63 years) and recorded data for 64 nights (Figure 1). The number of nights per participant varied from 5 to 21. The average nightly sleep duration via MW8 per participant varied from 5.52 to 6.73 hours, and the participants differed in the regularity of their sleep durations.

XSL•FO
RenderX

Figure 1. Participants' information and sleep duration measured using the MotionWatch 8 actigraph (MW8; 6 participants, 64 nights).

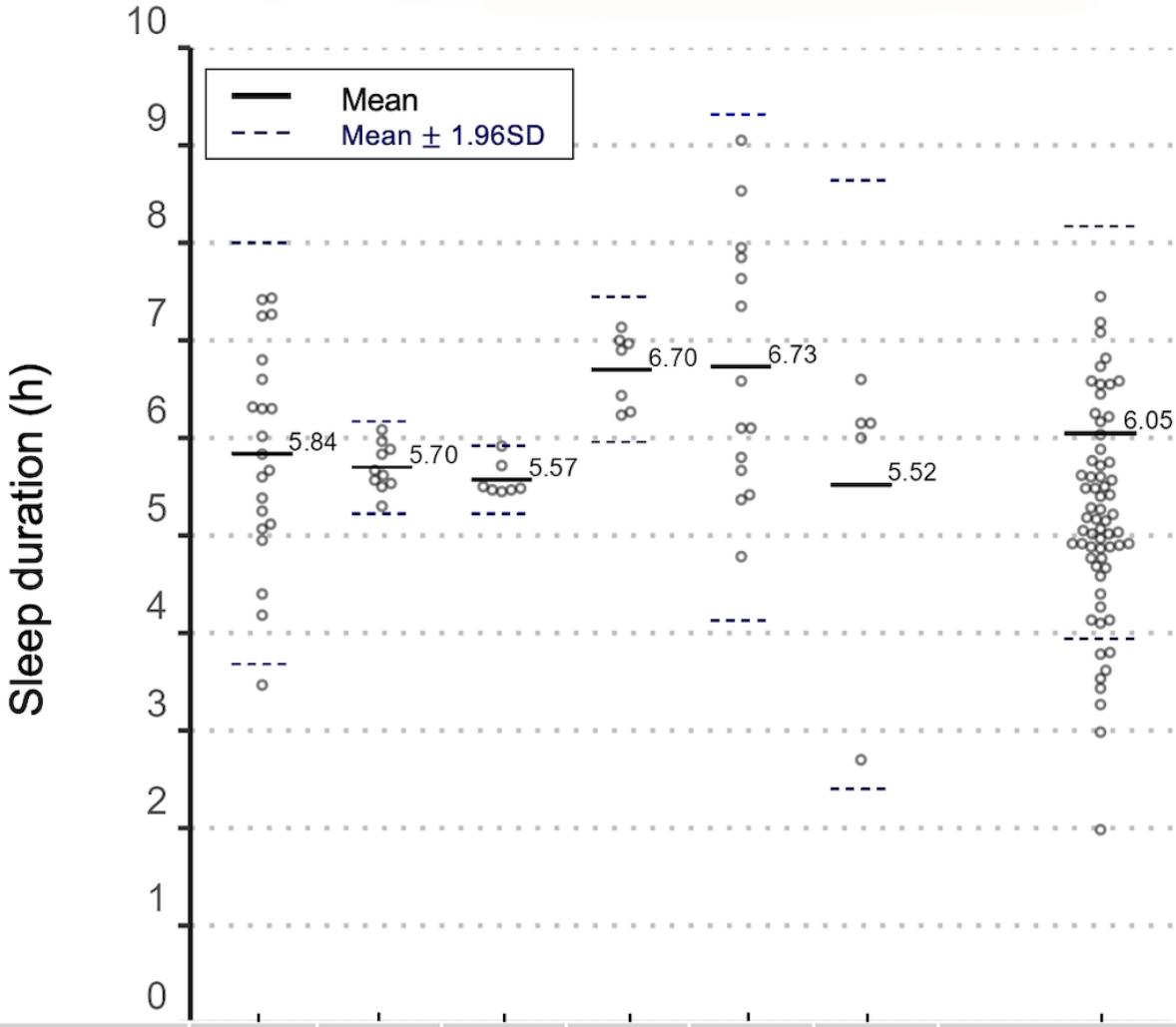

ID	4	9	10	7	11	12	Total
Nights, n	21	10	7	7	14	5	64
Age (yrs)	24	50	54	63	55	35	
Sex	M	F	F	M	M	F	
Sleep duration (h), mean (95%CI)	5.84 [5.30-6.34]	5.70 [5.52-5.87]	5.57 [5.41-5.74]	6.70 [6.35-7.06]	6.73 [5.96-7.49]	5.52 [3.54-7.50]	6.05 [5.78-6.32]

Measurements

The iAide2 underestimated duration relative to the MW8 and showed moderate ICC of 0.69 (95% CI –0.07 to 0.91; Figure 2A) [26]. Relative performance was excellent (ICC=0.91, 95% CI 0.86-0.95; 2-way mixed effects, consistency, single measurement) [26]. The mean, maximum, and minimum differences between iAide2 and MW8 were –0.86, 0.22, and –2.22 hours, respectively (Figure 2B).

Similarly, the iAide2 underestimated TST and showed moderate ICC of 0.56 (95% CI –0.07 to 0.82; Figure 2C) and a good consistency ICC (2-way mixed effects, consistency, single measurement) of 0.78 (95% CI 0.67-0.86). The mean, maximum, and minimum differences between iAide2 and MW8 were –0.93, 0.80, and –3.20 hours, respectively (Figure 2D).

XSL•FO
RenderX

Figure 2. Scatter plot and Bland-Altman plot for measured sleep duration and total sleep time using iAide2 and the MotionWatch 8 actigraph (MW8; 64 nights).

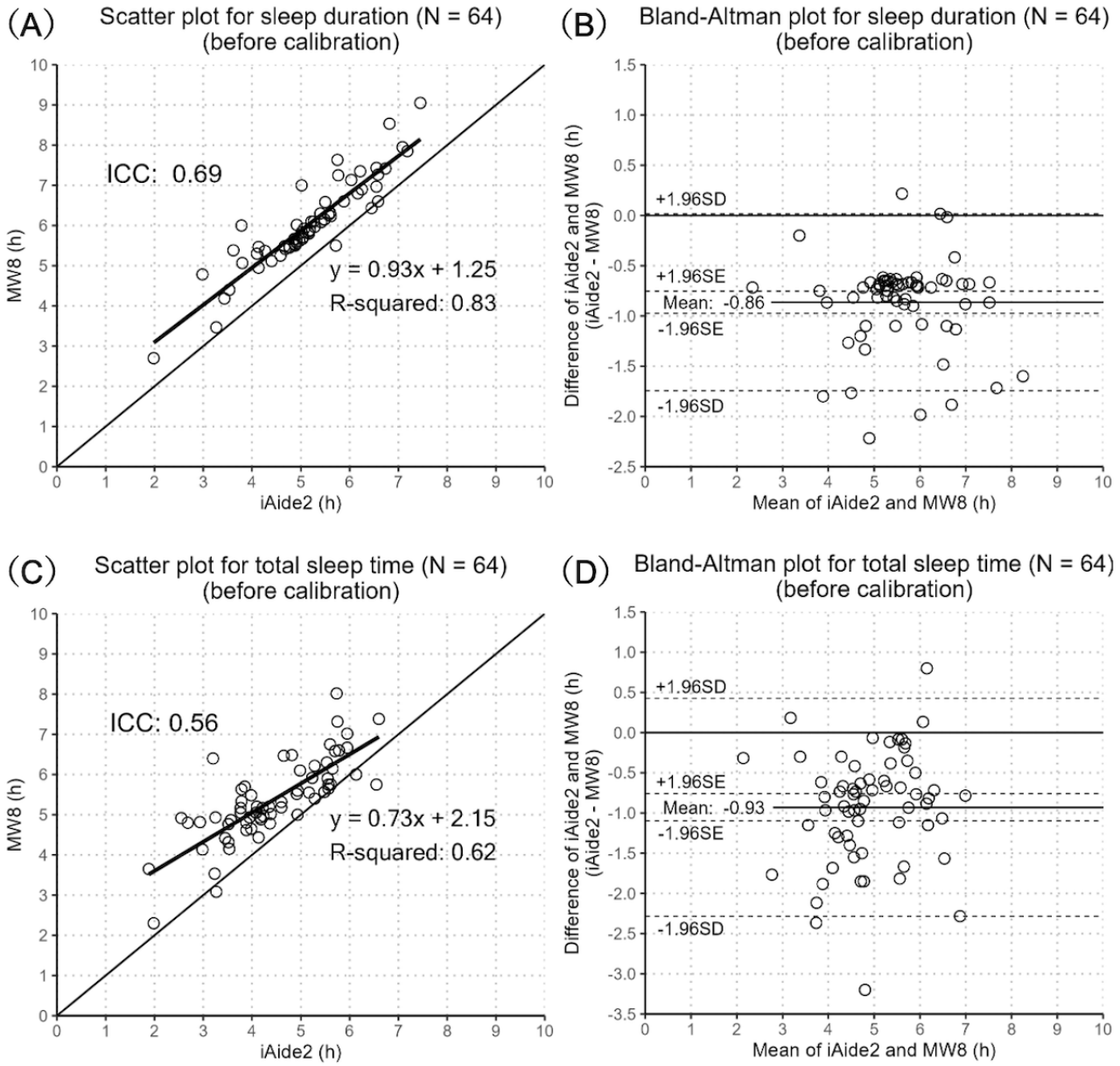

Calibration

Our iAide2 data showed very small sleep offset latencies (mean 0.03 hours) but significant overestimates of sleep onset latency, so we corrected the sleep duration by the difference between the iAide2-measured sleep onset latency and an iAide2 sleep onset latency calibrated to the MW8 measurements via ordinary least squares (OLS). We added a small (<5 minute) bias correction, selected to optimize ICC, to account for small offset latency biases. To prevent physically impossible results, we limited the calibrated duration to be no longer than the TIB. The resulting calibration equation is D' = min {0.59S + D + 0.26, TIB} (hours), where D is the measured duration, S is the measured sleep onset latency, and D' is the calibrated duration.

For TST, we used OLS to calibrate the measured TST value, limiting the result to be not greater than the calibrated sleep duration. The resulting equation is T' = min {0.73T + 2.15, D'} (hours), where T is the measured TST and T' is the calibrated TST.

Calibration greatly improved accuracy. Sleep duration had excellent ICC (0.96, 95% CI 0.94-0.98; Figure 3A). The mean difference almost disappeared (–0.01 hours) and the difference between the maximum (0.72 hours) and minimum (–0.96 hours) values decreased (Figure 3B). TST similarly had greatly improved accuracy, with good ICC (0.82, 95% CI 0.71-0.88; Figure 3C). The mean difference became almost zero (–0.04 hours) and the difference between the maximum (1.15 hours) and minimum (–1.93 hours) values decreased (Figure 3D).

XSL•FO
RenderX

Figure 3. Scatter plot and Bland-Altman plot for calibrated sleep duration and total sleep time using iAide2 and the MotionWatch 8 actigraph (MW8; 64 nights).

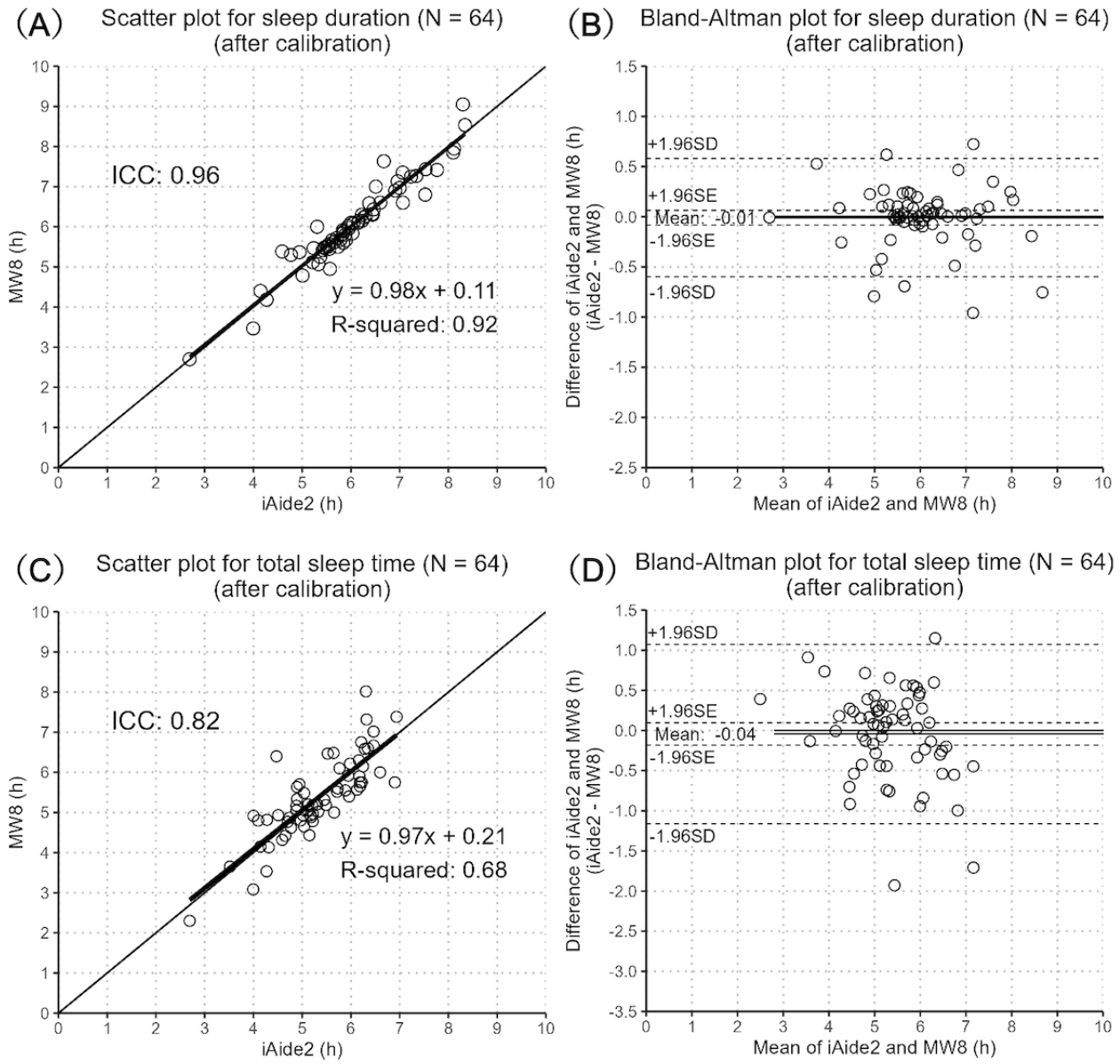

Because our sample size is small with a varying number of nights per participant, we conducted a sensitivity analysis to determine how the ICC agreement rate changes based on the included participants. First, to show that our ICCs are not sensitive to the exact individuals, we performed trials excluding the data of one participant at a time, recalibrating for each trial. The calibrated ICC range was 0.96-0.97 for sleep duration, effectively unchanged from 0.96 in the baseline analysis, and 0.79-0.86 for TST, similar to 0.82 in the baseline analysis, demonstrating that our results are not highly influenced by the variability in number of nights per participant. To demonstrate the generality of our calibration equations, we applied the calibration derived from all participants' data to trials excluding one participant at a time. The ICC range was 0.96-0.97 for sleep duration and 0.79-0.87 for TST, again similar to baseline analysis values, showing that our calibration has high generality and is not highly dependent on any one participant's data.

Discussion

Principal Findings

The absolute accuracies of uncalibrated iAide2 measurements for sleep duration (ICC=0.69) and TST (ICC=0.56) are relatively poor. Uncalibrated data could be useful in applications needing relative measurements, with good values for the consistency version of the ICC (2-way mixed effects, consistency, single measurement) of 0.91 for sleep duration and 0.78 for TST.

The calibrated measurements have excellent absolute accuracy. The ICCs for sleep duration (0.96) and TST (0.82) are comparable to or better than those reported previously (0.67 and 0.68 for sleep duration and 0.77 and 0.84 for TST [28,29]). Sensitivity analysis revealed essentially unchanged performance when applying this calibration to data while removing one participant at a time and when recalibrating while removing one participant at a time. The calibration methods are simple

XSL•FO
RenderX

and have good statistical bases. The calibrated iAide2 measurements of sleep duration and TST seem usable in a wide range of applications.

Limitations

The MW8, our validation standard, has its own errors, with a reported underestimation of sleep onset latency (mean difference 8.9 minutes) and overestimation of TST (mean difference 47.1 minutes) relative to PSG [30]. Validation against other standards, including PSG, is warranted.

We used more nights (N=64) than some PSG-based validation studies (N=20, 37, and 38) but fewer than actigraph-based research (N=140) [17,18,31,32]. Our convenience sample was small (N=6); however, it included a mix of ages, sex, and sleeping regularity. Our results may not fully apply to other populations; and larger studies are warranted.

Conclusions

The iAide2 has interesting measurement capabilities with convenient features, and validating its sleep measurement accuracy adds to its usefulness. The absolute accuracies are low, but good relative accuracy may be suitable for some applications. Its postcalibration accuracies are excellent and well validated against the MW8.

Acknowledgments

We thank Daniel Lane for manuscript editing and scientific discussions.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request for noncommercial purposes.

Authors' Contributions

KW was responsible for the overall design and evaluation of this study. The model was designed and implemented by YB. Data were collected and organized by YB, RN, and MW. Experiments were conducted by YB, RN, and MW. KW and RS supervised the work. The first draft of the manuscript was written by MW. All authors read and approved the final manuscript.

Conflicts of Interest

None.

References

1. iAide2 (iAide2-W). Pharmaceuticals and Medical Devices Agency. 2021. URL: https://www.info.pmda.go.jp/downfiles/md/PDF/340205/340205_301AGBZX00020000_A_01_02.pdf [accessed 2024-02-26]
2. iAide. Tokai Corporation. URL: https://www.iaide-system.jp/iAide/Pages/Login.aspx [accessed 2024-04-10]
3. Watch-type activity meters that can be used in various businesses Silmee W22. TDK Product Center. URL: https://product.tdk.com/ja/products/biosensor/biosensor/silmee_w22/index.html [accessed 2024-02-26]
4. Miyamoto K, Hashimoto K, Kasaoka M, Kakumu M. Wearable sensors corresponding to various applications in medical/healthcare field. 2020 Presented at: 27th International Workshop on Active-Matrix Flatpanel Displays and Devices (AM-FPD); September 1-4, 2020; Kyoto, Japan. [doi: 10.23919/am-fpd49417.2020.9224474]
5. Detection of atrial fibrillation using a deep learning with wearable pulse wave sensor. 2019 Presented at: APHRS (ASIA Pacific Heart Rhythm Society scientific session); Oct 24-27, 2019; Bangkok, Thailand. [doi: 10.1002/joa3.12267]
6. Biometric Sensors: Business collaboration to commercialize 24-hour remote monitoring system for medical and nursing care using wristband-type wearable devices. TDK Corporation. 2017. URL: https://www.tdk.com/ja/news_center/press/20171002_01.html [accessed 2024-02-26]
7. Kimura N, Aso Y, Yabuuchi K, Ishibashi M, Hori D, Sasaki Y, et al. Modifiable lifestyle factors and cognitive function in older people: a cross-sectional observational study. Front Neurol 2019 Apr 24;10:401 [FREE Full text] [doi: 10.3389/fneur.2019.00401] [Medline: 31068892]
8. Tazawa Y, Liang K, Yoshimura M, Kitazawa M, Kaise Y, Takamiya A, et al. Evaluating depression with multimodal wristband-type wearable device: screening and assessing patient severity utilizing machine-learning. Heliyon 2020 Feb;6(2):e03274 [FREE Full text] [doi: 10.1016/j.heliyon.2020.e03274] [Medline: 32055728]
9. Kadoya Y, Khan MSR, Watanapongvanich S, Binnagan P. Emotional status and productivity: evidence from the Special Economic Zone in Laos. Sustainability 2020 Feb 19;12(4):1544. [doi: 10.3390/su12041544]
10. Tryon WW. Nocturnal activity and sleep assessment. Clin Psychol Rev 1996 Jan;16(3):197-213. [doi: 10.1016/0272-7358(95)00059-3]
11. Asgari Mehrabadi M, Azimi I, Sarhaddi F, Axelin A, Niela-Vilén H, Myllyntausta S, et al. Sleep tracking of a commercially available smart ring and smartwatch against medical-grade actigraphy in everyday settings: instrument validation study. JMIR Mhealth Uhealth 2020 Nov 02;8(10):e20465. [doi: 10.2196/20465] [Medline: 33038869]

XSL•FO
RenderX

12. Fekedulegn D, Andrew ME, Shi M, Violanti JM, Knox S, Innes KE. Actigraphy-based assessment of sleep parameters. Ann Work Expo Health 2020 Apr 30;64(4):350-367 [FREE Full text] [doi: 10.1093/annweh/wxaa007] [Medline: 32053169]
13. Ancoli-Israel S, Cole R, Alessi C, Chambers M, Moorcroft W, Pollak CP. The role of actigraphy in the study of sleep and circadian rhythms. Sleep 2003 May 01;26(3):342-392. [doi: 10.1093/sleep/26.3.342] [Medline: 12749557]
14. MotionWatch 8. CamNtech Ltd. URL: https://www.camntech.com/motionwatch-8/ [accessed 2024-02-26]
15. Morgan PL, Hampton S, Karatziotou A, Zaslona J, Middleton B. Validation of two new activity monitors: motionwatch 8 and pro-diary motion. 2012 Presented at: 21st Congress of the European-Sleep-Research-Society; September 4-8, 2012; Paris, France.
16. Topalidis P, Florea C, Eigl E, Kurapov A, Leon CAB, Schabus M. Evaluation of a low-cost commercial actigraph and its potential use in detecting cultural variations in physical activity and sleep. Sensors (Basel) 2021 May 29;21(11) [FREE Full text] [doi: 10.3390/s21113774] [Medline: 34072347]
17. Shah VV, Brumbach BH, Pearson S, Vasilyev P, King E, Carlson-Kuhta P, et al. Opal actigraphy (activity and sleep) measures compared to ActiGraph: a validation study. Sensors (Basel) 2023 Feb 18;23(4):2296 [FREE Full text] [doi: 10.3390/s23042296] [Medline: 36850896]
18. Elbaz M, Yauy K, Metlaine A, Martoni M, Leger D. Validation of a new actigraph motion watch versus polysomnography on 70 healthy and suspected sleep-disordered subjects. CamNtech Ltd. URL: https://www.camntech.com/Engineering/MotionWare/MotionWatchPSG.pdf [accessed 2023-10-19]
19. Dorffner G, Kemethofer M, Saletu-Zyhlarz G, Simeoni M, Rainer L, Parapatics S, et al. Measuring sleep: a comparison of polysomnography (PSG), wrist actigraphy and a reduced PSG solution. International Society for CNS Clinical Trials and Methodology. URL: https://isctm.org/public_access/Autumn2018/Abstracts/Dorffner-abstract.pdf [accessed 2023-10-17]
20. Yavuz-Kodat E, Reynaud E, Geoffray M, Limousin N, Franco P, Bourgin P, et al. Validity of actigraphy compared to polysomnography for sleep assessment in children with autism spectrum disorder. Front Psychiatry 2019 Aug 2;10:551 [FREE Full text] [doi: 10.3389/fpsyt.2019.00551] [Medline: 31428003]
21. Deantoni M, Reyt M, Baillet M, Dourte M, De Haan S, Lesoinne A, et al. Napping and circadian sleep-wake regulation during healthy aging. Sleep 2024 May 10;47(5):2023 [FREE Full text] [doi: 10.1093/sleep/zsad287] [Medline: 37943833]
22. Illingworth G, Sharman R, Harvey C, Foster RG, Espie CA. The Teensleep study: the effectiveness of a school-based sleep education programme at improving early adolescent sleep. Sleep Med X 2020 Dec;2:100011 [FREE Full text] [doi: 10.1016/j.sleepx.2019.100011] [Medline: 33458648]
23. Nédélec M, Chauvineau M, Guilhem G. On the road to Camarón: the sleep of an ultra-endurance athlete cycling 10,000 km in 24 days. Int J Environ Res Public Health 2022 Apr 09;19(8):4543 [FREE Full text] [doi: 10.3390/ijerph19084543] [Medline: 35457410]
24. The MotionWatch User Guide. CamNtech. 2024. URL: https://www.camntech.com/Products/MotionWatch/The%20MotionWatch%20User%20Guide.pdf [accessed 2024-02-26]
25. MotionWare Software User Guide. CamNtech. 2024. URL: https://www.camntech.com/Products/MotionWatch/The%20MotionWare%20Software%20User%20Guide.pdf [accessed 2024-02-26]
26. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016 Jun;15(2):155-163. [doi: 10.1016/j.jcm.2016.02.012] [Medline: 27330520]
27. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychol Methods 1996 Mar;1(1):30-46. [doi: 10.1037/1082-989X.1.1.30]
28. Plekhanova T, Rowlands AV, Davies MJ, Hall AP, Yates T, Edwardson CL. Validation of an automated sleep detection algorithm using data from multiple accelerometer brands. J Sleep Res 2023 Jun 31;32(3):e13760. [doi: 10.1111/jsr.13760] [Medline: 36317222]
29. Kainec KA, Caccavaro J, Barnes M, Hoff C, Berlin A, Spencer RMC. Evaluating accuracy in five commercial sleep-tracking devices compared to research-grade actigraphy and polysomnography. Sensors (Basel) 2024 Jan 19;24(2) [FREE Full text] [doi: 10.3390/s24020635] [Medline: 38276327]
30. Ameen MS, Cheung LM, Hauser T, Hahn MA, Schabus M. About the accuracy and problems of consumer devices in the assessment of sleep. Sensors (Basel) 2019 Sep 25;19(19) [FREE Full text] [doi: 10.3390/s19194160] [Medline: 31557952]
31. Kanady JC, Ruoff L, Straus LD, Varbel J, Metzler T, Richards A, et al. Validation of sleep measurement in a multisensor consumer grade wearable device in healthy young adults. J Clin Sleep Med 2020 Jun 15;16(6):917-924 [FREE Full text] [doi: 10.5664/jcsm.8362] [Medline: 32048595]
32. Kosmadopoulos A, Sargent C, Darwent D, Zhou X, Roach GD. Alternatives to polysomnography (PSG): a validation of wrist actigraphy and a partial-PSG system. Behav Res Methods 2014 Dec;46(4):1032-1041. [doi: 10.3758/s13428-013-0438-7] [Medline: 24442593]

Abbreviations

ICC: intraclass correlation coefficient
MW8: MotionWatch 8 actigraph
OLS: ordinary least squares

XSL•FO
RenderX

PSG: polysomnography
TIB: time in bed
TST: total sleep time

Edited by A Mavragani; submitted 23.06.24; peer-reviewed by Y Minami, JE Lee; comments to author 25.10.24; revised version received 11.11.24; accepted 24.11.24; published 06.01.25.

Please cite as:
Waki M, Nakada R, Waki K, Ban Y, Suzuki R, Yamauchi T, Nangaku M, Ohe K
Validation of Sleep Measurements of an Actigraphy Watch: Instrument Validation Study
JMIR Form Res 2025;9:e63529
URL: https://formative.jmir.org/2025/1/e63529
doi:10.2196/63529
PMID:

©Mari Waki, Ryohei Nakada, Kayo Waki, Yuki Ban, Ryo Suzuki, Toshimasa Yamauchi, Masaomi Nangaku, Kazuhiko Ohe. Originally published in JMIR Formative Research (https://formative.jmir.org), 06.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Differences in eHealth Access, Use, and Perceived Benefit Between Different Socioeconomic Groups in the Dutch Context: Secondary Cross-Sectional Study

Lucille Standaar[1,2,3], BSc, MSc; Lilian van Tuyl[2], MBA, PhD; Anita Suijkerbuijk[1], PhD; Anne Brabers[2], PhD; Roland Friele[2,3], Prof Dr

[1]Department of Population Health and Health Services Research, Centre for Public Health, Healthcare and Society, National Institute for Public Health and the Environment, Bilthoven, Netherlands

[2]Department Organisation and Quality of Care, Netherlands Institute for Health Services Research, Utrecht, Netherlands

[3]Tranzo Scientific Center for Care and Wellbeing, Tilburg University, Tilburg, Netherlands

Corresponding Author:
Lucille Standaar, BSc, MSc
Department of Population Health and Health Services Research
Centre for Public Health, Healthcare and Society
National Institute for Public Health and the Environment
PO Box 1
Bilthoven, 3723 BA
Netherlands
Phone: 31 302743243
Fax: 31 302744407
Email: lucille.standaar@rivm.nl

Abstract

Background: There is a growing concern that digital health care may exacerbate existing health disparities. Digital health care or eHealth encompasses the digital apps that are used in health care. Differences in access, use, and perceived benefits of digital technology among socioeconomic groups are commonly referred to as the digital divide. Current research shows that people in lower socioeconomic positions (SEPs) use eHealth less frequently.

Objective: This study aims to (1) investigate the association between SEP and eHealth access to, use of, and perceived benefit within the adult Dutch population and (2) evaluate disparities in eHealth access, use, and perceived benefit through three socioeconomic variables—education, standardized income, and the socioeconomic status of the neighborhood.

Methods: A secondary analysis was conducted on data from the Nivel Dutch Health Care Consumer Panel (response rate 57%, 849/1500), to assess access to, use of, and perceived benefits from eHealth. These data were collected to monitor eHealth developments in the Netherlands. eHealth was examined through two concepts: (1) eHealth in general and (2) websites, apps, and wearables. Results were stratified into 9 SEP populations based on 3 indicators—education, standardized income, and socioeconomic status level of the neighborhood. Logistic regression analyses were performed to evaluate whether the outcomes varied significantly across different SEP groups. Age was included as a covariate to control for confounding.

Results: This study confirms the association between eHealth and SEP and shows that low SEP respondents have less access (odds ratio [OR] 5.72, 95% CI 3.06-10.72) and use (OR 4.96, 95% CI 2.66-9.24) of eHealth compared to medium or high SEP respondents. Differences were most profound when stratifying for levels of education.

Conclusions: The access to and use of eHealth has a socioeconomic gradient and emphasizes that SEP indicators cannot be used interchangeably to assess eHealth access and use. The results underline the importance of activities and policies aimed at improving eHealth accessibility and usage among low SEP groups to mitigate disparities in health between different socioeconomic groups.

(JMIR Form Res 2025;9:e49585) doi:10.2196/49585

XSL•FO
RenderX

KEYWORDS

eHealth; digital divide; socioeconomic factors; education; income; neighborhood; health disparities; cross-sectional studies; digital health care; health equity; Netherlands

Introduction

Digital health care is expected to provide benefits for health care systems, providers, and patients and is considered a solution to address workforce shortages and rising health care costs [1-4]. Moreover, digital health care is anticipated to enhance the quality of care, stimulate patient self-management, and improve health accessibility and equity [1-4]. Digital health care or eHealth is defined as the digital apps that are used in health care. Health care policies focus on increased use of and dependency on the eHealth apps [4-6]. Concerns have been raised that digital health care may not be equally accessible for all [7-13]. Populations with lower socioeconomic positions (SEPs) are more likely to encounter financial, skill, or cognitive barriers to accessing and using eHealth, such as limited access to devices, limited digital health skills, or limited ability to take the initiative in using eHealth [7,11-16]. Additionally, these populations experience challenges in comprehending and implementing health information and healthy behaviors in daily life [17-21].

The demand for health care services is often higher in low-SEP populations, as people with a low SEP endure more often from chronic illnesses [18,19,22,23]. Research shows that low-SEP populations often have different views on health and the possible benefits of healthy behaviors compared to high-SEP populations [24-27]. Studies find that low-SEP populations have less time, more stress, and limited financial capacities to implement healthy behaviors [26]. Next to this, it is theorized that health beliefs often find origin in the health beliefs of previous generations [24-27]. In the context of low SEP, the expectation of a shorter life and the belief that their own behavior has limited influence on their longevity pose barriers to adhering to healthy behaviors [20]. Therefore, the digitalization of health care could seriously impact the access and use of health care for those who need it most [5,6,13,28-31].

National digital connectivity and policies that stimulate the transition toward a digital health care system could improve the implementation and accessibility of eHealth. The European Union and its member states deploy policies to realize the digital transition of health care systems [5,6]. Some states, such as the Netherlands, have been experimenting with digital health care for over a decade [4,6]. Most Dutch households (98%) have fast broadband coverage (2020) and 88% of the Dutch population uses mobile broadband (2019), which indicates the use of a mobile phone or other device with mobile internet access [5]. The level of connectivity in the Netherlands could facilitate the implementation of digital health care [5].

Several countries have developed national eHealth monitoring programs to monitor the uptake and effects of eHealth among health care professionals and citizens [32-34]. From a citizen's perspective, the monitoring programs focus on the use and evaluation of eHealth that involve citizen interaction [32-34]. This includes applications such as websites, apps, and wearables that citizens can use independently or involve eHealth tools that facilitate digital communication with a health care professional, such as video calls or messaging via patient portals [32-34]. This study is based on secondary data analysis of the Dutch eHealth monitoring program (2021) [29], which collects data about access, use, and perceived benefits from eHealth among Dutch citizens through questionnaires.

There is still limited understanding of the relationship between SEP and eHealth access, use, and perceived benefit and that understanding is generally limited to either access, use, or perceived benefit from eHealth, specific eHealth apps or specific subpopulations. Research showed the relation between SEP and the use of personal health records [35] and mobile apps [36]. Other research focuses on either the benefit from [37] or the use of eHealth [38,39] or specific patient groups such as cancer survivors [40,41], or citizens bound to specific locations [42]. To our knowledge, insight into differences in eHealth access, use, and perceived benefits and how different indicators for SEP display these differences within the Dutch general population aged 18 years and older are largely unknown. This study aims to assess differences in eHealth-related access, usage, and perceived benefits for different socioeconomic populations, based on education, standardized income, and socioeconomic status (SES) level of the neighborhood in the Netherlands. The findings of this study give insight into the disparities in access to, use of, and perceived benefit from eHealth in a highly connected country with an increasingly digitalized health care system. The results are insightful for other contexts that aim for or experience the same ambitions to transition to a digital health care system.

Methods

Panel

Data from the Dutch Health Care Consumer Panel (DHCCP) were used [43]. The DHCCP is a panel managed by Nivel (the Netherlands Institute for Health Services Research) and currently (as of September 2023) consists of approximately 11,500 panel members aged 18 years and older [43]. For this study, a study sample of 1.500 panel members was drawn by researchers from DHCCP. The study sample was representative of the Dutch population aged 18 years and older regarding age and sex [43]. Background characteristics of panel members, including their sex, age, level of education, net monthly income per household, and 4-digit postal codes were known. The panel was periodically renewed to ensure representative samples of the adult population in the Netherlands can be drawn. New panel members were recruited by buying an address file from an address supplier [43]. As a result, possible new members were sampled at random from the general population in the Netherlands [43]. The panel could only be joined through invitation. It was not possible for people to sign up on their own initiative [43]. Upon membership, panel members were informed of the purpose, scope, method,

XSL•FO
RenderX

and use of the panel [43]. Based on this information, participants could give permission to participate in the panel [43].

Ethical Considerations

According to the Dutch legislation, neither obtaining informed consent nor approval by a medical ethics committee was obligatory for doing research within the DHCCP [43,44]. Data analysis was conducted with pseudonymized data, according to the privacy regulations of the DHCCP, in compliance with the General Data Protection Regulation [44]. The privacy of the panel members was protected. All data were carefully stored by Nivel [43]. Personal information such as addresses was stored separately from the data of the questionnaires [43]. The privacy of the panel members in the study sample was guaranteed by DHCCP [43]. The researcher (LS) who analyzed the data had no access to the personal information of the panel members [43]. A written or digital informed consent was obtained at the time of registration of a new member to the panel [43]. Panel members were asked to participate approximately 4 or 5 times per year [43]. Participation was voluntary.

Data Collection

Data on the population's perspective on eHealth were collected via the DHCCP as part of a larger monitoring study into the perceptions, experiences, and usage of eHealth in the Netherlands [43]. A questionnaire was developed and reviewed by a team of representatives from the health care field in the Netherlands. The questionnaire was based on earlier distributed questionnaires of the monitoring study and was adjusted to reflect market developments [34]. The questionnaire was distributed via email and post (according to the preferences of the panel members) in May 2021. A digital reminder was sent after 1, 2, and 3 weeks after the start of the questionnaire and 1 written reminder was sent after 2 weeks. Panel members had 4 weeks to respond.

Socioeconomic Position Indicators

Overview

The concept of SEP is complex, as it results from the interaction between individual, social, economic, cultural, and societal factors [45,46]. In this study, 3 different operationalizations of SEP were used to study the digital divide in the context of eHealth—education as a historic starting point, standardized household income as a measure of current wealth, and SES level of the neighborhood to include environmental influence [46-49].

Education

The education levels were defined as low (none, primary school, or prevocational education); medium (secondary or vocational education level 1, 2, 3, or 4); and high (professional higher education or university) [50].

Standardized Income

Standardized income was defined as the net monthly income of the household adjusted for number of household members. The net income was converted to the equivalent of the net income of a single adult household by using equivalence factors from Statistics Netherlands (the Dutch Institute for Population Statistics) [51]. Some respondents acknowledged having children or other adults living in their household apart from their partner or children older than 18 years of age but did not specify the number. In the case of an unknown amount of children, 1.57 children were assumed. In case of an unknown amount of extra adults 1 extra adult was assumed. Assumptions were derived from the averages in Dutch households [50,52,53]. Information gathered from the panel members about their monthly net income was in ranges, and the mean of the range was taken as the monthly net income. Standardized income was divided into three categories (1) low (between €0 and €1659 [US $1718.56] per month; The used conversion rate was applicable on May 1, 2021), (2) medium (between €1660 [US $1719.59] and €2332 [US $2415.72] per month), and (3) high (more than €2332 [US $2415.72] per month). The categories were derived from the quartile distribution of the net income of the Dutch households (2020) [54].

Socioeconomic Status Level of the Neighborhood

The SES level of the neighborhood of all respondents was determined using the Social Economic Status-Wealth Education Employment (SES-WOA) score (2019) from the Statistics Netherlands. The SES-WOA score was based on the wealth, educational status, and recent employment history of households in the neighborhood [55,56]. The SES-WOA score was matched to the respondent by the 4-digit postal code. The neighborhoods of the respondents were categorized as (1) low (first tertile of SES score: –0.89 to 0.042), (2) medium (second tertile of SES score: 0.043-0.21), and (3) high (third tertile of SES score: 0.21-0.71). The average score of Dutch neighborhoods was 0.092 (SD 0.23; range –0.89 to 0.71) [55,56].

eHealth and the Digital Divide

For the interpretation of the data, the digital divide model was used. The digital divide model published by van Dijk et al [57] conceptualized that individuals' SEPs influence the available resources to access, use, and benefit from new digital media. In this study, the digital media in focus was eHealth.

Two concepts of eHealth were examined: (1) eHealth in general and (2) websites, apps, and wearables. Items in the questionnaire that informed these 2 concepts were matched to the 3 levels of the digital divide model [57], namely level 1—access, level 2—use, and level 3—perceived benefit. Operationalization of the variables measuring access, use, and perceived benefit for eHealth in general and websites, apps, and wearables can be found in Multimedia Appendix 1.

First, eHealth in general was studied to gain insight into an overall interest toward digital apps in health care. The levels of digital divide that were studied are access and the perceived benefit. Access was measured by motivation. Here respondents were asked what their general thoughts are about digital apps used in health care. The perceived benefit was operationalized by measuring to what extent the respondents perceived themselves as making more conscious decisions about their health as a result of eHealth use.

Second, eHealth in terms of websites, apps, and wearables was studied to gain insight into the use of these specific eHealth tools to improve health or provide support in coping with a disease. The digital divide levels that were taken into account

XSL•FO
RenderX

are access and use. Access was measured by motivation and physical access. Motivation was measured by asking the respondents if they have used or would like to use websites, apps, and wearables for their health. For physical access, the respondents were asked if they have access to an electronic device with internet.

The digital divide concept use was measured by barriers in use, diversity of use, and frequency of use. For the concept barriers in use, the respondents were asked if they experienced barriers in the use of websites, apps, and wearables. The use of websites, apps, and wearables was further operationalized in 2 variables—diversity of use and frequency of use. For these variables, the respondents were asked about 16 different websites, apps, or wearables if they used the app (once or more than once). Diversity entails the variety of websites, apps, and wearables used, while the frequency of use operationalizes the number of times (more than once) the apps were used.

Statistical Analysis

Descriptive analyses were used to describe the demographics and the outcome of the variables measuring the access, use, and perceived benefit for eHealth in general and websites, apps, and wearables. The variables measuring access, use, and perceived benefit were constructed by combining items from the original questionnaire. The operationalization of these variables can be found in Multimedia Appendix 1.

The outcomes were stratified by the 3 variables of SEP—educational level, standardized income, and SES level of the neighborhood. The differences in eHealth access and usage between SEP populations were investigated using logistic regression analysis. Ordered logistic regression was used for testing the differences in perceived benefit between SEP populations. Age was included in the analysis to test for confounding, as age is associated with both health and familiarity and use of digital media [10,58,59]. The correlation between the independent variables was determined via the Spearman rank order correlation coefficient. Data analysis was conducted using the Stata Statistical Software release (version 16.1; StataCorp) [60]. A $P<.05$ was considered statistically significant. Variables of physical access and diversity of use were not included in the logistic regression analysis because there were too few cases in outcome categories to meet the assumptions of the logistic regression analysis. The univariate outcomes are presented in Multimedia Appendix 2.

Results

Sample Characteristics

In total, 849 panel members responded to the questionnaire, resulting in a response rate of 56.6% (849/1500). Among the panel population of 1500, 8.9% (133/1500), 40% (600/1500), and 48.7% (731/1500) had a low, medium, and high level of education, respectively. Of these groups, 55.6% (74/133), 59.3% (356/600), and 54.3% (397/731) responded to the questionnaire. Regarding standardized income 35.8% (537/1500), 31.8% (477/1500), and 27.9 (419/1500) had low, medium, and high levels of standardized income, respectively. Of these groups, 57.2% (307/537), 56.4% (269/477), and 54.9% (230/269) had responded to the questionnaire. Finally, for SES-level of the neighborhood, 42.8% (642/1500), 32.6% (489/1500), and 23.1% (346/1500) had low, medium, and high levels of SES-level of the neighborhood in the panel population, respectively. Of these groups, 56.1% (360/642), 58.1% (284/489), and 55.2% (191/346) responded to the questionnaire.

The demographics of the study population can be found in Table 1. An overview of the study population stratified by education, standardized income, and SES level of the neighborhood can be found in Multimedia Appendix 3. Overall, the sample contained the same distribution of males or females as in the general population. When stratified for age category, our sample contained slightly more respondents aged 40 years and older and fewer respondents aged 18-39 years as compared to the general population. A frequency table of males and females in 3 age categories from the study population and the general Dutch population can be found in Multimedia Appendix 4 [61]. Compared to the Dutch general population, the study population had more males (11.3% general population, 14.9% study population) and females (13.1% general population, 15.4% study population) in the age category of 65 years and older. In terms of sex, the distribution was equal between both populations (Dutch general population of 49.3% male and 50.7% female, study population male 48.7% and female 51.3%). The mean age was 54 (SD 16.96) years. The lowest number of respondents was in the low educational level subpopulation (74/849, 8.72%) and the highest in the high educational level subpopulation (397/849, 46.76%). A high educational level was significantly and positively associated with a high standardized income level, as indicated by the correlations of educational level and standardized income level: $\rho=0.37$ ($P<.001$), educational level and SES level of the neighborhood: $\rho=0.021$ ($P=.55$), and standardized income level and SES level of the neighborhood: $\rho=0.035$ ($P=.33$) [62].

XSL•FO
RenderX

Table 1. Demographic description of the study population (n=849). Study population was sampled (2021) from a representative population (N=1500) of general Dutch population aged 18 years and older.

	Total (n=849)
Sex, n (%)	
Male	413 (48.6)
Female	435 (51.2)
Missing	1 (0)
Age (years)	
Mean (SD)	54 (16.96)
Range	19-92
Average household size	
Mean (SD)	2.29 (1.11)
Range	1-7
Number of households with children younger than 18 years, n (%)	
No	604 (71.1)
Yes	235 (27.7)
Missing	10 (1.2)
Number of households with children older than 18 years, n (%)	
No	763 (89.9)
Yes	76 (9)
Missing	10 (1.2)
Education	
Mean (SD)	2.39 (0.65)
Range	1-3
Low, n (%)	74 (8.7)
Medium, n (%)	356 (41.9)
High, n (%)	397 (46.8)
Missing, n (%)	22 (2.6)
Standardized income	
Mean (SD)	1.90 (0.81)
Range	1-3
Low, n (%)	307 (36.2)
Medium, n (%)	269 (31.7)
High, n (%)	230 (27.1)
Missing, n (%)	43 (5.1)
SES[b] level of the neighborhood	
Mean (SD)	1.80 (0.79)
Range	1-3
Low, n (%)	360 (42.4)
Medium, n (%)	284 (33.5)
High, n (%)	191 (22.5)
Missing, n (%)	14 (1.7)

[a]Education level—low (none, primary school or prevocational education); medium (secondary or vocational education level 1, 2, 3, or 4); and high (professional higher education or university). Standardized income was divided into 3 categories—low (between €0 and €1659 [US $1718.56] per month); medium (between €1660 [US $1719.59] and €2332 [US $2415.72] per month); and high (more than €2332 [US $2415.72] per month). The

XSL•FO
RenderX

SES level of the neighborhood was determined using the Social Economic Status-Wealth Education Employment (SES-WOA) score (2019) from Statistics Netherlands. The SES-WOA score was based on the wealth, educational status, and recent employment history of households in the neighborhood [55,56]. Categories: low (first tertile of SES score: –0.89 to 0.042); medium (second tertile of SES score: 0.043-0.21); and high (third tertile of SES score: 0.21-0.71). Not all values add up to 100% due to missing values.

[b]SES: socioeconomic status.

Results of the Relation Between the Digital Divide Levels and the Socioeconomic Position Indicators

Overview

An overview of the measured and analyzed variables can be found in Figure 1. The frequencies of these outcomes can be found in Multimedia Appendix 2. Table 2 presents a descriptive overview of access, use, and perceived benefit, stratified by SEP indicators. The results showed that the outcome differed the most when the population was stratified by education. Figure 2 presents the associations between access; use; perceived benefit; low, medium, and high education; standardized income; and SES level of the neighborhood populations. The underlying data used for Figure 2 is presented in Multimedia Appendix 5.

Figure 1. Overview of the concepts of access, use, and perceived benefit, the variables matched to these concepts and the socioeconomic position indicators used in this study. Legend indicates whether the variables were matched for eHealth in general (bold), websites apps and wearables (underlined) or both. Second, the legend indicates whether the variables were viable, either included (black) or not included (grey), for logistic (ordered) regression analysis [57]. SES: socioeconomic status.

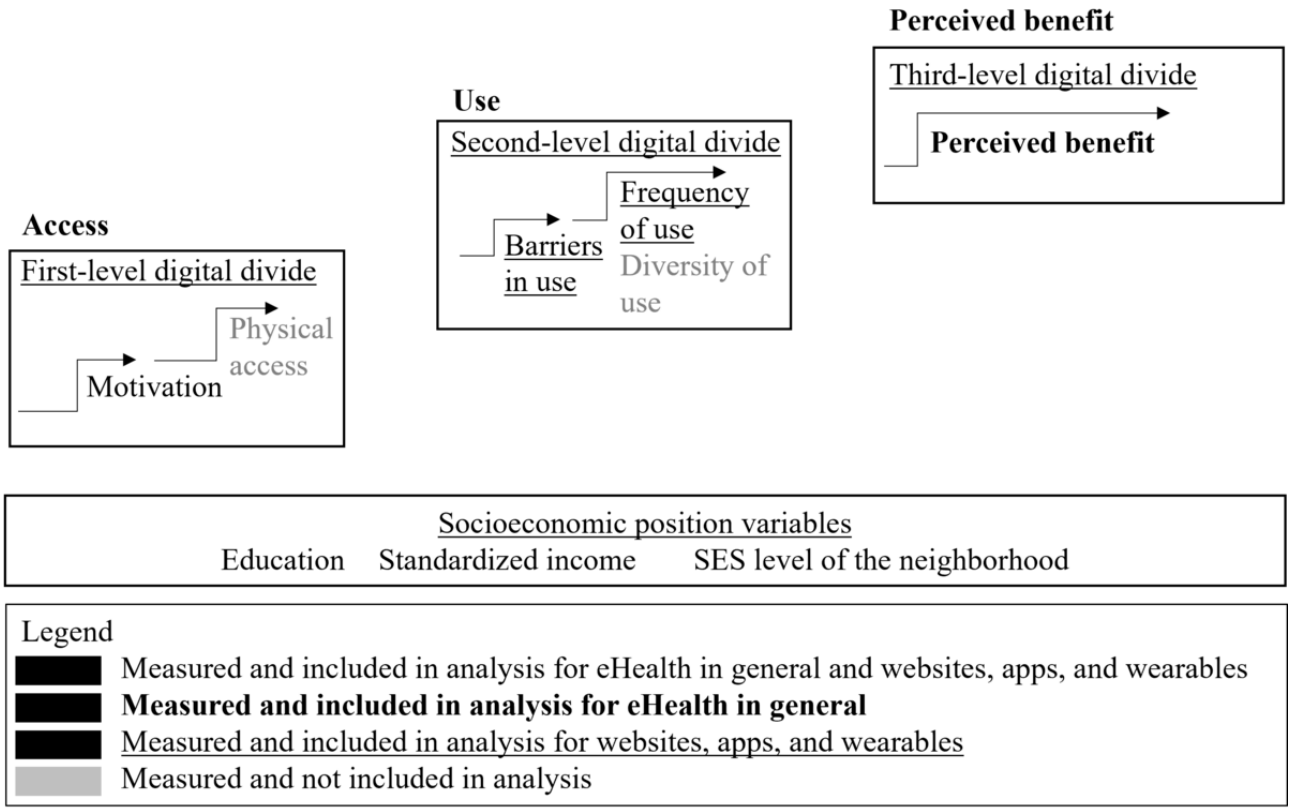

XSL•FO
RenderX

Figure 2. The relation between access, use, and benefit of eHealth and low, medium, and high levels of 3 SEP indicators—education, standardized income, and SES-level of the neighborhood. Results from a questionnaire (2021) answered by a representative study population (n=849) for the general Dutch population aged 18 years and older. Access: motivation, use: barriers in use and frequency of use, and perceived benefit: perceived benefit. For each SEP indicator, a bar graph of the results of logistic (ordered) regression analysis was presented. Each bar shows the odds ratio (OR) and the 95% CI for the difference between SEP levels for each digital divide concept. For each SEP indicator, the comparisons made were medium compared to low (dark grey), high compared to low (light grey), and high compared to medium (medium grey). The number of respondents (n) included in the analysis for each digital divide concept was indicated. The variable barriers in use were recoded to ensure that a positive score (1) reflected the outcome: no experienced barriers. Positive ORs (OR>1) should be interpreted that the primary group in the comparison, in the case of this study either medium or high SEP, had more likelihood of not experiencing barriers in use. **P*>.05, ***P*>.01, ****P*>.001. SEP: socioeconomic position; SES: socioeconomic status.

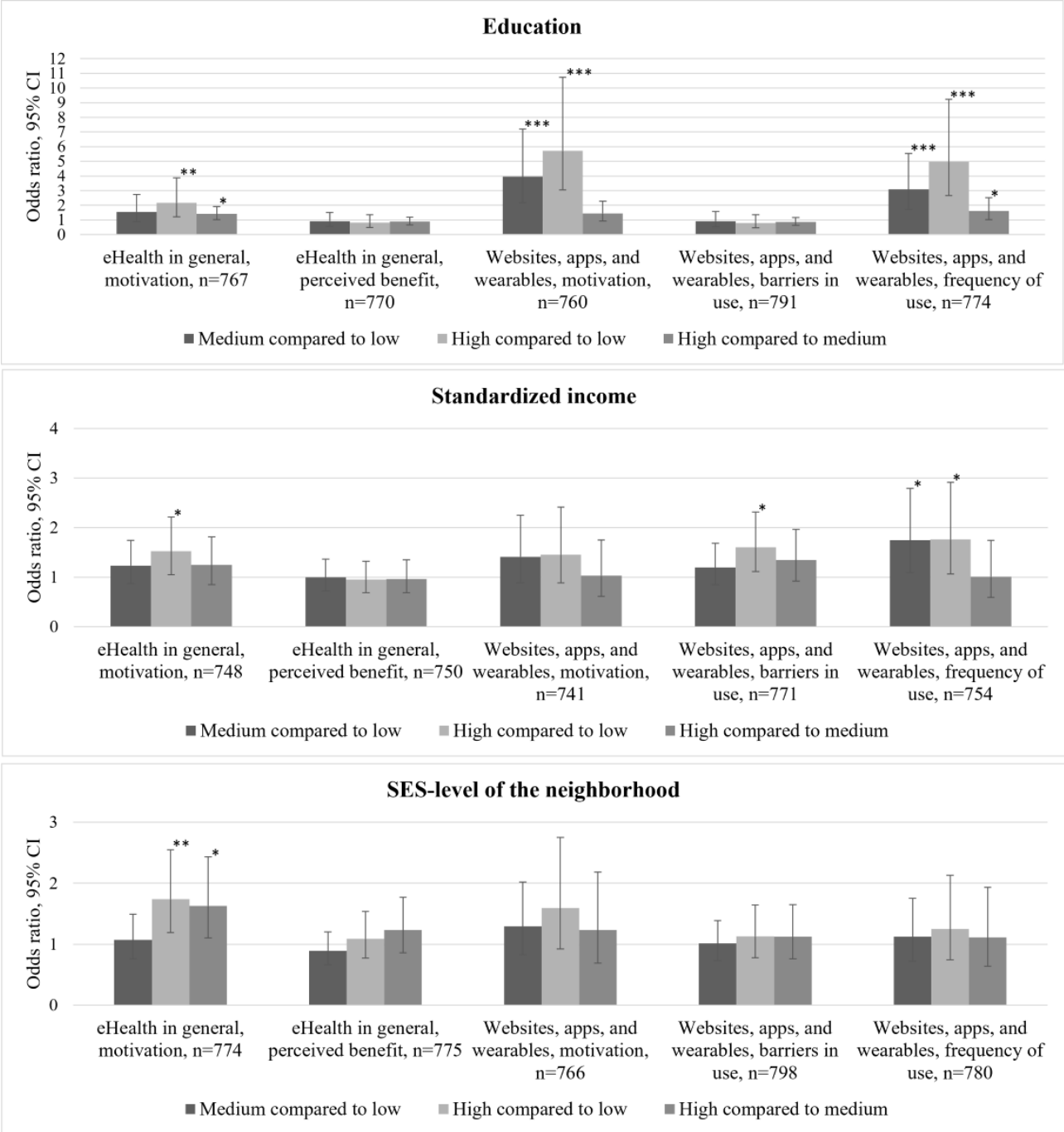

XSL•FO
RenderX

Table 2. Frequency distribution of access, use, and perceived benefit for (1) eHealth in general and (2) websites, apps, and wearables, stratified by education, standardized income, and SES[a] level of the neighborhood. Frequencies are derived from results of a questionnaire (2021) conducted among a sample of the general Dutch population (N=1500), final study population (n=849)[b,c].

	Total	Education			Standardized income			SES level of the neighborhood		
		Low	Med[d]	High	Low	Med	High	Low	Med	High
Population, n	849	74	356	397	307	269	230	360	284	191
eHealth in general										
Access—motivation, n (%)										
No motivation	386 (45.5)	46 (62.2)	177 (49.7)	152 (38.3)	153 (49.8)	124 (46.1)	90 (39.1)	174 (48.3)	141 (49.6)	68 (35.6)
Motivation	401 (47.2)	21 (28.4)	153 (43.0)	218 (54.9)	131 (42.7)	127 (47.2)	123 (53.5)	157 (43.6)	127 (44.7)	107 (56)
Perceived benefit—perceived benefit, n (%)										
Totally disagree—disagree	211 (24.9)	17 (23.0)	80 (22.4)	108 (27.2)	70 (22.8)	68 (25.3)	65 (28.3)	85 (23.6)	76 (26.7)	46 (24.1)
Not agree nor disagree	370 (43.6)	33 (44.5)	171 (48.0)	159 (40.1)	144 (46.9)	119 (44.2)	89 (38.7)	158 (43.9)	134 (47.2)	75 (39.3)
Agree—totally agree	205 (24.1)	17 (23.0)	81 (22.8)	104 (26.2)	69 (22.5)	65 (24.2)	61 (26.5)	87 (24.2)	61 (21.5)	53 (27.7)
Websites, apps, and wearables										
Access—motivation, n (%)										
No motivation	143 (16.8)	35 (47.3)	62 (17.4)	41 (10.3)	59 (19.2)	43 (16.0)	34 (14.8)	65 (18.1)	50 (17.6)	23 (12.0)
Motivation	636 (75.0)	29 (39.2)	265 (74.4)	328 (82.6)	221 (72.0)	206 (76.6)	178 (77.4)	263 (73.1)	215 (75.7)	150 (78.5)
Use—barriers in use, n (%)										
Experienced barriers	339 (40.0)	29 (39.2)	140 (39.3)	159 (40.1)	135 (44.0)	109 (40.5)	76 (33.0)	145 (40.3)	119 (41.9)	71 (37.2)
No experienced barriers	472 (55.6)	39 (52.7)	201 (56.5)	223 (56.2)	157 (51.1)	149 (55.4)	145 (63.0)	197 (54.7)	156 (54.9)	110 (57.6)
Use—frequency of use, n (%)										
No frequent use	131 (15.4)	29 (39.2)	60 (16.9)	38 (9.6)	59 (19.2)	36 (13.4)	29 (12.6)	57 (15.8)	47 (16.5)	24 (12.6)
Frequent use	662 (78.0)	34 (45.9)	274 (77.0)	339 (85.4)	222 (72.3)	219 (81.4)	189 (82.2)	277 (76.9)	226 (79.6)	149 (78.0)

[a]SES: socioeconomic status.

[b]The variables physical access and diversity of use for websites, apps, and wearables were not presented in Table 2 because these variables could not meet the assumptions for logistic regression analysis. The frequencies of these outcomes can be found in Multimedia Appendix 2.

[c]Not all values add up to 100% due to missing values.

[d]Med: medium level.

Access and Motivation

For eHealth in general, as well as for websites, apps and wearables, differences in motivation were found between different levels of education. Differences in motivation were most profound between low or medium versus highly educated respondents in both eHealth in general (odds ratio [OR] 2.18, 95% CI 1.22-3.88) and websites, apps, and wearables (OR 5.72, 95% CI 3.06-10.72). Regarding standardized income, a difference in motivation for eHealth in general was found between high and low standardized incomes (OR 1.52, 95% CI 1.05-2.21). A significant difference in motivation between low (OR 1.74, 95% CI 1.19-2.55) and medium (OR 1.63, 95% CI 1.1-2.43) versus high SES level of the neighborhood was found for eHealth in general.

Use, Barriers in Use, and Experienced Barriers

High standardized income was associated with no experience of barriers in use in comparison to low and medium levels of standardized income. This implies that fewer respondents with a high standardized income experienced barriers while using eHealth websites, apps, and wearables (OR 1.60, 95% CI 1.11-2.31). The frequency of eHealth use also differed between respondents with a low, medium, or high level of education,

XSL•FO
RenderX

with the most significant difference between high and low educational levels (OR 4.96, 95% CI 2.66-9.24). In terms of standardized income, there were significant differences between low and high (OR 1.76, 95% CI 1.06-2.91) and low and medium (OR 1.74, 95% CI 1.09-2.79) standardized income levels. High SEP respondents were more likely to frequently use an eHealth app, website, or wearable compared to medium or low SEP respondents.

Perceived Benefit

There were no significant differences found regarding the perceived benefits between low, medium, or high-SEP populations.

Discussion

Principal Findings

This study shows that low SEP respondents have less access to and use of eHealth compared to medium or high SEP respondents. The most significant digital divide observed in this study is related to educational background. The results of this paper contribute on a population level to previous findings that the access and use of eHealth has a socioeconomic gradient. Additionally, the results emphasize that SEP indicators cannot be used interchangeably to assess eHealth access and use.

The results of this study highlight that, across all 3 SEP indicators, the most substantial differences are found in access through motivation. Respondents from higher socioeconomic categories expressed greater motivation to use eHealth, including websites, apps, and wearables, in comparison to those from lower SEPs. Health equity researchers emphasize that comprehending how people perceive health and health care is a complex issue influenced by several societal, contextual, social, and biological individual factors. Therefore, a multidimensional and multicausal approach is necessary to comprehend these disparities [24,63,64]. Weiss et al [28] discuss existing literature and multiple theories as to why differences in eHealth access and use exist between socioeconomic groups. The literature described that the social position of individuals and the influence of the context and organizations surrounding the individual play a role in whether individuals choose to consume digital health care or not [28]. Other literature emphasized that the diffusion of digital health care in society will decrease the digital divide gap as the low-SEP population is assumed to be the latest to adopt [28]. However, the role of health care organizations, social policies, and political decisions that impact individuals' motivations toward eHealth has not been adequately studied. Current national Dutch policies concerning eHealth are focused on the development and interoperability of eHealth apps, the digital skills of health care professionals and the use of eHealth by older people at home, resulting in eHealth becoming an essential part of the health care system [65]. In countries where health care digitalization is progressing, it would be valuable to examine the potential influence of the government, health care organizations, and businesses on the motivation of low socioeconomic populations toward eHealth. Such research could offer valuable insights into the societal and policy changes required to make eHealth more appealing to low-SEP populations.

In this study, use was examined by studying the barriers in use and frequency of use of eHealth websites, apps, and wearables. The results demonstrated that respondents with a high standardized income level infrequently experienced barriers to using eHealth websites, apps, and wearables compared to lower standardized income levels. Previous studies show that highly educated people often have higher health literacy levels and digital skills and are more in contact with the digital world via their education or profession [10,31,59,66-70]. Other studies show that in the development of eHealth new apps are often pilot-tested by highly educated respondents and, therefore, might be more tailored to the needs of highly educated individuals [9,71,72]. In contrast, other studies point out the variety of health behaviors and health care use within SEP groups. De Boer et al [73] show that low SEP groups have more health care costs but that healthy lifestyle behavior such as smoking or being member of a sports club are attributed greatly to the variety in health care use in each socioeconomic group. In the light of eHealth, Agachi et al [74] showed that the user interface and the type of eHealth offered attributes to the use of eHealth between socioeconomic groups. Results revealed that for the same eHealth program, people living in a low socioeconomic neighborhood use the app-based tool more than people living high in a socioeconomic neighborhood. For the web-based version, results show the opposite emphasizing the importance of the user interface and the accessibility to digital devices, as is also theorized in the digital divide model [49,74]. Both De Boer [73] and Agachi [74] show that behavior-related and technical-related factors play a role in the use of health care and eHealth. The results of this study and other studies showed that understanding and creating insight into the existence of and possible solutions for health disparities is dependent on multiple dimensions. Future research into how the socioeconomic gradient in eHealth access and use are associated with other behavioral and technical factors is important to create an in-depth understanding of disparities in eHealth access and use that can inspire research, policy, and practice.

Results pointed out that high education and high standardized income levels were associated with frequent use of websites, apps, and wearables. This is in accordance with previous research, which shows that a high level of education and income is found to be associated with more access to and use of eHealth [12,13,31]. Surprisingly, no difference was found in the perceived benefits (making more conscious decisions in health because of eHealth use) in any of the SEP indicators, although the frequency of use of websites, apps, and wearables was high (78%) and did show significant differences. This implies that respondents with more frequent use of websites, apps, and wearables had the same perceived benefit, namely, making more conscious decisions due to eHealth use, compared to respondents who have not used websites, apps, and wearables once. Evaluation studies show that high SEP respondents have better outcomes from eHealth use than low SEP respondents [10,30,31,59]. The results of this study might indicate that in the real-life context, even though current eHealth apps might be more suitable for highly educated individuals, eHealth is not used appropriately or with similar discipline as in the clinical trial context.

XSL•FO
RenderX

Strengths and Limitations

The strengths of this study were the use of a large and representative sample of the Dutch population and the use of 3 SEP operationalizations to provide a broad insight into the effects of SEP on the digital divide. This study also has some limitations. Despite a large number of respondents, the skewed distribution of outcomes across SEP levels hampered the performance of multivariable logistic regression analysis [62]. Next to this, the SEP indicators used are focused on social demographic and economic aspects of SEP. Cultural and other social aspects, such as social networks and cultural background, have not been taken into account.

The questionnaire used in the study was not designed to measure all the different aspects of the digital divide model and is a secondary analysis of the data gathered. Although the majority of digital divide levels could be well matched with questionnaire items, data on the second level digital divide for eHealth in general and the third digital divide level for websites, apps, and wearables was lacking. In some cases, with emphasis on the concept barriers in use for websites, apps, and wearables more in-depth insights into the digital divide levels would have been desirable to improve the validation of the findings. Van Dijk [57] provides concepts to further define the digital divide levels. Access, the first level of digital divide, is described to entail the concepts of motivational access and physical access. Use, the second level of digital divide entails the concepts of digital skills and usage. Usage here encompasses both frequency and diversity of digital media use. Perceived benefit, the third level digital divide, is conceptualized by personal outcomes that are a result of the use of digital media [57]. In this study, the concept barriers in use were used instead of digital skills, as experienced barriers are not limited to barriers formed by a lack of digital skills.

Additionally, for the first and the second level digital divide, the technical design and the information and communication technology of eHealth are of importance [57]. The technical design and the information and communication technology imply factors such as accessibility, usability, mobility, quality, and accessibility of internet access and automation (self-learning devices or software tailored to serve the consumer better and automatically) of devices and apps. These factors are important to facilitate adherence and appropriate use of eHealth apps [57]. The questionnaire provided no insight into these factors.

Conclusions

The results of this study revealed that differences in motivation for eHealth use are most profound between different socioeconomic populations in the Dutch society, in which low-educated people are likely to be disadvantaged. A successful transition toward digital health care is a social issue that is dependent on the motivation to use eHealth and specific apps. It is imperative that future studies within academia and within the health care field focus on the motivations and needs associated with digital health care, specifically for low-SEP populations. Research on the societal changes stemming from the digital health care transition and the technical and design studies of digital care apps in single-intervention studies are both vital for the realization of an inclusive and comprehensive digital health care system. If eHealth takes a predominant role in the Dutch health care system, it might affect access and use of health care for the citizens who need it the most. The results of this study underscore the importance of policies aimed at facilitating and supporting low-SEP populations in the use of eHealth to reduce differences in health.

Acknowledgments

The authors would like to thank E de Vries and L van Deursen for providing comments and suggestions. This study was funded by the Dutch Ministry of Health, Welfare, and Sport. The funders had no role in the study design, data collection and analysis, data interpretation, writing of the paper, or approval for publication.

Data Availability

The datasets generated and analyzed during this study are available upon reasonable request from Prof Judith D de Jong (j.dejong@nivel.nl), project leader of the Dutch Health Care Consumer Panel, or the panel's secretary (conusmentenpanel@nivel.nl).

Conflicts of Interest

None to declare.

Multimedia Appendix 1
Operationalization of access, use, and perceived benefit of eHealth in general and websites, apps, and wearables.
[DOCX File , 24 KB - formative_v9i1e49585_app1.docx]

Multimedia Appendix 2
Frequency table of the outcomes of access, use, and perceived benefit for eHealth in general and websites, apps, and wearables gathered via a questionnaire (2021).
[DOCX File , 24 KB - formative_v9i1e49585_app2.docx]

XSL•FO
RenderX

Multimedia Appendix 3
Demographic description of the study population and the study population stratified by levels of education, standardized income, and SES level of the neighborhood.
[DOCX File , 31 KB - formative_v9i1e49585_app3.docx]

Multimedia Appendix 4
Frequency table of males and females aged above 18 in three age categories of the Dutch general population (2021) and the study population (2021).
[DOCX File , 14 KB - formative_v9i1e49585_app4.docx]

Multimedia Appendix 5
Results from logistic (ordered) regression analysis.
[DOCX File , 21 KB - formative_v9i1e49585_app5.docx]

References

1. Rutten P, Pruim J, Zijl NV, Merckelbach S. Digitale zorg in Nederland. Digitale zorg in Nederland. 2020. URL: https://www.mckinsey.com/nl/our-insights/digitale-zorg-in-nederland [accessed 2023-05-05]
2. Abernethy A, Adams L, Barrett M, Bechtel C, Brennan P, Butte A, et al. The promise of digital health: then, now, and the future. NAM Perspect 2022 Jun 27:202206e [FREE Full text] [doi: 10.31478/202206e] [Medline: 36177208]
3. Pagliari C, Sloan D, Gregor P, Sullivan F, Detmer D, Kahan JP, et al. What is eHealth (4): a scoping exercise to map the field. J Med Internet Res 2005 Mar 31;7(1):e9 [FREE Full text] [doi: 10.2196/jmir.7.1.e9] [Medline: 15829481]
4. Dutch digitisation strategy 2.0. Ministery of Economic Affairs and Climate Policy. 2019. URL: https://www.nederlanddigitaal.nl/documenten/publicaties/2019/11/13/english-version-of-the-dutch-digitalisation-strategy-2.0 [accessed 2023-05-05]
5. Digital economy and society index (DESI) 2021: The Netherlands. European Commission. 2021. URL: https://ec.europa.eu/newsroom/dae/redirection/document/80493 [accessed 2022-02-01]
6. 2030 digital compass: the European way for the digital decade. European Commission. 2021. URL: https://eufordigital.eu/library/2030-digital-compass-the-european-way-for-the-digital-decade/ [accessed 2022-02-01]
7. Boers SN, Jongsma KR, Lucivero F, Aardoom J, Büchner FL, de Vries M, et al. SERIES: eHealth in primary care. Part 2: Exploring the ethical implications of its application in primary care practice. Eur J Gen Pract 2019 Oct 30;26(1):26-32 [FREE Full text] [doi: 10.1080/13814788.2019.1678958] [Medline: 31663394]
8. European C. Protecting fundamental rights in the digital age - 2021 annual report on the application of the EU charter of fundamental rights. European Commission. 2021. URL: https://ec.europa.eu/commission/presscorner/detail/en/ip_21_6699 [accessed 2023-04-06]
9. Goedhart NS, Zuiderent-Jerak T, Woudstra J, Broerse JEW, Betten AW, Dedding C. Persistent inequitable design and implementation of patient portals for users at the margins. J Am Med Inform Assoc 2021 Feb 15;28(2):276-283 [FREE Full text] [doi: 10.1093/jamia/ocaa273] [Medline: 33463691]
10. Latulippe K, Hamel C, Giroux D. Social health inequalities and eHealth: a literature review with qualitative synthesis of theoretical and empirical studies. J Med Internet Res 2017 Apr 27;19(4):e136 [FREE Full text] [doi: 10.2196/jmir.6731] [Medline: 28450271]
11. McAuley A. Digital health interventions: widening access or widening inequalities? Public Health 2014 Dec;128(12):1118-1120. [doi: 10.1016/j.puhe.2014.10.008] [Medline: 25458115]
12. Kontos E, Blake KD, Chou WYS, Prestin A. Predictors of eHealth usage: insights on the digital divide from the Health Information National Trends Survey 2012. J Med Internet Res 2014 Jul 16;16(7):e172 [FREE Full text] [doi: 10.2196/jmir.3117] [Medline: 25048379]
13. Perez SL, Kravitz RL, Bell RA, Chan MS, Paterniti DA. Characterizing internet health information seeking strategies by socioeconomic status: a mixed methods approach. BMC Med Inform Decis Mak 2016 Aug 09;16:107. [doi: 10.1186/s12911-016-0344-x] [Medline: 27506607]
14. Scherer R, Siddiq F. The relation between students' socioeconomic status and ICT literacy: findings from a meta-analysis. Comput Educ 2019;138:13-32. [doi: 10.1016/j.compedu.2019.04.011]
15. Azzopardi-Muscat N, Sørensen K. Towards an equitable digital public health era: promoting equity through a health literacy perspective. Eur J Public Health 2019;29(Supplement_3):13-17 [FREE Full text] [doi: 10.1093/eurpub/ckz166] [Medline: 31738443]
16. Wong BLH, Maaß L, Vodden A, van Kessel R, Sorbello S, Buttigieg S, European Public Health Association (EUPHA) Digital Health Section. The dawn of digital public health in Europe: implications for public health policy and practice. Lancet Reg Health Eur 2022;14:100316 [FREE Full text] [doi: 10.1016/j.lanepe.2022.100316] [Medline: 35132399]
17. Raad VVES. Een eerlijke kans op gezond leven. Raad voor Volksgezondheid en Samenleving. 2021. URL: https://www.raadrvs.nl/documenten/publicaties/2021/04/07/een-eerlijke-kans-op-gezond-leven [accessed 2021-10-13]

XSL•FO
RenderX

18. de Boer WIJ, Buskens E, Koning RH, Mierau JO. Neighborhood socioeconomic status and health care costs: a population-wide study in the Netherlands. Am J Public Health 2019;109(6):927-933. [doi: 10.2105/AJPH.2019.305035] [Medline: 30998412]
19. Loef B, Meulman I, Herber GM, Kommer GJ, Koopmanschap MA, Kunst AE, et al. Socioeconomic differences in healthcare expenditure and utilization in the Netherlands. BMC Health Serv Res 2021;21(1):643 [FREE Full text] [doi: 10.1186/s12913-021-06694-9] [Medline: 34217287]
20. Petrovic D, de Mestral C, Bochud M, Bartley M, Kivimäki M, Vineis P, et al. The contribution of health behaviors to socioeconomic inequalities in health: a systematic review. Prev Med 2018;113:15-31 [FREE Full text] [doi: 10.1016/j.ypmed.2018.05.003] [Medline: 29752959]
21. Phelan JC, Link BG, Tehranifar P. Social conditions as fundamental causes of health inequalities: theory, evidence, and policy implications. J Health Soc Behav 2010;51 Suppl:S28-S40. [doi: 10.1177/0022146510383498] [Medline: 20943581]
22. Agerholm J, Bruce D, Ponce de Leon A, Burström B. Socioeconomic differences in healthcare utilization, with and without adjustment for need: an example from Stockholm, Sweden. Scand J Public Health 2013;41(3):318-325. [doi: 10.1177/1403494812473205] [Medline: 23406653]
23. Lemstra M, Mackenbach J, Neudorf C, Nannapaneni U. High health care utilization and costs associated with lower socio-economic status: results from a linked dataset. Can J Public Health 2009;100(3):180-183 [FREE Full text] [doi: 10.1007/BF03405536] [Medline: 19507718]
24. Mackenbach JP. The persistence of health inequalities in modern welfare states: the explanation of a paradox. Soc Sci Med 2012;75(4):761-769. [doi: 10.1016/j.socscimed.2012.02.031] [Medline: 22475407]
25. Oude Groeniger J, Kamphuis CBM, Mackenbach JP, Beenackers MA, van Lenthe FJ. Are socio-economic inequalities in diet and physical activity a matter of social distinction? A cross-sectional study. Int J Public Health 2019;64(7):1037-1047 [FREE Full text] [doi: 10.1007/s00038-019-01268-3] [Medline: 31187165]
26. Pampel FC, Krueger PM, Denney JT. Socioeconomic disparities in health behaviors. Annu Rev Sociol 2010;36:349-370. [doi: 10.1146/annurev.soc.012809.102529] [Medline: 21909182]
27. Singh-Manoux A, Marmot M. Role of socialization in explaining social inequalities in health. Soc Sci Med 2005;60(9):2129-2133 [FREE Full text] [doi: 10.1016/j.socscimed.2004.08.070] [Medline: 15743660]
28. Weiss D, Rydland HT, Øversveen E, Jensen MR, Solhaug S, Krokstad S. Innovative technologies and social inequalities in health: a scoping review of the literature. PLoS One 2018;13(4):e0195447 [FREE Full text] [doi: 10.1371/journal.pone.0195447] [Medline: 29614114]
29. van der VR, van DL, Standaar LMB, Wouters EJM, Suijkerbuijk AWM, van TL. E-healthmonitor 2021: stand van zaken digitale zorg. 2022. URL: https://www.rivm.nl/documenten/e-healthmonitor-2021-stand-van-zaken-digitale-zorg [accessed 2023-05-05]
30. Western MJ, Armstrong MEG, Islam I, Morgan K, Jones UF, Kelson MJ. The effectiveness of digital interventions for increasing physical activity in individuals of low socioeconomic status: a systematic review and meta-analysis. Int J Behav Nutr Phys Act 2021;18(1):148 [FREE Full text] [doi: 10.1186/s12966-021-01218-4] [Medline: 34753490]
31. Hansen AH, Bradway M, Broz J, Claudi T, Henriksen Ø, Wangberg SC, et al. Inequalities in the use of eHealth between socioeconomic groups among patients with type 1 and type 2 diabetes: cross-sectional study. J Med Internet Res 2019;21(5):e13615 [FREE Full text] [doi: 10.2196/13615] [Medline: 31144669]
32. Verhellen A, Jewell C, Van Gils M, Jacobs A, Steenberghs E. Ehealthmonitor Citizens. 2019. URL: https://researchportal.vub.be/en/publications/ehealthmonitor-2019-citizens [accessed 2023-09-13]
33. Vehko T, Ruotsalainen R, Hyppönen H, Aalto A, Terveyden ja hyvinvoinnin laitos. E-health and e-welfare of Finland: check point 2018. 2019. URL: https://oula.finna.fi/Record/oy.9917519693906252?lng=en-gb [accessed 2023-09-13]
34. Huygens MWJ, Voogdt-Pruis HR, Wouters M, Meurs MM, van Lettow B, Kleijweg C, et al. The uptake and use of telemonitoring in chronic care between 2014 and 2019: nationwide survey among patients and health care professionals in the Netherlands. J Med Internet Res 2021;23(5):e24908 [FREE Full text] [doi: 10.2196/24908] [Medline: 33938808]
35. Paccoud I, Baumann M, Le Bihan E, Pétré B, Breinbauer M, Böhme P, et al. Socioeconomic and behavioural factors associated with access to and use of personal health records. BMC Med Inform Decis Mak 2021;21(1):18 [FREE Full text] [doi: 10.1186/s12911-020-01383-9] [Medline: 33435970]
36. Bol N, Helberger N, Weert JCM. Differences in mobile health app use: a source of new digital inequalities? Inf Soc 2018;34(3):183-193. [doi: 10.1080/01972243.2018.1438550]
37. Van Deursen AJM, Helsper EJ. The third-level digital divide: who benefits most from being online. In: Robinson, L, Cotten SR, Schulz J, Hale TM, Williams A, editors. Communication and Information Technologies Annual. Leeds, England: Emerald Group Publishing Limited; 2015:29-52.
38. Chagpar AB. Sociodemographic factors affecting telemedicine access: a population-based analysis. Surgery 2022;171(3):793-798 [FREE Full text] [doi: 10.1016/j.surg.2021.08.059] [Medline: 34848076]
39. Aldhahir AM, Alqahtani JS, Althobiani MA, Alghamdi SM, Alanazi AF, Alnaim N, et al. Current knowledge, satisfaction, and use of E-Health mobile application (Seha) among the general population of Saudi Arabia: a cross-sectional study. J Multidiscip Healthc 2022;15:667-678 [FREE Full text] [doi: 10.2147/JMDH.S355093] [Medline: 35399807]

XSL•FO
RenderX

40. Moon Z, Zuchowski M, Moss-Morris R, Hunter MS, Norton S, Hughes LD. Disparities in access to mobile devices and e-health literacy among breast cancer survivors. Support Care Cancer 2022;30(1):117-126 [FREE Full text] [doi: 10.1007/s00520-021-06407-2] [Medline: 34236506]
41. Melhem SJ, Nabhani-Gebara S, Kayyali R. Digital trends, digital literacy, and E-Health engagement predictors of breast and colorectal cancer survivors: a population-based cross-sectional survey. Int J Environ Res Public Health 2023;20(2):1472 [FREE Full text] [doi: 10.3390/ijerph20021472] [Medline: 36674237]
42. Özdemir R, Bektemur G, Keles E, Baydili KN. Internet use, e-Heath literacy, and associated factors in Istanbul, Turkey: a cross-sectional study. J Consumer Health Internet 2023;27(1):1-11 [FREE Full text] [doi: 10.1080/15398285.2022.2129178]
43. Brabers AEM, de JJD. Nivel Consumentenpanel Gezondheidszorg Basisrapport met informatie over het panel 2022. NIVEL. 2022. URL: https://www.nivel.nl/nl/publicatie/nivel-consumentenpanel-gezondheidszorg-basisrapport-met-informatie-over-het-panel-2022 [accessed 2023-04-06]
44. Your research: is it subject to the WMO or not? Central Committee on Research Involving Human Subjects. URL: https://english.ccmo.nl/investigators/legal-framework-for-medical-scientific-research/your-research-is-it-subject-to-the-wmo-or-not [accessed 2024-06-13]
45. Giddens A. The Constitution of Society: Outline of the Theory of Structuration. Berkeley: University of California Press; 1984.
46. Bourdieu P. Outline of a Theory of Practice. Cambridge: Cambridge University Press; 1977.
47. Krieger N, Williams DR, Moss NE. Measuring social class in US public health research: concepts, methodologies, and guidelines. Annu Rev Public Health 1997;18:341-378. [doi: 10.1146/annurev.publhealth.18.1.341] [Medline: 9143723]
48. Scheerder A, van Deursen A, van Dijk J. Determinants of internet skills, uses and outcomes. A systematic review of the second- and third-level digital divide. Telematics Inform 2017;34(8):1607-1624. [doi: 10.1016/j.tele.2017.07.007]
49. Vrooman J, Boelhouwer J, Iedema J, van der T. Summary contemporary inequality. The Netherlands Institute for Social Research. 2023. URL: https://english.scp.nl/publications/publications/2023/03/07/summary-contemporary-inequality [accessed 2023-05-09]
50. Opleidingsniveau. Statistics Netherlands. 2019. URL: https://www.cbs.nl/nl-nl/nieuws/2019/33/verschil-levensverwachting-hoog-en-laagopgeleid-groeit/opleidingsniveau [accessed 2022-01-27]
51. What is my spendable income? Statistics Netherlands. 2018. URL: https://www.cbs.nl/en-gb/background/2008/50/what-is-my-spendable-income- [accessed 2023-04-06]
52. Geboorte; kerncijfers. StatLine. 2022. URL: https://opendata.cbs.nl/statline/#/CBS/nl/dataset/37422ned/table [accessed 2022-02-20]
53. Ruim 140 duizend samenwonende volwassen kinderen met ouders op leeftijd. Centraal Bureau van de Statistiek. 2014. URL: https://www.cbs.nl/nl-nl/achtergrond/2014/27/ruim-140-duizend-samenwonende-volwassen-kinderen-met-ouders-op-leeftijd [accessed 2022-08-02]
54. Income distribution (standardised income). Centraal Bureau van de Statistiek. 2019. URL: https://www.cbs.nl/en-gb/visualisations/income-distribution [accessed 2021-08-04]
55. Statusscore per wijk en buurt o.b.v. welvaart, opleidingsniveau en arbeid. Centraal Bureau van de Statistiek. 2022. URL: https://www.cbs.nl/nl-nl/achtergrond/2022/14/statusscore-per-wijk-en-buurt-o-b-v-welvaart-opleidingsniveau-en-arbeid [accessed 2022-08-01]
56. Sociaal-economische status; scores per wijk en buurt, regio-indeling 2021. Centraal Bureau van de Statistiek. 2022. URL: https://www.cbs.nl/nl-nl/cijfers/detail/85163NED [accessed 2022-08-01]
57. Van Dijk J. The Digital Divide. Cambridge: Polity Press; 2020.
58. Elena-Bucea A, Cruz-Jesus F, Oliveira T, Coelho PS. Assessing the role of age, education, gender and income on the digital divide: evidence for the European union. Inf Syst Front 2020;23(4):1007-1021. [doi: 10.1007/s10796-020-10012-9]
59. Reiners F, Sturm J, Bouw LJW, Wouters EJM. Sociodemographic factors influencing the use of eHealth in people with chronic diseases. Int J Environ Res Public Health 2019;16(4):645 [FREE Full text] [doi: 10.3390/ijerph16040645] [Medline: 30795623]
60. Stata statistical software: release 16. StataCorp. 2019. URL: https://www.stata-press.com/data/r16/ [accessed 2024-06-20]
61. Bevolking op 1 januari en gemiddeld; geslacht, leeftijd en regio. Statistics Netherlands. 2021. URL: https://www.cbs.nl/nl-nl/cijfers/detail/03759ned?dl=39E0B [accessed 2024-05-15]
62. Akoglu H. User's guide to correlation coefficients. Turk J Emerg Med 2018;18(3):91-93 [FREE Full text] [doi: 10.1016/j.tjem.2018.08.001] [Medline: 30191186]
63. Øversveen E, Rydland HT, Bambra C, Eikemo TA. Rethinking the relationship between socio-economic status and health: making the case for sociological theory in health inequality research. Scand J Public Health 2017;45(2):103-112. [doi: 10.1177/1403494816686711] [Medline: 28078944]
64. Eikemo TA, Øversveen E. Social inequalities in health: challenges, knowledge gaps, key debates and the need for new data. Scand J Public Health 2019;47(6):593-597. [doi: 10.1177/1403494819866416] [Medline: 31512565]
65. Stimuleren gebruik e-health. Ministry of Public Health, Wellbeing and Sports. URL: https://www.rijksoverheid.nl/onderwerpen/e-health/overheid-stimuleert-e-health [accessed 2023-09-08]

XSL•FO
RenderX

66. Hiltz SR, Turoff M. Education goes digital: the evolution of online learning and the revolution in higher education. Commun ACM 2005;48(10):59-64. [doi: 10.1145/1089107.1089139]
67. Grossman LV, Masterson Creber RM, Benda NC, Wright D, Vawdrey DK, Ancker JS. Interventions to increase patient portal use in vulnerable populations: a systematic review. J Am Med Inform Assoc 2019;26(8-9):855-870 [FREE Full text] [doi: 10.1093/jamia/ocz023] [Medline: 30958532]
68. Norman CD, Skinner HA. eHealth literacy: essential skills for consumer health in a networked world. J Med Internet Res 2006;8(2):e9 [FREE Full text] [doi: 10.2196/jmir.8.2.e9] [Medline: 16867972]
69. Bejaković P, Mrnjavac Z. The importance of digital literacy on the labour market. ER 2020;42(4):921-932. [doi: 10.1108/er-07-2019-0274]
70. Vasilescu MD, Serban AC, Dimian GC, Aceleanu MI, Picatoste X. Digital divide, skills and perceptions on digitalisation in the European union-towards a smart labour market. PLoS One 2020;15(4):e0232032 [FREE Full text] [doi: 10.1371/journal.pone.0232032] [Medline: 32324786]
71. Wagenaar KP, Hakim N, Broekhuizen BDL, Jaarsma T, Rutten FH, Hoes AW. Representativeness of participants in heart failure E-health trials: a report from the E-vita HF study. J Card Fail 2017;23(1):88-89 [FREE Full text] [doi: 10.1016/j.cardfail.2016.08.006] [Medline: 27565046]
72. Al-Dhahir I, Reijnders T, Faber JS, van den Berg-Emons RJ, Janssen VR, Kraaijenhagen RA, et al. The barriers and facilitators of eHealth-based lifestyle intervention programs for people with a low socioeconomic status: scoping review. J Med Internet Res 2022;24(8):e34229 [FREE Full text] [doi: 10.2196/34229] [Medline: 36001380]
73. de Boer WIJ, Dekker LH, Koning RH, Navis GJ, Mierau JO. How are lifestyle factors associated with socioeconomic differences in health care costs? Evidence from full population data in the Netherlands. Prev Med 2020;130:105929 [FREE Full text] [doi: 10.1016/j.ypmed.2019.105929] [Medline: 31778685]
74. Agachi E, Bijmolt THA, Mierau JO, van Ittersum K. Adoption of the website and mobile app of a preventive health program across neighborhoods with different socioeconomic conditions in the Netherlands: longitudinal study. JMIR Hum Factors 2022;9(1):e32112 [FREE Full text] [doi: 10.2196/32112] [Medline: 35107433]

Abbreviations

DHCCP: Dutch Health Care Consumer Panel
OR: odds ratio
SEP: socioeconomic position
SES-WOA: Social Economic Status-Wealth Education Employment
SES: socioeconomic status

Edited by A Mavragani; submitted 03.06.23; peer-reviewed by J Mierau, H Traino, PhD, MPH, M Earl; comments to author 29.08.23; revised version received 02.10.23; accepted 21.07.24; published 07.01.25.

Please cite as:
Standaar L, van Tuyl L, Suijkerbuijk A, Brabers A, Friele R
Differences in eHealth Access, Use, and Perceived Benefit Between Different Socioeconomic Groups in the Dutch Context: Secondary Cross-Sectional Study
JMIR Form Res 2025;9:e49585
URL: https://formative.jmir.org/2025/1/e49585
doi:10.2196/49585
PMID:

©Lucille Standaar, Lilian van Tuyl, Anita Suijkerbuijk, Anne Brabers, Roland Friele. Originally published in JMIR Formative Research (https://formative.jmir.org), 07.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Trifecta of Industry, Academic, and Health System Partnership to Improve Mental Health Care Through Smartphone-Based Remote Patient Monitoring: Development and Usability Study

C Neill Epperson[1], MD; Rachel Davis[1], MD; Allison Dempsey[1], PhD; Heinrich C Haller[1], BS; David J Kupfer[2], MD; Tiffany Love[1], PhD; Pamela M Villarreal[3], MS; Mark Matthews[4], PhD; Susan L Moore[5], PhD; Kimberly Muller[5], MSc, JD; Christopher D Schneck[1], MD; Jessica L Scott[1], MA, MPH; Richard D Zane[3,6], MD; Ellen Frank[2,4], PhD

[1]Department of Psychiatry, School of Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO, United States

[2]Department of Psychiatry, School of Medicine, University of Pittsburgh, Pittsburgh, PA, United States

[3]UCHealth CARE Innovations Center, Anschutz Medical Campus, University of Colorado Health System, Aurora, CO, United States

[4]Health Rhythms, New York, NY, United States

[5]CU Innovations, University of Colorado Anschutz Medical Campus, Aurora, CO, United States

[6]Department of Emergency Medicine, School of Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO, United States

Corresponding Author:
C Neill Epperson, MD
Department of Psychiatry
School of Medicine
University of Colorado Anschutz Medical Campus
1890 N Revere Ct
Suite 4003, Mail Stop F546
Aurora, CO, 80045
United States
Phone: 1 303 724 4940
Email: neill.epperson@cuanschutz.edu

Abstract

Background: Mental health treatment is hindered by the limited number of mental health care providers and the infrequency of care. Digital mental health technology can help supplement treatment by remotely monitoring patient symptoms and predicting mental health crises in between clinical visits. However, the feasibility of digital mental health technologies has not yet been sufficiently explored. Rhythms, from the company Health Rhythms, is a smartphone platform that uses passively acquired smartphone data with artificial intelligence and predictive analytics to alert patients and providers to an emerging mental health crisis.

Objective: The objective of this study was to test the feasibility and acceptability of Rhythms among patients attending an academic psychiatric outpatient clinic.

Methods: Our group embedded Rhythms into the electronic health record of a large health system. Patients with a diagnosis of major depressive disorder, bipolar disorder, or other mood disorder were contacted online and enrolled for a 6-week trial of Rhythms. Participants provided data by completing electronic surveys as well as by active and passive use of Rhythms. Emergent and urgent alerts were monitored and managed according to passively collected data and patient self-ratings. A purposively sampled group of participants also participated in qualitative interviews about their experience with Rhythms at the end of the study.

Results: Of the 104 participants, 89 (85.6%) completed 6 weeks of monitoring. The majority of the participants were women (72/104, 69.2%), White (84/104, 80.8%), and non-Hispanic (100/104, 96.2%) and had a diagnosis of major depressive disorder (71/104, 68.3%). Two emergent alerts and 19 urgent alerts were received and managed according to protocol over 16 weeks. More than two-thirds (63/87, 72%) of those participating continued to use Rhythms after study completion. Comments from participants indicated appreciation for greater self-awareness and provider connection, while providers reported that Rhythms provided a more nuanced understanding of patient experience between clinical visits.

XSL•FO
RenderX

Conclusions: Rhythms is a user-friendly, electronic health record–adaptable, smartphone-based tool that provides patients and providers with a greater understanding of patient mental health status. Integration of Rhythms into health systems has the potential to facilitate mental health care and improve the experience of both patients and providers.

(JMIR Form Res 2025;9:e57624) doi:10.2196/57624

KEYWORDS

digital health; mobile intervention; telepsychiatry; artificial intelligence; psychiatry; mental health; depression; mood; bipolar; monitor; diagnostic tool; diagnosis; electronic health record; EHR; alert; notification; prediction; mHealth; mobile health; smartphone; passive; self-reported; patient generated

Introduction

Mental disorders are predicted to be the leading cause of disease burden worldwide by 2030 [1]. However, there is a tremendous mismatch between the number of psychiatrists, psychologists, and other mental health care providers and individuals in need of treatment for mental illness, including substance use disorders. In the United States alone, 160 million people reside where there is a shortage of mental health professionals [2]. Roughly 56% of counties are without a practicing psychiatrist, while 33% of counties are without a licensed psychologist [3]. The workforce is expected to be short 38,821 psychiatrists by 2024 [4]. In the United States, the majority of individuals seeking treatment for depression, anxiety, or any mental illness receive care from their primary care physician (PCPs), and PCPs provide almost a third of the care for serious mental illness [5].

On average, a PCP addresses between 3 to 5 problems, including psychiatric problems, during a typical encounter [6,7]. Furthermore, the majority of PCPs cannot see patients as frequently as may be necessary to adequately monitor a patients' mental illness and response to treatment, nor do they have a mental health specialist to whom they can refer their more concerning patients [8]. In the context of the current mental health landscape, remote monitoring of patient mental well-being between visits is poised to play a major role in the redesigning of mental health care, similar to what is occurring with remote management of myriad other conditions [9,10].

Ideally, remotely acquired data would be integrated into the electronic health record (EHR) and readily available to the patient's provider for interpretation and intervention if needed. With remote patient monitoring, the documentation of history since the previous appointment is quicker, the challenges of patient recall with respect to their signs and symptoms are obviated, and there is a foundation for patients to participate to a greater degree in their own psychiatric care by tracking their symptoms alongside their provider.

Clinical, patient-generated, app-acquired health data that are embedded and presented into the EHR workflow provide the opportunity for interprofessional use. This is particularly important when patients have multiple comorbid medical conditions that impact brain health and vice versa. Furthermore, patient engagement with health apps, including those that are mental health related, declines dramatically over the first few weeks of use, in some cases dropping below 10% within 2 weeks [11,12]. An ideal app may rely less upon active user engagement for data collection and to a greater degree on passively acquired data that reduce patient burden, which then trigger the completion of symptom self-reports using nationally accepted standardized ratings when passively acquired patient-generated health data (PGHD) indicate a relapse is about to occur.

To address these deficits in the digital mental health space, an academic-industry collaborative team was formed to test the feasibility and acceptability of industry start-up Health Rhythms' smartphone solution called Rhythms for use across a large health system. Previous research has demonstrated that Rhythms' proprietary artificial intelligence (AI) algorithm applied to passively collected data indicative of behavioral vital signs (eg, sleep patterns, activity measures, and sociability) can predict a relapse of depression 7 days before the patient's self-report of relapse with 90% accuracy [13-15]. This collaborative team reports the results of the EHR integration, results of the feasibility testing of the urgent and emergent alert algorithms, effectiveness of the 24/7 virtual patient monitoring system, user acceptability in terms of both patient use of and satisfaction with Rhythms, and provider impressions of the impact of Rhythms on patient care.

Methods

Collaborative Team

Health Rhythms is a technology company that was founded by experienced researchers and clinicians (EF and DK) in the assessment, treatment, and research of mental illness, particularly bipolar disorder and depression, in partnership with international leaders (MM) in behavioral sensing technology focused on mental health. The University of Colorado (CU) Anschutz Medical Campus Department of Psychiatry is a large department (500+ faculty) with a strong commitment to the incorporation of digital health and technology in the care of patients with serious mental illness. CU Innovations brings together industry partners, entrepreneurs, and investors to help CU researchers create biomedical technology that improves care delivery, health outcomes, and patient quality of life worldwide. University of Colorado Health (UCHealth) is a not-for-profit health care system across Colorado, southern Wyoming, and western Nebraska with an academic medical center, the University of Colorado Hospital (UCH), that is affiliated with the CU School of Medicine. The CARE Innovation Center at UCHealth partners with CU to create a comprehensive platform of resources both to conduct clinical and biomedical research and to transform real-world clinical practice together with industry partners or faculty entrepreneurs.

XSL•FO
RenderX

App Development

The Rhythms app was developed using a participatory design process [16], an approach that seeks to deeply involve end users in the design process. A weekly meeting between the Health Rhythms team and the UCHealth clinical team was the primary means to engage in this process. It focused on understanding end-user needs, existing clinical workflows, and seeking feedback on early prototypes.

A key decision resulting from this process was to integrate the Health Rhythms software development kit (SDK) into UCHealth's MyChart mobile app, rather than providing a distinct separate app that participants would have to download. Health Rhythms' SDK is a modular piece of software that can be integrated easily into native iOS or Android apps. Although this decision required significant technical work, the team anticipated that it would significantly streamline the patient end-user experience, since they would only have to install and use 1 app, and would help maintain user privacy and security.

Once integrated into the UCHealth MyChart app and end users have provided explicit permissions during onboarding, the SDK collects data continuously in the background from the activity, location, and display smartphone sensors. These data can then be used to automatically assess behavior patterns derived from these data streams including how much time a person is spending at home, levels of physical activity, estimates of sleep timing, and duration and interaction patterns with the smartphone itself.

Safeguarding patient privacy was paramount. Users were uniquely identified using a nonidentifiable identifier that was passed to Health Rhythms. The Rhythms app did not collect any identifiable data that falls under the protections of the Health Insurance Portability and Accountability Act (HIPAA) such as birth dates, home addresses, or social security numbers, and the SDK did not collect any data from texts, calls, photos, internet browsing, or other content. All data, including GPS location, are encrypted using state of the art encryption (ie, a 256-bit key) on the device, in transmission and in the cloud. The sensor data for each patient were computed into daily nonidentifiable behavioral summaries that were then transmitted into the EHR using standard Fast Healthcare Interoperability Resources end points. No unprocessed sensor data, including GPS, were transferred into the EHR.

While there is always a risk of data breach in such systems, the following measures were implemented to mitigate this risk. The team applied the principle of data minimization, reducing the collection, storage, and use of personal data to what is strictly necessary to assess individual behavioral health. Patient end users had the agency to rescind permissions for sensor data collection at any time by a "Manage Settings" screen, and transparent, easy-to-understand messaging during onboarding communicated clearly what data were collected and how they would be used. Finally, Health Rhythms maintained General Data Protection Regulation and California Consumer Privacy Act certification and conducted regular security tests by professional third parties.

AI Methods

The Rhythms system collects passive sensing data, comprising location, pedometer, activity, and device data, from the onboard sensors in smartphones. Each day, these data are processed into 65 behavioral inferences for each person. These inferences encompass measures related to physical and social activity, stability or instability of the person's weekly routine, sleep, and smartphone use. For example, time at home is an inference derived from location data about how much time a person is spending at home.

These behavioral inferences are used in 2 ways. First, heuristics related to these inferences are used to trigger a request for a self-report from participants. A total of 15 static heuristics, empirically derived from previous datasets, dynamically prompted individuals to complete self-reported assessments. For example, if an individual had a diagnosis of major depression and spent a long time at home for over 7 days, then a self-report was triggered. Second, the behavioral inferences were fed into a logistic regression machine learning model that provided a continuous and passive estimate of a person's risk for a relapse in depression, defined as scoring greater than 10 on the Patient Health Questionnaire 9-Item (PHQ-9). This model aims to be predictive of how a given person would answer specific questions on the PHQ-9. Model outputs with a high probability for a high PHQ-9 score were used to flag individuals as a medium alert on the dashboard if there was no corroborating self-report.

Participant Criteria and Recruitment

Potential participants were residents of Colorado; aged between 18 and 89 years; were the primary user of a mobile phone that supported iOS12 or above; enrolled in MyHealthConnection (MHC), which is UCH's EHR patient portal; and were willing to enroll their device into Rhythms.

The UCH's EHR (EPIC) was used to search the caseload of participating clinicians in the CU Department of Psychiatry to identify patients who carry the diagnosis of depressive disorder including major depressive disorder (MDD), dysthymia, depressive disorder due to another medical condition, or bipolar disorder. Providers reviewed the list to confirm that individuals were current patients and would be clinically appropriate for contact about Rhythms.

Recruitment was conducted virtually. Patients were contacted initially through MHC with a general message describing the study. Due to poor initial response to the MHC outreach, a personalized email letter was sent from their provider. Potential participants were contacted up to 3 times by a study coordinator, and screening appointments were scheduled if they expressed interest in the study. After consenting and screening, the study coordinator reviewed how to download the app and interact with the platform and placed an order for Rhythms in EPIC. Participants were considered lost to follow-up if they did not respond to 3 contacts by the study team after enrollment.

Throughout the study, participants provided data by completing electronic surveys sent by the study team (baseline, end of study [6 weeks], and when triggered by algorithm), as well as by active and passive use of the Rhythms platform. A purposively

XSL•FO
RenderX

sampled [17,18] group of participants participated in qualitative interviews with a member of the study team at the end of the study. An overview of the study design and procedures is presented in Figure 1.

Figure 1. Study flow chart: (A) potential participants identified using the electronic health record in collaboration with providers; (B) potential participants contacted up to three times; (C) participants (who met criteria and signed consent) downloaded Rhythms and were instructed in its use; (D) participants completed baseline assessments; (E) participants continued their daily activities, with Rhythms passively collecting data for 6 weeks; (F) participants completed baseline assessment again at end of study; (G) participants completed a survey of likes and dislikes with respect to Rhythms; and (H) selected group underwent qualitative interviews about Rhythms use.

Data Collected by Rhythms

Passively Measured Outcomes

Rhythms acquired proxy information regarding sleep quality, activity level, and sociability. The AI algorithm uses passively collected data to detect deviations from the individual's baseline and triggers a push to complete the self-report ratings described below. During the course of this study, the participant could not view their own passively acquired data.

Self-Report Measures

Per study protocol, self-report measures were administered at baseline and end of study. In addition, self-report measures were pushed by Rhythms to the participant during the study period when the AI algorithm detected a deviation in passively collected data indicative of worsening of symptoms of depression, anxiety, or mania. Participant self-ratings included the following:

PHQ-9 Instrument

The PHQ-9 asks the individual to rate the severity of symptoms over the past 2 weeks on a 4-point scale ranging from "not at all" to "nearly every day." Endorsement of either depressed mood or anhedonia (PHQ-2) triggers further evaluation with the PHQ-9, which assesses for the presence of other criteria for MDD. In multiple studies, PHQ-9 scores >10 had a sensitivity of 88% and a specificity of 88% for MDD [19-21].

Columbia-Suicide Symptom Severity Rating Scale

The Columbia-Suicide Symptom Severity Rating Scale (C-SSRS) is used to assess severity and immediacy of suicide risk using four constructs: (1) severity of suicidal ideation, (2) intensity of ideation, (3) behavior, and (4) lethality. The instrument has been reported to have a 67% sensitivity and 76% specificity for identifying suicidal behaviors [22,23].

Generalized Anxiety Disorder 7-Item

The Generalized Anxiety Disorder 7-Item (GAD-7) is used to measure the severity of generalized anxiety disorder symptoms by asking individuals to rate the severity of symptoms on a 4-point scale with ranging from "not at all" to "nearly every day." Scores of 5, 10, and 15 represent cutoff points for mild, moderate, and severe anxiety, respectively [24].

Altman Self-Rating Mania Scale

The Altman Self-Rating Mania Scale (ASRM) is a 5-item, self-reported scale to assess positive mood, self-confidence, sleep patterns, speech patterns and amount, and motor activity, which are rated on a scale of "0" or "normal" to "4" or "overtly manic." Total scores range from 0 to 20, with scores equal to or higher than 6 indicating a higher likelihood of manic or hypomanic symptoms [25].

Participant-Reported Measures Collected by Study Team

Postintervention Feasibility and Satisfaction Survey

At study end, participants completed a survey about what they liked or disliked most about Rhythms, their interest in continued use, challenges or barriers to use, and any concerns that they may have had regarding alerting and monitoring through the use of Rhythms.

Qualitative Interviews

The qualitative in-depth interviews were conducted by Zoom (Zoom Video Communications) using a semistructured interview guide to ensure key topics were discussed while also allowing for emergence and exploration of topics important to the participants. The main interview questions for patients focused on their perceived benefits from and drawbacks of the Rhythms' system and suggestions for improving Rhythms' features and use.

Providers also completed a feedback survey at study end. Main questions for providers included their experience integrating the Rhythms system into their workflow, how the system was perceived to impact the quality of care, and the efficiency of providing personalized treatment for patients.

Determining Alert Structure

Based upon previous research with Rhythms [14,15,17,26] and consensus of subject matter experts in the treatment of MDD, anxiety, and bipolar disorder, Rhythms was programmed to send urgent or emergent alerts to UCHealth's Virtual Behavioral Health Center (VBHC) or the Department of Psychiatry's care coordinator based upon passively acquired data and participant self-ratings (Textbox 1).

XSL•FO
RenderX

Textbox 1. Urgent and emergent alerts.

Urgent alert

- Physician Health Questionnaire-9 item score is 20-27, when the previous score was lower.
- Columbia-Suicide Severity Rating Scale question 3=yes.
- Altman Self-Rating for Mania Scale score is ≥11.
- Altman Self-Rating for Mania Scale score is <11 and score for question 3=3 or 4.
- If bipolar disorder diagnosis and Rhythms detect a 40% decrease in sleep.
- If bipolar disorder diagnosis and Rhythms monitoring shows no sleep in 2 days.
- Generalized Anxiety Disorder-7 item score ranges from 15-21, when the previous score was not in this range.

Emergent alert

- Columbia-Suicide Severity Rating Scale questions 4, 5, or 6=yes.
- Altman Self-Rating for Mania Scale total score >11 and answer to question 3=4.

The following specific self-rating questions pertained to the alert structure: C-SSRS question 3: Active suicidal ideation with any methods (not plan) without intent to act; C-SSRS question 4: Active suicidal ideation with some intent to act, without specific plan; C-SSRS question 5: Active suicidal ideation with specific plan and intent; C-SSRS question 6: Have you ever done anything, started to do anything, or prepared to do anything to end your life?; and ASRM question 3: Regarding Sleep Patterns: score of 3="I frequently need less sleep than usual." and score of 4="I can go all day and night without any sleep and not feel tired."

Per protocol, a care coordinator from the VBHC or the Department of Psychiatry called the participant up to 3 times within 24 hours in response to an *urgent* alert. The VBHC monitor covered alerts over weekends, holidays, and after hours, while the care coordinator supporting the participant's primary mental health care provider covered during regular work hours. When the participant was reached by the VBHC, a brief safety assessment was conducted; participants were offered the chance to speak to an on-call clinician for the VBHC and referred to their current provider if appropriate. In the case of clinic care coordinators, the participant was assessed with respect to safety, possible medication changes, and need for earlier appointment with their provider.

In response to *emergent* alerts, Health Rhythms sent a page to the VBHC monitor. The pager beeped every 5 minutes until the alert was acknowledged by the monitor in the EHR. All emergent alerts were managed by the VBHC monitor with the expectation that a call to the patient would be initiated within 15 minutes of the alert. Upon reaching the patient, a safety assessment was conducted, and an appropriate disposition was managed. If after 3 attempts there was no contact with the patient, the VBHC monitor called emergency services to conduct a welfare check at the address listed for the participant.

Initially, participants received a message each time they completed self-report measures that they could receive a call from the VBHC based upon how they answered the questionnaires. This caused confusion for participants and the process was changed such that only participants who had triggered an emergent or urgent alert would get the message that they would receive a call.

Data Analysis

Demographics and baseline behavioral rating scores as well as number and nature of alerts were analyzed descriptively with means (SDs) and frequencies (percentages). Postintervention survey data were summarized descriptively for fixed-choice items. Open-ended survey items and interview transcripts were analyzed qualitatively. Human-adjudicated transcription of interview recordings was performed with Otter.ai, a proprietary AI-powered note taking engine, to generate initial transcripts, after which a member of the study team reviewed the transcripts while simultaneously listening to the recordings to ensure there were no errors in transcription.

A rapid qualitative analysis approach was used for content analysis of open-ended survey and interview transcript data. We first deductively explored and identified common themes that appeared in participants' responses. Two members of the research team separately reviewed the data and identified keywords and topics that are relevant to our research questions (eg, most liked and disliked features of the Rhythms' system, facilitators and barriers to using the Rhythms system, and suggestions to improve the system). During this process, each reviewer independently summarized how frequently common items were reported by study participants. The reviewers compared their results, discussed and resolved any discrepancies, reached consensus on the latent themes that had emerged in analysis, and generated a summary.

Ethical Considerations

This project was approved by the University of Colorado Multiple Institution Review Board (COMIRB; 21-4903), and participants gave electronic informed consent. Participants received US $25 in compensation for completing the screening, enrollment, and baseline surveys, and an additional US $50 for completing the postintervention surveys as described below. All data were deidentified for analysis and presented anonymously.

XSL•FO
RenderX

Results

Participants

Of 394 patients who were informed about the study through MHC or email, 178 (45.2%) did not respond, 26 (6.7%) were contacted but lost to follow-up before enrollment, 86 (21.8%) were screened but did not enroll, and 104 (26.4%) were enrolled within 8 weeks of study initiation. The most common reason (60/86, 69.8%) for not meeting study criteria was having a phone using the Android operating system.

The study sample comprised 104 individuals; the majority identified as women (n=72 69.2%), non-Hispanic (n=100 96.2%), and White (n=84 80.8%). The average age was 42.1 (SD 15.9) years. The majority of participants had MDD (n=71 68.3%), while 23 (22.2%) individuals had either bipolar I or II disorder and 10 individuals (9.6%) had "other mood disorder" (Table 1).

Between enrollment and study completion, 17 (16.3%) of the 104 enrolled participants were lost to follow-up, withdrew from the study, or did not complete the final survey. Of note, among the 87 (83.7%) completers out of 104, a total of 72% (63/87) continued using Rhythms despite no longer being compensated for doing so.

Table 1. Participant characteristics and baseline behavioral ratings.

Variable	Value (n=104)
Age (years), mean (SD)	42.1 (15.9)
Gender, n (%)	
Women	72 (69.2)
Men	29 (27.9)
Transgender or nonbinary	3 (2.9)
Ethnicity, n (%)	
Hispanic	3 (2.9)
Non-Hispanic	100 (96.2)
Not answered	1 (1)
Race, n (%)	
Asian	9 (8.6)
Black or African American	1 (1)
American Indian or Alaska Native	1 (1)
White	84 (80.8)
More than one race	4 (3.8)
Other or prefer not to answer	5 (4.8)
Diagnosis, n (%)	
Major depressive disorder	71 (68.3)
Bipolar 1 disorder	12 (11.5)
Bipolar II disorder	11 (10.6)
Other mood disorder	10 (9.6)
Ratings, mean (SD)	
Physician Health Questionnaire 9-Item	6.5 (5.3)
Generalized Anxiety Disorder 7-Item	5.9 (4.8)
Altman Self-Rating Mania Scale	2.2 (2.7)

Urgent and Emergent Alerts

Over the course of the entire 16-week study period, the monitoring of 104 individuals with Rhythms for 6 weeks each resulted in 2 emergent alerts and 19 urgent alerts. The largest number of participants being monitored at 1 time was 92 and the average was 64. The greatest number of alerts on any given day was 3 (all urgent, 1 day only) when there were 58 participants being monitored (Figure 2).

XSL•FO
RenderX

Figure 2. Urgent and emergent alerts across time and by number of active participants.

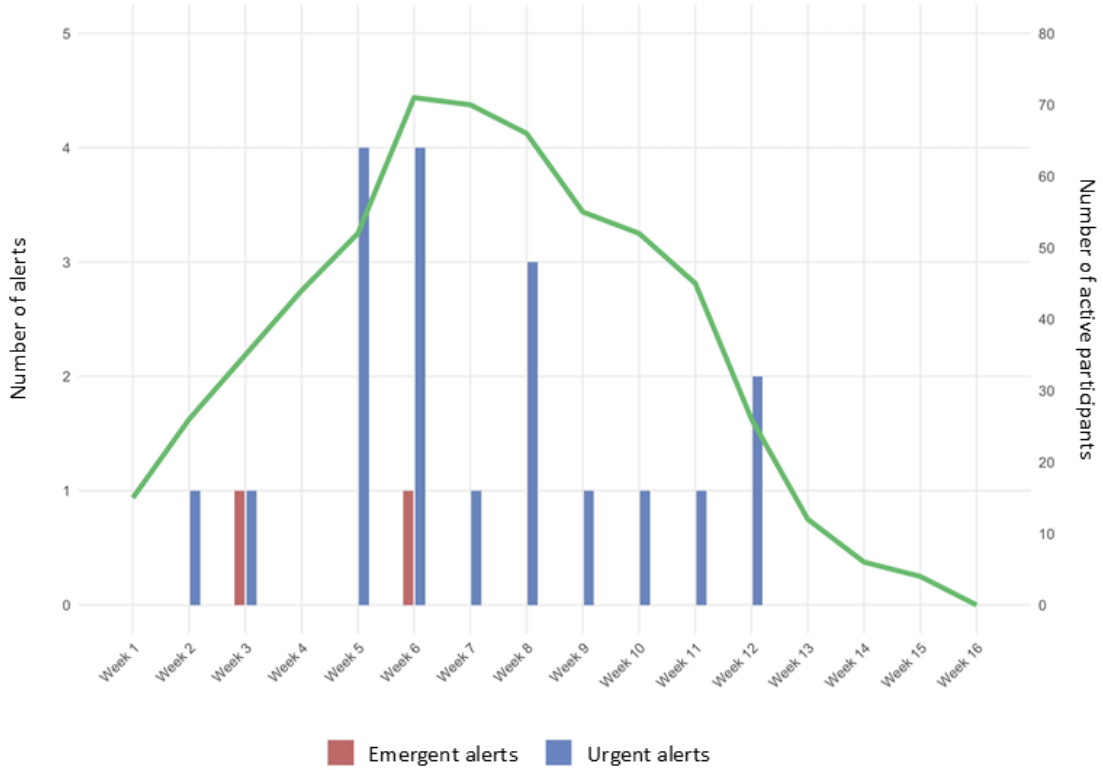

The first emergent alert was triggered by a participant answering ASRM question 3 about sleep with a "4," indicating "I can go all day and night without any sleep and not feel tired." The second emergent alert was triggered by the participant answering yes to C-SSRS question 6, indicating that they "had ever done anything, started to do anything or prepared to do anything to end their life." The primary (n=15) trigger for urgent alerts was answering "yes" to question 3 on the C-SSRS, indicating that the participant had active suicidal ideation without method or intent to act. The second most common (n=12) reason for urgent alerts was a GAD-7 score indicating severe anxiety. Three urgent alerts occurred because of the ASRM indicating mania or hypomania based upon total ASRM or response to question 3 regarding need for sleep.

All calls to participants were managed within the specified alert time. Eleven (58%) of the 19 urgent alerts came through during work hours and were managed by the Department of Psychiatry's care coordinator. The 2 emergent alerts were managed by the VBHC, neither of which required an emergency room referral or in-person well-being check.

Postintervention Survey: Patients and Providers

Of the 87 participants who completed the postintervention feasibility and satisfaction survey, the median score for ease of starting the system, ease of use, and surveys being delivered at convenient times were 9, 10, and 9, respectively, on a scale of 0 (strongly disagree) to 10 (strongly agree). Median score for participant satisfaction with the Rhythms system within their MHC app was 8.

Themes from the free text indicated that respondents most liked that Rhythms gave them a greater awareness of their own mental health status, the feeling of being taken care of by their provider, and its ease of use. The most common complaint about Rhythms (mentioned by n=18, 21%) was having to keep the app open. Only 3 (3%) reported concerns about being "tracked."

When providers were surveyed about their experience with Rhythms, they noted that the platform provided them with information they would not otherwise have had and information that was directly actionable. For example, 1 provider commented that "...the sleep section was the most useful for me to review. I am very surprised that one of my patients is only averaging about 5 hours of sleep per night, so that is definitely something we will talk about." Another commented that "The more frequent assessment was great. I don't think patients remember their condition for more than a few days. I was able to reach out to a couple of patients based on worsening in [their] data."

Qualitative Interviews: Patients

Qualitative interviews with 19 (22%) of the 87 completers revealed that all but 2 were well established in the CU and UCH system for their mental health care, having received care there for at least 3 years. Most agreed to participate in the study because they felt positively about research (n=10, 11%), thought it would be easy (n=4, 5%), or thought it would be helpful to either themselves (n=3, 3%) or others (n=3, 3%). Setup and initial use of Rhythms was reported by 13 (68%) of those interviewed as being "simple, intuitive, and unobtrusive," although 5 (26%) reported some technical challenges, including difficulty with changing settings, granting permissions, or finding specific features. With respect to completing the self-report measures and their frequency, 63% (n=12) of respondents reported completing them at the time of the prompt. In general, respondents were okay with the frequency of measures (n=3, 3%) or agreed that weekly (n=3, 3%) or every 2 weeks (n=4, 5%) would be acceptable, although a few (n=2,

XSL•FO
RenderX

2%) preferred monthly or felt surveys were sent too often (n=1, 1%).

Overall, interview participants reported positive perceptions of Rhythms, finding it easy to use (n=8, 9%) and enjoying its use (n=5, 6%). With respect to what was liked most about Rhythms, there was consensus that the platform promoted greater provider connection (n=4, 5%) and provider awareness of patients' health status (n=6, 7%), contributing to a feeling of being "looked out for" (n=4, 5%). Participants also reported improved self-awareness of one's own mental health status (n=5, 6%). What was liked least about Rhythms was having to keep the app open (n=7, 8%), with related concern for battery life (n=4, 5%); the wording of a response message to completed surveys notifying patients that they might be contacted based on their responses (n=6, 7%); concerns about what would happen if answers to questions were negative (n=3, 3%); not receiving immediate feedback from Rhythms; and accuracy of the data if the person did not have their phone with them or were traveling outside of Colorado.

Roughly half of those interviewed endorsed having had a helpful discussion about their Rhythms data with their provider, while others indicated that their provider did not mention their data during their appointment, or they did not have an appointment during the 6 weeks of tracking. The most common recommendation for system improvement was the ability to access and track their own data in a visual format, an improvement that is now available to users. Illustrative quotations about positive aspects of Rhythms were as follows:

> *…I had one two-week period where it was like, everything's catastrophic…which solicited the call from [my provider], through the patient portal. "So I saw your responses. What's, the scoop?"…I liked the fact that somebody was looking out for me. It just—I'm, I'm an introvert. I don't want to bother people. You know, am I really so bad that I need to interrupt somebody today? I don't want to be a bother. But I was able to experience the effectiveness of the system, because [my provider] reached out to me. So, from a patient point of view, I can tell you, that particular item worked.*
>
> *And then there was one survey when I was starting to feel a little more anxious than I normally do. I got asked to complete a survey. And I did. And I know when I completed that survey, [my provider] reached out to me and said, "Hey, I noticed that you're feeling a little more anxious than you seem to be normally." So, I was. I'm not sure what prompted that. But I thought it was interesting that at a time I kind of had a need to communicate with [my provider], the system seemed to have recognized that.*
>
> *Overall, I think that it is a wonderful tool to have in one's toolbox to help support mental wellness.*

Illustrative quotations regarding concern about Rhythms were as follows:

> *Also, I remember first time and a few times after that, too, I got - because I know sometimes when you answer things in a way that is concerning, they say someone will call after every survey, I got that message. Sometimes it was someone may call and sometimes someone will call…I remember because one time I left my phone at home and I went out and I was like, oh my god, someone's gonna call and they're gonna message. Yeah, that message was a little scary sometimes…*
>
> *I didn't like hearing that I may or may not receive a call from a provider, I would have liked it if it was more definitive.*
>
> *Like, I don't sleep with my mobile device…and I don't always exercise with my phone.*

Reliability

The Rhythms system performed well during the pilot study with no outages and no major bugs detected.

Discussion

Principal Findings

Here, we describe the use, including acceptability, of the smartphone platform Rhythms by patients and providers in an academic department of psychiatry. Overall, the system integrated well with the health system's EHR and clinical workflow and was able to be deployed with very few technical issues to a group of 104 participants. The algorithm for urgent and emergent alerts allowed the team to identify concerning changes in patient mental health status and may be considered by other clinical groups wishing to use Rhythms. The lack of adverse outcomes during this period suggests that the algorithm was set conservatively enough to capture those who are in need of outreach, but not so sensitive as to lead to inappropriate alerts that could unnecessarily burden the patient or overwhelm the monitoring team.

An impressive percentage (63/87, 72%) of study participants remained on the platform and continued to use it without any compensation beyond the end of the 6-week study, suggesting they found the app helpful. This is in distinction to the findings from Kopka and colleagues [11] showing that use of 4 out of 10 popular mental health apps dropped below 20% within the 2 weeks of initiation. None of the 10 apps chosen by participants in Kopka and colleagues [11] collected passively acquired data or were integrated into the individual's health record. In our study, poststudy interviews revealed that passively collected data lead to a greater connection with one's provider and was a major strength of Rhythms. Participants also provided important feedback regarding the functionality (ie, ability to track ratings visually over time and receiving immediate feedback) of Rhythms that has now been incorporated into the platform. Importantly, privacy was not a major concern expressed by participants in this study, unlike those in some [26-29] but not all previous studies [12].

The concern about having to leave the app open is real, but this did not deter the majority of individuals from remaining on the app. While providers reported that Rhythms is a useful tool in their clinical practice, some participants reported that their provider did not discuss their Rhythms data during their

XSL•FO
RenderX

appointment. This may reflect the timing of Rhythms use and the availability of collected data in relation to participants' appointments, or it may be indicative of differing levels of provider adoption of the technology and how they incorporated it into their own practices. The use of objective measurements for patient care, which is quite common in most fields of medicine, is only recently making its way into the practice of psychiatry. We are encouraged that when the Rhythms data were brought into the session, the patients reported a benefit. While scalable technology holds great promise for improving psychiatric treatment, the acceptance and use of technology in psychiatry would require a culture change if measurement assisted care is to become a reality. Since the COVID-19 pandemic, many psychiatrists and other mental health specialists are now comfortable using videoconferencing for patient visits. However, incorporating data collected from apps and wearables is an unfamiliar step in the therapeutic interaction.

Likewise, many health systems have not yet embraced linkage of apps used to collect PGHD for mental health care purposes with their EHR. When there is collection of these data through medical records standards initiative such as Fast Healthcare Interoperability Resources and other platforms [30], information is not easily collated across multiple patient-derived sources. Recent advances have been made in this area through platforms, such as mindLAMP from the Digital Psychiatry Program at Beth Israel Deaconess Medical Center of Harvard Medical School, which has the power to incorporate PGHD collected by multiple apps and wearables and offers tools to help configure data such that they are more readily usable for research or clinical care [31]. Within this context, the benefit of Rhythms is that it can be seamlessly integrated into the patient's EHR such that the provider can view passively acquired data as well as ratings with 1 click in the patient's chart. The alert algorithm is strong and can help guide others in their use of Rhythms to detect emerging patient crises.

Those psychiatrists who incorporated Rhythms data into their clinical care reported that it improved their understanding of their patients' clinical status and that they used these data in their treatment planning. Likewise, patients appreciated the added connection to their provider, and there were no reports of increased provider burden across the study.

Several limitations should be considered. In total, 45.2% (178/394) of potential participants did not respond to either a general inquiry sent through MHC or an email letter sent by their provider. It is unclear why they did not respond to messages through the EHR, although previous research suggests that patient preferences for communication method are influenced by the type of information to be received [32]. Willingness to participate in a study that requires downloading an app onto one's personal phone may require direct interaction with one's provider and a clear discussion of the potential benefits for their treatment. The sample predominantly identified as women and White, limiting generalizability of these findings to other populations. The relatively short duration of this initial trial (6 weeks for each participant) and the relatively small number of patient and provider participants was also a limitation. Though the clinical value of Rhythms was quickly apparent, a longer period of monitoring is needed to determine the economic impact of remote patient monitoring of mental health status on health care use and costs. All study participants were currently in active mental health treatment and, on the whole, relatively stable. How Rhythms would perform in a more acutely ill population or those with psychosis within the greater health system is unknown but should be considered in future studies, as psychiatric disorders can negatively impact one's motivation and ability to engage with an app and their overall health care [33,34].

In summary, Rhythms is a user-friendly, digital platform that can be embedded into a health system's EHR and used to identify patients who are experiencing a worsening of their mental health between visits with their mental health care provider. The AI algorithms are based upon passively acquired data, limiting the need for patient surveys except when deviations in data suggest a deterioration in the patient's clinical status. The culture surrounding the use of technology as part of measurement assisted mental health care will become more welcoming as studies such as ours show the ease and clinical benefits of using remote patient monitoring tools.

Acknowledgments

This work was funded in part by a research grant from Health Rhythms. The authors thank Dr. Chloe E Page for assistance with submission and copyediting.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

CNE contributed to the study conception and design, data collection, analysis and interpretation of results, and manuscript preparation. RD contributed to study design, data collection, and manuscript preparation. AD contributed to data collection, interpretation of results, and manuscript preparation. HCH contributed to data collection, data management, and manuscript preparation. DJK participated in study conception and manuscript preparation. TL contributed to study design, data collection, and manuscript preparation. PMV participated in study design, data collection, and manuscript preparation. MM participated in data collection and manuscript preparation. SLM participated in study conception and design, data collection, analysis and interpretation of results, and manuscript preparation. KM participated in study conception and preparation. CDS participated in

XSL•FO
RenderX

study design, data collection, and manuscript preparation. JLS participated in study design, data collection, analysis and interpretation of results, and manuscript preparation. RDZ contributed to manuscript preparation. EF contributed to study conception and design and manuscript preparation.

Conflicts of Interest

CNE is a consultant to Skyland Trail, BabyScripts, and EmbarkNeuro. CNE was the principal investigator on this study, which was supported in part by a grant from Health Rhythms. EF is an employee of and holds equity in Health Rhythms. TL, MM, CNE, and DJK hold equity in Health Rhythms. AD received research grant support from Health Rhythms. KM is an investor in Health Rhythms.

References

1. Arias D, Saxena S, Verguet S. Quantifying the global burden of mental disorders and their economic value. EClinicalMedicine 2022;54:101675 [FREE Full text] [doi: 10.1016/j.eclinm.2022.101675] [Medline: 36193171]
2. Counts N. Understanding the US behavioral health workforce shortage. The Commonwealth Fund. 2023. URL: https://www.commonwealthfund.org/publications/explainer/2023/may/understanding-us-behavioral-health-workforce-shortage [accessed 2024-11-02]
3. Beck AJ, Page C, Buche J, Rittman D, Gaiser M. Estimating the distribution of the U.S. psychiatric subspecialist workforce. Health Workforce Technical Assistance Center. 2018. URL: https://www.healthworkforceta.org/research-alerts/estimating-the-distribution-of-the-us-psychiatrist-subspecialist-workforce/ [accessed 2024-11-08]
4. Satiani A, Niedermier J, Satiani B, Svendsen DP. Projected workforce of psychiatrists in the United States: a population analysis. Psychiatr Serv 2018;69(6):710-713. [doi: 10.1176/appi.ps.201700344] [Medline: 29540118]
5. Jetty A, Petterson S, Westfall JM, Jabbarpour Y. Assessing primary care contributions to behavioral health: a cross-sectional study using medical expenditure panel survey. J Prim Care Community Health 2021;12:21501327211023871 [FREE Full text] [doi: 10.1177/21501327211023871] [Medline: 34109860]
6. Kyanko KA, A Curry L, E Keene D, Sutherland R, Naik K, Busch SH. Does primary care fill the gap in access to specialty mental health care? a mixed methods study. J Gen Intern Med 2022;37(7):1641-1647 [FREE Full text] [doi: 10.1007/s11606-021-07260-z] [Medline: 34993864]
7. Beasley JW, Hankey TH, Erickson R, Stange KC, Mundt M, Elliott M, et al. How many problems do family physicians manage at each encounter? a WReN study. Ann Fam Med 2004;2(5):405-410 [FREE Full text] [doi: 10.1370/afm.94] [Medline: 15506571]
8. Richman EL, Lombardi BM, Zerden LD. Mapping colocation: using national provider identified data to assess primary care and behavioral health colocation. Fam Syst Health 2020;38(1):16-23. [doi: 10.1037/fsh0000465] [Medline: 32202831]
9. Drincic A, Prahalad P, Greenwood D, Klonoff DC. Evidence-based mobile medical applications in diabetes. Endocrinol Metab Clin North Am 2016;45(4):943-965 [FREE Full text] [doi: 10.1016/j.ecl.2016.06.001] [Medline: 27823614]
10. Su D, Michaud TL, Estabrooks P, Schwab RJ, Eiland LA, Hansen G, et al. Diabetes management through remote patient monitoring: the importance of patient activation and engagement with the technology. Telemed J E Health 2019;25(10):952-959. [doi: 10.1089/tmj.2018.0205] [Medline: 30372366]
11. Kopka M, Camacho E, Kwon S, Torous J. Exploring how informed mental health app selection may impact user engagement and satisfaction. PLOS Digit Health 2023;2(3):e0000219 [FREE Full text] [doi: 10.1371/journal.pdig.0000219] [Medline: 36989237]
12. Baumel A, Muench F, Edan S, Kane JM. Objective user engagement with mental health apps: systematic search and panel-based usage analysis. J Med Internet Res 2019;21(9):e14567 [FREE Full text] [doi: 10.2196/14567] [Medline: 31573916]
13. Frank E, Wallace ML, Matthews MJ, Kendrick J, Leach J, Moore T, et al. Personalized digital intervention for depression based on social rhythm principles adds significantly to outpatient treatment. Front Digit Health 2022;4:870522 [FREE Full text] [doi: 10.3389/fdgth.2022.870522] [Medline: 36120713]
14. Malkoff-Schwartz S, Frank E, Anderson BP, Hlastala SA, Luther JF, Sherrill JT, et al. Social rhythm disruption and stressful life events in the onset of bipolar and unipolar episodes. Psychol Med 2000;30(5):1005-1016. [doi: 10.1017/s0033291799002706] [Medline: 12027038]
15. Abdullah S, Murnane ELM, Matthews M, Choudhury T. Circadian computing: sensing, modeling, and maintaining biological rhythms. In: Rehg J, Murphy S, Kumar S, editors. Mobile Health: Sensors, Analytic Methods, and Applications. Cham, Switzerland: Springer; 2017:35-58.
16. Schuler D, Namioka A. Participatory Design: Principles and Practices. Boca Raton, FL: CRC press; 1993.
17. Patton MQ. Qualitative Research and Evaluation Methods. 3rd ed. Thousand Oaks, CA: Sage Publications; 2002.
18. Creswell JW, Clark VLP. Designing and Conducting Mixed Method Research. 2nd ed. Thousand Oaks, CA: Sage Publications; 2011.
19. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]

XSL•FO
RenderX

20. Kroenke K, Spitzer RL. The PHQ-9: a new depression diagnostic and severity measure. Psychiatric Annals 2002;32(9):509-515. [doi: 10.3928/0048-5713-20020901-06]
21. Gilbody S, Richards D, Brealey S, Hewitt C. Screening for depression in medical settings with the Patient Health Questionnaire (PHQ): a diagnostic meta-analysis. J Gen Intern Med 2007;22(11):1596-1602 [FREE Full text] [doi: 10.1007/s11606-007-0333-y] [Medline: 17874169]
22. Posner K, Brown GK, Stanley B, Brent DA, Yershova KV, Oquendo MA, et al. The Columbia-Suicide Severity Rating Scale: initial validity and internal consistency findings from three multisite studies with adolescents and adults. Am J Psychiatry 2011;168(12):1266-1277 [FREE Full text] [doi: 10.1176/appi.ajp.2011.10111704] [Medline: 22193671]
23. Mundt JC, Greist JH, Jefferson JW, Federico M, Mann JJ, Posner K. Prediction of suicidal behavior in clinical research by lifetime suicidal ideation and behavior ascertained by the electronic Columbia-Suicide Severity Rating Scale. J Clin Psychiatry 2013;74(9):887-893. [doi: 10.4088/JCP.13m08398] [Medline: 24107762]
24. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
25. Altman EG, Hedeker D, Peterson JL, Davis JM. The Altman Self-Rating Mania Scale. Biol Psychiatry 1997;42(10):948-955. [doi: 10.1016/S0006-3223(96)00548-3] [Medline: 9359982]
26. Matthews M, Abdullah S, Gay G, Choudhury T. Tracking mental well-being: balancing rich sensing and patient needs. Computer 2014;47(4):36-43. [doi: 10.1109/mc.2014.107]
27. Melcher J, Camacho E, Lagan S, Torous J. College student engagement with mental health apps: analysis of barriers to sustained use. J Am Coll Health 2022;70(6):1819-1825. [doi: 10.1080/07448481.2020.1825225] [Medline: 33048626]
28. Di Matteo D, Fine A, Fotinos K, Rose J, Katzman M. Patient willingness to consent to mobile phone data collection for mental health apps: structured questionnaire. JMIR Ment Health 2018;5(3):e56 [FREE Full text] [doi: 10.2196/mental.9539] [Medline: 30158102]
29. Zhou L, Bao J, Watzlaf V, Parmanto B. Barriers to and facilitators of the use of mobile health apps from a security perspective: mixed-methods study. JMIR Mhealth Uhealth 2019;7(4):e11223 [FREE Full text] [doi: 10.2196/11223] [Medline: 30990458]
30. Hassol A, Walker JM, Kidder D, Rokita K, Young D, Pierdon S, et al. Patient experiences and attitudes about access to a patient electronic health care record and linked web messaging. J Am Med Inform Assoc 2004;11(6):505-513 [FREE Full text] [doi: 10.1197/jamia.M1593] [Medline: 15299001]
31. Mandel JC, Kreda DA, Mandl KD, Kohane IS, Ramoni RB. SMART on FHIR: a standards-based, interoperable apps platform for electronic health records. J Am Med Inform Assoc 2016;23(5):899-908 [FREE Full text] [doi: 10.1093/jamia/ocv189] [Medline: 26911829]
32. Vaidyam A, Halamka J, Torous J. Enabling research and clinical use of patient-generated health data (the mindLAMP Platform): digital phenotyping study. JMIR Mhealth Uhealth 2022;10(1):e30557 [FREE Full text] [doi: 10.2196/30557] [Medline: 34994710]
33. Torous J, Nicholas J, Larsen ME, Firth J, Christensen H. Clinical review of user engagement with mental health smartphone apps: evidence, theory and improvements. Evid Based Ment Health 2018;21(3):116-119 [FREE Full text] [doi: 10.1136/eb-2018-102891] [Medline: 29871870]
34. Stiles-Shields C, Montague E, Lattie EG, Kwasny MJ, Mohr DC. What might get in the way: barriers to the use of apps for depression. Digit Health 2017;3:2055207617713827 [FREE Full text] [doi: 10.1177/2055207617713827] [Medline: 29942605]

Abbreviations

AI: artificial intelligence
ASRM: Altman Self-Rating Mania Scale
C-SSRS: Columbia-Suicide Symptom Severity Rating Scale
COMIRB: University of Colorado Multiple Institution Review Board
CU: University of Colorado
EHR: electronic health record
GAD-7: Generalized Anxiety Disorder 7-Item
HIPAA: Health Insurance Portability and Accountability Act
MDD: major depressive disorder
MHC: MyHealthConnection
PCP: primary care physician
PGDH: patient-generated health data
PHQ-9: Patient Health Questionnaire 9-Item
SDK: software development kit
UCH: University of Colorado Hospital
UCHealth: University of Colorado Health
VBHC: Virtual Behavioral Health Center

XSL•FO
RenderX

Edited by A Mavragani; submitted 21.02.24; peer-reviewed by D King, PH Liao; comments to author 22.06.24; revised version received 23.07.24; accepted 06.09.24; published 07.01.25.

Please cite as:
Epperson CN, Davis R, Dempsey A, Haller HC, Kupfer DJ, Love T, Villarreal PM, Matthews M, Moore SL, Muller K, Schneck CD, Scott JL, Zane RD, Frank E
The Trifecta of Industry, Academic, and Health System Partnership to Improve Mental Health Care Through Smartphone-Based Remote Patient Monitoring: Development and Usability Study
JMIR Form Res 2025;9:e57624
URL: https://formative.jmir.org/2025/1/e57624
doi:10.2196/57624
PMID:

©C Neill Epperson, Rachel Davis, Allison Dempsey, Heinrich C Haller, David J Kupfer, Tiffany Love, Pamela M Villarreal, Mark Matthews, Susan L Moore, Kimberly Muller, Christopher D Schneck, Jessica L Scott, Richard D Zane, Ellen Frank. Originally published in JMIR Formative Research (https://formative.jmir.org), 07.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Co-Design of an Escape Room for e-Mental Health Training of Mental Health Care Professionals: Research Through Design Study

Joyce J P A Bierbooms[1], PhD; Wouter R J W Sluis-Thiescheffer[2], PhD; Milou Anne Feijt[3], MSc; Inge M B Bongers[1], PhD

[1]Tilburg University, Tilburg, Netherlands
[2]HAN University of Applied Sciences, Arnhem and Nijmegen, Netherlands
[3]Eindhoven University of Technology, Eindhoven, Netherlands

Corresponding Author:
Joyce J P A Bierbooms, PhD
Tilburg University
PO box 90153
Tilburg, 5000 LE
Netherlands
Phone: 31 13 466 31 39
Email: j.j.p.a.bierbooms@tilburguniversity.edu

Abstract

Background: Many efforts to increase the uptake of e-mental health (eMH) have failed due to a lack of knowledge and skills, particularly among professionals. To train health care professionals in technology, serious gaming concepts such as educational escape rooms are increasingly used, which could also possibly be used in mental health care. However, such serious-game concepts are scarcely available for eMH training for mental health care professionals.

Objective: This study aims to co-design an escape room for training mental health care professionals' eMH skills and test the escape room's usability by exploring their experiences with this concept as a training method.

Methods: This project used a research through design approach with 3 design stages. In the first stage, the purpose, expectations, and storylines for the escape room were formulated in 2 co-design sessions with mental health care professionals, game designers, innovation staff, and researchers. In the second stage, the results were translated into the first escape room, which was tested in 3 sessions, including one web version of the escape room. In the third stage, the escape room was tested with mental health care professionals outside the co-design team. First, 2 test sessions took place, followed by 3 field study sessions. In the field study sessions, a questionnaire was used in combination with focus groups to assess the usability of the escape room for eMH training in practice.

Results: An escape room prototype was iteratively developed and tested by the co-design team, which delivered multiple suggestions for adaptations that were assimilated in each next version of the prototype. The field study showed that the escape room creates a positive mindset toward eMH. The suitability of the escape room to explore the possibilities of eMH was rated 4.7 out of 5 by the professionals who participated in the field study. In addition, it was found to be fun and educational at the same time, scoring 4.7 (SD 0.68) on a 5-point scale. Attention should be paid to the game's complexity, credibility, and flexibility. This is important for the usefulness of the escape room in clinical practice, which was rated an average of 3.8 (SD 0.77) on a 5-point scale. Finally, implementation challenges should be addressed, including organizational policy and stimulation of eMH training.

Conclusions: We can conclude that the perceived usability of an escape room for training mental health care professionals in eMH skills is promising. However, it requires additional effort to transfer the learnings into mental health care professionals' clinical practice. A straightforward implementation plan and testing the effectiveness of an escape room on skill enhancement in mental health care professionals are essential next steps to reach sustainable goals.

(JMIR Form Res 2025;9:e58650) doi:10.2196/58650

XSL•FO
RenderX

KEYWORDS

serious gaming; mental health care professionals; e-mental health; skill enhancement; training

Introduction

Background

Despite a growing body of evidence for the effectiveness and benefits of e-mental health (eMH), many studies report a relatively low use of eMH in practice [1-3]. eMH means mental health services delivered or supported by digital technologies, such as video calling, self-management and self-tracking apps, web-based treatment modules, and virtual reality (VR) [4,5]. An important factor hampering the use of eMH is its relatively slow adoption by mental health care professionals [5-7]. To a large extent, this slow adoption results from a lack of knowledge and skills regarding eMH among mental health care professionals [7-9].

This lack of skills entails general digital skills but, more importantly, the ability to use an adequate communication approach compensating for a potential lack of nonverbal cues and contextual information. In addition, skills are needed to guide a professional in handling boundaries (eg, expectations regarding reaction time and attainability), choosing appropriate communication channels for each situation, and up-to-date knowledge of the availability of technological possibilities to enrich treatment in mental health care [10-12]. The COVID-19 pandemic caused a significant increase in video calling and, subsequently, the skills to apply such technologies for therapeutic purposes. However, mental health care professionals are still hesitant to pursue broader incorporation of eMH (ie, using different eMH tools in a context where remote treatment is a choice rather than an imposition) into their therapeutic practices because of the skills required [13]. Thus, to increase the adoption of a broad range of eMH tools, there is a persistent need to enhance mental health care professionals' skills in using eMH.

Many training options have recently been developed to capacitate mental health care professionals in using eMH [14,15]. The introduction and development of web-based treatment platforms have brought different programs to acquire the technical skills to use such a platform. Examples are applying web modules, written client interactions, and video sessions instead of real-life therapeutic interactions. Besides the classical training approach, there are several eLearning options to support mental health care professionals in digital skill development [16]. However, research on the effectiveness of different learning and training strategies in health care has shown that traditional forms, such as classical instructions and eLearning, generate insufficient learning outcomes [17,18]. This can be explained by a lack of experiential learning in conventional learning methods; there is a need for a "learning by doing" approach. The learning by doing approach is more engaging, more realistic, and contributes to a larger self-efficacy [17,18]. Thus, there is a need to pursue new, more innovative, and experiential solutions to train mental health care professionals in using eMH.

To gain a feeling of self-efficacy regarding the application of eMH, a potential strategy is to offer mental health care professionals training possibilities based on the concept of serious gaming [19]. Serious games are physical or digital games applied to learn or acquire skills and usually offer a combination of educational content and engagement [19,20]. Escape rooms have emerged as a popular instructional approach within serious gaming, attracting growing interest from researchers and educators [21]. An escape room is a game in which participants are confronted with a storyline, usually locked into a room, and need to solve puzzles, search for information, and follow leads to free themselves from the room within a limited amount of time. Escape room games are increasingly used for learning [22-24].

The benefit of an escape room may be that it offers the option to train skills by gaining hands-on experiences (experiential learning). Escape rooms create realistic learning environments that foster meaningful participation and provide a supportive atmosphere for experimentation [25]. In contrast, other methods (eg, classical instructions or eLearning) mainly focus on providing instructions on using an artifact or service [18,26]. In addition, an escape room is focused on the learning process while simultaneously being educational and fun, which makes it much more engaging to participate in training [27,28].

The underlying learning mechanisms of an educational escape room (EER) are often described in terms of the self-determination theory (SDT) [29,30] and the flow theory [30,31]. The SDT refers to fundamental psychological needs such as autonomy, competence, and relatedness to others [29,30]. EERs provide a safe environment that can be tuned to the player's competencies, and an appropriate storyline helps players relate the competencies demanded in the EER puzzles to competencies required in real-world situations. Together, they increase a player's sense of autonomy, and they can create a positive experience in self-determination. The flow theory explains the occurrence of someone's complete immersion into a specific task based on positive experiences [30,31]. This "flow" can be achieved in serious gaming if the challenges meet the player's knowledge and abilities. By attempting to address these fundamental needs in a playful environment, users can have a positive learning experience [32].

Theoretically, these mechanisms provide a solid argument for the use of EERs to train mental health care professionals' skills in successfully integrating eMH into their therapeutic practice [22,23]. However, as Vorderobermeier et al [30] indicate, the success of EERs is hardly ever tracked back to these mechanisms. To contribute to the emerging field of research into EERs, we include an evaluation of the participants' learning experience in relation to these learning mechanisms.

EERs in eMH

In eMH, the concept of EERs is recognized as a safe and interactive environment to explore eMH in different possible therapeutic situations [30,32] while providing a safe and social environment for training in mental health care. EERs provide

XSL•FO
RenderX

an opportunity to experiment with otherwise complex or expensive therapeutic situations [33,34]. As EERs are an emerging phenomenon, there are also some recent examples of research into EERs in the context of eMH.

In the context of raising awareness of severe mental illnesses (SMIs), Rodriguez-Ferrer et al [35] studied the effectiveness of an EER in educating first-year nursing students (a nonspecialized audience) in experiencing the stigmas associated with SMI. The EER was developed with professionals in SMI and the effect of raising awareness was tested using a randomized controlled trial experiment. Apart from the primary objective (raising awareness of SMI), the EER objectives were to determine whether an entirely web-based version was effective and whether participants were sufficiently immersed in the experience. The EER in this study is different in the sense that it is aimed at experiential learning, involving both digital and physical interactions. Moreover, the EER developed in this study is aimed at a professional audience, which means that the challenges are tuned toward expert knowledge and an in-depth skillset rooted in ample experience in treating mental illnesses [35].

Aragon [36] designed and studied the use of an EER for mental health nursing students attending an accelerated associate degree program. The topic of learning was communication and collaboration strategies in mental health. The design of the EER took a pragmatic, digital form to simplify the strategy for the development, distribution, and deployment of the EER in the educational program. The authors studied the potential learning effect of the digital EER but found no significant difference compared to students who were not exposed to the EER. The EER in this paper is different, similar to the EER studied by Rodriguez-Ferrer et al [35] mentioned earlier, as it targets a professional audience and aims to provide a more physical, collaborative learning experience. Moreover, the EER by Aragon [36] was solely designed by one researcher, whereas this paper proposes a co-design process to include the views, needs, and goals of the target audience in the design of the EER. Finally, the work by Aragon [36] briefly discusses the underlying mechanisms of the EER, namely (1) the potentially positive effects of collaboratively solving nursing challenges and (2) the strong relatedness of the puzzles with the curriculum thus providing the students with alternative means to consolidate what was learned during the program [36]. The EER in this paper makes a clear choice to define the underlying mechanisms in the SDT theory and the flow theory.

There are more EERs for eMH, such as the one proposed in the study by Petkari and Calvo [37], and the number of EERs is likely to increase over time as its potential is currently widely researched. Similar to the studies discussed earlier, Petkari and Calvo [37] focused on EERs as alternative or additional elements for existing curricula, targeted at soon-to-be professionals. In those contexts, the design of the EER is mostly in the hands of educators or researchers, while the professionals are only involved in evaluations [37].

In a previous paper, we described the careful considerations we took to aim for a successful design and subsequent deployment of an EER that is tailored toward experiential learning about (innovative) digital elements suitable for blended therapies [38]. This paper describes the execution of a preliminary version of the EER.

This study aims to (1) further pursue the ambitions of a co-designed EER for training mental health care professionals in eMH, (2) investigate whether an EER would be an acceptable solution for this purpose, and (3) determine whether the SDT theory and flow theory are recognized as effective mechanics based on the experience of professionals, thus making a contribution to the field of eMH and the emerging field of EERs in the context of health care.

Co-Designing an EER for eMH Therapy

On the basis of these encouraging arguments and inspiring examples of eMH EERs in pursuing the development of an escape room, we gauged practitioners' first reactions and expectations regarding this idea. This delivered mostly positive responses: mental health care professionals expect an escape room to be innovative, engaging, inviting, and much more reasonable than the traditional and familiar eMH training methods. Moreover, based on our user requirements analysis [39], we aimed to develop a physical escape room. One argument is that a primary obstacle mental health care professionals reported in the user requirement analysis is that they are unacquainted with the possibilities of eMH. On the basis of this, we assumed that a physical room, with a facilitator present to moderate and explain, if necessary, would yield better results. A physical escape room provides more opportunities for collaboration [40]. Moreover, we learned from the user requirements analysis that users preferred a learning context in which they met others and could discuss with colleagues. Finally, playing an escape room is not a stand-alone solution. It should be part of a holistic experience, including a clear introduction and a reflection afterward, where participants can transfer the value of their eMH experience in the escape room to their real-life work practice [39].

As we aimed to develop a serious game for training mental health care professionals in using eMH, it is important to classify an escape room according to the gameplay/purpose/scope model for serious games [41]. This allows a better understanding of an escape room as a serious game and whether it meets our aim of developing a game with educational content. According to the definition by Djaouti et al [41], which distinguishes game-based from play-based activities by noting that the latter lacks a clear goal, we can, we can classify the gameplay of our intended escape room as game-based. The goal of the escape room is to enhance knowledge of the use of eMH by mental health professionals and train them to use eMH tools in different situations. That immediately brings us to the purpose of the escape room, which, in the classification by Djaouti et al [41], would be training. The intention of the escape room is to improve the performance of mental health professionals in using eMH, including how to use certain techniques as well as how to use these techniques in a particular context. We could also argue that there is a data exchange component in our escape room, as we also aim to encourage the users to help each other solve the puzzles and learn from each other's experiences. The scope of the escape room is health care, with mental health

XSL•FO
RenderX

professionals as the main target audience. However, there is also the possibility that an escape room to train eMH skills is incorporated into educational content at vocational or higher professional education [41].

Objectives

This paper presents the co-design process of developing an escape room for training mental health care professionals in eMH. By assessing the expectations and experiences regarding an EER in mental health care, we aim to evaluate the perceived usability and to gauge the effect of underlying mechanisms rooted in SDT and flow theory for training mental health care professionals' eMH skills.

Methods

Study Design

To shape the development and explore the perceived usability of an escape room, we used a process of 3 design stages, in which a testable prototype of an escape room for mental health care professionals was developed. In this "research through design" process, the design of the escape room is part of a research project in which the design activities contribute to generating knowledge [42]. Zimmerman et al [43] define research through design as "a research approach that employs methods and processes from design practice as a legitimate method of inquiry." This knowledge is applied at each stage to iteratively develop prototypes within the design process [42,43], leading to the final prototype. To explore the potential of an escape room, users' experiences and ideas are gathered on whether the escape room works as intended and whether it lives up to the promise of a concept for knowledge and skill enhancement. With this research through design approach, we seek to contribute to evidence-based decisions on the implementation, including potential adaptations and strategies to maximize the users' acceptance of the concept. This means that the design of an escape room is explored for suitability in the practice of mental health care and for its potential to address the aim of enhancing professionals' skills in using eMH. We organized 9 co-design sessions in 3 design stages (Figure 1). The escape room's purpose, expectations, and storyline ideas were formulated in the first design stage. These ideas were further translated into a gaming concept (ie, an escape room prototype), assessed for suitability in the second stage. In the third stage, a prototype was tested with end users (ie, mental health care professionals). Data from all stages within this process contributed to exploring the concept's perceived usability. In the Exploration of an Escape Room in 3 Design Stages subsection in the Results section, we report on the specific procedures and results of the co-design sessions. In doing this, we integrate the procedures and results to describe each design stage as a whole.

Figure 1. Three design stages in the development of an escape room for e-mental health (eMH) training.

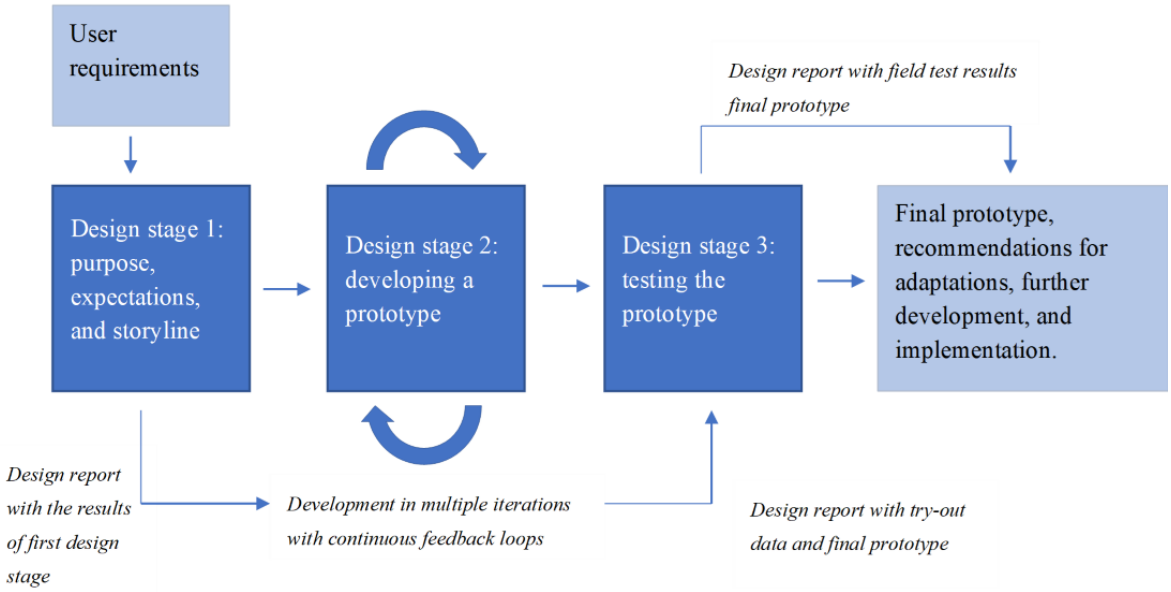

Research Context

The study was conducted at a mental health care organization in Eindhoven (Geestelijke Gezondheidszorg Eindhoven en De Kempen, GGzE), in the southern part of the Netherlands. GGzE delivers specialized mental health care for people with various psychiatric and psychological issues (eg, depression, anxiety disorders, psychosis, autism, and personality disorders). Mental health care professionals working at GGzE include (specialized) psychologists, psychiatrists, nurses, and social workers. GGzE treats approximately 12,000 clients annually and has 2368 employees [44]. The research involved researchers from Tilburg University, Eindhoven Technical University, and Fontys University of Applied Sciences. Through this collaboration between universities, we combined expertise from social science, human and technology interaction, and game design. This combination of expertise vouches for a project in which all necessary aspects of a useful design process (eg, technique, human attitudes and behavior, and serious gaming concepts) and their interaction are considered.

XSL•FO
RenderX

Participants

Stages 1 and 2

The co-design sessions, including 1 web version in stage 2, were held with 5 mental health care professionals (psychologists: n=2, 40%; nurses: n=2, 40%; and social worker: n=1, 20%), 1 innovation staff member, 3 researchers, 3 research laboratory leaders, and 2 game designers. Together, they formed the co-design team for this stage in the design process (Table 1). In line with the existing literature on co-design, we involved stakeholders from the three areas of expertise that were needed for the design project [45-47]: (1) the mental health care professionals that represent the end users of the escape room, the innovation staff that will be involved in launching the escape room in practice; (2) the game designers who have expertise in the development of an escape room; and (3) researchers that are involved in exploring the usability of and experiences with an escape room. The mental health care professionals and the innovation staff members worked at GGzE. They were recruited through an announcement on the organization's internal communication platform and through network contacts of the project team, that is, a convenience sampling method was used. Mental health care professionals with varying levels of expertise were included in this stage of the development process. After their initial agreement, the participants were further informed about the project, the purpose of their participation, and how their data would be handled.

Table 1. Participants' roles in the development of an escape room for e-mental health (eMH) training.

Participants	Design stage 1	Design stage 2	Web version (design stage 2)	Design stage 3
Psychologists	• Participants (n=2)	• Participants (n=2)	—[a]	• Participant field study session (n=3)
Nurses and therapists	• Participant (n=1)	• Participants (n=2)	Participant (n=1)	• Participant field study session (n=2)
Social workers	• Participant (n=1)	—	—	• Participant field study session (n=5)
Mental health professionals' team for web-based treatment	—	—	—	• Participant test session 2 (n=4)
Innovation staff members	• Participant (n=1)	• Participants (n=2)	Participant (n=1)	• Participant test session 1 (n=2) • Facilitator field study sessions 1 to 3 (n=1)
Researchers	• Participants (n=3)	• Participant (n=1) • Observers (n=2)	Participant (n=1)	• Observant field study sessions 1-3 (n=2)
Research laboratory leaders	• Facilitator (n=1) • Participant (n=1) • Observer (n=1)	• Observers (n=3)	Participant (n=1)	—
Game designers	• Facilitator (n=1) • Participant (n=1)	• Facilitator (n=1) • Observant (n=1)	Facilitator (n=1)	—

[a]Not applicable.

Stage 3

In the third stage of the design project, we first asked a purposively sampled selection of mental health care professionals, namely 4 highly experienced eMH therapists from the team for web-based treatment and 2 members of the innovation staff of GGzE, to participate in 2 test sessions. The 2 test sessions secured a sufficient perceived flow of the escape room workshop (ie, including introduction and reflection). They eliminated further technical and practical issues concerning the logistics escape room (Table 1). Subsequently, we invited therapists from all disciplines within ambulatory specialized mental health care to participate in 3 field study sessions to test whether the final prototype of the escape room would be suitable as a concept to enhance mental health care professionals' skills in using eMH (Table 1). The first session included 2 psychologists, 1 nurse specialist, and 1 social worker. The second session involved 5 therapists engaged in social work and supporting clients' daily activities. Finally, a clinical psychologist and a verbal therapist tested the escape room in the third session. All mental health care professionals in the 3 field study sessions had some essential experience with eMH (ie, using a web-based module and video calling on a nonfrequent basis).

Data Collection and Analysis

Detailed notes were taken for all co-design sessions in all stages and were used to compose written reports. These reports were checked with the participants, after which themes were identified to describe, organize, and interpret the data [48]. In addition, video recordings were made of the sessions in stage 3, which were used to verify and complement the written reports. In the

XSL•FO
RenderX

first stage, the 2 sessions were analyzed individually because of the difference in nature of the sessions. In the second stage, the written reports were summarized into recommendations for the design team, which they used in the next iteration of the design. In the third stage, the first 2 test sessions were analyzed together, and the 3 field study sessions were also combined for analysis.

We used the Standards for Quality Improvement Reporting Excellence reporting guidelines to check the completeness and improve the quality of our manuscript [49]. A completed checklist can be found in Multimedia Appendix 1.

Ethical Considerations

This project was approved by the ethics review board of Tilburg University (EC-2013.14) and by the scientific committee of mental health care in Eindhoven (the Netherlands), where the participants were recruited. All participants were informed about the research and allowed to withdraw at any moment. All participants signed the informed consent form. Data in this project were anonymized and cannot be linked back to participants. Participation was voluntary, without financial compensation.

Results

Exploration of an Escape Room in 3 Design Stages

We describe the 3 stages of our research through the design process subsequently (Figure 1). As explained in the Methods section, we elaborate on the specific procedures we followed in each design stage to gather the data, the results generated, and how the findings within all 3 stages provide insights into the usability of an escape room as a game-based solution for eMH training for mental health care professionals (Figure 1).

Design Stage 1: Purpose, Expectations, and Storyline

Overview

In the first co-design session of design stage 1, which took place in December 2019, a joint purpose and common ground for the escape room were formulated. Participants were invited to work in subgroups to discuss the following: (1) the current situation, including the mental health care professionals' attitude regarding eMH; (2) a vision of how future mental health care professionals should be equipped and accommodated to integrate eMH into their daily practice; and (3) the requirements regarding an escape room to transition from the current to the future mental health care professional, in which eMH plays a central role, including the success and failure factors for this concept.

This was done by exchanging ideas, associating concepts, and explaining the relationship between eMH and the work of mental health care professionals.

Following this, a second co-design session was organized, aimed to deliver the storyline for the envisioned escape room. Information was gathered on the particulars of clients in mental health care, the mental health care setting in general, the various treatments and services available, including eMH, and the different client journeys that occur. This provided information on the requirements regarding the content of the game. The designers delineated a fictitious client and a story to form the basis of the escape room game. This initial storyline was verified with the mental health care professionals of the co-design team and was adjusted accordingly.

Purpose

In the first co-design session, the participants mentioned that many mental health care professionals are still convinced that the best way to treat a client is through face-to-face interaction. There is still a feeling of eMH being imposed by financial incentives, offering very limited job enrichment for mental health care professionals and very limited options to enrich treatment possibilities for clients. As a result, practitioners often need more interest or time to explore the possibilities of eMH. An important driver for mental health care practitioners to use eMH in the future would be that they see a clear-cut advantage in it for their clients, such as feeling less of a "patient" and gaining self-control and autonomy. This could be established by gaining positive experiences using eMH, for example, in training. Another aspect of the current attitude of many mental health care professionals mentioned in the co-design session was the lack of knowledge and skills to use eMH. Enhancing this knowledge requires time that needs to be facilitated by the organization they work for.

Despite this attitude, mental health care practitioners see the need to integrate eMH into therapy. The future mental health care professional is seen as someone for whom eMH is a natural part of the job. The session revealed that mental health care professionals consider their profession a craft, and some may see innovation as detrimental to this craft. According to the participants in the session, an escape room should demonstrate the opposite: instead of devaluating the craft, eMH has to reinforce it.

Expectations

On the basis of the purpose of the escape room, it is essential that sharing positive experiences and building confidence in using eMH are emphasized when designing a prototype. In the co-design session, it was stated that examples of positive experiences with eMH in clinical practice should be visible in the escape room, along with opportunities for professionals to share their positive experiences. Furthermore, the educational content of the escape room should seamlessly fit various real-life scenarios, making a transfer to everyday practice easy. Specific and well-defined learning goals should provide the opportunity to gain confidence in using technology, and the escape room should consist of customizable elements to fit the particular learning needs. Finally, the escape room should be engaging and contain continuous feedback loops, a recognizable storyline, and multiplayer options for a shared learning experience.

In addition to game-specific criteria, conditional requirements, such as perceived ease of use, a robust (technological) system, and perceived usefulness were considered significant. Furthermore, the social influence of colleagues, the perception of available time, endorsement by management, and an excellent incentive to engage in the escape room are issues that need to be considered. Finally, the escape room should be a complete learning experience with a clear introduction and reflection after

XSL•FO
RenderX

playing the game. Mental health care professionals must transfer from the learning experience to their therapeutic practice.

Storylines

On the basis of the input in the first session, the designers crafted a storyline for the escape room, which was presented to the co-design team. In their reflection on the storyline, the co-design team remarked that the mental health problems of the fictitious client should be more complex to be eligible for specialized mental health care. It should also contain more details to be sufficiently recognizable for therapists. In this stage of the design process, the co-design team also concluded that it would be valuable if there were sets of multiple storylines revolving around different settings or problems in mental health care to enhance the reach among the potential users of the escape room (Figure 2).

Figure 2. Results of the first design iteration in the development of an escape room for e-mental health (eMH) training.

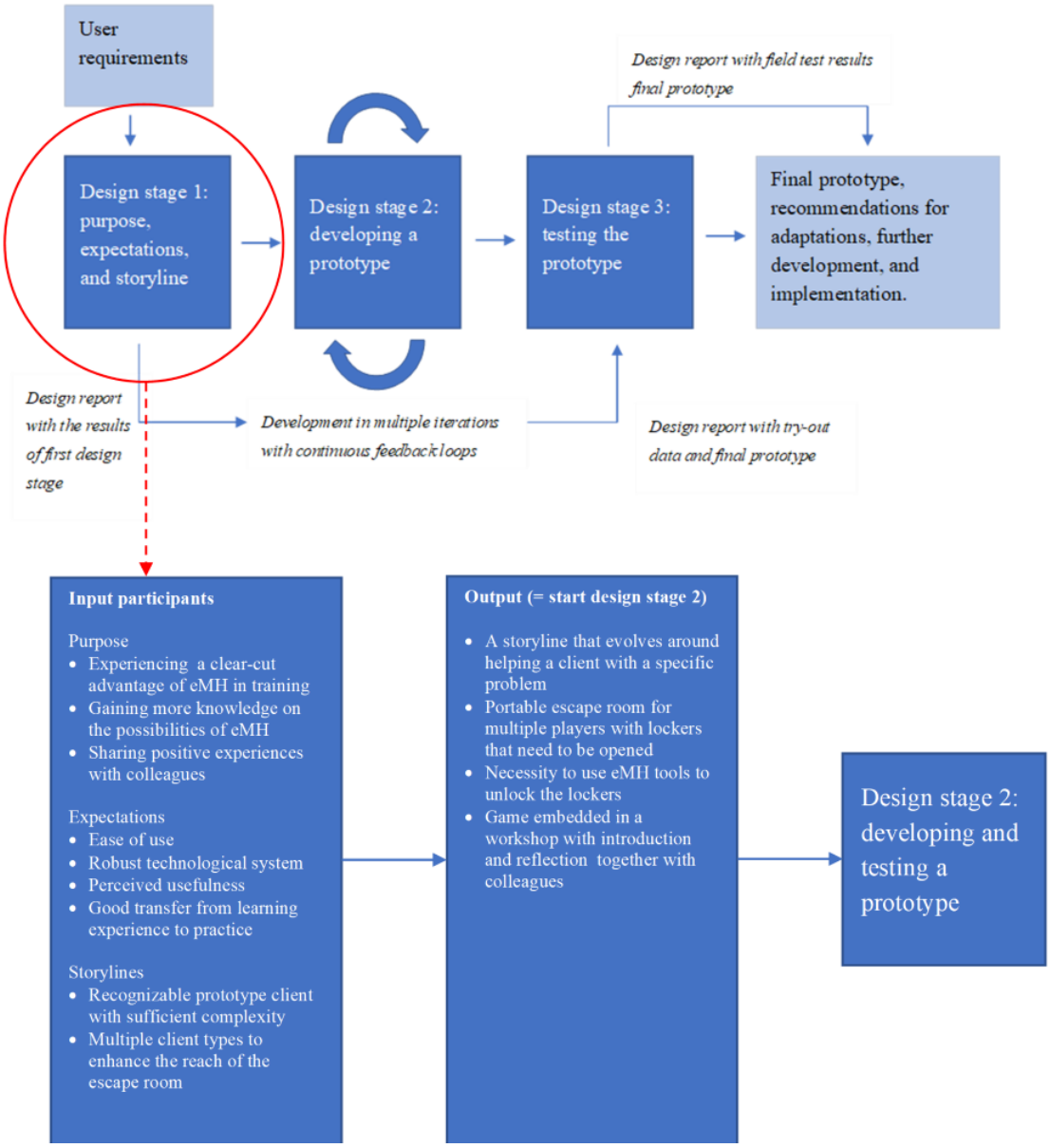

Design Stage 2: Developing a Prototype

Overview

The prototype was presented to the co-design team in February 2020. The prototype of the escape room consists of 2 storylines, which are played separately. In each storyline, the participants are introduced to a client who experiences mental health problems. The challenge in the escape room is to answer questions in the client's electronic patient file, leading to a final solution to help the client take the next step. These questions can only be answered by finding clues in several locked safes, thoroughly analyzing the information given about the client, and using the eMH tools they see in the room. The eMH tools the participants provided to "help their client" were an electronic diary, a VR application for exposure therapy, biofeedback, a chat function, and web-based treatment platform. All materials used in the escape room are portable, which means playing the escape room is not dependent on a fixed place (Figure 3). The actual game is embedded in a workshop with an introduction to familiarize with the topic, goals, and expectations regarding

XSL•FO
RenderX

the escape room and a reflection afterward to discuss the experiences and take-home messages.

In a try-out session, the co-design team tested this prototype by playing the escape room in 3 small groups of 3 participants. This delivered additional data for our research regarding the expectations and experiences of an escape room. At the same time, it provided the designers with feedback for a second prototype. In each round, the participants were given 15 minutes to play the game and solve the puzzles. Each subgroup reflected individually on the prototype directly after playing the escape room. At the end of the session, a plenary discussion took place to summarize the results. The designers developed a second prototype for the escape room based on the feedback. In this phase, a second storyline was also developed. Different eMH tools were added and various connections were made between these eMH tools, the storyline, and the puzzles.

Figure 3. Picture of an escape room setting.

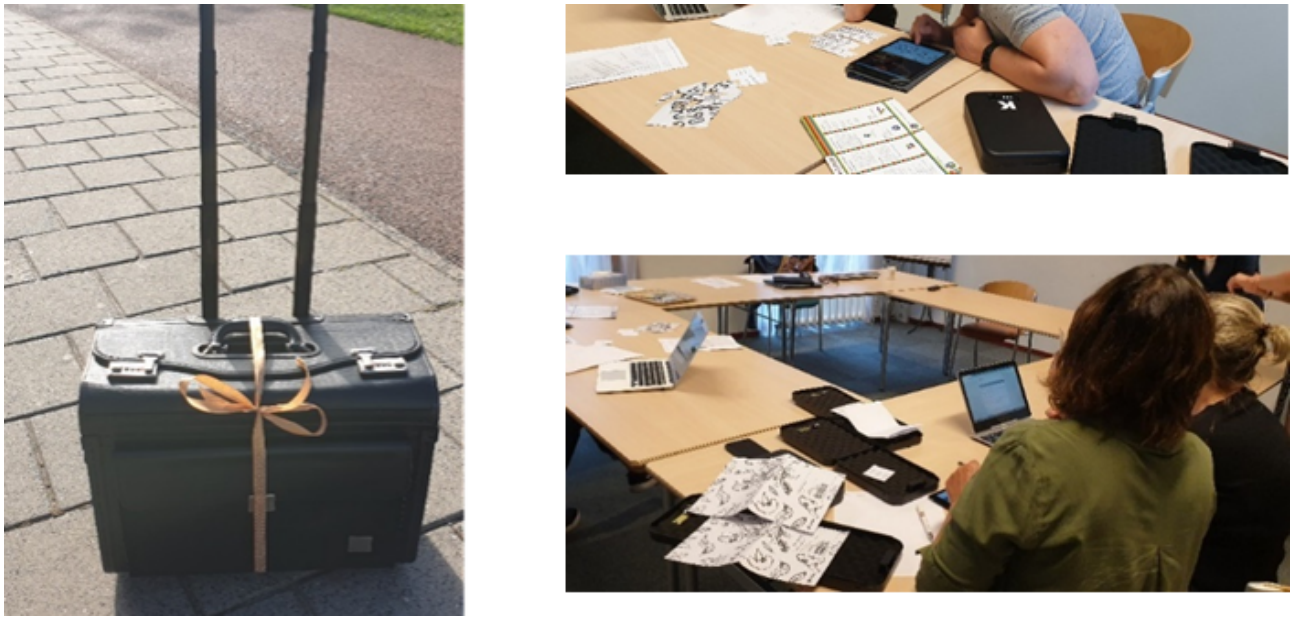

In March 2020, the world was confronted with an unprecedented global pandemic that caused a significant disruption of life as we knew it. Burdened by the consequences of the pandemic and the measures that national governments took to control the virus (eg, social distancing, lockdown, and quarantine), the activities in our design process were also heavily disturbed. Switching to this new situation, we sought to continue the design process by developing a web-based alternative to test the second prototype of the escape room in April 2020. This web version was a digital copy of the physical room, containing the same elements, puzzles, and eMH tools, except for the VR goggles. This second prototype included 2 storylines, each with a fictitious client, and was tested by the co-design team members. One storyline was tested while being moderated by the designers, the other storyline was tested by each member individually, without being moderated.

Try-Out Sessions

The feedback gathered during the try-out sessions delivered several expectations and experiences of mental health care professionals regarding an escape room. First, storylines in an escape room benefit from a good match with a client's pathway in mental health care and reflect truly relatable situations. Moreover, it is essential to focus on mental health problems as an integrated part of everyday life situations. This is important because otherwise, the suggested eMH tools cannot be appropriately linked, and a genuine eMH experience will be missed. This suggestion was also given in response to the final storyline in the previous phase of the design process. Second, there should not be too much focus on solving the puzzles in the escape room, as it diminishes the awareness of eMH in the game. More importantly, the eMH tools should be linked explicitly to the puzzles. The participants should experience a (positive) consequence when choosing, for example, finding the next clue when trying a specific eMH tool or experiencing that someone (ie, the fictitious client) benefits from you as a practitioner choosing to use eMH. Third, variation and differentiation in an escape room vouches for a stronger and fuller user experience. This could be established by multiple storylines (as was stated in the previous phase of the design process), different sets of eMH tools that can be used, and variations in the level of difficulty in the escape room. In this way, the escape room can be played multiple times and addresses different levels of expertise, either regarding eMH or playing an escape room. Fourth, there should be a clear conclusion to the escape room and, consequently, a euphoric feeling when finishing the game. To this end, the purpose at the beginning of the game may need to be more apparent. Fifth, it should be beyond any doubt that the experience of eMH in the escape room can also be transferred to the daily practice of a mental health care professional. Otherwise, the escape room will not be considered credible.

Web Version

From the tryout, it became clear that the experience of a web version of an escape room does not closely resemble the experience of an escape room in real life. Due to several technical hassles, the game's flow was hindered multiple times.

XSL•FO
RenderX

Most importantly, it impeded the possibility of team interactivity, working together, and exchanging ideas about eMH, which was mentioned multiple times in the design process as one of the critical success factors for an escape room.

Final Prototype

After receiving the feedback on the web version of the escape room, the designers continued developing a prototype that was fit for testing by end users. The co-design team subsequently assessed a final prototype in July 2020. After minor adaptations, this prototype was finalized in September 2020. Following this, a field study was planned (stage 3) to gather the user experiences of the escape room among the end users, that is, mental health care professionals from various disciplines (Multimedia Appendix 2).

Design Stage 3: Testing the Final Prototype

Test Sessions

In 2 test sessions (September and October 2020), we first checked the flow of the escape room workshop and the functioning of the materials and technology. This delivered a positive outcome regarding the potential of the concept for exploring eMH and, in that way, enhancing mental health care professionals' knowledge about the possibilities of technology in therapy. The participants of these sessions shared several thoughts on how to improve the workshop before starting the field study sessions with end users.

The first finding in the test sessions was the importance of good functioning of the content and technical aspects of the escape room to ensure a good flow in the game. Besides a good flow in the game, a facilitator should lead a good introduction and reflection to complement the learning experience.

According to the participants in the test sessions, a good explanation of the purpose and relevance of the escape room and presenting a game manual to maximize the effect of playing the escape room is necessary. One participant particularly mentioned the importance of stressing the educational purpose and preventing the feeling that the workshop is merely something for fun. After playing the escape room, reflecting on the purpose is important. According to the participants, there should be a clear take-home message, discussion on the experiences gained in the escape room, and acquired knowledge to share with colleagues afterward.

Another issue raised during the 2 test sessions was the need to continuously adapt the storylines or create the flexibility to insert additional storylines in the escape room game. Two storylines may be too few to address all therapeutic contexts, all levels of complexity in mental health care, and all eMH possibilities that could add value. The participants also marked a sense of reality in the storylines. Opportunism should be avoided because it should not imply that technology will always vouch for a positive treatment outcome. For example, in VR, the escape room should not pretend that it is something that instantly makes everything better. This was regarded as an essential issue to warrant the credibility of the escape room. However, the same session also mentioned that experiencing eMH tools in this manner created awareness about escape room's potential. At the same time, the biofeedback application also showed that it is possible to visually track panic attacks in such an application. The final aspect regarding the mentioned storylines is using the correct professional jargon in the escape room. At one point, a participant stated that the way things were written in the escape room did not reflect the language a therapist would use. In contrast, language and jargon appeared very personal, making it difficult to always be accurate in an escape room. This emphasizes the fact that there is a large diversity among mental health care professionals and the clients they are treating.

Finally, the participants discussed the complexity of the game itself. Some participants found figuring out the puzzles and clues challenging, but others remarked it was too simple. Some solutions were easily found without using the eMH tools in the escape room, resulting in an ineffective learning experience. Some participants in the test sessions suggested creating multiple levels of complexity in the escape room to address several types of game players (ie, level of experience with escape room games) and other kinds of therapists.

On the basis of the results of these 2 test sessions, minor adjustments were made regarding the complexity of the puzzles to ensure that clues and solutions could not be found without using eMH tools. Furthermore, the introduction and reflection were refined to embed the escape room game in a workshop format (Figure 4).

XSL•FO
RenderX

Figure 4. Results of the test and field study sessions (third iteration) in the development of an escape room for e-mental health (eMH) training.

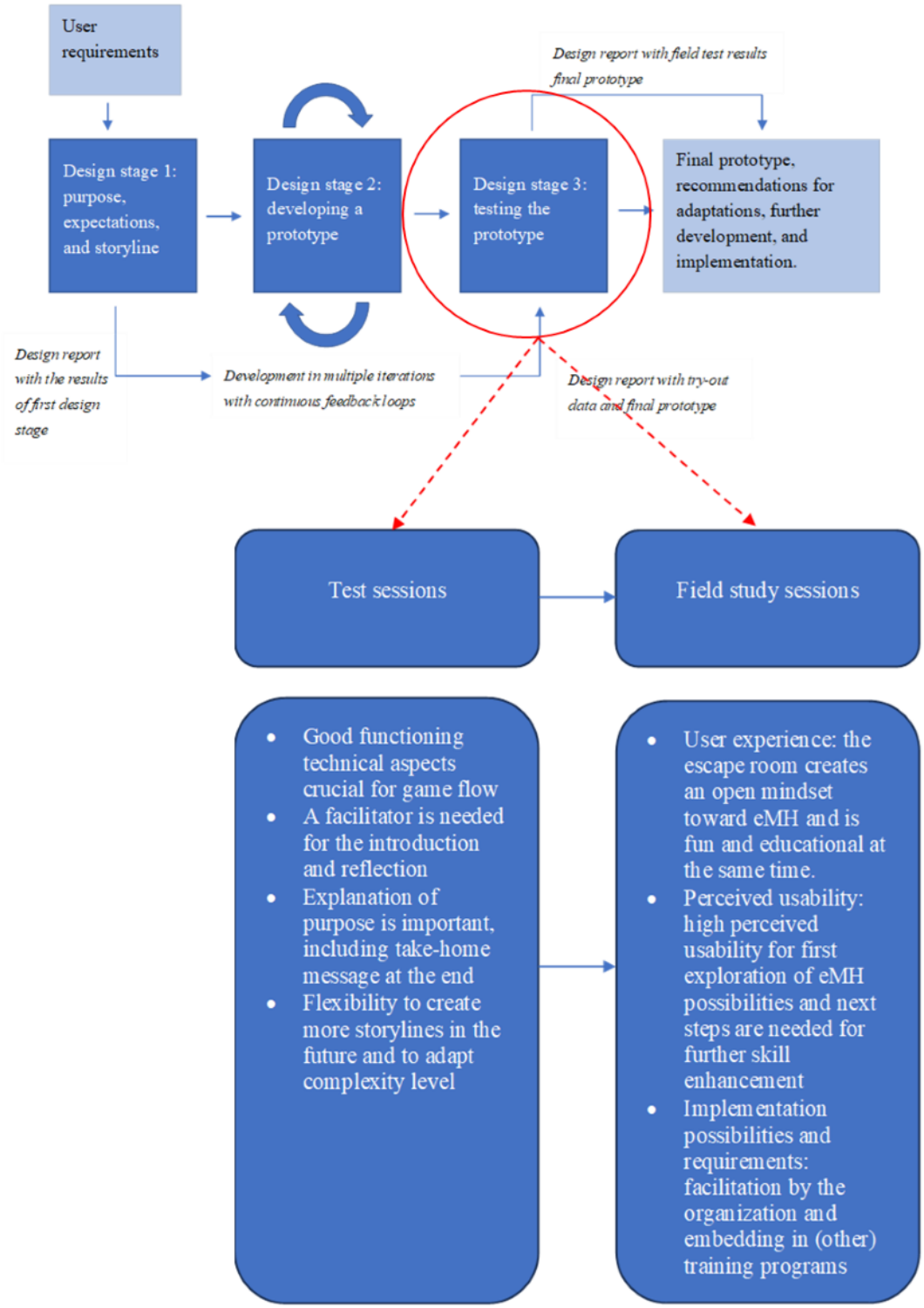

Field Study Sessions

Overview

Hereafter, the workshop was tested in 3 sessions with mental health care professionals from different areas of expertise. A total of 11 mental health care professionals participated in these field study sessions (November and December 2020 and June 2021) and were asked to fill in a questionnaire (Multimedia Appendix 3) at the end of the workshop. Their answers formed the basis for a discussion about their user experiences. The questionnaire consisted of an overall experience rating (question 1), three parts to reflect on (1) the quality of the introduction (questions 2-4), (2) the game itself (questions 5-11), and (3) the reflection (questions 12-15), overall questions regarding the perceived usability (questions 16-19), and a question on strengths and suggestions (question 20). The questions in the 3 parts of the questionnaire were constructed based on the results of the co-design sessions in design stage 1. As we explained in design stage 1, mental health care professionals use eMH mainly because of its apparent benefits. To encourage this, the escape

XSL•FO
RenderX

room should therefore highlight positive experiences with eMH in clinical practice, in line with the flow theory that states that a positive learning experience can be reached if the challenges meet the abilities of the users and if they can share this with colleagues [31]. To explore the extent to which the escape room was experienced positively, we formulated questions 1, 3, 9, 10, and 13. Furthermore, from design stage 1 and SDT, we learned that the educational content must align with real-life scenarios, making it easy to apply in practice [29]. Questions 6, 8, and 15 were therefore constructed to assess whether the storylines are recognizable and whether the training with eMH tools in these storylines connects to their daily practice. In addition, questions 2 and 5 were formulated to explore if the learning goals become clear, and questions 7 and 12 assessed whether the escape room helps to build confidence in using technology, with customizable elements tailored to specific needs. Finally, the game should enable multiplayer options for shared learning [31]. To assess whether the escape room was experienced that way, we added questions 4, 11, and 14. Together this led to the questionnaire, which had 20 targeted questions. The answers to the questions were subsequently used as a guide in discussion with the participants to gain a more in-depth understanding of the possibilities of an escape room as an environment for mental health care professionals to experiment and practice with eMH.

User Experiences

From the questionnaire and discussion, we learned that the participants in the field study sessions were enthusiastic about the workshop they participated in. They stated that the workshop is inspiring and creates a positive mindset toward eMH. They recognized the potential to remove barriers they had at first, showing that eMH is accessible and not always that difficult to use. The participants experienced significant fun in the escape room, making it feel less obligatory. This is expected to help people engage more easily. In addition, the alternation between playing and reflecting was mentioned as meaningful. In total, 2 (18%) of the 11 participants mentioned they were pleased that there was no explicit element of competition in the escape room. According to these participants, too much competition, for example, ranking who has solved the escape room the fastest, would distract the users from the storyline and the eMH possibilities they come across.

All but one participant expressed that they experienced the storylines in the escape room as realistic. One mental health care professional did not recognize the storylines in the escape room. Another aspect related to this was the “credibility” of the suggested effect of eMH in the escape room. Although most (9/11, 82%) of the participants could describe the use of eMH to certain benefits for mental health care, it should be carefully suggested in the escape room. However, most (9/11, 82%) participants experienced how they could practically apply eMH in therapeutic situations. For example, VR was experienced in the escape room, resulting in a discussion on how VR could be used for clients in specific treatment situations. The participants were able to connect their VR experience in the workshop to real-life situations and could see the added value of using eHealth. All kinds of eHealth suggestions were exchanged, which illustrates the effect of being given the possibility to explore and experiment with eHealth in this type of playful setting.

Participants wondered whether an escape room would appeal enough to therapists who were skeptical about eMH. The participants also mentioned that the current storylines and eMH tools may not be recognizable for all mental health care professionals. It was considered important to explore eMH tools that are also accessible in real life to be able to transfer knowledge from training to reality. In contrast, it is also mentioned that mental health care professionals with vast experience in eMH will find it more attractive to explore more experimental eMH tools.

Perceived Usability of an Escape Room for eMH Training

To explore the perceived usability of an escape room for skill enhancement in eMH, we asked the participants to indicate whether the workshop provided them with a more elaborate and precise picture of the possibilities of eMH (Table 2). The feedback on this issue varied among the participants. In total, 2 (18%) of the 11 mental health care professionals reported that they had not gained a better picture of the options of eMH in the escape room. This was because they already had a relatively good basic knowledge about eMH beforehand. The participants who were less experienced users of eMH reflected that they had a much better idea about the possibilities of eMH after participating in the escape room workshop. It was suggested that the current escape room prototype would benefit from further development to create more options to try additional eMH tools because, in the current prototype, not all mental health care professionals would find tools suitable for their therapeutic practice.

Table 2. Perceived usability of an escape room (ER) for e-mental health (eMH) training (questionnaire results).

Items	Score on a scale of 1-5, mean (SD)
Suitability of an ER to explore possibilities of eMH	4.7 (0.92)
Possibilities for collaboration in the ER	4.7 (0.47)
Clarity of the storylines	4.3 (0.92)
Clarity of learning goal in the ER	4.6 (0.67)
Balance educational content and fun	4.7 (0.68)
Usefulness for daily practice	3.8 (0.77)

XSL•FO
RenderX

The participants largely agreed on the suitability of an escape room concept for exploring eMH (average score of 4.7, SD 0.92). They perceived the learning goal in the workshop as evident (average score of 4.6, SD 0.67). They also experienced a good balance between the educational and the fun part in the escape room (average score of 4.7, SD 0.68). It was mentioned in the discussion that suitability is supported when the bridge between the escape room and everyday practice is explicitly discussed in the workshop as well as the steps that can be taken after completing the workshop and returning to the workplace. Follow-up training is essential because otherwise the acquired knowledge will not be secured. According to the participants, this escape room offers an excellent first experience with the possibilities of eMH; more than that, it is just another educational workshop, "It is more 'getting to know eMH' than 'learning how to apply eMH." It was also suggested that more focus on the client groups that people are treating would be helpful in the reflection part of the workshop. The participants advocated that involving professionals with similar client groups or treatment contexts would be more effective in stimulating the exchange of experiences with eMH and the transferability of the experiences in the escape room to daily practice.

In line with the findings about follow-up training, the participants elaborately reflect on how to apply and secure knowledge on eMH once the first experiences are gained in the escape room. While this workshop helps to talk about eMH, keeping the conversation going in everyday practice is equally important. According to the participants in our study, mental health care professionals are focused on the content of their treatment and support, and the knowledge gained in the eMH escape room might lose its impact. It was stated that the organization should provide continuous information about what is new in the field of eMH and should provide time to invest in exploring new tools. The concept of an escape room or a game offers a very persuasive opportunity to invest in keeping knowledge up to date and motivating people. All workshops suggested that this would be an excellent fit for a teambuilding activity.

Implementation Possibilities and Requirements

In addition to gathering experiences with the escape room workshop, we asked the mental health care professionals who participated to indicate what they deemed essential for successfully implementing it. The first thing mentioned was that professionals in mental health care should be confronted with eMH from various directions, that is, in an escape room, but also by clients, colleagues, and organizational policies aimed at being innovative, requiring innovative and digitally skilled employees. Incorporating an escape room in an introductory meeting might be possible when people are new to the organization.

Another issue that emerged from the workshops concerns the approachability of the escape room. It is experienced as convenient that the escape room is currently a mobile version, which does not necessarily require people to go somewhere and spend much time traveling. It would be even more convenient if the escape room could be played without an external facilitator so that participants could be more flexible in their time planning. However, this remains the main concern of the participants in the workshops: will mental health care professionals be provided with enough time to explore eMH in the future, and will their production targets be adjusted accordingly? In addition, it should be decided who will provide training services and materials in the organization. Finally, the participants stated that a strong vision by the policy makers on the use of eMH in the organization and the expectations regarding the future mental healthcare professional is indispensable to warrant a successful implementation of an escape room or any serious gaming concept for eMH training in mental health care.

Discussion

Principal Findings

In this research through design study, we investigated the perceived usability of an escape room for training mental health care professionals in using eMH. On the basis of the results in each stage of our research and the overall findings, we can conclude that an escape room concept might be suitable for this type of training. We subsequently discuss our findings by reflecting on the experiences with our escape room prototype and what this means in light of the purpose of our design project: developing new training methods for mental health care professionals, with the ultimate goal of contributing to increasing the adoption of eMH. We conclude with an overall view of the possibilities of game-based training in mental health care based on the results of this design study.

What we know from earlier research, which also emerged in this study, is that many mental health care professionals are not convinced of the benefits of eMH, resulting in dismissive attitudes and behavior [50,51]. The escape room was experienced as both fun and educational; however, most importantly, it induced a more positive stance toward eMH tools. While an open mindset is a prerequisite for change, the escape room may offer a potential concept as the first step to changing attitudes and behavior among mental health care professionals regarding eMH [36,40].

For the escape room to be accepted as a valuable training method by its end users, the storylines, game elements (puzzles and clues), and eMH tools in the escape room needed to be credible and recognizable for mental health care professionals. At the same time, it became clear that it is tough to design storylines that seamlessly match the practices of mental health care therapists. In one of the first versions of the prototype, there was too much focus on everyday life situations instead of mental health problems. This is crucial because the participants need a genuine eMH experience to link to their treatment context properly. This feedback was given in different iterations in the design process, which marks the difficulty of designing a good and representative storyline and the importance of involving multiple disciplines in a co-design team [46,47].

Moreover, it became clear that mental health care professionals are diverse in terms of their client population, type of mental health care services, and level of adoption in eMH. This requires a flexible and adaptable training concept to expand the reach of the escape room. Greater diversification regarding the

XSL•FO
RenderX

storylines and available eMH tools in the escape room may be needed, along with various difficulty levels. The participants in the field study sessions indicated that the escape room in its current form might not be suitable for people who are already frequent users of eMH. This implies that the gains of the current escape room are probably highest for people who are relatively inexperienced eMH users, which applied to most mental health care professionals when we started the design process. Adding new, more innovative tools could, however, expand the reach of the escape room.

The participants welcomed the escape room as an engaging concept and noted that its design needs continuation and further development. In its current form, the escape room offers an environment for exploration and the first acquaintance with eMH rather than actual skill development in applying these tools. We believe that, whereas most mental health care professionals are—partly because of the coronavirus pandemic—familiar with standard eMH tools that are also used in everyday life (eg, videoconferencing and SMS text messaging), they remain unaware of the possibilities of more novel technologies, such as VR and biofeedback, and may be biased about its use in their daily clinical practice. Therefore, creating a different mindset remains an important foundation for further skill development in the future and increasing adoption in general. An extensive research base reports on various barriers that prevent accelerating the adoption of eMH [5,51,52]. This escape room addresses one of the critical barriers, a perceived lack of knowledge. However, issues such as compensating for the loss of nonverbal information, the lack of trust in technological systems, and the availability of time and resources to gain positive experiences in using eMH also need to be addressed.

When exploring ideas for further training for specific eMH tools, the question is whether an escape room would also be suitable for this particular type of training. Looking at escape rooms in health care, we find that escape rooms are most appropriate for exploration and awareness and less ideal for training-specific skills [23,24]. This was also why we chose an escape room concept in the first place; it suited the results of our user requirements analysis that revealed a specific need for exploring and enhancing knowledge rather than specific hands-on skills. Training these hands-on skills should be addressed for the next step, particularly because the participants in our study also indicated this need. Moreover, for both the possibility of exploration and specific skill training, the support in terms of time and resources provided by the organization and its management is crucial. The organization's commitment to a digital transition and the necessity for mental health care professionals to be equipped with 21st-century skills must be at the forefront. Finally, follow-up research is required to test the escape room's effectiveness regarding knowledge increase and skill enhancement.

The Perceived Usability of Serious Gaming for eMH Training in Mental Health Care Based on Users' Experiences

The results of our study show positive experiences with an escape room as a serious gaming concept for eMH training in mental health care. These positive experiences might be explained in terms of the SDT [29] and the flow theory [31]. By attempting to address the needs of competence and autonomy, the users may have had a positive experience. The flow of the escape room may have positively contributed to their motivation to enhance their knowledge of eMH and their general attitude toward using eMH in their clinical practice [30].

In the design of the escape room, we addressed several important game mechanics that we found to have influenced our users' experiences, for example, an adequate level of complexity in the game [19,53,54]. The participants in our study emphasized that when a puzzle in the escape room takes too much time or is too difficult, it may lead to frustration and an adverse effect. The opposite of this also works negatively; when clues are too obvious, people may not feel they are taken seriously. In general, our escape room was experienced as having an appropriate level of difficulty, which led to a positive experience. We also paid much attention to the importance of the story or narrative in the escape room. When this narrative is too farfetched or not relatable to clinical practice, it loses strength; when people can identify with the story in the game, it may be much more effective than any other learning method [29-31]. In addition, the aspect of social interaction [30] was predominantly present in our escape room and experienced as one of the critical factors for its success.

Strengths and Limitations

In this study, we particularly choose a research through design approach to develop and test an escape room as a serious gaming solution for eMH training among mental health care professionals. The strength of this approach is that it enabled us to use formative evaluations [55] between the stages of design to provide input for the next design stage. We strongly believe that without these intermediate assessments, we could not design a prototype suitable to the practice of mental health care providers. More importantly, the involvement of mental health care professionals in our co-design team ensured a prototype that matched the therapeutic practice we pursued to reflect in our escape room. However, there are some limitations. First of all, we used an inductive analysis method, which allowed us to gather any information deemed necessary for the development of the escape room but may have lacked the required structure to establish an evaluation of the effectiveness of the escape room, in terms of skill enhancement, and for other researchers to replicate the analyses. Analysis procedures that would allow an assessment of the escape room's success could be to use a pre- and posttest quasi-experimental design for the evaluation [56], a controlled experiment with naturalistic and summative evaluations to assess whether the purpose of the game is indeed reached [57] or a heuristic evaluation using predefined rules of thumb to measure the usability of the escape room [58]. A second limitation of this study is that we included a small sample of mental health care professionals who all worked at the same mental health care organization. In other words, we did not include participants from a diverse mental health care setting other than the differentiation in mental health care contexts within an organization. In addition, we are aware that for projects like this, where participants are recruited through open invitation, people who are already enthusiastic about the topic,

XSL•FO
RenderX

in this case, eMH, tend to be the ones who sign up for participation. In future research, it would be worthwhile to test the escape room or any other game-based training solution among mental health care professionals who are skeptical about eMH to see whether also, in this group, the effect of creating an open mindset through exploration and experience-based learning is achieved.

Conclusions

In general, we can conclude that experience-based training, such as game-based training in an escape room, helps people experience what they are being trained to do or use. This experience may be much stronger than the experience of learning through traditional learning (eg, classroom, textbook, or e-learning). When training is more fun and engaging, this will increase the learning experience. Moreover, it increases the participants' motivation and creates an open mindset. In our study, we learned that an escape room as a serious gaming concept stimulates the willingness to participate in training and be open to change. This openness to change is essential in increasing the use of eMH. However, it is important that there is a clear next step to this first exploration, providing professionals with the opportunity to thoroughly train their skills in specific eMH tools. Subsequently, it would be valuable to evaluate the effectiveness of game-based training in an escape room for enhancing particular skills. Furthermore, designing is a cooperation between different experts. The design process should be viewed from different angles and areas of expertise. The contribution of mental health care professionals was necessary to create a credible game in which both the storyline and set of integrated eMH tools can be transferred to their therapeutic practice. Finally, the success of a game-based training method, such as an escape room, depends on the commitment of the organization in which it is implemented. People need to feel that they have the time and support of their management to engage in training. This commitment is more than just telling them to "go and participate in training." The whole organization must enhance specific knowledge based on common ground and a shared vision of what is needed regarding competencies and skills by mental health care professionals.

Acknowledgments

The authors would like to thank all mental health care professionals and other participants who participated in this project. This work is supported by the Netherlands Organisation for Scientific Research grant 055.16.141, under the project titled Psychosocial Games Supporting Mental Health Professionals in Stress Reduction and Empathic Interactions in Remote/online Psychotherapy, under the Serious Games for Professional Skills Program.

Data Availability

The datasets generated and analyzed during this study are not publicly available due to agreements about confidentiality and exchange of data that were made with the respondents in the study. The consent given by the participants does not allow for public availability. The data can be made available from the corresponding author on reasonable request and with the consent of the participants in this study.

Authors' Contributions

JB, MAF, and RJWST organized and conducted the co-design sessions. JB coded and analyzed the data and composed an initial draft for the manuscript. All authors contributed to multiple revisions. All authors agreed with the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
SQUIRE (Standards for Quality Improvement Reporting Excellence) checklist.
[DOCX File , 34 KB - formative_v9i1e58650_app1.docx]

Multimedia Appendix 2
Results of the second design iteration in the development of an escape room for e-Mental health training.
[DOCX File , 23 KB - formative_v9i1e58650_app2.docx]

Multimedia Appendix 3
Evaluation questions (English translation).
[PDF File (Adobe PDF File), 164 KB - formative_v9i1e58650_app3.pdf]

References

XSL•FO
RenderX

1. Andersson G, Cuijpers P, Carlbring P, Riper H, Hedman E. Guided internet-based vs. face-to-face cognitive behavior therapy for psychiatric and somatic disorders: a systematic review and meta-analysis. World Psychiatry 2014 Oct 01;13(3):288-295 [FREE Full text] [doi: 10.1002/wps.20151] [Medline: 25273302]
2. de Grood C, Raissi A, Kwon Y, Santana M. Adoption of e-health technology by physicians: a scoping review. J Multidisciplin Healthc 2016 Aug;Volume 9:335-344. [doi: 10.2147/jmdh.s103881]
3. Glueckauf RL, Maheu MM, Drude KP, Wells BA, Wang Y, Gustafson DJ, et al. Survey of psychologists' telebehavioral health practices: technology use, ethical issues, and training needs. Prof Psychol Res Pract 2018 Jun;49(3):205-219. [doi: 10.1037/pro0000188]
4. Lal S, Adair CE. E-mental health: a rapid review of the literature. Psychiatr Serv 2014 Jan 01;65(1):24-32. [doi: 10.1176/appi.ps.201300009] [Medline: 24081188]
5. Feijt MA, de Kort YA, Bongers IM, IJsselsteijn WA. Perceived drivers and barriers to the adoption of eMental health by psychologists: the construction of the levels of adoption of eMental health model. J Med Internet Res 2018 Apr 24;20(4):e153 [FREE Full text] [doi: 10.2196/jmir.9485] [Medline: 29691215]
6. Vis C, Kleiboer A, Prior R, Bønes E, Cavallo M, Clark SA, et al. Implementing and up-scaling evidence-based eMental health in Europe: the study protocol for the MasterMind project. Internet Interv 2015 Nov;2(4):399-409. [doi: 10.1016/j.invent.2015.10.002]
7. Wozney L, Newton AS, Gehring ND, Bennett K, Huguet A, Hartling L, et al. Implementation of eMental Health care: viewpoints from key informants from organizations and agencies with eHealth mandates. BMC Med Inform Decis Mak 2017 Jun 02;17(1):78 [FREE Full text] [doi: 10.1186/s12911-017-0474-9] [Medline: 28577543]
8. Kilbourne AM. E-health and the transformation of mental health care. Psychiatr Serv 2012 Nov;63(11):1059. [doi: 10.1176/appi.ps.631103] [Medline: 23117506]
9. Ross J, Stevenson F, Lau R, Murray E. Factors that influence the implementation of e-health: a systematic review of systematic reviews (an update). Implement Sci 2016 Oct 26;11(1):146 [FREE Full text] [doi: 10.1186/s13012-016-0510-7] [Medline: 27782832]
10. Chan CV, Kaufman DR. A framework for characterizing eHealth literacy demands and barriers. J Med Internet Res 2011 Nov 17;13(4):e94 [FREE Full text] [doi: 10.2196/jmir.1750] [Medline: 22094891]
11. Norman C. eHealth literacy 2.0: problems and opportunities with an evolving concept. J Med Internet Res 2011 Dec 23;13(4):e125 [FREE Full text] [doi: 10.2196/jmir.2035] [Medline: 22193243]
12. Schalken F. Handboek Online Hulpverlening: Met Internet Zorg en Welzijn Verbeteren. Houten, The Netherlands: Bohn Stafleu Van Loghum B.V; 2013.
13. Bierbooms JJ, van Haaren M, IJsselsteijn WA, de Kort YA, Feijt M, Bongers IM. Integration of online treatment into the "new normal" in mental health care in post-COVID-19 times: exploratory qualitative study. JMIR Form Res 2020 Oct 08;4(10):e21344 [FREE Full text] [doi: 10.2196/21344] [Medline: 33001835]
14. Crawford A, Gratzer D, Jovanovic M, Rodie D, Sockalingam S, Sunderji N, et al. Building eHealth and telepsychiatry capabilities: three educational reports across the learning continuum. Acad Psychiatry 2018 Dec 11;42(6):852-856. [doi: 10.1007/s40596-018-0989-0] [Medline: 30311178]
15. Echelard JF, Méthot F, Nguyen HA, Pomey MP. Medical student training in eHealth: scoping review. JMIR Med Educ 2020 Sep 11;6(2):e20027 [FREE Full text] [doi: 10.2196/20027] [Medline: 32915154]
16. Timman A, Caris JA. Leren in een andere realiteit. Onderwijs en Gezondheidszorg 2022;1:20-23. [doi: 10.24078/oeng.2022.2.128459]
17. Connolly TM, Boyle EA, MacArthur E, Hainey T, Boyle JM. A systematic literature review of empirical evidence on computer games and serious games. Comput Educ 2012 Sep;59(2):661-686. [doi: 10.1016/j.compedu.2012.03.004]
18. Squire K, Jenkins H. Harnessing the power of games in education. Insight 2003;3:5-33.
19. Wang R, DeMaria SJ, Goldberg A, Katz D. A systematic review of serious games in training health care professionals. Simul Healthc 2016 Feb;11(1):41-51. [doi: 10.1097/SIH.0000000000000118] [Medline: 26536340]
20. Susi T, Johannesson M, Backlund P. Serious games – an overview. Serious Games Cluster and Business Network Project. 2007 Feb 5. URL: https://www.diva-portal.org/smash/get/diva2:2416/FULLTEXT01.pdf [accessed 2024-12-16]
21. Makri A, Vlachopoulos D, Martina RA. Digital escape rooms as innovative pedagogical tools in education: a systematic literature review. Sustainability 2021 Apr 20;13(8):4587. [doi: 10.3390/su13084587]
22. Ritterfeld U, Cody M, Vorderer P. Serious Games: Mechanisms and Effects. Milton Park, UK: Taylor & Francis; 2009.
23. Adams V, Burger S, Crawford K, Setter R. Can you escape? Creating an escape room to facilitate active learning. J Nurses Prof Dev 2018;34(2):E1-E5. [doi: 10.1097/nnd.0000000000000433]
24. Mullen M, Seiler SO. Escaping the confines of traditional instruction: the use of the escape room in a transition to practice program. J Nurses Prof Dev 2019;35(6):354-356. [doi: 10.1097/nnd.0000000000000571]
25. Veldkamp A, van de Grint L, Knippels MC, van Joolingen WR. Escape education: a systematic review on escape rooms in education. Educ Res Rev 2020 Nov;31:100364. [doi: 10.1016/j.edurev.2020.100364]
26. Kolb DA, Boyatzis RE, Mainemelis C. Experiential learning theory: previous research and new directions. In: Sternberg RJ, Zhang LF, editors. Perspectives on Thinking, Learning, and Cognitive Styles. New York, NY: Routledge; 2014.

XSL•FO
RenderX

27. Prensky M. Fun, play and games: what makes games engaging. In: Digital Game-Based Learning. New York, NY: McGraw-Hill Education; 2001:5-31.
28. Grande-de-Prado M, García-Martín S, Baelo R, Abella-García V. Edu-escape rooms. Encyclopedia 2020 Dec 22;1(1):12-19. [doi: 10.3390/encyclopedia1010004]
29. Deci EL, Ryan RM. Handbook of Self-Determination Research. New York, NY: University of Rochester Press; 2004.
30. Vorderobermeier A, Abel J, Sailer M. Theoretical foundations and approaches in research on educational escape rooms: a systematic review. Educ Res Rev 2024 Aug;44:100625. [doi: 10.1016/j.edurev.2024.100625]
31. Csikszent M. Flow: The Psychology of Optimal Experience. New York, NY: HarperCollins; 1991.
32. Deen M. G.A.M.E., Games autonomy motivation & education: how autonomy-supportive game design may improve motivation to learn. Technische Universiteit Eindhoven. 2015. URL: https://pure.tue.nl/ws/files/3895781/784813.pdf [accessed 2024-12-07]
33. Okuda Y, Bryson EO, DeMaria SJ, Jacobson L, Quinones J, Shen B, et al. The utility of simulation in medical education: what is the evidence? Mt Sinai J Med 2009 Aug 29;76(4):330-343. [doi: 10.1002/msj.20127] [Medline: 19642147]
34. Reeves B, Read JL. Total Engagement: How Games and Virtual Worlds Are Changing the Way People Work and Businesses Compete. Brighton, MA: Harvard Business Review Press; 2009.
35. Rodriguez-Ferrer JM, Manzano-León A, Cangas AJ, Aguilar-Parra JM. A web-based escape room to raise awareness about severe mental illness among university students: randomized controlled trial. JMIR Serious Games 2022 May 05;10(2):e34222 [FREE Full text] [doi: 10.2196/34222] [Medline: 35511232]
36. Aragon CJ. Determining the effectiveness of an escape room in a mental health nursing course: a quantitative study. Northcentral University. 2023. URL: https://www.proquest.com/openview/afaa70dd57076e8ae4bab3fa3631ac20/1?pq-origsite=gscholar&cbl=18750&diss=y [accessed 2024-12-16]
37. Petkari E, Calvo A. Escaping psychodiagnostics: implementation of a Digital Educational Escape Room with future mental health professionals. Interact Learn Environ 2024 May 23:1-10. [doi: 10.1080/10494820.2024.2354429]
38. Bierbooms J, Feijt M, IJsselsteijn W, Sluis-Thiescheffer W, Weijmans M, van der Burg E, et al. Serious games for professional skills: the design of an escape room to explore the possibilities of eMental health. In: Proceedings of the Companion Publication of the 2020 ACM Designing Interactive Systems Conference. 2020 Presented at: DIS' 20 Companion; July 6-10, 2020; Eindhoven, The Netherlands. [doi: 10.1145/3393914.3395888]
39. Bierbooms J, Feijt MA, IJsselsteijn WA, Bongers IM. Design of a game-based training environment to enhance mental health care professionals' skills in using e-mental health: multiple methods user requirements analysis. JMIR Serious Games 2022 Jul 27;10(3):e34700 [FREE Full text] [doi: 10.2196/34700] [Medline: 35896032]
40. Pan R, Lo H, Neustaedter C. Collaboration, awareness, and communication in real-life escape rooms. In: Proceedings of the 2017 Conference on Designing Interactive Systems. 2017 Presented at: DIS '17; June 10-14, 2017; Edinburgh, UK. [doi: 10.1145/3064663.3064767]
41. Djaouti D, Alvarez J, Jessel JP. Classifying serious games: the G/P/S model. In: Handbook of Research on Improving Learning and Motivation through Educational Games: Multidisciplinary Approaches. Hershey, PA: IGI Global; 2011.
42. Stappers PJ, Giaccardi E. Research through design. In: Soegaard M, Friis-Dam R, editors. The Encyclopedia of Human-Computer Interaction. Aarhus, Denmark: The Interaction Design Foundation; 2014.
43. Zimmerman J, Forlizzi J, Evenson S. Research through design as a method for interaction design research in HCI. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2007 Presented at: CHI '07; April 28-May 3, 2007; San Jose, CA. [doi: 10.1145/1240624.1240704]
44. Jaarverslag GGzE 2020: werken aan mentale vooruitgang. GGzE. URL: https://www.jaarverslagggze.nl/ [accessed 2021-10-22]
45. Donetto S, Pierri P, Tsianakas V, Robert G. Experience-based co-design and healthcare improvement: realizing participatory design in the public sector. Des J 2015 May 07;18(2):227-248. [doi: 10.2752/175630615X14212498964312]
46. Jessen S, Mirkovic J, Ruland CM. Creating gameful design in mHealth: a participatory co-design approach. JMIR Mhealth Uhealth 2018 Dec 14;6(12):e11579 [FREE Full text] [doi: 10.2196/11579] [Medline: 30552080]
47. Robert G. Participatory action research: using experience-based co-design to improve the quality of healthcare services. In: Ziebland S, Coulter A, Calabrese JD, Locock L, editors. Understanding and Using Health Experiences: Improving Patient Care. Oxford, UK: Oxford University Press; 2013.
48. Boyatzis RE. Transforming Qualitative Information: Thematic Analysis and Code Development. Ann Arbor, MI: The University of Michigan; 1998.
49. Ogrinc G, Davies L, Goodman D, Batalden P, Davidoff F, Stevens D. SQUIRE 2.0 (Standards for QUality Improvement Reporting Excellence): revised publication guidelines from a detailed consensus process. BMJ Qual Saf 2016 Dec;25(12):986-992 [FREE Full text] [doi: 10.1136/bmjqs-2015-004411] [Medline: 26369893]
50. Connolly SL, Miller CJ, Lindsay JA, Bauer MS. A systematic review of providers' attitudes toward telemental health via videoconferencing. Clin Psychol (New York) 2020 Jan 06;27(2):10.1111/cpsp.12311 [FREE Full text] [doi: 10.1111/cpsp.12311] [Medline: 35966216]

XSL•FO
RenderX

51. Davies F, Shepherd HL, Beatty L, Clark B, Butow P, Shaw J. Implementing web-based therapy in routine mental health care: systematic review of health professionals' perspectives. J Med Internet Res 2020 Jul 23;22(7):e17362 [FREE Full text] [doi: 10.2196/17362] [Medline: 32706713]
52. Stoll J, Müller JA, Trachsel M. Ethical issues in online psychotherapy: a narrative review. Front Psychiatry 2019 Feb 11;10:993 [FREE Full text] [doi: 10.3389/fpsyt.2019.00993] [Medline: 32116819]
53. Ijaz A, Khan MY, Ali SM, Qadir J, Kamel Boulos MN. Serious games for healthcare professional training: a systematic review. Eur J Biomed Inform 2019;15(1):14-30 [FREE Full text]
54. Rooney P. Creating serious games at third level: evaluating the implications of an in-house approach. In: Proceedings of the European Conference on Game Based Learning. 2012 Presented at: ECGBL 2012; October 4-5, 2012; Cork, Ireland.
55. van Gemert-Pijnen JE, Nijland N, van Limburg M, Ossebaard HC, Kelders SM, Eysenbach G, et al. A holistic framework to improve the uptake and impact of eHealth technologies. J Med Internet Res 2011 Dec 05;13(4):e111 [FREE Full text] [doi: 10.2196/jmir.1672] [Medline: 22155738]
56. Moizer J, Lean J, Dell'Aquila E, Walsh P, Keary A, O'Byrne D, et al. An approach to evaluating the user experience of serious games. Comput Educ 2019 Jul;136:141-151. [doi: 10.1016/j.compedu.2019.04.006]
57. Pistono AM, dos Santos AM, Baptista RJ, Mamede HS. Framework for adaptive serious games. Comput Appl Eng Educ 2024 Mar 07;32(4):e22731. [doi: 10.1002/cae.22731]
58. Karavidas L, Apostolidis H, Tsiatsos T. Usability evaluation of an adaptive serious game prototype based on affective feedback. Information 2022 Sep 08;13(9):425. [doi: 10.3390/info13090425]

Abbreviations

EER: educational escape room
eMH: e-Mental health
GGzE: Geestelijke Gezondheidszorg Eindhoven en de Kempen
SDT: self-determination theory
SMI: severe mental illness
VR: virtual reality

Edited by A Mavragani; submitted 21.03.24; peer-reviewed by A Frey, IK Singgih, A Ferworn, E Vallefuoco; comments to author 10.05.24; revised version received 22.08.24; accepted 21.10.24; published 07.01.25.

Please cite as:
Bierbooms JJPA, Sluis-Thiescheffer WRJW, Feijt MA, Bongers IMB
Co-Design of an Escape Room for e-Mental Health Training of Mental Health Care Professionals: Research Through Design Study
JMIR Form Res 2025;9:e58650
URL: https://formative.jmir.org/2025/1/e58650
doi:10.2196/58650
PMID:

©Joyce J P A Bierbooms, Wouter R J W Sluis-Thiescheffer, Milou Anne Feijt, Inge M B Bongers. Originally published in JMIR Formative Research (https://formative.jmir.org), 07.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Measuring Bound Attention During Complex Liver Surgery Planning: Feasibility Study

Tim Schneider[1*], MSc; Timur Cetin[1*], Dr rer nat; Stefan Uppenkamp[2], PD, Dr rer nat; Dirk Weyhe[1], Prof Dr Med; Thomas Muender[3], Dr-Ing; Anke V Reinschluessel[3,4], Dr-Ing; Daniela Salzmann[1], Dr phil; Verena Uslar[1], Dr rer nat

[1]University Hospital for Visceral Surgery, PIUS-Hospital, Department for Human Medicine, Faculty VI, University of Oldenburg, Oldenburg, Germany

[2]Department of Medical Physics and Acoustics, Faculty VI, University of Oldenburg, Oldenburg, Germany

[3]Digital Media Lab, Department of Mathematics and Computer Science, Faculty III, University of Bremen, Bremen, Germany

[4]Department of Computer and Information Science, University of Konstanz, Konstanz, Germany

[*]these authors contributed equally

Corresponding Author:
Tim Schneider, MSc
University Hospital for Visceral Surgery, PIUS-Hospital
Department for Human Medicine, Faculty VI
University of Oldenburg
Ammerländer Heerstraße 114-118
Oldenburg, 26129
Germany
Phone: 49 4412292922
Email: tim.schneider@uol.de

Abstract

Background: The integration of advanced technologies such as augmented reality (AR) and virtual reality (VR) into surgical procedures has garnered significant attention. However, the introduction of these innovations requires thorough evaluation in the context of human-machine interaction. Despite their potential benefits, new technologies can complicate surgical tasks and increase the cognitive load on surgeons, potentially offsetting their intended advantages. It is crucial to evaluate these technologies not only for their functional improvements but also for their impact on the surgeon's workload in clinical settings. A surgical team today must increasingly navigate advanced technologies such as AR and VR, aiming to reduce surgical trauma and enhance patient safety. However, each innovation needs to be evaluated in terms of human-machine interaction. Even if an innovation appears to bring advancements to the field it is applied in, it may complicate the work and increase the surgeon's workload rather than benefiting the surgeon.

Objective: This study aims to establish a method for objectively determining the additional workload generated using AR or VR glasses in a clinical context for the first time.

Methods: Electroencephalography (EEG) signals were recorded using a passive auditory oddball paradigm while 9 participants performed surgical planning for liver resection across 3 different conditions: (1) using AR glasses, (2) VR glasses, and (3) the conventional planning software on a computer.

Results: The electrophysiological results, that is, the potentials evoked by the auditory stimulus, were compared with the subjectively perceived stress of the participants, as determined by the National Aeronautics and Space Administration-Task Load Index (NASA-TLX) questionnaire. The AR condition had the highest scores for mental demand (median 75, IQR 70-85), effort (median 55, IQR 30-65), and frustration (median 40, IQR 15-75) compared with the VR and PC conditions. The analysis of the EEG revealed a trend toward a lower amplitude of the N1 component as well as for the P3 component at the central electrodes in the AR condition, suggesting a higher workload for participants when using AR glasses. In addition, EEG components in the VR condition did not reveal any noticeable differences compared with the EEG components in the conventional planning condition. For the P1 component, the VR condition elicited significantly earlier latencies at the Fz electrode (mean 75.3 ms, SD 25.8 ms) compared with the PC condition (mean 99.4 ms, SD 28.6 ms).

Conclusions: The results suggest a lower stress level when using VR glasses compared with AR glasses, likely due to the 3D visualization of the liver model. Additionally, the alignment between subjectively determined results and objectively determined results confirms the validity of the study design applied in this research.

XSL•FO
RenderX

(JMIR Form Res 2025;9:e62740) doi:10.2196/62740

KEYWORDS

workload measurement; virtual reality; VR; augmented reality; AR; electroencephalography; EEG; event-related potential; ERP; auditory evoked potential; AEP; oddball experiment; National Aeronautics and Space Administration Task Load Index; NASA-TLX; surgical planning

Introduction

In modern surgery, new techniques are increasingly being adopted. One example of such innovative technologies is minimally invasive procedures, which aim to reduce surgical trauma and enhance patient safety [1]. However, while these techniques benefit the patient, they pose significant challenges—both cognitively and physically—for the surgical team, particularly the surgeon, due to limited visibility and increased physical strain. This, in turn, may jeopardize patient safety [2].

Another example of novel technologies applied in surgery is the use of augmented reality (AR) and virtual reality (VR) simulations. These simulations display 3D models of anatomical structures on a 3D screen, potentially aiding in their examination and analysis. Glasses that project these simulations can be used during the preoperative phase to assist with surgical planning.

To incorporate technologies such as AR and VR applications into everyday clinical practice or the operating room, these technologies must first be evaluated for safety, ergonomics, and particularly for their impact on workplace strain [3]. So far, AR and VR glasses have primarily been assessed for mental workload using subjective questionnaires, such as the National Aeronautics and Space Administration-Task Load Index (NASA-TLX), or by measuring the user's performance quality [4]. This study aims to assess mental workload both objectively, through electrophysiological recordings, and subjectively, using established questionnaires, to facilitate a comparison between the 2. The study is not intended to evaluate or rate the different techniques used.

From a neuroergonomic perspective, replacing subjective questionnaires with objective measurements offers distinct advantages. Subjective questionnaires rely on self-reporting, which can be influenced by various biases. By contrast, objective measurements provide more accurate and reliable data by directly assessing neurophysiological and behavioral responses in real time [5]. By minimizing reliance on subjective questionnaires, objective measurements enhance the scientific validity of neuroergonomic research, resulting in more robust findings and practical implications that could optimize human-machine interactions.

AR and VR glasses enable more accurate planning of invasive procedures through 3D visualization of anatomical structures. This allows the medical team to anticipate potential anomalies and develop a tailored plan for complex surgeries that considers the patient's individual anatomy more precisely than traditional preparation using conventional radiological images. Such technologies are particularly valuable in challenging interventions with limited direct visibility, such as laparoscopic procedures [6]. Additionally, VR and AR can serve as effective training tools for inexperienced surgeons, helping them gain proficiency and improve safety during interventions.

Several studies have subjectively evaluated the reduction of stress and cognitive burden during the use of VR and AR [7-9]. The results suggest that VR and AR can alleviate stress and cognitive load for users. These applications have been used to clarify anatomical conditions and facilitate the learning of practical tasks, thus reducing stress by providing presurgical training. Building on these findings, we hypothesized that in this study, using AR and VR simulations would be less taxing for participants when preparing for surgery, compared with the traditional method of using classical radiological images such as magnetic resonance imaging (MRI), CT, and X-rays.

Ghani et al [10] demonstrated that a passive oddball condition can be used to objectively assess mental workload through participants' electroencephalography (EEG) [10]. By combining subjective questionnaires with the objective evaluation of workload via EEG, we hypothesize that in this study, the objective results on mental workload will align with the subjective data from the NASA-TLX questionnaire [11].

Bernhardt et al [6], Nicolau et al [12], and Dixon et al [13] described significantly improved visualization capabilities through the use of AR and VR, particularly in minimally invasive procedures. Based on these findings, we hypothesize that AR and VR will facilitate mental visualization of anatomical conditions more effectively than traditional 2D images, due to their 3D displays, ultimately leading to better outcomes, such as improved surgical performance.

To test the following 3 hypotheses, a passive oddball experiment with continuous EEG monitoring was conducted with 9 visceral surgeons as participants. They performed mock preoperative planning for the resection of a cancerous liver tumor using AR and VR headsets, as well as traditional radiological images (MRI and computed tomography scans, resolution: 5 mm; arterial and venous phases available) displayed on a standard computer screen.

- Hypothesis 1: AR and VR simulations will reduce mental workload compared with traditional radiological images.
- Hypothesis 2: The mental workload measured objectively using EEG will match with subjective reports from the NASA-TLX questionnaire.
- Hypothesis 3: AR and VR use will improve anatomical visualization and surgical performance compared with traditional 2D imaging.

To successfully test hypothesis 1, we are using 2 real, anonymized surgical cases that have been adapted for display in traditional radiological, VR, and AR environments.

To evaluate hypothesis 2, a state-of-the-art mobile EEG setup is used on the same cases, after which participants are asked to

XSL•FO
RenderX

complete a commonly used, validated questionnaire on workload strain.

To evaluate performance in hypothesis 3, participants were required to assess the resectability of the tumor in each case, considering both its size and location to make an informed decision regarding surgical feasibility.

These hypotheses require both objective and subjective workload assessments, along with a comparative evaluation of task performance under different imaging conditions.

The design of the EEG task follows a standard passive auditory oddball paradigm, in which both a frequent and a rare auditory stimulus are presented. The EEG response to the target stimulus is analyzed, as it is expected to produce 2 distinct potentials: in addition to the early neuronal responses to the auditory stimulus (P1, N1, P2, and N2, collectively referred to as auditory evoked potentials [AEP]), a later response to the oddball modality—that is, the shift from a frequent to a rare stimulus—is expected to manifest as the P3 potential (hereafter referred to as event-related potential [ERP]). While the early AEPs are associated with basic neuronal processing, primarily reflecting the activity of working memory, the later ERPs are linked to more complex, higher-order cognitive processing of the presented stimulus [14].

The degree of these potentials is used to draw conclusions about the objective workload, specifically in terms of the attention required for a primary task. While the auditory stimuli are presented passively in the background, participants focus actively on a primary task—in our case, planning a surgery in different modalities. It has been demonstrated multiple times that EEG potentials are more pronounced when the primary task is less demanding, that is, requiring fewer neuronal resources: when the primary task is not resource-intensive, more mental resources are allocated to the passive background task, leading to higher EEG potentials for both AEPs and ERPs. If the primary task is more complex, fewer mental resources are allocated to the secondary, passive background task, resulting in less pronounced EEG potentials [10,14].

The objective workload in the 3 different conditions will be compared with the subjective workload measured using the NASA-TLX questionnaire. Additionally, the surgeons will be asked to perform simple tasks during the surgical planning process, such as measuring the diameter of lesions, to assess the quality of their work.

Methods

Ethical Considerations

All participants provided informed consent before the study. Participation was voluntary, and no form of compensation was offered. The study was approved by the Medical Ethics Committee (reference number 2020-150), the Data Protection and Information Security Management Unit at the University of Oldenburg, and the staff council at PIUS-Hospital. All participants were recruited through word of mouth by the medical staff of the University of Oldenburg's Department of Visceral Surgery, located at PIUS-Hospital, Oldenburg. The study was conducted in accordance with the Declaration of Helsinki and the guidelines of Good Clinical Practice. Data were treated confidentially at all times. The database, containing anonymized patient data, was stored on a password-protected computer, ensuring access to authorized personnel only. Data were stored in a pseudonymized manner, with each participant assigned a numerical code.

Participants

Nine participants with prior experience in liver surgery were recruited for this study. Although the sample size was relatively small, all participants were carefully selected from a well-defined target group—surgeons with experience in liver surgery—highlighting the study's quality. Inclusion criteria included attending status (or residents in their fourth year of training), signed informed consent, no significant coffee consumption, and no history of neurological damage. Exclusion criteria for participants included excessive caffeine consumption, the use of psychoactive drugs or medications, and any mental or neurological disorders. These criteria were assessed in advance using a standard questionnaire for EEG measurements.

All participants were male, right-handed individuals with an average age of 36.8 (SD 5.59) years. Six of the participating surgeons were attending surgeons (including 2 senior surgeons), and 3 were residents in the later stages of their training. On average, before the study, participants had assisted in an average of 43 (SD 65.63) liver surgeries and had led an average of 16 (SD 29.7) surgeries.

Stimuli

The acoustic oddball stimuli consisted of 2 sinusoidal tone pips, each with a duration of 50 ms and a rise/fall time of 5 ms. The standard tone, with a frequency of 500 Hz, had a probability of 88%, while the rarer target tone, at a frequency of 1 kHz, had a probability of 12%. Both tones were presented via loudspeakers (iLoud MTM; IK Multimedia), connected to an amplifier (the t.amp E4-130; Thomann), at a volume of 75 dB sound pressure level, with an interstimulus interval of 1 second. The tones were randomized with the constraint that a target tone was always followed by a standard tone. The stimuli were similar to those used by Williams et al [15]. The auditory stimuli were digitally created using MATLAB (version R2021a; The MathWorks, Inc.) and presented using Presentation Software (version 23.1; Neurobehavioral Systems).

Conditions

Three different conditions were tested to assess their impact on focused attention during liver surgery planning. These conditions were (1) an AR headset, (2) a VR headset, and (3) a PC with traditional radiological images. The VR setup included the VIVE Pro headset (HTC Corporation), 1 controller, and 2 tracking devices, known as lighthouses. For AR, the Microsoft HoloLens 2 was used (Microsoft Corp.). The software used for visualization in VR was a custom-built setup previously used in our department [16]. For the AR setup, a demonstrator developed as part of this project was used. Traditional radiological images were displayed using the Xero Viewer (Agfa Health care) on a conventional PC. Screenshots of the Xero Viewer software and videos of the VR and AR conditions are

XSL•FO
RenderX

available in Multimedia Appendices 1-4, respectively. In the AR and VR conditions, participants were presented with previously segmented 3D models in combination with radiological images. In the PC condition, participants were only presented with the radiological images. In all 3 conditions, participants were instructed to plan a liver resection using all available functions specific to each condition. In the AR and VR conditions, participants could draw on the 3D model, set markers, and manually measure anatomical structures. In the VR condition, additional functionalities were available, such as volume cropping and the placement of vascular clips. In both the AR and VR conditions, participants could also view or superimpose the corresponding MRI images onto the 3D model. The control mechanisms differed between the 2 conditions due to distinct hardware setups: the VR condition was controlled using dedicated controllers, while the AR condition relied on gesture-based interaction. In all 3 conditions, participants were asked to assess the size and segment of the largest tumor, as well as the resectability of the tumor. The primary goal was to keep participants engaged for at least 3 minutes in each condition, using a basic task (ie, diagnosing based on the images). This task was consistent across all conditions, regardless of the specific interface or available functionalities. Evaluating the usability of the respective software and its functionalities was beyond the scope of this study, particularly because both the AR and VR setups were custom-built and still under development.

Experimental Setup

A mobile EEG amplifier (Smarting mobile; mbt) was connected to a standard 10/20 EEG cap (EasyCap GmbH) and recorded using software (Smarting Streamer v3.4.2; mbt) on a desktop computer running Windows 10 (Microsoft Corp.). The 2 speakers were placed next to the recording computer, with participants positioned 2 m in front of the speakers. During the experiment, care was taken to position the participants consistently in the same marked location across all 3 conditions to ensure that the stimuli were presented at the same sound level in each condition. A laptop was used to connect wirelessly to the AR glasses via Miracast, allowing the instructor to monitor what the participants were doing and seeing. The laptop was placed out of sight of the participants to avoid distractions. A second desktop computer was used to run the VR setup, with the VR headset connected to this computer and the 2 VR lighthouses mounted on tripods, enabling the participants to move freely while wearing the VR headset. To ensure that participants remained in the same position during all 3 conditions, the traditional radiological images presented in the PC condition were displayed on a computer connected to the hospital information system. This computer was placed on a mobile, height-adjustable table, with the screen positioned at eye level. This setup was designed to prevent the EEG cap from slipping due to excessive head movement and to allow participants to perform all conditions while standing (Figure 1). Controllers were used to operate the VR condition and were placed in the participants' hands. In the AR condition, control was achieved through gesture-based interaction using the participants' hands. No standardized training was provided before the experiment. However, if the conditions were unfamiliar to the participants or if they had any questions, a test run was conducted in which they could familiarize themselves with the controls and functions until they felt confident in handling the technology.

Figure 1. Experimental setup. 1: main computer for recording electroencephalography (EEG); 2: amplifier; 3: computer virtual reality (VR); 4: computer augmented reality (AR); 5: lighthouses for the VR headset; 6: VR headset; 7: EEG on marked position; 8: computer for conventional surgical planning; and 9: AR headset.

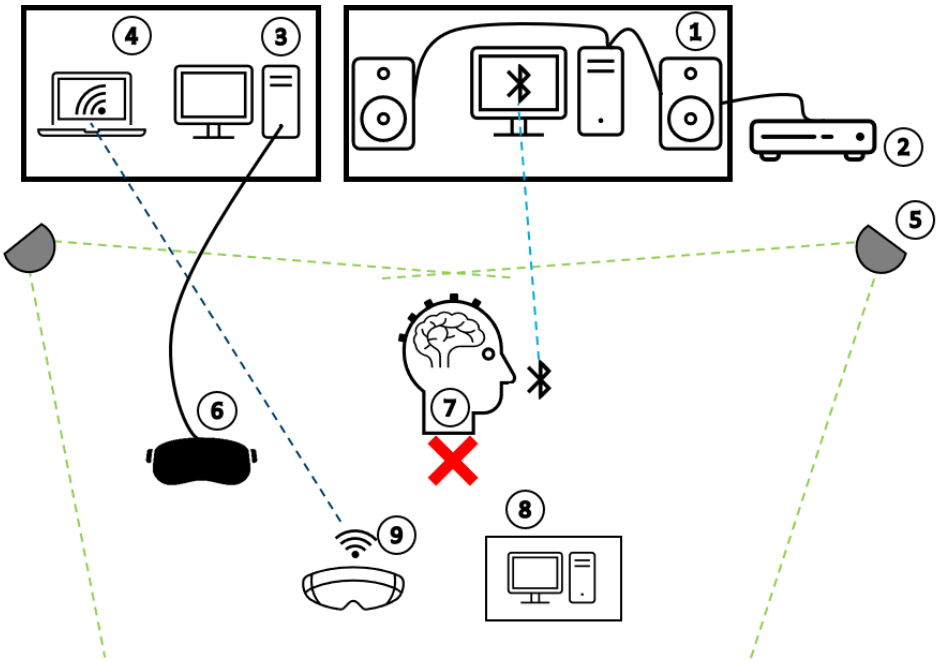

XSL•FO
RenderX

Procedure

Mobile EEG recordings (sampling rate: 250 Hz, Smarting Mobi; mbt) were obtained using a standard EasyCap EEG recording cap with 24 electrodes: Fp1, Fp2, AFz, F7, F3, Fz, F4, F8, T7, C3, Cz, C4, T8, CPz, M1, M2, P7, P3, Pz, P4, P8, POz, O1, and O2. The reference electrode was placed at a fronto-central position. Continuous EEG data were stored in .xdf format before being processed for analysis. For AEP analysis, EEG data were analyzed using EEGlab (V2020.0, open-source toolbox for MATLAB R2020a; MathWorks) on a standard PC running Windows 10 (Microsoft Corp.). For ERP analysis, EEG data were first converted to Brain Vision Analyzer format using MATLAB and then analyzed using Brain Vision Analyzer Software (version 2.2.1.8266, Professional Edition; Brain Products GmbH). EEG electrode impedances were adjusted to below 10 kΩ using 70% rubbing alcohol and abrasive conductive gel (EasyCap Abalyte Hi-Cl), following the manufacturer's recommended protocol. After the electrodes were properly placed, 2 minutes of resting-state EEG was recorded with the eyes open and eyes closed before actual testing began. For the testing, 1 of the 3 conditions—"AR," "VR," or "radiological images"—was selected in randomized order. In each of the 3 conditions, participants were required to plan 2 different surgeries, which included measuring the dimensions of the tumors and identifying their anatomical locations within the liver to determine whether the tumors were resectable. The 2 patient cases were presented in randomized order (see Figure 2 for the experimental procedure). By the end of the session, each participant had completed 6 sets of surgical planning (2 cases for each of the 3 conditions). The task for the participants was to plan a surgical procedure to remove cancerous tissue from the liver. The processing time for each condition (AR, VR, or radiology) was 10 minutes, with 5 minutes allocated to each patient case. Throughout the task, auditory stimuli were presented to the participants via 2 speakers in the background. Before switching to the next condition, the participants' subjective workload was assessed using the NASA-TLX questionnaire.

Figure 2. Experimental procedure flowchart: randomized order of conditions. AR: augmented reality; NASA-TLX: National Aeronautics and Space Administration-Task Load Index; VR: virtual reality.

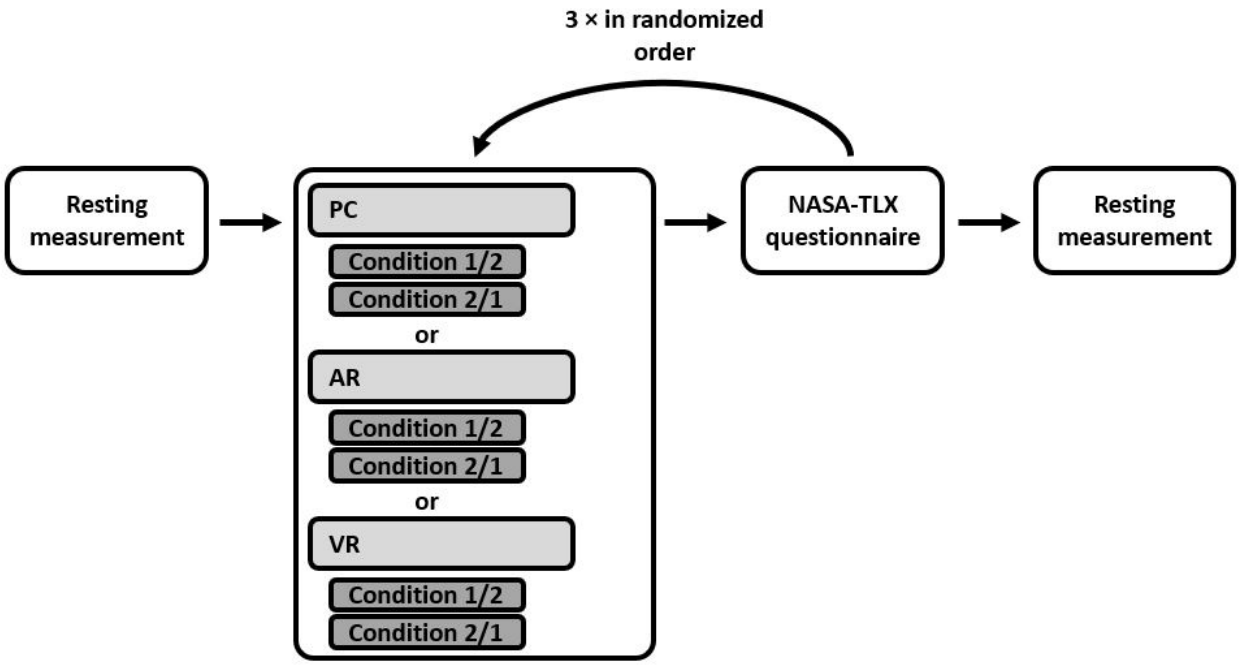

NASA-TLX

The NASA-TLX workload questionnaire has been validated as an effective measure of subjective workload in medical work environments [17].

The NASA-TLX is a widely used tool for assessing subjective workload. Originally developed for aviation settings by NASA, it has since been adapted for use in various fields, including health care [17]. The NASA-TLX measures 6 dimensions of workload: mental demand, physical demand, temporal demand, effort, frustration, and perceived performance. Each dimension is rated on a scale from 0 (low) to 100 (high) [18].

Data Preprocessing of EEG Data

To retain the frequency components relevant for EEG analysis, the recordings were filtered using a bandpass filter (low cutoff: 0.05 Hz, high cutoff: 30 Hz, and filter order: 4). A notch filter was also applied at 50 Hz to remove power line interference. The continuous EEG data were divided into 1000 ms epochs, with 200 ms before and 800 ms after the stimulus. The data were then categorized into 2 categories: standard stimulus (nontarget: 0.5 kHz) and target stimulus (1 kHz). Automatic artifact rejection was performed by discarding all epochs where fluctuations greater than +100 µV or –100 µV occurred within 100 ms. Additionally, all epochs underwent manual visual inspection, focusing on signals from electrodes FP1 and FP2 to identify strong artifacts due to eye movements. Averages were

XSL•FO
RenderX

then calculated for the epochs recorded during the presentation of the target stimulus in each of the 3 conditions: 1 average for the EEG response to the target stimulus during AR, 1 for the EEG response during VR, and 1 for the EEG response during surgical planning using traditional radiological images. Averages were calculated for each participant to facilitate further statistical analyses. Baseline correction was applied to prevent potential linear drifts in the recorded signals and to ensure that the grand averages across participants and conditions were comparable. For AEP analysis, the data were re-referenced to electrodes T5 and T6, while for ERP analysis, the data remained referenced to the physical fronto-central reference.

Results

Subjective Workload

The median physical demand score for the AR condition was 30 (IQR 25-70) on the NASA-TLX scale, higher than the scores for the VR (median 25, IQR 15-45) and PC (median 10, IQR 5-20) conditions. For the perceived performance and temporal demand categories, all 3 conditions yielded the same NASA-TLX scores: 25 for temporal demand and 30 for perceived performance. Additionally, for the mental demand, effort, and frustration categories, the AR condition showed the highest median scores: 75 (IQR 70-85) for mental demand, 55 (IQR 50-75) for effort, and 40 (20-80) for frustration (Figure 3).

Figure 3. National Aeronautics and Space Administration-Task Load Index (NASA-TLX) scores. AR: augmented reality; VR: virtual reality.

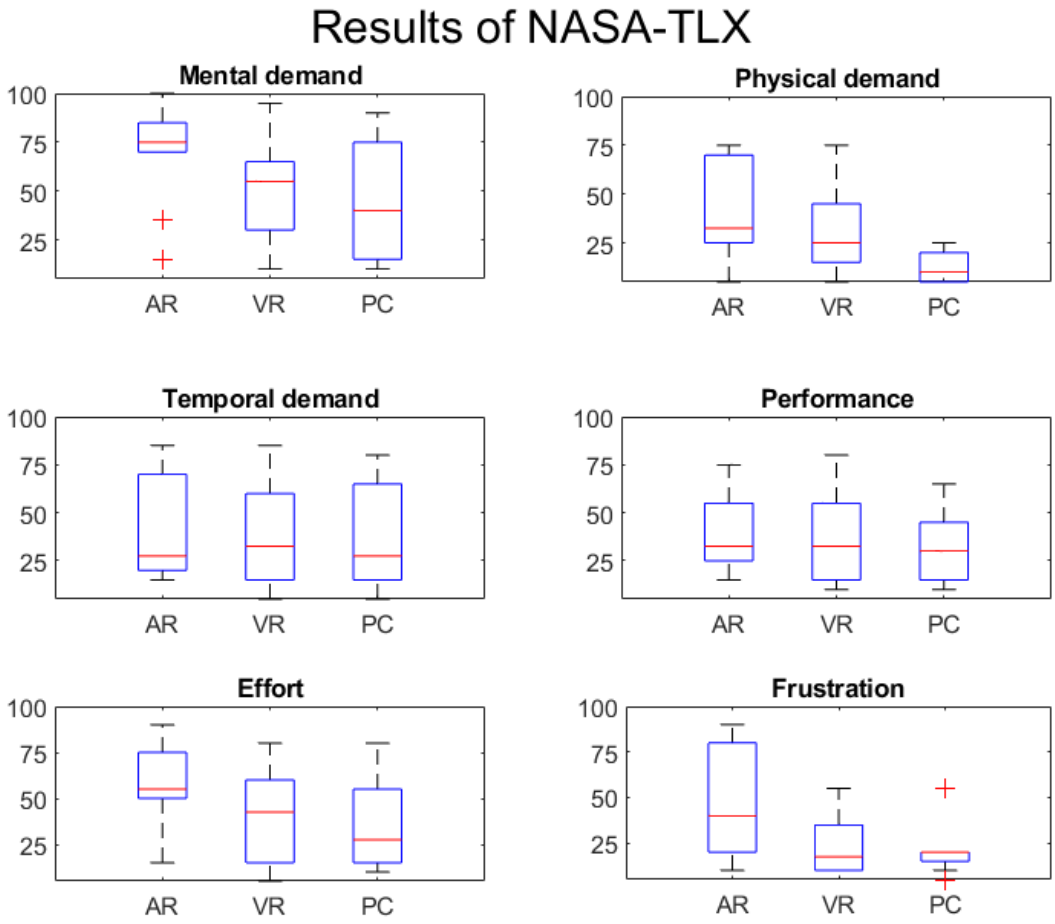

Examination of the box plots reveals that the mental demand score for the AR headset (median 75, IQR 70-85) was higher than those for both the VR headset (median 55, IQR 30-65) and the PC (median 40, IQR 15-75) conditions. Similarly, higher scores for effort and frustration were observed in the NASA-TLX evaluations when using the AR headset (median 55, IQR 50-75 for effort; and median 40, IQR 20-80 for frustration) compared with using the VR headset (median 40, IQR 15-60 for effort; and median 15, IQR 10-35 for frustration) or the PC condition (median 25, IQR 15-55 for effort; and median 20, IQR 15-20 for frustration). In terms of physical demand, the score was lower in the PC condition (median 10, IQR 5-20) compared with the AR (median 30, IQR 25-70) or VR (median 25, IQR 15-45) condition.

A Friedman test was performed to assess significance (significance level of .05), as the data did not follow a normal distribution according to the Kolmogorov-Smirnov test. To account for multiple comparisons, the Benjamini-Hochberg method was applied for correction using a web-based calculator [19]. Assuming a 95% CI, no significant differences (P=.10) were found.

XSL•FO
RenderX

Electrophysiological Recordings—Early Components

AEPs were primarily observed on the electrodes along the central row and midline (AFz to Pz, F3/C3, and F4/P4). Grand averages were calculated for all electrodes (Figure 4).

Figure 4. Event-related potentials as response to the target tone for all electrodes and all 3 conditions.

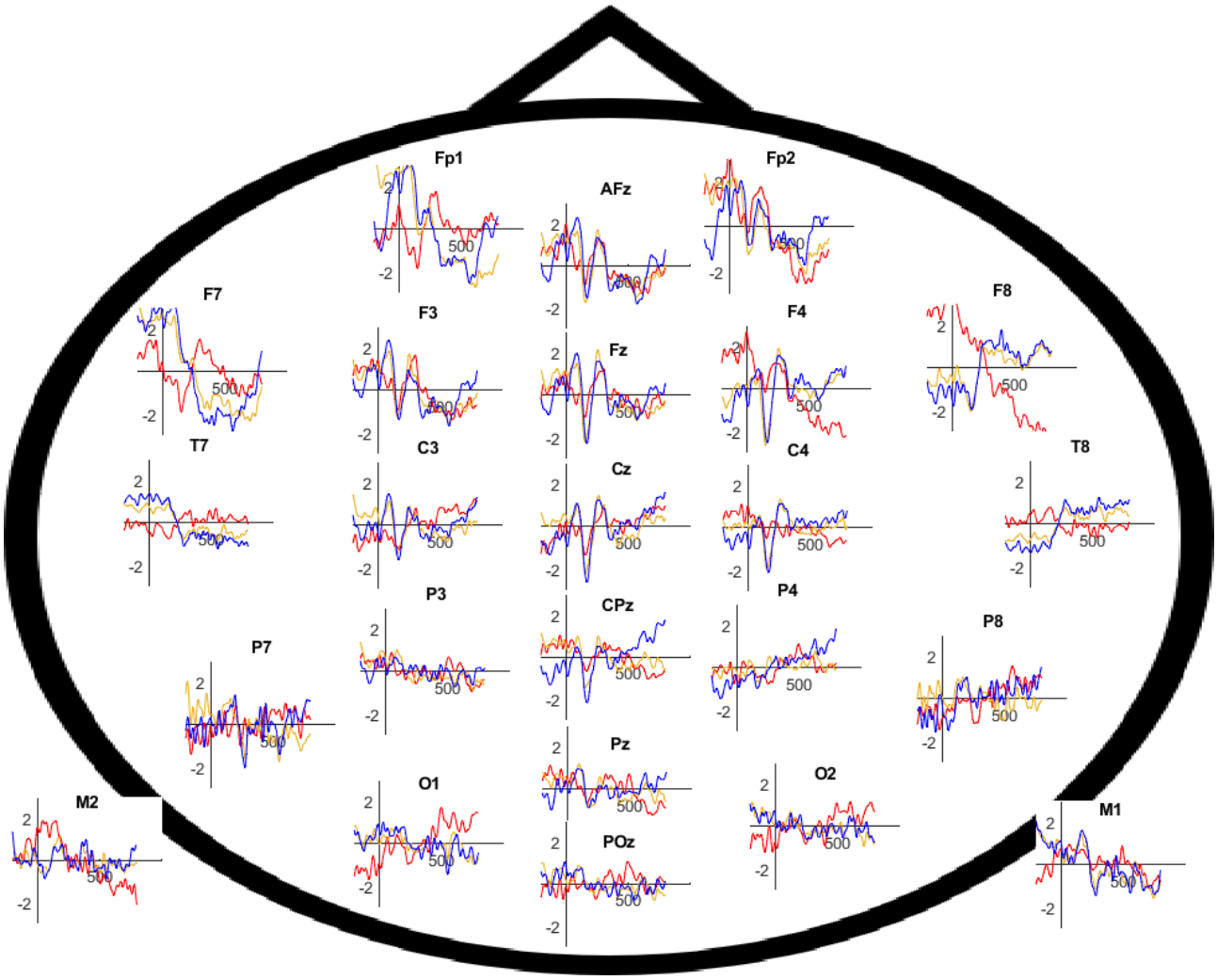

To statistically analyze the amplitudes for each participant, the peaks and their corresponding latencies were identified for all conditions at the central electrodes Fz, Cz, and Pz (Figure 5B-D). To statistically verify differences in the amplitudes of the grand averages, the approach outlined by Handy [20] was used as a guide. The intervals for P1, N1, and P2 were marked accordingly (Figure 5A). The resulting peaks and latencies for each participant were then analyzed using a repeated measures analysis of variance (rmANOVA) in SPSS (version 28.0.1.0; IBM Corp.) with the factors patient case, condition, electrode, and ERP components. A significant difference was observed between the ERP components ($F_{1.718,13.743}$=46.754, P<.001).

XSL•FO
RenderX

Figure 5. Grand averages of event-related potentials as a response to the target tone for statistically analyzed electrodes under all 3 conditions: (A) temporal windows for the event-related components P1, N1, and P2, exemplarily shown for the electrode position Fz; (B) grand averages of the electrode Fz; (C) grand averages of the electrode Cz; (D) grand averages of the electrode Pz. AR: augmented reality; VR: virtual reality.

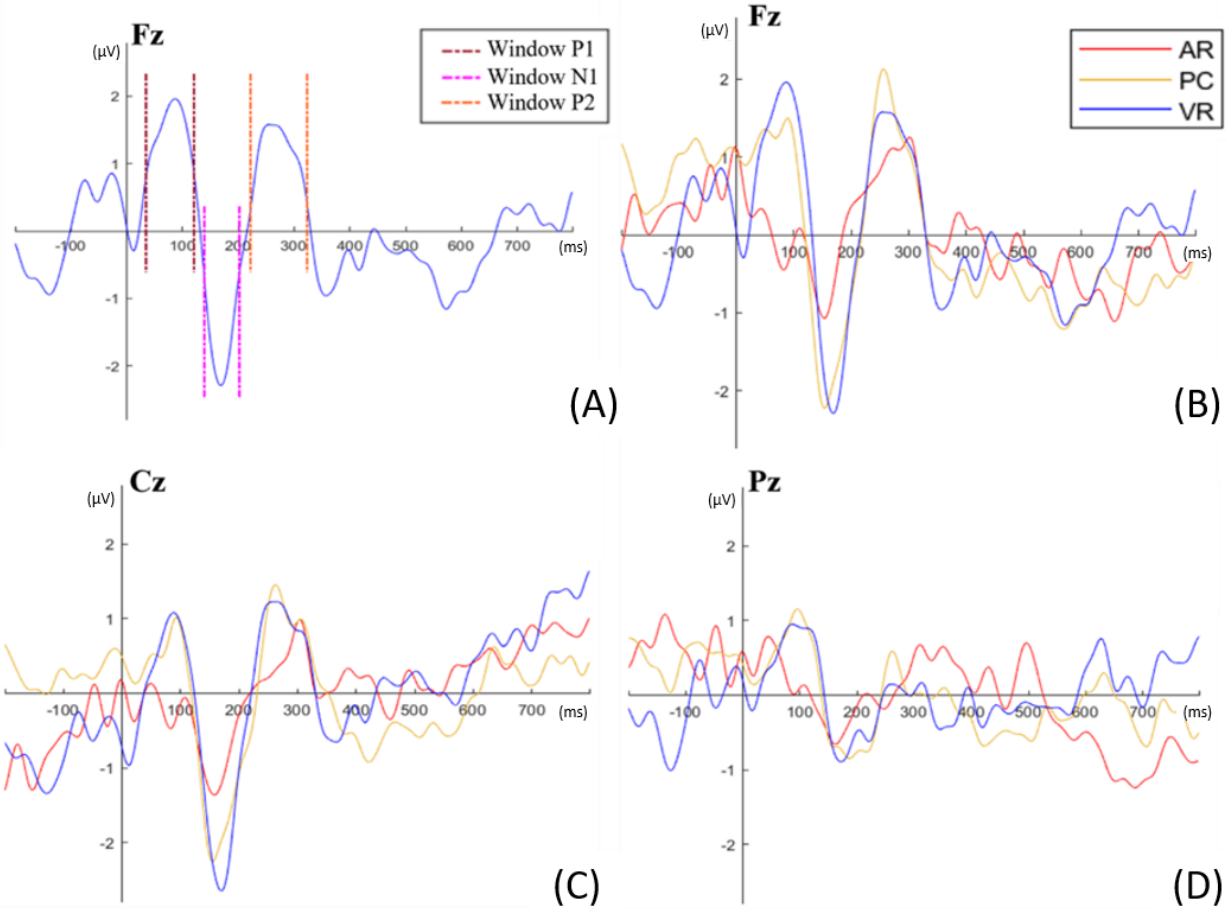

The tendencies observed in the grand averages were not supported by the statistical analysis, as no significant differences were found between patient cases (*P*=.57), conditions (*P*=.71), or electrodes (*P*=.49). To determine the latencies of the AEP components (P1, N1, and P2), the method described by Handy [20] was also applied. The latency of the identified peaks within the defined intervals was measured and recorded for each participant. Subsequently, an rmANOVA was conducted with the factors patient case, condition, electrode, and ERP components to identify significant differences in the temporal occurrence of these components.

The tests of within-subject effects revealed a significant interaction among condition, electrode, and ERP components ($F_{3.401,27.211}$=6.747, *P*=.001). The results of the post hoc pairwise comparisons are presented in Table 1. According to these comparisons, the P1 component at the Fz electrode occurs significantly (*P*=.01) earlier in the VR condition (mean 75.3 ms, SD 25.8 ms) compared with when using traditional radiological images (mean 99.4 ms, SD 28.6 ms). The Cz electrode also demonstrated a significant (*P*=.04) difference in the P1 component between the VR condition (mean 111.1 ms, SD 9.9 ms) and the AR condition (mean 80.4 ms, SD 32.1 ms). Additionally, the P2 component in the AR condition (mean 290.6 ms, SD 23.5 ms) occurred significantly (*P*=.03) later than in the traditional radiological images condition (mean 265.3 ms, SD 13.1 ms). Significant differences were also observed for the P1 component at the Pz electrode, with the VR condition (mean 62.8 ms, SD 24.8 ms) occurring significantly earlier (*P*=.04) than both the AR condition (mean 89.1 ms, SD 14.6 ms) and the traditional radiological images condition (mean 92.8 ms, SD 28.4 ms; *P*=.04). No significant (*P*=.12) interactions were found for the N2 component.

XSL•FO
RenderX

Table 1. Significant results of the post hoc tests (pairwise comparisons).

Electrode and ERP[a] component	Condition (I)	Condition (J)	Latency (ms), mean		Difference between means (I–J)	*P* value
			Condition (I)	Condition (J)		
Fz						
P1	VR[b]	PC[c]	75.3	99.4	–24.1	.01
Cz						
P1	VR	AR[d]	111.11	80.4	30.71	.04
P2	AR	PC	280.7	265.3	25.4	.03
Pz						
P1	VR	AR	62.9	89.1	–26.2	.04
		PC	62.9	92.9	–30	.04

[a]ERP: event-related potential.

[b]VR: surgery planning using the virtual reality setup.

[c]PC: surgery planning using traditional radiological images on a desktop computer.

[d]AR: surgery planning using the augmented reality setup.

Electrophysiological Recordings (P3)

For the P3 component of the ERP, peak amplitudes within the time range of interest (300-400 ms poststimulus) were calculated. A grand average for all 3 conditions (radiology, VR, and AR) was computed using data from all electrodes (Figure 6). A clear trend was observed, with the classical radiological approach eliciting the lowest P3 amplitude (mean 2.45 µV, SD 2.26 µV), VR eliciting the second largest P3 amplitude (mean 3.08 µV, SD 2.80 µV), and the AR condition eliciting the highest P3 amplitude (mean 6.20 µV, SD 6.84 µV). SDs were high, and no statistical significance was found using an rmANOVA in SPSS (version 28.0.1.0) with the factors patient case, condition, and P3 amplitude ($F_{1.94,10.58}$=1.469, P=.26).

XSL•FO
RenderX

Figure 6. Amplitude of the P3 component of the event-related potential across all 3 conditions: the augmented reality (AR) approach (left), which showed the highest P3 amplitude (mean 6.20 µV, SD 6.84 µV); the virtual reality (VR) condition (center), with a mean amplitude of 3.08 µV (SD 2.80 µV); and the classical approach (right), which exhibited the lowest P3 amplitude (mean 2.45 µV, SD 2.26 µV). No statistically significant differences were found ($F_{(1.94,10.58)}$=1.469; P=.26).

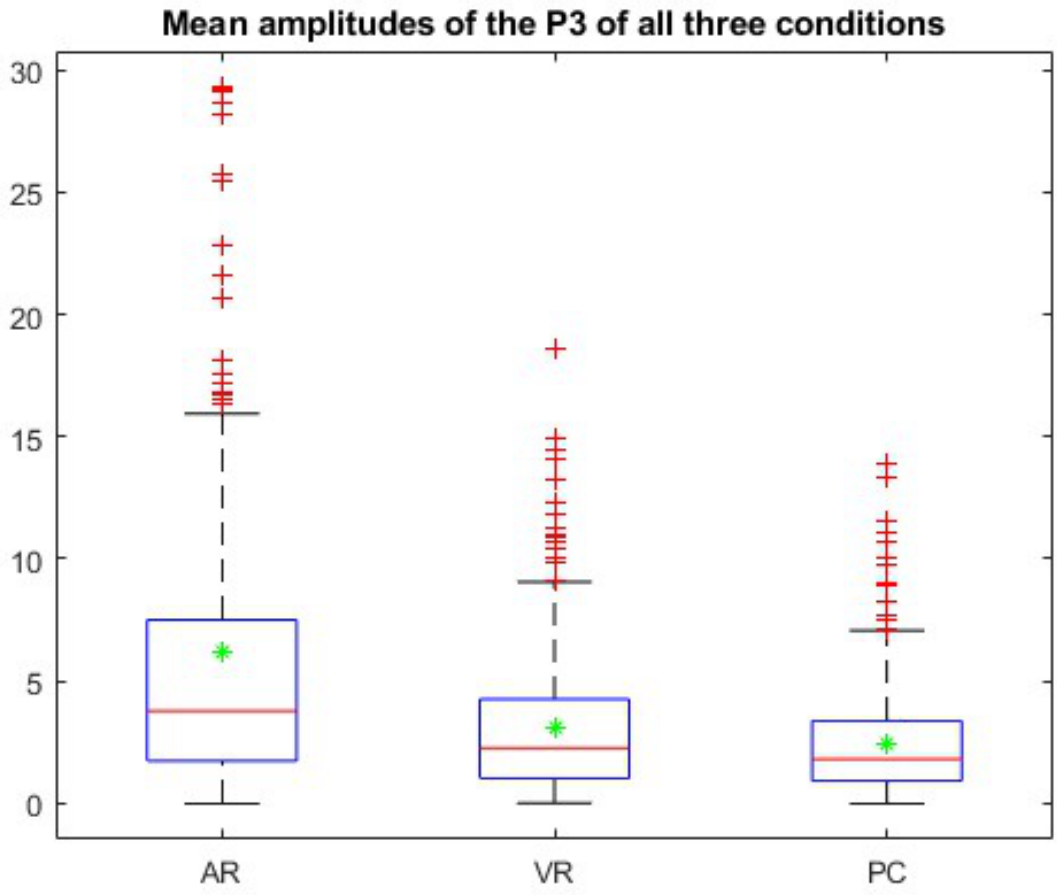

Post Hoc Power Analysis

Given the small sample size (N=9), a post hoc power analysis was conducted to determine the effect size and assess whether a larger sample might yield different results. Partial eta-squared (η^2=0.114) was calculated for the interaction between experimental conditions using SPSS. To estimate the sample size required for significant results, a web-based calculator was utilized [19], indicating that a sample size of 101 participants would be needed to achieve a statistical power of 0.8 at a significance level of α=.05.

Correlation

A correlation analysis was conducted between the N1 component, which demonstrates a decreased amplitude with increased cognitive workload [10], and the "Mental Demand" and "Effort" dimensions of the NASA-TLX questionnaire. The results revealed a correlation coefficient of r=–0.0042 between the N1 component and "Mental Demand," and r=0.066 between the N1 component and "Effort."

Performance

All participants correctly identified the resectability of the tumors in all conditions and for every patient case, based on their accurate assessment of tumor size and localization.

Discussion

Findings and Implications

The aim of this study was to establish a method for objectively determining the additional workload generated using AR or VR glasses in a clinical context for the first time. For this purpose, 2 runs of planning a liver surgery were performed under each of the 3 conditions. In the background, a passive oddball experiment was conducted to produce clear early AEP components (P1, N1, and P2) as well as the P3 component of the ERP in the EEG recordings, guided by the work of Ghani et al [10].

Subjective Workload

NASA-TLX scores indicate a tendency toward higher scores, suggesting a greater subjective workload for AR planning, particularly in terms of "mental demands" and "frustration." However, this trend does not reach statistical significance when compared with VR or the traditional approach using radiological images, likely due to the small sample size. Moreover, the large variances observed in the data could have contributed to the lack of significant findings, as they indicate substantial variability in subjective workload perceptions among participants. The observed tendencies may be attributed to the

XSL•FO
RenderX

more complex and unfamiliar gesture controls of the AR headset. In addition, a poor internet connection may have resulted in higher latencies, potentially affecting gesture control and thus leading to higher workloads when using the AR demonstrator.

Higher physical demands were expected in the AR and VR setups compared with the traditional approach. A tendency was observed in the subjective ratings, as using a stationary mouse on a desktop PC required significantly less physical effort compared with using an AR or VR headset.

Objective Workload

The expectation that using an AR or VR setup would generally make planning surgery easier due to its 3D projection of the anatomical situation was not confirmed.

EEG data indicate a trend with more pronounced early AEPs when using VR and classical radiological images compared with AR, suggesting that AR results in the highest workload for the surgeon, while VR does not increase workload compared with classical radiological images. Differences are especially visible in the ERPs P1 and N1 components: no P1 component was found for the AR condition, whereas the N1 component showed a significant difference for the AR condition compared with radiological images and VR on the level of grand averages. This aligns with the finding of Ghani et al [10], indicating that the N1 component becomes less pronounced with increased task difficulty due to the more complicated gesture control of the AR headset and limited opportunity to familiarize with the gestures required to handle 3D objects in AR.

We would like to emphasize that the significance at the electrode Pz should be treated with caution, as the amplitudes here are generally lower. The results for the electrode Pz are clearly noisier, likely due to its position being obstructed by the fastening of the glasses, which are attached to the head near the position of electrode Pz. These results should therefore be interpreted with caution.

When examining the significances of the latencies of the different early AEP components, it is noticeable that there are occasional significances that do not follow a coherent pattern.

Concerning the event-related P3 wave in response to the target stimulus of the passive oddball paradigm, the pattern found differs from that seen in the early potentials: a trend from a low P3 in the radiology condition to a high P3 in the AR condition can be observed. The studies conducted by Thees et al [7], Küçük et al [8], and Chao et al [9], which asserted that AR and VR headsets result in reduced cognitive load, can thus only be partly substantiated in this investigation.

Based on our data, we conclude that the use of the passive oddball paradigm is a suitable tool to measure the objective workload between the radiology, AR, and VR setups in planning a complex surgery. There appear to be measurable differences in the neuronal processing of 2D screen images when compared with VR and AR images. However, the answer to why the different modalities of EEG components—early AEP versus ERP—yielded different findings is complex and likely rooted in distinct, complex neuronal and psychological processes responsible for producing these potentials. To better understand the different psychological processes behind the processing of 2D screen images versus VR and AR images, in addition to increasing the number of participants, a more in-depth analysis of the P3 potential is needed. In our analysis, we only looked at the grand average of the peak amplitude. However, it is also important to investigate the location of the most pronounced P3 potentials, as well as the dynamics—that is, the distribution of P3 peak latencies across different electrode localizations and how this distribution changes over time. Current research indicates that the P3 potential arises as a combination of different subcomponents, P3a and P3b, which may be involved in distinct neuropsychological processes. This differentiation needs to be addressed in further studies.

A simple but feasible explanation for the different findings between early AEPs and P3-ERPs could be attributed to the different mental processes involved: while the complex cognitive processing of AR pictures, represented by the P3 potential, might be less demanding for the user, aligning with findings from Thees et al [7], Küçük et al [8], and Chao et al [9], the difficult or less familiar navigation in the AR—and to a lesser extent, VR—environment may lead to higher demands on basic neuronal tasks, such as visuomotor coordination, which are more strongly represented by early neurological potentials than by late cognitive potentials. While this remains speculative at this point, we believe that our study should be regarded as a proof-of-principle, demonstrating a foundation for further research with a larger sample size and in more complex settings. This will allow for a deeper investigation into this issue.

The lack of a significant correlation between the N1 component and the subjective measures of mental demand and effort is likely influenced by the extremely small sample size in this study. However, this result also suggests that cognitive workload, as measured by physiological markers such as EEG, may not necessarily align with the subjectively perceived workload. This finding underscores the importance of using objective measures in assessing workplace stress and cognitive load. Relying solely on subjective assessments may not fully capture the cognitive demands placed on individuals, highlighting the need for more robust and comprehensive approaches in future research.

The use of a mobile EEG was particularly useful in this study. The high degree of mobility provided by the use of portable smartphones as recording devices allowed for recordings to be made in situations closely aligned with the operation without restricting participants' freedom of action. The validity of corresponding mobile EEGs has been previously described by Wascher et al [21]. Moreover, the work of El Basbasse et al [22], among others, has confirmed that reliable data can be derived using mobile EEG recordings, including the implementation of the EasyCap system in conjunction with VR headsets [21,22]. The results of this study support this research. Additionally, our findings demonstrate that the study design, when combined with mobile EEG technology, enables the investigation of the additional workload induced by various medical and clinical devices. Notably, even head-worn devices cannot be unequivocally excluded from this measurement approach.

XSL•FO
RenderX

The AR setup was perceived and measured as the most difficult of the 3 setups. As mentioned, this may be due to the complexity of its controls. While the classical approach was well-known to the participants, controlling the VR setup appeared to be more intuitive than using the AR setup. However, it is crucial to consider that AR headsets represent a relatively nascent technology, and gesture recognition is still in its developmental stages, with anticipated advancements in future iterations of AR headsets. It appeared that participants were overwhelmed by handling the headset without a controller or corresponding control element, as gesture control was a new experience for most of them. This difficulty may be overcome by more accurate training of users before working with the AR setup. We believe that our work could be used as a tool to assess how well users are trained in using the AR setup and may be used to generally investigate the need for training when navigating in an AR and VR environment, even beyond its medical applications. Additionally, more complex, subconscious explanations could be considered: The VR approach might benefit from the possibility of exclusively displaying the anatomical model, while the AR setup might strain participants' mental capabilities more due to the simultaneous display of the computer-generated 3D model along with the real-world environment. Furthermore, factors such as the different displays of the HoloLens and the VIVE Pro could contribute to the differences in participants' workload. However, this requires more detailed investigations. It must be considered in this discussion that the 2D viewing habits trained over decades are completely abandoned. New 3D viewing habits must be relearned in AR and VR, along with the integration of voice and gesture control.

This study makes a significant contribution to the existing body of research by specifically focusing on the mental workload of medical professionals. Unlike the studies listed in Table 2, which mainly explore cognitive load in nonmedical settings or involve participants lacking domain-specific expertise, our research is uniquely positioned within the context of clinical practice. By using objective EEG measurements, we were able to assess workplace strain in real-world scenarios, providing insights into the cognitive demands experienced by surgeons when using AR and VR technologies. This focus on medical professionals, coupled with the use of advanced neuroergonomic methods, highlights the practical relevance and novelty of our findings. Our study not only expands the understanding of how AR and VR technologies impact mental workload in high-stakes environments but also provides valuable data that could inform the optimization of these technologies for clinical use.

XSL•FO
RenderX

Table 2. Comparison of recent research studies investigating mental workload using EEG[a] measurements in AR[b] and VR[c] applications.

Study	Population	EEG application	Key findings
Alessa et al [23]	28 Healthy males (mean age 32.12 years)	EEG power analysis	• AR instructions reduced task time compared with paper instructions, particularly for high-demand tasks. • AR increased mental workload and information processing (germane load).
Binias et al [24]	Pilots in short-haul flight simulator sessions	EEG power analysis	• Positive correlation between theta power (frontal lobe) and reaction time. • Significant event-related changes in the frontal lobe for theta and beta waves.
Ji et al [25]	Pilots in VR-simulated missions	Focusing on β rhythms	• β rhythm changes correlated with high-difficulty tasks. • A significant relationship between mental workload and β rhythms.
Aksoy et al [26]	General population	Visual event–related potentials such as N1, P1, and P3	• ERP[d] waveforms were similar between VR HMD[e] and CS[f] • P3 amplitudes were comparable between VR HMD and CS. • N1 amplitudes were significantly higher in VR HMD compared with CS. • P1 amplitudes were higher at the occipital region in VR HMD.
Giorgi et al [27]	General population	Global field power, alpha, and theta	• A higher Workload Index was observed in the museum compared with VR. • Overall, the museum showed higher engagement potential, while VR potentially supported greater embodiment.
Marucci et al [28]	General population	P300 potentials with respect to the unimodal (visual) and visual and tactile bimodal stimulation	• Multisensory stimuli significantly improved performance and reduced EEG-based workload only in high load conditions. • Multimodal stimulation (visual-audio-tactile) decreased latency and increased P300 amplitude, indicating faster and more effective stimulus processing.
Atici-Ulusu et al [29]	Employees on an automobile assembly line	EEG and NASA-TLX were used to measure cognitive load	• Cognitive load (measured by EEG) was lower with AR glasses. • NASA-TLX scores decreased by 10% when using AR glasses, indicating reduced cognitive load. • No significant difference in cognitive load between the first and last days of AR glasses usage.
Trupti et al [30]	General population	EEG signals were analyzed using autoregressive modeling, fuzzy c-means clustering, hidden Markov model, and deep recurrent neural network classification	• The proposed method for classifying cognitive workload using EEG signals achieved 97.8% accuracy. • Handcrafted features combined with temporal dynamics modeling outperformed state-of-the-art methods. • The optimal performance was achieved with 4 electrodes.
Luque et al [31]	General population	Event-related potentials such as P3 and contingent negative variation	• VR effectively mimicked real-world cognitive load and decision-making processes. • Significant ERPs related to attention and decision-making were elicited in VR.

XSL•FO
RenderX

Study	Population	EEG application	Key findings
Mondellini et al [32]	General population	Calculated the MWL index[h]	• Different task complexity levels significantly affected MWL, increasing subjective measures and decreasing performance. • HMD generally resulted in lower EEG-derived MWL, indicating reduced cognitive load for some tasks.

[a]EEG: electroencephalography.
[b]AR: augmented reality.
[c]VR: virtual reality.
[d]ERP: event-related potential.
[e]HMD: head-mounted display.
[f]CS: computer screen.
[g]NASA-TLX: National Aeronautics and Space Administration-Task Load Index.
[h]MWL: mental workload.

To provide an outlook beyond the scope of this study, we believe that AR and VR can be very useful technologies in the clinical context when used for the appropriate tasks. For instance, VR could be more suitable in a teaching setting or for preoperative planning, whereas AR holds immense potential beyond preoperative planning, particularly in the operating room. Its ability to provide an open view of the real environment while maintaining sterility allows for virtual overlays of anatomical structures on a patient's body during surgery, enhancing visualization and accuracy. Whether the significant benefits of these technologies outweigh the potential additional workload for surgeons should be further investigated in future studies.

Furthermore, investigating the impact of training, the continuous development of software and hardware, and the increasing familiarity of the younger generation of physicians with these new technologies on measures of workload would be of great interest. Such research could provide valuable insights into the potential benefits and challenges posed by integrating advanced technologies such as AR and VR in medical practice. Further exploration in this area may also facilitate the design of tailored training programs and technological interventions to optimize the utilization of these innovations and improve overall health care outcomes.

Limitations

Concerning our hypothesis that the use of AR or VR might improve surgical performance, no conclusive results were found. Basic tasks, such as deciding whether a tumor is resectable or not, were performed with the same quality across all experimental conditions. This might indicate that AR and VR technology does not inherently lead to better performance in trained surgeons for simple tasks, but this needs to be evaluated further. Future studies should include more specialized criteria for evaluating performance, particularly to assess finer aspects of clinical decision-making in AR and VR environments. This could involve more complex cases or tasks that better reflect the challenges faced in surgical planning and execution, allowing for a deeper understanding of how these technologies impact surgical performance and outcomes. This was beyond the aim of this study. However, based on our findings, we suggest that increased performance might be found in a more challenging task, but not concerning simple anatomical measurements.

On a more general outlook, one might speculate that increased workload over a prolonged period might eventually lead to increased fatigue, potentially decreasing a surgeon's performance after an extended period. This is purely speculative and needs to be investigated in detail through further studies, possibly utilizing our suggested approach for workload measurement—correlating electrophysiological and performance results.

This study focused solely on the workload of the 3 conditions. It should be made clear that there are scenarios where a higher workload is acceptable when weighed against the benefits that such an approach would yield. While the VR setup might be suitable for surgical planning due to its lack of visibility into the real environment, the AR setup may be the more suitable approach during surgery—such as projecting radiological images as 3D models directly into the patient's situs during a procedure to localize, for example, tumors or blood vessels. In these situations, however, technologies such as AR glasses can offer significant advantages due to their see-through capabilities, allowing a slightly higher workload, as observed in this study, to be acceptable if parameters such as patient safety are enhanced through the use of such technologies.

Interpretation of our results should be conducted with caution due to the very low number of participants (N=9), resulting in limited statistical power. This work should be seen as a feasibility study on how to measure objective workload while using AR or VR in a surgical simulation setting. However, we believe the study still provides valuable data, as experienced surgeons in a clinical setting used real-life patient cases with state-of-the-art equipment for AR and VR presentation.

Conclusions

We aimed at developing a possible setup for objective measurement of workplace strain when using AR and VR technologies in the medical profession. We conclude that the use of a passive oddball paradigm is a feasible approach for this objective. The measurement of workload during surgical planning, and thereby likely also during other medical

XSL•FO
RenderX

applications when using AR or VR technology, can be sufficiently done using an oddball paradigm. This approach proves to be a good opportunity to objectively determine workplace strain in a medical setting.

In future studies, the use of this objective measurement could potentially replace subjective questionnaires and significantly enhance the significance of the measured workplace strain. This approach, with the objectivity provided, may even become a commercially available application for medical education and surgical training, utilizing automated EEG analysis. Furthermore, we consider that AR or VR approaches to displaying radiological images are not inherently superior to the classical approach. The use of AR technology, in particular, seems to require a certain amount of training, likely due to the novelty of gesture control. However, we are confident that with future developments in AR technology and gesture recognition, these issues will become less significant over time. Regardless, the differences in workload between AR and VR highlight the need for intuitive user interaction and ease of use for any new technology, especially when applied in a medical setting.

Additionally, integrating machine learning techniques could indeed enhance the analysis of physiological data, allowing for the development of predictive models that may generalize across different tasks and populations. Such approaches could provide deeper insights into cognitive workload and potentially lead to the creation of real-time monitoring tools to optimize performance in high-stakes environments such as surgery. This could eventually result in commercially available solutions for use in medical education or surgical training.

In conclusion, this paper demonstrates the feasibility of utilizing a mobile EEG setup in conjunction with a passive oddball paradigm to assess the potential additional cognitive load imposed by emerging technologies such as AR and VR headsets in a clinical setting.

Acknowledgments

We express our gratitude to Kilian Gloy for his contributions to the planning of the conducted study. His expertise and dedication have played a significant role in shaping the research design and ensuring its successful implementation. Our thanks also go to Fraunhofer MEVIS, Bremen, Germany, namely, Andrea Schenk and Christian Schumann, for providing us with segmented liver models used in this study. This study was conducted within the scope of the VIVATOP project and has thus received funding from the German Ministry of Education and Research (BMBF) in the grant program “Gesundes Leben” (healthy living; grant 16SV8079).

Data Availability

The data presented in this study are available on request from the corresponding author.

Conflicts of Interest

DS was employed in the involved research group at the University Clinic for Visceral Surgery at the University of Oldenburg during the project, and subsequently transitioned to apoQlar (Hamburg, Germany), a company also involved in the project for the development of the AR demonstrator. Other authors have no conflicts of interest to declare.

Multimedia Appendix 1
Screenshot of the PC condition.
[PNG File , 201 KB - formative_v9i1e62740_app1.png]

Multimedia Appendix 2
Video part 1 of the virtual reality condition.
[MP4 File (MP4 Video), 8008 KB - formative_v9i1e62740_app2.mp4]

Multimedia Appendix 3
Video part 2 of the virtual reality condition.
[MP4 File (MP4 Video), 5818 KB - formative_v9i1e62740_app3.mp4]

Multimedia Appendix 4
Video of the augmented reality condition.
[MP4 File (MP4 Video), 143163 KB - formative_v9i1e62740_app4.mp4]

References

1. Dick J, Gebhard F. Zukunft der intraoperativen Bildgebung und Navigation. OP-JOURNAL 2015 Apr 23;30(03):188-195. [doi: 10.1055/s-0034-1383340]

XSL•FO
RenderX

2. Bogdanova R, Boulanger P, Zheng B. Depth perception of surgeons in minimally invasive surgery. Surg Innov 2016 Oct 09;23(5):515-524. [doi: 10.1177/1553350616639141] [Medline: 27009686]
3. Matern U, Büchel D. Gebrauchstauglichkeit von Medizinprodukten. In: Kramme R, editor. Medizintechnik. Berlin/Heidelberg, Germany: Springer; 2011:99-108.
4. Funk J, Schmidt L. Evaluation of an augmented reality in-struction for a complex assembly task. i-com 2021;20(1):63-72. [doi: 10.1515/icom-2021-0005]
5. Rosanne O, Albuquerque I, Cassani R, Gagnon J, Tremblay S, Falk TH. Adaptive filtering for improved EEG-based mental workload assessment of ambulant users. Front Neurosci 2021 Apr 7;15:611962. [doi: 10.3389/fnins.2021.611962] [Medline: 33897342]
6. Bernhardt S, Nicolau SA, Soler L, Doignon C. The status of augmented reality in laparoscopic surgery as of 2016. Med Image Anal 2017 Apr 18;37:66-90 [FREE Full text] [doi: 10.1016/j.media.2017.01.007] [Medline: 28160692]
7. Thees M, Kapp S, Strzys MP, Beil F, Lukowicz P, Kuhn J. Effects of augmented reality on learning and cognitive load in university physics laboratory courses. Computers in Human Behavior 2020 Jul;108:106316. [doi: 10.1016/j.chb.2020.106316]
8. Küçük S, Kapakin S, Göktaş Y. Learning anatomy via mobile augmented reality: effects on achievement and cognitive load. Anat Sci Educ 2016 Oct;9(5):411-421. [doi: 10.1002/ase.1603] [Medline: 26950521]
9. Chao C, Wu S, Yau Y, Feng W, Tseng F. Effects of three - dimensional virtual reality and traditional training methods on mental workload and training performance. Hum Ftrs & Erg Mfg Svc 2017 May 24;27(4):187-196. [doi: 10.1002/hfm.20702]
10. Ghani U, Signal N, Niazi I, Taylor D. A novel approach to validate the efficacy of single task ERP paradigms to measure cognitive workload. Int J Psychophysiol 2020 Dec;158:9-15. [doi: 10.1016/j.ijpsycho.2020.09.007] [Medline: 33045292]
11. Hart S, Stavenland LE. Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. In: Human Mental Workload. Amsterdam, The Netherlands: Elsevier; 1988:139-183.
12. Nicolau S, Soler L, Mutter D, Marescaux J. Augmented reality in laparoscopic surgical oncology. Surg Oncol 2011 Sep;20(3):189-201. [doi: 10.1016/j.suronc.2011.07.002] [Medline: 21802281]
13. Dixon BJ, Daly MJ, Chan H, Vescan A, Witterick IJ, Irish JC. Augmented real-time navigation with critical structure proximity alerts for endoscopic skull base surgery. Laryngoscope 2014 Apr;124(4):853-859. [doi: 10.1002/lary.24385] [Medline: 24122916]
14. Polich J. Updating P300: an integrative theory of P3a and P3b. Clin Neurophysiol 2007 Oct;118(10):2128-2148 [FREE Full text] [doi: 10.1016/j.clinph.2007.04.019] [Medline: 17573239]
15. Williams LM, Simms E, Clark CR, Paul RH, Rowe D, Gordon E. The test-retest reliability of a standardized neurocognitive and neurophysiological test battery: "neuromarker". Int J Neurosci 2005 Dec;115(12):1605-1630. [doi: 10.1080/00207450590958475] [Medline: 16287629]
16. Reinschluessel AV, Muender T, Salzmann D, Döring T, Malaka R, Weyhe D. Virtual reality for surgical planning - evaluation based on two liver tumor resections. Front Surg 2022;9:821060 [FREE Full text] [doi: 10.3389/fsurg.2022.821060] [Medline: 35296126]
17. Colligan L, Potts HWW, Finn CT, Sinkin RA. Cognitive workload changes for nurses transitioning from a legacy system with paper documentation to a commercial electronic health record. Int J Med Inform 2015 Jul;84(7):469-476. [doi: 10.1016/j.ijmedinf.2015.03.003] [Medline: 25868807]
18. Park S, Yoo J, Lee Y, DeGuzman PB, Kang M, Dykes PC, et al. Quantifying emergency department nursing workload at the task level using NASA-TLX: an exploratory descriptive study. Int Emerg Nurs 2024 Jun;74:101424. [doi: 10.1016/j.ienj.2024.101424] [Medline: 38531213]
19. Hemmerich W. Rechner zur adjustierung des α-niveaus. StatistikGuru. 2016. URL: https://statistikguru.de/rechner/adjustierung-des-alphaniveaus.html [accessed 2024-08-01]
20. Handy T. In: Handy T, editor. Event-Related Potentials. A Methods Handbook. Cambridge, MA: MIT Press; 2005:33-56.
21. Wascher E, Reiser J, Rinkenauer G, Larrá M, Dreger FA, Schneider D, et al. Neuroergonomics on the go: an evaluation of the potential of mobile EEG for workplace assessment and design. Hum Factors 2023 Feb;65(1):86-106 [FREE Full text] [doi: 10.1177/00187208211007707] [Medline: 33861182]
22. El Basbasse Y, Packheiser J, Peterburs J, Maymon C, Güntürkün O, Grimshaw G, et al. Walk the plank! Using mobile electroencephalography to investigate emotional lateralization of immersive fear in virtual reality. R Soc Open Sci 2023 May 31;10(5):221239 [FREE Full text] [doi: 10.1098/rsos.221239] [Medline: 37266038]
23. Alessa FM, Alhaag MH, Al-Harkan IM, Ramadan MZ, Alqahtani FM. A neurophysiological evaluation of cognitive load during augmented reality interactions in various industrial maintenance and assembly tasks. Sensors (Basel) 2023 Sep 06;23(18):7698 [FREE Full text] [doi: 10.3390/s23187698] [Medline: 37765755]
24. Binias B, Myszor D, Binias S, Cyran KA. Analysis of relation between brainwave activity and reaction time of short-haul pilots based on EEG data. Sensors (Basel) 2023 Jul 17;23(14):6470 [FREE Full text] [doi: 10.3390/s23146470] [Medline: 37514762]
25. Ji L, Zhang C, Li H, Zhang N, Guo C, Zhang Y, et al. A comprehensive experimental framework based on analysis of the pilot's EEG and NASA-TLX questionnaire in a VR environment. Proc Inst Mech Eng H 2023 Jul;237(7):869-878. [doi: 10.1177/09544119231179842] [Medline: 37313977]

XSL•FO
RenderX

26. Aksoy M, Ufodiama CE, Bateson AD, Martin S, Asghar AUR. A comparative experimental study of visual brain event-related potentials to a working memory task: virtual reality head-mounted display versus a desktop computer screen. Exp Brain Res 2021 Oct 04;239(10):3007-3022 [FREE Full text] [doi: 10.1007/s00221-021-06158-w] [Medline: 34347129]
27. Giorgi A, Menicocci S, Forte M, Ferrara V, Mingione M, Alaimo Di Loro P, et al. Virtual and reality: a neurophysiological pilot study of the sarcophagus of the spouses. Brain Sci 2023 Apr 07;13(4):635 [FREE Full text] [doi: 10.3390/brainsci13040635] [Medline: 37190600]
28. Marucci M, Di Flumeri G, Borghini G, Sciaraffa N, Scandola M, Pavone EF, et al. The impact of multisensory integration and perceptual load in virtual reality settings on performance, workload and presence. Sci Rep 2021 Mar 01;11(1):4831 [FREE Full text] [doi: 10.1038/s41598-021-84196-8] [Medline: 33649348]
29. Atici-Ulusu H, Ikiz YD, Taskapilioglu O, Gunduz T. Effects of augmented reality glasses on the cognitive load of assembly operators in the automotive industry. International Journal of Computer Integrated Manufacturing 2021 Mar 21;34(5):487-499. [doi: 10.1080/0951192x.2021.1901314]
30. Taori TJ, Gupta SS, Gajre SS, Manthalkar RR. Cognitive workload classification: towards generalization through innovative pipeline interface using HMM. Biomedical Signal Processing and Control 2022 Sep;78:104010. [doi: 10.1016/j.bspc.2022.104010]
31. Luque F, Armada V, Piovano L, Jurado-Barba R, Santamaría A. Understanding pedestrian cognition workload in traffic environments using virtual reality and electroencephalography. Electronics 2024 Apr 11;13(8):1453. [doi: 10.3390/electronics13081453]
32. Mondellini M, Pirovano I, Colombo V, Arlati S, Sacco M, Rizzo G, et al. A multimodal approach exploiting EEG to investigate the effects of VR environment on mental workload. International Journal of Human–Computer Interaction 2023 Sep 19;40(20):6566-6578. [doi: 10.1080/10447318.2023.2258017]

Abbreviations

AEP: auditory evoked potential
AR: augmented reality
EEG: electroencephalography
ERP: event-related potential
MRI: magnetic resonance imaging
NASA-TLX: National Aeronautics and Space Administration-Task Load Index
rmANOVA: repeated measures analysis of variance
VR: virtual reality

Edited by A Mavragani; submitted 10.06.24; peer-reviewed by Y Gogna; comments to author 30.08.24; revised version received 14.10.24; accepted 07.11.24; published 08.01.25.

Please cite as:
Schneider T, Cetin T, Uppenkamp S, Weyhe D, Muender T, Reinschluessel AV, Salzmann D, Uslar V
Measuring Bound Attention During Complex Liver Surgery Planning: Feasibility Study
JMIR Form Res 2025;9:e62740
URL: https://formative.jmir.org/2025/1/e62740
doi:10.2196/62740
PMID:

©Tim Schneider, Timur Cetin, Stefan Uppenkamp, Dirk Weyhe, Thomas Muender, Anke V Reinschluessel, Daniela Salzmann, Verena Uslar. Originally published in JMIR Formative Research (https://formative.jmir.org), 08.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Development of Personas and Journey Maps for Artificial Intelligence Agents Supporting the Use of Health Big Data: Human-Centered Design Approach

Yoon Heui Lee[1], RN, MSN; Hanna Choi[2], RN, PhD; Soo-Kyoung Lee[3], RN, PhD

[1]Department of Nursing, Graduate School, Keimyung University, Daegu, Republic of Korea

[2]Department of Nursing Science, Nambu University, Gwangju, Republic of Korea

[3]Big Data Convergence and Open Sharing System, Seoul National University, Seoul, Republic of Korea

Corresponding Author:
Soo-Kyoung Lee, RN, PhD
Big Data Convergence and Open Sharing System
Seoul National University
1 Gwanak-ro
Seoul, 08826
Republic of Korea
Phone: 82 880 2344
Email: soo1005@snu.ac.kr

Abstract

Background: The rapid proliferation of artificial intelligence (AI) requires new approaches for human-AI interfaces that are different from classic human-computer interfaces. In developing a system that is conducive to the analysis and use of health big data (HBD), reflecting the empirical characteristics of users who have performed HBD analysis is the most crucial aspect to consider. Recently, human-centered design methodology, a field of user-centered design, has been expanded and is used not only to develop types of products but also technologies and services.

Objective: This study was conducted to integrate and analyze users' experiences along the HBD analysis journey using the human-centered design methodology and reflect them in the development of AI agents that support future HBD analysis. This research aims to help accelerate the development of novel human-AI interfaces for AI agents that support the analysis and use of HBD, which will be urgently needed in the near future.

Methods: Using human-centered design methodology, we collected data through shadowing and in-depth interviews with 16 people with experience in analyzing and using HBD. We identified users' empirical characteristics, emotions, pain points, and needs related to HBD analysis and use and created personas and journey maps.

Results: The general characteristics of participants (n=16) were as follows: the majority were in their 40s (n=6, 38%) and held a PhD degree (n=10, 63%). Professors (n=7, 44%) and health care personnel (n=10, 63%) represented the largest professional groups. Participants' experiences with big data analysis varied, with 25% (n=4) being beginners and 38% (n=6) having extensive experience. Common analysis methods included statistical analysis (n=7, 44%) and data mining (n=6, 38%). Qualitative findings from shadowing and in-depth interviews revealed key challenges: lack of knowledge on using analytical solutions, crisis management difficulties during errors, and inadequate understanding of health care data and clinical decision-making, especially among non–health care professionals. Three types of personas and journey maps—health care professionals as big data analysis beginners, health care professionals who have experience in big data analytics, and non–health care professionals who are experts in big data analytics—were derived. They showed a need for personalized platforms tailored to the user level, appropriate direction through a navigation function, a crisis management support system, communication and sharing among users, and expert linkage service.

Conclusions: The knowledge obtained from this study can be leveraged in designing an AI agent to support future HBD analysis and use. This is expected to further increase the usability of HBD by helping users perform effective use of HBD more easily.

(JMIR Form Res 2025;9:e67272) doi:10.2196/67272

XSL•FO
RenderX

KEYWORDS

analysis; health big data; human-centered design; persona; user journey map; artificial intelligence; human-AI; interviews; users' experiences

Introduction

Since COVID-19, the new normal era has brought innovative advances in digital technology, resulting in various changes in all areas of our lives. Health care is rapidly emerging as one of the most representative and top-priority areas of advanced digital technology [1-3]. Digital health is a generic term for health care–related fields that combine digital technology and medical care, including telehealth, telemedicine, mobile health, wearable and internet of things–based health care, and big data and artificial intelligence (AI)–based health care [4-6]. The purpose of digital health is to create specialized services for consumers based on data- and evidence-based judgment, and its goal is to realize precision health care [3,7,8] through sustainable digital transformation of health care [9].

Precision health care is a new health care paradigm that involves analyzing various information such as genetic and environmental factors, disease history, lifestyle, and individual health information to prevent diseases (preventive health care), maximize treatment effects, and minimize treatment side effects (personalized health care); this also involves the participation of patients in the decision-making process (participatory health care) [10-13]. Having an accurate prediction based on various health data is important in realizing such precision health care [14]. Health care is one of the fastest-growing sectors worldwide and requires an efficient and continuous digital transformation [15]. The digital transformation of health care should consider technological, political, social, cultural, and economic aspects and is a key and strategic field for improving health care quality and creating new value [16].

According to IBM, an individual generates more than 1100 terabytes of health data in their lifetime [17], and this huge amount of data that was previously just discarded can now be shared, transmitted, and stored with advances in the cloud, AI, and social media. Integrating and analyzing such health big data (HBD) can derive new insights for predicting, preventing, and treating diseases [18,19]. Korea has a huge amount of HBD; hence, the development of various government-level policy support and related technologies continues [20,21] and has favorable conditions for the use of HBD. However, the usability of HBD is low. The use of big data is mainly hindered by the absence of experts related to big data analysis, inconsistent data quality, and difficulty in analysis procedures [22,23].

Currently, there are only a few professionals who can manage and analyze big data in Korea. As interest in big data expands, analytic expert training courses are continuously being opened; however, they are still far from enough to meet demands [24]. In addition, considering future demands where the use and influence of big data will further increase, the workforce imbalance is also expected to increase [22,25]. Therefore, training more people with universal capabilities based on an understanding of data can be an effective alternative to meet insufficient workforce demand and increase the use of big data [26]. To this end, a system that supports the overall process of big data analysis is needed so that many people can access HBD more easily for analysis and use. Recently, human-centered design methodology, a field of user-centered design, has been expanded and used not only to develop types of products but also technologies or services [27,28]. This helps in specifically identifying the problem based on the user's experience, identifying the user's needs, and reflecting them in the human-centered design [29,30].

AI applications are becoming increasingly common in almost all areas of human life. In addition, the medical field is also accelerating to reflect AI. The rapid proliferation of AI requires new approaches for human-AI interfaces that are different from classic human-computer interfaces [31]. Persona and journey mapping techniques have been used in the literature and industry to elicit requirements for system development. Most of the software developers, users, and managers consider that the journey map and persona techniques are effective in helping understand the software requirements to be developed by the development teams [32].

In developing a system that supports the analysis and use of HBD, reflecting the empirical characteristics of users who actually perform big data analysis is the most important procedure to consider. Therefore, this study aims to understand the empirical characteristics and feelings of users related to big data analysis by building persona and user journey map included in the human-centered design methodology and to understand users' needs for systems that support HBD analysis and use. This can be used as a design concept for AI agents to support future HBD analysis and use.

Furthermore, this study aims to design an AI agent to support future HBD analysis and use by integrating and analyzing users' experiences along the HBD analysis and use journey and applying the human-centered design methodology as follows: (1) the persona type is derived by investigating the user's experience focusing on the HBD analysis and use journey and (2) a customer journey map is created according to the derived persona to identify the needs of users presented by type.

Methods

Research Design

This is a methodological study that synthesizes and analyzes users' experiences according to the big data analysis and use journey; generally, the human-centered design methodology is used before developing an AI agent that supports the analysis and use of HBD.

Participants

Research participants were secured through convenience and snowball sampling methods for adults aged 19 years or older living in Korea who have experience in big data analysis using HBD for research or work purposes or have experience using

XSL•FO
RenderX

the results of the big data analysis. Participants for this study were recruited voluntarily by sending a recruitment notice to individuals identified as meeting the study's objectives from among those who had participated in a prior demand survey on the analysis and use of HBD [33]. This survey served as a preliminary study for this research, and the individuals identified through it were selected as the primary participants. Additional participants were then recruited through referrals from these primary participants. To ensure diversity within the participant pool, individuals were classified into 4 categories based on their level of analysis experience (beginner or proficient) and their occupational field (health and medical fields or other fields). Efforts were made to select participants evenly across these categories. There were 16 participants in this study. The number of participants was evaluated based on previous research that suggested that a saturation point without new discoveries could be reached when interviewing 6 to 12 participants, and recruitment of participants was terminated at the point when new topics were no longer derived [34]. Data were collected until no new topics emerged from the interviews, and the recruitment of new participants was stopped at the time of saturation [35].

Procedures

This study consisted of 2 stages based on a review of the big data analysis and use procedure through desk research and a case study on the HBD platform (Figure 1). The first stage, discover, involves examining and understanding the problem situation to find the difficulties and hidden needs of those who use the service. Data research was conducted through shadowing and in-depth interviews with participants working on HBD analysis. The second stage, define, involves deriving major issues related to the problem based on analysis and contextual understanding of user survey results and setting service goals. Based on the needs of users explored through persona and user journey map, the AI agent's design concept was proposed.

Figure 1. Research progress process: this figure outlines the research process, which comprises 2 stages: the first stage focuses on identifying users' challenges and hidden needs, while the second stage involves defining the problem and setting service goals to address those issues.

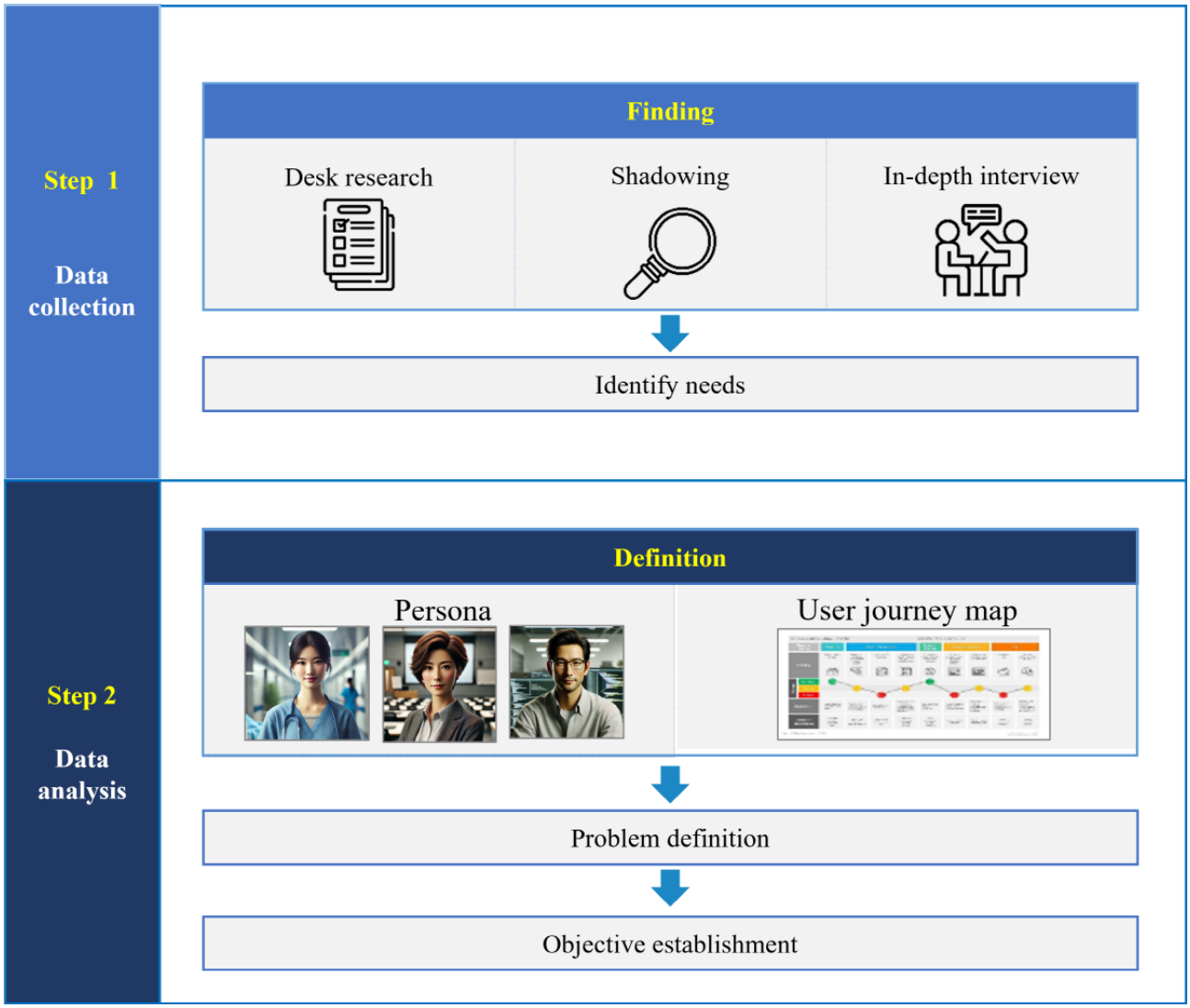

Data Collection

Desk research examined prior documents and public institution announcements and papers searching for HBD and big data analysis. Platforms in overseas (PubMed and Google Scholar) and domestic (DBPia, Riss, and Kiss) databases were also searched. In-depth interviews took place over 3 weeks from February 1 to 21, 2023. The interview was conducted using a

XSL•FO
RenderX

guideline consisting of 6 short-answer questions on the participant's basics and open questions related to their HBD analysis and use journey. Based on the related literature [36] and expert experience, some main categories of difficulties and coping methods experienced in each stage of big data analysis arose, as well as desired support services used to solve the problem (Textbox 1). This was done after the initial formulation and underwent expert review.

Textbox 1. In-depth interview guideline: open-ended questions regarding users' challenges and support needs at each stage of health big data analysis.

Collection

- Talk about the type of data you used, and how to select and collect data.
- What was the most difficult part of the data collection process?
- How did you solve the difficulty?
- If you had a magic wand to solve any problems during the data collection phase, what would you solve?

Preprocess

- Talk about the process of preprocessing collected data (integration, conversion, and refinement).
- What were the difficulties you experienced during the preprocessing process and what was the most difficult thing?
- How did you solve the difficulties that occurred during the preprocessing stage?
- If you had a magic wand that could solve any problems during the preprocessing phase, what would you solve?

Analyzing

- Did you conduct the analysis yourself? If not, why?
- Was there any problem when you asked an expert to analyze it?
- What were the difficulties you experienced in the analysis stage, and what was the most difficult thing?
- How did you solve the difficulties that occurred?
- If you had a magic wand to solve any problems during the analysis phase, what would you solve?

Visualizing

- Was there any difficulty in interpreting and applying the derived results?
- What were the difficulties you experienced in the analysis stage, and what was the most difficult thing?
- What support do you need most during the result interpretation and visualization phase?

Use

- What features do you think should be included in artificial intelligence agents that support the use of health big data analysis?

To increase the realistic value of the data, the researcher visited the user's studio where the computer was used by the participant for actual big data analysis and conducted one-on-one interviews. The individual interviews took about 1.5 hours. The process concluded when responses became repetitive, and no new information was generated. For aspects that were challenging to understand through verbal explanations alone, participants were asked to provide illustrative examples during the interview. With prior consent from the participants, the interviews were audio-recorded, and significant concepts identified during the discussions were documented in field notes. Each participant was interviewed once, and if further information was required during the data analysis phase, additional data were collected through follow-up phone calls or emails with the participants.

For 2 weeks, the researcher visited and shadowed 1 participant, a nurse researcher, who is currently conducting HBD analysis. While shadowing her in the laboratory, the researcher observed the participant's behavior, and they captured the contact point of inconvenience and the moment when a problem occurred. To check for difficulties encountered by participants under different stages, from preprocessing to visualization of HBD, the researcher visited 3 times for a total of 1 hour per visit and interviewed the participant when important issues were found. Observed behaviors and environmental factors were recorded in field notes.

Data Analysis

The interview data were transcribed by the researcher on the same day they were collected, and the transcriptions were analyzed at the levels of codes, subcategories, categories, and themes using qualitative content analysis procedures [37]. In the first stage, redundant content was excluded from the meaning units derived from the interviews, and codes were created to represent the core ideas using relevant concepts or phrases. In the second stage, the implicit meanings or content captured in the codes were abstracted to form subcategories. In the third stage, these subcategories were grouped into broader categories. Finally, in the fourth stage, themes were developed by

XSL•FO
RenderX

integrating participants' key experiences related to the HBD analysis process. To ensure the rigor of the analysis, the data were reviewed, analyzed, and discussed iteratively with the assistance of the second author (HC), a nursing professor with extensive experience in qualitative research. The findings were further validated for their universality and content validity through feedback from 2 research participants.

Based on insights confirmed through interviews and shadowing, personas and customer journey maps were designed to represent the participants' experiences. A persona is a vivid and specific description of a potential end user and a fictional character representing the type of group to use the service; this is created by adding attributes to multiple subgroups identified within a potential user group [38]. Identifying users' needs through personas formed by systematically classifying and embodying potential user groups can help present specific goals and directions for service development [39]. User journey maps are visualizations of the overall process and emotions of a persona experiencing the service over time; this helps in identifying and understanding the context of the problematic situation [30,40].

Ethical Considerations

This study was approved by the institutional review board of the University of Keimyung (40525-202210-HR-059-03). All participants received a detailed information sheet about the research, were given the opportunity to ask questions, and were informed of their right to withdraw at any time without any negative consequences. Written informed consent was obtained from all participants, and to ensure confidentiality, all data were anonymized. To ensure confidentiality, all audio-recorded data were anonymized by removing personal information and assigning unique identification codes to distinguish participants. These data were securely stored on the researcher's personal computer, equipped with security software, and were used solely for research purposes. Measures were taken to prevent any breach of personal information, violations of human rights, or ethical concerns, and the data will be permanently deleted after the study's conclusion. Upon the completion of interviews and observations, a small honorarium was provided to the participants.

Results

Participants' General Characteristics

Participants' general characteristics are as follows (Table 1). Regarding the age of the participants, 6 (38%) were in their 40s. In terms of education, 6 (38%) held a master degree, and 10 (63%) held a PhD degree. As far as professions, professors (n=7, 44%) and researchers (n=4, 25%) accounted for the majority. Among the participants, 10 (62%) were health care personnel, and 6 (38%) were non–health care personnel. Regarding the experience of big data analysis, 4 (25%) were beginners, 6 (38%) had minor experience, and 6 (38%) had extensive experience. Formal data from public institutions were the most common type of data used (n=10, 62%). The analysis methods were statistical analysis (n=7, 44%), data mining (n=6, 38%), and text mining (n=5, 31%). Big data were used mostly for academic and research purposes (n=14, 88%).

XSL•FO
RenderX

Table 1. General characteristics of participants about age, occupation, education, and experience in health big data analysis (N=16).

Characteristics and categories	Values, n (%)
Age (years)	
≤29	1 (6)
30-39	4 (25)
40-49	6 (38)
50-59	4 (25)
≥60	1 (6)
Degree of education	
Master degree	6 (38)
Doctoral degree	10 (62)
Job	
Doctor	1 (6)
Nurse	2 (12)
Professor	7 (44)
Researcher	4 (25)
Company worker	2 (12)
Experience of analysis	
Beginner	4 (25)
Minority	6 (38)
Majority	6 (37)
Method of analysis[a]	
Statistical analysis	7 (44)
Data mining	6 (38)
Text mining	5(31)
Type of used data[a]	
Structured data (public institution)	10 (62)
Dielectric data	2 (13)
Text data	5 (31)
Purpose of use	
Scholarship or research	14 (88)
Product or service development	2 (13)

[a]Multiple responses.

Shadowing

As a result of observing the preprocessing of text data, which were unstructured data, the natural language processing stage had more procedures for manual verification than automatic tasks; this involved visually checking numerous data individually using analytical solutions. Researchers call this work manual labor due to the enormous amount of time spent on the work and considerable energy consumption. This pretreatment is an important task that can affect the outcome later if it is not completely refined at that stage. However, it has been found that nonanalysis professionals often deal with huge amounts of work without understanding the step-by-step progress, which often leads to situations wherein important parts are omitted.

Most users are trained on how to use the analytical solution before conducting big data analysis; however, they often forget the necessary commands or use because they do not have sufficient knowledge and are unfamiliar with the command line interface. In this case, the user wants to inquire about the information to an analysis-related expert, but there is no expert around to answer; hence, it was found that it was mainly solved by books or internet searches. However, this method took too much time, and in most cases, the desired information could not be found.

XSL•FO
RenderX

In-Depth Interview

Overview

Based on the results of the in-depth interview, the user's difficulties in the process of analyzing HBD showed a clear difference depending on the job. In the case of health care personnel, there was a slight difference in their experience when performing analysis work directly and requesting analysis work from analysts; yet, in both cases, there was no difficulty in understanding health data or finding potential meanings in analysis results and applying them to clinical situations. However, it was found that the major problems commonly experienced involve selection, installation, and use of analytical solutions and lack of knowledge and understanding of step-by-step analysis tasks that vary widely by data. Meanwhile, for non–health care personnel, difficulties in decision-making at various stages derived from a lack of health care knowledge have been identified as a major problem. These difficulties can be divided into 3: lack of knowledge on how to use analytical solutions (health care personnel), lack of crisis management ability (health care personnel), and lack of understanding of health care data and clinical decision-making ability (non–health care personnel).

Type 1: Lack of Knowledge on How to Use Analytical Solutions

In the case of health care personnel, beginners who did not have much experience in analyzing big data answered that they had difficulty in selecting and installing appropriate analytical solutions. Moreover, in the preprocessing and analysis stages, they did not know the package or library required for each step or did not know the syntax or function. One participant stated that:

> *I thought it would be over if I installed the analytical solution, but it was really painful to have to reinstall the library I needed for each task. I don't know which package to use and how to install it, and how to enter the command after installing it. I don't remember what I learned in the past, and it's faster to do things manually, so most of the preprocessing were done manually.* [Participant 4]

Type 2: Lack of Crisis Management Skills

Participants responded that they had difficulty finding the cause and solution when an error occurred during big data analysis. In particular, if the error was not resolved, they could no longer proceed to the next step, and the whole process was completely interrupted; sometimes, even very simple problems that could be easily solved could not be found, so they spent days and days resolving them. One participant stated that:

> *I definitely put in the command as I learned, but I kept getting error messages. I don't know what's wrong, and I can't ask anyone because there's no one around me with this knowledge. I found Google and YouTubers to solve it, but it doesn't work out. "Should I really give up here?" I had all sorts of thoughts. Later, through the expert, I found that the cause of the error was a mistake in the command. When I corrected it, it was executed in less than five minutes, but I was disappointed. I wasted three days without doing anything because of that and tried all sorts of ways by Googling.* [Participant 7]

Type 3: Lack of Understanding of Health Care Data and Clinical Decision-Making Ability

In the case of non–health care personnel, it was difficult for participants to determine the clinical significance or usability in the stage of selecting the analysis topic. In addition, for some domestic public data, participants found the data collection procedure too complicated. They responded that it was difficult to grasp the meaning of various variables included in the collected data due to a lack of understanding of health data and that it was cumbersome to merge data separated by year or variable. Regarding preprocessing, it was found that difficulties in preprocessing procedures such as normalization and missing value treatment, which changed the health care meaning and resulted in difficulty in decision-making, were greater than the difficulties related to the analysis itself. In addition, it was confirmed that it was difficult for participants to select variables to be used for analysis. One participant stated that:

> *To use data from the National Health Insurance Service, you have to submit IRB in the research plan, which is so complicated and difficult. The data collection stage was more difficult than the analysis. "Why did the industrial complex data divide the files into so many?" The examination data are difficult to combine because the items conducted every year are slightly different. There is an explanation file for variables, but there are too many variables in the data, so it is difficult to grasp. The test results do not clearly show the extent of the numerical value, so I can't judge well. All the teachers who work in the hospital are busy, and there is no one close enough to approach personally and seek healthcare advice.* [Participant 12]

Definition of Persona and User Journey Map

Overview

In this study, when combining situations based on users' actual experiences obtained through in-depth interviews, it was confirmed that analysis experience, occupation, and direct performance of analysis work act as the main factors affecting the user's experience. Accordingly, 3 personas were set up: "health care professionals as big data analysis beginners," "health care professionals who have experience in big data analytics," and "non–health care professionals who are experts in big data analytics." Based on this, needs and pain points from the user's perspective were identified (Figure 2).

XSL•FO
RenderX

Figure 2. Three types of personas: persona 1: health care professionals as big data analysis beginners, persona 2: health care professionals who have experienced big data analytics, and persona 3: non–health care professionals who are experts in big data analytics.

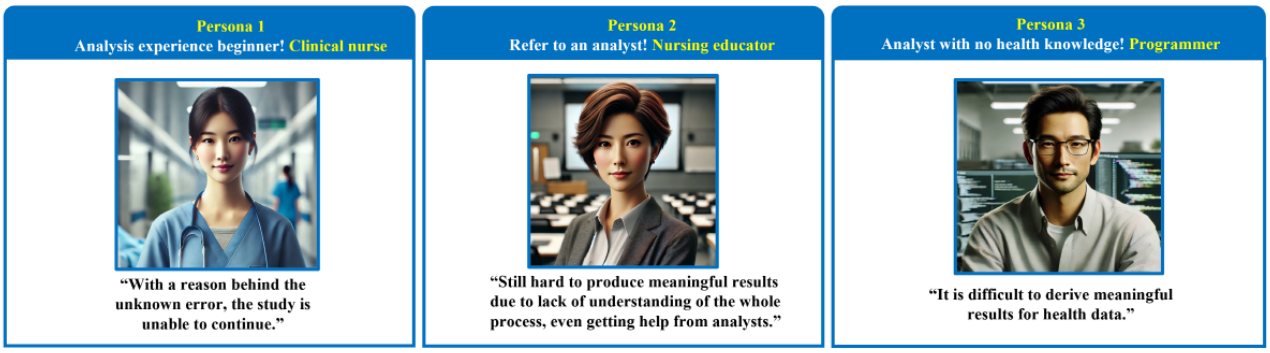

Persona 1: Health care professionals as big data analysis beginners

Persona 1 is a persona that represents a health care practitioner with entry-level analytical experience (Figure 3). They lack knowledge of analytical solutions, experience frequent errors in progress due to unfamiliar use, and have difficulty finding solutions. Therefore, they need a support system that can provide information on how to use the analytical solutions and query errors or to be able to ask questions real time during the progress stage.

Figure 3. Persona 1: health care professionals as big data analysis beginners: represents novice medical personnel with limited experience in big data analysis. This persona lacks knowledge of analysis tools and solutions and has minimal experience using them.

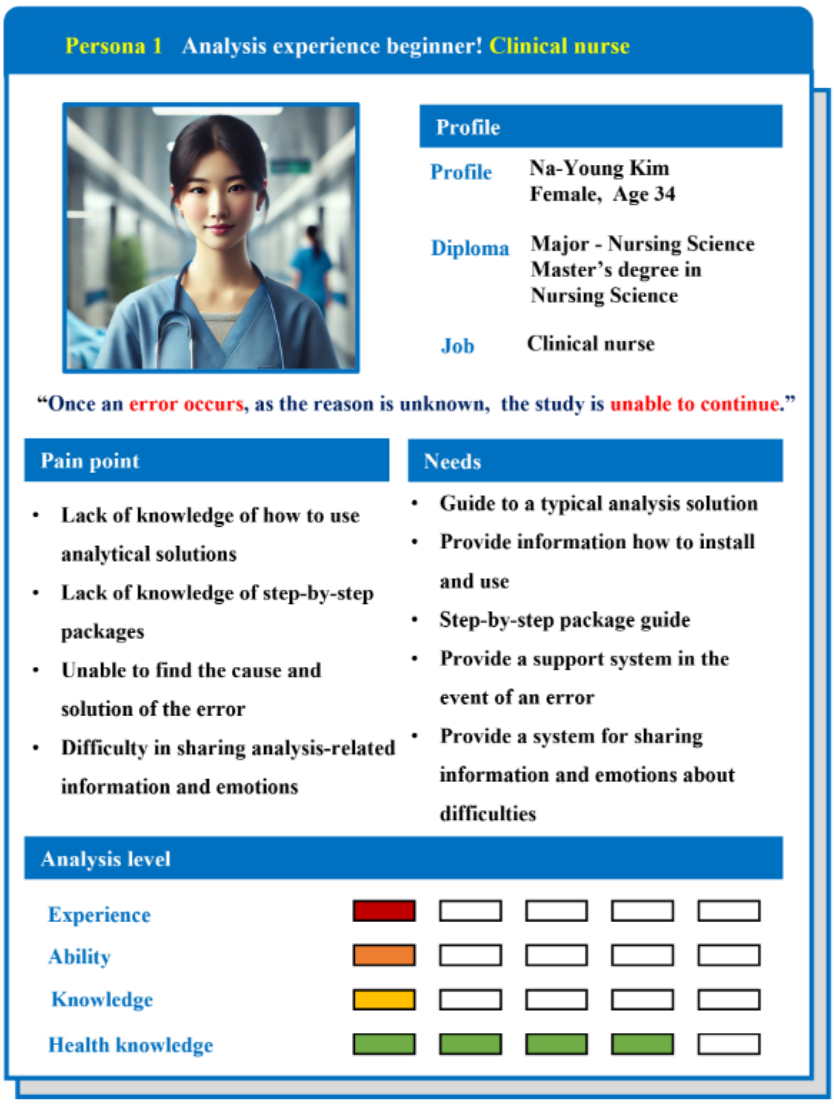

The user journey map prepared based on this persona is shown in Figure 4. "Na-Young Kim," who works as a clinical nurse at a university hospital, recently conducted a potential class analysis study on "risk factors for metabolic syndrome in

XSL•FO
RenderX

single-person households." Since she has never performed big data analysis before, she completed analysis-related training before conducting the research. She was unfamiliar with how to use the analytical solution in the preprocessing process, so she had to proceed with most of it manually, which consumes a lot of time and energy. In the complete analysis stage, an error occurred during the command input process, but the progress was stopped because the cause of the error could not be found. She searched for information through Google and tried to modify it by referring to the provided notice, but it was not solved, so she spent days and days working on the problem. The cause of the error, which was confirmed through experts who were barely connected after many twists and turns, was unexpectedly too simple, and the problem was solved in 5 minutes. She thinks several times that it would be nice to have someone who she can ask for advice on something. After many twists and turns, the analysis was completed, but with the idea that such a difficult experience is enough once in a lifetime, she decided that she will never do big data analysis research again.

Figure 4. User journey map of persona 1: highlights frequent errors occurring during the data preprocessing and analysis stages. The inability to identify the causes of these errors leads to significant time consumption and frustration. Q&A: question and answer.

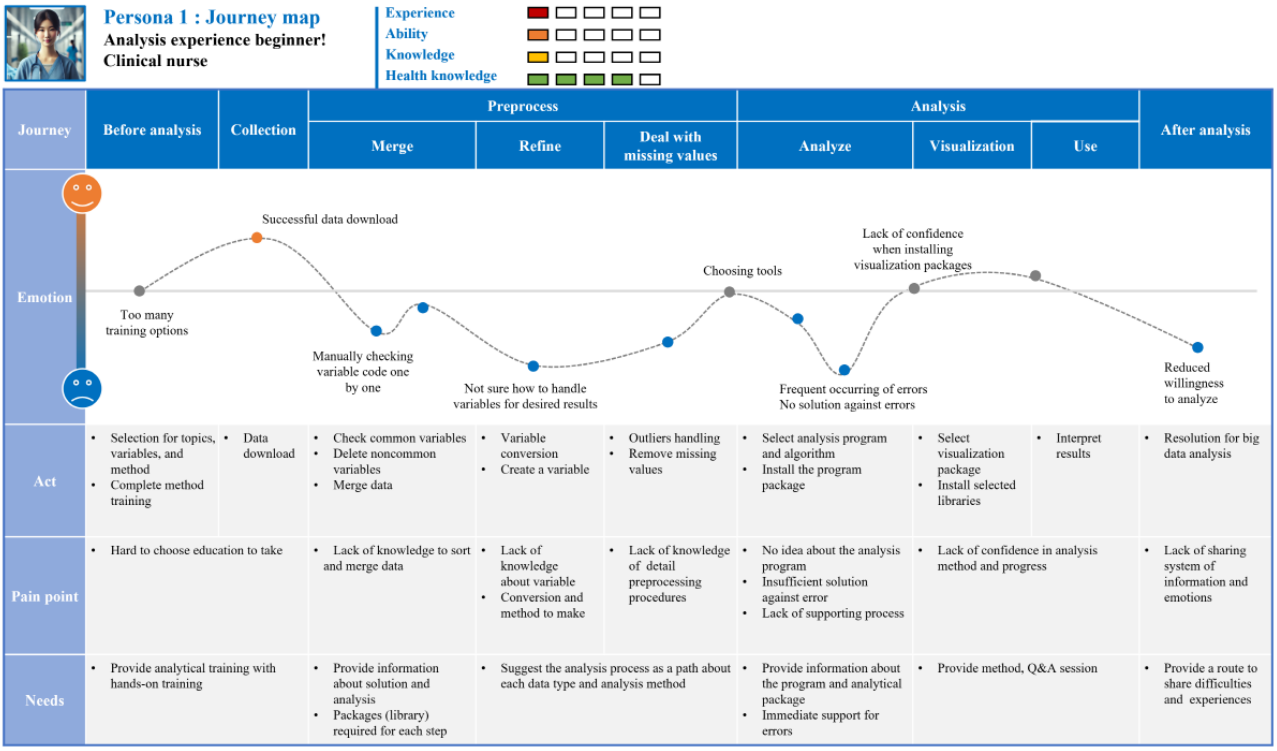

Which part of it is wrong? I don't know, there's no one I can ask, and once an error occurs like this, everything will be completely stopped, so it's really hopeless. I can't believe I held on to a problem that would be solved in five minutes for four days. It's so frustrating.

Persona 2: Health Care Professionals Who Have Experienced Big Data Analytics

Persona 2 is a persona that represents the experience of a health care professional who commissioned an analysis expert to perform the analysis (Figure 5). In this case, if the client lacks an understanding of the overall process of big data analysis or does not communicate sufficiently in advance, disagreement may occur during the process. Therefore, it was found that users who conduct big data analysis through collaboration with analysts need a system that provides information on the contents and precautions to be discussed in advance when requesting analysis, guides the entire process of big data analysis, and provides explanations.

XSL•FO
RenderX

Figure 5. Persona 2: health care professionals who have experience in big data analytics: represents medical personnel who outsource big data analysis tasks to experts. This persona lacks an understanding of the overall analysis process and methods.

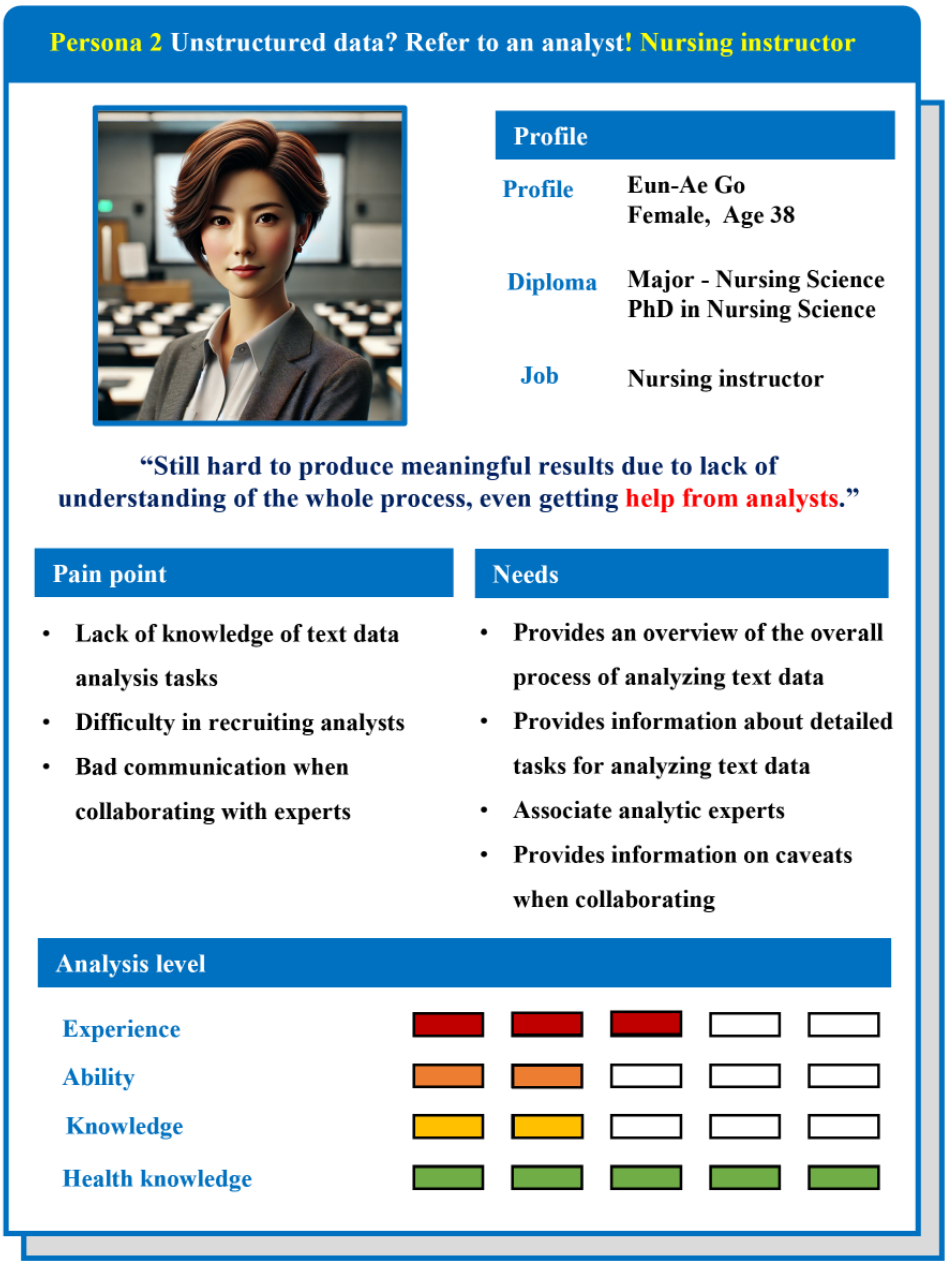

The itinerary map for persona 2 is shown in Figure 6. "Eun-Ae Go" is a lecturer who is currently teaching students at a nursing university and recently conducted a study to identify "severe COVID-19 risk factors" using HBD. Although she has previously conducted several studies using HBD, it was her first time analyzing a network using text data; hence, she completed several training sessions on using network analytical solutions before starting the analysis. She did not have an expert in big data analysis around her, and she could not make a decision because she did not trust the companies she found on the internet. She finally got in touch with an engineering major through a few acquaintances and asked them to do both data preprocessing and analysis. It was found after the end of the analysis that the resulting values were affected by texts that were not completely refined during the preprocessing process. She wanted to reanalyze by repurifying, but the analyst did not agree to the reanalysis due to time and cost issues. Through this, she realized that sufficient communication and consultation must precede when requesting analysis work from others; there is also a limit to discovering and interpreting the meaning of the data only with the analysis results conducted by the AI agent without understanding the analysis process or method. Next time, she intends to perform the pretreatment and analysis herself, even if she is not an expert.

XSL•FO
RenderX

Figure 6. User journey map of persona 2: illustrates difficulties in communication and disagreements during the collaboration process, stemming from limited knowledge of the analysis process and methods. These issues result in constraints on the quality and scope of the analysis outcomes.

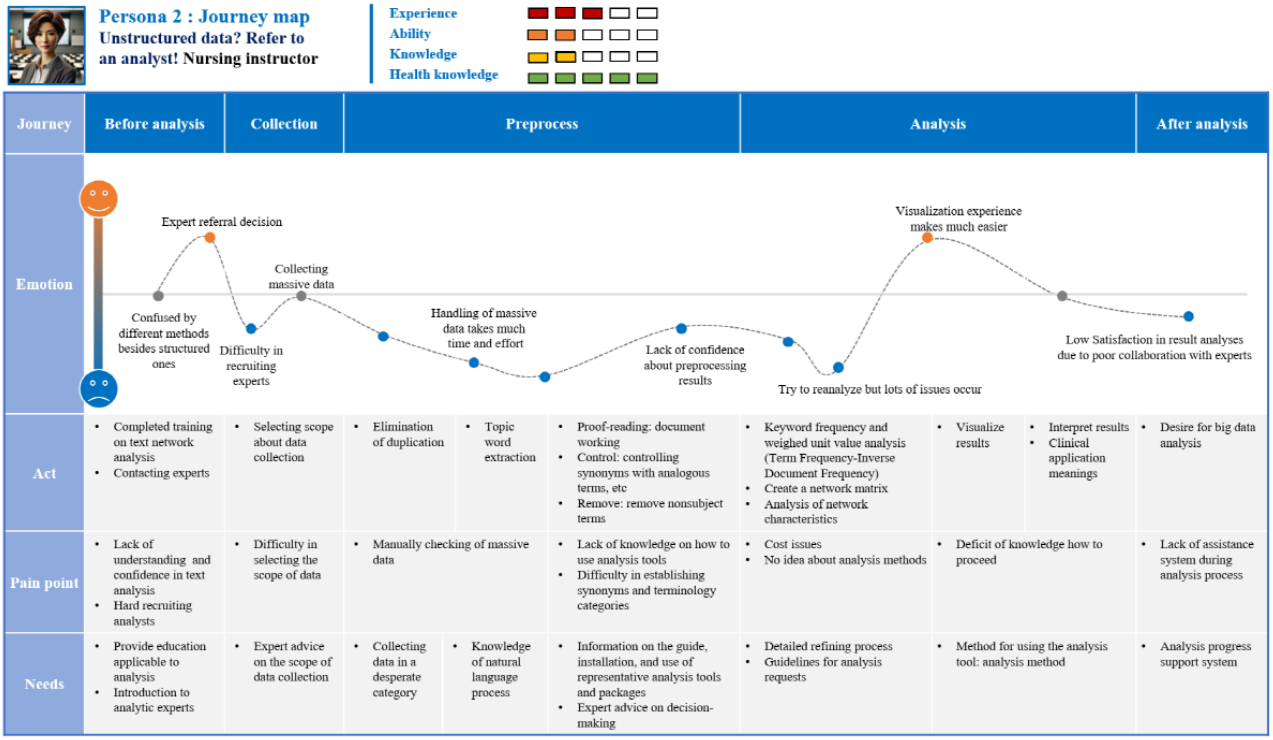

You need to know something when you ask others to do it. Even if we ask experts to analyze big data, we can produce meaningful results only when we have a basic understanding and knowledge of the overall analysis process and method.

Persona 3: Non–Health Care Professionals Who Are Experts in Big Data Analytics

Persona 3 is a persona representing people who are experts in big data analysis but have no knowledge of health care (Figure 7). They are good at big data and analysis, but they experience continuous doubts and confusion during the analysis process due to a lack of health care knowledge. Through this, it was found that non–health care users who are fluent in analysis need a system that provides information on health care data. They also need to seek advice from health care experts in decision-making situations that occur during the analysis process.

XSL•FO
RenderX

Figure 7. Persona 3: non–health care professionals who are experts in big data analytics: represents a nonmedical professional proficient in big data analysis but lacking expertise in health care. This persona struggles with understanding health data and clinical knowledge.

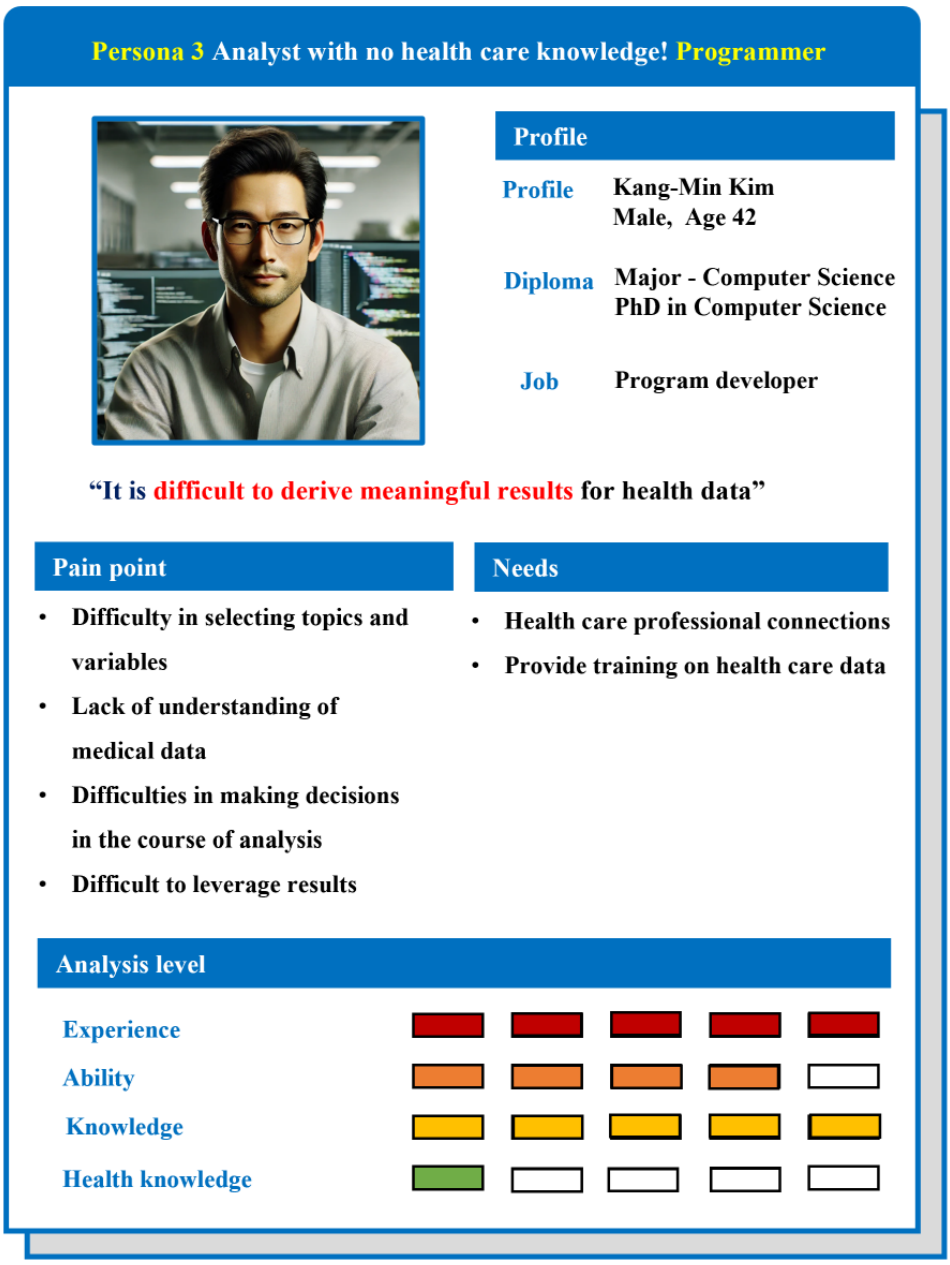

The itinerary map for persona 3 is shown in Figure 8. Currently working as a programmer, "Kang-Min Kim" is a computer science major and a big data analysis expert. Several studies related to big data analysis in various fields have been conducted. Recently, interest in HBD with infinite use value has increased, and research on "building a diabetes prediction model" has been conducted using the National Health Insurance Service's screening data. Since "Kang-Min Kim" is good at analyzing big data, he predicted that there will be no special variables or difficulties in analyzing big data in the process of HBD analysis. However, unlike previous data analysis work, he felt uncertainty and doubt throughout the analysis process due to the nature of the health data that the analysis findings of the machine were not 100% reliable. Due to a lack of understanding of the meaning of each variable in the health data, the impact of normalization and missing value processing methods on future analysis results, the selection of input variables when applying machine learning, and clinical meaning and application of derived results, among others, he had doubts that he could not be sure of the results of the work that he decided and implemented alone without communication with related experts. At the end of the work, it was deduced that clinical judgment and opinions of health care experts must be combined in the analysis process to properly use HBD. In the future, he intends to conduct HBD analysis through collaboration with health care experts.

XSL•FO
RenderX

Figure 8. User journey map of persona 3: describes experiences of confusion and uncertainty when situations requiring clinical judgment and decision-making arise at various stages, including preprocessing, analysis, and postanalysis.

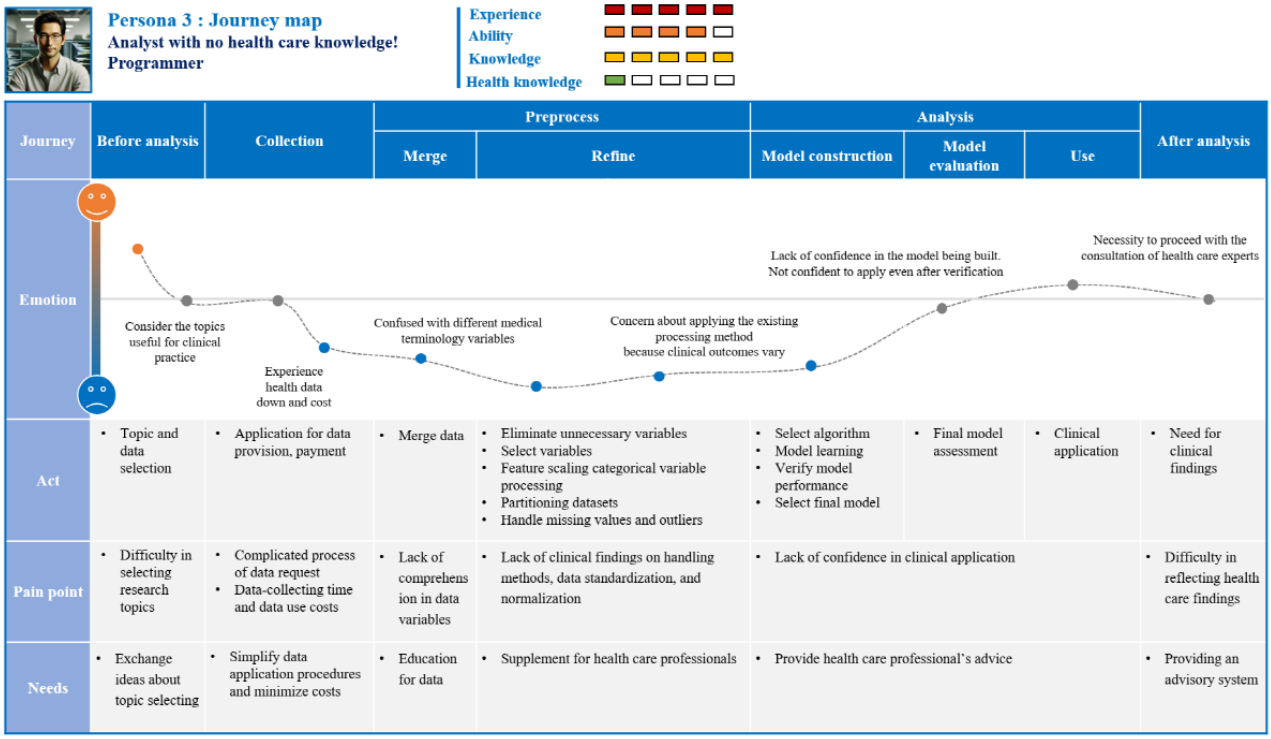

This numerical value is an abnormal range, how much does this abnormal mean? Does this process affect the clinical meaning of data values? I can't judge because I don't have healthcare knowledge. A proper analysis of health big data requires not only being good at analysis itself.

Deduction of Needs

The user's needs derived for AI agent design that supports the use of big data analysis are as follows.

First, there is a need for a service that provides a guide for analytical solutions, which can be easily used by the public and provides information on how to use the solution to proceed with each step. For nonanalysis professionals, the knowledge of the various libraries (packages) and execution methods required for several tasks to be performed at each stage of preprocessing, analysis, and visualization was unclear; even those with experience did not know how to use the solution. Therefore, users will be able to conduct big data analysis more easily if they have a guide for analytical solutions that are easily available to the public and provide detailed explanations on how to use the solutions for step-by-step tasks according to the progress procedures for each data type.

Second, there is a need for a service that guides the overall process of big data analysis in detail. Big data analysis takes place through 4 stages: data collection, preprocessing, analysis, and visualization and use. Most of the big data analysis platforms in other fields currently operate these 4 stages and provide services such as data introduction and transaction services, data analysis tools introduction and installation guidance, analysis-related training guidance, data analysis and use case sharing. However, there is no system that guides detailed work procedures and specific methods required for each stage. Ordinary people, who are not big data analysis experts, need a service that guides them in their overall analysis journey and provides detailed explanations of the specific tasks to be performed at each stage, just like with navigation, wherein a person arrives at the destination if guided.

Third, there is a need for a system where one can ask or seek real-time advice from analysis experts about problem situations or questions that occur during analysis work. Users may encounter problem situations in the analysis process, and there may be questions that they want to be confirmed by experts. These problems should be dealt with quickly because if they are not solved, they cannot proceed to the next step. However, it is not easy for nonexperts in the field to find the cause of the problem, resulting in a significant waste of time.

Fourth, there is a need for a service that introduces and links analysis experts who can request analysis work or health care experts who can provide health care advice. There may be situations wherein health care personnel need to seek advice on big data analysis and non–health care personnel need to seek health care advice, but it is not easy for users to find these people on their own. It is necessary to provide a channel to meet necessary professionals and at the same time provide guidance on the contents of prior consultations or precautions necessary for collaboration.

Based on the earlier-mentioned problems and user needs, we present five design concepts: (1) personalized platform for the user level, (2) appropriate direction through a navigation function, (3) crisis management support system through question and answer, (4) communication and sharing opportunities between users, and (5) expert-linked services. In addition, it is

XSL•FO
RenderX

believed that these concepts will be able to meet users' needs by adding experience or information related to HBD analysis, providing guidance on health data introduction and access procedures, giving useful big data–related education introduction, and having one-on-one consultations.

Discussion

Principal Findings

To design appropriate and useful technologies or services, it is important to focus on the applicable user's experience and consider the user's individual needs [41]. Recently, this user-centered approach has been actively used in health care–related service development, such as in studies on the design of eHealth AI services [42], smart home system design [43], and web-based agents to support physical and mental health self-management in patients with chronic obstructive pulmonary disease [44]. In this study, before developing an AI agent that supports the use of HBD analysis, the experience and needs of users who have performed big data analysis by applying a human-centered design methodology were identified. The discussions on the 5 service concepts derived through this and the goals and contents to be achieved through each service are as follows.

First, the service is a personalized platform that fits the level of the first user. As a result of the study, the demand for a system that supports big data analysis was different for each user type. The analytical solution of the analysis environment provided in the existing big data platform focuses only on creating an analytic environment and building necessary analytical solutions rather than meeting the needs of users [45,46]. These platforms are not a problem for analysts to use, but in the case of ordinary users, they do not even know how to use a given analytical solution, so the platform created is practically irrelevant or meaningless. Therefore, it is important to provide an analytical environment that fits the user's level and includes detailed information required by the user [47].

In a study on web-based platform development to support participants and caregivers with memory problems, it was also found that the evaluation of the usefulness of the function was different for each group due to different priorities for platform functions between user groups. As a result, a customized platform was designed to meet the needs of a specific group [48]. Guiding cost-free, easily available analytical solutions, and providing necessary information according to the user's needs such as installation methods, necessary libraries, and commands, will reduce the fear of using analytical solutions by inexperienced users. This is supported by examples of successful results in weight, blood pressure, and blood sugar management by characterizing users in the development of a meal planning platform that supports healthy eating development and providing services such as customized nutrition assessment for each user, personalization of meal plans, and changing the environment of individual food purchases [49].

Second, the service provides an appropriate direction presentation system through a navigation function. Most health care professional users were found to lack knowledge on detailed big data analysis processes that vary by data type. If a detailed guide is provided for the analysis progress procedure for each type of representative data, like a road guide navigation, even inexperienced beginners can follow the guide and conduct the analysis on their own. In addition, this can improve understanding of the overall analysis process. In a digital platform designed for youth mental health services, the path of each service needs to be presented, and the content of the service provided by each route is different so that users map along the treatment path determined by case and are guided by the platform to continuously implement it according to the process [50].

Third, the service is a crisis management ability support system through question and answer. A service that allows users to raise queries and answer questions or errors that arise during analysis operations helps users solve problems quickly and reduces unnecessary time and energy waste. A platform development study, which supported operators who manage big data information processing systems, was also equipped with an intelligent web-based system based on AI that helps system operators cope with problems that occur occasionally and obtain the necessary information [51]. In the development of a mobile platform to support patient participation and reporting on lung cancer surgery, a channel was provided to ask patients using the program about questions or concerns that may arise during and after surgery or to report various suspected exacerbations. This allows users to hear immediate health care team explanations, promoting their ability to solve problems [52].

Fourth, the service creates a venue for communication and sharing that provides opportunities for exchange between users. There is a need for a channel of communication that can exchange experiences or information related to big data analysis and share effective big data use cases. Through this, the user may obtain important information and achieve new collaboration through communication. PatientsLikeMe (PatientsLikeMe Inc), a platform that provides information on rare diseases, provides a channel for users to share various information such as their health status and symptoms, drug effects, and side effects. It is not only useful to other users, but it has also resulted in an increase in the number of subscribers to the platform and further expansion of the platform's capabilities, with accumulated data being sold back to necessary institutions [5,6]. The survey on user needs for digital mental health platforms also revealed that the community was a major topic. It has been found that users do not hide their information about diseases or treatments but rather want to connect and share experiences with others, thereby developing the platform's capabilities and believing that new use cases can be developed [53]. In a digital platform designed to enhance cancer treatment support for adult survivors of cancer, survivors of cancer were also found to receive psychological support by expressing various interests such as eating habits and mental health to peer and patient support groups and sharing experiences [5,8].

Finally, the service provides a connection service with experts. Users who want to request analysis work, analysts, or those who need advice from health care experts are required to provide services that link related experts, and these services enable active

XSL•FO
RenderX

use of big data through collaboration. PatientsLikeMe also uses services that link institutions that run clinical trials on new drug development with patients who want the opportunity to use a drug being tested.

Strengths and Limitations

This study adopted a human-centered design approach to enhance the usability of HBD, successfully identifying the unique needs and challenges faced by various user groups. Through in-depth interviews and shadowing techniques, we gained significant insights into the emotional and cognitive aspects of users' experiences during the analysis process. The development of 3 distinct personas and journey maps revealed that each user group has specific requirements. For instance, novice health care professionals require more guidance and support during critical situations, whereas experienced data analysts prefer advanced features and customized settings tailored to various data types. These findings indicate that services provided by the system cannot be uniform and underscore the necessity for personalized user support. Additionally, while many users seek expert advice or communication with peers who have similar experiences, there is a notable lack of accessible communities dedicated to health data analysis and use. This highlights the need for developing user-friendly systems capable of offering practical support to users in the future. Our study has several limitations. First, the in-depth interview participants were primarily users who analyze and use widely used health data, which may limit the generalizability of the findings to users working with less common or emerging types of health data. As new types of health data continue to emerge, the specific demands and analytical experiences associated with these data have not been fully incorporated into the current personas and user journey maps. Additionally, this study may not be generalizable to users outside of the South Korean environment, as the findings are context-specific.

When the number of participants is expanded in future research, the personas and journey maps developed in this study may require reorganization, modification, or expansion beyond the 3 types presented. Like all in-depth studies, the results of this research cannot be standardized but aim to broadly capture the complex and challenging experiences of users regarding HBD analysis and use. Nevertheless, the personas and journey maps created in this study can serve as a foundational guide, gradually expanding and adapting to meet the needs of a wider range of users and data environments. Future research should conduct in-depth demand analyses targeting users of emerging data types to more comprehensively reflect diverse user experiences.

Conclusions

The ultimate purpose of precision health care is to provide optimal treatment and treatment tailored to consider individual genetic, environmental, and lifestyle factors based on health care information. This can be realized through an effective use of big data. This study proposes a human-centered design to develop an AI agent supporting the use of HBD to increase use. Based on the needs of users through shadowing and in-depth interviews, 5 key implications for HBD analysis and use support system (personal platform, navigation function, crisis management support, communication and sharing, and experts) were revealed.

In the future, we propose a follow-up study to develop a system that supports the overall process of big data analysis based on the implications derived through this study. This is expected to further increase big data use by helping many users who hesitate to use big data due to difficulties in the analysis process to analyze big data more easily.

Acknowledgments

The authors would like to acknowledge all participants who provided us with their time and shared their experiences in the interviews. This work was supported by the National Research Foundation of Korea; grant funded by the Korean government (Ministry of Science and ICT; 2021R1A2C2095271).

Data Availability

The datasets generated and analyzed during this study are not publicly available due to privacy restrictions but are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

References

1. De Santis KK, Jahnel T, Sina E, Wienert J, Zeeb H. Digitization and health in Germany: cross-sectional nationwide survey. JMIR Public Health Surveill 2021;7(11):e32951 [FREE Full text] [doi: 10.2196/32951] [Medline: 34813493]
2. Golinelli D, Boetto E, Carullo G, Nuzzolese AG, Landini MP, Fantini MP. Adoption of digital technologies in health care during the COVID-19 pandemic: systematic review of early scientific literature. J Med Internet Res 2020;22(11):e22280 [FREE Full text] [doi: 10.2196/22280] [Medline: 33079693]
3. Shin JY. What we have to prepare for being smart in healthcare: based on the Korean current situation and long-term direction of the national health insurance. HIRA Policy Brief. 2021. URL: https://repository.hira.or.kr/handle/2019.oak/2534 [accessed 2023-05-30]

XSL•FO
RenderX

4. Han JA, Kim EJ. Smart health care. KISTEP Technol. Trends Brief. 2020. URL: https://www.kistep.re.kr/board.es?mid=a10306060000&bid=0031&b_list=10&act=view&list_no=34992&nPage=4&keyField=&orderby= [accessed 2023-05-20]
5. Kostkova P. Grand challenges in digital health. Front Public Health 2015;3:149-199. [doi: 10.3389/fpubh.2015.00134]
6. Digital health center of excellence. US Food and Drug Administration. URL: https://www.fda.gov/medicaldevices/digitalhealth/#mobileapp [accessed 2023-06-02]
7. Choi YS. How will we respond to changes in the Age of digital technology? Global smart healthcare industry. In: Global Smart Healthcare Industry. Cheongju: Trend Publishing; Jan 12, 2016:1-12.
8. Gourraud PA, Henry RG, Cree BA, Crane JC, Lizee A, Olson MP, et al. Precision medicine in chronic disease management: the multiple sclerosis BioScreen. Ann Neurol 2014;76(5):633-642 [FREE Full text] [doi: 10.1002/ana.24282] [Medline: 25263997]
9. Garcia-Perez A, Cegarra-Navarro JG, Sallos MP, Martinez-Caro E, Chinnaswamy A. Resilience in healthcare systems: cyber security and digital transformation. Technovation 2023;121:102583. [doi: 10.1016/j.technovation.2022.102583]
10. Adams SC, Petersen C. Precision medicine: opportunities, possibilities, and challenges for patients and providers. J Am Med Inform Assoc 2016;23(4):787-790 [FREE Full text] [doi: 10.1093/jamia/ocv215] [Medline: 26977101]
11. Duffy DJ. Problems, challenges and promises: perspectives on precision medicine. Brief Bioinform 2016;17(3):494-504. [doi: 10.1093/bib/bbv060] [Medline: 26249224]
12. Jung IY, Goo WM. Data integration for healthcare innovation. Science Technology Policy. 2018. URL: https://www.ibric.org/bric/trend/bio-report.do?mode=view&articleNo=8692173&title=#!/list [accessed 2023-06-20]
13. Kim KH, Ko MS, Kim H. A study of establishment of medical CRM model in the post-corona era: focusing on the primary-level hospital. J Inf Technol Appl Manag 2021:28-12. [doi: 10.21219/jitam.2021.28.1.001]
14. Hopp WJ, Li J, Wang G. Big data and the precision medicine revolution. Prod Oper Manag 2018;27(9):1647-1664. [doi: 10.1111/poms.12891]
15. Dal Mas F, Massaro M, Rippa P, Secundo G. The challenges of digital transformation in healthcare: an interdisciplinary literature review, framework, and future research agenda. Technovation 2023;123:102716. [doi: 10.1016/j.technovation.2023.102716]
16. Santarsiero F, Schiuma G, Carlucci D, Helander N. Digital transformation in healthcare organisations: the role of innovation labs. Technovation 2023;122:102640. [doi: 10.1016/j.technovation.2022.102640]
17. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. Health Inf Sci Syst 2014;2:3 [FREE Full text] [doi: 10.1186/2047-2501-2-3] [Medline: 25825667]
18. Lee D, Kim S. Digital healthcare innovation trends and policy implications. Science and Technology Policy Institute. 2018. URL: https://eiec.kdi.re.kr/policy/domesticView.do?ac=0000144167 [accessed 2023-06-13]
19. Wang L, Alexander CA. Big data analytics in medical engineering and healthcare: methods, advances and challenges. J Med Eng Technol 2020;44(6):267-283. [doi: 10.1080/03091902.2020.1769758] [Medline: 32498594]
20. Lee EJ. Applications of blockchain technology in healthcare: some notable cases abroad. HIRA. 2020. URL: https://repository.hira.or.kr/handle/2019.oak/2306 [accessed 2023-07-24]
21. Lim GC. Status and challenges of health and welfare data utilization. KIHASA 2020 Aug 01;8:2-5. [doi: 10.23062/2020.08.1]
22. Report for big data market. Big data platform newsletter. K-ICT Big Data Center. 2019. URL: https://kbig.kr/portal/kbig/knowledge/files.page [accessed 2023-07-15]
23. Hulsen T, Jamuar SS, Moody AR, Karnes JH, Varga O, Hedensted S, et al. From big data to precision medicine. Front Med (Lausanne) 2019;6:34 [FREE Full text] [doi: 10.3389/fmed.2019.00034] [Medline: 30881956]
24. Choi H, Choi YJ, Choi JM, Yoon SY. Big data education process of university optimal in the age of 4th Industrial Revolution. J Korea Institute Inf Commun Eng 2020;24:440-441.
25. Song YA. A case study on the big data analysis curriculum for the efficient use of data. J Pract Eng Educ 2020;12:23-29. [doi: 10.14702/JPEE.2020.023]
26. Park YS, Lee SJ. Study on the direction of universal big data and big data education-based on the survey of big data experts. J Korean Assoc Inf Educ 2020;24:201-214. [doi: 10.14352/jkaie.2020.24.2.201]
27. Karpathakis K, Libow G, Potts HW, Dixon S, Greaves F, Murray E. An evaluation service for digital public health interventions: user-centered design approach. J Med Internet Res 2021;23(9):e28356 [FREE Full text] [doi: 10.2196/28356] [Medline: 34494965]
28. Stickdorn M, Hormess ME, Lawrence A, Schneider J. This Is Service Design Doing: Applying Service Design Thinking in the Real World. Sebastopol, CA: O'Reilly Media, Inc; 2018.
29. Babione JN, Ocampo W, Haubrich S, Yang C, Zuk T, Kaufman J, et al. Human-centred design processes for clinical decision support: a pulmonary embolism case study. Int J Med Inform 2020;142:104196. [doi: 10.1016/j.ijmedinf.2020.104196] [Medline: 32947115]
30. Micheaux A, Bosio B. Customer journey mapping as a new way to teach data-driven marketing as a service. J Market Educ 2018;41(2):127-140. [doi: 10.1177/0273475318812551]
31. Holzinger A, Kargl M, Kipperer K, Regitnig P, Plass M, Muller H. Personas for artificial intelligence (AI) an open source toolbox. IEEE Access 2022 Feb 25;10(2):23732-23747. [doi: 10.1109/ACCESS.2022.3154776] [Medline: 20044162]

XSL•FO
RenderX

32. Canedo ED, Calazans ATS, Silva GRS, Costa PHT, Masson ETS. Use of journey maps and personas in software requirements elicitation. Int J Soft Eng Knowl Eng 2023;33(3):313-342. [doi: 10.1142/s0218194023300014]
33. Lee YH, Jang YJ, Lee SK. Obstacles to health big data utilization based on the perceptions and demands of health care workers in South Korea: web-based survey study. JMIR Form Res 2023;7:e45913 [FREE Full text] [doi: 10.2196/45913] [Medline: 37052992]
34. Glaser BA, Strauss AL. The Discovery of Grounded Theory: Strategies for Qualitative Research. 2nd Edition. London: A Division of Transaction Publishers; 2017.
35. Dash S, Shakyawar SK, Sharma M, Kaushik S. Big data in healthcare: management, analysis and future prospects. J Big Data 2019;6(1):1-25. [doi: 10.1186/s40537-019-0217-0]
36. The guideline for big data analysis and usage version 1.2. K-ICT Big Data Center. 2014. URL: https://kbig.kr/portal/ [accessed 2024-12-19]
37. Graneheim UH, Lundman B. Qualitative content analysis in nursing research: concepts, procedures and measures to achieve trustworthiness. Nurse Educ Today 2004;24(2):105-112. [doi: 10.1016/j.nedt.2003.10.001] [Medline: 14769454]
38. Williams I, Brereton M, Donovan J, Mcdonald K, Millard T, Tam A, et al. A collaborative rapid persona-building workshop: creating design personas with health researchers. Int J Sociotechnol Knowl Dev 2014;6(2):17-35. [doi: 10.4018/ijskd.2014040102]
39. Miaskiewicz T, Kozar KA. Personas and user-centered design: how can personas benefit product design processes? Des Stud 2011;32(5):417-430. [doi: 10.1016/j.destud.2011.03.003]
40. Meyer C, Schwager A. Understanding customer experience. Harvard Business Rev 2007;85:116 [FREE Full text]
41. Yang B, Liu Y, Liang Y, Tang M. Exploiting user experience from online customer reviews for product design. Int J Inform Manag 2019;46:173-186. [doi: 10.1016/j.ijinfomgt.2018.12.006]
42. He Q, Du F, Simonse LW. A patient journey map to improve the home isolation experience of persons with mild COVID-19: design research for service touchpoints of artificial intelligence in eHealth. JMIR Med Inform 2021;9(4):e23238. [doi: 10.2196/23238] [Medline: 33444156]
43. Kang HJ, Han J, Kwon GH. Determining the intellectual structure and academic trends of smart home health care research: coword and topic analyses. J Med Internet Res 2021;23(1):e19625 [FREE Full text] [doi: 10.2196/19625] [Medline: 33475514]
44. Easton K, Potter S, Bec R, Bennion M, Christensen H, Grindell C, et al. A virtual agent to support individuals living with physical and mental comorbidities: co-design and acceptability testing. J Med Internet Res 2019;21(5):e12996 [FREE Full text] [doi: 10.2196/12996] [Medline: 31148545]
45. Kim B. Big data platform case analysis and deployment strategies to revitalize the data economy. Converg Secur J 2021:21-78. [doi: 10.33778/kcsa.2021.21.1.073]
46. McPadden J, Durant TJ, Bunch DR, Coppi A, Price N, Rodgerson K, et al. Health care and precision medicine research: analysis of a scalable data science platform. J Med Internet Res 2019;21(4):e13043 [FREE Full text] [doi: 10.2196/13043] [Medline: 30964441]
47. Kim CW. Digital healthcare platform initiatives also designed from the perspective of users, not providers. Digital Health Innov 2016;215(2):116-126 [FREE Full text]
48. Zafeiridi P, Paulson K, Dunn R, Wolverson E, White C, Thorpe JA, et al. A web-based platform for people with memory problems and their caregivers (CAREGIVERSPRO-MMD): mixed-methods evaluation of usability. JMIR Form Res 2018;2(1):e4 [FREE Full text] [doi: 10.2196/formative.9083] [Medline: 30684403]
49. Shea B, Bakre S, Carano K, Scharen J, Langheier J, Hu EA. Changes in glycemic control among individuals with diabetes who used a personalized digital nutrition platform: longitudinal study. JMIR Diabetes 2021;6(4):e32298 [FREE Full text] [doi: 10.2196/32298] [Medline: 34661545]
50. Iorfino F, Ho N, Carpenter JS, Cross SP, Davenport TA, Hermens DF, et al. Predicting self-harm within six months after initial presentation to youth mental health services: a machine learning study. PLoS One 2020;15(12):e0243467 [FREE Full text] [doi: 10.1371/journal.pone.0243467] [Medline: 33382713]
51. Yan X, Yan J. Design and implementation of interactive platform for operation and maintenance of multimedia information system based on artificial intelligence and big data. Comput Intell Neurosci 2022;2022:4620930 [FREE Full text] [doi: 10.1155/2022/4620930] [Medline: 35571710]
52. Kneuertz PJ, Jagadesh N, Perkins A, Fitzgerald M, Moffatt-Bruce SD, Merritt RE, et al. Improving patient engagement, adherence, and satisfaction in lung cancer surgery with implementation of a mobile device platform for patient reported outcomes. J Thorac Dis 2020;12(11):6883-6891 [FREE Full text] [doi: 10.21037/jtd.2020.01.23] [Medline: 33282391]
53. Torous J, Wisniewski H, Bird B, Carpenter E, David G, Elejalde E, et al. Creating a digital health smartphone app and digital phenotyping platform for mental health and diverse healthcare needs: an interdisciplinary and collaborative approach. J Technol Behav Sci 2019;4(2):73-85 [FREE Full text] [doi: 10.1007/s41347-019-00095-w]

Abbreviations

AI: artificial intelligence

XSL•FO
RenderX

HBD: health big data

Edited by A Mavragani; submitted 07.10.24; peer-reviewed by A Holzinger, L He; comments to author 15.11.24; revised version received 28.11.24; accepted 28.11.24; published 08.01.25.

Please cite as:
Lee YH, Choi H, Lee SK
Development of Personas and Journey Maps for Artificial Intelligence Agents Supporting the Use of Health Big Data: Human-Centered Design Approach
JMIR Form Res 2025;9:e67272
URL: https://formative.jmir.org/2025/1/e67272
doi:10.2196/67272
PMID:

©Yoon Heui Lee, Hanna Choi, Soo-Kyoung Lee. Originally published in JMIR Formative Research (https://formative.jmir.org), 08.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effect of a Narrative-Based Online Course Aimed at Reducing Stigma Toward Transgender Children and Adolescents: Longitudinal Observational Study

Merlin Greuel[1], MD, MGH; Van Kinh Nguyen[1], PhD; Doron Amsalem[2], MD; Maya Adam[1,3,4*], MD, PhD; Till Bärnighausen[1,5,6*], MD, MSci, MSci, SCD, Dr med

[1]Heidelberg Institute of Global Health, Heidelberg University, Heidelberg, Germany
[2]Department of Psychiatry, Irving Medical Center, Columbia University, New York, NY, United States
[3]Department of Pediatrics, Stanford School of Medicine, Stanford University, Stanford, CA, United States
[4]Center for Digital Health, Stanford School of Medicine, Stanford University, Stanford, CA, United States
[5]Department of Global Health and Population, Harvard T H Chan School of Public Health, Harvard University, Cambridge, MA, United States
[6]Africa Health Research Institute, Somkhele, South Africa
[*]these authors contributed equally

Corresponding Author:
Merlin Greuel, MD, MGH
Heidelberg Institute of Global Health
Heidelberg University
Im Neuenheimer Feld 130/3
Heidelberg, 69120
Germany
Phone: 49 6221565344
Email: greuel@stud.uni-heidelberg.de

Abstract

Background: Stigma toward transgender children and adolescents negatively impacts their health and educational outcomes. Contact with members of stigmatized groups can dismantle stereotypes and reduce stigma by facilitating exposure to the unique cognitive and emotional perspectives of individuals within the group. Recent evidence suggests that video-based contact interventions can be as effective as face-to-face encounters, but challenges lie in protecting the identities of transgender youth, since many of them live in stealth.

Objective: This study aims to evaluate the impact of an animated online course, rooted in authentic, personal narratives, on course participants' stigma toward transgender youth.

Methods: The online course was offered free of charge on Coursera and contained 19 teaching videos (3-7 minutes each), intermittent practice quizzes, and discussion prompts. Using real voice recordings of transgender children and their caregivers, the videos were designed to elicit empathy and transmit knowledge. All videos conveying the narratives of transgender youth were animated to protect their identities. A total of 447 course participants, distributed around the globe, completed pre- and postcourse surveys. While the course primarily targeted parents and caregivers of transgender youth, it was open to anyone with a Coursera account. The survey was based on the Transgender Attitudes and Beliefs Scale but modified to reflect the context of parents and caregivers. Using a 5-point Likert scale, it contained 5 questions that captured participants' levels of transgender stigma. Results of the pre- and postcourse surveys were then compared.

Results: The results were obtained in January 2023. Baseline levels of stigma were relatively low (18/25 across all questions, with 25 representing the lowest possible levels of stigma) and decreased further after completion of the course (to 19/25 across all questions, $P<.001$). A multivariate ordinal probit regression showed that, depending on the question, participants were 7%-34% more likely to endorse statements that indicated the lowest levels of stigma after completing the course. The course was equally effective across all demographics represented in our participant population.

Conclusions: Our findings document a significant reduction in stigma toward transgender youth in participants who chose to enroll in the first animated, open online gender health course, rooted in the authentic narratives of transgender youth. Stigma levels decreased significantly after taking the course, even among participants whose baseline levels of stigma were low. Future

XSL•FO
RenderX

interventions should include participants with more variable baseline levels of stigma, ideally in the setting of a randomized controlled trial. Despite its limitations, this evaluation adds to the existing evidence that digital, contact-based antistigma interventions, animated to protect the identity of the narrators, can effectively reduce stigma toward transgender youth.

(JMIR Form Res 2025;9:e59605) doi:10.2196/59605

KEYWORDS

stigma; transgender; children; adolescents; mental health; online course evaluation; entertainment-education; narratives

Introduction

Adolescence is a period marked by change, self-discovery, and significant vulnerability to mental health issues [1]. Children and young adults who realize that their gender identity differs from their sex assigned at birth are especially prone to poor mental health outcomes: transgender youth endure elevated rates of depression, self-harm, eating disorders, and suicidality compared with their nontransgender peers [2-4]. In addition, the educational outcomes of gender-diverse children often lag behind those of their cisgender peers [5]. Thus, the stressors that affect gender-diverse children and adolescents can have a profound effect on individuals' academic performance and the course of their future lives.

Stigma toward transgender youth contributes to the observed differences in mental health. According to the gender minority stress theory, many of the health disparities observed between transgender and cisgender individuals can be attributed to stressors emerging from the stigma faced by transgender people, also called "transphobia" [6,7]. The chronic stress associated with transphobia results in poor mental health outcomes with repercussions even on physical health, as evidenced by elevated rates of diabetes, hypertension, and death by suicide [7]. This is especially true for individuals with multiple disadvantaged statuses such as low-income, transgender women of color [7,8]. Stigma can affect individuals' health directly, for example, through chronic activation of the body's stress response and increased cortisol output [7], or indirectly by decreasing access to the housing market, education, and employment [7,9], all critical social determinants of health.

Depending on the academic discipline, stigma has been studied from different angles, using diverse terminology [7,9]. Stigma may be defined as "a severe social disapproval due to believed or actual individual characteristics, beliefs or behaviors that are against norms, be they economic, political, cultural, or social" [10]. More recently, neuroscience has offered a fresh perspective on stigma, describing its cognitive, emotional, and behavioral dimensions [11-13]. Stigma can be considered a defensive reaction resulting from an innate fear of individuals who differ from the self and collectively form an "outgroup" [11]. Functional neuroimaging studies show that stigma [14,15] and perceived outgroup membership [16,17] can decrease the levels of empathy felt for others. As a consequence, members of the outgroup are depersonalized and viewed as more homogeneous than ingroup members, thus encouraging prejudice, negative stereotyping, and antisocial behaviors [18].

Contact with a member of a stigmatized community can dismantle stereotypes and facilitate entry of the previously stigmatized person into the in-group [19,20]. Intergroup contact is considered the most effective means of reducing stigma [20,21]. Experimental evidence suggests that video-based contact interventions may be as effective as face-to-face encounters while being less expensive and easier to disseminate [20,22-24]. The conceptual framework of Jankowski et al [20] for social contact-based video interventions highlights the importance of addressing the cognitive, emotional, and behavioral dimensions of stigma [20]. By targeting all of these components, viewers can empathize and identify with the protagonists, thereby reducing intergroup anxiety [20] and with it the perceived barriers between groups.

By allowing the public to experience the cognitive and emotional perspectives of transgender youth, authentic, personal narratives lend themselves to antistigma interventions [18]. Narratives can provide a nonthreatening context to establish vicarious social contact with others who might previously have been met with anxiety or mistrust. Narrative stories have been shown to elicit more compassion, favorable attitudes, and beneficial behavioral intentions toward stigmatized groups when compared with nonnarrative stories [25]. High capacity for engagement, as well as the compassionate and affective reactions evoked by the narrative format, are thought to trigger empathic attitudes in the audience [25] that may translate into nonstigmatizing behavior. Through empathy and knowledge transfer, transgender individuals may leave the "outgroup" and enter the "ingroup," allowing their nontransgender peers to view and treat them as one of their own. Yet, few studies have explored the use of video-based narratives for reducing the stigma against transgender youth. These studies found that social contact–based videos have the potential to increase knowledge about transgender youth and to decrease transphobia [26,27]. However, care needs to be taken to protect the identities of transgender youth who share their stories, since many of them live in stealth [28]. To our knowledge, no study has examined the effect of integrating animated short videos, featuring authentic voices and narratives, into an online course aimed at reducing transphobia. Thus, this article contributes to the existing literature by integrating evidence-based methods to reduce transgender stigma into a novel, digitally scalable format that protects the identities of the narrators. The benefit of an online course over interventions with exposure to a single video is the possibility to expose participants to content over a longer time period, enabling a more in-depth discussion of course material, repetitions of core concepts, and assessments that stimulate subjects to actively engage with the topic of interest.

Here, we evaluate the effect of the online course "Health Across the Gender Spectrum," rooted in entertainment narratives, on participants' attitudes toward transgender youth. The aim of the

XSL•FO
RenderX

course was to promote a gender-inclusive society by establishing virtual contact with the transgender community. The narratives were unique in that they included real voice recordings of transgender children as well as their parents and caregivers and they were animated to conceal the identities of the storytellers. The course was made freely available to the public, with some targeting toward parents and caregivers of children whose gender identity differs from their sex assigned at birth. This target group was chosen because adolescents' "coming out" as transgender can significantly deteriorate family relations [2] due to a lack of acceptance of family members. Family rejection of transgender children and adolescents is strongly correlated with mental health issues, substance use, suicidality, and sexual risk [29-31]. On the other hand, family support of transgender youth protects against depression and improves their quality of life [32]. We analyzed pre- and postcourse survey data to determine whether participants' attitudes toward transgender youth changed after completing the course. Our hypothesis was that subjects would show lower levels of stigma toward transgender youth upon completion of the course. This evaluation may help pave the way for other innovative, empirically verified interventions to reduce family and caregiver stigma toward transgender individuals, thus improving their health and life outcomes [29-32].

Methods

Evaluation Design and Approval

This online course evaluation was designed as a virtual contact-based exposure study with a pre-post design, using a convenience sample. The CHERRIES (Checklist for Reporting Results of Internet E-Surveys) [33] was used to guide this article.

Online Course and Participants

The course "Health Across the Gender Spectrum," still available free of charge on the Coursera online platform as of publication [34], was set up as a 3-week program (8 hours of instruction in total) consisting of 3 modules: "What is Gender Identity?" "What is the Gender Spectrum?" and "How Do We Create a Gender-Inclusive Society?" Coursera is a global massive open online course provider based in the United States and used by universities and similar organizations to offer online courses, certificate programs, and degree courses, some of which are free of charge while others require a tuition fee. Participants had no fixed time schedule and could decide the pace of their progress on their own. The language of instruction was English. Conceived by the Stanford School of Medicine under the leadership of our coauthor (MA), learners were introduced to the experiences of 6 transgender children and their families. Through illustrated stories and 19 short teaching videos, participants explored the topics of gender identity and the gender spectrum. Stanford physicians, K-12 educators, and transgender faculty members offered practical advice for parents, health care providers, teachers, and others interested in exploring gender identity in children and adolescents.

Apart from a Coursera account, there were no prerequisites for enrollment. That is, while the course was geared toward parents and caregivers of transgender and gender-diverse children and adolescents, anyone was eligible to enroll. The course was announced to learners enrolled in the Stanford Child Nutrition and Cooking Course, also available on Coursera. Coursera also independently announces new courses to registered account holders. Each of the 3 modules was divided into subsections that sought to answer common questions such as "What's the difference between gender identity and sexual orientation?" The answers were then provided by a series of short, animated educational videos, interspersed with the narrative videos. Each subsection contained one practice quiz with one or two true-false questions, which required a response to move on to the next video. Before advancing to the next subsection, learners had the option to answer a voluntary discussion prompt. At the end of the last section of each module, participants were asked to complete a graded quiz of 4-6 questions. A score of 80% or higher was required to move on. Learners could repeat the quiz as often as they wished. An overview of all modules, sections, and titles of the videos and discussion prompts can be found in Multimedia Appendix 1.

We did not store any participant information other than the survey responses submitted by the learners. We had access to anonymized data from Coursera, such as the IP addresses of those who submitted surveys. These data did not personally identify any participants. Subjects were informed about the purpose of the surveys and their approximate length (3 minutes).

Video Animations

A total of 19 teaching videos (3-7 minutes each) were developed by our coauthor (MA) at the Stanford School of Medicine, with input from faculty at the Pediatric and Adolescent Gender Clinic. The videos featured universally accessible animations, appropriate for all cultural, ethnic, and racial contexts. This raised the likelihood that viewers could identify with the protagonists no matter their own cultural identity and origin. In some of the videos, MA interviewed faculty experts from the Stanford Pediatric Gender Clinic as well as 2 transgender faculty members, who appeared in person. The transgender faculty members shared personal anecdotes and experiences relevant to the course theme. However, depending on the 3 modules of the course, the animated content predominated and ranged between 75% and 100%. The narrative videos' hallmark was their story-telling approach, using the real voices of transgender children, parents, and caregivers to make the animations come alive. Rather than being didactic, these videos were designed to establish vicarious social contact between the viewers and transgender children as well as their families and caregivers, in a nonthreatening context. The use of animation effectively protected the identity of the young people who shared their stories and allowed us to visually emphasize the importance of diversity and inclusion in all vignettes. The emotional appeal of the video narratives relied heavily on the characters' descriptions of their feelings, which were mirrored in the animations and underscored by customized soundtracks. The aim was to enable viewers to experience the cognitive and emotional perspectives of the protagonists, thereby triggering compassionate and affective reactions and empathy. Some of the videos focused more on individuals' personal experiences, while others transmitted knowledge about gender identity and related topics. The language was simple and avoided medical jargon to reach viewers with diverse educational backgrounds.

XSL•FO
RenderX

Figure 1 shows select scenes from the narrative videos. A complete list of course videos can be found in Multimedia Appendix 2.

Figure 1. Sample scenes from narrative course videos with titles and run-times.

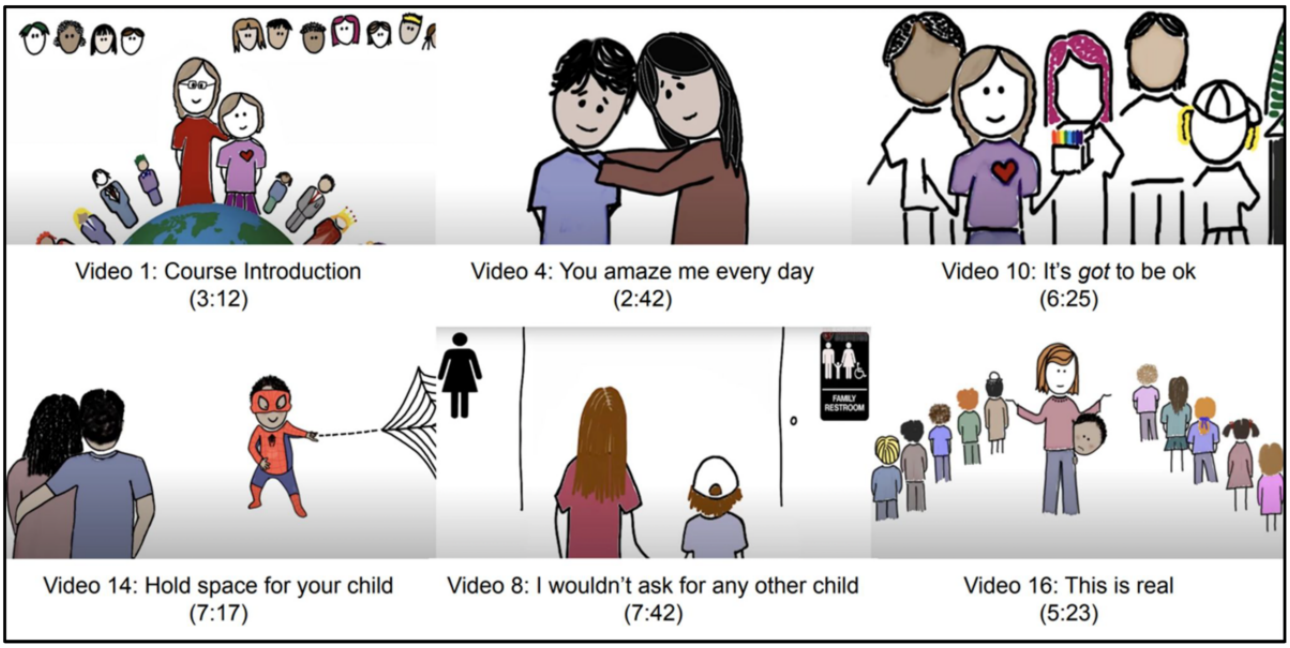

Survey

After enrollment, participants received an email in which they were asked to complete an open precourse survey. The survey was hosted by Stanford School of Medicine's Qualtrics account and appeared on a single page. Demographic information was collected on the second page. Qualtrics is a platform for online experiments and data collection [35]. To match pre- and postcourse surveys, we used Coursera's user ID hashtags. In addition, participants were asked to generate a unique survey ID as a backup. We also used IP addresses, locations, demographic information, and dates of submission to identify any falsely matched surveys or multiple submissions. The questionnaire was informed by the Transgender Attitudes and Beliefs Scale (TABS) [36]. We chose to abbreviate and modify the TABS for the context of parents and caregivers because the original TABS contains 29 items, which we deemed too long for a survey taken on a voluntary basis, with no financial incentives or compensation. Our questionnaire contained 5 items aimed at measuring preexisting stigma against transgender children and adolescents. The statements were based on the first 2 subscales of the TABS, which explore interpersonal comfort with transgender individuals and sex or gender beliefs. We modified the items to better reflect the perspective of parents and caregivers of transgender children and adolescents. For example, one of the items in the original TABS reads "All adults should identify as either male or female" (Q3.13, subscale 2) [36]. We changed this item to "It is important that all children learn to identify themselves as either a boy or a girl" (item 5 in our questionnaire). Participants responded through a 5-point Likert scale and an additional not applicable option. The answer choices ranged from "strongly disagree" to "strongly agree." Please refer to Multimedia Appendix 3 for the complete questionnaire, and to Multimedia Appendix 4 for an item-by-item overview of how the original TABS items were modified. The technical functionality of the survey was tested before its launch. The data were collected between 2017 and 2023.

After completing the course, learners received another email asking them to fill out the same questionnaire, thus allowing for pre-post comparisons. In both surveys, participants were instructed to generate a unique, anonymous identifier to match the surveys in the analysis.

Data Analysis

Whenever possible, the pre- and postcourse surveys were matched for each participant, creating a fully matched sample of individuals. We used a multivariate ordinal probit regression for matched data, in which the 5 questions were modeled jointly [37]. This approach treats the Likert response as an ordinal scale providing a better theoretical interpretation and statistical inferences (compared with a nonparametric or normally distributed response assumption, eg, in a change point analysis) [38]; furthermore, potential correlations between the questions are taken into account [39]. Matched individuals were modeled with a random intercept to allow for individual heterogeneity (Multimedia Appendix 5). An ordinal probit regression can model the relationship between one ordinal-dependent variable (in our case, the answer choices of the Likert scale) and one or more independent variables [40]. A multivariate ordinal probit model also takes into account the potential correlation among the responses across the 5 questions [37]. First, we defined that an answer was in the "positive" direction when it reflected low or very low levels of transgender stigma. For the first 3 questions, a "positive" answer meant that participants agreed (score of 4) or strongly agreed (score of 5) with the respective statements; for the last 2 survey items, a response in the "positive" direction was reflected by respondents' disagreeing (score of 4) or strongly disagreeing (score of 5) with the

XSL•FO
RenderX

statements. We then calculated the cumulative probabilities for each Likert category in the pre- and postcourse surveys; in other words, we determined the likelihood that a subject chose a specific answer (Likert category) for each question. Finally, we calculated the ratio between the probabilities for each question-and-answer choice (Likert category) in the post- versus precourse survey. This ratio describes the relative effect of the intervention. For instance, a relative effect (ratio) of "1" indicates that respondents' likelihood of choosing a respective answer choice after completing the course was just as high as before the course. A relative effect of "0.5" means that participants were half as likely to choose a respective answer after completing the course, compared with the precourse condition. Conversely, a relative effect of "2" indicates that subjects were twice as likely to select a respective answer choice after the course.

Since respondents often included erroneous identifiers, not all surveys could be matched. Thus, as a robustness check, we performed a secondary analysis on a propensity score 1:1 matched sample [41] (Multimedia Appendix 6). We used a probit regression as described above. The matching procedure used nearest neighbor matching and sampling with replacement [42].

Ethical Considerations

We obtained ethics approval from Stanford University's institutional review board (protocol #76972). Completion of the survey was optional and not a prerequisite for beginning the course, and no incentives were offered to participants. The data collected were anonymized and did not allow for the personal identification of subjects.

Results

Demographics

By January 2023, 3878 learners completed the course. Of these, we obtained 1591 survey records. A total of 447 participants (12% of all 3878 learners) could be fully matched based on subjects' unique survey IDs. Coursera's user-ID hashtags were not used due to a technical issue on the part of the platform. We also used data such as IP addresses, location, submission dates, and demographic information to identify potential mismatches or multiple submissions between pre- and postcourse surveys. If we found any mismatch or multiple submissions, we did not include the submission. Of the 447 learners, 377 (84%) were female, 406 (91%) had college-level education, and 283 (63%) identified as White. A total of 229 (51%) of participants completed the survey in 2017. A total of 200 participants (45%) completed both, the precourse and the postcourse survey within one day. Table 1 shows an overview of the subjects' social demographics for the matched sample. Figure 2 depicts the geographic distribution of all recorded participants, containing both pre- and postsurvey data.

XSL•FO
RenderX

Table 1. Descriptive social demographics and responses of the fully matched sample.

Characteristics	Values
Age (years), mean (SD)	37.29 (14.85)
Sex, n (%)	
Female	377 (0.84)
Male	47 (0.11)
Gender-diverse	23 (0.05)
Education, n (%)	
College	406 (0.91)
High school	41 (0.09)
Race, n (%)	
Asian	45 (0.10)
Black	14 (0.03)
Latino	79 (0.18)
Mixed or other	26 (0.06)
White	283 (0.63)
US-based, n (%)	
No	280 (0.63)
Yes	167 (0.37)
Year of survey completion, n (%)	
2017	229 (0.51)
2018	55 (0.12)
2019	17 (0.04)
2020	70 (0.16)
2021	37 (0.08)
2022	28 (0.06)
2023	11 (0.02)
Time between pre- and postsurveys, n (%)	
Day	200 (0.45)
Week	66 (0.15)
Month	129 (0.29)
Month or more	52 (0.12)
Total score, median (IQR)[a]	
Precourse	18 (14-20)
Postcourse	19 (16-20)

[a]Wilcoxon signed rank *P* value <.001.

XSL•FO
RenderX

Figure 2. Geographic distribution of the sample, showing both pre- and postcourse participants, including those that could not be matched regarding their pre- and postcourse surveys.

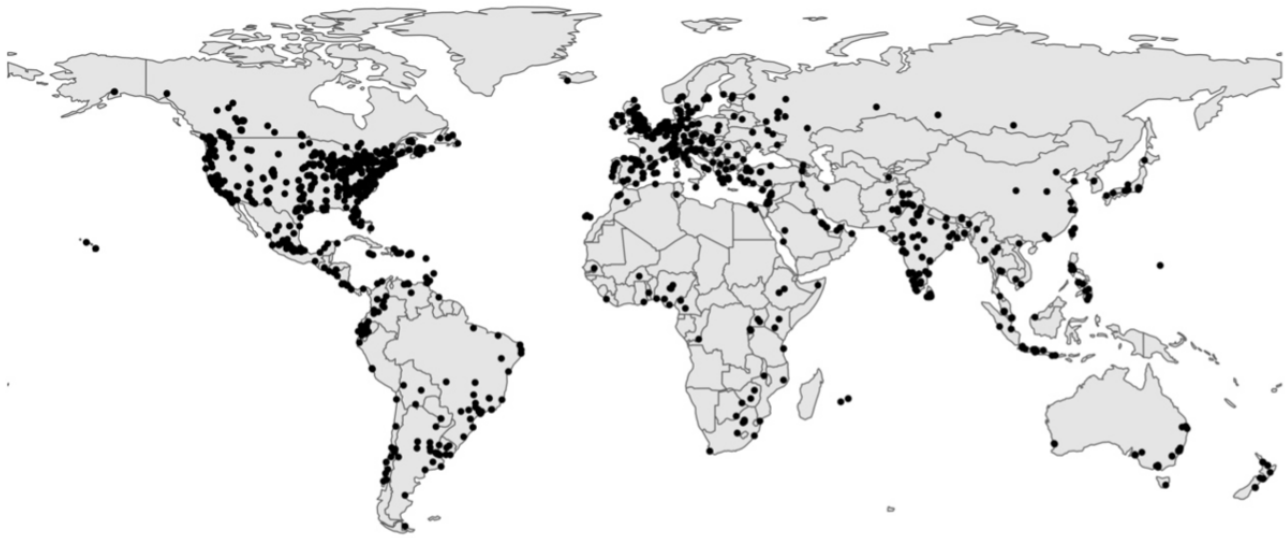

Of the participants in the fully matched sample, 167 (36%) were located in the United States, the country most highly represented in the course. The US-based subjects were distributed across the country, following a pattern that roughly reflects the country's population density (Figure 2). However, 280 (63%) learners were located outside of the United States, distributed across all continents with larger clusters in Europe and Latin America.

Levels of Stigma

The 5 questions had a Cronbach α of 0.83, reflecting the questionnaire's high level of internal consistency. Given that each of the 5 questions allowed for a maximum possible stigma score of 5 (and a minimum of 1), the maximum possible score for all 5 questions combined (indicating the lowest possible levels of stigma) was 25. The median of the total score across all 5 questions before the course was 18 (IQR 14-20). This median increased significantly to 19 (IQR 16-20) after the course (Wilcoxon signed rank test, $P<.001$).

Figure 3 shows the effect of the course on the fully matched data, showing separate results for each of the survey items. After the course, participants had a higher likelihood of answering very positively (strong agreement with the first 3 statements, strong disagreement with the remaining 2 items). At the same time, they were also less likely to provide negative, neutral, or less positive answers. The effect of the course varied largely between individual participants. While acknowledging that our sample was quite homogeneous in terms of demographic variables (refer to Table 1 and Discussion below), we found no significant differences in efficacy across the demographics represented in our participant population. The results of the secondary analysis on a propensity score 1:1 matched sample were consistent with those presented above (Multimedia Appendix 6).

XSL•FO
RenderX

Figure 3. Marginal effect of the course on the responses to the 5 questions. The relative effect was computed as the ratio between the postcourse and the precourse predicted probabilities for each survey item.

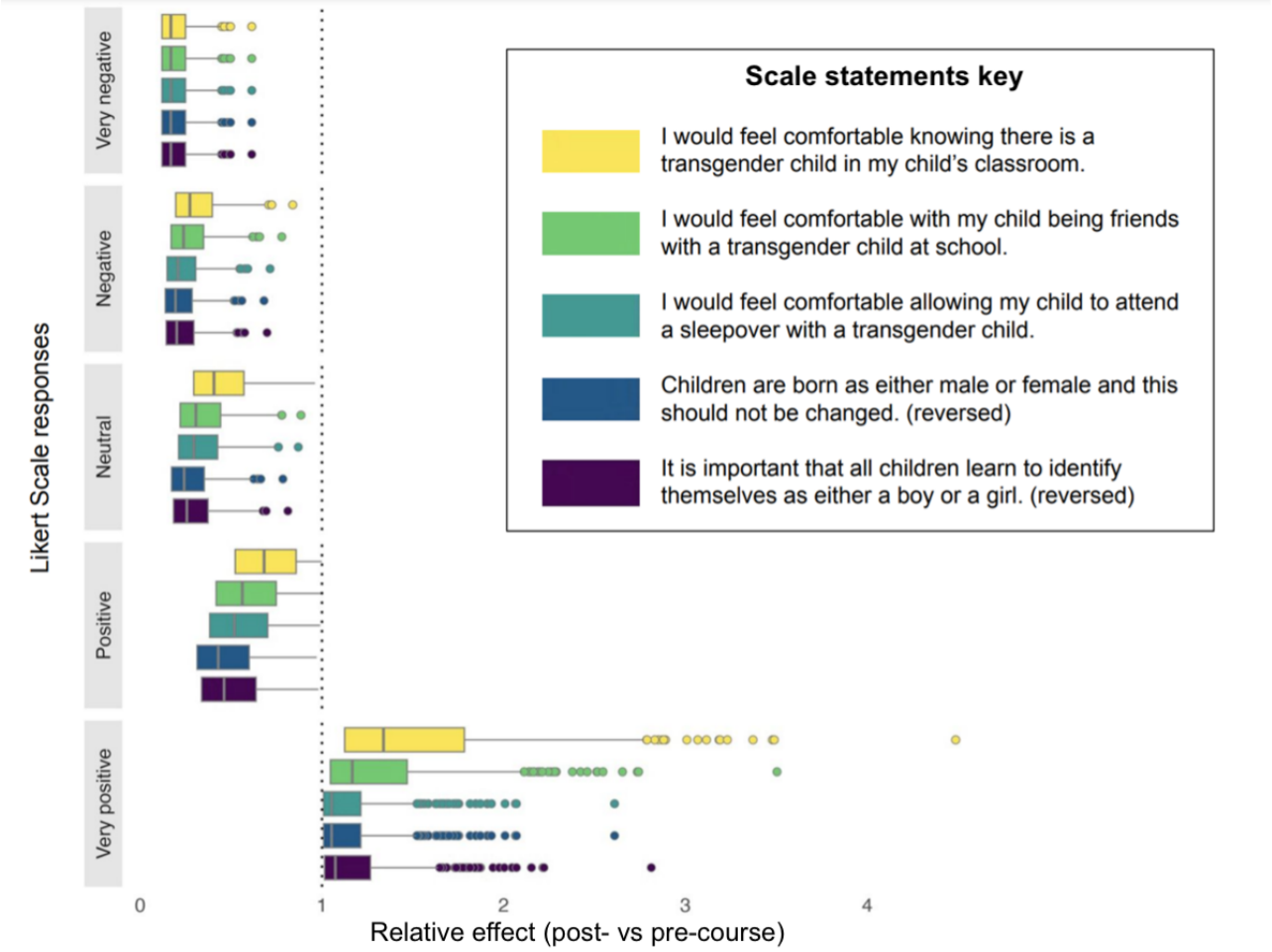

Discussion

Principal Findings

In this survey-based evaluation of the social contact–based online course Health Across the Gender Spectrum, we observed a significant decrease in participants' stigma toward transgender children and adolescents. Baseline levels of stigma were low but decreased further upon completion of the course. This effect was observed for all survey items and across all demographic groups. Despite its limitations, the findings of this evaluation provide the first documented evidence that an online course rooted in animated narratives may effectively reduce participants' stigma toward transgender youth. In addition, the fact that the online course reached subjects across all levels of education, gender, age, ethnicity, or location is another important revelation of this study with the potential to guide future interventions.

The course significantly reduced stigma toward transgender children and adolescents across all question items, despite low baseline stigma levels. After completion of the course, participants were more likely to respond "very positively" and less likely to choose answers in the other categories. This means that there was a general shift in responses toward the "very positive" category, which reflected the lowest levels of stigma. This effect was observed for all questions but especially pronounced for the statements regarding the innateness of gender and the importance that children learn to identify as a boy or a girl. These 2 statements were arguably the most complex to understand for a learner who had not been exposed to the course. Moreover, they represented the highest level of escalation together with the question regarding the acceptability of sleepovers with transgender children, meaning that they had been designed to maximize the discomfort of participants with high levels of stigma. The fact that the strongest change was observed for these statements implies that subjects with higher preexisting levels of stigma or less informed participants may have benefited the most from the course.

Not surprisingly, learners who enrolled in the course showed relatively low levels of transgender stigma even before completing the intervention. This makes sense because the course likely attracted learners with socially progressive attitudes and an openness to learn more about transgender individuals. This is a limitation of this study, as individuals with higher levels of transphobia are presumably less likely to enroll in this course. Nonetheless, the fact that we were able to document a measurable reduction in stigma, even among this enlightened group, suggests that the stigma-reducing effect of the course could be even greater in populations that are less accepting of gender diversity in children. Future studies, ideally in the form of randomized controlled trials, may explore the stigma-reducing effect of this course or similar content within a more

XSL•FO
RenderX

heterogeneous sample that is representative of the population as a whole.

The findings of our evaluation confirm previous research that demonstrated the stigma-reducing effect of social contact–based antistigma interventions [26,27]. This article adds to the literature by demonstrating that an online course with real learners (rather than participants recruited in a controlled experimental setting) achieved similar results. Furthermore, this is the first open online gender health course to experiment with the use of animation to protect the identities of transgender youth sharing authentic stories. Research in the field of neuroscience may help to explain why the course was effective. Stigma is thought to encompass a cognitive, emotional, and behavioral dimension [11-13]. The power of social contact–based video interventions is that they can address all of these components [20]. Emotion is a core feature of stigma, which often manifests as reflexive disgust, fear, and disapproval [11]. The brain regions responsible for emotional responses, the amygdala and insula, generally have a shorter latency to respond to aversive stimuli (eg, those inducing disgust) than cortical regions [11,43,44]. Thus, with the emotional response preceding cognitive modulation, emotions play a primary role in shaping individuals' proneness to exert stigmatizing behavior.

Direct contact with a member of a stigmatized community can reduce the negative emotions associated with stigma [20]. However, virtual contact may also be effective, as exemplified by an intervention that established intergroup contact with individuals enduring schizophrenia using an electronic chat [20,45]. Participants who had been in virtual contact with a schizophrenia patient exhibited lower levels of anger, fear, and stereotyping compared with those who had been in virtual contact with a person who did not suffer from a mental illness [45]. In fact, even just imagining a positive encounter with a schizophrenia patient can reduce anxiety and negative stereotypes [46], demonstrating the power of social contact—be it direct, virtual, or imagined—in shaping our emotions. In a similar way, the video-based virtual contact in our online course might have alleviated the negative emotions associated with transgender stigma, thereby contributing to the reduction in stigma scores observed in the postsurvey.

While emotions are key, stigma is more than just an affective response. It also includes cognition. Implicit stigma describes attitudes and biases that are outside of conscious control [47,48]. It is impulsive and depends on associative links [11]. In contrast, explicit stigma occurs within conscious awareness and is self-reportable and regulatable through cognitive processes [11,48]. Explicit stigma relies on factual knowledge and reflection [11]. Thus, an effective contact-based online course should also transmit knowledge. Contact-based interventions have proven effective when they portray characters that deviate from negative stereotypes in moderate ways; that is, by focusing on the characters' struggles and stories of hope instead of their accomplishments [20]. This is precisely what our videos do. In the video segments, animated children (narrated by real transgender youth) describe their journeys—their initial doubts about their gender identities, their feelings of shame and guilt, their fear of "coming out" and all of its consequences. Interviews with 2 transgender Stanford faculty role models give caregivers and parents hope by sharing how they successfully navigate the challenges of being transgender in academia. Finally, short clips embedded between the video sequences of transgender experiences also transmit facts about the concepts of gender identity, sexual orientation, the gender spectrum, the medical classification of transgender, gender-affirming management options, and the creation of a gender-inclusive society. A lack of subject knowledge supports the formation of stigma [49], so the aim of these supplementary videos was to provide facts that were easily comprehensible to a global audience. Thus, the effect of our video-based online course may have been underscored by the fact that its contact-based approach addressed both the emotion and cognition underlying stigma. The interaction between emotion and cognition gives rise to the concept of empathy—the ability to be sensitive to, comprehend, and vicariously experience the thoughts and feelings of others.

Fear of individuals who differ from the self are categorized internally into an "outgroup" [11]. Those who exert stigmatizing attitudes view "them" as decisively different from "us" [50,51]. This assumption prevents individuals from identifying with others: people use their self-knowledge, mental states, and experiences to infer what others might think or feel [50]; that is, to engage in the theory of mind [52]. If others are deemed fundamentally different, self-referential thinking is reduced [50], and with it, empathy. A neurobiological correlate of stigma is a reduction in the activity of the medial prefrontal cortex [52], a part of the brain that is thought to modulate empathy as it is involved in judgments about others and in viewing others as "human" [52]. As a consequence, members of the outgroup are depersonalized and viewed as more homogeneous than ingroup members, thus encouraging prejudice, negative stereotyping, and antisocial behaviors [18].

Since neural projections connect the amygdala and insula (responsible for emotion) to the ventral medial prefrontal cortex (modulating empathy), it is likely that affective responses are integrated into the theory of mind and empathy processes [52]. Moreover, emotions and empathy influence stigma-related cognition and behavior (and vice versa) [52,53] through a complex interplay involving other areas of the brain: cortical structures such as the temporal lobes and the inferior frontal gyrus encode and store stereotype concepts as semantic or social object (person) memory, which can be selected and transferred into working memory to influence judgments and behaviors [52]. The anterior temporal lobe, particularly involved in the representation of social knowledge (attributes describing people), is densely interconnected with the medial prefrontal cortex (the neural center for empathy) and the amygdala (implicated in emotion) [52,54]. Thus, it becomes evident that the various subcomponents of stigma—implicit versus explicit, with emotional, cognitive, and behavioral elements—can be traced to neurobiological correlates that work in a concerted action, determining individuals' capacity for empathy.

The videos in our course were designed to trigger a strong empathic response in viewers, containing emotionally salient stimuli such as the use of transgender children's original voices that explained their experiences, frustrations, and feelings of distress. Compelling visuals and soundtracks were intended to further support an empathic response in viewers. At the same

XSL•FO
RenderX

time, the knowledge transmitted throughout the course provided participants with the necessary context to understand the medical and social background of transgender adolescents. By acknowledging the multidimensional nature of stigma, the course notably decreased participants' stigma against transgender children and hopefully contributed to reducing the third component of stigma—the behavioral (enacted) one. Future studies designed as more rigorous randomized controlled trials may build on this article and include a behavioral endpoint. Furthermore, the emotional and empathic response of viewers may be tested in laboratory-based fMRI studies or by using online platforms that track viewers' facial reactions through a webcam. The former may elucidate the real-time impact of the intervention on the brain regions involved in stigma in a limited number of individuals, while the latter may be easier to implement in large-scale interventions with a greater sample size.

Limitations

This evaluation has several limitations. First of all, participants were matched ad hoc, resulting in a large number of subjects that could not be matched. Thus, the size of the matched sample represented only a third of the overall number of records. We partly alleviated this issue by performing a secondary analysis using propensity score matching to include more participants. This analysis supported our results. Second, our evaluation lacked a control group. Future studies should include a control group that is not exposed to the course. In addition, the time period between the pre- and postcourse surveys was variable among participants. It is unclear how this may have affected subjects' levels of stigma, although variable completion rates likely reflect the reality of most online courses. Another limitation was that participants enrolled voluntarily in the course. Learners who enrolled in the course likely shared characteristics that differ from the population as a whole, such as holding more socially progressive attitudes or an elevated interest in transgender issues. This is reflected by the fact that our sample was largely female, college-educated, and White. Those who enrolled had low baseline levels of stigma, and the course was likely unappealing to individuals with high levels of transphobia. Future studies should ensure random sampling and the inclusion of participants with diverse attitudes toward the transgender community and more balanced demographic features. Furthermore, as described above, stigma is inherently complex, and the questionnaire used is a simplistic method that might not capture the full dimension of stigma. The questionnaire had not been validated before this evaluation. We found no validated surveys that were appropriate for measuring transphobia toward youth within the parent-child context of the course content. To the best of our knowledge, there are no widely used, validated surveys measuring transphobia toward children and adolescents within this context. Another limitation of this intervention is that it remains unclear how much of the course's effect may be attributed to the narrated, animated content versus the appearance of guest speakers, discussion prompts, or quizzes. Yet, given that the majority of the content was animated, it is likely that the narrated, animated format of the course was the main active ingredient in the outcomes observed here. Finally, our intervention lacked any follow-up assessment. Future studies should include follow-up evaluations to explore if subjects' improvements in transgender stigma are stable over time.

Conclusion

This evaluation of transphobia in online learners enrolled in an animated, story-based course adds to the limited existing evidence that digital, contact-based antistigma interventions can be as effective as face-to-face approaches. Our findings make a case for readily scalable, digital contact-based interventions that use animation to protect the identities of individuals sharing their lived experiences. Ongoing advances in the neuroscience of stigma will further increase our knowledge of how animated educational content can be used to address stigma in historically stigmatized groups. Future interventions may build on the present one and feature more robust samples and study designs. The fact that attitudes toward transgender individuals vary highly across cultures provides an opportunity to assess the efficacy of interventions in diverse cultural contexts. We believe that stigma against transgender youth should be explored in more detail—first and foremost due to the demonstrated burden on the mental health of stigmatized children, who deserve the chance to express themselves authentically and thrive in an accepting society.

Acknowledgments

This study was funded by a faculty teaching innovation grant from the Stanford School of Medicine, a research grant from the Stanford Center for Digital Health, the Alexander von Humboldt Foundation, and the Heidelberg Institute of Global Health. For the publication fee, we acknowledge financial support from Heidelberg University.

Data Availability

The datasets generated during and/or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

MG contributed to the study design, implementation, and analysis, as well as the drafting of this manuscript. MA conceived and designed the online course and all animated content. MA, KN, DA, and TB advised on the study design and data analysis. KN performed the statistical analysis. MG and DA provided context on transgender stigma.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Overview of course content.
[DOCX File , 19 KB - formative_v9i1e59605_app1.docx]

Multimedia Appendix 2
Overview of video content.
[PDF File (Adobe PDF File), 469 KB - formative_v9i1e59605_app2.pdf]

Multimedia Appendix 3
Transgender stigma survey.
[PNG File , 345 KB - formative_v9i1e59605_app3.png]

Multimedia Appendix 4
Adaptation of the Transgender Attitudes and Beliefs Scale (TABS) to the context of parents and caregivers.
[DOCX File , 13 KB - formative_v9i1e59605_app4.docx]

Multimedia Appendix 5
Pre- and post-course survey matching based on the unique identifier.
[DOCX File , 16 KB - formative_v9i1e59605_app5.docx]

Multimedia Appendix 6
Pre- and post-course survey matching using a propensity score approach.
[DOCX File , 249 KB - formative_v9i1e59605_app6.docx]

References

1. Blakemore SJ. Adolescence and mental health. Lancet 2019;393(10185):2030-2031. [doi: 10.1016/S0140-6736(19)31013-X] [Medline: 31106741]
2. Clark TC, Lucassen MF, Bullen P, Denny SJ, Fleming TM, Robinson EM, et al. The health and well-being of transgender high school students: results from the New Zealand Adolescent Health Survey (Youth'12). J Adolesc Health 2014;55(1):93-99 [FREE Full text] [doi: 10.1016/j.jadohealth.2013.11.008] [Medline: 24438852]
3. Connolly MD, Zervos MJ, Barone CJ, Johnson CC, Joseph CL. The mental health of transgender youth: advances in understanding. J Adolesc Health 2016;59(5):489-495. [doi: 10.1016/j.jadohealth.2016.06.012] [Medline: 27544457]
4. Reisner SL, Vetters R, Leclerc M, Zaslow S, Wolfrum S, Shumer D, et al. Mental health of transgender youth in care at an adolescent urban community health center: a matched retrospective cohort study. J Adolesc Health 2015;56(3):274-279 [FREE Full text] [doi: 10.1016/j.jadohealth.2014.10.264] [Medline: 25577670]
5. Wilkinson L, Shifrer D, Pearson J. Educational outcomes of gender-diverse youth: a national population-based study. Gend Soc 2021;35(5):806-837 [FREE Full text] [doi: 10.1177/08912432211038689] [Medline: 34602743]
6. Hendricks ML, Testa RJ. A conceptual framework for clinical work with transgender and gender nonconforming clients: an adaptation of the minority stress model. Prof Psychol Res Pr 2012;43(5):460-467. [doi: 10.1037/a0029597]
7. White Hughto JM, Reisner SL, Pachankis JE. Transgender stigma and health: a critical review of stigma determinants, mechanisms, and interventions. Soc Sci Med 2015;147:222-231 [FREE Full text] [doi: 10.1016/j.socscimed.2015.11.010] [Medline: 26599625]
8. Denise EJ. Multiple disadvantaged statuses and health: the role of multiple forms of discrimination. J Health Soc Behav 2014;55(1):3-19. [doi: 10.1177/0022146514521215] [Medline: 24578393]
9. Major B, O'Brien LT. The social psychology of stigma. Annu Rev Psychol 2005;56:393-421. [doi: 10.1146/annurev.psych.56.091103.070137] [Medline: 15709941]
10. Lauber C. Stigma and discrimination against people with mental illness: a critical appraisal. Epidemiol Psichiatr Soc 2011;17(1):10-13. [doi: 10.1017/s1121189x0000261x]
11. Almeida OFX, Sousa N. Leveraging neuroscience to fight stigma around mental health. Front Behav Neurosci 2021;15:812184 [FREE Full text] [doi: 10.3389/fnbeh.2021.812184] [Medline: 35295248]
12. Rössler W. The stigma of mental disorders: a millennia-long history of social exclusion and prejudices. EMBO Rep 2016;17(9):1250-1253 [FREE Full text] [doi: 10.15252/embr.201643041] [Medline: 27470237]
13. Thornicroft G, Rose D, Kassam A, Sartorius N. Stigma: ignorance, prejudice or discrimination? Br J Psychiatry 2007;190:192-193. [doi: 10.1192/bjp.bp.106.025791] [Medline: 17329736]

XSL•FO
RenderX

14. Decety J, Echols S, Correll J. The blame game: the effect of responsibility and social stigma on empathy for pain. J Cogn Neurosci 2010;22(5):985-997. [doi: 10.1162/jocn.2009.21266] [Medline: 19425830]
15. Azevedo RT, Macaluso E, Viola V, Sani G, Aglioti SM. Weighing the stigma of weight: an fMRI study of neural reactivity to the pain of obese individuals. Neuroimage 2014;91:109-119. [doi: 10.1016/j.neuroimage.2013.11.041] [Medline: 24287441]
16. Hein G, Silani G, Preuschoff K, Batson CD, Singer T. Neural responses to ingroup and outgroup members' suffering predict individual differences in costly helping. Neuron 2010;68(1):149-160 [FREE Full text] [doi: 10.1016/j.neuron.2010.09.003] [Medline: 20920798]
17. Xu X, Zuo X, Wang X, Han S. Do you feel my pain? Racial group membership modulates empathic neural responses. J Neurosci 2009;29(26):8525-8529 [FREE Full text] [doi: 10.1523/JNEUROSCI.2418-09.2009] [Medline: 19571143]
18. Chung AH, Slater MD. Reducing stigma and out-group distinctions through perspective-taking in narratives. J Commun 2013;63(5):894-911. [doi: 10.1111/jcom.12050]
19. Corrigan P. The Stigma Effect. New York, NY: Columbia University Press; 2019.
20. Jankowski SE, Yanos P, Dixon LB, Amsalem D. Reducing public stigma towards psychosis: a conceptual framework for understanding the effects of social contact based brief video interventions. Schizophr Bull 2023;49(1):99-107 [FREE Full text] [doi: 10.1093/schbul/sbac143] [Medline: 36190348]
21. Reinke RR, Corrigan PW, Leonhard C, Lundin RK, Kubiak MA. Examining two aspects of contact on the stigma of mental illness. J Soc Clin Psychol 2004;23(3):377-389. [doi: 10.1521/jscp.23.3.377.35457]
22. Amsalem D, Valeri L, Jankowski SE, Yang LH, Bello I, Nossel I, et al. Reducing public stigma toward individuals with psychosis across race and gender: a randomized controlled trial of young adults. Schizophr Res 2022;243:195-202. [doi: 10.1016/j.schres.2022.03.011] [Medline: 35397250]
23. Amsalem D, Yang LH, Jankowski S, Lieff SA, Markowitz JC, Dixon LB. Reducing stigma toward individuals with schizophrenia using a brief video: a randomized controlled trial of young adults. Schizophr Bull 2021;47(1):7-14 [FREE Full text] [doi: 10.1093/schbul/sbaa114] [Medline: 33484269]
24. Amsalem D, Markowitz JC, Jankowski SE, Yang LH, Valeri L, Lieff SA, et al. Sustained effect of a brief video in reducing public stigma toward individuals with psychosis: a randomized controlled trial of young adults. Am J Psychiatry 2021;178(7):635-642. [doi: 10.1176/appi.ajp.2020.20091293] [Medline: 33900809]
25. Oliver MB, Dillard JP, Bae K, Tamul DJ. The effect of narrative news format on empathy for stigmatized groups. J Mass Commun Q 2012;89(2):205-224. [doi: 10.1177/1077699012439020]
26. Martin A, Celentano J, Olezeski C, Halloran J, Penque B, Aguilar J, et al. Collaborating with transgender youth to educate healthcare trainees and professionals: randomized controlled trial of a didactic enhanced by brief videos. BMC Public Health 2022;22(1):2427 [FREE Full text] [doi: 10.1186/s12889-022-14791-5] [Medline: 36572857]
27. Amsalem D, Halloran J, Penque B, Celentano J, Martin A. Effect of a brief social contact video on transphobia and depression-related stigma among adolescents: a randomized clinical trial. JAMA Netw Open 2022;5(2):e220376 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.0376] [Medline: 35212749]
28. Olson KR, Durwood L, DeMeules M, McLaughlin KA. Mental health of transgender children who are supported in their identities. Pediatrics 2016;137(3):e20153223 [FREE Full text] [doi: 10.1542/peds.2015-3223] [Medline: 26921285]
29. Newcomb ME, LaSala MC, Bouris A, Mustanski B, Prado G, Schrager SM, et al. The influence of families on LGBTQ youth health: a call to action for innovation in research and intervention development. LGBT Health 2019;6(4):139-145 [FREE Full text] [doi: 10.1089/lgbt.2018.0157] [Medline: 30844341]
30. Bouris A, Guilamo-Ramos V, Pickard A, Shiu C, Loosier PS, Dittus P, et al. A systematic review of parental influences on the health and well-being of lesbian, gay, and bisexual youth: time for a new public health research and practice agenda. J Prim Prev 2010;31(5-6):273-309. [doi: 10.1007/s10935-010-0229-1] [Medline: 21161599]
31. Ryan C, Huebner D, Diaz RM, Sanchez J. Family rejection as a predictor of negative health outcomes in white and Latino lesbian, gay, and bisexual young adults. Pediatrics 2009;123(1):346-352. [doi: 10.1542/peds.2007-3524] [Medline: 19117902]
32. Simons L, Schrager SM, Clark LF, Belzer M, Olson J. Parental support and mental health among transgender adolescents. J Adolesc Health 2013;53(6):791-793 [FREE Full text] [doi: 10.1016/j.jadohealth.2013.07.019] [Medline: 24012067]
33. Eysenbach G. Improving the quality of web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004;6(3):e34 [FREE Full text] [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
34. Coursera. Stanford. Health across the gender spectrum. 2017. URL: https://www.coursera.org/learn/healthgenderspectrum?dl=XdOxMRjKxyKR4AnAzJ1rUkCVWB1p7xuA0w08sgw=1&utm_medium=partners&utm_source=impact&utm_campaign=319988&utm_content=b2&gclsrc=1 [accessed 2024-12-06]
35. QualtricsXM. URL: http://www.qualtrics.com [accessed 2024-09-05]
36. Kanamori Y, Cornelius-White JHD, Pegors TK, Daniel T, Hulgus J. Development and validation of the transgender attitudes and beliefs scale. Arch Sex Behav 2017;46(5):1503-1515. [doi: 10.1007/s10508-016-0840-1] [Medline: 27571743]
37. Schauberger G, Tutz G. Multivariate ordinal random effects models including subject and group specific response style effects. Stat Model 2021;22(5):409-429. [doi: 10.1177/1471082x20978034]
38. Bürkner PC, Vuorre M. Ordinal regression models in psychology: a tutorial. Adv Methods Pract Psychol Sci 2019;2(1):77-101. [doi: 10.1177/2515245918823199]

XSL•FO
RenderX

39. DeYoreo M, Kottas A. Bayesian nonparametric modeling for multivariate ordinal regression. J Comput Graph Stat 2017;27(1):71-84. [doi: 10.1080/10618600.2017.1316280]
40. Ananth CV, Kleinbaum DG. Regression models for ordinal responses: a review of methods and applications. Int J Epidemiol 1997;26(6):1323-1333. [doi: 10.1093/ije/26.6.1323] [Medline: 9447413]
41. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983;70(1):41-55. [doi: 10.1093/biomet/70.1.41]
42. Ho DE, Imai K, King G, Stuart EA. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Polit Anal 2017;15(3):199-236. [doi: 10.1093/pan/mpl013]
43. Krendl AC, Kensinger EA, Ambady N. How does the brain regulate negative bias to stigma? Soc Cogn Affect Neurosci 2012;7(6):715-726 [FREE Full text] [doi: 10.1093/scan/nsr046] [Medline: 21896496]
44. Krendl AC, Zucker HR, Kensinger EA. Examining the effects of emotion regulation on the ERP response to highly negative social stigmas. Soc Neurosci 2017;12(3):349-360 [FREE Full text] [doi: 10.1080/17470919.2016.1166155] [Medline: 26982459]
45. Maunder RD, White FA, Verrelli S. Modern avenues for intergroup contact: using E-contact and intergroup emotions to reduce stereotyping and social distancing against people with schizophrenia. Group Process Intergr 2018;22(7):947-963. [doi: 10.1177/1368430218794873]
46. Stathi S, Tsantila K, Crisp RJ. Imagining intergroup contact can combat mental health stigma by reducing anxiety, avoidance and negative stereotyping. J Soc Psychol 2012;152(6):746-757. [doi: 10.1080/00224545.2012.697080] [Medline: 23057193]
47. Stull LG, McGrew JH, Salyers MP, Ashburn-Nardo L. Implicit and explicit stigma of mental illness: attitudes in an evidence-based practice. J Nerv Ment Dis 2013;201(12):1072-1079 [FREE Full text] [doi: 10.1097/NMD.0000000000000056] [Medline: 24284643]
48. Nosek BA, Hawkins CB, Frazier RS. Implicit social cognition: from measures to mechanisms. Trends Cogn Sci 2011;15(4):152-159 [FREE Full text] [doi: 10.1016/j.tics.2011.01.005] [Medline: 21376657]
49. Simmons L, Jones T, Bradley E. Reducing mental health stigma : the relationship between knowledge and attitude change. Eur J Ment Health 2017;12(1):25-40. [doi: 10.5708/ejmh.12.2017.1.2]
50. Shin WG, Woo CW, Jung WH, Kim H, Lee TY, Decety J, et al. The neurobehavioral mechanisms underlying attitudes toward people with mental or physical illness. Front Behav Neurosci 2020;14:571225 [FREE Full text] [doi: 10.3389/fnbeh.2020.571225] [Medline: 33281570]
51. Link BG, Phelan JC. Conceptualizing stigma. Annu Rev Sociol 2001;27(1):363-385. [doi: 10.1146/annurev.soc.27.1.363]
52. Amodio DM. The neuroscience of prejudice and stereotyping. Nat Rev Neurosci 2014;15(10):670-682. [doi: 10.1038/nrn3800] [Medline: 25186236]
53. DeSteno D, Dasgupta N, Bartlett MY, Cajdric A. Prejudice from thin air: the effect of emotion on automatic intergroup attitudes. Psychol Sci 2004;15(5):319-324. [doi: 10.1111/j.0956-7976.2004.00676.x] [Medline: 15102141]
54. Thiebaut de Schotten M, Dell'Acqua F, Valabregue R, Catani M. Monkey to human comparative anatomy of the frontal lobe association tracts. Cortex 2012;48(1):82-96. [doi: 10.1016/j.cortex.2011.10.001] [Medline: 22088488]

Abbreviations

CHERRIES: Checklist for Reporting Results of Internet E-Surveys
TABS: Transgender Attitudes and Beliefs Scale

Edited by A Mavragani; submitted 17.04.24; peer-reviewed by D Shumer, Y Perry; comments to author 14.08.24; revised version received 10.09.24; accepted 10.11.24; published 09.01.25.

Please cite as:
Greuel M, Nguyen VK, Amsalem D, Adam M, Bärnighausen T
Effect of a Narrative-Based Online Course Aimed at Reducing Stigma Toward Transgender Children and Adolescents: Longitudinal Observational Study
JMIR Form Res 2025;9:e59605
URL: https://formative.jmir.org/2025/1/e59605
doi:10.2196/59605
PMID:

©Merlin Greuel, Van Kinh Nguyen, Doron Amsalem, Maya Adam, Till Bärnighausen. Originally published in JMIR Formative Research (https://formative.jmir.org), 09.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The

XSL•FO
RenderX

complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

User Outcomes for an App-Delivered Hypnosis Intervention for Menopausal Hot Flashes: Retrospective Analysis

Katherine Scheffrahn[1], BS; Claire Hall[2], BS; Vanessa Muñiz[1], MA; Gary Elkins[1], PhD

[1]Mind-Body Medicine Lab, Department of Psychology and Neuroscience, Baylor University, Waco, TX, United States

[2]Mindset Health, Melbourne, Australia

Corresponding Author:
Gary Elkins, PhD
Mind-Body Medicine Lab
Department of Psychology and Neuroscience
Baylor University
801 Washington Ave
Waco, TX, 76701
United States
Phone: 1 254 296 0824
Email: gary_elkins@baylor.edu

Abstract

Background: Hypnotherapy has been shown to be a safe, nonhormonal intervention effective for treating menopausal hot flashes. However, women experiencing hot flashes may face accessibility barriers to in-person hypnotherapy. To solve this issue, a smartphone app has been created to deliver hypnotherapy. The Evia app delivers audio-recorded hypnotherapy and has the potential to help individuals experiencing hot flashes.

Objective: This study aims to determine user outcomes in hot flash frequency and severity for users of the Evia app.

Methods: This study is a retrospective analysis of a dataset of Evia app users. Participants were divided into 2 groups for analysis. The first group reported daytime hot flashes and night sweats, while the second group was asked to report only daytime hot flashes. The participants in the first group (daytime hot flashes and night sweats) were 139 women with ≥3 daily hot flashes who downloaded the Evia app between November 6, 2021, and June 9, 2022, with a baseline mean of 8.330 (SD 3.977) daily hot flashes. The participants in the second group (daytime hot flashes) were 271 women with ≥3 daily hot flashes who downloaded the Evia app between June 10, 2022, and February 5, 2024, with a baseline mean of 6.040 (SD 3.282) daily hot flashes. The Evia program included a 5-week program for all participants with daily tasks such as educational readings, hypnotic inductions, and daily hot-flash tracking. The app uses audio-recorded hypnosis and mental imagery for coolness, such as imagery for a cool breeze, snow, or calmness.

Results: A clinically significant reduction, defined as a 50% reduction, in daily hot flashes was experienced by 76.3% (106/139) of the women with hot flashes and night sweats and 56.8% (154/271) of the women with daily hot flashes from baseline to their last logged Evia app survey. On average, the women with hot flashes and night sweats experienced a reduction of 61.4% (SD 33.185%) in their hot flashes experienced at day and night while using the Evia app, and the women with daily hot flashes experienced a reduction of 45.2% (SD 42.567%) in their daytime hot flashes. In both groups, there was a large, statistically significant difference in the average number of daily hot flashes from baseline to end point (women with hot flashes and night sweats: Cohen d=1.28; t_{138}=15.055; P<.001; women with daily hot flashes: Cohen d=0.82; t_{270}=13.555; P<.001).

Conclusions: Hypnotherapy is an efficacious intervention for hot flashes, with the potential to improve women's lives by reducing hot flashes without hormonal or pharmacological intervention. This study takes the first step in evaluating the efficacy of an app-delivered hypnosis intervention for menopausal hot flashes, demonstrating the Evia app provides a promising app delivery of hypnotherapy with potential to increase accessibility to hypnotherapy.

(JMIR Form Res 2025;9:e63948) doi:10.2196/63948

KEYWORDS

hypnosis; hypnotherapy; hot flash; menopause; women's health; gynecology; smartphone app; applications; mobile health; mHealth; user; outcome; intervention; alternative; complementary; mind-body; mobile phone

XSL•FO
RenderX

Introduction

Background

During the menopause transition, approximately 80% of women experience hot flashes, and approximately 20% of women report hot flashes to be very bothersome [1,2]. Hot flashes negatively impact women's quality of life [3], and because the period during which women experience hot flashes lasts an average of 7.4 years, the impacts are long lasting [4].

In the past, the standard treatment for hot flashes, both premenopausal and postmenopausal, has been hormone therapy. However, hormone therapy has been associated with an increased risk of cancer, strokes, and thromboembolism [5-7]. Due to these concerns, many women do not wish to use hormone therapy; moreover, experts do not recommend hormone therapy due to these potential health issues [8].

Most of the options for nonhormonal treatments are either not as efficacious as would be desired or increase the burden on women significantly. Selective serotonin reuptake inhibitor, serotonin-norepinephrine reuptake inhibitor antidepressants, gabapentin, and paroxetine mesylate are nonhormonal medications approved by the US Food and Drug Administration for reducing hot flashes, but all these medications are associated with negative side effects [9-15]. A medication newly approved by the US Food and Drug Administration for menopausal hot flashes, fezolinetant (Veozah), has demonstrated evidence of effectiveness in reducing hot flashes [16,17]. However, fezolinetant's potential side effects include liver injury; to guard against this, patients must undergo routine blood work during the first 9 months of using this drug [17]. This risk of liver injury and the blood work required places undue burden on women seeking a beneficial nonhormonal treatment for hot flashes.

When turning to nonpharmacological, nonhormonal intervention options, a commonly used intervention is cognitive behavioral therapy, which does decrease the level of distress women feel from hot flashes; however, cognitive behavioral therapy does not reduce the number of hot flashes experienced [18,19].

Clinical hypnotherapy is a mind-body intervention that has been shown to be effective in reducing both the frequency and severity of menopausal hot flashes [15,20,21]. Hypnosis is a state of consciousness involving focused attention and reduced peripheral awareness characterized by an enhanced capacity for response to suggestion [22]. Hypnosis is a safe, nonhormonal, nonpharmacological treatment with minimal to no adverse events [15,20,21].

However, face-to-face hypnotherapy for hot flashes has limited accessibility for many women. Challenges such as financial concerns, a lack of proximity to a hypnotherapist, or the time commitment needed for appointments may all factor into women being unable to access hypnotherapy [23,24]. A smartphone app-delivered hypnotherapy intervention for menopausal hot flashes can help overcome these intervention implementation barriers. Mobile delivery of hypnotherapy provides convenient and affordable hypnotherapy for women experiencing hot flashes. With the significant growth of the mobile health industry, smartphone apps have become a commonly used method for intervention dissemination, and this approach can also be used to expand access to hypnotherapy for those in need [25].

To this end, a smartphone app to deliver hypnotherapy for hot flashes has been developed. Created by Mindset Health, the Evia app delivers hypnotic inductions and educational information for hot flashes. The Evia program was developed based on the hypnosis protocol for hot flashes tested in multiple randomized clinical trials [20,21]. The app offers a 5-week program with daily tasks such as educational readings, hypnotic inductions, and hot flash tracking. The app uses audio-recorded hypnotherapy and mental imagery for coolness (ie, imagery for cool breeze, snow, and calmness).

The Evia app holds significant potential to help individuals experiencing hot flashes and has already greatly increased women's access to hypnotherapy for hot flashes due to its availability through smartphone app stores. However, research on the app's effectiveness is lacking. A previous study analyzed Evia app users' characteristics and demographics and demonstrated that the age of Evia app users on average is consistent with the age of menopause onset and the emergence of menopausal symptoms [26]. In addition, this previous study demonstrated that on average app users experienced ≥5 hot flashes per day, a frequency similar to that of participants in previous randomized clinical trials.

Objectives

This study, which will be the first to provide data on Evia app users' outcomes, is a retrospective analysis of data collected from participants who downloaded and used the Evia app. The data, collected by Mindset Health, are from the questionnaire that users fill out before beginning the program, as well as the hot flash logs that users are encouraged to fill out throughout the program. There were 2 groups of participants. The first group was asked about daily hot flashes experienced during the day and night sweats (hot flashes that occur at night), while the second group was asked to report only daily hot flashes (daytime hot flashes). The data from each group were separately analyzed to fulfill this study's aims.

This study aims to (1) determine self-reported user outcomes regarding whether users experienced a clinically significant (50%) reduction in daily hot flashes, (2) determine self-reported user outcomes regarding changes in hot flash severity, (3) describe user demographics and characteristics, (4) describe self-reported Evia app use patterns, (5) determine factors associated with intervention outcomes, and (6) determine the association of self-reported hot flash outcomes with self-reported use patterns.

Methods

Overview

This study is a retrospective, cross-sectional analysis of data from the Evia app collected by its developers, Mindset Health. The data were collected at 1 time point from self-report questionnaires and app use data. See Multimedia Appendix 1 for the Checklist for Reporting Results of Internet E-Surveys (CHERRIES) for this study [27]. All individuals who

XSL•FO
RenderX

downloaded the app were prompted to fill out questionnaires using an open survey method (anyone could download the app and fill out the questionnaire). Evia app users were required by the app design to fill out the baseline questionnaire when they downloaded the app (before beginning the intervention program); the study authors were not provided any information regarding whether there were potential Evia app users who did not complete the baseline questionnaire and therefore did not continue the program. Participants were prompted to optionally fill out hot flash tracker logs throughout the program. All questionnaires were filled out using the Evia app. The data used as the end point measure for this study were derived from the latest hot flash log filled out by each participant. The data were collected from November 6, 2021, to February 5, 2024.

The Evia app offers a 5-week program with daily tasks such as educational readings, hypnotic inductions, and hot flash tracking. The app uses audio-recorded hypnotherapy to deliver suggestions to induce mental imagery for coolness. During the program, users can access the following hypnotherapy audio files: moving snow, cool mountain stream, rock by a lake, cooling blue light, internal thermostat, hillside breeze, snowy mountain path, walking in summer rain, eating a cool treat, cold misty lake, air-conditioned cafe, lying in a hammock, cooling face mask, frosty morning, cold day at the beach, and relaxing in a pool. Each hypnotherapy audio file has a duration of 13 to 15 minutes. App users are encouraged to listen to a hypnotherapy audio file at least once per day.

All questionnaires were developed by Mindset Health for the Evia app, and the questionnaires were in use before the time frame from which this study used data; thus, all questionnaires were provided in a way that was usable for prior app users. Over the course of this study, the user interface and intervention program remained essentially consistent. During the time span of data collection, there was a change within the Evia app in how participants were asked to report their hot flashes. One group was asked to report daily hot flashes and night sweats (hot flashes that occur at night), and a second group was asked to report only daily hot flashes (hot flashes that occur during the day). The 2 groups are discussed further, with an explanation provided for why participants were divided into 2 groups in our analysis, in the Participant Groups subsection.

Ethical Considerations

The study protocol was reviewed by the institutional review board at Baylor University (2167670) and determined to be exempt from review because it qualified as non–human subjects research; therefore, institutional review board approval was not required. The data analyzed were provided in a deidentified form and contained no personal information. Users provided informed consent upon agreeing to the Evia app's terms and conditions, which included notification that their deidentified data may be provided for research focused on aggregated user data rather than individual information.

Participants

The participants are a convenience sample of women who downloaded and used the Evia app between November 6, 2021, and February 5, 2024. Given the contents of the Evia app, the respondents of the survey questionnaire are women experiencing hot flashes. A total of 835 Evia app users completed the initial questionnaire after downloading the app. Of these 835 women, 410 (49.1%) met the inclusion criteria and were included in the data analysis. Participants were included in the final analysis if they met the following criteria: (1) were aged ≥18 years and used the Evia app, defined as listening to >1 hypnotherapy audio file; (2) reported ≥3 daily hot flashes at baseline; and (3) filled out at least 1 hot flash diary log. These criteria ensured that the sample in the analysis only included women who used the Evia app past the initial download of the app, who experienced persistent hot flashes, and who had end point data to compare to baseline data. However, to better understand the differences between the people who downloaded and used the app and those who downloaded and did not use the app, demographic and characteristic data of the excluded users are also reported. Participants were not solicited or recruited for the study because the sample consists of women who downloaded and used the app. Participants provided consent digitally when downloading the Evia app.

After excluding app users based on the eligibility criteria, a box plot (Figure 1) was used to identify outliers based on the hot flash frequency change; this method was chosen because it is a standardized way to identify outliers when the data are not normally distributed. Of the 416 participants, 6 (1.4%) were identified as outliers and removed from the analysis. We chose to remove the statistical outliers because doing so allows for better concentration on the relevant data to fulfill the study aims. This study is meant to provide an understanding of what most women experience while using the Evia app and any patterns that may be present. Removing the statistical outliers ensures that the analysis is not skewed by extreme values. This resulted in 410 (49.1%) of the 835 Evia app users who completed the initial app questionnaire being included as participants.

XSL•FO
RenderX

Figure 1. Simple box plot of hot flash frequency change outliers. After excluding app users based on the eligibility criteria, a box plot was created to view the distribution of the changes in hot flash frequency from baseline to end point. This method revealed 6 participants who were statistical outliers. To best represent the relevant data to answer our study aims, these 6 outliers were removed from the final analysis in a retrospective data analysis on app use and outcomes in women experiencing menopausal hot flashes who downloaded the Evia app between November 6, 2021, and February 5, 2024.

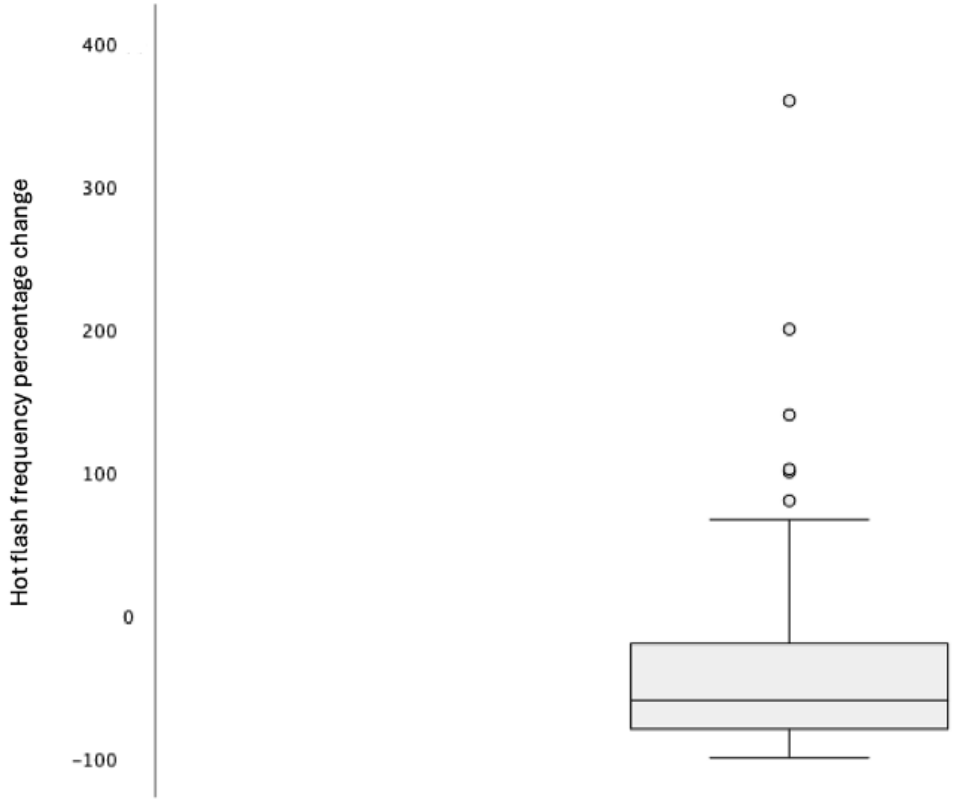

Participant Groups

During the time span of data collection, the Evia app underwent several small changes to the entry questionnaire and the hot flash logs that users are prompted to respond to at the time of app download. When users downloaded the app, they were asked to report how many hot flashes they experienced. Users who downloaded the app up until June 9, 2022, were asked 2 questions about their hot flashes:

- “How many hot flashes do you experience each day?”
- “How many night sweats do you experience during the night?”

After June 9, 2022, users downloading the Evia app were only asked 1 question:

- “How many hot flashes do you experience each day?”

Due to this change in the app survey made by the developers during data collection, the 410 participants were divided into 2 groups: the first group (hot flashes and night sweats) included 139 (33.9%) women, and the second group (daily hot flashes) consisted of 271 (66.1%) women. This was necessary to accurately describe the data. This study reports the results of the 2 groups separately because their baseline hot flashes were measured differently.

Measures

Aim 1

Throughout the Evia program, participants were prompted to log their daily hot flashes by entering a number in the hot flash logs. Participants in both groups were prompted to report the daily number of hot flashes, while participants with hot flashes and night sweats were prompted to report night sweats (hot flashes experienced at night) as well. In the app, the daily hot flash log was optional for users, but due to the aims of this study, only individuals who had filled out at least 1 daily hot flash log were included in the final analysis. For the participants with hot flashes and night sweats, the total number of hot flashes experienced during the day and at night were summed together for each participant at baseline and end point to provide a final number of hot flashes experienced throughout a 24-hour period. The baseline number of daily hot flashes was reported by the user in the entry questionnaire, and the end point number of daily hot flashes was the number indicated by the user on their last recorded hot flash log.

Aim 2

When responding to the baseline questionnaire, before beginning the app-delivered intervention, participants provided the average intensity of their hot flashes. Participants were asked, “How would you rate the intensity of your daily hot flashes (on average)?” The response options included “mild,” “moderate,” “severe,” and “very severe.” This question provided descriptions of what each level of intensity would entail. The participants were given the option of logging their daily hot flash intensity

XSL•FO
RenderX

throughout the 5-week program with the hot flash tracker. The hot flash severity indicated on their last recorded hot flash log was considered the participants' hot flash severity at end point.

Aim 3

In the initial questionnaire, participants filled out questions on demographics and characteristics. The order of the items on the baseline questionnaire was not randomized, and 1 item per page was displayed on the app screen, with a total of 20 pages. Throughout the questionnaire, participants were able to go back and change any answers if needed. Participants were asked to report their age and whether their menstrual cycles were "regular," "irregular but haven't stopped," "stopped within the past year," "or "stopped over a year ago." Participants were also asked, "How would you classify your stage of menopause?" The response options included "perimenopause," "menopause," "postmenopause," and "I'm not sure." Participants were asked whether they were currently taking hormone replacement therapy and were given the options "yes," "yes, but weaning off it," or "no." Participants were asked at baseline, "How do you currently manage your hot flashes?" Answer options included "natural supplements or OTC drug," "other prescribed medications," "hormone therapy (HRT)," "cooling tools (water spray, fans, etc)," "soy consumption (soy isoflavones)," "lifestyle changes (exercise, etc)," "antidepressants (for hot flashes)," and "other." Participants were asked, "How important to you is a nonhormonal tool to help with hot flashes?" The answer options included "extremely important," "somewhat important," and "not important." Participants were asked, "Have you tried hypnotherapy before?" The answer options were "yes" or "no."

Aim 4

Each time an Evia app user listened to a hypnotherapy audio file on the app, the event was logged. To determine Evia app use patterns, this study used these listening logs to determine the number of times an individual listened to a hypnotherapy audio file as well as the number of days on which an individual listened to at least 1 audio file. Only listening events that occurred before each participant's end point were included in the analysis.

Aim 5

The aforementioned characteristics and demographics, including age, menopausal stage, whether they had tried hypnosis before, the importance of a nonhormonal tool, and other intervention methods, were used as measures to determine the characteristics associated with positive outcomes in hot flash reduction.

Aim 6

For this aim, the listening logs for each individual will be used in conjunction with the self-reported initial daily hot flashes and the self-reported daily hot flashes indicated in the end point hot flash tracker.

Data Analyses

In analyzing these data, descriptive statistics (means, SDs, and frequencies) were calculated for the variables of interest. For aim 1, a 2-tailed paired samples *t* test was used. Aim 2 used a Wilcoxon signed rank test. To carry out aims 5 and 6, Spearman rank correlation tests were used to test associations between the variables.

Results

Aim 1: Hot Flash Reduction Outcomes

This study included 410 women who used the Evia app and had ≥3 daily hot flashes. The 410 participants were divided into 2 groups: the first group (hot flashes and night sweats) included 139 (33.9%) women, and the second group (daily hot flashes) consisted of 271 (66.1%) women. At baseline, the women with hot flashes and night sweats experienced a mean of 8.330 (SD 3.977; range 3-24) daily hot flashes, while the women with daily hot flashes experienced a mean of 6.040 (SD 3.282; range 3-20) daily hot flashes. At end point, the women with hot flashes and night sweats experienced a mean of 3.070 (SD 2.883; range 0-15) daily hot flashes, while the women with daily hot flashes experienced a mean of 2.970 (SD 2.377; range 0-15) daily hot flashes. Tables 1 and 2 show the distribution, respectively, of total baseline and endpoint hot flashes reported by the group of participants with daily hot flashes and night sweats. Tables 3 and 4 show the distribution, respectively, of total baseline and endpoint hot flashes reported by the group of participants with daily hot flashes.

A paired samples *t* test indicated that there was a large, statistically significant difference in the average number of daily hot flashes experienced by the women with hot flashes and night sweats from baseline to end point (Cohen d=1.28; t_{138}=15.055; 2-tailed P<.001), as well as in the average number of daily hot flashes experienced by the women with daily hot flashes from baseline to end point (Cohen d=.82, t_{270}=13.555; 2-tailed P<.001).

Clinical significance in the reduction of hot flashes is defined as a reduction of 50% in daily hot flashes [5,28]. In this study, 76.3% (106/139) of the women with hot flashes and night sweats experienced at least a 50% reduction in daily hot flashes from when they completed their entry questionnaire to the latest hot flash log included in the dataset, while 56.8% (154/271) of the women with daily hot flashes experienced a clinically significant 50% reduction in daily hot flashes. The average percentage reduction in daily hot flashes for the women with hot flashes and night sweats was 61.4% (SD 33.185%), while for the women with daily hot flashes, it was 45.2% (SD 42.567%).

XSL•FO
RenderX

Table 1. Distribution of baseline hot flashes for women with daily hot flashes and night sweats (n=139).

Baseline total hot flashes	Participants, n (%)
3	8 (5.8)
4	10 (7.2)
5	21 (15.1)
6	15 (10.7)
7	6 (4.3)
8	13 (9.4)
9	12 (8.6)
10	39 (28.1)
11	1 (0.7)
12	1 (0.7)
13	1 (0.7)
14	0 (0)
15	5 (3.6)
16	0 (0)
17	0 (0)
18	0 (0)
19	0 (0)
20	6 (4.3)
21	0 (0)
22	0 (0)
23	0 (0)
24	1 (0.7)

Table 2. Distribution of end point hot flashes for women with daily hot flashes and night sweats (n=139).

End point total hot flashes	Participants, n (%)
0	22 (15.8)
1	21 (15.1)
2	28 (20.1)
3	22 (15.8)
4	17 (12.2)
5	12 (8.6)
6	3 (2.2)
7	2 (1.4)
8	4 (2.9)
9	0 (0)
10	5 (3.6)
11	0 (0)
12	1 (0.7)
13	0 (0)
14	0 (0)
15	2 (1.4)

XSL•FO
RenderX

Table 3. Distribution of baseline hot flashes with women with daily hot flashes (n=271).

Baseline total hot flashes	Participants, n (%)
3	30 (11.1)
4	35 (12.9)
5	135 (49.8)
6	14 (5.2)
7	3 (1.1)
8	8 (3)
9	0 (0)
10	29 (10.7)
11	0 (0)
12	2 (0.7)
13	0 (0)
14	1 (0.4)
15	8 (3)
16	1 (0.4)
17	0 (0)
18	0 (0)
19	0 (0)
20	5 (1.8)

Table 4. Distribution of end point hot flashes for women with daily hot flashes (n=271).

End point total hot flashes	Participants, n (%)
0	37 (13.7)
1	39 (14.4)
2	68 (25.1)
3	30 (11.1)
4	35 (12.9)
5	21 (7.7)
6	21 (7.7)
7	6 (2.2)
8	19 (7)
9	0 (0)
10	4 (1.5)
11	0 (0)
12	0 (0)
13	0 (0)
14	0 (0)
15	1 (0.4)

Aim 2: Hot Flash Severity Outcomes

Participants were asked to report their hot flash severity at baseline and end point. For participants with hot flashes and night sweats, 2 participants declined to answer regarding their hot flash severity at baseline, 42 (30.7%) reported mild hot flashes, 65 (47.4%) moderate hot flashes, 22 (16.1%) severe hot flashes, and 8 (5.8%) very severe hot flashes. At end point, of the 139 participants with hot flashes and night sweats, 22 (15.8%) reported no hot flash severity due to the absence of hot flashes, 58 (41.7%) reported mild hot flashes, 46 (33.1%)

XSL•FO
RenderX

reported moderate hot flashes, 11 (7.9%) reported severe hot flashes, and 2 (1.4%) reported very severe hot flashes.

At baseline, of the 271 participants with daily hot flashes, 73 (26.9%) reported mild hot flashes, 116 (42.8%) moderate hot flashes, 64 (23.6%) severe hot flashes, and 18 (6.6%) very severe hot flashes. At end point, of the 271 participants with daily hot flashes, 37 (13.7%) reported no hot flash severity due to the absence of hot flashes, 147 (54.2%) reported mild hot flashes, 74 (27.3%) reported moderate hot flashes, and 13 (4.8%) reported severe hot flashes. None reported very severe hot flashes.

For the participants with hot flashes and night sweats, a Wilcoxon signed rank test indicated that there was a significant ($P<.001$) difference between baseline and end point in daily hot flash severity, and the test demonstrated that 48.9% (67/137) of the women experienced a decrease in the severity of their daily hot flashes, 11.7% (16/137) had an increase in severity, and 39.4% (54/137) experienced no change in their daily hot flash severity.

For the participants with daily hot flashes, a Wilcoxon signed rank test indicated that there was a significant ($P<.001$) difference between baseline and end point in daily hot flash severity, and the test demonstrated that 64.2% (174/271) of the women experienced a decrease in the severity of their daily hot flashes, 4.4% (12/271) had an increase in severity, and 31.4% (85/271) experienced no change in their daily hot flash severity.

Aim 3: User Demographics and Characteristics

The mean age of the participants with hot flashes and night sweats was 52.600 (SD 5.577; range 38-72) years. The mean age of the participants with daily hot flashes was 52.400 (SD 5.483; range 30-77) years. When asked to classify their stage of menopause, of the 139 participants with hot flashes and night sweats, 51 (36.7%) indicated that they were unsure, 34 (24.5%) reported that they were perimenopausal, 23 (16.5%) reported that they were menopausal, and 31 (22.3%) reported that they were postmenopausal, while of the 271 participants with daily hot flashes, 96 (35.4%) indicated that they were unsure, 86 (31.7%) reported that they were perimenopausal, 29 (10.7%) reported that they were menopausal, and 60 (22.1%) reported that they were postmenopausal. When asked about their menstrual cycles, of the 77 women with hot flashes and night sweats who responded to this question, 35 (45%) reported that their menstrual cycles had stopped more than a year ago, 18 (23%) reported that they had stopped within the past year, 20 (26%) indicated that their cycles were irregular but had not stopped, and 4 (5%) reported having regular cycles. When asked about their menstrual cycles, of the 121 women with daily hot flashes who responded to this question, 62 (51.2%) reported that their menstrual cycles had stopped more than a year ago, 30 (24.8%) reported that they had stopped within the past year, 27 (22.3%) indicated that their cycles were irregular but had not stopped, and 2 (1.7%) reported having regular cycles.

Regarding using hormone therapy for hot flashes, of the 139 participants with hot flashes and night sweats, 128 (92.1%) were not using hormone therapy, 10 (7.2%) were using hormone therapy, and 1 (0.7%) was using hormone therapy but was weaning off it, while of the 271 participants with daily hot flashes, 234 (86.3%) were not using hormone therapy, 34 (12.5%) were using hormone therapy, and 3 (1.1%) were using hormone therapy but were weaning off it. In answer to the question “How important to you is a nonhormonal tool to help with hot flashes?” of the 63 women with hot flashes and night sweats who provided an answer, 44 (70%) responded “extremely important,” 17 (27%) responded “somewhat important,” and 2 (3%) responded “not important,” while of the 150 women with hot flashes who provided an answer, 104 (69.3%) responded “extremely important,” 39 (26%) responded “somewhat important,” and 7 (4.7%) responded “not important.”

Of the 410 participants in both groups, 209 (51%) provided information at baseline on what treatments they were currently using to manage their hot flashes. Of these 209 women, 111 (53.1%) used cooling tools such as a water spray or fans, 95 (45.5%) adopted lifestyle changes such as exercise, 76 (36.4%) used natural supplements or over-the-counter drugs, 30 (14.4%) consumed soy isoflavones, 14 (6.7%) used antidepressants for hot flashes, 10 (4.8%) took other prescribed medications, and 31 (14.8%) selected “other.” The majority of the participants (136/181, 75.1%) had not tried hypnotherapy before, while the rest (45/181, 24.9%) had tried it before.

Demographics and characteristics surveyed in the baseline questionnaire were also analyzed for the 425 users who downloaded the Evia app between November 6, 2021, and February 5, 2024, but these users were not included in the full data analysis due to not using the Evia app beyond downloading it, not experiencing at least 3 daily hot flashes, or not having end point data. It is important to note that not all individuals in this group filled out all questionnaire items; therefore, the percentages are based on those who provided an answer. The mean age of these users was 51.100 (SD 6.339; range 18-83) years. When asked to classify their stage of menopause, of the 401 users who responded, 118 (29.4%) indicated that they were unsure, 151 (37.7%) reported that they were perimenopausal, 50 (12.5%) reported that they were menopausal, 79 (19.7%) reported that they were postmenopausal, and 3 (0.7%) reported that they were not menopausal. When asked about their menstrual cycles, of the 231 women who responded, 103 (44.6%) reported that their menstrual cycles had stopped more than a year ago, 42 (17.7%) reported that they had stopped within the past year, 61 (26.4%) indicated that their cycles were irregular but had not stopped, and 25 (10.8%) reported having regular cycles. Regarding using hormone therapy for hot flashes, of the 419 users who responded, 362 (86.4%) were not using hormone therapy, 45 (10.7%) were using hormone therapy, and 12 (2.9%) were using hormone therapy but were weaning off it. In answer to the question “How important to you is a nonhormonal tool to help with hot flashes?” of the 166 users who provided an answer, 94 (56.6%) responded “extremely important,” 58 (34.9%) responded “somewhat important,” and 14 (8.4%) responded “not important.” Not all participants responded to this question because it was added to the Evia questionnaire partway through the data collection period. The majority of the participants (96/146, 65.8%) had not tried hypnotherapy before, while the rest (50/146, 34.2%) had tried it before.

XSL•FO
RenderX

Aim 4: Evia App Use Patterns

Evia app use can be evaluated by the number of days on which a user listened to at least 1 hypnotherapy audio file. Participants' end point data, based on their last logged hot flash diary, varied in the time elapsed since baseline (when the app was downloaded). For the participants with hot flashes and night sweats, end point ranged from day 1 (baseline) to day 730, with the average length of time from baseline being 110.31 (SD 152.86) days. For the participants with daily hot flashes, end point ranged from day 1 (baseline) to day 461, with the average length of time from baseline being 67.62 (SD 78.25) days. This differed from the number of days on which the participants used the app. The average number of distinct days on which the participants with hot flashes and night sweats listened to at least 1 audio file was 35.370 (SD 44.638; range 1-307). The average number of hypnotherapy audio files listened to overall by the participants with hot flashes and night sweats was 40.380 (SD 50.741; range 2-344). The average number of distinct days on which the participants with daily hot flashes listened to at least 1 audio file was 27.140 (SD 35.663; range 1-390). The average number of hypnotherapy audio files listened to overall by the participants with daily hot flashes was 31.280 (SD 39.816; range 2-414).

Aim 5: Factors Associated With Outcomes

The factors tested for association with the Evia app users' outcomes were participants' age, menopausal stage, hormone use, and past hypnotherapy use. A Spearman rank correlation test on the data of the participants with hot flashes and night sweats found no significant associations between participants' age, menopausal stage, hormone use, or past hypnotherapy use and the outcomes in hot flash frequency reduction. The results of the correlation test are shown in Table 5.

Table 5. Spearman rank correlation test for user characteristics and hot flash frequency percentage reduction among the participants with hot flashes and night sweats[a].

Characteristics	Correlation coefficient
Age (n=139)	–0.029
Menopausal stage (n=139)	0.027
Hormone use (n=139)	–0.113
Past hypnotherapy use (n=31)	0.070

[a]This table shows the correlational analysis results for the associations between user characteristics and user outcomes in a retrospective data analysis on app use and outcomes in women who experience menopausal hot flashes and downloaded the Evia app between November 6, 2021, and June 9, 2022.

A Spearman rank correlation test on the data of the participants with daily hot flashes found no significant associations between participants' age, hormone use, or past hypnotherapy use and their outcomes in hot flash frequency reduction. However, there was a small, positive, significant correlation between participants' menopausal stage and percentage reduction in hot flashes. The results of the correlation test are shown in Table 6.

Table 6. Spearman rank correlation test for user characteristics and hot flash frequency percentage reduction among the participants with daily hot flashes (n=271)[a].

Characteristics	Correlation coefficient
Age	–0.092
Menopausal stage	0.239[b]
Hormone use	0.047
Past hypnotherapy use	0.051

[a]This table shows the correlational analysis results for the associations between user characteristics and user outcomes in a retrospective data analysis on app use and outcomes in women who experience menopausal hot flashes and downloaded the Evia app between June 10, 2022, and February 5, 2024.

[b]$P<.001$.

Aim 6: Use Pattern Associations With Outcomes

A Spearman rank correlation test to assess the association between app use and percentage reduction in hot flashes among the participants with hot flashes and night sweats found a small, positive, statistically significant association between the number of days on which a user listened to at least 1 hypnotherapy audio file and the percentage reduction in hot flashes from baseline to end point. There was also a small, positive, statistically significant association between the total number of hypnotherapy audio files that users listened to and the percentage reduction in hot flashes from baseline to end point. The results of the correlation test are presented in Table 7.

XSL•FO
RenderX

Table 7. Spearman rank correlation test for app use and hot flash frequency percentage reduction among the participants with hot flashes and night sweats (n=139)[a].

App use	Correlation coefficient
Number of distinct days listened	0.209[b]
Number of audio file listens	0.199[c]

[a]This table shows the correlational analysis results for the associations between user outcomes and users' app use in a retrospective data analysis on app use and outcomes in the women with hot flashes and night sweats who experience menopausal hot flashes and downloaded the Evia app between November 6, 2021, and June 9, 2022. There was a small significant correlation between increased app use and a greater percentage reduction in hot flashes.

[b]P=.01.

[c]P=.02.

A Spearman rank correlation test to assess the association between app use and percentage reduction in hot flashes among the participants with daily hot flashes found no significant association between the number of days on which they listened to at least 1 hypnotherapy audio file and the percentage reduction in hot flashes from baseline to end point. There was also no statistically significant association between the total number of hypnotherapy audio files that users listened to and the percentage reduction in hot flashes from baseline to end point. The results of the correlation test are presented in Table 8.

Table 8. Spearman rank correlation test for app use and hot flash frequency percentage reduction among the participants with daily hot flashes (n=271)[a].

App use	Correlation coefficient
Number of distinct days listened	0.091
Number of audio file listens	0.080

[a]This table shows the correlational analysis results for the associations between user outcomes and users' app use in a retrospective data analysis on app use and outcomes in the women with daily hot flashes who experience menopausal hot flashes and downloaded the Evia app between June 10, 2022, and February 5, 2024. There were no significant correlations between increased app use and a percentage reduction in hot flashes.

Discussion

Principal Findings

The results demonstrated that the majority of Evia app users experienced a clinically significant reduction of at least 50% in their daily hot flashes, with 76.3% (106/139) of the participants with hot flashes and night sweats and 56.8% (154/271) of the participants with daily hot flashes achieving mean reductions of 61.4% and 45.2%, respectively. The difference in daily hot flashes from baseline to end point was statistically significant (P<.001) for both groups. This result demonstrates the potential benefits of the Evia app. Although this result should be interpreted in light of the fact that there was no control group in this dataset to compare with, the majority of women who began using the Evia app saw a decrease in their hot flashes of at least 50%. Using the measure of clinical significance to interpret these results will allow clinicians to provide clear information when discussing how the Evia app might benefit their patients. The participants with hot flashes and night sweats had a greater proportion reaching clinical significance due to more accurate reporting of hot flashes experienced during a 24-hour period (at day and night); thus, this group had a higher reported baseline mean of hot flashes than the participants with daily hot flashes. Although the baseline mean number of daily hot flashes for both groups differed (participants with hot flashes and night sweats: mean 8.330, SD 3.977; participants with daily hot flashes: mean 6.040, SD 3.282) because the participants with daily hot flashes were not being asked to report night sweats separately, both groups had a similar mean number of daily hot flashes at end point (participants with hot flashes and night sweats: mean 3.070, SD 2.883; participants with daily hot flashes: mean 2.970, SD 2.377).

In this study, many Evia app users also experienced decreases in their hot flash severity. Hot flash severity decreased in 48.9% (67/137) of the women with hot flashes and night sweats and 64.2% (174/271) of the women with daily hot flashes. Crucially, participants' severe and very severe hot flashes decreased: of the 410 women across both groups, 86 (21%) reported severe hot flashes at baseline, but only 24 (5.9%) reported severe hot flashes at end point; similarly, at baseline, 26 (6.3%) reported very severe hot flashes, but only 2 (0.5%) reported very severe hot flashes at end point. This result demonstrates that even women who do not experience a clinically significant reduction in their hot flash frequency may still experience significant benefit by a decrease in hot flash severity. The significant decreases in both hot flash frequency and severity are encouraging due to the need for accessible and nonpharmacological interventions for menopausal hot flashes. These results warrant further investigation in a randomized controlled trial of the Evia app to better understand if this is an efficacious mobile delivery method for hypnotherapy in treating menopausal hot flashes.

This study found a small, positive, significant association between app use and hot flash outcomes among the participants with hot flashes and night sweats. This may indicate that the more an individual uses the Evia app hypnotherapy audio files, the greater the percentage reduction in hot flashes. However, the correlation was small; therefore, this result should be

XSL•FO
RenderX

interpreted with caution. In addition, there was not a similarly significant result among the participants with daily hot flashes. App use varied widely between participants (ranging from 1 to 390 unique days), and this study did not evaluate whether participants used the app on consecutive days. As the Evia program is designed to be used daily for 5 weeks, it may be that some participants' lack of consistency made it challenging to understand the effects of use on outcome. A future study may be able to address this by examining within a more standardized study design whether use is associated with hot flash outcomes. Furthermore, this dataset relied on single-day measures of hot flash frequency for both baseline and end point. This may also have impacted the lack of a significant association between use and outcomes because women will naturally have some variability in their hot flash frequency from day to day.

Likewise, the lack of significant associations between participant characteristics such as age, hormone use, and past hypnotherapy use and outcomes is likely due to the lack of standardization across participants. There was a small, positive, significant association between participants' menopausal stage and hot flash reduction among the participants with daily hot flashes, with women further along in menopausal stage being more likely to have greater hot flash reductions. However, this association should be interpreted with caution, given that it is a small correlation within a large sample size, and there was no control group; no significant correlation was found among the participants with hot flashes and night sweats. A future study should further investigate what characteristics might be associated with achieving a clinically significant outcome through a randomized controlled trial design.

When comparing the users who met the inclusion criteria with users who did not continue using the app, the demographics and characteristics were generally comparable between the groups. One difference is that the mean number of daily hot flashes (6.820, SD 3.691) among the included users (n=410) at baseline was higher than the mean number of daily hot flashes (3.310, SD 3.799) among the excluded users (n=425). It may be that those who experienced more daily hot flashes were more likely to persist in using the app beyond the initial download.

Comparisons With Prior Work

In the retrospective study by Snyder and Elkins [26] on the demographics and characteristics of Evia app users, the authors indicated that a surprising number of women (3048/9103, 33.48%) reported that they were unsure of their menopausal stage. Our study found this proportion to be slightly higher when looking at the sample included in the analysis, with 35.9% (147/410) indicating that they were unsure of their menopausal stage. There may be a knowledge gap in the general population of women on the topic of menopause. An increase in educational content on menopausal stages in the Evia app may provide a needed resource for women.

This study found the mean number of daily hot flashes to be 8.330 (SD 3.977) among the participants with hot flashes and night sweats and 6.040 (SD 3.282) among the participants with daily hot flashes at baseline, both of which are lower than those found in an earlier clinical trial conducted by Elkins et al [21], where the authors found the mean number of daily hot flashes to be 10. However, the distribution of the baseline hot flash frequency is consistent with that in previous studies, with more women experiencing ≥5 hot flashes [21,26]. This may be due to the differences in measuring hot flashes between this study and the one by Elkins et al [21]. Our study had a single-item measure, with participants being asked, "How many hot flashes do you experience each day?" In the study by Elkins et al [21], participants tracked their daily hot flashes for a week, which was then averaged, thus providing a more accurate method due to women experiencing variable numbers of hot flashes from day to day. Future studies of the Evia app should explore the use of a weekly hot flash log.

As our study did not have a control group for comparison, we lack the ability to know how a control group's hot flash frequency and severity might have changed over time compared to the intervention group. However, by comparing our findings to those of prior research that included control groups, we can better understand the typical pattern of hot flash severity and frequency over time. Prior research has demonstrated that there is not a large change in hot flash frequency or severity over time for a control group. A 2013 randomized controlled trial examining face-to-face hypnosis for hot flashes found that the control group's hot flash score (a product of hot flash frequency × hot flash severity) only demonstrated a mean reduction of 8.32% compared to the mean reduction of 71.36% in the hot flash score of the hypnosis group [21]. Future studies should examine Evia app users compared to a matched control to understand whether the outcomes are significantly different.

Limitations

A primary limitation of this study concerns the change in the baseline questionnaire that resulted in 1 group of Evia app users not being specifically asked about their night sweats. This meant that the participants with daily hot flashes might have underreported their total number of hot flashes experienced during a 24-hour period (at day and night). However, we still observed that both the participants with hot flashes and night sweats and those with daily hot flashes had a very similar mean number of daily hot flashes at end point (3.070, SD 2.883 and 2.970, SD 2.377, respectively). In addition, since the completion of the data collection for this study, the Evia app has been updated to once again ask users specifically about night sweats, separate from daytime hot flashes.

Another primary limitation of this study is the lack of end point data for information collected in the baseline survey. Due to the retrospective nature of this analysis, participants only completed a baseline survey, after which they filled out hot flash tracker logs throughout the program; as a result, only hot flash frequency and severity could be compared between baseline and end point.

The data collected included the number of times a user listened to a hypnotherapy audio file. Listens were only logged in the app once a user had completed listening to the audio file. Thus, there are no data captured for users who may have partially listened to a hypnotherapy audio file. While the intervention is designed to have users listen to the hypnotherapy audio files in their entirety, future studies might consider including partial listens to the data collected.

XSL•FO
RenderX

In addition, this study is unable to fully assess the impacts of potential participant attrition. Participants who were more motivated to use all elements of the app, such as the hot flash diary logs (which served as end point measures), may have experienced greater benefit. Alternatively, those who benefited from the intervention may have reduced app use due to the absence of ongoing hot flashes. Given this limitation, future replication of this study is warranted.

Finally, this study was also limited by the lack of standardization across participants. Due to the retrospective nature of this analysis, the data included participants with varying levels of Evia app use. Some of the participants completed the 5-week Evia program, while others did not; some of the participants used the Evia app frequently, while others did not. A future study with a greater level of standardization across participants would be able to better provide a clearer understanding of the benefits of the Evia app.

Future Directions

Considering the positive clinical outcomes noted in this study, there is a need for future studies to address this study's limitations by conducting a randomized controlled trial of the Evia app for menopausal hot flashes. A randomized controlled trial will be able to clearly demonstrate whether the Evia app is an effective mobile delivery method for hypnotherapy in treating menopausal hot flashes. The randomized controlled trial should compare the Evia intervention to a placebo control. A future study should administer postintervention measures to provide end point data, ensure that all participants complete the 5-week program to reduce the variability in use among participants, and use weekly hot flash diaries to gain a more accurate measure of hot flash frequency. In addition, a randomized controlled trial of the Evia app should examine when during the 5-week program women begin to see results because this study only examined baseline and end point data.

Conclusions

This study is the first to report the hot flash frequency and severity outcomes of Evia app users. The results demonstrated that 76.3% (106/139) of the Evia app users who were asked about hot flashes during the day and night experienced a clinically significant reduction in their daily hot flashes, while 56.8% (154/271) of the Evia app users who were only asked about daily hot flashes achieved a clinically significant reduction in their daily hot flashes. In addition, 48.9% (67/137) of the women with hot flashes and night sweats and 64.2% (174/271) of the women with daily hot flashes experienced a decrease in their hot flash severity from baseline to end point. This study demonstrates the potential benefits of the Evia app as an accessible and nonpharmacological intervention for menopausal hot flashes. However, additional research is needed to explore the efficacy of the app in randomized controlled trial as well as to better understand the factors that moderate and mediate beneficial therapeutic outcomes.

Acknowledgments

The authors would like to acknowledge Mindset Health and its employees for collaborating in this project. Mindset Health funded the publication cost for this open access article but was not involved in the results or data analysis.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

GE is a consultant to Mindset Health but was not involved in data processing. GE receives a consultant salary from Mindset Health and holds stock options in the company. However, this salary and equity are not dependent on the results of this research. CH is an employee of Mindset Health and receives compensation, however CH was not involved in any of the data analysis of this project. All other authors declare no other conflicts of interest.

Multimedia Appendix 1
Checklist for Reporting Results of Internet E-Surveys (CHERRIES).
[DOCX File , 22 KB - formative_v9i1e63948_app1.docx]

References

1. Gold EB, Colvin A, Avis N, Bromberger J, Greendale GA, Powell L, et al. Longitudinal analysis of the association between vasomotor symptoms and race/ethnicity across the menopausal transition: study of women's health across the nation. Am J Public Health 2006 Jul;96(7):1226-1235. [doi: 10.2105/AJPH.2005.066936] [Medline: 16735636]
2. Kronenberg F. Hot flashes: epidemiology and physiology. Ann N Y Acad Sci 1990;592:52-86; discussion 123-33. [doi: 10.1111/j.1749-6632.1990.tb30316.x] [Medline: 2197954]
3. Stein KD, Jacobsen PB, Hann DM, Greenberg H, Lyman G. Impact of hot flashes on quality of life among postmenopausal women being treated for breast cancer. J Pain Symptom Manage 2000 Jun;19(6):436-445 [FREE Full text] [doi: 10.1016/s0885-3924(00)00142-1] [Medline: 10908824]

XSL•FO
RenderX

4. Avis NE, Crawford SL, Greendale G, Bromberger JT, Everson-Rose SA, Gold EB, et al. Duration of menopausal vasomotor symptoms over the menopause transition. JAMA Intern Med 2015 Apr;175(4):531-539 [FREE Full text] [doi: 10.1001/jamainternmed.2014.8063] [Medline: 25686030]
5. Loprinzi CL, Michalak JC, Quella SK, O'Fallon JR, Hatfield AK, Nelimark RA, et al. Megestrol acetate for the prevention of hot flashes. N Engl J Med 1994 Aug 11;331(6):347-352. [doi: 10.1056/NEJM199408113310602] [Medline: 8028614]
6. North American Menopause Society. The 2012 hormone therapy position statement of: the North American Menopause Society. Menopause 2012 Mar;19(3):257-271 [FREE Full text] [doi: 10.1097/gme.0b013e31824b970a] [Medline: 22367731]
7. Rossouw JE, Anderson GL, Prentice RL, LaCroix AZ, Kooperberg C, Stefanick ML, et al. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results from the Women's Health Initiative randomized controlled trial. JAMA 2002 Jul 17;288(3):321-333. [doi: 10.1001/jama.288.3.321] [Medline: 12117397]
8. Chlebowski RT, McTiernan A. Elements of informed consent for hormone replacement therapy in patients with diagnosed breast cancer. J Clin Oncol 1999 Jan;17(1):130-142. [doi: 10.1200/JCO.1999.17.1.130] [Medline: 10458226]
9. Nelson HD, Vesco KK, Haney E, Fu R, Nedrow A, Miller J, et al. Nonhormonal therapies for menopausal hot flashes: systematic review and meta-analysis. JAMA 2006 May 03;295(17):2057-2071. [doi: 10.1001/jama.295.17.2057] [Medline: 16670414]
10. Shams T, Firwana B, Habib F, Alshahrani A, Alnouh B, Murad MH, et al. SSRIs for hot flashes: a systematic review and meta-analysis of randomized trials. J Gen Intern Med 2014 Jan;29(1):204-213 [FREE Full text] [doi: 10.1007/s11606-013-2535-9] [Medline: 23888328]
11. Sun Z, Hao Y, Zhang M. Efficacy and safety of desvenlafaxine treatment for hot flashes associated with menopause: a meta-analysis of randomized controlled trials. Gynecol Obstet Invest 2013;75(4):255-262. [doi: 10.1159/000348564] [Medline: 23548358]
12. Wei D, Chen Y, Wu C, Wu Q, Yao L, Wang Q, et al. Effect and safety of paroxetine for vasomotor symptoms: systematic review and meta-analysis. BJOG 2016 Oct;123(11):1735-1743. [doi: 10.1111/1471-0528.13951] [Medline: 27062457]
13. Freeman EW, Guthrie KA, Caan B, Sternfeld B, Cohen LS, Joffe H, et al. Efficacy of escitalopram for hot flashes in healthy menopausal women: a randomized controlled trial. JAMA 2011 Jan 19;305(3):267-274 [FREE Full text] [doi: 10.1001/jama.2010.2016] [Medline: 21245182]
14. Yoon SH, Lee JY, Lee C, Lee H, Kim SN. Gabapentin for the treatment of hot flushes in menopause: a meta-analysis. Menopause 2020 Apr;27(4):485-493. [doi: 10.1097/GME.0000000000001491] [Medline: 32049930]
15. Maclaughlan David S, Salzillo S, Bowe P, Scuncio S, Malit B, Raker C, et al. Randomised controlled trial comparing hypnotherapy versus gabapentin for the treatment of hot flashes in breast cancer survivors: a pilot study. BMJ Open 2013 Sep 10;3(9):e003138 [FREE Full text] [doi: 10.1136/bmjopen-2013-003138] [Medline: 24022390]
16. Johnson KA, Martin N, Nappi RE, Neal-Perry G, Shapiro M, Stute P, et al. Efficacy and safety of Fezolinetant in moderate to severe vasomotor symptoms associated with menopause: a phase 3 RCT. J Clin Endocrinol Metab 2023 Jul 14;108(8):1981-1997 [FREE Full text] [doi: 10.1210/clinem/dgad058] [Medline: 36734148]
17. FDA approves novel drug to treat moderate to severe hot flashes caused by menopause. U.S. Food & Drug Administration. 2023 May 16. URL: https://www.fda.gov/news-events/press-announcements/fda-approves-novel-drug-treat-moderate-severe-hot-flashes-caused-menopause [accessed 2024-01-08]
18. Mann E, Smith MJ, Hellier J, Balabanovic JA, Hamed H, Grunfeld EA, et al. Cognitive behavioural treatment for women who have menopausal symptoms after breast cancer treatment (MENOS 1): a randomised controlled trial. Lancet Oncol 2012 Mar;13(3):309-318 [FREE Full text] [doi: 10.1016/S1470-2045(11)70364-3] [Medline: 22340966]
19. Ayers B, Smith M, Hellier J, Mann E, Hunter MS. Effectiveness of group and self-help cognitive behavior therapy in reducing problematic menopausal hot flushes and night sweats (MENOS 2): a randomized controlled trial. Menopause 2012 Jul;19(7):749-759. [doi: 10.1097/gme.0b013e31823fe835] [Medline: 22336748]
20. Elkins G, Marcus J, Stearns V, Perfect M, Rajab MH, Ruud C, et al. Randomized trial of a hypnosis intervention for treatment of hot flashes among breast cancer survivors. J Clin Oncol 2008 Nov 01;26(31):5022-5026 [FREE Full text] [doi: 10.1200/JCO.2008.16.6389] [Medline: 18809612]
21. Elkins GR, Fisher WI, Johnson AK, Carpenter JS, Keith TZ. Clinical hypnosis in the treatment of postmenopausal hot flashes: a randomized controlled trial. Menopause 2013 Mar;20(3):291-298 [FREE Full text] [doi: 10.1097/gme.0b013e31826ce3ed] [Medline: 23435026]
22. Elkins GR, Barabasz AF, Council JR, Spiegel D. Advancing research and practice: the revised APA Division 30 definition of hypnosis. Int J Clin Exp Hypn 2015;63(1):1-9. [doi: 10.1080/00207144.2014.961870] [Medline: 25365125]
23. Hasan SS, Vasant D. The emerging new reality of hypnosis teletherapy: a major new mode of delivery of hypnotherapy and clinical hypnosis training. Int J Clin Exp Hypn 2023;71(2):153-164. [doi: 10.1080/00207144.2023.2185527] [Medline: 37040191]
24. Hussain FA. Hypnotherapy as an adjunct to medical treatment: highlighting effectiveness and identifying barriers to further integrative treatment. J Health Manag 2021 May 07;23(2):185-196. [doi: 10.1177/09720634211011557]
25. mHealth apps market size, share and global growth report. Fortune Business Insights. URL: https://www.fortunebusinessinsights.com/mhealth-apps-market-102020 [accessed 2024-04-10]

XSL•FO
RenderX

26. Snyder M, Elkins GR. Characteristics of users of a digital hypnotherapy intervention for hot flashes: retrospective study. JMIR Form Res 2024 Mar 14;8:e53555 [FREE Full text] [doi: 10.2196/53555] [Medline: 38483465]
27. Eysenbach G. Improving the quality of web surveys: the checklist for reporting results of internet e-surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34 [FREE Full text] [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
28. Sloan JA, Loprinzi CL, Novotny PJ, Barton DL, Lavasseur BI, Windschitl H. Methodologic lessons learned from hot flash studies. J Clin Oncol 2001 Dec 01;19(23):4280-4290. [doi: 10.1200/JCO.2001.19.23.4280] [Medline: 11731510]

Abbreviations

CHERRIES: Checklist for Reporting Results of Internet E-Surveys

Edited by A Mavragani; submitted 03.07.24; peer-reviewed by J Marcus, Z Kekecs; comments to author 06.08.24; revised version received 11.10.24; accepted 24.11.24; published 09.01.25.

Please cite as:
Scheffrahn K, Hall C, Muñiz V, Elkins G
User Outcomes for an App-Delivered Hypnosis Intervention for Menopausal Hot Flashes: Retrospective Analysis
JMIR Form Res 2025;9:e63948
URL: https://formative.jmir.org/2025/1/e63948
doi:10.2196/63948
PMID:

©Katherine Scheffrahn, Claire Hall, Vanessa Muñiz, Gary Elkins. Originally published in JMIR Formative Research (https://formative.jmir.org), 09.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Development and Beta Validation of an mHealth-Based Hearing Screener (SRESHT) for Young Children in Resource-Limited Countries: Pilot Validation Study

Vidya Ramkumar[1], MASLP, PhD; Deepashree Joshi B[1], BASLP, MSc.Aud; Anil Prabhakar[2], PhD; James W Hall[3], PhD; Ramya Vaidyanath[1], PhD

[1]Faculty of Audiology and Speech Language Pathology, Sri Ramachandra Institute of Higher Education and Research, Chennai, India
[2]Department of Electrical Engineering, Indian Institute of Technology Madras, Chennai, India
[3]Osborne College of Audiology, Salus University, Pennsylvania, PA, United States

Corresponding Author:
Deepashree Joshi B, BASLP, MSc.Aud
Faculty of Audiology and Speech Language Pathology
Sri Ramachandra Institute of Higher Education and Research
Sri Ramachandra Nagar, Porur
Chennai, 600116
India
Phone: 91 9894170609
Email: deepasrijoshi25@sriramachandra.edu.in

Abstract

Background: The prevalence of hearing loss in infants in India varies between 4 and 5 per 1000. Objective-based otoacoustic emissions and auditory brainstem response have been used in high-income countries for establishing early hearing screening and intervention programs. Nevertheless, the use of objective screening tests in low- and middle-income countries (LMICs) such as India is not feasible. Mobile health (mHealth) solutions have been demonstrated to be a viable option for hearing screening in LMICs.

Objective: This study aims to develop and beta-validate an affordable hearing screener for children younger than 6 years of age to identify moderately severe or higher degrees of hearing loss.

Methods: In phase 1, a mHealth-based hearing screener (SRESHT) was developed using a single board computer with wireless commercial headphones and speakers as transducers, which were calibrated according to the standard procedure. Three subjective hearing screening modules were conceptualized and developed for different age groups: (1) behavioral observation audiometry–screening for infants aged from 0 to 1 year; (2) speech spectrum awareness task–screening for children 1 to 3 years old; and (3) speech recognition task–screening for children 3 to 6 years old. Different auditory stimuli for the screening modules were generated and suitability was assessed: (1) noisemakers, animal sounds, and environmental sounds for infants (birth to 1 year old); (2) animal sounds and nonsense syllables for children (1 to 3 years old); and (3) eighteen picturable spondee words for children (3 to 6 years old). In phase 2, the SRESHT screener was beta-validated in children aged below 6 years to establish the agreement between SRESHT modules and the gold-standard procedure in identifying moderately severe and higher degrees of hearing loss.

Results: Off-the-shelf commercial speakers and headphones were selected and calibrated. On comparison of stimuli for behavioral observation audiometry on 15 children, Noisemaker stimuli were found suitable based on the average minimum response levels. On comparison of different stimuli for speech spectrum awareness task on 15 children, animal sounds were found to be suitable. On familiarity check of 18 spondee words for speech recognition task among 20 children, 12 spondee words had the eligibility cutoff (85%) and a presentation level of 5 dB SL (re-pure tone threshold) was sufficient to achieve 80% psychometric function. In phase 2, a total of 55 children aged 0 to 6 years (31 normal hearing and 24 hearing impairment) underwent SRESHT screening for beta validation. Cohen κ indicated that the overall SRESHT screener had a very good agreement (κ=0.82) with gold-standard audiometric screening for identifying moderately severe and higher degrees of hearing loss.

Conclusions: The development and beta validation of the SRESHT screener using the selected auditory stimuli showed that the stimuli were suitable for screening children.

XSL•FO
RenderX

(JMIR Form Res 2025;9:e53460) doi:10.2196/53460

KEYWORDS

audiometry; mHealth; devices; wireless; tablet-based screening; childhood hearing loss; early hearing detection and intervention; tablets; children; neonates; hearing loss; infants; development; validation; mobile phones

Introduction

Childhood hearing loss can have wide-ranging effects on a child's language, speech, education, and social integration, making this a crucial issue to address. To achieve optimal outcomes using amplification devices, it is necessary to screen, identify, and intervene children before 6 months of age [1]. In high-income countries, there are often policies for newborn hearing screening and hearing screening of young school-aged children [1]. On the other hand, in low- and middle-income countries (LMICs), such mandatory policies do not exist.

In India, the prevalence of hearing loss among neonates and infants is 5 per 1000 [2], and it ranges between 6% and 16% among children [3]. However, procuring and maintaining the recommended objective screening instruments like otoacoustic emissions and auditory brainstem response has not been viable at scale. Even in several other LMICs, only paper-pencil–based checklists are used to screen hearing loss, for example, Bangladesh [4], Thailand [5], Ecuador [6], Kenya [7], and Brazil [8]. The problem is further magnified by the lack of adequate ear and hearing care providers for early identification and rehabilitation of children with hearing loss. For example, in India, the ratio of audiologists to the population as per recent data from the Rehabilitation Council of India is 4.41 per 100,000 population [9] and in Bangladesh, the ratio of audiometrists to the population is 0.77% [10].

Task shifting to community workers or health care volunteers is a strategy that has been recommended [11] and explored [12-14] to overcome the stark service gaps. Such task shifting has resulted in successful implementation by improving program outcomes like coverage, refer rate, and follow-up rate [12,14]. This has been well supported by innovations in mobile health (mHealth). mHealth apps have decentralized hearing care from tertiary and secondary care hospitals to primary and community settings in some of the LMICs [14-16]. However, many of the mHealth solutions are focused on adult hearing loss and hearing aid fitting [17-20]. Smartphone or tablet-based apps for children are only available for those over the age of 4 years [21-24] and use either pure tones [22], digit in noise [25], or speech stimuli [21]. Therefore, there is a need for an affordable mHealth screener for younger children that can be used by community workers to accurately screen hearing loss and support early intervention in LMICs.

In general, subjective measures are less preferred for screening due to their lower reliability. For example, behavioral observation audiometry (BOA) is reported to be less accurate as it is often performed informally using uncalibrated stimuli without standardization of intensity, frequency, and distance [11,26-28]. Yet, recognizing the inequity in the availability of testing tools, the World Health Organization's guidelines for implementing community-based childhood hearing screening include the recommendation of using subjective-based screening tools [29].

Therefore, this study describes the development of a tablet-based subjective hearing screener (SRESHT screener) to identify children younger than 6 years of age with moderately severe hearing loss or higher degrees of hearing loss as per the definition of the Global Experts Group [30]. Since the variability of subjective measures is higher only for lower degrees of hearing loss [31,32], measures such as BOA and speech awareness and speech recognition–based screening modules were considered affordable alternatives to screen for hearing loss (HL; 60 dB and greater in the better ear). This preliminary level (60 dB HL and greater) of screening was chosen as part of a larger feasibility study to screen for hearing losses that are eligible for publicly funded welfare schemes in India. Similarly, in many other LMICs (eg, Bangladesh, Brazil), moderate or higher degrees of hearing loss are considered for welfare schemes [33], aids, and appliances with public funds. Therefore, the screening device is designed for application in similar contexts.

The specific objectives were to (1) identify the most appropriate auditory stimuli to screen for hearing loss in children aged between 0 and 6 years and (2) to evaluate the agreement between the pass or refer results of from behavioral screening using gold standard audiometer and of screening device in screening for moderately severe and above degrees of hearing loss.

Methods

Ethical Considerations

This cross-sectional study was approved by the institutional ethics committee of Sri Ramachandra Institute of Higher Education and Research (IEC-NI/20/OCT/76/113). Informed consent was obtained from parents of children who participated in the study. The subjects were included on a voluntary basis and no monetary benefits were given to the participants. All subjects were anonymized during data entry, analysis, and reporting to maintain confidentiality.

Development of SRESHT Hearing Screener

Hardware and Software

A single board computer was used as the hardware with software coded using Python (Python Software Foundation) language. App modules were developed for hearing screening for children from birth to 6 years of age. The user interface was designed in both English and the local language with simple texts, images, and symbols such that nontechnical screening personnel can use the software. Three hearing screening test procedures were included: BOA-based screening (for children younger than 1 year), speech spectrum awareness task (SSAT)–based screening (for children older than 1 year and younger than 3 years), and

XSL•FO
RenderX

speech recognition task (SRT)–based screening (for children older than 3 years and younger than 6 years).

The test module gets automatically selected based on the age selection in the app by the tester. The test modules include screening frequencies of 500 Hz, 1000 Hz, 2000 Hz, and 4000 Hz at a specified moderate intensity level (selected based on calibration). For each test procedure, age-appropriate stimuli were developed (described under each test procedure in detail) and uploaded to the software. An inbuilt noise monitoring software was included which provides access to the screening module only when the ambient noise levels are less than 50 dB (A). A cloud-based back-end data management was integrated via Google Drive. The screening result is displayed automatically based on the response entered for each frequency.

Selection of Transducers

Suitable commercial transducers appropriate to the test module and age were selected based on frequency response at 500 Hz to 4 kHz, output intensity range, and distortions at intensities up to 80 dB SPL (sound pressure level). At first, 3 commercial grade supra-aural headphones (Sony WHCH510 headphones, Sennheiser HD 450 BT, and Boatrockers 410), and 3 speakers (JBL Go 2, Sony SRSXB12, and Fire Boltt Xplode 1500) were selected for output comparison based on manufacturer information on frequency, intensity range, cost, and ease of availability in the market.

Headphone

Headphones were connected via Bluetooth to the SRESHT screener at maximum volume setting. The headphone diaphragm was coupled to the calibration microphone with a 6CC coupler connected to the Larsen Davis sound level meter (SLM).

Speaker

Bluetooth speakers were connected to the SRESHT screener at maximum volume setting and placed at a distance of 1 m and at 0-degree azimuth from the SLM microphone. A subject was seated in a chair and the ear level was marked from the floor. The subject was then removed and the SLM microphone was suspended from a stand approximately at the ear level marked.

For all the transducers, warble tone stimuli were presented at 500 Hz, 1000 Hz, 2000 Hz, and 4000 Hz at different volume settings (%) of the SRESHT screener hardware, and the corresponding dB SPL values were measured in the SLM. The software was initially coded to produce the desired output level for moderate intensity levels in SPL (40 dB, 50 dB, and 60 dB) across the frequencies.

The Sony WHCH510 headphones and JBL Go 2 pro speakers were selected based on their accuracy of output, consistency of output, and output range (described in Development of the SRESHT Screener section).

Reference Equivalent Threshold Sound Pressure Level of Transducers

Sony WHCH510 Headphones

The mean thresholds of 10 normal hearing individuals were measured on a standard audiometer using TDH39 headphones from 500 Hz to 4000 Hz. The same individuals' threshold volume levels (%) were then obtained using the SRESHT hearing screener and Sony WHCH510 headphones. Larson Davis SLM was used to calculate the dB SPL value corresponding to the threshold volume level for each frequency. The Wilber method was used to estimate the reference equivalent threshold sound pressure level (RETSPL) values for the Sony WHCH510 headphones [34].

JBL Go2 RETSPL

The standard RETSPL values for speakers at 0-degree azimuth [35] were applied for 500 Hz to 4 kHz. The software was recoded as per the RETSPL value derivation.

Calibration Tone

A 1000 Hz warble tone was generated and used as a calibration tone, which is introduced in the app prior to the commencement of the screening. The minimum audible level of the tone at a moderate noisy environment of 40 to 45 dBA for 10 normal hearing subjects was measured and used to set the calibration tone.

Stimulus Validation

BOA-Based Hearing Screening

Both tonal and nontonal stimuli were assessed for BOA [32,36]. Studies suggested that children's responses to nontonal stimuli were as effective as tonal stimuli [36]. Nontonal stimuli included environmental sounds, animal sounds, and noisemakers that were contextually relevant to screen children. The stimuli were downloaded from an open-source website [37,38]. The acoustic parameters of the stimuli were then modified and corrected or normalized using Audacity (version 3.5.0; Audacity Team) and Apple Logic Pro (version 10.6; Apple Inc). The stimuli and the characteristics are summarized in Table 1.

XSL•FO
RenderX

Table 1. Nontonal stimuli categorized into noisemakers, environmental sounds, and animal sounds with center frequencies of 500 Hz, 1000 Hz, 2000 Hz, and 4000 Hz and a bandpass filter with 500-Hz bandwidth (filter slope 24 dB/octave).

Frequencies (center frequency, Hz)	Noisemakers	Environmental sounds	Animal sounds	Bandwidth required	Duration (seconds)
500	Drum or wooden drum	Horn	Cow mooing	250 Hz to 750 Hz (18 or 24 dB/octave slope)	2-3
1000	Wooden rattles	Applause	Dog barking	750 Hz to 1250 Hz (24 dB/octave slope)	2-3
2000	Metallic rattle or squeaking toy	Laughing	Whinnying horse and goat meh	1750 to 2250 Hz (24 dB/octave slope)	2-3
4000	Bell	Telephone ringing	Bird chirp and cat meow	3750 to 4250 Hz (24 dB/octave slope)	2-3

Procedure for Identifying Suitable Stimuli for BOA

A sound-treated room was preferred for conducting the screening test at the SRIHER clinic or laboratory. Participants were children with normal hearing and hearing loss between 0 to 1 year of age. Participants' details are described in Table 2.

A standard diagnostic audiometer (Madsen Astera[2]) with calibrated speakers (Martin Audio C115) was used to conduct the test using the newly developed stimuli (warble, noisemakers, environmental sounds, and animal sounds) at 500 Hz, 1000 Hz, 2000 Hz, and 4000 Hz for children. Behavioral responses such as eyeblink, eye widening, searching, and head turn toward the source were considered acceptable responses. This was noted by an audiology intern. Initially, the stimuli presentation started at 20 dB HL in 500 Hz for children with normal hearing and at a higher intensity level for children with hearing loss. If a response was obtained, the stimuli presentation was presented thrice in the same intensity level to confirm the response and if not, the presentation was increased by 5 dB steps (ascending method). Whenever the child got distracted or bored, tangible or visual reinforcement was given by the audiology assistant. The same procedure was repeated for each stimulus at each frequency. Minimum response levels or thresholds were obtained for each stimulus at each frequency and compared to identify the stimuli that resulted in the lowest thresholds.

Table 2. Characteristics of children involved in BOA[a] stimuli validation.

Participants	Sex	Age	Hearing age	Diagnosis	Type of amplification
Child 1	Male	6 months	6 months	Normal hearing	N/A[b]
Child 2	Female	4 months	4 months	Normal hearing	N/A
Child 3	Male	6 months	6 months	Normal hearing	N/A
Child 4	Male	6 months	6 months	Normal hearing	N/A
Child 5	Male	6 months	6 months	Normal hearing	N/A
Child 6	Female	4 months	4 months	Normal hearing	N/A
Child 7	Female	8 months	8 months	Normal hearing	N/A
Child 8	Male	8 months	8 months	Normal hearing	N/A
Child 9	Male	3 months	3 months	Normal hearing	N/A
Child 10	Female	9 months	9 months	Normal hearing	N/A
Child 11	Female	5 months	5 months	Normal hearing	N/A
Child 12	Male	2 years 5 months	8 months	Bilateral: Severe to profound hearing loss	Right: Cochlear implant
Child 13	Male	2 years 6 months	10 months	Bilateral: Severe to profound hearing loss	Right: Cochlear implant
Child 14	Female	1 year 8 months	6 months	Bilateral: Moderately severe hearing loss	Bilateral: Hearing aid
Child 15	Female	2 years 2 months	8 months	Bilateral: Moderately severe hearing loss	Bilateral: Hearing aid

[a]BOA: behavioral observation audiometry.

[b]N/A: not applicable.

XSL•FO
RenderX

SSAT-Based Hearing Screening

For children between 1 and 3 years of age, speech-like animal sounds, such as a cows' *moo*, dogs' *bow-wow*, goats' *meh*, and cats' *meow*, and nonsense syllables like *mamama* and *bababa* were included.

The spectrum of the animal sounds using PRAAT software (version 6.3.2; Softsonic International) was compared to the spectrum of consonants and vowels to verify that it resembled the properties of speech. The formant frequencies (F1, F2, and F3) of animal sounds were approximately similar to the corresponding formant frequencies of the target vowels. Table 3 shows the formant frequencies of recorded animal sounds and existing normative vowel formant frequencies, as adapted from Raphael et al [39].

Table 3. Comparison of formant frequencies of animal sounds and normative formant frequencies of vowels matched approximately[a].

Formant frequencies	Animal sounds			
	Cow moo/ /u/	Dog bark/ /a/	Goat meh/ /ae/	Cat meow/ /i/
F1	300 Hz/300 Hz	853 Hz/ 730-850 Hz	1565 Hz/ 850-1100 Hz	1566 Hz/ 1500 Hz
F2	693 Hz/800 Hz	1180 Hz/1100-1220 Hz	2060 Hz/2050-2300 Hz	2217 Hz/2500 Hz
F3	986 Hz/2100 Hz	1616 Hz/2100-2400 Hz	2507 Hz/ 2800-3200 Hz	2835 Hz/ 3000 Hz

[a]/u/,/a/,/ae/,/i/ are English vowel sounds represented in international phonetic alphabet (IPA).

Procedure for Identifying Suitable Stimuli for SSAT

For children between 1 and 3 years of age, SSAT-based screening was conducted with the standard audiometer (Madsen Astera[2]), with TDH 39 headphones with newly developed stimuli. The participants included children with normal hearing and hearing impairment between the ages of 1 and 3 years. Behavioral responses included head turn, locating the source of the sound, and imitating the sound. Stimuli was presented initially at 1 kHz at 60 dB HL for normal hearing and 80 dB HL for children with hearing impairment. The Hughson-Westlake method of threshold tracking was used to identify the minimum response levels for each stimulus at each frequency. The minimum response levels for these stimuli were compared and the stimuli yielding better responses (thresholds) were finalized for the screening test procedure.

SRT-Based Hearing Screening

SRT-based screening was developed for children between 3 and 6 years of age. Eighteen picturable spondee words in the local language (Tamil) were excerpted from a standardized spondee list of 50 words [40] that is routinely used in clinical practice. The words were selected based on the relevance of the word in present-day context, and its picturability. Pictures corresponding to the 18 words were downloaded from the open-source web-based repository named Wikimedia Commons [41].

Procedure for Identifying Suitable Stimuli for SRT

A familiarity test of all the words and their corresponding picture was conducted. Participants were both children with normal hearing and children with hearing impairment above the age of 3 years. For familiarity testing of words, each child was asked to point out the appropriate picture when the word was read out at the conversation level. Similarly, for picture familiarity, each child was asked to name the picture.

These 18 spondee words were further divided into 3 lists of 6 words each. Then, these spondee words were recorded with a female voice centered at 1000 Hz and with loudness balanced at –3 dB peak value, having an interstimulus interval of 5 seconds. A closed-choice response set was developed with 1 target picture and 3 other pictures. The 4 pictures were arranged in a sequential manner with an equal number of repetitions.

Procedure for Determining Suitable Presentation Level for SRT Screening in Children

Participants were children with normal hearing and children with hearing impairment between the age group of 3 and 9 years. Using the standard audiometer (Madsen Astera[2]) with TDH 39 headphones, each child's pure tone thresholds (PTA) were obtained using the Hughson-Westlake method. Then, each child underwent SRT with picture identification as response. This was performed at different sensation levels (0 dB SL, 5 dB SL, 10 dB SL, and 20 dB SL with re-PTA) to identify the level at which 80% correct response (5/6 words) was obtained.

Beta Validation of SRESHT Screening Modules

This beta validation was conducted prior to the full validation in which the suitability of the stimuli was evaluated and agreement between new stimuli versus standard stimuli was compared.

Sample

Two groups of children participated in the study. The first group consisted of children under 6 years of age with normal hearing. The other group consisted of children with hearing impairment having a hearing age of less than 6 years and these children were younger than 10 years of chronological age. Children with any other comorbid conditions or any other known disabilities were excluded from the study (eg, children with developmental delay, vision impairment, autism spectrum disorders).

Children were recruited from the outpatient clinic of the tertiary care hospital attached to the institution where the study was conducted, as well as in camps organized in 2 rural districts of Tamil Nadu for the period of 3 months.

Informed written consent was obtained from the parents of the children before the commencement of data collection.

XSL•FO
RenderX

Procedure

The procedure was conducted in any quiet room near the outpatient department or the campsites with an ambient noise level of less than 50 dB (A) as recorded through the noise monitoring software of the SRESHT screener. SRESHT hearing screening with the finalized stimuli and screening modules for each age group was performed for all children using age-appropriate screening procedures at 60 dB HL.

Then, the minimum response level was estimated for 500 Hz, 1 kHz, 2 kHz, and 4 kHz with the standard diagnostic audiometer (Madsen Astera[2]) for each test procedure, as a gold standard. Screening results for the SRESHT hearing screener at 60 dB HL (pass or refer) were compared to audiometric thresholds more or less than 60 dB HL.

Analysis

Nonparametric tests were used to analyze the data as the results of the Shapiro-Wilk test showed that the data were not normally distributed.

The Friedman test was used to compare the children's responses to different stimuli in BOA. The Wilcoxon signed rank test was used to compare the children's responses to different stimuli for SSAT. Percentage analysis was used for word familiarity for SRT screening. Agreement statistics were used to beta-validate the screening results.

Results

Development of SRESHT Hearing Screener

Transducers

The summary of output characteristics of various speakers and headphones is shown in Table 4. The Sony WHCH510 headphones and JBL Go 2 pro speakers were selected based on their accuracy of output, consistency of output, output range, ease of use, and cost.

Table 4. Characteristics of each transducer (speakers or headphones) for the hearing screening module.

Transducers	Frequency response	Minimum output	Maximum output	Distortion
Sony WHCH510 headphones	No frequency splattering from 500 Hz to 4 kHz	40 dB SPL	100 dB SPL	Within permissible limits (ANSI[a], 2020)
Sennheiser HD450BT headphones	No frequency splattering from 500 Hz to 4 kHz	40 dB SPL	102 dB SPL	Within permissible limits (ANSI, 2020)
Boat rockers 410 headphones	Frequency splattering present at 4 kHz	50 dB SPL	104 dB SPL	Distortion present at 2 kHz and 4 kHz at 80 dB SPL
JBL Go 2 speakers	No frequency splattering from 500 Hz to 4 kHz	40 dB SPL	80 dB SPL	Within permissible limits (ANSI, 2020)
Sony SRSXB12 speakers	Appropriate from 500 Hz to 4 kHz	40 dB SPL	77 dB SPL	Within permissible limits (ANSI, 2020)
Fire Boltt Xplode1500 speakers	Frequency splattering present 2 kHz and 4 kHz	50 dB SPL	87 dB SPL	Distortion presented 2 kHz and 4 kHz at 80 dB SPL

[a]ANSI: American National Standards Institute [42].

RETSPL of Transducers

The RETSPL values derived for the Sony WHCH510 headphones are shown in Table 5.

Table 5. Sony WHCH510 headphone RETSPL[a] values.

Frequency	RETSPL (dB)
500 Hz	11
1 kHz	1
2 kHz	3
4 kHz	–2

[a]RETSPL: reference equivalent threshold sound pressure level.

Stimuli Validation

BOA-Based Hearing Screening

Fifteen children (11 Children with normal hearing and 4 Children with hearing impairment) in the age range of 0 to 1 year of age with a mean age of 6 (SD 1.9) months for children with normal hearing and a mean hearing age of 8 (SD 1.6) months for children with hearing impairment (Table 2), underwent the BOA using the newly developed stimuli. The Friedman test revealed no significant difference in the average minimum response levels of warble tone versus average

XSL•FO
RenderX

minimum response levels of noisemakers, animal sounds, and environmental sounds (n=15; χ^2_3=15.44; *P*=.089).

A comparison of the average minimum response levels for different stimuli for children with normal hearing and children with hearing impairment revealed that the average minimum responses to noisemakers for children with hearing impairment were superior to the responses to other stimuli, that is, animal sounds and environmental sounds (Figure 1). Therefore, noisemaker-based stimuli were selected for this age group, even though all 3 categories of non-tonal stimuli were appropriate for hearing screening.

Figure 1. Comparison of average minimum response levels for tonal and nontonal stimuli among children with normal hearing and children with hearing loss.

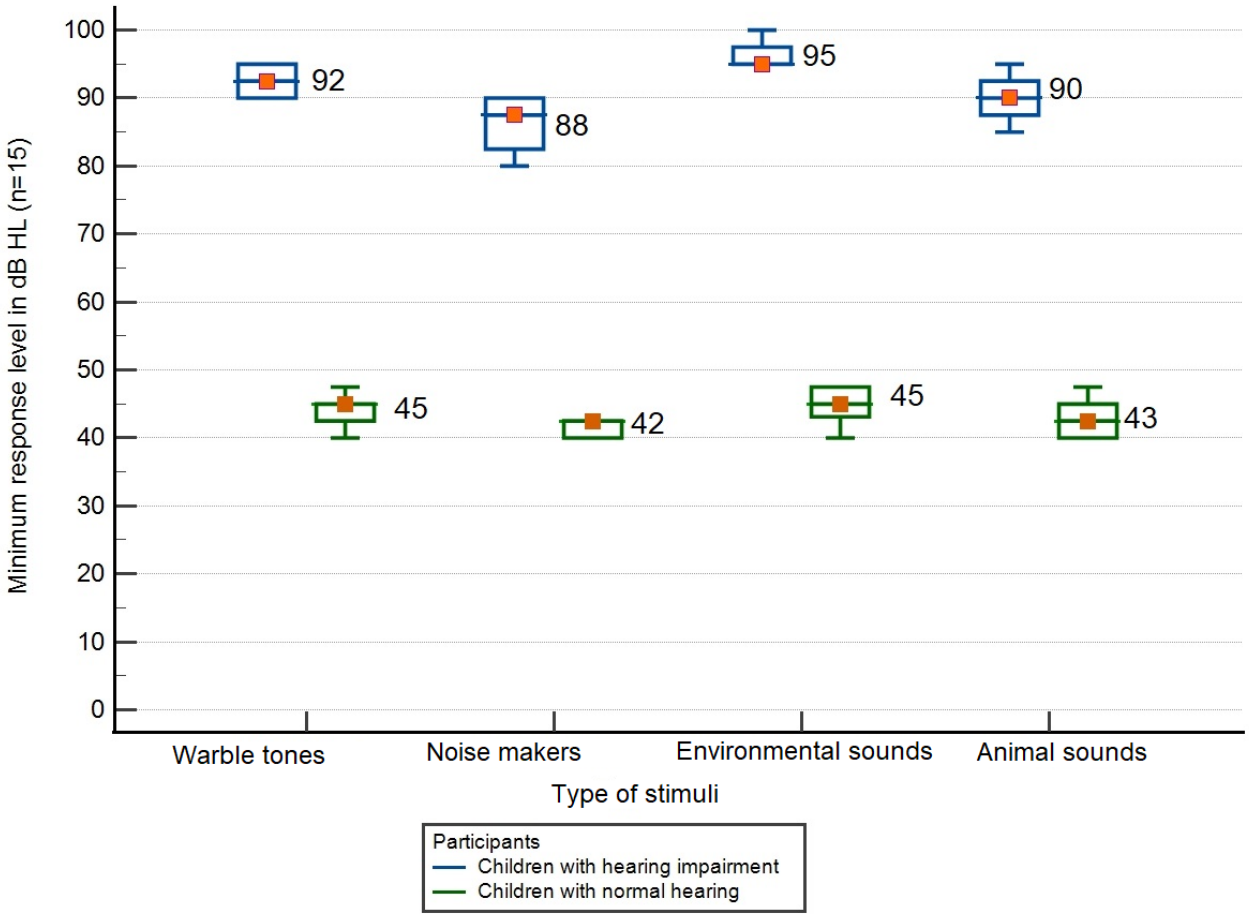

SSAT-Based Screening

A total of 15 children (10 children with normal hearing and 5 children with hearing impairment) in the age range of 1 to 3 years with a mean age of 2.3 (SD 0.37) years for children with normal hearing and 1.7 (SD 0.26) years of hearing age for children with hearing impairment (Table 6) underwent the SSAT. Using animal sound stimuli centered as frequencies corresponding to different speech sounds. Wilcoxon signed rank test showed a significant difference in the threshold for the nonsense syllables response and animal sounds response (*z*=–3.59; *P*<.001). The mean minimum response level difference between animal sounds was 5 dB better than the mean minimum response levels of nonsense syllables (Figure 2). Therefore, animal sounds were finalized for this age group.

XSL•FO
RenderX

Figure 2. Comparison of average minimum response levels for different stimuli among children with normal hearing and hearing loss (HL).

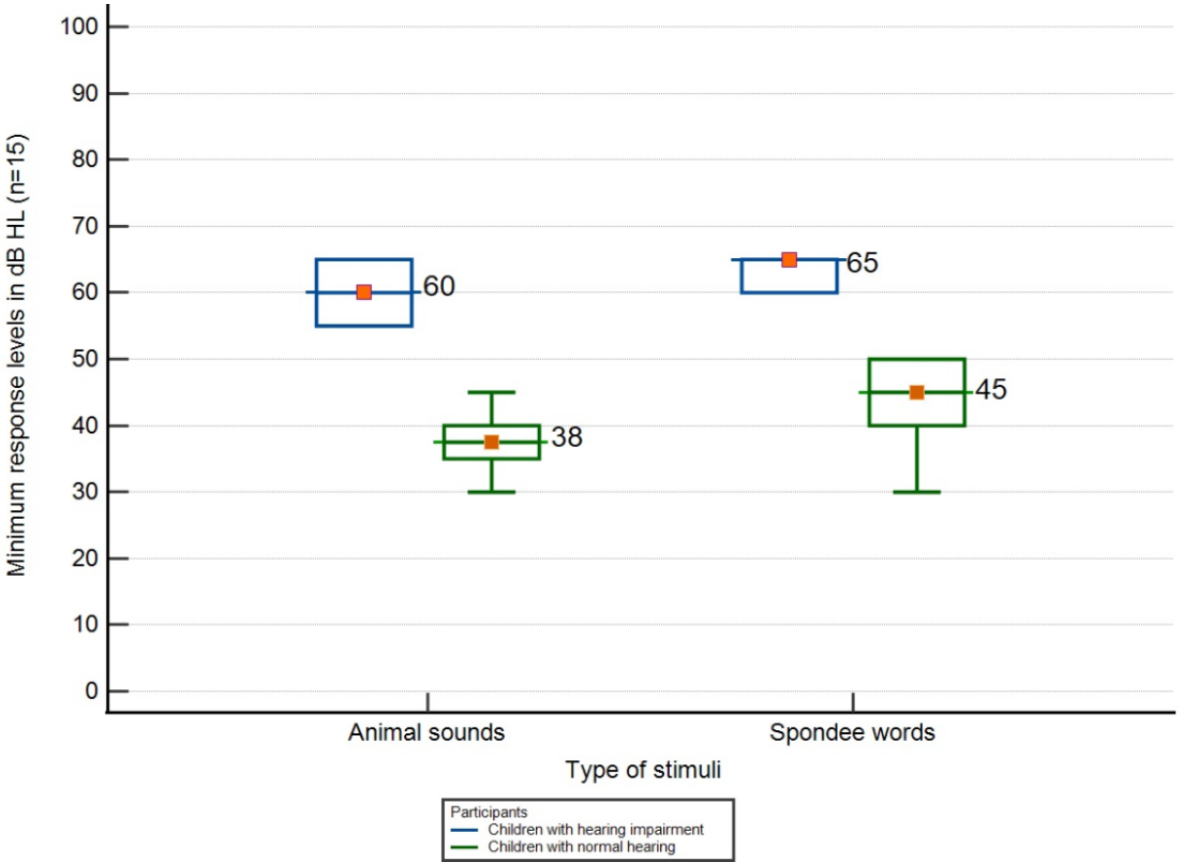

Table 6. Participants characteristics of children involved in SSAT[a] stimuli validation.

Participants	Sex	Chronological age (in years)	Hearing age (in years)	Diagnosis	Type of amplification
Child 1	Female	2.1	2.1	Normal hearing	N/A[b]
Child 2	Female	2.3	2.3	Normal hearing	N/A
Child 3	Male	2.5	2.5	Normal hearing	N/A
Child 4	Male	2.5	2.5	Normal hearing	N/A
Child 5	Male	2.5	2.5	Normal hearing	N/A
Child 6	Male	1.9	1.9	Normal hearing	N/A
Child 7	Female	2	2	Normal hearing	N/A
Child 8	Male	2.8	2.8	Normal hearing	N/A
Child 9	Male	2.9	2.9	Normal hearing	N/A
Child 10	Female	1.8	1.8	Normal hearing	N/A
Child 11	Male	3.7	1.5	Bilateral: Moderately severe hearing loss	Bilateral: Hearing aid user
Child 12	Female	4	1.8	Bilateral: Severe to profound hearing loss	Right: Cochlear implant
Child 13	Male	3.8	1.7	Bilateral: Severe hearing loss	Right: Cochlear implant; left: Hearing aid
Child 14	Female	4.9	2.2	Bilateral: Severe to profound hearing loss	Left: Cochlear implant
Child 15	Female	4.2	1.7	Bilateral: Moderately severe hearing loss	Bilateral: Hearing aid

[a]SSAT: speech spectrum awareness task.
[b]N/A: not applicable.

XSL•FO
RenderX

SRT-Based Screening

Twenty children (15 children with hearing impairment and 5 children with normal hearing between the ages of 3 to 9 years with a mean age of 3.7 (SD 0.44) years for children with normal hearing and 3.8 (SD 0.94) years of hearing age for children with hearing impairment (Table 7) underwent the speech recognition task using the spondee words.

Twelve of the 18 words could be identified by 85% of the normal-hearing children. Therefore, these words were included as 2 lists of 6 words each (Figure 3).

Thresholds for warble tone were compared with the speech recognition threshold to find the suitable presentation level for the spondee words for 80% psychometric function, 5 dB SL (re PTA) was found to be sufficient (Figure 4).

Table 7. Participants characteristics of children involved in SRT[a] stimuli validation.

Participants	Sex	Age (years)	Hearing age (years)	Diagnosis	Type of amplification
Child 1	Female	3.3	3.3	Normal hearing	N/A[b]
Child 2	Female	4.3	4.3	Normal hearing	N/A
Child 3	Male	4.2	4.2	Normal hearing	N/A
Child 4	Male	3.5	3.5	Normal hearing	N/A
Child 5	Male	3.6	3.6	Normal hearing	N/A
Child 6	Female	5.8	3	Bilateral: Severe to profound hearing loss	Right: Cochlear implant
Child 7	Male	4.1	3	Bilateral: Severe to profound hearing loss	Left: Cochlear implant
Child 8	Female	4.5	3.1	Bilateral: Moderate to moderately severe hearing loss	Bilateral: Hearing aid user
Child 9	Female	5	3.5	Bilateral: Profound hearing loss	Left: Cochlear implant
Child 10	Female	4.2	3	Right: Moderately severe hearing loss; left: Severe hearing loss	Left: Cochlear implant
Child 11	Female	8.5	5.2	Bilateral: Moderately severe hearing loss	Bilateral: Hearing aid user
Child 12	Male	7.1	4.5	Bilateral: Severe to profound hearing loss	Right: Cochlear implant
Child 13	Male	8	5.5	Bilateral: Moderately severe hearing loss	Right: Cochlear implant; left: Hearing aid
Child 14	Male	6.5	4.1	Left: Moderately severe hearing loss; right: Severe hearing loss	Bilateral: Hearing aid user
Child 15	Female	5	3.5	Bilateral: Moderately severe hearing loss	Bilateral: Hearing aid user
Child 16	Female	9	5	Bilateral: Severe to profound hearing loss	Left: Cochlear implant
Child 17	Female	7.5	3.2	Bilateral: Moderately severe hearing loss	Right: Cochlear implant
Child 18	Male	4.6	3	Bilateral: Severe hearing loss	Left: Cochlear implant
Child 19	Male	5.2	3.1	Bilateral: Severe to profound hearing loss	Right: Cochlear implant
Child 20	Male	4	3	Bilateral Severe to profound hearing loss	Left: Cochlear implant

[a]SRT: speech recognition task.

[b]N/A: not applicable.

XSL•FO
RenderX

Figure 3. Average percentage of correct responses by children for each word.

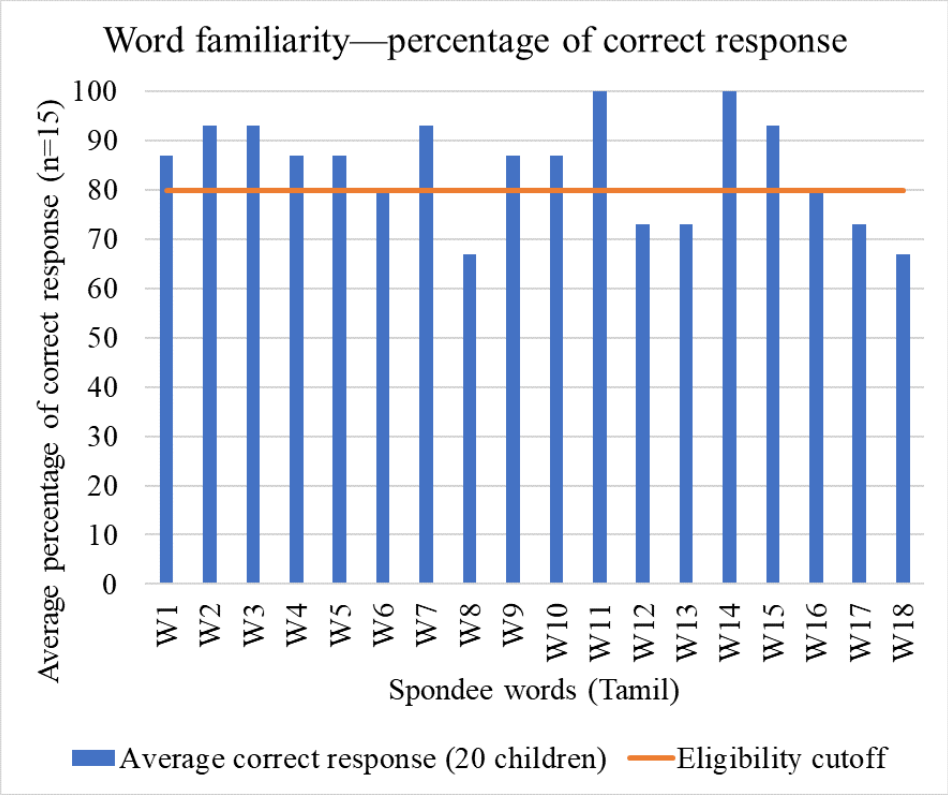

Figure 4. Comparison of percentage correct responses versus speech recognition threshold of 20 children. PTA: pure tone threshold.

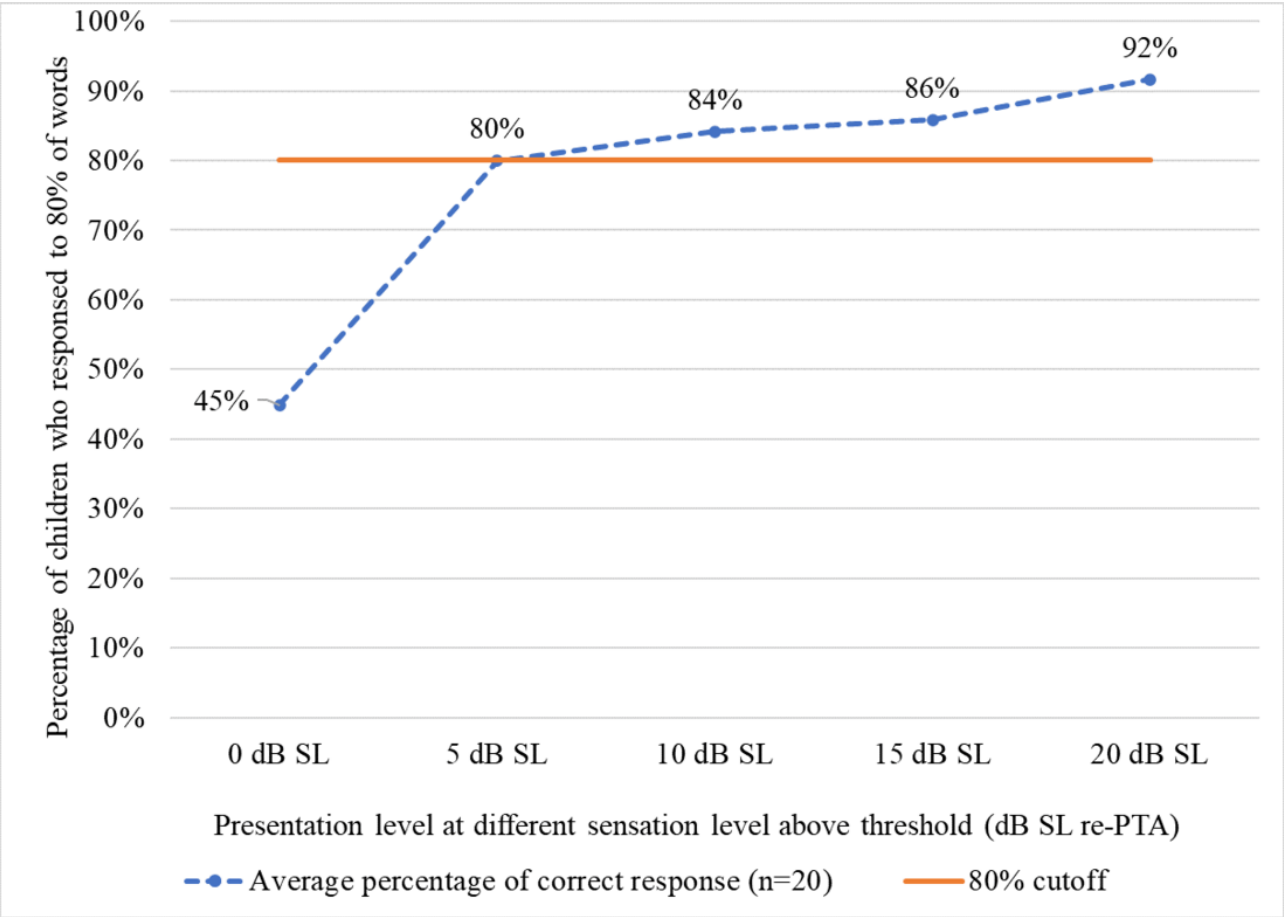

XSL•FO
RenderX

Accuracy of SRESHT Screening Modules: Beta Validation

A total of 55 children (31 children with normal hearing and 24 children with hearing impairment underwent hearing screening using the SRESHT screener device. The chronological age range for children with normal hearing participants was 2 days to 6 years with a mean age of 3.2 (SD: 1.94) years, whereas the chronological age range for children with hearing impairment participants was 1 year to 9 years with a mean age of 4.9 (SD 1.42) years.

Of the 24 children with hearing impairment, 14 were unilateral cochlear implantees, 8 were bilateral hearing aid users, and 2 used bimodal hearing devices. The diagnostic reports of all 24 children with hearing impairment were reviewed to verify their unaided and most recent aided thresholds. Table 8 shows the agreement results.

Therefore, the SRESHT screener modules with new stimuli were as good as the standard audiometer in identifying moderately severe and higher degrees of hearing loss of 60 dB HL or higher.

Table 8. Agreement in findings between SRESHT screener and gold standard audiometric screening.

Test frequencies	κ value	95% CI	*P* value	Agreement
500 Hz	1.00	1.00-1.00	0.001	Perfect agreement
1000 Hz	0.87	0.75-1.00	0.001	Very good agreement
2000 Hz	0.72	0.47-0.97	0.005	Good agreement
4000 Hz	0.85	0.75-1.00	0.001	Very good agreement
Overall	0.82	0.66-0.97	0.001	Very good agreement

Discussion

Principal Findings

This study focused on the development and beta validation of an affordable subjective hearing screener (SRESHT) for identifying hearing losses greater than 60 dB HL in children younger than 6 years of age. The SRESHT hearing screener was developed in a single-board computer with calibrated off-the-shelf transducers. For children between 0 and 1 year of age, the BOA module was developed with noisemakers as suitable stimuli. For children between 1 and 3 years of age, the SSAT module with animal sounds was found as a suitable stimulus. For children between 3 and 6 years of age, the SRT module with 12 Tamil spondee words was found suitable. On beta validation of the SRESHT screener with age-appropriate modules among children between 0 to 6 years of age, it was found that the SRESHT screener was suitable to identify hearing losses of moderately severe and higher degrees.

The SRESHT screener is an affordable alternative with age-appropriate subjective hearing screening modules for children from birth to 6 years of age. The cost estimate is expected to be one-tenth of an objective hearing screener. Among the available smartphone-based hearing apps, only a few are designed specifically for children, but all are applicable only for children older than 4 years of age [18,21,24,43]. So far, to our best knowledge, the clinical population of children with hearing impairment has not been included in studies that describe mHealth hearing screeners.

In high-income countries, objective screening methods such as otoacoustic emissions and auditory brainstem response are routinely used as part of universal hearing screening as they are more accurate [26]. In LMICs, while there are efforts to use objective screeners by private hospitals and research teams, nationwide programs are negligible [2,44,45]. In a pilot government-initiated newborn hearing screening program in a southern state in India, objective screeners were initially procured, however, its long-term viability was called into question due to the high cost of screening equipment and frequent repair and consumables costs [46-48].

The semiautomated SRESHT screener is currently designed to screen “moderately severe” or higher degrees of hearing loss [49]. As a first step, this higher level of screening was chosen as part of a feasibility study to make a valid tool available to screen for 60 dB HL in the better ear (eligibility for publicly funded welfare schemes) within the public health system of a southern state in India. Not only in India, but in several LMICs (Bangladesh, Pakistan, and Kenya), a criterion of 50 dB HL or higher is used to provide welfare schemes, aid, and appliances with public funds. As a result, affordable tools that support this level of screening are useful in identifying this group of children who would otherwise miss out on welfare measures.

In the past, subjective screening using distraction and whisper tests in the United Kingdom [50,51], informal clap screening in Africa [52], noisemaker-based screening in India [53], and questionnaire-based screening in Thailand [5] and Kenya [7] have been used to support mass screening in the community. Some of the drawbacks of these screening methods are the lack of calibrated stimuli, the requirement of 2 personnel to administer the test, lack of standardization of intensity, frequency, distance, and test-retest reliability [11,26-28]. The SRESHT screening with age-appropriate calibrated stimuli reduced the extent of subjectivity. For the youngest babies, noisemakers are known to be robust [31,36,54] and this was reconfirmed from our beta validation. Stimuli such as animal sounds are known to be appealing to younger children for hearing testing [55], therefore, animal sounds were mapped to speech spectrums to develop the speech spectrum awareness test. These 2 modules are language independent and therefore can be used in several resource-limited settings across the world to support community-based screening efforts.

For the youngest age group (0 to 1 year), noisemakers are known to be robust [31,54], and this was reconfirmed from the results

XSL•FO
RenderX

of phase 1 of the study and used for the BOA screening module. In this study, BOA screening was conducted with noisemakers for infants under the age of 1 year. Visual reinforcement audiometry screening or distraction test is recommended for children aged 6 months to 1 year in a few studies [54] and guidelines [11,56], however, since the SRESHT screener is conceptualized for use by community workers and therefore the level of complexity and training required was an important consideration to choose BOA over visual reinforcement audiometry.

The initial version of the screening test for the age group of 1 year to 3 years consisted of a visual reinforcement task using warble tones. A simple light-emitting toy with a manual switch control was also developed for visual reinforcement. However, during field testing, the grassroots-level workers could not perform the screening task efficiently and provided feedback on practical challenges. This led to the modification of the module to a speech spectrum awareness task. The task was also tried in commercial grade insert earphones. However, insert earphones were not found suitable as children between 1 and 3 years of age were not comfortable to wear them.

For older children, the SRESHT screener uses picturable age-appropriate spondee words in the local language. While many apps have used pure tones [18,57]; warble tones [22,43], and very few exist with speech stimuli [21,24]. Picturable spondee words of different languages can be added to make the screener available elsewhere.

Affordable screeners are much needed to support nationwide early intervention programs in LMICs for long-term viability. The cost of screening a child for hearing loss using objective screening methods is estimated to be between US $17 and US $26 across high-income countries and LMICs like Brazil [58], India [59], the United States [60], and China, [61,62]. This cost can be lowered when device costs are less. Hearing screening apps for adults and older children range in price from US $600 to US $2000 [63], and some require standard audiometric headphones [18,22,24]. Commercial headphones with precise calibrations, on the other hand, can produce consistent output [16,17,21]. To keep costs low, the SRESHT screener was developed using off-the-shelf transducers that produce standard output with RETSPL derivation thereby reducing the cost to one-tenth the price of objective screening tools.

In community-based, school-based or mass screening programs, in-built noise monitoring is crucial due to the lack of sound-treated booths. The SRESHT screener performs noise monitoring prior to the commencement of each of the screening tests to limit the contamination of results by high ambient noise levels. Such features exist in some of the app-based screeners already [23,43,64].

The SRESHT screener is currently limited to permanent hearing losses of moderately severe and higher degrees in children under 6 years of age. Hence, further study is required to extend the scope of the screener to all types and degrees of hearing loss. While the scope of the current study is limited to the development of the screening modules with suitable beta validation, the full validation of the hearing screener with these finalized parameters has been undertaken and will be reported as a separate paper.

Future Directions

Full validation against the gold standard objective test is undertaken for this device and the sensitivity, specificity, and positive and negative predictive values will be reported in a separate publication. A project is also being undertaken to lower the screening intensity levels to screen for lower degrees of hearing loss.

Acknowledgments

This study was conducted as a part of a larger project funded by the Department of Biotechnology/Wellcome Trust India Alliance.

Data Availability

The datasets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

VR and DJB researched articles and came up with the design of the software. AP was involved in the hardware and software development. Data collection and analysis was done by DJB. Reviewing of the analysis was done by VR. Manuscript writing was done by DJB and VR. The manuscript review was done by JWH and RV.

Conflicts of Interest

None declared.

References

1. American Academy of Pediatrics, Joint Committee on Infant Hearing. Year 2007 position statement: Principles and guidelines for early hearing detection and intervention programs. Pediatrics 2007 Oct;120(4):898-921. [doi: 10.1542/peds.2007-2333] [Medline: 17908777]

XSL•FO
RenderX

2. Joshi B, Ramkumar V, Nair LS, Kuper H. Early hearing detection and intervention (EHDI) programmes for infants and young children in low-income and middle-income countries in Asia: a systematic review. BMJ Paediatr Open 2023 Jan;7(1):1-7 [FREE Full text] [doi: 10.1136/bmjpo-2022-001752] [Medline: 36720503]
3. Verma R, Konkimalla A, Thakar A, Sikka K, Singh A, Khanna T. Prevalence of hearing loss in India. NMJI 2022 Jan 29;34:216-222 [FREE Full text] [doi: 10.25259/nmji_66_21]
4. Berg A, Papri H, Ferdous S, Khan NZ, Durkin MS. Screening methods for childhood hearing impairment in rural Bangladesh. Int J Pediatr Otorhinolaryngol 2006 Jan;70(1):107-114. [doi: 10.1016/j.ijporl.2005.05.029] [Medline: 16023223]
5. Poonual W, Navacharoen N, Kangsanarak J, Namwongprom S. Risk factors for hearing loss in infants under universal hearing screening program in Northern Thailand. J Multidiscip Healthc 2016;9:1-5 [FREE Full text] [doi: 10.2147/JMDH.S92818] [Medline: 26766912]
6. Muñoz K, Luzuriaga E, Callow-Heusser C, Ortiz E, White K. Evaluation of a hearing screening questionnaire for use with Ecuadorian school-aged children. Int J Audiol 2015;54(9):587-592. [doi: 10.3109/14992027.2015.1015689] [Medline: 25804301]
7. Newton VE, Macharia I, Mugwe P, Ototo B, Kan SW. Evaluation of the use of a questionnaire to detect hearing loss in Kenyan pre-school children. Int J Pediatr Otorhinolaryngol 2001 Mar;57(3):229-234. [doi: 10.1016/s0165-5876(00)00453-5] [Medline: 11223455]
8. Samelli AG, Rabelo CM, Vespasiano APC. Development and analysis of a low-cost screening tool to identify and classify hearing loss in children: a proposal for developing countries. Clinics (Sao Paulo) 2011;66(11):1943-1948 [FREE Full text] [doi: 10.1590/s1807-59322011001100015] [Medline: 22086526]
9. 35th Annual Report. Rehabilitation Council of India. 2022. URL: https://rehabcouncil.nic.in/node/467 [accessed 2024-11-25]
10. Al Imam MH, Jahan I, Das MC, Muhit M, Akbar D, Badawi N, et al. Situation analysis of rehabilitation services for persons with disabilities in Bangladesh: identifying service gaps and scopes for improvement. Disabil Rehabil 2022 Sep;44(19):5571-5584. [doi: 10.1080/09638288.2021.1939799] [Medline: 34176400]
11. Hearing screening: considerations for implementation. World Health Organization. 2021. URL: https://www.who.int/publications/i/item/9789240032767 [accessed 2021-09-03]
12. de Kock T, Swanepoel DW, Hall JW. Newborn hearing screening at a community-based obstetric unit: Screening and diagnostic outcomes. Int J Pediatr Otorhinolaryngol 2016 May;84:124-131. [doi: 10.1016/j.ijporl.2016.02.031] [Medline: 27063767]
13. Shinn JR, Jayawardena ADL, Patro A, Zuniga MG, Netterville JL. Teacher Prescreening for Hearing Loss in the Developing World. Ear Nose Throat J 2021 Jun;100(3_suppl):259S-262S [FREE Full text] [doi: 10.1177/0145561319880388] [Medline: 31608682]
14. Yousuf Hussein S, Swanepoel DW, Mahomed F, Biagio de Jager L. Community-based hearing screening for young children using an mHealth service-delivery model. Glob Health Action 2018;11(1):1467077 [FREE Full text] [doi: 10.1080/16549716.2018.1467077] [Medline: 29764328]
15. Chu Y, Cheng Y, Lai Y, Tsao Y, Tu T, Young ST, et al. A Mobile Phone-Based Approach for Hearing Screening of School-Age Children: Cross-Sectional Validation Study. JMIR Mhealth Uhealth 2019 Apr 01;7(4):e12033 [FREE Full text] [doi: 10.2196/12033] [Medline: 30932870]
16. Dawood N, Mahomed Asmail F, Louw C, Swanepoel DW. Mhealth hearing screening for children by non-specialist health workers in communities. Int J Audiol 2021 Apr;60(sup1):S23-S29. [doi: 10.1080/14992027.2020.1829719] [Medline: 33043733]
17. Renda L, Selçuk ÖT, Eyigör H, Osma Ü, Yılmaz MD. Smartphone Based Audiometric Test for Confirming the Level of Hearing; Is It Useable in Underserved Areas? J Int Adv Otol 2016 Apr;12(1):61-66 [FREE Full text] [doi: 10.5152/iao.2016.1421] [Medline: 27340985]
18. Khoza-Shangase K, Kassner L. Automated screening audiometry in the digital age: exploring uhear™ and its use in a resource-stricken developing country. Int J Technol Assess Health Care 2013 Jan;29(1):42-47. [doi: 10.1017/S0266462312000761] [Medline: 23298579]
19. Tao KFM, Brennan-Jones CG, Jayakody DMP, Swanepoel DW, Fava G, Bellekom SR, et al. Validation of teleaudiology hearing aid rehabilitation services for adults: a systematic review of outcome measurement tools. Disabil Rehabil 2022 Aug;44(16):4161-4178. [doi: 10.1080/09638288.2021.1900928] [Medline: 33787435]
20. Samelli AG, Rabelo CM, Sanches SGG, Aquino CP, Gonzaga D. Tablet-Based Hearing Screening Test. Telemed J E Health 2017 Sep;23(9):747-752. [doi: 10.1089/tmj.2016.0253] [Medline: 28328389]
21. Dillon H, Mee C, Moreno JC, Seymour J. Hearing tests are just child's play: the sound scouts game for children entering school. Int J Audiol 2018 Jul;57(7):529-537. [doi: 10.1080/14992027.2018.1463464] [Medline: 29703099]
22. Kam ASC, Li LKC, Yeung KNK, Wu W, Huang Z, Wu H, et al. Automated hearing screening for preschool children. J Med Screen 2014 Jun;21(2):71-75. [doi: 10.1177/0969141314528322] [Medline: 24622246]
23. Swanepoel DW, Myburgh HC, Howe DM, Mahomed F, Eikelboom RH. Smartphone hearing screening with integrated quality control and data management. Int J Audiol 2014 Dec;53(12):841-849. [doi: 10.3109/14992027.2014.920965] [Medline: 24998412]

XSL•FO
RenderX

24. Yimtae K, Israsena P, Thanawirattananit P, Seesutas S, Saibua S, Kasemsiri P, et al. A Tablet-Based Mobile Hearing Screening System for Preschoolers: Design and Validation Study. JMIR Mhealth Uhealth 2018 Oct 23;6(10):e186 [FREE Full text] [doi: 10.2196/mhealth.9560] [Medline: 30355558]
25. Mahomed-Asmail F, Swanepoel D, Eikelboom RH, Myburgh HC, Hall J. Clinical Validity of hearScreen™ Smartphone Hearing Screening for School Children. Ear Hear 2016;37(1):e11-e17. [doi: 10.1097/AUD.0000000000000223] [Medline: 26372265]
26. Yoshinaga-Itano C, Manchaiah V, Hunnicutt C. Outcomes of universal newborn screening programs: systematic review. J Clin Med 2021;10(13):2784 [FREE Full text] [doi: 10.3390/jcm10132784] [Medline: 34202909]
27. Mott A, Emond A. What is the role of the distraction test of hearing? Arch Dis Child 1994 Jan;70(1):10-13 [FREE Full text] [doi: 10.1136/adc.70.1.10] [Medline: 8109999]
28. Weir CG. Use of behavioural tests in early diagnosis of hearing loss. Acta Otolaryngol Suppl 1985;421:86-92. [doi: 10.3109/00016488509121761] [Medline: 3862333]
29. World report on hearing. World Health Organization. 2021. URL: https://www.who.int/publications/i/item/9789240020481 [accessed 2021-03-03]
30. Olusanya BO, Davis AC, Hoffman HJ. Hearing loss grades and the international classification of functioning, disability and health. Bull World Health Organ 2019;97(10):725-728 [FREE Full text] [doi: 10.2471/BLT.19.230367] [Medline: 31656340]
31. Ramesh A, Jagdish C, Nagapoorinima M, Suman Rao PN, Ramakrishnan AG, Thomas GC, et al. Low cost calibrated mechanical noisemaker for hearing screening of neonates in resource constrained settings. Indian J Med Res 2012;135(2):170-176 [FREE Full text] [Medline: 22446858]
32. Karikoski JO, Marttila TT, Jauhiainen T. Behavioural observation audiometry in testing young hearing-impaired children. Scand Audiol 1998;27(3):183-187. [doi: 10.1080/010503998422700] [Medline: 9728780]
33. Bangladesh Persons with Disablity Welfare Act. World Health Organization. 2001. URL: https://extranet.who.int/mindbank/item/3767 [accessed 2011-12-31]
34. Wilber LA, Kruger B, Killion MC. Reference thresholds for the ER-3A insert earphone. J Acoust Soc Am 1988 Feb;83(2):669-676. [doi: 10.1121/1.396162] [Medline: 3351125]
35. ASA/ANSI S3.6-2018 (R2023): Specification For Audiometers. American National Standards Institute. 2018. URL: https://blog.ansi.org/asa-ansi-s3-6-2018-r2023-audiometers-specification/ [accessed 2025-01-08]
36. Massie R, Dillon H, Ching T, Birtles G. Birtles G (2005) The evaluation of nontonal stimuli for hearing assessment of young children. Aust N Z J Audiol 2005;27(1):10-17 [FREE Full text] [doi: 10.1375/audi.2005.27.1.10]
37. Free sound effects downloads. ZapSplat. URL: https://www.zapsplat.com/ [accessed 2024-09-24]
38. Royalty free stock music. Mixkit. URL: https://mixkit.co/free-stock-music/ [accessed 2024-09-24]
39. Raphael LJ, Borden GJ, Harris KS. Speech Science Primer: Physiology, Acoustics, and Perception of Speech. Philadelphia, PA: Lippincott Williams and Wilkins; 2011.
40. Dayalan J. Development and Standardization of Phonetically Balanced Test Materials in Tamil Language [dissertation]. Mysore, India: Mysore University; 1976.
41. Wikimedia Commons. URL: https://commons.wikimedia.org/wiki/Main_Page [accessed 2024-12-12]
42. Procedure for Measuring the Ambient Noise Level in a Room. American National Standards Institute. 2020. URL: https://webstore.ansi.org/preview-pages/ASA/preview_ANSI+ASA+S12.72-2015.pdf [accessed 2022-03-05]
43. Yeung J, Javidnia H, Heley S, Beauregard Y, Champagne S, Bromwich M. The new age of play audiometry: prospective validation testing of an iPad-based play audiometer. J Otolaryngol Head Neck Surg 2013;42(1):21 [FREE Full text] [doi: 10.1186/1916-0216-42-21] [Medline: 23663317]
44. Olusanya BO. Neonatal hearing screening and intervention in resource-limited settings: an overview. Arch Dis Child 2012;97(7):654-659. [doi: 10.1136/archdischild-2012-301786] [Medline: 22611062]
45. Olusanya BO. Highlights of the new WHO Report on Newborn and Infant Hearing Screening and implications for developing countries. Int J Pediatr Otorhinolaryngol 2011 Jun;75(6):745-748. [doi: 10.1016/j.ijporl.2011.01.036] [Medline: 21728253]
46. Galhotra A, Sahu P. Challenges and solutions in implementing hearing screening program in India. Indian J Community Med 2019;44(4):299-302 [FREE Full text] [doi: 10.4103/ijcm.ijcm_73_19]
47. Olusanya B. Screening for neonatal deafness in resource-poor countries: challenges and solutions. RRN 2015 May:51. [doi: 10.2147/rrn.s61862]
48. Krishnan LA, Donaldson LK. Newborn hearing screening in developing countries: understanding the challenges and complexities of implementation. Perspect Glob Issues Commun Sci Relat Disord 2013;3(2):54-61 [FREE Full text] [doi: 10.1044/gics3.2.54]
49. Informal Working Group on Prevention of Deafness and Hearing Impairment Programme Planning, World Health Organization. Report of the Informal Working Group on Prevention of Deafness and Hearing Impairment Programme Planning, Geneva, 18-21 June 1991. World Health Organization. 1991. URL: https://iris.who.int/handle/10665/58839 [accessed 1991-07-21]

XSL•FO
RenderX

50. Owen M, Webb M, Evans K. Community based universal neonatal hearing screening by health visitors using otoacoustic emissions. Arch Dis Child Fetal Neonatal Ed 2001 May;84(3):F157-F162 [FREE Full text] [doi: 10.1136/fn.84.3.f157] [Medline: 11320040]
51. Pirozzo S, Papinczak T, Glasziou P. Whispered voice test for screening for hearing impairment in adults and children: systematic review. BMJ 2003 Oct 25;327(7421):967 [FREE Full text] [doi: 10.1136/bmj.327.7421.967] [Medline: 14576249]
52. Olusanya B. Early detection of hearing impairment in a developing country: What options? Int J Audiol 2001 Jan;40(3):141-147 [FREE Full text] [doi: 10.3109/00206090109073109]
53. Singh AK, Kumar R, Mishra CK, Khera A, Srivastava A. Moving from Survival to Healthy Survival through Child Health Screening and Early Intervention Services Under Rashtriya Bal Swasthya Karyakram (RBSK). Indian J Pediatr 2015 Nov;82(11):1012-1018. [doi: 10.1007/s12098-015-1823-2] [Medline: 26199076]
54. Valencia K. Testing small children's hearing with little or no equipment. Community Ear Hear Health 2015;12(16):1 [FREE Full text]
55. Nolte A, Coninx F, Müller F, Hess M, Wiesner T, Dudek N, et al. Frequency-specific Animal Sound Test (FAST) 4: A valid method for hearing screening. Int J Pediatr Otorhinolaryngol 2016 Feb;81:68-79. [doi: 10.1016/j.ijporl.2015.12.004] [Medline: 26810294]
56. Clinical practice guidelines: Childhood hearing screening. American Academy of Audiology. 2011. URL: https://www.audiology.org/practice-guideline/clinical-practice-guidelines-childhood-hearing-screening/ [accessed 2022-03-07]
57. Wu W, Lü J, Li Y, Kam ACS, Fai Tong MC, Huang Z, et al. A new hearing screening system for preschool children. Int J Pediatr Otorhinolaryngol 2014 Feb;78(2):290-295. [doi: 10.1016/j.ijporl.2013.11.026] [Medline: 24360950]
58. Bevilacqua MC, Alvarenga KDF, Costa OA, Moret ALM. The universal newborn hearing screening in Brazil: from identification to intervention. Int J Pediatr Otorhinolaryngol 2010 May;74(5):510-515. [doi: 10.1016/j.ijporl.2010.02.009] [Medline: 20303604]
59. Ramkumar V, John KR, Selvakumar K, Vanaja CS, Nagarajan R, Hall JW. Cost and outcome of a community-based paediatric hearing screening programme in rural India with application of tele-audiology for follow-up diagnostic hearing assessment. Int J Audiol 2018 Jun;57(6):407-414. [doi: 10.1080/14992027.2018.1442592] [Medline: 29490519]
60. Shulman S, Besculides M, Saltzman A, Ireys H, White KR, Forsman I. Evaluation of the universal newborn hearing screening and intervention program. Pediatrics 2010 Aug;126 Suppl 1:S19-S27. [doi: 10.1542/peds.2010-0354F] [Medline: 20679316]
61. Wenjin W, Xiangrong T, Yun LI, Jingrong L, Jianyong C, Xueling W, et al. Neonatal hearing screening in remote areas of China: a comparison between rural and urban populations. J Int Med Res 2018 Feb;46(2):637-651 [FREE Full text] [doi: 10.1177/0300060517706643] [Medline: 28606020]
62. Dai P, Huang L, Wang G, Gao X, Qu C, Chen X, et al. Concurrent Hearing and Genetic Screening of 180,469 Neonates with Follow-up in Beijing, China. Am J Hum Genet 2019 Oct 03;105(4):803-812 [FREE Full text] [doi: 10.1016/j.ajhg.2019.09.003] [Medline: 31564438]
63. Bright T, Pallawela D. Validated Smartphone-Based Apps for Ear and Hearing Assessments: A Review. JMIR Rehabil Assist Technol 2016 Dec 23;3(2):e13. [doi: 10.2196/rehab.6074] [Medline: 28582261]
64. Maclennan-Smith F, Swanepoel D, Hall JW. Validity of diagnostic pure-tone audiometry without a sound-treated environment in older adults. Int J Audiol 2013 Feb;52(2):66-73. [doi: 10.3109/14992027.2012.736692] [Medline: 23140522]

Abbreviations

BOA: behavioral observation audiometry
HL: hearing loss
LMIC: low- and middle-income country
mHealth: mobile health
PTA: pure tone threshold
RETSPL: reference equivalent threshold sound pressure level
SLM: sound level meter
SRT: speech recognition task
SSAT: speech spectrum awareness task

XSL•FO
RenderX

Edited by A Mavragani; submitted 09.10.23; peer-reviewed by K Ganapathy, C Frisby; comments to author 04.09.24; revised version received 24.09.24; accepted 10.10.24; published 13.01.25.

Please cite as:
Ramkumar V, Joshi B D, Prabhakar A, Hall JW, Vaidyanath R
Development and Beta Validation of an mHealth-Based Hearing Screener (SRESHT) for Young Children in Resource-Limited Countries: Pilot Validation Study
JMIR Form Res 2025;9:e53460
URL: https://formative.jmir.org/2025/1/e53460
doi:10.2196/53460
PMID:

©Vidya Ramkumar, Deepashree Joshi B, Anil Prabhakar, James W Hall, Ramya Vaidyanath. Originally published in JMIR Formative Research (https://formative.jmir.org), 13.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Healthy Kai (Food) Checker Web-Based Tool to Support Healthy Food Policy Implementation: Development and Usability Study

Magda Rosin[1,2], PhD; Cliona Ni Mhurchu[1,2], PhD; Elaine Umali[3], MPH; Sally Mackay[1], PhD

[1]Department of Epidemiology and Biostatistics, School of Population Health, Faculty of Medical and Health Sciences, University of Auckland, Auckland, New Zealand

[2]Centre for Translational Health Research: Informing Policy and Practice (TRANSFORM), Faculty of Medical and Health Sciences, University of Auckland, Auckland, New Zealand

[3]National Institute for Health Innovation, School of Population Health, Faculty of Medical and Health Sciences, University of Auckland, Auckland, New Zealand

Corresponding Author:
Magda Rosin, PhD
Department of Epidemiology and Biostatistics
School of Population Health, Faculty of Medical and Health Sciences
University of Auckland
Building 507
22-30 Park Ave, Grafton
Auckland, 1023
New Zealand
Phone: 64 093737599 ext 88733
Email: m.rosin@auckland.ac.nz

Abstract

Background: Public health programs and policies can positively influence food environments. In 2016, a voluntary National Healthy Food and Drink Policy was released in New Zealand to improve the healthiness of food and drinks for hospital staff and visitors. However, no resources were developed to support policy implementation.

Objective: This study aimed to design, develop, and test a new web-based tool to support food providers implementing the National Healthy Food and Drink Policy in New Zealand.

Methods: The Double Diamond model, a structured framework with 4 design phases, was used to design and develop a web-based tool. Findings from our previous research, such as (1) systematic review of barriers and facilitators to workplace healthy food policy implementation; (2) scoping review of current tools and resources available in New Zealand, Australia, and Canada; (3) interviews with food providers and public health nutrition professionals; and (4) food and drink availability audit results in New Zealand hospitals were used in the “Discover” (understanding of current gaps) and “Define” (prioritizing functions and features) phases. Subsequent phases focused on generating ideas, creating prototypes, and testing a new web-based tool using Figma, a prototyping tool. During the “Develop” phase, project stakeholders (11 public health nutrition professionals) provided feedback on the basic content outline of the initial low-fidelity prototype. In the final “Deliver” phase, a high-fidelity prototype resembling the appearance and functionality of the final tool was tested with 3 end users (public health nutrition professionals) through interactive interviews, and user suggestions were incorporated to improve the tool.

Results: A new digital tool, Healthy Kai (Food) Checker—a searchable database of packaged food and drink products that classifies items according to the Policy’s nutritional criteria—was identified as a key tool to support Policy implementation. Of 18 potential functions and features, 11 were prioritized by the study team, including basic and advanced searches for products, sorting list options, the ability to compile a list of selected products, a means to report products missing from the database, and ability to use on different devices. Feedback from interview participants was that the tool was easy to use, was logical to navigate, and had an appealing color scheme. Suggested visual and usability improvements included ensuring that images represented the diverse New Zealand population, reducing unnecessary clickable elements, adding information about the free registration option, and including more frequently asked questions.

Conclusions: Comprehensive research informed the development of a new digital tool to support implementation of the National Healthy Food and Drink Policy. Testing with end users identified features that would further enhance the tool’s acceptability and

XSL•FO
RenderX

usability. Incorporation of more functions and extending the database to include products classified according to the healthy school lunches program policy in the same database would increase the tool's utility.

(JMIR Form Res 2025;9:e60447) doi:10.2196/60447

KEYWORDS

healthy food policy; policy implementation; nutrition; food environment; hospital; workplace; web-based tool; digital tool; database; user testing; food; drink; nutrition professionals; acceptability; usability

Introduction

Background

Despite the importance of healthy eating in the prevention of chronic diseases, the nutritional quality of diets worldwide does not align with evidence-based dietary guidelines, with widespread overconsumption of unhealthy and ultraprocessed foods and drinks [1,2]. However, foods are chosen in the context of local food environments [3,4], which in New Zealand are dominated by less healthy choices [5]. Workplace food environments, including hospitals, have been recognized as an important factor for employees' health promotion and well-being due to their potential for broad and sustainable population reach [6], as well as the fact that the majority of the adult population works [7-9]. When this study began, the New Zealand health system was structured into 20 district health boards (DHBs) responsible for delivering national health care services through hospitals and clinical centers. However, in 2022, the DHB system was disestablished as part of a national health system reform [10]. The provision of health care services was centralized under the newly established Health New Zealand—Te Whatu Ora agency that directly employs around 90,000 staff, making it a major nationwide employer [11].

A voluntary National Healthy Food and Drink Policy (the Policy) was released in New Zealand in 2016 [12] to encourage hospitals to provide healthier food and drink options for staff and visitors (not inpatients). The Policy includes a traffic light classification system for food and drinks, where *healthy* (*Green*) options should make up at least 55% of available items; the *less healthy* (*Amber*) options that might still provide some nutritive value should be less prevalent (<45% of available items); and the *unhealthy* (*Red*) items, such as confectionery, deep-fried foods and sugar-sweetened drinks, are not permitted [12]. Yet, 2021 evaluation findings showed that 5 years after the introduction of the voluntary Policy in 2016, only 8 of the 20 DHBs had adopted the Policy, resulting in regional inconsistency [13]. Furthermore, on-site audits of food and drink availability in 2021/2022 indicated that the implementation was inconsistent and largely unsuccessful, with *Red* and *Amber* choices predominating [14].

Previous Work

We conducted a systematic literature review of facilitators and barriers to implementing healthy food and drink policies in public sector workplaces. The review indicated that tailored supportive tools and resources facilitate the implementation of policies promoting healthy food environments [15]. A key reason why such support is required is because identifying healthier items can be challenging for food providers, who may find the policy's nutrition criteria too complex to understand without nutrition expertise [15]. The supportive tools identified in the systematic review [15] were closely aligned with the supporting materials recommended by the World Health Organization [16] in their 2021 action framework for the implementation of effective healthy food policies in public settings. Examples of useful tools included administrative materials (eg, contract templates and standardized procedural guidelines); customer communication resources, compliance monitoring tools; and databases of compliant recipes, suppliers, and products [15,16]. The most useful tools were policy specific, readily available, and centrally provided, limiting the burden on individual food providers [15]. Importantly, supportive tools needed to be practical, up-to-date, user-friendly, and useful to those implementing a healthy food policy [15].

As part of a scoping review, we identified and evaluated the currently available tools used in public sector workplace policy implementation in New Zealand, Australia, and Canada [17]. Just 2 paper-based tools were related to the New Zealand Policy: guidelines to make better pies [18] and an event and fundraiser guide aimed at schools and community groups [19], and they were likely insufficient to facilitate implementation of the Policy [17]. No other tools were developed to support the New Zealand Policy implementation in contrast to most Australian food environment policies. Nevertheless, the scoping review [17] provided information on the features, usability, and quality of the available tools, which can be used to guide the development of new tools or update of existing tools. For example, for digital tools, which were less common than paper-based tools, logical navigation, clear search and result structure, consistent style, and unambiguous indication of products' compliance in a web-based database were perceived as favorable features [17].

To identify which implementation tools were needed in New Zealand, we interviewed hospital food providers and public health nutrition professionals who were members of the National Food and Drink Environments Network (the Network) supporting implementation of the Policy in their respective DHBs [20]. Apart from customer communication materials, which are in development by Health New Zealand (the Network virtual meetings, 2023), interview participants suggested 2 digital tools that could support the New Zealand Policy implementation: a food and drink audit tool and an up-to-date database showing the traffic light classification of packaged products. Our research team previously developed a digital monitoring tool [21] with an embedded traffic light classification algorithm to systematically conduct the on-site audits of food and drinks available in New Zealand hospitals [14]. The second tool, a web-based product database, was requested because food providers and public health nutrition professionals spent

XSL•FO
RenderX

significant time looking for products on the web and manually assessing individual products against the Policy criteria [20]. Furthermore, there has been no platform for food manufacturers and suppliers to communicate with the hospital food providers about food and drink products that may comply with the Policy [20].

Rationale and Objective

Usability and appearance are likely to impact the usage of web-based tools. To develop the most optimal and feasible tool that meets the needs of its end users, user experience and user interface (UX/UI) design principles and user testing are needed [22]. A user-centered design approach is commonly used in fields that develop and use digital tools such as health care management [23], grocery shopping websites [22], and physical activity apps [24]. However, there is limited research on developing and testing tools designed to support the implementation of healthy food and drink policies.

The tool development was a component of the Healthy Policy Evaluation (HYPE) study that comprehensively evaluated the implementation of the National Healthy Food and Drink Policy in New Zealand and determined which tools and resources were needed to support Policy implementation. This paper aims to describe the development and user testing of a new prototype web-based tool to facilitate classification of packaged products according to the National Healthy Food and Drink Policy traffic light criteria.

Methods

Overview

This research was underpinned by pragmatism as the methodological paradigm [25-27]. We focused on practical and carefully considered decisions to develop the most effective and efficient prototype tool to facilitate implementation of the New Zealand Policy by incorporating a user-centered approach that emphasized the needs and preferences of product end users (hospital food providers and Network members [public health nutrition professionals]) [28,29]. The design and development of the new tool were guided by the Double Diamond model (Figure 1) created by the UK Design Council [30]. The model outlines an iterative and flexible process of identifying a real-world problem or gap, defining a possible solution, and finally developing a tangible, valuable, and well-informed outcome addressing the identified problem or gap [30]. The framework is not linear or rigid, and the phases and diamonds can overlap to produce the most optimal tool [30,31] and represent the evolving innovation process and continually changing digital space [30]. The Double Diamond model has been revised over the years by integrating additional steps within the framework, although the principles and the 4 major phases (Discover, Define, Develop, and Deliver) remain largely unchanged [32]. The Double Diamond phases were completed with the assistance of a subcontracted UX/UI designer, and study authors reviewed and provided feedback on all design and development outputs to ensure accurate representation and alignment with end user needs and the aims and scope of the Policy.

Figure 1. The Double Diamond model, illustrating the 4 phases (Discover, Define, Develop, and Deliver) used to guide the development of the Healthy Kai (Food) Checker web-based tool. Arrows pointing toward the tips of the diamonds demonstrate the diverging phases of the process characterized by creative thinking and the expansion and exploration of different ideas and possibilities. Arrows pointing away from the tips of the diamonds reflect the converging phases of the process where progressive decision-making occurs by focusing, prioritizing, and narrowing down the ideas and possibilities toward the most feasible and relevant solution. Circular arrows within the diamonds illustrate the iterative nature of the design process (adapted from the UK Design Council's Double Diamond model [30] under Creative Common License CC BY 4.0 [52]).

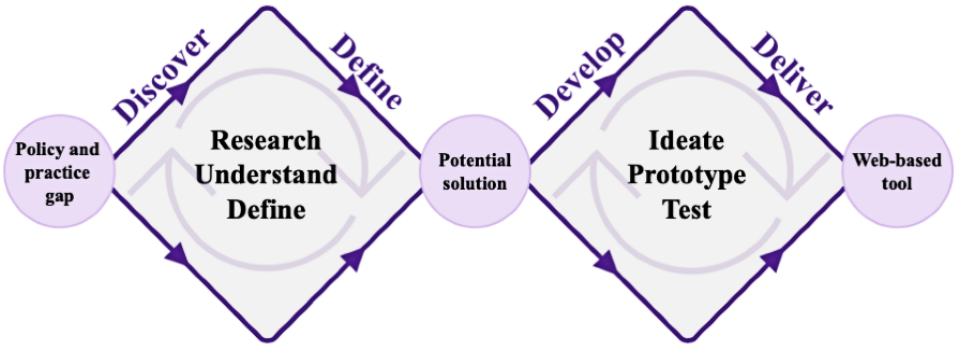

The UX/UI design approach could be broadly defined as "the process of designing physical or digital products that are useful, easy to use, and provide a great experience in interacting with them" [28]. Key UX/UI design principles [28,29,34] applied in this study were as follows:

1. User-centered design: designing and developing focused on the needs and preferences of end users (prioritizing accessibility and usability) and collaborating and seeking feedback from key stakeholders throughout the process.
2. Clear information hierarchy: arranging and presenting information in the web-based tool in a logical and organized way to ensure easy and user-friendly navigation and understanding of the content.
3. Visual and interaction consistency: maintaining a consistent and harmonious visual style to achieve a cohesive and professional look and ensuring predictable and consistent performance of the same style elements.
4. Feedback and confirmation: providing instant feedback and confirmation messages on successful user actions and warnings to prevent user errors (eg, deletion).

XSL•FO
RenderX

5. Mobile responsiveness: creating a seamless experience on both desktop and mobile platforms by optimizing the design to adapt to various devices.

Ethical Considerations

Ethical approval for the HYPE study was granted by the Auckland Health Research Ethics Committee (reference AH2519). Subsequent locality approvals were sought from individual DHBs. All usability testing participants provided written consent to participate in the virtual interactive interviews. We approached potential participants up to 3 times by email between May and August 2023, explaining the purpose of the study and providing a detailed participant information sheet explaining the testing procedure. Participants who agreed to take part were sent a consent form, which they signed and returned by email. Participants had the option to have their *whānau* (family) present during the interview and were reimbursed for their time with a New Zealand $50 (US $30) grocery voucher. Study data were deidentified to ensure confidentiality and privacy.

Discover Phase

In this phase, we compiled findings from other components of the HYPE study outlined in the *Introduction* section [14,15,17,20] relevant to designing and developing the new tool. The findings included thematic analysis codes generated during the coding of studies in the systematic review [15] and interview data [20], as well as the features, usability, and quality of the tools assessed in the scoping review [17]. Furthermore, we obtained screenshots of the digital tools included in the scoping review [17] and the websites of the major food suppliers, distributors, and retailers in New Zealand. We collated the findings using the whiteboard function in the web-based Miro software tool (RealtimeBoard, Inc).

Define Phase

In the second phase, we analyzed and refined findings and outputs compiled in the Discover phase to define a feature and function requirement blueprint for the new tool. First, we created a user persona of a fictional end user; defined challenges they might face (pain points); and described their needs, actions they want to take, and expected benefits of using the new tool (a user story) [28,34]. Second, we further reviewed similar available web-based tools [17] in Australia [35,36] and Canada [37] and compared them on their UX/UI aspects of "first impressions," "website visual design" features, and "website interaction" (features, accessibility, user flow, and navigation) [28,29,34].

In the next step, we used the Must have, Should have, Could have, Won't have (MoSCoW) prioritization framework to define features and functions that were deemed "Must have" (essential and critical), "Should have" (important but not critical), "Could have" (desirable but not critical), and "Won't have" (out of scope and excluded in the current design but could be considered in the future) [38]. We considered features identified under the "Must have" and "Should have" as core elements to be incorporated into the new prototype web-based tool.

In the final step, we focused on understanding and carefully planning the path a user would take when interacting with the web-based tool from the initial starting point (eg, home page) to the final step (eg, finding suitable products and checking their classification) [28,34]. The user flow diagram was an important starting point for creating an efficient and easy-to-use web-based tool [39]. The corresponding site map outlined the hierarchical structure of the tool's pages, indicating how pages are linked, grouped, and prioritized in the structural tree [40], and we progressively simplified it to essential pages that aligned with the user flow diagram content.

Develop Phase

We used Figma, a prototyping and design tool, for the remaining Develop and Deliver phases. In the Develop phase, we created a wide range of ideas for the prototype web-based tool underpinned by the outputs from the Define phase and the UX/UI principles outlined earlier. We started with wireframes, a basic arrangement of text, images, and buttons on the web pages before any additional visual elements or content were included [28].

The subsequent design included low-fidelity prototypes containing more details than wireframes, allowing us to test and refine ideas and concepts [28,41]. We engaged with the Network (n=36 members, consisting of public health professionals and dietitians representing their respective DHBs, as well as nutrition and public health advisors from the Ministry of Health and New Zealand Heart Foundation) as key project stakeholders and asked for their feedback and input on the low-fidelity prototypes before committing to a specific design. The Network was provided with a document via email containing different design options alongside a small set of questions about their preferred features, structure, layout, and content (Multimedia Appendix 1). Feedback was collated during the subsequent monthly Network virtual meeting (11 members were present) and used to refine the prototype further.

Deliver Phase

Overview

In the Deliver phase, we developed a comprehensive high-fidelity prototype with interactive and functional elements, such as clickable content, buttons, and menus [28,41]. The tool's pages were linked, mirroring the navigational flow of a live website and providing a realistic user experience, which allowed us to test user interaction with the prototype and refine the user journey [28,41]. We tested the prototype multiple times and once with a colleague to further improve and refine the high-fidelity prototype before testing it with end users.

Interactive Interviews to Test Usability of the Prototype Web-Based Tool

We conducted usability testing that involved interactive virtual interviews in which user interaction and engagement with the prototype were evaluated to determine user preferences and behaviors related to the prototype tool [42]. The testing also identified UX/UI challenges and pain points that needed addressing or redesigning to improve the tool [42].

XSL•FO
RenderX

Participant Selection and Recruitment

Potential participants (end users) were food providers (for staff and visitors) and members of the Network (public health nutrition professionals) within any of the New Zealand hospitals who had previously participated in the stakeholder interviews (n=12) [20] and who expressed willingness to test the new web-based tool, as well as other key stakeholders who had previously used the Policy and were familiar with its traffic light classification system.

Data Collection

The usability test interview combined 2 methods: performance testing [43], in which participants were given tasks to complete while interacting with the prototype, and visual design testing [44], in which participants were asked to evaluate visual aspects of the prototype tool. The usability interview guide (Multimedia Appendix 2) included questions about the prototype tool's engagement, intuitive usage, functionality, navigation, quality, and user-friendliness, as well as scenarios and tasks related to the tool's functions.

At the beginning of the virtual interviews, we explained the research purpose (testing the usability and user experience of the design product and not the participants' ability to use the prototype), process, and participants' rights; answered any questions; and verbally confirmed the permission to video record the interviews (as previously agreed to on the consent form). We provided participants with a digital link to access the Figma prototype and asked them to share their screen to demonstrate their interaction with the prototype's features and functions. Participants were encouraged to think aloud during testing and to comment on the design, their preferences, and how the tool could be improved. All interviews were transcribed verbatim, and participants were given the opportunity to review and edit their transcripts.

Data Analysis

One author watched interview recordings multiple times and reviewed transcripts to analyze how participants interacted with the tool; how intuitive, engaging, efficient, and accessible the prototype tool was; and what changes were needed to improve the user experience. Participants' suggestions, comments, and interactions were recorded on the printouts of the prototype. The results were shared and discussed with the UX/UI designer, who incorporated the feedback to improve and refine the design of the prototype. Textbox 1 summarizes the methods used and outputs in the phases of the Double Diamond model.

Textbox 1. Phases and outputs of the Double Diamond model used in the design and development of the new prototype Healthy Kai (Food) Checker web-based tool to support implementation of the National Healthy Food and Drink Policy in New Zealand hospitals.

Discover—divergent research phase **(understanding user needs and challenges they face, exploring similar tools available)**

- A systematic review [15] and scoping review findings [17] about tools and resources used internationally in implementation of healthy food and drink policies.
- Healthy Policy Evaluation (HYPE) study interview findings [20] specific to the characteristics and needs of the target audience (food providers and Network members [public health nutrition professionals]).
- HYPE study food and drinks audit findings [14], including types of products available.
- Layout and structure characteristics of existing digital tools [17] and major retailer and supplier websites.

Define—convergent synthesis phase **(analyzing and refining insights generated in the Discover phase)**

- User persona of a fictional food provider, including their pain points and user story.
- Review and comparison of the user experience and user interface features of similar tools [35-37].
- Requirements blueprint for the new tool (refined by considering project timelines and the available budget).
- New tool features prioritized with the Must have, Should have, Could have, Won't have (MoSCoW) framework [38].
- Visual user flow diagram (outlining actions, decisions, and steps users take through the tool).
- Structural site map of the pages in the tool.

Develop—divergent ideation phase **(creating basic outlines of the prototype interface)**

- Wireframes, a basic skeletal outline of the website interface and elements.
- Low-fidelity prototype (a simple, monotone, and static representation of a prototype).
- Informal feedback from project stakeholders on the low-fidelity prototype.

Deliver—convergent implementation phase **(refining of the prototype and usability testing)**

- High-fidelity prototype (detailed, interactive, and realistic representation closely resembling the final web-based tool).
- Interactive usability testing with end users.

XSL•FO
RenderX

Results

Define Phase

The user persona of a fictional food provider, including their characteristics, challenges, and user story, is summarized in Figure 2. Based on the user persona, a new web-based tool, Healthy Kai (Food) Checker—a searchable database of packaged food and drink products showing their traffic light classification—was identified as a key tool to support the New Zealand Policy implementation. Textbox 2 summarizes positive UX/UI attributes [17] we identified based on the review of similar available tools from Australia, Healthy Food Finder [36] in New South Wales, and FoodChecker [35] in Victoria, as well as a Canadian tool, Brand Name Food List [37], available in British Columbia. The function and feature requirements of the Healthy Kai Checker prioritized with the MoSCoW framework as "Must have" and "Should have" are indicated in Textbox 3.

The Healthy Kai Checker user flow (Figure 3) and functionality design included elements closely related to the "Must have" and "Should have" features outlined in Textbox 3 [38], for example, registration to create a free account, personalization to apply search filters and create lists of favorite products, educational content to explain terms and provide tips on how to use the tool, and a feedback mechanism to report an issue or suggest improvements. The corresponding Healthy Kai Checker site map is available in Multimedia Appendix 3.

Figure 2. User persona, pain points, and user story of a fictional hospital food provider in New Zealand. The figure illustrates the characteristics, challenges, and narrative of a typical hospital food provider, highlighting their needs and experiences in implementing the National Healthy Food and Drink Policy and expected benefits of using the new tool.

XSL•FO
RenderX

Textbox 2. Favorable user experience and user interface features of web-based tools with product databases based on the review and comparison of similar existing web-based tools [17] in Australia [35,36] and Canada [37].

First impressions

- Visually appealing (clear, fresh, and crisp look) and clean design
- Good use of white space for a less compact layout
- Colorful
- Modern feel

Website visual design

- Consistent visual style
- Consistent and clear branding
- Clear font
- Simple design
- Good color scheme and use of images
- A good indication of clickable elements

Website interaction

- Automatic suggestion of products in the search bar (predictive text)
- A clear indication of product compliance/traffic light classification
- Results displayed based on relevance to the search term (eg, for misspelled words)
- Product information in separate columns (name, brand, and size)
- Good column sorting function (both directions)
- Relevant search filters
- Explanation of why the product does not meet the Policy criteria
- Option to report missing products
- Ability to create own product lists
- Ability to return to previous screens without loss of content
- Confirmation of completed action
- Intuitive navigation

XSL•FO
RenderX

Textbox 3. Features and functions of the Healthy Kai (Food) Checker web-based tool prioritized with the Must have, Should have, Could have, Won't have (MoSCoW) framework [38]. "Must have" features and functions were considered essential and critical, "Should have" were important but not critical, "Could have" were desirable but not critical, and "Won't have" were out of scope and excluded in the current design but could be considered in the future. "Must have" and "Should have" were core elements to be incorporated into the new prototype web-based tool to support implementation of the National Healthy Food and Drink Policy in New Zealand hospitals.

Must have

- A searchable list of packaged products (including *Red* products so that end users can quickly identify items that do not comply with the Policy instead of spending time classifying them manually).
- Ability to search by name, keyword, and brand.
- Basic and advanced search options (traffic light criteria and food group filters).
- A clear indication of the traffic light classification (text and color).
- Up-to-date product database (a database of products was created during the food and drink audits conducted as part of the HYPE study [14] that can be integrated).
- Engaging information about the Policy and Healthy Kai Checker.

Should have

- Product information separated into columns (name, size, and brand).
- Sort results function (alphabetically, product size, and traffic light classification).
- Ability to report a missing or incorrect products.
- Free registration requirement to view all products and unlock full access.
- Free registration requirement to create, save, export, and print product lists.
- No paid registration (free for all users).
- Mobile responsiveness for different devices.

Could have

- Detailed information about each product (eg, nutritional, ingredient list, and allergens)
- Clear product photographs.
- Filters related to allergens, suitability for vending machines.
- Explanation or indication of why a product does not meet the *Green* criteria.

Won't have

- Self-assessment of products and recipes against the Policy criteria.
- Integration of healthy food and drink policies for New Zealand schools and early childhood education centers, or government-funded healthy school lunches program.
- Web-based Policy-relevant training modules.

XSL•FO
RenderX

Figure 3. User flow diagram of the Healthy Kai (Food) Checker web-based tool. This diagram illustrates the user journey through the tool, detailing steps from initial log-in to final product selection. The user flow is designed to support users in implementing the National Healthy Food and Drink Policy in New Zealand hospitals. The green ovals show actions taken by the users, while the light purple diamonds indicate points where users choose their next action. Grayed-out rectangles show features classified as "Could have" or "Won't have" according to the Must have, Should have, Could have, Won't have (MoSCoW) prioritization framework, reflecting their lower priority and exclusion from the initial prototype tool.

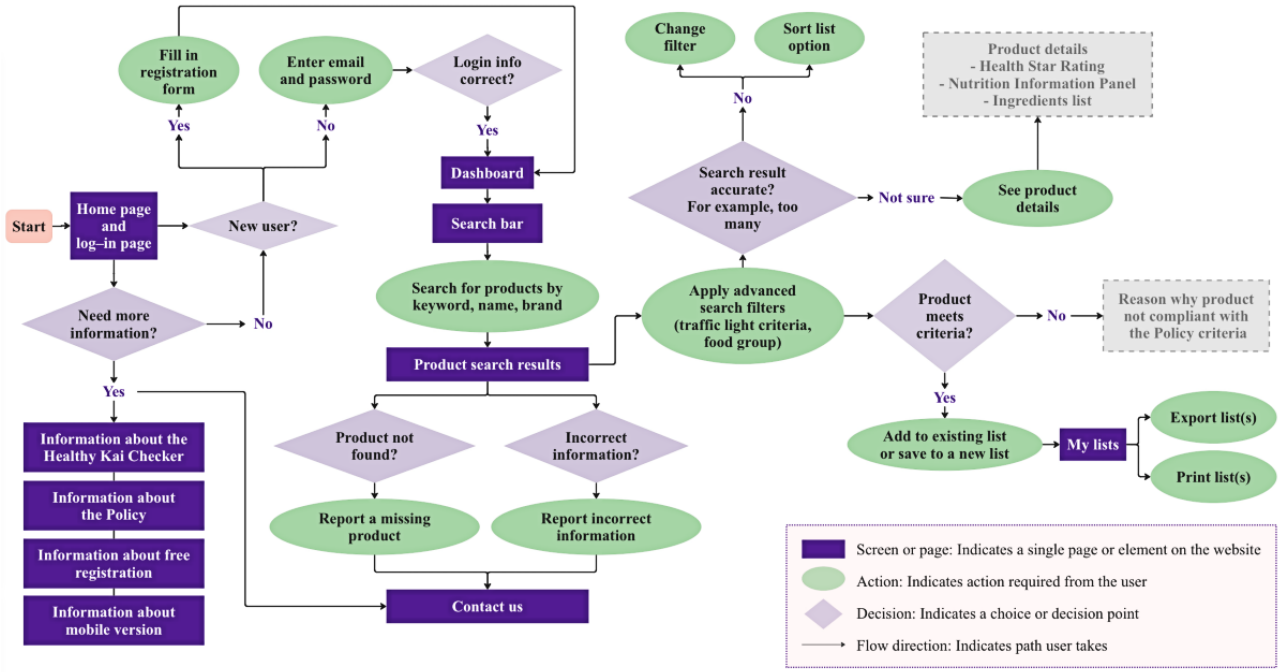

Develop Phase

The initial wireframe outline and the low-fidelity prototype for the home page are shown in Figure 4. The low-fidelity prototypes of the 2 design options for the product search and result dashboard are available in Multimedia Appendix 1. Feedback from the 11 Network members and the research team favored the option where the search box and result filters are at the top rather than on the side of the search page.

XSL•FO
RenderX

Figure 4. Wireframes and low-fidelity prototype of the Healthy Kai (Food) Checker web-based tool home page. This figure represents the initial design concepts for the tool, outlining the layout and structure of the home page and showcasing key elements of the interface design and basic functionality. The feedback from 11 public health nutrition professionals on the low-fidelity prototype was used to refine the web-based tool to support implementation of the National Healthy Food and Drink Policy in New Zealand hospitals.

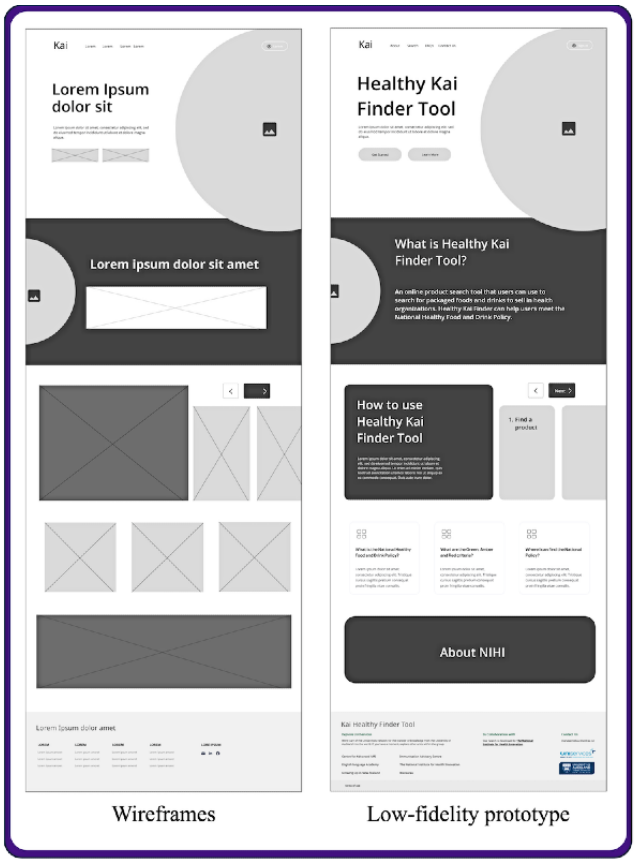

Deliver Phase (Usability Testing)

Twelve individuals were invited to take part in the interactive interviews, of whom 1 declined due to lack of time, 2 were no longer employed in the same role, and 6 did not respond after 3 invitation emails. The final interview sample included 2 Network members (public health nutrition professionals) and 1 stakeholder who used the Policy in public health research on food environments. The interviews took an average of 49 (range 40-59) minutes to complete.

Participants generally found the Healthy Kai Checker prototype easy to use and logical to navigate. They appreciated the hyperlinks in the prototype that directed users to external web pages, such as the Ministry of Health website where the Policy was published. There was also an appreciation that the different pages within the tool could be reached through multiple buttons or links (ie, having multiple paths to 1 page), and participants provided additional suggestions on where specific buttons connecting to the various web pages of the tool or external links should be incorporated. All participants liked the color scheme and the purple and green palette, which made the tool look vibrant and fresh and differentiated it from other public health or government websites that typically use more muted colors. One suggestion was to ensure that images were representative of the diverse New Zealand population. The Healthy Kai Checker was positively perceived due to its simple and concise interface, which did not overwhelm users with too much information.

On the home page, participants suggested including more information about the free website registration and its benefits and placing that component closer to the top of the web page. The section informing users about the ability to use the tool on mobile or tablet devices was positively perceived and thought to be necessary in the current technology-driven working environments. However, a feature in the middle of the home page, “How to use the Healthy Kai Product Database” that contained a clickable carousel of moving parts from right to left, was deemed confusing. Participants suggested reducing the steps showing information on how to use the tool from 4 to 3 and displaying all of them at once on the main page, thus improving usability by eliminating the need to click on the various components to see the entire content.

On the product search and results page and the pages showing products saved in the user’s individual lists, the indication of products’ categorization as *Green*, *Amber*, or *Red* was perceived as unambiguous. The favorable features were the ability to apply advanced search filters (traffic light classification and the Policy’s food group) to display only those products with the specified characteristics. Three suggestions for improvements were incorporating an indication of the number of products in the search results and personalized product lists, making the “My Lists” feature more prominent, and making some of the buttons and writing bigger to improve visibility. Participants also thought that allowing manufacturers and suppliers to submit their products for classification and inclusion in the database,

XSL•FO
RenderX

along with product photographs, was a good idea, and having a "Contact Us" page on the website that clearly outlined the steps required would facilitate this.

For the remaining pages, participants suggested some frequently asked questions that users would likely be interested in, since the Network members had experience working closely with food providers to implement the Policy. Furthermore, minor aesthetic adjustments, such as reducing plain text and adding graphical elements or photographs, were suggested to make the pages more visually appealing. Screenshots of the most recent version of the Healthy Kai Checker prototype tool are shown in Figure 5.

Figure 5. Screenshots of the various pages of the Healthy Kai (Food) Checker prototype web-based tool designed to support implementation of the National Healthy Food and Drink Policy in New Zealand hospitals. The tool is a searchable database of packaged food and drink products that classifies items according to the Policy's nutritional criteria, with functions such as basic and advanced searches for products, sorting list options, the ability to compile a list of selected products, a means to report products missing from the database, and responsiveness for usage on different devices. FAQ: frequently asked questions.

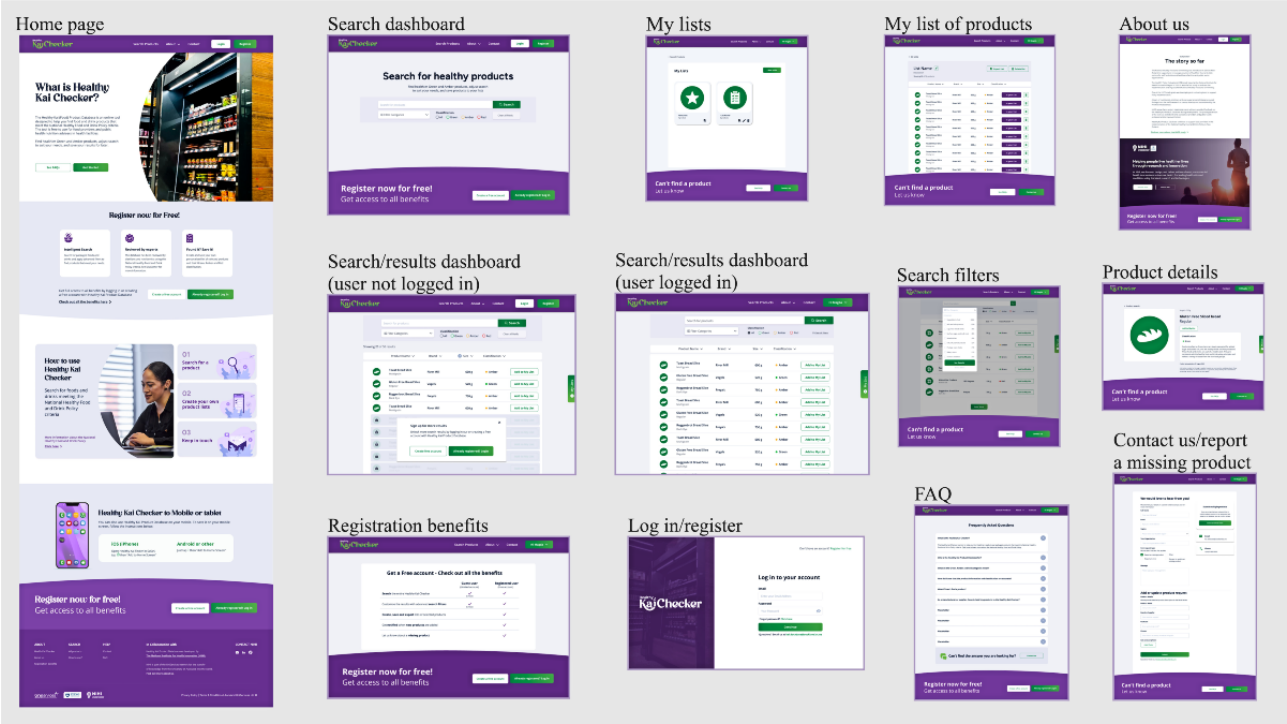

Discussion

Principal Findings

We aimed to develop and test a new evidence-based, web-based prototype tool to support the adoption and implementation of the National Healthy Food and Drink Policy in New Zealand. The new Healthy Kai Checker web-based tool will reduce the burden on individual food providers and Network members associated with identifying and classifying products according to the Policy's traffic light nutrition criteria. Until now, individual food providers and organizations had to undertake this essential implementation task manually, which is a repetitive and resource-intensive process.

The choice of a web-based tool in this study was driven by the increasingly digital or web-based practices and systems [45] that already saturate the grocery and retail market [22] and are increasingly used for individual behavioral change approaches [46] and research on food supply [47] and physical activity [24]. The scarcity of web-based or digital tools in New Zealand [17] for quickly checking the classification of packaged food and drink products represents an opportunity for the new Healthy Kai Checker to fill this significant gap and facilitate implementation of the Policy. Healthy Kai Checker was designed as an intuitive and logical web-based tool suitable for end users who may have limited nutrition knowledge and experience, and understanding of technical nutrition jargon, which can be overwhelming and complex [48-50]. Furthermore, the aim was to optimize user paths, accommodate varying levels of technological proficiency, and minimize the steps end users had to take [39] to search and identify suitable products.

The design and development of the Healthy Kai Checker followed the rigorous but flexible Double Diamond model [30], was guided by key UX/UI design principles [28,29,34], and was based on the comprehensive research our team conducted as part of the HYPE study [14,15,17,20]. While other design and development strategies could have resulted in a different or better tool, we are unable to make a direct comparison since the design and development process of similar tools in Australia [35,36] and Canada [37] had not been outlined in the literature. However, our approach provided key advantages of emphasizing and understanding user needs and stakeholder engagement, mitigating potential risks and challenges early in the process, and allowing iteration and creativity, leading to a unique, tailored, and effective solution to the identified resource gap [30,31].

XSL•FO
RenderX

Implementation Considerations

To ensure a seamless user experience for all intended end users, Healthy Kai Checker would need to be further tested with food providers in New Zealand hospitals. Furthermore, expanding the current features to include other healthy food policies in New Zealand, such as the government-funded healthy school lunches program policy (Ka Ora, Ka Ako) [51], would require user testing with food providers working in school environments. However, it should be noted that usability testing and the incorporation of end user feedback are an ongoing process in the evolving digital space and a critical component of the product's road map [30]. After Healthy Kai Checker is developed and released as a fully functioning web-based tool, it would need to be periodically updated to ensure that it maintains user-centered design and continues to offer advantages over other potential tools in the market.

Another key aspect of the Healthy Kai Checker is the ongoing management and maintenance of the food and drink product database that underpins the tool, as well as new updates of the algorithm embedded into the tool to facilitate the classification of products as *Green*, *Amber*, or *Red* [21]. The current database contains information on food and drinks collected during food availability audits in New Zealand hospitals as part of the HYPE study [14]. However, these data were captured in 2021/2022, and it is likely that, at some point, they will no longer accurately represent the supply on the market. Further manual field data collection, as used for the HYPE study, is likely too resource-intensive and costly to do regularly. Reaching out to the major New Zealand suppliers to encourage them to provide nutritional information regularly on their products could be a cost-effective strategy to streamline data collection and maintain the accuracy and currency of the dataset.

Another avenue includes obtaining product information directly from Healthy Kai Checker end users, as seen with crowdsourcing of supermarket data from the FoodSwitch [47] app users in Australia. The Healthy Kai Checker includes a feature where users can directly provide the necessary information and photographs about a missing or incorrect product for evaluation. An additional emerging method of data collection is the "web harvesting" or "data mining" approach used to automatically extract information from websites, such as web-based supermarket stores [52], and could also be explored to source data for the Healthy Kai Checker.

Regardless of the source of product data, classifying individual products into *Green*, *Amber*, and *Red* categories using the Policy's criteria can be complex because some of the required information is not provided on the food packaging. The challenging classification of some products with the Policy criteria was observed during food and drink availability audits as part of the HYPE study [14], although a robust algorithm was developed and included in the web-based audit tool to make the process systematic [21]. Simplifying the Policy's criteria could increase food provider and supplier understanding of the guidelines and streamline data collection to keep the product database up-to-date. Importantly, a classification algorithm would need to be revised if the Policy undergoes an update of its nutritional criteria, representing an additional cost associated with ongoing maintenance of the Healthy Kai Checker.

Strengths and Limitations

There are several strengths in this study. First, the design and development of the Healthy Kai Checker were informed by comprehensive research findings and aligned with the needs of the Policy implementers. Second, the UX/UI designer used the latest website design techniques and best practices to ensure the user-friendliness of the tool. Third, feedback and suggestions provided by stakeholders and participants were incorporated into the tool to improve its usability. However, there were some limitations. Although the Healthy Kai Checker prototype has been tested with 1 group of potential end users, testing has not yet been conducted with the New Zealand food providers because they did not respond to the invitation to take part in the research. Future research should explore the usability of the prototype tool with food providers and stakeholders associated with other policies or settings. User testing should also be conducted once the Healthy Kai Checker is a fully functioning website. Since the tool is still a prototype, this study did not explore the effectiveness and impact of the Healthy Kai Checker as a tool in the Policy implementation. Research on the effectiveness of tools and resources to support healthy food policy implementation is generally lacking, and future studies should focus on evaluating the effectiveness of tools in this space.

Conclusions

This study comprehensively designed and developed a new web-based Healthy Kai Checker tool, a searchable database of packaged food and drink products, to facilitate the implementation of the National Healthy Food and Drink Policy in New Zealand. The development of the tool was guided by the Double Diamond model and UX/UI design principles. Feedback received from project stakeholders was incorporated into the prototype, and comprehensive user-testing results further improved the design and usability of the prototype tool. The Healthy Kai Checker could provide valuable support to New Zealand hospital food providers and others who support implementation of the Policy.

Acknowledgments

The authors would like to thank and acknowledge Alexandra Rengifo, user experience and user interface (UX/UI) designer, for her help in conceptualizing the tool and developing the user-testing guide and the prototype tool, and thank the participants for taking part in the study. This research was funded by the Healthier Lives National Science Challenge as part of the Healthy Policy Evaluation (HYPE) study.

XSL•FO
RenderX

Data Availability

The interview data analyzed in this study are not publicly available due to presence of sensitive and personally identifiable qualitative data that could compromise the anonymity of the participants. The remaining datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

MR, CNM, and SM conceptualized the study. MR recruited participants and collected, managed, and analyzed the data. MR and EU conceptualized the tool and developed user-testing guide. MR wrote the first draft and is responsible for the paper's final content. All authors contributed to the interpretation of results, critically revised the paper, and approved the final version.

Conflicts of Interest

MR and CNM were members of the National DHB Food & Drink Environments Network. The Network had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. SM and EU declare no conflict of interest.

Multimedia Appendix 1
Low-fidelity prototype feedback document.
[PDF File (Adobe PDF File), 577 KB - formative_v9i1e60447_app1.pdf]

Multimedia Appendix 2
Usability testing of the Healthy Kai Checker interactive interview guide.
[PDF File (Adobe PDF File), 186 KB - formative_v9i1e60447_app2.pdf]

Multimedia Appendix 3
Healthy Kai Checker site map.
[PDF File (Adobe PDF File), 190 KB - formative_v9i1e60447_app3.pdf]

References

1. Imamura F, Micha R, Khatibzadeh S, Fahimi S, Shi P, Powles J, et al. Dietary quality among men and women in 187 countries in 1990 and 2010: a systematic assessment. Lancet Glob Health 2015;3(3):e132-e142 [FREE Full text] [doi: 10.1016/S2214-109X(14)70381-X] [Medline: 25701991]
2. Miller V, Webb P, Cudhea F, Shi P, Zhang J, Reedy J, et al. Global dietary quality in 185 countries from 1990 to 2018 show wide differences by nation, age, education, and urbanicity. Nat Food 2022;3(9):694-702. [doi: 10.1038/s43016-022-00594-9] [Medline: 37118151]
3. Healthy food environments. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/nutrition/healthy-food-environments/ [accessed 2024-12-16]
4. Influencing food environments for healthy diets. Food and Agriculture Organization (FAO). 2016. URL: https://www.fao.org/fsnforum/resources/reports-and-briefs/influencing-food-environments-healthy-diets [accessed 2023-10-15]
5. Mackay S, Garton K, Gerritsen S, Sing F, Swinburn B. How Healthy Are Aotearoa New Zealand's Food Environments? Assessing the Impact of Recent Food Policies 2018–2021. Auckland, New Zealand: University of Auckland; 2021. URL: https://tinyurl.com/v2hysyht
6. Chu C, Breucker G, Harris N, Stitzel A, Gan X, Gu X, et al. Health-promoting workplaces—international settings development. Health Promot Int 2000;15(2):155-167. [doi: 10.1093/heapro/15.2.155]
7. Chu C, Driscoll T, Dwyer S. The health-promoting workplace: an integrative perspective. Aust N Z J Public Health 1997;21(4 Spec No):377-385 [FREE Full text] [doi: 10.1111/j.1467-842x.1997.tb01718.x] [Medline: 9308202]
8. Labour market statistics March 2023 quarter summary. Stats New Zealand. 2023. URL: https://tinyurl.com/54bzedhu [accessed 2023-10-15]
9. 2018 Census place summaries. Stats New Zealand. 2018. URL: https://www.stats.govt.nz/tools/2018-census-place-summaries/new-zealand#health [accessed 2023-09-08]
10. Changing the system. Te Whatu Ora—Health New Zealand. URL: https://www.tewhatuora.govt.nz/whats-happening/changing-the-system/ [accessed 2023-09-21]
11. Te Whatu Ora—Health New Zealand. Health workforce plan 2023/24. 2023. URL: https://www.tewhatuora.govt.nz/publications/health-workforce-plan-202324/ [accessed 2023-08-28]
12. National Healthy Food and Drink Policy (2nd ed). Ministry of Health. National District Health Board Food and Drink Environments Network. 2019. URL: https://www.health.govt.nz/publications/national-healthy-food-and-drink-policy [accessed 2020-10-10]

XSL•FO
RenderX

13. Gerritsen S, Kidd B, Rosin M, Shen S, Mackay S, Te Morenga L, et al. 2021 assessment of New Zealand district health boards' institutional healthy food and drink policies: the HealthY Policy Evaluation (HYPE) study. N Z Med J 2022;135(1560):67-76. [doi: 10.26635/6965.5698] [Medline: 35999800]
14. Ni Mhurchu C, Shen S, Kidd B, Jiang Y, Umali E, Gerritsen S, et al. Food and drink availability in New Zealand hospitals 2021/22: alignment with a voluntary National Healthy Food and Drink Policy. 2023 Presented at: ISBNPA 2023 Abstract Booklet. International Society of Behavioral Nutrition and Physical Activity; 2023 Jun 14-17; Uppsala, Sweden.
15. Rosin M, Mackay S, Gerritsen S, Te Morenga L, Terry G, Ni Mhurchu C. Barriers and facilitators to implementation of healthy food and drink policies in public sector workplaces: a systematic literature review. Nutr Rev 2024;82(4):503-535 [FREE Full text] [doi: 10.1093/nutrit/nuad062] [Medline: 37335857]
16. World Health Organization. Action Framework for Developing and Implementing Public Food Procurement and Service Policies for a Healthy Diet. Geneva, Switzerland: World Health Organization; 2021. URL: https://www.who.int/publications/i/item/9789240018341
17. Rosin M, Mackay S, Ni Mhurchu C. Tools and resources used to support implementation of workplace healthy food and drink policies: a scoping review of grey literature. Nutr Diet 2023;80(5):452-462. [doi: 10.1111/1747-0080.12844] [Medline: 37743098]
18. Appendix 3: how to make better pies guidelines (in the National Healthy Food and Drink Policy). The Pie Group. 2019. URL: https://www.health.govt.nz/publications/national-healthy-food-and-drink-policy [accessed 2020-10-10]
19. Healthy events and fundraiser guide. Canterbury District Health Board. URL: https://www.cph.co.nz/your-health/love-kai/ [accessed 2023-05-17]
20. Rosin M, Ni Mhurchu C, Mackay S. Implementing healthy food policies in health sector settings: New Zealand stakeholder perspectives. BMC Nutr 2024;10(1):119. [doi: 10.1186/s40795-024-00924-z] [Medline: 39244614]
21. Shen S, Kidd B, Mackay S, Te Morenga L, Gerritsen S, Jiang Y, et al. Designing an audit tool to evaluate the national healthy food and drink policy: the HYPE study. 2022 Presented at: Nutrition Society of New Zealand Annual Conference; 2021 December 2-3; New Zealand p. 27. [doi: 10.3390/msf2022009027]
22. Merritt K, Zhao S. An innovative reflection based on critically applying UX design principles. Journal Open Innov Technol Market Complexity 2021;7(2):129. [doi: 10.3390/joitmc7020129]
23. Stola K. User experience and design thinking as a global trend in healthcare. J Med Sci 2018;87(1):28-33. [doi: 10.20883/jms.2018.281]
24. Sanders JP, Gokal K, Thomas JJC, Rawstorn JC, Sherar LB, Maddison R, et al. Development of a mobile health snacktivity app to promote physical activity in inactive adults (SnackApp): intervention mapping and user testing study. JMIR Form Res 2023;7:e41114 [FREE Full text] [doi: 10.2196/41114] [Medline: 37213198]
25. Morgan DL. Integrating Qualitative and Quantitative Methods: A Pragmatic Approach. Thousand Oaks, CA: SAGE Publications, Inc; 2014.
26. Creswell JW, Poth CN. Qualitative Inquiry and Research Design: Choosing Among Five Approaches. Thousands Oaks, CA: SAGE; 2018.
27. Kaushik V, Walsh CA. Pragmatism as a research paradigm and its implications for social work research. Soc Sci 2019;8(9):255. [doi: 10.3390/socsci8090255]
28. Canziba E. Hands-on UX Design for Developers: Design, Prototype, and Implement Compelling User Experiences From Scratch. Birmingham, United Kingdom: Packt Publishing; 2018.
29. What are design principles? Interaction Design Foundation. 2016. URL: https://www.interaction-design.org/literature/topics/design-principles [accessed 2023-10-11]
30. The Double Diamond. Design Council. URL: https://www.designcouncil.org.uk [accessed 2023-08-29]
31. Design in the public sector? Design Council. URL: https://tinyurl.com/3x5jmntn [accessed 2023-10-11]
32. Nessler D. Medium. 2018. URL: https://tinyurl.com/aawvyx6x [accessed 2023-10-11]
33. Attribution 4.0 International (CC BY 4.0). Creative Commons. URL: https://creativecommons.org/licenses/by/4.0/ [accessed 2024-12-20]
34. Rosenzweig E. Successful User Experience: Strategies and Roadmaps. Edinburgh, London: Elsevier Science; 2015.
35. Healthy eating advisory service—FoodChecker. State of Victoria. 2016. URL: https://foodchecker.au/ [accessed 2023-07-25]
36. Healthy food finder. New South Wales Ministry of Health. URL: https://www.foodfinder.health.nsw.gov.au/ [accessed 2022-02-14]
37. Province of British Columbia, HealthLinkBC. Brand name food list. URL: https://bnfl.healthlinkbc.ca/ [accessed 2022-02-11]
38. Chapter 10: MoSCoW prioritisation framework (The DSDM Agile Project Framework 2014 onwards). Agile Business Consortium. 2014. URL: https://www.agilebusiness.org/dsdm-project-framework/moscow-prioririsation.html [accessed 2023-08-29]
39. Frey C. Why user flows are pivotal and how to properly outline them. Medium. 2022. URL: https://tinyurl.com/3b238xwz [accessed 2023-10-12]
40. Jhang J. How to create a UX sitemap: a simple guideline. Medium. 2020. URL: https://tinyurl.com/ycx7e4ep [accessed 2023-10-12]

XSL•FO
RenderX

41. Kasturika K, Skrok D, Briggs C. 14 UX deliverables: what will I be making as a UX designer? Interaction Design Foundation. 2020. URL: https://www.interaction-design.org/literature/article/7-ux-deliverables-what-will-i-be-making-as-a-ux-designer [accessed 2023-10-12]
42. Moran K. Usability (user) testing 101. NN/g Nielsen Norman Group. 2019. URL: https://tinyurl.com/2rfr83et [accessed 2023-10-13]
43. Schroeter E. Usability testing: everything you need to know (methods, tools, and examples). Interaction Design Foundation. 2021. URL: https://careerfoundry.com/en/blog/ux-design/usability-testing-guide/#performance-testing [accessed 2023-10-13]
44. Whitenton K. How to test visual design. NN/g Nielsen Norman Group. 2018. URL: https://tinyurl.com/yr8fs8fn [accessed 2023-10-13]
45. Xu M, David JM, Kim SH. The fourth industrial revolution: opportunities and challenges. IJFR 2018;9(2):90-95. [doi: 10.5430/ijfr.v9n2p90]
46. Schoeppe S, Alley S, Rebar AL, Hayman M, Bray NA, van Lippevelde W, et al. Apps to improve diet, physical activity and sedentary behaviour in children and adolescents: a review of quality, features and behaviour change techniques. Int J Behav Nutr Phys Act 2017;14(1):83 [FREE Full text] [doi: 10.1186/s12966-017-0538-3] [Medline: 28646889]
47. Dunford E, Trevena H, Goodsell C, Ng KH, Webster J, Millis A, et al. FoodSwitch: a mobile phone app to enable consumers to make healthier food choices and crowdsourcing of national food composition data. JMIR Mhealth Uhealth 2014;2(3):e37 [FREE Full text] [doi: 10.2196/mhealth.3230] [Medline: 25147135]
48. Mackey MA, Metz M. Ease of reading of mandatory information on Canadian food product labels. Int J Consumer Studies 2009;33(4):369-381. [doi: 10.1111/j.1470-6431.2009.00787.x]
49. Dodson EA, Eyler AA, Chalifour S, Wintrode CG. A review of obesity-themed policy briefs. Am J Prev Med 2012;43(3 Suppl 2):S143-S148 [FREE Full text] [doi: 10.1016/j.amepre.2012.05.021] [Medline: 22898164]
50. Kastner M, Bhattacharyya O, Hayden L, Makarski J, Estey E, Durocher L, et al. Guideline uptake is influenced by six implementability domains for creating and communicating guidelines: a realist review. J Clin Epidemiol 2015;68(5):498-509 [FREE Full text] [doi: 10.1016/j.jclinepi.2014.12.013] [Medline: 25684154]
51. Nutrition Standards for Ka Ora, Ka Ako | Healthy school lunches programme. New Zealand Ministry of Education. 2022. URL: https://kaorakaako.education.govt.nz/file-download/Nutrition%20Standards%202022.pdf [accessed 2023-02-13]
52. Ahmed M, Schermel A, Lee J, Weippert M, Franco-Arellano B, L'Abbé M. Development of the food label information program: a comprehensive Canadian branded food composition database. Front Nutr 2021;8:825050 [FREE Full text] [doi: 10.3389/fnut.2021.825050] [Medline: 35187026]

Abbreviations

DHB: district health board
HYPE: Healthy Policy Evaluation
MoSCoW: Must have, Should have, Could have, Won't have
UX/UI: user experience and user interface

Edited by A Mavragani; submitted 11.05.24; peer-reviewed by L Cranney; comments to author 10.09.24; revised version received 15.09.24; accepted 15.11.24; published 13.01.25.

Please cite as:
Rosin M, Ni Mhurchu C, Umali E, Mackay S
Healthy Kai (Food) Checker Web-Based Tool to Support Healthy Food Policy Implementation: Development and Usability Study
JMIR Form Res 2025;9:e60447
URL: https://formative.jmir.org/2025/1/e60447
doi:10.2196/60447
PMID:

©Magda Rosin, Cliona Ni Mhurchu, Elaine Umali, Sally Mackay. Originally published in JMIR Formative Research (https://formative.jmir.org), 13.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Exploring the General Acceptability and User Experience of a Digital Therapeutic for Cognitive Training in a Singaporean Older Adult Population: Qualitative Study

Siong Peng Kwek[1,2,3], BSocSci; Qiao Ying Leong[1,2], MPH; V Vien Lee[1,2], PhD; Ni Yin Lau[1,2], BSocSci; Smrithi Vijayakumar[1,2], PhD; Wei Ying Ng[1,2], BSocSci; Bina Rai[2,3], PhD; Marlena Natalia Raczkowska[1,2], PhD; Christopher L Asplund[1,2,3,4,5*], PhD; Alexandria Remus[1,2,3,6*], PhD; Dean Ho[1,2,3,7,8,9*], PhD

[1]The Institute for Digital Medicine (WisDM), Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore
[2]The N.1 Institute for Health, National University of Singapore, Singapore, Singapore
[3]Department of Biomedical Engineering, College of Design and Engineering, National University of Singapore, Singapore, Singapore
[4]Division of Social Sciences, Yale-NUS College, National University of Singapore, Singapore, Singapore
[5]Department of Psychology, College of Humanities and Sciences, National University of Singapore, Singapore, Singapore
[6]Yong Loo Lin School of Medicine, Heat Resilience & Performance Centre, National University of Singapore, Singapore, Singapore
[7]Department of Pharmacology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore
[8]The Bia-Echo Asia Centre for Reproductive Longevity and Equality, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore
[9]Singapore's Health District @ Queenstown, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore
[*]these authors contributed equally

Corresponding Author:
Alexandria Remus, PhD
Yong Loo Lin School of Medicine
Heat Resilience & Performance Centre
National University of Singapore
27 Medical Drive, #03-01
Singapore, 117510
Singapore
Phone: 65 66017766
Email: alexandria.remus@nus.edu.sg

Abstract

Background: Singapore's large aging population poses significant challenges for the health care system in managing cognitive decline, underscoring the importance of identifying and implementing effective interventions. Cognitive training delivered remotely as a digital therapeutic (DTx) may serve as a scalable and accessible approach to overcoming these challenges. While previous studies indicate the potential of cognitive training as a promising solution for managing cognitive decline, understanding the attitudes and experiences of older adults toward using such DTx platforms remains relatively unexplored.

Objective: This study aimed to characterize the general acceptability and user experience of CURATE.DTx, a multitasking-based DTx platform that challenges the cognitive domains of attention, problem-solving, and executive function in the Singaporean older adult population.

Methods: A total of 15 older adult participants (mean age 66.1, SD 3.5 years) were recruited for a 90-minute in-person session. This session included a 30-minute playtest of CURATE.DTx, followed by a 60-minute semistructured interview to understand their overall attitudes, experience, motivation, and views of the intervention. Interviews were audio-recorded and transcribed verbatim, then analyzed using an inductive approach. Thematic analysis was used to identify emerging patterns and insights.

Results: A total of 3 main themes, and their respective subthemes, emerged from the interviews: comprehension, with subthemes of instruction and task comprehension; acceptability, with subthemes of tablet usability, engagement and enjoyment, and attitude and perceived benefits; and facilitators to adoption, with subthemes of framing and aesthetics, motivation recommendations and the role of medical professionals. Our findings revealed that participants encountered some challenges with understanding certain elements of CURATE.DTx. Nevertheless, they were still highly engaged with it, finding the challenge to be enjoyable. Participants

XSL•FO
RenderX

also showed a strong awareness of the importance of cognitive training and expressed a keen interest in using CURATE.DTx for this purpose, especially if recommended by medical professionals.

Conclusions: Given the positive engagement and feedback obtained from Singaporean older adults on CURATE.DTx, this study can serve as a basis for future platform iterations and strategies that should be considered during implementation. Future studies should continue implementing an iterative codesign approach to ensure the broader applicability and effectiveness of interventions tailored to this demographic.

(JMIR Form Res 2025;9:e63568) doi:10.2196/63568

KEYWORDS

older adults; cognitive training; digital therapeutic; DTx; remote; usability; acceptance; interviews; gerontology; geriatric; elderly; experiences; attitudes; opinions; perceptions; perspectives; interview; cognition; digital health; qualitative; thematic

Introduction

As individuals age, cognitive abilities mature at different time points and eventually decline [1]. This age-related cognitive decline commonly affects memory, learning ability, fluid intelligence, and multitasking skills [2,3]. In Singapore, aggregate declines may be particularly significant given its rapidly aging population, whereby nearly 1 in 4 Singaporeans will be over the age of 65 by 2030 [4]. This decline in cognitive abilities not only poses challenges for the workforce but also imposes a heavy burden on the health care system [5]. Given these factors, exploring effective methods to manage the decline of cognitive capabilities within this population would be essential to overcome these challenges.

Research has shown that cognitive training can improve cognitive functioning in older adults and even potentially delay the onset or progression of neurodegenerative disorders such as Alzheimer disease [6,7]. In Singapore, where there is high internet and mobile device penetration [8], cognitive training can be administered as a digital therapeutic (DTx) at scale. DTx are evidence-based software designed to prevent, manage, or treat health conditions [9]. They can be delivered remotely and often incorporate rewards and gamification elements to enhance adherence [10].

We have developed CURATE.DTx, a DTx platform that delivers personalized digital cognitive training by a tablet-ready adaptation of the Multi-Attribute Task Battery (MATB) platform, a flight deck simulator initially developed by the National Aeronautics and Space Administration (NASA) [11] and later further adapted by the United States Air Force (USAF) [12,13]. The MATB platform requires users to simultaneously manage 4 different tasks, which include responding to indicator lights and changing scales, tracking a target using a joystick, following verbal instructions to tune to specific radio frequencies and channels, and managing fuel tank levels. It is a naturalistic multitasking paradigm that has commonly been used as a mental workload measure for different task intensities [14,15] and has strong face validity as, by design, it involves attention, problem-solving, and coordination amongst tasks. Using the platform challenges the executive function by demanding sustained vigilance and goal-directed action for managing multiple tasks while also training attention and problem-solving for each specific task [16]. MATB has also been shown to evoke local and network-level brain responses, characteristic of multitasking [17], and such training may enhance cognitive efficiency and coordination [18].

CURATE.DTx also harnesses CURATE.AI, an adaptive approach that has previously been validated for drug dose optimization in clinical conditions, to personalize a user's digital cognitive training by correlating difficulty level doses (inputs) to performance outcomes (outputs) using only their own data [19,20]. While this was initially developed for cognitive training in healthy adult and postradiotherapy patient populations, the same application may serve as a potential solution for managing cognitive decline in older adults [19-22].

At its core, CURATE.DTx was designed to retain the primary tasks and functionality of MATB. However, the interface was modernized with gamification, and usability elements were also incorporated into the platform to improve user engagement and overall experience, such as the detailed instructions screens explaining the MATB tasks and a tablet form factor with a touchscreen for intuitive control. The inclusion of these usability features could also improve the accessibility of CURATE.DTx for older adults. Indeed, the widespread adoption of tablets as the medium for game-based interventions for older adults underscores their versatility for this population [23]. In addition, while the use of DTx in older adults is a relatively recent development, initial studies have shown promising outcomes [24,25]. As older adults often have impaired cognitive performance compared to young adults, especially in multitasking-related tasks under high task loads [26], it would be useful for CURATE.DTx to be adapted for this population. Finally, older adults tend to show greater levels of disengagement than young adults when faced with higher task difficulty [27]. Such disengagement could also impact their enjoyment, which is a predictor of motivation when engaging with an intervention [28]. Considering these factors, it is important to understand older adults' attitudes and experiences with CURATE.DTx, as their feedback is also helpful for improving future iterations of the platform. Therefore, this study aimed to examine the general acceptability and user experience of the CURATE.DTx platform among older Singaporean adults.

Methods

Overview

The Consolidated Criteria for Reporting Qualitative Studies (COREQ) guidelines were followed for the reporting of this

XSL•FO
RenderX

study [29], with the COREQ checklist presented in Multimedia Appendix 1.

CURATE.DTx Platform

The CURATE.DTx platform is comprised of 4 tasks: Systems Monitoring, Communications, Tracking, and Resource Management, as depicted in Figure 1. Systems Monitoring requires the user to monitor the lights and gauges and respond as quickly as possible when they illuminate or change color. Communications requires the user to listen for the call sign "NGT504" and set the radio station and frequency according to the audio instructions, ignoring instructions for other call signs. Tracking requires the user to keep the ball as close to the center of the cross-hair as possible, dragging the ball back when it deviates away. Finally, Resource Management requires the user to maintain the two fuel tank levels within the score area by opening and closing the arrow valves, making appropriate adjustments when the valves occasionally turn off or fail. All 4 tasks can occur at the same time, and the user has to manage them simultaneously until the session ends.

Figure 1. Screenshot of the CURATE.DTx's user interface. Top left: Systems Monitoring—respond to indicator lights and gauges. Top right: Communications—follow verbal instructions to tune to specific radio frequencies and channels. Bottom left: Tracking—keep the target as close to the center of the crosshair as possible. Bottom right: Resource Management—maintain the fuel tank levels.

Recruitment

This study was approved by the National University of Singapore's institutional review board (reference NUS-IRB-2021-95). Participants were recruited through purposive sampling by age group from a community senior activity center in Singapore. The activity center coordinator identified potential participants, provided an overview of the study, and arranged for informed consent and study sessions. Participants were also told about the research team's involvement in the study and the reasons for doing the research. The research team (SPK, VVL, SV, and AR) consisted of full-time researchers, 3 female postdoctoral fellows, and 1 male research assistant, and had previous training in qualitative data collection and analyses. They screened potential participants against the following inclusion criteria: (1) Singaporean older adults aged 60 years old and older and (2) fluent in English. Participants with significant hearing, visual, or cognitive impairment that would affect their ability to complete the study were excluded. No relationships with the participants were established before the interview sessions. Recruitment took place over a period of 4 months and ended when the target sample of 15 was obtained. No participants refused to participate or dropped out after consenting to the study.

Data Collection

This study used semistructured interview sessions to understand the participants' experiences and attitudes toward CURATE.DTx. Each session was approximately 90 minutes and consisted of a 30-minute playtest where participants play-tested CURATE.DTx on a Samsung Galaxy Tab A7 (SM-T500) 10.4-inch tablet, followed by a 60-minute semistructured interview on their thoughts and experience. To reduce the risk of bias and enhance the credibility and robustness of the results, methods triangulation was used [30]. We compared the participants' actions observed during their playtest with their verbal qualitative feedback provided in the subsequent interview to ensure consistency of findings across different mediums. All sessions were in-person at the senior activity center's premises and were conducted in English with at least 2 members of the research team and no nonparticipants present.

Before the session commenced, the research team provided an overview of the study, and informed consent was obtained. All participants consented to the audio recording of the session for transcription purposes, while the participants' demographic background details were also collected. While performance score data from the sessions were automatically collected, they were not accessed or analyzed as they did not relate to the purpose of this study. Open-ended questions were asked during the interview based on a semistructured interview guide, which was developed to include guiding topics that broadly cover the factors of the Technology Acceptance Model (TAM) framework [31]. The guiding topics used for the interview discussion are presented in Table 1. While the research team's past familiarity with CURATE.DTx could have influenced them to engage with

XSL•FO
RenderX

participants about it in a more positive manner, these guiding topics ensured that both positive and negative feedback were fully captured. Short field notes were also taken where necessary to ensure understanding and clarity of participants' responses. At the end of the session, all participants were reimbursed for their time. All data collected, including signed consent forms, interview recordings, and demographic data, were stored in a secure database and deidentified for the purpose of this study.

Table 1. Semistructured interview topic guide.

Category	Guiding topics for discussion
Interface-specific	• Discuss the purpose of the tasks. • Discuss understanding of the tasks. • Follow up on slips and errors from the playtest session. • Discuss the ease of use and difficulties with using the interface.
Program-specific	• Discuss the expectations of joining this program. • Discuss factors that would increase motivation toward adoption. • Discuss barriers and concerns of adoption. • Discuss alternatives to this program.

Data Analysis

The interview recordings were transcribed verbatim, and inductive thematic analysis was used to identify patterns and emerging or recurring themes [32]. Initial coding was conducted on Microsoft Word. Participants' responses were descriptively labeled and summarized into codes independently by the 4 members of the research team. Any discrepancies in the codes were resolved through discussion. Codes were then organized into potential themes using Microsoft Excel and further reviewed and refined in relation to the entire data set. The final codes and broader themes were generated from multiple discussions and iterations by the research analysis team (SPK, VVL, SV, and AR). Data saturation was achieved when there were no new themes identified and confirmed through discussion within the research team. There were no repeat interviews, transcripts were not returned to participants for comments, and participants did not provide feedback on the findings.

Ethical Considerations

This study was approved by the National University of Singapore's institutional review board (NUS-IRB reference: NUS-IRB-2021-95). All participants provided informed consent and were informed of their right to withdraw from the study at any time. All data collected were deidentified for the purpose of this study. At the end of their sessions, participants were reimbursed SGD 45 (US $34) for their time and participation.

Results

Participant Characteristics

A total of 15 participants completed the study (10 males, 5 females; mean age 66.1, SD 3.5, range 62-73 years). Participants were from a range of education level backgrounds, with 5 O Levels (10+ years), 6 Diploma (13+ years), and 4 University (16+ years). All participants reported regular use of technology, such as mobile phones, computers, televisions, and tablets. Full participant demographics and characteristics are presented in Table 2.

XSL•FO
RenderX

Table 2. Participant demographics and characteristics.

Participant	Age (years)	Gender	Highest education level[a]	Current technology usage
1	63	Male	Diploma	Mobile phone, computer, television
2	67	Male	Diploma	Mobile phone, computer, television
3	67	Male	University	Mobile phone
4	68	Male	O Levels	Mobile phone
5	68	Male	Diploma	Mobile phone, computer, television, tablet
6	67	Male	O Levels	Mobile phone
7	72	Male	Diploma	Mobile phone
8	62	Female	University	Mobile phone
9	66	Male	University	Mobile phone, television, tablet
10	63	Female	Diploma	Mobile phone, television
11	64	Female	University	Mobile phone
12	69	Male	O Levels	Mobile phone, computer, tablet
13	73	Male	O Levels	Mobile phone, computer
14	73	Female	O Levels	Mobile phone
15	64	Female	Diploma	Mobile phone, television, tablet

[a]The years of education based on the Singapore education system is as follows: O Levels=10+ years, Diploma=13+ years, University=16+ years.

Interview Data

From the participants' responses, 3 main themes emerged—comprehension, acceptability, and facilitators to adoption. Comprehension describes the participants' understanding of the CURATE.DTx platform, acceptability covers their thoughts and attitudes towards the CURATE.DTx platform and the tablet hardware, and facilitators to adoption encompass feedback on several recommendations discussed with the participants.

Theme 1: Comprehension

Under the theme of comprehension, 2 subthemes emerged from participants' responses, namely instruction comprehension, which is the feedback specific to the "How To Play" instructions page on the web-based platform, and task comprehension, their feedback on the individual MATB tasks and overall gameplay.

Instruction Comprehension

Participants initially struggled to understand CURATE.DTx instructions, mainly regarding 2 aspects: the complexity of the instructions and a misconception regarding simultaneous task management.

Participants expressed that the provided instructions were long-winded and difficult to understand, often perceiving them as complicated. Participant 11 mentioned the length of the instructions that, "Firstly, the instructions are very long." Participant 13 further emphasized their initial struggle, stating, "It is quite complicated actually when you first look at it, I actually struggled initially to try and grasp what this whole thing is about."

In addition, participants misunderstood the instructions as they did not realize that all tasks had to be played simultaneously. Although it was explicitly mentioned early in the instructions, the sequential presentation of task instructions led some participants to expect CURATE.DTx is to be conducted in a single-task manner. Participant 5 explained this misunderstanding, "The first instruction says four tasks are running at the same time, and it shows during the explanation the four tasks will be coming out, but it is explaining one task at a time, [so I thought] okay so this [CURATE.DTx] is one task at a time."

Task Comprehension

Participants conveyed that it was challenging to understand the more complicated tasks from just the instructions, and that it took some time using the platform to get used to them. As Participant 11 expressed, "I never read instructions fully. I skim. So it's only as I play [that] I more or less get the gist of it, I know what's happening." However, participants also voiced concerns it might potentially be too complicated and challenging for some in their age group. Participant 9 cautioned, "I guess the elderly, if you give them too many instructions and too many things for multitasking, they might just switch off their mind, they find it a bit difficult for them."

Participants mentioned that using CURATE.DTx itself helped them understand how the tasks worked. They relied on the in-game task feedback to figure out if they were doing it correctly. Participant 9 remarked, "Initially, I just don't understand what this is, this whole thing. But after a while, oh yeah, I got the thing." Participant 5 emphasized the importance of effective task feedback, stating, "In terms of my understanding of the game, the task will give a person better feedback on what to [do], how to react."

However, participants still expressed a desire for additional assistance and guidance when they faced difficulties in

XSL•FO
RenderX

understanding CURATE.DTx. Participant 4 mentioned, “I think like when I’m jammed at a particular part, I ask you all, you all must help me then I [can clearly] understand the game.”

Theme 2: Acceptability

Under the theme of acceptability, 3 subthemes emerged from participants’ responses. The subthemes included table usability, engagement, and enjoyment of CURATE.DTx, and their attitude and perceived benefits of the intervention.

Tablet Usability

Participants mentioned that the usability of the tablet, namely its size and weight, were important factors to consider for the acceptability of CURATE.DTx in this population. While most were satisfied with the physical attributes of the provided tablet, some expressed concerns about other usability aspects of the tablet. Participant 12 shared that, “I think it’s okay. Too big, my age group will have a problem holding. So, lightweight [is better].” However, Participant 10 mentioned the potential issue of screen reflectivity, saying, “Just now the light was glaring, so I have to hold it up so that I can see better.”

Engagement and Enjoyment

Participants exhibited high engagement with CURATE.DTx, maintaining a strong level of focus throughout the entire playtest session. Participant 3 remarked afterward, “Like it keeps me on my toes. You didn’t see? If you look at this you will see me very attentive, listening and all this.”

Regarding their enjoyment of various aspects of the DTx, participants acknowledged that certain tasks were more enjoyable than others, with a general preference for the more challenging tasks over the easier ones. Participant 10 stated, “It’s nice to play... Challenging is nice.” Participant 12 had a similar opinion about a simpler task, stating, “This one [Systems Monitoring] is actually very simple, too simple already. Just pressing a button. No fun at all!”

Furthermore, some participants appreciated the presence of a mix of easier and more difficult tasks, as it added variety and enhanced their enjoyment. Participant 13 highlighted this aspect, “In a way, the four [tasks], there is variety... Not all like this [Systems Monitoring], this [is] kind of standard, then I think it's not so enjoyable.”

Attitude and Perceived Benefits

Participants expressed that they found CURATE.DTx as an intervention could be potentially beneficial as it was tailored to their prospective cognitive needs, with many considering cognitive training to be important for maintaining their mental capacity in their senior years. Participant 5 emphasized the significance of cognitive stimulation, “For me, my purpose is to stimulate the thinking, the brain, or else your brain deteriorates over time.” Participants viewed CURATE.DTx as a tool for cognitive stimulation and expressed strong intrinsic motivation and interest in using it for cognitive training to challenge themselves. Participant 15 remarked, “I might play this one, for training. Better than the sudoku.”

Theme 3: Facilitators to Adoption

Under the theme of facilitators to adoption, 3 subthemes emerged from participants’ responses. The subthemes include framing and aesthetics recommendations, motivation recommendations, and the role of medical professionals.

Framing and Aesthetics Recommendations

When discussing CURATE.DTx’s visual design, participants suggested that it would be more attractive to older adults if the tasks were framed as real-world, daily life analogies, as they found the current aviation-related tasks to be less relatable. Participant 11 illustrated, “I look at this one, you’ve got petrol, that one is the meter, maybe you can see this car dashboard? And then this one is your air-con slider. And from there at least I can associate, the person also can associate.”

Participants also expressed that the general visual appeal of CURATE.DTx could be improved. They found the color palette to be dull and suggested incorporating more visually appealing elements. Participant 8 elaborated, “If you say that you're going to do it in this color, it is very boring. Because for me, I am a very bright color person, blue, blue, blue? I don’t know.”

Motivation Recommendations

When discussing motivation, most participants showed strong intrinsic motivation towards using CURATE.DTx for cognitive training. However, participants also expressed approval for the addition of extrinsic motivation factors, such as the inclusion of a scoring system to CURATE.DTx. They believed that a scoring system would provide them with feedback on their performance and motivate them to strive for improvement. Participant 12 illustrated this viewpoint, saying, “For example, today I score 80, I want to aim for 85. 85, got to aim for 90, 90, 95! People will continue to chase. So if you don't have this, it's [just] another game.”

Some participants thought that CURATE.DTx could be boring over time due to the lack of variety of tasks and levels. As Participant 3 described it, “[it] can be quite boring, to play this everyday. Like there is no changes. Like you see the other games, you see the clearing the blocks game, that one every time change level. That one is different.” They also mentioned that CURATE.DTx lacked some common and attractive features compared to other more attractive existing games in the market. Participant 12 noted, “Most of the time, I play with the music. You see, [the music] attract[s] you, you know?”

The potential gameplay duration of each CURATE.DTx session also emerged as a factor that affected participants’ motivation to use the platform. Overall, most participants were satisfied with the playtest session’s duration, with Participant 12 considering it “not too long, not too short” and Participant 9 saying that, “I think it’s just right.” Participants also emphasized the importance of CURATE.DTx in fitting into their daily schedule. Participant 6 mentioned, “My time is very tight... If I got the time, I don’t mind.” Participant 10 echoed this opinion, saying, “Yes, I would if I have time.”

Role of Medical Professionals

When discussing the role of doctors and medical professionals in interventions, most participants expressed a strong belief in

XSL•FO
RenderX

the recommendations of their medical personnel, emphasizing that they would heed their advice. Participant 10 succinctly conveyed this sentiment, stating, “Anyone with medical [knowledge] who is looking after me medically, yes, I will take their advice seriously.”

Furthermore, several participants were hopeful that the intervention could serve as a potential alternative to medication. Participant 3 expressed this sentiment, stating, “If this game can help, then why not? Instead of medication.” Since the intervention was in the form of a game, participants generally did not expect any adverse effects. Participant 8 also commented on the absence of adverse effects typically associated with games, stating, “Generally games, the negative effects would be addiction, but I think this one doesn't have.”

Furthermore, participants demonstrated a higher willingness to adopt the platform if it was scientifically proven to be beneficial for their health, even if they wouldn't have considered using it otherwise. Participant 3 elaborated on this point, explaining, “If you say by playing this twice a week, every day, for half an hour, you will improve your mental state of health, then I will definitely play since it is scientifically proven.”

Discussion

Principal Findings

This study aimed to understand the acceptability and user experience of older Singaporean adults using CURATE.DTx, a digital cognitive training DTx platform. Overall, we found that participants faced some difficulties with understanding the CURATE.DTx platform. However, engagement with the platform was generally still strong, and participants were also enthusiastic about using it for cognitive training, especially if advised by medical professionals. In light of these findings, there is a need to address several key factors for the next iteration of the CURATE.DTx platform to ensure greater adoption and sustained usage among this older adult demographic group, with these considerations broadly applicable to the development of other similar DTx platforms in this space as well.

Comprehension Considerations

We identified the importance of considering the comprehension of the instructions on overall experience when developing a cognitive training DTx. Aging has been shown to be associated with declining cognitive abilities that could affect information processing, such as in visual attention [33] and working memory [34]. Specifically, older adults may have a harder time processing unfamiliar visual information and retaining it in memory [35]. We observed this in some of our participants, who struggled to fully understand the game instructions provided on CURATE.DTx and perceived them to be too long-winded and complicated. Some of this difficulty can be attributed to the complexity of several task explanations, as well as the continuous sequential presentation of the instructions. Therefore, older adults would benefit from having complex concepts broken down into smaller portions so that each step is within their capacity for processing, and this information should be summarized again for each task to ensure that it is adequately rehearsed and retained in their memory [36]. Providing accessible bite-sized chunks of information that are reiterated at various points of the instructions will ensure that they are clear and concise, which would prevent misunderstanding and improve their overall experience when using CURATE.DTx. In addition, participants may also benefit from the implementation of instructions explained as an interactive tutorial with the opportunity to practice all 4 tasks simultaneously at the end. This would provide them with the opportunity to familiarize themselves with the tasks and gain a better understanding of the mechanisms involved.

Our findings also showed that providing in-game feedback can be a useful tool for helping older adult users understand the tasks. In line with Nielsen’s usability heuristics, which emphasizes recognition rather than recall [37], incorporating visual elements that provide small reminders of information can aid users in navigating the task without requiring them to remember specific task instructions. Our study participants themselves mentioned a preference for learning CURATE.DTx mechanics through gameplay, using in-game feedback to gauge their performance and adjust their strategies. This is in line with the literature, where feedback in educational games can provide players with schemas to help them learn from their misunderstandings [38]. Indeed, providing regular feedback about their performance enables users to engage in more critical learning [39], and clear and specific feedback—such as indicating correct and incorrect actions, highlighting areas for improvement, and offering suggestions or tips—can direct them to develop more effective game strategies. Furthermore, it is also important to present feedback in a way that is easily understandable to older adults, taking into account their cognitive abilities and visual acuity. Hence, designing a user-friendly interface with intuitive controls, clear visual cues, and accessible feedback would enhance their interaction and learning with CURATE.DTx, thus improving their overall experience.

Acceptability Considerations

Our findings showed that there was strong interest in using CURATE.DTx for cognitive training. Participants recognized the importance of cognitive training for maintaining cognitive functioning and believed that the CURATE.DTx platform would be useful and beneficial for this purpose. These stated views align with previous research showing that older adults in Singapore are open to adopting technology to achieve their health goals [40]. Furthermore, the intended and perceived benefits of the training platform emerged as the most common facilitators of its use. This finding aligns with the principles of the TAM framework, which suggests that the perceived usefulness of a technological tool strongly influences its acceptance as a learning and training resource [41]. Given the existing intrinsic motivation and expressed interest in CURATE.DTx, ensuring sustained engagement, would thus be key to its long-term adoption in older adults.

Engagement and enjoyment were also found to affect the acceptability of the CURATE.DTx platform. Specifically, participants expressed a preference for more challenging tasks, like Resource Management, which requires them to develop an

XSL•FO
RenderX

optimal strategy for balancing resources over simpler tasks, like Systems Monitoring, which merely requires them to click on lights when they illuminate. As facing challenges in a game has previously been reported to directly enhance learning and increase engagement with the game [42], incorporating more challenging tasks can offer participants a sense of accomplishment and keep them motivated. However, we also found that some tasks were reported to be overly complex and difficult to understand for this age group. Thus, striking a "sweet spot" trade-off between challenge and ease of understanding can lead to flow—the sense of engagement and immersion in the interactive experience [43]. While the use of an adaptive training approach such as CURATE.AI can partially aid in this by adjusting the difficulty to an optimal level based on personal performance [19], improving the design of the task elements based on participant feedback would help to ensure a more cohesively enjoyable experience.

The source of the challenge is also an important consideration. While there can be a motivating in-game frustration that causes the player to want to persevere and overcome the challenge, lack of comprehension of CURATE.DTx's tasks could instead lead to disheartening at-game frustration that causes them to disengage [44,45]. As these seemingly contradictory aspects are rated most highly in terms of importance for older adults [46], successfully finding this balance can make the platform highly appealing to this demographic group.

Finally, participants generally viewed using tablets as the physical medium for CURATE.DTx positively. However, the size and weight of the tablets were highlighted as important factors to consider, underscoring the need for a device that is easy to manage and handle. Despite some concerns about screen reflectivity, the overall design of the tablet was well-received. This is in line with existing literature reporting high satisfaction among older individuals using tablets in clinical settings [47] as well as their perceived benefit not just for cognitive training but also for enjoyment [48]. Given that the effective deployment of a tablet-based DTx in a remote setting use case requires addressing additional factors such as ensuring household internet access, it is important to note that in Singapore, where 99% of resident households, including 93% of senior-only households, have internet access [49], this is less of a concern compared with the digital divide at the global level [50], with disparities in internet access often linked poverty and unemployment [51]. Therefore, understanding the specific needs and circumstances of the target group is crucial for the successful adoption of the tablet-based platform.

Facilitators to Adoption Considerations

Our findings revealed that enhancing the framing and aesthetics of the current CURATE.DTx platform would make it more appealing and engaging to the older Singaporean population. Specifically, incorporating visually appealing elements and features found in popular games in the current market, such as vibrant colors and attractive visuals, can help to enhance enjoyment. Adhering to design principles, such as judicious color use with no more than 2 or 3 fully saturated intense colors, can assist in accomplishing this [52]. Participants also expressed a preference for tasks framed to be more relatable and familiar to their daily lives, such as a car dashboard, rather than the current MATB flight deck simulator interface incorporated into CURATE.DTx. Again, DTx developers should consider Nielsen's usability heuristics, specifically the principle of matching the system to the real world [37], to make the platform's task more intuitive. Using familiar symbols, illustrations, or other analogies that resonate with older adults' day-to-day lived experiences will ensure that they are able to understand the task requirements more easily. Such aesthetic changes can help to make the platform more appealing to older adults and increase their engagement with the platform [53].

Our findings also highlighted the importance of motivation in the DTx platform design. While participants already demonstrated intrinsic motivation in using the platform, they also welcomed extrinsic motivational factors to further strengthen their interest. Participants were particularly taken with the idea of gamification elements such as a scoring system, reporting that knowledge of their scores would drive them to improve. Enhancing scoring with a leaderboard may be an option to consider in a senior-friendly iteration of CURATE.DTx, as they have previously been shown to motivate players to strive for higher-difficulty goals [54]. Nevertheless, it is important to ensure that extrinsic motivation factors do not undermine participants' existing intrinsic motivation to use the CURATE.DTx platform but rather enhance it [55].

Participants also supported the introduction of a greater variety of tasks for CURATE.DTx. This can be a means to prevent boredom and also promote meta-learning, a concept of applying knowledge from past experiences to new, related tasks [56]. While the concept is mainly applied in machine learning and artificial intelligence, recent research has proposed that this idea of drawing from past learned experience can be applied in cognitive training, in particular with cognitive control [57]. Promoting such learning could be beneficial to the older adults' mental health while also increasing the appeal of the CURATE.DTx platform and improving its long-term engagement and adoption.

Finally, our findings revealed the role of medical professionals to be pivotal in the adoption process. In Singapore, health care providers are highly trusted sources for learning about health platforms [58,59], thus playing a critical role in influencing their patient's decision to adopt new technologies. Participants also highlighted the importance of scientific evidence demonstrating the benefits of the DTx, likely reflecting this population's trust in research [60]. They also viewed the DTx as a potential alternative to medication, indicating their willingness to explore nonpharmacological approaches to cognitive maintenance. Thus, collaborating with health care professionals and conducting further research studies to demonstrate the efficacy and cognitive benefits of the game could significantly increase trust and promote broader adoption among older Singaporean adults.

Limitations

Several limitations should be considered when interpreting the findings of this study. First, the study was restricted to English-speaking older adults, which may not fully capture the experiences and perspectives of older adults from

XSL•FO
RenderX

non–English-speaking backgrounds. Singapore is a multilingual country, and considering the diverse linguistic landscape, it is important to extend the research to include participants who speak languages other than English. This would provide a more comprehensive understanding of older adults' acceptability and experiences with the cognitive training digital therapeutics platform.

In addition, it is worth highlighting that this study is focused solely on the particular DTx cognitive training platform used in this research. It's important to acknowledge that this platform was not originally tailored for older adult users and that moving forward, it is recommended that future iterations be codesigned with the target demographic to develop a more bespoke solution. Future research should explore a wider range of cognitive training interventions and digital therapeutics platforms to gain a more comprehensive understanding of older adults' acceptability and experiences across different interventions.

Conclusion

Overall, this study investigated the general acceptability and user experience of a cognitive training DTx in an older adult Singapore population. Despite facing some difficulties in instructional and task comprehension, participants were still highly engaged with the platform and found the challenge provided by the DTx to be enjoyable. Participants also expressed interest and recognized the importance of cognitive training for their population. We also found that certain aesthetics and motivational features and the incorporation of medical professionals into the onboarding process would facilitate greater adoption and long-term usage in this population. These findings highlight several important considerations that can serve as a useful reference to inform the design and implementation strategies of the future iteration of CURATE.DTx as well as other cognitive training interventions for this population.

Acknowledgments

We gratefully acknowledge funding from the Singapore Cancer Society (grant SCS-GRA-2019-00063) for funding this study and had no influence on any part of it. DH gratefully acknowledges funding from the National Research Foundation Singapore under its AI Singapore Programme (award AISG-GC-2019-002) and the Singapore Ministry of Health's National Medical Research Council under its Open Fund- Large Collaborative Grant ('OF-LCG'; grant MOH-OFLCG18May-0028), Institute for Digital Medicine (WisDM) Translational Research Programme (grant R-719-000-037-733) at the Yong Loo Lin School of Medicine, National University of Singapore, Ministry of Education Tier 1 FRC Grant (grant R-397-000-333-114) and the Next-Generation Brain-Computer-Brain Platform—A Holistic Solution for the Restoration & Enhancement of Brain Functions (NOURISH) project from the RIE2020 ADVANCED MANUFACTURING AND ENGINEERING (AME) PROGRAMMATIC FUND (grant A20G8b0102/A-0002199-02-00). All funders have no influence on the study design, collection, management, analysis, interpretation of data, writing of the report, and the decision to submit the report for publication. We would also like to thank Jason Labbe for his assistance in the development of the CURATE.DTx platform and C&B for enabling a collaborative environment, and Lion Befrienders and Joshann Lee for their kind assistance during the data collection process.

Conflicts of Interest

CLA and DH are coinventors of intellectual property pertaining to personalized digital therapeutics and CURATE.AI. DH is a shareholder of KYAN Therapeutics, which has licensed intellectual property pertaining to AI-based oncology drug development. He has also filed a provisional patent on data-driven behavior change.

Multimedia Appendix 1
Consolidated Criteria for Reporting Qualitative Studies (COREQ) checklist.
[PDF File (Adobe PDF File), 416 KB - formative_v9i1e63568_app1.pdf]

References

1. Hartshorne JK, Germine LT. When does cognitive functioning peak? The asynchronous rise and fall of different cognitive abilities across the life span. Psychol Sci 2015;26(4):433-443 [FREE Full text] [doi: 10.1177/0956797614567339] [Medline: 25770099]
2. Bettio LEB, Rajendran L, Gil-Mohapel J. The effects of aging in the hippocampus and cognitive decline. Neurosci Biobehav Rev 2017;79:66-86. [doi: 10.1016/j.neubiorev.2017.04.030] [Medline: 28476525]
3. Kievit RA, Davis SW, Mitchell DJ, Taylor JR, Duncan J, Cam-CAN Research Team, Cam-CAN Research Team. Distinct aspects of frontal lobe structure mediate age-related differences in fluid intelligence and multitasking. Nat Commun 2014;5(1):5658 [FREE Full text] [doi: 10.1038/ncomms6658] [Medline: 25519467]
4. PMO | PM lee hsien loong at the 'Singapore ageing: issues and challenges ahead' Book Launch. Prime Minister's Office Singapore. Prime Minister's Office Singapore. URL: https://www.pmo.gov.sg/Newsroom/PM-Lee-Hsien-Loong-at-the-Singapore-Ageing-Issues-and-Challenges-Ahead-Book-Launch [accessed 2023-04-11]
5. Tan CC, Lam CSP, Matchar DB, Zee YK, Wong JEL. Singapore's health-care system: key features, challenges, and shifts. Lancet 2021;398(10305):1091-1104. [doi: 10.1016/S0140-6736(21)00252-X] [Medline: 34481560]

XSL•FO
RenderX

6. Rahe J, Petrelli A, Kaesberg S, Fink G, Kessler J, Kalbe E. Effects of cognitive training with additional physical activity compared to pure cognitive training in healthy older adults. CIA 2015;10:297. [doi: 10.2147/cia.s74071]
7. Brookmeyer R, Johnson E, Ziegler-Graham K, Arrighi HM. Forecasting the global burden of alzheimer's disease. Alzheimers Dement 2007;3(3):186-191. [doi: 10.1016/j.jalz.2007.04.381] [Medline: 19595937]
8. World Bank. Individuals using the Internet (% of population) - Singapore. World Bank Open Data. URL: https://data.worldbank.org [accessed 2023-11-16]
9. Digital Therapeutics Alliance.. What is a DTx? Digital Therapeutics Alliance. URL: https://dtxalliance.org/understanding-dtx/what-is-a-dtx/ [accessed 2023-11-16]
10. Mahncke HW, Kim SJ, Rose A, Stasio C, Buckley P, Caroff S, et al. Evaluation of a plasticity-based cognitive training program in schizophrenia: results from the eCaesar trial. Schizophr Res 2019;208:182-189 [FREE Full text] [doi: 10.1016/j.schres.2019.03.006] [Medline: 30930034]
11. Comstock JR, Arnegard RJ. The Multi-Attribute Task Battery for Human Operator Workload and Strategic Behavior Research. 1992. URL: https://ntrs.nasa.gov/citations/19920007912 [accessed 1992-01-01]
12. Miller W. Air Force Research Laboratory. The US air force-developed adaptation of the multi-attribute task battery for the assessment of human operator workload and strategic behavior. 2010. URL: https://apps.dtic.mil/sti/citations/ADA537547 [accessed 2010-07-01]
13. Miller W, Schmidt K, Estepp J, Bowers M, Davis I. DTIC Document. An updated version of the US air force multi-attribute task battery (AF-MATB). URL: https://apps.dtic.mil/sti/citations/ADA611870 [accessed 2014-08-01]
14. Liu S, Nam CS. Quantitative modeling of user performance in multitasking environments. Computers in Human Behavior 2018;84:130-140. [doi: 10.1016/j.chb.2018.02.035]
15. Pontiggia A, Gomez-Merino D, Quiquempoix M, Beauchamps V, Boffet A, Fabries P, et al. MATB for assessing different mental workload levels. Front Physiol 2024;15:1408242 [FREE Full text] [doi: 10.3389/fphys.2024.1408242] [Medline: 39108543]
16. Kee T, Weiyan C, Blasiak A, Wang P, Chong JK, Chen J, et al. Harnessing CURATE.AI as a digital therapeutics platform by identifying N‐of‐1 learning trajectory profiles. Advanced Therapeutics 2019;2(9):1900023. [doi: 10.1002/adtp.201900023]
17. Lam TK, Vartanian O, Hollands JG. The brain under cognitive workload: neural networks underlying multitasking performance in the multi-attribute task battery. Neuropsychologia 2022;174:108350. [doi: 10.1016/j.neuropsychologia.2022.108350] [Medline: 35988804]
18. von Bastian CC, Belleville S, Udale RC, Reinhartz A, Essounni M, Strobach T. Mechanisms underlying training-induced cognitive change. Nat Rev Psychol 2022;1(1):30-41. [doi: 10.1038/s44159-021-00001-3]
19. Blasiak A, Khong J, Kee T. CURATE.AI: optimizing personalized medicine with artificial intelligence. SLAS Technol 2020;25(2):95-105. [doi: 10.1177/2472630319890316] [Medline: 31771394]
20. Remus A, Tadeo X, Kai GNS, Blasiak A, Kee T, Vijayakumar S, et al. CURATE.AI COR-Tx platform as a digital therapy and digital diagnostic for cognitive function in patients with brain tumour postradiotherapy treatment: protocol for a prospective mixed-methods feasibility clinical trial. BMJ Open 2023;13(10):e077219 [FREE Full text] [doi: 10.1136/bmjopen-2023-077219] [Medline: 37879700]
21. Wang P, Leong QY, Lau NY, Ng WY, Kwek SP, Tan L, et al. N-of-1 medicine. Singapore Med J 2024;65(3):167-175 [FREE Full text] [doi: 10.4103/singaporemedj.SMJ-2023-243] [Medline: 38527301]
22. Tan SB, Tan J, Raczkowska MN, Chean Wen Lee J, Rai B, Remus A, et al. Digital game-based interventions for cognitive training in healthy adults and adults with cognitive impairment: protocol for a two-part systematic review and meta-analysis. BMJ Open 2023;13(5):e071059 [FREE Full text] [doi: 10.1136/bmjopen-2022-071059] [Medline: 37142320]
23. Koivisto J, Malik A. Gamification for older adults: a systematic literature review. Gerontologist 2021;61(7):e360-e372 [FREE Full text] [doi: 10.1093/geront/gnaa047] [Medline: 32530026]
24. Al-Saleh S, Lee JK, Rogers WA, Insel KC. Translation of a successful behavioral intervention to a digital therapeutic self-management system for older adults. Ergon Des 2024;32(2):5-13 [FREE Full text] [doi: 10.1177/10648046211066409] [Medline: 38487251]
25. Bohle H, Rimpel J, Schauenburg G, Gebel A, Stelzel C, Heinzel S, et al. Behavioral and neural correlates of cognitive-motor interference during multitasking in young and old adults. Neural Plast 2019;2019 [FREE Full text] [doi: 10.1155/2019/9478656] [Medline: 31582967]
26. Tye-Murray N, Spehar B, Mauze E, Cardinal C. Hearing health care digital therapeutics: patient satisfaction evidence. Am J Audiol 2022;31(3S):905-913. [doi: 10.1044/2022_aja-21-00236]
27. Ennis GE, Hess TM, Smith BT. The impact of age and motivation on cognitive effort: implications for cognitive engagement in older adulthood. Psychol Aging 2013;28(2):495-504 [FREE Full text] [doi: 10.1037/a0031255] [Medline: 23421325]
28. Boot WR, Souders D, Charness N, Blocker K, Roque N, Vitale T. In: Zhou J, Salvendy G, editors. The Gamification of Cognitive Training: Older Adults? Perceptions ofAttitudes Toward Digital Game-Based Interventions. United States: Springer International Publishing; 2016.

XSL•FO
RenderX

29. Booth A, Hannes K, Harden A, Noyes J, Harris J, Tong A. COREQ (Consolidated Criteria for Reporting Qualitative Studies). In: Moher D, Altman DG, Schulz KF, Simera I, Wager E, editors. Guidelines for Reporting Health Research: A User's Manual. United States: Wiley; 2014.
30. Patton MQ. Enhancing the quality and credibility of qualitative analysis. Health Serv Res 1999;34(5 Pt 2):1189-1208 [FREE Full text] [Medline: 10591279]
31. Holden RJ, Karsh BT. The technology acceptance model: its past and its future in health care. J Biomed Inform 2010 Mar;43(1):159-172 [FREE Full text] [doi: 10.1016/j.jbi.2009.07.002] [Medline: 19615467]
32. Braun V, Clarke V. Using thematic analysis in psychology. Qualitative Research in Psychology 2006;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
33. Jefferies LN, Roggeveen AB, Enns JT, Bennett PJ, Sekuler AB, Di Lollo V. On the time course of attentional focusing in older adults. Psychol Res 2015;79(1):28-41. [doi: 10.1007/s00426-013-0528-2] [Medline: 24337971]
34. Read CA, Rogers JM, Wilson PH. Working memory binding of visual object features in older adults. Neuropsychol Dev Cogn B Aging Neuropsychol Cogn 2016;23(3):263-281. [doi: 10.1080/13825585.2015.1083937] [Medline: 26344033]
35. Ebaid D, Crewther SG. Visual information processing in young and older adults. Front Aging Neurosci 2019;11:116 [FREE Full text] [doi: 10.3389/fnagi.2019.00116] [Medline: 31156422]
36. Sharit J, Czaja SJ. Overcoming older adult barriers to learning through an understanding of perspectives on human information processing. J Appl Gerontol 2020;39(3):233-241 [FREE Full text] [doi: 10.1177/0733464818794574] [Medline: 30160573]
37. Nielsen J, Molich R. Heuristic evaluation of user interfaces. 1990 Presented at: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems; 1990 April 5; Washington USA. [doi: 10.1145/97243.97281]
38. Clark RC, Mayer RE. e-Learning and the Science of Instruction: Proven Guidelines for Consumers and Designers of Multimedia Learning. Hoboken: Wiley; 2016.
39. Erhel S, Jamet E. Digital game-based learning: Impact of instructions and feedback on motivation and learning effectiveness. Computers & Education 2013;67:156-167. [doi: 10.1016/j.compedu.2013.02.019]
40. Low STH, Sakhardande PG, Lai YF, Long ADS, Kaur-Gill S. Attitudes and perceptions toward healthcare technology adoption among older adults in singapore: a qualitative study. Front Public Health 2021;9:588590 [FREE Full text] [doi: 10.3389/fpubh.2021.588590] [Medline: 33659232]
41. Granić A, Marangunić N. Technology acceptance model in educational context: a systematic literature review. Brit J Educational Tech 2019;50(5):2572-2593. [doi: 10.1111/bjet.12864]
42. Hamari J, Shernoff DJ, Rowe E, Coller B, Asbell-Clarke J, Edwards T. Challenging games help students learn: an empirical study on engagement, flow and immersion in game-based learning. Comput Hum Behav 2016;54:170-179. [doi: 10.1016/j.chb.2015.07.045]
43. Abuhamdeh S, Csikszentmihalyi M. The importance of challenge for the enjoyment of intrinsically motivated, goal-directed activities. Pers Soc Psychol Bull 2012;38(3):317-330. [doi: 10.1177/0146167211427147] [Medline: 22067510]
44. Gilleade KM, Dix A. Using frustration in the design of adaptive videogames. 2004 Presented at: Proceedings of the 2004 ACM SIGCHI International Conference on Advances in Computer Entertainment Technology; June 5, 2005; Singapore. [doi: 10.1145/1067343.1067372]
45. Miller MK, Mandryk RL. Differentiating in-game frustration from at-game frustration using touch pressure. 2016 Presented at: Proceedings of the 2016 ACM International Conference on Interactive Surfaces and Spaces; November 9, 2016; United States. [doi: 10.1145/2992154.2992185]
46. Salmon JP, Dolan SM, Drake RS, Wilson GC, Klein RM, Eskes GA. A survey of video game preferences in adults: building better games for older adults. Entertain Comput 2017;21:45-64. [doi: 10.1016/j.entcom.2017.04.006]
47. Ramprasad C, Tamariz L, Garcia-Barcena J, Nemeth Z, Palacio A. The use of tablet technology by older adults in health care settings-is it effective and satisfying? A systematic review and meta analysis. Clin Gerontol 2019;42(1):17-26. [doi: 10.1080/07317115.2017.1322162] [Medline: 28665775]
48. Chen K, Lou VWQ, Lo SSC. Exploring the acceptance of tablets usage for cognitive training among older people with cognitive impairments: a mixed-methods study. Appl Ergon 2021;93:103381. [doi: 10.1016/j.apergo.2021.103381] [Medline: 33578065]
49. Infocomm Media Development Authority. Household access to computer by household composition, 2017 - 2022. URL: https://www.imda.gov.sg/about-imda/research-and-statistics/digital-society [accessed 2023-11-16]
50. Seifert A, Cotten SR, Xie B. A double burden of exclusion? Digital and social exclusion of older adults in times of COVID-19. J Gerontol B Psychol Sci Soc Sci 2021;76(3):e99-e103 [FREE Full text] [doi: 10.1093/geronb/gbaa098] [Medline: 32672332]
51. Stringer S. Census and the city: Overcoming NYC's digital divide in the 2020 census. 2019. URL: https://comptroller.nyc.gov/reports/census-and-the-city/ [accessed 2019-07-23]
52. Sutcliffe A. Designing for User Engagment: Aesthetic and Attractive User Interfaces. Cham: Springer International Publishing; 2022.
53. Sayago S, Rosales A, Righi V, Ferreira SM, Coleman GW, Blat J. On the conceptualization, design, and evaluation of appealing, meaningful, and playable digital games for older people. Games and Culture 2015;11(1-2):53-80. [doi: 10.1177/1555412015597108]

XSL•FO
RenderX

54. Landers RN, Bauer KN, Callan RC. Gamification of task performance with leaderboards: a goal setting experiment. Comput Hum Behav 2017;71:508-515. [doi: 10.1016/j.chb.2015.08.008]
55. Hennessey B, Moran S, Altringer B, Amabile TM. Extrinsic and intrinsic motivation. Wiley Encyclopedia of Management. 2015. URL: https://onlinelibrary.wiley.com/doi/10.1002/9781118785317.weom110098 [accessed 2024-12-21]
56. Botvinick M, Ritter S, Wang JX, Kurth-Nelson Z, Blundell C, Hassabis D. Reinforcement learning, fast and slow. Trends Cogn Sci 2019;23(5):408-422 [FREE Full text] [doi: 10.1016/j.tics.2019.02.006] [Medline: 31003893]
57. Botvinick M, Wang JX, Dabney W, Miller KJ, Kurth-Nelson Z. Deep reinforcement learning and its neuroscientific implications. Neuron 2020;107(4):603-616 [FREE Full text] [doi: 10.1016/j.neuron.2020.06.014] [Medline: 32663439]
58. Lee ST, Dutta MJ, Lin J, Luk P, Kaur-Gill S. Trust ecologies and channel complementarity for information seeking in cancer prevention. J Health Commun 2018;23(3):254-263. [doi: 10.1080/10810730.2018.1433253] [Medline: 29436966]
59. Lee VV, Vijayakumar S, Lau NY, Blasiak A, Siah KTH, Ho D. Understanding the user: patients' perception, needs, and concerns of health apps for chronic constipation. Digit Health 2022;8 [FREE Full text] [doi: 10.1177/20552076221104673] [Medline: 35663236]
60. Striley CW, Lloyd S, Varma D, Vaddiparti K, Cottler LB. 3487 trust in research among older adults. J. Clin. Trans. Sci 2019;3(s1):98-98. [doi: 10.1017/cts.2019.223]

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Studies
DTx: digital therapeutics
MATB: Multi-Attribute Task Battery
NASA: National Aeronautics and Space Administration
TAM: Technology Acceptance Model
USAF: United States Air Force

Edited by A Mavragani; submitted 17.07.24; peer-reviewed by N Buawangpong; comments to author 20.11.24; revised version received 04.12.24; accepted 09.12.24; published 13.01.25.

Please cite as:
Kwek SP, Leong QY, Lee VV, Lau NY, Vijayakumar S, Ng WY, Rai B, Raczkowska MN, Asplund CL, Remus A, Ho D
Exploring the General Acceptability and User Experience of a Digital Therapeutic for Cognitive Training in a Singaporean Older Adult Population: Qualitative Study
JMIR Form Res 2025;9:e63568
URL: https://formative.jmir.org/2025/1/e63568
doi:10.2196/63568
PMID:

©Siong Peng Kwek, Qiao Ying Leong, V Vien Lee, Ni Yin Lau, Smrithi Vijayakumar, Wei Ying Ng, Bina Rai, Marlena Natalia Raczkowska, Christopher L Asplund, Alexandria Remus, Dean Ho. Originally published in JMIR Formative Research (https://formative.jmir.org), 13.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Proposed mHealth Intervention to Address Patient Barriers to Colposcopy Attendance: Qualitative Interview Study of Clinic Staff and Patient Perspectives

Jennifer R Hemler[1,2], PhD; Rachel B Wagner[3,4], MA; Brittany Sullivan[2], MPH; Myneka Macenat[5], MPH; Erin K Tagai[6], MPH, PhD; Jazmarie L Vega[6], BA; Enrique Hernandez[6], MD; Suzanne M Miller[6], PhD; Kuang-Yi Wen[7], PhD; Charletta A Ayers[8], MPH, MD; Mark H Einstein[9], MD; Shawna V Hudson[1,2,3], PhD; Racquel E Kohler[3,4], MSPH, PhD

[1]Department of Family Medicine and Community Health, Rutgers Robert Wood Johnson Medical School, New Brunswick, NJ, United States

[2]Center Advancing Research and Evaluation for Person-Centered Care, Rutgers Robert Wood Johnson Medical School, New Brunswick, NJ, United States

[3]Center for Cancer Health Equity, Rutgers Cancer Institute, New Brunswick, NJ, United States

[4]School of Public Health, Rutgers University, Piscataway, NJ, United States

[5]Cancer Prevention and Outcomes Data Support, Rutgers Cancer Institute, New Brunswick, NJ, United States

[6]Cancer Prevention and Control, Fox Chase Cancer Center, Temple University Health System, Philadelphia, PA, United States

[7]Department of Medical Oncology, Thomas Jefferson University, Philadelphia, PA, United States

[8]Department of Obstetrics, Gynecology and Reproductive Sciences, Rutgers Robert Wood Johnson Medical School, New Brunswick, NJ, United States

[9]Department of Obstetrics & Gynecology and Women's Health, Montefiore Medical Center, Bronx, NY, United States

Corresponding Author:
Jennifer R Hemler, PhD
Department of Family Medicine and Community Health
Rutgers Robert Wood Johnson Medical School
303 George St.
Suite 301, Rm 305
New Brunswick, NJ, 08901
United States
Phone: 1 848 932 0209
Email: hemlerje@rwjms.rutgers.edu

Abstract

Background: Cervical cancer disparities persist among minoritized women due to infrequent screening and poor follow-up. Structural and psychosocial barriers to following up with colposcopy are problematic for minoritized women. Evidence-based interventions using patient navigation and tailored telephone counseling, including the Tailored Communication for Cervical Cancer Risk (TC3), have modestly improved colposcopy attendance. However, the efficacious TC3 intervention is human resource-intense and could have greater reach if adapted for mobile health, which increases convenience and access to health information.

Objective: This study aimed to describe feedback from clinic staff members involved in colposcopy processes and patients referred for colposcopy regarding adaptions to the TC3 phone-based intervention to text messaging, which addresses barriers among those referred for colposcopy after abnormal screening results.

Methods: Semistructured depth qualitative interviews were conducted over Zoom [Zoom Communications, Inc] or telephone with a purposive sample of 22 clinic staff members (including clinicians and support staff members) and 34 patients referred for colposcopy from 3 academic obstetrics and gynecology (OB-GYN) clinics that serve predominantly low-income, minoritized patients in different urban locations in New Jersey and Pennsylvania. Participants were asked about colposcopy attendance barriers and perspectives on a proposed text message intervention to provide tailored education and support in the time between abnormal cervical screening and colposcopy. The analytic team discussed interviews, wrote summaries, and consensus-coded transcripts, analyzing output for emergent findings and crystallizing themes.

XSL•FO
RenderX

Results: Clinic staff members and patients had mixed feelings about a text-only intervention. They overwhelmingly perceived a need to provide patients with appointment reminders and information about abnormal cervical screening results and colposcopy purpose and procedure. Both groups also thought messages emphasizing that human papillomavirus is common and cervical cancer can be prevented with follow-up could enhance attendance. However, some had concerns about the privacy of text messages and text fatigue. Both groups thought that talking to clinic staff members was needed in certain instances; they proposed connecting patients experiencing complex psychosocial or structural barriers to staff members for additional information, psychological support, and help with scheduling around work and finding childcare and transportation solutions. They also identified inadequate scheduling and reminder systems as barriers. From this feedback, we revised our text message content and intervention design, adding a health coaching component to support patients with complex barriers and concerns.

Conclusions: Clinic staff members and patient perspectives are critical for designing appropriate and relevant interventions. These groups conveyed that text message-only interventions may be useful for patients with lesser barriers who may benefit from reminders, basic educational information, and scheduling support. However, multimodal interventions may be necessary for patients with complex barriers to colposcopy attendance, which we intend to evaluate in a subsequent trial.

(JMIR Form Res 2025;9:e55043) doi:10.2196/55043

KEYWORDS

cervical cancer screening; colposcopy; HPV; human papillomavirus; mHealth; health communication; qualitative research; cancer screening; cancer; cervical cancer; screening; women; clinic staff; barrier; messaging; privacy; text message; qualitative; colposcopic; mhealth intervention; mobile phone

Introduction

Over the last few decades, the incidence of invasive cervical cancer in the United States has decreased due to the high uptake of preventive screening and access to human papillomavirus (HPV) vaccination [1-3]. However, there are large and persistent disparities in cervical cancer incidence [4,5] and mortality [6] among minoritized populations [7-11]. Importantly, cancer cases are higher among women living in socially deprived communities (eg, high proportions living below poverty) with high proportions of racial and ethnic minorities and historical racial segregation [4,12,13]. There is variation in screening by race and ethnicity, geography, and insurance [14-18], and low follow-up after an abnormal test result, especially for Black and Hispanic women [19-21]. Indeed, the majority of invasive cervical cancer cases are diagnosed among women who have never or rarely screened [8,22-24]. Guideline-concordant screening has decreased over the last 2 decades, leaving many women under-screened and at risk [25,26]. Poor follow-up behaviors and barriers have also been documented among women diagnosed with cancer [8,27].

Follow-up management for abnormal screening results among minoritized groups is influenced by a myriad of barriers. Patients' low awareness or lack of knowledge [9,28-33], feelings of anxiety [29,33], and fear of receiving a fatal cancer diagnosis [31,34] are well-documented barriers to follow-up attendance. Patient confusion after receiving and discussing abnormal Papanicolaou or cotesting (ie, Papanicolaou and HPV tests) results with providers is common [35,36]. Delays in results delivery and challenges in appointment scheduling also contribute to low follow-up. Structural barriers such as lack of time, transportation, and childcare have also hindered colposcopy attendance [30]. Neighborhood factors, specifically living in more resource-deprived communities, and distance to care also influence time to follow-up, receipt of treatment, and ultimately outcomes, particularly among Black and Hispanic women [13,37-39]. Addressing the multiple, complex barriers to colposcopy remains critical to cervical cancer prevention.

Interventions using patient navigation [40,41] and tailored telephone counseling [42-44] have modestly improved colposcopy attendance. However, these human resource-intense interventions may have greater reach if they are adapted for new mobile health (mHealth) delivery platforms, which increase convenience and access to health information through automated direct communication [45,46]. Interventions using mHealth strategies have improved other aspects of cervical cancer prevention [47]. For example, online patient education tools administered before appointments increased knowledge about HPV or Papanicolaou testing and follow-up treatment [48,49]. In resource-limited settings, text messaging and app-based interventions were acceptable reminders [50], with some successfully increasing triage attendance [51,52] and others reducing psychosocial barriers to recommended follow-up [53-55]. However, many interventions have been developed outside of the United States and use one-way text messages to deliver cervical cancer screening results, which have not improved follow-up [54,55]. These studies highlight the need for additional features and content to address barriers and promote attendance. The flexibility [56,57] and ease of tailoring [58] interactive text messages have potential to increase an intervention's impact, but acceptability and appropriateness must be evaluated for women referred for colposcopy.

Therefore, as part of a multisite effort to adapt an efficacious phone-based counseling intervention, Tailored Communication for Cervical Cancer Risk (TC3) [42,59], to be delivered through preprogrammed text messages [60,61], we sought to incorporate feedback from those engaged in colposcopy services—clinicians and support staff members—and patients who had been referred for colposcopy in clinics that serve minoritized populations. In this manuscript, we report the major findings from interviews where participants provided insight on critical barriers to colposcopy attendance and feedback on the proposed SMS text

XSL•FO
RenderX

message intervention, useful functions the intervention should provide, and preferred delivery mode.

Methods

Intervention Adaptation

Our adapted intervention of the TC3 phone-based intervention, "Health Enhancement Resource System to address urban, cervical cancer disparities," is a pilot SMS text message intervention designed to deliver tailored educational and counseling messages to women who were referred for colposcopy after receiving abnormal cervical screening results. In TC3, patients completed a telephone-based session with a trained health educator. Using a computer-assisted telephone interviewing system, the health educator asked the patient a series of questions related to five Cognitive-Social Health Information-Processing (C-SHIP)–informed topic areas: (1) knowledge and perceptions of the implications of positive HPV and abnormal Papanicolaou tests, cervical cancer, and colposcopy; (2) expectancies and beliefs concerning colposcopy; (3) interfering affect in relation to cancer and colposcopy; (4) values and goals related to follow-up and physical well-being; and (5) self-regulation in terms of overcoming and coping with potential barriers to colposcopy attendance, including structural barriers [43,62]. Based on patients' responses, the computer-assisted telephone interviewing system algorithm gave the health educator a message to read to the patient for each C-SHIP barrier. In the Health Enhancement Resource System, patients awaiting colposcopy would receive a series of preprogrammed text messages adapted for text message delivery from TC3. For each domain, participants would receive a text asking them to assess their knowledge, level of concern, or perceived barrier, depending on the topic, on a 5-point scale. Participants would receive an intervention text message tailored to their response. For example, if expressing low knowledge or high concern about the barrier, they would receive an educational or counseling message; if expressing high knowledge or low concern about the barrier, they would receive an affirmation and move to the next text in the series. Participants could potentially receive all texts within 1 texting session or within a 2-week-long period before their colposcopy appointment.

Study Design and Setting

Qualitative semistructured depth interviews developed within a constructivist approach [63] were used to discuss facilitators and barriers to colposcopy attendance and to elicit feedback about the fit of the adapted text message intervention with clinic staff members and patients. We developed different interview guides for clinic staff members and patients, including additional questions on the patient interview guide in relation to domains of the C-SHIP model (Multimedia Appendix 1). The objectives for both sets of interviews were to determine whether the adapted intervention aligned with patients' needs and preferences and clinics' communication and counseling processes. For this study, we recruited a purposive sample of clinic staff members, including physicians, and patients. We asked participants about their perceptions and either personal or known barriers to colposcopy attendance as well as their thoughts about what would be useful in a text-based intervention aiming to help patients attend their colposcopy appointments. We included questions about demographics and social characteristics. In addition, we asked patient participants pointed questions about intervention modality preference, frequency, and timing of receiving messages.

The study included 3 obstetrics and gynecology (OB-GYN) and colposcopy clinics affiliated with 2 universities that serve patients predominantly of low socioeconomic status and minoritized groups in different urban locations in New Jersey (Site A, clinics 1 and 2) and Pennsylvania (Site B, clinic 3). The 3 clinics within the 2 sites were selected with the intent to pilot the adapted text message intervention after receiving feedback from this interview study. The clinics have similar patient care management processes and perform similar rates of screening and follow-up care, including colposcopy.

Participant Recruitment and Data Collection

Potential clinic staff member participants were referred to the study team by a clinic contact, who sent the team a list of clinic staff members involved in colposcopy services. At Site A, the clinics also sent the team a list of patients screened for eligibility and participation interest. Site B identified eligible patients through the electronic medical records, reviewing patients recently scheduled for colposcopy appointments. Clinic staff members were eligible to participate in interviews if they played a role in colposcopy services and were not a part of the research team. Patients were eligible if they were 18-70 years old, owned a cell phone with texting capability, and had received a colposcopy referral in the past year (completing the colposcopy was not a requirement). Patients were excluded if they were pregnant at the time of recruitment, had a history of malignancy, or had been diagnosed with invasive carcinoma of the cervix. Both groups of participants needed to be able to consent to participate. Site A participants were eligible if they spoke English or Spanish, whereas Site B participants were eligible if they spoke English.

A team of experienced health services researchers (JH, RB, BS, MM, JV, and Nicole Hernandez) trained in qualitative methods by a PhD health services researcher (JH and ET) and overseen by the larger team of cancer prevention and control, methodological, and clinical experts conducted qualitative semistructured depth interviews from July 2021 through March 2022 at Site A and until January 2023 at Site B. The team made up to 5 outreach attempts by phone and email to enroll potential participants into the study and emailed potential participants the study consent form to review ahead of the scheduled interview. Researchers answered all participant questions before asking participants for their consent. The trained researchers, all female and one bilingual in English and Spanish, conducted interviews in the participant's preferred language. Interviews were held over the telephone or via Zoom.

We received contact information for 50 eligible clinic staff members (n=16 at Site A; n=34 at Site B) and 145 eligible patient participants (n=29 at Site A; n=116 at Site B). Across sites, we were unable to reach 17 clinic staff members and 61 patients within the prescribed number of attempts. A total of 39 patients declined and 4 dropped out of the study after

XSL•FO
RenderX

providing consent. We were not able to schedule 4 clinic staff members and 7 patient interviews within the immediate time frame; saturation was achieved before being able to schedule these few interviews. Across sites, we interviewed 22 clinic staff members and 34 patients.

On average, clinician interviews were 53 minutes and patient interviews were 45 minutes. All interviews were recorded, professionally transcribed, translated from Spanish to English if necessary, and deidentified.

Data Analysis

At each site, the same team of health services researchers who conducted the interviews met routinely to discuss interviews on a rolling basis, writing summaries and creating a matrix of participant responses to C-SHIP–informed barriers and perspectives on the content and mode of the proposed intervention. From these materials, we created a codebook of a priori and emergent topics. Interviews were concluded upon reaching saturation at each site, which was determined upon discussing interviews and finding that information aligned with previous interviews and did not offer substantially different perspectives. After this, the teams concretized the code book and coded in pairs using Atlas.ti software (Lumivero; version 9) to reach consensus upon code application. Each site's analytic team discussed output and emergent themes with the larger study team, which included clinicians and health services researchers with expertise in cervical cancer prevention. The 2 sites compared and contrasted thematic findings, iterating themes until reaching crystallization [63]. The authors referred to the COREQ (Consolidated Criteria for Reporting Qualitative Research) [64] in reporting this research.

Ethical Considerations

This study was approved by Rutgers University's and Fox Chase Cancer Center's institutional review boards (Pro2020002750 and 21-1080, respectively). All participants provided informed consent prior to being interviewed. Data were deidentified. Participants were given a US $30 gift card for taking part in the study.

Results

Sample Characteristics

The 22 clinic staff members interviewed across clinics and sites were similar (Table 1). Most were physicians, female, and non-Hispanic White. Clinic participants from Site A averaged 48 years of age; however, age was not assessed at Site B. Clinic staff members from both sites reported serving various roles related to colposcopy services; some physicians performed both screening (ie, HPV and Papanicolaou tests) and colposcopies while others primarily performed colposcopies in specialized clinics or settings (ie, colposcopy clinic and gynecologic oncology team). Support staff members typically scheduled colposcopies for patients, while physicians or advanced practice providers communicated with patients directly about their colposcopy results and the next steps thereafter.

The 34 patients interviewed across the clinics were similar in most regards. The majority were non-Hispanic Black, worked full-time, and were single, divorced, or separated. We noted variation among patients in terms of age and education across the clinics. At Site A, Clinic 1, participants averaged 46 years old and were more highly educated than the other 2 clinics. Clinic 2 participants were younger on average than the other 2 clinics, averaging 32 years old; these participants completed lower levels of education than the other 2 clinics. Site B, Clinic 3 participants averaged 41 years of age and varied across income and education (Table 2).

All patient participants, from both sites and the 3 clinics, had received an abnormal Papanicolaou test or positive HPV test result and were referred for colposcopy in the past year. Half were referred for their first colposcopy. All had smartphones and reported using their cell phones for multiple activities, like talking, texting, making appointments, using health and fitness and other apps, and using social media. While all reported text messaging as one of the main activities they did with their phone, their estimated phone time varied widely.

Table 1. Obstetrics and gynecology (OB-GYN) and colposcopy clinic staff members' demographics by participating sites.

Characteristics	Site A		Site B	Total (N=22)
	Clinic 1 (n=8), n (%)	Clinic 2 (n=5), n (%)	Clinic 3 (n=9), n (%)	
Race and ethnicity, n (%)				
Non-Hispanic Black	0 (0)	2 (40.0)	2 (22.2)	4 (18.2)
Non-Hispanic White	5 (62.5)	2 (40.0)	6 (66.7)	13 (59.1)
Non-Hispanic Asian	2 (25.0)	0 (0)	0 (0)	2 (9.1)
Hispanic	1 (12.5)	1 (20.0)	0 (0)	2 (9.1)
More than one race	0 (0)	0 (0)	1 (10.1)	1 (4.5)
Job title, n (%)				
Physician	5 (62.5)	3 (60.0)	7 (77.8)	15 (68.2)
Nurse	1 (12.5)	0 (0)	2 (22.2)	3 (13.6)
Advanced practice provider	1 (12.5)	1 (20.0)	0 (0)	2 (9.1)
Patient support	1 (12.5)	1 (20.0)	0 (0)	2 (9.1)

XSL•FO
RenderX

Table 2. Demographic characteristics for patients referred for colposcopy at participating sites.

Characteristics	Site A		Site B	Total (N=34)
	Clinic 1 (n=9)	Clinic 2 (n=6)	Clinic 3 (n=19)	
Age (mean years, SD)	46.11 (15.01)	31.67 (7.39)	41.32 (10.06)	40.88 (11.88)
Race and ethnicity, n (%)				
Non-Hispanic Black	3 (33.3)	4 (66.6)	14 (73.7)	21 (61.8)
Non-Hispanic White	2 (22.2)	1 (16.7)	2 (10.5)	5 (14.7)
Non-Hispanic Asian	2 (22.2)	0 (0)	0 (0)	2 (5.9)
Hispanic	1 (11.1)	1 (16.7)	3 (18.8)	5 (14.7)
More than one race	1 (11.1)	0 (0)	0 (0)	1 (2.9)
Marital status, n (%)				
Single or never married	2 (22.2)	4 (66.6)	13 (68.4)	19 (55.9)
Married or living with partner	5 (55.6)	1 (16.7)	2 (10.5)	8 (23.5)
Divorced or separated or widowed	2 (22.2)	1 (16.7)	4 (21.1)	7 (20.6)
Employment status, n (%)				
Employed full-time	4 (44.4)	3 (50.0)	9 (47.4)	16 (47.0)
Employed part-time	0 (0)	1 (16.7)	4 (21.0)	5 (14.7)
Unemployed	1 (11.1)	2 (33.3)	4 (21.0)	7 (20.6)
Disabled	2 (22.2)	0 (0)	2 (10.5)	4 (11.8)
Retired	2 (22.2)	0 (0)	0 (0)	2 (5.9)
Annual income, n (%)				
US $15,000	0 (0)	0 (0)	6 (31.6)	6 (17.6)
US $15,001-US $30,000	0 (0)	4 (66.6)	6 (31.6)	10 (29.5)
>US $30,000	6 (66.7)	2 (33.3)	7 (36.8)	15 (44.1)
Don’t know or refused	3 (33.3)	0 (0)	0 (0)	3 (8.8)
Educational attainment, n (%)				
Less than a high-school diploma	0 (0)	0 (0)	2 (10.6)	2 (5.9)
High-school diploma or equivalent	0 (0)	5 (83.3)	8 (42.1)	13 (38.2)
Some college	0 (0)	0 (0)	4 (21.0)	4 (11.8)
Vocational school	0 (0)	0 (0)	3 (15.8)	3 (8.8)
Bachelor’s degree or higher	9 (100)	1 (16.7)	2 (10.6)	12 (35.3)
Follow-up appointment is for patient’s first abnormal cervical test, n (%)	5 (55.6)	4 (66.6)	8 (42.1)	17 (50.0)

Qualitative Findings

Below, we highlight the critical feedback participants gave regarding the text-based intervention. Clinicians and patients, despite patients’ demographic differences across clinics, were remarkably similar in their perspectives on the barriers they thought women encounter in attending their colposcopy appointments and in their thoughts on our proposed intervention concept, including the types of messages we should include. They had mixed preferences for the mode by which they thought counseling and informational content should be delivered. We report these themes and nuances below.

Address Variable Colposcopy Attendance Barriers

Clinic staff members and patients posited that women who do not attend their colposcopy appointments experience variable socioemotional and structural barriers (Table 3). They thought fear of having a biopsy or being afraid of receiving cancerous results prevented some women from attending appointments but might compel other women to follow up. Many clinic staff members stated that they thought patients would attend appointments if they understood the importance of the procedure in preventing cervical cancer. However, most of the patients we interviewed attended their appointments without feeling knowledgeable or prepared; they explained that they attended appointments to lessen their fear and take care of their health. Patients and clinic staff members acknowledged that structural barriers were difficult for many women to overcome, especially

XSL•FO
RenderX

inflexible or hourly work and lack of childcare, transportation, or inadequate health insurance. They noted that clinics not having evening or weekend appointment hours and needing patients to schedule appointments several months ahead likely increased the potential for patients not attending appointments. In addition, clinic staff members admitted their scheduling systems for making appointments were unnecessarily cumbersome, especially when patients needed to call back after receiving screening results rather than being able to schedule their colposcopy appointment during the same call. Patients thought being able to schedule through an automated online system, like a portal or website, would be more efficient and easier than having to call a scheduler, especially because they could use automated systems after clinic hours.

Table 3. Clinic and patient perspectives on common colposcopy attendance barriers.

Theme	Clinicians	Patients
Fear or avoidance	*Sometimes people are scared of the biopsies and sometimes they don't think that it's [the colposcopy] necessary.* [Site A, Clinic 1, Clinician, non-Hispanic Asian, 43 years old]	*They're fearful that it could be cancer, so that's why some people say, oh, just let it be. . . . They don't want to discover that they have positive results.* [Site A, Clinic 1, non-Hispanic Asian, 66 years old]
Structural barriers	*They face a lot of challenges and a lot of struggles, just being able to get to their appointments . . . . They're making sure the appointment is going to be covered with their insurance. . . or they don't have insurance. . . . being able to take off of work to come to clinic.* [Site B, Clinic 3, Clinician, non-Hispanic White, age not assessed]	*The childcare, because I'm a single mother and then I can't schedule it [at certain times]. . . . [In addition,] it takes me an hour to drive to the hospital.* [Site A, Clinic 2, non-Hispanic White, 28 years old]
Scheduling process	*Part of it is that our schedules book out a little bit far. Our colposcopy services, they're only offered at certain time of the day… scheduling can be important to our patients, especially ones who work non-traditional schedules.* [Site B, Clinic 3, Clinician, non-Hispanic Black, age not assessed]	*I was literally on the phone with her for ten minutes . . . between my work schedule and then you can't go there while you're having your period. And then I was also going on vacation. So really? I think we bounced between – I don't know – 10 different dates*. [Site A, Clinic 1, non-Hispanic Black, 40 years old]

Provide Practical Health Information and Supportive Messaging

All participants expressed a need for an intervention to provide pragmatic, educational, and counseling messages during the time between receipt of abnormal cervical screening results and colposcopy appointments to help patients overcome socioemotional and knowledge-based barriers to attendance. They thought sending appointment reminders through text was essential. They also suggested that sharing information with patients that emphasized prevention, prepared patients for the procedure, and reduced anxiety would be key to improving colposcopy attendance (Table 4).

Table 4. Clinic and patient perspectives on the types of messages intervention should include.

Theme	Clinic staff members	Patients
Provide patient appointment reminders	*I think our whole patient-reminder system, which is just a phone call with a message left, is probably not the most effective. . . . Text messages are better. . . . 99 percent of the other appointments I go to, I get a text message that says, "reply C to confirm.* [Site A, Clinic 1, Clinician, non-Hispanic White, 48 years old]	*It's sometimes unclear about when to follow up or if it's necessary . . . what I should do when I leave the doctor. . . . I think it helps just to – yeah, just like as a reminder and knowing more about what's going on with your appointments.* [Site A, Clinic 1, Biracial, non-Hispanic Black and White, 27 years old]
Emphasize prevention	*But even if there was a short little thing … would help them to understand why they need to go because I find that so many of our patients . . . they, a lot of times, are like, "Oh, I don't know. I just got a call that I have this appointment.* [Site B, Clinician, Biracial, non-Hispanic White and Asian, age not assessed]	*I don't know whether "cancer" will scare people off [but] maybe if you have an abnormal Pap smear and then you need a colposcopy maybe you need to tell them. You need this because this is a test that allows us to detect whether you have the beginnings, right.* [Site A, Clinic 1, non-Hispanic Black, 63 years old]
Reinforce information and prepare patients	*[A]ny reinforcement is excellent. . . . [A]s soon as you leave the office, you realize that you've got a question that you wanted to ask. And this may be something that helps.* [Site A, Clinic 2, Clinician, non-Hispanic Asian, 58 years old]	*I had to look everything up on my own. . . . [S]ome people might not even know where to start, or they may be too nervous to look it up. So if it was right there or provided, it may be a little easier.* [Site B, Clinic 3, non-Hispanic Black, 32 years old]
Reduce anxiety and provide support	*They've been anxious about all of these questions at home [and] these [texts] can relieve that. . . . This is exactly what we counsel them in the clinic….* [Site A, Clinic 1, Clinician, non-Hispanic Asian, 38 years old]	*I was so afraid of that appointment. So having a way to calm my nerves, to make me feel all right about the procedure, that is 100% for me.* [Site B, Clinic 3, Hispanic, 44 years old]

XSL•FO
RenderX

Provide Patient Appointment Reminders

When first asked about a potential text message intervention, most participants, clinic staff members, and patients assumed that the texts would be reminder texts. While clinics used patient portals, none had a text messaging system and relied on phone calls and letters to patients that clinic staff felt were "not the most effective." Clinic staff members thought text reminders could alert patients to schedule childcare, find transportation or make other arrangements that would help enable them to attend their appointments. In addition, some patients noted that reminders would be helpful for them to track follow-up needs. Some said they were confused about when, if, and what kind of follow-up appointment they needed. A text could help know "when to follow-up or if it's necessary . . . what I should do when I leave the doctor. . . ."

Emphasize That Cervical Cancer Can Be Prevented

Clinic staff members and patients also agreed that focusing on prevention and early detection was critical to increasing colposcopy attendance. Multiple participants noted that receiving text messages that explained why colposcopy was needed and how the procedure would prevent cancer–"help them understand 'why' they need to go"–would increase the likelihood of making an appointment. Clinic staff members and patients expressed that the texts needed to strike a delicate balance, presenting an adequate amount of information regarding cancer risk without instilling fear or raising new concerns; as 1 patient said, "I don't know whether [using the word] 'cancer' will scare people off…but maybe you need to tell them." Clinic staff members also suggested that messages could be an opportunity to promote HPV vaccination for cancer prevention and should be included in the intervention.

Reinforce Information Given by the Clinic and Prepare Patients for the Procedure

Clinic staff members said that the proposed messaging aligned with patients' frequently asked questions and what they tell patients. They thought redelivering this information was important, as they believed patients did not retain information about colposcopy when receiving their abnormal screening results: "any reinforcement is excellent." Patients said that texts about what to expect at the visit, including how long the procedure would take, if they would experience pain, and if they would need assistance post procedure, would help patients better prepare for the visit. Clinic staff members and patients thought including postcolposcopy instructions, such as how long to avoid sex if a biopsy is taken and the potential for light spotting to occur, would be helpful in preparing patients. Some patients wanted to know what follow-up plans could entail; however, some clinic staff members thought the purpose of the texts should be to encourage women to attend their immediate appointment and thought that projecting too far into the future would be too complicated and case-specific to capture in a brief text.

Reduce Anxiety and Provide Support

Patients reported not receiving enough information when they received their abnormal screening results and said that having more information would have eased some of the anxiety they felt while waiting for their colposcopy. They wanted information from a trustworthy source, such as their health care provider's office, instead of having to search online, and they wanted detailed information on types of HPV, HPV transmission, and potential consequences of their abnormal screening results. Younger women wanted information on fertility, especially. Clinic staff members acknowledged that patients have anxiety–"they've been anxious about all of these questions at home"–and that the texts may help address that anxiety: "this is exactly what we counsel them in the clinic." However, while some clinic staff members thought explaining some of these topics over text might be too complex or not generalizable to everyone receiving the texts, they agreed that explaining how the HPV virus may lie dormant for years and therefore may not be a direct result of a recent infidelity would help reduce the stigma of sexually transmitted infections. Some patients thought the intervention should include messages offering encouraging words or affirmations, which would help reduce stress: "I was so afraid… having a way to calm my nerves… that is 100% for me."

Enhance Text Messages With Additional Forms of Information or Support

While clinic staff members and patients responded positively to a patient-facing intervention that provided reminders, health information, and supportive messages, they had mixed feelings about a text-only intervention. Many in our sample preferred texts because they wanted "very direct," "short, quick pieces of information"; they thought texts would help patients "know quickly what [they] needed to do" (Table 5). Others liked the idea of texts, but suggested texts be supplemented with infographics and diagrams or links to credible websites where more detailed information could be found. Both clinic staff members and patients thought texts should provide a phone number to a clinic contact in the event a patient had additional questions and suggested linking patients to additional counseling services or a support group should patients experience elevated concern or distress. Some clinic staff members noted that patients could call the clinic if they needed help with insurance or transportation services, as well; some clinics have personnel like social workers, patient navigators, or front desk staff members who can work with patients to help them find a way to attend their appointments.

Some participants raised concerns about text messages in terms of fatigue and privacy. Some patients wanted the ability to opt in or out of threads, and a few preferred to have information sent by email or accessible through an app so they could peruse the intervention information in one place at one time at their convenience. Younger patients worried that someone else might see a message notification containing sensitive information on their phone screen. Clinic staff members worried about HIPAA (Health Insurance Portability and Accountability Act) violations or privacy and access issues for patients who shared phones with relatives. Some staff members thought older patients would not be as receptive to texts as younger patients, but we found no evidence for this from older patients in our sample.

XSL•FO
RenderX

Table 5. Clinic and patient preferences for mode of message delivery.

Theme	Clinicians	Patients
Text messages	*I think texting is a really good way to get short, quick pieces of information communicated.* [Site A, Clinic 1, Clinician, Non-Hispanic White, 48 years old]	*Texts tend to be shorter so I guess I would choose text. It's easier to read and then if I needed to do something, the message would be very direct so I would know quickly what I needed to do.* [Site B, Clinic 3, Non-Hispanic White, 37 years old]
Texts plus links to additional information	*Maybe text message ahead a link that you could click for information or also if you had any questions.* [Site A, Clinic 2, Support Staffs, Non-Hispanic White, 46 years old]	*Oh, that would be great, reliable information. . . . If there's a link that can explain to you what's happening and the reason why it's happening, that would be so helpful.* [Site B, Clinic 3, Hispanic, 44 years old]
Texts plus a person who can counsel	*I think certain things might need clarification. I just find that when I call patients . . . I don't know how that would translate through text. And I also wonder if it would feel a little impersonal as a patient to get like counseling through a text message.* [Site A, Clinic 1, Support Staff, Hispanic, 31 years old]	*There's some situations that text messages is not going to work. . . . [If] you need some information, I can send you something on text messages. You can read about it. [But] If you have more doubts and you feel like you need some support, I have a counselor that can assist you. . . .* [Site A, Clinic 1, Hispanic, 33 years old]

Discussion

Principal Results

We conducted this study to elicit feedback from clinic staff members and patients with colposcopy experience. Despite differences in our sample between clinic locations, participant roles, and patient age and educational characteristics, participants overwhelmingly agreed that an intervention providing colposcopy appointment reminders, abnormal screening results and colposcopy procedure information, and psychosocial and structural barriers counseling is needed for patients awaiting colposcopy appointments. As noted in the literature and echoed throughout this study, patients may experience varied and often multiple barriers to attendance [30]. In our adapted SMS text message intervention, we proposed to assess and address C-SHIP–related [62] barriers by providing information and counseling texts related to barriers patients endorse. This intervention has the potential to address many of the patient needs participants discussed, which align with C-SHIP constructs of knowledge and perception, expectancies and beliefs, and interfering affect. In addition, although there was interest in a text message intervention, participants expressed that texts alone may not be sufficient for or desired by all patients. Participants conveyed that patients who want additional information or who need additional support in overcoming barriers should be linked to counselors or other resources to enhance the intervention's potential to increase patient attendance.

mHealth strategies have been successful in improving prevention and health education among underserved minoritized populations [65], the population experiencing the greatest cervical cancer disparities. We heard from clinic staff members that health education is critical for follow-up attendance [66]. Our proposed text intervention is foundational for activated patients [67] who need basic information to assist in their decision-making about their colposcopy appointment. In our previous study, patients described being confused about the meaning of their screening results and having limited knowledge of HPV risk and colposcopy; seeking information on their own often increased anxiety and confusion [36]. In this study, clinic staff members reported that patients often do not absorb the information provided at the time of abnormal screening results delivery. Both patients and clinic staff members thought that messaging to reinforce the commonality of abnormal results but also the importance of colposcopy in cervical cancer prevention would be critical for patients seeking information to enhance their decision-making. Providing information of this kind may help reduce anxiety and increase knowledge, leading to increased colposcopy attendance for these types of patients. However, similar to other studies noting privacy concerns [68], the intervention messages will not contain patient information about their specific screening results.

Our findings also suggested a need for additional communication through other delivery modes for other kinds of patients. Texts may be limited in their ability to engage and support all patients and address all relevant topics. Participants acknowledged that patients with more complex or numerous psychosocial [69] or structural barriers [30] likely need additional support in overcoming their attendance barriers than text messages can provide. This group of patients would benefit from being connected to a point person at the practice or a professional counselor or support group, in addition to having the text message serve as a primer. Some clinic staff members noted that they can help connect patients to local resources for structural barriers like transportation or childcare or help navigate charity care or insurance needs. However, as staffing resources vary by clinic, this solution may not be scalable.

Importantly, the barrier experienced most by the patients we interviewed was scheduling, which is beyond the capacity of our proposed intervention. Work and childcare conflicts make scheduling difficult but having to schedule appointments months out aggravates these and other challenges especially pertinent to patients who have limited capacity to foresee or control their futures [18,70]. Clinic scheduling systems themselves were cumbersome for some patients. Including automated scheduling in conjunction with text reminders may help patients attend their follow-up appointments by reducing barriers to scheduling [71]. Patient-centered scheduling systems are essential structural elements that health systems can incorporate to improve

XSL•FO
RenderX

appointment attendance in general and colposcopy appointments specifically.

Future Research

Due to the strength and commonality of perspectives across clinic staff members and patient groups, and additional feasibility testing [61], the team expanded the intervention design to include brief health coaching sessions over the phone. Designing with efficiency and scalability in mind [72,73], this additional intervention component will be reserved for patients who experience complex barriers to attendance and missed appointments. Health coaching interventions, which are patient-centered and goal-focused, have improved health outcomes by empowering patients to self-manage their health or illness across multiple conditions and contexts [74-77]. The purpose of these calls will be to assess barriers, guide patients to existing resources, including connections to scheduling and structural barrier supports, and help them make a plan to attend the colposcopy appointment. The multimodal intervention will be tested this year.

Limitations

This study had several limitations to consider. As mentioned above, we interviewed patients referred for colposcopy, many of whom had already attended a colposcopy appointment. The patients and staff members we interviewed spoke about others they knew and could imagine scenarios in which they themselves would experience such barriers [42,60,78]. The small number of patients we recruited who did not attend their appointments (n=5) noted similar barriers and provided similar feedback on how to improve the intervention. We anticipate health coaching professionals may help this group address attendance barriers. Also, we interviewed participants from 2 urban academic OB-GYN sites that have support services like care coordinators for their patients. As this was a purposive sample to gather perspectives on our proposed intervention, we found these perspectives helpful. However, our results may not be generalizable to clinics and practices with lower attendance rates and fewer supports for patients.

Conclusions

Clinic staff members and patient perspectives are critical for designing appropriate and relevant interventions. These groups conveyed that text message-only interventions are useful for some patients but not for others. Multimodal interventions may be necessary for patients with complex barriers to colposcopy attendance, while patients with fewer barriers may benefit from reminders, basic educational information, and scheduling support.

Acknowledgments

This work was supported by a grant (R01 MD017675-01) and a Cancer Health Equity Pilot Award (REK and SVH) through Rutgers Cancer Institute of New Jersey, which is supported in part by an NIH/NCI Cancer Center Support Grant (P30CA072720). SVH was supported by the National Center for Advancing Translational Sciences (UL1TR003017). SMM was supported by the National Institutes of Health's National Cancer Institute (NIH/NCI) grant (P30CA006927) and the Marcia and Frank Carlucci Charitable Foundation Award (Prevent Cancer Foundation). REK was supported by the National Cancer Institute (K22CA258675). We would like to thank the patients who took part in this research as well as the staff at the OB-GYN practices, without whom this work would not have been possible. We would like to thank Nicole Hernandez for her work interviewing participants in Spanish.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Sample Interview Questions from Patient and Healthcare Staff Interview Guides.
[DOCX File , 42 KB - formative_v9i1e55043_app1.docx]

References

1. Adegoke O, Kulasingam S, Virnig B. Cervical cancer trends in the United States: a 35-year population-based analysis. J Womens Health (Larchmt) 2012 Oct;21(10):1031-1037 [FREE Full text] [doi: 10.1089/jwh.2011.3385] [Medline: 22816437]
2. Mix JM, Van Dyne EA, Saraiya M, Hallowell BD, Thomas CC. Assessing impact of HPV vaccination on cervical cancer incidence among women aged 15-29 years in the United States, 1999-2017: an ecologic study. Cancer Epidemiol Biomarkers Prev 2021 Jan;30(1):30-37 [FREE Full text] [doi: 10.1158/1055-9965.EPI-20-0846] [Medline: 33082207]
3. Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. CA Cancer J Clin 2023 Jan;73(1):17-48 [FREE Full text] [doi: 10.3322/caac.21763] [Medline: 36633525]
4. Roche LM, Niu X, Henry KA. Invasive cervical cancer incidence disparities in New Jersey--a spatial analysis in a high incidence state. J Health Care Poor Underserved 2015 Nov;26(4):1173-1185. [doi: 10.1353/hpu.2015.0127] [Medline: 26548671]
5. SEER*Explorer: an interactive website for SEER cancer statistics [Internet].: Surveillance Research Program, National Cancer Institute; 2024 Apr 17. URL: https://seer.cancer.gov/statistics-network/explorer/ [accessed 2024-09-06]

XSL•FO
RenderX

6. Beavis AL, Gravitt PE, Rositch AF. Hysterectomy-corrected cervical cancer mortality rates reveal a larger racial disparity in the United States. Cancer 2017 May 15;123(6):1044-1050 [FREE Full text] [doi: 10.1002/cncr.30507] [Medline: 28112816]
7. Benard VB, Lawson HW, Eheman CR, Anderson C, Helsel W. Adherence to guidelines for follow-up of low-grade cytologic abnormalities among medically underserved women. Obstet Gynecol 2005 Jun;105(6):1323-1328. [doi: 10.1097/01.AOG.0000159549.56601.75] [Medline: 15932824]
8. Benard VB, Jackson JE, Greek A, Senkomago V, Huh WK, Thomas CC, et al. A population study of screening history and diagnostic outcomes of women with invasive cervical cancer. Cancer Med 2021 Jun;10(12):4127-4137 [FREE Full text] [doi: 10.1002/cam4.3951] [Medline: 34018674]
9. Chase DM, Osann K, Sepina N, Wenzel L, Tewari KS. The challenge of follow-up in a low-income colposcopy clinic: characteristics associated with noncompliance in high-risk populations. J Low Genit Tract Dis 2012 Oct;16(4):345-351 [FREE Full text] [doi: 10.1097/LGT.0b013e318249640f] [Medline: 22622340]
10. Gaffney DK, Hashibe M, Kepka D, Maurer KA, Werner TL. Too many women are dying from cervix cancer: Problems and solutions. Gynecol Oncol 2018;151(3):547-554 [FREE Full text] [doi: 10.1016/j.ygyno.2018.10.004] [Medline: 30301561]
11. Olusola P, Banerjee HN, Philley JV, Dasgupta S. Human papilloma virus-associated cervical cancer and health disparities. Cells 2019;8(6):622 [FREE Full text] [doi: 10.3390/cells8060622] [Medline: 31234354]
12. Krieger N, Wright E, Chen JT, Waterman PD, Huntley ER, Arcaya M. Cancer stage at diagnosis, historical redlining, and current neighborhood characteristics: breast, cervical, lung, and colorectal cancers, Massachusetts, 2001-2015. Am J Epidemiol 2020;189(10):1065-1075 [FREE Full text] [doi: 10.1093/aje/kwaa045] [Medline: 32219369]
13. Sokale IO, Oluyomi AO, Montealegre JR, Thrift AP. Racial/Ethnic disparities in cervical cancer stage at diagnosis: mediating effects of neighborhood-level socioeconomic deprivation. Cancer Epidemiol Biomarkers Prev 2023;32(6):818-824 [FREE Full text] [doi: 10.1158/1055-9965.EPI-23-0038] [Medline: 37067295]
14. Suk R, Hong YR, Rajan SS, Xie Z, Zhu Y, Spencer JC. Assessment of US preventive services task force guideline-concordant cervical cancer screening rates and reasons for underscreening by age, race and Ethnicity, sexual orientation, rurality, and insurance, 2005 to 2019. JAMA Netw Open 2022;5(1):e2143582 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.43582] [Medline: 35040970]
15. Moss JL, Geyer NR, Lengerich EJ. Patterns of cancer-related healthcare access across pennsylvania: analysis of novel census tract-level indicators of persistent poverty. Cancer Epidemiol Biomarkers Prev 2024;33(4):616-623. [doi: 10.1158/1055-9965.EPI-23-1255] [Medline: 38329390]
16. McDaniel CC, Hallam HH, Cadwallader T, Lee HY, Chou C. Persistent racial disparities in cervical cancer screening with Pap test. Prev Med Rep 2021 Dec;24:101652 [FREE Full text] [doi: 10.1016/j.pmedr.2021.101652] [Medline: 34976700]
17. Sabatino SA, Thompson TD, White MC, Villarroel MA, Shapiro JA, Croswell JM, et al. Up-to-date breast, cervical, and colorectal cancer screening test use in the United States, 2021. Prev Chronic Dis 2023 Oct 26;20:E94 [FREE Full text] [doi: 10.5888/pcd20.230071] [Medline: 37884318]
18. Fuzzell LN, Perkins RB, Christy SM, Lake PW, Vadaparampil ST. Cervical cancer screening in the United States: challenges and potential solutions for underscreened groups. Prev Med 2021 Mar;144:106400. [doi: 10.1016/j.ypmed.2020.106400] [Medline: 33388330]
19. Ford S, Tarraf W, Williams KP, Roman LA, Leach R. Differences in cervical cancer screening and follow-up for black and white women in the United States. Gynecol Oncol 2021 Feb;160(2):369-374. [doi: 10.1016/j.ygyno.2020.11.027] [Medline: 33323276]
20. Boitano TKL, Ketch P, Maier JG, Nguyen CT, Huh WK, Michael Straughn J, et al. Increased disparities associated with black women and abnormal cervical cancer screening follow-up. Gynecol Oncol Rep 2022 Aug;42:101041 [FREE Full text] [doi: 10.1016/j.gore.2022.101041] [Medline: 35898199]
21. Spencer JC, Kim JJ, Tiro JA, Feldman SJ, Kobrin SC, Skinner CS, et al. Racial and ethnic disparities in cervical cancer screening from three U.S. healthcare settings. Am J Prev Med 2023 Oct;65(4):667-677 [FREE Full text] [doi: 10.1016/j.amepre.2023.04.016] [Medline: 37146839]
22. National Institutes of Health. National Institutes of Health Consensus Development Conference Statement: cervical cancer, April 1-3, 1996. National Institutes of Health Consensus Development Panel. J Natl Cancer Inst Monogr 1996(21):vii-xix. [Medline: 9023821]
23. Leyden WA, Manos MM, Geiger AM, Weinmann S, Mouchawar J, Bischoff K, et al. Cervical cancer in women with comprehensive health care access: attributable factors in the screening process. J Natl Cancer Inst 2005 May 04;97(9):675-683. [doi: 10.1093/jnci/dji115] [Medline: 15870438]
24. Landy R, Pesola F, Castañón A, Sasieni P. Impact of cervical screening on cervical cancer mortality: estimation using stage-specific results from a nested case-control study. Br J Cancer 2016 Oct 25;115(9):1140-1146 [FREE Full text] [doi: 10.1038/bjc.2016.290] [Medline: 27632376]
25. Rosenblum HG, Gargano JW, Cleveland AA, Dahl RM, Park IU, Whitney E, et al. U.S. women with invasive cervical cancer: characteristics and potential barriers to prevention. J Womens Health (Larchmt) 2024 May;33(5):594-603. [doi: 10.1089/jwh.2023.0462] [Medline: 38608239]

XSL•FO
RenderX

26. Holt HK, Peterson CE, MacLaughlan David S, Abdelaziz A, Sawaya GF, Guadamuz JS, et al. Mediation of racial and ethnic inequities in the diagnosis of advanced-stage cervical cancer by insurance status. JAMA Netw Open 2023 Mar 01;6(3):e232985 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.2985] [Medline: 36897588]
27. Subramaniam A, Fauci JM, Schneider KE, Whitworth JM, Erickson BK, Kim K, et al. Invasive cervical cancer and screening: what are the rates of unscreened and underscreened women in the modern era? J Low Genit Tract Dis 2011 Apr;15(2):110-113 [FREE Full text] [doi: 10.1097/LGT.0b013e3181f515a2] [Medline: 21263352]
28. Kulkarni A, Glynn S, Gamble CR, Shen MJ, Cantillo E, Frey MK, et al. Understanding perceived barriers to colposcopy follow-up among underserved women at an urban teaching hospital: a qualitative study. J Low Genit Tract Dis 2023;27(1):87-92. [doi: 10.1097/LGT.0000000000000700] [Medline: 36074132]
29. Hui SKA, Miller SM, Wen KY, Fang Z, Li T, Buzaglo J, et al. Psychosocial barriers to follow-up adherence after an abnormal cervical cytology test result among low-income, inner-city women. J Prim Care Community Health 2014;5(4):234-241 [FREE Full text] [doi: 10.1177/2150131914529307] [Medline: 24718518]
30. Glynn S, Cantillo E, Vasquez K, Shen M, Litvin D, Frey M, et al. Understanding knowledge gaps and perceived barriers to follow up after abnormal Pap smear among low-income women in NYC: a qualitative study. Gynecol Oncol 2021;162(1):S320-S321 [FREE Full text] [doi: 10.1016/S0090-8258(21)01259-2]
31. Nelson K, Geiger AM, Mangione CM. Effect of health beliefs on delays in care for abnormal cervical cytology in a multi-ethnic population. J Gen Intern Med 2002;17(9):709-716 [FREE Full text] [doi: 10.1046/j.1525-1497.2002.11231.x] [Medline: 12220368]
32. Nolan J, Renderos TB, Hynson J, Dai X, Chow W, Christie A, et al. Barriers to cervical cancer screening and follow-up care among black women in Massachusetts. J Obstet Gynecol Neonatal Nurs 2014;43(5):580-588. [doi: 10.1111/1552-6909.12488] [Medline: 25139164]
33. Percac-Lima S, Aldrich LS, Gamba GB, Bearse AM, Atlas SJ. Barriers to follow-up of an abnormal Pap smear in latina women referred for colposcopy. J Gen Intern Med 2010;25(11):1198-1204 [FREE Full text] [doi: 10.1007/s11606-010-1450-6] [Medline: 20652647]
34. Radecki Breitkopf C, Pearson HC. A theory-based approach to understanding follow-up of abnormal Pap tests. J Health Psychol 2009;14(3):361-371 [FREE Full text] [doi: 10.1177/1359105308101674] [Medline: 19293297]
35. Head KJ, Imburgia TM, Zimet GD, Shew ML. Women's understanding of their Pap and HPV test results: implications for patient–provider communication. J Healthc Commun 2017;10(1):37-46. [doi: 10.1080/17538068.2017.1282085]
36. Kohler RE, Hemler J, Wagner RB, Sullivan B, Macenat M, Tagai EK, et al. Confusion and anxiety in between abnormal cervical cancer screening results and colposcopy: "The land of the unknown". Patient Educ Couns 2023;114:107810. [doi: 10.1016/j.pec.2023.107810] [Medline: 37244133]
37. Spees LP, Brewster WR, Varia MA, Weinberger M, Baggett C, Zhou X, et al. Examining urban and rural differences in how distance to care influences the initiation and completion of treatment among insured cervical cancer patients. Cancer Epidemiol Biomarkers Prev 2019;28(5):882-889 [FREE Full text] [doi: 10.1158/1055-9965.EPI-18-0945] [Medline: 30733307]
38. Lim JW, Ashing-Giwa KT. Examining the effect of minority status and neighborhood characteristics on cervical cancer survival outcomes. Gynecol Oncol 2011;121(1):87-93. [doi: 10.1016/j.ygyno.2010.11.041] [Medline: 21183210]
39. Plascak JJ, Llanos AA, Pennell ML, Weier RC, Paskett ED. Neighborhood factors associated with time to resolution following an abnormal breast or cervical cancer screening test. Cancer Epidemiol Biomarkers Prev 2014;23(12):2819-2828 [FREE Full text] [doi: 10.1158/1055-9965.EPI-14-0348] [Medline: 25205516]
40. Glick SB, Clarke AR, Blanchard A, Whitaker AK. Cervical cancer screening, diagnosis and treatment interventions for racial and ethnic minorities: a systematic review. J Gen Intern Med 2012;27(8):1016-1032 [FREE Full text] [doi: 10.1007/s11606-012-2052-2] [Medline: 22798213]
41. Paskett ED, Dudley D, Young GS, Bernardo BM, Wells KJ, Calhoun EA, PNRP Investigators. Impact of patient navigation interventions on timely diagnostic follow up for abnormal cervical screening. J Womens Health (Larchmt) 2016;25(1):15-21 [FREE Full text] [doi: 10.1089/jwh.2014.5094] [Medline: 26625131]
42. Miller SM, Hui SA, Wen KY, Scarpato J, Zhu F, Buzaglo J, et al. Tailored telephone counseling to improve adherence to follow-up regimens after an abnormal pap smear among minority, underserved women. Patient Educ Couns 2013;93(3):488-495 [FREE Full text] [doi: 10.1016/j.pec.2013.08.008] [Medline: 24007767]
43. Miller SM, Siejak KK, Schroeder CM, Lerman C, Hernandez E, Helm CW. Enhancing adherence following abnormal Pap smears among low-income minority women: a preventive telephone counseling strategy. J Natl Cancer Inst 1997;89(10):703-708. [doi: 10.1093/jnci/89.10.703] [Medline: 9168185]
44. Lerman C, Hanjani P, Caputo C, Miller S, Delmoor E, Nolte S, et al. Telephone counseling improves adherence to colposcopy among lower-income minority women. J Clin Oncol 1992;10(2):330-333. [doi: 10.1200/JCO.1992.10.2.330] [Medline: 1732434]
45. Yang Q, Van Stee SK. The comparative effectiveness of mobile phone interventions in improving health outcomes: meta-analytic review. JMIR Mhealth Uhealth 2019;7(4):e11244 [FREE Full text] [doi: 10.2196/11244] [Medline: 30942695]

XSL•FO
RenderX

46. Marcolino MS, Oliveira JAQ, D'Agostino M, Ribeiro AL, Alkmim MBM, Novillo-Ortiz D. The impact of mHealth interventions: systematic review of systematic reviews. JMIR Mhealth Uhealth 2018;6(1):e23 [FREE Full text] [doi: 10.2196/mhealth.8873] [Medline: 29343463]
47. Bhochhibhoya S, Dobbs PD, Maness SB. Interventions using mHealth strategies to improve screening rates of cervical cancer: a scoping review. Prev Med 2021;143:106387. [doi: 10.1016/j.ypmed.2020.106387] [Medline: 33383069]
48. Lee HY, Koopmeiners JS, Rhee TG, Raveis VH, Ahluwalia JS. Mobile phone text messaging intervention for cervical cancer screening: changes in knowledge and behavior pre-post intervention. J Med Internet Res 2014;16(8):e196 [FREE Full text] [doi: 10.2196/jmir.3576] [Medline: 25164545]
49. Thiel de Bocanegra H, Dehlendorf C, Kuppermann M, Vangala SS, Moscicki AB. Impact of an educational tool on young women's knowledge of cervical cancer screening recommendations. Cancer Causes Control 2022;33(6):813-821 [FREE Full text] [doi: 10.1007/s10552-022-01569-8] [Medline: 35312891]
50. Romocki LS, Des Marais A, Cofie L, Anderson C, Curington T, Smith JS. Use of short messaging service to improve follow-up for abnormal pap test results in minority and medically underserved women in north carolina: questionnaire on attitudes and acceptability. JMIR Form Res 2019;3(3):e12675 [FREE Full text] [doi: 10.2196/12675] [Medline: 31389334]
51. Lee YH, Huang LH, Chen SH, Shao JH, Lai CH, Yang NP. Effects of mobile application program (App)-assisted health education on preventive behaviors and cancer literacy among women with cervical intraepithelial neoplasia. Int J Environ Res Public Health 2021;18(21):11603 [FREE Full text] [doi: 10.3390/ijerph182111603] [Medline: 34770117]
52. Arrossi S, Paolino M, Antelo VS, Thouyaret L, Kohler RE, Cuberli M, ATICA Study team. Effectiveness of an mHealth intervention to increase adherence to triage of HPV DNA positive women who have performed self-collection (the ATICA study): a hybrid type I cluster randomised effectiveness-implementation trial. Lancet Reg Health Am 2022;9:100199 [FREE Full text] [doi: 10.1016/j.lana.2022.100199] [Medline: 35655914]
53. Erwin E, Aronson KJ, Day A, Ginsburg O, Macheku G, Feksi A, et al. SMS behaviour change communication and eVoucher interventions to increase uptake of cervical cancer screening in the Kilimanjaro and Arusha regions of Tanzania: a randomised, double-blind, controlled trial of effectiveness. BMJ Innov 2019;5(1):28-34 [FREE Full text] [doi: 10.1136/bmjinnov-2018-000276] [Medline: 31645991]
54. Joseph NT, Namuli A, Kakuhikire B, Baguma C, Juliet M, Ayebare P, et al. Implementing community-based human papillomavirus self-sampling with SMS text follow-up for cervical cancer screening in rural, southwestern Uganda. J Glob Health 2021;11:04036 [FREE Full text] [doi: 10.7189/jogh.11.04036] [Medline: 35003710]
55. Linde DS, Andersen MS, Mwaiselage J, Manongi R, Kjaer SK, Rasch V. Effectiveness of one-way text messaging on attendance to follow-up cervical cancer screening among human papillomavirus-positive tanzanian women (Connected2Care): parallel-group randomized controlled trial. J Med Internet Res 2020;22(4):e15863 [FREE Full text] [doi: 10.2196/15863] [Medline: 32238335]
56. Kannisto KA, Koivunen MH, Välimäki MA. Use of mobile phone text message reminders in health care services: a narrative literature review. J Med Internet Res 2014;16(10):e222 [FREE Full text] [doi: 10.2196/jmir.3442] [Medline: 25326646]
57. Willcox JC, Dobson R, Whittaker R. Old-fashioned technology in the era of "Bling": is there a future for text messaging in health care? J Med Internet Res 2019;21(12):e16630 [FREE Full text] [doi: 10.2196/16630] [Medline: 31859678]
58. Suffoletto B. Text message behavioral interventions: from here to where? Curr Opin Psychol 2016;9:16-21 [FREE Full text] [doi: 10.1016/j.copsyc.2015.09.012] [Medline: 26665157]
59. Evidence-Based Cancer Control Programs (EBCCP) 2023. National Cancer Institute. 2023. URL: https://ebccp.cancercontrol.cancer.gov/programDetails.do?programId=26767713 [accessed 2024-09-09]
60. Sansoni M, Tagai EK, Lapitan E, Wen KY, Xu J, Belfiglio A, et al. Development of a text message-based intervention for follow-up colposcopy among predominately underserved Black and Hispanic/Latinx women. Cancer Causes Control 2022;33(6):861-873 [FREE Full text] [doi: 10.1007/s10552-022-01573-y] [Medline: 35334016]
61. Wen KY, Dayaratna S, Slamon R, Granda-Cameron C, Tagai EK, Kohler RE, et al. Chatbot-interfaced and cognitive-affective barrier-driven messages to improve colposcopy adherence after abnormal Pap test results in underserved urban women: A feasibility pilot study. Transl Behav Med 2024 Jan 11;14(1):1-12. [doi: 10.1093/tbm/ibad064] [Medline: 38014626]
62. Miller SM, Diefenbach MA. The cognitive-social health information-processing (C-SHIP) model: a theoretical framework for research in behavioral oncology. In: Technology and Methods in Behavioral Medicine. Mahwah: Lawrence Erlbaum Associates; 1998:219-244.
63. Crabtree BF, Miller WL. Doing Qualitative Research. Third Edition. Washington DC: SAGE Publications, Inc; 2023.
64. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
65. Anderson-Lewis C, Darville G, Mercado RE, Howell S, Di Maggio S. mHealth technology use and implications in historically underserved and minority populations in the United States: systematic literature review. JMIR Mhealth Uhealth 2018;6(6):e128 [FREE Full text] [doi: 10.2196/mhealth.8383] [Medline: 29914860]
66. Shokar NK, Calderon-Mora J, Molokwu J, Byrd T, Alomari A, Mallawaarachchi I, et al. Outcomes of a multicomponent culturally tailored cervical cancer screening intervention among underserved Hispanic women (De Casa en Casa). Health Promot Pract 2021;22(1):112-121. [doi: 10.1177/1524839919893309] [Medline: 31874564]

XSL•FO
RenderX

67. Miller SM, Tagai EK, Wen K, Lee M, Hui SKA, Kurtz D, et al. Predictors of adherence to follow-up recommendations after an abnormal pap smear among underserved inner-city women. Patient Educ Couns 2017;100(7):1353-1359 [FREE Full text] [doi: 10.1016/j.pec.2017.01.020] [Medline: 28190541]
68. Sanchez Antelo V, Kohler RE, Curotto M, Viswanath KV, Paolino M, Arrossi S. Developing SMS content to promote Papanicolaou triage among women who performed HPV self-collection test: qualitative study. JMIR Form Res 2020 Mar 06;4(3):e14652 [FREE Full text] [doi: 10.2196/14652] [Medline: 32032940]
69. Ilic I, Babic G, Dimitrijevic A, Sipetic Grujicic S, Ilic M. Psychological distress in women with abnormal Pap smear results attending cervical cancer screening. J Low Genit Tract Dis 2023 Oct 01;27(4):343-350. [doi: 10.1097/LGT.0000000000000761] [Medline: 37535069]
70. Atlas SJ, Tosteson ANA, Burdick TE, Wright A, Breslau ES, Dang TH, et al. Primary care practitioner perceptions on the follow-up of abnormal cancer screening test results. JAMA Netw Open 2022;5(9):e2234194 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.34194] [Medline: 36173627]
71. Feldman S, Lykken JM, Haas JS, Werner CL, Kobrin SC, Tiro JA, et al. Factors associated with timely colposcopy following an abnormal cervical cancer test result. Prev Med 2022;164:107307 [FREE Full text] [doi: 10.1016/j.ypmed.2022.107307] [Medline: 36270434]
72. Guastaferro K, Collins LM. Achieving the goals of translational science in public health intervention research: the multiphase optimization strategy (MOST). Am J Public Health 2019 Feb;109(S2):S128-S129. [doi: 10.2105/AJPH.2018.304874] [Medline: 30785800]
73. Guastaferro K, Collins LM. Optimization methods and implementation science: an opportunity for behavioral and biobehavioral interventions. Implement Res Pract 2021;2:26334895211054363 [FREE Full text] [doi: 10.1177/26334895211054363] [Medline: 37089990]
74. Hutchison AJ, Breckon JD. A review of telephone coaching services for people with long-term conditions. J Telemed Telecare 2011;17(8):451-458. [doi: 10.1258/jtt.2011.110513] [Medline: 22025743]
75. Kivelä K, Elo S, Kyngäs H, Kääriäinen M. The effects of health coaching on adult patients with chronic diseases: a systematic review. Patient Educ Couns 2014 Nov;97(2):147-157. [doi: 10.1016/j.pec.2014.07.026] [Medline: 25127667]
76. Lin WC, Chien HL, Willis G, O'Connell E, Rennie KS, Bottella HM, et al. The effect of a telephone-based health coaching disease management program on Medicaid members with chronic conditions. Med Care 2012 Jan;50(1):91-98. [doi: 10.1097/MLR.0b013e31822dcedf] [Medline: 21993059]
77. Simmons LA, Wolever RQ. Integrative health coaching and motivational interviewing: synergistic approaches to behavior change in healthcare. Glob Adv Health Med 2013 Jul;2(4):28-35 [FREE Full text] [doi: 10.7453/gahmj.2013.037] [Medline: 24416683]
78. Spaulding S. Simulation theory. In: The Routledge Handbook of Philosophy of Imagination. New York, NY: Routledge; 2016:262-273.

Abbreviations

C-SHIP: Cognitive Social Health Information Processing
COREQ: Consolidated Criteria for Reporting Qualitative Research
HIPAA: Health Insurance Portability and Accountability Act
HPV: human papillomavirus
mHealth: mobile health
OB-GYN: Obstetrics and gynecology
TC3: Tailored Communication for Cervical Cancer Risk

Edited by A Mavragani; submitted 30.11.23; peer-reviewed by S Xiong, SS Shaik, O Ezeigwe; comments to author 16.04.24; revised version received 08.06.24; accepted 04.09.24; published 14.01.25.

Please cite as:
Hemler JR, Wagner RB, Sullivan B, Macenat M, Tagai EK, Vega JL, Hernandez E, Miller SM, Wen KY, Ayers CA, Einstein MH, Hudson SV, Kohler RE
A Proposed mHealth Intervention to Address Patient Barriers to Colposcopy Attendance: Qualitative Interview Study of Clinic Staff and Patient Perspectives
JMIR Form Res 2025;9:e55043
URL: https://formative.jmir.org/2025/1/e55043
doi:10.2196/55043
PMID:39808485

XSL•FO
RenderX

©Jennifer R Hemler, Rachel B Wagner, Brittany Sullivan, Myneka Macenat, Erin K Tagai, Jazmarie L Vega, Enrique Hernandez, Suzanne M Miller, Kuang-Yi Wen, Charletta A Ayers, Mark H Einstein, Shawna V Hudson, Racquel E Kohler. Originally published in JMIR Formative Research (https://formative.jmir.org), 14.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Acceptance, Safety, and Effect Sizes in Online Dialectical Behavior Therapy for Borderline Personality Disorder: Interventional Pilot Study

Ruben Vonderlin[1,2], PhD; Tali Boritz[3,4], Prof Dr; Carola Claus[5], MSc; Büsra Senyüz[1,2], MSc; Saskia Mahalingam[1,2], MSc; Rachel Tennenhouse[4], BSc; Stefanie Lis[2,6], Prof Dr; Christian Schmahl[1,2], Prof Dr; Jürgen Margraf[5], Prof Dr; Tobias Teismann[5], Prof Dr; Nikolaus Kleindienst[1,2], Prof Dr; Shelley McMain[4,7], Prof Dr; Martin Bohus[1,2,5,8], Prof Dr

[1]Department of Psychosomatic Medicine and Psychotherapy, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany

[2]German Center for Mental Health (DZPG), Partner Site Mannheim - Heidelberg - Ulm, Mannheim, Germany

[3]Department of Psychology, York University, Toronto, ON, Canada

[4]Centre for Addiction and Mental Health, Borderline Personality Disorder Clinic, Toronto, ON, Canada

[5]Mental Health Research and Treatment Center, Ruhr-University Bochum, Bochum, Germany

[6]Department of Clinical Psychology, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany

[7]Department of Psychiatry, University of Toronto, Toronto, ON, Canada

[8]McLean Hospital, Harvard Medical School, Boston, MA, United States

Corresponding Author:
Ruben Vonderlin, PhD
Department of Psychosomatic Medicine and Psychotherapy
Central Institute of Mental Health
Medical Faculty Mannheim, Heidelberg University
J5
Mannheim, 68159
Germany
Phone: 49 621 1703 4445
Email: ruben.vonderlin@zi-mannheim.de

Abstract

Background: The potential of telehealth psychotherapy (ie, the online delivery of treatment via a video web-based platform) is gaining increased attention. However, there is skepticism about its acceptance, safety, and efficacy for patients with high emotional and behavioral dysregulation.

Objective: This study aims to provide initial effect size estimates of symptom change from pre- to post treatment, and the acceptance and safety of telehealth dialectical behavior therapy (DBT) for individuals diagnosed with borderline personality disorder (BPD).

Methods: A total of 39 individuals meeting the *DSM-5* (*Diagnostic and Statistical Manual of Mental Disorders* [Fifth Edition]) criteria for BPD received 1 year of outpatient telehealth DBT at 3 sites in Germany and Canada. Effect size estimates were assessed using pre-post measures of BPD symptoms, dissociation, and quality of life. Safety was evaluated by analyzing suicide attempts and self-harm. Additionally, acceptance and feasibility, satisfaction with treatment, useability of the telehealth format, and the quality of the therapeutic alliance were assessed from both therapists' and patients' perspectives. All analyses were conducted on both the intention-to-treat (ITT) and according-to-protocol (ATP) samples.

Results: Analyses showed significant and large pre-post effect sizes for BPD symptoms (d=1.13 in the ITT sample and d=1.44 in the ATP sample; P<.001) and for quality of life (d=0.65 in the ITT sample and d=1.24 in the ATP sample). Dissociative symptoms showed small to nonsignificant reductions. Self-harm behaviors decreased significantly from 80% to 28% of all patients showing at least 1 self-harm behavior in the last 10 weeks (risk ratio 0.35). A high dropout rate of 38% was observed. One low-lethality suicide attempt was reported. Acceptance, feasibility, and satisfaction measures were high, although therapists reported only moderate useability of the telehealth format.

XSL•FO
RenderX

Conclusions: Telehealth DBT for BPD showed large pre-post effect sizes for BPD symptoms and quality of life. While the telehealth format appeared feasible and well-accepted, the dropout rate was relatively high. Future research should compare the efficacy of telehealth DBT with in-person formats in randomized controlled trials. Overall, telehealth DBT might offer a potentially effective alternative treatment option, enhancing treatment accessibility. However, strategies for decreasing drop-out should be considered.

Trial Registration: German Clinical Trials Register DRKS00027824; https://drks.de/search/en/trial/DRKS00027824

(JMIR Form Res 2025;9:e66181) doi:10.2196/66181

KEYWORDS

dialectical behavior therapy; borderline personality disorder; online psychotherapy; virtual psychotherapy; telehealth; personality disorders; mental; psychotherapy; online; internet; telemedicine; psychiatry; psychiatric; acceptance

Introduction

Background

Borderline personality disorder (BPD) is a severe mental disorder characterized by high instability in affect, identity, and interpersonal relationships [1,2]. The prevalence of BPD is estimated to be between 0.7% and 2.7% in community samples [3,4] and between 11% and 35.6% in clinical samples [4]. BPD is associated with significant health care use and costs, as well as profound individual suffering and impaired quality of life [5-7].

According to international guidelines [8,9], psychotherapy is the first-line treatment for BPD. However, the complexity of symptoms such as high levels of suicidality, and interpersonal difficulties present challenges for psychological treatments [10,11]. Dialectical behavior therapy (DBT) [12-14] is a specialized treatment that was developed to support therapists in dealing with the complex symptom constellations that characterize BPD. DBT includes a dynamic hierarchy of treatment foci and a large set of therapeutic strategies and skills-based interventions to support individuals in building a life worth living. DBT is a multimodal psychological treatment that consists of individual and group therapy. A comprehensive DBT treatment includes both individual and group sessions that occur weekly over a 1-year period. DBT is an empirically supported treatment that has a robust research base demonstrating its efficacy in the treatment of BPD [15].

The Potential Benefit of Telehealth DBT for BPD

While there are effective psychological treatments for BPD, DBT and other evidence-based treatments are difficult to access [16,17]. One significant barrier to accessing DBT is that given the complexity of the treatment (ie, its multiple components, length, and team-based model) the treatment is often localized in specialized treatment centers in urban settings; DBT-trained therapists and comprehensive DBT treatment programs are often not available for clients in remote or rural settings. This problem is even more pronounced in developing countries or low- and middle-income countries [18].

One solution for improving access to DBT is to expand the delivery options of the treatment to include telehealth treatment (ie, the online delivery of treatment via a video web-based platform). Propelled in large part by the COVID-19 pandemic, DBT practitioners around the world have already begun to shift to the telehealth delivery of the treatment, with many clinicians in both health care centers and in private practice currently offering telehealth DBT [19-21]. However, concerns have been raised regarding the efficacy, safety, and feasibility of telehealth-delivered psychotherapy, especially for patients with high levels of emotional and behavioral dysregulation. Although research has started to explore the benefits, barriers, and subjective experiences of clients and therapists engaged in telehealth DBT in routine care, to date there are no rigorous research trials that have examined the efficacy, safety, acceptance, and feasibility of telehealth DBT [19-29].

Challenges of Telehealth Psychotherapy for BPD

Of particular concern is the limited ability of therapists to effectively deal with crisis situations and suicidal behavior in telehealth treatments [20,25,30]. Zalewski et al [28] conducted a qualitative study that shed light on various telehealth treatment challenges with patients with BPD, including patients missing sessions, turning off their cameras prematurely, logging off unexpectedly, or engaging in distracting behaviors like smoking or drinking alcohol during telehealth group therapy sessions. They also reported that some patients appeared dissociative during group sessions. Dissociation poses a particular challenge to the safety and efficacy of telehealth treatments for BPD. Previous research has underscored how dissociative experiences can lead to neurobiological alterations and impede emotional learning, thereby hindering successful psychotherapeutic outcomes [31-33]. Ideally, a therapist noticing dissociation during a session would intervene and provide guidance on how to use distress tolerance skills to decrease or interrupt dissociation. However, implementing these skills often involves strong sensory input or physical activity, which make be harder to facilitate telehealth versus in-person. More research is needed to investigate if and how dissociation and other therapy-interfering behaviors can be managed in telehealth treatments for BPD [34].

Another critical aspect of telehealth psychotherapy is the quality of the therapeutic alliance [20]. The interpersonal difficulties that are common in BPD, such as mistrust and rejection sensitivity, can make it harder to develop a strong alliance [35]. Previous DBT research suggests that a weaker therapeutic alliance is associated with higher therapy dropout rates [36] and that alliance rupture and repair processes in the early stage of treatment differ between recovered and unrecovered patients with BPD [37]. There is an ongoing debate as to whether the telehealth setting hampers the effective establishment of a

XSL•FO
RenderX

therapeutic alliance and increases the likelihood of alliance ruptures. While some studies show comparability between telehealth and in-person treatment formats [38,39], others suggest that telehealth therapies may be less conducive to developing an alliance than in-person therapies. Despite the pivotal role of the therapeutic alliance in DBT, at this point, there is no conclusive evidence as to how well a therapeutic alliance can be established and maintained in telehealth DBT.

Previous Research on Telehealth DBT

Two scoping reviews have been conducted to document the research evidence associated with the online delivery of DBT for individuals with BPD [24,40]. van Leeuwen et al [40] conducted their review to document the research evidence regarding the efficacy and clinical use of telepsychology in DBT. The studies they included in their review were heterogenous and ranged from the online delivery of treatment (eg, telehealth defined as videoconferencing for individual and group sessions, to the use of telephone coaching between sessions, to the integration of mobile apps as stand-alone or hybrid interventions, to the integration of virtual reality elements in treatment). Overall, they included 41 studies; however, only 2 of these trials investigated telehealth DBT applied by therapists via videoconferencing systems for patients with BPD specifically [41,42]. In a second review, Lakeman et al [24] similarly reviewed the online delivery of DBT for a range of different mental disorders. They showed that attendance tended to be higher in telehealth format and clinical improvements were comparable to in-person formats. However, many of the original trials also highlighted challenges related to risk management, therapist preparedness, and technical difficulties. The authors concluded that despite various challenges, the telehealth delivery of DBT programs is more accessible and feasible and equally acceptable, safe, and effective as the in-person delivery. However, of the 11 trials included in this review, only 1 trial was conducted with patients with BPD [43], therefore it remains unclear whether their conclusions can be generalized to BPD-specific populations.

While the aforementioned results suggest the telehealth delivery of DBT may be promising, a closer look into the primary studies included in the 2 reviews reveals a number of limitations that impact the conclusions we can draw from them. For example, Lopez et al [41] conducted a nonrandomized trial comparing attendance, client satisfaction, and group cohesion measures in a telehealth DBT skills group (n=20) compared to an in-person skills group (n=15). They found that attendance rates were significantly higher in the telehealth format (91% vs 75%), though there were no significant differences in satisfaction between the groups. Cohesion measures between group members and their therapists showed no differences, although cohesion between group members was significantly higher in the in-person group. Additionally, clients were not randomized but instead allowed to select their preferred group format. Finally, their study did not report on clinical outcomes related to borderline symptoms, which means we cannot compare the clinical effectiveness of the 2 treatment formats on this domain.

Salamin et al [42] analyzed diary card data of 7 patients with BPD within a treatment that had to be shifted from in-person to telehealth format due to the COVID-19 pandemic. In the first 8 weeks of treatment, patients received weekly in-person individual sessions as well as group sessions. Due to COVID-19 restrictions, treatment shifted to telehealth, and patients received weekly individual sessions by telephone or videoconferencing and group therapy was replaced with personal coaching. Weekly monitoring of problem behaviors showed a decrease in binge-eating behaviors and a trend toward a decrease in alcohol consumption following the shift to telehealth treatment. Results also indicated that while clients reported a decrease in feelings of shame or guilt, fear and tension following the move to a telehealth format, their general level of distress increased. Given the small sample size and lack of a control group, as well as the significant confound of the start of the COVID-19 pandemic that spurred the change in format to begin with, results from this study cannot reliably be attributed to the treatment or delivery format.

Alavi et al [43] investigated the effectiveness of an email-based DBT intervention compared to an in-person DBT skills group in 107 patients with BPD. While there were no significant differences between the treatment formats, these findings must be interpreted with caution: the DBT intervention consisted of only 15 sessions, and less than 50% of patients in either group completed treatment and were included in the analysis.

In summary, there is limited evidence regarding the feasibility and effectiveness of telehealth comprehensive DBT.

Objective of This Study

In this study, we aimed to investigate the acceptance, safety, and effect sizes of a 1-year comprehensive outpatient DBT program delivered telehealth for individuals with BPD within a controlled research setting. We were interested in the effect sizes for BPD symptoms and quality of life, and whether self-harm and dissociative symptoms could be targeted and effectively managed in a telehealth setting. To understand the acceptability and feasibility of telehealth DBT, we collected clients' and therapists' perspectives on the delivery of telehealth treatment. Finally, to determine whether a strong therapeutic relationship could be established in a telehealth treatment, we measured both therapists' and clients' ratings of the quality of the alliance throughout treatment.

Methods

Study Design

The data originate from a pilot psychotherapy trial comparing 2 different forms of DBT (standard DBT and trauma-focused DBT) within a telehealth setting. This pilot trial was designed to test the feasibility, safety, and acceptance of telehealth DBT and to prepare for a larger randomized controlled trial (RCT) of trauma-focused DBT. In order to ensure comparability with existing studies in the field, we chose to solely analyze the study arm that delivered telehealth standard (comprehensive) DBT, as developed by Marsha Linehan [12-14]. The study was conducted across 2 sites in Germany (Mannheim and Bochum) and 1 site in Canada (Toronto) between February 2022 and June 2023. The pilot study was registered in the German Clinical

XSL•FO
RenderX

Trials Register (DRKS00027824) and received approval from the respective ethical review boards at each participating site.

Intervention and Treatment

Participants received a 12-month outpatient standard DBT intervention, which comprised weekly individual therapy sessions (each lasting 50 minutes) and weekly group sessions (each lasting 100 minutes). All individual and group sessions were delivered via telehealth using videoconferencing platforms. The treatment was based on Marsha Linehan DBT treatment manual [13,14]. The group therapy focused on skill-building in mindfulness, distress tolerance, emotion regulation, and interpersonal effectiveness. Individual session agendas were informed by a review of clients' diary cards. All skills were taught twice (ie, 2 rounds of a 20-session curriculum). In individual therapy, treatment targets were prioritized based on the dynamic hierarchy focus of DBT that addresses life-threatening behavior, therapy-interfering behavior, and quality-of-life-interfering behavior, as needed. After completing the 12-month therapy period, clients were offered up to 5 booster sessions, which could be scheduled flexibly within the 6 months following treatment termination. Treatment was considered concluded once the booster sessions were finished. No specific adjustments were made to the treatment content to accommodate the telehealth format. However, patients completed their diary cards electronically via an app on their personal smartphones to facilitate diary card review. Participants who missed 4 consecutive weeks of individual therapy were categorized as treatment dropouts. Completion according to protocol (ATP) was defined as attending a minimum of 32 individual sessions, which equaled 80% of the expected dosage. All therapists (n=16) were trained in DBT with at least 2 days of theoretical DBT training completed at the beginning of the study (mean 4.7, SD 2.2 days), and at least half a year of clinical DBT experience (mean 3.6, SD 3.4 years). All therapists participated in a DBT consultation team to guide the treatment, which occurred on a weekly basis, consisting of 1.5-hour sessions per week.

Participants

Participants were recruited from outpatient clinic waiting lists in Mannheim, Bochum, and Toronto, as well as through advertisements and therapist referrals. Inclusion criteria were individuals aged 18-65 years with a diagnosis of BPD, and meeting at least 5 of 9 criteria outlined in the *DSM-5* (*Diagnostic and Statistical Manual of Mental Disorders* [Fifth Edition]). Additionally, participants needed to commit to 1 year of outpatient treatment and be willing to engage in telehealth therapy (with internet-enabled devices provided to participants, if needed). Proficiency in English (for the Toronto site) or German (for the Mannheim and Bochum sites) and written informed consent were further prerequisites for participation. Exclusion criteria represented symptoms that indicated a different treatment model or setting was likely needed, such as a lifetime diagnosis of schizophrenia, bipolar I disorder, intellectual disability (IQ<70), posttraumatic stress disorder (PTSD), a BMI below 17.5, severe substance use disorder according to *DSM-5* requiring medical withdrawal support, and dementia. Additionally, individuals with more than 8 weeks of DBT in the preceding year were excluded.

As depicted in the flowchart (Figure 1), 258 individuals underwent telephone screening for eligibility, with 126 undergoing full assessment. Ultimately, 81 (64%) individuals met the inclusion or exclusion criteria and were randomized to receive either standard DBT treatment (n=39; 48%) or a newly adapted DBT treatment (trauma-focused DBT; n=42, 52%). As this study only analyzed acceptance, safety, and effect sizes for standard telehealth DBT, our final sample comprised 39 participants enrolled in the 1-year telehealth standard DBT program. Among these, 15 (38%) individuals discontinued treatment and were classified as dropouts. One client ended the treatment as an early responder. Additionally, 12 (31%) participants did not provide assessments postintervention and were excluded from the analysis ATP. Consequently, 18 (46%) participants were included in the ATP analysis post intervention. The intent-to-treat (ITT) sample comprised all 39 participants who were randomly assigned to telehealth standard DBT.

XSL•FO
RenderX

Figure 1. The flowchart shows the progression of participants throughout the study. ATP: according to protocol; BPD: borderline personality disorder; DBT: dialectical behavior therapy; ITT: intention-to-treat.

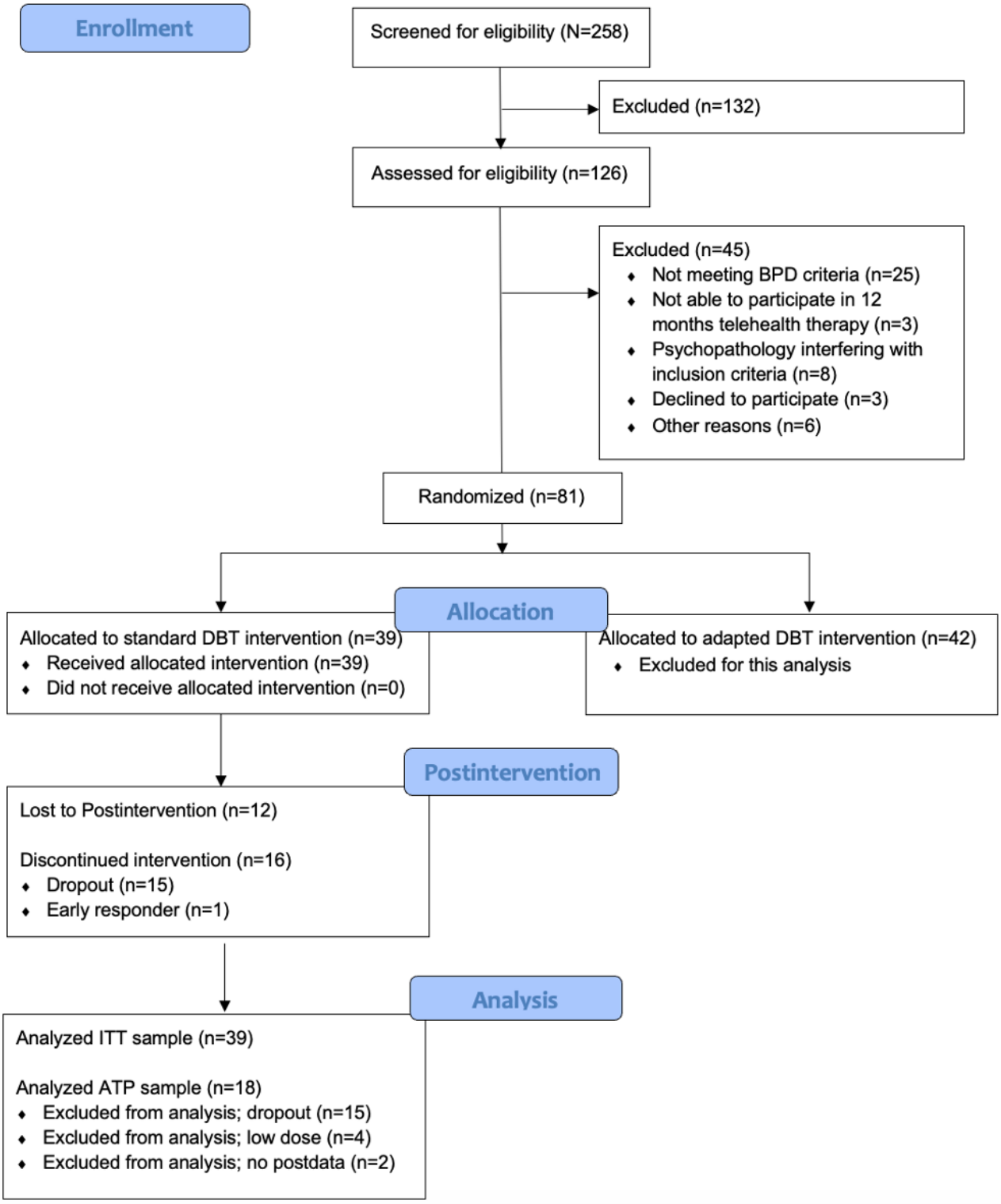

Measures

Primary Clinical Outcome

Borderline Symptom List-23

The Borderline Symptom List-23 (BSL-23) [44] was used to assess the borderline symptoms of participants throughout the trial. This self-report instrument comprises 23 items assessing borderline-typical symptoms on a 5-point Likert Scale, ranging from 0=not at all to 4=very strong. Widely used, the BSL-23 has demonstrated good psychometric properties [44,45]. In our sample, Cronbach α for the BSL-23 was 0.92. Assessment points included pre- and postintervention points, as well as every 3 months during the treatment period.

XSL•FO
RenderX

Secondary Clinical Outcomes

Deliberate Self-Harm Inventory

The Deliberate Self-Harm Inventory (DSHI) [46] is a self-report questionnaire designed to assess nonsuicidal self-injuries (NSSIs). Comprising 17 items, the DSHI captures data on the frequency, severity, duration, and type of self-harm behaviors. This instrument was administered both before and post intervention, as well as every 3 months during the treatment phase. In these analyses, we used the frequency score of the DSHI to determine the proportion of clients showing any NSSI in the last 10 weeks.

Dissociative Symptom Scale

The Dissociative Symptom Scale (DSS) [47] is a 20-item self-report questionnaire aimed at assessing dissociative symptoms. Participants are asked how often they have experienced dissociative symptoms over the last week ranging from 0=not at all to 4=more than once a day. It consists of 4 dimensions: depersonalization, derealization, gaps in awareness or memory, and dissociative reexperiencing. Since we were interested in the effect sizes on dissociative symptoms as a secondary outcome, we analyzed a global mean of the DSS across all dimensions. The DSS was administered before and post interventions as well as every 3 months during treatment.

Feasibility and Acceptability Outcomes

Acceptability of Intervention Measure

The Acceptability of Intervention Measure (AIM) [48] is a 15-item scale designed to assess therapists' perspectives on the acceptability, feasibility, and appropriateness of the delivered treatment. Widely used in implementation research, AIM has exhibited robust psychometric properties across various studies [48]. In our sample, intercorrelations between dimensions ranged from r=0.81 to 0.90. Therefore, we computed a mean score by combining the 3 dimensions for analysis. Administration occurred after 6 months of treatment and post intervention.

Client Satisfaction Questionnaire

The Client Satisfaction Questionnaire (CSQ-8) [49] is an 8-item scale that measures the client's satisfaction with the treatment and treatment provider on a 4-point Likert scale. The CSQ-8 was administered after 6 months of treatment as well as post intervention.

Usability of Technology

The unified theory of acceptance and use of technology (UTAUT) [50] assesses acceptance toward digital health care interventions based on the UTAUT framework. The questionnaire consists of 20 items assessing behavioral intention (4 items), performance expectancy (3 items), effort expectancy (3 items), social influence (2 items), facilitating conditions (2 items), internet anxiety (3 items), and experience with the internet (3 items) on a 5-point Likert scale ranging from 0=not at all to 4=very strong. We adapted items to also assess their usability throughout the treatment. The UTAUT was administered before and post interventions as well as at 6 months (midtreatment). Intercorrelations between dimensions ranged from r=0.80 to 0.91. Consequently, we computed a mean score to combine dimensions for analysis.

Severe Adverse Events

Severe adverse events (SAEs) were defined in accordance with the guidelines of good clinical practice. They encompassed suicide attempts or other events leading to death or posing an acute threat to life, where the study participant is in imminent danger of death at the time of the SAE. Therapists were queried weekly regarding the occurrence of potential SAEs. In the event of an SAE, therapists provided additional details to assess its intensity, outcome, and relationship to psychotherapeutic treatment. For safety purposes, all observed SAEs were promptly reported to the responsible ethics committee at the respective study site.

Working Alliance Inventory

The Working Alliance Inventory (WAI) [51] was used to assess the therapeutic alliance from both therapists' and clients' perspectives. The scale comprises 12 items assessing the therapeutic alliance on 3 dimensions: bonding, agreement on tasks, and agreement on goals. Therapists and their clients rate these items on a 7-point Likert scale ranging from 1=never to 7=always. Each therapist received separate questionnaires for each of their clients if they treated more than 1 client. The WAI was administered every 3 months during the treatment period. In our sample, intercorrelations between dimensions ranged from r=0.81 to 0.87 on the client level and from r=0.80 to 0.90 on the therapist level. Consequently, we computed a mean score to combine the 3 dimensions for analysis.

Statistical Analyses

To determine treatment effects, we analyzed BPD symptoms on the BSL-23 as the primary outcome. As secondary outcomes, we analyzed the quality of life, dissociation, and self-harm of participants throughout the treatment. Given the nonlinearity of treatment effects over time, we used a nonparametric Friedman test to assess differences among repeated measures. The Friedman test is robust and capable of handling nonnormal distributions and outliers. Effect sizes for pre-post comparisons for each outcome were analyzed using Cohen d, providing insights into the magnitude of treatment effects. We also calculated 95% CIs to test the robustness of effect size estimates. To compare the proportion of patients showing at least 1 NSSI in the last 10 weeks between pre- and posttreatments, we calculated risk ratio (RR) according to Altman [52] with the following equation, where a represented the number of patients with at least 1 NSSI at postintervention, b represented the number of patients without NSSI at postintervention, c represented the number of patients with at least 1 NSSI at preintervention, and d represented the number of patients without self-harm at preintervention.

To assess the safety of the telehealth trial, we examined SAEs (eg, serious and life-threatening suicide attempts) throughout the study, NSSI as well as individual symptom trajectories of clients on the BSL-23 to identify potential symptom deterioration throughout the treatment.

To evaluate the feasibility and acceptance of the telehealth trial, we conducted descriptive analyses of the client and therapist

XSL•FO
RenderX

feasibility and acceptance measures, as well as client and therapist ratings of the therapeutic alliance.

To analyze the intention-to-treat (ITT) sample, including all participants who had been randomized to the treatment, we imputed missing data by multiple imputation procedures [53]. Multiple imputation was based on the scale-level values of other assessment points and the dropout status (yes or no) to impute missing data. Since the missing pattern was not monotone, the Markov Chain Monte Carlo method [54] was used. Multiple imputation was based on the SAS (version 9.4; SAS Institute) multiple imputation procedures (1000 runs) and MIANALYZE. *P* values<.05 (2-tailed) were considered statistically significant.

Ethical Considerations

This study protocol was reviewed and approved by the applicable ethics committees at each study center, Medical Faculty Mannheim at Heidelberg University in Mannheim of the leading committee for all German sites including Ruhr University Bochum (2021-628-MA), and Center for Addiction and Mental Health, Toronto, Ontario, Canada (023-2023). A written informed consent was obtained from all participants in the study. At the start of the trial, all participants were assigned a randomly generated study code, ensuring that all data were processed and analyzed in a pseudonymized format. Participants did not receive financial compensation for data assessments conducted during the treatment. However, those who discontinued the treatment prematurely were compensated €20 (US $21) for each subsequent assessment time point. At the Canadian study site in Toronto, all clients received CAD $15 (US $10.5) per assessment hour.

Results

Participants

The mean age of participants was 31.1 (SD 10.6) years, with 76.9% (n=30) born female. On average, participants met 6.18 (SD 1.07) criteria for BPD according to *DSM-5*. Additionally, 74.4% of participants had a comorbid affective disorder, 53.8% had a comorbid anxiety disorder, and 23.1% (n=9) had a comorbid eating disorder. Psychotropic medication was received by 97.5% of clients. Table 1 presents the demographic characteristics of the sample.

XSL•FO
RenderX

Table 1. Demographic and clinical characteristics of participants.

Baseline characteristics	Full Sample	
	Values	Range
Age (years), mean (SD)	31.1 (10.6)	19-58
Sex (female), n (%)	30 (76.9)	N/A[a]
Highest educational level, n (%)		
No graduation	0 (0)	N/A
Lower secondary school (hauptschule)	6 (15.4)	N/A
Intermediate secondary school (realschule)	11 (28.2)	N/A
Higher secondary school (abitur)	13 (33.3)	N/A
University or postgraduate degree	9 (23)	N/A
Marital status, n (%)		
Single	26 (66.7)	N/A
Married or partnered	8 (20.5)	N/A
Divorced or widowed	5 (12.8)	N/A
Employment, n (%)		
Unemployed	5 (12.8)	N/A
Student	4 (10.3)	N/A
Employed	28 (71.8)	N/A
Other[b]	2 (5.1)	N/A
BPD[c] criteria (IPDE[d]), mean (SD)	6.18 (1.07)	5-8
Comorbidities[e], n (%)		
Affective disorders	29 (74.4)	N/A
Anxiety disorders	18 (46.2)	N/A
Eating disorders	9 (23.1)	N/A
Obsessive compulsive disorder	5 (12.8)	N/A
Psychotropic medications, n (%)		
Any psychotropic medication	31 (79.5)	N/A
Antidepressants	22 (56.4)	N/A
Neuroleptics	14 (35.9)	N/A
Mood stabilizers	1 (2.6)	N/A
Benzodiazepines	2 (5.1)	N/A
Others	16 (41)	N/A
1 and more suicide attempt (lifetime), n (%)	21 (54)	N/A
1 and more nonsuicidal self-injury in the last month, n (%)	31 (80)	N/A

[a]N/A: not applicable.

[b]n=1 identified as part-time student, n=1 reported current reintegration in the labor market.

[c]BPD: borderline personality disorder.

[d]IPDE: International Personality Disorder Examination.

[e]Comorbidities were categorized as follows: affective disorders (depressive disorder and bipolar II disorder); anxiety disorders (panic disorder, agoraphobia, social anxiety disorder and generalized anxiety disorder); and eating disorders (anorexia, bulimia, and binge eating).

Effect Size Estimations

The effect size estimations revealed a significant decrease in BPD symptoms (BSL-23), our primary outcome, with large effect sizes observed in both the ATP (χ^2_4=24.9; P<.001; d=1.44) and the ITT sample (χ^2_4=38.8; P<.001; d=1.04; Figure 2). As illustrated in Figure 3, the most substantial reduction in BPD symptoms occurred within the first 3 months of treatment,

XSL•FO
RenderX

with a further albeit smaller decrease observed throughout the remainder of the treatment period. Post–BSL-23 scores were a mean 0.91 (SD 0.60) in the ATP sample and a mean 1.20 (SD 0.65) in the ITT sample, indicating mild BPD symptoms in the ATP sample and moderate BPD symptoms in the ITT sample after treatment, respectively.

Figure 2. Course of borderline symptoms as the primary outcome in the ITT sample (n=39) as well as the ATP sample (n=18). Error bars represent 95% CIs using the bootstrapping method. ATP: according to protocol; BSL: Borderline Symptom List; ITT: intention-to-treat.

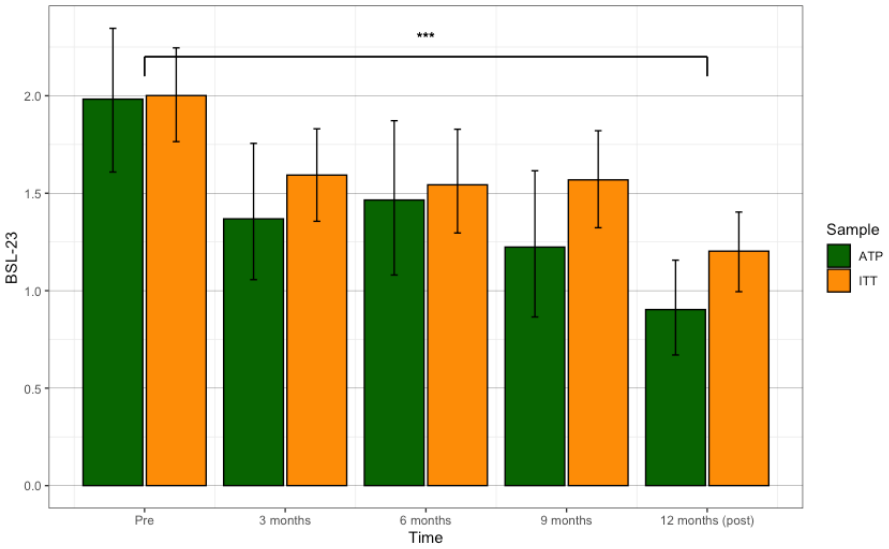

Figure 3. Individual data of participants' borderline symptoms over the course of the whole treatment. BSL: Borderline Symptom List-23.

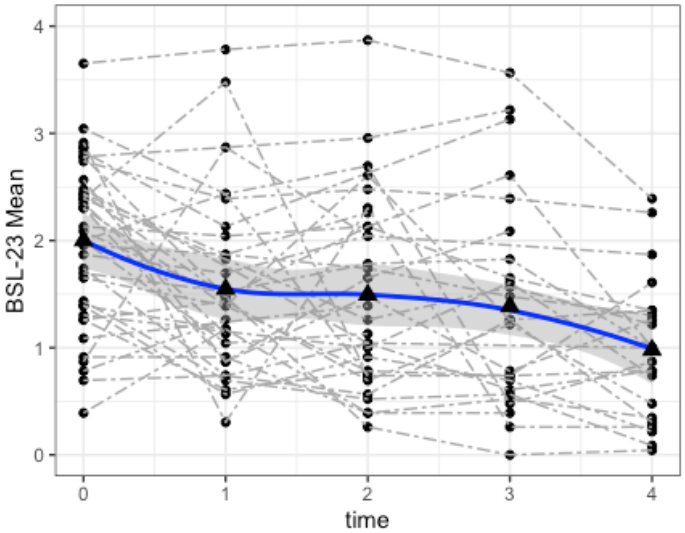

Dissociative symptoms exhibited a small, but significant decline in the ATP (χ^2_4=11.1; *P*=.02; *d*=0.37) and ITT samples (χ^2_4=12.3; *P*=.02; *d*=0.45) with clients reporting fewer dissociative symptoms throughout the treatment course. However, all 95% CIs of dissociation effect sizes included zero. Self-harm (DSHI scores) significantly decreased over the course of treatment, with 80% of clients engaging in at least 1 self-harm event in the last 10 weeks before treatment compared to 28% after treatment (χ^2_4=16.60; *P*<.001; *RR* 0.35).

The significant reduction in clinical symptoms was mirrored by an increase in quality of life, displaying a large effect size in the ATP (χ^2_4=14.5; *P*<.001; *d*=1.24) and a medium effect size in the ITT sample (χ^2_4= 31.0; *P*<.001; *d*=0.65). The pre-post changes of borderline symptoms and quality of life were significantly intercorrelated by *r*=–0.55 (t_{37}=–4.04; *P*<.001). Detailed results are presented in Table 2.

XSL•FO
RenderX

Table 2. Mean, SD, Friedman Test, and effect size statistics for study variables for ITT[a] and ATP[b] samples[c].

Variable	Baseline, mean (SD)	3 months, mean (SD)	6 months, mean (SD)	9 months, mean (SD)	12 months, mean (SD)	Friedman test		Effect size *d* (pre-post 95% CI)
						Chi-square (*df*)	*P* value	
Effectiveness measures								
ITT (*n*=39)								
Borderline symptoms (BSL-23[d])	2.00 (0.76)	1.59 (0.78)	1.54 (0.84)	1.57 (0.86)	1.20 (0.65)	38.8 (4)	<.001	1.13 (0.64-1.61)
Quality of life (ReQoL[e])	1.48 (0.60)	1.83 (0.55)	2.01 (0.74)	1.71 (0.67)	1.93 (0.76)	31.0 (4)	<.001	0.65 (1.12-0.19)
Dissociation (DSS[f])	0.91 (0.61)	0.81 (0.59)	0.80 (0.68)	0.79 (0.44)	0.65 (0.55)	12.3 (4)	.015	0.45 (–0.01 to 0.90)
Per Protocol (ATP; n=18)								
Borderline symptoms (BSL-23)	1.98 (0.85)	1.38 (0.80)	1.47 (0.96)	1.22 (0.87)	0.91 (0.60)	24.9 (4)	<.001	1.44 (0.70-2.17)
Quality of life (ReQoL)	1.55 (0.58)	1.93 (0.48)	2.15 (0.81)	2.05 (0.65)	2.33 (0.68)	14.5 (4)	<.001	1.24 (0.64-1.61)
Dissociation (DSS)	0.99 (0.67)	0.83 (0.64)	0.88 (0.86)	0.68 (0.50)	0.72 (0.75)	11.1 (4)	.023	0.37 (–0.29 to 1.04)
NSSI[g] (%; DSHI[h])	0.80 (0.41)	0.68 (0.48)	0.40 (0.50)	0.42 (0.50)	0.28 (0.46)	16.6 (4)	<.001	0.35 (RR[i])
Acceptance Measures								
Client measures								
Client satisfaction (1-4; CSQ[j])	—[k]	—	3.33 (0.54)	—	3.40 (0.74)	—	—	—
Useability of technology (0-4; UTAUT[l])	2.81 (0.59)	—	2.83 (0.72)	—	3.13 (0.67)	—	—	—
Therapeutic alliance (1-7; WAI[m])	—	5.52 (1.08)	5.89 (0.75)	5.65 (1.28)	—	—	—	—
Therapist measures								
Acceptance (1-5; AIM[n])	—	—	4.21 (0.74)	—	4.53 (0.61)	—	—	—
Useability of technology (0-4; UTAUT)	2.48 (0.84)	—	—	—	2.57 (0.71)	—	—	—
Therapeutic alliance (1-7; WAI)	—	5.24 (0.98)	5.30 (0.99)	5.40 (0.88)	—	—	—	—

[a]ITT: intention-to-treat.

[b]ATP: according-to-protocol.

[c]Effect Sizes for continuous data were calculated using Cohen *d*. Effect sizes for categorical data (ie, NSSI) were calculated using risk ratio estimates.

[d]BSL-23: Borderline Symptom List-23.

[e]ReQoL: recovering quality of life.

[f]DSS: Dissociative Symptom Scale.

[g]NSSI: nonsuicidal self-injury.

[h]DSHI: Deliberate Self-Harm Inventory.

[i]RR: risk ratio

[j]CSQ: Client Satisfaction Questionnaire.

[k]Not applicable because variable was not assessed at this time-point

[l]UTAUT: unified theory of acceptance and use of technology.

[m]WAI: Working Alliance Inventory.

[n]AIM: Acceptability of Intervention Measure.

XSL•FO
RenderX

Safety

During the course of treatment, 1 SAE was recorded. One patient was hospitalized following a suicide attempt involving an overdose in week 12 of treatment. The therapist rated this SAE as moderate and potentially linked to the psychotherapeutic treatment. In addition, individual patient trajectories in BPD symptoms were observed graphically for safety assessment, as illustrated in Figure 3. The majority of clients demonstrated a consistent decrease in BPD symptoms throughout the treatment duration.

Acceptance, Feasibility, and Usability of Telehealth Treatment

The results regarding the acceptance and feasibility of clients are summarized in Table 2 and depicted in Multimedia Appendix 1. Clients reported high levels of satisfaction with the treatment. 97% (29/30) of clients reported high satisfaction with telehealth DBT treatment midway through therapy (mean 3.33, SD 0.54 on the CSQ-8) and 87% (20/23) reported high satisfaction at the end of treatment (mean 3.40, SD 0.74). Prior to treatment, 69% (27/39) of clients anticipated that the telehealth format would be very or mostly usable for receiving DBT treatment (mean 2.81, SD 0.59 on the UTAUT). This percentage increased slightly during treatment to 77% (23/30) at midtherapy (mean 2.83, SD 0.72) and further to 83% (19/23) at the end of treatment (mean 3.13, SD 0.67).

The results regarding the acceptance and feasibility of therapists are summarized in Table 2 and depicted in Multimedia Appendix 2. From therapists' perspective, 86% (12/14) found the telehealth DBT treatment to be very acceptable, feasible, and appropriate for their clients (mean 4.21, SD 0.74 on the AIM). This percentage rose to 91% (10/11) after the completion of treatment (mean 4.53, SD 0.61). Prior to treatment, 53% (8/15) of therapists anticipated that the telehealth format would be very much or mostly usable for delivering DBT treatment (mean 2.48, SD 0.84 on the UTAUT). However, this percentage slightly decreased during treatment to 45% (5/11), with therapists rating the telehealth format as very or mostly usable for delivering DBT after treatment (mean 2.57, SD 0.71).

Therapeutic Alliance

The therapeutic alliance, as assessed by both clients and therapists, is summarized in Table 2 and depicted in Multimedia Appendix 1 for clients and Multimedia Appendix 2 for therapists. From the client perspective, 82% (28/34) reported experiencing a (very) high affective bond with their therapists and demonstrated (very) high agreement on treatment-related tasks and goals (mean 5.52, SD 1.08 on the WAI). This percentage increased to 96% (27/28) at midtherapy (mean 5.89, SD 0.75) before slightly decreasing to 84% at 9 months into treatment (mean 5.65, SD 1.28).

Similarly, therapist perspectives revealed that 80% (28/25) reported a (very) high affective bond with their clients and rated (very) high agreement on treatment-related tasks and goals (mean 5.24, SD 0.98 on the WAI). This percentage increased slightly to 87% (16/19) at midtherapy (mean 5.30, SD 0.99) before decreasing slightly to 84% (21/25) at 9 months into treatment (mean 5.40, SD 0.88).

A moderate yet significant intercorrelation was observed between therapist and client ratings, with r=0.52 (t_{78}=5.35; P<.001) across all subscales and all 3 assessment times.

Discussion

Primary Results

This pilot study aimed to provide initial insights into the acceptance, safety, and effect sizes of standard (comprehensive) DBT treatment delivered via telehealth (ie, the online delivery of treatment via a video web-based platform) within a controlled research setting. Our findings indicate large pre-post effect sizes on BPD symptoms (d=1.13-1.44) and quality of life (d=0.65-1.24) in both the ITT and ATP samples, despite a relatively high dropout rate of 38% (n=15). One SAE was recorded during the study, attributed to hospitalization following a suicide attempt. Client ratings reflected high levels of satisfaction and usability, with 97% (29/30) expressing high satisfaction with the treatment and 83% (19/23) reporting high usability of the telehealth format. Therapists also expressed strong acceptance of the treatment, with 91% (10/11) indicating high acceptance. Interestingly, therapists rated the usability of the telehealth format comparatively lower than clients, with only 53% (8/15) reporting high usability. The majority of clients (27/28, 96%) and therapists (16/19, 87%) were able to establish a strong therapeutic relationship through telehealth contact.

Comparison With Prior DBT Studies

The dropout rate of 38% (n=15) in this telehealth trial was higher than the average dropout rate of 28% reported in the meta-analysis on in-person DBT trials [55]. One possible explanation for this is the influence of the telehealth setting. Given both the heterogeneity in dropout rates across psychotherapy trials and the lack of an in-person control group in this study to contextualize this result, this finding should be interpreted with caution. Past research has yielded conflicting results regarding attendance and dropout rates in telehealth versus in-person treatments, with some studies indicating no differences between telehealth and in-person DBT, while others have suggested even higher attendance rates in the telehealth format [26,41]. Future studies should aim to directly compare dropout rates between telehealth and in-person DBT programs within RCTs as a broader framework is necessary to determine which clients are more likely to benefit from telehealth settings.

The estimation of effect sizes in this trial demonstrated a significant reduction in BPD symptoms, with large effect sizes of d=1.13 in the ITT sample and d=1.44 in the ATP sample throughout the treatment period. According to the interpretation guidelines for the BSL-23 proposed by Kleindienst et al [56], participants experienced a decrease in average borderline symptom severity from high at baseline to moderate at post-intervention in the ITT sample, and mild in the ATP sample, respectively. Comparing these outcomes with previous in-person DBT trials using a similar treatment dosage, our results suggest that clinical effectiveness in our trial is comparable or even slightly higher. For instance, McMain et al [57] reported a pre-post reduction of d=0.67. Similarly, Priebe et al [58] observed a significant decrease in BPD symptoms, as

XSL•FO
RenderX

assessed by the Zanarini Rating Scale, with an effect size of *d*=0.71 from pre- to postintervention in a 12-month in-person DBT treatment. McMain et al [59] found that a 12-month in-person DBT treatment yielded a pre-post effect size of *d*=1.23 on the Zanarini Rating Scale. Sinnaeve et al [60] demonstrated a significant decrease in the Borderline Personality Disorder Symptom Index [61], with an effect size of *d*=1.15 from pre- to post-intervention in a 12-month outpatient in-person DBT treatment.

Of note, in our study, dissociative symptoms showed only small effect sizes of *d*=0.37 in the ATP and *d*=0.45 in the ITT sample with the 95% CIs containing 0. This finding may be due, in part, to the relatively low baseline dissociative scores of mean 0.91 (SD 0.61) and mean 0.99 (SD 0.67) likely related to our exclusion of patients with BPD with comorbid PTSD, who typically present higher dissociative symptoms compared to those without PTSD [62]. Our dissociation findings therefore continue the ongoing debate on the effectiveness of DBT in reducing dissociative experiences. According to the current Cochrane Review, DBT has a medium effect size for dissociation (*d*=0.43) across primary studies compared to waitlist and treatment-as-usual control groups [15]. Moreover, research suggests that greater effects on dissociation can be achieved by adapting a trauma-focused approach within DBT and explicitly teaching clients antidissociative skills, which were not included in our standard DBT protocol [63,64]. Consequently, our study does not allow us to conclude whether and how dissociative symptoms can be effectively addressed through the telehealth format. Future research should address this question by directly comparing the management of dissociative symptoms in telehealth versus in-person settings.

Self-harm and suicidal behavior in our study significantly decreased with a large reduction of participants' NSSI of RR 0.35 in the ATP sample. While 80% of participants reported engaging in at least 1 self-harm episode in the months preceding treatment initiation, this number decreased significantly to 28% postintervention. This finding aligns with prior research indicating the efficacy of DBT in reducing NSSI behaviors by empowering individuals to attain behavioral control [60,65,66]. Importantly, our study suggests that the reduction of self-harm can be also effectively achieved in a telehealth format. Notably, there was 1 SAE, a suicide attempt by overdose, during the treatment period, with no recorded completed suicides. In previous in-person DBT trials, the prevalence of suicide attempts requiring medical attention (eg, emergency department visits) over a 12-month treatment period has ranged from 3.5% to 6.7% [57,67,68]. Based on these prevalence rates, we would expect 1 to 3 suicide attempts in a cohort of 39 patients. Thus, the single suicide attempt recorded in our study aligns with this expected range and does not suggest an increased occurrence of suicide attempts in telehealth treatment settings.

In terms of the therapeutic alliance, our study showed that over 80% of clients and therapists reported a very strong alliance. Mean scores for the therapeutic alliance ranged from mean 5.24 (SD 0.98) to mean 5.89 (SD 0.75) on a 7-point Likert scale of the WAI. These scores are comparable to, or even higher than, those reported in studies administering the same questionnaire in in-person DBT trials [36,69]. However, this finding may be biased as participants who dropped out of treatment also discontinued completing the alliance measure, potentially inflating alliance ratings over the course of treatment. Nonetheless, our results indicate that a strong therapeutic alliance can be achieved in telehealth DBT settings. This finding challenges the argument that telehealth treatment leads to heightened disengagement or weak therapeutic alliances between clients and therapists [20,70].

Satisfaction and acceptance ratings in our sample were high, with 91% of therapists rating the acceptance of the treatment as high and 98% of participants indicating satisfaction with the treatment. These findings align with previous research demonstrating high acceptance and satisfaction rates of telehealth psychotherapy for patients with personality disorders [71]. From the participants' perspective, useability ratings of the telehealth format for delivering DBT were also high, with 69%-83% indicating high usability. Interestingly, participants' useability ratings significantly increased throughout the treatment, potentially indicating that those with initially low useability ratings may have dropped out of treatment. Unfortunately, our sample size was insufficient to robustly analyze useability as a predictor of therapy dropout during telehealth treatment. Future studies should explore potential predictors for treatment success related to the usability of the telehealth format. In contrast, therapists' perception of telehealth format usability was notably lower, with only half of the therapists (45%-53%) indicating high usability for delivering psychotherapy via telehealth. This discrepancy may be attributed to the greater adjustments required by therapists, transitioning from in-person to telehealth psychotherapy delivery. This finding is consistent with prior research identifying various barriers, challenges, and concerns among therapists delivering DBT via telehealth [20]. Thus, future research should explore strategies to enhance therapists' usability experiences during telehealth treatment. We would also argue that telehealth psychotherapy need not replace in-person psychotherapy, but rather can be used as a complement for clients in need. Both modalities offer valuable options for delivering effective and accessible treatment. Future studies should consider how to tailor treatment delivery format to individual preferences and circumstances.

Limitations

There are some limitations that should be considered when interpreting the study's results. First, the exclusion of patients with comorbid PTSD limits the representativeness of the studied population, particularly its applicability and generalizability to all patients with BPD. Therefore, we caution against generalizing these results to patients with BPD and comorbid PTSD. Second, the absence of a control group directly comparing telehealth and in-person conditions within the same trial and cohort impedes drawing inferential conclusions regarding the relative efficacy of telehealth versus in-person DBT. Lastly, the small sample size of 39 clients did not allow us to analyze potential moderator effects on treatment outcomes or dropout rates. Given the limitations of this pilot and feasibility study, future research should include RCTs with larger sample sizes to directly compare telehealth and in-person DBT treatments.

XSL•FO
RenderX

Conclusions

Our findings provide a first indication that telehealth DBT for BPD appears clinically efficacious, safe, feasible, and well accepted. However, the dropout rate was higher compared to other studies investigating in-person delivered DBT. Future research should compare the efficacy of telehealth DBT with in-person formats in RCTs. Overall, telehealth-delivered DBT offers a potentially effective alternative to in-person DBT, enhancing treatment accessibility and overcoming barriers to treatment engagement. This knowledge may be particularly useful for increasing access to treatment for those in rural areas or low- and middle-income countries where traveling distances to specialized DBT practitioners or treatment programs may be a large barrier.

Acknowledgments

The authors acknowledge the contribution and support of the staff at the study coordinating center and at the participating study sites, Leonie Staufner, Thirza Meyer, Hannah Stiefelhagen, and Eva-Maria Bütow from the Department of Psychosomatic Medicine and Psychotherapy, Central Institute of Mental Health Mannheim, Medical Faculty Mannheim, Heidelberg University; Jasmine Shirvani, Maya Mehrotra, and Askshayan Vimalanathan from the Center for Addiction and Mental Health, Toronto, Ontario, Canada. The authors thank the study therapists, as well as the supervisors, Hans Gunia and Angela Oermann.

Data Availability

All data generated or analyzed during this study are included in this article. Further inquiries can be directed to the corresponding author and will be made available on request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Acceptance, satisfaction, useability, and therapeutic alliance measures of clients during telehealth DBT. DBT: dialectical behavior therapy.
[PNG File , 307 KB - formative_v9i1e66181_app1.png]

Multimedia Appendix 2
Acceptance, satisfaction, useability, and therapeutic alliance measures of therapists during telehealth DBT. DBT: dialectical behavior therapy.
[PNG File , 229 KB - formative_v9i1e66181_app2.png]

References

1. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders. 5th edition. Washington, DC: American Psychiatric Association; 2013.
2. Bohus M, Stoffers-Winterling J, Sharp C, Krause-Utz A, Schmahl C, Lieb K. Borderline personality disorder. Lancet 2021;398(10310):1528-1540. [doi: 10.1016/s0140-6736(21)00476-1]
3. Eaton NR, Greene AL. Personality disorders: community prevalence and socio-demographic correlates. Curr Opin Psychol 2018;21:28-32. [doi: 10.1016/j.copsyc.2017.09.001] [Medline: 28961462]
4. Ellison WD, Rosenstein LK, Morgan TA, Zimmerman M. Community and clinical epidemiology of borderline personality disorder. Psychiatr Clin North Am 2018;41(4):561-573. [doi: 10.1016/j.psc.2018.07.008] [Medline: 30447724]
5. Bode K, Vogel R, Walker J, Kröger C. Health care costs of borderline personality disorder and matched controls with major depressive disorder: a comparative study based on anonymized claims data. Eur J Health Econ 2017;18(9):1125-1135. [doi: 10.1007/s10198-016-0858-2] [Medline: 27990594]
6. Soeteman DI, Verheul R, Busschbach JJ. The burden of disease in personality disorders: diagnosis-specific quality of life. J Pers Disord 2008;22(3):259-268. [doi: 10.1521/pedi.2008.22.3.259] [Medline: 18540798]
7. van Asselt A, Dirksen CD, Arntz A, Severens JL. The cost of borderline personality disorder: societal cost of illness in BPD-patients. Eur Psychiatry 2007;22(6):354-361. [doi: 10.1016/j.eurpsy.2007.04.001] [Medline: 17544636]
8. Borderline personality disorder: treatment and management. National Institute for Health and Care Excellence. 2009. URL: https://www.nice.org.uk/guidance/cg78 [accessed 2024-12-18]
9. [S3 guideline Borderline personality disorder]. Deutsche Gesellschaft für Psychiatrie und Psychotherapie, Psychosomatik und Nervenkeilkunde. URL: https://register.awmf.org/assets/guidelines/038-015l_S3_Borderline-Persoenlichkeitsstörungen_2022-11.pdf [accessed 2024-12-10]
10. Mehlum L. Clinical challenges in the assessment and management of suicidal behaviour in patients with bordeline personality disorder. Epidemiol Psichiatr Soc 2009;18(3):184-190. [doi: 10.1017/s1121189x00000440] [Medline: 20034194]

XSL•FO
RenderX

11. Pompili M, Girardi P, Ruberto A, Tatarelli R. Suicide in borderline personality disorder: a meta-analysis. Nord J Psychiatry 2005;59(5):319-324. [doi: 10.1080/08039480500320025] [Medline: 16757458]
12. Linehan MM, Kehrer CA. Borderline personality disorder. In: Barlow DH, editor. Clinical Handbook of Psychological Disorders: A Step-by-Step Treatment Manual. New York, NY: Guilford Press; 1999.
13. Linehan M. Skills Training Manual for Treating Borderline Personality Disorder. New York, NY: Guilford Press; 1993.
14. Linehan M. Cognitive-Behavioral Treatment of Borderline Personality Disorder. New York, NY: Guilford Press; 1993.
15. Storebø OJ, Stoffers-Winterling JM, Völlm BA, Kongerslev MT, Mattivi JT, Jørgensen MS, et al. Psychological therapies for people with borderline personality disorder. Cochrane Database Syst Rev 2020;5(5):CD012955 [FREE Full text] [doi: 10.1002/14651858.CD012955.pub2] [Medline: 32368793]
16. Lohman MC, Whiteman KL, Yeomans FE, Cherico SA, Christ WR. Qualitative analysis of resources and barriers related to treatment of borderline personality disorder in the United States. Psychiatr Serv 2017;68(2):167-172 [FREE Full text] [doi: 10.1176/appi.ps.201600108] [Medline: 27691382]
17. Spina DS, Levy KN. The availability of dialectical behavior therapy in partial hospitalization and residential services for borderline personality disorder: an exploratory longitudinal study of the National Mental Health Services Survey from 2014 to 2021. J Consult Clin Psychol 2024;92(3):176-186. [doi: 10.1037/ccp0000870] [Medline: 38059945]
18. Malhotra S, Chakrabarti S, Shah R. Telepsychiatry: promise, potential, and challenges. Indian J Psychiatry 2013;55(1):3-11 [FREE Full text] [doi: 10.4103/0019-5545.105499] [Medline: 23441027]
19. Bock MM, Graf T, Woeber V, Kothgassner OD, Buerger A, Plener PL. Radical acceptance of reality: putting DBT®-A skill groups online during the COVID-19 pandemic: a qualitative study. Front Psychiatry 2022;13:617941 [FREE Full text] [doi: 10.3389/fpsyt.2022.617941] [Medline: 35546945]
20. Cooney EB, Walton CJ, Gonzalez S. Getting DBT online down under: the experience of Australian and New Zealand dialectical behaviour therapy programmes during the Covid-19 pandemic. PLoS One 2022;17(10):e0275636 [FREE Full text] [doi: 10.1371/journal.pone.0275636] [Medline: 36201507]
21. Hood P, Turner CJ, Beggs B, Owens L, Chapman AL. Making lemonade out of lemons: dialectical behavior therapy via telehealth during a pandemic. Behav Ther 2023;54(5):876-891 [FREE Full text] [doi: 10.1016/j.beth.2023.03.007] [Medline: 37597964]
22. Conroy K, Kehrer SM, Georgiadis C, Hare M, Ringle VM, Shaw AM. Learning from adolescents and caregivers to enhance acceptability and engagement within virtual dialectical behavior therapy for adolescents skills groups: a qualitative study. Child Psychiatry Hum Dev 2024:1-15. [doi: 10.1007/s10578-023-01641-7] [Medline: 38183575]
23. Dunn N, Walton CJ, Matsunaga E, Williams C, Dimeff LA. Acceptability of telehealth for multidiagnostic suicidal patients in a real-world dialectical behavior therapy clinic during the COVID-19 pandemic. Telemed J E Health 2023;29(4):593-601. [doi: 10.1089/tmj.2022.0205] [Medline: 35984847]
24. Lakeman R, King P, Hurley J, Tranter R, Leggett A, Campbell K, et al. Towards online delivery of dialectical behaviour therapy: a scoping review. Int J Ment Health Nurs 2022;31(4):843-856 [FREE Full text] [doi: 10.1111/inm.12976] [Medline: 35048482]
25. Stetsiv K, Rebmann K, Wilks CR. Fears and challenges of dialectical behavior therapists using telehealth. J Behav Cogn Ther 2023;33(1):57-64. [doi: 10.1016/j.jbct.2023.02.001]
26. Walton CJ, Gonzalez S, Cooney EB, Leigh L, Szwec S. Engagement over telehealth: comparing attendance between dialectical behaviour therapy delivered face-to-face and via telehealth for programs in Australia and New Zealand during the Covid-19 pandemic. Borderline Personal Disord Emot Dysregul 2023;10(1):16 [FREE Full text] [doi: 10.1186/s40479-023-00221-4] [Medline: 37208784]
27. Wilks CR, Gurtovenko K. Telehealth delivery of dialectical behavior therapy. In: Weinberg H, Rolnick A, Leighton A, editors. Advances in Online Therapy: Emergence of a New Paradigm. New York, NY: Routledge; 2022.
28. Zalewski M, Walton CJ, Rizvi SL, White AW, Gamache Martin C, O'Brien JR, et al. Lessons learned conducting dialectical behavior therapy via telehealth in the age of COVID-19. Cogn Behav Pract 2021;28(4):573-587 [FREE Full text] [doi: 10.1016/j.cbpra.2021.02.005] [Medline: 34629837]
29. Hyland KA, McDonald JB, Verzijl CL, Faraci DC, Calixte-Civil PF, Gorey CM, et al. Telehealth for dialectical behavioral therapy: a commentary on the experience of a rapid transition to virtual delivery of DBT. Cogn Behav Pract 2022;29(2):367-380. [doi: 10.1016/j.cbpra.2021.02.006]
30. Stoll J, Müller JA, Trachsel M. Ethical issues in online psychotherapy: a narrative review. Front Psychiatry 2019;10:993 [FREE Full text] [doi: 10.3389/fpsyt.2019.00993] [Medline: 32116819]
31. Kleindienst N, Limberger MF, Ebner-Priemer UW, Keibel-Mauchnik J, Dyer A, Berger M, et al. Dissociation predicts poor response to dialectial behavioral therapy in female patients with borderline personality disorder. J Pers Disord 2011;25(4):432-447. [doi: 10.1521/pedi.2011.25.4.432] [Medline: 21838560]
32. Kleindienst N, Priebe K, Görg N, Dyer A, Steil R, Lyssenko L, et al. State dissociation moderates response to dialectical behavior therapy for posttraumatic stress disorder in women with and without borderline personality disorder. Eur J Psychotraumatol 2016;7(1):30375 [FREE Full text] [doi: 10.3402/ejpt.v7.30375] [Medline: 27396380]

XSL•FO
RenderX

33. Seitz KI, Sicorello M, Schmitz M, Valencia N, Herpertz SC, Bertsch K, et al. Childhood maltreatment and amygdala response to interpersonal threat in a transdiagnostic adult sample: the role of trait dissociation. Biol Psychiatry Cogn Neurosci Neuroimaging 2024;9(6):626-634. [doi: 10.1016/j.bpsc.2024.01.003] [Medline: 38280631]
34. Lyssenko L, Schmahl C, Bockhacker L, Vonderlin R, Bohus M, Kleindienst N. Dissociation in psychiatric disorders: a meta-analysis of studies using the dissociative experiences scale. Am J Psychiatry 2018;175(1):37-46. [doi: 10.1176/appi.ajp.2017.17010025] [Medline: 28946763]
35. Miano A, Barnow S, Wagner S, Roepke S, Dziobek I. Dyadic emotion regulation in women with borderline personality disorder. Cogn Ther Res 2021;45(6):1077-1092. [doi: 10.1007/s10608-021-10206-8]
36. Steuwe C, Berg M, Driessen M, Beblo T. Therapeutic alliance and dropout in patients with borderline pathology receiving residential dialectical behavior therapy. BMC Psychiatry 2023;23(1):605 [FREE Full text] [doi: 10.1186/s12888-023-05061-8] [Medline: 37596568]
37. Boritz T, Barnhart R, Eubanks CF, McMain S. Alliance rupture and resolution in dialectical behavior therapy for borderline personality disorder. J Pers Disord 2018;32:115-128. [doi: 10.1521/pedi.2018.32.supp.115] [Medline: 29388901]
38. Flückiger C, Del Re AC, Wampold BE, Horvath AO. The alliance in adult psychotherapy: a meta-analytic synthesis. Psychotherapy 2018;55(4):316-340 [FREE Full text] [doi: 10.1037/pst0000172] [Medline: 29792475]
39. Kaiser J, Hanschmidt F, Kersting A. The association between therapeutic alliance and outcome in internet-based psychological interventions: a meta-analysis. Comput Hum Behav 2021;114:106512. [doi: 10.1016/j.chb.2020.106512]
40. van Leeuwen H, Sinnaeve R, Witteveen U, Van Daele T, Ossewaarde L, Egger JIM, et al. Reviewing the availability, efficacy and clinical utility of telepsychology in dialectical behavior therapy (Tele-DBT). Borderline Personal Disord Emot Dysregul 2021;8(1):26 [FREE Full text] [doi: 10.1186/s40479-021-00165-7] [Medline: 34717772]
41. Lopez A, Rothberg B, Reaser E, Schwenk S, Griffin R. Therapeutic groups via video teleconferencing and the impact on group cohesion. Mhealth 2020;6:13 [FREE Full text] [doi: 10.21037/mhealth.2019.11.04] [Medline: 32270005]
42. Salamin V, Rossier V, Joye D, Nolde C, Pierrehumbert T, Gothuey I, et al. [Adaptations of an outpatient dialectical behavioral therapy during the COVID-19 pandemic and consequences of the confinement on patients with borderline personality disorder]. Ann Med Psychol 2021;179(2):131-136 [FREE Full text] [doi: 10.1016/j.amp.2020.08.006] [Medline: 32843771]
43. Alavi N, Stephenson C, Rivera M. Effectiveness of delivering dialectical behavioral therapy techniques by email in patients with borderline personality disorder: nonrandomized controlled trial. JMIR Ment Health 2021;8(4):e27308 [FREE Full text] [doi: 10.2196/27308] [Medline: 33835936]
44. Wolf M, Limberger MF, Kleindienst N, Stieglitz R, Domsalla M, Philipsen A, et al. [Short version of the borderline symptom list (BSL-23): development and psychometric evaluation]. Psychother Psychosom Med Psychol 2009;59(8):321-324. [doi: 10.1055/s-0028-1104598] [Medline: 19274605]
45. Bohus M, Kleindienst N, Limberger MF, Stieglitz R, Domsalla M, Chapman AL, et al. The short version of the borderline symptom List (BSL-23): development and initial data on psychometric properties. Psychopathology 2009;42(1):32-39. [doi: 10.1159/000173701] [Medline: 19023232]
46. Gratz K. Measurement of deliberate self-harm: preliminary data on the deliberate self-harm inventory. J Psychopathol Behav Assess 2001;23:253-263. [doi: 10.1037/t04163-000]
47. Carlson EB, Waelde LC, Palmieri PA, Macia KS, Smith SR, McDade-Montez E. Development and validation of the Dissociative Symptoms Scale. Assessment 2018;25(1):84-98. [doi: 10.1177/1073191116645904] [Medline: 27178761]
48. Weiner BJ, Lewis CC, Stanick C, Powell BJ, Dorsey CN, Clary AS, et al. Psychometric assessment of three newly developed implementation outcome measures. Implement Sci 2017;12(1):108 [FREE Full text] [doi: 10.1186/s13012-017-0635-3] [Medline: 28851459]
49. Attkisson CC, Greenfield TK. The Client Satisfaction Questionnaire (CSQ) scales and the Service Satisfaction Scale-30 (SSS-30). In: Sederer LL, Dickey B, editors. Outcomes Assessment in Clinical Practice. Baltimore, MD: Williams & Wilkins; 1996:120-127.
50. Philippi P, Baumeister H, Apolinário-Hagen J, Ebert DD, Hennemann S, Kott L, et al. Acceptance towards digital health interventions—model validation and further development of the unified theory of acceptance and use of technology. Internet Interv 2021;26:100459 [FREE Full text] [doi: 10.1016/j.invent.2021.100459] [Medline: 34603973]
51. Horvath AO, Greenberg LS. Development and validation of the Working Alliance Inventory. J Counsel Psychol 1989;36(2):223-233. [doi: 10.1037//0022-0167.36.2.223]
52. Altman DG. Practical Statistics for Medical Research. New York, NY: Chapman and Hall/CRC; 1990.
53. Rubin DB. Multiple imputation after 18+ years. J Am Stat Assoc 1996;91(434):473-489. [doi: 10.1080/01621459.1996.10476908]
54. Schafer JL. Analysis of Incomplete Multivariate Data. New York, NY: Chapman and Hall/CRC; 1997.
55. Dixon LJ, Linardon J. A systematic review and meta-analysis of dropout rates from dialectical behaviour therapy in randomized controlled trials. Cogn Behav Ther 2020;49(3):181-196. [doi: 10.1080/16506073.2019.1620324] [Medline: 31204902]
56. Kleindienst N, Jungkunz M, Bohus M. A proposed severity classification of borderline symptoms using the Borderline Symptom List (BSL-23). Borderline Personal Disord Emot Dysregul 2020;7(1):11 [FREE Full text] [doi: 10.1186/s40479-020-00126-6] [Medline: 32514359]

XSL•FO
RenderX

57. McMain SF, Chapman AL, Kuo JR, Dixon-Gordon KL, Guimond TH, Labrish C, et al. The effectiveness of 6 versus 12 months of dialectical behavior therapy for borderline personality disorder: a noninferiority randomized clinical trial. Psychother Psychosom 2022;91(6):382-397 [FREE Full text] [doi: 10.1159/000525102] [Medline: 35738244]
58. Priebe S, Bhatti N, Barnicot K, Bremner S, Gaglia A, Katsakou C, et al. Effectiveness and cost-effectiveness of dialectical behaviour therapy for self-harming patients with personality disorder: a pragmatic randomised controlled trial. Psychother Psychosom 2012;81(6):356-365 [FREE Full text] [doi: 10.1159/000338897] [Medline: 22964561]
59. McMain SF, Links PS, Gnam WH, Guimond T, Cardish RJ, Korman L, et al. A randomized trial of dialectical behavior therapy versus general psychiatric management for borderline personality disorder. Am J Psychiatry 2009;166(12):1365-1374. [doi: 10.1176/appi.ajp.2009.09010039] [Medline: 19755574]
60. Sinnaeve R, van den Bosch LM, Hakkaart-van Roijen L, Vansteelandt K. Effectiveness of step-down versus outpatient dialectical behaviour therapy for patients with severe levels of borderline personality disorder: a pragmatic randomized controlled trial. Borderline Personal Disord Emot Dysregul 2018;5(1):12 [FREE Full text] [doi: 10.1186/s40479-018-0089-5] [Medline: 30002832]
61. Arntz A, van den Hoorn M, Cornelis J, Verheul R, van den Bosch WM, de Bie AJ. Reliability and validity of the borderline personality disorder severity index. J Pers Disord 2003;17(1):45-59. [doi: 10.1521/pedi.17.1.45.24053] [Medline: 12659546]
62. Cackowski S, Neubauer T, Kleindienst N. The impact of posttraumatic stress disorder on the symptomatology of borderline personality disorder. Borderline Personal Disord Emot Dysregul 2016;3(1):1-7 [FREE Full text] [doi: 10.1186/s40479-016-0042-4] [Medline: 27482383]
63. Harned MS, Korslund KE, Linehan MM. A pilot randomized controlled trial of dialectical behavior therapy with and without the dialectical behavior therapy prolonged exposure protocol for suicidal and self-injuring women with borderline personality disorder and PTSD. Behav Res Ther 2014;55:7-17 [FREE Full text] [doi: 10.1016/j.brat.2014.01.008] [Medline: 24562087]
64. Bohus M, Kleindienst N, Hahn C, Müller-Engelmann M, Ludäscher P, Steil R, et al. Dialectical behavior therapy for posttraumatic stress disorder (DBT-PTSD) compared with cognitive processing therapy (CPT) in complex presentations of PTSD in women survivors of childhood abuse: a randomized clinical trial. JAMA Psychiatry 2020;77(12):1235-1245 [FREE Full text] [doi: 10.1001/jamapsychiatry.2020.2148] [Medline: 32697288]
65. Carter GL, Willcox CH, Lewin TJ, Conrad AM, Bendit N. Hunter DBT project: randomized controlled trial of dialectical behaviour therapy in women with borderline personality disorder. Aust N Z J Psychiatry 2010;44(2):162-173. [doi: 10.3109/00048670903393621] [Medline: 20113305]
66. van den Bosch LM, Koeter MW, Stijnen T, Verheul R, van den Brink W. Sustained efficacy of dialectical behaviour therapy for borderline personality disorder. Behav Res Ther 2005;43(9):1231-1241. [doi: 10.1016/j.brat.2004.09.008] [Medline: 16005708]
67. Linehan MM, Comtois KA, Murray AM, Brown MZ, Gallop RJ, Heard HL, et al. Two-year randomized controlled trial and follow-up of dialectical behavior therapy vs therapy by experts for suicidal behaviors and borderline personality disorder. Arch Gen Psychiatry 2006;63(7):757-766. [doi: 10.1001/archpsyc.63.7.757] [Medline: 16818865]
68. Linehan MM, Korslund KE, Harned MS, Gallop RJ, Lungu A, Neacsiu AD, et al. Dialectical behavior therapy for high suicide risk in individuals with borderline personality disorder: a randomized clinical trial and component analysis. JAMA Psychiatry 2015;72(5):475-482. [doi: 10.1001/jamapsychiatry.2014.3039] [Medline: 25806661]
69. Hirsh JB, Quilty LC, Bagby RM, McMain SF. The relationship between agreeableness and the development of the working alliance in patients with borderline personality disorder. J Pers Disord 2012;26(4):616-627. [doi: 10.1521/pedi.2012.26.4.616] [Medline: 22867511]
70. Norwood C, Moghaddam NG, Malins S, Sabin-Farrell R. Working alliance and outcome effectiveness in videoconferencing psychotherapy: a systematic review and noninferiority meta-analysis. Clin Psychol Psychother 2018;25(6):797-808. [doi: 10.1002/cpp.2315] [Medline: 30014606]
71. Bailey RC, Knowles NG, Grenyer BF. Efficacy and recommendations for the delivery of telehealth psychotherapy for people with personality disorder. Australas Psychiatry 2024;32(2):125-134 [FREE Full text] [doi: 10.1177/10398562231222768] [Medline: 38103036]

Abbreviations

AIM: Acceptability of Intervention Measure
ATP: according to protocol
BPD: borderline personality disorder
BSL-23: Borderline Symptom List-23
CSQ-8: Client Satisfaction Questionnaire
DBT: dialectical behavior therapy
DSHI: Deliberate Self-Harm Inventory
DSM-5: *Diagnostic and Statistical Manual of Mental Disorders* (Fifth Edition)
DSS: Dissociative Symptom Scale

XSL•FO
RenderX

ITT: intention-to-treat
NSSI: nonsuicidal self-injury
PTSD: posttraumatic stress disorder
RCT: randomized controlled trial
RR: risk ratio
SAE: severe adverse event
UTAUT: unified theory of acceptance and use of technology
WAI: Working Alliance Inventory

Edited by A Mavragani; submitted 06.09.24; peer-reviewed by B Mohajerin, DR Bartsch; comments to author 14.10.24; revised version received 28.10.24; accepted 01.11.24; published 14.01.25.

Please cite as:
Vonderlin R, Boritz T, Claus C, Senyüz B, Mahalingam S, Tennenhouse R, Lis S, Schmahl C, Margraf J, Teismann T, Kleindienst N, McMain S, Bohus M
Acceptance, Safety, and Effect Sizes in Online Dialectical Behavior Therapy for Borderline Personality Disorder: Interventional Pilot Study
JMIR Form Res 2025;9:e66181
URL: https://formative.jmir.org/2025/1/e66181
doi:10.2196/66181
PMID:

©Ruben Vonderlin, Tali Boritz, Carola Claus, Büsra Senyüz, Saskia Mahalingam, Rachel Tennenhouse, Stefanie Lis, Christian Schmahl, Jürgen Margraf, Tobias Teismann, Nikolaus Kleindienst, Shelley McMain, Martin Bohus. Originally published in JMIR Formative Research (https://formative.jmir.org), 14.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Cardiomeds, an mHealth App for Self-Management to Support Swiss Patients With Heart Failure: 2-Stage Mixed Methods Usability Study

Lisa Simioni[1]; Elena Tessitore[1,2], MD; Hamdi Hagberg[3]; Aurélie Schneider-Paccot[2]; Katherine Blondon[1,4], MD, PhD; Liliane Gschwind[5], PhD; Philippe Meyer[2], MD; Frederic Ehrler[3], PhD

[1]Faculty of Medicine, University of Geneva, Geneva, Switzerland
[2]Department of Cardiology, University Hospital of Geneva, Geneva, Switzerland
[3]Department of Computer Science, University Hospital of Geneva, Geneva, Switzerland
[4]Medical Directorate, University Hospital of Geneva, Geneva, Switzerland
[5]Department of Pharmacy, University Hospital of Geneva, Geneva, Switzerland

Corresponding Author:
Lisa Simioni
Faculty of Medicine
University of Geneva
Université de Genève Rue du Général-Dufour 24 CH
1211 Genève 4
Geneva
Switzerland
Phone: 41 786420657
Email: lisa.simioni@hotmail.com

Abstract

Background: Mobile health apps have shown promising results in improving self-management of several chronic diseases in patients. We have developed a mobile health app (Cardiomeds) dedicated to patients with heart failure (HF). This app includes an interactive medication list; daily self-monitoring of symptoms, weight, blood pressure, and heart rate; and educational information on HF delivered through various formats.

Objective: This study aimed to perform a mixed methods usability study of Cardiomeds.

Methods: Smartphone users with HF were recruited from the HF outpatient clinic at the University Hospital of Geneva. The usability test was conducted in 2 stages, with modifications made to the app after the first stage to address major usability issues. Each stage required 10 participants to perform 14 tasks, such as entering vital signs, entering a new medication and time of intake, or finding information about HF. Each task was timed, sessions were recorded, and all data were anonymized. After completing the tasks, patients completed the System Usability Scale 10-item questionnaire and answered 5 open questions about their perceptions of Cardiomeds.

Results: Twenty patients with HF, 75% (15/20) of whom were men, with a mean age of 55 years, were included in this study. The average time to complete all 14 tasks was 18 (SD 5.7) minutes. Manual medication entry was the most time-consuming task, taking an average of 154.40 (SD 68.08) seconds in the first stage, 103.10 (SD 42.76) seconds in the second stage, and 128 (SD 63) seconds overall. The mean overall success rate was 77% (SD 0.23%) for the first stage and 94% (SD 0.07%) for the second stage. A total of 30% (3/10) of participants in the first stage completed all tasks without any help compared with 50% (5/10) of participants during the second stage. The average System Usability Scale score was 80% (SD 17%), showing a slight increase from 79% (SD 16%) in the first stage to 80% (SD 28%) in the second stage, which qualifies the app as “good” in terms of usability. Between the 2 stages, part of the app interface was redesigned to address the key issues identified in the first stage. Despite these improvements, problems related to guidance were frequent and comprised 36% (8/22) of the problems in the first stage and 40% (6/15) in the second stage. In response to open questions, 85% (17/20) of the participants responded that they would like to use the app when it became available.

XSL•FO
RenderX

Conclusions: The usability test indicated that Cardiomeds is a suitable and user-friendly app for patients with HF. The app will be further tested in a randomized clinical trial (2022-00731) after acute HF hospitalization to assess its impact on patients' knowledge about HF, self-care, and quality of life.

(JMIR Form Res 2025;9:e63941) doi:10.2196/63941

KEYWORDS

usability; medication; mobile health; mHealth; Cardiomeds; mobile app; patient empowerment; eHealth; smartphone; heart failure; HF; chronic disease; interactive; self-monitoring; usability test; mobile phone

Introduction

Background

Heart failure (HF) impacts around 2% of the adult population in Europe and is the leading cause of hospital admissions among individuals aged >65 years. Within a year after being hospitalized for acute HF, over half of the patients are readmitted, and nearly a quarter of them die. This situation poses an enormous medical and economic challenge [1]. The cornerstone of HF management is pharmacotherapy, which involves the combination of multiple drug classes [2]. Patients with HF need to make lifestyle adjustments to prevent deterioration, keep track of symptom changes, and take daily medications. Support interventions for self-management behavior generally aim to provide patients with the skills to actively manage their chronic condition. These interventions focus on promoting symptom monitoring and improving problem-solving and decision-making abilities for medical treatment and a healthy lifestyle [3]. According to the latest guidelines of the European Society of Cardiology, self-management strategies are strongly recommended to reduce the risk of hospitalization and mortality related to HF [2]. Various self-management strategies are possible, such as multidisciplinary care, peer meetings, and the use of logs for symptom monitoring or setting goals. The longer these interventions are maintained, the greater the reduction in mortality, with a 1% to 4% decrease in risk for each additional month. In addition, these interventions work better when accompanied by regular contact with health professionals [4].

With advancements in technology, the introduction of therapeutic monitoring with telephone support is becoming increasingly prominent [5], which includes features to register symptoms, enter a treatment card, organize appointments, record alarms for taking medications, or support behavior reinforcement through educational materials [6].

Research consistently supports the effectiveness of structured telephone support and noninvasive telemonitoring. Several systematic reviews have concluded that these approaches significantly reduce all-cause mortality and HF-related hospitalizations [7-9]. Moreover, these interventions also showed enhancements in health-related quality of life, knowledge of HF, and self-care behaviors [7,10].

Smartphone-guided interventions in HF are growing in significance and effectiveness, with studies incorporating monitoring not only through text messages and phone calls but also via mobile health (mHealth) apps [11,12].

Thus, many mHealth apps are being developed and are the subject of various studies [5,6,10,12,13].

Several mHealth apps, providing valuable support to patients with HF are already available. For instance, the WOW ME 2000mg—Heart Failure Self-management Tool (AtlantiCare) focuses on helping patients and caregivers track symptoms and medications [6,14,15]. WebMD (WebMD Health Corp) offers a wide range of health resources, including HF management [14,16]. Similarly, My Cardiac Coach (American Heart Association) is designed for post–heart attack care and self-monitoring. Finally, "Heart Failure Health Storylines" (Heart Failure Society of America) aids in symptom tracking and treatment adherence [6,14,17]. Nevertheless, none of these apps refer to the Swiss medication database.

Experience with these existing mHealth apps has demonstrated a significant improvement in therapeutic adherence among patients requiring daily treatment [18,19]. In addition, patients understood their illness better, which allows them to feel more secure [10].

Despite the potential benefits of mHealth apps, studies have shown that patients with HF often face barriers to achieving good adherence. For many older patients, difficulties begin with technical issues, such as challenges in installing or navigating the apps on their smartphones [20]. These issues are often exacerbated by limited digital literacy, a lack of confidence in using technology, or cognitive impairments commonly associated with HF. In addition, usability concerns, such as complex interfaces, insufficient personalization, or lack of reminders, can further reduce the likelihood of long-term adherence [21]. Overcoming these barriers requires designing apps with simplicity, user-friendliness, and accessibility in mind while providing continuous support for patients [22].

In Switzerland, according to a study conducted by Pro Senectute, 69% of citizens older than 65 years own a smartphone, 81% of whom use it daily. Moreover, this study also discovered that their interest in mHealth apps is consequent [23]. Thus, the familiarity of older adults with smartphones makes the use of an mHealth app credible for conducting a study on patients with HF [24].

In this context, the cardiology department of the University Hospital of Geneva (HUG) has undertaken the development of a new mHealth app tailored for patients with HF, known as Cardiomeds. Designed to complement the physician-patient relationship, the app has been developed following a user-centric approach. We collected the information required to develop the app through interviews and focus groups that included patient groups, clinicians, and IT professionals. We created personas

XSL•FO
RenderX

and user journeys and refined our prototype iteratively using feedback from the stakeholders.

Therefore, it was important to conduct a usability test of this app before offering it to patients with HF. Indeed, these tests make it possible to point out the design problems of the apps, to improve the uncovering of opportunities, and to learn about the target user's behavior and preferences [25].

Objectives

The study aimed to evaluate the usability of the Cardiomeds mHealth app through qualitative and quantitative usability metric in laboratory settings. The main focus of the usability assessment was to test the app's ergonomic design, user-friendliness, comprehensibility, and appropriateness for the target population.

Methods

Study Design

The app's usability was evaluated through a usability test requiring participants to complete a sequence of tasks. The study was performed in 2 stages. After the completion of the first usability test by 10 participants, modifications were implemented in the app based on their feedback. Subsequently, a second usability test, featuring a new cohort of 10 new patients with the same protocol was conducted.

Participants and Setting

The study was conducted from May 6, 2022, to December 21, 2022, in a multipurpose room at the HUG to standardize the intervention. The evaluation was done through user-investigator interaction, deliberately omitting the user's real environments. Tasks were performed on an iPhone 6s smartphone (Apple Inc) at a resolution of 1334×750 pixels. For each of the 2 stages, we recruited 2 groups of 10 participants. Ten participants enabled to increase the number of comments and results in accordance with the recommendations which propose at least 5 participants for one arm of a study [26]. Participants were recruited on a voluntary basis from the HUG HF outpatient clinic. Participants had to be adults (aged ≥18 years), speak French, have and use a smartphone, and be a local resident. All enrolled participants had a diagnosis of HF.

The study was divided into 2 stages, each comprising 10 participants (resulting in a total of 20 participants), as proposed by the Nielsen Norman Group [27].

Cardiomeds Mobile App Overview

Cardiomeds is an mHealth app being developed since 2017 by HUG's cardiology, pharmacy, medical directorate, and IT departments, with the help of computer scientists, a nurse, pharmacists, physicians, and a student [28]. The app has been designed to fulfill the needs of patients with HF, allowing them to gather and record pertinent data concerning their condition, including treatment regimens, vital sign measurements, and the progression of their symptoms. In addition, the app facilitates data sharing with health care providers and enhances patients' understanding of their condition through informative quizzes and simplified content. The app also includes the option of setting reminders or alarms to ensure regular and accurate treatment, as well as monitoring patients' daily medication compliance.

The app's ultimate objective is to empower patients by fostering a comprehensive understanding of their medical condition, promoting effective self-management of daily treatments, facilitating continuous clinical monitoring, and enhancing patients' knowledge of their condition. Figures 1-5 provide an overview of the main menu and the app's appearance as tested by patients.

XSL•FO
RenderX

Figure 1. Screenshot of the main menu section of the Cardiomeds app. In this section, patients enter their vital signs of the day, program reminders, and validate their medication intake. Moreover, if their vital signs are not in a healthy range, the app suggests contacting their physician.

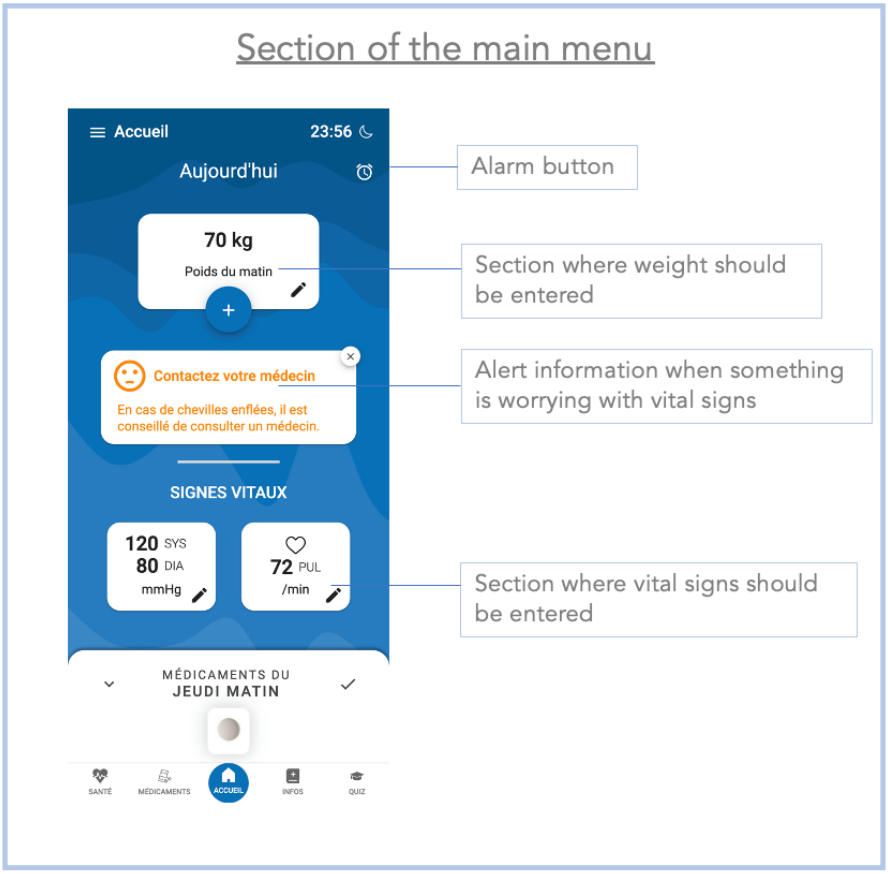

Figure 2. Screenshot of the section where vital signs are entered in the Cardiomeds app. Patients have to enter their systolic and diastolic blood pressure, their pulse, and the day's date.

XSL•FO
RenderX

Figure 3. Screenshot of the information section of the Cardiomeds app. Here, patients can find information about their illness, signs of worsening heart failure, their medications, and recommended lifestyle.

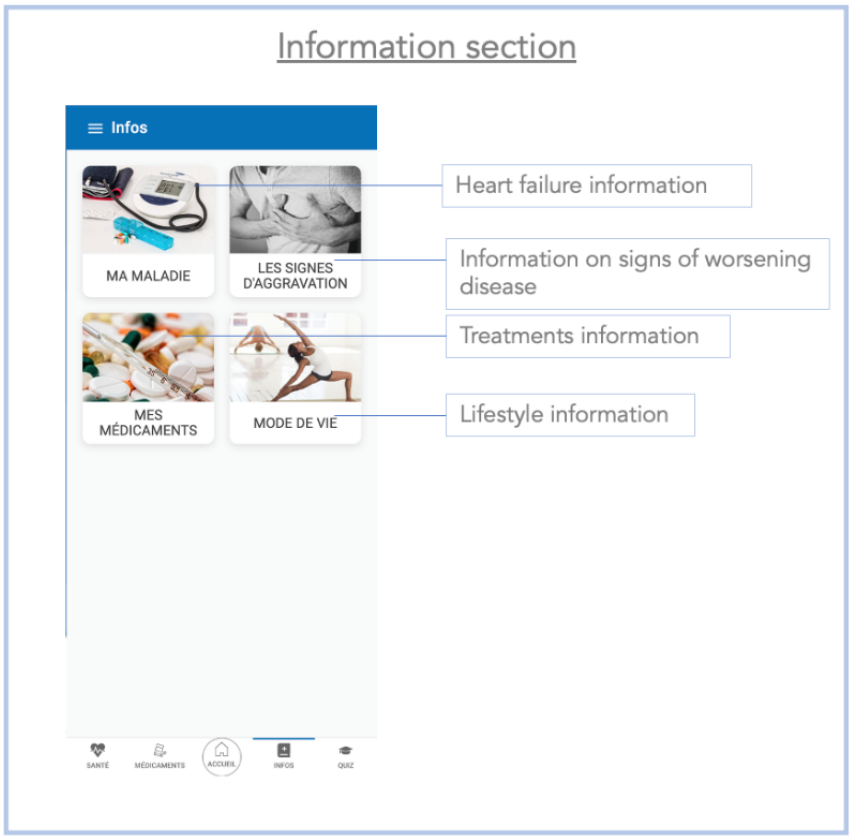

Figure 4. Screenshot of the disease course curve section of the Cardiomeds app. In this section, patients can see the course over the past few days for weight, blood pressure, and pulse. A button to export the curves is also present.

XSL•FO
RenderX

Figure 5. Screenshot of the alarm reminder section of the Cardiomeds app. This section allows the user to set an alarm for taking medication or measuring their vital signs.

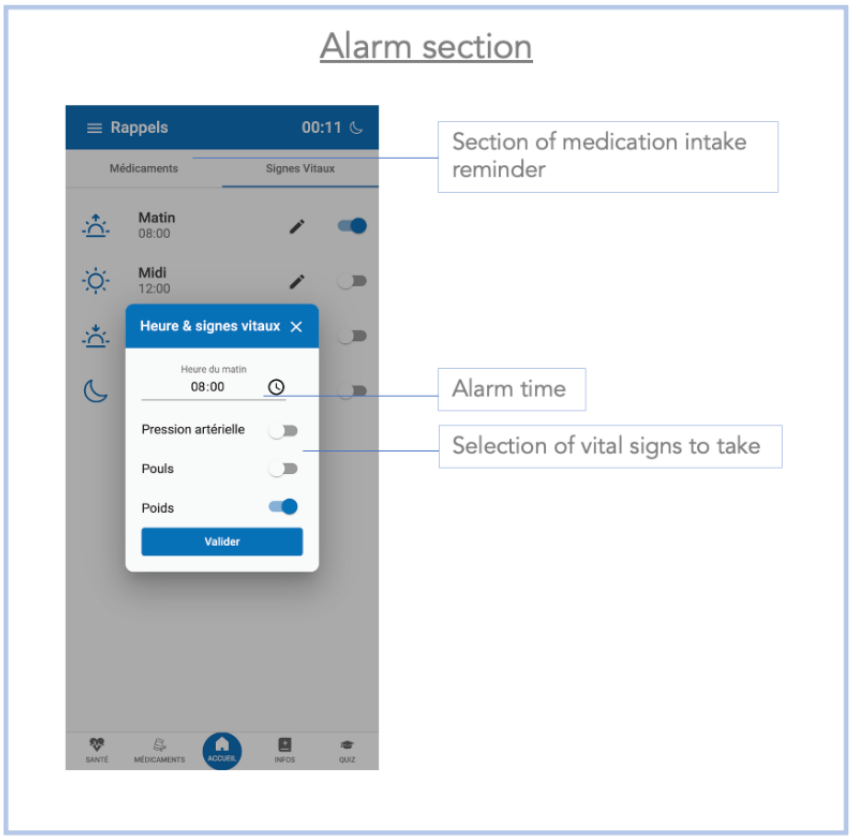

Procedure

All participants were recruited from the HF outpatient clinic of the HUG. They were all invited to join the study via a phone call by a student (LS). Moreover, patients who were selected, were under hospital monitoring, with some having just completed the cardiac rehabilitation program following cardiac decompensation. All interested patients were invited on a given date to take the test. Participants were welcomed within a dedicated room at the HUG, where 2 assigned investigators, a student (LS) and a medical professional (ET), were responsible for overseeing the test procedure. Upon arrival and initial briefing by the investigators, participants signed a consent form, which explicitly stipulated the anonymization of their data.

All participants were provided with a loaned iPhone for test purposes. Then, the test was executed according to the protocol (Textbox 1). Collected measures included task completion timing, recording the activities performed on the smartphone screen, and capturing both the participants' vocal expressions and the investigator's commentary. The test comprised 14 predefined tasks exploring the functionalities of the app (Textbox 2). The investigator read out these tasks and precisely timed their execution. Task outcomes were categorized as follows:

- "Passed" if completed within a maximum of 2 minutes without requiring assistance.
- "Partially successful" if completed in more than 2 minutes or necessitating clues provided by the investigator.
- "Failed" if the task was abandoned or remained unsuccessful after 5 minutes.

Upon test completion, each participant was invited to complete the System Usability Scale (SUS) questionnaire [29,30], as well as 5 open questions about their personal feelings about the app and possible improvements (Textbox 3).

Subsequently, data collected from the tests were analyzed using Microsoft Excel. For in-depth insights into the procedures governing each test, please refer to the comprehensive procedure (Textbox 1), as well as the questionnaires distributed to participants (Textbox 3).

XSL•FO
RenderX

Textbox 1. Procedure of the usability test.

- Welcoming of the user of the app by the investigator.
- Arrival of the user and the worker in a booth.
- Verification of the user's identity and consent by the investigator.
- Asking for the user's patient ID, age, gender, education level, smart phone use time per day, and health app use (yes or no)
- Description of the booth:
  - The booth will contain a desk as well as 2 chairs side by side where the user and the investigator can sit.
  - On the desk, 2 cell phones (a Samsung [Samsung Electronics Co, Ltd] phone and an iPhone) will be placed. Therefore, the user will be able to choose the phone whose operating mode they know and with which they are most comfortable.
  - In addition, a sheet on which the tasks to be carried out during the session will be placed on the desk for the user to guide them.
  - Medication packages will also be placed on the desk to carry out the session's tasks.
- Explanation of the session to the user by the investigator:
  - The investigator will inform the user that their screen and their voice will be recorded (consent).
  - The investigator will explain to the user that the latter will have to carry out the requested tasks, describe their actions out loud, and inform the investigator of the start and end of their tasks.
  - The investigator will explain to the user that after 2 minutes, he or she will gradually give them clues specific to each task to help the user be able to finish the task. If after 5 minutes, the user is unable to complete the task despite help, the task will be marked as failed by the investigator. If the task is completed, the worker will mark it as successful. Once a task is completed, the user can move on to the next task.
  - The investigator will inform the user that he or she can interrupt the session at any time for any reason.
- Start of session:
  - At the start of the session the investigator will set up a screenshot as well as an audio recording if the user gives their consent.
  - The investigator will read the tasks aloud one after the other to the user.
  - Before starting to perform the tasks, the user will have 2 minutes to familiarize themselves with the design of the app.
  - At the beginning, the only questions the investigator will answer will be questions of understanding, if there is a misunderstanding of terminology.
  - The user can reread the task on the sheet placed on the desk for them if necessary.
  - As soon as the user starts each task, he or she will mention it out loud.
  - As soon as the user finishes each task, he or she will mention it out loud.
  - If the user has comments, the investigator will listen to them.
- Tasks: read aloud 14 tasks for the user. Please refer to the tasks below in Textbox 2.

XSL•FO
RenderX

Textbox 2. Description of the 14 goal-oriented test tasks to be performed by users during the usability study.

- Task 1: set alarm for vital signs (set a reminder to measure your vital signs at 8 AM).
- Task 2: decision-making (go to the home page and enter your morning weight of 65.0 kg). You must then answer 3 questions as follows:
  - Short of breath? As usual.
  - Swollen ankles? More than usual.
  - Any other symptoms? None of the above.
- Task 3: enter weight (you forgot to record your weight on 04.01.2022, and you now want to enter this value, which was 62.3 kg).
- Task 4: enter vital signs (enter your daily vital signs: blood pressure 120/80 mm Hg and pulse rate 70/min).
- Task 5: enter medication (add the medication to the app's medication list using the available package. Scan the barcode for Beloc ZOK 25 mg. The prescription for Beloc ZOK 25 mg is 2 1⁄2 tablets (2.5 tablets) in the morning every day. In addition, add the number of tablets found in the package).
- Task 6: enter medication without medication package (add a second medication to the app's medication list. You do not have the medication box. The prescription is as follows: Fosamax [bisphosphonate] 70 mg, 1 tablet on Sunday mornings only; 1 per week. The quantity in the pack is 4 tablets).
- Task 7: enter pro re nata (PRN; taken as needed without a set schedule) medication, without medication package (add a third medication to the app's medication list. You do not have the medication package. The prescription is as follows: zolpidem (any type proposed) 10 mg, maximum 1 tablet per day, in the evening if necessary. The number of tablets in the pack is 10).
- Task 8: set alarm for medication intake (set a reminder to take your medication at 8:00 AM).
- Task 9: confirm your medication intake (confirm that you have taken your morning medication).
- Task 10: check medication history (check your medication history).
- Task 11: find medication information (you no longer know why you are taking Beloc ZOK [β-blocker]. Where can you find this information? There are 2 ways to find this information).
- Task 12: export vital signs (export your health data [weight, blood pressure, and pulse] to your doctor by email).
- Task 13: quiz (test your knowledge on heart failure with the quiz).
- Task 14: find alcohol information (find information on the possibility of consuming alcohol in the context of your disease).

Textbox 3. Five open questions questionnaire given to users after passing the test.

1. What is your overall impression of the Cardiomeds app?
2. Do you plan to use Cardiomeds when the app is available, and why?
3. What improvements could we add to the Cardiomeds app?
4. Do you have any other comments about the Cardiomeds app or the usability test?
5. What differences have you noticed between Cardiomeds and other health apps?

Data Collection

Participants were audio-recorded during the test, and the screen of the test iPhone was recorded. The recording allowed us to retrospectively classify the usability problems encountered with the Bastian and Scapin criteria [31]. The audio was recorded using an iPad-mini-Dictaphone by a single investigator. Moreover, during the test, the investigator timed the tasks with another smartphone. The screen was recorded using the screen recording option built in the iPhone on which the participants were taking the tests. Subsequently, the usability metrics were transcribed onto Microsoft Excel spreadsheets. The audio records were analyzed by the investigator using thematic analysis. The SUS paper questionnaires were collected immediately after the intervention and transcribed onto Microsoft Excel spreadsheets. The investigator analyzed the success rates of each task and their duration. All of the collected data were anonymized. All data described above were collected and analyzed by a single investigator for the 20 participants. ChatGPT (GPT-4 version, OpenAI) was used as an assistance tool to generate text drafts, as a translation tool, and to synthesize concepts. The results were subsequently reviewed and edited by the authors to ensure accuracy and relevance to the topic.

Usability Analysis

Quantitative Evaluation

The participant's task-based performance was measured by several metrics. Effectiveness is defined by the number of tasks completed by each participant accurately [32]. In this study, effectiveness was calculated in 2 ways: "task completion rate" (TCR) per participant is the percentage of tasks successfully completed [33]. This was calculated using the following equation: TCR per participant=(number of tasks completed

XSL•FO
RenderX

successfully/total number of tasks undertaken)×100. (equation 1) When a task could not be started and evaluated (ie, because of a problem with the Wi-Fi connection), it was coded as "unavailable." TCR per participant was categorized according to 3 different levels of achievement as follows: (1) the task is considered completed when the user has successfully completed the task without any errors or difficulties; (2) completed with difficulty, but with difficulties that could have been solved with the help of the instructor in 5 minutes or less; and (3) failed to complete when the task was left incomplete, abandoned, or the participant took too much time to complete the task (>5 min) [33].

"Distribution of task success by task" was defined as the proportion of participants completing a task according to 3 possible levels of achievement: (1) the task is considered completed when the user has successfully completed the task without any errors or difficulties; (2) completed with difficulty, but with difficulties that could have been solved with the help of the instructor in 5 minutes or less; and (3) failed to complete and the task is left incomplete, abandoned, or the participant took too much time to complete the task (>5 min). When a task could not be started and evaluated (ie, because of a problem with the Wi-Fi connection), it was coded as unavailable [33].

"Efficiency" was defined as the time required for users to achieve specified goals in relation to accuracy and completeness [32]. In this study it was calculated as follows: "time on task" was defined as the average amount of time taken (in seconds) to complete a given task from the moment the participant finished hearing the instructions until the task was completed (whether with ease or with difficulty) or abandoned. When a task could not be started and evaluated (ie, because of a problem with the Wi-Fi connection), it was coded as unavailable.

Satisfaction was measured through the SUS questionnaire. After the completion of the usability test, the SUS questionnaire [29,30] was submitted to each participant. The SUS is a validated questionnaire designed to measure the usability of diverse products and services, including mHealth apps. This questionnaire, conceived by Brooke [29], comprises 10 questions, each offering 5 possible answers, ranging from "strongly disagree" to "strongly agree." In this ordinal scale, "strongly disagree" corresponds to a numerical value of 1, while "strongly agree" corresponds to a numerical value of 5. The even-numbered questions within the SUS are subjected to negative rotation, and their scores are calculated by subtracting the chosen value from 5. In contrast, the odd-numbered questions undergo positive rotation, and their scores are derived by subtracting 1 from their value. Once the individual scores for all 10 questions were calculated, they are added together and multiplied by a factor of 2.5. Consequently, the resultant SUS score falls within the range of 0 to 100. Usability is considered favorable when the SUS score attains a threshold of ≥70, signifying good usability for the tested product.

Qualitative Evaluation

Qualitative data were recorded through an open questionnaire. After completing the SUS questionnaire, participants were asked to answer a series of 5 open questions. The purpose of these questions, defined by the investigators, was to explore the participants' overall impression allowing us to better understand the strengths and weaknesses of our tools. The 5 open questions are described in Textbox 3.

The responses to the first question were classified into positive responses, negative responses, and no responses. The responses to the second question were classified into 4 categories: yes, no, maybe, and do not know. Questions 3 and 4 were analyzed using the frequency of propositions. All answers to the questions were recorded, sorted, and analyzed by a single person.

Bastien and Scapin Classification

We also used the criteria published by Bastien and Scapin [31] in 1993 to evaluate the problems of the app. These criteria cover 8 dimensions and enable the classification of problems between users and the interface.

Ethical Considerations

On April 1, 2022, a formal request was submitted to the Ethics Commission of the Canton of Geneva to conduct usability tests (Req-2022-00423). This request was subsequently approved on April 4, 2022. The app encompassed the testing protocol, a consent form to be executed by the study participants, and an evaluation sheet for each participant.

Results

Participant Characteristics

Between May and December 2022, a total of 20 participants participated in the study. Baseline demographic characteristics of the 2 study stages are shown in Table 1.

Demographic characteristics of study participants (N=10).

XSL•FO
RenderX

Table 1. Characteristics of the participants separated into the 2 stages of the study (N=20).

	Stage 1, n (%)	Stage 2, n (%)	Total, n (%)
Gender			
Women	3 (30)	2 (20)	5 (25)
Men	7 (70)	8 (80)	15 (75)
Age categories (years)			
31-40	1 (10)	0 (0)	1 (5)
41-50	3 (30)	1 (10)	4 (20)
51-60	4 (40)	5 (50)	9 (45)
61-70	0 (0)	4 (40)	4 (20)
71-80	2 (20)	0 (0)	2 (10)
Type of phone			
iPhone	3 (30)	1 (10)	16 (80)
Samsung	7 (70)	9 (90)	4 (20)
Educational level			
PRE[a] graduate school	5 (50)	5 (50)	10 (50)
Graduate school	2 (20)	1 (10)	3 (15)
University	3 (30)	4 (40)	7 (35)
Use of other medical apps			
Yes	4 (40)	2 (20)	6 (30)
No	6 (60)	8 (80)	14 (70)
Time of phone use (h)			
<1	1 (10)	0 (0)	1 (5)
1-3	8 (80)	7 (70)	15 (75)
>3	1 (10)	3 (30)	4 (20)

[a]PRE: federal diploma of vocational education and training.

First Stage of the Study

Quantitative Evaluation

Effectiveness per Participant

In the first part of the study, the overall completion rate (tasks completed, partially completed, and failed to complete) was 100% (140/140; Figure 6). The mean overall complete success rate was 50% (5/10), meaning that 50% of the participants solved all the tasks without making any mistakes (they did not fail to complete any task) and 30% (3/10) of the participants completed all the tasks without any help. Participant number 6 failed 50% (7/14) of the tasks and completed only 20% (3/14) of them without any help. Finally, on average, participants completed 77% (11/14; SD 0.23%) of tasks without any help. All these measurements are exposed and supported in Figure 6.

XSL•FO
RenderX

Figure 6. Task completion rate per participant for the 14 assigned tasks during the first stage of the study. "Task completed" represents the percentage of tasks successfully completed by a participant. "Failed to complete" is the percentage of tasks that participants failed to complete. "Partially completed" is defined as tasks completed, but with difficulty that could have been solved with the help of the instructor in 5 minutes or less. "Nonavailable" represents the percentage of missing data when a task could not be started and evaluated.

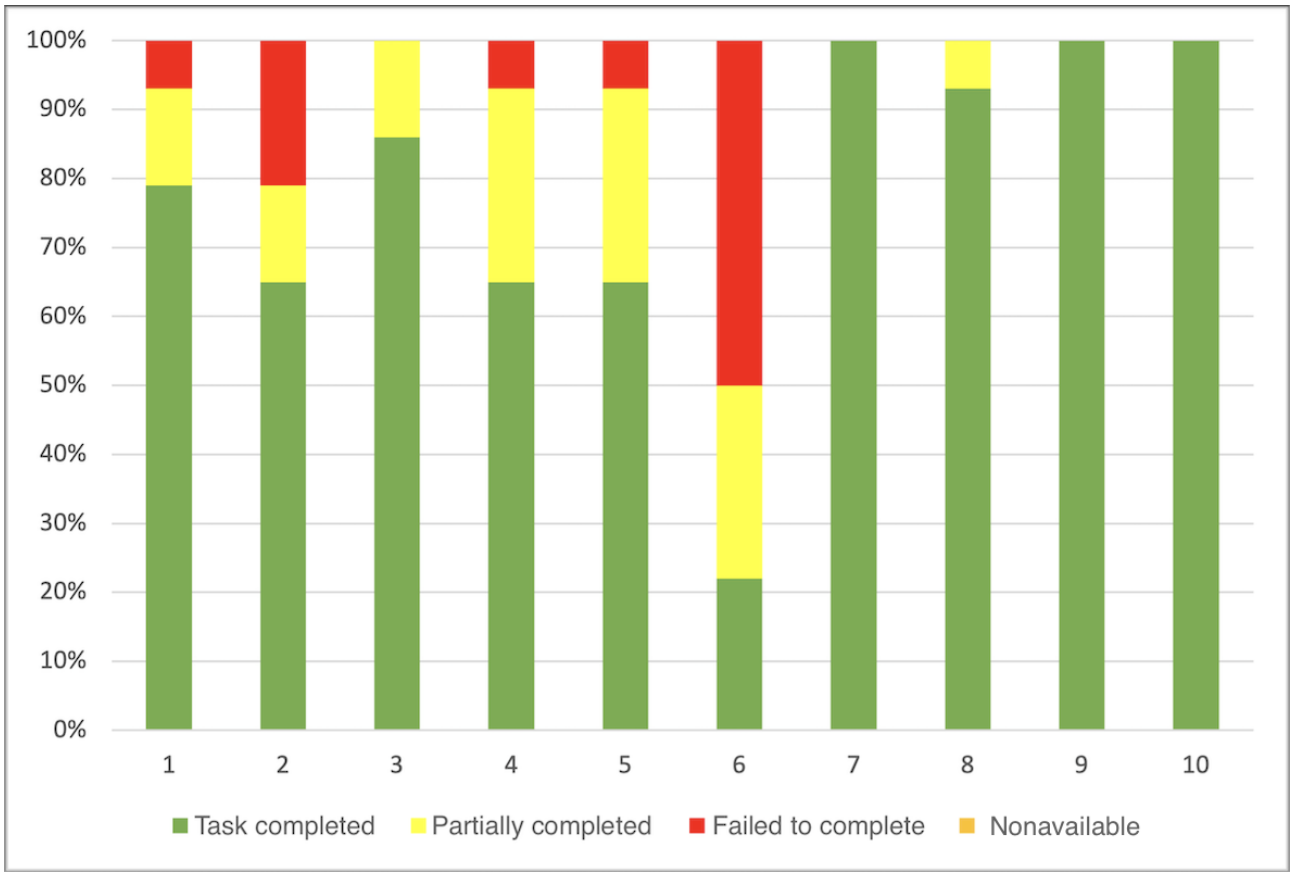

Task Success Distribution per Task

In the first part of the study, overall 77.1% (108/140) of tasks were completed successfully, 13.6% (19/140) were completed partially, and 9.3% (13/140) were not completed. A total of 21% (3/14) of the tasks (4, 10, and 14) were completed by all the participants without any help (10/10, 100%), followed by task 2 (9/10, 90%). Apart from tasks 4, 10, and 14, 79% (11/14) of the tasks led to difficulties for up to 60% (6/10) of the participants (task 6). At least 1 (1/10) participant failed to complete 8 (8/14) of the tasks (1, 5, 6, 7, 8, 9, 11, and 12), with up to 30% (3/10) of the participants failing to complete task 11, which consisted of finding information about medications in 2 different ways. All these measurements are exposed and supported in Figure 7.

XSL•FO
RenderX

Figure 7. The success distribution per task (N=10 participants) for the first stage of the study. "Complete" represents the percentage of tasks successfully completed by a participant. "Failed to complete" is the percentage of tasks that participants failed to complete. "Partially complete" is defined as tasks completed, but with difficulty that could have been solved with the help of the instructor in 5 minutes or less. "Nonavailable" represents the percentage of missing data when a task could not be started and evaluated.

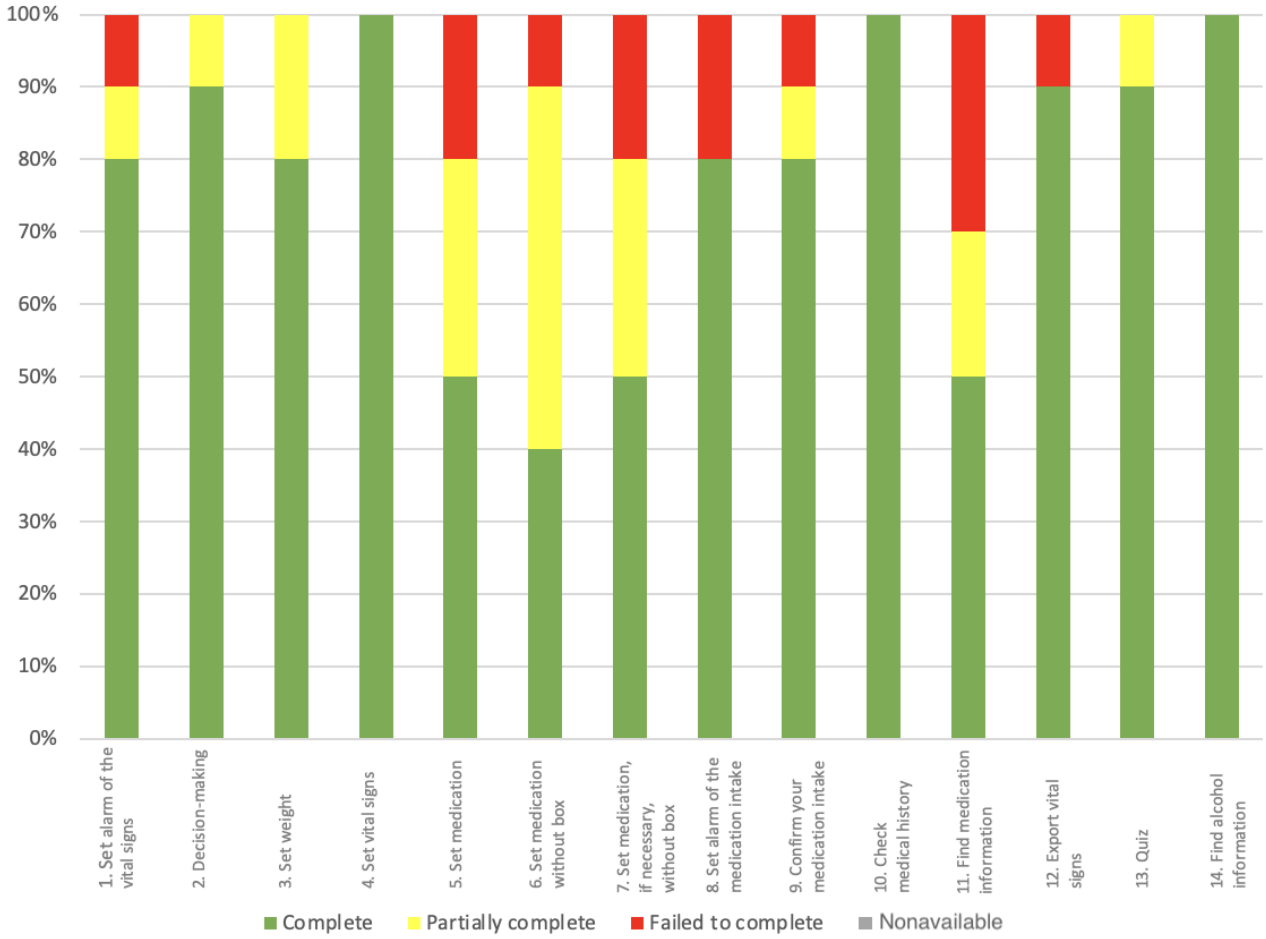

Time on Task per Study Task

In the first stage of the study, task 6, which required the participants to enter medication, took the longest time with an average of 154.40 (SD 68.08) seconds (Figure 8). The fastest task completed was checking medication history (task 10) with an average of 4.8 (SD 4.60) seconds.

XSL•FO
RenderX

Figure 8. Boxplot of the time taken (in seconds) to perform each task in stage 1.

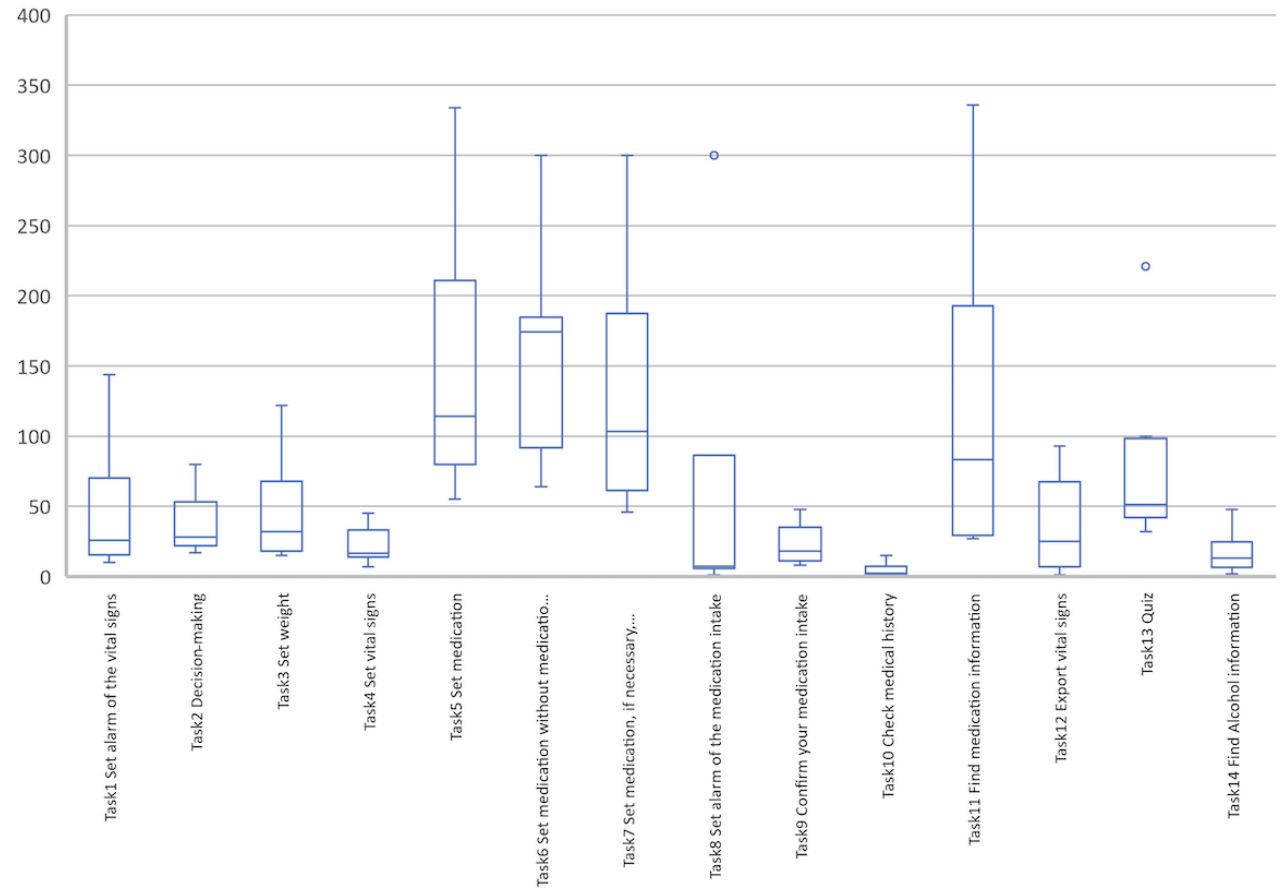

Qualitative Evaluation

Five Open Questions Questionnaire

In stage 1, 80% (8/10) of the participants answered positively, 10% (1/10) answered negatively, and 10% (1/10) gave no answer about their overall impression. A total of 70% (7/10) of the participants said that they would like to use the app in the future, 20% (2/10) answered they would not, and 10% (1/10) did not answer the question. The participants who did not want to use the app mentioned that they already had an mHealth app or found that the app was too complicated for them. Participants suggested 3 different types of improvements for the app as follows: first, to increase the font size of the app; second, to change the medication input; and third, to improve the explanation given about the HF drugs. These comments are more described in Textbox 4.

Textbox 4. List of comments made by participants (N=10) that are useful in the short- and long-term for improving the app during the first stage of the test.

Useful comments for short-term improvements

1. The 2 (20%) oldest participants mentioned that the font in the app was too small to understand, which may be problematic for older users.
2. One (10%) participant was interested in transferring his data, but also in having contact with his cardiologist directly via the app.
3. Six (60%) participants found that entering the information on medications was complicated.
4. One (10%) participant wished to have the emergency number, the number of his physician, his cardiologist, and his pharmacy inside the app.

Useful comments for long-term improvements

1. One (10%) participant was interested in a search function in the information part.
2. One (10%) participant wished to have interactive blood pressure and saturation connected with his smartwatch.
3. One (10%) participant wanted this app to be connected with his other apps linked to the University Hospital of Geneva Swissmeds app.

Error Classification With the Bastien and Scapin Criteria

In stage 1, we recorded 22 problems during the completion of the 14 tasks. We have classified the problems into 9 categories according to the criteria of Bastien and Scapin [31] (Table 2). Task 6, requiring entering a medication, was the most problematic, with 6 (N=10, 60%) users experiencing problems.

XSL•FO
RenderX

The screen enabling manual medication recording was considered too dense with information.

Table 2. List of problems encountered by participants (N=10) during the 14 tasks and description, categorization, and frequency of occurrence of these problems during task completion.

Item	Category (number according to the Bastien and Scapin classification)	Description of the problem	Task in which it occurs (frequency of occurrence)
1	1. Guidance	Five (50%) users did not find the place of the button to put a reminder. For most of them, it was too small, and its location was not easy to find.	Set alarm for the vital signs (5)
2	1. Guidance	Three (30%) users said the "information" symbol was too small next to the treatments list and that they did not find it easily	Find medication information (3)
3	7. Significance of codes	Two (20%) users did not find the scanning possibility option in the app explicit enough	Enter medication (2)
4	2.2 Information density	Six (60%) participants had difficulties recording medication manually in the app. They say they had to put too much information on one page.	Enter medication without medication package (6)
5	7. Significance of codes	One (10%) participant did not understand names of some drugs because the name of the medication was not written next to the active substance	Find medication information (1)
6	1.3 Immediate feedback	One (10%) participant found the validation of medication intake not explicit enough	Confirm your medication intake (1)
7	7. Significance of codes	Two (20%) participants did not understand the term "if necessary," such as medication to take in reserve	Enter pro re nata, if necessary, without medication package (2)
8	8. Compatibility	One (10%) participant asked to set an end date for his treatment, but this function does not exist	Enter medication (1)
9	5.2 Error message quality	When entering the date, one participant (10%) did not write it correctly (separated by "/" symbols and not "." symbols). Therefore, the date could not be validated and no explanation appeared	Enter weight (1)

Satisfaction: SUS Questionnaire

At the end of each test, all participants were asked to complete the SUS questionnaire [29,30]. The mean SUS score of the first stage of the study was 79.28 (SD 16.35), (Figure 9; Table 3).

Figure 9. Cardiomeds score on the System Usability Scale (SUS) at stage 1.

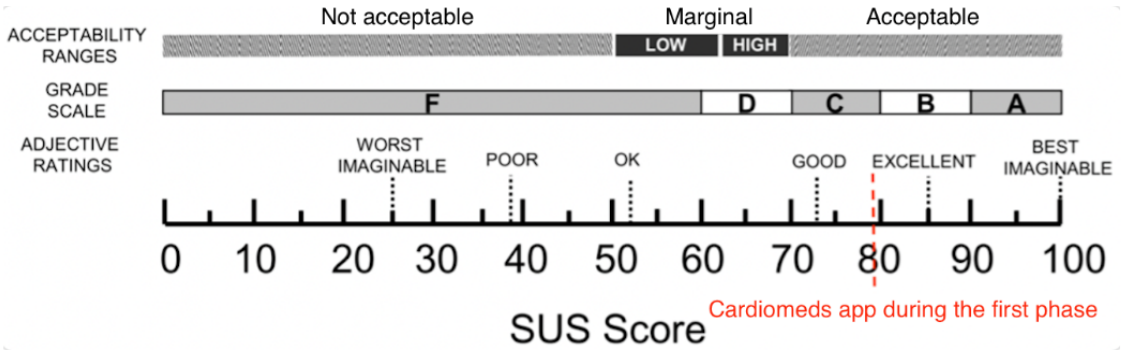

XSL•FO
RenderX

Table 3. System Usability Scale (SUS) questionnaire responses in the first stage of the study.

	Question 1	Question 2	Question 3	Question 4	Question 5	Question 6	Question 7	Question 8	Question 9	Question 10	SUS score (sum×2.5; maximum 100)
Participant 1	4	4	3	4	2	2	3	4	3	3	80
Participant 2	3	3	3	3	3	4	3	4	3	1	75
Participant 3	4	4	4	4	4	4	4	4	4	4	100
Participant 4	4	1	2	1	4	3	3	3	3	2	65
Participant 5	4	4	4	3	3	4	3	4	4	4	92.5
Participant 6	4	1	0	0	4	4	4	0	3	0	50
Participant 7	4	4	4	4	4	4	0	4	4	4	90
Participant 8	4	4	4	4	4	4	4	4	4	2	95
Participant 9	2	3	3	4	3	3	3	2	4	4	77.5
Participant 10	2	4	4	4	4	4	3	4	4	4	92.5
Mean (SD)	3.5 (0.82)	3.2 (1.19)	3.1 (1.22)	3.1 (1.40)	3.5 (0.81)	3.6 (0.82)	3 (1.1)	3.3 (1.29)	3.6 (0.52)	2.8 (1.40)	79.28 (16.35)
Median (IQR)	4 (4-3.25)	4 (4-3)	3.5 (4-3)	4 (4-3)	4 (4-3)	4 (4-3.25)	3 (3.75-3)	4 (4-3.25)	4 (4-3)	3.5 (4-2)	80 (92.5-70)

Improvements Made Between the 2 Stages of the Study

After completing the test with the first 10 users, we realized that many users encountered difficulties entering medication without scanning the package (task 6). In addition, the users provided comments suggesting the need to facilitate the recording of medications. During stage 1, the manual recording of medications was carried out in a single page (Figure 9). To simplify this process, the ergonomist proposed splitting the process into 3 pages, each requiring a validation (Figures 10-12).

We also made some minor changes, such as modifying the localization and visibility of the reminder and information sections to make them more accessible and more visible from the main menu. Once the modifications were done, we began stage 2 of the study. However, we did not increase the font size in the app, despite the many comments we received as mentioned earlier. This did not seem necessary because the font on the smartphone can be changed in its own settings, allowing each user to adapt it to their desires and needs.

XSL•FO
RenderX

Figure 10. Screenshot of the rolling menu for new medication, before improvements were made, during stage 1 of the study. All the process was presented on one page. First, the patient had to insert the name of their medication and then the time of the intake. There was also a scan button to scan the barcode of the medication.

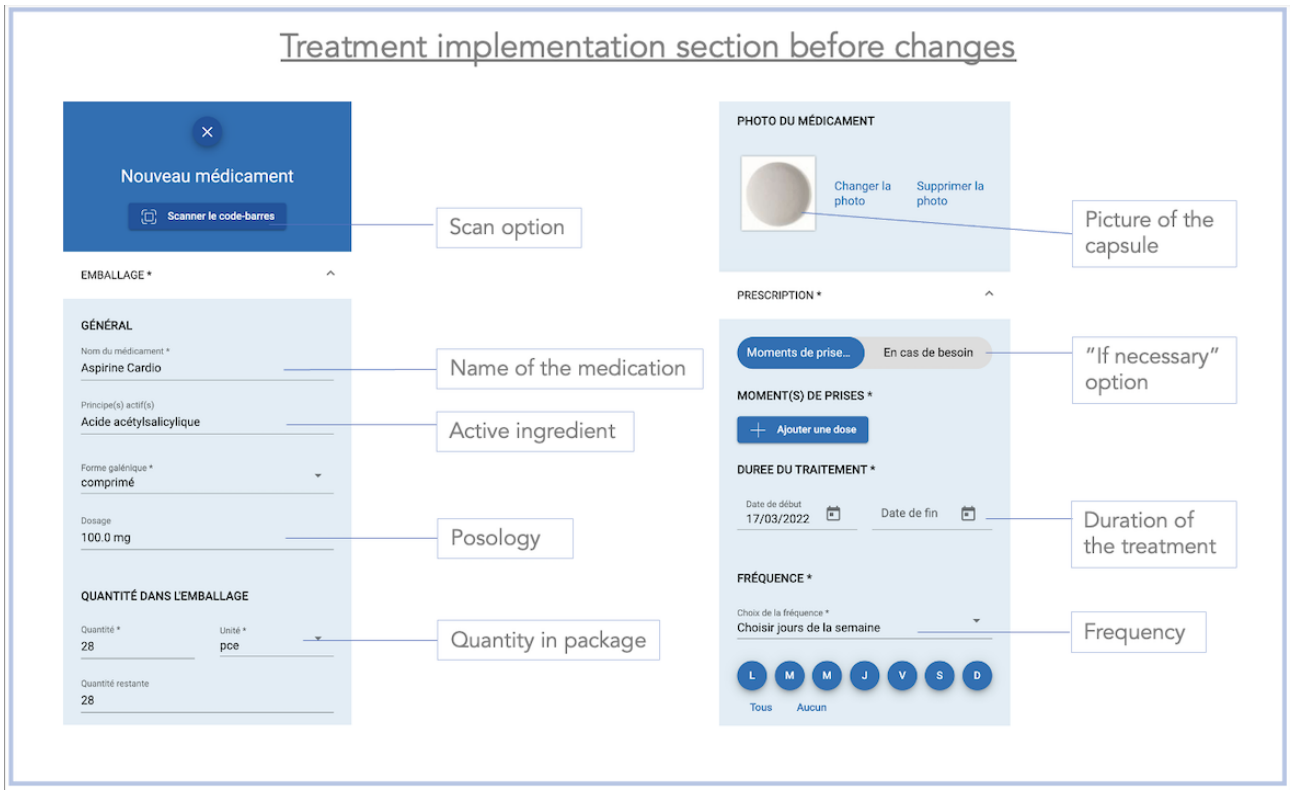

Figure 11. Screenshot of the medication intake section of the Cardiomeds app after the improvements were made during stage 2. First, the patient had to enter the name of the medication or scan the barcode. Then, they pressed the "next" button to go to the next section.

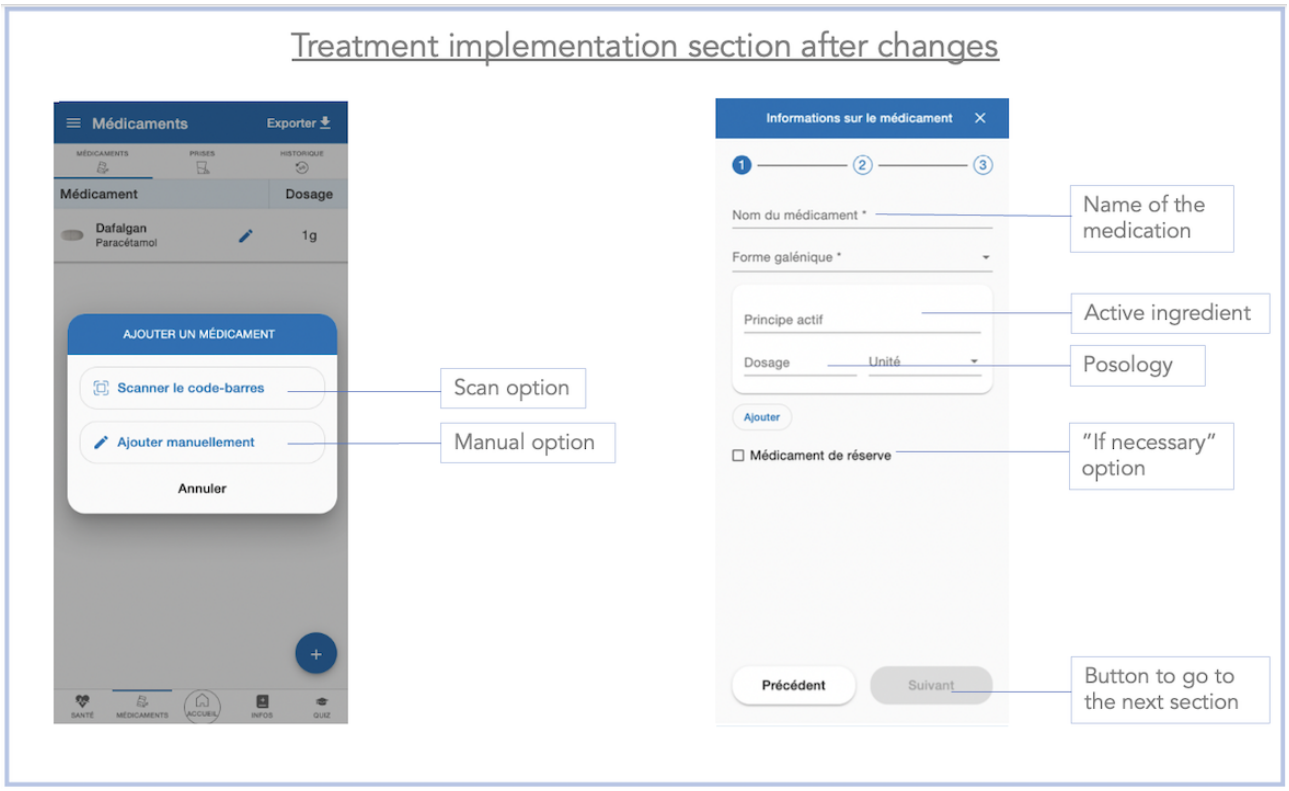

XSL•FO
RenderX

Figure 12. Screenshot of the medication section of the Cardiomeds app after implementation of the changes during phase 2. The patient reached this section after entering the name of the medication or after scanning the medication barcode. Here, the patient had to select the time to take the medication, the dosage, and the frequency.

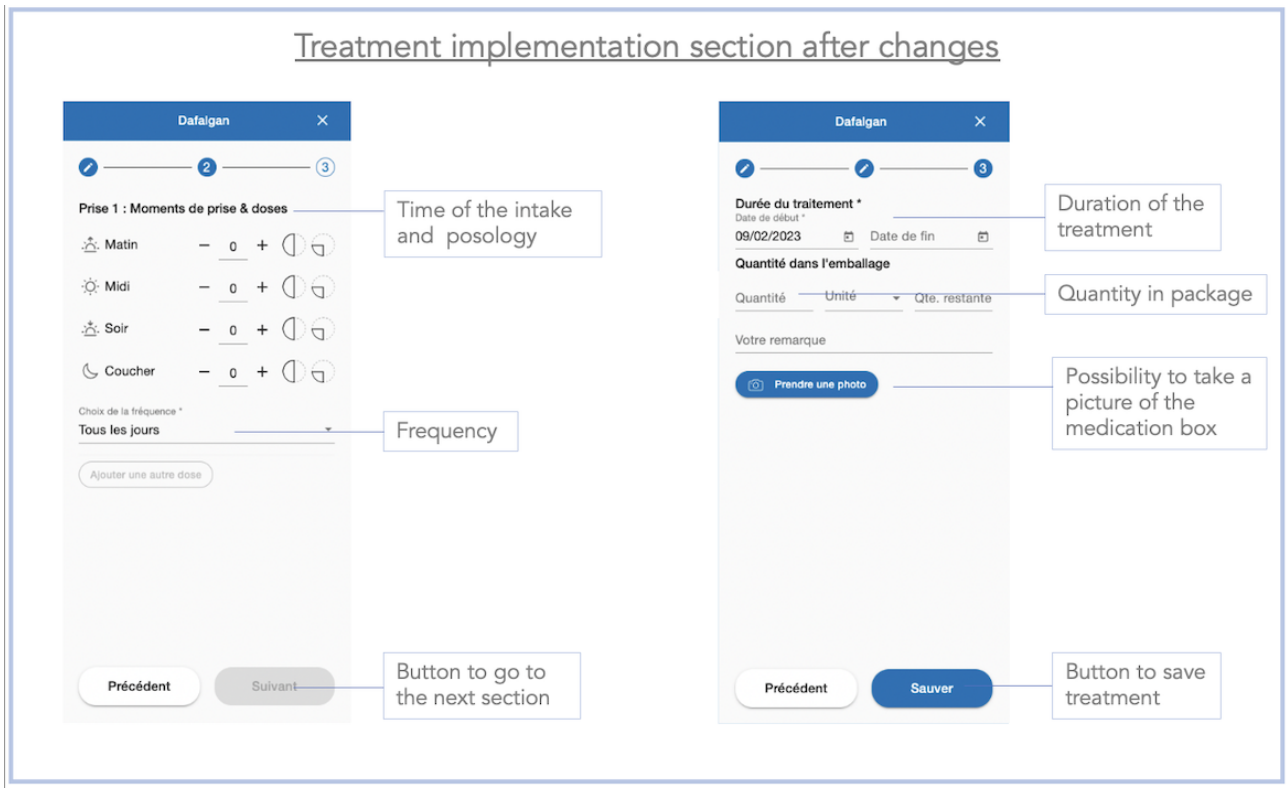

Second Stage of the Study

Quantitative Evaluation

Effectiveness per Participant

In stage 2, the overall completion rate of the test (tasks completed, partially completed, and failed to be completed) was 99.3% (139/140). Participant 4 did not perform task 14 during his test because of a software bug. None of the participants failed to complete any task. Only 50% (5/10) of the participants needed help to complete at least one task. Participant 2 needed the most help to finish the test, with 20% (3/14) of the tasks completed with help. All these measurements are exposed and supported in Figure 13.

XSL•FO
RenderX

Figure 13. Task completion rate per participant for the 14 assigned tasks during the second stage of the study. "Task completed" represents the percentage of tasks successfully completed by a participant. "Failed to complete" is the percentage of tasks that participants failed to complete. "Partially completed" is defined as tasks completed, but with difficulty that could have been solved with the help of the instructor in 5 minutes or less. "Nonavailable" represents the percentage of missing data when a task could not be started and evaluated.

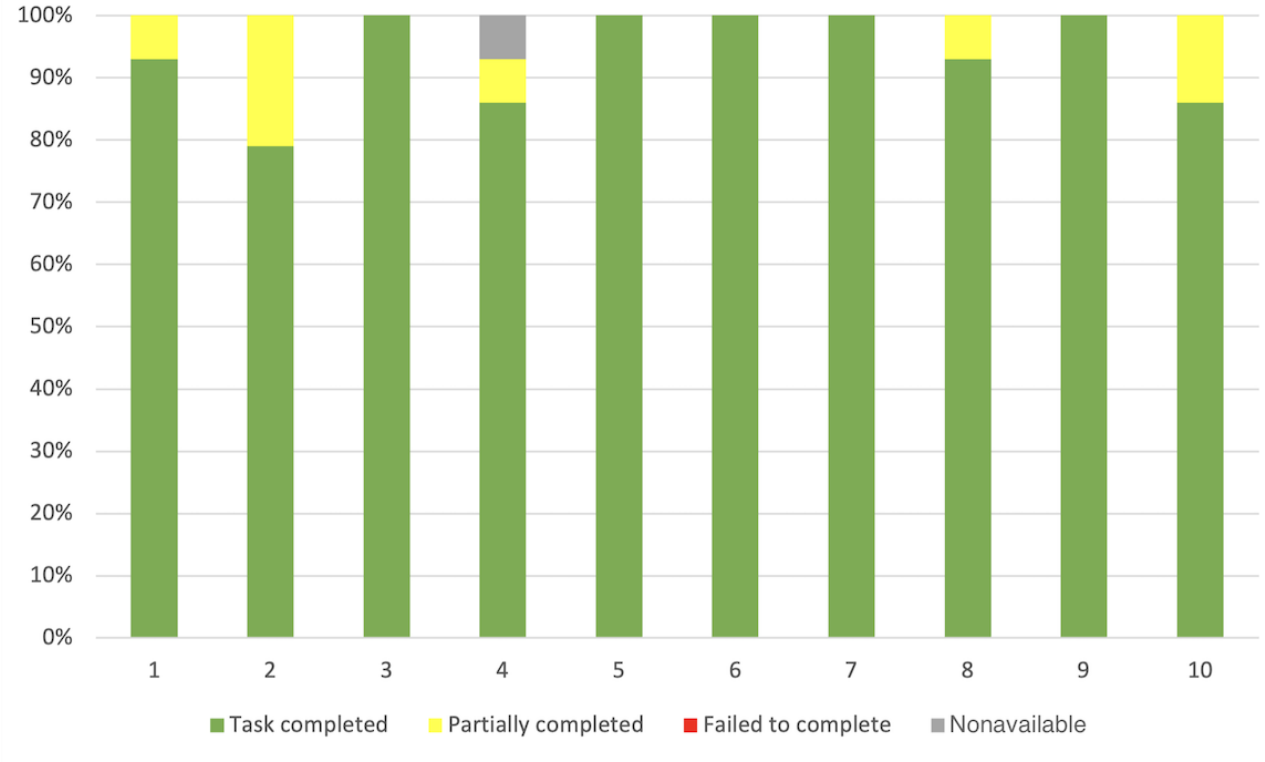

Task Success Distribution per Task

In stage 2, overall 93.6% (131/140) of tasks were completed successfully, 5.7% (8/140) were completed partially, and 0% (0/140) failed. A total of 64% (9/14) of tasks 2, 3, 4, 7, 8, 9, 10, 12, and 14 were completed successfully by all participants (100%, 10/10), followed by task 13 (9/10, 90%). A total of 29% (4/14) of the tasks 1, 6, 11, and 13 were completed with difficulty with a rate ranging from 70% to 90%. In this stage of the study, no task completion error was made. Task 5 (enter medication) is the only one with 10% (1/10) missing data because of a technical problem during the test. During this task, the photo scan of the phone did not work. All these measurements are exposed and supported in Figure 14.

XSL•FO
RenderX

Figure 14. Task success distribution per task (N=10 participants) for the second stage of the study. "Complete" represents the percentage of participants who completed the task with ease. "Partially complete" represents the percentage of participants who completed the task with difficulty that could have been solved with the help of the instructor in 5 minutes or less. "Failed to complete" is defined as the percentage of participants who failed to complete the task. "Nonavailable" represents the percentage of missing data when a task could not be started and evaluated.

Time on Task per Study Task

In the second part of the study, entering medication details (task 6) remained the longest in duration with an average of 103.10 (SD 42.76) seconds (Figure 15). The task that was completed the fastest was task 8 with an average of 21.10 (SD 25.69) seconds.

XSL•FO
RenderX

Figure 15. Boxplot of the time taken (in seconds) to perform each task in stage 2.

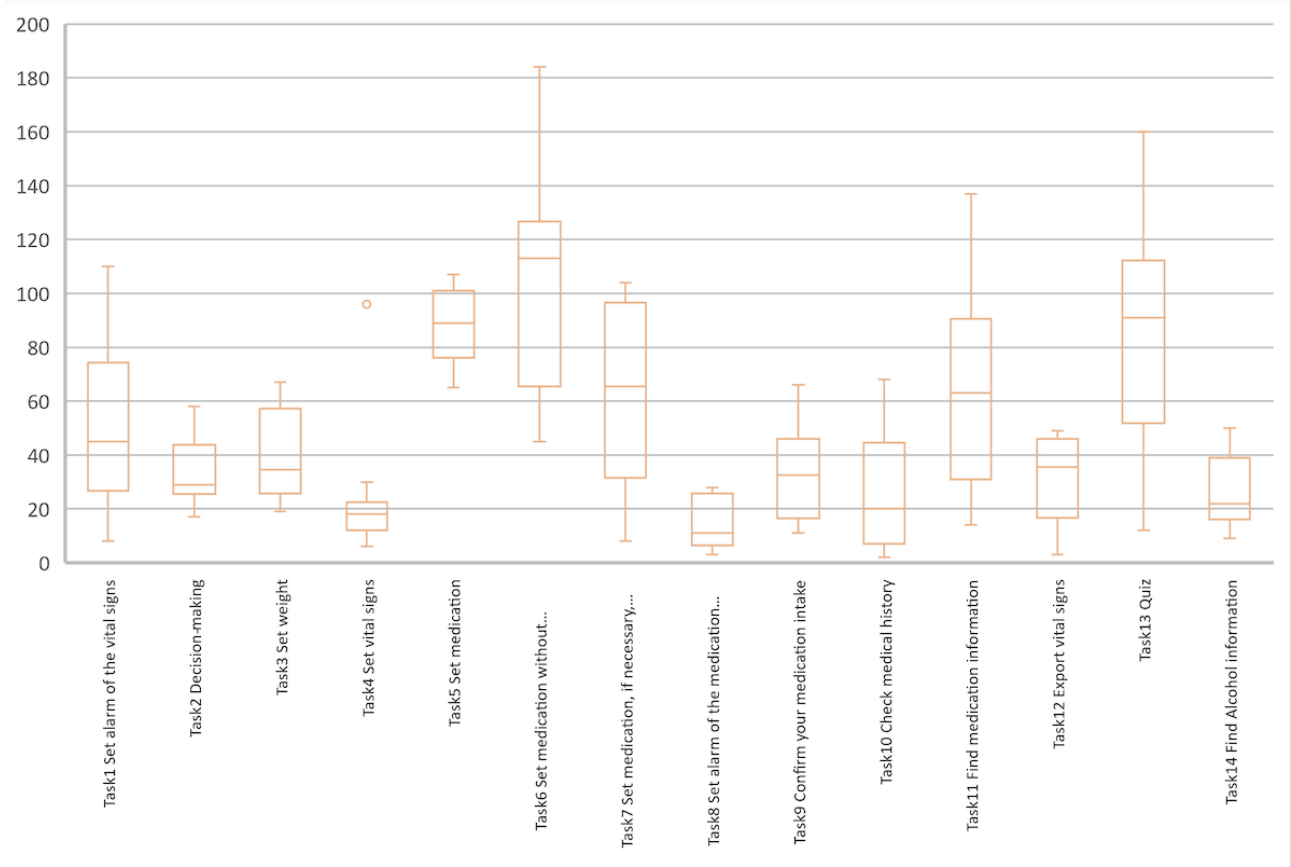

Qualitative Evaluation

Five Open Questions Questionnaire

In stage 2, 80% (8/10) of participants answered that their impression of the app was "positive," 0% answered negatively, and 20% (2/10) did not answer this question.

Furthermore, 100% (10/10) of the participants answered that they would like to use the app.

During the second part of the study, participants suggested making the validation of medication intake more explicit, adding some emergency numbers inside the app, and putting the name of the drug next to the name of the active ingredient of the drug to increase patients' understanding of their treatments (Textbox 5).

Textbox 5. List of comments made by participants (N=10) that are useful in the short- and long-term for improving the app during the second stage of the test.

Useful comments for short-term improvements

- One (10%) participant suggested making the validation of drug intake more explicit
- One (10%) participant wished to have the emergency number and the number of his physician, his cardiologist, and his pharmacy inside the app

Useful comments for long-term improvements

- One (10%) participant wanted to have the name of the medication next to the active principle in the medication information

Error Classification With the Bastien and Scapin Criteria

In the second part of the study, we recorded a total of 15 problems. We classified them into 7 categories according to Bastien and Scapin's criteria [31] (Table 4). The most problematic were tasks 9 (confirm your medication intake) and 12 (export vital signs), with 4 (N=10, 40%) users experiencing problems with each task. Participants reported that the validation of daily medication intake was not explicit enough. The participants who encountered a problem with task 12 found that the "export" button was too small and therefore not legible and that the location of this button was not well placed.

XSL•FO
RenderX

Table 4. List of problems encountered by participants (N=10) of the second part of the study during the 14 tasks and description, categorization, and frequency of occurrence of these problems during task completion.

Item and problem	Category (number according to the Bastien and Scapin classification)	Description of the problem	Task in which it occurs (frequency of occurrence)
1	1. Guidance	Five (50%) users found the front text size too small and the localization of some symbols not easy to find	Enter medication (1) and export vital sign (4)
2	1.3 Immediate feedback	Four (40%) participants did not easily understand how to validate a medication intake	Confirm your medication intake (4)
3	7. Significance of codes	One (10%) participant did not understand some names of drugs because the name of the medication was not written next to the active substance	Find medication information (1)
4	1. Guidance	One (10%) user did not find the medication information in the principal menu	Find medication information (1)
5	5.2 Error message quality	When entering the date, two (20%) participants did not write it correctly (separated by "/" symbols and not "." symbols). Therefore, the date could not be validated, and no explanation appeared	Enter weight (2)
6	5.2 Error message quality	One (10%) participant made a mistake while entering his vital signs and could not go back, so he had to restart the task from the beginning	Enter vital signs (1)
7	2.2 Concision	One (10%) participant was troubled by the fact that there are two places to find information on drugs	Find medication information (1)

Satisfaction: SUS Questionnaire

The mean SUS score of the second stage of the study was 80 (SD 28.3), described in Table 5 and classified in Figure 16.

Table 5. Responses to the System Usability Scale (SUS) questionnaire in the second stage of the study.

	Question 1	Question 2	Question 3	Question 4	Question 5	Question 6	Question 7	Question 8	Question 9	Question 10	SUS score (sum×2.5; maximum 100)
Participant 1	3	3	3	4	3	3	3	3	3	3	77.5
Participant 2	3	0	3	3	3	4	2	3	3	0	60
Participant 3	3	4	4	4	4	4	4	4	4	4	97.5
Participant 4	3	4	3	4	4	4	4	4	4	3	92.5
Participant 5	4	4	4	4	4	4	4	4	4	4	100
Participant 6	4	3	3	3	3	3	3	3	3	3	77.5
Participant 7	3	4	4	4	3	4	3	4	4	4	92.5
Participant 8	3	4	3	4	3	4	3	4	3	4	87.5
Participant 9	4	3	3	4	3	4	3	4	4	1	82.5
Participant 10	3	1	1	1	3	3	3	2	3	1	52.5
Mean (SD)	3.3 (0.47)	3 (1.34)	3.1 (0.83)	3.5 (0.93)	3.3 (0.46)	3.7 (0.5)	3.2 (0.6)	3.5 (0.68)	3.5 (0.52)	2.7 (1.52)	80 (28.3)
Median (IQR)	3 (3.75-3)	3.5 (4-3)	3 (3.75-3)	4 (4-3.25)	3 (3.75-3)	4 (4-3.25)	3 (3.75-3)	4 (4-3)	3 .5 (4-3)	3 (4-1.5)	80 (90-70)

XSL•FO
RenderX

Figure 16. System Usability Scale (SUS) score on the SUS [35] of the Cardiomeds app at stage 2.

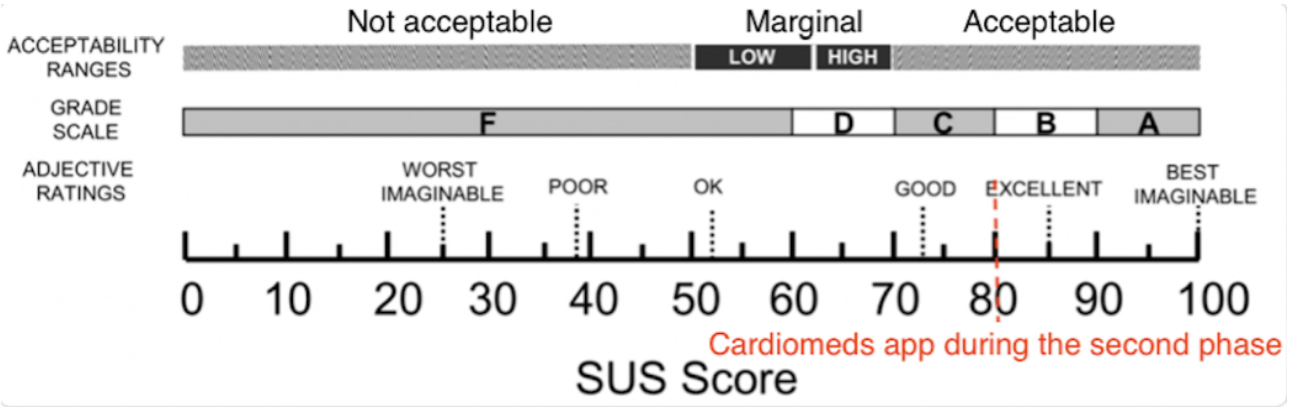

Comparison of First and Second Stages

Quantitative Evaluation

Task Success

On average, participants completed 86% (12/14; SD 0.19%) of tasks without any help overall. Conversely, on average, participants needed help with 14% (SD 0.11%) of tasks. As depicted in Figure 15, tasks 4, 10, and 14 maintained a success rate of 100% (20/20) in both stages of the study. Tasks 2, 3, 7, 8, 9, and 12 had a successful completion rate without any help of 100% in the second stage of the study. We also noticed that the failure rate for all tasks combined was reduced to 0% in the second stage of the study. All these measurements are exposed and supported in Figure 17.

Figure 17. Task success distribution per task (N=20 participants) for the overall study. "Completed" represents the percentage of participants who completed the task with ease. "Partially completed" represents the percentage of participants who completed the task with difficulty that could have been solved with the help of the instructor in 5 minutes or less. "Failed to complete" is defined as the percentage of participants who failed to complete the task. "Nonavailable" represents the percentage of missing data when a task could not be started and evaluated.

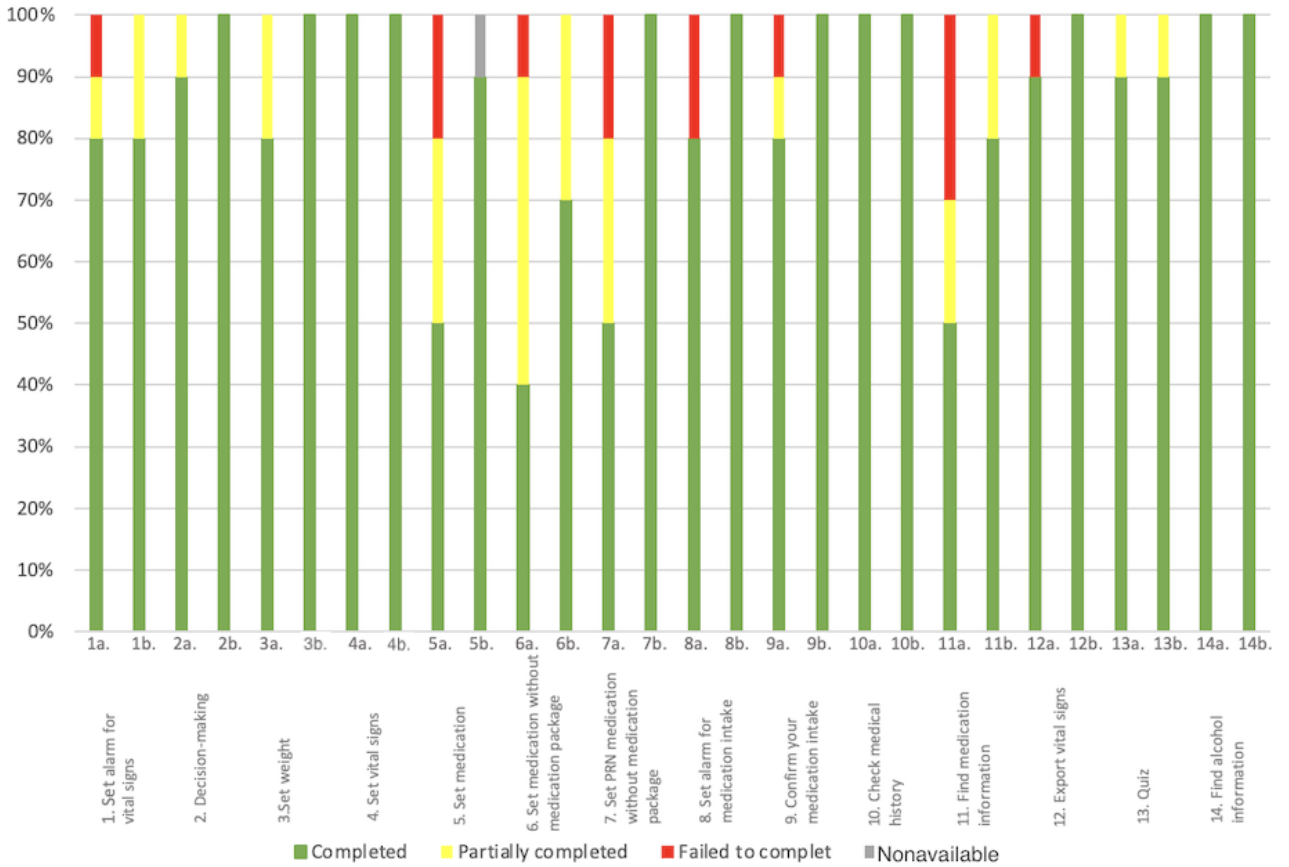

Time on Task per Study Task

The mean overall time on task for all tasks was 57.19 (SD 40.05) seconds. Tasks 5, 6, and 7 had a longer duration than the other tasks in both stages of the study (Figure 18). Nevertheless, task 11 was also one of the longest during the first stage. These findings showed that the task with the most failures (ie, task 6 [enter medication without medication package]) was also the

XSL•FO
RenderX

most time-consuming in both stages of the study, although the time taken was reduced during stage 2.

Figure 18. Boxplot of the comparison of the efficiency of the 2 stages of the study. Stage 1 is shown in blue and stage 2 is shown in orange.

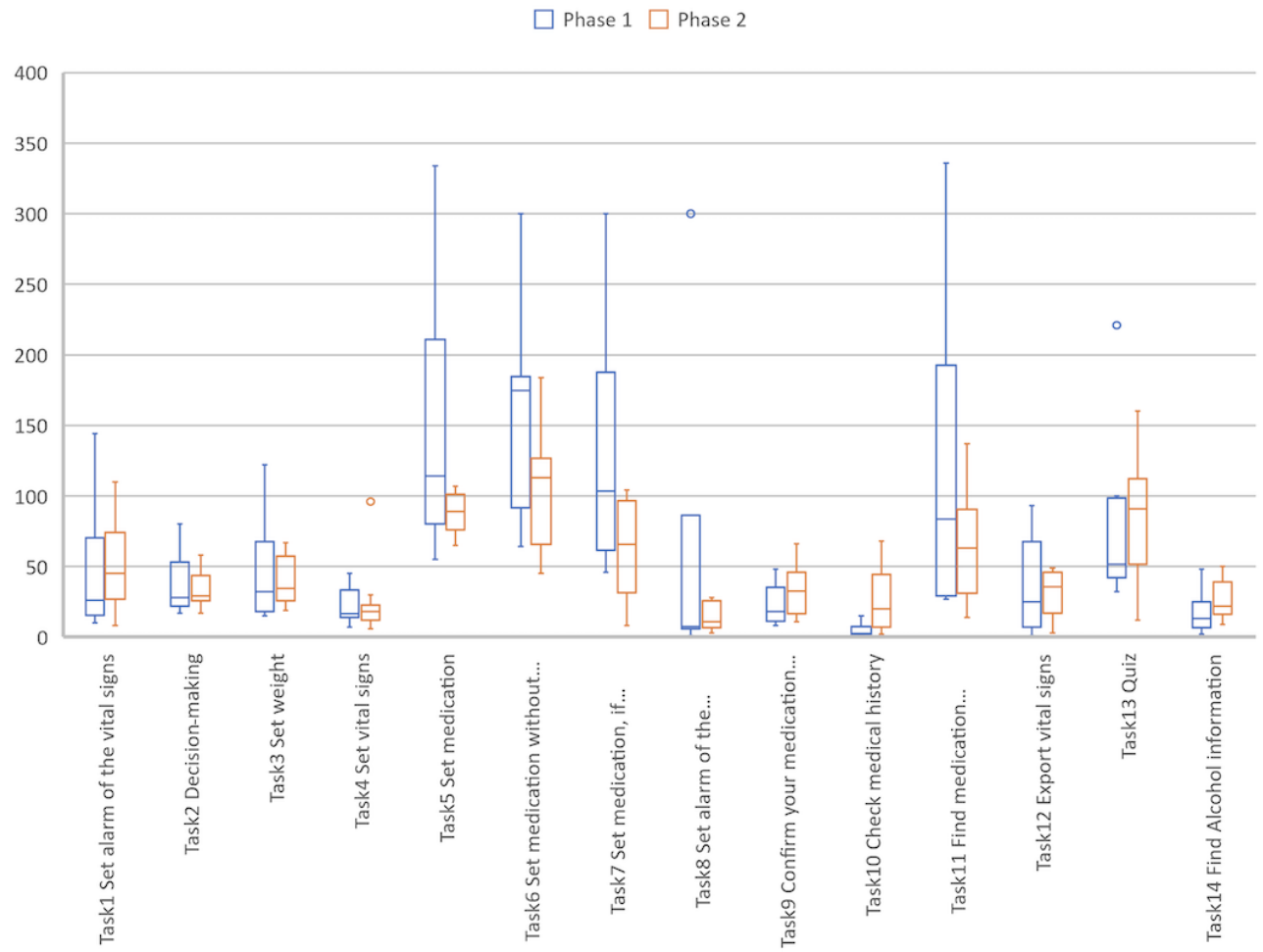

Qualitative Evaluation

Five Open Questions Questionnaire

Regarding the qualitative part of the study, 80% (16/20) of participants responded positively to our open questions about general feelings in both stages of the study. Between the 2 stages, the number of participants who had negative feelings was reduced from 1 participant to 0 participant. In addition, 85% (17/20) of participants were inclined to use the app once it became available, indicating its ease of use. Another comment that emerged in both stages of the study was the wish to access various contact numbers (4/20, 20%), such as the emergency, physician, cardiologist, or pharmacy.

Error Classification

The first stage of evaluation highlighted various ergonomic problems as defined by the Bastien and Scapin criteria [31]. The reduction, from 22 problems of 6 different types in the first stage of the study to 15 problems of 5 different types identified during the second stage, demonstrated the benefits of the improvements made to the app.

We observed that in both stages of the study, the most frequent problem remained the problem of guidance (first stage; 8/22, 36%, and second stage; 6/15, 40%).

It seems that the change in localization of the reminder and information section led to a reduction of guidance problems, although 2 different types of problems remained, including 1 new problem (Table 6). Given that guidance problems appeared several times in both stages of the study for tasks 8, 11, and 12, we could improve the app further by making the buttons of the reminder and the information section even more visible in a future version (reminder, export vital signs, and find medication information). A recommendation would be to add a label under the reminder button and increase its size. As for the button in the information section, we would like to change its color and increase its size.

XSL•FO
RenderX

Table 6. Type of problems during stage 1, number of occurrences, number of new problems, and total type of problems in stage 2.

Category (number according to the Bastien and Scapin [31] classification)	Stage 1	Stage 2		
	Types of problems	Solved problems	New problems	Types of problems
1. Guidance	2	1	1	2
5.2 Error message quality	1	0	1	2
7. Significance of codes	3	2	0	1
8. Compatibility	1	1	0	0
2.2 Information density	1	1	0	0
1.3 Immediate feedback	1	0	0	1
2.2 Concision	0	0	1	1

Moreover, we observed that the problem of the significance of codes decreased from 22% (5/22) to 7% (1/15) and from 3 such problems to only 1 (Table 6). These problems appeared during the task in which the participants had to scan the medication package, which we have made easier and more visible within the app, but also about the appellation by the component of certain medications that we have also tried to popularize as much as possible to reach a wider audience, for example, we have inserted the names of the drugs (metozerok) next to the class of the drug (β-blocker).

Problems related to error message quality (from 1/22, 5% to 3/15, 20%) and immediate feedback (from 1/22, 5% to 4/15, 26%) increased, although there was no new problem of the immediate feedback type, unlike the error message quality for which a new type of problem appeared. These problems did not increase significantly and are strongly dependent on the participant samples. Finally, a new problem of concision was mentioned during stage 2 by a single participant concerning the location of information concerning medications which could be found in 2 different places.

Satisfaction: SUS Questionnaire

Averaging the 2 results of the SUS questionnaire gave a score of 80 for the entire study. This shows that the usability of the Cardiomeds app was perceived as good in both stages of the study with a slight improvement during the second stage after some modifications. The scale in Figure 19 is used to interpret these results [34].

Figure 19. Cardiomeds score on the System Usability Scale (SUS) for the overall study.

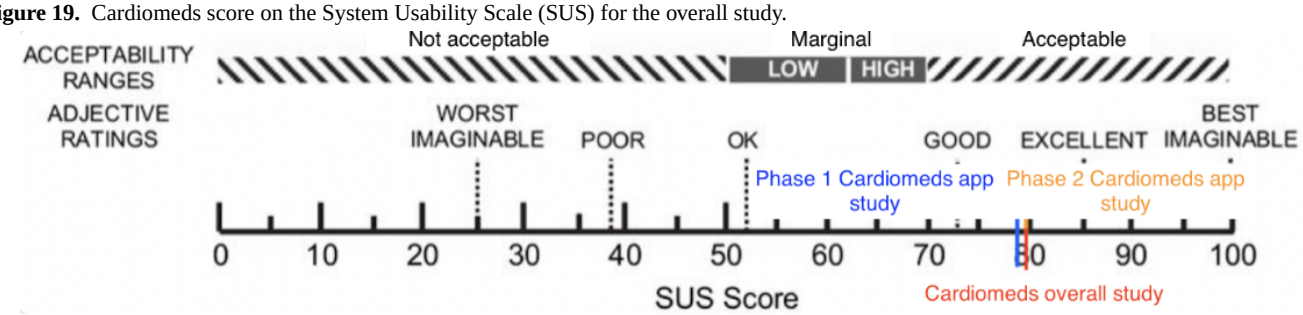

Discussion

Principal Findings

Participants took 18 minutes on average to complete all 14 tasks, which means just over 1 minute per task. However, certain tasks took much longer than others. The shortest task took an average of 4.80 (SD 4.60) seconds to be completed in the first stage (task 10) and 21.10 (SD 25.69) seconds in the second stage (task 8). One possible reason that tasks 8 and 10 were the shortest in the study could be that these 2 tasks required minimal navigation within the app. In addition, for task 8 (set alarm for medication intake), participants had already set an alarm for task 1 (set an alarm for vital signs), which likely made this task easier. Similarly, task 10 (check medication history) followed task 9 (confirm your medication intake), where the same menu also provided access to the medication intake history button. Finally, because both tasks followed task 6, which was the longest in the study, participants may have become more familiar with the app's navigation, helping them complete the later tasks more efficiently. Furthermore, between the 2 stages of the study, we modified the localization and visibility of the reminder icon to make it more accessible and more visible from the main menu. This might explain why task 8 (set alarm for the medication intake) was the fastest task. However, we were unable to determine why the duration of task 10 (check medical history) was longer in the second part of the study than the first because we did not make any change that could influence retrieval time.

The maximum average time was 154.40 (SD 68.08) seconds during the first stage (task 6) and 103.10 (SD 42.76) seconds during the second stage (task 6). The entry of medications (task 6) took the longest time in both stages of the study, although we observed an improvement between the first and the second stages, suggesting that the app had been improved. Improving the medication entry process was crucial to reduce the time patients spent inputting or modifying their treatment in the app. We know that patients with HF with reduced left ventricular

XSL•FO
RenderX

ejection fraction usually take at least 4 different medications. In addition, the older the patients, the more they suffer from different comorbidities, such as type 2 diabetes [35] or hypertension [36], which further increases the number of medications required. Furthermore, the treatment for patients with HF is likely to change often, so we wanted to make this part as fluid, ludic, and easy to use as possible. Knowing these considerations, the time spent on the app could significantly increase which could discourage the target audience from using it. Therefore, it was important to improve the ergonomics of drug recording to make this process as short as possible. Feedback collected on task 6 (enter medication without medication package) led to app modifications aimed at simplifying the medication entry process. The main change was to divide the drug recording process into 3 different pages rather than having a single scrollable page. Nevertheless, we believe that after repeated use, the time taken will diminish, and the patient will become increasingly comfortable to use the app. Also, although entering medications is an important initial process in using the app, it is not needed in daily use.

Concerning the quantitative results of the study, we observed that the mean overall success rate, which represents the effectiveness (completed the task without any help and completed the task with help) was 77.1% (108/140) for the first stage and 93.6% (131/140) for the second stage of the study. Furthermore, only 30% (3/10) of the participants in the first stage of the study managed to complete all tasks without any help versus 50% (5/10) during the second stage. We wonder if the differences observed between the 2 stages are partly due to the difference in characteristics of the participants in the 2 groups. Stage 1 included more participants aged between 71 and 80 years (1/10, 10%) than stage 2 (0/10, 0%).

We observed that 8 tasks (8/14, 57%) were failed by at least one participant during stage 1 compared to none during the second stage. The tasks failed or completed with difficulty in the first stage of the study included tasks 5 (5/10, 50%), 6 (6/10, 60%), and 7 (5/10, 50%). All these tasks were related to the recording of the medication. These same tasks induced fewer problems during the second stage of the study, with only 30% (3/10) of the participants completing task 6 with difficulty, and none of the participants making any errors for tasks 5 and 7. As these tasks aimed at the recording of medications, we can reasonably assume that the changes made to the app after the first stage contributed to the reduction of problems encountered during the second stage of the study.

Finally, we observe in Figure 19 that the efficiency of task 10 (check medical history) seems worse in the second stage of the study (mean time for the overall participants 27.10, SD 21.70 seconds) than in the first stage (mean time for the overall participants 4.80, SD 4.60 seconds). On the basis of our observations, we suspect that during the first stage of the study, participants took more time to complete the previous tasks, which gave them more time to navigate through the app. Thus, we hypothesize that the medication history page was already localized by the participants in stage 1.

It is surprising to observe that the significant increase in effectiveness per participant, from 50% (5/10) in the first stage of the study to 100% (10/10) in the second stage, is not reflected in an increased SUS score. Indeed, the Cardiomeds app was first evaluated with usability measured as “good” with a SUS score of 79% on average. It improved by 1% to reach 80% after modifications were made to the app between the 2 stages of the study. Nevertheless, when we look at the distribution of scores in the 2 groups in Figure 20, we see that they are very similar. We concluded that the participants mainly judged the concept of the app rather than the difficulty of the tasks carried out during the usability tests through the SUS. Therefore, it is logical that despite the changes made to the app, the SUS score did not change much.

Figure 20. System Usability Scale score distribution in both groups.

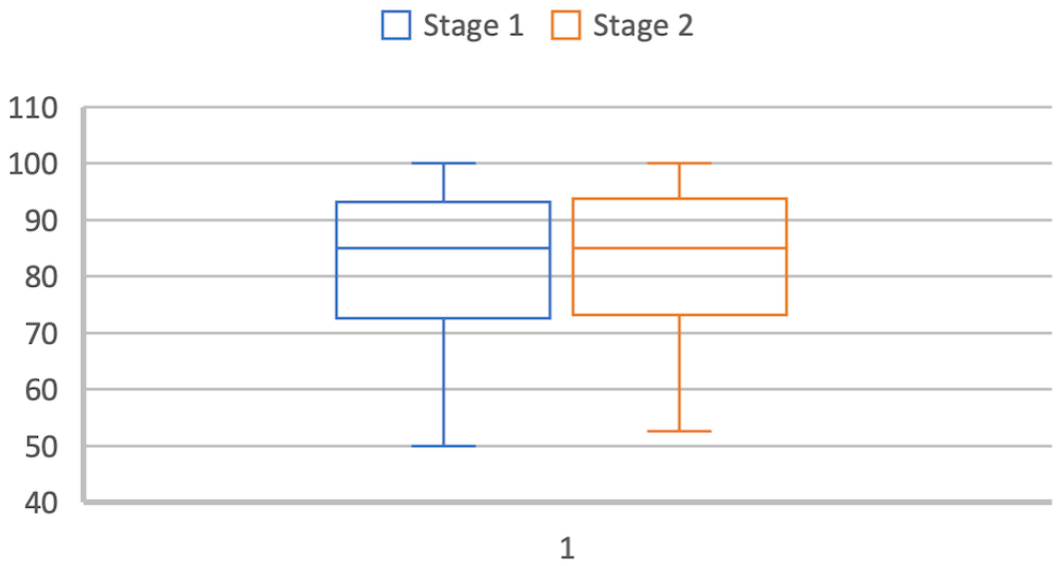

During the tests, we collected various participants’ comments. In the first stage of the study, the most common comment (2/10, 20%) was the extremely small font size, which may be a problem for older users. We did not change the font size in the app because it is possible to do so globally in the settings of

XSL•FO
RenderX

each device. During the second stage of the study, this remark did not emerge again, although we did not change the font size.

Furthermore, we can observe in Table 6 that between the 2 stages of the study, the problems of information density and compatibility were solved. Indeed, these problems mainly appeared while entering medications without a medication package (task 6), whose interface was improved between the 2 stages of the study. As mentioned earlier, this stage has been improved because we believe that the recording of medications must be simple given the number of treatments necessary and their regular change depending on the clinical condition of the patients and the course of their illness. We decided not to categorize guidance problems because it was not clear whether the guidance problems were localization problems or legibility problems. Indeed, the guidance problems were found in task 8 (set alarm for medication intake), 11 (find medication information), and 12 (export vital signs). It was not clear if the button was too small or poorly located, which would not make it visible enough.

It is important to point out that despite the availability of mHealth health interventions, most patients still wish to have follow-up appointments with their physician to maintain supervision and human contact. Many patients express the fear of being harmed when relying only on a health app [37]. However, with the current shortage of physicians and nurses in Switzerland's urban areas [38], the management of these patients is complex. The follow-up made possible with an app, such as Cardiomeds, could probably provide relief and help physicians to focus more on the relationship with the patient during medical appointments. Moreover, when patients are involved in their care and carry out self-checks, they are also more involved in their health, thus enabling in-depth monitoring of their illness to best prevent cardiac decompensations and avoid hospitalizations. Therefore, it is clear that physicians need to add eHealth to their daily practice by integrating it into their management, to offer optimal long-term follow-up.

Strengths

The first strong point of our study was that although common guidelines assume that a sample of 5 users is sufficient to reveal 85% of usability problems and 15 users is sufficient to discover almost 97% of problems [39], we evaluated the usability of the Cardiomeds app with 2 groups of 10 people each. In addition, we split the study into 2 stages, which enabled us to make modifications based on the feedback and then test the app with a new group to see if the changes improved the app's usability.

Second, we used mixed methods combined with usability testing methods to report and analyze quantitative and qualitative data in this study, which is recognized as a strong point in studies analyzing the usability of apps [40,41].

Finally, in our usability testing protocol, users had to complete tasks like those they would perform alone at home. Studies have shown that this is a good way to test an app's usability. In addition, the inclusion criteria for our study were broad, allowing a large panel of patients with HF to participate. In addition, even if most participants did not have an iPhone, it was still possible for Android users to test the app.

Limitations

A few limitations need to be considered. First, we can see that the most represented age group in our study (50-60 years) is younger than the age of the audience targeted by the app. Indeed, HF is more prevalent in patients >65 years of age [42]. Second, in both groups, women were underrepresented, with an average of 25% (5/20) of women, which is less than the prevalence of HF in women, which reaches 50% [42]. Third, only one investigator oversaw data collection in this study. Indeed, the usability tests, the SUS score, and the 5 open questions questionnaire were evaluated by one investigator, and all the recorded data were analyzed and timed by a single investigator. As only one person measured the times, there is a risk of bias in the results; however, this would be minimal given that the same investigator performed all the recordings.

Finally, the SUS questionnaire has limitations in assessing the usability of apps. Indeed, the SUS questionnaire does not provide specific insights into which parts of the app may cause issues for users. Moreover, it is a subjective measure of perceived usability. By adding some open questions to this questionnaire, we were able to obtain more information, but this does not reduce the inherent deficiencies of the SUS [43,44].

Conclusions

According to the SUS score obtained in both stages of the study, our usability study rates the Cardiomeds app as "good," with a slight improvement of 1% between the 2 stages of the study, to reach 80% in stage 2. The effectiveness per participant (the number of tasks completed or partially completed by each participant), went from 50% in stage 1% to 100% in stage 2. The most time-consuming task in both stages of the study was manual medication recording (task 6). Nevertheless, the improvement made after the first stage of the study enabled a reduction in the necessary time to complete this task. The most frequent ergonomic problems according to the Bastien and Scapin criteria remained problems of location and legibility within both stages of the study, with a slight improvement in stage 2. The problem of information density was solved because of the improvements made to this app by the IT professionals regarding medication recording. Most participants who tested the app recommended it and said they would use it when available. The multidisciplinary care required by patients with chronic HF is often complex because of the presence of many comorbidities. Therefore, it would be interesting to enable them to better understand their disease and increase their involvement, improve their therapeutic adherence, and finally enhance their quality of life through the future use of this mHealth app.

XSL•FO
RenderX

Acknowledgments

This study would not have been possible without the support of the University Hospital of Geneva. Funding for this usability study was provided by the cardiology department of the University Hospital of Geneva, and we thank the University of Geneva for their financial support.

We used the generative artificial intelligence tool, ChatGPT by OpenAI [45] to translate from French to English and to rephrase some sentences from the Introduction and Discussion of the study.

Conflicts of Interest

None declared.

References

1. Savarese G, Becher PM, Lund LH, Seferovic P, Rosano GM, Coats AJ. Global burden of heart failure: a comprehensive and updated review of epidemiology. Cardiovasc Res 2023 Jan 18;118(17):3272-3287 [FREE Full text] [doi: 10.1093/cvr/cvac013] [Medline: 35150240]
2. McDonagh TA, Metra M, Adamo M, Gardner RS, Baumbach A, Böhm M, ESC Scientific Document Group. 2021 ESC guidelines for the diagnosis and treatment of acute and chronic heart failure. Eur Heart J 2021 Sep 21;42(36):3599-3726 [FREE Full text] [doi: 10.1093/eurheartj/ehab368] [Medline: 34447992]
3. Barlow J, Wright C, Sheasby J, Turner A, Hainsworth J. Self-management approaches for people with chronic conditions: a review. Patient Educ Couns 2002;48(2):177-187 [FREE Full text] [doi: 10.1016/s0738-3991(02)00032-0] [Medline: 12401421]
4. Jonkman NH, Westland H, Groenwold RH, Ågren S, Anguita M, Blue L, et al. What are effective program characteristics of self-management interventions in patients with heart failure? An individual patient data meta-analysis. J Card Fail 2016 Nov;22(11):861-871 [FREE Full text] [doi: 10.1016/j.cardfail.2016.06.422] [Medline: 27374838]
5. Visseren FL, Mach F, Smulders YM, Carballo D, Koskinas KC, Bäck M, ESC National Cardiac Societies, ESC Scientific Document Group. 2021 ESC Guidelines on cardiovascular disease prevention in clinical practice. Eur Heart J 2021 Sep 07;42(34):3227-3337 [FREE Full text] [doi: 10.1093/eurheartj/ehab484] [Medline: 34458905]
6. Achury Saldaña DM, Gonzalez RA, Garcia A, Mariño A, Aponte L, Bohorquez WR. Evaluation of a mobile application for heart failure telemonitoring. Comput Inform Nurs 2021 May 17;39(11):764-771. [doi: 10.1097/CIN.0000000000000756] [Medline: 33993153]
7. Inglis SC, Clark RA, Dierckx R, Prieto-Merino D, Cleland JG. Structured telephone support or non-invasive telemonitoring for patients with heart failure. Cochrane Database Syst Rev 2015 Oct 31;2015(10):CD007228 [FREE Full text] [doi: 10.1002/14651858.CD007228.pub3] [Medline: 26517969]
8. McAlister FA, Stewart S, Ferrua S, McMurray JJ. Multidisciplinary strategies for the management of heart failure patients at high risk for admission: a systematic review of randomized trials. J Am Coll Cardiol 2004 Aug 18;44(4):810-819 [FREE Full text] [doi: 10.1016/j.jacc.2004.05.055] [Medline: 15312864]
9. Zhu Y, Gu X, Xu C. Effectiveness of telemedicine systems for adults with heart failure: a meta-analysis of randomized controlled trials. Heart Fail Rev 2020 Mar;25(2):231-243 [FREE Full text] [doi: 10.1007/s10741-019-09801-5] [Medline: 31197564]
10. Lefler LL, Rhoads SJ, Harris M, Funderburg AE, Lubin SA, Martel ID, et al. Evaluating the use of mobile health technology in older adults with heart failure: mixed-methods study. JMIR Aging 2018 Dec 04;1(2):e12178 [FREE Full text] [doi: 10.2196/12178] [Medline: 31518257]
11. Gandapur Y, Kianoush S, Kelli HM, Misra S, Urrea B, Blaha MJ, et al. The role of mHealth for improving medication adherence in patients with cardiovascular disease: a systematic review. Eur Heart J Qual Care Clin Outcomes 2016 Oct 01;2(4):237-244 [FREE Full text] [doi: 10.1093/ehjqcco/qcw018] [Medline: 29474713]
12. Heiney SP, Donevant SB, Arp Adams SA, Parker PD, Chen H, Levkoff S. A smartphone app for self-management of heart failure in older African Americans: feasibility and usability study. JMIR Aging 2020 Apr 03;3(1):e17142 [FREE Full text] [doi: 10.2196/17142] [Medline: 32242822]
13. Bakogiannis C, Tsarouchas A, Mouselimis D, Lazaridis C, Theofillogianakos E, Billis A, et al. A patient-oriented app (ThessHF) to improve self-care quality in heart failure: from evidence-based design to pilot study. JMIR Mhealth Uhealth 2021 Apr 13;9(4):e24271 [FREE Full text] [doi: 10.2196/24271] [Medline: 33847599]
14. Masterson Creber RM, Maurer MS, Reading M, Hiraldo G, Hickey KT, Iribarren S. Review and analysis of existing mobile phone apps to support heart failure symptom monitoring and self-care management using the mobile application rating scale (MARS). JMIR Mhealth Uhealth 2016 Jun 14;4(2):e74 [FREE Full text] [doi: 10.2196/mhealth.5882] [Medline: 27302310]
15. WOW ME 2000mg - heart failure self-management tool for patients and caregivers. CNET Download. URL: https://download.cnet.com/wow-me-2000mg-heart-failure-self-management-tool-for-patients-and-caregivers/3000-2129_4-76773751.html [accessed 2024-09-13]

XSL•FO
RenderX

16. WebMD: symptom checker. Google Play. URL: https://play.google.com/store/apps/details?id=com.webmd.android&hl=en [accessed 2024-09-13]
17. 8 smart mobile apps and gadgets for heart health. PCNA. 2018. URL: https://pcna.net/8-smart-mobile-apps-and-gadgets-for-heart-health/ [accessed 2024-09-13]
18. Mertens A, Brandl C, Miron-Shatz T, Schlick C, Neumann T, Kribben A, et al. A mobile application improves therapy-adherence rates in elderly patients undergoing rehabilitation: a crossover design study comparing documentation via iPad with paper-based control. Medicine (Baltimore) 2016 Sep;95(36):e4446 [FREE Full text] [doi: 10.1097/MD.0000000000004446] [Medline: 27603339]
19. Park JY, Li J, Howren A, Tsao NW, De Vera M. Mobile phone apps targeting medication adherence: quality assessment and content analysis of user reviews. JMIR Mhealth Uhealth 2019 Jan 31;7(1):e11919 [FREE Full text] [doi: 10.2196/11919] [Medline: 30702435]
20. Cajita MI, Gleason KT, Han HR. A systematic review of mHealth-based heart failure interventions. J Cardiovasc Nurs 2016;31(3):E10-E22 [FREE Full text] [doi: 10.1097/JCN.0000000000000305] [Medline: 26544175]
21. Bertolazzi A, Quaglia V, Bongelli R. Barriers and facilitators to health technology adoption by older adults with chronic diseases: an integrative systematic review. BMC Public Health 2024 Feb 16;24(1):506 [FREE Full text] [doi: 10.1186/s12889-024-18036-5] [Medline: 38365698]
22. Ahmad NA, Mat Ludin AF, Shahar S, Mohd Noah SA, Mohd Tohit N. Willingness, perceived barriers and motivators in adopting mobile applications for health-related interventions among older adults: a scoping review protocol. BMJ Open 2020 Mar 16;10(3):e033870 [FREE Full text] [doi: 10.1136/bmjopen-2019-033870] [Medline: 32184309]
23. Étude digital seniors. Pro Senectute Schweiz. URL: https://www.prosenectute.ch/fr/espace-pro/publications/etudes/digital-seniors.html [accessed 2023-03-01]
24. Trois Suisses sur quatre ont adopté le smartphone. Comparis. URL: https://fr.comparis.ch/comparis/press/medienmitteilungen/artikel/2016/telecom/smartphone-studie-2016/smartphone-verbreitungsstudie-2016 [accessed 2023-03-01]
25. Usability testing 101. Nielsen Norman Group. URL: https://www.nngroup.com/articles/usability-testing-101/ [accessed 2024-09-13]
26. Test d'utilisabilité: Réaliser un test d'utilisabilité (ou Test utilisateur) afin d'évaluer un logiciel ou une application. Usabilis: Conseil UX et Ergonomie Digitale. URL: https://www.usabilis.com/ressources/test-d-utilisabilite/ [accessed 2024-03-17]
27. Usability 101: introduction to usability. Nielsen Norman Group. URL: https://www.nngroup.com/articles/usability-101-introduction-to-usability/ [accessed 2024-03-17]
28. Ehrler F, Gschwind L, Meyer P, Blondon K. Swiss-Meds: an app fostering medication adherence of Swiss patient. Stud Health Technol Inform 2019;259:71-76. [Medline: 30923276]
29. Aubagna M. Comment utiliser SUS pour mesurer l'expérience utilisateur ? Skeepers. URL: https://skeepers.io/fr/blog/questionnaire-system-usability-scale-experience-client/ [accessed 2023-03-01]
30. System usability scale (SUS). U.S. General Services Administration. URL: https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html [accessed 2023-03-01]
31. Les critères Heuristiques de Bastien et Scapin. Le bloc-notes, UX & Design d'expérience utilisateur. URL: https://blocnotes.iergo.fr/concevoir/les-criteres-heuristiques-de-bastien-et-scapin/ [accessed 2024-03-17]
32. ISO 9241-11:1998(en) Ergonomic requirements for office work with visual display terminals (VDTs) — part 11: guidance on usability. International Organization for Standardization. URL: https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-1:v1:en [accessed 2024-03-17]
33. Ben Ramadan AA, Jackson-Thompson J, Schmaltz CL. Usability assessment of the Missouri cancer registry's published interactive mapping reports: round two. Online J Public Health Inform 2019;11(2):e3 [FREE Full text] [doi: 10.5210/ojphi.v11i2.9483] [Medline: 31632597]
34. Brooke J. SUS: a retrospective. J User Exp 2013;8(2):29-40 [FREE Full text]
35. Boyle L, Grainger R, Hall RM, Krebs JD. Use of and beliefs about mobile phone apps for diabetes self-management: surveys of people in a hospital diabetes clinic and diabetes health professionals in New Zealand. JMIR Mhealth Uhealth 2017 Jun 30;5(6):e85 [FREE Full text] [doi: 10.2196/mhealth.7263] [Medline: 28666975]
36. Kang H, Park HA. A mobile app for hypertension management based on clinical practice guidelines: development and deployment. JMIR Mhealth Uhealth 2016 Feb 02;4(1):e12 [FREE Full text] [doi: 10.2196/mhealth.4966] [Medline: 26839283]
37. Frederix I, Caiani EG, Dendale P, Anker S, Bax J, Böhm A, et al. ESC e-cardiology working group position paper: overcoming challenges in digital health implementation in cardiovascular medicine. Eur J Prev Cardiol 2019 Jul;26(11):1166-1177 [FREE Full text] [doi: 10.1177/2047487319832394] [Medline: 30917695]
38. Bernheim L, Balavoine JF. Pénurie de médecins en Suisse : mythe ou réalité ? Rev Med Suisse 2016;504:280-281 [FREE Full text] [doi: 10.53738/revmed.2016.12.504.0280]
39. Faulkner L. Beyond the five-user assumption: benefits of increased sample sizes in usability testing. Behav Res Methods Instrum Comput 2003 Aug;35(3):379-383 [FREE Full text] [doi: 10.3758/bf03195514] [Medline: 14587545]
40. Zapata BC, Fernández-Alemán JL, Idri A, Toval A. Empirical studies on usability of mHealth apps: a systematic literature review. J Med Syst 2015 Feb;39(2):1 [FREE Full text] [doi: 10.1007/s10916-014-0182-2] [Medline: 25600193]

XSL•FO
RenderX

41. Alwashmi MF, Hawboldt J, Davis E, Fetters MD. The iterative convergent design for mobile health usability testing: mixed methods approach. JMIR Mhealth Uhealth 2019 Apr 26;7(4):e11656 [FREE Full text] [doi: 10.2196/11656] [Medline: 31025951]
42. Shah RU, Klein L, Lloyd-Jones DM. Heart failure in women: epidemiology, biology and treatment. Womens Health (Lond) 2009 Sep;5(5):517-527 [FREE Full text] [doi: 10.2217/whe.09.50] [Medline: 19702451]
43. What every client should know about SUS scores. Bentley University. URL: https://www.bentley.edu/centers/user-experience-center/what-every-client-should-know-about-sus-scores [accessed 2024-09-12]
44. Pros and cons of the system usability scale (SUS). Research Collective. URL: https://research-collective.com/pros-and-cons-of-the-system-usability-scale-sus/ [accessed 2024-09-13]
45. ChatGPT. URL: https://chatgpt.com [accessed 2024-10-14]

Abbreviations

HF: heart failure
HUG: University Hospital of Geneva
mHealth: mobile health
SUS: System Usability Scale
TCR: task completion rate

Edited by A Mavragani; submitted 10.07.24; peer-reviewed by R Rabiei, J Xia; comments to author 30.08.24; revised version received 18.09.24; accepted 15.10.24; published 15.01.25.

Please cite as:
Simioni L, Tessitore E, Hagberg H, Schneider-Paccot A, Blondon K, Gschwind L, Meyer P, Ehrler F
Cardiomeds, an mHealth App for Self-Management to Support Swiss Patients With Heart Failure: 2-Stage Mixed Methods Usability Study
JMIR Form Res 2025;9:e63941
URL: https://formative.jmir.org/2025/1/e63941
doi:10.2196/63941
PMID:39813081

©Lisa Simioni, Elena Tessitore, Hamdi Hagberg, Aurélie Schneider-Paccot, Katherine Blondon, Liliane Gschwind, Philippe Meyer, Frederic Ehrler. Originally published in JMIR Formative Research (https://formative.jmir.org), 15.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Teaching in the Digital Age—Developing a Support Program for Nursing Education Providers: Design-Based Research

Stefan Walzer[1*], MSc; Carolin Barthel[2*], MSc; Ronja Pazouki[2], MSc; Helga Marx[3], MSc; Sven Ziegler[3], MSc; Peter Koenig[1], Prof Dr; Christiane Kugler[2], Prof Dr; Stefan Jobst[2], MSc

[1]Care and Technology Lab, Furtwangen University, Furtwangen im Schwarzwald, Germany
[2]Institute of Nursing Science, Faculty of Medicine, University of Freiburg, Freiburg, Germany
[3]Nursing Direction, Medical Center, University of Freiburg, Freiburg, Germany
[*]these authors contributed equally

Corresponding Author:
Stefan Walzer, MSc
Care and Technology Lab
Furtwangen University
Robert-Gerwig-Platz 1
Furtwangen im Schwarzwald, 78120
Germany
Phone: 49 7723 920 2957
Email: stefan.walzer@hs-furtwangen.de

Abstract

Background: Health care systems and the nursing profession worldwide are being transformed by technology and digitalization. Nurses acquire digital competence through their own experience in daily practice, but also from education and training; nursing education providers thus play an important role. While nursing education providers have some level of digital competence, there is a need for ongoing training and support for them to develop more advanced skills and effectively integrate technology into their teaching.

Objective: This study aims to develop a needs-based support program for nursing education providers to foster digital competencies and to test this intervention.

Methods: We used a design-based research approach, incorporating iterative development with expert consultation to create and evaluate a support program for nursing education providers. Focus groups were conducted online to assess needs, and thematic content analysis was used to derive key insights. The support program was then refined through expert feedback and subjected to a feasibility and satisfaction test, with participant evaluations analyzed descriptively.

Results: Six main categories emerged from the focus groups, highlighting key areas, including the use of digital technology, ongoing support needs, and the current state of digitalization in nursing education. The support program was developed based on these findings, with expert validation leading to adjustments in timing, content prioritization, and platform integration. Preliminary testing showed good overall satisfaction with the support program, although participants suggested improvements in content relevance and digital platform usability.

Conclusions: Although the feasibility test showed high satisfaction with the support program, low participation rates and limited perceived knowledge gain were major concerns. The results suggest that while the program was well received, further refinements, including a focus on competency-based approaches and addressing workplace barriers, are needed to increase participation and effectiveness of such interventions. The findings of this research can be used as a basis for the development of similar programs in other educational and health care contexts.

(JMIR Form Res 2025;9:e66109) doi:10.2196/66109

KEYWORDS

digital competencies; nursing education; support program; needs assessment; design-based research; feasibility study; nursing education provider; qualitative research; nurse; health care; focus group; digital age; expert consultation; thematic content analysis; feasibility test; satisfaction; competency-based approach; workplace barrier; health care digitalization; digital technology

XSL•FO
RenderX

Introduction

Background

Technology and digitalization are transforming health care systems and the nursing profession globally [1]. The range of modern technologies used in health care practice is broad, with varying levels of complexity and includes, among others, electronic health records, sensors, tracking, artificial intelligence, and robotics [2,3]. Digital competence consists of knowledge, attitudes, and skills "required to use new technologies in a meaningful way and as a tool for learning, working and leisure time, understanding the essential phenomena of digital technologies in society as well as in one's own life, and the motivation to participate in the digital world as an active and responsible actor" (pp. 670-671) [4].

Digital competence for nurses is acquired through their own experience in daily practice, but also in continuing education and training, where the providers of this education play an important role [5,6]. In Germany, nursing is an apprenticeship-based profession. Educational activities during this apprenticeship are performed by nurse educators, who provide theoretical education to nursing students, and clinical mentors, who provide supervision and mentoring of trainees, students, or new professionals; these roles are referred to together as "nursing education providers" in this paper. The key aspects of the training and the characteristics of these job designations are summarized in Multimedia Appendix 1. Nursing education providers are responsible for designing and delivering educational programs that include the use of digital technologies [7]. These responsibilities involve developing digital teaching materials and resources for use in the classroom or online, in addition to providing hands-on training and support to help students or new professionals develop their digital competencies [8]. There are already several digital technologies being used in theoretical and practical nursing education, for example, laptops, headphones, and video cameras for online or hybrid teaching, digital blackboards, simulation dolls, and digital platforms [9]. The use of these technologies requires that nursing education providers are capable of imparting the necessary competencies and that they are also able to use digital technologies themselves through the demonstration of pedagogical digital competence in their work [10].

Internationally, several studies have investigated the evidence regarding the digital competence of nursing education providers. A study by Forman et al [11] showed barriers against the implementation of technology in teaching, as well as the need for better support for nursing education providers who want to use technology in educational activities. Furthermore, they pointed out a missing consensus about the definition of minimum levels of digital competence [11]. Similarly, Männistö et al [8] described competence as a multifaceted phenomenon, which depends on various factors. It is proven by multiple studies that a lack of knowledge regarding technology among nursing education providers leads to worse outcomes in the preparation of nursing students [12].

A survey of 169 nursing education providers in a German medical center assessed their self-perceived competence in using digital technologies for teaching [13]. The results showed that nurses generally feel competent and positive about using digital technology in educational activities. While a basic setup of digital teaching tools is available, respondents expressed a need for further training in digital competence. The study's quantitative approach provided insights into digital competence, but the authors recommend qualitative methods to deepen understanding and develop practical recommendations.

Overall, the evidence suggests that while nursing education providers have some level of digital competence, there is a need for ongoing training and support to help them develop more advanced skills and effectively incorporate technology into their teaching. Although the topic is attracting increasing international interest, there still seem to be no established training programs focused on digitalization in nursing education in several countries, including Germany [14].

Aim

The aims of this study were (1) to develop a needs-based support program for nursing education providers to foster digital competencies in a specific context at a university hospital in Germany and (2) to test this program regarding feasibility and acceptance.

Methods

Overview

This research report was written according to the GUIDED (Guidance for Reporting Intervention Development Studies in Health Research) [15] and the TIDieR (Template for Intervention Description and Replication) checklists [16]. Both checklists can be found in Multimedia Appendices 2 and 3.

Design

The design of this study was based on the design-based research methodology as described by Anderson and Shattuck [17]. Design-based research is an iterative approach that aims to improve educational practices through the design, implementation, and refinement of interventions in real-world settings [17]. This methodology emphasizes stakeholder collaboration, iterative refinement, and empirical testing, making it well suited to the complexities of developing and testing a support program for nursing education providers [17] and has already proven effective for digital innovation in education [18]. Within design-based research methodology, the development of the support program followed a step-by-step, iterative, and participatory process with the consultative involvement of experts of the target group, based on the recommendations for the development of educational programs by Schneiderhahn et al [19], Schlutz [20], and the Medical Research Council framework for the development and evaluation of complex interventions in health care [21]. Schlutz [20] emphasizes the importance of developing educational services that respond to the specific needs of learners within their professional contexts. His focus on participatory approaches helps to ensure that the support program we develop is not only relevant but also tailored to the challenges faced by nursing education providers in the digital age. Similarly, Schneiderhan et al [19] present a systematic guide to curriculum development that emphasizes

XSL•FO
RenderX

the need to align educational objectives with learner needs and competencies. Their emphasis on stakeholder involvement and iterative revision is consistent with our design-based research methodology, enhancing the program's effectiveness and ensuring its relevance to nursing education. The Medical Research Council framework [20] further contributes to our study by providing a structured approach to the development and evaluation of complex interventions. It ensures that our support program is contextually appropriate, and incorporates iterative feedback and refinement throughout the development process.

This study was conducted and coordinated by a group of nurse researchers with expertise in the fields of digitization (SW and SZ), adult education (CB and HM), and intervention development (SJ). One group member (SJ) was already involved in the previous quantitative survey mentioned above [13].

Procedures

In the context of this study, the phases for curriculum development according to Schneiderhan et al [19] and Schlutz [20] were integrated into the process of design-based research for the detailed design of the learning situation. The phases and the iterative approach are illustrated in Figure 1 and are outlined in the following sections.

Figure 1. Support program development steps.

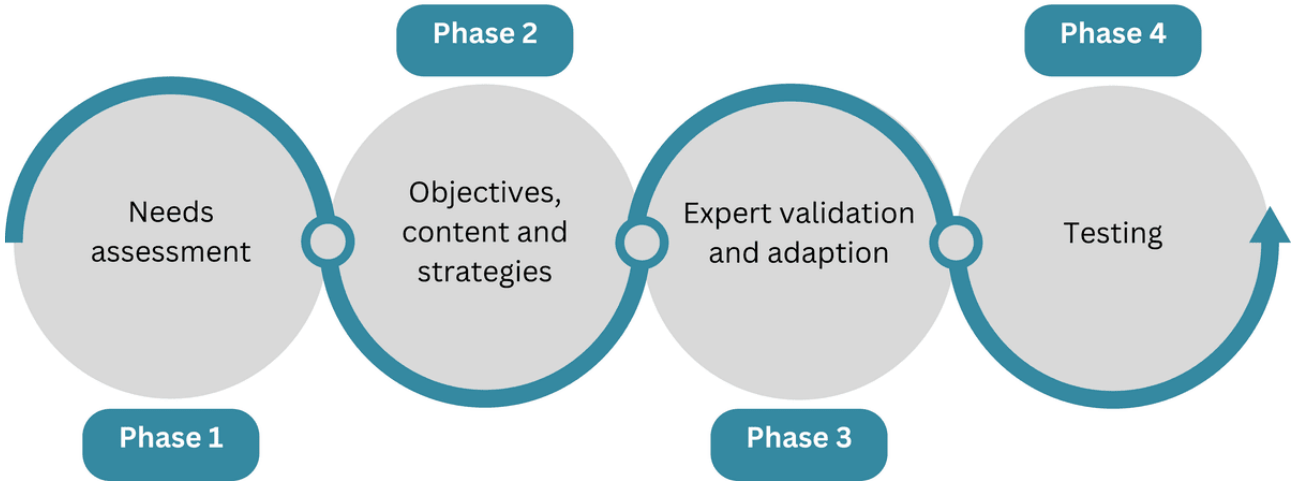

Phase 1: Needs Assessment

Overview

Based on an explanatory sequential design [20], this phase adopted a qualitative research approach using focus groups to assess the needs of nursing education providers based on the results of the mentioned quantitative survey [13]. Focus groups are a special type of group interview where a moderator leads a discussion focused on a particular subject [22]. They are useful for obtaining information about beliefs, attitudes, thoughts, and perceptions on a specific topic from multiple individuals in a short amount of time. Participant interaction within the group plays a vital role in generating the data [22-24].

An interview guideline was developed by the research group following the recommendations of Kallio et al [25]. A semistructured interview guideline was chosen because it focused on the concepts of interest while still allowing flexibility. Sequential data integration [26] was performed by using the structure and content of the guideline as a base for the structure of the written survey questionnaire used in the study of Jobst et al [13] and the core findings to be explored in greater depth. During the development phase, preliminary versions of the questionnaire went through several internal reviews and adjustments by all members of the research team. The themes of the final guideline were as follows: (1) existence and use of digital technologies; (2) assessment of digital competencies; (3) support; and (4) age, gender, and specialty area. Each topic area was introduced in advance by a brief presentation of corresponding results from the written survey to provide context for participants and to encourage them to discuss their experiences related to the topic of interest. Due to hygiene regulations (caused by the COVID-19 pandemic) at the time the study was conducted, focus groups were scheduled as online sessions. For this purpose, the videoconferencing service WebEx (Cisco) was used which operated via the secure servers of the institution. The research team programmed technical support if necessary. Although the members of the research team and potential participants belong to the same professional group, there is no overlap between their respective fields of work. Thus, it was assumed that there is no dependency relationship or conflict of interest between the research team and the participants.

Participants

Nursing education providers were determined as the target group for participation in this study, as they had specific knowledge and experience related to the objectives under investigation. The teaching activity in nursing education can be considered as a common characteristic of the participants, which facilitates the creation of a pleasant atmosphere for discussion and exchange of opinions, as well as promoting interaction between them. A purposive sampling approach [27] was used to identify and recruit nursing education providers working at the above-mentioned medical center and the associated nursing school. Prerequisites for participation in the focus groups were membership in the professional group of nursing education providers employed at the University Medical Center, being

XSL•FO
RenderX

aged at least 18 years, providing a signed consent form, and having to access to technology to use the videoconferencing service. Prior to starting the recruitment process, permission was obtained from the institution's staff council, which approved the involvement of nursing education providers. Potential participants were contacted via email using internal institutional mailing lists and asked to respond in writing if they were interested in participating in the study. Interested individuals received detailed written information about the study and participation, as well as a written informed consent form.

Data Collection

The two focus groups were led by two of the authors (SW and SJ), of whom one already had experience in conducting interviews. They were assisted by two other members of the research team who took field notes or asked supplementary questions. The focus group discussions were audio recorded and transcribed using a denaturalized approach in which idiosyncratic speech elements were omitted [28]. Field notes written during the discussions were merged and converted into a digital text format. Transcripts and field notes were pseudonymized prior to analysis. The participants received a demographic questionnaire prior to the focus group discussion.

Analysis

The qualitative data in textual form collected from the transcribed focus group discussions were analyzed using MAXQDA 2022 (VERBI Software). Content analysis using a deductive-inductive approach using the method of thematic qualitative text analysis was performed [29]. To analyze the data, a deductive categorization matrix was developed with four main categories based on the themes of the interview guideline. Content was assigned to these main categories and respective subcategories were inductively formed in the process. Codes that did not fit into the categorization matrix were combined into inductively formed new main categories and subcategories during the analysis process [30]. Prior to the coding process, the categorization matrix was pilot-tested by two members of the research team to check its fit and feasibility. Six coders divided into three groups sequentially coded the transcripts, each group building on the interpretive work of the previous ones. First, initial codes were developed in a process of condensation, followed by an abstraction process, where codes were then combined into categories that were allocated to the main categories of the predefined categorization matrix. After the end of this phase, coders met to discuss the resulting codes and categories, their classification in the categorization matrix, and any discrepancies, and to make modifications. Finally, summaries in narrative form were written for each category. Based on these summaries, thematic descriptions were created for the main categories [31]. For the development of the support program, content from categories that pointed out existing needs and requirements was used. The essential information contained in these descriptions is provided in the Results section of this report.

Phase 2: Objectives, Content, and Strategies

Overview

Phase 2 involved defining the content and objectives of the support service, as well as the format and methods with which the support service was to be implemented. The procedure as described in the model of Schneiderhan et al [19] was applied in an adapted form: steps 2 to 4—determining and prioritizing content, writing goals and objectives, and selecting teaching or educational strategies—were combined into one, following the example of Schlutz [20], as the content, objectives, and methods of a learning situation are highly interdependent.

The development of the support program has been guided by Bloom's [32] taxonomy, which categorizes learning into cognitive, affective, and social domains. To ensure support and acceptability of the support program by the targeted audience and key personnel (eg, nurse managers, leading clinical mentors, and school administration), current conditions, resources, and the results of both empirical studies (survey and focus groups) were also considered during the development of the support program.

Data Collection

The goals, objectives, and content were derived from the relevant results of the survey and the focus groups using sequential data integration [26]. The research group met regularly over a period of three months to plan the pilot implementation and testing of the support program. The following data sources were used: results of the web-based survey, results of the focus groups, experts within the institution, and knowledge of internal workflows at the hospital that would facilitate the testing of the support program.

Analysis

The research team produced a draft based on a matrix identifying the relevant components of a support program. This structure was linked to the steps of Schlutz [20]. It included information on the theoretical aspects of the support program [19,33], as well as information on practical implementation. Based on the results of the web-based survey and the focus groups, learning objectives were developed, which served as the basis for the content design of the support program. The analyses of the focus groups revealed the participants' learning needs for several topics. These were prioritized according to their importance and realizability. After that, topics were developed in a creative and iterative group process as frames for the organization of the content of the support program. Topics were operationalized using brief descriptions. At the same time, an educational strategy was defined in terms of format, teaching methods and media, quantity, and timing. Both, topic descriptions and educational strategies, were then entered into the matrix. Experts from the institution were identified and asked to act as potential lecturers at the planned times. Based on the subject areas and descriptions in the manual, they prepared short presentations according to the specified parameters (format, time). The content and methodology were agreed with the members of the research team.

XSL•FO
RenderX

Phase 3: Expert Validation and Adaptation

Overview

This phase included several feedback rounds in which the draft of the support program was presented to various experts from the target group. The feedback from the experts was intended to ensure the feasibility of the support service, as well as to increase the acceptance in the target group.

Participants

A total of five experts, all working at a University Medical Center, were consulted. They had the following expertise: adult education in health care, clinical mentoring, practical and theoretical training of nurses, or digitalization and e-learning.

Data Collection

The experts were interviewed in a total of four appointments, on three of the four appointments the experts were interviewed individually, on one appointment two clinical mentors were interviewed simultaneously. The interviews were conducted face to face and were structured as follows: first, the experts were presented with the draft of the support service, after which they were able to comment on it openly. Each meeting, and especially the feedback by the experts, was protocolled by a member of the research group.

Analysis

The protocols were anonymized; their contents were compiled and then analyzed using qualitative content analysis. Themes and categories resulting from this analysis were discussed within the research group and used for adjustments of the draft produced in phase 2 and finally expanded into a manual with detailed descriptions of all elements for implementing the support program.

Phase 4: Feasibility Testing (Feasibility and Acceptance)

Overview

In phase 4, the support program was implemented according to the manual to assess its feasibility and acceptance. Feasibility in this context refers to the extent to which the intervention can be practically implemented as intended. This includes factors such as the manageability of the program, the availability of necessary resources, and the ability of the target group to use the program as designed [33]. Acceptance was operationalized as participant satisfaction, meaning that the focus was on how well the program was received by users and whether it met their needs and expectations [33].

Prior to the implementation of the support program, experts from the institution were identified and asked to act as potential lecturers at the planned times. Based on the subject areas and descriptions in the manual, they prepared short presentations according to the specified parameters (format, time). The content and methodology were agreed with the members of the research team. A member of the research team was present at each session to provide technical support and moderation. Each short presentation was recorded with the consent of the lecturers so that it could be made available asynchronously at a later date. In order to comply with data protection regulations, the participants' cameras and microphones were deactivated during the short lecture. After the recording was finished, at the end of the lecture, the participants' cameras and microphones were activated so that they could take part in the discussion round.

Participants

All nursing education providers at the medical center had the opportunity to take part in the support program as part of the feasibility test. Gatekeepers forwarded an invitation to the target group via email with information on the support program. Participation in the support program and the survey at the end of phase 4 was voluntary.

Data Collection

A survey design was used to evaluate the support program afterward. A new questionnaire was developed by the research team especially for this study, as no suitable instrument was available in German. The development of the questionnaire was guided by the conceptualizations of Proctor et al [34], Sidani and Epstein [35], and integrated elements of the ABC-SAT questionnaire to assess affective, behavioral, and cognitive domains of satisfaction [36]. The questionnaire was designed as a web-based questionnaire comprising 18 questions on the domains of satisfaction, feasibility, and appropriateness, each of which could be rated using a 5-point Likert scale. Attendance at the relevant events was also requested and there was an option to enter a free text. The demographic information collected in the questionnaire was the participants' main area of work, age group, and gender. In addition, anonymized data on the number of participants during the sessions and the number of views of the video recordings of the sessions was collected by means of a protocol.

Analysis

Numerical data were analyzed descriptively using Microsoft Excel 2016. Textual data (comments of participants) were analyzed using qualitative content analysis.

Setting

The study was conducted at the University Medical Center Freiburg.

Ethical Considerations

The conceptualization and implementation of this study were based on the principles of the Declaration of Helsinki. By German law, survey studies with a focus on employees must be approved by the employee council at the respective institution. This body serves to fulfill ethical requirements and safeguard the rights of employees in this country. These responsibilities also refer to the protection of personal rights and data protection in the context of surveys. For this reason, a project description and information materials for potential participants were submitted to the relevant council for approval. The responsible committee of the employee council of the Freiburg University Medical Centre provided approval for this study in written format. Participation in the focus groups was voluntary, no personal data were collected, and anonymity was always maintained. All potential participants received written information on the study (reason for the study, objective,

XSL•FO
RenderX

processes, data protection), were informed about the decision of the employee council and its subcommittees, and had the opportunity to contact the investigators in case of questions at any time during the study. Informed consent to participate was assumed if individuals completed the questionnaire and were confirmed (by ticking a box) at the beginning of the questionnaire.

Results

The presentation of the results of the study follows the structure of the 4 steps of the development process of the support program.

Phase 1: Needs Assessment

Overview

Two focus group sessions with a total of 7 participants (clinical mentors: n=4; nurse educators: n=3) took place in April and May 2022. The first focus group discussion lasted 1 hour and 15 minutes, the second focus group discussion lasted 56 minutes. The characteristics of the participants are provided in Table 1.

Table 1. Sample characteristics.

Characteristics	Participants (N=7), n (%)
Job designation	
Clinical mentor	4 (57)
Nurse educator	3 (43)
Sex	
Female	4 (57)
Male	3 (43)
Age (in years)	
18-35	2 (29)
36-49	3 (43)
≥50	2 (29)

Six main categories emerged from the data, each with several subcategories: (1) existence and use of digital technologies; (2) assessment of (own) digital competencies; (3) support; (4) age, gender, and field of specialization; (5) attitude toward new technologies; and (6) current state of digitization. The last two main categories were derived through an inductive approach from the qualitative data, while the remaining main categories were derived from the interview guide. The development of the support program was based on the analysis of the relevant categories (1), (3), and (6). An overview of these categories and their subcategories is presented in Textbox 1.

Textbox 1. Categorization matrix of the content analysis of the focus groups with relevant content used for the development of the support program.

Main category 1: existence and use of digital technologies (deductive category)

- Existing and used technologies
- Existing and unused technologies
- Desired technologies
- Expectations of technologies

Main category 2: support (deductive category)

- Content-related aspects
- Organization and design of support
- Ongoing input and support
- Experiences with existing support

Main category 3: current state of digitalization (inductive category)

- Best practice
- Opportunities
- Poor practice
- Barriers

XSL•FO
RenderX

Existence and Use of Digital Technologies

This main category covers various aspects related to the use of technology in educational activities. Focus group discussions revealed a wide range of technologies that were already being used, such as digital blackboards, digital learning platforms, laptops, and software for online meetings. However, some existing technologies, like smart boards, were not used, or not used optimally due to a lack of knowledge or partially nonfunctional hardware:

> *Actually, the other parts [smart boards] are essentially used like whiteboards for writing on. You write on them and then wipe it away. And that approach is not really what one could do.* [Focus group 1, nurse educator]

Participants expressed a desire for better equipment in the rooms for hybrid teaching, including more laptops, tablets, simulation manikins, and robots. Specific expectations for technologies used in education were also mentioned. Participants want technology to be useful and have a reasonable cost-benefit ratio. Furthermore, it should increase students' interest and motivation and support the learning process. Additionally, the use of technology should be easy and accessible for collaborative work. The participants also mentioned the benefits of technologies that allow for the easy integration of media into teaching and the documentation of examination or instruction situations.

Support

Four subcategories emerged from the data as a result of the discussion about support. This category included participants' desire for ongoing input and support, their experience with existing support, as well as their ideas, wishes, and needs related to content, organizational, and design aspects of (new) support.

Several participants were already aware of support programs and training on new technologies provided by their institution and had taken advantage of it in individual cases. In this context, even small gains in knowledge were of great importance to the participants. Among other things, access to support programs through a complicated registration process or limited scheduling was viewed critically. Online formats and purely written information were perceived as challenging. The lack of practical exercises was also highlighted as a criticism. In addition, uneven levels of difficulty and limited applicability were named.

It was mentioned that support for nursing education providers during the implementation of new technologies in teaching must be oriented toward different elements, that is, inspiration and support for the correct use of technology in teaching, as well as training in the appropriate use of these technologies. In particular, opportunities for practice were desired. When organizing support, special attention should be paid to interprofessional group formation, different levels of competence, and the requirements of the particular professional groups in regard to new technologies:

> *I often had the feeling that meanwhile there was a huge gap. Either I get a very marginal introduction ... Or I have something high-end, which is not user-related.* [Focus group 1, nurse educator]

In addition, participants expressed a desire to contact persons who can provide support in the event of technical problems. Further training should also be programmed on an ongoing basis in addition to regular refresher training. Participants pointed out the challenge of reconciling a busy schedule with training on new technologies.

Current State of Digitalization

All participants gave examples and anecdotes of the current state of digitalization and related competencies in the institutions in which they work. In line with the two subcategories, the reports of participants can be divided into "best practice" examples and opportunities, as well as "poor practice" examples and barriers.

In regards to "best practice," the use of online platforms for communication and information exchange was reported. Other positive aspects mentioned were the training opportunities, internal staff with responsibility for training in certain technologies, and support and mutual collegial support in case of technical problems.

> *...for example, recently I did a video call into the dialysis unit. And I did the lesson with the colleague via webinar together. Something like that. It all works. That's the advantage. Thanks to Corona.* [Focus group 1, nurse educator]

Opportunities for improvement included the need for further development of existing training and high-quality hardware equipment, as well as expanding the monthly work-group meeting to include technical aspects. Participants noted that in order to increase motivation to use new technologies, a recognizable added value must be evident along with comprehensible introductions. In the case of the latter, one participant would like to see better training methods.

Clinical mentors reported examples of poor practice, such as the lack of simulation dolls due to high costs, as well as a lack of training opportunities and responsibilities, and underutilization of existing software despite having the knowledge for an appropriate use. Nursing education providers discussed inadequate use of digital blackboards, limited dates for training, and insufficient differentiation based on competence levels in the training programed. All participants identified common barriers to the use of new technologies, such as high costs, inadequate equipment, lack of time, and lack of practical training space.

> *Because when I do hybrid teaching with a laptop that is on the table in front, I then occasionally run over and ask if he [a student, note by the authors] can hear and see everything, which of course he cannot...* [Focus group 11, nurse educator]

Participants also noted individual barriers, such as negative attitudes and lack of experience among colleagues, as well as the challenge of implementing technical solutions into large companies or hospitals due to existing federalist structures. They highlighted the organizational effort of implementing technology, particularly in terms of adapting teaching methods. Additionally, the design and delivery of general IT training from instructors without a background in teaching or nursing

XSL•FO
RenderX

was identified as a barrier as the sessions failed to address the needs of the participants.

The study revealed that while a range of tools such as digital blackboards and online platforms are used, some technologies are underused due to limited knowledge or nonfunctional hardware. Participants expressed a need for better equipment, ongoing training, and support, particularly in practical applications of technology and in overcoming barriers such as high costs, inadequate time, and organizational challenges. They emphasized the importance of tailored, accessible support that considers different competence levels and professional needs. Additionally, participants noted both positive and negative experiences with digitalization in their institutions, highlighting the potential for improvement in training and infrastructure.

Phase 2: Objectives, Content, and Strategies

The following three learning objectives were formulated for the support program, which guided the development process:

- Raising awareness of the topic of digitalization in one's own field of work
- Promoting motivation to use technologies in nursing education
- Creating opportunities for networking between nursing education providers from different areas of work

In line with Bloom et al [32], participants are expected to develop competencies within the cognitive domain, by gaining essential knowledge about digitalization in nursing education and gaining an understanding of how to apply digital tools to enhance teaching practices. They are also anticipated to cultivate competence within the affective domain by fostering positive attitudes toward digital technologies and increasing their motivation to engage with these tools. Finally, social competence is expected to be strengthened through collaboration and critical engagement with digital solutions, enabling participants to contribute meaningfully to professional discourse on digitalization in nursing education. Based on the objectives, initial topic descriptions were developed for the support program. The draft of the newly developed support program consisted of a series of short online meetings (ie, sessions), which were planned to be programmed once a week over a period of five weeks via the videoconferencing program WebEx, each lasting 15-20 minutes. Each session would focus on one of the topics. The session would start with a topic-centered keynote speech (max 10 min) by a lecturer with expertise in the respective topic followed by a moderated discussion round with the opportunity to ask questions and exchange views on the topic. Every session would be recorded and would later be made available on an institutional digital learning platform. Participants would thus have the option of attending directly or asynchronously. In addition, a forum would be set up on the learning platform to enable participants to interact between sessions.

Phase 3: Expert Validation and Adaptation

The experts provided feedback on the following aspects of the support program: framework conditions, themes, format, timing, motivation, and potential obstacles. The draft of the support program was generally rated favorably by the experts surveyed. The online format, which does not require registration, the use of software already familiar to the institution, the interactivity of the program (forum and discussion group), and the possibility of addressing questions directly to the relevant experts were highlighted as particularly beneficial features. In particular, the provision of recorded sessions was emphasized, which allows great flexibility for the target group. According to the experts, fixed dates and the short duration of the sessions would enhance motivation and predictability of participation. Experts noted that scheduling the sessions between 1:30 PM and 2 PM would ensure the largest possible number of potential participants could be reached. In addition, experts recommended addressing opportunities for digital collaboration between nurse educators and clinical mentors, some of which are unknown. Based on this feedback, the timing of the program was adjusted accordingly and the topic "Collaboration on a digital learning platform" was added to the lecture series.

Finally, the following five topic descriptions were developed for the support program:

- New nursing technologies at the University Medical Center Freiburg
- Collaboration on a digital learning platform
- Further training programs or further training catalogs for teaching nurses
- Nuts and bolts of Cisco WebEx
- Overview of functions of a locally available online training platform

The manual for the support program was updated accordingly and the anchoring of the support program on the institutional digital teaching-learning platform was finalized.

Phase 4: Testing (Feasibility and Acceptance)

A total of 25 people enrolled in the course for the support program. Attendance at the four keynote presentations varied, with the second session (collaboration on a digital learning platform) and fourth session (nuts and bolts of Cisco WebEx) being the most attended (n=2), and the third session (educational programs or catalog for teaching nurses) receiving no attendance. The video of the first keynote (new nursing technologies) was viewed most often (n=12), with declining views for every following keynote. With one exception, the testing was conducted as initially planned. The last keynote speech could not be held due to the illness of the lecturer. Overall, the evaluation of the support program involved seven participants, consisting of four male participants, two female participants, and one nonbinary person. Participants included 4 clinical mentors and 2 nurse educators, with one abstention. The median age range was 40-49 years.

Satisfaction with the support program was rated overall as good. Positive responses noted a supportive atmosphere during presentations, high instructor motivation, and well-prepared instructors. However, cognitive feedback indicated less perceived knowledge gain and mixed opinions on content relevance and session length adequacy. Regarding feasibility, participants found the format suitable but noted challenges in integrating it into their daily routines. The digital conference format was generally viewed as appropriate, but the forum and

XSL•FO
RenderX

platform usability were deemed less satisfactory. Open-text responses highlighted strengths such as competent speakers and efficient time management, while suggestions for improvement included enhancing educational content relevance and platform usability.

Discussion

Principal Findings

In this study, a design-based research methodology was used to develop and test a needs-based support program for nursing education providers to foster digital competencies. The study was conducted in four phases: phase 1 used focus groups to assess the current use of digital technologies and the state of digitalization in nursing education. In phase 2, the objectives, content, and strategies for a support program were developed based on previous survey results and the focus group insights, leading to three main learning objectives: raising awareness of digitalization, motivating the use of technology, and fostering networking among nursing educators. This phase was framed by Bloom et al's [32] taxonomy, which emphasizes the development of learning in the cognitive, affective, and social domains—each of which is crucial to the development of digital competencies. Phase 3 involved expert validation through multiple interviews, resulting in the refinement of 5 key topics and the finalization of the support program. Finally, phase 4 tested the feasibility and acceptance of the support program using a web-based questionnaire based on the ABC-SAT [36], which showed high satisfaction but low participation and insufficient perceived learning outcomes.

In several countries, including Germany, there are no formal training programs regarding digitalization for nursing education providers [14]. The feasibility test of this support program represents the first step toward the implementation of such a formal training program. This program can be classified as level 1—"reaction" in Kirkpatrick's model of the "four levels of training" [37]. This means that this program remains rather rudimentary, but has the potential to be expanded and developed further using additional resources. In the absence of a specific guideline for the development of such a support program, the design-based research methodology was used to structure and guide the process of this study. This proved to be extremely helpful, as the steps were comprehensible and the quality of the support program could be improved through iterative and participatory processes. Involving representatives of the target group in the development process brought several advantages: not only did it contribute to increasing the quality of the support program, it also ensured the feasibility of the support program and provided the possibility of easy and low-threshold contact and information for the target group about the planned support program.

The support program aimed to address the entire target group of nursing education providers, which was legitimized with the aim of promoting cooperation between the subgroups of nurse educators and clinical mentors. When designing the content of the support program, a topic-centered approach was taken and an attempt was made to find topics that would appeal to both subgroups of the target group. We chose this approach for several reasons. First, it allowed us to address the immediate interests and needs of both subgroups within our target audience. By focusing on engaging topics, we aimed to maximize initial interest and participation, which is crucial for the success of any educational program. Our goal was to create a stimulating environment where participants would be motivated to actively engage with the content. As a consequence of the topic-oriented approach, it was not possible to design a competence-oriented approach. Although this is not evident from the qualitative and quantitative survey data, there are indications that a competence-oriented design would have been more suitable for the support program. To address this, potential improvements could include introducing competence assessments and providing additional resources while applying a broader variety of learning methods and concepts. This would lead to more needs-oriented programs, especially for the respective target group, and among other things, result in a differentiation of the presented topics into different competence levels (eg, beginner, advanced, and expert) or specific addressing of one of the professional groups. As noted by Sillat et al [38], the increasing number of technologies, frameworks, and strategies underscores the need for assessing and evaluating digital competence in various settings. Given the challenges in self-assessment, which are not limited to nursing settings, the provision of a screening or assessment of digital competencies before choosing a training level should be considered.

The extent to which differentiation into different competence levels would have led to higher participation figures and greater satisfaction with the support program remains to be clarified. With regard to the number of participants, it is noticeable that the recordings of the keynote speeches, which were made available online asynchronously, had higher participation rates than the synchronous online presentations. One possible reason for this could be the stressful and sometimes unpredictable working day of nursing education providers, especially if they work in clinical practice. In addition, the respondents previously reported that they had hardly any time resources, which may be another reason for the low level of participation. This is also confirmed by Dornan [39] stating that employees with a care mandate generally prioritize these tasks higher than participation in continuing education courses. After or during a stressful working day, some individuals may lack the mental capacity to absorb further input in the form of training, as they may already be in cognitive overload [40]. Another reason could be a lack of motivation to participate. In this respect, it is known from adult education that the willingness and motivation to participate in further training depends very much on the general environment of the workplace [41].

These are, for example, the values and mission statement of the organization, as well as the demands and requirements placed on employees in the context of lifelong learning. Learning behavior and thus participation in further training is also strongly dependent on individual learning behavior, learning habits, and personal attitudes [41]. Noteworthy in this context are the different levels of involvement, ranging from enthusiastic to completely absent. While the stakeholders, who also belong to the target group of nursing education providers but tend to work less in direct training and increasingly take on management

XSL•FO
RenderX

tasks, were very interested and committed to the development of the support program, there was almost no interest at all on the part of the nurses being involved in direct patient care. We can only speculate about the reason for this great difference in motivation and participation. Since this research project was part of a large, 5-year project with various other research projects, one can also speculate about a kind of fatigue or desensitization to other research projects. Such a phenomenon has already been described by health care professionals [42].

The evaluation recorded only a few participants. However, this number is certainly justifiable and positive in the context of a feasibility test. The evaluation results show an ambivalent picture. In addition to the low number of participants, the results showed little to no increase in knowledge among the participants, yet they stated that they were generally satisfied with the program. A further more in-depth survey of the target group could provide clarity here and answer questions such as which people did or did not take part in the sessions and for what reasons, and whether a different format or a different time would have been more suitable. To address the low participation rate, measures such as enhanced recruitment strategies (eg, engagement of leadership) and flexible participation options (eg, live, asynchronous, and on-demand) should be considered to encourage higher involvement. Additionally, a focus on both intrinsic motivation (eg, emphasizing the value of lifelong learning, team culture, professional identity, and commitment to quality care) and extrinsic motivation (eg, incentives) should be incorporated [43].

Implications for Practice and Research

Based on the results of this study, several implications for practice can be derived. The next step should be to determine the reasons for participation or nonparticipation in the support program, for example, through a follow-up survey of the target group. This should also address obstacles and facilitating factors.

The support program developed here should also be further developed on the basis of design-based research. When further developing existing support programs or designing new ones, workplace-related factors such as workload, time resources, and working hours should be taken into account. It is also recommended—especially for complex and wide-ranging topics—to pursue a competence-oriented approach in order to best meet the needs of the target group.

Limitations

This study has several strengths and limitations that should be considered when interpreting the results. First, the sample of this study was relatively small and consisted of individuals from only one medical center. This sample may restrict the generalizability of the findings. However, the evaluation showed an acceptable number of participants for a feasibility study and the issues and findings raised in this study may be still relevant on a national and international level, especially for nursing education and training programs that prioritize the digital competencies of nursing education providers. In this study, the sample was highly specific which improves the credibility [44].

Second, despite a review of existing instruments, no suitable questionnaire could be found, as existing instruments were mostly too generic. Therefore, a self-created questionnaire had to be used for the evaluation, which was based on the ABC-SAT questionnaire [36] and conceptualizations of feasibility and acceptance, but was not validated in its final form.

Third, it is important to be aware of the potential bias of the researchers in this study, as data collection and analysis were conducted by researchers with a background in nursing education [45]. This situation may lead to a better understanding of the context but also affect the interpretation of the data. To compensate for these biases, triangulation was performed during the analysis.

Conclusions

The research design, consisting of a quantitative survey, a qualitative exploration, and subsequent development with participatory elements of a support program, facilitated both the structured collection of nursing education providers' views and experiences regarding digital competence and the targeted development of a needs-based support program. However, the likely willingness of the target audience to use the programming in the future appears to be severely limited, as reflected in the low attendance during the synchronous keynote presentations. The reasons for this poor response are varied and may be due to factors such as lack of information or insufficient motivation on the part of the target group. Overall satisfaction was high, but concerns were expressed about the lack of opportunities for professional discourse and individual knowledge gain. Despite these challenges, the overall satisfaction with the pilot implementation suggests that digital and innovative programs have potential and can serve as a basis for further development. Moreover, the insights gained from this study, despite being derived from a localized setting, could be applied to guide the development and implementation of similar support programs in diverse educational and health care contexts.

Acknowledgments

We thank Sadhbh Byrne for data processing, statistical consultation, and proofreading the manuscript; as well as the stakeholders for their support in designing the support program. This study was conducted at the Center of Implementing Nursing Care Innovations, Freiburg, Germany. This joint project on the development and research of new care technologies in Germany is funded by the Federal Ministry of Education and Research. At the Freiburg Center, the focus is on integrating and evaluating innovative nursing technologies in the acute care setting. This entails evaluating the competencies of the users, and if necessary, supporting their development (grants 16SV7887, 16SV7888, and 16SV7886K).

XSL•FO
RenderX

Data Availability

The data sets generated during and analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

Presentation of the authors' contributions followed the recommendations of the Contributor Roles Taxonomy (CRediT) [46]. SW was responsible for the conceptualization of the study, developed the methodology, conducted the investigation, performed formal analysis, curated the data, and wrote the original draft of the manuscript. CB contributed to the study's conceptualization and methodology, participated in the investigation, performed formal analysis, curated the data, and wrote the original draft. RP curated the data and contributed to the writing of the manuscript by reviewing and editing it. HM worked on the conceptualization and methodology of the study, conducted the investigation, and contributed to the writing and editing of the manuscript. SZ played a role in the conceptualization and methodology, conducted the investigation, and contributed to the writing and editing of the manuscript. PK contributed to the conceptualization of the study, provided resources, contributed to the writing and editing of the manuscript, and supervised the project. CK was involved in the conceptualization, provided resources, contributed to writing and editing, and supervised the project. SJ contributed to the conceptualization, provided resources, contributed to the writing and editing of the manuscript, and supervised the study. All authors reviewed and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Key aspects of the training and the characteristics of educating nurses in Germany.
[DOCX File , 16 KB - formative_v9i1e66109_app1.docx]

Multimedia Appendix 2
Guidance for reporting intervention development studies in health research (GUIDED) checklist.
[PDF File (Adobe PDF File), 780 KB - formative_v9i1e66109_app2.pdf]

Multimedia Appendix 3
Template for Intervention Description and Replication (TIDieR) checklist.
[PDF File (Adobe PDF File), 502 KB - formative_v9i1e66109_app3.pdf]

References

1. Rouleau G, Gagnon MP, Côté J, Payne-Gagnon J, Hudson E, Dubois CA. Impact of information and communication technologies on nursing care: results of an overview of systematic reviews. J Med Internet Res 2017;19(4):e122 [FREE Full text] [doi: 10.2196/jmir.6686] [Medline: 28442454]
2. Krick T, Huter K, Domhoff D, Schmidt A, Rothgang H, Wolf-Ostermann K. Digital technology and nursing care: a scoping review on acceptance, effectiveness and efficiency studies of informal and formal care technologies. BMC Health Serv Res 2019;19(1):400 [FREE Full text] [doi: 10.1186/s12913-019-4238-3] [Medline: 31221133]
3. Huter K, Krick T, Domhoff D, Seibert K, Wolf-Ostermann K, Rothgang H. Effectiveness of digital technologies to support nursing care: results of a scoping review. J Multidiscip Healthc 2020;13:1905-1926 [FREE Full text] [doi: 10.2147/JMDH.S286193] [Medline: 33328736]
4. Ilomäki L, Paavola S, Lakkala M, Kantosalo A. Digital competence—an emergent boundary concept for policy and educational research. Educ Inf Technol 2014;21(3):655-679. [doi: 10.1007/s10639-014-9346-4]
5. Kaufman DM. Applying educational theory in practice. ABC BMJ 2003;326(7382):213-216 [FREE Full text] [doi: 10.1136/bmj.326.7382.213] [Medline: 12543841]
6. Bierema LL. Adult learning in health professions education. New Dir Adult Contin Educ 2018;2018(157):27-40. [doi: 10.1002/ace.20266]
7. Nes AAG, Steindal SA, Larsen MH, Heer HC, Lærum-Onsager E, Gjevjon ER. Technological literacy in nursing education: a scoping review. J Prof Nurs 2021;37(2):320-334 [FREE Full text] [doi: 10.1016/j.profnurs.2021.01.008] [Medline: 33867086]
8. Männistö M, Mikkonen K, Kuivila HM, Koskinen C, Koivula M, Sjögren T, et al. Health and social care educators' competence in digital collaborative learning: a cross-sectional survey. Sage Open 2020;10(4):215824402096278. [doi: 10.1177/2158244020962780]
9. Grimwood T, Snell L. The use of technology in healthcare education: a literature review. MedEdPublish 2020;9(1):137. [doi: 10.15694/mep.2020.000137.1]

XSL•FO
RenderX

10. From J. Pedagogical digital competence—between values, knowledge and skills. High Educ Stud 2017;7(2):43 [FREE Full text]
11. Forman TM, Armor DA, Miller AS. A review of clinical informatics competencies in nursing to inform best practices in education and nurse faculty development. Nurs Educ Perspect 2020;41(1):E3-E7. [doi: 10.1097/01.NEP.0000000000000588] [Medline: 31860501]
12. Webb L, Clough J, O'Reilly D, Wilmott D, Witham G. The utility and impact of information communication technology (ICT) for pre-registration nurse education: a narrative synthesis systematic review. Nurse Educ Today 2017;48:160-171. [doi: 10.1016/j.nedt.2016.10.007] [Medline: 27816862]
13. Jobst S, Lindwedel U, Marx H, Pazouki R, Ziegler S, König P, et al. Competencies and needs of nurse educators and clinical mentors for teaching in the digital age—a multi-institutional, cross-sectional study. BMC Nurs 2022;21(1):240. [doi: 10.1186/s12912-022-01018-6] [Medline: 36031618]
14. Tischendorf T, Hasseler M, Schaal T, Ruppert SN, Marchwacka M, Heitmann-Möller A, et al. Developing digital competencies of nursing professionals in continuing education and training—a scoping review. Front Med (Lausanne) 2024;11:1358398 [FREE Full text] [doi: 10.3389/fmed.2024.1358398] [Medline: 38947234]
15. Duncan E, O'Cathain A, Rousseau N, Croot L, Sworn K, Turner KM, et al. Guidance for reporting intervention development studies in health research (GUIDED): an evidence-based consensus study. BMJ Open 2020;10(4):e033516 [FREE Full text] [doi: 10.1136/bmjopen-2019-033516] [Medline: 32273313]
16. Hoffmann TC, Glasziou PP, Boutron I, Milne R, Perera R, Moher D, et al. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. BMJ 2014;348:g1687 [FREE Full text] [doi: 10.1136/bmj.g1687] [Medline: 24609605]
17. Anderson T, Shattuck J. Design-based research: a decade of progress in education research? Educ Res 2012;41(1):16-25. [doi: 10.3102/0013189x11428813]
18. Hall T. Bridging practice and theory: the emerging potential of design-based research (DBR) for digital innovation in education. Educ Res Perspect 2020;47:157-173 [FREE Full text]
19. Schneiderhan J, Guetterman TC, Dobson ML. Curriculum development: a how to primer. Fam Med Community Health 2019;7(2):e000046 [FREE Full text] [doi: 10.1136/fmch-2018-000046] [Medline: 32148703]
20. Schlutz E. Bildungsdienstleistungen und Angebotsentwicklung. Germany: Waxmann Verlag; 2006.
21. Skivington K, Matthews L, Simpson SA, Craig P, Baird J, Blazeby JM, et al. A new framework for developing and evaluating complex interventions: update of medical research council guidance. BMJ 2021;374:n2061 [FREE Full text] [doi: 10.1136/bmj.n2061] [Medline: 34593508]
22. Plummer P. Focus group methodology. Part 1: design considerations. Int J Ther Rehabil 2017;24(7):297-301. [doi: 10.12968/ijtr.2017.24.7.297]
23. Doody O, Slevin E, Taggart L. Focus group interviews in nursing research: part 1. Br J Nurs 2013;22(1):16-19. [doi: 10.12968/bjon.2013.22.1.16] [Medline: 23299206]
24. Kuckartz U, Rädiker S. Analyzing Qualitative Data With MAXQDA: Text, Audio, and Video. New York, NY: Springer International Publishing; 2019.
25. Kallio H, Pietilä AM, Johnson M, Kangasniemi M. Systematic methodological review: developing a framework for a qualitative semi-structured interview guide. J Adv Nurs 2016;72(12):2954-2965. [doi: 10.1111/jan.13031] [Medline: 27221824]
26. Fetters MD, Curry LA, Creswell JW. Achieving integration in mixed methods designs-principles and practices. Health Serv Res 2013;48(6 Pt 2):2134-2156 [FREE Full text] [doi: 10.1111/1475-6773.12117] [Medline: 24279835]
27. Campbell S, Greenwood M, Prior S, Shearer T, Walkem K, Young S, et al. Purposive sampling: complex or simple? Research case examples. J Res Nurs 2020;25(8):652-661 [FREE Full text] [doi: 10.1177/1744987120927206] [Medline: 34394687]
28. Azevedo V, Carvalho M, Fernandes-Costa F, Mesquita S, Soares J, Teixeira F, et al. Interview transcription: conceptual issues, practical guidelines, and challenges. Rev Enf Ref 2017;IV Série(Nº14):159-168 [FREE Full text]
29. Graneheim UH, Lindgren B, Lundman B. Methodological challenges in qualitative content analysis: a discussion paper. Nurse Educ Today 2017;56:29-34. [doi: 10.1016/j.nedt.2017.06.002] [Medline: 28651100]
30. Elo S, Kyngäs H. The qualitative content analysis process. J Adv Nurs 2008;62(1):107-115. [doi: 10.1111/j.1365-2648.2007.04569.x] [Medline: 18352969]
31. Lindgren B, Lundman B, Graneheim UH. Abstraction and interpretation during the qualitative content analysis process. Int J Nurs Stud 2020;108:103632. [doi: 10.1016/j.ijnurstu.2020.103632] [Medline: 32505813]
32. Bloom BS, Engelhart MD, Furst EJ, Hill WH, Krathwohl DR. Taxonomy of Educational Objectives: The Classification of Educational Goals. Green, NY: Longmans; 1964.
33. Sidani S, Braden C. Nursing and Health Interventions: Design, Evaluation and Implementation. 2nd Edition. Hoboken, NJ: Wiley-Blackwell; 2021.
34. Proctor E, Silmere H, Raghavan R, Hovmand P, Aarons G, Bunger A, et al. Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda. Adm Policy Ment Health 2011;38(2):65-76 [FREE Full text] [doi: 10.1007/s10488-010-0319-7] [Medline: 20957426]

XSL•FO
RenderX

35. Sidani S, Epstein DR. Toward a conceptualization and operationalization of satisfaction with nonpharmacological interventions. Res Theory Nurs Pract 2016;30(3):242-257. [doi: 10.1891/1541-6577.30.3.242] [Medline: 28304269]
36. Meng M, Peter D, Mattner F, Igel C, Kugler C. Development and psychometric pilot-testing of a questionnaire for the evaluation of satisfaction with continuing education in infection control nurses. Nurse Educ Pract 2018;31:77-82. [doi: 10.1016/j.nepr.2018.05.003] [Medline: 29800763]
37. Kirkpatrick JD, Kirkpatrick WK. Kirkpatrick's Four Levels of Training Evaluation: Association for Talent Development. Alexandria, VA: ATD Press; 2016.
38. Sillat LH, Tammets K, Laanpere M. Digital competence assessment methods in higher education: a systematic literature review. Educ Sci 2021;11(8):402. [doi: 10.3390/educsci11080402]
39. Dornan T. Workplace learning. Perspect Med Educ 2012;1(1):15-23 [FREE Full text] [doi: 10.1007/s40037-012-0005-4] [Medline: 23316455]
40. McElhatton J, Drew C. Hurry-up syndrome. ASRS directline. 1993. URL: https://asrs.arc.nasa.gov/publications/directline/dl5_hurry.htm [accessed 2024-08-08]
41. West CP, Shanafelt TD. The influence of personal and environmental factors on professionalism in medical education. BMC Med Educ 2007;7:29 [FREE Full text] [doi: 10.1186/1472-6920-7-29] [Medline: 17760986]
42. Roxburgh M. An exploration of factors which constrain nurses from research participation. J Clin Nurs 2006;15(5):535-545. [doi: 10.1111/j.1365-2702.2006.01374.x] [Medline: 16629962]
43. Kalenda J, Kočvarová I. “Why don’t they participate?” Reasons for nonparticipation in adult learning and education from the viewpoint of self-determination theory. Eur J Res Educ Learn Adults 2022;13(2):193-208 [FREE Full text]
44. Malterud K, Siersma VD, Guassora AD. Sample size in qualitative interview studies: guided by information power. Qual Health Res 2016;26(13):1753-1760. [doi: 10.1177/1049732315617444] [Medline: 26613970]
45. Kyngäs H, Mikkonen K, Kääriäinen M, editors. The Application of Content Analysis in Nursing Science Research. New York, NY: Springer International Publishing; 2020.
46. Brand AE, Allen L, Altman M, Hlava MMK, Scott J. Beyond authorship: attribution, contribution, collaboration, and credit. Learn Pub 2015;28(2):151-155 [FREE Full text]

Abbreviations

GUIDED: Guidance for reporting intervention development studies in health research
TIDieR: Template for Intervention Description and Replication

Edited by A Mavragani; submitted 04.09.24; peer-reviewed by S Kim, J Beaudin; comments to author 01.10.24; revised version received 18.10.24; accepted 31.10.24; published 15.01.25.

Please cite as:
Walzer S, Barthel C, Pazouki R, Marx H, Ziegler S, Koenig P, Kugler C, Jobst S
Teaching in the Digital Age—Developing a Support Program for Nursing Education Providers: Design-Based Research
JMIR Form Res 2025;9:e66109
URL: https://formative.jmir.org/2025/1/e66109
doi:10.2196/66109
PMID:

©Stefan Walzer, Carolin Barthel, Ronja Pazouki, Helga Marx, Sven Ziegler, Peter Koenig, Christiane Kugler, Stefan Jobst. Originally published in JMIR Formative Research (https://formative.jmir.org), 15.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Sexual and Gender Minority Adolescents' Preferences for HIV Pre-Exposure Prophylaxis Social Marketing Campaigns: Qualitative Preimplementation Study

Kathryn Macapagal[1,2,3], PhD; Juan Pablo Zapata[1,2], PhD; Junye Ma[4], MA, MS; Jacob D Gordon[1,5], MSW, PhD; Christopher Owens[6], MPH, PhD; Silvia Valadez-Tapia[7], MA; Peter Cummings[2,3], MA, MPH; Nathan Walter[8], PhD; Jim Pickett[9]

[1]Institute for Sexual and Gender Minority Health and Wellbeing, Northwestern University, Chicago, IL, United States
[2]Department of Medical Social Sciences, Feinberg School of Medicine, Northwestern University, Chicago, IL, United States
[3]Department of Psychiatry & Behavioral Sciences, Feinberg School of Medicine, Northwestern University, Chicago, IL, United States
[4]Joint Doctoral Program in Clinical Psychology, San Diego State University/University of California San Diego, San Diego, CA, United States
[5]College of Nursing, University of Cincinnati, Cincinnati, United States
[6]Department of Health Behavior, School of Public Health, Texas A&M University, College Station, TX, United States
[7]AIDS Foundation Chicago, Chicago, IL, United States
[8]Department of Communication Studies, School of Communication, Northwestern University, Evanston, IL, United States
[9]Jim Pickett Consulting, Chicago, IL, United States

Corresponding Author:
Kathryn Macapagal, PhD
Institute for Sexual and Gender Minority Health and Wellbeing
Northwestern University
625 N Michigan Ave, Suite 1400
Chicago, IL, 60611
United States
Phone: 1 312 503 3605
Fax: 1 312 503 6700
Email: kathryn.macapagal@northwestern.edu

Abstract

Background: Sexual and gender minority (SGM) adolescents in the United States are disproportionately affected by HIV. Pre-exposure prophylaxis (PrEP) is a highly effective biomedical HIV prevention method, but its awareness and uptake among SGM adolescents are low. There are no adolescent-centered PrEP social marketing campaigns in the United States that have the potential to increase awareness and interest in PrEP.

Objective: To address this gap, this qualitative study aims to examine SGM adolescents' needs and preferences regarding adolescent-centered PrEP social marketing campaigns.

Methods: SGM adolescents from Chicago and its surrounding areas participated in web-based asynchronous focus groups from February to May 2021. Questions elicited their preferences for content, design, and delivery of SGM adolescent–centered PrEP campaigns. We used rapid qualitative data analysis and organized the findings around key components of social marketing, known as the 4 Ps: *product, price, place,* and *promotion.*

Results: Participants (N=56) were aged 14 to 19 years (mean 18.16, SD 1.22 y), and 64% (36/56) of them identified as a racial or ethnic minority. Among the 56 participants, 70% (n=39) were aware of PrEP; however, 95% (n=53) did not know that PrEP could be prescribed to those aged under 18 years. Adolescents expressed a need for PrEP campaign messaging that provides simple, accurate, and easily accessible information (eg, what is PrEP, for whom PrEP is indicated, and where and how to access PrEP). For product and price, SGM adolescents wanted a campaign to address barriers to, costs of, and how to access PrEP and desired to know about other adolescents' PrEP experiences to improve campaign relatability. For place and promotion, participants preferred digital campaigns on social media to reduce the possibility of embarrassment and stigma and increase the accessibility of health content.

XSL•FO
RenderX

Conclusions: These findings lay the groundwork for designing adolescent-centered educational PrEP campaigns that prioritize both user preferences in PrEP marketing design and strategies to overcome common barriers to PrEP awareness.

(JMIR Form Res 2025;9:e60398) doi:10.2196/60398

KEYWORDS

social marketing campaigns; sexual and gender minority; adolescent; HIV; pre-exposure prophylaxis; PrEP; human-centered design; implementation science; dissemination

Introduction

Background

Adolescents and young adults aged 13 to 24 years in the United States are disproportionately affected by HIV, accounting for 19.3% of new diagnoses in 2022 [1]. In 2019, of the 25,552 new HIV infections among sexual minority men in the United States, 6386 (25%) were aged 13 to 24 years [2]. Surveillance data in Cook County, Illinois, an Ending the HIV Epidemic priority area that includes Chicago and its surrounding areas, show that 82.3% of HIV cases among youths aged 13 to 29 years were among men who have sex with men, 59.7% of the cases were among Black sexual and gender minority (SGM) youths, and 20.3% of the cases were among Latino SGM youths [3].

Pre-exposure prophylaxis (PrEP) refers to HIV prevention medication regimens that currently include oral emtricitabine-tenofovir disoproxil fumarate (brand name Truvada) or emtricitabine-tenofovir alafenamide (brand name Descovy) and long-acting injectable cabotegravir (brand name Apretude). PrEP is a highly effective biomedical prevention strategy, preventing up to 99% of HIV infections when taken as prescribed [4,5,6]. Despite the US Food and Drug Administration's approval of oral PrEP for individuals aged under 18 years in 2018 and long-acting injectable PrEP in 2021 [7], PrEP uptake remains subpar in this age group [8,9]. One study that used US pharmacy claims data to estimate adolescent PrEP use [10] found that of the 6444 youths aged 13 to 19 years who were prescribed PrEP in 2021, just 12% were aged under 18 years, and most were adolescent girls. Across several studies using convenience samples, self-reported PrEP uptake among SGM youth aged 14 to 17 years has been consistently low, ranging from 2% to 5.9% [8,11-14].

Awareness and knowledge of PrEP are 2 significant barriers to its use among adolescents. In studies based in the United States, PrEP awareness among SGM adolescent samples has varied from 15% to 55% [12-15], with higher rates observed in studies consisting of participants who had previously taken part in HIV prevention research where they learned about PrEP. Several demographic, social, cultural, and structural factors relate to SGM youth PrEP awareness, including older age, greater educational attainment, and better access to sexual health care [16-20]. Research also suggests that adolescents have limited knowledge about PrEP beyond the fact that it prevents HIV infection and assume that PrEP is only available for cisgender men or those aged ≥18 years [11,12]. When adolescents are aware of PrEP, they report moderate to high levels of willingness to take it; for example, 2 survey studies showed that 64.9% [9] to 78.5% [8] of adolescent sexual minority men aged 13 to 18 years were willing to use PrEP. Furthermore, even when SGM adolescents are aware of PrEP, their willingness to use it may be moderated by a variety of developmental, socioeconomic, and sociocultural barriers to obtaining and adhering to PrEP. These include concerns about parental involvement, affordability, the need for regular clinic visits but the difficulty in accessing them independently, side effect concerns, and a lack of health care provider training in adolescent PrEP provision [8,10,12,17,21-25].

Social marketing campaigns, which integrate marketing, psychology, and public health principles, have immense potential to promote PrEP awareness and uptake among SGM adolescents, creating positive changes in attitudes, behaviors, and awareness for the greater good of society [26,27]. The key components of social marketing are often referred to as the 4 Ps [28,29]. *Product* includes messages about the desired behavior and its attributes, its health benefits (eg, PrEP awareness, uptake, and HIV prevention), and perceived nonhealth benefits of the behavior (eg, fulfilling social needs). *Price* refers to real and perceived financial or practical costs or other barriers to engaging in the desired behavior from the consumer's perspective and strategies to lower these costs. *Place* involves determining where the focal audience engages in the behavior (in this case, where one obtains PrEP) and implementing ways to make the desired behavior, service, or product more accessible. *Promotion* entails the ways persuasive messages are designed and delivered to the desired audience [28]. Other models also include a fifth P, *policy*, which includes policy-level barriers and facilitators to engaging in the desired behavior that should be communicated to the audience [29]. For instance, in some US states, PrEP is explicitly included in laws permitting adolescents to access certain types of health care without parental consent (eg, mental health and substance use treatment; prevention, testing, and treatment of sexually transmitted infections [STI] and HIV; and other sexual and reproductive health care) [30].

Social marketing campaigns take effect on both population and individual levels [31]. In the case of adolescents, a PrEP social marketing campaign could reach them quickly and directly through multiple forms of media (eg, social media, websites, and posters) that circumvent the need to rely on parents, health care providers, and educators who may or may not deliver or even know about this information. One example of such a campaign is PrEP4Love [32,33], a Chicago-based PrEP awareness initiative designed for Black adult gay and bisexual men and transgender and cisgender women. PrEP4Love reached 41 million unique viewers and was associated with a 3-fold increase in calls to a PrEP hotline over the 2 months following the campaign launch, most of whom were interested in learning

XSL•FO
RenderX

more about or being linked to PrEP [32,33]. The success of the campaign in Chicago led to its adoption by a French AIDS service organization to raise PrEP awareness across France [34].

Although adolescent-inclusive sexual health campaigns exist, such as Get Yourself Tested [35] and National Youth HIV/AIDS Awareness Day [36], few specifically include information about or an emphasis on PrEP. Social marketing draws from exchange theory, which suggests that individuals consider perceived costs and benefits when making decisions and focuses on segmenting the population into distinct audiences (ie, market segmentation) that are more likely to respond to a particular campaign or campaign component [37,38]. As such, social marketing campaigns can be designed to be developmentally appropriate for adolescents to highlight their specific preferences and needs. For example, the content and imagery of PrEP4Love [32,33] were clearly sensual and sexual in nature, which may make many adolescents uncomfortable; approaches that elicit different emotions, such as empowerment or affirmation, could be more resonant with younger audiences [39,40]. Moreover, where campaigns are placed is critical in ensuring that the campaign reaches its intended audience. For instance, PrEP4Love was often promoted in person (eg, in health clinics, public transit, and nightlife establishments) with a website, hotline, and social media presence. However, SGM adolescents may not be able to access the same physical or web-based places (eg, bars and adult-focused sexual health clinics) or may not pay attention to campaign materials in public spaces that portray intimacy among adults.

This Study

In this study, we used a "designing for dissemination and sustainability" approach, which uses methods and principles from dissemination and implementation science, communications, marketing, community engagement, and human-centered design. This interdisciplinary approach enables the tangible products of public health research to better meet the long-term needs of the intended audience and settings [41]. Following this approach, we directly engaged SGM adolescents in preparatory implementation research via web-based focus groups to inform whether and how to undertake an SGM adolescent–centered PrEP social marketing campaign in the Chicago area. Focus group questions sought to understand their perspectives on existing PrEP and sexual health campaigns and their preferences for campaign content, design, and implementation. At the end of the focus groups, we invited SGM adolescents to develop initial prototypes of digital marketing materials that they thought may resonate with a teenaged audience [42,43].

Methods

These data were collected in the context of a larger study that focused on lesbian, gay, bisexual, transgender, and queer (LGBTQ+) adolescents and supportive adults' perspectives on the need for an adolescent-centered PrEP campaign in the Chicago area and what each group desired out of such a campaign. Findings from interviews with adult participants are reported elsewhere [24].

Recruitment

SGM adolescents in Cook County, Illinois, were recruited from February to May 2021 via social media advertisements (eg, Instagram [Meta Platforms] and Snapchat [Snap Inc]); the Illinois PrEP Working Group listserv; registries of individuals who had expressed interest in research study participation through Northwestern University's Institute for Sexual and Gender Minority Health and Wellbeing and the Third Coast Center for AIDS Research; and through the research team's partnership with AIDS Foundation Chicago, a local HIV advocacy and service organization. Social media advertisements were designed to reach adolescents interested in SGM culture (eg, LGBTQ+ organizations and important figures), and study staff reached out to individuals in the registries via SMS text message or email to inform them of a new study opportunity. Social media advertisements and registry outreach directed prospective participants to the study's web-based screener. Eligible adolescents included those who (1) were aged 13 to 19 years; (2) lived in Cook County, Illinois; (3) identified as a sexual or gender minority individual; (4) self-reported HIV negative or unknown status; (5) had access to the internet; and (6) could read in English at an eighth-grade level. We used purposive sampling to enroll up to 20% transgender and nonbinary youths and at least 80% cisgender sexual minority youths to ensure representation of youths across the LGBTQ+ spectrum and aimed for a sample composition of at least 60% racial and ethnic minority youths to reflect the demographics of the Chicago area.

Ethical Considerations

All study procedures were approved by the Northwestern University Institutional Review Board (STU00213036). A waiver of parental permission was obtained for minor adolescents, as the study was determined to pose minimal risk. Following completion of the web-based screener, those who were preliminarily eligible were given a web-based consent form and responded to 4 multiple-choice items to assess their comprehension of the study procedures; those who failed these items twice after being given the opportunity to reread the consent form were deemed ineligible. Then, study staff contacted preliminarily eligible adolescents to participate in a Zoom (Zoom Communications, Inc) call, during which staff confirmed that the respondent was indeed an adolescent, met the eligibility criteria, understood the study procedures, and agreed to a code of conduct for focus group behavior (eg, no harassment or bullying, respectful language, and keeping discussion confidential). To protect their privacy, participants were encouraged to select a username for the focus groups that they did not use elsewhere on the Internet and were not permitted to include identifying information in their usernames. Throughout the study, participants were reminded that they could contact the study team at any time via study SMS text message or email with questions or concerns. At the end of the study, information about study payment (US $40 Amazon gift card sent via email) and HIV and PrEP resources were provided. Data presented here are deidentified for privacy.

XSL•FO
RenderX

Procedures

After enrolling, participants were sent a link to a 25-minute web-based baseline survey via email or text. Participants could select "I don't want to answer" for any item in the survey. Study staff then sent adolescents instructions to access an asynchronous focus group website developed using vBulletin (Internet Brands), web-based forum software that the study team had used in previous research [11,12]. Web-based focus groups have long been used in social science research due to their accessibility and flexibility (eg, they can be conducted synchronously or asynchronously), including in the research team's own work [44,45]. They have also been shown to provide data quality comparable to that of in-person focus groups [46]. Participants accessed the focus group website with unique log-ins. There, they could view links to an introductory forum and 7 forums focused on different study topics, with 1 to 5 questions within each topic (refer to Focus Group Guide section). The introductory forum included a video introducing the study team to participants, an overview of ground rules, and an icebreaker. Participants contributed by typing their responses or uploading content at their convenience. The focus group website was available for 7 days, and participants were strongly encouraged to complete all questions within the first 2 to 3 days to ensure sufficient time for the moderators to ask follow-up questions and for participants to interact with each other's written responses. To increase the likelihood of participant responses, probes used the @username function to trigger notifications. In addition, daily text messages or email reminders that included a direct link to questions and probes participants had not yet responded to were used to increase their engagement. We conducted 5 focus groups, with each group including 11 to 17 participants. Each group was facilitated by 2 moderators who were trained in qualitative methods. Adolescents who answered all questions on ≥4 of the 7 topics were considered completers and eligible for payment.

Focus Group Guide

The focus group guide assessed participants' reactions to existing PrEP campaigns and preferences for future adolescent PrEP campaigns. Questions were adapted from an unpublished needs assessment previously conducted by local community leaders working in HIV prevention and care, including one of the coauthors (JP), to inform the development of adult-focused English and Spanish PrEP social marketing campaigns in Chicago. Additional questions were added by the study team to be inclusive of adolescent experiences based on findings from empirical research on adolescent barriers and facilitators to PrEP and adolescents' preferences for receiving information about sexual health [11,13,25]. Textbox 1 provides examples of focus group questions and topics.

First, we presented existing PrEP and HIV campaigns (PrEP4Love [32-34], Get Yourself Tested [35], and the National Youth HIV/AIDS Awareness Day Campaign [36]) to give participants a reference point for public health campaigns focused on HIV and sexual health and those focused on youths versus adults, given that no PrEP campaigns for adolescents existed at the time. Questions assessed participants' reactions to these campaigns, areas of improvement, and comparisons. Next, we assessed adolescents' preferences for message delivery and dissemination, preferences on campaign design and content, developmental and cultural considerations, and perspectives on messaging for parents and injectable PrEP (not described in this study). Finally, a creative activity prompted adolescents to design mock PrEP campaign content; provide a sample image or video of a campaign they liked and an explanation; or develop a mood board reflecting colors, feelings, or content that might inspire a future campaign. At the end of the group, participants were invited to share their feedback, ask questions, or share additional thoughts they may have had that were not covered. Each focus group transcript was saved as a deidentified Microsoft Word document on password-protected servers. Upon downloading focus group transcripts, study staff disabled further user comments and visibility of the group to all users.

XSL•FO
RenderX

Textbox 1. Focus group topics and sample questions.

Topic 1: Your thoughts on the PrEP4Love campaign

- What works well in this campaign? What doesn't work so well in this campaign?
- PrEP4Love was designed for adults. What changes would you make to the campaign to make it more appealing and interesting to LGBTQ+ teens?

Topic 2: Your thoughts on the Get Yourself Tested and National Youth HIV/AIDS Awareness Day campaigns

- Compare and contrast the PrEP4Love, GYT, and NYHAAD campaigns. What aspects of each campaign do you like the best? What do you not like?
- Which of the aspects you mentioned above are the most important to you to include in a PrEP campaign for LGBTQ+ teens?
- What, if anything, might be missing from any of these campaigns that would be important to include in a PrEP campaign for LGBTQ+ teens?

Topic 3: Where should PrEP campaigns go?

- What in-person spaces would be good places to post ads/share info about PrEP for LGBTQ teens?
- What online spaces/platforms would be good places to post ads/share info about PrEP for LGBTQ teens?
- If you were to design a PrEP video for TikTok, what would your video look like?
- If you were to design a PrEP post or story on Instagram, what would it look like?

Topic 4: What should PrEP messages for LGBTQ+ teens in Chicago look like

- What words and/or phrases would capture your and other LGBTQ+ teens' attention?
- What types of images would you want to include in your PrEP campaign?
- If you wanted to show people in your PrEP campaign, what type of people would you show?

Topic 5: LGBTQ+ teens' feelings and reactions to PrEP campaigns

- How would you want a new PrEP campaign to make you and other LGBTQ+ teens feel?
- How (if at all) should teen PrEP campaigns differ for younger versus older LGBTQ+ teens? LGBTQ+ teens who ever had sex versus those who never had sex? LGBTQ+ teens who are out versus those who are not out? Teens who identify as LGBTQ+ versus teens who are questioning or do not identify as LGBTQ+ but have sex with same sex/gender partners?

Topic 6: PrEP messaging for adults and other forms of PrEP

- How do you think PrEP messages for parents and other adults need to be DIFFERENT from those aimed at LGBTQ teens? How should they be SIMILAR?
- In your opinion, what information is most important in an ad about teens and PrEP designed for parents/guardians?
- If you were designing a campaign for injectable PrEP, what is the one piece of information you'd want in it? How would it be different from a PrEP campaign focused on pills?

Topic 7: Your examples of LGBTQ+ teen PrEP campaigns!

- Option 1: Find an example of health-related (or PrEP/HIV/sexual health-related) content that you LIKE or LOVE! Then, explain what you like about it, and how it relates to your ideas for what a PrEP campaign for teens should be like.
- Option 2: The DIY route. Here are some ideas to get you started, but if you have other ideas feel free to run with it! Use apps like Instagram or Snapchat to edit/draw on/type on the campaigns we shared with you, create a mood board/collage to share the vibe and look you'd want your campaign to have, draw something on your phone.

Measures

Sociodemographic Characteristics

Participants self-reported their age, race, ethnicity, sex assigned at birth, sexual orientation identity, gender identity, outness to parents or guardians, parents' or guardians' acceptance of SGM identity, and sources of sexual health information. Additional investigator-created items assessed participants' HIV and STI testing history, previous sexual experience, and engagement in HIV transmission risk behavior, and other items assessed perceived HIV risk [47].

General PrEP Questions

Investigator-created items assessed perceived norms of PrEP uptake ("I'd like you to imagine 100 adolescents like you. How many of these 100 adolescents do you think are likely to use PrEP"), PrEP awareness before the study (*yes* or *no*), and lifetime uptake (*yes*; *no*). Participants also self-reported perceived difficulty (1=*not at all difficult* to 5=*very difficult*)

XSL•FO
RenderX

and confidence in their ability to access PrEP (1=*not at all confident* to 5=*very confident*).

PrEP Attitudes and Stigma

A scale developed by Walsh [48] in 2019 was used to assess participants' attitudes and perceived stigma toward PrEP. A total of 5 items evaluated perceived effectiveness, safety, and adherence to PrEP, while another 5 items addressed stigma concerns associated with PrEP (eg, "People who take PrEP are promiscuous"). Responses were on a 5-point Likert scale (1=*strongly disagree* to 5=*strongly agree*). PrEP attitude scores ranged from 5 to 25, with higher scores indicating a positive view of PrEP. Stigma scores also ranged from 5 to 25, with higher scores indicating greater perceived stigma.

Data Analysis

Descriptive analyses for all quantitative variables were conducted in R software (x64, version 3.6.1; R Foundation for Statistical Computing). Focus group data were analyzed deductively using rapid qualitative analysis [49], a method commonly used in public health and implementation research to quickly identify actionable themes in qualitative data. The results of such analyses have been shown to correlate highly with traditional qualitative analysis [50,51]. First, study staff exported participants' responses from the Microsoft Word group-level transcripts to tables in Microsoft Excel, with different tabs created for each focus group topic, and responses were cross-validated for accuracy. Two coders reduced data by emphasizing informative sections of participants' responses and removing redundant or filler words. This process resulted in concise data that spoke to the study objectives. Following data reduction, the coders independently examined all excerpts, assigned codes to all excerpts, refined the code definitions iteratively, and held discussions to reconcile differences. Data were then organized conceptually and reported according to the 4 Ps model. Quotes presented in the Results section include participants' age, gender, race, ethnicity, and sexual orientation.

Results

Screening, Enrollment, and Analytic Sample Characteristics

Overall, 348 individuals accessed the eligibility screener, of whom 126 (36.2%) completed the screener. Of the 126 participants who completed the screener, 96 (76.2%) were preliminarily eligible. Of the preliminarily eligible individuals, 68% (65/96) were enrolled after the Zoom call. Then, 94% (61/65) completed the baseline survey and 92% (56/61) answered ≥1 question in the focus group, and 89% (50/56) completed the focus group (ie, responded to all questions in the majority; 4/7, 57% of the topics); 94% (47/50) participants responded to all questions in all 7 topics. The characteristics of the analytic sample are in Table 1.

XSL•FO
RenderX

Table 1. Participant demographics (N=56).

Variables	Values
Age (y), mean (SD)	18.16 (1.22)
Age (y), n (%)	
14	1 (2)
15	2 (4)
16	5 (9)
17	16 (29)
18	16 (29)
19	16 (29)
Sex assigned at birth, n (%)	
Male	28 (50)
Female	28 (50)
Race and ethnicity, n (%)	
Asian	6 (11)
Black	9 (16)
Hispanic	12 (21)
Multiracial	9 (16)
Non-Hispanic White	19 (34)
No answer	1 (2)
Gender identity, n (%)	
Agender	1 (2)
Cisgender man	20 (36)
Cisgender woman	12 (21)
Transgender man	5 (9)
Transgender woman	3 (5)
Genderqueer	4 (7)
Gender nonconforming	3 (5)
Questioning	6 (11)
No answer	2 (4)
Sexual orientation identity, n (%)	
Asexual	1 (2)
Bisexual or pansexual	23 (41)
Gay	20 (36)
Lesbian	3 (5)
Queer	6 (11)
Questioning	3 (5)
Prior sexual experience (allowed multiple options), n (%)	
People with penises	33 (59)
People with vaginas	17 (30)
Never had penetrative sex	19 (34)

Participants could indicate multiple sources from where they heard of the study. Most participants heard of the study through paid social media advertisements (23/56, 41% via Instagram; 10/56, 18% via Facebook; and 6/56, 11% via Snapchat). Other sources included a registry email or text message (7/56, 13%), an LGBTQ+ organization (5/56, 9%), a friend (5/56, 9%), or a

XSL•FO
RenderX

clinic or hospital (2/56, 4%), and 2 (4%) participants indicated that they heard about the study elsewhere (a family member and an email newsletter for a different study for LGBTQ youth). Participants said that their main sources of sexual health information were social media (30/56, 54%), LGBTQ+ websites (15/56, 27%), and porn websites (14/56, 25%). More than half (37/56, 66%) of the participants had prior sexual experience, 48% (27/56) reported condomless penetrative sex, and 13% (7/56) had sex with someone who tested positive for HIV or who was of unknown status. Less than half self-reported that they had previously tested for HIV (24/56, 43%) or STIs (18/56, 32%).

One-third (19/56, 34%) of the participants worried about testing positive for HIV, and 24% (13/56) perceived themselves at risk for HIV infection. ANOVAs revealed significant group differences in perceived risk for HIV infection by gender identity ($F_{7,55}$=2.47; P=.03); Bonferroni-corrected pairwise comparisons showed that those who identified as men perceived greater risk than those who identified as women (mean 2.56, SD 1.26 vs mean 1.4, SD 0.63, respectively). There were also significant group differences in sexual orientation ($F_{5,55}$=2.79; P=.03) whereby participants identifying as gay perceived greater risk than those identifying as bisexual or pansexual (mean 2.80, SD 1.20 vs mean 1.78, SD 0.90, respectively). There were no significant differences in perceived HIV risk by age, race, ethnicity, sex assigned at birth, or having ever been tested for HIV or STIs.

Most (39/56, 70%) participants were aware of PrEP. Only 4% (2/56) had taken it, and 95% (53/56) were unaware that it could be prescribed to those aged <18 years. In addition, 38% (21/56) of the participants found accessing PrEP easy, while 50% (28/56) were confident in their ability to do so. Adolescents viewed PrEP positively regarding effectiveness, safety, and adherence (mean 19.93, SD 2.91; range 5-25) and had moderate stigma levels (mean 11.91, SD 4.39; range 5-25). More details are provided in Table 2.

Table 2. Pre-exposure prophylaxis (PrEP) awareness, use, and attitudes (N=56).

Variables	Values
Perceived norms (“Out of 100 adolescents like you how many do you think are likely to use PrEP?”), mean (SD)	37.88 (24.11)
PrEP attitudes, mean (SD; range)[a]	19.93 (2.91; 5-25)
PrEP stigma, mean (SD; range)[b]	11.91 (4.39; 5-25)
Participants’ awareness of PrEP before participating in the study, n (%)	
Yes	39 (70)
No	16 (29)
No answer	1 (2)
Do you know any teens aged under age 18 who are taking PrEP? n (%)	
Yes	3 (5)
No	53 (95)
Lifetime PrEP use, n (%)	
Yes	2 (4)
No	54 (96)
Perceived difficulty in accessing PrEP, n (%)	
Very easy	7 (12)
Somewhat easy	14 (25)
Neither difficult nor easy	17 (30)
Somewhat difficult	17 (30)
Very difficult	1 (2)
Confidence in the ability to access PrEP, n (%)	
Not at all confident	4 (7)
Somewhat unconfident	9 (16)
Neither confident nor unconfident	15 (27)
Somewhat confident	22 (39)
Very confident	6 (11)

[a]Higher scores indicate more positive attitudes toward PrEP.

[b]Higher scores indicate greater stigma against PrEP use.

XSL•FO
RenderX

Reactions to Existing PrEP Campaigns

Participants were presented with the PrEP4Love, Get Yourself Tested, and the National Youth HIV/AIDS Awareness Day Campaigns and were asked to react to the design, images, and written messages of each campaign. Adolescents generally favored campaigns that highlighted "diversity" in sexual and gender identities, body type, and couple types (eg, couples depicting transgender and cisgender individuals) and disliked campaigns that lacked representation of these groups, including adolescents:

> *I really like that the images show many different couples, regardless of race, gender, sexuality, etc.* [Asian gay man aged 19 years]

That said, they noted that the intentionally sensual nature of the PrEP4Love campaign was off-putting:

> *It seems too sexual/sensual and likely would make me look away before reading the [tag]line.* [Non-Hispanic White bisexual or pansexual man aged 17 years]

Regarding campaign taglines and copy, SGM adolescents favored "straightforward" messages, particularly those that challenged HIV stigma:

> *The simplicity of the campaign definitely works well. I think that it conveys simple information...while simultaneously being informative enough.* [Hispanic bisexual man aged 17 years]

Respondents expressed dislike for messages that were too broad, lacked clarity, or failed to provide PrEP-related resources or information. Adolescents noted that the National Youth HIV/AIDS Awareness Day campaign messages were "too busy and wordy" (non-Hispanic White bisexual or pansexual woman aged 17 years), and that the Get Yourself Tested campaign had a clear call to action but lacked further resources:

> *For GYT...I do think that there could have been a little bit more information.* [Hispanic questioning woman aged 19 years]

Participants preferred existing campaigns with "bright, eye-catching designs" and disliked campaigns with fonts that were difficult to read or those that did not feature teenaged models.

Preferences for Future Adolescent-Centered PrEP Campaigns

In this section, participants' preferences regarding the design of future adolescent-centered PrEP campaigns were organized based on the 4 Ps framework [29].

Product

Participants' comments suggested that social marketing campaigns should specify certain details about PrEP and HIV prevention for adolescents; mention its benefits among youths who have used PrEP; include messages that normalize the health, experiences, and relationships of SGM adolescents; and provide ample resources about PrEP, including information about how to talk to a health care provider about PrEP.

Participants described wanting campaigns to give information regarding the purpose and effectiveness of PrEP and how to access it:

> *I think it should include more...about HIV, what PrEP is, and the cost of PrEP.* [White queer, genderqueer adolescent assigned male at birth aged 19 years]

Participants also desired more tangible information about PrEP statistics and adolescents' lived experiences obtaining and being on PrEP:

> *Including statistics, quotes, and real experiences from adolescents would make it more relatable and emphasize the importance.* [Hispanic White questioning woman aged 19 years]

Other participants also emphasized the importance of including resources about PrEP access, such as offering a comprehensive list of local adolescent-friendly PrEP providers, information about talking to a provider about PrEP, and resources about insurance coverage and paying for PrEP. A , mentioned as follows:

> *I think you're missing financial information [is it free?], contact information...and you can say "talk to your doctor" all you want but like, when I was [younger], I didn't have a doctor. Who would I go to?* [Black transgender man questioning their sexual orientation aged 19 years]

A participant summarized common sentiments among the group regarding product information and how to frame this information to appeal to the audience:

> *I think that the dosing, how to get it, why it's good to take, and information regarding how it works [in a simplified manner], are all good things to include. If you just have a poster saying, "Take PrEP," nobody will take it, but if you say something closer to "Take PrEP, a once daily medication that keeps you 99% safe from HIV," it will be more effective.* [White bisexual woman aged 19 years]

Respondents noted that PrEP messaging centered on the needs of SGM adolescents, and not exclusively focusing on SGM populations, would likely be beneficial to adolescents regardless of background or identity. This was not only perceived to help normalize the health, experiences, and relationships of SGM adolescents but also showed cisgender heterosexual adolescents that PrEP could be for them as well (eg, campaign materials that say "PrEP is for everyone" and depict adolescents of various sexual orientations and genders). This focus on inclusive yet relatable messaging is reflected as follows:

> *The way that my friends explain PrEP to me is...you take the pill once a day, kinda like birth control [which is reassuring because we all know someone on birth control].* [Black gay man aged 18 years]

Furthermore, youths described how portraying SGM adolescents authentically, "just being themselves, like watching TV or enjoying outdoor activities" (Black gay man aged 17 years), can facilitate this sentiment in PrEP campaigns. Finally, another benefit of adolescent-centered PrEP campaigns is that they can

XSL•FO
RenderX

foster a sense of inclusion and empowerment among adolescents, which can help motivate them to consider taking action to use PrEP:

> *I want to feel valued, supported, and seen, where people genuinely care about my sexual health and well-being.* [Asian gay man aged 18 years]

Price

Most comments identified in this domain referenced the price of PrEP (eg, "A big part would be price of PrEP...and any side effects," as mentioned by a Hispanic gay man aged 16 years), the effort involved in accessing it, and the possibility of stigma and embarrassment with engaging with the campaign or seeking PrEP. Fewer adolescents expressed concerns about side effects or the medication regimen itself.

One transgender adolescent explained the potential concerns regarding PrEP that a campaign may help address by relating PrEP to their experiences seeking gender-affirming care:

> *Like when I was younger, I really wanted to go on hormones but...it felt like it would be impossible to get because I didn't fully understand how to access it and a lot of people were telling me how complicated it was. So, I think teens need reassurance that it is simple/easy to obtain prep and maybe even include a step-by-step guide on how to go on it?* [White bisexual transgender man aged 19 years]

Other participants described how engaging with a PrEP campaign or talking about PrEP to a provider or clinic staff member could be embarrassing or elicit judgment from others, which campaigns should proactively address:

> *I don't want to feel ashamed of talking about the topic. I want to feel respected and that I don't need to fear that others might judge me.* [Black gay woman aged 18 years]

Similarly, other youths reported parental involvement as a potential barrier to accessing PrEP:

> *If we go to our doctors for a routine checkup, we'd be scared to talk about sex in general...because most teens are going to be intruded on by parents...it's surely a roadblock for many teens.* [Black gay man aged 18 years]

Moreover, another participant described ensuring that the campaign had multiple ways to speak with a person about PrEP, such as a text line or web-based chat function, given young people's reluctance to speak on the phone:

> *I would include specific ways to get into contact about getting PrEP besides just a phone call because many teens would be way too anxious to call a stranger about that.* [Non-Hispanic White gay man aged 19 years]

Participants were asked to articulate possible drawbacks of adolescent PrEP campaigns focused on specific populations (eg, SGM individuals, racial and ethnic minority individuals, and younger adolescents). Although campaigns specifically for SGM individuals can ensure that PrEP awareness is increased among this population as well as address certain concerns that these populations may have in relation to PrEP ("Black and Latinx people lack medical access and are wary of the medical system because it has mistreated them lots of times," as mentioned by a Hispanic bisexual woman aged 18), participants described that this approach could downplay HIV susceptibility among cisgender heterosexual adolescents and perpetuate stigma within and toward the SGM community. A Black gay woman aged 18 years summarized this concern, stating the following:

> *...it may convey the message that HIV solely affects the LGBTQ community, rather than being a universal virus that impacts all of us.*

Others suggested how campaigns should be attuned to developmental differences in adolescents, with older adolescents being more comfortable with more explicit sexual information or references and younger adolescents possibly experiencing stigma in searching for sexual health information:

> *I feel like some younger people may feel embarrassment or shame in looking into preventative measures.* [Non-Hispanic White bisexual woman aged 19 years]

Place

For the domain of place, participants primarily provided recommendations on how specific individual "champions" or locations that can support youth well-being and health behavior could effectively communicate where and how adolescents could access PrEP, particularly for those who may be less familiar with navigating sexual health care. The following participant quoted below and others described how key influential adults and organizations (eg, schools and accepting churches) could play a role in supporting adolescents' access to PrEP, even if these locations did not offer PrEP themselves:

> *...every school should have guidance counselors or at least a support group for LGBTQ+ teens. I go to a catholic school and recently this year we made a new group called "Unidos" ['Together' in English]. That group helped me so much with resources and it made me feel represented.* [Hispanic gay man aged 15 years]

In addition, physical spaces recommended by participants included schools and public and entertainment venues, such as malls, public libraries, public transportation, and restaurants. These locations were favored due to their accessibility and convenience for SGM adolescents. For example, some youths identified public libraries as places that are accessible to and safe for adolescents, where they could learn more about PrEP with relative privacy:

> *I also think that libraries could work well because they are very trustworthy and when teenagers go to libraries, they are often alone to study or find books.* [White bisexual woman aged 17 years]

Participants also recommended advertising in LGBTQ+ serving organizations, clinics such as Planned Parenthood, pharmacies, and health care provider waiting rooms. These participants

XSL•FO
RenderX

indicated that the placement of marketing materials in waiting rooms could be beneficial:

> *Family doc waiting rooms...if you have to talk to your doctor about PrEP a reminder before going in would be good.* [White bisexual woman aged 18 years]

The placement of marketing materials in pharmacies and hospitals was echoed by other participants as follows:

> *...because it will generate a lot of trust from teens because it is in a healthcare facility. When you go get a check-up it will be easy to remember, and teens might even like to bring it up with their doctor in private.* [Black gay woman aged 18 years]

Promotion

Participants' feedback predominantly focused on the element of promotion, and they articulated strategies that can raise awareness about PrEP among SGM adolescents through carefully crafted content, appealing graphic design, and strategic placement. Finally, participants were also asked to give examples of what these campaign materials might look like.

Campaign Content

This element was distinct from "product" in that it focuses more on message framing, tone, word choice, and market segmentation than the features or benefits of the product itself. Regarding the phrasing of adolescent PrEP messages, participants expressed a preference for concise messages presented in casual language to increase engagement. Furthermore, adolescents highlighted that a campaign should have an affirming and empowering tone that is community-oriented, with several adolescents offering phrases such as "your health matters" or "let's stop HIV together."

Considering the influence of peers and role models on adolescents, participants also suggested including celebrities and social media influencers in future PrEP campaigns to increase campaign appeal:

> *I would feature a well-known celebrity who is attractive and has been a strong supporter of the LGBTQ community.* [Hispanic gay man aged 16 years]

Participants frequently recommended involving adult authority figures, such as scientists, health care providers, and educators, for their credibility and expertise in addressing concerns about PrEP side effects or potential long-term effects, provided their delivery was empowering and supportive rather than patronizing:

> *...doctors/healthcare providers. This would help make PrEP campaigns feel more reliable and trustworthy.* [Black bisexual or pansexual woman aged 17 years]

Moreover, SGM adolescents suggested featuring parents and other adults who may play influential roles in adolescents' health care (eg, "parents who feel more comfortable with their adolescents taking PrEP," as mentioned by an Asian gay man).

Although participants felt that adolescents should get the same access to basic PrEP information regardless of where they lived or who they were, they acknowledged the importance of content segmented by population to address key differences:

> *As someone who went to school in the 'burbs but lives in the city, it tends to be a bit more conservative up there, so I would just be aware of that.* [White queer man aged 19 years]

> *you'd have to advertise in different types of locations, as suburban teens don't take public transportation for example.* [Hispanic bisexual man aged 17 years]

This included creating Spanish language materials to reach the large Spanish-speaking community in Chicago, which could ultimately improve access:

> *...this community, which I'm also a part of, could benefit [by] having ads that are also in Spanish. If possible, just adapting ads to be more approachable and informative for other languages other than English.* [Hispanic bisexual woman aged 18 years]

Campaign Design

Regarding campaign design, adolescents preferred eye-catching graphics, including cartoons, avatars, and infographics:

> *...well thought out color schemes and easy-to-read fonts, along with statistical evidence that explains the effectiveness of PrEP [such as pie charts or bar graphs].* [White gay man aged 18 years]

Adolescents suggested that campaigns should avoid overly sexual images for discretion, with a Hispanic bisexual or pansexual woman aged 18 years stating, "staying away from anything that reinforces stereotypes or is overly sexual (due to parents)." Moreover, participants were clear that the campaign should avoid seeming too slick or corporate, which was viewed as untrustworthy, as described by a participant:

> *...you need to get that attention in a way that is personable, to show that this isn't just something being peddled by a brand, and that it is something genuine and important, from queer folk, to queer folk...Make it real, make it raw, adding polish at the end removes a level of authenticity that is necessary for this level of trust.* [White queer transgender woman aged 19 years]

Although adolescents acknowledged that there are benefits to market segmentation, they were emphatic that the campaign design should also be inclusive and diverse as they believed everyone could benefit from learning about PrEP:

> *I wouldn't necessarily exclude white bodies, but I would put an emphasis on people of color...it hurts literally no one to include them more and really only benefits. also, I think there shouldn't be much distinction between cis and trans [people regarding campaign content], but I do think that we need a bit more trans representation.* [Non-Hispanic White queer man aged 19 years]

Campaign Placement

Participants emphasized that an adolescent-centered PrEP campaign should predominantly be digital to reach this age group regardless of their location. They suggested using smartphone apps and LGBTQ+ websites; social media, such as

XSL•FO
RenderX

Instagram, TikTok, and Tumblr; dating apps (eg, Grindr and Tinder); music streaming apps (eg, Spotify); and mobile games. Among these, social media, particularly TikTok, emerged as the most widely discussed digital platform for campaign dissemination as it was perceived to have "a significant reach and tends to be LGBTQ+ friendly" (Hispanic questioning woman aged 19 years). Participants suggested avoiding Facebook ("I don't know many adolescents who actually use it. It's more popular with adults and the older generation," as mentioned by a Hispanic bisexual or pansexual woman aged 18 years) and digital spaces that feature skippable advertisements (eg, "ads on video streaming services because many of us skip them and don't even pay attention to them," as mentioned by a Hispanic questioning man aged 16 years) or have a reputation for tolerating more toxic Internet discourse (eg, Reddit and Twitter (subsequently rebranded X).

Regardless of medium, participants suggested that physical and digital campaign assets all lead to centralized resources:

> *Having a QR Code that leads to a website with more information on physical advertisements...the idea of a hotline is fantastic.* [White gay man aged 18 years]

They also suggested that such resources should be comprehensive enough so that adolescents do not have to hunt for information elsewhere:

> *The website should include all possible information about PrEP so that it's a one stop shop and so there is no way to walk into it blind...it will inform and persuade because they can just look at it and know, rather than having...to look all of it up on their own.* [White bisexual woman aged 19 years]

Example Campaigns From Focus Group Participants

At the end of the focus group, participants were given the following prompt: "We just asked you a bunch of questions about what a PrEP campaign for LGBTQ+ teens might look like. But some folks might prefer to share their thoughts visually."

Participants were given the option to share graphics or existing sexual health content that evoked what they wanted from a PrEP campaign or to design an example from scratch. In this section, we provide several representative examples and their rationales.

Participants who submitted an original entry most frequently gave examples formatted for social media, such as Instagram or TikTok. This participant described clarity, directness, and inclusion in their design:

> *The idea behind this one [Figure 1] was an Instagram ad that has simplicity, there's a hook at the top, only 2 things to read, cute stickers, an option to swipe up for more info. And the two things I found most important were the hashtag that I was able to make rainbow! And the symbols for the other social media. I was able to get my point across that prep is for everyone, and a neat little reference is that the background somewhat resembles the bisexual flag, hinting that it doesn't matter if you have sex with men or women [or both or anywhere in between] and that prep is still for you.* [Hispanic gay man aged 17 years]

Others submitted examples that included information in both English and Spanish, referencing the conversation earlier in the focus group about the inclusion of Spanish speakers (Figure 2):

Figure 1. Social media pre-exposure prophylaxis (PrEP) campaign asset reflecting inclusive messaging and calls to action.

XSL•FO
RenderX

Figure 2. Social media pre-exposure prophylaxis (PrEP) campaign asset reflecting inclusive messaging and English- and Spanish-language information.

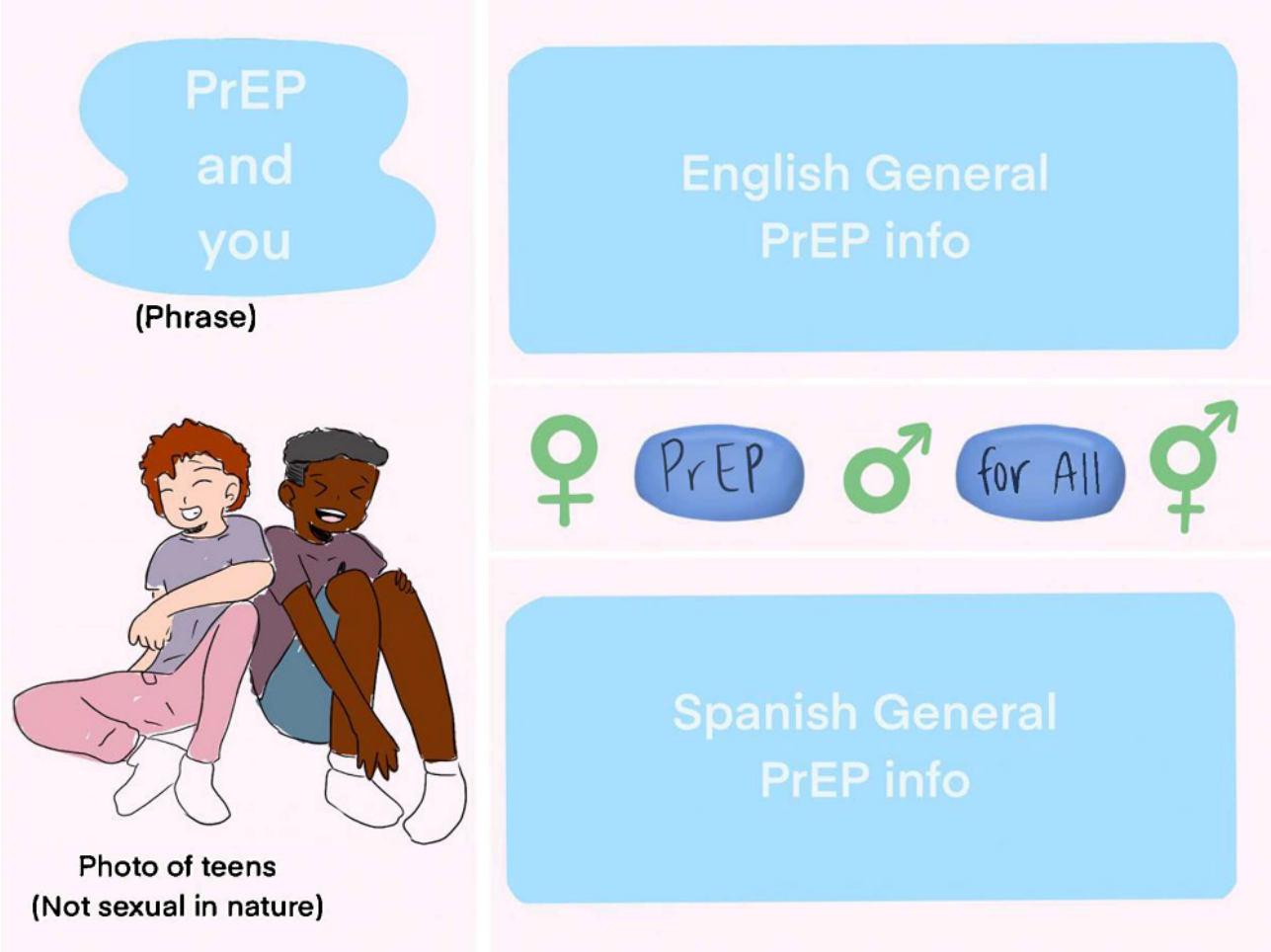

> *I figured that incorporating teens in the campaign in a nonsexual nature would...draw attention—these teens are also different, appearance-wise. For the information, I saw on a different forum that there should probably be information in both Spanish and English. Separating the information are gender symbols and PrEP pills that say, "PrEP for all" on them.* [Black gay woman aged 18 years]

One participant created a script for a commercial reminiscent of the US Trojan Man condom commercials [52] of the mid to late 1990s (Textbox 2):

> *I wrote a quick script for a hypothetical commercial. It's aimed at a [younger?] teen audience and ideal for platforms like TikTok, Snapchat, and Instagram. I think it could be either live-action or animated. This one would maybe be best aimed at teens who have not yet started having sex considering that's the situation in the commercial. The characters can be any gender/body type/sexual orientation. I also wrote it to be slightly raunchy since that will certainly appeal to/get teens attention.* [White bisexual man aged 19 years]

XSL•FO
RenderX

Textbox 2. Excerpt of proposed commercial or brief video for social marketing campaign.

INT. A TEEN BOY'S ROOM LATE AT NIGHT

Two teens make out in bed until TEEN 1 pulls away.

TEEN 1

You sure you're ready?

TEEN 2

(smiling warmly)

Mhm.

TEEN 2 leans into continue kissing TEEN 1 when CAPTAIN PrEP, a caped, over-the-top superhero with an anti-HIV logo or a pill logo on their shield, flies through the wall. The terrified teens scream.

CAPTAIN PrEP

Fear not, young horndogs. It is only Captain PrEP here to make sure you don't get HIV.

TEEN 1

What the f*ck, you destroyed my room!

Captain PrEP is unfazed.

CAPTAIN PrEP

(earnestly)

We all make sacrifices. But your sexual health shouldn't be one of them.

Discussion

Principal Findings

PrEP has the potential to dramatically reduce the number of HIV infections among adolescents in the United States. However, its use among young people remains low, with individual studies reporting that <0.5% of adolescents who may benefit from PrEP have used it [53,54]. Social marketing campaigns have the potential to improve PrEP awareness and uptake [24,26,32]; however, limited research has explored what adolescents want out of adolescent-centered PrEP campaigns. To our knowledge, this is among the first studies on SGM adolescents' perspectives on existing and future PrEP marketing and implementation. Our findings, framed around the 4 Ps of social marketing, suggest that previous sexual health campaigns tailored for SGM adults, or adolescents in general, do not meet the needs of SGM adolescents and that youths themselves can be involved in conceptualizing the content and implementation of PrEP campaigns [43]. In this section, we discuss concrete takeaways from the focus groups that public health practitioners and implementers may consider incorporating into adolescent PrEP campaigns.

Overall, what adolescents wanted from PrEP campaigns aligned with our findings from interviews with adults who worked with or parented LGBTQ+ adolescents about what should go into an adolescent-centered PrEP campaign [24]. For example, adults wanted adolescent-centered PrEP campaigns to offer basic information about PrEP (eg, safety, how it works, access, and costs) that was inclusive, destigmatizing, and aligned PrEP with adolescents' priorities (eg, romantic relationships). In addition, adults wanted concrete, actionable information for adolescents to act on (eg, call this hotline and go to this clinic), delivered through social media and physical spaces where adolescents spend most of their time. For instance, regarding *product* and *price*, SGM adolescents expressed a desire for basic PrEP information in these campaigns that address potential barriers and costs adolescents may encounter. Several years after PrEP's Food and Drug Administration approval for adolescents, access to consumer-facing, adolescent-specific information on PrEP in the United States remains sparse. Although there are some examples of consumer-facing information that indicates that adolescents can take PrEP [55], there are few details addressing the developmental issues that surfaced in this study and previous research on adolescents and PrEP [8,11,12]. An adolescent-focused social marketing campaign could offer answers to the following questions, some of which may need to be tailored to reflect current state laws and policies: Do I need to involve my parents in the decision to take PrEP, because in some states, PrEP is explicitly included in mature minors' right to health care, and in others, it is unclear or implied [56]? What health care providers or clinics can prescribe PrEP to minors, as web-based PrEP locator tools lack information about whether locations serve adolescents? I have never been to a health care provider on my own, so how do I prepare for a PrEP appointment? How do I pay for PrEP without my parents knowing?

Moreover, because PrEP requires consistent health care engagement (eg, for oral PrEP, quarterly appointments are recommended for HIV testing and lab work to check for side effects), campaigns should offer resources that help youth feel more comfortable independently accessing care given that they may have limited experience doing so. Options to cover or defray PrEP medication and related follow-up care should be made explicit (eg, coupons from pharmaceutical companies and requirements for insurance companies to cover PrEP prescriptions and services via the United States Affordable Care Act).

XSL•FO
RenderX

In general, participants in this study were more concerned about how to access PrEP than its side effects, suggesting that apart from basic information on PrEP to increase its awareness, campaigns should focus on demystifying the process of obtaining, paying for, and sustaining PrEP for adolescents. Highlighting real adolescents' experiences pursuing PrEP can make a social campaign more relatable, according to our results. In a different qualitative study by our team, 100 adolescents who had successfully accessed PrEP described different strategies, barriers, and facilitators they encountered as well as their experiences of using PrEP [57]. Sharing anecdotes from these youths and others could be incorporated into a future PrEP campaign to ground the messages in reality [57]. Regarding *place* and *promotion*, participants described wanting a primarily digital campaign implemented via social media. SGM adolescents are heavy users of social media [58-60], where they often seek and consume health content, and consequently, a digital-first campaign could reduce adolescents' embarrassment and stigma (which was relatively high in this sample) by reaching them where they already are. Moreover, additional suggestions indicated that participants wanted a campaign that acted as a "one-stop shop," short of visiting a physical clinic. In other words, adolescents' responses suggested that they wanted a campaign that not only raised awareness and offered education, but also included resources that could facilitate their ability to ask questions relatively anonymously and on demand (eg, a text line) as well as connections to PrEP care (eg, telehealth PrEP consultations). Nonetheless, adolescents also identified potential in-person spaces that could promote a PrEP campaign or connect adolescents to PrEP services, such as gender-sexuality alliances, offices of family medicine physicians or pediatricians, or public libraries. However, given rapidly increasing legislation proposing to restrict sexual, reproductive, gender-affirming, and LGBTQ+-inclusive care across the United States [61,62], one must consider the local social and cultural climate and policies to determine whether it is feasible, acceptable, and safe to implement campaign materials promoting adolescent sexual health services.

One important consideration for promotion is that at the time this study was conducted, self-service social media advertising platforms (eg, Meta, which includes Facebook and Instagram) permitted users to target advertisements to minors based on their demographics, such as gender, presumed sexual orientation identity, and race or ethnicity. However, since early 2023, in the United States, significant restrictions have been placed on social media advertising to minors that would preclude the ability to do market segmentation, in other words, focus educational campaigns on young populations who need it most. Thus, public health professionals wanting to reach minors via digital spaces may need to share their campaigns more broadly with adolescents and hope that LGBTQ+-inclusive messaging and imaging will reach their intended audiences. Alternatively, they will need to identify different means of sharing health campaign materials that do not solely rely on social media advertising, which is also increasingly more expensive. These means could include reaching youth-serving organizations with a digital presence instead of attempting to reach individual adolescents themselves or returning to a focus on in-person events where adolescents are present.

Regarding how content is displayed, participants advised against including sexualized images that are common in adult-focused sexual health campaigns. Instead, participants suggested bright, youthful imagery with concise and affirming messaging delivered by trusted messengers (eg, supportive parental figures, adolescents, and near peers). Results also indicated the need to strike a balance between purposefully inclusive, "PrEP is for all" campaign messaging and information intended to reach populations in demographic groups at higher HIV transmission risk or those with unique medical or health care access considerations, such as transgender, nonbinary, and rural or suburban youth. This preference stands in contrast to PrEP awareness campaigns that tend to adopt a targeted approach to specific populations at greatest risk of HIV transmission. Consistent with literature showing that adolescents value diversity in media [63], participants showed a strong desire for a PrEP campaign that emphasizes diversity across multiple dimensions. However, without some market segmentation that considers local context and needs regarding PrEP, a campaign can run the risk of seeming too generic and impersonal and, consequently, may not reach those who need the resources the most. Moreover, participants repeatedly expressed a distaste for marketing materials that appeared too "corporate" and polished, which suggested that community members who were young, LGBTQ+, and Black, Indigenous, and people of color were not as involved in campaign design. This speaks to the need for adolescent HIV prevention campaign designers to involve community members who are the intended recipients throughout the process, which can also increase trust and buy-in when the campaign is implemented.

Finally, to better understand what participants envisioned adolescent-centered PrEP campaign materials might look like, we included a participatory design exercise at the end of the focus group. Such approaches have been used previously in adolescent sexual health intervention design and have been suggested to improve both youth well-being and youth-centered programming [64,65]. This exercise enabled us to begin to understand how participants' text responses in the focus group data manifested in their design choices and priorities. Ultimately, these youth-generated materials can be used to inspire the direction of PrEP campaign components, ensuring that design choices are not guided solely by researchers' interpretations of focus group data but by adolescents' own visions as well. In fact, these findings, in part, formed the basis of our own adolescent-centered PrEP social marketing campaign, PrEP4Teens [66], launched in late 2023, the development of which will be described in a future publication.

Limitations

While this study has valuable insights, it is important to acknowledge its limitations. The participants were predominantly adolescents who had heard of PrEP before, with a larger proportion of transgender and nonbinary teenagers as well as cisgender sexual minority girls, compared to what is typically seen in HIV prevention research focusing on those at the highest behavioral risk for HIV acquisition (eg, sexually active sexual minority male individuals and transfeminine individuals who engage in unprotected anal sex). It is possible that adolescents who have not heard of PrEP before might want

XSL•FO
RenderX

different information or have different concerns than those who are already somewhat familiar with PrEP. Moreover, those in groups at the highest risk for acquiring HIV may have different perspectives on the campaigns. However, arguably, as nationwide clinical guidelines suggest that all sexually active adolescents and adults should be informed about PrEP regardless of sexuality or gender [67], this sample composition can also be seen as a strength. As this study took place in the Chicago area, which is a large urban area with a historically progressive social and political climate, findings may not generalize to other environments. Furthermore, as this study was conducted in February to May 2021, when most in-person social behavioral research studies had not yet resumed during the COVID-19 pandemic, it is possible that our use of asynchronous web-based focus groups may not have been able to capture certain nuances that live, in-person focus groups would have (eg, body language and tone of voice).

Conclusions

Developmentally tailored, SGM adolescent–inclusive information about PrEP and where to access PrEP services is critical to contribute to US efforts to end the domestic HIV epidemic. Social marketing campaigns that directly reach adolescents have great potential to address these needs quickly and equip them with the tools to make informed decisions about their sexual health. Public health professionals can consider developing their own adolescent-centered PrEP campaigns guided by our findings and make their campaigns responsive to local needs in partnership with local adolescents and adult stakeholders who support or parent adolescents. Public health dissemination and implementation researchers should consider ways to measure and evaluate the reach and impact of PrEP campaigns on adolescents' PrEP awareness, interest, and uptake—perhaps, in part, by adding questions about social marketing campaign exposure and PrEP uptake to large public health surveys of adolescents (eg, Youth Risk Behavior Surveillance Survey by Centers for Disease Control and Prevention [68])—and explore how such patterns may differ regionally and by social-political climate. Having said that, as parents, health care providers, and other supportive adults are also critical gatekeepers of adolescents' access to health care and health information, additional research should examine and design educational PrEP campaigns for these groups.

Acknowledgments

This project was made possible with support from an administrative supplement to the Third Coast Center for AIDS Research, a center funded by the National Institutes of Health (P30AI117943).

The authors would like to thank the participants in this study for sharing their time and for the valuable insights they provided. The authors would also like to thank Dr Sybil Hosek, Hadeis Safi, Jennifer Leininger, and Elijah McKinnon for their expert guidance.

Data Availability

The datasets generated during and analyzed during this study are not publicly available as they contain identifiable information about a socially marginalized group of minor adolescents, but deidentified data are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

References

1. Understanding the current HIV epidemic. AIDSVu. URL: https://aidsvu.org/local-data/united-states/ [accessed 2024-01-14]
2. HIV Surveillance Report, 2020; vol. 33. Centers for Disease Control and Prevention. 2022 May. URL: https://stacks.cdc.gov/view/cdc/121127/cdc_121127_DS1.pdf [accessed 2024-12-25]
3. HIV Surveillance Update Reports. Illinois Department of Public Health. URL: https://dph.illinois.gov/topics-services/diseases-and-conditions/hiv-aids/hiv-surveillance/update-reports.html [accessed 2025-1-10]
4. Preventing HIV with PrEP. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/hiv/prevention/prep.html [accessed 2024-12-23]
5. Fonner VA, Dalglish SL, Kennedy CE, Baggaley R, O'Reilly KR, Koechlin FM, et al. Effectiveness and safety of oral HIV preexposure prophylaxis for all populations. AIDS 2016 Jul 31;30(12):1973-1983 [FREE Full text] [doi: 10.1097/QAD.0000000000001145] [Medline: 27149090]
6. McCormack S, Dunn DT, Desai M, Dolling DI, Gafos M, Gilson R, et al. Pre-exposure prophylaxis to prevent the acquisition of HIV-1 infection (PROUD): effectiveness results from the pilot phase of a pragmatic open-label randomised trial. Lancet 2016 Jan 02;387(10013):53-60 [FREE Full text] [doi: 10.1016/S0140-6736(15)00056-2] [Medline: 26364263]
7. U.S. Food and Drug Administration approves expanded indication for Truvada® (emtricitabine and tenofovir disoproxil fumarate) for reducing the risk of acquiring HIV-1 in adolescents. Gilead Sciences, Inc. 2018. URL: https://www.gilead.com/news/news-details/2018/us-food-and-drug-administration-approves-expanded-indication-for-truvada-emtricitabine-and-tenofovir-disoproxil-fumarate-for-reducing-the-risk-of-acquiring-hiv-1-in-adolescents [accessed 2024-12-23]

XSL•FO
RenderX

8. Moskowitz DA, Moran KO, Matson M, Alvarado-Avila A, Mustanski B. The PrEP cascade in a national cohort of adolescent men who have sex with men. J Acquir Immune Defic Syndr 2021 Apr 15;86(5):536-543. [doi: 10.1097/QAI.0000000000002613] [Medline: 33399311]
9. Dunville R, Harper C, Johns MM, Heim Viox M, Avripas S, Fordyce E, et al. Awareness and willingness to use biomedical prevention strategies for HIV among sexual and gender minority youth: results from a national survey. J Adolesc Health 2021 Jan;68(1):199-206 [FREE Full text] [doi: 10.1016/j.jadohealth.2020.05.032] [Medline: 32693984]
10. Kimball AA, Zhu W, Leonard J, Wei W, Ravichandran I, Tanner MR, et al. HIV preexposure prophylaxis provision among adolescents: 2018 to 2021. Pediatrics 2023 Nov 01;152(5):2018 [FREE Full text] [doi: 10.1542/peds.2023-062599] [Medline: 37899721]
11. Macapagal K, Nery-Hurwit M, Matson M, Crosby S, Greene GJ. Perspectives on and preferences for on-demand and long-acting PrEP among sexual and gender minority adolescents assigned male at birth. Sex Res Social Policy 2021 Mar;18(1):39-53 [FREE Full text] [doi: 10.1007/s13178-020-00441-1] [Medline: 33456624]
12. Matson M, Nery-Hurwit M, Crosby S, Greene GJ, Macapagal K. Awareness and knowledge of pre-exposure prophylaxis among sexual and gender minority adolescents assigned male at birth in the United States. Arch Sex Behav 2021 May;50(4):1701-1711 [FREE Full text] [doi: 10.1007/s10508-020-01904-7] [Medline: 33982212]
13. Moskowitz DA, Macapagal K, Mongrella M, Pérez-Cardona L, Newcomb ME, Mustanski B. What if my dad finds out!?: assessing adolescent men who have sex with men's perceptions about parents as barriers to PrEP uptake. AIDS Behav 2020 Sep;24(9):2703-2719 [FREE Full text] [doi: 10.1007/s10461-020-02827-z] [Medline: 32157491]
14. Owens C, Moran K, Mongrella M, Moskowitz DA, Mustanski B, Macapagal K. "It's very inconvenient for me": a mixed-method study assessing barriers and facilitators of adolescent sexual minority males attending PrEP follow-up appointments. AIDS Behav 2022 Jan;26(1):21-34 [FREE Full text] [doi: 10.1007/s10461-021-03313-w] [Medline: 34081237]
15. Bauermeister JA, Meanley S, Pingel E, Soler JH, Harper GW. PrEP awareness and perceived barriers among single young men who have sex with men. Curr HIV Res 2013 Oct;11(7):520-527 [FREE Full text] [doi: 10.2174/1570162x12666140129100411] [Medline: 24476355]
16. Holloway IW, Dougherty R, Gildner J, Beougher SC, Pulsipher C, Montoya JA, et al. Brief report: PrEP uptake, adherence, and discontinuation among California YMSM using geosocial networking applications. J Acquir Immune Defic Syndr 2017 Jan 01;74(1):15-20 [FREE Full text] [doi: 10.1097/QAI.0000000000001164] [Medline: 27552158]
17. Khanna AS, Michaels S, Skaathun B, Morgan E, Green K, Young L, et al. Pre-exposure prophylaxis (PrEP) awareness and use in a population-based sample of young Black men who have sex with men. JAMA Intern Med 2016 Jan 01;176(1):136-138. [doi: 10.1001/jamainternmed.2015.6536] [Medline: 26571368]
18. Morgan E, Ryan DT, Newcomb ME, Mustanski B. High rate of discontinuation may diminish PrEP coverage among young men who have sex with men. AIDS Behav 2018 Nov;22(11):3645-3648 [FREE Full text] [doi: 10.1007/s10461-018-2125-2] [Medline: 29728950]
19. Strauss BB, Greene GJ, Phillips G, Bhatia R, Madkins K, Parsons JT, et al. Exploring patterns of awareness and use of HIV pre-exposure prophylaxis among young men who have sex with men. AIDS Behav 2017 May;21(5):1288-1298 [FREE Full text] [doi: 10.1007/s10461-016-1480-0] [Medline: 27401537]
20. Thoma BC, Huebner DM. Brief report: HIV pre-exposure prophylaxis engagement among adolescent men who have sex with men: the role of parent-adolescent communication about sex. J Acquir Immune Defic Syndr 2018 Dec 01;79(4):453-457 [FREE Full text] [doi: 10.1097/QAI.0000000000001837] [Medline: 30371531]
21. Macapagal K, Coventry R, Arbeit MR, Fisher CB, Mustanski B. "I won't out myself just to do a survey": sexual and gender minority adolescents' perspectives on the risks and benefits of sex research. Arch Sex Behav 2017 Jul 28;46(5):1393-1409 [FREE Full text] [doi: 10.1007/s10508-016-0784-5] [Medline: 27469352]
22. Ma J, Chase GE, Black A, Klaphake J, Garcia-Myers K, Baker JV, et al. Attitudes toward and beliefs in the effectiveness of biomedical HIV prevention strategies among emerging and young adult sexual minority men. Int J Behav Med 2023 Dec 19:10.1007/s12529-023-10244-4 [FREE Full text] [doi: 10.1007/s12529-023-10244-4] [Medline: 38114707]
23. Moore K, Dell S, Oliva M, Morgan A, Rothman R, Hsieh YH, et al. Does parental insurance impact willingness to take prep in adolescents and young adults? J Adolesc Health 2018 Feb;62(2):S23-S24 [FREE Full text] [doi: 10.1016/j.jadohealth.2017.11.047]
24. Ma J, Owens C, Valadez-Tapia S, Brooks JJ, Pickett J, Walter N, et al. Adult stakeholders' perspectives on the content, design, and dissemination of sexual and gender minority adolescent-centered PrEP campaigns. Sex Res Soc Policy 2023 May 31;21(1):339-351 [FREE Full text] [doi: 10.1007/s13178-023-00826-y]
25. Macapagal K, Kraus A, Korpak AK, Jozsa K, Moskowitz DA. PrEP awareness, uptake, barriers, and correlates among adolescents assigned male at birth who have sex with males in the U.S. Arch Sex Behav 2020 Jan 10;49(1):113-124 [FREE Full text] [doi: 10.1007/s10508-019-1429-2] [Medline: 31602584]
26. Macapagal K, Ma J, Owens C, Valadez-Tapia S, Kraus A, Walter N, et al. 64. #PrEP4Teens: LGBTQ+ adolescent perspectives on content and implementation of a teen-centered PrEP social marketing campaign. J Adolesc Health 2022 Apr;70(4):S34 [FREE Full text] [doi: 10.1016/j.jadohealth.2022.01.024]
27. Grier S, Bryant CA. Social marketing in public health. Annu Rev Public Health 2005 Apr 21;26(1):319-339. [doi: 10.1146/annurev.publhealth.26.021304.144610] [Medline: 15760292]

XSL•FO
RenderX

28. Chau JY, McGill B, Thomas MM, Carroll TE, Bellew W, Bauman A, et al. Is this health campaign really social marketing? A checklist to help you decide. Health Promot J Austr 2018 Apr;29(1):79-83 [FREE Full text] [doi: 10.1002/hpja.13] [Medline: 29700937]
29. McDermott RJ. Social marketing: a tool for health education. Am J Health Behav 2000 Jan 01;24(1):6-10. [doi: 10.5993/AJHB.24.1.2]
30. Minor consent and confidentiality laws for HIV treatment and PrEP: Alabama. National Alliance of State and Territorial AIDS Directors. 2022. URL: https://nastad.org/sites/default/files/2022-02/PDF_Minor_Consent_and_Confidentiality_Laws.pdf [accessed 2024-12-23]
31. Evans WD. How social marketing works in health care. BMJ 2006 May 20;332(7551):1207-1210 [FREE Full text] [doi: 10.1136/bmj.332.7551.1207-a] [Medline: 16710002]
32. Keene LC, Dehlin JM, Pickett J, Berringer KR, Little I, Tsang A, et al. #PrEP4Love: success and stigma following release of the first sex-positive PrEP public health campaign. Cult Health Sex 2021 Mar;23(3):397-413. [doi: 10.1080/13691058.2020.1715482] [Medline: 32212962]
33. Dehlin JM, Stillwagon R, Pickett J, Keene L, Schneider JA. #PrEP4Love: an evaluation of a sex-positive HIV prevention campaign. JMIR Public Health Surveill 2019 Jun 17;5(2):e12822 [FREE Full text] [doi: 10.2196/12822] [Medline: 31210141]
34. #PrEP4Love. AIDSeS. URL: https://www.aides.org/campagne/prep4love [accessed 2024-12-27]
35. Friedman AL, Brookmeyer KA, Kachur RE, Ford J, Hogben M, Habel MA, et al. An assessment of the GYT: Get Yourself Tested campaign: an integrated approach to sexually transmitted disease prevention communication. Sex Transm Dis 2014 Mar;41(3):151-157 [FREE Full text] [doi: 10.1097/OLQ.0000000000000100] [Medline: 24521718]
36. National youth HIV/AIDS awareness day. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/healthyyouth/youth_hiv/resources.htm#:~:text=National%20Youth%20HIV%20%26%20AIDS%20Awareness%20Day%20(NYHAAD)%20is%20observed,health%20and%20safety%20of%20youth [accessed 2022-10-14]
37. Forthofer MS, Bryant CA. Using audience-segmentation techniques to tailor health behavior change strategies. Am J Health Behav 2000;24(1):36-43 [FREE Full text] [doi: 10.5993/AJHB.24.1.6]
38. Icard L, Bourjolly JN, Siddiqui N. Designing social marketing strategies to increase African Americans' access to health promotion programs. Health Soc Work 2003 Aug;28(3):214-223 [FREE Full text] [doi: 10.1093/hsw/28.3.214] [Medline: 12971285]
39. Walter N, Demetriades SZ, Murphy ST. Just a spoonful of sugar helps the messages go down: using stories and vicarious self-affirmation to reduce e-cigarette use. Health Commun 2019 Mar 13;34(3):352-360 [FREE Full text] [doi: 10.1080/10410236.2017.1407275] [Medline: 29236554]
40. Nabi RL, Walter N, Oshidary N, Endacott CG, Love-Nichols J, Lew ZJ, et al. Can emotions capture the elusive gain-loss framing effect? A meta-analysis. Commun Res 2019 Jul 13;47(8):1107-1130 [FREE Full text] [doi: 10.1177/0093650219861256]
41. Kwan BM, Brownson RC, Glasgow RE, Morrato EH, Luke DA. Designing for dissemination and sustainability to promote equitable impacts on health. Annu Rev Public Health 2022 Apr 05;43:331-353 [FREE Full text] [doi: 10.1146/annurev-publhealth-052220-112457] [Medline: 34982585]
42. Biroscak BJ, Schneider T, Martinez Tyson D, Aguado Loi CX, Bryant CA. Applying tools from human-centered design to social marketing planning. Soc Mark Q 2018 Mar 05;24(2):63-73 [FREE Full text] [doi: 10.1177/1524500418761624]
43. Riley WT, Glasgow RE, Etheredge L, Abernethy AP. Rapid, responsive, relevant (R3) research: a call for a rapid learning health research enterprise. Clin Transl Med 2013 May 10;2(1):10 [FREE Full text] [doi: 10.1186/2001-1326-2-10] [Medline: 23663660]
44. Stewart K, Williams M. Researching online populations: the use of online focus groups for social research. Qual Res 2005 Nov 01;5(4):395-416 [FREE Full text] [doi: 10.1177/1468794105056916]
45. Synnot A, Hill S, Summers M, Taylor M. Comparing face-to-face and online qualitative research with people with multiple sclerosis. Qual Health Res 2014 Mar;24(3):431-438 [FREE Full text] [doi: 10.1177/1049732314523840] [Medline: 24572010]
46. Woodyatt CR, Finneran CA, Stephenson R. In-person versus online focus group discussions: a comparative analysis of data quality. Qual Health Res 2016 May;26(6):741-749 [FREE Full text] [doi: 10.1177/1049732316631510] [Medline: 26935719]
47. Napper LE, Fisher DG, Reynolds GL. Development of the perceived risk of HIV scale. AIDS Behav 2012 May;16(4):1075-1083 [FREE Full text] [doi: 10.1007/s10461-011-0003-2] [Medline: 21785873]
48. Walsh JL. Applying the information-motivation-behavioral skills model to understand PrEP intentions and use among men who have sex with men. AIDS Behav 2019 Jul;23(7):1904-1916 [FREE Full text] [doi: 10.1007/s10461-018-2371-3] [Medline: 30554396]
49. Hamilton AB, Finley EP. Qualitative methods in implementation research: an introduction. Psychiatry Res 2019 Oct;280:112516 [FREE Full text] [doi: 10.1016/j.psychres.2019.112516] [Medline: 31437661]
50. Watkins DC. Rapid and rigorous qualitative data analysis: the “RADaR” technique for applied research. Int J Qual Methods 2017 Jun 08;16(1):1-9 [FREE Full text] [doi: 10.1177/1609406917712131]

XSL•FO
RenderX

51. Taylor B, Henshall C, Kenyon S, Litchfield I, Greenfield S. Can rapid approaches to qualitative analysis deliver timely, valid findings to clinical leaders? A mixed methods study comparing rapid and thematic analysis. BMJ Open 2018 Oct 08;8(10):e019993 [FREE Full text] [doi: 10.1136/bmjopen-2017-019993] [Medline: 30297341]
52. Marketing campaign case studies: Trojan man campaign. Marketing Campaign Case Studies. 2008 May 13. URL: https://marketing-case-studies.blogspot.com/2008/05/trojan-man-campaign.html [accessed 2023-07-01]
53. Magnuson D, Hawkins T, Mera R. Adolescent use of Truvada (FTC/TDF) for HIV pre-exposure prophylaxis (PrEP) in the United States: (2012-2017). In: Proceedings of the 22nd International AIDS Conference. 2018 Presented at: AIDS 2018; July 23-27, 2018; Amsterdam, The Netherlands.
54. Mustanski B, Saber R, Macapagal K, Matson M, Laber E, Rodriguez-Díaz C, et al. Effectiveness of the SMART Sex Ed program among 13–18 year old English and Spanish speaking adolescent men who have sex with men. AIDS and Behavior 2023 Feb 27:733-44.
55. Let's talk about it: preventing HIV with PrEP. US Preventive Services Task Force. URL: https://www.uspreventiveservicestaskforce.org/uspstf/sites/default/files/inline-files/hiv-prep-prevention-discussion-guide.pdf [accessed 2024-12-23]
56. Underhill K, Nelson KM. Consent laws for minors regarding sexually transmitted infection and HIV services-reply. JAMA 2022 Dec 20;328(23):2364 [FREE Full text] [doi: 10.1001/jama.2022.19040] [Medline: 36538312]
57. Gordon J, Wongsomboon V, Alvarado Avila A, Lorenzo J, Mustanski B, Macapagal K. Experiences of successful PrEP uptake among adolescent sexual minority men in the United States: a qualitative exploration. AIDS Educ Prev 2024 Feb;36(1):1-15 [FREE Full text] [doi: 10.1521/aeap.2024.36.1.1] [Medline: 38349349]
58. Phillips G2, Magnus M, Kuo I, Rawls A, Peterson J, Jia Y, et al. Use of geosocial networking (GSN) mobile phone applications to find men for sex by men who have sex with men (MSM) in Washington, DC. AIDS Behav 2014 Sep 29;18(9):1630-1637. [doi: 10.1007/s10461-014-0760-9] [Medline: 24682866]
59. Vogels EA, Gelles-Watnick R, Massarat N. Teens, social media and technology 2022. Pew Research Center. 2022 Aug 10. URL: https://www.pewresearch.org/internet/2022/08/10/teens-social-media-and-technology-2022/ [accessed 2023-06-14]
60. McConnell E, Néray B, Hogan B, Korpak A, Clifford A, Birkett M. "Everybody puts their whole life on Facebook": identity management and the online social networks of LGBTQ Youth. Int J Environ Res Public Health 2018 May 26;15(6):1078 [FREE Full text] [doi: 10.3390/ijerph15061078] [Medline: 29861439]
61. Mapping attacks on LGBTQ rights in U.S. state legislatures in 2024. American Civil Liberties Union. URL: https://www.aclu.org/legislative-attacks-on-lgbtq-rights-2024?state=NV [accessed 2024-03-01]
62. State legislation tracker: major developments in sexual and reproductive health. Guttmacher Institute. URL: https://www.guttmacher.org/state-legislation-tracker [accessed 2024-03-01]
63. Ellithorpe ME, Bleakley A. Wanting to see people like me? Racial and gender diversity in popular adolescent television. J Youth Adolesc 2016 Jul 12;45(7):1426-1437 [FREE Full text] [doi: 10.1007/s10964-016-0415-4] [Medline: 26759131]
64. Ozer EJ. Youth-led participatory action research: developmental and equity perspectives. Adv Child Dev Behav 2016;50:189-207 [FREE Full text] [doi: 10.1016/bs.acdb.2015.11.006] [Medline: 26956074]
65. Fakoya I, Cole C, Larkin C, Punton M, Brown E, Ballonoff Suleiman A. Enhancing human-centered design with youth-led participatory action research approaches for adolescent sexual and reproductive health programming. Health Promot Pract 2022 Jan;23(1):25-31 [FREE Full text] [doi: 10.1177/15248399211003544] [Medline: 33858267]
66. PrEP4Teens homepage. PrEP4Teens. URL: https://prep4teens.com/about [accessed 2024-04-04]
67. Preexposure prophylaxis for the prevention of HIV infection in the United States – 2021 update. Centers for Disease Control and Prevention. 2021. URL: https://www.cdc.gov/hiv/pdf/risk/prep/cdc-hiv-prep-guidelines-2021.pdf [accessed 2022-12-01]
68. Youth risk behavior surveillance system (YRBSS). Centers for Disease Control and Prevention. URL: https://www.cdc.gov/yrbs/about/index.html [accessed 2024-06-28]

Abbreviations

LGBTQ+: lesbian, gay, bisexual, transgender, and queer
PrEP: pre-exposure prophylaxis
SGM: sexual and gender minority
STI: sexually transmitted infection

XSL•FO
RenderX

Edited by A Mavragani; submitted 16.05.24; peer-reviewed by PJ Coulaud, Y Cui; comments to author 21.08.24; revised version received 03.11.24; accepted 26.11.24; published 17.01.25.

Please cite as:
Macapagal K, Zapata JP, Ma J, Gordon JD, Owens C, Valadez-Tapia S, Cummings P, Walter N, Pickett J
Sexual and Gender Minority Adolescents' Preferences for HIV Pre-Exposure Prophylaxis Social Marketing Campaigns: Qualitative Preimplementation Study
JMIR Form Res 2025;9:e60398
URL: https://formative.jmir.org/2025/1/e60398
doi:10.2196/60398
PMID:

©Kathryn Macapagal, Juan Pablo Zapata, Junye Ma, Jacob D Gordon, Christopher Owens, Silvia Valadez-Tapia, Peter Cummings, Nathan Walter, Jim Pickett. Originally published in JMIR Formative Research (https://formative.jmir.org), 17.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Mobile Health Intervention to Support Collaborative Decision-Making in Mental Health Care: Development and Usability

Kristin Lie Romm[1,2], MD, PhD; Mari Skoge[1], ClinPsy; Elizabeth Ann Barrett[1], ClinPsy, PhD; Lars-Christian Berentzen[3], MD; Dagfinn Bergsager[4], BSc; Pål Fugelli[4], MEd, PhD; Thomas Bjella[2,5], MSc; Erlend Strand Gardsjord[3], MD, PhD; Kristine Kling[1], MA; Sindre Hembre Kruse[1], BSc; Kari Jorunn Kværner[6,7], MD, PhD; Ingrid Melle[2,8], MD, PhD; Erlend Mork[1], ClinPsy, PhD; Henrik Myhre Ihler[9], MD, PhD; Eline Borger Rognli[10], ClinPsy, PhD; Carmen Simonsen[1], ClinPsy, PhD; Tor Gunnar Værnes[1], ClinPsy, PhD; Sofie Ragnhild Aminoff[1], ClinPsy, PhD

[1]Early Intervention in Psychosis Advisory Unit for South-East Norway, Division of Mental Health and Addiction, Oslo University Hospital, Oslo, Norway

[2]Institute of Clinical Medicine, University of Oslo, Oslo, Norway

[3]Section for Early Intervention in Psychosis, Division of Mental Health and Addiction, Oslo University Hospital, Oslo, Norway

[4]University Center for Information Technology, University of Oslo, Oslo, Norway

[5]Department of Research and Innovation, Division of Mental Health and Addiction, Oslo University Hospital, Oslo, Norway

[6]Centre for Connected Care (C3), Oslo University Hospital, Oslo, Norway

[7]BI Norwegian Business School, Oslo, Norway

[8]Section for Clinical Psychosis Research, Division of Mental Health and Addiction, Oslo University Hospital, Oslo, Norway

[9]Nydalen DPS, Division of Mental Health and Addiction, Oslo University Hospital, Oslo, Norway

[10]Section for Clinical Addiction Research, Division of Mental Health and Addiction, Oslo University Hospital, Oslo, Norway

Corresponding Author:
Kristin Lie Romm, MD, PhD
Early Intervention in Psychosis Advisory Unit for South-East Norway
Division of Mental Health and Addiction
Oslo University Hospital
Sognsvannsveien 21
Oslo, 0372
Norway
Phone: 47 91 50 27 70
Email: k.l.romm@medisin.uio.no

Abstract

Background: Shared decision-making between clinicians and service users is crucial in mental health care. One significant barrier to achieving this goal is the lack of user-centered services. Integrating digital tools into mental health services holds promise for addressing some of these challenges. However, the implementation of digital tools, such as mobile apps, remains limited, and attrition rates for mental health apps are typically high. Design thinking can support the development of tools tailored to the needs of service users and clinicians.

Objective: This study aims to develop and beta test a digital tool designed for individuals with severe mental disorders or substance use disorders to facilitate shared decision-making on treatment goals and strategies within mental health services.

Methods: We used a user-centered design approach to develop iTandem, an app facilitating collaborative treatment between service users and clinicians. Through qualitative interviews and workshops, we engaged 6 service users with severe mental disorders or substance use disorders, 6 clinicians, and 1 relative to identify and design relevant app modules. A beta test of iTandem was conducted to refine the app and plan for a pilot trial in a clinical setting. After 6 weeks of app use, 5 clinicians and 4 service users were interviewed to provide feedback on the concept, implementation, and technical issues. Safety and ethical considerations were thoroughly discussed and addressed.

Results: To avoid overload for the service users, we applied a pragmatic take on module content and size. Thus, iTandem includes the following 8 modules, primarily based on the needs of service users and clinicians: Sleep (sleep diary), Medication

XSL•FO
RenderX

(intake and side effects), Recovery (measures, including well-being and personal recovery, and exercises, including good things and personal strengths), Mood (mood diary and report of daily feelings), Psychosis (level of positive symptoms and their consequences and level of negative symptoms), Activity (goal setting and progress), Substance use (weekly use, potential triggers or strategies used to abstain), and Feedback on therapy (of individual sessions and overall rating of the past week). For the beta testing, service users and clinicians collaborated in choosing 2-3 modules in iTandem to work with during treatment sessions. The testing showed that the app was well received by service users, and that facilitation for implementation is crucial.

Conclusions: iTandem and similar apps have the potential to enhance treatment outcomes by facilitating shared decision-making and tailoring treatment to the needs of service users. However, successful implementation requires thorough testing, iterative development, and evaluations of both utility and treatment effects. There is a critical need to focus on how technology integrates into clinical settings—from development to implementation—and to conduct further research on early health technology assessments to guide these processes.

(JMIR Form Res 2025;9:e57614) doi:10.2196/57614

KEYWORDS

eHealth; shared decision-making; user involvement; user-centered design; mental disorder; mobile technology; illness course; recovery; mobile apps; mHealth

Introduction

> *...in what way has it (iTandem) been helpful?*
>
> *Like right now, I just went into the "substance use" module to write that I managed to say 'no' to buying heroin. Or if I see needles and stuff, I choose to try and get away – move away from where there are needles and that kind of stuff available.* [Beta tester of iTandem]

The ability to communicate effectively with service users is essential in health care. Therapeutic communication can lead to several benefits, including a more accurate description of symptoms, improved detection of emotional states, more appropriate therapeutic measures, better identification of the service user's needs and understanding of their disorder and treatment, and an increased likelihood of shared decision-making between clinicians and service users [1-3]. Engaging individuals with severe mental disorders, substance use disorders, or both in their recovery process can be challenging. Various factors contribute to this difficulty, including a lack of trust in health care services [4], poor therapeutic alliance [5], reduced insight into their illness [6], diminished cognitive functioning [7], and issues with motivation and apathy [8,9].

Priebe and colleagues [10] outlined 5 guiding principles for effective communication in mental health services: focusing on the service users' concerns, demonstrating positive regard and personal respect, appropriately involving service users in decision-making, showing genuineness with a personal touch, and using a psychological treatment model. Emphasizing service users' concerns aligns closely with the concept of "patient-centeredness" in health care. A recent review [11] defined centeredness as encompassing attributes such as being treated as unique, being heard, and sharing responsibility. However, Feldthusen and colleagues [11] emphasize that centeredness is a complex concept, with its essence rooted in the interactions and communication between clinicians and service users. For individuals with severe mental illness, achieving a sense of positive regard and personal respect can be challenging. Therefore, adopting a nonjudgmental approach has been shown to be particularly valuable [12]. Shared decision-making is a widely recognized concept in medicine, but it remains challenging to implement effectively in mental health care. For individuals with severe mental disorders, barriers include clinicians' paternalistic approaches to care [13] and service users' internalized stigma [14]. Increasing service user involvement in activities such as goal setting, addressing values and preferences, and working with motivational factors is essential [15]. Additionally, exploring shared risk-taking can further enhance communication [16]. Genuineness and a personal touch involve openness and sharing but also reflect the therapist's behavior and characteristics, such as warmth and flexibility [17]. Finally, adopting a coherent client-centered therapy approach, such as solution-focused therapy, problem-solving therapy, or cognitive behavioral therapy, is often preferable [10].

Digitalization offers a potential solution to some of these challenges by making treatment more focused and efficient. There has been growing interest in the use of patient-reported outcomes and patient-reported outcome measures (PROMs) in health care [18], which are frequently digitalized. The inclusion of validated assessment tools in digital formats provides clear advantages. However, a significant barrier for health professionals is the time and expertise required to interpret and integrate the extensive findings into their workflow [19]. Additionally, Nguyen and colleagues [19] identified perceived irrelevance and lack of value for the service user as barriers to completing PROMs. As a result, both service users and clinicians emphasize the need for tools that are functional, straightforward, and concise, with as few questions as possible.

There has been a surge in smartphone apps designed to support individuals with mental health challenges, readily available on platforms such as the App Store (Apple Inc.) or Google Play (Alphabet Inc.). Most of these apps are standalone offerings in the free market, focusing on symptoms such as depression, anxiety, or stress. However, they are rarely developed to meet the safety standards typically required in health care settings, making integration into clinical practice difficult. Additionally, while many apps appear acceptable and feasible initially, attrition rates over time are often high [20]. Individuals with mental disorders may face unique challenges that demand

XSL•FO
RenderX

flexible solutions, as therapeutic targets can shift throughout the course of illness. Thus, assessing attrition alone is often insufficient without a deeper understanding of the context. Proposed solutions include placing greater emphasis on external motivating factors and user experience (UX) design, as well as integrating digital mental health solutions into existing care pathways [21,22].

KISS—"keep it simple, stupid"—is a design principle [23] frequently used in design processes and should be applied to service design as well [24]. This principle emphasizes that most systems function optimally when kept simple rather than overly complex [25]. Simpler innovations are more likely to be successfully deployed. Aligning with this concept, human-centered design has gained significant attention in recent years. It involves creative approaches that enable end users to influence the design of technology, tailoring it to better meet their needs [26]. One such approach is participatory design. A fundamental principle in this approach is that people affected by a decision should have the opportunity to influence it, with the goal of designing alternatives to improve quality of life [27]. Despite extensive research on the importance of applying design principles to technology to make it effective, only a limited number of studies in mental health care report using co-design and expert designers in their developmental process [26,28].

A previous service design project focused on the establishment of a low-threshold service for people with psychosis in health care services [29]. Based on insights from service users, caregivers, and clinicians, easy access to a first assessment of psychosis was established. During this project, it became evident that individuals with lived experience were interested in new ways of collaborating with health care services. They sought tools that could enhance the effects of treatment, make better use of the time between sessions, and help them stay focused on their goals. As one service user remarked, "I live my life between the appointments." This highlighted the need to explore how technology could strengthen the focus on treatment between appointments.

The main aim of the iTandem development was to investigate the needs described by service users and clinicians and to integrate them into an app designed to enhance shared decision-making for people with psychotic, bipolar, or substance use disorders in treatment. We applied design thinking principles to incorporate user needs and feedback throughout the process [30,31]. Our long-term goal is to explore the barriers and enablers for integrating technology into shared decision-making in mental health care and to examine how the integration of digital tools affects the mode of service delivery. In this paper, we will present (1) the design process behind the development of iTandem and (2) the results of beta testing. Finally, we will discuss the importance of future research on the digital enhancement of service delivery.

Methods

Project Setting

iTandem was developed at the Early Intervention in Psychosis Advisory Unit for South East Norway (TIPS SE) at Oslo University Hospital (OUH). TIPS SE works to support mental health services in detecting and treating patients with early psychosis in accordance with national guidelines, with a particular emphasis on how these services are organized.

Design Process

Overview

External funding from the South-Eastern Norway Regional Health Authority enabled the hiring of experts in service design and interaction design to lead the process. The process progressed through 3 phases, involving multiple stakeholders including patients, clinicians, developers, data protection officers, and clinicians, in addition to the original project team consisting of researchers and clinicians from OUH (KLR, EAB, KK, EM, EBR, CS, TGV, and SRA). The phases involved are discussed in the following subsections.

Preparation Phase

Qualitative interviews were conducted to inform the design process by understanding the needs and preferences regarding app content. Six patients diagnosed with psychotic or bipolar disorder, as well as 1 relative, were interviewed individually. In addition, 2 focus groups were held with 6 clinicians: 1 nurse practitioner, 1 clinical psychologist, 1 psychiatrist, and 1 medical doctor (all from an early intervention in a psychosis team); as well as 1 psychologist from a general psychiatric outpatient service and 1 from the child and adolescent services for early psychosis.

Insight Phase

A private design company managed the design process and developed a clickable app prototype to inform the developers. The design company, in collaboration with researchers, agreed upon a project plan that included several meetings with clinicians (the same clinicians who had been involved in phase 1) and 4 patients (the same as in phase 1). The designers utilized the insights collected during phase 1 and knowledge generated through workshops to focus on themes that appeared to be both significant and feasible to work with, given the project's budget and time constraints.

Developmental Phase

This phase was carried out in collaboration with the Mobile App team at the Center for Information Technology (USIT) at the University of Oslo (UiO), which developed a working app capable of transferring encrypted data to secure storage within Services for Sensitive Data (Tjenester for Sensitive Data or TSD). This was necessary to meet the safety requirements for the storage of sensitive data, as mandated by the data protection officer at OUH. A risk and vulnerability analysis was conducted at OUH. During this phase, 4 patients and 2 clinicians were invited to test the iTandem app during development to ensure it could be adjusted based on short iterations.

Project Organization

Overview of the Project Group

The project group included researchers and clinicians working with early psychosis treatment. TIPS SE also has a board composed of people with lived experience of bipolar and

XSL•FO
RenderX

psychotic disorders. These individuals were either currently in treatment or had previously been in treatment at mental health services in OUH during the development project. The roles of the different groups are presented below.

Researchers

During phase 1, researchers conducted individual qualitative interviews with 6 service users and 1 relative and led 1 focus group with 6 clinicians. The purpose was to explore the challenges participants believed digital tools could address. During phase 2, the researchers served as a bridge between the clinic and the design team, facilitating workshops, collaborating with user representatives, and ensuring ethical issues were handled appropriately. They also suggested improvements to the app content to ensure it aligned with clinical guidelines and recommendations regarding treatment. During phase 3, the researchers facilitated and oversaw the developmental process of the app at USIT.

Designers

During phase 2, designers planned and conducted workshops with clinicians and researchers. Throughout the project, they interviewed clinicians and user representatives, and user tested different options in the studio with both clinicians and user representatives in an iterative process. The process was based on design thinking, working with divergent and convergent thinking according to the Double Diamond framework [32]. During phase 3, UX designers, engaged in making products or services usable, enjoyable, and accessible, were involved at USIT to further develop the interface and add functionality based on tests from the prototyping phase.

Service Users

Service users with severe mental disorders or substance use disorders were involved throughout all phases of the project. Service users (aged 18-30 years) and 1 relative (parent) provided input on functionality and were invited to user test different prototypes. This involvement had an impact on the final app interface and functionality. Some of these service users were among those interviewed during phase 1. Additionally, we met with the board of user representatives in the Department of Substance Use and Addiction at OUH to discuss app content relevant to them. Service users not directly involved in development were invited to beta test iTandem.

Clinicians

During the process, involved clinicians highlighted important aspects essential for implementation, and shared their ideas on how iTandem could be valuable in their daily clinical work. Additionally, they user tested the app during iterative cycles and provided feedback on how they believed the app would be received by their patients.

Software

The iTandem app was developed using the React Native Open-Source UI Software Framework (Meta Platforms, Inc.) and is available for both iOS and Android platforms. It can be downloaded from the App Store or Google Play. Data are transferred via HTTPS as JSON files through a secure web survey provider (Nettskjema), hosted by USIT at the UiO. The “Activity,” “List of Medication,” and “Diary of Good Things” modules are stored on the user’s phone for easy access, while aggregated data from all modules are encrypted and transferred to a secure server (TSD) at UiO once a module is completed. These data generate a report that can be accessed by clinicians during consultations and provide structured data available for researchers.

Ethical Considerations

The study was approved by the Data Protection Officer at OUH (approval numbers 2016/14408 and 20/18434). The regional ethical committee found that the development and testing fell outside their mandate (reference number 105851). All clinicians and service users who were interviewed signed informed consent forms. Service users who were engaged during the early phases of development, in collaboration with the service designers, also signed informed consent for their involvement in this work. All data were handled securely at TSD and by a secure research server at OUH. Service users were compensated with approximately US $23 per hour. All data used in this paper have been anonymized.

Results

The Development Process

Designing iTandem for Clinical Integration and Shared Decision-Making

We aimed to make the use of iTandem as true to a clinical setting as possible, keeping both the service user’s and clinicians’ experiences in mind when framing the app content. Our primary goal was to develop a tool to enhance shared decision-making. Besides supporting collaboration, we believe this is crucial for compliance. iTandem is not a standalone tool, and both its usage and the reasons for any attrition should be explored during treatment. We used phrases that closely mirror those used by clinicians. Furthermore, we chose to incorporate only a minimum of validated self-report measures, as the rationale behind the app was to serve as a communicative tool rather than a strict research assessment instrument.

The described needs were in line with current research, and based on the current literature, the rationale behind each module is outlined below (Table 1).

XSL•FO
RenderX

Table 1. Description of the modules of iTandem and their intended purposes.

Module	Frequency of notifications	Aim	Contents	Basis of measure/intervention
Sleep	• Twice daily	• Map sleep patterns and quality, link to the level of function	• Sleep diary	• Consensus sleep diary
Medication	• Once or twice daily	• Support service users in keeping track of and remembering to take medication at the right time	• Registration of medications taken and potential side effects • List of medications	• Partly developed by the project group and partly adapted from the UKU Side Effect Scale • Developed by the project group
Recovery	• Daily/weekly • Monthly	• Do positive psychology exercises • Map and focus on personal recovery and positive aspects of mental health	• Good things • Personal strengths • Psychological well-being • Emotional well-being • Cognitive well-being	• Exercises adapted from "Positive psychotherapy for psychosis" • Questionnaire about the Process of Recovery • Warrick-Edinburgh Mental Well-Being Scale • Lehman Life Satisfaction Scale
Mood	• Daily	• Map feelings and mood variability, link to sleep, medication, experiences, and other symptoms	• Mood diary • Emotions	• Adapted from the paper version of the mood diary from the Systematic Treatment Enhancement Program for Bipolar Disorder study • Developed by the project group
Psychosis	• Weekly • Monthly	• Follow symptom development and relation to difficulties in reaching goals, link to other variables, guide treatment	• Symptoms of psychosis • Negative symptoms	• Developed by the project group • Self-evaluation of negative symptoms (Self-Evaluation of Negative Symptoms Scale)
Activity	• Daily	• Increase/regulate activity level, set goals, plan activities, motivate to complete, bring personal experiences into treatment	• Registration of goals • Registration of activity	• Developed by the project group
Substance use	• Weekly	• Gain a better understanding of how substance use impacts symptoms and daily life, guide coping strategies	• Registration of substance use and coping strategies	• Developed by the project group
Feedback on treatment	• Before and after consultations	• Facilitate feedback to the clinician to strengthen patient involvement and quality of treatment, provide an overview of the treatment course	• Before consultation • After consultation	• Developed by the project group

Sleep

Sleep disturbances are common in psychotic and bipolar disorders [33]. In bipolar disorder, sleep disturbances are frequently observed during and between episodes and are associated with more severe symptoms [34]. Therefore, being able to detect and monitor disturbed sleep patterns and implement treatments that focus on improving sleep may be crucial. Both service users and clinicians considered it valuable to have a tool that collects daily data on sleep to avoid recall bias. Additionally, clinicians expressed a need for a tool to assist in diagnosing sleep disorders. The Sleep module was based on the consensus sleep diary, which enables the monitoring of subjective sleep measures and the diagnosing of common sleep disorders [35]. The diary consists of a questionnaire in the morning about how well you slept and a questionnaire in the evening where users rate how their day was. The intended purpose of the Sleep module is to provide an overview of service users' sleep patterns and, based on this, initiate relevant sleep treatment if necessary. Additionally, the module includes a link to a website [36] with information about common sleep problems in severe mental disorders and sleep hygiene advice.

Medication

Both service users and clinicians wanted a tool that could assist with reminders (notifications) for taking medication and monitoring medication side effects. Medication is a key component in the treatment of both psychotic and bipolar disorders. Antipsychotics are effective, especially during the early and acute phase of psychosis [37], and are also commonly used in people with bipolar disorder [38], along with mood stabilizers [39]. However, these medications may have side

XSL•FO
RenderX

effects that are not always properly addressed [40]. Recent studies have highlighted the importance of considering the user's experience level, autonomy processes, values and risk preferences, knowledge, and needs to accurately understand the service users' experience with medication [41]. The module was designed to enable service users to create a (revisable) list of their medications, set reminders for doses, and report side effects. Previous research has demonstrated that the language used to describe side effects can influence users' expectations and may contribute to experiencing nocebo-induced side effects [42]. Webster and colleagues [42] found this to be particularly significant for people with negative beliefs about side effects. To mitigate this, we used UX design to integrate the patient-rated UKU Side Effect Scale [43], which consists of 48 single symptom items, into the Medication module. Instead of requiring the service user to rate 48 individual items, we chose to present the scale as a visualization of the human body. Service users can click on the body part where they experience side effects and select the relevant side effect(s).

Recovery

Clinicians were primarily interested in how treatment impacted clinical recovery in terms of symptom remission and adequate functioning. However, they also emphasized the importance of considering the impact on quality of life. Service users, by contrast, highlighted the need to focus more on their positive outcomes in line with personal recovery. This is commonly defined as the subjective experience of connectedness, hope, identity, meaning in life, and empowerment [44]. Personal recovery has received increased attention in mental health research in recent years, particularly in relation to psychotic disorders [45,46]. This highlights the importance of monitoring treatment outcomes in terms of well-being, personal recovery, and clinical recovery. Positive psychology has been applied to the treatment of mental disorders and, more recently, to psychotic disorders through positive psychotherapy for psychosis [47]. The Recovery module has 2 parts: First, it monitors personal recovery using a validated assessment measure [48], focusing on good feelings and thoughts with a validated measure of mental well-being (Warrick-Edinburgh Mental Well-Being Scale) [49], and life satisfaction with The Life Satisfaction Scale [50]. Second, the module includes 2 positive psychology exercises, namely, (1) Good Things, which is a positive diary of good experiences. The diary allows for daily notes to be stored on the phone, which can be revisited by the service user. It is based on positive psychology, suggesting that gratitude interventions can positively affect well-being, happiness, and life satisfaction [51]. (2) Personal Strength Exercise, which emphasizes the need for strength-based interventions for people with severe mental disorders [52]. The service user is presented with 24 strengths visualized by icons. They are encouraged to select 4 of their personal strengths and plan activities that utilize these strengths. This approach is adapted from the Positive Psychotherapy for Psychosis group intervention developed by Slade et al [47], in accordance with the WELLFOCUS study [53].

Mood

Detecting fluctuations in mood symptoms, such as elevated and depressed moods, is crucial in the treatment and relapse prevention of bipolar disorder, as emphasized by the service users. However, mood symptoms are not limited to bipolar disorder; they also play a significant role in other mental disorders, as highlighted by all the service users and clinicians who were interviewed [54,55]. Monitoring mood and affective symptoms is also relevant for personal recovery [56]. The service users wanted the opportunity to log affective states or emotions as they emerge in daily life, with the option of linking such experiences to external triggers. Service users with bipolar disorder may benefit from using daily mood charts to monitor their symptoms [57]. This feature was incorporated into the Mood module based on a paper version adapted from the Systematic Treatment Enhancement Program for Bipolar Disorder [58]. Additionally, we developed a module to monitor daily emotions, potential triggers, and an evaluation of fluctuating or stable emotions.

Psychosis

The clinicians were interested in measures to detect how psychotic symptoms fluctuate and also the ability to detect triggers. This functionality was less important to the service users. This might seem paradoxical but aligns with the research literature on recovery [59]. However, most service users agreed that it could be relevant to log their symptoms to register the effects of medication, although they argued that the questions needed to be brief and relevant. As we defined iTandem as a tool aimed at facilitating dialogue in treatment sessions and not for research assessment, we chose to include questions about psychosis as typically phrased by clinicians in clinical practice. Thus, the Psychosis module includes a list of positive psychosis symptoms to choose from and brief follow-up questions regarding how symptoms affect daily function, triggers, etc. This may serve as a means to monitor treatment progression regarding positive psychotic symptoms. Additionally, the clinicians wanted to monitor negative symptoms. Negative symptoms are challenging to assess due to their overlapping features with, for example, depression and medication side effects [60]. However, the Self-Evaluation of Negative Symptoms Scale has been shown to be a valid self-report scale, and we chose to incorporate this into the Psychosis module [61].

Activity

Increasing goal-directed activity is important in the recovery process for severe mental disorders. However, achieving this in treatment can be challenging. The service users and clinicians wanted functionality that could be incorporated as part of treatment and motivate for change. This aligns with previous research, which states that goal planning is an important element in mental health service delivery, but we lack standardized approaches to support it [62]. Furthermore, both working alliance and goal achievement support the individual recovery process, and a strong working alliance seems to make people more capable of reaching their goals [63]. Potentially, digital tools could supplement and enhance this alliance. This module aims to support motivation for activity by allowing the service user to set long- and short-term goals. When an activity that

XSL•FO
RenderX

supports the goal has been executed, the service user selects an icon representing this goal. Each completed activity creates a streak.

Substance Use

Substance use and abuse are of major concern in both psychotic [64] and bipolar disorders [65]. In addition, people with substance use disorder have a higher prevalence of mental disorders than the general population [66]. However, even though both service users and clinicians agree that this is a major problem, it became apparent that the service users did not want to focus on detailed intake of substances. They stated that thorough monitoring of their weekly substance use did not motivate change and argued for brief questioning on drug intake. However, they found it useful to be able to convey which strategies they had used to abstain from taking drugs during the week. Clinicians were more prone to ask for detailed accounts of substance use to get an overview but were also concerned that this could lead to demotivation and disengagement. In addition, we were guided by principles from motivational interviewing and aimed to engage the service user in strategies to abstain from using drugs when feeling a craving [67]. Hence, we chose the same strategy as with the Psychosis module and included only brief questions about use and abstaining strategies, along with a link to a recommended user-friendly web page about substance use.

Treatment

The final module of iTandem was included to adhere to the Norwegian guidelines for psychosis treatment, which underline the need for better tools to evaluate the experience of treatment. PROMs and patient-reported experience measures (PREMs) are widely used [68]. PROMs are primarily used for the assessment of treatment outcomes, while PREMs are used to evaluate patients' experiences with the service they receive [69]. They allow for more direct feedback to the provider. In a tool designed for therapeutic collaboration and alliance building, the service users' experience and ability to provide feedback on the sessions were considered important for iTandem. Feedback on treatment is also a fundamental principle in cognitive behavioral therapy, which is one of the most frequently used treatment approaches in mental health care. Therefore, including this module would also support cognitive behavioral therapy. The module consisted of questions about how the last week had been, and questions about the individual treatment sessions, to be answered before and after each session (Figure 1).

XSL•FO
RenderX

Figure 1. Screenshots illustrating the home page in iTandem and the page where you choose which modules you want to use (setting goals). The other pages show the various modules. Left: sleep, recovery, psychosis, and substance use. Right: medication, mood, activity, and feedback on treatment.

Report

Both clinicians and service users wanted a report to be able to log patterns. However, due to data security issues, only a limited amount of data could be stored on the phone. Except for the positive diary, activity, and medication list, data from all the modules were presented in the report. The positive diary was excluded from the report to allow service users to record positive events without feeling compelled to share them with clinicians. The Activity module incorporated elements of gamification, such as weekly streaks, which were believed to be a sufficient motivator. The other data entered were transferred to TSD. A report was available for clinicians, with the purpose of being shared with service users during sessions. Design principles were applied to make the report as easy to read as possible. Patients were not able to log into the report themselves. This decision was influenced by both technical and regulatory challenges, as well as how iTandem was intended to function

XSL•FO
RenderX

as a tool for collaboration and shared decision-making. By integrating the report reading into the sessions, we believe it better supports the main goal (Figure 2).

Figure 2. Illustration of 2 reports based on data logged into iTandem and how they will look on a screen. To the left is a mood diary illustrating data capturing the highest and the lowest mood during a day. To the right is a sleep diary, visualizing sleep during a week. The sleep period and sleep effectiveness are marked in different colors. A higher resolution version of this image is available as Multimedia Appendix 1.

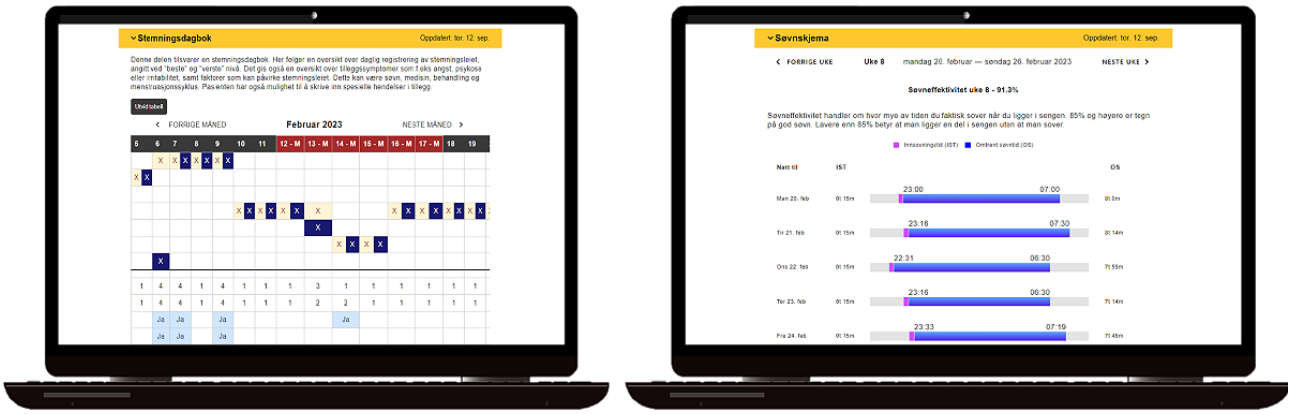

Beta Testing

Overview

Six clinicians (2 psychologists and 4 psychiatrists) were invited to participate in beta testing and attended a 3-hour training session on how to use iTandem. During this session, we demonstrated the different modules and their intended purposes. In addition, we provided an online protocol where participating clinicians and service users could access written information about the content and how to use the app. In collaboration with clinicians, service users were encouraged to choose 2-3 modules relevant to their treatment. It was not emphasized that this was set as an upper limit, but both clinicians and designers were concerned about overwhelming the patients with too many notifications. This could also undermine the overall goal of helping to focus sessions on issues considered particularly relevant to the patient. The patients used the app for 6 weeks (Figure 3).

Figure 3. A workflow with iTandem. After setting initial goals, the service user logs data, which are assessed during sessions. After monitoring sleep for 4 weeks, the service user and the clinician choose to change modules as they want to focus more on activities and medication.

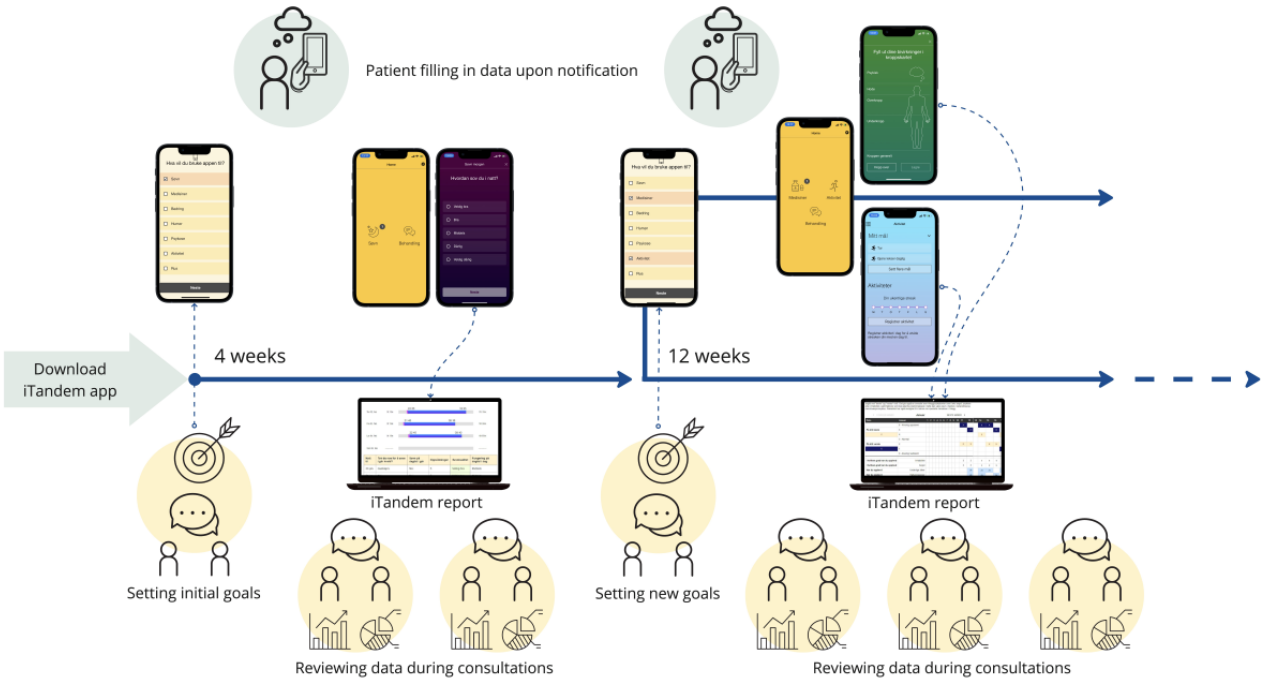

After beta testing, 4 service users and 5 clinicians agreed to be interviewed about their experiences with iTandem. This served as a preliminary proof of concept to determine whether the concept appeared to have potential. We also wanted feedback to adjust both iTandem and the educational implementation program before proceeding with a larger test in a clinical setting. Several issues were addressed, and the responses were sorted based on what appeared to be the most relevant attitudes and sentiments following the test.

XSL•FO
RenderX

Feedback on Concept by Service Users

The service users were mainly positive toward iTandem and the possibilities it offered. They appreciated the ability to select different modules that captured issues beyond the symptoms of their primary diagnosis, working alongside their clinician. This approach made it easier for them to focus on subjects and issues that were important to them and to convey this effectively to their clinician, thus supporting collaboration on goal setting and shared decision-making. They had concerns about how the report was used during sessions and emphasized the importance of clinicians utilizing the logged data; otherwise, the aspect of collaboration would be undermined. Some service users felt that the questions in the app were a bit repetitive (mainly the validated forms), and called for more opportunities to use free text to convey important information to their clinicians. It should be noted that there were service users who signed the informed consent to participate but did not follow through. According to clinicians, this was related to either paranoid ideas, high general symptom load, or being transferred to other departments.

Feedback on the Concept by Clinicians

Clinicians found the data from iTandem in the report to be useful; it provided important insights for their therapeutic work. Several were surprised by how the focus of treatment shifted when they received new information through the app, aligning with the goal of shared decision-making and increased collaboration. However, they found it was not always easy to remember to access and read the report, and they called for the report to be made accessible in the regular hospital records. By contrast, when this was not an option, they found the final solution to be manageable.

Implementation Issues

Many service users were eager to try several modules at once. However, notifications from multiple modules were perceived as overwhelming, highlighting the importance of working with 2-3 modules at the same time. The time spent on each module was deemed appropriate. Clinicians also felt overwhelmed by the challenge of integrating a digital tool into treatment sessions while finding space for it in their clinical routine. They found it stressful not to be fully familiar with the tool, making it difficult for them to effectively explain its use to the service user.

Technical Issues

Several technical issues arose regarding both the data sent from the app to TSD and the presentation in the report. This was partly due to errors in the app, but also because the report functionality was newly developed by the web app group at USIT and had not been properly tested beforehand. This led to several periods of downtime for the report. Even though this was seen as a major problem that persisted for several weeks, participants continued to enter data and found it meaningful. Furthermore, there were issues regarding the user-friendliness of the clock in the Sleep module. It was too easy to enter the wrong date, which obviously resulted in incorrect sleep patterns in the report. We also discovered missing data in the report.

Revision After Beta Testing and Plans for Pilot Testing

Based on the beta testing, we managed to resolve the technical issues. We revised the information material and teaching manual, omitting text and focusing on video films to show and tell how the app is intended to be used. In addition, we only recommend using a limited number of modules at a time to avoid overwhelming notifications. The new information material, manual, and the revised iTandem have since been further tested in a pilot trial conducted during winter 2023. In this trial, we tested iTandem as part of an early intervention service in a rural area in Norway. Data from this trial are currently being analyzed and will be part of a PhD thesis on how digitalization impacts service delivery in mental health care.

Furthermore, several studies are planned, incorporating iTandem as part of studies that use single modules and investigate implementation research connected to digitalization in mental health care.

Discussion

We have developed and beta tested a digital tool, iTandem, based on principles from design thinking. The main aim has been to utilize technology to strengthen and focus shared decision-making between service users and clinicians in mental health care and to enhance goal setting. The demand for simple functionality was underlined by both service users and clinicians.

We believe the design process was helpful in designing a tool that accommodates shared decision-making. We found the tool to follow many of the principles recommended by Priebe and colleagues [10], such as a heightened focus on the service users' concerns, positive regard and personal respect, appropriate involvement of service users in decision-making, genuineness with personal touch, and the use of a psychological treatment model [10]. By allowing service users to choose modules to aid focus in treatment, we aim to empower them during their recovery process.

What is meaningful for the user should be the priority in treatment. This does not mean that clinical evaluation is less important, but to build an alliance and secure a therapeutic relationship, the service users' interests must remain in focus throughout therapy. This aligns with the demand for more patient-centered care. It is also in line with the shift from "patient-centered" to "person-centered" care, as discussed by Boardman and Dave [70]. There is a change in how we view the therapeutic relationship, where the therapeutic alliance is more about co-production as a key element in the clinician-service user collaboration.

Furthermore, we believe that the ability to choose themes to work on and the possibility to give feedback on treatment sessions, including questions about being met with respect, will underpin service users' experience of being respected and treated in a nonjudgmental way. We hope this tool will enhance clinicians' motivation to make changes in accordance with the service users' experienced needs. By using design both as a creative process and applied to the visual elements, we have strived to make the tool engaging and less "clinical." This may increase the feeling of a more personal touch. To accommodate

XSL•FO
RenderX

a more generic therapeutic use, the elements of iTandem are well suited to fit cognitive behavioral therapy, focusing on agenda setting, goal setting, and the execution of activities. Unsurprisingly, a design process led to solutions that align well with this therapy approach, as it is collaborative, pragmatic, structured, and focused on current problems.

We have previously worked with Boundary Objects Theory, where using a digital object can be seen as a way to transfer knowledge between different knowledge domains [71]. Based on this work, for an object to function effectively, it needs to inherit features such as dynamism, flexibility, standardization, and a shared structure [72]. We believe iTandem has the potential to work as a boundary object. It accommodates the needs of the service user, and it is simple to engage in, hence adhering to the KISS principle. iTandem allows for the transfer of knowledge between service user and clinician with an inherent flexibility yet standardized approach. The ability to collaborate via this object allows for a deeper understanding of the problem, leading to more focused problem-solving.

iTandem differs from the numerous assessment tools offered as PROMs in Norwegian mental health care services at the moment. The positive aspects of PROMs lie in their basis on validated measures that aim to track progress in therapy. However, the extensive number of questions to be answered can be overwhelming, and the amount of data may be challenging for clinicians to manage and oversee effectively. PROMs often cover broad categories, and their questions can sometimes seem less relevant to the service users' specific problems. Therefore, it may be necessary to scale down the assessments to focus on a minimum that allows for qualified evaluations of how treatment progresses. This approach can alleviate the burden on clinicians and service users by combining different tools, such as apps like iTandem and more focused classic PROMs. Utilizing language and questions that are closely aligned with a clinical setting might be sufficient to establish a rapport and gather relevant information about progress.

Another important aspect is how we provide flexibility in services. While we offer options such as video consultations, PROMs, apps, and similar tools, the solutions presented to service users often rely on traditional approaches, as if one size fits all. Alternating between different tools and formats during treatment sessions could, in our view, make the treatment more effective. Insights from user reviews of mental health apps indicate that highly rated apps tend to offer a variety of options, functionalities, and content for users to choose from [73]. This is particularly relevant to attrition rates, as different tools may be needed during various phases of the recovery process. A lack of use could even be a positive sign, indicating that the problem has been resolved and the user has moved on. Additionally, Alqahtani et al [73] identified several common weaknesses of mental health apps, including poor usability, lack of personalization, credibility issues, security concerns, and inadequate customer service. We believe these findings align with the need for greater flexibility in services.

To address the future needs of health care services, we believe it is essential to examine how digital tools can be integrated into regular workflows. Greater insight is needed into the barriers and facilitators of implementation, as well as how health technology assessment methods can be used to evaluate which technologies benefit service users, clinicians, and organizations while also being economically effective.

Acknowledgments

The development of iTandem was funded by the South-Eastern Norway Regional Health Authority (reference number 17/00264-14). We want to thank all the service users and clinicians who willingly have offered their time and interest in this development process. We also want to thank the agile team at USIT, which is not represented as co-authors in this paper, for good collaboration. We especially thank the senior developer, Magnus Alderslyst Nygaard, and the highly skilled, creative designers. Finally, we will thank Jonathan Romm for the illustrations and Trine Vik Lagerberg for valuable support along the way.

Data Availability

Because of the small sample size and limitations in the written informed consent, the data are not available to other researchers.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Higher Resolution version of Figure 2.
[PNG File , 409 KB - formative_v9i1e57614_app1.png]

References

1. Brédart A, Bouleuc C, Dolbeault S. Doctor-patient communication and satisfaction with care in oncology. Curr Opin Oncol 2005 Jul;17(4):351-354. [doi: 10.1097/01.cco.0000167734.26454.30] [Medline: 15933466]
2. Arora NK. Interacting with cancer patients: the significance of physicians' communication behavior. Soc Sci Med 2003 Sep;57(5):791-806. [doi: 10.1016/s0277-9536(02)00449-5] [Medline: 12850107]

XSL•FO
RenderX

3. Lee SJ, Back AL, Block SD, Stewart SK. Enhancing physician-patient communication. Hematology Am Soc Hematol Educ Program 2002:464-483. [doi: 10.1182/asheducation-2002.1.464] [Medline: 12446437]
4. Aftab A, LaGrotta C, Zyzanski SJ, Mishra P, Mehdi SMA, Brown K, et al. Impact of psychiatric hospitalization on trust, disclosure and working alliance with the outpatient psychiatric provider: a pilot survey study. Cureus 2019 Apr 22;11(4):e4515 [FREE Full text] [doi: 10.7759/cureus.4515] [Medline: 31259124]
5. Rübig LL, Fuchshuber J, Köldorfer P, Rinner A, Fink A, Unterrainer H. Attachment and therapeutic alliance in substance use disorders: initial findings for treatment in the therapeutic community. Front Psychiatry 2021;12:730876 [FREE Full text] [doi: 10.3389/fpsyt.2021.730876] [Medline: 34858223]
6. Reddy MS. Insight and psychosis. Indian J Psychol Med 2015;37(3):257-260 [FREE Full text] [doi: 10.4103/0253-7176.162909] [Medline: 26664072]
7. Harvey PD, Bosia M, Cavallaro R, Howes OD, Kahn RS, Leucht S, et al. Cognitive dysfunction in schizophrenia: an expert group paper on the current state of the art. Schizophr Res Cogn 2022 Sep;29:100249 [FREE Full text] [doi: 10.1016/j.scog.2022.100249] [Medline: 35345598]
8. Lyngstad SH, Gardsjord ES, Engen MJ, Haatveit B, Ihler HM, Wedervang-Resell K, et al. Trajectory and early predictors of apathy development in first-episode psychosis and healthy controls: a 10-year follow-up study. Eur Arch Psychiatry Clin Neurosci 2020 Sep;270(6):709-722 [FREE Full text] [doi: 10.1007/s00406-020-01112-3] [Medline: 32130475]
9. Ihler HM, Lyngstad SH, Mørch-Johnsen LE, Lagerberg TV, Melle I, Romm KL. A transdiagnostic approach to negative symptoms: exploring factor structure and negative symptoms in bipolar disorders. Front Psychiatry 2023;14:1136097 [FREE Full text] [doi: 10.3389/fpsyt.2023.1136097] [Medline: 37398600]
10. Priebe S, Dimic S, Wildgrube C, Jankovic J, Cushing A, McCabe R. Good communication in psychiatry--a conceptual review. Eur Psychiatry 2011 Oct;26(7):403-407. [doi: 10.1016/j.eurpsy.2010.07.010] [Medline: 21571504]
11. Feldthusen C, Forsgren E, Wallström S, Andersson V, Löfqvist N, Sawatzky R, et al. Centredness in health care: a systematic overview of reviews. Health Expect 2022 Jun;25(3):885-901 [FREE Full text] [doi: 10.1111/hex.13461] [Medline: 35261138]
12. Brown K, Parry S. How do people with first episode psychosis experience therapeutic relationships with mental health practitioners? A narrative review. Psychosis 2022 Dec 22;16(2):155-166. [doi: 10.1080/17522439.2022.2160487]
13. Hamann J, Heres S. Adapting shared decision making for individuals with severe mental illness. Psychiatr Serv 2014 Dec 01;65(12):1483-1486. [doi: 10.1176/appi.ps.201400307] [Medline: 25756970]
14. Simmons M, Hetrick S, Jorm A. Shared decision-making: benefits, barriers and current opportunities for application. Australas Psychiatry 2010 Oct;18(5):394-397. [doi: 10.3109/10398562.2010.499944] [Medline: 20863175]
15. Zisman-Ilani Y, Barnett E, Harik J, Pavlo A, O'Connell M. Expanding the concept of shared decision making for mental health: systematic search and scoping review of interventions. MHRJ 2017 Sep 11;22(3):191-213. [doi: 10.1108/mhrj-01-2017-0002]
16. Zisman-Ilani Y, Lysaker PH, Hasson-Ohayon I. Shared risk taking: shared decision making in serious mental illness. Psychiatr Serv 2021 Apr 01;72(4):461-463. [doi: 10.1176/appi.ps.202000156] [Medline: 33557595]
17. Castonguay LG, Constantino MJ, Holtforth MG. The working alliance: Where are we and where should we go? Psychotherapy (Chic) 2006;43(3):271-279. [doi: 10.1037/0033-3204.43.3.271] [Medline: 22122096]
18. Weldring T, Smith SMS. Patient-reported outcomes (PROs) and patient-reported outcome measures (PROMs). Health Serv Insights 2013;6:61-68 [FREE Full text] [doi: 10.4137/HSI.S11093] [Medline: 25114561]
19. Nguyen H, Butow P, Dhillon H, Sundaresan P. A review of the barriers to using patient-reported outcomes (PROs) and patient-reported outcome measures (PROMs) in routine cancer care. J Med Radiat Sci 2021 Jun;68(2):186-195 [FREE Full text] [doi: 10.1002/jmrs.421] [Medline: 32815314]
20. Nwosu A, Boardman S, Husain MM, Doraiswamy PM. Digital therapeutics for mental health: is attrition the Achilles heel? Front Psychiatry 2022;13:900615 [FREE Full text] [doi: 10.3389/fpsyt.2022.900615] [Medline: 35982936]
21. Szinay D, Jones A, Chadborn T, Brown J, Naughton F. Influences on the uptake of and engagement with health and well-being smartphone apps: systematic review. J Med Internet Res 2020 May 29;22(5):e17572 [FREE Full text] [doi: 10.2196/17572] [Medline: 32348255]
22. Mordecai D, Histon T, Neuwirth E, Heisler WS, Kraft A, Bang Y, et al. How Kaiser permanente created a mental health and wellness digital ecosystem. NEJM Catalyst 2021 Jan;2(1):1 [FREE Full text] [doi: 10.1056/cat.20.0295]
23. Dalzell T. The Routledge Dictionary of Modern American Slang and Unconventional English. London, UK: Routledge; Jul 25, 2008:595.
24. Alwin DF, Beattie BA. The KISS principle in survey design. Sociological Methodology 2016 Jun 21;46(1):121-152. [doi: 10.1177/0081175016641714]
25. Rich BR. Clarence Leonard (Kelly) Johnson 1910-1990: A Biographical Memoir. Washington, DC: The National Academies Press; 1995:221-241.
26. Vial S, Boudhraâ S, Dumont M. Human-centered design approaches in digital mental health interventions: exploratory mapping review. JMIR Ment Health 2022 Jun 07;9(6):e35591 [FREE Full text] [doi: 10.2196/35591] [Medline: 35671081]
27. Halskov K, Hansen N. The diversity of participatory design research practice at PDC 2002–2012. International Journal of Human-Computer Studies 2015 Feb;74:81-92 [FREE Full text] [doi: 10.1016/j.ijhcs.2014.09.003]

XSL•FO
RenderX

28. Bergin AD, Vallejos EP, Davies EB, Daley D, Ford T, Harold G, et al. Preventive digital mental health interventions for children and young people: a review of the design and reporting of research. NPJ Digit Med 2020;3:133 [FREE Full text] [doi: 10.1038/s41746-020-00339-7] [Medline: 33083568]
29. Romm KL, Gardsjord ES, Gjermundsen K, Aguirre Ulloa M, Berentzen L, Melle I. Designing easy access to care for first-episode psychosis in complex organizations. Early Interv Psychiatry 2019 Oct 27;13(5):1276-1282. [doi: 10.1111/eip.12802] [Medline: 30919597]
30. Roberts JP, Fisher TR, Trowbridge MJ, Bent C. A design thinking framework for healthcare management and innovation. Healthc (Amst) 2016 Mar;4(1):11-14. [doi: 10.1016/j.hjdsi.2015.12.002] [Medline: 27001093]
31. Altman M, Huang TT, Breland JY. Design thinking in health care. Prev Chronic Dis 2018 Sep 27;15:E117 [FREE Full text] [doi: 10.5888/pcd15.180128] [Medline: 30264690]
32. The Design Council. The Double Diamond. The Design Council. URL: https://www.designcouncil.org.uk/our-resources/the-double-diamond/ [accessed 2024-12-13]
33. Laskemoen JF, Simonsen C, Büchmann C, Barrett EA, Bjella T, Lagerberg TV, et al. Sleep disturbances in schizophrenia spectrum and bipolar disorders - a transdiagnostic perspective. Compr Psychiatry 2019 May;91:6-12 [FREE Full text] [doi: 10.1016/j.comppsych.2019.02.006] [Medline: 30856497]
34. Sylvia LG, Chang WC, Kamali M, Tohen M, Kinrys G, Deckersbach T, et al. Sleep disturbance may impact treatment outcome in bipolar disorder: a preliminary investigation in the context of a large comparative effectiveness trial. J Affect Disord 2018 Jan 01;225:563-568. [doi: 10.1016/j.jad.2017.08.056] [Medline: 28881294]
35. Carney CE, Buysse DJ, Ancoli-Israel S, Edinger JD, Krystal AD, Lichstein KL, et al. The consensus sleep diary: standardizing prospective sleep self-monitoring. Sleep 2012 Feb 01;35(2):287-302 [FREE Full text] [doi: 10.5665/sleep.1642] [Medline: 22294820]
36. www.psykose-bipolar.no. Psykose/Bipolar. Oslo, Norway: Oslo University Hospital URL: https://www.psykose-bipolar.no/bipolar-lidelse/tidlige-tegn-pa-bipolar-lidelse/ [accessed 2024-12-17]
37. Zhu Y, Li C, Huhn M, Rothe P, Krause M, Bighelli I, et al. How well do patients with a first episode of schizophrenia respond to antipsychotics: a systematic review and meta-analysis. Eur Neuropsychopharmacol 2017 Sep;27(9):835-844. [doi: 10.1016/j.euroneuro.2017.06.011] [Medline: 28669774]
38. Keramatian K, Chakrabarty T, Saraf G, Yatham LN. New developments in the use of atypical antipsychotics in the treatment of bipolar disorder: a systematic review of recent randomized controlled trials. Curr Psychiatry Rep 2021 May 08;23(7):39. [doi: 10.1007/s11920-021-01252-w] [Medline: 33963957]
39. Kishi T, Ikuta T, Matsuda Y, Sakuma K, Okuya M, Mishima K, et al. Mood stabilizers and/or antipsychotics for bipolar disorder in the maintenance phase: a systematic review and network meta-analysis of randomized controlled trials. Mol Psychiatry 2021 Aug;26(8):4146-4157 [FREE Full text] [doi: 10.1038/s41380-020-00946-6] [Medline: 33177610]
40. Haddad PM, Correll CU. The acute efficacy of antipsychotics in schizophrenia: a review of recent meta-analyses. Ther Adv Psychopharmacol 2018 Nov;8(11):303-318 [FREE Full text] [doi: 10.1177/2045125318781475] [Medline: 30344997]
41. Bjornestad J, Lavik KO, Davidson L, Hjeltnes A, Moltu C, Veseth M. Antipsychotic treatment - a systematic literature review and meta-analysis of qualitative studies. J Ment Health 2020 Oct;29(5):513-523. [doi: 10.1080/09638237.2019.1581352] [Medline: 30862219]
42. Webster RK, Weinman J, Rubin GJ. How does the side-effect information in patient information leaflets influence peoples' side-effect expectations? A cross-sectional national survey of 18- to 65-year-olds in England. Health Expect 2017 Dec;20(6):1411-1420 [FREE Full text] [doi: 10.1111/hex.12584] [Medline: 28618140]
43. Lindström E, Lewander T, Malm U, Malt UF, Lublin H, Ahlfors UG. Patient-rated versus clinician-rated side effects of drug treatment in schizophrenia. Clinical validation of a self-rating version of the UKU Side Effect Rating Scale (UKU-SERS-Pat). Nord J Psychiatry 2001;55 Suppl 44:5-69. [doi: 10.1080/080394801317084428] [Medline: 11860666]
44. Leamy M, Bird V, Le Boutillier C, Williams J, Slade M. Conceptual framework for personal recovery in mental health: systematic review and narrative synthesis. Br J Psychiatry 2011 Dec;199(6):445-452. [doi: 10.1192/bjp.bp.110.083733] [Medline: 22130746]
45. Slade M, Amering M, Farkas M, Hamilton B, O'Hagan M, Panther G, et al. Uses and abuses of recovery: implementing recovery-oriented practices in mental health systems. World Psychiatry 2014 Feb;13(1):12-20 [FREE Full text] [doi: 10.1002/wps.20084] [Medline: 24497237]
46. Åsbø G, Ueland T, Haatveit B, Bjella T, Flaaten CB, Wold KF, et al. The time is ripe for a consensus definition of clinical recovery in first-episode psychosis: suggestions based on a 10-year follow-up study. Schizophr Bull 2022 Jun 21;48(4):839-849. [doi: 10.1093/schbul/sbac035] [Medline: 35419608]
47. Slade M, Brownell T, Rashid T, Schrank B. Positive Psychotherapy for Psychosis: A Clinician's Guide and Manual. London, UK: Routledge; Dec 19, 2016.
48. Neil ST, Kilbride M, Pitt L, Nothard S, Welford M, Sellwood W, et al. The questionnaire about the process of recovery (QPR): a measurement tool developed in collaboration with service users. Psychosis 2009 Aug;1(2):145-155. [doi: 10.1080/17522430902913450]

XSL•FO
RenderX

49. Tennant R, Hiller L, Fishwick R, Platt S, Joseph S, Weich S, et al. The Warwick-Edinburgh Mental Well-Being Scale (WEMWBS): development and UK validation. Health Qual Life Outcomes 2007 Nov 27;5:63 [FREE Full text] [doi: 10.1186/1477-7525-5-63] [Medline: 18042300]
50. Lehman A. A quality of life interview for the chronically mentally ill. Evaluation and Program Planning 1988 Jul 03;11(1):51-62 [FREE Full text] [doi: 10.1016/0149-7189(88)90033-X]
51. Dickens LR. Using gratitude to promote positive change: a series of meta-analyses investigating the effectiveness of gratitude interventions. Basic and Applied Social Psychology 2017 May 30;39(4):193-208. [doi: 10.1080/01973533.2017.1323638]
52. Tse S, Tsoi EW, Hamilton B, O'Hagan M, Shepherd G, Slade M, et al. Uses of strength-based interventions for people with serious mental illness: a critical review. Int J Soc Psychiatry 2016 May;62(3):281-291. [doi: 10.1177/0020764015623970] [Medline: 26831826]
53. Riches S, Schrank B, Rashid T, Slade M. WELLFOCUS PPT: modifying positive psychotherapy for psychosis. Psychotherapy (Chic) 2016 Mar;53(1):68-77. [doi: 10.1037/pst0000013] [Medline: 25961372]
54. Upthegrove R, Marwaha S, Birchwood M. Depression and schizophrenia: cause, consequence, or trans-diagnostic issue? Schizophr Bull 2017 Mar 01;43(2):240-244 [FREE Full text] [doi: 10.1093/schbul/sbw097] [Medline: 27421793]
55. Alsheikh AM, Elemam MO, El-Bahnasawi M. Treatment of depression with alcohol and substance dependence: a systematic review. Cureus 2020 Oct 26;12(10):e11168 [FREE Full text] [doi: 10.7759/cureus.11168] [Medline: 33133799]
56. Skar-Fröding R, Clausen H, Šaltytė Benth J, Ruud T, Slade M, S Heiervang K. Associations between personal recovery and service user-rated versus clinician-rated clinical recovery, a cross-sectional study. BMC Psychiatry 2022 Jan 18;22(1):42 [FREE Full text] [doi: 10.1186/s12888-022-03691-y] [Medline: 35042494]
57. Yasui-Furukori N, Nakamura K. Bipolar disorder recurrence prevention using self-monitoring daily mood charts: case reports from a 5 year period. NDT 2017 Mar;Volume 13:733-736. [doi: 10.2147/ndt.s132355]
58. Sachs GS, Guille C, McMurrich SL. A clinical monitoring form for mood disorders. Bipolar Disord 2002 Oct;4(5):323-327. [doi: 10.1034/j.1399-5618.2002.01195.x] [Medline: 12479665]
59. Slade M. Implementing shared decision making in routine mental health care. World Psychiatry 2017 Jun;16(2):146-153 [FREE Full text] [doi: 10.1002/wps.20412] [Medline: 28498575]
60. Galderisi S, Mucci A, Dollfus S, Nordentoft M, Falkai P, Kaiser S, et al. EPA guidance on assessment of negative symptoms in schizophrenia. Eur Psychiatry 2021 Feb 18;64(1):e23 [FREE Full text] [doi: 10.1192/j.eurpsy.2021.11] [Medline: 33597064]
61. Dollfus S, Mach C, Morello R. Self-evaluation of negative symptoms: a novel tool to assess negative symptoms. Schizophr Bull 2016 May;42(3):571-578 [FREE Full text] [doi: 10.1093/schbul/sbv161] [Medline: 26564898]
62. Stewart V, McMillan SS, Hu J, Ng R, El-Den S, O'Reilly C, et al. Goal planning in mental health service delivery: a systematic integrative review. Front Psychiatry 2022;13:1057915 [FREE Full text] [doi: 10.3389/fpsyt.2022.1057915] [Medline: 36601527]
63. Rose G, Smith L. Mental health recovery, goal setting and working alliance in an Australian community-managed organisation. Health Psychol Open 2018;5(1):2055102918774674 [FREE Full text] [doi: 10.1177/2055102918774674] [Medline: 29785279]
64. Hartz SM, Pato CN, Medeiros H, Cavazos-Rehg P, Sobell JL, Knowles JA, Genomic Psychiatry Cohort Consortium. Comorbidity of severe psychotic disorders with measures of substance use. JAMA Psychiatry 2014 Mar;71(3):248-254 [FREE Full text] [doi: 10.1001/jamapsychiatry.2013.3726] [Medline: 24382686]
65. Hunt GE, Malhi GS, Cleary M, Lai HMX, Sitharthan T. Prevalence of comorbid bipolar and substance use disorders in clinical settings, 1990-2015: systematic review and meta-analysis. J Affect Disord 2016 Dec;206:331-349. [doi: 10.1016/j.jad.2016.07.011] [Medline: 27476137]
66. Richert T, Anderberg M, Dahlberg M. Mental health problems among young people in substance abuse treatment in Sweden. Subst Abuse Treat Prev Policy 2020 Jun 24;15(1):43 [FREE Full text] [doi: 10.1186/s13011-020-00282-6] [Medline: 32580732]
67. Galvani S. Motivational Interviewing: Helping People Change (3rd edition). Social Work Education 2014 Mar 04;33(6):854-855. [doi: 10.1080/02615479.2014.894351]
68. Friedel AL, Siegel S, Kirstein CF, Gerigk M, Bingel U, Diehl A, et al. Measuring patient experience and patient satisfaction-how are we doing it and why does it matter? A comparison of European and U.S. American approaches. Healthcare (Basel) 2023 Mar 08;11(6):797 [FREE Full text] [doi: 10.3390/healthcare11060797] [Medline: 36981454]
69. Kingsley C, Patel S. Patient-reported outcome measures and patient-reported experience measures. BJA Education 2017 Apr;17(4):137-144. [doi: 10.1093/bjaed/mkw060]
70. Boardman J, Dave S. Person-centred care and psychiatry: some key perspectives. BJPsych Int 2020 Aug;17(3):65-68 [FREE Full text] [doi: 10.1192/bji.2020.21] [Medline: 34287426]
71. Saidi T, Mork E, Aminoff S, Lobban F, Romm KL. Crossing boundaries in the delivery of healthcare - a qualitative study of an eHealth intervention in relation to boundary object theory. Digit Health 2023;9:20552076231196970 [FREE Full text] [doi: 10.1177/20552076231196970] [Medline: 37662674]
72. Star SL, Griesemer JR. Institutional Ecology, `Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. Soc Stud Sci 1989 Aug 01;19(3):387-420. [doi: 10.1177/030631289019003001]

XSL•FO
RenderX

73. Alqahtani F, Orji R. Insights from user reviews to improve mental health apps. Health Informatics J 2020 Sep;26(3):2042-2066 [FREE Full text] [doi: 10.1177/1460458219896492] [Medline: 31920160]

Abbreviations

KISS: keep it simple, stupid
OUH: Oslo University Hospital
PREM: patient-reported experience measure
PROM: patient-reported outcome measure
TIPS SE: Early Intervention in Psychosis Advisory Unit for South East Norway
TSD: Services for Sensitive Data Tjenester for Sensitive Data
UiO: University of Oslo
USIT: University of Oslo's Center for Information Technology
UX: user experience

Edited by A Mavragani; submitted 22.02.24; peer-reviewed by A Oostdyk, A Kurilla; comments to author 14.10.24; revised version received 28.10.24; accepted 06.11.24; published 17.01.25.

Please cite as:
Romm KL, Skoge M, Barrett EA, Berentzen LC, Bergsager D, Fugelli P, Bjella T, Gardsjord ES, Kling K, Kruse SH, Kværner KJ, Melle I, Mork E, Ihler HM, Rognli EB, Simonsen C, Værnes TG, Aminoff SR
A Mobile Health Intervention to Support Collaborative Decision-Making in Mental Health Care: Development and Usability
JMIR Form Res 2025;9:e57614
URL: https://formative.jmir.org/2025/1/e57614
doi:10.2196/57614
PMID:

©Kristin Lie Romm, Mari Skoge, Elizabeth Ann Barrett, Lars-Christian Berentzen, Dagfinn Bergsager, Pål Fugelli, Thomas Bjella, Erlend Strand Gardsjord, Kristine Kling, Sindre Hembre Kruse, Kari Jorunn Kværner, Ingrid Melle, Erlend Mork, Henrik Myhre Ihler, Eline Borger Rognli, Carmen Simonsen, Tor Gunnar Værnes, Sofie Ragnhild Aminoff. Originally published in JMIR Formative Research (https://formative.jmir.org), 17.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Mindfulness-Based App Intervention for Pregnant Women: Qualitative Evaluation of a Prototype Using Multiple Case Studies

Silvia Rizzi[1], MD; Maria Chiara Pavesi[2], MD; Alessia Moser[2], MD; Francesca Paolazzi[2], MD; Michele Marchesoni[3], MD; Stefania Poggianella[1], MD; Erik Gadotti[2], MD; Stefano Forti[4], MD

[1]Digital Health Research, Centre for Digital Health & Wellbeing, Fondazione Bruno Kessler, Trento, Italy
[2]Istituto Pavoniano Artigianelli, Trento, Italy
[3]Digital Health Innovation Lab, Centre for Digital Health & Wellbeing, Fondazione Bruno Kessler, Trento, Italy
[4]Centre for Digital Health & Wellbeing, Fondazione Bruno Kessler, Trento, Italy

Corresponding Author:
Silvia Rizzi, MD
Digital Health Research
Centre for Digital Health & Wellbeing
Fondazione Bruno Kessler
Via Sommarive 18
Trento, 38123
Italy
Phone: 39 0461312415
Email: srizzi@fbk.eu

Abstract

Background: Pregnancy is a complex period characterized by significant transformations. How a woman adapts to these changes can affect her quality of life and psychological well-being. Recently developed digital solutions have assumed a crucial role in supporting the psychological well-being of pregnant women. However, these tools have mainly been developed for women who already present clinically relevant psychological symptoms or mental disorders.

Objective: This study aimed to develop a mindfulness-based well-being intervention for all pregnant women that can be delivered electronically and guided by an online assistant with wide reach and dissemination. This paper aimed to describe a prototype technology-based mindfulness intervention's design and development process for pregnant women, including the exploration phase, intervention content development, and iterative software development (including design, development, and formative evaluation of paper and low-fidelity prototypes).

Methods: Design and development processes were iterative and performed in close collaboration with key stakeholders (N=15), domain experts including mindfulness experts (n=2), communication experts (n=2), and psychologists (n=3), and target users including pregnant women (n=2), mothers with young children (n=2), and midwives (n=4). User-centered and service design methods, such as interviews and usability testing, were included to ensure user involvement in each phase. Domain experts evaluated a paper prototype, while target users evaluated a low-fidelity prototype. Intervention content was developed by psychologists and mindfulness experts based on the Mindfulness-Based Childbirth and Parenting program and adjusted to an electronic format through multiple iterations with stakeholders.

Results: An 8-session intervention in a prototype electronic format using text, audio, video, and images was designed. In general, the prototypes were evaluated positively by the users involved. The questionnaires showed that domain experts, for instance, positively evaluated chatbot-related aspects such as empathy and comprehensibility of the terms used and rated the mindfulness traces present as supportive and functional. The target users found the content interesting and clear. However, both parties regarded the listening as not fully active. In addition, the interviews made it possible to pick up useful suggestions in order to refine the intervention. Domain experts suggested incorporating auditory components alongside textual content or substituting text entirely with auditory or audiovisual formats. Debate surrounded the inclusion of background music in mindfulness exercises, with opinions divided on its potential to either distract or aid in engagement. The target users proposed to supplement the app with some face-to-face meetings at crucial moments of the course, such as the beginning and the end.

XSL•FO
RenderX

Conclusions: This study illustrates how user-centered and service designs can be applied to identify and incorporate essential stakeholder aspects in the design and development process. Combined with evidence-based concepts, this process facilitated the development of a mindfulness intervention designed for the end users, in this case, pregnant women.

(JMIR Form Res 2025;9:e58265) doi:10.2196/58265

KEYWORDS

mindfulness; promoting well-being; pregnancy; eHealth; mHealth; mobile apps; development; usability; user-centered design; well-being; maternal health; digital health; intervention; design; preliminary testing; technology-based; interview; multidisciplinary approach; mother; women; WhatsApp; email; midwife

Introduction

Background

In a woman's life, the transition to motherhood is a time of profound change, requiring reorganizing her identity [1]. A woman's general well-being is affected by pregnancy, as various changes are expected at the physical, mental, and social levels [2]. Several studies are presented in the literature showing the presence of psychological symptoms in pregnant women. The presence of symptoms of anxiety, stress, and depression are most commonly emphasized [3,4]. Regarding anxiety and psychosocial stress, it is estimated that between 12.2% and 39% of pregnant women may suffer from these conditions [5]. Psychosocial anxiety is defined as an imbalance resulting precisely from the challenges of the new condition [6]. Therefore, a woman's quality of life and her psychological health are affected by how she reacts to all these changes. The vulnerable period of pregnancy and early motherhood also often involves increased worries and fears related to managing the impending changes associated with childbirth and parenting [7,8]. Woods and colleagues [9] found a spectrum of stress symptoms ranging from 6% (91/1522; high levels) to 78% (1190/1522; medium-low stress levels). Only 16% (241/1522) of the sample reported nothing. Another clinically significant condition in pregnant women is perinatal depression [10]. This occurs in 14.5% during pregnancy and reaches up to 49% in the first year after delivery [3].

Although psychoeducational interventions aimed at the psychological well-being of pregnant women are still scarce, ensuring a positive pregnancy experience is crucial, as it lays the foundation for cognitive motivation and a healthy future relationship between mother and child [11]. The use of digital tools represents a promising solution. Indeed, digital interventions offer accessibility and flexibility and overcome social stigma, making them an optimal choice for women seeking support [12,13].

Recognizing the critical need to promote psychological well-being during the transition to motherhood, especially in the digital age, becomes imperative.

According to Kabat-Zinn [14], mindfulness is defined as "the awareness that arises from paying attention, on purpose, in the present moment, and non-judgmentally." This practice involves actively noticing one's thoughts, emotions, and bodily sensations and fostering an attitude of openness and acceptance. Kabat-Zinn [15] emphasizes the significance of mindfulness in reducing stress and enhancing overall well-being. Mindfulness practices enable pregnant women to manage stress, anxiety, and depression actively, equipping them with valuable skills for various situations, including childbirth [14,16-19]. For this reason, mindfulness interventions delivered through web-based platforms, for example, have shown promise in improving the psychological well-being of pregnant women [17,20,21].

Therefore, psychoeducational interventions using mindfulness-based digital solutions have significant potential to promote psychological well-being, foster women's psychological adjustment during physiological pregnancies, and prevent clinically relevant symptoms [22].

This study illustrates the creation of a tailored mindfulness intervention for pregnant women using technology and a dedicated mobile app. Building on established mindfulness concepts proven effective in face-to-face interventions [23-25], the study adopted a user-centered design approach to ensure alignment of the intervention with users' needs and contextual use [26,27]. The fundamental principle of user-centered design is to actively involve users in the design and development stages, ensuring that they influence the evolution of the product [28].

The user-centered design approach ensures a holistic understanding of user perspectives by incorporating various methods that facilitate user participation at different development stages and involvement levels [29]. In addition, the study incorporated service design principles, emphasizing whole ecosystem experiences, such as how and when it is used, by whom, and where, rather than focusing exclusively on the end product [30]. This fusion of user-centered and service design approaches aimed to involve stakeholders, including pregnant women, health care providers, mindfulness and communication experts, psychologists, and midwives, throughout the design and development process. The overall goal was to create a final product that was easy to use, practical, engaging, and motivating, and fit seamlessly into the larger context of daily life and the challenges faced during pregnancy.

Mindfulness-Based Childbirth and Parenting

The Mindfulness-Based Childbirth and Parenting (MBCP) protocol was created in 1998 by Nancy Bardacke, a nurse, midwife, and mindfulness trainer. Through participation in the MBCP protocol, extrapolated from Kabat-Zinn's [15] Mindfulness-Based Stress Reduction (MBSR) protocol, pregnant women and their partners have the opportunity to learn mindfulness techniques to cope with the states of anxiety and stress ordinarily present during pregnancy, pain, and fear during childbirth, as well as to develop couple cooperation and parenting sensitivity [24,25,31]. MBCP was initially a group

XSL•FO
RenderX

course held for 3 hours once a week for 9 weeks. The 9-week MBCP course includes a day of silent meditation between weeks 6 and 7 and an additional reunion gathering after all women have given birth. Participants begin MBCP in the late second or early third trimester and are asked to commit to home practice using guided meditation CDs for 30 minutes a day, 6 days a week throughout the course. Formal mindfulness meditation instruction is given in each class, including sitting meditation, body scan, mindful yoga, and loving-kindness meditation, all tailored to include attention to mental and physical aspects of pregnancy (eg, noting fetal movements during the body scan) [24]. For a description of the main topics covered week by week, refer to Multimedia Appendix 1.

The literature indicates the extraordinary effectiveness of the MBCP protocol both during [23,32-36] and after pregnancy [20,37-39]. Creswell [22] found a moderate effect size (Cohen d=0.5) for mindfulness interventions in reducing anxiety and depression in pregnant women. Duncan et al [32] reported a pooled effect size of d=0.48 for mindfulness-based interventions in enhancing psychological well-being during pregnancy.

Some of the benefits of mindful birthing are (1) an increase in confidence and decrease in fear of childbirth; (2) tapping into deep inner resources for working with pain; (3) improving couple communication, connection, and cooperation; and (4) providing stress-reducing skills for greater joy and well-being [40].

Results suggest the effects of a group-based psychosocial intervention during pregnancy may endure for years, well beyond the initial perinatal period. Investing in prevention and intervention efforts for mental health during pregnancy may have long-term benefits for the well-being of women [39].

There are also several potential benefits to participating in a mindfulness program individually, especially during pregnancy. Participants can choose when and where to practice, making it easier to incorporate mindfulness into a busy schedule or daily routine. Practicing mindfulness alone can also help women concentrate more deeply on their thoughts, feelings, and sensations without distractions from group dynamics, encouraging deeper self-reflection and self-awareness, fostering a better understanding of personal experiences and emotions. For those who may feel anxious in group settings, individual participation can provide a more comfortable environment to explore mindfulness. Individuals can also select specific practices or techniques that resonate with them, creating a more personalized experience; this can empower participants to develop their mindfulness skills independently, fostering a sense of ownership over their mental well-being.

There are several digital apps that are based on MBCP principles. For example, Mindful Mama (Mindful Mamas Club) [41], an app that provides resources and guided meditations designed to support mothers during pregnancy and parenting, incorporating mindfulness techniques, or Mind the Bump (Smiling Mind) [42], which offers mindfulness practices, educational resources, and a community support feature for parents-to-be, integrating MBCP concepts. None of these involve the use of an online coach.

Our Study

MBCP intervention, which has already been validated and tested, will be fully available to users through digital tools. In particular, it will be delivered by a mobile app and guided by an online assistant, MAIA (developed by FBK). This research aims to assess the prototype of the MAIA app. This paper describes the prototype technology-based mindfulness intervention's design and development process for pregnant women, including the exploration phase, intervention content development, and iterative software development (including design, development, and formative evaluation of paper- and low-fidelity prototypes).

Methods

Overview

The design and development process encompassed a multidisciplinary approach and continuous, systematic evaluation throughout, as recommended in the Center for eHealth Research and Disease Management comprehensive roadmap approach to improve the uptake and impact of eHealth technologies [43].

The study's principal investigator (SR), a researcher and psychologist, led the intervention development work. During the design and development phase, the multidisciplinary project team, consisting of psychologists, mindfulness experts, communication experts, and information technology developers, had biweekly meetings. User-centered and service design methodologies were applied to ensure user involvement throughout the design and development process.

The mindfulness intervention was developed in iterative processes, following the ORBIT model [44], as shown in Figure 1, through a combination of (1) intervention content development (identified and adjusted from the MBCP program) and (2) iterative software development of a prototypes and formative evaluation. The development and iterative processes of the mindfulness intervention are shown in Figure 2.

XSL•FO
RenderX

Figure 1. The ORBIT model.

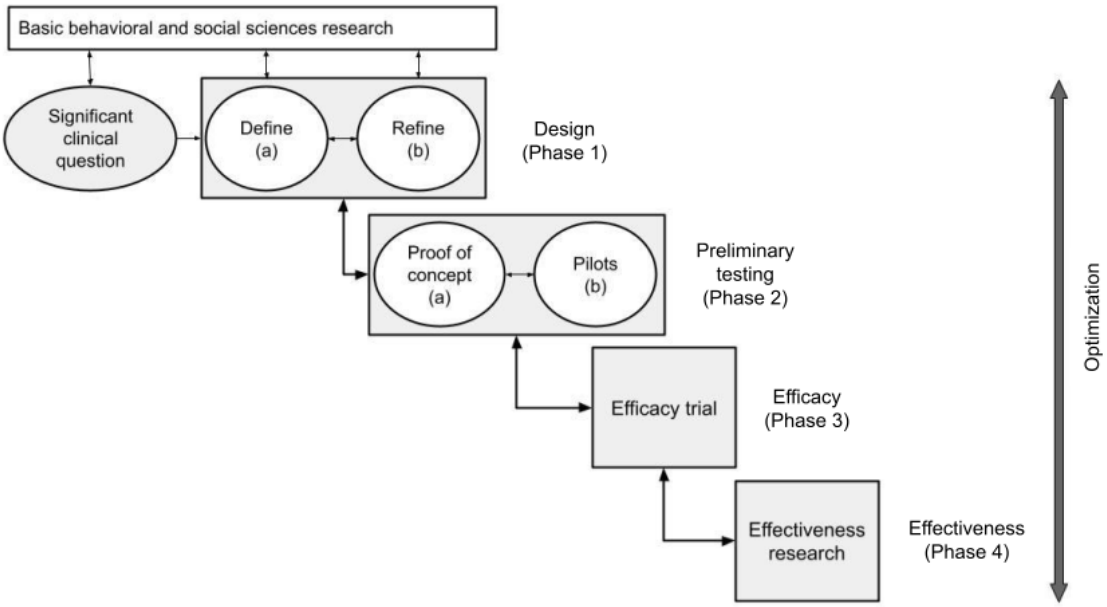

Figure 2. Development and iterative processes of the mindfulness intervention (Phase 1 of the ORBIT model—Design). MBCP: Mindfulness-Based Childbirth and Parenting.

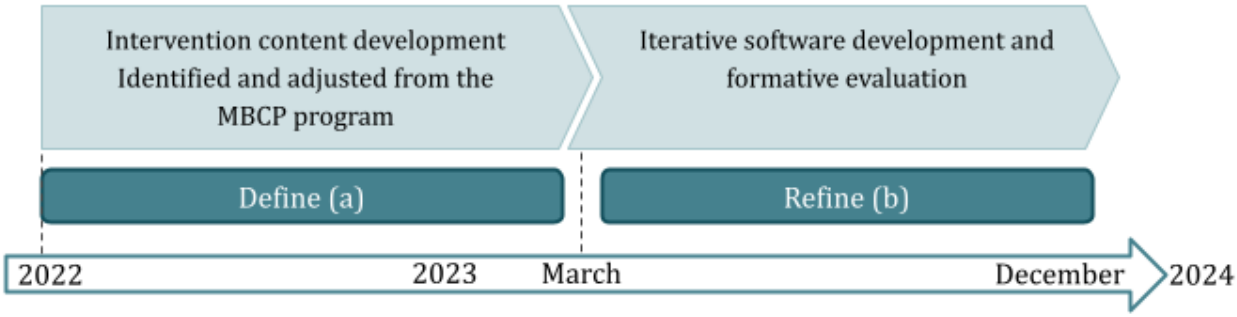

Intervention Content Development

Adapting the protocol into a digital format involved an initial phase of studying the literature and the protocol itself. Once the main themes being presented week by week were identified, a team of psychologists with communication skills worked to adapt the content into an individual digital format delivered by chatbot. First, the team developed the online coach's script and then multimedia content (videos, audio, and images) useful for delivering critical information in a way that is easier for women to understand than just text format.

The initial intervention content for this study was first developed by the primary investigator, then adapted and tailored by the entire research team through iterative processes to fit an 8-session-based intervention in electronic format through text, audio, video, and images. The intervention was user-tested to make the content easily accessible, confirm adaptation to an app format, and ensure that the scientific foundation for the intervention was intact. Table 1 lists and briefly describes the 8 sessions and the topics covered in each session.

The intervention was constructed in such a way that MAIA is always the first to interact with the user. It is currently not possible for the woman to ask specific questions. The chatbot provides information but also asks questions to help the user in self-reflection. The user's response text is not analyzed at this stage of the study as it is not useful for the purpose of the study.

XSL•FO
RenderX

Table 1. MAIA intervention overview of sessions (8 sessions preceded by an introductory one) and their content divided into formal practices, informal practices, and readings or homework to be carried out independently between sessions.

Session	Formal practices	Informal practices	Readings or homework
S0	• Welcome	• Mindfulness definition. Formal practice and informal practice. • Rituality, posture, and intentionality.	Preparation of materials and space
S1	• Awareness of breath • Raisin meditation	• Being with the child	__[a]
S2	• Body scan • Awareness of breath	• Mindful eating	"The Body scan"
S3	• Body scan • Awareness of breath	• Mindfulness in daily life (brushing your teeth, dishes, cooking, and so on)	"Mindfulness in daily life," "The pain of childbirth," enjoyable events calendar
S4	• Body scan • Awareness of breath • Mindfulness meditation-related pain	• Informal pain practice	Calendar of unpleasant events, "Mindfulness practices for being with pain"
S5	• Body scan • Awareness of breath • Mindfulness meditation-related pain • Mindful touch-care meditation	• Physical activity and informal practice, being with the child	—
S6	• Body scan • Sitting meditation • Walking meditation	—	BRANN[b] (make informed decisions) and Mindfulness day planning
S7[c]	• Body scan • Sitting meditation • Loving-kindness meditation	• Inquiry families of origin	—
S8	• Personal choice (choose one or more meditation) • Loving-kindness meditation	• Continue informal practices	"Breastfeeding" and Mindfulness slogans, tips for continuing on your own

[a]Not applicable.

[b]BRANN: Benefits, Risk, Alternatives, Nothing, Now.

[c]A day of mindfulness practice was held between S6 and S7.

Iterative Software Development and Formative Evaluation

Adapting the MBCP interventions by implementing the MAIA chatbot provides a novel approach to delivering psychological support. Users can engage in a comprehensive and effective intervention through personalized sessions. Specifically, in the initial session, users are asked to select 2 days of the week on which they want to be contacted by MAIA to conduct mindfulness sessions, along with their preferred time slot (morning, afternoon, or evening). In addition, on the basis of some answers given, MAIA provided them with customized information (eg, MAIA asks whether or not the user wants to do a review of the previous session, whether they prefer to do a certain mindfulness practice now or later, and so on and on the basis of the answers the dialogue tree takes different directions). Therefore, considering different sections of the app that could be integrated was critical to supporting user engagement, adherence, and overall well-being.

In the first iteration, in spring 2023, 2 psychologists and 2 communication experts tested preliminary. They gave feedback on the prototype to ensure that the intervention program was logically built.

At this co-design stage, the involvement of experienced professionals, in addition to end users, is of paramount importance. For this reason, 3 other psychologists, 2 mindfulness experts, and 2 communication experts were involved in a later study (test 1) by creating a paper prototype. Later, after the development of a low-fidelity prototype, end users (4 women and 4 clinicians) were also involved (test 2). Thus, a multistage evaluation of the MAIA prototype was conducted between August and December 2023, involving 15 participants, as elaborated in Figure 3.

XSL•FO
RenderX

Figure 3. Timeline of software development and formative evaluation specifying the people involved (N=19).

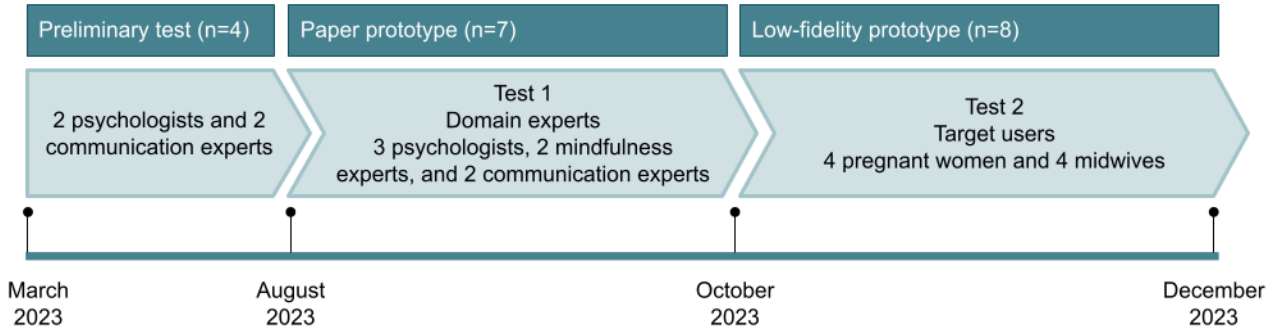

Identification of Variables

To gather the necessary information from tests 1 and 2 with the paper prototype and the low-fidelity prototype, key variables were identified for investigation—communication, session structure, materials, and content. Given the study's objectives, a list of subvariables was chosen to be created ad hoc for the research. Table 2 shows the chosen variables and their respective subvariables. It was also decided not to ask each person everything. Still, based on the domain, a grid was created to determine which subvariables to investigate with the people involved.

Table 2. List of variables and subvariables investigated and people involved in evaluating individual variables.

Variables and subvariables	Psychologists	Mindfulness experts	Communication experts	Target	Midwives
Communication					
Empathy and listening	✓	✓	✓	✓	✓
Smoothness and fluidity	✓	✓	✓	✓	✓
Chatbot interaction	✓	__[a]	✓	✓	✓
Lexicon	✓	✓	✓	✓	✓
Session structure					
Interaction length	✓	✓	✓	✓	✓
Notifications and reminders	✓	—	✓	✓	—
external links	—	—	✓	✓	✓
Materials					
Diary	✓	✓	✓	✓	✓
Audio tracks	✓	✓	✓	✓	✓
Infographics and videos	✓	✓	✓	✓	✓
Content					
Adherence to MBCP[b] protocol	—	✓	—	—	—
content evaluation	✓	✓	—	✓	✓
content clarity	—	✓	✓	✓	✓

[a]Not applicable.

[b]MBCP: Mindfulness-Based Childbirth and Parenting.

These variables were assessed through the semantic differential tool [45] and a semistructured interview. The semantic differential is an instrument consisting of a series of scales, each composed of a pair of bipolar adjectives between which a rating scale (generally 5- or 7-position) is placed.

Interviews, instead, are suitable for a more in-depth investigation of users' attitudes and preferences toward new technological solutions since open-ended discussions with users can help researchers better understand the issues and concerns related to the possible future adoption of these solutions.

Iterative Development and Prototypes

Based on needed content adjustments, the first prototype version of the software was developed. This initial paper prototype consisted of the scripted dialogue of the chatbot and screens presenting the different sections of the app. All professionals

XSL•FO
RenderX

involved were assigned specific days of the mindfulness sessions to view. They were then asked to complete an online questionnaire based on the semantic differential. In this study, we used a 4-point rating instead of 5 or 7 points to prevent neutral feedback allocation. The rating scale displayed semantically opposing adjectives with negative connotations on the left and positive connotations on the right, representing the extremes. The pairs of adjectives chosen explicitly for this study correspond to 1 of the 4 macrovariables under investigation. Participants were tasked with indicating, for each adjective pair, which adjective better characterized the specified variable related to the chatbot. For instance, if the adjective pair is "passive listening-active listening," associated with the "communication" variable, participants would choose point 1 or 2 if they perceive the chatbot's listening as passive and point 3 or 4 if they perceive it as more active. Following the questionnaire, a short semistructured interview was conducted to investigate participants' impressions. The interviewer posed specific questions, and then an opportunity was given for the interviewee to share additional open reflections. To facilitate subsequent analyses, the interview was recorded after the participant's direct consent was obtained. Multimedia Appendix 2 shows the list of topics and questions posed during the interviews.

To recruit participants, we adopted criterion-based sampling. Specifically, participants had to be psychologists, mindfulness instructors, or communication experts. Participants were recruited through instant messaging tools (WhatsApp [Meta]) and email.

In the first iteration, psychologists, mindfulness, and communication experts tested and gave feedback on the prototype to ensure that the intervention program was logically built. Participants were given oral and written information about the study and provided written informed consent before user testing.

At this stage, the Personas methodology was also found useful. Personas is a user-centered and service design method used to create and visualize fictional representations of the target group [46]. Personas is an effective method for all project team members to understand the target group for which the app is built better. The personas in this study contained information about the pregnancy background, challenges, and technology use. Refer to Multimedia Appendix 3 for illustrated examples of study Personas. Participants had to identify with one of these Personas before the reading.

The data obtained from the administered questionnaires underwent aggregated analysis. Therefore, participants' evaluations on a 4-point scale for each adjective pair were summed. Once summed, the arithmetic mean for each adjective pair's evaluation was calculated. After obtaining all the average scores, a dot chart was generated using Datawrapper (Datawrapper GmbH), a web-based charting and mapping tool [47], to observe the overall trend of the scores better and, consequently, the semantic pole toward which they tend to orient.

The information obtained during the interviews has been transcribed for a more detailed analysis. The respondents' answers were then categorized by topics within the 4 initially identified variables—communication, session structure, materials, and content. In addition to these variables, a new theme emerged concerning the "postintervention" phase. The categorized responses within the variables were further divided into their respective subvariables (for the complete list, refer to Table 2). Finally, they were classified into positive and negative aspects or suggestions (areas for improvement). By "suggestions," we refer to those aspects of the app that might need modification in a forward-looking perspective.

Following minor adjustments, the paper prototype was transformed into an electronic tool to simulate the chatbot and app concept. The intervention was delivered through a software platform called TreC-PerseO (FBK), which supports the delivery of digital behavior change interventions. This proprietary software architecture is designed to ensure privacy and certified workflows in a production environment while also facilitating prototyping by adapting to the steps outlined in the chosen methodology. In this phase, Telegram (Telegram Messenger Inc) chat was used to deliver the intervention (to validate the content without the need to collect personal user data). During this second iteration (test 2), the final version of the low-fidelity prototype was tested by 2 pregnant women, 2 mothers, and 4 midwives. The participants were assigned different days of the mindfulness sessions to review, after which they completed an online questionnaire based on the semantic differential. A short interview followed the questionnaire.

To recruit participants, we adopted criteria-based sampling. Specifically, participants had to meet the following criteria: be pregnant or mothers with children aged less than 1 year or be midwives. Participants were recruited through instant messaging tools (WhatsApp) and email. The sample collected was 8 participants—2 pregnant women, 2 mothers, and 4 midwives. Participants were given oral and written information about the study and provided written informed consent before user testing.

The data collected during the interviews has been transcribed for a more detailed analysis. The respondents' answers were then categorized by topics within the 4 initially identified variables. The categorized responses within the variables were further divided into their respective subvariables, and then they were classified into positive and negative aspects or suggestions (areas for improvement). The resulting participant's feedback data were again used to evaluate, refine, iteratively adjust, and upgrade the prototype.

Ethical Considerations

At the time of enrollment, all participants who freely decided to take part in the study were asked to sign an informed consent form after a careful explanation of the study, its aims, the level of involvement required, and the duration of the research, as well as all ethical issues concerning confidentiality. All data collected were anonymized before analysis.

The participants will then be informed of the study's results through the dissemination of articles.

In line with Italian Law No. 3/2018, ethical review is mandatory for clinical trials, therapeutic procedures, or studies involving health risks or sensitive data. This study, being nonclinical and

XSL•FO
RenderX

focused on prototype development, did not meet these criteria, justifying the absence of ethics committee involvement.

No compensation was provided for participants.

Results

Intervention Content Development

The iterative content development processes were parallel to app programming, and adjustments were made based on usability testing. Therefore, the app is designed to include 5 sections: the chatbot, Practices, Gallery, To Do List, and Diary.

Figure 4 shows some mockups of the app with its different sections.

Figure 4. Mockups of the MAIA app with its different sections.

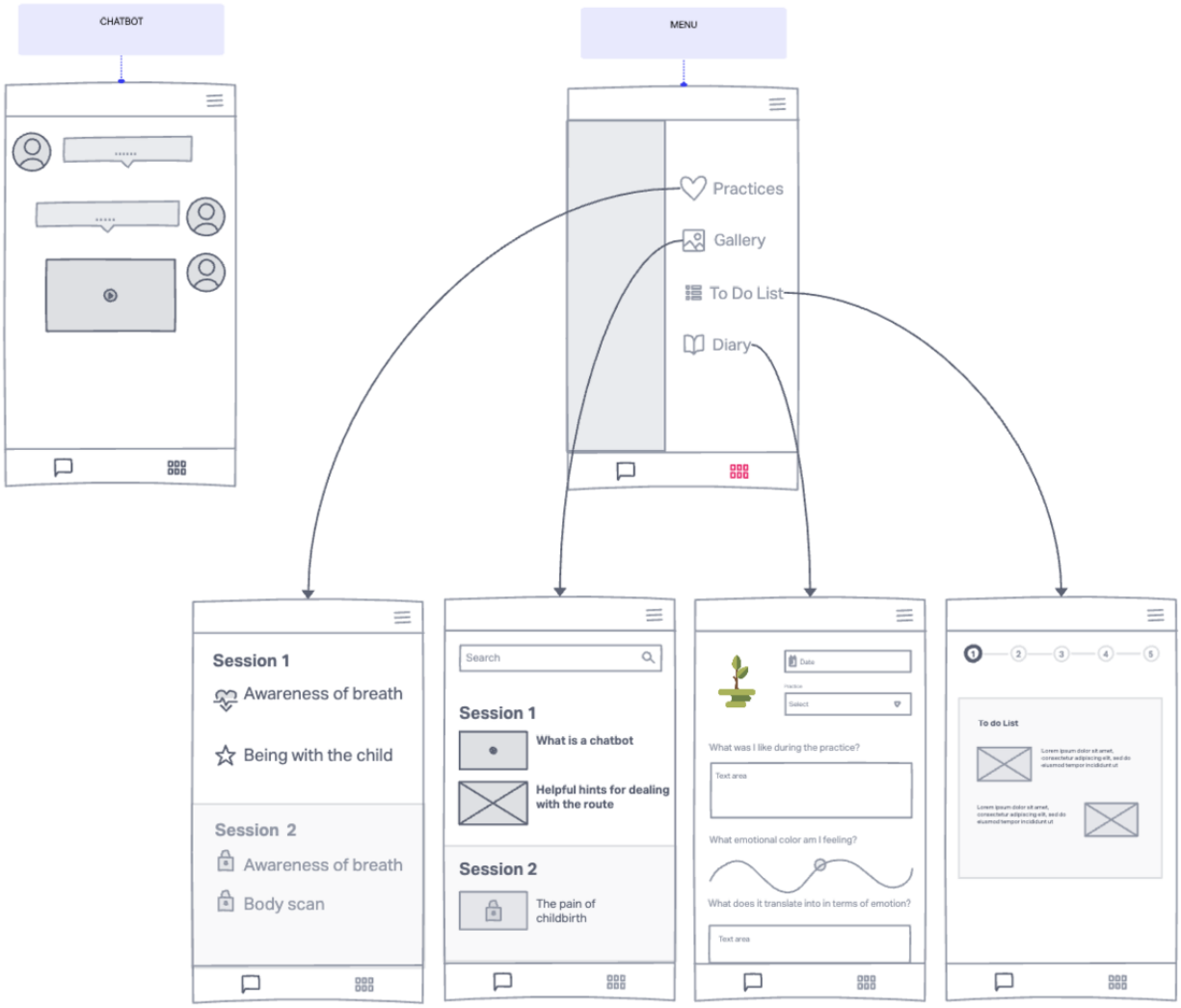

MAIA assigns homework during the session that the users should carry out during the subsequent days, so the app Diary section will support self-monitoring. We have also provided for a tree that grows as the woman progresses. All the exercises women need to do during the week are listed in the To-Do List section to help her remember what she needs to do. All the practices, divided week per week, are shown in the Practices section. The Gallery page features all multimedia material delivered by MAIA, such as educational videos, images, and intervention introductions.

Iterative Software Development and Formative Evaluation

Iterative Development and Preliminary Test

A total of 2 psychologists and 2 communication experts took part in this preliminary test phase. The objective of this evaluative phase was to test, through role-playing, the communicative style of dialogues—empathetic, welcoming, fluent, and smooth.

Therefore, it was possible to evaluate and revise MAIA's communication style through role-playing.

XSL•FO
RenderX

Regarding fluency and smoothness, the evaluation was primarily positive. To make the protocol more and more realistic, the dialogues were revised in some parts. Changes were made grammatically and syntactically to make the text more fluent and read smoothly. The changes also considered empathy. This makes it possible to create a feeling of trust in the relationship. Through empathic communication, the person can feel in a safe space, within which he or she is reassured and protected. Initially, the feeling while reading the dialogues could have been more apathetic communication. Therefore, multiple-choice questions were included about the individual practice (eg, How was this experience?) and the general course of action (eg, How did you do with mindfulness practices last time?). This allows MAIA to respond appropriately and more accurately based on the user's experience. Some examples are illustrated in Figure 5.

Figure 5. Examples of new multiple-choice questions included after preliminary tests.

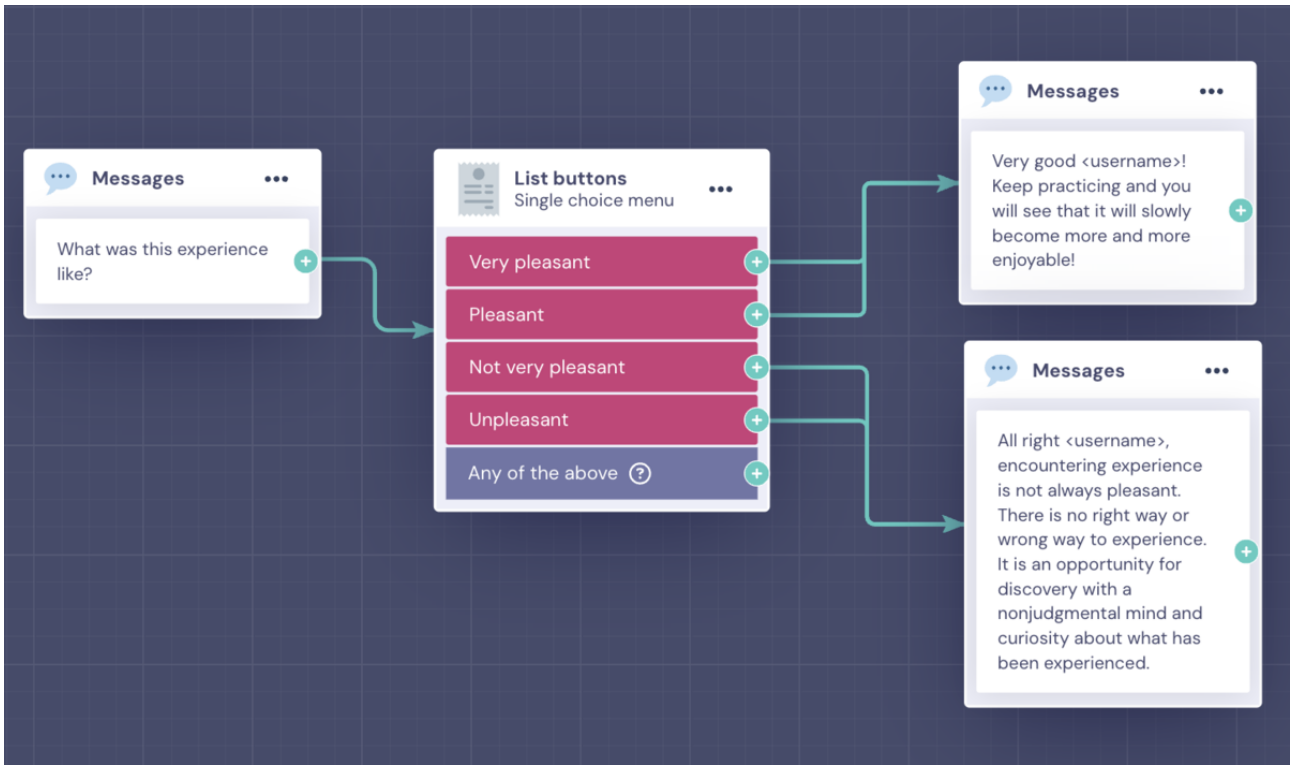

A new practice not provided for in the MBCP initiate protocol, namely the practice of 3 breaths before each session, was also included. The user receiving notification from MAIA can be anywhere and under any conditions. To help the person focus on the present moment and be ready for the practices offered later, it was decided to provide this new initial practice at each session.

Finally, the correctness of the videos and images being sent was checked. For each video link, consistency was checked against the instructions, and the topic was addressed. Everything was correct, so no changes had to be made.

Iterative Development and Paper Prototype

After modifying some parts of the dialogue, as described above, the first paper prototype was developed. In this phase, the sample involved consisted of 7 professionals—3 psychologists, 2 mindfulness instructors, and 2 communication experts.

Using a semantic differential-based questionnaire facilitates observing the respondents' average positioning with respect to the 4 macrovariables under investigation.

As can be seen from Figure 6, a clear tendency toward the positive semantic pole (right pole) emerges in the feedback. This means that a significant portion of the chatbot-related feature pairs were evaluated in the positive rather than the negative sense. For instance, the vocabulary used by the chatbot was judged as understandable rather than abstruse. However, feedback related to 4 pairs of features occupied a middle position between the positive and negative poles, that is, passive listening–active listening, cold-warm, troublesome-welcome, and soporific-activator.

XSL•FO
RenderX

Figure 6. Graphical representation of mean values derived from semantic differential administered to domain experts (psychologists, mindfulness instructors, and communication experts).

In total, 2 psychologists (experts in communication) conducted the qualitative interview to gather additional information. Table 3 briefly highlights key positive and negative aspects that emerged during the domain experts' interviews.

In particular, the analysis of the interviews unearthed several notable observations that garnered attention due to their recurrent mention by more experts. Certain interviewees advocated augmenting written textual content with an auditory counterpart or substituting it entirely with auditory or audiovisual formats. The diary, initially conceived as a personal and confidential tool, prevents external control. However, the process of diary entry may elicit personal experiences, necessitating potential consultation with a designated professional if the need arises.

A further debated aspect is the inclusion of background music in mindfulness exercise tracks. According to certain interviewees, such music might prove distracting, diverting attention away from the individual's introspective focus. Conversely, others contend that it could serve as an aid, particularly in the absence of direct human interaction. In summation, these insights are deemed fundamental for the following stages of research, wherein the perspectives of pregnant women directly engaged in this study can be further investigated. In addition, we underscore the presence of numerous positive aspects delineated in the interviews, including the empathetic and welcoming communication style, the fluidity of discourse, and the use of imparted content, among others.

XSL•FO
RenderX

Table 3. Positive and negative aspects emerged from interviews with domain experts (psychologists, mindfulness instructors, and communication experts).

Variables and subvariables	Positive aspects	Negative aspects (and suggestions)
Communication		
Empathy and listening	• Empathic communication is comprehensive, conveys calm, reassuring, warm, welcoming, and puts at ease • Nice dialogue, relaxing, helpful • Customized	• Include more sentences that also address the child • Responses are repetitive and somewhat mechanical • "How did it go (the process)?" is too judgmental
Smoothness and fluidity	• Very smooth • Enjoyable to read	—[a]
Chatbot interaction	• If I read, I can go at the speed I want • It is not artificial, it works	• Needs to be more engaging • Reading can be tedious • Audio-only or audio and video or written and audio
Lexicon	• Accessible to everyone: intuitive, simple, straightforward, and understandable • Gentle, suitable, and consistent terms	• The term "discomfort" may cause concern • The term "pain" is repeated
Session structure		
Interaction length	• Not demanding • Self-management of dialogue times	—
Notifications and reminders	• Helpful in saving and reviewing it without entering the app	—
Materials		
Diary	• Putting into writing what is felt allows for processing and gaining clarity	• Option to print the diary and write it by hand
Audio tracks	• Useful content • "Being with the pain": very helpful for preparing for childbirth • Use of different voices	• Banal video graphics • Remove the music • Add music • Lack of initial welcoming moment • Option to choose a male or female voice
Infographics and videos	• Beautiful graphics, practical, and charming • Nice color used • Option to print PDFs	—
Content		
Adherence to MBCP[b] protocol	• Positive and negative experiences of other women • Liked the 8 attitudes	• Providing two options (same content presented differently) based on past pregnancy experiences
Content evaluation	• Interesting content • Practices dedicated to both mother and baby	• Having too much time to think is only sometimes positive
Content clarity	• Engaging protocol	—
Other	—	• Understand how it went postpartum in light of these weeks • Hold an in-person meeting at the end of the protocol to make it more engaging

[a]Not available.

[b]MBCP: Mindfulness-Based Childbirth and Parenting.

XSL•FO
RenderX

Iterative Development and Low-Fidelity Prototypes

In total, 2 pregnant women, 2 moms, and 4 midwives participated in this second test phase. To better understand the app and the entire intervention, more realistic mockups were created (Figure 7). This allowed users to represent the app in its final appearance better.

In this case, using a semantic differential-based questionnaire facilitates the observation of the respondents' average positioning with respect to the 4 variables under investigation.

As illustrated in Figure 8, a tendency toward the positive semantic pole (right pole) emerges in the feedback also in these groups of participants. However, feedback related to 5 pairs of features occupied a middle position between the positive and negative poles for the pregnant woman and moms' group and 5 for the midwives group.

Figure 7. More realistic mockups of the MAIA app with its different sections.

XSL•FO
RenderX

Figure 8. Graphical representation of the mean values obtained from administering the semantic differential to target users.

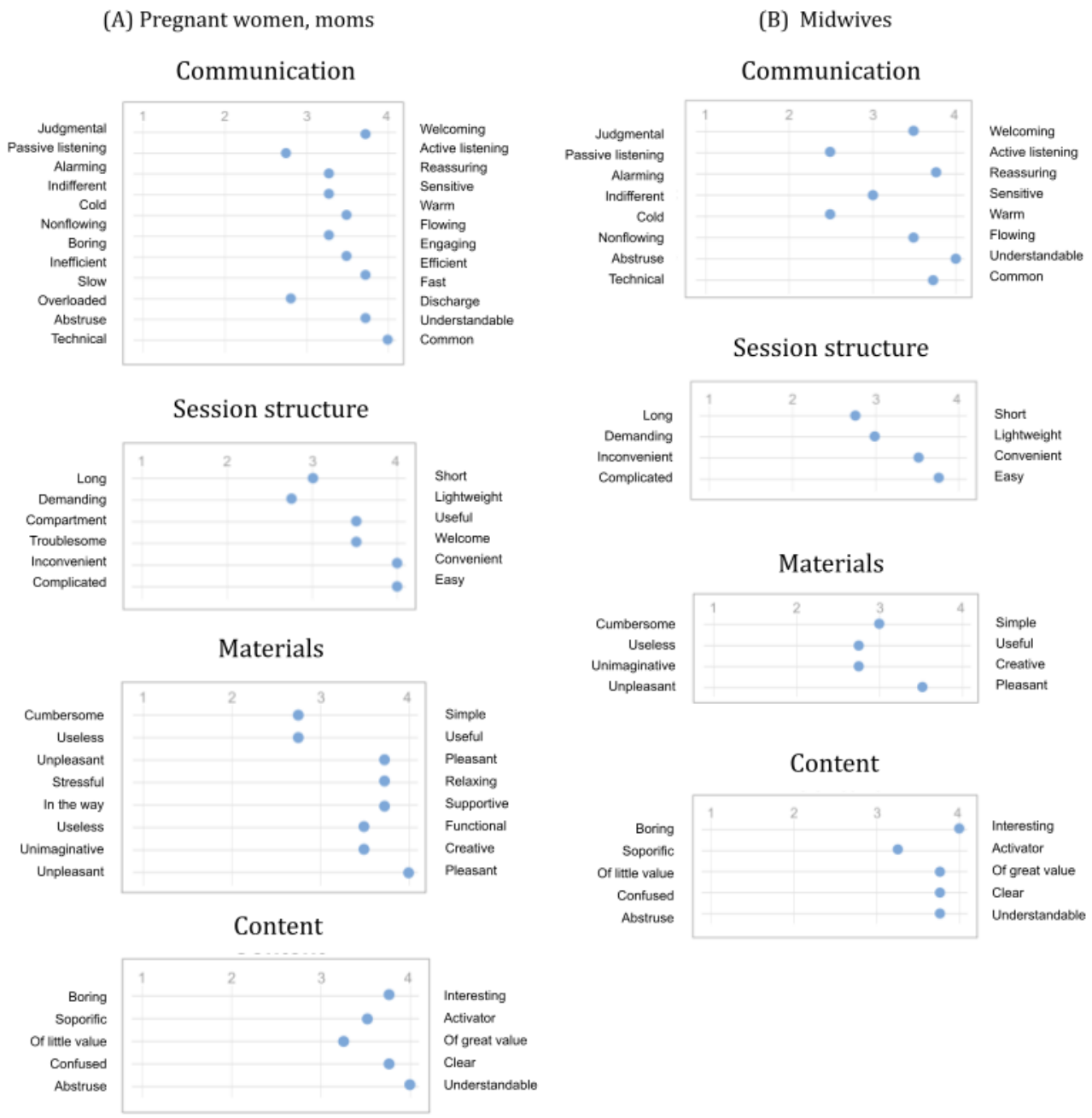

Furthermore, 2 psychologists (experts in communication) conducted the qualitative interview to gather additional information. Table 4 briefly highlights key positive and negative aspects that emerged during user interviews.

All respondents found the protocol content of great value, considering it interesting and useful. They appreciated the alternation of simple and streamlined proposals with more time-consuming ones. They consider it an added value to manage the timing of the course autonomously; with the app, they can postpone the performance of activities, reassured by the fact that they can catch up at another time. They also mostly consider the diary to be useful and usable as long as it is left as an option.

One of them suggested supplementing the stand-alone activities with a couple of face-to-face meetings (an initial one to explain the meaning and experience as a group and a central one for the more engaging exercises, such as the sultana and ice cube exercises).

The midwives dwelt more on the contents and found them correct and set out in a simple and comprehensible manner for all, giving useful advice to improve the video part in a couple of cases. One of them strongly urged the translation of the protocol into several languages.

Among the materials, the infographics were highly appreciated by both categories for their incisiveness and the fact that they were printable.

XSL•FO
RenderX

Table 4. Positive and negative aspects emerged from user interviews (pregnant women, moms, and midwives).

Variables and subvariables	Positive aspects	Negative aspects (and suggestions)
Communication		
Empathy and listening	• Nice dialogue • Communication puts at ease • Customized, nonjudgmental	• Responses are repetitive, somewhat mechanical • When the exercise does not work, it could be a question that goes into depth analyses and also advises for future exercises
Smoothness and fluidity	• Very smooth • Enjoyable to read	—[a]
Chatbot interaction	• It is not artificial, it works • It is fast and light, particularly when it proposes questions with multiple-choice answers	—
Lexicon	• Accessible to everyone: intuitive, simple, clear, and understandable	—
Session structure		
Interaction length	• Not demanding • Self-management of dialogue times	—
Notifications and reminders	• Useful even if you have no time at that moment	• Sometimes stressful because even at the scheduled time, you cannot do the exercises
Materials		
Diary	• Useful if optional • Interesting to focus on different exercises • Very useful to promote awareness	• It depends on the available time
Audio tracks	• Useful content • Good timing • Good voice	• Add music • Sometimes they are too long
Infographics and videos	• Beautiful graphics, effective, and charming • Very useful Infographics to focus on the most important information, and to repeat the exercise • Very useful images in the videos, especially for people with visual memory	• The pregnant woman is too thin • The eighth-week video on pain has a very heavy atmosphere and does not reflect the real situation in the delivery room (the bed is used very little, and the woman is never alone)
Content		
Content evaluation	• Interesting content • Engaging protocol • Very valuable content • The content on breastfeeding was very well done, it says the right things in the right way • The daily practical exercises are very interesting and engaging (eg, walking and teeth washing) • It is engaging because the commitment required is growing • The practice of three breaths is very useful and effective	—
Content clarity	• Very clear and simple to follow	• At first, it seems like the practice will always last 5 minutes

XSL•FO
RenderX

Variables and subvariables	Positive aspects	Negative aspects (and suggestions)
Other	• Very comfortable to have all you need in the app	• Desire to have other two weeks after the childbirth • Hold an in-person meeting at the beginning and at the end of the protocol to make it more engaging • It could be easier to follow the most demanding exercises in the team, with a guide

[a]Not available.

Discussion

Principal Findings

This study illustrates how user-centered and service designs can be applied to identify and incorporate essential stakeholder aspects in the design and development process. Combined with evidence-based concepts, this process facilitated the development of a mindfulness intervention designed for the end users, in this case, pregnant women.

Importing and adapting the intervention was a comprehensive and meticulous process. According to the ORBIT model, we examined a wide range of scientific studies on the efficacy of mindfulness during pregnancy and the postpartum period, focusing on the MBSR and the MBCP protocols. By integrating elements from both methodologies, we tailored key aspects of each protocol to create an intervention specifically designed for the needs of pregnant women. In addition, we considered other evidence and best practices in the field of mindfulness, ensuring the incorporation of empirically supported and culturally sensitive approaches (ie, the practice of the 3 breaths to anchor oneself in the present moment).

This integrative approach allowed us to develop a protocol that is scientifically grounded, practical, and suitable for the unique context of pregnancy.

Paying attention to the semantic differentials with lower values, we can see that in the case of listening, the answers of the users and midwives agree. In the interviews, 3 out of 4 women stated that they see the chatbot as the necessary tool to dispense material. Furthermore, 2 of them consider the dialogues to be valuable when it is fast and nondemanding (eg, when a multiple-choice answer is expected), which is also confirmed by the results on the overloaded or discharged differential. However, as seen from the other semantic differentials, the chatbot is welcoming, reassuring, and polite.

Consider the pair “passive listening–active listening,” referring to whether one feels actively or passively listened to by the chatbot; domain experts, on average, did not distinctly categorize the listening as active. This judgment may be due to the inherent limitation of digital apps—lacking a natural person on the other side of the screen responding to each user’s comment. The structure of this type of interaction, therefore, makes it difficult to feel fully heard, particularly in an active way.

One of the suggestions we received from users was to add audio in addition to the written text. This suggestion stems from sociocultural considerations. Indeed, nowadays, people increasingly use social networks, interacting with content that is short in duration and predominantly in the form of video rather than text. Reading may, therefore, appear more boring for some. Another focal point pertains to the diary feature, particularly the need to facilitate interaction with a designated professional.

Strengths, Limitations, and Future Directions

One of the main strengths of this study lies in the fact that it engaged both domain experts and target users from the very early stages. The suggestions that were gathered highlight different but equally important points of view for refining the intervention. Having midwives involved, in addition to pregnant women, once again allowed us to have an additional point of view. The midwives could give us suggestions (eg, editing a video so that it was more aligned with what is happening in the delivery room) that none of the other actors involved would have been able to provide.

One limitation, however, mainly concerns the development of the low-fidelity prototype and the number of users involved. Along with that prototype, we provided users with a mockup of the app, but interacting with such a prototype or with the final app is a very different experience. The purpose of this study, however, was not to evaluate the user experience with the app but to gather impressions of content, vocabulary, and communication style. For that reason, the tool used was deemed good enough. The next step will be to develop the application following the suggestions gleaned from this first study and then to set up a proof-of-concept study, following the orbit model, involving about 50 women to evaluate the acceptability, feasibility, and usability of the MAIA app [48].

Based on the data collected during this second study, we expect to have to make further modifications to the intervention and application. Following these modifications, we will finally be ready to field a randomized control trial to evaluate the efficacy of the intervention (the study protocol will be finalized by mid-2025). After conducting this randomized controlled trial, we also plan to conduct a real-world study in order to evaluate the effectiveness of the intervention and the app itself.

Conclusions

Developing a digital intervention is a complex procedure that requires work in different areas and involves different professionals. We aimed to build an intervention that was as suitable as possible for the target audience, both in terms of content and design. The contribution of specialists is undoubtedly an added value from the early stages in a coconstruction perspective. This survey is the first step in a

XSL•FO
RenderX

series of studies to validate the intervention. At this stage, it is still difficult to state how this research contributes to the broader evidence base supporting the use of mindfulness-based practices among pregnant women, as we are still at a preliminary stage. However, we can state that the impression of the users and experts involved is positive; as the results seem to be encouraging, the next step will be to conduct a study to assess the experience and engagement of pregnant women.

Acknowledgments

The research team would like to thank all the women involved, including mothers, midwives, and domain experts, who contributed to the success of this study in various capacities.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author upon reasonable request. Before disclosure, data are aggregated and anonymized.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Brief description of the main topics covered week by week in the Mindfulness-Based Childbirth and Parenting original program.
[PNG File , 463 KB - formative_v9i1e58265_app1.png]

Multimedia Appendix 2
Questions posed to participants attending the evaluation.
[PNG File , 126 KB - formative_v9i1e58265_app2.png]

Multimedia Appendix 3
Examples of personas.
[PNG File , 80 KB - formative_v9i1e58265_app3.png]

References

1. Stern DN. The Motherhood Constellation. United Kingdom: Routledge; 2018. [doi: 10.4324/9780429482489-12]
2. Lutkiewicz K, Bieleninik Ł, Cieślak M, Bidzan M. Maternal-infant bonding and its relationships with maternal depressive symptoms, stress and anxiety in the early postpartum period in a Polish sample. Int J Environ Res Public Health 2020;17(15):5427 [FREE Full text] [doi: 10.3390/ijerph17155427] [Medline: 32731490]
3. Gavin NI, Gaynes BN, Lohr KN, Meltzer-Brody S, Gartlehner G, Swinson T. Perinatal depression: a systematic review of prevalence and incidence. Obstet Gynecol 2005;106(5 Pt 1):1071-1083. [doi: 10.1097/01.AOG.0000183597.31630.db] [Medline: 16260528]
4. Goodman JH, Guarino A, Chenausky K, Klein L, Prager J, Petersen R, et al. CALM pregnancy: results of a pilot study of mindfulness-based cognitive therapy for perinatal anxiety. Arch Womens Ment Health 2014;17(5):373-387 [FREE Full text] [doi: 10.1007/s00737-013-0402-7] [Medline: 24449191]
5. Dunkel Schetter C. Psychological science on pregnancy: stress processes, biopsychosocial models, and emerging research issues. Annu Rev Psychol 2011;62:531-558. [doi: 10.1146/annurev.psych.031809.130727] [Medline: 21126184]
6. Zarenejad M, Yazdkhasti M, Rahimzadeh M, Tourzani ZM, Esmaelzadeh-Saeieh S. The effect of mindfulness-based stress reduction on maternal anxiety and self-efficacy: a randomized controlled trial. Brain Behav 2020;10(4):e01561 [FREE Full text] [doi: 10.1002/brb3.1561] [Medline: 32162450]
7. Conde A, Figueiredo B, Tendais I, Teixeira C, Costa R, Pacheco A, et al. Mother's anxiety and depression and associated risk factors during early pregnancy: effects on fetal growth and activity at 20-22 weeks of gestation. J Psychosom Obstet Gynaecol 2010;31(2):70-82 [FREE Full text] [doi: 10.3109/01674821003681464] [Medline: 20236029]
8. Li H, Bowen A, Bowen R, Muhajarine N, Balbuena L. Mood instability, depression, and anxiety in pregnancy and adverse neonatal outcomes. BMC Pregnancy Childbirth 2021;21(1):583 [FREE Full text] [doi: 10.1186/s12884-021-04021-y] [Medline: 34429072]
9. Woods SM, Melville JL, Guo Y, Fan MY, Gavin AR. Psychosocial stress during pregnancy. Obstetric Anesthesia Digest 2010;30(4):237 [FREE Full text] [doi: 10.1097/01.aoa.0000389617.09252.47]
10. Fonseca A, Gorayeb R, Canavarro MC. Women s help-seeking behaviours for depressive symptoms during the perinatal period: socio-demographic and clinical correlates and perceived barriers to seeking professional help. Midwifery 2015;31(12):1177-1185. [doi: 10.1016/j.midw.2015.09.002] [Medline: 26433622]

XSL•FO
RenderX

11. Rilling JK. The neural and hormonal bases of human parental care. Neuropsychologia 2013;51(4):731-747. [doi: 10.1016/j.neuropsychologia.2012.12.017] [Medline: 23333868]
12. Barrera AZ, Wickham RE, Muñoz RF. Online prevention of postpartum depression for Spanish- and English-speaking pregnant women: a pilot randomized controlled trial. Internet Interv 2015;2(3):257-265 [FREE Full text] [doi: 10.1016/j.invent.2015.06.002] [Medline: 26273567]
13. Philippe TJ, Sikder N, Jackson A, Koblanski ME, Liow E, Pilarinos A, et al. Digital health interventions for delivery of mental health care: systematic and comprehensive meta-review. JMIR Ment Health 2022;9(5):e35159 [FREE Full text] [doi: 10.2196/35159] [Medline: 35551058]
14. Kabat-Zinn J. Full catastrophe living: using the wisdom of your body and mind to face stress, pain, and illness. New York City: New York Dell Publishing Co; 1990.
15. Kabat-Zinn J. Wherever You Go, There You Are. London, United Kingdom: Piatkus; 1994. [doi: 10.1037/a0012744]
16. Dhillon A, Sparkes E, Duarte RV. Mindfulness-based interventions during pregnancy: a systematic review and meta-analysis. Mindfulness (N Y) 2017;8(6):1421-1437 [FREE Full text] [doi: 10.1007/s12671-017-0726-x] [Medline: 29201244]
17. Sevilla-Llewellyn-Jones J, Santesteban-Echarri O, Pryor I, McGorry P, Alvarez-Jimenez M. Web-based mindfulness interventions for mental health treatment: systematic review and meta-analysis. JMIR Ment Health 2018;5(3):e10278 [FREE Full text] [doi: 10.2196/10278] [Medline: 30274963]
18. Lönnberg G, Jonas W, Unternaehrer E, Bränström R, Nissen E, Niemi M. Effects of a mindfulness based childbirth and parenting program on pregnant women's perceived stress and risk of perinatal depression-results from a randomized controlled trial. J Affect Disord 2020;262:133-142. [doi: 10.1016/j.jad.2019.10.048] [Medline: 31733457]
19. Vieten C, Astin J. Effects of a mindfulness-based intervention during pregnancy on prenatal stress and mood: results of a pilot study. Arch Womens Ment Health 2008;11(1):67-74. [doi: 10.1007/s00737-008-0214-3] [Medline: 18317710]
20. Pan WL, Chang CW, Chen SM, Gau ML. Assessing the effectiveness of mindfulness-based programs on mental health during pregnancy and early motherhood - a randomized control trial. BMC Pregnancy Childbirth 2019;19(1):346 [FREE Full text] [doi: 10.1186/s12884-019-2503-4] [Medline: 31601170]
21. Dayan J, Creveuil C, Herlicoviez M, Herbel C, Baranger E, Savoye C, et al. Role of anxiety and depression in the onset of spontaneous preterm labor. Am J Epidemiol 2002;155(4):293-301. [doi: 10.1093/aje/155.4.293] [Medline: 11836191]
22. Creswell JD. Mindfulness interventions. Annu Rev Psychol 2017;68:491-516. [doi: 10.1146/annurev-psych-042716-051139] [Medline: 27687118]
23. Zhang D, Tsang KWK, Duncan LG, Yip BHK, Chan DCC, Lee EKP, et al. Effects of the Mindfulness-Based Childbirth and Parenting (MBCP) program among pregnant women: a randomized controlled trial. Mindfulness 2023;14(1):50-65 [FREE Full text] [doi: 10.1007/s12671-022-02046-8]
24. Duncan LG, Shaddix C. Mindfulness-Based Childbirth and Parenting (MBCP): innovation in birth preparation to support healthy, happy families. Int J Birth Parent Educ 2015;2(2):30-33. [Medline: 29051821]
25. Van der Meulen RT, Veringa-Skiba I, Van Steensel FJA, Bögels SM, De Bruin EI. Mindfulness-Based Childbirth and Parenting for pregnant women with high fear of childbirth and their partners: outcomes of a randomized controlled trial assessing short- and longer-term effects on psychological well-being, birth and pregnancy experience. Midwifery 2023;116:103545 [FREE Full text] [doi: 10.1016/j.midw.2022.103545] [Medline: 36375411]
26. Abras C, Maloney-Krichmea D, Preece J. User-centered design. In: Bainbridge W, editor. Berkshire Encyclopedia of Human-Computer Interaction When Science Fiction Becomes Science Fact. Great Barrington, Massachusetts: Berkshire Publishing Group; 2004:757-768.
27. Dabbs ADV, Myers BA, Mc Curry KR, Dunbar-Jacob J, Hawkins RP, Begey A, et al. User-centered design and interactive health technologies for patients. Comput Inform Nurs 2009;27(3):175-183 [FREE Full text] [doi: 10.1097/NCN.0b013e31819f7c7c] [Medline: 19411947]
28. Gulliksen J, Göransson B, Boivie I, Blomkvist S, Persson J, Cajander A. Key principles for user-centred systems design. Behaviour & Information Technology 2003;22(6):397-409 [FREE Full text] [doi: 10.1080/01449290310001624329]
29. Sanders EBN, Stappers PJ. Co-creation and the new landscapes of design. CoDesign 2008;4(1):5-18 [FREE Full text] [doi: 10.1080/15710880701875068]
30. Stickdorn M, Schneider J. This is Service Design Thinking: Basics, Tools, Cases. Amsterdam: Bis Publishers; 2010.
31. Bardacke N. Mindfulness in Gravidanza. Perignano, Tuscany: Edra; 2023.
32. Duncan LG, Cohn MA, Chao MT, Cook JG, Riccobono J, Bardacke N. Benefits of preparing for childbirth with mindfulness training: a randomized controlled trial with active comparison. BMC Pregnancy Childbirth 2017;17(1):140 [FREE Full text] [doi: 10.1186/s12884-017-1319-3] [Medline: 28499376]
33. Warriner S, Crane C, Dymond M, Krusche A. An evaluation of mindfulness-based childbirth and parenting courses for pregnant women and prospective fathers/partners within the UK NHS (MBCP-4-NHS). Midwifery 2018;64:1-10. [doi: 10.1016/j.midw.2018.05.004] [Medline: 29843001]
34. Sbrilli MD, Duncan LG, Laurent HK. Effects of prenatal mindfulness-based childbirth education on child-bearers' trajectories of distress: a randomized control trial. BMC Pregnancy Childbirth 2020;20(1):623 [FREE Full text] [doi: 10.1186/s12884-020-03318-8] [Medline: 33059638]

XSL•FO
RenderX

35. Veringa-Skiba IK, de Bruin EI, van Steensel FJA, Bögels SM. Fear of childbirth, nonurgent obstetric interventions, and newborn outcomes: a randomized controlled trial comparing mindfulness-based childbirth and parenting with enhanced care as usual. Birth 2022;49(1):40-51 [FREE Full text] [doi: 10.1111/birt.12571] [Medline: 34250636]
36. Lönnberg G, Niemi M, Salomonsson B, Bränström R, Nissen E, Jonas W. Exploring the effects of mindfulness-based childbirth and parenting on infant social-emotional development. Mindfulness 2021;12(8):2009-2020 [FREE Full text] [doi: 10.1007/s12671-021-01658-w]
37. Kantrowitz-Gordon I, Abbott S, Hoehn R. Experiences of postpartum women after mindfulness childbirth classes: a qualitative study. J Midwifery Womens Health 2018;63(4):462-469. [doi: 10.1111/jmwh.12734] [Medline: 29924917]
38. Lönnberg G, Niemi M, Salomonsson B, Bränström R, Nissen E, Jonas W. Correction to: exploring the effects of mindfulness-based childbirth and parenting on infant social-emotional development. Mindfulness 2021;12(8):2021 [FREE Full text] [doi: 10.1007/s12671-021-01669-7]
39. Roubinov DS, Epel ES, Coccia M, Coleman-Phox K, Vieten C, Adler NE, et al. Long-term effects of a prenatal mindfulness intervention on depressive symptoms in a diverse sample of women. J Consult Clin Psychol 2022;90(12):942-949 [FREE Full text] [doi: 10.1037/ccp0000776] [Medline: 36441994]
40. Resources. Mindful Birthing and Parenting. URL: https://www.mindfulbirthing.org/resources [accessed 2023-11-29]
41. Self-care and mindfulness for moms. Mindful Mamas. URL: https://mindfulmamasclub.com/ [accessed 2024-10-14]
42. Mind the Bump. URL: https://www.mindthebump.org.au/ [accessed 2024-12-13]
43. van Gemert-Pijnen JEWC, Nijland N, van Limburg M, Ossebaard HC, Kelders SM, Eysenbach G, et al. A holistic framework to improve the uptake and impact of eHealth technologies. J Med Internet Res 2011;13(4):e111 [FREE Full text] [doi: 10.2196/jmir.1672] [Medline: 22155738]
44. Czajkowski SM, Powell LH, Adler N, Naar-King S, Reynolds KD, Hunter CM, et al. From ideas to efficacy: the ORBIT model for developing behavioral treatments for chronic diseases. Health Psychol 2015;34(10):971-982 [FREE Full text] [doi: 10.1037/hea0000161] [Medline: 25642841]
45. Charles EO, Suci GJ, Tannenbaum PH. The Measurement of Meaning. United States: University Of Illinois Press; 1957.
46. Adlin T, Pruitt J. The Essential Persona Lifecycle: Your Guide to Building and Using Personas. Cambridge, Massachusetts, United States: Morgan Kaufmann; 2010. [doi: 10.1016/b978-0-12-381418-0.00006-1]
47. Create charts and maps with Datawrapper. Datawrapper. URL: https://www.datawrapper.de/ [accessed 2019-01-30]
48. Rizzi S, Poggianella S, Pavesi M, Gios L, Bincoletto G, Scolari I, et al. A mindfulness-based app intervention for pregnant women: protocol for a pilot feasibility study. JMIR Res Protoc 2024;13:e53890 [FREE Full text] [doi: 10.2196/53890] [Medline: 38567964]

Abbreviations

MBCP: Mindfulness-Based Childbirth and Parenting
MBSR: Mindfulness-Based Stress Reduction

Edited by A Mavragani; submitted 26.03.24; peer-reviewed by L Paulos-Guarnieri, CME Fletcher; comments to author 05.10.24; revised version received 21.10.24; accepted 18.11.24; published 17.01.25.

Please cite as:
Rizzi S, Pavesi MC, Moser A, Paolazzi F, Marchesoni M, Poggianella S, Gadotti E, Forti S
A Mindfulness-Based App Intervention for Pregnant Women: Qualitative Evaluation of a Prototype Using Multiple Case Studies
JMIR Form Res 2025;9:e58265
URL: https://formative.jmir.org/2025/1/e58265
doi:10.2196/58265
PMID:

©Silvia Rizzi, Maria Chiara Pavesi, Alessia Moser, Francesca Paolazzi, Michele Marchesoni, Stefania Poggianella, Erik Gadotti, Stefano Forti. Originally published in JMIR Formative Research (https://formative.jmir.org), 17.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Reflections of Foster Youth Engaging in the Co-Design of Digital Mental Health Technology: Duoethnography Study

Ifunanya Ezimora[1*], BA; Tylia Lundberg[1*], BA; Dylan Miars[1], BA; Jeruel Trujeque[1]; Ashley Papias[1], AA; Margareth V Del Cid[1], PhD; Johanna B Folk[1], PhD; Marina Tolou-Shams[1], PhD

Department of Psychiatry and Behavioral Sciences, University of California, San Francisco, San Francisco, CA, United States

[*]these authors contributed equally

Corresponding Author:
Johanna B Folk, PhD
Department of Psychiatry and Behavioral Sciences
University of California, San Francisco
1001 Potrero Avenue
San Francisco, CA, 94110
United States
Phone: 1 4156029521
Email: johanna.folk@ucsf.edu

Abstract

Background: Current research on digital applications to support the mental health and well-being of foster youth is limited to theoretical applications for transition-aged foster youth and support platforms developed without intentional input from foster youth themselves. Centering the lived expertise of foster youth in digital solutions is crucial to dismantling barriers to care, leading to an increase in service access and improving mental health outcomes. Co-design centers the intended end users during the design process, creating a direct relationship between potential users and developers. This methodology holds promise for creating tools centered on foster youth, yet little is known about the co-design experience for foster youth. Understanding foster youth's experience with co-design is crucial to identifying best practices, knowledge of which is currently limited.

Objective: The aim of this paper is to reflect on the experiences of 4 foster youth involved in the co-design of FostrSpace, a mobile app designed through a collaboration among foster youth in the San Francisco Bay Area; clinicians and academics from the Juvenile Justice Behavioral Health research team at the University of California, San Francisco; and Chorus Innovations, a rapid technology development platform specializing in participatory design practices. Key recommendations for co-designing with foster youth were generated with reference to these reflections.

Methods: A duoethnography study was conducted over a 1-month period with the 4 transition-aged former foster youth co-designers of FostrSpace via written reflections and a single in-person roundtable discussion. Reflections were coded and analyzed via reflexive thematic analysis.

Results: In total, 4 main themes were identified from coding of the duoethnography reflections: power and control, resource navigation, building community and safe spaces, and identity. Themes of power and control and resource navigation highlighted the challenges FostrSpace co-designers experienced trying to access basic needs, support from caregivers, and mental health resources as foster youth and former foster youth. Discussions pertaining to building community and safe spaces highlighted the positive effect of foster youth communities on co-designers, and discussions related to identity revealed the complexities associated with understanding and embracing foster youth identity.

Conclusions: This duoethnography study highlights the importance of centering the lived expertise of co-designers throughout the app development process. As the digital health field increasingly shifts toward using co-design methods to develop digital mental health technologies for underserved youth populations, we offer recommendations for researchers seeking to ethically and effectively engage youth co-designers. Actively reflecting throughout the co-design process, finding creative ways to engage in power-sharing practices to build community, and ensuring mutual benefit among co-designers are some of the recommended core components to address when co-designing behavioral health technologies for youth.

(JMIR Form Res 2025;9:e53231) doi:10.2196/53231

XSL•FO
RenderX

KEYWORDS

foster youth; digital health technology; co-design; app development; mental health; adolescent; young adult; mobile health; mHealth; foster care; duoethnography

Introduction

Background

Over 391,000 youth are in foster care in the United States, with behavioral health being their largest unmet need [1]. Foster youth are young people removed from their caregivers due to safety risks or unforeseen circumstances (eg, death of caregivers) and placed in out-of-home care that aims to be safe and stable. Out-of-home placement, such as group homes and placement with a foster parent, can precede reunification with their original caregiver, permanent placement with a legal guardian or adoptive family, or other planned permanency arrangements [2]. Contact with the foster care system at key periods of development and abrupt placement with new caregivers can contribute to foster youth experiencing adverse behavioral health outcomes, increased risk of not having an accountable caregiver, and insufficient access to services that support their well-being [3]. Changes in placement can result in service disruption, further exacerbating unmet mental health needs. However, digital health technology may provide an intermediary behavioral health support solution while youth transition between placements or act as a longer-term support that can go with them anywhere from placement to placement. There is a need to target widespread inequities in behavioral health service access and involve foster youth voices in the development of new resources. Currently, there is a gap in the literature on the development, deployment, and evaluation of mobile apps designed for foster youth. Foster youth have been excluded from conversations regarding tools designed to support their well-being despite their lived expertise being crucial in targeting barriers that drive adverse mental health outcomes and underuse of services [3]. Co-design holds promise for creating youth-centered tools that are responsive to the needs of specific communities such as foster youth. Despite the potential benefits, little is known about how best to facilitate co-design with youth from underserved groups, such as foster youth, or their experiences working on co-design teams.

Co-Designing Mobile Apps

Interventions to address mental health and resource needs have increasingly shifted toward digital health technologies. Digital health methods hold great potential for addressing gaps in services for underserved communities, and using co-design methods may better meet the needs of these communities [4]. Co-design processes create a direct relationship between potential users and designers, where intended users from target communities actively participate in designing technologies for their community [4,5]. Co-design centers users in the design process and lies at one end of the user-centered design spectrum. Co-design processes with young people are meant to center youth involvement throughout the design process, but many technologies claiming to use co-design processes with youth use the term *co-design* to reference a continuum of user involvement ranging from minimal youth involvement at the level of consultation to more integrated involvement through mutual partnership (true co-design) [4].

Recent co-designed applications include mood-monitoring applications for youth experiencing depression; integrated digital crisis planning tools for youth; personal informatics apps; and open access, anonymous platforms with evidence-based, self-guided interventions for youth mental health [6-9]. Co-designed digital health technologies for both youth and adults have used human-centered co-design with studio methodology (ie, an iterative process of young people sketching out an idealized app and its features), adolescent-led and initiated co-design processes at a higher level in the "ladder of participation," and participatory technology development incorporating community-partnered research principles with technology platforms (eg, Chorus) [6,8,10,11].

Researchers have begun designing digital health technologies with greater intentionality in their work with young people by including them as early in the research design process as possible [5,12] and have called on others to do the same. However, there is still a dearth of mobile apps designed by young people for young people. Current data estimate that >10,000 existing mental health apps are available for download [13]. However, a 2020 systematic review identified only 30 mental health technologies that were co-designed digital mental health technologies for children and young people [4]. Another systematic review in 2020 surfaced only 10 publications specifically on mobile apps and web applications for young people, with most being focused on the dissemination of mental health information to specific populations [14]. With the increased use of co-design for the development of web-based mental health resources and the lack of literature on co-design with foster youth, it is imperative to focus efforts on pinpointing strategies that help center the voices of the intended beneficiaries in the creation of these technologies.

Apps and web-based platforms designed to be used by the foster youth community (and caregivers) exist. For example, FosterClub and iFoster are 2 web-based communities targeted toward empowering current and former foster youth through resources and opportunities [15,16]. Currently, iFoster is led by a former foster youth with team members with lived expertise working with children in care and transition-aged youth. FosterClub was founded by a caregiver of 2 foster youth to "help fill the gap she perceived her boys—and other children and youth in foster care in America—were experiencing: a lack of access to a peer support group and information to help them navigate the foster care system" [16]. However, these websites were designed and created by resource providers and professionals (some of whom were former foster youth that had long aged out of the system) and did not directly partner with current or transition-aged foster youth to address the emotional wellness needs of the foster youth community at creation. A recent study with current and former foster youth analyzed a conceptual framework for developing a mobile app to assist

XSL•FO
RenderX

youth in transitioning from foster care to adulthood; however, this study did not result in the formal creation of the app [17]. Thus, there remains a need to center the lived expertise of young adults presently in or recently aged out of the foster care system in designing and building emotional wellness tools. Particularly in this era defined by technology and increased accessibility, there is no room to exclude minoritized young people, such as foster youth, from participating in the creation of digital health technology meant to support them.

FostrSpace

FostrSpace is a mobile app designed and created through a collaboration among San Francisco Bay Area foster youth; clinicians and academics from the Juvenile Justice Behavioral Health (JJBH) research team at the University of California, San Francisco (UCSF); and Chorus Innovations, a rapid technology development platform specializing in participatory design practices. FostrSpace was created to take the stigma out of accessing mental health care and resources for young people who have been in foster care by offering access to digital psychoeducational materials, a personal resource navigator, and licensed clinicians who provide tele–mental health services. To our knowledge, FostrSpace is the only mobile app focused on improving emotional wellness resources for the foster youth community that was co-designed from inception to production to iteration with foster youth. FostrSpace is currently live and can be accessed at its website [18].

Aim of This Study

The aim of this paper is to reflect on the co-design experience of the 4 foster youth involved in the development of FostrSpace. The focus of this manuscript is not on the methodology of co-design as a practice but, rather, on the foster youth experience participating in the co-design process, including how lived experience in the foster care system informed what was cocreated and how the co-design process impacted the foster youth co-designers. These reflections are later leveraged to provide recommended practices for academics aiming to facilitate effective co-design spaces with foster youth. The preceding section briefly outlined the FostrSpace app and its features. The aim of this manuscript is to offer insights into the co-design process from the perspective of foster youth. Rather than providing an exhaustive description or outcomes of the resulting app, our emphasis remains on reflecting on the collaborative design experience. The FostrSpace co-design process reflects the critical duty of academics to collaborate with minoritized communities when developing technological resources. In general, academics without lived expertise in the foster care system can be at higher risk of recreating experiences that result in mental health and resource disparities or creating web-based spaces that exacerbate trauma experienced by foster youth. The duoethnography practice used in this study built on research by Porche et al [5] that aimed to promote more authentic partnerships between academics and minoritized communities. The ultimate goal is to dismantle barriers to resource access while nurturing mutually beneficial relationships among all participants. Through the duoethnography methodology, Porche et al [5] focused on the reflections of 4 psychologists and psychiatrists in academic medicine, 2 of whom are also authors of this paper. Porche et al [5] offer broad co-design reflections across various projects (including FostrSpace), highlighting barriers and challenges in sharing co-design responsibility with co-design partners. The paper by Porche et al [5] focuses solely on reflections from academic co-designers and does not include perspectives from community co-designers. This paper expands on the work by Porche et al [5] by examining the co-design process from the youth perspective and specifically with foster youth. The use of duoethnography in examining the FostrSpace co-design process helped reveal candid reflections on foster youth's lived expertise and involvement in co-design. These reflections underscore how positive dynamics within the co-design team led to an open and collaborative partnership.

Methods

Co-Design

The design of FostrSpace used a community-partnered participatory co-design process to ensure power sharing between co-designers. Techquity, the process of achieving health equity through strategic development and deployment of technology, guided the co-design [19]. Digital health technologies can help connect young people with care, addressing service use deficits and cost barriers. Techquity applies a critical lens to the inclusion and exclusion of individuals during the technology creation process, recognizing that excluding certain voices leads to gaps in technological advancements.

The co-design process used youth participatory action research (YPAR) principles with an antiracist framework. YPAR challenges assumptions about who is allowed to conduct research, placing young people in a central role to investigate social conditions (eg, the foster care system) that impact their lives [20]. YPAR principles underscore the significance of establishing a secure and inclusive environment for involving young people in research processes. Our commitment to safety and inclusion was manifested through practices such as ensuring that participants could make informed decisions about their engagement, highlighting voluntary participation, upholding the confidentiality of shared information, fostering active involvement in decision-making, and implementing continuous feedback and reflection mechanisms. Similarly, we were mindful of racialized research positionality and cultural and developmental contexts, deferring to the FostrSpace Advisory Board (FAB)—a name chosen by the members—co-designers for language and framing feedback whenever possible on the app. YPAR is a valuable approach to support young people wanting to make a positive impact, proving especially impactful for those facing marginalization rooted in racism, sexism, homophobia, transphobia, classism, ableism, or other oppressive factors. With this in mind, we chose YPAR as a framework to underscore the expertise of our foster youth co-designers and carefully manage power dynamics to promote power sharing. While all research projects within the JJBH research team incorporate antiracist principles, it requires a deliberate and sustained effort and, thus, was an intentional focus in the FostrSpace co-design. Particular efforts were made to value multiple ways of knowing and honor foster youth similar and

XSL•FO
RenderX

dissimilar lived expertise, as well as counter paternalism to ensure that participating foster youth were empowered to direct FostrSpace's creation from design to mechanics.

FostrSpace co-design included transition-aged foster youth, clinicians, academics, and technology developers in the same co-design space rather than holding separate parallel co-design processes, as is common in app development. By having all co-designers in the room, everyone could learn from one another and collaborate to integrate distinct experiences into the FostrSpace design process. Youth co-designers came with lived expertise and histories of foster youth advocacy work, whereas academic co-designers came with experience of co-designing with young people in other research projects [5]. Clinicians, some of whom held dual roles as academics, possessed clinical expertise in supporting youth well-being and facilitating group processes. Technology developers came with expertise creating digital health technologies through participatory design.

Co-design of FostrSpace started in the planning stages, with young people aged 16 to 26 years who were previously or currently involved in the foster care system being invited to participate in idea-sharing work groups. The work groups were advertised through organizations and resource providers that have ongoing collaborations with the JJBH research team (eg, the San Francisco Unified School District), who shared recruitment flyers within their networks. In total, 2 work groups took place in September 2021, qualitatively exploring how a mobile app could help support foster youth in identifying and navigating mental health and wellness resources, supports, and interventions. The first work group had 3 participants, and the second work group had 2 participants; each work group lasted 90 minutes and was held with a different group of foster youth. A call was put out to the 5 participants over the 2 work groups to join the co-design team in a more formal staff capacity. Once the team was finalized (ie, the FAB members were hired), formal co-design meetings for FostrSpace took place on a weekly basis.

Duoethnography

Overview

Duoethnography, as defined by Sawyer and Norris [21], was used to guide the analysis. This research method aims to look at "the cultural contexts of autobiographical experiences in order to gain insight into their current perspectives and experiences of issues related to personal and professional identities" [22], allowing researchers to use their "life curriculum" to collaboratively examine and apply meaning to the experience of interest. In this context, the duoethnography process focuses on the experience of 4 self-identified foster youth (the FAB members) who codeveloped the mental wellness app FostrSpace with a team of academics, clinicians, and technology developers. Duoethnography was regarded as the most appropriate approach for this study as it allowed for centering the voices of the FAB. Furthermore, the duoethnography reflection process promoted further collaboration between the FAB and the academic co-design team. The FAB's sharing of their lived experiences that motivated involvement in FostrSpace highlighted the importance of having young people with lived expertise contribute to interventions aimed at their community, as well as the practices that were considered to have made the co-design of FostrSpace a productive, reflective space.

Participants

Coauthors of this paper include 4 FAB co-designers (TL, DM, JT, and AP), with whom the duoethnography was performed, and 4 academic co-designers (IE, JBF, MVDC, and MT-S), who did not provide reflections for the duoethnography. As part of the duoethnography process, FAB co-designers provided both written reflections and verbal responses via a roundtable discussion regarding their experiences co-designing and cobuilding FostrSpace. All FAB members also contributed to the writing of the duoethnography in a multitude of ways (eg, co–first author and author) through providing data to shape the thematic analysis (TA), being involved in TA (eg, coding and writing), or providing feedback on the themes derived (eg, whether quotes were used appropriately or themes were derived appropriately). The collaborative academic co-designers and authors (n=4) were involved in providing structure to the co-design of the FostrSpace app, the reflection prompts, and the roundtable discussion. All academic co-designers were involved in the writing of this manuscript, in TA (eg, coding and writing), or in providing feedback on the themes derived. While the sample size of 4 FAB members contributing to the data collected for the TA is small, it reflects the number of young people with lived expertise in foster care who contributed to the development of FostrSpace and their reflections on the co-design process. Duoethnography does not have a set number of collaborators needed to engage in the methodology, rather focusing on the ability to maintain "intimacy, trust, and commitment" [22] to the tenets of duoethnography (eg, self-reflexivity) among 2 or more researchers. Similarly, reflexive TA dictates that the participant sample size for analysis depends on the size and scope of the project [23]. The FAB members' reflections provided a basis for engaging with foster youth in collaborative development of digital solutions and a framework for centering lived expertise in the development of and research on mental health technology.

The FAB included a White nonbinary foster youth, a White-passing mixed Asian-Hispanic foster youth, a Mexican-Latino foster youth, and a Chicana foster youth. The FAB members are a group of current and recently graduated multilingual, first-generation college students who, together, have experiences in both the traditional foster care system and kinship or relative care, making them intimately aware of the disparities present in the child welfare system. At the time of the duoethnography, 75% (3/4) of the FAB members identified as former foster youth, and 25% (1/4) identified as transition-aged foster youth; they maintained a consensus that they were not current foster youth due to being aged over 18 years but were still impacted by the identity. Through their upbringings, FAB members experienced poverty, low-income, working-class, and middle-class statuses, living in group homes, kinship caregiver households, community housing centers, campus housing, and individual housing. As a group, they hold both cisgender and queer identities, various levels of connection and separation from their cultural identities, different levels of disability status, and part-time positions in both FostrSpace and

XSL•FO
RenderX

other jobs in the community while maintaining statuses as students or full-time employees.

The academic co-designers included a first-generation Nigerian Black pangender and pansexual research coordinator; a first-generation Guatemalan American, neurodivergent, cisgender, and straight light-skinned Latina clinical psychologist; a White Jewish genderqueer and gay clinical psychologist; and a White-passing Iranian immigrant cisgender and heterosexual clinical psychologist. All are conducting community-engaged research to increase behavioral health service access for youth and families in contact with the juvenile legal and child welfare systems and have had personal ties to these communities.

Through connected but distinct lived expertise and various roles that we occupy in community and academic spheres, all the coauthors of the paper have an awareness of the types of hardship that foster youth can experience from childhood to adulthood and how those hardships interact with intersectional identities. These hardships included social and emotional factors such as parentification from a young age as well as structural and systemic barriers to resource navigation in the behavioral health, medical, and educational spheres. Our own identities and lived expertise drive us to address these barriers to improve conditions for foster youth and provide a lens for interpreting our results.

Procedure

Following duoethnography guidelines, FAB members reflected on their experiences in the foster care system and how this related to their experience in the FostrSpace app development co-design process (eg, what they personally got out of the process and how they conceptualized different features on the app) via written reflections and verbal discussions. Members spent 1 month (April 2022-May 2022) writing their own reflections and reading and responding to their fellow FAB members' reflections. Following the writing period, an in-person roundtable discussion about the co-design experience was held to allow members to verbally share their reflections as well (eg, because not all co-designers felt as comfortable with written reflections). The 4 FAB members engaged in an hour-long discussion with JBF as a facilitator and IE as notetaker. The facilitator and notetaker were academic co-designers that contributed to the FostrSpace app development and duoethnography paper writing; neither held a formal supervisory role over the FAB members (ie, the supervisor on record was not involved in the duoethnography process). We acknowledge that the power dynamics of having academic co-designers facilitate the roundtable discussion among the FAB members about their co-design experience could have impacted what was and was not shared. Keeping this in mind, academic co-designers provided structure to the discussion through the prompting of discussion questions while allowing space for the co-designers to talk with one another freely on the topics with limited interjection. In addition, FAB members and academic co-designers built mutual, trusting partnerships with one another during the co-design process, and the duoethnography took place after the bulk of the co-design process was concluded and the app was functional. As a result of the mitigation of the potential consequences of negative power dynamics through these methods, FAB members felt comfortable sharing sensitive data in the roundtable, as indicated by the vulnerability and honesty exhibited in the quotes used in the results as well as within co-design meeting discussions and one-on-one check-ins with staff. Reflection prompts were collaboratively developed by the coauthors before starting the duoethnography process, keeping in mind the research aim of understanding the experience of foster youth co-designing an app in collaboration with academics, clinicians, and technology developers. The same prompts were used for the written and verbal reflections. The prompts are available in Textbox 1.

Textbox 1. Reflection prompts used in the FostrSpace duoethnography with FostrSpace Advisory Board co-designers over the 1-month reflection period.

- Why did you get involved in this process? What encouraged you to keep going? After attending the first meeting, what kept you going to the second, third, and so on?
- How do you feel about the way the co-design was handled? In what ways would you have liked the process to go differently? What were your experiences of power dynamics in the co-design process and in what you bring to this work?
- Why is it important to center lived experience in this work? What parts of yourself and your experiences in foster care did you bring into this work? What inspired you throughout your time in foster care to have interest in what we are doing? What support did you lack in foster care? Remember: not every story is the same—not everyone would do the same things growing up and everyone's feelings are valid.
- How have you changed throughout the course of the co-design process? How have your goals or aspirations or mindset changed since starting this project?
- Has this project amplified your voice or feelings toward reform within the foster care system? In what ways?
- Why do you think this is a good resource for other foster youth to have?

Data Analysis

Reflexive TA [23-27] was used with 9 pieces (n=8, 89% written reflections and peer responses to reflections and n=1, 11% roundtable transcription) from the 4 FAB members. Reflexive TA was specifically chosen (as opposed to a coding reliability TA or a codebook approach to TA) to be congruent with the co-design value of honoring participant lived expertise. Reflexive TA allows for the analysis of participant subjective data through the lens of researcher reflexive interpretations [26]. An inductive, "bottom up" approach under an essentialist and realist paradigm was used by the first author coders (IE and TL) to ensure that identified patterns within the data were pulled from the data itself [23-27]. Initial codes aimed to provide a rich description of the full dataset to capture the breadth of

XSL•FO
RenderX

factors that should be at the forefront of academics' minds when co-designing with individuals with lived expertise in the foster care system. Themes were developed using a semantic approach in which surface meanings of the data were focused on in the initial coding process. This distinction was important for the coders as a way to acknowledge the prevalent history in the foster youth community of feeling disempowered by individuals and groups in power. By using a semantic approach to theme development, coders hoped that the process would be more empowering and respectful to the FAB members' experiences. A realist and essentialist paradigm rather than a constructionist approach was used to focus on individual psychologies as well as differences and similarities in lived expertise among the FAB members.

In line with the 6 guidelines by Braun and Clarke [23-27] for reflexive TA, the coders first familiarized themselves with the dataset before initiating coding. As an FAB member and participant in the verbal discussion and written prompts and reflections, TL was intimately familiar with the dataset. Similarly, as lead research coordinator on FostrSpace, IE read the FAB reflections as they were being written and acted as the notetaker during the roundtable discussion. All qualitative analysis was conducted in ATLAS.ti (ATLAS.ti Scientific Software Development GmbH) [28]. The 2 coders independently generated initial codes related to the co-design process and the foster youth experience. Preliminary codes were discussed collaboratively among the coders and coauthors (IE, TL, JBF, and MT-S) to develop a final list of codes. A total of 18 codes (initial broad themes) were developed and used to conduct final coding through which themes were developed (Textbox 2). Analysis involved more than one coder and, therefore, necessitated collaborative discussion.

Textbox 2. The 18 codes developed from thematic analysis and applied to the 4 FostrSpace Advisory Board members' reflections on the co-design process.

Codes

- App as a place for transparency
- App as a safe space
- Barriers to accessing resources that are not foster youth specific
- Barriers to accessing resources specifically for foster youth
- Co-design as a safe collaborative space
- Future goals for the app
- Identity (negative)—external gatekeeping
- Identity (negative)—internal gatekeeping
- Identity (positive)—finding, accepting, and embracing
- Initial app involvement
- Intersectionality
- Lack of clarity—foster care system
- Motivation for the app—paying it forward
- Negative experiences in kinship care
- Positive community support
- Positive impact of co-design participation on the self
- Resource stigma
- Stigma toward foster youth

The results include final analyses of the selected extracts and relevant themes, which were reviewed and approved by the coauthors of this paper, including the entire FAB team, who had the option of retracting extracts of their data that they did not feel comfortable sharing with the public. All FAB members endorsed the information presented in the manuscript and opted not to retract data.

Ethical Considerations

The institutional review board of the UCSF determined that this research is exempt from institutional review board oversight according to the federal regulations summarized in Title 45 of the Code of Federal Regulations Part 46.102(l). Participants from the work groups were compensated with a US $75 gift card for their participation. A total of 80% (4/5) of the individuals from the work groups applied for and joined the final co-design team, making up the FAB. FAB members were then hired as UCSF employees to co-design the FostrSpace app and were compensated at an hourly rate based on their individual compensation package, which was determined by educational level and experience as set by UCSF Human Resources for all the time spent completing tasks related to UCSF employment (eg, co-designing the FostrSpace app and engaging in the duoethnography reflections). The FAB as employees of UCSF must abide by federal and state employment laws as well as have access to human resources to support any workplace issues.

XSL•FO
RenderX

Results

Overview

Coders derived 4 main themes from organizing codes around core commonalities through reflexive TA. The final themes addressed the experience of taking part in the co-design process, including the FAB members' broad experiences in the foster care system, which directly influenced their approach to and navigation through the co-design process and app development. The number of quotes presented in the results from each FAB member ranged from 5 to 15 excerpts. Themes included (1) power and control (loss of power, reclamation of power, and power dynamics), (2) resource navigation, (3) building community and safe spaces, and (4) identity. All themes were rooted in explicit content from the coded data.

Power and Control

Loss of Power

FAB members described their experiences with a loss of power due to the nature of their foster youth status through childhood, adolescence, and even extending into young adulthood. Systems (ie, the child welfare system) allegedly designed to empower and protect them withheld crucial information necessary for FAB members to feel like they had agency over their own lives. To illustrate, the lack of transparency regarding monthly payments made to foster parents (both the monetary amount and transaction timeline) led to frustration and disempowerment stemming from FAB members feeling that they were left in the dark about their own care:

> *So, for a while, I was just like, "I'm homeless." Or I'm not homeless, "I'm poor." I was just like, "I'm poor. I can't ask for anything." And then 10-plus years later to find out that that person was still getting paid to take care of me? And I'm like—because they're family, you know what I mean, it's [sic] was like they didn't really owe you much?* [FAB member 3]

Existing foster care systems and policies (eg, foster youth not having a say in placement with family members) constantly take away power and control from foster youth, making decisions for them without their input. Decisions made by guardians without FAB members' input negatively impacted the ability of FAB members to access vital information necessary for their care:

> *I asked my grandparents, "Who is my social worker?" They were like, "Oh, we shredded all those documents because it didn't apply because they weren't coming anymore. And we were good." I'm like, "What?" So, I tried calling the county, they never got back to me. They just put me in this whole rigamarole.* [FAB member 2]

In addition, though part of a larger network of youth in the child welfare system, FAB members noted the isolation prevalent in the foster care system:

> *Growing up foster kids were never given a platform to speak and ask for help. I personally was never given resources growing up that could help connect me to other foster youth or provide mental health resources when I needed them.* [FAB member 1]

Reclamation of Power

By participating in the co-design of FostrSpace, FAB members were able to reclaim some of the power they had lost and their decision-making ability. Members felt empowered to reclaim control over their own trajectories as well as lay the groundwork for foster youth coming after them to assert agency. Through FostrSpace, FAB members felt that they could equip current and future generations of foster youth with information and resources, which could allow them to take control of their own lives and mental health:

> *I agree that the app is a great place for current and past foster youth to find a supportive community to share with as it is not always easy to reach out to the people around you for help. Especially when you're searching for a safe space to share struggles with mental health as so much stigma still surrounds mental illness in the United States as well as being a former or current foster youth.* [FAB member 4]

In addition to promoting resource access, co-designing FostrSpace strengthened FAB members' feelings of power and control by educating youth about the foster care system. Growing up in the isolation of the foster care system led to increased appreciation for the transparency available in FostrSpace:

> *[My caregiver] villainized everybody and it was like we cut contact off with everyone. It was really like isolating. So, I like that we have this resource where we can go search it up and be like, "Okay, well, what does it mean to be a foster kid? Where can I find a resource to go get some money for food?"* [FAB member 1]

FAB members emphasized the importance of giving users ample information to equip them with knowledge, encouraging them to be informed consumers. FAB members appreciated how FostrSpace leans toward having too much rather than too little information:

> *That's the thing—like the app. I remember I always like where we specify everything down to a T. Sometimes it gets a little wordy and that's annoying for some users of the app, but it explains everything to you clearly.* [FAB member 1]

With the opportunity to use their lived expertise in the development of the web-based platform, FAB members felt that they could reclaim the narrative over their life. FostrSpace gave them the opportunity to regain control after years of experiencing intentional withholding of control by the system:

> *It really points to just this greater idea of mitigating suffering. This greater idea of moving beyond and I think that's really important. I've always been focused on this as a greater abstract idea, but this is one of the ways I've actually nailed down providing support and making the world better, in a limited way that I*

XSL•FO
RenderX

can. And so, I've really appreciated having the space to allow me to actually do this. [FAB member 2]

Power Dynamics

The interplay of power dynamics was key for FAB members regarding life experiences as detrimental versus supportive. Harmful power dynamics manifested in both navigating governmental assistance and the family unit. While navigating already mystifying systems, FAB members had to advocate for themselves to individuals in positions of power when submitting applications for food assistance, housing, and more. FAB members also expressed feeling as if their own realities and experiences were minimized or denied when trying to access resources:

> *...I had to call back for my EBT [electronic benefits transfer] because they denied it last month. I remember calling them and I was like, "Hey, you guys denied it. Even the app says I didn't have any money entered." And she's checking her records and she's like, "No. We put it in." And I was like, "Lady, like no you didn't."...I remember just being like, Come on. I'm already telling you I had to use money out of my own pocket that I've been trying to save for rent. I know what I'm using my money for.* [FAB member 1]

Government and school staff alike also denied FAB members' foster youth identities and ability to access foster youth–specific resources because they did not have proper documentation labeling them as foster youth:

> *In school, I remember I was constantly denied. I asked if they could get help finding foster youth resources and scholarships, and I literally had staff tell me, "You're not a foster kid. I don't have documentation."* [FAB member 1]

Consistently facing roadblocks and antagonism from the systems designed to help them led to feelings of disempowerment among the FAB members. Having to go up the chain of command to receive a response or going through several layers of forms, applications, and paperwork only for these to be rejected accumulated experiences of trauma and negatively impacted the motivation to access these resources for some FAB members:

> *I am and have been eligible for EBT [electronic benefits transfer], but after doing it for a year I don't do it...how many of these are actually foster youth-friendly? Because we can tell people to sign up for SNAP, but when we look at these programs the blatant villainization or willful ignorance to what we've gone through and the struggles we are going through makes it so it's harder to reach out for these resources...they don't take me seriously when I tell them about the trauma that comes from it.* [FAB member 2]

FAB members experienced kinship care as mirroring the unsupportive power dynamics in governmental assistance, with kinship care "family" behaving differently from how a "traditional" family or chosen family or community may be thought to behave. FAB members described feeling alienated due to the dynamics of being in a kinship care household:

> *...even though foster youth may have access to family in foster care through family members being their foster parent or having their sibling with them, it's still easy to feel alone. I personally felt that as the oldest sibling of three I couldn't be vulnerable when I was struggling as everyone else was also having a difficult time. In my experience being in foster care made it difficult for me to ask for support, I was uncomfortable being in a new parental dynamic with people I didn't know too well so it was hard to ask for things.* [FAB member 4]

FAB members experienced kinship caregivers withholding resources and information about the foster care system that FAB members needed even though they were meant to be trusted adults who could provide care:

> *She just wanted to do stuff where again she was the only one in control. And with me going and finding these resources? She didn't help me go do that. She was leaving me on my own and like, "Go do that." Again, it's just like she's leaving me out there and alone. It was really sad.* [FAB member 1]

Similarly, FAB members felt the weight of being the different party in a new caregiver-youth dynamic with mixed signals:

> *But looking back on [growing up], it was a heavy load kind of me just to like—well, my family is proud of me? No, they're not proud of me because I'm smarter than I guess they thought I'd be. So, for a while, it was just like wow. I was just so conflicted where it was like, "Am I doing the right thing? Should I keep pursuing my education?"...I'm like..."Why are you talking down on me and making me feel bad?...Isn't this what y'all wanted for me?"* [FAB member 3]

The ability of the co-design process to empower FAB members was in great part due to the intentional way in which power dynamics were attended to in the process:

> *I remember telling everybody like, "Dude, it's equal here...they don't talk on top of each other. They don't cut you off. They let you finish your idea even if it's dumb because it might be." [laughs]...here everyone is on an equal playing field, even your bosses. They're all chill with you...they're always, they're engaged. They want to listen, they want to do something, and they want to help the community. I think that's the most important part is the want to do something and then actually doing it.* [FAB member 1]

Notably, as individuals who had experienced systems designed to uplift them repeatedly ignoring their voices, FAB members went into FostrSpace without hopeful expectations:

> *I didn't necessarily anticipate having a big role in the development of this app, but I also didn't have many expectations about its potential either. I became ecstatic when I received my official offer letter to join*

XSL•FO
RenderX

> *UCSF and begin the process of developing an app as an ambassador.* [FAB member 3]

Over time in the co-design space, FAB members became comfortable, learning that the participating academics were working for the betterment of the community and uplifting individual voices within the group:

> *...the co-design team was wonderful to work with as it felt all the foster youth giving feedback were heard and treated as equals by superiors. Overall, the friendly space that was created by the entire team during meetings and the positive influence the finished app would have in the foster youth community keep me returning to the meetings.* [FAB member 4]

Resource Navigation

FAB members shared situational factors that contributed to inaccessibility of resources, including lack of knowledge of available resources and how to access them, responsive resources, adult figures to guide them toward resources, and documentation. FAB members expressed frustration over being stuck in a system with ambiguous directions:

> *Yes. Oh, my god. [collective agreement] All the [redacted name] county forms are like, "Worker ID this, ID this." Then you're like, okay, who do I call? Leaving voicemails for the workers because they never want to answer their phones. Then they never call you back. It's just stuff like that where you're just sitting there like, "Okay, well, what do I do?"* [FAB member 1]

Furthermore, FAB members mentioned that, when they did access resources, these were often not responsive to their lived expertise and status as foster youth, instead following a one-size-fits-all structure. FAB members recounted situations in which resources did not meet their needs due to limitations in access (eg, time frame):

> *I remember telling my whole life story at five years old to this lady and just telling her, "Oh, this is what happened," again the gory situation and everything, and she helped me cope with it. It was really good. I would walk out feeling refreshed. I felt like everything was just off my shoulders. Then after maybe like a few sessions it just stopped. And then I never got any resources.* [FAB member 1]

A lack of reliable adult figures during key periods of development was a common experience for FAB members, who suddenly lost support people who could direct them to needed resources. FAB members expressed that FostrSpace resonated with them as it attempts to connect foster youth with resources for mental health and basic needs:

> *I had received an email about outreach being done through UCSF to develop an app to better support the needs of foster youth, particular in the field of mental health and connecting youth to basic needs. I was previously in foster care in my early teens and lost both my parents in my teens, these experiences made me feel like I had no adult guidance to show me what resources I had available to me as well as support me when I was dealing with mental health issues.* [FAB member 4]

When resources required FAB members to provide documentation (eg, case IDs, permanent address, or social worker's name) that they did not have access to, they were hindered in submitting applications:

> *...they were doing scholarships or support funds for foster youth, and even like transition age. I remember I applied to it and they were like, "What's your social worker's name? What's your case number?" All this stuff. And I didn't have that information.* [FAB member 2]

FAB members also had difficulty identifying resources that they were entitled to as foster youth, such as Court Appointed Special Advocate workers:

> *Like, I'm going to be honest. I don't think I ever had a CASA person. So, just like you guys talking about CASA, it's like there's so many different aspects of just being a foster youth where we all have different experiences.* [FAB member 3]

The resource navigation challenges experienced by FAB members were described in contrast to the FostrSpace approach, which could benefit foster youth experiencing the same deterrents to resource access. The web-based platform provides foster youth with a directory of resources and live help on emotional wellness or resource navigation without the need for documentation. Due to FostrSpace's more straightforward approach to resource navigation, FAB members expressed that FostrSpace helps reinforce self-advocacy and facilitates positive resource navigation experiences:

> *Moreover, having good experiences accessing resources should further encourage foster youth to engage with resources as I have experienced and seen people turn away from help when their experience is negative.* [FAB member 4]

Building Community and Safe Spaces

FAB members discussed the importance of centering building community and safe spaces for foster youth. Membership in an established foster youth community positively impacted FAB members:

> *But I remembered when I got to UC Berkeley, my brother immediately was like, "Oh, my god, there's this program called Berkeley Hope Scholars"... I signed up—and [Director name redacted] like a week before was just like, "Hey, this is what we have. These are the resources. We're going to help you with everything." And I immediately felt a sense of just comfortability. I felt safe there and I finally found a community where I could find other people who are like me or who experienced some of the same stuff as me.* [FAB member 1]

Similarly, the FAB community was beneficial for creating a space to collaboratively give back to the foster youth community through FostrSpace:

XSL•FO
RenderX

> *After attending the first meeting and meeting my co-workers I instantly felt as if I had found a great community of people fighting for change for a community that is often hidden in the shadows. I was excited to finally have my voice heard and to speak for my friends who experienced the system in a multitude of ways.* [FAB member 1]

There was shared comradery among FAB members as FostrSpace allowed them to unite under a common goal: to curate a safe space for heterogeneous foster youth tied by a common thread of lived expertise:

> *We're dedicated to just creating a space so that people don't have to further suffer the way we might have already, or in ways that we may not have experienced, but we don't want anyone to ever risk going through.* [FAB member 2]

The importance of considering intersectionality was highlighted when building community and safe spaces for foster youth as foster youth experiences are shaped by how multitudes of identities, global world views, and inclusion interact. Through this duoethnography process, FAB members came to the realization that future iterations of FostrSpace's resource directory need to take into account other identities that foster youth may hold (eg, race and gender) to better target nuances of resource inaccessibility:

> *...I'm always talking about my multitude of identities and intersecting identities. But that's something where I don't want resources just to default onto foster youth because sometimes they're for White foster youth, Asian foster youth, or Black foster youth. It's like specific. So, I'd really like to find more resources for trans youth, and for other races, as well, like Black, White, Asian, Latino—because there's a huge Latino population in California. But just stuff like that where it's resources for women's reproductive health because again, I'm a woman and that's something I have to go through...Like just where it covers those identities and takes care of kids that not only align under foster youth, but are foster youth AND, you know foster youth-included.* [FAB member 1]

FAB members also felt fortunate to have found a foster youth community as well as a common desire to develop a community within FostrSpace, expressing that FostrSpace could help connect more foster youth with a supportive community:

> *But then when I actually saw what y'all were doing, I was really moved by it. Like providing mental health and wellness resources for like—a I just felt like it spoke to an earlier me that was in desperate need for those and needed support and a community aspect.* [FAB member 2]

Identity

FAB members engaged in rich discussions regarding the decision to identify as foster youth, the journey of coming to that decision, and the resulting consequences. Discussions highlighted that carrying a foster youth identity led to grappling with both positive and negative experiences. Within the context of negative experiences regarding identity, FAB members disclosed the strain that shame and stigma have on a foster youth's ability to embrace their identity. Shame was described as something that was both ascribed by others (eg, strangers and family members) and internalized. FAB members associated the identity with traumatic upbringing:

> *I was never told that I was even a foster kid. When I said—like when me and my brothers said that name, my aunt would roll her eyes and be like, "What's that? You're just making stuff up. You're dramatic." So, it was like things like that. It had a very negative connotation around it.* [FAB member 1]

While kinship care is recognized as a part of the foster care system, FAB members noted internal shame regarding placement in kinship care:

> *I think part of it is because I still feel a bit of shame about my identity where there's a part of me that's just like, "Yeah, but you were raised by your grandparents. And even though you were in the system, do you really count?"* [FAB member 2]

In addition, they experienced external shame when actively aligning with the foster youth identity:

> *...when I created my website, I was really open about it, I've written a lot about it. Then one day, one of my distant uncles calls me. I'll never forget this. I was getting ready and he was just like, "I saw that you posted an article on your Facebook about saying that you were a foster youth. You're not a foster youth." He was like, "How do you think your grandparents feel if you told them that you identify this way or that you say this?"* [FAB member 2]

Stigma was evident in the mischaracterization of the foster youth identity and pitying by others:

> *...when we were handing out flyers, one of my friends was like, "Watch out, they're going to say sorry a lot." I was like, "What do you mean?" And we were handing out flyers and the students were like, "Oh, I'm so sorry." They'll look at you with puppy-dog eyes and you're just like, "I'm good. I'm chilling, bro. I just got some free lemonade." [group laughter] It's just stuff where you feel the name "foster youth"—just you feel like has this weight on it. And that's what people add to it. It's just like not like that.* [FAB member 1]

While negative associations with identity stemmed from general community (eg, schoolmates), family (eg, kinship caregivers), and internalized anti–foster youth dialogue from living with kinship caregivers, positive associations with their foster youth identity were built through engagement in the FostrSpace co-design process and with a foster youth community. Being a part of the co-design process offered FAB members space to reflect on negative experiences related to identity as well as build a new, positive association with the foster youth identity (eg, through community building in meetings):

> *...this process has...given [me] an ability to understand and process the uniqueness of what I have*

XSL•FO
RenderX

> *gone through, but also the commonality of the struggles and how no matter who you are, we all share these ideas, these feelings. And so why not try together to overcome them?* [FAB member 2]

Moreover, experience in an established foster youth community encouraged FAB members to embrace and accept identity, as well as providing a venue to receive support regarding struggles of identity. Co-designers expressed that being members of the FAB and working on the development of FostrSpace together validated their identity and allowed them to recognize that they were not isolated in their identity or experiences:

> *So, talking with you guys has always been great because every meeting is like, "Damn, I didn't know that," or, "Yes, I've been there." Not that I'm glad that someone else has went through that with me, but it's just like knowing that now collectively we can put that insight and experience to fruition.* [FAB member 3]

FostrSpace also acted as a platform to support the dismantling of stigma regarding the foster youth identity as FAB members were given the space to call on their lived expertise in the development of FostrSpace:

> *I felt as if I had just gained a new purpose in life, which was to share my lived experience and encourage others in the foster care system to embrace their stories and share it with pride.* [FAB member 3]

Discussion

Principal Findings

Co-design processes hold significant potential to create meaningful interventions for underserved populations such as current and former foster youth. The co-design process for FostrSpace exemplifies the value of centering the lived expertise of foster youth in the creation of digital tools designed to support them. The FAB members actively brought their own lived experiences navigating resources in their personal lives into the co-design space and used those experiences in developing the resource library for FostrSpace. As active co-designers, FAB members identified potential barriers to accessing resources that foster youth may face as well as brainstorming and problem-solving ways to inform the FostrSpace approach to resource navigation. For example, one barrier to accessing resources that FAB members identified was applications requiring details that foster youth may not have access to (ie, their social worker's ID, particularly because some foster youth are not alerted to the fact that they have a social worker). Due to this, the FAB reflected as part of the co-design that FostrSpace needed to commit to providing resources made with foster youth in mind. Ideally, foster youth–tailored resources would not require information that foster youth do not have access to. To address this, the FAB recommended providing brief highlights on the information required to access certain resources (ie, CalFresh). This would equip foster youth with knowledge of potential barriers before trying to access the application. In line with this, during co-design, the FAB advocated for ensuring that foster youth did not face similar barriers and restrictions to accessing FostrSpace to those they faced in accessing other resources meant for foster youth (eg, as such, FostrSpace does not require foster youth to upload proof of status as a foster youth or social worker ID numbers).

Furthermore, reflections from the FAB members indicated that the co-design process was a safe and constructive space where all members worked together for the betterment of the wider foster youth community. As FAB member 3 stated, "Together, we can change the system and it all starts with collective effort of informing the broader community," showing the drive to collectively work toward the shared goal of creating FostrSpace. This duoethnographic process adds to the literature of co-designing with young people and specifically to the scholarship of co-designing formal mental health applications with foster youth.

Recommended Practices

Overview

The 4 main themes shared across the 4 FAB members demonstrate their reflections on their experiences with the FostrSpace co-design process, including motivations for participation and potential reasons why FAB members could not access the very resources they aimed to make more accessible through FostrSpace. From this reflection process, we recommend 3 key practices when co-designing with young people. Although we did not explicitly seek direct feedback from the FAB on these practices as part of the duoethnography process, the richness of the duoethnography data affirms the value of these practices in creating an enriching co-design space. In addition, while a number of different practices were brought into the FostrSpace co-design space, the key practices that we chose to list here are inspired by what was highlighted in the FAB members' duoethnography quotes in addition to their endorsements in co-design meetings and check-ins. These recommendations are built from practices that were implemented in the creation of FostrSpace and are believed to have made the co-design process successful and worthwhile for co-designers:

1. Engage in active reflection on the co-design process both during and after co-design.
2. Find creative ways to engage in power-sharing practices to build community across diverse experiences and positions of power and privilege.
3. Ensure mutual benefit among co-designers.

Engage in Active Reflection of the Co-Design Process Both During and After Co-Design

Engaging in active reflection throughout the co-design process allows for the intervention (in this case, FostrSpace, a digital health technology) to become more actively shaped by thoughtful reflections. In the co-design of FostrSpace, brief icebreakers were used at the start of every meeting to ease individuals into lighthearted discussion before talking about a specific feature of the app. In addition, brief check-in questions were used at the end of meetings as well as in occasional one-on-one check-ins between FAB members and UCSF academic staff to evaluate co-designer comfort and feedback on the app. Although these engagement and reflection methods were highly useful in building community among the

XSL•FO
RenderX

co-designers and allowing for free discussion, many of the reflections uncovered through the duoethnography were highly distinct and had not surfaced through previous reflection tools used during the co-design process. Specifically, by having a space designed solely for FAB co-designers to talk with one another and respond to questions not directly related to building the app through the in-person duoethnography process, we were able to reaffirm the most important features of the current app and important aspects to consider for iterations moving forward (eg, fully detailing a community element and the different ways in which this was crucial for this community).

Notably, in FostrSpace, we aimed to provide a robust resource directory for foster youth with food insecurity, legal or housing concerns, employment goals, or transportation goals, among others. We felt confident that we had developed strong resource lists within the counties that we initially served (eg, by tailoring resource pages with resource eligibility criteria, relevant website links, summaries, and office hours); however, we found through the duoethnography process that we were missing an essential element. The resources, though helpful when successfully acquired, often contained application or eligibility processes that were not foster youth friendly. This discovery, made possible only due to FAB members discussing their own personal frustrations with accessing public aid (eg, being given the "run-around," having their needs denied, or being shamed for their needs) throughout the duoethnography, highlighted the need to seek out resource providers knowledgeable of the foster youth experience. It also reinforced the need to continue pushing the personal navigator to the forefront of the app as someone "who takes the role of goal-collaborator and accountabili-buddy by providing connections to resources so folks don't have to navigate life alone" and ensuring that, for resources that are not foster youth friendly, app users have someone to support them throughout the process. Furthermore, FAB members' experiences navigating public resources highlighted the importance of self-advocacy when dealing with resource providers lacking knowledge about the foster youth experience; personal and peer navigators (FAB members) can connect with users to help build self-advocacy to support resource access. FAB members as peer navigators and continued contributors to the app have since made content on sharing knowledge with current and former foster youth about skills they learned transitioning into adulthood, essentially creating a "how to" guide to accessing some common resources (eg, electronic benefits transfer).

Find Creative Ways to Engage in Power-Sharing Practices to Build Community

As noted previously, co-design meetings began with icebreakers, some led by UCSF academics, some led by Chorus designers and developers, and some led by the FAB members. Co-designers were encouraged to think of icebreakers in between meetings, topics they were curious about in their coworkers. Many came to the meetings excited to share their icebreakers, as FAB member 1 shared in the duoethnography roundtable:

> *I think the intro questions are my favorite part. I think that really showed it, too. They were like, "Hey, what's your favorite taco spot?" It was just like small questions like that where you were like, Oh, okay, this is a cool space. Every time before each time, I'm like, Okay, what cool icebreaker questions will I ask my coworkers?*

These icebreakers aided the development of a relaxed space before diving into the FostrSpace app design and development. All members had a chance to contribute and share something before transitioning into potentially daunting topics regarding designing an app. Furthermore, engaging in conversations about non–work-related topics served to shift the focus away from work titles and hierarchies, allowing participants to connect on a more personal level beyond the confines of the work environment. In addition, as co-design meetings were often in the early evening after a potentially draining work or school day for co-designers, the icebreakers helped in providing a check-in for co-designer energy levels, which influenced how many breaks were taken in a meeting and what agenda items were prioritized. FostrSpace was an app built through a co-design process that involved internal meetings among UCSF academics and clinicians; collaborative meetings with UCSF and Chorus designers and technology staff; and co-design meetings with UCSF, Chorus, and the FAB. Many ideas went through various iterations, with the FAB voice being taken into the highest consideration. FAB members provided different viewpoints and made decisions on design elements such as name and color scheme, as well as content elements such as specific questions and text language. Furthermore, having a technology partner with extensive experience in co-designing digital health technologies with diverse communities and ensuring equitable power sharing in the development of FostrSpace contributed a great deal to the success of the co-design space.

Ensure Mutual Benefit Among Co-Designers

The maintenance of reciprocal relationships during the co-design process is ideal, with all participants (academics, foster youth, clinicians, and developers) benefiting, whether through greater understanding of the self, community building, professional development, or furthering research. This relationship can allow for continued engagement in co-design and progression of the product, elucidation of honest and often sensitive or personal feedback, and novelty and creativity in the design process. Although FAB members held their own legitimacy as individuals with lived expertise in the foster care system, UCSF academics went through the process of creating staff positions for the FAB members, formalizing their roles and providing a professional title to add to their career trajectories and resume. FAB members also actively engaged in professional development opportunities such as presentations to community members, colleges, and donors to discuss the FostrSpace development. In hiring the FAB as staff, UCSF academics reaffirmed that FostrSpace would be a longer-term commitment and partnership. In turn, UCSF academics derived benefit from the stable consistency of FAB staff, who not only were committed to their work but also had the stability through financial support that allowed for continued engagement in the work. All co-designers were able to benefit by learning from one another (eg, UCSF academics and clinicians and FAB members learning about the vast capabilities of Chorus),

XSL•FO
RenderX

building community, and having a safe space to share. All co-designers also benefited from building an emotional wellness app for foster youth—the FAB members by developing an app tailored to their needs and the needs of their fellow community members, UCSF academics by adding to the literature and development of foster youth interventions and services, and Chorus by partnering with another community to leverage participatory design practices.

Limitations

The foster youth lived expertise is highly heterogeneous, and thus, there are limits to the generalizability found in the themes we derived. The cohort of co-design partners was diverse in ethnic origin, academic background, sexuality, and lived expertise; due to the small size of the FAB group (4 members), there are limitations to broader application as lived expertise in the foster care system can vary widely. Within the small group of FAB members, there was notable overlap in their lived expertise, especially due to the overrepresentation of students (all 4 FAB members were students in higher-education institutions, and 3/4, 75% were in the same college). The common profiles of tertiary education institutions and academic studies in the San Francisco Bay Area provide another limit to the generalizability of these findings to foster youth outside of the region. All FAB members had also been placed in kinship care, suggesting that barriers to mental health resources that are faced in other forms of foster care placement may not be adequately highlighted. Expanding the co-design process to include more diverse academic backgrounds (including youth who are not pursuing higher education), partners from regions beyond Northern California, and different forms of foster care placements appears to be a promising next step for improving intervention impacts and understanding how foster youth contribute to co-design processes.

Conclusions

Lessons learned from this duoethnographic process can inform future researchers seeking to co-design digital health technologies with foster youth, as well as young people more broadly. The FAB's determination of roles and level of participation underwent constant revision to accommodate and support FAB members as their lives developed over the course of the project. As suggested by Porche et al [5], by tailoring the nature and style of power sharing to the interests and thematic motivations of the co-design partners rather than just increasing the degree of power sharing and authority alone, a greater depth of engagement can be achieved. To this end, early and concurrent partnership with the intervention population to improve impact and ensure not only a satisfactory co-design process for the partners but also a tailored end product is essential [29]. Future research may consider the nuances of the themes derived from this duoethnography and devise methods to ensure that future co-design processes address those themes both through the actual co-design and in the developed intervention or app. Foster youth who have had extensive experiences with negative power dynamics in different spaces, for example, may benefit more from greater intentionality toward the cultivation of safety and empowerment in the co-design space than may be typically granted for other young people. Similarly, foster youth such as those who participated in this duoethnography process may have increased levels of negative experiences accessing resources and may be put off from the idea of trying to access future resources through a different platform; in this case, someone creating a resource bank may seek to provide more details on the resources being offered, tips and tricks for successful resource access, and potentially a difficulty rating or timescale indicating how much bandwidth a foster youth would have to commit in trying to access the resource.

Due to the diversity within the foster youth community, it may be difficult to develop a single plan that ensures adequate empowerment and recognition of each co-designer's lived expertise in a co-design process, thereby reinforcing the need for active reflection throughout the co-design process. Continuous reflection co-design practices can be refined over time to meet the needs of a specific co-design team. The themes and subthemes gathered from the FostrSpace co-design team's duoethnography may provide a helpful framework in starting to think about how to best serve foster youth throughout the co-design process. Co-design practices are vital for creating digital health technologies that are helpful, intuitive, and healing for their intended users. Co-design practices can draw out nuanced community knowledge and result in a thoughtful, vetted product. We call on researchers to continue the work of co-designing with young people and specifically foster youth, with a commitment to active and in-depth reflection throughout the co-design process.

Acknowledgments

FostrSpace was funded by an anonymous private donor, and the time spent on this paper was funded in part by grants K24DA046569 (principal investigator: MT-S), R61DA059860 (principal investigator: MT-S), and K23DA050798 (principal investigator: JBF). The authors would like to thank Michelle Porche, EdD, for consultation on qualitative analysis plan development; Chorus Innovations for partnership related to technology development; and the wider FostrSpace team for their contributions in developing FostrSpace.

Data Availability

The datasets generated during and analyzed during this study is not publicly available due to ethical concerns about possible reidentification with this small study sample. Study metadata (eg, reflection questions and codebook) are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Authors' Contributions

IE, TL, JBF, DM, JT, AP, and MT-S conceptualized the paper. IE and TL carried out the selection of articles, data analysis, and data interpretation. IE, TL, DM, JT, and AP contributed to writing the original draft. IE, TL, JBF, DM, JT, AP, MVDC, and MT-S were involved in writing, reviewing, and editing the manuscript. MT-S and JBF contributed equally to the senior authorship role and as such are listed as co-senior authors.

Conflicts of Interest

None declared.

References

1. National foster care month outreach toolkit. Child Welfare Information Gateway. URL: https://www.childwelfare.gov/resources/national-foster-care-month-outreach-toolkit/ [accessed 2023-06-09]
2. Title IV-E foster care. Administration for Children & Families, U.S. Department of Health & Human Services. 2012 May 17. URL: https://www.acf.hhs.gov/cb/grant-funding/title-iv-e-foster-care [accessed 2023-04-28]
3. Kerker BD, Dore MM. Mental health needs and treatment of foster youth: barriers and opportunities. Am J Orthopsychiatry 2006 Jan;76(1):138-147 [FREE Full text] [doi: 10.1037/0002-9432.76.1.138] [Medline: 16569139]
4. Bevan Jones R, Stallard P, Agha SS, Rice S, Werner-Seidler A, Stasiak K, et al. Practitioner review: co-design of digital mental health technologies with children and young people. J Child Psychol Psychiatry 2020 Aug;61(8):928-940 [FREE Full text] [doi: 10.1111/jcpp.13258] [Medline: 32572961]
5. Porche MV, Folk J, Tolou-Shams M, Fortuna LR. Researchers' perspectives on digital mental health intervention co-design with marginalized community stakeholder youth and families. Front Psychiatry 2022;13:867460 [FREE Full text] [doi: 10.3389/fpsyt.2022.867460] [Medline: 35530032]
6. Hetrick SE, Robinson J, Burge E, Blandon R, Mobilio B, Rice SM, et al. Youth codesign of a mobile phone app to facilitate self-monitoring and management of mood symptoms in young people with major depression, suicidal ideation, and self-harm. JMIR Ment Health 2018 Jan 23;5(1):e9 [FREE Full text] [doi: 10.2196/mental.9041] [Medline: 29362208]
7. Hodson E, Dadashi N, Delgado R, Chisholm C, Sgrignoli R, Swaine R. Co-design in mental health; Mellow: a self-help holistic crisis planning mobile application by youth, for youth. Des J 2019 May 31;22(sup1):1529-1542 [FREE Full text] [doi: 10.1080/14606925.2019.1594975]
8. Potapov K, Marshall P. LifeMosaic: co-design of a personal informatics tool for youth. In: Proceedings of the Interaction Design and Children Conference. 2020 Presented at: IDC '20; June 21-24, 2020; London, UK. [doi: 10.1145/3392063.3394429]
9. Shroff A, Roulston C, Fassler J, Dierschke NA, Todd JS, Ríos-Herrera Á, et al. A digital single-session intervention platform for youth mental health: cultural adaptation, evaluation, and dissemination. JMIR Ment Health 2023 Feb 14;10:e43062 [FREE Full text] [doi: 10.2196/43062] [Medline: 36787180]
10. Arevian AC, O'Hora J, Jones F, Mango J, Jones L, Williams PG, et al. Participatory technology development to enhance community resilience. Ethn Dis 2018;28(Suppl 2):493-502 [FREE Full text] [doi: 10.18865/ed.28.S2.493] [Medline: 30202203]
11. Arevian AC, O'Hora J, Rosser J, Mango JD, Miklowitz DJ, Wells KB. Patient and provider cocreation of mobile texting apps to support behavioral health: usability study. JMIR Mhealth Uhealth 2020 Jul 29;8(7):e12655 [FREE Full text] [doi: 10.2196/12655] [Medline: 32723714]
12. Palomo-Carrión R, Romay-Barrero H, Pinero-Pinto E, Romero-Galisteo RP, López-Muñoz P, Martínez-Galán I. Early intervention in unilateral cerebral palsy: let's listen to the families! What are their desires and perspectives? A preliminary family-researcher co-design study. Children (Basel) 2021 Aug 30;8(9):750 [FREE Full text] [doi: 10.3390/children8090750] [Medline: 34572182]
13. Torous J, Roberts LW. Needed innovation in digital health and smartphone applications for mental health: transparency and trust. JAMA Psychiatry 2017 May 01;74(5):437-438. [doi: 10.1001/jamapsychiatry.2017.0262] [Medline: 28384700]
14. Liverpool S, Mota CP, Sales CM, Čuš A, Carletto S, Hancheva C, et al. Engaging children and young people in digital mental health interventions: systematic review of modes of delivery, facilitators, and barriers. J Med Internet Res 2020 Jun 23;22(6):e16317 [FREE Full text] [doi: 10.2196/16317] [Medline: 32442160]
15. The national network for youth in foster care. The National Foster Care Youth & Alumni Policy Council. URL: https://www.nationalpolicycouncil.org/ [accessed 2023-04-23]
16. iFoster homepage. iFoster. URL: https://www.ifoster.org/ [accessed 2023-04-27]
17. Miller JJ, Chih MY, Washington E. Conceptualizing a mobile app for foster youth transitioning to adulthood: a mixed-method approach. J Technol Human Serv 2016 May 06;34(2):145-170. [doi: 10.1080/15228835.2015.1108260]
18. FostrSpace homepage. FostrSpace. URL: https://www.joinchorus.com/fostrspace [accessed 2025-01-01]
19. Rhee K, Dankwa-Mullan I, Brennan V, Clark C. What is TechQuity? J Health Care Poor Underserved 2021;32(2S):xiii-viii. [doi: 10.1353/hpu.2021.0045]

XSL•FO
RenderX

20. Aldana A, Richards-Schuster K. Youth-led antiracism research: making a case for participatory methods and creative strategies in developmental science. J Adolesc Res 2021 Sep 18;36(6):654-685 [FREE Full text] [doi: 10.1177/07435584211043289]
21. Sawyer RD, Norris J. Duoethnography: Understanding Qualitative Research. Oxford, UK: Oxford University Press; 2012.
22. Breault RA. Emerging issues in duoethnography. Int J Qual Stud Educ 2016 Apr 04;29(6):777-794. [doi: 10.1080/09518398.2016.1162866]
23. Thematic analysis. The University of Auckland. URL: https://www.thematicanalysis.net/ [accessed 2022-10-01]
24. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
25. What is thematic analysis? Victoria Clarke YouTube page. URL: https://www.youtube.com/watch?v=4voVhTiVydc [accessed 2024-12-19]
26. Byrne D. A worked example of Braun and Clarke's approach to reflexive thematic analysis. Qual Quant 2021 Jun 26;56(3):1391-1412. [doi: 10.1007/s11135-021-01182-y]
27. Braun V, Clarke V. Reflecting on reflexive thematic analysis. Qual Res Sport Exerc Health 2019 Jun 13;11(4):589-597. [doi: 10.1080/2159676x.2019.1628806]
28. ATLAS.ti Mac and Windows (version 22) [Qualitative data analysis software]. ATLAS.ti Scientific Software Development GmbH. URL: https://atlasti.com [accessed 2022-12-13]
29. Ding X, Wuerth K, Sakakibara B, Schmidt J, Parde N, Holsti L, et al. Understanding mobile health and youth mental health: scoping review. JMIR Mhealth Uhealth 2023 Jun 16;11:e44951 [FREE Full text] [doi: 10.2196/44951] [Medline: 37220197]

Abbreviations

FAB: FostrSpace Advisory Board
JJBH: Juvenile Justice Behavioral Health
TA: thematic analysis
UCSF: University of California, San Francisco
YPAR: youth participatory action research

Edited by A Mavragani; submitted 29.09.23; peer-reviewed by M Simmons, J D'Arcey; comments to author 18.12.23; revised version received 29.02.24; accepted 20.11.24; published 20.01.25.

Please cite as:
Ezimora I, Lundberg T, Miars D, Trujeque J, Papias A, Del Cid MV, Folk JB, Tolou-Shams M
Reflections of Foster Youth Engaging in the Co-Design of Digital Mental Health Technology: Duoethnography Study
JMIR Form Res 2025;9:e53231
URL: https://formative.jmir.org/2025/1/e53231
doi:10.2196/53231
PMID:

©Ifunanya Ezimora, Tylia Lundberg, Dylan Miars, Jeruel Trujeque, Ashley Papias, Margareth V Del Cid, Johanna B Folk, Marina Tolou-Shams. Originally published in JMIR Formative Research (https://formative.jmir.org), 20.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Developing a Machine Learning–Based Automated Patient Engagement Estimator for Telehealth: Algorithm Development and Validation Study

Pooja Guhan[1], PhD; Naman Awasthi[1], PhD; Kathryn McDonald[2], BS; Kristin Bussell[3], PhD; Gloria Reeves[2], MD; Dinesh Manocha[1], PhD; Aniket Bera[4], PhD

[1]Department of Computer Science, University of Maryland, College Park, MD, United States
[2]Department of Psychiatry, Child and Adolescent Division, University of Maryland, Baltimore, MD, United States
[3]School of Nursing, University of Maryland, Baltimore, MD, United States
[4]Department of Computer Science, Purdue University, West Lafayett, IN, United States

Corresponding Author:
Pooja Guhan, PhD
Department of Computer Science
University of Maryland
8125 Paint Branch Dr
College Park, MD, 20742
United States
Phone: 1 2406309133
Fax: 1 (301) 405 6707
Email: pguhan@umd.edu

Abstract

Background: Patient engagement is a critical but challenging public health priority in behavioral health care. During telehealth sessions, health care providers need to rely predominantly on verbal strategies rather than typical nonverbal cues to effectively engage patients. Hence, the typical patient engagement behaviors are now different, and health care provider training on telehealth patient engagement is unavailable or quite limited. Therefore, we explore the application of machine learning for estimating patient engagement. This can assist psychotherapists in the development of a therapeutic relationship with the patient and enhance patient engagement in the treatment of mental health conditions during tele–mental health sessions.

Objective: This study aimed to examine the ability of machine learning models to estimate patient engagement levels during a tele–mental health session and understand whether the machine learning approach could support therapeutic engagement between the client and psychotherapist.

Methods: We proposed a multimodal learning-based approach. We uniquely leveraged latent vectors corresponding to affective and cognitive features frequently used in psychology literature to understand a person's level of engagement. Given the labeled data constraints that exist in health care, we explored a semisupervised learning solution. To support the development of similar technologies for telehealth, we also plan to release a dataset called Multimodal Engagement Detection in Clinical Analysis (MEDICA). This dataset includes 1229 video clips, each lasting 3 seconds. In addition, we present experiments conducted on this dataset, along with real-world tests that demonstrate the effectiveness of our method.

Results: Our algorithm reports a 40% improvement in root mean square error over state-of-the-art methods for engagement estimation. In our real-world tests on 438 video clips from psychotherapy sessions with 20 patients, in comparison to prior methods, positive correlations were observed between psychotherapists' Working Alliance Inventory scores and our mean and median engagement level estimates. This indicates the potential of the proposed model to present patient engagement estimations that align well with the engagement measures used by psychotherapists.

Conclusions: Patient engagement has been identified as being important to improve therapeutic alliance. However, limited research has been conducted to measure this in a telehealth setting, where the therapist lacks conventional cues to make a confident assessment. The algorithm developed is an attempt to model person-oriented engagement modeling theories within machine learning frameworks to estimate the level of engagement of the patient accurately and reliably in telehealth. The results are encouraging and emphasize the value of combining psychology and machine learning to understand patient engagement. Further testing in the real-world setting is necessary to fully assess its usefulness in helping therapists gauge patient engagement during

XSL•FO
RenderX

online sessions. However, the proposed approach and the creation of the new dataset, MEDICA, open avenues for future research and the development of impactful tools for telehealth.

(JMIR Form Res 2025;9:e46390) doi:10.2196/46390

KEYWORDS

machine learning; mental health; telehealth; engagement detection; patient engagement

Introduction

Overview

The World Health Organization defines mental health as a "state of well-being" that allows a person to lead a fulfilling and productive life and contribute to society [1]. In addition, the World Health Organization estimates that one-fourth of the adult population is affected by a mental disorder [2]. However, there are only approximately 9 psychiatrists per 100,000 people in higher-income countries and only approximately 0.1 psychiatrists for every 1,000,000 people in lower-income countries [3,4]. As stress and pressure continue to increase, leading to poor mental health outcomes, the need for improved tele–mental health care has become critical. Tele–mental health care is the process of providing psychotherapy remotely, typically using Health Insurance Portability and Accountability Act (HIPAA)–compliant videoconferencing [5]. It offers an effective means of accessing mental health services and treatment, transcending geographical and cultural boundaries worldwide, and helps address the chronic shortage of psychotherapists. These services not only remove practical barriers to care, such as transportation, but also offer affordability and direct access to qualified therapists. Therefore, there has been an upward trend in the demand for such services [6]. Despite these undeniable benefits, this emerging treatment modality raises new challenges in patient engagement compared to in-person care. By "engagement," we refer to the connection between a therapist and patient, characterized by a sense of basic trust and a willingness or interest to collaborate. This connection is essential for the therapeutic process and fosters the development of a strong therapeutic relationship.

Background

Patient engagement is a critical priority in behavioral health care as it involves establishing effective rapport between health care providers and patients. In the context of mental health, patient engagement is an indicator of a successful therapy session. This multidimensional concept [7,8] encompasses the interaction of cognitive and emotion-related components of the patient's psychological state. However, one of the challenges in fully understanding and evaluating patient engagement lies in the difficulty of establishing a standardized measurement approach. Patient engagement is a complex and dynamic construct that defies a one-size-fits-all measurement. The diverse factors influencing patient engagement, such as individual characteristics, and health care contexts make it challenging to develop a universally accepted standard of measuring its effectiveness. Traditional quantitative metrics (appointment attendance rates, treatment adherence, and patient satisfaction) often fall short in capturing the depth and richness of patient engagement. While they can provide numerical data, they may overlook the subjective experiences of the interaction and do not necessarily reflect health care provider-patient alliance in treatment. In contrast, qualitative methods such as patient narratives and feedback provide valuable insights but lack scalability, objectivity, and uniformity across different health systems. Therefore, to address these limitations, machine learning techniques are being explored to complement and enhance these traditional methods by leveraging computational power, scalability, and ability to identify patterns, relationships, and insights that may not be otherwise apparent. Some prior works in engagement detection have focused on using a single modality, such as facial expressions [9,10], speech [11], body posture [12], gaze direction [13], or head pose [14], to detect engagement. Combining different modalities has been observed to improve engagement detection accuracy [15-17]. The authors expanded the work of Stanford PBL Lab's eRing [18] by including information streams such as facial expressions, voice, and other biometric data. Monkaresi et al [19] proposed an approach to detect engagement levels in students during a writing task by not only making use of facial features but also features obtained from remote video-based detection of heart rate. The dataset used was generated by the authors, and they used self-reports instead of external annotation for classification purposes. Chang et al [20] made use of facial expressions as well as body posture for detecting engagement in learners. Fedotov et al [21] proposed the use of audio, facial, and body pose features to detect engagement and disengagement for an in-the-wild dataset that had been created using videos obtained from uncontrolled environments or situations. The distribution of class labels in this dataset is not balanced.

Despite the existence of a variety of such algorithms to perform engagement detection, the results obtained from these approaches (particularly single modality–based approaches) could be misleading in a telehealth setting due to factors such as camera position and resistant or guarded clients. In telehealth appointments, therapists have limited visual data available to them, as they can only view the patient's face rather than their full body. Asking the patient to adjust their position to gain a better view of their body during telehealth appointments is a potential solution for therapists. However, this approach may lead to communication difficulties, as it can impact the patient's proximity to the microphone, potentially affecting the quality of audio communication between the patient and the therapist. Therefore, therapists must rely on verbal communication strategies to engage patients than in-person care because they cannot use typical nonverbal cues to convey interest and be responsive to the patient (eg, a handshake at the beginning of a session, adjusting the distance between the patient and health care provider by moving a chair closer or further away, and observing a patient's response to questions while maintaining eye contact). It is also more difficult for therapists to convey

XSL•FO
RenderX

attentiveness because eye contact requires the therapist to look at a camera rather than observe or look at a person. They may also have limited training on telehealth patient engagement, leading to a lack of guidance on measuring patient engagement. Therefore, this highlights the need for a system that can provide feedback on patient engagement by just using the data that are easily accessible, namely, text, audio, and face visuals. Engagement is critical for both retention of patients in care as well as effectiveness of mental health treatment. Developing such a system will make it possible to enhance the quality of tele–mental health and ultimately improve patient outcomes.

Objective

Taking all this into consideration, we propose and explore a machine learning–based approach that takes visual, audio, and text data as input and estimates the engagement level of a patient with mental health problems through a regression-based approach. We demonstrate the effectiveness of our method in tele–mental health sessions and provide a new dataset called Multimodal Engagement Detection in Clinical Analysis (MEDICA) to advance mental health research by understanding patient engagement levels during therapy sessions. MEDICA consists of 1229 short video clips from mock mental health therapy sessions used by medical schools in their psychiatry teaching curriculum. To the best of our knowledge, MEDICA is the first multimodal dataset that focuses on mental health research and consists of annotations useful for understanding the psychological state of patients with mental health problems during a therapy session.

Methods

Datasets

MEDICA Dataset

Overview

Engagement is an overloaded term, and the definition varies with the application, making it difficult and expensive to collect, annotate, and analyze such data. As a result, we find too few multimodal-based engagement detection datasets currently available for us to use. Our problem statement revolves specifically around detecting patient engagement during a tele–mental health session. In such a setting, the only information we can work with includes the patient's face and speech (audio and text). There exist datasets such as Carnegie Mellon University Multimodal Corpus of Sentimental Intensity (CMU-MOSI) [22], Carnegie Mellon University Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) [23], and Stanford Emotional Narratives Dataset (SEND) [24] that capture such settings. However, they are not specifically for engagement detection. Given the lack of a dataset that allows researchers to use multimodal features (video, text, and audio) for engagement, we propose MEDICA, a novel dataset developed specifically to cater to engagement detection using tele–mental health session videos. To use these data to address a broader range of issues related to mental health, we also include labels pertaining to stress and emotions.

To the best of our knowledge, this dataset is one of the first publicly available datasets specifically catering to multimodal research in patient engagement in mental health. Table 1 presents a comparison between MEDICA and other related datasets. Despite the rise in telehealth services and the poor patient-to-therapist ratios for individuals with mental health problems, there are no datasets that even try modeling telehealth sessions to give the community an opportunity to innovate and develop new technologies. MEDICA is a humble attempt by us to kick-start interesting research opportunities. The MEDICA dataset is unique in that it contains multimodal data from multiple sensors including speech and video. These data have been annotated with the engagement levels exhibited by each participant at different points in time.

XSL•FO
RenderX

Table 1. Comparison of the Multimodal Engagement Detection in Clinical Analysis (MEDICA) dataset with other related datasets[a].

Dataset name	Samples	Unique speakers	Modes	Emotion labels	Engagement information	Other mental health cues
RECOLA[b] [25]	3400	27	{v,a}	Yes	No	Physiological (electrocardiogram and electrothermal activity)
CMU MOSEAS[c] [26]	715	Multiple	{v,a,t}	Yes	No	—[d]
CMU-MOSI[e] [22]	2199	Multiple	{v,a,t}	Yes	No	—
CMU-MOSEI[f] [23]	3228	1000	{v,a,t}	Yes	No	—
SEND[g] [24]	193	49	{v,a,t}	Yes	No	—
DAiSEE[h] [27]	9068	112	{v}	No	Yes	—
HBCU[i] [9]	120	34	{v}	No	Yes	—
In the wild [28]	195	78	{v}	No	Yes	—
SDMATH [29]	20	20	{v,a}	Yes	Yes	—
MEDICA	1229[b]	13	{v,a,t}	Yes	Yes	Hesitation, stress, and attentiveness

[a]Modes indicate the subset of modalities present from visual (v), audio (a), and text (t).
[b]RECOLA: Remote Collaborative and Affective.
[c]CMU MOSEAS: Carnegie Mellon University Multimodal Opinion Sentiment, Emotions and Attributes.
[d]Not applicable.
[e]CMU-MOSI: Carnegie Mellon University Multimodal Corpus of Sentiment Intensity.
[f]CMU-MOSEI: Carnegie Mellon University Multimodal Opinion Sentiment and Emotion Intensity.
[g]SEND: Stanford Emotional Narratives Dataset.
[h]DAiSEE: Dataset for Affective States in E-Environments.
[i]HBCU: historically black college and university.
[j]Current status of the dataset. The size of the dataset will be increased.

Acquisition

MEDICA has been developed by collecting publicly available mock therapy session videos created by different psychiatry medical schools for training their students. The patients in these videos are being advised for depression, social anxiety, and posttraumatic stress disorder. We collected 13 videos, each having a duration of approximately 20 to 30 minutes. We limit the videos to the setup where both the therapist and the patient are not visible together in the same frame. In addition, we take only those videos where there is only 1 patient. Each video has a unique English-speaking patient. The therapists in these videos are actual therapists, while the patients are trained actors who have been dictated to act and react in a particular way. Figure 1 shows the snapshots taken from various video clips in the MEDICA dataset.

Figure 1. Examples from the Multimodal Engagement Detection in Clinical Analysis (MEDICA) dataset created for mental health research. This dataset has been created using publicly available videos that are usually used for training purposes by different medical schools. The faces of the individuals have been blurred to protect their identity.

Processing and Annotation

Our primary focus was to create a dataset that captures the behavior of patients with mental health problems during their sessions. Therefore, we selected only portions of the videos where only the patient was visible in the frames, which appeared at various intervals and durations throughout the recordings. We took these scattered clips and divided them into smaller clips of 3 seconds each, resulting in a dataset of size 1229. We use Moviepy and speech-recognition libraries to extract audio and text from the video clips, respectively. Each video was annotated for attentiveness, stress, and engagement, which were

XSL•FO
RenderX

scored on a Likert scale ranging from –3 to 3; hesitation was a binary target variable (yes or no). Humans tend to have multiple emotions with varying intensities while expressing their thoughts and feelings. Therefore, the videos have been labeled for 8 emotions related to mental health: happy, sad, irritated, neutral, anxious, embarrassed, scared, and surprised. This will enable us to develop systems capable of understanding the various interacting emotions of the users.

Real-World Data

The videos collected to develop MEDICA were created by the psychiatry schools in controlled settings. By controlled, we mean that the therapists in these videos were real, while the patients were trained method actors who had been instructed in a very detailed manner regarding what to expect and how they should react. While these videos were approved by a team of psychotherapists as being inclusive of potential scenarios and patient reactions, the work would still be meaningless if not tested in the wild, that is, in settings where we have real patients who have not been instructed what to do or how to behave during the session with their real therapist. To accomplish this, we collaborated with 8 child psychotherapists. These psychotherapists are individuals who were trained and licensed as psychotherapists for both children and adults. The administrative staff identified patients that were scheduled for sessions with any of the participating therapists and provided the caregiver contact information. Our research assistant contacted eligible caregivers and provided information related to the purpose, potential risks and benefits, and responsibilities of study participation. They were informed of the study protocol that involved recording 1 to 2 telemental sessions to test our proposed approach. A total of 20 caregivers provided consent to participate in the study. They were provided with instructions on the setup of the equipment that we provided to ensure a clean recording. “Clean” refers to a recording executed with a camera of good quality with appropriate lighting conditions. The equipment included a smartphone with a good quality camera, a ring light with a stand to ensure that the session was recorded in a well-lit environment, and a hot spot internet connection to ensure that the session occurred smoothly without any network glitches. They were also given the assurance regarding preserving the confidentiality of the data being collected. All video recordings were labeled only by study number, with all identifiable information removed and stored in a double password–protected database. Participants were also informed that any facial images other than the caregiver who participated in the session would be deidentified (using methods such as blurring, etc). These steps were taken to ensure that only the caregiver’s level of engagement was measured and analyzed. In addition, this step protected the confidentiality of nonparticipants who did not provide consent.

On average, each of these sessions lasted approximately 20 minutes. The demographics of the caregivers who participated in our real-world experiments can be found in Multimedia Appendix 1. The entire data collection process can be divided into three parts:

1. Before each tele–mental health session of a caregiver with their therapist, a research assistant helped the caregiver with setting up the equipment to record their session. The assistant also ensured that the caregivers were comfortable using the equipment.
2. During the session, we made sure that the tele–mental health session proceeded in the same manner as a typical session. The caregivers were requested to use the equipment provided if possible, but there were no restrictions imposed by us on either the therapist or the caregiver regarding how the session should be conducted. After the presession process, the research assistant would log off. Therefore, during the session, it would be just the therapist and the caregiver having a conversation. No one else from the study would be a part of it. The only thing different about this session was that the caregiver was being recorded using the smartphone given to them. We did not record the therapist.
3. After the session, after the session was complete, a research assistant guided the participant regarding the steps to stop the recording and save the data collected.

Upon completion of the session, the therapist scored the quality of the collaborative relationship (therapeutic alliance) established between them during the session with the Working Alliance Inventory (WAI) [30,31]. It is used for both child and adult therapy and is a widely studied therapeutic alliance measure. The WAI was modeled on the theoretical work conducted by Bordin [32]. It captures 3 dimensions of the alliance: bond, task, and goals. Extensive tests showed 12 items per dimension to be the minimum length for effective representations of the inventory. A composite score is computed based on these 12 items for each of the sessions conducted. Henceforth, we refer to this score as the WAI score. The WAI scores can range from 12 to 84 [33,34]. Figure 2 shows frames from a few of the video recordings collected.

XSL•FO
RenderX

Figure 2. A few frames from the real-world videos we collected. The faces have been blurred here to protect the identity of the patients. However, the consent of the patients was taken to use their unblurred faces as input to our model.

While efforts were taken to ensure clean video recordings, we still could not use 7 out of the 20 videos collected for any type of analysis. The reasons for it ranged from the video being extremely shaky due to the continuous movement of some participants during their session to not having the permissions to process and analyze videos where the child spoke for the entire session instead of the caregiver. The remaining videos were processed in the same way as MEDICA. At the end of data processing, we obtained approximately 438 clips, each lasting 3 seconds. Unlike MEDICA where we had annotations for every 3 seconds, we had only a single WAI score for the entire session. Therefore, all the 3-second clips obtained from a session were annotated with the WAI score obtained for the session. Any analysis conducted on these data treated all the 438 clips independently.

Proposed Model Design

Figure 3 presents an overview of our engagement estimation model. It has 2 key components: *multimodal feature extractor* and *semisupervised learning network*. We discuss each of them in detail in the subsequent sections.

XSL•FO
RenderX

Figure 3. Overall block diagram of the proposed architecture. The features obtained from the 2 modalities, cognitive and affective, are combined to form hT (real) of shape 655×1. The generator takes in random noise (Gaussian) and tries to generate a fake feature map that looks similar to hfakeT that looks similar to hT (real). The discriminator tries to distinguish between hT and hfakeT. The yellow bars in the discriminator and generator refer to linear layers with activation function as leaky rectified linear activation unit. The shape of the linear layers used has been written on the bars. The number of linear layers used can be changed depending upon the dataset size and task at hand.

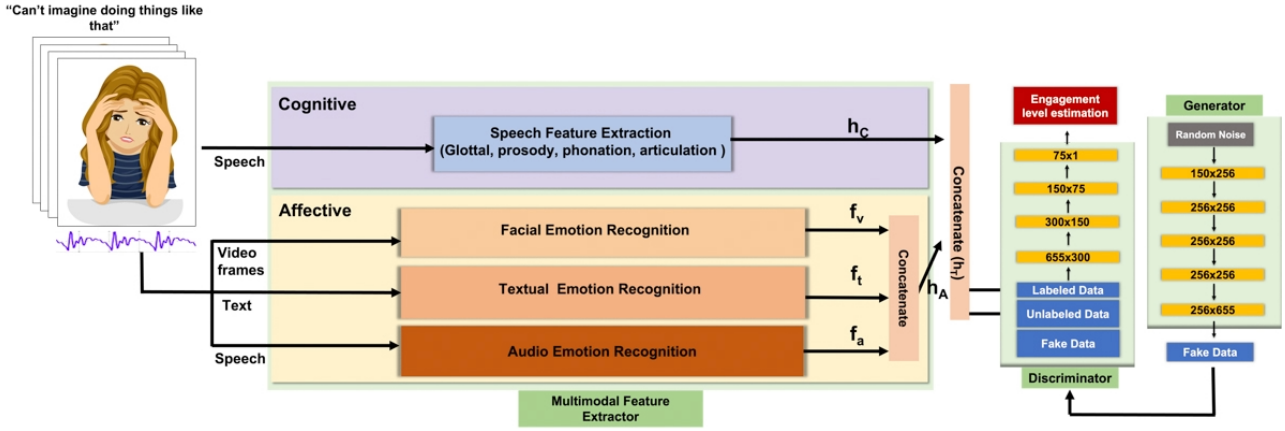

Multimodal Feature Extractor

Overview

Multimodal feature extractor in general refers to a system or algorithm that extracts and combines features from multiple modalities, such as audio, video, text, and other types of data, to provide a more comprehensive representation of input data. The goal of multimodal feature extraction is to capture complementary information from different sources, which can lead to better performance in various tasks. In our model, the multimodal feature extractor is used to comprehend the psychological state of the patient with mental health problems. We take inspiration from literature in psychology and psychiatry to build an extractor that can enable the learning network to recognize, understand, and evaluate engagement as close as possible to that of a psychotherapist. Therefore, we proposed a multicomponent approach to build the feature extractor and introduce 2 modules, namely, cognitive and affective states, to capture the different cues used by psychotherapists to assess their patients. As mentioned earlier, the visual data available are limited. Therefore, apart from the visuals, we also need to make use of verbal information, that is, understanding the speaking style of the person (audio) and what the person said (text).

Therefore, given this information corresponding to a participant, the multimodal feature extractor outputs a feature vector h_T. h_T is obtained after concatenating the feature vectors h_C and h_A obtained from the cognitive and affective state modalities (components of the multimodal feature extractor), respectively. We now discuss the cognitive and affective state modalities in more detail in the subsequent sections.

Cognitive State Modality

The cognitive state involves comprehending complex concepts and issues and acquiring difficult skills. It conveys deep (rather than surface level) processing of information, where the person gains a critical or higher-order understanding of the subject matter and solves challenging problems.

Psychotherapists usually measure and evaluate the cognitive state of the person using the mental status examination, typically conducted via in-person interviews or self-evaluations, to gauge memory, thinking, and the extent of understanding of the topic of discussion. There has been a lot of work around determining biomarkers for detecting signs of a person's cognitive state. However, these methods are either offline or fail to consider various essential perceptual indicators. Recently, there has been a lot of work around using speech as a potential biomarker for detecting cognitive decline. For instance, stress negatively affects the cognitive functions of a person, and this can be easily detected using speech signals. Moreover, speech-based methods are attractive because they are nonintrusive, inexpensive, and can be done in real time. Four distinct features have proven to be highly effective in identifying signs of cognitive impairment and are increasingly utilized to detect conditions like Alzheimer's and Parkinson's diseases. These features include glottal (f_g), prosody (f_{pr}), phonation (f_{ph}), and articulation measures (f_{ar}), each offering unique insights into speech characteristics associated with cognitive decline.

Glottal features are acoustic measures used to characterize speech under stress. The glottal is the space between the vocal cords in the larynx, and the opening and closing of the vocal cords create sound during speech. During periods of stress, there is an aberration in the amount of tension applied in the opening (abduction) and closing (adduction) of the vocal cords. Prosody features characterize the speaker's intonation and speaking styles. Under this feature, we analyze variables such as timing, pitch, and loudness during the production of speech. Phonation refers to the production of vocal sounds by the vibration of the vocal cords. In individuals with cognitive decline, phonation is often characterized by bowing and inadequate closure of the vocal cords. These changes can lead to instability and irregularity in the vibration of the vocal cords, which in turn affects the production of speech. They are analyzed in terms of features related to perturbation measures, such as jitter (temporal perturbations of the fundamental frequency), shimmer (temporal perturbation of the amplitude of the signal), amplitude perturbation quotient, and pitch perturbation quotient. Apart from these, the degree of unvoiced is also included. Articulation

XSL•FO
RenderX

features refer to the movements of the lips, tongue, and jaw during speech production. Reduced amplitude and velocity of these movements can impact speech intelligibility. The analysis of articulation is primarily based on the first 2 vocal formants F1 and F2, which correspond to the frequency of the sound waves produced by the vocal tract.

For a given audio input, we extract these features using *librosa* [35] and *praat* [36] libraries. Therefore, we define features corresponding to cognitive state as the concatenation of these 4 audio features. Therefore, cognitive state features

$$h_c = \text{concat}\ (f_g, f_{pr}, f_{ph}, f_{ar}) \quad \textbf{(1)}$$

Affective State Modality

The affective state encompasses emotional reactions such as excitement, boredom, curiosity, and anger. The range of affective expressions will vary based on individual demographic factors (eg, age), cultural backgrounds or norms, and mental health symptoms.

To understand the affective state, we check if there exists any inconsistency between the emotions perceived and the statement the person made. Balomenos et al [37] and Porter and ten Brinke [38] suggest that when different modalities are modeled and projected onto a common space, they should point to similar affective cues; otherwise, the incongruity suggests distraction, deception, etc. In other words, if E1, E2, and E3 represent the emotions perceived individually from what the patient said (text), the way they said it or sounded (audio), and how they looked or expressed (visuals), respectively, then the patient would be considered engaged if E1, E2, and E3 are similar; otherwise, they would be considered disengaged. Therefore, we adopt pretrained emotion recognition models to extract affective features corresponding to audio, visuals, and text from each video sample separately.

Audio features (f_a) are extracted using mel-frequency cepstral coefficients (MFCC) from audio clips available in the dataset. MFCC is a commonly used audio feature extraction technique in signal processing and machine learning. It is based on the concept of the mel scale, which is a nonlinear transformation of frequency that is more perceptually relevant to human hearing. By applying MFCC to audio signals, we can extract features that capture important aspects of the sound, such as pitch, timbre, and spectral shape. These features are often used in speech and music analysis as well as in various machine learning applications, including emotion recognition and speaker identification. In our work, the extracted MFCC features from audio are then passed as input to a multilayer perceptron network trained for emotion recognition in speech, specifically using the data available in the Crowd-source Emotional Multimodal Actors Dataset (CREMA-D) to extract affective features from the audio data. Through this method, we obtained a feature vector f_a, for each audio clip, which provides valuable insights into the emotional characteristics of the speech.

Visual features (f_v) correspond to the affective features extracted from video frames to understand the emotions expressed through facial cues of the patient. We used a deep learning architecture called Visual Geometry Group-B (VCG-B), which has been pretrained as described in the study by Arriaga et al [39], to extract the affective features. We modified the output dimensions of the second last layer of the network to produce a feature vector f_v of length 100.

Text (f_t), characterizing the emotions, sentiment, and other affective cues present in the text, is obtained using a Bidirectional Encoder Representations from Transformers(BERT)–based model, which has been pretrained on the GoEmotions dataset [40]. The Bidirectional Encoder Representations from Transformers model works by processing the text input through a multilayer neural network that is trained to predict the context of each word based on its surrounding words. This allows the model to capture the complex relationships between words and their meanings, which are essential for understanding the affective context of the text.

Therefore, we represent the affective state of the patient as a concatenation of f_a, f_v, and f_t. Hence, affective state features

$$h_A = \text{concat}\ (f_a, f_v, f_t) \quad \textbf{(2)}$$

Therefore, the overall output of the multimodal feature extractor is a feature vector h_T*which c*an be defined as

$$h_T = \text{concat}\ (h_C, h_A)$$

The learning network discussed in the next section makes use of this.

Learning Network

In this section, we discuss the second part of our proposed model, that is, the semisupervised learning network. Different machine learning techniques can be explored to solve the task of predicting the level of engagement. However, obtaining a large amount of high-quality labeled data to train a robust model for predicting patient engagement is inevitably laborious and requires expert medical knowledge. Considering that unlabeled data are relatively easy to collect, we propose a semisupervised learning–based solution. Semisupervised learning enables us to deploy machine learning systems in real-life applications (eg, image search [41], speech analysis [42,43], and natural language processing) where we have few labeled data samples and a lot of unlabeled data. Some prior works have also explored semisupervised learning for engagement detection in nonmedical domains. One of the earliest studies in this area is by Alyuz et al [44], where they consider the development of an engagement detection system, more specifically emotional or affective engagement of students, in a semisupervised fashion to personalize systems such as intelligent tutoring systems according to their needs. Nezami et al [45] conducted experiments to detect user engagement using a facial feature–based semisupervised model. Most state-of-the-art semisupervised learning methods use generative adversarial nets (GANs) [46].

GANs are a class of machine learning models and typically have 2 neural networks competing with each other to generate more accurate predictions. These 2 neural networks are referred to as the generator and the discriminator. The generator's goal is to artificially manufacture outputs (ie, fake data) that could easily be mistaken as real data. The goal of the discriminator is

XSL•FO
RenderX

to identify the real from the artificially generated data. In trying to generate high-quality outputs, the generator learns to capture the different possible variations in the input variables, and therefore, the data manifold well. This is extremely helpful when we may not be able to access data containing a wide variety of similar engagement-related cues visible across different patients. In general, the training process of GANs involves a dynamic interplay and competition between the generator and discriminator networks. Both the generator and discriminator networks improve their abilities iteratively. The generator aims to generate more realistic samples, while the discriminator aims to become better at distinguishing between real and generated (artificial) data. The eventual goal of the training process is for the generator and discriminator to reach Nash equilibrium [47]. Once the equilibrium is achieved, no player (generator or discriminator) has any incentive to deviate toward a more optimal position, that is, no player has a better strategy given the choices of the other player.

Taking all this into consideration, we propose using a multimodal semisupervised GAN as our learning network for regressing on the levels of patient engagement. This type of machine learning model combines multiple modalities (in this case, cognitive and affective modalities) to generate realistic samples (or features) that can fool the discriminator. The term *semisupervised* refers to the network's ability to use a limited amount of labeled data alongside a larger amount of unlabeled data to enhance its accuracy. Such a network will generalize better and be more robust compared to previously defined semisupervised learning approaches. It is different than the semisupervised GAN framework SR-GAN proposed by Olmschenk et al [48] in 2 main ways. First, our network is multimodal, whereas SR-GAN is not. Second, unlike the SR-GAN generator, which focuses on generating realistic images, our GAN's generator is designed to create realistic feature maps that mimic the multimodal feature vector h_T, derived from the fusion of cognitive and affective state modalities (as discussed in the previous section). The generator takes Gaussian noise as input and produces a synthetic feature vector h_{fakeT}, aiming to deceive the discriminator into classifying it as h_{fakeT}. The discriminator's role, in turn, is to distinguish h_T as a fake representation and correctly identify h_T as the genuine feature vector.

Overall, our multimodal semisupervised GAN can therefore be divided into 3 parts: multimodal feature extractor, generator, and discriminator.

For the purpose of training, we make use of the following four standard loss functions, namely, labeled loss (L_{lab}), unlabeled loss (L_{un}), fake loss (L_{fake}), and generator loss (L_{gen}):

1. L_{lab}—this loss is similar to typical supervised learning regression losses, where the goal is to assess the network's ability to estimate the engagement level accurately, as close as possible to the ground truth (ie, the actual engagement level annotated for the data sample). Therefore, we compute the mean squared error of model output (, ie, predicted level of engagement) with ground truth (y_t, ie, actual level of engagement).

2. L_{un}—in semisupervised GANs, the labeled dataset contains a small portion of data with labels, while most of the data are unlabeled. The goal of the L_{un} loss function is to improve the performance of the generator by minimizing the distance between the feature spaces of the labeled and unlabeled datasets. This helps to ensure that the generator produces feature maps that are consistent with both labeled and unlabeled data, thereby improving the generalization performance of the model. By minimizing the distance between the feature spaces of labeled and unlabeled datasets, the model can effectively learn from both labeled and unlabeled data, which is particularly useful in cases where labeled data is scarce.

3. L_{fake}—it is a loss function used to update the discriminator network. The objective of this loss is to maximize the distance between the feature space of unlabeled dataset and fake features generated by the generator. This enables the discriminator to distinguish the real and fake data features better.

4. L_{gen}—the generator network's objective is to produce feature vectors h_T (fake) that can be passed off as real feature vectors h_T: L_{gen} updates the generator network to do just that. L_{gen} encourages the generator to learn the underlying distribution of unlabeled data features (real) and generate fake features that match the feature statistics of the real features.

Similar to SR-GAN, we also make use of a gradient penalty (P) to keep the gradient of the discriminator in check, which helps convergence. The gradient penalty is calculated with respect to a randomly chosen point on the convex manifold connecting the unlabeled feature vector h_T to the fake feature vector h_T. We compute it as

Here, $p_{interpolate}$ examples are generated by $\alpha p_{unlabeled} + (1 - \alpha)p_{fake}$ for $\alpha \sim U(0,1)$.

$U(0,1)$ represents a uniform distribution over the range from 0 to 1.

The overall loss function used for training the network is:

$$L = L_{lab} + L_{un} + L_{fake} + L_{gen} + \lambda P \quad \textbf{(8)}$$

Overall Pipeline

In this section, we present an overview of how the various components of our proposed approach interact to estimate the engagement level of a patient with mental health problems during a tele–mental health session. We provide insights into the working mechanism of our model and shed light on the

XSL•FO
RenderX

underlying processes involved in engagement level estimation in a multimodal setting.

As depicted in Figure 3, the proposed model involves data preprocessing, which includes extracting video frames, audio signal, and text data from the input video clip. To capture engagement dynamics over small time scales, we analyze engagement at the microlevel, considering video clips of a few seconds duration. This aligns with the person-oriented analysis suggested by Sinatra et al [49]. Specifically, for the MEDICA dataset and real-world data, we consider video clips of 3 seconds duration. All samples, labeled and unlabeled, from the dataset are treated as *real* data for the semisupervised GAN. The multimodal feature extractor (consisting of the cognitive and affective state modalities) is used to obtain the h_T features from the preprocessed visual, audio, and text data. h_T is considered real. In parallel, the generator network creates h_{fakeT} (ie, fake h_T) features using gaussian noise as input. The labeled and unlabeled h_T features obtained from the multimodal feature extractor and the h_{fakeT} features obtained from the generator are then input to the discriminator network. In addition to distinguishing between real (h_T) and fake (h_{fakeT}) features, the discriminator network outputs the engagement level of the patient with mental health problems as a continuous value.

The subsequent sections outline the different studies being conducted to verify our approach.

Study 1: Testing Our Proposed Approach on MEDICA

The purpose of the first study is to test the applicability of our approach to estimate the levels of engagement exhibited by the participants present in the videos. The test was conducted on the MEDICA dataset. In addition, we compared the performance of the proposed model against other similar frameworks to understand its effectiveness. To test the models, we partitioned the dataset in a ratio of 70:10:20 for training, testing, and validation. Motivated by recent works in clinical psychotherapy [50-52], we used the standard evaluation metric of root mean square error (RMSE) to evaluate our approach. Smaller the RMSE value, better is the performance of the model for the engagement estimation task.

Study 2: Ablation Studies

The proposed model design involves the introduction of 2 modalities, namely, affective and cognitive. In this study, we explored the contribution of these 2 modalities for the purpose of engagement detection. We aimed to test the effectiveness of the model with and without these modalities. Therefore, we ran the overall pipeline described earlier by removing either one of the modalities corresponding to affective and cognitive states and reported our findings. The tests were conducted on the MEDICA dataset. Similar to our previous study, we performed the evaluation using RMSE.

Study 3: Analysis on Real-World Data

Evaluating the model only on MEDICA is not sufficient to prove the usability of the model in the real-world setting. Therefore, this study aimed to test our approach on data collected from real-world scenarios. The study involved testing the proposed model that has been trained on MEDICA using real-world data. We also compared the performance of our proposed model against other baselines explored in study 1 on the real-world data. Apart from testing our model's usability for the real world, we also tested its capability to perform on unseen real-world data, which is different from the data (MEDICA) it has been trained on. Contrary to previous experiments where we compared the performance on engagement estimation values, in this test, we will study the correlation of the model-estimated engagement values with the WAI score. This will help us to also test whether the scores estimated by our model are clinically useful for therapists.

Ethical Considerations

Both the MEDICA dataset and the real-world data collection processes have been approved by the institutional review boards (#HP-00092966) at both the University of Maryland College Park's computer science department and University of Maryland Baltimore medical school.

Participation of the caregivers while collecting the real-world data was completely voluntary. They could withdraw anytime without any negative consequences. They were informed about the purpose of the study and the analysis that would be conducted using the data collected.

The therapist and the participating caregiver were given the option of allowing us to store videotaped recordings for development of future studies or to request to have all videotape sessions destroyed after a 2-year period. The video recordings were stored in our laboratory databases and subject to strict University of Maryland privacy protocol, including encryption and password protection. While the research is ongoing, only select project personnel with internal clearance can have access to the data. In addition, we ensured that the data are untampered using standard cryptographic hash functions. During the video storage process, we deidentified any facial images beside the health care provider or caregiver who are recorded (eg, sibling interrupts session) so that they are not recognizable on the recorded video. All data related to enrollment, demographics, and questionnaires will be stored in an electronic database at the University of Maryland and will be double password protected with only select research members having access. The participating caregiver was given an individual ID number, and their data were deidentified to ensure it could not be linked back to them.

While the participating caregiver may not directly benefit from the study, they received 1-month free internet hot spot and a Fitbit, which were provided as part of the real-world study equipment. They were compensated with a US $20 gift card after returning the equipment.

Results

The semisupervised models were trained on an NVIDIA GeForce GTX 1080 Ti GPU with a batch size of 512 and a learning rate of 0.0001 for 50 epochs.

Study 1: Testing Our Proposed Approach on MEDICA

The purpose of the first study is to demonstrate the ability of our model to estimate the level of engagement exhibited by the

XSL•FO
RenderX

patient in the video. This study was performed on the MEDICA dataset. As our proposed methodology leverages a semisupervised approach, we extracted labeled samples from MEDICA and unlabeled samples from the CMU-MOSEI dataset. After preprocessing, we extracted 12,854 unlabeled data points from CMU-MOSEI. We split the 1229 labeled data points from MEDICA into 70:10:20 for training, validation, and testing, respectively. Therefore, the split of the labeled training data to unlabeled training data points was 860:12,854. In addition, the ground truth engagement levels corresponding to the training data points in the MEDICA dataset were normalized to fall within a range of 0 and 1. We compared our model with the following state-of-the-art methods for engagement detection:

1. Kaur et al [28] used a deep multiple instance learning–based framework for detecting engagement in students. They extracted local binary pattern on three orthogonal planes (LBP-TOP) features from the facial video segments and performed linear regression using a deep neural network to estimate the engagement scores.

2. Nezami et al [45] performed a semisupervised engagement detection using a semisupervised support vector machine.

In addition to being state-of-the-art, these methods can be used in a telehealth setting like ours. We used the publicly available implementation proposed by Kaur et al [28] and trained the entire model on MEDICA. Nezami et al [45] do not have a publicly available implementation. We reproduced the method to the best of our understanding.

Table 2 summarizes the RMSE values obtained for methods described by Kaur et al [28] and Nezami et al [45], including ours. We observed an improvement of at least 40%. Our approach is one of the first methods of engagement estimation built on the principles of psychotherapy. The modules used, specifically cognitive and affective states, help the overall model to effectively mimic the ways a psychotherapist perceives the patient's level of engagement. Similar to psychotherapists, these modules also look for specific engagement-related cues exhibited by the patient in the video.

Table 2. Comparison of RMSE values obtained from various methods for estimating levels of engagement on the Multimodal Engagement Detection in Clinical Analysis dataset.

Method	RMSE[a] for engagement
Kaur et al [28]	0.96
Nezami et al [45]	0.17
Our approach	0.10

[a]RMSE: root mean square error.

Study 2: Ablation Studies

To show the importance of the different components (affective and cognitive) used in our approach, we run our method on MEDICA by removing either one of the modules corresponding to affective or cognitive states and report our findings. Table 3 summarizes the results obtained from the ablation experiments. We observe that the ablated model (ie, only using affective [A] or cognitive [C] modules) does not perform as well as the model that includes both modules. To understand and verify the contribution of these modules further, we leveraged the other labels (stress, hesitation, and attention) available in MEDICA and performed regression tasks using our proposed architecture on all of them. We observed that mode C performed better when predicting stress and hesitation values. Mode A performed better in estimating a patient's level of attentiveness. These results agree with our understanding of cognitive state and affective state. Therefore, the combination of affective and cognitive state modes helps in efficiently predicting the engagement level of the patient.

Table 3. Ablation experiments on the Multimodal Engagement Detection in Clinical Analysis dataset. We ran our proposed model using only 1 modality at a time (either affective or cognitive) and compared the performance with that of using both modalities together.

Modality	RMSE[a] for engagement	RMSE for stress	RMSE for hesitation	RMSE for attention
Affective	0.24	0.15	0.146	0.07
Cognitive	0.3	0.13	0.16	0.08
Affective + cognitive (our approach)	0.10	0.12	0.14	0.1

[a]RMSE: root mean square error.

Study 3: Analysis on Real-World Data

Our model that has been trained for estimating engagement levels using the MEDICA dataset was tested on the processed real-world data. WAI scoring is based on certain observations the therapist makes during the session with the patient. The score obtained from our model was different from the WAI score, but we claim that similar to WAI, our estimates also captured the engagement levels of the patient well. If this is indeed the case, then both WAI and our estimates should be correlated. As discussed earlier, a single WAI score is reported by the therapist (health care provider) for the entire session. Unlike WAI, which reports a single score per session, our method, along with those proposed by Kaur et al [28] and Nezami et al [45], performs microanalysis, generating engagement estimates at multiple time points within a session.

XSL•FO
RenderX

To enable a fair comparison, we calculated the mean of the engagement estimates across each session for all methods. The mean engagement estimates across sessions were as follows: Kaur et al [28] had a mean of 0.50 (SD: 7.82e-17), Nezami et al [45] had a mean of 0.64 (SD: 0.017), and our approach had a mean of 0.65 (SD: 0.006). We subsequently examined each of their correlations with the corresponding WAI scores across sessions. Given that WAI scores and engagement estimates are measured on different scales, WAI scores were normalized to a 0-1 range for consistency with the engagement values from the methods. We found that Kaur et al [28] obtained a correlation score of –0.03, Nezami et al [45] obtained a score of –0.24 and our approach obtained a much better score of 0.38. We plot the engagement level estimations obtained for each session against its corresponding WAI score, as shown in Figure 4. The plot corresponding to our approach shows a noticeably better alignment with the WAI score trends compared to other methods [28,45].

Figure 4. Scatterplot between the engagement level estimations obtained from each method for each session: (A) the plot between engagement level estimations obtained from our approach and the Working Alliance Inventory (WAI) scores, (B) the plot between the engagement level estimations obtained by Kaur et al [28] for each session and the WAI scores, (C) the plot between the engagement level estimations for each session obtained by Nezami et al [45] and the WAI scores.

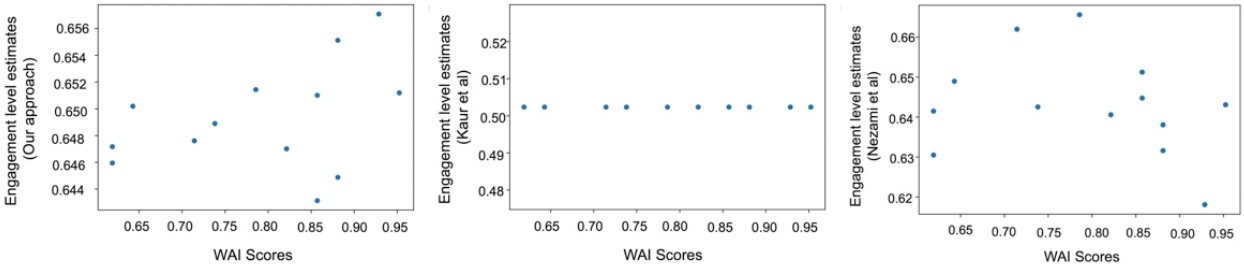

Additionally, instead of just taking the mean, we also took the median of the engagement level estimates available at different instances of the sessions. The average median of the engagement estimates across sessions for different methods were as follows: Kaur et al [28] had an average median of 0.50 (IQR 0.0), Nezami et al [45] had an average median of 0.64 (IQR 0.02), and our approach had an average median of 0.65 (IQR 0.007).

We then examined the median of the engagement estimates for each session, obtained using different methods, and assessed their correlation with the WAI scores. In this case, Kaur et al [28] obtained a correlation score of 0, Nezami et al [45] obtained a score of –0.18 and our approach obtained a relatively high score of 0.40. These findings reinforce that our model captures WAI patterns more effectively than prior methods. Its stronger positive correlation with the WAI scores indicates its capability to better estimate engagement levels during therapy sessions.

The conceptual model of our proposed approach is also supported by the theoretical work by Bordin [32]. According to this theory, the therapist-provider alliance is driven by 3 factors: bond, agreement on goals, and agreement on tasks; these factors align well with the features identified in this work. While bond would correspond with the affective state, goals and task agreement correspond with the cognitive state. The merit of the approach by Bordin [32] is that it has been used for child and adult therapy, and it is one of the more widely studied therapeutic alliance measures. Therefore, it is no surprise that our approach can work well to provide an estimate of engagement levels in a tele–mental health session.

Discussion

Principal Findings

In this study, we developed a machine learning–based model that integrates features related to cognitive and affective psychological states to predict patient engagement levels during tele–mental health sessions. The input to our proposed algorithm is the visual, audio, and text data available, while the output is the engagement level of the patient. Our investigation demonstrated the significant promise of our proposed method, achieving an average improvement of 40% in RMSE for predicting engagement levels on the newly introduced MEDICA dataset compared to prior works. In real-world settings, our method provided patient engagement estimates that closely aligned with the trends observed in WAI scores assigned by therapists based on their sessions [30,31]. This study highlights the potential of our multimodal engagement estimation algorithm to enhance the quality of telehealth services and ultimately improve patient outcomes.

Comparison With Prior Work

Patient engagement is recognized as a pivotal determinant of a successful therapy session, particularly in the realm of mental health care [53]. Current psychiatric literature [54-56] has extensively examined a range of strategies and methodologies designed to foster [57] and sustain patient engagement during therapy sessions. These strategies are crucial for achieving optimal therapeutic outcomes. There has been a growing interest in applying artificial intelligence (AI) to mental health [58-62], with innovative approaches being explored to enhance patient care. However, despite the proliferation of AI-based methods in mental health applications, there remains a noticeable gap in research specifically focused on understanding and quantifying engagement. This gap is even more evident in the context of tele–mental health [63]. The transition from traditional in-person therapy to digital platforms has introduced unique challenges, as conventional indicators of engagement, such as body language [64], eye contact [65], and physical presence, maybe less discernible or entirely absent in web-based settings.

XSL•FO
RenderX

Our research addresses this critical gap by pioneering an approach that estimates patient engagement specifically in the tele–mental health context. This work is the first of its kind to target engagement estimation in this setting, offering a novel perspective that has largely been overlooked in previous studies. While the broader concept of engagement has been explored by AI in various domains, existing methods [28,45,66] have not been adapted or tested within the unique demands of mental health care, particularly in a remote or web-based environment. From an algorithmic standpoint, previous works have explored both unimodal and multimodal learning-based networks to understand engagement. Unimodal approaches focus on individual modalities, such as speech, body posture, biometric data (eg, heart rate variability and skin conductance) [67,68], gaze direction [69], and head pose [70]. Multimodal approaches [71] integrate various signals, leveraging the complementary strengths of each modality to provide a more holistic understanding of engagement. In contrast to these works, our research focuses on using the types of information that are readily available during a tele–mental health session, specifically audio, video, and text data. This approach is designed to be practical and easily implementable in real-world telehealth settings. This also ensures that the engagement estimation process is minimally disruptive to the therapeutic experience during the telehealth session. We validated our approach using our newly introduced dataset, MEDICA. In addition, we tested it on data collected from real-world therapy sessions. Our method achieved state-of-the-art performance in both cases.

Strengths and Implications

In this work, we explored the multidimensional and temporally dynamic nature of patient engagement [72]. Our method estimates engagement at regular short intervals. These characteristics of our approach align perfectly with the person-oriented analysis discussed by Sinatra et al [49]. We released a new dataset, MEDICA, to enhance mental health research, specifically toward understanding the engagement levels of patients attending therapy sessions. To the best of our knowledge, there is no other multimodal dataset that caters specifically to the needs of the mental health–based research. In addition, while there are some image-based [9,27,28] or sensor information–based datasets, there is no dataset that addresses the possibility of exploring engagement detection using visual, audio, and text modalities. We verified the usefulness of our machine learning method using 3 studies. Study 1 demonstrates the ability of our model to estimate the engagement level exhibited by the patient in video. It was done to understand the usefulness of the proposed model in comparison to other existing engagement estimation methods such as those by Nezami et al [45] and Kaur et al [28]. Our approach reported an RMSE [50-52] of 0.1, which on an average is 40% less than the RMSE reported by other methods (Table 1). Our study is one of the first to estimate engagement specifically for patients with mental health problems in a tele–mental health setting. The modalities used, specifically cognitive and affective states, propel the overall algorithm to mimic the way a psychotherapist perceives the patient's level of engagement. As part of study 2, we were interested to understand the contribution of the 2 modalities (cognitive and affective states). The experiments conducted as part of this study are also termed as *ablation* as it involves running the network pipeline using only 1 of the 2 modalities available at a time. We leveraged the annotations in MEDICA on stress, hesitation, and attention to dig deeper regarding the contribution of the cognitive and affective state modalities. We ran the same ablation experiments on not only patient engagement but also stress, hesitation, and attention estimation. While the combination of affective and cognitive states gives better results (RMSE of 0.1), it was interesting to note that the observations obtained supported the theoretical discussion about the 2 states. Cognitive state modality was able to understand stress (RMSE of 0.13) better. Affective state that was built on the concept of capturing distraction using emotion inconsistency between visual, audio, and text data helped relate better with the understanding of attention (RMSE of 0.07). These results agree with our understanding of cognitive state and affective state discussed earlier.

WAI is a popular measure of patient engagement used by many mental health care providers. Therefore, in study 3, we wished to understand and check if the values being estimated by our method related well with the trends observed in WAI. We did so by computing the correlation scores between WAI and the values estimated by our proposed model. In addition, to test the effectiveness of our approach, we also compared the correlation scores obtained between WAI and patient engagement values estimated by other existing methods. Positive values are preferred because they indicate similarity in distribution behavior. Negative values would indicate that the trends observed in WAI and the engagement values predicted are different and opposite. The results (Table 2) revealed that our model was able to capture the trends in WAI much better than other engagement estimation methods, reinforcing its clinical relevance. The correlation strength observed between WAI and our method's estimated engagement values is better than other approaches, but it also indicates the need for further investigation. The trend is positive, suggesting that we have successfully modeled engagement to some extent. Further analysis will provide opportunities to enhance and strengthen the relationship between WAI and the engagement values estimated by machine learning.

Limitations and Future Work

One of the primary limitations of the proposed algorithm is that engagement predictions may not be optimal in case of occlusions or missing modalities, data corruption due to low internet bandwidth, and integration of wearable devices with our model. We recognize that this situation is likely to occur in a tele–mental health session and plan to incorporate solutions for it in our future work. In addition, we aim to explore ways to make the predictions more explainable, allowing psychotherapists to receive evidence-based suggestions to support their final decisions.

Conclusions

Telehealth behavioral services, delivered via videoconferencing systems, have emerged as a cost-effective, dependable, and secure option for mental health treatment, particularly in recent times. However, gauging patient engagement, a critical mental

XSL•FO
RenderX

health care standard, remains challenging in telehealth due to the absence of nonverbal cues and sensory data such as heart rate. To address this, we propose a novel multimodal semisupervised GAN that leverages affective and cognitive features from psychology to estimate engagement levels using visuals, audio, and text from video calls. This approach can significantly enhance social interactions and assist psychotherapists during tele–mental health sessions. Engagement is typically assessed through patient reports, which are susceptible to response bias, and current measures such as "show rate" and "patient satisfaction" do not accurately reflect the health care provider–patient alliance. Our model demonstrates its effectiveness on both our newly introduced dataset called MEDICA as well as real-world data. Given the lack of systematic engagement measurement and the limited training for health care providers on telehealth engagement, our proposed system offers a promising solution to these challenges, promoting better patient-health care provider interactions and making telehealth more effective.

Acknowledgments

This research was fully supported by the MPower grant.

Data Availability

The datasets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Demographic information on the real-world experiment participants.
[DOCX File , 14 KB - formative_v9i1e46390_app1.docx]

References

1. WHO highlights urgent need to transform mental health and mental health care. World Health Organization. 2022 Jun 17. URL: https://www.who.int/news/item/17-06-2022-who-highlights-urgent-need-to-transform-mental-health-and-mental-health-care [accessed 2024-11-12]
2. The world health report 2001: mental health; new understanding, new hope. World Health Organization. 2001. URL: https://iris.who.int/handle/10665/42390 [accessed 2024-11-12]
3. Murray CJ, Vos T, Lozano R, Naghavi M, Flaxman AD, Michaud C, et al. Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet 2012 Dec 15;380(9859):2197-2223. [doi: 10.1016/S0140-6736(12)61689-4] [Medline: 23245608]
4. Oladeji BD, Gureje O. Brain drain: a challenge to global mental health. BJPsych Int 2016 Aug;13(3):61-63 [FREE Full text] [doi: 10.1192/s2056474000001240] [Medline: 29093905]
5. Telemental health/providers. Anxiety & Depression Association of America. URL: https://adaa.org/find-help/treatment-help/telemental-health-providers [accessed 2024-11-12]
6. Greenwood H, Krzyzaniak N, Peiris R, Clark J, Scott AM, Cardona M, et al. Telehealth versus face-to-face psychotherapy for less common mental health conditions: systematic review and meta-analysis of randomized controlled trials. JMIR Ment Health 2022 Mar 11;9(3):e31780 [FREE Full text] [doi: 10.2196/31780] [Medline: 35275081]
7. Kelders SM, van Zyl LE, Ludden GD. The concept and components of engagement in different domains applied to eHealth: a systematic scoping review. Front Psychol 2020;11:926 [FREE Full text] [doi: 10.3389/fpsyg.2020.00926] [Medline: 32536888]
8. Yardley L, Spring BJ, Riper H, Morrison LG, Crane DH, Curtis K, et al. Understanding and promoting effective engagement with digital behavior change interventions. Am J Prev Med 2016 Nov;51(5):833-842. [doi: 10.1016/j.amepre.2016.06.015] [Medline: 27745683]
9. Whitehill J, Serpell Z, Lin YC, Foster A, Movellan JR. The faces of engagement: automatic recognition of student engagement from facial expressions. IEEE Trans Affective Comput 2014 Jan 1;5(1):86-98. [doi: 10.1109/taffc.2014.2316163]
10. Murshed M, Dewan MA, Lin F, Wen D. Engagement detection in e-learning environments using convolutional neural networks. In: Proceedings of the IEEE International Conference on Dependable, Autonomic and Secure Computing, International Conference on Pervasive Intelligence and Computing, International Conference on Cloud and Big Data Computing, International Conference on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech). 2019 Presented at: DASC/PiCom/CBDCom/CyberSciTech 2019; August 5-8, 2019; Fukuoka, Japan. [doi: 10.1109/dasc/picom/cbdcom/cyberscitech.2019.00028]
11. Yu C, Aoki PM, Woodruff A. Detecting user engagement in everyday conversations. arXiv Preprint posted online on October 13, 2004 [FREE Full text] [doi: 10.21437/interspeech.2004-327]

XSL•FO
RenderX

12. Sanghvi J, Castellano G, Leite I, Pereira A, McOwan PW, Paiva A. Automatic analysis of affective postures and body motion to detect engagement with a game companion. In: Proceedings of the 6th International Conference on Human-Robot Interaction. 2011 Presented at: HRI '11; March 6-9, 2011; Lausanne, Switzerland. [doi: 10.1145/1957656.1957781]
13. Nakano YI, Ishii R. Estimating user's engagement from eye-gaze behaviors in human-agent conversations. In: Proceedings of the 15th International Conference on Intelligent User Interfaces. 2010 Presented at: IUI '10; February 7-10, 2010; Hong Kong, China. [doi: 10.1145/1719970.1719990]
14. Sharma P, Joshi S, Gautam S, Maharjan S, Khanal SR, Reis MC, et al. Student engagement detection using emotion analysis, eye tracking and head movement with machine learning. arXiv Preprint posted online on September 18, 2019 [FREE Full text] [doi: 10.1007/978-3-031-22918-3_5]
15. Psaltis A, Apostolakis KC, Dimitropoulos K, Daras P. Multimodal student engagement recognition in prosocial games. IEEE Trans Games 2018 Sep;10(3):292-303. [doi: 10.1109/tciaig.2017.2743341]
16. Grafsgaard JF, Wiggins JB, Boyer KE, Wiebe EN, Lester LC. Embodied affect in tutorial dialogue: student gesture and posture. In: Proceedings of the 16th International Conference on Artificial Intelligence in Education. 2013 Presented at: AIED 2013; July 9-13, 2013; Memphis, TN. [doi: 10.1007/978-3-642-39112-5_1]
17. Aslan S, Cataltepe Z, Dundar O, Esme AA, Ferens R. Learner engagement measurement and classification in 1:1 learning. In: Proceedings of the 13th International Conference on Machine Learning and Applications. 2014 Presented at: ICMLA 2014; December 3-6, 2014; Detroit, MI. [doi: 10.1109/icmla.2014.111]
18. Ma J, Fruchter R. eRing: Body Motion Engagement Detection and Feedback in Global Teams. Stanford, CA: Stanford University; Apr 2015.
19. Monkaresi H, Bosch N, Calvo RA, D'Mello SK. Automated detection of engagement using video-based estimation of facial expressions and heart rate. IEEE Trans Affective Comput 2017 Jan 1;8(1):15-28. [doi: 10.1109/taffc.2016.2515084]
20. Chang C, Zhang C, Chen L, Liu Y. An ensemble model using face and body tracking for engagement detection. In: Proceedings of the 20th ACM International Conference on Multimodal Interaction. 2018 Presented at: ICMI '18; October 16-20, 2018; Boulder, CO. [doi: 10.1145/3242969.3264986]
21. Fedotov D, Perepelkina O, Kazimirova E, Konstantinova M, Minker W. Multimodal approach to engagement and disengagement detection with highly imbalanced in-the-wild data. In: Proceedings of the Workshop on Modeling Cognitive Processes from Multimodal Data. 2018 Presented at: MCPMD '18; October 16, 2018; Boulder, CO. [doi: 10.1145/3279810.3279842]
22. Zadeh A, Zellers R, Pincus E, Morency LP. MOSI: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. arXiv Preprint posted online on June 20, 2016 [FREE Full text]
23. Zadeh AB, Liang PP, Poria S, Cambria E, Morency LP. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. 2018 Presented at: ACL 2018; July 15-20, 2018; Melbourne, Australia. [doi: 10.18653/v1/p18-1208]
24. Ong DC, Wu Z, Zhi-Xuan T, Reddan M, Kahhale I, Mattek A, et al. Modeling emotion in complex stories: the Stanford Emotional Narratives Dataset. IEEE Trans Affect Comput 2021 Jul 1;12(3):579-594 [FREE Full text] [doi: 10.1109/taffc.2019.2955949] [Medline: 34484569]
25. Ringeval F, Sonderegger A, Sauer J, Lalanne D. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In: Proceedings of the 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG). 2013 Presented at: FG 2013; April 22-26, 2013; Shanghai, China. [doi: 10.1109/fg.2013.6553805]
26. Zadeh A, Cao YS, Hessner S, Liang PP, Poria S, Morency LP. CMU-MOSEAS: a multimodal language dataset for Spanish, Portuguese, German and French. Proc Conf Empir Methods Nat Lang Process 2020 Nov;2020:1801-1812 [FREE Full text] [doi: 10.18653/v1/2020.emnlp-main.141] [Medline: 33969362]
27. Gupta A, D'Cunha A, Awasthi K, Balasubramanian V. DAiSEE: towards user engagement recognition in the wild. arXiv Preprint posted online on September 7, 2016 [FREE Full text]
28. Kaur A, Mustafa A, Mehta L, Dhall A. Prediction and localization of student engagement in the wild. arXiv Preprint posted online on April 3, 2018 [FREE Full text] [doi: 10.1109/dicta.2018.8615851]
29. Sathayanarayana S, Satzoda RK, Carini A, Lee M, Salamanca L, Reilly J. Towards automated understanding of student-tutor interactions using visual deictic gestures. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2014 Presented at: CVPRW 2014; June 23-28, 2014; Columbus, OH. [doi: 10.1109/cvprw.2014.77]
30. Berger T, Boettcher J, Caspar F. Internet-based guided self-help for several anxiety disorders: a randomized controlled trial comparing a tailored with a standardized disorder-specific approach. Psychotherapy (Chic) 2014 Jun;51(2):207-219. [doi: 10.1037/a0032527] [Medline: 24041199]
31. Gómez Penedo JM, Berger T, Grosse Holtforth M, Krieger T, Schröder J, Hohagen F, et al. The Working Alliance Inventory for guided Internet interventions (WAI-I). J Clin Psychol 2020 Jun;76(6):973-986. [doi: 10.1002/jclp.22823] [Medline: 31240727]
32. Bordin ES. The generalizability of the psychoanalytic concept of the working alliance. Psychother Theory Res Pract 1979;16(3):252-260. [doi: 10.1037/h0085885]
33. Herrero R, Vara MD, Miragall M, Botella C, García-Palacios A, Riper H, et al. Working alliance inventory for online interventions-short form (WAI-TECH-SF): the role of the therapeutic alliance between patient and online program in

XSL•FO
RenderX

therapeutic outcomes. Int J Environ Res Public Health 2020 Aug 25;17(17):6169 [FREE Full text] [doi: 10.3390/ijerph17176169] [Medline: 32854381]
34. Shinno S, Matsumoto K, Hamatani S, Inaba Y, Ozawa Y, Kawasaki Y, et al. Feasibility of guided internet-based cognitive behavioral therapy for panic disorder and social anxiety disorder in Japan: pilot single-arm trial. JMIR Form Res 2024 Feb 29;8:e53659 [FREE Full text] [doi: 10.2196/53659] [Medline: 38421717]
35. McFee B, Raffel C, Liang D, Ellis DP, McVicar M, Battenberg E, et al. librosa: audio and music signal analysis in python. In: Proceedings of the 14th Python in Science Conference. 2015 Presented at: SciPy 2015; July 6-12, 2015; Austin, TX. [doi: 10.25080/majora-7b98e3ed-003]
36. Magdin M, Sulka T, Tomanová J, Vozár M. Voice analysis using PRAAT software and classification of user emotional state. Int J Interact Multimed Artif Intell 2019;5(6):33-42. [doi: 10.9781/ijimai.2019.03.004]
37. Balomenos T, Raouzaiou A, Ioannou S, Drosopoulos A, Karpouzis K, Kollias S. Emotion analysis in man-machine interaction systems. In: Proceedings of the First International Workshop on Machine Learning for Multimodal Interaction. 2004 Presented at: MLMI 2004; June 21-23, 2004; Martigny, Switzerland. [doi: 10.1007/978-3-540-30568-2_27]
38. Porter S, ten Brinke L. Reading between the lies: identifying concealed and falsified emotions in universal facial expressions. Psychol Sci 2008 May;19(5):508-514. [doi: 10.1111/j.1467-9280.2008.02116.x] [Medline: 18466413]
39. Arriaga O, Valdenegro-Toro M, Plöger P. Real-time convolutional neural networks for emotion and gender classification. arXiv Preprint posted online on October 20, 2017.
40. Demszky D, Movshovitz-Attias D, Ko J, Cowen A, Nemade G, Ravi S. GoEmotions: a dataset of fine-grained emotions. arXiv Preprint posted online on May 1, 2020 [FREE Full text] [doi: 10.18653/v1/2020.acl-main.372]
41. Lu K, Zhao J, Xia M, Zeng J. Semi-supervised learning for image retrieval using support vector machines. In: Proceedings of the Second International Symposium on Neural Networks. 2005 Presented at: ISNN 2005; May 30-June 1, 2005; Chongqing, China. [doi: 10.1007/11427391_108]
42. Yu D, Varadarajan B, Deng L, Acero A. Active learning and semi-supervised learning for speech recognition: a unified framework using the global entropy reduction maximization criterion. Comput Speech Lang 2010 Jul;24(3):433-444. [doi: 10.1016/j.csl.2009.03.004]
43. Liu Y, Kirchhoff K. Graph-based semi-supervised learning for phone and segment classification. In: Proceedings of the INTERSPEECH 2013. 2013 Presented at: INTERSPEECH 2013; August 25-29, 2013; Lyon, France URL: https://www.isca-archive.org/interspeech_2013/liu13e_interspeech.html# [doi: 10.21437/interspeech.2013-453]
44. Alyuz N, Okur E, Oktay E, Genc U, Aslan S, Mete SE, et al. Semi-supervised model personalization for improved detection of learner's emotional engagement. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction. 2016 Presented at: ICMI '16; November 12-16, 2016; Tokyo, Japan. [doi: 10.1145/2993148.2993166]
45. Nezami OM, Richards D, Hamey L. Semi-supervised detection of student engagement. In: Proceedings of the PACIS 2017. 2017 Presented at: PACIS 2017; July 16-20, 2017; Langkawi, Malaysia.
46. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. In: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2. 2014 Presented at: NIPS'14; December 8-13, 2014; Montreal, QC.
47. Saeed AQ, Sheikh Abdullah SN, Che-Hamzah J, Abdul Ghani AT. Accuracy of using generative adversarial networks for glaucoma detection: systematic review and bibliometric analysis. J Med Internet Res 2021 Sep 21;23(9):e27414 [FREE Full text] [doi: 10.2196/27414] [Medline: 34236992]
48. Olmschenk G, Zhu Z, Tang H. Generalizing semi-supervised generative adversarial networks to regression using feature contrasting. arXiv Preprint posted online on on November 27, 2018 [FREE Full text]
49. Sinatra GM, Heddy BC, Lombardi D. The challenges of defining and measuring student engagement in science. Educ Psychol 2015 Feb 27;50(1):1-13. [doi: 10.1080/00461520.2014.1002924]
50. Békés V, Aafjes-van Doorn K, Zilcha-Mano S, Prout T, Hoffman L. Psychotherapists' acceptance of telepsychotherapy during the COVID-19 pandemic: a machine learning approach. Clin Psychol Psychother 2021 Nov;28(6):1403-1415 [FREE Full text] [doi: 10.1002/cpp.2682] [Medline: 34723404]
51. Vail AK, Girard J, Bylsma L, Cohn J, Fournier J, Swartz H. Goals, tasks, and bonds: toward the computational assessment of therapist versus client perception of working alliance. In: Proceedings of the 16th IEEE International Conference on Automatic Face and Gesture Recognition. 2021 Presented at: FG 2021; December 15-18, 2021; Jodhpur, India. [doi: 10.1109/fg52635.2021.9667021]
52. Tracey TJ, Kokotovic AM. Factor structure of the Working Alliance Inventory. Psychol Assess J Consult Clin Psychol 1989 Sep;1(3):207-210. [doi: 10.1037/1040-3590.1.3.207]
53. Ewalds Mulliez AP, Pomey MP, Bordeleau J, Desbiens F, Pelletier JF. A voice for the patients: evaluation of the implementation of a strategic organizational committee for patient engagement in mental health. PLoS One 2018 Oct 24;13(10):e0205173 [FREE Full text] [doi: 10.1371/journal.pone.0205173] [Medline: 30356239]
54. Sagen JS, Smedslund G, Simonsen AE, Habberstad A, Kjeken I, Dagfinrud H, et al. Patient engagement in the development and delivery of healthcare services: a systematic scoping review. BMJ Open Qual 2023 Jun 27;12(2):e002309 [FREE Full text] [doi: 10.1136/bmjoq-2023-002309] [Medline: 37369560]

XSL•FO
RenderX

55. Khosravi M, Azar G. Factors influencing patient engagement in mental health chatbots: a thematic analysis of findings from a systematic review of reviews. Digit Health 2024 Apr 22;10:20552076241247983 [FREE Full text] [doi: 10.1177/20552076241247983] [Medline: 38655378]
56. Remmert JE, Destefano L, Chinman M, Oslin DW, Mavandadi S. Development of a peer-delivered primary care intervention to improve veteran mental health treatment engagement. Psychol Serv 2024 Aug 01. [doi: 10.1037/ser0000883] [Medline: 39088007]
57. Jørgensen K, Lerbæk B, Frederiksen J, Watson E, Karlsson B. Patient participation in mental health care - perspectives of healthcare professionals and patients: a scoping review. Issues Ment Health Nurs 2024 Aug 20;45(8):794-810. [doi: 10.1080/01612840.2024.2358931] [Medline: 38900284]
58. Elyoseph Z, Levkovich I. Comparing the perspectives of generative AI, mental health experts, and the general public on schizophrenia recovery: case vignette study. JMIR Ment Health 2024 Mar 18;11:e53043 [FREE Full text] [doi: 10.2196/53043] [Medline: 38533615]
59. Elyoseph Z, Levkovich I, Shinan-Altman S. Assessing prognosis in depression: comparing perspectives of AI models, mental health professionals and the general public. Fam Med Community Health 2024 Jan 09;12(Suppl 1):e002583 [FREE Full text] [doi: 10.1136/fmch-2023-002583] [Medline: 38199604]
60. Thieme A, Hanratty M, Lyons M, Palacios J, Marques RF, Morrison C, et al. Designing human-centered AI for mental health: developing clinically relevant applications for online CBT treatment. ACM Trans Comput Hum Interact 2023 Mar 17;30(2):1-50. [doi: 10.1145/3564752]
61. Rollwage M, Habicht J, Juechems K, Carrington B, Viswanathan S, Stylianou M, et al. Using conversational AI to facilitate mental health assessments and improve clinical efficiency within psychotherapy services: real-world observational study. JMIR AI 2023 Dec 13;2:e44358 [FREE Full text] [doi: 10.2196/44358] [Medline: 38875569]
62. Roy K, Khandelwal V, Goswami R, Dolbir N, Malekar J, Sheth A. Demo alleviate: demonstrating artificial intelligence enabled virtual assistance for telehealth: the mental health case. Proc AAAI Conf Artif Intell 2023 Jun 26;37(13):16479-16481. [doi: 10.1609/aaai.v37i13.27085]
63. Siegel A, Zuo Y, Moghaddamcharkari N, McIntyre RS, Rosenblat JD. Barriers, benefits and interventions for improving the delivery of telemental health services during the coronavirus disease 2019 pandemic: a systematic review. Curr Opin Psychiatry 2021 Jul 01;34(4):434-443 [FREE Full text] [doi: 10.1097/YCO.0000000000000714] [Medline: 33928918]
64. Foley GN, Gentile JP. Nonverbal communication in psychotherapy. Psychiatry (Edgmont) 2010 Jun;7(6):38-44 [FREE Full text] [Medline: 20622944]
65. Mitzkovitz C, Dowd SM, Cothran T, Musil S. The eyes have it: psychotherapy in the era of masks. J Clin Psychol Med Settings 2022 Dec 03;29(4):886-897 [FREE Full text] [doi: 10.1007/s10880-022-09856-x] [Medline: 35118604]
66. Wang Y, DeVito Dabbs A, Thomas TH, Campbell G, Donovan H. Measuring engagement in provider-guided digital health interventions with a conceptual and analytical framework using nurse WRITE as an exemplar: exploratory study with an iterative approach. JMIR Form Res 2024 Jul 22;8:e57529 [FREE Full text] [doi: 10.2196/57529] [Medline: 39037757]
67. Fragueiro A, Debroize RP, Bannier E, Cury C. Eye-tracking and skin conductance to monitor task engagement during neurofeedback sessions. In: Proceedings of the 9th Graz Brain Computer Interface Conference. 2024 Presented at: BCI 2024; September 9-12, 2024; Graz, Austria.
68. Ventura RB, Porfiri M. Galvanic skin response as a measure of engagement during play in virtual reality. In: Proceedings of the ASME 2020 Dynamic Systems and Control Conference. 2020 Presented at: ASME 2020; October 5-7, 2020; Virtual Event. [doi: 10.1115/DSCC2020-3177]
69. Varma M, Washington P, Chrisman B, Kline A, Leblanc E, Paskov K, et al. Identification of social engagement indicators associated with autism spectrum disorder using a game-based mobile app: comparative study of gaze fixation and visual scanning methods. J Med Internet Res 2022 Feb 15;24(2):e31830 [FREE Full text] [doi: 10.2196/31830] [Medline: 35166683]
70. Goldberg P, Sümer Ö, Stürmer K, Wagner W, Göllner R, Gerjets P, et al. Attentive or not? Toward a machine learning approach to assessing students' visible engagement in classroom instruction. Educ Psychol Rev 2019 Dec 18;33:27-49. [doi: 10.1007/s10648-019-09514-z]
71. Booth BM, Bosch N, D'Mello SK. Engagement detection and its applications in learning: a tutorial and selective review. Proc IEEE 2023 Oct;111(10):1398-1422. [doi: 10.1109/jproc.2023.3309560]
72. Bernard A, de Ossorno Garcia S, Salhi L, John A, DelPozo-Banos M. Patterns of engagement in a digital mental health service during COVID-19: a cohort study for children and young people. Front Psychiatry 2023 Jul 27;14:1143272 [FREE Full text] [doi: 10.3389/fpsyt.2023.1143272] [Medline: 37575580]

Abbreviations

AI: artificial intelligence
CMU-MOSEI: Carnegie Mellon University Multimodal Opinion Sentiment and Emotion Intensity
CMU-MOSI: Carnegie Mellon University Multimodal Corpus of Sentimental Intensity
GAN: generative adversarial net
HIPAA: Health Insurance Portability and Accountability Act

XSL•FO
RenderX

MEDICA: Multimodal Engagement Detection in Clinical Analysis
MFCC: mel-frequency cepstral coefficients
RMSE: root mean square error
WAI: Working Alliance Inventory
CREMA-D: Crowd-source Emotional Multimodal Actors Dataset
VGG-B: Visual Geometry Group-B
BERT: Bidirectional Encoder Representations from Transformers
LBP-TOP: local binary pattern on three orthogonal planes

Edited by A Mavragani; submitted 09.02.23; peer-reviewed by R Marshall, T Huang; comments to author 13.04.23; revised version received 30.06.23; accepted 03.09.24; published 20.01.25.

Please cite as:
Guhan P, Awasthi N, McDonald K, Bussell K, Reeves G, Manocha D, Bera A
Developing a Machine Learning–Based Automated Patient Engagement Estimator for Telehealth: Algorithm Development and Validation Study
JMIR Form Res 2025;9:e46390
URL: https://formative.jmir.org/2025/1/e46390
doi:10.2196/46390
PMID:

©Pooja Guhan, Naman Awasthi, Kathryn McDonald, Kristin Bussell, Gloria Reeves, Dinesh Manocha, Aniket Bera. Originally published in JMIR Formative Research (https://formative.jmir.org), 20.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Psychological and Behavioral Insights From Social Media Users: Natural Language Processing–Based Quantitative Study on Mental Well-Being

Xingwei Yang[1], PhD; Guang Li[2], PhD

[1]Information Technology Management, Ted Rogers School of Management, Toronto Metropolitan University, Toronto, ON, Canada
[2]Smith School of Business, Queen's University, Kingston, ON, Canada

Corresponding Author:
Xingwei Yang, PhD
Information Technology Management
Ted Rogers School of Management
Toronto Metropolitan University
55 Dundas St W
Toronto, ON, M5G 2C3
Canada
Phone: 1 416 979 5044
Email: nancy.yang@torontomu.ca

Abstract

Background: Depression significantly impacts an individual's thoughts, emotions, behaviors, and moods; this prevalent mental health condition affects millions globally. Traditional approaches to detecting and treating depression rely on questionnaires and personal interviews, which can be time consuming and potentially inefficient. As social media has permanently shifted the pattern of our daily communications, social media postings can offer new perspectives in understanding mental illness in individuals because they provide an unbiased exploration of their language use and behavioral patterns.

Objective: This study aimed to develop and evaluate a methodological language framework that integrates psychological patterns, contextual information, and social interactions using natural language processing and machine learning techniques. The goal was to enhance intelligent decision-making for detecting depression at the user level.

Methods: We extracted language patterns via natural language processing approaches that facilitate understanding contextual and psychological factors, such as affective patterns and personality traits linked with depression. Then, we extracted social interaction influence features. The resultant social interaction influence that users have within their online social group is derived based on users' emotions, psychological states, and context of communication extracted from status updates and the social network structure. We empirically evaluated the effectiveness of our framework by applying machine learning models to detect depression, reporting accuracy, recall, precision, and F_1-score using social media status updates from 1047 users along with their associated depression diagnosis questionnaire scores. These datasets also include user postings, network connections, and personality responses.

Results: The proposed framework demonstrates accurate and effective detection of depression, improving performance compared to traditional baselines with an average improvement of 6% in accuracy and 10% in F_1-score. It also shows competitive performance relative to state-of-the-art models. The inclusion of social interaction features demonstrates strong performance. By using all influence features (affective influence features, contextual influence features, and personality influence features), the model achieved an accuracy of 77% and a precision of 80%. Using affective features and affective influence features also showed strong performance, achieving 81% precision and an F_1-score of 79%.

Conclusions: The developed framework offers practical applications, such as accelerating hospital diagnoses, improving prediction accuracy, facilitating timely referrals, and providing actionable insights for early interventions in mental health treatment plans.

(JMIR Form Res 2025;9:e60286) doi:10.2196/60286

XSL•FO
RenderX

KEYWORDS

social media; natural language processing; social interaction; decision support system; depression detection

Introduction

Depression is characterized by a constant depressed mood and the inability to anticipate happiness or pleasure; [1] it has emerged as one of the major challenges globally [2]. Approximately 350 million people experience depression worldwide [3]. Without timely psychiatric interventions, a patient may have worsening symptoms, leading to more severe consequences such as self-harm, physical disability, or even suicide [4]. This problem has been exacerbated by the COVID-19 pandemic. Fear of COVID-19 infection, lack of social support, isolation, and limited access to medical resources during the COVID-19 pandemic have increased psychological distress [5,6]. Compared to the prepandemic period, depression rates almost doubled in 2020 to 2021 [7]. Despite its severity and seriousness, depression remains significantly underdiagnosed and undertreated [8]. Thus, there is an increasing need for intelligent clinical decision support tools to offer timely depression diagnoses and referral services so that people can receive early intervention or treatment before developing more serious mental health conditions.

Traditional methods of depression detection and treatment are based on questionnaires and personal interviews, which usually require long one-on-one physical engagements where psychiatrists gather information from patients and then recommend medications or treatments. This process can be both time consuming and costly, not only because of its complexity but also because of the expenses associated with the required trained personnel [9]. Thus, it is challenging for mental health professionals to offer immediate detection, referral services, proactive interventions, or treatments for those at risk and prepare for the next public health emergency, such as COVID-19, particularly when in-person assessments are difficult or costly. This challenge has motivated the exploration of more cost-effective and efficient methods, such as leveraging data from social media [10]. Platforms such as Meta are commonly used to share feelings and emotions. This makes social media a new venue for depression investigations [11-13] because they provide an unbiased exploration of language use and behavioral patterns of individuals [14]. Moreover, research has shown that depression can affect how individuals express themselves in their social media posts [15]. Therefore, there is an increasing trend toward using abundant text data and information from social media to examine individuals' sentiments and behaviors in mental health studies [16,17]. However, manually analyzing texts is not practical for timely psychiatric support due to the exponential growth of text data on social media and the increasing number of cases of mental illnesses. Thus, there is a rising need to use machine learning (ML) and natural language processing (NLP) techniques to deal with the large volume of data. Existing research in this domain primarily focuses on finding the key indicators, such as linguistic and sentiment features, to enhance the accuracy of depression detection [18,19]. For example, many studies focus on extracting textual content from social media data using linguistic features such as semantics and syntactic features [20], a bag of words [21], and n-grams [22]. Although these features can help yield reasonable results in depression detection, they overlook several critical aspects.

First, these linguistic features do not thoroughly reflect an individual's psychological patterns and states. Affective patterns (affect refers to emotions that change or influence the behaviors or thoughts of an individual) in language expressions are important dimensions of psychological patterns. They can reveal an individual's moods, sentiments, and emotional factors beyond simple linguistic features for analyzing mental disorders on social media. Studies show that frequent negative affect expression has higher connections with depressive states, while positive affect tends to have a lower connection to depression [23-25]. Thus, emotions play an essential role in diagnosing mental illnesses such as depression [11]. In addition, an individual's psychological states are directly related to depression. Personality has been used to distinguish social media users' activities, status updates content, and language use habits [26,27]. Moreover, Klein et al [28] and Hakulinen et al [29] identified the links between depression and personality types and found that certain personality traits may serve as predictive factors and influence the treatment of depression. Second, contextual information about the words is another critical dimension. Because pretrained language models have shown strong contextual modeling ability, they are used in understanding texts in various applications [30,31]. By considering the sequence of all words in the documents, sequence-level semantics can be learned to reflect the diverse contexts and meanings of words in sentences. For example, in the sentence "My car was left on the left-hand side," the word "left" is used 2 times, but it carries 2 different meanings. The first "left" is a verb suggesting the action performed on the car, while the second "left" is a noun specifying the car's location. The contextual meaning (eg, the meaning derived from the words' surroundings) can vary for the same word. Thus, contextual information is also a crucial aspect of understanding individuals' mental states.

Third, existing research often overlooks the impact of social interactions when analyzing features for mental disorders on social media. Social interactions reflect how users influence each other within an online social group. According to social influence theory, individuals in social environments significantly shape each other's attitudes, beliefs, and actions [32]. Recent psychological studies show that family, friends, or colleagues can directly or indirectly affect mental well-being [33]. For instance, in internet-based communities, recreational discussions, or chat groups, a negative mood can spread among different individuals [34,35]. Mental illnesses are influenced by an individual's daily social interactions [36]. This impact can be even more substantial on social media, where individuals are influenced by their friends' shared conventions and norms. Therefore, investigating the influence of social interaction is crucial for monitoring psychological patterns and states on social media. Studies by Murarka et al [37], Cui et al [38], and Ghosal

XSL•FO
RenderX

and Jain [39] analyzed emotions, personality, and pretrained models for mental health problems, but they did not account for the impact of social interactions. To study the mental well-being of social media users, an individual's mental health is reflected by more than just the content of their status updates. Thus, in our study, we not only account for psychological patterns and states and contextual information but also incorporate the influence of social interactions by considering the social network structure within these feature patterns.

This study aimed to offer a novel methodological language framework to enhance intelligent decision-making in the mental health domain. Our contributions are outlined as follows:

- We developed a unified language framework using NLP and ML techniques to enhance the detection of depression at the user level, making it more effective and accurate.
- Our framework advances previous studies by extracting and analyzing language patterns to comprehensively understand the contextual and psychological features (eg, affective patterns and personality) associated with depressive symptoms.
- We extracted features of social interaction influence, capturing how users influence their online social group. These features consider users' emotions, psychological states, and communication context from status updates and social network structure, differentiating individuals with depression from those without.

Our study aims for early-stage detection and initial screening to identify individuals at risk of getting depression and narrow down those needing closer observation or follow-up. It is designed to complement, not replace, clinical diagnoses, offering insights that support timely interventions and enhance patient care. Our framework can enrich the reliability and knowledge base of decision support systems for health care providers, enabling more precise diagnoses and timely referrals, interventions, and treatment plans. The framework is designed to be adaptable, enabling researchers to apply similar methods to their own datasets and explore psychological, contextual, and social influence factors. While it provides valuable insights, the framework is not intended to draw definitive conclusions but rather to offer a structured approach to understanding social interactions and their impact. Beyond its focus on mental illness, it can be extended to other domains, generating interpretable findings tailored to specific contexts.

The paper is organized as follows. In the Methods section, we introduce our proposed framework and present the language feature discovery processes and the influence analysis. In the Results section, we empirically evaluate the effectiveness of our framework. We discuss the implications of our results, suggest directions for future research, and conclude the paper in the Discussion section.

Methods

Psychological and Contextual Analysis NLP Framework

This section presents our methods for inferring language patterns and conducting an analysis of social interactions from social media content. Figure 1 shows a conceptual representation of the proposed psychological and contextual analysis NLP framework to extract affective, personality, contextual, and social interaction influence features. Given the data collected from social media platforms, our framework starts with extensive data preparation and cleaning procedures. In addition, we establish a user-friendship network based on users' social connections on social media. This network shows if one user is connected to others (ie, if this user is a friend of other users).

Figure 1. The proposed psychological and contextual analysis natural language processing framework. It starts with extensive data preparation and cleaning procedures and establishes a user-friendship network based on users' social connections on social media. We then apply machine learning models to effectively manage high-volume and high-speed user-generated data to analyze the depression states of users within social media environments. This framework can be used in various applications.

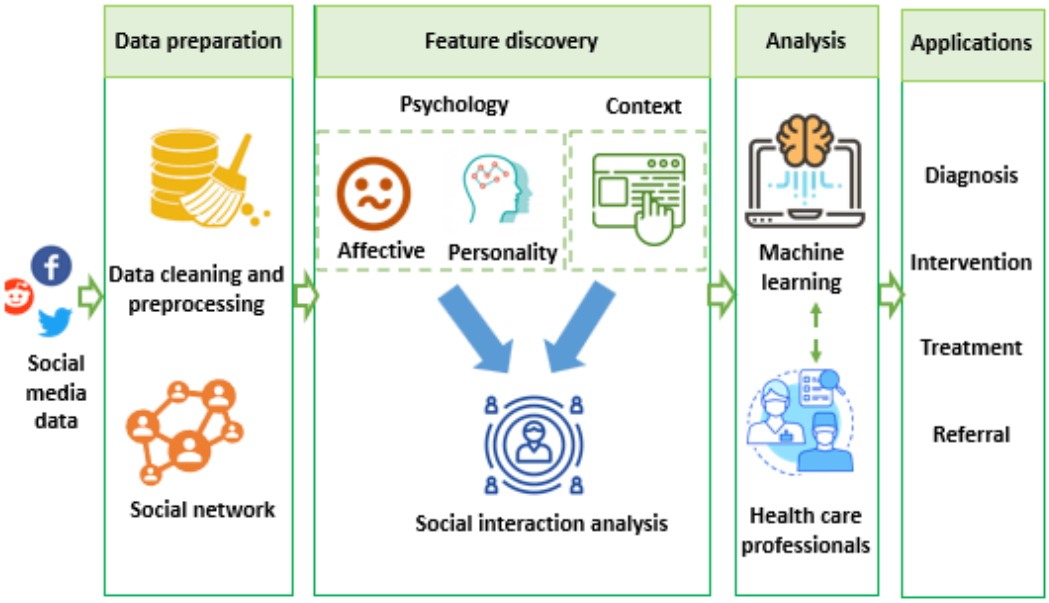

XSL•FO
RenderX

In the feature discovery phase, our framework deduces depressive behaviors from social media users, considering both contextual and psychological perspectives. We first focus on extracting features, including (1) affective features—these suggest users' psychological patterns such as mood, sentiment, and emotional factors as expressed in their social media postings, (2) personality features—these reveal how users' psychological states (ie, personalities) affect depression states, and (3) contextual features—these show sequence-level semantics learned from users' social media postings to reflect different contexts and meanings of each word within sentences by considering the sequence of all words. On the basis of these features, we proceeded to perform a social interaction analysis to understand how the interactions between individuals with and without depression influence each other within their user-friendship network. We considered both the social network structure and users' psychological and contextual features for extracting social interaction influence features, where the structure of the user-friendship network shows the strength of connections among users in the social media environment. Then, we applied ML models to effectively manage high-volume and high-speed user-generated data to analyze the depression states of users within social media environments.

Data Cleaning and Preparation

We illustrated and evaluated our proposed framework using datasets from the study by Kosinski et al [40], which contain 197,230 Facebook (Meta Platforms, Inc) status updates from 1047 users, along with their associated Center for Epidemiologic Studies Depression (CES-D) Scale questionnaire scores. These datasets also include user postings, network connections, and personality responses. Note that datasets from other general social networks or online discussion forums can also be used. To minimize the privacy impact regarding using text data, we followed privacy-preserving procedures [41] to ensure anonymity throughout our analysis.

We obtained status updates from all users, regardless of their demographics, as our focus was on a general observation of social network users. Each user's status updates were then aggregated based on their anonymous IDs. We cleaned and preprocessed the data to ensure consistent content. Text preprocessing was conducted in Python (Python Software Foundation), involving correcting spelling errors with SpellChecker (package in Python), converting text to lowercase, removing special characters and numbers, and expanding contractions. We used the NLTK package (Natural Language Toolkit, a Python library for NLP) to remove stop words such as "the," "is," and "our," which are the most common words that appear in a sentence and do not contribute to its meaning. Thus, we removed stop words without sacrificing the meaning of the sentence. Our analysis assesses users' personality features based on the Big Five Factor Model using their responses to the International Personality Item Pool (IPIP) [42]. We determined users' depression levels by obtaining their responses to the CES-D Scale questionnaire, which is a commonly used approach to diagnose depression [43]. It is a 20-item screening questionnaire that captures users' responses to questions related to depressive states. These questions assess the frequency of feelings and behaviors over the past week, including loneliness, sadness, and restless sleep. The higher the score, the more likely an individual is to experience depression. We adopted a similar threshold of 25, as suggested by Lewinsohn et al [44] and Julian et al [45], to label users with depression (total CES-D score >25) and without depression.

We begin by extracting affective, personality, and contextual features from status updates, followed by deriving social interaction influence features based on social network connections. The subsequent sections detail the procedure to identify psychological (eg, affective and personality) and contextual features. These features provide insights into how emotional states, personality traits, and contextual factors interact with social influences. By analyzing these features, we can better understand the dynamics of social interactions and their broader implications for applications such as behavior modeling and targeted interventions.

Discover Affective Features

This section outlines how we extracted affective features. Emotions are essential in diagnosing mental illnesses [46]. Several studies show that depressive symptoms are more related to negative affect, while individuals without depression are more related to positive affect [23,24]. In our study, we extracted affective features to reveal individuals' emotional states and understand users' moods, sentiments, and emotional factors beyond simple linguistic features. Multimedia Appendix 1 shows the process of extracting emotion intensity features and emotion word features to represent the affective features. Because different words can convey affect to various degrees (eg, intensities), we used the NRC Emotion Intensity Lexicon [47] to extract the emotion words from users' social media postings. The lexicon includes >10,000 words related to emotions, including surprise, anger, joy, fear, trust, sadness, anticipation, and disgust, with different intensities. Both common English words and prominent words in social media are covered. For each word t, the intensity of a particular emotion e falls within the range of 0 to 1. A value of "1" indicates the highest intensity of a particular emotion e conveyed by the word t, while "0" indicates the lowest intensity of emotion e for the word t, where $e \in$ {fear, sadness, anger, joy, anticipation, trust, disgust, and surprise}. Thus, each emotion word has an emotion intensity vector $v(t)$ showing the intensities of 8 emotion dimensions. We first calculated how often each emotion word appears within individuals' social media postings to obtain an affective feature vector from a user's posting. Then, for each emotion word t, we multiplied the emotion intensity vector $v(t)$ with the frequency. After that, we computed the average intensity for all emotion words to obtain the affective feature vector.

For example, a user uses 3 emotion words "threat," "illness," and "sunny" in social media postings with a frequency of $freq(threat)=2$, and $freq(illness)=freq(sunny)=1$, respectively. The emotion intensity vector of each emotion word is $v(threat)=<0.742, 0.604, 0, 0, 0, 0, 0, 0>$, $v(illness)=<0, 0.609, 0.688, 0, 0, 0, 0, 0>$, and $v(sunny)=<0, 0, 0, 0, 0.562, 0.5, 0>$. For each emotion word, we multiplied the emotion intensity vector with the frequency of the word and then computed the average emotion intensity for all emotion words. The average of all emotion word intensity for this user is <0.495, 0.605, 0.229, 0,

XSL•FO
RenderX

0, 0.187, 0.167, 0 >. We repeated the process for each user to obtain the user-level average emotion intensity to obtain affective features for all users.

Research shows that the occurrence and distribution of particular words within a document can represent its main content [48]. This suggests that the occurrence of a particular emotion word appearing in social media postings can indicate individuals' emotional state, sentiment, and mood. Thus, we extracted emotion words from individual with and without depression to illustrate their respective emotional states and make a comparison between them. We divided the status updates into 2 datasets: one containing only the status updates of the individuals with depression and the other containing only the status updates of individuals without depression. We then computed the term frequencies of each emotion word based on these datasets from individuals with and without depression:

where $c_g(t)$ indicates the frequency of emotion words t in group g ($g=d$ for the depressed group and $g=nd$ for the nondepressed group).

Then, we find df (t), the ratio differences in relative term frequency for each emotion word t that appears in the datasets from individuals with and without depression as follows:

$$df(t) = tf_d(t) - tf_{nd}(t) \quad \textbf{(2)}$$

We then sorted the ratio differences of all emotion words to find the most distinct emotion words across the depressed and nondepressed groups. We identify the top-k emotion words as distinct features of the depressed group and the bottom-k emotion words as distinct features of the nondepressed group, which help differentiate between the 2 groups.

Figure 2 illustrates an example of the top 10 and bottom 10 emotions words. The emotion words such as "like," "love," "life," "ill," and "heart" have the most positive disparities in term frequencies between the depressed group and the nondepressed group. These words indicate the depressed group's distinct emotion features. In contrast, "tomorrow," "weekend," "morning," "home," and "jeopardy" have the most negative differences, indicating the distinct emotion features of the nondepressed group. The distinct emotion features discovered through term frequency computation highlighted the contrasting emotional states between the depressed and nondepressed groups. These insights serve as invaluable resources for clinically intelligent decision-making, enhancing diagnostic accuracy and offering support tailored to the unique emotional needs of individuals, which, in turn, promotes more effective therapeutic outcomes and enhances overall well-being.

Figure 2. The top 10 distinct emotion words (positive) were associated with the group with depression, and the bottom 10 distinct emotion words (negative) were associated with the group without depression.

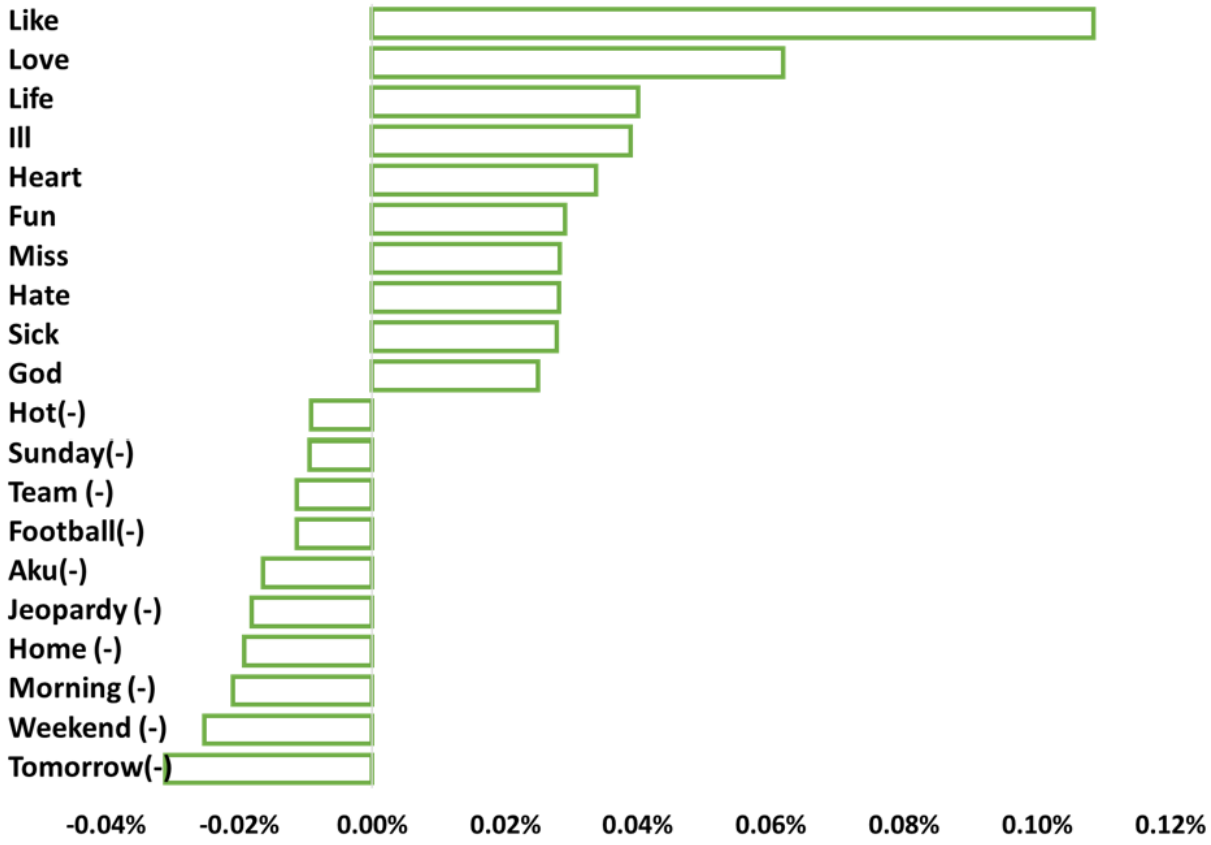

XSL•FO
RenderX

Once we identify the most distinct emotion features, we use term frequency and inverse document frequency (TFIDF) to indicate the importance of a word *t* to a document D (eg, users' aggregated status updates) in a corpus [49]. Each user will have 1 feature vector with all emotion words based on the TFIDF (t, D) value:

$$TFIDF\ (t, D) = TF\ (t, D) \times IDF\ (t) \quad \textbf{(3)}$$

where TF (t, D) stands for the term frequency of word *t* in a specific document *D*. The higher TF *(t, D)*, the more important the word *t* is for that document D. IDF (t) stands for inverse document frequency and shows how significant term *t* is in the entire collection of documents, with rarer words having higher IDF values.

Discover Personality Features

This section outlines how we extracted personality features. Personality shows the patterns of individuals' thoughts, feelings, and actions as well as their attitudes and reactions to external social environments. Numerous studies have shown a link between personality and depressive symptoms [28,29,50]. Therefore, many studies adopt personality for diagnosing depression [32,51].

Our study uses the IPIP questionnaire [42] to measure individuals' personalities based on 5 dimensions outlined in the Big Five Factor Model [48]. This model is the most commonly accepted personality theory in various research studies and clinical practice, offering a comprehensive approach to characterizing an individual's unique characteristics and patterns of behaviors. The model shows how people think, feel, and behave with 5 dimensions: extraversion, agreeableness, conscientiousness, neuroticism, and openness [52,53]. Preoţiuc-Pietro et al [54] used language-derived and user-level "big five" personality factor dimensions to predict the depression state of a user based on social media postings. The personality features show strong performance in differentiating between users with depression and without depression. Yang et al [55] also used personality indicators for depression detection, affirming their strong correlation with traits such as neuroticism, extraversion, and conscientiousness. The IPIP questionnaire includes 50 assessment questions designed to assess diverse individual differences. Participants indicate the extent of truthfulness regarding themselves on a scale from 1=disagree to 5=agree. Higher scores on a personality dimension indicate greater alignment with the respective trait. Multimedia Appendix 2 shows a detailed breakdown of each trait. Five numerical features are derived through responses to the IPIP questionnaire, each corresponding to a participant's personality trait.

Discover Contextual Features

This section shows how we extracted contextual features (contextual meaning of words derived from the words' surroundings in a sentence), which is another critical component of user-generated content. To extract contextual-level language features to reflect different contexts and meanings of words in sentences, we considered the sequence of words in users' social media postings to learn sequence-level semantics. Given that social media posts are often written in a general format, characterized by informal or everyday language, pretrained models—typically trained on general datasets—are particularly effective in capturing these patterns. Their strong contextual modeling capabilities have made pretrained language models well suited for analyzing such data, including the detection of different mental health conditions [51,56,57]. Thus, we take advantage of them to extract contextual-level language features from user generated content. Moreover, considering contextual information, we conducted cluster analysis to visualize specific words used by users with depression and without depression for further analysis by clinicians.

A state-of-the-art model such as Bidirectional Encoder Representations from Transformers (BERT) [58] is a pretrained language model designed to learn deep bidirectional representations of natural language from extensive unsupervised text corpora. BERT has the ability to capture information from both the preceding and following words, enabling it to grasp a more comprehensive understanding of the language context. BERT generates a representation for each word, called word embedding, which can be used for further ML tasks such as classification or clustering. However, to perform ML tasks such as classifying sentences, BERT cannot directly generate independent sentence embeddings (eg, representation for sentences). Thus, we adopted Sentence BERT (SBERT) [59], which effectively represents texts with >1 term. The SBERT model was fine-tuned and pretrained on large datasets [60,61]. Input sentences can be transformed into fixed-sized vectors (768 dimensions), serving as sentence representations. Moreover, sentences that are semantically similar will be close to each other; that is, if we use similarity measures such as Euclidean distance measure to compute the differences between 2 sentences, the more similar the 2 sentences are, the shorter the distance between them.

Multimedia Appendix 3 outlines the proposed procedure to discover contextual features. We first aggregated each user's social media postings. Then, we applied SBERT for each user's aggregated posting [59] to learn sentence embeddings based on each word's contextual information in a sentence to obtain user-level contextual features.

User-level contextual features alone may not offer interpretable information for aiding clinicians in making informed decisions. Hence, based on the sentence embedding, we conducted a cluster analysis to examine the content discussed by different user groups. First, we split the sentence embeddings (eg, user-level contextual features) into 2 datasets (depressed and nondepressed). Next, we applied K-means clustering separately to the datasets of user groups with depression and without depression to cluster sentence embeddings from each group. This helps differentiate the contents or topics discussed by different user groups.

To accomplish this, K-means clustering is conducted across a range of cluster sizes, computing the corresponding silhouette scores—a well-known measure indicating how effectively the resulting clusters separate from each other. The best model is chosen based on the highest silhouette score, with the associated number of clusters representing the optimal cluster count for K-means clustering on the dataset.

XSL•FO
RenderX

For each user group, users belonging to the same cluster are grouped together to obtain the aggregated social media postings. Next, we identified the main topics that user groups with depression and without depression talk about within each cluster. This is achieved by extracting the top-*k* words from each cluster based on term frequency, indicating how often a term appears in a specific cluster and group. Subsequently, we analyzed the words and main topics frequently discussed by different user groups.

Discover Social Interaction Influence Features

This section shows how we extracted social interaction influence features. Studies show that social interactions within social groups can directly or indirectly impact an individual's mental state [33]. Online social interactions are considered as intentional social actions that individuals engage in within the context of participation [62]. These interactions include users expressing their feelings and emotions by posting status updates and comments to share conventions, norms, or languages with positive or negative moods in online social groups. Research indicates that social interaction can facilitate the transfer of negative moods between individuals [34]. Thus, individuals' psychological patterns and the specific words and sentences they use in their online postings in different contexts may affect how friends or social connections in their online social groups react or express themselves on the internet. We conducted social interaction analysis to extract the social interaction influence features for each user in the social network and examine the impact from the psychological and contextual perspective of other users in social groups through social connections. Social interaction influences features that consider users' psychological and contextual features and network structure from social connections.

Multimedia Appendix 4 shows procedures for obtaining social interaction influence features. For each user, we obtained feature vectors from psychological and contextual features described in the Discover Affective Features, Discover Personality Features, and Discover Contextual Features sections. For each type of feature *f* (eg, emotion, personality, or context features), we have 2 feature vectors, $FV^f_{U,d}$, and $FV^f_{U,nd}$ for each user *U* where $g \in \{d, nd\}$. We then found the social connections of each pair of users via the user-friendship network that shows if one user is connected with the other. If user *U* is a friend of another user, *Y*, they are socially connected. Next, we examined the influence of users with depression and without depression on the target user *U*, respectively, based on the different features to obtain emotion, personality, and context influence features. We argue that the emotions, personality, and context information of users with depression and without depression on social media affect how other users express themselves through immediate and nonimmediate social connections and interactions. Therefore, if one user is more influenced by another, their features will be more similar. Moreover, the target user is more likely to be influenced by their close neighbors in a social network regarding distance. Thus, a more immediate neighbor influences the target user more than a faraway neighbor in the social network. Hence, we extract distance-weighted social interaction influence features from users with depression and without depression by considering both network structure (eg, distances between users) and different features (eg, emotions, personality, and contextual features) simultaneously to reflect the neighbors' impact on the target user. We first run the shortest paths for the whole social network between each pair of users. For target user *U*, we obtained the distance measures of the target user and the neighbor user *Y* as *Distance* (*U, Y)*. Then, we assigned weight to each neighbor user *Y*.

Then, the influence of depressed and nondepressed neighbors on the target user was computed using the weight and feature vectors. Consequently, we derived 2 separate social interaction influence features for the target user to represent the degree of similarities between the target user and their depressed and nondepressed neighbors. The same procedure was applied to all social network users.

We illustrate how the analysis works with an example for a particular user, Jenny. Figure 3 shows a partial user-friendship network Jenny. She has a direct connection to Rachel (immediate neighbor) and indirect connections to Paul, Helen, Daniel (second-level neighbors), and Linda (third-level neighbor) in her user-friendship network. Because the shortest path between Jenny and Rachel (first-level neighbor) is 1, we consider their distance to be 1 (ie, *Distance* [*Jenny, Rachel*]=1). Because the shortest path between Jenny and her second-level neighbors (ie, Paul, Helen, and Daniel) is 2, we consider the distance as 2. Likewise, the distance between Jenny and her third-tier neighbor (ie, Linda) is defined as 3, and so forth. Because we argue that the target user is more likely to be influenced by their immediate neighbors in a social network, we can infer that the first-level neighbor, Rachel, has a greater potential to impact Jenny than the second-level neighbors, Paul, Helen, and Daniel. The second-level neighbors have a stronger influence over Jenny than the third-level neighbor, Linda. Consequently, we allocated a greater influence on the first-level neighbor than the second-level neighbor to reflect the higher impact first-level neighbors have on the target user. Similarly, we assigned a higher weight to second-level neighbors than third-level neighbors. For each neighbor *Y* of user *U*, we assigned scores based on the following equation:

XSL•FO
RenderX

Figure 3. Partial user-friendship network of Jenny showing a direct connection to Rachel (immediate neighbor) and indirect connections to Paul, Helen, Daniel (second-level neighbors), and Linda (third-level neighbor) in her user-friendship network.

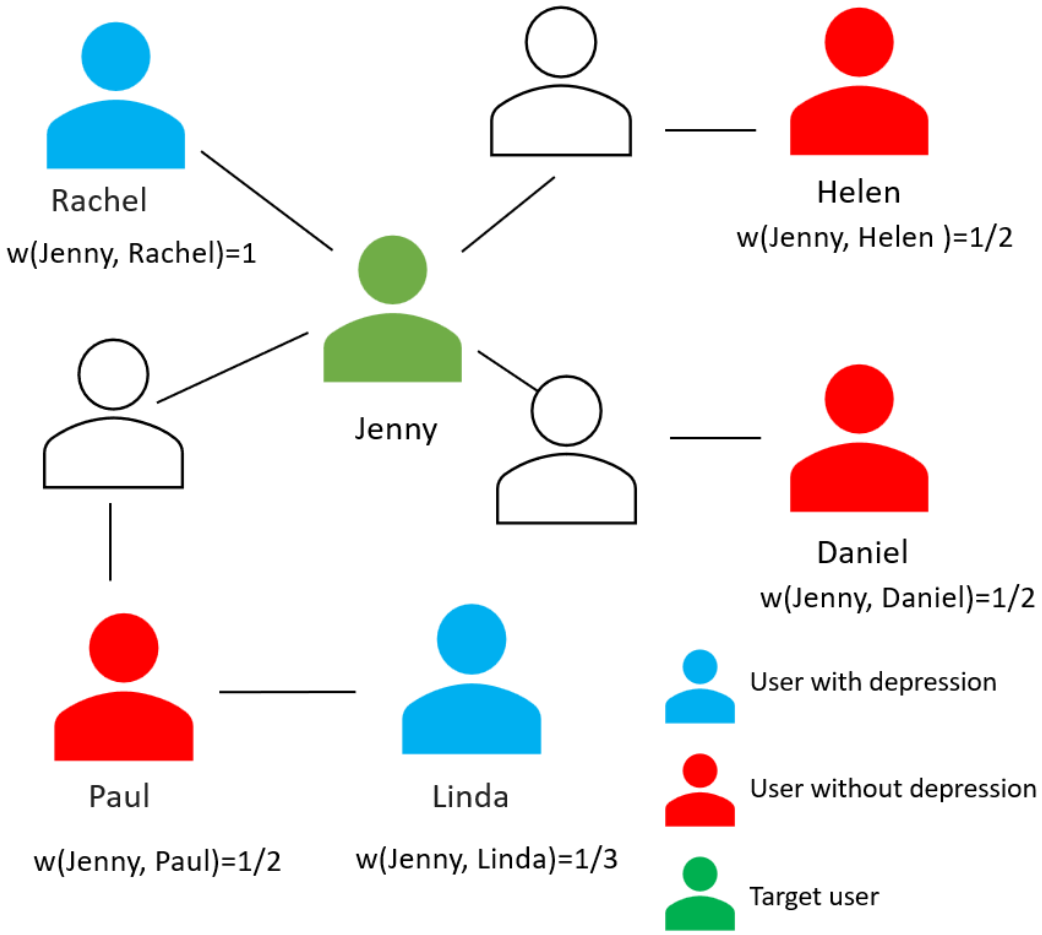

Thus, we assigned a weight of 1 to Rachel; a weight of 1/2 to Paul, Helen, and Daniel; and a weight of 1/3 to Linda to capture the decreasing influence with increasing distance between users.

In our analysis, we considered that the level of influence one user has on another is correlated with the similarity of features between the 2 users. Thus, we used Euclidean distance to assess the similarities of user *U* with their depressed and nondepressed neighbors separately. These similarities are determined by evaluating the depressed and nondepressed social media content of user *U* and their neighbor *Y* for each type of feature *f* (eg, emotions, personality, and contextual features). The similarity is calculated as the absolute difference between the feature vectors:

$Sim\ (U, Y)^f$ is calculated for all users in a social network with confirmed depression states. Then, we combined similarity measures with weights to obtain the social interaction score:

where α_Y is an indicator and α_Y=1 when the neighbor belongs to user group *g*. The higher the values of $SIC_g(U, Y)$, the more impact the specific neighbor user group (depressed or nondepressed) has on the target user in the social network. Multimedia Appendix 5 shows the pseudocode of the proposed social interaction analysis.

Ethical Considerations

The data used in this study were originally collected by researchers at Johns Hopkins University and received approval from their institutional review board. Permission to use the data solely for research purposes was obtained under a confidentiality agreement. This paper details the privacy-preserving procedures used to maintain participant anonymity in the Data Cleaning and Preparation section. Because this study only reanalyzed secondary, anonymized data, additional ethics approval for this study was not required. The original data host has obtained informed consent from all participants who voluntarily donated their data for research purposes, permitting reuse for academic research. The dataset used for this secondary analysis consisted of fully anonymized data and was restricted to noncommercial academic research purposes. This secondary analysis did not involve any direct participant interaction or additional compensation. There is no identification of individual participants or users in any images or supplementary material included in the manuscript. The dataset was deidentified, ensuring no identifiable information was presented.

XSL•FO
RenderX

Results

Overview

In this section, we follow the feature discovery procedure discussed in the Methods section and analyze the proposed features. We then conduct experiments to evaluate the proposed framework based on the datasets [40], which include Facebook users' status updates, network connections, users' associated depression scores measured by CES-D questionnaires, and personality measured by the IPIP personality questionnaire. We demonstrate how we extract psychological and contextual features as well as the resultant social interaction influence features. We explore the understandability of the developed features in explaining and understanding users' sentiments, moods, emotional behaviors, psychological states, and influence from social interactions in social networks. Our framework can be extended to include any social media and social platforms to enhance the capability of mental health professionals and clinicians to make explainable and timely diagnoses, referrals, interventions, and treatments for their potential patients.

Analysis of Affective Features

Given the procedure outlined in the Discover Affective Features section, we first obtained user-level emotion intensity by averaging the emotion intensities of the 8 emotion dimensions in users' social media postings. We further explored emotions, sentiments, and moods across different user groups by examining the top 30 and bottom 30 emotion words as distinct features, which were extracted from the depressed and nondepressed groups, respectively. Figure 4 illustrates the frequency distributions for each user group. Words with higher frequency values were more distinct emotion terms in each user group. The user group with depression frequently used emotion words with negative sentiments, such as "sick," "hurt," "bad," "alone," "ill," "lost," and "hate," and some swear words, which align with those reported in previous studies [63]. In contrast, users without depression expressed positive sentiments, including "good," "wonderful," "incredible," and "lovely." Words such as "family," "friends," and "daughter" were frequently mentioned in status updates by individuals in the user group with depression, indicating specific individuals or individuals that are related to them. In contrast, events such as "exams" and "vacation"; specific food, drinks, or items related to events including "beer," "wine," "football," "chocolate," and "hot"; and outcomes of events such as "win," "break," and "jeopardy" were discussed by the user group without depression.

Figure 4. Emotion frequency distributions of the user groups with depression and without depression. The top 30 distinct emotion words (positive) are associated with the depressed group, and the top 30 distinct emotion words (negative) are associated with the group without depression.

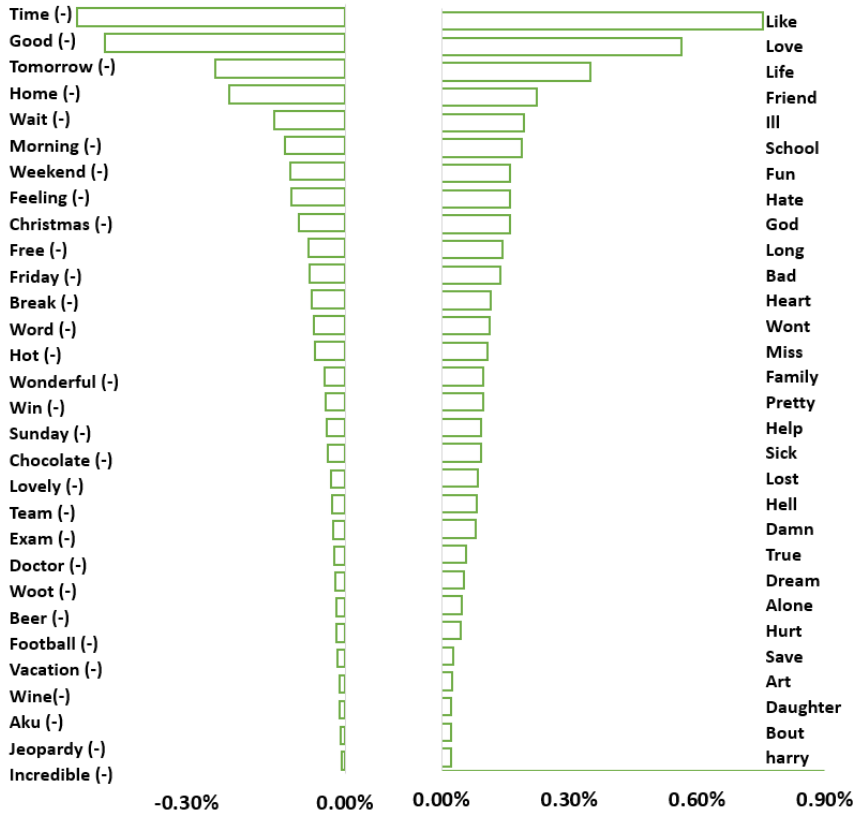

From Figure 4, we observe that the user group without depression is more likely to use emotion words associated with time, such as "time," "tomorrow," "morning," "weekend," "friday," "christmas," and "sunday." In contrast, the user group with depression expresses feelings by using words such as "like," "love," and "fun." They also care about more internal things such as "god," "life," "hear," and "dreams."

We calculated TFIDF (t, D) for each emotion word *t* in each user's social media postings *D* to generate the feature vectors

XSL•FO
RenderX

based on the most distinct depressed and nondepressed words as inputs for classifying the users with and without depression.

We conducted further analysis to examine the emotion intensities for each user group separately. To achieve this, we calculated the average intensity of each word in the emotion dimensions for the top 30 words in the depressed user group and the bottom 30 words in the nondepressed user group. The set of distinct emotion words used by user group g is $\{T_{g1}, T_{g2}, ..., T_{g30}\}$, where $g \in \{d, nd\}$. For each emotion word $t_g \in \{T_{g1}, T_{g2}, ..., T_{g30}\}$ in group g, we denoted the total count of word t_g's appearance in all users' social media postings in group g by TC_{tg} and the total number of words for all users as TC_U. Then, the average intensity of in group g is computed as: Sum of:

$$(TC_{tg} * V(t_g))/TC_U \quad \textbf{(7)}$$

We compared the average intensities of the most distinct emotions of these 2 user groups, as shown in Figure 5. Distinct depressed emotion words have high intensities in all emotion categories, suggesting that the user group with depression tends to have more emotional behavioral changes than the user group without depression. The user group with depression shows significant intensities in sadness, anger, fear, disgust, and trust compared to the user group without depression, suggesting that negative emotions are frequently expressed through social media postings by the user group with depression. The user group with depression also shows a higher intensity of joy than the user group without depression. This suggests that users with depression may try to use words related to joy to cheer themselves up, or they are seeking joy.

Figure 5. Comparison of average emotion intensities of the most distinct emotions of user groups with depression and without depression.

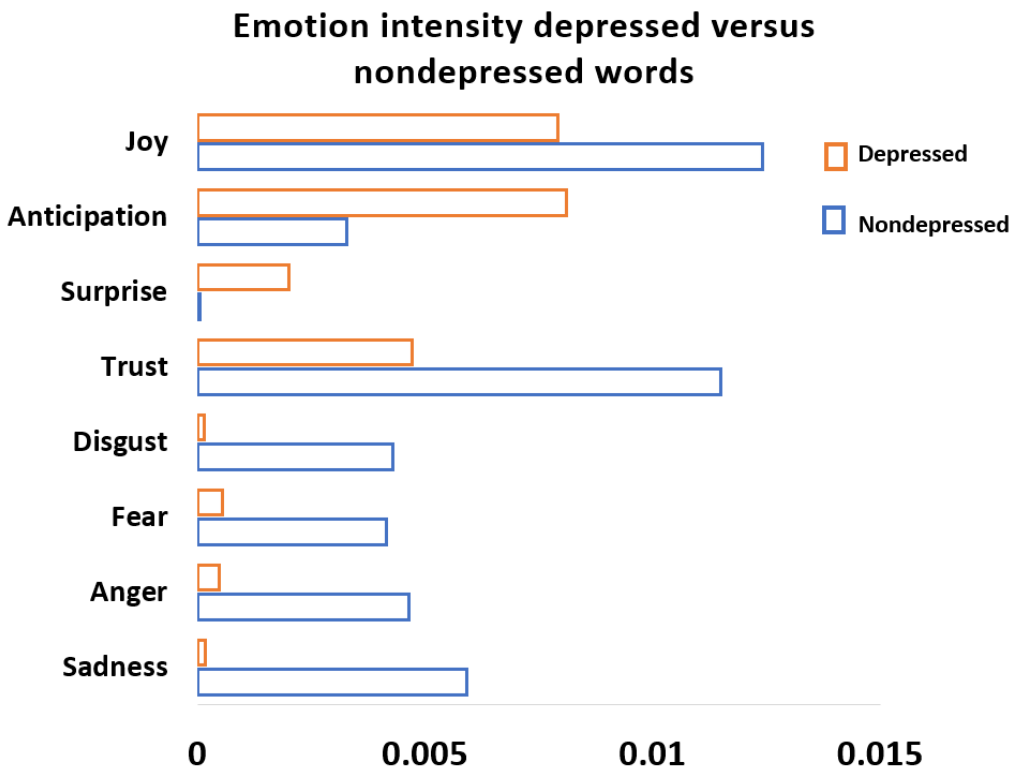

Analysis of Personality Features

As discussed in the Discover Personality Features section, we used the IPIP questionnaire to assess individuals on 5 dimensions of personality according to the Big Five Factor Model. For each user, we obtained a score for each of the 5 personality dimensions to form a feature vector for the personality feature of that user. Table 1 provides an example illustrating how we obtain the feature vector for each user. This personality feature vector comprises 5 values, each representing 1 of the 5 personality dimensions. The score for each dimension ranges from 1 to 5. For example, as shown in Table 1, the first user, U_1, scores 4 for openness, 2.25 for conscientiousness, 2.2 for extraversion, 3.6 for agreeableness, and 2.8 for neuroticism. Thus, U_1's feature vector for personality is <4, 2.25, 2.2, 3.6, 2.8 >. Similarly, we obtained the feature vector for U_2 as <3.8, 3.75, 3.4, 4.4, 2.15> and for U_3 as <3.55, 3.2, 3.2, 3.35, 2.9>, and so on. The feature vector for personality will be used to run experiments for depression detection.

XSL•FO
RenderX

Table 1. Examples of personality feature vectors for 3 users across 5 personality traits.

User	The feature vector <openness, conscientiousness, extraversion, agreeableness, neuroticism> is a numerical representation of personality traits based on the Big Five model used to describe and analyze individual characteristics in a structured format.
U_1[a]	<4, 2.25, 2.2, 3.6, 2.8>
U_2	<3.8, 3.75, 3.4, 4.4, 2.15>
U_3	<3.55, 3.2, 3.2, 3.35, 2.9>

[a]U_1 represents first user, U2 represent second user, and so on.

Analysis of Contextual Features

As outlined in the Discover Contextual Features section, we first used SBERT [59] to learn sentence embeddings based on each word's contextual information in a sentence for each user's social media posting to obtain user-level contextual features. The sentences users discuss in their social media postings are transformed into fixed-sized vectors (768 dimensions), which are the representation of the sentences. We conducted cluster analysis based on the SBERT feature vectors to obtain more valuable insights to support clinical decisions. We conducted K-means clustering separately for the user groups with and without depression to analyze different contents or topics different user groups discuss on social media. We found the optimal number of clusters for each user group based on the "silhouette" method. We chose 2 and 4 to be the optimal number of clusters for the user groups with and without depression, respectively, by experimenting with silhouette scores with regard to a range of clusters (ie, we choose the number of clusters with the highest silhouette score as the optimal number).

We then grouped users belonging to the same cluster together and obtained the aggregated social media postings for each cluster. Because the nondepressed group had fewer clusters than the depressed group, we obtained the top 20 words from each depressed cluster based on the term frequency and the top 40 words for each nondepressed cluster to have an equal number of top words from each user group for comparison.

The top words discussed in each cluster for each user group are shown in Multimedia Appendices 6 and 7.

On the basis of the top words, we conducted further analysis. Swear words and negative sentiments, such as "ass," "f**king," and "tear," were unique to the user group with depression, as shown in Textbox 1. This group also uses words related to family members, such as "dad" and "daughter," and spiritual words, such as "pray," "bless," and "trust." This may suggest that the user group with depression tends to seek comfort from family or the spiritual world. They also talk about school-related terms such as "teacher" and "college." This might suggest that the user group with depression has depressed symptoms due to pressure from school. The non-depressed group typically uses fewer negative sentiment words. The user group without depression uniquely mentioned words related to being forward-looking or future, such as "future," "forward," "chance," and "step"; optimistic actions, such as "pass," "win," and "worth"; or positive words, such as "power" and "perfect." They also used words suggesting the beginning of something, such as "awake" and "begin."

Textbox 1. Comparison of unique words used by users with and without depression.

Unique words used by users with depression

- "Awake", "begin", "burn", "chance", "comment", "cut", "exam", "Facebook", "fear", "final", "forward", "future", "human", "pass", "paste", "perfect", "point", "power", "problem", "self", "side", "step", "thought", "truth", "weather", "win", and "worth".

Unique words used by users without depression

- "Ass", "band", "beat", "bless", "dad", "daughter", "chocolate", "college", "easy", "fly", "fxxking", "kind", "omg", "pray", "repost", "seriously", "tear", "teacher", "text", and "trust".

Analysis of Social Interaction Influence Features

Next, we used the proposed psychological and contextual features, along with the social network structure derived from social connections, to extract social interaction influence features. We created social interaction influence features for each user and each type of proposed feature to show the extent of influence exerted by user groups with and without depression on the target user. As shown in Table 2, the example illustrates the affective, personality, and contextual influence of neighbors with and without depression, respectively, on 4 users, U_1 (user 1), U_2 (user 2), U_3 (user 3), and U_4.(user 4) We first extracted each user's affective, personality, and contextual features. For each user, U_i, and each feature, f, we obtained feature vectors $FV^f_{U,d}$ and $FV^f_{U,nd}$ that represent the features from user U_i 's social connections with and without depression, respectively.

XSL•FO
RenderX

Table 2. Examples of affective, personality, and contextual influences on four users, stemming from interactions with neighbors with depression and without depression, respectively. Each value represents the proportion or impact (as a fraction of 1) of the specific influence type (affective, personality, contextual) on the given user from neighbors with or without depression.

User	Feature type					
	Affective		Personality		Contextual	
	Depressed	Nondepressed	Depressed	Nondepressed	Depressed	Nondepressed
U_1[a]	0.736	0.264	0.677	0.323	0.668	0.332
U_2	0.566	0.434	0.514	0.486	0.396	0.604
U_3	0.282	0.718	0.405	0.595	0.464	0.536
U_4	0.206	0.794	0.191	0.809	0.265	0.735

[a]U_1 represent first user, U2 represent second user, and so on.

These feature vectors are inputs for extracting the social interaction influence features. In the case of affective features, the scores obtained from the users' group with depression are higher than those from the users' group without depression for U_1 and U_2. This suggests a stronger social interaction influence from the user with depression compared to the user without depression for U_1 and U_2. As a result, these 2 users exhibited emotions that are closer to those of their friends with depression and further from those of their friends without depression. In contrast, U_3 and U_4 exhibited greater similarity to their friends without depression compared to their friends with depression, suggesting they experienced a stronger influence from the friends without depression. Thus, $SIC_d(U,Y)$ affective is larger than $SIC_{nd}(U,Y)$ affective for user U_1 and U_2 for emotion features. Likewise, for personality and contextual features, higher scores suggest a higher influence. By comparing the scores between the user groups with depression and without depression, a higher score for a feature in 1 group suggests that the users are more aligned with that particular user group regarding that feature. This observation suggests that within the social networks, the user groups with depression and without depression may have different effects on their friends. Users who are more influenced by the user group without depression tend to display a higher degree of resemblance in affective, personality, or contextual information with the user group without depression, and vice versa.

Experiment Results: Depression Prediction Using ML Methods

The proposed psychological and contextual analysis NLP framework has the potential to enhance intelligent decision-making in the mental health domain. Our framework can enhance the reliability and knowledge base of decision support systems for health care providers, facilitating more accurate diagnoses and timely referrals and interventions, while also simplifying treatment plans. We found that incorporating affective, contextual, and personality features outperformed traditional baselines (improving performance compared to traditional baselines with an average improvement of 6% in accuracy and 10% in F_1-score) and achieved performance comparable to state-of-the-art baselines. Furthermore, the addition of social interaction influence enhanced performance across all metrics.

We conducted experiments using ML methods with the proposed features for depression detection to demonstrate the performance of our framework. All ML implementations were conducted using Python's *scikit-learn* library. The experiments were conducted on a ThinkPad (X1 Yoga Gen 8 Intel) with an Intel Core i7 CPU, 16 GB of RAM, running Windows 10. We used 5-fold cross-validation for model evaluation. Four ML methods were used: (1) logistic regression, (2) neural network, (3) k-nearest neighbors (KNN), and (4) random forest. We used an 80/20 split for training and testing the models. Specifically, we reported the accuracy of depression detection as part of our evaluation:

where TP=true positives, TN=true negatives, FP=false positives, and FN=false negatives.

In addition, we reported recall and precision scores. The recall represents the ratio of correctly predicted users with depression among all actual users with depression, while precision measures the proportion of users with depression correctly identified as depressed relative to the total number of users predicted as depressed, that is:

We also reported F_1-score, which is the combination of precision and recall:

To assess the predictive capability of our proposed framework, we leveraged psychological (eg, affective and personality), contextual, and social interaction influence features to detect depression. We used the Linguistic Inquiry and Word Count (LIWC) [64] feature as the baseline, which is easily replicable to provide a comparative benchmark with the same data. The LIWC feature is widely used in depression detection because it offers reliable, unbiased, and effective measures [65-68].

We first conducted experiments using psychological (eg, affective and personality) and contextual features. The performance of various ML methods for different sets of features is illustrated in Figure 6. KNN showed the highest accuracy among all other ML models, with 69% for the LIWC feature,

XSL•FO
RenderX

76.6% for the affective features, 73% for contextual features, and 79% for the personality features. For precision, although neural network showed a higher performance of 67% compared to 66% under KNN for the affective features, KNN showed better performance compared to other ML methods for all other features. KNN also indicated the highest recall and F_1-score compared to other ML methods among all feature sets. Overall, KNN performed the best among all ML methods used in this study. Therefore, we conducted an additional set of experiments with different feature sets (F1-F9 correspond to different feature sets) listed in Table 3, using KNN, to assess the predictive power of depression detection based on the proposed features for each feature set.

Figure 6. Performance comparison of different machine learning methods across 4 metrics using Linguistic Inquiry and Word Count (LIWC), personality (psychology), contextual, and affective features. KNN: k-nearest neighbors; LR: logistic regression; NN: neural network; RF: random forest.

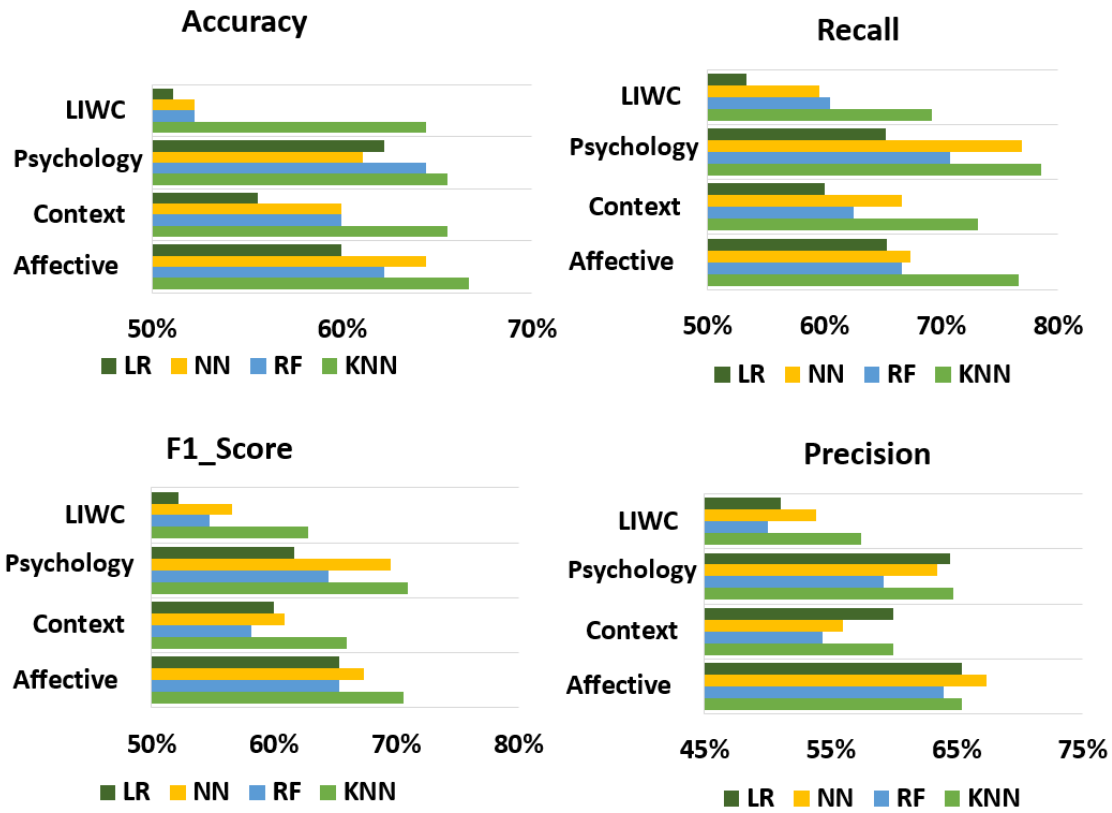

XSL•FO
RenderX

Table 3. Different feature sets using KNNa to evaluate the predictive power of depression detection.

Feature short name	Feature variables
Baseline1	LIWC[b] features
Baseline2	FastText
Baseline3	CNN[c]
F1	Affective features
F2	Contextual features
F3	Personality features
F4	Affective features, contextual features, and personality features
F5	Affective influence features, contextual influence features, and personality influence features
F6	Affective features and affective influence features
F7	Contextual features and contextual influence features
F8	Personality features and personality influence features
F9	Affective features, contextual features, and personality features and affective influence features, contextual influence features, and personality influence features

[a]KNN: k-nearest neighbors.

[b]LIWC: Linguistic Inquiry and Word Count.

[c]CNN: convolutional neural network. F1-F9 correspond to different feature sets listed in Table 3, F1 is feature set 1 which is Affective Features as listed in Table 3, F2 is feature set 2 which is listed as Contextual Features in Table 3 and so on.

We treated the LIWC feature as baseline 1 and compared it against the proposed affective, contextual, personality, and corresponding social interaction influence features (eg, affective influence, contextual influence, and personality influence features) using KNN. We also compared our model with state-of-the-art approaches that used pretrained models for text processing combined with ML techniques for detection, such as FastText with XGBoost (baseline 2, as used in the study by Marriwala and Chaudhary [69]) and deep learning architectures, including convolutional neural networks for text processing and detection (baseline 3, as described in the study by Ghosal and Jain [39]).

As shown in Figure 7, the feature sets incorporating the proposed features significantly improved the detection capability compared to the LIWC feature across all performance measures. Our results are comparable to those achieved by baseline 2 (FastText with XGBoost) and baseline 3 (convolutional neural networks); however, the combination of all affective, personality, and contextual features, along with their influence features, yields even stronger performance. The inclusion of emotional and personality-driven factors, combined with social interaction features within our framework, offers a unique advantage. This approach delivers richer insights and a higher interpretive value compared to previous models, while maintaining high interpretability—a crucial aspect for mental health applications.

XSL•FO
RenderX

Figure 7. Performance comparison across different feature sets and baselines.

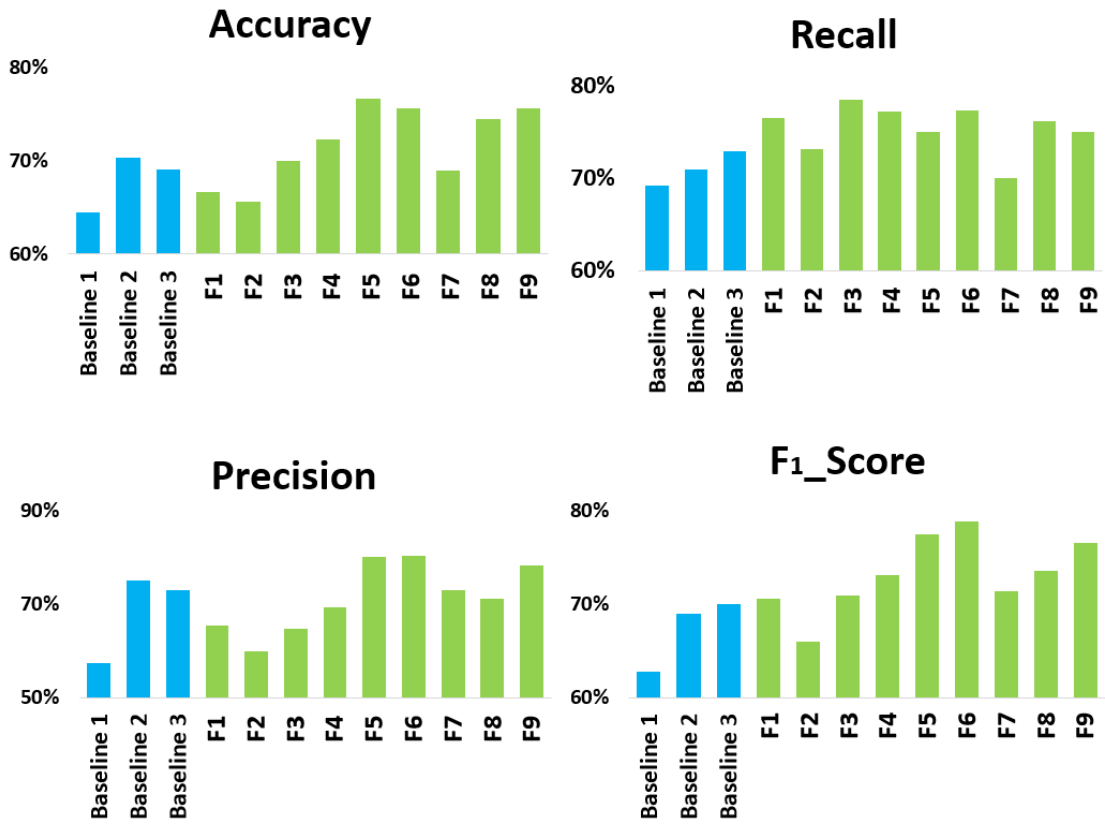

Among the proposed features, the model with affective features (F1) and the model with personality features (F3) performed better than the model with contextual features (F2). The model with affective features, personality features, and contextual features (F4) had the best performance in all measures except recall compared to F1, F2, and F3, while F3 showed better performance in recall compared to F4. The relatively strong performance of F2 showed the model's ability to effectively distinguish users with depression, suggesting its capacity to identify depression-specific language patterns. When considering the social interaction influence features for each feature type, the model with affective influence, personality influence, and contextual influence features (F5) showed higher performance compared to F4 with respect to accuracy, precision, and F_1-score. F3 and F4 showed better recall compared to F5. Compared to affective features (F1), the combination of affective features and the corresponding social interaction influence of affective features—affective influence features (F6)—performed better across all performance metrics. The same pattern applies for F7 with the combination of contextual features and the corresponding contextual influence features and for F8 with the combination of personality features and the corresponding personality influence features; they both performed better compared to the model with contextual features (F2) and personality features (F3) alone across all performance metrics. The model with all proposed features, including affective, personality, contextual, and the corresponding social interaction influence features (F9), showed high performance but not the highest among all models; this implies that simply including all features may not necessarily enhance performance. All proposed features outperformed baselines; notably, F5 and F6 demonstrated the strongest overall performance.

To examine how each feature impacts model performance, we performed permutation importance. This approach provides insights into how each feature affects the model's accuracy by measuring the change in performance when individual features are randomly shuffled [70]. Multimedia Appendix 8 shows the top 10 influential features, highlighting those most impactful for depression detection. The results indicate that personality features such as neuroticism, extraversion, and agreeableness are highly significant, followed by social interaction influence based on emotion, personality, and contextual measures. Finally, the affective feature, anticipation score, is the most influential among all emotions. F1-F9 represent short name of 9 feature sets we used in the experiments

Discussion

Principal Findings

We contribute by proposing a methodological framework using NLP and ML techniques to extract features and patterns, thereby comprehensively understanding depression-related symptoms from psychological and contextual perspectives. Building on the foundations of social influence theory, which points out that the way individuals behave, think, or act is influenced by their social environment, we extracted social interaction influence features based on social media postings. These features considered both psychological and contextual features (eg, users' emotions, psychological states, and context of communication) from users' social media postings and the social network structure of social media. They reflected the influence that a

XSL•FO
RenderX

user's social connections may have on their mental well-being. The combination of all affective, personality, and contextual features, along with their influence features, yields strong performance compared to baselines. F5 (which represent affective influence features, contextual influence features, and personality influence features in Table 3) and F6 (affective features and affective influence features) show particularly strong performances. Our study, designed to complement clinical diagnoses, delivers results that are sufficiently robust to support timely interventions and practical applications in mental health monitoring. By achieving comparable results to state-of-the-art methods, our approach demonstrates strong performance while adding value through richer insights and high interpretability. These patterns bridge the gap between model accuracy and interpretability, providing more actionable and inseparable insights that can aid clinicians in better understanding users' mentality and supporting their decision-making processes. This framework has the potential to expedite the diagnostic process and provide timely referral services to individuals at high risk, facilitating early intervention or treatment before more severe mental health conditions develop. With an improved understanding of users' mental states reflected in the psychological and contextual aspects, clinicians can more easily detect significant changes in users' mental well-being. This allows clinicians to provide well-supported guidance during early intervention and facilitate follow-ups in treatment plans.

Research in psychology has verified the theoretical connections between patients' mental health conditions and certain linguistic features. Various studies establish that individuals experiencing certain mental illnesses exhibit specific verbal behavior. Therefore, there is an increasing trend toward using rich sources of text data and information from social media to examine users' sentiments and behaviors in mental health studies [16,17]. Studies show that frequent negative affect expression has higher connections with depressive states, while positive affect tends to have a lower connection to depression [23-25]. Thus, emotions play an essential role in diagnosing mental illnesses such as depression [11]. Various studies have characterized emotions from language use of users disclosing their mental illnesses on social media to detect depression [68,71,72]. Personality is another important aspect of analyzing depression. Personality has been used to distinguish social media users' activities, posting content, and language use habits [26,27,73]. Several studies show a link between personality types and depression, and personality features can be used in depression diagnosis and treatments [28,50,55].

In addition to psychological factors, pretrained language models have shown strong contextual modeling ability; they are used in understanding texts in various applications [58]. For example, Haque et al [74] leveraged pretrained language models to detect early signs of suicidal thoughts, improve the diagnosis process, and provide people with appropriate treatment on time. Jiang et al [75] used deep contextualized pretrained word representation to build strong predictive models to detect different mental health problems from social media. Murarka et al [37] built a multiclass model using the pretrained language model to classify 5 prominent types of mental illnesses by analyzing unstructured user data on Reddit. Social media plays a critical role in shaping attitudes, beliefs, and behaviors through interactions [76,77]. Social influence theory suggests that individuals' mental states and social interactions affect their language and behaviors on social media [32,78]. While previous research has focused on linguistic, emotional, and personality factors, it often overlooks the impact of social interactions [37,68]. Our study aims to bridge this gap by analyzing social influence and interactions alongside psychological and contextual factors. This comprehensive framework can aid in improving mental health diagnosis and treatment plans.

Our study has several practical implications. By providing mental health professionals with a deeper and more comprehensive insight into a user's mental state, our framework can help accelerate the depression diagnosis process. It also enables timely referral services to individuals who may be unaware of their mental states or have limited access to mental health resources. Moreover, it is beneficial to alert health care management to offer early intervention or treatment to individuals at risk of developing more severe mental health conditions. Furthermore, by gaining a deeper insight into users' behaviors reflected in different aspects, mental health professionals can identify the significant changes in users' emotional, psychological, and mental health states to provide more justified recommendations during the early intervention and facilitate easier follow-ups in treatment plans.

Other domains that analyze social interactions and language use can also adopt our framework. In the general health domain, our framework offers valuable assistance in discovering positive and negative opinions by studying user-generated content from health-based social platforms. These platforms serve as spaces where users, including patients, physicians, and caregivers, can pose questions and seek advice from other users. By analyzing emotions, personality, and contextual information and the influence of social interaction, medical professionals and caregivers can effectively monitor different health conditions among patients and individuals at risk. This allows them to provide timely advice, referral services, or interventions, ultimately enhancing health care quality in these online communities. In the finance domain, for example, StockTwits is an online platform where investors exchange ideas, opinions, and sentiments regarding financial markets [79]. Our framework can be used to examine how emotional shifts, personality differences, and contextual information within sentences influence investors' decision-making processes. It also enables the examination of how social interactions among different investors may influence their investment decisions.

Limitations

While analyzing users' postings on social media enables the identification of mental states through means unavailable offline, our research does not consider the socioeconomic status, idiosyncratic behaviors "behind the scenes," or dynamic language data. For future studies, complementary sources of behavioral data and dynamic data, such as activity logs or search history, could be integrated with health records, such as records of prescriptions or health care claims data, to offer a more comprehensive understanding of the mental health status of an individual. Moreover, the expression of individuals with

XSL•FO
RenderX

depression may vary across different intensity levels of their depression. As a result, investigating how language use differs and evolves across different stages of depression could be a promising future direction, potentially leading to more personalized treatment plans. In addition, we plan to incorporate domain-specific lexicons and more specialized datasets to further enhance the model's accuracy in capturing nuanced depressive language. We also aim to fine-tune SBERT embeddings using mental health–specific corpora, which could improve performance in detecting depression-related language. Furthermore, after comprehensively understanding the patterns and trends of users with depression, a database system can be developed to store significant behaviors or language use changes. Updated classifiers could be trained using real-time data from online platforms or other sources. The system could automatically detect users who are at risk of depression even when they are unaware of their mental conditions and recommend suitable treatments based on information extracted from previous instances. We also plan to incorporate more recent and real-world datasets with similar features, extending the analysis beyond mental health contexts to various domains of social network analysis. This broader approach will facilitate a more comprehensive exploration across diverse scenarios, enhancing the framework's relevance and adaptability.

Conclusions

Our study illustrates a psychological and contextual analysis NLP framework that uses affective, personality, and contextual information. We further explore the influence of social interaction shaped by these elements to monitor depression behaviors on social media using NLP and ML. The proposed framework highlights the potential of using emotion intensities, personality measurement, pretrained language models, and social interactions among users on social media as features or patterns for predicting depression in individuals. Our research shows the potential of developing a better foundation for predicting the risk of depression and offering referrals, early intervention, and treatment plans via the analysis of social media data. It presents promising potential as an intelligent decision-making tool integrating information from multiple channels, such as social networks.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The process of extracting emotion intensity features and emotion word features to represent the affective features. We extracted emotion words from social media posts of users with and without depression and calculated the ratio differences in relative term frequency to identify distinct emotion words. We also combined the intensity levels across 8 emotion dimensions with emotion word frequency, computing the average intensity to form an emotion intensity feature.
[PNG File , 93 KB - formative_v9i1e60286_app1.png]

Multimedia Appendix 2
The Big Five Factor Model—personality traits with descriptions, showing the behavioral tendencies associated with low and high scores for each trait.
[PDF File (Adobe PDF File), 13 KB - formative_v9i1e60286_app2.pdf]

Multimedia Appendix 3
Discovery process of contextual features. We aggregated each user's social media postings and applied Sentence Bidirectional Encoder Representations (SBERT) from Transformers to generate sentence embeddings, capturing the contextual information of each word within a sentence. These embeddings serve as user-level contextual features. Then, we performed cluster analysis using K-means clustering, applied separately to the sentence embeddings of user groups with depression and without depression to analyze the content discussed by different user groups.
[PNG File , 108 KB - formative_v9i1e60286_app3.png]

Multimedia Appendix 4
Discovery process of social interaction influence features. For each user, we obtained feature vectors from the psychological and contextual features identified earlier. We then analyzed the influence of users with depression and without depression on the target user separately, based on the emotion, personality, and context influence features.
[PNG File , 19 KB - formative_v9i1e60286_app4.png]

Multimedia Appendix 5
Algorithm for discovering social interaction influence features.

XSL•FO
RenderX

[PNG File , 104 KB - formative_v9i1e60286_app5.png]

Multimedia Appendix 6
The most frequently discussed words across the four clusters for the user group with depression.
[PNG File , 254 KB - formative_v9i1e60286_app6.png]

Multimedia Appendix 7
The most frequently discussed words across the two clusters for the user group without depression.
[PNG File , 244 KB - formative_v9i1e60286_app7.png]

Multimedia Appendix 8
The top 10 influential features, highlighting the most impactful features for depression detection.
[PNG File , 21 KB - formative_v9i1e60286_app8.png]

References

1. Williams JW, First M. Diagnostic and statistical manual of mental disorders. In: McNutt JG, editor. Encyclopedia of Social Work. Oxford, UK: Oxford University Press; 2013.
2. Zhang T, Schoene AM, Ji S, Ananiadou S. Natural language processing applied to mental illness detection: a narrative review. NPJ Digit Med 2022 Apr 08;5(1):46 [FREE Full text] [doi: 10.1038/s41746-022-00589-7] [Medline: 35396451]
3. Depression. World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/depression [accessed 2023-07-13]
4. Evans-Lacko S, Aguilar-Gaxiola S, Al-Hamzawi A, Alonso J, Benjet C, Bruffaerts R, et al. Socio-economic variations in the mental health treatment gap for people with anxiety, mood, and substance use disorders: results from the WHO World Mental Health (WMH) surveys. Psychol Med 2018 Jul;48(9):1560-1571 [FREE Full text] [doi: 10.1017/S0033291717003336] [Medline: 29173244]
5. Yang K, Zhang T, Ananiadou S. A mental state knowledge–aware and contrastive network for early stress and depression detection on social media. Inf Process Manage 2022 Jul;59(4):102961. [doi: 10.1016/j.ipm.2022.102961]
6. Motrico E, Mateus V, Bina R, Felice E, Bramante A, Kalcev G, et al. Good Practices in perinatal mental health during the COVID-19 pandemic: a report from task-force RISEUP-PPD COVID-19. Clínica y Salud 2020;31(3):155-160. [doi: 10.5093/clysa2020a26]
7. Coronavirus and depression in adults, Great Britain. Office for National Statistics. URL: https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/articles/coronavirusanddepressioninadultsgreatbritain/julytoaugust2021 [accessed 2023-07-29]
8. Sheehan DV. Depression: underdiagnosed, undertreated, underappreciated. Manag Care 2004 Jun;13(6 Suppl Depression):6-8. [Medline: 15293765]
9. Thombs BD, Kwakkenbos L, Levis AW, Benedetti A. Addressing overestimation of the prevalence of depression based on self-report screening questionnaires. CMAJ 2018 Jan 15;190(2):E44-E49 [FREE Full text] [doi: 10.1503/cmaj.170691] [Medline: 29335262]
10. Lin H, Jia J, Nie L, Shen G, Chua T. What does social media say about your stress? In: Proceedings of the 25th International Joint Conference on Artificial Intelligence. 2016 Presented at: IJCAI '16; July 9-15, 2016; New York, NY p. 3775-3781 URL: https://dl.acm.org/doi/abs/10.5555/3061053.3061147
11. Zhang T, Yang K, Ji S, Ananiadou S. Emotion fusion for mental illness detection from social media: a survey. Inf Fusion 2023 Apr;92:231-246. [doi: 10.1016/j.inffus.2022.11.031]
12. Zhang T, Yang K, Alhuzali H, Liu B, Ananiadou S. PHQ-aware depressive symptoms identification with similarity contrastive learning on social media. Inf Process Manag 2023 Sep;60(5):103417. [doi: 10.1016/j.ipm.2023.103417]
13. Issaka B, Aidoo EA, Wood SF, Mohammed F. "Anxiety is not cute" analysis of twitter users' discourses on romanticizing mental illness. BMC Psychiatry 2024 Mar 21;24(1):221 [FREE Full text] [doi: 10.1186/s12888-024-05663-w] [Medline: 38515062]
14. De Choudhury M. Role of social media in tackling challenges in mental health. In: Proceedings of the 2nd international workshop on Socially-aware multimedia. 2013 Presented at: SAM '13; October 21, 2013; Barcelona, Spain p. 49-52 URL: https://dl.acm.org/doi/10.1145/2509916.2509921 [doi: 10.1145/2509916.2509921]
15. Reece AG, Reagan AJ, Lix KL, Dodds PS, Danforth CM, Langer EJ. Forecasting the onset and course of mental illness with Twitter data. Sci Rep 2017 Oct 11;7(1):13006 [FREE Full text] [doi: 10.1038/s41598-017-12961-9] [Medline: 29021528]
16. Chancellor S, De Choudhury M. Methods in predictive techniques for mental health status on social media: a critical review. NPJ Digit Med 2020 Mar 24;3(1):43 [FREE Full text] [doi: 10.1038/s41746-020-0233-7] [Medline: 32219184]
17. Ríssola EA, Losada DE, Crestani F. A survey of computational methods for online mental state assessment on social media. ACM Trans Comput Healthcare 2021 Mar 17;2(2):1-31. [doi: 10.1145/3437259]

XSL•FO
RenderX

18. Amanat A, Rizwan M, Javed AR, Abdelhaq M, Alsaqour R, Pandya S, et al. Deep learning for depression detection from textual data. Electronics 2022 Feb 23;11(5):676. [doi: 10.3390/electronics11050676]
19. Ghosh S, Ekbal A, Bhattacharyya P. A multitask framework to detect depression, sentiment and multi-label emotion from suicide notes. Cogn Comput 2021 Feb 05;14(1):110-129. [doi: 10.1007/S12559-021-09828-7]
20. Yoo M, Lee S, Ha T. Semantic network analysis for understanding user experiences of bipolar and depressive disorders on Reddit. Inf Process Manag 2019 Jul;56(4):1565-1575. [doi: 10.1016/j.ipm.2018.10.001]
21. Aragón ME, López Monroy AP, González-Gurrola L, Montes M. Detecting depression in social media using fine-grained emotions. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019 Presented at: NAACL '19; June 2-7, 2019; Minneapolis, Minnesota p. 1481-1486 URL: https://aclanthology.org/N19-1151.pdf [doi: 10.18653/v1/n19-1151]
22. Shickel B, Siegel S, Heesacker M, Benton S, Rashidi P. Automatic detection and classification of cognitive distortions in mental health text. In: Proceedings of the 20th International Conference on Bioinformatics and Bioengineering. 2020 Presented at: BIBE '20; October 26-28, 2020; Cincinnati, OH p. 21-55 URL: https://ieeexplore.ieee.org/document/9288097 [doi: 10.1109/bibe50027.2020.00052]
23. Settanni M, Marengo D. Sharing feelings online: studying emotional well-being via automated text analysis of Facebook posts. Front Psychol 2015;6:1045 [FREE Full text] [doi: 10.3389/fpsyg.2015.01045] [Medline: 26257692]
24. Cavazos-Rehg PA, Krauss MJ, Sowles S, Connolly S, Rosas C, Bharadwaj M, et al. A content analysis of depression-related Tweets. Comput Human Behav 2016 Jan 01;54:351-357 [FREE Full text] [doi: 10.1016/j.chb.2015.08.023] [Medline: 26392678]
25. Park J, Lee DS, Shablack H, Verduyn P, Deldin P, Ybarra O, et al. When perceptions defy reality: the relationships between depression and actual and perceived Facebook social support. J Affect Disord 2016 Aug;200:37-44. [doi: 10.1016/j.jad.2016.01.048] [Medline: 27126138]
26. Golbeck J, Robles C, Turner K. Predicting personality with social media. In: Proceedings of the 2011 Conference on Human Factors in Computing Systems. 2011 Presented at: CHI EA '11; May 7-12, 2011; Vancouver, BC p. 253-262. [doi: 10.1145/1979742.1979614]
27. Tadesse MM, Lin H, Xu B, Yang L. Personality predictions based on user behavior on the Facebook social media platform. IEEE Access 2018;6:61959-61969. [doi: 10.1109/ACCESS.2018.2876502]
28. Klein DN, Kotov R, Bufferd SJ. Personality and depression: explanatory models and review of the evidence. Annu Rev Clin Psychol 2011;7:269-295 [FREE Full text] [doi: 10.1146/annurev-clinpsy-032210-104540] [Medline: 21166535]
29. Hakulinen C, Elovainio M, Pulkki-Råback L, Virtanen M, Kivimäki M, Jokela M. Personality and depressive symptoms: individual participant meta-analysis of 10 cohort studies. Depress Anxiety 2015 Jul 26;32(7):461-470 [FREE Full text] [doi: 10.1002/da.22376] [Medline: 26014798]
30. Nguyen T, Yates A, Zirikly A, Desmet B, Cohan A. Improving the generalizability of depression detection by leveraging clinical questionnaires. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. 2022 Presented at: ACL '22; May 22-27, 2022; Virtual Event p. 8446-8459 URL: https://aclanthology.org/2022.acl-long.578.pdf [doi: 10.18653/v1/2022.acl-long.578]
31. Unlu A, Truong S, Tammi T, Lohiniva AL. Exploring political mistrust in pandemic risk communication: mixed-method study using social media data analysis. J Med Internet Res 2023 Oct 20;25:e50199 [FREE Full text] [doi: 10.2196/50199] [Medline: 37862088]
32. Kelman HC. Compliance, identification, and internalization three processes of attitude change. J Conflict Resolut 1958 Mar 01;2(1):51-60. [doi: 10.1177/002200275800200106]
33. About mental illness. Government of Canada. 2017. URL: https://www.canada.ca/en/public-health/services/about-mental-illness.html [accessed 2024-01-29]
34. Neumann R, Strack F. "Mood contagion": the automatic transfer of mood between persons. J Pers Soc Psychol 2000 Aug;79(2):211-223. [doi: 10.1037//0022-3514.79.2.211] [Medline: 10948975]
35. Gleick ME. The Information: A History, a Theory, a Flood. New York, NY: Pantheon; 2012.
36. Mental health myths and facts. MentalHealth.gov. 2020. URL: https://www.samhsa.gov/mental-health/myths-and-facts [accessed 2023-07-23]
37. Murarka A, Radhakrishnan B, Ravichandran S. Classification of mental illnesses on social media using RoBERTa. In: Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis. 2021 Presented at: HTMIA '21; April 19, 2021; Virtual Event p. 59-68 URL: https://aclanthology.org/2021.louhi-1.7.pdf
38. Cui B, Wang J, Lin H, Zhang Y, Yang L, Xu B. Emotion-based reinforcement attention network for depression detection on social media: algorithm development and validation. JMIR Med Inform 2022 Aug 09;10(8):e37818 [FREE Full text] [doi: 10.2196/37818] [Medline: 35943770]
39. Ghosal S, Jain A. Depression and suicide risk detection on social media using fastText embedding and XGBoost classifier. Procedia Comput Sci 2023;218:1631-1639. [doi: 10.1016/j.procs.2023.01.141]
40. Kosinski M, Matz SC, Gosling SD, Popov V, Stillwell D. Facebook as a research tool for the social sciences: opportunities, challenges, ethical considerations, and practical guidelines. Am Psychol 2015 Sep;70(6):543-556. [doi: 10.1037/a0039210] [Medline: 26348336]

XSL•FO
RenderX

41. Mehmood A, Natgunanathan I, Xiang Y, Hua G, Guo S. Protection of big data privacy. IEEE Access 2016;4:1821-1834. [doi: 10.1109/ACCESS.2016.2558446]
42. Goldberg LR, Johnson JA, Eber HW, Hogan R, Ashton MC, Cloninger CR, et al. The international personality item pool and the future of public-domain personality measures. J Res Pers 2006 Feb;40(1):84-96. [doi: 10.1016/j.jrp.2005.08.007]
43. Radloff LS. The CES-D scale: a self-report depression scale for research in the general population. Appl Psychol Meas 1977 Jun 01;1(3):385-401. [doi: 10.1177/014662167700100306]
44. Lewinsohn PM, Seeley JR, Roberts RE, Allen NB. Center for Epidemiologic Studies Depression scale (CES-D) as a screening instrument for depression among community-residing older adults. Psychol Aging 1997 Jun;12(2):277-287. [doi: 10.1037//0882-7974.12.2.277] [Medline: 9189988]
45. Julian LJ, Gregorich SE, Tonner C, Yazdany J, Trupin L, Criswell LA, et al. Using the Center for Epidemiologic Studies Depression Scale to screen for depression in systemic lupus erythematosus. Arthritis Care Res (Hoboken) 2011 Jun;63(6):884-890 [FREE Full text] [doi: 10.1002/acr.20447] [Medline: 21312347]
46. Bollen J, Gonçalves B, Ruan G, Mao H. Happiness is assortative in online social networks. Artif Life 2011;17(3):237-251. [doi: 10.1162/artl_a_00034] [Medline: 21554117]
47. Mohammad SM. Word affect intensities. In: Proceedings of the 11th Edition of the Language Resources and Evaluation Conference. 2018 Presented at: LREC '18; May 7-12, 2018; Miyazaki, Japan URL: https://aclanthology.org/L18-1027.pdf
48. Morinaga S, Yamanishi K, Tateishi K, Fukushima T. Mining product reputations on the web. In: Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2002 Presented at: KDD '02; July 23-26, 2002; Edmonton, AB p. 341-349 URL: https://dl.acm.org/doi/10.1145/775047.775098 [doi: 10.1145/775094.775098]
49. Rajaraman A, Ullman JD. Mining of Massive Datasets. Cambridge, MA: Cambridge University Press; 2011.
50. Bagby RM, Psych C, Quilty LC, Ryder AC. Personality and depression. Can J Psychiatry 2008 Jan;53(1):14-25. [doi: 10.1177/070674370805300104] [Medline: 18286868]
51. Ghosh S, Ekbal A, Bhattacharyya P. VAD-assisted multitask transformer framework for emotion recognition and intensity prediction on suicide notes. Inf Process Manag 2023 Mar;60(2):103234. [doi: 10.1016/j.ipm.2022.103234]
52. McCrae RR, Costa PT. Personality in Adulthood: A Five-Factor Theory Perspective. New York, NY: The Guilford Press; 1990.
53. McCrae RR, John OP. An introduction to the five-factor model and its applications. J Pers 1992 Jun;60(2):175-215. [doi: 10.1111/j.1467-6494.1992.tb00970.x] [Medline: 1635039]
54. Preoţiuc-Pietro D, Eichstaedt J, Park G, Sap M, Smith L. The role of personality, age, and gender in tweeting about mental illness. In: Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. 2015 Presented at: CLPsych '15; June 5, 2015; Denver, CO p. 21-30 URL: https://aclanthology.org/W15-1203.pdf [doi: 10.3115/v1/w15-1203]
55. Yang X, McEwen R, Ong LR, Zihayat M. A big data analytics framework for detecting user-level depression from social networks. Int J Inf Manage 2020 Oct;54:102141. [doi: 10.1016/j.ijinfomgt.2020.102141]
56. Muñoz S, Iglesias CA. A text classification approach to detect psychological stress combining a lexicon-based feature framework with distributional representations. Inf Process Manag 2022 Sep;59(5):103011. [doi: 10.1016/j.ipm.2022.103011]
57. Yao X, Chen S, Yu G. Effects of members' response styles in an online depression community based on text mining and empirical analysis. Inf Process Manag 2023 Mar;60(2):103198. [doi: 10.1016/j.ipm.2022.103198]
58. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021 Presented at: NAACL-HLT '21; June 2-7, 2019; Virtual Event p. 4171-4186 URL: https://aclanthology.org/N19-1423.pdf [doi: 10.18653/v1/N19-1423]
59. Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using siamese BERT-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. 2019 Presented at: EMNLP-NLP '19; November 3-7, 2019; Hong Kong, China p. 3982-3992 URL: https://aclanthology.org/D19-1410.pdf [doi: 10.18653/v1/d19-1410]
60. Bowman SR, Angeli G, Potts C, Manning CD. A large annotated corpus for learning natural language inference. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015 Presented at: EMNLP '15; September 17-21, 2015; Lisbon, Portugal p. 632-642 URL: https://aclanthology.org/D15-1075.pdf [doi: 10.18653/v1/d15-1075]
61. Williams A, Nangia N, Bowman SR. A broad-coverage challenge corpus for sentence understanding through inference. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2018 Presented at: NAACL '18; June 1-6, 2018; New Orleans, LA p. 1112-1122 URL: https://aclanthology.org/N18-1101.pdf [doi: 10.18653/v1/n18-1101]
62. Bagozzi RP, Dholakia UM. Intentional social action in virtual communities. J Interact Mark 2002 May 01;16(2):2-21. [doi: 10.1002/dir.10006]
63. Slonim T. Verbal behavior in individuals with generalized anxiety disorder and depressive disorders. City University of New York. 2014. URL: https://academicworks.cuny.edu/gc_etds/386 [accessed 2024-07-23]

XSL•FO
RenderX

64. Liwc2015: how it works. Linguistic Inquiry and Word Count. 2020. URL: http://liwc.wpengine.com/how-it-works/ [accessed 2021-01-23]
65. Coppersmith G, Dredze M, Harman C. Quantifying mental health signals in Twitter. In: Proceedings of the 2014 Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. 2014 Presented at: CLPsych '14; June 27, 2014; Baltimore, MD p. 51-60 URL: https://aclanthology.org/W14-3207.pdf [doi: 10.3115/v1/w14-3207]
66. Coppersmith G, Dredze M, Harman C, Hollingshead K. From ADHD to SAD: analyzing the language of mental health on twitter through self-reported diagnoses. In: Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. 2015 Presented at: CLPsych '15; June 5, 2015; Denver, CO p. 1-10 URL: https://aclanthology.org/W15-1201.pdf [doi: 10.3115/v1/w15-1201]
67. De Choudhury M, Gamon M, Counts S, Horvitz E. Predicting depression via social media. Proc Int AAAI Conf Web Soc Media 2021 Aug 03;7(1):128-137. [doi: 10.1609/icwsm.v7i1.14432]
68. Chen X, Sykora MD, Jackson TW, Elayan S. What about mood swings: identifying depression on Twitter with temporal measures of emotions. In: Proceedings of the 2018 Conference on World Wide Web. 2018 Presented at: WWW '18; April 23-27, 2018; Lyon, France p. 1653-1660 URL: https://dl.acm.org/doi/10.1145/3184558.3191624 [doi: 10.1145/3184558.3191624]
69. Marriwala N, Chaudhary D. A hybrid model for depression detection using deep learning. Measur Sens 2023 Feb;25:100587. [doi: 10.1016/j.measen.2022.100587]
70. Breiman L. Random forests. Machine Learn 2001;45:5-32 [FREE Full text] [doi: 10.1023/A:1010933404324] [Medline: 21816105]
71. Marerngsit S, Thammaboosadee S. A two-stage text-to-emotion depressive disorder screening assistance based on contents from online community. In: Proceedings of the 8th International Conference on Electrical Engineering Congress. 2020 Presented at: iEECON '20; March 4-6, 2020; Chiang Mai, Thailand p. 1-8 URL: https://ieeexplore.ieee.org/document/9077452 [doi: 10.1109/ieecon48109.2020.229524]
72. Wu W, Wu M, Yu K. Climate and weather: inspecting depression detection via emotion recognition. In: Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing. 2022 Presented at: ICASSP '22; May 23-27, 2022; Singapore, Singapore p. 6262-6266 URL: https://ieeexplore.ieee.org/document/9746634 [doi: 10.1109/icassp43922.2022.9746634]
73. Sumner C, Byers A, Shearing M. Determining personality traits and privacy concerns from Facebook activity. Black Hat Brief 2011;11:197-221 [FREE Full text]
74. Haque F, Nur RU, Jahan SA, Mahmud Z, Shah FM. A transformer based approach to detect suicidal ideation using pre-trained language models. In: Proceedings of the 23rd International Conference on Computer and Information Technology. 2020 Presented at: ICCIT '20; December 19-21, 2020; Dhaka, Bangladesh p. 1-5 URL: https://ieeexplore.ieee.org/document/9392692 [doi: 10.1109/iccit51783.2020.9392692]
75. Jiang Z, Levitan SI, Zomick J, Hirschberg J. Detection of mental health from Reddit via deep contextualized representations. In: Proceedings of the 2020 International Workshop on Health Text Mining and Information Analysis. 2020 Presented at: HTMIA '20; November 20, 2020; Virtual Event p. 147-156 URL: https://aclanthology.org/2020.louhi-1.16.pdf [doi: 10.18653/v1/2020.louhi-1.16]
76. Kleinman SS. Social identification in a computer-mediated group for women in science and engineering. Sci Commun 2000 Jun 01;21(4):344-366. [doi: 10.1177/1075547000021004002]
77. McKenna KY, Bargh JA. Plan 9 from cyberspace: the implications of the internet for personality and social psychology. Pers Soc Psychol Rev 2000 Feb 01;4(1):57-75. [doi: 10.1207/s15327957pspr0401_6]
78. Park M, McDonald D, Cha M. Perception differences between the depressed and non-depressed users in Twitter. Proc Int AAAI Conf Web Soc Media 2021 Aug 03;7(1):476-485. [doi: 10.1609/icwsm.v7i1.14425]
79. Mahmoudi N, Docherty P, Moscato P. Deep neural networks understand investors better. Decis Support Syst 2018 Aug;112:23-34. [doi: 10.1016/j.dss.2018.06.002]

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
CES-D: Center for Epidemiologic Studies Depression
IPIP: International Personality Item Pool
KNN: k-nearest neighbors
LIWC: Linguistic Inquiry and Word Count
ML: machine learning
NLP: natural language processing
SBERT: Sentence Bidirectional Encoder Representations from Transformers
TFIDF: term frequency and inverse document frequency

XSL•FO
RenderX

Edited by A Mavragani; submitted 07.05.24; peer-reviewed by J Abbas, P Zadeh, S Shatabda; comments to author 25.10.24; revised version received 05.11.24; accepted 05.12.24; published 20.01.25.

Please cite as:
Yang X, Li G
Psychological and Behavioral Insights From Social Media Users: Natural Language Processing–Based Quantitative Study on Mental Well-Being
JMIR Form Res 2025;9:e60286
URL: https://formative.jmir.org/2025/1/e60286
doi:10.2196/60286
PMID:

©Xingwei Yang, Guang Li. Originally published in JMIR Formative Research (https://formative.jmir.org), 20.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Attitudes of German General Practitioners Toward eHealth Apps for Dementia Risk Reduction: Qualitative Interview Study

Adrian Schultz[1], BSc; Melanie Luppa[1], PD, Dr Rer Med; Markus Bleckwenn[2], Prof Dr Med; Steffi G Riedel-Heller[1*], MPH, Prof Dr Med; Andrea Zuelke[1*], Dr Rer Nat

[1]Institute of Social Medicine, Occupational Health and Public Health, Leipzig University, Leipzig, Germany

[2]Institute of General Practice, Leipzig University, Leipzig, Germany

[*]these authors contributed equally

Corresponding Author:
Andrea Zuelke, Dr Rer Nat
Institute of Social Medicine, Occupational Health and Public Health
Leipzig University
Philipp Rosenthal Str. 55
Leipzig, 04103
Germany
Phone: 49 03419715483
Email: andrea.zuelke@medizin.uni-leipzig.de

Abstract

Background: eHealth interventions constitute a promising approach to disease prevention, particularly because of their ability to facilitate lifestyle changes. Although a rather recent development, eHealth interventions might be able to promote brain health and reduce dementia risk in older adults.

Objective: This study aimed to explore the perspective of general practitioners (GPs) on the potentials and barriers of eHealth interventions for brain health. Understanding the perspective of GPs allows us to identify chances and challenges for implementing eHealth apps for dementia risk reduction.

Methods: We conducted semistructured expert interviews with 9 GPs working in an outpatient setting in and near Leipzig, Germany. Data were fully transcribed and analyzed using a process model of qualitative content analysis with codes and categories being constructed inductively and deductively.

Results: We found generally favorable but balanced views of eHealth apps for brain health. Eight themes were identified and elaborated on in the data as follows: “addressing dementia,” “knowledge about dementia,” “need for information,” “potential for prevention,” “chances for apps for prevention,” “development of apps for prevention,” and “barriers of apps for prevention.” GPs talked mostly about how and when to address dementia and the requirements for their use of eHealth apps for dementia prevention. GPs stated that they only addressed dementia once abnormalities were already present or less frequently when a patient or relative expressed a direct wish, while individual dementia risk or standardized diagnostic during routine check-ups were mentioned much less frequently. According to GPs, knowledge about dementia in patients was low; therefore, patients expressed little need for information on dementia risk factors and prevention in GP practices. Most patients wished for quick information regarding diagnostics, treatment options, and progression of the disease. GPs mentioned a lack of overview of the available eHealth apps and their content. They also expressed a fear of inducing health anxiety when talking to patients about risk factors and prevention.

Conclusions: GPs want patients to receive relevant and individualized information. Prerequisites for the use of eHealth apps for dementia prevention were app characteristics related to design and content. GPs need to address dementia more routinely, assess relevant risk factors, and aid patients in a preventive role. Concerns were expressed over limited effectiveness, overwhelming patients, limited use in clinical practice, and only targeting patients with an already low risk of dementia.

(JMIR Form Res 2025;9:e56310) doi:10.2196/56310

KEYWORDS

eHealth; dementia; primary care; lifestyle; risk factor; older adults; prevention; brain health

XSL•FO
RenderX

Introduction

Background

Dementia is one of the most common neurodegenerative diseases. In Germany alone, dementia is one of the 5 most common diseases prevalent at the time of death [1]. At the same time, dementia is connected with a high burden of illness and excess costs of illness that are 2.5 times higher when compared to people without dementia [2]. Without available curative therapies, risk factors for dementia are a primary focus for dementia prevention. According to the latest report of the Lancet Commission on Dementia Prevention, Intervention, and Care, 45% of dementia cases can be explained by 14 modifiable risk factors (low levels of education, hearing impairment, hypertension, smoking, obesity, depression, physical inactivity, diabetes, higher alcohol consumption, traumatic brain injury, air pollution, social isolation, high levels of low-density lipoprotein cholesterol, and visual impairment [3]). This underlines the potential of dementia risk reduction by risk factor modification. In Germany, 38% of dementia cases are estimated to be due to 11 modifiable risk factors [4]. The use of eHealth apps has garnered considerable attention as a potential tool to assist in promoting brain health and mitigating dementia risk [5].

General practitioners (GPs) working in primary care usually consult their patients regularly and could potentially identify risk factors and early markers of dementia. GPs are also perceived to be an important support body with a high level of competence and trust [6] and play a central role in advising patients on health-related matters. Therefore, the primary care setting is crucial for the assessment of dementia risk factors and early dementia diagnosis. However, existing research has shown that dementia risk reduction is not routinely conducted in primary care settings [7,8]. Only few existing studies examined the views of GPs on dementia risk reduction.

While older adults are the most rapidly growing share of internet users in Germany [9], their use of eHealth tools has been found to be significantly lower than in other (European) countries [10]. The general use of health apps was low in a population-based sample of older (aged ≥60 years) German adults conducted in 2016, with only 16.5% reporting previous use of health apps. Concerns about data protection and lack of trust in eHealth tools were the most commonly named barriers [11]. Although these findings indicate a rather skeptical view of older German adults on eHealth tools, findings from the @ktiv-trial, testing an eHealth intervention for depression, found higher levels of adherence in older than in younger users, indicating the potential benefits of eHealth interventions for older adults [12].

Digital interventions aimed at enhancing brain health and mitigating dementia risk factors are currently an important focus of research [13-15], which explores potential benefits and challenges of digital apps for dementia risk reduction, highlighting their capacity to deliver personalized interventions and facilitate self-management on a large scale. However, despite these advancements in the field, only limited attention has been directed toward understanding the perspective of GPs, who play a critical role in adoption and implementation of these apps into clinical practice. Therefore, knowing the potentials and barriers before the development and implementation of such measures can be important to enhance potential. To the best of our knowledge, no study so far has examined the attitudes of German GPs toward eHealth apps for dementia risk reduction.

This Study

Consequently, we conducted 9 qualitative expert interviews with German GPs practicing in the Leipzig area to gain insight into GPs' subjective experience with older patients, how GPs address topics of dementia risk factors, and the possibilities of eHealth apps for prevention and managing dementia risk. Qualitative research can provide an understanding of subjective experiences, attitudes, and challenges GPs face. Qualitative studies can help explore and explain social relations between the health care system and the way patients and providers interact. In addition, qualitative approaches allow depicting aspects that might be lost by transformation of experiences into numerical forms for proper statistical analysis [16,17]. We aimed to identify possible chances and hindrances of eHealth apps for prevention of dementia and challenges for implementation. Furthermore, we aimed to provide specific recommendations toward goals, content, challenges, and possible implementation of eHealth apps to reduce dementia risk.

Methods

Study Design

This study uses a qualitative semistructured interview design. We conducted interviews with GPs via telephone using guided interviews. Telephone-based interviews were conducted to increase convenience and flexibility for participating GPs. Furthermore, we aimed to reduce the risk of social desirability bias and keep technical requirements for GPs to a minimum, arguing for telephone- instead of video-based interviews. Guidelines for the interview were designed to cover a prespecified topic list focused on dementia prevention, attitudes toward eHealth apps for dementia prevention, and technology-related factors. Although participants were reminded that the interview focuses on dementia prevention, GPs frequently steered the conversation toward identifying and managing dementia. This divergence from the intended focus of prevention was noted and addressed in real time by interviewers. Participants were encouraged to address further topics not covered by the interview guidelines at the end of the interview. We used a process model of qualitative content analysis as developed by Mayring [18] to identify key themes using inductive and deductive coding.

Sample

Participants were GPs with practices in an outpatient setting (individual practice, joint practice, or medical center) in the city of Leipzig, Germany, and its surrounding region. No further inclusion or exclusion criteria were applied. Eligible GP practices were recruited from a database at the Institute of Social Medicine, Occupational Health and Public Health, including 53 GPs who had participated in previous studies or expressed interest in participating in future research. This sample contained

XSL•FO
RenderX

GP practices from all regions of the city of Leipzig, comprising urban and suburban areas, as well as practices from the environs of Leipzig. All 53 GPs were contacted via telephone and email, of which 5 (9%) refused participation. Of the remaining 48 (91%) GPs, 12 (23%) expressed interest in participation. In 3 (6%) cases, interviews could not be scheduled due to time constraints on the side of GPs; therefore, the final sample included 9 (17%) GPs. On the basis of previous experiences conducting qualitative interviews, as well as narrowly defined research questions and a relatively homogeneous study sample [19,20], we expected to reach data saturation after 8 interviews. Interviews were conducted in a timeframe of roughly 25 minutes each. Participating GPs received detailed study information, an informed consent form, account data forms for the transfer of the expense allowance, and a sociodemographic questionnaire in advance via postal service. At the beginning of the interviews, interviewers informed participants about the method of expert interviews and encouraged them to share their experiences and opinions openly.

Analysis

All interviews followed the same general structure. Before the interviews, participants were given a short introduction on the interview topic identical to the previously received detailed study information. Interviewers informed GPs that the goal of the study was to gather information on how informed older patients are about the risk and protective factors of dementia and their willingness to use eHealth apps to manage dementia risk. In addition, they were informed that there was an interview guideline and encouraged to share their own experiences and opinions. The interview guideline was developed for this study; a full English translation is provided in Multimedia Appendix 1.

First, interviewers asked GPs whether they address the topic of dementia and cognitive ability with their older patients. If so, they were further asked with which groups of patients they bring the topic up and whether older patients addressed dementia at appointments themselves. Second, GPs were asked to assess the knowledge of patients about dementia and respective risk or protective factors. In the third section of the interview, interviewers inquired if patients wished to receive information on dementia prevention. In the fourth section, interviewers gave a brief description of how modifiable risk factors, for example, managing cardiovascular risk, increasing physical and social activity, or optimizing diet, have a positive impact on cognitive ability. Interviewers then asked participants which risk factors they considered especially promising for dementia risk reduction, whether they bring up modifiable risk factors when talking about dementia with their patients, and which risk factors patients might already know. In the following section, participants were asked to assess the potential of eHealth apps for managing cognitive decline and reducing the risk of dementia. In the sixth section, participants were asked to imagine a respective eHealth tool that addressed modifiable risk factors (eg, diet, physical and social activity, cardiovascular risk) via standardized questionnaires and that allowed patients to receive individualized notifications and information on how to reduce their personal risk for dementia. Following this, GPs were asked to assess whether older patients would make use of this program, which patients they would recommend the program to, and whether they would recommend all of it to their own patients. In section 7, GPs were asked to elaborate on facilitators and barriers for such a program. In the last section, interviewers asked for the GP's opinion on the implementation of respective eHealth apps for the prevention of dementia and which information they would need to assess the use of the service. Technical requirements, content, criteria of quality, measurable effect, privacy, and cost were given as anchors. Closing the interview, participants were thanked for their collaboration and asked whether they had an opinion, question, or if any topic was not covered in the interview.

All interviews were audio-recorded, fully transcribed and content-analyzed using MAXQDA 2022. We formed codes inductively and deductively according to the process model developed by Mayring [18] and the analytic pathway 1 as identified by Raskind et al [21] on qualitative data analysis practices in health education and health behavior research. Three trained scientific project members with experience in qualitative research independently coded and revised all interviews. The coding scheme was derived from an initial coding template based on literature, the interview guide (deductive coding), and emerging codes from the interview material (inductive coding, second-level coding within a domain, and openness for open coding during this stage). Methodological rigor was ensured by agreeing on a rule and process model-based approach before coding, iterative processing of reviewing coded data, discussing coded data in a group setting to develop and reach agreement on salient themes, repeated comparisons among coded data, combining similar or related codes into themes, and extracting themes from within codes [18,21]. After the final group discussion, categories were reviewed and given examples from the transcripts.

Ethical Considerations

This study was performed in accordance with the principles of the Declaration of Helsinki in its revised version from 2000. The Ethics Committee of the Medical Faculty of Leipzig University, Germany, approved the study (587/21-ek). Participating GPs were informed of their right to withdraw from the study at any time without consequences and provided written informed consent before participation. Each participating GP was assigned an identification number and all data were deidentified for analyses. GPs received a reimbursement of €100 (US $103).

Results

Sample Characteristics

Table 1 shows the collected sociodemographic data and information for the GPs in the sample.

XSL•FO
RenderX

Table 1. Sample characteristics (N=9).

Characteristics	Values
Sex, n (%)	
Male	4 (44)
Female	5 (56)
Age (y), mean (SD; range)	46.3 (8.8); 33-57
Duration of practice as resident (y), mean (SD)	10.3 (5.5)
11-25, n (%)	3 (33)
26-50, n (%)	4 (44)
51-75, n (%)	2 (22)
Older patients treated per quarter, n (%)	
51-100	1 (11)
101-500	4 (44)
501-1000	3 (33)
>1000	1 (11)
Estimated percentage of older patients with dementia, mean (SD)	20.11 (11.1)
Form of practice, n (%)	
Individual practice	4 (44)
Joint practice (*Gemeinschaftspraxis*)	3 (33)
Medical center	2 (22)

Themes and Subthemes Identified

We identified 8 themes, each including 2 to 12 subthemes. Table 2 shows the overall frequency of quotes per subcategory.

The theme mentioned most frequently was addressing dementia. On the level of subthemes, requirements for the implementation of eHealth apps were mentioned more frequently than other subthemes. GPs emphasized personal requirements for using an eHealth app for dementia prevention in their care. Themes and subthemes with respective key quotations are described in Multimedia Appendix 2.

XSL•FO
RenderX

Table 2. Frequency of themes (N=8) and subthemes (N=28).

Themes and subthemes	Frequency, n[a]
Addressing dementia	54
Cause: abnormalities	17
Cause: high risk of dementia	1
Cause: routine checkup	3
Cause: affected relatives	3
Cause: initiative of patient or relatives	12
Expressed wishes toward GP[b]	18
Knowledge about dementia	13
High knowledge	2
Low knowledge	11
Need for information	17
For diagnostics, treatment options or progression	11
Psychosocial aspects	4
For risk factors	1
No need for information	1
Potential for risk reduction	35
Most promising risk factors	12
Addressing risk factors	11
Risk factors known by patients	9
Risk factors are disregarded	3
Chances for eHealth apps for dementia risk reduction	21
Advantages for patients	11
Advantages for practitioners	9
Low interest of patients	1
Development of eHealth apps for dementia risk reduction	25
Characteristics qualifying patients for apps for dementia risk reduction	13
Characteristics rendering patients unsuitable for eHealth apps for dementia risk reduction	10
Characteristics of eHealth apps for dementia risk reduction	2
Barriers to eHealth apps for dementia risk reduction	17
General barriers	3
Concerns	10
Limited use	4
Implementation of eHealth apps for dementia risk reduction	33
Prerequisite: accessibility	13
Prerequisites: characteristics of eHealth apps	10
Integration into GP care	10

[a]Frequency of reference to the respective theme or subtheme in the total sample (n=9 GPs); multiple mentions of themes or subthemes per GP possible.
[b]GP: general practitioner.

Addressing Dementia

GPs overwhelmingly addressed dementia in their care only when abnormalities were already present and noticeable by either the GP or their staff. Routine checkups or concerns of relatives played a minor role with only a singular mention of high dementia risk being the cause for addressing dementia in routine care:

XSL•FO
RenderX

> *So we address that routinely when doing check-ups, starting at 70. And in patients under 70 who have personal concerns or an increased risk of dementia from my perspective.* [GP5]

The initiative of patients or relatives was named frequently. GPs mentioned that patients rarely take initiative action regarding dementia risk reduction. Instead, GPs stated that relatives noticed changes in patients and referred them to GPs with a wish for diagnosis:

> *Oh well, when relatives register for consultation hours, they nearly always name somewhat of a reason, if it's acute. And then they'll say, "Dad is odd," or "My husband is odd" And surely also "can't we take a look?" and actively ask whether it's the case [dementia]. So yes, it's talked about quite directly.* [GP2]

Congruently, wishes expressed toward GPs had the most quotations within the theme. Wishes varied and ranged from explicit wishes for fast diagnosis to less specifically expressed emotional needs:

> *They [patients] surely want to hear that they are not forgetful when they are. Or can't remember certain things. And whether there is medication. And get their fears off their chest, which they surely can't articulate like that at home. Because it is a kind of weakness, how they feel about that. So this [GP practice] is a protected environment, where you can slowly move forward and ask whether it's still normal. So they slowly advance, maybe talk about an occurrence.* [GP4]

Knowledge About Dementia

All GPs described patients' knowledge about dementia as low. These included symptoms of dementia, especially in contrast to age-related cognitive decline, mental disorders in general, as well as types and progression of dementia:

> *I think patients don't know much about that [early signs of dementia]. Especially separating normal signs of aging and dementia is tough for a layman, I think. And in my experience, patients that worry a lot about having dementia more commonly suffer from senile depression or an anxiety disorder and not dementia. While the patients that do have dementia often negate that or trivialize that and think it's normal when there's surely an illness and a need for therapy, or at least help, from my perspective.* [GP5]

One GP mentioned that the level of education, personal experience, and degree of personal growth are relevant factors for high knowledge about dementia:

> *[Knowledge of elderly patients about dementia] is dependent on the level of education. It differs. Those who have the ability for differentiated thinking can assess that in a completely different way. And some confuse dementia with something else and simply say "Well, you're surely demented" or something like that to their relative even though he isn't. It's not only related to the level of education but personal experience that they've made, with other patients. And to their own personality.* [GP1]

Need for Information

In the experience of GPs, most patients want to receive information on personal diagnosis, treatment, and course of illness. Patients wish for standardized diagnostic as well as personally relevant assessments made by the GP. These assessments include personalized advice related to psychosocial factors, for example, living situation, relationships with family members, or concerns. Information regarding treatment options and the course of illness were mentioned more frequently than diagnostics. The need for information was only mentioned in regard to early diagnosis of dementia with only a singular mention of risk factors in the context of general health recommendations during check-ups:

> *For many it's about the question: If I had dementia, how would I continue to lead my life? Can I stay independent? Do I need to change my living situation? What happens to my relatives or close ones, commonly spouses? That's what I think concerns patients and is why dementia is omnipresent in our practice.* [GP5]

One GP elaborated that it is common for patients to have received misinformation about the use and limitations of pharmaceutical treatment for dementia by specialized health care providers. It is noteworthy that despite being included as an example in the interview guide and the inductive category system, no GP mentioned a need for information on the causes or risk factors of dementia:

> *If you suffer from dementia and see a psychiatrist, you get one or two different kinds of medication and that's that. There is nothing about lifestyle, nothing about counseling the [social] surroundings, nothing with alternative therapeutic options, what is still possible. You get your pills prescribed and that was that.* [GP6]

Potential for Dementia Risk Reduction

When asked about the most promising factors to be targeted for reducing dementia risk, GPs named social, cognitive, and physical activity. Cardiovascular risk factors were frequently named as part of routine care but are commonly not put into the context of dementia:

> *Measurable factors like blood sugar, blood pressure, and arteriosclerosis, that is definitely part of my job as a GP and an essential part of preventing dementia. The topic of social interactions, social contacts, exercise, I think those are very important. I talk about that, too, but I see my influence there as smaller because more initiative of the patient is required.* [GP5]

GPs stressed their concern over socially isolated older patients and their perceived limited influence on psychosocial factors (eg, social support). All participants stated that they addressed common risk factors in their practice but added that they are

XSL•FO
RenderX

often unsure whether patients follow their recommendations. In the experience of GPs, patients often know about broad components of a healthy lifestyle and cardiovascular risk factors for illnesses, but not for dementia in particular:

> *[I do not mention risk factors] at every contact, but I especially use check-ups and so on to ask: how do you exercise. I always ask about alcohol, and, of course, their daily routine and such things. There, I often talk about what would be good to do more of and what to do less of. But if me addressing it [risk and protective factors for dementia] is enough, that is always hard to say.* [GP8]

One GP mentioned that they addressed lifestyle factors in general but did not give patients any recommendations regarding diet, due to limited potential for improvement:

> *[Regarding diet] I wouldn't say that they [patients] should limit themselves or do less of anything. Disregarding alcohol, but nothing else. So, I wouldn't say they need to up their vitamins, because of Vitamin C... I don't see relevant potential there.* [GP9]

Chances for eHealth Apps for Dementia Risk Reduction

GPs perceived eHealth apps for dementia risk reduction mostly as a chance to standardize patient education and integrate social factors into care. Strengthening self-efficacy and motivating patients was another point mentioned by participating GPs:

> *It really is mostly about relatives who would be very interested in helping their spouse, for example. Or help their parents. And then, of course, if a patient isn't able to educate themselves about this application or read everything, with the help of relatives it would be very appreciated. I think that makes sense.* [GP3]

One GP mentioned the ability to self-administer screening tests before appointments as a chance to improve care:

> *[I would use it] just how I recommend a good depression scale if I know one, check yourself for depression. There it says a bit more, for example, moderate depression, and then I would offer to check with me again.* [GP9]

Development of eHealth Apps for Dementia Risk Reduction

GPs characterized suitable patients for eHealth apps for prevention as high in health literacy and self-efficacy. Especially younger patients and patients with high media literacy were named as suitable for eHealth apps for dementia risk reduction. Accordingly, patients who would be unsuitable for such measures were characterized as older individuals or those who lived alone without help:

> *But the older people who are alone, I don't think they could manage to do that. If I imagine two 85 year old seniors here, no. And one of those has dementia... They rather have questions like, how do I turn on the phone? And then getting to an application, that is really hard.* [GP6]

Barriers to eHealth Apps for Dementia Risk Reduction

GPs mentioned they were worried that early preventive measures for at-risk patients could increase anxiety and worries about dementia in patients. They also argued that patients who would be likely to use these measures are both already overly anxious and therefore already active. This perception would limit the effectiveness and likeliness of use for GPs severely.

One GP elaborated that it would be important for them to use eHealth only for supporting already existing care and explicitly communicate barriers of eHealth and limitations of prevention for dementia. As an example, this GP listed individualized recommendations for lifestyle changes from an eHealth app.

Implementation of eHealth apps for Dementia Risk Reduction

To implement eHealth apps for dementia risk reduction into their care, GPs mentioned that they would need to have an overview of the information provided by the app to administer its use. Furthermore, GPs emphasized a high degree of accessibility in terms of ease of use and low complexity. Reliable privacy policies and free use were mentioned as key factors for recommending such measures to patients. Most GPs advocated supervised use of preventive measures by qualified personnel (eg, GPs, psychologists, and nurses):

> *I imagine a certain type of older person at their computer, entering their data and the application says: "you have moderately severe dementia." And I don't want to imagine that patient alone with their computer.* [GP5]

Further ideas of GPs were integrating recommendations for eHealth apps for prevention into routine check-ups or cooperating with nursing services that can assess the need for lifestyle changes. Finally, one GP mentioned a lack of information regarding existing preventive measures for GPs and wished for vocational training, published guidance, and personalized letters or emails.

Discussion

Principal Findings

To the best of our knowledge, our study is the first to use a qualitative approach for examining GPs' views on eHealth apps for dementia risk reduction, interviewing 9 German GPs working in an outpatient setting. Overall, GPs named more positive than negative aspects of eHealth apps for dementia risk reduction and displayed a balanced but positive attitude toward them. It is noteworthy that participants frequently steered the conversation toward early dementia diagnosis or dementia care management, although the questions focused on dementia prevention. This was also reflected in the coding system, as GPs frequently noted that they only addressed dementia when abnormalities were present and that patients' need for information concerns only diagnostic and treatment options. Existing research has established that GPs are reluctant to perform early dementia diagnosis due to uncertainty, which seems to extend to dementia prevention [22]. Our findings may aid in the development and implementation of eHealth apps for

XSL•FO
RenderX

dementia prevention and to inform discussions about acceptance and attitudes of eHealth apps for dementia risk reduction with primary care givers in Germany.

According to the typology of eHealth implementation theories developed by Heinsch et al [23], respective theories can fall into the 5 categories of centering agency, structure, relations, meaning, and norms. Agency-centered theories focus on an individual's actions, beliefs, and attitudes as opposed to other theories that center, for example, on organizational environments or interactional processes. The topics covered by our participants are highly compatible with agency-centered theories, as there was no mention of structural, organizational, or other factors. Particularly, the Technology Acceptance Model (TAM) and the Unified Theory of Acceptance and Use of Technology (UTAUT) were reflected in our coding system. TAM proposes that technology acceptance and use are affected by perceived ease of use, perceived usefulness, and subjective norms, all of which were most frequently mentioned by our participants. UTAUT proposes that the intention to use eHealth apps is affected by effort expectancy, performance expectancy, social influence, facilitating conditions, and habit [23]. Effort expectancy and performance expectancy were factors GPs talked about in-depth, while social, environmental, and habitual factors only played a minor role, for example, for practical implementation purposes, such as delegating the use of eHealth app to staff. Addressing dementia due to high risk or during routine checkups was only mentioned 4 times. Therefore, many GPs seem to show little to no regard for primary prevention and dementia risk reduction, which is in line with recent findings from the Netherlands [6] and the United Kingdom [8], where GPs stated that they did not address dementia prevention or risk reduction with their patients in routine care.

According to the interviewed GPs, patients want to receive individualized information and apps they can work at their tempo. The social environment and relatives should be included in the use of the app. In addition, apps should be designed sensitively so as to not cause or reinforce worries regarding dementia and give realistic goals and expectations.

There seems to be a need for information on the part of patients, as GPs predominantly perceived their patients' knowledge about dementia as low. This indicates potential for increasing knowledge, for example, via eHealth apps for dementia risk reduction. Patients seem to have no interest in information on prevention of dementia, and risk factors can be attributed either to no interest in these aspects or, in our eyes more likely, to no knowledge about modifiable dementia risk factors [14,24]. When patients are unaware of a link between lifestyle and dementia risk, they have no reason to seek out their GP for information.

In an earlier study, older patients expressed the wish for eHealth services to be integrated into usual GP treatment and had favorable opinions toward eHealth services when they supplemented care [25]. Congruently, GPs were generally also in favor of individualized risk assessment and lifestyle recommendations. In a recent study [26], only a small number of interviewed GPs saw self-monitoring as a useful area for implementation of eHealth apps for prevention measures.

In accordance with these findings, the use of eHealth apps for dementia risk reduction was also only addressed by one GP in our study, while the use for psychosocial care was generally favored. In existing research on the perception of internet-based and digital health apps in general, physicians see a variety of advantages and potential uses, particularly for preventive measures and lifestyle modifications [26-29]. A meta-analysis reported small to medium effect sizes for eHealth interventions for brain health on subjective and cognitive performance as well as risk profiles for dementia [13]. More recent studies showed beneficial intervention effects on risk factors for dementia in middle aged and older adults [30,31]. Communicating respective findings regarding eHealth interventions for dementia risk reduction to GPs might increase practitioners' interest and convince them of their potential effectiveness. Against this background, raising awareness for the potential of dementia risk reduction and the role that eHealth interventions might play in this regard among GPs seems necessary to maximize the potential of risk reduction strategies.

Tying in with earlier studies [26], GPs stressed free access and ease of use. In addition, GPs felt unsure navigating the variety of digital health apps and wished for scientifically evaluated effectiveness to find appropriate apps. Despite the currently explorative stage of research, our results fit well into existing studies and allow us to draw narrow conclusions about the preventive use of internet-based or digital health apps for dementia risk reduction. A previous review identified wrongful data use, cost, and lack of quality of care as topics that concern patients the most regarding digital health apps and this fear seems to be warranted [32]. Therefore, potential eHealth apps for dementia risk reduction should prioritize no costs for patients in their financial planning if they want to ensure maximum use by GPs and patients alike. Similar to the studies that attribute the low use of digital health apps in practice to concerns regarding safety, interviewed GPs in our study also frequently mentioned security concerns and privacy issues. The most frequently named concerns in our study (limited use, low effects, and inducing anxiety and worries) can be explained both within the theoretical frameworks of TAM and UTAUT, as they seem to reflect the factors of perceived usefulness (TAM) and expected performance (UTAUT). A recent study with a similar design to our study from the United Kingdom also found that the fear of inducing health anxiety seems to be common in GPs regarding dementia prevention [8].

Another perceived barrier not mentioned in previous studies was that eHealth apps for prevention measures would only be used by patients with (1) high anxieties and worries regarding dementia, (2) high self-efficacy, or (3) high health literacy, thereby limiting potential. Generally, existing research suggests that digital health interventions improve health literacy [33,34]. However, the use of digital health apps is affected by structural inequalities, such as low socioeconomic status, ethnicity, and education status, which contribute to disparities in health literacy as well as use of additional health education programs [35]. In contrast, eHealth interventions can use various channels to provide information, for example, audiovisual or interactive formats, which may facilitate the inclusion of older adults with lower levels of education or health literacy [36]. This view was

XSL•FO
RenderX

also expressed by one of the GPs in our study (GP8). Previous studies indicate that the effects of digital health app use on health behaviors could be mediated by self-efficacy [37], which is a different effect than the one frequently proposed by our participating GPs (eHealth apps increase knowledge for already motivated patients) and could be communicated to GPs interested in the mechanisms of future internet-based preventive measures. An earlier study focusing on older adults in a low socioeconomic community found increased computer health literacy and self-efficacy because of a targeted computer intervention [38]. The GPs in our interviews also focused on increased self-efficacy and participation of patients and their social environments as one of the most promising factors for the use of eHealth apps for prevention measures. The barriers discussed by our participants and the concerns over implementation into routine care seem to reflect a fear of a lower quality of care or a loss of trust between the physician and the patient when relying too heavily on digital apps. The review by Ramachandran et al [39] on the impact of eHealth on relationships and trust in primary care shows that eHealth apps can even increase trust when patients perceive that they are receiving personalized and collaborative care. This indicates that GPs could be more in favor of eHealth apps for dementia prevention when used in a hybrid care model, where GPs use eHealth prevention tools when assessing high dementia risk in a patient. Our results allow a potential investigation of the role of eHealth apps that GPs would prefer, although the integration into routine care is a longstanding debate with many more factors to consider [40-42]. It would be interesting to see future evaluations of eHealth apps for prevention measures include constructs, such as self-efficacy, perceived usefulness, physician-patient trust, and measurement of possible covariates that might influence the effect of eHealth apps for prevention measures. Research based on social cognitive theory has consistently found that self-efficacy facilitates a change and uptake of change in health behaviors [43] and is, therefore, a relevant outcome for evaluative studies that test digital health apps.

GPs mentioned both a lack of overview of existing apps and their respective content and were unsure about how to integrate them into their clinical practice. Therefore, eHealth apps should be designed in a way that frequently refers to the GP in a primary care setting, for example, via having tips that direct patients to their GP. Furthermore, referral to eHealth apps by trusted sources like GPs has been suggested as a promising approach to increase older adults' interest in eHealth tools [35]. This could serve to strengthen the role of the GP in providing care and reduce unrealistic expectations of patients regarding the possibilities of the app, reducing 2 important barriers at once. To maximize cooperation with GPs, internet-based prevention measures should serve a clear and identifiable role in the care of patients, for example, as part of comprehensive psychosocial care provided by the GP or part of self-monitoring in-between appointments. Developers of eHealth interventions aimed at brain health and dementia risk reduction might benefit from cooperation with GPs early on in the process of designing interventions. Furthermore, offering concise information for GPs on the device, that is, contents, target group, administration in GP practice, and tutorials for GP practice could further enhance acceptance of eHealth tools among GPs and strengthen cooperation between researchers, developers, and practitioners toward the goal of dementia risk reduction [4]. A possible application for eHealth interventions for dementia risk reduction and other digital health apps should further focus on supplementing in-person health care to increase health literacy and mend care in areas with weak health infrastructure or particularly hard to reach patients [34,38,45,46]. Textbox 1 summarizes recommendations for the development and implementation of eHealth tools for dementia risk reduction, as expressed by participating GPs.

Textbox 1. Recommendations for design and implementation of eHealth tools for dementia risk reduction.

- Ease of use, low complexity: relevant information is readily traceable within the eHealth tool, intuitive navigation
- Individualized recommendations: provide guidance and recommendations based on individual dementia risk profiles and stage of motivation for behavior change
- Sincerity about the potential for dementia prevention: avoid overemphasizing individual means of risk reduction (eg, "risk reduction" instead of "prevention")
- Inclusion of partners and relatives: provide means for partners or other relatives to interact with the eHealth tool, for example, to coach patients in using the app
- Privacy and data protection: ensure privacy of user data, prohibit wrongful transfer of participant data to third parties
- Accessibility: provide the eHealth tool free of cost, easily findable in app stores
- Inclusion of general practitioners (GPs) and other health professionals: provide structured information on available eHealth tools for dementia risk reduction, including evidence of effectiveness, for GPs. Train health professionals (GPs, medical assistants, nurses, psychologists, etc) in the use of the app to facilitate supervision of patients' use of eHealth tools

Discussion of Methods

Qualitative research does not allow direct comparability via quantified scores and readily defined constructs. However, this tradeoff allows a more accurate depiction of ambivalences, subtleties, and possibly contradicting opinions that would be lost if focusing on rigorously established scales. Therefore, we decided on a structured, but explorative design to gain as broad of an insight as possible. Using the qualitative research approach of Mayring [18] in formulating the interview guide and data analysis with a very high degree of structure, validity, reliability, and generalizability, we hope to provide robust information. This procedure seems particularly appropriate, given the novelty of eHealth apps for dementia risk reduction.

XSL•FO
RenderX

Strengths and Limitations

Our study assessed a timely topic for means of dementia risk reduction, that is, the use of eHealth apps. GPs can play a major role in providing recommendations for lifestyle changes to reduce dementia risk. Therefore, assessing GPs' attitudes toward eHealth for dementia risk reduction may likely provide useful evidence for both the design and implementation of apps, as well as uptake and acceptance by patients. Still, certain limitations need to be considered when interpreting our findings. Although the interviews focused on primary prevention, that is, dementia risk reduction, several GPs referred to the use of eHealth for secondary prevention in our study. Furthermore, heterogeneity in participating GPs' age may have influenced our findings. It might be possible that younger participants may generally be more familiar with eHealth tools, while GPs in their 50s might be less accustomed to apps, possibly rendering them more skeptical toward eHealth for dementia risk reduction. To ensure the timely execution of interviews and due to time constraints of the project, GPs were recruited from established networks of GPs, raising the risk of selection bias and resulting in a rather homogeneous sample. Because we aimed to keep effort for GPs to a minimum, we only assessed limited sociodemographics and further information about GPs. A larger, more diverse sample recruited across various channels may improve the generalizability of findings in future studies. However, this study aimed to provide a first insight into German GPs' attitudes toward eHealth for a fairly new topic, that is, dementia risk reduction. We are confident that our findings provide valuable insights that can inform larger-scale surveys on the topic as well as the design of eHealth interventions aimed at reducing dementia risk.

Implications and Recommendations for Future Research

Future eHealth interventions for brain health could be developed and implemented as part of a blended care approach, similar to strategies already used in psychiatry and primary care [47-49]. Considering the intended form of use (eg, directly supervised, supported use in between appointments) before developing the app and its features would allow increased performance expectations and better evaluation. As lifestyle interventions aim to induce behavioral change, future studies for developing and evaluating eHealth apps for brain health should include measures that can assess the constructs that are intended to induce this behavioral change alongside relevant outcome measures. Along with general efficacy and effectiveness evaluations, this could allow a better understanding of not just if eHealth apps for prevention measures work but how they work, as few available interventions include comparable constructs (eg, changing health beliefs, increasing self-efficacy, and introducing or improving existing health behavior). A proof-of-concept study evaluating the Dutch MijnBreincoach app indicated no change in beliefs and attitudes despite positive evaluations [50], whereas in another study, the same app improved adherence to the Mediterranean diet, BMI, and knowledge of dementia risk reduction [30]. We recommend that future researchers rely on unified theory-based questionnaires for health beliefs, such as the Motivation to Change Lifestyle and Health Behaviors for Dementia Risk Reduction scale [51].

We further propose that researchers and developers consider the needs and concerns of GPs named in this study and similar studies to increase acceptance and adoption by practitioners. Similarly, the factors influencing user acceptance should be assessed to make future interventions valuable to practitioners and patients alike. The development of future interventions should prioritize tailoring and personalization of eHealth apps for dementia risk reduction. Features, recommendations, and information given should be relevant to the individual patient and provide a positive approach toward change of lifestyle factors, without raising or increasing fear of dementia in users. Finally, future research should pay attention to general health inequalities and access to eHealth apps for dementia risk reduction [34,35,46]. Understanding existing disparities in access, use, and outcomes among populations with different socioeconomic backgrounds can allow equitable distribution and ensure that interventions reach those who need them.

Acknowledgments

This work was funded by a junior research grant by the Medical Faculty of Leipzig University, granted to AZ. The funder had no role in the design and conduct of the study, collection, analyses, and interpretation of data, drafting of the manuscript, or the decision to submit the results. The authors wish to thank all participating general practitioners for their contribution to this study. They acknowledge support from Leipzig University within the program of open-access publishing.

Data Availability

The datasets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

AZ, ML, and SGR-H participated in the conceptualization of the study. AS and AZ participated in developing the methodology. AS and AZ were responsible for validation. AS was responsible for formal analyses and investigation. SGR-H was responsible for procuring resources. AS and AZ participated in data curation. AS was responsible for writing the original draft. AS, ML, MB, SGR-H, and AZ were responsible for writing, reviewing, and editing; AS participated in visualization. ML, SGR-H, and AZ provided supervision. AZ, ML, SGR-H participated in project administration and AZ contributed to funding acquisition. No generative AI tools were used to prepare this manuscript.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview guide.
[DOCX File , 21 KB - formative_v9i1e56310_app1.docx]

Multimedia Appendix 2
Themes, subthemes, and quotations.
[DOCX File , 25 KB - formative_v9i1e56310_app2.docx]

References

1. Doblhammer G, Fritze T, Reinke C, Fink A. Can dementia become the most prevalent disease at the time of death in Germany? Projections up to the year 2060 for the five most important diseases at the time of death. J Popul Ageing 2022 May 16;15(2):523-540. [doi: 10.1007/s12062-022-09365-7]
2. Neubert L, König HH, Löbner M, Luppa M, Pentzek M, Fuchs A, et al. Excess costs of dementia in old age (85+) in Germany: results from the AgeCoDe-AgeQualiDe study. J Econ Ageing 2021 Oct;20:100346. [doi: 10.1016/j.jeoa.2021.100346]
3. Livingston G, Huntley J, Liu KY, Costafreda SG, Selbæk G, Alladi S, et al. Dementia prevention, intervention, and care: 2024 report of the Lancet standing Commission. Lancet 2024 Aug 10;404(10452):572-628. [doi: 10.1016/S0140-6736(24)01296-0] [Medline: 39096926]
4. Blotenberg I, Hoffmann W, Thyrian JR. Dementia in Germany: epidemiology and prevention potential. Dtsch Arztebl Int 2023 Jul 10;120(27-28):470-476 [FREE Full text] [doi: 10.3238/arztebl.m2023.0100] [Medline: 37226316]
5. Solomon A, Stephen R, Altomare D, Carrera E, Frisoni GB, Kulmala J, et al. Multidomain interventions: state-of-the-art and future directions for protocols to implement precision dementia risk reduction. A user manual for Brain Health Services-part 4 of 6. Alzheimers Res Ther 2021 Oct 11;13(1):171 [FREE Full text] [doi: 10.1186/s13195-021-00875-8] [Medline: 34635167]
6. Wangler J, Jansky M. Factors influencing general practitioners' perception of and attitude towards dementia diagnostics and care-results of a survey among primary care physicians in Germany. Wien Med Wochenschr 2021 May;171(7-8):165-173 [FREE Full text] [doi: 10.1007/s10354-020-00803-9] [Medline: 33443611]
7. Godbee K, Gunn J, Lautenschlager NT, Curran E, Palmer VJ. Implementing dementia risk reduction in primary care: a preliminary conceptual model based on a scoping review of practitioners' views. Prim Health Care Res Dev 2019 Oct 23;20:e140 [FREE Full text] [doi: 10.1017/S1463423619000744] [Medline: 31640836]
8. Jones D, Drewery R, Windle K, Humphrey S, de Paiva AF. Dementia prevention and the GP's role: a qualitative interview study. Br J Gen Pract 2024 Apr;74(741):e242-e249 [FREE Full text] [doi: 10.3399/BJGP.2023.0103] [Medline: 37549993]
9. Müller A. D21-digital-index 2017/2018. Merz Medien Erziehung 2018 Apr 02;62(2):3 [FREE Full text] [doi: 10.21240/merz/2018.2.3]
10. Merkel S, Hess M. The use of internet-based health and care services by elderly people in Europe and the importance of the country context: multilevel study. JMIR Aging 2020 Jun 03;3(1):e15491 [FREE Full text] [doi: 10.2196/15491] [Medline: 32490837]
11. Rasche P, Wille M, Bröhl C, Theis S, Schäfer K, Knobe M, et al. Prevalence of health app use among older adults in Germany: national survey. JMIR Mhealth Uhealth 2018 Jan 23;6(1):e26 [FREE Full text] [doi: 10.2196/mhealth.8619] [Medline: 29362211]
12. Pabst A, Löbner M, Stein J, Luppa M, Kersting A, König HH, et al. Internet-based cognitive behavior therapy only for the young? A secondary analysis of a randomized controlled trial of depression treatment. Front Psychiatry 2020 Jul 24;11:735 [FREE Full text] [doi: 10.3389/fpsyt.2020.00735] [Medline: 32848915]
13. Wesselman LM, Hooghiemstra AM, Schoonmade LJ, de Wit MC, van der Flier WM, Sikkes SA. Web-based multidomain lifestyle programs for brain health: comprehensive overview and meta-analysis. JMIR Ment Health 2019 Apr 09;6(4):e12104 [FREE Full text] [doi: 10.2196/12104] [Medline: 30964438]
14. Heger I, Deckers K, van Boxtel M, de Vugt M, Hajema K, Verhey F, et al. Dementia awareness and risk perception in middle-aged and older individuals: baseline results of the MijnBreincoach survey on the association between lifestyle and brain health. BMC Public Health 2019 Jun 03;19(1):678 [FREE Full text] [doi: 10.1186/s12889-019-7010-z] [Medline: 31159779]
15. Richard E, Moll van Charante EP, Hoevenaar-Blom MP, Coley N, Barbera M, van der Groep A, et al. Healthy ageing through internet counselling in the elderly (HATICE): a multinational, randomised controlled trial. Lancet Digit Health 2019 Dec;1(8):e424-e434 [FREE Full text] [doi: 10.1016/S2589-7500(19)30153-0] [Medline: 33323224]

XSL•FO
RenderX

16. Mykhalovskiy E, Eakin J, Beagan B, Beausoleil N, Gibson BE, Macdonald ME, et al. Beyond bare bones: critical, theoretically engaged qualitative research in public health. Can J Public Health 2018 Dec;109(5-6):613-621 [FREE Full text] [doi: 10.17269/s41997-018-0154-2] [Medline: 30465286]
17. Greenhalgh T, Annandale E, Ashcroft R, Barlow J, Black N, Bleakley A, et al. An open letter to The BMJ editors on qualitative research. BMJ 2016 Feb 10;352:i563 [FREE Full text] [doi: 10.1136/bmj.i563] [Medline: 26865572]
18. Mayring P. Qualitative content analysis: demarcation, varieties, developments. Forum Qual Soc Res 2019;20(3). [doi: 10.17169/fqs-20.3.3343]
19. Schladitz K, Förster F, Löbner M, Welzel F, Stein J, Luppa M, et al. [Grief and loss in elderly people: a qualitative study regarding the user acceptance of an internet-based self-help program from user and expert perspective]. Z Evid Fortbild Qual Gesundhwes 2020 Apr;150-152:112-123. [doi: 10.1016/j.zefq.2020.01.007] [Medline: 32460967]
20. Baker SE, Edwards R. How many qualitative interviews is enough? Expert voices and early career reflections on sampling and cases in qualitative research. National Centre for Research Methods. URL: https://eprints.ncrm.ac.uk/id/eprint/2273/4/how_many_interviews.pdf [accessed 2024-09-26]
21. Raskind IG, Shelton RC, Comeau DL, Cooper HL, Griffith DM, Kegler MC. A review of qualitative data analysis practices in health education and health behavior research. Health Educ Behav 2019 Feb 18;46(1):32-39 [FREE Full text] [doi: 10.1177/1090198118795019] [Medline: 30227078]
22. Wangler J, Jansky M. Factors influencing general practitioners' perception of and attitude towards dementia diagnostics and care—results of a survey among primary care physicians in Germany. Wien Med Wochenschr 2021 Jan 14;171(7-8):165-173. [doi: 10.1007/s10354-020-00803-9]
23. Heinsch M, Wyllie J, Carlson J, Wells H, Tickner C, Kay-Lambkin F. Theories informing eHealth implementation: systematic review and typology classification. J Med Internet Res 2021 May 31;23(5):e18500 [FREE Full text] [doi: 10.2196/18500] [Medline: 34057427]
24. Zülke AE, Luppa M, Köhler S, Riedel-Heller SG. Knowledge of risk and protective factors for dementia in older German adults: a population-based survey on risk and protective factors for dementia and internet-based brain health interventions. PLoS One 2022 Nov 7;17(11):e0277037 [FREE Full text] [doi: 10.1371/journal.pone.0277037] [Medline: 36342935]
25. Nymberg VM, Bolmsjö BB, Wolff M, Calling S, Gerward S, Sandberg M. 'Having to learn this so late in our lives…' Swedish elderly patients' beliefs, experiences, attitudes and expectations of e-health in primary health care. Scand J Prim Health Care 2019 Mar;37(1):41-52 [FREE Full text] [doi: 10.1080/02813432.2019.1570612] [Medline: 30732519]
26. Wangler J, Jansky M. Gesundheits-apps als instrumente der prävention? – Eine interviewstudie zu potenzialen für das hausärztliche setting. Präv Gesundheitsf 2020 Mar 31;15(4):340-346. [doi: 10.1007/s11553-020-00769-x]
27. Bittner J. So denken Ärzte über digital health: eine synopse der aktuellen umfragen. Bertelsmann Stiftung. 2017 Aug 21. URL: https://blog.der-digitale-patient.de/synopse-aerztebefragungen-digital-health/ [accessed 2023-04-21]
28. Realitäts-check – nutzen Ärzte health apps? Coliquio. 2017 Jul 31. URL: https://www.coliquio-insights.de/nutzen-aerzte-health-apps-exklusive-umfrage/ [accessed 2023-04-21]
29. Rohlender B, Reinhardt K. Gesundheit 4.0 – wie Ärzte die digitale Zukunft sehen. Hartmannbund. 2017 Jun 8. URL: http://www.hartmannbund.de/fileadmin/user_upload/Downloads/Umfragen/2017_HB-Bitkom_Start-ups.pdf [accessed 2023-04-21]
30. Van Asbroeck S, Köhler S, Wimmers SC, Muris JW, van Boxtel MP, Deckers K. Multidomain dementia risk reduction in primary care is feasible: a proof-of-concept study. J Alzheimers Dis 2024;99(4):1455-1471 [FREE Full text] [doi: 10.3233/JAD-240229] [Medline: 38759017]
31. McMaster M, Kim S, Clare L, Torres SJ, Cherbuin N, D Este C, et al. Lifestyle risk factors and cognitive outcomes from the multidomain dementia risk reduction randomized controlled trial, body brain life for cognitive decline (BBL-CD). J Am Geriatr Soc 2020 Nov;68(11):2629-2637. [doi: 10.1111/jgs.16762] [Medline: 32909259]
32. Crossnohere NL, Weiss B, Hyman S, Bridges JF. Patient preferences for health information technologies: a systematic review. J Hosp Manag Health Policy 2021 Sep;5:25 [FREE Full text] [doi: 10.21037/jhmhp-20-105]
33. Jacobs RJ, Lou JQ, Ownby RL, Caballero J. A systematic review of eHealth interventions to improve health literacy. Health Informatics J 2016 Jun;22(2):81-98 [FREE Full text] [doi: 10.1177/1460458214534092] [Medline: 24916567]
34. Bodie GD, Dutta MJ. Understanding health literacy for strategic health marketing: eHealth literacy, health disparities, and the digital divide. Health Mark Q 2008;25(1-2):175-203. [doi: 10.1080/07359680802126301] [Medline: 18935884]
35. Wilson J, Heinsch M, Betts D, Booth D, Kay-Lambkin F. Barriers and facilitators to the use of e-health by older adults: a scoping review. BMC Public Health 2021 Aug 17;21(1):1556 [FREE Full text] [doi: 10.1186/s12889-021-11623-w] [Medline: 34399716]
36. Meppelink CS, van Weert JC, Haven CJ, Smit EG. The effectiveness of health animations in audiences with different health literacy levels: an experimental study. J Med Internet Res 2015 Jan 13;17(1):e11 [FREE Full text] [doi: 10.2196/jmir.3979] [Medline: 25586711]
37. Choi M. Association of eHealth use, literacy, informational social support, and health-promoting behaviors: mediation of health self-efficacy. Int J Environ Res Public Health 2020 Oct 28;17(21):7890 [FREE Full text] [doi: 10.3390/ijerph17217890] [Medline: 33126469]

XSL•FO
RenderX

38. Chu A, Huber J, Mastel-Smith B, Cesario S. "Partnering with seniors for better health": computer use and internet health information retrieval among older adults in a low socioeconomic community. J Med Libr Assoc 2009 Jan;97(1):12-20 [FREE Full text] [doi: 10.3163/1536-5050.97.1.003] [Medline: 19159002]
39. Ramachandran M, Brinton C, Wiljer D, Upshur R, Gray CS. The impact of eHealth on relationships and trust in primary care: a review of reviews. BMC Prim Care 2023 Nov 03;24(1):228 [FREE Full text] [doi: 10.1186/s12875-023-02176-5] [Medline: 37919688]
40. Carey M, Noble N, Mansfield E, Waller A, Henskens F, Sanson-Fisher R. The role of eHealth in optimizing preventive care in the primary care setting. J Med Internet Res 2015 May 22;17(5):e126 [FREE Full text] [doi: 10.2196/jmir.3817] [Medline: 26001983]
41. Stevens WJ, van der Sande R, Beijer LJ, Gerritsen MG, Assendelft WJ. eHealth apps replacing or complementing health care contacts: scoping review on adverse effects. J Med Internet Res 2019 Mar 01;21(3):e10736 [FREE Full text] [doi: 10.2196/10736] [Medline: 30821690]
42. Tossaint-Schoenmakers R, Versluis A, Chavannes N, Talboom-Kamp E, Kasteleyn M. The challenge of integrating eHealth into health care: systematic literature review of the Donabedian model of structure, process, and outcome. J Med Internet Res 2021 May 10;23(5):e27180 [FREE Full text] [doi: 10.2196/27180] [Medline: 33970123]
43. Warner LM, Schwarzer R. Self-efficacy and health. In: Gulliver SB, Cohen LM, editors. The Wiley Encyclopedia of Health Psychology. Hoboken, NJ: John Wiley & Sons; 2020.
44. Steyaert J, Deckers K, Smits C, Fox C, Thyrian R, Jeon YH, et al. Putting primary prevention of dementia on everybody's agenda. Aging Ment Health 2021 Aug;25(8):1376-1380. [doi: 10.1080/13607863.2020.1783514] [Medline: 32590910]
45. Hantgan SL, Jariwala SP. User-centered mobile health applications for asthma. Ann Allergy Asthma Immunol 2023 Feb;130(2):159-160 [FREE Full text] [doi: 10.1016/j.anai.2022.11.011] [Medline: 36400353]
46. Beacom AM, Newman SJ. Communicating health information to disadvantaged populations. Fam Community Health 2010;33(2):152-162. [doi: 10.1097/FCH.0b013e3181d59344] [Medline: 20216358]
47. Grätzel P. Blended treatment: theoretisch gut, in der umsetzung mühsam. InFo Neurologie Psychiatrie 2020 Mar 26;22(3):57-58. [doi: 10.1007/s15005-020-1314-5]
48. Maehder K, Löwe B, Härter M, Heddaeus D, Scherer M, Weigel A. Management of comorbid mental and somatic disorders in stepped care approaches in primary care: a systematic review. Fam Pract 2019 Jan 25;36(1):38-52. [doi: 10.1093/fampra/cmy122] [Medline: 30535053]
49. Löbner M, Pabst A, Stein J, Dorow M, Matschinger H, Luppa M, et al. Computerized cognitive behavior therapy for patients with mild to moderately severe depression in primary care: a pragmatic cluster randomized controlled trial (@ktiv). J Affect Disord 2018 Oct 01;238:317-326. [doi: 10.1016/j.jad.2018.06.008] [Medline: 29902736]
50. Heger I, Deckers K, de Vugt M, Verhey F, Oenema A, van Boxtel M, et al. Using mHealth for primary prevention of dementia: a proof-of-concept study on usage patterns, appreciation, and beliefs and attitudes regarding prevention. J Alzheimers Dis 2023;94(3):935-948 [FREE Full text] [doi: 10.3233/JAD-230225] [Medline: 37355903]
51. Kim S, Sargent-Cox K, Cherbuin N, Anstey KJ. Development of the motivation to change lifestyle and health behaviours for dementia risk reduction scale. Dement Geriatr Cogn Dis Extra 2014 May;4(2):172-183 [FREE Full text] [doi: 10.1159/000362228] [Medline: 25028583]

Abbreviations

GP: general practitioner
TAM: Technology Acceptance Model
UTAUT: Unified Theory of Acceptance and Use of Technology

Edited by A Mavragani; submitted 12.01.24; peer-reviewed by A Paiva, K Bond; comments to author 19.09.24; revised version received 24.10.24; accepted 14.11.24; published 22.01.25.

Please cite as:
Schultz A, Luppa M, Bleckwenn M, Riedel-Heller SG, Zuelke A
Attitudes of German General Practitioners Toward eHealth Apps for Dementia Risk Reduction: Qualitative Interview Study
JMIR Form Res 2025;9:e56310
URL: https://formative.jmir.org/2025/1/e56310
doi:10.2196/56310
PMID:

©Adrian Schultz, Melanie Luppa, Markus Bleckwenn, Steffi G Riedel-Heller, Andrea Zuelke. Originally published in JMIR Formative Research (https://formative.jmir.org), 22.01.2025. This is an open-access article distributed under the terms of the

XSL•FO
RenderX

Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Discrimination of Radiologists' Experience Level Using Eye-Tracking Technology and Machine Learning: Case Study

Stanford Martinez[1], PhD; Carolina Ramirez-Tamayo[1], MSc; Syed Hasib Akhter Faruqui[2], PhD; Kal Clark[3], MD; Adel Alaeddini[4], PhD; Nicholas Czarnek, PhD; Aarushi Aggarwal[3], MD; Sahra Emamzadeh[3], MD, MPH; Jeffrey R Mock[5], PhD; Edward J Golob[5], PhD

[1]Department of Mechanical Engineering, The University of Texas at San Antonio, San Antonio, TX, United States
[2]Department of Engineering Technology, Sam Houston State University, Huntsville, TX, United States
[3]Department of Radiology, University of Texas Health Science Center at San Antonio, San Antonio, TX, United States
[4]Department of Mechanical Engineering, Southern Methodist University, Dallas, TX, United States
[5]Department of Psychology, The University of Texas at San Antonio, San Antonio, TX, United States

Corresponding Author:
Adel Alaeddini, PhD
Department of Mechanical Engineering
Southern Methodist University
3101 Dyer Street
Dallas, TX, 75205
United States
Phone: 1 214 768 3050
Email: aalaeddini@smu.edu

Abstract

Background: Perception-related errors comprise most diagnostic mistakes in radiology. To mitigate this problem, radiologists use personalized and high-dimensional visual search strategies, otherwise known as search patterns. Qualitative descriptions of these search patterns, which involve the physician verbalizing or annotating the order he or she analyzes the image, can be unreliable due to discrepancies in what is reported versus the actual visual patterns. This discrepancy can interfere with quality improvement interventions and negatively impact patient care.

Objective: The objective of this study is to provide an alternative method for distinguishing between radiologists by means of captured eye-tracking data such that the raw gaze (or processed fixation data) can be used to discriminate users based on subconscious behavior in visual inspection.

Methods: We present a novel discretized feature encoding based on spatiotemporal binning of fixation data for efficient geometric alignment and temporal ordering of eye movement when reading chest x-rays. The encoded features of the eye-fixation data are used by machine learning classifiers to discriminate between faculty and trainee radiologists. A clinical trial case study was conducted using metrics such as the area under the curve, accuracy, F_1-score, sensitivity, and specificity to evaluate the discriminability between the 2 groups regarding their level of experience. The classification performance was then compared with state-of-the-art methodologies. In addition, a repeatability experiment using a separate dataset, experimental protocol, and eye tracker was performed with 8 participants to evaluate the robustness of the proposed approach.

Results: The numerical results from both experiments demonstrate that classifiers using the proposed feature encoding methods outperform the current state-of-the-art in differentiating between radiologists in terms of experience level. An average performance gain of 6.9% is observed compared with traditional features while classifying experience levels of radiologists. This gain in accuracy is also substantial across different eye tracker–collected datasets, with improvements of 6.41% using the Tobii eye tracker and 7.29% using the EyeLink eye tracker. These results signify the potential impact of the proposed method for identifying radiologists' level of expertise and those who would benefit from additional training.

Conclusions: The effectiveness of the proposed spatiotemporal discretization approach, validated across diverse datasets and various classification metrics, underscores its potential for objective evaluation, informing targeted interventions and training strategies in radiology. This research advances reliable assessment tools, addressing challenges in perception-related errors to enhance patient care outcomes.

XSL•FO
RenderX

(JMIR Form Res 2025;9:e53928) doi:10.2196/53928

KEYWORDS

machine learning; eye-tracking; experience level determination; radiology education; search pattern feature extraction; search pattern; radiology; classification; gaze; fixation; education; experience; spatio-temporal; image; x-ray; eye movement

Introduction

Lung cancer is the leading cause of cancer death, claiming 139,000 American lives yearly [1]. To mitigate its impact, the US Preventative Task Force recommends annual radiological screening for at-risk individuals [2]. Radiologists identify suspicious lung lesions (nodules) from patient chest images and recommend further management, including biopsy, continued surveillance, or further workup. Radiological surveillance reduces population mortality from lung cancer, but it is estimated that radiologists will make errors on 33% of abnormal chest exams, eliminating the chance for patients to start lifesaving treatment [3]. The predominant source of these errors is not deficient medical knowledge. Rather, errors primarily stem from the methods radiologists use to visually inspect the image, referred to as perceptual errors [4]. In other words, perceptual errors in radiology are mistakes that occur during the visual inspection and interpretation of medical images. They are distinct from cognitive errors, which involve incorrect reasoning or decision-making based on observed information. There are 2 primary patterns for overlooking a disease due to perceptual errors:

- Examining the affected area but ignoring the disease: This occurs when the radiologist inspects the region with the abnormality but fails to recognize it, possibly due to subtle presentation, distractions, or visual fatigue.
- Not examining the affected area: This happens when the radiologist misses the region with the abnormality entirely, often due to inefficient search patterns, incomplete scanning, or being misled by more prominent findings elsewhere.

Kundel [5] investigated the effects of perceptual errors in radiology and concluded that decisions and outcomes improve when radiologists' experiences are enhanced.

Radiologists and radiology educators understand the stakes associated with missed diagnoses due to perceptual errors but have limited tools to combat these errors. Classical educational texts include general concepts, for example, "...scan the areas of least interest first, working toward the more important areas" [6], which, unfortunately, are inadequate to improve radiologist performance meaningfully.

Eye-tracking technology has been previously proposed as a tool to evaluate radiologist perception. Eye trackers are powerful because they provide high (>30 Hz) temporal and spatial resolution (approximately 1 degree of error). With the aid of eye tracking, quantitative analyses can be performed to understand the cognitive and perceptual processes better. Eye-tracking technology has previously proven relevant in evaluating decision-making processes [7], attention interruption [8], skill level determination [9], and impact of search pattern education [10].

In 2017, van der Gijp et al [11] performed a systematic literature review outlining the current state of science concerning visual perception in radiology. A key tenet is the global-focal search model [12-14], which can be summarized as the generation of an initial, fast global impression followed by a more detailed focal search. Eye-tracking technology allows these principles to be tested and potentially optimized to evaluate all clinically relevant portions of the exam in greater detail. Of the 22 relevant articles van der Gijp et al [11] reviewed, a consensus "traditional" feature set consisting of 5 features that could be experimentally measured was found to be associated with expertise.

Despite the development of this consensus feature set, visual search complexity may not be adequately captured by simple, low-dimensional features that do not fully describe how visual perception relates to skill. Machine learning is well-suited to provide deeper insight into radiologist visual search behavior and how this relates to radiologist performance. Waite et al [3] highlighted the importance of understanding perceptual expertise in radiology and the potential use of eye-tracking and perceptual learning methods in medical training to improve diagnostic accuracy. Lim et al [15] identified several features that can be extracted from eye-tracking data, including pupil size, saccade, fixations, velocity, blink, pupil position, electrooculogram, and gaze point, to be used in machine learning models. Among these features, fixation was the most commonly used feature in the studies reviewed.

Shamyuktha et al [16] developed a machine learning framework using eye gaze data such as saccade latency and amplitude to classify expert and nonexpert radiologists. Harezlak et al [17] investigated eye movement traits to differentiate experts and laymen in a similar study. Akshay et al [18] proposed a machine learning algorithm to identify eye movement metrics using raw eye-tracking data. Rizzo et al [19] used machine learning to detect cognitive interference based on eye-tracking data. Öder et al [20] applied machine learning to classify familiar web users based on eye-tracking data. Indeed, these techniques can be used to enhance competency assessment and feedback techniques in radiologists.

Eye tracking also holds the potential for understanding the longitudinal aspects of competency progression in medical education, allowing for examining how interpretive and diagnostic skills develop over time. Karargyris et al [21] and Bigolin Lanfredi et al [22] created and validated chest x-ray datasets with eye-tracking data and report dictation for developing such artificial intelligence systems. These datasets aim to support the research community in developing more complex support tools for radiology research.

In this study, we use machine learning to compare the discriminability of 2 radiologists of different skill levels using, first, the aforementioned "traditional" gaze-based features (such

XSL•FO
RenderX

as time to scan, saccade length, the total number of fixations, and total regressive fixations) [11] and second, the "proposed" features that we developed to describe high-dimensional visual search patterns thoroughly and quantitatively. We curate the traditional feature sets to those that could be practically acquired without laborious manual ground truthing of exams, as this would permit large-scale deployment of this technology to health care institutions. To highlight the use of eye-tracking data and artificial intelligence, we term our general approach "biometric radiology artificial intelligence."

The driving hypothesis behind the work presented in this paper is that gaze patterns measurably differ among radiologists as a function of their experience level. To test this hypothesis, we proposed a novel discretized feature encoding method that condenses fixation data into a few representative spatiotemporal bins for descriptive and predictive analytics purposes (Figure 1). With spatiotemporal binning, fixations are divided into a predefined number of temporal segments (bins). Within each temporal bin, the fixations are counted within spatial subdivisions of the image. This process results in a vector that captures detailed and structured information about both where and when fixations occurred. By splitting fixations into temporal bins, we capture the evolution of the visual search process over time, providing insights into how radiologists allocate their attention during different phases of image inspection. Also, spatial binning allows us to understand which regions of the image are being focused on and how frequently. In addition, this method transforms raw fixation data into structured features that can be effectively used by machine learning models.

Figure 1. Overall algorithm: the steps required to generate proposed features from the raw dataset and build the proposed machine learning model.

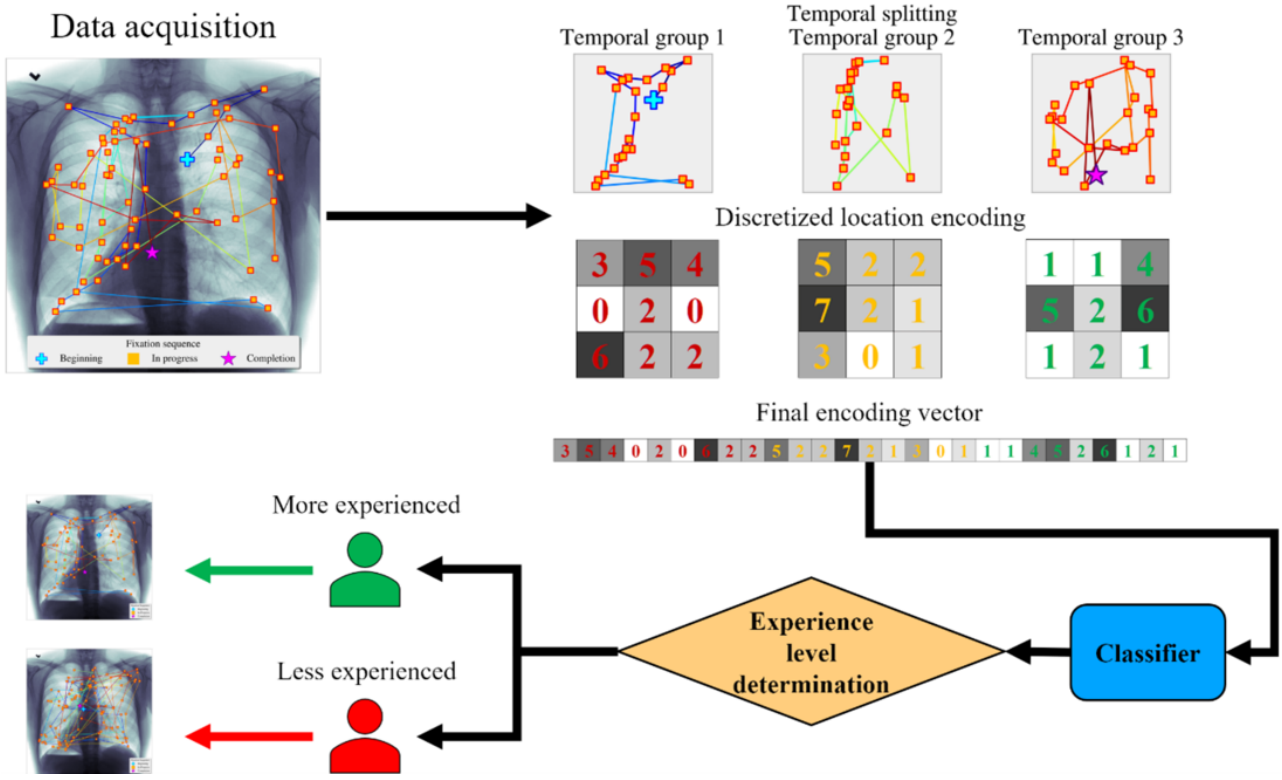

We collected the gaze fixation data from radiologists while they were reading the x-rays. These data were then segmented into fixed temporal groups before discretizing them to convert them into final encoded vectors. The final encoded features were then used in training machine learning algorithms to classify radiologists.

Collecting data from 2 participants—1 faculty member (expert) and 1 resident (trainee)—we analyzed their behavior and level of experience using the proposed approach. Using stratified cross-validation over 10 folds, we compared the area under the curve (AUC) performance of several classifiers using the proposed methodology with the AUC performance of those same classifiers when using a traditional feature set (Table 1). We then confirmed our results using data from a second similarly designed, larger study evaluating 8 participants—4 faculty members (expert) and 4 residents (trainee). The remainder of the paper is structured as follows: *Methods* presents the data collection and preparation procedures and details of the proposed method; *Results* describes the simulation study and interpretation; and *Discussion* presents the discussion, concluding remarks, and advice to practitioners.

XSL•FO
RenderX

Table 1. State-of-the-art features from 22 relevant studies.

Attribute (per trial)	Attribute description	Association with high level of expertise (percentage of the total number of included studies)
Total time to scan	Measures the total duration spent scanning the chest x-ray image, indicating the thoroughness of the visual inspection	Decrease (45.45%)
Regressive fixation count	Counts the number of distinct locations revisited during the scan, suggesting areas of uncertainty or interest	Increase (4.55%) or decrease (4.55%)
Fixation count	Total number of fixations, reflecting the intensity of the visual scrutiny	Decrease (18.18%)
Total saccade length	Sum of all saccade lengths, indicating the extent and pattern of the visual search (time between fixations) [23] captured in a single chest x-ray scan	Increase (9.09%) or decrease (4.55%)
Coverage	Percentage of salient regions covered by the gaze, reflecting the comprehensiveness of the examination	Increase (9.09%) or decrease (9.09%)

Methods

Study Design, Data Collection, and Preparation

Overview

The study design was prospective, controlled, block-randomized, and Institutional Review Board (IRB) approved. Each study participant completed 4 roughly 1-hour sessions in a radiology reading room, including tutorial, calibration, assessment, and annotation periods. The tutorial included an overview of the assessment period and instructions on how to perform dictation and annotation consistently. Calibration was performed to ensure that recorded and actual gaze were consistent based on a 9-point custom calibration mapping script.

Nodule and normal cases were derived from the Shiraishi 2000 chest radiograph dataset [24], which includes 154 chest radiographs with 5 degrees of subtlety from level 1 (extremely subtle) to level 5 (obvious). Distractor cases were derived from the VinDr chest radiograph dataset [25]. A total of 3 sets of 6 nodule cases from the Japanese Society of Radiological Technology dataset, 1 set each from the intermediate difficulty levels (2, 3, and 4), and 1 set of 9 normal cases from the Japanese Society of Radiological Technology dataset were randomly sampled without replacement. In total, 2 cases each of pneumothorax, cardiomegaly, and consolidation from the VinDr dataset were randomly sampled without replacement to serve as distractor cases. These distractor cases functioned mainly to prevent control subject bias to the nodule detection task. Each participant reviewed the exams only once during the trial, and all study participants reviewed the same set of cases.

A custom software tool was developed to automatically display the study images and capture time-stamped bilateral gaze, bilateral pupil, head pose, voice, annotation, and image display configuration data. No chin rest was used to ensure that the study was performed in a manner that was as close as possible to a clinical setting. After each session, data were transferred to a database for further analysis.

Data Acquisition

In the first study, the EyeLink 1000 eye tracker and software were used to collect eye-tracking data [26]. A total of 2 participants—1 faculty member (9 years of faculty experience) and 1 resident (3 years of trainee experience)—observed a series of chest x-ray images, which contained a balanced class composition of normal scans (no abnormalities), abnormal scans (mass or nodule present), or abnormal scans with pleural effusion. A total of 110 trials (55 trials were studied by each participant) were recorded. We leveraged the EyeLink suite to remove most artifacts, such as blinks, from the eye-tracking data captured in each participant's trial and manually filtered remaining artifacts, such as off-screen distractions left unprocessed (eg, the far-displaced fixations in Figure 2).

XSL•FO
RenderX

Figure 2. Example of eye-tracking fixations for 1 trial processed by the EyeLink software. The fixations illustrated include participants 1 (blue) and 2 (red) superimposed on the image displayed during the trial. The "invalid" fixations that were not successfully filtered out are shown as "x" markers and were manually removed during data processing.

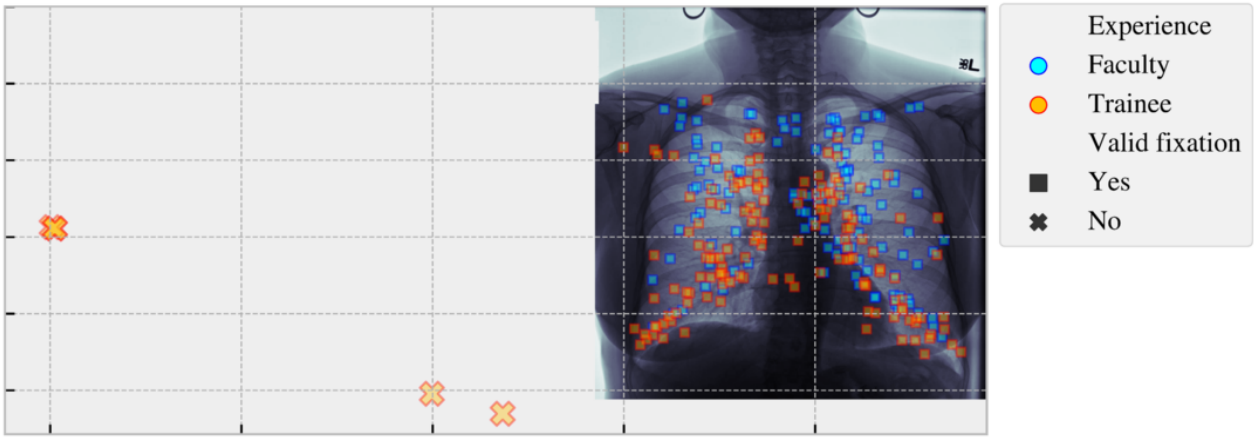

In the second study, a Tobii 5L eye tracker was used [27]. This second dataset included 8 participants (4 faculty with an average of 12.75 years of faculty experience and 4 trainees with 2.25 years of trainee experience), each scanning the same set of 27 images. The Tobii gaze data were unprocessed to evaluate the robustness of the proposed method to fixation postprocessing.

Common (Traditional) Features

To establish a necessary baseline to which the proposed methodology can compare, several attributes were established based on a meta-analysis done by van der Gijp et al [11] in 2017. The baseline is used for two main reasons: (1) traditional features are well-known and correlate with radiologist expertise, serving as a necessary reference point to evaluate our proposed method's effectiveness, and (2) comparing our novel discretized vector encoding method against this baseline demonstrates the added value, improved classification accuracy, and robustness of our new approach. We separated those features based on if they required ground truthing of exams. Meanwhile, ground truthing of medical exams is costly and time-consuming as it involves manual annotation by experts, which also limits the scalability of the method. In addition, reliance on ground truth annotations can introduce biases and errors, as the annotations themselves might vary between experts, which limits applicability and transferability to real-world applications. Consequently, features that required knowledge of the image abnormalities' ground truth location (ie, area of interest) were removed: fixation duration on the area of interest, number of fixations on the area of interest, and the time between trial start and the first fixation on the area of interest. This ensures that the proposed approach can be more easily and widely applied in clinical settings without the need for extensive preparatory work. It also helps mitigate potential variability and subjectivity in the training data, leading to more robust and generalizable models. Furthermore, by excluding features tied to known locations abnormally, we can better capture inherent differences in visual search patterns between experienced and less experienced radiologists.

Table 1 summarizes the remaining attribute names, descriptions, and expected association with levels of expertise. All features were used as originally defined except for coverage. Salient regions refer to areas of an image that are not part of a peripheral black background. This is typically necessary because users may be viewing scans with different amounts of background area. As noted previously, we used the Tobii gaze data without fixation postprocessing. For evaluating traditional features using fixations in the Tobii dataset, we substituted raw gaze data with fixation data. For purposes of clarity and brevity, we use fixations and gaze interchangeably for the remainder of the paper.

Proposed Approach: Discretized Vector Encoding for Fixation Data

Here, we describe the proposed method for directly using the fixation patterns as an alternative approach to using the current and previously described attributes in Table 1. The proposed strategy aims to extract information from fixations in the following 2 ways: first, geometric alignment: this involves mapping the coordinates of eye fixations on the chest x-ray images into a Cartesian grid. Each fixation is assigned to a specific grid cell based on its position, such as the Cartesian locations of the fixations when displayed on a chest x-ray image. Second, temporal order in which the fixations appear: the order in which fixations occur is crucial. Fixations are split into temporal bins, preserving the sequence of visual inspection. For each trial with recorded fixation data, we split the fixations into t number of temporal bins (each bin covers "total time divided by the number of bins in seconds") or groups before counting the number of fixations captured within square grids or subdivisions of size x. Then, the t number of x-by-x grids is encoded into a single vector of size 1-by-$(x \times x \times t)$. The overall procedure is described in pseudocode in algorithm 1 (Textbox 1) and illustrated in Figure 3.

XSL•FO
RenderX

Textbox 1. Discrete vector encoding for fixation data.

Algorithm 1: Vector encoding for fixation data

- Input
- n-Fixation coordinates of a single trial, F[n×2], number of x- and y-axis subdivisions, (x,y), number of temporal groups, (t)
- Output
- Encoded vector, V[1×(t×x×y)]
- Initialize
- Create array A[t×x×y] and centroids C[(x×y)×2] corresponding to the center of each grid subdivision (defined by second and third indices of A
- Evenly split fixations into t-groups, T = ([F1, F2,]1, , [, F(n-1) Fn]t)
- Procedure
- For $i = 1 \rightarrow >t$ do:

  $f = T_i$

  For $j = 1 \rightarrow len\ (f)$ do:

  $C^* = argmin\ (\| C - f_j \|)$

  $A\ [i, C_x^*, C_y^*] \mathrel{+}= 1$

  $V = vec(A)$

  Return V

Figure 3. Proposed discretized vector encoding for fixation data. Bins 1, 2, and 3 capture fixations in a preserved spatial dimension across different temporal windows. Each row represents a temporal bin, and within each bin, the chest x-ray image is divided into spatial grids. The fixations are counted within each grid cell, providing a detailed representation of the radiologist's visual search pattern over time.

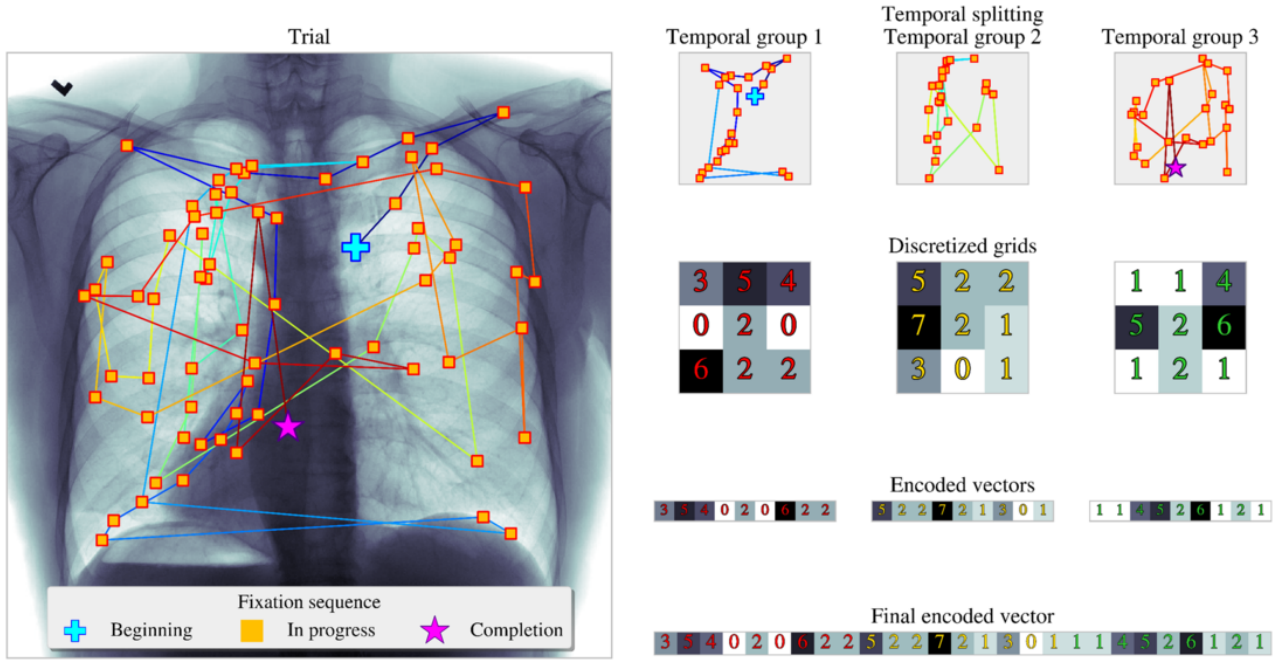

In Figure 3, participant 1 inspects a single chest x-ray image; the processed fixations are captured as illustrated gold squares with red edges, with the first fixation labeled as a blue cross and the last fixation visualized as a magenta star. In this example, the fixations are split into 3 temporal segments (Step 1), in which 3-by-3 grids count the number of fixations within them (Step 2). Then, the proposed algorithm outputs the final encoding vector as the flattening and concatenation of the set of 3-by-3 grids (Step 3). For a given (square) grid of size *x* and *t* number of temporal segments, the final yielded output vector is of length, regardless of trial temporal duration. Segmenting the raw data into fixed temporal segments is one of the benefits of this approach and a strategy developed and imposed to generate consistent numbers of variables on the encoding output across different trials. As the number of fixations across each trial can vary between participants, fixing the number of temporal segments allows the capturing of trial duration while conforming to a prescribed number of grid subdivisions and

XSL•FO
RenderX

temporal groups. For example, with Figure 3 as a reference, the final encoded vector will yield a vector with larger values therein for longer trials and yield a sparse vector (with lower values or values of zero) for shorter trials. Users can increase the fidelity of the grid and the number of time groups to represent a continuous spatiotemporal domain more closely. It is notable that the proposed methodology possesses the capability for tensor configuration for use in deep learning architecture by using *t* layers of grids. This tensor configuration is not studied in the paper due to the small sample size. The introduced technique is designed with more accessible or simpler classifiers in mind.

Performance Metric and Simulation Setup

To evaluate the discriminability of participants using the proposed approach, we use a stratified-fold cross-validation to calculate the AUC metric for several classification models, where each of the folds contains 5 trials from both levels of experience as the hold-out set. The study was performed on the data acquired by the EyeLink and Tobii equipment separately, and the following sections will contain an elaboration on their respective results. We performed cross-validations on a full-factorial configuration of 5, 7, 10, and 15 square grid subdivisions and 3, 5, 10, and 20 temporal groupings and selected the settings for each classifier that yielded the best results. In the presentation of these results, the average scores were calculated by computing the AUC metrics at the lowest level (data acquisition method, classifier, data type, feature extraction method, grid-size, temporal-group, and cross-validation seed) and averaged to the presented levels of granularity. Given the small sample size of 110 (EyeLink dataset) and 216 (Tobii dataset) trials, and high dimensionality in the chosen configurations (up to 4500 encoded variables in our study), there are available pathways that we have used to alleviate the effects of the curse of dimensionality present [28], such as principal component analysis (PCA) [29] and kernel principal component analysis (KPCA) [30]. The feature extraction and dimensionality reduction methods used include reducing the input data to 2 dimensions (with varying amounts of explained variance) and fixing the amount of variance explained to 50%, 90%, and 99% (with varying numbers of dimensions). These techniques were used not only to reduce the density of the data but also to introduce an additional preprocessing step that leverages the spectral decomposition of data collected from each participant.

Some of the major reasons for considering PCA and KPCA instead of the other alternatives include the following: PCA and KPCA are among the most popular method of dimensionality reduction; most technical practitioners, especially in the field of medicine, are familiar with PCA and KPCA; PCA and KPCA have rigorous mathematical properties and commonly used baseline methods in statistical analysis; and PCA and KPCA have relatively low computational complexity compared with many of the other shallow and deep alternatives.

All the codes were written using Python. The used libraries and versions are as follows: *matplotlib* (3.7.1), *seaborn* (0.12.2), *tqdm* (4.65.0), *scipy* (1.8.0), *scikit-learn* (1.0.2), *xgboost* (1.7.5), *GPy* (1.10.0), *numpy* (1.21.6), *pandas* (2.0.1), and *joblib* (1.2.0).

Ethical Considerations

This study was conducted in full compliance with human participant research ethics and was reviewed and approved by the University of Texas Health San Antonio Institutional Review Board (20190533HU). All participants were fully informed about the purpose and procedures of the study, and informed consent was obtained before their inclusion. To ensure the privacy and confidentiality of participant data, all identifying information was removed to anonymize the dataset before analysis. Furthermore, participants were compensated $400 USD for their time and involvement in the study.

Results

Competing Algorithms and Training

In this study, we use the Gaussian process, logistic regression, and k-nearest neighbors classifiers from the *Scikit-learn* [31] package; the extreme gradient boosting (XGBoost) [32] tree-based ensemble classifier; and a modified AlexNet [33] deep learning classifier. The *Scikit-learn* classifiers were selected for their accessibility to users, while the XGBoost and AlexNet-like neural networks were chosen as more complex classifiers. The logistic regression, k-nearest neighbors, and XGBoost classifiers used *Scikit-learn*'s *StratifiedKFold* and *GridSearchCV* packages to train on the balanced accuracy loss function (also defined by *Scikit-learn*), while the Gaussian process methodology used Laplace approximation as detailed in their documentation [31]. Finally, the AlexNet-like classifier used sparse categorical cross-entropy [34] for training.

EyeLink Dataset

Figure 4 illustrates the average AUC across each classifier tasked with distinguishing between 2 participants (particularly between 2 levels of experience) using either the traditional or the proposed encoded data types (features). Along with the original data types, we include the average AUC of the classifiers based on the usage of select feature extraction configurations. The encoded features extracted from the raw dataset, shown on the left, illustrate a consistently high AUC score compared with the traditional features shown on the right, implying that the model performance for each classifier (except for certain feature extraction configurations of the AlexNet model) has high discriminatory power under optimal spatiotemporal encoding settings.

XSL•FO
RenderX

Figure 4. Numerical study results on the area under the curve metric reported for each classifier when consuming the EyeLink dataset, organized by the aggregated average of classifier, data type, and select feature extraction levels using the original dataset, principal component analysis (PCA), and kernel principal component analysis (KPCA). Alex: AlexNet-like neural network classifier; Average: average of all classifiers; GP: Gaussian process; KNN: k-nearest neighbors; LR: logistic regression; XGBoost: extreme gradient boosting;.

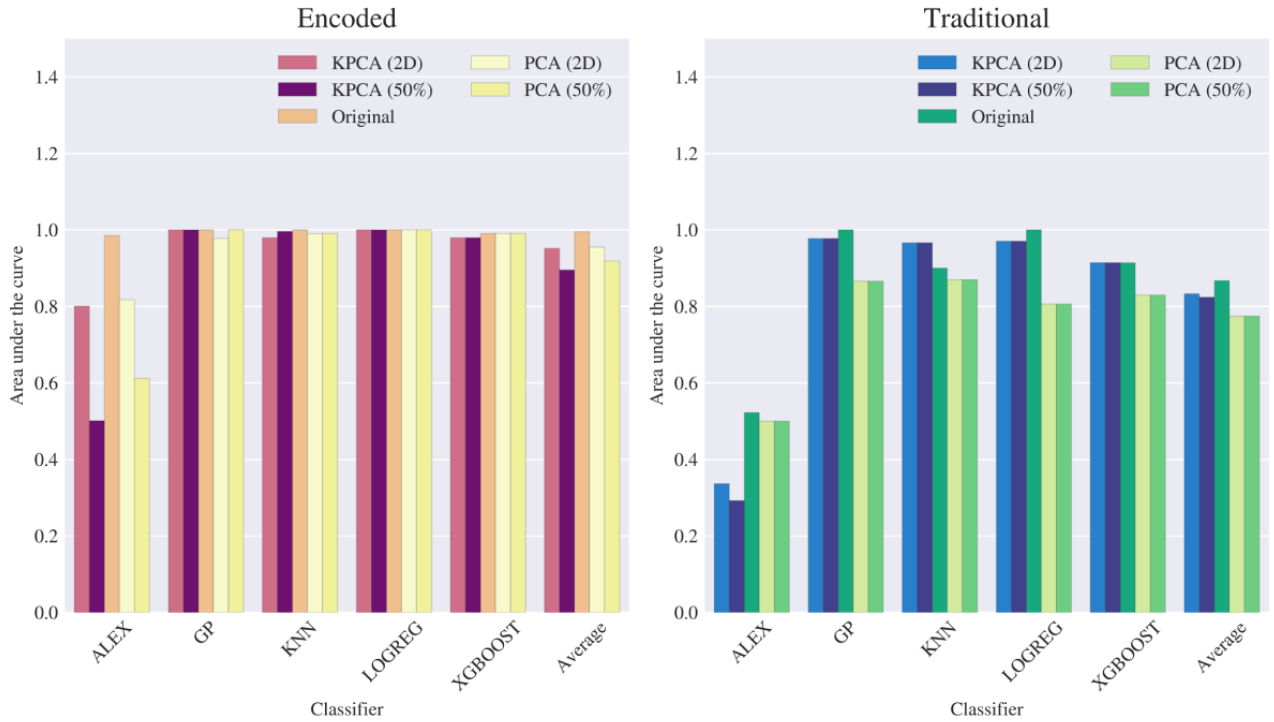

We also present the AUC metrics from Figure 4 below in Table 2. The performance of the classifiers using the encoded data type consistently yielded higher discriminatory power than those using the traditional data type across all feature extraction methods. Encoding the fixation data into the proposed spatiotemporal elements provides more information each classifier can use to determine the experience level of a given participant more effectively than using traditional attributes. This table illustrates the original encoded data to possess the highest performance, with AUC scores consistently above 0.98 across all classifiers. However, usage of the traditional data yields roughly 0.522 at worst, as seen in the reported results for the AlexNet classifier. This trend of encoded data providing better results is also seen when using feature extraction; although a performance decrease is observable when reducing dimensions either through an information covariance matrix (PCA) or spatial relation (KPCA), the use of encoded data still outperforms those corresponding to the use of the traditional data. This suggests that the loss in information due to dimensionality reduction can be considered negligible in light of the benefits of using spatiotemporal encoding. The lower relative performance of the AlexNet-like classifier is likely caused by the number of training samples available in this study. The report on AUC in the table for the classifier is higher for the encoded data type, where it is observable that using the data without dimensionality reduction provides the best performance. This effect has been studied in Sumner and Alaeddini [35], in which neural networks already perform feature extraction throughout each present layer; this supportively evidences the reported results here, whereas (besides the small dataset) performing feature extraction beforehand may not provide enough information for the network to use its architecture to its fullest potential.

XSL•FO
RenderX

Table 2. Numerical tabulation of area under the curve scores across each classifier and data type and select feature extraction methods.

Feature extraction method and data type	Classifier for the EyeLink dataset, AUC[a]				
	AlexNet	Gaussian process	KNN[b]	LR[c]	XGBoost[d]
KPCA[e] (2D)					
Encoded	0.801	1.000	0.980	1.000	0.980
Traditional	0.336	0.978	0.967	0.970	0.914
KPCA (50%)					
Encoded	0.501	1.000	0.996	1.000	0.980
Traditional	0.292	0.978	0.967	0.970	0.914
Original					
Encoded	0.985	1.000	1.000	1.000	0.991
Traditional	0.522	1.000	0.900	1.000	0.914
PCA[f] (2D)					
Encoded	0.818	0.978	0.991	1.000	0.991
Traditional	0.500	0.866	0.870	0.806	0.830
PCA (50%)					
Encoded	0.611	1.000	0.991	1.000	0.991
Traditional	0.500	0.866	0.870	0.806	0.830

[a]AUC: area under the curve.
[b]KNN: k-nearest neighbor.
[c]LR: logistic regression.
[d]XGBoost: extreme gradient boosting.
[e]KPCA: kernel principal component analysis.
[f]PCA: principal component analysis.

By using the encoded vectors for classification, differences in eye-tracking patterns can more consistently be distinguished between the 2 participants. Figures 5 and 6 illustrate one such difference in search pattern behavior. The more experienced participant (participant 1, Figure 5) shows a more uniformly distributed search pattern across the chest x-ray. In contrast, the less-experienced participant (participant 2, Figure 6) focuses on regions where they suspect abnormalities. It is clear from a visual inspection that the behavior between these participants is markedly different and using the correct spatiotemporal configurations to capture the differences between the 2 participant's behavior by leveraging the proposed methodology (as reported numerically in Table 2) provides a consistent improvement of classification accuracy.

XSL•FO
RenderX

Figure 5. Scan of chest x-ray by participant 1 (faculty).

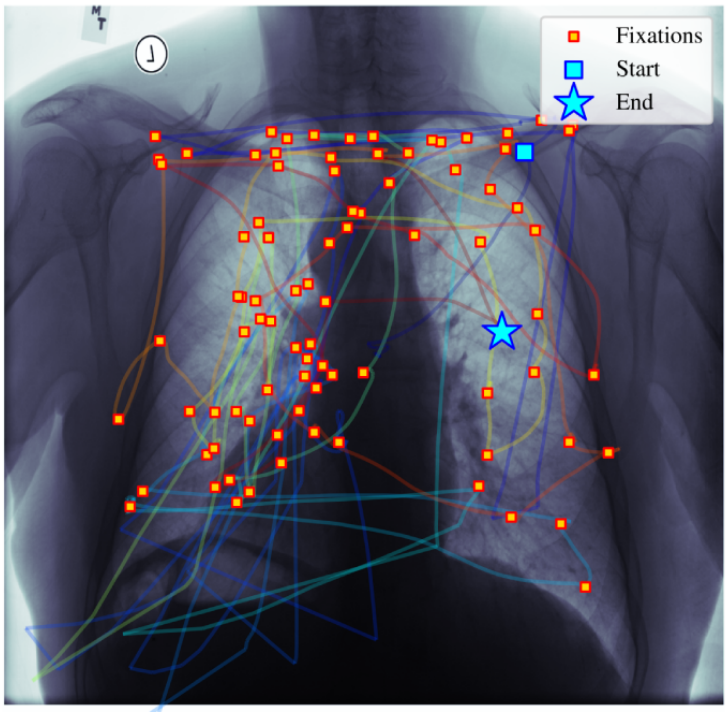

Figure 6. Scan of chest x-ray by participant 2 (trainee).

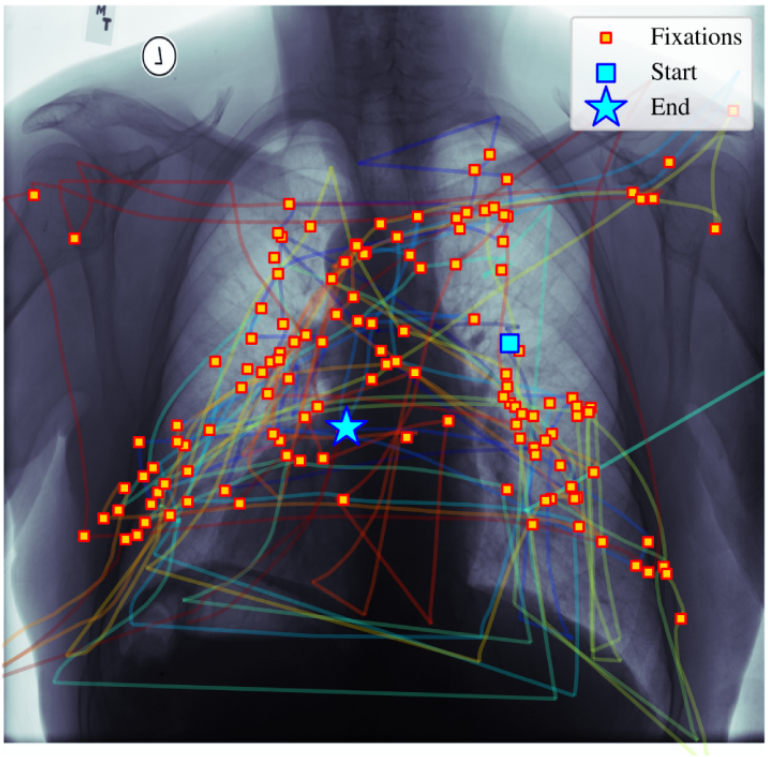

Tobii Dataset

We have performed the same analysis on the data acquired using Tobii eye-tracking equipment. It is notable that although the AUC scores from the EyeLink dataset are consistently high, natural anticipations allow one to observe more variation in classifier performance when more individual participants (classified as either a more-experienced faculty or less-experienced trainee) are introduced to the study. Figure 7 illustrates a report on AUC in a similar fashion to that in Figure 4, with lower scores across all classifying models for both data types. As seen in Figure 4, Figure 7 also suggests that the best performance for the encoded data on average is attained when using it without feature extraction, although, for several cases, we can observe that some form of feature extraction yields better results than their respective traditional dataset counterparts.

XSL•FO
RenderX

Figure 7. Numerical study results on the area under the curve metric reported for each classifier when consuming the Tobii dataset, organized by the aggregated average of classifier, data type, and selected feature extraction levels using the original dataset, principal component analysis (PCA), and kernel principal component analysis (KPCA). Alex: AlexNet-like neural network classifier; Average: average of all classifiers; GP: Gaussian process; KNN: k-nearest neighbors; LR: logistic regression; XGBoost: extreme gradient boosting;.

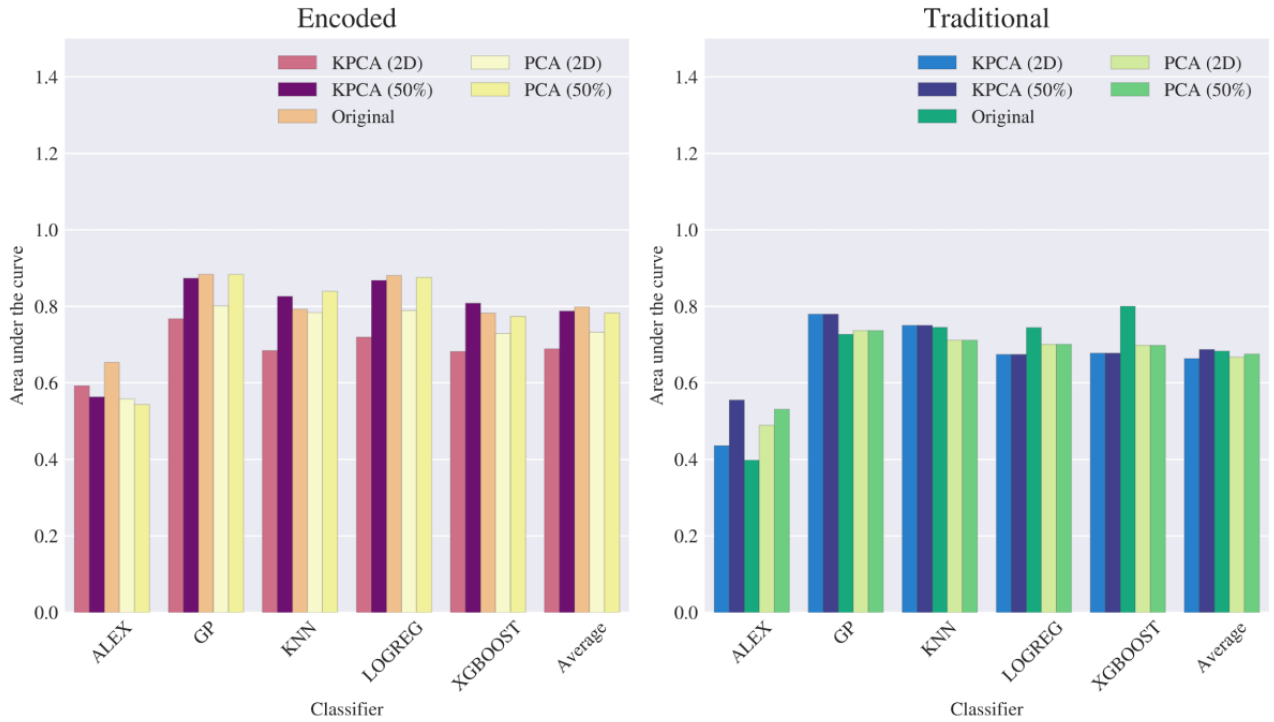

When inspecting Tables 3, we can numerically inspect the average (and variance of) AUC, F_1-score, accuracy, specificity, and sensitivity of each classifier when consuming each data type in both datasets. Within the Tobii dataset, the encoded data type generally outperformed (shown in italics) the traditional data type across most metrics and models. Although the encoded data type that was consumed within the Tobii dataset possessed more discriminatory capability than that in the traditional data, the performance gap was less pronounced than those observable in the EyeLink dataset. For example, the Tobii average (and variance) AUC scores for the encoded data type ranged from 0.55 (0.05) to 0.82 (0.04), while the traditional data type ranged from 0.51 (0.07) to 0.76 (0.05), and the EyeLink average (variance) AUC scores for the same data types ranged from 0.63 (0.07) to 1.0 (0.0) and from 0.52 (0.08) to 0.96 (0.01), respectively. There is a consistent trend across datasets that support the encoded data are capable of providing higher values of accuracy and performance; the F_1-score for the Gaussian process, k-nearest neighbors, logistic regression, and XGBoost were consistently higher when using the encoded data than when using traditional attributes in classification. This highlights the ability of the proposed encoding procedure to improve the balance between precision and recall in the classifiers and, as a result, the overall effectiveness of each model's predictions. In terms of specificity, the encoded data type is also shown to have a competitive edge in boosting a classifier's ability to correctly identify true negative class labels (experienced participants). As seen in Table 3, the average range of improvement lies between 0.01 to 0.04 for the EyeLink dataset and between –0.2 to +0.15 for the Tobii dataset; the negative value of the improvement is seen with the AlexNet-like model, which, as explained before, may have difficulty fitting well for classification on small datasets, made more difficult by the variation in subconscious behavior between participants that are recorded in spatiotemporal encodings by the proposed methodology.

XSL•FO
RenderX

Table 3. Numerical tabulation of mean and variance area under the curve, F_1-score, accuracy, specificity, and sensitivity across each data acquisition method, classifier, and data type.

Metric and classifier	Data acquisition system			
	EyeLink		Tobii	
	Data type		Data type	
	Encoded, average (SD)	Traditional, average (SD)	Encoded, average (SD)	Traditional, average (SD)
Area under the curve				
AlexNet	*0.63 (0.07)* [a]	0.52 (0.08)	*0.55 (0.05)*	0.51 (0.07)
GP[b]	*1 (0)*	0.96 (0.01)	*0.82 (0.04)*	0.76 (0.05)
KNN[c]	*0.96 (0.01)*	0.93 (0.01)	0.73 (0.05)	*0.74 (0.04)*
LR[d]	*1 (0)*	0.93 (0.01)	*0.82 (0.04)*	0.71 (0.08)
XGBoost[e]	*0.97 (0.0)*	0.91 (0.01)	*0.73 (0.05)*	0.71 (0.05)
F_1-score				
AlexNet	*0.43 (0.1)*	0.41 (0.09)	*0.39 (0.1)*	0.23 (0.1)
GP	*0.98 (0)*	0.90 (0.03)	*0.73 (0.07)*	0.71 (0.05)
KNN	*0.90 (0.03)*	0.88 (0.02)	0.61 (0.1)	*0.71 (0.04)*
LR	*0.98 (0)*	0.82 (0.04)	*0.74 (0.06)*	0.58 (0.09)
XGBoost	*0.96 (0)*	0.87 (0.02)	*0.68 (0.07)*	0.67 (0.05)
Accuracy				
AlexNet	*0.52 (0.02)*	0.49 (0.04)	*0.51 (0.0)*	0.5 (0.01)
GP	*0.99 (0.0)*	0.93 (0.01)	*0.76 (0.03)*	0.7 (0.03)
KNN	*0.93 (0.01)*	0.9 (0.01)	0.69 (0.03)	*0.7 (0.02)*
LR	*0.98 (0.0)*	0.88 (0.01)	*0.77 (0.03)*	0.64 (0.03)
XGBoost	*0.96 (0.0)*	0.9 (0.01)	*0.72 (0.03)*	0.67 (0.03)
Specificity				
AlexNet	*0.43 (0.22)*	0.42 (0.22)	0.46 (0.16)	*0.66 (0.12)*
GP	*0.98 (0)*	0.97 (0.01)	*0.80 (0.04)*	0.65 (0.07)
KNN	*0.97 (0.01)*	0.93 (0.01)	*0.76 (0.06)*	0.67 (0.04)
LR	*0.99 (0.0)*	0.97 (0.01)	*0.81 (0.04)*	0.67 (0.08)
XGBoost	*0.95 (0.01)*	0.92 (0.01)	*0.76 (0.05)*	0.63 (0.08)
Sensitivity				
AlexNet	*0.64 (0.22)*	0.57 (0.2)	*0.56 (0.17)*	0.32 (0.12)
GP	*0.99 (0)*	0.89 (0.04)	0.72 (0.07)	*0.76 (0.05)*
KNN	*0.89 (0.05)*	0.87 (0.03)	0.61 (0.1)	*0.74 (0.04)*
LR	*0.98 (0.01)*	0.77 (0.06)	*0.72 (0.06)*	0.6 (0.09)
XGBoost	*0.98 (0)*	0.86 (0.03)	0.68 (0.07)	*0.71 (0.05)*

[a]Superior values are italicized.
[b]GP: Gaussian process.
[c]KNN: k-nearest neighbor.
[d]LR: logistic regression.
[e]XGBoost: extreme gradient boosting.

XSL•FO
RenderX

Discussion

Principal Findings

In this study, we demonstrated the capacity of eye-tracking technology, combined with machine learning algorithms, to discriminate between radiologists' experience levels. For this purpose, we developed a novel feature encoding technique where fixations are first spatially arranged according to their Cartesian coordinates on chest x-ray images and temporally ordered. The fixations are then subdivided into predefined temporal bins, and within each bin, we count the number of eye fixations within each subdivision. These counted bins are then concatenated to form a vector encoding to be used as feature input for machine learning algorithms. Our experiments showed that the discretized vector encoding creates discriminative features that are not captured by conventional metrics. Using the encoding approach allows classifiers to better distinguish between participants in terms of experience level, which highlights performance gains (when compared with using traditional features for discrimination) of 6.9%, 7.11%, 9.14%, 9.59%, and 5.61% for AUC, accuracy, F_1-score, sensitivity, and specificity, respectively, aggregated across both EyeLink and Tobii datasets in Table 3. The Tobii dataset exhibits a lower performance gain (6.41%, 7.48%, 8.62%, 5.11%, and 9.45%) than observed using the EyeLink dataset (7.29%, 6.83%, 9.54%, and 13.13%) due to using a more diverse roster of participants; however, the trend in using the proposed eye-tracking encoding approach possessing the competitive edge is still present, highlighting the effectiveness of spatiotemporal assortment in the introduced method. These results validate our initial hypothesis that when appropriately encoded, eye-tracking data can provide nuanced insights into the difference between radiologist's expertise levels.

We can also observe the perceptual strategies radiologists use during diagnostic evaluations. Previous research has often focused on more general eye-tracking metrics without leveraging the full potential of machine learning to analyze the data. For example, studies by van der Gijp et al [11] and Waite et al [3] explored how visual search patterns correlate with diagnostic accuracy and expertise. With the help of the proposed encoding method, such machine learning models can be developed to determine expertise level and has the potential to identify and track potential features from eye fixations or gaze fixations.

Limitations

While our study has shown promising results and potential benefits, it is important to acknowledge limitations that may have a degree of effect on our findings. One such limitation is sample size; across both the EyeLink and Tobii datasets, there were 2 participants in one study (EyeLink) and 8 participants in the other (Tobii), with both containing small numbers of images scanned by each participant. Another condition involves the variation in data acquisition. In total, 2 different eye-tracking devices (EyeLink and Tobii) were used for data collection, and while serving the same overarching purpose of collecting data, some additional variability in the findings are notably attributed to the usage of 2 different hardware-software configurations. Another important consideration includes the difference in traditional feature sets between the EyeLink and Tobii datasets. Coupled with data acquisition differences, some features from the EyeLink software were not congruent with the Tobii dataset, such as the usage of fixations (EyeLink) versus gaze (Tobii). When applying the encoding approach to these datasets, the Tobii dataset had larger yielded values in each output vector. This did not affect the results substantially; however, it underscores the challenge of directly comparing data from 2 sources. One final consideration was our decision to remove certain metrics related to the location of abnormalities in chest x-rays as features in the traditional data type during performance evaluation. For example, we did not consider the time to first fixation on the region of abnormality. This and other like attributes possess statistical significance in previous works; however, their inclusion necessitates extensive labeling, validation, and other processing in order to establish ground truth information for each image scanned by each participant.

Conclusions

Despite the limitations above, this study holds significant promise and offers a range of benefits worthy of attention and consideration for use. By extracting spatiotemporal features from eye-tracking data, the proposed approach has demonstrated the capacity to differentiate users based on eye-tracking behavior alone instead of traditional methods and can be extended for use in fields ranging from medical to educational. The approach enables discriminability between users and offers a pathway to gaining deeper insights into generalized differences between levels of expertise. By exploring these pathways, this approach holds the potential to establish more effective educational programs that can assist users optimize their search patterns. Respective to the study conducted, by providing support to radiologists to find abnormalities quickly and accurately in chest x-rays, our approach seeks to reduce perceptual errors in medical diagnoses. In fields where the development of unique and precise search patterns is important, the proposed approach offers a valuable source of knowledge transfer. Transmission of expertise from more-experienced individuals to less-experienced individuals can be facilitated and result in increased streamlining during the learning process and yield more efficient and accurate search patterns. The potential benefits can apply to professionals and trainees or students alike.

In summary, we have shown the potential for spatiotemporal features extracted from eye-tracking data to be useful in discriminating between radiologists of different skill levels and opening the door to improving education. We plan to augment this research by increasing the number of radiologists to demonstrate generalizability and exploring additional types of spatiotemporal analyses. The implications of our findings extend beyond radiology, suggesting that similar methodologies could revolutionize training and assessment protocols in various fields that rely on visual cognition like aviation and ground transportation. Further research could explore the integration of these techniques into real-time training tools, potentially transforming educational paradigms in professions requiring visual expertise.

XSL•FO
RenderX

Acknowledgments

This research is funded by San Antonio Medical Foundation, and principal investigators KC (University of Texas Health Science Center at San Antonio) and EJG (University of Texas at San Antonio).

Data Availability

Data collected and analyzed in this study are available upon request. Interested researchers can obtain the data by completing a data request form, which includes their planned project outline and the details of the person responsible for handling the data. This process ensures that the data are used responsibly and for scientifically valid purposes. The request forms are reviewed on a case-by-case basis to ensure alignment with ethical and research standards. Images and annotations were obtained from the National Institutes of Health Clinical Center and can be downloaded on the web [36]. The Institutional Review Board restricts sharing of the gaze data. In case further information, please contact KC (clarkkl@uthscsa.edu).

Conflicts of Interest

KC holds equity in Zauron Labs Inc and NC is a co-owner of Zauron Labs.

References

1. An update on cancer deaths in the United States. Centers for Disease Control and Prevention. 2022 Feb 28. URL: https://stacks.cdc.gov/view/cdc/119728 [accessed 2022-02-28]
2. Screening for lung caner: U.S. Preventitive Services Task Force recommendation statement. Lung Cancer Research foundation. 2014. URL: https://www.lungcancerresearchfoundation.org/screening-for-lung-cancer-u-s-preventive-services-task-force-recommendation-statement/ [accessed 2024-09-23]
3. Waite S, Grigorian A, Alexander RG, Macknik SL, Carrasco M, Heeger DJ, et al. Analysis of perceptual expertise in radiology - current knowledge and a new perspective. Front Hum Neurosci 2019;13:213 [FREE Full text] [doi: 10.3389/fnhum.2019.00213] [Medline: 31293407]
4. Bruno MA, Walker EA, Abujudeh HH. Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction. Radiographics 2015;35(6):1668-1676. [doi: 10.1148/rg.2015150023] [Medline: 26466178]
5. Kundel HL. Perception errors in chest radiography. Seminars in Respiratory Medicine 1989;10(03):203-210. [doi: 10.1055/s-2007-1006173]
6. Goodman LR. Felson's Principles of Chest Roentgenology: A Programmed Text. 2nd ed. Radiology 2000 Mar;214(3):848. [doi: 10.1148/radiology.214.3.r00fe55848]
7. Tanoubi I, Tourangeau M, Sodoké K, Perron R, Drolet P, Bélanger ME, et al. Comparing the visual perception according to the performance using the eye-tracking technology in high-fidelity simulation settings. Behav Sci (Basel) 2021;11(3):31 [FREE Full text] [doi: 10.3390/bs11030031] [Medline: 33807673]
8. Wu CC, Wolfe JM. Eye movements in medical image perception: a selective review of past, present and future. Vision (Basel) 2019;3(2):32 [FREE Full text] [doi: 10.3390/vision3020032] [Medline: 31735833]
9. Kelly BS, Rainford LA, Darcy SP, Kavanagh EC, Toomey RJ. The development of expertise in radiology: in chest radiograph interpretation, "Expert" search pattern may predate "Expert" levels of diagnostic accuracy for pneumothorax identification. Radiology 2016;280(1):252-260. [doi: 10.1148/radiol.2016150409] [Medline: 27322975]
10. Wolfe JM, Lyu W, Dong J, Wu CC. What eye tracking can tell us about how radiologists use automated breast ultrasound. J Med Imaging (Bellingham) 2022;9(4):045502 [FREE Full text] [doi: 10.1117/1.JMI.9.4.045502] [Medline: 35911209]
11. van der Gijp A, Ravesloot CJ, Jarodzka H, van der Schaaf MF, van der Schaaf IC, van Schaik JPJ, et al. How visual search relates to visual diagnostic performance: a narrative systematic review of eye-tracking research in radiology. Adv Health Sci Educ Theory Pract 2017;22(3):765-787 [FREE Full text] [doi: 10.1007/s10459-016-9698-1] [Medline: 27436353]
12. Drew T, Evans K, Võ MLH, Jacobson FL, Wolfe JM. Informatics in radiology: what can you see in a single glance and how might this guide visual search in medical images? Radiographics 2013;33(1):263-274 [FREE Full text] [doi: 10.1148/rg.331125023] [Medline: 23104971]
13. Kundel HL, Nodine CF. Interpreting chest radiographs without visual search. Radiology 1975;116(3):527-532. [doi: 10.1148/116.3.527] [Medline: 125436]
14. Swensson RG. A two-stage detection model applied to skilled visual search by radiologists. Perception & Psychophysics 1980;27(1):11-16. [doi: 10.3758/bf03199899]
15. Lim JZ, Mountstephens J, Teo J. Eye-tracking feature extraction for biometric machine learning. Front Neurorobot 2022;15:796895 [FREE Full text] [doi: 10.3389/fnbot.2021.796895] [Medline: 35177973]
16. Shamyuktha RS, Amudha J, Aiswariya Milan K. A machine learning framework for classification of expert and non-experts radiologists using eye gaze data. : IEEE; 2022 Presented at: 2022 IEEE 7th International Conference on Recent Advances

XSL•FO
RenderX

and Innovations in Engineering (ICRAIE); December 01-03, 2022; Mangalore, India. [doi: 10.1109/icraie56454.2022.10054277]
17. Harezlak K, Kasprowska P, Kasprowska S. Eye movement traits in differentiating experts and laymen. 2017 Presented at: Man-Machine Interactions 5: 5th International Conference on Man-Machine Interactions, ICMMI 2017; October 3-6, 2017; Kraków, Poland. [doi: 10.1007/978-3-319-67792-7_9]
18. Akshay S, Megha YJ, Shetty CB. Machine learning algorithm to identify eye movement metrics using raw eye tracking data. : IEEE; 2020 Presented at: 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT); August 20-22, 2020; Tirunelveli, India p. 949-955. [doi: 10.1109/icssit48917.2020.9214290]
19. Rizzo A, Ermini S, Zanca D, Bernabini D, Rossi A. A machine learning approach for detecting cognitive interference based on eye-tracking data. Front Hum Neurosci 2022;16:806330 [FREE Full text] [doi: 10.3389/fnhum.2022.806330] [Medline: 35572006]
20. Öder M, Eraslan S, Yesilada Y. Automatically classifying familiar web users from eye-tracking data: a machine learning approach. Turkish Journal of Electrical Engineering and Computer Sciences 2022;30(1):233-248. [doi: 10.3906/elk-2103-6]
21. Karargyris A, Kashyap S, Lourentzou I, Wu JT, Sharma A, Tong M, et al. Creation and validation of a chest X-ray dataset with eye-tracking and report dictation for AI development. Sci Data 2021;8(1):92 [FREE Full text] [doi: 10.1038/s41597-021-00863-5] [Medline: 33767191]
22. Bigolin Lanfredi R, Zhang M, Auffermann WF, Chan J, Duong PAT, Srikumar V, et al. REFLACX, a dataset of reports and eye-tracking data for localization of abnormalities in chest x-rays. Sci Data 2022;9(1):350 [FREE Full text] [doi: 10.1038/s41597-022-01441-z] [Medline: 35717401]
23. Robinson KA. Dictionary of eye terminology. BMJ 1997;81(11) [FREE Full text] [doi: 10.1136/bjo.81.11.1021c]
24. Shiraishi J, Katsuragawa S, Ikezoe J, Matsumoto T, Kobayashi T, Komatsu K, et al. Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. AJR Am J Roentgenol 2000;174(1):71-74. [doi: 10.2214/ajr.174.1.1740071] [Medline: 10628457]
25. Nguyen HQ, Lam K, Le LT, Pham HH, Tran DQ, Nguyen DB, et al. VinDr-CXR: an open dataset of chest X-rays with radiologist's annotations. Sci Data 2022;9(1):429 [FREE Full text] [doi: 10.1038/s41597-022-01498-w] [Medline: 35858929]
26. EyeLink 1000 plus. SR-Research. 2017. URL: https://www.sr-research.com/eyelink-1000-plus/ [accessed 2024-09-18]
27. Tobii eye tracker 5. Tobii. 2020. URL: https://gaming.tobii.com/product/eye-tracker-5/ [accessed 2024-09-18]
28. Aggarwal CC, Hinneburg A, Keim DA. On the surprising behavior of distance metrics in high dimensional space. 2001 Presented at: Database Theory - ICDT 2001; January 4-6, 2001; London, United Kingdom p. 8. [doi: 10.1007/3-540-44503-x_27]
29. Wold S, Esbensen K, Geladi P. Principal component analysis. Chemometrics and Intelligent Laboratory Systems 1987;2(1-3):37-52. [doi: 10.1016/0169-7439(87)80084-9]
30. Schölkopf B, Smola A, Müller KR. Kernel principal component analysis. 1997 Presented at: Artificial Neural Networks — ICANN ’97: 7th International Conference; October 8–10, 1997; Lausanne, Switzerland. [doi: 10.1007/bfb0020217]
31. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in python. The Journal of Machine Learning Research 2011;12:2825-2830 [FREE Full text]
32. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. 2016 Presented at: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13-17, 2016; San Francisco, CA. [doi: 10.1145/2939672.2939785]
33. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun. ACM 2017;60(6):84-90. [doi: 10.1145/3065386]
34. Cybenko G, O'Leary DP, Rissanen J. The Mathematics of Information Coding, Extraction and Distribution. Germany: Springer Science & Business Media; 1998.
35. Summer J, Alaeddini A. Analysis of feature extraction methods for prediction of 30-day hospital readmissions. Methods of Information in Medicine 2019;58(6):213-221. [doi: 10.1055/s-0040-1702159] [Medline: 32349155]
36. CXR8. NIH Clinical Center Box. URL: https://nihcc.app.box.com/v/ChestXray-NIHCC [accessed 2024-12-02]

Abbreviations

AUC: area under the curve
IRB: institutional review board
PCA: principal component analysis
KPCA: kernel principal component analysis
XGBoost: extreme gradient boosting

XSL•FO
RenderX

Edited by A Mavragani; submitted 24.10.23; peer-reviewed by N Silva, M Hirota; comments to author 05.12.23; revised version received 31.05.24; accepted 22.06.24; published 22.01.25.

Please cite as:
Martinez S, Ramirez-Tamayo C, Akhter Faruqui SH, Clark K, Alaeddini A, Czarnek N, Aggarwal A, Emamzadeh S, Mock JR, Golob EJ
Discrimination of Radiologists' Experience Level Using Eye-Tracking Technology and Machine Learning: Case Study
JMIR Form Res 2025;9:e53928
URL: https://formative.jmir.org/2025/1/e53928
doi:10.2196/53928
PMID:

©Stanford Martinez, Carolina Ramirez-Tamayo, Syed Hasib Akhter Faruqui, Kal Clark, Adel Alaeddini, Nicholas Czarnek, Aarushi Aggarwal, Sahra Emamzadeh, Jeffrey R Mock, Edward J Golob. Originally published in JMIR Formative Research (https://formative.jmir.org), 22.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Use of Go-Beyond as a Self-Directed Internet-Based Program Supporting Veterans' Transition to Civilian Life: Preliminary Usability Study

Karolina Katarzyna Alichniewicz[1,2], PhD; Sarah Hampton[1,3], DPsych(Clin); Madeline Romaniuk[2], DPsych(Clin); Darcy Bennett[1], BPsych(Hons); Camila Guindalini[1], PhD

[1]Greenslopes Private Hospital, Gallipoli Medical Research, Brisbane, Australia

[2]The University of Queensland, Faculty of Health and Behavioural Sciences, Brisbane, Australia

[3]School of Medicine, The University of Queensland, Brisbane, Australia

Corresponding Author:
Karolina Katarzyna Alichniewicz, PhD
Greenslopes Private Hospital
Gallipoli Medical Research
Newdegate Street
Greenslopes
Brisbane, QLD 4120
Australia
Phone: 61 7 33947613
Fax: 61 7 33297744
Email: alichniewiczkarolina@gallipoliresearch.org.au

Abstract

Background: The transition from military service to civilian life presents a variety of challenges for veterans, influenced by individual factors such as premilitary life, length of service, and deployment history. Mental health issues, physical injuries, difficulties in relationships, and identity loss compound the reintegration process. To address these challenges, various face-to-face and internet-based programs are available yet underused. This paper presents the preliminary evaluation of "Go-Beyond, Navigating Life Beyond Service," an internet-based psychoeducational program for veterans.

Objective: The study aims to identify the reach, adoption, and engagement with the program and to generate future recommendations to enhance its overall impact.

Methods: This study exclusively used data that were automatically and routinely collected from the start of the Go-Beyond program's launch on May 24, 2021, until May 7, 2023. When accessing the Go-Beyond website, veterans were asked to complete the Military-Civilian Adjustment and Reintegration Measure (M-CARM) questionnaire, which produces a unique M-CARM profile of results specifying potential areas of need on the 5 domains of the measure. Users were then automatically allocated to Go-Beyond modules that aligned with their M-CARM profile. Additionally, quantitative and qualitative data were collected from a survey on aesthetics, interactivity, user journey, and user experience, which was optional for users to complete at the end of each module.

Results: Results show a conversion rate of 28.5% (273/959) from the M-CARM survey to the Go-Beyond program. This rate is notably higher compared with similar internet-based self-help programs, such as VetChange (1033/22,087, 4.7%) and resources for gambling behavior (5652/8083, 14%), but lower than the MoodGYM program (82,159/194,840, 42.2%). However, these comparisons should be interpreted with caution due to the limited availability of published conversion rates and varying definitions of uptake and adoption across studies. Additionally, individuals were 1.64 (95% CI 1.17-2.28) more likely to enroll when they express a need in Purpose and Connection, and they were 1.50 (95% CI 1.06-2.18) times more likely to enroll when they express the need Beliefs About Civilians, compared with those without these needs. The overall completion rate for the program was 31% (85/273) and modules' individual completion rates varied from 8.4% (17/203) to 20% (41/206). Feedback survey revealed high overall user satisfaction with Go-Beyond, emphasizing its engaging content and user-friendly modules. Notably, 94% (88/94) of survey respondents indicated they would recommend the program to other veterans, family, or friends.

Conclusions: The Go-Beyond program may offer promising support for veterans transitioning to civilian life through digital technology. Our study reveals insights on user engagement and adoption, emphasizing the need for ongoing evaluation to further

XSL•FO
RenderX

address the diverse needs of military personnel. Future research should explore predictors of engagement, the addition of peer or facilitator support, and the use of outcome measures for effectiveness assessment.

(JMIR Form Res 2025;9:e60868) doi:10.2196/60868

KEYWORDS

military transition; web-based interventions; military-civilian adjustment; Go-Beyond; internet-based program; civilian; military service; veteran; premilitary life; mental health issues; physical injuries; adoption; quantitative analysis; survey; family; support; digital technology; user engagement; effectiveness; assessment

Introduction

The transition from military service to civilian life is a multifaceted process that can present challenges for veterans [1-3]. Upon leaving the military, veterans often face a period of adjustment as they navigate a vastly different sociocultural and professional landscape [4-6]. The process of transitioning from military service to civilian life is not uniform, as each veteran's experience is shaped by individual factors such as premilitary life, length of service, deployment history, and personal resilience [7]. While many individuals successfully navigate this change, a substantial proportion experience difficulty in adjusting to civilian life, with these difficulties persisting over the years [1,2,4].

In the literature, there is recognition of various challenges faced by military personnel when they transition into civilian life. Among these challenges, mental health issues such as posttraumatic stress disorder (PTSD), substance use disorders, depression, and anxiety are frequently reported [1]. In addition, physical injuries, including chronic pain and traumatic brain injuries, play a significant role [8]. Furthermore, difficulties in establishing and maintaining friendships and family relationships [4,5], loss of identity and purpose [2,3,9], disparities between military and civilian cultures, as well as employment factors [6] are all critical elements. Collectively, these multiple and dynamic factors have a significant impact on the reintegration of veterans into civilian life. It is important to note that the presence of hardships in various domains can compound the effects and exacerbate functional impairments beyond those caused by individual stressors alone [10]. Consequently, understanding the factors that influence transition and identifying effective accessible interventions is critical in supporting military personnel during this crucial phase.

Internet-based approaches can complement and expand the support systems available to service members and veterans [11-13]. By harnessing the potential of digital technology, personalized and accessible interventions that cater to the diverse needs of military personnel during their transition to civilian life can be provided. Moreover, internet-based interventions could present an effective way to overcome some of the barriers to care, such as long waiting periods, stigma, or lack of access to treatment due to remote location [12]. Internet-based interventions hold significant potential in supporting veterans during the critical transitioning phase. However, ensuring their uptake and effectiveness requires robust evaluations, covering aspects like utility, user satisfaction, usability, acceptability, and potential in addressing the diverse needs of transitioning military personnel [14]. Through such evaluations, policy makers, researchers, and support organizations can gain valuable insights into the most beneficial strategies for veterans [15]. Furthermore, a comprehensive assessment of internet-based interventions may enable tailored interventions, improvement of support mechanisms, and the development of evidence-based practices [16]. Therefore, understanding the impact of internet-based interventions is paramount in enhancing the well-being and successful integration of military veterans into civilian society.

In this paper, we present the preliminary evaluation of an internet-based psychoeducational program "Go-Beyond, Navigating Life Beyond Service" designed for veterans who are in the process of reintegration to civilian life. The aim of the study was to perform a user-centered evaluation of the Go-Beyond program using user engagement data routinely collected and assess the reach, adoption, engagement, and completion rates. Furthermore, the study aimed to gather user feedback to enhance the concept version of the program's modules, identify opportunities for further improvement, and generate future recommendations.

Methods

Program Description

"Go-Beyond, Navigating Life Beyond Service" is an evidence-informed, self-directed, web-based program designed to support veterans' adjustment to civilian life after separation from military service. The program was developed by Gallipoli Medical Research (GMR), a not-for-profit organization that has the mission to improve the quality of life for veterans and their families through research, translation, and implementation, in partnership with Returned and Services League of Australia Queensland (RSLQ), and was officially launched on May 24, 2021. Over the 24 months after the launch, all interested veterans were provided unlimited access to the Go-Beyond program. Former service members were made aware of the Go-Beyond program through a national implementation task force initiative that included flyers, posters, presentations, and a social media campaign and was conducted by GMR, RSLQ, Department of Veteran Affairs, Open Arms, and Commonwealth Super Corporation. The program was also promoted to various ex-serving organizations. Furthermore, Go-Beyond was accompanied by an applied practitioner portal that provided training on the use of the program and targeted clinicians, especially rehab providers. The implementation of the Go-Beyond program was part of a larger research agreement with RSLQ, and the associated costs for its development and preliminary implementation have been covered by the funding organization. This funding covered essential elements such as

XSL•FO
RenderX

platform design, content creation, technical support, and user testing to ensure the program's accessibility for veterans. In addition, future strategies are being developed to ensure the program's sustainability beyond the funding period in Australia.

When accessing the Go-Beyond website, veterans are asked to complete the Military-Civilian Adjustment and Reintegration Measure (M-CARM) questionnaire, which produces a unique M-CARM profile of results specifying potential areas of need on the 5 factors of the measure. The M-CARM is a valid and reliable self-report questionnaire designed to assess adjustment and reintegration to civilian life after service. It contains 21 questions, rated from 1 to 5, and takes approximately 5 minutes to complete. The measure is freely available on a dedicated website [17]. Users are automatically allocated to Go-Beyond modules that align with their M-CARM profile.

The Go-Beyond program aims to (1) provide education about topics that are important for psychological adjustment and reintegration to civilian life; (2) build psychological insight into personal challenges adjusting to civilian life; (3) offer practical first steps and tools to facilitate stronger adjustment and reintegration; and (4) provide links to additional services and support available in Australia.

The Go-Beyond program was developed by 2 clinical psychologists at GMR who have both research and clinical experience with ex-service personnel in Australia. The content of the program is based on previous research describing empirically derived psychological and cultural factors associated with adjustment and reintegration to civilian life, including (1) Purpose and Connection, (2) Help Seeking, (3) Beliefs About Civilians, (4) Resentment and Regret, and (5) Regimentation [18]. These factors were derived as part of the development of the M-CARM [18]. Go-Beyond consists of seven interactive modules entitled (1) Finding Purpose, (2) Help Seeking, (3) Regimentation, (4) Resentment and Regret, (5) Social Connection, (6) Beliefs About Civilians, and (7) an optional module available to anyone who signed up to Go-Beyond: (7) Effective Communication in Civilian Life. They are presented to participants in the above order. If a participant is not required to do a particular module, it gets omitted in that order. The session outline of the program is presented in Textbox 1. The program content uses evidence-based psychological therapeutic approaches including cognitive behavioral therapy and acceptance and commitment therapy. Both cognitive behavioral therapy and acceptance and commitment therapy have demonstrated transdiagnostic treatment efficacy within veteran populations [19]. Each module includes (1) psychoeducation for each topic; (2) hypothetical case studies and personal, video-recorded stories from Australian ex-service members; (3) personal reflection opportunities or written tasks; (4) helpful activities; and (5) direction to more specific services and resources for Australian ex-service personnel. The approximate time needed to complete each module is 45-60 minutes.

XSL•FO
RenderX

Textbox 1. Descriptions of Go-Beyond modules.

Beliefs About Civilians

- The module introduces the concept of beliefs, social identity theory, and common unhelpful beliefs about civilians that can hinder adjustment to civilian life. Participants engage in self-reflection to evaluate their beliefs and learn techniques to adapt them, including thought challenging, defusing exercises, and a "take your thought to court" activity. The importance of practicing these techniques is emphasized throughout.

Finding Purpose

- The module begins by exploring the importance of purpose, particularly in the military context, and the sense of loss that can occur during the transition to civilian life. It helps participants clarify their values, distinguish between values and goals, and engage in activities like the life compass, values-guided actions, and goal setting. The module also emphasizes finding purpose through volunteering or paid employment and provides information on services and resources that support veterans in securing meaningful work.

Help Seeking

- The module provides psychoeducation on the importance of help-seeking for mental health, addressing common barriers and highlighting the rates of help-seeking among Australian veterans. It covers how to find evidence-based treatments and recognize when help is needed. Through the "Staying Well" activity, participants learn to identify indicators of good mental health, and signs of decline, and create an action plan. Practical steps for accessing high-quality mental health support are also provided.

Social Connections

- This module covers the importance of social connections for health and well-being, focusing on their role in adjusting to transitions, especially for the "Defence family." It addresses common barriers such as cultural differences, misconceptions, and communication issues. Strategies for finding mateship outside the military, strengthening existing ties, and building new ones are discussed, alongside opportunities for social connection and volunteering.

Resentment and Regret

- This module explores resentment, covering its definition, functions, and impacts, and includes activities to identify underlying feelings, use self-compassion, and resolve related injustices. It also addresses regret, discussing its nature, impacts, and irrationality, with activities to manage regret by confronting hindsight bias, accepting imperfections, and clarifying values and goals.

Regimentation

- This module covers regimentation; its role in Defence; and its effects on personal life, family, and work. It includes activities to identify and evaluate regimentation's impact and emphasizes the importance of psychological flexibility and adaptability. Practical strategies are provided for embracing change, trying new things, and adjusting routines and thought processes.

Effective Communication in Civilian Life

- This module covers communication styles (aggressive, passive, passive-aggressive, and assertive) and includes a quiz to identify your style. It provides skills for meaningful conversations, including asking questions, self-disclosure, managing stereotypes, active listening, and overcoming barriers. The module also explores using compliments and empathy to build social connections.

Data Collection

This study exclusively used data that were automatically and routinely collected starting from the program's launch on May 24, 2021, until May 7, 2023, through websites [17] for M-CARM and [20] for Go-Beyond. Both websites were freely available to everyone, both veterans and civilian populations, in Australia. No specific inclusion or exclusion criteria were applied to this study. The reach of Go-Beyond was defined as the number of users who accessed and completed the M-CARM measure and the adoption was defined as the number of users who signed up to the Go-Beyond program. Completion was measured by the number of users who accessed and reached the end of the module, as tracked by the Absorb LMS system (Absorb Software Inc). The system records whether a user has started, is in progress, and has fully completed the module, including exiting the module properly. The modules were designed to be completed in 45-60 minutes.

In addition, quantitative and qualitative data were collected from a survey on aesthetics, interactivity, user journey, and user experience, which was optional for users to complete at the end of each module. The questions asked can be found in Table 1.

XSL•FO
RenderX

Table 1. User experience survey.

Question	Rating options
How did you find the duration of the module?	Option for comments
How easy was the Go-Beyond Platform to use? Were you able to navigate to what you needed? Please provide details.	Option for comments
How likely are you to implement changes in your life following this module?	Likert scale 1-5 • 1-Very Unlikely • 2-Unlikely • 3-Likely • 4-Very Likely • 5-Extremely Likely
If yes (likely as above), what changes would you make? If no (unlikely or very unlikely), why not?	Option for comments
How many minutes did it take for you to complete this module?	Option for comments
How useful did you find this module?	Likert scale 1-5 • 1-Completely dissatisfied • 2-Dissatisfied • 3-Neutral • 4-Satisfied • 5-Completely satisfied
How would you rate the aesthetics (layout, graphics, or visuals)?	Likert scale 1-5 • 1-Completely dissatisfied • 2-Dissatisfied • 3-Neutral • 4-Satisfied • 5-Completely satisfied
How would you rate the gestural design (interactions, taps, swipes, and scrolling) functions?	Likert scale 1-5 • 1-Completely dissatisfied • 2-Dissatisfied • 3-Neutral • 4-Satisfied • 5-Completely satisfied
How would you rate the performance of the module (features and components)?	Likert scale 1-5 • 1-Completely dissatisfied • 2-Dissatisfied • 3-Neutral • 4-Satisfied • 5-Completely satisfied
How would you rate the quality of content?	Likert scale 1-5 • 1-Completely dissatisfied • 2-Dissatisfied • 3-Neutral • 4-Satisfied • 5-Completely satisfied
How would you rate the usability or navigation of this module?	Likert scale 1-5 • 1-Completely dissatisfied • 2-Dissatisfied • 3-Neutral • 4-Satisfied • 5-Completely satisfied

XSL•FO
RenderX

Question	Rating options
Overall how engaging did you find the module (entertainment, interest, relatability, and interactivity)?	Likert scale 1-5 • 1-Completely dissatisfied • 2-Dissatisfied • 3-Neutral • 4-Satisfied • 5-Completely satisfied
Overall what would you rate this module?	Likert scale 1-5 • 1- Poor • 2-Satisfactory • 3-Good • 4-Very Good • 5-Excellent
What device did you use to access the module (desktop, mobile, or tablet)?	Options to choose from: • Laptop • Mobile • Tablet • Desktop • Combination
Would you recommend this module to other veterans or family or friends of a veteran?	Options to choose from: Yes, No
Did the audio and visual presentations resonate with you and your personal experiences?	Options to choose from: No, Somewhat, Yes, No comment
How did you find the use of personal stories, videos, animations, and illustrations?	Option for comments
Please make any further comments regarding the engagement and communication of the module.	Option for comments
Please provide any additional comments on the functionality of the module.	Option for comments
Was the visual information accurate (concepts are logical and correct)? Please provide any comments.	Option for comments
What further recommendations or comments would you make to improve the Go-Beyond Platform and the module?	Option for comments
What was the least helpful part of the module?	Option for comments
What was the most helpful part of the module?	Option for comments

Statistical Analysis

To determine the reach and adoption of the Go-Beyond program, a user-centered evaluation was performed using information routinely collected to monitor and improve the quality of the program. The system engagement data were automatically collected from users who accessed the M-CARM and Go-Beyond websites on a voluntary basis.

The data analysis was conducted using R (R Core Team) [21]. Descriptive statistics were used to gain insights into the distribution and central tendencies of the collected data. Subsequently, logistic regression analysis was used to discern factors influencing users' likelihood to progress to Go-Beyond following completion of the M-CARM.

To analyze the qualitative data gathered through the voluntary online surveys, a systematic approach was used. Thematic analysis was used to recognize patterns and identify core themes. Themes were extracted from the data and grouped under the questions posed adhering to the thematic analysis approach by Braun and Clarke [22]. The initial coding of responses was conducted manually by the primary researcher, who independently reviewed the data to develop preliminary codes. Responses were coded into categories and were organized thematically, allowing for the identification of recurring topics. These codes were then reviewed and refined through consensus discussions among the research team, ensuring that differing perspectives were considered. The researchers acknowledge that their perspectives and experiences may influence their interpretations, and they have taken steps to reflect on and mitigate potential biases throughout the research process. The positionality of the researchers was acknowledged throughout the analysis process, with the team reflecting on how their backgrounds, experiences, and potential biases might influence the interpretation of the data and this qualitative analysis [23]. All researchers were committed to evaluating a web-based intervention aiming for it to be not only useful but also feasible and acceptable to the intended users. The primary researcher brings over 10 years of experience working with clients from military backgrounds, providing a unique perspective informed by extensive professional practice. However, it is important to note that the primary and senior researchers were not involved in the creation or implementation of the version of the program under study, positioning them as outsiders in this context. This

XSL•FO
RenderX

outsider status allows for a more detached analysis, free from potential biases related to program design and execution. Consensus discussions among the research team also ensured that individual biases were minimized and enhanced the credibility of the interpretations. Representative quotations and exemplars were extracted to illustrate each identified theme.

Ethical Considerations

The Department of Defense and Veterans' Affairs Human Research Ethics Committee has deemed this study to be a quality assurance activity that upholds the principles of the National Statement on Ethical Conduct in Human Research and the Ethical Consideration in Quality Assurance and Evaluation Activities (approval number 516-23). In the privacy policy statement that participants have access to when deciding to sign in to the program, they are informed about the organization's commitment to comply with the Privacy Act 1988 (Cth) (Privacy Act), including the Australian Privacy Principles. The data were extracted from the Absorb LMS system by the LMS Administrator and deidentified before they were provided to the research team. After data cleaning and integration, analysis was conducted by the research team on deidentified data. All collected data were held confidential at all times, in a password-protected database, on a server that requires a unique log-in and password by all team members. Only the project team had access to the data. No compensation was provided to users as this service evaluation used routinely collected data.

Results

Go-Beyond User and M-CARM–Only User Profile

Out of the total of 959 users who voluntarily completed M-CARM, there were 686 who did not proceed to sign up for the Go-Beyond program (M-CARM–only users) and 273 who signed up for Go-Beyond, resulting in a conversion rate of 28.5% (273/959) and a preprogram dropout rate of 71.5% (686/959). Table 2 summarizes the demographic characteristics of both user groups. Only upon signing up for the Go-Beyond program, the data regarding time since discharge from the Australian Defence Force were recorded. Out of 273 Go-Beyond users, 215 were discharged from the Australian Defence Force and 43 reported that they were still serving. The median number of years since transition for Go-Beyond users who were discharged equaled 2.74 years, with a large proportion of users (165/273, 60.5 %) reporting being within 0-5 years from discharge.

XSL•FO
RenderX

Table 2. Demographic characteristics of the Go-Beyond users and Military-Civilian Adjustment and Reintegration Measure (M-CARM)–only users.

Variables	Go-Beyond users (n=273)	M-CARM–only users (n=686)
Sex, n (%)		
Male	192 (70.3)	498 (72.6)
Female	67 (25)	178 (26)
Prefer not to answer	1 (0.2)	10 (2)
N/A[a]	13 (4.8)	(0)
Age range (years), n (%)		
18-25	20 (7.3)	74 (11)
26-35	65 (24)	187 (27.3)
36-45	59 (21)	162 (23.6)
46-55	73 (26.7)	144 (21)
56-65	32 (12)	93 (14)
66-75	8 (3)	15 (2)
>75	2 (1)	6 (1)
Prefer not to answer	1 (0.2)	5 (1)
N/A	13 (4.8)	(0)
Service, n (%)		
Army	153 (56)	384 (56)
Navy	48 (18)	96 (14)
Airforce	31 (11)	117 (17)
Prefer not to answer	24 (9)	77 (11)
Army+Airforce	2 (1)	3 (0.4)
Army+Navy	1 (0.2)	1 (0.2)
Navy+Airforce	0 (0)	2 (0.3)
N/A	14 (5.1)	0 (0)
Type of needs as shown on M-CARM, n (%)[b]		
Beliefs About Civilians	197 (75.5)	441 (64.3)
Regimentation	182 (69.7)	452 (65.9)
Resentment and Regret	176 (67.4)	396 (57.7)
Purpose and Connection	146 (55.9)	287 (41.8)
Help Seeking	76 (29.1)	184 (26.8)

[a]N/A: not applicable.

[b]One person can have multiple needs.

There were no statistically significant differences observed between the Go-Beyond and M-CARM–only user groups concerning age, sex, or service type (all $P>.05$). However, a significantly higher number of individuals in the Go-Beyond user group were identified as having needs in the areas of Purpose and Connection ($\chi^2_4=14.59$, $P<.001$), Resentment and Regret ($\chi^2_4=7.05$, $P=.008$), and Beliefs About Civilians ($\chi^2_4=10.273$, $P=.001$) compared with the M-CARM–only users.

A logistic regression was then performed to assess the effects of M-CARM outcomes and service type on the likelihood of enrolling in the Go-Beyond program. Participant's age, sex, and service type were controlled for. The outcomes of logistic regression showed that participants with the need for Purpose and Connection and Beliefs About Civilians were found to be significantly more likely to sign up for Go-Beyond ($P=.003$ and $P=.03$ respectively). Individuals expressing a need for Purpose and Connection were approximately 1.64 times more likely to enroll (odds ratio 1.64, 95% CI 1.17-2.28) compared with those who did not. Similarly, those with the need Beliefs About Civilians were 1.50 times more likely to enroll (odds ratio 1.50, 95% CI 1.06-2.18) compared with those without this need. Individuals who were identified with a need for Resentment and Regret were not found to be significantly more likely to sign up for Go-Beyond ($P>.05$).

XSL•FO
RenderX

Go-Beyond User Engagement

Out of 273 users who signed up for Go-Beyond, 85 (31%) users completed at least 1 assigned module. The data on the enrollment status at the time of the data extraction for each of the modules and corresponding completion rates can be found in Table 3. Out of participants who completed at least 1 module, nearly all users who had been assigned 1 module, completed it (36/38, 94% users). Among users who were assigned 2 modules, 37% (14/38) have completed them both. For those who had been assigned 3 modules, 26% (9/34) have completed all 3. Of users who had been assigned 4 modules, 10% (3/31) have completed them, and of users who had been assigned 5 modules, 3% (2/63) have completed all 5. A total of 22% (10/45) of users who had assigned 6 modules have completed all 6. Overall, 3% (8/247) of participants who had been assigned other modules also completed the Communication module. Out of Go-Beyond users, who did not have any modules assigned, but still accessed the program (26 users), 4 (15%) users completed 1 module, 1 (4%) user completed 3 modules, 2 (8%) users completed 4 modules, 1 (4%) user completed all available modules, and 3 (12%) completed the Communication module. The module with the highest completion rate was Finding Purpose with 41 (20%) out of 206 users who signed up to this module, followed by Beliefs About Civilians with 48 (19.7%) out of 244 who signed up, Resentment and Regret with 38 (17.6%) out of 216 who signed up, Regimentation with 33 (16%) out of 205 who signed up, Help Seeking with 21 (15.5%) out of 156 who signed up, and Social Connection with 17 (8.4%) out of 220 signed up users. The lowest completion rate was the additional module Effective Communication in Civilian Life, in which everyone was enrolled automatically, with 11 (4%) out of 273 users who signed up.

Table 3. Enrollment status for each Go-Beyond module at the time of data extraction.

Module and status	Not started, n (%)	In progress, n (%)	Complete, n (%)
Beliefs About Civilians (n=244)	151 (61.9)	45 (18.4)	48 (19.7)
Finding Purpose (n=206)	128 (62)	37 (18)	41 (20)
Resentment and Regret (n=216)	163 (75.5)	15 (6.9)	38 (17.6)
Regimentation (n=205)	166 (81)	6 (3)	33 (16)
Help Seeking (n=156)	110 (81.5)	4 (3)	21 (15.5)
Social Connection (n=220)	175 (86)	11 (5.4)	17 (8.4)
Communication[a] (n=273)	262 (96)	0 (0)	11 (4)

[a]This module is available to everyone as an additional module regardless of their Military-Civilian Adjustment and Reintegration Measure profile.

The “time spent” refers to the time participants spent engaging with each module. This metric could reflect multiple factors, including module complexity, participant interest, or difficulty with the content. The module with the highest median time spent by the users who did complete it was Beliefs About Civilians (67, IQR 23.4-110.6 minutes). It was followed by Finding Purpose (median time of 32, IQR 9-55 minutes), Regimentation (median time of 25, IQR 11-39 minutes), Resentment and Regret (median time of 17, IQR 0.25-33.75 minutes), Social Connection (median time of 13, IQR 5.5-20.5 minutes) and Help Seeking (median time of 11, IQR 0-31.5 minutes). The median time for the additional module Effective Communication in Civilian Life was 19 (IQR 10.75-27.25 minutes) minutes.

No statistically significant differences were observed (all $P>.05$) in relation to age, sex, service type, or type of needs (Purpose and Connection, Help Seeking, Regimentation, Resentment and Regret, and Beliefs About Civilians) between participants who had completed at least 1 module and those who had not started or were still in progress at the time of data extraction.

Go-Beyond User Experience Survey Data

There were 94 self-report surveys assessing the Go-Beyond internet-based program completed. The survey responses provided insights regarding the program’s subjective usefulness, user engagement, the platform’s usability, and user recommendations.

Perceived Usefulness and Satisfaction

Out of the 94 internet-based self-report surveys evaluating Go-Beyond, 80% (74/94) of respondents acknowledged the usefulness of the module. Moreover, 85% (80/94) of the users were “satisfied” or “completely satisfied” with the program’s features and components, and 79% (73/94) of survey respondents stated that they were “satisfied” or “completely satisfied” with the quality of the content. Furthermore, 86% (81/94) of survey respondents indicated a likelihood of implementing changes in their lives following the module, and 94% (88/94) reported they would recommend the program to other veterans or their family or friends.

User Assessment of Platforms Engaging Features

The users were asked how engaging overall the modules were in terms of entertainment, interest, relatability, and interactivity of the modules (see Textbox 1 for questions). A total of 78% (73/94) of the survey respondents found the modules engaging overall. In addition, 86% (81/94) of participants found the time required to complete the program to be satisfactory.

Platform Usability

A total of 91% (86/94) of users reported a high degree of satisfaction with the Go-Beyond Platform’s usability. The audio and visual presentations resonated or somewhat resonated with 94% (88/94) of users. In relation to the use of personal stories, videos, animations, and illustrations, users reported that they were: “Very helpful and informative” and commented that they

XSL•FO
RenderX

enhanced the program: “I thought the quotes, illustrations, and videos were used really well in this module and very realistic.” And helped them self-reflect: “It was good. Made me look at myself a little and accept that it is ok to be flexible.” The survey indicated strong endorsements of the platform’s aesthetics, rated as “satisfied” and completely satisfied” by 86% (81/94) of users. Furthermore, 85% (80/94) of users were “satisfied” or “completely satisfied” with gestural design functions of the platform.

User Recommendations

When questioned about the most beneficial features of the program, the majority of responses were related to (1) relatable psychoeducational content (ie, “The information behind emotions and why we feel them” and “The personal insights it provided along with leaving me with a sense of not being alone in feeling this way”), (2) opportunities for self-reflection (ie, “Given the option to self-reflect” and “Providing places to self-reflect on the information given”), and (3) practical strategies and techniques (ie, “strategies and thought techniques to widen perspectives and assist with positive actions”). On the other hand, the least helpful features included (1) a long reference list and links to resources (ie, “The links end up with more links that take you further and further from the main content”), (2) a lack of diversity in the personal videos (ie, “I don’t recall any useful or memorable audio clips with information or personal stories, although I can appreciate that it is not simple to source authentic, vulnerable, and engaging stories”). Furthermore, some technical issues were also identified (ie, “No previous work was recorded. I had to wait to I got access to a printer to complete the module and all my previous work had not been saved. Having a complete module button that takes you back to the home page would be great,” “Return to other courses button when completing the module would be great,” and “Better audio for the doctor videos. Very echo’y”).

Discussion

Principal Findings

The study aimed to assess the reach, adoption, and engagement with the “Go-Beyond, Navigating Life Beyond Service Program,” a web-based program designed to aid veterans in transitioning to civilian life, incorporating evidence-informed content on psychological and cultural factors. The program aims to provide education, build psychological insight, offer practical tools, and offer users links to available support. The findings underscore the potential value of the Go-Beyond program as a web-based initiative, leveraging digital technology to provide a personalized and easily accessible resource for veterans navigating the transition to civilian life.

Most of the Go-Beyond users were male (192/273, 70.3%), within 46-55 years of age, and served in the Army (153/273, 56%). Median years since the transition for Go-Beyond users who were discharged equaled 2.74 years, with a large proportion of users (165/273, 60.5 %) reporting being within 0-5 years from discharge. It is noteworthy that the first few years after discharge are widely recognized as an extremely vulnerable period for ex-serving members. Suicide risk, in particular, is notably high during the initial year of transitioning to civilian life [24]. In addition, evidence in the literature points out that many individuals experience reintegration difficulties within 6 years after separation [25], underscoring the importance of the program during this critical phase. Furthermore, the analysis of users’ needs highlights the need for a program that can address a wide array of transitional challenges and provide essential support during the transition process [18].

The overall conversion rate from completing the M-CARM questionnaire to signing up for the Go-Beyond program was 28.5% (273/959). Interestingly, among the analyzed factors, the study identified that having needs in Beliefs About Civilians and Purpose and Connection domains could be considered relevant predictors of engagement with the program. Users identified with these needs were up to almost 2 times more likely to sign up for Go-Beyond after completing the M-CARM survey. Interestingly, aligning with these findings, the module Finding Purpose had the highest completion rate (41/206, 20%), followed by Beliefs About Civilians (48/244, 19.7%). There is a limited number of studies providing conversion rates for internet-based, self-help programs for mental health, with the definition of uptake or adoption varying widely based on program structure and objectives [26]. A study assessing the public implementation of a web-based program for veterans dealing with risky alcohol use and PTSD, named VetChange, reported a conversion rate of 4.7% (1033/22,087) from unique homepage visits resulting in full account registrations [27]. Another study examining the impact of self-help resources on gambling behavior, found a conversion rate of 14% (5652/8083) from the program registration page to create an account. On the other hand, a study evaluating the MoodGYM program, which aims to prevent and manage symptoms of depression and anxiety, demonstrated a 42.2% (82,159/194,840) registration rate, considering the number of website visitors. Although further research is necessary to thoroughly assess ongoing participation after the initial sign-up to Go-Beyond, the conversion rate found in this study appears to be mostly higher than comparable published programs and indicates a decent level of interest and engagement among veterans.

The time spent on each module of the Go-Beyond Program indicates varying degrees of user engagement, with a notable proportion (31%) completing at least 1 assigned module. These outcomes also correspond to the findings of other studies [27]. A systematic review of digital self-help interventions for depression, low mood, or anxiety demonstrated that real-world uptake and engagement present a significant variability and may differ from adherence in controlled trials [27]. Across the studies analyzed in that research, minimal usage of the intervention ranged from 21% to 88%, with users either accessing it at least once or completing one module. However, completion or sustained use, defined as finishing all modules or the final assessment, or continuing to use the apps after a prolonged period of time was lower and ranged from 0.5% to 28.6%. Research has demonstrated that factors such as motivation, previous experience, and the perceived relevance of course content can significantly influence engagement and completion rates in online interventions [28]. Jardine et al [29] identified uncertainty about the intervention’s usefulness and usability as the main barriers to its uptake. Previous studies also

XSL•FO
RenderX

revealed that forgetting about the intervention, not finding time for it, and not finding it useful posed main barriers to early engagement [28]. Further barriers to engagement included technological challenges, severe mental health issues, lack of personalization [30], lack of privacy, or difficulty in relating to the content [30,31]. Conversely, facilitators include among others user-friendly interfaces, personalized content, social connectedness facilitated by the intervention, and the perception of gaining better insight into the condition [30,31].

Beatty and Binnion [31] examined 36 studies to determine predictors for adherence to web-based interventions, such as demographic, psychological, characteristics of the presenting problem, and intervention and computer-related factors. The findings indicated that the female sex, higher treatment expectancy, sufficient time, and personalized intervention content were associated with increased adherence. Further, not having enough time, dissatisfaction with program content, perceiving content as impersonal, and computer difficulties were found to decrease adherence [32]. Sex was not found to play a significant role in predicting the completion of any module in the Go-Beyond program in our study. However, as identified in the users' feedback, technical challenges experienced by Go-Beyond users, such as difficulties with saving completed worksheets, accessing the homepage with all relevant modules, and audio quality, may have influenced veterans' frustration levels and their willingness or ability to complete the modules. On the other hand, the highly personalized content of the program, which offers users modules tailored to their identified M-CARM areas of need, may have contributed to the completion rates and engagement with the module evidenced by high ratings of quality of content and usefulness of the program. Moreover, the results of the user survey underscore overall user satisfaction with the Go-Beyond program, encompassing various aspects of module effectiveness, user engagement, platform usability, and user recommendations. They suggest that the Go-Beyond internet-based program engages users, providing valuable content with aesthetically pleasing and functionally efficient modules. Users also express a high likelihood of recommending the program to others, indicating a potential helpful impact on a broader audience. However, further studies evaluating specific barriers and facilitators for engagement with Go-Beyond are necessary.

While considering factors impacting adherence, it needs to be noted that Wojtowicz et al [32] suggest that levels of guidance or support through phone or email contribute to increased adherence to internet-based interventions. It is suggested that guided support may enhance motivation to participate or foster a greater sense of accountability to adhere [33]. Peer support and professional-assisted delivery of web-based interventions also pose a potential solution to the enhancement of engagement in web-based interventions [34]. Research suggests that peer-assisted internet-based interventions have proven to be feasible in the veteran population [35,36]. Since Go-Beyond was designed as a completely self-help tool, an entirely self-help format may be challenging for the ex-Defence cohort where low motivation and avoidance are key features in common mental health disorders experienced by veterans such as depression, anxiety, and PTSD [37-39].

The findings of this preliminary evaluation provide several lessons learned from the preliminary implementation and lay the groundwork for further investigation and refinement of the Go-Beyond program. In this context, it is crucial to emphasize the importance of using real-world data when evaluating adoption and engagement in digital mental health interventions. As previously highlighted, real-world settings often differ from controlled trial conditions, as individuals' interaction with these interventions can be influenced by factors such as accessibility, usability, and personal preferences. Therefore, comparing uptake and usage patterns across various interventions using real-world data could offer invaluable insights, enabling researchers and developers to identify trends, successes, and areas for improvement in the field of digital mental health. This, in turn, would facilitate making informed decisions to enhance user experience and outcomes [26].

Limitations and Future Directions

There are several limitations to this study. Veterans who chose to participate in the Go-Beyond program may not represent the entire veteran population, leading to selection bias. Since the M-CARM measure as well as the Go-Beyond program are available to anyone, there is a possibility that some of the users might have been members of the general population interested in the program. Those who completed the M-CARM and signed up for the Go-Beyond might have unique characteristics or experiences that differ from nonparticipants, impacting the generalizability of the findings. Therefore, the existing literature is the only source of information to draw conclusions as to why some chose to enroll in the program and complete assigned modules. Furthermore, only 31% of users who signed up for Go-Beyond completed at least 1 assigned module and provided feedback. As a result, the favorable quantitative and qualitative feedback received regarding the program may not fully represent the overall user experience. This highlights the importance of considering the perspectives of those who did not fully engage with or did not complete the program in future evaluations another limitation of our study is related to the interpretation of the "time spent" metric. While "time spent" refers to the duration participants engaged with each module, this measure does not account for various underlying factors that could influence the observed time. For instance, a longer duration may indicate that a module was more complex, more engaging, or potentially more challenging for participants. This variability can complicate the interpretation of whether differences in "time spent" are due to the inherent characteristics of the modules or other participant-specific factors.

Finally, this was a preliminary examination of the reach, user engagement, and completion rates of the Go-Beyond program, and further evaluations of the barriers and facilitators, as well as the effectiveness of the internet-based program on veterans' readjustment to civilian life are needed.

Conclusion

In conclusion, the Go-Beyond program may be a promising initiative in leveraging digital technology to support veterans during their transition to civilian life. The study's outcomes contribute valuable insights into user engagement, completion rates, and the program's overall adoption. As the program

XSL•FO
RenderX

continues to evolve, ongoing evaluation and refinement will be crucial to enhancing its effectiveness and addressing the diverse needs of military personnel in their reintegration journey. Our study supports the notion that web-based interventions hold substantial promise in enhancing the experience of transition into the civilian life of veterans. Future research should examine factors that predict the use and nonuse of Go-Beyond, how use changes over time, and what factors can serve as predictors of use. Furthermore, it is crucial to incorporate relevant outcome measures to gain a thorough understanding of its effectiveness and potential benefits.

Acknowledgments

We gratefully acknowledge Returned and Services League of Australia Queensland (RSLQ) for their committed support and ongoing partnership. RSLQ has funded a multiphase research project aimed at determining the factors that contribute to a successful and psychologically healthy transition. This collaboration has led to the development of the M-CARM and Go-Beyond program. Generative artificial intelligence was not used in any portion of the manuscript writing.

Data Availability

The datasets generated during and analyzed during this study are not publicly available due to ethical and privacy considerations but are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

References

1. Resnik L, Bradford DW, Glynn SM, Jette AM, Johnson Hernandez CJ, Wills S. Issues in defining and measuring veteran community reintegration: proceedings of the working group on community reintegration, VA Rehabilitation Outcomes Conference, Miami, Florida. J Rehabil Res Dev 2012;49(1):87-100 [FREE Full text] [doi: 10.1682/jrrd.2010.06.0107] [Medline: 22492341]
2. Romaniuk M, Kidd C. The psychological adjustment experience of reintegration following discharge from military service: a systemic review. Journal of Military and Veterans' Health 2018;26(2):60.
3. Sachdev S, Dixit S. Military to civilian cultural transition experiences of retired military personnel: a systematic meta-synthesis. Mil Psychol 2023:1-14 [FREE Full text] [doi: 10.1080/08995605.2023.2237835] [Medline: 37490333]
4. Black T, Papile C. Making it on civvy street: an online survey of canadian veterans in transition. Canadian journal of counselling and psychotherapy 2010;44(4):383.
5. Blackburn D. Out of uniform: psychosocial issues experienced and coping mechanisms used by veterans during the military–civilian transition. Journal of Military, Veteran and Family Health 2017;3(1):62-69 [FREE Full text] [doi: 10.3138/jmvfh.4160]
6. Romaniuk M, Kidd C, Banfield M, Batterham PJ. Obtaining and retaining employment post military service: a qualitative analysis of challenges experienced by Australian veterans. Journal of Employment Couns 2022;60(2):72-88 [FREE Full text] [doi: 10.1002/joec.12199]
7. Misca G, Augustus J, Russell J, Walker J. Meaning(s) of transition(s) from military to civilian life at the intersection with mental health: implications for clinical settings. Front Psychol 2023;14:1142528 [FREE Full text] [doi: 10.3389/fpsyg.2023.1142528] [Medline: 37228334]
8. Moriarty H, Winter L, Robinson K, True G, Piersol C, Vause-Earland T, et al. Exploration of individual and family factors related to community reintegration in veterans with traumatic brain injury. J Am Psychiatr Nurses Assoc 2015;21(3):195-211 [FREE Full text] [doi: 10.1177/1078390315591879] [Medline: 26156059]
9. Ahern J, Worthen M, Masters J, Lippman SA, Ozer EJ, Moos R. The challenges of afghanistan and iraq veterans' transition from military to civilian life and approaches to reconnection. PLoS One 2015;10(7):e0128599 [FREE Full text] [doi: 10.1371/journal.pone.0128599] [Medline: 26132291]
10. Pease JL, Billera M, Gerard G. Military culture and the transition to civilian life: suicide risk and other considerations. Soc Work 2016;61(1):83-86 [FREE Full text] [doi: 10.1093/sw/swv050] [Medline: 26898003]
11. Bröcker E, Suliman S, Olff M, Seedat S. Feasibility, acceptability, and effectiveness of web-based and mobile PTSD Coach: a systematic review and meta-analysis. Eur J Psychotraumatol 2023;14(2):2209469 [FREE Full text] [doi: 10.1080/20008066.2023.2209469] [Medline: 37226885]
12. Hobfoll SE, Blais RK, Stevens NR, Walt L, Gengler R. Vets prevail online intervention reduces PTSD and depression in veterans with mild-to-moderate symptoms. J Consult Clin Psychol 2016;84(1):31-42 [FREE Full text] [doi: 10.1037/ccp0000041] [Medline: 26322788]
13. Kuhn E, van der Meer C, Owen JE, Hoffman JE, Cash R, Carrese P, et al. PTSD coach around the world. Mhealth 2018;4:15 [FREE Full text] [doi: 10.21037/mhealth.2018.05.01] [Medline: 29963560]

XSL•FO
RenderX

14. McMahon B, Guindalini C, Mellor R. Computerised health interventions targeting Australian veterans and their families: a scoping review. Health Promot J Austr 2024;35(4):875-890 [FREE Full text] [doi: 10.1002/hpja.832] [Medline: 38082422]
15. Zanaboni P, Ngangue P, Mbemba GIC, Schopf TR, Bergmo TS, Gagnon MP. Methods to evaluate the effects of internet-based digital health interventions for citizens: systematic review of reviews. J Med Internet Res 2018;20(6):e10202-e10202 [FREE Full text] [doi: 10.2196/10202] [Medline: 29880470]
16. Fraynt R, Cooper D, Edwards-Stewart A, Hoyt T, Micheel L, Pruitt L, et al. An evaluation of mobile applications designed to assist service members and veterans transitioning to civilian life. Psychol Serv 2018;15(2):208-215 [FREE Full text] [doi: 10.1037/ser0000205] [Medline: 29723023]
17. Military-Civilian Adjustment and Reintegration Measure. 2021. URL: https://www.m-carm.org [accessed 2021-01-11]
18. Romaniuk M, Fisher G, Kidd C, Batterham PJ. Assessing psychological adjustment and cultural reintegration after military service: development and psychometric evaluation of the post-separation military-civilian adjustment and reintegration measure (M-CARM). BMC Psychiatry 2020;20(1):531 [FREE Full text] [doi: 10.1186/s12888-020-02936-y] [Medline: 33167907]
19. Lang AJ, Schnurr PP, Jain S, He F, Walser RD, Bolton E, et al. Randomized controlled trial of acceptance and commitment therapy for distress and impairment in OEF/OIF/OND veterans. Psychol Trauma 2017;9(Suppl 1):74-84 [FREE Full text] [doi: 10.1037/tra0000127] [Medline: 27322609]
20. Go-Beyond, navigating life beyond service®. 2021. URL: https://gobeyond.org.au/ [accessed 2024-11-23]
21. Team R. R: A language and environment for statistical computing. R Foundation for Statistical Computing 2021;12. [doi: 10.1201/b17972-15]
22. Braun V, Clarke V. Thematic Analysis In APA Handbook of Research Methods in Psychology. In: Research Designs: Quantitative, Qualitative, Neuropsychological, and Biological. Washington, DC: American Psychological Association; 2012.
23. Shaw RM, Howe J, Beazer J, Carr T. Ethics and positionality in qualitative research with vulnerable and marginal groups. Qualitative Research 2019;20(3):277-293 [FREE Full text] [doi: 10.1177/1468794119841839]
24. Ravindran C, Morley SW, Stephens BM, Stanley H, Reger MA. Association of suicide risk with transition to civilian life among US military service members. JAMA Netw Open 2020;3(9):e2016261 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.16261] [Medline: 32915235]
25. Sayer NA, Frazier P, Orazem RJ, Murdoch M, Gravely A, Carlson KF, et al. Military to civilian questionnaire: a measure of postdeployment community reintegration difficulty among veterans using Department of Veterans Affairs medical care. J Trauma Stress 2011;24(6):660-670 [FREE Full text] [doi: 10.1002/jts.20706] [Medline: 22162082]
26. Fleming T, Bavin L, Lucassen M, Stasiak K, Hopkins S, Merry S. Beyond the trial: systematic review of real-world uptake and engagement with digital self-help interventions for depression, low mood, or anxiety. J Med Internet Res 2018;20(6):e199-e199 [FREE Full text] [doi: 10.2196/jmir.9275] [Medline: 29875089]
27. Enggasser JL, Hermos JA, Rubin A, Lachowicz M, Rybin D, Brief DJ, et al. Drinking goal choice and outcomes in a web-based alcohol intervention: results from VetChange. Addict Behav 2015;42:63-68 [FREE Full text] [doi: 10.1016/j.addbeh.2014.10.036] [Medline: 25671224]
28. Peng R, Li X, Guo Y, Ning H, Huang J, Jiang D, et al. Barriers and facilitators to acceptance and implementation of eMental-health intervention among older adults: a qualitative systematic review. Digit Health 2024;10:20552076241234628 [FREE Full text] [doi: 10.1177/20552076241234628] [Medline: 38444518]
29. Jardine J, Nadal C, Robinson S, Enrique A, Hanratty M, Doherty G. Between rhetoric and reality: real-world barriers to uptake and early engagement in digital mental health interventions. ACM Trans. Comput.-Hum. Interact 2024;31(2):1-59 [FREE Full text] [doi: 10.1145/3635472]
30. Borghouts J, Eikey E, Mark G, de Leon C, Schueller SM, Schneider M, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. J Med Internet Res 2021;23(3):e24387 [FREE Full text] [doi: 10.2196/24387] [Medline: 33759801]
31. Beatty L, Binnion C. A systematic review of predictors of, and reasons for, adherence to online psychological interventions. Int J Behav Med 2016;23(6):776-794 [FREE Full text] [doi: 10.1007/s12529-016-9556-9] [Medline: 26957109]
32. Wojtowicz M, Day V, McGrath PJ. Predictors of participant retention in a guided online self-help program for university students: prospective cohort study. J Med Internet Res 2013;15(5):e96 [FREE Full text] [doi: 10.2196/jmir.2323] [Medline: 23697614]
33. Possemato K, Wu J, Greene C, MacQueen R, Blonigen D, Wade M, et al. Web-based problem-solving training with and without peer support in veterans with unmet mental health needs: pilot study of feasibility, user acceptability, and participant engagement. J Med Internet Res 2022;24(1):e29559 [FREE Full text] [doi: 10.2196/29559] [Medline: 35023846]
34. Romaniuk M, Evans J, Kidd C. Evaluation of the online, peer delivered 'Post War: Survive to Thrive Program' for veterans with symptoms of posttraumatic stress disorder. Journal of Military and Veterans' Health 2019;27(2):55.
35. Nelson CB, Abraham KM, Walters H, Pfeiffer PN, Valenstein M. Integration of peer support and computer-based CBT for veterans with depression. Computers in Human Behavior 2014;31:57-64 [FREE Full text] [doi: 10.1016/j.chb.2013.10.012]

XSL•FO
RenderX

36. Boscarino JA, Hoffman SN, Pitcavage JM, Urosevich TG. Mental health disorders and treatment seeking among veterans in non-VA facilities: results and implications from the veterans' health study. Mil Behav Health 2015;3(4):244-254 [FREE Full text] [doi: 10.1080/21635781.2015.1077179] [Medline: 26640743]
37. Orcutt HK, Reffi AN, Ellis RA. Emotion in posttraumatic stress disorder. Etiology, Assessment, Neurobiology, and Treatment 2020:409-436 [FREE Full text] [doi: 10.1016/b978-0-12-816022-0.00014-4]
38. Smith B. Depression and motivation. Phenom Cogn Sci 2012;12(4):615-635 [FREE Full text] [doi: 10.1007/s11097-012-9264-0]
39. Svendsen MJ, Wood KW, Kyle J, Cooper K, Rasmussen CDN, Sandal LF, et al. Barriers and facilitators to patient uptake and utilisation of digital interventions for the self-management of low back pain: a systematic review of qualitative studies. BMJ Open 2020;10(12):e038800 [FREE Full text] [doi: 10.1136/bmjopen-2020-038800] [Medline: 33310794]

Abbreviations

GMR: Gallipoli Medical Research
M-CARM: Military-Civilian Adjustment and Reintegration Measure
PTSD: posttraumatic stress disorder
RSLQ: Returned and Services League of Australia Queensland

Edited by A Mavragani; submitted 23.05.24; peer-reviewed by A Barker, K Carra; comments to author 09.08.24; revised version received 13.09.24; accepted 17.10.24; published 23.01.25.

Please cite as:
Alichniewicz KK, Hampton S, Romaniuk M, Bennett D, Guindalini C
Use of Go-Beyond as a Self-Directed Internet-Based Program Supporting Veterans' Transition to Civilian Life: Preliminary Usability Study
JMIR Form Res 2025;9:e60868
URL: https://formative.jmir.org/2025/1/e60868
doi:10.2196/60868
PMID:39847427

©Karolina Katarzyna Alichniewicz, Sarah Hampton, Madeline Romaniuk, Darcy Bennett, Camila Guindalini. Originally published in JMIR Formative Research (https://formative.jmir.org), 23.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Usability and Usefulness of a Symptom Management Coaching System for Patients With Cancer Treated With Immune Checkpoint Inhibitors: Comparative Mixed Methods Study

Savannah Lucia Caterina Glaser[1,2], MSc; Itske Fraterman[3], MSc; Noah van Brummelen[1,4], MSc; Valentina Tibollo[5], MSc; Laura Maria Del Campo[6], PhD; Henk Mallo[7], MSN; Sofie Wilgenhof[7], MD, PhD; Szymon Wilk[8], PhD; Vitali Gisko[9], BA; Vadzim Khadakou[9], BA; Ronald Cornet[1,2], PhD, IR; Manuel Ottaviano[10], PhD; Stephanie Medlock[1,11], DVM, PhD

[1]Department of Medical Informatics, Amsterdam UMC - University of Amsterdam, Amsterdam, Netherlands
[2]Methodology & Digital Health, Amsterdam Public Health, Amsterdam, Netherlands
[3]Department of Psychosocial Research and Epidemiology, The Netherlands Cancer Institute, Amsterdam, Netherlands
[4]Digital Health, Amsterdam Public Health, Amsterdam, Netherlands
[5]Laboratory of Informatics and Systems Engineering for Clinical Research, Istituti Clinici Scientifici Maugeri SpA SB IRCCS, Pavia, Italy
[6]Associazione Italiana Malati di Cancro, Rome, Italy
[7]Department of Medical Oncology, Antoni van Leeuwenhoek, Amsterdam, Netherlands
[8]Institute of Computing Science, Poznan University of Technology, Poznan, Poland
[9]BITSENS JSC, Vilnius, Lithuania
[10]Life Supporting Technologies, Universidad Politécnica de Madrid, Madrid, Spain
[11]Methodology & Aging & Later Life, Amsterdam Public Health, Amsterdam, Netherlands

Corresponding Author:
Stephanie Medlock, DVM, PhD
Department of Medical Informatics
Amsterdam UMC - University of Amsterdam
Amsterdam
Netherlands
Phone: 31 205669111
Email: s.k.medlock@amsterdamumc.nl

Abstract

Background: The prognosis for patients with several types of cancer has substantially improved following the introduction of immune checkpoint inhibitors, a novel type of immunotherapy. However, patients may experience symptoms both from the cancer itself and from the medication. A prototype of the eHealth tool Cancer Patients Better Life Experience (CAPABLE) was developed to facilitate symptom management, aimed at patients with melanoma and renal cell carcinoma treated with immunotherapy. Better usability of such eHealth tools can lead to improved user well-being and reduced risk of harm. It is unknown for usability evaluations whether certain usability problems would only be evident to patients whose condition closely resembles the target population, or if a broader group of patients would lead to the identification of a broader range of potential usability issues.

Objective: This study aims to evaluate the CAPABLE prototype by conducting tests to assess usability, user experience, and perceived acceptability among end users, and to assess any agreements or differences in the results of our wide range of participants.

Methods: This usability study was executed by interviewing participants with a melanoma or renal cell carcinoma diagnosis who have received immunotherapy and participants without direct experience with the targeted cancer types who have not received immunotherapy. Participants were asked to review the concept of the tool, perform think-aloud tasks, and complete the System Usability Scale and a Perceived Usefulness questionnaire. Usability problems were extracted from the interview data by independent coding and mapped to an eHealth Usability Problem Framework.

Results: We included 21 participants in the study, aged 29 to 73 years; 13 participants who had received immunotherapy and 8 participants who had not received immunotherapy. In total, 76 usability problems were identified. A total of 22 usability problems were in the task-technology fit category of the usability framework, mostly regarding the coaching and symptom functionality of the prototype. Critical problems regarding the symptom monitoring functionality were mainly found by participants who had

XSL•FO
RenderX

received immunotherapy. For 8 out of 10 statements in the Perceived Usefulness questionnaire, more than 75% of participants agreed or strongly agreed. The overall mean System Usability Scale score was 80 out of 100 (SD 11.3).

Conclusions: Despite identified usability issues, participants responded positively to the Perceived Usefulness questionnaire regarding the evaluated tool. Further analysis of the usability problems indicates that it was essential to include participants who matched the target end users. Participants treated with immunotherapy, specifically with previous experience in immune-related adverse events, encountered critical problems with symptom reporting that would not have been identified if these participants were not included. For other tasks and functionalities, it seems likely that loosening the inclusion criteria would have resulted in sufficient feedback without critical missing usability issues.

(JMIR Form Res 2025;9:e57659) doi:10.2196/57659

KEYWORDS

oncology; usability; usefulness; symptom management; coaching system; patients with cancer; immune checkpoint inhibitors; comparative qualitative study; medication; eHealth applications; caregivers; cancer treatment; patient education; well-being interventions; acceptability; melanoma; renal cell carcinoma; immunotherapy

Introduction

The prognosis for patients with several types of cancer has substantially improved following the introduction of immune checkpoint inhibitors, a novel form of immunotherapy that incites the patient's own immune system to attack the cancer cells [1-4]. The treatment has improved prognosis, but also often incites side effects, ranging from mild to potentially life-threatening. In addition, patients may experience symptoms from the cancer itself, as well as the stress of the disease and treatment on their mental well-being. Consequently, patients may experience a diminished health-related quality of life [2-5]. Previous research has shown that patients with cancer often have unmet care and information needs during their treatment and follow-up. eHealth apps may support health providers in addressing these needs [6,7].

For example, eHealth apps in cancer treatment are used to facilitate timely symptom reporting by the collection of patient-reported outcomes, by providing information for patients and caregivers on diagnosis, treatment, and side effects, and by giving patients access to home interventions for physical and mental well-being [8-12]. While there is conclusive evidence on the impact of eHealth on perceived support and knowledge levels, there are inconsistent findings for outcomes related to quality of life, self-management, and physical or mental well-being [6,13,14]. These potential benefits of eHealth apps are partially dependent on their ease of use, Perceived Usefulness, and eventual user acceptance [15,16].

It is known that user-centered design (UCD) processes for these types of eHealth apps may benefit system usability and user acceptance [16-18]. Better usability can lead to benefits such as enhanced user well-being and reduced risk of harm [19]. A significant part of UCD is evaluating apps using a usability evaluation method (UEM). Most UEMs highlight the importance of doing so with the intended end users of the product in order to accurately extract and understand the usability problems of target users [20]. The International Organization for Standardization standard for health and wellness apps states if the app is specifically designed to cater to individuals with a particular health condition, testing should involve participants with that condition [21]. However, the characteristics of the intended end users for health apps might not be strictly defined, or the intended end users might consist of a broadly defined, heterogeneous patient population [22]. Thus, it is not always clear which user characteristics should be considered when recruiting representative participants. In addition, it may be difficult for researchers to find patients who fit a particular profile, forcing them to adjust their inclusion criteria.

Prior to this study, we developed a prototype of the eHealth tool Cancer Patients Better Life Experience (CAPABLE). This tool facilitates symptom reporting, patient education, and well-being interventions [23]. The prototype was developed in an iterative manner based on UCD principles [6]. At the start of prototype development, preliminary and semistructured interviews were conducted with patients, caregivers, and health care professionals, to elicit their specific support needs and requirements for an eHealth tool [7]. These requirements were translated into functionalities of the prototype, followed by a preliminary usability test using heuristic evaluation.

The CAPABLE tool is intended for melanoma and renal cell carcinoma patients treated with immunotherapy, and their health care providers. During development, we aimed to consider various aspects of implementation for these different patient populations in different countries. The pilot study of CAPABLE will focus on these specific patient populations [24]. However, it is intended to be useful for a more general population of patients with cancer who are undergoing treatment. It is not known whether certain usability problems would only be evident to patients whose condition closely resembles patients in the trial, or if a broader group of patients would lead to the identification of a broader range of potential usability issues. This is particularly relevant for systems like CAPABLE, intended for use by patients with serious health complaints, who ethically should only be asked to participate in tasks where their time and effort are truly needed.

Therefore, we aimed to evaluate the CAPABLE prototype by conducting tests to assess usability, user experience, and perceived acceptability among end users. A secondary objective of the study is to assess any agreements or differences in the results of our wide range of participants in the usability studies, considering the target population of our eHealth intervention (patients with melanoma and renal cell carcinoma treated with immunotherapy) and participants that are part of a broader

XSL•FO
RenderX

population (patients with cancer with other types of cancer and informal caregivers).

Methods

Overview

The CAPABLE prototype was developed by a multidisciplinary Consortium, as part of the CAPABLE Project [23]. The overall aim of the project is to explore the effect, usability, and feasibility of the CAPABLE tool in a pilot study with patients with melanoma and renal cell carcinoma, during treatment with immune-checkpoint inhibitors [24]. This usability study was performed as part of the UCD process, through a think-aloud approach where participants were asked to execute scenario-based tasks using the CAPABLE prototype [20,25]. We followed the COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist for reporting our results (Multimedia Appendix 1).

The CAPABLE Prototype

The CAPABLE prototype consists of a mobile app for patients and a web-based dashboard for their health care providers. This usability study focuses on the mobile app for patients. Software from Invision was used to create a full clickable prototype of the CAPABLE app. See Figure 1 for an excerpt of screens from the CAPABLE prototype app screens [26].

The mobile app for patients consists of five different sections to facilitate symptom reporting, well-being interventions, and patient education (Table 1).

Figure 1. Excerpt of screens from the CAPABLE mobile app prototype evaluated during the usability study. CAPABLE: Cancer Patients Better Life Experience.

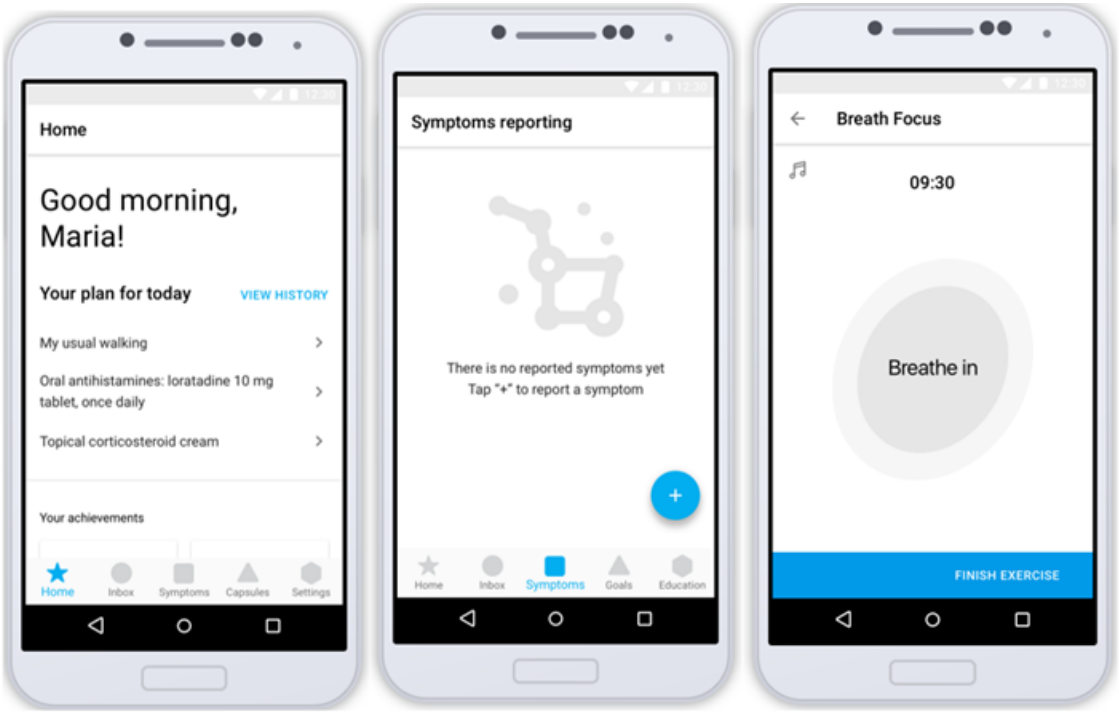

Table 1. Sections of the CAPABLE mobile app with their goals and content for supporting patients treated with immunotherapy, including symptom reporting, well-being interventions, and patient education.

Sections	Goals and contents
Home page	Contains a daily plan for patients to follow, including their hospital appointments, links to patient questionnaires, and suggestions for well-being interventions.
Inbox	Contains messages, recommendations, and reminders regarding their symptoms and well-being interventions.
Symptom reporting	Allows users to report their symptoms and provides feedback based on implemented computer-interpretable clinical guidelines for the management of symptoms related to immunotherapy side effects. Users report symptoms by selecting from a predefined set of descriptions that detail both the severity of symptoms and their impact on daily activities. After reporting, they receive feedback such as self-care instructions or advice to contact their health care provider.
Goals	Provides well-being goals such as improving sleep, physical well-being, and mental well-being, and offers corresponding theory-driven digital behavior change interventions including breathing exercises, meditation, and walking activities [27].
Education	Contains information about melanoma and renal cell carcinoma, treatments including targeted therapy and immunotherapy, side effects, supportive care, and nutrition.

XSL•FO
RenderX

Recruiting and Inclusion

Three recruiting organizations participated in the study, the Netherlands Cancer Institute-Antoni van Leeuwenhoek (NKI-AvL) in The Netherlands, Istituti Clinici Scientifici Maugeri (ICSM) in Italy, and the Italian Association of Cancer Patients, Relatives and Friends (AIMAC).

We aimed to recruit three participant groups; individuals diagnosed with melanoma previously treated with immunotherapy, those with renal cell carcinoma previously treated with immunotherapy, and participants without direct experience with the targeted cancer types nor with immunotherapy (thus patients with other types of cancer or informal caregivers). For each participant group, our goal was to recruit seven to nine participants. We did not specifically aim for data saturation as we considered the recommended sample sizes found in the literature, which range from five to ten participants. We chose seven to nine participants per group to reduce the risk of missing usability problems while being sensitive to time and resource constraints [28,29]. The inclusion criteria and recruitment strategies are described in Table 2.

Table 2. Inclusion criteria and recruitment strategies across the three participating centers.

	NKI-AvL[a]	ICSM[b]	AIMAC[c]
Participant	• Adult patients	• Adult patients	• Adult AIMAC members or their informal caregivers
Diagnosis	• High risk (resectable stage III) or advanced (stage IV and unresectable stage III) melanoma	• Renal cell carcinoma	• Any type of cancer
Treatment	• During or after treatment with immune checkpoint inhibitors	• During or after treatment with immune checkpoint inhibitors	• Any type of treatment
Language	• Sufficient understanding of the Dutch language	• Sufficient understanding of the Italian language	• Sufficient understanding of the Italian language
Recruitment	• Purposive sampling strategy to obtain a sample that varied in age • Participants were invited by their treating clinician, face-to-face or by telephone	• Purposive sampling strategy to obtain a sample that varied in age • Participants were invited by their treating clinician, face-to-face or by telephone	• Open enrollment recruitment from volunteering patient network

[a]NKI-AvL: the Netherlands Cancer Institute-Antoni van Leeuwenhoek.

[b]ICSM: Istituti Clinici Scientifici Maugeri in Italy.

[c]AIMAC: the Italian Association of Cancer Patients, Relatives, and Friends.

Ethical Considerations

This study was approved by the medical ethics committee of Amsterdam Universitair Medisch Centrum (Amsterdam; ID 2023.0944). Written informed consent was obtained from all participants prior to the interviews. All data collected for this study were pseudonymized with a new identifier. The research participants were not compensated for their study participation.

Interviews

Overview

The interviews were performed by two research teams. The interviews at NKI-AvL were performed in Dutch by two female PhD candidates with prior interviewing experience (IF and SLCG). The interviews at ICSM and AIMAC were performed in Italian by a female researcher (VT), a male senior researcher (MO), and a male research assistant with previous experience as a health care professional (Federico Dagostin). Two participants from NKI-AvL had participated in a previous interview with the NKI-AvL research team to elicit specific needs and requirements for the CAPABLE tool. The researchers had no clinical relationship with nor did know the remaining participants. A collaborative training session was held with all interviewers to streamline the UEMs used in the interviews as much as possible.

Due to the COVID-19 pandemic, interviews were done digitally using videoconferencing tools Teams (Microsoft Corp) and Zoom (Zoom Video Communications). The interviews were recorded using the screen- and audio-capture functionalities of these tools. The planned duration of the interviews was 45-60 minutes. The interviews were conducted in three phases: (1) the introduction, where participants were asked their opinion on the CAPABLE concept; (2) completing think-aloud tasks, and (3) filling in questionnaires. Participants were aware that the interviewers were researchers involved in the development of the CAPABLE prototype. See Multimedia Appendix 2 for the interview protocol.

Phase I: Introduction

The interview started with a few questions regarding patient demographics, and their previous experience with smartphones and technology. This was followed by a short presentation about the CAPABLE system and a few open-ended questions regarding the participant's opinion on the proposed goals and functionalities of the mobile app. Then, the interviewer explained the concept of the think-aloud phase of the interview. The participants were informed that the purpose of the

XSL•FO
RenderX

think-aloud tasks was to evaluate the app's performance and that the method required them to talk about what they were doing and thinking whilst using the app.

Phase II: Think-Aloud Tasks

The participants were asked to complete five tasks in total. These tasks were developed in cooperation with research team members and the developer. The tasks were (1) go through the introduction, check notifications, and report an activity; (2) report an itch symptom in the patient role; (3) report a fever symptom in the caregiver role; (4) find and perform a deep breathing exercise; and (5) find and review information about skin toxicity.

After performing the tasks, participants were asked about their final opinions, suggestions, or functionalities that they missed in the app.

Phase III: Questionnaires

Next, we administered the System Usability Scale (SUS) survey, a 10-item questionnaire aimed to assess the usability of a system considering effectiveness, efficiency, and satisfaction [30]. Finally, the participants were asked to fill out an 11-item questionnaire on Perceived Usefulness, based on the Technology Acceptance Model [31]. For both questionnaires, each item contains a statement that the participant is asked to rate based on a 5-point scale, ranging from strongly agree to strongly disagree (Multimedia Appendix 3).

Data Analysis

Participant Characteristics

The participants' characteristics were analyzed using descriptive statistics. We divided participants into two groups: treated with immunotherapy and not treated with immunotherapy.

Task Completion

We assessed the effectiveness of the participants performing the tasks according to three measures: (1) completed with ease, (2) completed with difficulty, and (3) failed to complete. We defined completed with difficulty as completed while needing to receive hints from the interviewer. Giving hints was done to ensure that the participant would have the opportunity to review all the content of the app and allow the researchers to obtain additional information about the usability of the app regardless of the completion status of the task.

Usability Problems

The Dutch interview recordings at NKI-AvL were transcribed verbatim (SLCG) as source data for our coding and analysis of the usability problems. During the Italian interviews at ICSM and AIMAC, notes were made by the interviewer that summarized the participants' feedback. These notes were revised and completed after the interview upon reviewing the recording. The transcripts and notes were not returned to the participants.

As two research teams performed the interviews and the data used to code the interviews varied (ie, transcripts and summarizing notes), we conducted a data quality check. Two Italian interviews and two Dutch interviews were transcribed and independently coded by researchers from both research teams (SLCG and VT). The codes based on the transcript were compared to those based on the summarizing notes of the Italian interviews to evaluate their consistency. In addition, the independently coded interviews were compared to ensure whether the interpretation of the data was similar between the two research teams. Transcripts and notes were considered comparable if both reviewers agreed that there were no substantial differences or omissions in the code's results.

Coding and Analysis of Interview Recordings

Several steps were taken to code and analyze the interview recordings. This was executed by the same researchers who conducted the Dutch interviews (SLCG and IF), plus NB, an MSc in Medical Informatics. The coding and analysis were done in English. See Figure 2 for an overview of these steps.

XSL•FO
RenderX

Figure 2. Steps executed for open coding of the usability interview transcripts and notes and analysis of the usability problems by mapping to the eHealth Usability Problem Framework. FD: Federico Dagostin.

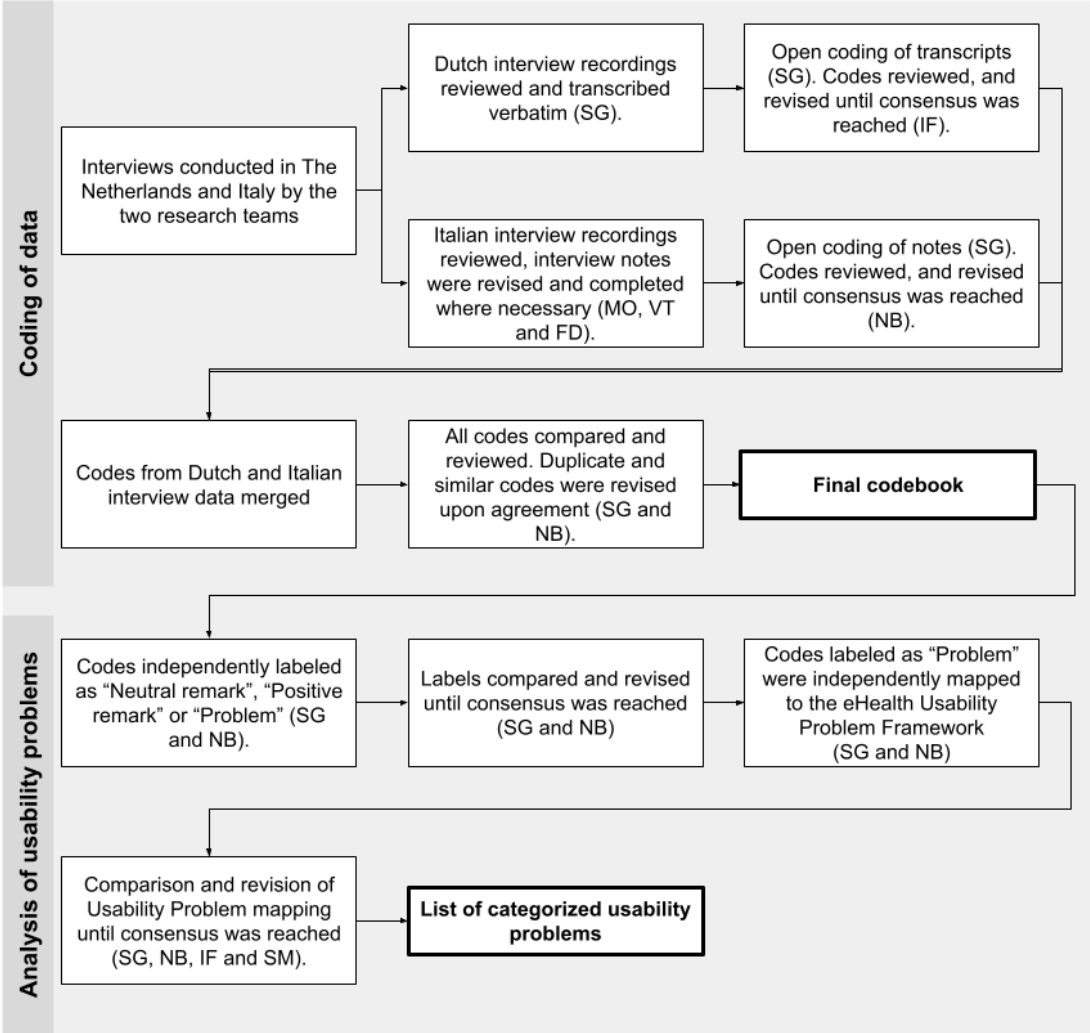

First, the interviews were open-coded by SLCG. The codes from the Dutch and Italian interviews were reviewed by IF and NB, respectively, until a consensus was reached. After merging the data and codes from the Dutch and Italian interviews, an extra revision step was executed to ensure that duplicate and similar codes were revised, upon agreement of SLCG and NB. This resulted in the final codebook. All codes in the codebook were independently labeled by SLCG and NB as "Positive remark," "Neutral remark," or "Problem" to allow us to focus on the usability problems indicated by the participants. Then, we mapped the "Problem" codes in the codebook to an eHealth Usability Problem Framework developed by Broekhuis et al [32]. This framework specifies 21 usability factors in the following categories: basic system performance, task-technology fit, accessibility, interface design, navigation and structure, information and technology, guidance and support, and satisfaction. This mapping process was done by SLCG and NB, in three different cycles, independently mapping and comparing the mappings, reaching consensus. Any discrepancies were discussed with IF and SM.

To gain insight into the usability problems and the frequency with which they were encountered, we counted the distinct problems associated with each usability factor (eg, the number of issues related to the usability factor "design clarity" under the category "interface design." To address our secondary objective, we noted which participants had observed specific problems, and counted how many were identified by both groups (those treated with immunotherapy and those not treated with immunotherapy) versus problems identified by only one group. Furthermore, for each distinct usability problem, we noted the number of participants who identified the problem and compared how frequently the problem was reported across the two participant groups.

We provide a more detailed account of the usability problems in the "Qualitative Assessment" section. We adopted a UCD approach and a postpositivist perspective. The eHealth Usability Problem Framework was used as a theoretical framework for our analysis. We included the problems most frequently identified, as well as those where there was a notable difference in identification between the two participant groups.

The qualitative analysis was conducted without the use of any specialized software. The quantitative analysis of counts and frequency was performed using Excel (Microsoft Corp).

XSL•FO
RenderX

Perceived Usefulness and SUS Questionnaires

The SUS survey results were evaluated using the standard SUS scoring algorithm. We followed the general guideline for interpretation of the results, with scores below 68 indicating poor usability, and scores above 68 indicative of good usability. The Perceived Usefulness results were analyzed using frequency analysis.

Results

Participant Characteristics

We conducted 21 interviews in total. The interviews took place between June 2021 and April 2022. The average age of the participants was 53 (SD 12.1) years. We interviewed 7 patients with melanoma, 6 patients with renal cell carcinoma, 6 patients with other types of cancer (breast cancer, urinary cancer, lymphoma, and vestibular schwannoma), and 2 caregivers. See Table 3 for an overview of the participants' characteristics. The participants with melanoma and renal cell carcinoma were treated with immune checkpoint inhibitors. All participants used a smartphone and were familiar with mobile apps. Five participants mentioned having used a health-related mobile app before.

Table 3. Characteristics of participants that were interviewed for this usability study (between June 2021 and April 2022).

	Treated with immunotherapy?	
	Yes (n=13)	No, N/A[a] (caregiver; n=8)
Sex, n (%)		
Male	6 (29)	6 (29)
Female	7 (33)	2 (10)
Age (years)		
Mean (SD)	52 (11)	56 (14)
Range	32-71	29-73
Type of cancer, n (%)		
Melanoma III-IV	7 (33)	0 (0)
Kidney I-II	4 (19)	0 (0)
Kidney III-IV	2 (10)	0 (0)
Breast I-II	0 (0)	1 (5)
Breast III-IV	0 (0)	2 (10)
Lymphoma	0 (0)	1 (5)
Urinary IV	0 (0)	1 (5)
Vestibular schwannoma I-II	0 (0)	1 (5)
N/A (caregiver)	0 (0)	2 (10)
Treatment status, n (%)		
On treatment	8 (38)	4 (19)
Off treatment	5 (24)	2 (10)
N/A (caregiver)	0 (0)	2 (10)
Living situation, n (%)		
Alone	2 (10)	1 (5)
With relatives	11 (52)	6 (29)
Shared housing	0 (0)	1 (5)
Type of smartphone, n (%)		
Android	6 (29)	5 (24)
iOS	7 (33)	3 (14)

[a]Not applicable.

XSL•FO
RenderX

Task Completion

We measured the effectiveness of the participants performing tasks by completion rates (Table 4). Participants found Task 1, "Go through introduction, check notifications, and report an activity," and Task 3, "Report a fever symptom in a caregiver's role," the most challenging. Task 5, "Find and review information about skin toxicity," had the highest completion rate. Participants who were not treated with immunotherapy seemed to struggle more with the symptom-reporting tasks (Tasks 2 and 3).

Table 4. Task completion rates of tasks performed during the think-aloud part of the usability interview, divided by participant group (participants previously treated with immunotherapy and not treated with immunotherapy).

	Task 1: introduction and home page		Task 2: symptom reporting and response inbox (itch as patient)		Task 3: symptom reporting and response inbox (fever as caregiver)		Task 4: coaching, goals, and activities		Task 5: patient education	
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No
Completed chemotherapy with ease, n (%)	8 (62)	5 (63)	11 (85)	4 (50)	10 (77)	3 (38)	9 (69)	5 (63)	11 (85)	8 (100)
Completed chemotherapy with difficulty, n (%)	5 (38)	3 (37)	2 (15)	4 (50)	3 (23)	5 (62)	4 (31)	3 (37)	2 (15)	0 (0)

Usability Problems

In the four interviews used to assess data quality, no substantial differences were noted in the results from the Italian and Dutch research teams, nor were differences between the analyses based on transcripts compared to the analyses from notes. The participants identified 76 distinct usability problems and proposed 15 additional features missing in the CAPABLE app. The first task "Go through introduction, check notifications, and report an activity" resulted in 31 issues; the second and third tasks "Report an itch symptom in a patients' role" and "Report a fever symptom in a caregivers' role" resulted in 24 issues; the fourth task "Find and perform a breathing exercise" resulted in 16 issues; and the final fifth task "Find and review information about skin toxicity" resulted in 12 issues. Eight problems recurred in multiple tasks.

Table 5 shows an overview of the eHealth Usability Problem Framework [32] and the count of usability problems found per usability factor. The category with the highest number of usability problems is the task-technology fit category, which relates to the match between the system on the one hand, and the user, their context of use, and their health goals on the other hand. Of these, 23 usability problems were found by both participants who received immunotherapy and participants who did not receive immunotherapy, 31 problems were solely found by participants who received immunotherapy, and 22 problems were solely found by participants who did not receive immunotherapy. Table 6 shows an overview of all usability problems that were indicated by more than one participant, resulting in 32 problems. See Multimedia Appendix 4 for the complete table of usability problems. We discuss the usability problems that participants most frequently indicated during the study in the following section.

XSL•FO
RenderX

Table 5. Number of distinct usability problems identified during the think-aloud part of the usability interview[a].

Usability factors	Number of issues found by both groups	Number of issues found only by the immunotherapy group	Number of issues found only by the nonimmunotherapy group	Number of total issues
Basic system performance	1	1	0	2
General system interaction	1	1	0	2
Task-technology fit	6	9	7	22
Fit between system and context of use	1	2	1	4
Fit between system and health goals	3	2	3	8
Fit between system and user	2	5	3	10
Interface design	2	5	4	11
Design clarity	0	2	0	2
Interface organization	1	1	1	3
Readability of texts	1	0	2	3
Symbols, icons, and buttons	0	2	1	3
Navigation and structure	4	0	1	5
Navigation	3	0	1	4
Structure	1	0	0	1
Information and terminology	6	3	6	15
Health-related information	4	3	5	12
System information	2	0	1	3
Guidance and support	1	7	1	9
Procedural health-related information	1	6	1	8
Procedural system information	0	1	0	1
Satisfaction	3	6	3	12
Satisfaction with system	2	2	1	5
Satisfaction with the system's ability to support health goals	1	4	2	7
Total	23	31	22	76

[a]The counts are presented per usability factor and usability factor category.

XSL•FO
RenderX

Table 6. List of usability problems including the count of participants that identified these usability problems, divided by group (treated with immunotherapy vs not treated with immunotherapy)[a].

	Treated with immunotherapy?	
	Yes	No
Introduction and home page		
Task-technology fit		
Current list of hobbies not sufficient	2	2
Expects automatic detection of activities by smartwatch	3	1
Time of going to bed is different every day	2	0
Unclear what the added value is of recording exercises or activities	2	0
Allow user to write down why they did or did not like the challenge	2	0
Not possible to select multiple hobbies	1	1
Need for balance between coaching, support for symptoms, and support for cancer treatment in content	0	2
Interface design		
Text is too long	5	2
Participant prefers graphical explanations to textual explanations	0	2
Navigation and structure		
Link between times of waking up or going to bed and the symptom management unclear	1	1
Information and terminology		
Unclear what vital functions in home page are	2	0
Satisfaction		
Tone in the introduction text is not appreciated	1	1
Symptom reporting and response inbox		
Basic system performance		
Not clear that scrolling was necessary to view everything in screen	6	5
Task-technology fit		
Information missing about medication in recommendation (dosage, need for prescription, where to get it)	7	1
Symptom descriptions do not match experience of itch of the patient, would be difficult to choose	6	0
Feedback missing after report	1	5
Navigation and structure		
Participant cannot find symptom section easily	3	1
Information and terminology		
Term caregiver is unclear, caregiver can be a professional or family or friends visiting	1	1
Guidance and support	0	0
Unclear if the clinician will view the report, if patient will be contacted, and what is expected of patient	4	0
Missing from instructions that caregivers can report symptoms	2	0
Satisfaction		
Symptom reporting process seems long or steps redundant	1	1
Participant does not trust the feedback from the app	1	1
Coaching, goals, and well-being interventions		
Task-technology fit		
Expects automatic detection of activities by smartwatch	0	3
Unclear how users can set their own goals	0	2
Unmet expectation of the app recommending a schedule with activities, which can be personalized	2	2

XSL•FO
RenderX

	Treated with immunotherapy?	
	Yes	No
Navigation and structure		
Participant cannot find the exercise easily	7	2
Information and terminology		
Names of types of breathing exercises are not self-explanatory	4	1
Menu term “objectives or goals” does not match with content found	2	1
Guidance and support		
Unclear how goals relate to the activities	4	0
Invitation for well-being intervention is not clear, not self-explanatory what it is and how to proceed	10	7
Satisfaction		
Content, purpose, and benefits of well-being interventions not sufficient currently	3	0
Patient education		
Interface design		
Reorganize the categories and structure of the educational content list	1	3
Navigation and structure		
Participant cannot find educational section easily (clicks on symptoms first)	1	3
Information and terminology		
First part of the text is difficult to understand without in-depth knowledge	1	2
Education not the correct term for section	1	2
Rash and itching are not translated	1	1
Total	128	85

[a]Problems included in this table were identified by at least two participants.

Qualitative Assessment

Introduction and Home Page

Task-technology fit: the CAPABLE app asks the user to select their hobbies from a predefined list. This selection is used to recommend similar activities. Participants could not find their own hobbies in this list and were unable to select more than one hobby.

Interface design: in the introduction, the text to explain the purpose and functionalities of the CAPABLE app was deemed too long by some participants and participants stated a preference for more graphical explanations.

> *Yes, it [the text] was long, and the concentration of some patients might be decreased, so a video might be easier.* [Participant 20]

Symptom Reporting and Feedback Inbox

In general, it was not always clear to participants that scrolling was necessary to view everything on screen.

Task-technology fit: participants who received immunotherapy treatment predominantly experienced problems when reporting an “Itch” symptom using the symptom reporting function in the app. The app presents a set of symptom descriptions to choose from. Nearly half of the participants in the immunotherapy group found it difficult to choose one of those descriptions, which did not match their own experiences with itching caused by immunotherapy:

> *It’s very black-and-white, isn’t it? Yes, I had constant itching, all day long, but I could still do activities. And that’s not reflected in there. You really should give a lot more options ... I had constant itching, but of course, I could still go grocery shopping, and wash myself, and, sleeping was really difficult though, whole nights without sleep.* [Participant 21]

During the think-aloud task, the app recommended an emollient cream for the itch symptom reported by the participants. While participants appreciated the advice, a majority of the participants in the immunotherapy group found the message incomplete, as they were not sure what the exact dosage of the cream would be, if a prescription would be needed, and where to acquire the medication.

Guidance and support: after reporting the symptom, participants were confused about the next steps in the process. A subset of participants from the immunotherapy group expressed uncertainty on whether a clinician would view the report, in which cases they would be contacted by the clinician, and what was expected of them as a patient.

> *That may still need to be coordinated, like, when should the patient or caregiver take action themselves if they notice things are not going well, and when*

XSL•FO
RenderX

> *does the team proactively step in or call to check how things are going or if they can offer any help? would have expected if you report [symptoms] in this manner, that your care team would be informed.* [Participant 15]

Coaching, Goals, and Activities

Task-technology fit: the section "Goals" in the app contains a list of exercises and activities. These can be filtered by their goal, including supporting mental well-being, physical well-being, sleep, and acceptance. The connection between the name of the section and the list of activities was perceived as unclear and confusing by some participants. As a result, when asked to find a breathing exercise in the app, it was not always obvious for participants to click on the "Goals" button from the home page. In addition, while participants appreciated the ability to easily select and do an exercise, it was unclear how often to do these activities and how to schedule activities in the daily plan on the home page of the app:

> *[I expected] some sort of automatic coach within the system that informs you: you've chosen mental well-being as your focus, so every night before going to sleep, do fifteen minutes of meditation.* [Participant 8]

Participants were also confused about wearing a smartwatch while the app requires manual registration of exercises such as walking.

Guidance and support: users receive an invitation in the CAPABLE app inbox to participate in an activity, such as a daily walking challenge. While users are immediately prompted to accept this challenge, most participants found it unclear in the invitation what is expected of them, how they should proceed, where they can see the content of this program in the app, and whether and where they should register to do these activities.

Patient Education

Navigation and structure: participants were asked to find information about skin adverse events due to immunotherapy. Some participants had trouble finding the correct section in the app and clicked on the "Symptom management" section instead of the "Education" section.

Information and terminology: some participants indicated that the text in the example section shown during the interview was challenging to understand without in-depth knowledge.

Perceived Usefulness and SUS Questionnaires

The participants filled in a questionnaire aiming to measure their acceptance and perceived usefulness of the CAPABLE app. Overall, participants were positive about nearly all statements. For eight out of ten questions, more than 75% of participants agreed or strongly agreed (Figure 3). The statements that were most widely agreed upon were that the system would easily fit in users' daily routines, could help health care professionals to follow up on patients' well-being, could improve communication with the care team, could help users cope with their treatment, and could help users support their quality of life. Participants were more skeptical about the CAPABLE app helping to manage emotions such as anxiety and stress and the ability of the app to help cope with daily life problems. The results of the questionnaire were comparable between the participant group treated with immunotherapy, versus the group not treated with immunotherapy. Finally, the participants were asked to fill in the SUS questionnaire. The mean SUS score was 80 (SD 11), which is classified as "good" in terms of usability.

Figure 3. Outcomes of the Perceived Usefulness questionnaire filled in at the end of the usability interview by participants, presented as a stacked bar chart. CAPABLE: Cancer Patients Better Life Experience; HCP: health care provider.

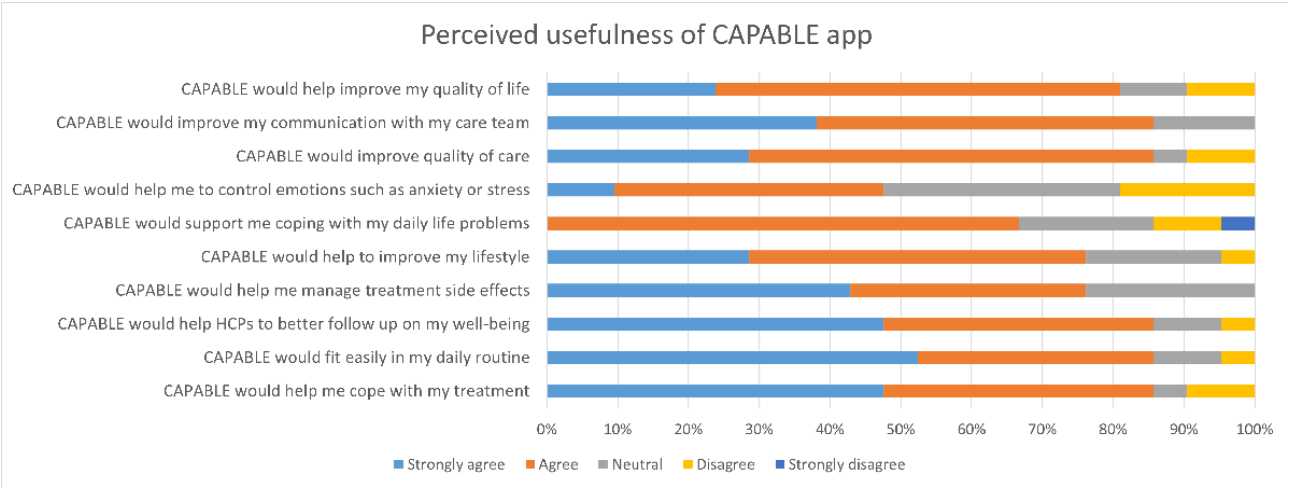

Discussion

Principal Findings

We evaluated the prototype of the CAPABLE app, a symptom monitoring and coaching system, with 21 participants divided into two groups, participants that had received immunotherapy (patients with melanoma and renal cell carcinoma) and that had not received immunotherapy (other cancer types), from The Netherlands and Italy. This evaluation was executed by use of think-aloud interviews and two questionnaires. In total, 76 usability issues were identified. Specifically, 23 usability issues were identified by both groups. The immunotherapy group

XSL•FO
RenderX

found 31 additional issues, and the nonimmunotherapy group found 22 additional issues.

Most usability problems were in the task-technology fit category of the eHealth Usability Problem Framework [33]. This is reflected in problems encountered in the coaching section of the app. The participants' mental model, defined as what a user knows, believes about, and expects from a system [34], did not match with the design and presentation of the digital behavior change interventions. Participants expected to be able to set their own goals, and thought that the app would suggest a recommended weekly schedule of activities that can be personalized according to their needs. Problems with task-technology fit were also encountered in creating a symptom report and with the subsequent advice given. Participants who had experienced itching due to immunotherapy found it difficult to choose one of the descriptions given to make a symptom report, as none of them matched their own experiences with the symptom. Additionally, participants felt that the advice provided was insufficient because it lacked information about the prescription, use, and dosage of the recommended medication, as well as whether to contact their clinician. This set of issues was mainly indicated by participants who had received immunotherapy. This is logical, since patients who received immunotherapy are more likely to have experienced a situation similar to the scenario used in the test, and could assess the fit of the app with their own lived experience. No differences were noted in the number and type of usability issues identified by caregivers compared to patients.

Despite identified usability issues, participants responded positively to the perceived acceptance and usefulness questionnaire regarding the CAPABLE app. Specifically, they would expect that the CAPABLE app would facilitate at-home monitoring, help patients cope with treatment, and support their quality of life.

Overall, our analysis of the usability problems seems to indicate a necessity to include participants with the characteristics of the intended end users for the evaluation of certain functionalities. In our case, this meant having the symptom-reporting functionality evaluated by patients who had been treated with immunotherapy, specifically with previous experience in immune-related adverse events. For the other tasks and functionalities, it seems likely that loosening the inclusion criteria would have resulted in sufficient feedback without critical missing usability issues.

Interpretation and Impact

This study adds to previously published information by extracting a specific description of the usability problems encountered by intended end users during the evaluation of symptom management and coaching eHealth intervention. These findings are relevant for researchers developing and evaluating tools like symptom monitoring apps, especially for patients with cancer. More generally, we noted that patients who had taken immunotherapy noticed problems that other patients with cancer (and caregivers) did not. This is useful information for researchers who are considering whether patients who have experienced the specific health problem under study need to be recruited. It can be that these patients are quite ill, and researchers do not want to burden them. Our results suggest that the participation of these patients added valuable information, at least in this study, and thus asking for their participation in usability studies is justified.

Comparison With Prior Work

This study found positive perspectives from participants on measured perceived acceptance and perceived usefulness. This might seem contradictory to the number and severity of usability issues reported. This is reflected in other studies, where participants are enthusiastic about the functionalities and perceived future benefits, while simultaneously encountering difficulties during the usability evaluation [35-37]. Patients' wants and needs seem to be identified clearly in research, while a gap between user needs and eHealth tool implementation might be caused by practical considerations, the adaptability of the tool to local context, complexity factors, and health professionals' uptake of the eHealth app [38].

In addition, we investigated the impact of different user characteristics on usability problems found during a think-aloud evaluation. Previous studies have researched the impact of domain knowledge, specified as the familiarity or expertise an individual has with a particular topic or subject area, on finding usability problems [39]. In these studies, the output of novices and experts, with previous expertise with the software or tools evaluated, was compared. In some cases, novices found more usability problems, less usability problems, or less usability problems but more severe problems, compared to experts [40-42]. However, this study would define previous domain knowledge not as previous experience with the software, but as previous experience with the treatment (immunotherapy) or previous experience with certain side effects (itch). Previous research found a significant difference between the novice and expert group with previous domain knowledge defined as previous pregnancy [43]. More research is needed on the impact of previous experience with a disease or treatment in the oncology field as part of the users' profile on usability evaluation outcomes.

Strengths and Limitations

In this study, we were able to recruit a large and diverse participant group from multiple centers in both The Netherlands and Italy. This was a benefit to our goal of collecting broad feedback on the CAPABLE tool and aiming to find a complete overview of potential usability issues. We had researchers from two different research teams performing the interviews. Nearly all researchers had previous interviewing experience, and a collaborative training session was held with all interviewers to streamline the UEMs used in the interviews as much as possible to prevent bias. The interviews were qualitative in nature; however, the data from the think-aloud tasks was specifically used to identify usability problems, rather than for thematic analysis. No pilot testing was performed, but no changes were needed during the course of the interview rounds.

The raw data used for coding the interviews varied: interviews with participants with renal cell carcinoma and other types of cancer were coded using summarizing interview notes, while interviews with melanoma participants were coded using

XSL•FO
RenderX

transcriptions. This discrepancy may have led to missing data from incomplete interview notes. However, our data quality check revealed that although transcripts provided more context for the codes, the codes derived from the interview notes were consistent with those from the transcripts. In addition, we had multiple researchers involved in the coding and analysis phase of the study, to reduce the evaluator effect [44].

We did not take health, smartphone literacy, or physical or mental impairments into account during recruitment. There might be a selection bias of participants agreeing to the interview as they have an interest in using their mobile phones, or invitees refusing to participate because of limited smartphone literacy. However, we recruited a varied group of 21 participants, with varying ages, different diagnoses, past treatments, and countries.

Future Work

The usability results of this study were used to improve the CAPABLE prototype. A prospective exploratory pilot study involving the final version of CAPABLE is currently being held in The Netherlands and Italy, with different patient populations treated with immune checkpoint inhibitors, including patients with melanoma and renal cell carcinoma [24]. The clinical impact and usability of the CAPABLE tool will be evaluated.

Our comparison of the usability problems generated by participants treated with immunotherapy versus the nonimmunotherapy group indicated a difference in output mainly for the symptom-reporting functionality of the CAPABLE tool. This comparison was based on qualitative data. Multiple factors may have been involved in the differences of usability problems found in addition to previous experience with the treatment or side effects, such as age, diagnosis, stadium of disease, health literacy, and country. However, our findings implore future studies to focus on exploring and comparing the usability evaluation output of participants with varying previous experiences. In addition, future research is needed to determine for which types of software, or which specific functionalities, there would be an added benefit or need for participants with specific characteristics.

Conclusions

While participants identified usability problems regarding task-technology fit, interface design, and overall satisfaction, they responded positively regarding the perceived impact of CAPABLE in monitoring patients from home, helping to cope with treatment, and supporting quality of life. Further analysis of the usability problems indicates that it was essential to include participants who matched the target end users. Participants treated with immunotherapy, specifically with previous experience in immune-related adverse events encountered critical problems with symptom reporting that would not have been identified if these participants were not included. For other tasks and functionalities, it seems likely that loosening the inclusion criteria would have resulted in sufficient feedback without critical missing usability issues. Future research is needed to determine for which types of software and which specific functionalities, there would be an added benefit or need for participants with specific characteristics.

Acknowledgments

This work was supported by the European Union's Horizon 2020 research and innovation program (grant agreement 875052). The authors would like to thank all the participants who took part in the interviews.

Data Availability

The aggregated data is included in Multimedia Appendix 4. The unaggregated datasets generated during this study are not publicly available due to our institutions' privacy policies but are available from the corresponding author upon reasonable request and with approval from the respective institution.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Consolidated criteria for reporting qualitative studies (COREQ): 32-item checklist.
[DOCX File , 22 KB - formative_v9i1e57659_app1.docx]

Multimedia Appendix 2
Think-aloud interview protocol.
[DOCX File , 27 KB - formative_v9i1e57659_app2.docx]

Multimedia Appendix 3
System Usability Scale and perceived impact questionnaire.
[DOCX File , 24 KB - formative_v9i1e57659_app3.docx]

Multimedia Appendix 4
Overview of all usability problems and suggestions.

XSL•FO
RenderX

[DOCX File , 38 KB - formative_v9i1e57659_app4.docx]

References

1. Larkin J, Chiarion-Sileni V, Gonzalez R, Grob JJ, Rutkowski P, Lao CD, et al. Five-year survival with combined nivolumab and ipilimumab in advanced melanoma. N Engl J Med 2019;381(16):1535-1546 [FREE Full text] [doi: 10.1056/NEJMoa1910836] [Medline: 31562797]
2. Robert C. A decade of immune-checkpoint inhibitors in cancer therapy. Nat Commun 2020;11(1):3801 [FREE Full text] [doi: 10.1038/s41467-020-17670-y] [Medline: 32732879]
3. Hargadon KM, Johnson CE, Williams CJ. Immune checkpoint blockade therapy for cancer: an overview of FDA-approved immune checkpoint inhibitors. Int Immunopharmacol 2018;62:29-39. [doi: 10.1016/j.intimp.2018.06.001] [Medline: 29990692]
4. Shiravand Y, Khodadadi F, Kashani SMA, Hosseini-Fard SR, Hosseini S, Sadeghirad H, et al. Immune checkpoint inhibitors in cancer therapy. Curr Oncol 2022;29(5):3044-3060 [FREE Full text] [doi: 10.3390/curroncol29050247] [Medline: 35621637]
5. Rogiers A, Boekhout A, Schwarze J, Awada G, Blank C, Neyns B. Long-term survival, quality of life, and psychosocial outcomes in advanced melanoma patients treated with immune checkpoint inhibitors. J Oncol 2019;2019:5269062 [FREE Full text] [doi: 10.1155/2019/5269062] [Medline: 31182961]
6. Slev VN, Mistiaen P, Pasman HRW, Verdonck-de LIM, van UCF, Francke AL. Effects of eHealth for patients and informal caregivers confronted with cancer: a meta-review. Int J Med Inform 2016;87:54-67. [doi: 10.1016/j.ijmedinf.2015.12.013] [Medline: 26806712]
7. Fraterman I, Glaser S, Wilgenhof S, Medlock S, Mallo H, Cornet R, et al. Exploring supportive care and information needs through a proposed eHealth application among melanoma patients undergoing systemic therapy: a qualitative study. Support Care Cancer 2022;30(9):7249-7260 [FREE Full text] [doi: 10.1007/s00520-022-07133-z] [Medline: 35589878]
8. Kotronoulas G, Kearney N, Maguire R, Harrow A, Di Domenico D, Croy S, et al. What is the value of the routine use of patient-reported outcome measures toward improvement of patient outcomes, processes of care, and health service outcomes in cancer care? A systematic review of controlled trials. J Clin Oncol 2014;32(14):1480-1501 [FREE Full text] [doi: 10.1200/JCO.2013.53.5948] [Medline: 24711559]
9. Lubberding S, van Uden-Kraan CF, Te Velde EA, Cuijpers P, Leemans CR, Verdonck-de Leeuw IM. Improving access to supportive cancer care through an eHealth application: a qualitative needs assessment among cancer survivors. J Clin Nurs 2015;24(9-10):1367-1379. [doi: 10.1111/jocn.12753] [Medline: 25677218]
10. Iivanainen S, Alanko T, Vihinen P, Konkola T, Ekstrom J, Virtanen H, et al. Follow-Up of cancer patients receiving anti-PD-(L)1 therapy using an electronic patient-reported outcomes tool (KISS): prospective feasibility cohort study. JMIR Form Res 2020;4(10):e17898 [FREE Full text] [doi: 10.2196/17898] [Medline: 33112242]
11. Hægermark EA, Kongshaug N, Raj SX, Hofsli E, Faxvaag A. Design, development, and evaluation of an mHealth app for reporting of side effects during cytostatic treatment: usability test and interview study. JMIR Form Res 2023;7:e47374 [FREE Full text] [doi: 10.2196/47374] [Medline: 37856183]
12. Schläpfer S, Schneider F, Santhanam P, Eicher M, Kowatsch T, Witt CM, et al. Engagement with a relaxation and mindfulness mobile app among people with cancer: exploratory analysis of use data and self-reports from a randomized controlled trial. JMIR Cancer 2024;10:e52386 [FREE Full text] [doi: 10.2196/52386] [Medline: 38819907]
13. Chen J, Ou L, Hollis SJ. A systematic review of the impact of routine collection of patient reported outcome measures on patients, providers and health organisations in an oncologic setting. BMC Health Serv Res 2013;13:211 [FREE Full text] [doi: 10.1186/1472-6963-13-211] [Medline: 23758898]
14. van Deursen L, Versluis A, van der Vaart R, Standaar L, Struijs J, Chavannes N, et al. eHealth interventions for Dutch cancer care: systematic review using the triple aim lens. JMIR Cancer 2022;8(2):e37093 [FREE Full text] [doi: 10.2196/37093] [Medline: 35699991]
15. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989;13(3):319. [doi: 10.2307/249008]
16. Ardito V, Golubev G, Ciani O, Tarricone R. Evaluating barriers and facilitators to the uptake of mHealth apps in cancer care using the consolidated framework for implementation research: scoping literature review. JMIR Cancer 2023;9:e42092 [FREE Full text] [doi: 10.2196/42092] [Medline: 36995750]
17. De Vito Dabbs A, Myers BA, Mc Curry KR, Dunbar-Jacob J, Hawkins RP, Begey A, et al. User-centered design and interactive health technologies for patients. Comput Inform Nurs 2009;27(3):175-183 [FREE Full text] [doi: 10.1097/NCN.0b013e31819f7c7c] [Medline: 19411947]
18. Saparamadu AADNS, Fernando P, Zeng P, Teo H, Goh A, Lee JMY, et al. User-centered design process of an mHealth app for health professionals: case study. JMIR Mhealth Uhealth 2021;9(3):e18079 [FREE Full text] [doi: 10.2196/18079] [Medline: 33769297]
19. Maramba I, Chatterjee A, Newman C. Methods of usability testing in the development of eHealth applications: a scoping review. Int J Med Inform 2019;126:95-104. [doi: 10.1016/j.ijmedinf.2019.03.018] [Medline: 31029270]

XSL•FO
RenderX

20. Jaspers MWM. A comparison of usability methods for testing interactive health technologies: methodological aspects and empirical evidence. Int J Med Inform 2009;78(5):340-353. [doi: 10.1016/j.ijmedinf.2008.10.002] [Medline: 19046928]
21. ISO/TS 82304-2:2021(en). Health software part 2: health and wellness apps. Quality and reliability [Internet]. URL: https://www.iso.org/obp/ui/en/#iso:std:iso:ts:82304:-2:ed-1:v1:en [accessed 2023-01-21]
22. Neal D, Engelsma T, Tan J, Craven MP, Marcilly R, Peute L, et al. Limitations of the new ISO standard for health and wellness apps. Lancet Digital Health 2022;4(2):e80-e82 [FREE Full text] [doi: 10.1016/S2589-7500(21)00273-9] [Medline: 35090677]
23. Home—CAPABLE [Internet]. URL: https://capable-project.eu/ [accessed 2021-04-13]
24. Fraterman I, Wollersheim BM, Tibollo V, Glaser SLC, Medlock S, Cornet R, et al. An eHealth app (CAPABLE) providing symptom monitoring, well-being interventions, and educational material for patients with melanoma treated with immune checkpoint inhibitors: protocol for an exploratory intervention trial. JMIR Res Protoc 2023;12:e49252 [FREE Full text] [doi: 10.2196/49252] [Medline: 37819691]
25. Jaspers MWM, Steen T, van den Bos C, Geenen M. The think aloud method: a guide to user interface design. Int J Med Inform 2004;73(11):781-795. [doi: 10.1016/j.ijmedinf.2004.08.003] [Medline: 15491929]
26. InVision. Collaborate better. URL: https://www.invisionapp.com/ [accessed 2023-10-13]
27. Lisowska A, Wilk S, Peleg M. SATO (IDEAS expAnded wiTh BCIO): workflow for designers of patient-centered mobile health behaviour change intervention applications. J Biomed Inform 2023;138:104276. [doi: 10.1016/j.jbi.2022.104276] [Medline: 36586499]
28. Faulkner L. Beyond the five-user assumption: benefits of increased sample sizes in usability testing. Behav Res Methods Instrum Comput 2003;35(3):379-383. [doi: 10.3758/bf03195514] [Medline: 14587545]
29. Nielsen J. Estimating the number of subjects needed for a thinking aloud test. Int J Hum-Comput Stud 1994;41(3):385-397. [doi: 10.1006/ijhc.1994.1065]
30. Bangor A, Kortum PT, Miller JT. An empirical evaluation of the system usability scale. Int J Hum-Comput Interact 2008;24(6):574-594. [doi: 10.1080/10447310802205776]
31. Davis FD. Sloan School of Management. A technology acceptance model for empirically testing new end-user information systems: theory and results [Thesis]. 1985. URL: https://dspace.mit.edu/handle/1721.1/15192 [accessed 2023-07-18]
32. Broekhuis M, van Velsen L, Peute L, Halim M, Hermens H. Conceptualizing usability for the eHealth context: content analysis of usability problems of eHealth applications. JMIR Form Res 2021;5(7):e18198 [FREE Full text] [doi: 10.2196/18198] [Medline: 34313594]
33. Peute LW, Driest KF, Marcilly R, Bras Da Costa S, Beuscart-Zephir M, Jaspers MWM. A framework for reporting on human factor/usability studies of health information technologies. Stud Health Technol Inform 2013;194:54-60. [Medline: 23941930]
34. Nielsen Norman group. Experience WL in RBU. Mental Models and User Experience Design. URL: https://www.nngroup.com/articles/mental-models/ [accessed 2024-02-12]
35. Mirkovic J, Kaufman DR, Ruland CM. Supporting cancer patients in illness management: usability evaluation of a mobile app. JMIR Mhealth Uhealth 2014;2(3):e33 [FREE Full text] [doi: 10.2196/mhealth.3359] [Medline: 25119490]
36. O'Malley DM, Davis SN, Devine KA, Sullivan B, Bator A, Clemow L, et al. Development and usability testing of the e-EXCELS tool to guide cancer survivorship follow-up care. Psychooncology 2020;29(1):123-131 [FREE Full text] [doi: 10.1002/pon.5222] [Medline: 31626397]
37. Duman-Lubberding S, van Uden-Kraan CF, Peek N, Cuijpers P, Leemans CR, Verdonck-de Leeuw IM. An eHealth application in head and neck cancer survivorship care: health care professionals' perspectives. J Med Internet Res 2015;17(10):e235 [FREE Full text] [doi: 10.2196/jmir.4870] [Medline: 26489918]
38. Ross J, Stevenson F, Lau R, Murray E. Factors that influence the implementation of e-health: a systematic review of systematic reviews (an update). Implement Sci 2016;11(1):146 [FREE Full text] [doi: 10.1186/s13012-016-0510-7] [Medline: 27782832]
39. Ju B. Does domain knowledge matter: mapping users' expertise to their information interactions. J Am Soc Inf Sci 2007;58(13):2007-2020. [doi: 10.1002/asi.20680]
40. Liu Y, Osvalder AL, Karlsson M. Considering the Importance of User Profiles in Interface Design. Croatia: NTECH; 2010.
41. Clarke MA, Belden JL, Kim MS. Determining differences in user performance between expert and novice primary care doctors when using an electronic health record (EHR). J Eval Clin Pract 2014;20(6):1153-1161. [doi: 10.1111/jep.12277] [Medline: 25470668]
42. Liu Y, Osvalder A, Karlsson M. User's expertise differences when interacting with simple medical user interfaces. In: Holzinger A, editor. HCI and Usability for Medicine and Health Care. Berlin, Heidelberg: Springer; 2007:441.
43. van den Berg L, van Beukering M, Velu A, Kok M. The impact of user characteristics on think aloud method results. Evaluation of a mHealth application for pregnancy-related work advice and information. URL: https://scripties.uba.uva.nl/ [accessed 2023-11-13]
44. Hertzum M, Jacobsen NE. The evaluator effect: a chilling fact about usability evaluation methods. Int J Hum-Comput Interact 2003;15(1):183-204. [doi: 10.1207/s15327590ijhc1501_14]

XSL•FO
RenderX

Abbreviations

AIMAC: Italian Association of Cancer Patients, Relatives and Friends
CAPABLE: Cancer Patients Better Life Experience
COREQ: Consolidated Criteria for Reporting Qualitative Research
ICSM: Istituti Clinici Scientifici Maugeri
NKI-AvL: Netherlands Cancer Institute-Antoni van Leeuwenhoek
SUS: System Usability Scale
UCD: user-centered design
UEM: usability evaluation method

Edited by A Mavragani; submitted 28.02.24; peer-reviewed by K Le Du, M Derksen; comments to author 09.09.24; revised version received 26.09.24; accepted 13.11.24; published 23.01.25.

Please cite as:
Glaser SLC, Fraterman I, van Brummelen N, Tibollo V, Del Campo LM, Mallo H, Wilgenhof S, Wilk S, Gisko V, Khadakou V, Cornet R, Ottaviano M, Medlock S
Usability and Usefulness of a Symptom Management Coaching System for Patients With Cancer Treated With Immune Checkpoint Inhibitors: Comparative Mixed Methods Study
JMIR Form Res 2025;9:e57659
URL: https://formative.jmir.org/2025/1/e57659
doi:10.2196/57659
PMID:

©Savannah Lucia Caterina Glaser, Itske Fraterman, Noah van Brummelen, Valentina Tibollo, Laura Maria Del Campo, Henk Mallo, Sofie Wilgenhof, Szymon Wilk, Vitali Gisko, Vadzim Khadakou, Ronald Cornet, Manuel Ottaviano, Stephanie Medlock. Originally published in JMIR Formative Research (https://formative.jmir.org), 23.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Improving Mental Health and Well-Being Through the Paradym App: Quantitative Study of Real-World Data

Athina Marina Metaxa[1], MSc; Shaun Liverpool[2], PhD; Mia Eisenstadt[3], PhD; John Pollard[4], MSc; Courtney Carlsson[4], MBA

[1]Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, United Kingdom
[2]Faculty of Health, Social Care and Medicine, Edge Hill University, Ormskirk, United Kingdom
[3]Evidence Based Practice Unit, Anna Freud National Centre for Children and Families, University College London, London, United Kingdom
[4]Paradym, London, United Kingdom

Corresponding Author:
Athina Marina Metaxa, MSc
Nuffield Department of Primary Care Health Sciences
University of Oxford
Radcliffe Observatory Quarter, Woodstock Rd
Oxford, OX2 6GG
United Kingdom
Phone: 30 6930228210
Email: athina.metaxa@hmc.ox.ac.uk

Abstract

Background: With growing evidence suggesting that levels of emotional well-being have been decreasing globally over the past few years, demand for easily accessible, convenient, and affordable well-being and mental health support has increased. Although mental health apps designed to tackle this demand by targeting diagnosed conditions have been shown to be beneficial, less research has focused on apps aiming to improve emotional well-being. There is also a dearth of research on well-being apps structured around users' lived experiences and emotional patterns and a lack of integration of real-world evidence of app usage. Thus, the potential benefits of these apps need to be evaluated using robust real-world data.

Objective: This study aimed to explore usage patterns and preliminary outcomes related to mental health and well-being among users of an app (Paradym; Paradym Ltd) designed to promote emotional well-being and positive mental health.

Methods: This is a pre-post, single-arm evaluation of real-world data provided by users of the Paradym app. Data were provided as part of optional built-in self-assessments that users completed to test their levels of depression (Patient Health Questionnaire-9), anxiety (Generalized Anxiety Disorder Questionnaire-7), life satisfaction (Satisfaction With Life Scale), and overall well-being (World Health Organization-5 Well-Being Index) when they first started using the app and at regular intervals following initial usage. Usage patterns, including the number of assessments completed and the length of time between assessments, were recorded. Data were analyzed using within-subjects *t* tests, and Cohen *d* estimates were used to measure effect sizes.

Results: A total of 3237 app users completed at least 1 self-assessment, and 787 users completed a follow-up assessment. The sample was diverse, with 2000 users (61.8%) being located outside of the United States. At baseline, many users reported experiencing strong feelings of burnout (677/1627, 41.6%), strong insecurities (73/211, 34.6%), and low levels of thriving (140/260, 53.8%). Users also experienced symptoms of depression (mean 9.85, SD 5.55) and anxiety (mean 14.27, SD 6.77) and reported low levels of life satisfaction (mean 12.14, SD 7.42) and general well-being (mean 9.88, SD 5.51). On average, users had been using the app for 74 days when they completed a follow-up assessment. Following app usage, small but significant improvements were reported across all outcomes of interest, with anxiety and depression scores improving by 1.20 and 1.26 points on average, respectively, and life satisfaction and well-being scores improving by 0.71 and 0.97 points, respectively.

Conclusions: This real-world data analysis and evaluation provided positive preliminary evidence for the Paradym app's effectiveness in improving mental health and well-being, supporting its use as a scalable intervention for emotional well-being, with potential applications across diverse populations and settings, and encourages the use of built-in assessments in mental health app research.

(JMIR Form Res 2025;9:e68031) doi:10.2196/68031

XSL•FO
RenderX

KEYWORDS

well-being; awareness; mental health; formative; mobile phone; well-being; apps; quantitative evaluation; real-world data; emotional well-being; pre-post; single arm; quantitative data

Introduction

Background

Emotional well-being has been widely recognized as a crucial factor in predicting overall health and quality of life [1]. Increased emotional well-being has been robustly associated with a lower risk of both physical and mental health problems, according to numerous studies [2,3]. Research findings from different countries indicate that emotional well-being has been declining within the general population over the last few years [4,5], a trend that was first observed following the beginning of the COVID-19 pandemic in 2020. Since then, problems with well-being and general satisfaction with life have intensified globally, and simultaneously, the demand for easily accessible, affordable, and effective support services has increased [4]. While research suggests that one effective way to enhance emotional well-being is by increasing emotional self-awareness, such a goal is not easily obtainable for many people facing mental health and well-being difficulties. Specifically, conventional face-to-face psychotherapeutic treatments are often inaccessible, usually due to high costs, long waiting times, and inflexible treatment schedules that are not accommodating to people's daily lives and work-related commitments [4,6]. As a result, mental health apps have been highlighted as a promising solution for delivering these services, given their affordability, ease of access, scalability, and ability to offer anonymous support [6].

In addition, mental health and well-being apps have also begun to be integrated into clinical practice; some therapists use such apps to enhance in-person treatment and provide patients with a tool they can use outside of face-to-face sessions to gain psychoeducational insights, monitor their symptoms, and reflect on their progress [7]. Therapists have expressed positive views about the integration of online mental health tools into their practice, as they can provide a streamlined way to organize and conduct therapy-related activities, improve patient's accessibility to therapy, and reduce the stigma associated with visiting a mental health professional [8]. Specific types of therapy, such as art therapy, can especially benefit from online mental health tools that enable the creation and sharing of visual arts media, which is a key method to fostering self-expression and well-being for patients [9]. However, concerns have been raised focusing on the high implementation costs associated with the integration of digital apps, both for the therapist and the patient, as well as the privacy, safety, and anonymity of app users, as apps collect sensitive personal information related to patients' mental and physical health and experiences [10]. In addition, clinicians have expressed some concerns about the robustness of the evidence base behind mental health apps, as some apps' content and structure are not clearly based on validated therapeutic techniques and have not been investigated in rigorous research studies [10,11]. However, emotional well-being apps have been increasingly studied and have mixed results with effectiveness. Of note, some studies show that, overall, such apps are valued by users for the privacy, personalization, and instant support that they can provide [12].

Emotional Well-Being

Emotional well-being is a key aspect of mental health that encompasses overall happiness and life satisfaction [13]. Emotional self-awareness, which refers to recognizing and expressing one's emotions as they occur [14], is crucial for achieving emotional well-being, as low levels of emotional awareness have consistently been linked to diminished emotional well-being, reduced quality of life, problems in interpersonal relationships, and a higher risk of mental health disorders like depression and anxiety [15,16]. In contrast, increased emotional awareness has been linked to better emotional regulation, increased positive emotions and quality of life, improvements in social functioning, and reductions in symptoms of depression and anxiety [17]. Based on these findings, both researchers and policymakers have highlighted emotional well-being as a means to improve overall mental and physical health and reduce the costs linked to poor mental health and the treatment of mental health disorders [1,2].

Emotional Well-Being Apps

In recent years, a growing number of apps have focused on the connection between emotional self-awareness and emotional well-being has led to a range of digital interventions that focus on this approach [18,19]. These apps, which can be easily accessed on smartphones and tablets, allow users to improve their emotional well-being at their convenience, offering privacy and flexibility, while they are usually either free or have a much lower annual cost than traditional face-to-face therapy [18,19].

A comprehensive review of 52 well-being apps revealed that most of these tools focus on helping users improve their emotional well-being and emotional awareness primarily through techniques such as mindfulness, meditation, cognitive behavioral therapy, mood tracking, and journaling [20]. Mindfulness-based apps, such as Headspace and Calm (Calm.com Inc), have gained popularity by encouraging users to incorporate meditation and breathing exercises into their daily routines. These practices are designed to heighten emotional awareness and help users manage their thoughts and feelings more effectively [21-23]. Similarly, other apps have integrated cognitive behavioral therapy strategies, which are aimed at helping users challenge and reframe negative thought patterns and engage in positive reappraisal [24,25]. Moreover, mood-tracking features in some apps enable users to record and monitor their emotions over time, often providing personalized suggestions for activities that can help regulate and improve their moods [26,27]. This combination of approaches highlights the diverse methods through which digital tools are addressing emotional well-being.

Despite the variety of techniques used, many well-being apps have been critiqued for not fully incorporating lived experiences or storytelling elements in their design [20]. When these

XSL•FO
RenderX

elements are included, they typically take the form of fictional narratives or real-life testimonials from individuals who have successfully navigated difficult emotional situations or life events [28]. These storytelling techniques can be powerful, offering users relatable examples and motivating them to engage more deeply with the app's content [29]. Research indicates that the incorporation of storytelling in therapeutic interventions can improve patient engagement and destigmatize mental health treatment [30]. Effective storytelling has also been shown to enable the delivery of novel health care interventions and capture the unmet needs of patients across demographic strata and treatment areas [31]. However, the limited use of these approaches suggests that there is untapped potential in using storytelling and lived experiences more extensively. In addition, very few apps encourage users to explore their emotional patterns over time. Such exploration could provide valuable insights into their personality traits and behavioral tendencies, offering a deeper understanding of their emotional landscape beyond the immediate moment [20].

Given the rapid growth of well-being apps and the limited research on those that use storytelling, lived experiences, or focus on emotional patterns, there is a clear opportunity for further exploration, which could help identify new strategies to increase user engagement and satisfaction with digital well-being tools, leading to better outcomes in emotional health management. As digital interventions continue to evolve, exploring these areas could play a crucial role in developing more holistic and user-centered tools that better meet the diverse needs of those seeking to improve their emotional well-being [20,32].

The Paradym App

The Paradym app was created in response to the increasing demand for scalable mental health solutions, driven by global challenges such as the COVID-19 pandemic and increased societal awareness of mental well-being. Paradym addresses these challenges by providing an accessible, evidence-based platform to help users increase their emotional awareness by exploring their emotional patterns, ultimately aiming to enhance emotional well-being, boost self-awareness, and improve overall life satisfaction [32]. The Paradym app introduces users to psychoeducational content through storytelling, covering key aspects of life such as love and relationships, body image, work and success, and identity. In addition, it incorporates elements of lived experience, which involves the user's personal, first-hand experiences and the significance these experiences bring to their current situations [33].

Paradym's unique use of emotional pattern recognition and storytelling aims to offer a novel approach within the field of digital mental health and use this combination of lived experience and evidence-based interventions to boost user engagement and emotional awareness. Finally, in contrast to many of the widely used mental health and well-being apps available, Paradym incorporates validated mental health questionnaires in its interface, such as the Satisfaction with Life Scale (SWLS), which allows users to monitor their mood and well-being as they go through the app's contents. To ensure effective implementation and guide further development, the app's design considered the principles of the Technology Acceptance Model [34]. This model suggests that users' engagement with a new digital tool is influenced by their perceptions of its usefulness and ease of use. In addition, the model recommends monitoring user anxiety, as previous research has identified anxiety as an external factor that can impact technology acceptance [35]. A 2021 acceptability and engagement study of the Paradym app [32] revealed that the majority of users found the app easy to use (64%) and used it daily (79%), and post hoc statistical analyses showed a significant increase in the well-being measures (d=0.73 for SWLS scores and d=0.51 for World Health Organization-5 Well-Being Index [WHO-5] scores) and reduction in symptoms of depression (d=–0.38 for Patient Health Questionnaire-9 [PHQ-9] scores) following app use. Importantly, no adverse effects were reported, providing encouraging preliminary evidence for the Paradym app's effectiveness and usability.

Aim of This Study

The aim of this study was to expand Paradym's evidence base by evaluating the app's effectiveness in addressing symptoms of anxiety and depression and promoting overall well-being and life satisfaction in a real-world context.

Methods

Participants and Procedure

Real-world data were collected from an international pool of Paradym app users, who had downloaded and were using the mobile app from the App Store or Google Play. The participant pool for this study consisted of the total of Paradym users who had completed the app's built-in assessments and questionnaires (Generalized Anxiety Disorder-7 [GAD-7], PHQ-9, WHO-5, and SWLS) at least once. These questionnaires are an optional part of the Paradym app experience, and users can complete them as many times as they wish while using the app to track their progress in their own well-being and life satisfaction. Owing to the fact that assessments are part of the app intervention a specific recruitment procedure was not required to obtain participants' data for this study. Instead, before completing the questionnaires, all Paradym app users were informed by a written statement that their data would be used for research purposes related to the effects of Paradym on anxiety, depression, general well-being, and satisfaction with life. Users provided consent when signing up to the app and informed that surveys were part of the app user experience. Following the principle of data minimization, users did not provide detailed demographic data beyond the country of residence. The data that informed this study from the in-app assessments was collected between March 2023 and June 2024. Participants did not receive any compensation for filling out the questionnaires as the measures were part of the app intervention.

The Checklist for Reporting Results of Internet E-Surveys (CHERRIES) [36] was used to track the details of this study's methodology, this provides further detail on recruitment and administration and is provided in in Multimedia Appendix 1.

XSL•FO
RenderX

Intervention Overview

Paradym is an app designed to promote emotional well-being among the general population, specifically targeting adults aged 18 years and older. Its primary goal is to help users cultivate greater emotional awareness and achieve higher levels of life satisfaction. In its early development, Paradym was funded through crowdfunding, a choice made to avoid the influence of commercial interests during the initial stages of creating the intervention.

Paradym was developed based on evidence-based techniques and was tailored to address themes like success, body, identity, love, and relationships, which were identified by users as important. The platform's development involved clinical psychologists, coaches, and researchers who collaborated on crafting the content. This included selecting and reviewing psychoeducational materials, appropriate evidence-based strategies, and various activities like journaling and note-taking [32].

The foundation of Paradym is an integrative theoretical approach, enriched by therapeutic strategies like acceptance and commitment therapy and schema therapy, which shape the lived experience and storytelling aspects of the platform. User feedback further refined these approaches, leading to the integration of additional theory-driven strategies to generate content that resonated with the needs expressed by users. The core strategies of Paradym encompass these elements, providing a comprehensive toolkit for personal development.

Digital Lessons

Paradym's psychoeducational digital lessons are organized around 5 central themes: awareness, success, love, identity, and body. These areas were selected based on feedback gathered from users during the initial phase of development [32]. The educational content covers several topics, including building emotional awareness, aiding users in recognizing and understanding their emotions, and helping them identify and reinforce their personal values.

Each lesson starts by presenting the educational concept through narratives based on real-life experiences. The use of personal stories is intended to make the material more accessible and relatable, enhance empathy, and encourage users to see themselves in the stories being told. This storytelling approach is known to increase user engagement with both psychological interventions and digital applications. The lessons are available in various formats, including text chapters, audio recordings, and videos, catering to different learning preferences.

Emotional Patterns

At the conclusion of each digital lesson, users are prompted to examine their own emotional patterns in context with the app's material, incorporating elements from schema therapy, in order to enhance emotional insight. To facilitate this, the app presents various "emotional patterns" and encourages users to reflect on which pattern they resonate with most at the time of finishing the lesson. This allows users to track and potentially modify their emotional patterns as they progress through the lessons.

Reflections

Users receive reflections through push notifications, designed to enhance their daily interaction with the app by introducing a new reflection each day. Research has highlighted numerous advantages to such reflection exercises, including aiding individuals in maintaining focus on tasks amidst stressful situations, and reducing tendencies to ruminate, which is often associated with internal struggles and social conflicts. In addition, the use of notifications has been proven to boost user interaction with apps [32], and studies suggest that mobile apps delivering psychological insights are most effective when coupled with proactive user involvement.

Data Collection Tools

Demographics

In order to maintain users' anonymity and confidentiality, only information regarding their geographic location was collected.

Well-Being Measure

The WHO-5 was used to assess well-being in this study. The WHO-5 consists of a brief assessment (under 1 minute) of well-being over a 2-week period [37], in which individuals are asked to indicate how they felt over the past 2 weeks for each of 5 statements using a 6-point Likert scale, ranging from 0="at no time" to 5="all of the time." A higher total score on the WHO-5 indicates a higher level of well-being. The WHO-5 is derived from a 28-item questionnaire based on the Zung scales for depression, distress, and anxiety, as well as from the General Health Questionnaire and the Psychological General Well-Being Scale [38,39]. The WHO-5 has been validated as a measure of depression in both adolescents and older adults, with high measurement invariance [38,39].

Satisfaction With Life Measure

The SWLS [40] is a brief (approximately 1 minute) assessment consisting of 5 items designed to measure the global cognitive assessment of satisfaction with life. The SWLS has been shown to have very high construct validity, with Cronbach α=0.85-0.87 [41] and moderately high reliability (Cronbach α=0.78) [42]. A higher score on the SWLS indicates a higher level of life satisfaction.

Anxiety Measure

The GAD-7 is a 7-item, brief, widely-used clinical measure that assesses the presence and severity of generalized anxiety disorder. The self-report scale asks how often during the last 2 weeks individuals experienced symptoms of generalized anxiety disorder. Total scores range from 0-21, with cutoff scores of 5, 10, and 15 indicating mild, moderate, and severe anxiety, respectively. Increasing scores on the GAD-7 are strongly associated with greater functional impairment in real-world settings. Sample items are rated from 0 (not at all) to 3 (nearly every day). Scores over 10 have good sensitivity (89%) and specificity (82%) for diagnosis of generalized anxiety disorder by interview, and the scale has high internal reliability, as suggested by Cronbach α=0.92 [43]. The scale has been widely used and considered a valid and reliable screening tool in previous research, presenting good reliability, factorial, and

XSL•FO
RenderX

concurrent validity [44]. A higher GAD-7 score indicates more severe anxiety.

Depression Measure

The PHQ-9 is a widely used depression scale that scores each of the 9 Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition criteria as 0 (not at all) to 3 (nearly every day). The PHQ-9 has been validated for use in primary care [45]. It is not a screening tool for depression but can monitor the severity of symptoms and response to treatment. Scores over 10 have good sensitivity (88%) and specificity (88%) for the diagnosis of major depression by interview. It has high internal reliability, and Cronbach α=0.89 [46]. The construct validity of the PHQ-9 is also high for community and clinical samples [47]. A high score obtained from the PHQ-9 is indicative of a severe level of depression.

Data Analysis

Available demographic information (ie, geographic location) and emotional patterns (ie, burnout, sense of thriving, and insecurity) for app users were descriptively analyzed using count and percentages (n, %). Engagement was calculated based on the number of times users visited the app and the length of time the user had been registered on the Paradym app. The results for primary outcomes were largely descriptive in nature, and we evaluated success based on improvements in scores over time. Within-subjects *t* tests were conducted to estimate differences between time 1 (ie, the start of app use) and follow-up. A *P* value of <.001 was considered significant and Cohen *d* estimates were used to measure effect sizes. All statistical analyses were conducted using the SPSS (IBM Corp) software [48].

Ethical Considerations

This study was conducted with a strong commitment to ethical principles, ensuring the protection of participants' rights, safety, dignity, and well-being [49]. Data were gathered remotely from individuals who were not part of a clinical setting, and who freely gave their informed consent, allowing their anonymized and nonidentifiable in-app data to be used for research purposes. The research team made their contact information available, offering support and directing participants to additional well-being resources as needed. Participants did not receive any compensation for filling out the questionnaires as the measures were part of the app intervention.

The study was considered to be user testing or service evaluation, and as such, it was not subject to the usual requirements for formal ethical approval by the National Health Service Health Research Authority, aligning with practices in similar studies [50,51]. The certificate of exemption is provided in Multimedia Appendix 2. Since the evaluation included a nonclinical group, it was exempt from registration in a public trials registry, similar to other studies where formal ethical oversight was deemed unnecessary due to the low-risk nature of the research and its nonclinical focus [52,53]. The detailed Terms and Conditions regulations are provided in Multimedia Appendix 3; through these, users of the app agree to their data being used in research.

Results

Descriptive Data

A total of 3237 unique users of the Paradym app provided data included in this analysis, between March of 2023 and June of 2024. User uniqueness was determined based on the principal investigator and cookie data. The data were anonymized to protect participants' identities, so details on their age, gender, and race were not recorded. However, the users had the option to provide information on their geographic location (Table 1).

Table 1. Baseline demographic information of Paradym app users whose data were used in this real-world evaluation (N=3237).

Country	Participants, n (%)
United States	1237 (38.2)
Philippines	547 (16.9)
Great Britain	210 (6.5)
Canada	204 (6.3)
Australia	139 (4.3)
Other	900 (27.8)

Users had the option to self-report their perceived levels of burnout, sense of thriving, and level of insecurity before starting to use the app. A total of 1627 users provided information on their level of burnout, and almost half (n=677, 41.6%) reported experiencing strong burnout, while 641 (39.4%) experienced moderate levels of burnout and 309 (19%) experienced low levels of burnout. A total of 260 users rated their sense of thriving, and the majority reported either slight (n=140, 53.8%) or moderate (n=104, 40%) thriving, with only 16 (6.2%) reporting a strong sense of thriving. Among the 211 users who rated their level of insecurity, 34.6% (n=73) and 29.4% (n=62) reported strong and moderate insecurity, respectively, with 36.0% reporting slight insecurity (n=76; Table 2).

XSL•FO
RenderX

Table 2. Baseline data on Paradym app users' levels of burnout (n=1627), thriving (n=260), and insecurity (n=211).

Domain	Participants, n (%)
Burnout	
Slight	309 (19)
Moderate	641 (39.4)
Strong	677 (41.6)
Thriving	
Slight	140 (53.8)
Moderate	104 (40)
Strong	16 (6.2)
Insecurity	
Slight	76 (36)
Moderate	62 (29.4)
Strong	73 (34.6)

Efficacy Data

After downloading the app, the first-time users completed questionnaires assessing their levels of anxiety (GAD-7), depression (PHQ-9), life satisfaction (SWLS), and well-being (WHO-5). App usage among the participants in this study varied, but on average, users had been registered for 74 days at the time the data were collected and had visited the Paradym app 10 times, with over 30% of users reporting that they had been using the app somewhat frequently or very often. Of the 3237 app users, just over 20% (n=787) of app users returned to the app more than once to complete mental health and well-being assessments. This was also a significant indicator of those who used the app more frequently (119 times vs 379 times) and had been using the app over a longer period of time (mean 215 days vs mean 36 days). Notably, those who returned to monitor their mental health and well-being indicators at least once showed significant improvements in all 4 mental health and well-being measures. Using available data, within-subjects *t* tests were applied, and the improvement in anxiety (t_{521}=5.288, P<.001), depression (t_{521}=6.667, P<.001), satisfaction with life (t_{521}=3.191, P<.001), and general well-being (t_{521}=4.947, P<.001) scores were all significant at follow-up (Table 3; Figure 1).

Table 3. Baseline (first questionnaire completion) and follow-up (subsequent questionnaire completion) data on Paradym app users' levels of anxiety, depression, satisfaction with life, and general well-being (n=787).

Outcome measure	Baseline, mean (SD)	Follow-up, mean (SD)	Average baseline to follow-up score change	Effect size (Cohen *d*)
GAD-7[a]	14.27 (6.77)	13.07 (6.63)	1.20	0.231
PHQ-9[b]	9.85 (5.55)	8.6 (5.55)	1.26	0.292
SWLS[c]	12.14 (7.42)	12.85 (7.46)	0.71	0.140
WHO-5[d]	9.88 (5.51)	10.85 (5.5)	0.97	0.217

[a]GAD-7: Generalized Anxiety Disorder-7.

[b]PHQ-9: Patient Health Questionnaire-9.

[c]SWLS: Satisfaction With Life Scale.

[d]WHO-5: World Health Organization-5 Well-Being Index.

XSL•FO
RenderX

Figure 1. Baseline and follow-up data on participants' levels of anxiety, depression, satisfaction with life, and general well-being (n=787). GAD-7: Generalized Anxiety Disorder-7; PHQ-9: Patient Health Questionnaire-9; SWLS: Satisfaction with Life Scale; WHO-5: World Health Organization-5 Well-Being Index.

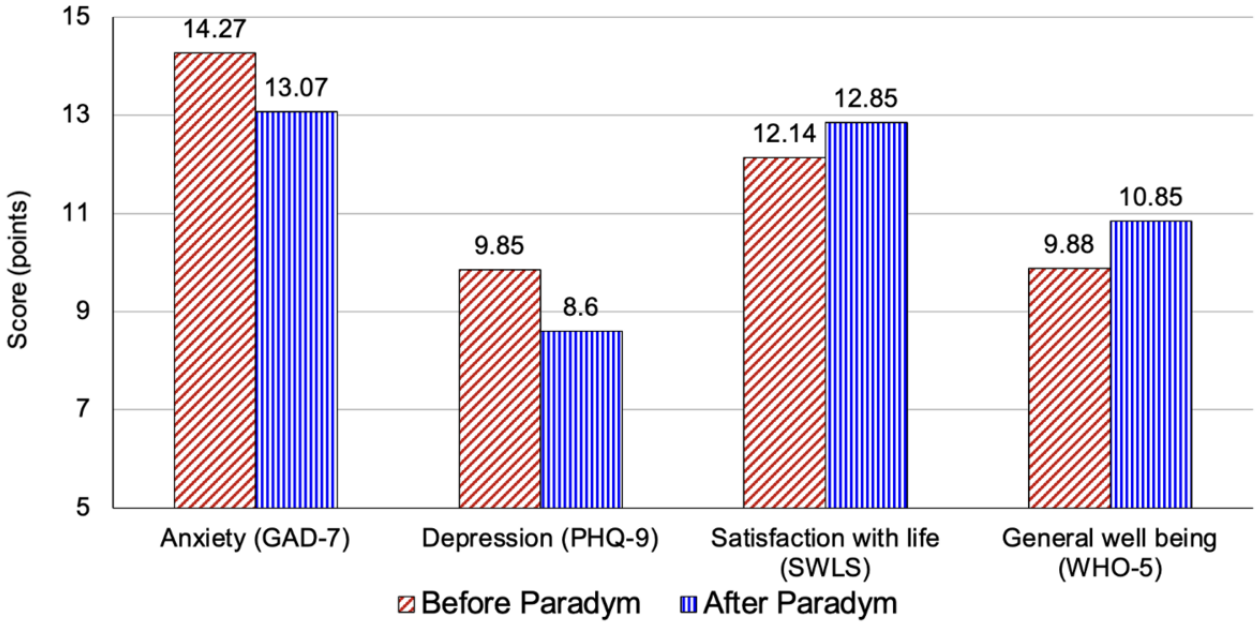

Discussion

Principal Findings

The aim of this study was to investigate the effects of the well-being app Paradym on the users' levels of anxiety, depression, satisfaction with life, and general well-being. The findings indicated that users who accessed the app's contents and completed the built-in mental health and well-being assessments (N=3237) generally reported problems with burnout, insecurity, and a lack of thriving in their lives. In addition, the app's user base scored highly on scales of depression (PHQ-9) and anxiety (GAD-7) but reported low levels of life satisfaction (SWLS) and well-being (WHO-5). A total of the 787 users chose to complete follow-up assessments of these measures, in which significant decreases in average scores of depression and anxiety were observed. In addition, both life satisfaction and well-being scores increased significantly, providing real-world data on Paradym's benefits for users with a relatively large sample. However, as this was not a controlled study, it is not possible to make any causal claims or explicitly attribute the findings to the use of the app. Nonetheless, these preliminary findings are overall positive and indicate that Paradym's digital format allows for wide accessibility and scalability, positioning it as an effective tool for both individual and public mental health strategies.

Despite these positive findings, the number of users who opted to not complete the follow-up mental health and well-being assessments was high (2450/3237), which is commonly observed in studies of mental health and well-being apps [54]. However, this dropout rate only refers to users who did not use the app's built-in questionnaires. Thus, it is possible that many of these users continued using the app and engaging with its contents, but did not feel the need to assess their progress with the app's built-in questionnaires. Alternatively, users who experienced higher levels of distress or more severe depression and anxiety symptoms found it more difficult to engage with the app or respond to the outcome measures, as has been observed in other similar studies [55]. In addition, mental health apps generally have high dropout rates, and some studies have found that once users learn a skill or knowledge from a particular app, they stop using it [56]. Further research would be required to understand the factors that influence users to complete the outcome measures beyond the first measure.

Comparison With Previous Work

This study's findings on Paradym's effectiveness in improving mental health and well-being are in line with those observed in studies of other similar apps and add to the growing literature supporting apps' benefits as mental health tools [57,58].

A significant decrease in symptoms of anxiety and depression has been observed in multiple other studies of popular apps, such as Headspace and Calm, using the same questionnaires as this study [22,59-61]. The same is true for studies examining well-being and life satisfaction [15,62]. In addition, the effect sizes observed for depression and anxiety are consistent with those reported in other studies and meta-analyses of mental health apps [21-23,58,60-63].

However, very few studies have examined the effects that multiple compared to stand-alone assessments or the timing of those assessments could have on the participants, and we did not identify any studies that studied this association using naturalistic data through assessments that were built into the apps' experience.

Strengths and Limitations

A key strength of the present analysis rests on its use of a large sample of users under real-world conditions from a diverse user

XSL•FO
RenderX

base over a long period of time. Naturalistic data were collected, as users were completing assessments as part of the app experience, and not as part of a controlled study environment. Such analyses could prompt app developers to integrate rigorous data collection processes within their apps' interface and core features. In addition, the high levels of burnout, insecurity, anxiety, and depression, and low levels of life satisfaction, thriving, and general well-being reported by users at baseline act as a strong indicator that Paradym reached users who were seeking support for their mental health and well-being.

Since no personal data related to the users' gender, race, or age were collected by the app, in order to ensure users' anonymity and confidentiality and follow the data minimization principle some limitations in our data analysis are acknowledged. Due to the absence of such detailed demographic data**,** subgroup analyses based on such characteristics were not conducted as part of this study. Subgroup analyses of effectiveness could provide important insights into how demographic factors affect the outcomes of interest, as has been observed in other studies [64,65].

Further, this study's single-arm design did not allow for the inclusion of comparators, which could limit the findings' robustness. Change scores were small but significant so this should be interpreted with caution. The change scores and resulting effect sizes were lower than those commonly accepted to indicate a clinically meaningful difference (approximately 4 points on the GAD-7 [66] and 3 points on the PHQ-9 [67]). However, it should be noted that these thresholds for clinically meaningful change are based on studies of clinical populations, including patients diagnosed with anxiety or depression who had mild to severe symptoms [68]; Paradym is not exclusively aimed at diagnosed patients, and thus, it is expected that the aforementioned clinically meaningful thresholds are not directly applicable to the app's users, who may have benefited substantially from smaller improvements in questionnaire scores [69].

Finally, while the number of users who dropped out in regards to the questionnaire filling was captured in the data, the exact patterns of app usage were not measured. For instance, some users may have stopped completing the questionnaires, but they maintained a subscription and continued using the app and engaging with its content daily. Collecting more specific data on user engagement patterns would allow for a clearer view of which features of the Paradym app users found more useful, beyond the questionnaires.

Future Directions

Future studies ought to gain more understanding of the mechanisms underpinning the observed reductions in anxiety and depression symptoms and the increase in life satisfaction and well-being that Paradym users experience. Assessments to capture the specific components of the app's content and structure that have a greater impact on users could be incorporated into future analyses of user data.

In addition, more detailed information on users' experiences with burnout, low levels of thriving, anxiety, and depression would help us better understand the needs of the app's user base, and tailor the app accordingly. For example, questions exploring the specific domains users struggle with (eg, work and personal relationships) could shed light on what kind of content they would benefit the most from (eg, self-assessments, reflection exercises, and guided lessons).

Paradym's benefits for mental health and well-being could be investigated through randomized controlled trials in the future, in order to understand whether improvements in the users' mental health and well-being are attributed to ongoing app usage. Comparison groups using similar apps or face-to-face therapy could also be introduced to enhance the results' validity, robustness, and generalizability.

While not collecting demographic data for this study was a deliberate choice made to preserve the users' anonymity, future studies could consider collecting data related to age, gender, and race, in order to allow for subgroup analyses of Paradym's efficacy based on these characteristics. This will be a significant next step in order to answer questions around "what worked, for whom, and under which circumstances."

Further studies could also explore Paradym's integration within broader systems, such as health care or educational settings, to evaluate the app's effectiveness at scale. For instance, patients could be offered access to Paradym if they are placed on a waiting list for in-person therapy, or if they do not feel comfortable initiating face-to-face therapy. In addition, future research could explore the app's impact across different populations and contexts, including workplace wellness programs or health care partnerships.

Finally, this study's findings indicate that the anxiety, depression, well-being, and life satisfaction data collection tools incorporated within the app are beneficial to the users. As a result, additional tools, including qualitative data collection measures, could be introduced to the app's interface, allowing participants to provide valuable and detailed information that could provide insight into their emotional and mental health journey while using the app, as well as a way for them to track their own progress. Feedback measures could also be added to capture the reasons why some participants decide to complete the existing assessments more or less frequently, and why some opt to not complete them at all. This would help to elucidate potential barriers to user engagement with these data collection tools, and whether these are related to the app's content and structure or to external factors that prevent users from fully engaging with Paradym's user journey.

Conclusions

Overall, this analysis highlighted the value of conducting formative research on mental health and well-being apps. The findings suggest that built-in assessments in a mental health app can lead to the collection of valid real-world data from users and promote organic engagement with the app's features. Among users who completed follow-up assessments in the Paradym app, levels of anxiety and depression decreased significantly, while life satisfaction and general well-being improved significantly. This study provided positive evidence for Paradym's effectiveness in improving mood and well-being and highlighted the value of the app's content, functionality,

XSL•FO
RenderX

and underlying principles. These encouraging results suggest the scope for further research on Paradym and the integration of built-in assessments in mental health apps.

Acknowledgments

The authors thank the participants who took the time to participate in this study. This study was funded by Paradym Ltd. Apart from commissioning a call for evidence in this area, CC and JP had no role in data collection, data analysis, or data interpretation. However, CC and JP were responsible for the oversight of the project and were involved in the study design and editing of the drafts of the manuscript.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

AMM, ME, CC, JP, and SL conceptualized the study and contributed to the study design. All authors contributed to user testing. To reduce the risk of bias in favor of the intervention, SL independently analyzed the data and interpreted the output with input from ME and AMM. All authors contributed to the writing and editing of the manuscript. All authors were involved in reviewing, editing, and final approval of the manuscript before submission.

Conflicts of Interest

CC is the founder and Chief Executive Officer of Paradym Ltd, and JP is the Head of Development lead at Paradym Ltd. AMM, ME, and SL were commissioned as external consultants during the analysis and write-up of this evaluation.

Multimedia Appendix 1
Checklist for Reporting Results of Internet E-Surveys (CHERRIES).
[DOCX File , 19 KB - formative_v9i1e68031_app1.docx]

Multimedia Appendix 2
NHS Health Research Authority exemption certificate.
[PDF File (Adobe PDF File), 175 KB - formative_v9i1e68031_app2.pdf]

Multimedia Appendix 3
Captures of the Paradym app's interface, including the onboarding screen, the privacy policy, and the Terms and Conditions.
[PDF File (Adobe PDF File), 527 KB - formative_v9i1e68031_app3.pdf]

References

1. Agteren J, Iasiello M. Advancing our understanding of mental well-being and mental health: the call to embrace complexity over simplification. Australian Psychologist 2021;55(4):307-316. [doi: 10.1111/ap.12440]
2. Huppert F. Psychological well‐being: evidence regarding its causes and consequences. Applied Psych Health & Well 2009;1(2):137-164. [doi: 10.1111/j.1758-0854.2009.01008.x]
3. Chida Y, Steptoe A. Positive psychological well-being and mortality: a quantitative review of prospective observational studies. Psychosom Med 2008;70(7):741-756. [doi: 10.1097/PSY.0b013e31818105ba] [Medline: 18725425]
4. Kaufman KR, Petkova E, Bhui KS, Schulze TG. A global needs assessment in times of a global crisis: world psychiatry response to the COVID-19 pandemic. BJPsych Open 2020;6(3):e48 [FREE Full text] [doi: 10.1192/bjo.2020.25] [Medline: 32250235]
5. Salari N, Hosseinian-Far A, Jalali R, Vaisi-Raygani A, Rasoulpoor S, Mohammadi M, et al. Prevalence of stress, anxiety, depression among the general population during the COVID-19 pandemic: a systematic review and meta-analysis. Global Health 2020;16(1):57 [FREE Full text] [doi: 10.1186/s12992-020-00589-w] [Medline: 32631403]
6. Bakker D, Kazantzis N, Rickwood D, Rickard N. Mental health smartphone apps: review and evidence-based recommendations for future developments. JMIR Ment Health 2016;3(1):e7 [FREE Full text] [doi: 10.2196/mental.4984] [Medline: 26932350]
7. Okoro YO, Ayo-Farai O, Maduka CP, Okongwu CC, Sodamade OT. The role of technology in enhancing mental health advocacy: a systematic review. Int. j. appl. res. soc. sci 2024;6(1):37-50. [doi: 10.51594/ijarss.v6i1.690]
8. Shojaei F, Osorio Torres J, Shih PC. Exploring the integration of technology in art therapy: insights from interviews with art therapists. Art Therapy 2024:1-7. [doi: 10.1080/07421656.2024.2383826]
9. Zubala A, Kennell N, Hackett S. Art therapy in the digital world: an integrative review of current practice and future directions. Front Psychol 2021;12:595536 [FREE Full text] [doi: 10.3389/fpsyg.2021.600070] [Medline: 33897522]

XSL•FO
RenderX

10. Berardi C, Antonini M, Jordan Z, Wechtler H, Paolucci F, Hinwood M. Barriers and facilitators to the implementation of digital technologies in mental health systems: a qualitative systematic review to inform a policy framework. BMC Health Serv Res 2024;24(1):243 [FREE Full text] [doi: 10.1186/s12913-023-10536-1] [Medline: 38408938]
11. Smith KA, Blease C, Faurholt-Jepsen M, Firth J, Van Daele T, Moreno C, et al. Digital mental health: challenges and next steps. BMJ Ment Health 2023;26(1):e300670 [FREE Full text] [doi: 10.1136/bmjment-2023-300670] [Medline: 37197797]
12. Pieritz S, Khwaja M, Faisal A, Matic A. Personalised recommendations in mental health apps: the impact of autonomy and data sharing. 2021 Presented at: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems; 2021 May 07; Yokohama Japan. [doi: 10.1145/3411764.3445523]
13. Keyes CLM. Mental health as a complete state: how the salutogenic perspective completes the picture. In: Bauer GF, Hammig O, editors. Bridging Occupational, Organizational and Public Health. New York: Springer; 2014.
14. Agnoli S, Mancini G, Andrei F, Trombini E. The relationship between trait emotional intelligence, cognition, and emotional awareness: an interpretative model. Front Psychol 2019;10:1711 [FREE Full text] [doi: 10.3389/fpsyg.2019.01711] [Medline: 31417455]
15. Hendryx MS, Haviland MG, Shaw DG. Dimensions of alexithymia and their relationships to anxiety and depression. J Pers Assess 1991;56(2):227-237. [doi: 10.1207/s15327752jpa5602_4] [Medline: 2056418]
16. Kauer S. School of Psychological Science, The University of Melbourne. Emotional self-awareness and depressive symptoms: an investigation of an early intervention mobile phone self-monitoring program for adolescents. URL: https://minerva-access.unimelb.edu.au/handle/11343/37899 [accessed 2025-12-01]
17. Guendelman S, Medeiros S, Rampes H. Mindfulness and emotion regulation: insights from neurobiological, psychological, and clinical studies. Front Psychol 2017;8:220 [FREE Full text] [doi: 10.3389/fpsyg.2017.00220] [Medline: 28321194]
18. McKay FH, Wright A, Shill J, Stephens H, Uccellini M. Using health and well-being apps for behavior change: a systematic search and rating of apps. JMIR mHealth uHealth 2019;7(7):e11926 [FREE Full text] [doi: 10.2196/11926] [Medline: 31274112]
19. Roffarello AM, Russis LD. The race towards digital wellbeing. 2019 Presented at: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems; 2019 May 9; Glasgow Scotland UK p. 4-9. [doi: 10.1145/3290605.3300616]
20. Eisenstadt M, Liverpool S, Infanti E, Ciuvat RM, Carlsson C. Mobile apps that promote emotion regulation, positive mental health, and well-being in the general population: systematic review and meta-analysis. JMIR Ment Health 2021;8(11):e31170 [FREE Full text] [doi: 10.2196/31170] [Medline: 34747713]
21. Avalos LA, Aghaee S, Kurtovich E, Quesenberry C, Nkemere L, McGinnis MK, et al. A mobile health mindfulness intervention for women with moderate to moderately severe postpartum depressive symptoms: feasibility study. JMIR Ment Health 2020;7(11):e17405 [FREE Full text] [doi: 10.2196/17405] [Medline: 33180028]
22. Comtois KA, Mata-Greve F, Johnson M, Pullmann MD, Mosser B, Arean P. Effectiveness of mental health apps for distress during COVID-19 in US unemployed and essential workers: remote pragmatic randomized clinical trial. JMIR mHealth uHealth 2022;10(11):e41689 [FREE Full text] [doi: 10.2196/41689] [Medline: 36191176]
23. Kubo A, Aghaee S, Kurtovich EM, Nkemere L, Quesenberry CP, McGinnis MK, et al. mHealth mindfulness intervention for women with moderate-to-moderately-severe antenatal depressive symptoms: a pilot study within an integrated health care system. Mindfulness (N Y) 2021;12(6):1387-1397 [FREE Full text] [doi: 10.1007/s12671-021-01606-8] [Medline: 33723491]
24. Bakker D, Rickard N. Engagement with a cognitive behavioural therapy mobile phone app predicts changes in mental health and wellbeing: MoodMission. Aust Psychol 2019;54(4):245-260. [doi: 10.1111/ap.12383]
25. Birney AJ, Gunn R, Russell JK, Ary DV. MoodHacker mobile web app with email for adults to self-manage mild-to-moderate depression: randomized controlled trial. JMIR mHealth uHealth 2016;4(1):e8 [FREE Full text] [doi: 10.2196/mhealth.4231] [Medline: 26813737]
26. Throuvala MA, Griffiths MD, Rennoldson M, Kuss DJ. Mind over matter: testing the efficacy of an online randomized controlled trial to reduce distraction from smartphone use. Int J Environ Res Public Health 2020;17(13):4842 [FREE Full text] [doi: 10.3390/ijerph17134842] [Medline: 32635650]
27. Bakker D, Rickard N. Engagement in mobile phone app for self-monitoring of emotional wellbeing predicts changes in mental health: MoodPrism. J Affect Disord 2018;227:432-442. [doi: 10.1016/j.jad.2017.11.016] [Medline: 29154165]
28. Daugherty DA, Runyan JD, Steenbergh TA, Fratzke BJ, Fry BN, Westra E. Smartphone delivery of a hope intervention: another way to flourish. PLoS One 2018;13(6):e0197930 [FREE Full text] [doi: 10.1371/journal.pone.0197930] [Medline: 29856805]
29. Zhu D, Al Mahmud A, Liu W. Digital storytelling intervention for enhancing the social participation of people with mild cognitive impairment: co-design and usability study. JMIR Aging 2024;7:e54138 [FREE Full text] [doi: 10.2196/54138] [Medline: 38231541]
30. Davidson T, Moreland A, Bunnell BE, Winkelmann J, Hamblen JL, Ruggiero KJ. Reducing stigma in mental health through digital storytelling. In Deconstructing stigma in mental health. USA: IGI Global; 2018.
31. Mehta J, Long E, Bynagari V, Shojaei F, Shojaei F, O'Brien ARW, et al. Creating demand for unmet needs: agile storytelling. Front Health Serv 2024;4:1376695 [FREE Full text] [doi: 10.3389/frhs.2024.1376695] [Medline: 39544454]

XSL•FO
RenderX

32. Eisenstadt A, Liverpool S, Metaxa AM, Ciuvat RM, Carlsson C. Acceptability, engagement, and exploratory outcomes of an emotional well-being app: mixed methods preliminary evaluation and descriptive analysis. JMIR Form Res 2021;5(11):e31064 [FREE Full text] [doi: 10.2196/31064] [Medline: 34569466]
33. Lindseth A, Norberg A. A phenomenological hermeneutical method for researching lived experience. Scand J Caring Sci 2004;18(2):145-153. [doi: 10.1111/j.1471-6712.2004.00258.x] [Medline: 15147477]
34. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly 1989;13(3):319. [doi: 10.2307/249008]
35. Dönmez-Turan A, Kır M. User anxiety as an external variable of technology acceptance model: a meta-analytic study. Procedia Computer Science 2019;158:715-724. [doi: 10.1016/j.procs.2019.09.107]
36. Eysenbach G. Improving the quality of web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004;6(3):e34 [FREE Full text] [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
37. Johansen KS. The use of well-being measures in primary health care/The DepCare Project. World Health Organization, Regional Office for Europe. URL: https://www.euro.who.int/__data/assets/pdf_file/0016/130750/E60246.pdf [accessed 2025-01-13]
38. Krieger T, Zimmermann J, Huffziger S, Ubl B, Diener C, Kuehner C, et al. Measuring depression with a well-being index: further evidence for the validity of the WHO well-being index (WHO-5) as a measure of the severity of depression. J Affect Disord 2014;156:240-244. [doi: 10.1016/j.jad.2013.12.015] [Medline: 24412323]
39. Topp CW, Østergaard SD, Søndergaard S, Bech P. The WHO-5 well-being index: a systematic review of the literature. Psychother Psychosom 2015;84(3):167-176 [FREE Full text] [doi: 10.1159/000376585] [Medline: 25831962]
40. Diener E, Emmons RA, Larsen RJ, Griffin S. The satisfaction with life scale. J Pers Assess 1985;49(1):71-75. [doi: 10.1207/s15327752jpa4901_13] [Medline: 16367493]
41. The satisfaction with life scale examining construct validity. Heerlen: Statistics Netherlands; 2012.
42. Vassar M. A note on the score reliability for the satisfaction with life scale: an RG study. Soc Indic Res 2007;86(1):47-57. [doi: 10.1007/s11205-007-9113-7]
43. Spitzer RL, Kroenke K, Williams JB, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
44. Sousa TV, Viveiros V, Chai MV, Vicente FL, Jesus G, Carnot MJ, et al. Reliability and validity of the Portuguese version of the Generalized Anxiety Disorder (GAD-7) scale. Health Qual Life Outcomes 2015;13:50 [FREE Full text] [doi: 10.1186/s12955-015-0244-2] [Medline: 25908249]
45. Mazzocco PJ, Green MC, Sasota JA, Jones NW. This story is not for everyone: transportability and narrative persuasion. Social Psychological and Personality Science 2010;1(4):361-368. [doi: 10.1177/1948550610376600]
46. Kroenke K, Spitzer RL, Williams JBW. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001;16(9):606-613. [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
47. Kocalevent R, Hinz A, Brähler E. Standardization of the depression screener patient health questionnaire (PHQ-9) in the general population. Gen Hosp Psychiatry 2013;35(5):551-555. [doi: 10.1016/j.genhosppsych.2013.04.006] [Medline: 23664569]
48. IBM SPSS Statistics for Windows, version 27. URL: https://www.ibm.com/support/pages/downloading-ibm-spss-statistics-27 [accessed 2021-05-21]
49. Koocher G, Keith-Spiegel P. Ethics in Psychology and the Mental Health Professions: Standards and Cases (3rd ed). New York: Oxford University Press; 2008.
50. Ly KH, Dahl J, Carlbring P, Andersson G. Development and initial evaluation of a smartphone application based on acceptance and commitment therapy. Springerplus 2012;1:11 [FREE Full text] [doi: 10.1186/2193-1801-1-11] [Medline: 23961342]
51. Morris ME, Kathawala Q, Leen TK, Gorenstein EE, Guilak F, Labhard M, et al. Mobile therapy: case study evaluations of a cell phone application for emotional self-awareness. J Med Internet Res 2010;12(2):e10 [FREE Full text] [doi: 10.2196/jmir.1371] [Medline: 20439251]
52. Ahtinen A, Mattila E, Välkkynen P, Kaipainen K, Vanhala T, Ermes M, et al. Mobile mental wellness training for stress management: feasibility and design implications based on a one-month field study. JMIR mHealth uHealth 2013;1(2):e11 [FREE Full text] [doi: 10.2196/mhealth.2596] [Medline: 25100683]
53. Ly KH, Ly A, Andersson G. A fully automated conversational agent for promoting mental well-being: a pilot RCT using mixed methods. Internet Interv 2017;10:39-46 [FREE Full text] [doi: 10.1016/j.invent.2017.10.002] [Medline: 30135751]
54. Schwartz K, Ganster FM, Tran US. Mindfulness-based mobile apps and their impact on well-being in nonclinical populations: systematic review of randomized controlled trials. J Med Internet Res 2023;25:e44638 [FREE Full text] [doi: 10.2196/44638] [Medline: 37540550]
55. Lui JH, Marcus DK, Barry CT. Evidence-based apps? A review of mental health mobile applications in a psychotherapy context. Professional Psychology: Research and Practice 2017;48(3):199-210. [doi: 10.1037/pro0000122]
56. Yardley L, Spring BJ, Riper H, Morrison LG, Crane DH, Curtis K, et al. Understanding and promoting effective engagement with digital behavior change interventions. Am J Prev Med 2016;51(5):833-842. [doi: 10.1016/j.amepre.2016.06.015] [Medline: 27745683]

XSL•FO
RenderX

57. Omylinska-Thurston J, Aithal S, Liverpool S, Clark R, Moula Z, Wood J, et al. Digital psychotherapies for adults experiencing depressive symptoms: systematic review and meta-analysis. JMIR Ment Health 2024;11:e55500 [FREE Full text] [doi: 10.2196/55500] [Medline: 39348177]
58. Linardon J, Cuijpers P, Carlbring P, Messer M, Fuller-Tyszkiewicz M. The efficacy of app-supported smartphone interventions for mental health problems: a meta-analysis of randomized controlled trials. World Psychiatry 2019;18(3):325-336 [FREE Full text] [doi: 10.1002/wps.20673] [Medline: 31496095]
59. Horwitz AG, Mills ED, Sen S, Bohnert ASB. Comparative effectiveness of three digital interventions for adults seeking psychiatric services: a randomized clinical trial. JAMA Netw Open 2024;7(7):e2422115. [doi: 10.1001/jamanetworkopen.2024.22115] [Medline: 39023893]
60. Strauss C, Dunkeld C, Cavanagh K. Is clinician-supported use of a mindfulness smartphone app a feasible treatment for depression? A mixed-methods feasibility study. Internet Interv 2021;25:100413 [FREE Full text] [doi: 10.1016/j.invent.2021.100413] [Medline: 34401372]
61. Huberty J, Eckert R, Larkey L, Kurka J, Rodríguez De Jesús SA, Yoo W, et al. Smartphone-based meditation for myeloproliferative neoplasm patients: feasibility study to inform future trials. JMIR Form Res 2019;3(2):e12662 [FREE Full text] [doi: 10.2196/12662] [Medline: 31033443]
62. Mahlo L, Windsor TD. Feasibility, acceptability, and preliminary efficacy of an app-based mindfulness-meditation program among older adults. Gerontologist 2021;61(5):775-786. [doi: 10.1093/geront/gnaa093] [Medline: 32663286]
63. Fish MT, Saul AD. The gamification of meditation: a randomized-controlled study of a prescribed mobile mindfulness meditation application in reducing college students' depression. Simulation & Gaming 2019;50(4):419-435. [doi: 10.1177/1046878119851821]
64. Lipschitz JM, Connolly SL, Miller CJ, Hogan TP, Simon SR, Burdick KE. Patient interest in mental health mobile app interventions: demographic and symptom-level differences. J Affect Disord 2020;263:216-220. [doi: 10.1016/j.jad.2019.11.083] [Medline: 31818779]
65. Petros NG, Hadlaczky G, Carletto S, Martinez SG, Ostacoli L, Ottaviano M, et al. Sociodemographic characteristics associated with an eHealth system designed to reduce depressive symptoms among patients with breast or prostate cancer: prospective study. JMIR Form Res 2022;6(6):e33734 [FREE Full text] [doi: 10.2196/33734] [Medline: 35675116]
66. Toussaint A, Hüsing P, Gumz A, Wingenfeld K, Härter M, Schramm E, et al. Sensitivity to change and minimal clinically important difference of the 7-item Generalized Anxiety Disorder Questionnaire (GAD-7). J Affect Disord 2020;265:395-401. [doi: 10.1016/j.jad.2020.01.032] [Medline: 32090765]
67. Lynch CP, Cha EDK, Jenkins NW, Parrish JM, Mohan S, Jadczak CN, et al. The minimum clinically important difference for Patient Health Questionnaire-9 in minimally invasive transforaminal interbody fusion. Spine 2021;46(9):603-609. [doi: 10.1097/BRS.0000000000003853] [Medline: 33290370]
68. Kounali D, Button KS, Lewis G, Gilbody S, Kessler D, Araya R, et al. How much change is enough? Evidence from a longitudinal study on depression in UK primary care. Psychol Med 2020;52(10):1875-1882. [doi: 10.1017/s0033291720003700]
69. Doi S, Ito M, Takebayashi Y, Muramatsu K, Horikoshi M. Factorial validity and invariance of the 7-item Generalized Anxiety Disorder scale (GAD-7) among populations with and without self-reported psychiatric diagnostic status. Front Psychol 2018;9:1741 [FREE Full text] [doi: 10.3389/fpsyg.2018.01741] [Medline: 30283386]

Abbreviations

GAD-7: Generalized Anxiety Disorder-7
PHQ-9: Patient Health Questionnaire-9
SWLS: Satisfaction With Life Scale
WHO-5: World Health Organization-5 Well-Being Index

Edited by A Mavragani; submitted 26.10.24; peer-reviewed by F Shojaei, TF Ojo; comments to author 11.11.24; revised version received 02.12.24; accepted 18.12.24; published 23.01.25.

Please cite as:
Metaxa AM, Liverpool S, Eisenstadt M, Pollard J, Carlsson C
Improving Mental Health and Well-Being Through the Paradym App: Quantitative Study of Real-World Data
JMIR Form Res 2025;9:e68031
URL: https://formative.jmir.org/2025/1/e68031
doi:10.2196/68031
PMID:

XSL•FO
RenderX

©Athina Marina Metaxa, Shaun Liverpool, Mia Eisenstadt, John Pollard, Courtney Carlsson. Originally published in JMIR Formative Research (https://formative.jmir.org), 23.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Digital Mindfulness Training for Burnout Reduction in Physicians: Clinician-Driven Approach

Lia Antico[1], PhD; Judson Brewer[1], PhD, MD

Brown University, Department of Behavioral and Social Sciences, Providence, RI, United States

Corresponding Author:
Lia Antico, PhD
Brown University
Department of Behavioral and Social Sciences
121 S Main St
Providence, RI, 02903
United States
Phone: 1 401 328 0763
Email: lia_antico@brown.edu

Abstract

Background: Physician burnout is widespread in health care systems, with harmful consequences on physicians, patients, and health care organizations. Mindfulness training (MT) has proven effective in reducing burnout; however, its time-consuming requirements often pose challenges for physicians who are already struggling with their busy schedules.

Objective: This study aimed to design a short and pragmatic digital MT program with input from clinicians specifically to address burnout and to test its efficacy in physicians.

Methods: Two separate nonrandomized pilot studies were conducted. In the first study, 27 physicians received the digital MT in a podcast format, while in the second study, 29 physicians and nurse practitioners accessed the same training through a free app-based platform. The main outcome measure was cynicism, one dimension of burnout. The secondary outcome measures were emotional exhaustion (the second dimension of burnout), anxiety, depression, intolerance of uncertainty, empathy (personal distress, perspective taking, and empathic concern subscales), self-compassion, and mindfulness (nonreactivity and nonjudgment subscales). In the second study, worry, sleep disturbances, and difficulties in emotion regulation were also measured. Changes in outcomes were assessed using self-report questionnaires administered before and after the treatment and 1 month later as follow-up.

Results: Both studies showed that MT decreased cynicism (posttreatment: 33% reduction; $P \leq .04$; $r \geq 0.41$ and follow-up: 33% reduction; $P \leq .04$; $r \geq 0.45$), while improvements in emotional exhaustion were observed solely in the first study (25% reduction, $P=.02$, $r=.50$ at posttreatment; 25% reduction, $P=.008$, $r=.62$ at follow-up). There were also significant reductions in anxiety ($P \leq .01$, $r \geq 0.49$ at posttreatment; $P \leq .01$, $r \geq 0.54$ at follow-up), intolerance of uncertainty ($P \leq .03$, $r \geq .57$ at posttreatment; $P<.001$, $r \geq 0.66$ at follow-up), and personal distress ($P=.03$, $r=0.43$ at posttreatment; $P=.03$, $r=0.46$ at follow-up), while increases in self-compassion ($P \leq .02$, $r \geq 0.50$ at posttreatment; $P \leq .006$, $r \geq 0.59$ at follow-up) and mindfulness (nonreactivity: $P \leq .001$, $r \geq 0.69$ at posttreatment; $P \leq .004$, $r \geq 0.58$ at follow-up; nonjudgment: $P \leq .009$, $r \geq 0.50$ at posttreatment; $P \leq .03$, $r \geq 0.60$ at follow-up). In addition, the second study reported significant decreases in worry ($P=.04$, $r=0.40$ at posttreatment; $P=.006$, $r=0.58$ at follow-up), sleep disturbances ($P=.04$, $r=0.42$ at posttreatment; $P=.01$, $r=0.53$ at follow-up), and difficulties in emotion regulation ($P=.005$, $r=0.54$ at posttreatment; $P<.001$, $r=0.70$ at follow-up). However, no changes were observed over time for depression or perspective taking and empathic concern. Finally, both studies revealed significant positive correlations between burnout and anxiety (cynicism: $r \geq 0.38$; $P \leq .04$; emotional exhaustion: $r \geq 0.58$; $P \leq .001$).

Conclusions: To our knowledge, this research is the first where clinicians were involved in designing an intervention targeting burnout. These findings suggest that this digital MT serves as a viable and effective tool for alleviating burnout and anxiety among physicians.

Trial Registration: ClinicalTrials.gov NCT06145425; https://clinicaltrials.gov/study/NCT06145425

(JMIR Form Res 2025;9:e63197) doi:10.2196/63197

XSL•FO
RenderX

KEYWORDS

burnout; anxiety; empathy fatigue; physician; mindfulness; digital therapeutics; app; smartphone; podcast; compassion; health care provider; training; physician burnout; cynicism; efficacy; treatment; meditation; chronic; workplace stress; digital health; mHealth; mobile phone

Introduction

Background

Burnout is an occupational phenomenon that is highly prevalent among physicians globally. While the rates have increased because of the COVID-19 pandemic, this was only the acceleration of an existing trend over the last decade [1-4]. In a 2021 survey of 2500 US physicians, 62.8% reported at least 1 symptom of burnout, compared with 38.2% in 2020, 43.9% in 2017, 54.4% in 2014, and 45.5% in 2011 [4].

Defined by the World Health Organization as "a syndrome resulting from chronic workplace stress that has not been managed successfully," burnout is characterized by 3 dimensions: emotional exhaustion, cynicism, and reduced personal accomplishment [5]. *Emotional exhaustion* includes feeling overwhelmed by work demands and experiencing depletion of emotional and physical resources. *Cynicism* represents the interpersonal aspect of burnout and refers to callous and detached attitudes toward one's job and people. *Reduced personal accomplishment* refers to poor professional self-esteem and efficacy [6].

The consequences of burnout on physicians, patients, and health care organizations are significant and far reaching. A large number of studies showed that physicians reporting high rates of burnout are also those who report high levels of anxiety [7], sleep disturbances [8-10], depression, and substance abuse [11,12]. Suicide rates among physicians are twice as high as those in the general public [13]. The emotional toll is so high that in a 2020 survey, 31% of physicians planned to reduce their work hours, and 24% of physicians intended to leave their job within 2 years [14]. An estimation of the direct costs of physician turnover and reduced clinical work is approximately US $4.6 billion each year in the United States [15]. Meta-analyses found that physician burnout is also associated with increased risk of patient safety incidents and malpractice claims, poorer quality due to low professionalism, and diminished and ineffective communication between physicians and patients [16,17]. Indirect costs from medical errors, reduced patient satisfaction, damage to the organization's reputation, and decreased patient loyalty raise the burnout bill to approximately US $7.6 million annually per employed physician [15].

Associated Factors of Physician Burnout

Physician burnout is complex and multifactorial. Burnout among physicians arises from both institutional and individual factors that create an imbalance between job demands and available resources. For instance, high workload and increased administrative tasks (eg, the electronic medical record systems) lead to overworking and deprioritizing self-care. A study involving 7288 US physicians found that for every additional hour worked per week beyond 51.8 hours, burnout symptoms rose by approximately 2% [18]. Similarly, research on 1490 US oncologists revealed that each additional hour spent working at home led to a 2% rise in burnout [2]. Another major contributor to burnout is the loss of flexibility, autonomy, and control that occurs when health care organizations and leadership fail to support individual work goals. This leaves physicians excluded from decisions about patient visit durations and treatment approaches [19].

It may not be a coincidence that the term "burnout" was first used in 1974 to describe the exhaustion observed in people in "healing professions" [20]. Indeed, physicians are constantly exposed to highly stressful and demanding situations that are intrinsic to the clinical practice. Many aspects of medicine are uncertain and ambiguous. For example, diagnoses are often unclear, the natural course of illness is unpredictable, responses of patients to treatments vary, and the clinical information is complex. A growing body of literature indicates that physicians' reactions to uncertainty and their ability to tolerate it can predict stress [21]. Cross-sectional studies reported negative correlations between intolerance of uncertainty and well-being of physicians [22,23]. In particular, physicians with a high intolerance for uncertainty are more likely to experience higher rates of burnout and anxiety, make more referrals, and report lower job satisfaction [24]. However, to date, no studies have directly tested the effects of interventions aimed at reducing physician burnout on intolerance of uncertainty.

While alleviating patients' distress is among the main reasons why many physicians enter the healing professions, caring for the sick is not without consequences. Repeated exposure to illness, death, or social inequities may cause empathic distress that leads to empathy fatigue (sometimes called inappropriately compassion fatigue [25]) and burnout.

Empathy is defined as the capacity to understand and share others' emotions, feelings, and mental states by imagining or taking another person's perspective [25]. Clinical empathy is an essential element of quality care, associated with improved patient satisfaction; increased adherence to treatment; better surgical recovery; shorter hospital stays [26-30]; and increased physician health, well-being, and professional satisfaction [31,32]. However, overempathizing can lead to being overwhelmed by the patient's distress, which might result in withdrawal and contribute to empathy fatigue. Conversely, maintaining a professional perspective with compassionate concern allows physicians to engage with the patient's distress without taking it personally, fostering stronger connections while reducing the risk of burnout.

Interventions for Physician Burnout

Combating burnout is a shared responsibility between the institution and the individual. Effective initiatives have been developed and implemented at both levels [33]. Organization-directed interventions target workload, foster communication between members of the health care system,

XSL•FO
RenderX

and cultivate a sense of team cohesion and job control [33,34]. Individual-focused approaches aim to decrease perceived stress, increase resilience to stressful work environments, and enhance work engagement [33-35]. In particular, this paper focuses on the tangible support that could be provided to physicians at the individual level.

Mindfulness Training

Several studies have reported that mindfulness training (MT) is one of the most studied individual-focused interventions for reducing stress and enhancing well-being among physicians [33,36-40]. Mindfulness refers to being fully aware and present in the current moment, on purpose and nonjudgmentally, of thoughts, feelings, and sensations [41]. By fostering self-awareness and attitudes of openness and "nonjudgment," MT facilitates acceptance of the challenging components of experiences and reduces emotional reactivity [42], both of which are crucial in health care. Most studies have observed that MT reduces rumination, worry, and emotional reactivity among health care providers [40,43]. For instance, nonreactivity to inner experiences has been inversely associated with perceived stress in health care providers [44].

In addition, a meta-analysis found that MT also increases self-compassion in health care providers, with moderate to high effect sizes [45]. Self-compassion involves being touched by one's own distress and being motivated to alleviate it with kindness and understanding [46,47]. It has been theorized that self-compassion is pertinent to physicians' well-being at work and care for patients [48], as it was inversely associated with burnout and empathy fatigue among health care professionals [49-52]. For example, it has been found that physicians who were more self-compassionate felt less exhausted due to work demands and experienced greater work engagement and satisfaction with their professional life [53].

These findings suggest that MT may be promising as an intervention for physician burnout. Different formats have been used in clinical settings [54]; some of them are intensive, including weekly groups sessions of 2.5 hours, home practices of 45 minutes for 8 weeks, and a 7-hour retreat [39]. Therefore, shorter and more practical formats are needed to avoid inadvertently contributing to overload. For example, in a study, 44% of health care professionals randomized to an MT dropped the program because of "lack of time" [55]. Even when health care providers are willing to attend MT, work conflicts can often arise and affect attendance and home practice adherence [35]. In a qualitative study, physicians confirmed the role of mindfulness interventions in the development of self-awareness and highlighted the importance of adapting the course timing and materials to meet their schedule, thereby improving participation and adherence [56].

Furthermore, engaging physicians in the design and implementation of interventions might increase their sense of control and engagement ("user-centered design") [57] and might make the interventions more feasible and tailored to their specific needs. However, to our knowledge, no studies have yet addressed this issue. This study is one of the few that involve clinicians in designing an intervention to reduce physician burnout. Successively, we conducted 2 independent single-arm studies and assessed the impact of digital MT training on physician burnout. In study 1, the intervention was delivered in a podcast format, while in study 2, it was delivered via a free app-based platform.

Study 1

Overview

To develop a short and tailored intervention, we used a user-centered design approach by involving clinicians and focusing on their needs in each phase of the development process. After iterative content development, we delivered the digital intervention in the format of an audio podcast, with the reasoning that content could be consumed during one's commute to and from work and, thus, not add more to the physician's workload. The primary goal of this single-arm study was to examine whether the intervention reduces cynicism (ie, one dimension of burnout) in physicians. The secondary goal was to assess its effects on other variables, such as emotional exhaustion (ie, another dimension of burnout), anxiety, depression, mindfulness (ie, the aspects of nonreactivity and nonjudging of inner experiences), intolerance of uncertainty, empathy (ie, the dimensions of empathic concern, perspective taking, and personal distress), and self-compassion. Finally, we explored the relationship between the 2 dimensions of burnout (ie, cynicism and emotional exhaustion) and the other secondary outcomes.

We evaluated the following hypotheses:

- Hypothesis 1: there will be a main effect of time on cynicism. Physicians will report lower scores for cynicism at postintervention assessments compared to baseline.
- Hypothesis 2: there will be a main effect of time on the other variables. Physicians will report lower scores for emotional exhaustion, anxiety, depression, intolerance of uncertainty, and personal distress and higher scores for nonreactivity, nonjudging of inner experiences, empathic concern, perspective taking, and self-compassion at postintervention assessments compared to baseline.
- Hypothesis 3: cynicism and emotional exhaustion will positively correlate with anxiety.

Methods

Ethical Considerations

The study was approved by the Brown University Institutional Review Board (protocol 2022003296) and was conducted in accordance with the Declaration of Helsinki for experiments involving human subjects. Eligible individuals were directed to the informed consent form on Qualtrics (Qualtrics International, Inc). After enrolling by selecting "yes" (a waiver of documentation of informed consent was obtained), they were redirected to complete the web-based survey. Participants were compensated with a US $25 Amazon gift card for the completion of each survey (US $75 for 3 surveys). All collected data were anonymized, with each participant being assigned a unique alphanumeric code. For statistical analyses, the final database contained only these alphanumeric codes to ensure privacy.

XSL•FO
RenderX

Participants

Physicians were recruited using university hospital mailing lists and flyer advertisements targeting physicians who were interested in testing whether a short training program was effective at reducing burnout. We used the following text in the recruitment message: "Would you like to help us test a short mindfulness-based training to reduce burnout? If you are a physician and are interested in helping us learn about burnout, please contact us."

Inclusion criteria were as follows: direct patient interaction; currently employed as a physician; fluency in English; endorsed willingness to listen to a mindfulness audio course for 15 minutes per day for 7 days; and endorsed willingness to complete 3 web-based surveys, before and after the completion of the training and 1 month later as follow-up. The only exclusion criteria were changing the dose of psychotropic medication in the past 6 weeks and prior exposure to this specific podcast or its content.

Intervention

The mindfulness-based training teaches physicians how to acknowledge and support patients' suffering without becoming overwhelmed by it. It also teaches them to identify and work with empathy fatigue habits loops as well as anxious and worry thoughts that contribute to burnout (eg, not being able to stop thinking about a problem, worrying about patients, and taking work home). Specifically, this program trains awareness and curiosity by targeting fundamental learning processes, such as reinforcement learning, to alter entrenched habits [58]. Previous studies have demonstrated that these changes can lead to significant, clinically relevant effects [58-61].

This experiential education is delivered via an audio podcast, which consists of 7 modules of brief didactic and experience-based MT, each approximately 15 minutes long, and a summary module at the end. Each module includes real-world vignettes and practical mindfulness exercises designed to help identify habit loops in clinical practice and work-related tasks and to provide strategies for breaking these unhelpful loops (Textbox 1). We also provided links to guided meditations, including noting, breathing, grounding practices, as well as curiosity and loving-kindness meditations. Although participants were encouraged to practice these guided meditations, they were optional and not mandatory for participation in this study.

Textbox 1. Overview of the audio course: burnout to resilience themes and content.

Module 1: goals, empathy fatigue, and unhelpful habits

- Set goals and introduce how important it is to be empathetic with patients in their recovery and its limits. Help physicians reflect on the different coping strategies used to protect themselves from patients' distress that become habits and contribute to exhaustion. It discusses specific strategies for stepping out of unhelpful habit loops.

Module 2: empathy versus compassion

- Addresses specifically the difference between empathy and compassion in patterns of activations in the body and mind, and it introduces compassion as a new way to feel patients' distress without being emotionally exhausted by it.

Module 3: intrusive thoughts

- Explains that intrusive thoughts are distracting and can cause great distress. This module also provides strategies for handling them and coming back to the present moment.

Module 4: self-judgment and kindness

- Describes how the mind can be caught up in the loop of self-judgments that contribute to exhaustion by getting in the way of doing one's own tasks and interacting with patients. This module also focuses on how kindness can be a detox of self-judgment.

Module 5: stress and anxiety

- Discusses why the brain does not like uncertainty and how this causes stress and anxiety. This module also provides strategies to change habit patterns and reduce stress and anxiety.

Module 6: emotional contagion and curiosity

- Introduces emotional contagion and its impact on the interactions with patients and colleagues. This module also explains how curiosity can be used as a vaccine against emotional contagion.

Module 7: bias in the diagnosis and treatment of pain

- Describes the complexity of pain, including factors and biases that influence the diagnosis and treatment of pain in patients, and provides strategies for recognizing and dealing with those biases.

The content of this intervention was developed using a user-centered design approach, an iterative design process in which designers focus on the users and their needs in each phase of the design process. On the basis of the extant literature and interviews of the physicians, we identified common individual empathy fatigue "habits" that contribute to burnout [62-64]. Scripts were written by a physician with >20 years of clinical and mindfulness practice and >10 years of experience in

XSL•FO
RenderX

developing MT programs [59-61]. We developed a "minimum viable product" consisting of script-based content that users read, tried out, and commented. We refined the audio course and tested its suitability for a wide range of physicians and other health care providers by collecting feedback from 40 clinicians in 2 rounds. In round 1, a total of 10 physicians from different subspecialties were emailed a script, instructions, and a link to a feedback form. The feedback form included open-ended questions about what participants learned, what was helpful in their clinical practice and personal life, and any confusion or difficulties with understanding and completing home practices and slider questions (ranging from 0 to 100) about how well the module met their needs and those of their colleagues, the usefulness of the real-life stories shared, and the clarity of the home practice instructions. Once each module was completed (as indicated by a completed feedback form), the next module was emailed to the pilot tester. Feedback was collated, and common themes were extracted for training refinement. Vignettes from this round of pilot testers were also incorporated for story-based learning [65]. In round 2, a total of 30 physicians and other clinicians were recruited to repeat this process to confirm content fit and gather additional real-world vignettes for potential incorporation.

Intervention Orientation and Engagement

Participants were instructed to complete the modules at their own pace, with a maximum of 1 module per day. The project director reached out with check-in messages every 15 days from the start of the training to address technical difficulties and encourage continued engagement. During these check-ins, participants were asked about their experience with the podcast since the last contact. If participants expressed difficulties, efforts were made to resolve the problem.

Measures

Primary and secondary outcomes were assessed at 3 time points: baseline (before training), postintervention (after training completion), and follow-up (1 month after training completion).

Primary Outcome

Cynicism, a key dimension of burnout, was measured using a single item from the Maslach Burnout Inventory. This approach is supported by research from West et al [66,67], which demonstrated the validity of using a single item in relation to the full 22-item questionnaire. Participants rated how often they felt "more callous toward people since starting this job" on a 7-point scale (0="never" and 6="every day"). Higher scores indicate greater cynicism [68].

Secondary Outcomes

Emotional exhaustion was measured using a single Maslach Burnout Inventory item. Participants rated how often they felt "burned out from their work" on a 7-point scale, with higher scores indicating greater exhaustion [68].

Anxiety was measured using the *Generalized Anxiety Disorder-7* on a 4-point scale (0="not at all" and 3="nearly every day"). Scores range from 0 to 21, with higher scores indicating more severe anxiety. Total scores of 5 to 9, 10 to 14, and ≥15 correspond to cutoff points of mild, moderate, and severe anxiety, respectively [69].

Depression was measured using the *Patient Health Questionnaire-2* assessing depressive symptoms over the past 2 weeks. Participants rated each item on a 4-point scale (0="not at all" and 3="nearly every day"), with higher scores indicating more severe symptoms [70].

Intolerance *of uncertainty* was assessed by the *Intolerance of Uncertainty scale-12* measuring intolerance of uncertainty with 12 items rated on a 5-point scale (1="not at all" to 5="entirely"). Total scores range from 12 to 60, with higher scores indicating greater intolerance [71].

Self-compassion was measured by the *Self-Compassion Scale-Short Form* assessing self-compassion using 12 items rated on a 5-point scale (1="almost never" to 5="almost always"). Total scores range from 12 to 60, with higher scores indicating greater self-compassion [72].

Mindfulness was assessed using the *Five Facet Mindfulness Questionnaire* measuring nonreactivity and nonjudgment, with subscales rated on a 5-point scale (1="never true" to 5="always true"). Nonreactivity total scores range from 7 to 35, with higher scores indicating a greater ability to not react to or get caught up in thoughts or emotions. Nonjudgment total scores range from 8 to 40, with higher scores indicating a greater ability to not judge or criticize thoughts or emotions [73].

Empathy was assessed using the *Interpersonal Reactivity Index* that measured empathy using 4 subscales (empathic concern, perspective taking, personal distress, and fantasy). Items were rated on a 5-point scale (0="does not describe me" to 4="describes me well"). Each subscale consists of 7 items, and total scores range from 0 to 28, with higher scores indicating greater empathy across subscales. The fantasy subscale was excluded as not relevant [74].

Statistical Analysis

Statistical analyses were conducted in R (version 4.2.2; R Foundation for Statistical Computing). Friedman ANOVA, a nonparametric test, was used to analyze the overall change in cynicism scores (primary outcome) and the other scores (secondary outcomes) at the 3 time points due to the data having a nonnormal distribution. Post hoc analyses between the individual time points were analyzed using Wilcoxon signed rank tests and were corrected for multiple comparisons using the Benjamini-Hochberg procedure. The relationship between the dimensions of burnout and the other variables was evaluated using the Spearman rank correlation coefficient. Effect sizes were reported as Kendall $W=\chi^2/N(K-1)$ for Friedman tests and as Pearson r=z/sqrt (N) for pairwise Wilcoxon signed rank tests. Kendall W and Pearson r use Cohen criteria where 0.1 to <0.3 is a small effect, 0.3 to <0.5 is a medium effect, and ≥0.5 is a large effect [75].

Results

Participants

A total of 40 physicians replied to our recruitment message, and 32 (80%) participants met eligibility criteria and consented

XSL•FO
RenderX

to participate between August 2022 and February 2023. Out of this group who received the allocated intervention, of the 32 physicians, 27 (84%) completed assessments at all 3 time points and were included in the final analysis (Figure 1). The participant population comprised 20 women and 7 men who worked in health care for an average of 15 (SD 9.76) years. The average age was 50.6 (SD 9.74) years (Table 1).

Table 1. Demographic characteristics of study 1.

	Participants (N=27)
Age (y), mean (SD)	50.6 (9.74)
Gender, n (%)	
Man	7 (26)
Woman	20 (74)
Race and ethnicity, n (%)	
Asian	5 (19)
Black or African American	1 (4)
Hispanic, Latinx, or Spanish	0 (0)
White	21 (77)
Work status, n (%)	
Employed full time	22 (82)
Employed part time	5 (8)
Years working in health care, mean (SD)	15 (9.76)
Medical specialty , n (%)	
Emergency medicine	2 (7)
Pediatric emergency medicine	3 (12)
Pediatric hematology oncology	2 (7)
Pulmonary and critical care medicine	2 (7)
Family medicine	6 (22)
Internal medicine	5 (19)
Neurology	1 (4)
Neurocritical care	1 (4)
Psychiatry	2 (7)
Infectious disease	1 (4)
Obstetrics-gynecology	2 (7)

XSL•FO
RenderX

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) flow diagram of study procedures of study 1.

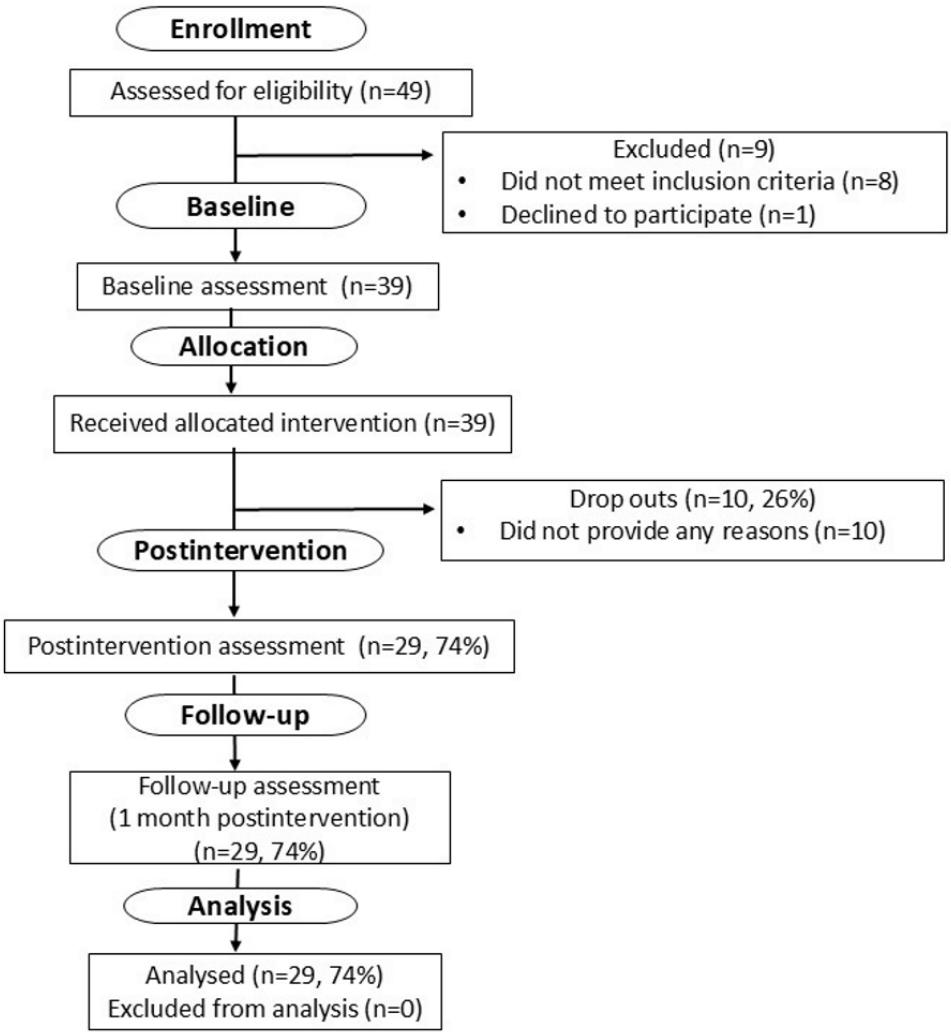

Primary and Secondary Outcomes

The average time taken by participants to complete the program was 30 (SD 16) days.

Table 2 shows the outcome scores at each assessment point compared with baseline.

Friedman ANOVA tests indicated significant changes over time for cynicism (χ^2_2=6.96; *P*=.03; W=0.13), emotional exhaustion (χ^2_2=12.6; *P*=.002; W=0.23), anxiety (χ^2_2=10.7; *P*=.005; W=0.20), intolerance of uncertainty (χ^2_2=12.1; *P*=.002; W=0.22), self-compassion (χ^2_2=9.88; *P*=.007; W=0.18), nonreactivity (χ^2_2=16.7; *P*<.001; W=0.31), and nonjudgment of inner experiences (χ^2_2=19.5; *P*<.001; W=0.36) and personal distress (a subdimension of empathy; χ^2_2=6.74; *P*=.03; W=0.13). No significant change was observed across the 3 time points for the other subdimensions of empathy, such as perspective taking (χ^2_2=5.75; *P*=.06; W=0.106) and empathic concern (χ^2_2=90.3; *P*=.86; W=0.006), and for depression (χ^2_2=1.72; *P*=.42; W=0.032).

Post hoc Wilcoxon signed rank tests revealed 33% reduction of cynicism from baseline to postintervention (W_{26}=134; *P*=.04; *r*=0.41) and 1 month later (W_{26}=140; *P*=.04; *r*=0.47), 25% reduction of emotional exhaustion at both points (postintervention: W_{26}=116; *P*=.02; *r*=0.50 and follow-up: W_{26}=154; *P*=.008; *r*=0.62), and 43% reduction of anxiety at postintervention (W_{26}=244; *P*=.004; *r*=0.62) and 29% one month later (W_{26}=228; *P*=.01; *r*=0.54; Figure 2).

Wilcoxon signed rank tests revealed 26% reduction of intolerance of uncertainty at postintervention (W_{26}=314; *P*<.001; *r*=0.67) and 1 month after the end of the intervention (W_{26}=242; *P*<.001; *r*=0.70), 15% increase of self-compassion at both points (postintervention: W_{26}=73.5; *P*=.02; *r*=0.50 and follow-up: W_{26}=36; *P*=.006; *r*=0.59), 16% increase of nonreactivity (postintervention: W_{26}=22.5; *P*=.001; *r*=0.69 and follow-up: W_{26}=50; *P*=.004; *r*=0.58), and 15% increase of nonjudgment of inner experiences at postintervention (W_{26}=5; *P*<.001; *r*=0.81) and 12% increase 1 month later (W_{26}=47; *P*=.003; *r*=0.60; Figure 3).

XSL•FO
RenderX

Table 2. Descriptive statistics for primary and secondary outcomes (N=27 physicians).

	Values, mean (SD)	Values, median (IQR)	*P* value	Effect sizes (*r*)
Cynicism				
Baseline	2.93 (1.77)	3 (2)	__[a]	—
Postintervention	2.26 (1.26)	2 (1.5)	—	—
1-mo postintervention	2.07 (1.38)	2 (2)	—	—
Δ[b] at postintervention (%)	−0.67 (−23)	−1 (−33)	.04	0.41
Δ at 1-mo postintervention (%)	−0.86 (−29)	−1 (−33)	.04	0.47
Emotional exhaustion				
Baseline	3.89 (1.31)	4 (2)	—	—
Postintervention	3.22 (1.31)	3 (2)	—	—
1-mo postintervention	2.93 (1.41)	3 (2)	—	—
Δ at postintervention (%)	−0.67 (−17)	−1 (−25)	.02	0.50
Δ at 1-mo postintervention (%)	–0.96 (−25)	−1 (−25)	.08	0.62
GAD-7[c]				
Baseline	7.63 (5.25)	7 (8)	—	—
Postintervention	4.59 (4.45)	4 (4.5)	—	—
1-mo postintervention	4.67 (3.85)	5 (4)	—	—
Δ at postintervention (%)	−3.04 (−40)	−3 (−43)	.004	0.62
Δ at 1-mo postintervention (%)	−2.96 (−39)	−2 (−29)	.01	0.54
PHQ-2 [d]				
Baseline	0.56 (0.70)	0 (1)	—	—
Postintervention	0.63 (1.12)	0 (1)	—	—
1-mo postintervention	0.52 (0.80)	0 (1)	—	—
Δ at postintervention (%)	−0.07 (−12)	0 (0)	.99	0.03
Δ at 1-mo postintervention (%)	−0.04 (−7)	0 (0)	.97	0.11
Intolerance of uncertainty				
Baseline	32.6 (10)	34 (12.5)	—	—
Postintervention	27.4 (9.31)	25 (10)	—	—
1-mo postintervention	27.4 (8.97)	25 (13)	—	—
Δ at postintervention (%)	−5.2 (16)	−9 (−26)	<.001	0.67
Δ at 1-mo postintervention (%)	−5.2 (16)	−9 (−26)	<.001	0.70
Self-compassion				
Baseline	34.3 (7)	33 (11)	—	—
Postintervention	38.1 (6.97)	38 (7)	—	—
1-mo postintervention	39.2 (7.87)	38 (11.5)	—	—
Δ at postintervention (%)	3.8 (11)	5 (15)	.02	0.50
Δ at 1-mo postintervention (%)	4.9 (14)	5 (15)	.006	0.59
Nonreactivity				
Baseline	19.3 (4.65)	19 (7)	—	—
Postintervention	22.6 (3.90)	22 (4)	—	—
1-mo postintervention	21.9 (4.26)	22 (4)	—	—
Δ at postintervention (%)	3.3 (17)	3 (16)	.001	0.69

XSL•FO
RenderX

	Values, mean (SD)	Values, median (IQR)	*P* value	Effect sizes (*r*)
Δ at 1-mo postintervention (%)	2.6 (13)	3 (16)	.004	0.58
Nonjudgment				
Baseline	24 (6.66)	26 (10.5)	—	—
Postintervention	29.7 (5.46)	30 (5)	—	—
1-mo postintervention	28.3 (5.51)	29 (6)	—	—
Δ at postintervention (%)	5.7 (24)	4 (15)	<.001	0.81
Δ at 1-mo postintervention (%)	4.3 (18)	3 (12)	.003	0.60
Personal distress				
Baseline	8.52 (4.43)	8 (5.5)	—	—
Postintervention	6.96 (4.57)	7 (6)	—	—
1-mo postintervention	6.74 (4.64)	6 (4.5)	—	—
Δ at postintervention (%)	−1.56 (−18)	−1 (−12)	.03	0.43
Δ at 1-mo postintervention (%)	−1.78 (−21)	−2 (−25)	.03	0.46
Empathic concern				
Baseline	21.5 (5.02)	22 (6)	—	—
Postintervention	22.1 (3.47)	22 (5.5)	—	—
1-mo postintervention	21.6 (3.96)	22 (6)	—	—
Δ at postintervention (%)	0.6 (2.8)	0 (0)	.81	0.07
Δ at 1-mo postintervention (%)	0.1 (0.5)	0 (0)	.81	0.04
Perspective taking				
Baseline	20.3 (4.57)	21 (4.5)	—	—
Postintervention	21.2 (4.31)	21 (5.5)	—	—
1-mo postintervention	21.7 (4.31)	22 (5)	—	—
Δ at postintervention (%)	0.9 (4)	0 (0)	.18	0.26
Δ at 1-mo postintervention (%)	1.4 (7)	1 (5)	.07	0.43

[a]Not applicable.

[b]Δ: change between baseline and stated time point.

[c]GAD-7: Generalized Anxiety Disorder-7.

[d]PHQ-2: Patient Health Questionnaire-2.

Figure 2. Box and whisker plots at baseline, at postintervention, and 1 month after intervention completion for (A) Generalized Anxiety Disorder-7 scores, (B) cynicism scores, and (C) emotional exhaustion scores from the Maslach Burnout Inventory. **P*=.05, ***P*=.01.

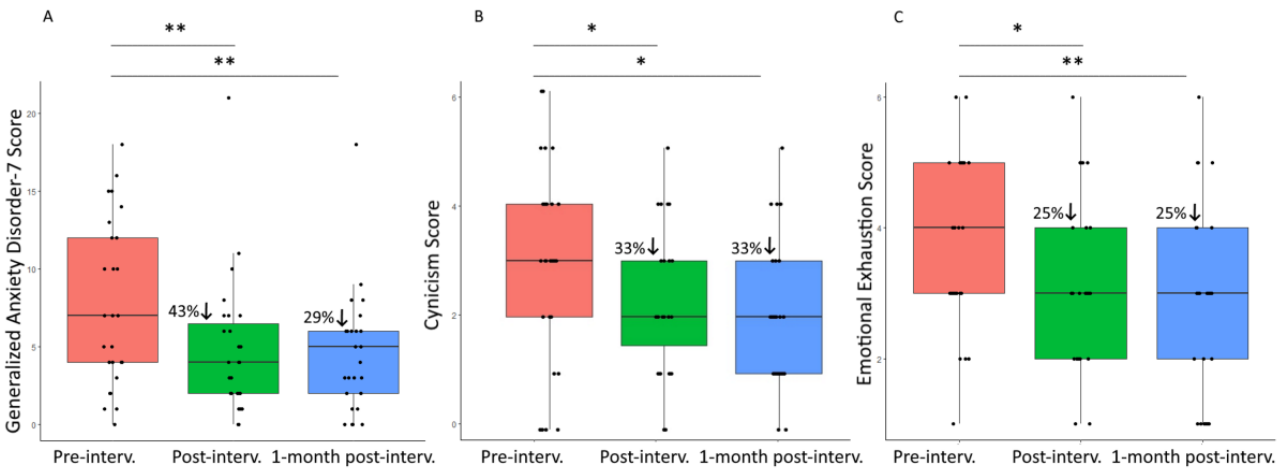

XSL•FO
RenderX

Figure 3. Box and whisker plots at baseline, at postintervention, and 1 month after intervention completion for (A) intolerance of uncertainty scores, (B) self-compassion scores, (C) nonreactivity, and (D) nonjudgment of inner experiences scores. *P=.05, **P=.01, ***P=.001.

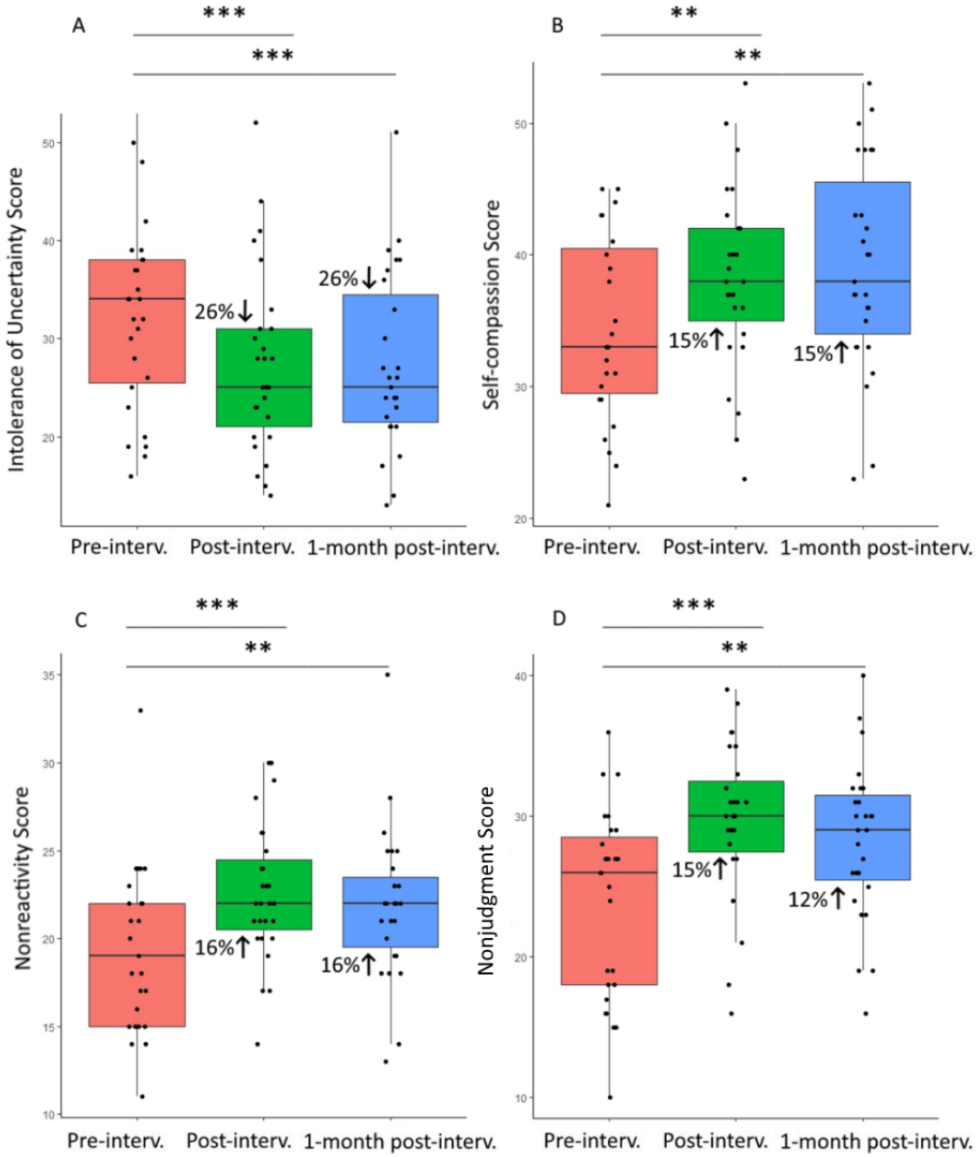

Finally, Wilcoxon signed rank tests showed a 12% reduction in personal distress scores from baseline to postintervention (W_{26}=217; P=.03; r=0.43) and a 25% reduction at 1 month after the end of the intervention (W_{26}=200; P=.03; r=0.46; Figure 4).

XSL•FO
RenderX

Figure 4. Box and whisker plots at baseline, at post-intervention, and 1 month after intervention completion, for personal distress scores. Significance level is denoted by asterisks: *P=.05.

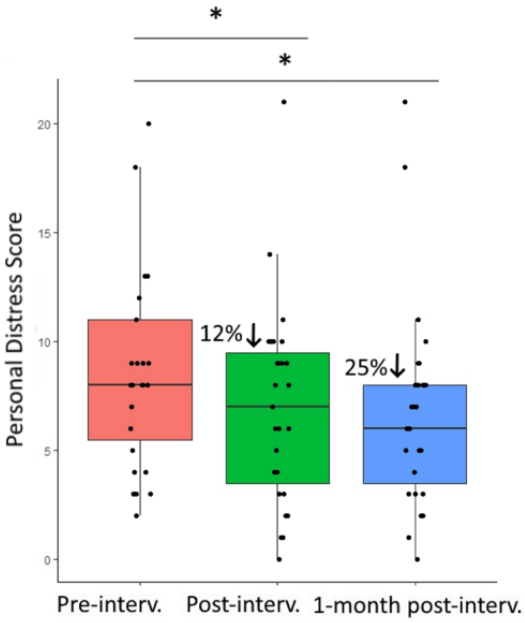

Correlations Between Burnout and Other Outcomes

Table 3 shows the correlations between the 2 burnout dimensions of cynicism and emotional exhaustion and the other outcomes at each assessment point. There were significant correlations between burnout and anxiety at baseline and postintervention (cynicism: $r \geq 0.56$; $P \leq .03$ and emotional exhaustion: $r \geq 0.53$; $P \leq .004$) but not 1 month after the end of the intervention. Burnout scores were also correlated with nonreactivity scores at baseline and postintervention (cynicism: $r \geq 0.39$; $P \leq .04$ and emotional exhaustion: $r \geq -0.40$; $P \leq .04$) but not at follow-up. Emotional exhaustion score was also correlated with self-compassion score only at baseline ($r = -0.39$; $P \leq .043$). We did not observe any other correlations ($r \leq |0.37|$; $P \geq |.06|$).

XSL•FO
RenderX

Table 3. Correlations between burnout scores and other outcomes at baseline, postintervention, and 1-month postintervention.

Variables	Baseline	Postintervention	1-mo postintervention
GAD-7[a] × cynicism			
Value	0.56	0.41	0.24
P value	.002	.03	.22
GAD-7 × emotional exhaustion			
Value	0.60	0.53	0.32
P value	<.001	.004	.99
PHQ-2[b] × cynicism			
Value	0.25	0.12	0.11
P value	.21	.56	.60
PHQ-2 × emotional exhaustion			
Value	0.17	0.08	0.21
P value	.39	.70	.28
IU[c] × cynicism			
Value	0.37	0.02	0.16
P value	.06	.91	.44
IU × emotional exhaustion			
Value	0.15	0.36	0.12
P value	.44	.07	.55
Self-compassion × cynicism			
Value	−0.22	−0.16	−0.05
P value	.27	.44	.80
Self-compassion × emotional exhaustion			
Value	−0.39	−0.17	−0.03
P value	.04	.41	.87
Nonreactivity × cynicism			
Value	−0.39	−0.69	−0.35
P value	.04	<.001	.07
Nonreactivity × emotional exhaustion			
Value	−0.40	−0.49	−0.35
P value	.04	.01	.07
Nonjudgment × cynicism			
Value	−0.34	−0.13	−0.17
P value	.09	.52	.41
Nonjudgment × emotional exhaustion			
Value	−0.37	−0.02	−0.06
P value	.06	.90	.76
Personal distress × cynicism			
Value	0.10	0.08	0.34
P value	.63	.67	.10
Personal distress × emotional exhaustion			
Value	0.06	0.15	0.23

XSL•FO
RenderX

Variables	Baseline	Postintervention	1-mo postintervention
P value	.77	.45	.24
Empathic concern × cynicism			
Value	−0.07	−0.01	−0.24
P value	.75	.95	.23
Empathic concern × emotional exhaustion			
Value	0.27	0.24	0.14
P value	.17	.24	.48
Perspective taking × cynicism			
Value	−0.17	−0.09	−0.10
P value	.38	.67	.62
Perspective taking × emotional exhaustion			
Value	−0.03	−0.25	−0.15
P value	.85	.21	.46

[a]GAD-7: Generalized Anxiety Disorder-7.

[b]PHQ-2: Patient Health Questionnaire-2.

[c]IU: intolerance of uncertainty.

Study 2

Overview

The results of study 1 provided evidence that physicians reported lower cynicism and emotional exhaustion at postintervention assessments compared to baseline. This was also the case for reduced anxiety, intolerance of uncertainty, and personal distress and increased self-compassion and mindfulness after the intervention. No difference was observed for depression and 2 subdimensions of empathy (ie, empathic concern and perspective taking) across time. This lack of change may be attributed to participants reporting low levels of depression and high levels of empathy at baseline.

To confirm the validity of these initial findings from study 1, we conducted a preregistered replication study. We used that same intervention but delivered it via a free smartphone-based platform called Unwinding by Sharecare (Behavioral Health team at Sharecare), which allows a progression of 7 daily modules of brief didactic; 1 summary at the end; and access to further features, such as experience-based MT, app-triggered check-ins, guided meditations (5-15 minutes), and clinician life stories.

Similar to study 1, participants completed web-based surveys at 3 time points: baseline, postintervention, and 1 month later as follow-up. To be more inclusive and expand to a broader range of clinicians, we recruited physicians and nurse practitioners, who provide direct care and medication prescriptions to their patients.

The same primary and secondary outcomes were included as in study 1, with the exception of the Interpersonal Reactivity Index, which was removed due to its lack of appropriateness for physicians, as indicated by the results of study 1.

In addition, based on the strong reduction of anxiety observed in study 1, we aimed to explore the effects of the intervention on further variables that have been suggested to be relevant for the onset and maintenance of anxiety, such as worry [76], sleep disturbances [77], and difficulties in emotion regulation [78].

We evaluated the following hypotheses:

- Hypothesis 1: there will be a main effect of time on cynicism. Physicians and nurse practitioners will report lower scores for cynicism at postintervention assessments compared to baseline.
- Hypothesis 2: there will be a main effect of time on the other variables. Physicians and nurse practitioners will report lower scores for emotional exhaustion, anxiety, depression, intolerance of uncertainty, worry, sleep disturbances, and difficulties in emotion regulation and higher scores for nonreactivity, nonjudging of inner experiences, and self-compassion at postintervention assessments compared to baseline.
- Hypothesis 3: cynicism and emotional exhaustion will positively correlate with anxiety.

Methods

Ethical Considerations

Similar to study 1, this study was approved by the Brown University Institutional Review Board (protocol 2022003296) and followed the Declaration of Helsinki. The same consent process was used: participants were directed to the informed consent form on Qualtrics, and after selecting “yes” (with a waiver of documentation), they completed the web-based survey. Data were anonymized with alphanumeric codes, and the final database used only these codes. Likewise, participants were compensated with a US $25 Amazon gift card for the completion of each survey (US $75 for 3 surveys). The study was preregistered at ClinicalTrials.gov (NCT06145425), except for

XSL•FO
RenderX

the exploratory outcome of difficulties in emotion regulation that we added later.

Participants

The same study design, recruitment process, and inclusion and exclusion criteria were used as in study 1, except for the additional recruitment of nurse practitioners. To be more inclusive, we recruited physicians and nurse practitioners, who are equally skilled and knowledgeable in providing direct care and medication prescriptions to their patients.

Intervention

The same intervention was delivered via the smartphone app Unwinding by Sharecare, which included 7 daily modules of MT, a final summary, app-triggered check-ins, guided meditations (5-15 minutes), and clinician life stories. Participants were encouraged to explore other app features; however, this was optional. Similar to study 1, participants completed 1 module per day at their own pace, with check-in messages from the project director every 15 days to address technical issues and encourage engagement.

Engagement

Data related to program engagement, as measured by the number of modules completed, were obtained directly from the Unwinding by Sharecare app.

Outcome Measures

The same approach was used as described in study 1: participants completed web-based surveys at 3 time points: baseline, postintervention, and 1 month later as follow-up. The same outcomes were included as described in study 1, with the following exceptions: the Interpersonal Reactivity Index was removed because the results of study 1 showed that this measure might not be appropriate for physicians. The Patient-Reported Outcomes Measurement Information System (PROMIS), the Penn State Worry Questionnaire, and the Difficulties in Emotion Regulation Scale were added as secondary outcomes to explore the effects of the intervention on sleep disturbances, worry, and difficulties in emotion regulation, respectively.

PROMIS was used to measure worry-related sleep disturbances. Five items from PROMIS were preselected based on their direct assessment of worry interfering with sleep [79]. Items were rated on a 5-point Likert scale (1="not at all" and 5="very much"). Scores range from 5 to 25, with higher scores indicating more sleep disturbances and greater sleep-related impairment [79,80].

Penn State Worry Questionnaire is a 16-item questionnaire to assess worry on a 5-point Likert scale (1="not at all typical of me" and 5="very typical of me"). Scores range from 16 to 80, with higher total scores indicating increased levels of worry [81].

Difficulties in Emotion Regulation Scale is a 36-item questionnaire used to assess emotion regulation challenges on a 5-point Likert scale (1="almost never" and 5="almost always"). Scores range from 36 to 180, with higher total scores indicating increased greater emotion regulation difficulty [82].

Statistical Analysis

The same data analysis approach described in study 1 was used.

Results

Participants

A total of 40 physicians and 10 nurse practitioners replied to our recruitment message, and 39 (78%) participants (29 physicians and 10 nurse practitioners) met the eligibility criteria and consented to participate between June 2023 and May 2024. Out of this group who received the allocated intervention, of the N = 39 participants, 29 (74%, n=20, 69% physicians and n=9, 90% nurse practitioners) completed the 3-time point assessments and were included in the final analysis (Figure 5). The participant population comprised 23 women and 6 men who worked in health care for an average of 16.4 (SD 10.1) years. The average age was 49.6 (SD 9.8) years (Table 4).

XSL•FO
RenderX

Figure 5. CONSORT (Consolidated Standards of Reporting Trials) flow diagram of study procedures of study 2.

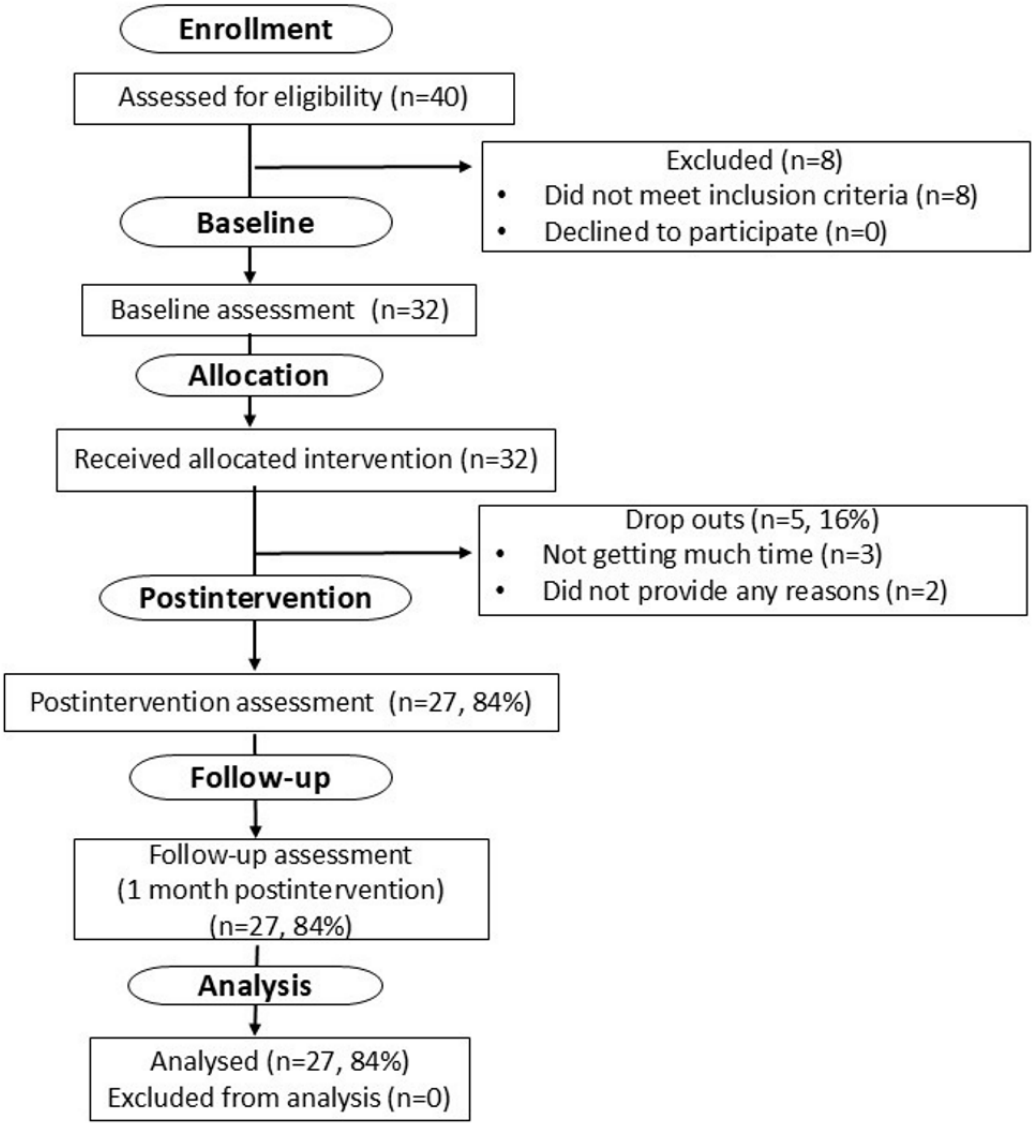

XSL•FO
RenderX

Table 4. Demographic characteristics of study 2.

	Physicians+nurse practitioners (n=29)	Physicians (n=20)	Nurse practitioners (n=9)
Age (y), mean (SD)	49.6 (9.8)	49.8 (10.9)	49.1 (6.9)
Gender, n (%)			
Men	6 (21)	5 (25)	1 (11)
Women	23 (79)	15 (75)	8 (89)
Race and ethnicity, n (%)			
Asian	3 (10)	3 (15)	0 (0)
Black or African American	2 (7)	1 (5)	1 (11)
Hispanic, Latinx, or Spanish	0 (0)	0 (0)	0 (0)
White	24 (83)	16 (80)	8 (89)
Work status, n (%)			
Employed full time	27 (93)	18 (90)	9 (100)
Employed part time	2 (7)	2 (10)	0 (0)
Working duration in health care (y), mean (SD)	16.41 (10.14)	14.85 (8.83)	19.89 (11.86)
Medical specialty, n (%)			
Emergency medicine	3 (10.2)	2 (10)	1 (11)
Pediatrics	6 (21)	2 (10)	4 (45)
Oncology	1 (3.4)	0 (0)	1 (11)
Palliative medicine	2 (7)	2 (10)	0 (0)
Family medicine	2 (7)	1 (5)	1 (11)
Internal medicine	4 (14)	4 (20)	0 (0)
Nephrology	1 (3.4)	1 (5)	0 (0)
Hospitality	1 (3.4)	1 (5)	0 (0)
Psychiatry	1 (3.4)	0 (0)	1 (11)
Infectious disease	1 (3.4)	1 (5)	0 (0)
Obstetrics-gynecology	3 (10.2)	3 (15)	0 (0)
Endocrinology	1 (3.4)	1 (5)	0 (0)
Clinical immunology	1 (3.4)	1 (5)	0 (0)
Physical medicine	1 (3.4)	1 (5)	0 (0)
Orthopedics	1 (3.4)	0 (0)	1 (11)

Engagement

At posttreatment, 100% (29/29) of the participants completed the 7-module program; only 3 (10%) participants did not listen to the summary at the end. The average time taken by participants to complete the program was 23.8 (SD 18.7) days.

Primary and Secondary Outcomes

Multimedia Appendix 1 shows the outcome scores at each assessment point compared with baseline. Friedman ANOVA tests indicated significant changes over time for cynicism (χ^2_2=13.1; *P*=.001; W=0.23), emotional exhaustion (χ^2_2=2.39; *P*=.30; W=0.04), anxiety (χ^2_2=14.6; *P*<.001; W=0.25), intolerance of uncertainty (χ^2_2=15.6; *P*<.001; W=0.27), self-compassion (χ^2_2=22.2; *P*<.001; W=0.38), nonreactivity (χ^2_2=24.3; *P*<.001; W=0.42), and nonjudgment of inner experiences (χ^2_2=10.3; *P*=.006; W=0.18). Friedman ANOVA tests revealed significant changes over time for worry (χ^2_2=7.53; *P*=.02; W=0.13), sleep disturbances (χ^2_2=7.07; *P*=.03; W=0.122), and difficulties in emotion regulation (χ^2_2=15.4; *P*<.001; W=0.27). No significant change across the 3 time points was observed for depression (χ^2_2=1.77; *P*=.41; W=0.03).

Post hoc Wilcoxon signed rank tests revealed 33% reduction of cynicism at both postintervention (W_{28}=184; *P*=.008; *r*=0.58) and 1 month after the end of the intervention (W_{28}=196; *P*=.04; *r*=0.45) and 29% reduction of anxiety scores from baseline to postintervention (W_{28}=317; *P*=.01; *r*=0.49) and 57% reduction 1 month later (W_{28}=340; *P*=.006; *r*=0.59; Figure 6). We

XSL•FO
RenderX

observed stronger results when we analyzed only the data from physicians: a significant change across the 3 time points for cynicism (χ^2_2=16.3; *P*<.001; W=0.41) and a 50% reduction in scores from baseline to postintervention (W_{19}=105; *P*=.003; *r*=0.81) and 33% reduction 1 month later (W_{19}=128; *P*=.02; *r*=0.58). Likewise, we observe similar results for anxiety over time (χ^2_2=10.5; *P*=.005; W=0.26) and 33% reduction from baseline to postintervention (W_{19}=150; *P*=.04; *r*=0.51) and 60% reduction 1 month later (W_{19}=165; *P*=.02; *r*=0.63).

Wilcoxon signed rank tests found 12% reduction of intolerance of uncertainty scores from baseline to postintervention (W_{26}=338; *P*=.003; *r*=0.57) and 8% reduction 1 month after the end of the intervention (W_{26}=310; *P*<.001; *r*=0.66) and 21% increase in self-compassion scores from baseline to postintervention (W_{28}=32.5; *P*<.001; *r*=0.69) and 27% increase 1 later (W_{28}=2; *P*<.001; *r*=0.84). We observed similar results with data collected only from physicians: significant changes across the 3 time points for intolerance of uncertainty (χ^2_2=9.24; *P*=.01; W=0.23) and for self-compassion (χ^2_2=15.5; *P*<.001; W=0.39). Intolerance of uncertainty scores reduced by 22% at postintervention (W_{19}=162; *P*=.01; *r*=0.59) and by 24% one month later (W_{19}=158; *P*=.003; *r*=0.72) and self-compassion scores increased by 31% at postintervention (W_{19}=9.5; *P*=.001; *r*=0.74) and 38% one month later (W_{19}=1.5; *P*<.001; *r*=0.82).

Wilcoxon signed rank tests also found 33% increase in nonreactivity scores from baseline to postintervention (W_{28}=38.5; *P*<.001; *r*=0.72) and 39% increase 1 month after the end of the intervention (W_{28}=9; *P*<.001; *r*=0.83) and 8% increase in nonjudgment scores at postintervention (W_{28}=60.5; *P*=.009; *r*=0.50) and 12% increase 1 month later (W_{28}=38.5; *P*=.003; *r*=0.62; Figure 7). We observed similar results with data only from physicians: a significant change across the 3 time points for nonreactivity (χ^2_2=18.8; *P*<.001; W=0.47) and for nonjudgment (χ^2_2=7.10; *P*=.03; W=0.18) and 28% increase in nonreactivity scores at postintervention (W_{19}=21; *P*=.003; *r*=0.71) and 39% increase 1 month later (W_{19}=4; *P*<.001; *r*=0.85) and 10% increase in nonjudgment scores at postintervention (W_{19}=40.5; *P*=.08; *r*=0.43) and 12% increase 1 month later (W_{19}=19.5; *P*=.02; *r*=0.62).

Figure 6. Box and whisker plots at baseline, at postintervention, and 1 month after intervention completion for (A) Generalized Anxiety Disorder-7 scores, (B) cynicism scores, and (C) emotional exhaustion scores from the Maslach Burnout Inventory in 29 physicians and nurse practitioners. **P*=.05, ***P*=.01.

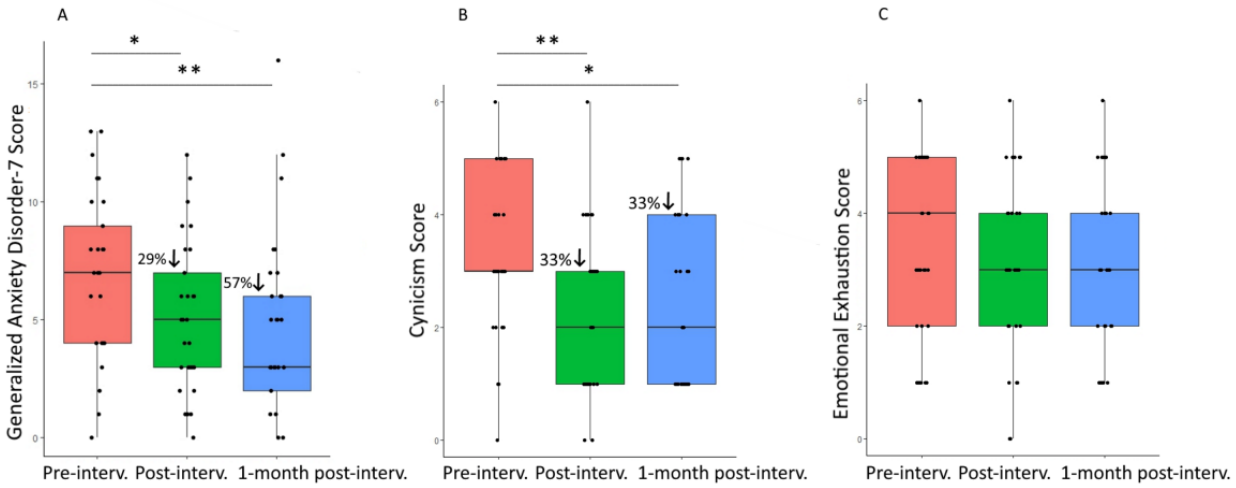

XSL•FO
RenderX

Figure 7. Box and whisker plots at baseline, at postintervention, and 1 month after intervention completion for (A) intolerance of uncertainty scores, (B) self-compassion scores, and (C) nonreactivity toward inner experiences scores in 29 physicians and nurse practitioners. *P=.05, **P=.01,***P=.001.

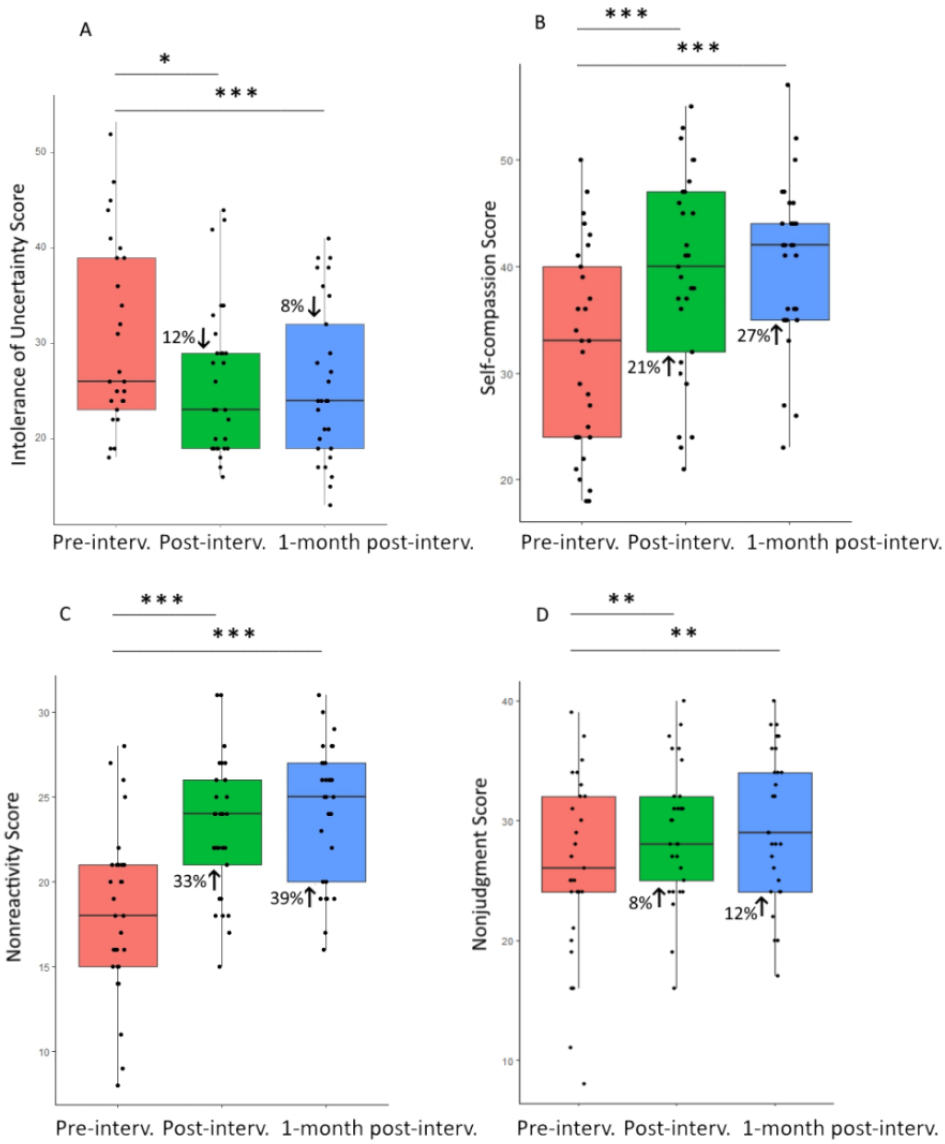

Finally, Wilcoxon signed rank tests revealed 14% reduction in worry scores from baseline to postintervention (W_{28}=247; P=.04; r=0.40) and 12% reduction 1 month after the end of the intervention (W_{28}=319; P=.006; r=0.58), 10% reduction in sleep disturbances at postintervention (W_{28}=283; P=.04; r=.42) and 20% reduction 1 month later (W_{28}=268; P=.01; r=0.53), and 4% reduction in difficulties in emotion regulation scores at postintervention (W_{28}=353; P=.005; r=0.54) and 8% reduction 1 month later (W_{28}=392; P<.001; r=0.70; Figure 8). We observe less strong results with data only from physicians: a trend change across the 3 time points for worry (χ^2_2=5.30; P=.07; W=0.13) and significant changes for sleep (χ^2_2=7.53; P=.02; W=0.188) and emotion regulation (χ^2_2=9.29; P=.009; W=0.23). We only observed 21% reduction in worry at 1 month after the end of the intervention (W_{19}=168; P=.01; r=0.66) but not at postintervention (W_{19}=116; P=.09; r=0.37). We found 22% reduction in sleep disturbances scores from baseline to postintervention (W_{19}=156; P=.02; r=0.54) and 30% reduction 1 month later (W_{19}=152; P=.01; r=0.64) and 8% reduction in difficulties in emotion regulation at postintervention (W_{19}=176; P=.01; r=0.59) and 11% reduction 1 month later (W_{19}=192; P=.04; r=0.73).

XSL•FO
RenderX

Figure 8. Box and whisker plots at baseline, at postintervention, and 1 month after intervention completion for (A) worry scores, (B) sleep disturbances scores, and (C) difficulties in emotion regulation scores in 29 physicians and nurse practitioners. *P=.05, **P=.01,***P=.001.

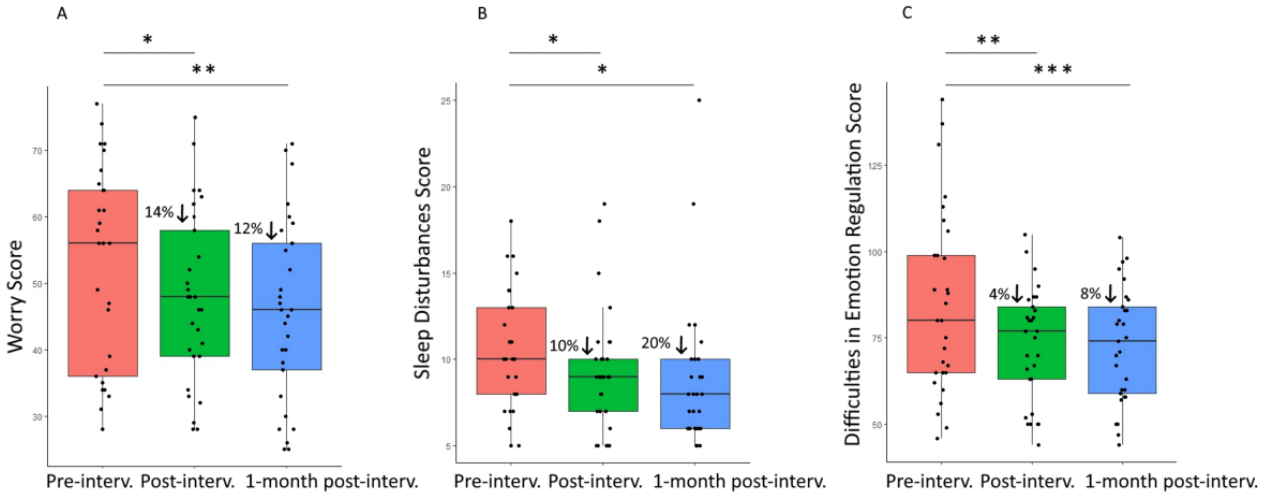

Correlations Between Burnout and Other Outcomes

Table 5 shows the correlations between the 2 burnout dimensions of cynicism and emotional exhaustion and the other outcomes at each assessment point. There were significant correlations between anxiety and burnout at baseline and 1 month after the end of the intervention (cynicism: $r \geq 0.39$; $P \leq .04$ and emotional exhaustion: $r \geq 0.58$; $P \leq .001$) but not right after the intervention. Burnout were correlated also with self-compassion scores (cynicism: $r \geq -0.42$; $P \leq .02$ and emotional exhaustion: $r = -0.56$; $P = .002$), intolerance of uncertainty scores (cynicism: $r \geq 0.37$; $P \leq .05$ and emotional exhaustion: $r \geq 0.40$; $P \leq .03$), and depression (cynicism: $r = 0.41$; $P = .03$ and emotional exhaustion: $r = 0.41$; $P = .03$). Burnout scores were also correlated with worry (cynicism: $r \geq 0.45$; $P \leq .01$ and emotional exhaustion: $r \geq 0.51$; $P \leq .004$) and emotion regulation (cynicism: $r \geq 0.46$; $P \leq .01$ and emotional exhaustion: $r = 0.40$; $P = .03$). Cynicism scores were also correlated with nonjudgment at baseline ($r = -0.49$; $P = .006$) and with sleep after the intervention ($r = 0.38$; $P = .04$). Emotional exhaustion scores were also correlated with nonreactivity scores only at baseline ($r = -0.56$; $P = .002$). We did not observe any other correlations ($r \leq |0.36|$; $P \geq |.054|$).

XSL•FO
RenderX

Table 5. Correlations between anxiety and burnout scores at baseline, postintervention, and 1-month postintervention.

Variables	Baseline	Postintervention	1-mo postintervention
GAD-7[a] × cynicism			
Value	0.56	0.41	0.24
P value	.002	.03	.22
GAD-7 × emotional exhaustion			
Value	0.60	0.53	0.32
P value	<.001	.004	.10
PHQ-2[b] × cynicism			
Value	0.22	0.05	0.41
P value	.25	.81	.03
PHQ-2 × emotional exhaustion			
Value	0.35	0.23	0.41
P value	.06	.24	.03
IU[c] × cynicism			
Value	0.52	0.28	0.37
P value	.004	.13	.05
IU × emotional exhaustion			
Value	0.41	0.45	0.4
P value	.03	.01	.03
Self-compassion × cynicism			
Value	−0.52	−0.42	−0.62
P value	.004	.002	<.001
Self-compassion × emotional exhaustion			
Value	−0.35	−0.17	−0.56
P value	.06	.37	.002
Nonreactivity × cynicism			
Value	−0.27	−0.20	−0.3
P value	.15	.29	.11
Nonreactivity × emotional exhaustion			
Value	−0.56	−0.3	−0.26
P value	.002	.12	.12
Nonjudgment × cynicism			
Value	−0.49	−0.3	−0.34
P value	.006	.11	.07
Nonjudgment × emotional exhaustion			
Value	−0.25	−0.18	−0.34
P value	.19	.35	.08
Worry × cynicism			
Value	−0.49	−0.45	−0.64
P value	.007	.01	.001
Worry × emotional distress			
Value	0.51	0.32	0.54

XSL•FO
RenderX

Variables	Baseline	Postintervention	1-mo postintervention
P value	.004	.09	.003
Difficulties in emotion regulation × cynicism			
Value	0.49	0.33	0.53
P value	.01	.08	.003
Difficulties in emotion regulation × emotional exhaustion			
Value	0.35	0.31	0.4
P value	.07	.10	.03
Sleep disturbances × cynicism			
Value	−0.22	−0.38	−0.25
P value	.26	.04	.19
Sleep disturbances × emotional exhaustion			
Value	−0.29	−0.2	0.36
P value	.12	.30	.60

[a]GAD-7: Generalized Anxiety Disorder-7.

[b]PHQ-2: Patient Health Questionnaire-2.

[c]IU: intolerance of uncertainty.

Discussion

Principal Findings

To our knowledge, this is the first research to assess the efficacy of a digital MT program designed with input from clinicians specifically to address burnout. Two independent single-arm studies were conducted to evaluate its impact on physicians. In the first study, MT was delivered via podcast, while in the second study, it was administered through a free mobile app.

Results from both studies demonstrated that MT led to reductions in burnout, anxiety, intolerance of uncertainty, and personal distress, while also fostering increases in self-compassion and mindfulness, with effect sizes ranging from medium to large. In addition, the second study revealed decreases in worry, sleep disturbances, and difficulties in emotion regulation. However, no significant changes were observed over time for depression or perspective taking and empathic concern, potentially due to the already low levels of depression and high levels of perspective taking and empathic concern reported at baseline. Furthermore, significant positive correlations were found between burnout and anxiety at baseline in both studies, at postintervention in study 1, and at follow-up in study 2, thus confirming the relationship between these 2 variables.

Reduction of Physician Burnout

Preliminary findings suggest that this training could be an efficacious, accessible way for busy physicians to combat burnout. Both studies showed that it reduced cynicism, with the first also reducing emotional exhaustion. This may be due to the focus of the intervention on individual well-being, particularly addressing the interpersonal aspect of burnout. However, individual-focused interventions alone may not fully address physician burnout, which often requires collaborative efforts at both the individual and organizational levels [33,83].

Increase of Mindfulness and Self-Compassion

This study builds upon prior research demonstrating that in-person mindfulness-based interventions have been effective in reducing burnout and anxiety among physicians [37-40]. Importantly, we were able to replicate these findings using a digital program that is both free and less time consuming than the intensive in-person MT used in previous studies. This digital intervention may therefore offer a more feasible option for integration into the already hectic schedules of physicians. In addition, this intervention contributes to the growing field of digital therapeutics by demonstrating clinically meaningful effects [7,60,61,84].

Consistent with numerous studies, we found that MT increased nonreactivity and nonjudgmental attitudes toward thoughts, sensations, and emotions, while also enhancing self-compassion. Self-compassion has been theorized to provide self-care during times of distress, failure, and difficulty and to enable individuals to offer compassionate care to others. This concept comprises 3 interconnected components: mindfulness, self-kindness, and common humanity [46]. In practical terms, self-compassion can facilitate several key aspects: (1) recognizing and accepting feelings as they arise without suppression or avoidance (ie, mindfulness), (2) treating oneself with kindness and understanding (ie, self-kindness), and (3) acknowledging one's own pain and distress as part of the broader human experience (ie, common humanity [46]).

Therefore, self-compassion may enable health care providers, who often struggle with perfectionism and self-criticism in high-stakes environments where errors can have grave consequences, to better navigate the challenges they face. Such feelings of inadequacy have been linked to anxiety, depression, and even suicidal ideation among medical students [85,86] and physicians [87]. For instance, interventions promoting self-compassion among athletes have proven successful in

XSL•FO
RenderX

helping them manage concerns about mistakes, self-criticism, and rumination during setbacks and uncertainties, thereby improving both athletic performance and psychological well-being [88].

Reduction of Personal Distress and Intolerance of Uncertainty

Regarding empathy, we only observed a significant reduction in personal distress during postintervention assessments. This finding aligns with existing literature, which has consistently reported that personal distress is the empathic component positively associated with burnout [89]. However, we did not observe any effects on the other 2 dimensions of empathy: perspective taking and empathic concern. This outcome may be due to a ceiling effect, as previous research with health care professionals has also reported [90]. The high baseline scores suggest that the chosen questionnaire may not have been sensitive enough to capture changes in these dimensions among physicians.

Although previous literature has shown a link between high intolerance of uncertainty and an increased risk or presence of physician burnout [24,91] and work-related stress and satisfaction [21,92], these 2 studies are the first to investigate the effects of MT on intolerance of uncertainty. We found a significant reduction with a large effect size. Recognizing uncertainty and developing coping strategies are crucial in the field of medicine and are recognized as essential clinical competencies to be fostered throughout medical education and training [93,94].

For example, when faced with clinical situations where outcomes are difficult to predict, physicians with high intolerance of uncertainty may experience excessive worry and anxiety about the consequences of their decisions. They may also struggle with self-doubt and insecurity regarding their professional competence, leading to a sense of being “stuck” in uncertainty and hesitancy in decision-making [95,96]. It is estimated that 17% of the excessive costs in medical care result from physicians’ anxiety related to uncertainty management, leading to increased tendencies to order further tests [97], failures to adhere to evidence-based guidelines [98], and fear of malpractice litigation and defensive practices [99].

Reduction of Anxiety and Related Variables

Importantly, we observed that MT reduced anxiety by 40% postintervention, with moderate to large effect sizes. Furthermore, we confirmed the positive correlation between burnout and anxiety, although this correlation was not significant across all 3 time points in both studies. These results build upon previous findings regarding the relationship between burnout and anxiety [7].

Interestingly, MT also reduced variables previously associated with the onset and maintenance of anxiety, such as worry, sleep disturbances, and difficulties with emotion regulation. In previous research, app-based MT significantly reduced anxiety in individuals with generalized anxiety disorders [61] and improved sleep disturbances [84]. These effects were mediated by a reduction in worry [61,84]. MT appears to assist individuals in interrupting habitual worry patterns by fostering present-focused awareness of thoughts and emotions. Guiding physicians to observe and manage repetitive worry patterns rather than reacting to and reinforcing them could alleviate anxiety and burnout.

In addition, research indicates that being aware of one’s own emotions and regulating them in stressful work environments can not only reduce personal distress but also promote prosocial behavior [100,101]. For instance, when physicians encounter patients with sudden worsening symptoms, they need to manage their own discomfort to provide appropriate care for their patients considered distressed.

Limitations

These studies offer several strengths and significant findings, such as a freely accessible training program designed for real-world application, preregistered outcomes, and the inclusion of a diverse range of medical specialties in the samples. However, they also have several limitations.

First, both studies are single-arm trials with relatively short follow-up periods (ie, 1 month after completing the training). To validate the efficacy of this program and ascertain its long-term effects, future studies using a waiting list or active control intervention with longer follow-up periods are warranted. These studies could also investigate the underlying psychological mechanisms.

Second, the study samples are relatively small and not fully representative, with predominantly White and female physician participants. While future studies with balanced representation across sexes are necessary for generalizability, it is worth noting that women are more likely to be affected by burnout and anxiety compared to men [102-104]. In addition, it is important to recognize that burnout affects not only physicians but also other health care professionals and students. Moving forward, it is critical to adapt and test interventions across a wider population within the health care system. Addressing clinician burnout early in professional development and acknowledging specific stressors in the learning environment are essential steps in mitigating its impact. In addition, both studies involved self-selection of participants, meaning that those who chose to participate were likely already predisposed to favor the program. This suggests that the intervention might have been particularly beneficial for individuals who were attracted to the mindfulness component of the program. As a result, we do not know whether the intervention would affect those who were less interested in mindfulness or less inclined to participate.

Finally, the impact of the intervention was assessed exclusively through self-report measures and individual-related variables, without considering the individual-organization relationship. As we have noted, both individual and organizational factors contribute to burnout. Similar to recent findings in the field [105], future research should incorporate measures such as work engagement, teamwork, and feelings of being valued by the organization, in addition to patient-related ecological outcomes. While collecting data on mortality rates, complication rates, and disease progression might be challenging across different health care institutions, other measures such as patient compliance, the strength of the patient–health care provider

XSL•FO
RenderX

relationship, compassionate care, and testing and prescribing behaviors can be effectively implemented. Understanding these broader outcomes is crucial for developing effective interventions and strategies to address physician burnout and improve patient care and health outcomes.

Conclusions

Despite these limitations, both studies demonstrated that physicians were open to trying a pragmatically designed and delivered digital MT program and could potentially benefit from it. They suggest that digital MT holds promise as an accessible, evidence-based tool for addressing burnout and related aspects of anxiety among physicians. However, it is important to note that causal claims cannot be definitively made until future randomized controlled studies are conducted. These proof-of-concept studies serve as an initial exploration of potential mechanisms and effect size calculations for larger randomized controlled trials.

Furthermore, this evidence underscores the need for both individual-focused and organizational-focused approaches to addressing burnout, and it highlights the importance of involving physicians in the design of new interventions to create effective and sustainable strategies for preventing or mitigating burnout and promoting professional well-being.

Acknowledgments

The authors are grateful to Lauren Allister, Jacqui Barnett, Courtney Bilodeau, Kenneth Botelho, Robin Boudette, Ruth Chen, Christine Evans, Elizabeth Freire, Alisa Hale, Donna Hill Howes, Ana Hope Logghe, Katherine King, Lucia Larson, Nancy Lasson, Beth Lee-Herbert, Laurie Marbas, Gemma Majewski, Mary McLellan, Diana Meeks, Veronica Modise, Shantell Moreno, Chika Nwokenna, Samba Nyirenda, Nkenna Odom, Saley Pels, Heidi Rayala, Cynthia Rivera, Nickholas Ross, Nancy Quesada-Rodriguez, Chiara Saracci, Greg Salgueiro, Paul Sidandi, Rob Suhoza, David Scott Upton, Vinicius Vieira Neves, and Hannah Whitmore for their contribution to the development of the mindfulness intervention used in these studies. This research was financed by the Swiss National Science Foundation (P500PS_202924) and was made possible also in part by a grant from the Peter G Peterson Foundation (#008225). The statements and views expressed are solely the responsibility of the authors.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

LA and JB were responsible for the conceptualization of the study, and both authors were involved in securing funding for the research. The investigation process was carried out by LA and JB, who also played key roles in developing the methodology and overseeing the project administration. LA managed data curation and formal analysis, ensuring the accuracy and interpretation of the data. In addition, LA wrote the first draft of the manuscript, which was subsequently reviewed and edited by both authors.

Conflicts of Interest

JB owns stock in MindSciences, the company that developed the mindfulness app used in this study. This financial interest has been disclosed to Brown University and is being managed in accordance with its Conflict of Interest and Conflict of Commitment policies. All other authors declare no other conflicts of interest.

Multimedia Appendix 1
Descriptive statistics for primary and secondary outcomes (N=29 physicians+nurse practitioners).
[DOCX File , 20 KB - formative_v9i1e63197_app1.docx]

References

1. Shanafelt TD, Boone S, Tan L, Dyrbye LN, Sotile W, Satele D, et al. Burnout and satisfaction with work-life balance among US physicians relative to the general US population. Arch Intern Med 2012 Oct 08;172(18):1377-1385. [doi: 10.1001/archinternmed.2012.3199] [Medline: 22911330]
2. Shanafelt TD, Gradishar WJ, Kosty M, Satele D, Chew H, Horn L, et al. Burnout and career satisfaction among US oncologists. J Clin Oncol 2014 Mar 01;32(7):678-686 [FREE Full text] [doi: 10.1200/JCO.2013.51.8480] [Medline: 24470006]
3. Shanafelt TD, Hasan O, Dyrbye LN, Sinsky C, Satele D, Sloan J, et al. Changes in burnout and satisfaction with work-life balance in physicians and the general US working population between 2011 and 2014. Mayo Clin Proc 2015 Dec;90(12):1600-1613. [doi: 10.1016/j.mayocp.2015.08.023] [Medline: 26653297]
4. Shanafelt TD, West CP, Dyrbye LN, Trockel M, Tutty M, Wang H, et al. Changes in burnout and satisfaction with work-life integration in physicians during the first 2 years of the COVID-19 pandemic. Mayo Clin Proc 2022 Dec;97(12):2248-2258 [FREE Full text] [doi: 10.1016/j.mayocp.2022.09.002] [Medline: 36229269]

XSL•FO
RenderX

5. Burn-out an "occupational phenomenon": international classification of diseases. World Health Organization. 2019. URL: https://www.who.int/news/item/28-05-2019-burn-out-an-occupational-phenomenon-international-classification-of-diseases [accessed 2024-06-07]
6. Maslach C, Leiter MP. New insights into burnout and health care: strategies for improving civility and alleviating burnout. Med Teach 2017 Feb;39(2):160-163. [doi: 10.1080/0142159X.2016.1248918] [Medline: 27841065]
7. Roy A, Druker S, Hoge EA, Brewer JA. Physician anxiety and burnout: symptom correlates and a prospective pilot study of app-delivered mindfulness training. JMIR Mhealth Uhealth 2020 Apr 01;8(4):e15608 [FREE Full text] [doi: 10.2196/15608] [Medline: 32234708]
8. Melamed S, Ugarten U, Shirom A, Kahana L, Lerman Y, Froom P. Chronic burnout, somatic arousal and elevated salivary cortisol levels. J Psychosom Res 1999 Jun;46(6):591-598. [doi: 10.1016/s0022-3999(99)00007-0] [Medline: 10454175]
9. Vela-Bueno A, Moreno-Jiménez B, Rodríguez-Muñoz A, Olavarrieta-Bernardino S, Fernández-Mendoza J, De la Cruz-Troca JJ, et al. Insomnia and sleep quality among primary care physicians with low and high burnout levels. J Psychosom Res 2008 Apr;64(4):435-442. [doi: 10.1016/j.jpsychores.2007.10.014] [Medline: 18374744]
10. Weaver MD, Robbins R, Quan SF, O'Brien CS, Viyaran NC, Czeisler CA, et al. Association of sleep disorders with physician burnout. JAMA Netw Open 2020 Oct 01;3(10):e2023256 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.23256] [Medline: 33125492]
11. Ryan E, Hore K, Power J, Jackson T. The relationship between physician burnout and depression, anxiety, suicidality and substance abuse: a mixed methods systematic review. Front Public Health 2023 Mar 30;11:1133484 [FREE Full text] [doi: 10.3389/fpubh.2023.1133484] [Medline: 37064688]
12. Shanafelt TD, Balch CM, Dyrbye L, Bechamps G, Russell T, Satele D, et al. Special report: suicidal ideation among American surgeons. Arch Surg 2011 Jan;146(1):54-62. [doi: 10.1001/archsurg.2010.292] [Medline: 21242446]
13. Schernhammer ES, Colditz GA. Suicide rates among physicians: a quantitative and gender assessment (meta-analysis). Am J Psychiatry 2004 Dec;161(12):2295-2302. [doi: 10.1176/appi.ajp.161.12.2295] [Medline: 15569903]
14. Sinsky CA, Brown RL, Stillman MJ, Linzer M. COVID-related stress and work intentions in a sample of US health care workers. Mayo Clin Proc Innov Qual Outcomes 2021 Dec;5(6):1165-1173 [FREE Full text] [doi: 10.1016/j.mayocpiqo.2021.08.007] [Medline: 34901752]
15. Han S, Shanafelt TD, Sinsky CA, Awad KM, Dyrbye LN, Fiscus LC, et al. Estimating the attributable cost of physician burnout in the United States. Ann Intern Med 2019 Jun 04;170(11):784-790. [doi: 10.7326/M18-1422] [Medline: 31132791]
16. Dewa CS, Loong D, Bonato S, Trojanowski L. The relationship between physician burnout and quality of healthcare in terms of safety and acceptability: a systematic review. BMJ Open 2017 Jun 21;7(6):e015141 [FREE Full text] [doi: 10.1136/bmjopen-2016-015141] [Medline: 28637730]
17. Salyers MP, Bonfils KA, Luther L, Firmin RL, White DA, Adams EL, et al. The relationship between professional burnout and quality and safety in healthcare: a meta-analysis. J Gen Intern Med 2017 Apr;32(4):475-482 [FREE Full text] [doi: 10.1007/s11606-016-3886-9] [Medline: 27785668]
18. Dyrbye LN, Varkey P, Boone SL, Satele DV, Sloan JA, Shanafelt TD. Physician satisfaction and burnout at different career stages. Mayo Clin Proc 2013 Dec;88(12):1358-1367. [doi: 10.1016/j.mayocp.2013.07.016] [Medline: 24290109]
19. Southwick FS, Southwick SM. The loss of a sense of control as a major contributor to physician burnout: a neuropsychiatric pathway to prevention and recovery. JAMA Psychiatry 2018 Jul 01;75(7):665-666. [doi: 10.1001/jamapsychiatry.2018.0566] [Medline: 29799941]
20. Freudenberger HJ. Staff burn ‐ out. J Soc Issues 2010 Apr 14;30(1):159-165. [doi: 10.1111/j.1540-4560.1974.tb00706.x]
21. Iannello P, Mottini A, Tirelli S, Riva S, Antonietti A. Ambiguity and uncertainty tolerance, need for cognition, and their association with stress. A study among Italian practicing physicians. Med Educ Online 2017;22(1):1270009 [FREE Full text] [doi: 10.1080/10872981.2016.1270009] [Medline: 28178917]
22. Begin AS, Hidrue M, Lehrhoff S, Del Carmen MG, Armstrong K, Wasfy JH. Factors associated with physician tolerance of uncertainty: an observational study. J Gen Intern Med 2022 May;37(6):1415-1421 [FREE Full text] [doi: 10.1007/s11606-021-06776-8] [Medline: 33904030]
23. Strout TD, Hillen M, Gutheil C, Anderson E, Hutchinson R, Ward H, et al. Tolerance of uncertainty: a systematic review of health and healthcare-related outcomes. Patient Educ Couns 2018 Sep;101(9):1518-1537. [doi: 10.1016/j.pec.2018.03.030] [Medline: 29655876]
24. Cooke GP, Doust JA, Steele MC. A survey of resilience, burnout, and tolerance of uncertainty in Australian general practice registrars. BMC Med Educ 2013 Jan 07;13(1):2 [FREE Full text] [doi: 10.1186/1472-6920-13-2] [Medline: 23294479]
25. Klimecki O, Singer T. Empathic distress fatigue rather than compassion fatigue? Integrating findings from empathy research in psychology and social neuroscience. In: Oakley B, Knafo A, Madhavan G, Wilson DS, editors. Pathological Altruism. Oxford, UK: Oxford University Press; 2012:368-384.
26. Burns DD, Nolen-Hoeksema S. Therapeutic empathy and recovery from depression in cognitive-behavioral therapy: a structural equation model. J Consult Clin Psychol 1992 Jun;60(3):441-449. [doi: 10.1037//0022-006x.60.3.441] [Medline: 1619098]

XSL•FO
RenderX

27. Del Canale S, Louis DZ, Maio V, Wang X, Rossi G, Hojat M, et al. The relationship between physician empathy and disease complications: an empirical study of primary care physicians and their diabetic patients in Parma, Italy. Acad Med 2012 Sep;87(9):1243-1249. [doi: 10.1097/ACM.0b013e3182628fbf] [Medline: 22836852]
28. Hojat M, Louis DZ, Markham FW, Wender R, Rabinowitz C, Gonnella JS. Physicians' empathy and clinical outcomes for diabetic patients. Acad Med 2011 Mar;86(3):359-364. [doi: 10.1097/ACM.0b013e3182086fe1] [Medline: 21248604]
29. Pereira L, Figueiredo-Braga M, Carvalho IP. Preoperative anxiety in ambulatory surgery: the impact of an empathic patient-centered approach on psychological and clinical outcomes. Patient Educ Couns 2016 May;99(5):733-738. [doi: 10.1016/j.pec.2015.11.016] [Medline: 26654958]
30. Rakel DP, Hoeft TJ, Barrett BP, Chewning BA, Craig BM, Niu M. Practitioner empathy and the duration of the common cold. Fam Med 2009;41(7):494-501 [FREE Full text] [Medline: 19582635]
31. Mercer SW, Reynolds WJ. Empathy and quality of care. Br J Gen Pract 2002 Oct;52 Suppl(Suppl):S9-12 [FREE Full text] [Medline: 12389763]
32. Shanafelt TD, West C, Zhao X, Novotny P, Kolars J, Habermann T, et al. Relationship between increased personal well-being and enhanced empathy among. J Gen Intern Med 2005 Jul;20(7):559-564. [doi: 10.1007/S11606-005-0102-8]
33. West CP, Dyrbye LN, Erwin PJ, Shanafelt TD. Interventions to prevent and reduce physician burnout: a systematic review and meta-analysis. Lancet 2016 Nov 05;388(10057):2272-2281. [doi: 10.1016/S0140-6736(16)31279-X] [Medline: 27692469]
34. Ruotsalainen JH, Verbeek JH, Mariné A, Serra C. Preventing occupational stress in healthcare workers. Cochrane Database Syst Rev 2015 Apr 07;2015(4):CD002892 [FREE Full text] [doi: 10.1002/14651858.CD002892.pub5] [Medline: 25847433]
35. Luberto CM, Wasson RS, Kraemer KM, Sears RW, Hueber C, Cotton S. Feasibility, acceptability, and preliminary effectiveness of a 4-week mindfulness-based cognitive therapy protocol for hospital employees. Mindfulness (N Y) 2017 Dec;8(6):1522-1531 [FREE Full text] [doi: 10.1007/s12671-017-0718-x] [Medline: 29335671]
36. Cascales-Pérez ML, Ferrer-Cascales R, Fernández-Alcántara M, Cabañero-Martínez MJ. Effects of a mindfulness-based programme on the health- and work-related quality of life of healthcare professionals. Scand J Caring Sci 2021 Sep;35(3):881-891. [doi: 10.1111/scs.12905] [Medline: 32865258]
37. Goodman MJ, Schorling JB. A mindfulness course decreases burnout and improves well-being among healthcare providers. Int J Psychiatry Med 2012;43(2):119-128. [doi: 10.2190/PM.43.2.b] [Medline: 22849035]
38. Gracia Gozalo RM, Ferrer Tarrés JM, Ayora Ayora A, Alonso Herrero M, Amutio Kareaga A, Ferrer Roca R. Application of a mindfulness program among healthcare professionals in an intensive care unit: effect on burnout, empathy and self-compassion. Med Intensiva (Engl Ed) 2019 May;43(4):207-216. [doi: 10.1016/j.medin.2018.02.005] [Medline: 29544729]
39. Krasner MS, Epstein RM, Beckman H, Suchman AL, Chapman B, Mooney CJ, et al. Association of an educational program in mindful communication with burnout, empathy, and attitudes among primary care physicians. JAMA 2009 Sep 23;302(12):1284-1293. [doi: 10.1001/jama.2009.1384] [Medline: 19773563]
40. Martín-Asuero A, García-Banda G. The Mindfulness-based Stress Reduction program (MBSR) reduces stress-related psychological distress in healthcare professionals. Span J Psychol 2010 Nov 10;13(2):897-905. [doi: 10.1017/s1138741600002547] [Medline: 20977037]
41. Kabat-Zinn J. Full Catastrophe Living: Using the Wisdom of Your Body and Mind to Face Stress, Pain, and Illness. New York, NY: Bantam Books; 1991.
42. Lindsay EK, Creswell JD. Mechanisms of mindfulness training: Monitor and Acceptance Theory (MAT). Clin Psychol Rev 2017 Feb;51:48-59 [FREE Full text] [doi: 10.1016/j.cpr.2016.10.011] [Medline: 27835764]
43. Watson T, Walker O, Cann R, Varghese AK. The benefits of mindfulness in mental healthcare professionals. F1000Res 2021;10:1085 [FREE Full text] [doi: 10.12688/f1000research.73729.2] [Medline: 35387270]
44. Benzo RP, Anderson PM, Bronars C, Clark M. Mindfulness for healthcare providers: the role of non-reactivity in reducing stress. Explore (NY) 2018 Nov;14(6):453-456 [FREE Full text] [doi: 10.1016/j.explore.2018.03.008] [Medline: 30292600]
45. Wasson RS, Barratt C, O'Brien WH. Effects of mindfulness-based interventions on self-compassion in health care professionals: a meta-analysis. Mindfulness (N Y) 2020;11(8):1914-1934 [FREE Full text] [doi: 10.1007/s12671-020-01342-5] [Medline: 32421083]
46. Neff KD. Self-compassion: an alternative conceptualization of a healthy attitude toward oneself. Self Identity 2003 Apr;2(2):85-101. [doi: 10.1080/15298860309032]
47. Neff KD, Hsieh YP, Dejitterat K. Self-compassion, achievement goals, and coping with academic failure. Self Identity 2005 Jul;4(3):263-287. [doi: 10.1080/13576500444000317]
48. Sinclair S, Beamer K, Hack TF, McClement S, Raffin Bouchal S, Chochinov HM, et al. Sympathy, empathy, and compassion: a grounded theory study of palliative care patients' understandings, experiences, and preferences. Palliat Med 2017 May;31(5):437-447 [FREE Full text] [doi: 10.1177/0269216316663499] [Medline: 27535319]
49. Duarte J, Pinto-Gouveia J, Cruz B. Relationships between nurses' empathy, self-compassion and dimensions of professional quality of life: a cross-sectional study. Int J Nurs Stud 2016 Aug;60:1-11. [doi: 10.1016/j.ijnurstu.2016.02.015] [Medline: 27297364]

XSL•FO
RenderX

50. Kemper KJ, McClafferty H, Wilson PM, Serwint JR, Batra M, Mahan JD, Pediatric Resident Burnout-Resilience Study Consortium. Do mindfulness and self-compassion predict burnout in pediatric residents? Acad Med 2019 Jun;94(6):876-884. [doi: 10.1097/ACM.0000000000002546] [Medline: 30520809]
51. Montero-Marin J, Zubiaga F, Cereceda M, Piva Demarzo MM, Trenc P, Garcia-Campayo J. Burnout subtypes and absence of self-compassion in primary healthcare professionals: a cross-sectional study. PLoS One 2016;11(6):e0157499 [FREE Full text] [doi: 10.1371/journal.pone.0157499] [Medline: 27310426]
52. Olson K, Kemper KJ, Mahan JD. What factors promote resilience and protect against burnout in first-year pediatric and medicine-pediatric residents? J Evid Based Complementary Altern Med 2015 Jul 18;20(3):192-198. [doi: 10.1177/2156587214568894] [Medline: 25694128]
53. Babenko O, Mosewich AD, Lee A, Koppula S. Association of physicians' self-compassion with work engagement, exhaustion, and professional life satisfaction. Med Sci (Basel) 2019 Feb 12;7(2):29 [FREE Full text] [doi: 10.3390/medsci7020029] [Medline: 30759845]
54. Scheepers RA, Emke H, Epstein RM, Lombarts KM. The impact of mindfulness-based interventions on doctors' well-being and performance: a systematic review. Med Educ 2020 Feb;54(2):138-149 [FREE Full text] [doi: 10.1111/medu.14020] [Medline: 31868262]
55. Shapiro SL, Astin JA, Bishop SR, Cordova M. Mindfulness-based stress reduction for health care professionals: results from a randomized trial. Int J Stress Manag 2005 May;12(2):164-176. [doi: 10.1037/1072-5245.12.2.164]
56. Valley M, Stallones L. A thematic analysis of health care workers' adoption of mindfulness practices. Workplace Health Saf 2018 Nov;66(11):538-544. [doi: 10.1177/2165079918771991] [Medline: 29806801]
57. Swensen S, Kabcenell A, Shanafelt T. Physician-organization collaboration reduces physician burnout and promotes engagement: the mayo clinic experience. J Healthc Manag 2016;61(2):105-127. [Medline: 27111930]
58. Brewer J. Mindfulness training for addictions: has neuroscience revealed a brain hack by which awareness subverts the addictive process? Curr Opin Psychol 2019 Aug;28:198-203. [doi: 10.1016/j.copsyc.2019.01.014] [Medline: 30785066]
59. Brewer JA, Mallik S, Babuscio TA, Nich C, Johnson HE, Deleone CM, et al. Mindfulness training for smoking cessation: results from a randomized controlled trial. Drug Alcohol Depend 2011 Dec 01;119(1-2):72-80 [FREE Full text] [doi: 10.1016/j.drugalcdep.2011.05.027] [Medline: 21723049]
60. Mason AE, Jhaveri K, Cohn M, Brewer JA. Testing a mobile mindful eating intervention targeting craving-related eating: feasibility and proof of concept. J Behav Med 2018 Apr 16;41(2):160-173 [FREE Full text] [doi: 10.1007/s10865-017-9884-5] [Medline: 28918456]
61. Roy A, Hoge EA, Abrante P, Druker S, Liu T, Brewer JA. Clinical efficacy and psychological mechanisms of an app-based digital therapeutic for generalized anxiety disorder: randomized controlled trial. J Med Internet Res 2021 Dec 02;23(12):e26987 [FREE Full text] [doi: 10.2196/26987] [Medline: 34860673]
62. Gleichgerrcht E, Decety J. Empathy in clinical practice: how individual dispositions, gender, and experience moderate empathic concern, burnout, and emotional distress in physicians. PLoS One 2013 Apr 19;8(4):e61526 [FREE Full text] [doi: 10.1371/journal.pone.0061526] [Medline: 23620760]
63. Gleichgerrcht E, Decety J. The relationship between different facets of empathy, pain perception and compassion fatigue among physicians. Front Behav Neurosci 2014 Jul 11;8:243 [FREE Full text] [doi: 10.3389/fnbeh.2014.00243] [Medline: 25071495]
64. Neumann M, Edelhäuser F, Tauschel D, Fischer MR, Wirtz M, Woopen C, et al. Empathy decline and its reasons: a systematic review of studies with medical students and residents. Acad Med 2011 Aug;86(8):996-1009. [doi: 10.1097/ACM.0b013e318221e615] [Medline: 21670661]
65. Rowe JP, Shores LR, Mott BW, Lester JC. Integrating learning and engagement in narrative-centered learning environments. In: Proceedings of the 10th International Conference on Intelligent Tutoring Systems. 2010 Presented at: ITS '10; June 14-18, 2010; Pittsburgh, PA p. 166-177 URL: https://link.springer.com/chapter/10.1007/978-3-642-13437-1_17 [doi: 10.1007/978-3-642-13437-1_17]
66. West CP, Dyrbye LN, Sloan JA, Shanafelt TD. Single item measures of emotional exhaustion and depersonalization are useful for assessing burnout in medical professionals. J Gen Intern Med 2009 Dec 3;24(12):1318-1321 [FREE Full text] [doi: 10.1007/s11606-009-1129-z] [Medline: 19802645]
67. West CP, Dyrbye LN, Satele DV, Sloan JA, Shanafelt TD. Concurrent validity of single-item measures of emotional exhaustion and depersonalization in burnout assessment. J Gen Intern Med 2012 Nov 24;27(11):1445-1452 [FREE Full text] [doi: 10.1007/s11606-012-2015-7] [Medline: 22362127]
68. Maslach C, Jackson SE, Leiter MP. Maslach burnout inventory. In: Zalaquett CP, Wood RJ, editors. Evaluating Stress: A Book of Resources. New York, NY: Scarecrow Education; 1997:191-218.
69. Spitzer RL, Kroenke K, Williams JB, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
70. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
71. Carleton RN, Norton MA, Asmundson GJ. Fearing the unknown: a short version of the Intolerance of Uncertainty Scale. J Anxiety Disord 2007 Jan;21(1):105-117. [doi: 10.1016/j.janxdis.2006.03.014] [Medline: 16647833]

XSL•FO
RenderX

72. Raes F, Pommier E, Neff KD, Van Gucht D. Construction and factorial validation of a short form of the Self-Compassion Scale. Clin Psychol Psychother 2011 Jun 08;18(3):250-255. [doi: 10.1002/cpp.702] [Medline: 21584907]
73. Baer RA, Smith GT, Lykins E, Button D, Krietemeyer J, Sauer S, et al. Construct validity of the five facet mindfulness questionnaire in meditating and nonmeditating samples. Assessment 2008 Sep;15(3):329-342. [doi: 10.1177/1073191107313003] [Medline: 18310597]
74. Davis MH. Measuring individual differences in empathy: evidence for a multidimensional approach. J Pers Soc Psychol 1983 Jan;44(1):113-126. [doi: 10.1037/0022-3514.44.1.113]
75. Cohen J. A power primer. Psychol Bull 1992 Jul;112(1):155-159. [doi: 10.1037//0033-2909.112.1.155] [Medline: 19565683]
76. Roemer L, Salters K, Raffa SD, Orsillo SM. Fear and avoidance of internal experiences in GAD: preliminary tests of a conceptual model. Cogn Ther Res 2005 Feb;29(1):71-88. [doi: 10.1007/S10608-005-1650-2]
77. Cox RC, Olatunji BO. Sleep in the anxiety-related disorders: a meta-analysis of subjective and objective research. Sleep Med Rev 2020 Jun;51:101282. [doi: 10.1016/j.smrv.2020.101282] [Medline: 32109832]
78. Cisler JM, Olatunji BO, Feldner MT, Forsyth JP. Emotion regulation and the anxiety disorders: an integrative review. J Psychopathol Behav Assess 2010 Mar 16;32(1):68-82 [FREE Full text] [doi: 10.1007/s10862-009-9161-1] [Medline: 20622981]
79. Yu L, Buysse DJ, Germain A, Moul DE, Stover A, Dodds NE, et al. Development of short forms from the PROMIS™ sleep disturbance and sleep-related impairment item banks. Behav Sleep Med 2011 Dec 28;10(1):6-24 [FREE Full text] [doi: 10.1080/15402002.2012.636266] [Medline: 22250775]
80. Buysse DJ, Yu L, Moul DE, Germain A, Stover A, Dodds NE, et al. Development and validation of patient-reported outcome measures for sleep disturbance and sleep-related impairments. Sleep 2010 Jun;33(6):781-792 [FREE Full text] [doi: 10.1093/sleep/33.6.781] [Medline: 20550019]
81. Meyer TJ, Miller ML, Metzger RL, Borkovec TD. Development and validation of the Penn State worry questionnaire. Behav Res Ther 1990 Jan;28(6):487-495. [doi: 10.1016/0005-7967(90)90135-6] [Medline: 2076086]
82. Gratz KL, Roemer L. Multidimensional assessment of emotion regulation and dysregulation: development, factor structure, and initial validation of the difficulties in emotion regulation scale. J Psychopathol Behav Assess 2008 Oct 17;30(4):315. [doi: 10.1007/S10862-008-9102-4]
83. Carrau D, Janis JE. Physician burnout: solutions for individuals and organizations. Plast Reconstr Surg Glob Open 2021 Feb;9(2):e3418 [FREE Full text] [doi: 10.1097/GOX.0000000000003418] [Medline: 33680666]
84. Gao M, Roy A, Deluty A, Sharkey KM, Hoge EA, Liu T, et al. Targeting anxiety to improve sleep disturbance: a randomized clinical trial of app-based mindfulness training. Psychosom Med 2022 Jun 01;84(5):632-642 [FREE Full text] [doi: 10.1097/PSY.0000000000001083] [Medline: 35420589]
85. Eley DS, Bansal V, Leung J. Perfectionism as a mediator of psychological distress: implications for addressing underlying vulnerabilities to the mental health of medical students. Med Teach 2020 Nov;42(11):1301-1307. [doi: 10.1080/0142159X.2020.1805101] [Medline: 32805157]
86. Enns MW, Cox BJ, Sareen J, Freeman P. Adaptive and maladaptive perfectionism in medical students: a longitudinal investigation. Med Educ 2001 Nov 04;35(11):1034-1042. [doi: 10.1046/j.1365-2923.2001.01044.x] [Medline: 11703639]
87. Craiovan PM. Correlations between perfectionism, stress, psychopathological symptoms and burnout in the medical field. Procedia Soc Behav Sci 2014 Apr;127:529-533. [doi: 10.1016/j.sbspro.2014.03.304]
88. Mosewich AD, Crocker PR, Kowalski KC, Delongis A. Applying self-compassion in sport: an intervention with women athletes. J Sport Exerc Psychol 2013 Oct;35(5):514-524. [doi: 10.1123/jsep.35.5.514] [Medline: 24197719]
89. Delgado N, Delgado J, Betancort M, Bonache H, Harris L. What is the link between different components of empathy and burnout in healthcare professionals? A systematic review and meta-analysis. Psychol Res Behav Manag 2023;16:447-463 [FREE Full text] [doi: 10.2147/PRBM.S384247] [Medline: 36814637]
90. Boellinghaus I, Jones FW, Hutton J. The role of mindfulness and loving-kindness meditation in cultivating self-compassion and other-focused concern in health care professionals. Mindfulness 2012 Oct 21;5(2):129-138. [doi: 10.1007/S12671-012-0158-6]
91. Kuhn G, Goldberg R, Compton S. Tolerance for uncertainty, burnout, and satisfaction with the career of emergency medicine. Ann Emerg Med 2009 Jul;54(1):106-13.e6. [doi: 10.1016/j.annemergmed.2008.12.019] [Medline: 19201058]
92. Bovier PA, Perneger TV. Stress from uncertainty from graduation to retirement--a population-based study of Swiss physicians. J Gen Intern Med 2007 May 9;22(5):632-638 [FREE Full text] [doi: 10.1007/s11606-007-0159-7] [Medline: 17443371]
93. Englander R, Cameron T, Ballard AJ, Dodge J, Bull J, Aschenbrener CA. Toward a common taxonomy of competency domains for the health professions and competencies for physicians. Acad Med 2013 Aug;88(8):1088-1094. [doi: 10.1097/ACM.0b013e31829a3b2b] [Medline: 23807109]
94. Simpson JG, Furnace J, Crosby J, Cumming AD, Evans PA, Friedman Ben David M, et al. The Scottish doctor--learning outcomes for the medical undergraduate in Scotland: a foundation for competent and reflective practitioners. Med Teach 2002 Mar;24(2):136-143. [doi: 10.1080/01421590220120713] [Medline: 12098432]
95. Lally J, Cantillon P. Uncertainty and ambiguity and their association with psychological distress in medical students. Acad Psychiatry 2014 Jun 10;38(3):339-344. [doi: 10.1007/s40596-014-0100-4] [Medline: 24718768]

XSL•FO
RenderX

96. Simpkin AL, Schwartzstein RM. Tolerating uncertainty - the next medical revolution? N Engl J Med 2016 Nov 03;375(18):1713-1715. [doi: 10.1056/NEJMp1606402] [Medline: 27806221]
97. Allison JJ, Kiefe CI, Cook EF, Gerrity MS, Orav EJ, Centor R. The association of physician attitudes about uncertainty and risk taking with resource use in a Medicare HMO. Med Decis Making 1998;18(3):320-329. [doi: 10.1177/0272989X9801800310] [Medline: 9679997]
98. Ghosh AK. On the challenges of using evidence-based information: the role of clinical uncertainty. J Lab Clin Med 2004 Aug;144(2):60-64. [doi: 10.1016/j.lab.2004.05.013] [Medline: 15322499]
99. Benbassat J, Pilpel D, Schor R. Physicians' attitudes toward litigation and defensive practice: development of a scale. Behav Med 2001;27(2):52-60. [doi: 10.1080/08964280109595771] [Medline: 11763825]
100. Eisenberg N, Fabes RA, Murphy B, Karbon M, Maszk P, Smith M, et al. The relations of emotionality and regulation to dispositional and situational empathy-related responding. J Pers Soc Psychol 1994 Apr;66(4):776-797. [doi: 10.1037//0022-3514.66.4.776] [Medline: 8189352]
101. Ochsner KN, Gross JJ. The cognitive control of emotion. Trends Cogn Sci 2005 May;9(5):242-249. [doi: 10.1016/j.tics.2005.03.010] [Medline: 15866151]
102. Catuzzi JE, Beck KD. Anxiety vulnerability in women: a two-hit hypothesis. Exp Neurol 2014 Sep;259:75-80. [doi: 10.1016/j.expneurol.2014.01.023] [Medline: 24518489]
103. Hoff T, Lee DR. Burnout and physician gender: what do we know? Med Care 2021 Aug 01;59(8):711-720. [doi: 10.1097/MLR.0000000000001584] [Medline: 34081678]
104. Lyubarova R, Salman L, Rittenberg E. Gender differences in physician burnout: driving factors and potential solutions. Perm J 2023 Jun 15;27(2):130-136 [FREE Full text] [doi: 10.7812/TPP/23.023] [Medline: 37303223]
105. Epstein RM, Marshall F, Sanders M, Krasner MS. Effect of an intensive mindful practice workshop on patient-centered compassionate care, clinician well-being, work engagement, and teamwork. J Contin Educ Health Prof 2022 Jan 01;42(1):19-27. [doi: 10.1097/CEH.0000000000000379] [Medline: 34459443]

Abbreviations

MT: mindfulness training
PROMIS: Patient-Reported Outcomes Measurement Information System

Edited by A Mavragani; submitted 12.06.24; peer-reviewed by MS Krasner, R Epstein; comments to author 27.08.24; revised version received 15.09.24; accepted 04.10.24; published 24.01.25.

Please cite as:
Antico L, Brewer J
Digital Mindfulness Training for Burnout Reduction in Physicians: Clinician-Driven Approach
JMIR Form Res 2025;9:e63197
URL: https://formative.jmir.org/2025/1/e63197
doi:10.2196/63197
PMID:

©Lia Antico, Judson Brewer. Originally published in JMIR Formative Research (https://formative.jmir.org), 24.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Factors Influencing Telemedicine Adoption Among Health Care Professionals: Qualitative Interview Study

Fiona Schürmann[1*], PhD; Daniel Westmattelmann[1*], PhD; Gerhard Schewe[1], PhD

Center for Management, University of Münster, Münster, Germany

[*]these authors contributed equally

Corresponding Author:
Daniel Westmattelmann, PhD
Center for Management
University of Münster
Schlossplatz 3
Münster, 48149
Germany
Phone: 49 2518325010
Email: d.west@uni-muenster.de

Abstract

Background: Telemedicine is transforming health care by enabling remote diagnosis, consultation, and treatment. Despite rapid adoption during the COVID-19 pandemic, telemedicine uptake among health care professionals (HCPs) remains inconsistent due to perceived risks and lack of tailored policies. Existing studies focus on patient perspectives or general adoption factors, neglecting the complex interplay of contextual variables and trust constructs influencing HCPs' telemedicine adoption. This gap highlights the need for a framework integrating risks, benefits, and trust in telemedicine adoption, while addressing health care's unique dynamics.

Objective: This study aimed to adapt and extend the extended valence framework (EVF) to telemedicine, deconstructing factors driving adoption from an HCP perspective. Specifically, it investigated the nuanced roles of perceived risks, benefits, and trust referents (eg, technology, treatment, technology provider, and patient) in shaping behavioral intentions, while integrating contextual factors.

Methods: We used a qualitative research design involving semistructured interviews with 14 HCPs experienced in offering video consultations. The interview data were analyzed with deductive and inductive coding based on the EVF. Two coders conducted the coding process independently, achieving an intercoder reliability of 86.14%. The qualitative content analysis aimed to uncover the nuanced perspectives of HCPs, identifying key risk and benefit dimensions and trust referents relevant to telemedicine adoption.

Results: The study reveals the complex considerations HCPs have when adopting telemedicine. Perceived risks were multidimensional, including performance risks such as treatment limitations (mentioned by 7/14, 50% of the participants) and reliance on technical proficiency of patients (5/14, 36%), privacy risks related to data security (10/14, 71%), and time and financial risks associated with training (7/14, 50%) and equipment costs (4/14, 29%). Perceived benefits encompassed convenience through reduced travel time (5/14, 36%), improved care quality due to higher accessibility (8/14, 57%), and operational efficiency (7/14, 50%). Trust referents played a pivotal role; trust in technology was linked to functionality (6/14, 43%) and reliability (5/14, 36%), while trust in treatment depended on effective collaboration (9/14, 64%). Transparency emerged as a critical antecedent of trust across different referents, comprising disclosure, clarity, and accuracy. In addition, the study highlighted the importance of context-specific variables such as symptom characteristics (10/14, 71%) and prior professional experience with telemedicine (11/14, 79%).

Conclusions: This study expands the EVF for telemedicine, providing a framework integrating multidimensional risks, benefits, trust, and contextual factors. It advances theory by decomposing trust referents and transparency into actionable subdimensions and emphasizing context-specific variables. Practically, the findings guide stakeholders: policy makers should prioritize transparent regulations and data security, health care organizations should provide training and support for HCPs, and technology developers must design telemedicine solutions aligning with trust and usability needs. This understanding equips health care to address barriers, optimize adoption, and leverage telemedicine's potential for sustainable clinical integration.

XSL•FO
RenderX

(JMIR Form Res 2025;9:e54777) doi:10.2196/54777

KEYWORDS

trust referents; multidimensional risk; benefits; transparency; technology adoption; telemedicine; extended valence framework

Introduction

Background

Generally, a rise in technological innovations is omnipresent and particularly impacts the health care sector [1,2]. Nowadays, telemedicine can provide a wide range of services, which is attributed to the continuous evolution of innovations [3]. Remote patient care, diagnosis, and treatment enabled by leveraging information and communication technologies offer health care services in addition to traditional face-to-face encounters of health care professionals (HCPs) with patients [4]. The use of technology allows new opportunities to be exploited in a wide range of medical fields [5]. Among researchers, the added value, such as cost efficiency, is widely agreed upon. As an example, the percentage of global gross domestic product accounted for by health care costs is around 10% and rising [6]. Correspondingly, studies have shown that low-cost improvements in patients' health may be achieved by using remote health services for consultation, evaluation, or treatment [7,8]. Furthermore, the implementation of telemedicine allows for increased accessibility to health care. Thus, by eliminating the proximity factor, patients can access basic and particularly specialist care from their home, which allows instant assessment or treatment [7,9-11]. The relevance of the accessibility gained via telemedicine is highlighted in diverse respects. Especially in rural areas, a shortage of health care providers represents a challenge that can be bridged by telemedicine [7]. Considering the disparity in health care among nations, the increase in telemedicine adoption has the potential to provide great benefits not only to patients but also to medical staff worldwide [10].

Despite the multitude of presented advantages, the adoption rate of telemedicine remains low [2]. Regarding diffusion, the recent COVID-19 pandemic had a strong impact on the adoption rate of telemedicine. The emergence of COVID-19 in 2020 and the restrictions that were enacted resulted in far-reaching challenges worldwide [12,13]. Telemedicine provided a viable solution to many of those challenges by avoiding direct contact, reducing the danger of transmission, and ultimately ensuring continuous care [5,14]. Consequently, the circumstances enhanced the adoption of telemedicine, inducing a rapid upscale of diffusion in 2020. However, many HCPs, especially in countries lagging in technology, remain suspicious, resulting in low telemedicine adoption rates [15]. For example, in countries such as Germany (industrialized economy), India (resource-limited economy), and Brazil (emerging economy), telemedicine has been slower to establish itself than in other countries such as Israel (industrialized economy) [16]. These countries are examples of countries lagging in technology in telemedicine adoption due to barriers such as regulatory hurdles, insufficient digital infrastructure, skepticism among HCPs, and limited access to technology in rural areas, despite the potential benefits [16,17].

Extensive research has been conducted on relevant factors for telemedicine adoption [5,18,19]. In this regard, trust has been repeatedly highlighted as a key determinant of acceptance and use of telemedicine services [20,21]. Relatedly, perceived usefulness and ease of use have also been found to influence telemedicine adoption substantially [22]. The adoption of technologies in health care brings uncertainties, particularly related to data security and privacy [23,24]. Such uncertainties often lead to perceived risks that pose significant barriers to adoption intentions [25,26]. Factors such as the HCP-patient relationship, cultural aspects, and the digital divide also serve a critical role in the overall adoption of telemedicine [27-29]. Taken together, the literature on trust, perceived risk, perceived benefits, and other factors related to telemedicine adoption has been researched in a variety of settings and from differing perspectives. Kuen et al [30], for example, highlighted the multidimensionality of risks and trust transfer effects between several trust referents in a telemedicine context from the patient's perspective. However, it is still to be determined to what extent these constructs are also relevant from an HCP's perspective.

The information systems (IS) literature perspective is particularly relevant for telemedicine adoption as it explores how technology is effectively implemented, managed, and accepted within organizations. Because telemedicine heavily relies on digital systems and interactions, IS literature provides valuable insights into user acceptance and integration of these technologies in health care. Indeed, several systematic literature reviews have shown that technology adoption models such as the technology acceptance model (TAM) and the unified theory of acceptance and use of technology (UTAUT) are the dominant theoretical frameworks used to study telemedicine adoption among HCPs [31,32]. While these models offer important insights into factors such as perceived usefulness and ease of use, they are limited in their ability to account for the full complexity of telemedicine adoption in health care settings. Specifically, they fail to adequately address the critical role of perceived risks, which are especially relevant due to the sensitive nature of health care services [31]. Moreover, these models often assume that the decision-making process is primarily driven by functional considerations, such as ease of integration and productivity improvements [33]. However, HCPs operate in a more complex environment, where ethical concerns, patient safety, and trust in the technology and technology provider play significant roles [30,34]. Traditional models such as TAM and UTAUT also overlook the emotional and psychological dimensions of decision-making, including anxiety about new technology or concerns about the erosion of the HCP-patient relationship [35]. In addition, these models are generally ill-suited to capture the intricate professional and legal responsibilities faced by HCPs, which can directly influence their adoption behavior [35]. In the health care context, the stakes are higher, and the consequences of technology adoption

XSL•FO
RenderX

can directly impact patient outcomes, creating a level of risk that these models fail to fully incorporate [36].

Therefore, these frameworks do not offer the depth needed to explore the multilayered, context-specific factors that influence telemedicine adoption among HCPs. For example, while Bakshi and Tandon [37] explored how various dimensions of risk—such as financial, social, technological, and privacy risks—affect physicians' intentions to adopt telemedicine, their work does not fully contextualize these risks or examine the factors that drive these perceptions. Furthermore, some studies that identify key adoption factors in telemedicine do not use specific theoretical frameworks or examine the relationships between these factors [38]. This creates a gap in the literature, as the existing models and frameworks (eg, TAM and UTAUT) do not sufficiently incorporate the multidimensional risks that are crucial in the health care context [31].

This gap underscores the need for a more comprehensive framework. While technology adoption models provide one perspective, they do not account for the broader factors that influence decision-making in health care. As Hong et al [39] argue, theoretical models must be adapted to specific contexts to provide meaningful insights. In health care, the adoption of telemedicine goes beyond technical functionality and is shaped by professional ethics, regulatory environments, and context-specific risks. Therefore, a framework that integrates these elements is needed to better explain telemedicine adoption from an HCP's perspective. Therefore, we pose the following research question: "Which factors are relevant in the context of telemedicine adoption from an HCP's perspective, and how can these be incorporated into an overarching framework?"

Objectives

The research goal of this paper is to identify the factors influencing the adoption of telemedicine in terms of perceived risk, perceived benefits, and trust from an HCP's perspective. The aim is to derive a contextualized framework for the telemedicine context, which can serve as a guideline for future research. We draw on the extended valence framework (EVF) by Kim et al [40], which was initially developed for the consumer context but has demonstrated applicability to technology adoption in various domains (eg, mobile payment solutions [41] and remote anti-doping testing [42]), including consumer acceptance of online health information services [43]. We adapt it to the context-specific characteristics of telemedicine, which include the exchange of sensitive and personal data, situations with high uncertainty and dependency for patients, the significant responsibility placed on HCPs, and the adoption of new technology.

Theoretical Lens

To answer the research question, we draw on previous literature on adoption behavior and provide a brief introduction to the relevant definitions of the related constructs. We argue that the EVF is well suited as an overarching framework to explore relevant factors for technology adoption as it provides a broad perspective on individuals' attitudes and perceptions toward a new technology [40]. Unlike traditional models such as the TAM and the UTAUT, which primarily emphasize positive factors such as perceived usefulness, the EVF integrates both positive drivers (eg, perceived benefits) and negative drivers (eg, perceived risks), offering a more balanced understanding of adoption decisions. This is particularly critical in health care, where concerns over patient safety, data security, and ethical implications play a central role in decision-making. The EVF integrates perceived risks, perceived benefits, trust, and behavioral intentions, all of which are critical to the adoption of new technologies [44]. Figure 1 [40] depicts the layout of the EVF in which trust positively affects perceived benefits and behavioral intentions but negatively affects perceived risks. In addition, perceived risks have a negative relationship, and perceived benefits have a positive relationship with behavioral intentions.

Figure 1. Extended valence framework (EVF) for telemedicine adoption (this depiction of the EVF is based on the study by Kim et al [40]).

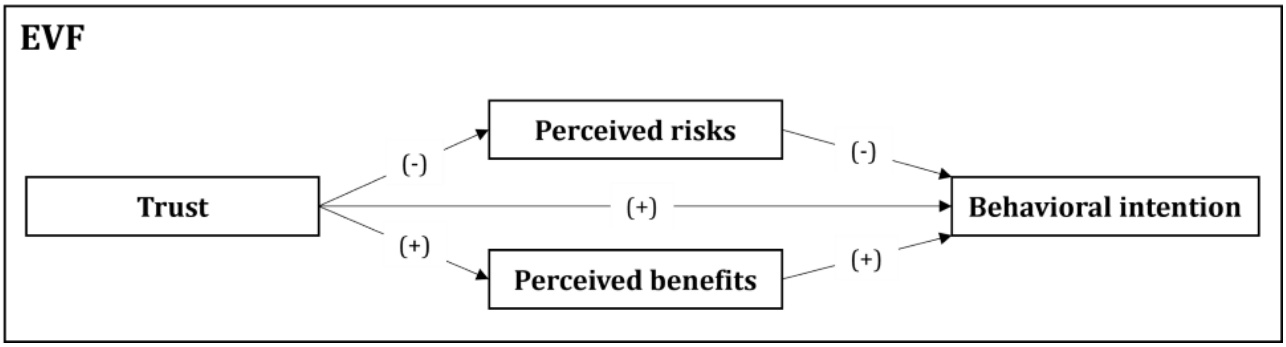

In the context of telemedicine, the EVF is particularly appropriate because it not only captures the benefits of new technology but also addresses the risks associated with it, such as misdiagnoses, technical reliability, and patient confidentiality, which are often underexplored in models such as TAM and UTAUT. Perceived risk refers to the uncertainty and potential severity of negative consequences, a definition established by Dowling and Staelin [45] as well as Featherman and Pavlou [26]. Various dimensions of risk have been identified in the literature, underscoring its multifaceted nature [46]. While the risk dimensions used in this study were initially developed in consumer e-services and purchasing contexts [26,47,48], research has extended their applicability to health care and telemedicine [30]. Table 1 provides an overview of the risks identified by previous research and their respective definitions.

XSL•FO
RenderX

Table 1. Definitions of risk dimensions relevant to telemedicine adoption.

Risk dimension	Definition	Study
Performance risk	Reflects the perception that a product or service fails to meet the anticipated performance standards, consequently not delivering the desired benefits	Grewal et al [47], 1994
Privacy risk	Reflects the "potential loss of control over personal information"	Featherman and Pavlou [26], 2003
Financial risk	Reflects the potential monetary outlay due to the product's or service's costs	Featherman and Pavlou [26], 2003
Psychological risk	Reflects the mental stress that possibly arises through the use of technology	Jacoby and Kaplan [48], 1972, and Featherman and Pavlou [26], 2003
Time risk	Reflects the potential loss of time when making a bad purchasing decision by wasting time researching and making the purchase, learning how to use a technology or service only to must replace it if it does not perform to expectations	Featherman and Pavlou [26], 2003
Social risk	Reflects the potential loss of status in one's social group because of adopting a product or service, looking foolish or untrendy	Featherman and Pavlou [26], 2003

Contrary to the multidimensional definition of perceived risk, perceived benefit represents the extent to which an individual believes that they will benefit from using a product or service [23]. It is characterized by the user's motivation to maximize the expected value associated with a technology [40]. Therefore, in this study, perceived benefit is defined as the potential value that a user associates with the use of telemedicine. As telemedicine represents a context that is different from other topics addressed in IS research, commonly used constructs cannot be easily transferred [49]. Previous research on telemedicine has shown, for example, the impact of efficiency on perceived benefits in health contexts such as autism and diabetes [7,8], as well as the relevance of the perceived quality of the specific service or product under consideration [50-52]. A study on the patient's perspective toward telemedicine adoption additionally highlights convenience as an impacting factor [53].

Within the EVF, trust is integrated as another relevant variable, as studies have shown that it reduces risk perceptions and can increase perceived benefits [54-56]. Thus, we include it in the scope of our contextual decomposition approach and define trust according to Mayer et al [57] as "the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the truster, irrespective of the ability to monitor or control that other party." Within this definition, a relationship is considered to exist with another identifiable party that acts and reacts with volition toward the truster. Both parties—the truster and the trustee—are not limited to individuals; they can also include other entities, both physical and non-physical, such as technology [58]. Interpersonal trust is considered to be a trusting relationship between individuals. According to Mayer et al [57], individuals' trust is determined by their perception of the trustworthiness of the other party. Trustworthiness is considered a multidimensional construct, the dimensions (ability, benevolence, and integrity) that influence trust in another individual, groups of individuals, or an organization and lead to a higher willingness to take risks in the respective relationship [57]. In addition to interpersonal trust, the need to distinguish trust in technology emerged. While trust in an individual entails a moral and conscious interpersonal dynamic, technology is an artificial creation with limited capabilities and no moral agency. Therefore, further research on trust developed a multidimensional construct of trusting beliefs in technology [55,59]. Table 2 shows a compromised overview of the relevant dimension definitions.

As this study aims to decompose the EVF, we look at several potential and context-specific trust referents. In research on trust in telemedicine, 4 main stakeholders have been identified (technology, treatment, patient, and provider) [60]. The concept of trustworthiness and trusting beliefs can be applied to patients and providers (trustworthiness) as well as technology (trusting beliefs). As treatment is understood to be a process, it does not fall under the 2 trust constructs. However, trust in treatment can be characterized as an individual's belief that the telemedical treatment effectively addresses medical concerns, is clear, and results from a collaborative decision-making process between the patient and HCP [30,61].

XSL•FO
RenderX

Table 2. Definitions of trust dimensions for interpersonal and technological contexts.

Trust construct and dimension	Definition	Study
Interpersonal trust/trustworthiness		Mayer et al [57], 1995
Ability	The skills, competencies, and attributes that enable a party to have an impact in a specific domain.	
Benevolence	The degree to which a trustee intends goodwill toward the truster, alongside their perceived positive attitude toward the truster.	
Integrity	The truster's perception that the trustee adheres to a set of principles that the truster finds acceptable.	
Trust in technology/trusting beliefs		McKnight et al [59], 2011
Functionality	The belief that a specific technology has the appropriate functionalities or features to fulfill the requirement.	
Reliability	The belief that the specific technology will consistently operate properly.	
Helpfulness	The belief that adequate and responsive help is provided by the technology.	

Methods

Study Design

To determine how perceived risks, perceived benefits, and trust are related in the specific context of telemedicine, we referred to the guidelines by Hong et al [39] on context-specific theorizing. Accordingly, we chose a qualitative research approach and conducted semistructured interviews with HCPs who already offer video consultations to their patients, which is considered a widely applied method to obtain contextual and authentic accounts [62]. HCPs represent the primary party in adopting telemedicine, as they must be willing to offer telemedicine to enable patients to actively opt for it. Hence, they have the opportunity to shape the broad adoption of telemedicine [63]. Given the limited adoption of this technology, interviewing HCPs with extensive experience in telemedicine is expected to provide greater value, as their use and familiarity with it are essential for overall social acceptance.

As part of the study's rigorous design, a pilot study was conducted before the actual data collection to ensure the clarity and relevance of the interview guide [64]. This pilot phase involved conducting preliminary interviews with 4 HCPs who had prior experience with telemedicine. On the basis of this feedback, the interview guide only had to be marginally adapted. The interview guide was mainly guided by the EVF (Figure 1), covering key constructs such as trust, perceived risks, perceived benefits, use behavior, and intention (Table S1 in Multimedia Appendix 1). It also included contextual factors to capture additional insights that may not be covered by the EVF. The guide consisted of open-ended questions designed to explore how these factors influence telemedicine adoption from the HCP's perspective. This structure allowed for flexibility, enabling follow-up questions to be asked during the interviews as needed to explore emerging themes.

On the basis of the EVF, a combination of deductively derived and inductively elaborated relevant constructs was chosen as the research approach [65]. Thus, in line with the study by Hong et al [39], the various constructs can be decomposed, the appropriateness of the EVF for the specific context can be examined, and all relevant constructs can then be transferred into a context-specific variation of the EVF. In addition, newly emerging antecedents and context-specific variables are added.

Data Collection

To ensure the interview partners were sufficiently qualified, only HCPs who (1) have completed formal education in the medical or therapeutic field, (2) have experience with telemedicine through video consultations, and (3) practice in Germany (to ensure cultural comparability) were contacted. All medical specializations were included. As psychotherapists represent about 25% of HCPs in Germany, this group was included in our analysis [66]. This selection deliberately focused on gaining insights from HCPs who have practical experience with telemedicine, enabling a deeper exploration of its real-world effectiveness and challenges. Involving only experienced users ensures that the study captures practical insights that are directly relevant to improving the future implementation of telemedicine. Obtaining consent from interviewees and assuring anonymization, we audio-recorded the interviews and then transcribed them verbatim [67,68]. We chose to recruit via the medical website Jameda [69]. A total of 142 HCPs who met the requirements as experts were contacted by email. Out of these 142 professionals, 26 (18.3%) expressed interest in participating. The interviews were conducted online via Zoom (Zoom Video Communications) during March and April 2023. After 14 interviews, theoretical saturation was reached, suggesting that further interviews would be unlikely to add any new insights [70].

Sample Description

The resulting interview sample included 14 experts, half of whom were male (7/14, 50%) or female (7/14, 50%) and were, on average, aged 47 (range 34-68) years. The length of the interviews varied between 13 and 53 (mean 27, SD 10) minutes. Table 3 gives an overview of the relevant descriptives.

XSL•FO
RenderX

Table 3. Characteristics of the 14 health care professionals interviewed.

Number	Age (y)	Highest education	Field of practice
1	43	Doctoral degree	Palliative care
2	49	High school diploma	Psychotherapy
3	55	High school diploma	Psychotherapy
4	48	Doctoral degree	Gynecology
5	68	Master's degree	Psychotherapy
6	45	Doctoral degree	Psychiatry and psychotherapy
7	52	Doctoral degree	General medicine
8	50	Doctoral degree	Endocrinology
9	36	Bachelor's degree	Physiotherapy
10	42	Doctoral degree	Gynecology
11	68	Doctoral degree	General medicine
12	37	State examination	Psychotherapy
13	35	Doctoral degree	General medicine
14	34	State examination	General medicine

Data Analysis

We conducted a qualitative content analysis [65] using MAXQDA 2022 (version 22.7.0) software. The analysis was conducted by 2 authors (coder A and coder B) separately and using deductive as well as inductive coding methods. The deductive codes were derived from the relevant constructs already outlined in the theoretical lens based on the EVF. First, all transcripts were coded by coder A according to these categories. In addition, coder A inductively derived subcategories to allow for more in-depth explanations and context-specific extension, which were then discussed with coder B. Following coding, coding rules, definitions, and anchor examples were developed to ensure consistency for each category. Coder B performed the coding process independently following coder A. Subsequently, the coding was iteratively extended and analyzed in several steps. Transparency, for example, was identified as an antecedent to trust and coded following the dimensions defined by Schnackenberg and Tomlinson [71] (disclosure, clarity, and accuracy). We adhered to the standards set by Venkatesh et al [72] to ensure the quality of our qualitative research. The validity of our research design was confirmed through a thorough account of our research context, which increased the relevance and trustworthiness of our conclusions. To ensure a valid coding process, we took steps to ensure inferential validity. The interauthor agreement in the coding process resulted in an intercoder reliability of 86.14%. To account for the potential of random agreement at the segment level, we assessed coding reliability and calculated Cohen κ, which yielded a value of 0.86. These results suggest an "almost perfect" agreement between the 2 coders [73].

Ethical Considerations

The study received ethics approval from the ethics committee of the School of Business and Economics at the University of Münster (2020–02). The research adhered to ethical guidelines for human subject research. All participants provided written informed consent before participating in the study. The consent process included an explanation of the study objectives, procedures, and the right to withdraw at any time without any consequences. To protect participants' privacy, all interview data were anonymized during transcription. Identifiable information was removed, and participants were assigned unique codes to ensure confidentiality. Data were securely stored on encrypted devices accessible only to the research team. No monetary or material compensation was provided to participants for their involvement in the study. This study does not include any images of participants or supplementary material that could potentially identify individuals. As a precaution, the interview transcripts were carefully reviewed to exclude any inadvertent references that could lead to participant identification.

Results

Outcome Evaluation

By decomposing the constructs of the EVF, our results demonstrated their individual sublevels and relevance to the context of telemedicine. We found that HCPs clearly perceived both the risks and benefits of telemedicine. They were able to reflect on and consider these extensively. In addition to the decomposed constructs of the EVF, we identified context-specific variables; transparency was an inductively derived factor that preceded trust. The following sections presents our result according to each construct of the EVF. Although certain inductive codes were mentioned infrequently, their inclusion is justified by the nature of qualitative research. Qualitative analysis values the depth of insights over frequency, meaning even low-frequency mentions can offer significant contextual relevance. These rare codes can highlight emerging or underexplored challenges within telemedicine adoption, particularly in niche areas of health care practice [74]. Specific quote examples for the inductively generated level 3 codings can be found in Tables S2-S6 in Multimedia Appendix 1.

XSL•FO
RenderX

Perceived Risks

Limited treatment options due to visual restrictions were mentioned with regard to *performance risk*; that is, no holistic insight into the patient can be gained. Furthermore, physical contact is omitted, and further examinations for diagnosis might be neglected. This results in the "...danger...that something is overlooked in the diagnosis" (interviewee 11). Internet connection issues were additionally noted, especially concerning patients in rural areas. An interviewee expressed the following:

> *...the great issue is the poor internet connection of the clients most of the time....* [Interviewee 4]

In addition, *financial risks* were recognized, particularly associated with acquiring high-technology equipment. While the general cost of equipment was not seen as a risk, the purchase of advanced, high-technology tools was perceived differently. An interviewee stated the following:

> *You could also control it remotely and make a much better zoom than with the webcam. Such a tool is expensive, of course.* [Interviewee 6]

Concerning *time risks*, the investment of time required for staff to familiarize themselves with technology and its functions was underlined. One respondent mentioned that "[Telemedicine] is indeed a considerable investment of time, as you have to train the staff as a whole" (interviewee 10), pointing out the significant time commitment necessary for staff training. As far as *psychological risk* is concerned, it was rarely addressed and, if so, was associated with an uncomfortable feeling and a certain weight in digital communication. Concerning *social risks*, it was inferred that negative comments are mainly received from the social and professional environment. However, *privacy risk* was most frequently addressed by the interviewees. The main focus was on the perceived risk posed by patient data security:

> *The basic question is always that of data security since patient data is exchanged.* [Interviewee 1]

Table 4 shows the deductive and inductive codes for perceived risks according to the number of interviews in which these were addressed.

XSL•FO
RenderX

Table 4. Results of the qualitative content analysis for perceived risks (N=14).

Levels 1[a], 2[b], and 3[c]	Interviews in which this aspect was mentioned, n (%)
Perceived risk	
Performance risk	
Restriction in treatment [d]	7 (50)
Restriction in patient-HCP [e] *connection*	7 (50)
Internet connection issues	5 (36)
Technical issues	4 (29)
Technical proficiency of patient	5 (36)
Financial risk	
Need for equipment	4 (29)
Need for staff	1 (7)
Insurance coverage	3 (21)
Limited user group	2 (14)
Inadequate compensation	1 (7)
Time risk	
Familiarization with the technology	7 (50)
Adaptation to the new workflow	2 (14)
Research for alternatives	1 (7)
Psychological risk	
Uncomfortable feeling	3 (21)
Weight of communication	3 (21)
External interfering factors	2 (14)
Social risk	
Negative comments	5 (36)
Privacy risk	
Data security	10 (71)
HCPs' privacy	2 (14)
Patients' privacy	4 (29)

[a]Category derived from the study by Kim et al [40].

[b]Categories derived from the study by Featherman and Pavlou [26].

[c]Categories are inductive codes based on specifications from respondents.

[d]Inductive codes are italicized.

[e]HCP: health care professional.

Perceived Benefits

Interviewees expressed multiple benefits related to the *convenience of telemedicine*. A notable advantage was the reduction in travel time, which benefited HCPs by allowing them to manage their schedules more efficiently and handle a higher volume of consultations. One participant stated that video consultation had the following benefits: "No more traveling. Rehabilitation clinics, in particular, are always a little further out of town as far as our field is concerned." (interviewee 9). Moreover, convenience was accentuated in treating patients located in rural areas. Another interviewee emphasized that these patients particularly value telemedicine as it allows them to receive consultations at home. In addition, interviewees commented on the medical implications of reduced travel, mentioning that longer travel durations "may also not be harmless due to the illness" (interviewee 6). They suggested that maintaining distance, especially if patients are contagious, is sensible for both patient and HCP. Apart from convenience, there was a shared sentiment that telemedicine augments the *quality of care*. Accessibility was often spotlighted as a distinct advantage of telemedicine, with interviewees noting its significant role in improving patient care. In addition to the accessibility of medical treatment and consultation, the possibility of maintaining a safe space was mentioned several

XSL•FO
RenderX

times, particularly in the psychological field, as many patients might feel more comfortable at home:

> *...the more distance there is between medical staff and patients,...it gives patients a low-barrier opportunity to mention things that they might not dare to mention face-to-face.* [Interviewee 9]

Consequently, HCPs obtained more relevant information, thus ensuring a higher quality of diagnosis and treatment. Furthermore, telemedicine was observed to enhance HCP's *efficiency*. The emphasis was largely on the significant time savings for HCPs, allowing them to provide timely care while easily coordinating and managing their schedules through health platforms. This streamlined process enables HCPs to handle more appointments and optimize their workflow, ultimately improving their productivity. Table 5 shows the deductive and inductive codes for perceived benefits according to the number of interviews in which these were addressed.

Table 5. Results of the qualitative content analysis for perceived benefits (N=14).

Levels 1[a], 2[b], and 3[c]	Interviews in which this aspect was mentioned, n (%)
Perceived benefit	
Convenience of telemedical use [d]	
Space independence	8 (57)
Reduction of travel time	5 (36)
Time flexibility	6 (43)
Protection of HCPs' [e] *well-being*	4 (29)
Protection of patients' well-being	6 (43)
Quality of care	
Accessibility	8 (57)
Safe space for sensitive topics	4 (29)
Enhanced perception	3 (21)
Efficiency of telemedical service	
Time saving	7 (50)
More precise working time	2 (14)
Financial advantage	4 (29)

[a]Category derived from the study by Kim et al [40].

[b]Categories derived from the study by Featherman and Pavlou [26].

[c]Categories are inductive codes based on specifications from respondents.

[d]Inductive codes are italicized.

[e]HCP: health care professional.

Trust Referents

Trust referents were distinguished by the HCPs, and all were acknowledged, although not all were considered to be of high importance. Trust in technology was typically mentioned in terms of its functionality, such as the ability to provide audio and visual connection, schedule appointments, and even collect some anamnesis data; for example,

> *[Take] a quick picture of the throat with a smartphone so that you can get an idea. Are the tonsils swollen now? Are they coated? That works.* [Interviewee 13]

With regard to *trust in the treatment process*, 3 subcategories emerged, which can be subdivided into the initial necessary consensus, the subsequent effective collaboration of patient and HCP, and the final effectiveness of the treatment. The patient-HCP connection was particularly relevant for HCPs:

> *Definitely, if you manage to maintain the personal factor further [using video consultation].* [Interviewee 9]

> *Trust in the provider was supported above all by the data security it offers, which also corresponds to the high perceived risk of data security:*

> *I trust my specialist provider, so to speak, and assume that the solution he offers has also been checked with the data protection authorities in terms of data protection law.* [Interviewee 2]

It also became clear that the provider is expected to have both technical competence and knowledge of medical procedures, that is, an understanding of the particular application context. Compared to the other trust referents, the HCPs' perception of *trust in the patient* was rather limited. It was primarily based on the digital affinity assigned to the patient, which is relevant for the HCP to be able to use telemedicine effectively together with the patient. Table 6 presents the deductive and inductive codes for all trust referents.

XSL•FO
RenderX

Table 6. Results of the qualitative content analysis for trust referents (N=14).

Levels 1[a], 2[b], and 3[c]	Interviews in which this aspect was mentioned, n (%)
Technology as a trust referent	
Technology's functionality	
Audio connection [d]	5 (36)
Visual connection	6 (43)
Appointment scheduling	1 (7)
Anamnesis data taken	3 (21)
Technology's reliability	
Stable connection	5 (36)
Stable platform	4 (29)
Treatment as a trust referent	
Consensus understanding of the treatment process	
Practicality	4 (29)
Enabling adequate dialogue	4 (29)
Effectiveness of collaboration	
Consent of patient	5 (36)
Enabling compliance	4 (29)
HCP [e]*-patient connection*	9 (64)
Patient feedback	2 (14)
Effectiveness of treatment	
Provision of medical reports	3 (21)
Work facilitation	4 (29)
Possible guidance to self-examination	2 (14)
Provider as a trust referent	
Provider ability	
Contextual knowledge	2 (14)
Technical knowledge	2 (14)
Provider benevolence	
Proximity	1 (7)
Provider integrity	
Data security	11 (79)
Responsiveness	3 (21)
Certification	3 (21)
Patient as a trust referent	
Patient ability	
Digital affinity	5 (36)

[a]Category derived from the studies by Van Velsen et al [60] and Kim et al [40].
[b]Categories derived from the studies by Mayer et al [57] and McKnight et al [59].
[c]Categories as inductive codes based on respondents' specifications.
[d]Inductive codes are italicized.
[e]HCP: health care professional.

XSL•FO
RenderX

Transparency

In the coding process, transparency emerged as an important antecedent to several trust referents for HCPs. This is in line with previous literature that identified transparency as an antecedent to trust, as information shared in this way indicates the ability, benevolence, and integrity of a truster [75,76]. The definition by Schnackenberg and Tomlinson [71] is widely accepted and defines transparency as "the perceived quality of intentionally shared information from a sender." We followed the accepted multidimensional approach and accordingly subdivided transparency into its 3 dimensions: disclosure, clarity, and accuracy [71,76,77]. Thereby, we define "disclosure" as the belief that all relevant information is shared in a timely manner, "clarity" as the level of correspondence between the intended and the understood meaning of information, and "accuracy" as the extent to which information given matches reality and is free from intentional distortions [71,76].

Transparency was emphasized in the information provided about technology, treatment, and providers. The patient as a trust referent was not mentioned when it came to transparency. In addition to transparency in relation to the provider, the relevance for disclosure on technology also became apparent: "Now I'm not a technician and can't assess what it could do with this data, but I feel well informed." (interviewee 3) Similarly, the importance of transparency in the treatment process between the patient and the HCP was highlighted: "When you feel like you've had a lot explained to you and you have a lot of knowledge about it and you know how things work" (interviewee 9). Table 7 shows the deductive codes for the transparency recipients according to the number of interviews in which these were addressed for each transparency dimension.

Table 7. Results of the qualitative content analysis for transparency dimensions (N=14).

Levels 1, 2[a], and3[b]	Interviews in which this aspect was mentioned, n (%)
Transparency	
Transparency on technology	
Disclosure	2 (14)
Clarity	1 (7)
Accuracy	3 (21)
Transparency on treatment	
Disclosure	4 (29)
Clarity	2 (14)
Accuracy	3 (21)
Transparency on provider	
Disclosure	7 (50)
Clarity	7 (50)
Accuracy	5 (36)

[a]Categories are in accordance to the trust referents based on the studies by Van Velsen et al [60] and Kuen et al [30].

[b]Categories derived from the study by Schnackenberg and Tomlinson [71].

Context-Specific Variables

Regarding the need for contextual specification mentioned by Hong et al [39], we were able to identify variables that apply specifically to the telemedicine context. Especially, the relevance of prior consideration of the symptom characteristics and their suitability for treatment with telemedicine services, such as video consultation, was highlighted:

> *I can take a lot of responsibility for doing this via telemedicine or not....And I have a bad feeling when I hear patients report that...the patient is being sealed off even more.* [Interviewee 12]

It is not so much the distinction between mental and physical that plays a role here, but rather the necessary examination proximity and physical treatment approach:

> *What is more difficult is to examine you....Let's say one would have to palpate it or something else, then it stops.* [Interviewee 1]

Another context-specific variable is the previous experience or familiarization of the HCP and their patients with telemedicine:

> *[Video consultation] has also now worked out quite well over the last two years.* [Interviewee 8]

Frequent positive experiences, therefore, increase both the willingness to adopt and the level of trust in the technology. Table 8 shows the inductively derived codes for context-specific variables.

XSL•FO
RenderX

Table 8. Results of the qualitative content analysis for context-specific variables affecting telemedicine adoption.

Levels 1 and 2[a]	Interviews in which this aspect was mentioned, n (%)
Context variable	
Symptom characteristics (10)	10 (71)
Experience (11)	11 (79)

[a]Categories derived inductively.

In summary, our proposed expansion of the EVF is shown in Figure 2 [26,30,40,60,71]. We consider transparency a key multidimensional antecedent, along with context-specific variables such as symptom characteristics and experience with telemedicine services.

Figure 2. Telemedicine-contextualized valence framework. The extended valence framework is based on the study by Kim et al [40]. Perceived risk categories follow the study by Featherman and Pavlou [26]. Trust referents are derived from the studies by Van Velsen et al [61] and Kuen et al [30], and transparency dimensions are based on the study by Schnackenberg and Tomlinson [71].

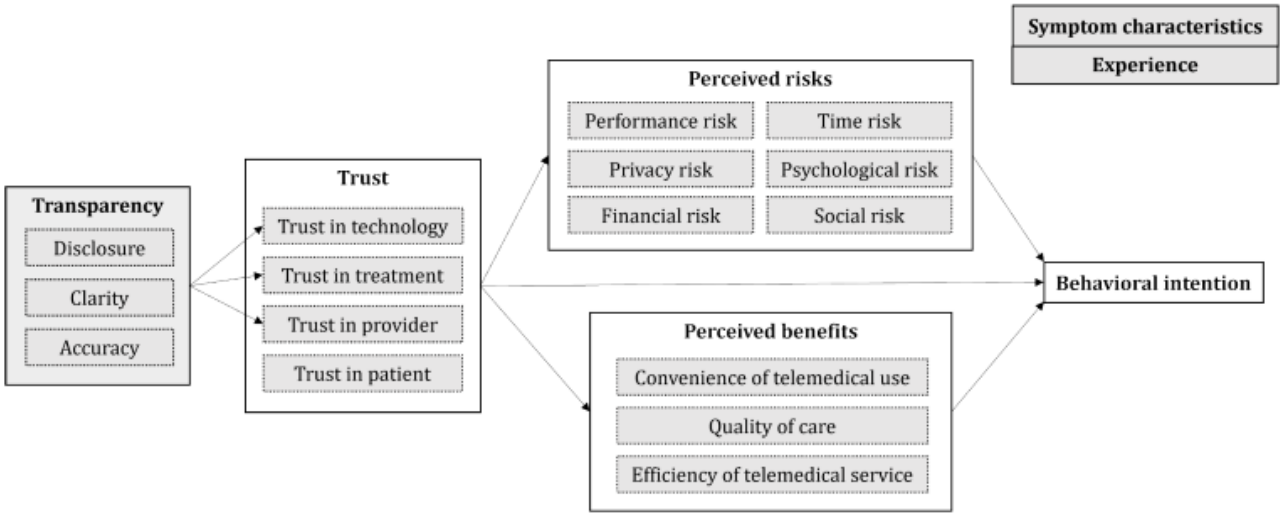

Discussion

Interpretation of Results

Our study provides a multifaceted examination of the considerations HCPs have when adopting and using telemedicine services, revealing various facets of perceived risks and benefits, trust referents, transparency, and context-specific variables.

First, it is evident that HCPs have various concerns about telemedicine. Concerns primarily relate to performance risks such as visibility limitations that restrict holistic insight into the patient's well-being. Lack of physical contact and potential confusion in diagnosis were highlighted by the HCPs, pointing to the critical role of traditional face-to-face interaction in medical assessments. Financial risks, such as the time required to learn to use technology, and privacy risks, such as concerns about data security, emphasize the challenges that HCPs face. Previous studies have identified technological issues, data security, and privacy as primary risks associated with the telemedicine context [30,78]. Consistent with earlier research, this study reaffirms the existence and multidimensionality of perceived risk [26,30], further delineating it into subdimensions of performance risks (eg, visual restrictions), privacy risks (eg, data security), and financial risks (eg, uncertainty about insurance coverage) as the risks most often mentioned. The addition of these context-specific subdimensions is a nuanced insight, adding depth to the understanding of practical challenges faced by HCPs.

Second, among the perceived benefits, the convenience of the telemedical approach, quality of care, and efficiency of the telemedical service were identified as key factors. Particular emphasis was placed on minimizing the HCP's travel time and improving the protection of the patient's well-being by reducing the need for transportation. This convenience, along with increased comfort and protection through minimal physical contact, establishes a better experience and better safety using telemedicine. In addition, our findings indicate that telemedicine can foster an environment that leads to increased efficiency of the services the HCP offers. Telemedicine was often complimented for its convenience, accessibility, and potential to deliver health care services to remote or underserved areas [7]. Increased patient engagement and satisfaction were also commonly cited benefits [10]. Our results align with existing literature, emphasizing convenience, quality, and efficiency as significant benefits and adding context-specific dimensions by highlighting the potential for streamlined operations and enhanced HCP efficiency in subdimensions.

Third, we found that trust is a key driver of adoption behavior, with multiple trust referents such as technology, treatment process, provider, and patient. Our findings illustrate that while technology and treatment process were strongly and diversely

XSL•FO
RenderX

perceived trust referents, trust in patients was solely measured by the patient's abilities in terms of their digital affinity. This underscores the important role that technical reliability and effective treatment collaboration have in successful telemedicine practices. These findings corroborate prior research, signifying trust as a multifaceted concept involving technology, treatment processes, and patients [30,60]. However, our results contradict the relevance of trust in patients from an HCP's perspective, as they represent a given parameter in the general treatment process for the HCP, and changing to an online setting seems to adapt the relationship only to the extent that patients must be able to use the new technology in question. Furthermore, the subdivisions of the treatment process need to be considered. Therefore, relevancy for trust is seen in particular in the second part, the consensus of collaboration. This is consistent with the assumption that the highest level of interaction and reliance on functioning communication between patient and HCP occurs here. Therefore, this study offers a more nuanced perspective by identifying the different relevancy of trust referents from the HCP's perspective.

Fourth, transparency emerged as an antecedent for trust, indicating the need for disclosure, clarity, and accuracy in technological processes and interactions with providers to foster a trustworthy telemedicine environment. This underscores the critical role of clear communication and information sharing. Transparency, specifically regarding data use and security, has been identified as crucial in previous research in differing contexts [79-81]. In addition, transparency is often linked to increased trust and acceptance of telemedicine services [60]. Notably, the need for transparency was identified for all trust referents except the patient. This could be explained by the personal view of the HCPs, in that they do not feel any dependency on their patients, as they perceive themselves as service providers on whom the patients depend. This highlights the special nature of the medical sector and the attitude and unilateral dependent relationship between HCPs and patients. Our results align with prior research, highlighting the importance of transparency as an antecedent for certain trust referents. This underscores the need to tailor the information provided according to the HCPs' relevant trust referents when fostering a trustworthy telemedicine environment. We also show that the view of transparency as multidimensional and the 3 defined dimensions (disclosure, clarity, and accuracy) [71] are transferable to the context under consideration and that, with a qualitative research approach, no major deviations in the relevancy of the individual dimensions are apparent.

Fifth, context-specific variables such as symptom characteristics and prior experience with telemedicine have a significant role in its use and efficacy. An HCP's intention to use telemedicine is influenced by various contextual factors such as the characteristics of symptoms, patient demographics, and technological familiarity among users [82-84]. The results from our interview study resonated with these findings, emphasizing the impact of symptom characteristics and previous telemedicine experiences on its use and perceived effectiveness. In addition, adoption behaviors may vary across medical specializations, depending on how telemedicine software interacts with clinical workflows and the patient-practitioner dynamic. For example, therapy with clinical psychologists might involve a more personal and sensitive experience compared to discussions of physical symptoms with HCPs, which may not require the same level of intimacy. The distinct benefits telemedicine can offer in fields such as infectious diseases further highlight the need to consider how different clinical contexts impact telemedicine adoption. Thereby, we specify which type of symptom characteristic is decisive for the usability of telemedical services such as video consultation. We abandon a division into mental and physical symptoms, an approach taken in prior research [30], as we found no significant difference between mental and physical symptoms. Instead, distinctions should be made according to the severity of the need for direct intervention. This is similar to a distinction of symptom severity, which has also already been introduced [85], but should not be interpreted as synonymous with it. The demonstrated impact of repetition, leading to increased experience and familiarity, is consistent with other studies, such as the study by Venkatesh et al [33], which highlight the importance of the repetition cycle in building trust [86].

Implication for Research

Several theoretical implications can be derived from the results of our study. Arguably, the first and most straightforward is the applicability and relevance of the guidelines by Hong et al [39] for context-specific theorizing. Our results show that a systematic decomposition of the relevant constructs of adoption behavior should address context-specific characteristics, especially because the context of telemedicine is characterized by a special relationship among the patient, HCP, and technology used for the treatment process and its provider.

Our second theoretical contribution is that this study addresses the limitations of existing technology adoption models such as TAM and UTAUT by incorporating telemedicine-specific factors into a contextualized framework. While TAM and UTAUT focus primarily on general constructs such as perceived usefulness and ease of use, they often fail to capture critical dimensions relevant to health care, such as the multifaceted risks and trust referents that are crucial for HCPs. The developed contextualized model overcomes these limitations by integrating additional factors such as transparency and symptom characteristics, offering a more comprehensive understanding of telemedicine adoption from an HCP perspective. This framework provides a valuable additional perspective to existing technology adoption models, serving as a foundation for future research on telemedicine adoption. Moving forward, future research can build on this by conducting quantitative validation and refining the framework further.

Third, our study emphasizes the multidimensionality of relevant factors in telemedicine, such as risks [26,30] and trust [59,60]. Recognizing the multidimensional nature of these factors enables a more nuanced understanding and specific contextualization for telemedicine. This perspective can guide future research to explore these dimensions in different health care settings and with different stakeholders.

Fourth, the developed framework offers starting points for further analysis of stakeholders in health care. For example, the perspectives of patients or technology providers, who have

XSL•FO
RenderX

already been identified by Van Velsen et al [60] as trust referents in the context of telemedicine adoption, may be examined in future studies. Understanding their perspectives and relevant context-specific factors can enrich the theoretical understanding of telemedicine adoption and its broader implications in health care.

Practical Implications

Several practical implications for HCPs and their use of telemedicine can be derived from our findings. First, the need for additional training of medical staff in telemedicine technologies can be addressed by ensuring that HCPs can focus on medical examinations and treatments without being distracted by technical issues. Therefore, health care facilities and general practices should consider hiring specialized telemedicine support staff. This approach enables a more efficient division of labor and ensures that technical problems do not affect the quality of medical care. In addition, specific training programs can be developed or used to help staff mitigate risks and fully understand and leverage the benefits of telemedicine. Training should thus include best practice approaches and encourage exchange with other medical facilities. In addition, telemedicine should be formally integrated into medical school curriculums and clinical placements to prepare future HCPs for the evolving landscape of digital health care. This will ensure that medical students are equipped with both the technical and clinical skills necessary for using telemedicine effectively in their practice. However, it is important to note that the cost and feasibility of these recommendations may vary, particularly for smaller or resource-constrained organizations. Ideally, telemedicine software should be intuitive enough for existing staff, such as IT or administrative personnel, to manage with minimal training, although this may not always be feasible.

Second, the individually identified risks that are particularly relevant can be addressed in a more targeted manner. To better manage such risks, risk management guidelines could be developed to provide guidance and protocols to counteract the perceived risks of telemedicine, for example, by reducing the performance risk through enhanced remote examination with improved imaging technologies or remote monitoring tools to mitigate visibility limitations and diagnostic confusion. Data security protocols can also be strengthened, thus minimizing the privacy risk, as insurance with compliance privacy regulations sends a trustworthy signal.

Third, transparency is a significant antecedent that underlines the necessity to choose an appropriate provider and the right technology for the planned medical approach. On the basis of our results, HCPs should choose telemedicine platforms and technologies that fit their most important trust factors, for example, those that offer reliable, user-friendly interfaces and robust security features. Thus, professionals should look for providers that offer comprehensive support and training. In addition, health care institutions should aid HCPs in selecting the appropriate telemedicine platform and technology by choosing their recommendations based on the criteria that emerged as particularly relevant in this study.

Fourth, as not every symptom characteristic can be treated equally well or at all with telemedicine, providers should focus on developing solutions that are specialized for certain medical conditions and symptoms and can optimally support these. By tailoring their technologies and programs to diverse symptom characteristics, HCPs may more effectively manage treatment via telemedicine. However, developing specific policies and guidelines on when it is appropriate to use telemedicine may be a more practical approach, given the constraints of resources and the limitations of digital communication. These guidelines can help ensure that telemedicine is used optimally without compromising patient care.

Limitations and Further Research

First, the amendments to the framework are based on a qualitative approach that aimed to identify relevant constructs for the considered context. Next, they should be tested quantitatively for the strength of the proposed causal relationships and applied to other telemedicine applications, such as telemonitoring or videoconferencing between HCPs that are not diagnostic or a treatment. We have added the new context-adapted subcategories to already established constructs, such as multiple risks. Therefore, future research should review and adapt existing scales to ensure that they cover the relevant new subcategories. However, the current framework does not fully address key human factors such as task load burden and the user experience of interacting with telemedicine software in daily clinical practice. These factors represent ongoing, structural impacts on health care delivery rather than the transitional costs of implementation. Although this was not a focus of this study, it is important to acknowledge this limitation and consider it as an avenue for future research.

Second, we adjusted the sample size to suit the rarity of the niche population (ie, HCPs in Germany offering video consultations). Due to the low adoption rate, HCPs with an affinity for technology are more likely to use telemedicine. Therefore, the benefits could be overestimated and the perceived risks underestimated. In addition, the use of volunteer sampling may introduce self-selection bias, as those who choose to participate are likely to hold more favorable and supportive views of telemedicine. This could further increase the risk of overestimating its benefits and underestimating its risks. Thus, it has to be argued that there is a possible lack of generalizability of our results, which should be addressed in future studies by considering larger samples and subject triangulation [87]. However, we intentionally only interviewed HCPs with experience, ensuring that the perceived benefits and risks were based on real experience rather than assumptions. Moreover, we ensured a diverse sample in terms of age, gender, and specialization to take a variety of perspectives into account. Future research should compare our findings with the perspectives of HCPs who have not adopted telemedicine as well as their transferability to a patient's perspective. Furthermore, the transferability of our results to the patient's perspective ought to be examined, given that previous studies showed a greater disparity between HCPs and patients [88]. This will provide deeper insights into the barriers and facilitators to further implementation of telemedicine.

Third, our sample is limited to the German context with its national regulations and funding opportunities. Cultural factors

XSL•FO
RenderX

specific to Germany, such as high individualism and low uncertainty avoidance [89], may also influence perceptions of telemedicine. These biases could differ significantly in other cultural environments, particularly in low-income countries, where telemedicine adoption might be shaped by different cultural attitudes and resource constraints. Telemedicine is already more established in other countries, which might differ in terms of regulatory and organizational conditions [90,91]. Our study shows the specificity of early adopters, who may still face more prejudices, insecurities, and lack of habits. The results of this study can be adapted to other cultural environments with more established telemedicine offerings and reveal differing cultural-specific results. To address the relevance of the governmental framework, future research should analyze how recent and anticipated changes in health policy and regulations affect the adoption and effectiveness of telemedicine and how this might affect the constructs proposed in this study's adoption framework. In addition, the demographics and communities served by telemedicine users should be considered, as populations less affected by the digital divide may perceive lower risks and greater benefits compared to the groups considered disadvantaged.

Conclusions

Context-specific consideration is useful and shows which additional factors from generally applicable frameworks or theories are particularly relevant. It also reveals which factors, if any, do not seem relevant in the context of telemedicine from an HCP's perspective. Our results show that multidimensional transparency as an antecedent to trust, as well as the context-specific metavariables (symptom characteristics and experience), impact telemedicine adoption. The complexity and higher risk perception associated with telemedicine also lead to a multitude of perceived risks. For future research, the results of this study imply that prospective studies in the field of telemedicine should refer to the design of a modified telemedicine-contextualized valence framework to conduct a fully comprehensive analysis. The role of a broader systematic and organizational context should also be considered to understand how external factors shape telemedicine adoption. Especially when using quantitative approaches, existing scales should address telemedicine-specific characteristics of constructs such as trust referents, risks, and benefits, as well as antecedents such as transparency in their items. If they do not, the scales should be adapted accordingly.

Acknowledgments

The authors acknowledge the use of ChatGPT for assisting with grammar and language checks during the final preparation of this manuscript. This study did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability

The datasets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

FS was responsible for conceptualization, methodology, writing, project administration, and formal analysis. DW was responsible for conceptualization, methodology, writing, and supervision.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview guide and coding schemes.
[DOCX File , 61 KB - formative_v9i1e54777_app1.docx]

References

1. Forkner-Dunn J. Internet-based patient self-care: the next generation of health care delivery. J Med Internet Res 2003;5(2):e8 [FREE Full text] [doi: 10.2196/jmir.5.2.e8] [Medline: 12857664]
2. Iyanna S, Kaur P, Ractham P, Talwar S, Najmul Islam A. Digital transformation of healthcare sector. What is impeding adoption and continued usage of technology-driven innovations by end-users? J Bus Res 2022 Dec;153:150-161. [doi: 10.1016/j.jbusres.2022.08.007]
3. Schreiweis B, Pobiruchin M, Strotbaum V, Suleder J, Wiesner M, Bergh B. Barriers and facilitators to the implementation of eHealth services: systematic literature analysis. J Med Internet Res 2019 Nov 22;21(11):e14197 [FREE Full text] [doi: 10.2196/14197] [Medline: 31755869]
4. Bashshur RL, Reardon TG, Shannon GW. Telemedicine: a new health care delivery system. Annu Rev Public Health 2000;21:613-637. [doi: 10.1146/annurev.publhealth.21.1.613] [Medline: 10884967]

XSL•FO
RenderX

5. Weißenfeld MM, Goetz K, Steinhäuser J. Facilitators and barriers for the implementation of telemedicine from a local government point of view - a cross-sectional survey in Germany. BMC Health Serv Res 2021 Sep 06;21(1):919 [FREE Full text] [doi: 10.1186/s12913-021-06929-9] [Medline: 34488753]
6. World development indicators. World Bank. URL: https://databank.worldbank.org/source/world-development-indicators [accessed 2023-11-06]
7. Harris MA, Freeman KA, Duke DC. Seeing is believing: using Skype to improve diabetes outcomes in youth. Diabetes Care 2015 Aug;38(8):1427-1434. [doi: 10.2337/dc14-2469] [Medline: 26033508]
8. Vismara LA, McCormick C, Young GS, Nadhan A, Monlux K. Preliminary findings of a telehealth approach to parent training in autism. J Autism Dev Disord 2013 Dec;43(12):2953-2969. [doi: 10.1007/s10803-013-1841-8] [Medline: 23677382]
9. Dorsey ER, Achey MA, Beck CA, Beran DB, Biglan KM, Boyd CM, et al. National randomized controlled trial of virtual house calls for people with Parkinson's disease: interest and barriers. Telemed J E Health 2016 Jul;22(7):590-598 [FREE Full text] [doi: 10.1089/tmj.2015.0191] [Medline: 26886406]
10. Scott Kruse C, Karem P, Shifflett K, Vegi L, Ravi K, Brooks M. Evaluating barriers to adopting telemedicine worldwide: a systematic review. J Telemed Telecare 2018 Jan;24(1):4-12 [FREE Full text] [doi: 10.1177/1357633X16674087] [Medline: 29320966]
11. Williams A, Larocca R, Chang T, Trinh NH, Fava M, Kvedar J, et al. Web-based depression screening and psychiatric consultation for college students: a feasibility and acceptability study. Int J Telemed Appl 2014;2014:580786 [FREE Full text] [doi: 10.1155/2014/580786] [Medline: 24799895]
12. Laato S, Islam AN, Farooq A, Dhir A. Unusual purchasing behavior during the early stages of the COVID-19 pandemic: the stimulus-organism-response approach. J Retail Consum Serv 2020 Nov;57:102224. [doi: 10.1016/j.jretconser.2020.102224]
13. Talwar S, Talwar M, Tarjanne V, Dhir A. Why retail investors traded equity during the pandemic? An application of artificial neural networks to examine behavioral biases. Psychol Mark 2021 Jul 26;38(11):2142-2163. [doi: 10.1002/mar.21550]
14. Monaghesh E, Hajizadeh A. The role of telehealth during COVID-19 outbreak: a systematic review based on current evidence. BMC Public Health 2020 Aug 01;20(1):1193 [FREE Full text] [doi: 10.1186/s12889-020-09301-4] [Medline: 32738884]
15. Webster P. Virtual health care in the era of COVID-19. Lancet 2020 Apr 11;395(10231):1180-1181 [FREE Full text] [doi: 10.1016/S0140-6736(20)30818-7] [Medline: 32278374]
16. Zeltzer D, Einav L, Balicer R. Healthcare delivery: what the rise of telemedicine means for you. World Economic Forum. URL: https://www.weforum.org/agenda/2021/07/telemedicine-covid-coronavirus-pandemic-healthcare/ [accessed 2024-09-28]
17. Bestsennyy O, Gilbert G, Harris A, Rost J. Telehealth: a quarter-trillion-dollar post-COVID-19 reality? McKinsey & Company. URL: https://www.mckinsey.com/industries/healthcare/our-insights/telehealth-a-quarter-trillion-dollar-post-covid-19-reality [accessed 2024-09-28]
18. Almathami HK, Win KT, Vlahu-Gjorgievska E. Barriers and facilitators that influence telemedicine-based, real-time, online consultation at patients' homes: systematic literature review. J Med Internet Res 2020 Feb 20;22(2):e16407 [FREE Full text] [doi: 10.2196/16407] [Medline: 32130131]
19. Dos Santos AD, Mata-Machado AT, Melo MD, Fonseca Sobrinho D, Araújo LL, Silva É, et al. Implementation of telehealth resources in primary care in Brazil and its association with quality of care. Telemed J E Health 2019 Oct;25(10):996-1004. [doi: 10.1089/tmj.2018.0166] [Medline: 30592699]
20. Katta M, Shao C, Herbey I, Modi S, Gleason L, Marques I, et al. IBD and the patient-provider relationship: does trust impact telemedicine utilization? Gastroenterology 2022 Feb;162(3):S83. [doi: 10.1053/j.gastro.2021.12.174]
21. Zheng S, Chang PY, Chen J, Chang YW, Fan HC. An investigation of patient decisions to use eHealth. J Organ End User Comput 2021;34(4):1-24. [doi: 10.4018/JOEUC.289433]
22. Harst L, Lantzsch H, Scheibe M. Theories predicting end-user acceptance of telemedicine use: systematic review. J Med Internet Res 2019 May 21;21(5):e13117 [FREE Full text] [doi: 10.2196/13117] [Medline: 31115340]
23. Kim DJ, Ferrin DL, Rao HR. A trust-based consumer decision-making model in electronic commerce: the role of trust, perceived risk, and their antecedents. Decis Support Syst 2008 Jan;44(2):544-564. [doi: 10.1016/j.dss.2007.07.001]
24. Wu TC, Ho CT. Barriers to telemedicine adoption during the COVID-19 pandemic in Taiwan: comparison of perceived risks by socioeconomic status correlates. Int J Environ Res Public Health 2023 Feb 16;20(4):23 [FREE Full text] [doi: 10.3390/ijerph20043504] [Medline: 36834205]
25. Jalali MS, Bruckes M, Westmattelmann D, Schewe G. Why employees (Still) click on phishing links: investigation in hospitals. J Med Internet Res 2020 Jan 23;22(1):e16775 [FREE Full text] [doi: 10.2196/16775] [Medline: 32012071]
26. Featherman MS, Pavlou PA. Predicting e-services adoption: a perceived risk facets perspective. Int J Hum Comput Stud 2003 Oct;59(4):451-474. [doi: 10.1016/S1071-5819(03)00111-3]
27. Alajlani M, Clarke M. Effect of culture on acceptance of telemedicine in Middle Eastern countries: case study of Jordan and Syria. Telemed J E Health 2013 Apr;19(4):305-311. [doi: 10.1089/tmj.2012.0106] [Medline: 23540280]

XSL•FO
RenderX

28. Luo J, Tong L, Crotty BH, Somai M, Taylor B, Osinski K, et al. Telemedicine adoption during the COVID-19 pandemic: gaps and inequalities. Appl Clin Inform 2021 Aug;12(4):836-844 [FREE Full text] [doi: 10.1055/s-0041-1733848] [Medline: 34496419]
29. Li Y, James L, McKibben J. Trust between physicians and patients in the e-health era. Technol Soc 2016 Aug;46:28-34. [doi: 10.1016/j.techsoc.2016.02.004]
30. Kuen L, Schürmann F, Westmattelmann D, Hartwig S, Tzafrir S, Schewe G. Trust transfer effects and associated risks in telemedicine adoption. Electron Mark 2023 Jul 21;33(1):6. [doi: 10.1007/S12525-023-00657-0]
31. Garavand A, Aslani N, Nadri H, Abedini S, Dehghan S. Acceptance of telemedicine technology among physicians: a systematic review. Inform Med Unlocked 2022;30:100943. [doi: 10.1016/j.imu.2022.100943]
32. Ghiwaa T, Khan I, White M, Beloff N. Telemedicine adoption for healthcare delivery: a systematic review. Int J Adv Comput Sci Appl 2023;14(11):1230-1243. [doi: 10.14569/IJACSA.2023.01411125]
33. Venkatesh V, Thong JY, Xu X. Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology. MIS Q 2012;36(1):157. [doi: 10.2307/41410412]
34. Nittari G, Khuman R, Baldoni S, Pallotta G, Battineni G, Sirignano A, et al. Telemedicine practice: review of the current ethical and legal challenges. Telemed J E Health 2020 Dec;26(12):1427-1437 [FREE Full text] [doi: 10.1089/tmj.2019.0158] [Medline: 32049608]
35. Shachak A, Kuziemsky C, Petersen C. Beyond TAM and UTAUT: future directions for HIT implementation research. J Biomed Inform 2019 Dec;100:103315 [FREE Full text] [doi: 10.1016/j.jbi.2019.103315] [Medline: 31629923]
36. Meinert E, Alturkistani A, Brindley D, Knight P, Wells G, de Pennington N. Weighing benefits and risks in aspects of security, privacy and adoption of technology in a value-based healthcare system. BMC Med Inform Decis Mak 2018 Nov 13;18(1):100 [FREE Full text] [doi: 10.1186/s12911-018-0700-0] [Medline: 30424753]
37. Bakshi S, Tandon U. Understanding barriers of telemedicine adoption: a study in North India. Syst Res Behav Sci 2021 Jan 04;39(1):128-142. [doi: 10.1002/sres.2774]
38. Ly BA, Labonté R, Bourgeault IL, Niang MN. The individual and contextual determinants of the use of telemedicine: a descriptive study of the perceptions of Senegal's physicians and telemedicine projects managers. PLoS One 2017;12(7):e0181070 [FREE Full text] [doi: 10.1371/journal.pone.0181070] [Medline: 28732028]
39. Hong W, Chan FK, Thong JY, Chasalow LC, Dhillon G. A framework and guidelines for context-specific theorizing in information systems research. Inf Syst Res 2014 Mar;25(1):111-136. [doi: 10.1287/isre.2013.0501]
40. Kim DJ, Ferrin DL, Rao HR. Trust and satisfaction, two stepping stones for successful e-commerce relationships: a longitudinal exploration. Inf Syst Res 2009 Jun;20(2):237-257. [doi: 10.1287/isre.1080.0188]
41. Chin AG, Harris MA, Brookshire R. An empirical investigation of intent to adopt mobile payment systems using a trust-based extended valence framework. Inf Syst Front 2020 Nov 03;24(1):329-347. [doi: 10.1007/S10796-020-10080-X]
42. Stoffers B, Lanfer J, Westmattelmann D, Schewe G. One app to test them all -opportunities and challenges of implementing a remote sampling system in anti-doping work. In: Proceedings of the 2023 International Conference on Information Systems. 2023 Presented at: ICIS '23; December 10-13, 2023; Hyderabad, India p. 1-17 URL: https://www.researchgate.net/publication/375029604_One_App_to_Test_Them_All_-Opportunities_and_Challenges_of_Implementing_a_Remote_Sampling_System_in_Anti-Doping_Work
43. Mou J, Shin DH, Cohen J. Health beliefs and the valence framework in health information seeking behaviors. Inf Technol People 2016 Nov 07;29(4):876-900. [doi: 10.1108/ITP-06-2015-0140]
44. Venkatesh V, Thong JY, Xu X. Unified theory of acceptance and use of technology: a synthesis and the road ahead. J Assoc Inf Syst 2016;17(5):328-376. [doi: 10.17705/1JAIS.00428]
45. Dowling GR, Staelin R. A model of perceived risk and intended risk-handling activity. J Consum Res 1994 Jun;21(1):119. [doi: 10.1086/209386]
46. Lim N. Consumers' perceived risk: sources versus consequences. Electron Commer Res Appl 2003 Sep;2(3):216-228. [doi: 10.1016/S1567-4223(03)00025-5]
47. Grewal D, Gotlieb J, Marmorstein H. The moderating effects of message framing and source credibility on the price-perceived risk relationship. J Consum Res 1994 Jun;21(1):145. [doi: 10.1086/209388]
48. Jacoby J, Kaplan LB. The components of perceived risk. In: Proceedings of the 3rd Annual Conference of the Association for Consumer Research. 1972 Presented at: ACR '72; November 3-5, 1972; Chicago, IL p. 382-393 URL: https://www.researchgate.net/publication/247814928_The_Components_Of_Perceived_Risk
49. Van Velsen L, Wentzel J, Van Gemert-Pijnen JE. Designing eHealth that matters via a multidisciplinary requirements development approach. JMIR Res Protoc 2013 Jun 24;2(1):e21 [FREE Full text] [doi: 10.2196/resprot.2547] [Medline: 23796508]
50. Harris K, Schultz J, Feldman R. Measuring consumer perceptions of quality differences among competing health benefit plans. J Health Econ 2002 Jan;21(1):1-17. [doi: 10.1016/s0167-6296(01)00098-4] [Medline: 11845919]
51. Pitic L, Brad S, Pitic D. Study on perceived quality and perceived fair price. Procedia Econ Financ 2014;15:1304-1309. [doi: 10.1016/S2212-5671(14)00592-9]
52. Shah A. The impact of quality on satisfaction, revenue, and cost as perceived by providers of higher education. J Mark High Educ 2009 Dec 16;19(2):125-141. [doi: 10.1080/08841240903451324]

XSL•FO
RenderX

53. Call VR, Erickson LD, Dailey NK, Hicken BL, Rupper R, Yorgason JB, et al. Attitudes toward telemedicine in urban, rural, and highly rural communities. Telemed J E Health 2015 Aug;21(8):644-651. [doi: 10.1089/tmj.2014.0125] [Medline: 25839334]
54. Kim DJ, Song YI, Braynov S, Rao H. A multidimensional trust formation model in B-to-C e-commerce: a conceptual framework and content analyses of academia/practitioner perspectives. Decis Support Syst 2005 Aug;40(2):143-165. [doi: 10.1016/j.dss.2004.01.006]
55. McKnight DH, Chervany NL. What trust means in e-commerce customer relationships: an interdisciplinary conceptual typology. Int J Electron Commer 2014 Dec 23;6(2):35-59. [doi: 10.1080/10864415.2001.11044235]
56. Massey PA, Montoya-Weiss M. Unraveling the temporal fabric of knowledge conversion: a model of media selection and use. MIS Q 2006;30(1):99-114. [doi: 10.2307/25148720]
57. Mayer RC, Davis JH, Schoorman FD. An integrative model of organizational trust. Acad Manag Rev 1995 Jul;20(3):709. [doi: 10.2307/258792]
58. Kelton K, Fleischmann KR, Wallace WA. Trust in digital information. J Am Soc Inf Sci 2007 Dec 04;59(3):363-374. [doi: 10.1002/asi.20722]
59. Mcknight DH, Carter M, Thatcher JB, Clay PF. Trust in a specific technology: an investigation of its components and measures. ACM Trans Manage Inf Syst 2011 Jul 01;2(2):1-25. [doi: 10.1145/1985347.1985353]
60. Van Velsen L, Wildevuur S, Flierman I, Van Schooten B, Tabak M, Hermens H. Trust in telemedicine portals for rehabilitation care: an exploratory focus group study with patients and healthcare professionals. BMC Med Inform Decis Mak 2016 Jan 27;16:11 [FREE Full text] [doi: 10.1186/s12911-016-0250-2] [Medline: 26818611]
61. van Velsen L, Tabak M, Hermens H. Measuring patient trust in telemedicine services: development of a survey instrument and its validation for an anticoagulation web-service. Int J Med Inform 2017 Jan;97:52-58. [doi: 10.1016/j.ijmedinf.2016.09.009] [Medline: 27919395]
62. Schultze U, Avital M. Designing interviews to generate rich data for information systems research. Inf Organ 2011 Jan;21(1):1-16. [doi: 10.1016/j.infoandorg.2010.11.001]
63. van Velsen L, Flierman I, Tabak M. The formation of patient trust and its transference to online health services: the case of a Dutch online patient portal for rehabilitation care. BMC Med Inform Decis Mak 2021 Jun 12;21(1):188 [FREE Full text] [doi: 10.1186/s12911-021-01552-4] [Medline: 34118919]
64. Thabane L, Ma J, Chu R, Cheng J, Ismaila A, Rios LP, et al. A tutorial on pilot studies: the what, why and how. BMC Med Res Methodol 2010 Jan 06;10:1 [FREE Full text] [doi: 10.1186/1471-2288-10-1] [Medline: 20053272]
65. Mayring P. Qualitative content analysis. In: Flick U, von Kardoff E, Steinke I, editors. A companion to qualitative research. Thousand Oaks, CA: Sage Publications; 2004:159-176.
66. Tolle M, Stratkötter A. Die Geschlechtszugehörigkeit von Therapeutinnen und Therapeuten in der psychotherapeutischen Arbeit: Ein integratives Modell. In: Breuer F, editor. Qualitative Psychologie: Grundlagen, Methoden und Anwendungen eines Forschungsstils. Cham, Switzerland: Springer; 2022:251-266.
67. Myers MD, Newman M. The qualitative interview in IS research: examining the craft. Inf Organ 2007 Jan;17(1):2-26. [doi: 10.1016/j.infoandorg.2006.11.001]
68. Sarker S, Xiao X, Beaulieu T. Guest editorial: qualitative studies in information systems: a critical review and some guiding principles. MIS Q 2013;37(4):iii-xviii [FREE Full text]
69. Jameda GmbH. URL: https://pro.jameda.de/ [accessed 2024-04-29]
70. Bowen GA. Naturalistic inquiry and the saturation concept: a research note. Qual Res 2008 Feb 01;8(1):137-152. [doi: 10.1177/1468794107085301]
71. Schnackenberg AK, Tomlinson EC. Organizational transparency. J Manag 2016 Jul 09;42(7):1784-1810. [doi: 10.1177/0149206314525202]
72. Venkatesh V, Brown SA, Bala H. Bridging the qualitative-quantitative divide: guidelines for conducting mixed methods research in information systems. MIS Q 2013 Jan 1;37(1):21-54. [doi: 10.25300/MISQ/2013/37.1.02]
73. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159. [doi: 10.2307/2529310]
74. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
75. Tomlinson EC, Schnackenberg A. The effects of transparency perceptions on trustworthiness perceptions and trust. J Trust Res 2022 Apr 18;12(1):1-23. [doi: 10.1080/21515581.2022.2060245]
76. Schnackenberg AK, Tomlinson E, Coen C. The dimensional structure of transparency: a construct validation of transparency as disclosure, clarity, and accuracy in organizations. Hum Relat 2020 Jun 27;74(10):1628-1660. [doi: 10.1177/0018726720933317]
77. Lee YW, Strong DM, Kahn BK, Wang RY. AIMQ: a methodology for information quality assessment. Inf Manag 2002 Dec;40(2):133-146. [doi: 10.1016/S0378-7206(02)00043-5]
78. Orrange S, Patel A, Mack WJ, Cassetta J. Patient satisfaction and trust in telemedicine during the COVID-19 pandemic: retrospective observational study. JMIR Hum Factors 2021 Apr 22;8(2):e28589 [FREE Full text] [doi: 10.2196/28589] [Medline: 33822736]

XSL•FO
RenderX

79. Sabani A. Investigating the influence of transparency on the adoption of e-Government in Indonesia. J Sci Technol Policy Manag 2020 Nov 23;12(2):236-255. [doi: 10.1108/jstpm-03-2020-0046]
80. Märtins J, Kehl R, Westmattelmann D, Schewe G. A multi-dimensional theory of transparency, trust, and risk in technology adoption. In: Proceedings of the 2022 Conference on International Conference on Information Systems. 2022 Presented at: ICIS '22; December 16-20, 2022; Copenhagen, Denmark p. 1-17 URL: https://www.researchgate.net/publication/365490186_A_Multi-Dimensional_Theory_of_Transparency_Trust_and_Risk_in_Technology_Adoption
81. Oldeweme A, Märtins J, Westmattelmann D, Schewe G. The role of transparency, trust, and social influence on uncertainty reduction in times of pandemics: empirical study on the adoption of COVID-19 tracing apps. J Med Internet Res 2021 Feb 08;23(2):e25893 [FREE Full text] [doi: 10.2196/25893] [Medline: 33465036]
82. Rahimi B, Nadri H, Lotfnezhad Afshar H, Timpka T. A systematic review of the technology acceptance model in health informatics. Appl Clin Inform 2018 Jul;9(3):604-634 [FREE Full text] [doi: 10.1055/s-0038-1668091] [Medline: 30112741]
83. Sintonen S, Immonen M. Telecare services for aging people: assessment of critical factors influencing the adoption intention. Comput Human Behav 2013 Jul;29(4):1307-1317. [doi: 10.1016/j.chb.2013.01.037]
84. Serrano KM, Mendes GH, Lizarelli FL, Ganga GM. Assessing the telemedicine acceptance for adults in Brazil. Int J Health Care Qual Assur (Forthcoming) 2020 Dec 24. [doi: 10.1108/IJHCQA-06-2020-0098] [Medline: 33369378]
85. Wu D, Gu H, Gu S, You H. Individual motivation and social influence: a study of telemedicine adoption in China based on social cognitive theory. Health Policy Technol 2021 Sep;10(3):100525. [doi: 10.1016/j.hlpt.2021.100525]
86. Bohnet I, Huck S. Repetition and reputation: implications for trust and trustworthiness in the short and in the long run. SSRN Journal. Preprint posted online February 6, 2004 [FREE Full text] [doi: 10.2139/ssrn.474544]
87. Patton MQ. Qualitative Evaluation and Research Methods. 2nd edition. Thousand Oaks, CA: Sage Publications; 2001.
88. Loriot A, Larceneux F, Guillard V, Bertocchio JP. Patients' representations of perceived distance and proximity to telehealth in France: qualitative study. J Med Internet Res 2024 Apr 26;26:e45702 [FREE Full text] [doi: 10.2196/45702] [Medline: 38669676]
89. Hofstede G. Culture's Consequences: Comparing Values, Behaviors, Institutions and Organizations Across Nations. Thousand Oaks, CA: Sage Publications; 2001.
90. Nittari G, Savva D, Tomassoni D, Tayebati SK, Amenta F. Telemedicine in the COVID-19 era: a narrative review based on current evidence. Int J Environ Res Public Health 2022 Apr 22;19(9):23 [FREE Full text] [doi: 10.3390/ijerph19095101] [Medline: 35564494]
91. Luciano E, Mahmood MA, Mansouri Rad P. Telemedicine adoption issues in the United States and Brazil: perception of healthcare professionals. Health Informatics J 2020 Dec;26(4):2344-2361. [doi: 10.1177/1460458220902957] [Medline: 32072843]

Abbreviations

EVF: extended valence framework
HCP: health care professional
IS: information systems
TAM: technology acceptance model
UTAUT: unified theory of acceptance and use of technology

Edited by A Mavragani; submitted 21.11.23; peer-reviewed by D Singh, D Hua, G Sebastian, JP Bertocchio; comments to author 08.08.24; revised version received 01.10.24; accepted 01.12.24; published 27.01.25.

Please cite as:
Schürmann F, Westmattelmann D, Schewe G
Factors Influencing Telemedicine Adoption Among Health Care Professionals: Qualitative Interview Study
JMIR Form Res 2025;9:e54777
URL: https://formative.jmir.org/2025/1/e54777
doi:10.2196/54777
PMID:

©Fiona Schürmann, Daniel Westmattelmann, Gerhard Schewe. Originally published in JMIR Formative Research (https://formative.jmir.org), 27.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Digital Tool (Technology-Assisted Problem Management Plus) for Lay Health Workers to Address Common Mental Health Disorders: Co-production and Usability Study in Pakistan

Maham Saleem[1,2], MSPH; Shamsa Zafar[3], Prof Dr Med; Thomas Klein[4], MSc; Markus Koesters[5], Prof Dr; Adnan Bashir[6], MCS; Daniela C Fuhr[1], Prof Dr; Siham Sikander[7], FCPS, PhD; Hajo Zeeb[1,2], Prof Dr Med

[1]Leibniz Institute for Prevention Research and Epidemiology – BIPS, Bremen, Germany
[2]Leibniz Science Campus Digital Public Health, Bremen, Germany
[3]Fazaia Medical College, Air University, Islamabad, Pakistan
[4]Department of Psychiatry and Psychotherapy II, District Hospital Guenzburg, Ulm University, Ulm, Germany
[5]Center for Evidence-Based Healthcare, Medical Faculty Carl Gustav Carus, Technische Universität Dresden, Dresden, Germany
[6]Health Information Systems Program, Islamabad, Pakistan
[7]Department of Primary Care and Mental Health, University of Liverpool, Liverpool, United Kingdom

Corresponding Author:
Maham Saleem, MSPH
Leibniz Institute for Prevention Research and Epidemiology – BIPS
Achterstraße 30
Bremen, 28359
Germany
Phone: 49 421 218 ext 56760
Email: saleem@leibniz-bips.de

Abstract

Background: Mental health remains among the top 10 leading causes of disease burden globally, and there is a significant treatment gap due to limited resources, stigma, limited accessibility, and low perceived need for treatment. Problem Management Plus, a World Health Organization–endorsed brief psychological intervention for mental health disorders, has been shown to be effective and cost-effective in various countries globally but faces implementation challenges, such as quality control in training, supervision, and delivery. While digital technologies to foster mental health care have the potential to close treatment gaps and address the issues of quality control, their development requires context-specific, interdisciplinary, and participatory approaches to enhance impact and acceptance.

Objective: We aimed to co-produce Technology-Assisted Problem Management Plus (TA-PM+) for "lady health workers" (LHWs; this is the terminology used by the Lady Health Worker Programme for lay health workers) to efficiently deliver sessions to women with symptoms of common mental health disorders within the community settings of Pakistan and conducted usability testing in community settings.

Methods: A 3-stage framework was used for co-producing and prototyping the intervention. Stage 1 (evidence review and stakeholder consultation) included 3 focus group discussions with 32 LHWs and 7 in-depth interviews with key stakeholders working in the health system or at the health policy level. Thematic analyses using the Capability, Opportunity, and Motivation for Behavioral Change (COM-B) model were conducted. Stage 2 included over eight online workshops, and a multidisciplinary intervention development group co-produced TA-PM+. Stage 3 (prototyping) involved 2 usability testing rounds. In round 1 conducted in laboratory settings, 6 LHWs participated in role plays and completed the 15-item mHealth Usability App Questionnaire (MUAQ) (score range 0-7). In round 2 conducted in community settings, trained LHWs delivered the intervention to 6 participants screened for depression and anxiety. Data were collected using the MUAQ completed by LHWs and the Patient Satisfaction Questionnaire (PSQ) (score range 0-46) completed by participants.

Results: Qualitative analysis indicated that a lack of digital skills among LHWs, high workload, resource scarcity for digitization (specifically internet bandwidth in the community), and need for comprehensive training were barriers for TA-PM+ implementation in the community through LHWs. Training, professional support, user guidance, an easy and automated interface, offline functionalities, incentives, and strong credibility among communities were perceived to enhance the capability, opportunity, and motivation of LHWs to implement TA-PM+. TA-PM+ was co-produced with features like an automated interface, a personal

XSL•FO
RenderX

dashboard, guidance videos, and a connected supervisory panel. The mean MUAQ score was 5.62 in round 1 of usability testing and improved to 5.96 after incorporating LHW feedback in round 2. The mean PSQ score for TA-PM+ was 40 in round 2.

Conclusions: Co-production of TA-PM+ for LHWs balanced context and evidence. The 3-stage iterative development approach resulted in high usability and acceptability of TA-PM+ for LHWs and participants.

(JMIR Form Res 2025;9:e59414) doi:10.2196/59414

KEYWORDS

digital mental health; digital technology; digital intervention; Problem Management Plus; lay health worker programme; common mental health disorders; low- and middle-income countries; co-production

Introduction

Background

Recent research has indicated that mental health remains among the top 10 leading causes of disease burden globally and constitutes 32.4% of years lived with disability [1]. Common mental health disorders are conditions that impact mood, thought processes, and behavior, often leading to challenges in daily functioning and overall well-being. These include disorders, such as depression, anxiety, and bipolar disorder, each characterized by unique symptoms and impacts on the lives of individuals [2]. Depression, affecting over 300 million people globally, is the most prevalent mental health disorder [3]. A major part of this burden (about 70%) is concentrated in low- and middle-income countries (LMICs). Additionally, a shortage of mental health specialists in LMICs contributes to a mental health treatment gap of 90% for almost all mental health conditions, meaning that only 1 out of 10 conditions requiring mental health treatment receive it from a mental health specialist [3]. Various factors contribute to the mental health treatment gap, including lack of human resources for mental health, stigma, accessibility, uneven mental health service distribution, and lack of a perceived need for treatment [4]. To address this gap, the World Health Organization (WHO) developed the Mental Health Gap Action Programme (mhGAP) by providing evidence-based interventions for 8 priority mental health disorders. These guidelines recommend involving lay health workers and community members in the early detection of mental health disorders and low-intensity psychological care of patients with mental health disorders. This can help LMICs address the treatment gap and health resource shortfall, and achieve WHO Mental Health Action Plan targets [3,4]. Building on the WHO's mhGAP, several initiatives in LMICs have been developed to address the mental health treatment gap [5-7]. These programs empower nonspecialist health workers to deliver mental health care through structured culturally adapted interventions. Examples include Problem Management Plus (PM+), a transdiagnostic intervention for managing stress and depression, and the Thinking Healthy Programme, which addresses perinatal depression through simplified behavioral strategies. Community-based programs like Friendship Bench in Zimbabwe further extend the reach of mhGAP by embedding mental health support within local settings, demonstrating significant improvements in access and patient outcomes. The Disease Control Priorities (DCP3) report endorses such scalable models as effective affordable solutions that bolster mental health care capacity in LMICs [6,7]. A comprehensive review [7] of 208 studies focusing on various mhGAP interventions in LMICs demonstrated their overall efficacy. The mhGAP interventions were shown to significantly improve the symptoms of common mental health disorders, reduce substance use, and positively impact parenting and child outcomes. Furthermore, mhGAP training across these interventions enhanced the clinical skills, knowledge, and competency of health workers, contributing to improved quality of care after training [7].

PM+ is a low-intensity brief psychological and transdiagnostic intervention for common mental health disorders. It is an evidence-based intervention that can address various mental health issues and address the limitations posed by scarce specialized treatments in low-resource settings [8-14]. PM+ has been designed for depression, anxiety, stress, grief, and self-identified practical issues (interpersonal conflict and unemployment) and is delivered by lay health workers [15]. The intervention consists of 5 weekly face-to-face sessions involving stress management and problem-solving approaches [15]. PM+ has been tested for effectiveness and cost-effectiveness in various LMICs [8-14]. It has proven to be particularly effective in reducing the symptoms of anxiety, depression, and general distress in settings affected by significant adversity, including conflict-affected regions or under-resourced settings [16]. PM+ is recommended for stepped-care models, especially in low-resource settings where it serves as an effective first-line intervention. In these contexts, PM+ addresses common mental health needs at a low cost, allowing higher-intensity services to be reserved for more severe cases. This approach optimizes resource use and enhances accessibility in areas with limited mental health infrastructure. Notably, PM+ is available in several languages including Urdu, and it has been contextualized and rigorously evaluated in Pakistan [10,12,13].

However, the implementation of PM+ can be challenging for lay health workers and can be hampered by issues of quality control in training, supervision, and delivery [8-11]. Lengthy sessions and the complex nature of systematic stress and patient problem management strategies further compound these challenges, along with the struggle to recall all core strategies and protocols [8-11]. This specific phenomenon is termed "program drift" in the field of implementation science, which refers to a situation where the quality and fidelity of an intervention diminish because of deviations from the manual protocol, resulting in a reduction in the potential benefits of the intervention [17]. These challenges and barriers can be overcome by using digital tools and training lay health workers for them [18]. A recent systematic review of studies from LMICs revealed

XSL•FO
RenderX

that the integration of digital technologies into mental health interventions led by nonspecialists typically results in positive outcomes and has the potential to address the following three gaps highlighted in mental health literature: (1) the treatment gap, (2) the treatment-quality gap by improving the competencies and knowledge of nonspecialists, and (3) the service utilization gap by increasing the use of mental health care services [18].

Few initiatives have pursued digital adaptations of PM+ to enhance accessibility and scalability, particularly in low-resource settings. The step-by-step intervention was initially conceptualized as an online self-help adaptation of PM+. While the step-by-step intervention initially incorporated the principles of PM+, the intervention diverged as the complexity of implementing problem-solving components within a digital app proved challenging. Consequently, the step-by-step intervention evolved into an independent transdiagnostic intervention focused on managing stress and addressing depression and anxiety through behavioral activation principles [19]. In response to the COVID-19 pandemic, additional digital training frameworks were developed to enable remote training of lay health workers, ensuring uninterrupted mental health support amid physical distancing measures [20].

PM+ has been rigorously evaluated in Pakistan, showing effectiveness in reducing depression and anxiety symptoms [10,12,13]. Despite higher initial costs than the standard care, it has demonstrated long-term cost-effectiveness and is recommended for integration into a stepped-care model in resource-limited settings like Pakistan [12]. In a stepped-care model [21], individuals begin with lower-intensity support and progress to more intensive treatments only if necessary, optimizing resources and tailoring care to individual needs. This approach is particularly relevant in Pakistan, where mental health services are limited and mostly concentrated in major urban centers [22]. Pakistan's health care system, with its community outreach through "lady health workers" (LHWs; this is the terminology used by the Lady Health Worker Programme for lay health workers) [23], who have established supervision and referral mechanisms, provides a strong foundation for implementing stepped-care mental health interventions like PM+. To address the challenges in implementing PM+ by applying technological assistance and to align with the recommended stepped-care model for Pakistan, we aimed to develop a technological platform Technology-Assisted Problem Management Plus (TA-PM+). TA-PM+ consists of an app designed to support nonspecialists in delivering the PM+ intervention within Pakistan's health system, with an established backend for supervision, monitoring, and referral processes. TA-PM+ seeks to ensure the consistent delivery of therapeutic components at the appropriate dosage and with high fidelity while being delivered face-to-face, fostering a therapeutic alliance, which is essential for maintaining the problem-management focus required for effective intervention delivery. This paper describes the stepwise co-production of TA-PM+ and its key features.

Objective

We aimed to co-produce TA-PM+ for LHWs to use in efficiently delivering sessions to women within community settings in Pakistan.

Methods

Setting and Context

The study was conducted in the Federal District of Islamabad Capital Territory (ICT), Pakistan. The study considered the Union Council (UC) of Tarlai, which has a population of 100,000 and is located about 10 kilometers outside the ICT. The UC is a predominantly low-income and lower middle–class area, with a significant number of migrants from different parts of Pakistan. The crude birth rate is estimated to be around 40 births per 1000 population, with families typically having 4-8 children. Traditional gender roles prevail in families. Among the poorer sections of the population, several adversities are common, including malnutrition, irregular income, and low literacy rates [24].

Each UC is served by the primary health care unit, which consists of a medical doctor, vaccinators, and a pharmacist, and serves as the base for the community outreach program of LHWs. The LHWs are supervised and monitored by a facility-based supervisor, known as the lady health supervisor (LHS). The LHWs provide family planning, and maternal and child preventive and promotive services to 250-300 households in the catchment area and establish referrals for high-risk cases. More recently, maternal and adolescent mental health awareness and prevention have been added to the curriculum. The UC of Tarlai has 32 LHWs who cover nearly 80% of the population of Tarlai.

Ethical Considerations

This study has been approved by the Institutional Review Board of Fazaia Medical College, Islamabad, Pakistan (approval number: IBD/FMC/1341/IRB), in accordance with the ethical guidelines for human subject research and the Declaration of Helsinki. Informed consent was obtained from all participants, and for secondary analyses, the original consent permitted data use without requiring additional permissions. All data were anonymized to ensure privacy and were securely stored on a protected cloud platform accessible only to the research team. Participants received modest compensation, such as financial allowances and travel reimbursements (US $5-7). No identifiable participant information has been included in the manuscript or supplementary materials.

TA-PM+ Co-production

In line with the 3-stage framework of co-production and prototyping of public health interventions [25], TA-PM+ was designed in three stages: (1) qualitative investigation, (2) co-production, and (3) prototyping and usability testing. Qualitative investigation used semistructured interviews and was designed and reported in line with the Consolidated Criteria for Reporting Qualitative Research (COREQ) checklist (Multimedia Appendix 1).

XSL•FO
RenderX

Stage 1: Qualitative Investigation

A qualitative study was conducted to inform the development of TA-PM+. The main objectives were to explore implementation barriers; understand user needs (including training and supervision), particularly those related to technology use assisting intervention delivery; evaluate monitoring; and identify TA-PM+ functionalities that could be used to strategically address barriers and aspects of delivering PM+ sessions. The sample for qualitative investigation consisted of LHWs, an LHS, and participants working in the health system or at the health policy level. Focus group discussions and in-depth interviews were conducted as part of qualitative investigation. All participants were given information about the research aims, and rapport was established.

Focus Group Discussions

A trained female researcher conducted 3 focus groups between June and August 2022 with the LHWs (n=32). They were purposefully sampled from a range of settings to ensure diversity in demographic factors, age, and education. The focus group discussions were held at a primary health care facility in Islamabad, Pakistan. One focus group discussion lasted between 70 and 85 minutes. The focus group discussions were held in Urdu, and data on mental health, digitization, and intervention functions perceived to be useful for TA-PM+ were elicited using semistructured techniques (Multimedia Appendix 2).

In-Depth Interviews

Interviews were conducted with a purposive sample of 7 participants working in the health system (n=2) or at the health policy level (n=5) between July and August 2022. The participants included an LHS, doctors at the primary health care unit of the health system, health system specialists, individuals in Director General programs, program officers of noncommunicable diseases, and digital health and information specialists at the policy level. The interviews were conducted in Urdu and face-to-face at a location of the participant's choice by a trained female researcher. The duration of the interviews ranged from 35 to 45 minutes, during which a semistructured interview guide was employed (Multimedia Appendix 2).

Data Analysis

The focus group discussions and interviews were audio recorded and transcribed verbatim. The framework analysis approach was used for analyses consisting of familiarization, identification of the thematic framework, charting, mapping, and interpretation. A second author (SZ) was engaged to independently code a randomly selected subset of 20% of the transcripts. The deductive framework was constructed using the Capability Opportunity and Motivation for Behavioral Change framework and the Theoretical Domain Framework (TDF) [26]. The TDF was implemented in deductive analyses to include various elements that otherwise might have been ignored. However, the researcher realized that not all constructs would be used to code the data, and constructs that did not yield any findings were excluded. The thematic framework was refined through repetitive discussions between the researchers MS and SZ. To ensure the authenticity of the data and to minimize the risk of data loss, the coding process was carried out on the transcript written in Urdu. Once indexing and charting were completed using MaxQDA, the coded sections of the transcript were double translated into English.

Stage 2: Co-production of TA-PM+

During stage 2, TA-PM+ was collaboratively developed by an intervention development group comprising both local (n=5) and international (n=5) members. This group included professionals with diverse backgrounds, including software developers, clinical psychiatrists, epidemiologists, public health researchers, and health system specialists.

Co-production Workshops

Eight co-production workshops were conducted online on Zoom between July and October 2022. Each co-production workshop lasted 120 to 150 minutes and was conducted with the aim of planning TA-PM+ features and delivery mechanisms. The findings from stage 1 were presented, ideas were sought from all members, and a storyboard was formulated. Subsequently, software designers presented concepts, prototypes, and avatar samples, and this was followed by the collection of feedback and iterative refinement, ultimately resulting in the finalization of features for TA-PM+. All the workshops were recorded, and after each workshop, a researcher (MS) summarized the findings decided during the workshop, which included technological features, scripts of the videos from the PM+ manual, and text for different interfaces of sessions. Additionally, member checking [27] (ie, sharing the findings with all members of the workshop) was carried out to verify the authenticity of the summarized findings. The finalized summary was shared with the software designers who prepared diagrams, vignettes, icon pictures, and interface mock-ups and was presented at the next workshop where feedback was sought again. This process was replicated across all 8 workshops, resulting in the finalization of all features, interface mock-ups, and videos for TA-PM+.

TA-PM+ Technological Platform Development

TA-PM+ was developed using the Agile methodology fostering flexibility and adaptability throughout the app's development and maintenance phases [28] from October 2022 till May 2023. The Agile methodology is an iterative approach to software development that prioritizes flexibility, customer collaboration, and continuous improvement. It emphasizes adaptive planning, incremental delivery, and embracing change. Agile teams work in short iterations to deliver working software increments, gather feedback, and adapt to evolving requirements. TA-PM+ is designed to operate on standard mobile devices, including iOS and Android platforms, underpinned by a robust hardware infrastructure. The mobile app and web dashboard are hosted on servers that necessitate a configuration with a minimum of 4 GB of RAM and 20 GB of storage capacity, ensuring the smooth operation of the app. MySQL is employed as the database management system on the server, providing a reliable platform for data management.

The software components of TA-PM+ are developed using advanced programming languages. PHP 7.3 and PHP 8.0 are leveraged for web development, while React Native is used for the mobile app and Laravel 8.40 is used for web development.

XSL•FO
RenderX

The TA-PM+ mobile app, developed using React Native, serves as the primary interface for users. The server-side component, developed using PHP (Laravel), involves a RESTful API and is responsible for authenticating users, conducting assessments, and facilitating data retrieval. The web dashboard, built using PHP (Laravel), is accessible via web browsers and is primarily intended for administrators. It provides a user-friendly interface for content management, configuration settings, and report viewing.

Stage 3: Prototyping and Usability Testing

Prototyping

After the co-production, external experts (n=2) in the PM+ intervention and in clinical psychiatry prototyped the content, resources, and videos of TA-PM+. The content was updated according to their feedback. The intervention development team examined the application for heuristic evaluation employing Jakob Nielsen heuristics [29].

Cultural Adaptation

We used the culturally adapted and Urdu-translated PM+ manual developed for the Pakistan context [10,12,13]. To further enhance cultural relevance, psychoeducational videos for patients were developed with detailed adjustments, including culturally appropriate avatars, dress codes, names, video backgrounds, and case scenarios. Voiceovers were provided by a native Urdu speaker, and the videos were reviewed by clinical psychiatrists (n=2) in Pakistan for language refinement. LHWs (n=6) participated in usability testing of TA-PM+, providing feedback on wording, icon labels, and flowcharts in the app. Additionally, women (n=6) from the community together with LHWs (n=6) evaluated the videos for cultural appropriateness and sensitivity. All feedback was incorporated to optimize clarity, usability, and cultural alignment.

Usability Testing

To evaluate the preliminary acceptance and utility of TA-PM+, usability testing sessions were carried out with a convenience sample of LHWs (n=6) in 2 rounds. Nelson [30] has established that a total of 5 participants is sufficient to detect usability issues.

Recruitment

Six LHWs were randomly recruited from the primary health care unit at Tarlai for usability testing. The LHWs had a comprehensive training program of 15 hours, which focused on their proficiency in using TA-PM+. The training agenda included introduction to common mental health disorders, confidentiality, PSYCHLOPS (Psychological Outcome Profiles), introduction to the PM+ intervention, content of the first session of PM+ on stress management, and delivery of the first session using TA-PM+. Subsequently, for round 2 of usability testing, recruitment was conducted in community settings, and the LHWs identified individuals exhibiting symptoms indicative of a common mental health problem from their catchment areas. A researcher screened these women for the symptoms of common mental health disorders. The following two screening tools were used (specific cutoffs were applied): (1) General Health Questionnaire-12 (GHQ-12): a score of 3 or higher on the GHQ-12, which is a 12-item questionnaire measuring general psychological distress on a 4-point scale (possible range of 0 to 36) and (2) WHO Disability Assessment Schedule (WHODAS): a score of 16 or higher on the WHODAS 2.0, which is a screener for functional impairment consisting of 12 items measured on a scale ranging from 1 to 5 (possible range of 12 to 60).

A total of 6 women from community settings were included for round 2 of usability testing.

Data Collection

The LHWs completed round 1 of usability testing through role plays following the completion of the training, using the mHealth App Usability Questionnaire (MUAQ) for standalone apps for health care providers [31]. The MUAQ comprises 18 items rated on a 7-point scale, with a higher score indicating greater usability. To assess the usability of TA-PM+, the total average score of the responses to all statements was calculated. With a maximum score of 7, a higher overall average score indicates superior app usability. Feedback was sought from LHWs, and TA-PM+ was improved with feedback gathered before round 2 of usability testing was conducted in the community with 6 participants. After recruiting participants, 6 trained LHWs delivered the first session of TA-PM+ to the positively screened individuals. After session delivery, the MUAQ [31] was administered to the LHWs, while the Patient Satisfaction Questionnaire (PSQ) involving an mHealth intervention [32] adapted for TA-PM+ was administered to the participants. The PSQ comprises a total of 14 items. On a 5-point Likert scale, respondents indicate the degree of agreement with each item. Ten items are presented in the positive, while 4 are stated in the negative. In the analysis, the negatively stated items are inverted; therefore, a higher score indicates increased satisfaction with the intervention. The highest possible score is 46. Another round of feedback was gathered from the LHWs and integrated into the final version of the TA-PM+ app.

Results

Overview

This section is divided into three parts describing the results of (1) qualitative investigation, (2) co-production, and (3) prototyping and evaluation.

Qualitative Investigation

Participant Characteristics

A total of 40 participants (mean age 49 years; range 40-60 years) were recruited, out of which 34 were female and 6 were male. Among the 40 participants, 30 completed 10 years of education, 2 completed 14 years of education, and 8 completed over 16 years of education.

Perception of Factors Influencing the Implementation of TA-PM+ and the Features of TA-PM+ to Address These Factors Within the Capability, Opportunity, and Motivation-Behavior Model

Figure 1 provides an overview of the health system's barriers to the implementation of TA-PM+ and the TA-PM+ features

XSL•FO
RenderX

to facilitate implementation that were mapped under the constructs of the TDF and the components of the Capability, Opportunity, and Motivation for Behavioral Change (COM-B) model. The code chart with the description of codes is provided in Multimedia Appendix 3.

Figure 1. Perceived barriers and facilitators for the implementation of Technology-Assisted Problem Management Plus (TA-PM+) and the corresponding digital facilitators of TA-PM+ identified to address the barriers. Thematic analysis conducted using the Capability, Opportunity, and Motivation for Behavioral Change (COM-B) framework. TDF: Theoretical Domain Framework.

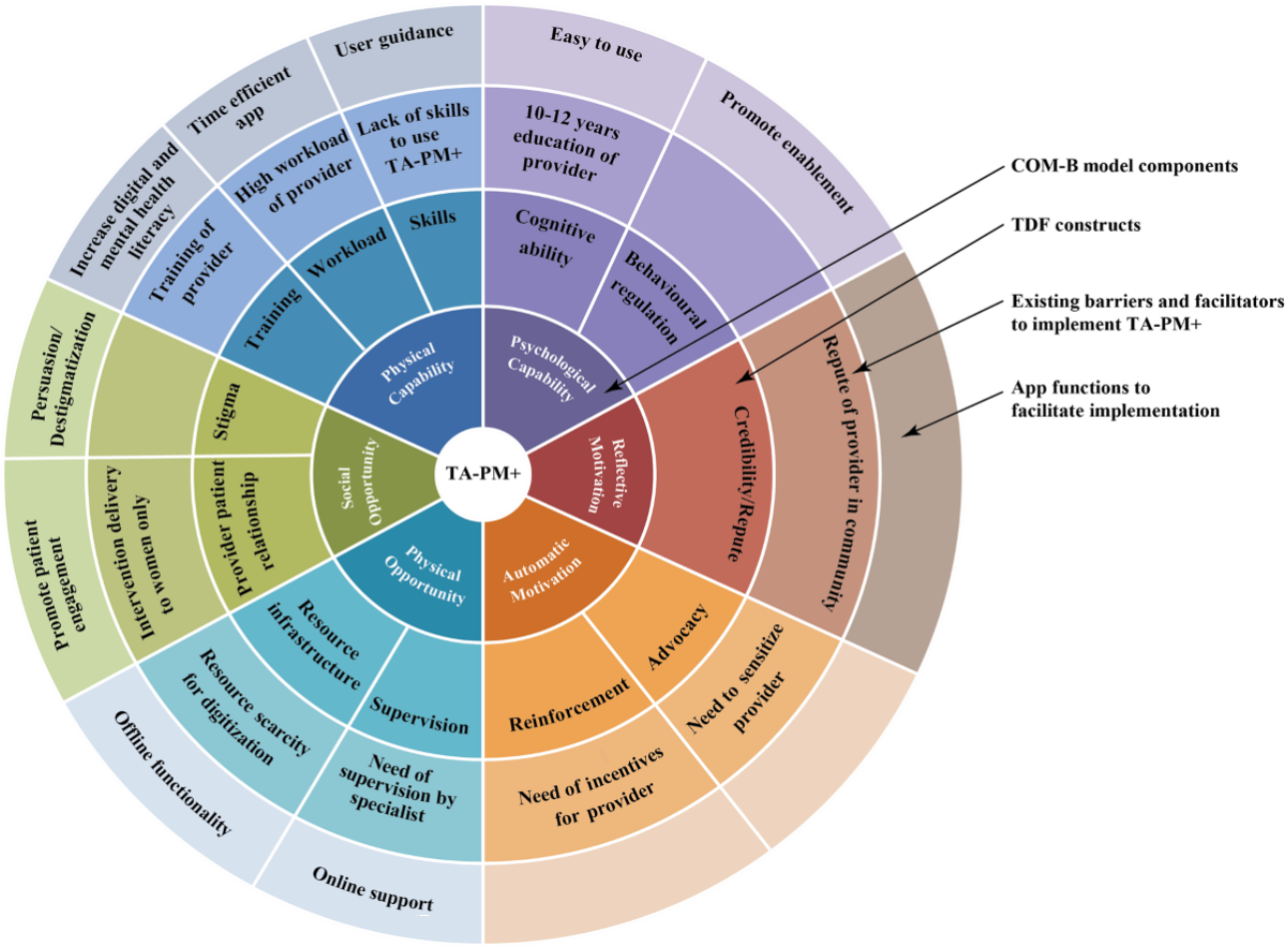

Barriers and Facilitators of the Health System Capacity to Implement TA-PM+ and the Features of TA-PM+ to Improve the Capacity Within the COM-B Framework

Knowledge and Skills

The significance of the knowledge and skill proficiency of LHWs in the effective implementation of TA-PM+ was highlighted by multiple participants. Although the digitization of the outpatient departments (OPDs) of LHWs has resulted in the development of some competencies to use digital tools, this aspect assumes utmost importance in the implementation process. Quotes have been provided, and the position of each quote in the transcript has been indicated.

> *There are some challenges for this intervention. Although the LHWs are given tablets, and their OPD is digitized…they receive proper digital training. But then again there might be problems regarding digital skills of LHWs to implement TA-PM+.* [Health policy participant_5, Position 10]
>
> *The challenges that can be faced include HR (Human Resources) and capacity issues.* [Health policy participant_3, Position 50]

Many participants perceived that in addition to training, the provision of user guidance and the incorporation of a fully automated guided app could effectively address the competency barrier.

> *When we open this app, it guides us from beginning till end. For example, when we are checking a child in a family, we open it all data comes and as we keep progressing, we are instructed what to do like its written 'check' and the name of the medicine, we click it, and it moves forward and tells us what to do next. This will make it easier for us.* [Focus group_1, LHW participant_6, Position 110]

Time/Workload of the Provider

All the participants agreed that a significant barrier to the implementation of TA-PM+ is the workload of LHWs. They were extensively involved in health promotion, disease prevention, and care delivery to women and children in their catchment area. The integration of a complex psychological intervention was perceived as an additional burden. Few respondents believed that the added workload associated with the additional digital tool complexity would outweigh the

XSL•FO
RenderX

assistance TA-PM+ is expected to provide in delivering the intervention.

> *We visit seven to eight homes daily, work on various aspects. We spend around 40 to 50 minutes in one household. In our app, which is dedicated to dengue, we send pictures. We also monitor the weight of children and assess the health of both mothers and children. When we encounter pregnant women, we provide them with cards and refer them accordingly. We give guidance, suggesting that they go to the center and obtain medication from there. In addition to all of this, we also record everything in our registers, and we enter data in the tablets.* [Focus group_3, LHW participant_1, Position 3]

The proposed design and strategies aimed at mitigating the strain experienced by LHWs were primarily focused on improving the time efficiency of TA-PM+. The participants strongly agreed on the importance of minimizing the amount of time spent using TA-PM+. Moreover, the participants emphasized the need for a simplified method that alleviates LHWs from the burden of typing. The ideal app, according to the participants, would be one that is fully automated and requires minimal input from LHWs. Almost all participants agreed to reduce session durations by leveraging the capabilities of the digital platform.

> *First, it should let them do everything with just a click. Why not create a form in such a way that these poor women don't need to type? So, I would suggest that you consider developing an app that is quick for them to use. The reason behind this suggestion is that these women are dealing with dengue, pneumonia, polio, health education, breast feeding, breast cancer, family planning, and post-abortion matters. So, think about how it could benefit them.* [Health system participant_6, Position 100]

Training

The importance of LHW training was acknowledged by the participants as a critical component of successful implementation. They emphasized the significance of comprehensive training covering the PM+ intervention as well as training to use TA-PM+. Some participants also highlighted the importance of field usability testing prior to implementation.

> *They should be guided on how to use it properly, and proper training and facilitation will be required. Train them for problem management plus and for the digital App. I think they shouldn't encounter any problems, but essentially, it's about sensitizing them first, then explaining the application thoroughly and conducting proper training. And, useful to conduct field test, then it should work well.* [Health policy participant_5, Position 16]

Regarding TA-PM+ functionalities, the participants felt that it should improve the TA-PM+ digital literacy of LHWs. Suggestions included the use of infographics or the provision of TA-PM+ introduction videos to serve as a refresher training for LHWs. This functionality was thought to be a helpful tool that LHWs could access at any time to address challenges in using the app.

> *Infographics should be designed in such a way that the LHWs can easily understand and get refresher training. For example, the LHW goes into the community, and sometimes someone asks her how to make ORS (Oral Rehydration Solution), and she doesn't remember how to make ORS. So, if she opens her learning platform and she can see how to make it. If there's a small infographic video uploaded, it will make it easier for community workers. Sometimes they may not remember everything, but with such resources, they can easily learn.* [Health policy participant_3, Position 56]

Cognitive Capacity

The cognitive capacity, including memory and attention, of LHWs has been identified as a crucial factor in the successful implementation of the TA-PM+ intervention. The majority of LHWs have 10 years of formal education. After training, LHWs need to remember and retain the information provided, pay attention to the needs and concerns of the individuals receiving the intervention, and actively engage with TA-PM+ to effectively deliver the intended support. Given the complexity of TA-PM+ and the cognitive demands associated with delivering it through a digital platform, it is crucial to develop an easy-to-use app that is not overly complex and employs plain language.

> *These poor women are not IT experts, they are all over 40, and while it seems easy, at this age, learning something new is challenging for them because they are accustomed to their old ways. So, I would suggest that you provide them with an app that is easy to download and use.* [Health system participant_6, Position 92]

Behavioral Regulation

One of the most important functionalities of TA-PM+ highlighted by LHWs was to promote enablement, which refers to the app's capacity to enhance precision in intervention delivery. To ensure that the intervention is implemented successfully, the TA-PM+ design must be optimized to promote and support the targeted actions.

> *At first, the name should appear, then their age, followed by their phone number, and then some other contact information. Also, session information like you've started a session, an entry has been made for that session or our visit data is recorded in it. The next session and date should be visible and whenever we click on patient name, the next session will come up.* [Focus group_3, LHW participant_6, Position 145]

XSL•FO
RenderX

Barriers and Facilitators for the Opportunity in Health Systems to Implement TA-PM+ and the Features of TA-PM+ to Increase the Opportunity Within the COM-B Framework

Resources/Infrastructure

The effective implementation of TA-PM+ relies on the availability of sufficient infrastructure and resources. The ongoing health reforms and digitization of health care in Pakistan provide a favorable environment for the implementation of a digital mental health intervention. However, the digitization of primary health systems is still in its early stages, and LHWs face numerous challenges. Despite the availability of digital tablets and the use of digital applications by LHWs for service delivery, concerns remain regarding device insurance, internet availability in different areas, and limited funds for internet access.

> *When you go to the community, sometimes it happens that you don't have access to the internet, and we cannot submit data. This internet, which is making life easy for people, has become a burden for us. It has become a source of trouble.* [Focus group_1, LHW participant_8, Position 254]

One of the key features considered important by all participants was offline functionality for TA-PM+. It was unanimously agreed that TA-PM+ should not rely on internet connectivity for its operation.

Monitoring, Supervision, and Support

The provision of supervision and continuous monitoring of LHWs was seen as crucial to the effective delivery of TA-PM+ to mental health patients. Furthermore, it is of utmost importance to ensure adequate support in case of any issues arising with the app.

The current model, which involves the supervision of LHWs by LHSs and referral from the community to primary health care facilities as needed, has been identified as a favorable approach for implementation. For TA-PM+ functionality, the online monitoring and referral features should be included to enhance the supervision process and ensure efficient communication between LHWs and health care facilities.

> *There should be a referral system in place where there is someone available who can assess if your patient is not benefiting from sessions and if their condition is severe, they should be referred through app. There should be a doctor available at the center, and there should be a clear referral process. It should also be made convenient so that if someone comes for help, it is beneficial for them, and there is no embarrassment for us.* [Focus group_3, LHW participant_2, Position 276]

Patient and Provider Relationship

TA-PM+ will be delivered by LHWs, and all participants perceived that the delivery can only be planned for women residing in the areas. Delivering the intervention to men is not possible given the focus of the LHW program on women's health. Furthermore, it was also perceived as a facilitator to implementation because LHWs have a positive reputation among the women of the community and the entire household has trust in them. This will facilitate the identification of cases of depression and stress by the LHW, as she will be fully informed of the circumstances a woman is experiencing.

> *When you visit someone's home, you get an idea of the internal situation there. You can estimate how they are doing, for instance... by observing their tone, silence, arguments, and fights. You can figure out that if a woman is distressed now, it might be because of her husband, or maybe because of her mother-in-law. In this case, they can obtain history and they can provide this lady intervention. And I think that it's not easy to go to a psychologist or psychiatrist for stress. But lady health worker can deal with this individual very comfortably.* [Health system participant_6, Position 108]

Participants indicated that features that improve patient engagement with the intervention were needed in TA-PM+ functionality. Features that are specifically tailored for patients and actively encourage and support patient engagement should be included in TA-PM+ to enhance the effectiveness of the intervention and further strengthen the patient-provider relationship.

> *We have worked on another digital project, there were videos for patients in that. We used to show them videos initially, and patient would understand quite well and used to be very convinced and focused.* [Focus group_3, LHW participant_2, Position 130]

Stigma

Stigma surrounding mental health disorders persists within the society in Pakistan, leading many individuals to perceive mental health treatment as stigmatized. Mental health services are generally provided at tertiary health hospitals, and psychiatric wards are locally known by stigmatized labels, reinforcing mental health's negative connotations. However, an alternative approach, that is, TA-PM+ provision in the community, specifically in the patient's home with LHWs, may reduce stigma and encourage treatment. This is because the community trusts these LHWs. One important component of TA-PM+ that has been identified is the use of persuasive strategies to reduce mental health stigma. In addition to promoting user engagement, TA-PM+ should also destigmatize mental health disorders and convince patients to accept support.

> *I really liked the feature of videos, especially for LHWs dealing with depression patients. This is very good. Additionally, there should be a face-to-face component in it. But overall, if you create something that includes engaging and persuading videos for patients, I think it would be very easily available to them all the time through the digital platform. I think it's good enough and efficient enough.* [Health policy participant_7, Position 11-12]

XSL•FO
RenderX

Barriers and Facilitators of Motivation for Health Systems to Implement TA-PM+ and the Features of TA-PM+ to Enhance Motivation Within the COM-B Framework

Credibility

Critical determinants of reflective motivation include credibility and reputation within the community. An LHW provides service to 150-200 households in the community where she resides. Improving her reputation in the community is directly linked to the provision of modern tools and expanded services, which leads to a sense of empowerment. It was perceived that her reputation and position in the community would be further enhanced by the implementation of the TA-PM+ program using a digital tablet to address depression and stress.

> *One thing to remember is that when we provide a tool to LHW, I give an example like this: if we introduce a new intervention that we are implementing throughout Pakistan, where they check blood pressure and refer patients with hypertension. When you provide them with a tool, their credibility in the community improves. In this, basically counseling, is included. But when you give them a digital tool, it motivates them and allows them to feel a bit empowered.* [Health policy participant_2, Position 42]

Co-production of TA-PM+

During the co-production of TA-PM+, the intervention development team reflected on the findings from stage 1 and decided on the key features of TA-PM+. Table 1 summarizes the TA-PM+ features planned during co-production workshops informed by the themes identified in the qualitative analyses of stage 1.

Table 1. Functionality of Technology-Assisted Problem Management Plus (TA-PM+) identified through a qualitative investigation and features developed in a co-production workshop.

TA-PM+[a] functionality identified in a qualitative investigation	TA-PM+ features developed in a co-production workshop
User guidance	• Various prompts and alerts in TA-PM+ to guide LHWs[b] and confirm completion of tasks • Instruction manual for TA-PM+
Easy and time efficient	• Basic consistent design across every session • Fully automated
Increase digital and mental literacy	• Two-minute informational videos for LHWs that provide information and outline the purpose of the session, its steps, the technique used, and how to use TA-PM+ to conduct the session • PDFs of the PM+[c] Urdu manual
Reduce cognitive load	• Fully automated PSYCHLOPS[d] displays questions in accordance with the scoring protocol • No typing required and audio recording available
Promote competence/enablement	• Personalized dashboard for each LHW, which includes patient details, name, address, session details, and next appointment date • Red, yellow, and green color codes to encourage regular appointments; the green code indicates that the session has just been delivered, the yellow code indicates that it is almost due, and the red code indicates that it is urgently due
Internet independent	• To enable use in places where internet access may be costly or where bandwidth is limited or unpredictable, the content's file size is kept as small as possible • All videos are stored in TA-PM+ and can be viewed without the internet
Supervision and monitoring	• Web-based supervisor backend; information is accessible regarding the session's date, duration, PSYCHLOPS responses, and score • Notification if the PSYCHLOPS score rises in subsequent sessions or suicidal ideation is noted in the patient • A locking feature that prevents session data from being altered once it is submitted • WhatsApp support group for urgent queries related to TA-PM+
Promote engagement	• Psychoeducation videos for patients that explain the physiological symptoms of anxiety and depression to the patients
Persuasion/destigmatization	• Psychoeducation videos for patients that include destigmatized vocabulary for symptoms • Psychoeducation videos for patients narrated by a female doctor to boost persuasiveness

[a]TA-PM+: Technology-Assisted Problem Management Plus.

[b]LHW: lady health worker.

[c]PM+: Problem Management Plus.

[d]PSYCHLOPS: Psychological Outcome Profiles.

XSL•FO
RenderX

TA-PM+ User Interface

TA-PM+ is an app that guides LHWs to deliver the PM+ intervention in a systematic manner. TA-PM+ features a personalized password-protected account and personal dashboard for each LHW (Figure 2), presenting essential patient details, session specifics, and upcoming appointment dates. Using a color-coded system, the app encourages the regular scheduling of appointments, indicating the urgency of upcoming sessions. By clicking on a patient name in the dashboard, the session home screen of the patient opens. TA-PM+ ensures the security and integrity of data through its locking features, preventing the alteration of submitted data.

TA-PM+ includes 5 informational videos, referred to as "Session Aid" (Figure 3), designed to provide LHWs with an outline of each session's purpose, the steps involved, and the techniques employed for each session. Moreover, the videos offer guidance on how to effectively use TA-PM+ to conduct these sessions. The videos have a duration of 2 minutes to promote utilization anytime uncertainty arises, without concern for schedule delays. Additionally, TA-PM+ provides access to the PM+ Urdu manual in PDF format.

To enhance understanding and engagement, and address mental health stigma, a series of psychoeducational videos (Figure 4) were developed and included in the session interface. These videos are narrated by a fictional character Dr Ayesha and employ a narrative approach, presenting the physiological symptoms of anxiety and depression in layperson's terms. The content is tailored to respect the cultural nuances and demographics of target communities and thus ensures maximum relevance. Using destigmatized vocabulary, the videos serve as a potent educational tool, addressing patient concerns and mitigating potential embarrassment or fear associated with their conditions. The primary objective is to develop trust and empathy, thereby promoting patient engagement and fostering acceptance of the sessions. Additionally, these videos play a vital role in efficiently explaining the active ingredients of PM+ sessions to the patients, empowering LHWs to establish context, and minimizing the time required to convey background information. The duration of each video is 4-5 minutes, and the content is designed to be concise and memorable for a diverse audience. The inclusion of a female doctor's narration adds a layer of persuasiveness, reinforcing and promoting mental health awareness. One video is developed for each session to deliver the active therapeutic ingredients for fidelity and ensure delivery at the right dose.

This design approach not only ensures the accurate delivery of active elements through patient-computer interaction, but also maintains the crucial patient-facilitator interaction for counseling and empathy throughout the intervention process. In addition, TA-PM+ enhances the counseling skills of LHWs and the therapeutic alliance between LHWs and patients.

Figure 2. User interface of the Technology-Assisted Problem Management Plus (TA-PM+) app developed to assist lady health workers in delivering mental health interventions within community settings in Pakistan. The first screen presents a dashboard with color-coded indicators (red, yellow, and green) to track and prioritize patient sessions. Subsequent screens illustrate session navigation, patient data entry, and automated PSYCHLOPS (Psychological Outcome Profiles) assessments in Urdu. Features have been co-designed for cultural relevance and usability among lady health workers in Pakistan.

XSL•FO
RenderX

Figure 3. Screenshots from session-aid videos developed for lady health workers to enhance digital literacy and support the delivery of Technology-Assisted Problem Management Plus (TA-PM+). These videos, designed in Urdu, provide step-by-step instructions for navigating the TA-PM+ app and conducting sessions effectively.

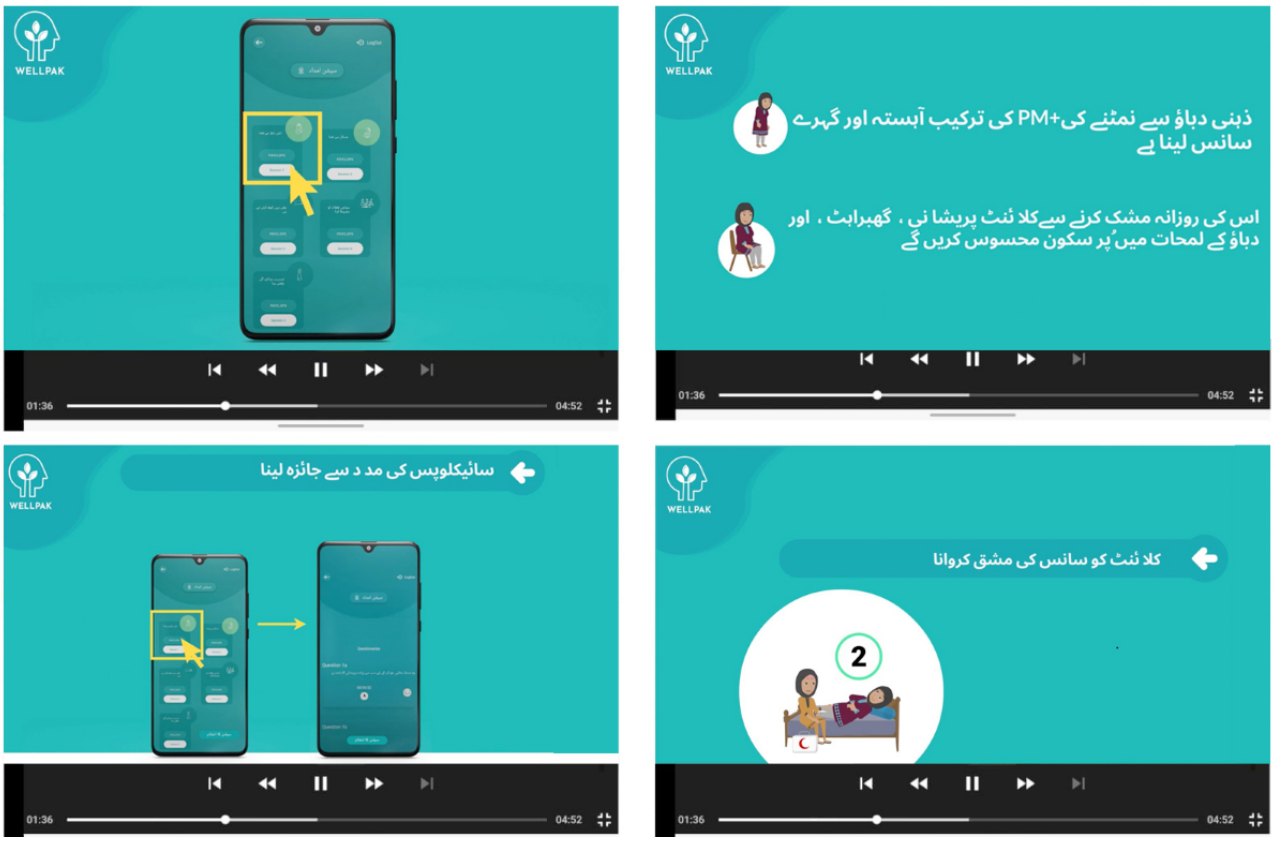

Figure 4. Screenshots from psychoeducational videos for patients, showing the characters Dr Ayesha, lady health worker Shazia, and patient Saima, which have been developed for sessions of Technology-Assisted Problem Management Plus (TA-PM+).

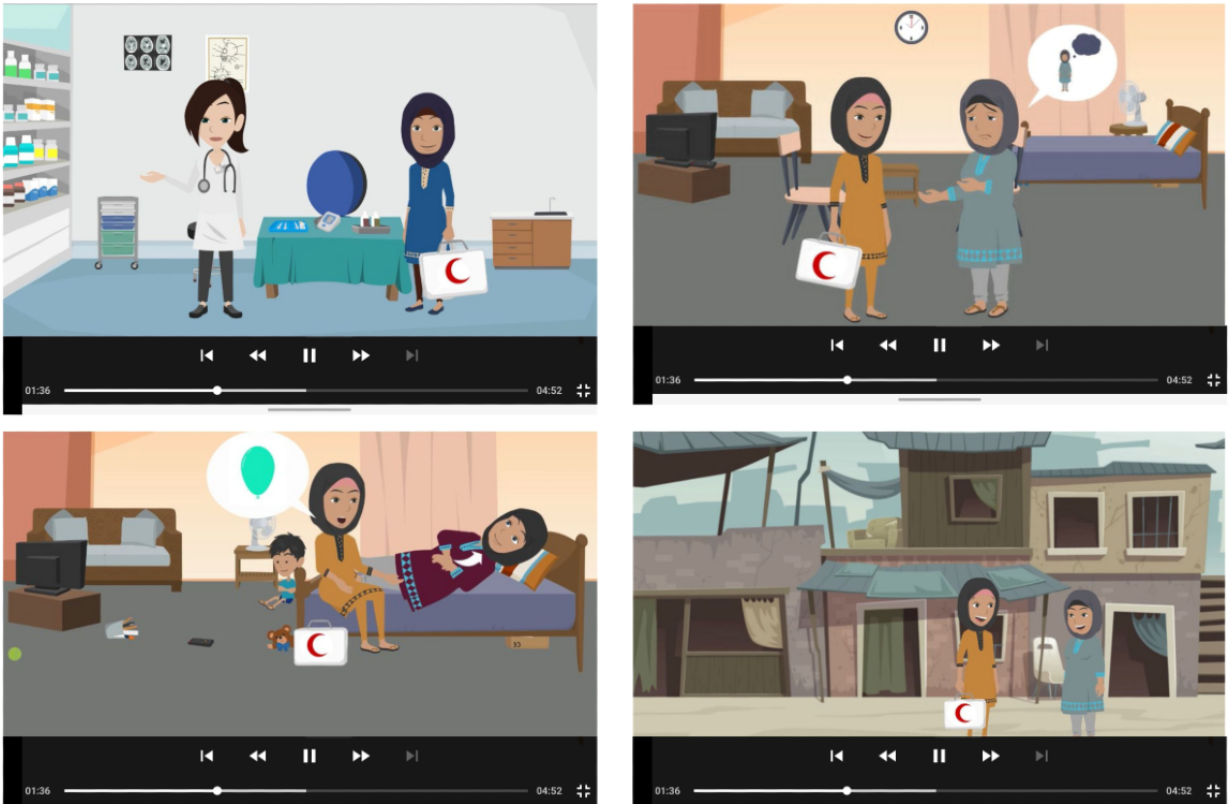

XSL•FO
RenderX

Web Dashboard for Supervisors

The online dashboard provides a user-friendly interface tailored for supervisors, streamlining the data from TA-PM+. This dashboard is structured to include an individual interface for each LHW, which in turn provides access to specific patient details. Supervisors can consistently monitor key information, such as session duration, delivery dates, and PSYCHLOPS scores. Moreover, the system incorporates notifications to alert supervisors in the event of a patient expressing suicidal thoughts, facilitating active case monitoring and prompt referral initiation.

Prototyping and Usability Testing

Employing the MUAQ designed for standalone mHealth apps used by health care providers, usability testing was performed twice, and it involved a cohort of 6 LHWs.

In the initial evaluation, 6 participating LHWs (mean age 44 years; range 41-58 years) were invited to fill the questionnaire and provide their feedback, which highlighted the issues of user interface, language, and pop-up alert settings.

The MUAQ item scores are shown in Table 2. After training, the mean usability score was recorded at 5.62 (highest possible score of 7). In the subsequent field assessment, the usability score increased to 5.96, which indicates improvement resulting from the incorporation of input from the first round. The usability scores demonstrated the system's overall usefulness.

Table 2 displays the average overall scores of the MUAQ according to the following 3 categories: ease of use, interface and satisfaction, and usefulness for delivering TA-PM+. After incorporating feedback from round 1 of usability testing into TA-PM+, scores of the categories ease of use and interface satisfaction increased in round 2. However, scores of the category usefulness for delivering PM+ did not change. This may be because comments were largely sought on the usability, interface, and functionality of TA-PM+ and not on its content, which was adapted from the PM+ manual.

In the subsequent assessment phase, field usability testing was conducted with 6 participants who were screened for depression using the GHQ-12 questionnaire and WHODAS. The 6 screened participants (mean age 29 years; range 25-37 years) were invited to fill the mHealth Satisfaction Questionnaire after TA-PM+ session delivery by LHWs. The TA-PM+ session client satisfaction score was 40 (highest possible score of 46) (Table 3), indicating that the patients expressed contentment with the session.

XSL•FO
RenderX

Table 2. Usability scores of the mHealth Usability App Questionnaire (MUAQ) for the Technology-Assisted Problem Management Plus (TA-PM+) app.

MUAQ[a] item	Score, mean (SD)	
	Round 1 (posttraining)	Round 2 (field)
Ease of use	4.33 (0.31)	5.13 (0.47)
The app was easy to use.	4.50 (1.97)	4.33 (2.65)
It was easy for me to learn to use the app.	5.00 (1.78)	5.83 (1.83)
The navigation was consistent when moving between screens.	3.00 (2.28)	5.16 (2.56)
The interface of the app allowed me to use all the functions (such as looking at the dashboard, filling the PSYCHLOPS[b], opening different sessions, and viewing videos) offered by the app.	4.16 (2.40)	4.00 (2.50)
Whenever I made a mistake using the app, I could recover easily and quickly.	5.00 (1.67)	6.33 (1.63)
Interface and satisfaction	6.01 (0.67)	6.21 (0.51)
I like the interface of the app.	5.16 (2.10)	5.83 (2.85)
The information in the app was well organized, so I could easily find the information I needed.	5.33 (1.86)	6.33 (1.63)
The app adequately acknowledged and provided information to let me know the progress of my action (delivered sessions).	6.33 (1.63)	6.33 (1.54)
I feel comfortable using this app while delivering the session to a patient.	5.83 (1.83)	6.33 (1.21)
The amount of time involved in using this app has been fitting for me.	6.33 (1.63)	6.33 (1.63)
I would use this app again.	7.00 (0.00)	6.33 (1.63)
Overall, I am satisfied with this app.	6.33 (1.63)	6.33 (1.63)
Usefulness	6.11 (0.89)	6.11 (0.61)
The app would be useful for my health care practice.	7.00 (0.00)	6.33 (1.63)
The app improved my access to delivering mental health service.	7.00 (0.00)	6.33 (1.63)
The app helped me manage my patients' mental health effectively.	6.33 (1.21)	6.16 (1.60)
This app has all the functions and capabilities I expected it to have.	6.00 (1.26)	6.50 (1.22)
I could use the app even when the internet connection was poor.	3.66 (2.65)	4.50 (2.25)
This mobile health app provides an acceptable way to deliver the PM+[c] intervention, such as accessing educational materials, tracking my own activities, and performing assessments.	6.66 (0.51)	6.83 (0.40)
Overall usability of the TA-PM+[d] app	5.62 (0.81)	5.96 (0.59)

[a]MUAQ: mHealth Usability App Questionnaire.

[b]PSYCHLOPS: Psychological Outcome Profiles.

[c]PM+: Problem Management Plus.

[d]TA-PM+: Technology-Assisted Problem Management Plus.

XSL•FO
RenderX

Table 3. Client satisfaction (mHealth Satisfaction Questionnaire) with the videos and sessions delivered by lady health workers using the Technology-Assisted Problem Management Plus (TA-PM+) app.

mHealth Satisfaction Questionnaire items	Score, mean (SD)
The video was helpful in receiving the sessions.	4.83 (0.40)
I enjoyed receiving the sessions through the app and videos.	5.00 (0.00)
The time spent receiving the session through the app was acceptable.	4.83 (0.40)
The language used in the videos was difficult to understand.	5.00 (0.00)
The information given in the video was in small chunks, which I was able to retain.	4.83 (0.40)
Videos were too time consuming.	2.33 (2.06)
I didn't like the videos or app usage during the session, and it felt like interruption.	1.66 (1.63)
It was so boring that I lost my interest during the sessions.	1.66 (0.81)
It was like a lecture, a disturbance, or too difficult to understand.	1.66 (0.00)
I can recommend it to others.	4.33 (1.03)
The avatars helped me to engage with the program. The scene settings were relatable, and it has motivated me to do the practices recommended for my well-being.	4.83 (1.02)
The examples in the videos helped me in understanding the information and helped me understand how the sessions will help my mental well-being.	4.83 (0.40)
The examples in the videos helped me understand how I need to implement session strategies in my life to feel better.	4.66 (0.51)
The videos and sessions have helped me set personal goals for my lifestyle habits in a way that I could not have done on my own.	4.83 (0.40)
Overall client satisfaction	40

Discussion

Principal Findings

The TA-PM+ app was designed to support LHWs in delivering the transdiagnostic PM+ intervention to patients with symptoms of stress, anxiety, and depression in Islamabad, Pakistan. The development process followed a thorough 3-stage methodology involving co-production with local stakeholders. Throughout the development of TA-PM+, careful consideration was given to the local health system perspective, available resources, and cultural sensitivities by involvement of experts. Usability testing confirmed that TA-PM+ is well-received by both LHWs and participants in community settings.

Urdu-translated PM+ was further adapted for technology-assisted delivery in our study using a comprehensive methodology involving local stakeholders and experts, a strategy recommended in the existing literature for PM+ adaptation. According to Heim et al [33], PM+ requires many adaptations to fit a given context before testing or implementing PM+ in that setting. Even changes to peripheral components of the intervention, such as instructional materials and illustrations, require the involvement of numerous experts in planning, translation, and cultural adaptation [34]. This procedural diligence is equally important when adapting the delivery style, for example, when PM+ is predominantly administered through web-based communication or when the length of treatment deviates from the originally specified 5 PM+ sessions [34].

The TA-PM+ app was developed through a co-production methodology, with the chosen delivery format being face-to-face interaction. The primary objective of the TA-PM+ app is to support LHWs in delivering the PM+ intervention in person, and patients only interact with the app to access psychoeducational videos. This particular design choice was carefully considered during the co-production to ensure a smooth integration with contextual elements and local digital and health resources, and the potential for a wider impact. Existing literature emphasizes that the success of digital technologies in mental health is dependent upon their alignment with sociopolitical conditions and cultural contexts. Failure to align with these factors often results in reduced user engagement and higher attrition rates. The term "nonalignment" encompasses various factors, such as values, mental health treatment methods, and specific health needs of the population [33,35].

The barriers to implementing TA-PM+ that were highlighted during the first round of qualitative investigation included skills and knowledge of LHWs to use TA-PM+, workload, time to implement TA-PM+, stigma associated with mental health, and lack of resources like internet availability. These barriers are in line with the current literature. A survey to explore barriers to implementation of digital mental health interventions in primary health care settings identified the availability of supporting infrastructure, organizational readiness and capacity, health care professional skills, and cultural appropriateness [36]. Research on digital mental health implementation [34] and digital health interventions through health systems [37] has shown that resource availability is a major barrier. The challenge posed by limited resource availability is particularly heightened in LMICs that are already struggling with resource constraints. A notable challenge in LMICs pertains to the availability of internet access and sufficient bandwidth. Addressing this barrier in digital health technologies is imperative, as highlighted both in our

XSL•FO
RenderX

study and in a concurrent investigation conducted in Nepal [38,39], advocating for the implementation of interventions that are independent of internet connectivity. One of the most frequent recommendations for TA-PM+ that emerged through the qualitative investigation was to shorten the sessions from 90 minutes to 45-60 minutes to increase feasibility. This is consistent with the existing research on the reasons for high attrition in PM+. Prior research has indicated that participants commonly had difficulties finishing all 5 sessions of PM+ because of the lengthy sessions, which frequently interfered with finding or maintaining a job, performing domestic duties, or raising children [8-10]. Top of Form

During the co-production workshop, customization of the content, especially tailoring of the psychoeducational videos to respect the cultural nuances and demographics of the target communities, was planned for TA-PM+, to address stigma and improve engagement. Likewise, recent reviews on engagement with digital mental health interventions [40,41] identified that tailored digital technologies with personalized videos and avatars increased user engagement and demonstrated effectiveness. To augment persuasiveness in TA-PM+, there was a strategic decision during co-production workshops to have the videos narrated by a fictional doctor avatar. This design feature of TA-PM+ is consistent with the self-help digital intervention recently developed by the WHO for depression, wherein all the videos feature a doctor as the narrator [19].

Limitations

There are limitations in this study, which have been divided into methodological limitations and design limitations of TA-PM+.

Methodological limitations include the potential lack of inclusivity and diversity in the co-production workshop sample, where people with lived experiences were not involved in the development of TA-PM+. Another limitation is the usability testing, which was conducted on a limited sample size and in the first session of TA-PM+, assuming interface uniformity across all sessions. The subsequent stage of our evaluation (feasibility study) will assess the feasibility, effectiveness, and usability of TA-PM+.

One of the main limitations in the design of TA-PM+ is that it requires the internet for the backend data storage function. Internet availability and bandwidth were identified as major barriers to the implementation of TA-PM+ in qualitative investigation. However, owing to backend storage of data and online monitoring, TA-PM+ cannot function without the internet, although some features requiring a high bandwidth, such as video playback, work without internet access. An additional limitation is that TA-PM+ will be administered by LHWs who provide services to women only. Thus, it will not be applicable to men, even though men require mental health care. TA-PM+ is currently in Urdu, necessitating technical and cultural adaptations to ensure its relevance and effectiveness in various sociocultural settings.

Conclusions

We successfully developed the TA-PM+ app for LHWs, with the potential to be integrated into primary health care to address common mental health conditions. This work has several implications even beyond the Pakistan setting. These include (1) an understanding of the process of developing a resource-appropriate digital health tool for high uptake; (2) a model for adaptation that is flexible and can be used to create similar tools for PM+ in other low-resource settings globally; and (3) the ability of the system to be expanded to function as a national tool to capture the prevalence of common mental health conditions at the population level.

Further feasibility testing and eventual up-scaling are envisaged based on the thorough development process outlined in this paper.

Acknowledgments

This work was supported by the German Federal Ministry of Education and Research (BMBF) grant 01KA2105A as part of the German Alliance for Global Health Research (GLOHRA).

Data Availability

All data generated and analyzed during this study are included in the main manuscript and its supplementary materials. The data are not deposited in a public repository due to the sensitive nature of participant information and ethical restrictions. Subject to ethical guidelines and institutional review board approval, the corresponding author can provide additional supporting data, such as comprehensive usability scores and satisfaction ratings, upon reasonable request.

Authors' Contributions

MS, SS, and HZ conceived and planned the study. MS and SZ conducted the qualitative interviews and analyses. MS, SZ, MK, TK, AB, DCF, SS, and HZ were involved in the co-production workshops and the planning of the TA-PM+ features. AB developed the software for TA-PM+. MS and SZ trained the lady health workers and conducted usability testing. MS and DCF wrote the manuscript. All authors commented on the manuscript.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
COREQ (COnsolidated criteria for REporting Qualitative research) Checklist.
[PDF File (Adobe PDF File), 403 KB - formative_v9i1e59414_app1.pdf]

Multimedia Appendix 2
Interview guides.
[DOCX File , 30 KB - formative_v9i1e59414_app2.docx]

Multimedia Appendix 3
Code chart.
[DOCX File , 21 KB - formative_v9i1e59414_app3.docx]

References

1. GBD 2019 Mental Disorders Collaborators. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet Psychiatry 2022 Feb;9(2):137-150 [FREE Full text] [doi: 10.1016/S2215-0366(21)00395-3] [Medline: 35026139]
2. American Psychiatric Association. Diagnostic and statistical manual of mental disorders: DSM-5. Washington, DC: American Psychiatric Publishing, Inc; 2013.
3. Depression and Other Common Mental Disorders: Global Health Estimates. World Health Organization. 2017. URL: https://iris.who.int/bitstream/handle/10665/254610/WHO-MSD-MER-2017.2-eng.pdf [accessed 2024-12-30]
4. WHO and Ministry of Health Pakistan. WHO-AIMS Report on Mental Health System in Pakistan. World Health Organization. 2009. URL: https://extranet.who.int/mindbank/item/1305 [accessed 2024-12-30]
5. Saxena S, Funk M, Chisholm D. World Health Assembly adopts Comprehensive Mental Health Action Plan 2013-2020. Lancet 2013 Jun 08;381(9882):1970-1971 [FREE Full text] [doi: 10.1016/S0140-6736(13)61139-3] [Medline: 23746771]
6. Verhey IJ, Ryan GK, Scherer N, Magidson JF. Implementation outcomes of cognitive behavioural therapy delivered by non-specialists for common mental disorders and substance-use disorders in low- and middle-income countries: a systematic review. Int J Ment Health Syst 2020;14:40 [FREE Full text] [doi: 10.1186/s13033-020-00372-9] [Medline: 32514304]
7. Keynejad RC, Spagnolo J, Thornicroft G. Mental healthcare in primary and community-based settings: evidence beyond the WHO Mental Health Gap Action Programme (mhGAP) Intervention Guide. Evid Based Ment Health 2022 Dec;25(e1):e1-e7 [FREE Full text] [doi: 10.1136/ebmental-2021-300401] [Medline: 35473750]
8. Acarturk C, Uygun E, Ilkkursun Z, Yurtbakan T, Kurt G, Adam-Troian J, et al. Group problem management plus (PM+) to decrease psychological distress among Syrian refugees in Turkey: a pilot randomised controlled trial. BMC Psychiatry 2022 Jan 04;22(1):8 [FREE Full text] [doi: 10.1186/s12888-021-03645-w] [Medline: 34983461]
9. de Graaff AM, Cuijpers P, McDaid D, Park A, Woodward A, Bryant RA, et al. Peer-provided Problem Management Plus (PM+) for adult Syrian refugees: a pilot randomised controlled trial on effectiveness and cost-effectiveness. Epidemiol Psychiatr Sci 2020 Aug 18;29:e162. [doi: 10.1017/s2045796020000724]
10. Khan MN, Hamdani SU, Chiumento A, Dawson K, Bryant RA, Sijbrandij M, et al. Evaluating feasibility and acceptability of a group WHO trans-diagnostic intervention for women with common mental disorders in rural Pakistan: a cluster randomised controlled feasibility trial. Epidemiol Psychiatr Sci 2019 Feb;28(1):77-87 [FREE Full text] [doi: 10.1017/S2045796017000336] [Medline: 28689511]
11. Maercker A, Heim E, Kirmayer LJ. Cultural Clinical Psychology and PTSD. Göttingen, Germany: Hogrefe Publishing GmbH; 2019.
12. Hamdani SU, Huma Z, Rahman A, Wang D, Chen T, van Ommeren M, et al. Cost-effectiveness of WHO Problem Management Plus for adults with mood and anxiety disorders in a post-conflict area of Pakistan: randomised controlled trial. Br J Psychiatry 2020 Nov 28;217(5):623-629. [doi: 10.1192/bjp.2020.138] [Medline: 32720628]
13. Rahman A, Riaz N, Dawson KS, Usman Hamdani S, Chiumento A, Sijbrandij M, et al. Problem Management Plus (PM+): pilot trial of a WHO transdiagnostic psychological intervention in conflict-affected Pakistan. World Psychiatry 2016 Jun;15(2):182-183 [FREE Full text] [doi: 10.1002/wps.20312] [Medline: 27265713]
14. Jordans MJD, Kohrt BA, Sangraula M, Turner EL, Wang X, Shrestha P, et al. Effectiveness of Group Problem Management Plus, a brief psychological intervention for adults affected by humanitarian disasters in Nepal: A cluster randomized controlled trial. PLoS Med 2021 Jun;18(6):e1003621 [FREE Full text] [doi: 10.1371/journal.pmed.1003621] [Medline: 34138875]
15. Dawson KS, Bryant RA, Harper M, Kuowei Tay A, Rahman A, Schafer A, et al. Problem Management Plus (PM+): a WHO transdiagnostic psychological intervention for common mental health problems. World Psychiatry 2015 Oct 25;14(3):354-357 [FREE Full text] [doi: 10.1002/wps.20255] [Medline: 26407793]
16. Schäfer S, Thomas LM, Lindner S, Lieb K. World Health Organization's low-intensity psychosocial interventions: a systematic review and meta-analysis of the effects of Problem Management Plus and Step-by-Step. World Psychiatry 2023 Oct 15;22(3):449-462 [FREE Full text] [doi: 10.1002/wps.21129] [Medline: 37713578]

XSL•FO
RenderX

17. Chambers DA, Glasgow RE, Stange KC. The dynamic sustainability framework: addressing the paradox of sustainment amid ongoing change. Implement Sci 2013 Oct 02;8:117 [FREE Full text] [doi: 10.1186/1748-5908-8-117] [Medline: 24088228]
18. Mudiyanselage KWW, De Santis KK, Jörg F, Saleem M, Stewart R, Zeeb H, et al. The effectiveness of mental health interventions involving non-specialists and digital technology in low-and middle-income countries - a systematic review. BMC Public Health 2024 Jan 03;24(1):77 [FREE Full text] [doi: 10.1186/s12889-023-17417-6] [Medline: 38172713]
19. Carswell K, Harper-Shehadeh M, Watts S, Van't Hof E, Abi Ramia J, Heim E, et al. Step-by-Step: a new WHO digital mental health intervention for depression. Mhealth 2018;4:34 [FREE Full text] [doi: 10.21037/mhealth.2018.08.01] [Medline: 30225240]
20. McBride K, Harrison S, Mahata S, Pfeffer K, Cardamone F, Ngigi T, et al. Building mental health and psychosocial support capacity during a pandemic: the process of adapting problem management plus for remote training and implementation during COVID-19 in New York City, Europe and East Africa. Intervention 2021;19(1):37. [doi: 10.4103/INTV.INTV_30_20]
21. Rivero-Santana A, Perestelo-Perez L, Alvarez-Perez Y, Ramos-Garcia V, Duarte-Diaz A, Linertova R, PECUNIA Group. Stepped care for the treatment of depression: a systematic review and meta-analysis. J Affect Disord 2021 Nov 01;294:391-409 [FREE Full text] [doi: 10.1016/j.jad.2021.07.008] [Medline: 34315101]
22. Hussain SS, Khan M, Gul R, Asad N. Integration of mental health into primary healthcare: perceptions of stakeholders in Pakistan. East Mediterr Health J 2018 May 03;24(2):146-153 [FREE Full text] [Medline: 29748943]
23. Hafeez A, Mohamud B, Shiekh M, Shah S, Jooma R. Lady health workers programme in Pakistan: Challenges, achievements and the way forward. J Pak Med Assoc 2011 Mar 1;61(3):210-215 [FREE Full text]
24. Population Census. Pakistan Bureau of Statistics. URL: http://www.pbs.gov.pk/content/population-census [accessed 2024-12-30]
25. Hawkins J, Madden K, Fletcher A, Midgley L, Grant A, Cox G, et al. Development of a framework for the co-production and prototyping of public health interventions. BMC Public Health 2017 Sep 04;17(1):689 [FREE Full text] [doi: 10.1186/s12889-017-4695-8] [Medline: 28870192]
26. Atkins L, Francis J, Islam R, O'Connor D, Patey A, Ivers N, et al. A guide to using the Theoretical Domains Framework of behaviour change to investigate implementation problems. Implement Sci 2017 Jun 21;12(1):77 [FREE Full text] [doi: 10.1186/s13012-017-0605-9] [Medline: 28637486]
27. Thomas DR. Feedback from research participants: are member checks useful in qualitative research? Qualitative Research in Psychology 2016 Aug 02;14(1):23-41. [doi: 10.1080/14780887.2016.1219435]
28. Williams L. Agile software development methodologies and practices. Advances in Computers 2010;80:1-44. [doi: 10.1016/S0065-2458(10)80001-4]
29. Joyce G, Lilley M. Towards the Development of Usability Heuristics for Native Smartphone Mobile Applications. In: Marcus A, editor. Design, User Experience, and Usability. Theories, Methods, and Tools for Designing the User Experience. DUXU 2014. Lecture Notes in Computer Science, vol 8517. Cham: Springer; 2014:465-474.
30. Nielsen J. How Many Test Users in a Usability Study? NN/g. URL: https://www.nngroup.com/articles/how-many-test-users/ [accessed 2024-12-30]
31. Zhou L, Bao J, Setiawan IMA, Saptono A, Parmanto B. The mHealth App Usability Questionnaire (MAUQ): Development and validation study. JMIR Mhealth Uhealth 2019 Apr 11;7(4):e11500 [FREE Full text] [doi: 10.2196/11500] [Medline: 30973342]
32. Melin J, Bonn SE, Pendrill L, Trolle Lagerros Y. A questionnaire for assessing user satisfaction with mobile health apps: development using Rasch Measurement Theory. JMIR Mhealth Uhealth 2020 May 26;8(5):e15909 [FREE Full text] [doi: 10.2196/15909] [Medline: 32452817]
33. Heim E, Kohrt BA. Cultural adaptation of scalable psychological interventions: a new conceptual framework. CPE 2019 Dec 17;1(4):37679. [doi: 10.32872/cpe.v1i4.37679]
34. Williams MG, Stott R, Bromwich N, Oblak SK, Espie CA, Rose JB. Determinants of and barriers to adoption of digital therapeutics for mental health at scale in the NHS. BMJ Innov 2020 May 07;6(3):92-98. [doi: 10.1136/bmjinnov-2019-000384]
35. Connelly K, Stein KF, Chaudry B, Trabold N. Development of an ecological momentary assessment mobile app for a low-literacy, Mexican American population to collect disordered eating behaviors. JMIR Public Health Surveill 2016 Jul 14;2(2):e31 [FREE Full text] [doi: 10.2196/publichealth.5511] [Medline: 27418020]
36. Brantnell A, Temiz S, Baraldi E, Woodford J, von Essen L. Barriers to and facilitators of the implementation of digital mental health interventions as perceived by primary care decision makers: content analysis of structured open-ended survey data. JMIR Hum Factors 2023 Jun 26;10:e44688 [FREE Full text] [doi: 10.2196/44688] [Medline: 37358902]
37. Neher M, Nygårdh A, Broström A, Lundgren J, Johansson P. Perspectives of policy makers and service users concerning the implementation of eHealth in Sweden: interview study. J Med Internet Res 2022 Jan 28;24(1):e28870 [FREE Full text] [doi: 10.2196/28870] [Medline: 35089139]
38. Pokhrel P, Karmacharya R, Taylor Salisbury T, Carswell K, Kohrt BA, Jordans MJD, et al. Perception of healthcare workers on mobile app-based clinical guideline for the detection and treatment of mental health problems in primary care: a qualitative study in Nepal. BMC Med Inform Decis Mak 2021 Jan 19;21(1):21 [FREE Full text] [doi: 10.1186/s12911-021-01386-0] [Medline: 33468120]

XSL•FO
RenderX

39. Luitel N, Neupane V, Lamichhane B, Koirala G, Gautam K, Karki E, et al. Experience of primary healthcare workers in using the mobile app-based WHO mhGAP intervention guide in detection and treatment of people with mental disorders: A qualitative study in Nepal. SSM Ment Health 2023 Dec 15;4:100278 [FREE Full text] [doi: 10.1016/j.ssmmh.2023.100278] [Medline: 38155812]
40. Saleem M, Kühne L, De Santis KK, Christianson L, Brand T, Busse H. Understanding engagement strategies in digital interventions for mental health promotion: scoping review. JMIR Ment Health 2021 Dec 20;8(12):e30000 [FREE Full text] [doi: 10.2196/30000] [Medline: 34931995]
41. Morrison LG, Yardley L, Powell J, Michie S. What design features are used in effective e-health interventions? A review using techniques from Critical Interpretive Synthesis. Telemed J E Health 2012 Mar;18(2):137-144 [FREE Full text] [doi: 10.1089/tmj.2011.0062] [Medline: 22381060]

Abbreviations

COM-B: Capability, Opportunity, and Motivation for Behavioral Change
GHQ-12: General Health Questionnaire-12
LHS: lady health supervisor
LHW: lady health worker
LMIC: low- and middle-income country
mhGAP: Mental Health Gap Action Programme
MUAQ: mHealth App Usability Questionnaire
OPD: outpatient department
PM+: Problem Management Plus
PSQ: Patient Satisfaction Questionnaire
TA-PM+: Technology-Assisted Problem Management Plus
TDF: Theoretical Domain Framework
UC: Union Council
WHO: World Health Organization
WHODAS: WHO Disability Assessment Schedule

Edited by A Mavragani; submitted 11.04.24; peer-reviewed by W Zhang; comments to author 16.10.24; revised version received 06.11.24; accepted 02.12.24; published 28.01.25.

Please cite as:
Saleem M, Zafar S, Klein T, Koesters M, Bashir A, Fuhr DC, Sikander S, Zeeb H
A Digital Tool (Technology-Assisted Problem Management Plus) for Lay Health Workers to Address Common Mental Health Disorders: Co-production and Usability Study in Pakistan
JMIR Form Res 2025;9:e59414
URL: https://formative.jmir.org/2025/1/e59414
doi:10.2196/59414
PMID:39874072

©Maham Saleem, Shamsa Zafar, Thomas Klein, Markus Koesters, Adnan Bashir, Daniela C Fuhr, Siham Sikander, Hajo Zeeb. Originally published in JMIR Formative Research (https://formative.jmir.org), 28.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Technological Adjuncts to Streamline Patient Recruitment, Informed Consent, and Data Management Processes in Clinical Research: Observational Study

Jodie Koh[1], MS; Stacey Caron[2*], BSN; Amber N Watters[2,3*], MD, MS; Mahesh Vaidyanathan[2,3*], MD; David Melnick[2*], BS; Alyssa Santi[3*], BSN; Kenneth Hudson[2,3*], BS, MS; Catherine Arguelles[3*], BSN; Priyanka Mathur[4,5*], MD; Mozziyar Etemadi[2,3*], MD, PhD

[1]Kellogg School of Management, Northwestern University, Evanston, IL, United States
[2]Northwestern Medicine, Chicago, IL, United States
[3]Feinberg School of Medicine, Northwestern University, Chicago, IL, United States
[4]Department of Obstetrics and Gynecology, Brigham and Women's Hospital, Boston, MA, United States
[5]Department of Obstetrics and Gynecology, Massachusetts General Hospital, Boston, MA, United States
[*]these authors contributed equally

Corresponding Author:
Jodie Koh, MS
Kellogg School of Management
Northwestern University
2211 Campus Drive
Evanston, IL, 60208
United States
Phone: 1 847 491 3300
Email: jodie.koh@kellogg.northwestern.edu

Abstract

Background: Patient recruitment and data management are laborious, resource-intensive aspects of clinical research that often dictate whether the successful completion of studies is possible. Technological advances present opportunities for streamlining these processes, thus improving completion rates for clinical research studies.

Objective: This paper aims to demonstrate how technological adjuncts can enhance clinical research processes via automation and digital integration.

Methods: Using one clinical research study as an example, we highlighted the use of technological adjuncts to automate and streamline research processes across various digital platforms, including a centralized database of electronic medical records (enterprise data warehouse [EDW]); a clinical research data management tool (REDCap [Research Electronic Data Capture]); and a locally managed, Health Insurance Portability and Accountability Act–compliant server. Eligible participants were identified through automated queries in the EDW, after which they received personalized email invitations with digital consent forms. After digital consent, patient data were transferred to a single Health Insurance Portability and Accountability Act–compliant server where each participant was assigned a unique QR code to facilitate data collection and integration. After the research study visit, data obtained were associated with existing electronic medical record data for each participant via a QR code system that collated participant consent, imaging data, and associated clinical data according to a unique examination ID.

Results: Over a 19-month period, automated EDW queries identified 20,988 eligible patients, and 10,582 patients received personalized email invitations. In total, 1000 (9.45%) patients signed consents to participate in the study. Of the consented patients, 549 unique patients completed 779 study visits; some patients consented to the study at more than 1 time period during their pregnancy.

Conclusions: Technological adjuncts in clinical research decrease human labor while increasing participant reach and minimizing disruptions to clinic operations. Automating portions of the clinical research process benefits clinical research efforts by expanding and optimizing participant reach while reducing the limitations of labor and time in completing research studies.

(JMIR Form Res 2025;9:e58628) doi:10.2196/58628

XSL•FO
RenderX

KEYWORDS

digital health; patient recruitment; consent; technological adjuncts; data management; clinical research processes; automation; digital platforms; data warehouse; patient data; imaging data; pregnancy; clinical research methods

Introduction

Clinical research processes, specifically in patient recruitment, are rife with logistical complexities that often determine the success or failure of a clinical trial [1]. Successful participant enrollment depends on organization, personnel, and participant factors as well as the features of the clinical trial at hand [2-5]. McDonald et al [6] cited reasons for research study failure as including heavy clinical and research workload of research team members, perceived imbalance between patient incentive and risk, and low status typically conferred to recruitment work. Of 114 studies in the review of McDonald et al [6], only 31% met their original recruitment goals, while 53% required extended periods for completion.

The resource-intensiveness of recruitment activities often dictates whether studies are completed successfully [7]. Recruitment processes have traditionally relied heavily on trained manual labor [8,9], which is costly. Moreover, these research personnel often face logistic barriers when recruiting in high-volume clinical settings, where their research activities may be restricted by clinical workflows and tight patient turnaround [10]. In general, in-person recruiting is personnel-intensive, as it requires research staff to repetitively explain the research study to each potential participant accurately and representatively [11,12]. Finally, traditional methods for recruiting patients rely on collecting and filing printed paperwork, especially in managing patient records and study results, which oftentimes require a high cognitive workload to manage and safeguard [2].

Following these observations, health care and federal institutions have increasingly demonstrated support for the integration of digital technological capabilities within clinical research [13-15], especially leveraging the capabilities brought about by automating clinical processes [16,17]. In 2019, the National Heart, Lung, and Blood Institute; National Institutes of Health; and National Science Foundation hosted a workshop calling for the digitization of clinical research using advanced analytics to facilitate patient screening and data management to increase the representation of diverse patient populations [18]. In a separate review of recruitment methods across 61 clinical trials, digital recruitment was found to be 52% more effective than conventional offline recruitment [19].

Such technological advancements present opportunities for overcoming the onerous complexities of clinical research workflows while also reducing the operational costs of research (by reducing research staff labor-intensiveness and expanding potential participant reach) as well as the potential for human error in recruitment processes (by digitizing data management).

In this paper, we share our experience using technological adjuncts to streamline participant identification, recruitment, consent, and data management processes for an observational clinical research study. These strategies allowed us to complete our study recruitment goal ahead of schedule and in a cost-effective, resource-optimized way.

Methods

Ethical Considerations

All elements of the research study—which, in this portion, involved patient recruitment to achieve the goals of our broader study—were approved by Northwestern University's Institutional Review Board Office (study ID: STU00215717). This approval confirms that all aspects of the study meet the ethical standards of the review board. Invited patients were asked to provide consent if they agreed to participate in our study and were also given the option to opt out of the study with no consequences to the care provided to them. Patients were not financially compensated for their participation in our study, and all data collected were subsequently deidentified.

Purpose of Broader Research Study

The digitized recruitment processes that we describe in this paper are part of a prospective, observational research study. The research study's primary aim is to develop a database of ultrasonography images for use in the development of artificial intelligence technology with specific applications to perinatal care. To fulfill this goal, we planned to enroll 1000 patients across various time periods of pregnancy from within a single institution. To accomplish this task, we designed and executed a digitized process of patient recruitment and data management.

Given our ultimate goal to build algorithms capable of providing diagnostic recommendations and predictions of various maternal and fetal health outcomes, we engaged in strategic and targeted patient recruitment to ensure that we recruited an appropriate number of patients with specific health conditions across various outcomes while abiding by constraints including a temporally bounded funding structure. Our clinical research processes relied on 3 main technological adjuncts: a secured, centralized database of electronic medical records (enterprise data warehouse [EDW]), a research data management tool (REDCap [Research Electronic Data Capture]; Vanderbilt University), and a locally managed, Health Insurance Portability and Accountability Act–compliant QR code–based system that allowed us to coordinate all patient information, data, and enrollment-related documentation on a single, secured server. An overview of our clinical research processes is depicted in Figure 1. Each step of these processes is detailed below.

XSL•FO
RenderX

Figure 1. Workflow for clinical research processes across digital adjuncts. API: application programming interface; EDW: enterprise data warehouse; REDCap: Research Electronic Data Capture.

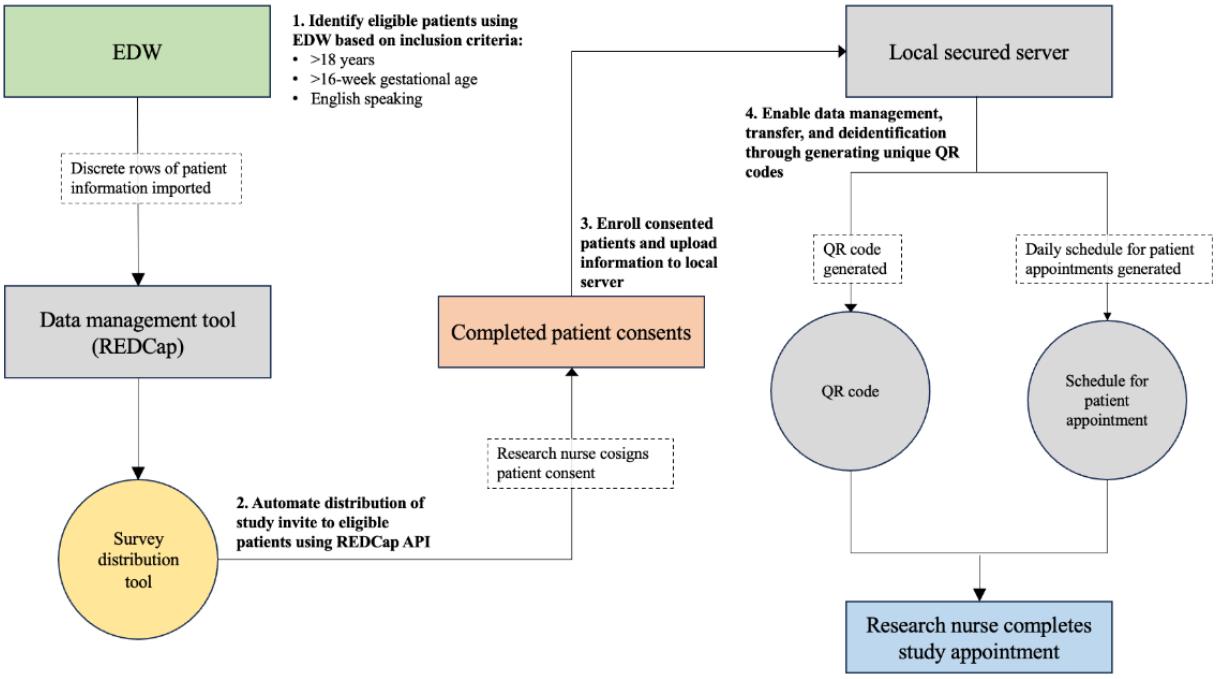

Identify Eligible Patients Using EDW Based on Inclusion Criteria

The patient recruitment process leveraged access to patient medical records collated from Epic and stored in a local EDW to identify eligible patients for the study. In doing so, we performed a SQL query, filtering for patients based on the inclusion and exclusion criteria and limited to participating clinical units. The inclusion criteria were patients aged 18 years and older, English speaking, and at least 16 weeks of gestational age. This query derived multiple rows of eligible patients to be invited to participate in our study. We set up this query function as an automated process with the list of eligible patients refreshed every 24 hours. This query generated a list of patient appointments that had been scheduled within the next 30 days across participating clinics.

Automate Distribution of Study Invite to Eligible Patients Using the REDCap Application Programming Interface

We used a REDCap-specific application programming interface to import the rows of eligible patient appointments generated from the EDW into the REDCap database. Each discrete row provided all necessary details for patient identification and recruitment including patient medical record numbers, personal and clinical information, contact information, and details including patient notes on upcoming appointments. Patients provide universal consent when first receiving care at our institution, thereby granting access to contact information for use in recruitment for research studies.

Within REDCap, we conducted an additional manual check on the list of eligible patients against our inclusion criteria using a filtering function. This was done to mitigate any errors that may have been made as a result of data compatibility issues during transfer between platforms. Given that patients are exclusively electronically recruited for this study, we also filtered patients based on the presence of an email address on file. Finally, we filtered to exclude patients who previously indicated disinterest in participating in this study, resulting in a narrowed list of eligible patients to recruit.

We used REDCap's survey distribution tool function to distribute invitations to eligible patients. We sent personalized emails with patient names as designated in their charts followed by all relevant information pertaining to the study including the title and purpose of clinical research, risks and benefits of participation, participant expectations for the study, and contact information for any inquiries. Patients were provided unique links that directed them to their personalized electronic consent forms embedded in REDCap. The email invitation text may be reviewed in Multimedia Appendix 1.

Enroll Consented Patients and Upload Information to Local Server

As the research nurse on the team cosigned patient consent forms, using a second SQL query, we uploaded the list of consented patients onto a local server managed by the project team. On the server, expired patient appointments were removed to optimize memory space. With the list of consented patients, we deployed a third SQL script that generated a list of patients who had an appointment in the clinical units each day. This script also generated a unique QR code for each patient to associate the collected data with the patient's medical records on the local server and in REDCap.

With the list of consented patients, the research nurse scheduled data collection appointments by prioritizing patients with outcomes of interest to our study, and also inconsideration of

XSL•FO
RenderX

where each patient was located and the providers' preference for data collection should be carried out. Scheduling accounted for travel time between appointment rooms and clinical units and the time taken to prepare for and wrap up data collection. This routine enabled the research nurse to carry data collection in multiple clinical units in a single day.

Enable Data Management, Transfer, and Deidentification Through Generating Unique QR Codes

As is standard for most data collection studies, the study workflow must culminate in generating a fully deidentified database. Alongside generating the QR code, a unique research examination ID that is not present in the patient's medical record is generated as a function for identifying collected data without associating it with any protected health information (PHI).

At the beginning of each data collection appointment, a research nurse would scan the unique QR code that would route the collected ultrasound scans to the appropriate folder on the local server to be stored. The folders on the server are labeled using the unique examination ID. This system allows the QR code to serve as a means for identification and reference as we link deidentified patient files with the associated ultrasound scans. Aside from a single "master list" linking the patient's "real" PHI information to their study information (as is standard in data collection studies), routing and parsing relevant data points are fully automated and rely on research-specific identifiers (unique examination ID) and QR codes, thus reducing the risk of study documentation being lost or otherwise mishandled. This contrasts current, manual approaches that oftentimes rely on names or date of births written on documents to associate data to a patient, which subsequently must be shredded and disposed of.

Results

Between December 1, 2021, and June 13, 2023 (19 months), we achieved our study goal of enrolling 1000 patients for the study. Our automated EDW queries identified 20,988 eligible patient appointments across 3 obstetric clinics throughout that time period. Each query was automatically run every 24 hours and generated between 1000 and 2000 eligible patient appointments.

Logistic constraints including limited study personnel (1 research nurse on the project) and limited work schedule availability to collect data resulted in only approximately 50.4% (10,582/20,988) of identified eligible patients receiving an invitation to participate in the study. We also restricted invites to 2 or 3 potential patients with overlapping appointment times across all clinical sites. Data collection appointments were only scheduled on weekdays between 8 AM and 4 PM.

Based on these constraints, personalized email invitations were subsequently issued to 10,582 potential participants. We enrolled 1000 patients, yielding a 9.5% enrollment rate over 19 months. Among those who were enrolled in the study, 779 patients completed the research study visit within the reported time frame. The difference between enrolled patients and those with complete data collection is a result of some consented patients not having completed their study appointments within our reported time frame. This is largely attributed to the rescheduling of data collection to a subsequent appointment to accommodate the time or logistic constraints of our 1 research nurse performing study data collection. During the study interval reported here, a total of 549 unique individual patients participated in the study; some patients participated more than once during their pregnancy.

Table 1 presents the descriptive demographics of participating patients for race, ethnicity, age, gestational age, and insurance payor. The patient recruitment and enrollment process are summarized in Figure 2.

XSL•FO
RenderX

Table 1. Demographic data of the consented patient population at the time of study enrollment.

	Study population (raw), n (%)	Institution population (raw), n (%)
Race		
Asian	44 (8.0)	3958 (7.2)
Black	74 (13.5)	5452 (9.9)
White	284 (51.7)	35,930 (65.1)
Other	40 (7.3)	3363 (6.1)
Not specified	107 (19.5)	8028 (14.5)
Ethnicity		
Hispanic	80 (14.6)	11,403 (20.6)
Non-Hispanic	396 (72.1)	40,665 (73.6)
Not specified	73 (13.3)	3160 (5.7)
Age group (years)		
<20	4 (0.7)	1309 (2.4)
20-30	101 (18.4)	16,770 (30.4)
30-40	404 (73.6)	34,474 (62.4)
>40	40 (7.3)	2675 (4.8)
Gestational age (years)		
<16	2 (0.3)	794 (2.8)
16-20	167 (16.7)	3781 (18.2)
21-25	163 (16.3)	2666 (12.9)
26-30	196 (19.6)	3120 (15)
31-35	315 (31.5)	5502 (26.5)
36-40	157 (15.7)	4693 (22.6)
>40	0 (0)	183 (0.8)
Insurance payor		
Public	126 (23)	1286 (33.1)
Private	423 (77)	2595 (66.9)

Figure 2. Overview of patient recruitment data across the recruitment and enrollment process.

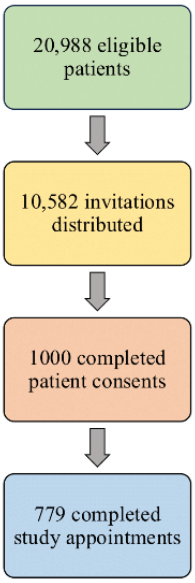

XSL•FO
RenderX

Discussion

Principal Findings

Overall, relying on digital adjuncts in this study helped us to achieve our patient enrollment goal in less than 19 months (as compared to an initial projection of 24 months) while optimizing operational costs and minimizing disruptions to participating clinical sites. For this study, using SQL automation within the EDW allowed us to accurately identify and regularly update lists of eligible participants without ongoing human input or intervention. Digitizing recruitment and consent expanded our reach to many more eligible participants than 1 research nurse could have realistically approached via in-person recruitment, thereby shifting the research nurse's efforts and availability to collecting ultrasound scans and ensuring a positive patient experience participating in the study. Finally, using a digital interface simplified and contained data management for the study to a single server, which reduced the human burden of data management and consequently improved data integrity.

Strengths

Consistent with the existing literature, we found digitizing clinical research processes to be effective in attaining goal recruitment rates, minimizing disruptions to clinic workflows, and ensuring transparency of information regarding research studies [9,19,20]. We also experienced the benefits of integrating the technological capabilities of the electronic medical record for broad-based identification of eligible participants [21-23].

Digitizing patient recruitment and enrollment allowed us the flexibility to strategically target patients with specific outcomes or appointments in specific locations. This also meant that our recruitment data were available to be analyzed frequently, thereby providing us with real-time feedback to improve our recruitment methods. Based on this feedback, we adjusted the time interval between sending an invitation and the scheduled appointment date to optimize participant consent rates.

Reducing aspects of clinical research that are highly dependent on human labor helped to alleviate financial constraints that have previously crippled clinical research and instead focused efforts on participant experience throughout the study [24]. Over 19 months of active patient recruitment with only 1 research nurse, our approach to patient recruitment enabled the research nurse to successfully collect data from patients across 3 clinical units that were in close geographic proximity, completing data collection appointments with up to 5 participants per day. In comparison, a study looking at consent rates using traditional in-person recruitment found that 9 recruiters were needed to approach 2498 patients over a 1-year period in a primary care physician office [24]. If we extended those ratios to our study, we would have needed 18 recruiters over 2 years to approach the 10,582 patients we reached by personalized email invitation.

In clinic settings where multiple clinical research studies are conducted concurrently, front-loading recruitment activities and consent limited disruption to the clinic workflows and minimized patient recruitment fatigue or coercion. We found that limiting research personnel presence in clinics with high volumes of patients and quick patient turnaround increased clinics' willingness to collaborate as recruitment sites for our study.

For patients, shifting recruitment processes onto digital platforms allowed them unlimited time to review holistically all the information pertaining to the study, their participation in the study, and address any data privacy concerns before agreeing to participate in our study. Digitizing patient recruitment inherently standardized the information shared with all eligible participants, giving us confidence in the ethical execution of informed consent.

Finally, our QR code system alleviated pressures on research nurses to maintain cumbersome paperwork associated with data management and instead focus on collecting data and the patient's experience in the study. Additionally, digitizing data management leveraged the compatibility between already digitized patient information (on Epic) and routed relevant study data, thus simplifying the deidentification process of PHI, especially when sharing data with external parties.

Limitations

A limitation of patient recruitment via digital platforms is the potential for selection bias, in our case by restricting recruitment to patients with an associated email address. To mitigate this, research teams may wish to complement digital recruitment activities with telephone or in-person outreach to directly target groups at risk of exclusion.

We recognize that our study was minimally invasive and posed no risk to participants. As such, our results using technological adjuncts for clinical research processes may not be generalizable to clinical research with different study features. Additionally, the scope of this study did not include patient feedback on the recruitment process, and patient experiences with digital recruitment processes are a valuable area for future consideration.

Finally, using digital adjuncts in clinical research inherently means that research teams must include team members well-versed in the technical skills and know-how for setting up digitized recruitment systems and troubleshooting when needed. This option may not be readily available for all research programs or may require consideration of the costs of such expertise.

Conclusions

Digital adjuncts are promising tools that can assist in streamlining patient identification, recruitment, and enrollment in clinical research studies. In our experience, digital adjuncts allowed us to reach out to many more patients than would have been possible with traditional one-on-one, in-person–based recruitment. Additionally, we believe that the benefit of privately reviewing information pertaining to the study without time constraints or any in-person influences optimizes an ethical consent process for patients. Finally, streamlining the data management necessary for our imaging-based observational study has reduced the personnel and resources required for the completion of the study. Using these digital tools could ease research and make it more equitable, leading to higher probabilities of study completion.

XSL•FO
RenderX

Acknowledgments

The study in which this digitized patient recruitment and enrollment process is for is funded by the Bill and Melinda Gates Foundation and Google.

Authors' Contributions

All authors contributed equally to the development of this paper. SC and DM were responsible for building the digitized patient recruitment process. SC, DM, KH, and AS were responsible for providing data for this paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Personalized invitation sent to patients to participate in study.
[DOCX File , 15 KB - formative_v9i1e58628_app1.docx]

References

1. Chaudhari N, Ravi R, Gogtay N, Thatte U. Recruitment and retention of the participants in clinical trials: challenges and solutions. Perspect Clin Res 2020;11(2):64-69 [FREE Full text] [doi: 10.4103/picr.PICR_206_19] [Medline: 32670830]
2. Frank G. Current challenges in clinical trial patient recruitment and enrollment. SoCRA Source. (February) 2004;2(February):30-38. [doi: 10.1093/acprof:oso/9780195035681.003.0014]
3. Ohmann C, Kuchinke W. Meeting the challenges of patient recruitment: a role for electronic health records. Int J Pharm Med 2007;21(4):263-270. [doi: 10.2165/00124363-200721040-00002]
4. Thoma A, Farrokhyar F, McKnight L, Bhandari M. Practical tips for surgical research: how to optimize patient recruitment. Can J Surg 2010;53(3):205-210 [FREE Full text] [Medline: 20507795]
5. Adams M, Caffrey L, McKevitt C. Barriers and opportunities for enhancing patient recruitment and retention in clinical research: findings from an interview study in an NHS academic health science centre. Health Res Policy Syst 2015;13:8 [FREE Full text] [doi: 10.1186/1478-4505-13-8] [Medline: 25971302]
6. McDonald AM, Knight RC, Campbell MK, Entwistle VA, Grant AM, Cook JA, et al. What influences recruitment to randomised controlled trials? A review of trials funded by two UK funding agencies. Trials 2006;7:9 [FREE Full text] [doi: 10.1186/1745-6215-7-9] [Medline: 16603070]
7. Nipp RD, Hong K, Paskett ED. Overcoming barriers to clinical trial enrollment. Am Soc Clin Oncol Educ Book 2019;39:105-114 [FREE Full text] [doi: 10.1200/EDBK_243729] [Medline: 31099636]
8. Briel M, Elger BS, McLennan S, Schandelmaier S, von Elm E, Satalkar P. Exploring reasons for recruitment failure in clinical trials: a qualitative study with clinical trial stakeholders in Switzerland, Germany, and Canada. Trials 2021;22(1):844 [FREE Full text] [doi: 10.1186/s13063-021-05818-0] [Medline: 34823582]
9. Fitzer K, Haeuslschmid R, Blasini R, Altun FB, Hampf C, Freiesleben S, et al. Patient recruitment system for clinical trials: mixed methods study about requirements at ten university hospitals. JMIR Med Inform 2022;10(4):e28696 [FREE Full text] [doi: 10.2196/28696] [Medline: 35442203]
10. Fogel DB. Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: a review. Contemp Clin Trials Commun 2018;11:156-164 [FREE Full text] [doi: 10.1016/j.conctc.2018.08.001] [Medline: 30112460]
11. Goodyear-Smith F, York D, Petousis-Harris H, Turner N, Copp J, Kerse N, et al. Recruitment of practices in primary care research: the long and the short of it. Fam Pract 2009;26(2):128-136. [doi: 10.1093/fampra/cmp015] [Medline: 19251761]
12. Manca DP, O'Beirne M, Lightbody T, Johnston DW, Dymianiw DL, Nastalska K, et al. The most effective strategy for recruiting a pregnancy cohort: a tale of two cities. BMC Pregnancy Childbirth 2013;13:75 [FREE Full text] [doi: 10.1186/1471-2393-13-75] [Medline: 23521869]
13. Brøgger-Mikkelsen M, Ali Z, Zibert JR, Andersen AD, Thomsen SF. Online patient recruitment in clinical trials: systematic review and meta-analysis. J Med Internet Res 2020;22(11):e22179 [FREE Full text] [doi: 10.2196/22179] [Medline: 33146627]
14. Abernethy A, Adams L, Barrett M, Bechtel C, Brennan P, Butte A, et al. The promise of digital health: then, now, and the future. NAM Perspect 2022;2022 [FREE Full text] [doi: 10.31478/202206e] [Medline: 36177208]
15. Rosa C, Marsch LA, Winstanley EL, Brunner M, Campbell AN. Using digital technologies in clinical trials: current and future applications. Contemp Clin Trials 2021;100:106219 [FREE Full text] [doi: 10.1016/j.cct.2020.106219] [Medline: 33212293]
16. Inan OT, Tenaerts P, Prindiville SA, Reynolds HR, Dizon DS, Cooper-Arnold K, et al. Digitizing clinical trials. NPJ Digit Med 2020;3:101 [FREE Full text] [doi: 10.1038/s41746-020-0302-y] [Medline: 32821856]

XSL•FO
RenderX

17. Zayas-Cabán T, Okubo TH, Posnack S. Priorities to accelerate workflow automation in health care. J Am Med Inform Assoc 2022;30(1):195-201 [FREE Full text] [doi: 10.1093/jamia/ocac197] [Medline: 36259967]
18. Zayas-Cabán T, Haque SN, Kemper N. Identifying opportunities for workflow automation in health care: lessons learned from other industries. Appl Clin Inform 2021;12(3):686-697 [FREE Full text] [doi: 10.1055/s-0041-1731744] [Medline: 34320683]
19. Digital clinical trials workshop: creating a vision for the future. National Institutes of Health, National Heart, Lung, and Blood Institute. 2019. URL: https://www.nhlbi.nih.gov/events/2019/digital-clinical-trials-workshop-creating-vision-future [accessed 2023-10-19]
20. Mazurenko O, Sanner L, Apathy NC, Mamlin BW, Menachemi N, Adams MCB, et al. Evaluation of electronic recruitment efforts of primary care providers as research subjects during the COVID-19 pandemic. BMC Prim Care 2022;23(1):95 [FREE Full text] [doi: 10.1186/s12875-022-01705-y] [Medline: 35484491]
21. Embi PJ, Jain A, Harris CM. Physicians' perceptions of an electronic health record-based clinical trial alert approach to subject recruitment: a survey. BMC Med Inform Decis Mak 2008;8:13 [FREE Full text] [doi: 10.1186/1472-6947-8-13] [Medline: 18384682]
22. Dugas M, Lange M, Berdel WE, Müller-Tidow C. Workflow to improve patient recruitment for clinical trials within hospital information systems—a case-study. Trials 2008;9:2 [FREE Full text] [doi: 10.1186/1745-6215-9-2] [Medline: 18186949]
23. Dugas M, Lange M, Müller-Tidow C, Kirchhof P, Prokosch H. Routine data from hospital information systems can support patient recruitment for clinical studies. Clin Trials 2010;7(2):183-189. [doi: 10.1177/1740774510363013] [Medline: 20338903]
24. Felsen CB, Shaw EK, Ferrante JM, Lacroix LJ, Crabtree BF. Strategies for in-person recruitment: lessons learned from a New Jersey Primary Care Research Network (NJPCRN) study. J Am Board Fam Med 2010;23(4):523-533 [FREE Full text] [doi: 10.3122/jabfm.2010.04.090096] [Medline: 20616295]

Abbreviations

EDW: enterprise data warehouse
PHI: protected health information
REDCap: Research Electronic Data Capture

Edited by A Mavragani; submitted 12.04.24; peer-reviewed by Y Wang, JM Kerver; comments to author 28.11.24; revised version received 05.12.24; accepted 06.12.24; published 29.01.25.

Please cite as:
Koh J, Caron S, Watters AN, Vaidyanathan M, Melnick D, Santi A, Hudson K, Arguelles C, Mathur P, Etemadi M
Technological Adjuncts to Streamline Patient Recruitment, Informed Consent, and Data Management Processes in Clinical Research: Observational Study
JMIR Form Res 2025;9:e58628
URL: https://formative.jmir.org/2025/1/e58628
doi:10.2196/58628
PMID:39879093

©Jodie Koh, Stacey Caron, Amber N Watters, Mahesh Vaidyanathan, David Melnick, Alyssa Santi, Kenneth Hudson, Catherine Arguelles, Priyanka Mathur, Mozziyar Etemadi. Originally published in JMIR Formative Research (https://formative.jmir.org), 29.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Understanding Individual Differences in Happiness Sources and Implications for Health Technology Design: Exploratory Analysis of an Open Dataset

Edel Ennis[1*], PhD; Raymond Bond[2*], PhD; Maurice Mulvenna[2], PhD; Colm Sweeney[1], MSc

[1]School of Psychology, Ulster University, Coleraine, United Kingdom

[2]School of Computing, Ulster University, Belfast, United Kingdom

[*]these authors contributed equally

Corresponding Author:
Edel Ennis, PhD
School of Psychology
Ulster University
Cromore Road
Coleraine, BT52 1SA
United Kingdom
Phone: 44 2870123892
Email: e.ennis@ulster.ac.uk

Abstract

Background: Psychologists have developed frameworks to understand many constructs, which have subsequently informed the design of digital mental health interventions (DMHIs) aimed at improving mental health outcomes. The science of happiness is one such domain that holds significant applied importance due to its links to well-being and evidence that happiness can be cultivated through interventions. However, as with many constructs, the unique ways in which individuals experience happiness present major challenges for designing personalized DMHIs.

Objective: This paper aims to (1) present an analysis of how sex may interact with age, marital status, and parental status to predict individual differences in sources of happiness, and (2) to present a preliminary discussion of how open datasets may contribute to the process of designing health-related technology innovations.

Methods: The HappyDB is an open database of 100,535 statements of what people consider to have made them happy, with some people asking to consider the past 24 hours (49,831 statements) and some considering the last 3 months (50,704 statements). Demographic information is also provided. Binary logistic regression analyses are used to determine whether various groups differed in their likelihood of selecting or not selecting a category as a source of their happiness.

Results: Sex and age interacted to influence what was selected as sources of happiness, with patterns being less consistent among female individuals in comparison with male individuals. For marital status, differences in sources of happiness were predominantly between married individuals and those who are divorced or separated, but these were the same for both sexes. Married, single, and widowed individuals were all largely similar in their likelihood of selecting each of the categories as a source of their happiness. However, there were some anomalies, and sex appeared to be important in these anomalies. Sex and parental status also interacted to influence what was selected as sources of happiness.

Conclusions: Sex interacts with age, marital status, and parental status in the likelihood of reporting affection, bonding, leisure, achievement, or enjoying the moment as sources of happiness. The contribution of an open dataset to understanding individual differences in sources of happiness is discussed in terms of its potential role in addressing the challenges of designing DMHIs that are ethical, responsible, evidence based, acceptable, engaging, inclusive, and effective for users. The discussion considers how the content design of DMHIs in general may benefit from exploring new methods informed by diverse data sources. It is proposed that examining the extent to which insights from nondigital settings can inform requirements gathering for DMHIs is warranted.

(JMIR Form Res 2025;9:e65658) doi:10.2196/65658

XSL•FO
RenderX

KEYWORDS

happiness; sexes; age; marital status; parents; affections; achievements; datasets; digital health; well-being; mental health; digital mental health interventions; regression analyses; evidence based

Introduction

Exploring Individual Differences in Sources of Happiness, and the Role of Open Datasets in Health Technology Design

Well-being is founded upon positive emotions, engagement, relationships, meaning, and accomplishments [1]. Different people will derive well-being from each of these 5 building blocks to varying degrees [1]. The premise is that individuals need to focus on positive emotions, engage with life and the activities around them, strive for meaningful relationships and social connections, try to find meaning in life, and have goals and ambitions. Later discussions of the model have concluded life is easier and more enjoyable when our health is optimal. Much literature uses terms such as "happiness," "subjective well-being," "thriving," and "flourishing" interchangeably [2,3]. There is no complete consensus on the relationship between happiness and well-being, but happiness is at least considered to be a component of well-being.

At the very basic level, happiness is an example of a positive emotion that contributes to well-being, with joy, love, and gratitude representing additional examples [4]. The links between happiness and well-being have led to great interest in the science of happiness. This investigates what constitutes happiness; what makes people happy; and what we can do, to ourselves or others, to feel happier [5]. While genetics and life circumstances (eg, finances, job, and material belongings) play roles, 40% of our happiness is determined by behaviors that are under our control [6].

Open data represent a crucial element of the open science movement [7]. These datasets are publicly available, without any restrictions on their use or distribution [7]. The growth of the open science movement has led to an increased availability of open datasets and consideration of how we can reap maximum benefits from these [7]. Example benefits have included transparency and the capacity to amalgamate large datasets, which can give comprehensive insights to inform data-driven decision-making, for example, informing targeted interventions or data-driven resource allocation [7]. Open datasets also allow analysis from a variety of diverse perspectives and thus represent a powerful resource for developing insights into the underlying meaning of psychological constructs [8]. This in turn facilitates enhanced precision and clarity within theories and definitions, an understanding of the meaning of the construct across different cultures and contexts, the development of improved means of measurement, and the development of more effective interventions and treatment options [8].

The HappyDB is an example of a data source that has advanced our understanding of the individual meaning and sources of happiness. The HappyDB is a dataset containing over 100,000 happy statements [9] that is freely available on the web [10]. Some people were asked to consider what had brought them happiness in the past 24 hours, and some were asked to consider the last 3 months. Individuals could make more than 1 entry, but not more than 3 entries. Happiness does not always mean the same thing. Short-term happiness (24 hours) tends to be associated with more daily tasks such as "Exercise," "Nature," and "Leisure," whereas longer-term happiness is more likely to come from loved ones or achievements [9]. Sources of happiness vary cross-culturally [11].

Based on studying happiness within a Latin-American sample, Pena-Lopez et al [12] stressed the need to understand the individual determinants of happiness. Focusing on the words used to express happiness within the HappyDB as opposed to the categories of the sources of happiness, Mohamed and Mostafa [13] used logistic regression, gradient boosting, and the fastText deep learning algorithm, and they found that demographics did influence the meaning and sources of happiness. For example, male individuals expressed greater interest in games and gadgets as sources of happiness, while female individuals discussed family and friends to a greater extent [13]. Parents placed more discussion around the well-being of their families and children, whereas nonparents placed their emphasis on friends, games, eating out, pets, and watching television [13]. Unmarried people focused mostly on dating, friendship, food, and exercise as sources of happiness, whereas married people focused mainly on children and family [13]. Age brackets also differed in how they expressed happiness [13]. Other research studies have also shown differences between demographic groups in the factors associated with happiness levels [14,15].

As noted earlier, sex influences the sources of happiness. Age, marital status, and parenthood status also influence the sources of happiness. The effects of age and marital status would also appear to interact with sex to influence the importance of individual sources of happiness [14]. This is not surprising given that they have been shown to interact in predicting overall levels of happiness or well-being. For example, some research has suggested that the transition to parenthood links to mental well-being, and this differs across the sexes [16,17]. However, other research suggests that the sex gap in the emotional implications of parenthood has dissipated [18]. On a similar note, sex has been found to interact with marital status in predicting levels of happiness [19].

This paper expands on existing research by examining how sex and time frame (last 24 hours vs last 3 months) interact with demographics like age, marital status, and parenthood in predicting sources of happiness. It offers new insights not only by exploring how these factors influence well-being over different time periods but also by considering them in relation to the 5 primary sources of happiness. Mohamed and Mostafa [13] examined the topics of sources of happiness, through natural language processing, and Asai et al [9] identified that statements of happiness can be classified within 7 categories (Table 1). Two of these are endorsed much less frequently than

XSL•FO
RenderX

the others (across 24-hour and 3-month time periods; Table 1), but the five most popular sources of happiness are (1) affection, (2) bonding, (3) leisure, (4) achievement, and (5) enjoying the moment [9]. This research focuses on these 5 most common sources of happiness.

These 5 sources shall be focused upon within this paper. The importance of affection aligns with research showing links between romantic relationships and mental health [20]. The importance of achievement concurs with Walsh et al [21] in their conclusion that happiness is associated with career success [21]. Indeed, while much research is correlational, they conclude that happiness actually often precedes career success, in that positive emotions lead to improved outcomes in the workplace [21].

Table 1. The categories of happy moments: definitions, examples, and frequency of occurrence.

Category	Definition	Examples	Happiness statements within the 24-hour category (n=49,831), n (%)	Happiness statements within the 3-month category (n=50,704), n (%)
Achievement	With extra effort to achieve a better-than-expected result	• Finish work • Complete marathon	15,398 (30.9)	18,507 (36.5)
Affection	Meaningful interaction with family, loved ones, and pets	• Hug • Cuddle • Kiss	16,295 (32.7)	17,797 (35.1)
Bonding	Meaningful interaction with friends and colleagues	• Have meals with coworker • Meet friends	5182 (10.4)	5476 (10.8)
Enjoy the Moment	Being aware or reflecting on present environment	• Have a good time • Mesmerize	6628 (13.3)	4513 (8.9)
Leisure	An activity done regularly in one's free time for pleasure	• Play games • Watch movies • Bake cookies	4335 (8.7)	3093 (6.1)
Nature	In the open air and in nature	• Garden • Beach • Sunset • Weather	1046 (2.1)	761 (1.5)
Exercise	With intent to exercise or workout	• Run • Bike • Do yoga • Lift weights	747 (1.5)	406 (0.8)

Objectives

This paper uses the HappyDB database to examine whether age category, marital status, and parenthood status predict the likelihood of selecting or not selecting each of the 5 most popular sources of happiness (affection, bonding, leisure, achievement, and enjoying the moment) and whether this differs by sex and time frame. Results will be discussed within a preliminary discussion of how open datasets such as HappyDB may represent a novel method to inform the content design of positive psychology interventions (PPIs), and indeed digital mental health interventions (DMHIs) in general. This domain brings great potential, but the ethical and responsible design of DMHIs brings many considerations and calls for creative approaches [22,23]. Best practice guidelines for the development of DMHIs recommend evidence-based approaches; theory-driven design; tailoring the DMHIs to the population and the context; and consideration of engagement, inclusivity, accessibility, and flexibility [22,24,25]. This represents a significant challenge, as collaboration is needed across many disciplines, and open datasets may have valuable contributions to make.

Methods

Materials

There are 100,535 statements within the HappyDB database, with 49,831 statements relating to what made the individual happy in the past 24 hours and 50,704 relating to what made the individual happy within the last 3 months [9]. Table 1 illustrates definitions of each of the 7 categories and examples of the topics discussed within each category. Examples of statements include "My son gave me a big hug in the morning when I woke him up," "I finally managed to make 40 pushups," "I had dinner with my husband," "Morning started with the chirping of birds and the pleasant sun rays," "The event at work was fun. I loved spending time with my good friends and laughing," or "I went to the park with the kids. The weather was perfect" [9].

XSL•FO
RenderX

Ethical Considerations

HappyDB is an open, publicly available dataset that has been used in many other research studies. No personally identifiable information is included in the datasets. Nonetheless, this study received ethical approval from the Filter Committee of the School of Psychology at Ulster University (FCPSYCH2020EE).

Participants

Sample details are outlined in Table 2.

Table 2. Sample details.

Variable	Overall sample (N=100,535), n (%)	24-hour data (n=49,831), n (%)	3-month data (n=50,704), n (%)
Sex			
Male	57,661 (57.4)	28,483 (49.4)	20,937 (49.8)
Female	42,036 (41.8)	29,178 (50.6)	21,099 (50.2)
Marital status			
Married	41,325 (41.2)	20,925 (42.1)	20,400 (40.3)
Divorced or separated	4446 (4.4)	2142 (4.3)	2302 (4.6)
Widowed	477 (0.5)	228 (0.5)	249 (0.5)
Single	54,091 (53.9)	26,441 (53.2)	27,650 (54.6)
Parental status			
No children	60,911 (60.6)	29,700(59.7)	31,211 (61.6)
Children	39,499 (39.3)	20,072 (40.3)	19,427 (38.4)
Age group (years)			
18-25	24,138 (24.1)	11,960 (24)	12,178 (24.1)
26-64	74,932 (74.7)	37,161 (74.7)	37,771 (74.6)
65 and older	1264 (1.3)	609 (1.2)	655 (1.8)

Analyses

Analyses took the form of binary regression analyses. Based on Asai et al [9] noting that sources of happiness varied across the time periods, the 24-hour data and 3-month data were analyzed separately. Analyses also considered male and female individuals separately. As noted earlier, exercise and nature were the least frequently endorsed sources of happiness (Table 1). Due to the small group sizes associated with these 2 categories, analyses focused on considering group differences within the remaining top-5 sources of happiness. Both sex and parental status were treated as dichotomous variables. Age was considered an ordinal variable, as it has a natural order but unequal intervals. Marital status, being categorical with no inherent order or ranking, was classified as nominal.

There were 20 binary logistic regression analyses in all. The overall dataset was separated into four separate datasets: (1) sources of happiness for female individuals over the previous 24 hours; (2) sources of happiness for male individuals over the previous 24 hours; (3) sources of happiness for female individuals over the previous 3 months; and (4) sources of happiness for male individuals over the previous 3 months. Within each of the 4 datasets, 5 binary logistic regression analyses were conducted. Each one of these five regressions considered one of the sources of happiness as a dependent variable: (1) the presence or absence of affection as a source of happiness; (2) the presence or absence of bonding as a source of happiness; (3) the presence or absence of leisure as a source of happiness; (4) the presence or absence of achievement as a source of happiness; and (5) the presence or absence of enjoying the moment as a source of happiness. The analysis examined whether each of these could be predicted by the demographic factors of age category, marital status, and parenthood status (predictors) and whether this differed by sex and time frame. Basically, there were 5 forms of regression, and each one was done within each of the 4 datasets, giving rise to 20 binary logistic regression analyses. For age as an independent variable, 18-25 years represented the reference category. For marital status as a predictor, married individuals represented the reference category. For parental status as a predictor, those with no children represented the reference category. The assumptions for the analyses were addressed as follows.

- Binary logistic regression assumes that observations are independent. While individuals could provide multiple reports about their sources of happiness, each report was treated as a separate data entry, ensuring no repeated measures
- Multicollinearity among independent variables was examined using variance inflation factor (VIF) values. VIF values below 5 indicate low multicollinearity, while values above 10 suggest a high degree of multicollinearity [26]. In this study, all VIF values were below 2, confirming no multicollinearity issues.
- The linearity of the logit: assumption applies to continuous independent variables. Since all independent variables in this study are categorical, this assumption is not relevant.
- Outlier detection is not applicable here due to the categorical nature of all variables.

XSL•FO
RenderX

Results

Affection

Predictors of affection as a source of happiness are outlined in Table 3. The first category (not selecting affection) was the reference category.

Within male individuals, the different age groups did not differ in their likelihood of selecting or not selecting affection as a source of their happiness. Within female individuals, all age groups were significantly more likely than the 18-25 years age group to select affection as a source of their happiness. This was evident within both the 24-hour (*P*<.001) and 3-month (*P*<.01) reports of sources of happiness.

Within both the male and female data, divorced or separated individuals were significantly more likely than married individuals to select affection as a source of their happiness (*P*<.001). However, across both male and female individuals, those who were widowed did not differ from married individuals in their likelihood of selecting affection as a source of happiness. Within male individuals, those who were single did not differ from married individuals in their likelihood of selecting affection as a source of happiness. Within the female individuals, this was the same within the 3-month data. However, within the data where female individuals had reported their sources of happiness within the last 24 hours, single female individuals were significantly (1.38 times; *P*<.05) more likely than married female individuals to select affection as a source of their happiness.

Within both the male and female data, those with children were significantly less likely than those without children to select affection as a source of their happiness (*P*<.001).

Table 3. ORs[a] and significance for logistic regression on choosing "affection" as a source of happiness by demographics and time frame[b].

	Male, OR (95% CI)		Female, OR (95% CI)	
	24-hour data	3-month data	24-hour data	3-month data
Age group (years) with "18-25 years" as reference				
26-64	1.13 (0.89-1.45)	1.14 (0.91-1.44)	2.13 (1.64-2.75)***	1.43 (1.12-1.83)**
65 and older	1.06 (0.84-1.35)	1.17 (0.94-1.46)	1.55 (1.21-1.98)***	1.21 (0.95-1.52)
Marital status with "Married" as reference				
Divorced or separated	1.54 (1.43-1.66)***	1.48 (1.38-1.59)***	1.49 (1.38-1.61)***	1.34 (1.24-1.45)***
Widowed	1.00 (0.85-1.19	1.13 (0.96-1.32)	1.04 (0.91-1.19)	0.98 (0.87-1.11)
Single	1.31 (0.71-2.41)	1.32 (0.81-2.13)	1.38 (1.01-1.88)*	1.26 (0.92-1.71)
Parenthood with "No children" as reference				
Children	0.60 (0.56-0.65)***	0.63 (0.59-0.68)***	0.56 (0.52-0.60)***	0.51 (0.48-0.55)***

[a]OR: odds ratio.

[b]Reference category: no selection of affection.

*P<.05.

**P<.01.

***P<.001.

Bonding

Predictors of bonding as a source of happiness are outlined in Table 4. The first category (not selecting bonding) was the reference category.

For age as a predictor of bonding as a source of happiness, 18-25 years represents the reference category. Within male individuals, the different age groups did not differ in their likelihood of selecting or not selecting bonding as a source of their happiness. Within female individuals, the different age groups did not differ in their likelihood of selecting or not selecting bonding as a source of their happiness over the past 3 months. However, when considering what made them happy over the past 24 hours, female individuals aged 26-64 years were significantly more (1.94 times; *P*<.05) more likely than female individuals aged 18-25 years to select bonding as a source of their happiness. Within reports of happiness in the past 24 hours, female individuals aged 65 years old and older did not differ significantly from those aged 18-25 years in their likelihood of selecting bonding as a source of their happiness.

Within both the male and female data, divorced or separated individuals were significantly less likely than married individuals to select bonding as a source of their happiness (*P*<.001). Within male reports of happiness within the past 24 hours, widowed male individuals were significantly less likely than married male individuals to select bonding as a source of their happiness (*P*<.05). However, widowed and married male individuals did not differ in their likelihood of selecting bonding as a source of their happiness over the past 3 months. Widowed and married female individuals did not differ in their likelihood of selecting bonding as a source of their happiness. Across both male and female individuals, those who were single did not differ from married individuals in their likelihood of selecting bonding as a source of happiness.

Male individuals with and without children did not differ in their likelihood of selecting bonding as a source of their happiness over the past 24 hours. However, within 3-month

XSL•FO
RenderX

reports of sources of happiness, male individuals with children were much more likely to endorse something within the bonding category in comparison with male individuals with no children (*P*<.001). Within female individuals, those with children were much more likely to endorse something within the bonding category in comparison with those with no children (*P*<.001).

Table 4. ORs[a] and significance for logistic regression on choosing "bonding" as a source of happiness by demographics and time frame[b].

	Male, OR (95% CI)		Female, OR (95% CI)	
	24-hour data	3-month data	24-hour data	3-month data
Age group (years) with "18-25 years" as reference				
26-64	0.85 (0.58-1.24)	1.03 (0.69-1.55)	1.94 (1.16-3.24)*	1.03 (0.68-1.58)
65 and older	0.75 (0.52-1.08)	0.98 (0.66-1.46)	1.52 (0.92-2.50)	0.93 (0.62-1.40)
Marital status with "Married" as reference				
Divorced or separated	0.69 (0.62-0.78)***	0.68 (0.61-0.75)***	0.78 (0.69-0.88)***	0.82 (0.72-0.92)***
Widowed	0.78 (0.61-1.00)*	0.83 (0.65-1.06)	0.96 (0.77-1.18)	1.13 (0.93-1.38)
Single	0.83(0.32-2.11)	1.10 (0.55-2.24)	1.20 (0.73-1.96)	0.88 (0.51-1.51)
Parenthood with "No children" as reference				
Children	1.07 (0.96-1.20)	1.24 (1.11-1.39)***	1.26 (1.11-1.42)***	1.61 (1.43-1.82)

[a]OR: odds ratio.

[b]Reference category: no selection of affection.

*P<.05.

**P<.01.

***P<.001.

Leisure

Predictors of leisure as a source of happiness are outlined in Table 5. The first category (not selecting leisure) was the reference category.

Table 5. ORs[a] and significance for logistic regression on choosing "leisure" as a source of happiness by demographics and time frame[b].

	Male, OR (95% CI)		Female, OR (95% CI)	
	24-hour data	3-month data	24-hour data	3-month data
Age group (years) with "18-25 years" as reference				
26-64	1.31 (0.80-2.15)	0.85 (0.53-1.36)	0.26 (0.17-0.38)***	0.44 (0.25-0.78)**
65 and older	1.40 (0.86-2.28)	0.94 (0.59-1.49)	0.44(0.30-0.63)***	0.85(0.49-1.47)
Marital status with "Married" as reference				
Divorced or separated	1.18 (1.05-1.31)**	1.08 (0.95-1.22)	0.95 (0.83-1.09)	1.27 (1.08-1.49)**
Widowed	1.24 (0.97-1.58)	0.94 (0.70-1.27)	0.95 (0.75-1.21)	1.28 (0.98-1.68)
Single	3.11 (1.48-6.53)**	0.95 (0.35-2.63)	0.31 (0.13-0.77)**	0.26 (0.06-1.04)
Parenthood with "No children" as reference				
Children	1.84 (1.63-2.07)	1.56 (1.36-1.78)***	1.87 (1.63-2.15)***	2.35 (2.00-2.75)***

[a]OR: odds ratio.

[b]Reference category: no selection of affection.

*P<.05.

**P<.01.

***P<.001.

Results show that within the male individuals of the sample, the different age groups did not differ in their likelihood of selecting or not selecting leisure as a source of their happiness. Within the female individuals of the sample, those aged 26-64 years and those aged 65 years and older were both significantly less likely than those aged 18-25 years to select leisure as a source of their happiness (*P*<.001). This was evident in reports of happiness over the past 3 months and the past 24 hours.

In relation to marital status, divorced or separated male individuals were significantly more (1.18 times; *P*<.01) likely than married male individuals to cite leisure as a source of their happiness over the past 24 hours. However, divorced or

XSL•FO
RenderX

separated and married male individuals did not differ in their likelihood of citing leisure as a source of their happiness over the past 3 months. The pattern was reversed among female individuals. Specifically, divorced or separated and married female individuals did not differ in their likelihood of citing leisure as a source of their happiness over the past 24 hours. However, divorced or separated female individuals were significantly more (1.27 times; *P*<.01) likely than married female individuals to cite leisure as a source of their happiness over the past 3 months.

Male individuals with and without children did not differ in their likelihood of selecting leisure as a source of their happiness over the past 24 hours. However, within 3-month reports of sources of happiness, male individuals with children were much more likely to endorse something within the leisure category in comparison with male individuals with no children (*P*<.001). Within female individuals, those with children were much more likely to endorse something within the leisure category in comparison with those with no children (*P*<.001).

Achievement

Predictors of achievement as a source of happiness are outlined in Table 6. The first category (not selecting achievement) was the reference category.

Results show that within the male individuals of the sample, the different age groups did not differ in their likelihood of selecting or not selecting achievement as a source of their happiness. Within the female individuals of the sample, those aged 26-64 years were significantly less likely than those aged 18-25 years to select achievement as a source of their happiness. This was evident in reports of happiness over the past 3 months (*P*<.05) and the past 24 hours (*P*<.001). However, female individuals aged 65 years and older only differed significantly from those aged 18-25 years in terms of them being less likely (0.74 times; *P*<.05) to cite achievement as a source of their happiness over the past 24 hours. The 2 groups did not differ significantly in their likelihood of endorsing achievement as a source of their happiness over the past 3 months.

In relation to marital status, those who were divorced or separated (male and female individuals) were significantly less likely than their counterparts to select achievement as a source of their happiness (*P*<.001). This was evident within 24-hour reports and those of the past 3 months. Within the 24-hour and 3-month reports, widowed individuals did not differ from their married counterparts in the likelihood of selecting achievement as a source of their happiness. Compared to married male individuals, single male individuals were significantly less (0.39 times; *P*<.05) likely to cite achievement as a source of their happiness over the past 24 hours. However, single and married female individuals did not differ significantly in their likelihood of endorsing achievement as a source of their happiness over the past 3 months. Single and married male individuals did not differ significantly in their likelihood of endorsing achievement as a source of their happiness over the past 24 hours or 3 months.

Within male and female individuals, and within the 24-hour reports and the 3-month reports, those with children were significantly more likely than those without children to select achievement as a source of their happiness (*P*<.001).

Table 6. ORs[a] and significance for logistic regression on choosing "achievement" as a source of happiness by demographics and time frame[b].

	Male, OR (95% CI)		Female, OR (95% CI)	
	24-hour data	3-month data	24-hour data	3-month data
Age group (years) with "18-25 years" as reference				
26-64	0.96 (0.74-1.25)	0.88 (0.70-1.11)	0.64 (0.49-0.84)***	0.73 (0.57-0.94)*
65 and older	1.02 (0.80-1.32)	0.84 (0.67-1.05)	0.74 (0.58-0.96)*	0.81 (0.63-1.03)
Marital status with "Married" as reference				
Divorced or separated	0.83 (0.77-0.89)***	0.82 (0.77-0.88)***	0.83 (0.76-0.90)***	0.76 (0.70-0.82)***
Widowed	0.99 (0.84-1.15)	1.08 (0.93-1.25)	1.12 (0.97-1.28)	0.90 (0.80-1.03)
Single	0.39 (0.18-0.83)*	0.92 (0.57-1.49)	1.08 (0.78-1.50)	0.84 (0.61-1.16)
Parenthood with "No children" as reference				
Children	1.17 (1.09-1.26)***	1.16 (1.08-1.24)***	1.30 (1.20-1.41)***	1.24 (1.15-1.34)***

[a]OR: odds ratio.

[b]Reference category: no selection of affection.

*P<.05.

**P<.01.

***P<.001.

Enjoy the Moment

Predictors of enjoying the moment as a source of happiness are outlined in Table 7. The first category (not selecting enjoying the moment) was the reference category.

Within male and female individuals, the age categories did not differ in their likelihood of selecting enjoying the moment as a source of their happiness. This was evident within both the 24-hour and 3-month reports of sources of happiness. Within the 24-hour data, divorced or separated male and female individuals were significantly less likely than their counterparts

XSL•FO
RenderX

to cite enjoying the moment as a source of their happiness over the past 24 hours (*P*<.001). However, within the 3-month data, divorced or separated male and female individuals did not differ significantly from their counterparts in their likelihood of citing enjoying the moment as a source of their happiness over the past 24 hours. Within male and female individuals, neither the widowed individuals nor single individuals differed from their married counterparts in their likelihood of selecting enjoying the moment as a source of their happiness. This was evident within both the 24-hour and 3-month reports of sources of happiness.

Within 24-hour reports of happiness, those with children (male and female individuals) did not differ from their counterparts in their likelihood of selecting enjoying the moment as a source of their happiness. However, within 3-month reports of happiness, those with children (male individuals: *P*<.01 and female individuals: *P*<.001) were significantly more likely than their counterparts in their likelihood of selecting enjoying the moment as a source of their happiness.

Table 7. ORs[a] and significance for logistic regression on choosing "enjoying the moment" as a source of happiness by demographics and time frame[b].

	Male, OR (95% CI)		Female, OR (95% CI)	
	24-hour data	3-month data	24-hour data	3-month data
Age group (years) with "18-25 years" as reference				
26-64	1.29 (0.89-1.87)	1.54 (0.95-2.52)	0.92 (0.63-1.34)	141 (0.87-2.27)
65 and older	1.26 (0.88-1.82)	1.57 (0.97-2.55)	0.94 (0.65-1.35)	1.13 (0.71-1.80)
Marital status with "Married" as reference				
Divorced or separated	0.82 (0.75-0.91)***	0.97 (0.87-1.09)	0.79 (0.71-0.89)***	1.00 (0.88-1.14)
Widowed	1.02 (0.83-1.26)	0.79 (0.60-1.04)	0.86 (0.71-0.89)	1.17 (0.95-1.46)
Single	1.58 (0.78-3.21)	0.46 (0.15-1.47)	0.70 (0.42-1.15)	1.39 (0.83-2.32)
Parenthood with "No children" as reference				
Children	0.97 (0.88-1.07)	1.20 (1.07-1.35)**	1.07 (0.95-1.19)	1.31 (1.15-1.49)***

[a]OR: odds ratio.

[b]Reference category: no selection of affection.

*$P<.05$.

**$P<.01$.

***$P<.001$.

Discussion

Principal Findings

This paper set out to (1) present an analysis of how sex may interact with age, marital status, and parental status to predict individual differences in sources of happiness; and (2) present a preliminary discussion of how open datasets may contribute to the process of designing health-related technology innovations. In terms of a brief summary of findings, sex interacted with age, marital status, and parental status in predicting sources of happiness, with some patterns varying across time periods.

Earlier research has suggested that sex interacts with other demographic factors to predict individual differences in sources of happiness [14]. Results partially supported this suggestion. First, sex and age interacted in predicting sources of happiness. Clear trends were evident in the data for male individuals. Male individuals across the 3 age categories did not differ in their likelihood of reporting affection, bonding, leisure, achievement, or enjoying the moment as sources of happiness. Trends were not as consistent within the female data, with the time period under consideration also exerting significant effects on the likelihood of affection, bonding, leisure, and achievement being reported as a source of happiness. The exact direction of effects is described in the *Results* section. The only exception was in relation to enjoying the moment as a source of happiness wherein, like the male group, the different female age groups did not differ in their likelihood of reporting enjoying the moment as a source of happiness. This was regardless of the time period under consideration. Results can therefore be said to only partially agree with a previous review by Buijs et al [27], which concluded that chronological age did not relate to affection, status, and other behavioral factors as sources of happiness.

Results partially tied in with earlier conclusions that sex interacts with marital status in predicting levels of happiness [19]. Across the board, differences in sources of happiness were predominantly between married individuals and those who are divorced or separated. However, these differences were the same for both sexes. Married, single, and widowed individuals were all largely similar in their likelihood of selecting each of the categories as a source of their happiness. However, there were some anomalies and sex appeared to be important in these anomalies. For example, in comparison to married male individuals, single male individuals were less likely to report achievement as a source of their happiness in the previous 24 hours. On the contrary, married and single female individuals did not differ in their likelihood of reporting achievement as a source of their happiness in the previous 24 hours. Compared to married male individuals, single male individuals were significantly more likely to report leisure as a source of their

XSL•FO
RenderX

happiness in the past 24 hours. Conversely, in comparison with married female individuals, single female individuals were significantly less likely to report leisure as a source of happiness in the previous 24 hours. Such patterns were not evident in the 3-month data. It must be acknowledged that the current study only looks at marital status, whereas other factors such as commitment are also important [20]. Understanding how marital status ties to happiness is important, as there is a bidirectional relationship between relationships and mental health, which is particularly strong when relationship status is used as a predictor of health [20].

Parental status and sex also appeared to interact in predicting some, but not all, sources of happiness. Compared to those who were not parents, those who were parents were significantly less likely to consider affection as a source of their happiness and significantly more likely to consider achievement as a source of their happiness. However, differences were apparent within both the male and female data, and across both the 24-hour and 3-month time periods. Compared to male individuals who were not parents, those male individuals who were parents were significantly more likely to consider enjoying the moment as a source of their happiness over the past 3 months, but not the past 24 hours. This exact same pattern was evident in the female data. However, as can be seen from the results, the sexes did differ in how they reported bonding and leisure as sources of happiness. Some past research has suggested sex differences in how the transition to parenthood links to mental well-being [16,17], and some have suggested that the sex gap in the emotional implications of parenthood has dissipated [18]. The current findings are consistent with the inconclusive nature of past research, and the recognized need to consider the complexity of how and why parents experience happiness [28].

Limitations

Limitations of course must be considered. As the HappyDB data were gathered via crowdsourcing, the representativeness of the sample is unknown. Obviously, data collection was dependent upon access to technology. Many other demographic factors link to happiness, for example, health and socioeconomic status. Sample details are not available for factors such as socioeconomic status or health conditions, but the US background and the limited representation of older participants are clear limitations in terms of the cultural and age group representativeness of this analysis. Open datasets often lack details on psychological factors, such as personality traits, that influence individual differences in happiness [29]. To maximize their use in developing DMHIs, researchers must balance the breadth and depth of data. While large sample sizes offer generalizable insights and scalability, incorporating psychometric measures (eg, personality or emotional well-being scales), and providing richer, more personalized data. A balanced approach that integrates demographic diversity with targeted, detailed subsamples would ensure both robust and effective DMHIs. Similarly, within the Positive Emotions, Engagement, Relationships, Meaning, Accomplishment, and Health model, happiness is just 1 example of positive emotion, and many other components link to well-being (eg, engagement, relationships, meaning, accomplishments, and health). This paper analyzes each source of happiness independently. However, exploring correlations between sources within demographics could offer deeper insights. For instance, responses about affection, bonding, leisure, achievement, or enjoying the moment might reflect the same “event” over different time frames, with patterns varying by demographic. It would be interesting to examine how many individuals fall into the same happiness category for both the 24-hour and 3-month statements by demographic group, and whether certain demographics are more likely to report consistent happiness across time frames.

Conclusions

There currently are no best practice guidelines to support developers and designers in gathering and defining the stakeholder-driven or humanity-centric requirements, use cases, and goals for DMHIs in select domains. The ethical and responsible design of DMHIs should be evidence-based approaches; use theory-driven design; be tailored to the population and the context; and be considerate of engagement, inclusivity, accessibility, and flexibility, but this represents a significant challenge [22,24,25]. However, the identification of universal and domain-specific needs is also central to user engagement and retention. Coproduction would typically include interviews, focus groups, literature reviews, workshops, panel discussions, and clinical expertise [23]. Of course, such processes are very time, and as such resource intensive [23].

Within their systematic review, Brotherdale et al [23] note the need for creative methods to be adopted in the coproduction of digital mental health intervention, and the use of open datasets may be a fruitful approach in this endeavor. Currently, little attention has been paid to open datasets as a creative approach to requirements gathering and informing content development of PPIs, and indeed DMHIs in general. The HappyDB analysis can be presented as a sample application.

PPIs can be tailored to address diverse challenges and promote well-being, as well as prevent mental ill health by promoting protective factors, enhancing resilience, and fostering well-being [30]. PPIs often involve suggested activities to enhance happiness or well-being [9]. PPIs generally focus on one of five categories: (1) savoring experiences and sensations; (2) cultivating (and sometimes expressing) gratitude; (3) engaging in kind acts; (4) promoting positive relationship processes; and (5) pursuing hope and meaning [31]. Their mode of operation is typically either diaries for self-reflection or applications that make suggestions based on responses to set questions [9]. Some examples relevant to this discussion might be virtual reality interventions that seek to promote happiness by immersing users in positive environments or guided meditations; strength-based platforms that help users to identify and use their personal strengths; positive psychology games that are designed to promote positive emotions; or gratitude apps where users can keep a digital gratitude journal, but there would also be reminders and prompts to help them focus on positive aspects of their lives.

User engagement difficulties across the board represent a major challenge to PPIs, and indeed DMHIs in general, achieving their maximum potential benefits [32]. Designing content that can be personalized based on user data, preferences, and progress

XSL•FO
RenderX

to increase engagement and effectiveness represents a major challenge for design developers, with collaboration across several disciplines required [32]. However, as can be seen earlier, notwithstanding the recognized limitations, open datasets such as the HappyDB provide great insight into how demographics link to differences in sources of happiness. Certain PPIs or suggested activities may have increased meaning for certain subgroups, and open datasets may provide a starting point for understanding what these might be. Research on the science of happiness is also relevant to the concept of intrinsic capacity for healthy aging and holistic well-being. This concept remains undervalidated, and a deeper understanding of happiness could help validate and practically apply it in mental health interventions [33]. Open datasets may also come in various formats. For example, there are many sources that outline counseling dialogues, and how these are rated by users. Through the use of machine learning, these may give insights into useful content generation for DMHIs.

When discussing DMHI content design, it is important to consider the lack of context in interpreting the significance of insights derived from open-source data. This is a key limitation in health technology design. Evaluating user satisfaction and engagement is essential to determine whether a version created using open data is received more positively than one developed through other approaches. Metrics such as usability, satisfaction, and effectiveness should be compared to assess the value of incorporating open data in DMHI development.

Many platforms host open datasets, with the Open Science Framework, Mendeley Data, PsychOpen Community, Archive, and Metadata for Open Science, or Kaggle representing examples. Nonetheless, supporting researchers in providing findable, accessible, interoperable, and reusable data poses significant challenges [8]. For example, what might a framework to describe the nature of the data that are typically gathered across different domains look like? What would the best methods be for structuring data (and metadata) that are human and machine readable but also useful for designers and analysts and for interpretability across different channels? How would datasets be described in terms of purposes, goals, and communication patterns, and what might a structured representational repository look like?

Huston [7] discusses the challenges associated with open data in terms of three categories: (1) making the technological shift; (2) social, cultural, legal, and ethical issues; and (3) avoiding the pitfalls. Nonetheless, a discussion of the extent to which our understanding of data and interactions from nondigital settings may inform requirement gathering for DMHIs is warranted. While the challenges associated with open data are substantial and should not be underestimated, the potential benefits of open data are very exciting [7]. They may play a role in the development of content for digital PPIs that are engaging, effective, and grounded in scientific research.

Data Availability

The HappyDB is a dataset containing over 100,000 happy statements [9] that is freely available on the web [10]. The data set analyzed during this study is available on the Megagon Labs website [34].

Conflicts of Interest

None declared.

References

1. Seligman M. PERMA and the building blocks of well-being. J Posit Psychol 2018;13(4):333-335. [doi: 10.1080/17439760.2018.1437466]
2. Butler J, Kern ML. The PERMA-Profiler: a brief multidimensional measure of flourishing. Int J Wellbeing 2016;6(3):1-48. [doi: 10.5502/ijw.v6i3.526]
3. Medvedev ON, Landhuis CE. Exploring constructs of well-being, happiness and quality of life. PeerJ 2018;6:e4903 [FREE Full text] [doi: 10.7717/peerj.4903] [Medline: 29876148]
4. Seligman MEP. Flourish: A Visionary New Understanding of Happiness and Well-Being. Mumbai, India: Free Press; 2011.
5. Seligman MEP. Positive psychology: a personal history. Annu Rev Clin Psychol 2019;15:1-23. [doi: 10.1146/annurev-clinpsy-050718-095653] [Medline: 30525996]
6. Diener E, Suh EM, Lucas RE, Smith HL. Subjective well-being: three decades of progress. Psychol Bull 1999;125(2):276-302. [doi: 10.1037//0033-2909.125.2.276]
7. Huston P, Edge VL, Bernier E. Reaping the benefits of open data in public health. Can Commun Dis Rep 2019;45(11):252-256 [FREE Full text] [doi: 10.14745/ccdr.v45i10a01] [Medline: 31647060]
8. Sadeh Y, Denejkina A, Karyotaki E, Lenferink LIM, Kassam-Adams N. Opportunities for improving data sharing and FAIR data practices to advance global mental health. Glob Ment Health 2023;10:e14 [FREE Full text] [doi: 10.1017/gmh.2023.7] [Medline: 37860102]
9. Asai A, Evensen S, Golshan B, Halevy A, Li V, Lopatenko A, et al. HappyDB: A corpus of 100,000 crowdsourced happy moments. 2018 Presented at: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018); May 1, 2018; Miyazaki, Japan URL: https://arxiv.org/abs/1801.07746

XSL•FO
RenderX

10. HappyDB: a happiness database of 100,000 happy moments. GitHub. URL: https://github.com/megagonlabs/HappyDB [accessed 2025-01-07]
11. Oishi S, Gilbert EA. Current and future directions in culture and happiness research. Curr Opin Psychol 2016;8:54-58. [doi: 10.1016/j.copsyc.2015.10.005] [Medline: 29506804]
12. Pena-López A, Rungo P, López-Bermúdez B. The "Efficiency" effect of conceptual referents on the generation of happiness: a cross-national analysis. J Happiness Stud 2020;22(6):2457-2483. [doi: 10.1007/s10902-020-00328-3]
13. Mohamed E, Mostafa SA. Computing happiness from textual data. Stats 2019;2(3):347-370. [doi: 10.3390/stats2030025]
14. Alkan Ö, Kavalci K. Gender differences in happiness during COVID-19 pandemic: evidence from Türkiye. J Psychiatry Behav Sci 2023;13(3):153-165. [doi: 10.5455/pbs.20230512053548]
15. Karaaslan K. Analysis of factors affecting individuals' sources of happiness with multinomial logistic model. J Meas Eval Educ Psychol 2021;12(3):286-302 [FREE Full text] [doi: 10.21031/epod.925631]
16. Metzger S, Gracia P. Gender differences in mental health following the transition into parenthood: longitudinal evidence from the UK. Adv Life Course Res 2023;56:100550. [doi: 10.1016/j.alcr.2023.100550] [Medline: 38054884]
17. Nomaguchi K, Milkie MA. Parenthood and well-being: a decade in review. J Marriage Fam 2020;82(1):198-223 [FREE Full text] [doi: 10.1111/jomf.12646] [Medline: 32606480]
18. Preisner K, Neuberger F, Bertogg A, Schaub JM. Closing the happiness gap: the decline of gendered parenthood norms and the increase in parental life satisfaction. Gender Society 2019;34(1):31-55. [doi: 10.1177/0891243219869365]
19. Maharlouei N, Cobb S, Bazargan M, Assari S. Subjective health and happiness in the United States: gender differences in the effects of socioeconomic status indicators. J Ment Health Clin Psychol 2020;4(2):8-17 [FREE Full text] [doi: 10.29245/2578-2959/2020/2.1196] [Medline: 32568256]
20. Braithwaite S, Holt-Lunstad J. Romantic relationships and mental health. Curr Opin Psychol 2017;13:120-125. [doi: 10.1016/j.copsyc.2016.04.001] [Medline: 28813281]
21. Walsh LC, Boehm JK, Lyubomirsky S. Does happiness promote career success? Revisiting the evidence. J Career Assess 2018;26(2):199-219. [doi: 10.1177/1069072717751441]
22. van Zyl LE, Efendic E, Rothmann S, Shankland R. Best-practice guidelines for positive psychological intervention research design. In: van Zyl L, Rothmann S, editors. Positive Psychological Intervention Design and Protocols for Multi-Cultural Contexts. Cham, Switzerland: Springer; 2019.
23. Brotherdale R, Berry K, Branitsky A, Bucci S. Co-producing digital mental health interventions: a systematic review. Digit Health 2024;10:20552076241239172 [FREE Full text] [doi: 10.1177/20552076241239172] [Medline: 38665886]
24. Berardi C, Antonini M, Jordan Z, Wechtler H, Paolucci F, Hinwood M. Barriers and facilitators to the implementation of digital technologies in mental health systems: a qualitative systematic review to inform a policy framework. BMC Health Serv Res 2024;24(1):243 [FREE Full text] [doi: 10.1186/s12913-023-10536-1] [Medline: 38408938]
25. Lipschitz JM, Pike CK, Hogan TP, Murphy SA, Burdick KE. The engagement problem: a review of engagement with digital mental health interventions and recommendations for a path forward. Curr Treat Options Psychiatry 2023;10(3):119-135 [FREE Full text] [doi: 10.1007/s40501-023-00297-3] [Medline: 38390026]
26. Alkan Ö, Güney E. Investigation of factors that affect the frequency of alcohol use of employees in Turkey. J Subst Use 2020;26(5):468-474. [doi: 10.1080/14659891.2020.1846811]
27. Buijs VL, Jeronimus BF, Lodder G, Steverink N, de Jonge P. Social needs and happiness: a life course perspective. J Happiness Stud 2020;22(2):1953-1978. [doi: 10.1007/s10902-020-00287-9]
28. Nelson SK, Kushlev K, Lyubomirsky S. The pains and pleasures of parenting: When, why, and how is parenthood associated with more or less well-being? Psychol Bull 2014;140(3):846-895. [doi: 10.1037/a0035444] [Medline: 24491021]
29. Abdullahi AM, Orji R, Rabiu AM, Kawu AA. Personality and subjective well-being: towards personalized persuasive interventions for health and well-being. Online J Public Health Inform 2020;12(1):e1 [FREE Full text] [doi: 10.5210/ojphi.v12i1.10335] [Medline: 32547678]
30. Carr A, Cullen K, Keeney C, Canning C, Mooney O, Chinseallaigh E, et al. Effectiveness of positive psychology interventions: a systematic review and meta-analysis. J Posit Psychol 2021;16(6):749-769. [doi: 10.1080/17439760.2020.1818807]
31. Schueller SM, Parks AC. The science of self-help: translating positive psychology research into increased individual happiness. Eur Psychol 2014;19(2):145-155. [doi: 10.1027/1016-9040/a000181]
32. Saleem M, Kühne L, de Santis KK, Christianson L, Brand T, Busse H. Understanding engagement strategies in digital interventions for mental health promotion: scoping review. JMIR Ment Health 2021;8(12):e30000 [FREE Full text] [doi: 10.2196/30000] [Medline: 34931995]
33. Hoogendijk EO, Dent E, Koivunen K. Intrinsic capacity: An under-researched concept in geriatrics. Age Ageing 2023;52(10):afad183. [doi: 10.1093/ageing/afad183] [Medline: 37782890]
34. HappyDB: a happiness database of 100,000 happy moments. Megagon Labs. URL: https://megagon.ai/happydb-a-happiness-database-of-100000-happy-moments/ [accessed 2025-01-27]

Abbreviations

DMHI: digital mental health intervention

XSL•FO
RenderX

PPI: positive psychology intervention
VIF: variance inflation factor

Edited by A Mavragani; submitted 21.08.24; peer-reviewed by Ö Alkan, N Hird; comments to author 10.12.24; revised version received 17.12.24; accepted 18.12.24; published 29.01.25.

Please cite as:
Ennis E, Bond R, Mulvenna M, Sweeney C
Understanding Individual Differences in Happiness Sources and Implications for Health Technology Design: Exploratory Analysis of an Open Dataset
JMIR Form Res 2025;9:e65658
URL: https://formative.jmir.org/2025/1/e65658
doi:10.2196/65658
PMID:

©Edel Ennis, Raymond Bond, Maurice Mulvenna, Colm Sweeney. Originally published in JMIR Formative Research (https://formative.jmir.org), 29.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Experiences of Peer Mentoring Sexual and Gender Minority Emerging Adults Who Are at Risk for Suicide: Mixed Methods Study

Jennifer T Tran[1], PhD; Jessica Webster[1], MSc; James R Wolfe[1], MSc; Jennifer Ben Nathan[1], BA; Lindiwe Mayinja[2], BS; Marin Kautz[2], PhD; Maria A Oquendo[2], MD, PhD; Gregory K Brown[2], PhD; David Mandell[2], ScD; Danielle Mowery[3], PhD; José A Bauermeister[1], MPH, PhD; Lily A Brown[2], PhD

[1]Department of Family and Community Health, School of Nursing, University of Pennsylvania, Philadelphia, PA, United States

[2]Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States

[3]Department of Biostatistics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States

Corresponding Author:
Jennifer T Tran, PhD
Department of Family and Community Health
School of Nursing
University of Pennsylvania
418 Curie Blvd
Philadelphia, PA, 19104
United States
Phone: 1 2158983616
Fax: 1 215 573 8439.
Email: jtgtran@nursing.upenn.edu

Abstract

Background: Sexual and Gender Diverse Youth (SGDY) are at increased risk for suicide due to unique experiences including discrimination, family or friend rejection, and low positive affect. Peer mentors (PMs) may offer a unique opportunity for intervention but are underutilized for suicide prevention among SGDY.

Objective: Little is known about the training needed for PMs when working with SGDY at risk for suicide. We developed an intervention, Supporting Transitions to Adulthood and Reducing Suicide (STARS), to improve suicide prevention among SGDY and increase social support, coping, and positive effects. PMs were trained by a licensed clinical therapist and provided a manual. PMs meet virtually for 6 weeks, providing social support, strategies to diminish the impact of discrimination, connection to safe spaces, and reinforcement of intentions to use Safety Plans with mentees.

Methods: To understand PMs' experiences in their role, including distress, fidelity to the manual, and perceptions of feasibility and acceptability of STARS and mentees' Safety Plan, we collected survey data from mentees and PMs as well as in-depth interviews with PMs after the completion of the intervention.

Results: As of September 2024, all peer mentees (N=64) have completed the study and all PMs have finished providing sessions for peer mentees. PMs (n=5) reported overall high comfort (8.52) and low distress (1.93) during sessions. All 5 PMs had high fidelity (>90%) to the PM intervention training. All 5 PMs reported high feasibility (17.50), acceptability (20), and appropriateness (20) of the STARS intervention. Mentees (n=27) reported high confidence ratings (3.54) in speaking with their PMs.

Conclusions: Peer mentorship for SGDY who are at risk for suicide was feasible and acceptable by PMs and mentees alike. PMs reported that they felt comfortable and confident during the sessions. Mentees also reported confidence in working with their PMs. Future research should explore the optimal strategies to support PMs and mentees as they engage in suicide prevention work as well as incorporate feedback from the PMs in this study to ensure optimal outcomes.

Trial Registration: ClinicalTrials.gov NCT05018143; https://clinicaltrials.gov/study/NCT05018143

International Registered Report Identifier (IRRID): RR2-10.2196/48177

(JMIR Form Res 2025;9:e67814) doi:10.2196/67814

XSL•FO
RenderX

KEYWORDS

suicide prevention; peer mentorship; LGBTQIA health; mental health

Introduction

Emerging adulthood is a critical period of suicide risk for Sexual and Gender Diverse Youth (SGDY) people [1]. A meta-analysis indicated 11%-20% of sexual and gender diverse (SGD) individuals have a lifetime prevalence of suicide attempts compared with 4% of heterosexual adults [2]. Rates of suicide attempts may be higher; however, as these statistics only include SGDY who have self-reported their sexual and gender minority identity and ignore those who have not "come out" and are at even greater risk for suicidal ideation [2]. A meta-analysis of SGDY and emerging adults (12-20 years old) found that bisexual and transgender youth were at the most significant risk of attempting suicide (odds ratios of 4.87 and 5.87, respectively) compared with their cisgender heterosexual peers [3].

Sexually diverse emerging adults (eg, lesbian, gay, bisexual, transgender, and queer/questioning [LGBTQ]) may present with unique individual-level distal risk factors for suicide, such as the age of coming out, violence, exposure to sexual orientation change or conversion efforts, and discrimination, and proximal risk factors for suicide, such as internalized homophobia, expectations of rejection, and level of social support, which all interact to either increase or decrease the risk for suicide depending on the social environment and the individual's perceived belongingness within said environment [4-7]. Rooted in both the minority stress theory and the interpersonal theory of suicide, previous work has indicated that sexual and gender minority stress for adolescents and emerging adults is associated with suicidal ideation and attempts through perceived burdensomeness and thwarted belongingness [8]. To address feelings of burdensome and thwarted belongingness, increasing a sense of connectedness for SGM emerging adults can be an important protective factor against suicide, even when accounting for experiences of discrimination and victimization [9].

PM may serve as a strategy to provide social support, promote positive affect and feelings of belongingness, and build a connection to the community. Given at-risk SGM individuals' tendency to disclose suicidal ideation to peers rather than professionals, peers are crucial actors in suicide prevention [10,11]. Previous research demonstrated that most individuals contemplating suicide, particularly adolescents and young adults, do not seek help from formal support structures due to stigma concerns, but instead use informal resources like friends and family [12]. Therefore, educating peers about suicide, depression, and the resources available for at-risk individuals may reduce their reluctance to intervene and increase their ability to do so appropriately.

A review shows that peer mentoring can effectively promote health behavior changes in adolescents and emerging adults through reinforcing coping skills, incorporating skill-building activities, and delivering social support [13]. In studies of suicide prevention models, peer mentorship decreased stigma related to help-seeking behaviors and reduced rates of repeated psychiatric hospitalization [14,15]. A pilot trial of a peer-delivered safety planning intervention (SPI) found that adults who received the peer-delivered version had fewer emergency room visits during the follow-up period than those who received the provider-delivered version [16]. Those with sexual minority identities reported reluctance to access mental health services for fear of discrimination and dismissal of their emotional distress [17,18]. Therefore, approaches using peer mentorship that decrease stigma and decrease the need for repeated service engagement may be critical for reducing suicide risk in the SGM community.

For SGDY, effective peer mentor (PM) interventions should incorporate several design characteristics to prevent future suicide risk. First, they should target social support [19]. Second, they should involve skill-based peer-administered interventions instead of purely supportive interventions [20,21]. Third, they should market interventions as LBGTQ (lesbian, gay, bisexual, transgender, and queer/questioning)-affirming [22]. Fourth, they should emphasize the use of a safety plan, which reduces the risk of suicide attempts [23]. When SGM emerging adults are reluctant to use their safety plan, peer mentorship should be enhanced with principles of motivational interviewing (MI). MI is a counseling method designed to elicit behavior change [24-27]. MI can be successfully used in peer-based interventions [28-30]. Given the need for peer mentor interventions to consider a multitude of important design characteristics and targets, there is a need to understand the challenges for PMs. Research to understand the experiences of PMs who deliver suicide prevention interventions to SGDY would be instructive.

In this paper, we describe the protocol for training PMs to fidelity in a suicide prevention intervention (Supporting Transitions to Adulthood and Reducing Suicide [STARS]) and detail their experiences in delivering MI- and cognitive behavioral therapy (CBT)–based content to SGDY mentees. We hypothesized that PMs would retain high fidelity to the protocol throughout their participation in the project, operationalized as greater than 80% fidelity to critical components of each session. We report on peer mentor distress and comfort ratings (completed after each peer mentor session) and evaluate changes in distress and comfort throughout the trial. We hypothesized that as PMs gained more experience with the protocol and the population, distress ratings would significantly decrease over time and comfort ratings would increase over time. Finally, we completed qualitative interviews with PMs about their experiences serving in this role. We hypothesized that PMs would report that the STARS intervention was acceptable, feasible, and appropriate.

Methods

The STARS Intervention

This study was part of a larger university randomized control trial (NCT05018143) evaluating an app-based intervention STARS aimed to reduce suicidal ideation and behaviors among SGDY (ages 18-24 years) [31]. Eligible participants were

XSL•FO
RenderX

between the ages of 18 and 24 years, lived in the Philadelphia Metro area, reported no psychotic symptoms (hallucinations or delusions), had access to a smartphone and Wi-Fi, and had active suicidal ideation in the past month. Participants were randomized to receive an in-person brief, evidence-based SPI (control arm) or to receive the SPI plus access to STARS (intervention arm). Mentees randomized to the STARS intervention had access to a mobile app focused on the provision of life skills and their safety plan and 6 peer mentor sessions. More detailed information on the more extensive study can be found in the study by Brown et al [31].

PM Training Plan

PMs completed eight 2-hour training sessions with 2 licensed mental health providers. The first session included an overall introduction to MI. PMs received training in core MI concepts, with an emphasis on skill use. We provided a 2-hour training covering ambivalence, MI spirit (ie, acceptance, partnership, compassion, and evocation), and the righting reflex (the inclination helpers have to give advice, correct what they see as wrong information, or the wrong reasoning, and generally fix things for the person they are helping). OARS (ie, open-ended questions, affirmations, reflections, and summaries) skills were covered in a second 2-hour training. A third 2-hour training covered the 4 MI processes (ie, engaging, focusing, evoking, and planning) and the Helping Roadmap model [32]. PMs were assigned exercises after each training session to support their acquisition of skills. These trainings were adapted from materials used in the iReach trial [28] and were led by a masters-prepared research staff member (project manager and research coordinator) with a background in mental health counseling and delivering MI training for PMs.

After training and practice in the style of MI, PMs received 62-hour training sessions on the content of PM sessions. These trainings included a didactic overview with questions and answers, modeled roleplay, and then an observed experiential role-play between the PMs with feedback from the study clinicians. After training in the MI-style and CBT content of the peer mentor intervention, additional time was dedicated to navigating difficult scenarios, including exacerbations in suicide risk, risky behaviors in general, boundary crossings, and building rapport and connection (or repairing ruptures in the peer mentor relationship).

After the original training, PMs attended a weekly group consultation meeting that was attended by 1 to 3 study clinicians. In the weekly meeting, PMs provided an update on the mentees to whom they were assigned, discussed difficulties in the session, and provided examples of successes in the PM relationship. PMs provided feedback and ideas to each other, and the study clinicians provided validation, coaching, and support around content areas that PMs indicated a need for more support (ie, what to do in a crisis and how to engage participants).

Approximately 6 months into the start of study recruitment, the team dedicated two 30-minute supervision sessions to reviewing and refreshing PMs' MI style.

Participant Measures: Mentee Feedback

In their follow-up survey (at 2 months), mentees were asked questions about their experience with PM sessions, including: "PM sessions were offered at times that worked for my schedule," and "I feel confident talking to a STARS PM to discuss what's going on in my life." Questions were answered on a Likert scale from 1 (strongly disagree) to 4 (strongly agree).

PM Measures: PM Comfort and Distress

At the end of each PM session, PMs were asked to fill out a post-session form that assessed issues that arose and distress and comfort during the PM sessions. The PM comfort rating included two questions: (1) "On a scale from 1-10 how comfortable did you feel during this PM session?" ("10" is "very comfortable" and "1" is "the most uncomfortable") and (2) "Please rate the highest level of distress you experienced during this session on a scale of 1-10." ("10" is "very distressed" and "1" is "not at all distressed"). Scores on the second question were reverse coded, and the scores for both questions were averaged for a total comfort score ranging from 1 to 10, with higher scores indicating higher comfort.

Implementation Measures

Overview

To determine the extent to which the STARS PM intervention was acceptable, appropriate, and feasible, we adapted the Acceptability of Intervention Measure (AIM), Intervention Appropriateness Measure (IAM), and Feasibility of Intervention Measure (FIM) [33]. The AIM, IAM, and FIM demonstrated strong psychometric properties in a series of 3 studies conducted by Weiner et al [33]. The AIM, IAM, and FIM are 4-item measures asking participants to answer questions on a scale of "1" ("completely disagree") to "5" ("completely agree"). We adapted the scales for the STARS PM intervention; for example, 1 item from the AIM states, "The STARS peer mentoring intervention meets my approval."

PM Qualitative Interview

Semistructured interviews were conducted with all PMs to gain a deeper understanding of PM experiences and views on the STARS PM intervention. A member of the research team, outside of those providing training or supervising the implementation of the peer mentoring intervention, conducted the interviews with PMs virtually. The interviews lasted about 60 minutes and were audio recorded and transcribed verbatim.

We used a Template Analysis approach, a form of thematic analysis emphasizing hierarchical coding [34-36]. An a priori codebook was created informed by the interview guide. Out of 3 researchers (JT, TB, and JW) coded the first transcript using the codebook as a guide and template. After the first transcript, a consensus was reached, and the codebook was edited with final details. The remaining transcripts were divided between 2 researchers and individually coded [34-36].

Fidelity Monitoring

PM fidelity to the intervention was assessed using a checklist that evaluated their adherence to content and style [31]. Content was assessed based on whether PMs covered key elements of

XSL•FO
RenderX

the session (eg, agenda setting for the day's session, checking in about safety plan use since the last session, and teaching the designated CBT skill for the session). Possible total content scores for the sessions ranged from 16 to 29, based on how many elements were meant to be covered in each session. The style was assessed based on the PMs' use of MI skills OARS [25] to facilitate participant learning and engagement with the session content. The style was scored on a scale of 0-5 for whether PMs asked open-ended questions, used affirming statements, used reflections, and used summaries during the session.

PMs' fidelity to the intervention was assessed across all sessions for the first 2 mentees they saw, allowing for early detection of challenges to fidelity and corrective supervision. After completing the intervention with their first 2 mentees, PMs were assessed randomly for 1 in every 6 sessions per mentee. Sessions were randomly selected for fidelity assessment using a randomized list generator. Out of 2 research team members were tasked with completing fidelity ratings, 10% of which were scored twice to assess interrater reliability.

Ethical Considerations

The study was reviewed by the university's institutional review board (protocol # 849500) and was deemed a human-subjects study. Informed consent was obtained from participants in person during the baseline visit. Participants were informed of all the risks and benefits of the study. All information is dedeidentified and all personal information collected is kept private and confidential. Participants received up to US $170 (US $50 for baseline, US $30 for 2 months, US $40 for 4 months, and US $50 for 6 months) through physical or electronic ClinCard.

Results

Mentee Description

As of September 2024, all peer mentees (N=64) have completed the study. A total of 32 mentees were randomized to receive the STARS app intervention. One mentee opted not to participate in the PM sessions. Out of 4 mentees did not complete all 6 recommended PM sessions. Of the 4 mentees who did not complete all 6 sessions, 2 only completed PM session 1, 1 completed 4 PM sessions, and 1 completed 5 PM sessions. A total of 27 mentees completed all 6 PM sessions.

PM Description

PMs were recruited from the Greater Philadelphia, Pennsylvania area, through job announcements through area university student employment boards, university bulletin boards, and network connections. Applicants were interviewed based on previous education, experience with mentoring, mental health, and LGBTQ advocacy. PMs selected needed to have some previous experience in mentoring; however, other aspects were not required as we intentionally wanted a heterogeneity of PMs in educational or professional experiences parallel to participants in the study. Although we never asked specific questions of applicants or employees, given human resources guidelines, we can provide descriptive information about the 5 PMs who delivered sessions, which we obtained from working closely with them. At hiring, PMs were between the ages of 22 to 26, self-identified as a part of a sexual and gender minority community orally, and over half identified as a racial and ethnic minority.

Out of 5 PMs delivered sessions after completing their training. PMs were assigned to study mentees who were randomized to the STARS intervention. PMs worked with no more than 3 STARS mentees at a time. STARS participants were assigned to PMs based on a list and caseload to prevent overburdening PMs.

Fidelity to PM Training

PM fidelity checks were completed by 2 researchers (LB and JW), both trained in therapeutic peer mentorship. PM fidelity to content was calculated for each of the 6 peer mentor sessions. Fidelity percentages were high, ranging from 95.7% to 100%. PM fidelity to style was calculated for each of the 6 PM sessions and was high, ranging from 92.1% to 97.4%.

Mentee Feedback on Intervention

STARS mentees provided ratings for PM sessions at 2 months. Mentees (N=28) rated their confidence in talking with a PM as 3.54 (SD 0.64) out of a possible confidence rating of 4.0. Mentees agreed that sessions were offered at times that worked for their schedule as 3.74 (SD 0.53) out of a score of 4.0.

PM Outcomes

PM Implementation Outcomes

PMs (n=5) rated the sessions with a feasibility score (FIM) of 16.40 (SD 3.21), an appropriateness score (IAM) of 18.40 (SD 3.58), and an acceptability score (AIM) of 18.40 (SD 3.58) with a possible maximum score of 20 for all implementation outcomes (Table 1).

XSL•FO
RenderX

Table 1. Average fidelity ratings to the training of 6 peer mentor (PM) sessions of a mixed methods study on suicide prevention in sexual and gender-diverse youth.

Measure	Session Time Point					
	1	2	3	4	5	6
Content total scores						
Mean (SD)	15.21 (0.70	20.93 (1.49)	20.33 (1.11)	21.38 (1.76)	17.54 (1.13)	26.71 (0.73)
Maximum	16	22	21	23	18	29
Style total scores						
Mean (SD)	20 (0)	19.80 (0.77)	19.87 (0.35)	19.92 (0.28)	19.77 (0.60)	19.14 (1.83)
Maximum	20	20	20	20	20	20

PM Comfort and Distress

PMs indicated high comfort ratings after each session and low distress. Paired samples *t* tests indicated that there were no significant differences in comfort (*P*=.44) or distress ratings (*P*=.70) from session 1 to session 6 (Table 2).

Table 2. Average comfort and distress ratings self-reported by peer mentors (PMs) for each of the 6 PM sessions in a mixed methods study on a suicide prevention intervention for sexual and gender-diverse youth.

Measure	Session time point					
	1	2	3	4	5	6
Comfort ratings						
N	31	29	29	29	28	27
Mean (SD)	8.81 (0.79)	8.45 (1.33)	8.38 (1.40)	8.41 (1.18)	8.89 (1.07)	8.52 (1.60)
Distress ratings						
N	31	29	29	29	28	27
Mean (SD)	1.81 (1.22)	2.52 (1.79)	2.59 (2.15)	2 (0.85)	1.96 (1.04)	1.93 (0.87)

PM Experiences

Thematic analysis of the interviews with PMs categorized their experiences into 6 themes: decision for peer mentoring, training experiences, use of the pm manual, interpersonal relationship with mentee, internal struggles and growth, and STARS App Thoughts.

Decisions for Peer Mentoring

This theme represents background information on the PMs and their motivations for becoming involved in STARS. PMs worked with STARS for 8 months to 18 months. All PMs indicated they were interested in the role due to their career goals related to mental health, LGBTQ+ research, or clinical work. Many PMs reported wanting to give back or support the LGBTQ+ community, due to their connection to the community. One PM stated, “I was really interested in the study itself. I thought it just sounded like a really cool and meaningful idea that would definitely be helpful for the participants.” While another PM said “reason is my self-identification as a cisgender gay man, and that aligns with my value to get back to the community to support all of them.”

Training Experiences

This theme represented information on PMs’ experiences of PM training and how prepared they felt to provide PM sessions to mentees for STARS. All PMs reported that their training was helpful, specifically the role-playing sessions. One PM stated, “I felt like I was very prepared to do everything in relation to the like discussing safety plans and sort of working to brainstorm a safety plan that wasn’t quite working for somebody that I felt like we did so much kind of prep work on, and that was always one of that was like one of my favorite parts, always.” Another PM said, “I think doing the mock sessions was definitely helpful. Where I would be pretending to be the participant in one, and then [another PM] would do the opposite for me. I think that was definitely really helpful.” Some PMs spoke about the importance of having weekly meetings (3/5 PMs), particularly having that space to talk to supervisors and other PMs about their experiences and troubleshooting. A PM stated, “It was nice to have, you know, the weekly meeting and be able to share difficult things that would come up.” One PM indicated a possibility for more practice during training periods. Several PMs noted specific skills or topics they needed clarification on, such as engagement with their STARS mentees. All PMs noted that the continued training (MI) was helpful.

Use of PM Manual

PMs were asked to describe their experiences delivering the STARS sessions to trial mentees. This theme includes feedback on barriers, facilitators, and suggestions for improvement in future STARS manualized session content implementation. Overall, PMs described having “good experiences” delivering the session content. Out of 2 noted session lengths varied, attributed to content and the interactions with individual

XSL•FO
RenderX

mentees. One PM explains, “I think that some of the sessions felt a little bit more like jam-packed with content than others, but that also depended on the person that you were with, and some people really had nothing to say-- and so you could get through it pretty easily. And with some people that was harder.” One PM expressed concern that mentees “zoned out” when delivering content where they had to “talk too much” per the session script. Out of 3 PMs described feeling “more comfortable” delivering the session scripts with each new mentee experience. For most PMs, the main barrier to delivering session content was covering session content within the 30-minute timeframe. Out of 3 PMs described an imbalance in the length of content from session to session as a barrier. One explained, “Delivering the content. Yeah, that would be one thing that I feel like I usually take a long time in delivering session 5 or 6, and I find it hard to squeeze the time to like 25 or 30. So that can be one thing.”

Out of 3 PMs identified the script in the manual as a primary facilitator to content delivery, describing it as “the way I guided people” and “easy to follow along” during sessions. Another explained, “How it was broken up I feel like it made sense. There was like a good flow to it. And it was nice to have examples of how to phrase things, or how to talk about something or questions to ask. Yeah. So that was helpful.” PMs identified flexibility in content and session structure as potential areas for improving STARS implementation. One PM suggested differentiating between essential and optional session content to help adhere to the time limit of sessions. Similarly, another expressed that increasing content flexibility would allow PMs to make sessions more relevant for individual STARS mentees. One PM noted that “if there were things that I felt like we couldn’t just breeze over or just wasn’t in line with where the script was headed. I think things like that made it hard for me to make sure I was getting all the fidelity items done.”

Interpersonal Relationship With Mentees

This theme described any interpersonal challenges experienced in the mentoring role as well as positive changes and growths that they observed in STARS mentees. Some challenges described by PMs included difficulties with engagement, challenges related to participant personality or behavior, missing appointments, texting during sessions, trouble relating to session content, and not using the safety plan. Of the 3 PMs had experiences where they felt mentees were not engaging with session content. One PM explains, “Some of the mentees who I felt weren’t that engaged would be examples that they brought up…And I feel like some mentees would like, give, just like, very like surface-level, or things that weren’t significant or gonna help them. I don’t know if they were avoiding getting deep with me. And, you know, I didn’t like that we couldn’t kinda get beyond that.”

Mentee personalities and behaviors were sometimes a challenge for PMs. Some PMs found it difficult to balance covering the session content while giving space for sharing with more talkative STARS mentees. During the sessions, a PM described challenges with STARS mentees doing other things on their computers and phones. Out of 2 PMs described challenges with STARS mentees who frequently missed appointments and needed to reschedule. A PM stated, “There were some who were like chronically late, or would forget, needed to reschedule.” Resolutions included sending reminders before sessions, changing times, and discussing barriers to scheduling.

While there were some interpersonal challenges, PMs also described growth in their STARS mentees. Out of 2 PMs identified an increase in openness to sharing experiences as an area of growth among STARS mentees. One PM described an example in which a mentee was initially resistant to being on camera at the first meeting due to negative body image and concerns of being seen by others but eventually engaged fully in video sessions as the relationship developed, indicating a relationship of safety and trust. All 5 PMs identified a willingness to practice and apply session content outside of weekly meetings as a primary growth area among STARS mentees. One PM noted, “The participants, mostly all of them I had like they were actually implementing [session content]. Like all of the participants, that kind of surprised me a bit. “Cause they actually were using the things that [STARS] was wanting to do. Which is the goal. Yeah. It was cool.” In one example, a PM noticed that STARS mentees who gave positive session feedback shared that they were also using the app and practicing content between sessions. Out of 2 PMs described a noticeable shift in mentees’ motivation and engagement around setting and following through on achieving goals. One PM explained, “I mean, a lot of people were reporting, I think, less suicidal ideation or more ease in using a safety plan and more sort of like feeling like they were able to ride the waves of something which was great and hmm. Yeah, I think just a lot of people developed coping skills.”

Internal Struggles and Growth

This theme represented information on the STARS PMs’ internal experiences throughout the study, including emotional responses and reactions to their interactions with mentees, both positive and negative. When discussing concerns about providing peer mentorship to mentees with a history of suicide risk, all 5 PMs mentioned a fear that a mentee would actively be in crisis during a session. They were comforted knowing that they had a script and protocol to follow if something did happen. One PM noted, “One major concern was just like how to go about like if they feel like they’re in crisis while they’re with me. And then another one was how to respond if someone says they want to harm themselves. I remember learning during the sessions.” Despite the safeguards in place, the potential for suicidal crises remained a source of anxiety. Some other internal, emotional, or personal challenges that PMs encountered included worries about their STARS mentees’ health, difficulty hearing about mentees’ history with suicidal ideation, and not being sure if they were making sense or being engaging to their STARS mentees. One person said that these challenges subsided the more “exposure” they had. Another PM felt frustrated by needing to balance helping the STARS mentee with sticking to the script to keep their fidelity rating high.

While PMs expressed some internal struggles, they managed difficult emotions that came up by mentioning them to the principal investigators during weekly team meetings and referring to the script in moments of uncertainty. Out of 2 PMs

XSL•FO
RenderX

mentioned the comfort that the script brought them. A PM stated, “I'm sticking to the scripts, and hitting all the items I need to hit… When those goals would kind of work against what I think is most important, to really just see what is most useful for the participant and see what I could do to help them.” Out of 2 other PMs said it was helpful to remind themselves of the boundaries of this role and that the STARS mentees’ lives outside of PM sessions were not in their control.

PMs also indicated that the experience was rewarding seeing some improvement in their STARS mentees. This ranged from simply feeling helpful and effective in their sessions to feeling like they were imparting something useful to seeing STARS mentees use their safety plan between sessions. A PM stated, “I think it was just like rewarding to be able to feel like I’m imparting something useful to the peer mentees. So it was nice when we would have like, you know, a fruitful discussion that felt like it was something that was helpful. And that they could use in their real life and use to make changes that were helpful to them.” One PM felt rewarded by getting encouraging supervisor feedback and hearing STARS mentees express gratitude. In addition, another PM appreciated the opportunity to interact more with the queer community and learn about experiences different from their own.

Discussion

Principal Findings

This study aimed to evaluate the experiences of PMs who deliver suicide prevention content to SGDY. The study was designed to provide PMs with comprehensive training for working with SGDY at risk for suicide, evaluate PM experiences (reported ratings of comfort and distress), and assess PM fidelity to the protocol (PM training).

It has been well-documented that there is anxiety in providing care for patients at risk for suicide in behavioral health practitioners [37]. Therefore, it is important to assess comfortability and distress with PMs working in suicide prevention interventions. Overall, PMs described high comfort and low distress ratings after each PM session. The comfort and distress ratings did not change over time. These ratings were aligned with the reported experiences of PMs during the follow-up interviews. PMs described having “good experiences” in delivering the session content and several reported that they felt “more comfortable” as they saw more mentees. Many PMs attributed their comfort with sessions by mentioning them during weekly team meetings. Some PMs stated that having the manual was a comfort during the sessions as a resource for them to refer. Therefore, having regular supervision and a thorough resource (such as a manual) are important components to include in future PM based interventions for SGDY at risk for suicide. To our knowledge, this is the first study to report on PM comfort and distress levels when providing support to mentees in a suicide prevention intervention. These results highlight the importance of structured supervision for PMs to aid in providing support to mentees at risk of suicide knowing the anxiety that can surround such a stigmatizing topic [37].

In examining the fidelity of PM sessions with the training, PMs had high fidelity overall (greater than 90% in content and style fidelity ratings). Fidelity (the degree to which an intervention, ie, peer mentoring, was delivered as intended) [38] plays an important role in how well an intervention is considered successful. Fidelity to an intervention provides information that can help determine if the intervention was delivered as intended and therefore any impacts can be more confidently associated with the intervention. PMs provided sessions to STARS mentees with high fidelity; however, PMs noted some challenges. PMs were made aware of fidelity to the training and session content were being assessed and expressed concerns about completing all the fidelity items in the allotted time limit (30-minute sessions). PMs noted that there was a lot of content to cover during the sessions, but they also wanted to provide space and time for their STARS mentees to share updates. PMs expressed frustration with sticking to the script and addressing all the fidelity content items while also preserving fidelity style. However, we did not interview or collect data pertaining to the STARS mentees experience with their PMs and cannot confirm if this was a similar experience for STARS mentees. These concerns from PMs have been noted by mental health professionals who advocate for the use of a manual in treatment [39]. One noted suggestion is the idea of flexibility in fidelity which refers to the implementation of an intervention protocol that contains the core attributes for fidelity but also allows for flexibility for each individual client [40]. A suggestion for future implementation should include highlighting content that is necessary. Determining what content is necessary can come from feedback provided by licensed mental health professionals, PMs, and mentees. Future iterations of the intervention should include flexibility in fidelity [40] within the manual and training by reducing content and increasing space for PMs and STARS mentees to connect for social support.

The STARS PM intervention was developed with several components for a culturally informed design including (1) targeting social support, (2) skill-based training for PMs, (3) focusing on SGM-affirming context, and (4) highlighting safety plan use [19-23]. In our interviews with PMs, we noted how PMs experience delivering the PM intervention to STARS mentees and how they align with the 4 factors incorporated into the intervention design. For the first factor of social support, PMs reported that mentees expressed gratitude and that they noticed improvement in their mentees. However, as mentioned above, we did not directly collect data from STARS mentees to determine if they felt supported by their PM. Future iterations of the study should look to collected data from STARS mentees and their experiences. Second, PMs highlighted the importance and helpfulness of the skill-based training that was provided, with MI and role-playing. All PMs noted that initial and continual training in MI skills were helpful in conducting sessions with STARS mentees. Some even noted even more training in MI would be helpful. Third, PMs all noted that they wanted to be mentors because they wanted to provide and support the SGD community. While we did not assess whether STARS mentees felt affirmed, we can note that the content delivered and provided for were SGD-affirming. Finally, the fourth component of the safety plan use was reiterated within the training and fidelity of the PMs’ sessions with STARS

XSL•FO
RenderX

mentees. During all sessions, PMs were asked to check in with STARS mentees and their use of the safety plan. Some PMs noted the difficulty in how to engage STARS mentees with the safety plan, particularly when mentees noted that they had not looked at or used their safety plan.

Our study provides needed context and implementation information regarding peer-based interventions for suicide prevention for SGDY. In a recent scoping review, Bowersox et al [41] identified very few peer-based interventions for suicide prevention overall. Researchers identified only 9 peer-based interventions aimed at crisis and relapse prevention: however, none of these focused on SGD communities. In a PM program for older adults, Van Orden et al [42] trained senior companions (55 years and older) on topics of confidentiality, reporting requirements, accommodations for disabilities, and common physical and mental health conditions but not in suicide risk. Another program, PREVAIL (Peers for Valued Living), is a trial for adapting peer support delivered intervention for veterans recently hospitalized for suicidal thoughts or behaviors [43]. In qualitative interviews with key parties (veterans with current or recent suicide risk, suicide prevention coordinator, clinicians, peer specialists, and a director of inpatient psychiatry), researchers noted that peers should be trained in topics specific to suicide prevention as well as more general clinical approaches such as MI. These findings inspired the focus in our project of offering practical skills training for reducing suicide risk using a style of MI. There are possible applications of the mentoring model developed in our study (ie, training structure, supervision support, and MI skills) for other communities at risk of suicide, including veterans and older adults.

Limitations and Future Directions

Limitations of our study include the small sample of PMs (n=5). The PMs may not have been representative of the whole pool of persons that would be PMs for our study as we were unable to recruit directly stating PMs with lived experiences of suicidal ideation and identify as a part of sexual and gender minority communities. Another limitation of the study is the lack of feedback from mentees on their experiences with PM sessions. Future studies should include gathering qualitative data from mentees on their experiences. Future directions should include simultaneous mixed methods study design of qualitative interviews with PMs and participants to understand the relationship between peers as it relates to quantitative outcomes (ie, thwarted belongingness) of participants identifying as SGM emerging adults.

Conclusion

PMs are outstanding candidates for delivering suicide prevention services, particularly when working with historically oppressed communities, such as SGDY. In this study, we demonstrated the feasibility and acceptability of training and implementing peer mentorship for SGDY who are at higher risk for suicide. PMs reported high ratings of comfort and low ratings of distress when implementing the STARS sessions. In addition, PMs had high fidelity ratings throughout the intervention. Finally, PMs offered helpful suggestions for improving our STARS app and PM content and style to improve outcomes in the future.

Data Availability

The datasets generated and analyzed during this study are not publicly available due to the sensitive nature of the data collected but are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

References

1. Fish JN, Rice CE, Lanza ST, Russell ST. Is young adulthood a critical period for suicidal behavior among sexual minorities? Results from a US national sample. Prev Sci 2019;20(3):353-365 [FREE Full text] [doi: 10.1007/s11121-018-0878-5] [Medline: 29594980]
2. Hottes TS, Bogaert L, Rhodes AE, Brennan DJ, Gesink D. Lifetime prevalence of suicide attempts among sexual minority adults by study sampling strategies: a systematic review and meta-analysis. Am J Public Health 2016;106(5):e1-e12 [FREE Full text] [doi: 10.2105/ajph.2016.303088]
3. di Giacomo E, Krausz M, Colmegna F, Aspesi F, Clerici M. Estimating the risk of attempted suicide among sexual minority youths: a systematic review and meta-analysis. JAMA Pediatr 2018;172(12):1145-1152 [FREE Full text] [doi: 10.1001/jamapediatrics.2018.2731] [Medline: 30304350]
4. Blosnich JR, Henderson ER, Coulter RWS, Goldbach JT, Meyer IH. Sexual orientation change efforts, adverse childhood experiences, and suicide ideation and attempt among sexual minority adults, United States, 2016–2018. Am J Public Health 2020;110(7):1024-1030 [FREE Full text] [doi: 10.2105/ajph.2020.305637]
5. Hong JS, Espelage DL, Kral MJ. Understanding suicide among sexual minority youth in America: an ecological systems analysis. J Adolesc 2011;34(5):885-894 [FREE Full text] [doi: 10.1016/j.adolescence.2011.01.002] [Medline: 21303716]
6. Meyer IH, Russell ST, Hammack PL, Frost DM, Wilson BDM. Minority stress, distress, and suicide attempts in three cohorts of sexual minority adults: A U.S. probability sample. PLoS One 2021;16(3):e0246827 [FREE Full text] [doi: 10.1371/journal.pone.0246827] [Medline: 33657122]

XSL•FO
RenderX

7. Plöderl M, Sellmeier M, Fartacek C, Pichler EM, Fartacek R, Kralovec K. Explaining the suicide risk of sexual minority individuals by contrasting the minority stress model with suicide models. Arch Sex Behav 2014;43(8):1559-1570 [FREE Full text] [doi: 10.1007/s10508-014-0268-4] [Medline: 24573399]
8. Fulginiti A, Goldbach JT, Mamey MR, Rusow J, Srivastava A, Rhoades H, et al. Integrating minority stress theory and the interpersonal theory of suicide among sexual minority youth who engage crisis services. Suicide Life Threat Behav 2020;50(3):601-616 [FREE Full text] [doi: 10.1111/sltb.12623] [Medline: 32048340]
9. Busby DR, Horwitz AG, Zheng K, Eisenberg D, Harper GW, Albucher RC, et al. Suicide risk among gender and sexual minority college students: the roles of victimization, discrimination, connectedness, and identity affirmation. J Psychiatr Res 2020;121:182-188 [FREE Full text] [doi: 10.1016/j.jpsychires.2019.11.013] [Medline: 31837538]
10. Aseltine RH, DeMartino R. An outcome evaluation of the SOS suicide prevention program. Am J Public Health 2004;94(3):446-451 [FREE Full text] [doi: 10.2105/ajph.94.3.446] [Medline: 14998812]
11. Cross WF, Seaburn D, Gibbs D, Schmeelk-Cone K, White AM, Caine ED. Does practice make perfect? A randomized control trial of behavioral rehearsal on suicide prevention gatekeeper skills. J Prim Prev 2011;32(3-4):195-211 [FREE Full text] [doi: 10.1007/s10935-011-0250-z] [Medline: 21814869]
12. Michelmore L, Hindley P. Help-seeking for suicidal thoughts and self-harm in young people: a systematic review. Suicide Life Threat Behav 2012;42(5):507-524. [doi: 10.1111/j.1943-278X.2012.00108.x] [Medline: 22889130]
13. Petosa RL, Smith LH. Peer mentoring for health behavior change: a systematic review. American Journal of Health Education 2014;45(6):351-357 [FREE Full text] [doi: 10.1080/19325037.2014.945670]
14. Sledge WH, Lawless M, Sells D, Wieland M, O'Connell MJ, Davidson L. Effectiveness of peer support in reducing readmissions of persons with multiple psychiatric hospitalizations. Psychiatr Serv 2011;62(5):541-544 [FREE Full text] [doi: 10.1176/ps.62.5.pss6205_0541] [Medline: 21532082]
15. Wyman PA, Brown CH, Inman J, Cross W, Schmeelk-Cone K, Guo J, et al. Randomized trial of a gatekeeper program for suicide prevention: 1-year impact on secondary school staff. Journal of Consulting and Clinical Psychology 2008;76(1):104-115 [FREE Full text] [doi: 10.1037/0022-006x.76.1.104]
16. Wilson MP, Waliski A, Thompson RG. Feasibility of peer-delivered suicide safety planning in the emergency department: results from a pilot trial. Psychiatr Serv 2022;73(10):1087-1093 [FREE Full text] [doi: 10.1176/appi.ps.202100561] [Medline: 35502515]
17. McDermott E, Hughes E, Rawlings V. Norms and normalisation: understanding lesbian, gay, bisexual, transgender and queer youth, suicidality and help-seeking. Cult Health Sex 2018;20(2):156-172 [FREE Full text] [doi: 10.1080/13691058.2017.1335435] [Medline: 28641479]
18. Mirza SA, Rooney C. Center for American Progress, 18. Discrimination prevents LGBTQ people from accessing health care. URL: https://www.americanprogress.org/article/discrimination-prevents-lgbtq-people-accessing-health-care/ [accessed 2018-01-18]
19. McDonald K. Social support and mental health in LGBTQ adolescents: a review of the literature. Issues Ment Health Nurs 2018;39(1):16-29. [doi: 10.1080/01612840.2017.1398283] [Medline: 29333899]
20. Bryan AE, Arkowitz H. Meta-analysis of the effects of peer-administered psychosocial interventions on symptoms of depression. Am J Community Psychol 2015;55(3-4):455-471 [FREE Full text] [doi: 10.1007/s10464-015-9718-y] [Medline: 25861883]
21. Pfeiffer PN, Heisler M, Piette JD, Rogers MA, Valenstein M. Efficacy of peer support interventions for depression: a meta-analysis. Gen Hosp Psychiatry 2011;33(1):29-36 [FREE Full text] [doi: 10.1016/j.genhosppsych.2010.10.002] [Medline: 21353125]
22. Goldbach JT, Rhoades H, Green D, Fulginiti A, Marshal MP. Is there a need for LGBT-Specific suicide crisis services? Crisis 2019;40(3):203-208 [FREE Full text] [doi: 10.1027/0227-5910/a000542] [Medline: 30109965]
23. Stanley B, Brown GK, Brenner LA, Galfalvy HC, Currier GW, Knox KL, et al. Comparison of the safety planning intervention with follow-up vs usual care of suicidal patients treated in the emergency department. JAMA Psychiatry 2018;75(9):894-900 [FREE Full text] [doi: 10.1001/jamapsychiatry.2018.1776] [Medline: 29998307]
24. Budhwani H, Naar S. Training providers in motivational interviewing to promote behavior change. Pediatr Clin North Am 2022;69(4):779-794 [FREE Full text] [doi: 10.1016/j.pcl.2022.04.008] [Medline: 35934499]
25. Miller W, Rollnick S. Motivational Interviewing: Helping people change (3rd ed.). New York: Guilford Press; 2013.
26. Naar-King S. Motivational interviewing in adolescent treatment. Can J Psychiatry 2011;56(11):651-657 [FREE Full text] [doi: 10.1177/070674371105601103] [Medline: 22114919]
27. Naar-King S, Suarez M. Motivational interviewing with adolescents and young adults. New York: Guilford Press; 2011.
28. Bonar EE, Wolfe JR, Drab R, Stephenson R, Sullivan PS, Chavanduka T, et al. Training young adult peers in a mobile motivational interviewing-based mentoring approach to upstream HIV prevention. Am J Community Psychol 2021;67(1-2):237-248 [FREE Full text] [doi: 10.1002/ajcp.12471] [Medline: 33137221]
29. Mastroleo NR, Magill M, Barnett NP, Borsari B. A pilot study of two supervision approaches for peer-led alcohol interventions with mandated college students. J Stud Alcohol Drugs 2014;75(3):458-466 [FREE Full text] [doi: 10.15288/jsad.2014.75.458] [Medline: 24766758]

XSL•FO
RenderX

30. Naar-King S, Outlaw A, Green-Jones M, Wright K, Parsons JT. Motivational interviewing by peer outreach workers: a pilot randomized clinical trial to retain adolescents and young adults in HIV care. AIDS Care 2009;21(7):868-873 [FREE Full text] [doi: 10.1080/09540120802612824] [Medline: 20024744]
31. Brown LA, Webster JL, Tran JT, Wolfe JR, Golinkoff J, Patel E, et al. A suicide prevention intervention for emerging adult sexual and gender minority groups: protocol for a pilot hybrid effectiveness randomized controlled trial. JMIR Res Protoc 2023;12:e48177 [FREE Full text] [doi: 10.2196/48177] [Medline: 37773618]
32. Young ME. Learning the art of helping: building blocks and techniques (5th ed). Hoboken: Upper Saddle River, NJ: Prentice Hall; 2013.
33. Weiner BJ, Lewis CC, Stanick C, Powell BJ, Dorsey CN, Clary AS, et al. Psychometric assessment of three newly developed implementation outcome measures. Implement Sci 2017;12(1):108-112 [FREE Full text] [doi: 10.1186/s13012-017-0635-3] [Medline: 28851459]
34. Braun V, Clarke V. Reflecting on reflexive thematic analysis. Qualitative research in sport, exercise and health 2019;11(4):589-597. [doi: 10.1080/2159676x.2019.1628806]
35. Brooks J, McCluskey S, Turley E, King N. The utility of template analysis in qualitative psychology research. Qual Res Psychol 2015;12(2):202-222 [FREE Full text] [doi: 10.1080/14780887.2014.955224] [Medline: 27499705]
36. Byrne D. A worked example of braun and clarke's approach to reflexive thematic analysis. Qual Quant 2021;56(3):1391-1412 [FREE Full text] [doi: 10.1007/s11135-021-01182-y]
37. Becker-Haimes EM, Klein CC, Frank HE, Oquendo MA, Jager-Hyman S, Brown GK, et al. Clinician maladaptive anxious avoidance in the context of implementation of evidence-based interventions: a commentary. Front Health Serv 2022;2 [FREE Full text] [doi: 10.3389/frhs.2022.833214] [Medline: 36382152]
38. Carroll C, Patterson M, Wood S, Booth A, Rick J, Balain S. A conceptual framework for implementation fidelity. Implement Sci 2007;2:40-49 [FREE Full text] [doi: 10.1186/1748-5908-2-40] [Medline: 18053122]
39. Kendall PC, Beidas RS. Smoothing the trail for dissemination of evidence-based practices for youth: flexibility within fidelity. Professional Psychology: Research and Practice 2007;38(1):13-20 [FREE Full text] [doi: 10.1037/0735-7028.38.1.13]
40. Kendall PC, Frank HE. Implementing evidence-based treatment protocols: Flexibility within fidelity. Clin Psychol (New York) 2018;25(4):e12271 [FREE Full text] [doi: 10.1111/cpsp.12271] [Medline: 30643355]
41. Bowersox NW, Jagusch J, Garlick J, Chen JI, Pfeiffer PN. Peer-based interventions targeting suicide prevention: a scoping review. Am J Community Psychol 2021;68(1-2):232-248 [FREE Full text] [doi: 10.1002/ajcp.12510] [Medline: 33720444]
42. Van Orden KA, Stone DM, Rowe J, McIntosh WL, Podgorski C, Conwell Y. The senior connection: design and rationale of a randomized trial of peer companionship to reduce suicide risk in later life. Contemp Clin Trials 2013;35(1):117-126 [FREE Full text] [doi: 10.1016/j.cct.2013.03.003] [Medline: 23506973]
43. Lapidos A, Abraham KM, Jagusch J, Garlick J, Walters H, Kim HM, et al. Peer mentorship to reduce suicide attempts among high-risk adults (PREVAIL): rationale and design of a randomized controlled effectiveness-implementation trial. Contemp Clin Trials 2019;87:105850 [FREE Full text] [doi: 10.1016/j.cct.2019.105850] [Medline: 31525489]

Abbreviations

AIM: Acceptability of Intervention Measure
CBT: cognitive behavioral therapy
FIM: Feasibility of Intervention Measure
IAM: Intervention Appropriateness Measure
LGBTQ: lesbian, gay, bisexual, transgender, and queer/questioning
MI: motivational interviewing
OARS: open-ended questions, affirmations, reflections, and summaries
PM: peer mentor
PREVAIL: Peers for Valued Living
SGD: sexual and gender diverse
SGDY: sexual and gender diverse youth
SPI: safety planning intervention
STARS: Supporting Transitions to Adulthood and Reducing Suicide

XSL•FO
RenderX

Edited by A Mavragani; submitted 23.10.24; peer-reviewed by C Oko, AD Eaton, TF Ojo; comments to author 28.11.24; revised version received 09.12.24; accepted 11.12.24; published 29.01.25.

Please cite as:
Tran JT, Webster J, Wolfe JR, Ben Nathan J, Mayinja L, Kautz M, Oquendo MA, Brown GK, Mandell D, Mowery D, Bauermeister JA, Brown LA
Experiences of Peer Mentoring Sexual and Gender Minority Emerging Adults Who Are at Risk for Suicide: Mixed Methods Study
JMIR Form Res 2025;9:e67814
URL: https://formative.jmir.org/2025/1/e67814
doi:10.2196/67814
PMID:

©Jennifer T Tran, Jessica Webster, James R Wolfe, Jennifer Ben Nathan, Lindiwe Mayinja, Marin Kautz, Maria A Oquendo, Gregory K Brown, David Mandell, Danielle Mowery, José A Bauermeister, Lily A Brown. Originally published in JMIR Formative Research (https://formative.jmir.org), 29.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Self-Guided Smartphone App (Vimbo) for the Reduction of Symptoms of Depression and Anxiety in South African Adults: Pilot Quantitative Single-Arm Study

Sherrie Steyn[1*], BA, MSc; Meggan Slabbert[2*], MSocSci

[1]Vimbo Health SA (Pty) Ltd, Sandton, South Africa

[2]Private Practice, Ballito, South Africa

[*]all authors contributed equally

Corresponding Author:
Sherrie Steyn, BA, MSc
Vimbo Health SA (Pty) Ltd
20 Riverclub Mews, 7 Sycamore Avenue, Riverclub
Sandton, 2191
South Africa
Phone: 27 63 861 6780
Email: sherrie@vimbohealth.com

Abstract

Background: Barriers to mental health assessment and intervention have been well documented within South Africa, in both urban and rural settings. Internationally, evidence has emerged for the effectiveness of technology and, specifically, app-based mental health tools and interventions to help overcome some of these barriers. However, research on digital interventions specific to the South African context and mental health is limited.

Objective: This pilot study investigated the feasibility of using an app (Vimbo) to treat symptoms of anxiety and depression in South African adults recruited from a community sample. The Vimbo app is a self-guided, cognitive behavioral therapy–based digital intervention for common mental health difficulties developed for the South African context.

Methods: This pilot study used a naturalistic, single-arm design testing the Vimbo app over 12 weeks, from October 2020 to February 2021. Participants were recruited through the South African Depression and Anxiety Group and social media advertisements online. A 2-week retention period was used to allow for a minimum of 2 datasets. App usage and engagement metrics were extracted directly from the back end of the app. Based on the model, researchers expected many users to discontinue usage when their symptom levels entered a healthy range. Pre-post review of symptom levels was used to reflect on clinical recovery status at discontinuation after the retention period.

Results: A total of 218 applicants met study eligibility criteria and were invited to download the Vimbo app. Of these, 52% (114/218) of the participants registered with the app, who indicated multiple variances of depression and anxiety symptoms ranging in severity from mild to severe. Two participants users withdrew from the study. Moreover, 69% (77/112) of users were retained, including 8 who had technical issues with their treatment. When comparing broad uptake across all interested participants, chi-square analysis indicated significantly reduced uptake in participants identifying as "unemployed but seeking employment" (χ^2_4=10.47; N=251; *P*=.03). When considering app usage for the entire cohort (n=69, excluding participants with technical issues), there was a mean of 72.87 (SD 71.425) total module pages read, a mean of 30% (SD 29.473%) of prescribed content completed, and a mean of 19.93 (SD 27.517) times engaging with tools and skills.

Conclusions: Our findings support the case for continued exploration of app-based interventions for treating depression and anxiety in South Africa. Developing strategies to increase access and improve intervention uptake may prove essential to helping mobile health interventions make as significant an impact as possible. Future research should include a randomized controlled trial with a larger sample to further assess the efficacy of app-based interventions in treating mental health difficulties in South Africa.

(JMIR Form Res 2025;9:e54216) doi:10.2196/54216

XSL•FO
RenderX

KEYWORDS

treatment gap; mental health; health; depression; anxiety; South Africa; CBT; cognitive behavioral therapy; app-based intervention; mobile health; mental health app; smartphone; mobile phone

Introduction

South Africa currently has some of the poorest mental health outcomes globally, with 36% of the population feeling distressed or struggling [1]. Recent studies suggest that nationally, 25.7% (874/3402) of respondents reported scores ≥10 on the Patient Health Questionnaire-9 (PHQ-9) and 17.8% (606/3402) reported scores of >10 on the Generalized Anxiety Disorder-7 (GAD-7), indicating probable depression or anxiety [2]. Disorder prevalence varies widely by geographical province, which may reflect the variance in mental health resource availability and allocation [2,3]. These estimates are a marked increase from previous findings from the World Health Organization, which estimated that 4.6% of the population lived with a depressive disorder and 3.4% lived with an anxiety disorder [4]. This increase is part of a more significant global increase in cases of anxiety and major depressive disorders following the COVID-19 pandemic, including a 21.5% increase in the prevalence of these disorders in sub-Saharan Africa [5].

There is an ever-widening divide between increasing numbers of people experiencing mental health difficulties and access to adequate support or therapeutic intervention, known as the "treatment gap" [6,7]. Findings indicate that mental disorders often go untreated despite their high prevalence and associated impairment [6]. The treatment gap for depression and many anxiety disorders was more than 50%, increasing to as much as 80% in low- and middle-income countries for severe cases [6,8]. Government spending on mental health is below what is needed to adequately address this gap, with those in middle-income countries being the most affected [8].

South Africa is no exception. Despite having dedicated legislation, such as the Mental Health Care Act of 2002, there remains a significant treatment gap, especially in rural areas [9-11]. Recent studies suggest that for people in South Africa with mental health problems, only 1 in 10 receives the care they need [3]. This figure amounts to a staggering 92% treatment gap before the added burden of COVID-19. South Africa faces many barriers to bridging the treatment gap, including high levels of comorbidity and the relationship between common mental disorders (CMDs) and communicable diseases (such as HIV and tuberculosis), the apartheid legacy of institutional disparity, lack of adequate development of services, and a severe shortage of mental health professionals [9,11]. Research into cost-effective interventions is essential to inform policy decisions [9-11].

The use of technology and mental health apps in bridging barriers to this treatment gap is growing [7,9,12]. In 2018, approximately 99.5% of the population in South Africa had access to the 3G mobile network, and 85.7% had 4G/LTE coverage [13]. As of January 2020, South Africa had an estimated 3493 million active mobile web users [14].

The potential for app-based interventions to significantly reduce the cost of treatment, increase access in difficult-to-reach populations (such as rural communities), and reduce the need for clinician input and time is especially relevant to the South African context where these barriers have a more significant impact [6,9-11]. Several studies within South Africa have shown that mobile phones are already used as informal learning tools by patients and health care workers [15-20].

Despite this growing evidence and the number of mobile health (mHealth) apps, research is needed into apps designed specifically for the South African context and market. More generally, further research is needed to consider the cost-effectiveness and therapeutic efficacy of interventions for CMDs, especially mobile-based technology in South Africa [10,11].

Vimbo Health is a mental health assessment and treatment app, informed and developed based on a cognitive behavioral therapy (CBT) approach. The app streams users into courses for depression, anxiety and comorbid presentations using standardized psychometric measures, namely, the PHQ-9 and GAD-7. Subclinical users receive a well-being course focused on building awareness and resilience. CBT has proven effective in reducing symptom severity and functional impairment in both depressive and anxiety disorders [21-25].

This study aimed to investigate the feasibility of using the Vimbo app to reduce symptoms of anxiety and depression within an adult South African cohort. Using a single-arm, naturalistic design, we sought to pilot the app within a community sample as a self-guided intervention for those interested in supporting their mental health. We hypothesized that if offered a self-guided app (Vimbo), South African adults would voluntarily engage with the app to improve their mental health. We further hypothesized that Vimbo would reduce common symptoms of anxiety and depression symptom severity within 12 weeks of initial usage. Testing Vimbo's safety and efficacy provides valuable knowledge about using app-based mental health interventions in South Africa.

Methods

Study Design and Procedure

This study used a naturalistic, single-arm design to assess the acceptability of a self-guided, digital intervention to reduce symptoms of depression and anxiety in South African adults, using multiple repeat measures administered in the app over 12 weeks.

Interested individuals provided consent (Multimedia Appendix 1) and completed a demographic questionnaire and PHQ-9 online via a web form. Eligible participants were invited to participate via email, offered technical support, and given 2 weeks to download and register with the app (February 1-15, 2021). App registration included completing the PHQ-9, GAD-7, and Work and Social Adjustment Scale (WSAS). Participants self-managed app engagement with psychometric measures administered biweekly in the app. Data were collected directly

XSL•FO
RenderX

from the app over 12 weeks (ended May 10, 2021). At the study's end, participants received a user experience questionnaire to collect qualitative feedback and were offered an opportunity to debrief with a researcher. Participants had 2 weeks to complete and submit the user experience questionnaire on the web. This study does not include these qualitative data, which will be reported separately.

Ethical Considerations

This study received ethical clearance from the Human Sciences Research Council (HSRC) of South Africa on August 28, 2020. The HSRC Research Ethics Committee (REC) is registered with the South African National Health Research Ethics Council (REC-290808-015). The HSRC REC has the US Office for Human Research Protections Federalwide Assurance (organization 0000 6347). All participants signed a research study information and consent form via a web-based portal using a standardized form provided by the HSRC. Received consent forms were reviewed by researchers to ensure proper completion. All interested individuals were encouraged to contact the researchers should they have any question or wish to discuss their participation. Participation in the research was completely voluntary, with participants able to withdraw from the study at any point during the study period. The study was available to anyone comfortable communicating in English.

The Vimbo app is designed specifically for low data usage, both during download and regular use. The app has offline capability; as such, course content is downloaded once at the point of registration. Guided audio meditations are downloaded once when first played, so further playing does not consume any additional data. The app syncs to our back end when in daily use; however, expected use should not exceed 0.5 MB of data per day. The app is expected to use 126 MB over the 3-month study period. As such, participants were offered a once off reimbursement of R29.00 (approximately USD 2) to cover data usage based on local data costs.

Personally identifiable information stored in the mobile app back end was held on an encrypted Google cloud database, access to which was password protected and tightly controlled by the chief executive officer. Access to the back end is limited to only the chief executive officer and the software developer and restricted to critical use cases.

Data extracted from the back end were anonymized by removing or obfuscating personally identifiable information (name, email address, and phone number). Extracted data were kept in a password-protected and encrypted web-based folder hosted by Google with access granted only to the researchers. All electronic files generated by the research were maintained in the same electronic folders. Google Drive offers encryption and restricted access rights.

Informed consents and prescreening data were captured and stored electronically using the platform Jotform, which is password protected, is encrypted, and meets high security standards. Participants were anonymized through the use of coded identifiers, with any personally identifiable information including contact details stored in a separate, password-protected folder. Data were accessible only to the researchers and, due to necessity, the chief executive officer of Vimbo Health Ltd.

Participants

Participants were recruited through the South African Depression and Anxiety Group and paid social media advertisements on Facebook from October 2020 to February 2021. Advertisements informed potential volunteers that the study assessed the effectiveness of an app for anxiety and depression over 12 weeks. Participation inclusion criteria included the following:

1. 18 years of age or older
2. South African residency
3. Access to an Android phone for the duration of the study
4. Comfortable using the app in English

Participants considered at risk of suicidality or self-harm as measured by a response of 2 and greater on question 9 of the PHQ-9 were excluded from participation in the intervention. Follow-up risk assessments and signposting were offered to these individuals.

A total of 4 participants were excluded based on technical criteria and 3 withdrew. An additional 36 were excluded due to risk concerns (PHQ-9 question 9 score of ≥2). Details of the flow of participants through recruitment and study are shown in Figure 1.

XSL•FO
RenderX

Figure 1. Flowchart of participants. DP2 is at 2 weeks after registration. DP2: data point 2.

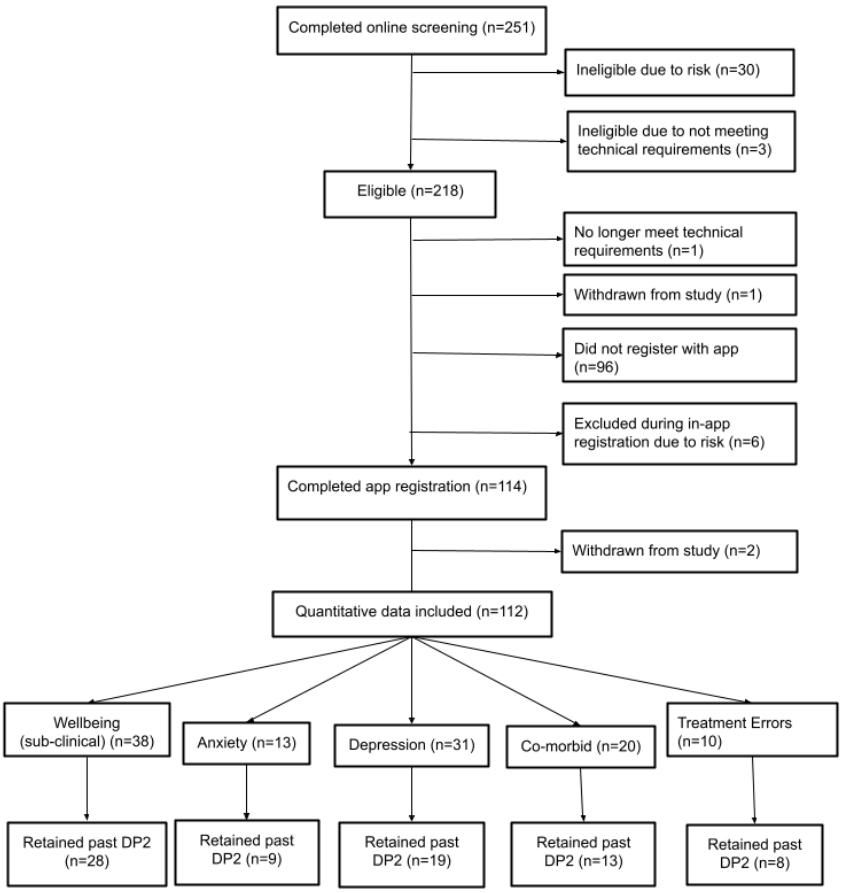

Instrument and Procedure Overview

Interested persons were directed via a web-based link to a Jotform page for further study information and consent. Following digital receipt of study consent, individuals were asked to complete a demographic questionnaire and the PHQ-9 to help assess suitability for study participation. Following review of the consent form, demographic questionnaire, and PHQ-9 for inclusion and exclusion criteria, individuals were advised of their suitability for the study via their preferred communication mode.

Those participants who met the inclusion criteria and none of the exclusion criteria were invited to participate in the intervention and given 2 weeks to download and register with the app (February 1-15, 2021). Technical support was offered to participants to aid this process.

The registration process included the following:

1. Offering some brief information on the frequency of mental health problems and about the app, namely, that it is CBT based and aimed at assessing and treating mental health.
2. Entering their mobile number to get started. This number is used to text message a security code to authenticate the device.
3. Registering their details, including preferred name, email address, age, gender, and employment status.
4. Reading and agreeing to the terms and conditions.
5. Completing the PHQ-9 and receiving feedback on their score.
6. Completing the GAD-7 and receiving feedback on their score.
7. Completing the WSAS and receiving feedback on their score.
8. Completing a brief questionnaire on well-being.
9. Receiving feedback on the overall presentation.

After completing the registration process, participants self-managed their app engagement except for the psychometric measures, administered through pop-ups in the app every 2 weeks. Data were collected directly from the app over 12 weeks (ended May 10, 2021).

Interventions

Following in-app registration, participants were placed into 1 of the following 6 treatment courses, based on their symptom profile:

1. Depression courses (PHQ-9 >10 or WSAS >20)
2. Anxiety course (GAD-7 >10)
3. Comorbid depression and anxiety course, addressing depression first (PHQ-9 and GAD-7 both >10)
4. Comorbid depression and anxiety course, addressing anxiety first (PHQ-9 and GAD-7 both >10)
5. Subclinical course focused on general well-being (for users who present with none to minimal symptom severity)
6. Low activity (subclinical scores, WSAS >20)

Users in the low activity group receive the same treatment course as those in the depression group.

XSL•FO
RenderX

Our courses are structured around CBT-based modules, which focus on skills development and encourage real-world practice. To support participants, we provide in-app tools and skills pages that facilitate the completion of suggested practice exercises, reflection, and skill integration. Measures are readministered fortnightly in the app. These allow the user to reflect on the impact their new skills are having on their symptom profiles. While no additional support is provided outside technical help, 21 random pop-up notifications are strategically used to encourage engagement. Details of course content and tools are included in Multimedia Appendix 2.

Outcome Measures

App usage and engagement were measured by tracking interactions with the various aspects of the app, including content, skills, and tools. We also considered the percentage of overall content completed with the specific treatment course. Unfortunately, we could not reliably use the time stamp information to calculate the overall time spent on the app and individual session data due to possible inconsistencies in how these back-end data were collected. Participants were considered retained if they entered at least 2 complete sets of measures (PHQ-9, GAD-7, and WSAS), including those completed during registration.

In-app outcome measures allow the users to track their progress and symptom changes as a result of their efforts. These include the PHQ-9, GAD-7, and the WSAS [26-30].

1. PHQ-9: 9 items, 4-point Likert scale, measures depression severity (cutoffs: 5, 10, 15, and 20).
2. GAD-7: 7 items, 4-point Likert scale, measures anxiety severity (cutoffs: 5, 10, and 15).
3. WSAS: 5 items, 8-point Likert scale, measures functional impairment (cutoffs: 10 and 20).

Both the PHQ-9 and the GAD-7, widely used in research and clinical practice in health care and community settings, have shown reliable sensitivity within the general population [31,32]. These measures have also been tested for reliability in South Africa [33-36]. The WSAS, on the other hand, has proven to be a reliable and valid measure of impaired functioning, offering enough sensitivity to be an effective outcome measure [36,37]. All 3 measures were self-administered within the app, at registration, and every 2 weeks after that for the duration of the study, ensuring the credibility of our results. Reviewing pre-post changes on these measures allowed us to look at how clinical recovery may impact discontinuation with the app, with researchers theorizing that clinical recovery may coincide with discontinuation of the app in line with the underpinning theoretical model. Participants met clinical recovery if their final scores were below 10 on the GAD-7 and PHQ-9 after having initially scored above 10 on either measure.

Statistical Analysis

All data analysis was conducted using IBM SPSS Statistics software (version 28; IBM Corp). Unless otherwise stated, the significance level for all tests was set to α=.05.

Demographic differences based on intervention uptake, retention, and recovery status were analyzed using chi-square and independent 2-tailed *t* tests. Frequency tables captured app usage and engagement. Unfortunately, due to technical issues, we could not accurately collate time-relative session data and, as such, could not reliably calculate the total number of usage sessions or the total amount of time spent on the app. Due to the small number of participants in the 2 comorbid streams, these were collapsed into a single group for analysis. Pre-post comparisons of symptoms levels on the PHQ-9 and GAD-7 allowed us to establish clinical recovery status. For these calculations, data collected during registration were compared with the last full set of primary measures completed in the app, regardless of the time point at which this was submitted (so entry and exit measures only). Therefore, any participants who completed only some of the primary measures would have their previous complete set of measures used for analysis.

Results

Participant Characteristics

A total of 218 people were invited to download and register the app. Moreover, 47% (104/218) of those invited did not register or no longer met criteria to participate in treatment. A total of 114 participants registered with the app; however, 2 withdrew from the study, leaving a total sample size of 112 participants.

The mean age of participants (n=112) was 35.21 (SD 9.6) years. The average PHQ-9 score was 9.58 (SD 6.02). In terms of ethnicity, 35% (39/112) identified as Black African, 34% (38/112) as White, 10% (11/112) as Coloured, 5% (6/112) as Indian, and 16% (18/112) as belonging to another ethnic group. The gender distribution was predominantly female, with 83% (90/112) identifying as female and 17% (21/112) as male.

Regarding primary language, 11% (12/112) of the participants spoke Afrikaans, 64% (72/112) spoke English, 11% (12/112) spoke isiZulu, and 14% (16/112) spoke another language. When asked about their current health condition, 43% (48/112) reported having a health condition, 53% (59/112) reported no health condition, and 5% (5/112) preferred not to say. Furthermore, 65% (31/48) of those who reported having a health condition reported a mental health issue. Employment status among the participants showed that 61% (69/112) were in some form of employment, 19% (21/112) were unemployed but seeking work, and 20% (22/112) fell into other categories.

When comparing broad uptake across all interested participants, chi-square analysis indicated significantly reduced uptake in participants identifying as “Unemployed but seeking employment” (χ^2_4=10.47; N=251; *P*=.03). Participants identifying with this employment category had the highest representation of those who did not register with the app (49/139, 35.3%). Also, those who categorized themselves as having a mental health diagnosis were significantly more likely to register (31/112, 28%) than those who did not identify as having a mental health diagnosis (81/112, 72%; χ^2_1=8.56; N=251; *P*=.003). No other significant demographic differences were found based on intervention uptake (all *P*>.05).

Moreover, 69% (77/112) of users were retained using a 2-week metric, with 31% (35/112) dropping out before completing a

XSL•FO
RenderX

second set of measures in the app. Also, 45% (50/112) of users submitted a third set of measures, 29% (33/112) a fourth set, 22% (25/112) a fifth set, 12% (14/112) a sixth, and with only 5% (6/112) submitting a sixth set of measures at 12 weeks.

Table 1 looks at retention rates based on treatment group and symptom severity level. Moreover, 6% (2/35) of those participants who dropped out were missing treatment course data due to technical problems. Analysis of demographic information found no statistically significant differences associated with study retention (all *P*>.05). These percentages correspond to the overall proportionality in each treatment group.

Ten participants were missing treatment course information or had been placed in the incorrect treatment course due to technical error. Two (20%) of these 10 participants dropped out and 8 (80%) were retained. Of those who were retained, 7 (88%) of the 8 met the clinical symptom severity level criteria—neither of the participants who dropped out met the criteria for clinical symptom severity level (0/2, 0%).

Table 1. Comparison of treatment group allocation and symptom severity levels between retained and dropped out study participants.

Variable	Retained (n=77)	Dropped out (n=35)
Participants (n=112), n (%)	77 (69)	35 (31)
Treatment group, n (%)		
Subclinical group	32 (42)	10 (29)
Depression group	21 (27)	12 (34)
Comorbid group	13 (17)	7 (20)
Anxiety group	11 (14)	4 (11)
Unknown due to technical error	N/A[a]	2 (6)
Depressive symptom severity, n (%)		
Minimal	13 (17)	7 (20)
Mild	25 (33)	7 (20)
Moderate	21 (27)	17 (49)
Moderately severe	12 (16)	4 (11)
Severe	6 (8)	N/A
Generalized anxiety disorder symptom severity, n (%)		
Minimal	17 (22)	9 (26)
Mild	26 (34)	13 (37)
Moderate	23 (30)	9 (26)
Severe	11 (14)	4 (11)

[a]N/A: not applicable.

App Usage

When considering overall app usage for the entire cohort (n=69, excluding participants with any technical issues), there was a mean of 72.87 (SD 71.425; range 0-278; mode 0) total module pages read, a mean of 3.36 (SD 2.169; range 0-10; mode 3) total modules accessed, and a mean of 19.93 (SD 27.517; range 0-126; mode 0) times engaging with tools and skills. Users (n=69) completed a mean of 30% (SD 29.473%; range 0%-100%; mode 0%) of prescribed content. Table 2 compares app usage and engagement across treatment groups.

There were no significant differences in app usage between those who met clinical recovery criteria and those who did not. Similarly, no significant correlations were found between changes in primary measures and app usage (all *P*>.05).

XSL•FO
RenderX

Table 2. App usage across treatment groups (total n=69, excluding users with technical errors).

Group	Module pages read (n)	Tools/skills usage (n)	Prescribed content completed (%)
Depression group (n=19)			
Mean (SD)	69.68 (58.357)	19.05 (22.225)	29.40 (24.623)
Range	210	86	89
Mode	0	8[a]	0[a]
Anxiety group (n=9)			
Mean (SD)	77.67 (84.109)	31.22 (37.181)	28.77 (31.151)
Range	263	119	97
Mode	0[a]	4	0[a]
Comorbid group (n=13)			
Mean (SD)	66.85 (73.546)	12.54 (28.070)	20.07 (22.086)
Range	260	104	78
Mode	18[a]	3	5[a]
Subclinical group (n=28)			
Mean (SD)	76.29 (77.676)	20.32 (27.356)	34.21 (34.832)
Range	223	126	100
Mode	0	0[a]	0

[a]Multiple modes exist. The smallest value is shown.

Symptom Status at Discontinuation

The intervention follows a least intervention first approach, meaning that users are expected to engage with the app as needed until their symptoms return to a healthy level. Since CBT is time limited, the app focuses on real-world integration of the skills and techniques taught, with users expected to reduce engagement or discontinue use as their symptoms improve. Considering this, we have included a review of the recovery status of users based on their last complete set of measures submitted in the app, as an indicator of symptom status at point of discontinuation (last set of complete measures submitted in the app).

Of the 69 users who indicated clinically severe symptom levels at registration, 25% (17/69) met clinical recovery, 35% (24/69) did not meet clinical recovery, and 41% (28/69) did not have initial scores high enough to be considered before discontinuing use (last set of measures). Of those with clinical symptom levels at registration, 42% (17/41) met clinical recovery and 59% (24/41) did not meet clinical recovery based on their last submitted measures.

Moreover, 58% (11/19) of the depression course, 44% (4/9) of the GAD course, and 31% (4/13) of the comorbid course met recovery on the GAD-7 in their final scores. Also, 53% (10/19) of the depression course, 44% (4/9) of the GAD course, and 46% (6/13) of the comorbid course met recovery on the PHQ-9 in their final scores.

Based on current evidence, UK National Institute for Clinical Excellence guidelines recommend offering low-intensity CBT, including computerized CBT, for those individuals presenting with mild to moderate symptom levels and not for those individuals with moderate to severe symptom levels [38-41]. Considering this, Table 3 shows retention and clinical recovery status at discontinuation, when removing participants with severe symptom levels at registration.

Table 3. Retention and clinical recovery status at discontinuation or study end (excluding participants with severe symptom levels).

	Total participants, n	Dropped out, n (%)	Retained past DP2[a], n (%)	Clinical recovery met, n (%)
Clinical cohort	51	20 (39)	31 (61)	16 (52)
Depression group	30	12 (40)	18 (60)	9 (50)
Anxiety group	4	1 (25)	3 (75)	3 (100)
Comorbid group	17	7 (41)	10 (59)	4 (40)

[a]DP2: Data point 2.

XSL•FO
RenderX

Discussion

Principal Findings

This pilot study sought to test the feasibility of a self-guided, CBT-focused, mental health app–based intervention for South African adults who are experiencing mild to severe symptoms of depression and anxiety. Using an app specifically designed for the South African environment, we recruited 112 South African adults online and through support services, who were interested in using a mental health app. Participant engagement with the app was completely self-driven, with only technical support offered to users. Study retention rates and app usage indicate great potential for mHealth-based treatment; however, further technical development may be necessary to encourage greater engagement with content and tool use. In line with the treatment model, data suggest that most users who discontinued usage beyond the retention period did so after meeting clinical recovery. Although small, this pilot study is one of the first to test the usage case for an app-based intervention designed, developed for, and delivered within the South African context using a CBT model. Overall, these results support the continued exploration of using app-based interventions within the South African context to support the diagnosis and treatment of depression and anxiety.

In their survey of a low-income community in Johannesburg, van Rie and colleagues [42] found that 23% (283/1230) of the participants reported mild to severe depression. In our study, 27% (21/77) of the participants reported clinically significant depressive symptoms and 14% (11/77) reported clinically significant anxiety levels. Moreover, 17% (13/77) of the participants reported clinically significant depression and anxiety symptoms. Comparing our findings with a 2022 national survey where 25.7% (874/3402) of respondents reported PHQ-9 scores of 10 and greater and 17.8% (606/3402) reported scores of 10 and greater on the GAD-7 allows us to see better the impact the COVID-19 lockdown had on mental health [2].

Several authors have reflected on the disparity between individual enrollment and initiation in self-guided web-based interventions, with a systematic review finding that 33% to 88% of users who downloaded the app used it at least once [43-45]. Our uptake and retention rates are comfortably on the higher end of this scale and offer further evidence supporting the intervention. Other mental health app–based intervention studies report retention rates ranging from 5.5% to 70% after 30 days [46,47]. Our retention rates ranged from 69% at 2 weeks, 29% at 8 weeks, and 5% at 12 weeks. This steady reduction in retention is in line with the app model, which focuses on real-world integration of skills and learning with users expected to use the app for only as long as they need to reduce their symptoms to healthy levels. It further reflects the differing lengths of courses and individual progression through the intervention.

There was significantly reduced uptake in those identified as "unemployed but seeking employment." This association may be due to issues with communication during the recruitment and app registration process, especially since subsequent analysis found no significant relationship between employment and retention. Several mHealth studies have found reduced retention in those who identified as unemployed [46,48]. The significantly increased uptake in those participants who reported having a mental health diagnosis suggests that already having a mental health diagnosis may act as a motivator for seeking support and engaging in research.

We found that users in a clinical course completed an average of 26% of their prescribed content, equating to approximately 2.1 modules completed per user. In a similar study, the participants completed an average of 2.6 (SD 2.7) modules. However, this included some access to coaching [49]. Aizenstros et al [46], in their study of a smartphone-delivered behavioral activation intervention, found that users completed an average of 5.46 (range 1-65; n=238) app-based activities over 30 days. Their findings were comparable with those of Watts et al [50], whose study participants' mean aggregated score for homework completion within the mobile-based intervention was 11 (range 6-17; n=35).

Although usage is comparable, if not higher than other app-based interventions, numerous studies suggest that outcomes may be further improved if a mental health care professional guides the digital intervention [51-53]. Anton et al [49] found increased treatment outcomes and engagement in participants who completed at least 4 modules compared with those who completed fewer, suggesting a minimal dosage effect. However, it is tricky to know since interventions may not be directly comparable in amount and length of content [49]. However, in a similar small-scale pilot, Moghimi et al [54] found significant improvements in the GAD-7 and PHQ-9 for participants of their 9-week mHealth intervention. Their study had better overall retention, suggesting room for improving intervention efficacy through increased treatment adherence. Considering the scalability and impact on accessibility that mHealth interventions offer, even small gains can lead to considerable public health gains [23]. As other authors have noted, this is especially relevant to vulnerable populations, including those who may be geographically isolated and those with limited access to suitable resources [19,55].

As mentioned, there is a significant treatment gap related to the assessment and treatment of mental health disorders, specifically those individuals with a comorbid diagnosis, that is, evidenced by 2 or more mental health difficulties or disorders experienced simultaneously. The applicability of mHealth for people accessing mental health services and the possibility of reducing hospital admissions and improving symptom severity, engagement, and treatment are undeniable for clinical practice. Furthermore, this applicability extends to using mHealth as a stand-alone process or a process used in conjunction with other mental health care interventions, such as individual and group therapy or support groups and psychiatric interventions. The potential for app-based interventions to significantly reduce the cost of treatment, increase access in difficult-to-reach populations (such as low-income or rural communities), and reduce the need for clinician input and time is especially relevant to the South African context where these barriers have a more significant impact [6,9-11,19,56].

XSL•FO
RenderX

The app allows users to track their symptom progression easily, is designed for low data usage, and includes offline functionality. Furthermore, the app can be accessed using a voucher code and text message verification, allowing almost immediate access without professional guidance. This feature increases immediate access to treatment and reduces barriers such as wait times, travel and transport difficulties, and stigma. This is especially noteworthy in light of the findings by Forman-Hoffman et al [57], whose investigation found no significant differences between participants in underserved rural or metropolitan areas regarding usage, satisfaction, and treatment outcomes.

The Vimbo app has multiple usage cases, including as a stand-alone intervention or as a therapeutic tool and resource. This adaptability allows it to be integrated at various care levels within a stepped-care system to effectively create easily personalized intervention options into current health systems [12,23]. Using this stepped-care implementation directly into current health systems encourages a reduced caseload for care providers and better use and distribution of resources [7,10]. Furthermore, research suggests that using a blended CBT model that integrates a digital program may improve working alliance and outcomes, suggesting that treatment as usual may benefit by including access to the Vimbo app [58].

Strengths and Limitations

This pilot study possesses several strengths. One of the primary strengths is its contextual relevance, as it is among the first to explore app-based digital interventions within the South African context. This offers valuable insights into the scalability of such interventions to larger settings and services. The study's naturalistic setting, community orientation, and diverse participant age range offer realistic and generalizable insights into everyday usage and engagement. The research points to future digital intervention developments for rural and urban South Africa and has provided Vimbo Health with data for app improvements.

However, it is important to note the study's limitations, which could impact the generalizability of the findings. While qualitative data on user experiences were collected, they were not included in this report due to the significant amount of quantitative data, with plans to publish these findings separately. The small sample size limits the statistical power of the study, making it challenging to extend the findings broadly. The context of the pandemic may have influenced user registration, engagement, and outcomes, indicating a need for further research in nonpandemic conditions. The reliance on self-reported data could introduce biases and affect the accuracy of the results. The app's availability only in English excluded non-English speakers. Future iterations should consider localizing the app to include all South African languages.

Conclusions

This pilot study takes the first critical steps to establish the feasibility of using self-guided, app-based treatment for anxiety and depression to help address the substantial mental health treatment gap in South Africa. The 69% (77/112) retention of users for a minimum of 2 weeks shows that mHealth interventions are a viable treatment modality, acceptable to the South African community. However, only 52% (114/218) of individuals invited to participate in the research registered with the app, suggesting the potential for greater impact if the access and onboarding process can be refined and improved to increase uptake. Overall, our results show promise for the use of digital technologies in addressing symptoms of anxiety and depression in South African adults struggling with their mental health. However, due to the pilot nature of this study and the limited number of other studies in this field within South Africa that have examined the use of digital interventions, further studies are needed to better understand treatment parameters including efficacy and dosage response. Large-scale, longitudinal studies are necessary to better assess the impact digital interventions have on mental health symptoms. Findings from this initial study offer support for app-based psychological interventions to bridge the ever-widening treatment gap.

Data Availability

All data generated or analyzed during this study are included in Multimedia Appendices 1-2.

Conflicts of Interest

Both authors are affiliated with Vimbo Health Ltd through personal investment in the company. The lead researcher is both cofounder and chief science officer for Vimbo Health Ltd. The coresearcher remains an advisor to Vimbo Health Ltd. Both have received a salary or payment for their work on this research. Research funded by Vimbo Health Ltd. A potential conflict of interest exists as Vimbo Health Ltd sponsored the research, and the lead researcher is a cofounder and chief science officer. Despite this, strict protocols ensured unbiased data processing and analysis. Raw data are available for further analysis and cross-checking.

Multimedia Appendix 1
Participant information sheet and consent form.
[PDF File (Adobe PDF File), 111 KB - formative_v9i1e54216_app1.pdf]

Multimedia Appendix 2
Vimbo version 1 app outline.
[PDF File (Adobe PDF File), 708 KB - formative_v9i1e54216_app2.pdf]

XSL•FO
RenderX

References

1. Newson JJ, Pastukh V, Thiagarajan TC. Assessment of population well-being with the mental health quotient: validation study. JMIR Ment Health 2022;9(4):e34105 [FREE Full text] [doi: 10.2196/34105] [Medline: 35442210]
2. Craig A, Rochat T, Naicker SN, Mapanga W, Mtintsilana A, Dlamini SN, et al. The prevalence of probable depression and probable anxiety, and associations with adverse childhood experiences and socio-demographics: a national survey in South Africa. Front Public Health 2022;10:986531 [FREE Full text] [doi: 10.3389/fpubh.2022.986531] [Medline: 36388391]
3. Docrat S, Besada D, Cleary S, Daviaud E, Lund C. Mental health system costs, resources and constraints in South Africa: a national survey. Health Policy Plan 2019;34(9):706-719 [FREE Full text] [doi: 10.1093/heapol/czz085] [Medline: 31544948]
4. World Health Organization. Depression and Other Common Mental Disorders: Global Health Estimates. Geneva, Switzerland: World Health Organization; 2017.
5. COVID-19 Mental Disorders Collaborators. Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. Lancet 2021;398(10312):1700-1712 [FREE Full text] [doi: 10.1016/S0140-6736(21)02143-7] [Medline: 34634250]
6. Demyttenaere K, Bruffaerts R, Posada-Villa J, Gasquet I, Kovess V, Lepine JP, et al. Prevalence, severity, and unmet need for treatment of mental disorders in the World Health Organization World Mental Health Surveys. JAMA 2004;291(21):2581-2590. [doi: 10.1001/jama.291.21.2581] [Medline: 15173149]
7. Prince M, Patel V, Saxena S, Maj M, Maselko J, Phillips MR, et al. No health without mental health. Lancet 2007;370(9590):859-877. [doi: 10.1016/S0140-6736(07)61238-0] [Medline: 17804063]
8. Saxena S, Thornicroft G, Knapp M, Whiteford H. Resources for mental health: scarcity, inequity, and inefficiency. Lancet 2007 Sep 08;370(9590):878-889. [doi: 10.1016/S0140-6736(07)61239-2] [Medline: 17804062]
9. Burns JK. The mental health gap in South Africa—a human rights issue. Equal Rights Trust. 2011. URL: https://www.equalrightstrust.org/ertdocumentbank/ERR06_special_Jonathan.pdf [accessed 2023-10-30]
10. Jack H, Wagner RG, Petersen I, Thom R, Newton CR, Stein A, et al. Closing the mental health treatment gap in South Africa: a review of costs and cost-effectiveness. Glob Health Action 2014;7:23431 [FREE Full text] [doi: 10.3402/gha.v7.23431] [Medline: 24848654]
11. Petersen I, Lund C. Mental health service delivery in South Africa from 2000 to 2010: one step forward, one step back. S Afr Med J 2011;101(10):751-757. [Medline: 22272856]
12. Wilhelm S, Weingarden H, Ladis I, Braddick V, Shin J, Jacobson NC. Cognitive-behavioral therapy in the digital age: presidential address. Behav Ther 2020;51(1):1-14 [FREE Full text] [doi: 10.1016/j.beth.2019.08.001] [Medline: 32005328]
13. O'Dea S. South Africa mobile coverage / smartphone penetration 2015-2018. Statista. 2023. URL: https://www.statista.com/statistics/1074377/south-africa-mobile-coverage-and-smartphone-penetration/ [accessed 2020-04-27]
14. Clement F. South Africa: digital population as of January. Statista. 2020. URL: https://www.statista.com/statistics/685134/south-africa-digital-population/ [accessed 2020-04-29]
15. Velghe F. Literacy acquisition, informal learning and mobile phones in a South African township. New York, NY: Association for Computing Machinery; 2013 Dec 07 Presented at: Sixth International Conference on Information and Communication Technologies and Development; December 7-10, 2013; Cape Town, South Africa p. 89-99. [doi: doi:10.1145/2516604.2516615]
16. Anstey Watkins JOT, Goudge J, Gómez-Olivé FX, Griffiths F. Mobile phone use among patients and health workers to enhance primary healthcare: a qualitative study in rural South Africa. Soc Sci Med 2018;198:139-147 [FREE Full text] [doi: 10.1016/j.socscimed.2018.01.011] [Medline: 29335160]
17. Tomlinson M, Rotheram-Borus MJ, Doherty T, Swendeman D, Tsai AC, Ijumba P, et al. Value of a mobile information system to improve quality of care by community health workers. S Afr J Inf Manag 2013;15(1):-9 [FREE Full text] [doi: 10.4102/sajim.v15i1.528] [Medline: 25147730]
18. Tsai AC, Tomlinson M, Dewing S, le Roux IM, Harwood JM, Chopra M, et al. Antenatal depression case finding by community health workers in South Africa: feasibility of a mobile phone application. Arch Womens Ment Health 2014;17(5):423-431 [FREE Full text] [doi: 10.1007/s00737-014-0426-7] [Medline: 24682529]
19. van Belle S, van Belle JPW, Weimann E. A South African mHealth app for depression management and treatment: a survey and requirements analysis. 2017 Presented at: African Conference on Information Systems & Technology (ACIST); September 12-13, 2024; Cape Town, South Africa URL: https://www.researchgate.net/publication/344453865_A_South_African_Mhealth_App_for_Depression_Management_and_Treatment_a_Survey_and_Requirements_Analysis_1
20. Barron P, Peter J, LeFevre AE, Sebidi J, Bekker M, Allen R, et al. Mobile health messaging service and helpdesk for South African mothers (MomConnect): history, successes and challenges. BMJ Glob Health 2018;3(Suppl 2):e000559 [FREE Full text] [doi: 10.1136/bmjgh-2017-000559] [Medline: 29713503]
21. Clark DM, Layard R, Smithies R, Richards DA, Suckling R, Wright B. Improving access to psychological therapy: initial evaluation of two UK demonstration sites. Behav Res Ther 2009;47(11):910-920 [FREE Full text] [doi: 10.1016/j.brat.2009.07.010] [Medline: 19647230]
22. Richards DA, Suckling R. Improving access to psychological therapy: the Doncaster demonstration site organisational model. Clin Psychol Forum 2008;181(181):9-16. [doi: 10.53841/bpscpf.2008.1.181.9]

XSL•FO
RenderX

23. Bennett-Levy J, Richards D, Farrand P, Christensen H, Lau MA, Proudfoot J, et al. Oxford Guide to Low Intensity CBT Interventions. New York, NY: Oxford University Press; 2010.
24. Dua T, Barbui C, Clark N, Fleischmann A, Poznyak V, van Ommeren M, et al. Evidence-based guidelines for mental, neurological, and substance use disorders in low- and middle-income countries: summary of WHO recommendations. PLoS Med 2011;8(11):e1001122 [FREE Full text] [doi: 10.1371/journal.pmed.1001122] [Medline: 22110406]
25. Overview: common mental health problems: identification and pathways to care: guidance. National Institute for Health and Care Excellence. 2011. URL: https://www.nice.org.uk/guidance/cg123 [accessed 2023-06-15]
26. Kroenke K, Spitzer RL, Williams JBW. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
27. Löwe B, Unützer J, Callahan CM, Perkins AJ, Kroenke K. Monitoring depression treatment outcomes with the patient health questionnaire-9. Med Care 2004;42(12):1194-1201. [doi: 10.1097/00005650-200412000-00006] [Medline: 15550799]
28. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
29. Kroenke K, Spitzer RL, Williams JBW, Monahan PO, Löwe B. Anxiety disorders in primary care: prevalence, impairment, comorbidity, and detection. Ann Intern Med 2007;146(5):317-325. [doi: 10.7326/0003-4819-146-5-200703060-00004] [Medline: 17339617]
30. Mundt JC, Marks IM, Shear MK, Greist JM. The work and social adjustment scale: a simple measure of impairment in functioning. Br J Psychiatry 2002;180:461-464. [doi: 10.1192/bjp.180.5.461] [Medline: 11983645]
31. Kocalevent RD, Hinz A, Brähler E. Standardization of the depression screener Patient Health Questionnaire (PHQ-9) in the general population. Gen Hosp Psychiatry 2013;35(5):551-555. [doi: 10.1016/j.genhosppsych.2013.04.006] [Medline: 23664569]
32. Löwe B, Decker O, Müller S, Brähler E, Schellberg D, Herzog W, et al. Validation and standardization of the generalized anxiety disorder screener (GAD-7) in the general population. Med Care 2008;46(3):266-274. [doi: 10.1097/MLR.0b013e318160d093] [Medline: 18388841]
33. Bhana A, Rathod SD, Selohilwe O, Kathree T, Petersen I. The validity of the Patient Health Questionnaire for screening depression in chronic care patients in primary health care in South Africa. BMC Psychiatry 2015;15:118 [FREE Full text] [doi: 10.1186/s12888-015-0503-0] [Medline: 26001915]
34. Cholera R, Gaynes BN, Pence BW, Bassett J, Qangule N, Macphail C, et al. Validity of the Patient Health Questionnaire-9 to screen for depression in a high-HIV burden primary healthcare clinic in Johannesburg, South Africa. J Affect Disord 2014;167:160-166 [FREE Full text] [doi: 10.1016/j.jad.2014.06.003] [Medline: 24972364]
35. Henn C, Morgan B. Differential item functioning of the CESDR-R and GAD-7 in African and white working adults. SA J Ind Psychol 2019;45(6):-10. [doi: 10.4102/sajip.v45i0.1663]
36. van Heyningen T, Honikman S, Tomlinson M, Field S, Myer L. Comparison of mental health screening tools for detecting antenatal depression and anxiety disorders in South African women. PLoS One 2018;13(4):e0193697 [FREE Full text] [doi: 10.1371/journal.pone.0193697] [Medline: 29668725]
37. Zahra D, Qureshi A, Henley W, Taylor R, Quinn C, Pooler J, et al. The Work and Social Adjustment Scale: reliability, sensitivity and value. Int J Psychiatry Clin Pract 2014;18(2):131-138. [doi: 10.3109/13651501.2014.894072] [Medline: 24527886]
38. Firth J, Torous J, Nicholas J, Carney R, Pratap A, Rosenbaum S, et al. The efficacy of smartphone-based mental health interventions for depressive symptoms: a meta-analysis of randomized controlled trials. World Psychiatry 2017;16(3):287-298 [FREE Full text] [doi: 10.1002/wps.20472] [Medline: 28941113]
39. Firth J, Torous J, Nicholas J, Carney R, Rosenbaum S, Sarris J. Can smartphone mental health interventions reduce symptoms of anxiety? A meta-analysis of randomized controlled trials. J Affect Disord 2017;218:15-22 [FREE Full text] [doi: 10.1016/j.jad.2017.04.046] [Medline: 28456072]
40. Karyotaki E, Efthimiou O, Miguel C, Bermpohl FMG, Furukawa TA, Cuijpers P, Individual Patient Data Meta-Analyses for Depression (IPDMA-DE) Collaboration, et al. Internet-based cognitive behavioral therapy for depression: a systematic review and individual patient data network meta-analysis. JAMA Psychiatry 2021;78(4):361-371 [FREE Full text] [doi: 10.1001/jamapsychiatry.2020.4364] [Medline: 33471111]
41. National Institute for Health and Care Excellence. Common mental health problems: identification and pathways to care. 2011. URL: https://www.nice.org.uk/guidance/cg123 [accessed 2024-06-04]
42. van Rie A, West NS, Schwartz SR, Mutanga L, Hanrahan CF, Ncayiyana J, et al. The unmet needs and health priorities of the urban poor: generating the evidence base for urban community health worker programmes in South Africa. S Afr Med J 2018;108(9):734-740. [doi: 10.7196/SAMJ.2018.v108i9.13054] [Medline: 30182898]
43. Eysenbach G. The law of attrition. J Med Internet Res 2005;7(1):e11 [FREE Full text] [doi: 10.2196/jmir.7.1.e11] [Medline: 15829473]
44. Fleming TM, de Beurs D, Khazaal Y, Gaggioli A, Riva G, Botella C, et al. Maximizing the impact of e-therapy and serious gaming: time for a paradigm shift. Front Psychiatry 2016;7:65 [FREE Full text] [doi: 10.3389/fpsyt.2016.00065] [Medline: 27148094]

XSL•FO
RenderX

45. Hanano M, Rith-Najarian L, Boyd M, Chavira D. Measuring adherence within a self-guided online intervention for depression and anxiety: secondary analyses of a randomized controlled trial. JMIR Ment Health 2022;9(3):e30754 [FREE Full text] [doi: 10.2196/30754] [Medline: 35343901]
46. Aizenstros A, Bakker D, Hofmann SG, Curtiss J, Kazantzis N. Engagement with smartphone-delivered behavioural activation interventions: a study of the moodMission smartphone application. Behav Cogn Psychother 2021;49(5):569-581. [doi: 10.1017/S1352465820000922] [Medline: 34396942]
47. Bakker D, Kazantzis N, Rickwood D, Rickard N. A randomized controlled trial of three smartphone apps for enhancing public mental health. Behav Res Ther 2018;109:75-83. [doi: 10.1016/j.brat.2018.08.003] [Medline: 30125790]
48. Kannisto KA, Korhonen J, Adams CE, Koivunen MH, Vahlberg T, Välimäki MA. Factors associated with dropout during recruitment and follow-up periods of a mHealth-based randomized controlled trial for mobile.Net to encourage treatment adherence for people with serious mental health problems. J Med Internet Res 2017;19(2):e46 [FREE Full text] [doi: 10.2196/jmir.6417] [Medline: 28223262]
49. Anton MT, Greenberger HM, Andreopoulos E, Pande RL. Evaluation of a commercial mobile health app for depression and anxiety (AbleTo Digital+): retrospective cohort study. JMIR Form Res 2021;5(9):e27570 [FREE Full text] [doi: 10.2196/27570] [Medline: 34546170]
50. Watts S, Mackenzie A, Thomas C, Griskaitis A, Mewton L, Williams A, et al. CBT for depression: a pilot RCT comparing mobile phone vs. computer. BMC Psychiatry 2013;13:49 [FREE Full text] [doi: 10.1186/1471-244X-13-49] [Medline: 23391304]
51. Duarte-Díaz A, Perestelo-Pérez L, Gelabert E, Robles N, Pérez-Navarro A, Vidal-Alaball J, et al. Efficacy, safety, and evaluation criteria of mHealth interventions for depression: systematic review. JMIR Ment Health 2023;10:e46877 [FREE Full text] [doi: 10.2196/46877] [Medline: 37756042]
52. Oehler C, Scholze K, Reich H, Sander C, Hegerl U. Intervention use and symptom change with unguided internet-based cognitive behavioral therapy for depression during the COVID-19 pandemic: log data analysis of a convenience sample. JMIR Ment Health 2021;8(7):e28321 [FREE Full text] [doi: 10.2196/28321] [Medline: 34115604]
53. Oser M, Wallace ML, Solano F, Szigethy EM. Guided digital cognitive behavioral program for anxiety in primary care: propensity-matched controlled trial. JMIR Ment Health 2019;6(4):e11981 [FREE Full text] [doi: 10.2196/11981] [Medline: 30946022]
54. Moghimi E, Stephenson C, Agarwal A, Nikjoo N, Malakouti N, Layzell G, et al. Efficacy of an electronic cognitive behavioral therapy program delivered via the online psychotherapy tool for depression and anxiety related to the COVID-19 pandemic: pre-post pilot study. JMIR Ment Health 2023;10:e51102 [FREE Full text] [doi: 10.2196/51102] [Medline: 37993984]
55. Forsythe AM, Venter C. Behavioral economics, motivating psycho-education improvements: a mobile technology initiative in South Africa. Front Psychol 2019;10:1560 [FREE Full text] [doi: 10.3389/fpsyg.2019.01560] [Medline: 31354576]
56. Kim J, Aryee LMD, Bang H, Prajogo S, Choi YK, Hoch JS, et al. Effectiveness of digital mental health tools to reduce depressive and anxiety symptoms in low- and middle-income countries: systematic review and meta-analysis. JMIR Ment Health 2023;10:e43066 [FREE Full text] [doi: 10.2196/43066] [Medline: 36939820]
57. Forman-Hoffman VL, Pirner MC, Flom M, Kirvin-Quamme A, Durden E, Kissinger JA, et al. Engagement, satisfaction, and mental health outcomes across different residential subgroup users of a digital mental health relational agent: exploratory single-arm study. JMIR Form Res 2023;7:e46473 [FREE Full text] [doi: 10.2196/46473] [Medline: 37756047]
58. Doukani A, Quartagno M, Sera F, Free C, Kakuma R, Riper H, et al. Comparison of the working alliance in blended cognitive behavioral therapy and treatment as usual for depression in Europe: secondary data analysis of the E-COMPARED randomized controlled trial. J Med Internet Res 2024;26:e47515 [FREE Full text] [doi: 10.2196/47515] [Medline: 38819882]

Abbreviations

CBT: cognitive behavioral therapy
CMD: common mental disorder
GAD-7: Generalized Anxiety Disorder Assessment-7
HSRC: Human Sciences Research Council
mHealth: mobile health
PHQ-9: Patient Health Questionnaire-9
REC: Research Ethics Committee
WSAS: Work and Social Adjustment Scale

XSL•FO
RenderX

Edited by A Mavragani; submitted 01.11.23; peer-reviewed by M Fernández-González; comments to author 18.02.24; revised version received 04.06.24; accepted 10.10.24; published 30.01.25.

Please cite as:
Steyn S, Slabbert M
Self-Guided Smartphone App (Vimbo) for the Reduction of Symptoms of Depression and Anxiety in South African Adults: Pilot Quantitative Single-Arm Study
JMIR Form Res 2025;9:e54216
URL: https://formative.jmir.org/2025/1/e54216
doi:10.2196/54216
PMID:

©Sherrie Steyn, Meggan Slabbert. Originally published in JMIR Formative Research (https://formative.jmir.org), 30.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Co-Designing a Digital Brief Intervention to Reduce the Risk of Prescription Opioid–Related Harm Among People With Chronic Noncancer Pain: Qualitative Analysis of Patient Lived Experiences

Rachel A Elphinston[1,2,3,4], PhD; Sue Pager[1,5], SpPath, MPH; Kelly Brown[1,4,6], MClinPsych; Michele Sterling[1,2], PhD; Farhad Fatehi[7,8], MD, PhD; Paul Gray[9,10], MBBS, PhD; Linda Hipper[4], Grad Dip, MPH; Lauren Cahill[1], BSc, BNurs; Jason P. Connor[10,11], PhD

[1]RECOVER Injury Research Centre, The University of Queensland, Herston, Australia

[2]National Health and Medical Research Council Centre for Research Excellence - Better Health Outcomes for Compensable Injury, The University of Queensland, Brisbane, Australia

[3]School of Psychology, The University of Queensland, Brisbane, Australia

[4]Metro South Addiction and Mental Health Services, Metro South Health Hospital and Health Service, Brisbane, Australia

[5]Health Equity & Access Unit, Metro South Hospital and Health Service, Brisbane, Australia

[6]Gallipoli Medical Research Foundation, Greenslopes Private Hospital, Brisbane, Australia

[7]Centre for Online Health, The University of Queensland, Brisbane, Australia

[8]School of Psychological Sciences & Turner Institute for Brain and Mental Health, Monash University, Melbourne, Australia

[9]Tess Cramond Pain and Research Centre, Metro North Hospital and Health Service, Brisbane, Australia

[10]Faculty of Medicine, The University of Queensland, Brisbane, Australia

[11]National Centre for Youth Substance Use Research, The University of Queensland, Brisbane, Australia

Corresponding Author:
Rachel A Elphinston, PhD
RECOVER Injury Research Centre
The University of Queensland
Level 7
296 Herston Road
Herston, 4006
Australia
Phone: 61 412662084
Email: rachelel@uq.edu.au

Abstract

Background: Opioid medications are important for pain management, but many patients progress to unsafe medication use. With few personalized and accessible behavioral treatment options to reduce potential opioid-related harm, new and innovative patient-centered approaches are urgently needed to fill this gap.

Objective: This study involved the first phase of co-designing a digital brief intervention to reduce the risk of opioid-related harm by investigating the lived experience of chronic noncancer pain (CNCP) in treatment-seeking patients, with a particular focus on opioid therapy experiences.

Methods: Eligible patients were those aged between 18 and 70 years with CNCP at a clinically significant level of intensity (a score of ≥4 of 10). Purposive sampling was used to engage patients on public hospital waitlists via mail or through the treating medical specialist. Participants (N=18; n=10 women; mean age 49.5 years, SD 11.50) completed semistructured telephone interviews. Interviews were transcribed verbatim, thematically analyzed using grounded theory, and member checked by patients.

Results: Eight overarching themes were found, listed in the order of their prominence from most to least prominent: limited treatment collaboration and partnership; limited biopsychosocial understanding of pain; continued opioid use when benefits do not outweigh harms; a trial-and-error approach to opioid use; cycles of hopefulness and hopelessness; diagnostic uncertainty; significant negative impacts tied to loss; and complexity of pain and opioid use journeys.

XSL•FO
RenderX

Conclusions: The findings of this study advance progress in co-designing digital brief interventions by actively engaging patient partners in their lived experiences of chronic pain and use of prescription opioid medications. The key recommendations proposed should guide the development of personalized solutions to address the complex care needs of patients with CNCP.

(JMIR Form Res 2025;9:e57208) doi:10.2196/57208

KEYWORDS

chronic noncancer pain; prescription opioid use; brief intervention; co-design; patient partners; lived experience; qualitative; digital health

Introduction

Opioid therapy is recommended as part of a multimodal approach to manage acute, postoperative, and cancer pain [1]. There is considerably less evidence on the effectiveness of opioid therapy for patients with chronic noncancer pain (CNCP) [2,3], and meta-analyses show modest CNCP relief [4]. The risks of adverse side effects [5], including opioid-related overdose death [6,7], remain a major public health concern [8-11], especially as (1) access to interdisciplinary rehabilitation is restricted by cost, treatment intensity, geography, demand [12-16], and patient fears about opioid discontinuation [17-19] and (2) the effectiveness of behavioral and pharmacological treatments to reduce the risks of opioid use in patients with CNCP is inconclusive [20,21].

Digital brief interventions (BIs) may offer a nonpharmacological option to reduce opioid-related harm [22,23]. BIs are widely used in substance use treatment, and they involve screening, feedback assessment, education, and goal setting, motivating those at risk to modify substance use and engage in further treatment if required [24]. BIs are effective in reducing substance use and related harm in various settings (eg, primary care and emergency departments) [25,26] and have shown similar effectiveness to more intensive psychological therapies [27,28]. Digital BIs are becoming an increasingly attractive, scalable technological innovation in the management of pain that may be just as effective as in-person BIs in reducing substance use and related harm [29]. The use of digital BIs in the CNCP population is still in its infancy, but there is interest among patients in these types of interventions [23,30].

Broadly speaking, as effective as digital health interventions (DHIs) are, they are not widely adopted in practice [31,32]. One of the main reasons is that key stakeholders (eg, patients and health professionals) have not been a part of the solution design [33-36]. Research co-design, that is, the meaningful involvement of end users in the design, implementation, and translation of research, has the potential to reduce the risk of research failing to translate into practice [37]. Co-design offers a means to increase the success of DHIs by aligning with user needs and preferences [38,39].

Co-design of DHIs for CNCP has not made significant progress over the past decade [40]. Most DHIs for pain or medication use are developed by the for-profit software industry [41], lack a theoretical framework [42] and evidence base [43]. Only 19% of studies involve end users in the development process, and this is usually in an ad hoc manner [44]; few studies support clinician access or involvement [42,44,45].

Using a co-design approach with patient partners makes it possible to design a fit-for-patient-purpose digital BI, which may accelerate its adoption and impact considerably. The first step of the co-design of DHIs is to obtain a rich and deep understanding of people's experiences within the disease management context while also sensitizing them to the problem space [46,47]. Despite the diverse and unique lived experiences of those with CNCP (eg, [48,49]), DHIs are rarely designed in partnership with this patient population. This study aimed to understand the lived experiences of patients with CNCP with a particular focus on opioid therapy—the first important step of co-design. The patients were informed of the overall goal to co-design a digital BI and then asked to describe their experiences of CNCP (past, present, and future) management, including opioid therapy. This new knowledge will inform future steps in the co-design process of digital BIs for the population that is most at risk of harmful use of opioid medicines.

Methods

Participants

Participants were eligible if they were aged between 18 and 70 years, living with CNCP (defined as a sensory and emotional experience that lasts or recurs for ≥3 months), and seeking treatment from public health specialist addiction or pain services. Participants were also required to report clinical levels of pain, as indicated by scoring ≥4 out of 10 on the Pain Numerical Rating Scale on average over the past week.

Participants were excluded if they were assessed as being highly distressed (based on scores ≥13 on the Kessler Psychological Distress Scale [50] and follow-up telephone risk assessments by the research team psychologists [RE and KB]); were non–English speaking; had a history of recent injecting drug use; or had currently tested positive with SARS-CoV-2. In addition, patients recruited from addiction services were excluded based on the clinical judgment from the addiction specialist if they had an acute, severe mental illness (eg, active psychosis).

Design

A co-design approach [51] was used. We applied the co-design framework by Sanders and Stappers [47] and updated by Noorbergen et al [46] during the co-design of the digital BI. It outlines 6 iterative co-design phases: predesign, generative, prototyping, evaluative, implementation, and postdesign. Figure 1 presents an overview of the co-design framework. This study reports on the first phase of the iterative co-design development process – the exploration of patient experiences as part of predesign (ie, contextual inquiry). A systematic review

XSL•FO
RenderX

examining the state of the evidence and an adjunctive study examining patient preferences for BIs to reduce pain, opioid use, and related harm was also conducted as part of the first phase and is reported elsewhere [23].

Figure 1. Overview of the iterative co-design framework used in this study. The predesign phase, denoted by the box highlighted in green, is reported in this study.

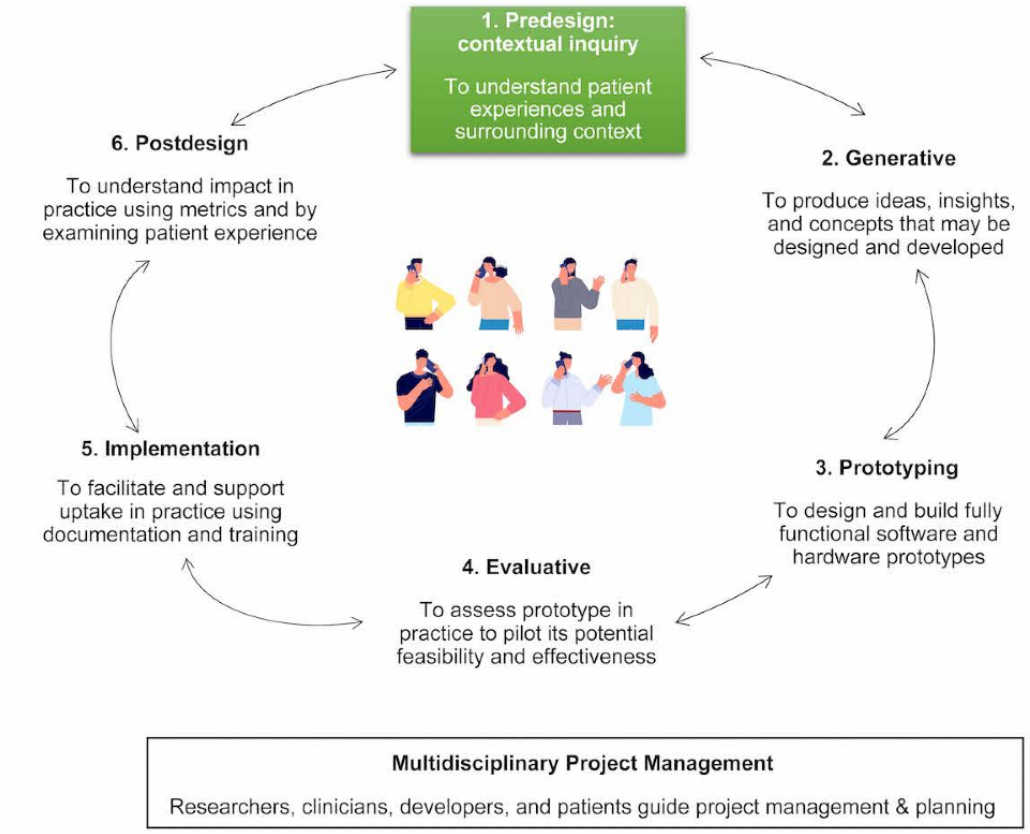

Recruitment

Treatment-seeking patients who were currently using prescription opioids or had a history of opioid use were purposively sampled. Participants with CNCP were recruited through a tertiary hospital specialist pain clinic or community addiction specialist service. Patients from the pain clinic were engaged by mailing an information pack inviting those on the waitlist to participate. Participants from the hospital addiction service were invited by the medical specialist and a member of the research team (RE) during their on-site appointment at the clinic. The final sample size was determined by thematic saturation using the approach outlined by Guest et al [52], applied during the data collection and inductive thematic analysis phase using 3 components: base size, run length, and new information threshold. The base size refers to the body of information identified in the initial dataset (ie, initial number of themes) and is used as a denominator of the saturation ratio. The run length refers to the number of interviews or other data collection events based on which new information or themes are identified. New information threshold refers to the level of information accepted as indicative of saturation. A new information threshold of ≤5% was selected to indicate that we had reached adequate saturation.

Procedure

Recruitment and interviews were completed from June to November 2020. Participants completed an initial telephone screening interview for us to determine eligibility. If eligible, patients completed a web-based questionnaire before the interview for descriptive purposes and to provide background to guide the interview. Interviews were audio-recorded, professionally transcribed verbatim, and thematically analyzed. Member checking of themes by patients who participated in the interviews was conducted during a follow-up workshop, which focused on the design components of the BI (not reported in this study). Figure 2 presents a summary of patient flow.

XSL•FO
RenderX

Figure 2. Patient flow.

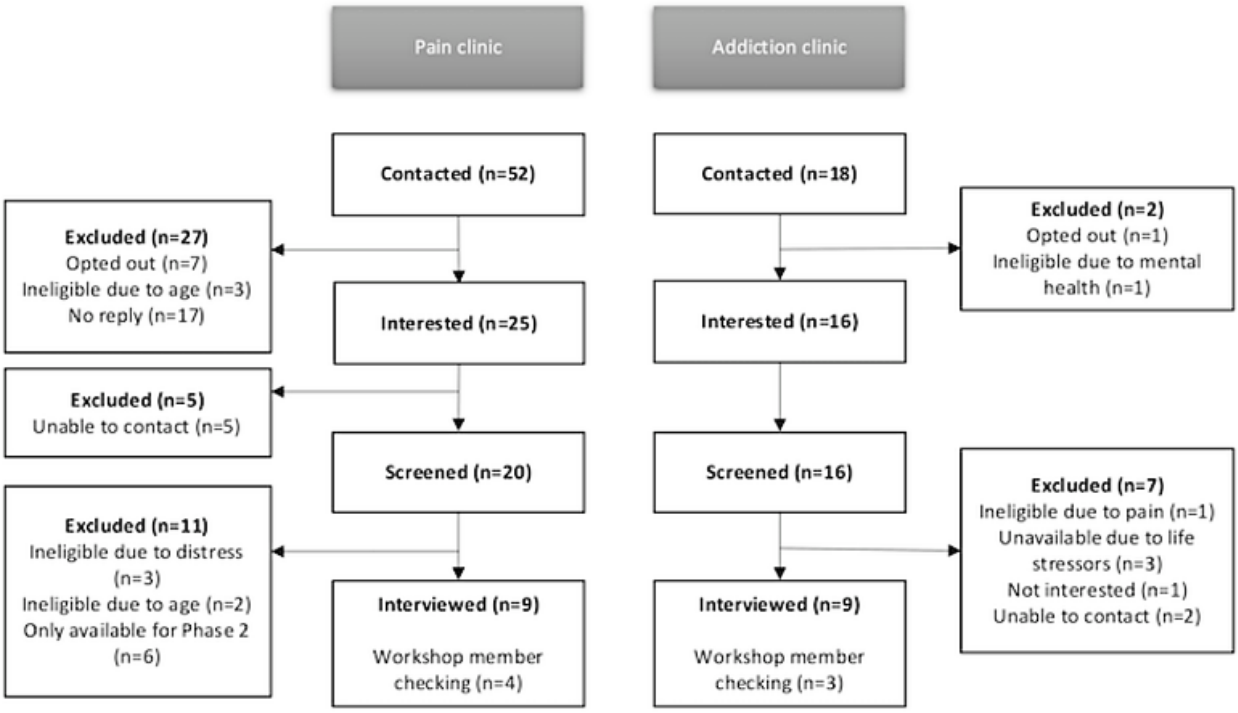

Measures

Demographics

Demographic data, including age, sex, and cultural background, as well as information on pain location, duration, intensity, and current medications were collected. Data on history of mental health or substance use disorder were also collected.

Brief Pain Inventory

The Brief Pain Inventory [53,54] is an 8-item measure that assesses the pain experience, including location of pain, pain intensity and interference on a scale from 0 (no pain) to 10 (pain as bad as you can imagine). Scores are summed to produce 2 separate scores of pain intensity and pain interference, with higher levels indicating greater intensity and interference. An example item is as follows: “Please rate your pain by circling the one number that best describes your pain at its worst in the last week.” The Brief Pain Inventory is a validated and widely used measure in clinical settings with efficacy in both patients with cancer pain and those with chronic, noncancer pain [53,54].

Kessler Psychological Distress Scale

The Kessler Psychological Distress Scale [50] is a 6-item measure used to assess a person’s level of psychological distress. Participants were asked how often they have felt a range of emotions over the past 4 weeks (eg, “How often did you feel anxious?”). Responses are scored on a 5-point scale (0=none of the time; 4=all of the time). Scores ≥13 indicate the presence of a possible mental illness [50]. This scale has demonstrated excellent reliability [50].

Depression, Anxiety, and Stress

The 21-item Depression Anxiety Stress Scales [55] asked participants to rate to what degree depressive, anxiety, or stress symptoms have impacted them in the previous week on a scale from 0 (did not apply to me at all) to 3 (very much or most of the time). Items from each subscale were summed and multiplied by 2 to denote a patient’s current level of depression, with higher scores reflecting higher levels of depression. An example item from the depression subscale is as follows: “I felt that I had nothing to look forward to.” The 21-item Depression Anxiety Stress Scales has previously been shown to be a reliable and valid tool in a chronic pain sample (Cronbach α=0.96 for depression, Cronbach α=0.89 for anxiety, and Cronbach α=0.95 for stress) [56].

Current Opioid Misuse Measure

The Current Opioid Misuse Measure (COMM) [57] consists of 17 items measured on a 0 (never) to 4 (very often) scale, which asks patients about their current opioid use behaviors. Scores are summed, with higher totals reflecting higher levels of opioid misuse. An example item is as follows: “In the past 30 days, how often have you taken your medication differently to how they are prescribed?” The COMM is a validated and reliable measure within chronic pain samples, with previous research reporting Cronbach α=0.86 [57-59].

Semistructured Interview

Semistructured interview questions explored patient experiences and their initial perspectives of a digital BI (Multimedia Appendix 1). Participants were first informed about the overall aim of the research, to co-design a digital BI to reduce risk of unsafe use of prescription opioid medications. They were then asked about their past, current, and future experiences of pain and pain management; access to treatment services and programs including psychologists and psychological treatment; and specifically, their experience of using prescription opioids. Then, suggestions for the design and implementation of the digital BI

XSL•FO
RenderX

were explored; however, these will be reported in a separate paper as part of the second (generative) phase of the co-design process [46,47]. All interviews were conducted by RE, a clinical psychologist, and KB, a psychologist, both with expertise in treatment of chronic pain and qualitative research. Field notes were made after the interviews to record the interviewer's observations and impressions.

Data Analysis

Transcripts were thematically analyzed using grounded theory [60]. Two independent coders (RE and SP) read the transcripts to establish familiarity. An inductive approach to coding was used, with codes added as they were identified in the data. The coding structure consisted of first level codes recorded line-by-line, second level codes of possible subthemes, and third level codes of higher order themes. Key quotes were extracted. The analysis was conducted iteratively, with codes cross-checked between coders approximately every 3 to 5 transcripts to identify any discrepancies, which were resolved by further discussion. Theme refinement was undertaken between the coders and another interviewer (KB), integrating the interviewer's field notes to aid interpretation and reduce coder bias.

Member checking [61] was conducted in a subsample (n=7) of interviewed participants in subsequent workshops to explore the validity of the themes as part of the assessment of data saturation approach, that is, to check whether the themes were consistent with the participant's experience. This involved facilitators providing verbal feedback on the themes integrated with the workshop activities. We sought disconfirming voices (objectivism) and provided opportunities for reflection on personal experiences, creating opportunities to add data (constructivism) [62].

Ethical Considerations

Ethics approval was obtained from the relevant institutional review boards (HREC/2020/QMS/60695). All participants provided written informed consent. Study data were deidentified for analysis, and any identifying information was removed before publication, including patient numbers, to ensure complete anonymity. All participants were offered an Aus $50 (US $33) gift voucher for each session attended to thank them for their time.

Results

Data Saturation Analysis

Table 1 presents a summary of the data saturation analysis. We reached adequate saturation (≤5% threshold) after 16^{+2} in-depth interviews. The final 2 extra interviews did not add substantially to the body of information collected; however, these ensured a purposive sampling approach included a balance of male and female participants and people from various culturally and linguistically diverse backgrounds.

Table 1. Data saturation analysis.

Step	Outcome
1. Number of themes for the base interviews[a]	8 × PID_P[b] interviews
2. Number of unique themes for the first run	5 new themes from 4 × PID_A[c] interviews
3. Saturation ratio	Number of new themes/run: 5/30 = 17% (>5% threshold); continue with interviews
4. Number of new, unique themes for the next run	2 new themes from 5 × PID_A interviews and interviewer reflections: 2/30=7%
5. Updated saturation ratio	7% (>5% threshold); continue with interviews
6. Number of new unique themes for the next run	1 new theme from 2 × workshop member checking: 1/30=3%
7. Updated saturation ratio	3%, less than ≤5% threshold
8. Number of new unique themes for the next run (to address gaps in demographics)	1 new theme from 1 × PID_A and 1 × PID_P interview: 1/30=3%
9. Updated saturation ratio	3%, ≤5% threshold

[a]Denominator in the equation.

[b]PID_P: patients from the pain clinic.

[c]PID_A: patients from specialist addiction services.

Sample Characteristics

The sample included 10 women and 8 men, with a mean age of 49.5 (SD 11.50; range 25-62) years. One-third of the patients self-identified an ethnicity in addition to Australian descent, including 2 from Aboriginal and Torres Strait Islander descent. Four (22%) patients were employed either part time or full time. Approximately half of the participants reported their highest level of education as grade 12 and 4 (22%) participants had at least an undergraduate degree. More than half of the patients were in a relationship.

The sample predominantly reported back or neck pain (89%). Duration of pain ranged from 6 months to 42 years (mean 11.57, SD 11.55 years; median 7 years). Average pain severity in the past week was 5.82 (SD 1.51), and average pain interference in the past week was 5.80 (SD 1.91). In total, 16 (89%) patients were currently using prescription opioids (range 1.15 mg/d [pro re nata] to 480 mg/d), and 2 (11%) patients had used opioid

XSL•FO
RenderX

therapy in the past. Half of the participants reported current opioid misuse (scores ≥9 on the COMM) [57]. Three (19%) patients were prescribed high-dose opioids (≥100 mg/d) [63]. Seven (39%) patients reported moderate depressive symptoms. Three (17%) reported severe or extremely severe depressive symptoms; these patients were included as they were assessed to have adequate supports and risk management plans, which supported their involvement. Table 2 presents a summary of individual characteristics (see Multimedia Appendix 2 for a more detailed summary of individual patient characteristics).

Table 2. Sample characteristics of patients.

Age (y)	Sex	Pain duration (y)	Pain interference (past week)	Pain severity (past week)	COMM[a]	OME[b]
33	Female	15	2.71	2.25	4	1.15
52	Male	2	9.14	7	5	30
60	Female	3	5.00	6	7	45
31	Male	5	4.14	5	14[c]	60
47	Female	0.5	4.71	6	—[d]	—[d]
61	Female	4	2.86	5	8	—[d]
53	Female	14	4.71	4	10[c]	70
47	Female	0.83	6.57	6.75	20[c]	22.8
58	Female	42	8.86	8	13[c]	143.4
60	Female	2.5	6.43	7.25	2	37.9
54	Male	20	7.43	7.25	25[c]	480
25	Male	3	5.83	6.5	1	4.2
58	Male	11	6.43	5.5	15[c]	N/A[e]
53	Male	20	5.29	5.75	13[c]	N/A[e]
56	Female	28	8.00	6.75	3	N/A[e]
48	Male	9	4.14	6.5	11[c]	90
33	Female	3.5	4.29	3	27[c]	480
62	Male	25	7.71	6.25	16[c]	20

[a]COMM: Current Opioid Misuse Measure.

[b]OME: oral morphine equivalence (daily).

[c]These scores meet the cutoff for the scale as risk of unsafe opioid use.

[d]Missing.

[e]N/A: not applicable; patients were using methadone and dose could not be calculated.

Thematic Analysis

Overview

Eight overarching themes and 13 subthemes were identified from the data, listed in the order of prominence from most to least prominent: limited treatment collaboration and partnership; limited biopsychosocial understanding of pain; continued opioid use when benefits do not outweigh harms; a trial-and-error approach to opioid use; cycles of hopefulness and hopelessness; diagnostic uncertainty; significant negative impacts tied to loss; and complexity of pain and opioid use journeys. Figure 3 provides an overview of the themes; larger circles reflect more prevalent themes. In subsequent sections, note that PID_P refers to patients from the pain clinic and PID_A refers to patients from addiction services. Patient numbers were removed before publication to ensure complete anonymity.

XSL•FO
RenderX

Figure 3. Overview of themes and subthemes. The overlap of circles represents connections between themes, and lines represent connections between subthemes.

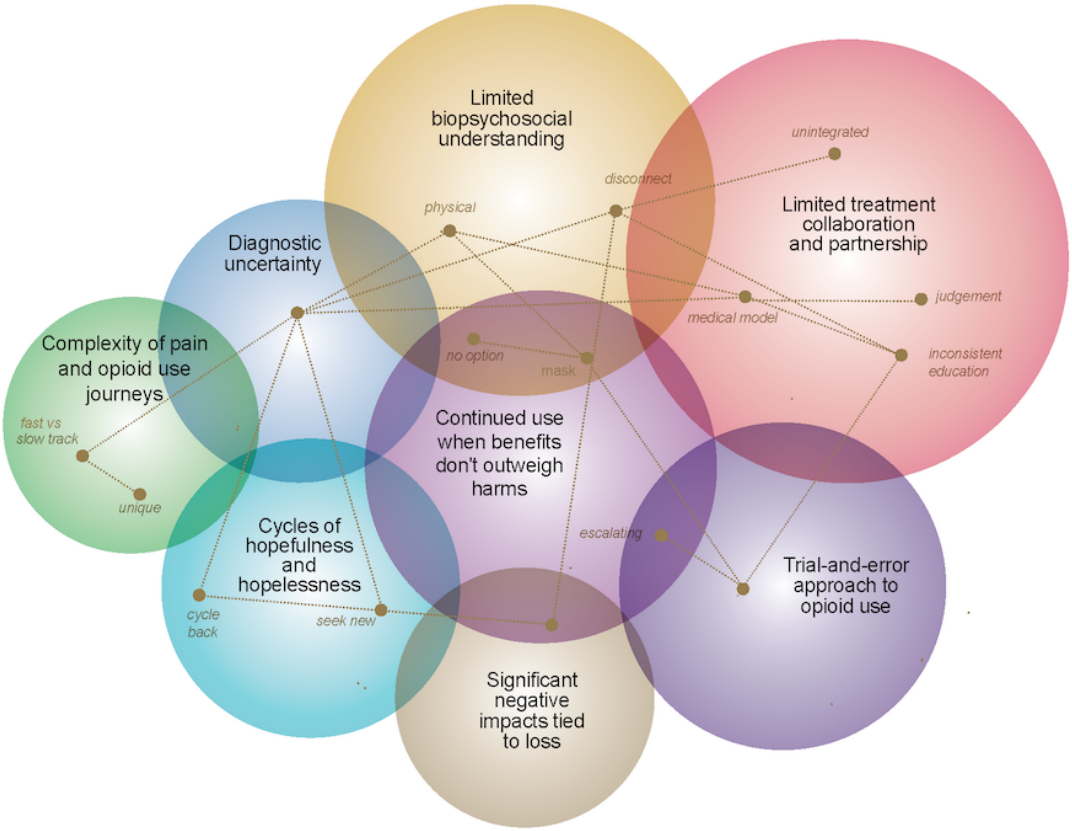

Theme 1: Limited Collaboration and Partnership During Treatment

This theme reflected the lack of a patient–health professional collaborative partnership and an integrated approach among health professionals in the care of patients.

Patients Initially Followed the Medical Model, With Faith and Trust

Initial opioid prescription was usually prompted by the treating physicians in response to an acute episode of pain or following surgery. Patients with a long history of pain described early prescribing practices as including more passive rather than active involvement of patients. One patient (PID_P) reported the following:

> *So the Dr says, look here, take this, this will stop the pain. So, you go, ok, you take it.*

Patients said they often continued to follow the medical model, with faith and trust in the treatment approach guided by their physician. This suggested a lack of shared decision-making about whether opioids were the best or right treatment option and whether the benefits outweighed potential harms—"I'm in the hands of the Doctors" (PID_A). As their pain persisted over time, trust in opioid or medical treatment decreased and participants tried to seek alternatives treatments to manage pain. Some patients reported remaining very trusting of their general practitioner (GP) and were more actively involved in discussing options with them. Others reported becoming distrustful and even angry.

Inconsistent Opioid Use Education, Monitoring, and Support

There was variation in the amount of education about opioid use that the patients received when first prescribed. Some patients said they received education from their physician or pharmacist about the potential harmful effects of prescription opioids (eg, side effects and addiction potential). Other participants could not recall receiving any education in this regard:

> *A lot of doctors don't have communication or you know, they'll just walk in, they'll give you the script and walk out. But I think it's very important for the doctors, as the ones prescribing to actually sit there and go, all right, here's the side effects.* [PID_A]

The reported level of ongoing education and medical monitoring of opioid use over time also varied. One patient (PID_P) suggested, "They should have an opiate diary and they should make people keep diaries of when they take them and how many, that would allow them to monitor usage because I get given a prescription, but nobody actually monitors my usage."

Patients also described the discussions they had with their physician about possible changes to their opioid use. Some patients felt that doctors prescribed opioids when they were not sure about or had exhausted other treatment options. For example, one patient (PID_P) reported, "I have tried to ask my doctor is there anything else I can try to see if it works better...his advice was there's not really anything else that's going to be able to help." Lack of education appeared to fuel questions later during treatment. Some patients questioned why

XSL•FO
RenderX

they were still using opioids, given that they initially expected opioids to be a relatively short-term solution for pain: "I personally thought this (opioids) might be taken 6-8 months maybe a year, we'll fix it, back to work" (PID_P). One patient said that doctors had different opinions and that the tapering plan differed to information on the internet, so they did not progress with it. Some reported being encouraged to taper in the context of recent policy changes to accessing opioids but did not always feel they had adequate support to do this. One patient (PID_A) said the following:

> *You know, and they keep saying to you try and cut down, you try and cut down so much. You're in pain, cutting down, you show that you can try and cut down.... I know I don't want to be on them, I should be off them by now if it had of gone right. I think you just feel isolated because we don't have anybody we can really talk to about our pain anymore.*

Some patients described that they had not discussed tapering with their physician as they were worried that pain may increase and were fearful of the physician's response. One patient (PID_P) said the following:

> *I was too scared to say to him "look I want to start reducing" because I just thought if I couldn't cope with the pain on a reduced dose he many not let me go back up again and then I would be stuck with pain. So that stopped me sort of officially asking for help.*

Health Professional and System Judgment, Distrust, and Stigma

This theme related to the patients' general feeling of disillusionment and distrust of mainstream medicine because of continued pain and seeking to apportion blame (eg, "I don't trust doctors. I don't trust physios. I don't trust anybody because of what they've done to me" [PID_A]). When patients described more active attempts to discuss pain relief, concerns about addiction reportedly impacted the clinician's decision to prescribe opioid medications. One patient (PID_P) stated the following:

> *Drs are more bothered about other people's addiction rather than prescribing me pain relief.... My pain was not her concern when I went and asked for strong pain killers.*

These reports of mistrust were also observed when patients discussed the impact of recent policy changes in Australia that restricted access to over-the-counter codeine medicines (February 2018) and medications for CNCP (June 2020). There was an understanding that the system and the message had changed, leaving them in a difficult position: "well the medical community's clearly been saying for a long time 'you need to take this, you need to take this', and then they say 'well no, you're not allowed to take them anymore', you're left with 'well, nothing else works, what are you going to do?'" (PID_A). For patients from addiction services, there were cases where their physician had ceased prescription, which prompted referral to an addiction specialist and engagement in an opioid replacement program. This was accompanied by feelings of abandonment toward their previous prescriber and, broadly, the medical system and ambivalence about accessing an addiction service. For some, this led to consideration of other treatment alternatives, including psychological therapy and cannabis (both prescribed and nonprescribed).

Several patients reported feeling increasingly sensitive to questioning and comments about their opioid use and chronic pain condition:

> *She would sort of say well what medications are you on? Oh, that's a high dose. oh that's been a long time. I feel like they make you feel like you're doing the wrong thing but nobody is offering any sort of solutions.* [PID_P]

Some patients reported finding this confronting and exhibited some defensiveness to the possibility of addiction or dependency. One patient (PID_A) said, "So I don't think I was addicted to it, I just took it because I got pain."

This appeared to contribute to the possible stigma associated with addiction more broadly. A patient (PID_A) explained as follows:

> *My doctor sent me there.... It kind of made me feel like I had a habit or had an addiction. That kind of really put me on the spot, and I didn't really want to go. But I went anyway. So if you put someone that's willing to cut down, but then you send them to a place like that they might feel attacked in a certain way. Like they're in the same category as those people but they're trying to do it for good reasons, not because they're addicts.*

One patient (PID_A) also described being on methadone as stigmatizing as "only heroin users use methadone." Discussions indicated that some patients felt that addiction services were not the right place for them to seek treatment and that they were *different* from other drug users. There was also a difference for people between acknowledging that they may be addicted to opioids versus the label of "addict" or "junkie." It is possible that this may contribute to an identity crisis for patients (linked to the theme of significant negative impacts linked to loss): if they are not engaging in medical treatment for a medical problem, then they either are a "junkie" or have a mental health problem ("it's all in my head".

Lack of Coordination and Integration of Care

Patients reported seeking many different types of treatments and supports for their pain. However, there was little reported integration or coordination of care between different providers, the public and private sectors, or traditional and alternative treatments. One participant shared a positive experience with a multidisciplinary rehabilitation program:

> *It was a triple pronged approach with having the psychologist, the physiotherapist and the exercise physiologist. That was really good.* [PID_P]

Another participant who was referred to the pain clinic by their GP was not receptive to an integrated approach: "the guy saying 'you'll see physio and you'll see psychologist and you'll see this and you'll see that' and I'm thinking I've tried that, been there, done that" (PID_A).

XSL•FO
RenderX

Theme 2: Limited Biopsychosocial Understanding of Pain

Focus on Physical Aspects of Pain

Most participants focused on the physical aspects of pain and minimized the potential role of psychosocial factors in exacerbating pain. One patient (PID_P) said, “Well I know where my pain is, my pain is a deterioration of the bones...so I don’t think it’s psychological.” There was also reference to experiences where one participant stated, “mine is physical pain not psychological pain” PID_P), with another participant describing pain as being perceived as separate from the mind:

> *Like if you’re in pain, how can your mind tell you that you’re not in pain? That’s how we’re looking at it. As if to say, well, you’re going to go get pain management, and they’re going to psychologically move your brain around to say that you’re not really in pain, you don’t really need these tablets.... but that still doesn’t have to do with my pain. And I try and keep the two separate because pain is one thing, and keeping my mind sane is another thing.* [PID_A]

Some participants did describe how psychosocial factors influenced their pain experience. This knowledge often emerged in hindsight or in a precontemplative state of considering treatment options other than opioids. Examples of the possible role of biopsychological factors were often inconsistent or lacked detail. For example, one patient (PID_P) said, “I think there is a big mental side to it as well.” Overall, patient connection with social influences was limited.

Disconnect Between Mental Health Problems and Pain and Engagement in Psychological Treatment

Several people described having a co-occurring mental health problem (eg, depression, anxiety, and posttraumatic stress disorder) for which they were receiving treatment from a psychiatrist or psychologist. This was also reflected in the clinical characteristics of the sample. Despite this, there was little evidence that they perceived their mental health as influencing the levels of pain experienced or vice versa.

Participants did not describe accessing psychological treatment specifically for pain or mention that psychological treatments were recommended. One participant (PID_P) reported:

> *With only recently going to the pain management sort of introduction that they told us that psychiatrist can...psychologist can sort of discuss pain with you and I just didn’t know that that had anything to do with being on pain killers....I think it’s fantastic because I think the brain is a powerful thing and I saw what cognitive behavioural therapy did with my son.*

Some patients reported positive experiences with psychological treatment for mental health problems but did not link how psychological treatments could be potentially effective for pain.

Theme 3: Continued Opioid Use When Benefits Do Not Outweigh Harms

Overview

Participants reported relatively more risks and consequences (eg, side effects and limited pain relief) versus benefits of their opioid use, but they continued to use them. Patients said the following:

> *I would reduce it down to nothing if I could...because I don’t like taking them or just the way it makes me feel.* [PID_P]
>
> *Well if I could maybe reduce some, it would be good, But I can’t see it happening.* [PID_P]

Patients also described frustration about taking opioids for pain, while at the same time, they reported holding out hope that there may be other options (eg, “it’s a catch 22...I have to think clearly if there’s a better way of doing it, than to be popping pills every day, then I’m open to anything” [PID_P]).

Opioids Mask the Pain

Some patients in this sample reported that opioids did not provide effective pain relief, but this varied from person to person; for example, “It’s not like taking painkillers is actually stopping or making the pain better” (PID_P) and “(opioids) doesn't knock all the pain over” (PID_A). There were reports that opioids “take (pain) down a few notches” (PID_P); “take the edge off” (PID_A and PID_P); or “drowns out the pain” (PID_P). Others did report that medications got rid of the pain (PID_A) and that opioids work better than other medications in the market (PID_A).

In the cases of limited pain relief, opioid medications were often not reportedly taken as prescribed (theme 4): “...even the opioids aren’t really helping with this pain. So that’s why I haven’t been taking them religiously, like I probably should have been” (PID_P). Another participant said, “it just puts me to sleep, I don’t feel anything...I don’t get the pain relief from the pill, you get pain relief from sleep” (PID_P).

No Options Other Than Opioids

Most participants reported that they had no other option but to use prescription opioids for pain. One patient (PID_P) said, “I can’t not take it, I don’t have any other option”; another patient (PID_A) said, “But with me, I’d rather not have them myself, you know, I don’t know any other way you can deal with the pain when an operation has gone bad.” For some patients, there was a view that they had no choice but to take opioids, which was related to feelings of frustration.

Escalating Dose and Need for Opioids

Participants described a history of escalating dose over time. Patients said the following: “it was two, then four, then six then eight Panadeine Forte...my body had totally adjusted to them...my body was saying, yep I’m having more and more of those” (PID_P); “That’s the first time I was put on an opioids, Endone think it was, and then started off through there, and as the pain got worse over the years the medication got higher” (PID_A). Patients often initiated dose increases with their health professional. One patient demonstrated awareness of this:

XSL•FO
RenderX

"because in the long-run I'm going to need a detox clinic, then I'm going to need someone from this clinic...psychologists and psychiatrists" (PID_P). Some patients may have been experiencing withdrawal: "sometimes I just forget my lunchtime dose and what reminds me is the severe drop in mood" (PID_P). Continued use despite negative consequences (eg, social, physical, or psychological problems) and features of tolerance and withdrawal were also reported.

Theme 4: Trial-and-Error Approach to Opioid Use

Some patients reported that they were taking opioid medicines as a "trial-and-error" approach to get relief. This appeared to be related to efforts to take control of their pain. Some people said that they took their medication when they were in pain rather than as prescribed:

> *I try not to take the opioid if I can help it. Wait to see how bad the pain gets before I take it (opioids)...the pain is unbearable. So, I get to the stage where nothing's helping.* [PID_P]
>
> *I do try to do without them and as long as I possibly can and then sometimes, the pain is so bad...then waste half a day in bed, I get pissed off at myself about it.* [PID_P]

These reports reflected patients' delayed use of opioids when in pain to minimize reliance on opioid therapy; that is, patients were hoping or taking the risk that their pain would not become more severe, so they did not have to use opioids.

Others described overuse of opioid medications:

> *Take more... Yeah. Yeah. I have four of what he gives me for that day and then there's one for the morning...My overuse of Oxycontin.* [PID_A]

Some patients reported a history of overuse: "Yes, it was taking too much Targin and before that it was too much morphine, I was actually on morphine and I was taking too much of it" (PID_A). Opioids were also used for other reasons such as sleep or improved mood. For instance, a patient (PID_P) reported the following:

> *It just puts me to sleep, I don't feel anything...I don't get the pain relief from the pill, you get pain relief from sleep.*

Patients also reported using stockpiled opioid medications:

> *I had some old medication that I had from old surgeries...so I would take them when my pain was like 9 out of 10.* [PID_P]

There were also examples of patients overusing other medicines (eg, benzodiapines and antiepileptics) in conjunction with opioids when they were in severe pain. There were examples where patients decided to self-taper their opioids, without health professional support:

> *I tried to go cold turkey initially and that was shocking...I thought I would try and reduce on my own.* [PID_P]

Theme 5: Cycles of Hopefulness and Hopelessness in Desperation to Find Pain Relief

Overview

Patient reports reflected cycles of hopelessness and hopefulness associated with pain treatment in a desperate effort to find relief. Those with increased pain or pain moving to different parts of the body over time reported a sense of hopelessness about the future: "In the last 3 years, it's gotten worse and I don't know what else is going to work" (PID_P). Desperation to find pain relief was commonly reported among participants, which was reflected in the sheer number and variety of other strategies tried and tested over time. This theme linked to the unknown cause of pain and uncertainty that prompts engagement in help-seeking behaviors to either find the source of pain or physical treatments or other strategies to relieve pain (theme 6). While patients tended to use less treatments or strategies over time due to feelings of defeat or failure, burden of treatment, and financial reasons, some seemed to be always looking for the next treatment opportunity.

Cycle Back to Previous Treatment Options

Patients described a pattern where they engaged in various pain treatments (eg, physiotherapy and chiropractic treatment) often earlier in their pain journey and that it did not help or became increasingly less effective. This was associated with feelings of hopelessness and defeat, resulting in them taking a break from treatment or trying other strategies until they were prompted to engage in physiotherapy again. One patient (PID_P) reported as follows:

> *To be told, I want you to go back and do physio and then stay on tramadol and come back in 10 months.... I have already been through all of that.*

Some patients reported significant frustration that went along with what they assessed as ineffective treatments.

Seeking New Treatments Despite Hopelessness

Participants appeared to hold feelings of hopelessness about current pain treatment and hopefulness that something still may help in the future, at the same time. Some patients were optimistic for a cure and still wanted to be "fixed." One patient (PID_P) reported, "all I'm asking for is just to get fixed and just go back to work and get my life back on track." Patients said that hope tended to decrease over time, with one participant reporting, "I was definitely more hopeful earlier on, I'm just thinking there's going to be something that's going to be sorted out" (PID_P). Despite this, some patients were open to the idea of receiving psychological treatments as well as medical or surgical interventions or alternative treatments (eg, stem cell therapy) in the *hope* that this may relieve their pain.

Theme 6: Diagnostic Uncertainty and Associated Help-Seeking Behaviors to "Fix"

This theme reflects the participants' uncertainty about the underlying causes of pain which were varied and often unexplained, contributing to ongoing attempts to understand the origins of their pain. Distress, anger, and a sense of unfairness was often tied to the triggering pain event or onset of pain, which was still apparent many years later.

XSL•FO
RenderX

For some individuals, the main explanations for the originating causes of their CNCP were seemingly unusual and involved either internalization (eg, diet and posture) or externalization (eg, weather and surgical or medical interventions going wrong) of blame. For others, there was no known cause for primary pain and patients had difficulty pinpointing the triggering event to their CNCP condition. One patient (PID_P) reported, "I couldn't tell you because it has always been there on one level, but over the course of the years it's gotten to the level where it is now." Diagnostic uncertainty contributed to ongoing attempts to understand their CNCP (cause) and find a "cure" (treatment) through more assessments and tests (eg, "...let's see what happens with this MRI" [PID_A]). Help-seeking behaviors were associated with hopeful or hopeless feelings about pain treatments and their effects (theme 5). Particularly, in cases where medical diagnoses did not explain the cause of pain, this resulted in patients being in a constant holding pattern, waiting for treatment or further tests.

Theme 7: Significant Negative Impacts Linked to Loss

Participants reported a significant number and range of negative impacts of their pain and pain management. These impacts were biological and physical (eg, dental problems); psychological and functional (eg, cognitive and memory problems, sleep disturbance, difficulties engaging in daily tasks, and low mood); or social and economic (eg, social isolation, financial burden of treatment, can't drive, and intimacy and relationship problems) in nature. One participant (PID_A) spoke about pain as a disability. Another shared that these impacts were often centered around grief and loss: "Most days I feel like I'm letting everybody down in my family because I'm not out there making money" (PID_P). The interviewer observed that participants become the most distressed in the interview when they compared their life before the pain to how it is now—and what they had lost or had to try to change or adapt. This was interpreted as a sense of loss of participating in life and normality.

For some, this sense of loss was also related to a loss or changing of identity:

> *If I come off the 10 tablets, am I going to be in more pain or am I going to be in the same pain as I am now...but have me back I'm me, have me back, just be myself.* [PID_P]

This included changes to their identity over time—to a "chronic opioid user" or "addict." Others noted a more recovery-focused interpretation of their identity with opioid use. A patient (PID_A) said the following:

> *Well, that's just part of what you did. It's not who you are. It's what you do. It doesn't have to define you. You can live a normal freaking life, with or without opioids. You've got a choice.*

Importantly, the varying ways people used prescription opioids (theme 4) was interwoven with these impacts; that is, the use or misuse of opioids was often to reduce the impacts of pain on their life.

Theme 8: Complexity of Pain and Opioid Use Journeys

Participants described complex pain journeys that involved variations in pain experience over time (got better or worse) and either a slow or fast track to opioid use. The uniqueness of the pain experience was also highlighted.

Slow Versus Fast Track to Opioid Use

The slow track to opioids often started with physiotherapy and primary care efforts to relieve pain and, in some cases, for specialists to "fact find." Over the course of their pain journey, there was a common feeling that patients were exhausted trying to investigate the physical or organic cause of pain. Because of failed treatments, patients were often initiated on opioids. One patient (PID_P) said, "cause he's the one that ... organised me to go to physiotherapy and all this, and that's when it didn't work. That's when I really started on this, using that." This pathway was common among patients from the pain clinic.

For other participants, a fast track to prescription opioids was initiated as part of acute pain management, associated with acute injury or surgery. These individuals were initiated on opioids in hospital and on discharge. A patient (PID_A) described as follows:

> *Pretty much straight away when I was in pain (in hospital).... Yes, they gave me medication to take home and then take the deep heat and that's how it started.*

Participants described either positive or negative experiences of being prescribed opioids for the first time. Some patients reported immediate pain relief or euphoria. This sense of relief was reported as sometimes occurring after previously trying multiple types of opioids. Other patients described experiencing adverse side effects and limited longer-term pain relief:

> *I knew a little bit about it but it didn't really deter me from it because, you know, not going to work, not paying bills, seems more of a downside than just taking something throughout the day that can help you continue on with your life.* [PID_A]

Everyone's Pain Experience Is Unique

Patients described the experience of pain as unique to each individual. One participant (PID_A) said the following:

> *Everyone is different... Because it's a very individual thing, I think. Handling pain, dealing with pain.*

Additional statements from the workshops provided further validation of this subtheme (eg, "well I've always had pain...pain is a personal thing" [PID_P]).

Interviewer Observations

During the patient interviews, there was evidence that some patients benefitted from the interactions with the interviewers (psychologists). This was evident even for those patients who reported that it was really challenging to describe their pain journey and they would not have been able to do this if it was years ago. There was a view that listening, asking nonjudgmental questions, and showing empathy resulted in some patients softening throughout the interview when they appeared to be more defensive at the outset:

XSL•FO
RenderX

> *You went about it the correct way. I used to get annoyed at people quite easily. So yeah, you done it fine, you ask questions nicely, you didn't talk over, you stopped when someone else is talking, which is a nice thing. So, you conducted it very well so it made it a lot easier.* [PID_A]

Alternatively, for one patient (PID_A), the interview brought attention to past pain experiences, and this was related to an increased experience of pain: "Like my backs starting to hurt just thinking about it."

There appeared to be increased interest in understanding and openness to the biopsychosocial model at the end of the interview (eg, compared with starting the discussion around the aim of the study). Some patients appeared to reflect upon and repeat the points on the biopsychosocial model throughout the interview.

Validity Checking

The process of member checking resulted in several changes to the theme structure. These included the following:

1. Tapering fears and lack of patient-physician partnering was integrated in theme 1. Furthermore, lack of coordination and integration of care among health professionals; influence of policy change judgment; and mistrust of medical profession were also included in theme 1.
2. We removed seeking certainty for cause of pain and looking within and looking outward for explanations; these were absorbed in theme 6. Furthermore, we refined the theme to focus on diagnostic uncertainty and associated help-seeking behaviors to capture these elements.
3. Theme 8, previously named multiple pain journeys and pathways to opioid use, was reconceptualized as complexity of pain and opioid use journeys. This was to capture the complexities of pain management (either fast or slow track to opioids) and also the unique experience of pain (a subtheme that emerged from workshop 1).

Discussion

Main Findings

This study is one of the first to describe the lived experience of individuals with CNCP in the context of co-designing a digital BI to increase the safe use of pain medicines. In the sample of patients interviewed, half of the patients met the current opioid misuse thresholds on validated psychometric scales (COMM [57]). Our main findings were the following 8 overarching themes that emerged in this sample related to the lived experience of patients with CNCP regarding pain management including opioid therapy: limited treatment collaboration and partnership; limited biopsychosocial understanding of pain; continued opioid use when benefits do not outweigh harms; trial-and-error approach to opioid use; cycles of hopefulness and hopelessness; diagnostic uncertainty; significant negative impacts tied to loss; and complexity of pain and the opioid use journey.

Overall, patients felt they had limited choice but to continue opioid therapy, as there were few perceived alterative treatment options available, a common concern also shared by physicians [64]. Patients' fears about making changes to their opioid use may have also played a role. This state of "stuckness" may have contributed to patients' sense of hopelessness about pain and its management. In many ways, patients may be rightly consumed by their pain in a way that they may be ambivalent about many aspects of pain management and life in general. Opioid therapy was also linked to behaviors that could elevate patient risk (eg, not taking medication as prescribed, self-tapering, and combining different opioids and classes of drugs) in a trial-and-error approach. Opioid use generally provided a pain "mask" with limited pain relief, sometimes; consistent with the results from meta-analyses [4]. However, some patients from addiction services described more positive experiences of opioid medications. Altogether, these experiences signal gaps in our current models of care and suggest that BIs may provide an avenue for (1) patients to explore thoughts and feelings about opioid use, (2) treating clinicians to support patient-physician partnering, (3) opioid use education (risks and harms), and (4) exploring patient ambivalence (risks versus benefits of status quo versus change) as a precursor to resolving the push and pull of behavioral change. Whether these techniques translates into behavioral change requires further testing. Certainly, there is evidence for the effectiveness of BIs in reducing substance-related harm [26,27] and growing evidence for their role in treating pain [65].

BIs provide another avenue to improve patient understanding of the biopsychosocial model of pain, critical to effective pain care [66]. The biopsychosocial model is broad and complex, and our explanation during the interview was brief and over the telephone. While participants may have received some education in the past, it was not obvious during the interview that they had a clear prior understanding of the influence of physical and psychosocial factors in the development and management of pain. When discussing the biopsychosocial model as a part of member checking, the use of visual tools to explain the concepts, the group peer setting, and reinforced messages may have contributed to patients' more positive responses to the model. Educating patients in the biopsychosocial model of pain is an important part of multidisciplinary rehabilitation and discipline-specific treatments that can be effective (eg, [67]). Efforts to enhance the education of health professionals in the biopsychological model could be important in contributing to improved patient understanding [68-70]. The consideration of patient health literacy and eHealth literacy [71] could help tailoring pain education to patient needs. These approaches are especially important, given that the lack of understanding of the biopsychosocial model may be related to ongoing diagnostic uncertainty and attempts to find a (biological or physical) source of pain and treatments that "fix."

Patients indicated that specific psychological treatment for pain was rarely offered, although many patients had previously engaged with psychologists or psychiatrists for other co-occurring mental health reasons. These reports are consistent with survey data from the United States indicating that 37% of patients are not aware of this treatment option and only 33% had ever worked with a pain psychologist [69]. This is despite psychological treatments being one of the recommended

XSL•FO
RenderX

evidence-based treatment options for the treatment of chronic pain and in conjunction with opioid therapy [3,8,72]. Approaches that raise awareness and provide education on the role of psychology in the treatment of pain are clearly needed not only at the patient level but also at the provider and public health systems level [73]. Most patients reported being open to this approach. Therefore, it is necessary for the psychologist workforce to meet this need for psychoeducation [13]. This will be especially important, given the psychologist workforce shortages and growing demand in the post–COVID-19 pandemic era [74].

Patient–health care professional communication and relationships had a significant impact on patients' emotional well-being and in seeking appropriate pain care. Patients felt that various aspects of opioid therapy (eg, monitoring, dose reduction, and policy change) could have been managed differently. It is acknowledged that we do not know about the patient's prescription instructions or if there was a management plan to taper them (as recommended in clinical guidelines). Our findings suggest that partnerships could be optimized by enhancing the patient–health care professional alliance and shared decision-making. Assuring patients that they will not be neglected by the treating practitioner in their medication management is essential [75], as is the need to make patients feel heard and understood. There were several examples of effective patient–health care professional communication and effective multidisciplinary management of the patients' pain and health concerns. Our study found that these examples of patient care were positively received by patients.

Questions in our interviews explored the patient's pain journey, including participation in psychological treatments as well as the pros and cons of opioid use, consistent with a motivational interviewing framework. Applying patient-identified motivational BI strategies may give patients with complex chronic conditions a space to share and feel heard. This can often be the first important step toward facilitating behavioral change [76,77]. Motivational interviewing BI techniques might also be particularly beneficial in exploring ambivalence, that is, teasing out where adverse effects and consequences of opioid use outweigh the benefits (ie, positive effects on pain or daily functioning); pain triggers; different pain management strategies; or highlighting unintended consequences to those who do not know they are at elevated risk of opioid-related harm.

Recommendations for the Design of BIs

On the basis of our research findings, we provide 6 core recommendations (not in any particular order) for co-designing and codeveloping digital BIs. Textbox 1 presents a summary of the recommendations. Future research should consider the potential barriers and enablers to implementing these recommendations during the design and development process. For example, patients may be reluctant to consider psychological therapies or to see a psychologist (recommendation 1), as their pain experience may be perceived as "all in their head." Health professionals (especially GPs) may believe that they do not have enough time to address aspects of pain management, such as the ups and downs of the pain experience and pain treatments (recommendation 2), given the demands of primary care.

XSL•FO
RenderX

Textbox 1. Summary of recommendations for the co-design and co-development of digital brief interventions.

Recommendation 1: educate about the role of psychologists and behavioral therapies in pain care

Educate patients and health professionals about the role of psychologists and mental health professionals and the evidence and clinical recommendations for use of psychological therapies in pain care. This is necessary to ensure there are no missed opportunities for patients to access psychological treatments earlier in their pain journey or as a first-line treatment.

Recommendation 2: normalize the hopefulness and hopelessness cycles in pain treatment

Normalize the cycle of pain treatment and associated feelings of hopefulness and hopelessness in the proposed brief intervention (BI) by discussing expectations and potential challenges, such as the idea of setbacks. Informing patients that they may go back to strategies that they have tried before may be helpful. The degree to which they are experiencing hope and optimism about treatment could be acknowledged and explored. Sharing what works with peers could enhance pain self-efficacy and social connectedness.

Recommendation 3: communicate with compassion and empower patients in patient–health care professional partnerships

Incorporate key principles of partnership and shared decision-making between health professionals and patients. Inviting patients to share their story could strengthen patient–health professional relationships as well as acknowledging feelings of loss and frustration. Strategies to empower patients and communicate with compassion are needed. Regular monitoring and feedback of opioid therapy by health professionals may assist patients. Digital solutions could also optimize this goal.

Recommendation 4: inform about the risks of unsafe opioid use and explore any ambivalence

Weighing the individual benefits versus harms of current opioid use may assist in exploring ambivalence about change. Educate patients about the possible harms of opioid use and contraindications. Developing insight into potentially unsafe opioid use patterns may be aided by monitoring, assessment, and feedback. Motivational interviewing may provide a guiding framework.

Recommendation 5: educate about the biopsychosocial model of pain

Educate patients about the biopsychosocial model of pain in a simple and engaging way; that is, the dynamic interaction among physiology, psychological, and social factors that can contribute to and perpetuate pain. Provide textual and visual cues and supporting materials to aid explanation. It will also be important to address ongoing patient efforts to seek a medical or physical explanation for the onset and maintenance of pain. A clear definition of psychological treatments is also required as well as consideration of marketing of the BI, particularly when patients have a history of failed pain management strategies.

Recommendation 6: acknowledge stigma and the uniqueness of pain and personalize management to different care pathways

Acknowledge stigma and personalize management to different care pathways (eg, slow or fast track to opioids and different referral or service perceptions). Patients with pain see themselves as different from other patients of addiction services. Notwithstanding, patients from pain clinics also expressed that they experience stigma related to chronic noncancer pain and opioid use. Acknowledge the uniqueness of each pain experience. Tailoring of content may also be required. Distinguishing concepts, such as addiction versus physical and psychological dependence, may also be important. Consideration of recovery-focused language may assist in exploring impacts of diagnostic labels on identity (internal or social conflicts; sick role).

Strengths and Limitations

Several strategies were used to increase the generalizability of the findings to the complex pain population. A broad sampling strategy was included to capture diversity, including age, culture, language, sex, and history of opioid therapy. Despite these attempts, culturally and linguistically diverse patients were not well-represented. Ensuring cultural diversity is incorporated as much as possible in the subsequent steps of co-design and co-development of BIs could increase customizability of the solution to fit a wider range of users. Our strategy also considered the sample size required for data saturation prospectively during data collection using a new assessment method [52]. The validity of the themes and subthemes was further enhanced through patient member checking. Future research could explore whether these core themes generalize to individuals living in a community (not actively seeking specialist treatment) in a replication study. Patients from specialist addiction and pain clinics were recruited using different strategies. This was necessary to integrate recruitment into the existing models of care, but it may have affected initial patient engagement. Patient reports may have also been influenced by their position in the treatment pathway when they were engaged in this study (eg, before multidisciplinary pain clinic care vs during the initial visit vs review visit). Our study was conducted in Australia, where patients have access to a relatively high standard of public health care and access to heavily subsidized medications. Investigations in countries with different health care systems are needed. There remains a power differential between patients, health care providers, and researchers. It is unknown how this may have influenced our findings. Future work could acknowledge and explore relationship power imbalances in the initial design steps and jointly develop guiding ethical principles. Inviting patients to be members of the research team will be important.

Conclusions

BIs have a long history in chronic relapsing conditions [24,25]. It remains to be discovered whether this treatment can be successfully applied to patients with CNCP who are at risk of opioid-related harm. However, we have made progress in effectively engaging *patient partners* in the co-design of a digital BI, the first necessary step. Exploring lived experiences as part of contextual inquiry could be considered a predesign step to co-design [78]. Our findings add to the growing literature on pre- or co-design in DHIs [79] and extend research into the CNCP field. Co-design has advantages in designing fit-for-purpose solutions that have the potential to enhance treatment effectiveness, adoption, and reach for individuals with CNCP at risk of prescription opioid–related harm.

XSL•FO
RenderX

This important first step of partnering with patients who have CNCP to understand their lived experience is integral to the co-design of new and innovative digital psychological treatments. Without this, many of the challenges identified by patients (eg, lack of trust and accessibility of treatment options) would compromise later co-design phases. The next phase in co-designing personalized digital BIs involves the generation of early design concepts and prototyping, by bringing together scientific, patient, clinical, and technical expertise to meet our current and foreseeable complex challenge of patient-centered, safe opioid therapy in CNCP care.

Acknowledgments

The authors would like to sincerely thank the patients for participating in this project and sharing their experiences. They would also like to thank the specialist pain and addiction services for collaborating on this project. The authors extend special thanks to Dr Andrew Claus, Dr Geraldine Chew, and Dr Tabibul Khan for their assistance in patient engagement. They would also like to thank Ms Chelsea Geary for her contribution to the development of the interview questions and internet-based surveys. The Metro South Research Support Scheme Early Career Researcher Grant (CIA RE; RSS_2020_035) provided funding for this project. The RECOVER Injury Research Centre, University of Queensland; Addiction and Mental Health Services, Metro South Health; and the Tess Cramond Multidisciplinary Pain and Research Centre, Metro North Health provided in-kind support. RE is supported by a Research Stimulus Fellowship grant administered by the University of Queensland.

Data Availability

The datasets generated during and analyzed during this study are not publicly available due to limitations of ethics approval but are available from the corresponding author on reasonable request.

Authors' Contributions

RE, JPC, MS, KB, FF, and SP contributed to the conception and design of the project. RE, KB, and SP collected and analyzed the data. RE drafted the manuscript, and all authors reviewed and provided edits and comments on the revised versions. All authors were involved in the interpretation of the results and have approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Patient semi-structured interview questions.
[DOCX File , 20 KB - formative_v9i1e57208_app1.docx]

Multimedia Appendix 2
Sample characteristics of patients.
[DOCX File , 21 KB - formative_v9i1e57208_app2.docx]

References

1. Chou R, Gordon DB, de Leon-Casasola OA, Rosenberg JM, Bickler S, Brennan T, et al. Management of postoperative pain: a clinical practice guideline from the American Pain Society, the American Society of Regional Anesthesia and Pain Medicine, and the American Society of Anesthesiologists' Committee on Regional Anesthesia, Executive Committee, and Administrative Council. J Pain 2016 Feb;17(2):131-157 [FREE Full text] [doi: 10.1016/j.jpain.2015.12.008] [Medline: 26827847]
2. Chou R, Turner JA, Devine EB, Hansen RN, Sullivan SD, Blazina I, et al. The effectiveness and risks of long-term opioid therapy for chronic pain: a systematic review for a National Institutes of Health Pathways to Prevention Workshop. Ann Intern Med 2015 Feb 17;162(4):276-286 [FREE Full text] [doi: 10.7326/M14-2559] [Medline: 25581257]
3. Busse JW, Craigie S, Juurlink DN, Buckley DN, Wang L, Couban RJ, et al. Guideline for opioid therapy and chronic noncancer pain. CMAJ 2017 May 08;189(18):E659-E666 [FREE Full text] [doi: 10.1503/cmaj.170363] [Medline: 28483845]
4. Busse JW, Wang L, Kamaleldin M, Craigie S, Riva JJ, Montoya L, et al. Opioids for chronic noncancer pain: a systematic review and meta-analysis. JAMA 2018 Dec 18;320(23):2448-2460 [FREE Full text] [doi: 10.1001/jama.2018.18472] [Medline: 30561481]
5. Els C, Jackson TD, Kunyk D, Lappi VG, Sonnenberg B, Hagtvedt R, et al. Adverse events associated with medium- and long-term use of opioids for chronic non-cancer pain: an overview of Cochrane Reviews. Cochrane Database Syst Rev 2017 Oct 30;10(10):CD012509 [FREE Full text] [doi: 10.1002/14651858.CD012509.pub2] [Medline: 29084357]

XSL•FO
RenderX

6. Scholl L, Seth P, Kariisa M, Wilson N, Baldwin G. Drug and opioid-involved overdose deaths — United States, 2013–2017. MMWR Morb Mortal Wkly Rep 2018 Dec 21;67(5152):1419-1427. [doi: 10.15585/mmwr.mm6751521e1] [Medline: 30605448]
7. Roxburgh A, Hall WD, Dobbins T, Gisev N, Burns L, Pearson S, et al. Trends in heroin and pharmaceutical opioid overdose deaths in Australia. Drug Alcohol Depend 2017 Oct 01;179:291-298. [doi: 10.1016/j.drugalcdep.2017.07.018] [Medline: 28826104]
8. Dowell D, Haegerich TM, Chou R. CDC guideline for prescribing opioids for chronic pain--United States, 2016. JAMA 2016 Apr 19;315(15):1624-1645 [FREE Full text] [doi: 10.1001/jama.2016.1464] [Medline: 26977696]
9. Response to the opioid crisis. United Nations Office on Drugs and Crime. URL: https://www.unodc.org/unodc/en/opioid-crisis/index.html [accessed 2023-10-08]
10. Morin KA, Eibl JK, Franklyn AM, Marsh DC. The opioid crisis: past, present and future policy climate in Ontario, Canada. Subst Abuse Treat Prev Policy 2017 Nov 02;12(1):45 [FREE Full text] [doi: 10.1186/s13011-017-0130-5] [Medline: 29096653]
11. HHS acting secretary declares public health emergency to address national opioid crisis. U.S. Department of Health & Human Services. 2017. URL: https://www.proquest.com/docview/1955885383?sourcetype=Wire%20Feeds [accessed 2023-08-12]
12. Eccleston C, Morley SJ, Williams AC. Psychological approaches to chronic pain management: evidence and challenges. Br J Anaesth 2013 Jul;111(1):59-63 [FREE Full text] [doi: 10.1093/bja/aet207] [Medline: 23794646]
13. Darnall BD, Carr DB, Schatman ME. Pain psychology and the biopsychosocial model of pain treatment: ethical imperatives and social responsibility. Pain Med 2017 Aug 01;18(8):1413-1415 [FREE Full text] [doi: 10.1093/pm/pnw166] [Medline: 27425187]
14. Hogg MN, Gibson S, Helou A, DeGabriele J, Farrell MJ. Waiting in pain: a systematic investigation into the provision of persistent pain services in Australia. Med J Aust 2012 Apr 02;196(6):386-390. [doi: 10.5694/mja12.10140] [Medline: 22471539]
15. Crouch TB, Wedin S, Kilpatrick RL, Christon L, Balliet W, Borckardt J, et al. Pain rehabilitation's dual power: treatment for chronic pain and prevention of opioid-related risks. Am Psychol 2020 Sep;75(6):825-839 [FREE Full text] [doi: 10.1037/amp0000663] [Medline: 32915026]
16. Nicholas MK, Asghari A, Sharpe L, Beeston L, Brooker C, Glare P, et al. Reducing the use of opioids by patients with chronic pain: an effectiveness study with long-term follow-up. Pain 2020 Mar;161(3):509-519. [doi: 10.1097/j.pain.0000000000001763] [Medline: 31764391]
17. Larance B, Campbell G, Moore T, Nielsen S, Bruno R, Lintzeris N, et al. Concerns and help-seeking among patients using opioids for management of chronic noncancer pain. Pain Med 2019 Apr 01;20(4):758-769. [doi: 10.1093/pm/pny078] [Medline: 29762767]
18. Frank JW, Levy C, Matlock DD, Calcaterra SL, Mueller SR, Koester S, et al. Patients' perspectives on tapering of chronic opioid therapy: a qualitative study. Pain Med 2016 Oct 20;17(10):1838-1847 [FREE Full text] [doi: 10.1093/pm/pnw078] [Medline: 27207301]
19. Goesling J, DeJonckheere M, Pierce J, Williams DA, Brummett CM, Hassett AL, et al. Opioid cessation and chronic pain: perspectives of former opioid users. Pain 2019 May 16;160(5):1131-1145 [FREE Full text] [doi: 10.1097/j.pain.0000000000001493] [Medline: 30889052]
20. Eccleston C, Fisher E, Thomas KH, Hearn L, Derry S, Stannard C, et al. Interventions for the reduction of prescribed opioid use in chronic non-cancer pain. Cochrane Database Syst Rev 2017 Nov 13;11(11):CD010323. [doi: 10.1002/14651858.CD010323.pub3] [Medline: 29130474]
21. Garland EL, Brintz CE, Hanley AW, Roseen EJ, Atchley RM, Gaylord SA, et al. Mind-body therapies for opioid-treated pain: a systematic review and meta-analysis. JAMA Intern Med 2020 Jan 01;180(1):91-105 [FREE Full text] [doi: 10.1001/jamainternmed.2019.4917] [Medline: 31682676]
22. Darnall BD. Psychological treatment for chronic pain: improving access and integration. Psychol Sci Public Interest 2021 Sep;22(2):45-51 [FREE Full text] [doi: 10.1177/15291006211033612] [Medline: 34541966]
23. Elphinston RA, Sterling M, Leung J, Gray PD, Scotti Requena S, Connor JP. Brief psychological interventions for reducing prescription opioid use, related harm, and pain intensity in patients with chronic pain: a systematic review and assessment of patient interest. Clin J Pain 2021 Apr 01;37(4):270-280. [doi: 10.1097/AJP.0000000000000908] [Medline: 33323791]
24. Brief intervention for hazardous and harmful drinking (AUDIT): a manual for use in primary care. World Health Organization. 2001 Aug 5. URL: https://tinyurl.com/mwnftddd [accessed 2023-04-05]
25. Kaner EF, Beyer FR, Muirhead C, Campbell F, Pienaar ED, Bertholet N, et al. Effectiveness of brief alcohol interventions in primary care populations. Cochrane Database Syst Rev 2018 Feb 24;2(2):CD004148 [FREE Full text] [doi: 10.1002/14651858.CD004148.pub4] [Medline: 29476653]
26. Neighbors CJ, Barnett NP, Rohsenow DJ, Colby SM, Monti PM. Cost-effectiveness of a motivational intervention for alcohol-involved youth in a hospital emergency department. J Stud Alcohol Drugs 2010 May;71(3):384-394 [FREE Full text] [doi: 10.15288/jsad.2010.71.384] [Medline: 20409432]

XSL•FO
RenderX

27. Schmidt CS, Schulte B, Seo HN, Kuhn S, O'Donnell A, Kriston L, et al. Meta-analysis on the effectiveness of alcohol screening with brief interventions for patients in emergency care settings. Addiction 2016 May;111(5):783-794. [doi: 10.1111/add.13263] [Medline: 26637990]
28. Poston JM, Hanson WE. Meta-analysis of psychological assessment as a therapeutic intervention. Psychol Assess 2010 Jun;22(2):203-212. [doi: 10.1037/a0018679] [Medline: 20528048]
29. Kaner EF, Beyer FR, Garnett C, Crane D, Brown J, Muirhead C, et al. Personalised digital interventions for reducing hazardous and harmful alcohol consumption in community-dwelling populations. Cochrane Database Syst Rev 2017 Sep 25;9(9):CD011479 [FREE Full text] [doi: 10.1002/14651858.CD011479.pub2] [Medline: 28944453]
30. Darnall BD, Ziadni MS, Krishnamurthy P, Flood P, Heathcote LC, Mackey IG, et al. "My surgical success": effect of a digital behavioral pain medicine intervention on time to opioid cessation after breast cancer surgery-a pilot randomized controlled clinical trial. Pain Med 2019 Nov 01;20(11):2228-2237 [FREE Full text] [doi: 10.1093/pm/pnz094] [Medline: 31087093]
31. Gilbody S, Littlewood E, Hewitt C, Brierley G, Tharmanathan P, Araya R, et al. Computerised cognitive behaviour therapy (cCBT) as treatment for depression in primary care (REEACT trial): large scale pragmatic randomised controlled trial. BMJ 2015 Nov 11;351:h5627 [FREE Full text] [doi: 10.1136/bmj.h5627] [Medline: 26559241]
32. Mohr DC, Weingardt KR, Reddy M, Schueller SM. Three problems with current digital mental health research . . . and three things we can do about them. Psychiatr Serv 2017 May 01;68(5):427-429 [FREE Full text] [doi: 10.1176/appi.ps.201600541] [Medline: 28412890]
33. Ioannidis JP. Why most clinical research is not useful. PLoS Med 2016 Jun;13(6):e1002049 [FREE Full text] [doi: 10.1371/journal.pmed.1002049] [Medline: 27328301]
34. Svendsen MJ, Wood KW, Kyle J, Cooper K, Rasmussen CD, Sandal LF, et al. Barriers and facilitators to patient uptake and utilisation of digital interventions for the self-management of low back pain: a systematic review of qualitative studies. BMJ Open 2020 Dec 12;10(12):e038800 [FREE Full text] [doi: 10.1136/bmjopen-2020-038800] [Medline: 33310794]
35. Fernandes LG, Devan H, Fioratti I, Kamper SJ, Williams CM, Saragiotto BT. At my own pace, space, and place: a systematic review of qualitative studies of enablers and barriers to telehealth interventions for people with chronic pain. Pain 2022 Feb 01;163(2):e165-e181. [doi: 10.1097/j.pain.0000000000002364] [Medline: 34433776]
36. Schlieter H, Marsch LA, Whitehouse D, Otto L, Londral AR, Teepe GW, et al. Scale-up of digital innovations in health care: expert commentary on enablers and barriers. J Med Internet Res 2022 Mar 11;24(3):e24582 [FREE Full text] [doi: 10.2196/24582] [Medline: 35275065]
37. Slattery P, Saeri AK, Bragge P. Research co-design in health: a rapid overview of reviews. Health Res Policy Syst 2020 Feb 11;18(1):17 [FREE Full text] [doi: 10.1186/s12961-020-0528-9] [Medline: 32046728]
38. van Gemert-Pijnen JE, Nijland N, van Limburg M, Ossebaard HC, Kelders SM, Eysenbach G, et al. A holistic framework to improve the uptake and impact of eHealth technologies. J Med Internet Res 2011 Dec 05;13(4):e111 [FREE Full text] [doi: 10.2196/jmir.1672] [Medline: 22155738]
39. Yardley L, Morrison L, Bradbury K, Muller I. The person-based approach to intervention development: application to digital health-related behavior change interventions. J Med Internet Res 2015 Jan 30;17(1):e30 [FREE Full text] [doi: 10.2196/jmir.4055] [Medline: 25639757]
40. Keogh E, Rosser BA, Eccleston C. e-Health and chronic pain management: current status and developments. Pain 2010 Oct;151(1):18-21. [doi: 10.1016/j.pain.2010.07.014] [Medline: 20674174]
41. Tabi K, Randhawa AS, Choi F, Mithani Z, Albers F, Schnieder M, et al. Mobile apps for medication management: review and analysis. JMIR Mhealth Uhealth 2019 Sep 11;7(9):e13608 [FREE Full text] [doi: 10.2196/13608] [Medline: 31512580]
42. Lalloo C, Jibb LA, Rivera J, Agarwal A, Stinson JN. "There's a pain app for that": review of patient-targeted smartphone applications for pain management. Clin J Pain 2015 Jun;31(6):557-563. [doi: 10.1097/AJP.0000000000000171] [Medline: 25370138]
43. Salazar A, de Sola H, Failde I, Moral-Munoz JA. Measuring the quality of mobile apps for the management of pain: systematic search and evaluation using the mobile app rating scale. JMIR Mhealth Uhealth 2018 Oct 25;6(10):e10718 [FREE Full text] [doi: 10.2196/10718] [Medline: 30361196]
44. Zhao P, Yoo I, Lancey R, Varghese E. Mobile applications for pain management: an app analysis for clinical usage. BMC Med Inform Decis Mak 2019 May 30;19(1):106 [FREE Full text] [doi: 10.1186/s12911-019-0827-7] [Medline: 31146739]
45. Devan H, Farmery D, Peebles L, Grainger R. Evaluation of self-management support functions in apps for people with persistent pain: systematic review. JMIR Mhealth Uhealth 2019 Feb 12;7(2):e13080 [FREE Full text] [doi: 10.2196/13080] [Medline: 30747715]
46. Noorbergen TJ, Adam MT, Teubner T, Collins CE. Using co-design in mobile health system development: a qualitative study with experts in co-design and mobile health system development. JMIR Mhealth Uhealth 2021 Nov 10;9(11):e27896 [FREE Full text] [doi: 10.2196/27896] [Medline: 34757323]
47. Sanders EB, Stappers PJ. Probes, toolkits and prototypes: three approaches to making in codesigning. CoDesign 2014 Mar 06;10(1):5-14. [doi: 10.1080/15710882.2014.888183]
48. Walker J, Holloway I, Sofaer B. In the system: the lived experience of chronic back pain from the perspectives of those seeking help from pain clinics. Pain 1999 Apr;80(3):621-628. [doi: 10.1016/S0304-3959(98)00254-1] [Medline: 10342423]

XSL•FO
RenderX

49. Liddle SD, Baxter GD, Gracey JH. Chronic low back pain: patients' experiences, opinions and expectations for clinical management. Disabil Rehabil 2007 Dec 30;29(24):1899-1909. [doi: 10.1080/09638280701189895] [Medline: 17852259]
50. Kessler RC, Andrews G, Colpe LJ, Hiripi E, Mroczek DK, Normand SL, et al. Short screening scales to monitor population prevalences and trends in non-specific psychological distress. Psychol Med 2002 Aug;32(6):959-976. [doi: 10.1017/s0033291702006074] [Medline: 12214795]
51. Bate P, Robert G. Experience-based design: from redesigning the system around the patient to co-designing services with the patient. Qual Saf Health Care 2006 Oct;15(5):307-310 [FREE Full text] [doi: 10.1136/qshc.2005.016527] [Medline: 17074863]
52. Guest G, Namey E, Chen M. A simple method to assess and report thematic saturation in qualitative research. PLoS One 2020 May 5;15(5):e0232076 [FREE Full text] [doi: 10.1371/journal.pone.0232076] [Medline: 32369511]
53. Kumar SP. Utilization of brief pain inventory as an assessment tool for pain in patients with cancer: a focused review. Indian J Palliat Care 2011 May;17(2):108-115 [FREE Full text] [doi: 10.4103/0973-1075.84531] [Medline: 21976850]
54. Tan G, Jensen MP, Thornby JI, Shanti BF. Validation of the Brief Pain Inventory for chronic nonmalignant pain. J Pain 2004 Mar;5(2):133-137 [FREE Full text] [doi: 10.1016/j.jpain.2003.12.005] [Medline: 15042521]
55. Lovibond PF, Lovibond SH. The structure of negative emotional states: comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories. Behav Res Ther 1995 Mar;33(3):335-343 [FREE Full text] [doi: 10.1016/0005-7967(94)00075-u] [Medline: 7726811]
56. Taylor R, Lovibond PF, Nicholas MK, Cayley C, Wilson PH. The utility of somatic items in the assessment of depression in patients with chronic pain: a comparison of the Zung Self-Rating Depression Scale and the Depression Anxiety Stress Scales in chronic pain and clinical and community samples. Clin J Pain 2005;21(1):91-100. [doi: 10.1097/00002508-200501000-00011] [Medline: 15599136]
57. Butler SF, Budman SH, Fernandez KC, Houle B, Benoit C, Katz N, et al. Development and validation of the current opioid misuse measure. Pain 2007 Jul;130(1-2):144-156 [FREE Full text] [doi: 10.1016/j.pain.2007.01.014] [Medline: 17493754]
58. Butler SF, Budman SH, Fanciullo GJ, Jamison RN. Cross validation of the current opioid misuse measure to monitor chronic pain patients on opioid therapy. Clin J Pain 2010;26(9):770-776 [FREE Full text] [doi: 10.1097/AJP.0b013e3181f195ba] [Medline: 20842012]
59. Meltzer EC, Rybin D, Saitz R, Samet JH, Schwartz SL, Butler SF, et al. Identifying prescription opioid use disorder in primary care: diagnostic characteristics of the Current Opioid Misuse Measure (COMM). Pain 2011 Feb;152(2):397-402 [FREE Full text] [doi: 10.1016/j.pain.2010.11.006] [Medline: 21177035]
60. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
61. Lincoln YS, Guba EG. Naturalistic Inquiry. Newbury Park, CA: SAGE Publications; 1985.
62. Charmaz K. Grounded theory. In: Smith JA, editor. Qualitative Psychology: A Practical Guide to Research Methods. Thousand Oaks, CA: SAGE Publications, Inc; 2003:81-110.
63. Prescribing drugs of dependence in general practice: part C2 - the role of opioids in pain management. The Royal Australian College of General Practitioners. URL: https://tinyurl.com/zs6ayr9y [accessed 2023-03-11]
64. Punwasi R, de Kleijn L, Rijkels-Otters JB, Veen M, Chiarotto A, Koes B. General practitioners' attitudes towards opioids for non-cancer pain: a qualitative systematic review. BMJ Open 2022 Feb 01;12(2):e054945 [FREE Full text] [doi: 10.1136/bmjopen-2021-054945] [Medline: 35105588]
65. Darnall BD, Roy A, Chen AL, Ziadni MS, Keane RT, You DS, et al. Comparison of a single-session pain management skills intervention with a single-session health education intervention and 8 sessions of cognitive behavioral therapy in adults with chronic low back pain: a randomized clinical trial. JAMA Netw Open 2021 Aug 02;4(8):e2113401 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.13401] [Medline: 34398206]
66. Gatchel RJ, Peng YB, Peters ML, Fuchs PN, Turk DC. The biopsychosocial approach to chronic pain: scientific advances and future directions. Psychol Bull 2007 Jul;133(4):581-624. [doi: 10.1037/0033-2909.133.4.581] [Medline: 17592957]
67. Siddall B, Ram A, Jones MD, Booth J, Perriman D, Summers SJ. Short-term impact of combining pain neuroscience education with exercise for chronic musculoskeletal pain: a systematic review and meta-analysis. Pain 2022 Jan 01;163(1):e20-e30. [doi: 10.1097/j.pain.0000000000002308] [Medline: 33863860]
68. Thompson K, Johnson MI, Milligan J, Briggs M. Twenty-five years of pain education research-what have we learned? Findings from a comprehensive scoping review of research into pre-registration pain education for health professionals. Pain 2018 Nov 26;159(11):2146-2158. [doi: 10.1097/j.pain.0000000000001352] [Medline: 30052574]
69. Darnall BD, Scheman J, Davin S, Burns JW, Murphy JL, Wilson AC, et al. Pain psychology: a global needs assessment and national call to action. Pain Med 2016 Feb;17(2):250-263 [FREE Full text] [doi: 10.1093/pm/pnv095] [Medline: 26803844]
70. Edmond SN, Heapy AA, Kerns RD. Engaging mental health professionals in addressing pain. JAMA Psychiatry 2019 Jun 01;76(6):565-566. [doi: 10.1001/jamapsychiatry.2019.0254] [Medline: 30942829]
71. Norman CD, Skinner HA. eHealth literacy: essential skills for consumer health in a networked world. J Med Internet Res 2006 Jun 16;8(2):e9 [FREE Full text] [doi: 10.2196/jmir.8.2.e9] [Medline: 16867972]

XSL•FO
RenderX

72. Qaseem A, Wilt TJ, McLean RM, Forciea MA, Clinical Guidelines Committee of the American College of Physicians, Denberg TD, et al. Noninvasive treatments for acute, subacute, and chronic low back pain: a clinical practice guideline from the American College of Physicians. Ann Intern Med 2017 Apr 04;166(7):514-530 [FREE Full text] [doi: 10.7326/M16-2367] [Medline: 28192789]
73. Suman A, Armijo-Olivo S, Deshpande S, Marietta-Vasquez J, Dennett L, Miciak M, et al. A systematic review of the effectiveness of mass media campaigns for the management of low back pain. Disabil Rehabil 2021 Dec;43(24):3523-3551. [doi: 10.1080/09638288.2020.1743777] [Medline: 32242464]
74. Psychological expertise where we need it most: GPs not the only answer to our mental health crisis. Australian Psychological Society. 2022 Mar 1. URL: https://tinyurl.com/bde43nvv [accessed 2022-11-01]
75. Matthias MS, Johnson NL, Shields CG, Bair MJ, MacKie P, Huffman M, et al. "I'm not gonna pull the rug out from under you": patient-provider communication about opioid tapering. J Pain 2017 Nov;18(11):1365-1373 [FREE Full text] [doi: 10.1016/j.jpain.2017.06.008] [Medline: 28690000]
76. Rollnick S, Miller WR. What is motivational interviewing? Behav Cogn Psychother 2009 Jun 16;23(4):325-334. [doi: 10.1017/S135246580001643X]
77. Rubak S, Sandbaek A, Lauritzen T, Christensen B. Motivational interviewing: a systematic review and meta-analysis. Br J Gen Pract 2005 Apr;55(513):305-312 [FREE Full text] [Medline: 15826439]
78. Bird M, McGillion M, Chambers EM, Dix J, Fajardo CJ, Gilmour M, et al. A generative co-design framework for healthcare innovation: development and application of an end-user engagement framework. Res Involv Engagem 2021 Mar 01;7(1):12 [FREE Full text] [doi: 10.1186/s40900-021-00252-7] [Medline: 33648588]
79. Vandekerckhove P, de Mul M, Bramer WM, de Bont AA. Generative participatory design methodology to develop electronic health interventions: systematic literature review. J Med Internet Res 2020 Apr 27;22(4):e13780 [FREE Full text] [doi: 10.2196/13780] [Medline: 32338617]

Abbreviations

BI: brief intervention
CNCP: chronic noncancer pain
COMM: Current Opioid Misuse Measure
DHI: digital health intervention
GP: general practitioner

Edited by A Mavragani; submitted 23.02.24; peer-reviewed by M Keller; comments to author 25.06.24; revised version received 13.08.24; accepted 24.09.24; published 30.01.25.

Please cite as:
Elphinston RA, Pager S, Brown K, Sterling M, Fatehi F, Gray P, Hipper L, Cahill L, Connor JP
Co-Designing a Digital Brief Intervention to Reduce the Risk of Prescription Opioid–Related Harm Among People With Chronic Noncancer Pain: Qualitative Analysis of Patient Lived Experiences
JMIR Form Res 2025;9:e57208
URL: https://formative.jmir.org/2025/1/e57208
doi:10.2196/57208
PMID:

©Rachel A Elphinston, Sue Pager, Kelly Brown, Michele Sterling, Farhad Fatehi, Paul Gray, Linda Hipper, Lauren Cahill, Jason P. Connor. Originally published in JMIR Formative Research (https://formative.jmir.org), 30.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Human-Centered Design of an mHealth Tool for Optimizing HIV Index Testing in Wartime Ukraine: Formative Research Case Study

Nancy Puttkammer[1], MPH, PhD; Elizabeth Dunbar[1,2], PhD; Myroslava Germanovych[3], MA; Mariia Rosol[3], MA; Matthew Golden[4,5], MPH, MD; Anna Hubashova[3], BA; Vladyslav Fedorchenko[3], MD; Larisa Hetman[3], MPA, MD; Liudmyla Legkostup[3], MS; Jan Flowers[1,6], BS, MS; Olena Nesterova[3], MD

[1]Digital Initiatives Group at I-TECH, Department of Global Health, University of Washington, Seattle, WA, United States

[2]Department of Human Centered Design and Engineering, College of Engineering, University of Washington, Seattle, WA, United States

[3]The State Institution «Public Health Center of the Ministry of Health of Ukraine», Kyiv, Ukraine

[4]Department of Medicine, Division of Allergy and Infectious Diseases, University of Washington, Seattle, WA, United States

[5]STD Control Program, Public Health Seattle King County, Seattle, WA, United States

[6]Biobehavioral Nursing and Health Informatics Department, School of Nursing, University of Washington, Seattle, WA, United States

Corresponding Author:
Nancy Puttkammer, MPH, PhD
Digital Initiatives Group at I-TECH
Department of Global Health
University of Washington
325 Ninth Ave, Box # 359932
Seattle, WA, 98104
United States
Phone: 1 206 898 5892
Email: nputt@uw.edu

Abstract

Background: Assisted partner services (APSs; sometimes called index testing) are now being brought to scale as a high-yield HIV testing strategy in many nations. However, the success of APSs is often hampered by low levels of partner elicitation. The Computer-Assisted Self-Interview (CASI)–Plus study sought to develop and test a mobile health (mHealth) tool to increase the elicitation of sexual and needle-sharing partners among persons with newly diagnosed HIV. CASI-Plus provides client-facing information on APS methods and uses a standardized, self-guided questionnaire with nonjudgmental language for clients to list partners who would benefit from HIV testing. The tool also enables health care workers (HCWs) to see summarized data to facilitate partner tracking.

Objective: The formative research phase of the CASI-Plus study aimed to gather client and HCW input on the design of the CASI-Plus tool to ensure its acceptability, feasibility, and usability.

Methods: This study gathered input to prioritize features and tested the usability of CASI-Plus with HCWs and clients receiving HIV services in public health clinics in wartime Ukraine. The CASI-Plus study's formative phase, carried out from May 2023 to July 2024, adapted human-centered design (HCD) methods grounded in principles of empathy, iteration, and creative ideation. The study involved 3 steps: *formative HCD*, including in-depth individual interviews with clients, such as men who have sex with men and people who inject drugs, and internet-based design workshops with HCWs from rural and urban HIV clinics in Chernihiv and Dnipro; *software platform assessment and heuristic evaluation*, including assessment of open-source mHealth platforms against CASI-Plus requirements, prototype development, and testing of the REDCap (Research Electronic Data Capture) prototype based on usability heuristics; and *usability walk-throughs*, including simulated cases with HCWs and clients.

Results: The formative phase of the CASI-Plus study included in-depth individual interviews with 10 clients and 3 workshops with 22 HCWs. This study demonstrated how simplified HCD methods, adapted to the wartime context, gathered rich input on prioritized features and tool design. The CASI-Plus design reflected features that are both culturally sensitive and in alignment with the constraints of Ukraine's wartime setting. Prioritized features included information about the benefits of HIV index testing; a nonjudgmental, self-guided questionnaire to report partners; client stories; and bright images to accompany the text. Two-way

XSL•FO
RenderX

SMS text messaging between clients and HCWs was deemed impractical based on risks of privacy breaches, national patient privacy regulations, and HCW workload.

Conclusions: It was feasible to conduct HCD research in Ukraine in a wartime setting. The CASI-Plus mHealth tool was acceptable to both HCWs and clients. The next step for this research is a randomized clinical trial of the effect of the REDCap-based CASI-Plus tool on the number of partners named and the rate of partners completing HIV testing.

(JMIR Form Res 2025;9:e66132) doi:10.2196/66132

KEYWORDS

human-centered design; mobile health; mHealth; Ukraine; HIV testing; war and humanitarian settings

Introduction

Background

In December 2016, the World Health Organization recommended scale-up of HIV assisted partner services (APSs) as a strategy to increase the identification and testing of partners of persons with HIV [1,2]. APSs offer persons with HIV assistance to confidentially notify their sexual and needle-sharing partners of their exposure and link the partners to testing and treatment. Randomized controlled trials have shown that APSs increase HIV testing and case finding [3-6] and are cost-effective [7,8]. APSs are scalable in routine practice, but routine APS programs have typically demonstrated lower HIV case finding than was observed in randomized controlled trials [9-12]. At least in part, the low case finding observed with APSs in most settings reflects the small number of sexual and needle-sharing partners named by index cases. As of 2020, persons with HIV receiving APSs in Ukraine named an average of only 1.14 partners despite training of APS program personnel that emphasized the need to elicit more than 1 partner for each index client [13].

Computer-Assisted Self-Interview (CASI)–Plus seeks to improve partner elicitation and testing through a mobile health (mHealth) client engagement tool using a CASI. CASI-Plus supports the initial APS encounter by providing information on how APSs work, eliciting names of all sexual or injection partners, screening for risk of intimate partner violence, and planning for partner notification. It also facilitates case management at the point of partner testing and via repeat follow-up surveys to assess for barriers to notification, interest in provider assistance, and occurrence of intimate partner violence. In surveys of sexual behavior across diverse settings, the computer-assisted self-interview approach reduced social desirability bias compared with in-person interviews [14-17]. However, the computer-assisted self-interview approach has not yet been used in APS programs in low- or middle-income countries. The CASI-Plus study sought to assess whether the tool improves partner elicitation and HIV testing in a routine APS program operating at scale, with the goal of increasing the number of persons with HIV who know their status and are linked to either HIV pre-exposure prophylaxis or HIV antiretroviral therapy services. The Russian Federation's full-scale invasion and ongoing war in Ukraine has underscored the need for simple and scalable interventions to support the control of the HIV epidemic in a disrupted health care system. CASI-Plus offers the potential to improve APS partner elicitation and HIV case finding in wartime Ukraine, as well as in many other settings where routine APS programs face challenges with high-fidelity APS implementation and low contact (partner) elicitation.

Objectives

For the CASI-Plus study, we first carried out formative research with health care workers (HCWs) and clients from a human-centered design (HCD) perspective. As applied to mHealth tools, HCD is a pragmatic research orientation that seeks to overcome barriers to usability and integration with workflows that arise when tool designers ignore the sociotechnical complexity of health care systems [18,19]. HCD principles and methods emphasize engaging with mHealth users, starting with problem definition before solution design; supporting divergence and convergence of ideas through the design process; using participatory co-design techniques, prototyping, and iteration; and applying values of creativity, empathy, tolerance for ambiguity, and willingness to fail [18,20-22]. There are few research studies or frameworks discussing the application of HCD perspectives in mHealth projects in wartime and humanitarian contexts. The participatory engagement and time required to carry out HCD-oriented research may be challenging in wartime contexts [23], and more scholarship is needed on realistic approaches to HCD in such contexts. During the CASI-Plus formative research phase, we adapted HCD methods to Ukraine's wartime context. The aims of our formative research were to (1) describe how HCW and client perspectives and the wartime context of Ukraine shaped the design of the CASI-Plus mHealth tool and (2) identify HCD methods and principles for effective formative mHealth research in a wartime setting.

Methods

Study Design

The formative phase of the CASI-Plus study (ClinicalTrials.gov NCT05826977) supported the design and software adaptation of an mHealth tool for use in Ukraine's APS program. During the first step of the formative research, we conducted internet-based workshops with HCWs from 3 health facilities, as well as in-depth interviews with clients enrolled in HIV care at 1 health facility. During the second step, we developed the CASI-Plus tool, rated its usability based on common usability heuristics or "rules of thumb," and made improvements to the prototype. During the third step, we obtained feedback on the CASI-Plus tool prototype through simulated use of the tool by health workers and clients to further adapt our design. Table 1

XSL•FO
RenderX

presents an overview of the activities across the 3 steps of our formative research.

Table 1. Summary of the Computer-Assisted Self-Interview (CASI)–Plus formative research design—participants, methods, and activities for research steps 1 to 3 in wartime Ukraine (May 2023-July 2024).

	Step 1: formative design workshops and interviews—elicit and prioritize CASI-Plus features		Step 2: heuristic evaluation—critique platforms and prototype	Step 3: simulated walk-throughs—end user prototype testing	
	Health care workers	HIV care clients	Study team	Health care workers	HIV care clients
Method	Design workshops with focus group	In-depth individual interviews	Compare tools and review CASI-Plus usability	Review of prototype and focus group	Review of prototype and individual interviews
Activities	• Preworkshop exercise • Critique ideas for CASI-Plus features • Feature prioritization game • Group synthesis and feedback	• Discuss current APS[a] barriers • Critique CASI-Plus mock-up • Review client script	• Compare 5 open-source platforms with CASI-Plus requirements from step 1 • Critique CASI-Plus REDCap[b] prototype using usability best practices	• Add 5 fictional clients to prototype and test features • Discuss feedback with the group	• 5 current HIV care clients walked through the client portion of CASI-Plus • Gather feedback

[a]APS: assisted partner service.

[b]REDCap: Research Electronic Data Capture (Vanderbilt University).

Ukrainian Setting and War

Ukraine is an Eastern European nation that regained its independence from the Soviet Union in 1991. On February 24, 2022, the Russian Federation invaded Ukraine, inflicting the largest military assault in Europe since World War II. Of a prewar population of 41.6 million, an estimated 7.5 million (18%) fled the country as refugees, 5.9 million (14.2%) were internally displaced, 1.6 to 2.8 million were forcefully deported to Russia, and >10,000 civilians were killed [24-26]. As of mid-2023, when this study was conducted, the population was estimated to be 41 million, with 12.4% being internally displaced persons and 11.6% being refugees who had returned to their homes [27]. War crimes have included direct attacks on health care facilities, resulting in damage to >1000 facilities, complete destruction of >200 facilities, and deaths of >100 HCWs (with approximately one-third killed while working) [28-31]. Our study sites included health facilities in the cities of Chernihiv and Dnipro. Chernihiv is in northern Ukraine and was occupied by Russian forces from March 2022 to April 2022, leading to widespread destruction and trauma [32]. Dnipro is a leading industrial city located in southeastern Ukraine that has served as a hub for wartime logistics and humanitarian aid for the eastern and southern regions most affected by war and occupation.

HIV in Ukraine

Ukraine has a significant HIV burden, with an estimated 245,000 persons with HIV living in the country in early 2022, of whom roughly 79% knew their HIV status [28]. In 2019, the Ukraine Ministry of Health began scaling APSs, reaching 7533 index clients and 5038 newly tested partners in 2020 [13]. Clients who are newly diagnosed with HIV, as well as clients who were previously diagnosed but who experienced interruptions in HIV treatment or other conditions leading to having an unsuppressed HIV viral load, are invited to participate in APSs on a voluntary basis. Index clients are asked to name all recent sexual and needle-sharing partners as well as biological children to indicate whether these partners have recently been tested for HIV and to select a notification method for each partner. Health workers then follow up over time with the index clients to ensure that each partner is notified and provided information about how to get tested for HIV. Persons who inject drugs and men who have sex with men have high HIV prevalence (20.9% and 3.9%, respectively) and are estimated to account for 29.9% and 2.9% of all persons with HIV in Ukraine [33-36]. However, these key populations appear underrepresented in APSs as persons who inject drugs made up only 14.9% and men who have sex with men made up only 1.8% of APS index clients in 2020 [13,37,38]. An additional barrier to epidemic control in Ukraine was that approximately one-third of named partners with unknown status (32.4%) in the APS program did not complete HIV testing within 60 days [13].

In the wartime context, HIV programs have been stretched, and conditions have favored increased HIV transmission [39]. HIV testing declined by 36% in 2022 compared to 2021 before rebounding to prewar levels in 2023 [28]. The Ukraine Ministry of Health estimated that reduced coverage of HIV testing services, together with wartime occupation and population displacement, contributed to a 13% increase in HIV incidence in 2023 compared to 2022 [28]. In the wartime context, APSs remain an expected part of the basic package of HIV services. However, their prioritization during the war given other pressing needs has not been thoroughly assessed, including in the context of our study.

HCD Methods for the Design of CASI-Plus

We sought to design CASI-Plus as a tool to increase partner elicitation and testing that would be both tailored to Ukraine's specific wartime context and generalizable to APS programs worldwide. We selected HCD methods given their utility in leveraging end-user expertise and taking a system thinking

XSL•FO
RenderX

approach while designing this health care innovation [19,22,40]. Melles et al [18] discuss the importance of system thinking regarding how an innovation interacts across 3 levels: micro (eg, using tools to perform tasks), meso (eg, health workers as a part of teams that coordinate to deliver care), and macro (eg, health workers and clients as a part of organizations and societies). With HCWs, our inquiry focused on eliciting task-based features (micro level), as well as how to integrate a tool within their health care delivery workflow (meso level) and how to tailor the design for the Ukrainian cultural context and wartime environment (macro level). For index clients, our study focused on considerations to ensure that the tool was accessible (micro), aligned with patient health care experiences (meso), and aligned with client values (macro).

Our original study, conceived before the full-scale invasion of Ukraine in 2022, proposed in-person, participatory HCD workshops and interviews with HCWs and clients led jointly by American and Ukrainian investigators. However, in the wartime context, we needed to reconsider our HCD methods based on disrupted ground transportation, safety of Public Health Center (PHC) of the Ministry of Health of Ukraine and University of Washington study team members to travel to the study sites, understaffing at the study sites, and unpredictable client demand for health care services due to mass population displacement and military mobilization. Table 1 provides an overview of HCD activities across 3 steps, whereas Table 2 provides a more detailed description of the internet-based, tele-conferencing sessions with HCWs and in-person engagements with index clients and how the HCD methods were adapted to the wartime context.

XSL•FO
RenderX

Table 2. Detailed description of human-centered design (HCD) methods used in the Computer-Assisted Self-Interview (CASI)–Plus study, including HCD principles and adaptations to Ukraine's wartime context (May 2023-July 2024).

Step, participants, and activity	Description	HCD principles	Adaptations to wartime context
Step 1—HCWs[a]			
Individual preworkshop exercise	HCWs created an ordered list of the steps of APSs[b] and described care activities, who was involved, and the data collected. HCWs were reminded of their expertise and that there were no right or wrong answers. The study team consolidated individual workflow maps to produce a consensus version for display during the workshop session.	Empowering end users as experts—presession exercises can prime users' engagement and allow them to reflect independently and without the social pressures of the workshop [41,42]; crowdsource responses from multiple participants—this helps consolidate core steps, which is critical to understand context and activities that the tool could support [18,43]	Reduced the duration of the design workshop and allowed participants to complete when they had availability (thereby reducing burden). Allowed the study team to consolidate information for review at the workshop. Included explicit discussion of how war has impacted clients and services.
Critique ideas for CASI-Plus features	Reviewed mock-ups of 12 ideas for CASI-Plus features. Each mock-up contained key discussion question prompts (eg, "what are the advantages of this feature?" and "what are your concerns?"). Invited participants to share additional feature ideas and describe ideal feature designs that would be most useful.	Ideation and creativity [20]; co-design process supported design and validation of a solution based first on diverging perspectives and then on converging perspectives	Conducted the workshop via videoconference (Zoom [Zoom Video Communications]) where participants joined individually to minimize safety risks and the inconvenience of in-person meetings. Web-based tools for ideation and creative work (eg, Figma) were deemed too complex for HCWs to learn and use, so the study team relied on slides and discussion. Visual mock-ups supported reactions and discussions and stimulated creative thinking.
Feature prioritization game	HCWs completed 3 rounds of voting on priority features, with each HCW having 3 votes per round. Between rounds of voting, HCWs discussed why they prioritized specific features. One group received 3 extra votes for the final round when they did not change choices between the first and second round.	The prioritization game helped focus the feature set for health worker top preferences. The prioritization game was adapted to voting on features rather than a money-based game of purchasing options by the study team per recommendations that voting would resonate more with participants.	Used verbal roll call for voting along with marking each vote on the screen to avoid the complexity of web-based voting tools. HCWs were engaged in exercises, discussion, and explaining their choices.
Group synthesis and feedback	Semistructured focus group discussions on how CASI-Plus would impact workflow and what barriers it would and would not address	Consider user needs and experiences [43]; system perspective	HCWs required extra time to familiarize themselves with and practice using the tool after the internet-based orientation.
Step 1—clients			
Discuss current APS barriers	Semistructured questions on barriers to naming partners, notifying partners, using APSs, and partners completing testing. Reflections on what works well and what can be improved in APSs.	Empathy [20]	IDI[c] questions asked about barriers in the wartime context.
Critique CASI-Plus features and review client script	Semistructured questions on CASI-Plus features, including opinions on 2-way texting and reflections on what barriers it would and would not address. Clients reviewed 13 cards with visual mock-ups of the client-facing screens of the CASI-Plus tool. They shared reactions, including things that were confusing or inappropriate. They provided alternative suggestions for visual design.	Co-design; creativity	A single round of design feedback was conducted to minimize burden on client and researcher time in wartime context.
Step 2—study team			

XSL•FO
RenderX

Step, participants, and activity	Description	HCD principles	Adaptations to wartime context
Heuristic evaluation to critique platforms and prototype	Evaluated 5 open-source software platforms against the CASI-Plus requirements developed by the study team and refined during step 1. Conducted heuristic evaluation (usability "rules of thumb").	Heuristic evaluations are common in HCD in which study team members compare a tool against heuristics to fix common usability issues [44,45].	Ukrainian and American study team members independently and asynchronously completed this activity and documented the results in a spreadsheet to minimize the need for synchronous web-based meetings.
Step 3—HCWs			
Usability walk-throughs	Provided "test scripts" for 5 fictional clients for HCWs to enter data and a template for documenting issues and recommendations.	Consider user needs and experiences [43].	The study team met via videoconference with HCWs to debrief their walk-throughs.
Step 3—clients			
Usability walk-throughs	A total of 5 clients participating in APSs tested CASI-Plus using a "talk aloud method." HCWs documented issues and recommendations in a structured template.	Consider user needs and experiences [43].	Clients rejected being video recorded when piloting the use of the tool.

[a]HCW: health care worker.

[b]APS: assisted partner service.

[c]IDI: in-depth interview.

Step 1: Formative HCD (May 2023-July 2023)

During step 1, with HCWs, we asked participants to (1) complete a preworkshop exercise to describe current APSs, including which personnel and what data collection or data management tools were involved; (2) discuss the current state of APSs, including barriers to partner notification and testing; (3) review 12 ideas for CASI-Plus features; and (4) prioritize features during a voting game in which participants selected top features, discussed their choices, and repeated 3 rounds of voting to allow them to change their priorities. We also carried out individual in-depth interviews with clients in which we asked them to (1) discuss current barriers to partner notification and testing and (2) critique mock-ups showing the client-facing screens and CASI-Plus tool design.

HCW workshop participants included a purposive sample of site managers, infectious disease physicians, nurses, psychologists, and social workers who deliver APSs and HIV care services. The first 2 HCW workshops included HCWs from the regional AIDS centers at hospitals in Chernihiv and Dnipro. The third workshop included HCWs from smaller rural primary care clinics that operate as subsidiary locations for the Chernihiv site. Because of safety concerns with travel of researchers to the study sites, the 3 HCW workshops were held via Zoom videoconference (Zoom Video Communications), with voluntary video and audio recording of the sessions. In total, 2 Ukrainian researchers facilitated the videoconference workshops using a facilitator's manual and slide decks. The English-speaking study team members listened in silently with synchronous translation to follow the proceedings and observe group dynamics. Before the workshops, participants were given an individual preworkshop assignment to describe APS consultation and case management activities. Between the first and second workshops, given the HCD approach, the study team iterated on potential CASI-Plus features to reflect feedback and better tailor ideas for the subsequent 2 HCW workshops.

For index clients, the study team held 30- to 45-minute in-person interviews at a single site because we could not assume that clients had access to the internet or familiarity with videoconferencing methods. We also reduced the number of client touch points during step 1 from 2 to 1 session to reduce participant burden during wartime. Clients were recruited from the Chernihiv oblast AIDS centers using a convenience sample of clients coming for routine medical visits. Inclusion criteria were (1) being aged ≥18 years and (2) having participated in APSs within the previous 12 months at the study clinic. Because we were particularly interested in the perspectives of clients from vulnerable, stigmatized groups, including men who have sex with men and persons who inject drugs, study personnel worked with HCWs to ensure that clients from these groups were referred to the study.

All workshops and client interviews were conducted in the state language, Ukrainian.

Step 2: Software Platform Assessment and Heuristic Evaluation (October 2023-May 2024)

During step 2, the study team evaluated 5 open-source software platforms against the CASI-Plus functional and technical requirements. The study team developed the requirements based on work in a preliminary study and updated this based on feedback from step 1. We then adapted the selected platform (REDCap [Research Electronic Data Capture; Vanderbilt University]) [46,47] to become the CASI-Plus tool and completed heuristic evaluation to assess usability. Each heuristic, or rule of thumb, for what makes for a pleasing and usable electronic tool was drawn from frameworks by Nielsen [44] and Kientz et al [45] using concepts such as "helps users recognize and recover from errors or relies on recognition rather than recall." We created a matrix tool for rating usability

XSL•FO
RenderX

according to each heuristic applied to each section of the CASI-Plus tool, with coding of each issue as having low, medium, or high impact on usability (Multimedia Appendix 1). Study team members then conducted walk-throughs of the tool with 5 fictional client "test scripts," rating each part of the CASI-Plus tool as passing or failing each heuristic. We collated all issues and addressed all modifiable high- and medium-impact issues through a round of improvements to the tool.

Step 3: Usability Walk-Throughs (July 2024)

Step 3 involved HCWs and clients conducting simulated walk-throughs of CASI-Plus at the Chernihiv study site using a live test version of the tool with the Android tablets procured for use of CASI-Plus at study sites. HCWs used the 5 fictional client test scripts, where they registered clients in the live version of CASI-Plus, entered client data, viewed the summary results, and edited client data as needed. After these walk-throughs, the PHC team held a focus group to gather feedback on navigation, system performance, and concerns using CASI-Plus in practice. In addition, the study coordinator completed walk-throughs with 5 actual clients, again using the live version of the tool and the tablets, followed by brief individual interviews to gather feedback, learn whether they would recommend the tool's use by others, and share suggested changes to the tool. Participants received similar incentives as in step 1.

Data Analysis

Ukrainian study team members created English-language rapid debrief reports for the data collected in step 1 using a template that followed the semistructured discussion guides to support rapid qualitative analysis [48-51]. Rapid qualitative analysis enabled actionable insights to emerge from each HCW workshop so that we could iterate on the ideas and visuals to consider input from each session. Later, full written transcripts were created and translated into English for additional analysis to reveal themes related to Ukraine's wartime and social context and inform the tool design. One analyst (NP) developed an inductive thematic codebook after an initial read-through of the transcripts and then applied the codebook to each transcript using the Dedoose software (SocioCultural Research Consultants) [52]. Data from steps 2 and 3 were tracked in spreadsheets and used in conjunction with the software requirements and design documents to guide modifications to the software prototype.

Ethical Considerations

The protocol for the CASI-Plus study was reviewed and approved by the University of Washington Human Subjects Division (00015808) and the institutional review board for the PHC (222; August 12, 2022). All participants provided written informed consent using a Ukrainian-language consent process and form. Health workers and clients were each offered an incentive of US $10 for participation in the CASI-Plus formative research. All data collected from study participants were deidentified before analysis and stored in secure, password-protected electronic storage locations.

Results

Overview

There were 22 HCWs and 10 clients participating in step 1 data collection (Table 3). All HCW participants (22/22, 100%) were female, most (13/22, 59%) were infectious disease or other physicians, and most (18/22, 82%) had provided HIV services for ≥5 years. Clients interviewed were mostly male (8/10, 80%) and included both persons who inject drugs (6/10, 60%) and men who have sex with men (4/10, 40%). Most clients interviewed (6/10, 60%) were aged 35 to 44 years.

XSL•FO
RenderX

Table 3. Study participants for step 1 of the Computer-Assisted Self-Interview–Plus formative research in wartime Ukraine (May 2023-July 2023).

Characteristic	Participants, n (%)
Health workers (n=22 individuals in 3 workshops)	
Location of work	
Oblast AIDS center	15 (68)
Peripheral health center	7 (32)
Sex	
Female	22 (100)
Male	0
Professional cadre	
Infectious disease physician	11 (50)
Nurse	7 (32)
Other physician	2 (9)
Psychologist	1 (5)
Social worker	1 (5)
Time working in HIV program (y)	
>10	8 (36)
5-10	10 (45)
2-5	4 (18)
Age group (y)	
<35	4 (18)
35-44	8 (36)
≥45	10 (45)
Clients (n=10 in-depth interviews)	
Sex	
Male	8 (80)
Female	2 (20)
Key population group	
Persons who inject drugs	6 (60)
Men who have sex with men	4 (40)
Age group (y)	
<35	2 (20)
35-44	6 (60)
≥45	2 (20)

Step 1: Formative HCD

APS Workflow

HCWs across the 3 workshops identified common, standard steps for the APS workflow, matching steps in the index client journey through initial and follow-up visits (Multimedia Appendix 2). The workflow activity resulted in a consensus on a standardized workflow that harmonized perspectives of HCWs within and across study sites. Noting that APS clients are newly engaged in care, HCWs discussed how some clients needed to build trust in the health care team before disclosing information about their partners, which underscored that CASI-Plus should be designed for longitudinal partner elicitation rather than elicitation at the initial APS enrollment visit. HCWs at one site reported eliciting information about household members and friends (in addition to HIV-exposed partners and biological children); however, the PHC researchers determined that "social network testing" was not part of the standard APS model and should not be a focus of the tool.

APS Barriers and Effects of Wartime

HCWs and clients noted barriers to APSs and challenges of wartime; Table 4 summarizes these barriers and challenges with illustrative quotes. Barriers to partner notification included fear of blame, rejection, family dissolution, and aggressive reactions

XSL•FO
RenderX

or abuse by partners, as well as clients' inability to come to terms with their HIV diagnosis. Both HCWs and clients discussed stigma and discrimination in health care by family physicians and other medical providers as being a barrier to HIV testing. Both groups also cited practical barriers to partner notification and testing, such as difficulty traveling to health facilities due to disrupted transportation networks, difficulty tracking partners due to widespread population displacement, and hesitancy to accept calls from unknown numbers given the efforts of military recruiters to enlist citizens in the armed forces. HCWs at the Chernihiv site explained that they did no HIV testing for 2 months while under Russian occupation in 2022. They estimated that HIV testing levels had rebounded to approximately 75% of the level in 2019, before the war and the COVID-19 pandemic, while noting that understaffing was an ongoing challenge in delivering APSs.

XSL•FO
RenderX

Table 4. Barriers to HIV assisted partner services (APSs) in wartime Ukraine—results from step 1 in-depth client interviews and health worker workshops (May 2023-July 2023).

Barrier	Themes	Illustrative quotes
Lack of acceptance of HIV diagnosis	1. Denial of HIV test result 2. Anger and blame of others	• "To be honest, I didn't immediately accept it, because I was on my own, I had the Internet at first, I was searching for all possible options why the test could be wrong. Well, I had a feeling that I wanted to turn this diagnosis inside out." [Client; male; man who has sex with men; aged 32 years] • "[A] person is offended that he/she was punished in this way, so they want to punish others, I believe. I had a situation like that...there was a time when I blamed everyone, I wanted to infect the whole world...[A] psychologist worked with me and everything is normal now, touch wood." [Client; female; person who injects drugs; aged 44 years]
Concern about disclosure of HIV status	1. Fear of "word getting out" about one's HIV status 2. Even an anonymous notification by an HCW[a] could allow a partner to deduce the source of the exposure	• "So, you [health worker] are calling and saying: 'Here, we have been informed...that you had sexual contact during last year.' The thing is, that we don't know how many were there, right? But he does. He knows that he was only with me." [Client; male; man who has sex with men; aged 32 years]
Fear of partners' reactions	1. Women and clients with newer partners may have more trouble notifying their partners 2. A client who was a man who has sex with men reported greater difficulty notifying bisexual partners because of fear of blame if wives and children become affected	• "[Y]ou're not, like, the same person anymore. You're HIV-positive. It's honestly stressing me out...[A] lot of people kind of say don't be afraid of this, but I'm already a second-class person." [Client; male; person who injects drugs; aged 44 years] • "[I]f, for example, he was single, yes [I could notify him]. Whatever was to happen let it happen. [But if] he is married, it's ok if everything is fine. But if not, it would be chaos in the family...I'm afraid of the reaction, so to speak." [Client; male; man who has sex with men; aged 32 years]
Clients have other priorities	1. Clients who are persons who inject drugs are generally difficult to engage due to their preoccupation with obtaining drugs or money for drugs if they are actively using drugs 2. Fatalistic wartime attitude 3. Low obligation to notify partner if they do not have an ongoing relationship	• "Due to the war, some clients have adopted a fatalistic attitude towards life and do not wish to inform their partners, making an excuse that they will live as long as they live, and what if they die tomorrow from a rocket." [HCW; female; 8 years of professional experience] • "[W]hen...I met my partner just once or twice...we met, he texted me or called me, I answered him, we met, and I don't even sign this [phone] number. Especially if I didn't like the partner, then, well, what's the point?" [Client; male; man who has sex with men; aged 32 years]
Difficulty disclosing partnerships to HCWs due to stigma	1. Shame to admit to health workers that clients have sex partners who are men who have sex with men or that they have shared needles 2. Social stigma can prevent partners from accessing HIV testing services because they do not want to be seen entering the HIV clinic for testing	• "[T]hey do not name their partners in all cases, they say 'I have no partner,' and then, after a while, we conduct conversations, we figure out that a partner of this or that client presented to us...It happens after some time, not immediately when a patient has just found out his status, rather after a while." [HCW; nurse; 8 years of professional experience] • "[T]here are some people who are coming here and try to hide the fact that they are using [drugs]...or they may say the opposite, 'I'm quitting' or something else...[T]hey will be afraid, if they have to tell the doctor, the doctor will just swear or will be negligent...like: 'Oh, they're drug addicts?...[T]hey're so...annoying,'...so they ignore it and don't talk about it." [Client; female; person who injects drugs; aged 43 years]
Ethical concerns	1. Concern about sharing partner contact details without asking them first (client) 2. Preference for routine HIV testing as part of primary health care, alleviating responsibility for notifying partners (client)	• "[P]ersonally, I don't provide phone numbers to anyone without asking the person if I can. I don't know if other people do it." [Client; male; man who has sex with men; aged 39 years]

XSL•FO
RenderX

Barrier	Themes	Illustrative quotes
Practical barriers to partner notification and testing	1. People may not pay attention to phone calls or messages attempting notification (HCWs and clients) 2. Massive displacement of the population during war creates difficulties tracing and contacting partners 3. Disrupted transportation networks make it difficult for citizens from rural and semiurban areas to access HIV testing	• "Due to the war, male clients and male partners are much more hesitant to answer calls from unknown numbers. They fear mobilization and are apprehensive about receiving calls from the military, so very often they don't pick up or don't trust strangers on the phone." [HCW; female; >10 years of professional experience]
Skeletal, overburdened HCW staff is challenged to provide APSs	1. HCWs have limited bandwidth to track partners over multiple attempts due to workload 2. Too few social workers to help physicians with APSs	• "I have 2 social workers here. One works part-time 25 per cent, which is 1 hour and 28 minutes, the other works part-time 75 per cent. A patient comes for ARVs...you need to find out about their sexual partners, you need to ask them about complaints, enter it [data] into the system of socially significant diseases, simultaneously enter it into the NHSU, into e-Health, so that we are paid for this case, write them out a prescription, refer them to tests… we are given 12 minutes per each patient...[Y]ou have to do your best, talk through everything quickly...There is such a race with each patient. And how to do all this, to ask them about all partners...they sit and try to remember, it can only be done properly in the social worker's office, if they are free..." [HCW; female; unit head; 23 years of professional service]

[a]HCW: health care worker.

Perceptions of the CASI-Plus Tool

Both HCWs and clients had an overall favorable perception of the CASI-Plus tool. HCWs and clients both felt that it would help clients discuss their partnerships without shame or embarrassment:

> *...there are patients who are sometimes ashamed to tell the doctor how many partners they have, because the person lives in a marriage and does not really want to reveal that they have someone anyone else, you now, and when it will be on the tablet, we will not look into their eyes, into their soul, as they say, perhaps the person will then simply open up. Perhaps a person will not name them, but at least they will say the number. As if it will be useful for us.* [HCW; female; >10 years of professional experience]

One client felt that it would help providers remember to discuss partner testing with each client:

> *...it's a great idea with a tablet, why, if the doctor, let's say, forgets to ask something, then that's it, I'm gone, I'll go home and think "[shoot], but I still wanted to talk about my problem, well, I won't, okay, later."* [Client; female; person who injects drugs; aged 43 years]

Of the 10 clients interviewed, all (100%) expressed willingness to use CASI-Plus on a tablet within the clinic, and only 1 (10%) expressed a definitive preference for speaking with an HCW about APSs rather than using CASI-Plus. HCWs expressed concern about the physical security and maintenance of tablet devices and made recommendations for preventing theft.

Preferred Features for the CASI-Plus Tool

Figure 1 shows the original concept for the CASI-Plus features, input from HCWs and clients about the features, and how we modified the design to address the feedback. The most important changes were (1) dropping 2-way texting with clients and (2) modifying the case management dashboard to focus on person-level information about partner disposition as reported by index clients rather than APS program performance metrics. Multimedia Appendix 2 summarizes feedback and shows how the design evolved across each step of the APS workflow, whereas Multimedia Appendix 3 shows the results of the HCW prioritization voting exercise.

During the 3 HCW workshops, HCWs prioritized features that support the initial APS encounter, including educational messaging about the rationale for APSs and how they work, as well as an initial partner elicitation questionnaire. Some HCWs and clients liked the idea of client stories that shared realistic examples of clients who pursued partner notification and testing. Clients liked the idea of CASI-Plus including 2-way texting because they felt that it would be a convenient way to communicate with their health care team. They felt that encrypted messaging (eg, via the Viber platform) would protect the privacy of communication. However, HCWs were concerned that 2-way texting would add to their workloads even if messaging was partially automated. They were also concerned about operating outside the current national legislative framework, which narrowly limits the collection and use of private information such as names and phone numbers.

There was less consensus among HCWs about CASI-Plus functionality to support APS case management. HCWs were enthusiastic about having a dashboard display of information that would help them track whether partners had completed HIV testing. However, they did not value a dashboard with performance indicators such as the total number of clients served or the average number of partners named per index client. HCWs said that they could not handle the additional data entry of case updates about contact attempts or partner testing disposition within the tool given their existing workload. They

XSL•FO
RenderX

recommended that clients capture partner follow-up data via a self-administered follow-up CASI-Plus questionnaire to be used whenever they returned to the clinic. Clients requested more health promotion information, such as messages normalizing HIV and explaining benefits of HIV pre-exposure prophylaxis and antiretroviral therapy. Design preferences included a large font for readability and a bright and hopeful visual design, with images that could help clients easily follow the text (Figure 2).

Figure 1. Evolution of the Computer-Assisted Self-Interview–Plus functional design based on step 1 formative research in wartime Ukraine (May 2023-July 2023).

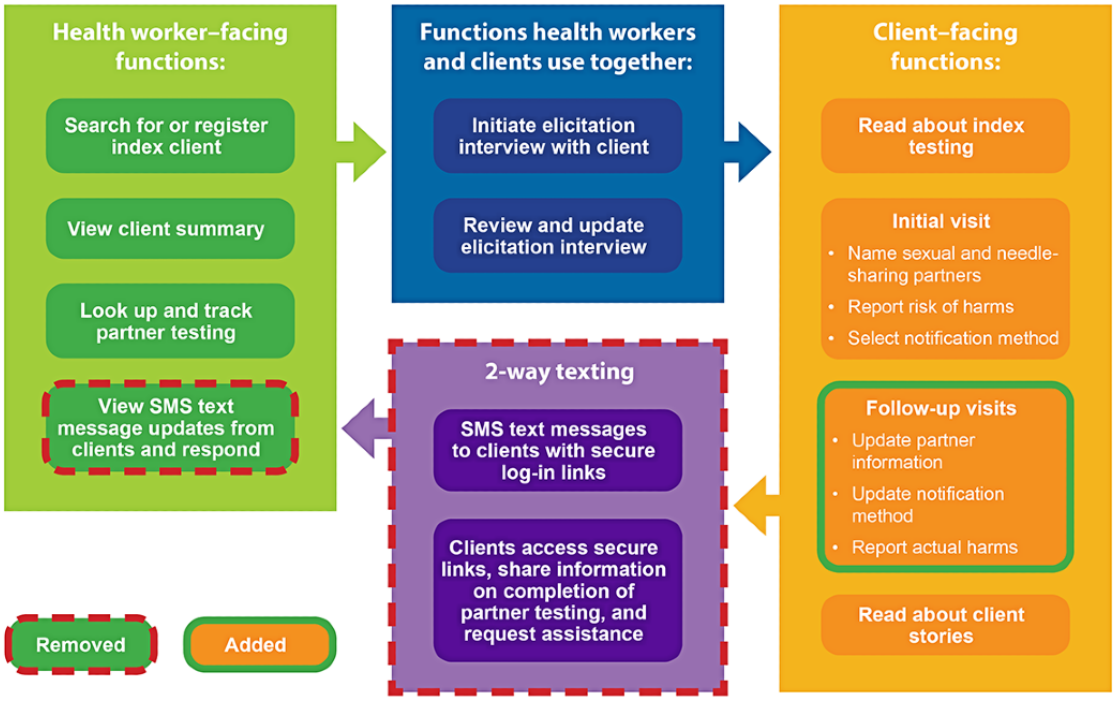

Figure 2. Evolution of the Computer-Assisted Self-Interview–Plus visual design based on step 1 formative research in wartime Ukraine (May 2023-July 2023).

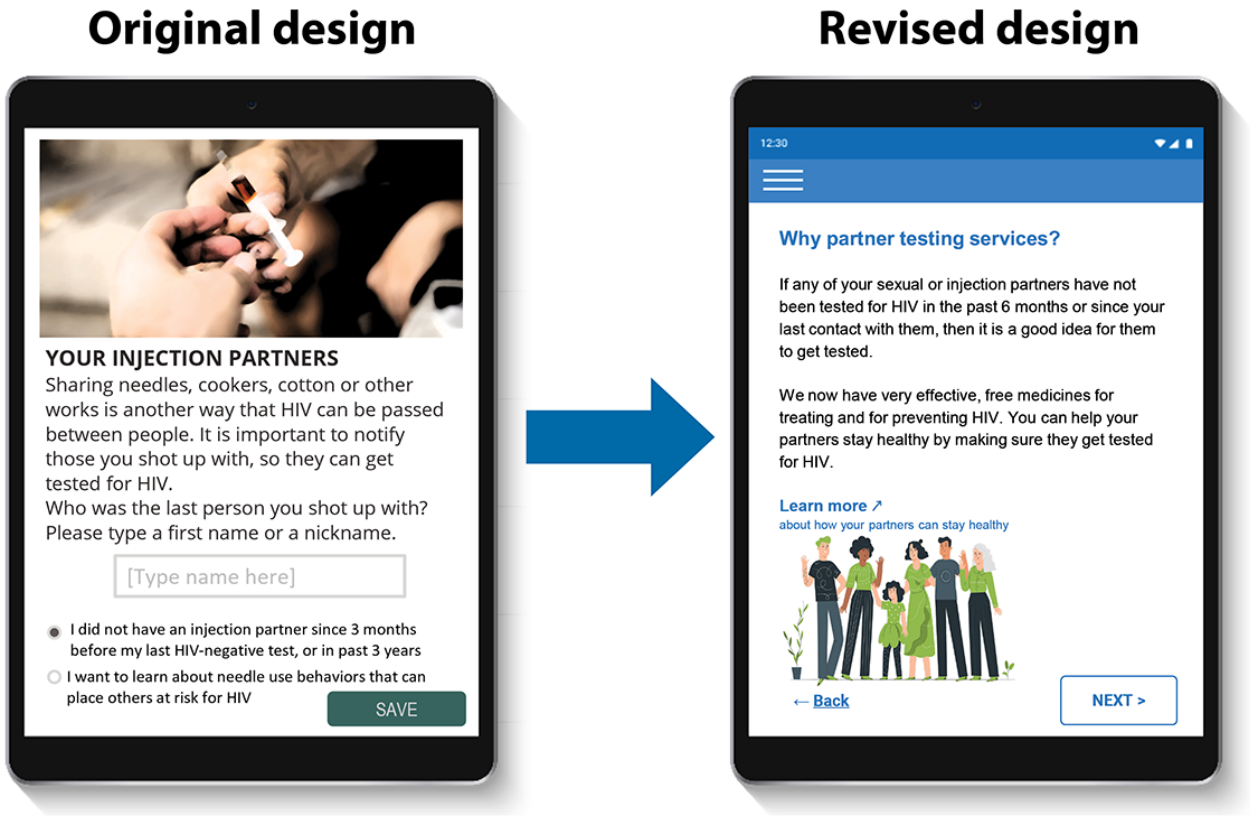

XSL•FO
RenderX

Step 2: Software Platform Assessment and Heuristic Evaluation

An evaluation of 5 widely used open-source technology platforms for digital health resulted in the initial selection and adaptation of OpenSRP (Ona Systems) [53] as the optimal platform to meet the detailed functional and technical requirements developed by the team for CASI-Plus. The other 4 platforms considered were the Community Health Toolkit [54], DHIS2 Tracker [55], OpenMRS [56,57], and REDCap [46,47,58]. However, after building CASI-Plus using OpenSRP, the study encountered potential delays in implementing the tool due to risks posed by the ongoing war. The use of technology in Ukraine, particularly in wartime, required a careful approach to ensure data security and prevent leaks, reflecting a broader state-level policy on protecting personal and medical data. In the context of Russian cyber warfare threats, hosting CASI-Plus on secure, trusted servers within Ukraine became a top priority. Therefore, the study team emphasized the specific operational needs of having a platform with an appropriate license, a Ukrainian-language interface, and reliable technical support. These requirements were critical for aligning with national policy and ensuring the project's successful execution. Consequently, the team shifted to developing the final CASI-Plus prototype in REDCap, a platform already well established in Ukraine, even though it lacked certain desirable features identified during the requirement-gathering and design stages. The REDCap tool was hosted on PHC servers and was accessible as a web application to authorized users, meaning that internet connectivity at the sites was required for its use.

A total of 5 members of the PHC research team and 3 members of the University of Washington research team completed heuristic testing of the REDCap tool. Resulting improvements to CASI-Plus usability included (1) more consistent ways to advance between screens, with messages to inform clients about progress through the questionnaire; (2) improved instructions for questionnaire items with single versus multiple response options; (3) removal of duplicate items and content; and (4) improved page loading of images (Table 5).

Table 5. Illustrative examples of findings of the usability heuristic evaluation—step 2 of the Computer-Assisted Self-Interview–Plus formative research in wartime Ukraine (October 2023-May 2024).

Heuristic	Finding (severity)	Change to software
Consistency and standards—users should not have to wonder whether different words, situations, or actions mean the same thing. Follow platform conventions.	Inconsistent messages at the end of each section of the tool before moving on to the next section (medium)	Harmonized language showing progress through the sections of the tool
Error prevention—even better than good error messages is a careful design that prevents a problem from occurring in the first place. Either eliminate error-prone conditions or check for them and present users with a confirmation option before they commit to the action.	Question flow allows users to enter a partner name while also indicating that they do not have a partner (eg, enter 2 mutually exclusive states at once; medium)	Modified question flow
Recognition rather than recall—minimize the user's memory load by making objects, actions, and options visible. The user should not have to remember information from one part of the dialogue to another. Instructions for use of the system should be visible or easily retrievable whenever appropriate.	For some questions, it is not obvious that multiple responses are allowed (low)	Added the description "Multiple answers allowed" to the questions with multiple response options
Aesthetic and minimalist design—dialogues should not contain information that is irrelevant or rarely needed. Every extra unit of information in a dialogue competes with the relevant units.	Confusing labeling and use of "show more" buttons in the section on methods of partner notification (high)	Changed style for "show more" buttons in REDCap[a]

[a]REDCap: Research Electronic Data Capture (Vanderbilt University).

Step 3: Usability Walk-Throughs

HCWs found that navigating the main menu to register and edit client records (an unmodifiable aspect of REDCap) was not intuitive but was learnable with practice and that navigating the client-facing sections of the tool was straightforward. They requested hosting separate instances of the REDCap tool for each study site to protect the confidentiality of client records across sites and prevent inappropriate editing of records by personnel from the other study sites. They recommended that future, scaled implementation of CASI-Plus would allow for selective access permissions by site, a feature built into the original OpenSRP-based tool but not available within REDCap. Overall, the HCWs were enthusiastic about testing the tool in clinical practice.

Clients also found CASI-Plus easy to use and understand. One user indicated that the tool "significantly simplifies notifying the partner; stories provide considerable support and confidence" (client; male).

There were several points of feedback. One client noted that the terms used for different partner notification methods were difficult to understand but that the explanations clarified the terminology. Another noted that it might be difficult to recall one's date of diagnosis. Client feedback did not require changes to the tool.

XSL•FO
RenderX

Discussion

Principal Findings

More interventions are needed to contain Ukraine's HIV epidemic and facilitate the effectiveness of large-scale APS programs in other low-resource settings. The CASI-Plus concept of an mHealth tool that gives persons with HIV a convenient way to learn about HIV index testing and report their partners and that helps HCWs track completion of partner testing was acceptable to both clients and HCWs in Ukraine. Prioritized features included information about the benefits of HIV index testing, a nonjudgmental self-guided questionnaire to report partners, client stories, and bright images to accompany text. Two-way SMS text messaging between clients and HCWs was not considered feasible to integrate within the workflow of understaffed public health clinics. Our final tool design and software platform choice (REDCap) reflected national policies and the principle of avoiding extra burden to clients and health workers, resulting in higher intention to engage with the tool.

Our focus was on developing a tool that would be of generalized relevance outside of Ukraine and that would also work well within the challenging conditions of Ukraine's wartime setting. This reflects the HCD design principle of "designing at the margins," in which tools that can address widespread problems are then adapted for marginalized use cases through an HCD process with the target population [59,60]. In our case, final choices for CASI-Plus features and design reflected the constraints of Ukraine's wartime setting but could be considered prudent choices for all contexts where health care resources are strained, workloads and time management are issues, and the target client population is stigmatized. The selection of the REDCap platform was driven by existing Ukraine Ministry of Health familiarity and capabilities. REDCap lacked several desired features that would have been possible with OpenSRP, such as hierarchical access permissions that would enable HCWs to view and edit the data only from their site while enabling scaled use across multiple sites and a data model supporting capture of record-level updates over time (rather than overwriting values through an editing process). Implementations of CASI-Plus in other settings might adapt the design but make a different technology choice.

This study contributes to the sparse literature on HCD methods for digital health interventions in war and humanitarian settings. The formative research phase of the CASI-Plus study demonstrated how simplified HCD methods adapted to the wartime context gathered rich input on the prioritized features and design of the CASI-Plus tool. There are few research studies or frameworks discussing HCD in mHealth projects in war or humanitarian settings. Existing published work has focused on tools for supporting the mental health of soldiers [61,62] or mHealth apps for health promotion among communities of refugees who have fled humanitarian crises [63-66]. Most existing work illustrates the role of the setting in shaping the problems and needs but not the specific ways in which the researchers adapted or shaped HCD methods based on a crisis context. For example, Beeman et al [65] described HCD methods applied to the design of menstrual health spaces for women living in a refugee settlement in Uganda. Their HCD methods involved standard steps of design research, rough prototyping, live prototyping, and piloting [65]. The resulting intervention design addressed the lack of privacy, tenuous access to disposable sanitary products, and other challenges facing women in the humanitarian setting. Another study of an mHealth app for Syrian refugees in Lebanon, a structured electronic health record with both client-facing features designed to improve health literacy and provider-facing features designed to support clinical decision-making, represents a cautionary tale in which design took place without participatory end-user input. The resulting tool failed to ease the workload of HCWs, and uptake was limited to approximately 20% of clinical consultations [67]. In another study, Bartlett et al [64] reported on an evaluation of an HCD process for a health care app for refugees. The evaluation findings noted power differentials in the design process and the need for diverse sampling techniques for client engagement so that design does not favor the perspectives of those with greater privilege [64].

Our formative study is unique in reporting on how we adapted mHealth HCD methods to a wartime context where the health system is under tremendous stress. Our study demonstrated several principles for successful HCD in a wartime context. First, we approached data collection so that it would be safe and feasible for both participants and study team members, all of whom were living through challenges of war, including displacement from their own homes; disrupted power supplies, including in winter; interrupted access to their offices, computers, and files; need to take shelter from bombs at a moment's notice; and having family members and friends risking their lives on the front lines. Drawing on the HCD principle of empathy, we relied on in-country investigators to guide the timing and design of participatory formative research activities, first brainstorming possible activities and then consulting with HCWs at the sites to confirm whether the activities would be feasible, hear their concerns, and modify plans. Second, we created a detailed facilitator guide for the HCD workshop and added guidance to workshop slides to support facilitators in real time in guiding the conversation toward the research aims. Third, we scheduled activities during the day (eg, in the afternoons before daily facility curfew closure times) and during weeks when wartime conditions (eg, fewer bombings and power outages) permitted.

Adoption of mHealth tools is always complex from a system perspective, affecting tasks by individuals, how teams work together, and how teams function within broader systems (the micro, meso, and macro levels discussed by Melles et al [18]). The complexity of absorbing system changes during war is amplified. Ukraine is struggling to maintain its health care delivery system in the face of direct attacks on health care infrastructure, departure of many health workers, and surging public health disease burden. Priorities for the health care system include carrying through with reforms started before the war (eg, decentralization of health care management and enabling of private providers to receive state funding to deliver approved HIV service packages) and adaptations resulting from the war (eg, modifications to the electronic data system for the HIV program [called the Information System for Socially Significant

XSL•FO
RenderX

Diseases] to enable remote use and tracking of patients receiving HIV care outside of Ukrainian borders). However, many adaptive responses in health care are still not formally recognized by the government, such as providers communicating with patients via messaging apps such as Viber or Telegram. In terms of CASI-Plus, our final design and prototype adapted to the existing APS workflows and formal regulations rather than introducing large changes that could be difficult to implement or sustain within understaffed public-sector clinics during wartime.

Future Priorities and Research

Future priorities for CASI-Plus technology development include features to reduce workloads for the already overworked HCWs, such as integrations with other existing routine data systems to avoid double data entry, and improved tracking of named partners. Future efforts are warranted to explore the use of 2-way SMS text messaging at a time when Ukrainian health policy makers can codify formal regulations for this type of technology. The next step in research on CASI-Plus for Ukraine's APS program is a randomized controlled trial of the effect of the REDCap-based CASI-Plus on the number of partners named and tested for HIV. CASI-Plus has the potential to complement other strategies being used in Ukraine to support HIV case finding in the wartime context. Previous studies have identified that APSs applied to social networks, particularly for persons who inject drugs, are effective for HIV case finding [33]. A recent study combined a social network approach to APSs with a machine learning algorithm applied to screening data from index clients to identify the networks most likely to yield new HIV cases [68]. Future research could study the use of CASI-Plus as part of social network testing and in conjunction with a prediction algorithm, as well as cultural adaptation of CASI-Plus in other countries.

Strengths and Limitations

Strengths of this study were the adaptation of HCD methods to Ukraine's wartime context, the inclusion of perspectives from 2 regions and both larger and smaller HIV service sites, and openness to HCW and client perspectives about the direction for CASI-Plus. Limitations of this study included the inability of HCWs and research team members to meet in person for creative co-design sessions; time pressure to complete the HCW workshops, where several participants had to drop out early when their worksites closed; and low engagement from rural, peripheral sites due to unfamiliarity with Zoom videoconferencing technology. Favorable opinions expressed about the CASI-Plus concept may reflect social desirability bias or a tendency to conform to the priorities of external funders because of the resources they provide [69]. Engaging participants from only 3 out of hundreds of primary health care sites may also have limited feature elicitation and utility across the spectrum of service delivery, particularly given the nuances in wartime settings. A limitation of pursuing CASI-Plus technology design during wartime was that national policies on technology platforms and patient electronic communications were less flexible than we anticipated when we conceptualized the study before the war, meaning that we were not able to implement 2-way SMS text messaging as part of the tool.

Conclusion

In conclusion, the CASI-Plus mHealth tool is an acceptable intervention to support HIV testing targeted at people most at risk of HIV exposure, namely, the sexual and needle-sharing partners of people who have been newly diagnosed with HIV. It was possible to adapt HCD methods grounded in principles of empathy, iteration, and creative ideation to Ukraine's wartime context using internet-based workshops with HCWs and individual interviews with clients. This work demonstrates the flexibility and value of HCD methods for adapting digital health tools in a wartime context.

Acknowledgments

The authors are grateful to Oleksandr Kurdus, head of the Information Technology Department, Public Health Center of the Ministry of Health of Ukraine, for assistance with implementation of the Computer-Assisted Self-Interview (CASI)–Plus tool in Ukraine and for review and input on this manuscript. They acknowledge Abert Namaya of the Digital Initiatives Group at the International Training and Education Center for Health at the University of Washington for contributions to the CASI-Plus technology assessment. This study would not have been possible without the support of all the health care workers and clients who chose to participate in the CASI-Plus formative study despite the disruptions and trauma of the war in Ukraine. The research reported in this publication was supported by the National Institute of Biomedical Imaging and Bioengineering of the National Institutes of Health under award R21EB032229. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Data Availability

The datasets generated during and/or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1

XSL•FO
RenderX

Usability heuristic template for step 2 Computer-Assisted Self-Interview–Plus heuristic evaluation in wartime Ukraine (October 2023-May 2024).
[DOCX File , 50 KB - formative_v9i1e66132_app1.docx]

Multimedia Appendix 2
Feedback on the Computer-Assisted Self-Interview–Plus proposed features and resulting design modifications based on the step 1 formative research in wartime Ukraine (May 2023-July 2023).
[DOC File , 49 KB - formative_v9i1e66132_app2.doc]

Multimedia Appendix 3
Prioritization of Computer-Assisted Self-Interview–Plus features by health care workers during step 1 formative workshops in wartime Ukraine (May 2023-July 2023).
[DOCX File , 19 KB - formative_v9i1e66132_app3.docx]

References

1. Guidelines on HIV self-testing and partner notification: supplement to consolidated guidelines on HIV testing services. World Health Organization. 2016. URL: https://iris.who.int/handle/10665/251655 [accessed 2024-12-09]
2. Consolidated guidelines on HIV testing services for a changing epidemic. World Health Organization. 2019 Nov 27. URL: https://www.who.int/publications/i/item/WHO-CDS-HIV-19.31 [accessed 2024-12-09]
3. Cherutich P, Golden MR, Wamuti B, Richardson BA, Ásbjörnsdóttir KH, Otieno FA, et al. Assisted partner services for HIV in Kenya: a cluster randomised controlled trial. The Lancet HIV 2017 Feb;4(2):e74-e82. [doi: 10.1016/s2352-3018(16)30214-4]
4. Rosenberg NE, Mtande TK, Saidi F, Stanley C, Jere E, Paile L, et al. Recruiting male partners for couple HIV testing and counselling in Malawi's option B+ programme: an unblinded randomised controlled trial. Lancet HIV 2015 Nov;2(11):e483-e491 [FREE Full text] [doi: 10.1016/S2352-3018(15)00182-4] [Medline: 26520928]
5. Brown LB, Miller WC, Kamanga G, Nyirenda N, Mmodzi P, Pettifor A, et al. HIV partner notification is effective and feasible in sub-Saharan Africa: opportunities for HIV treatment and prevention. J Acquir Immune Defic Syndr 2011 Apr 15;56(5):437-442 [FREE Full text] [doi: 10.1097/qai.0b013e318202bf7d] [Medline: 22046601]
6. Landis SE, Schoenbach VJ, Weber DJ, Mittal M, Krishan B, Lewis K, et al. Results of a randomized trial of partner notification in cases of HIV infection in North Carolina. N Engl J Med 1992 Jan 09;326(2):101-106. [doi: 10.1056/NEJM199201093260205] [Medline: 1445500]
7. Rutstein SE, Brown LB, Biddle AK, Wheeler SB, Kamanga G, Mmodzi P, et al. Cost-effectiveness of provider-based HIV partner notification in urban Malawi. Health Policy Plan 2014 Jan 15;29(1):115-126 [FREE Full text] [doi: 10.1093/heapol/czs140] [Medline: 23325584]
8. Sharma M, Smith JA, Farquhar C, Ying R, Cherutich P, Golden M, et al. Assisted partner notification services are cost-effective for decreasing HIV burden in western Kenya. AIDS 2018 Jan 14;32(2):233-241 [FREE Full text] [doi: 10.1097/QAD.0000000000001697] [Medline: 29135576]
9. Myers RS, Feldacker C, Cesár F, Paredes Z, Augusto G, Muluana C, et al. Acceptability and effectiveness of assisted human immunodeficiency virus partner services in Mozambique: results from a pilot program in a public, urban clinic. Sexual Trans Dis 2016;43(11):690-695. [doi: 10.1097/olq.0000000000000529]
10. Henley C, Forgwei G, Welty T, Golden M, Adimora A, Shields R, et al. Scale-up and case-finding effectiveness of an HIV partner services program in Cameroon: an innovative HIV prevention intervention for developing countries. Sex Transm Dis 2013 Dec;40(12):909-914 [FREE Full text] [doi: 10.1097/OLQ.0000000000000032] [Medline: 24220349]
11. Shamu S, Farirai T, Kuwanda L, Slabbert J, Guloba G, Khupakonke S, et al. Comparison of community-based HIV counselling and testing (CBCT) through index client tracing and other modalities: outcomes in 13 South African high HIV prevalence districts by gender and age. PLoS One 2019 Sep 6;14(9):e0221215 [FREE Full text] [doi: 10.1371/journal.pone.0221215] [Medline: 31490938]
12. Katbi M, Adegboye A, Adedoyin A, Yunusa F, Kayode G, Bello M, et al. Effect of clients Strategic Index Case Testing on community-based detection of HIV infections (STRICT study). Int J Infect Dis 2018 Sep;74:54-60 [FREE Full text] [doi: 10.1016/j.ijid.2018.06.018] [Medline: 30100537]
13. Ihnatiuk AP, Shapoval AY, Kazanzhy AP, Kuzin IV, Riabokon SV, Shotorbani S, et al. Distinctive model for HIV index testing (IT) in Eastern Europe: results of Ukraine's physician-led, integrated IT programme. BMJ Open 2023 Sep 20;13(9):e070542 [FREE Full text] [doi: 10.1136/bmjopen-2022-070542] [Medline: 37730414]
14. Adebajo S, Obianwu O, Eluwa G, Vu L, Oginni A, Tun W, et al. Comparison of audio computer assisted self-interview and face-to-face interview methods in eliciting HIV-related risks among men who have sex with men and men who inject drugs in Nigeria. PLoS One 2014 Jan 8;9(1):e81981 [FREE Full text] [doi: 10.1371/journal.pone.0081981] [Medline: 24416134]

XSL•FO
RenderX

15. Beauclair R, Meng F, Deprez N, Temmerman M, Welte A, Hens N, et al. Evaluating audio computer assisted self-interviews in urban South African communities: evidence for good suitability and reduced social desirability bias of a cross-sectional survey on sexual behaviour. BMC Med Res Methodol 2013 Jan 31;13(1):11 [FREE Full text] [doi: 10.1186/1471-2288-13-11] [Medline: 23368888]
16. NIMH Collaborative HIV/STD Prevention Trial Group. The feasibility of audio computer-assisted self-interviewing in international settings. AIDS 2007 Apr;21 Suppl 2(Suppl 2):S49-S58. [doi: 10.1097/01.aids.0000266457.11020.f0] [Medline: 17413264]
17. John SA, Petroll AE, Walsh JL, Quinn KG, Patel VV, Grov C. Reducing the discussion divide by digital questionnaires in health care settings: disruptive innovation for HIV testing and PrEP screening. J Acquir Immune Defic Syndr 2020 Nov 01;85(3):302-308 [FREE Full text] [doi: 10.1097/QAI.0000000000002459] [Medline: 32701822]
18. Melles M, Albayrak A, Goossens R. Innovating health care: key characteristics of human-centered design. Int J Qual Health Care 2021 Jan 12;33(Supplement_1):37-44 [FREE Full text] [doi: 10.1093/intqhc/mzaa127] [Medline: 33068104]
19. Evans L, Evans J, Pagliari C, Källander K. Scoping review: exploring the equity impact of current digital health design practices. Oxford Open Digit Health 2023;1:oqad006. [doi: 10.1093/oodh/oqad006]
20. Beres LK, Simbeza S, Holmes CB, Mwamba C, Mukamba N, Sharma A, et al. Human-centered design lessons for implementation science: improving the implementation of a patient-centered care intervention. J Acquir Immune Defic Syndr 2019 Dec;82 Suppl 3(3):S230-S243 [FREE Full text] [doi: 10.1097/QAI.0000000000002216] [Medline: 31764259]
21. Bazzano AN, Martin J, Hicks E, Faughnan M, Murphy L. Human-centred design in global health: a scoping review of applications and contexts. PLoS One 2017;12(11):e0186744 [FREE Full text] [doi: 10.1371/journal.pone.0186744] [Medline: 29091935]
22. Holeman I, Kane D. Human-centered design for global health equity. Inf Technol Dev 2019 Sep 29;26(3):477-505 [FREE Full text] [doi: 10.1080/02681102.2019.1667289] [Medline: 32982007]
23. Cheney C. How organizations can apply human-centered design in crisis response. Devex. 2022 Apr 8. URL: https://www.devex.com/news/how-organizations-can-apply-human-centered-design-in-crisis-response-102778 [accessed 2024-07-06]
24. Ukraine refugee situation. United Nations High Commissioner for Refugees Operational Data Portal. 2024. URL: https://data.unhcr.org/en/situations/ukraine [accessed 2024-01-08]
25. From now on, Ukrainians forcibly deported to the Russian Federation can issue a certificate for return to Ukraine. State Migration Service, Ministry of Internal Affairs of Ukraine. 2022. URL: https://tinyurl.com/yhk644sa [accessed 2024-12-09]
26. Janowski K. Civilian deaths in Ukraine war top 10,000, UN says. United Nations Ukraine. 2023 Nov 21. URL: https://tinyurl.com/y54atac8 [accessed 2024-07-06]
27. Ukraine — snapshot report: population figures and geographic distribution — general population survey round 13 (11 - 23 May). International Organization for Migration. 2023. URL: https://tinyurl.com/4zmcrjhn [accessed 2024-12-03]
28. National response of HIV, TB, viral hepatitis and SMT programmes in the context of full-scale Russian invasion. Public Health Center of the MOH of Ukraine. 2023. URL: https://phc.org.ua/sites/default/files/users/user90/National_response_HIV_TB_VH_SMT_war_2023_ENG.pdf [accessed 2024-10-10]
29. About registered criminal offenses and the results of their pre-trial investigation. Office of the Prosecutor General, Government of Ukraine. 2022. URL: https://tinyurl.com/487f3as5 [accessed 2024-01-25]
30. Levy BS, Leaning J. Russia's war in Ukraine - the devastation of health and human rights. N Engl J Med 2022 Jul 14;387(2):102-105. [doi: 10.1056/NEJMp2207415] [Medline: 35767514]
31. Surveillance system for attacks on healthcare (SSA). World Health Organization. URL: https://tinyurl.com/2ms8z99w [accessed 2024-11-19]
32. McCarthy S, George S, Haq SN, Macaya M, Hayes M, Chowdhury M, et al. April 3, 2022 Russia-Ukraine news. CNN World. 2022 Apr 4. URL: https://www.cnn.com/europe/live-news/ukraine-russia-putin-news-04-3-22/index.html [accessed 2024-10-10]
33. Williams LD, Korobchuk A, Smyrnov P, Sazonova Y, Nikolopoulos GK, Skaathun B, et al. Social network approaches to locating people recently infected with HIV in Odessa, Ukraine. J Int AIDS Soc 2019 Jun 27;22(6):e25330 [FREE Full text] [doi: 10.1002/jia2.25330] [Medline: 31245917]
34. Booth RE, Davis JM, Dvoryak S, Brewster JT, Lisovska O, Strathdee SA, et al. HIV incidence among people who inject drugs (PWIDs) in Ukraine: results from a clustered randomised trial. The Lancet HIV 2016 Oct;3(10):e482-e489. [doi: 10.1016/s2352-3018(16)30040-6]
35. The key population atlas. Joint United Nations Programme on HIV/AIDS. URL: https://kpatlas.unaids.org/dashboard [accessed 2024-10-10]
36. Puttkammer N, Ihnatiuk A, Shapoval A, Kazanzhy A, Secor A, Shotorbani S, et al. Profile of partners who completed HIV testing and received a new HIV diagnosis in Ukraine's HIV index testing program: a retrospective cohort study to inform program improvement. BMC Infect Dis 2023 May 05;23(1):291 [FREE Full text] [doi: 10.1186/s12879-023-08281-1] [Medline: 37147618]
37. Dumchev K, Kornilova M, Kulchynska R, Azarskova M, Vitek C. Improved ascertainment of modes of HIV transmission in Ukraine indicates importance of drug injecting and homosexual risk. BMC Public Health 2020 Aug 26;20(1):1288 [FREE Full text] [doi: 10.1186/s12889-020-09373-2] [Medline: 32843008]

XSL•FO
RenderX

38. Dumchev K, Sazonova Y, Salyuk T, Varetska O. Trends in HIV prevalence among people injecting drugs, men having sex with men, and female sex workers in Ukraine. Int J STD AIDS 2018 Nov 27;29(13):1337-1344. [doi: 10.1177/0956462418784096] [Medline: 30049259]
39. Friedman SR, Smyrnov P, Vasylyeva TI. Will the Russian war in Ukraine unleash larger epidemics of HIV, TB and associated conditions and diseases in Ukraine? Harm Reduct J 2023 Sep 01;20(1):119 [FREE Full text] [doi: 10.1186/s12954-023-00855-1] [Medline: 37658448]
40. LaFond A, Cherney M. A theory of change for guiding the integration of human-centered design into global health programming. Glob Health Sci Pract 2021 Nov 29;9(Supplement 2):S209-S216. [doi: 10.9745/ghsp-d-21-00334]
41. Jewell C, Salvetti F. Towards a combined method of web usability testing: an assessment of the complementary advantages of lab testing, pre-session assignments, and online usability services. In: Proceedings of the CHI '12 Extended Abstracts on Human Factors in Computing Systems. 2012 Presented at: CHI EA '12; May 5-10, 2012; Austin, TX. [doi: 10.1145/2212776.2223720]
42. Jewell C, Salvetti S. Using pre-session homework to elicit more insights during web usability testing. In: Proceedings of the HCI International 2011 Posters' Extended Abstracts. 2011 Presented at: HCI International 2011; July 9-14, 2011; Orlando, FL. [doi: 10.1007/978-3-642-22098-2_32]
43. Dopp AR, Parisi KE, Munson SA, Lyon AR. Aligning implementation and user-centered design strategies to enhance the impact of health services: results from a concept mapping study. Implement Sci Commun 2020 Feb 26;1(1):17 [FREE Full text] [doi: 10.1186/s43058-020-00020-w] [Medline: 32885179]
44. Nielsen J. 10 usability heuristics for user interface design. Nielsen Norman Group. 1994 Apr 24. URL: https://www.nngroup.com/articles/ten-usability-heuristics/ [accessed 2024-07-06]
45. Kientz JA, Choe EK, Birch B, Maharaj R, Fonville A, Glasson C, et al. Heuristic evaluation of persuasive health technologies. In: Proceedings of the 1st ACM International Health Informatics Symposium. 2010 Presented at: IHI '10; November 11-12, 2010; Arlington, VA. [doi: 10.1145/1882992.1883084]
46. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381 [FREE Full text] [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
47. Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O'Neal L, et al. The REDCap consortium: building an international community of software platform partners. J Biomed Inform 2019 Jul;95:103208 [FREE Full text] [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]
48. Vindrola-Padros C, Johnson GA. Rapid techniques in qualitative research: a critical review of the literature. Qual Health Res 2020 Aug 15;30(10):1596-1604. [doi: 10.1177/1049732320921835] [Medline: 32667277]
49. Taylor B, Henshall C, Kenyon S, Litchfield I, Greenfield S. Can rapid approaches to qualitative analysis deliver timely, valid findings to clinical leaders? A mixed methods study comparing rapid and thematic analysis. BMJ Open 2018 Oct 08;8(10):e019993 [FREE Full text] [doi: 10.1136/bmjopen-2017-019993] [Medline: 30297341]
50. Gale RC, Wu J, Erhardt T, Bounthavong M, Reardon CM, Damschroder LJ, et al. Comparison of rapid vs in-depth qualitative analytic methods from a process evaluation of academic detailing in the Veterans Health Administration. Implement Sci 2019 Feb 01;14(1):11 [FREE Full text] [doi: 10.1186/s13012-019-0853-y] [Medline: 30709368]
51. Simoni JM, Beima-Sofie K, Amico KR, Hosek SG, Johnson MO, Mensch BS. Debrief reports to expedite the impact of qualitative research: do they accurately capture data from in-depth interviews? AIDS Behav 2019 Aug 21;23(8):2185-2189 [FREE Full text] [doi: 10.1007/s10461-018-02387-3] [Medline: 30666522]
52. Dedoose hoempage. Dedoose. URL: https://www.dedoose.com/ [accessed 2024-04-17]
53. Open smart register platform homepage. OpenSRP. URL: https://opensrp.io/ [accessed 2024-10-10]
54. Tools - community health toolkit. Community Health Toolkit. URL: https://communityhealthtoolkit.org/tools [accessed 2024-10-10]
55. Individual data records with tracker. DHIS2. URL: https://dhis2.org/tracker/ [accessed 2024-10-10]
56. OpenMRS hoempage. OpenMRS. URL: https://openmrs.org/ [accessed 2024-10-10]
57. Mamlin BW, Biondich PG, Wolfe BA, Fraser H, Jazayeri D, Allen C, et al. Cooking up an open source EMR for developing countries: OpenMRS - a recipe for successful collaboration. AMIA Annu Symp Proc 2006;2006:529-533 [FREE Full text] [Medline: 17238397]
58. Research Electronic Data Capture (REDCap). Institute of Translational Health Sciences. URL: https://www.iths.org/investigators/services/bmi/redcap/ [accessed 2021-01-18]
59. Designing at the margins. Getting Smart. URL: https://www.gettingsmart.com/2023/10/05/designing-at-the-margins/ [accessed 2024-10-10]
60. Rigot A. Design from the margins. Belfer Center for Science and International Affairs. 2022. URL: https://tinyurl.com/3t9xk37s [accessed 2024-12-09]
61. Rizzo A, Pair J, McNerney PJ, Eastlund E, Manson B, Gratch J, et al. Development of a VR therapy application for Iraq war military personnel with PTSD. Stud Health Technol Inform 2005;111:407-413. [Medline: 15718769]
62. Mozgai SA, Femminella B, Hartholt A, Rizzo A. User-centered design model for mobile health (mHealth) applications: a military case study in rapid assessment process (RAP). In: Proceedings of the Extended Abstracts of the 2021 CHI Conference

XSL•FO
RenderX

on Human Factors in Computing Systems. 2021 Presented at: CHI EA '21; May 8-13, 2021; Yokohama, Japan. [doi: 10.1145/3411763.3443450]
63. Sikder A, Dickhoner J, Kysh L, Musheghyan L, Shekherdimian S, Levine B, et al. International collaboration to develop a remote monitoring web app for COVID-19 patients in Armenia: design and development with agile methodology. JMIR Hum Factors 2022 Nov 25;9(4):e40110 [FREE Full text] [doi: 10.2196/40110] [Medline: 36350739]
64. Bartlett R, Boyle JA, Simons Smith J, Khan N, Robinson T, Ramaswamy R. Evaluating human-centred design for public health: a case study on developing a healthcare app with refugee communities. Res Involv Engagem 2021 May 30;7(1):32 [FREE Full text] [doi: 10.1186/s40900-021-00273-2] [Medline: 34053451]
65. Beeman A, Kwesiga J, Ippoliti N, Bhandari T, Pandya G, Acam FA, et al. Using human-centered design to co-design dedicated menstrual health spaces with people who menstruate in Bidi Bidi refugee settlement, Uganda: learnings for further adaptation and scale in humanitarian settings. BMC Womens Health 2023 Jun 20;23(1):319 [FREE Full text] [doi: 10.1186/s12905-023-02421-0] [Medline: 37340385]
66. Black CJ, Berent JM, Joshi U, Khan A, Chamlagai L, Shrivastava R, et al. Applying human-centered design in global mental health to improve reach among underserved populations in the United States and India. Glob Health Sci Pract 2023 Feb 28;11(1):e2200312. [doi: 10.9745/ghsp-d-22-00312]
67. Doocy S, Paik K, Lyles E, Tam HH, Fahed Z, Winkler E, et al. Pilot testing and implementation of a mHealth tool for non-communicable diseases in a humanitarian setting. PLoS Curr 2017 Jun 05;9:ecurrents.dis.e98c648aac93797b1996a37de099be74 [FREE Full text] [doi: 10.1371/currents.dis.e98c648aac93797b1996a37de099be74] [Medline: 28744410]
68. Holt E. AI-assisted HIV case finding in Ukraine. The Lancet Microbe 2024 Jul;5(7):631. [doi: 10.1016/s2666-5247(24)00077-6]
69. Spicer N, Harmer A, Aleshkina J, Bogdan D, Chkhatarashvili K, Murzalieva G, et al. Circus monkeys or change agents? Civil society advocacy for HIV/AIDS in adverse policy environments. Soc Sci Med 2011 Dec;73(12):1748-1755. [doi: 10.1016/j.socscimed.2011.08.024] [Medline: 22036298]

Abbreviations

APS: assisted partner service
CASI-Plus: Computer-Assisted Self-Interview–Plus
HCD: human-centered design
HCW: health care worker
mHealth: mobile health
PHC: Public Health Center
REDCap: Research Electronic Data Capture

Edited by A Mavragani; submitted 18.10.24; peer-reviewed by A Zheluk, C Baxter; comments to author 24.11.24; revised version received 09.12.24; accepted 12.12.24; published 30.01.25.

Please cite as:
Puttkammer N, Dunbar E, Germanovych M, Rosol M, Golden M, Hubashova A, Fedorchenko V, Hetman L, Legkostup L, Flowers J, Nesterova O
Human-Centered Design of an mHealth Tool for Optimizing HIV Index Testing in Wartime Ukraine: Formative Research Case Study
JMIR Form Res 2025;9:e66132
URL: https://formative.jmir.org/2025/1/e66132
doi:10.2196/66132
PMID:

©Nancy Puttkammer, Elizabeth Dunbar, Myroslava Germanovych, Mariia Rosol, Matthew Golden, Anna Hubashova, Vladyslav Fedorchenko, Larisa Hetman, Liudmyla Legkostup, Jan Flowers, Olena Nesterova. Originally published in JMIR Formative Research (https://formative.jmir.org), 30.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Evaluating the Impact of Pharmacotherapy in Augmenting Quit Rates Among Hispanic Adults in an App-Delivered Smoking Cessation Intervention: Secondary Analysis of a Randomized Controlled Trial

Margarita Santiago-Torres[1], PhD; Kristin E Mull[1], MS; Brianna M Sullivan[1], MS; Ana Paula Cupertino[2], PhD; Ramzi G Salloum[3], PhD; Matthew Triplette[1,4,5], MD, PhD; Michael J Zvolensky[6,7], PhD; Jonathan B Bricker[1,8], PhD

[1]Division of Public Health Sciences, Fred Hutchinson Cancer Center, Seattle, WA, United States
[2]Department of Public Health Sciences, University of Rochester Medical Center, Rochester, NY, United States
[3]Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida, Gainesville, FL, United States
[4]Division of Pulmonary, Critical Care and Sleep Medicine, Department of Medicine, University of Washington, Seattle, WA, United States
[5]Clinical Research Division, Fred Hutchinson Cancer Center, Seattle, WA, United States
[6]Department of Psychology, University of Houston, Houston, TX, United States
[7]MD Anderson Cancer Center, University of Texas, Houston, TX, United States
[8]Department of Psychology, University of Washington, Seattle, WA, United States

Corresponding Author:
Margarita Santiago-Torres, PhD
Division of Public Health Sciences
Fred Hutchinson Cancer Center
1100 Fairview Avenue North
Seattle, WA, 98109
United States
Phone: 1 206 667 4780
Fax: 1 206 667 5977
Email: msantiag@fredhutch.org

Abstract

Background: Hispanic adults receive less advice to quit smoking and use fewer evidence-based smoking cessation treatments compared to their non-Hispanic counterparts. Digital smoking cessation interventions, such as those delivered via smartphone apps, provide a feasible and within-reach treatment option for Hispanic adults who smoke and want to quit smoking. While the combination of pharmacotherapy and behavioral interventions are considered best practices for smoking cessation, its efficacy among Hispanic adults, especially alongside smartphone app–based interventions, is uncertain.

Objective: This secondary analysis used data from a randomized controlled trial that compared the efficacy of 2 smoking cessation apps, iCanQuit (based on acceptance and commitment therapy) and QuitGuide (following US clinical practice guidelines), to explore the association between pharmacotherapy use and smoking cessation outcomes among the subsample of 173 Hispanic participants who reported on pharmacotherapy use. Given the randomized design, we first tested the potential interaction of pharmacotherapy use and intervention arm on 12-month cigarette smoking abstinence. We then examined whether the use of any pharmacotherapy (ie, nicotine replacement therapy [NRT], varenicline, or bupropion) and NRT alone augmented each app-based intervention efficacy.

Methods: Participants reported using pharmacotherapy on their own during the 3-month follow-up and cigarette smoking abstinence at the 12-month follow-up via web-based surveys. These data were used (1) to test the interaction effect of using pharmacotherapy to aid smoking cessation and intervention arm (iCanQuit vs QuitGuide) on smoking cessation at 12 months and (2) to test whether the use of pharmacotherapy to aid smoking cessation augmented the efficacy of each intervention arm to help participants successfully quit smoking.

Results: The subsample of Hispanic participants was recruited from 30 US states. They were on average 34.5 (SD 9.3) years of age, 50.9% (88/173) were female, and 56.1% (97/173) reported smoking at least 10 cigarettes daily. Approximately 22% (38/173) of participants reported using pharmacotherapy to aid smoking cessation at the 3-month follow-up, including NRT,

XSL•FO
RenderX

varenicline, or bupropion, with no difference between intervention arms. There was an interaction between pharmacotherapy use and intervention arm that marginally influenced 12-month quit rates at 12 months (*P* for interaction=.053). In the iCanQuit arm, 12-month missing-as-smoking quit rates were 43.8% (7/16) for pharmacotherapy users versus 28.8% (19/16) for nonusers (odds ratio 2.21, 95% CI 0.66-7.48; *P*=.20). In the QuitGuide arm, quit rates were 9.1% (2/22) for pharmacotherapy users versus 21.7% (15/69) for nonusers (odds ratio 0.36, 95% CI 0.07-1.72; *P*=.20). Results were similar for the use of NRT only.

Conclusions: Combining pharmacotherapy to aid smoking cessation with a smartphone app–based behavioral intervention that teaches acceptance of cravings to smoke (iCanQuit) shows promise in improving quit rates among Hispanic adults. However, this combined approach was not effective with the US clinical guideline–based app (QuitGuide).

Trial Registration: ClinicalTrials.gov NCT02724462; https://clinicaltrials.gov/study/NCT02724462

International Registered Report Identifier (IRRID): RR2-10.1001/jamainternmed.2020.4055

(JMIR Form Res 2025;9:e69311) doi:10.2196/69311

KEYWORDS

acceptance and commitment therapy; Hispanic or Latino; iCanQuit; QuitGuide; smartphone apps; smoking cessation; mobile phone

Introduction

Background

In the United States, 1 in 5 adults are Hispanic or Latino (referred to hereafter as Hispanic). Hispanic adults in the United States constitute the largest ethnic minority group [1,2]. In fact, it is estimated that by 2026, the Hispanic population will reach 111 million [3]. Presently, there are approximately 4 million Hispanic adults who smoke cigarettes [4]. Even though smoking rates are lower among Hispanic adults (7.7% vs 11.7%) when compared to the general population of adults in the United States who smoke [5], the prevalence of smoking varies across different Hispanic nationalities [6-8]. For instance, people from Puerto Rico and Cuba have higher than average smoking rates at 25.5% and 19.8%, respectively. Hispanic adults also face higher smoking-attributable morbidity and mortality compared to their non-Hispanic White counterparts [9-11]. Despite these health inequities in smoking and the rapid population growth of the Hispanic population in the United States, research on smoking cessation interventions for this group is sparse [12], with only 2 full-scale randomized controlled trials (RCTs) with 6-month follow-up or longer conducted to date [13,14].

While evidence-based smoking cessation treatments are generally within reach of the US general adult population [15,16], Hispanic adults often lack access to these treatments [17]. Reasons for these accessibility barriers include health providers being 50% less likely to offer quit-smoking advice to Hispanic adults compared to their non-Hispanic counterparts, which greatly contributes to significant inequities in smoking cessation efforts in this group [4,17]. Moreover, many Hispanic adults attempt to quit cigarette smoking cold turkey, which typically results in a very low likelihood of success of 2% to 5% [18,19]. Addressing these challenges requires research to develop culturally inclusive behavioral interventions that are within reach and tailored to the unique needs of Hispanic adults who smoke and want to quit smoking.

Free and remotely delivered digital interventions may improve access to evidence-based smoking cessation programs and thereby help improve smoking cessation outcomes [20]. Only 1 RCT, with a follow-up of 6 months or longer, has examined a digital smoking cessation intervention for Hispanic adults [21]. This study randomly assigned 457 English and Spanish-speaking Hispanic adults to receive either a culturally tailored mobile health smoking cessation intervention (Decídetexto) or standard of care, including written materials from the American Cancer Society and the National Cancer Institute (NCI), along with access to state-funded tobacco quitlines. Nicotine replacement therapies (NRTs; patches and gum) were offered to both arms at no cost. Results at 6 months showed higher (but nonsignificant) biochemically verified quit rates (Decídetexto: 33/229, 14.4% vs control group: 21/228, 9.2%; *P*=.09) and greater NRT use in the intervention group compared to standard of care (at least 1 day of NRT: 174/192, 90.6% vs 139/198, 70.2%; *P*=.01). However, it remains unclear if the use of pharmacotherapy to aid smoking cessation contributed to these quit rates, as both groups had access to free medication.

Digital interventions, like smartphone apps, offer an alternative approach to SMS text messaging or website-based interventions for smoking cessation [22], with over 500 smoking cessation apps currently available (R Nelson, personal communication, 2020). Although 91% of Hispanic adults in the United States own smartphones [23], to our knowledge, only 1 study has explored the potential efficacy of an app-based intervention for helping Hispanic adults quit smoking [24]. This study was a secondary analysis of a Hispanic adult subsample (n=210, 8.7% of the total sample) from a full-scale RCT comparing the iCanQuit app (Moby, Inc; based on acceptance and commitment therapy [ACT]) with the NCI's QuitGuide app (National Cancer Institute; based on US clinical practice guidelines [USCPGs]) for smoking cessation [24,25]. Results showed significantly higher 12-month quit rates in iCanQuit (28/82, 34.1%) versus QuitGuide (19/94, 20.2%; odds ratio [OR] 2.20, 95% CI 1.10-4.41; *P*=.02). This finding is significant, given the challenges faced by Hispanic adults in accessing evidence-based smoking cessation treatments, such as lack of medical insurance, ethnic discrimination, and cost [26,27].

Although app-based smoking cessation interventions hold promise, USCPGs emphasize that combining behavioral interventions with pharmacotherapy to aid smoking cessation

XSL•FO
RenderX

is more efficacious than either approach alone [28,29]. In the iCanQuit parent RCT, pharmacotherapy use was an exclusion criterion, but participants could opt to use it on their own after randomization [25]. In fact, both the iCanQuit and QuitGuide app–based interventions educate users on Food and Drug Administration (FDA)–approved pharmacotherapy to aid cessation [25]. In the full sample of the iCanQuit parent RCT, participants in the iCanQuit arm who used pharmacotherapy to aid smoking cessation on their own had significantly higher quit rates at 12 months compared to nonusers (users: 70/176, 39.8% vs nonusers: 192/691, 27.8%; P=.002) [30]. Whether the use of pharmacotherapy to aid cessation improved the quit rates of Hispanic participants, specifically, remains unknown.

Hispanic adults in the United States have the lowest rate of pharmacotherapy use to aid smoking cessation compared to other racial or ethnic groups [4,17,31]. For example, the rate of pharmacotherapy use is approximately 25.2% among non-Hispanic Black adults and 23.6% among non-Hispanic White adults but only 16.6% among Hispanic adults. In fact, reports from population-based studies of NRT use during quit attempts showed that an overwhelming 91% of Hispanic adults had never used NRT to aid their quit attempts [19]. There are various reasons why Hispanic adults may be more reluctant to use pharmacotherapy to aid their quit attempts. For instance, financial barriers and being medically uninsured (24% of US Hispanic adults) may play an important role [32,33]. Other reasons may include having a mistrust of the medical system [34] and having misconceptions about NRT use, such as NRT being addictive [31,33,35,36]. Prior research also highlights to a cultural emphasis on “voluntad propia” or the willpower to quit smoking that is unique to the Hispanic population [36]. In fact, when compared to non-Hispanic adults, Hispanic adults who smoke and have tried to quit smoking are twice as likely to cite willpower as their main reason for not trying pharmacotherapy to aid their quit attempts [19].

Objectives

To address existing knowledge gaps and build on prior studies [24,25], this secondary analysis used data from an RCT with long-term smoking cessation outcomes comparing 2 smoking cessation apps, iCanQuit (based on ACT) and QuitGuide (following USCPG) [25]. We explored the association between pharmacotherapy use and smoking cessation outcomes in the subsample of 173 English-speaking Hispanic adults recruited nationwide from 30 US states. Given the randomized design, we first tested the potential interaction of pharmacotherapy use and intervention arm on 12-month cigarette smoking abstinence. We then examined whether the use of any pharmacotherapy (ie, NRT, varenicline, or bupropion) augmented each app-based intervention efficacy to help Hispanic adults quit smoking. Finally, we compared NRT (eg, patch, gum, lozenge, and inhaler) users with nonusers, noting the scalability potential of mailed NRT for remote interventions compared to other smoking cessation medications.

Methods

Overview

The details of the iCanQuit RCT have been previously published [25]. Adults aged 18 years and older who reported smoking cigarettes every day, who had access to their own smartphones, and who had a desire to quit cigarette smoking within 30 days were eligible to participate. Exclusion criteria were an inability to read English, ongoing smoking cessation treatment including the use of pharmacotherapy at the time of enrollment only, prior use of iCanQuit or QuitGuide app–based interventions, or having a household member already enrolled in the trial. Participants were not excluded from the study if they chose to use pharmacotherapy to aid cessation on their own after randomization.

Ethical Considerations

All participants provided informed consent. Study procedures were approved by the Fred Hutchinson Cancer Center Institutional Review Board (IRB FHIRB0008317). Participants were compensated up to US $105 for completing data collection at the 3-, 6-, and 12-month follow-ups (US $35 per participant per follow-up). The data is de-identified.

Recruitment, Study Population, and Enrollment

Recruitment spanned from May 2017 to September 2018. Eligible participants were randomly assigned to iCanQuit or QuitGuide for 12 months. Follow-up data collection spanned from August 2017 to December 2019 via web-based surveys at the 3-, 6-, and 12-month follow-up after randomization. From the original 2415 adults in the iCanQuit parent RCT, a total of 173 Hispanic or Latino participants (173/2415, 7.2% of the total sample) who provided pharmacotherapy data at the 3-month follow-up were included in this paper’s secondary analysis (ie, pharmacotherapy data at the 3-month follow-up were missing for 37 Hispanic participants of 210 enrolled in the study).

Smartphone App–Based Interventions

The iCanQuit app, described in great detail elsewhere [25], comprises 8 ACT-based levels emphasizing two core processes of ACT: (1) acceptance of cravings to smoke, including physical sensations, emotions, and thoughts that cue smoking, and (2) aligning personal values, such as family, community, and health, with a smoke-free life. The “preparing to quit” phase fosters acceptance through mindfulness. It includes an “urge help” function and encourages users to track urges and cigarettes smoked daily. The “after you quit” phase focuses on motivation and relapse prevention.

QuitGuide, a free app developed by the NCI that is based on the USCPGs for smoking cessation [28], offers educational resources and techniques for quitting smoking and preventing relapse. It provides information on smoking cravings and quitting barriers. Unlike iCanQuit, which motivates through personal life values, QuitGuide focuses on health consequences and avoidance of cravings to smoke. Neither the content of iCanQuit nor QuitGuide was specifically designed for Hispanic adults who smoke. Neither arm provided pharmacotherapy to aid smoking cessation, coaching, or other interventions. Both

XSL•FO
RenderX

apps educated participants on FDA-approved medications to aid smoking cessation, highlighting benefits, where to obtain them, use instructions, and common side effects.

Measures

Baseline Characteristics

Participants completed baseline web-based surveys reporting sociodemographic, geographic, and smoking characteristics. Data collected included participants' self-reported age, sex, race, ethnicity, higher level of education attained, household annual income, employment status, marital status, sexual orientation, and household zip code [37-41]. Baseline smoking characteristics included daily cigarette consumption, cigarette dependence using the Fagerström Test for Cigarette Dependence [42], use of other nicotine-containing tobacco products including e-cigarettes, previous quit attempts, self-confidence in quitting, and friends and family who smoke. Alcohol consumption was assessed using the Quick Drinking Screen [43].

Pharmacotherapy Use

At the 3-month follow-up survey, participants were asked if they used NRTs or pharmacotherapy to aid smoking cessation since joining the study (ie, within the first 3 months following randomization). This time frame was selected because the content of each app is designed to be completed within approximately 6 weeks of initial use. Thus, assessing pharmacotherapy use by 3 months provides a relevant timeline to examine its potential synergistic effects with the content of the app-based interventions. Participants who reported using nicotine gum, patch, lozenge, or inhaler and varenicline or bupropion were classified as using pharmacotherapy. Participants selecting gum, patch, lozenge, or inhaler but not varenicline or bupropion were classified as using only NRT.

Smoking Abstinence

To assess cigarette smoking abstinence, participants were asked: "When was the last time you smoked, or even tried, a cigarette?" in the 12-month follow-up survey. Response choices were "earlier today," "24 hours ago," "2-7 days ago," "8-30 days ago," and "over 30 days ago" [25,44]. To be categorized as abstinence for the 30-day primary cessation outcomes, participants needed to report no cigarette consumption in the past 30 days at the time of the 12-month follow-up survey. The primary outcome was 30-day cigarette smoking abstinence analyzed through complete-case analysis and sensitivity analyses including missing-as-smoking and multiple imputation. Additional cessation outcomes included cessation of all nicotine and tobacco products (ie, cigarettes, e-cigarettes or vaping, chewing tobacco, snus, hookahs, cigars, cigarillos, tobacco pipes, and kreteks) and prolonged abstinence from cigarettes.

Statistical Analysis

Baseline characteristics were compared between groups by use of pharmacotherapy using Fisher exact tests for categorical variables and 2-sided *t* tests for continuous variables. To determine whether pharmacotherapy use or use of NRT alone by 3 months differentially impacted the effect of the app-based intervention arm on the primary 12-month cessation outcome, we used logistic regression models with a pharmacotherapy use or NRT use by intervention interaction term. Logistic regression models were then used to estimate the effect of pharmacotherapy use or NRT use on 12-month cessation, separately for each intervention arm [44]. For multiple imputation of missing 12-month smoking status, we used multivariate imputation by chained equations in the R package *mice* [45]. All models were adjusted for randomization stratification factors from the parent RCT: smoking frequency, education, and depression. Statistical analyses were carried out using R software (version 4.2.3; R Foundation for Statistical Computing), with 2-sided tests and α=.05.

Results

Baseline Characteristics

Hispanic participants were recruited from 30 US states, averaging 34.5 years in age, with 50.9% (88/173) being female; 20.2% (35/173) identifying as sexual or gender minority; and 30.6% (53/173) being married or partnered. Regarding race, 59% (102/173) identified as White, 15% (26/173) identified as multiracial, and 13.9% (24/173) did not report their race (Table 1). Regarding education, employment, and income, 46.8% (81/173) had a high school diploma or lower; 45.7% (79/173) were unemployed, disabled, or out of the labor force; and 42.8% (35/173) had annual household incomes of US $20,000 or less. More than half (94/173, 54.7%) screened positive for depression symptoms, 29.8% (51/173) for generalized anxiety, 31.8% (54/173) for panic disorder, 48.2% (82/173) for posttraumatic stress disorder, and 34.5% (59/173) for social anxiety disorder. Additionally, 19.1% (33/173) reported bipolar disorder or schizophrenia. On average, participants reported smoking an average of 16.2 (SD 15.9) cigarettes per day, 46.8% (81/173) had high cigarette dependence, 77.5% (134/173) smoked for at least 10 years, and 24.3% (42/173) were dual users of combustible cigarettes and e-cigarettes.

XSL•FO
RenderX

Table 1. Baseline characteristics of Hispanic participants.

Characteristic	Participants, n	Overall (N=173)	Did not use pharmacotherapy (n=135, 78%)	Used pharmacotherapy (n=38, 22%)	*P* value
Age (years), mean (SD)	173	34.5 (9.3)	34.2 (9.5)	35.5 (8.5)	.44
Sex, n (%)	173				.24
Female		88 (50.9)	68 (50.4)	20 (52.6)	
Male		82 (47.4)	65 (48.1)	17 (44.7)	
Transgender female		1 (0.6)	0 (0)	1 (2.6)	
Transgender male		2 (1.2)	2 (1.5)	0 (0)	
Race, n (%)	173				.19
American Indian or Alaska Native		8 (4.6)	8 (5.9)	0 (0)	
Black or African American		13 (7.5)	8 (5.9)	5 (13.2)	
Multiracial		26 (15)	21 (15.6)	5 (13.2)	
White		102 (59)	77 (57)	25 (65.8)	
Unknown (unreported) race		24 (13.9)	21 (15.6)	3 (7.9)	
LGBT[a], n (%)	173	35 (20.2)	28 (20.7)	7 (18.4)	.93
Education, n (%)	173				.93
GED[b], high school or less education		81 (46.8)	63 (46.7)	18 (47.4)	
Some college, no degree		63 (36.4)	50 (37)	13 (34.2)	
College degree or higher		29 (16.8)	22 (16.3)	7 (18.4)	
Employment status, n (%)	173				.04
Employed		94 (54.3)	68 (50.4)	26 (68.4)	
Unemployed		28 (16.2)	27 (20)	1 (2.6)	
Disabled		25 (14.5)	18 (13.3)	7 (18.4)	
Out of labor force		26 (15)	22 (16.3)	4 (10.5)	
Income (US $), n (%)	173				.17
Less than $20,000 per year		74 (42.8)	62 (45.9)	12 (31.6)	
$20,000-$54,999 per year		77 (44.5)	55 (40.7)	22 (57.9)	
$55,000 per year or more		22 (12.7)	18 (13.3)	4 (10.5)	
Married, n (%)	173	53 (30.6)	43 (31.9)	10 (26.3)	.65
Rural residence[c], n (%)	173	21 (12.1)	19 (14.1)	2 (5.3)	.23
Mental health positive screening, n (%)					
Depression (CES-D)[d]	172	94 (54.7)	70 (52.2)	24 (63.2)	.31
General anxiety (GAD-7)[e]	171	51 (29.8)	36 (27.1)	15 (39.5)	.20
Panic disorder (ANSQ)[f]	170	54 (31.8)	41 (31.1)	13 (34.2)	.87
PTSD (PCL-6)[g]	170	82 (48.2)	60 (45.5)	22 (57.9)	.24
Social anxiety (mini-SPIN)[h]	171	59 (34.5)	40 (30.1)	19 (50)	.04
Self-reported bipolar or schizophrenia[i], n (%)	173	33 (19.1)	26 (19.3)	7 (18.4)	>.99
Alcohol use, n (%)					
Drinks per drinking day, mean (SD)	166	2.3 (4.0)	2.3 (4.2)	2.4 (3.3)	.90
Heavy drinking[j], n (%)	166	29 (17.5)	22 (16.9)	7 (19.4)	.92
Smoking behavior					

XSL•FO
RenderX

Characteristic	Participants, n	Overall (N=173)	Did not use pharmacotherapy (n=135, 78%)	Used pharmacotherapy (n=38, 22%)	*P* value
Cigarettes smoked per day, mean (SD)	173	16.2 (15.9)	15.9 (16.2)	17.2 (15.1)	.67
FTCD score[k], mean (SD)	173	5.2 (2.0)	5.3 (1.9)	5.2 (2.5)	.97
High nicotine dependence (FTCD≥6), n (%)	173	81 (46.8)	60 (44.4)	21 (55.3)	.32
First cigarette within 5 minutes of waking, n (%)	173	75 (43.4)	58 (43)	17 (44.7)	.99
Smokes greater than one-half pack (10 cigarettes) per day, n (%)	173	97 (56.1)	72 (53.3)	25 (65.8)	.24
Smokes more than 1 pack (20 cigarettes) per day, n (%)	173	18 (10.4)	13 (9.6)	5 (13.2)	.74
Smoking for 10 or more years, n (%)	173	134 (77.5)	100 (74.1)	34 (89.5)	.07
Used e-cigarettes in the past month, n (%)	173	42 (24.3)	35 (25.9)	7 (18.4)	.46
Quit attempts in the past 12 months, mean (SD)	166	3.3 (15.8)	3.8 (17.8)	1.8 (3.2)	.51
Confidence to quit smoking[l], mean (SD)	173	67.9 (28.1)	68.0 (27.6)	67.8 (29.9)	.98
Friend and partner smoking					
Close friends who smoke, mean (SD)	173	2.8 (1.7)	2.7 (1.7)	2.9 (1.8)	.48
Housemates who smoke, mean (SD)	173	1.5 (1.1)	1.5 (1.1)	1.5 (0.8)	.96
Living with a partner who smokes, n (%)	173	56 (32.4)	44 (32.6)	12 (31.6)	>.99

[a]LGBT: lesbian, gay, bisexual, or transgender.

[b]GED: general education development.

[c]To ascertain the residential classification of participants into urban or rural areas within the United States, we used the *zipcode* R library [46] to link participants' zip codes to geographic locations. The classification was based on rural-urban commuting area codes at the subcounty level [37-41].

[d]Positive screening results for depression via Center for Epidemiologic Studies Depression Scale (CES-D)-20; threshold ≥16 [47].

[e]Positive screening results for generalized anxiety via General Anxiety Disorder (GAD)-7; threshold ≥10 [48].

[f]Positive screening results for panic disorder via Autonomic Nervous System Questionnaire (ANSQ) [49]. A positive screening was recorded for individuals reporting at least 1 panic attack in the past month, provided that at least 1 of these incidents occurred in a situation in which they were not in danger or the center of attention.

[g]Positive screening results for posttraumatic stress disorder (PTSD) via Post-Traumatic Checklist (PCL-6); threshold ≥14 [50].

[h]Positive screening results for social anxiety via Mini-Social Phobia Inventory (mini-SPIN); threshold ≥6 [51].

[i]Presence of bipolar disorder or schizophrenia was self-reported via question: "Do you have any of the following mental health disorders?" Multiselect response options included anxiety disorder, depression disorder, bipolar disorder, schizophrenia, alcohol abuse disorder, drug abuse disorder, or none of the above [52,53].

[j]Heavy drinking is defined as 4 or more drinks on a typical drinking day for women and 5 or more drinks on a typical drinking day for men within the past 30 days [43].

[k]Fagerström Test for Cigarette Dependence (FTCD) score [42].

[l]Range 0-100, where 0 indicates not at all confident and 100 indicates extremely confident.

Pharmacotherapy Use and Outcome Retention Rates

At the 3-month follow-up, 22% (38/173) of participants used pharmacotherapy (NRT, varenicline, or bupropion). There was no statistically significant difference in pharmacotherapy use between treatment arms (iCanQuit: 16/82, 19.5% vs QuitGuide: 22/91, 24.2%; *P*=.41). Approximately 73.6% (28/38) of participants who reported pharmacotherapy use opted for NRT only (ie, nicotine patch, gum, inhaler, or lozenge), whereas 26.3% (10/38) reported using a combination of NRT and medication (varenicline or bupropion). Outcome data retention at 12 months was high, with an overall rate of 93.6% (162/173). This rate did not differ by intervention arms (iCanQuit: 74/82, 90.2% vs QuitGuide: 88/91, 96.7%; *P*=.09) and pharmacotherapy use (users: 36/38, 94.7% vs nonusers: 126/135, 93.3%; *P*=.82).

Smoking Abstinence

Report of any FDA-approved pharmacotherapy use by 3 months had a marginal differential impact on 12-month quit rates by intervention arm (*P* for interaction=.053). In the iCanQuit arm, there was a signal indicating higher odds of quitting with pharmacotherapy. Specifically, 12-month cigarette smoking abstinence rates in the iCanQuit arm were 43.8% (7/16) for pharmacotherapy users versus 28.8% (19/66) for nonusers (OR 2.21, 95% CI 0.66-7.48; *P*=.20). In contrast, cigarette smoking abstinence rates in the QuitGuide arm were 9.1% (2/22) for pharmacotherapy users versus 21.7% (15/69) for nonusers (OR 0.36, 95% CI 0.07-1.72; *P*=.20). Similar trends were observed

XSL•FO
RenderX

for other secondary smoking cessation outcomes, but none reached statistical significance (Table 2).

The impact of using pharmacotherapy to aid cessation at 12 months was more pronounced when NRT was the sole form of pharmacotherapy used, although the interaction effect between NRT use and app-based intervention arm on the 12-month smoking cessation outcome was not statistically significant (*P*=.09). In the iCanQuit arm, 30-day cigarette smoking abstinence rates at 12 months were 45.5% (5/11) for NRT users versus 28.8% (20/69) for nonusers (OR 2.24, 95% CI 0.54-9.27; *P*=.26). In contrast, in the QuitGuide arm, 30-day cigarette smoking abstinence rates at 12 months were 11.8% (2/17) for NRT users versus 21.7% (15/69) for nonusers (OR 0.49, 95% CI 0.10-2.45; *P*=.36; Table 3).

Table 2. Association between any pharmacotherapy use and cessation outcomes in the iCanQuit and QuitGuide treatment arms among Hispanic participants.

12-Month cessation outcome	Interaction between treatment arm and use of pharmacotherapy		Treatment arm	Participants, n	Overall, n (%)	Did not use pharmacotherapy, n (%)	Used pharmacotherapy, n (%)	OR[a] (95% CI)[b]	*P* value
	β (SE)	*P* value							
30-Day PPA[c] from cigarettes, complete case	1.97 (1.02)	.053	iCanQuit	74	26 (35.1)	19 (31.7)	7 (50)	2.59 (0.69-9.73)	.16
30-Day PPA from cigarettes, complete case	1.97 (1.02)	.053	QuitGuide	88	17 (19.3)	15 (22.7)	2 (9.1)	0.34 (0.07-1.64)	.18
30-Day PPA from cigarettes, multiple imputation	1.86 (1.02)	.07	iCanQuit	820	290 (35.4)	214 (32.4)	76 (47.5)	2.27 (0.61-8.41)	.23
30-Day PPA from cigarettes, multiple imputation	1.86 (1.02)	.07	QuitGuide	910	177 (19.5)	157 (22.8)	20 (9.1)	0.34 (0.07-1.64)	.18
30-Day PPA from cigarettes, missing-as-smoking	1.79 (0.99)	.07	iCanQuit	82	26 (31.7)	19 (28.8)	7 (43.8)	2.21 (0.66-7.48)	.20
30-Day PPA from cigarettes, missing-as-smoking	1.79 (0.99)	.07	QuitGuide	91	17 (18.7)	15 (21.7)	2 (9.1)	0.36 (0.07-1.72)	.20
30-Day PPA from all nicotine and tobacco products[d]	1.90 (1.03)	.06	iCanQuit	74	21 (28.4)	15 (25)	6 (42.9)	2.79 (0.72-10.77)	.14
30-Day PPA from all nicotine and tobacco products[d]	1.90 (1.03)	.06	QuitGuide	88	16 (18.2)	14 (21.2)	2 (9.1)	0.37 (0.08-1.81)	.22
Prolonged abstinence from cigarettes[e]	1.06 (1.47)	.64	iCanQuit	54	10 (18.5)	8 (17.4)	2 (25)	1.67 (0.27-10.36)	.58
Prolonged abstinence from cigarettes[e]	1.06 (1.47)	.64	QuitGuide	69	6 (8.7)	5 (9.6)	1 (5.9)	0.58 (0.06-5.49)	.64

[a]OR: odds ratio.

[b]OR for use versus no use of pharmacotherapy by 3 months after randomization. All models were adjusted for factors used in stratified randomization: smoking >20 cigarettes per day, high school or lower education, and positive screening for depression.

[c]PPA: point prevalence abstinence.

[d]Including any kind of e-cigarettes or vaping, chewing tobacco, snus, hookahs, cigars, cigarillos, tobacco pipes, and kreteks.

[e]Prolonged abstinence is defined as no smoking since 3 months after randomization, using the self-reported date of the last cigarette.

XSL•FO
RenderX

Table 3. Association between use of nicotine replacement therapy (NRT) only and cessation outcomes in the iCanQuit and QuitGuide treatment arms among Hispanic participants.

12-Month cessation outcome	Interaction between treatment arm and use of NRT,		Treatment arm	Participants, n	Overall, n (%)	Did not use NRT, n (%)	Used NRT, n (%)	OR[a] (95% CI)[b]	*P* value
	β (SE)	*P* value							
30-Day PPA[c] from cigarettes, complete case	1.88 (1.11)	.09	iCanQuit	69	24 (34.8)	19 (31.7)	5 (55.6)	3.09 (0.62-15.41)	.17
30-Day PPA from cigarettes, complete case	1.88 (1.11)	.09	QuitGuide	83	17 (20.5)	15 (22.7)	2 (11.8)	0.47 (0.09-2.33)	.36
30-Day PPA from cigarettes, multiple imputation	1.69 (1.11)	.13	iCanQuit	770	270 (35.1)	214 (32.4)	56 (50.9)	2.46 (0.51-11.90)	.27
30-Day PPA from cigarettes, multiple imputation	1.69 (1.11)	.13	QuitGuide	860	177 (20.6)	157 (22.8)	20 (11.8)	0.47 (0.10-2.35)	.36
30-Day PPA from cigarettes, missing-as-smoking	1.56 (1.06)	.14	iCanQuit	77	24 (31.2)	19 (28.8)	5 (45.5)	2.24 (0.54-9.27)	.26
30-Day PPA from cigarettes, missing-as-smoking	1.56 (1.06)	.14	QuitGuide	86	17 (19.8)	15 (21.7)	2 (11.8)	0.49 (0.10-2.45)	.39
30-Day PPA from all nicotine and tobacco products[d]	1.66 (1.11)	.14	iCanQuit	69	19 (27.5)	15 (25)	4 (44.4)	2.73 (0.54-13.7)	.22
30-Day PPA from all nicotine and tobacco products[d]	1.66 (1.11)	.14	QuitGuide	83	16 (19.3)	14 (21.2)	2 (11.8)	0.51 (0.10-2.54)	.41
Prolonged abstinence from cigarettes[e]	0.29 (1.68)	.91	iCanQuit	51	9 (17.6)	8 (17.4)	1 (20)	1.20 (0.11-12.87)	.88
Prolonged abstinence from cigarettes[e]	0.29 (1.68)	.91	QuitGuide	64	6 (9.4)	5 (9.6)	1 (8.3)	0.90 (0.09-9.10)	.93

[a]OR: odds ratio.

[b]OR for use versus no use of NRT by 3 months after randomization. All models were adjusted for factors used in stratified randomization: smoking >20 cigarettes per day, high school or lower education, and positive screening for depression.

[c]PPA: point prevalence abstinence.

[d]Including any kind of e-cigarettes or vaping, chewing tobacco, snus, hookahs, cigars, cigarillos, tobacco pipes, and kreteks.

[e]Prolonged abstinence is defined as no smoking since 3 months after randomization, using the self-reported date of the last cigarette.

Discussion

Principal Findings

This paper aimed to address the critical knowledge gaps regarding the efficacy of smoking cessation app–based interventions combined with pharmacotherapy to help individuals quit cigarette smoking, particularly among Hispanic adults who tends to underuse evidence-based smoking cessation treatments, including pharmacotherapy [17,25,26,54]. The findings suggest that self-selected use of pharmacotherapy, including NRT, varenicline, or bupropion, improved 12-month quit rates among participants in the iCanQuit arm (users: 7/14, 50% vs nonusers: 19/60, 31.7%). However, this effect was not observed among participants in the QuitGuide arm (users: 2/22, 9.1% vs nonusers: 15/66, 22.7%).

The differential effect of using pharmacotherapy between participants in the iCanQuit and QuitGuide arms suggests that combining pharmacotherapy with an ACT-based intervention (iCanQuit) may have synergistically helped participants manage smoking cravings during quit attempts. The results suggest a combined and synergetic effect of managing short- and long-term withdrawal symptoms, resulting in a more effective treatment than ACT-based behavioral interventions alone. It is plausible that short-term withdrawal symptoms were ameliorated using pharmacotherapy, while the skills using iCanQuit to manage withdrawal symptoms may have helped in the long term. While the QuitGuide app–based intervention also mentions FDA-approved pharmacotherapy to aid cessation, the app usually portrays medications as separate rather than complementary to learning behavioral skills to quit smoking

XSL•FO
RenderX

[25,28]. It is then possible that participants in the QuitGuide arm may have not understood the potential complementary role of using pharmacotherapy to aid smoking cessation alongside the app. It is plausible that guidance is needed when pharmacotherapy is offered alongside behavioral interventions to have the desired synergetic effect [28]. Further investigation is warranted to understand the improved quit rates with pharmacotherapy in iCanQuit but not QuitGuide.

It is important to note that Hispanic adults in this study opted independently to use pharmacotherapy to aid smoking cessation, albeit at a nonsignificantly lower rate than non-Hispanic participants (38/173, 22% vs 581/1915, 30.3%; P=.33). This is contrary to population-based reports among US Hispanic adults being 50% less likely to use evidence-based smoking cessation treatments compared to the general population of adults who smoke in the United States [17,26,54]. Our results showed the potential of evidence-based smoking cessation interventions that are delivered remotely in improving awareness of FDA-approved pharmacotherapy to aid cessation.

Limitations

This study has several limitations. First, the reliance on secondary data inherently limits findings, as participants were not randomized to receive pharmacotherapy, resulting in associations rather than causal effects. Second, details on dosage, adherence, and duration of pharmacotherapy use are unknown; thus, the nature of pharmacotherapy usage remains uncertain. Third, Hispanic participants were English speakers who smoked daily, which limits the generalizability of the findings to Spanish-speaking Hispanic adults, or those who smoke less than daily [55]. Finally, reliance on self-reported use of pharmacotherapy and smoking abstinence introduces subjectivity and reporting bias. Future multimethod assessment protocols for smoking behavior would therefore be important.

This study also had several strengths. First, it is the first study to explore the effect of an app-based smoking cessation intervention combined with pharmacotherapy use as a potential treatment for Hispanic adults who smoke and wish to quit [17,26,54]. Second, the trial's long-term smoking cessation outcomes (12 months) and high retention rates that did not differ by intervention arm or pharmacotherapy use helped reduce any potential bias in the complete-case analysis and missing-as-smoking sensitivity analyses. Third, despite the modest sample size, Hispanic participants were recruited from 30 US states, underscoring the geographic diversity of the results and the potential of app-based interventions for widespread reach and broad dissemination.

The results from this secondary analysis have significant public health implications for designing future interventions to help Hispanic adults quit smoking. First, app-based smoking cessation interventions offer a viable approach for hard-to-reach population who are lacking access to evidence-based treatments, including 24% of uninsured Hispanic adults in the United States [32]. Second, app-based delivered interventions could help overcome accessibility barriers like transportation and time constraints. Third, our results show a notable increase in quit rates when using NRT alongside iCanQuit compared to nonuse (5/9, 55.6% vs 20/63, 31.7%). This is noteworthy, as over-the-counter NRT has shown high reach and efficacy [16,29]. Additionally, NRT is readily available and cost-effective and requires minimal medical oversight compared to other medications (varenicline or bupropion) [56,57]. Mailed NRT [58,59], as seen in previous trials, could hold potential for broad dissemination and scalability of smoking cessation programs at the population level. Moreover, future interventions for the Hispanic population could be offered in Spanish, expanding their reach to the 28% of Spanish-speaking Hispanic adults currently living in the United States [32].

Conclusions

This secondary analysis underscores the nuanced role of pharmacotherapy in enhancing the efficacy of app-based smoking cessation interventions among Hispanic adults. Pharmacotherapy use, particularly when paired with an ACT-based app (iCanQuit), appears to improve quit rates, highlighting a potential pathway to address disparities in smoking cessation outcomes. In contrast, the USCPG-based app (QuitGuide) did not result in a similar synergistic effect when combined with pharmacotherapy. Considering the scalability of NRT, future research should evaluate the efficacy and cost-effectiveness of remotely providing NRT alongside iCanQuit compared to iCanQuit alone.

Acknowledgments

The authors appreciate the tireless contributions of the entire study staff, most notably Eric Meier; Carolyn Ehret; Brianna Sullivan; the design services of Ayogo Inc; and the development services of Moby, Inc. The authors are very appreciative of the study participants. The study was funded by grant R01CA192849 awarded to JBB from the National Cancer Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

JBB, MS-T, and KEM conceptualized the study. MS-T led the manuscript writing. KEM led and conducted the data analysis. All authors assisted in manuscript writing and provided critical review. All authors have read and agreed to the published versions of the manuscript.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 966 KB - formative_v9i1e69311_app1.pdf]

References

1. Jones N, Marks R, Ramirez R, Ríos-Vargas M. Racial and ethnic diversity in the United States: 2010 census and 2020 census. U.S. Census Bureau. 2020. URL: https://www.census.gov/library/visualizations/interactive/racial-and-ethnic-diversity-in-the-united-states-2010-and-2020-census.html [accessed 2024-06-15]
2. Who is Hispanic? Pew Research Center. 2022. URL: https://www.pewresearch.org/fact-tank/2022/09/15/who-is-hispanic/ [accessed 2022-10-10]
3. Hispanic population to reach 111 million by 2060. U.S. Census Bureau. URL: https://www.census.gov/library/visualizations/2018/comm/hispanic-projected-pop.html [accessed 2018-10-09]
4. Babb S, Malarcher A, Schauer G, Asman K, Jamal A. Quitting smoking among adults—United States, 2000–2015. MMWR Morb Mortal Wkly Rep 2017;65(52):1457-1464 [FREE Full text] [doi: 10.15585/mmwr.mm6552a1] [Medline: 28056007]
5. Cornelius ME, Loretan CG, Jamal A, Davis Lynn BC, Mayer M, Alcantara IC, et al. Tobacco product use among adults—United States, 2021. MMWR Morb Mortal Wkly Rep 2023;72(18):475-483 [FREE Full text] [doi: 10.15585/mmwr.mm7218a1] [Medline: 37141154]
6. Kaplan RC, Bangdiwala SI, Barnhart JM, Castañeda SF, Gellman MD, Lee DJ, et al. Smoking among U.S. Hispanic/Latino adults: the Hispanic community health study/study of Latinos. Am J Prev Med 2014;46(5):496-506 [FREE Full text] [doi: 10.1016/j.amepre.2014.01.014] [Medline: 24745640]
7. Merzel CR, Isasi CR, Strizich G, Castañeda SF, Gellman M, Maisonet Giachello AL, et al. Smoking cessation among U.S. Hispanic/Latino adults: findings from the Hispanic Community Health Study/Study of Latinos (HCHS/SOL). Prev Med 2015;81:412-419 [FREE Full text] [doi: 10.1016/j.ypmed.2015.10.006] [Medline: 26515291]
8. Sisti JS, Jasek JP, Farley SM. Heterogeneity in current cigarette smoking among Hispanic/Latino Heritage Groups in New York City, 2003-2016. Ethn Dis 2020;30(1):97-108 [FREE Full text] [doi: 10.18865/ed.30.1.97] [Medline: 31969789]
9. Heron M. Deaths: leading causes for 2016. Natl Vital Stat Rep 2018;67(6):1-77. [doi: 10.15620/cdc:104186]
10. Henley SJ, Thomas CS, Sharapova SR, Momin B, Massetti GM, Winn DM, et al. Vital signs: disparities in tobacco-related cancer incidence and mortality—United States, 2004-2013. MMWR Morb Mortal Wkly Rep 2016;65(44):1212-1218 [FREE Full text] [doi: 10.15585/mmwr.mm6544a3] [Medline: 27832048]
11. Daviglus ML, Talavera GA, Avilés-Santa ML, Allison M, Cai J, Criqui MH, et al. Prevalence of major cardiovascular risk factors and cardiovascular diseases among Hispanic/Latino individuals of diverse backgrounds in the United States. JAMA 2012;308(17):1775-1784 [FREE Full text] [doi: 10.1001/jama.2012.14517] [Medline: 23117778]
12. Webb MS, Rodríguez-Esquivel D, Baker EA. Smoking cessation interventions among Hispanics in the United States: a systematic review and mini meta-analysis. Am J Health Promot 2010;25(2):109-118. [doi: 10.4278/ajhp.090123-lit-25]
13. Simmons VN, Sutton SK, Medina-Ramirez P, Martinez U, Brandon KO, Byrne MM, et al. Self-help smoking cessation intervention for Spanish-speaking Hispanics/Latinxs in the United States: a randomized controlled trial. Cancer 2022;128(5):984-994 [FREE Full text] [doi: 10.1002/cncr.33986] [Medline: 34679188]
14. Cartujano-Barrera F, Arana-Chicas E, Catley D, Cox LS, Diaz FJ, Ellerbeck EF, et al. Decídetexto: mobile cessation support for Latino smokers. Study protocol for a randomized clinical trial. Contemp Clin Trials 2020;99:106188 [FREE Full text] [doi: 10.1016/j.cct.2020.106188] [Medline: 33080379]
15. Smoking cessation: a report of the Surgeon General—executive summary. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health. 2020. URL: https://www.hhs.gov/sites/default/files/2020-cessation-sgr-executive-summary.pdf [accessed 2020-02-23]
16. Patnode CD, Henderson JT, Coppola EL, Melnikow J, Durbin S, Thomas RG. Interventions for tobacco cessation in adults, including pregnant persons: updated evidence report and systematic review for the US Preventive Services Task Force. JAMA 2021;325(3):280-298. [doi: 10.1001/jama.2020.23541] [Medline: 33464342]
17. Babb S, Malarcher A, Asman K, Johns M, Caraballo R, VanFrank B, et al. Disparities in cessation behaviors between Hispanic and non-Hispanic White adult cigarette smokers in the United States, 2000-2015. Prev Chronic Dis 2020;17:E10 [FREE Full text] [doi: 10.5888/pcd17.190279] [Medline: 31999539]
18. Prochaska JJ, Benowitz NL. The past, present, and future of nicotine addiction therapy. Annu Rev Med 2016;67(1):467-486 [FREE Full text] [doi: 10.1146/annurev-med-111314-033712] [Medline: 26332005]
19. Cook-Shimanek M, Burns EK, Levinson AH. Medicinal nicotine nonuse: smokers' rationales for past behavior and intentions to try medicinal nicotine in a future quit attempt. Nicotine Tob Res 2013;15(11):1926-1933 [FREE Full text] [doi: 10.1093/ntr/ntt085] [Medline: 23817584]

XSL•FO
RenderX

20. Bold KW, Garrison KA, DeLucia A, Horvath M, Nguyen M, Camacho E, et al. Smartphone apps for smoking cessation: systematic framework for app review and analysis. J Med Internet Res 2023;25:e45183 [FREE Full text] [doi: 10.2196/45183] [Medline: 37440305]
21. Cartujano-Barrera F, Cox LS, Catley D, Cai X, Diaz FJ, Arana-Chicas E, et al. Decídetexto: mobile cessation support for Latino adults who smoke: a randomized clinical trial. Chest 2024:04904-3 [FREE Full text] [doi: 10.1016/j.chest.2024.07.160] [Medline: 39134144]
22. Wasil AR, Gillespie S, Shingleton R, Wilks CR, Weisz JR. Examining the reach of smartphone apps for depression and anxiety. Am J Psychiatry 2020;177(5):464-465. [doi: 10.1176/appi.ajp.2019.19090905] [Medline: 32354266]
23. Mobile fact sheets: internet and broadband. Pew Research Center. URL: https://www.pewresearch.org/internet/fact-sheet/mobile/ [accessed 2025-01-04]
24. Santiago-Torres M, Mull KE, Sullivan BM, Zvolensky MJ, Bricker JB. Can a smartphone application help Hispanic/Latinx adults quit smoking? A randomized trial secondary analysis. J Contextual Behav Sci 2022;26:261-270 [FREE Full text] [doi: 10.1016/j.jcbs.2022.11.001] [Medline: 36437818]
25. Bricker JB, Watson NL, Mull KE, Sullivan BM, Heffner JL. Efficacy of smartphone applications for smoking cessation: a randomized clinical trial. JAMA Intern Med 2020;180(11):1472. [doi: 10.1001/jamainternmed.2020.4055]
26. Li L, Zhan S, Hu L, Wilson KM, Mazumdar M, Liu B. Examining the role of healthcare access in racial/ethnic disparities in receipt of provider-patient discussions about smoking: a latent class analysis. Prev Med 2021;148:106584. [doi: 10.1016/j.ypmed.2021.106584] [Medline: 33930432]
27. Cohen RA, Cha AE, Terlizzi EP, Martinez ME. Demographic variation in health insurance coverage: United States. Natl Health Stat Report 2019;2021(159):1-15. [doi: 10.15620/cdc:106462]
28. Clinical Practice Guideline Treating Tobacco Use and Dependence 2008 Update Panel, Liaisons, and Staff. A clinical practice guideline for treating tobacco use and dependence: 2008 update. A U.S. Public Health Service report. Am J Prev Med 2008;35(2):158-176 [FREE Full text] [doi: 10.1016/j.amepre.2008.04.009] [Medline: 18617085]
29. Hartmann-Boyce J, Hong B, Livingstone-Banks J, Wheat H, Fanshawe TR. Additional behavioural support as an adjunct to pharmacotherapy for smoking cessation. Cochrane Database Syst Rev 2019;6(6):CD009670 [FREE Full text] [doi: 10.1002/14651858.CD009670.pub4] [Medline: 31166007]
30. Bricker JB, Santiago-Torres M, Mull KE, Sullivan BM, David SP, Schmitz J, et al. Do medications increase the efficacy of digital interventions for smoking cessation? Secondary results from the iCanQuit randomized trial. Addiction 2024;119(4):664-676. [doi: 10.1111/add.16396] [Medline: 38009551]
31. Levinson AH, Pérez-Stable EJ, Espinoza P, Flores ET, Byers TE. Latinos report less use of pharmaceutical aids when trying to quit smoking. Am J Prev Med 2004;26(2):105-111. [doi: 10.1016/j.amepre.2003.10.012] [Medline: 14751320]
32. 11 Facts about Hispanic origin groups in the U.S. Pew Research Center. 2023. URL: https://tinyurl.com/2ukr5hyp [accessed 2024-01-31]
33. Levinson AH, Borrayo EA, Espinoza P, Flores ET, Pérez-Stable EJ. An exploration of Latino smokers and the use of pharmaceutical aids. Am J Prev Med 2006;31(2):167-171. [doi: 10.1016/j.amepre.2006.03.022] [Medline: 16829334]
34. Alexander AC, Hébert ET, Businelle MS, Kendzor DE. Everyday discrimination indirectly influences smoking cessation through post-quit self-efficacy. Drug Alcohol Depend 2019;198:63-69. [doi: 10.1016/j.drugalcdep.2019.01.033] [Medline: 30878768]
35. Pacek LR, McClernon FJ, Bosworth HB. Adherence to pharmacological smoking cessation interventions: a literature review and synthesis of correlates and barriers. Nicotine Tob Res 2018;20(10):1163-1172 [FREE Full text] [doi: 10.1093/ntr/ntx210] [Medline: 29059394]
36. Carpenter MJ, Ford ME, Cartmell K, Alberg AJ. Misperceptions of nicotine replacement therapy within racially and ethnically diverse smokers. J Natl Med Assoc 2011;103(9-10):885-894 [FREE Full text] [doi: 10.1016/s0027-9684(15)30444-2] [Medline: 22364057]
37. Larson EH, Andrilla CHA, Garberson LA, Evans DV. Geographic access to health care for rural medicare beneficiaries in five states: an update. Policy brief. WWAMI Rural Health Research Center, University of Washington. URL: https://familymedicine.uw.edu/rhrc/wp-content/uploads/sites/4/2021/09/RHRC_PBSEP2021_LARSON.pdf [accessed 2025-01-04]
38. Ratcliffe M, Burd C, Holder K, Fields A. Understanding and using American Community Survey data. U.S. Census Bureau. 2016. URL: https://www.census.gov/content/dam/Census/library/publications/2019/acs/ACS_rural_handbook_2019_ch01.pdf [accessed 2024-07-07]
39. Brooks MM, Mueller JT, Thiede BC. County reclassifications and rural-urban mortality disparities in the United States (1970-2018). Am J Public Health 2020;110(12):1814-1816. [doi: 10.2105/AJPH.2020.305895] [Medline: 33058708]
40. Chen X, Orom H, Hay JL, Waters EA, Schofield E, Li Y, et al. Differences in rural and urban health information access and use. J Rural Health 2019;35(3):405-417 [FREE Full text] [doi: 10.1111/jrh.12335] [Medline: 30444935]
41. Unger JM, Moseley A, Symington B, Chavez-MacGregor M, Ramsey SD, Hershman DL. Geographic distribution and survival outcomes for rural patients with cancer treated in clinical trials. JAMA Netw Open 2018;1(4):e181235 [FREE Full text] [doi: 10.1001/jamanetworkopen.2018.1235] [Medline: 30646114]

XSL•FO
RenderX

42. Heatherton TF, Kozlowski LT, Frecker RC, Fagerström KO. The Fagerström Test for Nicotine Dependence: a revision of the Fagerström Tolerance Questionnaire. Br J Addict 1991;86(9):1119-1127. [doi: 10.1111/j.1360-0443.1991.tb01879.x] [Medline: 1932883]
43. Roy M, Dum M, Sobell LC, Sobell MB, Simco ER, Manor H, et al. Comparison of the quick drinking screen and the alcohol timeline followback with outpatient alcohol abusers. Subst Use Misuse 2008;43(14):2116-2223. [doi: 10.1080/10826080802347586] [Medline: 18825590]
44. Santiago-Torres M, Mull KE, Sullivan BM, Bricker JB. Relative efficacy of an acceptance and commitment therapy-based smartphone app with a standard US Clinical Practice Guidelines-based app for smoking cessation in dual users of combustible and electronic cigarettes: secondary findings from a randomized trial. Subst Use Misuse 2024;59(4):591-600. [doi: 10.1080/10826084.2023.2293732] [Medline: 38098199]
45. van Buuren B, Groothuis-Oudshoorn K. mice: multivariate imputation by chained equations in R. J Stat Softw 2011;45(3):1-67. [doi: 10.32614/cran.package.mice]
46. Breen J. zipcode: U.S. ZIP Code database for geocoding. R package version 1.0. R Project. URL: https://CRAN.R-project.org/package=zipcode [accessed 2012-03-12]
47. Radloff LS. The CES-D scale: a self-report depression scale for research in the general population. Appl Psychol Meas 1977;1(3):385-401. [doi: 10.1177/014662167700100306]
48. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
49. Stein MB, Roy-Byrne PP, McQuaid JR, Laffaye C, Russo J, McCahill ME, et al. Development of a brief diagnostic screen for panic disorder in primary care. Psychosom Med 1999;61(3):359-364. [doi: 10.1097/00006842-199905000-00016] [Medline: 10367617]
50. Lang AJ, Wilkins K, Roy-Byrne PP, Golinelli D, Chavira D, Sherbourne C, et al. Abbreviated PTSD Checklist (PCL) as a guide to clinical response. Gen Hosp Psychiatry 2012;34(4):332-338 [FREE Full text] [doi: 10.1016/j.genhosppsych.2012.02.003] [Medline: 22460001]
51. Connor KM, Kobak KA, Churchill LE, Katzelnick D, Davidson JRT. Mini-SPIN: a brief screening assessment for generalized social anxiety disorder. Depress Anxiety 2001;14(2):137-140. [doi: 10.1002/da.1055] [Medline: 11668666]
52. Minimal data set. North American Quitline Consortium (NAQC). 2005. URL: https://www.naquitline.org/page/mds [accessed 2021-03-03]
53. Tedeschi GJ, Cummins SE, Anderson CM, Anthenelli RM, Zhuang Y, Zhu S. Smokers with self-reported mental health conditions: a case for screening in the context of tobacco cessation services. PLoS One 2016;11(7):e0159127 [FREE Full text] [doi: 10.1371/journal.pone.0159127] [Medline: 27391334]
54. Lee J, Contrera Avila J, Ahluwalia JS. Differences in cessation attempts and cessation methods by race/ethnicity among US adult smokers, 2016-2018. Addict Behav 2023;137:107523 [FREE Full text] [doi: 10.1016/j.addbeh.2022.107523] [Medline: 36279711]
55. Pagano A, Gubner N, Le T, Guydish J. Cigarette smoking and quit attempts among Latinos in substance use disorder treatment. Am J Drug Alcohol Abuse 2018;44(6):660-667 [FREE Full text] [doi: 10.1080/00952990.2017.1417417] [Medline: 29333890]
56. Cunningham JA, Kushnir V. Effectiveness of mass distribution of nicotine patches to promote tobacco control in rural versus urban settings. J Epidemiol Community Health 2017;71(5):519. [doi: 10.1136/jech-2016-208103] [Medline: 28396571]
57. Fu SS, van Ryn M, Nelson D, Burgess DJ, Thomas JL, Saul J, et al. Proactive tobacco treatment offering free nicotine replacement therapy and telephone counselling for socioeconomically disadvantaged smokers: a randomised clinical trial. Thorax 2016;71(5):446-453 [FREE Full text] [doi: 10.1136/thoraxjnl-2015-207904] [Medline: 26931362]
58. Kushnir V, Sproule BA, Cunningham JA. Mailed distribution of free nicotine patches without behavioral support: predictors of use and cessation. Addict Behav 2017;67:73-78. [doi: 10.1016/j.addbeh.2016.12.008] [Medline: 28039798]
59. Cunningham JA, Kushnir V, Selby P, Tyndale RF, Zawertailo L, Leatherdale ST. Effect of mailing nicotine patches on tobacco cessation among adult smokers: a randomized clinical trial. JAMA Intern Med 2016;176(2):184-190. [doi: 10.1001/jamainternmed.2015.7792] [Medline: 26809849]

Abbreviations

ACT: acceptance and commitment therapy
FDA: Food and Drug Administration
NCI: National Cancer Institute
NRT: nicotine replacement therapy
OR: odds ratio
RCT: randomized controlled trial
USCPG: US clinical practice guideline

XSL•FO
RenderX

Edited by A Mavragani; submitted 26.11.24; peer-reviewed by PS Olson; comments to author 24.12.24; revised version received 30.12.24; accepted 31.12.24; published 31.01.25.

Please cite as:
Santiago-Torres M, Mull KE, Sullivan BM, Cupertino AP, Salloum RG, Triplette M, Zvolensky MJ, Bricker JB
Evaluating the Impact of Pharmacotherapy in Augmenting Quit Rates Among Hispanic Adults in an App-Delivered Smoking Cessation Intervention: Secondary Analysis of a Randomized Controlled Trial
JMIR Form Res 2025;9:e69311
URL: https://formative.jmir.org/2025/1/e69311
doi:10.2196/69311
PMID:

©Margarita Santiago-Torres, Kristin E Mull, Brianna M Sullivan, Ana Paula Cupertino, Ramzi G Salloum, Matthew Triplette, Michael J Zvolensky, Jonathan B Bricker. Originally published in JMIR Formative Research (https://formative.jmir.org), 31.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effectiveness of Cognitive Behavioral Therapy Provided Through a Web Application for Subthreshold Depression, Subthreshold Insomnia, and Subthreshold Panic: Open-Labeled 6-Arm Randomized Clinical Trial Pilot Study

Kayoko Taguchi[1], PhD; Mirai Miyoshi[2]; Yoichi Seki[1], PhD; Shiori Baba[3]; Eiji Shimizu[2], PhD

[1]Research Center for Child Mental Development, Chiba University, Chiba, Japan

[2]Graduate School of Medicine, Chiba University, Chiba, Japan

[3]Chiba University Hospital, Chiba, Japan

Corresponding Author:
Kayoko Taguchi, PhD
Research Center for Child Mental Development
Chiba University
chuo-ku
Inohana 1-8-1
Chiba, 260-8670
Japan
Phone: 81 43 226 2027
Fax: 81 43 226 2028
Email: k.taguchi@chiba-u.jp

Abstract

Background: A common definition of "subthreshold" is that the diagnostic threshold is not met but the individuals are not asymptomatic. Some symptoms are present, causing significant difficulty in functioning and negatively impacting quality of life. Despite the attention given to subthreshold symptoms and the interventions for subthreshold symptoms being efficient in preventing the transition to psychiatric disease in primary care, reports on specific interventions are insufficient.

Objective: This study aimed to verify the effectiveness of internet-delivered cognitive behavioral therapy (ICBT) for subthreshold depression (SD), subthreshold insomnia (SI), and subthreshold panic (SP). Additionally, this study aimed to explore the minimally important change (MIC) of each subthreshold group's effectiveness outcome.

Methods: Participants aged 18-70 years from internet research monitors were categorized into SD, SI, and SP groups based on screening assessment. They were randomly assigned to intervention or control groups within each subthreshold symptom. The intervention groups worked on 4 weeks of nonguided ICBT ("Mentre"), while the control groups worked on a sham app. The primary outcome was the score change from screening (T1) to 4-week follow-up (T4) using the Center for Epidemiologic Studies Depression Scale (CESD) in the SD group, the Pittsburgh Sleep Quality Index (PSQI) in the SI group, and the Panic and Agoraphobia Scale (PAS) in the SP group. Secondary outcomes were score changes in the Generalized Anxiety Disorder-7 (GAD-7) scale, the Patient Health Questionnaire 9 (PHQ-9), the CESD, the PSQI, and the PAS, except the primary outcome in each group. Secondary outcomes were analyzed using complete-case analysis and repeated-measures ANOVA. Additionally, the MIC in the primary endpoint for each group was also calculated as an exploratory outcome.

Results: The SD, SP, and SI groups contained 846, 597, and 1106 participants, respectively. In the SD group, the difference in the CESD score change from baseline to follow-up between the intervention and control groups was significant (difference=0.52, 95% CI 1.29-4.66, $P<.001$). In the SI group, the difference in the PSQI score change was also significant (difference=0.53, 95% CI 0.11-0.94, $P=.01$). However, in the SP group, the difference in the PAS score change was not significant (difference=0.07, 95% CI –2.00 to 2.15, $P=.94$).

Conclusions: Our ICBT program Mentre contributes to the improvement of SI and SD. This suggests that nonguided ICBT may be effective in preventing SI and SD from progressing to the full threshold. However, appropriate definitions of subthreshold symptoms are necessary. In particular, it is difficult to define SP, and further research that considers the specific factors of each subthreshold symptom is necessary to accumulate evidence.

XSL•FO
RenderX

Trial Registration: University Hospital Medical Information Network (UMIN) UMIN000051280; https://tinyurl.com/2wyahhe3

(JMIR Form Res 2025;9:e63139) doi:10.2196/63139

KEYWORDS

minimally important change; nonguided cognitive behavioral therapy; subthreshold depression; subthreshold insomnia; subthreshold panic; cognitive behavioral therapy; CBT; psychiatric disease; primary care; interventions; depression; anxiety; insomnia; psychological therapy

Introduction

A common definition of "subthreshold" is that the diagnostic is not met but the individuals are not asymptomatic, that is, some symptoms are present, causing significant difficulty in functioning and negatively impacting the quality of life [1,2]. However, the concept of "subthreshold" symptoms remains unclear in the clinical setting. For instance, according to the latest scoping review of subthreshold depression (SD), there are many definitions based on the assessment tool cutoff point of the Center for Epidemiologic Studies Depression Scale (CESD), Beck's Depression Inventory II (BDI-II), and the Patient Health Questionnaire 9 (PHQ-9) and on the duration of symptoms. As a result, these definitions "seem to be arbitrary, with considerable overlap" [3]. Even the most popular subthreshold symptom is vague in its definition, and the other subthreshold symptoms, which are little mentioned, are even more so.

Although the definition is ambiguous, intervening for subthreshold symptoms is efficient in preventing the transition to psychiatric disease in primary care. Research has already suggested that shortening the duration of untreated illness (DUI) has a good therapeutic effect. The DUI is "measured as the interval between onset of the disorder and when the patient receives the first adequate treatment for that psychiatric disorder" [4], and many studies have claimed that the DUI affects treatment prognosis [5]. Furthermore, some studies have revealed the risk of transition from the subthreshold level to disease [6]. There is consensus that early intervention can be applied flexibly in different formats and across different target groups [7,8]. For instance, since SD and subthreshold anxiety occur twice as often as depression and anxiety [9], so appropriate definitions are, of course, essential for a concrete approach to primary care [10]. Furthermore, insomnia symptoms are closely associated with SD and subthreshold anxiety, and prevention programs for subthreshold insomnia (SI) have been reported to not only decrease depressive symptoms but also improve insomnia [11]. Namely, when considering primary care for subthreshold symptoms, overlapping symptoms also need to be considered; as a result, the treatment for specific subthreshold symptoms should also have the potential to cover a wide range of subthreshold symptoms. However, although there are some reports regarding SD interventions, there are insufficient studies on specific interventions in the subthreshold stage.

With limited reports, cognitive behavioral therapy (CBT) has been shown to be effective for subthreshold symptoms. CBT, which is a type of psychological therapy, improves symptoms by understanding patterns of maladaptive thinking and behavior and correcting the vicious cycle that perpetuates symptoms. It has been shown to be effective in improving diagnosed depression, anxiety, and insomnia [12-14]. However, nonguided internet-delivered cognitive behavioral therapy (ICBT) also has been shown to be effective for SD in a latest meta-analysis [15]. Since subthreshold symptoms are not severe, it is preferable that interventions be relatively short, easily accessible, and cost-effective. Nonguided CBT is more scalable and affordable [16,17], and it is suitable for people who have subthreshold symptoms but have no motivation to visit a hospital. However, there is insufficient evidence for ICBT as a subthreshold symptom treatment, and only 1 or 2 randomized controlled trials (RCTs) have verified that ICBT has valid effects on SD, SI, and subthreshold panic (SP). SI treatment verification over the internet has rarely been reported [18,19]. There are few reports on subthreshold anxiety, particularly reports on SP, which are mostly epidemiology and fact-finding surveys. Among the few papers on this topic, agreement has been reached concerning early intervention being valid and cost-effective for SP; however, they are not ICBTs [20,21].

Additionally, there are discussions of how patients perceive the significance of receiving ICBT for subthreshold symptoms. In recent years, interesting research results and suggestions concerning the minimally important change (MIC) have been reported [22]. The MIC is the smallest change or difference that patients perceive as significant. Researchers who reported on the MIC noted that the only satisfactory approach is for patients themselves to measure the health conditions they experience and know best. It is essential to discuss whether the MIC has implications for the significance of early intervention for subthreshold symptoms.

This study aimed to verify the effectiveness of the "Mentre" program as an ICBT for SD, SI, and SP in a 6-arm RCT. Additionally, this study aimed to explore the MIC of each subthreshold group's effectiveness outcome. These results will be useful as the first evidence accumulated about ICBT for subthreshold symptoms with stricter criteria and for knowledge of the MIC among patients with subthreshold symptoms.

Methods

Ethical Considerations

This study was conducted with the approval of the Institutional Review Board of Chiba University Hospital (approval number: G2022005). Additionally, the Clinical Research Ethics Review Committee oversaw proper implementation of the study at least once a year.

An explanatory document was presented to each web program site, and consent was obtained. Participants were informed that participation was voluntary, that they could decline at any time

XSL•FO
RenderX

for any reason, and that there would be no negative consequences for declining. Additionally, participants could withdraw their consent after providing informed consent.

For privacy, all participants' data were anonymized. Participants were granted points that would enable them to shop online as compensation for research participation. Each intervention group received 1000 points if the participants worked on the program and answered the outcome inventory completely, while the control groups received 800 points.

Trial Design and Procedure

The study was a prospective 6-arm, parallel-group RCT (Multimedia Appendix 1) [23]. The 6 arms were classified as SD, SI, and SP groups with 2 arms each. The trial was registered as "Effect of Mental Training Web Application for Subthreshold Insomnia, Subthreshold Depression and Subthreshold Panic Six Arms Randomized Controlled Trial" with the University Hospital Medical Information Network (UMIN; registration number: UMIN000051280). Participants were recruited from the internet research monitor of the NTT Com Online Marketing Solutions Corporation in May 2023. Participants who consented to the study completed a screening evaluation (T1) and a symptom persistence evaluation (T2) after 2 weeks. They were classified as having SD, SI, or SP and allocated to intervention or control groups. The intervention was a 4-week ICBT program Mentre, while the control program was a sham app, both of which started in June 2024 for each group. All participants were sent reminder emails every weekend to increase motivation and encourage continuous practice. After the program was complete, the participants were assessed for their symptoms postevaluation (T3), and 4 weeks later, follow-up evaluations were carried out (T4). Participants were awarded points for participating in the study and answering each time (T1-T4).

Participants

Participants were aged 18-70 years and did not visit a hospital for any disease. The exclusion and inclusion criteria for this study were as follows:

- Individuals who experienced little interest in things or had depressed thoughts almost every day for 2 weeks and met more than 5 items of depressive symptoms in the PHQ-9 were excluded as having depression. Of those who were not excluded, individuals with a CESD score of 16 or higher were eligible as those with SD.
- Individuals with SI who had PSQI scores greater than 6 and insomnia symptoms lasting less than 3 months were also eligible.
- Individuals with SP who had Panic and Agoraphobia Scale (PAS) scores greater than 9 and who experienced anxiety-related panic symptoms continuously lasting less than 1 month were also eligible. In the case of duplicate symptoms, if the panic score met the eligibility criteria, participants were allocated to the SP group preferentially because panic symptoms are more specific than depression and anxiety symptoms. SD sometimes overlaps with SI; hence, participants who showed SD and SI symptoms were allocated to the SD group.
- The exclusion criteria included individuals with a history of diagnosed depression, insomnia, or panic disorder and those who experienced difficulty or extreme difficulty in daily life owing to anxiety symptoms on the Generalized Anxiety Disorder-7 (GAD-7) scale or depressive thoughts on the PHQ-9.

Interventions

The intervention program Mentre was based on ICBT. We prepared 3 types of Mentre programs for each subthreshold symptom: SD, SI, and SP. Participants in the intervention groups undertook the relevant program for 4 weeks. The basic structure involved participants undergoing the first session at the beginning of the week and completing their homework throughout the remainder of the week. The control groups similarly recorded daily activities and weather conditions (sham app) in all subthreshold groups. All participants received guidance via email and accessed their program website through the URL provided on their mobile phones, personal computers, and tablets. The SD program included (1) cognitive restructuring about irrational thoughts, (2) psychoeducation for a depressive mood, (3) rumination distraction, and (4) behavioral activation to positive behavior. The SI program included (1) a sleep diary, (2) a review of one's sleeping behavior, (3) a review of one's sleep cognition, and (4) a review of one's sleep efficiency and sleep restriction methods. The SP program included (1) learning concerning panic disorder, (2) attention shift training for when facing a panic attack, (3) how to change the worst image of panic to a rational image, and (4) how to stop avoidant behavior. Each program involved paying attention to become aware of symptoms and to change irrational thoughts to rational thoughts and negative behavior to positive behavior. Participants could inquire about this study through each web program, and the researcher and manager of the web program responded accordingly. Reminder emails were sent to all participants at the end of each week to continue and not to forget working on the program.

Outcomes

Primary Endpoints

We set separate primary endpoints for each subthreshold symptom for the SD, SI, and SP groups: score change from screening (T1) to 4-week follow-up (T4) in each group:

- CESD in the SD group: The CESD is a brief self-report questionnaire for measuring the severity of depressive symptoms. It consists of 20 questions assessing numerous symptoms of depression experienced in the past week, with most items focusing on the emotional component of depression. Scores range from 0 to 60 points, with higher scores indicating more severe depressive symptoms. It has been used in many clinical trials involving a wide range of age groups [24,25].
- The Pittsburgh Sleep Quality Index (PSQI) in the SI group: The PSQI is a self-assessment consisting of 19 questions across 7 subscales (sleep quality, sleep latency, sleep duration, habitual sleep efficiency, sleep disturbances, hypnotic medication use, and daytime dysfunction). The total score indicates sleep quality and ranges between 0

XSL•FO
RenderX

(good sleep quality) and 21 (very poor sleep quality) points. The PSQI has been validated as a reliable and valid measure of subjective sleep quality in clinical practice and experimental studies [26,27].

- PAS in the SP group: The PAS is a 13-item questionnaire assessing panic attacks, composed of 5 factors (fear behavior, avoidance behavior, anticipatory anxiety, degree of disability, and health concerns) and evaluates the condition over the past week. Each item is scored from 0 to 4 points, with higher scores indicating more severe symptoms [28].

Secondary Endpoints

The secondary endpoints were the CESD, PSQI, and PAS score changes from screening (T1) to postevaluation (T3) and 4-week follow-up (T4) in all groups, as well as subsequent questionnaires. Additionally, the MIC in the primary endpoint for each group was also calculated.

- GAD-7 scale: It has been shown to have reliability, criterion, construct, factorial, and procedural validity. Cutoff points optimizing sensitivity (89%) and specificity (82%) were identified. The scale has 7 items assessing the severity of GAD in the past 2 weeks on a 4-point Likert scale (0=not at all, 1=1 episode, 2=on half or more days, and 3=almost daily). The minimum and maximum scores are 0 and 21, respectively (0-4, 5-9, 10-14, and 15-21 indicate no, mild, moderate, and severe symptoms, respectively). The cutoff score for clinically significant symptoms of anxiety is 10 [29].
- PHQ-9:It has diagnostic validity (for the diagnosis of 1 or more PHQ disorders; κ=0.65; overall accuracy=85,; sensitivity=75%, specificity=90%). It consists of 9 items scored on a 4-point Likert scale (0=not at all, 1=on several days, 2=on half or more days, and 3=almost daily). The minimum and maximum scores are 0 and 27, respectively (0-4, 5-9, 10-14, 15-19, and 20-27 indicate no, mild, moderate, moderate-to-severe, and severe symptoms, respectively). The PHQ cutoff score for clinically significant depressive symptoms is 10 [30].

Exploratory Endpoints

Minimally Important Change

The MIC is the minimum change threshold that the patient feels has improved or worsened and the change that is considered significant to the patient [20,29]. When patients evaluate their treatment effect using certain measures, it remains unclear how much change is meaningful. Hence, many studies suggest considering the MIC concept [32,33].

Sample Size

The sample size was calculated as follows: We needed to consider each subthreshold symptom, but there was insufficient evidence to estimate the effect size. We estimated the effect size of nonguided ICBT to be 0.2 based on a previous study [34]. Assuming an α level of .05 and a β level of .20, analysis was performed, resulting in a calculated sample size of 82. Based on the meta-analysis, more than 50% of the participants were expected to drop out. Hence, the sample size was set at 160 for each of the intervention and control groups.

Registration and Randomization

For case registration, participants who met the criteria were automatically registered after a preliminary screening on the server of an internet research company (NTT Com Online Marketing Solutions Corporation. This company assigned participants to the SI group, SD group, or SP group and further randomly assigned them to either the intervention or the control subgroup. The assigned control factors were CESD score≥23 or not for SD, PSQI score≥12 or not for SI, and PAS score≥22 or not for SP; sex was also assigned as a control factor for all groups. Participants were informed via email or other means about how to access the corresponding intervention program.

Protocol Deviations

Participants were provided the Mentre program through the web; therefore, dropouts and loss to follow-up were expected as main deviations. The number of not-evaluable cases was recorded. If any case reported a deviation due to an emergency crisis, the researcher immediately reported this to the Clinical Research Ethics Review Committee.

Statistical Methods

Main Analysis

Baseline variables were compared using the Fisher exact test and an unpaired 2-tailed *t* test for categorical and continuous variables, respectively. The significance level was set at .05. For the main analysis of treatment effects, the means of the least squares and their 95% CIs were estimated by repeated-measures ANOVA with each change in the CESD, PSQI, and PAS scores at the 4-week follow-up for each group. The model included intervention group, time, and intervention-by-time interaction as fixed effects. All comparisons were planned, and all *P* values were 2-sided. $P<.05$ was considered statistically significant. The primary outcome formed the basis of complete case analysis. No special complementary processing by statistical methods was performed for missing values. All statistical analyses were performed using R version 4.3.1 (R Foundation for Statistical Computing).

Secondary Analysis

To examine the MIC, we used an anchor. The patient-reported outcome (PRO) scores for each primary outcome were classified into the following anchor categories: “much better,” “slightly better,” “about the same,” “slightly worse,” and “much worse.” The average PRO score classified as “slightly better” was considered the MIC. We calculated the CESD, PSQI, and PAS scores of the MIC in this study.

Results

Participant Details

The flow of participants through the study is illustrated in Figure 1. This study was announced to 215,000 internet research monitors, and 25,418 (11.8%) monitors provided consent. Of them, 2549 (10%) participants with SD, SI, or SP symptoms were screened (T1) and assigned to 1 of 6 arms. Afterward,

XSL•FO
RenderX

2091 (82%) participants' symptoms continued after 2 weeks (T2). In total, 846 (33.2%) participants (n=424, 50.1%, intervention vs n=422, 49.9%, control participants) were classified as having SD, 1106 (43.4%) participants (n=552, 49.9%, intervention vs n=554, 50.1%, control participants) as having SI, and 597 (23.4%) participants (n=299, 50.1%, intervention vs n=298, 49.9%, control participants) as having SP according to screening evaluation scores. A total of 145 (34.2%) of 424 participants who were assigned to the SD group undertook the Mentre program, with 17 (11.7%) dropping out during the intervention. A total of 267 out of 552 participants who were assigned to the SI group undertook the Mentre program, with 74 dropping out during the intervention (27.7%). A total of 141 out of 299 participants who were assigned to the subthreshold panic group undertook the Mentre program, with 60 dropping out during the intervention (43.2%).

Figure 1. Flow diagram of the study participants.

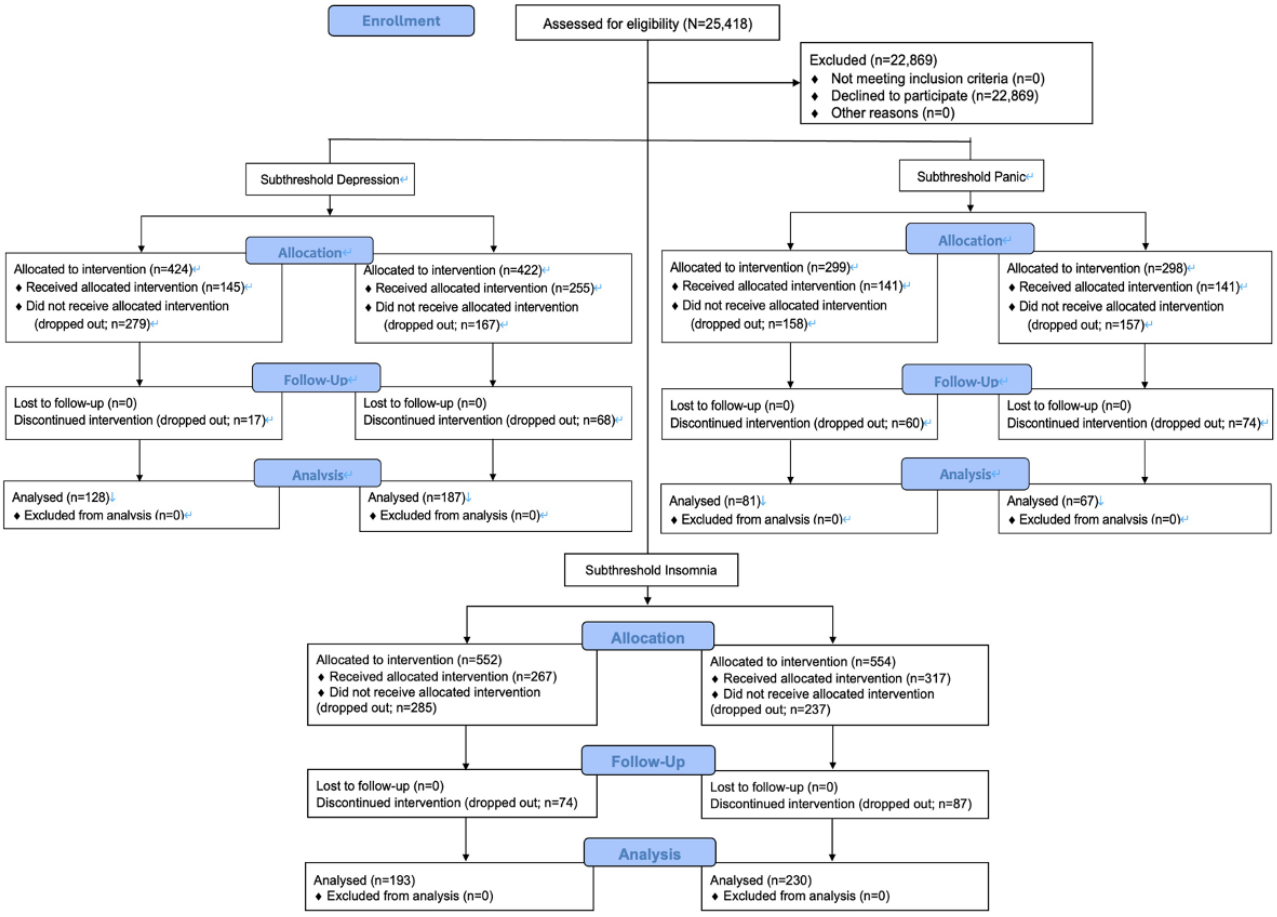

Table 1 presents the clinical characteristics of the patients in each group. In all 3 groups, no significant difference was found between the intervention and control subgroups in any average value, including sex, age, marital status, cohabitation rate, employment status, and income.

XSL•FO
RenderX

Table 1. Baseline demographic and clinical characteristics (N=2549).

Characteristics	SD[a]			SI[b]			SP[c]		
	Intervention (n=424)	Control (n=422)	*P* value	Intervention (n=552)	Control (n=554)	*P* value	Intervention (n=299)	Control (n=298)	*P* value
Sex (male), n (%)	130 (30.7)	130 (30.8)	.99	225 (40.7)	227 (41.0)	.951	106 (35.4)	107 (35.9)	.932
Age (years), mean (standard deviation)	41.3 (12.48)	42.5 (11.68)	.068	42.3 (11.74)	42.8 (12.10)	.546	41.2 (12.60)	42.1 (12.03)	.363
Marital status, n (%)									
Married	229 (54.0)	239 (56.6)	.448	318 (57.6)	331 (59.7)	.502	154 (51.5)	156 (52.3)	.870
Living with someone	322 (75.9)	317 (75.1)	.811	425 (77.0)	429 (77.4)	.886	236 (78.9)	229 (76.8)	.555
Education, mean (standard deviation)	14.8 (1.99)	14.7 (2.05)	.410	14.9 (1.95)	14.7 (1.92)	.059	14.6 (2.10)	14.3 (2.15)	.208
Employment, n (%)	284 (66.9)	291 (68.9)	.556	416 (75.3)	402 (72.6)	.304	195 (65.2)	186 (62.4)	.496
Yearly income (yen), n (%)	__[d]	—	.783	—	—	.320	—	—	.150
<1,000,000 (US $6322.53)[e]	131 (30.8)	119 (28.2)	—	126 (22.8)	143 (25.8)	—	105 (35.1)	112 (37.6)	—
1,000,001-5,000,000 (US $6322.54-$31,612.66)	203 (47.8)	214 (50.7)	—	284 (51.4)	268 (48.3)	—	137 (45.8)	135 (45.3)	—
5,000,001-10,000,000 (US $31,612.66-$63,225.31)	78 (18.4)	75 (17.7)	—	128 (23.2)	121 (21.8)	—	53 (17.7)	40 (13.4)	—
>10,000,000 (US $63,225.31)	12 (2.8)	14 (3.3)	—	14 (2.5)	22 (4)	—	4 (1.3)	11 (3.7)	—

[a]SD: subthreshold depression.

[b]SI: subthreshold insomnia.

[c]SP: subthreshold panic.

[d]Not applicable.

[e]An exchange rate of 1 yen=US $0.0063 has been applied.

Primary Outcome

Tables 2-4 and Figure 2 show the effects of the Mentre program as the primary outcome for each group. In the SD group, differences in the CESD score change from T1 to T4 between the intervention and control groups were significant (difference=0.52, 95% CI 1.29-4.66, *P*<.001). In the SI group, differences in the PSQI score change from T1 to T4 between the intervention and control groups were significant (difference=0.53, 95% CI 0.11-0.94, *P*=.01). In the SP group, differences in the PAS score change from T1 to T4 between the intervention and control groups were not significant (difference=0.07, 95% CI –2.00 to 2.15, *P*=.94).

XSL•FO
RenderX

Table 2. Primary outcome ANOVA results for the SD[a] group using the CESD[b].

Week[c]	Intervention (n=128), mean (standard deviation)	Control (n=187), mean (standard deviation)	Estimate (95% CI)	SE	*P* value	Cohen's d
T1	22.42 (9.03)	22.11 (8.48)	__[d]	—	—	—
T2	20.54 (9.70)	20.47 (8.97)	0.16 (–1.33 to 1.65)	0.76	.829	—
T3	18.41 (9.98)	20.18 (10.14)	2.01 (0.28 to 3.73)	0.88	.023[e]	0.18
T4	17.20 (9.33)	19.94 (10.05)	2.98 (1.29 to 4.66)	0.86	.001[f]	0.28

[a]SD: subthreshold depression.
[b]CESD: Center for Epidemiologic Studies Depression Scale.
[c]T1: screening; T2: symptom persistence evaluation; T3: postevaluation; T4: 4-week follow-up.
[d]Not applicable.
[e]$P<.05$.
[f]$P<.01$.

Table 3. Primary outcome ANOVA results for the SI[a] group using the PSQI[b].

Week[c]	Intervention (n=193), mean (standard deviation)	Control (n=230), mean (standard deviation)	Estimate (95% CI)	SE	*P* value	Cohen's d
T1	7.66 (1.63)	7.69 (1.87)	__[d]	—	—	—
T2	6.90 (2.26)	7.23 (2.49)	0.32 (–0.06 to 0.69)	0.19	.099	—
T3	6.48 (2.23)	7.00 (2.40)	0.50 (0.10 to 0.90)	0.20	.015[e]	0.22
T4	6.09 (2.26)	6.63 (2.42)	0.53 (0.11 to 0.94)	0.21	.013[e]	0.23

[a]SI: subthreshold insomnia.
[b]PSQI: Pittsburgh Sleep Quality Index.
[c]T1: screening; T2: symptom persistence evaluation; T3: postevaluation; T4: 4-week follow-up.
[d]Not applicable.
[e]$P<.05$.

Table 4. Primary outcome ANOVA results for the SP[a] group using the PAS[b].

Week[c]	Intervention (n=81), mean (standard deviation)	Control (n=67), mean (standard deviation)	Estimate (95% CI)	SE	*P* value	Cohen's d
T1	13.89 (4.73)	13.54 (4.40)	__[d]	—	—	—
T2	11.47 (6.68)	12.60 (7.32)	1.41 (–0.54 to 3.36)	0.99	.156	—
T3	12.30 (6.93)	10.43 (6.21)	–1.65 (–3.62 to 0.33)	1.00	.101	0.29
T4	10.69 (7.02)	10.52 (7.05)	0.07 (–2.00 to 2.15)	1.05	.945	0.02

[a]SP: subthreshold panic.
[b]PAS: Panic and Agoraphobia Scale.
[c]T1: screening; T2: symptom persistence evaluation; T3: postevaluation; T4: 4-week follow-up.
[d]Not applicable.

XSL•FO
RenderX

Figure 2. The effect of the Mentre program as the primary outcome for each group.

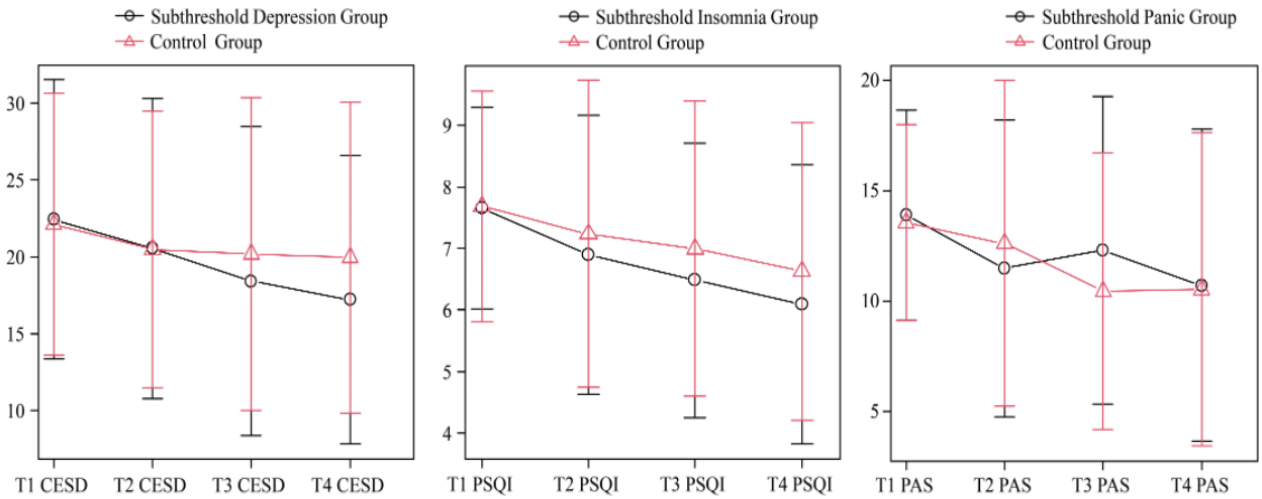

Secondary Outcomes

Tables 5-7 present the means of the secondary endpoints in each group and the differences between groups. In the SD group, differences in the mean PSQI and PAS scores at either time point were not significant; however, the difference in the PHQ-9 score at T4 was significant (difference=1.00, 95% CI 0.04-1.97, *P*=.04). Furthermore, at the time of postevaluation (T3), a significant difference was revealed in the PAS score (difference=2.05, 95% CI 0.02-4.08, *P*=.047). In the SI group, at the time of postevaluation (T3), a significant difference was found in the PSQI (difference=0.50, 95% CI 0.01-0.90, *P*=.02); however, between-group differences in changes in other endpoints were not significant at any time point. In the SP group, between-group differences in changes in all secondary endpoints were not significant.

XSL•FO
RenderX

Table 5. Secondary outcome ANOVA results for the SD[a] group.

Variables and weeks[b]	Intervention, mean (standard deviation)	Control group, mean (standard deviation)	SE (95% CI)	*P* value	Cohen's d
PSQI [c] (intervention: n=128; control: n=187)					
T1	6.01 (2.13)	6.47 (2.44)	__[d]	—	—
T3	5.83 (2.57)	6.53 (2.82)	0.26 (–0.12 to 0.89)	.13	0.26
T4	5.52 (2.48)	5.97 (3.13)	0.28 (–0.41 to 0.67)	.64	0.16
PAS [e] (intervention: n=38; control: n=48)					
T1	4.29 (3.70)	3.79 (2.21)	—	—	—
T3	4.24 (4.56)	5.94 (5.39)	1.02 (0.02 to 4.08)	.05	0.34
T4	4.29 (5.34)	4.60 (5.36)	1.09 (–1.50 to 2.84)	.54	0.06
PHQ-9 [f] (intervention: n=128; control: n=187)					
T1	7.81 (4.42)	7.80 (4.19)	—	—	—
T3	6.57 (5.18)	6.83 (5.23)	0.45 (–0.61 to 1.17)	.54	0.05
T4	5.69 (5.53)	6.68 (5.30)	0.49 (0.04 to 1.97)	.04[g]	0.18
GAD-7 [h] (intervention: n=128; control: n=187)					
T1	6.45 (4.26)	6.49 (4.23)	—	—	—
T3	5.09 (4.21)	5.42 (4.18)	0.39 (–0.46 to 1.07)	.43	0.08
T4	4.49 (4.14)	5.20 (4.22)	0.40 (–0.09 to 1.48)	.08	0.17

[a]SD: subthreshold depression.
[b]T1: screening; T3: postevaluation; T4: 4-week follow-up.
[c]PSQI: Pittsburgh Sleep Quality Index.
[d]Not applicable.
[e]PAS: Panic and Agoraphobia Scale.
[f]PHQ-9: Patient Health Questionnaire-9.
[g]$P<.05$.
[h]GAD-7: Generalized Anxiety Disorder-7.

XSL•FO
RenderX

Table 6. Secondary outcome ANOVA results for the SI[a] group.

Variables and weeks[b]	Intervention, mean (standard deviation)	Control group, mean (standard deviation)	SE (95% CI)	*P* value	Cohen's d
CESD [c] (intervention: n=193; control: n=230)					
T1	15.74 (7.74)	15.45 (7.18)	__[d]	—	—
T3	15.37 (7.99)	15.68 (8.46)	0.57 (–0.59 to 1.65)	.36	0.04
T4	14.49 (8.20)	15.02 (8.48)	0.60 (–0.44 to 1.92)	.22	0.06
PAS [e] (intervention: n=53; control: n=54)					
T1	3.91 (3.47)	3.43 (2.35)	—	—	—
T3	4.74 (4.72)	4.80 (4.59)	0.86 (–1.37 to 2.03)	.70	0.01
T4	4.26 (4.23)	3.83 (4.14)	0.78 (–1.75 to 1.34)	.79	0.10
PHQ-9 [f] (intervention: n=193; control: n=68)					
T1	3.98 (3.62)	5.01 (4.45)	—	—	—
T3	4.53 (4.27)	5.10 (4.85)	0.63 (–0.67 to 1.82)	.37	0.13
T4	4.13 (4.57)	4.19 (4.50)	0.65 (–1.21 to 1.35)	.91	0.01
GAD-7 [g] (intervention: n=193; control: n=230)					
T1	4.36 (4.13)	3.73 (3.46)	—	—	—
T3	3.54 (3.94)	3.54 (3.88)	0.31 (–0.22 to 0.99)	.22	0.00
T4	3.16 (3.68)	2.95 (3.68)	0.30 (–0.45 to 0.72)	.65	0.06

[a]SI: subthreshold insomnia.

[b]T1: screening; T3: postevaluation; T4: 4-week follow-up.

[c]CESD: Center for Epidemiologic Studies Depression Scale.

[d]Not applicable.

[e]PAS: Panic and Agoraphobia Scale.

[f]PHQ-9: Patient Health Questionnaire-9.

[g]GAD-7: Generalized Anxiety Disorder-7.

XSL•FO
RenderX

Table 7. Secondary outcome ANOVA results for the SP[a] group.

Variables and weeks[b]	Intervention, mean (standard deviation)	Control group, mean (standard deviation)	SE (95% CI)	*P* value	Cohen's d
PSQI [c] (intervention: n=99; control: n=98)					
T1	7.30 (2.76)	7.01 (2.79)	__[d]	—	—
T3	7.27 (2.98)	6.91 (2.93)	0.28 (–0.68 to 0.42)	.64	0.12
T4	6.80 (2.88)	6.73 (2.93)	0.28 (–0.39 to 0.72)	.56	0.02
CESD [e] (intervention: n=99; control: n=98)					
T1	24.68 (10.23)	21.80 (10.26)	—	—	—
T3	23.22 (9.53)	20.53 (9.21)	1.00 (–2.90 to 1.07)	.36	0.29
T4	20.33 (8.98)	20.51 (9.48)	1.02 (–0.16 to 3.87)	.07	0.02
PHQ-9 [f] (intervention: n=99; control: n=98)					
T1	6.04 (4.96)	4.92 (4.70)	—	—	—
T3	8.68 (5.83)	7.68 (5.71)	0.83 (–2.70 to 0.59)	.21	0.17
T4	8.01 (5.47)	7.30 (5.31)	0.78 (–2.35 to 0.72)	.30	0.13
GAD-7 [g] (intervention: n=99; control: n=98)					
T1	8.55 (4.64)	7.44 (3.96)	—	—	—
T3	7.26 (4.87)	6.28 (4.08)	0.52 (–1.32 to 0.73)	.57	0.22
T4	5.95 (4.23)	5.76 (3.98)	0.48 (–0.52 to 1.37)	.38	0.55

[a]SP: subthreshold panic.
[b]T1: screening; T3: postevaluation; T4: 4-week follow-up.
[c]PSQI: Pittsburgh Sleep Quality Index.
[d]Not applicable.
[e]CESD: Center for Epidemiologic Studies Depression Scale.
[f]PHQ-9: Patient Health Questionnaire-9.
[g]GAD-7: Generalized Anxiety Disorder-7.

Calculation of the MIC

Table 8 presents the MICs of the CESD, PSQI, and PAS. The mean MIC of the CESD for participants who undertook the Mentre program for SD was –5.23 (standard deviation 8.89). The mean MIC of the PSQI for participants who undertook the Mentre program for SI was –3.60 (standard deviation 6.34). The mean MIC of the PAS for participants who undertook the Mentre program for SP was –2.08 (standard deviation 2.44).

XSL•FO
RenderX

Table 8. MIC[a] of each primary endpoint.

CESD[b] in the SD[c] group			PSQI[d] in the SI[e] group		PAS[f] in the SP[g] group	
MIC	Participants, n (%)	Mean (standard deviation)	Participants, n (%)	Mean (standard deviation)	Participants, n (%)	Mean (standard deviation)
Better	23 (16.0)	–9.78 (6.381)	18 (8.4)	–3.78 (2.184)	11 (11.8)	–7.91 (5.991)
Slightly better	48 (33.3)	–5.23 (8.890)	71 (33.2)	–2.08 (2.436)	45 (48.4)	–3.60 (6.340)
About the same	66 (45.8)	–3.61 (7.831)	113 (52.8)	–1.1 (1.911)	33 (35.5)	–1.79 (5.732)
Slightly worse	4 (2.8)	–1.00 (1.414)	11 (5.1)	0.91 (2.548)	4 (4.3)	–3.00 (4.761)
Worse	3 (2.1)	8.00 (14.731)	1 (0.5)	0.00	0	__[h]
Total	144 (100.0)	–4.82 (8.470)	214 (100.0)	–1.59 (2.338)	93 (100.0)	–3.44 (6.220)

[a]MIC: minimally important change.
[b]CESD: Center for Epidemiologic Studies Depression Scale.
[c]SD: subthreshold depression.
[d]PSQI: Pittsburgh Sleep Quality Index.
[e]SI: subthreshold insomnia.
[f]PAS: Panic and Agoraphobia Scale.
[g]SP: subthreshold panic.
[h]Not applicable.

Adverse Events

None of the participants in either group experienced serious adverse events during the time of this study.

Discussion

Principal Findings

This study aimed to assess the effectiveness of the Mentre program, a 4-week ICBT, for SD, SI, and SP. We categorized participants into SD, SI, and SP groups based on their eligibility using previous research and conducted an RCT for each group. We verified differences in the score change from screening (T1) to 4-week follow-up (T4) as indicators of ICBT effectiveness. There was a significant difference between the SD and SI intervention and control groups but not in the SP groups. Additionally, significant differences in CESD score changes were observed from screening (T1) to postintervention (T3) in the SD intervention and control groups. The SI intervention and control groups also showed significant PSQI score changes from T1 to T3. Previous studies have yielded mixed results on ICBT for SD, indicating its effectiveness compared to usual treatment, with effectiveness maintained for more than 6 months [6,35]. Limited research on SI has also suggested the effectiveness of ICBT in improving insomnia symptoms [18]. Our findings of SD and SI are comparable to those of previous studies. However, there was no significant difference in PAS score changes in the SP intervention and control groups. One reason was the possibility that the eligibility of SP was not appropriate. Although no studies have yet properly defined and intervened on SP [36-40], a previous study [36] suggests that if agoraphobia is present, we should carefully attend to it even if it does not completely meet the panic disorder criteria. In another study [40], a participant was classified according to the severity based on the frequency of panic attacks. Our study participants were considered to have SP if they had no panic attacks and the duration of symptoms was less than 1 month but they scored more than 9 points on the PAS. Therefore, the Mentre program for SP did not work, because it focused on anticipatory anxiety or dealt with avoidance of a particular place.

At least in primary care, our Mentre app found that self-understanding depressive moods and thoughts and trying to change them into rational behavior are effective in SD and that recording and reviewing sleep are effective in SI. However, further research is needed for the definition of SP; in particular, it is necessary to verify whether ICBT programs, such as Mentre, are effective for SP accompanied by agoraphobia and panic attacks.

Regarding subthreshold criteria, it is not just SP that requires discussion. Previous studies [41,42] have emphasized the need to define the score indicating a healthy state for the definition of subthreshold symptoms. In our study, we considered the persistence of each subthreshold symptom for 2 weeks (T2), resulting in 458 participants no longer meeting the subthreshold criteria. Can we conclude that the 458 participants did not have subthreshold symptoms? It is worth considering whether the absence of symptoms for 2 weeks defines health. These patients were excluded from our study; however, we should consider the timing of the intervention for subthreshold symptoms, including medical histories, event factors, and the frequency of panic attacks or depressive episodes, in the definition of subthreshold symptoms in the future.

Given that the definition of the subthreshold state is ambiguous, it is difficult to estimate the appearance rate of subthreshold symptoms. In our study, we estimated 160 participants in each group as the target sample size, but in fact, there was considerable discrepancy. This may indicate that there are many more individuals with subthreshold symptoms than we predicted, possibly because there is substantial overlap across subthreshold states. In contrast, we also need to consider that the participants

XSL•FO
RenderX

earned reward points by registering over the internet, which may relate to the high dropout rate, and the participants' motivation to test the Mentre program should also be considered. Individuals with subthreshold symptoms may lack high motivation for treatment compared to diagnosed patients, because their symptoms are of a low level and do not impair everyday function. The Mentre program operates at the same level used for patients diagnosed with depression, insomnia, and panic disorder. Hence, daily participation may have been bothersome for the participants.

Future studies should be designed considering the concept of subthreshold symptoms and possible overlap across subthreshold conditions and having a structure that is manageable and appropriate for this particular patient population.

An essential consideration is the overlap across subthreshold symptoms. In this study, we verified all 3 subthreshold symptoms together, given the frequent comorbidity of depression and anxiety disorders, often with insomnia [43,44]. However, few previous studies have explored this overlap. A previous study [45] on universal prevention treatment for depression suggested that adolescents at low risk for depression benefit equally from preventive interventions as those at high risk; however, it did not specify a matching treatment for low risk [45]. In other words, patients with subthreshold symptoms may benefit from treatment, but the treatment is not specified, and a completely new approach may be necessary [46]. We hypothesize that if there is overlap across subthreshold symptoms, interventions targeting specific symptoms might also affect overlapping symptoms owing to their low severity. Unfortunately, our program only showed efficacy for the symptoms we pretargeted, but previous meta-analyses have shown that improvement in diagnosed insomnia is related to improved depressive symptoms [14]. Therefore, treatments focusing on identifying components of subthreshold symptoms and supplementing CBT techniques may be effective.

Additionally, our study provided evidence of the MIC for SD, SI, and SP, which is specific to the subthreshold condition. Despite being useful as an index for clinicians and patients, caution is necessary in generalizing to individuals, because the MIC provides a statistical measure of association and may not reflect clinical conditions [47]. Particularly, subthreshold symptoms were originally low in severity; hence, the MIC was also low. For the indicated MIC to be an important change, it is necessary to also examine the events and satisfaction levels that changed because of the MIC for patients who reported that they felt "slightly better."

Limitations

Several limitations of this study should be acknowledged. First, the participants were limited to internet research monitors; hence, it is possible that our sample may not have been representative of all individuals with subthreshold symptoms. Approximately 30% of the participants were not employed and had a low income, which also does not reflect the entire population. Second, the main analysis was a complete case analysis owing to concerns regarding the high dropout rate among patients with subthreshold symptoms. In this study, we needed to verify the effectiveness of the Mentre program for specialized subthreshold symptoms as a pilot study. In the future, more accurate efficacy verification research is needed by conducting RCTs on the general population and analyzing them using intent-to-treat methods.

Conclusion

Our study suggests that the use of the 4-week Mentre program as a nonguided ICBT is associated with improvements in symptoms of SD and SI in primary care. However, for SP, there is a need to first define it, and next, it is necessary to re-examine the effectiveness of ICBT. Additionally, although the MIC of subthreshold symptoms may be useful in clinical situations, individual adaptation needs to be approached with caution.

Acknowledgments

We would like to thank Editage for English language editing and all the researchers and staff who contributed to this study, as well as all the study participants. This work was supported by the Adaptable and Seamless Technology Transfer Program through Target-driven R&D (A-STEP) from the Japan Science and Technology Agency (JST; grant number: JPMJTR221A).

Data Availability

Raw data were generated at the research center of child mental developmental. Derived data supporting the findings of this study are available from the corresponding author (KT) upon request.

Authors' Contributions

KT and ES conceived and designed the study. KT, ES, and YS designed the Mentre program. KT calculated the sample size and developed an analysis plan. KT, MM. and SB wrote the first draft of the manuscript, and all other authors revised the manuscript critically. This manuscript was not used generative artificial intelligence in any part. All authors have approved the final version of the manuscript.

Conflicts of Interest

NTT Com Online Marketing Solutions Corporation was involved in this study, as mentioned in the manuscript. This company was in charge of recruitment, assignment randomization, data collection, and sending reminder emails to participants. The Mentre

XSL•FO
RenderX

program was developed by the authors, the subthreshold depression and subthreshold insomnia program was written using the programming code of JUNPUZI Co, Ltd, and the subthreshold panic program was written by NTT Communications.

Multimedia Appendix 1
CONSORT (Consolidated Standards of Reporting Trials) checklist.
[PDF File (Adobe PDF File), 131 KB - formative_v9i1e63139_app1.pdf]

References

1. Lipowski ZJ. Delirium (acute confusional states). JAMA 1987 Oct 02;258(13):1789-1792 [FREE Full text] [doi: 10.1001/jama.1987.03400130103041]
2. Rodríguez MR, Nuevo R, Chatterji S, Ayuso-Mateos JL. Definitions and factors associated with subthreshold depressive conditions: a systematic review. BMC Psychiatry 2012 Oct 30;12(1):181. [doi: 10.1186/1471-244x-12-181]
3. Volz H, Stirnweiß J, Kasper S, Möller HJ, Seifritz E. Subthreshold depression - concept, operationalisation and epidemiological data. A scoping review. Int J Psychiatry Clin Pract 2023 Mar 23;27(1):92-106 [FREE Full text] [doi: 10.1080/13651501.2022.2087530] [Medline: 35736807]
4. Albert U, Barbaro F, Bramante S, Rosso G, De Ronchi D, Maina G. Duration of untreated illness and response to SRI treatment in obsessive-compulsive disorder. Eur Psychiatry 2019 May 11;58:19-26 [FREE Full text] [doi: 10.1016/j.eurpsy.2019.01.017] [Medline: 30763828]
5. Vigne P, Fortes P, Dias RV, Laurito LD, Loureiro CP, de Menezes GB, et al. Duration of untreated illness in a cross-diagnostic sample of obsessive-compulsive disorder, panic disorder, and social anxiety disorder. CNS Spectr 2018 Nov 13;24(5):526-532. [doi: 10.1017/s1092852918001281]
6. Francesconi M, Minichino A, Carrión RE, Delle Chiaie R, Bevilacqua A, Parisi M, et al. Psychosis prediction in secondary mental health services. A broad, comprehensive approach to the "at risk mental state" syndrome. Eur Psychiatry 2017 Feb 16;40:96-104. [doi: 10.1016/j.eurpsy.2016.09.002] [Medline: 27992839]
7. Takagaki K, Okamoto Y, Jinnin R, Mori A, Nishiyama Y, Yamamura T, et al. Behavioral activation for late adolescents with subthreshold depression: a randomized controlled trial. Eur Child Adolesc Psychiatry 2016 Nov;25(11):1171-1182. [doi: 10.1007/s00787-016-0842-5] [Medline: 27003390]
8. Cuijpers P, Quero S, Dowrick C, Arroll B. Psychological treatment of depression in primary care: recent developments. Curr Psychiatry Rep 2019 Nov 23;21(12):129 [FREE Full text] [doi: 10.1007/s11920-019-1117-x] [Medline: 31760505]
9. Helmchen H. [Subthreshold psychological disorders]. Nervenarzt 2001 Mar 26;72(3):181-189. [doi: 10.1007/s001150050737] [Medline: 11268762]
10. Hanel G, Henningsen P, Herzog W, Sauer N, Schaefert R, Szecsenyi J, et al. Depression, anxiety, and somatoform disorders: vague or distinct categories in primary care? Results from a large cross-sectional study. J Psychosom Res 2009 Sep;67(3):189-197. [doi: 10.1016/j.jpsychores.2009.04.013] [Medline: 19686874]
11. Chan N, Li S, Zhang J, Lam SP, Kwok APL, Yu MWM, et al. A prevention program for insomnia in at-risk adolescents: a randomized controlled study. Pediatrics 2021 Mar;e2020006833(3):147. [doi: 10.1542/peds.2020-006833] [Medline: 33627370]
12. López-López JA, Davies SR, Caldwell DM, Churchill R, Peters TJ, Tallon D, et al. The process and delivery of CBT for depression in adults: a systematic review and network meta-analysis. Psychol Med 2019 Jun 10;49(12):1937-1947. [doi: 10.1017/s003329171900120x]
13. Papola D, Ostuzzi G, Tedeschi F, Gastaldon C, Purgato M, Del Giovane C, et al. CBT treatment delivery formats for panic disorder: a systematic review and network meta-analysis of randomised controlled trials. Psychol Med 2022 Dec 09;53(3):614-624. [doi: 10.1017/s0033291722003683]
14. Hertenstein E, Trinca E, Wunderlin M, Schneider CL, Züst MA, Fehér KD, et al. Cognitive behavioral therapy for insomnia in patients with mental disorders and comorbid insomnia: a systematic review and meta-analysis. Sleep Med Rev 2022 Apr;62:101597 [FREE Full text] [doi: 10.1016/j.smrv.2022.101597] [Medline: 35240417]
15. Karyotaki E, Efthimiou O, Miguel C, Bermpohl FMG, Furukawa TA, Cuijpers P, Individual Patient Data Meta-Analyses for Depression (IPDMA-DE) Collaboration, et al. Internet-based cognitive behavioral therapy for depression: a systematic review and individual patient data network meta-analysis. JAMA Psychiatry 2021 Apr 01;78(4):361-371 [FREE Full text] [doi: 10.1001/jamapsychiatry.2020.4364] [Medline: 33471111]
16. Karyotaki E, Riper H, Twisk J, Hoogendoorn A, Kleiboer A, Mira A, et al. Efficacy of self-guided internet-based cognitive behavioral therapy in the treatment of depressive symptoms: a meta-analysis of individual participant data. JAMA Psychiatry 2017 Apr 01;74(4):351-359 [FREE Full text] [doi: 10.1001/jamapsychiatry.2017.0044] [Medline: 28241179]
17. Fairburn CG, Patel V. The impact of digital technology on psychological treatments and their dissemination. Behav Res Ther 2017 Jan;88:19-25 [FREE Full text] [doi: 10.1016/j.brat.2016.08.012] [Medline: 28110672]
18. Denis D, Eley TC, Rijsdijk F, Zavos HM, Keers R, Espie CA, et al. Is digital cognitive behavioural therapy for insomnia effective in treating sub-threshold insomnia: a pilot RCT. Sleep Med 2020 Feb;66:174-183. [doi: 10.1016/j.sleep.2019.10.007] [Medline: 31901759]

XSL•FO
RenderX

19. Theppornpitak W, Hemrungrojn S, Thienwiwatnukul K, Muntham D, Chirakalwasan N, Srisawart P. Effectiveness of internet-based CBT-I for the treatment of chronic subthreshold to moderate insomnia. Front Neurol 2023 Jun 2;14:1180339 [FREE Full text] [doi: 10.3389/fneur.2023.1180339] [Medline: 37346166]
20. Meulenbeek P, Spinhoven P, Smit F, van Balkom A, Cuijpers P. Cognitive mediation of panic reduction during an early intervention for panic. Acta Psychiatr Scand 2010 Jul;122(1):20-29. [doi: 10.1111/j.1600-0447.2009.01530.x] [Medline: 20070849]
21. Ophuis RH, Lokkerbol J, Hiligsmann M, Evers SMAA. Early intervention for subthreshold panic disorder in the Netherlands: a model-based economic evaluation from a societal perspective. PLoS One 2018 Feb 21;13(2):e0193338 [FREE Full text] [doi: 10.1371/journal.pone.0193338] [Medline: 29466470]
22. Wang Y, Devji T, Carrasco-Labra A, King MT, Terluin B, Terwee CB, et al. A step-by-step approach for selecting an optimal minimal important difference. BMJ 2023 May 26;381:e073822. [doi: 10.1136/bmj-2022-073822] [Medline: 37236647]
23. Schulz KF, Altman DG, Moher D, CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMJ 2010 Mar 23;340:c332 [FREE Full text] [doi: 10.1136/bmj.c332] [Medline: 20332509]
24. Lewinsohn PM, Seeley JR, Roberts RE, Allen NB. Center for Epidemiologic Studies Depression Scale (CES-D) as a screening instrument for depression among community-residing older adults. Psychol Aging 1997;12(2):277-287. [doi: 10.1037/0882-7974.12.2.277]
25. Kim J, Chung H, Askew RL, Park R, Jones SMW, Cook KF, et al. Translating CESD-20 and PHQ-9 scores to PROMIS depression. Assessment 2017 Apr 28;24(3):300-307. [doi: 10.1177/1073191115607042] [Medline: 26423348]
26. Sato D, Sutoh C, Seki Y, Nagai E, Shimizu E. Treatment preferences for internet-based cognitive behavioral therapy for insomnia in Japan: online survey. JMIR Form Res 2019 May 15;3(2):e12635 [FREE Full text] [doi: 10.2196/12635] [Medline: 31094319]
27. Carpenter JS, Andrykowski MA. Psychometric evaluation of the Pittsburgh Sleep Quality Index. J Psychosom Res 1998 Jul;45(1):5-13. [doi: 10.1016/S0022-3999(97)00298-5]
28. Bandelow B. Assessing the efficacy of treatments for panic disorder and agoraphobia. II. The Panic and Agoraphobia Scale. Int Clin Psychopharmacol 1995 Jun;10(2):73-81. [doi: 10.1097/00004850-199506000-00003] [Medline: 7673659]
29. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
30. Spitzer RL, Kroenke K, Williams JB. Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. Primary Care Evaluation of Mental Disorders. Patient Health Questionnaire. JAMA 1999 Nov 10;282(18):1737-1744. [doi: 10.1001/jama.282.18.1737] [Medline: 10568646]
31. Terwee CB, Peipert JD, Chapman R, Lai J, Terluin B, Cella D, et al. Minimal important change (MIC): a conceptual clarification and systematic review of MIC estimates of PROMIS measures. Qual Life Res 2021 Oct;30(10):2729-2754 [FREE Full text] [doi: 10.1007/s11136-021-02925-y] [Medline: 34247326]
32. Angst F, Aeschlimann A, Angst J. The minimal clinically important difference raised the significance of outcome effects above the statistical level, with methodological implications for future studies. J Clin Epidemiol 2017 Feb;82:128-136 [FREE Full text] [doi: 10.1016/j.jclinepi.2016.11.016] [Medline: 27986610]
33. Dawson J, Boller I, Doll H, Lavis G, Sharp R, Cooke P, et al. Minimally important change was estimated for the Manchester-Oxford Foot Questionnaire after foot/ankle surgery. J Clin Epidemiol 2014 Jun;67(6):697-705. [doi: 10.1016/j.jclinepi.2014.01.003] [Medline: 24709030]
34. Imamura K, Kawakami N, Furukawa TA, Matsuyama Y, Shimazu A, Umanodan R, et al. Effects of an internet-based cognitive behavioral therapy intervention on improving work engagement and other work-related outcomes: an analysis of secondary outcomes of a randomized controlled trial. J Occup Environ Med 2015 May;57(5):578-584. [doi: 10.1097/JOM.0000000000000411] [Medline: 25749132]
35. He R, Wei J, Huang K, Yang H, Chen Y, Liu Z, et al. Nonpharmacological interventions for subthreshold depression in adults: a systematic review and network meta-analysis. Psychiatry Res 2022 Nov;317:114897. [doi: 10.1016/j.psychres.2022.114897] [Medline: 36242840]
36. Pané-Farré CA, Fenske K, Stender JP, Meyer C, John U, Rumpf H, et al. Sub-threshold panic attacks and agoraphobic avoidance increase comorbidity of mental disorders: results from an adult general population sample. J Anxiety Disord 2013 Jun;27(5):485-493. [doi: 10.1016/j.janxdis.2013.06.008] [Medline: 23911432]
37. Politis S, Bellos S, Hadjulis M, Gournellis R, Petrikis P, Ploumpidis D, et al. Epidemiology of panic disorder and subthreshold panic symptoms in the Greek general population. Psychiatriki 2020 Oct 1;31(3):201-215 [FREE Full text] [doi: 10.22365/jpsych.2020.313.201] [Medline: 33099461]
38. Oral E, Aydin N, Gulec M, Oral M. Panic disorder and subthreshold panic in the light of comorbidity: a follow-up study. Compr Psychiatry 2012 Oct;53(7):988-994. [doi: 10.1016/j.comppsych.2012.01.009] [Medline: 22425529]
39. Skapinakis P, Lewis G, Davies S, Brugha T, Prince M, Singleton N. Panic disorder and subthreshold panic in the UK general population: epidemiology, comorbidity and functional limitation. Eur Psychiatry 2011 Sep;26(6):354-362. [doi: 10.1016/j.eurpsy.2010.06.004] [Medline: 20813508]

XSL•FO
RenderX

40. Bystritsky A, Kerwin L, Niv N, Natoli JL, Abrahami N, Klap R, et al. Clinical and subthreshold panic disorder. Depress Anxiety 2010 Apr;27(4):381-389 [FREE Full text] [doi: 10.1002/da.20622] [Medline: 20013957]
41. Schotte K, Cooper B. Subthreshold affective disorders: a useful concept in psychiatric epidemiology? Epidemiol Psichiatr Soc 1999 Oct 11;8(4):255-261. [doi: 10.1017/s1121189x00008162] [Medline: 10709316]
42. Helmchen H, Linden M. Subthreshold disorders in psychiatry: clinical reality, methodological artifact, and the double-threshold problem. Compr Psychiatry 2000 Mar;41(2 Suppl 1):1-7. [doi: 10.1016/s0010-440x(00)80001-2] [Medline: 10746897]
43. Hirschfeld RMA. The comorbidity of major depression and anxiety disorders: recognition and management in primary care. Prim Care Companion J Clin Psychiatry 2001 Dec 01;3(6):244-254 [FREE Full text] [doi: 10.4088/pcc.v03n0609] [Medline: 15014592]
44. Choi K, Kim Y, Jeon H. Comorbid anxiety and depression: clinical and conceptual consideration and transdiagnostic treatment. Adv Exp Med Biol 2020;1191:219-235. [doi: 10.1007/978-981-32-9705-0_14] [Medline: 32002932]
45. Cardemil EV, Reivich KJ, Seligman MEP. The prevention of depressive symptoms in low-income minority middle school students. Prev Treat 2002;5(1):8. [doi: 10.1037//1522-3736.5.1.58a]
46. McLaughlin KA. The public health impact of major depression: a call for interdisciplinary prevention efforts. Prev Sci 2011 Dec 6;12(4):361-371 [FREE Full text] [doi: 10.1007/s11121-011-0231-8] [Medline: 21732121]
47. Rennard SI. Minimal clinically important difference, clinical perspective: an opinion. COPD 2005 Mar 24;2(1):51-55 [FREE Full text] [doi: 10.1081/copd-200050641] [Medline: 17136962]

Abbreviations

CBT: cognitive behavioral therapy
CESD: Center for Epidemiologic Studies Depression Scale
DUI: duration of untreated illness
GAD-7: Generalized Anxiety Disorder-7
ICBT: internet-delivered cognitive behavioral therapy
MIC: minimally important difference
PAS: Panic and Agoraphobia Scale
PHQ-9: Patient Health Questionnaire 9
PRO: patient-reported outcome
PSQI: Pittsburgh Sleep Quality Index
RCT: randomized controlled trial
SD: subthreshold depression
SI: subthreshold insomnia
SP: subthreshold panic

Edited by A Mavragani; submitted 17.06.24; peer-reviewed by L Yang, R Nooripour; comments to author 04.11.24; revised version received 11.12.24; accepted 19.12.24; published 03.02.25.

Please cite as:
Taguchi K, Miyoshi M, Seki Y, Baba S, Shimizu E
Effectiveness of Cognitive Behavioral Therapy Provided Through a Web Application for Subthreshold Depression, Subthreshold Insomnia, and Subthreshold Panic: Open-Labeled 6-Arm Randomized Clinical Trial Pilot Study
JMIR Form Res 2025;9:e63139
URL: https://formative.jmir.org/2025/1/e63139
doi:10.2196/63139
PMID:39899369

©Kayoko Taguchi, Mirai Miyoshi, Yoichi Seki, Shiori Baba, Eiji Shimizu. Originally published in JMIR Formative Research (https://formative.jmir.org), 03.02.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effectiveness of Electronic Quality Improvement Activities to Reduce Cardiovascular Disease Risk in People With Chronic Kidney Disease in General Practice: Cluster Randomized Trial With Active Control

Jo-Anne Manski-Nankervis[1,2,3], BSc (Hons), MBBS (Hons), PhD; Barbara Hunter[3], BA (Hons), PhD; Natalie Lumsden[3,4], BSc (Hon), PhD; Adrian Laughlin[3], BA (Hons), M (BioS); Rita McMorrow[2,3,5], MBBS, PhD; Douglas Boyle[2,3], BSc (Hons), PhD; Patty Chondros[3], BSc (Hons), MSc, PhD; Shilpanjali Jesudason[6], MBBS, PhD; Jan Radford[7], MBBS, MEd, MPsychMed; Megan Prictor[8], BA, MMus, PhD; Jon Emery[3], BA, MA, MBBS, PhD; Paul Amores[9], BEc (Hons); An Tran-Duy[9,10], MSc, PhD; Craig Nelson[4,11,12], MBBS, PhD

[1]Primary Care and Family Medicine, Lee Kong Chian School of Medicine, Singapore, Singapore
[2]Centre for Research Excellence in Interactive Digital Technology to Transform Australia's Chronic Disease Outcomes, Prahan, Australia
[3]Department of General Practice and Primary Care, University of Melbourne, Melbourne, Australia
[4]Western Health Chronic Disease Alliance, Western Health, Sunshine, Australia
[5]Department of General Practice, University College Cork, Cork, Ireland
[6]Central Northern Adelaide Renal and Transplantation Service, Royal Adelaide Hospital, University of Adelaide, Adelaide, Australia
[7]Launceston Clinical School, University of Tasmania, Launceston, Australia
[8]Melbourne Law School, University of Melbourne, Melbourne, Australia
[9]Centre for Health Policy, Melbourne School of Population and Global Health, University of Melbourne, Melbourne, Australia
[10]Australian Centre for Accelerating Diabetes, University of Melbourne, Melbourne, Australia
[11]Department of Medicine, University of Melbourne, Sunshine, Australia
[12]Department of Nephrology, Western Health, Sunshine, Australia

Corresponding Author:
Jo-Anne Manski-Nankervis, BSc (Hons), MBBS (Hons), PhD
Primary Care and Family Medicine
Lee Kong Chian School of Medicine
11 Mandalay Road
Singapore, 308232
Singapore
Phone: 65 65138572
Email: joanne.mn@ntu.edu.sg

Abstract

Background: Future Health Today (FHT) is a program integrated with electronic medical record (EMR) systems in general practice and comprises (1) a practice dashboard to identify people at risk of, or with, chronic disease who may benefit from intervention; (2) active clinical decision support (CDS) at the point of care; and (3) quality improvement activities. One module within FHT aims to facilitate cardiovascular disease (CVD) risk reduction in people with chronic kidney disease (CKD) through the recommendation of angiotensin-converting enzyme inhibitor inhibitors (ACEI), angiotensin receptor blockers (ARB), or statins according to Australian guidelines (defined as appropriate pharmacological therapy).

Objective: This study aimed to determine if the FHT program increases the proportion of general practice patients with CKD receiving appropriate pharmacological therapy (statins alone, ACEI or ARB alone, or both) to reduce CVD risk at 12 months postrandomization compared with active control (primary outcome).

Methods: General practices recruited through practice-based research networks in Victoria and Tasmania were randomly allocated 1:1 to the FHT CKD module or active control. The intervention was delivered to practices between October 4, 2021, and September 30, 2022. Data extracted from EMRs for eligible patients identified at baseline were used to evaluate the trial outcomes at the completion of the intervention period. The primary analysis used an intention-to-treat approach. The intervention

XSL•FO
RenderX

effect for the primary outcome was estimated with a marginal logistic model using generalized estimating equations with robust SE.

Results: Overall, of the 734 eligible patients from 19 intervention practices and 715 from 21 control practices, 82 (11.2%) and 70 (9.8%), respectively, had received appropriate pharmacological therapy (statins alone, ACEI or ARB alone, or both) at 12 months postintervention to reduce CVD risk, with an estimated between-trial group difference (Diff) of 2.0% (95% CI –1.6% to 5.7%) and odds ratio of 1.24 (95% CI 0.85 to 1.81; *P*=.26). Of the 470 intervention patients and 425 control patients that received a recommendation for statins, 61 (13%) and 38 (9%) were prescribed statins at follow-up (Diff 4.3%, 95% CI 0 to 8.6%; odds ratio 1.55, 95% CI 1.02 to 2.35; *P*=.04). There was no statistical evidence to support between-group differences in other secondary outcomes and general practice health care use.

Conclusions: FHT harnesses the data stored within EMRs to translate guidelines into practice through quality improvement activities and active clinical decision support. In this instance, it did not result in a difference in prescribing or clinical outcomes except for small changes in statin prescribing. This may relate to COVID-19–related disruptions, technical implementation challenges, and recruiting higher performing practices to the trial. A separate process evaluation will further explore factors impacting implementation and engagement with FHT.

Trial Registration: ACTRN12620000993998; https://www.anzctr.org.au/Trial/Registration/TrialReview.aspx?id=380119

(JMIR Form Res 2025;9:e54147) doi:10.2196/54147

KEYWORDS

primary care; general practice; clinical decision support; chronic kidney disease; cardiovascular disease; cardiovascular; clinical decision; decision support; support; kidney; kidney disease; electronic medical record; risk; risk reduction; pharmacological; pharmacological therapy; medical records; logistic model

Introduction

More than 4 in 5 Australians visit their general practice at least once per year and 2 million attend every week [1,2]. As medical knowledge continues to increase exponentially, it is crucial that this knowledge is translated into practice efficiently and effectively. Australian general practitioners (GPs) need to have a good working knowledge of 167 conditions to manage 85% of presentations [3], but there is no way for GPs to easily keep up to date with a multitude of guidelines that are stored in different locations while delivering person-centered care. This is critically important for people at risk of, or with, chronic diseases such as chronic kidney disease (CKD), where early detection and management have the potential to reduce disease progression and the development of complications, improve quality of life, and reduce the burden on the health care system [4].

CKD is a common condition, primarily managed in the general practice setting, that is defined as kidney damage reduced function, or both present for at least 3 months [5]. In Australia, CKD affects up to 2 million people and costs the Australian economy more than Aus $5 billion per year (a currency exchange rate as of October 1, 2021, of Aus $1=US $0.726 is applicable) [6]. Untreated CKD can lead to end-stage kidney failure, which may require dialysis or transplant therapies, but the most common cause of death in people with CKD is cardiovascular disease (CVD) [4]. Kidney Health Australia guidelines support the detection of CKD in high-risk populations and recommend pharmacological treatment with angiotensin-converting enzyme inhibitor inhibitors (ACEI) or angiotensin receptor blockers (ARB), or statins or both statins and an ACEI or ARB to reduce CKD progression and CVD risk [5]. It has been identified that additional strategies are required to optimize CVD risk management in general practice [7,8].

Quality improvement (QI) and clinical decision support (CDS) software may help to optimize the care provided to people with CKD in general practice. Building on systematic review evidence for effective elements of these interventions [9,10], we collaborated with GPs, general practice nurses, practice managers, and patients to codesign Future Health Today (FHT) [11]. The FHT software program integrates with the electronic medical records (EMRs) used by over 90% of Australian general practices. It consists of 4 components: a dashboard, a CDS tool, access to resources and QI activities.

This trial aimed to determine whether the FHT platform, packaged with QI support and a series of webinars provided using the Extension for Community Healthcare Outcomes (Project ECHO) program [12] through the Zoom videoconferencing platform, was effective in increasing the proportion of patients with CKD receiving guideline-concordant care to reduce CVD risk at 12 months postrandomization in intervention practices compared with control practices which received an FHT module focused on cancer care. This is a timely trial as policy informing the use of CDS in Australian general practice is currently under development [13].

Methods

Overview

We conducted a stratified cluster randomized head-to-head trial with general practices randomly assigned 1:1 to the FHT CKD program or an active control (FHT cancer program). The full protocol was published [14] and registered with the Australia and New Zealand Clinical Trial Registry (ACTRN12620000993998) [15].

Setting

The FHT trial was conducted in 39 general practices in Victoria and 1 practice in Tasmania, Australia. In Australia, general

XSL•FO
RenderX

practice provides continuous, longitudinal care across the lifespan. The Australian government funds general practice through the Medicare Benefit Schedule (MBS) as a fee-per-service model. The MBS is a list of the medical services for which the Australian Government will make a payment [16], although this may not cover the full cost of the consultation.

The COVID-19 pandemic had a significant impact on general practices, service provision, and patient access during the study period. The government strategy was COVID-19 elimination with frequent lockdowns and border closures. The sixth lockdown in Victoria lasted from August 5, 2021, to October 22, 2021, crossing over the commencement of this trial, which had already been delayed by 6 months due to the pandemic. Significant workload related to COVID-19 infection control procedures and vaccination [17], a shift to telehealth [18], and a significant reduction in pathology services utilization were observed [19].

Recruitment

Forty general practices were recruited between October 2020 and August 2021. Practices were recruited through the VicREN practice-based research and education network at the Department of General Practice and Primary Care, The University of Melbourne [20], the University of Tasmania's Northern Tasmania General Practice-based research network, and through newsletters sent to practices through Primary Health Networks. Following the referral of interested practices by these networks to the study team, the research team liaised with general practice staff to explain the study further and to arrange consent and the completion of an information technology checklist to ensure that the general practice technology infrastructure that was consistent with the requirements of the FHT software.

General practice eligibility criteria are summarized in Textbox 1. Practices needed to have at least 2500 active patients in their EMR system (ie, not marked as deceased or no longer attending the practice) and use either the Best Practice or Medical Director EMR systems. These 2 systems are estimated to be used by more than 90% of Victorian general practices [21]. Practices also needed to agree to install the GRHANITE data extraction tool [22] so that they could contribute deidentified data from their EMR to the Patron dataset [23] to facilitate analysis of the trial outcome measures.

Practices were compensated Aus $2250 for participating in the trial; additional payments were made to GPs and general practice staff if they were nominated as practice champions (Aus $200) or participated in interviews (Aus $50 per interview).

There was no individual patient-level recruitment or consent required as outcomes were ascertained using deidentified EMR data. Inclusion criteria for patient records: Adults aged 18 to 80 years, inclusive, that were not marked as inactive or deceased, with a recorded diagnosis or pathology tests consistent with CKD as defined by Kidney Health Australia [5] that may benefit from pharmacological therapy to reduce CVD risk at baseline. Individuals with a recorded history of renal transplant or chronic dialysis were excluded. Pregnant individuals were also excluded.

Textbox 1. General practice eligibility criteria.

Inclusion criteria

- Agreed to install the GRHANITE data extraction tool [22] to contribute deidentified data from their electronic medical records (EMRs) to the Patron dataset [23].
- Consultations with at least 35 adults aged ≥18 years per day, >2500 active adult patients (defined as patients who attended the general practice at least 3 times in the last 2 years and not marked as deceased) recorded in their EMR or have at least 50 patients that fit cohort definitions for people with chronic kidney disease not on optimal medications and abnormal test results and additional clinical features placing them at risk of an undiagnosed cancer.
- Employed a general practice nurse.
- Contributed (or willing to contribute) data to the Patron dataset, a repository of data from EMR shared by general practices and curated by the University of Melbourne.
- Used Best Practice or Medical Director EMR software to record clinical consultations, prescription of medications, and ordering and receiving pathology results.
- Could identify a workstation (i5/i7 and 16GB RAM or upgradeable to 16GB) with Windows 10 (operating system) that would have GRHANITE, the data extraction tool, installed.
- Computers had Edge or Chrome web browsers installed.

Exclusion criteria

- Had previously participated in other Future Health Today projects.
- Intended to change to another EMR software during the trial period.
- Used a cloud-based EMR system.

Randomization

General practices were allocated 1:1 to the FHT CKD program or the FHT cancer program (active control), using a computer-generated schedule, stratified by Index of Relative Socioeconomic Disadvantage (IRSD) terciles [24] and the number of full-time equivalent (FTE) GPs (4 or fewer vs greater than 4), using random permuted block sizes of 2 and 4 within the stratum. Practices were randomly allocated after they were all recruited, and practices' baseline measures were collected.

XSL•FO
RenderX

The random allocation schedule was generated by the statistician not involved in the practice recruitment or data collection. After practices were recruited, the statistician randomly allocated the deidentified practices to one of the 2 intervention groups and then informed the clinical liaison staff, who advised practices of their allocated group. Practices were informed that it was a pragmatic cluster randomized controlled trial with 2 active groups.

Blinding of general practice participants and the team providing QI support was not possible. The statistician conducting the analysis and study investigators not involved in practice support and engagement remained blinded, and allocation of practices was only revealed after the data were analyzed and results interpreted.

Intervention

Future Health Today Software

FHT is software installed in general practices to identify people who may benefit from the implementation of guideline-informed care. There are two main components: (1) a dashboard that can be used to create cohorts of patients that may benefit from review and potential optimization of care and (2) a CDS tool that deploys when a patient file is opened. Login details and passwords were provided for the dashboard. The CDS tool automatically deployed when a patient file was opened within the EMR without the need for the clinician to log in to FHT. FHT cohorts and CDS prompts are generated each night when code related to best practice guideline recommendations is run over EMR systems in the practice. Screenshots that illustrate the different components of FHT are provided in Multimedia Appendix 1. Implementation activities and timing for delivery are detailed in Figure 1. Following quality assurance testing of the FHT platform and the module algorithms in a virtual environment, FHT was installed in all practices (intervention and control) before the commencement of the trial. Practices were asked to install FHT on at least one GP and one practice nurse computer. A training FHT type 2 diabetes module was deployed and available to both intervention and control practices for approximately 4 weeks before the trial, to ensure that FHT was working correctly from a technical perspective. Pretrial training sessions covering the functionality of the FHT CDS and dashboard components were offered to all practices.

At trial commencement, the training FHT type 2 diabetes module was deactivated and the FHT CKD module was deployed to intervention practices. The CKD module consisted of 3 recommendations for medications that aim to reduce cardiovascular disease risk: “Consider initiation of ACEI or ARB,” “Consider initiation of statin for CKD management,” and “Consider initiation of ACEI or ARB (note: BP in target)” for patients who had an indication for these medications but had not been prescribed them. These recommendations were based on guidelines from Kidney Health Australia [5], the Royal Australian College of General Practitioners [25], and the National Vascular Disease Prevention Alliance [26] and were applied to patients who had a recorded diagnosis of CKD or pathology consistent with CKD, as defined by Kidney Health Australia guidelines [5]. Practices were asked to create a cohort of patients with these recommendations in the dashboard at baseline and asked to use FHT as they chose for the following 12 months.

Figure 1. Trial timeline and data sources (chronic kidney disease group).

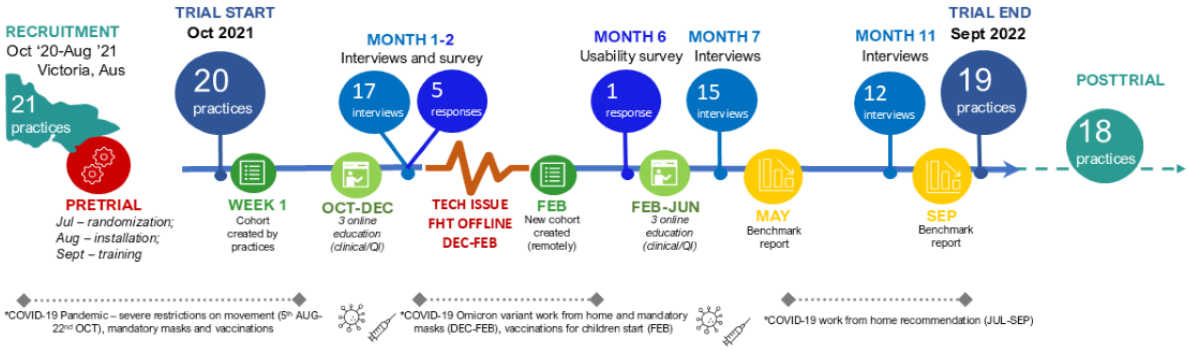

Quality Improvement Supports

A practice liaison staff member was assigned to support the intervention practices. This staff member provided software training to practices, and 2 benchmarking reports (6 months and 12 months from trial start) and offered practice guidance in undertaking a quality improvement clinical audit activity related to the intervention. The benchmarking reports summarized the number of patients with recommendations for a specific practice and then compared these to the overall intervention cohort. The practice liaison also coordinated a series of webinars using the Project ECHO program [12], an evidence-based platform that facilitates case-based learning networks with primary care, facilitated by academic medicine departments, and delivered by the Zoom videoconferencing platform. Three separate 1-hour sessions were delivered with a focus on CKD management and reduction of cardiovascular disease risk, and a further three 1-hour sessions were delivered on how to conduct quality improvement activities. The clinical audit and webinar series were accredited by the Royal Australian College of General Practitioners and the Australian College of Rural and Remote Medicine as continuing professional development (CPD) activities. All participants, including nonclinical staff, were provided with certificates of participation in these activities.

Two deviations from the initially proposed protocol occurred, both of which related to the availability of the FHT software, (1) there was a pause in the provision of the CKD module in FHT software from December 16, 2021, to February 14, 2022, due to complexity of the FHT module algorithms and mappings

XSL•FO
RenderX

from CKD not being appropriately implemented within the FHT code. This meant that the software was not available for the full intended 12 months, (2) one practice did not have access to FHT modules until 2 weeks after the trial started due to challenges with software installation.

Consumer Involvement

A consumer advisory group and general practice advisory group (consisting of GPs, general practice nurses, and practice managers) provided input to the trial intervention and interpretation of the trial results.

Endpoints and Data Collection

All outcomes were measured at 12 months postrandomization. The primary outcome was the proportion of eligible patients with a prescription of ACE inhibitors, ARBs statins, or both according to guidelines. Patients were considered to have received guideline-concordant care if they had a recorded prescription for the classes of medication indicated by FHT recommendations within the study period. If a patient received recommendations for both classes of medication at baseline, for the primary outcome, they were required to have been prescribed both classes of medication. If a prescription was not recorded in the EMR, patients were considered not to have received guideline-concordant care. The secondary outcomes related to the prescription of individual classes of these medications, pathology results (including estimated glomerular filtration rate, albumin creatinine ratio, and cholesterol), and systolic blood pressure; refer to Multimedia Appendix 2 for additional details. The number and billing of general practice consultations were also captured. Costs to the government associated with consultations occurring during the trial were based on the Medicare Benefits Schedule (MBS) rebate for the service items retrieved from the Patron database [27].

Baseline and outcome measures were obtained using patient data extracted from general practice EMRs and stored in the Patron database [28]. Patron data are stored at the University of Melbourne and were only accessible by the study statisticians.

Practice characteristics, including the number of GPs and other general practice staff, billing method, and regionality were collected through a survey before randomization.

We also collected survey and qualitative data relating to implementation, to be reported as a process evaluation elsewhere.

Sample Size

We determined that 2580 patients from 20 practices per group are required to detect an absolute 10% increase in the difference in the percentage of patients on optimal pharmacological management in the intervention group at 12 months postintervention compared with the control group at 80% power for a 2-sided 5% significance level. Based on electronic health records from 77 general practices that contribute data to the Patron dataset in Victoria, Australia [28], we expected 55% of active control patients with CKD would be on optimal pharmacological management with an intracluster correlation (ICC) coefficient of 0.03 to account for clustering at the practice level. Further, we expected on average 139 eligible patients per practice (fixed cluster size), with a coefficient of variation of 0.41 to account for variable cluster sizes across the practices. The target of 40 practices included 4 additional practices to allow for attrition by 12 months (eg, practice closure, withdrawals, mergers) and the addition of one extra practice per group, for a degree of freedom correction. The effect size of 10% was considered a minimally important difference and worth implementing and is consistent with another Australian general practice study that aimed to improve CVD risk management, albeit not limited to patients with CKD [29].

Statistical Analysis

Descriptive statistics were used to summarize general practice, clinician, and patient characteristics by study group. The primary analysis used an intention-to-treat approach with practices analyzed according to their allocated group. All general practice patients were included in the analysis, regardless of the occurrence of intercurrent events (eg, death, pregnancy). For the primary outcome, the OR and between-group differences in proportions were estimated with a generalized linear regression model, with the logit and identity link functions, respectively. Generalized estimating equations (GEEs) were used with an exchangeable correlation structure and robust SE to allow for the correlation of outcomes within general practice. The ICC was estimated using a one-way analysis of variance. Binary secondary outcomes used the same method as the primary outcome. For continuous outcomes, a linear mixed effects model was used to estimate the between-group difference in mean change scores with random effects for general practice and individual and adjusted for their respective baseline measure with study group means at baseline constrained to be equal. A proportional odds logistic regression model with GEEs and robust SE was used to estimate cumulative OR for the albumin creatinine ratio level at 12 months. The between-group difference in the rate of consultations per year and rate ratio of visits were modeled using a negative binomial mixed effects model, with random effects for general practice, with GEEs and robust SE.

A 2-part model consisting of a logistic regression model for the probability of positive Medicare benefit and a generalized linear model with a log link and gamma distribution for the positive Medicare benefit was used to assess the association between the intervention and the Medicare benefit. The marginal effect of the intervention was computed in both the logistic regression model (as percentage point change) and the generalized linear regression model (as dollar change).

All regression models included the study group and randomization stratification factors (GP, FTE, and IRSD tercile) as covariates. For secondary binary outcomes, the recommendation to initiate statin and ACEI or ARB was adjusted for the recommendation to consider ACEI or ARB and statin initiation, respectively. Sensitivity analyses included (1) complete cases only (for continuous outcomes, if the linear mixed effects model failed to converge, linear regression with robust SE was used) and (2) adjusting for patient age and sex and previous practice involvement in formalized QI programs. For the binary outcomes, if the model used to estimate the risk difference failed to converge, the additional covariates were

XSL•FO
RenderX

excluded (noted in the results tables). Adherence-adjusted analysis was not conducted as we were unable to measure adherence to the intervention.

Estimated intervention effects were reported with 95% CI and *P* values. No corrections were made for multiple comparisons.

All analyses were conducted using Stata 17.0 (StataCorp). The full statistical analysis plan has been uploaded to the ANZCTR.

Ethical Considerations

The FHT trial was approved by the Faculty of Medicine, Dentistry and Health Sciences Human Ethics Sub-Committee at the University of Melbourne (Ethics ID:2056564). General practices signed a study agreement to participate in the trial. Patients were not recruited for this study, with data from the Patron dataset used to calculate the primary and secondary trial outcomes. The Patron dataset has a waiver of consent granted by the Faculty of Medicine, Dentistry and Health Sciences Human Ethics Sub-Committee at the University of Melbourne (ethics number 2023-23358-37538-6). Data within the Patron dataset is anonymized. General practices were paid Aus $2250 to compensate them for any IT support, meeting or training attendance, generating patient lists, mailouts, monthly key staff informant interviews, and participation in quality improvement activities and ECHO education activities.

Results

Participating Practices

Between October 2020 and August 2021, there were 775 general practices assessed for eligibility, following which 50 practices were recruited to participate in the trial (refer to Figure 2). Six practices withdrew before randomization (all these practices were owned by the same group), and a practice withdrew after randomization but before they were notified of their group allocation. Two randomized practices did not receive the allocated intervention and 1 discontinued the intervention during the trial, leaving forty participating practices that completed the trial. Patron data was only provided for the forty practices that completed the study.

Baseline characteristics of the 40 practices, clinicians, and staff (Table 1) and their eligible patients (Table 2) were balanced between the study groups. A total of 734 and 715 eligible patients in the intervention and control groups, respectively, were identified in the EMR with a recorded diagnosis or pathology tests consistent with CKD that may benefit from pharmacological therapy to reduce CVD risk. Of these, 17% had a recommendation for both ACEI or ARB and satins, 45% for statins only, and the remaining 38% for ACEI/ARBS only.

XSL•FO
RenderX

Figure 2. Practice recruitment, allocation, follow-up, and data analysis.

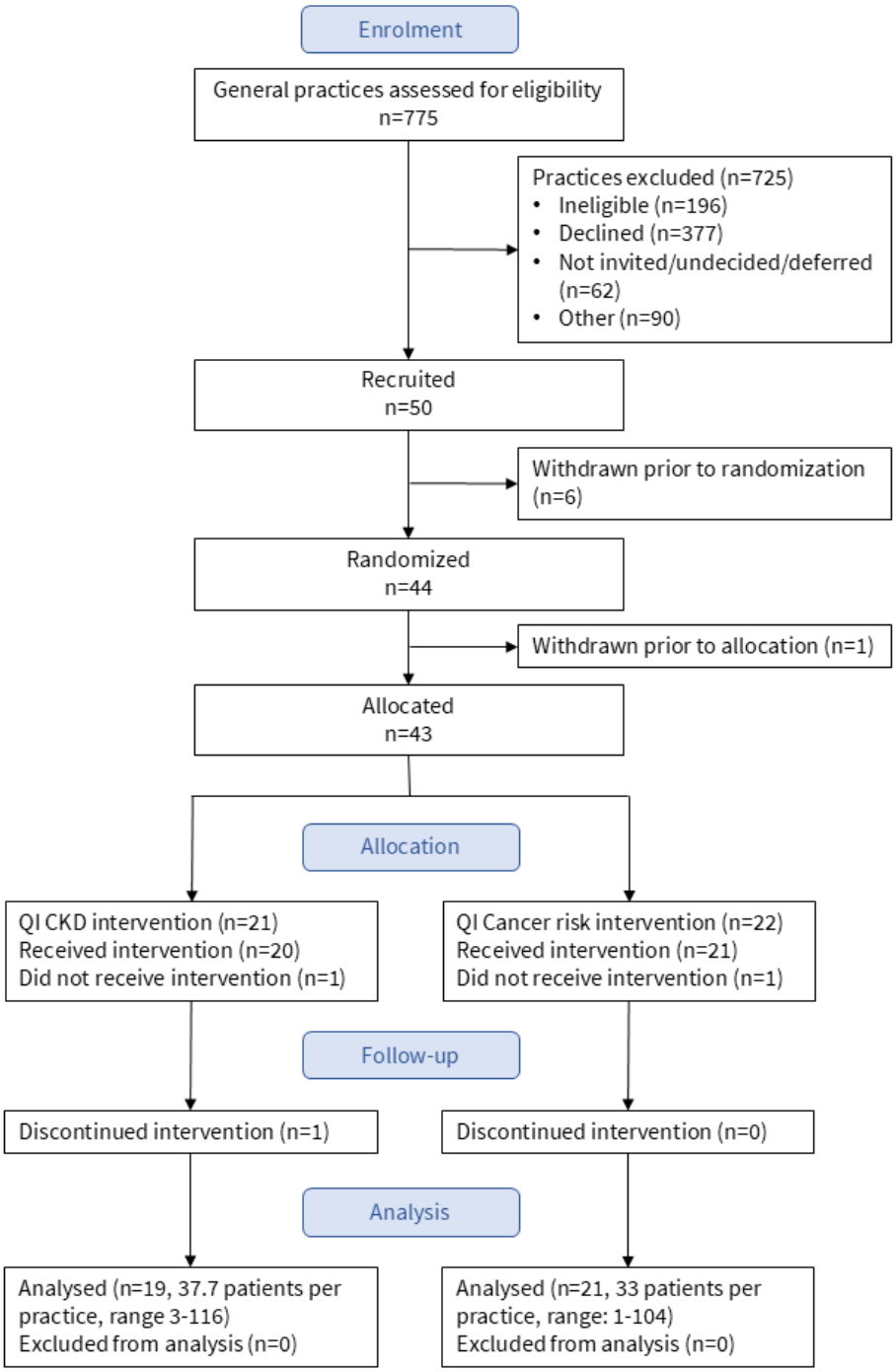

XSL•FO
RenderX

Table 1. Baseline characteristics of practices, clinicians, managers, and staff, overall and by study group.

	All practices (N=40)	Intervention group (N=19)	Control group (N=21)
Practice characteristics			
State, n (%)			
Victoria	39 (97.5)	19 (100)	20 (95.2)
Tasmania	1 (2.5)	0 (0)	1 (4.8)
Relative socioeconomic disadvantage index (terciles), n (%)			
1 (most disadvantaged)	12 (30)	6 (31.6)	6 (28.6)
2	12 (30)	6 (31.6)	6 (28.6)
3 (least disadvantaged)	16 (40)	7 (36.8)	9 (42.8)
Practice size			
4 or fewer FTE[a] GPs[b], n (%)	21 (52.5)	9 (47.4)	12 (57.1)
Greater than 4 FTE GPs, n (%)	19 (47.5)	10 (52.6)	9 (42.9)
Previous quality improvement program participation, n (%)	18 (45)	9 (47.4)	9 (42.9)
Eligible patients per practice, mean (range)	35.3 (1-116)	37.7 (3-116)	33.0 (1-104)
Workstations with Future Health Today per practice, mean (range)[c]	6.4 (1-20)	7.4 (1-20)	5.5 (1-19)
Future Health Today users per practice, mean (range)	31.8 (10-160)	30.2 (14-73)	33.2 (10-160)
Number of FTE GPs, median (IQR)	4.5 (3.5-6.0)	4.5 (3.5-7.5)	4.0 (3.5-5.5)
Number of FTE practice nurses, median (IQR)	1.8 (1.0-2.5)	1.5 (1.0-3.0)	2.0 (1.5-2.5)
Number of FTE practice managers or administrative staff, median (IQR)	4.0 (3.0-5.0)	4.5 (3.5-6.0)	3.5 (2.5-4.5)
General practitioners	309	164	145
Sex, n (%)			
Male	178 (57.6)	93 (56.7)	85 (58.6)
Female	131 (43.4)	71 (43.3)	60 (41.4)
Age, n (%)			
<35 years	40 (15)	19 (14.2)	21 (15.8)
35-50 years	129 (48.3)	56 (41.8)	73 (54.9)
>50 years	98 (36.7)	59 (44)	39 (29.3)
Nurses	123	64	59
Sex, n (%)			
Male	7 (5.7)	3 (4.7)	4 (6.8)
Female	116 (94.3)	61 (95.3)	55 (93.2)
Age[d], n (%)			
<35 years	60 (49.2)	33 (52.4)	32 (50)
35-50 years	31 (25.4)	14 (22.2)	17 (26.6)
>50 years	31 (25.4)	16 (25.4)	15 (23.4)
Practice managers or administrative staff	257	130	127
Sex, n (%)			
Male	27 (10.5)	14 (10.8)	13 (10.2)
Female	230 (89.5)	116 (89.2)	114 (89.8)
Age, n (%)			
<35 years	130 (51.6)	75 (60)	55 (43.3)
35-50 years	59 (23.4)	26 (20.8)	33 (26)

XSL•FO
RenderX

	All practices (N=40)	Intervention group (N=19)	Control group (N=21)
>50 years	63 (25)	24 (19.2)	39 (30.7)

[a]FTE: full-time equivalent.

[b]GPs: general practitioner.

[c]Six intervention and 2 control group practices ran a “terminal server” that pushed Future Health Today to all workstations.

[d]Percentages for age categories were calculated based on the number of staff who provided age data.

XSL•FO
RenderX

Table 2. Baseline characteristics of patients, overall and by study group.

Patient characteristics	All participants (N=1449)	Intervention group (n=734)	Control group (n=715)
Sex, n (%)			
Male	812 (56)	402 (54.8)	410 (57.3)
Female	637 (44)	332 (45.2)	305 (42.7)
Age (years), mean (SD)	65.9 (12.1)	66.3 (11.3)	65.4 (12.9)
Prescribed ACEI[a] or ARB[b] , n (%)			
ACEI only	215 (14.8)	109 (14.9)	106 (14.8)
ARB only	321 (22.2)	166 (22.6)	155 (21.7)
ACEI and ARB	6 (0.4)	4 (0.6)	2 (0.3)
None	907 (62.6)	455 (62)	452 (63.2)
Prescribed statins, n (%)			
Yes	424 (29.3)	192 (26.2)	232 (32.5)
No	1025 (70.7)	542 (73.8)	483 (67.5)
Guideline recommended initiation, n (%)			
Both ACEI or ARB and statins	246 (17)	131 (17.9)	115 (16.1)
ACEI or ARBs only	554 (38.2)	264 (36)	290 (40.5)
Statin only	649 (44.8)	339 (46.2)	310 (43.4)
Cardiovascular disease risk category, n (%)			
Total sample	1065 (73.5)	539 (73.4)	526 (73.8)
Low	85 (8)	43 (7.6)	42 (8)
Moderate	29 (2.7)	14 (2.7)	15 (2.9)
High	951 (89.3)	482 (89.7)	469 (89.2)
Type 2 diabetes n (%)	633 (43.7)	308 (42)	325 (45.5)
Systolic blood pressure (mmHg)			
Number (%)	843 (58.2)	434 (59.1)	409 (57.2)
Mean (SD)	138 (20)	138 (20)	137 (19)
Lipids (all mmol/L)			
Total cholesterol, n (%)	916 (63.2)	465 (63.4)	451 (63.1)
Total cholesterol, mean (SD)	4.8 (1.2)	4.8 (1.2)	4.8 (1.3)
Low-density lipoproteins, n (%)	844 (58.2)	431 (58.7)	413 (57.8)
Low-density lipoproteins, mean (SD)	2.7 (1.1)	2.7 (1.1)	2.6 (1.1)
High-density lipoproteins, n (%)	865 (59.7)	444 (60.5)	421 (58.9)
High-density lipoproteins, mean (SD)	1.3 (0.4)	1.3 (0.4)	1.3 (0.4)
Triglycerides, n (%)	914 (63.1)	464 (63.2)	450 (62.9)
Triglycerides mean (SD)	2.1 (1.8)	2.0 (1.4)	2.1 (2.1)
Urine albumin-creatinine ratio (mg/mmol)[c]			
Number (%)	702 (48.4)	323 (44)	379 (53)
Median (IQR)[c]	5.5 (1.6-21.7)	5.2 (1.4-22.5)	5.6 (1.7-20.2)
Macroalbuminuria, n (%)	137 (19.5)	64 (19.8)	73 (19.3)
Microalbuminuria, n (%)	343 (48.9)	150 (46.4)	193 (50.9)
Normal, n (%)	222 (31.6)	109 (33.8)	113 (29.8)
Estimated glomerular filtration rate (mL/min/1.73m^2)			

XSL•FO
RenderX

Patient characteristics	All participants (N=1449)	Intervention group (n=734)	Control group (n=715)
Number (%)	1175 (81.1)	592 (80.7)	583 (81.5)
Mean (SD)	55.3 (20.9)	53.5 (20.4)	57.1 (21.2)
G1 (≥90), n (%)	149 (12.7)	61 (10.3)	98 (15.1)
G2 (60-89), n (%)	224 (19.1)	105 (17.7)	119 (20.4)
G3a (45-59), n (%)	428 (36.4)	221 (37.3)	207 (35.5)
G3b (30-44), n (%)	267 (22.7)	146 (24.7)	121 (20.8)
G4 (15-29), n (%)	73 (6.2)	37 (6.3)	36 (6.2)
G5 (<15), n (%)	35 (2.9)	22 (3.7)	12 (2.1)

[a]ACEI: angiotensin-converting enzyme inhibitor.

[b]ARB: angiotensin receptor blocker.

[c]Median and IQR were reported due to skewed distribution.

Primary and Secondary Outcomes

Primary Outcome

Of patients identified as eligible for pharmacological therapy to reduce CVD risk, 11.2% (82/734) in the intervention group and 9.8% (70/715) in the control group had received appropriate pharmacological therapy at 12 months post-intervention, an increase of 2.0% (95% CI –1.6% to 5.7%) (Table 3), with the confidence bounds excluding our hypothesized minimally important difference of 10%. The ICC for the proportion of patients who were prescribed ACEI or ARB and Statin at baseline was estimated as 0.03 (95% CI 0.01 to 0.05).

Table 3. Primary outcome and secondary outcomes related to medication prescriptions

	All participants, n (%)	Intervention group, n (%)	Control group, n (%)	Estimated effect size (95% CI)		*P* value
				Diff[a] (95% CI), %	OR[b] (95% CI)	
Primary outcome						
Angiotensin-converting enzyme inhibitor or angiotensin receptor blocker and statins						
Number of participants	1449	734	715			
Primary analysis	152 (10.5)	82 (11.2)	70 (9.8)	2.0 (–1.6 to 5.7)	1.24 (0.85 to 1.81)	.26
Sensitivity analysis[c]	152 (10.5)	82 (11.2)	70 (9.8)	2.6 (0.0 to 6.1)	1.30 (0.90 to 1.87)	.17
Secondary outcomes						
Angiotensin-converting enzyme inhibitor or angiotensin receptor blocker						
Number of participants	800	395	405			
Primary analysis	100 (12.5)	48 (12.2)	52 (12.8)	0.0 (–3.7 to 4.6)	0.94 (0.62 to 1.42)	.76
Sensitivity analysis[a]	100 (12.5)	48 (12.2)	52 (12.8)	1.8 (–2.6 to 6.2)	1.04 (0.69 to 1.57)	.85
Statins						
Number of participants	895	470	425			
Primary analysis	99 (11.1)	61 (13)	38 (9)	4.3 (0 to 8.6)	1.55 (1.02 to 2.35)	.04
Sensitivity analysis[a]	99 (11.1)	61 (13)	38 (9)	4.4 (0 to 8.5)	1.56 (1.04 to 2.34)	.03

[a]Diff: difference in mean change score or percentages between the 2 groups.

[b]OR: odds ratio.

[c]Sensitivity analysis 1, as above, with adjustment for participant age and sex and general practitioner participation in formal quality improvement activities.

Secondary Outcomes

When we examined the patients with recommendations for statins and ACEI or ARBS, separately, 13% (61/470) of patients in the intervention group received a statin prescription compared to 9% (38/425) in the control group, an estimated between-group difference of 4.3% (95% CI 0%-8.6%). The sensitivity analyses

XSL•FO
RenderX

showed similar results. Primary analyses are generalized linear regression using generalized estimating equations with robust SE, adjusted for GP FTE and IRSD tercile.

Table 4 shows that there was no statistical evidence of an intervention effect for mean change in pathology measures from baseline between the intervention and control groups. Primary analyses for continuous measures used a linear mixed effects model with random effects for general practice and individuals, adjusted for baseline measure of the outcome. Primary analyses for binary outcomes used generalized linear regression using generalized estimating equations with robust standard errors. All primary analyses adjusted for GP FTE and IRSD tercile.

Table 5 summarizes the nonpathology measures of secondary outcomes. There was no evidence of an effect of the intervention on systolic blood pressure change, cardiovascular disease risk category, or rate of general practice visits of patients over the 12-month intervention period compared to the control group. Primary analyses for continuous measures used a linear mixed effects model with random effects for general practice and individuals and adjusted for baseline measures of the outcome. Primary analyses for ordinal outcomes used proportional odds logistic regression, using generalized estimating equations with robust standard errors. The primary analysis of the rate of visit data used a negative binomial mixed effects model with random effects for general practice. All primary analyses adjusted for GP FTE and IRSD tercile.

XSL•FO
RenderX

Table 4. Secondary outcomes for laboratory pathology measures.

	All participants	Intervention group	Control group	Estimated effect size, (95% CI)	*P* value
Cholesterol change (mmol/L), mean (SD)					
Sample size	563	260	303		
Primary analysis	0.14 (0.94)	0.07 (0.84)	0.20 (1.01)	Diff[a] –0.11 (–0.28 to 0.07)	.24
Sensitivity analysis[b]	0.14 (0.94)	0.07 (0.84)	0.20 (1.01)	Diff –0.09 (–0.27 to 0.09)	.32
Sensitivity analysis[c,d]	0.14 (0.94)	0.07 (0.84)	0.20 (1.01)	Diff –0.13 (–0.27 to 0.03)	.06
Low-density lipoproteins change (mmol/L), mean (SD)					
Sample size	484	230	254		
Primary analysis	0.13 (0.75)	0.05 (0.67)	0.20 (0.82)	Diff –0.10 (–0.26 to 0.06)	.21
Sensitivity analysis[b]	0.13 (0.75)	0.05 (0.67)	0.20 (0.82)	Diff –0.10 (–0.26 to 0.06)	.24
Sensitivity analysis[c]	0.13 (0.75)	0.05 (0.67)	0.20 (0.82)	Diff –0.13 (–0.26 to 0.00)	.05
High-density lipoproteins change (mmol/L), mean (SD)					
Sample size	518	247	271		
Primary analysis	0.00 (0.21)	–0.01 (0.20)	0.01 (0.22)	Diff –0.04 (–0.09 to 0.02)	.17
Sensitivity analysis[b]	0.00 (0.21)	–0.01 (0.20)	0.01 (0.22)	Diff –0.05 (–0.10 to 0.01)	.09
Sensitivity analysis[c]	0.00 (0.21)	–0.01 (0.20)	0.01 (0.22)	Diff –0.03 (–0.07 to 0.01)	.11
Triglycerides change (mmol/L), mean (SD)					
Sample size	560	258	302		
Primary analysis	0.07 (1.61)	0.06 (0.88)	0.07 (2.03)	Diff 0.09 (–0.11 to 0.29)	.36
Sensitivity analysis[b]	0.07 (1.61)	0.06 (0.88)	0.07 (2.03)	Diff 0.15 (–0.05 to 0.35)	.15
Sensitivity analysis[c]	0.07 (1.61)	0.06 (0.88)	0.07 (2.03)	Diff 0.09 (–0.11 to 0.29)	.39
Urine albumin-creatinine ratio change (mg/mmol), mean (SD)					
Sample size	398	178	220		
Primary analysis	–2.13 (41.80)	–6.56 (42.99)	1.45 (40.56)	Diff –9.11 (–21.36 to 3.14)	.15
Sensitivity analysis[b]	–2.13 (41.80)	–6.56 (42.99)	1.45 (40.56)	Diff –8.97 (–21.23 to 3.31)	.15
Sensitivity analysis[c]	–2.13 (41.80)	–6.56 (42.99)	1.45 (40.56)	Diff –6.69 (–14.33 to 0.95)	.09
Urine albumin-creatinine ratio 30% reduction					
Sample size	398	178	220		
Primary analysis, n (%)	117 (29.4)	54 (30.3)	63 (28.6)	Diff 3.1% (–5.7% to 12.0%); OR[e] 1.14 (.74 to 1.76)	.55
Sensitivity analysis[b], n (%)	117 (29.4)	54 (30.3)	63 (28.6)	Diff 2.4% (–6.8% to 11.5%); OR 1.09 (.69 to 1.73)	.70
Estimated glomerular filtration rate change (mL/min/1.73^2), mean (SD)					
Sample size	894	437	457		
Primary analysis	0.10 (8.47)	–0.27 (8.30)	0.44 (8.64)	Diff 1.35 (–1.15 to 3.86)	.29
Sensitivity analysis[b]	0.10 (8.47)	–0.27 (8.30)	0.44 (8.64)	Diff 1.04 (–1.37 to 3.45)	.40
Sensitivity analysis[c]	0.10 (8.47)	–0.27 (8.30)	0.44 (8.64)	Diff –0.56 (–1.75 to 0.64)	.36

[a]Diff: difference in mean change score or percentages between the 2 groups.

[b]Sensitivity analysis 1 additionally adjusts for participant age and sex, and practice participation in formal QI activities.

[c]Sensitivity analysis 2 using complete case only for outcomes with missing data.

[d]Linear regression model with robust SE used as an alternative to linear mixed effects model due to convergence issues.

[e]OR: odds ratio.

XSL•FO
RenderX

Table 5. Secondary outcomes for nonpathology measures.

	All participants	Intervention group	Control group	Estimated effect size (95% CI)	*P* value
Systolic blood pressure changes from baseline mmHg, mean (SD)					
Sample size	676	335	341		
Primary analysis	2.29 (20.26)	3.11 (20.66)	1.47 (19.86)	Diff[a] 0.75 (–1.71 to 3.21)	.55
Sensitivity analysis[b]	2.29 (20.26)	3.11 (20.66)	1.47 (19.86)	Diff 0.88 (–1.58 to 3.35)	.49
Sensitivity analysis[c]	2.29 (20.26)	3.11 (20.66)	1.47 (19.86)	Diff 1.00 (–1.49 to 3.48)	.43
Cardiovascular d isease risk category, n (%)					
Sample size	1019	516	503		
Primary analysis	—[d]	—	—	Cumulative OR[e] 0.89 (0.64 to 1.22)	.47
Low CVD[f] risk	101 (9.9)	48 (9.3)	53 (10.5)		
Medium CVD risk	39 (3.8)	23 (4.5)	16 (3.2)		
High CVD risk	879 (86.3)	445 (86.2)	434 (86.3)		
Sensitivity analysis[b]	—	—	—	Cumulative OR 0.81 (0.60 to 1.10)	.18
Rate of visits, n (%)					
Sample size	1447	734	713		
Number of visits, mean (SD)	17,237	7786	9451		
Primary analysis	11.91 (14.45)	10.61 (12.16)	13.26 (16.37)	Diff –0.14 (–0.38 to 0.11); RR[g] 0.87 (0.68 to 1.11)	.27
Sensitivity analysis[b]	11.91 (14.45)	10.61 (12.16)	13.26 (16.37)	Diff –0.15 (–0.38 to 0.08); RR 0.86 (0.69 to 1.08)	.20

[a]Diff: difference in mean change score or percentages between the 2 groups.

[b]Sensitivity analysis 1 additionally adjusts for participant age and sex and practice participation in formal quality improvement activities.

[c]Sensitivity analysis 2 using complete cases for outcomes with missing data.

[d]Not applicable.

[e]OR: odds ratio.

[f]CVD: cardiovascular disease.

[g]RR: rate ratio.

Health Care Costs

Table 6 summarizes the health economic outcomes. Of 734 patients in the intervention group, 541 (73.7%) had at least one general practice consultation associated with an MBS rebate, compared to 80.6% (576/715) patients in the control group. Patients meeting the criteria for the CKD FHT module in the practices receiving the intervention had a 6.7 percentage point lower probability of having at least one general practice consultation associated with an MBS rebate compared to those in the control group, but there was no evidence of a difference (95% CI –0.26 to 0.13).

Patients in the intervention group who claimed at least one MBS rebate during the trial period had Aus $38.55 lower Medicare benefits compared with the control group over 12 months, roughly equivalent to one standard consultation in general practice lasting up to 20 minutes. There was no evidence of a difference between groups (95% CI Aus $328.46 to Aus $251.35; *P*=0.79). Primary analyses for binary outcomes used logistic regression, with robust SE. Primary analysis of positive Medicare benefits was carried out with a generalized linear model with log link and gamma distribution, with robust standard errors. All primary analyses adjusted for GP FTE and IRSD tercile.

XSL•FO
RenderX

Table 6. Health economic outcomes.

	All participants	Intervention group	Control group	Estimated effect size (95% CI)	*P* value
Probability of incurring Medicare Benefits Schedule rebate (logistic regression), n (%)					
Sample size	1449	734	715		
Primary analysis	1117 (77.1)	541 (73.7)	576 (80.6)	Diff[a] –0.067 (–0.260 to 0.125); odds ratio 0.719 (0.285 to 1.82)	.49
Sensitivity analysis[b]	1117 (77.1)	541 (73.7)	576 (80.6)	Diff –0.094 (–0.337 to 0.148; odds ratio 0.659 (0.236 to 1.84)	.45
Positive Medicare Benefits Schedule rebate (generalized linear model), Aus $, mean (SE)					
Sample size	1117	541	576		
Primary analysis	847.28 (23.01)	850.96 (33.33)	843.84 (31.81)	Diff –38.555 (–328.463 to 251.353)	.79
Sensitivity analysis[b]	847.28 (23.01)	850.96 (33.33)	843.84 (31.81)	Diff 0.880 (–255.905 to 257.666)	.99

[a]Diff: difference in mean change score or percentages between the 2 groups.

[b]Sensitivity analysis 1 additionally adjusts for participant age and sex and practice participation in formal quality improvement activities (and a dummy variable for missing quality improvement practice participation).

Discussion

Principal Results

Our findings do not support that the intervention was effective in increasing the proportion of eligible patients with optimal pharmacological therapies at 12 months postintervention to reduce their CKD risk. Among general practice adult patients identified in the EMR at baseline with a recorded diagnosis or pathology tests consistent with CKD that may benefit from an ACEI or ARB or statin prescription or both to reduce the risk of CKD based on guidelines, we found that 11.2% and 9.8% of eligible patients received appropriate pharmacological therapy at 12 months post-randomization in those receiving the FHT CKD module compared to active control, an estimated 2% difference between the intervention and control groups.

For participants with a recommendation to initiate a statin, 13.0% received a prescription for statin medication in the intervention group compared to 9.0% of control patients, an estimated 4.3% (95% CI 0% to 8.6%) increase in statin prescribing when initiation was recommended. Although the confidence bound excluded the 10% effect size hypothesized as minimally important for optimal pharmacological therapies, the confidence bounds include improvement in statins prescribing when recommended which could potentially impact reducing cardiovascular events [30], particularly when considering the estimated effect size may be attenuated as a result of several extenuating circumstances.

Circumstances that may have impacted the trial findings include: pausing access of practices to the CKD module from December 16, 2021, to February 14, 2022, due to the complexity of the FHT module algorithms and mappings from CKD not being appropriately implemented within the FHT code; challenges with software installation in some practices with one practice not having access to FHT modules until two weeks after trial start; the number of patients identified in the module being approximately half that proposed in the target sample size; and variation in some of the coding implemented in FHT compared with that provided in the business requirements document. A sensitivity analysis conducted on the cohort of patients who would have been identified, had the algorithms been implemented without error, did not change the trial findings (data not shown).

Further, our trial was conducted during the COVID-19 pandemic which included stay-at-home orders and general practices needed to deliver COVID-19 vaccinations. As a result, there were unprecedented workforce pressures that may have impacted the ability of practices to use FHT and reduced patient attendance at general practice. GPs may have had more confidence in prescribing statins than ACEI or ARB as these medications do not require pathology tests postinitiation, which may have been an additional consideration during the pandemic.

Strengths and Limitations

Our study has several strengths. First, this was a pragmatic trial and so the results that are obtained are likely to reflect the impact of FHT in real-world practice. Forty practices were included in the analysis, although we had anticipated 36 after allowing for attrition. EMR data were used to capture patient data and outcome measures minimizing selection bias. Without this EMR data capture, the trial would have been logistically impossible due to the impact of COVID-19 on the general practice workforce and the COVID-19 restrictions that were in place restricted travel and access to medical practices by researchers during this period. Our light touch intervention was designed to overcome the known financial and scalability barriers to more traditional academic detailing. Finally, randomizing practices rather than individuals minimized the risk of contamination.

Study limitations included cluster sizes being smaller than anticipated and there was a lower percentage of eligible patients

XSL•FO
RenderX

receiving appropriate pharmacological therapy at 12 months in both groups than expected. Possible reasons include that we may have attracted higher performing practices to participate in the trial, so there were smaller proportions of patients with CKD who were not being managed according to guidelines; improvements in coding of the EMR data between the datasets used to estimate the parameters to determine the sample size calculations and then used to measure the primary outcome; variation in the quality of recording of data in EMRs within participating general practices; or fewer patients than anticipated may have attended the general practice during the COVID-19 pandemic due to the Victorian lockdowns. However, the smaller cluster sizes for the 40 practices and lower percentages receiving appropriate pharmacological therapy improved study efficiency, allowing us to address the research question and provide sufficient certainty to rule out the target 10% intervention effect. Further, among patients who did not receive appropriate pharmacological therapy, we were unable to determine if the GP did not consider that medications were indicated for other reasons or if the patient chose not to take the medications.

Comparison with Previous Work

QI activities and CDS form an important part of Australian and international primary care policy, with the aim of optimizing the care provided to patients through the strengthening of clinical decision-making and improving care processes. There is increased focus on QI programs in Australian general practices, with payments to practices for participating in QI activities commencing in 2019 [31] and exploration of the use and regulation of CDS in general practice as the focus of current Australian government consultation [13].

The advent of EMRs has resulted in the development of QI and CDS software internationally and across care settings. Australian general practices were early adopters of EMRs in the 1990s, with near universal computerization by 2006 [21]. Similar to our trial, the results of trials of the use of QI dashboards and clinical decision support software (CDSS) tools in primary care to facilitate guideline-recommended prescribing have been underwhelming. For example, the TORPEDO cluster randomized trial in Australian primary care found that the proportion of patients with a recording of risk factors for CVD increased, but the proportion of patients at high CVD risk that received recommended prescriptions was not significantly different from the control group [29]. Hespe's mixed methods descriptive study of a QI program in the Central and Eastern Sydney Primary Health Network area aimed to improve monitoring, prescribing practices, and attainment of BP and lipid targets for CVD risk reduction. The intervention included plan-do-study act cycles supported by a QI officer, workshops, an audit tool, and the HealthTracker guideline recommendation tool, but did not result in a change in the proportion of patients prescribed blood pressure and lipid-lowering medications over the 22-month intervention period.[24] The EBMeDS CDSS trial did not result in improvements in the management of diabetes in Belgian primary care despite high physician satisfaction [32].

Overall, the effectiveness of CDSS is generally small to moderate and highly variable [33] with the predictors of meaningful improvement unclear [34]. A systematic review of 11 trials incorporating clinical dashboards found limited evidence for their use, especially if they were not a component of a multifaceted intervention [35].

All of these studies highlight the challenges of implementation of QI dashboards and CDSS in general practice, including lack of time, practices' need for technical support, a perceived lack of value for QI work, difficulty disseminating knowledge across the practice team, tensions between the team and clinical staff, workforce, a need for strong IT infrastructure and technical support, organization of practices as providing reactive (as opposed to proactive, planned) care and suboptimal recording of patient data in the EMR [24,32,36].

Our experiences with this trial have resulted in more robust mechanisms to validate our algorithms that incorporate the use of comprehensive data from Patron. We think this, combined with greater utility through supporting a wider range of decision support capabilities, may increase utilization of FHT and offer a greater chance of demonstrating the effectiveness of the FHT CDSS going forward. We will explore factors relevant to the FHT trial in a process evaluation and implementation study to be published separately. Upcoming studies of different conditions, including the results of the cancer module active control in this trial, will provide further information regarding the implementation and effectiveness of FHT in Australian general practice.

Conclusions

In our trial, the CKD FHT intervention found no differences in trial primary or secondary outcomes except for a small improvement in prescribing of statin medication when initiation was recommended which may reduce future cardiovascular events. These results may be reflective of several factors, including the impact of the COVID-19 pandemic on general practice, technical implementation challenges, and sampling from high-performing practices. Which may have attenuated the intervention effect.

Acknowledgments

We would like to thank the general practices, health care professionals, and consumers that provided input to the development of FHT, piloted the tool in their practices, and participated in this trial. We are grateful to Christine Hallinan (University of Melbourne) who undertook the sample size calculations, and Antonia Arfaras, Ralph Audehm, Nicholas Brayshaw, Jade Cartwright, Deborah Daly, Ian Dennis, Julian Fidge, John Hodgson, George Jiang, Fiona Kleinitz, Eileen Lam, Ed Skinner and Richard Smith, who provided important insights to this trial and its interpretation through their involvement in our consumer and general practice advisory groups. Deidentified patient data from the Patron primary care data repository (extracted from consenting general

XSL•FO
RenderX

practices), which has been created and is operated by the Department of General Practice and Primary Care, The University of Melbourne was used to underpin trial feasibility, sample size calculations and to evaluate the outcomes of the trial.

The Paul Ramsay Foundation has provided funding for staff costs related to the technical development of FHT and the Implementation and Evaluation team members. They have also provided funding for professional staff support and general practice and consumer advisory groups. JDE is supported by an NHMRC (National Health and Medical Research Council) Investigator grant (APP1195302). The active control group of the trial was supported by the Cancer Australia-funded Primary Care Collaborative Cancer Clinical Trials Group (PC4). The trial is affiliated with the CanTest Collaborative (funded by Cancer Research UK C8640/A23385) of which JDE is an Associate Director. The design, management, analysis, and reporting of the study are entirely independent of the funders.

Generative AI was not used in any portion of the manuscript writing.

Data Availability

The datasets generated and analyzed during this study are not publicly available due to them containing sensitive health information but may be available from the Department of General Practice and Primary Care, University of Melbourne subject to ethical and data governance approvals.

Conflicts of Interest

FHT software was developed by The University of Melbourne in collaboration with Western Health. These affiliations were displayed on the FHT dashboard and training materials. CN, JE, and JAM-N are the coleads of the Future Health Today program and DB leads the technical team that developed the software. Intellectual property related to FHT is owned by The University of Melbourne.

Multimedia Appendix 1
Screenshots of the Future Health Today dashboard and clinical decision support tool.
[PPTX File , 2886 KB - formative_v9i1e54147_app1.pptx]

Multimedia Appendix 2
Primary, secondary, and health economics outcomes at 12 months postrandomization for quality improvement chronic kidney disease program.
[DOCX File , 23 KB - formative_v9i1e54147_app2.docx]

Multimedia Appendix 3
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 2582 KB - formative_v9i1e54147_app3.pdf]

References

1. Hayes P. No one knows you like your GP. newsGP. 2018. URL: https://www1.racgp.org.au/newsgp/racgp/no-one-knows-you-like-your-gp [accessed 2024-12-24]
2. Britt H, Henderson J, Bayram C, Harrison C, Valenti L, Pan Y, et al. General practice activity in Australia 2014-15. In: General Practice Series No. 38. Sydney: Sydney University Press; 2015.
3. Cooke G, Valenti L, Glasziou P, Britt H. Common general practice presentations and publication frequency. Aust Fam Physician 2013;42(1-2):65-68 [FREE Full text] [Medline: 23529466]
4. Cardiovascular disease, diabetes and chronic kidney disease - Australian facts: prevalence and incidence, in cardiovascular, diabetes and chronic kidney disease series no. 2. Australian Institute of Health and Welfare (AIHW). 2014. URL: https://www.aihw.gov.au/reports/heart-stroke-vascular-diseases/cardiovascular-diabetes-chronic-kidney-mortality/summary [accessed 2014-10-24]
5. Kidney Health Australia. Chronic kidney disease (CKD) management in primary care (4th ed). Melbourne: Kidney Health Australia; 2020.
6. National Strategic Action Plan for Kidney Disease. Kidney Health Australia. 2019. URL: https://kidney.org.au/get-involved/advocacy/national-strategic-action-plan-for-kidney-disease [accessed 2024-12-24]
7. Razavian M, Heeley EL, Perkovic V, Zoungas S, Weekes A, Patel AA, et al. Cardiovascular risk management in chronic kidney disease in general practice (the AusHEART study). Nephrol Dial Transplant 2012;27(4):1396-1402. [doi: 10.1093/ndt/gfr599] [Medline: 22053091]
8. Hespe CM, Campain A, Webster R, Patel A, Rychetnik L, Harris MF, et al. Implementing cardiovascular disease preventive care guidelines in general practice: an opportunity missed. Med J Aust 2020;213(7):327-328. [doi: 10.5694/mja2.50756] [Medline: 32880959]

XSL•FO
RenderX

9. Ivers N, Jamtvedt G, Flottorp S, Young JM, Odgaard-Jensen J, French SD, et al. Audit and feedback: effects on professional practice and healthcare outcomes. Cochrane Database Syst Rev 2012;2012(6):CD000259. [doi: 10.1002/14651858.CD000259.pub3] [Medline: 22696318]
10. Van de Velde S, Heselmans A, Delvaux N, Brandt L, Marco-Ruiz L, Spitaels D, et al. A systematic review of trials evaluating success factors of interventions with computerised clinical decision support. Implement Sci 2018;13(1):114 [FREE Full text] [doi: 10.1186/s13012-018-0790-1] [Medline: 30126421]
11. Hunter B, Biezen R, Alexander K, Lumsden N, Hallinan C, Wood A, et al. Future Health Today: codesign of an electronic chronic disease quality improvement tool for use in general practice using a service design approach. BMJ Open 2020;10(12):e040228 [FREE Full text] [doi: 10.1136/bmjopen-2020-040228] [Medline: 33371024]
12. Arora S, Geppert CMA, Kalishman S, Dion D, Pullara F, Bjeletich B, et al. Academic health center management of chronic diseases through knowledge networks: Project ECHO. Acad Med 2007;82(2):154-160 [FREE Full text] [doi: 10.1097/ACM.0b013e31802d8f68] [Medline: 17264693]
13. General practice data and electronic clinical decision support - consultation regulation impact statement. Australian Government Department of Health and Aged Care. 2022. URL: https://consultations.health.gov.au/primary-health-network/gp-data-and-ecds-cris/ [accessed 2023-05-06]
14. Manski-Nankervis J, Hallinan CM, Hunter BL, Lumsden N, Martinez Gutierrez J, Chima S, et al. Future Health Today: protocol for a pragmatic stratified cluster randomised head-to-head trial of quality improvement activities in general practice compared to active control. University of Melbourne 2022 [FREE Full text] [doi: 10.2196/preprints.54147]
15. Manski-Nankervis J. Future Health Today: a cluster randomised controlled trial of quality improvement activities in general practice. 2020. URL: https://www.anzctr.org.au/Trial/Registration/TrialReview.aspx?id=380119&isReview=true [accessed 2021-03-14]
16. What is Medicare? healthdirect. 2023. URL: https://www.healthdirect.gov.au/what-is-medicare [accessed 2023-05-06]
17. Kidd M, de Toca L. The contribution of Australia's general practitioners to the COVID-19 vaccine rollout. Aust J Gen Pract 2021;50(12):871-872 [FREE Full text] [doi: 10.31128/AJGP-11-21-6235] [Medline: 34845474]
18. McGillion M, Swoboda K, Jones S, Canfield D, Wolba B, Vaughan A, et al. Expansion of telehealth services. In: The Auditor-General Auditor-General Report No.10 2022-23 Performance Audit. Canberra: Australian National Audit Office; 2023.
19. Gillam MH, Roughead E, Tavella R, Dodd T, Beltrame J, Ryan R, et al. Impact of COVID-19 restrictions on pathology service utilisation. Intern Med J 2022;52(1):42-48 [FREE Full text] [doi: 10.1111/imj.15501] [Medline: 34432345]
20. Soós M, Temple-Smith M, Gunn J, Johnston-Ata'Ata K, Pirotta M. Establishing the Victorian primary care practice based research network. Aust Fam Physician 2010;39(11):857-862 [FREE Full text] [Medline: 21301660]
21. McInnes DK, Saltman DC, Kidd MR. General practitioners' use of computers for prescribing and electronic health records: results from a national survey. Med J Aust 2006;185(2):88-91. [doi: 10.5694/j.1326-5377.2006.tb00479.x] [Medline: 16842064]
22. Liaw S, Boyle D. Secure data linkage and information sharing with GRHANITE. In: The Person in the Centre. Melbourne: Health Informatics Conference; 2008.
23. Boyle D, Sanci L, Emery J, Gunn J, Hocking J, Manski-Nankervis J, et al. PATRON primary care research data repository. The University Of Melbourne 2019 [FREE Full text] [doi: 10.26188/5c52934b4aeb0]
24. Hespe CM, Giskes K, Harris MF, Peiris D. Findings and lessons learnt implementing a cardiovascular disease quality improvement program in Australian primary care: a mixed method evaluation. BMC Health Serv Res 2022;22(1):108 [FREE Full text] [doi: 10.1186/s12913-021-07310-6] [Medline: 35078460]
25. The Royal Australian College of General Practitioners, Guidelines for Preventive Activities in General Practice 9th edition. East Melbourne: RACGP; 2018.
26. National Vascular Disease Prevention Alliance, Guidelines for the Management of Absolute Cardiovascular Disease Risk. Melbourne: National Stroke Foundation; 2012.
27. MBS Online. Australian Government Department of Health and Aged Care. URL: http://www.mbsonline.gov.au/internet/mbsonline/publishing.nsf/Content/Home [accessed 2023-06-04]
28. Manski-Nankervis J, Canaway R, Chidgey C, Emery J, Sanci L, Hocking JS, et al. Data resource profile: primary care audit, teaching and research open network (Patron). Int J Epidemiol 2024;53(1):dyae002 [FREE Full text] [doi: 10.1093/ije/dyae002] [Medline: 38302745]
29. Peiris D, Usherwood T, Panaretto K, Harris M, Hunt J, Redfern J, et al. Effect of a computer-guided, quality improvement program for cardiovascular disease risk management in primary health care: the treatment of cardiovascular risk using electronic decision support cluster-randomized trial. Circ Cardiovasc Qual Outcomes 2015;8(1):87-95. [doi: 10.1161/CIRCOUTCOMES.114.001235] [Medline: 25587090]
30. Baigent C, Landray MJ, Reith C, Emberson J, Wheeler DC, Tomson C, SHARP Investigators. The effects of lowering LDL cholesterol with simvastatin plus ezetimibe in patients with chronic kidney disease (Study of Heart and Renal Protection): a randomised placebo-controlled trial. Lancet 2011;377(9784):2181-2192 [FREE Full text] [doi: 10.1016/S0140-6736(11)60739-3] [Medline: 21663949]

XSL•FO
RenderX

31. Department of Health. Practice Incentives Program Quality Improvement Incentive Guidelines. Canberra: Australian Government Department of Health; 2019.
32. Heselmans A, Delvaux N, Laenen A, Van de Velde S, Ramaekers D, Kunnamo I, et al. Computerized clinical decision support system for diabetes in primary care does not improve quality of care: a cluster-randomized controlled trial. Implement Sci 2020;15(1):5 [FREE Full text] [doi: 10.1186/s13012-019-0955-6] [Medline: 31910877]
33. Ivers N, Brown B, Grimshaw J. Clinical performance feedback and decision support. In: Wensing R, Grol R, Grimshaw J, editors. Improving Patient Care: The Implementation of Change in Health Care. Hoboken, NJ: Wiley-Blackwell; 2020:235-251.
34. Kwan JL, Lo L, Ferguson J, Goldberg H, Diaz-Martinez JP, Tomlinson G, et al. Computerised clinical decision support systems and absolute improvements in care: meta-analysis of controlled clinical trials. BMJ 2020;370:m3216 [FREE Full text] [doi: 10.1136/bmj.m3216] [Medline: 32943437]
35. Xie CX, Chen Q, Hincapié CA, Hofstetter L, Maher CG, Machado GC. Effectiveness of clinical dashboards as audit and feedback or clinical decision support tools on medication use and test ordering: a systematic review of randomized controlled trials. J Am Med Inform Assoc 2022;29(10):1773-1785. [doi: 10.1093/jamia/ocac094] [Medline: 35689652]
36. Patel B, Usherwood T, Harris M, Patel A, Panaretto K, Zwar N, et al. What drives adoption of a computerised, multifaceted quality improvement intervention for cardiovascular disease management in primary healthcare settings? A mixed methods analysis using normalisation process theory. Implement Sci 2018;13(1):140 [FREE Full text] [doi: 10.1186/s13012-018-0830-x] [Medline: 30419934]

Abbreviations

ACEI: angiotensin-converting enzyme inhibitor
ARB: angiotensin receptor blocker
CDS: clinical decision support
CDSS: clinical decision support software
CKD: chronic kidney disease
CPD: continuing professional development
CVD: cardiovascular disease
ECHO: Extension for Community Healthcare Outcomes
EMR: electronic medical record
FHT: Future Health Today
GEE: generalized estimating equation
GP: general practitioner
ICC: intracluster correlation
IRSD: Index of Relative Socioeconomic Disadvantage
MBS: Medicare Benefits Schedule
QI: quality improvement
SBP: systolic blood pressure

Edited by A Mavragani; submitted 31.10.23; peer-reviewed by J Walsh, J Sussman; comments to author 09.08.24; revised version received 12.09.24; accepted 03.12.24; published 03.02.25.

Please cite as:
Manski-Nankervis JA, Hunter B, Lumsden N, Laughlin A, McMorrow R, Boyle D, Chondros P, Jesudason S, Radford J, Prictor M, Emery J, Amores P, Tran-Duy A, Nelson C
Effectiveness of Electronic Quality Improvement Activities to Reduce Cardiovascular Disease Risk in People With Chronic Kidney Disease in General Practice: Cluster Randomized Trial With Active Control
JMIR Form Res 2025;9:e54147
URL: https://formative.jmir.org/2025/1/e54147
doi:10.2196/54147
PMID:

©Jo-Anne Manski-Nankervis, Barbara Hunter, Natalie Lumsden, Adrian Laughlin, Rita McMorrow, Douglas Boyle, Patty Chondros, Shilpanjali Jesudason, Jan Radford, Megan Prictor, Jon Emery, Paul Amores, An Tran-Duy, Craig Nelson. Originally published in JMIR Formative Research (https://formative.jmir.org), 03.02.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is

XSL•FO
RenderX

properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Nurses' Perspectives and Experiences of Using a Bed-Exit Information System in an Acute Hospital Setting: Mixed Methods Study

Stefan Walzer[1], MSc; Isabel Schön[2], MA; Johanna Pfeil[2], BA; Sam Klemm[3], BSc; Sven Ziegler[3], MSc; Claudia Schmoor[4], Dr rer nat; Christophe Kunze[1], Prof Dr

[1]Care and Technology Lab, Furtwangen University, Furtwangen im Schwarzwald, Germany
[2]AGP Social Research, FIVE e.V., Freiburg, Germany
[3]Center of Implementing Nursing Care Innovations Freiburg, Nursing Direction, Medical Center, University of Freiburg, Freiburg, Germany
[4]Clinical Trials Unit, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg, Germany

Corresponding Author:
Stefan Walzer, MSc
Care and Technology Lab
Furtwangen University
Robert-Gerwig-Platz 1
Furtwangen im Schwarzwald, 78120
Germany
Phone: 49 7723 920 2957
Email: stefan.walzer@hs-furtwangen.de

Abstract

Background: Technology that detects early when a patient at risk of falling leaves the bed can support nurses in acute care hospitals.

Objective: To develop a better understanding of nurses' perspectives and experiences with a bed-exit information system (BES) in an acute care hospital setting.

Methods: BES was implemented on 3 wards of a university medical center. Nurses completed 2 online surveys at each time point (P0 and P1) and participated in focus groups before (P0) and after (P1) implementation. Additional patient data were collected. Descriptive statistics summarized the survey results, while content analysis was applied to focus group data. Patient rates and adverse events in both phases were compared using negative binomial models. Reporting of this study adhered to the GRAMMS checklist.

Results: A total of 30 questionnaires were completed at P0 (30/72, 42%) and 24 at P1 (24/71, 33%). Of the participants, 15 completed both questionnaires (complete cases). At P1, 64% (9/14) of participants agreed that their perceived workload and strain in caring for patients with cognitive impairment was reduced by the use of the BES. The adverse event rate per patient per day was reduced by a factor of 0.61 (95% CI 0.393-0.955; P=.03). In addition, 11 nurses participated in 4 focus groups before and after the intervention. Participants found it challenging to operationalize the use of the BES due to the heterogeneity of care settings, but certain behaviors of patients with cognitive impairment were recognized as indicating a need for intervention. Negative experiences included information overload and alarm fatigue, leading to occasional removal of the system.

Conclusions: While BES provides some support in managing patients with cognitive impairment, its impact remains limited to specific scenarios and does not significantly reduce nurses' workload or strain. Our findings highlight the need to manage expectations of BES performance to ensure alignment between expected and actual benefits. To improve BES effectiveness and long-term implementation, future research should consider both objective measures of patient care and subjective factors such as nurse experience, structural conditions, and technical specifications. Improving information mechanisms within call systems could help reduce alarm fatigue and increase perceived usefulness. Overall, successful integration of BES in acute care settings will require close collaboration with nursing staff to drive meaningful healthcare innovation and ensure that the technology meets the needs of both patients and nurses.

Trial Registration: German Register for Clinical Studies DRKS00021720; https://drks.de/search/de/trial/DRKS00021720

XSL•FO
RenderX

(JMIR Form Res 2025;9:e64444) doi:10.2196/64444

KEYWORDS

cognitive impairment; bed-exit; technology; fall prevention; inpatient; hospital; mixed methods; nurse; information system; acute hospital; support; online questionnaire; cognitively impaired; workload

Introduction

Background

Globally, the number of older adults admitted to hospitals in emergency situations is increasing, and this trend is expected to continue as a result of current demographic changes [1]. Currently, older adults account for approximately two-thirds of all hospital inpatients, and up to 50% of this group show signs of cognitive impairment, including forms associated with dementia [2,3]. Patients with cognitive impairment often find it challenging to adhere to standardized care plans in the hospital setting [1]. They are at increased risk of developing responsive behaviors (responsive behavior is a term, which represents the effect of cognitive impairment on actions, words, and gestures of a person) and continued cognitive and functional decline [4]. Typical behaviors include an inverted day-night cycle or lack of adherence to care plans [5]. We refer to such behaviors as adverse events (AEs) in the context of nursing care when they affect the health or well-being of patients or increase the workload of nurses (see the Adverse Events in the Context of Nursing Care section). A particular challenge for nurses in hospitals is motor agitation in and around the bed, as well as related incidents such as falls or the disconnection of catheters or vascular access [6]. Kang and Hur [7] reported in a qualitative meta-synthesis that these symptoms were associated with a significantly higher workload [7]. In light of this, technological innovations have been discussed as a potential support for nurses and caregivers [8-10]. Bed-exit information systems (BES) serve as warning systems designed to monitor bed exits or patient movements in bed. This should enable nurses to prevent (or respond more quickly to) falls or other related events [11]. Nurses' perspectives and experiences on a particular technology play a central role in whether it will be successfully used and adopted over time [12]. Understanding and addressing these attitudes is essential to promoting successful and sustained technology adoption in the acute care hospital setting. In this context, little attention has been paid to nurses' perspectives and experiences on the use of a BES [13,14]. Previous studies have predominantly addressed the efficacy and effectiveness of BES in terms of reducing falls or fall rates [15,16].

Aim and Research Questions

Based on these findings, the aim of this study was to develop a better understanding of nurses' experiences, views, and needs regarding BES in an acute care hospital setting.

In terms of the aim, this study addresses the following research questions:

- What are nurses' perspectives and experiences of using a BES in an acute hospital setting?
- What is the perceived workload and strain of nurses dealing with patients with cognitive impairments and how does it change with the use of a BES?
- What is the number of AEs before and during the use of a BES?

Methods

Reporting of the study followed the Good Reporting of a Mixed Methods Study (GRAMMS) criteria proposed by O'Cathain et al [17].

Design and Setting

A parallel mixed methods triangulation design (Figure 1) with quantitative surveys and qualitative focus groups (FGs) was used to gain more insight from multiple perspectives [18]. A fundamental premise of mixed methods research is that the merging of quantitative and qualitative data provides a more complete understanding of the research question than using either data source in isolation. As defined by Creswell [19], mixed methods research is "research in which the investigator collects both quantitative (closed) and qualitative (open) data, integrates the 2, and then draws interpretations based on the strengths of both data sets to understand the research problem" [19]. Integration of methods occurred in the discussion chapter of the study, where the authors collaboratively interpreted and consolidated the results. The study was conducted in 3 wards assigned to the departments of neurosurgery, surgery, and radiology at a university medical center in southern Germany with 2 phases of data collection over a total period of 190 days. The first phase without the use of BES (P0) lasted 64 days. In the second phase (P1), a BES was available for 126 days as a tool to assist nurses in caring for patients with cognitive impairment. During the study period, 912 adult patients were treated in the first phase of the study (P0). In the second phase (P1), 1748 patients were treated.

XSL•FO
RenderX

Figure 1. Study design.

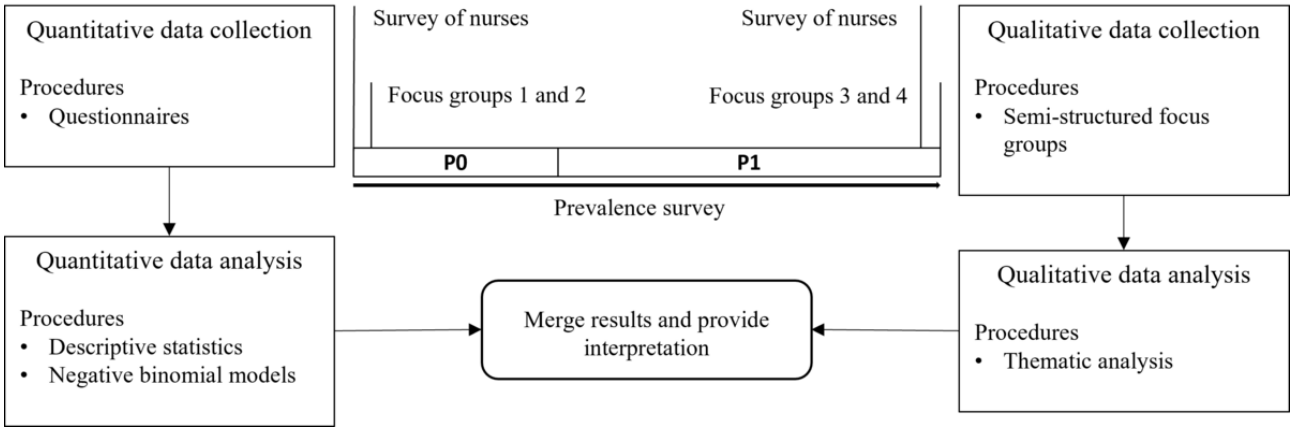

Bed-Exit Information Systems

After a rigorous, multistage selection process, the research group chose a sensor mat in the form of a mattress topper from a Swiss manufacturer. By lowering the body weight on the sensors, the device transmits information about (potential) bed exits to the call system through a radio transmitter. The manufacturer has developed this technology specifically for people with cognitive impairments such as delirium, dementia, and sleep disorders. The sensor mat, equipped with a radio transmitter, was placed under the bed mattress and the corresponding radio receiver was connected to the call system. The call system of the participating wards was not able to distinguish between alarms triggered by the sensor mat and other calls. In this study, the assessment of patient cognitive impairment was performed by nurses. The decision to use the BES was also made independently by the nurses, as not all patients with cognitive impairment exhibit symptoms such as motor agitation. Training sessions were conducted for all wards, focusing on the handling and use of the equipment according to the manufacturer's intended scenarios. These sessions were held on separate days during the handover period from the early to the late shift. If additional training was deemed necessary, additional training sessions could have been scheduled. One of the authors conducted the training sessions.

Patients With Cognitive Impairment

For this study, cognitive impairment is defined as temporary or persistent problems with mental performance, regardless of age. Typical complaints include increasing forgetfulness, decreased attention, difficulty concentrating, speech problems, disorientation, or memory loss. Cognitive impairment can occur after surgery, as part of an illness, or as a result of general age-related decline [20].

Adverse Events in the Context of Nursing Care

Responsive behaviors can lead to events that (1) disrupt acute inpatient care and can directly or indirectly endanger the affected patient or fellow patients and (2) are related to the unintentional (potential) leaving of the bed without supervision [21]. As part of the prevalence survey, the number of the following AEs was recorded: (1) finding a patient out of bed; (2) falling of a patient; (3) finding a patient in a helpless situation; (4) disconnection or dislocation of accesses, drains, or catheters; (5) use of restraints; (6) leaving the ward or the hospital unsupervised, and (7) disturbing or endangering of fellow patients. These AEs were previously identified as particularly important in 2 workshops with nurses at the university hospital, which is consistent with recent studies [1,22].

Participants

Overview

The surveys and FGs were conducted among nurses who met the inclusion criteria. The target group consisted of all nurses who met the predefined criteria listed in Textbox 1, while individuals who did not meet these criteria were excluded. At P0, the total number of registered nurses used in all wards was 72, which decreased to 70 at P1. Potential participants were informed about the surveys and the FGs during ward meetings and by email using internal institutional mailing lists and were invited to participate. Interested nurses received detailed written information about the study and what participation entailed (see the Ethical Considerations section for further details). The surveys and FGs were completed during working hours with the permission of the respective organization, and no additional incentives or rewards were offered to participants.

XSL•FO
RenderX

Textbox 1. Inclusion and exclusion criteria.

Inclusion criteria

- Age ≥18 years.
- Nurses with at least 3 years of apprenticeship or equivalent international training with professional recognition in Germany.
- Employees working on the included wards during the study period.

Exclusion criteria

- Employees of the included wards who belong to other professional groups.
- Employees of the reserve pool.
- Nurses from other wards helping out.
- Trainees.

Data Collection

The study followed a parallel mixed methods approach with the following elements:

1. Surveys and FGs of nurses at time points P0 and P1, before and after BES use
2. A prevalence survey on the number of patients with cognitive impairment and AEs related to their leaving the bed throughout the study period (P0/P1) and, in the intervention phase (P1), the number of patients with cognitive impairment who received the BES.

Surveys of Nurses

The surveys were conducted using a standardized questionnaire based on 2 questionnaires [23,24], supplemented by specific items related to the setting and technology. A total of 22 (P0) and 27 (P1) items assessed respondents' perceived strain in dealing with patients with cognitive impairment, expectations (P0) and experiences (P1) of the BES, and sociodemographic data. Items in the perceived strain category focused on AEs and the care process, which were associated with high levels of potential strain for nurses in the Isfort et al [23] study. With the exception of 2 unipolar-scaled items assessing general attitudes and acceptance of IT, items were expressed on a bipolar 7-point Likert scale.

Focus Groups

The 4 FGs were conducted by 2 moderators (SW and JP). Both moderators involved are experienced qualitative researchers, and one moderator (SW) is a registered nurse. FGs are a special type of group interview in which a moderator leads a discussion focused on a specific topic [25]. FGs are useful for obtaining information about beliefs, attitudes, thoughts, and perceptions about a particular topic from several people in a short period of time. The interaction of the participants within the group plays an important role in the generation of data [26]. A semistructured interview guide was used to structure the discussion. This guide was developed according to the recommendations of Kallio et al [27]. The structure and content of the guide are based on the current research findings mentioned above. A semistructured interview guide was chosen because it provided a framework for exploring the key concepts of interest while allowing for flexibility and openness in the discussion. During the development phase, several initial versions of the interview guide underwent multiple internal reviews and modifications by all members of the research team. Nurses were asked about their perspectives (P0) and experiences (P1), as well as barriers and opportunities for sustainable use of the BES, and the discussions were recorded and transcribed.

Prevalence Survey of Patients and Adverse Events

The prevalence survey was conducted retrospectively once a day for the last 24 hours by the nurses of the 3 wards. The number of patients with cognitive impairment and a tendency to get out of bed (first and second phases), the number of patients using a system (second phase), and the number and type of AEs that occurred in patients (first and second phases) were recorded anonymously on a documentation sheet (occupancy and fall prevalence statistics from the medical center's medical controlling and reporting systems were used as a reference for the quantitative data).

Analysis

Quantitative Data

The Clinical Trials Unit at the University Medical Center analyzed the quantitative data. Intraindividual changes in nurses' responses between phases were presented. AE rates of both phases were compared using negative binomial models [28]. Three different analyses were performed: (1) a raw overall comparison of both phases without considering individual wards, (2) a comparison of both phases separately for the 3 wards, and (3) an overall comparison of both phases adjusted for differences between wards. AE rates and multiplicative phase comparisons (ratio of AE rates in P1 vs P0) were obtained from the negative binomial models with 95% CIs, and tests of the hypothesis of equality of AE rates in P0 and P1 were performed at a 2-sided 5% significance level.

Qualitative Data

Content analysis was applied using a deductive-inductive approach [29]. The transcripts were read by 3 members of the research team (SW, IS, and JP) to familiarize themselves with the data. Codes were created through condensation and then abstracted into categories. The coders (SW and JP) reviewed the codes and categories, discussed any discrepancies, and made revisions. Finally, narrative summaries were written for each category, and thematic descriptions for the main categories were developed based on these summaries.

XSL•FO
RenderX

Setting

The study was conducted at the University Medical Center Freiburg.

Ethical Considerations

Approval was obtained from the Ethics Committee of the University of Freiburg (106/20 [MPG §23b]) and the Staff Council. Before the start of the study, all potential participants were provided with a comprehensive written document containing information about the study, including its purpose, objectives, procedures, and data protection measures. Participants were informed of the decision of the responsible committees and were given the opportunity to contact the investigators with any questions throughout the study. For participants who completed the surveys, they were explicitly informed at the beginning of the questionnaire that by submitting their responses they were indicating that they had read and understood the enclosed information letter and were giving their informed consent to participate in the survey. In the case of the FG, written informed consent was obtained from each participant at the beginning of each group session, confirming that they had been informed of the study details and were willing to participate in the study. Participation in the surveys and FGs was voluntary, no personal information was collected, and anonymity (surveys) or pseudonymization (FGs) was always maintained.

Results

Results of the Surveys

In total, 30 questionnaires were completed at P0 (30/72, 42%) and 24 at P1 (24/71, 33%). Among the total participants, 15 participants completed both questionnaires (complete cases). The sociodemographic characteristics of the participants are shown in Table 1. Approximately two-thirds of the respondents had at least 15 years of general nursing experience and the same amount of experience working with patients with cognitive impairment.

Table 1. Sociodemographic characteristics.

Characteristics	Before implementation (P0; n=30), n (%)	After implementation (P1; n=24), n (%)
Gender		
Women	20 (67)	12 (50)
Men	10 (33)	10 (42)
Age (years)		
≤40	10 (33)	9 (38)
>40	20 (67)	14 (58)
General work experience (years)		
≤15	8 (27)	6 (25)
>15	22 (73)	16 (67)
Work experience with patients with cognitive impairment (years)		
≤15	10 (33)	8 (33)
>15	20 (67)	15 (63)

The following results refer to the complete cases to allow a comparative view.

General Attitude Toward (Information) Technologies

The survey conducted at P0 showed that nurses generally had a positive attitude toward IT. This was reflected in the finding that 90% (27/30) of respondents recognized the benefits of IT, agreeing that it offered significant benefits and considering it essential in today's world.

Despite this overall positive outlook, some nurses expressed concerns and fears about using such technology. Overall, 14% (4/30) admitted that they feared making irreversible mistakes when using technology. In addition, 14% (4/30) found most technology-related issues challenging and 7% (2/30) said they were frightened by the possibility of using technology they had never worked with before.

The following results refer to the complete cases to allow a comparative view. Multimedia Appendix 1 shows that the participants rated their experience in dealing with IT or acceptance of IT rather positively (experience: median 7 (IQR 6-8); acceptance: median 8 (IQR 7-9).

Perceived Strain

The vast majority of participating nurses felt moderately to very strongly strained by patients with cognitive impairment, both before and after the use of BES (Figure 2). For example, 75% (9/12) in P0 and 100% (12/12) in P1 of the participating nurses felt strained by not being able to do right by the patients, and 92% (11/12) in P0 and P1 felt strained by the fact that patients with cognitive impairment could leave the bed unattended.

XSL•FO
RenderX

Figure 2. Perceived strain (ranging from moderate to very high).

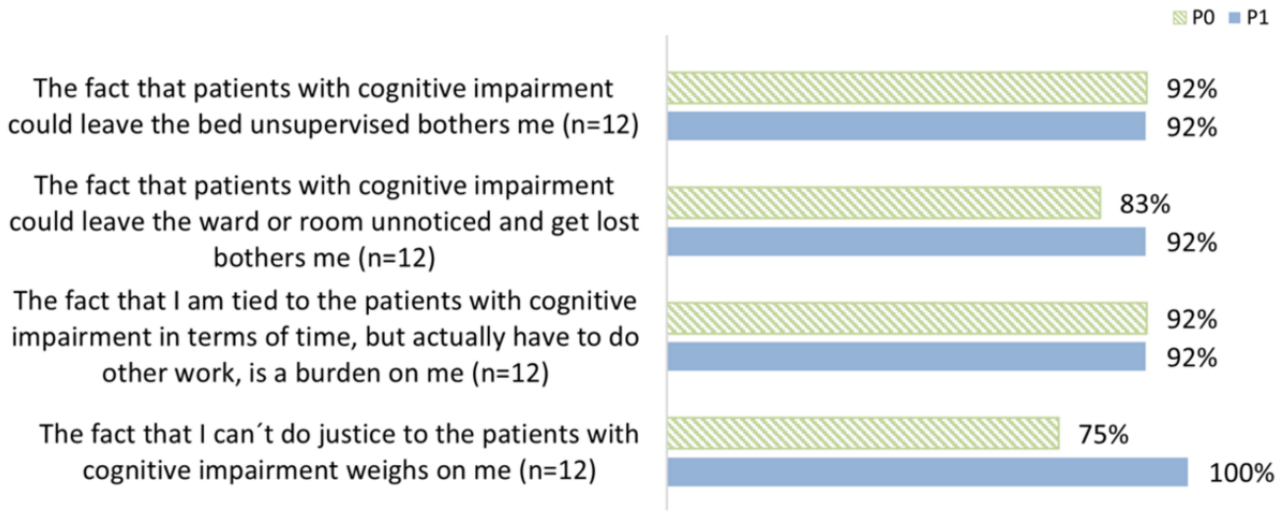

Expectations and Experiences Regarding the Use of Bed-Exit Information Systems

Among the nurses who participated in both surveys, their expectations regarding the use of BES did not match their experiences in most categories (Figure 3). For example, after the intervention, 64% (93%, P0) agreed that BES can have a positive effect on the management of patients with cognitive impairment and that daily work becomes easier, and 79% (93%, P0) expected that BES could be useful.

Figure 3. Expectations and experiences regarding the use of technology (statements ranging from somewhat agree to strongly agree).

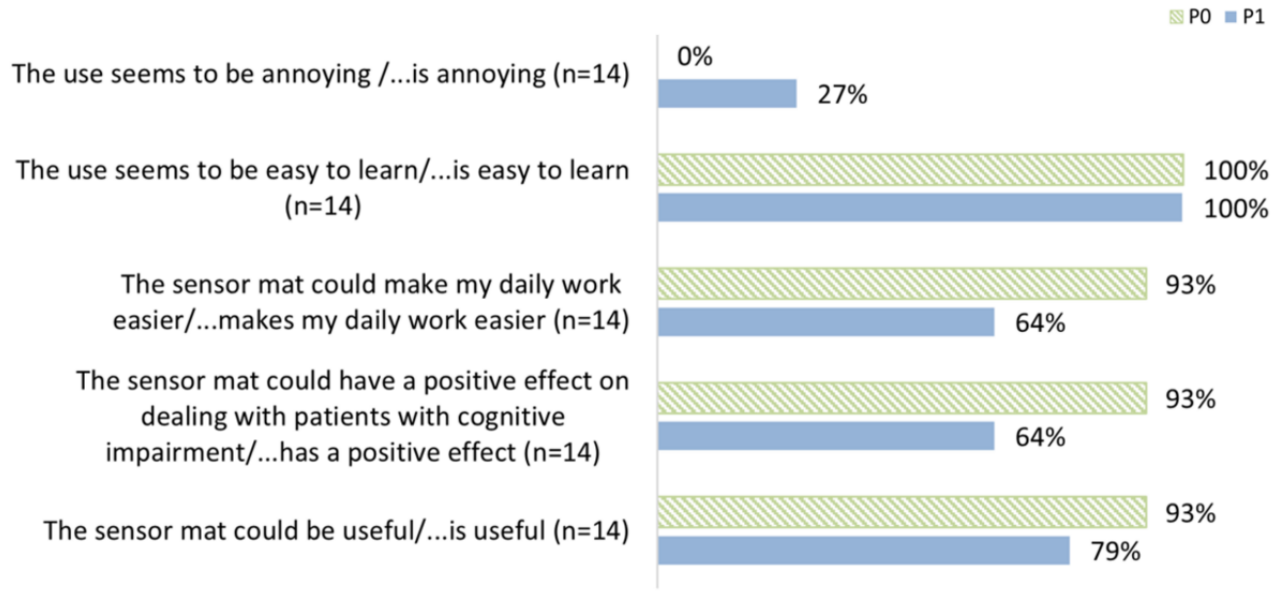

Results of the Prevalence Survey on Adverse Events

The collection of data through the questionnaires took place around 2 PM and referred to the period of the previous 24 hours. Analysis of the prevalence survey showed that in P0, patients with cognitive impairment were treated on 78% of the survey days (199/256). In P1, patients with cognitive impairment were treated on 60% of survey days (302/504), and BES was used by nurses on 20% of survey days (103 of 504). On 79 of 199 days in P0 (40%), 222 AEs were recorded in patients. In P1, 253 AEs were recorded on 82 of 302 days (27%).

The visualization in Figure 4 illustrates the occurrence of each defined AE. For each day, whether an event occurred at least once or not was recorded. It is important to note that the frequencies of the 3 categories presented can only be effectively compared within their respective contexts, as both the duration and the number of patients for each category vary. The most frequent event is the discovery of patients out of bed. This event may be due to a previous fall that left the patient in a vulnerable state. Conversely, a comparatively rare but meaningful event involves patients leaving the ward or clinic unattended, posing a potential risk of disorientation and “getting lost.”

XSL•FO
RenderX

Figure 4. Adverse events (type and number).

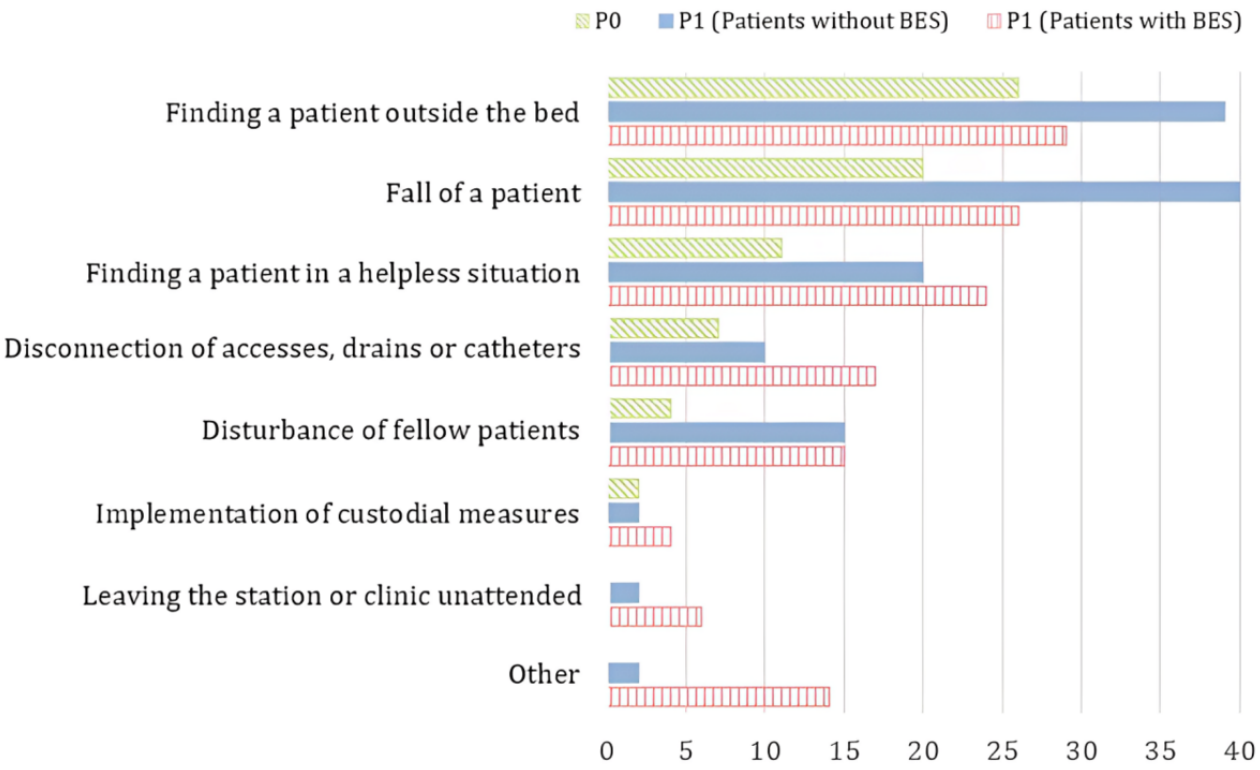

The average AE rates per patient per day in P0 and P1 and the ratio of AE rates in P1 versus P0 with 95% confidence intervals from the raw overall comparison without considering the individual wards shown in Table 2.

Table 2. Adverse events in the context of nursing care rates per patient and day in P0 and P1.

Phase	AE[a] rate (95% CI)	*P* value
0	0.49 (0.36-0.67)	—[b]
1	0.35 (0.27-0.46)	—
1 vs 0	0.73 (0.48-1.09)	.12

[a]AE: adverse event.

[b]Not applicable.

Table 3 provides a comparison of AE rates between P0 and P1 separately for the three wards.

XSL•FO
RenderX

Table 3. Adverse event rates per patient per day with interaction between phase and ward.

Ward	Phase	AE[a] rate (95% CI)	*P* value
1	0	0.29 (0.17-0.49)	—[b]
1	1	0.57 (0.39-0.82)	—
1	1 vs 0	1.97 (1.03-3.75)	.04
2	0	0.28 (0.13-0.60)	—
2	1	0.26 (0.08-0.08)	—
2	1 vs 0	0.92 (0.23-3.70)	.91
3	0	0.70 (0.47-1.05)	—
3	1	0.20 (0.13-0.29)	—
3	1 vs 0	0.28 (0.16-0.50)	<.001

[a]AE: adverse event; in the context of nursing care.

[b]Not applicable.

The picture was different for each ward. While ward 1 had a relatively low AE rate of 0.29 (95% CI 0.17-0.49) in P0 and increased by a factor of 1.97 (95% CI 1.03-3.75) in P1, ward 2 showed similarly low AE rates in P0 and P1. In ward 3 with 125 patient days in P0 and 248 patient days in P1, the relatively high AE rate in P0 decreased from 0.70 (95% CI 0.47-1.05) in P1 to 0.20 (95% CI 0.13-0.29). The statistical test for interaction between phase and ward showed a *P* value of .00. The overall comparison of AE rates between P0 and P1, adjusted for differences between the wards, is shown in Table 4.

Table 4. Adjusted adverse event rates per patient per day in P0 and P1.

Phase	AE[a] rate (95% CI)	*P* value
0	0.47 (0.33-0.67)	—[b]
1	0.29 (0.21-0.41)	—
1 vs 0	0.61 (0.39-0.95)	.03

[a]AE: adverse event; in the context of nursing care.

[b]Not applicable.

Thus, in a summary comparison of AE rates per cognitively impaired patient per day on each ward, there was a reduction in the AE rate by a factor of 0.613 (95% CI 0.393-0.955; *P*=.03).

Results of the FGs

A total of 11 nurses participated in the FG*s*. Two FGs with similar questions were held before (P0) and 2 were held after the use of the BES (P1). Each discussion lasted for 2 hours. The group size varied from 2 (see limitations for more details) to 5 participants, and 3 nurses participated in both P0 and P1. The key findings from the thematic descriptions are presented: (1) challenges in the use of BES and (2) opportunities in the use of BES; once from an expectation perspective (P0) and once from an experience perspective (P1).

Challenges

The nurses who participated in the preintervention FG identified several challenges in using BES. In the hospital setting, the exact causes of AEs are often not (or not fully) understood (also due to lack of biographical information). Therefore, concrete scenarios for the use of BES remain vague in advance. Consequently, the most specific expectations relate to the institutional framework: it is assumed that the system can be used in cases where continuous monitoring is needed, but no constant observers (sitters) and no other strategies (such as transfer to the intensive care unit) are available. Of course, nurses have also developed strategies to deal with AE independently of technical aids. Nontechnical alarms and safeguards such as mattresses on the floor in front of beds or plastic cups on door handles are used. However, the desire for sitters, who participants felt could best counteract AE, is mentioned very often.

> *[…] so, I see in general such a mat, a mat for movements ... I think, that [the BES] is the wrong way, because we actually need more people in the clinic […]* [P0_FG1]

After using the BES (P1), the participants confirmed that the diversity of conditions and symptoms makes operationalization a challenge. It became clear that the question, “Who gets a BES and for what reasons?” is not easy to answer. The heterogeneity of care situations also results in a variety of situations that are challenging for caregivers. Nevertheless, it was mostly specific behaviors of the patients with cognitive impairment that indicated intervention. One nurse summarized this as follows:

> *The cognitively altered [people], who are too mobile for the bed rail to protect [them]. [...] If this is someone who can perhaps still move a foot or even only half of it, he doesn’t necessarily need a sensor mat, because he can’t stand up anyway, yes. But [the*

XSL•FO
RenderX

> *one who is] just cognitively impaired, still very lively, very mobile, agile: he gets a mat.* [P1_FG1]

The discussions also revealed negative experiences with BES. Depending on the workload of the nurses, it became clear that the frequency of information could lead to delays and additional workload. After an initial rush to get information from the BES, information or alarm fatigue set in due to the frequency of information, and the systems were sometimes removed because nurses saw more disadvantages than benefits.

> […] *They said: 'Oh, that'll be the mat again.' But I mean - I think you still have to look. You can't say: 'My God, it's just the mat.' And just leave it there.* [P1_FG1]

Opportunities

Before using the BES (P0), the nurses discussed the need for more staff and sitters as mentioned earlier. Since sitters are not always available, despite the preference for sitters, a technological device such as the BES is seen as an alternative, especially at night.

> *And then we said: 'well better now than never' - sounds interesting and if it [BES] is of any use something […] then we'll cope with that. Because, I think for us, at night or also during the day depending on / - because sitters, that is another problem, you don't always get them and if this [BES] is of any help: great.* [P0_FG1]

As intended by the manufacturer, the BES is used for wandering patients. However, the participating nurses differentiated between whether the person felt a "harmless" urge to move, for example, to walk down the hallway, or whether there was an acute risk of falling. Participants indicated that BES was useful when patients with cognitive impairment were at risk of being lost if they left the building quickly and purposefully. In this context, BES has special potential at night.

> *I actually have the feeling that the situation [...] with people who absolutely have to go to the toilet at night and then they can't find the light - they don't even think about the bell-the night guards have already benefited from it and they also went into the room more quickly at night because they knew there was a mat in there […].* [P1_FG2]

After using the BES, the participating nurses were mostly positive about the reliability of the alarms and considered it acceptable that moderate technical problems occurred in a few cases, for example, false-positive alarms. From their point of view, the BES is a technical aid that can be used in individual situations, such as pressure mattresses or patient lifts, and as such has its place in everyday care.

> *I put that in once, plug it in and then it's there. When I put an infusion pump or an ultrasound or something else into the room, it's there, too. Maybe I'm too naive now, but of course it fits into the daily nursing routine. Because, as we have now experienced, it does prevent falls in some cases. Of course, it fits in there, for me.* [P1_FG1]

Even before use, participants identified potential support in the early detection of AE. This can help give them a sense of security and safety. BES can support and relieve nurses by alerting them to certain critical situations. One nurse summarized this as follows:

> *Then I already know how fast I can run. But that's the good thing, that I actually already know, ok, I don't hear it rumbling and then run, but it rings and I run and it doesn't rumble.* [P0_FG1]

This potential was also perceived by participants after use. In certain situations, the advantages of the BES may outweigh the disadvantages, such as constant alerting and the associated additional workload.

> *So, it was little more relaxed. Because you stop-if it rang, then there was something. Then you could go, more or less calmly. Sure, it rang all the time, but there wasn't this uncertainty of, 'What now?'*

Nonetheless, the question of overall benefit remains unanswered. Although nurses have had positive experiences, some question whether this technology is really what they need to improve their overall work situation and patient care.

Discussion

Principal Findings

This study used a parallel mixed methods triangulation design to explore nurses' perspectives and experiences of a BES in an acute care hospital. The BES was implemented in 3 wards in a university medical center. Data collection included standardized online surveys completed by nurses at two time points: before (P0) and after (P1) the implementation of the system. In addition, FGs were conducted at both time points, using a semistructured interview guide. Retrospective prevalence surveys anonymously documented patient-related data, including cognitive impairment, tendencies to get out of bed, and AEs such as falls, over 24-hour periods.

The results clearly show the emotional and psychological strain nurses experience when caring for and supporting patients with cognitive impairment. The reported workload was significantly higher than in a survey on the care of people with dementia in hospitals by Isfort et al [23]. In the survey, 66.9% of participants reported high levels of distress when patients with cognitive impairment were allowed to leave the ward unnoticed [23]. The same question in our survey resulted in over 90% agreement. The other statements also showed significantly higher reported distress. Although this is not a representative cross-section, the results are worth discussing. One explanation could be the different target group of respondents. Isfort et al [23] surveyed senior nurses who were at least no longer working full-time "at the bedside." The respondents in the current study were predominantly nurses working in direct patient care, so a greater potential for distress might be expected in this context. Another explanation may be that the participating units generally have an above-average number of patients with cognitive impairment. The nurses' general attitude and acceptance of IT and communication technology are positive. This is also consistent with the results of an online survey conducted by the German Center for Quality in Care [30].

XSL•FO
RenderX

Nearly half of nurses reported that the BES made their daily work easier or much easier. Depending on the situation, it can increase both patient safety and nurses' sense of safety. A BES can be a tool, but, like all technological innovations in highly complex environments, it does not solve complex, interrelated problems [12,31]. A BES can provide a sense of safety and increased security in selected situations, individually adapted to the setting and the patients. Looking at the complete cases, the experiences with the BES deviate negatively from the expectations. For example, the positive effect in dealing with patients with cognitive impairment, which 93% (13/14) of participants expected before using the BES, was observed by only 64% (9/14) of the participants after using the BES. This finding suggests that the experience of using the BES did not live up to previous expectations; helpfulness fell short of expectations. One way to deal with this is to manage expectations in advance (also on the part of the manufacturer) [32]. Whether the final perceived usefulness is sufficient for further use must be decided based on the fit between individual needs and benefits [33]. An improvement in perceived usefulness could possibly be achieved by implementing differentiated information mechanisms through the call system.

The most common AEs are getting patients out of bed and falls. A total of 66.9% (P0) and 59.9% (P1) of the respondents reported them as very stressful. These results are consistent with the findings of Isfort et al [23], who showed that these two events are the most stressful factors for nurses. Comparing the results of the cognitive impairment prevalence survey and AEs in the 2 phases, the rate of AEs per patient day decreased by a factor of 0.6. However, there were significant differences between the wards, with sometimes conflicting results. One explanation could be that the clinical conditions and comorbidities of the patients differ, which could lead to differences in the ability to use the BES very accurately.

Bed-leaving tendencies do not stand alone, but must always be evaluated in the context of other individual patient behaviors as well as the conditions of the particular ward. Focusing on a specific tool necessarily takes into account only a small part of the picture, which does not exist in this form for the nurses participating in the FGs. The situations described in the discussions show the complexity of supporting nurses in dealing with patients with cognitive impairment. Previous studies have predominantly addressed the efficacy and effectiveness of BES in terms of reducing falls or fall rates [15,34,35]. As a result, the primary focus of research has been on evaluating the functionality of the technology. However, there are many other factors that determine whether a technology will be successfully used and adopted over time in a particular setting [12,31]. This study provides a broader picture of the potential benefits of BES in the care and support of patients with cognitive impairments from a nurse's perspective.

The analysis of the FGs shows a generally open and pragmatic attitude toward technology in nursing. The benefits of BES from the nurses' point of view can be, for example, physical relief (eg, fewer rounds at night), psychological relief (eg, a greater sense of security), or an improvement in the quality of care (eg, avoidance of patient injury). At the same time, it is important to avoid new challenges, such as high alarm rates and limited response options, which can lead to alarm fatigue. In light of the survey results showing high expectations of the benefits of the BES and low experience of the benefits, the FGs showed that the high expectations were related to the reduction of falls. In the nurses' perception, this goal was not fully achieved. However, it helped to cope with the consequences and was therefore helpful in reducing stress. The FGs show that the BES is best suited for use at night and with patients at risk of falling and with mild cognitive impairment. This specification of use scenarios makes it possible to obtain benefits in the form of fall prevention or rapid response to falls without being overloaded by too many alarms. Our results suggest that the effectiveness of the intervention appears to be strongly influenced by several factors, such as the individual condition of the patient, the staff-to-patient ratio, and the timing of use.

Strengths and Limitations

This paper discusses the implementation of a specific BES. Due to the different systems and implementation aspects, it is not possible to generalize the results. The sample size, 30 people in P0, 24 in P1 and 15 complete cases, is a limitation for complex statistical procedures. However, the combination of quantitative and qualitative methods provides a comprehensive perspective and deeper insights into the research topic. We collected data for the prevalence survey using documentation forms at around 2 PM, referring to the previous 24-hour period. Overlaps occurred on 3 days per ward (study start, transition between study phases, and study end), but we ensured that patient and case counts were accurate to the day. The heterogeneous patient population, with complex and fluctuating symptom patterns and congruent but not equivalent wards, introduces potential bias in the evaluation of the prevalence survey. The significant reduction in the AE rate suggests that the use of the BES may be beneficial, although it does not allow causal conclusions. We must remember that leaving the bed unattended was also considered an AE. An increase in the AE rate could indicate a reduction in underreporting, paradoxically a desirable outcome. A more precise numerical presentation of the different AEs would have been helpful. Another limitation is the different conditions of use of the technology in different ward designs, which limits the direct applicability of the results to other wards. Since the focus of this study was to capture the perspectives of nursing staff specifically, the diversity of participants was limited by the study design. Future research could, however, consider including a broader range of health care professionals or expanding the diversity of nursing staff to explore how different perspectives, such as those from interdisciplinary teams, may influence the outcomes of BES implementation. Due to recruitment difficulties, we conducted a postintervention FG with only 2 participants, which may have changed the dynamics of the discussion, making it more like a dyadic interview. There are some similarities and differences between these methods [36]. In our case, each group discussion maintained the same thematic focus and similar dynamics.

Conclusions

Our results indicate that while the BES provides some assistance in managing patients with cognitive impairment, its impact is limited to specific scenarios and does not significantly reduce

XSL•FO
RenderX

nurses' workload or strain. In addition, our study highlights the importance of managing expectations regarding BES performance. Setting realistic expectations for both manufacturers and end users is essential to bridge the gap between expected and experienced benefits. To better evaluate the benefits of BES for nurses and increase the likelihood of successful long-term implementation, future studies should address not only objectively measurable changes in patient care but also subjective perspectives, structural factors, and technical specifications. Technical improvements, such as differentiated information mechanisms through the call system, could enhance its perceived usefulness and help reduce alarm fatigue. By examining these factors, a more comprehensive understanding of the BES's impact can be achieved, supporting its appropriate implementation in acute care settings. Finally, these findings emphasize the importance of integrating technology like the BES through close collaboration with nursing staff. Such collaboration is crucial for driving effective health care innovation and ensuring that new technologies meet the needs of both patients and nurses.

Acknowledgments

Our study team is particularly grateful to all the nurses in the participating wards for their cooperation. We would like to thank our colleagues and key persons of our project partners at the Center of Implementing Nursing Care Innovations, Freiburg, Germany. This joint project on the development and research of new care technologies in Germany is funded by the Federal Ministry of Education and Research. At the Freiburg Center, the focus is on integrating and evaluating innovative nursing technologies in the acute care setting (grants 16SV7888, 16SV7889, and 16SV7886K).

Authors' Contributions

SW contributed to conceptualization, methodology, investigation, formal analysis, data curation, and writing-original draft. IS was involved in conceptualization, methodology, investigation, formal analysis, data curation, and writing-original draft. JP contributed to data curation-reviewing and editing. SS contributed to the conceptualization, methodology, data curation, and investigation-reviewing and editing. SZ was involved in conceptualization, methodology, and investigation-reviewing and editing. CS contributed to conceptualization, supervision, resources-reviewing and editing. CK contributed to the conceptualization, supervision, resources-reviewing and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Self-assessment of general experience and acceptance (minimum 1, maximum 10).
[PNG File , 83 KB - formative_v9i1e64444_app1.png]

References

1. Mudge AM, Lee-Steere K, Treleaven E, Cahill M, Finnigan S, McRae P. Cognitive impairment in older hospital inpatients: prevalence, care needs and carer perceptions. Aust Health Rev 2022;46(2):244-250. [doi: 10.1071/AH20286] [Medline: 34856117]
2. Fogg C, Griffiths P, Meredith P, Bridges J. Hospital outcomes of older people with cognitive impairment: an integrative review. Int J Geriatr Psychiatry 2018;33(9):1177-1197 [FREE Full text] [doi: 10.1002/gps.4919] [Medline: 29947150]
3. Pinkert C, Faul E, Saxer S, Burgstaller M, Kamleitner D, Mayer H. Experiences of nurses with the care of patients with dementia in acute hospitals: a secondary analysis. J Clin Nurs 2018;27(1-2):162-172. [doi: 10.1111/jocn.13864] [Medline: 28426891]
4. Dookhy J, Daly L. Nurses' experiences of caring for persons with dementia experiencing responsive behaviours in an acute hospital: a qualitative descriptive study. Int J Older People Nurs 2021;16(4):e12379. [doi: 10.1111/opn.12379] [Medline: 33963674]
5. Motzek T, Büter K, Mächler K, Junge M, Marquardt G. Orientierungsschwierigkeiten, behaviorale und psychiatrische Symptome bei patienten mit kognitiven Einschränkungen im Krankenhaus. HBScience 2017;8(2):46-51. [doi: 10.1007/s16024-017-0297-8]
6. Choi SSW, Cajita MI, Gitlin LN. A review of measures of three common dementia-related behaviors: rejection of care, aggression, and agitation. Geriatr Nurs 2020;41(6):692-708. [doi: 10.1016/j.gerinurse.2020.04.003] [Medline: 32402574]
7. Kang Y, Hur Y. Nurses' experience of nursing workload-related issues during caring patients with dementia: a qualitative meta-synthesis. Int J Environ Res Public Health 2021;18(19):10448 [FREE Full text] [doi: 10.3390/ijerph181910448] [Medline: 34639748]
8. Torres RLS, Visvanathan R, Hoskins S, van den Hengel A, Ranasinghe D. Effectiveness of a batteryless and wireless wearable sensor system for identifying bed and chair exits in healthy older people. Sensors (Basel) 2016;16(4):546 [FREE Full text] [doi: 10.3390/s16040546] [Medline: 27092506]

XSL•FO
RenderX

9. Vandenberg AE, van Beijnum BJ, Overdevest VGP, Capezuti E, Johnson TM. US and Dutch nurse experiences with fall prevention technology within nursing home environment and workflow: a qualitative study. Geriatr Nurs 2017;38(4):276-282 [FREE Full text] [doi: 10.1016/j.gerinurse.2016.11.005] [Medline: 27956058]
10. Husebo BS, Heintz HL, Berge LI, Owoyemi P, Rahman AT, Vahia IV. Sensing technology to monitor behavioral and psychological symptoms and to assess treatment response in people with dementia. a systematic review. Front Pharmacol 2019;10:1699 [FREE Full text] [doi: 10.3389/fphar.2019.01699] [Medline: 32116687]
11. Seow JP, Chua TL, Aloweni F, Lim SH, Ang SY. Effectiveness of an integrated three-mode bed exit alarm system in reducing inpatient falls within an acute care setting. Jpn J Nurs Sci 2022;19(1):e12446. [doi: 10.1111/jjns.12446] [Medline: 34286920]
12. Greenhalgh T, Wherton J, Papoutsi C, Lynch J, Hughes G, A'Court C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017;19(11):e367 [FREE Full text] [doi: 10.2196/jmir.8775] [Medline: 29092808]
13. Hubbartt B, Davis SG, Kautz DD. Nurses' experiences with bed exit alarms may lead to ambivalence about their effectiveness. Rehabil Nurs 2011;36(5):196-199. [doi: 10.1002/j.2048-7940.2011.tb00195.x] [Medline: 21882797]
14. Wong Shee A, Phillips B, Hill K, Dodd K. Feasibility, acceptability, and effectiveness of an electronic sensor bed/chair alarm in reducing falls in patients with cognitive impairment in a subacute ward. J Nurs Care Qual 2014;29(3):253-262. [doi: 10.1097/NCQ.0000000000000054] [Medline: 24487696]
15. Cuttler SJ, Barr-Walker J, Cuttler L. Reducing medical-surgical inpatient falls and injuries with videos, icons and alarms. BMJ Open Qual 2017;6(2):e000119 [FREE Full text] [doi: 10.1136/bmjoq-2017-000119] [Medline: 29450285]
16. Røyset B, Talseth-Palmer BA, Lydersen S, Farup PG. Effects of a fall prevention program in elderly: a pragmatic observational study in two orthopedic departments. Clin Interv Aging 2019;14:145-154 [FREE Full text] [doi: 10.2147/CIA.S191832] [Medline: 30697039]
17. O'Cathain A, Murphy E, Nicholl J. The quality of mixed methods studies in health services research. J Health Serv Res Policy 2008;13(2):92-98. [doi: 10.1258/jhsrp.2007.007074] [Medline: 18416914]
18. Ingham-Broomfield B. A nurses' guide to mixed methods research. Aust J Adv Nurs 2016;33(4):46-52. [doi: 10.37464/2016.334.1559]
19. Creswell JW. A Concise Introduction to Mixed Methods Research. Washington DC: SAGE Publications; 2021:1-168.
20. Roy E. Cognitive Impairment. In: Gellman MD, Turner R, editors. Encyclopedia of Behavioral Medicine. New York, NY: Springer New York, NY; 2013:449-451.
21. Watkin L, Blanchard MR, Tookman A, Sampson EL. Prospective cohort study of adverse events in older people admitted to the acute general hospital: risk factors and the impact of dementia. Int J Geriatr Psychiatry 2012;27(1):76-82. [doi: 10.1002/gps.2693] [Medline: 21360591]
22. Bickel H, Hendlmeier I, Heßler JB, Junge MN, Leonhardt-Achilles S, Weber J, et al. The prevalence of dementia and cognitive impairment in hospitals. Dtsch Arztebl Int 2018;115(44):733-740 [FREE Full text] [doi: 10.3238/arztebl.2018.0733] [Medline: 30565543]
23. Isfort M, Klostermann J, Gehlen D, Siegling B. Pflege-Thermometer 2014 Eine bundesweite Befragung von leitenden Pflegekräften zur Pflege und Patientenversorgung von Menschen mit Demenz im Krankenhaus. Köln: Deutsches Institut für angewandte Pflegeforschung e.V; 2014.
24. Spagnolli A, Guardigli E, Orso V, Varotto A, Gamberini L. Measuring user acceptance of wearable symbiotic devices: validation study across application scenarios. In: Jacucci G, Gamberini L, Freeman J, Spagnolli A, editors. Symbiotic Interaction. Cham, Switzerland: Springer International Publishing; 2014:978-973.
25. Rädiker S, Kuckartz U. Analyzing qualitative data with MAXQDA. In: Text, Audio, and Video. Cham, Switzerland: Springer; 2019:XIII-X290.
26. Plummer P. Focus group methodology. Part 1: design considerations. Int J Ther Rehabil 2017;24(7):297-301. [doi: 10.12968/ijtr.2017.24.7.297]
27. Kallio H, Pietilä AM, Johnson M, Kangasniemi M. Systematic methodological review: developing a framework for a qualitative semi-structured interview guide. J Adv Nurs 2016;72(12):2954-2965. [doi: 10.1111/jan.13031] [Medline: 27221824]
28. Baxter A, Huddleston E. SAS/STAT® 15.1 User?s Guide. Cary, NC, USA: SAS Institute Inc; 2018.
29. Kuckartz U. Qualitative text analysis: a systematic approach. In: Kaiser G, editor. Compendium for early career researchers in mathematics education. Cham, Switzerland: Springer; 2019:181-197.
30. Kuhlmey A, Blüher S, Nordheim J, Zöllick J. Technik in der Pflege?Einstellungen von professionell Pflegenden zu Chancen und Risiken neuer Technologien und technischer Assistenzsysteme. Zentrum für Qualität in der Pflege, Berlin 2019.
31. Walzer S, Armbruster C, Kunze C, Farin-Glattacker E. NASSS-CAT-D: Komplexität bei der Einführung von Gesundheitstechnologien erkennen, verstehen und beherrschen. In: Krick T, Zerth J, Rothgang H, Klawunn R, Walzer S, Kley T, editors. Pflegeinnovationen in der Praxis. Wiesbaden, Italy: Springer Gabler; 2023:17-35.
32. Hebb A, Kistler M, George E, Zamboni B. Satisfaction and technology acceptance of staff regarding use of continuous video monitoring in comparison with sitters. J Nurs Adm 2021;51(2):60-62. [doi: 10.1097/NNA.0000000000000970] [Medline: 33449593]

XSL•FO
RenderX

33. Tubaishat A. Perceived usefulness and perceived ease of use of electronic health records among nurses: application of technology acceptance model. Inform Health Soc Care 2018;43(4):379-389. [doi: 10.1080/17538157.2017.1363761] [Medline: 28920708]
34. Balaguera HU, Wise D, Ng CY, Tso HW, Chiang WL, Hutchinson AM, et al. Using a medical intranet of things system to prevent bed falls in an acute care hospital: a pilot study. J Med Internet Res 2017;19(5):e150 [FREE Full text] [doi: 10.2196/jmir.7131] [Medline: 28473306]
35. Barker AL, Morello RT, Wolfe R, Brand CA, Haines TP, Hill KD, et al. 6-PACK programme to decrease fall injuries in acute hospitals: cluster randomised controlled trial. BMJ 2016;352:h6781 [FREE Full text] [doi: 10.1136/bmj.h6781] [Medline: 26813674]
36. Szulc J, King N. The practice of dyadic interviewing: strengths, limitations and key decisions. Forum Qual Soc Res 2022;23(2). [doi: 10.17169/fqs-22.2.3776]

Abbreviations

AE: adverse event
BES: bed-exit information system
FG: focus group
GRAMMS: Good Reporting of a Mixed Methods Study

Edited by A Mavragani; submitted 28.07.24; peer-reviewed by J Walsh; comments to author 30.10.24; revised version received 27.11.24; accepted 09.12.24; published 05.02.25.

Please cite as:
Walzer S, Schön I, Pfeil J, Klemm S, Ziegler S, Schmoor C, Kunze C
Nurses' Perspectives and Experiences of Using a Bed-Exit Information System in an Acute Hospital Setting: Mixed Methods Study
JMIR Form Res 2025;9:e64444
URL: https://formative.jmir.org/2025/1/e64444
doi:10.2196/64444
PMID:

©Stefan Walzer, Isabel Schön, Johanna Pfeil, Sam Klemm, Sven Ziegler, Claudia Schmoor, Christophe Kunze. Originally published in JMIR Formative Research (https://formative.jmir.org), 05.02.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Status of Digital Health Technology Adoption in 5 Vietnamese Hospitals: Cross-Sectional Assessment

Duc Minh Tran[1], PMD; Nguyen Thanh Dung[2*], MD; Chau Minh Duc[3*], MD, PhD; Huynh Ngoc Hon[4*], MD; Le Minh Khoi[5*], MD, PhD; Nguyen Phuc Hau[2*], BSc; Duong Thi Thu Huyen[6*], BSc; Huynh Thi Le Thu[4*], BSc; Tran Van Duc[5*], BSc; VITAL (Vietnam ICU Translational Applications Laboratory) Investigators[7*]; Lam Minh Yen[1], MD; C Louise Thwaites[1,8], BSc, MBBS, MD; Chris Paton[9,10], BMBS, BMedSci, MBA

[1]Oxford University Clinical Research Unit, Ho Chi Minh City, Vietnam
[2]Hospital for Tropical Diseases, Ho Chi Minh City, Vietnam
[3]Dong Thap General Hospital, Dong Thap, Vietnam
[4]Trung Vuong Hospital, Ho Chi Minh City, Vietnam
[5]University Medical Center, Ho Chi Minh City, Vietnam
[6]Department of Information Technology, National Hospital for Tropical Diseases, Hanoi, Vietnam
[7]See Acknowledgements,
[8]Centre for Tropical Medicine and Global Health, University of Oxford, Oxford, United Kingdom
[9]Nuffield Department of Medicine, University of Oxford, Oxford, United Kingdom
[10]Department of Information Science, University of Otago, Dunedin, New Zealand
[*]these authors contributed equally

Corresponding Author:
Duc Minh Tran, PMD
Oxford University Clinical Research Unit
764 Vo Van Kiet Street, Ward 1, District 5
Ho Chi Minh City
Vietnam
Phone: 84 2839237954
Email: ductm@oucru.org

Abstract

Background: Digital health technologies (DHTs) have been recognized as a key solution to help countries, especially those in the low- and middle-income group, to achieve the Sustainable Development Goals (SDGs) and the World Health Organization's (WHO) Triple Billion Targets. In hospital settings, DHTs need to be designed and implemented, considering the local context, to achieve usability and sustainability. As projects such as the Vietnam ICU Translational Applications Laboratory are seeking to integrate new digital technologies in the Vietnamese critical care settings, it is important to understand the current status of DHT adoption in Vietnamese hospitals.

Objective: We aimed to explore the current digital maturity in 5 Vietnamese public hospitals to understand their readiness in implementing new DHTs.

Methods: We assessed the adoption of some key DHTs and infrastructure in 5 top-tier public hospitals in Vietnam using a questionnaire adapted from the Vietnam Health Information Technology (HIT) Maturity Model. The questionnaire was answered by the heads of the hospitals' IT departments, with follow-up for clarifications and verifications on some answers. Descriptive statistics demonstrated on radar plots and tile graphs were used to visualize the data collected.

Results: Hospital information systems (HIS), laboratory information systems (LIS), and radiology information systems–picture archiving and communication systems (RIS-PACS) were implemented in all 5 hospitals, albeit at varied digital maturity levels. At least 50% of the criteria for LIS in the Vietnam HIT Maturity Model were satisfied by the hospitals in the assessment. However, this threshold was only met by 80% and 60% of the hospitals with regard to HIS and RIS-PACS, respectively. Two hospitals were not using any electronic medical record (EMR) system or fulfilling any extra digital capability, such as implementing clinical data repositories (CDRs) and clinical decision support systems (CDSS). No hospital reported sharing clinical data with other organizations using Health Level Seven (HL7) standards, such as Continuity of Care Document (CCD) and Clinical Document

XSL•FO
RenderX

Architecture (CDA), although 2 (40%) reported their systems adopted these standards. Of the 5 hospitals, 4 (80%) reported their RIS-PACS adopted the Digital Imaging and Communications in Medicine (DICOM) standard.

Conclusions: The 5 major Vietnamese public hospitals in this assessment have widely adopted information systems, such as HIS, LIS, and RIS-PACS, to support administrative and clinical tasks. Although the adoption of EMR systems is less common, their implementation revolves around data collection, management, and access to clinical data. Secondary use of clinical data for decision support through the implementation of CDRs and CDSS is limited, posing a potential barrier to the integration of external DHTs into the existing systems. However, the wide adoption of international standards, such as HL7 and DICOM, is a facilitator for the adoption of new DHTs in these hospitals.

(JMIR Form Res 2025;9:e53483) doi:10.2196/53483

KEYWORDS

electronic health record; electronic medical record; digital maturity; clinical decision support; digital infrastructure; Vietnam; health information technology; digital health technology; low- and middle-income country

Introduction

The COVID-19 pandemic demonstrated the need for countries worldwide to implement a national digital health infrastructure needed to respond to health emergencies and to achieve the Sustainable Development Goals (SDGs) [1]. As part of this digital transformation effort, digital health technologies (DHTs), such as electronic medical record (EMR) systems, laboratory information systems (LISs), and picture archiving and communication systems (PACSs), have become widespread. In recent years, Vietnam has developed and begun implementation of an ambitious national digital health strategy that includes the deployment of hospital-based EMRs and an electronic health insurance claim system [2].

As the cost of hardware continues to drop and the range and capabilities of DHTs continue to expand [3], digitization of health systems is becoming increasingly feasible in low- and middle-income countries (LMICs). New technologies, such as the Internet of Things, wearable devices, and artificial intelligence (AI), have begun to be adopted in health care settings and could be an important enabler for realizing the World Health Organization's (WHO) Triple Billion Targets (more than 1 billion people benefiting from universal health care; better protected from health emergencies; and enjoying better health and well-being) and to accelerate SDG achievement [4]. In 2023, WHO announced the Global Initiative for Digital Health to help LMICs achieve the aims of the Global Strategy on Digital Health by developing their national digital health infrastructure, adopting international data standards, and fostering increased international collaboration in digital health [1,4].

In global health research, there is increasing interest in implementing new technologies, such as AI or smart wearables, for clinical decision support [5,6]. However, for these systems to be practicable and to take advantage of increasingly available "big data," the local provision of technological infrastructure and implementation capabilities is necessary. For example, hospitals that have implemented EMR systems could use data from the EMRs to train AI models and enable the implementation of clinical decision support systems (CDSS) [7]. In addition, the adoption of international interoperability standards, such as Health Level Seven (HL7) Fast Healthcare Interoperability Resources (FHIR), is also necessary to allow efficient data exchange between information systems and new DHTs [8].

In hospital settings, digital maturity models, such as the Electronic Medical Record Adoption Model (EMRAM) from the Healthcare Information and Management Systems Society (HIMSS), are often used to guide DHT adoption and implementation [9]. These models can be used on a national or international scale and can target different dimensions, including technology, strategy, interoperability, analytics, and governance. Currently, there is no international consensus on how to best measure digital capabilities in health care institutions [9,10], and variations in digital maturity models across different countries may reflect an adaptation to national priorities and contexts, especially when being implemented in LMICs. In 2017, the Vietnam Ministry of Health (MoH) issued the Vietnam Health Information Technology (HIT) Maturity Model, originally known as Circular 54/2017/TT-BYT, as a roadmap for Vietnamese public hospitals toward a paperless and smart hospital model [2,11]. Although this guidance defines 7 digital maturity stages similar to EMRAM, the criteria to achieve a specific level have been significantly modified. For example, the Vietnam model includes the requirement for hospital information systems (HIS) to connect to the Vietnam Social Security claim portal and assesses the integration of the national standardized vocabularies into the digital systems [11].

The Vietnam ICU Translational Applications Laboratory (VITAL) project focuses on developing new technologies for improving outcomes of critically ill patients in Vietnam, such as CDSS using AI algorithms trained on local data [12-14]. In the national DHT policy landscape, regulations, infrastructure, and professional workflows are important elements [2,15] to ensure the sustainability and successful implementation of DHTs [6,16]. For this reason, in this study, we aimed to explore the current status and likely short-term development of the DHT infrastructure in 5 major public hospitals in Vietnam. Although the findings may not be generalizable to all Vietnamese hospitals regarding the state of their DHT adoption, we expect they can inform discussions around how new DHTs can be integrated with the existing IT infrastructure in Vietnamese public hospitals.

XSL•FO
RenderX

Methods

Study Design

This study used a structured questionnaire to explore the current adoption of DHTs in 5 public hospitals in Vietnam. The assessment items were developed based on the Vietnam HIT Maturity Model, which comprises 8 domains related to technology adoption and digital capabilities in hospitals [11]. The 8 domains are *IT infrastructure*, *HIS*, *LIS*, *radiology information systems–picture archiving and communication systems (RIS-PACS)*, *EMR*, *administrative and operation software*, *security and information safety*, and *nonfunctional criteria*. Except for security and information safety and nonfunctional criteria, each domain includes a set of infrastructure requirements, such as having a server room, or software functional criteria, such as having laboratory test management functionality. The model also sets out further digital capabilities focusing on assisting clinical practice, such as CDSS, clinical data repositories (CDRs), and medication management; these are named *extra capabilities*.

Following a stage-based approach, a hospital reaches a digital maturity level in a domain if a predefined set of criteria is met. The hospital can only move to a higher digital maturity level when all criteria of the lower levels have been satisfied. The domains of IT infrastructure and HIS have 7 digital maturity levels (1-7), while the other domains have 2 levels (*basic* and *advanced*). Multimedia Appendix 1 presents the digital maturity levels and their scoring criteria based on the domains and extra capabilities.

Questionnaire, Data Collection, and Analysis

Since the focus of this study was the adoption of digital health systems in hospitals, we developed an assessment tool based on the criteria from the IT infrastructure, HIS, LIS, RIS-PACS, EMR, and extra capabilities domains (see Multimedia Appendix 2). The administrative and operation software, security and information safety, and nonfunctional criteria domains were not included in this study as we considered these less relevant to the study's focus despite their importance to the overall digital maturity of a health care institution.

For each domain, we gathered information through the hospitals' IT departments to determine whether the hospitals met the component criteria or whether they planned to achieve the criteria in the next 3-5 years. Results of the assessments were discussed with the heads of the IT departments to ensure accuracy. Data collection was carried out from July 2022 to March 2023. Information about the digital maturity of each hospital was entered into Microsoft Excel for initial data cleaning and then analyzed using R software (R Foundation for Statistical Computing) [16].

Our aim was to provide a descriptive summary of what infrastructure and functionalities, or criteria, were available in each hospital across the domains rather than calculating scores for their digital maturity. The proportion of criteria that each hospital met in each domain was calculated. A radar plot was created using the *fmsb* package in R [17] to depict the difference in the proportions of criteria met among the hospitals by domain. Radar plots are regarded as an efficient tool to compare various groups on multiple variables, as discussed by Saary [18]. In addition to summary statistics, we further compared technology adoption between the 5hospitals by delving into the individual domains. Tile graphs were created using the *ggplot2* package in R [19] to illustrate the criteria that the hospitals had satisfied within each domain. These graphs allow for the visualization of individual data points (eg, which criterion was met by a specific hospital) as opposed to visualizing the counts or proportions of hospitals that met a particular criterion, as in the case of bar charts.

Study Population

Until the beginning of 2024, it was estimated that there were nearly 1500 public and over 300 private hospitals in Vietnam [20]. Five hospitals were purposively selected for this study based on previous interest in DHT innovation and participation in digital health research activities. Our approach aimed to sample hospitals where DHT innovation was likely to occur the soonest, rather than providing a comprehensive view of all government hospitals. Of the 5 hospitals that participated in this assessment, 4 (80%) were in Hanoi or Ho Chi Minh City, the 2 largest cities in Vietnam, and 1 (20%) was in a Mekong Delta (ie, rural) province. All are top-tier teaching public hospitals with a capacity of over 500 beds; 2 (40%) are tertiary hospitals that focus on a limited number of specialties, while 3 (60%) are general hospitals that provide care in a wide range of specialties. Most hospitals have between 8 and 12 IT staff per 1000 beds, except for 1 (20%) hospital that has an approximate ratio of over 50 per 1000 beds. Characteristics of the hospitals are shown in Table 1. To ensure confidentiality, we withheld information that could be used to identify the hospitals.

Table 1. Characteristics of the 5 hospitals participating in this assessment.

Hospital	Tier	Hospital type	Number of beds	Teaching hospital	IT staff:bed ratio
A	1	General hospital	>500	Yes	50:1000
B	1	Tertiary hospital	>500	Yes	13:1000
C	1	Tertiary hospital	>500	Yes	11:1000
D	1	General hospital	>500	Yes	8:1000
E	1	General hospital	>500	Yes	12:1000

XSL•FO
RenderX

Ethical Considerations

This project was exempted from ethical review as it constituted a service review, as defined by the Oxford Tropical Research Ethics Committee (OxTREC) [21,22]. The selected hospitals were identified from a professional network that the study's principal investigators (authors LT and CP) were part of. The hospitals were invited to participate in the assessment through an invitation letter explaining the purpose and scope of the study. Upon approval by the hospitals' authorities, the hospital IT directors were provided with an information sheet, the questionnaire, and an instruction document. Queries regarding the assessment, if any, were clarified by the first author (DMT).

The IT directors were informed in writing that they could refuse to provide any data if they wished and that the data collected would be securely managed and only be used for the purpose of the study. Participating departments were able to review and agree on the contents of the manuscript prior to submission. No personally identifiable data were collected during the research. The research data were managed as per the Oxford University Clinical Research Unit's data management policy. Information that can be used to identify the hospitals, such as names and the number of beds, was deidentified to ensure confidentiality. No compensation was given to the hospitals for participating in this study.

Information Systems in Vietnamese Public Hospitals

HIS, LIS, radiology information systems (RIS), PACS, and EMRs are popular information systems in health care institutions worldwide. However, their functionalities and use cases may differ to some extent depending on the national context that they are implemented in. In this section, we provide a Vietnam-oriented description of these information systems to support the interpretation of the study's findings.

Hospital Information Systems

Also known as hospital management systems, HIS are usually prioritized by Vietnamese hospitals among other information systems. They are typically used to manage outpatient services, as well as administrative and billing data emerging from all wards and departments in a hospital. HIS can be interfaced with RIS and LIS to send orders or receive imaging and laboratory results. A significant proportion of HIS in Vietnamese hospitals have been developed and provided by local vendors, such as the Corporation for Financing Promoting Technology (known as FPT), the Vietnam Posts and Telecommunications Group (known as VNPT), and Viettel [23].

Laboratory Information Systems

LIS allow clinical laboratories to track and use a wide range of data related to orders, specimens, results, and consumables. They reduce the turnaround time from ordering tests to collecting laboratory results through automatic processes, such as integrating with HIS to transfer orders to the laboratory and connecting with laboratory instruments to automatically receive results. Clinicians can use LIS installed in their wards or HIS interfaced with LIS to access laboratory results, when available. Many public hospitals in Vietnam have chosen LIS developed by local vendors, such as LABCONN by LABSoft (deployed in over 200 laboratories) and LABMDSOFT by MDsoft.

Radiology Information Systems–Picture Archiving and Communication Systems

RIS and PACS are often used together in medical imaging departments. RIS typically support managing orders, patients, and imaging procedure information. RIS can receive order information and send images and report back to HIS using HL7 messages. RIS can also send orders to imaging machines and receive processed images from PACS.

PACS comprise 4 components: imaging instruments, a secure network to transfer the images, workstations that allow doctors to view the images, and a storage system that can archive the images. PACS can thus facilitate image viewing, editing, storing, and sharing. With PACS, hospitals can eliminate physical film usage and provide multisite access to medical images through a web interface.

Electronic Medical Records

In Vietnam, EMR systems refer to systems implemented in hospitals, especially in inpatient wards, to manage patient care information, such as medical histories, problems, orders, tests, and medications, similar to paper charts [24]. EMRs are defined differently than electronic health records (EHRs) in Vietnam, which are defined as health records updated throughout a person's lifetime and can cover a comprehensive range of health information, such as allergies, vaccination, family history, and outpatient visits rather than specialized care information [25]. Although the data collected by EMR systems are independently managed by hospitals, EHR data are synchronized from multiple sources, such as hospital and primary care facility visits to a centralized platform managed by the MoH [26]. In the Vietnam HIT Maturity Model, EMR systems were assessed on 4 areas as follows:

- Clinical data management functions, such as medical histories, clinical documents, and test management
- Administrative and demographics functions, including managing information of health staff and patients and managing integrations between the EMR system and other information systems in a hospital
- Medical record storage capabilities, including the required record retention duration according to health care law, record synchronization, and medical record storage and restoring
- Technical administration functions, including security, supervision, standardized terminology management, standard-based data exchange, EMR workflow management, and database backup and recovery

Extra Capabilities

In the Vietnam HIT Maturity Model, each digital maturity level requires several extra digital capabilities, in addition to the principal information systems, such as HIS, LIS, and EMRs (see Multimedia Appendix 1). These capabilities can be grouped as follows:

- CDRs: These are centralized data stores that gather patient data generated from other clinical information systems. Data and resources from CDRs can also be used by other systems and applications in the same network, such as

XSL•FO
RenderX

standardized laboratory order vocabularies can be jointly deployed in EMRs and LIS. This assessment examined CDR implementation based on the following criteria: (1) Hospitals can establish a CDR containing standardized vocabularies, medications, orders, and laboratory tests; (2) information in CDRs can be shared with stakeholders involved in patient care; and (3) CDRs contain data on vital signs, nursing notes, and clinical procedures.

- CDSS are categorized into 3 levels: CDSS level 1 can support drug prescribing, including new prescription and represcribing; CDSS level 2 can send alerts for basic conflicts in ordering and prescribing; and CDSS level 3 can inform clinicians' treatment plans and outcomes via appropriately customized alerts.
- Electronic ordering and inpatient order management.
- Digitizing clinical notes using electronic structured templates.
- Closed loop medication management using identification technologies such as radiofrequency identification and barcoding.
- Sharing clinical data with stakeholders involved in patient care through standardized electronic transactions, such as the HL7 Clinical Document Architecture (CDA) and the HL7 Continuity of Care Document (CCD).
- Continuous summaries of service usage data from all the departments, such as inpatient, outpatient, and emergency departments.

Results

Assessment Criteria

Although we were not able to collect information about 3-5-year investment plans from the 5 hospitals, we were able to collect data on all the assessment criteria from the maturity model. The radar plot in Figure 1 presents the percentage of criteria the 5 hospitals met in each domain.

Figure 1. Percentage of HIT criteria satisfied in each domain across the 5 hospitals (IT infrastructure, EMR, LIS, HIS, RIS-PACS, and extra capabilities). EMR: electronic medical record; HIS: hospital information systems; HIT: health information technology; LIS: laboratory information systems; RIS-PACS: radiology information systems–picture archiving and communication systems.

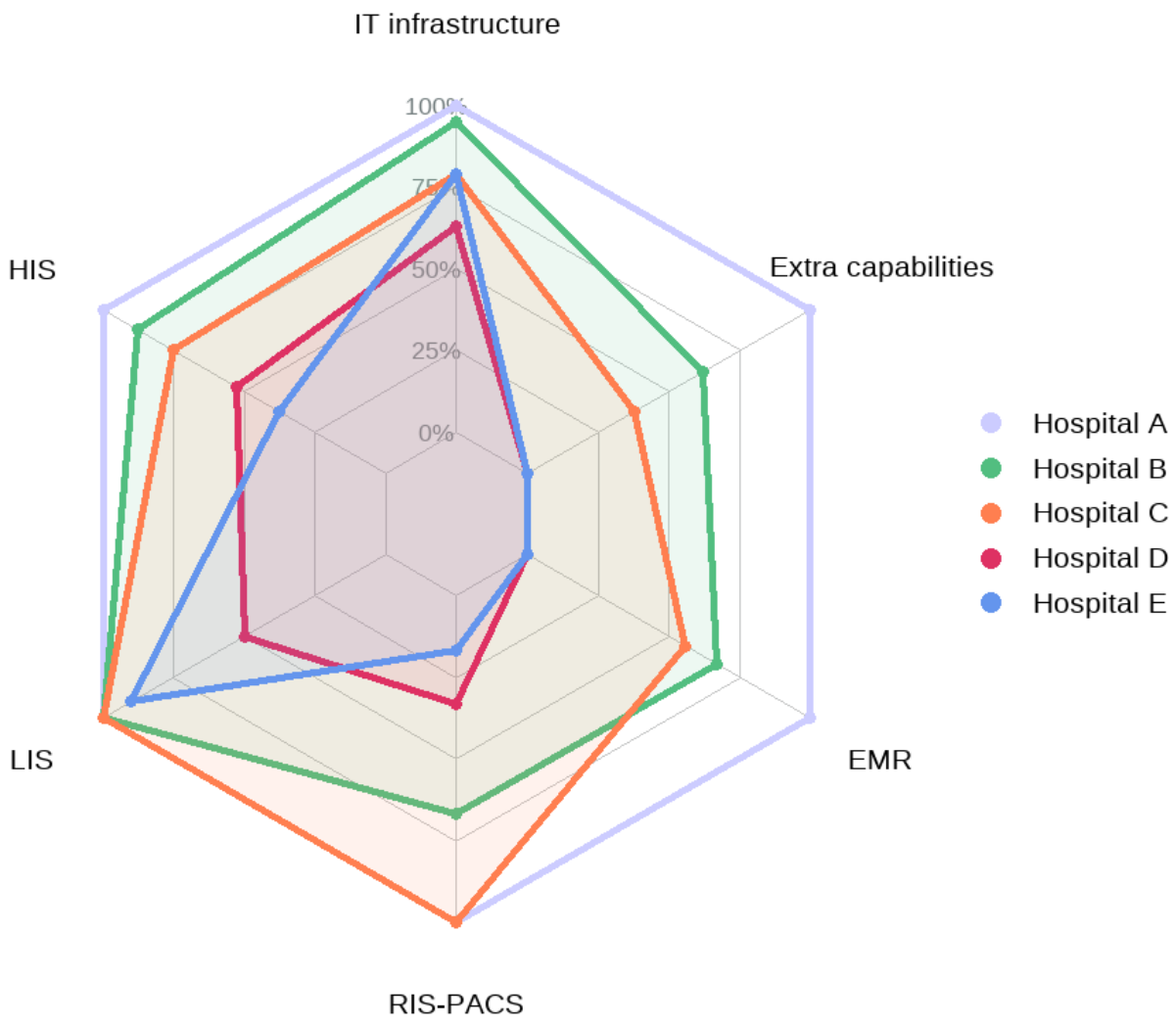

In general, HIS, LIS, and RIS-PACS were implemented in all 5 hospitals, albeit with varied digital maturity levels. Two hospitals were not using any EMR system or fulfilling any extra digital capability. One hospital reported meeting all the criteria across the 6 domains. In the following sections, we describe the

XSL•FO
RenderX

domains in more detail and compare the implementation strategies between the hospitals.

IT Infrastructure

Figure 2 shows the IT Infrastructure criteria that were met by the 5 hospitals. Level 1 criteria were met by all 5 hospitals, and most hospitals reported meeting level 2-5 criteria.

Figure 2. Criteria of IT infrastructure in the Vietnam HIT Maturity Model met by the 5 hospitals in this assessment, where a maximum of level 7 could be obtained. The color coding represents the maturity group that a criterion belongs to, not the entire domain's maturity level of a hospital. Although for official purposes, a hospital needs to meet all the criteria of a lower level to progress to a higher level, this figure represents any criteria attained. HIT: health information technology; LAN: local area network; NAS: network attached storage; SAN: storage area network.

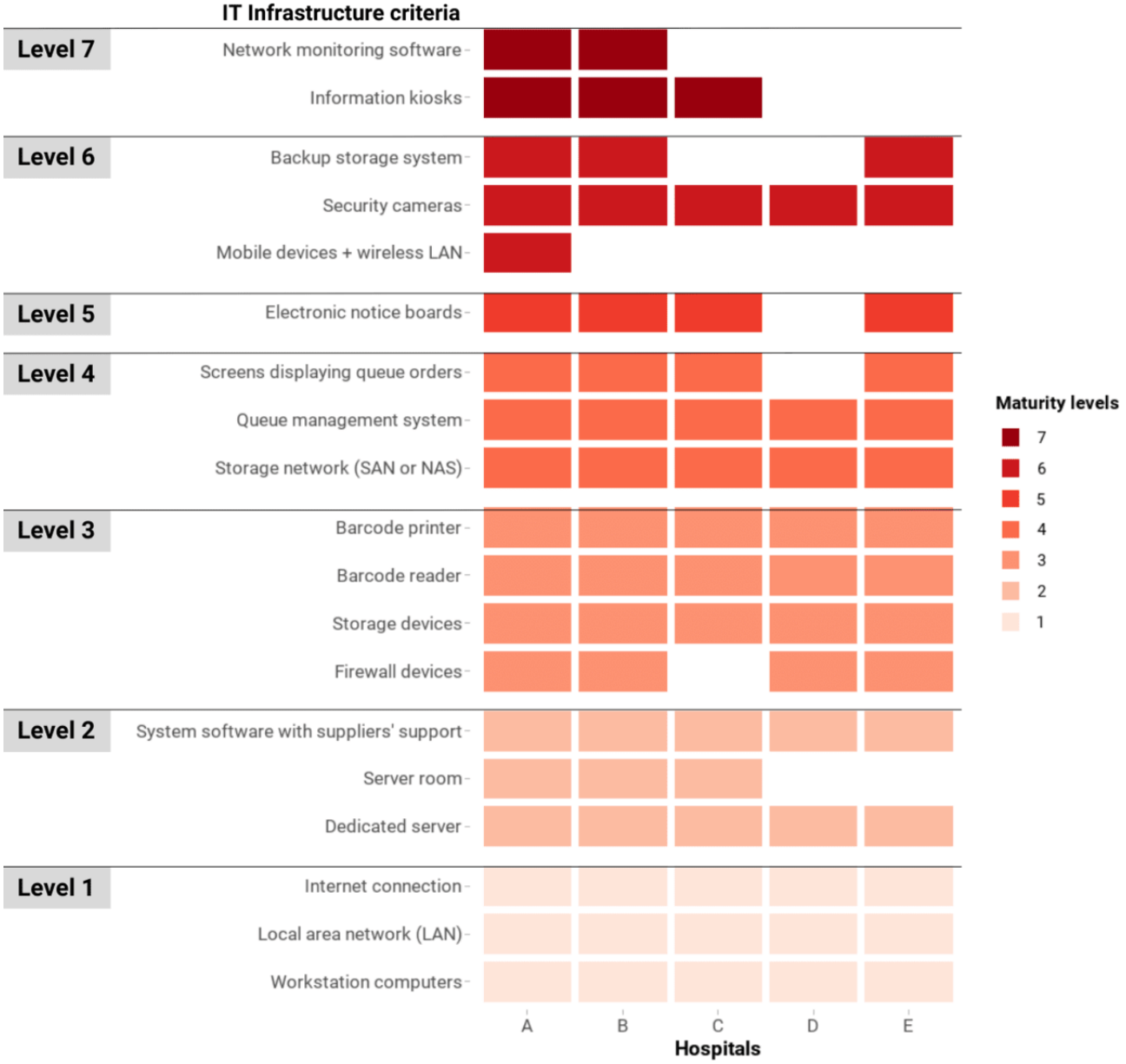

All 5 hospitals had sufficient workstations for staff. They were connected to the internet and local area networks. All 5 hospitals' digital systems and databases were run on dedicated servers, and 3 (60%) hospitals had their own on-premise server rooms with physical controls. The hospitals' data were stored on storage devices that were supported by storage networks, such as network attached storage (NAS) or storage area network (SAN) systems. Backup storage systems were available in 3 (60%) of the 5 hospitals.

Almost all hospitals in this assessment adopted technologies to help reduce errors and improve patient flow. These included barcode readers, barcode printers, queue management systems, screens displaying queue orders, and electronic notice boards. Information kiosks were less common as only 3 (60%) of the 5 hospitals implemented this technology.

Level 6 and 7 criteria were less available in the hospitals. Only 1 (20%) hospital reported adopting mobile devices, such as tablets and smartphones, and a wireless local area network (LAN) to support hospital activities.

XSL•FO
RenderX

Hospital Information Systems

All 5 hospitals reported that their HIS fully met the level 1 and 2 criteria (Figure 3). These systems were able to support patient registration, outpatient services, and pharmacy stock management. Hospital fees, including social health insurance payments, were also managed in these HIS, where insurance claims were formatted into XML files following a national standard and routinely submitted to the Vietnam Social Security portal. The HIS were reportedly able to integrate Vietnam's national coding systems to standardize the classifications of some administrative units, such as the Vietnam ethnic classification system, and clinical services, such as the WHO International Classification of Disease [27]. Medical orders and test results were assigned unique patient IDs and uploaded to the HIS. Inpatient service management and reporting functionalities (level 3) were also available in the HIS across the 5 hospitals.

Figure 3. Criteria of HIS in the Vietnam HIT Maturity Model met by the 5 hospitals in this assessment, where a maximum of level 7 can be obtained. The color coding represents the maturity group that a criterion belongs to, not the entire domain's maturity level of a hospital. Although for official purposes, a hospital needs to meet all the criteria of a lower level to progress to a higher level, this figure represents any criteria attained. EMR: electronic medical record; HIS: hospital information systems; HIT: health information technology; PACS: picture archiving and communication systems.

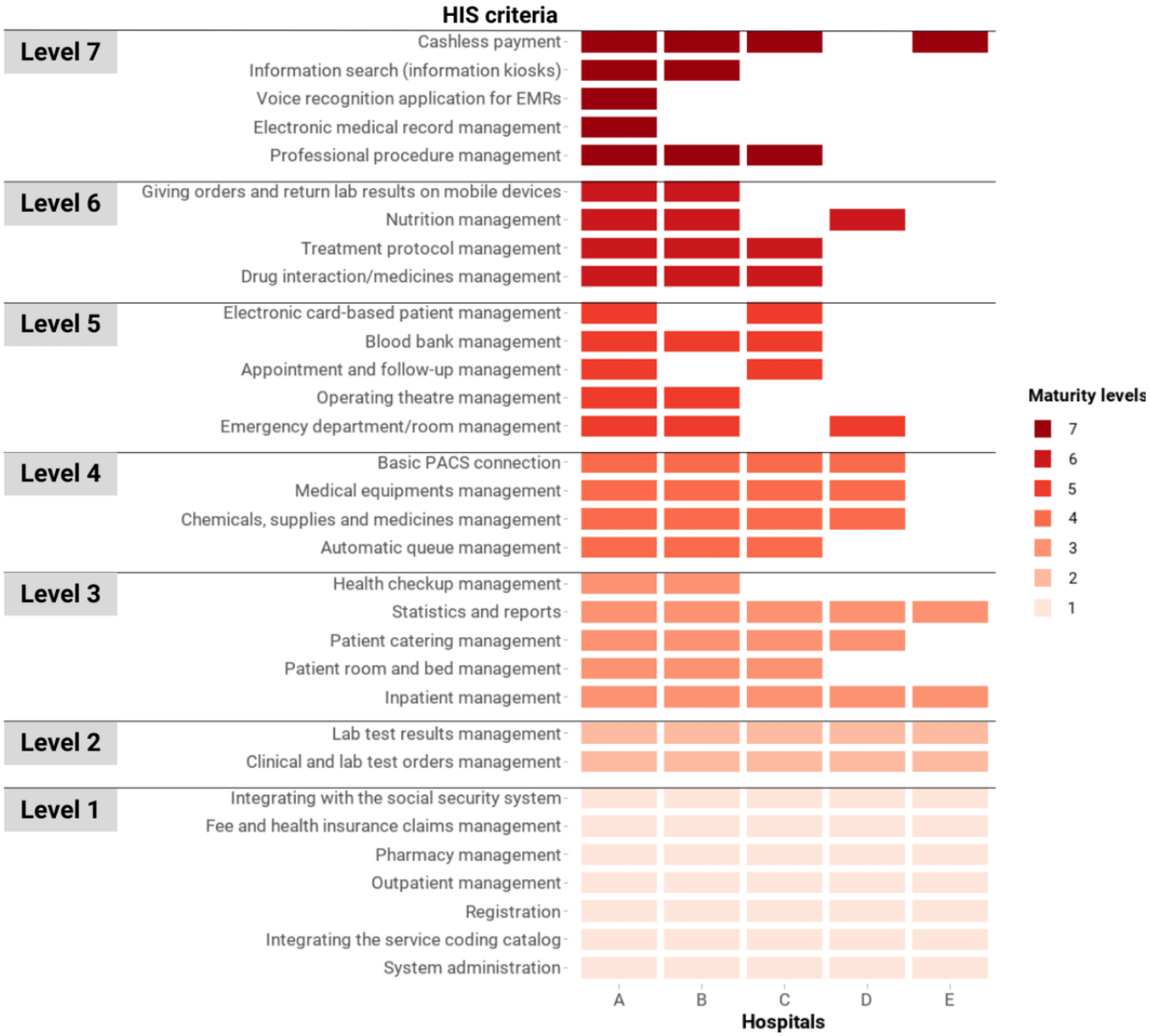

Of the 5 hospitals, 4 (80%) reported that their HIS integrated with PACS and cashless payment technology. However, modules for treatment protocols, professional procedures, and drug interaction management were implemented in only 3 (60%) hospitals.

We found the functionalities that were least adopted were the ones that required integration with EMR systems (ie, voice recognition for EMRs and EMR management, available in 1, 20%, hospital). In addition, only 2 (40%) hospitals reported that their HIS could interface with information kiosks, electronic patient cards, and mobile devices.

Laboratory Information Systems

Results of our assessment showed that all 5 hospitals reported that their LIS met the basic criteria (Figure 4). All the LIS could manage hospital laboratory orders and results, as well as generate reports. Most of these systems could automatically send orders to and receive results from laboratory machines.

XSL•FO
RenderX

One hospital reported that its LIS could receive results from laboratory instruments, but the orders could only be entered manually. The national dictionary system for laboratory tests was adopted across all 5 hospitals to standardize laboratory coding. The LIS of 4 (80%) hospitals satisfied all the advanced criteria, which included the ability to integrate with HIS for exchanging orders and results, set alert thresholds for laboratory results, manage chemical supply stocks, and manage laboratory specimens.

Figure 4. Criteria of LIS in the Vietnam HIT Maturity Model met by the 5 hospitals in this assessment, where a maximum of 2 levels can be obtained. The color coding represents the maturity group that a criterion belongs to, not the entire domain's maturity level of a hospital. Although for official purposes a hospital needs to meet all the criteria of a lower level to progress to a higher level, this figure represents any criteria attained. HIS: hospital information systems; HIT: health information technology; LIS: laboratory information systems.

Radiology Information Systems—Picture Archiving and Communication Systems

All 5 hospitals were using RIS-PACS at the time of this assessment, although with different degrees of implementation (Figure 5). Of the 5 hospitals, 2 (40%; A and C) had RIS-PACS that met all the basic and advanced criteria, 1 (20%) hospital (B) nearly satisfied all the basic and advanced criteria, and RIS-PACS from the remaining 2 (40%) hospitals (D and E) were largely operated by external vendors and only implemented a limited number of functionalities of RIS-PACS.

XSL•FO
RenderX

Figure 5. Criteria of RIS-PACS in the Vietnam HIT Maturity Model met by the 5 hospitals in this assessment, where a maximum of 2 levels can be obtained. The color coding represents the maturity group that a criterion belongs to, not the entire domain's maturity level of a hospital. Although for official purposes a hospital needs to meet all the criteria of a lower level to progress to a higher level, this figure represents any criteria attained. DICOM: Digital Imaging and Communications in Medicine; HIS: hospital information systems; HIT: health information technology; HL7: Health Level Seven; RIS-PACS: radiology information systems–picture archiving and communication systems.

RIS-PACS meeting the basic criteria were able to retrieve Digital Imaging and Communications in Medicine (DICOM) images of common imaging modalities, such as X-ray, magnetic resonance imaging, and ultrasound, through 2-way interfaces. The RIS, PACS, and HIS could exchange orders and images with one another based on HL7 messaging standards. The PACS could convert DICOM images to JPEG format, and any changes to images in the PACS could be promptly updated to the HIS. Orders and radiologists' reading result management, measurement, and reporting were also ensured by these RIS-PACS.

Hospitals meeting the advanced criteria of RIS-PACS could carry out multisite consultations through web-based access to DICOM images. One hospital reported that its PACS had adopted the HL7 FHIR standard.

Electronic Medical Records

Generally, the EMR status was less mature than other domains (Figure 6), with only 3 (60%) hospitals reporting that they were implementing EMR systems. Only 1 (20%) of these satisfied sufficient criteria in the regulations for EMR systems [24], which allowed the hospital to completely eliminate archiving of paper records for backup and legal purposes. Nevertheless, the other EMR systems could provide clinical data management functions, including managing medical histories, clinical documents, orders, laboratory and imaging results, treatments, and prescriptions.

XSL•FO
RenderX

Figure 6. Criteria of EMRs in the Vietnam HIT Maturity Model met by the 5 hospitals in this assessment, where a maximum of 2 levels can be obtained. The color coding represents the maturity group that a criterion belongs to, not the entire domain's maturity level of a hospital. Although for official purposes a hospital needs to meet all the criteria of a lower level to progress to a higher level, this figure represents any criteria attained. CCD: Continuity of Care Document; CDA: Clinical Document Architecture; EMR: electronic medical record; HIT: health information technology; HL7: Health Level Seven.

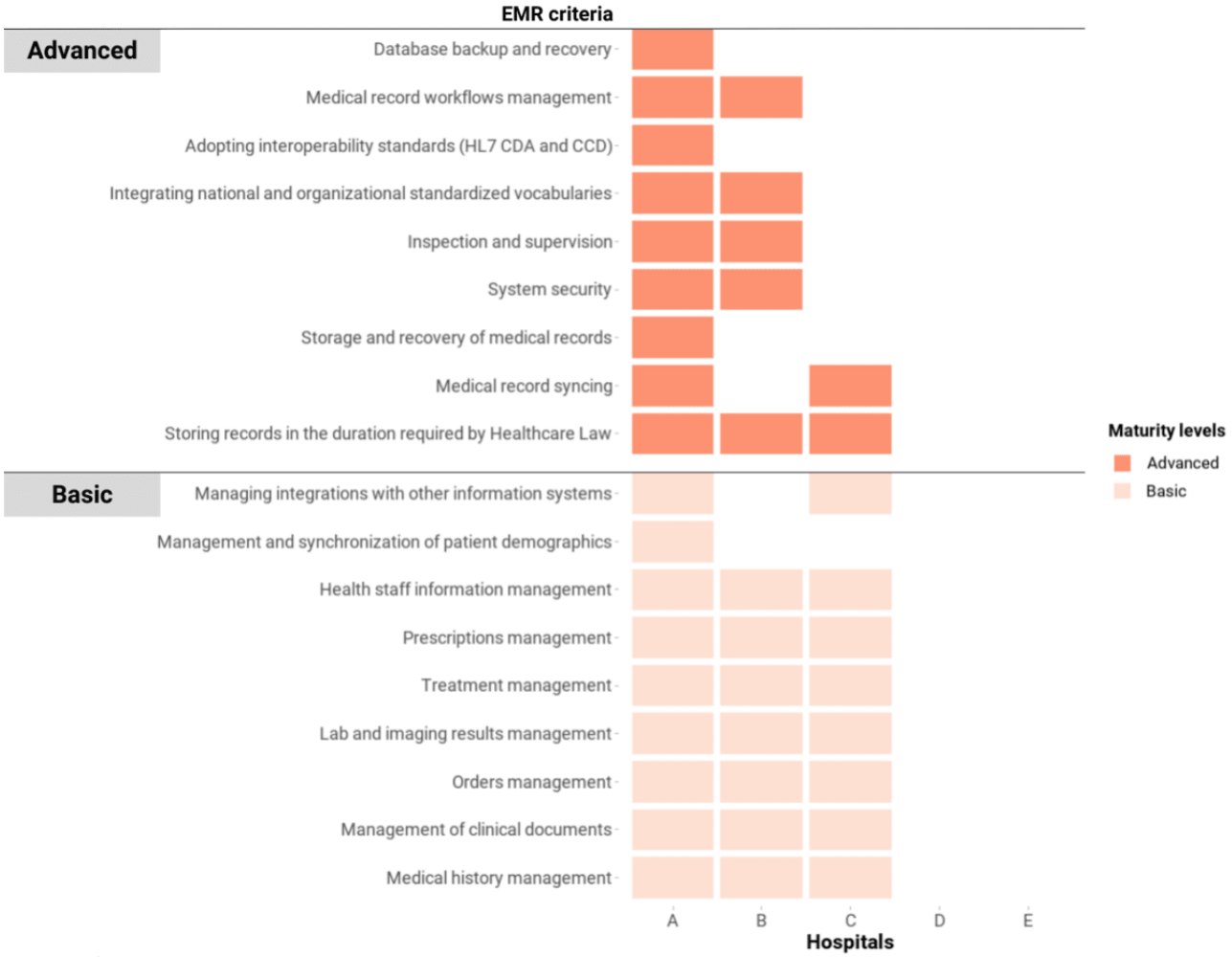

An EMR system fully compliant with the regulations for EMRs could ensure storage capacity, backup, and recovery capabilities. Interoperability standards, such as the HL7 CDA, the HL7 CCD, and HL7 FHIR, were implemented. Information security measures, such as authentication, audit trail, and encryption, were in place. Digital signatures were used to authenticate medical records and orders.

Extra Capabilities

The availability of these capabilities in the 5 hospitals is shown in Figure 7. Notably, 1 (20%) hospital reported achieving all the extra capabilities, with 2 (40%) hospitals confident about sharing clinical data using HL7 standards (although they were not currently sharing data at the time of the study). The adoption of CDRs and CDSS to support clinical tasks, such as giving orders, prescribing, and clinical decision-making, was only available in 1 (20%) hospital. However, all the hospitals with an EMR system reported being able to digitize all clinical documents and allow clinicians to give orders in the electronic environment.

XSL•FO
RenderX

Figure 7. Extra capability criteria in the Vietnam HIT Maturity Model met by the 5 hospitals in this assessment. CDR: clinical data repository; CDSS: clinical decision support systems; HIT: health information technology; HL7: Health Level Seven.

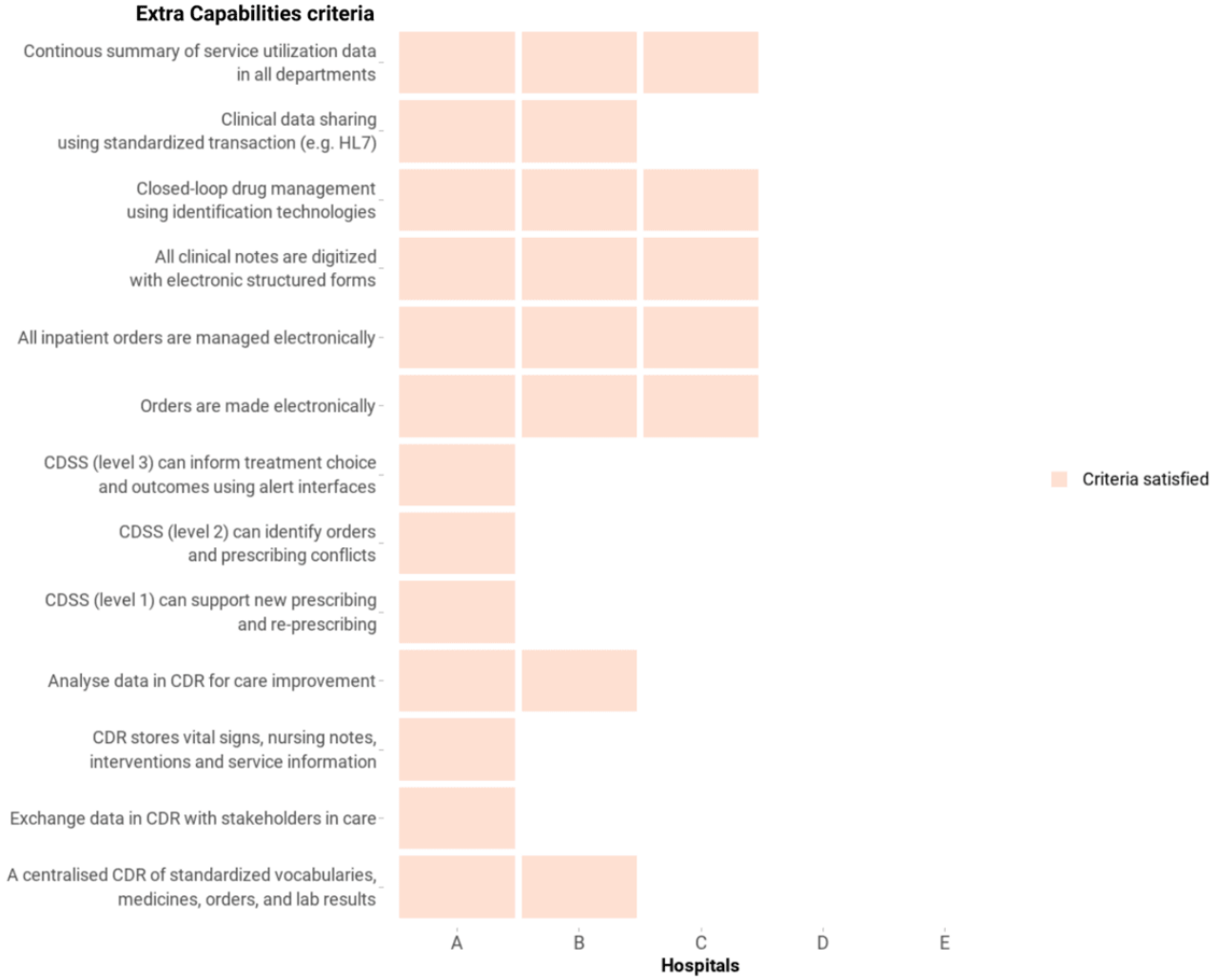

Discussion

Principal Findings

This study assessed the implementation of some key digital health systems in 5 top-tier hospitals in Vietnam. Overall, many processes related to administration, laboratory tests, and medical imaging were digitized in the hospitals through the implementation of HIS, LIS, and RIS-PACS on the basis of adequate basic IT infrastructures. The digital maturity of these systems, however, varied significantly between the hospitals, although all are large top-tier hospitals. EMR systems were implemented in 3 of the 5 hospitals. Although these systems could satisfy a wide range of data collection needs, their integration with CDRs and CDSS was sparse.

Comparison With Prior Works

This is the first original research that examined digital health system implementation in Vietnamese hospitals since the Vietnam MoH enacted the National EMR plan and the Vietnam HIT Maturity Model in 2018 as part of the sectorwide digital transformation agenda [11,21,24]. Academic publications on adoption and implementation of hospital-based digital health systems are limited [2]. Most of the previous studies have been conducted before 2014 and focused on the readiness to transition from paper-based processes to electronic systems rather than describing DHT adoption [28,29]. Other publications have described the development and pilots of technical solutions, such as LIS, data retrieval from medical devices, and data visualization for EMRs. We could not find any research exploring the digital systems currently used in hospitals in Vietnam [30-33].

Muinga et al [34] showed that most of the digital health systems adopted in Kenyan public hospitals from 2014 to 2016 were aimed at administrative and billing purposes rather than supporting clinical tasks. Systems such as LIS, RIS, and PACS, when available, were usually stand-alone systems and lacked interoperability with other systems in the same institution [34]. Our study found that all the Vietnamese hospitals assessed had gone beyond administrative systems and had used clinical information systems that are interoperable with each other, especially with HIS. EMRs, as a crucial clinical information system, had been implemented in 3 of the 5 hospitals to capture a wide range of data, such as medical histories, clinical documents, orders, and test results. However, the adoption of EMR capabilities, such as CDRs and CDSS, was limited, even though these are large teaching hospitals at the top tier of the Vietnamese health system. This is inferior to the digital maturity status of public hospitals in Turkey, as over 98% of the Turkish

XSL•FO
RenderX

hospitals surveyed reported having CDRs regardless of their size [35]. In addition, CDSS for medication orders and nonmedication orders were implemented in 71% and over 57% of Turkish public hospitals, respectively. The early use of EMRAM to guide the national digital health care transformation may be an important facilitator for this wide adoption of CDRs and CDSS in Turkish hospitals [36].

The Vietnam HIT Maturity Model [11] that this study's questionnaire was based on and the HIMSS EMRAM [9] share several similarities regarding assessment criteria. For example, CDRs, CDSS, closed-loop medication administration, and electronic documentation are included in both models, including the maturity stage that these systems belong to. However, EMRAM particularly focuses on EMR capabilities, in which systems such as laboratory, radiology, and pharmacy are seen as EMR ancillaries, while CDRs and CDSS are crucial components deciding the maturity of the EMR system. All the ancillary systems need to be implemented as early as in stage 1 of EMRAM. The Vietnam HIT Maturity Model, in contrast, splits the ancillaries into basic and more advanced maturity levels along the digital maturity roadmap. Systems such as CDRs and CDSS are the extra capabilities that should be met, alongside other domains, such as HIS, LIS, RIS, and PACS, so that a hospital can move up to a higher digital maturity level. Although the division of digital systems into smaller milestones may allow hospitals to better benchmark their digital maturity and plan for investments, the requirement to meet digital maturity criteria across multiple domains can be challenging for hospitals in a resource-limited setting such as Vietnam. Duncan et al [10] discussed that most published maturity models, including EMRAM, are not assessed in LMICs, questioning their applicability in cultures different from those of high-income countries, such as the United States and the United Kingdom. In addition, the implementation of EMRAM poses the risk of investing in complex systems without meeting the organization's local needs [37].

Integrating New DHTs Into the Existing System

Most hospitals participating in this assessment had a low adoption rate of EMRs, CDSS, and CDRs. The low number of complete EMR systems may be caused by the cost required to upgrade the infrastructure and a lack of the required technical capabilities to meet all the requirements in EMR regulations. It is difficult for hospitals with incomplete EMR systems and CDRs to embark on using CDSS. Integrating, implementing, and maintaining CDSS is usually complex and depends on multiple factors, such as the availability and interoperability of data, especially from EMRs; an updated knowledge base; and understanding of the hospital's workflow [38].

The wide availability of HIS, LIS, and RIS-PACS in the hospitals participating in our study suggests that digital systems for administrative and billing purposes and simpler systems to implement tend to be prioritized. Many hospitals may choose to implement clinical documentation functions before upgrading the other components to fully comply with EMR regulation. This leads to an increasing amount of medical data that are electronically available and useful for AI training and implementation even without full EMR implementation. However, the scarcity of CDSS, CDRs, and standard-based data-sharing experience found in this assessment suggests challenges that stakeholders may face when seeking to integrate AI-enabled solutions in Vietnamese hospitals. Although these challenges should be explored in further research, we believe a lack of national frameworks and detailed guidance for meaningful and secure clinical data exchange is 1 of the key barriers. In addition, recently, Chanh et al [39] mentioned exploring system interoperability, collaboration with local stakeholders from multiple disciplines, and understanding local data-sharing policies as some of the key considerations for applying AI in Vietnam.

Strengths and Limitations

We examined the adoption of digital health systems using the criteria published in the Vietnam HIT Maturity Model. This is a national benchmarking tool for HIT implementation in Vietnamese public hospitals, making it familiar to the hospitals participating in this study.

Only 5 public hospitals at the top-tier level were included in this assessment, so our findings do not represent the overall picture of HIT implementation in Vietnamese hospitals. The hospitals participating in this study were ones that had previous interest in DHT implementation and digital health innovation research. However, it is likely that hospitals at low tiers and located in less-funded regions have lower digital maturity across all the domains, especially in the EMR and extra capabilities domains.

Implications for Further Study

The adoption of digital health systems in this study was assessed based on technological criteria, such as the software functions that were available. Future work could address other dimensions related to the implementation of these systems, such as the organizational context and human factors [40]. The use of qualitative methods can be particularly suitable for such research questions as they can unveil the characteristics unique to technology adoption in hospitals in LMIC settings, such as Vietnam.

Conclusion

Several major public hospitals in Vietnam have a sound digital infrastructure in place to support fundamental administrative and clinical tasks, such as patient management, insurance claims, laboratory result management, and medical imaging inspection. Most of the hospitals have implemented EMR systems to a basic level that prioritizes data collection, management, and access. Nevertheless, the more advanced level of data management and use via CDRs and CDSS is not common. This can be seen as a barrier to the introduction of new DHTs in these hospitals. Along with the increased amount of data collected by the systems, the adoption of HL7, DICOM, and other international standards can be seen as a facilitator for new DHTs, such as AI-based CDSS, to be implemented in these hospitals.

XSL•FO
RenderX

Acknowledgments

This study was supported by funding from the Wellcome Trust United Kingdom (215010/Z/18/Z). The Vietnam Health Information Technology Maturity Model (Circular 54/2017/TT-BYT) was originally published in Vietnamese and was interpreted into English by DMT.

The members of *Vietnam ICU Translational Applications Laboratory (VITAL) investigators* are as follows (alphabetic order by surname):

OUCRU inclusive authorship list in Vietnam: Dang Phuong Thao, Dang Trung Kien, Doan Bui Xuan Thy, Dong Huu Khanh Trinh, Du Hong Duc, Ronald Geskus, Ho Bich Hai, Ho Quang Chanh, Ho Van Hien, Huynh Trung Trieu, Evelyne Kestelyn, Lam Minh Yen, Le Dinh Van Khoa, Le Thanh Phuong, Luu Hoai Bao Tran, Luu Phuoc An, Angela Mcbride, Nguyen Lam Vuong, Ngan Nguyen Lyle, Nguyen Quang Huy, Nguyen Than Ha Quyen, Nguyen Thanh Ngoc, Nguyen Thi Giang, Nguyen Thi Diem Trinh, Nguyen Thi Kim Anh, Nguyen Thi Le Thanh, Nguyen Thi Phuong Dung, Nguyen Thi Phuong Thao, Ninh Thi Thanh Van, Pham Tieu Kieu, Phan Nguyen Quoc Khanh, Phung Khanh Lam, Phung Tran Huy Nhat, Guy Thwaites, Louise Thwaites, Tran Minh Duc, Trinh Manh Hung, Hugo Turner, Jennifer Ilo Van Nuil, Vo Tan Hoang, Vu Ngo Thanh Huyen, Sophie Yacoub | Hospital for Tropical Diseases, Ho Chi Minh City: Cao Thi Tam, Ha Thi Hai Duong, Ho Dang Trung Nghia, Le Buu Chau, Le Mau Toan, Nguyen Hoan Phu, Nguyen Quoc Viet, Nguyen Thanh Dung, Nguyen Thanh Nguyen, Nguyen Thanh Phong, Nguyen Thi Cam Huong, Nguyen Van Hao, Nguyen Van Thanh Duoc, Pham Kieu Nguyet Oanh, Phan Thi Hong Van, Phan Tu Qui, Phan Vinh Tho, Truong Thi Phuong Thao | University of Oxford: Natasha Ali, James Anibal, David Clifton, Mike English, Ping Lu, Jacob McKnight, Chris Paton, Tingting Zhu | Imperial College London: Pantelis Georgiou, Bernard Hernandez Perez, Kerri Hill-Cawthorne, Alison Holmes, Stefan Karolcik, Damien Ming, Nicolas Moser, Jesus Rodriguez Manzano | King's College London: Liane Canas, Alberto Gomez, Hamideh Kerdegari, Andrew King, Marc Modat, Reza Razavi | University of Ulm: Walter Karlen | The University of Melbourne: Linda Denehy, Thomas Rollinson | Mahidol Oxford Tropical Medicine Research Unit (MORU): Luigi Pisani, Marcus Schultz

Data Availability

The data set generated and analyzed during this study is available upon request through the Oxford University Clinical Research Unit's managed data access scheme [41].

Authors' Contributions

All authors made substantial contributions to the conception and design of the study, drafted the work and revised it critically for important intellectual content, and provided final approval of the version to be published. DMT, NTD, CMD, HNH, LMK, NPH, DTTH, HTLT, TVD, and LMY were responsible for data acquisition and analysis, while DMT, CLT, and CP interpreted the data for the study. DMT, NTD, CMD, HNH, LMK, NPH, DTTH, HTLT, and TVD made the agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work were appropriately investigated and resolved. CLT and CP managed funding acquisition. The authors attest that there was no use of generative artificial intelligence technology in the generation of text, figures, or other informational content of this manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Vietnam HIT Maturity Model issued in 2017 by the Vietnam MoH. HIT: health information technology; MoH: Ministry of Health.
[DOCX File , 22 KB - formative_v9i1e53483_app1.docx]

Multimedia Appendix 2
The hospital health information systems assessment questionnaire.
[DOCX File , 21 KB - formative_v9i1e53483_app2.docx]

References

1. Global Initiative on Digital Health (GIDH) executive summary. World Health Organization. 2023 Feb 1. URL: https://cdn.who.int/media/docs/default-source/digital-health-documents/global-initiative-on-digital-health_executive-summary-31072023.pdf?sfvrsn=5282e32f_1 [accessed 2025-03-01]
2. Tran DM, Thwaites CL, Van Nuil JI, McKnight J, Luu AP, Paton C, Vietnam ICU Translational Applications Laboratory (VITAL). Digital health policy and programs for hospital care in Vietnam: scoping review. J Med Internet Res 2022 Feb 09;24(2):e32392 [FREE Full text] [doi: 10.2196/32392] [Medline: 35138264]

XSL•FO
RenderX

3. Moore GE. Cramming more components onto integrated circuits. Proc IEEE 1998 Jan;86(1):82-85. [doi: 10.1109/JPROC.1998.658762]
4. World Health Organization. Global Strategy on Digital Health 2020-2025. Geneva: World Health Organization; 2021.
5. Liverani M, Ir P, Perel P, Khan M, Balabanova D, Wiseman V. Assessing the potential of wearable health monitors for health system strengthening in low- and middle-income countries: a prospective study of technology adoption in Cambodia. Health Policy Plan 2022 Sep 13;37(8):943-951 [FREE Full text] [doi: 10.1093/heapol/czac019] [Medline: 35262172]
6. Schwalbe N, Wahl B. Artificial intelligence and the future of global health. Lancet 2020 May;395(10236):1579-1586. [doi: 10.1016/s0140-6736(20)30226-9]
7. Lee S, Kim H. Prospect of artificial intelligence based on electronic medical record. J Lipid Atheroscler 2021 Sep;10(3):282-290 [FREE Full text] [doi: 10.12997/jla.2021.10.3.282] [Medline: 34621699]
8. Bender D, Sartipi K. HL7 FHIR: an agile and RESTful approach to healthcare information exchange. 2013 Presented at: 26th IEEE International Symposium on Computer-Based Medical Systems; June 20-22, 2013; Porto, Portugal. [doi: 10.1109/cbms.2013.6627810]
9. EMRAM: a strategic roadmap for effective EMR adoption and maturity. HIMSS Eurasia. 2024 Sep 10. URL: https://himsseurasia.com/en/himss-emram-standards/ [accessed 2025-01-07]
10. Duncan R, Eden R, Woods L, Wong I, Sullivan C. Synthesizing dimensions of digital maturity in hospitals: systematic review. J Med Internet Res 2022 Mar 30;24(3):e32994 [FREE Full text] [doi: 10.2196/32994] [Medline: 35353050]
11. Thông t 54/2017/TT-BYT v B tiêu chí ng d ng công ngh thông tin t i các c s khám b nh, ch a b nh. Vietnam Ministry of Health. 2017. URL: https://thuvienphapluat.vn/van-ban/Cong-nghe-thong-tin/Thong-tu-54-2017-TT-BYT-Bo-tieu-chi-ung-dung-cong-nghe-thong-tin-tai-cac-co-so-kham-chua-benh-373292.aspx [accessed 2023-07-15]
12. Ming DK, Hernandez B, Sangkaew S, Vuong NL, Lam PK, Nguyet NM, Vietnam ICU Translational Applications Laboratory (VITAL) investigators. Applied machine learning for the risk-stratification and clinical decision support of hospitalised patients with dengue in Vietnam. PLOS Digit Health 2022 Jan 18;1(1):e0000005 [FREE Full text] [doi: 10.1371/journal.pdig.0000005] [Medline: 36812518]
13. The Vietnam ICU Translational Applications Laboratory (VITAL). Oxford University Clinical Research Unit. URL: http://vital.oucru.org/ [accessed 2023-07-01]
14. Van HMT, Hao NV, Phan Nguyen Quoc K, Hai HB, Khoa LDV, Yen LM, et al. Vital sign monitoring using wearable devices in a Vietnamese intensive care unit. BMJ Innov 2021 Mar 31;7(Suppl 1):s7-s11. [doi: 10.1136/bmjinnov-2021-000707]
15. Nguyen QH, Ming DK, Luu AP, Chanh HQ, Tam DTH, Truong NT, Vietnam ICU Translational Applications Laboratory (VITAL) investigators. Mapping patient pathways and understanding clinical decision-making in dengue management to inform the development of digital health tools. BMC Med Inform Decis Mak 2023 Feb 02;23(1):24 [FREE Full text] [doi: 10.1186/s12911-023-02116-4] [Medline: 36732718]
16. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2023.
17. Nakazawa M. Package 'fmsb': functions for medical statistics book with some demographic data. Comprehensive R Archive Network. 2023. URL: https://cran.r-project.org/web/packages/fmsb/fmsb.pdf [accessed 2025-01-07]
18. Saary M. Radar plots: a useful way for presenting multivariate health care data. J Clin Epidemiol 2008 Apr;61(4):311-317 [FREE Full text] [doi: 10.1016/j.jclinepi.2007.04.021] [Medline: 18313553]
19. Wickham H. ggplot2: Elegant Graphics for Data Analysis. New York, NY: Springer-Verlag; 2016.
20. Le N. C n c ch 70 b nh vi n dùng b nh án đi n t . VN Express. 2024. URL: https://vnexpress.net/ca-nuoc-chi-70-benh-vien-dung-benh-an-dien-tu-4732748.html [accessed 2024-09-10]
21. OxTREC application process. University of Oxford. 2024. URL: https://researchsupport.admin.ox.ac.uk/governance/ethics/apply/oxtrec#collapse404361 [accessed 2024-10-17]
22. Definition of terms: research, audit and service review. University of Oxford. 2024. URL: https://researchsupport.admin.ox.ac.uk/governance/ethics/apply/oxtrec/definitions [accessed 2024-10-18]
23. KPMG Department for International Trade, Oxford University Clinical Research Unit. Digital health in Vietnam. Market intelligence report. KPMG. 2020 Dec. URL: https://assets.kpmg.com/content/dam/kpmg/vn/pdf/publication/2021/digital-health-vietnam-2020-twopage.pdf [accessed 2025-01-07]
24. Thông t 46/2018/TT-BYT quy đ nh v h s b nh án đi n t . Vietnam Ministry of Health. 2018. URL: https://thuvienphapluat.vn/van-ban/Cong-nghe-thong-tin/Thong-tu-46-2018-TT-BYT-su-dung-va-quan-ly-ho-so-benh-an-dien-tu-391438.aspx [accessed 2023-07-15]
25. Quy t đ nh 5349/QĐ-BYT năm 2019 v phê duy t K ho ch tri n khai h s s c kh e đi n t . Vietnam Ministry of Health. 2019. URL: https://thuvienphapluat.vn/van-ban/The-thao-Y-te/Quyet-dinh-5349-QD-BYT-2019-phe-duyet-Ke-hoach-trien-khai-ho-so-suc-khoe-dien-tu-428071.aspx [accessed 2023-07-16]
26. Quy t đ nh 1923/QĐ-BYT năm 2023 phê duy t K ho ch tri n khai N n t ng H s s c kh e đi n t năm. Vietnam Ministry of Health. 2023. URL: https://thuvienphapluat.vn/van-ban/Cong-nghe-thong-tin/Quyet-dinh-1923-QD-BYT-2023-Ke-hoach-trien-khai-Nen-tang-Ho-so-suc-khoe-dien-tu-564088.aspx [accessed 2023-07-17]

XSL•FO
RenderX

27. Quy t đ nh 2035/QĐ-BYT năm 2013 công b Danh m c k thu t v ng d ng công ngh thông tin. Vietnam Ministry of Health. 2013. URL: https://thuvienphapluat.vn/van-ban/Cong-nghe-thong-tin/Quyet-dinh-2035-QD-BYT-nam-2013-Danh-muc-ky-thuat-ung-dung-cong-nghe-thong-tin-trong-y-te-199624.aspx [accessed 2024-09-01]
28. Hochwarter S, Cuong DD, Chuc NTK, Larsson M. Towards an electronic health record system in Vietnam: a core readiness assessment. J Health Inform Dev Countries 2014 Sep 21;8(2):93-103 [FREE Full text]
29. Nguyen KT. Electronic health record readiness assessment in Thai Binh Hospital, Vietnam. IJSR 2018 Mar;7(3):1277-1281. [doi: 10.21275/ART2018949]
30. Landgraf KM, Kakkar R, Meigs M, Jankauskas PT, Phan TTH, Nguyen VN, et al. Open-source LIMS in Vietnam: the path toward sustainability and host country ownership. Int J Med Inform 2016 Sep;93:92-102 [FREE Full text] [doi: 10.1016/j.ijmedinf.2016.06.010] [Medline: 27435952]
31. Vo C, CAO T, Doan N, Tu S, Nguyen A, Nguyen B, et al. EMRVisualization for patient progress tracking: interface design and system implementation. IJAE 2019;18(3):117-128. [doi: 10.5057/ijae.IJAE-D-18-00025]
32. Vu DH, Nguyen DT. Design of laboratory information system for healthcare in Vietnam BK-LIS. 2010 Presented at: International Conference on Communications and Electronics 2010; August 11-13, 2010; Nha Trang, Vietnam. [doi: 10.1109/icce.2010.5670692]
33. Hai VD, Thuan ND, Ngoc PP. Automatic retrieving data from medical equipment to create electronic medical records (EMRs) for an e-hospital model in Vietnam. 2011 Presented at: 5th International Symposium on Bio and Medical Informatics and Cybernetics (BMIC); July 19-22, 2011; Orlando FL.
34. Muinga N, Magare S, Monda J, English M, Fraser H, Powell J, et al. Digital health Systems in Kenyan public hospitals: a mixed-methods survey. BMC Med Inform Decis Mak 2020 Jan 06;20(1):2 [FREE Full text] [doi: 10.1186/s12911-019-1005-7] [Medline: 31906932]
35. Kose I, Rayner J, Birinci S, Ulgu MM, Yilmaz I, Guner S, HIMSS Analytics Team, Team MoH. Adoption rates of electronic health records in Turkish hospitals and the relation with hospital sizes. BMC Health Serv Res 2020 Oct 21;20(1):967 [FREE Full text] [doi: 10.1186/s12913-020-05767-5] [Medline: 33087106]
36. Miliard M. HIMSS helps Turkish hospitals go digital. Healthcare IT News. 2013 Nov 15. URL: https://www.healthcareitnews.com/news/himss-helps-turkish-hospitals-go-digital [accessed 2024-09-11]
37. Cresswell K, Sheikh A, Krasuska M, Heeney C, Franklin BD, Lane W, et al. Reconceptualising the digital maturity of health systems. Lancet Digital Health 2019 Sep;1(5):e200-e201. [doi: 10.1016/s2589-7500(19)30083-4]
38. Middleton B, Sittig DF, Wright A. Clinical decision support: a 25 year retrospective and a 25 year vision. Yearb Med Inform 2018 Mar 06;25(S 01):S103-S116. [doi: 10.15265/iys-2016-s034]
39. Chanh HQ, Ming D, Nguyen QH, Duc TM, An LP, Trieu HT, et al. Applying artificial intelligence and digital health technologies, Viet Nam. Bull World Health Org 2023 Jul 01;101(07):487-492. [doi: 10.2471/blt.22.289423]
40. Cresswell K, Williams R, Sheikh A. Developing and applying a formative evaluation framework for health information technology implementations: qualitative investigation. J Med Internet Res 2020 Jun 10;22(6):e15068 [FREE Full text] [doi: 10.2196/15068] [Medline: 32519968]
41. Data sharing policy. Oxford University Clinical Research Unit. URL: https://www.oucru.org/data-sharing-policy/ [accessed 2025-01-07]

Abbreviations

AI: artificial intelligence
CCD: Continuity of Care Document
CDA: Clinical Document Architecture
CDR: clinical data repository
CDSS: clinical decision support systems
DHT: digital health technology
DICOM: Digital Imaging and Communications in Medicine
EMR: electronic medical record
EMRAM: Electronic Medical Record Adoption Model
FHIR: Fast Healthcare Interoperability Resources
HIMSS: Healthcare Information and Management Systems Society
HIS: hospital information systems
HIT: health information technology
HL7: Health Level Seven
LAN: local area network
LIS: laboratory information systems
LMIC: low- and middle-income country
MoH: Ministry of Health

XSL•FO
RenderX

NAS: network attached storage
PACS: picture archiving and communication systems
RIS: radiology information systems
RIS-PACS: radiology information systems–picture archiving and communication systems
SAN: storage area network
SDG: Sustainable Development Goal
WHO: World Health Organization

Edited by A Mavragani; submitted 04.12.23; peer-reviewed by A Hadianfard, S Lloyd; comments to author 27.06.24; revised version received 11.09.24; accepted 13.11.24; published 06.02.25.

Please cite as:
Tran DM, Thanh Dung N, Minh Duc C, Ngoc Hon H, Minh Khoi L, Phuc Hau N, Thi Thu Huyen D, Thi Le Thu H, Van Duc T, VITAL (Vietnam ICU Translational Applications Laboratory) Investigators, Minh Yen L, Thwaites CL, Paton C
Status of Digital Health Technology Adoption in 5 Vietnamese Hospitals: Cross-Sectional Assessment
JMIR Form Res 2025;9:e53483
URL: https://formative.jmir.org/2025/1/e53483
doi:10.2196/53483
PMID:

©Duc Minh Tran, Nguyen Thanh Dung, Chau Minh Duc, Huynh Ngoc Hon, Le Minh Khoi, Nguyen Phuc Hau, Duong Thi Thu Huyen, Huynh Thi Le Thu, Tran Van Duc, VITAL (Vietnam ICU Translational Applications Laboratory) Investigators, Lam Minh Yen, C Louise Thwaites, Chris Paton. Originally published in JMIR Formative Research (https://formative.jmir.org), 06.02.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Adapting a Mobile Health App for Smoking Cessation in Black Adults With Anxiety Through an Analysis of the Mobile Anxiety Sensitivity Program Proof-of-Concept Trial: Qualitative Study

Marshall K Cheney[1], PhD; Adam C Alexander[2], PhD; Lorra Garey[3], PhD; Matthew W Gallagher[3,4], PhD; Emily T Hébert[5], PhD; Anka A Vujanovic[6], PhD; Krista M Kezbers[2], PhD; Cameron T Matoska[3], BS; Michael J Zvolensky[3,4,7], PhD; Michael S Businelle[2,8], PhD

[1]Department of Health and Exercise Science, University of Oklahoma, Norman, OK, United States

[2]Tobacco Settlement Endowment Trust (TSET) Health Promotion Research Center, University of Oklahoma Health Sciences Center, Oklahoma City, OK, United States

[3]Department of Psychology, University of Houston, Houston, TX, United States

[4]Helping Everyone Achieve a LifeTime of Health (HEALTH) Institute, University of Houston, Houston, TX, United States

[5]Department of Health Promotion and Behavioral Sciences, UTHealth School of Public Health, Austin, TX, United States

[6]Department of Psychological and Brain Sciences, Texas A&M University, College Station, TX, United States

[7]Department of Behavioral Science, The University of Texas MD Anderson Cancer Center, Houston, TX, United States

[8]Department of Family and Preventive Medicine, University of Oklahoma Health Sciences, Oklahoma City, OK, United States

Corresponding Author:
Marshall K Cheney, PhD
Department of Health and Exercise Science
University of Oklahoma
1401 Asp Ave
Norman, OK, 73019
United States
Phone: 1 4053255211
Email: marshall@ou.edu

Abstract

Background: At least half of smokers make a serious quit attempt each year, but Black adults who smoke are less likely than White adults who smoke to quit smoking successfully. Black adults who smoke and have high anxiety sensitivity (an individual difference factor implicated in smoking relapse and culturally relevant to Black adults) are even less successful. The Mobile Anxiety Sensitivity Program for Smoking (MASP) is a smoking cessation smartphone app culturally tailored to Black adults who smoke to increase smoking cessation rates by targeting anxiety sensitivity.

Objective: This study examined the acceptability and feasibility of the MASP smartphone app following a 6-week pilot test through postintervention qualitative interviews.

Methods: The MASP smoking cessation app was adapted from an evidence-based app by adding culturally tailored narration and images specific to the Black community, educational content on tobacco use in the Black community and the role of menthol, culturally tailored messages, and addressing tobacco use and racial discrimination. The MASP app was piloted with 24 adults with high anxiety sensitivity who identified as Black, smoked daily, and were not currently using medications or psychotherapy for smoking cessation. At the end of the 6-week pilot test, 21/24 participants (67% female; 95.2% non-Hispanic; mean age=47.3 years; 43% college educated; 86% single or separated) completed an audio-recorded semistructured interview assessing the acceptability and utility of the app, individual experiences, barriers to use, the cultural fit for Black adults who wanted to quit smoking, and identified areas for improvement. Transcribed interviews were coded using NVivo (Lumivero), and then analyzed for themes using an inductive, use-focused process.

Results: Most participants (17/21, 81%) had smoked for more than 20 years and 29% (6/21) of them smoked more than 20 cigarettes daily. Participants felt the MASP app was helpful in quitting smoking (20/21, 95%) and made them more aware of smoking thoughts, feelings, and behaviors (16/19, 84%). Half of the participants (11/21, 52%) thought the combination of medication and smartphone app gave them the best chance of quitting smoking. Themes related to participant experiences using

XSL•FO
RenderX

the app included establishing trust and credibility through the recruitment experience, providing personally tailored content linked to evidence-based stress reduction techniques, and self-reflection through daily surveys. The culturally tailored material increased app relevance, engagement, and acceptability. Suggested improvements included opportunities to engage with other participants, more control over app functions, and additional self-monitoring functions.

Conclusions: Adding culturally tailored material to an evidence-based mobile health (mHealth) intervention could increase the use of smoking cessation interventions among Black adults who want to quit smoking. Qualitative interviews provide mHealth app developers important insights into how apps can be improved before full study implementation and emphasize the importance of getting feedback from the target population throughout the development process of mHealth interventions.

Trial Registration: ClinicalTrials.gov NCT04838236; https://clinicaltrials.gov/ct2/show/NCT04838236

(JMIR Form Res 2025;9:e53566) doi:10.2196/53566

KEYWORDS

cultural tailoring; tailoring; African American; black; smoking cessation; mHealth; smartphone application; just in time adaptive intervention; qualitative; formative evaluation; app; application; anxiety; adult; qualitative analysis; smoking; mobile phone; tobacco

Introduction

Black or African American (hereafter referred to as Black) adults who smoke experience significant tobacco-related health disparities [1]. Black adults who smoke are at higher risk for tobacco-related death and disease, including lung cancer [2,3], heart disease, and stroke [4]. On average, Black adults who smoke initiate smoking later in life and smoke fewer cigarettes per day relative to their European American or White counterparts [5,6]. Yet, Black adults who smoke demonstrate greater cigarette dependence and are less likely to quit relative to other racial and ethnic groups [7,8], despite engaging in more quit attempts [8,9]. The difficulty this group faces in quitting successfully highlights the wide range of tobacco-related disparities that disproportionately impact Black adults who smoke (eg, tobacco advertising in predominately Black neighborhoods) [10-13].

One additional factor that can influence the success of a quit attempt and sustained smoking cessation among Black adults who smoke is anxiety sensitivity. Anxiety sensitivity is a relatively stable individual trait predisposing individuals to focus on internal bodily feelings of distress and to believe that anxiety and the bodily symptoms they experience from anxiety are uncommon or indicate an underlying threat to their health [14,15]. People who smoke and have high anxiety sensitivity can experience higher nicotine withdrawal symptoms and more intense cravings when they quit smoking [16], they are more prone to smoke when they have smoking-related thoughts, feelings, and sensations during cessation attempts, and they are more likely to smoke again soon after they have initiated a quit attempt [17-19]. Anxiety sensitivity is prevalent in up to a third of people who smoke across all races and ethnic groups [20,21]; however, Black people who smoke and have higher levels of anxiety sensitivity report experiencing the internal physical symptoms of nicotine withdrawal more intensely than Hispanic and White people who smoke [22,23]. Black people who smoke were also more likely to perceive these symptoms were out of their control or were worsened by environmental stressors such as racial discrimination [22,23]. Such processes could contribute to Black people being less successful in quitting smoking or put them more at risk of returning to smoking to help alleviate stress and anxiety [19].

One promising smoking cessation strategy for Black people with anxiety sensitivity is the delivery of tailored interventions through smartphone apps. Mobile health (mHealth) interventions can provide low-cost access to evidence-based smoking cessation strategies that extend the reach of traditional smoking cessation interventions. mHealth interventions can increase access for people who smoke but do not wish to engage in traditional smoking cessation interventions, those not interested in face-to-face interventions, and those who live in remote or underserved areas [24,25]. Smartphone smoking cessation apps have high acceptability in people who wish to quit smoking [25-27] and some smoking cessation apps have demonstrated efficacy in helping people to quit smoking, particularly for the most engaged app users [24,28-30]. While Black adults are underrepresented in mHealth studies [31], Black adults who have participated in mHealth studies viewed the interventions positively and found them to be acceptable and easy to use [32]. Thus, mHealth approaches may also reduce disparities in smoking cessation rates for Black people who smoke [33].

A benefit of mHealth smartphone apps is their potential to tailor intervention content to increase its relevance and acceptability for people who smoke. Tailoring involves selecting intervention content that corresponds to participant characteristics such as beliefs, identity, or readiness to change that can influence the success of a behavior change intervention and provide relevant content in the moments it is needed most [25,34-36]. Specifically, tailoring is the process of modifying the specific content a person receives in a mHealth smoking cessation intervention through *personalization* (based on regular participant assessments) and *feedback* (presenting participants with information about their behaviors over the course of the intervention) [34].

Cultural tailoring is an additional tailoring process of modifying an evidence-based mHealth behavioral intervention to increase the cultural relevance or cultural sensitivity of the intervention to a specific cultural group [37,38]. Cultural tailoring can include adjusting or adding intervention content to better match cultural and historical experiences, social norms and values, and

XSL•FO
RenderX

meanings of tobacco use to a specific cultural group [37]. Culturally tailored mHealth interventions can also address unique risk and protective factors, incorporate more inclusive language, and include culturally specific images and messages to decrease problems of acceptability, inclusion, and engagement [25,35,38,39]. Although it is not clear if culturally tailored materials increase the ultimate effectiveness of interventions, previous work has shown that Black people who smoke preferred culturally tailored content over standard content, which could ultimately increase acceptability and engagement with smoking cessation interventions, and improve retention rates [39-43]. Pilot testing culturally tailored interventions before full study implementation can verify if adapted content is acceptable and relevant to a cultural subgroup and can be used to obtain feedback on ways to improve the cultural relevance of the intervention [37,44].

The purpose of this study was (1) to assess participant reactions to an evidence-based smoking cessation smartphone app that was culturally tailored for Black people who smoke and have high anxiety sensitivity and (2) to provide recommendations for additional modifications to the app before full study implementation.

Methods

Mobile Anxiety Sensitivity Program for Smoking Adaptation

The Mobile Anxiety Sensitivity Program for Smoking (MASP) is a 6-week smoking cessation smartphone app for Black people with high anxiety sensitivity that was adapted from an evidence-based mHealth smartphone app for people with anxiety sensitivity [30,36,45,46]. The app contained a balance of standard intervention content for all participants (ie, on-demand tips, videos, and guided exercises) and individually tailored content based on the participant's responses to baseline, daily, and weekly assessments (ie, just-in-time messages; Figure 1) [46].

XSL•FO
RenderX

Figure 1. MASP (Mobile Anxiety Sensitivity Program) app components and cultural tailoring.

Component	Description	Cultural Tailoring	Participant Quotes (participant number)
Treatment Videos	16 4-6 minute intervention videos adapted from traditional anxiety sensitivity treatments: • relationship between stress and smoking • unhelpful thinking • myths about smoking • the importance of nicotine replacement therapy • strategies for cessation	Black male and female narrators for video content Education about menthol tobacco products, history of tobacco marketing to Black smokers, impact of racism and racial discrimination on stress and smoking	They gave me the history of African Americans of how things affect us differently…. How we wound up [with] menthol being the preferred cigarette… so that really made me wanna read more. P10
Coping Toolkit	• Guided relaxation and mindfulness exercises • challenging automatic thoughts • tips for coping with stress • managing uncomfortable sensations	Relating physical symptoms of the process of quitting smoking to anxiety, discrimination, and withdrawal.	just trying to get your body used to those sensations of being uncomfortable and being in that moment and it's OK to feel like this… it's gonna go away. P16
Quit Tips	On-demand education on nicotine replacement therapy and coping with smoking lapse triggers: • Nicotine medication tips • General Help for Quitting • Benefits of Quitting • Ways to Cope with Urges • Managing Stress/Mood • Motivate Me to Stay Quit • Harms of Smoking	Address concerns about pharmacotherapy, education about menthol	being around your spouse or your partner that smoke… social gatherings that smoke … it gave you an idea of how to deal with it. P5
Stress Management Trainings	3 x day participants received prompts for exercises, including: • over breathing • straw breathing • running in place • chair spinning	No tailoring – evidence-based standardized techniques	Just telling me how to handle this the stress that I was going through at the time and just breathing through it. P11
Daily Tailored Messages	Personally- tailored messages sent in response to EMA assessments. • inspirational messages • tailored to level of stress or craving • link to stress-management exercises	Mix of culturally tailored messages and evidence-based messages	it basically told me …how high the percentages from my culture and the bits of not smoking. P21

Based on the model proposed by Barrera and Castro [37], the MASP app was adapted for Black people who smoke with high anxiety sensitivity in an initial pilot study by incorporating culturally tailored content based on theoretical and empirical guidelines [47]. The MASP app was then reviewed by the University of Houston HEALTH Research Institute Community Research Advisory Board, which is part of a broader U54 addressing health equity. The Community Research Advisory Board provided input into video content and narration, clearer and more engaging language, and recruitment materials and strategies. From this input, key MASP app features were revised to include more culturally relevant content including treatment videos, messages, and on-demand content (Figure 1 and Multimedia Appendices 1 and 2). Examples of culturally tailored content included using Black male and female voices to narrate the videos, drawings, and illustrations of Black artists, the inclusion of additional videos addressing menthol tobacco products, the history of tobacco marketing to Black adults, and the impact of racism and discrimination on stress, smoking, and smoking relapse (Figure 1).

Expanded message content was created that related physical symptoms of quitting smoking to anxiety, discrimination, and severity of withdrawal symptoms, as well as messages relevant to Black culture (Textbox 1). Expanded educational content addressed topics relevant to Black people who smoke, including concerns about nicotine replacement therapy and medications to quit smoking. Several areas of the app were not modified, including skills-based approaches such as standard interceptive stress exposure videos, because of the universal applicability of the content [46]. The revised app reflected a balance of culturally tailored and standard material, consistent with other culturally adapted interventions [43].

XSL•FO
RenderX

Textbox 1. Examples of Mobile Anxiety Sensitivity Program culturally tailored messages for Black smokers.

- Your Tip: (FirstName), Did you know that 45,000 Black people die from smoking-related diseases each year in the US?
- Your Tip: (FirstName), 4 out of 5 Black smokers buy menthols. Chemicals are added to menthol cigarettes to give them a fresh, minty taste. This can make it easier for a smoker to inhale deeply, which may allow more harmful chemicals to enter the lungs.
- Your Tip: (FirstName), Did you know that the money that Black smokers spend on cigarettes in a single day could send 2500 Black students to college for an entire year?
- Your Tip: (FirstName), Tobacco companies have pushed menthol products on the Black community for decades. Did you know that there are significantly more menthol advertisements at stores in neighborhoods with a higher proportion of African American residents?
- Your Tip: (FirstName), Each year smoking kills more African Americans than AIDS, car accidents, fires, alcohol, drugs, and guns combined!
- Your Tip: (FirstName), Did you know African American children and adults are more likely to be exposed to secondhand smoke than any other racial or ethnic group? By quitting smoking you're not only making life better for yourself, but for the people around you too.
- Your Tip: (FirstName), Tobacco companies have targeted the African American community for decades. By quitting smoking, you're not only standing up for yourself but also for the members of your community.

Recruitment

Participants were recruited through Facebook (Meta) advertisements, mass emails through the University of Houston, and fliers posted in organizations and businesses in Houston communities with high concentrations of Black people. Individuals who saw the recruitment materials accessed the screening questionnaire through a posted link on the recruitment materials to assess their eligibility for the study. Individuals were eligible to participate in the pilot study if they (1) self-identified as Black or African American, (2) smoked daily (at least 10 cigarettes per day) for at least 2 years, (3) reported high Anxiety Sensitivity (Short Scale Anxiety Sensitivity Index score of at least 5), (4) endorsed a desire to quit smoking 2 weeks after completion of the baseline assessment, (5) were fluent in English, (6) had no current use of pharmacotherapy or psychotherapy for smoking cessation, and (7) had no cognitive impairment.

Fully enrolled participants (N=24) were instructed to download the INSIGHT (Insight Enterprises Inc) app to their personal smartphone, and those without a compatible phone were provided a study smartphone loaded with the MASP app. All participants were then asked to complete the baseline assessment in the app. Following baseline assessments and a final phone call appointment with project staff members, participants initiated the 6-week intervention. During the 6-week intervention, participants received 5 daily brief assessments (ie, 1 in the morning, 1 in the evening, and 3 at random times throughout the day) followed by individually tailored smoking cessation messages. Along with the daily surveys, participants were prompted to complete weekly follow-up assessments and self-reported smoking status. Participants also had access to stress management and relaxation videos, on-demand tips, strategies for smoking cessation and coping with cravings and urges, and no-cost nicotine replacement therapy. At the end of the intervention, participants used the app to complete a final assessment survey and were asked to participate in an online interview about their experiences with the app [46]. At the end of the 6-week intervention, participants were compensated US $45 for completing the week 6 survey, carbon monoxide breath test, and qualitative interview [46].

Qualitative Interviews

At the end of the 6-week pilot test, 21 of 24 (88%) participants agreed to participate in an interview about their experiences with the app. The interviews took place between March and August 2022. Two interviewers trained in qualitative interviewing techniques conducted the interviews over Zoom (Zoom Video Communications). The interviews with individual participants lasted an average of 44 minutes each. The interview question path was based on a previous study [48] and adapted to address the acceptability and utility of the app for smoking cessation, individual experiences with each component of the app, barriers to use, the relevance and fit for Black people who smoke, and identify areas for revision before full study implementation (Textbox 2 for question path excerpts).

Textbox 2. Excerpts from the Mobile Anxiety Sensitivity Program interview question path.

- What could have been done differently to help you complete more of the daily surveys?
- What are some specific things you liked about the app?
- What things did you like least about the app?
- If you had a slip-up during your quit attempt, was there ever a time that you didn't record that on the app?
- You received some different messages from the app during the study. Can you tell me about a time when you found the message to be really helpful when coping with stress?
- Can you tell me about a time when you found the message you received to be really helpful for quitting smoking?
- If you could add one thing to the app to make it more helpful to future smokers, what would you add?
- How relevant do you think the app is for Black Americans or African Americans who want to quit smoking?
- Do you have any suggestions for any way that we can improve the cultural tailoring of the app for Black Americans or African American smokers?

XSL•FO
RenderX

The audio-recorded interviews were transcribed verbatim, reviewed against the audio recording for accuracy, and then imported into NVivo (version 11; Lumivero) for data analysis. A qualitative researcher (MKC) first read through the transcripts, developed the codebook, and coded the transcripts in 2 coding passes. The codebook was developed using a usage-focused lens [49]. For the first coding pass, the individual interview questions were used as codes, which captured the specific functions of the MASP app, how participants interacted with the app, response to cultural tailoring, and the relevance of app content to Black people who wanted to quit smoking. The second coding pass used a pragmatism lens, coding for additional meanings and uses of the app across interview questions [50,51]. Following interview coding, a thematic analysis was conducted. The coded material was reviewed first within each code, then in groupings of related codes, and finally across all codes. Data displays were also created by grouping participant responses over a set of interview questions or codes to assess additional patterns of responses for individual participants [52]. A second qualitative researcher (AA) read the transcripts, reviewed the coding and primary data displays, and provided additional feedback. The qualitative researchers then arrived at the final determination of themes and quotes through an iterative process of feedback and discussion. A final review of the transcripts was then conducted for confirming and disconfirming evidence of themes [52]. Representative quotes were independently selected and distributed across participants.

Ethical Considerations

The University of Houston (reference number: 12747) and the University of Oklahoma Health Sciences Center (reference number: STUDY00000360) institutional review boards approved the protocol presented in this study. This trial was registered at ClinicalTrials.gov (NCT04838236).

Results

Participant Characteristics

Most (15/21, 71%) of the participants who completed the qualitative interviews were age 40 years or older (mean 47.33, SD=9.6), almost half (9/21, 43%) were college graduates, two-thirds were female (14/21, 67%), most participants were non-Hispanic (20/21, 95%), and reported being single (18/21, 86%). The majority smoked less than a pack of cigarettes per day (15/21, 71%), had smoked for 20 years or longer (17/21, 81%), smoked menthol cigarettes (20/21, 95%), and smoked an average of 15.48 (SD 5.02) cigarettes per day. Participants had moderate levels of nicotine dependence (mean 3.10, SD 1.45), and anxiety sensitivity (mean 11.00, SD 5.40). Most participants were highly motivated to quit smoking (Table 1).

XSL•FO
RenderX

Table 1. Interview participant characteristics (n=21).

Participant characteristics	Values
Age (years), mean (SD)	47.33 (9.6)
Age groups (years), n (%)	
30-39	6 (29)
40-49	6 (29)
50 or more	9 (43)
Education, n (%)	
Less than high-school degree	2 (10)
High-school graduate or General Educational Development degree	4 (19)
Some college	6 (29)
College or graduate degree	9 (43)
Sex, n (%)	
Female	14 (67)
Male	7 (33)
Ethnicity, n (%)	
Non-Hispanic	20 (95)
Hispanic	1 (5)
Marital status, n (%)	
Married or living with significant other	3 (14)
Single or separated	18 (86)
Employment, n (%)	
Regular full-time work	7 (33)
Regular part-time work	3 (14)
Stay-at-home caregiver	2 (10)
Unemployed or unable to work	9 (43)
Average number of cigarettes smoked per day, n (%)	
10-19	15 (71)
20 or more	6 (29)
Years smoked, n (%)	
10-19	4 (19)
20-29	12 (57)
30 or more	5 (24)
Motivation to quit (range 1-10), n (%)	
5-7.9	3 (14)
8-9.9	11 (52)
10	7 (33)

Before the start of the intervention, participants were asked about their preferred method for quitting smoking. While 17/21 indicated medications (81%) as their preferred cessation method, 11 (52%) also selected smartphone app. Half (n=11) of participants also indicated that the combination of medications and smartphone app would give them the best chance of quitting smoking (Table 2). Following the 6-week intervention, the majority of participants (16/19, 85%) thought that having the app and answering the questions increased their awareness of their smoking behaviors and 20/21 interviewees reported that the app was helpful in quitting smoking. The mean score on the African American Acculturation Scale was 15, indicating a medium level of Black cultural identification [53].

XSL•FO
RenderX

Table 2. Interview participant beliefs (n=21).

Participant beliefs	Values
Preintervention questions, n (%)	
If you were to quit smoking, which of the following would you prefer to receive?	
Medications	17 (81)
Group counseling	4 (19)
Smartphone app	11 (52)
In-person individual counseling	6 (29)
Smoking cessation telephone helpline	6 (29)
Which option would give you the best chance of quitting smoking?	
Medications	4 (19)
Both medications and counseling	4 (19)
Both smartphone apps and medications	11 (52)
Quitting “cold turkey” without counseling or medications	2 (10)
Postintervention questions, n (%)	
Consider the number of assessments that were automatically pushed by the smartphone app, was the number of assessments.	
Too high	11 (55)
About right	8 (40)
Not enough	1 (5)
Did carrying the phone and answering questions make you more aware of your thoughts, feelings, and behavior?	
Definitely or mostly no	3 (16)
Definitely or mostly yes	16 (84)
Overall, how helpful has the smartphone app been in helping you to quit smoking?	
Not at all helpful	1 (5)
Slightly or moderately helpful	5 (25)
Very or extremely helpful	14 (70)
African American Acculturation Scale score range (mean 15, SD 6.5, range 6-27)	
Low (score 1-9)	5 (24)
Medium (score 10-18)	9 (43)
High (score 19-27)	7 (33)

Themes

Thematic analysis identified 6 themes related to participant experiences using the app which are (1) recruitment experience, (2) overall experience with the app, (3) most helpful aspects of the app, (4) reaction to the culturally-tailored material, (5) experience with the daily surveys and other assessments, and (6) suggested changes to make the app more helpful or relevant to Black people who want to quit smoking (Figure 2 for a summary of themes).

XSL•FO
RenderX

Figure 2. Summary of themes.

Theme	Brief Summary	Example Quotes
Recruitment experience	Participants stressed the importance of determining study legitimacy and credibility through the professional appearance of the recruitment materials and linkage to an academic institution. Participants were also attracted to the novel intervention delivered by smartphone app.	*it was through the [University] … the lab was a legit thing, I felt much more confident. P12* *I've never seen a app to help you quit anything before* P17
Overall experience with the app	Ease of use and the flexibility and privacy the app provided. The supportive, non-judgmental tone of the messages and other content kept participants motivated and engaged even if they were not able to quit. Daily surveys and messages promoted self-reflection.	*it made me become accountable. It made me think about what I was doing… It made me just take a look at my patterns P10*
Most helpful aspects of the app	The integrated content linking messages and on-demand relaxation and stress reduction exercises increased coping skills in stressful situations and provided alternative strategies to smoking. The personally tailored daily messages increased the relevance of the intervention content. Having the app with them helped them avoid smoking in the moment.	*after I did my survey it was like do you want to do a coping exercise? P18* *I get in the house. First thing…and smoke a cigarette. With the app, I wouldn't do that. I would go to the app first …. using meditations on there. P8*
Reaction to culturally tailored material	Culturally tailored material increased relevance and engagement with the content. To increase relevance to Black smokers, participants suggested adding quotes or stories from Black former smokers, increasing the diversity of images, including more Black health care professionals, incorporate Black vernacular, and provide more information on tobacco use in the Black community.	*I think it's good is because you get the history, you get the fact that somebody understands. P9* *when you open up the app and being able to be it's talking like me it's looking like me. P12*
Experience with daily surveys and other assessments	Surveys increased self reflection and accountability. Suggestions to improve survey completion rates were to decrease survey frequency and question repetition, increase the window for response time, and make survey times more flexible to accommodate work schedules.	*it was more helpful to me answering them… made me not want to smoke. P15* *after so many weeks of doing the same thing ... that's when I started losing some focus. P11*
Suggested changes to improve the app	Provide an option to interact with other participants through the app. Increase self-monitoring functions to track progress.	*some way to be able to connect with …how has everyone done today? P18*

Recruitment Experience

Participants were asked about their initial reaction to the recruitment materials and what prompted them to complete the screening. Important factors in the recruitment process included the legitimacy of the study, the intervention was delivered using a smartphone app, and the inclusion of more familiar traditional nicotine replacement products.

Participants described the professional appearance of the recruitment materials, direct linkage with a university, and that the screening questions matched the stated purpose of the study as important factors in their initial participation. Participants were positive about being part of a real research study but did not want to be drawn into a for-profit research study. Two participants mentioned that they were skeptical but curious, looking up the university and the specific lab to make sure it was an actual study.

> *I thought this is probably a scam. But then I thought, okay, well I just won't give anything too personal, and as I started completing the survey, the questions*

XSL•FO
RenderX

> *were, they sounded legit. They were actually about my habit of smoking, and it just sounded scientific enough that I thought, okay yeah, this probably isn't a scam …. even though it was on Craigslist.* [Participant 12 (P12)]

Participants were interested in the study when they learned the intervention would be delivered using a smartphone app. This was a new strategy when others had not worked and would allow them more independence during the intervention.

Overall Experience With the App

Participants commented that they liked the app and found it easy to use and navigate, particularly for adults who are not "tech savvy" or who have not had much experience using smartphone apps. They reported that the app motivated them to keep trying to quit throughout the intervention, even if they were unable to completely quit smoking. Participants also said the supportive, encouraging, nonjudgmental tone of the app was very important to reporting their true behaviors and feelings.

> *the app … gives you the encouragement of why you don't have to do that [smoking]. There are other things you can do until you get better at not smoking… to help you along the way during that little minutes when it's difficult.* [P10]

> *it's not belittling you or anything and still giving you encouragement … they make you change your thought process into a positive one, so I think that's amazing.* [P16]

Over half of the participants also said that what they liked most was that the app helped them learn more about themselves and their smoking, made them aware of their smoking behaviors and habits, and "kept me accountable to myself."

> *The app made you think about the smoking you were doing. Normally you would just pick up a cigarette and smoke, but the app forces you like OK, do I really need that cigarette right now…. You're thinking about it more. Well, whereas smoking was just an extension of you before.* [P20]

About a third of participants mentioned what they liked the most was that the app was something they could do privately rather than with a group of others, trying different options when they were ready and at their own pace.

> *I was in a group before, and I didn't really get too much out of it… I'm by myself and I'm doing this alone. I'm with the app, I can think more clearly, and I think this is better than being in the group.* [P2]

> *It gives you anonymity. You don't have to tell anybody…It allowed me to be independent, and not have to speak to someone every day … it just allowed me to do it basically at my pace. I didn't feel pressured that I had to do it.* [P10]

A third of the participants mentioned the nicotine replacement products integrated into the intervention.

> *throughout the day, it would remind me to take more lozenges … that was cool.* [P6]

Most Helpful Aspects of the App

Participants found the tailored app content helpful for their current challenges with stress, anxiety, and smoking even if they were not able to quit smoking during the pilot study. Participants commented that the combination of app-directed content and access to complementary on-demand content was a real strength of the intervention. When participants mentioned a specific part of an app that was most helpful to them, they most often mentioned the relaxation and stress reduction exercises and videos, on-demand information that provided them with tips and suggestions, education about tobacco use, and the daily messages (Table 1).

> *They are targeted to what you put into the app yourself, you would say my anxiety is so much or I'm not able to deal with my anxiety. They would give you tips for that and I thought that was really helpful.* [P9]

> *It would tell you after you did the assessments, would you like to take a relaxation or would you like to go or just to go to your toolkit, and that would remind you after you took the assessments to actually do that.* [P14]

However, if the participant was at work, driving, or busy with family responsibilities, they found it difficult to read or act on a particular message at that time.

Across the different parts of the app, participants said that it was the content related to coping that was particularly helpful, either in the form of a tailored supportive message, developing a new skill to reduce stress or fight cravings, or through exercises to challenge unhelpful thoughts.

> *I feel like my anxiety leads into anger sometimes and that [the app] makes you sit there and think about it instead of you being in that moment and being so upset that you can't comprehend what's going on or be logical about your thinking … Just think about it…let's do this little activity.* [P16]

> *It would pose a question …It would give you a statement and ask you, do you believe this to be true … that really stuck with me. If you smoked one cigarette … do you believe that you could still not smoke the rest of the day … I really believe, once I've smoked one, I'm a smoker there's no stopping… I don't really challenge my way of thinking…There is a positive side if you would just open your eyes to it.* [P12]

Others said that the most helpful part of the intervention was having it with them all the time, to use at the moment they needed it.

> *Remove yourself from the situation… if you can just go to the app in that moment and have that put into your head, that's what's helpful … because it helps in that moment.* [P8]

> *When I was feeling stressed I pushed the button ... just stressed out like where you think you're gonna light up a cigarette and smoke … you push that button.* [P3]

XSL•FO
RenderX

Reactions to Cultural Tailoring

Participants were asked about the cultural tailoring of the app to reflect Black culture. Almost all participants said cultural tailoring contributed to their engagement with the app.

> *I really like this app being geared toward the Black people, I mean because it is a different thing. The high blood pressure and stress and the discrimination is a totally different thing than what White people face on a daily basis.* [P4]

Participants were then asked about the relevance of the tailored intervention components to smokers who are Black.

> *explaining about the history of cigarettes within the African American community, the effect that it had within the community … I feel like it's very relevant. There's a lot of facts in there that you wouldn't even think about as a Black person. There's a lot of things that were directed towards Black people…they portray it in media. I always thought that it was portrayed more for white people … for a Black person that actually smokes, I think it would be eye opening, it would be very, very helpful for them.* [P16]

> *A lot of the tips were directly towards African Americans.* [P17]

Participant recommendations to increase the relevance of the MASP app for Black people who want to quit smoking included hearing from Black former smokers, either through short videos added to the app, seeing former participant pictures added to the app, or brief quotes or stories from former participants as messages or on-demand content. These pictures and stories could help the participants connect to the project and the app and encourage them during the tough parts of the quitting process.

> *past users of the actual app that are African American that have been successful with it… if they [participants] have pictures of past people that have actually accomplished this, they'll be more inclined to use it…African American[s] really use this app, his picture's right there… it worked for him. It might work for me.* [P21]

Other suggestions to increase relevance for Black people who smoke included more overall diversity in photographs and videos throughout the app, especially more photographs and videos using Black people who smoke and former smokers, as well as Black health care professionals.

> *It's gotta be by Black people for black people to pay attention... I think a lot of African Americans are, if it's not something that looks familiar or sounds familiar, then they're scared to go around it.* [P12]

A few participants suggested additional information on menthol tobacco use that is highly prevalent.

> *I would just emphasize menthols and maybe Black and Milds? On top of cigarettes, you know that that those are just as deadly.* [P18]

Several participants also suggested incorporating Black vernacular in videos and messaging.

> *A more relatable language to the African American community …how we speak like there's certain words and lingo, maybe just incorporating that… it might appeal to more African Americans... they could read it and be like oh they know what I'm talking about.* [P16]

Experience With Daily Surveys and Other Assessments

All of the participants found the surveys easy to complete. Many of the participants thought the daily assessments were a part of the intervention content. Participants described the surveys as holding them more "accountable" and keeping them aware of their smoking behaviors and how much they were thinking about smoking. Several discussed how the survey questions helped them ask for help or remove themselves from a situation where they were more likely to smoke. Most responded quite positively to the idea of the daily surveys, but many participants did not like the amount of repetition of the survey questions and the number of surveys each day. The repetition of the questions over the 6 weeks of the study, which participants acknowledged as necessary to track their progress over time, was often experienced as frustrating or boring in the later weeks of the study.

> *I knew what questions were coming up and I didn't even have to read the questions anymore. I was just answering them.* [P20]

> *the amount of times you guys asked me a day if I'm depressed …it was a bit excessive. You're going to get the same answer 2 hours or three hours from now … so you might as well just ask me in the middle of the day … it was exhausting.* [P7]

Participants provided several suggestions to hold their interest in the surveys and help them complete more of the surveys throughout the study.

> *a different set of completely different set of questions weekly or change them every day.* [P4]

> *have more of a variety of questions…maybe add something, maybe switch them up a little bit.* [P18]

Participants expressed challenges with what they perceived as a short response window to start the surveys following notification, particularly the surveys that happened during the workday. Participants gave examples of missing surveys because they were driving, with customers, at work, or with family where they could not immediately stop to take the survey or press a button to indicate they would like to take the survey later. Participants also asked for more control over when to take the surveys during the day.

> *Give us more than five minutes to pull our cars over … there were quite a few times I'm on the highway or I'm on the other line with a doctor's office or kids school and I need more than five minutes.* [P7]

Suggested Changes to Improve the Mobile Anxiety Sensitivity Program for Smoking App

The most commonly suggested improvement was a feature to allow interaction with other participants. Other app improvements included more self-monitoring functions, more

XSL•FO
RenderX

pictures and color, and options for longer versions of the training, such as meditation and coping skills.

Hearing from other smokers trying to quit: Participants suggested connecting with other app users to chat, establish a buddy system within the app, or just talk through their challenges and hear from others. They suggested these contacts could take place within the app, be voluntary, and be anonymous to protect participant identities.

> *set up a buddy system when the person signs up for the study … get permission consent from both parties, if you need to talk to somebody live, this is your buddy.* [P10]

> *Maybe like a chat line…go into an open forum. Say we speak to 'em, just say whatever to chit chat, talk… moral support.* [P21]

Self-monitoring: Several participants asked if there was a way they could look at their survey data to see their trends in how many cigarettes they were smoking, how they were feeling at the time, and what seemed to trigger their smoking. Other participants suggested an optional journaling function where they could write about their experiences and review their entries later.

> *The option …to really track your smoking … more on an individual basis so you can actually see …how you were smoking or not or how long you were going without smoking, … especially your progress because I think that's the most motivating to see that you're actually doing it.* [P8]

Discussion

Principal Results

Black people are disproportionally affected by tobacco use and tobacco-related health conditions but historically have been underrepresented in tobacco cessation research [31]. The MASP smoking cessation app was shown to be acceptable and culturally fit with Black adults, and study participants provided recommendations to increase MASP's relevance to Black people who want to quit smoking.

Given the underrepresentation of Black adults in intervention research, generally, understanding what influences potential participants to complete study screening surveys or to contact study staff members is helpful in increasing participation in clinical trials among Black communities. Participants indicated that it is important that recruitment materials and app content be designed to establish study legitimacy. This includes the professional appearance of recruitment materials and linking the study with a specific university. One recommendation was to include testimonials from previous study participants, which is consistent with other mHealth studies [32,38,41].

Throughout the interviews, participants discussed trust and credibility regarding the intervention and its content. Similar to other studies focused on Black adults, the credibility of the information in the app was often assessed by how well the information in the app matched their own experiences and the source of the information [54,55]. This may be especially important for cultural groups such as Black adults who may experience suspicions or distrust of the medical community due to centuries of historical oppression and systemic racism [56]. To increase credibility and relevance to Black people who smoke, participants recommended including more images reflective of Black adults but also suggested increasing the diversity of the photos and videos beyond Black people to include more photographs and videos of Black health care professionals. However, participants in another study cautioned researchers not to use stock photos or stereotypical images, but to use images and photographs reflective of the app users [32,38].

MASP participants also suggested incorporating more language and statistics on Black people who smoke and more discussion of disparities and targeting by the tobacco industry into the intervention messages to increase engagement and acceptability of the app [32,41]. One novel strategy to increase the number of culturally relevant messages in other mHealth studies involved recruiting a group of Black people who smoke and those who have successfully quit smoking to generate motivational messages for other Black people trying to quit [57-59]. These messages could also be used to incorporate more culturally relevant content and Black vernacular into the app [59]. However, it should be noted that a small number of participants (2 out of 21 in the current study) did not react positively to the explicit cultural tailoring. Diversity of opinions and cultural identity would be expected in any cultural group, so testing messaging and educational content within the app and assessing a participant's degree of cultural identification at baseline would allow researchers to match culturally relevant messages by level of cultural identification so that participants could receive appropriately tailored messages for their level of cultural identification.

When asked what additional features they would add to the app, participants wanted a way to interact with other participants who were using the app to try to stop smoking. This was interesting, given participants' comments that they liked using the app because it allowed them privacy and flexibility. Interaction with other study participants within the app may increase cultural relevance as participants would interact with others from their cultural group who were embarking on a shared lived experience of smoking cessation. Other mHealth researchers have suggested a discussion board or group chats moderated by study staff members so that misinformation and unhealthy conversations could not be introduced [27,38]. A discussion board could be an easy option to allow participants to interact with each other [32,41,55,60]. These interactions could also be collected for additional qualitative data analysis. Researchers might want to consider a different discussion board for the control and intervention participants or only for intervention participants. This type of feature has not been well used in previous smoking cessation apps [61,62].

MASP participants also asked for tools that could summarize their progress within the app (eg, cigarette and other tobacco product tracking, money saved, and tracking triggers for smoking). This was also suggested by participants in other mHealth substance use studies [38,41,54,55,63]. Self-tracking or self-monitoring functions could help participants increase

XSL•FO
RenderX

their awareness of substance use and personal feelings of accountability for their behavior and develop plans to cope with personal smoking triggers [29]. Both personal accountability and increased awareness of behaviors were named by participants in this and other substance use mHealth studies as something they liked best about the study [54].

MASP participants also recommended prompting fewer surveys per day and adding or substituting questions within the surveys to increase variety and engagement throughout the study. One concern expressed by MASP participants was to make the surveys equally accessible to those who do not have as much control over their daily work schedule, as they cannot always stop what they are doing to complete surveys. This could apply to those who have changing shifts and scheduled work hours, those who drive a lot, are in customer service, and cannot stop serving customers to take surveys immediately, or those who are in hourly positions with less flexibility when they can take short breaks. This was also noted as a challenge in other mHealth interventions [30]. Similarly, MASP participants suggested an option to "try later" for surveys sent when participants could not or did not want to complete them; however, Bendotti et al [55] cautioned app developers to balance additional time to complete surveys with the number of daily notifications from the app as many users did not like frequent notifications. Rather than delaying surveys and messages, another option MASP participants suggested was to allow participants to keep the messages and surveys in a "bank" that they could come back to later that day. However, the MASP app prompted participants with 5 assessments and tailored messages throughout the day, so delaying them until the end of the day would mean participants would not be recording their experiences at the most stressful and potentially most meaningful moments during the day when tailored messages could be most needed to prevent smoking lapse.

Limitations

Limitations of the study's qualitative data analysis included the relatively small number of interviews, which limited generalizability to the larger population of Black people who smoke cigarettes. An additional limitation is that the 2 interviewers were not Black. This may have affected participant feedback, both constructive and positive. However, interviewers of the same cultural group as the participants can take unconscious "cultural shortcuts" in communication with participants or make assumptions about participant responses and move too quickly through participant responses without sufficient probing of responses. One suggestion is to have a diverse team of qualitative interviewers and analysts, which may lead to richer data collection and analysis.

Conclusions

In conclusion, MASP is a feasible and acceptable culturally tailored smoking cessation app for anxiety-sensitive Black adults who want to quit smoking. The development of evidence-based, culturally tailored mHealth apps such as MASP could help to reduce health disparities by improving the cultural relevance of mHealth interventions. Tailored mHealth interventions could decrease barriers to accessing smoking cessation interventions and could increase engagement and acceptability in population subgroups who experience higher burdens of health-harming behaviors and diseases [25,38].

Acknowledgments

AA, LG, MG, EH, AV, KK, CM, MZ, and MSB contributed to the intervention design and implementation; LG and CM collected the qualitative data and MKC and AA were the primary qualitative analysts with LG and CM. All authors contributed to the preparation of the manuscript.

This study was supported by funding from the National Institute on Minority Health and Health Disparities (U54MD015946: Principal Investigator Obasi; current study Multiple Principal Investigators (MPIs) Businelle and Zvolensky), National Cancer Institute (NCI) grant P30CA225520 awarded to the Stephenson Cancer Center, the Tobacco Settlement Endowment Trust (MSB), and the work used the Stephenson Cancer Center's mHealth Shared Resource. The manuscript preparation was additionally supported by grants to AA from the National Institute on Minority Health and Health Disparities (K01MD015295) and ETH from the National Institute on Drug Abuse (K99/R00DA046564). Work by LG on the manuscript was supported by the National Institute on Drug Abuse of the National Institutes of Health under Award Number R25DA054015 (MPIs: Obasi and Reitzel).

Conflicts of Interest

MSB is one of the inventors of the Insight mobile health platform and receives royalties related to the use of this platform by investigators external to the University of Oklahoma Health Sciences Center. As MSB is a multiple principal investigator in this study, he did not receive royalties for using the platform to create the Mobile Anxiety Sensitivity Program for Smoking app.

Multimedia Appendix 1
Examples of MASP (Mobile Anxiety Sensitivity Program) app menus.
[PNG File , 195 KB - formative_v9i1e53566_app1.png]

Multimedia Appendix 2
Sample MASP (Mobile Anxiety Sensitivity Program) video clip.
[MP4 File (MP4 Video), 2190 KB - formative_v9i1e53566_app2.mp4]

XSL•FO
RenderX

References

1. Martell BN, Garrett BE, Caraballo RS. Disparities in adult cigarette smoking-United States, 2002-2005 and 2010-2013. MMWR Morb Mortal Wkly Rep 2016;65(30):753-758 [FREE Full text] [doi: 10.15585/mmwr.mm6530a1] [Medline: 27491017]
2. Haiman CA, Stram DO, Wilkens LR, Pike MC, Kolonel LN, Henderson BE, et al. Ethnic and racial differences in the smoking-related risk of lung cancer. N Engl J Med 2006;354(4):333-342. [doi: 10.1056/NEJMoa033250] [Medline: 16436765]
3. Howlader N, Mariotto AB, Woloshin S, Schwartz LM. Providing clinicians and patients with actual prognosis: cancer in the context of competing causes of death. J Natl Cancer Inst Monogr 2014;2014(49):255-264 [FREE Full text] [doi: 10.1093/jncimonographs/lgu022] [Medline: 25417239]
4. Novotny TE, Warner KE, Kendrick JS, Remington PL. Smoking by blacks and whites: socioeconomic and demographic differences. Am J Public Health 1988;78(9):1187-1189. [doi: 10.2105/ajph.78.9.1187] [Medline: 3407817]
5. Caraballo RS, Giovino GA, Pechacek TF, Mowery PD, Richter PA, Strauss WJ, et al. Racial and ethnic differences in serum cotinine levels of cigarette smokers: third national health and nutrition examination survey, 1988-1991. JAMA 1998;280(2):135-139. [doi: 10.1001/jama.280.2.135] [Medline: 9669785]
6. Schoenborn CA, Adams PF, Peregoy JA. Health behaviors of adults: United States, 2008-2010. Vital Health Stat 10 2013(257):1-184 [FREE Full text] [Medline: 25116426]
7. Webb Hooper M, Baker EA, McNutt MD. Associations between coping, affect, and social support among low-income African American smokers. Addict Behav 2013;38(11):2736-2740. [doi: 10.1016/j.addbeh.2013.07.005] [Medline: 23934005]
8. Babb S, Malarcher A, Schauer G, Asman K, Jamal A. Quitting smoking among adults - United States, 2000-2015. MMWR Morb Mortal Wkly Rep 2017;65(52):1457-1464 [FREE Full text] [doi: 10.15585/mmwr.mm6552a1] [Medline: 28056007]
9. Bacio GA, Guzman IY, Shapiro JR, Ray LA. Differences in quit attempts between non-hispanic black and white daily smokers: the role of smoking motives. Addict Behav 2014;39(12):1769-1772 [FREE Full text] [doi: 10.1016/j.addbeh.2014.07.001] [Medline: 25123344]
10. Lee JGL, Henriksen L, Rose SW, Moreland-Russell S, Ribisl KM. A systematic review of neighborhood disparities in point-of-sale tobacco marketing. Am J Public Health 2015;105(9):e8-e18 [FREE Full text] [doi: 10.2105/AJPH.2015.302777] [Medline: 26180986]
11. Wheeler DC, Boyle J, Barsell DJ, Glasgow T, McClernon FJ, Oliver JA, et al. Associations of alcohol and tobacco retail outlet rates with neighborhood disadvantage. Int J Environ Res Public Health 2022;19(3):1134 [FREE Full text] [doi: 10.3390/ijerph19031134] [Medline: 35162162]
12. Cwalina SN, Ihenacho U, Barker J, Smiley SL, Pentz MA, Wipfli H. Advancing racial equity and social justice for black communities in US tobacco control policy. Tob Control 2023;32(3):381-384 [FREE Full text] [doi: 10.1136/tobaccocontrol-2021-056704] [Medline: 34526408]
13. Mattingly DT, Mezuk B, Elliott MR, Fleischer NL. Discrimination and tobacco use outcomes among US adults: effect modification by race/ethnicity. J Racial Ethn Health Disparities 2024;11(1):395-405 [FREE Full text] [doi: 10.1007/s40615-023-01527-3] [Medline: 36696079]
14. Leventhal AM, Zvolensky MJ. Anxiety, depression, and cigarette smoking: a transdiagnostic vulnerability framework to understanding emotion-smoking comorbidity. Psychol Bull 2015;141(1):176-212 [FREE Full text] [doi: 10.1037/bul0000003] [Medline: 25365764]
15. Langdon KJ, Farris SG, Hogan JB, Grover KW, Zvolensky MJ. Anxiety sensitivity in relation to quit day dropout among adult daily smokers recruited to participate in a self-guided cessation attempt. Addict Behav 2016;58:12-15 [FREE Full text] [doi: 10.1016/j.addbeh.2016.02.017] [Medline: 26896560]
16. Johnson KA, Stewart S, Rosenfield D, Steeves D, Zvolensky MJ. Prospective evaluation of the effects of anxiety sensitivity and state anxiety in predicting acute nicotine withdrawal symptoms during smoking cessation. Psychol Addict Behav 2012;26(2):289-297 [FREE Full text] [doi: 10.1037/a0024133] [Medline: 21644805]
17. Zvolensky MJ, Farris SG, Schmidt NB, Smits JAJ. The role of smoking inflexibility/avoidance in the relation between anxiety sensitivity and tobacco use and beliefs among treatment-seeking smokers. Exp Clin Psychopharmacol 2014;22(3):229-237 [FREE Full text] [doi: 10.1037/a0035306] [Medline: 24490706]
18. Johnson KA, Farris SG, Schmidt NB, Zvolensky MJ. Anxiety sensitivity and cognitive-based smoking processes: testing the mediating role of emotion dysregulation among treatment-seeking daily smokers. J Addict Dis 2012;31(2):143-157 [FREE Full text] [doi: 10.1080/10550887.2012.665695] [Medline: 22540436]
19. Assayag Y, Bernstein A, Zvolensky MJ, Steeves D, Stewart SS. Nature and role of change in anxiety sensitivity during NRT-aided cognitive-behavioral smoking cessation treatment. Cogn Behav Ther 2012;41(1):51-62. [doi: 10.1080/16506073.2011.632437] [Medline: 22375732]
20. Allan NP, Raines AM, Capron DW, Norr AM, Zvolensky MJ, Schmidt NB. Identification of anxiety sensitivity classes and clinical cut-scores in a sample of adult smokers: results from a factor mixture model. J Anxiety Disord 2014;28(7):696-703 [FREE Full text] [doi: 10.1016/j.janxdis.2014.07.006] [Medline: 25128664]

XSL•FO
RenderX

21. Farris SG, DiBello AM, Allan NP, Hogan J, Schmidt NB, Zvolensky MJ. Evaluation of the anxiety sensitivity index-3 among treatment-seeking smokers. Psychol Assess 2015;27(3):1123-1128 [FREE Full text] [doi: 10.1037/pas0000112] [Medline: 25894700]
22. Bello MS, Pang RD, Cropsey KL, Zvolensky MJ, Reitzel LR, Huh J, et al. Tobacco withdrawal amongst African American, hispanic, and white smokers. Nicotine Tob Res 2016;18(6):1479-1487 [FREE Full text] [doi: 10.1093/ntr/ntv231] [Medline: 26482061]
23. Reitzel LR, Childress SD, Obasi EM, Garey L, Vidrine DJ, McNeill LH, et al. Interactive effects of anxiety sensitivity and subjective social status on psychological symptomatology in black adults. Behav Med 2017;43(4):268-276 [FREE Full text] [doi: 10.1080/08964289.2016.1150805] [Medline: 26963636]
24. Smith AL, Carter SM, Chapman S, Dunlop SM, Freeman B. Why do smokers try to quit without medication or counselling? A qualitative study with ex-smokers. BMJ Open 2015;5(4):e007301 [FREE Full text] [doi: 10.1136/bmjopen-2014-007301] [Medline: 25933811]
25. Grundy Q. A review of the quality and impact of mobile health apps. Annu Rev Public Health 2022;43:117-134 [FREE Full text] [doi: 10.1146/annurev-publhealth-052020-103738] [Medline: 34910582]
26. Carreiro S, Newcomb M, Leach R, Ostrowski S, Boudreaux ED, Amante D. Current reporting of usability and impact of mHealth interventions for substance use disorder: a systematic review. Drug Alcohol Depend 2020;215:108201 [FREE Full text] [doi: 10.1016/j.drugalcdep.2020.108201] [Medline: 32777691]
27. Monney G, Penzenstadler L, Dupraz O, Etter JF, Khazaal Y. mHealth app for cannabis users: satisfaction and perceived usefulness. Front Psychiatry 2015;6:120 [FREE Full text] [doi: 10.3389/fpsyt.2015.00120] [Medline: 26379561]
28. Nierkens V, Hartman MA, Nicolaou M, Vissenberg C, Beune EJAJ, Hosper K, et al. Effectiveness of cultural adaptations of interventions aimed at smoking cessation, diet, and/or physical activity in ethnic minorities. a systematic review. PLoS One 2013;8(10):e73373 [FREE Full text] [doi: 10.1371/journal.pone.0073373] [Medline: 24116000]
29. Herbst E, McCaslin SE, Hassanbeigi Daryani S, Laird KT, Hopkins LB, Pennington D, et al. A qualitative examination of stay quit coach, a mobile application for veteran smokers with posttraumatic stress disorder. Nicotine Tob Res 2020;22(4):560-569. [doi: 10.1093/ntr/ntz037] [Medline: 30874289]
30. Hébert ET, Ra CK, Alexander AC, Helt A, Moisiuc R, Kendzor DE, et al. A mobile just-in-time adaptive intervention for smoking cessation: pilot randomized controlled trial. J Med Internet Res 2020;22(3):e16907 [FREE Full text] [doi: 10.2196/16907] [Medline: 32149716]
31. James DC, Harville C, Sears C, Efunbumi O, Bondoc I. Participation of African Americans in e-Health and m-Health studies: a systematic review. Telemed J E Health 2017;23(5):351-364. [doi: 10.1089/tmj.2016.0067] [Medline: 27792475]
32. Enyioha C, Loufman LM, Grewe ME, Cené CW, Khairat S, Goldstein AO, et al. Black smokers' preferences for features of a smoking cessation app: qualitative study. JMIR Form Res 2023;7:e43603 [FREE Full text] [doi: 10.2196/43603] [Medline: 37252777]
33. Ray R, Sewell AA, Gilbert KL, Roberts JD. Missed opportunity? Leveraging mobile technology to reduce racial health disparities. J Health Polit Policy Law 2017;42(5):901-924. [doi: 10.1215/03616878-3940477] [Medline: 28663182]
34. Hawkins RP, Kreuter M, Resnicow K, Fishbein M, Dijkstra A. Understanding tailoring in communicating about health. Health Educ Res 2008;23(3):454-466 [FREE Full text] [doi: 10.1093/her/cyn004] [Medline: 18349033]
35. Institute of Medicine. Speaking of Health: Assessing Health Communication Strategies for Diverse Populations. Washington, DC: The National Academies Press; 2002.
36. Businelle MS, Ma P, Kendzor DE, Frank SG, Vidrine DJ, Wetter DW. An ecological momentary intervention for smoking cessation: evaluation of feasibility and effectiveness. J Med Internet Res 2016;18(12):e321 [FREE Full text] [doi: 10.2196/jmir.6058] [Medline: 27956375]
37. Barrera M, Castro FG. A heuristic framework for the cultural adaptation of interventions. Clinical Psychology: Science and Practice 2006;13(4):311-316. [doi: 10.1111/j.1468-2850.2006.00043.x]
38. Trinidad SB, Shane AL, Guinn TR, Apok CR, Collier AF, Avey JP, et al. The cultural adaption of a sobriety support app for Alaskan Native and American Indian: qualitative feasibility and acceptability study. JMIR Form Res 2023;7:e38894 [FREE Full text] [doi: 10.2196/38894] [Medline: 36473107]
39. Webb MS. Does one size fit all African American smokers? The moderating role of acculturation in culturally specific interventions. Psychol Addict Behav 2008;22(4):592-596. [doi: 10.1037/a0012968] [Medline: 19071987]
40. Wall SP, Castillo P, Shuchat-Shaw F, Norman E, Brown D, Martinez-López N, et al. Targeting versus tailoring educational videos for encouraging deceased organ donor registration in black-owned barbershops. J Health Commun 2022;27(1):37-48 [FREE Full text] [doi: 10.1080/10810730.2022.2035021] [Medline: 35170401]
41. Brewer LC, Kumbamu A, Smith C, Jenkins S, Jones C, Hayes SN, et al. A cardiovascular health and wellness mobile health intervention among church-going African Americans: formative evaluation of the FAITH! app. JMIR Form Res 2020;4(11):e21450 [FREE Full text] [doi: 10.2196/21450] [Medline: 33200999]
42. Staffileno BA, Tangney CC, Fogg L. Favorable outcomes using an eHealth approach to promote physical activity and nutrition among young African American women. J Cardiovasc Nurs 2018;33(1):62-71. [doi: 10.1097/JCN.0000000000000409] [Medline: 28481823]

XSL•FO
RenderX

43. Matthews AK, Sánchez-Johnsen L, King A. Development of a culturally targeted smoking cessation intervention for African American smokers. J Community Health 2009;34(6):480-492 [FREE Full text] [doi: 10.1007/s10900-009-9181-5] [Medline: 19728056]
44. Resnicow K, Baranowski T, Ahluwalia JS, Braithwaite RL. Cultural sensitivity in public health: defined and demystified. Ethn Dis 1999;9(1):10-21. [Medline: 10355471]
45. Walters ST, Mun EY, Tan Z, Luningham JM, Hébert ET, Oliver JA, et al. Development and preliminary effectiveness of a smartphone-based, just-in-time adaptive intervention for adults with alcohol misuse who are experiencing homelessness. Alcohol Clin Exp Res 2022;46(9):1732-1741 [FREE Full text] [doi: 10.1111/acer.14908] [Medline: 35869820]
46. Businelle MS, Garey L, Gallagher MW, Hébert ET, Vujanovic A, Alexander A, et al. An integrated mHealth app for smoking cessation in black smokers with anxiety: protocol for a randomized controlled trial. JMIR Res Protoc 2022;11(5):e38905 [FREE Full text] [doi: 10.2196/38905] [Medline: 35635746]
47. Garey L, Hébert ET, Mayorga NA, Chavez J, Shepherd JM, Businelle MS, et al. Evaluating the feasibility and acceptability of a mobile-based health technology for smoking cessation: mobile anxiety sensitivity program. Br J Clin Psychol 2022;61 Suppl 1(Suppl 1):111-129 [FREE Full text] [doi: 10.1111/bjc.12294] [Medline: 33939190]
48. Casey M, Hayes PS, Glynn F, OLaighin G, Heaney D, Murphy AW, et al. Patients' experiences of using a smartphone application to increase physical activity: the SMART MOVE qualitative study in primary care. Br J Gen Pract 2014;64(625):e500-e508 [FREE Full text] [doi: 10.3399/bjgp14X680989] [Medline: 25071063]
49. Patton MQ. Utilization-Focused Evaluation. Thousand Oaks: Sage Publications, Inc; 2008.
50. Saldaña J. The Coding Manual for Qualitative Researchers. Thousand Oaks: Sage Publications, Inc; 2013.
51. Patton MQ. Qualitative Research and Evaluation Methods: Integrating Theory and Practice. Thousand Oaks: Sage Publications, Inc; 2014.
52. Miles MB, Huberman AM, Saldaña J. Qualitative Data Analysis. Thousand Oaks: SAGE Publications, Inc; 2019.
53. Snowden LR, Hines AM. A scale to assess African American acculturation. Journal of Black Psychology 1999;25(1):36-47. [doi: 10.1177/0095798499025001003]
54. Puddephatt JA, Leightley D, Palmer L, Jones N, Mahmoodi T, Drummond C, et al. A qualitative evaluation of the acceptability of a tailored smartphone alcohol intervention for a military population: information about drinking for ex-serving personnel (InDEx) app. JMIR Mhealth Uhealth 2019;7(5):e12267 [FREE Full text] [doi: 10.2196/12267] [Medline: 31127726]
55. Bendotti H, Lawler S, Ireland D, Gartner C, Hides L, Marshall HM. What do people want in a smoking cessation app? An analysis of user reviews and app quality. Nicotine Tob Res 2022;24(2):169-177. [doi: 10.1093/ntr/ntab174] [Medline: 34460922]
56. Brewer LC, Fortuna KL, Jones C, Walker R, Hayes SN, Patten CA, et al. Back to the future: achieving health equity through health informatics and digital health. JMIR Mhealth Uhealth 2020;8(1):e14512 [FREE Full text] [doi: 10.2196/14512] [Medline: 31934874]
57. Jackson DN, Sehgal N, Baur C. Benefits of mHealth co-design for African American and hispanic adults: multi-method participatory research for a health information app. JMIR Form Res 2022;6(3):e26764 [FREE Full text] [doi: 10.2196/26764] [Medline: 35262496]
58. Nagawa CS, Faro JM, Menon AJ, Ito Fukunaga M, Williams JH, Mourao D, et al. Written advice given by African American smokers to their peers: qualitative study of motivational messages. JMIR Form Res 2021;5(4):e21481 [FREE Full text] [doi: 10.2196/21481] [Medline: 33929332]
59. Sideropoulos V, Vangeli E, Naughton F, Cox S, Frings D, Notley C, et al. Mobile phone text messages to support people to stop smoking by switching to vaping: codevelopment, coproduction, and initial testing study. JMIR Form Res 2023;7:e49668 [FREE Full text] [doi: 10.2196/49668] [Medline: 37756034]
60. Struik L, Christianson K, Khan S, Sharma RH. Strengths and limitations of web-based cessation support for individuals who smoke, dual use, or vape: qualitative interview study. JMIR Form Res 2023;7:e43096 [FREE Full text] [doi: 10.2196/43096] [Medline: 38064266]
61. Barroso-Hurtado M, Suárez-Castro D, Martínez-Vispo C, Becoña E, López-Durán A. Smoking cessation apps: a systematic review of format, outcomes, and features. Int J Environ Res Public Health 2021;18(21):11664 [FREE Full text] [doi: 10.3390/ijerph182111664] [Medline: 34770178]
62. McClure JB, Heffner J, Hohl S, Klasnja P, Catz SL. Design considerations for mHealth programs targeting smokers not yet ready to quit: results of a sequential mixed-methods study. JMIR Mhealth Uhealth 2017;5(3):e31 [FREE Full text] [doi: 10.2196/mhealth.6845] [Medline: 28283465]
63. Attwood S, Parke H, Larsen J, Morton KL. Using a mobile health application to reduce alcohol consumption: a mixed-methods evaluation of the drinkaware track & calculate units application. BMC Public Health 2017;17(1):394 [FREE Full text] [doi: 10.1186/s12889-017-4358-9] [Medline: 28511698]

Abbreviations

MASP: Mobile Anxiety Sensitivity Program
mHealth: mobile health

XSL•FO
RenderX

NRT: Nicotine Replacement Therapy

Edited by A Mavragani; submitted 17.10.23; peer-reviewed by H Tateno; comments to author 23.03.24; accepted 24.04.24; published 07.02.25.

Please cite as:
Cheney MK, Alexander AC, Garey L, Gallagher MW, Hébert ET, Vujanovic AA, Kezbers KM, Matoska CT, Zvolensky MJ, Businelle MS
Adapting a Mobile Health App for Smoking Cessation in Black Adults With Anxiety Through an Analysis of the Mobile Anxiety Sensitivity Program Proof-of-Concept Trial: Qualitative Study
JMIR Form Res 2025;9:e53566
URL: https://formative.jmir.org/2025/1/e53566
doi:10.2196/53566
PMID:39918847

©Marshall K Cheney, Adam C Alexander, Lorra Garey, Matthew W Gallagher, Emily T Hébert, Anka A Vujanovic, Krista M Kezbers, Cameron T Matoska, Michael J Zvolensky, Michael S Businelle. Originally published in JMIR Formative Research (https://formative.jmir.org), 07.02.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Improving the Utility, Safety, and Ethical Use of a Passive Mood-Tracking App for People With Bipolar Disorder Using Coproduction: Qualitative Focus Group Study

Laurence Astill Wright[1,2], MBBCH; Matthew Moore[1,3]; Stuart Reeves[4], PhD; Elvira Perez Vallejos[1,3], PhD; Richard Morriss[1,3,5,6], PhD

[1]Institute of Mental Health, University of Nottingham, Nottingham, United Kingdom
[2]Centre for Academic Mental Health, Population Health Sciences, University of Bristol, Bristol, United Kingdom
[3]NIHR MindTech Medical Technology Collaborative, University of Nottingham, Nottingham, United Kingdom
[4]School of Computer Science, University of Nottingham, Nottingham, United Kingdom
[5]NIHR ARC East Midlands, University of Nottingham, Nottingham, United Kingdom
[6]Nottingham NIHR Biomedical Research Centre, University of Nottingham, Nottingham, United Kingdom

Corresponding Author:
Laurence Astill Wright, MBBCH
Institute of Mental Health
University of Nottingham
Jubilee Campus
Triumph Road
Nottingham, NG7 2TU
United Kingdom
Phone: 44 115 823 1294
Email: laurence.astillwright@nottingham.ac.uk

Abstract

Background: Coproduction with users of new digital technology, such as passive mood monitoring, is likely to improve its utility, safety, and successful implementation via improved design and consideration of how such technology fits with their daily lives. Mood-monitoring interventions are commonly used by people with bipolar disorder (BD) and have promising potential for digitization using novel technological methods.

Objective: This study aims to explore how a passive behavioral monitoring platform, Remote Assessment of Disease and Relapse, would meet the needs of people with BD by specifically considering purpose and function, diversity of need, personal preference, essential components and potential risks, and harms and mitigation strategies through an iterative coproduction process.

Methods: A total of 17 people with BD were recruited via national charities. We conducted 3 web-based focus groups as a part of an iterative coproduction process in line with responsible research and innovation principles and with consideration of clinical challenges associated with BD. Data were analyzed thematically. Results were cross-checked by someone with lived experience of BD.

Results: Focus groups were transcribed and analyzed using thematic analysis. Six themes were identified as follows: (1) the purpose of using the app, (2) desired features, (3) when to use the app, (4) risks of using the app, (5) sharing with family and friends, and (6) sharing with health care professionals.

Conclusions: People with BD who are interested in using passive technology to monitor their mood wish to do so for a wide variety of purposes, identifying several preferences and potential risks. Principally, people with BD wished to use this novel technology to aid them in self-managing their BD with greater insight and a better understanding of potential triggers. We discuss key features that may aid this functionality and purpose, including crisis plans and sharing with others. Future development of passive mood-monitoring technologies should not assume that the involvement of formal mental health services is desired.

(JMIR Form Res 2025;9:e65140) doi:10.2196/65140

XSL•FO
RenderX

KEYWORDS

mood monitoring; ecological momentary assessment; EMA; passive ecological momentary assessment; passive EMA; bipolar disorder; implementation; qualitative; mobile phone

Introduction

Background

Recent developments in technology have increased the possibility of determining the mental state of an individual via tracking behavioral data collected via smartphones [1]. This promises novel interventions based on passive ecological momentary assessment—a technology that can potentially frequently assess mood via data collected in the background of one's smartphone use [2]. While this hypothesis remains controversial, many research teams are trying to implement this new technology into usable and acceptable interventions [3]. One open-access research platform using passive monitoring technology to glean new insights into mental disorders is the Remote Assessment of Disease and Relapse (RADAR) [4]. Here, we explore the design and potential implementation of this technology, using the RADAR platform, in the form of a mood-tracking app for people with bipolar disorder (BD), considering purpose, function, components, and risks and harms in an iterative coproduction process centrally involving end users of the technology—people with BD interested in mood monitoring. However, clinical experts and experts in the design of digital technology involving human-computer interaction and responsible innovation were also involved in such coproduction, as we have previously proposed [5]. There is a tension between 2 models of technology explored here—one where mental state is predicted via smartphone use and another where technology enters our lives via coproduction. This paper builds on these conceptual ideas.

BD is a condition where individuals already apply various mood-tracking methods to try to gain insights into the disorder and prevent relapse [5,6]. There is some evidence suggesting that increasing awareness of mood fluctuations can improve insights into and the identification of early warning signs and can be useful in preventing relapse in BD [7]. It also suggests that novel technology increases the potential to harness this new technology to create more sophisticated and structured mood-tracking interventions for people with BD. Most of these existing techniques and methods of tracking mood require considerable effort from participants in logging their mood every day for many months [8,9] and may be subject to inaccurate information due to mood biases [10]. Many people with BD use approaches such as paper- and pen-based tracking procedures, mental and physical notation, conversations with their partner, or already existing active smartphone apps [5]. Passive data collection approaches (which are then interpreted actively by the user) may improve flexibility, usability, and acceptability and decrease the time requirement for people who monitor their moods. In a previous qualitative work, a passive mood-tracking app was perceived as potentially positive and helpful by people with BD [5], solving many of the issues with existing methods.

When applying novel technologies in potentially clinical and vulnerable populations, it is important to consider potential risks and harms [11,12]. Technical advances offer potential solutions to complex problems with lighter regulation than many existing approaches to mental health, for example, medication. There is an urgent need to embed responsible research and innovation (RRI) frameworks to maximize societal benefit and be able to anticipate and reflect on the medium- or long-term impact of technology on stakeholders as well as the environment [13]. For example, by engaging with users, developers can identify design properties that could minimize the potential for harm in a number of ways (eg, accessible interfaces, integration with conventional mental health services, and enhanced data privacy and security [14]) and also reduce other risks, for example, digital exclusion that might potentially increase inequalities [14-16].

RRI [11] provide an ethical approach that requires planning, anticipation, and co-design through technological development with end users and all those involved in developing and later using the technology. Along with mitigating harm and inequality, such a co-design process carried out in the framework helps to improve its acceptability and usability for all potential users and relevance to practice and provides some assurance of the safety of the product. For these reasons, a co-design process may lead to a faster spread of usable interventions and a quicker abandonment of less satisfactory developments. However, one element that has so far been given less consideration in the work of responsible innovation is the range of problems associated with different types of mental states when using different forms of technology, such as digital mood monitoring, particularly in conditions where change may be dramatically different, such as BD, where people with BD may experience more than one type of mental state, such as mania and depression. For instance, people with mania are likely to show dramatic changes in their social contacts and what they will communicate and share (particularly on social media) compared to when they are depressed or when they are in neither mental state. This paper applies principles from RRI to consider risk anticipation and risk mitigation in people with BD.

RADAR is an open-source platform and resource for self-tracking psychiatric and neurological conditions developed by a consortium of academic and industry organizations. Currently, the platform has been successfully developed for those with multiple sclerosis [17], depression [4], and epilepsy [18]. RADAR currently passively monitors the following smartphone data: phone use, local weather, step count, GPS location, app use and time duration, battery level, Bluetooth devices in the nearby vicinity, actigraphy, heart rate, call or SMS text message logs, and ambient light.

We are aiming to apply the technology to people with BD, to assess how it might be adapted and adjusted to meet the specific needs of this population. We will not only focus on the technology but also on the requirements needed to ensure it

XSL•FO
RenderX

fulfills the aspirations of people with BD, balanced against identifying potential risks. We will use expertise in responsible technology design to consider unintended consequences, potential risks, and how these might be mitigated.

Objectives

We conducted a series of focus groups to explore the opportunities and challenges of the app to ensure the acceptability and efficacy of passive mood-monitoring apps in the future. This work was part of an iterative coproduction process in line with the use of RRI. We aimed to determine opinions on the following: (1) the purpose and function of the app, (2) the essential and desired components of the digital tool and optimal methods of delivery and design frequency, and (3) the potential risks and harms of using the technology and how these might be mitigated

Methods

Overview

Participants were recruited via national and local charities (Bipolar UK and Bipolar LIFT CIC) via their research opportunity network and other snowballing techniques via an existing patient and public involvement (PPI) team based at the Institute of Mental Health and the University of Nottingham.

Previous research included 16 individuals with BD [5,19] in 2 workshops; however, this explored the utility of self-tracking in general before the development of the RADAR-based tool and did not explore the tool in detail. We chose to explore the study aims through 3 focus groups with potential end users of this technology to simulate discussion and exploration of these issues rather than conduct individual interviews. Previously, we recruited participants from a geographical area that was familiar with participation in research on BD and digital mental health. We wanted to check whether the findings would apply to participants who were not from an area where there was such exposure to research so that fresh ideas and challenges might emerge and also to check the transferability of previous findings. We wanted to test the credibility and confirmability by feeding back findings from previous workshops and focus groups. To facilitate this, one of the workshop leaders from our previous research (an individual with over a decade of experience in PPI and who also has experienced BD) took part in the development of the topic guides and write-up of the current project. All these approaches enhanced the trustworthiness of data collection and analysis. We hoped to recruit 5 to 8 individuals per focus group (>8 individuals would be inappropriate via video call). We planned to run as many focus groups as required until there was a saturation of themes emerging from the focus groups, and no new themes were emerging. Initially, we planned 3 focus groups based on our previous experience of carrying out workshops and focus groups.

The inclusion criteria were as follows: self-reported formal diagnosis of bipolar 1 disorder, bipolar 2 disorder, or bipolar spectrum disorder (BD not otherwise specified or cyclothymia) by a UK psychiatrist, owns and regularly uses a smartphone with an interest in self-tracking their mood via a digital tool, fluent in English, able to give informed consent, and aged >16 years. Exclusion criteria were as follows: participants who lacked the capacity to consent (eg, self-reported diagnosis of dementia and primary diagnosis of substance misuse) or were current inpatients in a psychiatric hospital.

The study was conducted between August and December 2023. We tried to ensure that our methods reflected current practices used by these organizations who organize workshops, focus groups, and group events, which was the reason behind our choice of data collection. Our 3 focus groups were held on the web to improve accessibility and to improve diversity recruitment. The COREQ (Consolidated Criteria for Reporting Qualitative Studies) checklist is presented in Multimedia Appendix 1.

Ethical Considerations

The study was subject to ethical review and approved by the faculty of medicine and health sciences at the University of Nottingham (FMHS 174-1222). Participants made first contact with the research team following advertising. Participants were initially provided with a participant information sheet and given an opportunity to discuss this and ask questions via phone call or email. After ≥48 hours, participants were invited to complete the consent form and again provided with an opportunity to ask questions via telephone call or an alternative method. The data in this study have been deidentified where applicable, and participants consented to the publishing of deidentified quotations. Participants were provided with a £30 (US $38) voucher following the completion of a single focus group. The participants had no relationship with the research team before joining the study. The participants knew about the reasons for conducting the research but had limited knowledge about the research team other than their qualifications. While the research team was diverse, including people with lived experience, some members were psychiatrists and therefore might have carried certain assumptions or biases about the delivery of interventions in people with BD in the National Health Service.

Focus Group Layout

Each focus group lasted 2 hours in total and followed a semistructured topic guide informed by previous research [5] and produced by 2 psychiatrists and 1 person with lived experience of BD (LAW, RM, and MM). Participants were given an explanation of the purpose of the focus groups before participation. The topic guide was used to ask open questions on (1) developing the function of the mood-tracking tool, (2) developing the content of the mood-tracking tool, and (3) how best to translate function and content into a usable and acceptable digital tool (Figure 1). The first 2 focus groups concentrated mostly on what participants wanted from the app and why, reaching saturation in relation to these themes. Therefore, the third focus group only briefly commented on these matters having been presented with these findings. The third focus group concentrated in more detail on how the app was delivered, although some of this content had been explored in the first 2 focus groups. Focus groups were video recorded followed by verbatim transcription. Focus groups were facilitated by LAW (male psychiatrist), who was trained in qualitative interviewing and focus group data collection. At the

XSL•FO
RenderX

time of the focus group, all participants knew LAW from their first contact when considering joining the study.

Figure 1. Overview of the main findings of the 6 themes.

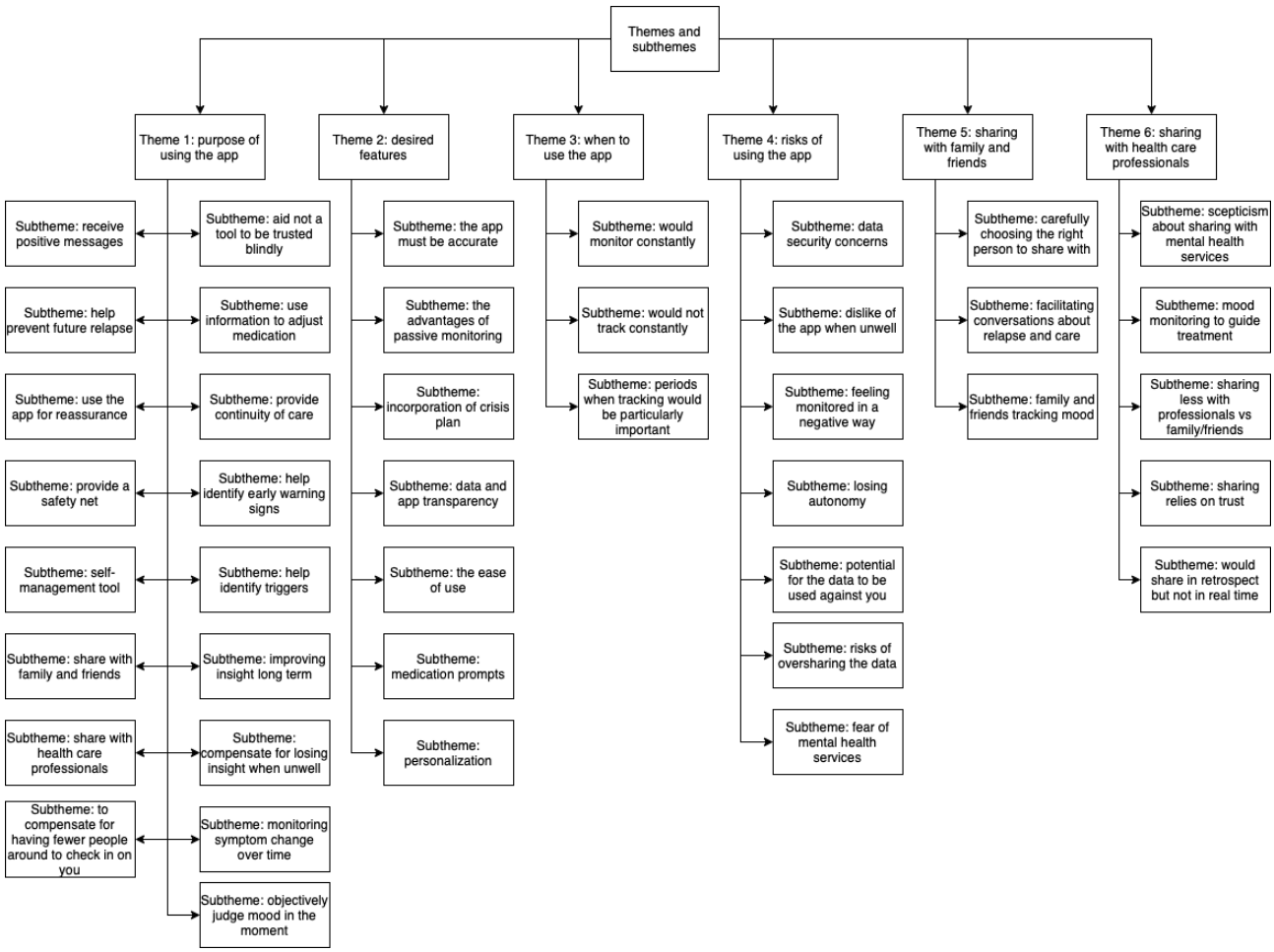

Participants in all focus groups were given a verbal description of the RADAR platform, including the type and amount of data it was possible to collect presented pictorially (Figure S1 in Multimedia Appendix 2). Participants in focus group 1 were presented with the results from the previous round of workshops. The results are summarized in Figures S2 and S3 in Multimedia Appendix 2.

Results

Overview

A total of 17 participants joined 3 focus groups. The demographic characteristics of participants are reported in Table 1. A total of 97 people initially made contact via recruitment methods.

XSL•FO
RenderX

Table 1. Demographics and clinical characteristics of the participants (N=17).

Demographic and clinical characteristics	Participants, n (%)
Age group (y)	
<18	0 (0)
18-24	1 (6)
25-34	4 (24)
35-44	3 (18)
45-55	3 (18)
55-64	4 (24)
>65	0 (0)
Sex	
Male	5 (29)
Female	9 (53)
Nonbinary	1 (6)
Category of bipolar diagnosis	
Bipolar 1 disorder	5 (29)
Bipolar 2 disorder	8 (47)
Cyclothymia	1 (6)
Bipolar disorder not specified	1 (6)
Years of lived experience of bipolar disorder	
<1	0 (0)
1-3	2 (12)
3-5	5 (29)
5-7	1 (6)
>7	7 (41)

Thematic analysis resulted in 6 themes as follows: (1) purpose of using the app, (2) desired features, (3) when to use the app, (4) risks of using the app, (5) sharing with family and friends, and (6) sharing with health care professionals. For each theme, participants described positives and negatives. Tables 2-4 report themes, subthemes, and the participant identifiers.

All the results were cross-checked by someone with lived experience of BD type 1 and over a decade of experience in PPI. Participants did not provide feedback on the findings. Coding via NVivo (Lumivero) [20] was performed by 1 researcher (LAW).

XSL•FO
RenderX

Table 2. Quotes related to theme 1 (purpose of using the app).

Subthemes	Quotes
Subtheme: aid not a tool to be trusted blindly	• "But then I think most people would rather have someone check in and say ohh you know, your app suggests there's been a change and then for them to say Oh no actually you know, I'm fine it's just that this has happened that's happened." [ID15].
Subtheme: understanding triggers	• "What I have used it for is we'll got on. I'm I'm a bit anxious for no particular reason. Am I just being bloody minded and annoyed? But a lot of the time, it's actually. I've tripled my workload. I haven't had sleep." [ID13]
Subtheme: improving insight	• "So actually, I didn't sleep well and this happened or I did sleep well and I was fine and ohh I've been spending a lot of money doing a lot of stuff and then this happened. But also what was happening with my medication and where was I in my cycle? Where was I? What exercise was my level of exercise good at high or low or yeah, so I think it's for your own picture and your own health and everybody's different." [ID11]
Subtheme: objectively judging the mood	• "I think one of the positives about it for people with bipolar is that objectivity and you know the objective data that comes because it also takes a bit of that pressure off the person." [ID12]
Subtheme: prevention	• "Yes, that is fair to say and to catch it quickly, you know, I mean sort of the sooner you can pick up on it and if you've got a monitoring device that can alert you, that's even better." [ID8]. • "Bit hyper, you're a bit, you know, you're talking really fast or you're having loads of different ideas about different things and stuff, and it might be that I missed that until it's fed back to me, reflected back to me so." [ID9]
Subtheme: safety net and reassurance	• "So something in the background. Something in the background sort of saying actually, you're doing X, Y & Z and that's not normal for you." [ID5] • "It's a sense of security in a way. You know, it's almost like having someone there constantly keeping a check on your mood, even though they're not there virtually." [ID8]
Subtheme: self-management	• "If I'm struggling, I'm at 2. What activities could I do to boost it so I'll go for a walk or art activities so I will go back to that book that we made and try the strategies first before I went to the doctor and got a PRN so that would be how I'd monitor it." [ID2]
Subtheme: to compensate for having fewer people around to check in on you	• "I live alone so I don't have anybody to notice. A mood tracker would help with that." [ID5]
Subtheme: use information to adjust medication	• "Yeah, lifestyle changes. Probably tweaking my medication. It could be that because sleep deprivation is a big trigger for me and it could be that I'm lacking in sleep, so the may need to give me some sleep medication." [ID8]

XSL•FO
RenderX

Table 3. Quotes related to theme 2 (desired features), 3 (when to use the app), and 4 (risks of using the app).

Themes and subthemes	Quotes
Theme 2: desired features	
Subtheme: accuracy is important	"Depends how good the app is, whether it does actually perform well 'cause it could be something that is obviously not correct. In a sense, it might not necessarily me being a period of mania, it could be picking up something that is questionable." [ID9]
Subtheme: the advantages of passive monitoring	"And quite often I'm not even that. I'm not that aware of. Maybe what I've put into it 'cause I get up, do it, and then you know and go about my business. So maybe something that's a little bit more engaging and if it's an app that's working in the background more rather than you know, and then I can perhaps turn to that and realize stuff that I've not seen before." [ID9]
Subtheme: crisis plan	"We have a plan in place like that. He's allowed to call the doctors and the mental health team and the GP that needs to be called." [ID17]
Subtheme: data and app transparency	"I think I would like to see all the individual pieces of data. I feel like if that app is collecting those data points, I kind of feel like I'm entitled to see them." [ID10]
Subtheme: the ease of use	"If it's really like clunky and not that accessible or it doesn't feel very friendly. It's gotta be something that fits, doesn't it? The individual." [ID9]
Subtheme: medication prompts	"When I'm not well then I'm absolutely rubbish at taking medication. So yes, prompts to take medication and stuff would probably be for me a good idea." [ID8]
Subtheme: personalization	"I think that people are gonna have very different preferences and people are going to want to use this in very different ways. And I think that one size isn't gonna fit all. And if you're gonna do this, then you need to build in that adaptability within it." [ID15]
Theme 3: when to use the app	
Subtheme: would monitor constantly	"I think for me it would be useful to have it kind of on in the background all the time, so that then you can kind of see those periods where you're not 100% sure what it was like. You can go back and say oh actually I've had this amount of time...especially when the mood's low you can I find that I might think, oh, I've only been OK for a week, but it would give you a bit more of a realistic insight to say oh actually you've had a few months when you've been okay." [ID7]
Subtheme: not track all the time	"As for using it every day, I probably wouldn't. I would probably just dip in if I felt my mood starting to get a bit lower, a bit high." [ID14]
Subtheme: periods when tracking is particularly important	"I guess the idea would be that the app would learn about your behavior, so it would be able to predict when you know things might be difficult for you, or you know if certain times of the year are more difficult than being able to provide that support for you and then being able to toggle things on and off." [ID10]
Theme 4: risks of using the app	
Subtheme: data security concerns	"I think for me anyway, because you've got a lot of data, I'd be a little bit worried about where that data was going if it was being shared." [ID7]
Subtheme: dislike of the app when unwell	"I also got really paranoid when I was ill...so you, you know, would you get to the point where you didn't trust the app where you thought that people were trying to get you, you know, sectioned or trying to get your money off you or, you know, trying to do some sort of nefarious activities, you know, against you at that time?" [ID 14]
Subtheme: feeling monitored in a negative way	"Obviously there's a danger where you feel that you're being monitored." [ID13]
Subtheme: losing autonomy	"If the app sensed you know that things weren't going too well, you know, in whatever kind of data point that is, then it kinda feels like my, I dunno, sense of agency is removed, or autonomy...it doesn't feel good, it doesn't necessarily feel like an empowering thing." [ID10]
Subtheme: potential for the data to be used against you	"But it depends in certain situations, like if you got an abusive partner or something like that, you might not want them to know so." [ID2]
Subtheme: the risks of oversharing data	"Massive risk is taken...Overshare...there's a real risk the if you've got this sort of information you could just say to anybody any stranger: look at all this. The inhibitors aren't there. You lose the self-control, the inner voice, when you're high." [ID13]
Subtheme: the fear of mental health services	"I'd be scared of what they'd make of that information and whether they'd make decisions behind my back about my care. That yeah, that I wouldn't agree with." [ID4]

XSL•FO
RenderX

Table 4. Quotes related to theme 5 (sharing with family and friends) and theme 6 (sharing with health care professionals).

Themes and subthemes	Quotes
Theme 5: sharing with family and friends	
Subtheme: carefully choosing the right person to share with	"So I think the ability to share with other people is is good, but as a everyone said, you need to be really careful and have the options of who you share it with." [ID14]
Subtheme: facilitating conversations about relapse and care	"Yeah, I think it could create some difficult conversations, but the friend would usually be able to tell pretty quickly, especially if it's mania they'd be able to tell just by giving me a quick phone call." [ID4]
Subtheme: family and friends tracking mood	"That's it really, with me. And of course my partner does pull me back, do you know what I mean, when he sees it running out of, do you know what I mean, running of control." [ID3]
Theme 6: sharing with health care professionals	
Subtheme: skepticism about sharing with mental health services	"They don't take at face value what I'm telling them and therefore they are making comments like 'Ohh sorry, we don't believe you' and all this sort of stuff, so I'm opening up saying I'm ill and they're saying but you're talking to me, you're presenting a very strong individual, go away." [ID13]
Subtheme: mood monitoring to guide treatment	"I will say it amazed me actually as in I could see my highs and my lows and I'd be like you know where as if I hadn't been doing that, I wouldn't have been able to put that across to the health professional because it would kind of just gone you know, so this...was actually really beneficial to me." [ID16]
Subtheme: sharing less with professionals versus family and friends	"Do it myself first and share that sort of more directly with all my husband or whatever rather than sharing it sort of with the mental health team straight off." [ID7]
Subtheme: sharing relies on trust	"I think you'd have to be at one with who you're sharing with so you wouldn't just be sharing indiscriminately with health care professionals who that don't know you...my consultant psychiatrist, I've had, you know for 16, almost 16 years, so you know, it's lucky for me that I know her and true her and she knows me." [ID12]
Subtheme: would share in retrospect but not in real time	"Retrospectively, like it would work, I guess in the same way how I've been like using like the previous apps, I think in real time, I guess it could be like quite a few problems." [ID10]

Participants articulated a wide variety of different needs and preferences for the use of the app; some were commonly endorsed, and some were only endorsed by a few but for specific reasons, for example, because they lived alone and needed more objectivity (Tables 3 and 4). Most (35/45, 78%) subthemes were articulated, even if very briefly, in all 3 focus groups. Naturally, some subthemes were discussed in much more depth in certain focus groups. This is likely to represent important and generalizable data [21]. A minority (13/45, 29%) of subthemes were present in 1 or 2 focus groups (eg, medication prompts, providing continuity of care, positive messages, reassurance, data security concerns, feeling monitored [negative], potential for abuse, the risks of oversharing data, facilitating conversations about relapse and care, skepticism about sharing with mental health services, monitoring constantly, self-rating of mood can be unreliable, and stop self-monitoring when unwell).

Theme 1: Purpose of Using the App

Views on the purpose of the app were mixed (Table 3), and the participants wished to use the app in a variety of different ways. Some participants reported multiple uses for the app; others wished to use it for one specific purpose. The most popular was using the information they inputted into and reflecting back from the app to get a sense of where their mood was at that moment and to self-manage their BD. This could take the form of scaling back commitments, communicating with one's support network, changing medication, ensuring sufficient sleep, and doing more exercise, depending on the mood state. Many people wished to share information about their mood with trusted individuals around them and some with their health care professional.

Many wished to use the app to prevent future relapse by identifying early warning signs and triggers and improving long-term insight by examining both where their mood was now and the long-term changes in mood over many months and years. Many wanted a safety net for when they lost insights into how low or high their mood was at that moment, to both prevent the worsening of their mood and to avoid the negative behavioral consequences of being unwell, for example, excessive spending leading to the accumulation of long-term debt. There was an acknowledgment that the app might occasionally be wrong and that this information would be used to inform decision-making but not be the sole driver of it.

Theme 2: Desired Features

Participants were complimentary of passive monitoring (provided it was accurate and gave usable data), particularly regarding the ease of tracking their data over many years and when they may not wish to do so, for example, when they were unwell (Table 4). Many identified that a key barrier to them tracking their mood at present was the amount of time and effort it took, which they felt passive approaches would solve. Function and usability were key—the app being accurate in its core functions, for example, roughly identifying mood state. Participants wished to incorporate a crisis plan and personal alert system into the app; suggestions included writing things a person wished to be reminded of when they were well (such as a crisis plan) that could then be shown to them when they

XSL•FO
RenderX

were unwell, for example, around self-management of the illness. Another suggestion was receiving a notification from the app simply informing a person that they might be unwell as well as potentially sharing that with a predetermined trusted individual (explored in more detail in Theme 5: Sharing With Family and Friends section). The ability to customize and personalize aspects of the app seemed key, for example, toggling on-off certain aspects of data collection or any sharing. It was acknowledged that individuals would wish to use the app in a variety of different ways with different preferences around data security and the way they manage their BD.

Theme 3: When to Use the App

Opinions regarding when participants might use the app were mixed (Table 4). Some wished to monitor it constantly, acknowledging that doing so may help them triangulate their mood around important life events in the hope of managing their illness. Others disliked the idea of tracking all the time and did not wish to do so when well. Some mentioned that there were times, for example, around particularly stressful life events, when tracking might be more important.

Theme 4: Risks of Using the App

Concerns were raised around data security and possible breaches of data (Table 4), including from the individual who was unwell, for example, oversharing their information or data with someone who could potentially harm them. Susceptibility to abuse because of the collected data were a concern and some participants felt as the stigma of their illness had been used against them previously. Some participants were concerned they might feel monitored and that they would develop a dislike of the app if they felt particularly paranoid when unwell, potentially deleting it or removing their data from it. In total, 2 participants expressed concerns that the app could decrease their sense of autonomy, particularly if it was directing a person to do certain things or sharing their information with others. Many expressed concerns and fear of the involvement of mental health services deriving from previous negative experiences, including not feeling heard, being detained under the Mental Health Act, and a lack of trust in their health care professional.

Theme 5: Sharing With Family and Friends

We explored sharing with others in detail as it appeared to be a particularly important and potentially difficult-to-implement aspect of the app (Table 4). Participants felt that carefully choosing the right person to share any data with was extremely important. Some participants felt that their family and friends would wish to see their mood charts over long periods, while others felt that that information would not be of interest. Some wished for a single chosen person or a couple of chosen people to receive a notification when they were unwell as a safety net—building on previous nuanced conversations around safety or crisis planning, relapse, self-management, the involvement of mental health services, etc. Others would not choose this functionality and would not wish to share their data.

Theme 6: Sharing With Health Care Professionals

Many individuals did not wish to share their mood data with their health care professional in real time, due to concerns about an overreaction from mental health services, coercion, and a preference for self-management (Table 4). This seemed to be determined by the degree of trust participants had in their health care professional. However, most would share the data about their mood in retrospect to help with diagnosis, assessment, and subsequent treatment, for example, medication monitoring.

Discussion

Principal Findings

This qualitative study explored what people with BD wanted from a passive, mood-tracking app and tool to help them manage their BD. We explored the purpose, desired features, and usability in the context of ethical development of technology considering potential risks and safety of use and misuse. We explored the potential methods of incorporating desired features into the app in ways that would not compromise safety or cause unintended consequences, for example, sharing data with others.

Fundamentally, people with BD wanted to use passive mood-monitoring technology for a wide variety of purposes using different methods. They wished to use the data in a diverse and pragmatic way to help them self-manage their BD with greater insight and a better understanding of potential triggers. We discuss key features that may aid this functionality, for example, sharing with trusted individuals and reminders of crisis plans or personal alert systems. We discuss the future development of passive mood-monitoring technologies and particularly state that it should not be assumed that the involvement of formal mental health services is desired.

Most individuals principally wished to use the app to aid the already existing self-management of their BD. The way individuals wished to self-manage varied, and a preference for personalization was clear to accommodate the variety of strategies of using the app and the data it generated. This accords with previous work where insight, understanding, and self-management [6] to aid coping and prevention of relapse were the most salient reasons for engaging with self-monitoring [5,22-24]. Some subthemes were only articulated by a minority of participants, for example, using the information from the app to adjust their medication, providing continuity of care, and receiving positive or encouraging messages from the app. These less frequently endorsed subthemes highlight the multiple, different, and individual ways that the data generated by the app could be used. This kind of technological intervention could present data to individuals who can then use these data flexibly to prompt a variety of more nuanced behavioral changes.

Sharing with health care professionals appeared to be a less important reason for developing the app; again, this chimes with previous findings [5]. Individuals in this study felt that sharing their data and facilitating conversations around their illness with someone and people they trust was important, reinforcing what previous findings had demonstrated [5]. Communication with trusted people to improve relationships and manage the condition together seems important for many people with BD, while for others who do not have a trusted individual to check on them, the app seemed to provide an alternative safety net.

Fundamentally the app needed to be easy to use and accurate when determining mood. This presents programming and

XSL•FO
RenderX

technical challenges around performance, validity, and human-computer interaction. The RADAR platform, as used in depression, multiple sclerosis, and epilepsy, uses passive monitoring to obtain a general detection of important changes in the underlying condition, which then triggers active data completion to provide more granularity and greater sensitivity and specificity [17,25,26]. Participants articulated that the app does not necessarily have to be accurate all the time (subtheme: aid not a tool to be trusted blindly) and that any data from the app would be triangulated with contextual information, such as their own personal judgment on their mood and the opinions of those around them. It does seem important to have an awareness of how accurate the app is likely to be and this will factor into any decision-making that is reliant on the app's data [27].

Participants in our study suggested feasible ways of incorporating desired features and mitigating the risks of this technology. Crisis plans and personal alert systems could be set up when an individual is well, and then they could be reminded of crisis plans and personal alert systems when they are unwell (similar methods have been used in diabetes care [28] and epilepsy [29]). The person with BD could nominate 1 individual to receive a notification if the app considered that the person with BD might be unwell—again, similar to recent developments with seizure or diabetes tracking apps [28,29]. These features can all be chosen actively and toggled on and off, including all individual data collection methods. This would incorporate some aspects of personalization that seem important to people with BD, both in this study and in others, to account for differing needs, patterns of illness, and self-management strategies [5,22,24,30]. Data security should be considered carefully, and we should assume that many individuals will be skeptical about sharing their data with mental health services.

Participants in our focus groups did not feel that the app would feel confronting and potentially exacerbate any negative thoughts [31] via increasing awareness of depressive symptoms; however, other studies have suggested that this may be the case with mood-monitoring apps [32,33]. Some participants were concerned about the risks of using the app and there was a concern about technology exacerbating the risks that people with BD already face, for example, vulnerability, abuse, a loss of autonomy and independence, and stigma.

Limitations

The limitations of the study are mentioned subsequently. The sample reported here is small and is unlikely to be representative of those with BD more broadly, partly because we specifically recruited individuals with an interest in smartphone-based interventions and mood tracking. However, it did show diversity, both in terms of demographics and subthemes and views articulated. The recruitment was based on the kind of people we expected would ultimately use the app therefore was a pragmatic sampling method to best inform the final intervention. As most of these individuals were interested in mood-monitoring interventions, the degree of engagement and detailed discussion was likely higher than if we had selected individuals who did not have a particular interest in mood-tracking or digital interventions. Another limitation of this work was that we did not systematically record ethnicity data or previous experience with mood-tracking platforms. We were limited to how much personal data we collected and reported so that the data, combined with direction quotations, would not identify participants who were assured of anonymity through ethical review and individual consent.

The app was discussed hypothetically with a demonstration or discussion of possible features or usability; therefore, participants did not use the app. It is possible that following use, opinions may change; however, mostly the discussion around potential features, risks, and usability was theoretical, hence we decided to include a hypothetical discussion. Further work could iteratively develop the app, incorporating this feedback and these views into its development. A series of rounds of qualitative work could be conducted to gradually develop a feasible, usable, and acceptable app for people with BD. Furthermore, these results are specific to the RADAR platform, but there are many commonalities between passive behavior-tracking apps. Therefore, it is reasonable to consider the applicability of other platforms. These focus groups highlight the importance of accuracy and performance of digital mood-tracking apps [32,34,35].

Using the principles of RRI, we have considered and reflected on the issues identified by stakeholders as well as mitigation practices in the iterative development of this app; however, this is largely hypothetical. As we develop the app further into a useable prototype and the clinical utility is increasingly explored, we will continue engaging with stakeholders to anticipate and act upon risk detection and risk surveillance to anticipate potential misuse or unintended use as well as prioritizing safety, inclusivity, and responsible innovation [11]. This is particularly important considering the potential risks or harms highlighted here and previous concerns around the potential adverse effects of mood-monitoring interventions [32,33].

Conclusions

This study suggests that people with BD, who are interested in using passive technology to monitor mood, wish to do so for a wide variety of purposes and using a variety of different methods of using the data. Principally, people with BD wished to use this novel technology to aid them in self-managing their BD with greater insight and a better understanding of potential triggers. There are a few key features that may aid this functionality, for example, sharing with trusted individuals and the reminder of crisis plans and personal alert systems. Future development of passive mood-monitoring technologies should not assume that the involvement of formal mental health services is desired.

XSL•FO
RenderX

Acknowledgments

This work was supported by a Wellcome Trust Fellowship to LAW. RM was the chief investigator on a grant from the UK National Institute for Health and Care Research (NIHR) Applied Research Collaboration East Midlands Mental Health and Well-being Theme.

Data Availability

The datasets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

LAW and RM were responsible for the original study design. LAW was responsible for data analysis and initial drafting of the manuscript. All authors were responsible for subsequent interpretation, editing, and rewriting of the manuscript.

Conflicts of Interest

RM has received funding from Novartis to serve on a data management and ethics committee for 2 trials on the treatment of depression.

Multimedia Appendix 1
COREQ (Consolidated Criteria for Reporting Qualitative Studies) checklist.
[DOCX File , 21 KB - formative_v9i1e65140_app1.docx]

Multimedia Appendix 2
Details of possible data collection and summary of previous works.
[DOCX File , 683 KB - formative_v9i1e65140_app2.docx]

References

1. Torous J, Bucci S, Bell IH, Kessing LV, Faurholt-Jepsen M, Whelan P, et al. The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. World Psychiatry 2021 Oct 09;20(3):318-335 [FREE Full text] [doi: 10.1002/wps.20883] [Medline: 34505369]
2. Torous J, Choudhury T, Barnett I, Keshavan M, Kane J. Smartphone relapse prediction in serious mental illness: a pathway towards personalized preventive care. World Psychiatry 2020 Oct 15;19(3):308-309 [FREE Full text] [doi: 10.1002/wps.20805] [Medline: 32931109]
3. Depp C, Torous J, Thompson W. Technology-based early warning systems for bipolar disorder: a conceptual framework. JMIR Ment Health 2016 Sep 07;3(3):e42 [FREE Full text] [doi: 10.2196/mental.5798] [Medline: 27604265]
4. Matcham F, Barattieri di San Pietro C, Bulgari V, de Girolamo G, Dobson R, Eriksson H, et al. Remote assessment of disease and relapse in major depressive disorder (RADAR-MDD): a multi-centre prospective cohort study protocol. BMC Psychiatry 2019 Feb 18;19(1):72 [FREE Full text] [doi: 10.1186/s12888-019-2049-z] [Medline: 30777041]
5. Majid S, Morriss R, Figueredo G, Reeves S. Exploring self-tracking practices for those with lived experience of bipolar disorder: learning from combined principles of patient and public involvement and HCI. In: Proceedings of the 2022 ACM Designing Interactive Systems Conference. 2022 Presented at: DIS '22; June 13-17, 2022; Virtual Event p. 1907-1920 URL: https://dl.acm.org/doi/10.1145/3532106.3533531 [doi: 10.1145/3532106.3533531]
6. Murnane E, Cosley D, Chang P, Guha S, Frank E, Gay G, et al. Self-monitoring practices, attitudes, and needs of individuals with bipolar disorder: implications for the design of technologies to manage mental health. J Am Med Inform Assoc 2016 May;23(3):477-484. [doi: 10.1093/jamia/ocv165] [Medline: 26911822]
7. Morriss R, Vinjamuri I, Faizal MA, Bolton CA, McCarthy JP. Training to recognise the early signs of recurrence in schizophrenia. Cochrane Database Syst Rev 2013 Feb 28(2):CD005147. [doi: 10.1002/14651858.CD005147.pub2] [Medline: 23450559]
8. Bauer M, Glenn T, Alda M, Grof P, Bauer R, Ebner-Priemer UW, et al. Longitudinal digital mood charting in bipolar disorder: experiences with chronorecord over 20 years. Pharmacopsychiatry 2023 Sep 07;56(5):182-187 [FREE Full text] [doi: 10.1055/a-2156-5667] [Medline: 37678394]
9. Post RM, Denicoff KD, Leverich GS, Altshuler LL, Frye MA, Suppes TM, et al. Morbidity in 258 bipolar outpatients followed for 1 year with daily prospective ratings on the NIMH life chart method. J Clin Psychiatry 2003 Jun 15;64(6):680-738. [doi: 10.4088/jcp.v64n0610] [Medline: 12823083]
10. Bos F, Schreuder M, Doornbos B, Snippe E, Bruggeman R, van der Krieke L, et al. Prospective early warning signals to detect transitions to manic and depressive episodes in bipolar disorder. Eur Psychiatry 2021 Aug 13;64(S1):S78-S79. [doi: 10.1192/j.eurpsy.2021.237]

XSL•FO
RenderX

11. Eyre HA, Ellsworth W, Fu E, Manji HK, Berk M. Responsible innovation in technology for mental health care. Lancet Psychiatry 2020 Sep;7(9):728-730 [FREE Full text] [doi: 10.1016/S2215-0366(20)30192-9] [Medline: 32828153]
12. Stilgoe J, Owen R, Macnaghten P. Developing a framework for responsible innovation. Res Policy 2013 Nov;42(9):1568-1580. [doi: 10.1016/j.respol.2013.05.008]
13. Jirotka M, Grimpe B, Stahl B, Eden G, Hartswood M. Responsible research and innovation in the digital age. Commun ACM 2017 Apr 24;60(5):62-68. [doi: 10.1145/3064940]
14. Lopez-Campos G, Gabarron E, Martin-Sanchez F, Merolli M, Petersen C, Denecke K. Digital interventions and their unexpected outcomes - time for digitalovigilance? Stud Health Technol Inform 2024 Jan 25;310:479-483. [doi: 10.3233/SHTI231011] [Medline: 38269849]
15. Middle R, Welch L. Experiences of digital exclusion and the impact on health in people living with severe mental illness. Front Digit Health 2022 Nov 22;4:1004547 [FREE Full text] [doi: 10.3389/fdgth.2022.1004547] [Medline: 36483535]
16. Greer B, Robotham D, Simblett S, Curtis H, Griffiths H, Wykes T. Digital exclusion among mental health service users: qualitative investigation. J Med Internet Res 2019 Jan 09;21(1):e11696 [FREE Full text] [doi: 10.2196/11696] [Medline: 30626564]
17. Sun S, Folarin AA, Zhang Y, Cummins N, Liu S, Stewart C, et al. The utility of wearable devices in assessing ambulatory impairments of people with multiple sclerosis in free-living conditions. Comput Methods Programs Biomed 2022 Dec;227:107204 [FREE Full text] [doi: 10.1016/j.cmpb.2022.107204] [Medline: 36371974]
18. Simblett S, Matcham F, Curtis H, Greer B, Polhemus A, Novák J, et al. Patients' measurement priorities for remote measurement technologies to aid chronic health conditions: qualitative analysis. JMIR Mhealth Uhealth 2020 Jun 10;8(6):e15086 [FREE Full text] [doi: 10.2196/15086] [Medline: 32519975]
19. Polhemus A, Novak J, Majid S, Simblett S, Morris D, Bruce S, et al. Data visualization for chronic neurological and mental health condition self-management: systematic review of user perspectives. JMIR Ment Health 2022 Apr 28;9(4):e25249 [FREE Full text] [doi: 10.2196/25249] [Medline: 35482368]
20. Castleberry A. NVivo 10 [software program]. Version 10. QSR international; 2012. Am J Pharm Educ 2014 Feb;78(1):25. [doi: 10.5688/ajpe78125]
21. Guba EG. Criteria for assessing the trustworthiness of naturalistic inquiries. Educ Technol Res Dev 1981 Jun;29(2):75-91 [FREE Full text] [doi: 10.1007/bf02766777]
22. Bos FM, Snippe E, Bruggeman R, Doornbos B, Wichers M, van der Krieke L. Recommendations for the use of long-term experience sampling in bipolar disorder care: a qualitative study of patient and clinician experiences. Int J Bipolar Disord 2020 Dec 01;8(1):38 [FREE Full text] [doi: 10.1186/s40345-020-00201-5] [Medline: 33258015]
23. Saunders KE, Bilderbeck AC, Panchal P, Atkinson LZ, Geddes JR, Goodwin GM. Experiences of remote mood and activity monitoring in bipolar disorder: a qualitative study. Eur Psychiatry 2017 Mar 27;41(1):115-121 [FREE Full text] [doi: 10.1016/j.eurpsy.2016.11.005] [Medline: 28135594]
24. Bos FM, Snippe E, Bruggeman R, Wichers M, van der Krieke L. Insights of patients and clinicians on the promise of the experience sampling method for psychiatric care. Psychiatr Serv 2019 Nov 01;70(11):983-991. [doi: 10.1176/appi.ps.201900050] [Medline: 31434558]
25. Matcham F, Leightley D, Siddi S, Lamers F, White KM, Annas P, et al. Remote Assessment of Disease and Relapse in Major Depressive Disorder (RADAR-MDD): recruitment, retention, and data availability in a longitudinal remote measurement study. BMC Psychiatry 2022 Feb 21;22(1):136 [FREE Full text] [doi: 10.1186/s12888-022-03753-1] [Medline: 35189842]
26. Simblett SK, Biondi A, Bruno E, Ballard D, Stoneman A, Lees S, et al. Patients' experience of wearing multimodal sensor devices intended to detect epileptic seizures: a qualitative analysis. Epilepsy Behav 2020 Jan;102:106717. [doi: 10.1016/j.yebeh.2019.106717] [Medline: 31785481]
27. Andrews JA, Craven MP, Jamnadas-Khoda J, Lang AR, Morriss R, Hollis C, et al. Health care professionals' views on using remote measurement technology in managing central nervous system disorders: qualitative interview study. J Med Internet Res 2020 Jul 24;22(7):e17414 [FREE Full text] [doi: 10.2196/17414] [Medline: 32706664]
28. Debong F, Mayer H, Kober J. Real-world assessments of mySugr mobile health app. Diabetes Technol Ther 2019 Jun;21(S2):S2-35-S2-40. [doi: 10.1089/dia.2019.0019]
29. Alzamanan MZ, Lim KS, Akmar Ismail M, Abdul Ghani N. Self-management apps for people with epilepsy: systematic analysis. JMIR Mhealth Uhealth 2021 May 28;9(5):e22489 [FREE Full text] [doi: 10.2196/22489] [Medline: 34047709]
30. Majid S, Reeves S, Figueredo G, Brown S, Lang A, Moore M, et al. The extent of user involvement in the design of self-tracking technology for bipolar disorder: literature review. JMIR Ment Health 2021 Dec 20;8(12):e27991 [FREE Full text] [doi: 10.2196/27991] [Medline: 34931992]
31. van Genugten CR, Schuurmans J, Lamers F, Riese H, Penninx BW, Schoevers RA, et al. Experienced burden of and adherence to smartphone-based ecological momentary assessment in persons with affective disorders. J Clin Med 2020 Jan 23;9(2):322 [FREE Full text] [doi: 10.3390/jcm9020322] [Medline: 31979340]
32. Dubad M, Winsper C, Meyer C, Livanou M, Marwaha S. A systematic review of the psychometric properties, usability and clinical impacts of mobile mood-monitoring applications in young people. Psychol Med 2018 Jan 23;48(2):208-228. [doi: 10.1017/S0033291717001659] [Medline: 28641609]

XSL•FO
RenderX

33. Palmier-Claus J, Lobban F, Mansell W, Jones S, Tyler E, Lodge C, et al. Mood monitoring in bipolar disorder: is it always helpful? Bipolar Disord 2021 Jun 19;23(4):429-431. [doi: 10.1111/bdi.13057] [Medline: 33570820]
34. Faurholt-Jepsen M, Munkholm K, Frost M, Bardram JE, Kessing LV. Electronic self-monitoring of mood using IT platforms in adult patients with bipolar disorder: a systematic review of the validity and evidence. BMC Psychiatry 2016 Jan 15;16(1):7 [FREE Full text] [doi: 10.1186/s12888-016-0713-0] [Medline: 26769120]
35. Ortiz A, Maslej MM, Husain MI, Daskalakis ZJ, Mulsant BH. Apps and gaps in bipolar disorder: a systematic review on electronic monitoring for episode prediction. J Affect Disord 2021 Dec 01;295:1190-1200. [doi: 10.1016/j.jad.2021.08.140] [Medline: 34706433]

Abbreviations

BD: bipolar disorder
COREQ: Consolidated Criteria for Reporting Qualitative Studies
PPI: patient and public involvement
RADAR: Remote Assessment of Disease and Relapse
RRI: responsible research and innovation

Edited by A Mavragani; submitted 06.08.24; peer-reviewed by M Mcinnis; comments to author 28.11.24; revised version received 04.12.24; accepted 15.12.24; published 07.02.25.

Please cite as:
Astill Wright L, Moore M, Reeves S, Vallejos EP, Morriss R
Improving the Utility, Safety, and Ethical Use of a Passive Mood-Tracking App for People With Bipolar Disorder Using Coproduction: Qualitative Focus Group Study
JMIR Form Res 2025;9:e65140
URL: https://formative.jmir.org/2025/1/e65140
doi:10.2196/65140
PMID:39918865

©Laurence Astill Wright, Matthew Moore, Stuart Reeves, Elvira Perez Vallejos, Richard Morriss. Originally published in JMIR Formative Research (https://formative.jmir.org), 07.02.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Social Media Use and Oral Health–Related Misconceptions in Saudi Arabia: Cross-Sectional Study

Rahaf Hamdan BinHamdan[1], BDS; Salwa Abdulrahman Alsadhan[1], BDS, MSc; Arwa Zohair Gazzaz[1], BDS, MPH, PhD; AlBandary Hassan AlJameel[1], BDS, MSc, PhD

Department of Periodontics and Community Dentistry, College of Dentistry, King Saud University, Riyadh, Saudi Arabia

Corresponding Author:
Rahaf Hamdan BinHamdan, BDS
Department of Periodontics and Community Dentistry
College of Dentistry
King Saud University
Diriyah 12372
Riyadh, 11545
Saudi Arabia
Phone: 966 0114677743
Email: rahafbinhamdan@gmail.com

Abstract

Background: Social media has become a central tool in health communication, offering both opportunities and challenges. In Saudi Arabia, where platforms like WhatsApp, Snapchat, and Instagram are widely used, the quality and credibility of oral health information shared digitally remain critical issues. Misconceptions about oral health can negatively influence individuals' behaviors and oral health outcomes.

Objective: This study aimed to describe the patterns of social media use and estimate the prevalence of oral health–related misconceptions among adults in Saudi Arabia. Additionally, it assessed the associations between engagement with oral health information, self-reported oral health, and the presence and count of these misconceptions.

Methods: A cross-sectional survey was conducted over 10 weeks, targeting adults aged 15 years and older in Saudi Arabia. Data were collected from a total sample size (n=387) via a questionnaire distributed through targeted advertisements on Instagram, TikTok, Snapchat, and X (Twitter). The prevalence of oral health–related misconceptions was estimated using descriptive statistics, including counts and percentages. Chi-square tests described sociodemographic, social media engagement, and self-reported oral health. Logistic and Poisson regression analyses were used to assess associations between engagement and self-reported oral health with misconceptions. Logistic regression models provided odds ratios and adjusted odds ratios with 95% CI to assess the presence of oral health misconceptions. Poisson regression was used to calculate mean ratios and adjusted mean ratios (AMRs) for the count of misconceptions.

Results: WhatsApp (n=344, 89.8%) and Instagram (n=304, 78.9%) were the most frequently used social media platforms daily. Common oral health misconceptions included beliefs that "Pregnancy causes calcium loss in teeth" (n=337, 87%) and "Dental treatment should be avoided during pregnancy" (n=245, 63.3%). Following dental-specific accounts was significantly associated with lower odds of having any misconceptions (adjusted odds ratio 0.41, 95% CI 0.22-0.78) and a lower count of misconceptions (AMR 0.87, 95% CI 0.77-0.98). Conversely, trust in social media as a source of oral health information was associated with a higher count of misconceptions (AMR 1.16, 95% CI 1.02-1.31).

Conclusions: Social media platforms are essential yet double-edged tools for oral health information dissemination in Saudi Arabia. Participants who followed dental-specific accounts had significantly lower misconceptions, while trust in social media as a source of information was linked to higher counts of misconceptions. These findings highlight the importance of promoting credible content from verified sources to combat misconceptions. Strategic collaborations with dental professionals are necessary to enhance the dissemination of accurate oral health information and public awareness and reduce the prevalence of oral health–related misconceptions.

(JMIR Form Res 2025;9:e70071) doi:10.2196/70071

XSL•FO
RenderX

KEYWORDS

social media; oral health; health misinformation; digital health; Saudi Arabia; public health; Instagram; Snapchat; TikTok; Twitter

Introduction

We live in a new era filled with rapid technological advancements that facilitate everyday life. The internet has been used worldwide for many purposes, and it has become an important tool that can improve the distribution of information and health care delivery and outcomes [1]. Web 2.0 is the second generation of the World Wide Web, emphasizing the shift from static web pages to more dynamic and participatory content, with social media becoming a key element shaping daily life [2]. Globally, social media platforms such as Facebook, Twitter, YouTube, and WhatsApp allow individuals to access and share information at a unique scale [3-5]. In Saudi Arabia, platforms like WhatsApp, Snapchat, and YouTube are heavily used for general communication and increasingly for seeking health-related information [6].

Social media platforms have emerged as powerful tools in health communication, offering both opportunities and challenges. In oral health, social media platforms can provide greater access to oral health information, engaging educational content, and web-based interaction with dental professionals [7,8]. For example, YouTube has been identified as a valuable resource for educating patients with leukemia on oral care [9]. Similarly, Snapchat-based interventions have effectively enhanced oral health knowledge among pregnant women in Saudi Arabia [10]. Oral health promotion campaigns delivered through social media platforms have been shown to improve oral health outcomes [11,12].

Despite social media benefits, information quality and reliability with sources often lack clear authorship or transparency, leading to difficulty in assessing information credibility for users [13-16]. This is particularly problematic in oral health, where misinformation—false or inaccurate health claims—could lead to negative oral health outcomes if it was taken by the users and distributed without consulting an expert [17,18]. Disinformation—information deliberately spread with the intent to harm—exacerbates these risks [19]. For example, practices such as using oil pulling to reduce mouth bacteria and prevent dental caries or relying on home remedy pastes to alleviate tooth pain are often shared digitally without scientific evidence [20]. This study uses “misconception” as a comprehensive term to encompass these various forms of inaccurate or misleading oral health information.

In addition to content concerns, sex-based differences in health behaviors, perceptions, and information-seeking patterns play a role in how individuals engage with oral health information on social media. It has been shown that women are more likely to seek web-based health information and engage with health care services compared to men [21]. They are also more likely to engage with and trust content shared on social media, making them a key demographic for addressing oral health misconceptions through these platforms [22]. Conversely, men may be less likely to search for web-based health information or disease prevention, which may lead to the persistence of misconceptions about oral health practices [23]. Sex differences explain that while women may be more susceptible to health misinformation due to their higher engagement, men may require distinct strategies to capture their attention and promote oral health awareness [24]. Such differences suggest that sex may influence the way individuals access and interpret oral health–related information on social media.

In Saudi Arabia, where 99% of the population are internet users actively engaging with social media platforms, this widespread use facilitates the diffusion of oral health misconceptions, leading users to accept inaccurate information and challenge health care professionals [25]. The challenge lies in the sheer quantity and variable quality of information available on social media platforms, which could lead to negative oral health outcomes if one misconception keeps spreading. For example, widespread beliefs about ineffective home remedies and misconceptions about neglecting primary teeth, dental treatment, and oral care during pregnancy continue to be prevalent among people in Saudi Arabia [26-28]. These misconceptions lead to public health implications, as untreated oral health issues can exacerbate systemic health conditions and increase health care costs [29]. To our knowledge, no previous study has specifically explored the context of oral health misconceptions on social media within this population, highlighting the need for this study.

This study aims to describe the patterns of social media use and to estimate the prevalence of oral health–related misconceptions among adults in Saudi Arabia. Additionally, the study aims to explore how social media engagement influences oral health misconceptions, with results stratified by sex to explore potential differences.

Methods

Study Design

This study is a cross-sectional survey. We collected the data over a 10-week period from September 9, 2024, to November 17, 2024, using an Arabic-language questionnaire developed on SurveyMonkey (SurveyMonkey Inc) among adults aged 15 years and older residing in Saudi Arabia. The questionnaire was initially developed in English and translated into Arabic to suit the target population. The questionnaire link was distributed through targeted web-based advertisements on the most popular social media platforms in Saudi Arabia [6], including Instagram, X (Twitter), Snapchat, and TikTok. Participants clicking on the advertisements were redirected to a dedicated web page containing a detailed study description and consent form, followed by the questionnaire. At the start of the survey, an introductory statement provided information on the study objectives, assured confidentiality, and included the researcher’s contact information. We obtained the consent via an “Agree” button, which participants clicked to access the survey. Participants were informed that the survey would take approximately five minutes to complete. They were assured

XSL•FO
RenderX

that participation is entirely voluntary, with the option to skip questions or withdraw at any time.

Sample Recruitment

The study was conducted in a digital setting with an open survey accessible to anyone who clicked on the advertisements, using widely used social media platforms in Saudi Arabia (Instagram, X [Twitter], Snapchat, and TikTok) for recruitment. These platforms were chosen due to their high penetration rates and active use in Saudi Arabia [6], ensuring ease of access and effective targeting through their advertising systems. We designed the social media–targeted advertisements to be culturally appropriate and visually engaging, highlighting the voluntary nature of participation and the study's focus on oral health–related information. The study used a simple random sampling approach facilitated by the advertisement algorithms. These algorithms were used to define the target group—adults aged 15 years and older residing in Saudi Arabia—ensuring that all individuals within this demographic had an equal chance of seeing the advertisement. According to 2022 population data from the General Authority of Statistics (GASTAT) [30] and 2023 internet penetration figures from the Communications, Space, and Technology Commission [6], the target population was estimated at 24,045,450 individuals, representing 99% of the total population in Saudi Arabia who are internet users within the targeted age group. This recruitment strategy aimed to: ensure representation of the population of interest by using an algorithm without prespecified interests, thereby targeting all individuals within the defined demographic; and maximize geographical diversity by specifying the algorithm to reach individuals residing in Saudi Arabia.

However, it is important to acknowledge that participation was voluntary, introducing self-selection bias that may affect the sample's generalizability. Despite this, the approach was a practical and effective method for recruiting a large, diverse population.

Effectiveness of the Selected Social Media Platforms

To effectively reach a broad and diverse sample, we launched advertisement campaigns across Instagram, X (Twitter), Snapchat, and TikTok. User interactions were monitored throughout the campaigns to evaluate reach and sample representativeness. Snapchat and TikTok provided us with strategic support to optimize campaign strategies. Snapchat emphasized targeting mechanisms such as predefined and custom audiences, advanced demographics, and accurate algorithmic delivery, which enhanced the ability to reach users based on demographics, location, and language. TikTok highlighted its robust targeting tools, including lookalike or custom audiences, ensuring advertisements were delivered to relevant users while monitoring user interactions like click-through and conversion rates. Although Instagram and X (Twitter) did not provide direct support, their platforms offered advanced targeting capabilities via Ads Manager tools. These tools allowed for audience segmentation by age, location, and sex, enabling effective recruitment across diverse demographics.

Sample Size

To estimate the prevalence of oral health–related misconceptions, with a 5% margin of error and a 95% CI, a sample size of 384 participants was required (assuming a 50% prevalence to maximize sample size requirements). Anticipating a low participation rate typical of web-based surveys, we assumed a 2% participation rate, leading us to target the reach of approximately 19,200 users in our advertisement campaigns across all platforms to ensure the required sample size. Reach refers to the estimated number of users who would see the advertisement, with metrics from Instagram, X (Twitter), Snapchat, and TikTok helping us track both views and link clicks.

Selection Criteria

We included both male and female participants aged 15 years and older residing in Saudi Arabia and those who provided informed consent to participate in the study. The age threshold of 15 was selected based on the classification by the GASTAT in Saudi Arabia, which classifies individuals aged 15 years and older as adults for statistical and demographic purposes [31]. Additionally, adolescents in this age group are active users of social media and are likely to be exposed to oral health information digitally, making their inclusion essential to assess the study objectives.

Questionnaire

We structured the questionnaire into five domains with a total of 29 fixed items distributed across six screens designed to align with the study's objectives and draw on previously published and validated cross-sectional studies. Face validity was established through expert review by a panel of dental public health specialists. Experts assessed the clarity, relevance, and domain alignment of each item, as well as the overall comprehensiveness of the questionnaire within the study's objectives. We implemented minor revisions to item wording and categorization based on their feedback (Multimedia Appendix 1). No adaptive questioning was used; all items required mandatory responses prior to proceeding to subsequent sections pages. Respondents could review and modify their answers using the "Back" button.

The Patterns of Social Media Use

The first domain assessed participants' use patterns and preferred platforms with one item [6,32-35]. Participants were asked to select their preferred social media platforms across the popular platforms in Saudi Arabia, which included X (Twitter), Snapchat, WhatsApp, Instagram, Facebook, TikTok, Telegram, YouTube, Line, and LinkedIn [6]. Responses were categorized as "more than once a day," "once a day," "2-3 times a week," "4-6 times a week," "once per week" or "once per month," and "I don't use this platform."

Engagement With Oral Health Information

The second domain, containing 6 items, focused on the engagement with oral health information obtained from social media [6,32,33,36,37]. Participants were asked binary (yes or no) questions, including whether they "use social media to search for oral health information," "ever received oral health

XSL•FO
RenderX

info from nondental professionals on social media," "following dental-specific accounts on social media," "trust in oral health information from social media," and "influence their trust by social media profile content." Perception of social media as a good source for oral health information was assessed with response options: "yes," "sometimes," and "no." We dichotomize it to combine "yes" and "sometimes" into a single category representing participants who perceived social media as a good source of oral health information.

Self-Reported Oral Health

To capture self-reported oral health, we included a single item in the third domain using the Global Self-Ratings of Oral Health question for participants to rate their own oral health status on a 3-point Likert scale as "good," "average," and "poor." [38]. For analysis, self-reported poor oral health was defined as participants selecting "poor," while better oral health perceptions included those reporting "good" oral health. Participants who neither rated their oral health as good nor poor, reflecting a moderate perception of their oral health, included those reporting "average" status.

Oral Health–Related Misconceptions

The fourth domain presented 14 statements on common oral health misconceptions in Saudi Arabia, derived from literature [27,28], with response options of "agree," "disagree," and "don't know." We combined responses of "disagree" and "don't know" into a single category representing participants who do not believe in the misconceptions. The category "agree" includes those who do agree with these misconceptions.

Sociodemographic Factors

In the fifth domain, we collected sociodemographic information across seven items: age, which was categorized into five groups: 15-24, 25-34, 35-44, 45-54, and 55+ years; sex; education, which was classified into "high school or less," "associate or intermediate diploma," "bachelor's degree," and "master's or PhD"; occupation, which was grouped into "student," "governmental worker," "private worker," and "unemployed"; marital status, "single," "married," or "other"; region of residence, included the 13 administrative regions of Saudi Arabia: "Riyadh," "Makkah," "Madinah," "Qassim," "Eastern Region," "Asir Region," "Tabuk," "Hail," "Northern Region," "Jazan," "Najran," "Al-Baha," and "Al-Jawf"; and nationality, "Saudi" and "Non-Saudi" within Saudi Arabia. Age and region categories were based on distributions from the GASTAT [30,39], and education levels followed the National Qualifications Framework 2023 [40].

Pilot

In a preliminary phase of the study, we conducted a pilot to confirm the validity and relevance of this study prior to full implementation. We targeted a sample group representing 10% of the total study sample size, with 10 participants recruited from each of the four social media platforms, which are Instagram, X (Twitter), Snapchat, and TikTok, resulting in a total of 40 participants. Participants responded positively to the questionnaire, navigating it without difficulty. Based on feedback, we made two key adjustments to improve the survey. First, we added a "sometimes" response option to the question in the second domain, "Do you think social media is a good source for obtaining oral health information?" to provide a wider range of responses. Second, we simplified certain Arabic terms in the misconception domain to make the statements clear to participants. This pilot phase was conducted from June 17, 2024, to September 2, 2024. Participants involved in this pilot were not included in the final study sample.

Statistical Analysis

Counts and percentages were reported to describe participants' sociodemographic characteristics, their engagement with oral health information, and their self-reported oral health. We used the chi-square test to examine potential associations between the 3 domains and sex, identifying statistically significant relationships where applicable.

For the multivariable analysis, the independent variables included engagement with the oral health information domain (the manner of participants' engagement with oral health–related information in social media measured via the six binary items, each item entered as a separate exposure in the model) and self-reported oral health (participants' subjective evaluation of their oral health status measured via 3-point Likert scale). The dependent variables were defined as:

- Presence of oral health misconceptions (binary): had at least one misconception based on the 14 statements on oral health misconceptions.
- Number of oral health misconceptions (count): the total number of misconceptions identified per participant.

To assess these relationships, we used logistic regression models to present both crude (odds ratio) and adjusted odds ratio (AOR), with respective 95% CI, to examine engagement with oral health information and self-reported oral health (exposure variables) with the presence of any oral health misconceptions among social media users (outcome). Logistic regression was selected due to its appropriateness for modeling binary outcomes, allowing us to examine the likelihood of participants holding at least one misconception. Poisson regression is used to report the mean ratio and adjusted mean ratio (AMR) of oral health misconceptions (outcome) by the engagement with oral health information and self-reported oral health (exposure variables). Poisson regression was chosen because it is suitable for modeling count data, particularly when the outcome variable represents a frequency count of misconceptions per participant. We confirmed the appropriateness of the Poisson model by assessing equidispersion using the Pearson chi-square statistic, which indicated no overdispersion. Sex, education (as a measure of socioeconomic status), and nationality were selected and adjusted as potential confounders based on their theoretical relevance [41] and observed associations with both the exposure and the outcome in exploratory analyses.

A sensitivity analysis was performed for both regression models to assess the potential impact on the results if the response "don't know" was reclassified as "agree" in the 14 statements addressing oral health misconceptions. We reclassified "don't know" responses based on the assumption that these individuals might be more susceptible to misconceptions due to a lack of knowledge or confidence. This provided a more conservative,

XSL•FO
RenderX

robust estimate of misconception prevalence, allowing us to examine the influence of uncertainty on our findings across different analyses. A 2-tailed α level of .05 was used to test relationships in both directions, and all statistical analyses were conducted using Stata/BE (version 17.0; StataCorp).

Ethical Considerations

This study was conducted in compliance with the ethical standards set by the institutional review board of King Saud University. Research ethical approval was obtained, and the study was registered with the Institutional Review Board from the Committee of Human and Social Research Ethics, King Saud University (registration KSU-HE-23848). Participants were informed about the study objectives, procedures, and their rights before participating in the survey. Informed consent was obtained electronically via a mandatory "Agree" button at the start of the questionnaire. Only individuals who provided consent were able to proceed with the survey. The survey content was specifically designed to be low-risk without including sensitive or intrusive questions. Data were collected anonymously, with no personally identifiable information linked to the responses. The platform (SurveyMonkey) complied with strict data security and privacy protocols to ensure the protection of participants' responses. IP addresses were monitored to identify duplicate entries within a 24-hour period and then omitted before analysis. Participation was voluntary, and no monetary or material compensation was provided to participants in this study. This study did not involve the use of images or media containing identifiable participants.

Results

Campaign Reach and Interaction Across Social Media Platforms

Using targeted social media campaigns across Instagram, X (Twitter), TikTok, and Snapchat, we observed notable differences in reach and interaction across platforms. TikTok achieved the highest reach (836,477) and link clicks (67,678) during its 30-day campaign, followed by Instagram, which reached (53,341) users and generated (1007) link clicks over 60 days. X (Twitter) and Snapchat had comparatively lower interactions. Across all platform advertisements, the overall view rate was 6.6%, and the participation rate was 0.7%.

Sociodemographic Characteristics

Out of 640 individuals who initially responded to the questionnaire, 387 (76%) individuals completed it fully; the remaining (n=253) either provided incomplete responses or exited the survey after clicking "disagree" to participate in the questionnaire. These partial responses were excluded from the analysis to ensure data completeness.

The final sample size consisted of 248 (64%) female and 139 (35.9%) male participants (Table 1). The largest age group was 15-24 years (n=221, 57.1%), followed by 25-34 years (n=77, 19.9%). Most participants held a bachelor's degree (n=184, 47.5%) or a high school diploma or less (n=138, 35.6%). Regarding occupation, more than half of the participants were students (n=199, 51.4%), while 97 (25%) participants were unemployed. In terms of marital status, 299 (77.2%) of participants were single, while 82 (21.1%) of participants were married. Saudi participants accounted for 72.3% (n=280) of the sample, and most participants resided in Riyadh (n=258, 66.6%). Significant sex-based differences were observed in occupation ($P<.001$) and marital status ($P=.02$), with male participants more likely to be employed and married compared to female participants, who were predominantly unemployed and single (Table 1).

XSL•FO
RenderX

Table 1. Sociodemographic characteristics and regional distribution among adults aged 15 years and older residing in Saudi Arabia (n=387), described by sex[a].

Sociodemographic characteristics	Overall (n=387, 100%), n (%)	Male (n=139, 35.9%), n (%)	Female (n=248, 64%), n (%)	*P* value
Age (in years)				.35
15-24	221 (57.1)	78 (35.2)	143 (64.7)	
25-34	77 (19.9)	22 (28.5)	55 (71.4)	
35-44	45 (11.6)	19 (42.2)	26 (57.7)	
45-54	36 (9.3)	16 (44.4)	20 (55.5)	
55+	8 (2)	4 (50)	4 (50)	
Education				.23
High school or less	138 (35.6)	49 (35.5)	89 (64.4)	
Associate—intermediate diploma	39 (10)	15 (38.4)	24 (61.5)	
Bachelor's degree	184 (47.5)	61 (33.1)	123 (66.8)	
Master's degree—PhD	26 (6.7)	14 (53.8)	12 (46.1)	
Occupation				<.001
Student	199 (51.4)	69 (34.6)	130 (65.3)	
Governmental worker	43 (11.1)	24 (55.8)	19 (44.1)	
Private worker	48 (12.4)	28 (58.3)	20 (41.6)	
Unemployed	97 (25)	18 (18.5)	79 (81.4)	
Marital status				.02
Single	299 (77.2)	98 (32.7)	201 (67.2)	
Married	82 (21.1)	40 (48.7)	42 (51.2)	
Other	6 (1.5)	1 (16.6)	5 (83.3)	
Nationality				.92
Saudi	280 (72.3)	101 (36)	179 (63.9)	
Non-Saudi	107 (27.6)	38 (35.5)	69 (64.4)	
Region of residence				.19
Riyadh	258 (66.6)	84 (32.5)	174 (67.4)	
Makkah	45 (11.6)	21 (46.6)	24 (53.3)	
Madinah	13 (3.3)	7 (53.8)	6 (46.1)	
Qassim	11 (2.8)	2 (18.1)	9 (81.8)	
Eastern Region	45 (11.6)	21 (46.6)	24 (53.3)	
Asir Region	5 (1.2)	1 (20)	4 (80)	
Tabuk	1 (0.2)	0 (0)	1 (100)	
Hail	2 (0.5)	0 (0)	2 (100)	
Northern Region	1 (0.2)	0 (0)	1 (100)	
Jazan	3 (0.7)	1 (33.3)	2 (66.6)	
Najran	1 (0.2)	1 (100)	0 (0)	
Al-Baha	1 (0.2)	1 (100)	0 (0)	
Al-Jawf	1 (0.2)	0 (0)	1 (100)	

[a]Data were collected in a cross-sectional survey on oral health misconceptions conducted from September to November 2024.

XSL•FO
RenderX

Engagement With Oral Health Information and Self-Reported Oral Health

Female participants were significantly more likely than male participants to use social media for oral health information (n=156, 68.7% compared to n=71, 31.2%; *P*=.02) and follow dental-specific accounts (n=126, 72% compared to n=49, 28%; *P*=.003; Table 2). For self-reported oral health, female participants were significantly more likely to rate their oral health as good (n=98, 73.1%) or average (n=125, 60.1%), while poor oral health was significantly slightly more common among male participants (n=20, 44.4%; *P*=.02).

Table 2. Engagement with oral health information and self-reported oral health described by sex among adults aged 15 years and older residing in Saudi Arabia (n=387)[a].

Characteristics	Overall (n=387, 100%), n (%)	Male (n=139, 35.9%), n (%)	Female (n=248, 64%), n (%)	*P* value
Engagement with oral health information				
Use of social media to search for oral health information	227 (58.6)	71 (31.2)	156 (68.7)	.02
Ever received oral health information from nondental professionals on social media	235 (60.7)	86 (36.6)	149 (63.4)	.73
Following dental-specific accounts on social media	175 (45.2)	49 (28)	126 (72)	.003
Trust in oral health information from social media	171 (44.1)	54 (31.5)	117 (68.4)	.11
Social media profile content influence on trust in oral health information	289 (74.6)	105 (36.3)	184 (63.6)	.77
Perception of social media as a good source for oral health information	349 (90.1)	126 (36.1)	223 (63.9)	.82
Self-reported oral health				.02
Good	134 (34.6)	36 (26.8)	98 (73.1)	
Average	208 (53.7)	83 (39.9)	125 (60.1)	
Poor	45 (11.6)	20 (44.4)	25 (55.5)	

[a]Data were collected in a cross-sectional survey on oral health misconceptions conducted from September to November 2024.

Patterns of Social Media Use

The patterns of social media use are summarized in Table 3. Of the total sample, the most frequently used platform on a daily basis was WhatsApp (n=344, 89.8%), followed by Instagram (n=304, 78.9%), Snapchat (n=239, 62%), and YouTube (n=196, 50.7%). The majority of our sample reported not using Line (n=374, 97.1%), Facebook (n=295, 76.4%), and LinkedIn (n=271, 70%). Telegram had a proportion of weekly users (n=134, 35.1%), while TikTok showed mixed use with 49.7% (n=192) reporting daily use and 31% (n=120) reporting no use.

XSL•FO
RenderX

Table 3. The patterns of social media use among adults aged 15 years and older residing in Saudi Arabia described by daily, weekly, monthly, and no use of commonly used platforms in Saudi Arabia[a].

Social media platforms	Daily use[b], n (%)	Weekly use[c], n (%)	Monthly use[d], n (%)	No use[e], n (%)
X (Twitter)	180 (46.8)	98 (25.5)	37 (9.6)	69 (17.9)
Snapchat	239 (62)	49 (12.7)	11 (2.8)	86 (22.3)
WhatsApp	344 (89.8)	32 (8.3)	2 (0.5)	5 (1.3)
Instagram	304 (78.9)	52 (13.5)	9 (2.3)	20 (5.1)
TikTok	192 (49.7)	61 (15.8)	13 (3.3)	120 (31)
Facebook	32 (8.2)	29 (7.5)	30 (7.7)	295 (76.4)
Telegram	146 (38.3)	134 (35.1)	51 (13.3)	50 (13.1)
YouTube	196 (50.7)	155 (40.1)	21 (5.4)	14 (3.6)
Line	4 (1)	5 (1.3)	2 (0.5)	374 (97.1)
LinkedIn	26 (6.7)	47 (12.1)	42 (10.8)	271 (70.2)

[a]Data were collected in a cross-sectional survey on oral health misconceptions conducted from September to November 2024.

[b]Daily use: users who use the platform once a day or more than once a day.

[c]Weekly use: users who use the platform once a week, 2-3 times a week, or 4-6 times a week. [d]Monthly use: users who use the platform once a month.

[e]No use: users who are not using the platform.

Oral Health Misconceptions

Table 4 highlights the prevalence of oral health misconceptions among participants. Overall, 96.9% (n=375) of participants reported having at least one or more of the identified misconceptions. When stratified by sex, 93.5% (n=130) of male and 98.8% (n=245) of female participants were found to hold at least one or more misconceptions. Common misconceptions included “During pregnancy, the baby absorbs calcium from the mother’s teeth and bones” (n=337, 87%) and “Pregnant women should only undergo dental treatment after childbirth” (n=245, 63.3%). Additionally, 59.9% (n=232) of participants believed that “You should not eat anything when going for tooth extraction.” Sex-based differences were observed for the misconception that “Using a hard-bristled toothbrush makes teeth whiter,” with male participants more likely to agree (n=38, 57.5%) than female participants (n=28, 43.4%; Table 4).

XSL•FO
RenderX

Table 4. The proportion of responses to common oral health misconceptions among adults aged 15 years and older residing in Saudi Arabia, described by sex[a].

Oral health misconceptions	Overall, n (%)	Male, n (%)	Female, n (%)
During pregnancy, the baby absorbs calcium from the mother's teeth and bones	337 (87)	108 (32)	229 (67.9)
Pregnant women should only undergo dental treatment after childbirth	245 (63.3)	89 (36.3)	156 (63.6)
You should not eat anything when going for tooth extraction	232 (59.9)	90 (38.7)	142 (61.2)
Scaling weakens the structure of the teeth	224 (57.8)	80 (35.7)	144 (64.2)
Upper teeth extraction affects the brain	218 (56.3)	66 (30.2)	152 (69.7)
Brushing with salt helps whiten teeth	183 (47.2)	63 (34.4)	120 (65.5)
It is better not to brush your teeth when you have bleeding gums	157 (40.5)	69 (43.9)	88 (56)
Extracted teeth do not need to be replaced with artificial teeth	146 (37.7)	59 (40.4)	87 (59.5)
Taking antibiotics will relieve the tooth pain	137 (35.4)	50 (36.6)	87 (63.5)
If there is no tooth pain, there is no need to visit the dentist	130 (33.5)	51 (39.2)	79 (60.7)
Leaving a milk bottle in the baby's mouth during sleep does not harm teeth	114 (29.4)	53 (46.4)	61 (53.5)
The appearance of wisdom teeth increases one's wisdom	82 (21.1)	31 (37.8)	51 (62.2)
Using a hard-bristled toothbrush makes teeth whiter	66 (17)	38 (57.5)	28 (43.4)
There is no need to care for primary teeth, as they will be replaced by permanent teeth anyway	64 (16.5)	38 (59.3)	26 (40.6)
Having at least one or more of the above-mentioned misconceptions	375 (96.90)	130 (93.53)	245 (98.79)

[a]Data were collected in a cross-sectional survey on oral health misconceptions conducted from September to November 2024.

Association Between Engagement With Oral Health Information and Self-Reported Oral Health With Oral Health Misconceptions

Logistic and Poisson regression analyses (Table 5) revealed significant associations between following dental-specific accounts on social media with lower odds of having any misconceptions (AOR 0.41, 95% CI 0.22-0.78) and a lower count of misconceptions (AMR 0.87, 95% CI 0.77-0.98). Trust in social media for oral health information was significantly associated with a higher count of misconceptions (AMR 1.16, 95% CI 1.02-1.31). Self-reported oral health showed that participants with poor oral health had a higher count of misconceptions (AMR 1.23, 95% CI 0.99-1.50), although the association was not statistically significant. After conducting the sensitivity analysis by combining "I don't know" with "agree" as a single variable, there was no difference in most of the observations, except that participants who perceived social media as a good source of oral health information had a statistically significant higher count of misconceptions (AMR 1.18, 95% CI 1.02-1.37). This reclassification suggests that uncertainty about oral health information may reflect susceptibility to misconception.

XSL•FO
RenderX

Table 5. Logistic and Poisson regression models of the association between engagement with oral health information and self-reported oral health with the presence and count of oral health misconceptions among adults aged 15 years and older residing in Saudi Arabia[a].

Characteristics	Any misconceptions		Number of misconceptions	
	Crude OR[b] (95% CI)	Adjusted OR (95% CI)	Crude MR[c] (95% CI)	Adjusted MR (95% CI)
Engagement with oral health information				
Use of social media to search for oral health information				
No	reference	reference	reference	reference
Yes	1.20 (0.67-2.16)	1.07 (0.59-1.97)	1.12 (0.99-1.28)	1.12 (0.99-1.28)
Ever received oral health info from nondental professionals on social media				
No	reference	reference	reference	reference
Yes	1.33 (0.74-2.39)	1.33 (0.73-2.44)	1.01 (0.89-1.15)	1.00 (0.89-1.14)
Following dental-specific accounts on social media				
No	reference	reference	reference	reference
Yes	0.49 (0.27-0.89)[d]	0.41 (0.22-0.78)[d]	0.85 (0.76-0.97)[d]	0.87 (0.77-0.98)[d]
Trust in oral health information from social media				
No	reference	reference	reference	reference
Yes	1.81 (0.98-3.35)	1.68 (0.89-3.17)	1.16 (1.02-1.31)[d]	1.16 (1.02-1.31)[d]
Social media profile content influence on trust in oral health information				
No	reference	reference	reference	reference
Yes	1.47 (0.79-2.77)	1.73 (0.89-3.35)	1.10 (0.95-1.28)	1.12 (0.97-1.30)
Perception of social media as a good source for oral health information				
No	reference	reference	reference	reference
Yes	0.95 (0.35-2.55)	0.93 (0.34-2.58)	1.21 (0.97-1.51)	1.20, (0.96-1.50)
Self-reported oral health				
Good	reference	reference	reference	reference
Average	1.31 (0.71-2.42)	1.33 (0.70-2.53)	1.17 (1.01-1.34)	1.14 (0.99-1.31)
Poor	2.01 (0.65-6.19)	2.11 (0.66-6.76)	1.26 (1.03-1.54)	1.23 (0.99-1.50)

[a]Data were collected in a cross-sectional survey on oral health misconceptions conducted from September to November 2024.

[b]OR: odds ratio.

[c]MR: mean ratio. Adjusted for sex, education, and nationality.

[d]Indicates statistical significance ($P<.05$).

Discussion

Main Findings

This study investigated the patterns of social media use and the prevalence of oral health misconceptions among Saudi adults, focusing on the relationship between engagement with oral health information on social media and self-reported oral health with oral health misconceptions. The study highlighted significant associations between following dental-specific accounts and lower oral health–related misconceptions, while higher trust in social media information was associated with a higher count of oral health–related misconceptions. Additionally, self-reported poor oral health was associated with higher counts of oral health–related misconceptions.

A significant portion of our sample consisted of students and women. The high proportion of students in our sample reflects that students are known to be highly active on social media and rely on these platforms for information seeking, including health-related information. This aligns with findings from a scoping review indicating that social media plays a dominant role in health education in Saudi Arabia and that younger individuals and women frequently seek medical and health-related advice through digital platforms [42].

The results of our targeted social media campaigns reveal distinct differences in audience reach and interaction across various platforms. TikTok, in particular, demonstrated a significant advantage in terms of reach. This aligns with previous studies that highlight TikTok's rapid rise as a dominant platform for user interaction, especially among younger audiences [43,44]. Instagram, known for its visual-centric content, also proved effective in reaching a wide audience and generating meaningful interactions in this study. Instagram has been shown to be effective for health communication because

XSL•FO
RenderX

of its ability to attract users through images, videos, and infographics, which enhance user interaction and retention [45]. In contrast, X (Twitter) and Snapchat showed lower interaction rates, which can be explained by their platform-specific characteristics. In contrast, both X (Twitter), with its focus on quick updates and real-time communication, and Snapchat, with its temporary content format appealing to younger users seeking informal and immediate interactions, limit their potential for sustained and broad interaction [46].

TikTok's effectiveness in disseminating health information can be attributed to its short-form video format, which caters to modern users' preference for quick, easily digestible content [47]. This format has not only captured the attention of younger audiences but also influenced other social media platforms to adopt similar features. The platform's popularity among younger demographics further enhances its role in health communication, as it serves as a primary source of information for this age group. Moreover, TikTok's participatory features, such as duets and challenges, encourage user participation and the viral spread of health-related content [48].

Our methodology assumed that using social media advertisement campaigns would effectively reach our target population, with advertisement reach aligning with the demographic distribution in Saudi Arabia, where men represent a larger proportion (61.2%) [49]. The term "ad recipient" refers specifically to individuals who were exposed to the advertisement (ie, saw the advertisement) without necessarily engaging with it by clicking. Men comprised the majority of advertisement recipients across all platforms: Snapchat (n=28,235, 56%), TikTok (n=625,665, 74.32%), Instagram (n=24,625, 57.2%), and X (Twitter; n=107,116, 61%), similar to the male demographic distribution in Saudi Arabia [49]. However, despite greater advertisement exposure among men, women represented the majority of questionnaire respondents. This difference in interaction may be explained by behavioral differences, as previous studies have found that women are more likely to interact with health-related advertisements and content on social media than men [50-52]. This trend may be driven by greater health awareness among women, a higher likelihood of managing family health matters, and cultural norms that encourage women to be more involved in health-related decision-making [53,54].

This study identified several common oral health–related misconceptions. The most prevalent misconceptions include the belief that pregnant women lose calcium during pregnancy, dental procedures should be postponed during pregnancy, eating before dental extractions is harmful, and scaling weakens tooth structure. These findings are consistent with earlier studies conducted in Saudi Arabia, which also documented widespread myths about oral health [26-28].

Many of these oral health misconceptions can negatively impact oral health by discouraging preventive behaviors and delaying necessary treatment. For example, myths about pregnancy and oral health, such as the belief that "Pregnancy causes calcium loss in teeth" or that "Dental treatment should be avoided during pregnancy," may lead to untreated dental issues, increasing the risk of adverse pregnancy outcomes like preterm birth and low birth weight [55]. Similarly, misconceptions about primary teeth being unimportant contribute to early childhood caries, which affects children's overall well-being and increases the likelihood of future dental problems [56]. Additionally, beliefs in ineffective home remedies, such as using salt for whitening or avoiding professional scaling, can lead to poor oral hygiene and a higher risk of periodontal disease [57,58]. Based on these observations, addressing these common oral health–related misconceptions through targeted public health campaigns and education is essential. Engaging the community to participate in the educational content, primarily through social media platforms, would be an active way of enhancing people's awareness.

The high prevalence of misconceptions underscores the challenges in addressing oral health information in the digital age. While social media platforms provide an avenue for disseminating information, their effectiveness depends on the quality and credibility of the content shared.

This study observed associations between trust in social media for oral health information and higher misconceptions are consistent with broader concerns about the role of social media in propagating misconceptions in health communication [59]. The user-driven nature of social media echo chambers, combined with algorithm-driven content personalization, can reinforce existing beliefs, making users more susceptible to content that aligns with their biases and less likely to evaluate or seek out alternative perspectives critically. Social media algorithms on platforms such as TikTok, Instagram, and Facebook select content based on users' past interactions, prioritizing posts that align with their preference history. As a result, users who have previously interacted with nonevidence-based health claims are more likely to be exposed to similar misconceptions [60,61]. Conversely, the protective effect of following dental-specific accounts aligns with prior studies that emphasize the importance of verified content from credible sources in mitigating health misconceptions [62-65]. On social media, accounts maintained by professionals or organizations with expertise in a given field act as anchors of credibility, reducing the likelihood that users will engage with inaccurate information [59].

Strengths and Limitations of the Study

Our research specifically investigates oral health–related misconceptions in Saudi Arabia, an area understudied despite social media's growing influence in shaping health beliefs. This study's multiplatform social media approach enabled a comprehensive analysis of social media's role in oral health information. The use of targeted advertisements across multiple platforms, including Snapchat, TikTok, Instagram, and X (Twitter), allowed for the random selection of participants, enhancing the representativeness of the sample. We provided robust insights into the associations between engagement behaviors and self-reported oral health with oral health misconceptions by using advanced statistical modeling, including logistic and Poisson regression analyses. However, the cross-sectional design limits the ability to establish temporal relationships between the variables. Moreover, the reliance on self-reported data introduces potential response and recall biases, as participants may have overstated their engagement with oral health information or underreported misconceptions due to

XSL•FO
RenderX

social desirability. Additionally, recall bias could have affected their ability to recall their pattern of social media use accurately. Furthermore, while we adjusted for key confounders, unmeasured variables may have influenced both the exposures and outcomes, potentially affecting the observed associations.

Implication of the Study

Our findings align with previous research that highlights social media's role as a valuable channel for disseminating health information and as a potential source of misconceptions [19,66]. Therefore, the highly used platforms like WhatsApp, Instagram, Snapchat, and YouTube, as observed in this study, highlight their potential as valuable tools for web-based dental public health campaigns in Saudi Arabia. These platforms are widely used daily, reflecting their integration into the community's digital habits and their capacity to reach broad and diverse audiences. To address oral health–related misconceptions, health authorities, dental professionals, and regulatory bodies should consider the following:

- Developing social media–targeted campaigns that collaborate with credible, well-followed dental professional creators to disseminate accurate and engaging oral health information. This can include posting participatory videos, photos, and Q and A sessions to address common oral health–related misconceptions identified in our findings. Use social media features tailored to user interaction—such as TikTok's targeting for reaching young adults, Instagram Reels for quick educational content, Snapchat Stories for oral health tips, and X (Twitter) threads to counter oral health–related misconceptions.
- Establishing Saudi health authorities verified social media pages that consistently share reliable oral health information, combat misconceptions, and engage with users. These dental-specific accounts should interact with trending dental topics to redirect audiences toward credible sources.
- Legalizing and monitoring both oral health–related advertisements and information in social media from Saudi regulatory bodies would help protect public health and ensure that individuals receive evidence-based information in social media. For example, the European Union's Digital Services Act mandates that web-based platforms prevent the spread of misinformation, including health-related content, by enforcing fact-checking measures and transparency requirements for web-based advertisements [67].

While this study provides valuable insights, future research is needed to investigate how different content formats (eg, videos, photos, and participatory posts) and features (eg, likes, comments, and shares) influence the dissemination and retention of oral health information on social media. Experimental studies can assess which formats are most effective in improving knowledge and correcting oral health–related misconceptions, while longitudinal research can explore how repeated exposure to accurate content impacts behavioral change over time.

Conclusions

The study underscores the role of social media in shaping oral health perceptions among adults in Saudi Arabia. Engagement with dental-specific accounts was associated with fewer misconceptions, while trust in oral health information in social media was associated with higher misconceptions. Platforms like WhatsApp, Instagram, TikTok, and Snapchat are widely used in Saudi Arabia and offer the potential for targeted dental public health campaigns. Strategic collaborations with credible dental professional creators to disseminate accurate and engaging oral health information. Creating Instagram reels for quick educational content or using TikTok's targeting features could enhance information retention and user interaction. Although this study focuses on Saudi Arabia, its findings have broader relevance, particularly in regions with high social media penetration and similar digital health-seeking behaviors, such as the Gulf countries. From a policy perspective, legalizing and monitoring oral health–related advertisements and information in social media should be implemented by regulatory bodies to protect public health.

Future research should build on these findings to investigate the effectiveness of varying social media platforms' content formats and features in improving oral health knowledge and correcting misconceptions through experimental studies. Longitudinal research should also explore the long-term impact of repeated exposure to accurate information on behavioral change. These insights provide a basis for stakeholders to formulate evidence-based, innovative strategies to improve public health in the context of the digital age.

Acknowledgments

The authors would like to thank the Deanship of Scientific Research at King Saud University for funding and supporting this research through the initiative of DSR Graduate Students Research Support. In addition, we would like to express our sincere gratitude to Muath Aldosari, BDS, MPH, DMSc, for his invaluable guidance throughout the analysis phase of this study. His expertise and willingness to provide feedback during multiple consultations were instrumental in navigating the complexities of data analysis.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to ethical restrictions and commitments that mandate strict confidentiality of participant information and prohibit the use of data for purposes beyond the approved research study outlined by the institutional review board at King Saud University, but are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Authors' Contributions

RHB conceived and designed the study, collected and analyzed the data, and wrote the manuscript draft. SAA contributed to the research design, methodology, results, and writing process. AZG contributed to the methodology, results, and discussion sections and edited the manuscript. AHA supervised the study, provided guidance on the research idea, and contributed to the research methodology, results, and discussion. All authors reviewed and revised the manuscript and approved the final version for submission.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Questionnaire used in the study to assess social media usage patterns, engagement with oral health information, self-reported oral health, oral health misconceptions, and sociodemographic characteristics among adults in Saudi Arabia.
[DOCX File , 22 KB - formative_v9i1e70071_app1.docx]

References

1. Baker L, Wagner TH, Singer S, Bundorf MK. Use of the internet and e-mail for health care information: results from a national survey. JAMA 2003;289(18):2400-2406. [doi: 10.1001/jama.289.18.2400] [Medline: 12746364]
2. Vuori M. Exploring uses of social media in a global corporation. J Syst Inf Technol 2012;14(2):155-170. [doi: 10.1108/13287261211232171]
3. Gottfried J. Americans' social media use. Pew Research Center. 2024. URL: https://www.pewresearch.org/internet/2024/01/31/americans-social-media-use/ [accessed 2024-01-31]
4. Grajales FJ, Sheps S, Ho K, Novak-Lauscher H, Eysenbach G. Social media: a review and tutorial of applications in medicine and health care. J Med Internet Res 2014;16(2):e13 [FREE Full text] [doi: 10.2196/jmir.2912] [Medline: 24518354]
5. Smailhodzic E, Hooijsma W, Boonstra A, Langley DJ. Social media use in healthcare: a systematic review of effects on patients and on their relationship with healthcare professionals. BMC Health Serv Res 2016;16(1):442 [FREE Full text] [doi: 10.1186/s12913-016-1691-0] [Medline: 27562728]
6. Saudi internet usage report 2023. Saudi Communications, Space, and Technology Commission. 2023. URL: https://cst.gov.sa [accessed 2024-11-20]
7. de Oliveira Júnior AJ, Oliveira JM, Bretz YP, Mialhe FL. Online social networks for prevention and promotion of oral health: a systematic review. Can J Dent Hyg 2023;57(2):83-97 [FREE Full text] [Medline: 37464997]
8. Sivrikaya EC, Yilmaz O, Sivrikaya P. Dentist-patient communication on dental anxiety using the social media: a randomized controlled trial. Scand J Psychol 2021;62(6):780-786. [doi: 10.1111/sjop.12769] [Medline: 34333783]
9. Şallı GA, Egil E. Are YouTube videos useful as a source of information for oral care of leukemia patients? Quintessence Int 2020;51(1):78-85. [doi: 10.3290/j.qi.a43665] [Medline: 31781693]
10. Aboalshamat K, Alharbi J, Alharthi S, Alnifaee A, Alhusayni A, Alhazmi R. The effects of social media (Snapchat) interventions on the knowledge of oral health during pregnancy among pregnant women in Saudi Arabia. PLoS One 2023;18(2):e0281908 [FREE Full text] [doi: 10.1371/journal.pone.0281908] [Medline: 36795745]
11. Deghatipour M, Ghorbani Z, Mokhlesi AH, Ghanbari S, Namdari M. Effect of oral health promotion interventions on pregnant women dental caries: a field trial. BMC Oral Health 2022;22(1):280 [FREE Full text] [doi: 10.1186/s12903-022-02292-1] [Medline: 35804346]
12. Zotti F, Dalessandri D, Salgarello S, Piancino M, Bonetti S, Visconti L, et al. Usefulness of an app in improving oral hygiene compliance in adolescent orthodontic patients. Angle Orthod 2016;86(1):101-107 [FREE Full text] [doi: 10.2319/010915-19.1] [Medline: 25799001]
13. Moorhead SA, Hazlett DE, Harrison L, Carroll JK, Irwin A, Hoving C. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. J Med Internet Res 2013;15(4):e85 [FREE Full text] [doi: 10.2196/jmir.1933] [Medline: 23615206]
14. Househ M, Borycki E, Kushniruk A. Empowering patients through social media: the benefits and challenges. Health Informatics J 2014;20(1):50-58 [FREE Full text] [doi: 10.1177/1460458213476969] [Medline: 24550564]
15. De Martino I, D'Apolito R, McLawhorn AS, Fehring KA, Sculco PK, Gasparini G. Social media for patients: benefits and drawbacks. Curr Rev Musculoskelet Med 2017;10(1):141-145 [FREE Full text] [doi: 10.1007/s12178-017-9394-7] [Medline: 28110391]
16. Wang Y, McKee M, Torbica A, Stuckler D. Systematic literature review on the spread of health-related misinformation on social media. Soc Sci Med 2019;240:112552 [FREE Full text] [doi: 10.1016/j.socscimed.2019.112552] [Medline: 31561111]
17. Sumayyia MD, Al-Madaney MM, Almousawi FH. Health information on social media. Perceptions, attitudes, and practices of patients and their companions. Saudi Med J 2019;40(12):1294-1298 [FREE Full text] [doi: 10.15537/smj.2019.12.24682] [Medline: 31828284]

XSL•FO
RenderX

18. Office of the Surgeon General (OSG). Confronting health misinformation: The U.S. Surgeon General's Advisory on building a healthy information environment. Washington (DC): US Department of Health and Human Services 2021:22 [FREE Full text] [Medline: 34283416]
19. Chou WYS, Gaysynsky A, Cappella JN. Where we go from here: health misinformation on social media. Am J Public Health 2020;110(S3):S273-S275. [doi: 10.2105/AJPH.2020.305905] [Medline: 33001722]
20. Lotto M, Jorge O, Machado MDA, Cruvinel T. Exploring online oral health misinformation: a content analysis. Braz Oral Res 2023;37:e049 [FREE Full text] [doi: 10.1590/1807-3107bor-2023.vol37.0049] [Medline: 37255069]
21. Bidmon S, Terlutter R. Gender differences in searching for health information on the internet and the virtual patient-physician relationship in Germany: exploratory results on how men and women differ and why. J Med Internet Res 2015;17(6):e156 [FREE Full text] [doi: 10.2196/jmir.4127] [Medline: 26099325]
22. Warner-Søderholm G, Bertsch A, Sawe E, Lee D, Wolfe T, Meyer J, et al. Who trusts social media? Comput Hum Behav 2018;81:303-315. [doi: 10.1016/j.chb.2017.12.026]
23. Stellefson M, Hanik B, Chaney B, Chaney D, Tennant B, Chavarria EA. eHealth literacy among college students: a systematic review with implications for eHealth education. J Med Internet Res 2011;13(4):e102 [FREE Full text] [doi: 10.2196/jmir.1703] [Medline: 22155629]
24. Pop LM, Iorga M, ipo LR, Iurcov R. Gender differences in healthy lifestyle, body consciousness, and the use of social networks among medical students. Medicina (Kaunas) 2021;57(7):648 [FREE Full text] [doi: 10.3390/medicina57070648] [Medline: 34202497]
25. Ruiz M, Kabani F, Cotter J. A review of the effects of oral health media hype on clients' perception of treatment. Can J Dent Hyg 2022;56(1):31-38 [FREE Full text] [Medline: 35401762]
26. Ajmal M, Assiri K, Al-Ahmari B. Oral hygiene myths and its association with gingival health status among patients in Aseer region of Saudi Arabia: a cross-sectional study. J Contemp Dent Pract 2021;22(5):506-510. [doi: 10.5005/jp-journals-10024-3062]
27. Al-Harthi MM, Holenarasipur AV, Al-Harthi HI, Al-Thobaiti FA, Al-Otaibi HA, Al-Dawood MM. Prevalence of myths regarding oral health, among patients attending a dental institute: a study from Taif, Saudi Arabia. JCDR 2019;13(9). [doi: 10.7860/jcdr/2019/41907.13125]
28. Gowdar IM, Alqahtani AM, Asiri AM, Aldossary SF, Alkhurayef IA, Alheneshi DI. Oral health myths among general population at Riyadh region, Saudi Arabia. J Pharm Bioallied Sci 2021;13(Suppl 1):S241-S245 [FREE Full text] [doi: 10.4103/jpbs.JPBS_700_20] [Medline: 34447085]
29. Duangthip D, Chu CH. Challenges in oral hygiene and oral health policy. Front Oral Health 2020;1:575428 [FREE Full text] [doi: 10.3389/froh.2020.575428] [Medline: 35047981]
30. Population estimates by gender, nationality, and region. General Authority for Statistics (GASTAT). 2022. URL: https://database.stats.gov.sa/home/indicator/535 [accessed 2024-11-20]
31. Manual of statistical definitions, concepts, and terms. General Authority for Statistics (GASTAT). 2024. URL: https://www.stats.gov.sa/sites/default/files/statistical_definitions_manual_english_2.pdf [accessed 2024-11-20]
32. Alhaddad MS. The use of social media among Saudi residents for medicines related information. Saudi Pharm J 2018;26(8):1106-1111 [FREE Full text] [doi: 10.1016/j.jsps.2018.05.021] [Medline: 30510470]
33. Almozainy M. Assessing the use of social media as a source of information related to dentistry in Saudi Arabia. JDHODT 2017;8(7):663-668. [doi: 10.15406/jdhodt.2017.08.00312]
34. Baik KM, Anbar G, Alshaikh A, Banjar A. Effect of social media on patient's perception of dental aesthetics in Saudi Arabia. Int J Dent 2022;2022:4794497 [FREE Full text] [doi: 10.1155/2022/4794497] [Medline: 35265132]
35. Barrera-Verdugo G, Villarroel-Villarroel A. Evaluating the relationship between social media use frequency and entrepreneurial perceptions and attitudes among students. Heliyon 2022;8(4):e09214 [FREE Full text] [doi: 10.1016/j.heliyon.2022.e09214] [Medline: 35434392]
36. Althunayan A, Alsalhi R, Elmoazen R. Role of social media in dental health promotion and bahavior change in qassim province, Saudi Arabia. Int J Med Health Res 2018;4:98-103 [FREE Full text]
37. ALrukban A. The health related uses of social media among users in Saudi Arabia. Int J Med Sci Public Health 2014;3(12):1492-1497. [doi: 10.5455/ijmsph.2014.170920141]
38. Lawal FB. Global self-rating of oral health as summary tool for oral health evaluation in low-resource settings. J Int Soc Prev Community Dent 2015;5(Suppl 1):S1-S6 [FREE Full text] [doi: 10.4103/2231-0762.156516] [Medline: 25984461]
39. Population by age groups and gender, mid-year 2020. General Authority for Statistics (GASTAT). 2020. URL: https://www.stats.gov.sa/sites/default/files/Population%20by%20Age%20Groups%20%2Cand%20Gender_0.pdf [accessed 2024-11-20]
40. National qualifications framework 2023: the national qualifications framework in the kingdom of Saudi Arabia. Education and Training Evaluation Commission (ETEC). 2023. URL: https://www.etec.gov.sa [accessed 2024-11-20]
41. Butani Y, Weintraub JA, Barker JC. Oral health-related cultural beliefs for four racial/ethnic groups: assessment of the literature. BMC Oral Health 2008;8:26 [FREE Full text] [doi: 10.1186/1472-6831-8-26] [Medline: 18793438]
42. El Kheir DYM, Boumarah DN, Bukhamseen FM, Masoudi JH, Boubshait LA. The Saudi experience of health-related social media use: a scoping review. Saudi J Health Syst Res 2021;1(3):1-12. [doi: 10.1159/000516473]

XSL•FO
RenderX

43. Lartey S. The future influence of TikTok on evolving social media trends: a comprehensive analysis. 2024. URL: https://tinyurl.com/5derawx6 [accessed 2024-12-12]
44. Gesmundo MAG, Jordan MDS, Meridor WHD, Muyot DV, Castano MCN, Bandojo AJP. TikTok as a platform for marketing campaigns: the effect of brand awareness and brand recall on the purchase intentions of millennials. JBMS 2022;4(2):343-361. [doi: 10.32996/jbms.2022.4.2.27]
45. Malik A, Khan ML, Quan-Haase A. Public health agencies outreach through instagram during the COVID-19 pandemic: crisis and emergency risk communication perspective. Int J Disaster Risk Reduct 2021;61:102346 [FREE Full text] [doi: 10.1016/j.ijdrr.2021.102346] [Medline: 36337987]
46. Alhabash S, Ma M. A tale of four platforms: motivations and uses of Facebook, Twitter, Instagram, and Snapchat among college students? Social Media + Society 2017;3(1):205630511769154. [doi: 10.1177/2056305117691544]
47. Sattora EA, Ganeles BC, Pierce ME, Wong R. Research on health topics communicated through TikTok: a systematic review of the literature. Journalism Media 2024;5(3):1395-1412. [doi: 10.3390/journalmedia5030088]
48. Basch CH, Hillyer GC, Jaime C. COVID-19 on TikTok: harnessing an emerging social media platform to convey important public health messages. Int J Adolesc Med Health 2022;34(5):367-369 [FREE Full text] [doi: 10.1515/ijamh-2020-0111] [Medline: 32776899]
49. Saudi census 2022: population summary report. General Authority for Statistics (GASTAT). URL: https://portal.saudicensus.sa/static-assets/media/content/20230531_GASTAT_Population_Report.pdf?crafterSite=gastat-portal [accessed 2024-11-19]
50. Maher CA, Lewis LK, Ferrar K, Marshall S, De Bourdeaudhuij I, Vandelanotte C. Are health behavior change interventions that use online social networks effective? a systematic review. J Med Internet Res 2014;16(2):e40 [FREE Full text] [doi: 10.2196/jmir.2952] [Medline: 24550083]
51. Ryan J, Lopian L, Le B, Edney S, Van Kessel G, Plotnikoff R, et al. It's not raining men: a mixed-methods study investigating methods of improving male recruitment to health behaviour research. BMC Public Health 2019;19(1):814 [FREE Full text] [doi: 10.1186/s12889-019-7087-4] [Medline: 31234825]
52. Asogwa CE, Okeke SV, Gever VC, Ezeah G. Gender disparities in the influence of social media advertisements on buying decision in Nigeria. Communicatio 2020;46(3):87-105 [FREE Full text] [doi: 10.1080/02500167.2020.1825100]
53. Courtenay WH. Key determinants of the health and well-being of men and boys. Int J Men's Health 2003;2(1):1-30. [doi: 10.3149/jmh.0201.1]
54. Addis ME, Mahalik JR. Men, masculinity, and the contexts of help seeking. Am Psychol 2003;58(1):5-14. [doi: 10.1037/0003-066x.58.1.5] [Medline: 12674814]
55. George A, Johnson M, Blinkhorn A, Ajwani S, Bhole S, Yeo AE, et al. The oral health status, practices and knowledge of pregnant women in South-Western Sydney. Aust Dent J 2013;58(1):26-33 [FREE Full text] [doi: 10.1111/adj.12024] [Medline: 23441789]
56. Anil S, Anand PS. Early childhood caries: prevalence, risk factors, and prevention. Front Pediatr 2017;5:157 [FREE Full text] [doi: 10.3389/fped.2017.00157] [Medline: 28770188]
57. Addy M, Hunter ML. Can tooth brushing damage your health? Effects on oral and dental tissues. Int Dent J 2003;53 Suppl 3:177-186. [doi: 10.1111/j.1875-595x.2003.tb00768.x] [Medline: 12875306]
58. Hujoel PP, Cunha-Cruz J, Banting DW, Loesche WJ. Dental flossing and interproximal caries: a systematic review. J Dent Res 2006;85(4):298-305. [doi: 10.1177/154405910608500404] [Medline: 16567548]
59. Puri N, Coomes EA, Haghbayan H, Gunaratne K. Social media and vaccine hesitancy: new updates for the era of COVID-19 and globalized infectious diseases. Hum Vaccines Immunother 2020;16(11):2586-2593 [FREE Full text] [doi: 10.1080/21645515.2020.1780846] [Medline: 32693678]
60. Del Vicario M, Bessi A, Zollo F, Petroni F, Scala A, Caldarelli G, et al. The spreading of misinformation online. Proc Natl Acad Sci U S A 2016;113(3):554-559 [FREE Full text] [doi: 10.1073/pnas.1517441113] [Medline: 26729863]
61. Getman R, Helmi M, Roberts H, Yansane A, Cutler D, Seymour B. Vaccine hesitancy and online information: the influence of digital networks. Health Educ Behav 2018;45(4):599-606. [doi: 10.1177/1090198117739673] [Medline: 29267129]
62. Oehler RL. On measles, vaccination, social media activism, and how to win back our role as our patients' best advocates. Clin Infect Dis 2020;70(2):338-340. [doi: 10.1093/cid/ciz656] [Medline: 31309980]
63. Chen Q, Min C, Zhang W, Wang G, Ma X, Evans R. Unpacking the black box: how to promote citizen engagement through government social media during the COVID-19 crisis. Comput Human Behav 2020;110:106380 [FREE Full text] [doi: 10.1016/j.chb.2020.106380] [Medline: 32292239]
64. Basch CE, Basch CH, Hillyer GC, Jaime C. The role of YouTube and the entertainment industry in saving lives by educating and mobilizing the public to adopt behaviors for community mitigation of COVID-19: successive sampling design study. JMIR Public Health Surveill 2020;6(2):e19145 [FREE Full text] [doi: 10.2196/19145] [Medline: 32297593]
65. D'Souza RS, D'Souza S, Strand N, Anderson A, Vogt M, Olatoye O. YouTube as a source of medical information on the novel coronavirus 2019 disease (COVID-19) pandemic. Glob Public Health 2020;15(7):935-942 [FREE Full text] [doi: 10.1080/17441692.2020.1761426] [Medline: 32397870]

XSL•FO
RenderX

66. Adebesin F, Smuts H, Mawela T, Maramba G, Hattingh M. The role of social media in health misinformation and disinformation during the COVID-19 pandemic: bibliometric analysis. JMIR Infodemiol 2023;3:e48620 [FREE Full text] [doi: 10.2196/48620] [Medline: 37728981]
67. The EU's digital services act. European Commission. URL: https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en [accessed 2022-10-27]

Abbreviations

AMR: adjusted mean ratio
AOR: adjusted odds ratio
GASTST: General Authority of Statistics

Edited by A Mavragani; submitted 14.12.24; peer-reviewed by R Musa, A Noor; comments to author 07.01.25; revised version received 20.01.25; accepted 22.01.25; published 10.02.25.

Please cite as:
BinHamdan RH, Alsadhan SA, Gazzaz AZ, AlJameel AH
Social Media Use and Oral Health–Related Misconceptions in Saudi Arabia: Cross-Sectional Study
JMIR Form Res 2025;9:e70071
URL: https://formative.jmir.org/2025/1/e70071
doi:10.2196/70071
PMID:

©Rahaf Hamdan BinHamdan, Salwa Abdulrahman Alsadhan, Arwa Zohair Gazzaz, AlBandary Hassan AlJameel. Originally published in JMIR Formative Research (https://formative.jmir.org), 10.02.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

mHealth App to Promote Healthy Lifestyles for Diverse Families Living in Rural Areas: Usability Study

Alejandra Perez Ramirez[1], MPH; Adrian Ortega[2], PhD; Natalie Stephenson[3], PhD; Angel Muñoz Osorio[1], BS; Anne Kazak[1,4], PhD; Thao-Ly Phan[1,4], MPH, MD

[1]Center For Healthcare Delivery Science, Nemours Children's Health, Wilmington, DE, United States
[2]Center for Behavior Intervention Technologies, Northwestern Feinberg School of Medicine, Northwestern University, Evanston, IL, United States
[3]Center for Health Delivery Innovation, Nemours Children's Health, Wilmington, DE, United States
[4]Department of Pediatrics, Sidney Kimmel Medical College, Thomas Jefferson University, Philadelphia, PA, United States

Corresponding Author:
Alejandra Perez Ramirez, MPH
Center For Healthcare Delivery Science
Nemours Children's Health
1600 Rockland Road
Wilmington, DE, 19803
United States
Phone: 1 302 358 5968
Email: alejandra.perezramirez@nemours.org

Abstract

Background: Mobile Integrated Care for Childhood Obesity is a multicomponent intervention for caregivers of young children with obesity from rural communities that was developed in collaboration with community, parent, and health care partners. It includes community programming to promote healthy lifestyles and address social needs and health care visits with an interdisciplinary team. A digital mobile health platform—the Healthy Lifestyle (Nemours Children's Health) dashboard—was designed as a self-management tool for caregivers to use as part of Mobile Integrated Care for Childhood Obesity.

Objective: This study aimed to improve the usability of the English and Spanish language versions of the Healthy Lifestyle dashboard.

Methods: During a 3-phased approach, usability testing was conducted with a diverse group of parents. In total, 7 mothers of children with obesity from rural communities (average age 39, SD 4.9 years; 4 Spanish-speaking and 3 English-speaking) provided feedback on a prototype of the dashboard. Participants verbalized their thoughts while using the prototype to complete 4 tasks. Preferences on the dashboard icon and resource page layout were also collected. Testing was done until feedback reached saturation and no additional substantive changes were suggested. Qualitative and quantitative data regarding usability, acceptability, and understandability were analyzed.

Results: The dashboard was noted to be acceptable by 100% (N=7) of the participants. Overall, participants found the dashboard easy to navigate and found the resources, notifications, and ability to communicate with the health care team to be especially helpful. However, all (N=4) of the Spanish-speaking participants identified challenges related to numeracy (eg, difficulty interpreting the growth chart) and literacy (eg, features not fully available in Spanish), which informed iterative refinements to make the dashboard clearer and more literacy-sensitive. All 7 participants (100%) selected the same dashboard icon and 71% (5/7) preferred the final resource page layout.

Conclusions: Conducting usability testing with key demographic populations, especially Spanish-speaking populations, was important to developing a mobile health intervention that is user-friendly, culturally relevant, and literacy-sensitive.

(JMIR Form Res 2025;9:e60495) doi:10.2196/60495

KEYWORDS

obesity; user testing; mHealth; mobile health; Spanish; child; rural population

XSL•FO
RenderX

Introduction

Childhood obesity persists as a significant public health concern in the United States [1,2]. Notably, 1 in every 5 children in the United States lives with obesity. Obesity in childhood is linked to increased risks of comorbid and future health conditions such as diabetes, cardiovascular diseases, cancer, and psychosocial concerns, leading to increased health care costs and utilization [1,3]. This is especially the case for children from communities of color, communities with fewer resources, and rural communities, who continue to have the highest rates of obesity [2]. Challenges in affording and accessing nutritious food, accessing safe spaces for physical activity, having transportation to get to health care services, and managing family stressors are a few of the many systemic barriers that perpetuate disparities in childhood obesity [4,5]. While lifestyle behavior interventions are first-line treatments for improving health behaviors and obesity outcomes, most available interventions do not address these systemic barriers and are often not effective when implemented in the communities with the greatest need [6,7]. Thus, more studies are needed that consider the needs and preferences of diverse patient families when designing an intervention.

Digital interventions for the treatment of pediatric obesity can increase the convenience and access to treatment through the removal of common barriers such as financial, travel, and time constraints [8,9]. A recent systematic review and meta-analysis revealed that digital interventions for the treatment of pediatric obesity have a significant yet small effect on BMI *z* scores [10]. Still, there are important limitations to digital interventions that require consideration, especially regarding their uptake and implementation. For one, digital interventions are subject to high nonusage attrition such as low uptake and disengagement [11,12]. Although some digital interventions for the treatment of pediatric obesity have included diverse samples in their studies, there remain concerns about their cultural sensitivity [10]. For example, most of these interventions do not include content related to systemic barriers, social drivers of health, or community resources [13]. Rurality or place of residence is also not often reported or considered when implementing digital interventions [14]. In addition, there is only 1 published study of a digital intervention for pediatric obesity available in Spanish, which is a major gap given the prevalence of Hispanic or Latino children with overweight or obesity in the United States [5,15,16]. Therefore, there is a need to incorporate human-centered design principles into the design of digital interventions, such as partnering with target users to enhance the usability of these interventions. This way, digital intervention designers can maximize engagement by building an intervention that centers on the needs and preferences of its users. Key principles of human-centered design include but are not limited to the active involvement of users with lived experiences relevant to the intervention in design activities as well as iteratively refining and updating intervention features based on user feedback [17].

Drawing on the advantages of using a digital intervention, a digital mobile health (mHealth) platform was included as part of a larger, multicomponent intervention called Mobile Integrated Care for Childhood Obesity (MICCO). MICCO was designed for caregivers of children with obesity from rural communities (Figure 1). MICCO combines evidence-based interdisciplinary obesity treatment, monthly community-based programming for families to engage in physical activity and have a healthy affordable meal together, and community health worker outreach to address psychosocial needs and barriers to achieving lifestyle goals. Importantly, MICCO was developed with the input of a Council of Parent Partners, which included 6 parents of children with obesity from rural communities, and a Council of Community Partners, which included 5 primary care providers from clinics serving rural communities and 4 community organizations serving rural communities. These Councils met routinely by videoconference throughout the entirety of the study to provide valuable feedback on the development and implementation of the MICCO intervention, drawing on their knowledge about the needs of families of children with obesity and resources in the community to support families.

Figure 1. Mobile Integrated Care for Childhood Obesity (MICCO) care model, developed by a research team for families of young children with obesity.

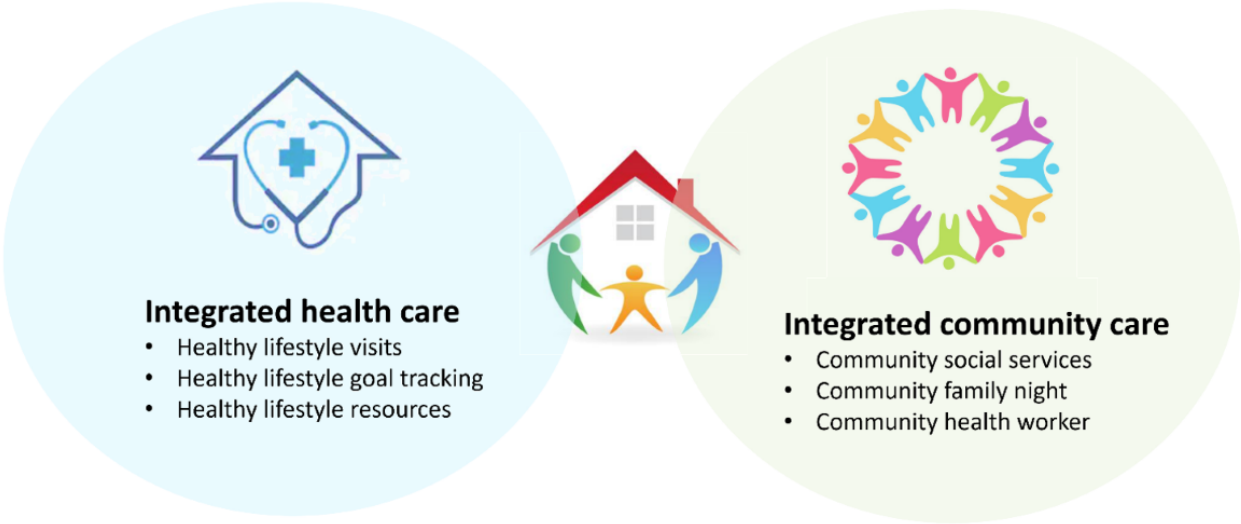

XSL•FO
RenderX

As an additional component of MICCO, a custom Healthy Lifestyle dashboard on a mobile app linked to the Nemours Children's Health electronic health record (EHR; Nemours App [Nemours Children's Health]) was built, with input from interdisciplinary specialists from the Nemours Healthy Weight and Wellness Clinic and Council of Community Partners and in collaboration with app designers from the Nemours Center for Health Delivery Innovation. Psychoeducational tools based on social cognitive theory [18], behavior change techniques [19], and the socioecological model of obesity [20] were included on the dashboard, including a self-monitoring function to track daily progress toward lifestyle goals formed during treatment visits and culturally relevant videos to promote positive parenting and healthy lifestyle behaviors. The video resources, including 6 that were created specifically for the MICCO intervention, provided content related to mood and behavior, physical activity, and healthy eating; were available in English and Spanish; and featured families and health care providers of diverse backgrounds. Through the dashboard, caregivers can also track their child's weight using an integrated Bluetooth scale, send messages to their health care team, and connect to telehealth appointments. The dashboard also included links to community programs (including healthy lifestyle and social need resources) and monthly reminders about the monthly community-based program. Since many of the patient families with obesity served by Nemours Children's Health are Hispanic or Latino, the dashboard was developed in English and Spanish. A parallel dashboard was created for health care providers to access data entered by patient families about their goals and weight progress to assist with clinical decision support.

User testing was conducted to test and refine the MICCO Healthy Lifestyle dashboard to enhance the usability of the intervention for families of diverse backgrounds. No hypotheses were made a priori.

Methods

Overview

Parents of children with obesity living in rural communities provided feedback on the usability of the Healthy Lifestyle dashboard through Think Aloud testing. Think Aloud testing is a research method applied to an iterative human-centered design process to discover usability problems with a product, incorporate feedback, revise the product, and retest to validate the design. This evidence-based method for obtaining in vivo feedback has been used in the development of digital health interventions [21]. In a typical usability study, the most common usability problems are uncovered after testing with 5 participants [22].

Sample and Recruitment

Target users of the MICCO intervention are legal guardians of children ages 4-12 years with obesity (BMI percentile ≥95) receiving primary care from 1 of 3 Nemours Children's Health practices in Delaware serving rural communities. Rurality was determined based on a score of 4 or greater on the US Department of Agriculture 2010 Rural-Urban Commuting Area Codes [23]. Therefore, we recruited participants matching these characteristics for user testing. Legal guardians also had to be able to read English or Spanish and have access to the internet. A primary aim of the study was to incorporate diverse perspectives in refining the Healthy Lifestyle dashboard; therefore, purposeful sampling was used to recruit a diverse group of parents including Spanish-speaking parents of Hispanic or Latino ethnicity and English-speaking parents of Black and White race. A list of potentially eligible participants was pulled from the EHR of Nemours Children's Health, which is a large pediatric health care system with 12 primary care clinics in Delaware. A total of 196 parents (63 Spanish-speaking and 133 English-speaking) from this EHR list received an SMS text message informing them of the study. In total, 34 parents (15 Spanish-speaking and 19 English-speaking) responded to the text. A research coordinator then followed up with a call to discuss the study and procedures with each parent. Of these, 7 agreed to participate (4 Spanish-speaking and 3 English-speaking), 2 declined after hearing about the study, and the remainder did not respond to any additional contact by the study team. For those who agreed to participate, the research coordinator scheduled a Think Aloud session. Consent was collected at the time of the session before any data collection (refer to Ethical Considerations for details regarding the consent process).

Data Collection

Think Aloud testing was conducted via video calls through a secure Zoom (Zoom Communications) platform, accessible to participants either on their mobile device or computer. A web link to an interactive, high-fidelity prototype of the Healthy Lifestyle dashboard was shared with participants. The online prototype mirrored the app and allowed participants to easily access the design for Think Aloud testing. The sessions were done through a video call, similar to other studies [24], to increase the convenience for parents of young children and to simulate the environmental context within which the actual mobile app would be experienced. It also allowed for screen recording so the research coordinator could see which parts were easy or hard to navigate.

The testing sessions were video-recorded, and the audio was transcribed. Spanish language recordings were translated and transcribed into English. Rarely, participants encountered minor technical issues with Zoom, such as slow connectivity or difficulty using Zoom (eg, how to join the call or share their screen), but the bilingual research coordinator on the call was able to guide them through these issues.

Testing sessions were conducted using a semistructured interview guide (Multimedia Appendix 1) to facilitate the completion of tasks. Introductory questions were used to establish a connection with the participant before user testing of the dashboard. During the testing portion of the interview, participants were introduced to the dashboard at the time of testing so that first impressions could be observed. This also allowed for genuine reactions as they became familiar with the platform and navigated the tools. Participants were then asked to complete four major tasks, as laid out by the interview guide: (1) enter daily goal completion and review their weekly progress, (2) enter the child's height and weight and review a growth curve, (3) contact their health care team, and (4) find

XSL•FO
RenderX

resources for their family. Families were also asked about specific design layouts using preference testing (ie, participants were shown 2 design options and asked to pick the one they preferred). Throughout the testing, participants were encouraged to Think Aloud or verbalize their thoughts as they navigated independently through the tasks. The participants shared their screens with the research coordinator, which allowed the coordinator to see their navigation when using the app. Immediately following the testing portion, participants were asked closing questions to collect additional general feedback about the dashboard. Each session lasted 30 to 60 minutes.

A 3-phased approach was used to refine the dashboard and validate that iterative changes were acceptable. In phase 1, Think Aloud sessions with 4 Spanish-speaking families were conducted until no new feedback was received. The research and Nemours app team met to discuss feedback, and rapid changes were made to the user interface based on participant feedback. In phase 2, Think Aloud sessions with 3 English-speaking families were completed with the revised prototype until no substantive changes were recommended. In phase 3, a final prototype based on all of the Think Aloud feedback was shown in a video call to the larger study's Council of Parent Partners, which was started after user testing was complete. The Council of Parent Partners included 4 parents from phase 1, one parent from phase 2, and one parent who did not participate in Think Aloud testing. This meeting was not formal user testing. Instead, it presented an opportunity to obtain final approval from parents, including many who had participated as individuals in the user testing. Meetings were facilitated by a bilingual research coordinator, with whom the parents had developed a rapport, and participants received an additional US $50 gift card for participating. Parents were advised during the group meeting that they could also provide feedback on the final prototype privately if desired.

Data Analysis

All Think Aloud transcripts were uploaded into Dedoose [25], a qualitative data analysis software, to complete a formal qualitative analysis. A preliminary codebook of a priori codes was established based on the interview guide. The goal of the codebook was to capture the usability, understandability, and acceptability of the dashboard and its specific features during the Think Aloud sessions (Multimedia Appendix 2). Furthermore, 3 members of the research team (TLP, APR, and AO) met regularly to review the transcripts and codebook. The first transcript was coded by LP and APR together to test the codebook. All 3 members then independently coded 2 more transcripts. The team met to review discrepancies and revise the codebook as needed. The fourth transcript was used to test reliability among the coders and the established codebook. Given the strong intercoder reliability (83.5%), the codebook was finalized, and then all the transcripts were coded with the final codebook by APR and TLP. They met regularly to resolve any discrepancies. After all transcripts were coded, feedback for each of the 4 specific tasks and the overall app was pulled. The research team noted specific challenges and features highlighted by participants. This feedback was analyzed for each phase given that there were changes in-between phases. Quantitative data regarding the number of participants who reported overall acceptability of the dashboard and usability and understandability of each task were also extracted by coding specific mentions and counting the frequency of the codes.

Ethical Considerations

Considerations to protect participants were taken into account when designing study procedures. Before recruitment, the Nemours IRB (#1812070) approved all study procedures. During recruitment, families were given the option to decline participation or opt out of further communication. Interested participants were asked to complete a consent form using an e-consent process. At a convenient time for the participant, the research coordinator trained in collecting informed consent, called the participant on the phone to review the full consent form. The consent form listed the study procedures and risks and benefits involved with the study. Participants were reminded that their participation was voluntary and that they could ask to withdraw at any time. Spanish-speaking participants were consented by a native Spanish-speaking research coordinator and with an approved Spanish-language form. Study procedures, such as having participants turn their cameras off at the time of the Think Aloud session and removing all identifying information from transcripts, were followed to protect the participants' identities. All participants received a US $50 gift card after completing the Think Aloud session.

Results

In total, 7 mothers (average age 39, SD 4.9 years) participated in Think Aloud sessions (4 in phase 1 and 3 in phase 2) and 6 mothers (average age 42, SD 4.9 years, 5 of whom participated in Think Aloud testing) participated in the Council of Parent Partners meeting (phase 3, Table 1). No additional participants were recruited given that user feedback did not identify any new usability problems after 2 rounds of testing. These sample sizes are in line with a typical usability study [22].

Overall, MICCO was accepted and liked by 100% (N=7) of the participants. Participants shared that the Healthy Lifestyle dashboard would be an easy-to-use tool to help them develop healthier habits as a family. As 1 participant described:

> *It is a tool that is within my reach. I can contact my son's doctor to know what he can or cannot eat, and not have to make an appointment to go to the hospital. I can use the application to solve the doubts I have about some things that my son can or cannot get. For me, it would be an easy tool to handle as a mom.*

While it was generally well-received, participants shared some challenges with specific tasks and highlighted features that were important to them, as described further in this study.

XSL•FO
RenderX

Table 1. Demographic information of caregivers of children with obesity from rural clinics in Delaware, participating in 3 iterative phases of Think Aloud Testing of a mobile dashboard to promote healthy lifestyles.

Characteristics	Phase 1 (n=4)	Phase 2 (n=3)	Phase 3[a] (n=6)
Parent age in years, mean (SD)	38.8 (3.0)	40.3 (7.6)	42.0 (4.9)
Parent sex (female), n (%)	4 (100)	3 (100)	6 (100)
Race, n (%)			
White	2 (50)	1 (33)	2 (33)
Black	__[b]	—	—
Other[c]	2 (50)	1 (33)	3 (50)
More than 1 race	—	1 (33)	1 (17)
Ethnicity (Hispanic or Latino), n (%)	4 (100)	2 (67)	6 (100)
Insurance status (public)[c], n (%)	4 (100)	2 (100)[d]	5 (83)

[a]Five participants also participated in phases 1 and 2.

[b]Not applicable.

[c]"Other" was not specified by participants.

[d]Missing n=1; public insurance refers to a program run by the government to pay for some or all health care costs.

Phase 1

Overview

Spanish-speaking participants reported challenges surrounding understandability and usability of certain features of the Healthy Lifestyle dashboard. For example, 3 participants did not understand the lifestyle goal progress feature, 4 did not understand the growth curve feature, and 3 had trouble using the tool to communicate with the health care team. Themes surrounding numeracy and literacy challenges (Table 2) emerged and specific changes were made to the dashboard to ease these challenges after phase 1 (Figures 2 and 3, respectively).

Table 2. Challenges and solutions identified in Think Aloud testing of a mobile dashboard to promote healthy lifestyles by Spanish-speaking caregivers of children with obesity from rural clinics in Delaware.

Area and challenge	Solution
Numeracy challenges	
Interpretation of weekly goal completion	Changed the percentage bar to thumbs-up icons and provided a count instead of percentage
Interpretation of child's growth chart	Changed growth chart to show weight change over time instead of BMI
Literacy challenges	
Some components of the Nemours app (ie, the message center) were not available in Spanish	A pop-up note was added to provide the user with instructions in Spanish and a note that the following page would only be available in English
Despite the dashboard being in Spanish, there were concerns about the reading level	Language was simplified throughout the app

XSL•FO
RenderX

Figure 2. Revisions to dashboard to address numeracy challenges identified by 4 Spanish-speaking caregivers of children with obesity from rural clinics in Delaware during phase 1 of Think Aloud testing of a mobile app to promote healthy lifestyles.

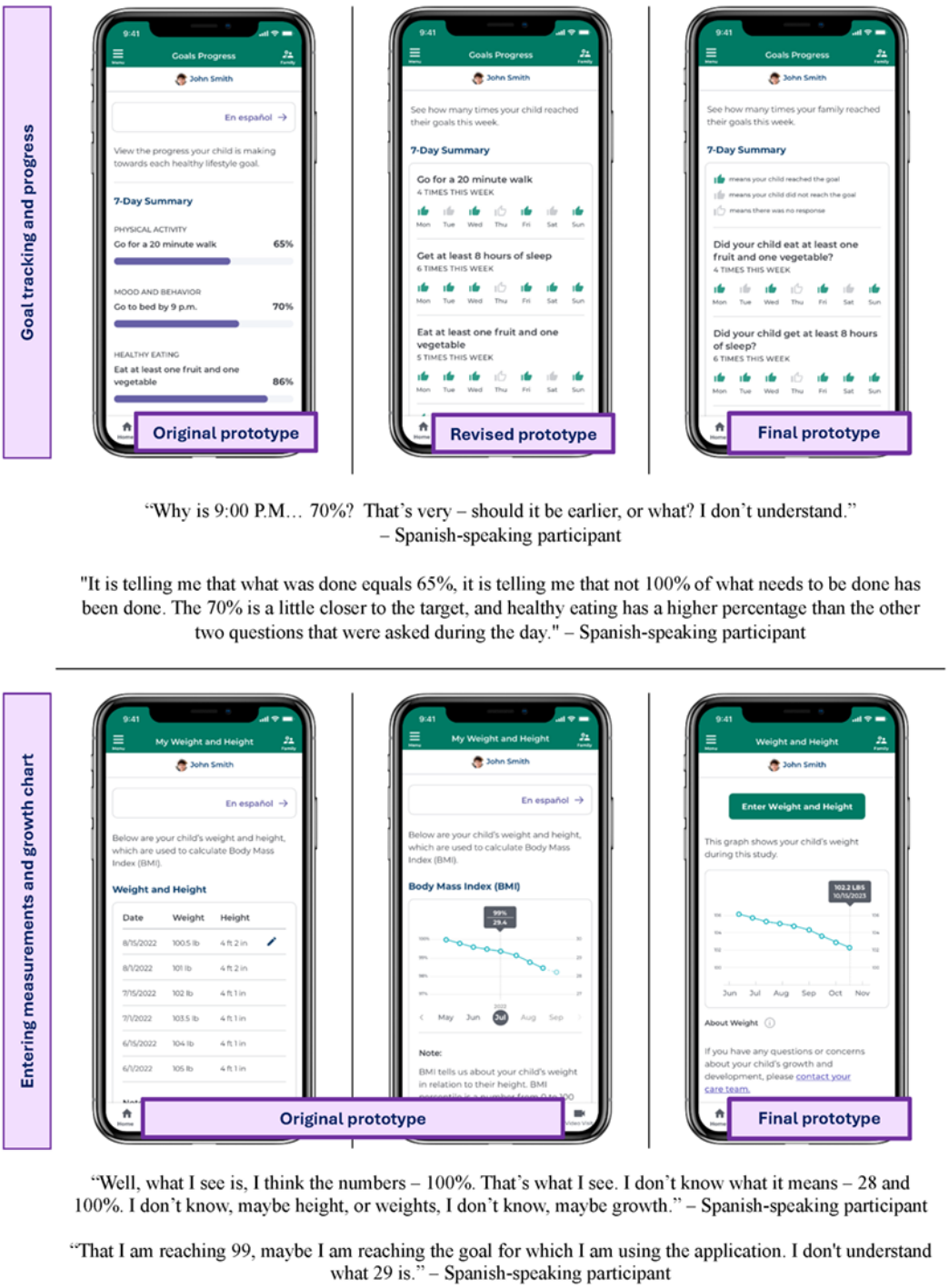

XSL•FO
RenderX

Figure 3. Revisions to address literacy challenges experienced during phase 1 Think Aloud testing of a mobile app to promote healthy lifestyles by Spanish-speaking caregivers of children with obesity from rural clinics in Delaware.

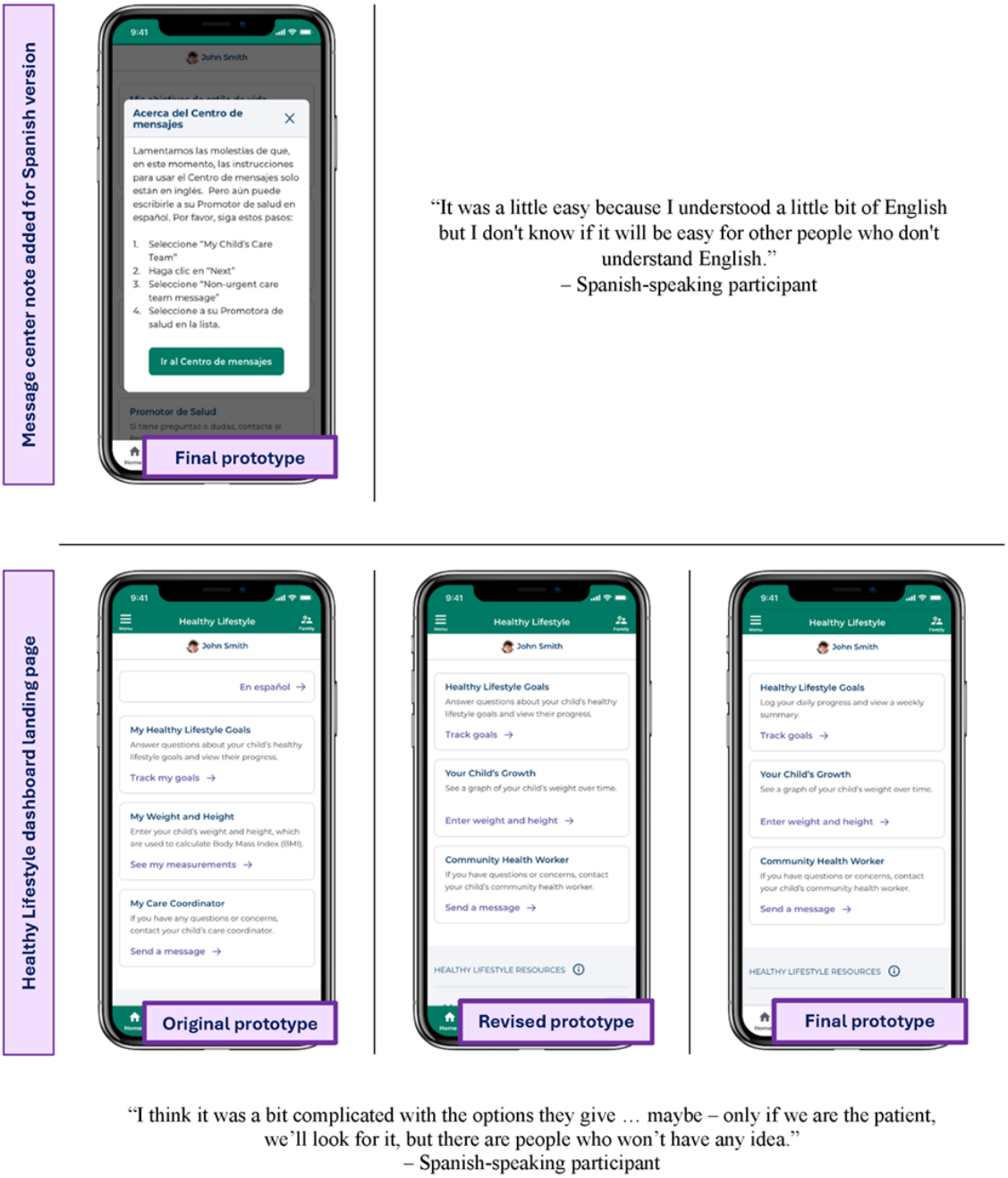

Numeracy Challenges

All 4 participants reported numeracy issues when reviewing the lifestyle goal progress and growth curve in the dashboard. In particular, some of the participants did not understand how to interpret the percentage of goals reached over the past 7 days and some interpreted it as a "grade" or "score." To simplify the weekly summary and ensure it would not be seen as punitive, the percentages were removed and substituted with a weekly view that used a green thumbs-up icon to indicate completion of the goal for that day, a grey thumbs-up if the goal was not completed, or a white thumbs-up if there was no data entered for that day. In the first iteration of the growth curve, participants were shown their child's BMI over time. Despite having a note that explained how BMI was calculated and what it meant, all 4 participants struggled with the interpretation of the growth curve. Given this, the growth curve was simplified to show weight instead of BMI and to have the entry of height and weight appear on the same page as the weight curve.

Literacy Challenges

While all of the content in the Healthy Lifestyle dashboard was available in Spanish, there were some features of the overall Nemours app that were only available in English at the time of

XSL•FO
RenderX

testing. This created some challenges. For example, participants were asked to send a message to their community health worker during Think Aloud testing. Since the Nemours app message center was only available in English, the language barrier caused some difficulty for participants. To provide families with more guidance when using the message center, a note in Spanish was added to the revised prototype to provide instructions on how to use the message center. In addition to this messaging task, 2 participants shared concerns about language barriers. As 1 participant stated:

> *For example, I have my mom … she definitely has trouble with technology. One, because she doesn't really work with technology, and secondly, because there is a language barrier. Even when things are in Spanish, she still feels uncomfortable.*

In recognition of these barriers, changes were made to simplify the language throughout the Healthy Lifestyle dashboard.

Phase 2

Overview

In phase 2, participants were shown a revised version of the Healthy Lifestyle prototype based on the feedback from phase 1. New changes were well-received by parents and 3 reported that the revised prototype was usable and understandable. Feedback during this phase highlighted the importance of certain features in the dashboard (Figure 4), with only minor suggestions provided.

XSL•FO
RenderX

Figure 4. Features that caregivers of children with obesity from rural clinics in Delaware considered important to include during Think Aloud testing of a mobile app to promote healthy lifestyles.

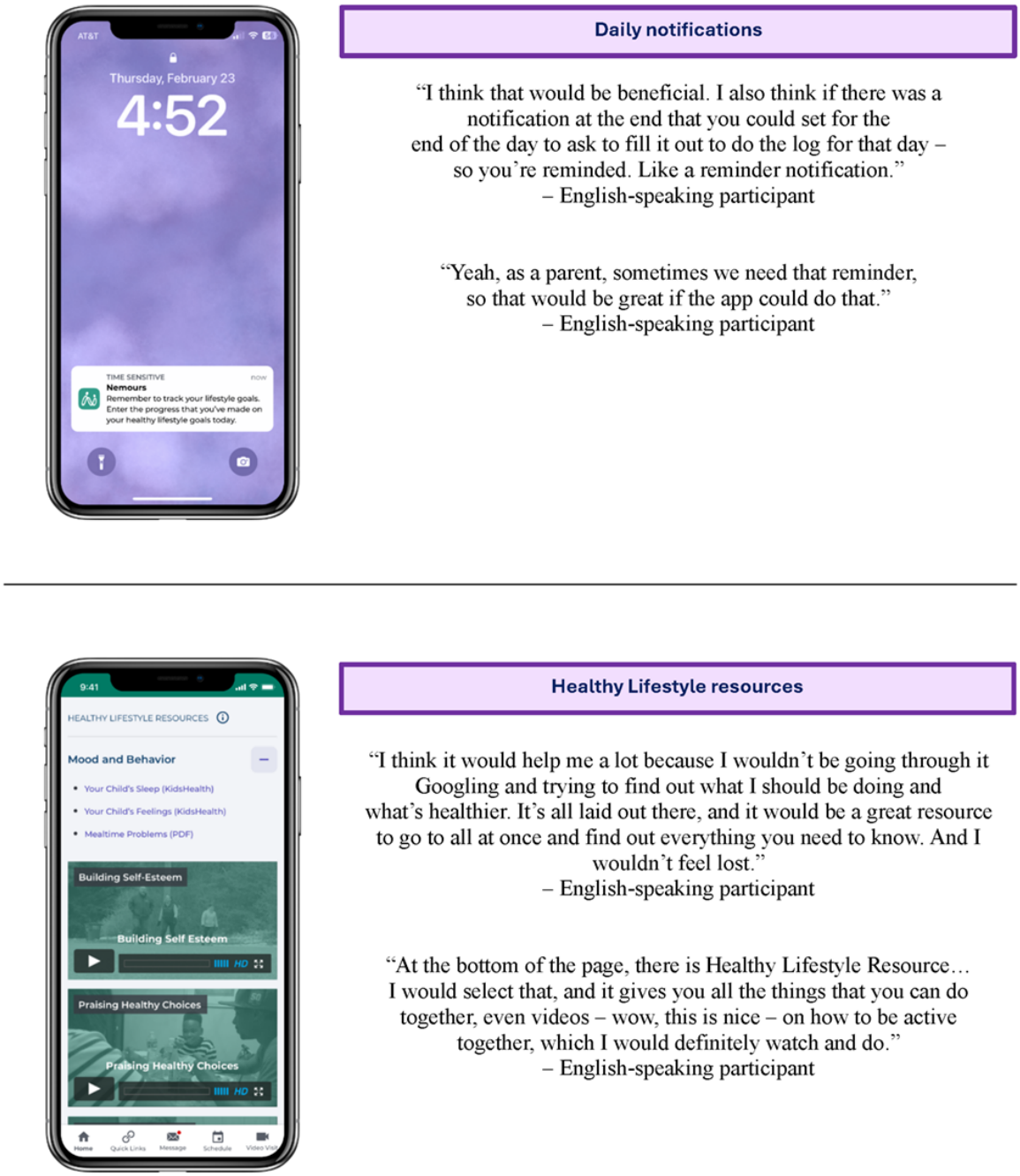

Important Features to Include

Participants liked many of the features already included in the Healthy Lifestyle dashboard and thought they would help motivate families. Participants found the multimedia resources (ie, videos, handouts, and links to external sources) embedded in the dashboard to be particularly useful. Participants especially liked the videos and thought they were a user-friendly way to view the resources. Having a set of curated resources in 1 centralized location was also an advantage because participants found it challenging to find reliable, evidence-based resources on their own. As 1 participant commented:

> *I would definitely promote all the things that app can help with – not just that it's about healthy living, healthy eating, and things like that, but also that it has resources for everything, because every family's different, and they deal with different things and different things within their children. There are great resources that you guys have and that you guys are offering.*

Finally, while participants liked the lifestyle goal feature and understood the importance of tracking their lifestyle goals, many participants talked about competing priorities that may get in the way of inputting their data. As a solution, they discussed

XSL•FO
RenderX

how an evening reminder to log their goals would help keep them on track.

Communication and Integration With Health Care Team

All participants discussed the importance of being able to communicate with their health care team through the dashboard, whether by sending messages or by having lifestyle goals and weight data available to their health care team. One participant highlighted how the transfer of the dashboard data to the child's care team could help keep families accountable for their goals:

> *I also think maybe to some parents, it would also be motivating to know – like if this is going to the provider, you want to show that as a parent, you're being proactive, and you're doing what you need to do for your child.*

Despite the appeal of communicating with their health care team through the app, participants expressed a need for more clarity about who would view their data and how and when the health care team would be communicating back to them, which was incorporated into the final prototype.

Preference Testing

Preference testing is a research method during which participants compare 2 variations of a design, choose 1, and explain their preference [21], and it was used in both phases 1 and 2. This method was used for selecting an icon to represent the Healthy Lifestyle dashboard and determining the preferred way to incorporate resources (Figure 5).

XSL•FO
RenderX

Figure 5. Results of preference testing done during Think Aloud testing of caregivers of children with obesity from rural clinics in Delaware to determine a dashboard icon and resource page layout for a mobile app to promote healthy lifestyles.

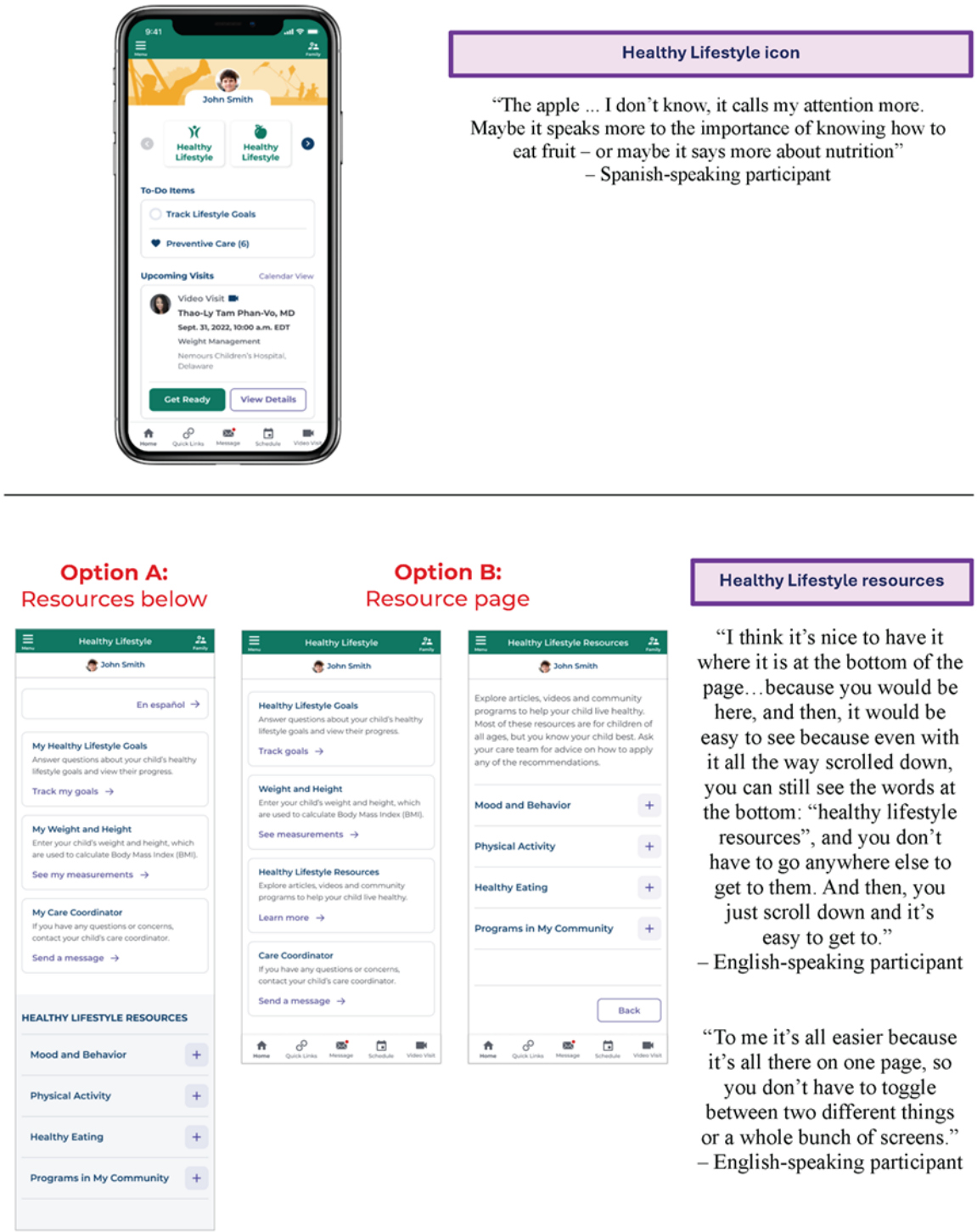

Healthy Lifestyle Dashboard Icon

Since the Healthy Lifestyle dashboard is incorporated into the Nemours app, which is a much broader app designed for a range of health care needs, an icon was needed to launch the dashboard. Designers on the Nemours app team explored a range of concepts for the icon. Symbols that represent different aspects of a healthy lifestyle, like physical activity and nutrition, were considered. The top 2 designs—a joyous person (left) and an apple (right)—were shared during phase 1 testing. Based on feedback from all the participants, the apple icon was chosen to be the branding element for the Healthy Lifestyle dashboard. Participants cited liking the apple more because they associated it with nutrition, which was the primary domain that participants associated with healthy lifestyles.

XSL•FO
RenderX

Resource Page Layout

The list of Healthy Lifestyle resources included 19 items across 4 domains (healthy eating, being active, mood and behavior, and community programs). Preference testing was used to understand how participants preferred to access the resources. In total, 2 variations of the Healthy Lifestyle resources were designed. Option A made the resources available at the bottom of each page (like a footer) with expandable sections. Whereas option B involved adding them to the dashboard as a fourth link, which would open a new page for this section. Participants compared the 2 options and then indicated their preference. Though 2 of the 7 participants thought option B looked better, the majority picked option A. They liked having all the resources easily accessible on every page instead of having to navigate back and forth through the pages.

Phase 3

All 6 parents (refer to Table 1 for demographic information) who participated in phase 3 were in agreement that the revisions to the dashboard addressed their concerns from phase 1 or 2. They reported that the final prototype was useful and easy to understand. Given no further feedback about the dashboard, it was finalized for use in future research studies.

Discussion

Principal Findings

MICCO is a comprehensive, multicomponent intervention that was developed in collaboration with community, parent, and health care partners to improve the relevance and accessibility of pediatric obesity interventions for families from underserved and rural communities. Digital self-management tools are a key component of the MICCO intervention and are made available to families through a mobile app dashboard integrated with the EHR. This dashboard component of MICCO addresses notable limitations in the mHealth for childhood obesity literature by including community resources, being available in Spanish, incorporating culturally relevant resources, and being integrated into health care delivery. Human-centered design methods were applied to ensure that the needs and preferences of families of diverse backgrounds were incorporated. The resulting dashboard is user-friendly, literacy-sensitive, and culturally responsive and incorporates features that will facilitate its usability in future research and clinical practice.

Through a human-centered approach to design, the research team was able to better understand the challenges that families, especially those with limited English proficiency and from rural neighborhoods, may experience in using an app to facilitate healthy lifestyle behaviors. Challenges navigating health information are not unique to this study and have been described among larger cohorts of Spanish-speaking individuals of Hispanic or Latino ethnicity [26,27]. Even when health information is available in Spanish, studies have found that reading levels are typically above a sixth-grade level which is higher than the recommended reading level for health information in any language [28,29]. Difficulty understanding health information can contribute to low rates of engagement with programs and services among Spanish-speaking populations [27,30,31]. Therefore, testing mHealth interventions with Spanish-speaking users is critical to ensuring that the intervention is understood and well-received by these populations.

In this study, numeracy and literacy barriers were identified during Think Aloud testing with Spanish-speaking participants. Therefore, pages depicting weight and goal progress were simplified to convey relevant health information more clearly. In addition, information was added to help Spanish-speaking families navigate the 1 component (messaging function) of the app that was not fully available in Spanish at the time of testing. Not only were these adaptations made as part of this research study, but findings from this line of research in conjunction with other Diversity, Equity, and Inclusion initiatives at Nemours Children's Health, prompted an overhaul of the entire Nemours App to be fully available in Spanish after testing.

Feedback received during usability testing also helped the team develop a deeper understanding of families' busy lifestyles. In particular, the majority of participants expressed challenges balancing competing priorities when taking care of young children. Not surprisingly, time constraints are a commonly cited barrier to adherence in childhood obesity treatment [32,33]. Because of this, many participants reported that an evening reminder about tracking lifestyle goals would be helpful, so a push notification was implemented to remind families to track their lifestyle goals in the app. Participants also liked the curated resources available through the dashboard, especially the brief videos. Having resources that were easily accessible added to the convenience of the app for families; in addition to automated reminders, this may be a way to keep families engaged in treatment.

Finally, participants endorsed an appreciation for the integration of the dashboard with their overall health care. For example, they reported a desire for information tracked through the dashboard to be shared with their health care team and wanted to know more about how to communicate with their health care team. The Healthy Lifestyle dashboard was designed to facilitate this bidirectional communication. This approach is different than most apps designed for childhood obesity, which are stand-alone. While there are advantages to this, findings from this study suggest that mHealth interventions may be more effective if integrated with health care delivery, especially for rural communities that may experience lower access to health care facilities. Not only can this enhance satisfaction and motivation on the part of families, but it can also inform clinical decision-making on the part of the health care team [34,35].

Limitations

There were some limitations to this study. While the recruitment strategy was broad and a diverse sample was recruited, it is possible that our sample was biased toward those with more digital literacy and broadband access. In the future, it will be important to evaluate the app across families from varying socioeconomic backgrounds and with those who face greater barriers to care. Using in-person testing may also allow for testing of patients with limited digital access and provide a better understanding of nonverbal cues that may not have been captured in the video recording. While our sample size was

XSL•FO
RenderX

small, it is consistent with best practices for usability testing but there may be limitations in our methods of obtaining final input through an Advisory Council (vs repeated formal usability testing). Also, while participants endorsed the acceptability of the dashboard, the study's focus on usability does not allow for an understanding of whether participants would engage with it on a routine basis. A future study might conduct field testing of the MICCO app, wherein participants can use the app continuously for a shortened duration (eg, a month) and provide acceptability and usability feedback before it is deployed in a longer trial. In addition, previous studies have identified that digital interventions are subject to low uptake and engagement levels; therefore, further research using implementation frameworks will be needed to better understand factors that will facilitate routine use of the dashboard as part of the larger intervention. Finally, while the dashboard (and intervention more broadly) was designed for caregivers of young children, future studies should consider refining and testing the dashboard for youth with obesity who may find a digital intervention even more acceptable.

Conclusion

Collaborating with community and patient partners and conducting usability testing with key demographic populations, especially Spanish-speaking populations, was important to developing an mHealth intervention that is user-friendly, culturally relevant, and literacy-sensitive to promote pediatric healthy lifestyles among diverse families. Designing an intervention that is usable for intended users and in intended settings is a critical first step to trialing and implementing a successful intervention.

Acknowledgments

We thank the National Institutes of Health/National Institute of General Medical Sciences (P20GM144270) for supporting this project through the Center of Biomedical Research Excellence for the REACH Center. We also want to thank the parents who generously gave us their time to participate in this study.

Data Availability

Deidentified data are available upon reasonable request. All requests should be directed to the corresponding author.

Authors' Contributions

APR collected and analyzed data and conceptualized, drafted, reviewed, and revised the manuscript. TLP designed the study, coordinated, and supervised data collection, assisted with data analysis, and conceptualized, drafted, reviewed, and revised the paper. AK designed the study and reviewed and revised the manuscript. AMO collected data and drafted, reviewed, and revised the manuscript. AO conducted data analyses and drafted, reviewed, and revised the manuscript. NS designed and refined the prototype and drafted, reviewed, and revised the manuscript. All authors approved the final manuscript as submitted and agree to be accountable for the work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Mobile Integrated Care for Childhood Obesity (MICCO) interview guide used to guide Think Aloud sessions with caregivers of young children with obesity.
[DOCX File , 33 KB - formative_v9i1e60495_app1.docx]

Multimedia Appendix 2
Mobile Integrated Care for Childhood Obesity (MICCO) – Think Aloud code book.
[DOCX File , 27 KB - formative_v9i1e60495_app2.docx]

References

1. Hales CM, Carroll MD, Fryar CD, Ogden CL. Prevalence of obesity among adults and youth: United States, 2015-2016. NCHS Data Brief 2017(288):1-8 [FREE Full text] [Medline: 29155689]
2. Ogden CL, Fryar CD, Hales CM, Carroll MD, Aoki Y, Freedman DS. Differences in obesity prevalence by demographics and urbanization in US children and adolescents, 2013-2016. JAMA 2018;319(23):2410-2418 [FREE Full text] [doi: 10.1001/jama.2018.5158] [Medline: 29922826]
3. Trasande L, Chatterjee S. The impact of obesity on health service utilization and costs in childhood. Obesity (Silver Spring) 2009;17(9):1749-1754 [FREE Full text] [doi: 10.1038/oby.2009.67] [Medline: 19300433]

XSL•FO
RenderX

4. Sallis JF, Floyd MF, Rodríguez DA, Saelens BE. Role of built environments in physical activity, obesity, and cardiovascular disease. Circulation 2012;125(5):729-737 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.110.969022] [Medline: 22311885]
5. Yusuf ZI, Dongarwar D, Yusuf RA, Bell M, Harris T, Salihu HM. Social determinants of overweight and obesity among children in the United States. Int J MCH AIDS 2020;9(1):22-33 [FREE Full text] [doi: 10.21106/ijma.337] [Medline: 32123625]
6. Galani C, Schneider H. Prevention and treatment of obesity with lifestyle interventions: review and meta-analysis. Int J Public Health 2007;52(6):348-359. [doi: 10.1007/s00038-007-7015-8] [Medline: 18368998]
7. Janicke DM, Steele RG, Gayes LA, Lim CS, Clifford LM, Schneider EM, et al. Systematic review and meta-analysis of comprehensive behavioral family lifestyle interventions addressing pediatric obesity. J Pediatr Psychol 2014;39(8):809-825. [doi: 10.1093/jpepsy/jsu023] [Medline: 24824614]
8. Tate EB, Spruijt-Metz D, O'Reilly G, Jordan-Marsh M, Gotsis M, Pentz MA, et al. mHealth approaches to child obesity prevention: successes, unique challenges, and next directions. Transl Behav Med 2013;3(4):406-415 [FREE Full text] [doi: 10.1007/s13142-013-0222-3] [Medline: 24294329]
9. Direito A, Carraça E, Rawstorn J, Whittaker R, Maddison R. mHealth technologies to influence physical activity and sedentary behaviors: behavior change techniques, systematic review and meta-analysis of randomized controlled trials. Ann Behav Med 2017;51(2):226-239. [doi: 10.1007/s12160-016-9846-0] [Medline: 27757789]
10. Azevedo LB, Stephenson J, Ells L, Adu-Ntiamoah S, DeSmet A, Giles EL, et al. The effectiveness of e-health interventions for the treatment of overweight or obesity in children and adolescents: a systematic review and meta-analysis. Obes Rev 2022;23(2):e13373. [doi: 10.1111/obr.13373] [Medline: 34747118]
11. Psihogios AM, Lane-Fall MB, Graham AK. Adolescents are still waiting on a digital health revolution: accelerating research-to-practice translation through design for implementation. JAMA Pediatr 2022;176(6):545-546. [doi: 10.1001/jamapediatrics.2022.0500] [Medline: 35404381]
12. Yardley L, Spring BJ, Riper H, Morrison LG, Crane DH, Curtis K, et al. Understanding and promoting effective engagement with digital behavior change interventions. Am J Prev Med 2016;51(5):833-842. [doi: 10.1016/j.amepre.2016.06.015] [Medline: 27745683]
13. Tully L, Burls A, Sorensen J, El-Moslemany R, O'Malley G. Mobile health for pediatric weight management: systematic scoping review. JMIR mHealth uHealth 2020;8(6):e16214 [FREE Full text] [doi: 10.2196/16214] [Medline: 32490849]
14. Partridge SR, Knight A, Todd A, McGill B, Wardak S, Alston L, et al. Addressing disparities: a systematic review of digital health equity for adolescent obesity prevention and management interventions. Obes Rev 2024;25(12):e13821. [doi: 10.1111/obr.13821] [Medline: 39256958]
15. Tsoi MF, Li HL, Feng Q, Cheung CL, Cheung TT, Cheung BMY. Prevalence of childhood obesity in the United States in 1999-2018: a 20-year analysis. Obes Facts 2022;15(4):560-569 [FREE Full text] [doi: 10.1159/000524261] [Medline: 35358970]
16. Soltero EG, Lopez C, Hernandez E, O'Connor TM, Thompson D. Technology-based obesity prevention interventions among Hispanic adolescents in the United States: scoping review. JMIR Pediatr Parent 2022;5(4):e39261 [FREE Full text] [doi: 10.2196/39261] [Medline: 36331547]
17. Maguire M. Methods to support human-centred design. International Journal of Human-Computer Studies 2001;55(4):587-634. [doi: 10.1006/ijhc.2001.0503]
18. Bandura A. Human agency in social cognitive theory. American Psychologist 1989;44(9):1175-1184. [doi: 10.1037/0003-066X.44.9.1175]
19. Michie S, Ashford S, Sniehotta FF, Dombrowski SU, Bishop A, French DP. A refined taxonomy of behaviour change techniques to help people change their physical activity and healthy eating behaviours: the CALO-RE taxonomy. Psychol Health 2011;26(11):1479-1498. [doi: 10.1080/08870446.2010.540664] [Medline: 21678185]
20. Davison KK, Birch LL. Childhood overweight: a contextual model and recommendations for future research. Obes Rev 2001;2(3):159-171 [FREE Full text] [doi: 10.1046/j.1467-789x.2001.00036.x] [Medline: 12120101]
21. Tomlin W. UX and Usability Testing Data: Combining Behavioral UX and Usability Testing Data to Optimize Websites. Berkeley, CA: Apress; 2018.
22. Nielsen J, Landauer T. A mathematical model of the finding of usability problems. 1993 Presented at: CHI '93: Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems; 1993 April 24 - 29; USA. [doi: 10.1145/169059.169166]
23. USDA ERS - Rural-Urban Commuting Area Codes. URL: https://www.ers.usda.gov/data-products/rural-urban-commuting-area-codes/ [accessed 2024-10-11]
24. Liinasuo M, Kuula T, Goriachev V, Helin K. Choosing between remote and face-to-face features in a test setting: methodology used in two usability and user experience case studies. Behaviour & Information Technology Taylor & Francis 2023;42(3):298-306. [doi: 10.1080/0144929x.2023.2165453]
25. Dedoose. Los Angeles, CA: SocioCultural Research Consultants, LLC; 2021.

XSL•FO
RenderX

26. Millar RJ, Sahoo S, Yamashita T, Cummins PA. Literacy skills, language use, and online health information seeking among Hispanic adults in the United States. Patient Educ Couns 2020;103(8):1595-1600. [doi: 10.1016/j.pec.2020.02.030] [Medline: 32115313]
27. Chavarria EA, Christy SM, Feng H, Miao H, Abdulla R, Gutierrez L, et al. Online health information seeking and eHealth literacy among Spanish language-dominant Latino adults receiving care in a community clinic: secondary analysis of pilot randomized controlled trial data. JMIR Form Res 2022;6(10):e37687 [FREE Full text] [doi: 10.2196/37687] [Medline: 35238785]
28. Karthik N, Barekatain K, Vu H, Wu DTY, Ehrlich JR. Readability comparison of online Spanish and English patient education materials about vision health. Ophthalmic Epidemiol 2022;29(2):182-188. [doi: 10.1080/09286586.2021.1910316] [Medline: 33832394]
29. Mazmudar RS, Sheth A, Tripathi R, Scott JF. Readability of online Spanish patient education materials in dermatology. Arch Dermatol Res 2021;313(3):201-204. [doi: 10.1007/s00403-020-02036-7] [Medline: 32020323]
30. Ohtani A, Suzuki T, Takeuchi H, Uchida H. Language barriers and access to psychiatric care: a systematic review. Psychiatr Serv 2015;66(8):798-805. [doi: 10.1176/appi.ps.201400351] [Medline: 25930043]
31. Marquine MJ, Jimenez D. Cultural and linguistic proficiency in mental health care: a crucial aspect of professional competence. Int Psychogeriatr 2020;32(1):1-3 [FREE Full text] [doi: 10.1017/S1041610219000541] [Medline: 32008601]
32. Skelton JA, Martin S, Irby MB. Satisfaction and attrition in paediatric weight management. Clin Obes 2016;6(2):143-153 [FREE Full text] [doi: 10.1111/cob.12138] [Medline: 27008068]
33. McMaster CM, Cohen J, Alexander S, Neal R, Gow ML, Calleja E, et al. Satisfaction and acceptability of paediatric weight management services amongst parents and carers: a mixed-methods study. Clin Obes 2020;10(6):e12391. [doi: 10.1111/cob.12391] [Medline: 32830905]
34. Moon K, Sobolev M, Kane JM. Digital and mobile health technology in collaborative behavioral health care: scoping review. JMIR Ment Health 2022;9(2):e30810 [FREE Full text] [doi: 10.2196/30810] [Medline: 35171105]
35. Dendere R, Slade C, Burton-Jones A, Sullivan C, Staib A, Janda M. Patient portals facilitating engagement with inpatient electronic medical records: a systematic review. J Med Internet Res 2019;21(4):e12779 [FREE Full text] [doi: 10.2196/12779] [Medline: 30973347]

Abbreviations

EHR: electronic health record
mHealth: mobile health
MICCO: Mobile Integrated Care for Childhood Obesity

Edited by A Mavragani; submitted 13.05.24; peer-reviewed by T O'Connor, K Haas, B Lancaster; comments to author 30.09.24; revised version received 27.11.24; accepted 24.12.24; published 11.02.25.

Please cite as:
Perez Ramirez A, Ortega A, Stephenson N, Muñoz Osorio A, Kazak A, Phan TL
mHealth App to Promote Healthy Lifestyles for Diverse Families Living in Rural Areas: Usability Study
JMIR Form Res 2025;9:e60495
URL: https://formative.jmir.org/2025/1/e60495
doi:10.2196/60495
PMID:39932772

©Alejandra Perez Ramirez, Adrian Ortega, Natalie Stephenson, Angel Muñoz Osorio, Anne Kazak, Thao-Ly Phan. Originally published in JMIR Formative Research (https://formative.jmir.org), 11.02.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Social Acceptance of Smart Glasses in Health Care: Model Evaluation Study of Anticipated Adoption and Social Interaction

Niek Zuidhof[1,2], PhD; Oscar Peters[3], PhD; Peter-Paul Verbeek[4], Prof Dr; Somaya Ben Allouch[5,6], Prof Dr

[1]Research Group Technology, Health & Care, Saxion University of Applied Sciences, Enschede, The Netherlands
[2]Research Group Employability Transition, Saxion University of Applied Sciences, Enschede, The Netherlands
[3]School of Finance and International Business, Saxion University of Applied Sciences, Enschede, The Netherlands
[4]Philosophy and Ethics of Science and Technology in a Changing World, University of Amsterdam, Amsterdam, The Netherlands
[5]Research Group Digital Interactions, University of Amsterdam, Amsterdam, The Netherlands
[6]Research Group Digital Life, Amsterdam University of Applied Sciences, Amsterdam, The Netherlands

Corresponding Author:
Niek Zuidhof, PhD
Research Group Technology, Health & Care
Saxion University of Applied Sciences
M.H. Tromplaan 28
Enschede, 7513 AB
The Netherlands
Phone: 31 682381157
Email: e.n.zuidhof@saxion.nl

Abstract

Background: Despite the growing interest in smart glasses, it is striking that they are not widespread among health care professionals. Previous research has identified issues related to social interactions involving the use of smart glasses in public settings, which may differ from those associated with their application in health care contexts.

Objective: Assuming that smart glasses mediate contact between the health care provider and patient, the objectives of this research are two-fold: (1) to develop an instrument that combines the adoption and mediation perspectives, and (2) to gain insights into how the intention to use is influenced through aspects of adoption and social interaction.

Methods: A questionnaire was administered to a target audience of health care professionals (N=450), with recruitment via MTurk. The sample primarily included male participants from the United States, with the majority aged 42 years or younger. Although a large portion of respondents were medical doctors, the sample also included nurses and other health care professionals. Data were analyzed by structural equation modeling.

Results: Regarding the aim of developing an instrument combining adoption and social interaction, the internal consistency was above the aspirational level (α>.70) for the instrument. Furthermore, regarding the second objective involving gaining insights into the influential constructs of the anticipated intention to use, the following results were highlighted: in testing the conceptual model, the measurement model generated a good fit and the respecified structural model also generated a good fit. The tested hypotheses confirmed that social interaction constructs could explain a higher variance of users' anticipated intention to use. Perceived social isolation and decreased attentional allocation did not have a significant effect on attitude. Furthermore, the intention to use smart glasses despite nonacceptance of smart glasses by the patient significantly influenced the anticipated intention to use. In summary, constructs that focus on social interaction could contribute to better explanation and prediction of the expected adoption of smart glasses in health care.

Conclusions: The empirical findings of this study provide new insights into how the mediation perspective can increase the explained variance compared to existing knowledge about adoption. Against expectations based on previous literature and despite the social issues raised earlier, these social aspects do play important roles for health care professionals but are ultimately not decisive for the intention to use. As a result, there are fewer threats to the adoption of smart glasses from the perspective of health care professionals than might be expected based on the previous literature. Therefore, the use of smart glasses can still be considered as an innovative way of working in health care.

(JMIR Form Res 2025;9:e49610) doi:10.2196/49610

XSL•FO
RenderX

KEYWORDS

smart glasses; technology adoption; social interaction; instrument development; structural equation modeling

Introduction

Background

The initial public reactions to smart glasses, notably Google Glass (Figure 1), were quite mixed. Google Glass was both praised for its innovative technologies, such as a see-through screen and hands-free use, and criticized for various reasons. Shortly after the introduction of Google Glass, cautionary stories about "Glassholes" emerged, a term coined to describe individuals who misuse the technology in socially inappropriate ways [1]. Examples of this phenomenon include people who wear Google Glass and ignore the people around them or stare off into space while operating their smart glasses. Google responded with rules to help users avoid these situations [2]. There are several examples of various emotional reactions to smart glasses in the media, and ethical questions are repeatedly raised about wearing smart glasses [3,4]. Although research into smart glasses is continuously expanding, current knowledge and theoretical foundations about the acceptance of smart glasses are still scarce.

Smart glasses have emerged as a growing research topic over the past years in various fields [5,6] and could bring about a revolution in health care [7]. Smart glasses are computerized head-mounted devices worn like eyeglasses, which can collect and provide information to the user and the environment. The majority of smart glasses incorporate a display through which information can be accessed while simultaneously maintaining visibility of the external environment. Furthermore, these devices often provide the advantage of hands-free operation [8]. Pilot studies on the use of smart glasses in health care have been conducted [7]; however, few researchers have addressed the issue of preadoption criteria, and identifying these criteria remains a critical issue [9].

Figure 1. Example of a health care professional wearing Google Glass smart glasses in a home environment to illustrate the appearance of the glasses.

Prior Research on the Adoption of Smart Glasses

The initial responses to smart glasses raise the question of how the adoption of smart glasses is progressing.

The growing body of research on the adoption of smart glasses was searched in the existing scientific literature. The following keywords were used: "smart glasses," "head-mounted display*," "wearable display*," "augmented reality (smart) glasses," "adoption," "acceptance," and "acceptability." The search was performed in the following databases on June 12, 2021: Scopus (11 hits), Web of Science (12 hits), ACM Digital Library (96 hits), IEEE Xplore (2 hits), and PubMed (0 hits). All synonyms were connected with the disjunction "OR," and both "smart glasses" and "adoption" were connected with the conjunction "AND." Keywords were searched in the title. The identified publications are listed in Multimedia Appendix 1.

Various methodologies have been employed by researchers to investigate the adoption of smart glasses. For an overview of these methods, please refer to Multimedia Appendix 1. The context of the studies varied, with the general context most often consisting of random individuals or students. A small number of studies were found in the context of health care and tourism; therefore, it can be concluded that knowledge about the adoption of smart glasses is still scarce in these fields. Adoption studies with industry, retail, and agriculture respondents are even more limited. Various methods have been applied, with the majority of studies being qualitative in nature. Most of the previous quantitative studies had a relatively low number of respondents.

XSL•FO
RenderX

The findings presented in Multimedia Appendix 1 offer foundational inputs for the development of a questionnaire. Furthermore, the social aspects surrounding smart glasses were mentioned multiple times in previous studies, and therefore, it could be assumed that social interaction is a relevant topic regarding the acceptance of smart glasses.

The Unified Theory of Acceptance and Use of Technology (UTAUT) [10], the Technology Acceptance Model (TAM) [11], and the Diffusion of Innovations [12] were found to be the most commonly accepted theories by scholars [13,14]. Some of the findings from prior research in Multimedia Appendix 1 might be related to the aforementioned models such as the influence on perceived usefulness, perceived ease of use, facilitating conditions, image, and perceived enjoyment [10,15]. However, other findings suggest themes regarding both the ethical and social contexts, which can be termed social acceptability [16]. Social acceptability can be divided into conversation partners, such as interpersonal communication, and spectators or third parties, with examples such as disturbance and controversial public use. The theme of privacy is also related to the social context and can be divided into the same categories: conversation partners and third parties. For example, wearing smart glasses is not only about the privacy of the user, but also about other people's privacy, and this can strongly influence the user's decision-making [17]. Figure 1 provides an illustrative example of an encounter with a health care professional wearing smart glasses. To summarize the findings of the prior research in Multimedia Appendix 1, it can be concluded that relevant aspects of the preadoption criteria should be informed by the current knowledge in adoption research, such as the TAM, and should also take into account the social context from the perspectives of both caregivers and patients.

Devices worn on the body in general influence the symbolic meaning in the natural environment of the person and the public surroundings and have psychological impacts on wearers and behavioral consequences for wearers [18]. Over the past decades, the TAM [19] had an incredible influence on empirical research. One might also say in general that current adoption models take a snapshot of adoption readiness, and the TAM consistently explains about 40% of the variance in the intention to use IT and the actual usage by individuals [15]. However, these adoption models have limitations [20], in which the current reality regarding smart glasses may not be fully involved. Current models are linear and strongly focused on the user's perspective alone. Moreover, according to recent studies, social influence, as it is considered in the UTAUT [10], may involve more than just the influence and desire of others in the future use of the one considering it [3]. Several studies have suggested that social context influences usage norms and that public use is controversial [21,22]. The use of smart glasses does not reproduce existing norms and understandings in turn-taking, knowledge, and identity and could reconfigure norms by creating new settings in which activities begin to compete and interfere with each other [23,24]. In addition, privacy is often mentioned to be influential [25]. The privacy of the user is important, and other people's privacy can also strongly influence the user's decision-making [17]. Therefore, smart glasses force us to take a broader view of adoption models, and we should not only consider the perspectives of potential users but also look at the surroundings. Hence, social influences might be currently underestimated in the context of the acceptance of smart glasses given the reports from various media and recent publications on the subject.

Turning to a perspective that takes a broader view of users and their surroundings and environment [26], the mediation perspective fits well with the identified gaps regarding the acceptance of smart glasses. The mediation perspective of the philosophy of technology considers the relationship between humans and technology from a phenomenological perspective, and it inspired the beginning of ubiquitous computing [27]. Furthermore, Mann [28] explained the use of smart glasses after over 35 years as a mediated version of reality. Those mediated relationships are reciprocal, and technology is perceived as the mediator between humans and the world. For example, a user looks *through* glasses rather than *at* glasses, and this is therefore termed as an embodiment relationship. In addition, smart glasses can represent information on a parallel screen while seeing the world, which is termed as a hermeneutic relationship [29]. A combined perspective of both adoption and mediation could complement each other in understanding the appropriation of smart glasses (Figure 2; derived and adapted from Zuidhof et al [30]). However, knowledge of the mediated experience of smart glasses is still scarce, and recent studies have reported strong influences of the social context in decision-making [17,21]. The theoretical framework of adoption and mediation can reveal both the user's perspective and the perspective of others who are related to the user, such as a patient.

In the field of the philosophy of technology, mediation is interpreted as determining human-technology relationships and interactions, and anticipating influences on human behavior [29]. Prior studies on the adoption of smart glasses, as illustrated in Multimedia Appendix 1, not only present familiar constructs but also have begun, since 2010, to delve into social themes from various perspectives, including those of the user, conversational partner, and third parties [26]. A large-scale survey in which the anticipated perspective of the health care professional and the anticipated social interaction are united is lacking in current knowledge. Based on previous research results and the theoretical framework of adoption and mediation (as part of a doctoral research project), it seems relevant to explore adoption themes and assess if social interaction influences the intention to use smart glasses.

Figure 2 proposes to adapt the theoretical framework of adoption and mediation to the context of the health care professional and the patient. It is considered that smart glasses for health care professionals are still in an early phase of dissemination, and therefore, it is needed to gather knowledge about anticipating the use of smart glasses. Moreover, health care professionals are expected to anticipate adoption while taking into account social interaction with patients, but they can also take into account the patient's perspective in social interaction. Therefore, this goes beyond mere adoption for self-interest and may lead to a higher or lower intention to use smart glasses.

XSL•FO
RenderX

Figure 2. Theoretical framework of anticipated adoption and social interaction.

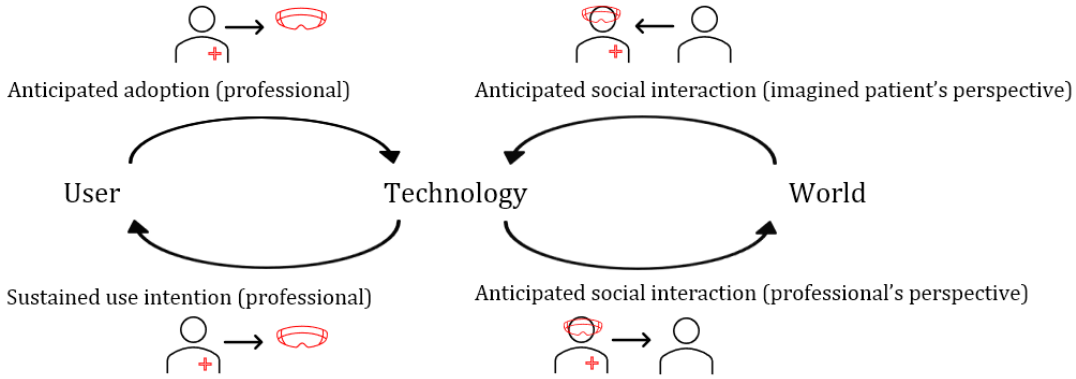

Objectives and Hypotheses

Prior research on the adoption of smart glasses has been mainly conducted in general contexts; however, some studies were carried out in the area of tourism or health care. Few of the recent qualitative studies on smart glasses in health care are part of the authors' doctoral research projects. Furthermore, it appears that there are no large-scale surveys conducted on the adoption of smart glasses and that most of the recent studies have an explorative and qualitative nature. Prior research on the adoption of smart glasses is presented in Multimedia Appendix 1. In addition, the emotional and ethical aspects that can play roles as issues to adoption described above are still involved to a limited extent in recent studies and might be especially relevant in professions where people are vulnerable or interact intensively like in health care.

The objectives of this study were two-fold: (1) to develop an instrument that combines the perspectives of adoption and mediation to measure the intention to use smart glasses by health care professionals, and (2) to obtain insights into how the intention to use is influenced by adoption and social interaction aspects. Regarding the state of the art and the objectives of this study, the following research question was formulated: "To what extent do constructs derived from a mediation perspective contribute to the established constructs of adoption to explain the intention to use smart glasses?"

Observed Variables Derived From Technology Acceptance Literature

In the proposed model (Figure 2), it was argued that technology adoption and social influence might influence each other and the user. The underlying idea here is that a decision to use smart glasses is explained by a linear model and is not only aimed at the self but also others, as is endorsed by the mediation perspective [30,31]. The recently developed conceptual model to explain the anticipated adoption and the social aspects of smart glasses has been outlined in Figure 3, and the aspects will be included.

Figure 3. Conceptual model with the hypotheses of the anticipated adoption and social interaction of smart glasses by health care professionals. H: hypothesis; TAM: Technology Acceptance Model.

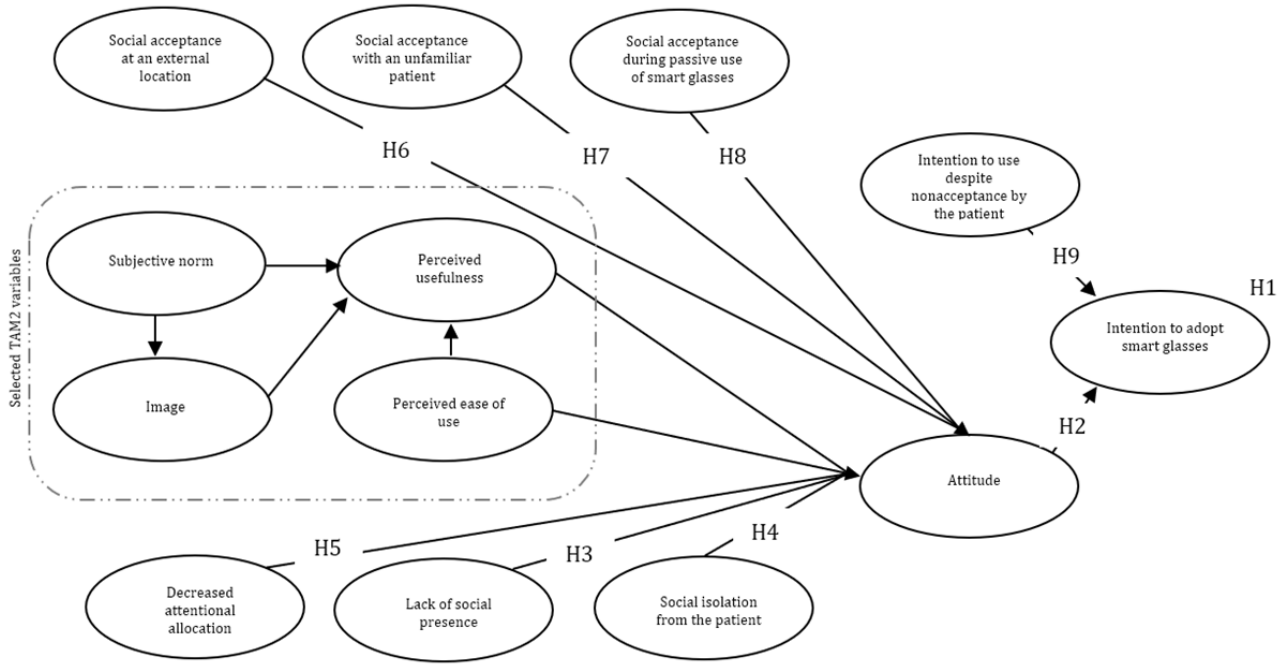

XSL•FO
RenderX

Observed Variables Derived From the Technology Acceptance Model 2

As one of the most accepted theories by scholars, in the TAM [13,14], intention to use is proposed as a key independent variable. The constructs "perceived usefulness" and "perceived ease of use" are mainly aimed at their considerations and do not yet tell us much about how this relates to others. Therefore, we found the Technology Acceptance Model 2 (TAM2) to be the most appropriate because it builds on the basic model and extends this model with constructs that take others into account, such as the constructs "subjective norm" and "image." Other constructs from the TAM2 model, such as experience, voluntariness, job relevance, output quality, and result demonstrability, have been excluded as they extend beyond the scope of this study. These constructs do not specifically address social interaction, which is the central emphasis of our research. By adding specifically more social constructs, we aim to gain a more holistic perspective of not only personal considerations in own use but also how much a health care professional takes patients into account and how this influences the anticipated use of technology.

Hypothesis 1 is as follows: Additional social interaction constructs can explain the higher variance of users' anticipated intention to use smart glasses compared with subjective norm, image, perceived usefulness, and perceived usefulness on anticipated intention to use.

The attitude toward technology is a debatable construct in adoption studies because the added value in the model might not be good enough [10], or it may theoretically differ little from the intention to use. Similar to the Theory of Planned Behavior [32], we regard attitude as a predictor of the intention to use smart glasses, and we added the construct "attitude" because of previous findings [31,33], which may be related to the early phase of the adoption of smart glasses where we can only measure the anticipated intention to use.

Hypothesis 2 is as follows: Users' attitudes directly influence the anticipated intention to use smart glasses.

Observed Variables Regarding Social Interaction With Smart Glasses

Previous studies found several constructs under the themes that seemed theoretically relevant, such as attentional shift, social presence, and social isolation [14,31,34-36]. The items were translated as accurately as possible and applied to the anticipation of smart glasses. This resulted in the following constructs: lack of social presence, social isolation of the patient, and reduced attentional allocation to the patient.

Hypothesis 3 is as follows: The lack of social presence due to wearing smart glasses during a physical contact moment influences the attitude directly.

Hypothesis 4 is as follows: Social isolation due to wearing smart glasses during a physical contact moment influences the attitude directly.

Hypothesis 5 is as follows: The decreased attentional allocation due to wearing smart glasses during a physical contact moment influences the attitude directly.

Observed Variables Regarding Social Scenarios With Smart Glasses in Health Care

Furthermore, in previous qualitative studies, themes related to the context of smart glass use were also found, such as location, an unknown patient, and wearing smart glasses but not actively using them [31]. These scenarios were linked to social presence [36], resulting in social presence in 3 scenarios, namely, providing care with smart glasses on in an external location (outpatient care), providing care with smart glasses while in passive use, and providing care with smart glasses to an unfamiliar patient.

Hypothesis 6 is as follows: The use of smart glasses at an external location during a physical contact moment directly influences the attitude.

Hypothesis 7 is as follows: The use of smart glasses with an unknown person during a physical contact moment directly influences the attitude.

Hypothesis 8 is as follows: The passive use of smart glasses during a physical contact moment directly influences the attitude.

Observed Variables Regarding Nonacceptance of Smart Glasses by the Patient

The mediation approach visualizes the relationship with the outside world during adoption, while technology acts as a mediator in the work of the health care professional and the patient. To understand both the social interaction with patients during the use of smart glasses and the sustained use of smart glasses, the unintended consequences of smart glass use might also be relevant [26,37-39]. In particular, the patient's nonacceptance of the use of smart glasses is interesting, simply because the patient's acceptance of smart glasses would probably not create tension in the intentional use of smart glasses. The mediation perspective has led to the following construct in the model: intention to use despite social nonacceptance by others.

Hypothesis 9 is as follows: The intention to use smart glasses despite the nonacceptance of smart glasses by the patient directly influences the anticipated intention to use.

Methods

Study Design

To achieve the aims of this study, a summated rating scale development approach for social sciences was used [40]. The initial steps were aimed at defining constructs and designing the scale. The resulting scale was evaluated through an online survey among a sample of students.

Defining Constructs and Item Pool Development

To develop an item pool, existing literature was consulted, and the state-of-the-art knowledge, as described in the Introduction, was combined with the theoretical framework of adoption and mediation. A core aspect of the framework (Figure 2) is the reciprocal relationship among the user, the technology, and the world. In the next step, the aspects from Figure 2 were substantiated with existing literature and translated into items.

XSL•FO
RenderX

See Multimedia Appendix 2 for the item pool. All measurements were recorded using 5-point Likert scales. Back translation was performed by 2 researchers as a validation method for the survey [41] and to develop a robust item pool.

Item Pool Refinement

To check for comprehensibility, readability, and internal consistency, 2 pretest moments were organized. First, the questions were reviewed by 2 potential respondents. Next, data collection (as a pretest) was conducted at a health care institution in the Netherlands (n=17). Although the pretest was conducted on a small sample, internal consistency was acceptable. Initially, reversed items were included, but inconsistent response patterns were observed for those items. Respondents reported too much confusion with reversed items, and therefore, these items were removed from the questionnaire.

Data Collection

Respondents (N=450) were recruited in July 2022 on the internet using Amazon's Mechanical Turk. The questionnaire was made available to respondents who met the following inclusion criteria: (1) language: English; (2) employment industry: health care; and (3) location: Northwest Europe (Austria, Belgium, Denmark, Finland, France, Germany, Iceland, Ireland, Luxembourg, Netherlands, Norway, Sweden, and Switzerland), Canada, United Kingdom, and United States. The questionnaire was closed once the desired number of respondents was reached.

Demographic Information

A summary of the demographic information of the respondents is presented in Table 1. The sample pool consisted mainly of male participants, with over three-quarters being 42 years of age or younger. The majority of the respondents were from the United States. With regard to occupation, the participants were distributed as follows: physicians, 192 (42.6%); nurses, 165 (36.7%); and others, 93 (20.7%). Other professions included a diverse group of medical assistants (n=14), medical technical professionals (n=11), managers (n=10), health care professionals (n=9), IT professionals (n=6), and home care professionals (n=4), among other various professions in health care.

XSL•FO
RenderX

Table 1. Demographic information of the respondents.

Characteristic	Value (N=450), n (%)
Gender	
Female	303 (32.4)
Male	146 (67.3)
Prefer not to say	1 (0.2)
Age (years)	
>72	1 (0.2)
63-72	15 (3.3)
53-62	42 (9.3)
43-52	45 (10.0)
33-42	124 (27.6)
23-32	222 (49.3)
18-22	1 (0.2)
Country of residence	
Canada	4 (0.9)
United States of America	442 (98.2)
Northwest Europe	3 (0.7)
United Kingdom	1 (0.2)
Highest degree	
High school or less	5 (1.1)
Some college but no degree	22 (4.9)
Associate degree	32 (7.1)
Bachelor’s degree	235 (52.2)
Master’s degree	145 (32.2)
Professional or doctoral degree	11 (2.4)
Current profession	
Physician	192 (42.6)
Nurse	165 (36.7)
Other	93 (20.7)
Employment status	
Employed full time	401 (89.1)
Employed part-time	35 (7.8)
Student	4 (0.9)
Self-employed	10 (2.2)
Years of work experience	
<10	224 (49.8)
11-20	115 (25.6)
21-30	74 (16.4)
31-40	28 (6.2)
>40	9 (2.0)

XSL•FO
RenderX

Ethical Considerations

Before starting the survey, respondents provided active online consent after reading the opening statement (Multimedia Appendix 3). Furthermore, pictures were used with permission from health care professionals and patients. Respondents were rewarded with US $2 when they completed the survey. Study data were deidentified and anonymously used for analysis. This study was reviewed and approved by the Ethical Review Board of the University of Twente (approval number 211121).

Procedure

After the opening statement and consent, respondents were asked to watch 2 short videos about smart glasses to get an idea of the technology and its use. The first video showed an augmented reality platform for health care with English-spoken animations [42], and the second video showed an example of Vuzix M400 smart glasses in the context of health care [43]. After watching both videos, respondents were asked to fill out the questionnaire. Demographic information was requested at the end of the questionnaire. Some photos used in the questionnaire are presented in Figure 4.

Figure 4. Photos used in the questionnaire about examples of using smart glasses consecutively: (A) contact with patients, (B, C) face-to-face interaction in outpatient care, (D) inpatient care, (E) passive use of smart glasses, and (F) active use of smart glasses (published with permission).

Data Analysis

Before the analyses, data were checked for normality. No significant deviation from normality was found. All skewness and kurtosis values were below 1.96, and values between –2 and +2 are considered acceptable to prove normal distribution [44]. Structural Equation Modeling was performed with maximum likelihood estimation, as advocated by [45,46], to test the conceptual model of the intention to adopt smart glasses by health care professionals. To test if the conceptual model fits the data well, the next absolute and incremental model fit indices were used as suggested by Holbert and Stephensen [47] and Hoyle and Panter [48]. The most common absolute fit index is the χ^2 goodness-of-fit test, and to minimize the impact of sample size on the model chi-square, the relative/normed chi-square ($\chi^{2/df}$) was used [49]. The standardized root mean square residual (SRMR) was reported as an absolute fit index. The following incremental fit indices have been reported: Tucker-Lewis index (TLI) and root mean square error of approximation (RMSEA). The practitioner recommendations for model evaluation were followed, suggesting a cutoff value close to 0.95 for the TLI in combination with a cutoff value close to 0.09 for the SRMR to evaluate model fit, and a cutoff value close to 0.06 for the RMSEA [50]. The Fornell and Larcker discriminant validity criterion was used to test discriminant validity. The Fornell and Larcker criterion is satisfied when a construct is more closely related to its own indicators than to other constructs [51].

Results

Diffusion of Smart Glasses in Health Care

The questionnaire on the anticipated adoption and mediation of smart glasses was designed to examine the perceptions of future users regarding smart glasses. It was also designed to get a better understanding of how specific situations in the use of smart glasses influence the intention to use them. The research question was: “To what extent do constructs derived from a mediation perspective contribute to the established constructs of adoption to explain the intention to use smart glasses?”

Participants were queried regarding their familiarity and experience with smart glasses. The constructs “knowledge” and “persuasion” served as constructs in assessing how health care professionals perceive the diffusion of smart glasses. Overall, the respondents reported having substantial knowledge about smart glasses (eg, I know what someone can do with smart glasses; mean score 4.02, SD 0.84) and expressed understanding of the concept of smart glasses (mean score 4.08, SD 0.79). Close to 75% of the respondents (n=339) indicated their willingness to try smart glasses if they were available to them (mean score 3.94, SD 0.93).

XSL•FO
RenderX

Explaining Anticipated Adoption of Smart Glasses: Model Comparison

To investigate the added value of the observed variables in the conceptual model, a comparison was made between validated TAM2 [52] constructs used to create a baseline and the additional social interaction variables in this study. The constructs "perceived usefulness" and "perceived ease of use" were adapted to the context of smart glasses, and subjective norm and image were included because they represent social interaction. Although the construct "attitude" was not included in the TAM, there was a reason to include this validated construct based on previous studies on the early phase of the diffusion of smart glasses.

Interpreting Selected TAM2 Effects: Intention to Use, Perceived Usefulness, Perceived Ease of Use, Subjective Norm, and Image

Measurement Model

The initial measurement model generated a good fit (χ^2_{142}=183.55; $\chi^{2/df}$=1.29; SRMR=0.03; TLI=0.99; RMSEA=0.03 [CI 0.01-0.04]). The internal consistency was above the aspirational level (α>.70).

Structural Model

The results obtained from testing the validity of a causal structure of earlier validated variables showed a good fit (χ^2_{14}=218.314; $\chi^{2/df}$=1.52; SRMR=0.04; TLI=0.98; RMSEA=0.03 [CI 0.02-0.04]).

Path Model

The path model with standardized path coefficients is featured in Figure 5. The standardized path coefficients show significant direct effects of perceived usefulness and perceived ease of use on the intention to use smart glasses. Furthermore, subjective norm had a direct effect on perceived usefulness. Figure 5 also shows that subjective norm has a direct effect on image. The direct effect of perceived ease of use on perceived usefulness was not significant. The direct effect of image on perceived usefulness was also not significant. Subjective norm was the only contributor to perceived usefulness. In addition, perceived usefulness was the strongest contributor to the intention to use smart glasses. Squared multiple correlations showed that image accounted for 61%, perceived usefulness accounted for 51%, and intention to use smart glasses accounted for 73% of the explained variance.

Figure 5. Standardized path coefficients of the TAM2 variables for anticipated use of smart glasses by health care professionals (N=450). Dotted lines represent nonsignificant paths, italics represents nonsignificant factor loadings, and underline represents squared multiple correlations. TAM: Technology Acceptance Model. ****P*<.001.

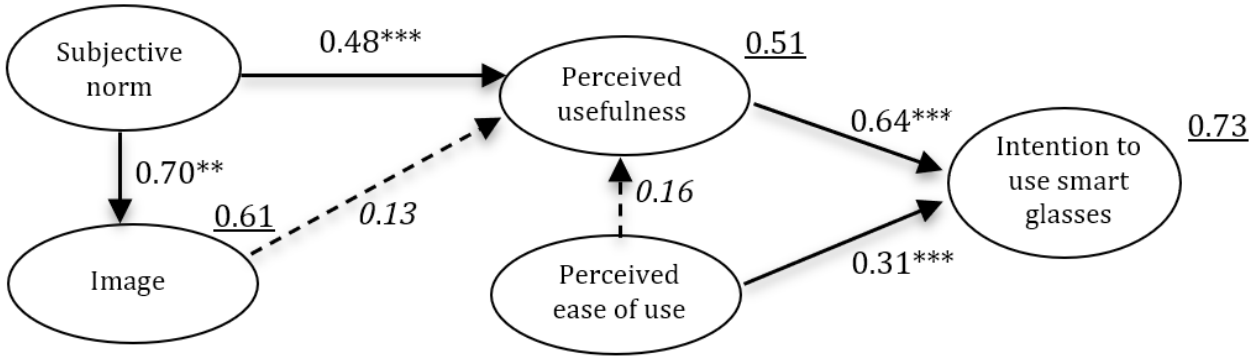

Interpreting Social Interaction Effects: Lack of Social Presence, Social Isolation, Decreased Attentional Allocation, External Location, Unfamiliar Person, Passive Use, Intention to Use Despite Nonacceptance by Others

Measurement Model

The initial measurement model generated a good fit (χ^2_{1461}=2146.696; $\chi^{2/df}$=1.47; SRMR=0.04; TLI=0.95; RMSEA=0.03 [CI 0.03-0.04]). The internal consistency was above the aspirational level (α>.70).

Structural Model

The results obtained from testing the validity of a causal structure of the conceptual model showed that the initial model did not fit the data (χ^2_{1525}=4896.900; $\chi^{2/df}$=3.21; SRMR=0.25; TLI=0.78; RMSEA=0.07 [CI 0.07-0.07]). Post-hoc modification indices suggested an improved fit by correlating error terms of perceived ease of use and social norm, social norm and image, lack of social presence and decreased attentional allocation, lack of social presence and social isolation, decreased attentional allocation and intention to use, external location and passive use, unfamiliar person and passive use, passive use and social norm, and passive use and intention to use despite social nonacceptance by others. Furthermore, post-hoc indication indices also suggested additional paths between perceived usefulness and intention to use, perceived ease of use and intention to use, perceived ease of use and image, perceived ease of use and social isolation, social norm and external location, social norm and lack of social presence, and lack of social presence and intention to use despite social nonacceptance by others. Some of the paths in the conceptual model were suggested differently, such as external location to intention to use instead of attitude, and the same applied for unfamiliar people and passive use. The paths suggested were from attitude to intention to use despite social nonacceptance by others and attitude to lack of social presence. The respecified model

XSL•FO
RenderX

generated a good fit (χ^2_{1337}=2113.035; $\chi^{2/df}$=1.58; SRMR=0.08; TLI=0.95; RMSEA=0.04 [CI 0.03-0.04]). Multimedia Appendix 2 summarizes the mean and SD, Cronbach α, factor loadings (β), and squared multiple correlations (R^2) of the observed variables to predict the anticipated intention to use smart glasses. The correlation matrix is available upon request.

Path Model

It was hypothesized that additional social interaction constructs could explain higher variance of users' anticipated intention to use compared to subjective norm, image, perceived usefulness, and perceived usefulness on anticipated intention to use (hypothesis 1). Squared multiple correlations are featured in Figure 6. It was shown that intention to use represented 96%, and it was raised with additional social interaction variables compared with squared multiple correlations, with representation of 73% from intention to use, perceived usefulness, perceived ease of use, subjective norm, and image. Attitude accounted for 89%, intention to use smart glasses despite social nonacceptance accounted for 65%, perceived usefulness accounted for 51%, and image accounted for 29%. These results supported the acceptance of hypothesis 1.

It was hypothesized that the user's attitudes directly influence the anticipated intention to use smart glasses (hypothesis 2). The standardized path coefficients in Figure 6 show a significant direct effect of attitude toward smart glasses on the intention to use smart glasses, which supported the acceptance of hypothesis 2.

It was stated that the lack of social presence due to wearing smart glasses during a physical contact moment influences the attitude directly (hypothesis 3). Figure 6 shows a significant negative direct effect of lack of social presence on attitude, and therefore, the results supported the acceptance of hypothesis 3. Surprisingly, the lack of social presence also had a positive direct effect on the intention to use despite social nonacceptance by patients.

Furthermore, it was hypothesized that social isolation due to wearing smart glasses during a physical contact moment influences the attitude directly. Social isolation did not have a significant effect on attitude (Figure 6), and this result supported the rejection of hypothesis 4.

In addition, it was hypothesized that the decreased attentional allocation due to wearing smart glasses during a physical contact moment influences the attitude directly (hypothesis 5). Figure 6 shows that decreased attentional allocation does not have a significant effect on attitude, and this result supported the rejection of hypothesis 5.

It was also hypothesized that the social acceptance of the use of smart glasses at an external location during a physical contact moment directly influences attitude (hypothesis 6). Figure 6 shows that use at an external location could not significantly explain attitude, and this result supported the rejection of hypothesis 6.

In addition, it was hypothesized that social acceptance regarding the use of smart glasses with an unknown person during a physical contact moment directly influences attitude. Figure 6 shows that there was no significant effect of the use of smart glasses with an unknown person on attitude, and this result supported the rejection of hypothesis 7.

Furthermore, it was hypothesized that the passive use of smart glasses during a physical contact moment directly influences attitude. Figure 6 shows that social acceptance regarding the passive use of smart glasses indeed directly influences attitude. Therefore, hypothesis 8 is accepted. In addition, social acceptance regarding the passive use of smart glasses has significant effects on the intention to use despite social nonacceptance by the patient, intention to use, and perceived usefulness.

Examining the nonacceptance of smart glasses by the patient, it was hypothesized that the intention to use despite the nonacceptance of smart glasses by the patient directly influences the anticipated intention to use (hypothesis 9). Figure 6 shows a significant effect, and therefore, hypothesis 9 is accepted. In addition, an overview of the accepted and rejected hypotheses can be found in Figure 7.

XSL•FO
RenderX

Figure 6. Standardized path coefficients of the adoption of smart glasses and the social interaction variables for the anticipated use of smart glasses by health care professionals (N=450). Dotted lines represent nonsignificant paths, italics represents nonsignificant factor loadings, and underline represents squared multiple correlations. TAM: Technology Acceptance Model. *P<.05; ***P<.001.

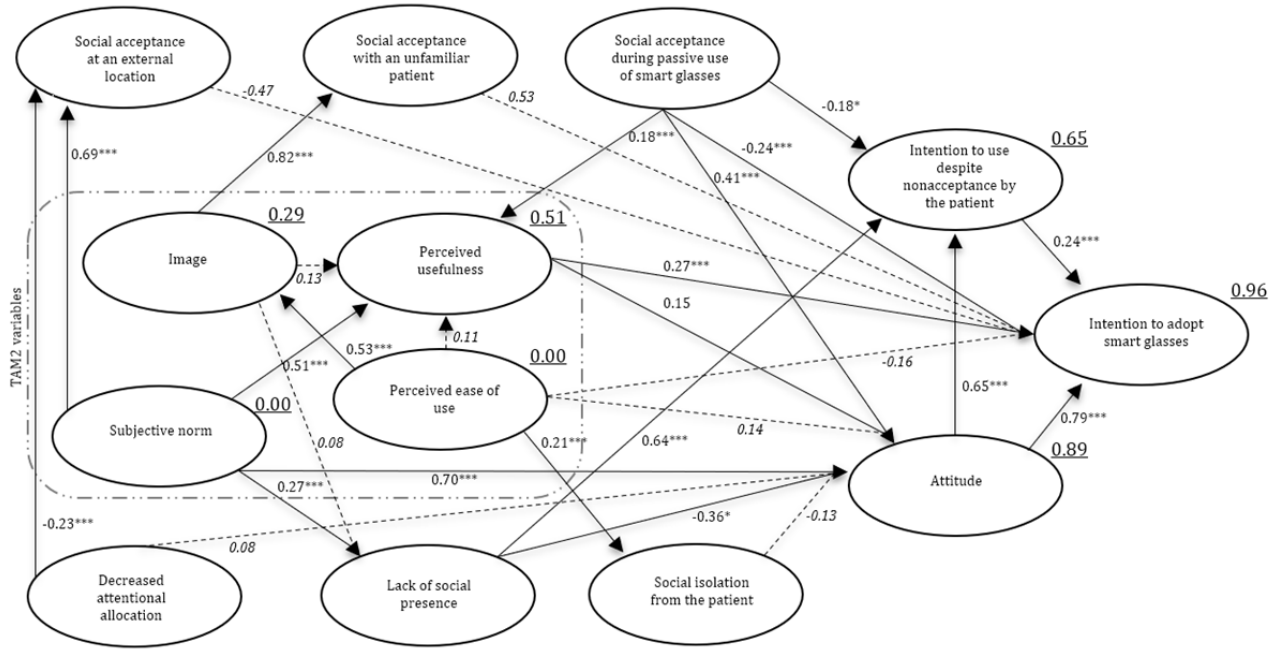

Figure 7. Overview of the tested hypotheses of the adoption of smart glasses and the social interaction variables for the anticipated use of smart glasses by health care professionals (N=450). Hypothesis 1 is underlined, and the hypothesis is accepted. Solid lines represent accepted hypotheses, and dotted lines represent rejected hypotheses. H: hypothesis.

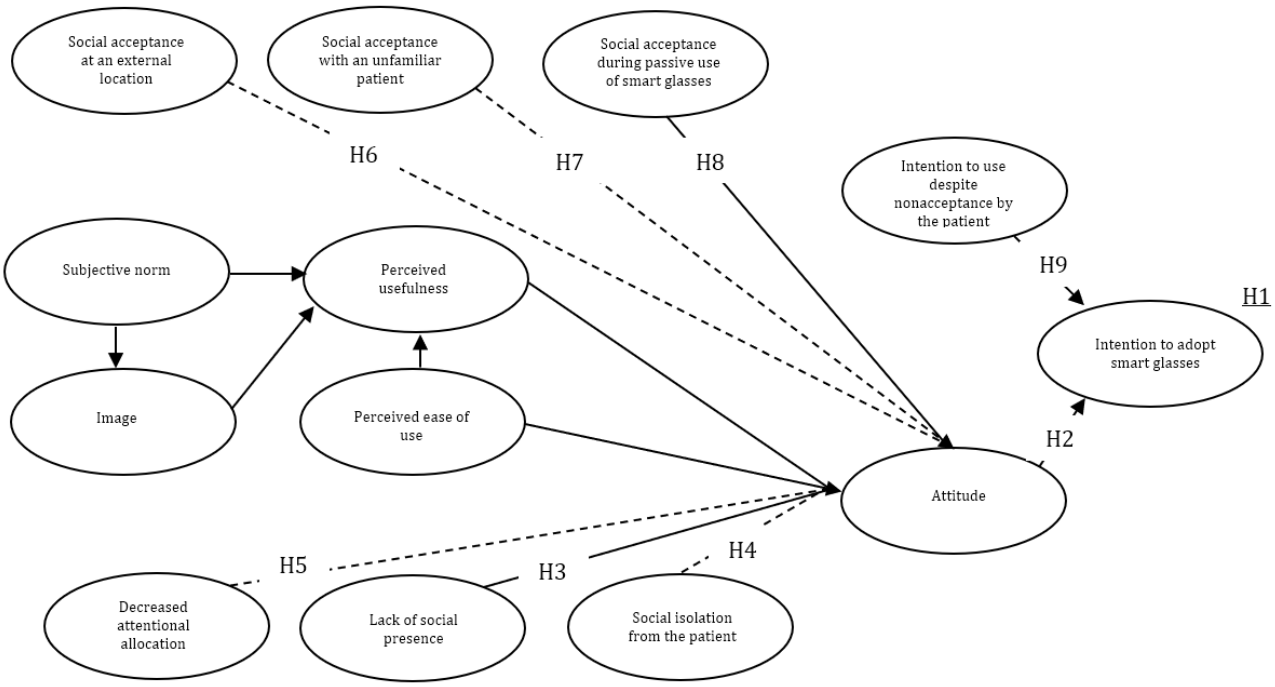

Discussion

Overview

Considering the current state of the art and the objectives of this study, we formulated the following research question: "To what extent do constructs derived from a mediation perspective contribute to the established constructs of adoption to explain the intention to use smart glasses?" This study had 2 primary objectives. The first objective was to develop an instrument that combines the perspectives of adoption and mediation to measure the intention to use smart glasses by health care professionals. A conceptual model was created by drawing on previous findings related to smart glass adoption and testing hypotheses about their anticipated effects on attitudes and intentions to use. The instrument developed integrates both adoption and social interaction perspectives, highlighting smart glasses as a mediating factor between individuals, and it was validated with health care professionals. The approach for this instrument

XSL•FO
RenderX

builds on prior research and represents a step forward in understanding the role of smart glasses in health care.

The second objective was to obtain insights into how the intention to use is influenced by adoption and social interaction aspects. The structural equation analysis showed that integrating social interaction factors with selected TAM2 constructs led to a higher explained variance in the intention to use smart glasses than TAM2 alone. Key factors influencing the intention to use included attitude, perceived usefulness, intention to use despite social nonacceptance, and passive use of smart glasses. Social presence had an unexpected effect: as social presence increased, attitudes toward smart glasses decreased. Despite social nonacceptance by patients, health care professionals still showed willingness to use smart glasses, suggesting that this factor may be less decisive in health care. Passive use was linked to perceived usefulness and intention to use, indicating that wearing smart glasses on top of the head (as presented in Figure 4) could be convenient to have smart glasses close by, even when not actively in use. These findings suggest that social interaction plays a nuanced role in adoption, and further research is recommended to explore these relationships, especially regarding social presence and passive use.

Principal Findings

Development of the Instrument

Regarding the first objective of this study, there was an initial endeavor to create an instrument for investigating the anticipated adoption and mediation (as expressed through social interactions) of smart glasses within the health care sector. Consequently, a conceptual model was derived, which (1) drew upon prior findings related to the adoption of smart glasses, (2) tested hypotheses pertaining to the anticipated effects on attitudes and use intention, and (3) was applied to a sample of health care professionals for validation. We believe that the proposed conceptual model of expected adoption and social interaction provides a strong theoretical basis for further development of the model. The structural model in this study adds a higher explained variance but can be improved in the parsimony of the model. Furthermore, the focus in the model can be shifted in later research studies to other bystanders, such as colleagues, who may be confronted with a user of smart glasses.

The initial objective of this research was to develop an instrument that integrates the perspectives of both adoption and mediation. The resultant instrument was crafted to align closely with the state-of-the-art knowledge of smart glass adoption while simultaneously addressing aspects of social interaction. Central to this development was the assumption that smart glasses serve as a mediating factor between individuals. While there exist numerous methods to investigate social interaction factors, this study opted to ground its approach in findings from previous research.

Explaining the Anticipated Adoption of Smart Glasses

The second aim of this study was to gain insights into how the intention to use is influenced by adoption and social interaction aspects. The outcomes of the structural equation analysis revealed that our path model (Figure 6) integrating constructs from both TAM2 and social interaction variables demonstrated a higher explained variance in the anticipated intention to use smart glasses, compared to the isolated TAM2 constructs that were selected. This finding indicates that social interaction concerning smart glasses plays a role in anticipating the adoption of smart glasses in health care besides the perceived usefulness and attitude toward smart glasses. More specifically, the intention to use smart glasses is directly influenced by attitude, perceived usefulness, intention to use despite social nonacceptance, and passive use of smart glasses.

The presented model showed that attitude and intention to use both had high squared multiple correlations. This gives a reason to consider whether there is a theoretical difference between these constructs. In the Introduction section, it was also indicated that the attitude toward technology is a debatable construct in adoption studies because the added value in the model would not be substantial [10]. Original adoption models were developed for computer use in the 1980s, but smart glasses are different in several ways: they are worn on the body, are in the field of view, and often contain a camera. In this model, attitude is a mediating variable for subjective norm and the lack of social presence and was found to be important for the intention to use. Follow-up studies may investigate this relationship further.

Intention to use despite social nonacceptance shows that the influence of others is important for the health care professional but that it is not decisive to eventually use smart glasses. This can be understood because care actions can often be unpleasant for the patient but necessary for health purposes. This construct may score very differently in other disciplines where social relationships may be more important. Furthermore, the anticipated intention to use smart glasses is indirectly influenced by the lack of social presence and subjective norm.

A notable result of the lack of social presence is the decline in attitude. If the lack of social presence decreases (with an increase in social presence), the attitude toward smart glasses also decreases. This might indicate that respondents do not expect that smart glasses will ever provide more social presence and would not use them for that reason. One explanation that supports this finding is the positive regression coefficient with intention to use despite social nonacceptance by patients. This suggests that professionals would use smart glasses despite social nonacceptance by the patient.

The results of this study also indicate that there is a complex relationship between the passive use of smart glasses and 4 variables: a direct relationship with intention to use and an indirect relationship through perceived usefulness, attitude, and intention to use despite social nonacceptance. This can have 4 explanations, namely that passive use has a negative relationship with the intention to use, which indicates, for example, that health care professionals would rather not passively use smart glasses. However, this seems to contradict the finding of a weaker relationship with perceived usefulness, which could mean, for example, that it is convenient to have the glasses quickly at hand by wearing them on top of the head, just as a doctor places a stethoscope on the neck. Passive use also leads to an increased positive attitude toward glasses, which in turn leads to increased intention. In addition, there was a significant

XSL•FO
RenderX

negative regression between passive use and intention to use despite social nonacceptance by the patient. This could indicate, for example, that the health care professional is willing not to use technology or, in this case, is willing to remove it if the smart glasses are not actively used and the patient is not comfortable with them.

The second objective was to gain insights into how the intention to use smart glasses is influenced by the aspects of both adoption and social interaction. Based on our findings, it can be asserted that the presented model provides a higher explained variance for the intention to use smart glasses. The majority of the hypotheses were confirmed. However, some results were unexpected, such as the influence of social presence on attitudes, as discussed previously. Contrary to expectations drawn from prior research [31,53], we found that a lack of social presence did not result in a negative attitude or diminished intention to use. Therefore, we recommend further exploration in subsequent research, potentially using methodologies such as in-depth interviews and quasiexperimental studies. The same applies to the construct of social acceptance during the passive use of smart glasses. The model indicated that this factor also impacts perceived usefulness, suggesting that passive use of smart glasses could be beneficial in a work context and could consequently enhance the intention to use. Further research should shed more light on the underlying reasons for this finding.

Limitations

This study focused on the combination of adoption and social interaction. A limitation of this study lies in the comparison with the full TAM2 model. A priori–selected TAM2 constructs based on relevance and the aim of the study were included, namely to investigate social interaction and assess to what extent themes from the mediation perspective add something to current models. A comparison with the full TAM2 model is recommended for a follow-up study.

Second, in this study, social acceptance was examined through scenarios and not directly as a variable. The reason for this was to tie in with the previous findings and become more concrete in the specific context. This ignores social acceptance as an independent construct, and it is advisable to do so now that it appears that social interaction plays a substantial role in the model.

Third, as discussed in the conclusion, attitude and intention to use as variables score high on the explained variance, and it is questionable whether they are distinctive enough. Nevertheless, this difference is based on previous literature. Attitude was assumed as a preliminary state in the conceptual model, and it also fits the final model. It is recommended to investigate this further in a subsequent study, and it might be related to the last limitation.

The sample was limited to individuals who have access to MTurk. The distribution of respondents was mainly centered in the United States, and the questionnaire may have been completed mainly by people who were enthusiastic about the topic. As other researchers have addressed potential issues with sample representativeness and quality of the data [54,55], the results are more difficult to generalize and are not necessarily representative of the population of health care professionals. Therefore, it is important to revalidate the results with a reliable and representative dataset.

Furthermore, participants were presented with 2 videos to watch before proceeding further with the questionnaire. These videos, accessible to the public, were specifically chosen to provide a balanced depiction of what smart glasses are and their capabilities. It is worth noting, however, that as commercial videos, they may be designed to persuade viewers of the utility of smart glasses. The researchers selected these videos because they provided examples that were realistic and applicable in real-life work contexts at the time of the study and represented an attempt to prevent unrealistic imaging from smart glasses.

Although this model has been created based on previous literature and has remained as close as possible to the conceptual model, it is not a parsimonious model. From Cutting [56], a series of criteria can be derived to discuss the model evaluation, which is reflected by the accuracy, scope, and simplicity of the model. Parameter counting is widely used to indicate simplicity: if 2 models are both equally valid, the one with the fewest parameters is the best [56]. Several nonsignificant regression coefficients were found in the results of this study, and in combination with the previous recommendations, a subsequent simpler model can be developed.

Conclusions

Anticipating Adoption and Appropriation

The main finding is the answer to the following research question: “To what extent do constructs derived from a mediation perspective contribute to the established constructs of adoption to explain the intention to use smart glasses?” In the Introduction, the assumption was made that the target audience would likely be in an early stage of adoption of smart glasses. However, the responses gathered concerning knowledge and experience revealed that the majority of respondents demonstrated a fair understanding of smart glasses, and approximately half of them had already experimented with the technology.

The mediation perspective has provided insights into potentially useful variables for the anticipated adoption and social interaction of smart glasses. Social interaction with technology may be important because smart glasses are worn on the body. Venkatesh and Davis [52] found that the influence of subjective norm was only significant in mandatory usage contexts and not in voluntary contexts [52]. In this study, no obligatory or voluntary context was specifically referred to, yet subjective norm had a significant association with intention to use. That moderator can be added for control in a subsequent study, but it is also expected that with the current technology, such as smart glasses that are worn on the body, this is different than with personal computers. With smart glasses, there is human interaction where technology differently mediates in the physical world. In addition, social influences would have a nonsignificant influence on intention to use without moderators such as gender, age, experience, and use voluntariness [10]. This study showed that subjective norm and the lack of social presence influence

XSL•FO
RenderX

attitude and indirectly influence intention to use smart glasses in health care and can therefore make better predictions for anticipated use intention for smart glasses. Furthermore, despite many of the social issues addressed in the literature [21,26,31] and previous studies, most of the issues incorporated in this study did not have an effect on the anticipated intention to use smart glasses in the future. The findings reported here shed new light on the intention to use smart glasses in health care, where it could be argued that social interaction is perceived as an important aspect in providing care but not the most important construct for predicting future use. While this study does not provide evidence regarding the issues in social interaction surrounding smart glasses in health care as seen in the literature, a limitation in this study might be that respondents did not experience the use of smart glasses in real time. On the other hand, it may also be that the quality of care is more important than social interaction, and even if a patient feels uncomfortable when it comes to health, an intervention is sometimes necessary. Respondents might assume that smart glasses will help them in their work and that they will add value, regardless of the outcomes of social interaction.

Furthermore, although it is suggested that social acceptance is regularly included in evaluation sections in smart glasses research [18,57], it was observed that the adoption was only partly included in research designs in some studies. For example, in some cases, technology acceptance was evaluated by 3 constructs: perceived usefulness, perceived ease of use, and intention to use [58]. However, the first TAM [11] included attitude as well, and the more recent developments of the model included 13 factors influencing perceived usefulness and perceived ease of use and thereafter intention to use [15,59]. This example and current research in technology acceptance in related fields [60,61] show that there is more to study in adoption than 3 factors, and we argue to use all relevant aspects of adoption and acceptance frameworks to obtain more valid answers in acceptance research.

In summary, several theoretical perspectives have been used in this study, and it has been found that despite the concerns raised about social interaction with smart glasses, these aspects play important roles for health care professionals but are ultimately not decisive for the intention to use. As a result, there are fewer threats to implementation for health care than might be expected based on previous literature. It can also be concluded that the empirical findings of this study provide a new understanding of how the mediation perspective can increase the explained variance compared to existing constructs from the literature. This is why the mediation perspective adds new insights into a theoretical framework that can also be applied in other professional fields or related technologies.

Acknowledgments

This study would not have been possible without the respondents from a health care organization in the Netherlands and the workers of Amazon's Mechanical Turk.

Data Availability

The datasets generated during or analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

NZ led the design, data collection, and data analysis, and drafted the initial manuscript. SBA, OP, and PPV contributed to the design and data collection. OP contributed to data analysis as well. All authors reviewed the protocols, contributed to the conceptualization and methodology, and reviewed the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Prior research on the adoption of smart glasses.
[DOCX File , 50 KB - formative_v9i1e49610_app1.docx]

Multimedia Appendix 2
Item pool, descriptive statistics, factor loadings, squared multiple correlations, and Cronbach alpha of the observed indicators to explain and predict the anticipated use of smart glasses.
[DOCX File , 121 KB - formative_v9i1e49610_app2.docx]

Multimedia Appendix 3
Opening statement.
[DOCX File , 14 KB - formative_v9i1e49610_app3.docx]

References

XSL•FO
RenderX

1. Get Ready For Even More Google Glasshole Sightings. TechCrunch. 2013. URL: https://techcrunch.com/2013/01/28/glassholes/ [accessed 2024-12-26]
2. Google Glass advice: how to avoid being a glasshole. The Guardian. 2014. URL: https://www.theguardian.com/technology/2014/feb/19/google-glass-advice-smartglasses-glasshole [accessed 2024-12-26]
3. Hein DWE, Jodoin JL, Rauschnabel PA, Ivens BS. Are Wearables Good or Bad for Society?: An Exploration of Societal Benefits, Risks, and Consequences of Augmented Reality Smart Glasses. In: Wearable Technologies: Concepts, Methodologies, Tools, and Applications. Hershey, PA: IGI Global Scientific Publishing; 2018:1313-1337.
4. Hofmann B, Haustein D, Landeweerd L. Smart-glasses: exposing and elucidating the ethical issues. Sci Eng Ethics 2017 Jun 18;23(3):701-721. [doi: 10.1007/s11948-016-9792-z] [Medline: 27432401]
5. Kress B. See Through Optical Architectures for Wearable Displays. 2014 Presented at: Imaging and Applied Optics; July 13-17, 2014; Seattle, WA. [doi: 10.1364/AIO.2014.JTu1A.1]
6. Wortmeier A, Sousa Calepso A, Kropp C, Sedlmair M, Weiskopf D. Configuring augmented reality users: analysing YouTube commercials to understand industry expectations. Behaviour & Information Technology 2023 Jan 09;43(2):371-386. [doi: 10.1080/0144929X.2022.2163693]
7. Kolodzey L, Grantcharov PD, Rivas H, Schijven MP, Grantcharov TP. Wearable technology in the operating room: a systematic review. BMJ Innov 2016 Dec 21;3(1):55-63. [doi: 10.1136/bmjinnov-2016-000133]
8. Zuidhof N, Ben Allouch S, Peters O, Verbeek P. Defining smart glasses: a rapid review of state-of-the-art perspectives and future challenges from a social sciences' perspective. Augment Hum Res 2021 Oct 31;6(1):15. [doi: 10.1007/s41133-021-00053-3]
9. Lu J, Yao JE, Yu C. Personal innovativeness, social influences and adoption of wireless Internet services via mobile technology. The Journal of Strategic Information Systems 2005 Sep;14(3):245-268. [doi: 10.1016/j.jsis.2005.07.003]
10. Venkatesh V, Morris M, Davis G, Davis F. User acceptance of information technology: toward a unified view. MIS Quarterly 2003;27(3):425. [doi: 10.2307/30036540]
11. Davis FD, Bagozzi RP, Warshaw PR. User acceptance of computer technology: a comparison of two theoretical models. Management Science 1989 Aug;35(8):982-1003. [doi: 10.1287/mnsc.35.8.982]
12. Rogers E. Diffusion of Innovations, 3rd edition. New York, NY: Free Press; 1983.
13. Akbari M, Rezvani A, Shahriari E, Zúñiga M, Pouladian H. Acceptance of 5 G technology: mediation role of trust and concentration. Journal of Engineering and Technology Management 2020;57:101585 [FREE Full text] [doi: 10.1016/j.jengtecman.2020.101585]
14. Lee Y, Kozar KA, Larsen KR. The technology acceptance model: past, present, and future. CAIS 2003;12:1250. [doi: 10.17705/1CAIS.01250]
15. Venkatesh V, Bala H. Technology acceptance model 3 and a research agenda on interventions. Decision Sciences 2008 May 09;39(2):273-315. [doi: 10.1111/j.1540-5915.2008.00192.x]
16. Koelle M, El Ali A, Cobus V, Heuten W, Boll S. All about Acceptability?: Identifying Factors for the Adoption of Data Glasses. In: CHI '17: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. 2017 Presented at: 2017 CHI Conference on Human Factors in Computing Systems; May 6-11, 2017; Denver, CO. [doi: 10.1145/3025453.3025749]
17. Rauschnabel PA, He J, Ro YK. Antecedents to the adoption of augmented reality smart glasses: A closer look at privacy risks. Journal of Business Research 2018 Nov;92:374-384. [doi: 10.1016/j.jbusres.2018.08.008]
18. Nam C, Lee Y. Validation of the wearable acceptability range scale for smart apparel. Fash Text 2020 May 05;7(1):13. [doi: 10.1186/s40691-019-0203-3]
19. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly 1989 Sep;13(3):319. [doi: 10.2307/249008]
20. Bagozzi R. The legacy of the technology acceptance model and a proposal for a paradigm shift. JAIS 2007 Apr;8(4):244-254. [doi: 10.17705/1jais.00122]
21. Koelle M, Kranz M, Möller A. Don't look at me that way!: Understanding User Attitudes Towards Data Glasses Usage. In: MobileHCI '15: Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services. 2015 Presented at: 17th International Conference on Human-Computer Interaction with Mobile Devices and Services; August 24-27, 2015; Copenhagen, Denmark. [doi: 10.1145/2785830.2785842]
22. Rauschnabel P, Brem A, Ivens B. Mixing Physical and Virtual Realities: Implications of Augmented Reality Smart Glasses for the Management of Innovation. 2015 Presented at: AMA Summer Marketing Educators' Conference 2015; August 14-18, 2015; Chicago, IL.
23. Due B. The social construction of a Glasshole: Google Glass and multiactivity in social interaction. PsychNology Journal 2015;13(2-3):149-178 [FREE Full text]
24. Liinasuo M, Kuula T, Goriachev V, Helin K. Choosing between remote and face-to-face features in a test setting: methodology used in two usability and user experience case studies. Behaviour & Information Technology 2023 Jan 13;42(3):298-306. [doi: 10.1080/0144929x.2023.2165453]

XSL•FO
RenderX

25. Goken M, Basoglu A, Dabic M. Exploring adoption of smart glasses: Applications in medical industry. 2016 Presented at: Portland International Conference on Management of Engineering and Technology (PICMET); September 04-08, 2016; Honolulu, HI. [doi: 10.1109/picmet.2016.7806835]
26. Montero C, Alexander J, Marshall M, Subramanian S. Would you do that?: understanding social acceptance of gestural interfaces. In: MobileHCI '10: Proceedings of the 12th International Conference on Human Computer Interaction with Mobile Devices and Services. 2010 Presented at: 12th International Conference on Human Computer Interaction with Mobile Devices and Services; September 7-10, 2010; Lisbon, Portugal. [doi: 10.1145/1851600.1851647]
27. Weiser M, Gold R, Brown JS. The origins of ubiquitous computing research at PARC in the late 1980s. IBM Syst. J 1999;38(4):693-696. [doi: 10.1147/sj.384.0693]
28. Mann S. What I've learned from 35 years of wearing computerized eyewear. IEEE Spectr. 2013. URL: https://www.nxtbook.com/nxtbooks/ieee/spectrum_na_0313/index.php?startid=23#/1 [accessed 2024-12-26]
29. Verbeek P. Beyond interaction: a short introduction to mediation theory. Interactions 2015 Apr 27;22(3):26-31. [doi: 10.1145/2751314]
30. Zuidhof N, Ben AS, Peters O, Verbeek P. A theoretical framework to study long-term use of smart eyewear. In: UbiComp/ISWC '19 Adjunct: Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers. 2019 Presented at: 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and ACM International Symposium on Wearable Computers; September 9-13, 2019; London, UK. [doi: 10.1145/3341162.3348382]
31. Zuidhof N, Ben Allouch S, Peters O, Verbeek P. Perspectives on the acceptance and social implications of smart glasses: a qualitative focus group study in healthcare. International Journal of Human–Computer Interaction 2022 Aug 26;40(2):149-159. [doi: 10.1080/10447318.2022.2111046]
32. Ajzen I. The theory of planned behavior. Organizational Behavior and Human Decision Processes 1991 Dec;50(2):179-211. [doi: 10.1016/0749-5978(91)90020-t]
33. Zuidhof N, Ben Allouch S, Peters O, Verbeek P. Anticipated Acceptance of Head Mounted Displays: a content analysis of YouTube comments. 2019 Presented at: IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops); March 11-15, 2019; Kyoto, Japan. [doi: 10.1109/percomw.2019.8730658]
34. Han DD, Tom Dieck MC, Jung T. Augmented Reality Smart Glasses (ARSG) visitor adoption in cultural tourism. Leisure Studies 2019 Apr 21;38(5):618-633. [doi: 10.1080/02614367.2019.1604790]
35. Harms C, Biocca F. Internal consistency and reliability of the networked minds social presence measure. In: Alcaniz M, Rey B, editors. Seventh Annual International Workshop: Presence. Valencia: Universidad Politecnica de Valencia; 2004:246-251.
36. Karahanna E, Limayem M. E-mail and v-mail usage: generalizing across technologies. J of Organizational Comp & Elec Commerce 2000 Mar 1;10(1):49-66. [doi: 10.1207/s15327744joce100103]
37. Tenner E. Why Things Bite Back, Technology and the Revenge of Unintended Consequences. New York, NY: Vintage Books; 1996.
38. Verbeek P. Op de vleugels van Icarus. Hoe techniek en moraal met elkaar meebewegen. Rotterdam: Lemniscaat; 2014.
39. Merton RK. The unanticipated consequences of purposive social action. American Sociological Review 1936 Dec;1(6):894. [doi: 10.2307/2084615]
40. Spector P. Summated Rating Scale Construction, 82nd edition. Newbury Park, CA: Sage; 1992.
41. Tyupa S. A theoretical framework for back-translation as a quality assessment tool. New Voices in Translation Studies 2011;7(1):35-46 [FREE Full text]
42. Gemvision Healthcare platform | Augmented Reality Smart Glasses. YouTube. URL: https://youtu.be/N_QoKStFn_Q [accessed 2024-12-26]
43. M400 Telemedicine Medical. YouTube. URL: https://youtu.be/TyA_NVyK9iw [accessed 2024-12-26]
44. George D, Mallery M. SPSS for Windows Step by Step: A Simple Guide and Reference, 17.0 Update. Boston, MA: Allyn & Bacon, Inc; 2010.
45. Jöreskog KG. A general method for estimating a linear structural equation system*. ETS Research Bulletin Series 2014 Aug 08;1970(2):i-41. [doi: 10.1002/j.2333-8504.1970.tb00783.x]
46. Kline R. Principles and practice of structural equation modeling, 4th edition. New York, NY: Guilford Press; 2016.
47. Holbert RL, Stephenson MT. Structural equation modeling in the communication sciences, 1995-2000. Human Comm Res 2002 Oct;28(4):531-551. [doi: 10.1111/j.1468-2958.2002.tb00822.x]
48. Hoyle R, Panter A. Writing about structural equation models. In: Hoyle RH, editor. Structural equation modeling: Concepts, issues, and applications. Thousand Oaks, CA: Sage Publications; 1995:158-176.
49. Hooper D, Coughlan J, Mullen M. Structural equation modelling: guidelines for determining model fit. Electronic Journal of Business Research Methods 2008;6(1):53-60 [FREE Full text]
50. Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling: A Multidisciplinary Journal 1999 Jan;6(1):1-55. [doi: 10.1080/10705519909540118]
51. Fornell C, Larcker DF. Evaluating structural equation models with unobservable variables and measurement error. Journal of Marketing Research 1981 Feb;18(1):39. [doi: 10.2307/3151312]

XSL•FO
RenderX

52. Venkatesh V, Davis FD. A theoretical extension of the technology acceptance model: four longitudinal field studies. Management Science 2000 Feb;46(2):186-204. [doi: 10.1287/mnsc.46.2.186.11926]
53. Uhde A, Tretter S, von Terzi P, Koelle M, Diefenbach S, Hassenzahl M. Interaction in the Public: Aesthetics, Social Acceptability, and Social Context. Mensch und Computer 2021, Workshopband, Workshop on Interaction in the Public: Aesthetics, Social Acceptability, and Social Context. 2021. URL: https://alarithuhde.com/publications/Workshop-Interaction-in-the-Public.pdf [accessed 2024-12-26]
54. Aguinis H, Villamor I, Ramani RS. MTurk research: review and recommendations. Journal of Management 2020 Nov 19;47(4):823-837. [doi: 10.1177/0149206320969787]
55. Paolacci G, Chandler J. Inside the Turk. Curr Dir Psychol Sci 2014 Jun 03;23(3):184-188. [doi: 10.1177/0963721414531598]
56. Cutting JE. Accuracy, scope, and flexibility of models. J Math Psychol 2000 Mar;44(1):3-19. [doi: 10.1006/jmps.1999.1274] [Medline: 10733855]
57. Lee L, Hui P. Interaction methods for smart glasses: a survey. IEEE Access 2018;6:28712-28732. [doi: 10.1109/access.2018.2831081]
58. Lam K, Hang LL, Braud T, Hui P. M2A: A framework for visualizing information from mobile web to mobile augmented reality. 2019 Presented at: 2019 IEEE International Conference on Pervasive Computing and Communications (PerCom); March 11-15, 2019; Kyoto, Japan. [doi: 10.1109/percom.2019.8767388]
59. Lai P. The literature review of technology adoption models and theories for the novelty technology. JISTEM USP 2017 Jun 08;14(1):21-38. [doi: 10.4301/S1807-17752017000100002]
60. Ben Allouch S. The design and anticipated adoption of ambient intelligence in the home. University of Twente. 2008. URL: https://ris.utwente.nl/ws/files/6085272/thesis_Ben_Allouch.pdf [accessed 2024-12-26]
61. de Graaf M, Ben Allouch S, van Dijk J. Why Do They Refuse to Use My Robot?: Reasons for Non-Use Derived from a Long-Term Home Study. In: HRI '17: Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction. 2017 Presented at: 2017 ACM/IEEE International Conference on Human-Robot Interaction; March 6-9, 2017; Vienna, Austria. [doi: 10.1145/2909824.3020236]

Abbreviations

RMSEA: root mean square error of approximation
SRMR: standardized root mean square residual
TAM: Technology Acceptance Model
TAM2: Technology Acceptance Model 2
TLI: Tucker-Lewis Index
UTAUT: Unified Theory of Acceptance and Use of Technology

Edited by A Mavragani; submitted 03.06.23; peer-reviewed by D Nogueira-Leite, C Romare; comments to author 19.07.23; revised version received 01.08.23; accepted 28.11.24; published 11.02.25.

Please cite as:
Zuidhof N, Peters O, Verbeek PP, Ben Allouch S
Social Acceptance of Smart Glasses in Health Care: Model Evaluation Study of Anticipated Adoption and Social Interaction
JMIR Form Res 2025;9:e49610
URL: https://formative.jmir.org/2025/1/e49610
doi:10.2196/49610
PMID:

©Niek Zuidhof, Oscar Peters, Peter-Paul Verbeek, Somaya Ben Allouch. Originally published in JMIR Formative Research (https://formative.jmir.org), 11.02.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Designing Patient-Centered Interventions for Emergency Care: Participatory Design Study

Woosuk Seo[1], BS; Shruti Jain[2], BS; Vivian Le[1], BS; Jiaqi Li[1], MS; Zhan Zhang[3], PhD; Hardeep Singh[4], MD, MPH; Kalyan Pasupathy[5], PhD; Prashant Mahajan[6], MPH, MBA, MD; Sun Young Park[7], PhD

[1]School of Information, University of Michigan, Ann Arbor, MI, United States

[2]Computer Science and Engineering, University of Michigan, Ann Arbor, United States

[3]Seidenberg School of Computer Science and Information Systems, Pace University, New York, NY, United States

[4]Center for Innovations in Quality, Effectiveness and Safety (IQuESt), Michael E. DeBakey Veterans Affairs Medical Center and Baylor College of Medicine, Houston, TX, United States

[5]Biomedical and Health Information Sciences, University of Illinois Chicago, Chicago, IL, United States

[6]Department of Emergency Medicine, University of Michigan Medical School, Ann Arbor, MI, United States

[7]School of Information, Stamps School of Art and Design, University of Michigan, Ann Arbor, MI, United States

Corresponding Author:
Woosuk Seo, BS
School of Information
University of Michigan
105 S State St
Ann Arbor, MI
United States
Phone: 1 2063109264
Email: seow@umich.edu

Abstract

Background: Emergency departments (EDs) are high-pressure environments where clinicians diagnose patients under significant constraints, including limited medical histories, severe time pressures, and frequent interruptions. Current ED care practices often inadequately support meaningful patient participation. Most interventions prioritize clinical workflow and health care provider communication, inadvertently overlooking patients' needs. Additionally, patient-facing technologies in EDs are typically developed without meaningful patient input, leading to solutions that may not effectively address patients' specific challenges. To enhance both patient-centered care practices and the diagnosis process in EDs, patient involvement in technology design is essential to ensure their needs during emergency care are understood and addressed.

Objective: This study aimed to invite ED patients to participatory design sessions, identify their needs during ED visits, and present potential design guidelines for technological interventions to address these needs.

Methods: We conducted 8 design sessions with 36 ED patients and caregivers to validate their needs and identify considerations for designing patient-centered interventions to improve diagnostic safety. We used 10 technological intervention ideas as probes for a needs evaluation of the study participants. Participants discussed the use cases of each intervention idea to assess their needs during the ED care process. We facilitated co-design activities with the participants to improve the technological intervention designs. We audio- and video-recorded the design sessions. We then analyzed session transcripts, field notes, and design sketches.

Results: On the basis of ED patients' feedback and evaluation of our intervention designs, we found the 3 most preferred intervention ideas that addressed the common challenges ED patients experience. We also identified 4 themes of ED patients' needs: a feeling of inclusion in the ED care process, access to sources of medical information to enhance patient comprehension, addressing patient anxiety related to information overload and privacy concerns, and ensuring continuity in care and information. We interpreted these as insights for designing technological interventions for ED patients. Therefore, on the basis of the findings, we present five considerations for designing better patient-centered interventions in the ED care process: technology-based interventions should (1) address patients' dynamic needs to promote continuity in care; (2) consider the amount and timing of information that patients receive; (3) empower patients to be more active for better patient safety and care quality; (4) optimize human resources, depending on patients' needs; and (5) be designed with the consideration of patients' perspectives on implementation.

XSL•FO
RenderX

Conclusions: This study provides unique insights for designing technological interventions to support ED diagnostic processes. By inviting ED patients into the design process, we present unique insights into the diagnostic process and design considerations for designing novel technological interventions to enhance patient safety.

International Registered Report Identifier (IRRID): RR2-10.2196/55357

(JMIR Form Res 2025;9:e63610) doi:10.2196/63610

KEYWORDS

emergency department; participatory design; patient; technology; intervention

Introduction

Background

In the high-pressure environment of emergency departments (EDs), diagnosing and managing patients is inherently complex. This complexity arises from extensive interactions among various stakeholders, including health care providers, family caregivers, and patients. The dynamic and time-sensitive context of EDs forces stakeholders to make critical medical decisions promptly. The urgent nature of emergency care thus poses substantial challenges to ED providers, compounding the difficulty of achieving timely and accurate diagnoses and raising concerns regarding patient safety. While exact rates of diagnostic errors remain elusive, a conservative estimate of 5% errors in adults out of the 131 million annual ED visits translates to about 7 million cases of ED-based diagnostic errors [1]. Notably, nearly half of these diagnostic errors have the potential to cause harm to patients.

Many existing technological interventions focus on supporting the diagnostic process and promoting patient safety in the ED. For example, an ED provider-facing mobile app currently exists to deliver real-time patient information and allows for collaborative patient management for care teams [2]. Artificial intelligence (AI)–based software [3] has also been developed to provide practical guidance to novice nurses or scan operators in the ED who have limited experience in echocardiography. In addition, AI-based systems [4,5] have been used to support documentation for the nursing team, indirectly promoting the efficiency of the care process. Such systems could support providers by easing the documentation burden (eg, parsing long notes and automatically populating relevant text) so that they could focus more on patient care.

On the other hand, some technological interventions have been developed that focus more on ED patients. An indoor navigation system [6] was developed to assist patients unfamiliar with the ED environment, and a mobile display [7] was designed to present an interactive report detailing patients' progress, care plans, and care teams throughout their stay in the ED. To reduce patients' anxiety, some prior studies highlighted opportunities to use music apps [8] or virtual reality applications [9]. More recently, a study presented the opportunity for an AI-based application to provide medical triage to patients more efficiently [10]. Compared with human clinicians, the AI-based application made probable triage decisions. While these technological interventions improve providers' diagnosing work or patient safety, direct input from patients and providers working at the frontline is often overlooked when developing such solutions.

Participatory design (PD) [11] is a methodology that engages all stakeholders in the design process to create solutions that address their needs. PD has been widely used by researchers in prior studies to co-design user-centered technologies with health care providers. For instance, Kusunoki et al [12] conducted PD workshops with trauma team members to understand the different needs of awareness support among the various roles of team members and identify concrete design strategies to manage these differences. Pollack et al [13] organized a design session with 11 clinicians to develop a clinical information tool using PD techniques. On the basis of the session's findings, the authors identified benefits (eg, a high level of domain knowledge can be used to anticipate how design ideas can be applied to clinical processes and workflow) and potential challenges (eg, power dynamics between physicians) of leveraging PD techniques in designing a clinical information tool. The authors also outlined guiding principles for implementing these methods in health care organizations interested in advancing health information technology. These prior studies have presented how PD is helpful and practical for designing human-centered technology in health care settings.

Despite the benefits of PD in designing health technology, limited work has adopted a user-centered PD approach to develop technological interventions for ED specifically to enhance patient safety. Østervang et al [14] conducted PD workshops with health care providers and patients to design an ED information system. In the study, providers brainstormed initial design ideas for an ED information system in a workshop; then, patients provided feedback in separate workshops. The authors presented how the PD approach helps yield insights from ED providers and patients to create a more user-centered system. However, ED patients in the study had limited participation because the workshops were conducted one-on-one, and patients were only asked to provide feedback on intervention ideas developed by health care providers.

Objectives

Building on previous research, we aimed to engage ED patients in validating their needs and co-designing novel interventions to promote patient safety and enhance the ED diagnostic process. Because patients are vulnerable to potential errors, their involvement in developing patient-centered interventions is important to improve the diagnostic process. Through identifying patients' needs during their ED visits, this study presents design guidelines for more patient-centered technological interventions. We used preliminary technological intervention concepts to facilitate discussions about participants' experiences and needs. This approach helped expand our understanding of patient experiences in the ED. Rather than

XSL•FO
RenderX

proposing specific solutions, we aimed to refine intervention concepts based on patient input and establish broader guidelines for designing patient-facing interventions.

Methods

Overview

This study is part of a larger project that aims to identify ED patients' and health care providers' perspectives on technological interventions to support the ED diagnostic process and to craft design guidelines for interventions that meet both stakeholders' needs. In this paper, we focus on validating the needs of ED patients and caregivers during ED visits and their perspectives on technological interventions. We conducted 8 design sessions with 36 ED patients or their caregivers. On the basis of the analysis of transcripts, design sketches, and field notes, we identified 4 themes representing patient needs, strategies to mitigate common challenges, and design guidelines for potential interventions.

Design Idea Generation Phase for ED Care Interventions

Before our PD study with ED patients, the research team interviewed 8 patients and caregivers to better understand their experiences and challenges during ED visits. From the interviews, we identified difficulties, emerging patterns of complaints, and general levels of satisfaction with different aspects of the care process (Table 1). Common themes were discovered from each problem category through analysis of the interview data ranging from patient and caregiver difficulties, complaints, and satisfaction levels with various aspects of the ED care process. In addition to the known problems, such as long wait times and insufficient or poor communication with providers, ED patients and caregivers also faced difficulties because of the absence of caregivers, lack of ED process literacy, or inadequate understanding of discharge information. On the basis of the findings, the research team brainstormed numerous design ideas for each problem category, focusing on the stakeholders (eg, patients, caregivers, and providers). We then merged the design ideas based on feasibility and usefulness. Finally, we narrowed the list and finalized the 10 most effective intervention ideas (Table 2). These ideas were used to validate patient needs during each PD session, presented as storyboards, as shown in Figure 1. Each intervention idea aims to address at least one problem category.

Figure 1. Two sample storyboards of technological intervention ideas presented to the participants during the sessions (top: table-top robot; bottom: virtual specialist). ED: emergency department.

XSL•FO
RenderX

Table 1. A list of identified problem categories from the previous study's patient interview data.

Problem category	Examples
Waiting room general challenges	• Patients feel forgotten in the waiting room because health care providers do not check up on them • Patients are left in pain in the waiting room
Overcrowding challenges	• There are patients with mild symptoms who make the ED[a] more crowded
Information presentation and overload challenges	• Patients are inadequately informed of the ED process during the visit (ie, lack of ED process literacy) • Test results can be confusing for patients to understand because of the use of medical terminology
Information sharing challenges	• Patients often feel that there are errors in the physician's notes (eg, it is different from what was discussed in the visit) • Health care providers at hospitals sometimes wait for information transfer from the patient's original hospital, which can take time
Communication challenges	• Patients find it hard to recount all personal medical details in the ED because they are sick, stressed, or unaware • Patients sometimes have trouble communicating with providers because of challenges like being a nonnative English speaker
Contagious disease transfer in ED	• Crowded waiting rooms can have many sick patients in close proximity, which makes patients concerned • Contact with surfaces can cause disease transfer
Discharge and postvisit challenges	• Lack of postvisit resources for patients • Patients feel that they have been wrongly discharged

[a]ED: emergency department.

As probes, the storyboards incorporated user requirements and insights about patient-centered interventions to help participants better understand the use cases and contexts where the proposed interventions could be implemented. Each storyboard consisted of a 3-panel illustration, describing an example of a patient facing one of the challenges mentioned above and using the intervention to mitigate or prevent the problem (see Figure 1 for sample storyboards). Details of the intervention's features and functionality were left vague. We wanted participants to focus on the need rather than specific features to encourage more discussion about improving the intervention concept. Each storyboard included a lead question to highlight the user's needs (ie, patients) outlined in each intervention and to facilitate participant discussion about whether they had experienced the need, rather than directly referencing the intervention. Using storyboards allowed participants to have less biased and more open-ended discussions in the PD sessions. Table 2 describes each technological intervention and the patient need it addresses, along with storyboard samples.

XSL•FO
RenderX

Table 2. A list of 10 intervention ideas with brief descriptions and discussion questions to validate patient needs.

Intervention number	Technological intervention	Description	Sample of lead question used for the discussion of patient need
I-1	Virtual presence of a caregiver	A communication device for ED[a] patients that enables the virtual presence of the caregiver for the patient and their ED provider when the caregiver cannot physically be with the patient	Have you experienced any difficulty visiting the ED alone without a caregiver?
I-2	Transparent interactive display for ED-related guidance	An interactive display for ED patients in a waiting room that displays ED-related guidance and acts as a separator between patients for personal hygiene	During your ED visit, have you worried that you might be coming in contact with contagious diseases while in the waiting room?
I-3	Table-top robot for distraction and entertainment	A table-top robot for ED patients in the waiting room that provides distraction and entertainment as they wait	Have you felt distraction and entertainment is helpful during a long wait in the ED?
I-4	Wearable device for monitoring real-time condition	A wearable device for ED patients in the waiting room that tracks their real-time data (eg, vitals, such as heart rate and blood oxygen) and notifies providers if vitals indicate a patient needs immediate help	Have you felt worried about your condition worsening while waiting, and being unable to inform healthcare providers properly?
I-5	Assistant robot in the waiting room	A mobile robot that assists and checks on ED patients in the waiting room	Have you felt forgotten while waiting because of busy ED providers not checking up on your condition?
I-6	ED-specific chatbot for ED practice literacy	A mobile device–compatible chatbot that can provide ED-specific information directly from the hospital to patients	Have you ever felt confused about the meaning of the medical information you received during ED care?
I-7	Virtual communication with a specialist	A communication device for ED patients and providers that enables a virtual consultation with a specialist during ED care	Have you felt the need to consult with a specialist during your ED visit?
I-8	Pain expression device	A tactile device that helps ED patients understand and describe pain through vibrations and haptic feedback	Have you had any difficulties in describing your pain verbally?
I-9	Patient-facing screen for sharing physician's notes	A screen display that presents care-related information for patients to view during consultation or follow-up with providers about diagnosis or test results	Have you ever noticed errors or confusing details on physician's notes (usually on discharge notes or after-visit summaries)?
I-10	AI[b] caller for providing after-visit support	A follow-up call powered by AI to help answer patients' postdischarge questions regarding their ED visit	Have you had any challenges or questions that needed to be properly addressed after your ED visit?

[a]ED: emergency department.

[b]AI: artificial intelligence.

Participant Recruitment

We recruited patients and caregivers who had visited the ED within 6 weeks from the point of contact. Participants were mainly recruited from the adult and pediatric EDs at a university-affiliated hospital in 3 ways: calling or emailing them after their most recent ED visit, posting fliers with the study's contact information in the adult ED and the pediatric ED, and approaching patients during their ED visit. The inclusion criteria for participants were as follows: (1) aged ≥18 years old, (2) previously visited ED within 6 weeks from the point of contact, and (3) comfortable speaking English. During a patient's ED visit, researchers mainly focused on recruiting those having a triage level of 3 to 5, with 1 being critical and 5 being nonurgent, to avoid disrupting or risking their care. We contacted more than 150 potential participants via email and phone, as participant retention was challenging because of last-minute cancellations from patients. In total, 36 ED patients and caregivers participated in 8 design sessions (see Table 3 for their demographic information). All but one session had 4 or 5 participants, and one session had 3 participants. As compensation for participation, each participant was provided with a US $100 gift card for the session.

XSL•FO
RenderX

Table 3. Demographic information of study participants (N=36).

Participant demographics	Participants, n (%)
Race	
Asian	1 (3)
Black	7 (19)
White	27 (75)
Unknown	1 (3)
Sex	
Male	12 (33)
Female	24 (67)
Participant type	
Patient	23 (64)
Caregiver	13 (36)

PD Session Procedure

We conducted 8 PD sessions. Each session was roughly divided into three parts: (1) session introduction, (2) storyboard critique, and (3) co-design as a group. Each session took approximately 2 hours.

In the introductory session, participants were each given a paper packet outlining the current ED care process timeline, the 10 technological intervention storyboards for reference, and a sheet to rank the interventions based on their preferences. A presentation of the ED care process timeline (Figure 2) was shown to prompt participants to recall their previous or most recent ED experience. We created this timeline based on our last interview study [15] and prior studies on the ED care framework [16]. We first asked participants to write down one major challenge they encountered during this experience and specifically indicate where it occurred in the current ED care process timeline in their packet. Each participant was then asked to discuss their experiences with the group, establishing the foundation for the main session activities. Allowing participants to reflect on their previous or most recent ED visit prefaced discussions about how their challenges might or might not be addressed with one or more of the 10 technological interventions and the co-design activity.

Figure 2. A timeline of the emergency department (ED) care process illustrating steps before, during, and after care. This timeline, based on prior studies, was provided to the participants as a reference for recalling their ED visits.

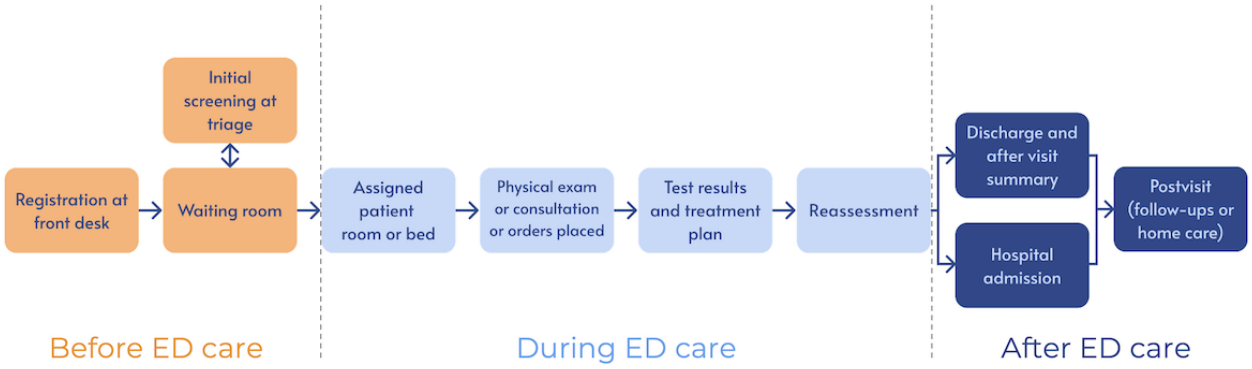

For the storyboard critique activity, 10 intervention design ideas were presented via storyboards to prompt participant discussion about each intervention's feasibility in the current ED care process. Researchers encouraged participants to initiate the conversation with lead questions and additional follow-up questions related to each scenario were asked, as necessary, to maintain the discussion within the group. The discussion of the 10 technological interventions provided a space for participants to speak about their experiences and needs during ED visits and for researchers to better understand the underlying challenges patients face. After all the intervention storyboards were presented, participants ranked their top 3 preferred interventions based on preference and potential feasibility in ED care. While the main task for each participant was to consider the interventions from the patient or caregiver perspective, participants naturally framed these interventions from the provider's perspective as well. We did not set any restrictions regarding this task during the discussions and design activities to solicit as much feedback as possible regarding the feasibility of each intervention in the ED care context.

Each group of participants discussed their most preferred 3 interventions and then selected one for the co-design activity. Then, as a group, participants had approximately 25 to 30 minutes to augment their chosen technological intervention to improve the ED experience (Figure 3). Participants were tasked with adding, changing, and removing features or modifying the context of the technological intervention to help it better align

XSL•FO
RenderX

with their preferred, ideal future implementation. Researchers emphasized the importance of focusing on the roles and functionality of technological interventions rather than the aesthetic quality of the designs. The co-design activity helped us understand the features of the intervention technologies participants desired to address patient needs, along with other fundamental and feature-specific aspects that may not have been revealed in the storyboard critique. Through this, participants could also discuss their decisions in detail, out loud, and reach a consensus about the most critical design features to incorporate into an ideal intervention solution. Following the co-design activity, participants were asked to provide a summary presentation to the researchers and explain their reasoning behind the design decisions they made together.

Figure 3. Co-design activity with the participants in the middle of a discussion about improving idea 7 (I-7 in Table 2). Participants collaboratively drew ideas on a whiteboard and wrote notes about imagined functionality.

Data Analysis

During the PD sessions, at least 2 researchers presented and took separate notes on points discussed among groups and overall interpretations of participant reactions to ideas. After every session, the researchers collaborated to reflect on session notes and compiled a document with observations on critical insights and sentiments toward each design idea. For each PD session, a researcher also took pictures of the sketches created during the design activity and on participant handouts, which contained participant notes, feedback, and intervention rankings. We used transcripts as the primary data source, and all other collected data were used as supplementary sources. We also incorporated the co-design activity sketches as supplementary data to validate what participants discussed during the session and clarify participants' design suggestions. The above data collection was conducted with the participant's permission.

After organizing the collected data, we coded the first 3 PD transcripts. At least 2 researchers coded each transcript, and each code was reviewed to account for missed information. The codes were compiled into a growing codebook organized by their corresponding intervention or put into a general category. On the basis of the codebook, the remaining transcripts were then coded. If any new codes were found, they were either added to the codebook as a new code or incorporated into the definition of an existing similar code.

We then used affinity diagramming [17] to identify the emerging themes from the codes. Similar codes from all 10 interventions were first grouped together into subthemes specific to each intervention. Then, a team of researchers had a series of discussions to build larger, emerging themes across interventions that revealed the most important patient needs from the study (see Multimedia Appendix 1 [18] for the screen capture of our affinity diagramming). For example, a larger theme of patient-provider communication contained various subthemes, such as patients being unable to notify providers and improving the communication of pain. Codes that were ambiguous or needed further support were refined by referencing participant quotes from transcripts and participants' sketches from PD sessions. After multiple group discussions, we identified 4 themes that described patients' key needs during their ED visits.

Ethical Considerations

We complied with the following ethical considerations. First, we recruited and obtained consent from all participants before the study sessions described in the multisite institutional review board approved by the University of Michigan (HUM00156261). Participants were informed that their participation was voluntary, the sessions would be recorded, the collected information would be deidentified and protected at a secured storage, and they would be compensated with US $100 for their participation. Second, we ensured that the collected transcript data were deidentified by replacing their names with pseudonyms, and

XSL•FO
RenderX

participants' faces in photos or videos were blurred. All collected data are stored in secured storage that requires the University of Michigan accounts. Third, all the researchers completed research compliance training such as the Collaborative Institutional Training Initiative or Program for the Education and Evaluation in Responsible Conduct of Research to become trained on best practices and ethical considerations of interacting with providers and patients.

Results

Overview

On the basis of the analysis of the collected data, we identified the 3 most preferred intervention ideas that participants preferred and 4 emerging themes that described the participants' needs in the ED care process.

Across all sessions, I-4 (ie, wearable device), I-7 (ie, virtual specialist), and I-1 (ie, virtual presence of a caregiver) emerged as the most preferred intervention ideas that the participants selected to help enhance the overall quality of patient care in the existing ED process. Among 36 participants, 25 (69%) chose I-4; 22 (61%) chose I-7; and 21 (58%) chose I-1 among their top 3 preferred interventions. These preferences represent which patient needs the participants considered most important. Through I-4, participants valued provider accountability and prioritization of patient conditions in the ED waiting room. For I-7, participants valued real-time virtual interactions with providers to better use their visit time while receiving the specialized care they needed. Furthermore, for I-1, participants valued the role of the caregiver for assistance and emotional support in the ED process even if they could not be physically present.

We also report 4 emerging themes of specific needs that participants had during their ED visits. These themes encompass broader perspectives of ED patients, extending beyond the technological intervention ideas provided. While describing these needs, we explain how participants addressed or wished to address them, along with their intervention ideas. Table 4 summarizes our key findings on patient needs and participants' suggestions for improving the intervention ideas.

Table 4. A summary of the 4 emerging themes based on patient needs during their ED[a] visit, discovered across all participatory design sessions, includes a description of each patient need and participant suggestions on how to address or improve the need.

Patient needs	Description	Patients suggestions for improvement
The feeling of inclusion in the ED care process	Patients feel that their voices are not heard by providers and want to be more involved in their care process	Create more communication channels with providers (eg, Help button)
Resources to improve patient comprehension of medical information	Patients feel inadequately literate about their health and want to have reliable ways to easily understand their condition, care plan, and treatment options	Use personalized information in intervention technology to enhance patients' understanding of their medical history
Relief of anxiety experienced by the patient over information overload and privacy concerns	Patients feel anxious because of privacy concerns and information overload in the ED. They strongly value emotional support from providers	Assist human-mediated emotional support, rather than replacing providers
Continuity in care and information	Patients perceive gaps in provider-to-provider communication that compromise care quality, emphasizing the need for improved technical infrastructure and procedures to ensure effective information delivery	Consider the accessibility of technology for different patient contexts

[a]ED: emergency department.

The Feeling of Inclusion in the ED Care Process

Throughout the 8 sessions, it is more than evident that participants wanted to be more included in the ED process. Participants often felt unheard, dismissed, and uninformed in their interactions with providers:

> *[The nurses] almost like, didn't believe me or didn't care, until they saw like, I have to pass out in front of them before they're going to do anything. And the second I passed out, I got moved to an actual room in the ED when I've been sitting in the hallway for hours.* [P13]

As described in the quote, participants sometimes felt that providers dismissed patient concerns unless they physically showed symptoms.

Their feeling of exclusion was caused mainly by the lack of timely and transparent communication with ED providers as they experienced providers not necessarily available to check on worsening patient conditions. Thus, it was challenging to inform providers of their health conditions:

> *Because there's a lot of times if you're on a two-hour wait and you're in the process of having a stroke, you're not able to get up and go back to the waiting room or the front desk, but it'd be good if something flagged that your condition was deteriorating.* [P24]

Because of the concern about potentially worsening conditions during the wait time, most participants favored I-4 (ie, wearable device) to track their conditions in real time and alert providers if necessary. In addition to the illustrated functionality, many participants suggested adding features such as pain-tracking and fall detection to the intervention, borrowing functionalities from other less popular intervention ideas, such as I-8 (ie, pain expression device) and I-5 (ie assistant robot). Some participants also thought that real-time tracking devices could be used to update patient priority in the posttriage stage:

> *If your condition is staying fine, or steady and stable and somebody else with a higher pain or priority*

XSL•FO
RenderX

> *could go first and you're still okay, you know something like that I could see being helpful.* [P3]

Participants felt that by using real-time tracking data, patients in critical condition could receive medical attention as soon as needed, ultimately improving the accuracy of the diagnosis.

Despite their general support for real-time tracking of symptoms, some participants expressed concerns about how providers could use such interventions. For instance, some participants were worried that providers would ignore tracking device emergency alerts in an already hectic environment. Others also expressed concern about providers becoming overreliant on tracking devices and potentially disregarding verbal patient feedback about worsening conditions if the device's data did not corroborate with patient concerns. Regardless of these worries, some participants pointed out that capturing real-time health data alerts could increase provider accountability in cases where patients felt ignored or misdiagnosed:

> *...if you're getting this information in real-time, then...people can be held to accountability. Because you go back and you look at their chart, you look at their record, "Hey, this person was in distress, and you neglected them," so I think in some ways, the AI could really assist in that because, again, there needs to be accountability to doctors and nurses and stuff that are just blowing you off.* [P18]

Participants also advocated for additional communication channels to create a more inclusive ED care process between patients and providers. A popular idea was incorporating a "help" button that participants could use to receive immediate medical attention from a provider, which they suggested could be implemented into one or many interventions (I-2: interactive display; I-3: table-top robot; I-4: wearable device; and I-5: assistant robot). Figure 4 demonstrates a sample drawing of an improved I-4 (ie, wearable device) with a "nurse help" button on the side. Moreover, many participants indicated that the I-4 (ie, wearable device) included a way to update their condition and pain levels as they changed during their visit, and some patients mentioned that concrete pain descriptors (eg, tactile or word based) could help them better describe their pain. Participants believed that having direct communication channels, such as a "help" button, with providers could lead to better care outcomes and higher trust in their care.

Figure 4. A sample drawing by a participant during a co-design activity. The drawing shows how I-4 (ie, a wearable device) could be improved with additional features such as the Nurse Help button.

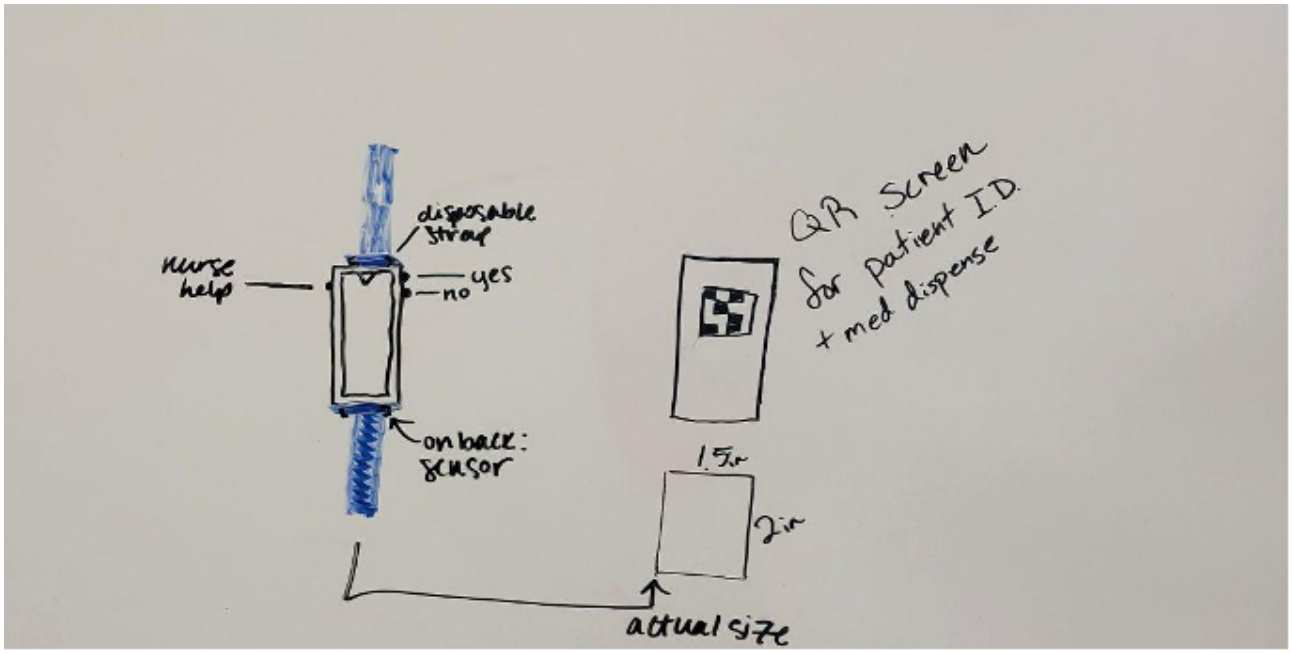

As our participant pool included caregivers and patients, the role of the caregiver in patient-provider communication was also emphasized. Caregivers are especially crucial as patients advocate for children and older adults, as they can significantly help patients who cannot correctly remember or communicate their condition in the ED. Participants were receptive to the idea of incorporating I-1 (ie, virtual caregiver) into the ED process when an in-person caregiver was unavailable, and some emphasized that a caregiver list could be incorporated into patient records for cases when the patient was too incoherent to call a caregiver.

Resources to Improve Patient Comprehension of Medical Information

Another theme we found was patients needing more information resources to improve their comprehension of care-related information. Participants expressed that their understanding of medical information required more alignment with their providers. Despite having access to their patient portal, most patients found it hard to familiarize themselves with the terminologies providers used in the ED when communicating about their condition or test results:

> *When I see...I think it's [a] diagnosis or something or like conditions I have...in my chart or portal or*

XSL•FO
RenderX

> *whatever. Like it says Latin...something just saying that my knees are bad...So I wish in parentheses it could say in English...the layman's term for that.* [P25]

A similar concern around patient health literacy arose when clarifying personal medical information. Participants said they often used Google (Google LLC) and other web resources to help understand their condition or diagnosis better while waiting for a provider's consultation. One participant expressed that it could be more accessible to conduct a Google search because patients can have a baseline understanding to prepare follow-up questions for providers. However, participants also noted that when seeking medical information on the internet in this way, they could be misled into thinking their condition had severe side effects or symptoms that were not identified or considered relevant to the provider's diagnosis.

To address the concern of potentially misleading diagnoses, participants frequently mentioned the necessity of incorporating personalized information into intervention technology to enhance patients' understanding of their medical history. Depending on the patient's level of health literacy, they might struggle with formulating questions or initiating conversations when communicating with their provider. In many design sessions, participants suggested that hospital-specific information can be curated based on the patient's medical history to provide relevant and reliable search results sourced directly from the ED that might otherwise be retrieved from a Google search. Participants specifically valued the features of the I-6 (ie, ED chatbot) and I-4 (ie, wearable device) for providing personalized patient information and helping effectively facilitate conversations from the patient to provider and with the patient themselves—such technology can break down complex medical terminology so that patients can follow up with providers in a more efficient way and aid in making informed decisions about their condition:

> *I think that it [I-6: ED chatbot] can be helpful to the emergency room staff or doctors because they could input the information that they want you [patients] to know...so they can control the information that's being given. Versus you [patients] googling it.* [P3]

Along with the difficulty in understanding medical information regarding their care, participants expressed the challenge of explaining their pain to providers. Issues with patient-provider communication arose when patients felt unable to properly convey pain levels, struggling with arbitrary scales from 1 to 10. Several participants mentioned that current pain scales do not provide adequate ways to express different types of pain:

> *Whenever they say on a scale of one to 10, I don't feel pain like other people.* [P15]

Some participants who struggled with chronic pain also recalled that current pain scales do not take into account that people might consider different levels of pain to be "normal." In such cases, the lack of medical information comprehension hinders patients from accurately describing their medical conditions (eg, pain) to providers. While many EDs ask about pain on a scale of 1 to 10, it is difficult to pinpoint or trace a patient's condition over time based on this abstract measurement:

> *The description of the pain varies for me.... And, I do live with pain.... All the rest of the pain is still there.* [P16]

Some participants found the intervention idea of an I-8 (ie, pain expression device) valuable in addressing this issue, particularly for patients who cannot communicate verbally with a provider or have a language barrier. Participants particularly liked the device's tactile capabilities as this feature could help anchor patients to a basic description or visualization of their pain to communicate with providers effectively. At the same time, some participants were also concerned about the ability of a patient to understand how a complex device works when in pain, further delaying their time in the ED and getting a proper diagnosis from providers:

> *If they [ED patients] are in pain already and we [the participants and researchers] have to stop and give an explanation of how to work it [a new intervention], I think it might present some difficulties, and there might actually end up being more of a barrier in terms of getting the information.* [P20]

Relief of Patient Anxiety Over Information Overload and Privacy Concerns

Many of our participants mentioned patient anxiety during their ED visit as one of their top concerns. On top of being admitted to the ED, participants experienced anxiety due to various factors beyond diagnosis, including long wait times, fear of contagious diseases (especially during and after the COVID-19 pandemic), and uncertainty about the subsequent steps in the ED process. They also noted that being overwhelmed with information as a patient in the ED aggravated their anxiety. Whether from a Google search or looking at test results, patients might not be able to understand all the information presented to them. Participants thought it was essential to consider the types of information that would be accessible—such as vitals or pain levels—to patients, providers, or both:

> *Do [patients] know how to read it? Do they know how to interpret the information that's given to them? Too much information for a patient can be bad. Then their mind wanders because they don't know.* [P30]

Potential anxiety induced by overwhelming information brought different perspectives regarding how much information should be displayed in the intervention ideas. Participants wanted to be informed but felt too much information could be unnecessary for patients who did not understand medical terminology:

> *I think that's a hard one. Because some people—it doesn't stress them out to know, and they like to know, other people can like to say their heart rates [are] high, but it can make it worse because then they're anxious about it.* [P13]

Participants across sessions discussed different scopes of information access available for each intervention idea. For example, patients might be presented with a fundamental view that only shows whether their vitals were in a normal range. At the same time, providers might have more detailed access to numeric values and other data that could help track patient conditions over time. Participants suggested this could be paired

XSL•FO
RenderX

with alerts on the I-4 (ie, wearable device) for the patient and provider, or providers only, to notify hospital staff in an emergency and better support care priority in the waiting room. Some participants even considered extending this idea for the entire duration of a patient's ED visit. Overall, participants suggested informing providers of a patient's condition would help mitigate patient anxiety in an ED setting.

The other factor that contributed to patient anxiety was related to patient privacy. Some participants felt that their health information was exposed to others in the waiting room when providers periodically checked in with them. Participants did not feel their privacy was protected during these discussions and thus felt their rights were being violated. Participants also attributed this feeling to more functional ED issues such as the waiting room's size, layout, structure, and overall ED design. With a small waiting room and open, shared space for beds in the ED, participants felt uncomfortable about exchanging confidential health information between patients and providers in open environments and public areas:

> *I heard everything that was going on. I could have told you what was going on....Because of our health and our health being we have rights for nobody to know.* [P16]

P20 also had a similar experience:

> *I'm hearing people's whole names being called out in this huge waiting room. I felt like the confidentiality was being breached because I knew that person was there getting care.*

To address this issue, many participants advocated for existing ED processes to be remediated at the structural level. This included the addition of more ED beds with closed walls or partitions, private spaces in the waiting room, and priority-based interventions. Overcrowded EDs induce long wait times and quickly fill up space in the waiting room where patient information is highly vulnerable to others. Some participants also suggested interventions, such as I-2 (ie, interactive display) and I-5 (ie, assistant robot), should avoid displaying personal information to protect patient privacy, primarily because these interventions would be concentrated in the waiting room.

To alleviate anxiety induced in general by multiple factors during ED visits, many participants stressed the importance of having human providers as an aspect of emotional support for patients in the ED care process. When presented with ideas that limited patient-to-provider interactions (eg, I-3: table-top robot and I-10: AI caller), participants expressed how technological interventions should support human providers rather than replace them. For this reason, interventions involving AI or robot technology were not popular with participants because they could have limitations in understanding patient requests or evaluating a patient's condition because of the lack of sophisticated technology in this space. Participants further highlighted empathy as another critical factor to consider because technology in the form of a robot would not convey as much sympathy and support as a human would in a stressful ED situation. Many participants thus preferred to have human support for alleviating their anxiety along with technological interventions because humans can communicate medical information and reassure patients in a manner that helps in coping with an anxiety-inducing environment:

> *I don't think people would have a lot of confidence in [I-5: assistant robot] either. You know what I'm saying—it needs to be people who still like to interact.* [P18]

Continuity in Care and Information

During group discussions, many participants shared their experiences with discontinuity in care and information. On the basis of their experience, participants perceived that discontinuity might have occurred due to the handoff process within the care team. They noted that the involvement of multiple providers in their care team meant that information would often get lost during handoffs. Because of such missing information, some participants experienced discrepancies in provider-to-provider communication that caused patient confusion:

> *Even a stay in the hospital, the doctors just rotate and roll over.... The first doctor said they wouldn't consider [discharging] until it [the protein level] goes below 10,000. Then the next doctor said, "Oh no, I'm not going to even think about it unless it's below 500." So, each time the shift changed—they told us something different.* [P10]

Such discrepancies in care also occurred between hospitals. For instance, P27 shared her experience when she received a different diagnosis in the second hospital:

> *So if he [P27's son] would have asked for a second opinion, at that point, you're leaving me in a lot of pain, and I still don't know what's going on. If he could have asked for that second opinion at that point, maybe he would have been diagnosed with a real issue five days earlier than he was.* [P27]

Reflecting on those situations where information discontinuity had occurred, most participants favored the I-7 (ie, virtual specialist) to improve the breakdowns. They insisted that for interventions such as I-7 (ie, virtual specialist), an ED physician should be present when the patient talks to the specialist so that the ED care team could have updated information. They envisioned that bringing specialists directly to the ED care process would improve diagnosis and documentation in patient care records with the second opinion. Participants also expressed the need to allow for secure but easy information sharing with providers outside the hospital when external providers were invited to the ED care process. For instance, a global access health care platform could be used across hospitals for patient medical data transfer. Therefore, they thought that I-7 (ie, virtual specialist) could be expanded to include medical professionals beyond certain specialists:

> *I think it depends on the context, the question, but for certain questions, it could be any nurse, it could be a PA [physician assistant], or someone who had familiarity with the patient. Based on the level of question, it may need to be a provider. Or a pharmacist.* [P27]

XSL•FO
RenderX

To promote continuity of care, participants commented that the technological interventions should be designed in an accessible way for different patient situations. For example, some patients wanted to use their existing wearable devices like smartwatches or fitness trackers instead of wearables provided by the ED to integrate live vitals tracking with historical data accurately. They also suggested that the I-6 (ie, ED chatbot) could be offered as an app for mobile devices that patients and caregivers could download when they arrived at the ED and use throughout their ED stay and after their ED care. These participants expected that using interventions with patients' devices would ensure that patient-collected health data would be used to its fullest extent during care in the ED and would allow providers to glimpse the patient's complete health profile over time rather than only view current condition. On the other hand, some other participants emphasized the need for care continuity for those who may not have the same access to technology (eg, smartphones and smartwatches) or may have different amounts of previously collected health data, for instance, patients with chronic illnesses may have more personal health data tracked than those without. Thus, it is important that intervention use did not become a privilege reserved only for a subset of patients who had access to certain technological and data resources:

> *I think that one, look at the person that comes in without any communications, no phone or no computer, you know, nothing.... I often go into situations and emergencies without any kind of feature.... I carry a phone—an old-fashioned flip phone with me. I can't access some of those things.* [P20]

Participants' desire for continuity of care was also shown during their co-design activity when they worked to improve the intervention designs. They envisioned an integrative, connected system where multiple intervention ideas were integrated with the current health portal or text notification system to flow health-related information continuously. For example, they envisioned a tablet that could be used for I-7 (ie, virtual specialist), displaying vitals synchronized with I-4 (ie, wearable device) or the patient's health device (Figure 4). Through such integration of features, participants envisioned inventions to enhance the continuity of care and information and to address their complex challenges in the ED care process. Although a few participants felt concerned that such an integrated intervention could delay ED stay time or contribute to patient information overload, many still thought our suggested interventions could be integrated with the existing patient portal and offered in an app format on various devices.

Discussion

Design Implications for Patient-Centered Interventions in ED

On the basis of the PD sessions, we identified ED patients' needs in the ED care process and expectations for future technological interventions. This section discusses how these findings may lead to design guidelines for more patient-centered interventions in EDs.

First, technology-based *interventions should address patients' dynamic needs to promote continuity in care*. Our participants found that their needs could be met by more than one intervention idea or multiple specific features because their complex needs may change depending on where they stand in the ED care process. For instance, many participants suggested integrating critical functions into one system in the co-design activity so that the system could provide necessary and relevant information to them, from arrival at the ED to postvisit follow-up. This desire was evident when some participants envisioned an integrated system of 2 different intervention ideas (I-7: virtual specialist and I-4: wearable device) that could track patients' status and share data with specialists during the diagnosis process. Such integrated interventions may also be used to support issues in patient-provider communication. Our findings demonstrated that patients lacked sufficient information and interactions with ED providers. Thus, our participants shared how integrated interventions may help fill this information gap by providing relevant communication channels or resources based on patients' dynamic needs during ED visits. However, mere integration of systems may lead to additional challenges for patients. As some of our participants highlighted, numerous resources may cause information overload that can lead to patient anxiety. In another case, inviting outside providers (ie, specialists) would require extra documentation, potentially resulting in additional patient burdens. To incorporate a patient-centered integrated system in the ED, we suggest that the system be capable of adapting to patients' needs throughout the ED care timeline rather than simply adding features into one system. Existing technologies are primarily designed to provide specific information to patients in a certain part of the ED process (eg, a mobile app for an interactive report [7] to inform patients about their real-time progress and care plans during their visits). Such technologies can be integrated with other features, such as I-10 (ie, AI caller) for after-visit support, to address the dynamically evolving needs of patients that extend beyond their stay in the ED. An integrated system that supports patients starting from triage to beyond their discharge would provide timely, relevant information to patients based on their status within the ED care process. This integration approach to promoting continuity in care resonates with prior studies that designed a multicomponent app for patients and caregivers, the Bone Marrow Transplant Roadmap [19]. The app included components that reflect the information needs of patients and caregivers during their clinic visits, such as an overview of the criteria needed for discharge, real-time laboratory results, personalized medication lists, and educational materials. This app was further expanded to include outpatient settings to better support caregivers [20]. Similar to these multicomponent approaches, we suggest the integration of various features to better meet ED patients' dynamic needs during or even after their ED visits.

Second, *interventions should account for both the amount and timing of information provided to patients*. Our participants shared their challenges in accessing the appropriate amount of information during their ED visits. They felt excluded from the care process when they received insufficient information, while overwhelming information heightened their anxiety. Thus, participants suggested that the patients should have the decision

XSL•FO
RenderX

to choose the timing and extent of information they receive for applicable interventions, such as opting for partial, complete, or no information access. One example of this suggestion was using a chatbot to provide patients with personalized care information. On the basis of the participants' ideas, I-6 (ie, ED chatbot) could be further developed to offer relevant information depending on the steps patients undergo in the care process and their current health condition. Prior studies have shown that chatbots can provide more accessible and relevant information to patients [21]. In cancer contexts, a chatbot provided answers to patients with breast cancer with similar satisfaction as physicians [22], while another chatbot collected patient-reported symptoms during chemotherapy and reduced ED visits and unscheduled hospitalizations [23]. While these studies presented the opportunities for chatbots to provide information to patients, our findings suggest more careful considerations on the timing of providing information to patients. The need for timely and appropriate information resonates with the findings of a prior study [24], which describes how temporal dependency can delay the information that ED patients need. As ED medical activities often depend on previous activities, the information patients need or want is usually delayed. Without having relevant information during the ED process, patients may experience communication challenges with ED providers, such as different perceptions of what information is critical to ED care [25]. Therefore, as our participants suggested, interventions should provide ED patients with optimized information relevant to the current step of the care process or options to control the amount and timing of information they expect to receive.

Third, *interventions should empower patients to be more active in their care process for better patient safety and care quality*. Our findings highlighted participants' need for patient autonomy in ED care. Unlike other patient care settings, the ED context is particularly intimidating for patients because ED visits are unexpected and urgent. Each visit may be different from the previous one. However, we found that participants in our study were eager to take a more active role in the ED care process. In the PD sessions, the need for empowerment was evident when most participants suggested features that allowed them to communicate assertively with providers about their symptoms. For instance, although I-4 (ie, wearable device) was initially designed to track patients' condition so they would not feel forgotten while waiting, many participants wanted to have another function, similar to a "help" button, to request support more actively from providers when their conditions worsened or changed. This suggestion indicates how ED patients want to be more active stakeholders in the ED care process rather than be considered passive stakeholders who are just being assessed for their symptoms. Patient-empowering interventions are implemented on the basis of the prior studies' insights from existing interventions, such as an inpatient portal [26] and a smartphone app [27]. For instance, a smartphone app called MySurgery [27] was developed to provide information about surgery risks and practical step-by-step advice for each risk. The app demonstrated a significant ability to enable patients to engage actively in discussions regarding their care and adopt behaviors related to their safety. Similarly, ED patients can be empowered with relevant information and advice, such as explaining medical terms or notices about the next step in the ED care process. This need for patient empowerment resonates with previous studies on shared decision-making in the ED [28,29]. These studies identified that ED patients want some degree of involvement in medical decision-making. Extending this line of study, we found contextual nuances and more concrete examples of ED interventions to promote patient involvement. For instance, a review study showed the opportunity for wearable devices to monitor patient vital signs and provide the collected data to providers to improve decision-making [30]. Our findings suggest how ED patients perceive such wearable devices as an opportunity to be more actively involved in the care process rather than letting the devices passively collect data. Therefore, we suggest that future ED interventions be designed to empower patients to be included in shared decision-making.

Fourth, *interventions should optimize human resources, depending on patients' needs*. We identified that some issues cannot be addressed through technological interventions alone. These issues may be related to distress or emotional needs of ED patients. For instance, our participants preferred human support rather than interactions with technical tools in alleviating their feeling of anxiety during ED visits. This preference implies that ED patients can manage their stress through human interactions, although the suggested interventions, such as I-5 (ie, assistant robot), can still provide medical information. Moreover, human resources for ED patients can include those who are not ED providers. Many participants expressed their need to connect with informal caregivers (eg, family members), specialists, or other providers during their ED visit. Their needs were particularly evident when patients visited the ED once because they did not know about the ED care process. On the basis of their previous visits, patients learned that some external resources (eg, family members and specialists) could provide more information to help assess their conditions or, in the case of family members, provide emotional support during the ED visit. This finding indicates that ED patients leverage resources they are more familiar with in the intimidating and unfamiliar ED settings. Prior studies in the medical literature have shown the opportunities and challenges of telehealth for patients [31-33]. However, those identified challenges (eg, unsteady or poorly framed video [33]) and strategies to overcome them may not be implemented in the fast-paced ED care process. Depending on the patient's current conditions, the right timing (eg, when a specialist can be contacted) and method (eg, simple video call or additional camera to show the patient's status) for providing external resources to patients should be determined. Thus, designing patient-centered interventions should optimize human resources to address ED-specific challenges, as we identified in our study. It will be critical to identify when and how interventions should provide external resources to patients. For instance, some of our participants wanted to be connected with specialists when they met ED providers after examinations so that they could have second opinions on their symptoms.

Finally, *interventions should be designed considering patients' perspectives on implementation*. Through co-design activities, we identified how our participants expected intervention ideas to be implemented to address their needs better. However, their expectations could differ from actual implementations,

XSL•FO
RenderX

potentially conflicting with what providers or the hospital expect. In such cases, it is critical to mitigate those expectations to implement patient-centered interventions. For instance, one of the potential reasons for our participants' preference for I-7 (ie, virtual specialist) was the assumption that specialists are available to have on-demand consultations. In contrast, in actual implementations, it would require other resources (eg, cost, time, and human resources) to implement a system to become connected with specialists. As a prior study shows, ED providers must modify their work practices to adopt the new interventions [34]. Moreover, some of our participants pointed out the accessibility of devices that ED patients may not have. While mobile device ownership is high among ED patients [35], our suggested intervention ideas involved devices that patients may not have, such as wearable trackers. Thus, it is essential to consider how ED patients perceive the implementation of technological interventions in the ED.

Limitations

This study has some limitations. Given their qualitative nature, the findings are specific to the context we examined. Providers and patients from other EDs may face different challenges. However, our findings will contribute to identifying potential interventions that address such challenges through a PD approach. In addition, we recruited participants with varying backgrounds from 2 ED units (pediatric and adult EDs), which could help with the study's generalizability. Another limitation is that we did not distinguish participants' previous experiences as ED patients or caregivers. While the ED experiences of patients and caregivers do not have noticeable differences, future work can explore caregiver-specific perspectives on interventions designed to address their challenges in the ED care process. Finally, we only presented concepts, not interactive system prototypes. Therefore, participants may not be quite familiar with some intervention ideas. They could only imagine how an intervention might work but could not see the real system or product. However, storyboards are a common Human-Computer Interaction methodology for eliciting user needs in the early system design stage. They are useful for validating user needs and expectations before implementing a system in a complex environment similar to an ED.

Conclusions

In this PD study, we invited ED patients to design sessions to assess their needs in the ED care process and co-design technical interventions. On the basis of the analysis of collected data, we identified 4 themes related to ED patients' needs: the feeling of exclusion from the ED care process, limited resources for patient comprehension of medical information, anxiety about overwhelming information and privacy, and discontinuity in care and information. Concerns were also expressed about the accuracy of the diagnosis. These findings also informed us to develop guidance for designing future technology-based patient-centered interventions to improve the diagnostic process in the ED. To refine the identified design guidelines, we aim to conduct a PD study with ED health care providers. By identifying similarities and differences between ED patients and providers, we expect to present concrete design guidelines for technological interventions to support the ED diagnostic process and patient safety.

Acknowledgments

This project was funded under grant number R18HS026622 from the Agency for Healthcare Research and Quality (AHRQ), U.S. Department of Health and Human Services (HHS). HS is supported in part by the Houston Veterans Administration (VA) Health Systems Research and Development (HSR&D) Center for Innovations in Quality, Effectiveness and Safety (CIN13–413), the VA National Center for Patient Safety, and the Agency for Healthcare Research and Quality (R01HS028595, R18HS029347, and R01HS029318). The authors are solely responsible for this document's contents, findings, and conclusions, which do not necessarily represent the views of AHRQ. Readers should not interpret any statement in this report as the official position or policy of the US government, AHRQ, HHS, or VA. None of the authors have any affiliation or financial involvement that conflicts with the material presented in this report.

Data Availability

The datasets generated and analyzed during this study are not publicly available because the consent form for the participants stated that the data could only be accessible to the research team.

Conflicts of Interest

None declared.

Multimedia Appendix 1
A screen capture of our affinity diagramming. We used the Miro board [18] for affinity diagramming.
[PDF File (Adobe PDF File), 161 KB - formative_v9i1e63610_app1.pdf]

References

1. Newman-Toker DE, Peterson SM, Badihian S, Hassoon A, Nassery N, Parizadeh D, et al. Diagnostic Errors in the Emergency Department: A Systematic Review. Rockville, MD: Agency for Healthcare Research and Quality (US); Dec 2022.

XSL•FO
RenderX

2. Ehrler F, Tuor C, Rey R, Siebert JN. A mobile app to improve patient management in emergency departments: caregiver needs analysis, design and early technology acceptance assessment. Stud Health Technol Inform 2021 Oct 27;285:233-238. [doi: 10.3233/SHTI210605] [Medline: 34734879]
3. Scheder-Bieschin J, Blümke B, de Buijzer E, Cotte F, Echterdiek F, Nacsa J, et al. Improving emergency department patient-physician conversation through an artificial intelligence symptom-taking tool: mixed methods pilot observational study. JMIR Form Res 2022 Feb 07;6(2):e28199 [FREE Full text] [doi: 10.2196/28199] [Medline: 35129452]
4. Murray L, Gopinath D, Agrawal M, Horng S, Sontag D, Karger DR. MedKnowts: unified documentation and information retrieval for electronic health records. In: Proceedings of the 34th Annual ACM Symposium on User Interface Software and Technology. 2021 Presented at: UIST '21; October 10-14, 2021; Virtual Event. [doi: 10.1145/3472749.3474814]
5. Greenbaum NR, Jernite Y, Halpern Y, Calder S, Nathanson LA, Sontag DA, et al. Improving documentation of presenting problems in the emergency department using a domain-specific ontology and machine learning-driven user interfaces. Int J Med Inform 2019 Dec;132:103981. [doi: 10.1016/j.ijmedinf.2019.103981] [Medline: 31605881]
6. Ho TW, Tsai CJ, Hsu CC, Chang YT, Lai F. Indoor navigation and physician-patient communication in emergency department. In: Proceedings of the 3rd International Conference on Communication and Information Processing. 2017 Presented at: ICCIP '17; November 24-26, 2017; Tokyo, Japan. [doi: 10.1145/3162957.3162971]
7. Vardoulakis LP, Karlson A, Morris D, Smith G, Gatewood J, Tan D. Using mobile phones to present medical information to hospital patients. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2012 Presented at: CHI '12; May 5-10, 2012; Austin, TX. [doi: 10.1145/2207676.2208601]
8. Chai PR, Schwartz E, Hasdianda MA, Azizoddin DR, Kikut A, Jambaulikar GD, et al. A brief music app to address pain in the emergency department: prospective study. J Med Internet Res 2020 May 20;22(5):e18537 [FREE Full text] [doi: 10.2196/18537] [Medline: 32432550]
9. Sikka N, Shu L, Ritchie B, Amdur RL, Pourmand A. Virtual reality-assisted pain, anxiety, and anger management in the emergency department. Telemed J E Health 2019 Dec 01;25(12):1207-1215. [doi: 10.1089/tmj.2018.0273] [Medline: 30785860]
10. Delshad S, Dontaraju VS, Chengat V. Artificial intelligence-based application provides accurate medical triage advice when compared to consensus decisions of healthcare providers. Cureus 2021 Aug;13(8):e16956 [FREE Full text] [doi: 10.7759/cureus.16956] [Medline: 34405077]
11. Muller MJ, Kuhn S. Participatory design. Commun ACM 1993 Jun;36(6):24-28. [doi: 10.1145/153571.255960]
12. Kusunoki D, Sarcevic A, Zhang Z, Yala M. Sketching awareness: a participatory study to elicit designs for supporting ad hoc emergency medical teamwork. Comput Support Coop Work 2015 Feb 01;24(1):1-38 [FREE Full text] [doi: 10.1007/s10606-014-9210-5] [Medline: 25870498]
13. Pollack AH, Miller A, Mishra SR, Pratt W. PD-atricians: leveraging physicians and participatory design to develop novel clinical information tools. AMIA Annu Symp Proc 2016;2016:1030-1039 [FREE Full text] [Medline: 28269900]
14. Østervang C, Lassen A, Schmidt T, Coyne E, Dieperink KB, Jensen CM. Development of a health information system to promote emergency care pathways: a participatory design study. Digit Health 2022 Dec 26;8:20552076221145856 [FREE Full text] [doi: 10.1177/20552076221145856] [Medline: 36601282]
15. Daniel M, Park S, Seifert CM, Chandanabhumma PP, Fetters MD, Wilson E, et al. Understanding diagnostic processes in emergency departments: a mixed methods case study protocol. BMJ Open 2021 Sep 24;11(9):e044194 [FREE Full text] [doi: 10.1136/bmjopen-2020-044194] [Medline: 34561251]
16. Mahajan P, Mollen C, Alpern ER, Baird-Cox K, Boothman RC, Chamberlain JM, et al. An operational framework to study diagnostic errors in emergency departments: findings from a consensus panel. J Patient Saf 2021 Dec 01;17(8):570-575. [doi: 10.1097/PTS.0000000000000624] [Medline: 31790012]
17. Scupin R. The KJ method: a technique for analyzing data derived from Japanese ethnology. Hum Org 2008 Jan 22;56(2):233-237. [doi: 10.17730/humo.56.2.x335923511444655]
18. Miro: The Innovation Workspace. Miro. URL: https://miro.com/ [accessed 2025-02-04]
19. Runaas L, Hanauer D, Maher M, Bischoff E, Fauer A, Hoang T, et al. BMT roadmap: a user-centered design health information technology tool to promote patient-centered care in pediatric hematopoietic cell transplantation. Biol Blood Marrow Transplant 2017 May;23(5):813-819 [FREE Full text] [doi: 10.1016/j.bbmt.2017.01.080] [Medline: 28132870]
20. Chaar D, Shin JY, Mazzoli A, Vue R, Kedroske J, Chappell G, et al. A mobile health app (Roadmap 2.0) for patients undergoing hematopoietic stem cell transplant: qualitative study on family caregivers' perspectives and design considerations. JMIR Mhealth Uhealth 2019 Oct 24;7(10):e15775 [FREE Full text] [doi: 10.2196/15775] [Medline: 31651402]
21. Clark M, Bailey S. Chatbots in health care: connecting patients to information. Can J Health Technol 2024 Jan 22;4(1). [doi: 10.51731/cjht.2024.818]
22. Bibault JE, Chaix B, Guillemassé A, Cousin S, Escande A, Perrin M, et al. A chatbot versus physicians to provide information for patients with breast cancer: blind, randomized controlled noninferiority trial. J Med Internet Res 2019 Nov 27;21(11):e15787 [FREE Full text] [doi: 10.2196/15787] [Medline: 31774408]
23. Huang MY, Weng CS, Kuo HL, Su YC. Using a chatbot to reduce emergency department visits and unscheduled hospitalizations among patients with gynecologic malignancies during chemotherapy: a retrospective cohort study. Heliyon 2023 May;9(5):e15798 [FREE Full text] [doi: 10.1016/j.heliyon.2023.e15798] [Medline: 37206031]

XSL•FO
RenderX

24. Park SY, Chen Y. Patient strategies as active adaptation: understanding patient behaviors during an emergency visit. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. 2017 Presented at: CHI '17; May 6-11, 2017; Denver, CO. [doi: 10.1145/3025453.3025978]
25. Park SY, Chen Y, Raj S. Beyond health literacy: supporting patient-provider communication during an emergency visit. In: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing. 2017 Presented at: CSCW '17; February 25-March 1, 2017; Portland, OR. [doi: 10.1145/2998181.2998357]
26. McAlearney AS, Fareed N, Gaughan A, MacEwan SR, Volney J, Sieck CJ. Empowering patients during hospitalization: perspectives on inpatient portal use. Appl Clin Inform 2019 Jan 13;10(1):103-112 [FREE Full text] [doi: 10.1055/s-0039-1677722] [Medline: 30759491]
27. Russ S, Latif Z, Hazell AL, Ogunmuyiwa H, Tapper J, Wachuku-King S, et al. A smartphone app designed to empower patients to contribute toward safer surgical care: community-based evaluation using a participatory approach. JMIR Mhealth Uhealth 2020 Jan 20;8(1):e12859 [FREE Full text] [doi: 10.2196/12859] [Medline: 31958067]
28. Schoenfeld EM, Kanzaria HK, Quigley DD, Marie PS, Nayyar N, Sabbagh SH, et al. Patient preferences regarding shared decision making in the emergency department: findings from a multisite survey. Acad Emerg Med 2018 Oct 19;25(10):1118-1128 [FREE Full text] [doi: 10.1111/acem.13499] [Medline: 29897639]
29. Schoenfeld EM, Goff SL, Downs G, Wenger RJ, Lindenauer PK, Mazor KM. A qualitative analysis of patients' perceptions of shared decision making in the emergency department: "let me know I have a choice". Acad Emerg Med 2018 Jul 07;25(7):716-727 [FREE Full text] [doi: 10.1111/acem.13416] [Medline: 29577490]
30. Nino V, Claudio D, Schiel C, Bellows B. Coupling wearable devices and decision theory in the United States emergency department triage process: a narrative review. Int J Environ Res Public Health 2020 Dec 21;17(24):9561 [FREE Full text] [doi: 10.3390/ijerph17249561] [Medline: 33371223]
31. Donelan K, Barreto EA, Sossong S, Michael C, Estrada JJ, Cohen AB, et al. Patient and clinician experiences with telehealth for patient follow-up care. Am J Manag Care 2019 Jan;25(1):40-44 [FREE Full text] [Medline: 30667610]
32. Powell RE, Henstenburg JM, Cooper G, Hollander JE, Rising KL. Patient perceptions of telehealth primary care video visits. Ann Fam Med 2017 May 08;15(3):225-229 [FREE Full text] [doi: 10.1370/afm.2095] [Medline: 28483887]
33. Wootton AR, McCuistian C, Legnitto Packard DA, Gruber VA, Saberi P. Overcoming technological challenges: lessons learned from a telehealth counseling study. Telemed J E Health 2020 Oct 01;26(10):1278-1283. [doi: 10.1089/tmj.2019.0191] [Medline: 31800368]
34. Park SY, Chen Y. Adaptation as design: learning from an EMR deployment study. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2012 Presented at: CHI '12; May 5-10, 2012; Austin, TX. [doi: 10.1145/2207676.2208361]
35. Kim E, Torous J, Horng S, Grossestreuer AV, Rodriguez J, Lee T, et al. Mobile device ownership among emergency department patients. Int J Med Inform 2019 Jun;126:114-117. [doi: 10.1016/j.ijmedinf.2019.03.020] [Medline: 31029252]

Abbreviations

AI: artificial intelligence
ED: emergency department
PD: participatory design

Edited by A Mavragani; submitted 25.06.24; peer-reviewed by K Fuji, A Londral; comments to author 30.08.24; revised version received 31.10.24; accepted 21.12.24; published 12.02.25.

Please cite as:
Seo W, Jain S, Le V, Li J, Zhang Z, Singh H, Pasupathy K, Mahajan P, Park SY
Designing Patient-Centered Interventions for Emergency Care: Participatory Design Study
JMIR Form Res 2025;9:e63610
URL: https://formative.jmir.org/2025/1/e63610
doi:10.2196/63610
PMID:

©Woosuk Seo, Shruti Jain, Vivian Le, Jiaqi Li, Zhan Zhang, Hardeep Singh, Kalyan Pasupathy, Prashant Mahajan, Sun Young Park. Originally published in JMIR Formative Research (https://formative.jmir.org), 12.02.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Effect of the Virtual Reality–Based Biofeedback Intervention DEEP on Stress, Emotional Tension, and Anger in Forensic Psychiatric Inpatients: Mixed Methods Single-Case Experimental Design

Lisa Klein Haneveld[1,2], MSc; Tessa Dekkers[1], PhD; Yvonne H A Bouman[2], PhD; Hanneke Scholten[3], PhD; Joanneke Weerdmeester[4], PhD; Saskia M Kelders[1], PhD; Hanneke Kip[1,2], PhD

[1]Centre for eHealth and Wellbeing Research, Department of Psychology, Health and Technology, University of Twente, Enschede, The Netherlands
[2]Department of Research, Transfore, Deventer, The Netherlands
[3]Department of Child and Adolescent Psychiatry and Psychology, Erasmus Medical Centre, Rotterdam, The Netherlands
[4]Research Group Creative Making Processes & Technology, HKU University of the Arts, Utrecht, The Netherlands

Corresponding Author:
Lisa Klein Haneveld, MSc
Centre for eHealth and Wellbeing Research
Department of Psychology, Health and Technology
University of Twente
Drienerlolaan 5
Enschede, 7522 NB
The Netherlands
Phone: 31 619210156
Email: l.kleinhaneveld@transfore.nl

Abstract

Background: Decreasing aggression through stress reduction is an important part of forensic psychiatric treatment. DEEP is an experience-based virtual reality intervention that uses biofeedback to train diaphragmatic breathing and increase relaxation. Although DEEP has shown promising results in reducing stress and anxiety in students and adolescents in special education, it has not been examined in forensic psychiatric populations.

Objective: This study aimed to evaluate DEEP's potential to reduce stress, emotional tension, and anger in forensic psychiatric inpatients.

Methods: A mixed methods, alternating treatment, single-case experimental design was conducted with 6 Dutch forensic inpatients. For 20 days, participants engaged in 4 DEEP sessions. Experience sampling was used for continuous monitoring of stress, emotional tension, and anger twice daily. A repeated linear mixed model was used as a primary statistical approach for analyzing the experience sampling data as well as visual analyses. Finally, semistructured interviews were conducted with participants and health care professionals to compare quantitative with qualitative results.

Results: Of the 6 participants, 3 (50%) completed all 4 DEEP sessions, while the other 3 (50%) missed one session due to technical difficulties or absence from the inpatient clinic. P1 showed a significant reduction of stress after session 2 (β=–.865; P=.005). No significant changes over time were found, although an experienced effect was reported during the interviews. P2 showed no significant results. They reported the sessions as being repetitive, with no experienced effect. P3 showed a momentary increase of emotional tension after the first session (β=–.053; P=.002), but no changes were observed over time. No experienced effects were reported in the interview. P4 did not show significant results over time, and was hesitant to report clear experienced effects. P5 showed a significant decline of emotional tension (β=–.012; P=.006), stress (β=–.014; P=.007), and anger (β=–.007; P=.02) over time. They also reported short-term experienced effects in the interview. P6 showed a significant decline of stress over time (β=–.029; P<.001) and reported experiencing substantial effects. Finally, health care professionals reported a relaxing effect of DEEP in their patients but did not expect many long-term effects because no clear behavioral changes were observed.

Conclusions: DEEP shows promise in teaching deep breathing techniques to forensic psychiatric inpatients, potentially decreasing stress, emotional tension, and anger in some patients. However, DEEP is not a one-size-fits-all intervention that supports every patient because the effectiveness on the outcome measures varied among participants. To increase effectiveness, emphasis should

XSL•FO
RenderX

be put on supporting patients to transfer deep breathing skills into their daily lives. This highlights the importance for the structural integration of DEEP into current treatment protocols.

(JMIR Form Res 2025;9:e65206) doi:10.2196/65206

KEYWORDS

virtual reality; VR; diaphragmatic breathing; biofeedback; DEEP; forensic psychiatry; mental health; stress; single-case experimental design; experience sampling method

Introduction

Background

Forensic mental health care is focused on the treatment of individuals who exhibit imminent aggressive or sexually inappropriate behavior due to 1 or more psychiatric illnesses [1,2]. Reducing aggression in treatment is complex because patients often show difficulties with reflection, reading, and writing due to low literacy and treatment motivation [3,4]. However, treatment is mostly focused on talking and thinking about behavior, guided by treatment frameworks such as cognitive behavioral therapy, with less attention for more experience-based or experimental approaches [5]. This shows a need for interventions that can support patients in preventing aggressive behavior in an experiential and engaging way.

Earlier research has shown the promise of virtual reality (VR) and serious games to better involve patients with lower cognitive skills and treatment motivation [5,6]. Previous research has also shown that the VR game DEEP fits the skills and needs of the forensic population [7]. In DEEP, the player moves through an immersive world where they have to use deep diaphragmatic breathing to navigate through and explore the world [8]. The user receives feedback on how they are breathing, which helps them to adjust their breathing and learn to apply diaphragmatic breathing. This technique of providing direct feedback on bodily signals is called biofeedback [9]. Diaphragmatic breathing has been shown to support relaxation, well-being, and anger regulation [10,11].

Previous research has shown the potential of DEEP to reduce negative emotions. First, a randomized controlled trial with DEEP in undergraduate students showed a decrease in anxiety symptoms after using DEEP 4 times [12]. However, these results cannot be generalized, as the study population consists of a relatively healthy, highly educated, and homogeneous target group, which is not the case for the forensic psychiatric population. Second, a single-case experimental design (SCED) study with adolescents in special education demonstrated DEEP's ability to decrease state anxiety and stress [11]. By using an SCED, this study was able to provide insights into both short-term and potential learning effects of DEEP. Moreover, their level of disruptive behavior in class seemed to decrease [11]. Although this population and the forensic target group have overlapping characteristics, the results are not entirely generalizable to a forensic setting. Namely, this study did not include state anger or aggression, which are important treatment outcomes in forensic mental health. There seems to be a need for more specific studies that investigate whether DEEP is fitting the characteristics of the forensic psychiatric setting, focusing on those specific outcomes measures. On the basis of these earlier studies and qualitative research in which DEEP was discussed with forensic patients and health care professionals, DEEP could be a fitting intervention for the hard-to-involve forensic psychiatric patient group. First, because of its appealing and immersive design in VR, DEEP might be able to engage populations with lower motivation, which is not always possible with regular breathing or relaxation interventions [7]. Second, DEEP uses diaphragmatic breathing, where an individual learns to breathe deeply using the diaphragm muscle, as opposed to shallowly through the rib cage and chest [13-15]. In addition to direct relaxation, the use of deep breathing during stressful situations can enhance emotion regulation and reduce stress [16]. By reducing stress, applying deep breathing may also reduce anger and aggression, because these constructs are important factors that may underlie outbursts of aggression [17-19]. By letting patients experience and adjust their own way of breathing in an engaging way, DEEP might be able to support them in increasing relaxation and well-being and decrease emotional tension, stress, and anger.

Objectives

This study aimed to investigate whether DEEP can support forensic inpatients in reducing stress, emotional tension, and anger.

In this mixed methods SCED study, the following research questions were formulated:

1. To what extent is there a significant improvement in self-reported stress, emotional tension, and anger in forensic psychiatric inpatients on days when DEEP is being used compared to days when it is not used?
2. To what extent is there a significant improvement in self-reported stress, emotional tension, and anger in forensic inpatients over time during the period when DEEP was used?
3. To what extent do patients and health care professionals observe or experience effects on emotional tension, stress, and anger that they attribute to DEEP?

Methods

Study Design

To answer the research questions, a mixed methods SCED was applied. Data on stress, emotional tension, and anger were collected quantitatively via experience sampling (ES) as well as qualitatively via interviews. In SCEDs, participants serve as both the intervention group and their own control group. Only a small number of participants are needed to provide sufficient statistical power (3-6), with a minimum of 3 data points per phase, including a baseline and follow-up [20]. Therefore, an

XSL•FO
RenderX

SCED is a fitting study design to use within the forensic psychiatric population, as they form a heterogeneous and complex-to-involve target group with high comorbidity. To study the effectiveness of DEEP, an alternating treatment SCED was chosen [20]. In this type of design, an intervention is not applied on certain days (A phase) and is implemented on other days (B phase), which is then repeated often in a brief period [21,22].

Setting and Participants

The study was conducted at Dutch inpatient clinics with various security levels. Five health care professionals (ie professionals providing social and behavioral support in mental health settings) were included. Of these, 4 (80%) identified as female and 1 (20%) as male, with ages ranging between 27 and 39 (mean 31.2, SD 7.1) years. Patients were included via convenience sampling, which meant that they were included via their health care professional, who identified them as suitable candidates and asked them to participate. A patient was included if (1) they were aged ≥18 years, (2) they were admitted to a forensic psychiatric hospital, (3) they were expected to reside in the same clinic for at least 5 weeks, (4) they were treated for aggression regulation problems, and (5) their health care professional deemed them physically and mentally capable to participate. A patient was excluded if (1) they experienced psychotic episodes or any other form of psychiatric crisis that could negatively impact participation, or (2) they experienced epileptic episodes. In total, 6 forensic inpatients were included. All (6/6, 100%) identified as male, with ages ranging between 23 and 45 (mean 33.4, SD 9.3) years. The data collection commenced 1 week after inclusion, allowing patients sufficient time to reconsider and withdraw from the study if they wished. The inclusion of participants was done in line with the guidelines from the Declaration of Helsinki and from the Medical Research Involving Human Subjects Act (wet medisch-wetenschappelijk onderzoek met mensen in Dutch).

Materials

Avicenna App

During the SCED, the participants were monitored through ES, providing insight into outcome measures, which were measured multiple times per day [23]. ES increases the ecological validity, possibilities, and reliability of frequent data collection [24]. The participants received questions regarding their experienced stress, emotional tension, and anger by use of the Ethica app (formerly known as the Avicenna app), which assists researchers in collecting real-time data [25]. During the study period, participants received 3 notifications per day (at 10:00 AM, 2:00 PM, and 9:00 PM) through the installed app on their mobile phones with a set of 3 statements about the study outcomes: “At this moment I feel [emotionally tense, stressed or angry].” The participants scored these statements on a 5-point Likert scale, where 1 was “not the case at all” and 5 was “very much the case.”

The DEEP Intervention

DEEP is a VR game based on scientific knowledge about anxiety and stress regulation [8]. The player finds themselves in an immersive virtual underwater world where they navigate using their diaphragmatic breathing, which is measured by an abdominal belt [8,26]. The player’s breathing is visualized in the VR environment through a breathing circle, as well as by the illumination and growth of plants and corals, which mirror the player’s breath (Figure 1). By using deep breathing, the player is able to increase their progress in the game, while shallow breathing, which often occurs while experiencing negative emotions, will lead to in-game stagnation. During this study, a beta-version of DEEP was used on the Oculus Rift headset. Oculus (currently known as Meta Quest) is a well-known provider of VR technology and is Conformité Européene certified [27]. In this study, the participants would sit on a spinning desk chair with the VR headset on their head and the abdominal belt around their diaphragm. Both the headset and the belt were attached to a laptop by wires.

Figure 1. A visual representation of a deep breathing exercise in DEEP, as measured with an abdominal belt, with the left image showing a breathing circle during the measured inhalation and right during exhalation.

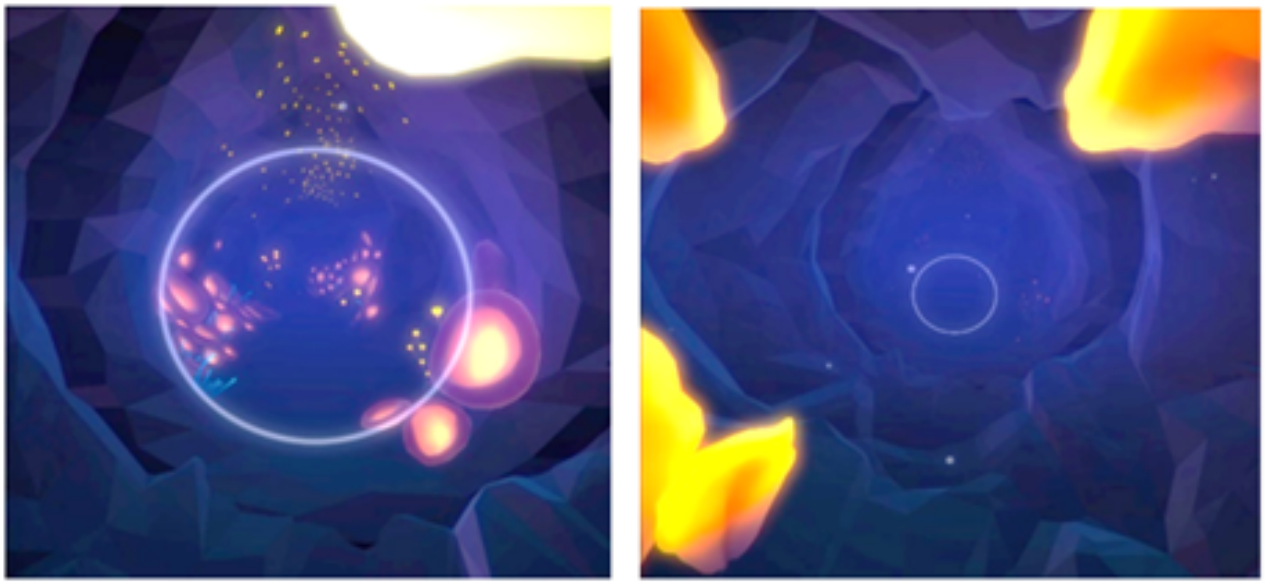

XSL•FO
RenderX

Procedure

Informing and Including Participants

First, the health care professional would ask the patient for approval to be approached by the researchers and to provide them a flyer and information letter. If the patient expressed interest in participating, an appointment was scheduled with the patient to discuss the study. During this meeting, the researcher informed the patient about the time commitment and measurement methods. The health care professional was not present at this meeting to avoid role conflicts that could potentially hinder the voluntary decision to participate due to the dependency relationship with their health care professional. If the patient felt adequately informed and decided to participate in the study, they signed an informed consent. Next, the researcher helped the patient in installing the Avicenna app and assisted them in completing the first set of questions regarding stress, emotional tension, and anger on Avicenna, for instance, by reading them aloud.

Study Procedure

The SCED was conducted between May 2022 and October 2022 by one researcher (LKH). An overview of the SCED procedure is provided in Table 1.

Table 1. Overview of the alternating treatment design conducted with forensic inpatients from a Dutch forensic psychiatric organization (N=6).

	3 days	1 day	3 days	1 day	3 days	1 day	3 days	1 day	3 days
P1	Baseline (A)	DEEP (B)	No DEEP (A)	DEEP (B)	No DEEP (A)	DEEP (B)	No DEEP (A)	DEEP (B)	No DEEP (A)
P2	Baseline (A)	DEEP (B)	No DEEP (A)	DEEP (B)	No DEEP (A)	DEEP (B)	No DEEP (A)	DEEP (B)	No DEEP (A)
P3	Baseline (A)	DEEP (B)	No DEEP (A)	DEEP (B)	No DEEP (A)	DEEP (B)	No DEEP (A)	DEEP (B)	No DEEP (A)

In the B phase, DEEP was introduced to the participant. During a period of 20 days, DEEP was introduced 4 times, based on a previous study on DEEP [11]. The researcher visited the location to prepare the DEEP session and supported the patient in its use. DEEP was started via the laptop, which meant that the researcher could watch along with what the participant was seeing through the headset. Before the first session, the researcher provided participants with a brief explanation of the DEEP concept, deep breathing, and what they would be doing during the DEEP sessions. During the first session, the researcher also provided some help if a participant would get stuck or felt lost in the game. The other 3 sessions were done by the patient independently, while the researcher was there to observe.

Semistructured Interviews

A week after the completion of the SCED, 2 separate semistructured interviews (mean 38, SD 4.23 min) were conducted with participating patients and their daily health care professionals to assess whether they experienced an effect of DEEP on the level of stress, emotional tension, and anger of the participant and if any other benefits or disadvantages for the participants and their behavior had been observed. This was further expanded by asking the participants whether DEEP had an effect during the session, the day of the session, and in their daily lives in general. The health care professionals were asked the same types of questions about the patient's experienced stress, emotional tension, and anger, and they were also asked about the behavior of the participant during the SCED.

Data Analysis

Visual Analyses

The responses to the ES questions resulted in a maximum of 60 data points per participant (3 times per day for 20 days). Time series graphs were constructed per participant in R (version 4.0.0; R Foundation for Statistical Computing) using the *scplot* package (version 0.3.8), which is an add-on package to the *scan* package (version 0.60.0) for visualizing single-case data [28]. Each graph illustrated stress, emotional tension, and anger over time, with days on which DEEP was used marked. Visual analysis was primarily used to gain an overall impression of the data [29,30]. Of the 6 aspects of visual analysis to consider in single case research [29], the level, variability, and trend of each phase were inspected. Level refers to the difference in mean values between phases. Variability refers to the fluctuation of the data, in this case reflected by the data's range. Trend refers to the slope of the best-fitting line for data within a phase and is used to examine whether the outcomes decreased or increased over time and whether the introduction of the intervention changed this trajectory [29].

Repeated Linear Mixed Model Analysis

The data were statistically analyzed using linear mixed models (LMMs) embedded in the MultiSCED web application [31]. MultiSCED is an interactive web application that provides R functionality. LMMs compute 4 parameters—intercept, time effect, intervention effect, and interaction of intervention over time—for each participant, enabling phase-wise comparisons at the participant level. Using MultiSCED, individual linear models were constructed for each participant to determine the individual effects of DEEP on stress, emotional tension, and anger. The additional use of SCED-specific fit indices to statistically examine the nonoverlap between data points across phases A and B such as nonoverlap of all pairs [32], treatment as usual–uncontrolled [33], and typicality of level change [34] were considered. However, this approach was abandoned as half of the participants did not reach the sufficient number of 5 data points in the B phase required for reliable effect sizes [34].

Semistructured Interviews

A deductive approach was used to analyze the interview data [35]. This way, potential findings from both the ES method (ESM) data and interviews complement each other. The interviews were recorded and transcribed verbatim. Afterward, the transcriptions were coded deductively for the outcome measures stress, emotional tension, and anger. Any relevant fragments regarding other outcomes measured were inductively coded. On the basis of these fragments, an initial version of the

XSL•FO
RenderX

coding scheme with main code and subcodes was established [36]. This coding scheme was then independently applied to the first 3 interviews by the first coder (LKH), which then was sent to the second coder (HK). An agreement rate of 81% was found, which is acceptable in the coding process [37].

Ethical Considerations

The Single-Case Reporting Guideline in Behavioral Interventions (SCRIBE) was used to design and report this study [38]. The checklist can be found in Multimedia Appendix 1 [38]. Ethics approval was given by the Ethics Committee of the University of Twente (Behavioral, Management and Social Sciences, no. 220190). Each patient was thoroughly informed about the study through information folders, information letters, and clarification from the researcher. Before participating, every patient signed an informed consent form. As compensation for their participation, all patients received 2 vouchers worth €10 (US $10) halfway through and at the end of the study. To ensure privacy and confidentiality, personal data regarding age, gender, and psychiatric disorder were stored on the secured research drive of the University of Twente. Moreover, names of patients and therapists were omitted from the transcripts, and only participant codes were used to connect quantitative and qualitative data, resulting in a pseudoanonymized dataset.

Results

Overview

The results of the ES and interviews are reported for each participant individually. The interview results per participant are reported using the codes regarding the outcomes measures, that is, stress, emotional tension, and anger. The coding schemes on group level can be found in Multimedia Appendix 1.

Descriptive Statistics

The mean scores on stress, emotional tension, and anger per phase are depicted in Table 2. Half of the participants (3/6, 50%) showed lower average levels of emotional tension (P4-P6), and most lower levels of anger (P1 and P4-P6) and stress (P1 and P3-P6). However, other participants reported more stress (P2), emotional tension (P2 and P3), and anger (P2) on DEEP days.

XSL•FO
RenderX

Table 2. Descriptive statistics of experience sampling data on emotional tension (ET), stress, and anger per participant per phase (scored with a 1-5–point Likert scale).

Participant and outcome	Reported range in Likert scale (1-5)	A phase, mean (SD)	B phase, mean (SD)	Total missing datapoints, for scoring stress, anger and ET combined, n (%)
P1				
Stress	3-5	4.00 (0.24)	3.75 (0.5)	—[a]
ET	3-4	3.92 (0.28)	4 (0)	20 (33)
Anger	2-4	3.39 (0.60)	3 (0.82)	—
P2				
Stress	2-5	2.28 (0.75)	3 (1.41)	—
ET	2-5	2.30 (0.76)	3 (1.41)	12 (20)
Anger	1-5	2.39 (0.93)	3 (1.41)	—
P3				
Stress	2-5	2.08 (0.49)	2 (0)	—
ET	1-4	2.03 (0.37)	2.5 (1)	18 (30)
Anger	2-4	2.05 (0.32)	2 (0)	—
P4				
Stress	2-4	2.12 (0.45)	2 (0)	—
ET	2-4	2.26 (0.66)	2 (0)	15 (25)
Anger	2-5	2.23 (0.68)	2 (0)	—
P5				
Stress	2-4	2.59 (0.69)	2.50 (0.71)	—
ET	2-4	2.42 (0.60)	2 (0)	4 (7)
Anger	1-3	2.13 (0.39)	2 (0)	—
P6				
Stress	1-4	2.64 (1.01)	2.25 (1.26)	—
ET	1-4	2.98 (0.93)	2.25 (1.26)	1 (2)
Anger	1-5	1.82 (1.14)	1.50 (0.58)	—

[a]Not applicable.

Results of ES and Interviews

The results of the visual analyses, LMMs, and interviews are described per participant. At the end of this section, a synthesis of the qualitative and quantitative results is presented. An overview of the LMM outcomes can be found in Multimedia Appendix 1.

Results for P1

P1 was a man aged 44 years who, according to himself, showed difficulty controlling his temper when others irritated or did not agree with him. During the study, P1 was enthusiastic about working with DEEP and was interested in how breathing in a VR environment would help him regulate his emotions. P1 completed all 4 DEEP sessions with no difficulty. Figure 2 illustrates that P1 showed consistent levels of high emotional tension and stress independent of DEEP use. In addition, a significant momentary reduction in stress was found after session 2 (β=–.865; *P*=.005). However, this effect disappeared over time (β=.023; *P*=.02). Although no significant changes over time were found, anger (β=–.001; *P*=.93) seemed to variate more than stress (β=–.002; *P*=.57) or emotional tension (β=–.002; *P*=.52).

XSL•FO
RenderX

Figure 2. Stress, emotional tension, and anger in P1 over time, measured with experience sampling. The light blue marking represents the period DEEP was introduced and withdrawn. The blue triangles indicate the first data point after a DEEP session.

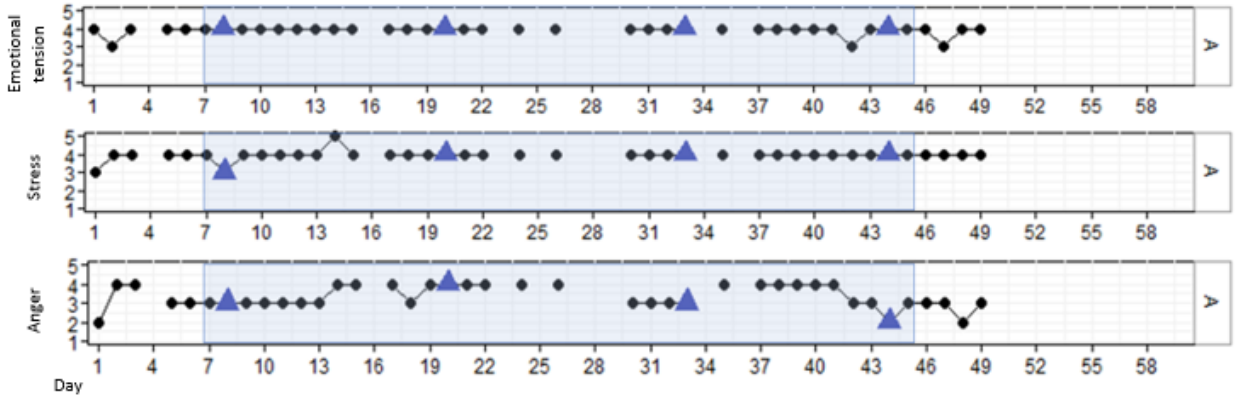

The interview outcomes and examples of quotes regarding those outcomes are shown in Table 3. During the interview, P1 indicated that he found DEEP a pleasant experience. P1 felt that DEEP was more of a helpful reference to understand deep breathing than a direct intervention in increasing emotion regulation. His advice was to not only offer DEEP as a stand-alone intervention but also provide patients with assignments, which can help to better apply deep breathing into daily life situations. Because P1 was trying out for assisted living, he had little contact with health care professionals during this study. Therefore, his health care professionals had no clear view on his daily emotions and behavior during this study.

Table 3. Results of the semistructured interview with P1 regarding the experienced effects on emotional tension, stress, and anger.

Outcome measure	Times mentioned	Example quotes from P1
Emotional tension	5	“DEEP became a reference for me. I do the session and think back on it. Like oh, I feel tense, let’s pretend that I am doing DEEP. Just in my head, feeling like a little fish.”
Stress	2	“And I’ve actually had quite a bit of trouble with panic attacks and stress. But it’s all pretty much gone now. And when I start feeling a bit off, I just focus on my breathing.”
Anger	1	“And it helped just as well with anger. Absolutely. Simply by breathing. Just going back. Back to basics, back to swimming and just breathing calmly.”

Results for P2

P2 was a male patient aged 23 years who, according to his health care professionals, was enthusiastic about his treatment but also impulsive and manipulative in nature. He was glad to participate in this study and was curious to find out whether the deep breathing skills he learned in therapy could support him during a DEEP session. P2 completed 3 out of 4 DEEP sessions. The first session was canceled due to a conflict between him and another inpatient. Figure 3 shows that P2 reported some variation in stress, emotional tension, and anger at the start of the study, which stabilized after data point 10 (emotional tension and stress) and data point 22 (anger). In addition, all outcomes temporarily increased after his last use of DEEP. However, no significant results in stress (β=.002; *P*=.57), emotional tension (β=–.006; *P*=.29), or anger (β=–.002; *P*=.78) were found over time or directly after DEEP sessions.

XSL•FO
RenderX

Figure 3. Stress, emotional tension, and anger in P2 over time, measured with experience sampling. The light blue marking represents the period DEEP was introduced and withdrawn. The blue triangles indicate the first data point after a DEEP session. The gray triangles indicate a DEEP session in which no measurements were taken.

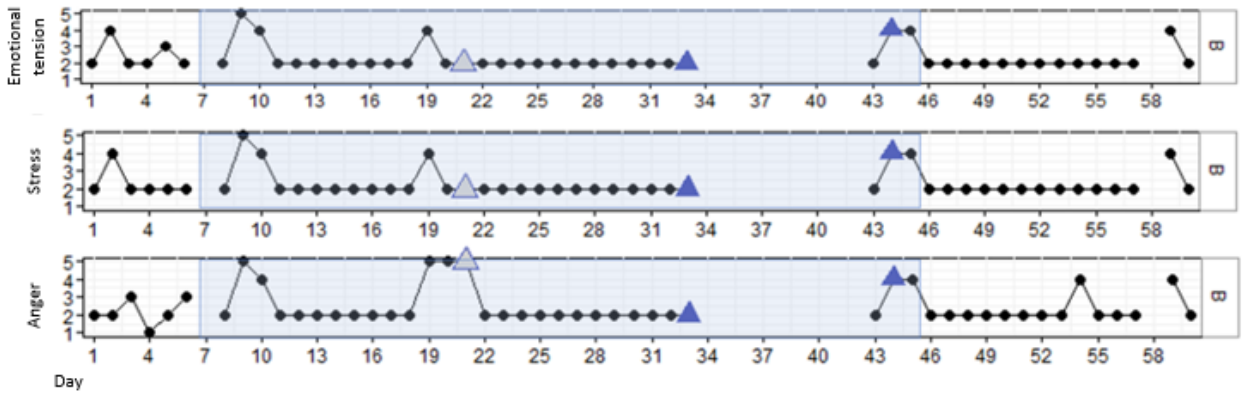

During the interview, P2 indicated that he really liked the first DEEP session but that the last session felt repetitive as he was doing the same thing each time. Moreover, he found the DEEP sessions relaxing, but did not find it to be helpful during moments of conflict. The health care professional of P2 did not really notice any changes in behavior after the DEEP sessions. She did acknowledge that it might have had a positive effect on his well-being, as he was more aware of his breathing. However, during a conflict, she did not notice that he was less angry or that he was able to regulate his emotions through deep breathing (Table 4).

Table 4. Results of the semistructured interview with P2 regarding the experienced effects on emotional tension, stress, and anger.

Outcome measure	Times mentioned	Example quotes from P2
Emotional tension	5	"I would recommend DEEP to other patients. Not everyone would maybe benefit, but it could help most of them with developing breathing skills and a better way to relax and become less tense."
Stress	1	"I did find the DEEP sessions to be effective in reducing my stress. Because I found that when I really focused on my breathing I became more calm and mindful."
Anger	1	"But I do notice that with DEEP, whenever I get stressed or angry or things like that, I do pay attention to my breathing. I don't realize it. It's already ingrained in my system [due to earlier treatment that used breathing exercises] that I do that."

Results for P3

P3 was a man aged 31 years. His health care professionals indicated that he sometimes struggles to access his emotions and talk about them. P3 was interested in participating because he has worked in the IT field and was intrigued by the technology behind the VR game and how the biofeedback in DEEP operates. P3 completed all 4 DEEP sessions. Figure 4 shows that P3 reported low levels and little fluctuation of stress, emotional tension, and anger throughout the study. In addition, his reported emotional tension was significantly increased after the first session (β=1.875; $P<.001$). However, this effect did not remain over time (β=–.053; $P=.002$), indicating only a momentary increase in tension with first-time use of DEEP. Finally, no changes over time in stress (β=–.006; $P=.34$), emotional tension (β=.002; $P=.71$), and anger (β=–.001; $P=.85$) were found.

Figure 4. Stress, emotional tension, and anger in P3 over time, measured with experience sampling. The light blue marking represents the period DEEP was introduced and withdrawn. The blue triangles indicate the first data point after a DEEP session.

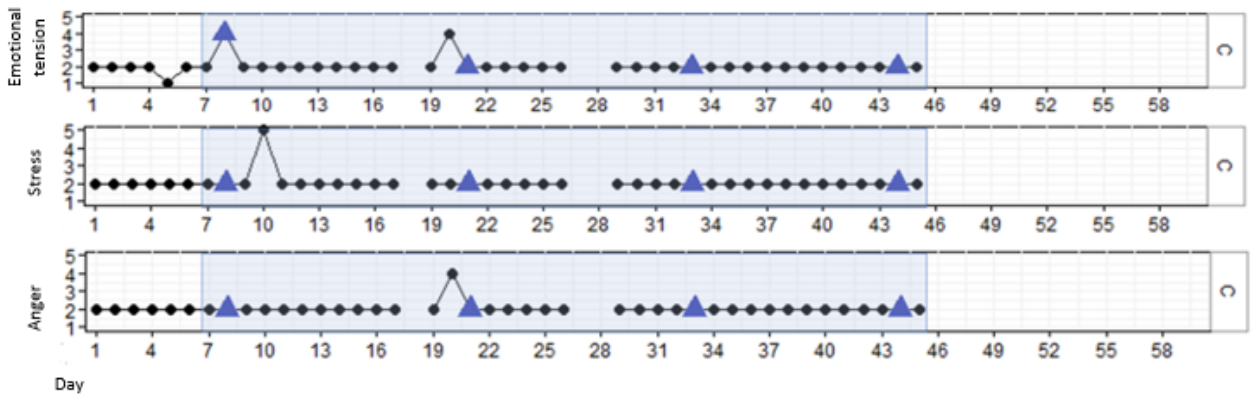

XSL•FO
RenderX

During the interview, P3 stated that he expected more from the DEEP sessions. He felt that the VR showed some technical defects and that it did not help him in reducing negative emotions. According to his health care professional, P3 was not entirely engaged in DEEP and what it might teach him. He was more involved in the technical part, such as the VR and abdominal belt, and expressed annoyance that it did not always work correctly (Table 5).

Table 5. Results of the semistructured interview with P3 regarding the experienced effects on emotional tension, stress, and anger.

Outcome measure	Times mentioned	Example quotes from P3
Emotional tension	2	"I was able to focus on my breathing during DEEP, which made me feel relaxed. But I already knew this from going to the gym and deep breath during my exercises."
Stress	1	"I don't know if DEEP helps with reducing stress. I would say that you would need more triggers to test that. Like that things get a bit scary in the game and that you have to keep breathing to calm yourself down."
Anger	2	"Something just came to mind. I don't even remember what it was about, but I was irritated recently. But expressed it calmly, and was immediately vocal about it, you know. But then, did I take a breath? No, that feels pointless."

Results for P4

P4 was a man aged 24 years. According to his health care professionals, he was a quiet man who had difficulty recognizing and articulating his own emotions. P4 completed 3 of 4 sessions. The third session was canceled due to the researcher being unavailable unexpectedly. As is shown in Figure 5, P4 reported low stress, emotional tension, and anger during the first half of the study. More variety in these outcomes was observed in the second half of the study. However, no significant results in stress (β=.004; *P*=.44), emotional tension (β=.000; *P*=.97), and anger (β=.004; *P*=.61) were found over time, nor was a direct effect observed after a DEEP session.

Figure 5. Stress, emotional tension, and anger in P4 over time, measured with experience sampling. The light blue marking represents the period DEEP was introduced and withdrawn. The blue triangles indicate the first data point after a DEEP session. The gray triangles indicate a DEEP session in which no measurements were taken.

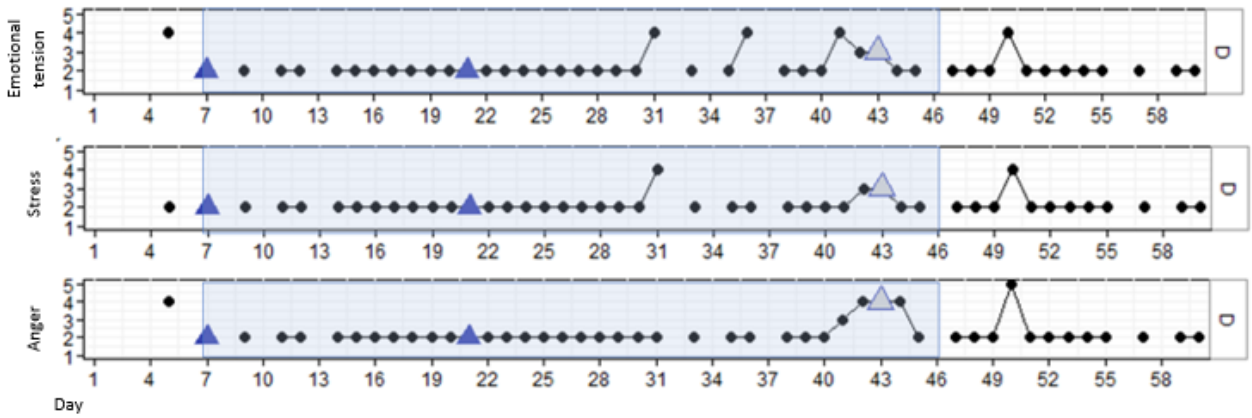

During the interview, P4 stated DEEP sessions to be a fun and relaxing experience. He also felt that after a DEEP session, he was able to stay calm the rest of the day, even when conflict occurred. However, P5 was hesitant to state that DEEP helped him using his deep breathing in daily life situations. Except for during a DEEP session, he did not use his deep breathing to cope with negative emotions, such as anger. His health care professional was also hesitant to ascribe a direct effect of DEEP on P5's behavior. She did find P5 to be calm and tired after a DEEP session and indicated that it provided him with more insight on deep breathing. However, she did not find him less tense, stressed, or angry during the SCED, as he was already behaving well in the clinic (Table 6).

Table 6. Results of the semistructured interview with P4 regarding the experienced effects on emotional tension, stress, and anger.

Outcome measure	Times mentioned	Example quotes from P4
Emotional tension	6	"I loved the DEEP sessions. It was nice and really relaxing. Just to be in a calm environment made me less tense already."
Stress	4	"DEEP gave me an empty mind. I had nothing to think or stress about. The mornings you [researcher] came around with DEEP were nice days. I could really relax."
Anger	4	"After the session we were cooking, and there was someone bossing people around, and normally I would say: do something yourself for once! Normally, I would react with irritation, but now I was very calm and could respond more calmly."

XSL•FO
RenderX

Results for P5

P5 was a man aged 37 years. According to both him and his health care professional, he did not always recognize when his anger was increasing and could suddenly explode. He was curious about how DEEP could help him in recognizing his increasing emotional tension and using deep breathing techniques in a timely manner. P5 completed 3 of 4 DEEP sessions. The first session was canceled due to a technical problem with the VR set. As is shown in Figure 6, P5 reported fluctuations on the 3 outcome measures. The DEEP sessions were consistently associated with the lowest scores on emotional tension, stress, and, to a lesser extent, anger. In addition, a significant decline over time in emotional tension (β=–.012; *P*=.006), stress (β=–.014; *P*=.007), and anger (β=–.007; *P*=.02) was found, with high to moderate levels of stress until day 9 and moderate to low levels of stress after that.

Figure 6. Stress, emotional tension, and anger in P5 over time, measured with experience sampling. The light blue marking represents the period DEEP was introduced and withdrawn. The blue triangles indicate the first data point after a DEEP session. The gray triangles indicate a DEEP session in which no measurements were taken.

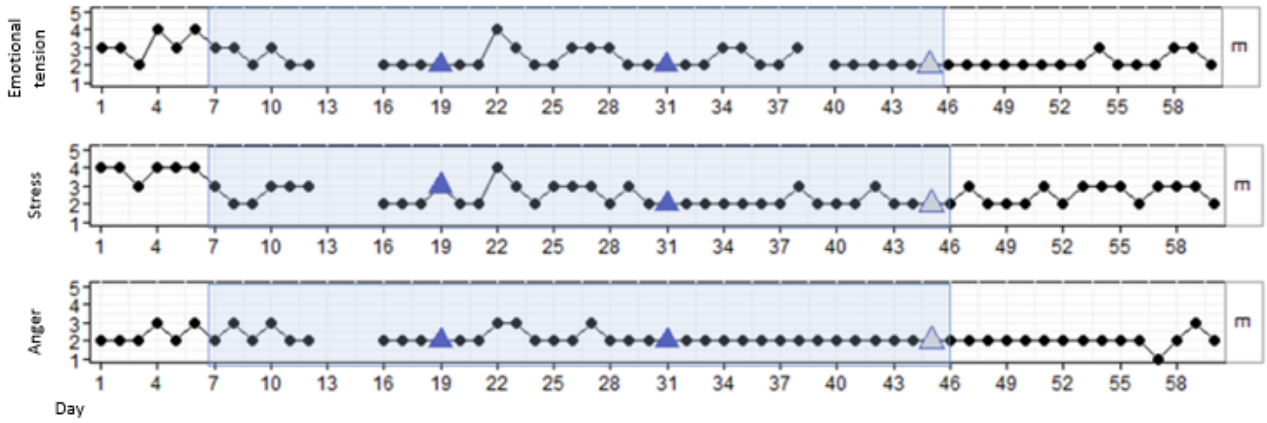

During the interview, P5 stated that DEEP was a helpful intervention for him to relax and focus on his own body. However, P5 was not able to use his deep breathing techniques and reduce his anger during conflicts or stressful situations. He stated that during those moments, he is so engaged in the conflict that he is not thinking about his bodily signals or what he should do to prevent aggression. His health care professional agreed with these statements. She acknowledged that P5 really felt happy and relaxed after each DEEP session and that he was able to keep calm during most conflicts. However, the health care professional did not find DEEP to be effective in supporting P5 during stressful moments or aggressive outbursts (Table 7).

Table 7. Results of the semistructured interview with P5 regarding the experienced effects on emotional tension, stress, and anger

Outcome measure	Times mentioned	Example quotes from P5
Emotional tension	5	"You know, one thing I've become more aware of is my breathing. Even before getting out of bed, I sometimes pay attention to my abdominal breathing. So, I'm conscious that breathing can be a momentary relief of tension to start and get through the day."
Stress	2	"I do feel less stressed, but I am not entirely sure if it can be linked to doing DEEP. I think I have to do more DEEP sessions to discover that."
Anger	2	"During DEEP I was in a calm environment to chill and get out of my head. But I don't know if those breathing exercises work if I am in a fight with a family member. I don't think so to be honest."

Results for P6

P6 was a man aged 33 years. According to his health care professional, he often struggled to regulate his negative emotions in a healthy way. DEEP appealed to P6 because he wanted to find distraction through it and calm his thoughts. P6 completed all 4 DEEP sessions. However, the second session for P6 was interrupted due to a signal disruption between the VR headset and the VR sensor after 8 minutes. The session was eventually completed but the participant did experience more discomfort than usual, because his vision kept shifting, causing difficulty to navigate properly. As is shown in Figure 7, P6 reported both highly fluctuating levels of stress, emotional tension, and anger. His findings are mixed, with DEEP sessions being associated with both the least (session 1; day 2) and highest (session 2; day 6) emotional tension. However, anger levels did seem to decrease with the use of DEEP, notably after session 3 (day 10), which was followed by a longer period of stability. In addition, a significant decline of stress was found over time (β=–.029; *P*<.001) but not in emotional tension (β=–.009; *P*=.22) or anger (β=.009; *P*=.33). In addition, as neither of the interaction effects are significant, these declines cannot be attributed directly to a momentary effect of DEEP.

XSL•FO
RenderX

Figure 7. Stress, emotional tension, and anger in P6 over time, measured with experience sampling. The light blue marking represents the period DEEP was introduced and withdrawn. The blue triangles indicate the first data point after a DEEP session.

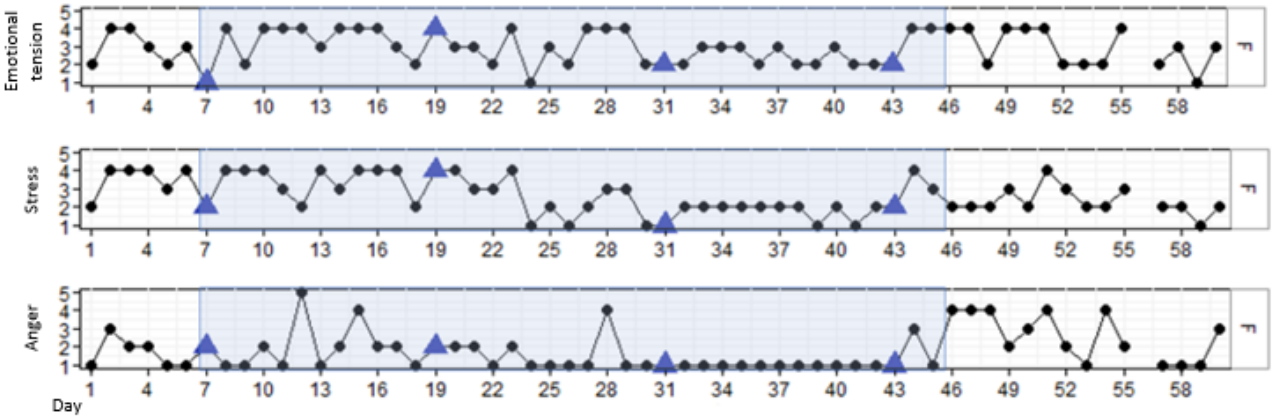

During the interview, P6 expressed being pleasantly surprised by how much he liked the DEEP sessions. He found the environment and music of DEEP to be a great distraction from his negative thoughts and felt that it helped him to cope with his negative emotions. According to his health care professional, P6 really experienced that he can use his deep breathing to reduce emotional tension and stress. She stated that, although there were some conflicts during the data collection period, P6 clearly was less angry and explosive toward the health care professionals (Table 8).

Table 8. Results of the semistructured interview with P6 regarding the experienced effects on emotional tension, stress, and anger.

Outcome measure	Times mentioned	Example quotes from P6
Emotional tension	6	"What really stuck with me was that I started every DEEP session tense and that I came out of the session calm and happy."
Stress	2	"I do remember that, during the DEEP sessions, I was able to breath in a deep and calm way which felt really nice and made my body relax as well."
Anger	3	"I've had quite a rough couple of weeks. And after the DEEP sessions, I noticed that I didn't immediately react to conflicts, but found a way to cope. I didn't explode right away but managed to keep myself together. Also, because I emerged from such a session feeling so relaxed."
Self-efficacy[a]	2	"I know that if I explode, I bring a lot of tension to the staff and the group. DEEP helped me to respond more calmly. But now the study has ended, and I don't know how I will do without the help of DEEP."

[a]This was not a predetermined outcome measure but additionally mentioned by P6.

Synthesis of the Results

In this study, we used multiple research methods to assess the effectiveness of DEEP on stress, emotional tension, and anger in forensic psychiatric patients. In Table 9, we juxtaposed the quantitative and qualitative results per participant. This provides an overview on group level and an individualized examination of whether they paint a consistent picture or present contradictory results. Some discrepancies were observed between quantitative and qualitative results. Two patients showed no significant changes quantitatively, but reported experiencing effects of DEEP in the interviews. Two other patients displayed short-term significant changes in emotional tension or stress, with one patient reporting matching experiences and the other reporting no effects. The final 2 patients showed significant changes over time, aligning with their reported experiences.

XSL•FO
RenderX

Table 9. Overview of quantitative and qualitative results per participant (N=6).

ID	Quantitative outcomes	Qualitative outcomes
P1	• Variation in stress, ET[a], and anger • Momentary decrease of stress • No significant changes over time	• Experienced short-term effect on ET and stress • No experienced long-term effect
P2	• Stabilization of stress, ET, and anger in second half SCED[b] • No significant changes over time	• Experienced short-term effects on ET • No experienced long-term effects
P3	• Low stress, ET, and anger throughout SCED • Momentary increase of ET • No significant changes over time	• No experienced short-term effects • No experienced long-term effects
P4	• More fluctuation of stress, ET, and anger in second half SCED • No significant changes over time	• Experienced short-term effect on ET, stress, and anger • Experienced long-term effect on ET
P5	• Fluctuation, but lower levels in stress, ET, and anger after DEEP sessions. • Significant reduction of stress, ET, and anger over time	• Experienced short-term effect on ET and stress • Experienced long-term effect on ET and stress
P6	• Elevated levels and fluctuation of stress, ET, and anger throughout SCED • Significant reduction of stress over time	• Experienced short-term effect on ET, stress, and anger • Experienced long-term effect on stress and anger

[a]ET: emotional tension.

[b]SCED: single-case experimental design.

Discussion

Principal Findings

This study aimed to investigate whether DEEP, through its immersive focus on relaxation and diaphragmatic breathing techniques, is able to decrease stress, emotional tension, and anger in forensic psychiatric inpatients. First, much variation in effectiveness was found between each participant. In 2 participants, a momentary effect on emotional tension and stress right after a DEEP session was found. Two other participants showed a decrease in stress, emotional tension, and anger over time, with no direct effect of DEEP. The last 2 participants did not show any significant effects. These results underline that DEEP is no one-size-fits-all solution and that users might benefit from DEEP in diverse ways or not at all. Second, most participants felt enthusiastic about DEEP and found it to be a helpful intervention to enhance short-term relaxation. Some participants also found DEEP to be able to teach them diaphragmatic breathing, which, in their experience, helped them reduce aggressive outbursts or conflicts. However, most patients were not yet able to put what they had learned in DEEP into practice, which was also indicated by their health care professionals. Finally, some discrepancies were found in the synthesis of the ES and qualitative results; although most patients experienced a positive effect of DEEP on their well-being, 67% (4/6) of them showed no significant change in emotional tension, stress, or anger.

Variation in Effectiveness

An important finding of this study is that DEEP seems able to reduce momentary emotional tension, stress, and anger and induce short-term relaxation in some but not all forensic inpatients. One possible explanation for this is that the participants in this study exhibit heterogeneity in their history of offenses, treatment objectives, cognitive abilities, and psychosocial characteristics [4,39], which might explain the differences in their experiences during the DEEP sessions and variation in effectiveness. This underlines earlier research, which states that there is no one-size-fits-all approach to support all psychiatric patients in their treatment and that a personalized approach is essential [40], which is in particular the case for forensic patients [3]. Furthermore, DEEP was provided as a stand-alone intervention within a research context without a more personalized instruction. The participants stated that they had trouble applying what they learned in DEEP into daily life situations. This shows that a structural integration of DEEP into existing treatment could support patients in generalizing deep breathing skills, making the intervention more effective as is supported by earlier studies. Finally, some technical difficulties with the abdominal waistband occurred during the study, which limited the impact biofeedback could have on participants. Future research could further investigate these reasons for variation in effectiveness, for example, by focusing on patient-, contextual-, and intervention-related factors [41-43].

Generalization of Deep Breathing Skills

Another finding of this study is that DEEP was able to induce short-term effects on stress, emotional tension, and anger. However, a long-term effect that could directly be ascribed to DEEP was not found in any participants. Similar to earlier evaluation studies of DEEP, the DEEP sessions were conducted as a “stand-alone” intervention, without an extensive introduction or debriefing from a health care professional [8,11,26]. In the interviews, patients and health care professionals stated that more information about diaphragmatic breathing could help them to actively use this as a coping

XSL•FO
RenderX

strategy to prevent aggression in their daily lives. Therefore, DEEP might require some form of psychoeducation and support with integration in daily life to generalize the skills that are taught in DEEP [7]. In addition, other studies recommended that technological interventions, such as VR, should be embedded within existing treatment as opposed to being used as a separate tool [6,44,45]. Therefore, future research should focus on determining the most optimal way to integrate DEEP in a treatment context where health care professionals are conducting the DEEP sessions.

Using Deep Breathing During Conflicts

A third finding of this study is that most participants often forgot to apply diaphragmatic breathing during conflicts because they were already too involved or angry. An explanation for this is that patients were often not aware of their increasing emotional tension or anger, which is a common problem in forensic psychiatric patients [46]. Consequently, necessary precondition for applying diaphragmatic breathing to prevent aggression seems to be that patients recognize their increasing tension early on [18,47,48]. This degree of recognition and trust in one's own bodily signals—which is called interoceptive awareness—is indeed found to be related to the improvement of emotion regulation skills [48]. Therefore, it would be interesting to determine whether the effect of DEEP or other emotion regulation interventions can be enhanced if the user's interoceptive awareness increases. Consequently, a recommendation for future research is to combine DEEP with other biofeedback devices such as wearables that can be used to increase interoceptive awareness in daily life [46,47].

Strengths and Limitations

First, by using a mixed methods single-case approach, this study was able to differentiate individual responses to a novel intervention. However, the data quality of the current SCED study limited opportunities to use SCED-specific effect sizes such as treatment as usual–uncontrolled, nonoverlap of all pairs, and typicality of level change and, as such, to be able to draw robust conclusions about the effectiveness of DEEP. To meet standards in future studies, at least 5 data points in each phase should be produced. This could be successfully reached by extending the DEEP phase by embedding the VR sessions in treatment as usual. This way more data points are easily collected and long-term changes in outcome measures might be reached. Second, by using convenience sampling as a sampling method, researchers and health care professionals were able to recruit and include participants, while ensuring a time- and cost-effective procedure [49]. However, it could have led to selection bias as health care professionals may have only recruited patients that they deemed motivated to participate to ensure that they were able to complete the study [49,50]. Therefore, it is important to remain careful when generalizing study outcomes, particularly as the vulnerable forensic psychiatric patients could not be involved in this study. Third, attention should be paid to the impact of technical issues, as some DEEP sessions were canceled or interrupted due to signal loss between the abdominal belt and the VR equipment. It should be noted that these issues could have inhibited the possible effectiveness of DEEP. Fourth, this study used ES, which is a suitable way to study real-time changes in mental states [51,52]. As a significant decrease in stress, emotional tension, and anger over time—without a direct effect of DEEP—was found in 2 patients, ES could have provided an unforeseen effect on their well-being, just by asking them to reflect on their emotions thrice a day. However, by using a mixed methods design, these possible unforeseen effects have been partly addressed in semistructured interviews because quantitative ESM data were supplemented by experienced effectiveness. Finally, emotional tension, stress, and anger are complex constructs to measure, particularly using ES [53]. Some participants mentioned effects of DEEP that were not reflected in the ESM data, raising questions about the extent to which the items were fully valid. Finally, participants scored their stress, emotional tension, and anger by using a 5-point Likert scale, an often-used scoring system in validated questionnaires [54,55]. However, this resulted in less detailed information about the emotional state of the participants; although an increase from 1 to 3 may seem significant, it simply indicates that the participant selected "(3) neutral/do not know," which does not directly reflect an improvement. To encounter this, a visual analog scale of 0 to 10 or 0 to 100 might have been a better way to assess the outcomes in a more specific, sensitive way.

Conclusions

This study has shown that DEEP seems a promising intervention for some, but not all, forensic psychiatric inpatients to practice deep breathing techniques in a novel and engaging way. DEEP is not a one-size-fits-all intervention that works for every patient, as the effectiveness in decreasing emotional tension, stress, and anger varied among the participants. Despite variation in quantitative results, both inpatients and the health care professionals found DEEP to be of added value in providing short-term relaxation. In addition, DEEP seems to be an engaging way to acquire deep breathing as a coping strategy. However, generalization to daily lives was difficult, which can be explained by problems with interoceptive awareness and lack of integration in treatment protocols. This underlines the need for a structural integration of interventions such as DEEP in existing treatment programs or psychoeducation. Future research should therefore focus not only on whether DEEP is effective, but also on why and for whom it is effective, by identifying patient-, context-, and intervention-related factors.

Acknowledgments

This study was supported by a grant from Kwaliteit Forensische Zorg to HK (KFZ 2020-146). The authors would like to thank the managers and health care professionals of the participating clinics for their time and effort to help conduct this study. The authors would also like to thank the patients for participating in this study.

XSL•FO
RenderX

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

HK and LKH contributed to the design and planning of the study. LKH contributed to data collection. All authors contributed to the interpretation and reporting of the results and approved the manuscript.

Conflicts of Interest

JW works with Explore DEEP Ltd as a scientific advisor. All other authors declare no other conflicts of interest.

Multimedia Appendix 1
Overview of Single-Case Reporting Guideline in Behavioral Interventions (SCRIBE) checklist, linear regression models, and coding schemes.
[DOCX File , 32 KB - formative_v9i1e65206_app1.docx]

References

1. Meynen G. Walls and laws: structural barriers to forensic psychiatric research. Eur Psychiatry 2017 Jul 23;44:208-209. [doi: 10.1016/j.eurpsy.2017.04.010] [Medline: 28646733]
2. Sygel K, Wallinius M. Immersive virtual reality simulation in forensic psychiatry and adjacent clinical fields: a review of current assessment and treatment methods for practitioners. Front Psychiatry 2021 May 28;12:673089 [FREE Full text] [doi: 10.3389/fpsyt.2021.673089] [Medline: 34122189]
3. Kip H, Bouman YH. Van denken en praten naar doen en ervaren: op zoek naar de toegevoegde waarde van technologie voor de forensische psychiatrie. Sancties 2020;1:249-260.
4. Kip H, Bouman YH. A perspective on the integration of eHealth in treatment of offenders: combining technology and the risk-need-responsivity model. Front Psychiatry 2021 Sep 1;12:703043 [FREE Full text] [doi: 10.3389/fpsyt.2021.703043] [Medline: 34539462]
5. Kip H, Oberschmidt K, Bierbooms J, Dijkslag D. Technologie in de forensische zorg - Crossing borders. Utrecht: Kwaliteit Forensische Zorg (KFZ). 2019. URL: https://www.researchgate.net/publication/334825702_Technologie_in_de_forensische_zorg_-_Crossing_borders [accessed 2024-04-29]
6. Klein Haneveld L, Kip H, Bouman YH, Weerdmeester J, Scholten H, Kelders SM. Een sprong in het diepe: de evaluatie en implementatie van DEEP, een virtual reality biofeedback game, in de forensische zorg. Utrecht: Kwaliteit Forensische Zorg. URL: https://kfz.nl/projecten/een-sprong-in-het-diepe-de-evaluatie-en-implementatie-van-deep [accessed 2024-04-29]
7. Klein Haneveld L, Kip H, Bouman YH, Weerdmeester J, Scholten H, Kelders SM. Exploring the added value of virtual reality biofeedback game DEEP in forensic psychiatric inpatient care-a qualitative study. Front Psychol 2023 Nov 9;14:1201485 [FREE Full text] [doi: 10.3389/fpsyg.2023.1201485] [Medline: 38023054]
8. Weerdmeester JW. Take a DEEP breath: exploring the potential of game-based biofeedback interventions for anxiety regulation. Radboud University. URL: https://repository.ubn.ru.nl/bitstream/handle/2066/239933/239933.pdf?sequence=2&isAllowed=y [accessed 2024-04-29]
9. Schwartz MS, Andrasik F. S. In: Biofeedback: A Practitioner's Guide. 4th edition. New York, NY: The Guilford Press; 2017.
10. Fitry AP, Ramdani VH, Suhanda S, Rosmiati R. Implementation of deep breathing relaxation techniques to control anger at violent behavior. KIAN J 2023 Mar 30;2(1):26-32. [doi: 10.56359/kian.v2i1.237]
11. Bossenbroek R, Wols A, Weerdmeester J, Lichtwarck-Aschoff A, Granic I, van Rooij MM. Efficacy of a virtual reality biofeedback game (DEEP) to reduce anxiety and disruptive classroom behavior: single-case study. JMIR Ment Health 2020 Mar 24;7(3):e16066 [FREE Full text] [doi: 10.2196/16066] [Medline: 32207697]
12. Weerdmeester JW, van Rooij MM, Maciejewski DF, Engels RC, Granic I. A randomized controlled trial assessing the efficacy of a virtual reality biofeedback video game: anxiety outcomes and appraisal processes. Technol Mind Behav 2021 Aug 05;2(2):23. [doi: 10.1037/tmb0000028]
13. Agrawal V, Naik V, Duggirala M, Athavale S. Calm a mobile based deep breathing game with biofeedback. In: Extended Abstracts of the 2020 Annual Symposium on Computer-Human Interaction in Play. 2020 Presented at: CHI PLAY '20; November 2-4, 2020; Virtual Event p. 153-157. [doi: 10.1145/3383668.3419876]
14. Chen YF, Huang XY, Chien CH, Cheng JF. The effectiveness of diaphragmatic breathing relaxation training for reducing anxiety. Perspect Psychiatr Care 2017 Oct 23;53(4):329-336. [doi: 10.1111/ppc.12184] [Medline: 27553981]
15. Hopper SI, Murray SL, Ferrara LR, Singleton JK. Effectiveness of diaphragmatic breathing for reducing physiological and psychological stress in adults: a quantitative systematic review. JBI Database System Rev Implement Rep 2019 Sep;17(9):1855-1876. [doi: 10.11124/JBISRIR-2017-003848] [Medline: 31436595]

XSL•FO
RenderX

16. Brzozowski A, White RG, Mitchell IJ, Beech AR, Gillespie SM. A feasibility trial of an instructed breathing course in prison to improve emotion regulation in people with substance use difficulties. J Forens Psychiatry Psychol 2020 Dec 10;32(2):308-325. [doi: 10.1080/14789949.2020.1859589]
17. Asberg K. Hostility/anger as a mediator between college students' emotion regulation abilities and symptoms of depression, social anxiety, and generalized anxiety. J Psychol 2013 Sep;147(5):469-490. [doi: 10.1080/00223980.2012.715601] [Medline: 24003591]
18. Pérez-Fuentes MD, Molero Jurado MD, Barragán Martín AB, Gázquez Linares JJ. Family functioning, emotional intelligence, and values: analysis of the relationship with aggressive behavior in adolescents. Int J Environ Res Public Health 2019 Feb 06;16(3):478 [FREE Full text] [doi: 10.3390/ijerph16030478] [Medline: 30736326]
19. Sprague J, Verona E, Kalkhoff W, Kilmer A. Moderators and mediators of the stress-aggression relationship: executive function and state anger. Emotion 2011 Feb;11(1):61-73. [doi: 10.1037/a0021788] [Medline: 21401226]
20. Krasny-Pacini A, Evans J. Single-case experimental designs to assess intervention effectiveness in rehabilitation: a practical guide. Ann Phys Rehabil Med 2018 May;61(3):164-179 [FREE Full text] [doi: 10.1016/j.rehab.2017.12.002] [Medline: 29253607]
21. Cariveau T, Fetzner D. Experimental control in the adapted alternating treatments design: a review of procedures and outcomes. Behav Interv 2022 Jan 25;37(3):805-818. [doi: 10.1002/bin.1865]
22. Sindelar PT, Rosenberg MS, Wilson RJ. An adapted alternating treatments design for instructional research. Educ Treat Children 1985;8(1):67-76.
23. Smith JD. Single-case experimental designs: a systematic review of published research and current standards. Psychol Methods 2012 Dec;17(4):510-550 [FREE Full text] [doi: 10.1037/a0029312] [Medline: 22845874]
24. van Berkel N, Ferreira D, Kostakos V. The experience sampling method on mobile devices. ACM Comput Surv 2017 Dec 06;50(6):1-40. [doi: 10.1145/3123988]
25. A reliable and featureful outcome assessment platform. Avicenna Research. URL: https://avicennaresearch.com/features/outcome-assessment [accessed 2024-04-29]
26. Van Rooij M, Lobel A, Harris O, Smit N, Granic I. DEEP: a biofeedback virtual reality game for children at-risk for anxiety. In: Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems. 2016 Presented at: CHI EA '16; May 7-12, 2016; San Jose, CA p. 1989-1997 URL: https://dl.acm.org/doi/10.1145/2851581.2892452 [doi: 10.1145/2851581.2892452]
27. Aan de slag met de Oculus Rift. Meta Quest. URL: https://www.oculus.com/rift-s/ [accessed 2024-04-29]
28. Wilbert J. Analyzing single-case data with R and scan. zenodo. URL: https://jazznbass.github.io/scan-Book/ [accessed 2024-04-29]
29. Kratochwill TR, Hitchcock J, Horner RH, Levin JR, Odom SL, Rindskopf DM, et al. Single-case design technical documentation. Institute of Education Sciences, U.S. Department of Education. 2010. URL: https://ies.ed.gov/ncee/WWC/Docs/ReferenceResources/wwc_scd.pdf [accessed 2024-04-29]
30. Kratochwill TR, Hitchcock JH, Horner RH, Levin JR, Odom SL, Rindskopf DM, et al. Single-case intervention research design standards. Remedial Spec Educ 2012 Aug 15;34(1):26-38. [doi: 10.1177/0741932512452794]
31. Declercq L, Cools W, Beretvas SN, Moeyaert M, Ferron JM, Van den Noortgate W. MultiSCED: a tool for (meta-)analyzing single-case experimental data with multilevel modeling. Behav Res Methods 2020 Feb 10;52(1):177-192. [doi: 10.3758/s13428-019-01216-2] [Medline: 30972557]
32. Parker RI, Vannest K. An improved effect size for single-case research: nonoverlap of all pairs. Behav Ther 2009 Dec;40(4):357-367. [doi: 10.1016/j.beth.2008.10.006] [Medline: 19892081]
33. Parker RI, Vannest KJ, Davis JL, Sauber SB. Combining nonoverlap and trend for single-case research: Tau-U. Behav Ther 2011 Jun;42(2):284-299. [doi: 10.1016/j.beth.2010.08.006] [Medline: 21496513]
34. Landman W, Bogaerts S, Spreen M. Typicality of level change (TLC) as an additional effect measure to NAP and Tau-U in single case research. Behav Modif 2023 Aug 31;48(1):51-74. [doi: 10.1177/01454455231190741]
35. Plano Clark VL, Huddleston-Casas CA, Churchill SL, O'Neil Green D, Garrett AL. Mixed methods approaches in family science research. J Fam Issues 2008 Apr 01;29(11):1543-1566. [doi: 10.1177/0192513x08318251]
36. Bree RT, Gallagher G. Using Microsoft Excel to code and thematically analyse qualitative data: a simple, cost-effective approach. All Ireland J High Educ 2016;8(2):2811-2824 [FREE Full text]
37. Miles MB, Huberman AM. Qualitative Data Analysis: An Expanded Sourcebook. Thousand Oaks, CA: Sage Publications; 1994.
38. Tate RL, Perdices M, Rosenkoetter U, Shadish W, Vohra S, Barlow DH, et al. The single-case reporting guideline in BEhavioural interventions (SCRIBE) 2016 statement. Remedial Spec Educ 2016 Aug 02;37(6):370-380. [doi: 10.1177/0741932516652893]
39. van Nieuwenhuizen C, Bogaerts S, de Ruijter EA, Bongers IL, Coppens M, Meijers RA. TBS-behandeling geprofileerd een gestructureerde casussenanalyse. Wetenschappelijk Onderzoek- en Datacentrum. 2011. URL: https://repository.wodc.nl/bitstream/handle/20.500.12832/1837/samenvatting_tcm28-71170.pdf?sequence=1&isAllowed=y [accessed 2021-02-06]

XSL•FO
RenderX

40. Pieterse M, Kip H, Martínez RR. The complexity of eHealth implementation: a theoretical and practical perspective. In: Kip H, Beerlage-de Jong N, van Gemert-Pijnen L, Sanderman R, Kelders SM, editors. eHealth Research, Theory and Development: A Multi-Disciplinary Approach. New York, NY: Routledge; 2018.
41. Nielsen SB, Lemire S, Tangsig S. Unpacking context in realist evaluations: findings from a comprehensive review. Evaluation 2021 Oct 25;28(1):91-112. [doi: 10.1177/13563890211053032]
42. Nielsen K, Miraglia M. What works for whom in which circumstances? On the need to move beyond the 'what works?' question in organizational intervention research. Hum Relat 2016 Nov 12;70(1):40-62. [doi: 10.1177/0018726716670226]
43. van Hooft SM, Been-Dahmen JM, Ista E, van Staa A, Boeije HR. A realist review: what do nurse-led self-management interventions achieve for outpatients with a chronic condition? J Adv Nurs 2017 Jun 09;73(6):1255-1271. [doi: 10.1111/jan.13189] [Medline: 27754557]
44. Kip H, Bouman YH, Kelders SM, van Gemert-Pijnen LJ. eHealth in treatment of offenders in forensic mental health: a review of the current state. Front Psychiatry 2018 Feb 21;9:42 [FREE Full text] [doi: 10.3389/fpsyt.2018.00042] [Medline: 29515468]
45. Maples-Keller JL, Bunnell BE, Kim S, Rothbaum BO. The use of virtual reality technology in the treatment of anxiety and other psychiatric disorders. Harv Rev Psychiatry 2017;25(3):103-113 [FREE Full text] [doi: 10.1097/HRP.0000000000000138] [Medline: 28475502]
46. Ter Harmsel JF, Noordzij ML, van der Pol TM, Swinkels LT, Goudriaan AE, Popma A. Exploring the effects of a wearable biocueing app (Sense-IT) as an addition to aggression regulation therapy in forensic psychiatric outpatients. Front Psychol 2023 Mar 10;14:983286 [FREE Full text] [doi: 10.3389/fpsyg.2023.983286] [Medline: 36968738]
47. de Looff PC, Noordzij ML, Nijman HL, Goedhard L, Bogaerts S, Didden R. Putting the usability of wearable technology in forensic psychiatry to the test: a randomized crossover trial. Front Psychiatry 2024 Jun 14;15:1330993 [FREE Full text] [doi: 10.3389/fpsyt.2024.1330993] [Medline: 38947186]
48. Klocek A, Řiháček T. The associations between interoceptive awareness, emotion regulation, acceptance, and well-being in patients receiving multicomponent treatment: a dynamic panel network model. Res Psychother 2023 Jul 26;26(2):e25 [FREE Full text] [doi: 10.4081/ripppo.2023.659] [Medline: 37503659]
49. Golzar J, Tajik O, Noor S. Convenience sampling. Int J Lang Stud 2022;1(2):72. [doi: 10.22034/ijels.2022.162981]
50. Emerson RW. Convenience sampling revisited: embracing its limitations through thoughtful study design. J Vis Impair Blind 2021 Jan 15;115(1):76-77. [doi: 10.1177/0145482x20987707]
51. Kramer I, Simons CJ, Hartmann JA, Menne-Lothmann C, Viechtbauer W, Peeters F, et al. A therapeutic application of the experience sampling method in the treatment of depression: a randomized controlled trial. World Psychiatry 2014 Feb 04;13(1):68-77 [FREE Full text] [doi: 10.1002/wps.20090] [Medline: 24497255]
52. von Klipstein L, Servaas MN, Schoevers RA, van der Veen DC, Riese H. Integrating personalized experience sampling in psychotherapy: a case illustration of the Therap-i module. Heliyon 2023 Mar;9(3):e14507 [FREE Full text] [doi: 10.1016/j.heliyon.2023.e14507] [Medline: 36967959]
53. Palmier-Claus JE, Myin-Germeys I, Barkus E, Bentley L, Udachina A, Delespaul PA, et al. Experience sampling research in individuals with mental illness: reflections and guidance. Acta Psychiatr Scand 2011 Jan;123(1):12-20. [doi: 10.1111/j.1600-0447.2010.01596.x] [Medline: 20712828]
54. Joshi A, Kale S, Chandel S, Pal D. Likert scale: explored and explained. Br J Appl Sci Technol 2015 Jan 10;7(4):396-403. [doi: 10.9734/bjast/2015/14975]
55. Lobsy J, Wetmore A. CDC coffee break: using Likert scales in evaluation survey work. Centers for Disease Control and Prevention. 2014. URL: https://stacks.cdc.gov/view/cdc/136903 [accessed 2024-04-29]

Abbreviations

ES: experience sampling
ESM: experience sampling method
LMM: linear mixed model
SCED: single-case experimental design
SCRIBE: Single-Case Reporting Guideline in Behavioral Interventions
VR: virtual reality

XSL•FO
RenderX

Edited by A Mavragani; submitted 08.08.24; peer-reviewed by S Bogaerts; comments to author 29.10.24; revised version received 18.11.24; accepted 11.12.24; published 12.02.25.

Please cite as:
Klein Haneveld L, Dekkers T, Bouman YHA, Scholten H, Weerdmeester J, Kelders SM, Kip H
The Effect of the Virtual Reality–Based Biofeedback Intervention DEEP on Stress, Emotional Tension, and Anger in Forensic Psychiatric Inpatients: Mixed Methods Single-Case Experimental Design
JMIR Form Res 2025;9:e65206
URL: https://formative.jmir.org/2025/1/e65206
doi:10.2196/65206
PMID:

©Lisa Klein Haneveld, Tessa Dekkers, Yvonne H A Bouman, Hanneke Scholten, Joanneke Weerdmeester, Saskia M Kelders, Hanneke Kip. Originally published in JMIR Formative Research (https://formative.jmir.org), 12.02.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Machine Learning Approach to Identifying Wrong-Site Surgeries Using Centers for Medicare and Medicaid Services Dataset: Development and Validation Study

Yuan-Hsin Chen[1*], MPH, MD; Ching-Hsuan Lin[2*], MPH, MD; Chiao-Hsin Fan[3], MS; An Jim Long[3], PhD; Jeremiah Scholl[3], PhD; Yen-Pin Kao[4], MD; Usman Iqbal[5], MBA, PharmD, PhD; Yu-Chuan Jack Li[4,6,7,8], MD, PhD

[1]Department of Surgery, Massachusetts General Hospital, Boston, MA, United States
[2]Center for the Evaluation of Value and Risk in Health, Tufts Medical Center, Boston, MA, United States
[3]AESOP Technology Inc, Taipei, Taiwan
[4]Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, New Taipei City, Taiwan
[5]School of Population Health, Faculty of Medicine and Health, University of New South Wales, Sydney, Australia
[6]International Center for Health Information and Technology, College of Medical Science and Technology, Taipei Medical University, Taipei, Taiwan
[7]Department of Dermatology, Wan Fang Hospital, Taipei, Taiwan
[8]Research Center for Artificial Intelligence in Medicine, Taipei Medical University, Taipei, Taiwan
[*]these authors contributed equally

Corresponding Author:
Yu-Chuan Jack Li, MD, PhD
Graduate Institute of Biomedical Informatics
College of Medical Science and Technology
Taipei Medical University
No. 301, Yuantong Rd. Zhonghe Dist
New Taipei City, 235
Taiwan
Phone: 886 266382736
Email: jack@tmu.edu.tw

Abstract

Background: Wrong-site surgery (WSS) is a critical but preventable medical error, often resulting in severe patient harm and substantial financial costs. While protocols exist to reduce wrong-site surgery, underreporting and inconsistent documentation continue to contribute to its persistence. Machine learning (ML) models, which have shown success in detecting medication errors, may offer a solution by identifying unusual procedure-diagnosis combinations. This study investigated whether an ML approach can effectively adapt to detect surgical errors.

Objective: This study aimed to evaluate the transferability and effectiveness of an ML-based model for detecting inconsistencies within surgical documentation, particularly focusing on laterality discrepancies.

Methods: We used claims data from the Centers for Medicare and Medicaid Services Limited Data Set (CMS-LDS) from 2017 to 2020, focusing on surgical procedures with documented laterality. We developed an adapted Association Outlier Pattern (AOP) ML model to identify uncommon procedure-diagnosis combinations, specifically targeting discrepancies in laterality. The model was trained on data from 2017 to 2019 and tested on 2020 orthopedic procedures, using *ICD-10-PCS* (*International Classification of Diseases, Tenth Revision, Procedure Coding System*) codes to distinguish body part and laterality. Test cases were classified based on alignment between procedural and diagnostic laterality, with 2 key subgroups (right-left and left-right mismatches) identified for evaluation. Model performance was assessed by comparing precision-recall curves and accuracy against rule-based methods.

Results: The findings here included 346,382 claims, of which 2170 claims demonstrated with significant laterality discrepancies between procedures and diagnoses. Among patients with left-side procedures and right-side diagnoses (603/1106), 54.5% were confirmed as errors after clinical review. For right-side procedures with left-side diagnoses (541/1064), 50.8% were classified as errors. The AOP model identified 697 and 655 potentially unusual combinations in the left-right and right-left subgroups, respectively, with over 80% of these cases confirmed as errors following clinical review. Most confirmed errors involved

XSL•FO
RenderX

discrepancies in laterality for the same body part, while nonerror cases typically involved general diagnoses without specified laterality.

Conclusions: This investigation showed that the AOP model effectively detects inconsistencies between surgical procedures and diagnoses using CMS-LDS data. The AOP model outperformed traditional rule-based methods, offering higher accuracy in identifying errors. Moreover, the model's transferability from medication-disease associations to procedure-diagnosis verification highlights its broad applicability. By improving the precision of identifying laterality discrepancies, the AOP model can reduce surgical errors, particularly in orthopedic care. These findings suggest that the model enhances patient safety and has the potential to improve clinical decision-making and outcomes.

(JMIR Form Res 2025;9:e68436) doi:10.2196/68436

KEYWORDS

patient safety; wrong site surgery; medical errors; machine learning; claim data

Introduction

Wrong-site surgery (WSS) is recognized as the fifth most severe medical error in the United States [1]. A recent study revealed that one-third of inpatient WSS incidents resulted in high-severity injuries, with 7.4% of these cases leading to death [2]. This preventable mistake not only causes a big impact on patients but also imposes considerable financial burdens on the health care system [3]. According to a review in the United States National Practitioner Data Bank, a total of US $1.3 billion was recorded in payouts for surgical never events for 2 decades (1990-2010) [4]. Despite efforts aimed at reducing surgical errors through preprocedure verification, site marking, and a preoperative timeout, WSS events continue to occur. Specifically, the services most frequently responsible for those cases were orthopedic, including spine and intervertebral disc surgery, arthroscopy, and surgery on muscles or tendons [2,5].

The underreporting of WSS further complicates efforts to address this issue. In the United States, only 2% of sentinel events are reported to The Joint Commission, contributing to the persistence of WSS [1]. Moreover, the World Health Organization's 2024 Global Patient Safety Report highlights that only 38% of countries have implemented systems to report preventable and catastrophic medical errors, leading to the widespread underestimation of their true frequency [6]. Among the factors to WSS, inconsistent documentation and inappropriate history and examination information contribute significantly to the occurrence of WSS [2,3]. For example, a patient may have a radiology report indicating a right renal mass from an outside hospital, while a referral document refers to a recent ultrasound report of a left renal mass, leading to a left nephrectomy [2]. Due to the inescapable nature of human error, the ability to automatically identify such discrepancies with the emerging technology could substantially reduce the incidence of WSS.

To address issues of inconsistent documentation and inappropriate history and examination information, a machine learning (ML) model that can automatically detect uncommon associations between diagnosis and surgical procedures may offer a viable solution for identifying and preventing WSS. Recent studies have shown promising results using ML algorithms to detect medication errors, using unsupervised association rule learning to identify unreasonable medication-disease combinations [7,8]. We hypothesized that the similar characteristics between procedure-diagnosis and medication-disease associations could be used to train a model capable of detecting surgical errors. Therefore, this study aims to assess the transferability of an adapted ML model and validate its performance in detecting inconsistencies in surgical procedure documentation.

Methods

Data Source

The Centers for Medicare and Medicaid Services Limited Data Set (CMS-LDS) files provide detailed information on health care services given to beneficiaries. This database includes claims data that record the medical services provided, along with related diagnosis and procedure codes [9]. This dataset is crucial for health care research and quality assessments, helping to measure health care outcomes in various settings [10-12]. We included claims with complete records of both medical and surgical procedures.

Model Construction

In this study, we constructed the adapted version of the Association Outlier Pattern (AOP) model using the ML algorithms, with methods described in a previous paper [13]. We used CMS-LDS data from 2017 to 2019 as a training set to develop the model. This AOP model is designed to automatically detect uncommon or rare associations between specific diseases and surgical procedures, with a particular focus on discrepancies between the diagnosed side (left or right) and the side operated on. The model determined a procedure to be substantiated if the index procedure could be explained by a relevant diagnosis. However, if there was a procedure that could not be explained by any of the diagnoses, then the procedure-diagnosis combination would be viewed as unsubstantiated. All data collection and analysis procedures in this study fully comply with the principles outlined in the "Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research" [14].

Test Set Development

The study used inpatient CMS-LDS data from 2020 as a test set. We limited our testing data to orthopedic procedures, identified by the Attending Physician Specialty Code for Orthopedic Surgery. The study population was selected only if the *ICD-10-PCS* (*International Classification of Diseases, Tenth*

XSL•FO
RenderX

Revision, Procedure Coding System) and *ICD-10-CM* (*International Classification of Diseases, Tenth Revision, Clinical Modification*) codes included identifiable body parts and laterality information. The fourth character of the *ICD-10-PCS* code was used to differentiate body parts and laterality. Cases in which the only procedure performed was the insertion of an infusion device were excluded, as the rationale for the insertion site is typically clinician-determined and often lacks a documented record.

Procedures and diagnoses were categorized by laterality: right, left, or both. "Right-side procedures" indicated that all procedures performed on the patient were on the right side, and "both-side procedures" indicated that the patient received both right- and left-side procedures. The same definitions applied to the laterality of diagnoses. Patients with complete laterality information for both procedures and diagnoses were classified into 9 subgroups based on the combination of procedural and diagnostic laterality, such as right-right, right-left, right-both, left-right, left-left, left-both, both-right, both-left, and both-both. The analysis focused on 2 subgroups with significant discrepancies between procedure and diagnosis laterality: right-side surgery with left-side diagnosis (right-left) and left-side surgery with right-side diagnosis (left-right).

Test Set Evaluation

The accuracy of these 2 subgroups was evaluated through a rigorous review process that examined the alignment between documented procedures and diagnoses. Two authors (YHC and CHL) independently reviewed the consistency between procedures and diagnoses. In cases of disagreement, both authors conducted a thorough secondary review of all available medical records and reached a consensus through discussion. A confirmed error case was defined as a scenario in which no direct association existed between the surgery performed and any recorded diagnosis. Clinical criteria for defining an error case included: (1) all diagnoses did not sufficiently justify the need for the surgery; (2) the anatomical areas mentioned in the diagnoses and procedures were either not the same, not in proximity (eg, shoulder joint and humerus are in proximity), or not appropriately related (eg, upper leg relates to femur bone); and (3) the anatomical areas were similar, but the laterality was opposite. These criteria ensured that discrepancies in laterality between diagnoses and surgical procedures were systematically identified.

Confirmed error cases were further categorized based on whether the procedure and diagnosis involved the same or different body parts. Confirmed nonerror cases were classified into five categories based on potential rationale: (1) same body part between procedure and diagnosis, but with the unspecified side in the diagnosis; (2) same body part between procedure and diagnosis, but with no laterality specified; (3) diagnosis without a specific body part that can explain the index procedure; (4) anatomical areas in the diagnoses and procedures were in proximity; and (5) the index procedure could be explained by a combination of diagnoses and concurrent procedures.

Model Performance Metrics

To assess the generalization and extrapolation performance of the model, we applied the trained AOP model to identify potentially unusual combinations of procedures and diagnoses among the 2 subgroups with significant discrepancies (right-left and left-right) in the test set. We plotted the precision-recall curves to obtain the optimal probability threshold to maximize the predictive capability of the AOP model. Precision-recall curves have been considered an effective metric for accessing the model, especially the data is an unbalanced dataset [15,16]. The optimal probability threshold is where a point can achieve high precision score while only sacrificing minimal recall (Figure S1 in Multimedia Appendix 1). By calculating the accurate prediction rate of the AOP model and the rule-based method (right-left and left-right), we can compare the performance between these 2 methods.

Ethical Considerations

This study used the CMS-LDS files, which provide beneficiary-level information for research aimed at improving the US health care system. Access to the data was granted through the standard Centers for Medicare and Medicaid Services (CMS) approval process, including entering into a Data Use Agreement with CMS. The study was exempt from human subject research ethics review, as the data used were deidentified in accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. Since the study involved secondary analyses of deidentified data, informed consent was not required, and no direct interaction with human participants occurred.

To ensure privacy and confidentiality, all analyses strictly adhered to CMS policies, including the cell suppression rule, which prohibits the publication of results with cell sizes of 10 or fewer (eg, patients or services). Furthermore, no attempts were made to reidentify individuals from the data, maintaining compliance with ethical and privacy standards.

Results

Baseline Study Population

Data from a total of 346,382 claims were included in the study. Among them, the data indicated that 1106 patients had a left-side procedure with right-side diagnoses, and 1064 patients had a right-side procedure with left-side diagnoses (Figure 1).

XSL•FO
RenderX

Figure 1. Selection of test set study population from Centers for Medicare and Medicaid Services Limited Data Set (CMS-LDS).

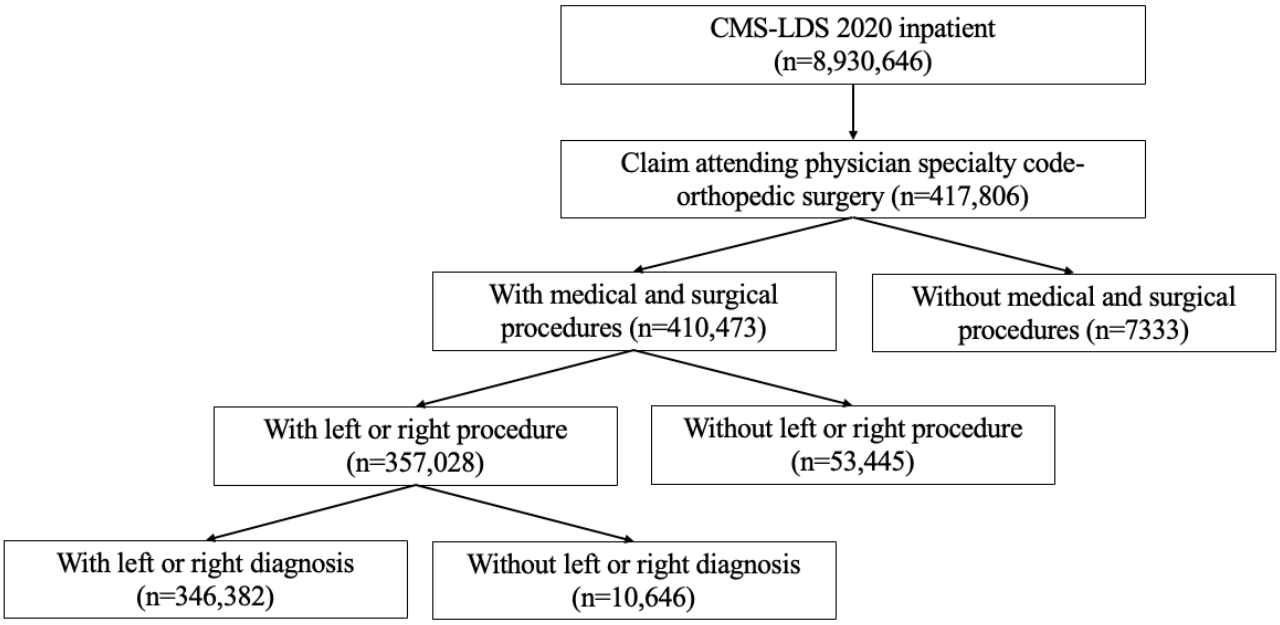

Subgroups based on the combination of procedural and diagnostic laterality.

Laterality	Left-side procedures	Right-side procedures	Both-side procedures
Left-side diagnoses	109,545	1064	1075
Right-side diagnoses	1106	128,101	871
Both-side diagnoses	48,147	50,013	6420

The study used inpatient CMS-LDS data from 2020 as the test set. Testing data was restricted to orthopedic procedures, and identified using the Attending Physician Specialty Code for Orthopedic Surgery. The study population included only cases where the *ICD-10* (*International Statistical Classification of Diseases, Tenth Revision*) procedure and diagnosis codes provided clear information about the body part and laterality.

Test Set Population Characteristics

Among the 1106 patients with left-side procedures with right-side diagnoses, 603/1106 (54.5%) cases were confirmed as errors, while 503/1106 (45.5%) were deemed reasonable combinations after applying the clinical review criteria. Similarly, among the 1064 patients with right-side procedures with left-side diagnoses, 541/1064 (50.8%) cases were confirmed as errors, while 523/1064 (49.2%) were deemed reasonable combinations (Table 1). Most confirmed errors involved the same body part in both procedure and diagnosis but with opposite laterality. Most nonerror cases were considered reasonable due to the presence of a relatively general diagnosis code with an unspecified body part, or a diagnosis with a specified body part but no laterality.

XSL•FO
RenderX

Table 1. Performance of rule-based method and association outlier pattern (AOP) model by laterality.

	Rule-based method[a]	AOP[b,c] model
Left-side procedure with right-side diagnoses		
Confirmed error	n=603 (54.5%)	n=576 (82.6%)
Same body part	n=421	n=404
Not same body part	n=182	n=172
Confirmed nonerror	n=503 (45.5%)	n=121 (17.4%)
Unspecified side of diagnosis[d]	n=39	n=17
No laterality for diagnosis[e]	n=131	n=34
No specific body part nor laterality[f]	n=251	n=58
Proximity area[g]	n=4	n=2
Others[h]	n=78	n=10
Right-side procedure with left-side diagnoses		
Confirmed error	n=541 (50.8%)	n=524 (80.0%)
Same body part	n=390	n=376
Not same body part	n=151	n=148
Confirmed nonerror	n=523 (49.2%)	n=131 (20.0%)
Unspecified side of diagnosis	n=50	n=17
No laterality for diagnosis	n=271	n=75
No specific body part nor laterality	n=129	n=25
Proximity area	n=4	n=4
Others	n=69	n=10

[a]Left-side procedure with right-side diagnoses: n=1106; right-side procedure with left-side diagnoses: n=1064.

[b]Left-side procedure with right-side diagnoses: n=697; right-side procedure with left-side diagnoses: n=655.

[c]AOP: association outlier pattern.

[d]Same body part between procedure and diagnosis, but with the unspecified side in the diagnosis.

[e]Same body part between procedure and diagnosis, but with no laterality specified.

[f]Diagnosis without a specific body part that can explain the index procedure.

[g]Anatomical areas in the diagnoses and procedures were in proximity.

[h]The index procedure could be explained by a combination of diagnoses and concurrent procedures.

This table summarizes the performance of the rule-based method and the AOP machine learning model in identifying surgical procedure-diagnosis inconsistencies in the CMS-LDS from 2020. The study focused on patients undergoing orthopedic procedures with discrepancies in laterality between procedures and diagnoses. Results are presented for 2 subgroups: left-side procedures with right-side diagnoses (n=1106) and right-side procedures with left-side diagnoses (n=1064). Clinical review outcomes are categorized into confirmed errors and nonerrors, with further classification based on the nature of the diagnosis and its alignment with the index procedure.

We focused on 2 subgroups with significant discrepancies regarding laterality between procedure and diagnosis, where combinations could be reasonable or unreasonable. For instance, a patient who underwent right knee joint arthroplasty but was diagnosed only with left knee osteoarthritis was deemed to have an error, as no diagnoses could justify the need for right knee joint replacement. Conversely, a nonerror case might involve a patient who underwent right lower leg detachment surgery with a diagnosis of diabetes-related gangrene (Table 2). Although the diagnosis did not specify the body part or laterality, it could explain the reason for the surgery and did not conflict with other diagnoses.

XSL•FO
RenderX

Table 2. Test set evaluation examples for clinical review of procedure-diagnosis discrepancies.

	Index procedures	Diagnoses
Confirmed error		
Same body part	• Right knee joint • 0SRC0J9-Replacement of right knee joint with synthetic substitute, cemented, open approach	• M1712-Unilateral primary osteoarthritis, left knee • F17210-Nicotine dependence, cigarettes, uncomplicated
Not same body part	• Left upper arm tendon, left shoulder joint • 0LS40ZZ-Reposition left upper arm tendon, open approach • 0RRK0JZ-Replacement of left shoulder joint with synthetic substitute, open approach	• M19021-Primary osteoarthritis and right elbow • E785-Hyperlipidemia, unspecified • I10-Essential (primary) hypertension • E7800-Pure hypercholesterolemia, unspecified
Confirmed non-error		
Unspecified body part and unspecified laterality	• Left hip joint • 0SPB0JZ-Removal of synthetic substitute from left hip joint, open approach • 0SRB049-Replacement of left hip joint with ceramic on polyethylene synthetic substitute, cemented, open approach	• T8484XA-Pain due to internal orthopedic prosthetic devices, implants and grafts, initial encounter • Y792-Prosthetic and other implants, materials and accessory orthopedic devices associated with adverse incidents • S0501XA-Injury of conjunctiva and corneal abrasion without foreign body, right eye, initial encounter • H10501-Unspecified blepharoconjunctivitis and right eye
Unspecified body part and unspecified laterality	• Right lower leg • 0Y6H0Z2-Detachment at right lower leg, mid, open approach	• E1152-Type 2 diabetes mellitus with diabetic peripheral angiopathy with gangrene • I96-Gangrene, not elsewhere classified • N186-End stage renal disease • I69354-Hemiplegia and hemiparesis following cerebral infarction affecting left nondominant side • E1122-Type 2 diabetes mellitus with diabetic chronic kidney disease • E785-Hyperlipidemia, unspecified • I252-Old myocardial infarction • Z85528-Personal history of other malignant neoplasm of kidney
Unspecified body part and unspecified laterality	• Right femoral shaft and right upper femur • 0QB80ZX-Excision of right femoral shaft, open approach, and diagnostic • 0QH606Z-Insertion of intramedullary internal fixation device into right upper femur, open approach	• C7951-Secondary malignant neoplasm of bone • C3402-Malignant neoplasm of left main bronchus • R200-Anesthesia of skin • R591-Generalized enlarged lymph nodes
No laterality for diagnosis	• Left knee • 0SPD08Z-Removal of spacer from left knee joint, open approach • 0SRD0J9-Replacement of left knee joint with synthetic substitute, cemented, open approach	• Z4733-Aftercare following explanation of knee joint prosthesis • E785-Hyperlipidemia, unspecified • I4510-Unspecified right bundle-branch block • I10-Essential (primary) hypertension

This table provides representative examples of confirmed errors and nonerrors from the test set of orthopedic procedures in the CMS-LDS data from 2020. Confirmed errors include cases where procedural laterality (eg, left or right) did not align with the diagnosis, while nonerrors involve reasonable justifications for the procedure based on provided diagnoses. Each example lists the index procedures, associated diagnoses, and rationale for classification based on clinical review.

AOP Model Performance Evaluation

Among the 1106 patients with left-side procedures with right-side diagnoses, the AOP model identified 697 potentially unusual combinations. After applying the clinical review criteria, 576 out of 697 (82.6%) cases were confirmed as errors, while 121 out of 697 (17.4%) were deemed reasonable. Similarly, among the 1064 patients with right-side procedures with left-side diagnoses, the AOP model identified 655 potentially unusual combinations. Of these, 524 out of 655 (80%) cases were confirmed as errors, and 131 out of 655 (20%) were reasonable combinations (Table 1). Most confirmed errors involved the

XSL•FO
RenderX

same body part in both procedure and diagnosis but with opposite laterality. Most nonerror cases were considered reasonable due to a general diagnosis with an unspecified body part or a diagnosis with a specified body part but no laterality.

An example of a confirmed error among AOP model-identified cases was a patient who received right shoulder joint arthroplasty, while the diagnoses only included a left humerus fracture. Since a left arm fracture could not justify right shoulder joint replacement, this case was deemed an error. Conversely, an example of a nonerror case was a patient who underwent left glenoid cavity replacement, with a diagnosis of unspecified side rotator cuff tear or rupture (Table 3). As an unspecified side shoulder injury could justify left shoulder surgery, this case was deemed a nonerror.

Table 3. Association outlier pattern (AOP) model evaluation examples for identifying unusual procedure-diagnosis combinations.

	Index procedures	Diagnoses
Confirmed error		
Same body part	• Left hip • 0SRB02A-Replacement of Left Hip Joint with Metal on Polyethylene Synthetic Substitute, Uncemented, Open Approach	• M1611 - Unilateral primary osteoarthritis, right hip • I10-Essential (primary) hypertension • F329-Major depressive disorder, single episode, unspecified
Not same body part	• Right shoulder joint • 0RRJ0J6-Replacement of right shoulder joint with synthetic substitute, humeral surface, open approach	• S42242A-4-part fracture of surgical neck of left humerus, initial encounter for closed fracture • Z880-Allergy status to penicillin • Z79899-Other long term (current) drug therapy
Confirmed non-error		
Unspecified body part and unspecified laterality	• Right hip bursa • 0MBL0ZZ-Excision of right hip bursa and ligament, open approach	• T8131XA-Disruption of external operation (surgical) wound, not elsewhere classified, initial encounter • G40909-Epilepsy, unspecified, not intractable, and without status epilepticus • I110-Hypertensive heart disease with heart failure • E669-Obesity, unspecified • G2581-Restless legs syndrome • Z96642-Presence of left artificial hip joint • B964-Proteus (mirabilis) (morganii) as the cause of diseases classified elsewhere • T8141XA-Infection following a procedure, superficial incisional surgical site, initial encounter • I428-Other cardiomyopathies • Y838-Other surgical procedures as the cause of abnormal reaction of the patient, or of later complication, without mention of misadventure at the time of the procedure
No laterality for diagnosis	• Left glenoid cavity • 0PR80JZ-Replacement of left glenoid cavity with synthetic substitute, open approach	• M75120-Complete rotator cuff tear or rupture of unspecified shoulder, not specified as traumatic • G8918-Other acute postprocedural pain • M109-Gout and unspecified • Z89511-Acquired absence of right leg below the knee

This table illustrates specific examples of confirmed errors and nonerrors identified by the AOP machine learning model in the CMS-LDS 2020 test set. Confirmed errors are characterized by laterality discrepancies between the procedure and diagnosis, while nonerrors reflect plausible procedure-diagnosis relationships despite general or unspecified diagnoses. Each case includes the index procedures, associated diagnoses, and the rationale for classification.

Discussion

Principal Findings

In this study, we evaluated the effectiveness of the AOP model in detecting whether surgical procedures are substantiated with diagnoses records using CMS-LDS data. We found the AOP model is more accurate in identifying inconsistencies in surgical procedure documentation when compared with the traditional rule-based method. Meanwhile, the adapted AOP model also showed the transferability from detecting the medication-disease association and expanding to verifying the procedure-diagnosis association.

Our study adds to the existing literature that ML has penetrated the medical field with great success. ML allows for more detail to be mined from the data, allowing for the development of better diagnostic and prognostic tools than traditional approaches [17,18]. Similarly, the AOP model can detect associations between specific diseases and surgical procedures by learning patterns from a large database, while the rule-based method depends on the laterality of procedures and diagnoses. For example, among the 2 subgroups with significant discrepancies

XSL•FO
RenderX

regarding laterality between procedure and diagnosis (left-right and right-left), a rule-based method classified all 2170 cases as inappropriate combinations due to the conflict in laterality. In contrast, the AOP model classified only 1352 of these cases as inappropriate. In terms of identifying inappropriate procedure-diagnosis combinations, the AOP model identified most errors similarly to the rule-based method (1100 for AOP vs 1144 for rule-based). In addition, the AOP performed with much higher accuracy (1100/1352, 81.4%) compared with the rule-based method (1144/2170, 52.7%).

The AOP model outperformed the traditional rule-based model, demonstrating a lower rate of false signals. Comparing the confirmed nonerror percentage between the rule-based method and the AOP model, there was a significantly lower rate of confirmed nonerror cases identified by the AOP model. In these cases, most confirmed nonerrors identified by the rule-based method could be explained by other diagnoses without mentioning body part or laterality. For example, cancer metastasis to bone may explain the reason for pelvic surgery, even without a specific pelvic metastasis diagnosis. This is because some diagnoses may not provide detailed information about the patient's condition. Similarly, a patient who underwent lower leg detachment surgery with a diagnosis of diabetes-related gangrene could have the surgery justified, even if the gangrene diagnosis did not specify the body part. The rule-based method focuses solely on the laterality of the diagnosis, whereas the AOP model uses ML methods to explain the association between procedures and diagnoses. Therefore, when the association is deemed normal, the AOP model does not label them as inappropriate.

This study also showed several noteworthy findings. First, cases with fewer diagnoses were more likely to be labeled as inappropriate procedure-diagnosis combinations. For example, a patient who underwent surgeries for the left upper arm and left shoulder had only 4 diagnoses, including a right elbow problem and other chronic diseases. This may indicate incomplete diagnosis coding or an error in the surgical site. Second, the AOP model effectively identified unusual procedure-diagnosis combinations within subgroups where traditional rule-based methods failed to do so. Specifically, procedures that aligned with their corresponding diagnoses were deemed appropriate by conventional methods, such as left-side procedures with left-side diagnoses (left-left) and right-side procedures with right-side diagnoses (right-right). However, the AOP model continued to uncover atypical combinations within these groups. For instance, the model identified a patient who underwent left hip arthroplasty despite having a diagnosis of left knee osteoarthritis only.

The strengths of our study include its being the first to explore whether a ML model could play a surveillance role in detecting wrong site surgery. The effectiveness and fast pace to review the inconsistencies of surgical documentation help increase reporting mistakes in a timely manner without putting too much workforce on this issue. The findings from our study raise important questions regarding whether these errors stem from documentation mistakes or represent actual incidents involving patients. Therefore, in addition to retrospective reviews, it is essential to leverage modern technology to facilitate real-time error prevention. The potential of our model could add value on the current technology to encompass proactive prediction and analysis of diagnoses and clinical evidence to inform accurate surgical decisions and improve documentation practices [19]. In addition, the potential adaptability of the AOP model extends beyond orthopedic surgery to other specialties where procedural and diagnostic codes require logical alignment. For example, in ophthalmic surgery, the model could identify mismatches such as a diagnosis of left-sided cataract paired with a procedure for right-eye cataract surgery. Similarly, otolaryngology and other surgical fields with specific laterality could benefit. However, certain specialties may pose unique challenges due to differences in coding systems or the absence of laterality indicators, requiring further validation and specialty-specific refinements.

The AOP model has limitations. There were still nonerror cases identified by the AOP model. For instance, a patient who underwent left glenoid cavity replacement surgery with a diagnosis of unspecified rotator cuff tear was marked as inappropriate by the AOP model. However, the diagnosis could explain the necessity of the surgery. The reason the AOP model labeled this as an inappropriate combination may be due to the low frequency of this procedure-diagnosis combination in the CMS dataset. If the training dataset has a limited number of surgeries or diagnoses, then the association may be viewed as low and therefore marked as an inappropriate combination.

Another limitation is that the AOP model failed to identify some error cases labeled by the rule-based method. For example, a patient received an internal fixation surgery of the right upper femur with diagnoses of secondary malignant neoplasm of bone and left femur osteonecrosis. While the diagnosis of secondary malignant neoplasm could justify the surgical necessity, the diagnosis of left femur osteonecrosis conflicted with the right femur surgery. In such cases, the AOP model uses the maximum association within all procedure-diagnosis combinations, meaning that if the association between the surgery and a diagnosis is high, the model will view the combination as appropriate, even if the other diagnosis is in conflict. Therefore, in cases where one diagnosis can justify surgery, but another is conflicting, the AOP model may not accurately identify the combination. However, this aspect of the AOP model presents an opportunity for optimization, and future adjustments incorporating more detailed information through advanced ML techniques could yield a more accurate model, especially in cases where diagnoses are vague, incomplete, or lack clear specifications of body parts or sides.

Building on these strengths and limitations, future steps for implementing the AOP model in electronic health record systems could proceed along 2 directions. First, the model could support retrospective reviews by efficiently identifying WSS events, reducing review times and workforce burden, and enabling more timely reporting. Second, it could facilitate real-time consistency checks when clinicians input procedure codes. For example, if a surgeon enters a right-sided knee arthroplasty for a patient diagnosed with left knee osteoarthritis, the AOP model could flag the discrepancy and issue an alert. Such real-time notifications may prevent WSS before it occurs. However, to achieve successful implementation, the model must

XSL•FO
RenderX

be optimized to address the limitations, and its performance must be validated across diverse electronic health record systems and datasets across different facilities.

Conclusion

Overall, the AOP model enhances patient safety by systematically analyzing the cooccurrence frequency of diseases and surgical procedures on opposite sides, thereby identifying potentially unusual procedure-diagnosis combinations. Our analysis demonstrated a significant improvement in accuracy rates when the AOP model was used, compared with scenarios where it was not applied. By improving precision in identifying and validating cases with discrepancies between the diagnosed and operated sides, the AOP model has the potential to reduce the risk of errors in orthopedic treatment planning. This enhanced accuracy enables health care providers to make more informed decisions, reducing the risk of WSS and ultimately contributing to improved patient outcomes in orthopedic health care settings.

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or non-profit sectors.

Generative artificial intelligence was not used in any portion of the manuscript writing.

Data Availability

This study used Centers for Medicare and Medicaid Services Limited Data Set (CMS-LDS) files. Access to these data is restricted and governed by CMS. Researchers interested in accessing the data must follow CMS's application process and enter into a Data Use Agreement. Due to these restrictions, the data cannot be deposited in publicly available repositories. Requests for data access should be directed to CMS.

Authors' Contributions

YHC, CHL, and YPK were contributed to conceptualization. YHC, CHF, AJL, and JS were handled methodology. YHC and CHL were writing–original draft preparation, formal analysis, and visualization. CHF and AJL were performed data curation. JS, YPK, and UI were writing–review and editing. YCJL managed supervision

Conflicts of Interest

JS, AJL, and Y-CJL are significant shareholders in, and continue to work with, AESOP Technology Inc, the company that developed and owns IP related to the AOP model in the study. C-HF is an employee at AESOP Technology Inc.

Multimedia Appendix 1
Supplement Figure 1. Precision-Recall curve of Association Outlier Pattern (AOP) Model in Test Set.
[PNG File , 87 KB - formative_v9i1e68436_app1.png]

References

1. Robinson TP, Bilimoria KY, Yang AD. Understanding a surgeon's worst nightmare: wrong site surgery. Jt Comm J Qual Patient Saf 2023;49(5):237-238. [doi: 10.1016/j.jcjq.2023.03.006] [Medline: 36990955]
2. Tan J, Ross JM, Wright D, Pimentel MPT, Urman RD. A contemporary analysis of closed claims related to wrong-site surgery. Jt Comm J Qual Patient Saf 2023;49(5):265-273. [doi: 10.1016/j.jcjq.2023.02.002] [Medline: 36925434]
3. Zil-E-Ali A, Laubscher L, Kourampi I, Tsagkaris C. Is surgery on the right track? The burden of wrong-site surgery. Proc (Bayl Univ Med Cent) 2023;36(5):657-660 [FREE Full text] [doi: 10.1080/08998280.2023.2231714] [Medline: 37614864]
4. Mehtsun WT, Ibrahim AM, Diener-West M, Pronovost PJ, Makary MA. Surgical never events in the United States. Surgery 2013;153(4):465-472. [doi: 10.1016/j.surg.2012.10.005] [Medline: 23257079]
5. Neily J, Mills PD, Eldridge N, Dunn EJ, Samples C, Turner JR, et al. Incorrect surgical procedures within and outside of the operating room. Arch Surg 2009;144(11):1028-1034. [doi: 10.1001/archsurg.2009.126] [Medline: 19917939]
6. Global patient safety report 2024, Geneva: World Health Organization. URL: https://iris.who.int/bitstream/handle/10665/376928/9789240095458-eng.pdf?sequence=1 [accessed 2024-05-30]
7. Chin YPH, Song W, Lien CE, Yoon CH, Wang WC, Liu J, et al. Assessing the international transferability of a machine learning model for detecting medication error in the general internal medicine clinic: multicenter preliminary validation study. JMIR Med Inform 2021;9(1):e23454 [FREE Full text] [doi: 10.2196/23454] [Medline: 33502331]
8. Huang CY, Nguyen PA, Yang HC, Islam MM, Liang CW, Lee FP, et al. A probabilistic model for reducing medication errors: a sensitivity analysis using electronic health records data. Comput Methods Programs Biomed 2019;170:31-38. [doi: 10.1016/j.cmpb.2018.12.033] [Medline: 30712602]
9. Limited Data Set (LDS) Files | CMS. URL: https://www.cms.gov/data-research/files-for-order/limited-data-set-lds-files [accessed 2025-10-25]

XSL•FO
RenderX

10. Hakkarainen TW, Arbabi S, Willis MM, Davidson GH, Flum DR. Outcomes of patients discharged to skilled nursing facilities after acute care hospitalizations. Ann Surg 2016;263(2):280-285 [FREE Full text] [doi: 10.1097/SLA.0000000000001367] [Medline: 26445466]
11. Avila A, Sato EH, Acuña AJ, Vasavada K, Kamath AF. Gender differences in reimbursement among orthopaedic surgeons: a cross-sectional analysis of medicare claims. J Am Acad Orthop Surg 2023;31(15):e570-e578. [doi: 10.5435/JAAOS-D-22-00823] [Medline: 37253194]
12. Adashi EY, Gruppuso PA. The centers for medicare and medicaid services (CMS) COVID-19 brief: unsettling racial and ethnic health disparities. J Am Board Fam Med 2021;34(Supplement):S13-S15. [doi: 10.3122/jabfm.2021.s1.200450]
13. Nguyen PA, Syed-Abdul S, Iqbal U, Hsu MH, Huang CL, Li HC, et al. A probabilistic model for reducing medication errors. PLoS One 2013;8(12):e82401 [FREE Full text] [doi: 10.1371/journal.pone.0082401] [Medline: 24312659]
14. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. J Med Internet Res 2016;18(12):e323 [FREE Full text] [doi: 10.2196/jmir.5870] [Medline: 27986644]
15. Muschelli J. ROC and AUC with a binary predictor: a potentially misleading metric. J Classif 2020;37(3):696-708 [FREE Full text] [doi: 10.1007/s00357-019-09345-1] [Medline: 33250548]
16. Zhou QM, Zhe L, Brooke RJ, Hudson MM, Yuan Y. A relationship between the incremental values of area under the ROC curve and of area under the precision-recall curve. Diagn Progn Res 2021;5(1):13 [FREE Full text] [doi: 10.1186/s41512-021-00102-w] [Medline: 34261544]
17. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform 2018;19(6):1236-1246 [FREE Full text] [doi: 10.1093/bib/bbx044] [Medline: 28481991]
18. Wang K, Shi Q, Sun C, Liu W, Yau V, Xu C, et al. Corrigendum: a machine learning model for visualization and dynamic clinical prediction of stroke recurrence in acute ischemic stroke patients: a real-world retrospective study. Front Neurosci 2023;17:1235340 [FREE Full text] [doi: 10.3389/fnins.2023.1235340] [Medline: 37492408]
19. Gloystein DM, Heiges BA, Schwartz DG, DeVine JG, Spratt D. Innovative technology system to prevent wrong site surgery and capture near misses: a multi-center review of 487 cases. Front Surg 2020;7:563337 [FREE Full text] [doi: 10.3389/fsurg.2020.563337] [Medline: 33195386]

Abbreviations

AOP: association Outlier Pattern
CMS: Centers for Medicare and Medicaid Services
CMS-LDS: Centers for Medicare and Medicaid Services Limited Data Set
HIPAA: Health Insurance Portability and Accountability Act
ICD-10: International Statistical Classification of Diseases, Tenth Revision
ICD-10-CM: International Classification of Diseases, Tenth Revision, Clinical Modification
ICD-10–PCS: International Classification of Diseases, Tenth Revision, Procedure Coding System
ML: machine learning
WSS: Wrong-site surgery

Edited by A Mavragani; submitted 06.11.24; peer-reviewed by S Bharati; comments to author 06.12.24; revised version received 20.12.24; accepted 12.01.25; published 13.02.25.

Please cite as:
Chen YH, Lin CH, Fan CH, Long AJ, Scholl J, Kao YP, Iqbal U, Li YCJ
Machine Learning Approach to Identifying Wrong-Site Surgeries Using Centers for Medicare and Medicaid Services Dataset: Development and Validation Study
JMIR Form Res 2025;9:e68436
URL: https://formative.jmir.org/2025/1/e68436
doi:10.2196/68436
PMID:

©Yuan-Hsin Chen, Ching-Hsuan Lin, Chiao-Hsin Fan, An Jim Long, Jeremiah Scholl, Yen-Pin Kao, Usman Iqbal, Yu-Chuan Jack Li. Originally published in JMIR Formative Research (https://formative.jmir.org), 13.02.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Exploring the Acceptance of Just-in-Time Adaptive Lifestyle Support for People With Type 2 Diabetes: Qualitative Acceptability Study

Eclaire A G Hietbrink[1,2], RN, MSc; Anouk Middelweerd[1], PhD; Wendy d'Hollosy[1], PhD; Laura K Schrijver[1,2], MSc; Gozewijn D Laverman[1,2], MD, PhD; Miriam M R Vollenbroek-Hutten[1,3], PhD

[1]Department of Biomedical Signals and Systems, University of Twente, Enschede, The Netherlands
[2]Department of Internal Medicine/Nephrology, Ziekenhuisgroep Twente, Almelo, The Netherlands
[3]Medisch Spectrum Twente, Enschede, The Netherlands

Corresponding Author:
Eclaire A G Hietbrink, RN, MSc
Department of Biomedical Signals and Systems
University of Twente
Drienerlolaan 5
Enschede, 7500 AE
The Netherlands
Phone: 31 53 4892760
Email: e.a.g.hietbrink@utwente.nl

Abstract

Background: The management of type 2 diabetes (T2D) requires individuals to adopt and maintain a healthy lifestyle. Personalized eHealth interventions can help individuals change their lifestyle behavior. Specifically, just-in-time adaptive interventions (JITAIs) offer a promising approach to provide tailored support to encourage healthy behaviors. Low-effort self-reporting via ecological momentary assessment (EMA) can provide insights into individuals' experiences and environmental factors and thus improve JITAI support, particularly for conditions that cannot be measured by sensors. We developed an EMA-driven JITAI to offer tailored support for various personal and environmental factors influencing healthy behavior in individuals with T2D.

Objective: This study aimed to assess the acceptability of EMA-driven, just-in-time adaptive lifestyle support in individuals with T2D.

Methods: In total, 8 individuals with T2D used the JITAI for 2 weeks. Participants completed daily EMAs about their activity, location, mood, overall condition, weather, and cravings and received tailored support via SMS text messaging. The acceptability of the JITAI was assessed through telephone-conducted, semistructured interviews. Interview topics included the acceptability of the EMA content and prompts, the intervention options, and the overall use of the JITAI. Data were analyzed using a hybrid approach of thematic analysis.

Results: Participants with a mean age of 70.5 (SD 9) years, BMI of 32.1 (SD 5.3) kg/m², and T2D duration of 15.6 (SD 7.7) years had high self-efficacy scores in physical activity (ie, 32) and nutrition (ie, 29) and were mainly initiating or maintaining behavior changes. The identified themes were related to the intervention design, decision points, tailoring variables, intervention options, and mechanisms underlying adherence and retention. Participants provided positive feedback on several aspects of the JITAI, such as the motivating and enjoyable messages that appeared well tailored to some individuals. However, there were notable differences in individual experiences with the JITAI, particularly regarding intervention intensity and the perceived personalization of the EMA and messages. The EMA was perceived as easy to use and low in burden, but participants felt it provided too much of a snapshot and too little context, reducing the perceived tailoring of the intervention options. Challenges with the timing and frequency of prompts and the relevance of some tailoring variables were also observed. While some participants found the support relevant and motivating, others were less inclined to follow the advice. Participants expressed the need for even more personalized support tailored to their specific characteristics and circumstances.

Conclusions: This study showed that an EMA-driven JITAI can provide motivating and tailored support, but more personalization is needed to ensure that the lifestyle support more closely fits each individual's unique needs. Key areas for improvement include developing more individually tailored interventions, improving assessment methods to balance active and passive data collection, and integrating JITAIs within comprehensive lifestyle interventions.

XSL•FO
RenderX

(JMIR Form Res 2025;9:e65026) doi:10.2196/65026

KEYWORDS

eHealth; just-in-time adaptive intervention; ecological momentary assessment; type 2 diabetes; behavior change; physical activity; nutrition; acceptability; formative evaluation; mobile phone

Introduction

Background

A healthy lifestyle has a pivotal role in effective management of type 2 diabetes (T2D) [1,2]. This approach includes addressing important contributors to T2D, such as insufficient physical activity, unhealthy dietary habits, inadequate sleep, and elevated stress levels. Substantial scientific evidence supports the substantial health benefits of lifestyle interventions [3-8], including studies indicating that lifestyle modifications can potentially reverse or achieve remission in T2D. In addition, a 10-year study demonstrated that lifestyle interventions could result in economic benefits in the form of reduced health care costs [9].

However, individuals with T2D often struggle to adhere to lifestyle guidelines [10-13]. One explanation is that social and environmental changes, such as the prevalence of sedentary jobs and the widespread availability of highly processed foods, have made it increasingly challenging for individuals to adopt and maintain a healthy lifestyle [14]. Moreover, managing a chronic condition poses additional challenges, including the time and effort required to manage symptoms, adhere to treatment plans, and cope with physical and psychosocial consequences [15,16]. Individuals need knowledge and self-management skills to implement sustainable lifestyle changes [2]. Due to variations in personal, environmental, and disease-related factors, self-management skills can differ substantially from person to person [17]. Therefore, a personalized approach is essential to guide each individual toward a healthier lifestyle, providing tailored support that meets their unique needs.

eHealth, which involves using technology to promote health, well-being, and health care [18], holds considerable potential in delivering personalized lifestyle support. Research has demonstrated that eHealth interventions can be both acceptable and effective for improving physical activity, diet, and health outcomes [19-23]. A promising approach to provide personalized support is just-in-time adaptive interventions (JITAIs). JITAIs are designed to dynamically provide the right support at the right time [24]. A key element of JITAIs is the need for insights into behavior and context to adapt the support accordingly. Today, wearables, smartphone sensors, and remote monitoring devices, such as activity trackers, GPS, and smartphone calendars, can be used for real-time monitoring and to inform intervention decisions [24,25]. In addition, low-effort self-reporting via ecological momentary assessment (EMA) can provide further insights into an individual's situation and emotional state, which may improve the support offered by JITAIs [26]. EMA involves repeated sampling of an individual's behaviors, cognitions, emotions, and environmental factors in their natural environment through short questionnaires that are prompted multiple times within a certain time frame [27]. EMAs allow for adapting support to individual needs, particularly for conditions that cannot be measured by sensors. Unlike typical eHealth interventions, JITAIs can provide near–real-time and contextually relevant support, tailored to an individual's needs and preferences [24].

JITAIs have been studied to determine their added value in supporting several lifestyle behaviors, such as physical activity, diet, and substance use. Studies have examined their acceptability; feasibility; and, to a lesser extent, effectiveness [28-32]. Overall, JITAIs appear to be acceptable, although some feasibility challenges were identified, including issues with smartphone battery life, sensor reliability, the punctuality of JITAI messages, and the incomplete automation of JITAI delivery [28,31]. People were more likely to follow JITAI messages and to find them more relevant than randomly selected messages [33,34]. Evidence on the effectiveness of JITAIs for behavior change is mixed, and there is insufficient data to draw well-founded conclusions [28,29,32]. JITAIs that use algorithms to tailor interventions based on past behavior and current needs, along with guidance from a human, are more effective [30].

Objectives

Our research team previously developed the E-Supporter, an evidence-based and tailored digital coach for people with chronic conditions, primarily T2D [35]. A previous feasibility study on the E-Supporter revealed that individuals with T2D need more tailored support to better align with their personal circumstances [36]. This finding highlights the need for interventions that can adapt to individual needs. JITAIs offer the opportunity to provide personalized, contextually relevant support when it is needed. In many JITAIs, tailoring is often based on objective parameters related to healthy lifestyles, such as wearable data [37-39]. However, these eHealth interventions do not frequently address the more nuanced aspects of individuals' experiences, such as disease-related challenges or environmental factors [28]. While a few JITAIs assess these conditions via EMA [26,32], none currently tailor support to the unique characteristics of individuals with T2D. Therefore, we developed an EMA-driven JITAI designed to offer tailored assistance for several personal and environmental factors that influence healthy behavior. Given its novelty, assessing acceptability is crucial for refinements and future development [18,40]. Hence, we conducted an acceptability assessment for the EMA-driven JITAI designed to support a healthy lifestyle for individuals with T2D.

Methods

Study Design

We performed a qualitative study to assess the acceptability of the JITAI over 2 weeks among individuals with T2D. The JITAI was designed to support physical activity and healthy nutrition in individuals with T2D. The study was conducted during a

XSL•FO
RenderX

national lockdown due to COVID-19, requiring all contact with participants to take place remotely.

Intervention Description

Overview

The foundation of the developed JITAI is rooted in the E-Supporter 1.0 [35], designed to promote physical activity and a healthy diet among individuals with chronic diseases, initially focusing on T2D. The E-Supporter 1.0 was participatory developed with input from both patients with T2D and health care professionals. It is tailored to several variables, including behavioral objectives, the phase of behavior change, type of chronic disease, time of day, and goal achievement. The behavior change strategies are based on the *Health Action Process Approach* [41] and theories to support behavior maintenance [42,43]. An in-depth description of the E-Supporter 1.0 can be found elsewhere [35].

We used a systematic approach to design the JITAI according to the principles outlined in the framework by Nahum-Shani et al [24]. This process involved developing the 6 components of a JITAI: distal outcome, proximal outcome, tailoring variables, intervention options, decision rules, and decision points (Figure 1 [24]). The development process involved iterative feedback on all components from a group of multidisciplinary experts, including a health scientist, sports scientist, computer scientist, and diabetologist.

Figure 1. Overview of the intervention components based on the conceptual model of just-in-time adaptive interventions. EMA: ecological momentary assessment.

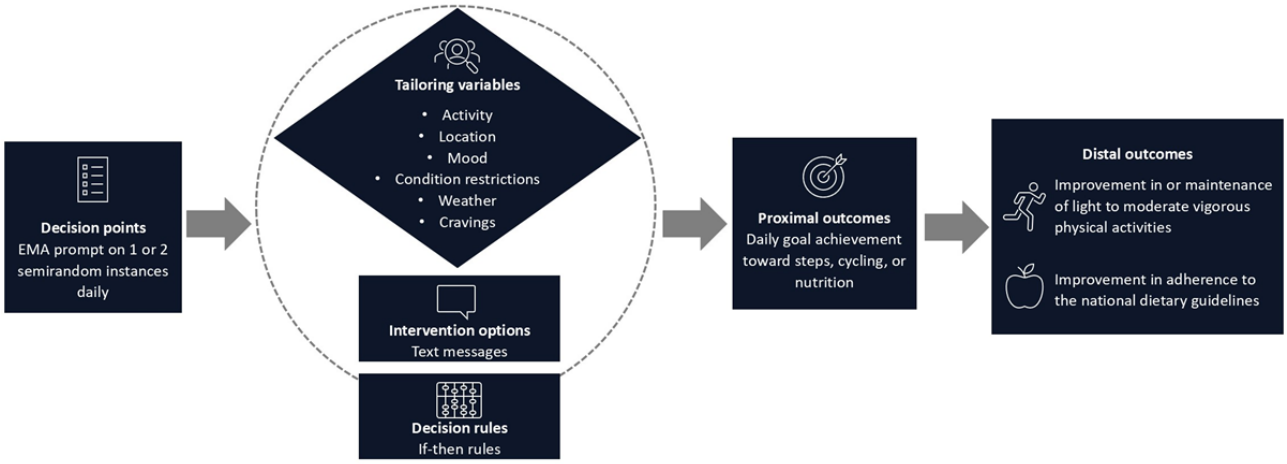

Distal Outcome

The distal outcome represents *the ultimate long-term goal* of the intervention, indicating the desired outcome that is expected to be achieved [24]. The distal outcome regarding the physical activity module is improvement in, or maintenance of, light to moderate vigorous physical activities in people with T2D. Regarding the nutritional module, the distal outcome is improvement in adherence to the national dietary guidelines [44]. A substantiation for the choice of the distal outcomes can be found in the study by Hietbrink et al [35].

Proximal Outcome

The proximal outcome represents *the short-term goal* that the intervention aims to achieve and indicates the progress to the distal outcome [24]. For this JITAI, we opted for a proximal outcome centered on actual behavior. The E-Supporter 1.0 allows users to set specific goals for steps, cycling, or nutrition, contributing to the attainment of distal outcomes [35]. Consequently, we focused on *daily goal achievement* as our proximal outcome, which is considered a meaningful indicator of short-term behavioral progress toward the distal outcome [24].

Tailoring Variables

Tailoring variables refer to individual-specific information used for personalization, determining when and how to intervene [24]. To identify tailoring variables, we conducted a survey assessing facilitators and barriers for a healthy lifestyle among individuals with a lifestyle-related chronic condition (L K Schrijver, MSc, unpublished data, March 2020). In the domain of physical activity, important dynamic factors included weather, mood, daily activities, fatigue, disease symptoms, pain, and social support. Regarding dietary choices, mood, location (eg, eating out in a restaurant), social occasions, and the presence of unhealthy foods emerged as dynamic factors. The research group clustered these factors into the tailoring variables for the JITAI. The following 6 tailoring variables were identified: activity, location, mood, physical or conditional restrictions, weather, and cravings. In this study, we used EMA as an active assessment method (ie, input from the user is required) to evaluate all tailoring variables. Each variable was addressed through a specific EMA question. The research team designed, evaluated, and refined the EMA questions and corresponding response options. Table 1 presents an overview of the tailored EMA questions designed to assess the identified variables.

XSL•FO
RenderX

Table 1. Ecological momentary assessment questions and corresponding response options designed to assess the tailoring variables during the study period.

Tailoring variable	Question	Response options	Coaching domain (ie, physical activity or nutrition)
Activity	What are you currently doing?	• At work or studying • Leisure activity • Social occasion • Traveling • Eating (ie, other than a social occasion) • Being physically active • Otherwise	Both
Location	What is your current location?	• At home • At someone else's house • At work • In a vehicle • In a restaurant or café or on a terrace • At a sports club or other sports location • Outside, not otherwise specified • Otherwise	Both
Mood	How are you feeling at the moment?	• I feel positive emotions (eg, happy, cheerful, loved, confident, or relaxed) • I feel negative emotions (eg, sad, angry, stressed, ashamed, or anxious) • I feel neutral (neither positive nor negative)	Both
Restrictions in physical or mental condition	Are you currently experiencing restrictions in your overall condition?	• I experience no restrictions • I experience discomfort due to my physical limitations • I experience pain complaints • I experience diabetes-specific symptoms (eg, hyperglycemia or hypoglycemia) • I experience fatigue • Otherwise	Physical activity
Weather	How do you feel about the weather for doing physical activity outdoors?	• The weather is nice to be physically active outside • The weather is not suitable (ie, too hot, cold, or wet) to be physically active outside	Physical activity
Cravings	Are you currently craving food?	• No, I do not feel like eating • Yes, I feel like eating healthy • Yes, I feel like eating unhealthy • Yes, but I do not crave a specific food	Nutrition

Intervention Options

Intervention options represent the diverse behavior change actions that can be taken in response to decision points [24]. The content of the intervention options was rooted in the theoretical framework of the E-Supporter 1.0 [35], structured around 3 behavioral change phases: initiation, action, and maintenance. For each phase, relevant determinants of behavior were identified, and corresponding behavior change techniques (BCTs) were integrated to influence behavior effectively. BCTs are active components included in interventions designed to change behavior through various mechanisms [45].

The JITAI contained a set of 215 short SMS text messages, each tailored to address the tailoring variables (Table 2). These messages aimed to build skills, motivate, prompt reflection, and provide encouragement. Input to develop the intervention options was gathered from the following sources: the literature, the Dutch Diabetes Fund, the Dutch Nutrition Center, and the Dutch guidelines for physical activity and healthy diet [44,46]. The determinants and BCTs from the E-Supporter 1.0 were used as basis to develop the JITAI intervention options [35]. In addition, new determinants and BCTs were added to align with the selected tailoring variables, such as activity and location, which use environmental cues and context to prompt and guide individuals in various settings (eg, work, home, or during specific activities) [47]. For a comprehensive overview of determinants and BCTs, please refer to Multimedia Appendix 1.

XSL•FO
RenderX

Table 2. Examples of intervention options provided during the study, featuring tailored text messages customized for each tailoring variable.

Tailoring variable	Answer option	Example intervention option
Activity	At work or studying	"Do you have a sedentary job? Scheduling a daily walk during your break can help. Maybe your colleagues would like to join you. It is not only good for physical activity but also a fun way to spend time together!"
Location	In a restaurant or café or on a terrace	"Hi! Eating out at a restaurant? Try to choose the healthier option on the menu. Here are some examples of fairly healthy main courses: lean meat (eg, chicken or turkey fillet) or fatty fish (eg, salmon), served with vegetables, potatoes, or whole grain rice."
Weather	The weather is nice to be physically active outside	"Hello *[name]*, with this beautiful weather, everything around you looks even nicer. Take a walk around the neighborhood and enjoy it!"
Mood	I feel negative emotions (think sad, angry, stressed, angry, ashamed, or anxious)	"Hello *[name]*, I am sorry you are not feeling great. Even when you are feeling down, a walk might help. Give it a try, it could do you some good!"
Physical restrictions	I experience fatigue	"Hi, Feeling tired? You can break up your daily physical activity into shorter sessions. That way, you will have time to rest in between."
Food cravings	Yes, I feel like eating unhealthy food	"If the temptation to snack is strong, doing something else can help, like going for a walk or talking to someone. A little distraction can make the craving fade!"

Decision Rules

The decision rules serve as the link between tailoring variables and intervention options, enabling personalized adaptation of the intervention based on when, for whom, and which intervention option to offer [24]. All intervention options were systematically coded according to a behavioral goal, phase of behavior change, determinants of behavior, BCTs, tailoring variable from the EMA, and time of day. In our intervention, we used rule-based decisions with if-then statements—if certain conditions were met, then a specific intervention option was provided. A fitting intervention option was randomly selected. In instances where none of the assessed tailoring variables influenced health behavior or the EMA was not filled out, a text message was sent from the message set derived from the previously developed E-Supporter 1.0.

Decision Points

A decision point refers to a specific moment in time when the delivery of interventions to the user is determined [24]. In this JITAI, intervention decisions were triggered by each EMA prompt. EMAs were initiated through time-based sampling, occurring at 1 or 2 semirandom instances daily—once in the morning and again in the afternoon or evening.

Participants and Recruitment

We aimed to recruit 8 to 10 individuals diagnosed with T2D for this study, as this group size typically yields valuable insights into the acceptability of technologies [48]. To be eligible for participation, individuals needed to have a diagnosis of T2D, be aged ≥18 years, own a mobile phone, have the ability to understand Dutch, and possess the capacity to understand the informed consent procedure. Participants were recruited from the observational Diabetes and Lifestyle Cohort Twente study, a longitudinal study to obtain lifestyle, glucose, and disease-related parameters in patients with T2D [49], from Ziekenhuisgroep Twente (ZGT), a regional hospital in the east of the Netherlands. ZGT has a diabetes outpatient clinic where internists, nurse practitioners, and diabetes nurses provide daily care for a large population of patients with T2D. This hospital is also actively engaged in extensive scientific research related to diabetes. Individuals from the Diabetes and Lifestyle Cohort Twente cohort who had previously expressed willingness to participate in additional scientific research were contacted by telephone by the researcher (EAGH) to determine their interest in participating in this study. Detailed study information was provided orally, and interested individuals were sent the participant information sheet via email. After a week, potential participants were contacted again by telephone to inquire about their decision to participate. Those who chose to participate provided their informed consent by signing a digital or paper consent form before study enrollment. Although participants were recruited from ZGT hospital, participants used the JITAI in their home situation.

Procedures

The researcher introduced herself and the aims of the study before the start of the intervention. Participants were instructed to use the intervention for a period of 2 weeks, with the option to receive coaching for physical activity, healthy nutrition, or both lifestyle domains. Before starting the intervention, participants completed a web-based survey, providing questionnaires related to self-efficacy and phase of behavior change (refer to the Data Collection and Analysis section). The information on participants' self-efficacy and phase of behavior change was used to identify the most suitable self-efficacy messages and phase of behavior change messages for delivering intervention options (eg, people in the action phase received fewer messages related to behavior initiation and more messages regarding the action and maintenance phase) [35].

Throughout the intervention, participants were prompted to fill in the EMAs at 1 or 2 semirandom time intervals each day—one in the morning and another in the afternoon or evening [26]. At the prompt time, participants received a text message containing a link to access the EMA questionnaire. Intervention options were delivered via texting, with a maximum of 2 messages per day.

Following the 2-week intervention period, semistructured interviews were conducted via telephone and audio recorded

XSL•FO
RenderX

by EAGH, a female PhD candidate in health sciences, to gather in-depth insights and feedback from participants regarding their experiences with the intervention. The interviewer followed a preestablished interview guide, had prior experience in conducting interviews, and had no prior relationship with the study participants. Each interview lasted between 20 and 40 minutes.

Data Collection and Analysis

Acceptability

After the 2-week intervention period, participants engaged in a semistructured interview to elicit their thoughts, experiences, and suggestions for improvement. We examined the comprehensiveness and ease of use regarding the use of the JITAI [18]. Furthermore, the focus was primarily on satisfaction with the experience and the perceived usefulness of the intervention [40]. The interview schedule (Multimedia Appendix 2) covered three core parts: (1) acceptability of the EMA content and prompts, (2) acceptability of the intervention options (ie, the text messages), and (3) use of the JITAI (ie, the combinations of EMA and the messages).

The interview questions regarding the EMA questionnaire section were based on recommendations to assess the acceptability of EMAs [50], ensuring questions regarding item relevance, comprehensiveness, and instrument comprehensibility. In addition, participants' perceptions of EMA burden in terms of length, intensity, and motivation for completion were addressed. To evaluate the intervention options and the use of the JITAI, the interview schedule was based on the *Unified Theory of Acceptance and Use of Technology 2* framework [51]. Questions aligned with the *Unified Theory of Acceptance and Use of Technology 2* determinants, such as performance expectancy, effort expectancy, use behavior, and behavioral intention.

The interviews were analyzed based on a hybrid approach of thematic analysis, as outlined by Bingham [52] using ATLAS.ti 23 (Lumivero, LLC). This method combines both inductive and deductive strategies, enabling the development of new insights from the data while grounding the findings in an existing theoretical framework. The inductive phase began with reviewing the transcripts to familiarize with participants' responses, followed by an open inductive coding approach. Each text segment offering a notable perspective on the research questions received a descriptive open code. EAGH coded all transcripts, while AM independently coded 25% (2/8) and cross-checked the remainder. Open codes were iteratively refined, and discrepancies between EAGH and AM were resolved through discussions until consensus was achieved. The process of axial coding followed, involving clustering and combining open codes to identify patterns within and across interviews. Themes were developed by incorporating axial codes into theme statements that captured the observed data patterns. In the deductive phase, the *conceptual model of JITAI components* proposed by Nahum-Shani et al [24] was used to refine and classify the theme statements based on JITAI's main intervention components. Initially, themes were broadly assigned to the JITAI component where the definitions showed the greatest overlap. This classification was iteratively revised, with themes being reassigned until all themes were grouped under the most appropriate intervention component within the model. Throughout these phases, the research team continually discussed and refined the findings until consensus was reached.

Self-Efficacy Level

Self-efficacy was measured using the Exercise Self-Efficacy Scale (ESES) [53], consisting of 10 items that assess self-confidence in performing physical activity. In addition, a modified version of the ESES was used to determine self-efficacy related to a healthy diet [54]. The assessment used a 4-point Likert scale, ranging from *1 (ie, not at all true)* to *4 (ie, always true)*, yielding sum scores between 10 and 40, with higher scores indicating higher self-efficacy levels. Self-efficacy scores were treated as a continuous variable, obtained by summing the scores from the 10 items of the ESES questionnaire. Mean (SD) was used to present self-efficacy scores.

Phase of Behavior Change

To assess the phase of behavior change, single-question self-assessment stages of change scale [55] was used. This scale uses a single question to inquire about an individual's engagement in healthy behavior, aligning with the 5 phases of behavior change described in the transtheoretical model [56]. The phase of behavior change was categorized as an ordinal variable, ranging from 1 to 5, representing the following phases: (1) maintenance, (2) action, (3) preparation, (4) contemplation, and (5) precontemplation. These 5 answer options were aligned with the 3 behavior change phases targeted in the E-Supporter 1.0 intervention to allow for tailoring of intervention content. Participants who selected answer options 3, 4, or 5 were in the initiation phase, participants who chose answer option 2 were classified in the action phase, and participants who selected answer option 1 were categorized in the maintenance phase.

Demographics

Information on age (years), sex (ie, male, female, or other), BMI (kg/m^2), duration of T2D diagnosis (years), medication use (ie, none, oral, or insulin), and diabetes-related complications (ie, none, retinopathy, nephropathy, or neuropathy) was obtained from the hospital's electronic patient record. Having a job (ie, yes or no) and highest level of education (ie, no formal education, elementary education, lower secondary education, upper secondary education, higher vocational education, preuniversity education, university education, or postgraduate education) were questioned. Demographic characteristics were reported as mean (SD) for continuous variables or frequency (%) for categorical variables.

Ethical Considerations

The study protocol was reviewed and approved by the Medical Ethics Research Committee Twente, Enschede, the Netherlands (K20-05). In addition, the local advisory committee of practical feasibility in ZGT hospital approved to conduct this study (ZGT17-39). We obtained informed consent from all study participants, either digitally or on paper. Participants did not receive financial compensation for their participation due to the short duration of the study. Audio recordings of the interviews

XSL•FO
RenderX

were deleted after transcription. All participant-identifiable data were removed or pseudonymized. The data are stored on a secure, encrypted server at the University of Twente, accessible only to the researchers.

Results

Participant Characteristics

In total, 12 individuals were invited to participate in the study. Three people declined participation due to personal circumstances or perceiving the study burden as too high. One participant initially agreed but withdrew prematurely due to family circumstances. The characteristics of the 8 participants who completed the study are presented in Table 3. The participants had a mean age of 70.5 (SD 9.0) years, a mean BMI of 32.1 (SD 5.3) kg/m^2, and a mean T2D diagnosis of 15.6 (SD 7.7) years. Microvascular complications included neuropathy in 50% (4/8) of the participants and nephropathy in 25% (2/8) of the participants. In addition, of the 8 participants, 5 (63%) were using insulin and 6 (75%) remained active in work or volunteer activities. The most common educational level was higher vocational education (4/8, 50%).

A total of 6 participants opted for coaching on both physical activity and nutrition, while 1 participant preferred coaching exclusively on physical activity and another participant preferred coaching solely on nutrition. The results indicate that participants reported higher self-efficacy levels in the domain of nutrition compared to physical activity, with mean scores of 32 (SD 4) and 29 (SD 7), respectively (Table 4). Regarding the phase of behavior change, a large proportion of participants were in the maintenance phase for both physical activity (3/7, 43%) and nutrition (5/7, 71%).

Table 3. Participant characteristics (n=8).

Characteristic	Participants
Age (y), mean (SD)	70.5 (9)
Sex, n (%)	
Male	6 (75)
Female	2 (25)
BMI (kg/m^2), mean (SD)	32.1 (5.3)
Type 2 diabetes duration (y), mean (SD)	15.6 (7.7)
Complications, n (%)	
Retinopathy	0 (0)
Neuropathy	4 (50)
Nephropathy	2 (25)
Insulin treatment, n (%)	5 (63)
Employed, n (%)	6 (75)
Highest level of education, n (%)	
No formal education	1 (13)
Elementary education	0 (0)
Lower secondary vocational education	1 (13)
Secondary vocational education	1 (13)
Upper secondary vocational education	0 (0)
Higher vocational education	4 (50)
Preuniversity education	0 (0)
University education	0 (0)
Postgraduate education	0 (0)
Other	1 (13)

XSL•FO
RenderX

Table 4. Self-efficacy levels and phase of behavior change of participants for the physical activity and nutrition domains.

	Physical activity (n=7)	Nutrition (n=7)
Self-efficacy level, mean (SD)	29 (7)	32 (4)
Phase of behavior change, n (%)		
Initiation	3 (43)	2 (29)
Action	1 (14)	0 (0)
Maintenance	3 (43)	5 (71)

Acceptability

Overview

Figure 2 [24] provides an overview of various JITAI components from the conceptual model of JITAIs by Nahum-Shani et al [24], with the themes identified from the interviews.

Table 5 presents the results of the interviews. The bold terms represent the various components from the conceptual model of JITAIs by Nahum-Shani et al [24], including the identified themes and their associated definitions that were identified from the interviews.

Figure 2. Overview of the just-in-time adaptive intervention (JITAI) components and associated themes identified from the interviews with individuals with type 2 diabetes. EMA: ecological momentary assessment.

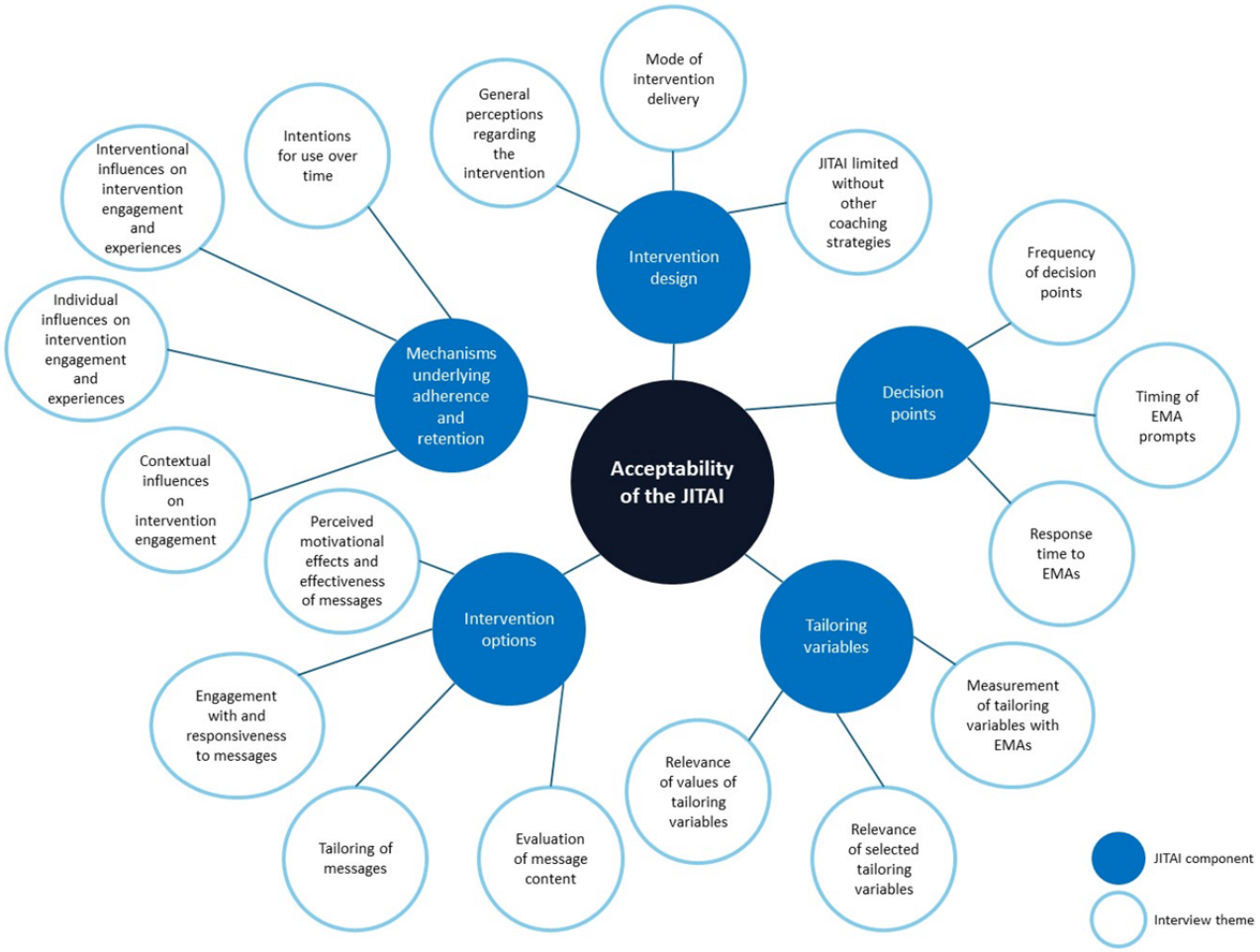

XSL•FO
RenderX

Table 5. Summary of the interview themes classified under JITAI[a] components related to the acceptability of the JITAI.

JITAI components and identified themes	Definitions
Intervention design	
General perceptions regarding the intervention	General perceptions about the intervention and associated components.
Mode of intervention delivery	Preferences regarding the method through which the intervention is delivered.
JITAI limited without other coaching strategies	Views on the need for additional support and the integration of other behavior change strategies.
Decision points	
Frequency of decision points	Perceptions and preferences regarding the frequency of EMA[b] prompts and intervention options.
Timing of EMA prompts	Perceptions and experiences regarding the appropriateness and predictability of the timing of EMA prompts.
Response time to EMAs	Behaviors and perceptions regarding the timeliness of their responses to EMA prompts.
Tailoring variables	
Measurement of tailoring variables with EMAs	Perspectives on the usefulness and constraints associated with using EMAs to measure tailoring variables.
Relevance of selected tailoring variables	Perceptions of the appropriateness and usefulness of the tailoring variables included in the EMAs.
Relevance of values of tailoring variables	Perceptions of the suitability and adequacy of the response options provided for the tailoring variables in the EMAs.
Intervention options	
Evaluation of message content	Perceptions and assessments of the content of the messages.
Tailoring of messages	Perceptions regarding the personalization and relevance of the messages they received.
Engagement with and responsiveness to messages	Interactions with and reactions to the messages.
Perceived motivational effects and effectiveness of messages	Perceptions of how the messages participants received influenced their motivation and behavior change efforts.
Mechanisms underlying adherence and retention	
Contextual influences on intervention engagement	External factors that impact participants' engagement with the intervention.
Individual influences on intervention engagement and experiences	Personal factors that affect participants' engagement and experiences with the intervention.
Interventional influences on intervention engagement and experiences	Various factors related to the intervention itself that influence participants' engagement and experiences.
Intentions for use over time	Intentions and willingness to continue using the JITAI system over a certain duration.

[a]JITAI: just-in-time adaptive intervention.

[b]EMA: ecological momentary assessment.

Intervention Design

All intervention components received *average* to *good* ratings, ranging from 7 to 9, with all components scoring approximately equally.

Overall, participants found this JITAI to be an easy-to-use and generally useful concept for supporting a healthy lifestyle. In addition to providing advice on a healthy lifestyle, participants felt that it also made them more aware of their lifestyle choices for the day.

In addition, most participants appreciated receiving technology-supported lifestyle guidance. In this study, SMS text messaging was used, which was perceived as an easy medium. Most participants preferred using an app, although 1 participant found using an app more complicated than SMS text messaging. Participant 8 had a clear preference for using an app; he stated the following:

> *I do everything via my phone. I find using a phone or tablet excellent. ... I would actually enjoy that [using an app], it is the easiest thing there is. I do everything on my mobile. It is convenient if personal goals and medical records are also available on it.* [Male, aged 59 years]

However, some participants preferred in-person support or guidance, such as support from a spouse as social support or professional guidance from a health care provider. The preference for in-person guidance mainly stemmed from the

XSL•FO
RenderX

ability to provide more context for certain events or feelings. Participant 3 illustrated this as follows:

> *For example, negative [mood] I will never indicate [in an EMA]. If I disagree with something, I will never put it in writing like that. If I have you on the phone personally, I will say what I feel, but on paper, that is not how I am. If I have to check on paper whether you are positive or negative, I will always indicate positive because if I indicate negative, I cannot provide any context, and then it is just how you want to read it and how you want to hear it. If I can provide an explanation to you, then I will more likely indicate that something is negative.* [Male, aged 77 years]

Some participants felt that relying solely on EMA-based messages to promote lifestyle was limited. They believed that coaching in this form alone lacked the pressure to actually change behavior. Examples to address this lack of pressure included setting goals and self-monitoring. Participant 1 had previously worked with goal setting and mentioned the following:

> *For example, the goal was: achieve 10,000 steps this week. And if I am missing 100, then I think: I will take a few more steps, I still want to reach those 10,000. It is an extra movement and pressure to do the exercise that I would not normally do, just to meet the set goal.* [Male, aged 78 years]

Other options mentioned were to make coaching more appealing by linking coaching to medical information about the participant or by providing healthy eating advice based on offers at the local supermarket.

Decision Points

For most participants, the number of prompts per day was acceptable, partly due to the low burden of completing the EMAs. However, some participants preferred a different prompt frequency, with options for both more and fewer EMA prompts mentioned. One participant suggested increasing the frequency from 3 to 4 EMA prompts per day, as it triggered engagement with a healthy lifestyle. Another participant preferred receiving the EMA prompts every other day, given their retired lifestyle with fewer varied activities.

The timing of the EMA prompts and messages was considered suboptimal by most participants. While they indicated that the timing was not disruptive, some prompts were deemed inconvenient, particularly when they occurred at times when participants could not respond, such as during work. Furthermore, some participants found the timing of the prompts predictable, as they often received them at approximately the same times each day, suggesting that more variation in timing would be preferred. Participant 5 elaborated on this, as follows:

> *For example, feeling hungry, that was often in the morning. But I had just eaten breakfast, so then I am not hungry and not craving anything. If you were to receive the questionnaire more in the evenings, that might be different, because then you're more likely watching TV and, well, then you might have a different answer.* [Female, aged 77 years]

The occasionally unsuitable timing of the prompts and the fact that not everyone consistently had their mobile phone with them affected the time frame used to complete the EMAs. Some participants deliberately made time to respond to the prompts upon receiving them, while others waited for a suitable moment. None of the participants intentionally delayed completing the EMAs.

Tailoring Variables

One participant misunderstood the purpose of the EMA and did not grasp why they had to fill out a questionnaire so frequently, but overall, participants were satisfied with using an EMA to gain insight into their situation. However, two main points of criticism were that (1) EMA responses provide only a snapshot and (2) there was insufficient opportunity to provide context or additional explanations for the response options in the EMA.

Regarding the first point, participants noted that while an EMA is a useful tool to assess the situation at a specific moment, it lacks insight into what happens for the rest of the day and therefore does not adequately account for rapid changes that may occur in one's situation, as explained by participant 3:

> *For example, I can be very negative in the morning, but after noon, I can be in a very good mood. I can bump my head and start swearing, and that can change at any moment. Therefore, it is a bit of a snapshot. That is why I often give standard answers, and I do not think you will gain much insight from that.* [Male, aged 77 years]

With regard to the second point, some participants felt there was insufficient opportunity to provide explanations for multiple-choice question answers. They would prefer an open field where they could freely elaborate on the multiple-choice questions.

In addition, participants felt a slight obligation to complete the EMAs because they felt researchers were waiting for their responses. Furthermore, some participants mentioned that they could not always answer completely freely due to feeling the need to provide socially desirable responses and being reluctant to give *negative* answers. However, all participants were comfortable sharing personal information, such as location or activity, via an EMA. The reasons for this included generally having no objection to sharing personal information, having given informed consent for this, and feeling a sense of autonomy by only providing information if they wished to do so.

Regarding the relevance of the tailoring variables, participants felt that the chosen variables were generally suitable for providing insights into their personal and environmental situations. Some participants mentioned that there could be room to include additional questions in the EMA without substantially increasing the burden. One participant suggested adding a specific question about experiencing hyperglycemia and hypoglycemia, so that this could also be taken into account in lifestyle advice. Furthermore, in some cases, certain questions were not relevant to a person, such as the mood question, which some participants mentioned always being in a good mood.

XSL•FO
RenderX

Most participants found the values of the tailoring variables, meaning the answer options in the EMA, sufficiently complete and clear. However, some improvement points were mentioned, such as making more distinctions in the answers for certain tailoring variables, for example, by including more types of weather conditions or also adding a neutral option in the *weather* tailoring variable. In addition, participant 5 indicated that the activity *eating* could not provide the researchers with enough information on the current activity. She made the following suggestion:

> *You could also fill in "eating," I have done that before. Those things, it is clear, but it could be a bit clearer... For example, what am I eating? If you are eating your regular meal or if you are eating a snack. It would be nice if you can provide more information about that.* [Female, aged 77 years]

Intervention Options

Most participants found the content of the messages enjoyable, funny, and varied, with clear tips for improving lifestyle. Although there was some variation in how enjoyable participants found the messages, the content was not perceived as intrusive. Most feedback was related to the content of the recipes referenced in the messages. For example, 1 participant mentioned that the recipes were quite complex, requiring many ingredients that are not commonly used, leading to items expiring. Furthermore, some minor suggestions were made to improve the content of the messages, such as adding reminders.

Participants were generally satisfied with the degree of tailoring in the messages and could see that their EMA responses were reflected in the messages. However, participants also felt that there was variation in how well the messages aligned with individual circumstances. Participants noticed tailoring the most when they were addressed by their first names, when messages matched their personal preferences, and when messages matched their location or activity at that time. Participant 6 illustrated this with the following example:

> *Yes, I noticed from the little tips you got, for example asking a colleague for a walk during the break or traveling by bike to work instead of by car when you are at work.* [Female, aged 55 years]

Some participants had a more neutral opinion on the tailoring of the messages, such as that the messages sometimes still contained too much general knowledge or did not sufficiently align with their specific situation. Participant 4 believed that his preferences were considered, but he found that the messages did not sufficiently address his physical limitations:

> *Here it says at one point that it is good to move. I completely agree with you, but for me, it is currently difficult to move.... We used to do a lot of cycling and walking. I have had trouble with my right leg for five years now.* [Male, aged 71 years]

Participants indicated that they always read the messages, and some of them were even curious about the message they would receive based on the EMA, as indicated by participant 6:

> *It was fun. I was always waiting for the response I got. When I was at work, I got a message like: ask your colleague or do something with your colleague. When I had been to the store, I got different answers again, that was fun.* [Female, aged 55 years]

It was not always possible to read the messages immediately, but participants did so as soon as they could. Participants also showed a preference for interactivity, as several participants responded to the messages via SMS text messaging and expressed enjoyment in giving a reaction to the messages.

The responses regarding the motivating effect and perceived effectiveness of the messages varied. Almost all participants found receiving messages motivating to work on their lifestyle. Some participants indicated that they did not always act on the advice, partly because the advice was not always applicable to them. Nevertheless, it helped some participants make healthier choices, as participant 8 indicated with the following examples:

> *It did trigger me. Sometimes in the evening I thought: let's get some fries. Then I thought: no, let's just peel two potatoes, with a piece of meat and vegetables. I did that twice.* [Male, aged 59 years]

> *For example, that church service. Are you going to church? How? In that questionnaire, you indicate that you are going to do an activity. Then you get a message with: how could you improve this activity? Do you think about this? Are you going by bike, car, or walking? I started cycling, otherwise I would have taken the car.* [Male, aged 59 years]

Mechanisms Underlying Adherence and Retention

A contextual factor influencing engagement with the intervention was the COVID-19 pandemic and the lockdown period in the Netherlands during which the study was conducted. For instance, 1 participant mentioned that daily activities had substantially reduced due to the COVID-19 pandemic, referring to limited opening hours of stores and restaurants during the lockdown. As a result, the participant reported providing nearly identical answers every day, which lowered his engagement with the intervention. Another participant mentioned that they had become less active during the COVID-19 pandemic due to disruptions in their regular routine.

In addition, participants indicated various individual factors that influenced the use of the intervention. Some participants reported making many lifestyle adjustments or losing weight before participating in the study, making the intervention information less useful for them. Other participants expressed that they felt no need to improve their lifestyle because they were satisfied with their current health and lifestyle. Finally, some participants felt a slight resistance to actually adjusting their lifestyle, which seemed to be independent from the indicated phase of behavior change (Table 4). This feeling partly stemmed from their unwillingness to make certain lifestyle changes (eg, a participant who did not want to take the stairs) and from a desire to also be allowed to make unhealthy choices. Participant 2 already made many improvements to his lifestyle in recent years but indicated that he sometimes wanted to make unhealthy choices:

XSL•FO
RenderX

> *I am very careful with my eating. In the past 2.5 years, I have lost 18-20 kilos. I try to maintain my current weight now. That is certainly better for my diabetes too. But I do still snack occasionally.* [Male, aged 78 years]

The intervention itself also influenced engagement and experiences with the intervention. An important factor was the intervention burden. All participants found the burden of completing the EMA minimal. Participants found the EMA instructions and questions clear and the questionnaire brief, and it took little time to complete the EMA. They also found the link to the EMA questionnaire in the text message easy to use. As a result, participants remained motivated. However, the lack of variation in the EMA questionnaire reduced engagement over time for some participants. Because participants received the exact same questions every day, they found it monotonous. Several participants mentioned that they started filling out the questions on *autopilot*, as participant 5 described:

> *Overall, there was enough motivation, but because it was always the same questionnaire, you start thinking about it less. Only for activities like drinking coffee, I sometimes gave a different answer, but otherwise, it was always the same answer. You do not have to think about it anymore.* [Female, aged 77 years]

Participants suggested rearranging the questions or asking different questions on different days to make the EMA more engaging and encourage careful consideration when completing it.

In addition, participants provided feedback on the continued use of the intervention, with most participants willing to use the intervention for >2 weeks. Some participants were willing to fill out EMAs and receive messages daily without indicating a maximum period of use, but others saw more value in a maximum period of several weeks to months.

Discussion

Principal Findings

With this study, we aimed to gain insight into the acceptability of our initial JITAI version among individuals with T2D. This preliminary phase is pivotal for gaining insights into the target group's experiences with the intervention, allowing us to assess and enhance its potential value based on their feedback [18]. Overall, participants expressed satisfaction with the intervention, as supported by their average to high ratings across its components. This opinion was evident from the positive feedback on the motivating and varied text messages, as well as some examples of personalized support that participants experienced.

It became evident that experiences varied substantially and were often influenced by individual and contextual factors that were not sufficiently captured with the EMAs. This variability was observed across all aspects of the intervention design. On the basis of the participants' feedback, we have identified three key areas for improvement: (1) optimizing tailoring, (2) refining assessment methods, and (3) emphasizing integration with other behavioral intervention strategies.

Regarding the optimization of tailoring, our study highlighted the need for further refinement in this area. Although tailoring variables were selected based on target group input, their relevance varied among participants. For instance, mood- or disease-related symptoms were relevant to some but not all individuals. The decision to offer 2 prompts per day was informed by previous E-Supporter studies [35,36], yet participant feedback on prompt frequency varied. Similarly, we noted nonoptimal timing of prompts, aligning with challenges observed in related studies [57]. This heterogeneity among individuals' experiences requires a shift toward individually tailored intervention designs. Realizing such individual tailoring can be approached in several ways. A straightforward method is to allow individuals to specify their preferences for tailoring variables, intervention options, as well as the frequency and timing of prompts before or during the intervention period [58,59]. Applying machine learning, such as reinforcement learning (RL), is a more advanced method for improving tailoring. RL learns and adapts from interaction and feedback. RL allows to tailor support strategies to individual needs and preferences regarding timing, frequency, and intervention types [60-63].

With respect to the assessment methods, participants mentioned the potential of EMAs to personalize intervention support, but they also identified limitations. These limitations included the repetitive nature of the EMA questionnaire and its snapshot insight into the current situation. Some participants reported reduced engagement over time, possibly attributed to the lack of variation in questions or the impact of the COVID-19 pandemic and associated restrictions during the study period [64]. Lockdown measures reduced social interactions and outdoor activities, leading to participants spending more time at home and engaging less in social or work-related activities. Consequently, responses to EMAs became predictable, and the participant's interest diminished over time. EMAs rely on user input to deliver personalized support [27]. Decreased user engagement may lead to either skipping questions or responding without due consideration. This may influence the reliability and validity of measurements. As Nahum-Shani et al [24] emphasized, balancing response burden with the need for frequent assessment for capturing dynamic changes in certain variables is critical. Exploring variability in measured constructs across EMAs and considering alternative measurement methods for tailoring variables, such as GPS tracking for location and weather application programming interface, could simplify the data collection [57,61,65,66]. Nevertheless, issues such as privacy concerns with GPS and individuals' perceptions of weather suitability for outdoor activities require consideration. While optimal measurement strategies for tailoring variables remain uncertain, we aim to find the right balance between passive and active methods for accommodating their variability in future interventions.

To realize sustainable behavior change, our study revealed that many participants found the intervention motivating, but some expressed no need, unwillingness, or resistance to follow the advice. The participants had T2D for an average of 15.6 (SD

XSL•FO
RenderX

7.7) years, which suggests they were already familiar with making lifestyle changes. This is supported by their relatively high self-efficacy scores, as individuals with high self-efficacy tend to feel confident in their ability to make behavioral changes. Consequently, their perceived need for external support from the intervention may have been reduced. Furthermore, many participants were already in the maintenance phase of behavior change, suggesting that they had already implemented some lifestyle changes. These individuals might require specific support for relapse prevention or, in some cases, no additional support at all [43]. A smaller group was in the initiation phase, indicating that they were just beginning to consider lifestyle changes. This group might require a stronger focus on motivational strategies to build commitment [1,2]. These differences highlight the need for interventions that can adapt to the wide variety of individual needs. However, JITAIs are generally designed to offer brief, just-in-time suggestions for immediate behavior adjustments [26,28,57]. While the JITAI may be effective for action-related changes, this design may fall short of addressing the more comprehensive strategies required. Therefore, we emphasize the importance of situating JITAIs within broader behavior change interventions. Integrating a JITAI within a comprehensive lifestyle intervention that includes extensive behavioral change strategies, as envisioned in our integration with the E-Supporter content, could provide the necessary support to facilitate sustainable behavior change.

Strengths and Limitations

A strength of this study was the application of a hybrid approach to thematic analysis for analyzing the interviews. An advantage of this approach is that we were able to combine the strengths of both inductive and deductive approaches. The hybrid approach allowed us to use the flexibility of inductive coding to create data-driven insights and also use the deductive method's ability to build upon existing knowledge, ensuring that the findings were both data driven and theoretically grounded.

A limitation of our study was the inability to measure the impact on proximal outcomes. Ideally, we would have supported the intervention with wearable activity monitoring and an app for dietary monitoring. However, due to COVID-19 restrictions preventing physical contact with participants, we were unable to distribute wearables or assist with app installations. In the future, our aim is to incorporate goal setting and self-monitoring into the intervention, as done in the E-Supporter. This approach can provide additional support in the behavior change process and enable examination of the intervention effects on proximal outcomes. As it was not the aim of this study to measure these outcomes, it does not affect the answer to the research question. Another limitation is the limited evaluation of intervention options for certain tailoring variables, such as negative mood or poor weather. As reported by participants, negative mood was seldom indicated in the EMA, resulting in no corresponding intervention options being sent. Furthermore, the evaluation was conducted during sunny weather, leading to minimal use of intervention options for poor weather conditions. Consequently, we lacked insight into the acceptability of these intervention options. However, the content and BCTs used in these intervention options are similar to others, so we anticipate similar perceptions of these messages compared to others. Finally, our sample size was relatively small, which may affect the robustness of the results. While this study provided valuable insights into the acceptability of the JITAI, our study population was relatively homogeneous, consisting predominantly of older individuals with a longer duration of diabetes. These characteristics appear to be a relatively good representation of the target group, characterized by an older average age, a slightly higher proportion of men than women, and including individuals with lower education levels, which are all sociodemographic factors associated with T2D [67,68]. However, the homogeneity limits the strength of our conclusions. Future research should focus on recruiting a more diverse population, including younger individuals and those with a shorter diabetes duration, to improve the generalizability of our findings.

Conclusions

This study presents the first version of an EMA-driven JITAI designed to support a healthy lifestyle in individuals with T2D. The intervention received positive feedback, particularly for its motivating messages that felt tailored. However, individual differences in experiences highlight the need for improvement in tailoring, assessment methods, and the integration of JITAI with other behavioral intervention strategies.

Acknowledgments

The authors would like to express their gratitude to all participants who were willing to participate in this study, especially because the COVID-19 pandemic made it necessary to conduct all communications remotely. The authors used the assistance of the generative artificial intelligence tool ChatGPT by OpenAI [69] to check English grammar and refine language fluency but not for data collection, data analysis, or writing the manuscript.

Data Availability

The data generated during this study are not publicly available because of the personal information shared during the interviews but are available from the corresponding author on reasonable request.

Authors' Contributions

EAGH designed the intervention, performed the data collection, and drafted the initial version of the manuscript. AM contributed to the data analysis and ongoing refinement of the manuscript. WH, LKS, GDL, and MMRV-H provided ongoing feedback related

XSL•FO
RenderX

to the intervention design, methods, results, and discussion of this study. All the authors critically evaluated and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
E-Supporter determinants and behavior change techniques.
[PDF File (Adobe PDF File), 178 KB - formative_v9i1e65026_app1.pdf]

Multimedia Appendix 2
Translated interview schedule.
[PDF File (Adobe PDF File), 206 KB - formative_v9i1e65026_app2.pdf]

References

1. Golovaty I, Ritchie ND, Tuomilehto J, Mohan V, Ali MK, Gregg EW, et al. Two decades of diabetes prevention efforts: a call to innovate and revitalize our approach to lifestyle change. Diabetes Res Clin Pract 2023 Apr;198:110195 [FREE Full text] [doi: 10.1016/j.diabres.2022.110195] [Medline: 36470316]
2. van Ommen B, Wopereis S, van Empelen P, van Keulen HM, Otten W, Kasteleyn M, et al. From diabetes care to diabetes cure-the integration of systems biology, eHealth, and behavioral change. Front Endocrinol (Lausanne) 2017;8:381 [FREE Full text] [doi: 10.3389/fendo.2017.00381] [Medline: 29403436]
3. Lean MEJ, Leslie WS, Barnes AC, Brosnahan N, Thom G, McCombie L, et al. Durability of a primary care-led weight-management intervention for remission of type 2 diabetes: 2-year results of the DiRECT open-label, cluster-randomised trial. Lancet Diabetes Endocrinol 2019 May;7(5):344-355 [FREE Full text] [doi: 10.1016/S2213-8587(19)30068-3] [Medline: 30852132]
4. Ried-Larsen M, Johansen MY, MacDonald CS, Hansen KB, Christensen R, Wedell-Neergaard AS, et al. Type 2 diabetes remission 1 year after an intensive lifestyle intervention: a secondary analysis of a randomized clinical trial. Diabetes Obes Metab 2019 Oct;21(10):2257-2266 [FREE Full text] [doi: 10.1111/dom.13802] [Medline: 31168922]
5. Pot GK, Battjes-Fries MC, Patijn ON, van der Zijl N, Pijl H, Voshol P. Lifestyle medicine for type 2 diabetes: practice-based evidence for long-term efficacy of a multicomponent lifestyle intervention (Reverse Diabetes2 Now). BMJ Nutr Prev Health 2020 Dec 18;3(2):188-195 [FREE Full text] [doi: 10.1136/bmjnph-2020-000081] [Medline: 33521528]
6. Chen L, Pei JH, Kuang J, Chen HM, Chen Z, Li ZW, et al. Effect of lifestyle intervention in patients with type 2 diabetes: a meta-analysis. Metabolism 2015 Feb;64(2):338-347 [FREE Full text] [doi: 10.1016/j.metabol.2014.10.018] [Medline: 25467842]
7. García-Molina L, Lewis-Mikhael AM, Riquelme-Gallego B, Cano-Ibáñez N, Oliveras-López MJ, Bueno-Cavanillas A. Improving type 2 diabetes mellitus glycaemic control through lifestyle modification implementing diet intervention: a systematic review and meta-analysis. Eur J Nutr 2020 Jun;59(4):1313-1328. [doi: 10.1007/s00394-019-02147-6] [Medline: 31781857]
8. Gostoli S, Raimondi G, Popa AP, Giovannini M, Benasi G, Rafanelli C. Behavioral lifestyle interventions for weight loss in overweight or obese patients with type 2 diabetes: a systematic review of the literature. Curr Obes Rep 2024 Jun;13(2):224-241 [FREE Full text] [doi: 10.1007/s13679-024-00552-5] [Medline: 38436787]
9. Espeland MA, Glick HA, Bertoni A, Brancati FL, Bray GA, Clark JM, et al. Impact of an intensive lifestyle intervention on use and cost of medical services among overweight and obese adults with type 2 diabetes: the action for health in diabetes. Diabetes Care 2014 Sep;37(9):2548-2556 [FREE Full text] [doi: 10.2337/dc14-0093] [Medline: 25147253]
10. Mirahmadizadeh A, Khorshidsavar H, Seif M, Sharifi MH. Adherence to medication, diet and physical activity and the associated factors amongst patients with type 2 diabetes. Diabetes Ther 2020 Feb 08;11(2):479-494 [FREE Full text] [doi: 10.1007/s13300-019-00750-8] [Medline: 31916213]
11. Hagedoorn IJM, den Braber N, Oosterwijk MM, Gant CM, Navis G, Vollenbroek-Hutten MMR, et al. Low physical activity in patients with complicated type 2 diabetes mellitus is associated with low muscle mass and low protein intake. J Clin Med 2020 Sep 25;9(10):3104 [FREE Full text] [doi: 10.3390/jcm9103104] [Medline: 32992990]
12. Jalving AC, Gant CM, Binnenmars SH, Soedamah-Muthu SS, Bakker SJL, Navis G, et al. Glycaemic control in the diabetes and Lifestyle Cohort Twente: a cross-sectional assessment of lifestyle and pharmacological management on Hba1c target achievement. Diabetes Obes Metab 2018 Oct;20(10):2494-2499 [FREE Full text] [doi: 10.1111/dom.13399] [Medline: 29862616]
13. Jarvie JL, Pandey A, Ayers CR, McGavock JM, Sénéchal M, Berry JD, et al. Aerobic fitness and adherence to guideline-recommended minimum physical activity among ambulatory patients with type 2 diabetes mellitus. Diabetes Care 2019 Jul;42(7):1333-1339 [FREE Full text] [doi: 10.2337/dc18-2634] [Medline: 31221698]

XSL•FO
RenderX

14. Lim SS, Vos T, Flaxman AD, Danaei G, Shibuya K, Adair-Rohani H, et al. A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet 2012 Dec 15;380(9859):2224-2260 [FREE Full text] [doi: 10.1016/S0140-6736(12)61766-8] [Medline: 23245609]
15. Schmidt SK, Hemmestad L, MacDonald CS, Langberg H, Valentiner LS. Motivation and barriers to maintaining lifestyle changes in patients with type 2 diabetes after an intensive lifestyle intervention (the U-TURN trial): a longitudinal qualitative study. Int J Environ Res Public Health 2020 Oct 13;17(20):7454 [FREE Full text] [doi: 10.3390/ijerph17207454] [Medline: 33066239]
16. Ahola AJ, Groop PH. Barriers to self-management of diabetes. Diabet Med 2013 Apr;30(4):413-420. [doi: 10.1111/dme.12105] [Medline: 23278342]
17. Ahmad F, Joshi SH. Self-care practices and their role in the control of diabetes: a narrative review. Cureus 2023 Jul;15(7):e41409 [FREE Full text] [doi: 10.7759/cureus.41409] [Medline: 37546053]
18. Kip H, van Gemert-Pijnen L, Jong NB, Sanderman R, Kelders SM. eHealth Research Theory and Development: A Multidisciplinary Approach. Milton Park, UK: Taylor & Francis; 2018.
19. Milne-Ives M, Lam C, de Cock C, van Velthoven MH, Meinert E. Mobile apps for health behavior change in physical activity, diet, drug and alcohol use, and mental health: systematic review. JMIR Mhealth Uhealth 2020 Mar 18;8(3):e17046 [FREE Full text] [doi: 10.2196/17046] [Medline: 32186518]
20. Romeo A, Edney S, Plotnikoff R, Curtis R, Ryan J, Sanders I, et al. Can smartphone apps increase physical activity? Systematic review and meta-analysis. J Med Internet Res 2019 Mar 19;21(3):e12053 [FREE Full text] [doi: 10.2196/12053] [Medline: 30888321]
21. Schoeppe S, Alley S, van Lippevelde W, Bray NA, Williams SL, Duncan MJ, et al. Efficacy of interventions that use apps to improve diet, physical activity and sedentary behaviour: a systematic review. Int J Behav Nutr Phys Act 2016 Dec 07;13(1):127 [FREE Full text] [doi: 10.1186/s12966-016-0454-y] [Medline: 27927218]
22. Villinger K, Wahl DR, Boeing H, Schupp HT, Renner B. The effectiveness of app-based mobile interventions on nutrition behaviours and nutrition-related health outcomes: a systematic review and meta-analysis. Obes Rev 2019 Oct;20(10):1465-1484 [FREE Full text] [doi: 10.1111/obr.12903] [Medline: 31353783]
23. Singh B, Ahmed M, Staiano AE, Gough C, Petersen J, Vandelanotte C, et al. A systematic umbrella review and meta-meta-analysis of eHealth and mHealth interventions for improving lifestyle behaviours. NPJ Digit Med 2024 Jul 05;7(1):179 [FREE Full text] [doi: 10.1038/s41746-024-01172-y] [Medline: 38969775]
24. Nahum-Shani I, Smith SN, Spring BJ, Collins LM, Witkiewitz K, Tewari A, et al. Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. Ann Behav Med 2018 May 18;52(6):446-462 [FREE Full text] [doi: 10.1007/s12160-016-9830-8] [Medline: 27663578]
25. Spruijt-Metz D, Nilsen W. Dynamic models of behavior for just-in-time adaptive interventions. IEEE Pervasive Comput 2014 Jul 09;13(3):13-17. [doi: 10.1109/MPRV.2014.46]
26. Goldstein SP, Evans BC, Flack D, Juarascio A, Manasse S, Zhang F, et al. Return of the JITAI: applying a just-in-time adaptive intervention framework to the development of m-health solutions for addictive behaviors. Int J Behav Med 2017 Oct;24(5):673-682 [FREE Full text] [doi: 10.1007/s12529-016-9627-y] [Medline: 28083725]
27. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol 2008;4:1-32. [doi: 10.1146/annurev.clinpsy.3.022806.091415] [Medline: 18509902]
28. Hardeman W, Houghton J, Lane K, Jones A, Naughton F. A systematic review of just-in-time adaptive interventions (JITAIs) to promote physical activity. Int J Behav Nutr Phys Act 2019 Apr 03;16(1):31 [FREE Full text] [doi: 10.1186/s12966-019-0792-7] [Medline: 30943983]
29. Perski O, Hébert ET, Naughton F, Hekler EB, Brown J, Businelle MS. Technology-mediated just-in-time adaptive interventions (JITAIs) to reduce harmful substance use: a systematic review. Addiction 2022 May;117(5):1220-1241 [FREE Full text] [doi: 10.1111/add.15687] [Medline: 34514668]
30. Wang L, Miller LC. Just-in-the-moment adaptive interventions (JITAI): a meta-analytical review. Health Commun 2020 Nov;35(12):1531-1544. [doi: 10.1080/10410236.2019.1652388] [Medline: 31488002]
31. Oikonomidi T, Ravaud P, LeBeau J, Tran VT. A systematic scoping review of just-in-time, adaptive interventions finds limited automation and incomplete reporting. J Clin Epidemiol 2023 Feb;154:108-116. [doi: 10.1016/j.jclinepi.2022.12.006] [Medline: 36521653]
32. Forman EM, Goldstein SP, Zhang F, Evans BC, Manasse SM, Butryn ML, et al. OnTrack: development and feasibility of a smartphone app designed to predict and prevent dietary lapses. Transl Behav Med 2019 Mar 01;9(2):236-245 [FREE Full text] [doi: 10.1093/tbm/iby016] [Medline: 29617911]
33. Rabbi M, Pfammatter A, Zhang M, Spring B, Choudhury T. Automated personalized feedback for physical activity and dietary behavior change with mobile phones: a randomized controlled trial on adults. JMIR Mhealth Uhealth 2015 May 14;3(2):e42 [FREE Full text] [doi: 10.2196/mhealth.4160] [Medline: 25977197]
34. Ding X, Xu J, Wang H, Chen G, Thind H, Zhang Y. WalkMore: promoting walking with just-in-time context-aware prompts. In: Proceedings of the IEEE Wireless Health. 2016 Presented at: WH 2016; October 25-27, 2016; Bethesda, MD. [doi: 10.1109/wh.2016.7764558]

XSL•FO
RenderX

35. Hietbrink EAG, Middelweerd A, van Empelen P, Preuhs K, Konijnendijk AAJ, Oude Nijeweme-d'Hollosy W, et al. A digital lifestyle coach (E-Supporter 1.0) to support people with type 2 diabetes: participatory development study. JMIR Hum Factors 2023 Jan 12;10:e40017 [FREE Full text] [doi: 10.2196/40017] [Medline: 36633898]
36. Hietbrink EAG, Oude Nijeweme-d'Hollosy W, Middelweerd A, Konijnendijk AAJ, Schrijver LK, Ten Voorde AS, et al. A digital coach (e-supporter 1.0) to support physical activity and a healthy diet in people with type 2 diabetes: acceptability and limited efficacy testing. JMIR Form Res 2023 Jul 28;7:e45294 [FREE Full text] [doi: 10.2196/45294] [Medline: 37505804]
37. Op den Akker H, Cabrita M, Op den Akker R, Jones VM, Hermens HJ. Tailored motivational message generation: a model and practical framework for real-time physical activity coaching. J Biomed Inform 2015 Jun;55:104-115 [FREE Full text] [doi: 10.1016/j.jbi.2015.03.005] [Medline: 25843359]
38. Lustria ML, Noar SM, Cortese J, van Stee SK, Glueckauf RL, Lee J. A meta-analysis of web-delivered tailored health behavior change interventions. J Health Commun 2013;18(9):1039-1069. [doi: 10.1080/10810730.2013.768727] [Medline: 23750972]
39. Chen J, Cade JE, Allman-Farinelli M. The most popular smartphone apps for weight loss: a quality assessment. JMIR Mhealth Uhealth 2015 Dec 16;3(4):e104 [FREE Full text] [doi: 10.2196/mhealth.4334] [Medline: 26678569]
40. Bowen DJ, Kreuter M, Spring B, Cofta-Woerpel L, Linnan L, Weiner D, et al. How we design feasibility studies. Am J Prev Med 2009 May;36(5):452-457 [FREE Full text] [doi: 10.1016/j.amepre.2009.02.002] [Medline: 19362699]
41. Schwarzer R. Health action process approach (HAPA) as a theoretical framework to understand behavior change. Actualidades Psicología 2016 Dec 05;30(121):119-130. [doi: 10.15517/ap.v30i121.23458]
42. Rothman AJ. Toward a theory-based analysis of behavioral maintenance. Health Psychol 2000 Jan;19(1S):64-69. [doi: 10.1037/0278-6133.19.suppl1.64] [Medline: 10709949]
43. Marlatt GA, Donovan DM. Relapse Prevention: Maintenance Strategies in the Treatment of Addictive Behaviors, Second Edition. New York, NY: The Guilford Press; 2005.
44. Kromhout D, Spaaij CJ, de Goede J, Weggemans RM. The 2015 Dutch food-based dietary guidelines. Eur J Clin Nutr 2016 Aug;70(8):869-878 [FREE Full text] [doi: 10.1038/ejcn.2016.52] [Medline: 27049034]
45. Michie S, Richardson M, Johnston M, Abraham C, Francis J, Hardeman W, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013 Aug;46(1):81-95 [FREE Full text] [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
46. Weggemans RM, Backx FJG, Borghouts L, Chinapaw M, Hopman MTE, Koster A, et al. The 2017 Dutch physical activity guidelines. Int J Behav Nutr Phys Act 2018 Jun 25;15(1):58 [FREE Full text] [doi: 10.1186/s12966-018-0661-9] [Medline: 29940977]
47. Johnston M, Carey RN, Connell Bohlen LE, Johnston DW, Rothman AJ, de Bruin M, et al. Development of an online tool for linking behavior change techniques and mechanisms of action based on triangulation of findings from literature synthesis and expert consensus. Transl Behav Med 2021 May 25;11(5):1049-1065 [FREE Full text] [doi: 10.1093/tbm/ibaa050] [Medline: 32749460]
48. Jaspers MWM. A comparison of usability methods for testing interactive health technologies: methodological aspects and empirical evidence. Int J Med Inform 2009 May;78(5):340-353. [doi: 10.1016/j.ijmedinf.2008.10.002] [Medline: 19046928]
49. Gant CM, Mensink I, Binnenmars SH, van der Palen JAM, Bakker SJL, Navis G, et al. Body weight course in the DIAbetes and LifEstyle Cohort Twente (DIALECT-1)-a 20-year observational study. PLoS One 2019 Jun 19;14(6):e0218400 [FREE Full text] [doi: 10.1371/journal.pone.0218400] [Medline: 31216324]
50. Degroote L, DeSmet A, de Bourdeaudhuij I, van Dyck D, Crombez G. Content validity and methodological considerations in ecological momentary assessment studies on physical activity and sedentary behaviour: a systematic review. Int J Behav Nutr Phys Act 2020 Mar 10;17(1):35 [FREE Full text] [doi: 10.1186/s12966-020-00932-9] [Medline: 32151251]
51. Venkatesh V, Thong JYL, Xu X. Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology. MIS Q 2012 Mar;36(1):157-178. [doi: 10.2307/41410412]
52. Bingham AJ. From data management to actionable findings: a five-phase process of qualitative data analysis. Int J Qual Methods 2023 Aug;22:1-11. [doi: 10.1177/16094069231183620]
53. Kroll T, Kehn M, Ho PS, Groah S. The SCI Exercise Self-Efficacy Scale (ESES): development and psychometric properties. Int J Behav Nutr Phys Act 2007 Aug 30;4:34 [FREE Full text] [doi: 10.1186/1479-5868-4-34] [Medline: 17760999]
54. Fokkema EMS. The development and evaluation of an eCoaching module to improve healthy nutrition in type 2 diabetes patients. University of Twente. 2019. URL: https://essay.utwente.nl/79927/ [accessed 2023-07-28]
55. Marttila J, Nupponen R. Assessing stage of change for physical activity: how congruent are parallel methods? Health Educ Res 2003 Aug;18(4):419-428. [doi: 10.1093/her/cyf034] [Medline: 12939124]
56. Prochaska JO, Diclemente CC, Norcross JC. In search of how people change: applications to addictive behaviors. In: Marlatt GA, VandenBos GR, editors. Addictive Behaviors: Readings on Etiology, Prevention, and Treatment. Washington, DC: American Psychological Association; 1997:671-696.

XSL•FO
RenderX

57. Mair JL, Hayes LD, Campbell AK, Buchan DS, Easton C, Sculthorpe N. A personalized smartphone-delivered just-in-time adaptive intervention (JitaBug) to increase physical activity in older adults: mixed methods feasibility study. JMIR Form Res 2022 Apr 07;6(4):e34662 [FREE Full text] [doi: 10.2196/34662] [Medline: 35389348]
58. Kandola A, Edwards K, Straatman J, Dührkoop B, Hein B, Hayes J. Digital self-management platform for adult asthma: randomized attention-placebo controlled trial. J Med Internet Res 2024 Apr 29;26:e50855 [FREE Full text] [doi: 10.2196/50855] [Medline: 38684084]
59. Chen YP, Woodward J, Bista D, Zhang X, Singh I, Obajemu O, et al. Investigating contextual notifications to drive self-monitoring in mHealth apps for weight maintenance. In: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. 2024 Presented at: CHI '24; May 11-16, 2024; Honolulu, HI. [doi: 10.1145/3613904.3641993]
60. Gönül S, Namlı T, Coşar A, Toroslu İ. A reinforcement learning based algorithm for personalization of digital, just-in-time, adaptive interventions. Artif Intell Med 2021 May;115:102062. [doi: 10.1016/j.artmed.2021.102062] [Medline: 34001322]
61. Vandelanotte C, Trost S, Hodgetts D, Imam T, Rashid M, To QG, et al. Increasing physical activity using an just-in-time adaptive digital assistant supported by machine learning: a novel approach for hyper-personalised mHealth interventions. J Biomed Inform 2023 Aug;144:104435 [FREE Full text] [doi: 10.1016/j.jbi.2023.104435] [Medline: 37394024]
62. Liao P, Greenewald K, Klasnja P, Murphy S. Personalized HeartSteps: a reinforcement learning algorithm for optimizing physical activity. Proc ACM Interact Mob Wearable Ubiquitous Technol 2020 Mar;4(1):18 [FREE Full text] [doi: 10.1145/3381007] [Medline: 34527853]
63. Grua EM, de Sanctis M, Malavolta I, Hoogendoorn M, Lago P. An evaluation of the effectiveness of personalization and self-adaptation for e-Health apps. Inf Softw Technol 2022 Jun;146:106841. [doi: 10.1016/j.infsof.2022.106841]
64. Ontwikkelingen coronavirus in 2020. Rijksoverheid. 2020. URL: https://www.rijksoverheid.nl/onderwerpen/coronavirus-tijdlijn/2020 [accessed 2024-05-17]
65. Thomas Craig KJ, Morgan LC, Chen CH, Michie S, Fusco N, Snowdon JL, et al. Systematic review of context-aware digital behavior change interventions to improve health. Transl Behav Med 2021 May 25;11(5):1037-1048 [FREE Full text] [doi: 10.1093/tbm/ibaa099] [Medline: 33085767]
66. Rajanna V, Lara-Garduno R, Behera DJ, Madanagopal K, Goldberg D, Hammond T. Step up life: a context aware health assistant. In: Proceedings of the Third ACM SIGSPATIAL International Workshop on the Use of GIS in Public Health. 2014 Presented at: HealthGIS '14; November 4, 2014; Dallas, TX. [doi: 10.1145/2676629.2676636]
67. Kontochristopoulou AM, Karatzi K, Karaglani E, Cardon G, Kivelä J, Wikström K, et al. Sociodemographic, anthropometric, and lifestyle correlates of prediabetes and type 2 diabetes in Europe: the Feel4Diabetes study. Nutr Metab Cardiovasc Dis 2022 Aug;32(8):1851-1862. [doi: 10.1016/j.numecd.2022.03.021] [Medline: 35610084]
68. Kyrou I, Tsigos C, Mavrogianni C, Cardon G, van Stappen V, Latomme J, et al. Sociodemographic and lifestyle-related risk factors for identifying vulnerable groups for type 2 diabetes: a narrative review with emphasis on data from Europe. BMC Endocr Disord 2020 Mar 12;20(Suppl 1):134 [FREE Full text] [doi: 10.1186/s12902-019-0463-3] [Medline: 32164656]
69. What can I help with? ChatGPT. URL: https://chatgpt.com/?ref=dotcom [accessed 2024-04-29]

Abbreviations

BCT: behavior change technique
EMA: ecological momentary assessment
ESES: Exercise Self-Efficacy Scale
JITAI: just-in-time adaptive intervention
RL: reinforcement learning
T2D: type 2 diabetes
ZGT: Ziekenhuisgroep Twente

Edited by A Mavragani; submitted 02.08.24; peer-reviewed by A Hornsten; comments to author 18.11.24; revised version received 06.12.24; accepted 18.12.24; published 19.02.25.

Please cite as:
Hietbrink EAG, Middelweerd A, d'Hollosy W, Schrijver LK, Laverman GD, Vollenbroek-Hutten MMR
Exploring the Acceptance of Just-in-Time Adaptive Lifestyle Support for People With Type 2 Diabetes: Qualitative Acceptability Study
JMIR Form Res 2025;9:e65026
URL: https://formative.jmir.org/2025/1/e65026
doi:10.2196/65026
PMID:39969969

XSL•FO
RenderX

©Eclaire A G Hietbrink, Anouk Middelweerd, Wendy d'Hollosy, Laura K Schrijver, Gozewijn D Laverman, Miriam M R Vollenbroek-Hutten. Originally published in JMIR Formative Research (https://formative.jmir.org), 19.02.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Nomophobia, Psychopathology, and Smartphone-Inferred Behaviors in Youth With Depression: Longitudinal Study

Tianyi Zhang[1], MIT; Andres Camargo[2], MCR; Lianne Schmaal[2], PhD; Vassilis Kostakos[1], PhD; Simon D'Alfonso[1], PhD

[1]School of Computing and Information Systems, University of Melbourne, Melbourne, Australia

[2]Centre for Youth Mental Health, University of Melbourne, Melbourne, Australia

Corresponding Author:
Tianyi Zhang, MIT
School of Computing and Information Systems
University of Melbourne
Level 4, Melbourne Connect (Building 290)
Melbourne, 3010
Australia
Phone: 61 449930128
Email: tianyiz4@student.unimelb.edu.au

Abstract

Background: Smartphones have become an indispensable part of people's lives, and the fear of being without them, what has been termed "no mobile phone phobia" (nomophobia), is a growing phenomenon. The rise of problematic smartphone use highlights the urgent need to explore the intricate relationship between smartphones and human behavior. However, the connections between nomophobia, mental health indicators, smartphone use patterns, and daily activities remain largely underexplored.

Objective: This study aimed to explore the relationship between young adults with depression and smartphones and investigate nomophobia by analyzing data obtained from a pilot study of depression in a youth cohort. Exploring nomophobia can enhance our understanding of the dynamics between young adults and smartphone use, potentially empowering them to manage and regulate their smartphone use more effectively.

Methods: During an 8-week period, data collected via smartphone sensors, such as locations and screen status, were gathered from a cohort of 41 individuals diagnosed with major depressive disorder. In addition to passive-sensing smartphone data, the study collected ecological momentary assessments and psychometric measures, including the Nomophobia Questionnaire, which formed the basis of our investigation. We explored statistical associations among smartphone-derived behavioral features, psychometric indicators, and nomophobia. In addition, we used behavioral and psychometric data to develop regression models demonstrating the prediction of nomophobia levels.

Results: Our findings revealed that the level of nomophobia was positively associated with depression and negative affect, lower geolocation movements, and higher comfort with smartphone sensing. The exploratory predictive linear regression models demonstrated the feasibility of predicting an individual's Nomophobia Questionnaire score based on their smartphone sensing data. These models effectively used input features derived from both a combination of smartphone sensing data and psychometric measures and from smartphone sensing data alone.

Conclusions: Our work is the first to explore the relationship between nomophobia and smartphone sensor data. It provides valuable insights into the predictors of nomophobia level, contributing to the understanding of the relationship between smartphones and human behavior and paving the way for future studies.

(JMIR Form Res 2025;9:e57512) doi:10.2196/57512

KEYWORDS

mobile sensing; nomophobia; digital phenotyping; depression; mental health; smartphone use; personal sensing; behavior analysis; machine learning; mobile health; mobile phone

XSL•FO
RenderX

Introduction

Background

Nomophobia is a portmanteau of "no mobile phone phobia" and describes the fear of being without a smartphone. The concept of nomophobia has been proposed and researched in several early works [1,2], and a subsequent paper [3] presented a validated quantitative scale to measure it, known as the Nomophobia Questionnaire (NMP-Q). The rapid proliferation of smartphones has made nomophobia a timely phenomenon to investigate.

By July 2022, more than two-thirds of the world's population used mobile phones [4]. However, the popularity of smartphones has raised concerns about their impact on individuals' well-being. Smartphone addiction, especially among youth populations, has become a prominent issue. Approximately 95% of teenagers (aged 13-18 years) use cell phones every day in the United States [5]. It is claimed that overattachment to smartphones can lead to actual harmful outcomes such as depression [6], anxiety [7], and reduction in physical activity in the young population.

The notions of problematic smartphone use and smartphone addiction remain unsettled and need further exploration to determine their status as genuine diagnostic phenomena. The last decade has seen various attempts to quantify these notions using self-report scales [8]; however, more work is needed to better understand and predict nomophobia. One promising avenue is the analysis of nomophobia in terms of *digital footprints*, a term used to denote the data trails generated from human interactions with digital devices (eg, smartphones) and the internet (eg, social media). The idea of computationally mining digital footprints to infer an individual's (or a group's) behavior, preferences, and mental states has given rise to a field of research associated with the terms *digital phenotyping* [9], *personal sensing* [10], and *psychoinformatics* [11].

In this study, we investigated the relationship among nomophobia, smartphone sensor data, several mental health measures, and several debriefing questions concerning participant comfort with personal data sensing. This investigation was based on the collection of passive smartphone sensing data, ecological momentary assessment (EMA) responses, and various psychometric measures collected at baseline and the 8-week follow-up of a mental health digital phenotyping study of 41 young people diagnosed with major depressive disorder (MDD). We studied this information in relation to accompanying results on the NMP-Q.

Related Work

Research suggests that nomophobia has been a prevalent phenomenon across youth populations. According to previous investigations, up to 99% of students or young people have some degree of nomophobia [12-16], and a quarter of university students experience severe nomophobia [17,18]. This, in turn, adversely affects their mental or physical health and performance [19], even to the extent that there have been proposals to include nomophobia in the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* [20].

As a nascent field, investigation into nomophobia is still emerging [19,21]. One study [22] used structural equation modeling and semantic network analysis to investigate the cognitive factors associated with nomophobia, showing that people with higher nomophobia tend to "perceive smartphones as their extended selves" [22] due to personal memories. Examining the relationships between nomophobia and psychological assessments, researchers have found various associations between nomophobia and social media use [23], loneliness [23-25], smartphone addiction [26], stress [25], problematic phone use [27], personality traits (Big Five personality traits) [28] and emotional stability [28], depression [29], anxiety [24,29], hyperactivity [29], social phobia [30], and oppositional symptoms [29]. One study [31] found that vertical collectivism, which involves sacrificing personal goals and submitting to authorities, is positively associated with nomophobia. Another study [32] took a qualitative approach and found a strong correlation between smartphone addiction and personal feelings such as social anxiety, antisocial behaviors, and loneliness.

In terms of nomophobia measurement, the basis of our work is the NMP-Q scale [3]. This validated scale consists of 4 dimensions: communication, connection, informativeness, and convenience. The 140-point total on the scale represents the sum of the responses to 20 questions, each rated on a Likert scale from 1 to 7. As per the scale developers, the total score can be categorized into 4 nomophobia severity levels: none and mild, moderate, and severe nomophobia. Since its release, this scale has been used in a handful of studies that affirm the reliance on mobile phones as an emerging issue for young people [16,32-34].

A recent study [35] investigated the NMP-Q scale in terms of various independent variables, particularly certain psychopathological measures, smartphone use, and lifestyle measurements. The study identified interpersonal sensitivity, obsession-compulsion, and the number of hours of smartphone use per day as predictors of nomophobia. While these results are of interest, the study was limited in terms of smartphone sensing. Only a few smartphone use–related inputs were obtained, such as the number of years using a smartphone, the number of daily hours using a smartphone, and the most used apps. Furthermore, the method to measure smartphone use was based on simply asking participants to subjectively provide estimates of these figures. While these results are of interest, the investigation of nomophobia against a greater variety of objectively collected smartphone sensing data is called for.

Thus, building upon this earlier work, in this study, we explored nomophobia outcomes against a range of objectively collected smartphone data inputs and psychometric scale results. Unlike work in digital phenotyping that uses smartphone data to infer psychopathology, our research on nomophobia uses objective phone data to directly characterize patterns of smartphone use.

XSL•FO
RenderX

Methods

Study Design

The data for this study were obtained from the SMARTSENSE-D project [36]. Participants were recruited from primary and specialized youth mental health services in Melbourne, Victoria, Australia, over a 2-year period between 2020 and 2021. Recruitment took longer than anticipated due to the COVID-19 pandemic situation in Melbourne. In total, 41 young people (aged 16-25 years) with a diagnosis of MDD consented to participate in data collection for 8 weeks.

With consent from their clinicians, we contacted potential participants who met the diagnosis criteria for MDD based on results from the Mini-International Neuropsychiatric Interview [37] and Patient Health Questionnaire–9 [38]. We included participants with an MDD diagnosis based on the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* [39], who regularly used a smartphone, understood smartphone data tracking, and were willing to participate. We excluded those at high risk of suicide, actively manic or actively psychotic, not familiar enough with smartphones to complete daily EMA surveys, or with insufficient ability to communicate in English.

Each participant provided demographic information, psychometric assessments (before and after the 8-week period), continuous smartphone sensing data (Android phones only), and data from a wrist actigraphy device during the 8-week trial period. We included in our investigation score results from the following measures and questionnaires that were administered in the study: (1) a demographic questionnaire (baseline assessment only) that captured demographic information such as age, gender, educational level, and other relevant background information; (2) the Depression, Anxiety, and Stress Scale–21 (DASS-21) [40,41], a set of 3 self-reported scales designed to measure the emotional states of depression, anxiety, and stress; (3) the Quick Inventory of Depressive Symptomatology (QIDS) [42,43], a brief self-report tool that assesses the severity of depressive symptoms; (4) the University of California, Los Angeles (UCLA), Loneliness Scale [44,45], which is designed to measure subjective feelings of loneliness and social isolation; (5) the Penn State Worry Questionnaire (PSWQ) [46,47], a self-report measure that assesses the tendency to worry excessively; (6) the Rumination Response Scale (RRS) [48,49], which assesses the extent to which individuals focus on their depressive symptoms and the possible causes and consequences of these symptoms; (7) the NMP-Q [3,50], which measures the severity of nomophobia, which is the fear of being without a mobile phone; and (8) a debriefing questionnaire (follow-up assessment only), which used a Likert scale from 1 (*strongly disagree*) to 7 (*strongly agree*) to capture participant attitudes toward such smartphone sensing and their perspectives on its feasibility and acceptability and also captured participant attitudes regarding each of the sensors being activated on their phones using a Likert scale from 1 (*extremely uncomfortable*) to 7 (*extremely comfortable*).

The AWARE-Light smartphone sensing app [51] was installed on each participant's Android phone and continuously collected geolocation, keyboard, communication (through calls and messages), screen status, app use, and touch operation data using sensors [52,53]. In addition, the AWARE-Light app administered EMA surveys twice daily at noon and 8 PM. The debriefing questionnaire items and EMA questions can be found in Multimedia Appendices 1 and 2, respectively.

Ethical Considerations

This study involved the collection of primary data directly from participants. All participants provided informed consent, which included detailed information about the purpose of the study, the types of data collected, how the data would be stored securely, and how they would be used. Participants were informed that their data would remain confidential and anonymized, and they had the option to withdraw from the study at any time without consequence. The study protocol was reviewed and approved by the University of Melbourne Human Research Ethics Committee (ID: 1955691.4). This study was carried out in accordance with the principles contained in the Declaration of Helsinki, the Australian National Health and Medical Research Council National Statement on Ethical Conduct in Human Research, and the National Health and Medical Research Council Australian Code for the Responsible Conduct of Research.

Participants were reimbursed Aus $100 (US $62.32) upon completion of the study, as well as an additional Aus $1.5 (US $0.93) for completion of each EMA survey (maximum of Aus $126 [US $78.53] if all surveys were completed).

Data Processing

Preprocessing

As participants installed the AWARE-Light sensing app at different times on their first study day, we discarded the data collected during the initial day of the 56-day data collection period. Therefore, the analysis was based on 55 full 24-hour days for each participant starting from the second day of installation.

We handled missing values in 2 steps. First, we only retained the sensor data for a participant if the data were recorded for at least 27 out of the 55 days; if a participant's sensor captured data for ≤26 days, then the data were excluded. The phone call and SMS text message sensors were exempt from this requirement as they may not be used every day. As such, we had different numbers of participants with valid data for each sensor dataset. Given this filtering, we ended up focusing on the smartphone sensor data of participants whose location and screen sensors met the threshold, resulting in a total of 27 valid sets of participant data.

Second, we used imputation to handle missing values in active data, including EMA questions and psychological assessments. In sum, <2% of the psychometric data were missing and imputed, and approximately 30% of EMA data were missing values. The days with missing EMA answers were excluded from feature calculations, such as the average and SD.

To resolve the missing data in the QIDS, we left the missing field as null to represent the missing data. Because the questions in these scales vary significantly, we believe that participants' attitude toward certain questions cannot be represented by the

XSL•FO
RenderX

ratings for other questions in the same scale. Note that the total value of the QIDS remained unchanged after resolving the missing data.

To resolve the missing data in assessments other than the QIDS, we first calculated the average ratings of nonmissing questions for each participant and then used the average score to replace the missing data. In this way, we assumed that participants had no special preference toward the questions they skipped.

For psychometric assessments, missing data arose from participants selecting "Prefer not to say," which we treated as missing data. This suggests that the missingness might be related to the participants' unwillingness to disclose certain information, implying that the data were missing not at random. Given that <2% of the data were missing, we opted for a straightforward imputation method. While this approach does not explicitly account for the potential missing-not-at-random nature of the data, the small proportion of missing data minimized the impact on our overall analysis. Nevertheless, we acknowledge that more sophisticated methods such as pattern-mixture models or sensitivity analysis could be considered in future studies to address any potential biases.

We verified the correlation and root mean square error (RMSE) values for the psychometric data and found that the lowest correlation between the original and imputed datasets was 0.89, whereas the highest correlation corresponded to an RMSE of 0, indicating a near-perfect match. Specifically, the RMSE values for the baseline DASS-21 total, baseline UCLA Loneliness Scale total, baseline PSWQ total, baseline RRS total, and follow-up RRS total were 0.61, 0.94, 2.15, 0.78, and 2.56, respectively. This indicates that the overall patterns remained robust and intact after imputation.

Feature Generation

The smartphone sensor data collected were used to generate a variety of information features. Message, keyboard, and screen status features were generated using the RAPIDS feature generation tool (AWARE Framework) [54], and the library by Doryab et al [55] was used to generate geolocation features such as average speed, maximum length staying at clusters, and number of frequent places [56]. Specifically, the Hierarchical Density-Based Spatial Clustering of Applications With Noise (HDBSCAN) [57] algorithm was used for cluster generation, effectively handling noise and clusters of varying densities.

We first generated daily smartphone sensor data and then averaged the data across the study period. For instance, to calculate the "maximum time spent at significant locations" feature, we first determined the maximum length of time that each participant stayed in various clusters per day, and then we computed the average duration over the 56 days for each participant.

In addition, we created our own Python scripts to work with certain sensors or generate features that were not available using the aforementioned tools, including touch operations, screen status, app use, and daily EMA questions. The fundamental features are the counts and average values of the interaction with the phone, such as counts of clicks and scrolling from touch operations, average and SD of EMA values, and the number of screen unlocks during each hour of the day.

The features derived from passively collected smartphone sensor data were preselected based on their relevance and potential to provide meaningful insights into user behavior and nomophobia. Location-based features offer detailed insights into the daily movement and habits of participants, encompassing various aspects of mobility, such as variability, speed, distance, and location diversity. Message traces and keyboard use for information retrieval or exchange indicate the extent of users' reliance on their smartphones. Screen use time and frequency, count of touch actions, and app use for particular app types not only reflect smartphone engagement but also serve as potential indicators of problematic smartphone use [27].

Regarding the psychometric measures investigated (DASS-21, QIDS, UCLA Loneliness Scale, PSWQ, RRS, and NMP-Q), apart from using the aggregate score of an assessment scale, in certain cases, we also investigated the subscores based on the dimensions or domains in each scale. For example, the four dimensions in the nomophobia scale [27] are (1) not being able to communicate (belonging and connectedness), (2) losing connectedness, (3) not being able to access information, and (4) giving up convenience. Similarly, several of the other psychometric assessments, such as the DASS-21, QIDS, and RRS (see the Models section), were also composed of different dimensions.

It is worth noting that we focused on analyzing the psychometric measures from the follow-up assessment for several reasons. First, the follow-up assessment results aligned better with the phone sensor data in terms of chronological order. The follow-up assessments were chronologically recorded after the passive sensing and could reveal more information about the entire study period. Second, the follow-up assessments from various questionnaires aligned better with each other as they were taken at the same time point. Third, the psychometric measures changed little throughout the 56-day study duration. The ratio of the absolute scale difference (calculated using the formula [after − before]/before) for the DASS-21, PSWQ, RRS, and QIDS was approximately 0.1, whereas the ratio for the UCLA Loneliness Scale was <0.05. This subtle difference between the baseline and follow-up absolute values indicates only minor variations across these scales, suggesting stability in the measured psychometric attributes over the study period.

Apart from the total scores for each psychometric measure, we explored the subfactors because they offered a more detailed and sensitive view of the data. While the total score provides a holistic perspective, the subfactors reveal detailed insights into different aspects of the psychometric constructs. These subfactors can highlight specific areas of change that may not be apparent in the overall scores, offering a more granular understanding. Therefore, we were interested in examining the correlation and predictive power of both the total scores and the subfactor scores of the psychometric measures.

Models

We used the Spearman rank correlation [58] to investigate the statistical associations between nomophobia scores and other

XSL•FO
RenderX

variables of interest. The Spearman correlation was chosen due to the nonnormal distribution of ordinal Likert-type questionnaire measures and continuous smartphone-collected features. We adjusted all *P* values using the original Benjamini-Hochberg method [59] to control the false discovery rate at 20%. This adjustment allows up to 20% of the rejected null hypotheses to be false positives and was chosen to ensure a reasonable balance between sensitivity and specificity in our exploratory analysis.

To test the correlations between smartphone sensor features and nomophobia, we only included a participant's dataset for a sensor if it had at least 27 out of the 55 days of availability (location and screen sensors were selected after this filtering process as they were the 2 sensors with sufficient data). Correlations between psychometric and debriefing questionnaire features and the NMP-Q were explored using valid data for 41 participants. Correlations between EMA features and nomophobia were explored using valid data for 85% (35/41) of the participants.

In total, we selected 48 correlation pairs to investigate, as shown in Textbox 1. The significant feature pairs are presented in the Results section.

We generated the smartphone sensor and psychometric features (Multimedia Appendix 3) shown in Textbox 2.

Textbox 1. Selected pairs of features for correlation investigation.

- A total of 9 location sensor features were each paired with the Nomophobia Questionnaire (NMP-Q) total score (Multimedia Appendix 3).
- A total of 3 screen sensor features were each paired with the NMP-Q total score (Multimedia Appendix 3).
- A total of 8 ecological momentary assessment (EMA) features were each paired with the NMP-Q total score. Each feature was obtained by averaging the morning and evening scores of an EMA affect question (Multimedia Appendix 3).
- A total of 2 debriefing questionnaire features—one based on the question concerning overall comfort with using AWARE-Light (statement: "I felt comfortable using the app") and the other based on the average of questions concerning comfort with the accelerometer, app, communication, location, light, keyboard, network, and screen sensors—were paired with the NMP-Q total score as well as the NMP-Q subdimension scores for a total combination of 10 pairs (Multimedia Appendix 3).
- A total of 4 Depression, Anxiety, and Stress Scale–21 (DASS-21) features were each paired with the NMP-Q total score (the DASS-21 total score, as well as the scores for the 3 subdimensions of depression, stress, and anxiety). Each subdimension or factor was added to the exploration given that they each represented separate mental health conditions. Furthermore, this investigation had a particular focus on depression given the cohort of young adults with depression (Multimedia Appendix 3).
- A total of 11 Quick Inventory of Depressive Symptomatology (QIDS) features were each paired with the NMP-Q total score (the QIDS total score, as well as the scores for the 10 subdimensions). As with the DASS-21, given the focus on depression, it was decided to investigate the subdimensions of this scale as well (Multimedia Appendix 3).
- One Rumination Response Scale total score feature was paired with the NMP-Q total score (Multimedia Appendix 3).
- One Penn State Worry Questionnaire total score feature was paired with the NMP-Q total score (Multimedia Appendix 3).
- One University of California, Los Angeles, Loneliness Scale total score feature was paired with the NMP-Q total (Multimedia Appendix 3).

XSL•FO
RenderX

Textbox 2. Smartphone sensor and psychometric features generated.

Location

- The RAPIDS tool (AWARE Framework) used Density-Based Spatial Clustering of Applications With Noise algorithms (HDBSCAN) to calculate location points into clusters. On the basis of the time spent in each location cluster, 8 features derived from location clusters were explored in correlation with the Nomophobia Questionnaire (NMP-Q) scale. In total, 5 of these features describe significant places (the minimum, maximum, average, and SD of the time spent per day at significant locations), as well as the number of visited significant places per day. The other 3 features include the average speed (indicating how fast a person moves) per day, total traveled distance per day, and time spent at home per day. There may be some overlap of information if the home is considered the most significant location for the person. The maximum period of staying at significant locations refers to the longest duration that a user stays at any identified significant location over a given period, identifying key locations that are important to the user, such as their home, workplace, or frequently visited places. Extended stays at certain locations can provide insights into routines and lifestyle patterns. The SD of the time spent in a cluster reflects the degree of variability in the time participants spend at different location clusters, with a higher SD indicating greater variation.

Screen

- A screen use episode is defined as the period from when the phone is unlocked until the screen is turned off. A total of 3 screen features were explored for correlation: the total duration of unlock episodes, considered as the time when smartphones are actively used; the longest duration of any unlock episode, representing the longest use episode per day; and the number of all unlock episodes, indicating how many times the person unlocked their smartphone in a day. For predictive models, 3 additional features were considered: the minimum, average, and SD of unlock episode duration in a day.

Messaging

- In total, 3 pairs of messaging features were explored for predictive models, including the total number of messages received from or sent to the contact who sent messages most frequently per day, the total number of messages received and sent by the participant per day, and the number of different contacts from whom the participant received and sent messages per day.

App use

- A total of 4 app use features were included in the candidate pool for the predictive model: the total number of times that social media apps were used by the participant per day, the total number of times that communication apps (eg, messaging and email) were used per day, the total number of times that entertainment apps (eg, games and videos) were used per day, and the total number of times that music or audio apps were used per day.

Touch actions

- One feature derived from collected smartphone touch data was the average number of items scrolled. This metric is calculated by tracking the number of distinct scroll actions a user performs on their smartphone screen during a specified period.

Psychometric subfactors

- Our study included both the total scores and subfactor scores of the following psychometrics: Depression, Anxiety, and Stress Scale–21 (DASS-21); Quick Inventory of Depressive Symptomatology (QIDS); University of California, Los Angeles (UCLA), Loneliness Scale; Penn State Worry Questionnaire (PSWQ); Rumination Response Scale (RRS); and NMP-Q. However, not all combinations across the subfactors were explored. We primarily focused on the total score on the NMP-Q for exploring correlations with smartphone sensor features and other psychometrics. In addition, for specific debriefing questions, we examined subfactors of participants' NMP-Q scores to understand why they felt comfortable with sensitive data being collected via smartphones. The DASS-21 is designed to be interpreted using 3 subscales: depression, anxiety, and stress. The QIDS also includes subfactors even though some subfactors are represented by a single question. In contrast, the RRS, PSWQ, and UCLA Loneliness Scale do not have subfactor representations and, thus, were not explored in this context.

In general, we investigated the correlation between each of these features and the NMP-Q total score. The one exception to this was the debriefing questionnaire features, which we also tested against the NMP-Q subdimensions because of our preassumption that using AWARE-Light and sensor activation might naturally correlate with certain aspects of nomophobia, such as the need for communication or convenience.

Apart from the correlation analysis, we also experimented with constructing some supervised machine learning models with the follow-up nomophobia scores as the outcome to be predicted. In our supervised learning models, we only included data from participants whose location and screen sensor data met the threshold of at least 27 out of the 55 days of availability (which resulted in 27/41, 66% of the participants included in the analysis). If a participant met this requirement for both sensors, then data for all their sensors, their psychometric results, and their EMA data were retained for regression.

After data cleaning, linear regression was applied to 47 extracted features (Multimedia Appendix 4) from 66% (27/41) of the participants, with filtering based on the combination of location and screen sensors. We focused on exploring two general models: (1) those whose input was only passive smartphone sensor features and (2) those whose input combined passive and active data, including EMA, phone sensors, and psychometrics. We implemented both wrapper methods and filter methods to test various combinations of features using R^2 to evaluate the models. Of the 47 preselected features, we included those that showed a potential improvement in model performance of >0.01 in terms of R^2 results. To predict NMP-Q scores, we used leave-one-out cross-validation for model

XSL•FO
RenderX

evaluation due to the small sample size. In addition to the R^2 score, we used mean absolute error (MAE), mean absolute percentage error (MAPE), and RMSE to assess the regression errors.

Results

Data Description

In total, 41 individuals took part in the study. Their demographic characteristics are presented in Table 1. The means and SDs of the psychometrics are shown in Table 2. The NMP-Q total scores of the participants only slightly changed between baseline and follow-up, as shown in Figure 1, with a Spearman test-retest reliability of 0.87 and a Cronbach α of 0.93. A total of 4 nomophobia severity levels as defined in the study by Yildirim and Correia [3] were used: no nomophobia (score of 0-20), mild nomophobia (score of 21-59), moderate nomophobia (score of 60-99), and severe nomophobia (score of 100-140). All participants (41/41, 100%) had an NMP-Q score of >20, with most participants (18/41, 44%) reporting a moderate level of nomophobia. A similar percentage of participants reported a mild (11/41, 27%) or severe (12/41, 29%) level of nomophobia.

Table 1. Demographic information of the young adults with depression involved in the 8-week longitudinal study (N=41).

Characteristic and categories	Participants, n (%)
Sex at birth	
Female	29 (71)
Male	12 (29)
Age group (y)	
16-18	11 (27)
19-22	17 (41)
23-25	13 (32)
Country of birth	
Australia	35 (85)
Overseas (South America, Asia, Europe, and New Zealand)	6 (15)

Table 2. Means, SDs, and score ranges of the psychometric measurements taken at the end of the 8-week study among 41 participants for the Depression, Anxiety, and Stress Scale–21 (DASS-21); Penn State Worry Questionnaire (PSWQ); Rumination Response Scale (RRS); University of California, Los Angeles (UCLA), Loneliness Scale; and Quick Inventory of Depressive Symptomatology (QIDS).

Psychometric scale	Scores, mean (SD; range)
DASS-21	30.6 (12.9; 0-126)
PSWQ	59.1 (12.9; 16-80)
RRS	26.9 (5.5; 4-40)
UCLA Loneliness Scale	57.0 (5.0; 20-80)
QIDS	15.0 (5.5; 0-27)

XSL•FO
RenderX

Figure 1. Comparison of Nomophobia Questionnaire (NMP-Q) total scores among 41 participants with major depressive disorder at baseline and at the end of the 8-week longitudinal study period, revealing minimal differences.

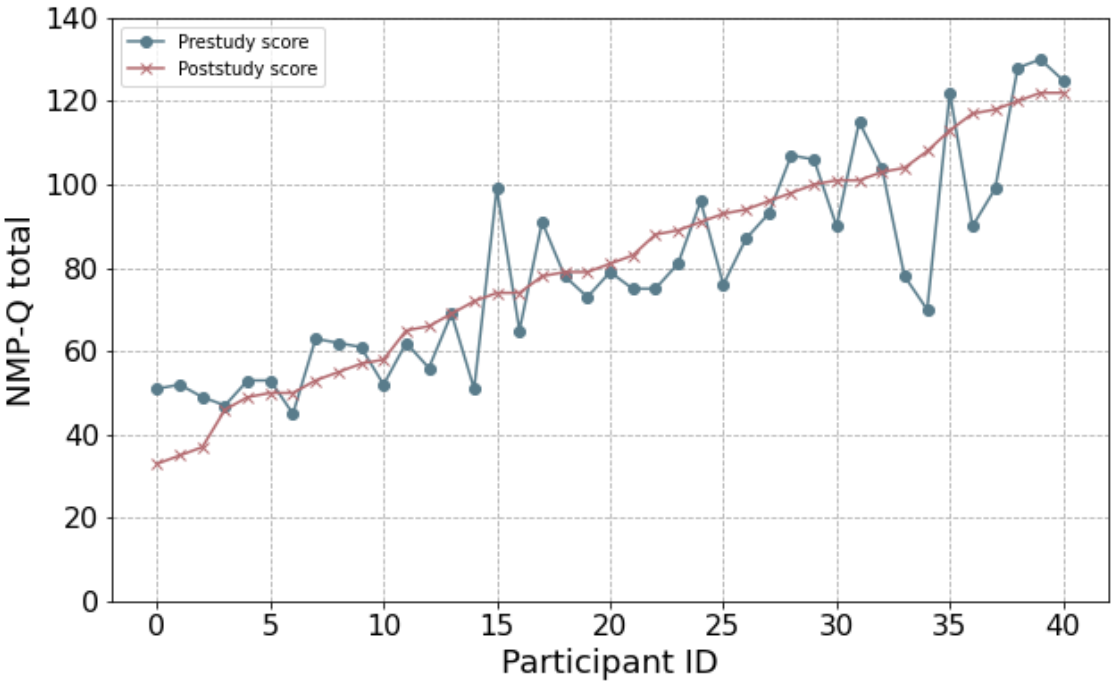

Statistical Analysis

Exploratory data analysis was conducted using the Spearman rank correlation analyzing follow-up NMP-Q results against the psychometric, questionnaire, passive smartphone sensing, and EMA data. As detailed previously, we examined 48 feature pairs, with each pair involving the NMP-Q total score or that of one of its subfactors. Of these 48 tested combinations, we found 11 (23%) significant results that are presented and analyzed in this section, with the remaining 37 (77%) tests yielding nonsignificant results (Multimedia Appendix 3).

To qualify for inclusion as significant, the results needed to have $P \leq .05$ and a correlation coefficient of >0.3 or <−0.3, in line with the standard by Cohen [60], according to which results of >0.3 are of moderate effect.

Overall, significant correlations with the NMP-Q were found for certain EMA items, certain geolocation features, certain debriefing questions, and certain psychometric scores. The following is a descriptive summary of these correlation results.

Regarding location data (Table 3), there was a positive correlation between nomophobia (NMP-Q total score) and time spent at home, time spent at the most significant location, the maximum time spent in a cluster (at a daily level, this feature captures the maximum time that a participant spent at any of their significant locations [as determined by the geolocation clustering algorithm] for the day), and the SD of the time spent in a cluster (significant location; at a daily level, this feature captures the SD of all the durations of time that a participant spent at their significant locations for the day).

Regarding debriefing questions (Table 4), a positive correlation was found between the comfort level of a participant using the sensing app and their nomophobia in terms of the NMP-Q factor of giving up convenience, and a positive correlation was found between the comfort level of a participant having certain sensors tracking their personal information and their nomophobia in terms of the NMP-Q factor of giving up convenience.

Regarding the DASS-21 (Table 5), there was a positive correlation between the depression score and nomophobia (NMP-Q total score).

Regarding the QIDS (Table 5), mood and fatigue in the depressive cohort were positively correlated with nomophobia (NMP-Q total score).

Regarding EMA (Table 6), participants who felt more relaxed and happier daily reported lower nomophobia (NMP-Q total score).

XSL•FO
RenderX

Table 3. Significant correlations between Nomophobia Questionnaire total scores and smartphone-based location sensor features among 27 participants with major depressive disorder during the 8-week study period.

Location feature	*P* value	Spearman correlation coefficient
The maximum period staying at significant locations	.02	0.35
The SD of the time spent in a cluster	.04	0.32
Time spent at home	.02	0.35
Time spent at the most significant location	.03	0.35

Table 4. Significant correlations between the "giving up convenience" subfactor of the Nomophobia Questionnaire and participants' responses to debriefing questions at the conclusion of the 8-week study period observed among 41 participants with major depressive disorder.

Feature	*P* value	Spearman correlation coefficient
Participant comfort with using AWARE-Light (item 4 in Multimedia Appendix 1)	.02	0.37
Participant comfort with sensor activation (an average of items 21-28 in Multimedia Appendix 1)	.01	0.40

Table 5. Significant correlations between Nomophobia Questionnaire total scores and various psychometric measures assessed among 41 participants with major depressive disorder at the conclusion of the 8-week study period.

Psychometric measure	*P* value	Spearman correlation coefficient
DASS-21[a] factor—depression	.02	0.38
QIDS[b] factor—mood	.046	0.31
QIDS factor—energy and fatigue	.02	0.37

[a]DASS-21: Depression, Anxiety, and Stress Scale–21.

[b]QIDS: Quick Inventory of Depressive Symptomatology.

Table 6. Significant correlations between Nomophobia Questionnaire total scores and responses to ecological momentary assessment (EMA) questions captured over the 8-week study period among 41 participants with major depressive disorder.

EMA measure	*P* value	Spearman correlation coefficient
Average of morning and evening EMA score for how happy participants felt	.03	–0.34
Average of morning and evening EMA score for how relaxed participants felt	.03	–0.337

Predicting Nomophobia

Tables 7 and 8 show the attributes of our best models for passively plus actively collected data and passively collected data only, respectively. Our best model achieved an R^2 value of 0.779 (MAE=9.42; MAPE=13.56%; RMSE=12.09). Figure 2 shows the distribution of predicted and actual NMP-Q scores based on both phone sensor features and psychometric measures input into a linear regression model. We also developed a model to predict individuals' NMP-Q scores using only smartphone sensor data. Our best model achieved an R^2 of 0.732 (MAE=10.88; MAPE=14.44%; RMSE=13.32), as illustrated in Figure 3.

Table 7. Overview of smartphone sensor data and survey features included in the linear regression model to predict Nomophobia Questionnaire total scores among 27 participants with depressive symptoms over the 8-week study period, reaching an R^2 value of 0.779.

Rank	Feature	Sensor or questionnaire	Attribute coefficient
1	Count of social media app use	App	−17.55
2	Sleep disturbance	QIDS[a]	−4.40
3	Count of sent messages to the most frequent contact	Message	−1.5
4	The total traveled distance	Location	−1.32
5	SD of unlock episode duration	Screen	0.81
6	The maximum period staying at significant locations	Location	0.03

[a]QIDS: Quick Inventory of Depressive Symptomatology.

XSL•FO
RenderX

Table 8. Overview of smartphone sensor features included in the linear regression model to predict Nomophobia Questionnaire total scores among 27 participants with depressive symptoms over the 8-week study period, reaching an R^2 value of 0.732.

Rank	Feature	Sensor or questionnaire	Attribute coefficient
1	Count of social media app use	App	−20.59
2	Count of sent messages to the most frequent contact	Message	−2.23
3	SD of unlock episode duration	Screen	0.81
4	The maximum period of staying at significant locations	Location	0.02
5	The area covered by a participant	Location	–0.0005
6	The total traveled distance	Location	–0.00000218
7	The average amount of items scrolled	Touch	0.00000158

Figure 2. Comparison of actual versus predicted Nomophobia Questionnaire (NMP-Q) total scores using a combination of psychometric measures and smartphone sensor data from 27 participants with depressive symptoms who had valid data during the 8-week study period. This highlights the potential of machine learning for accurate predictions.

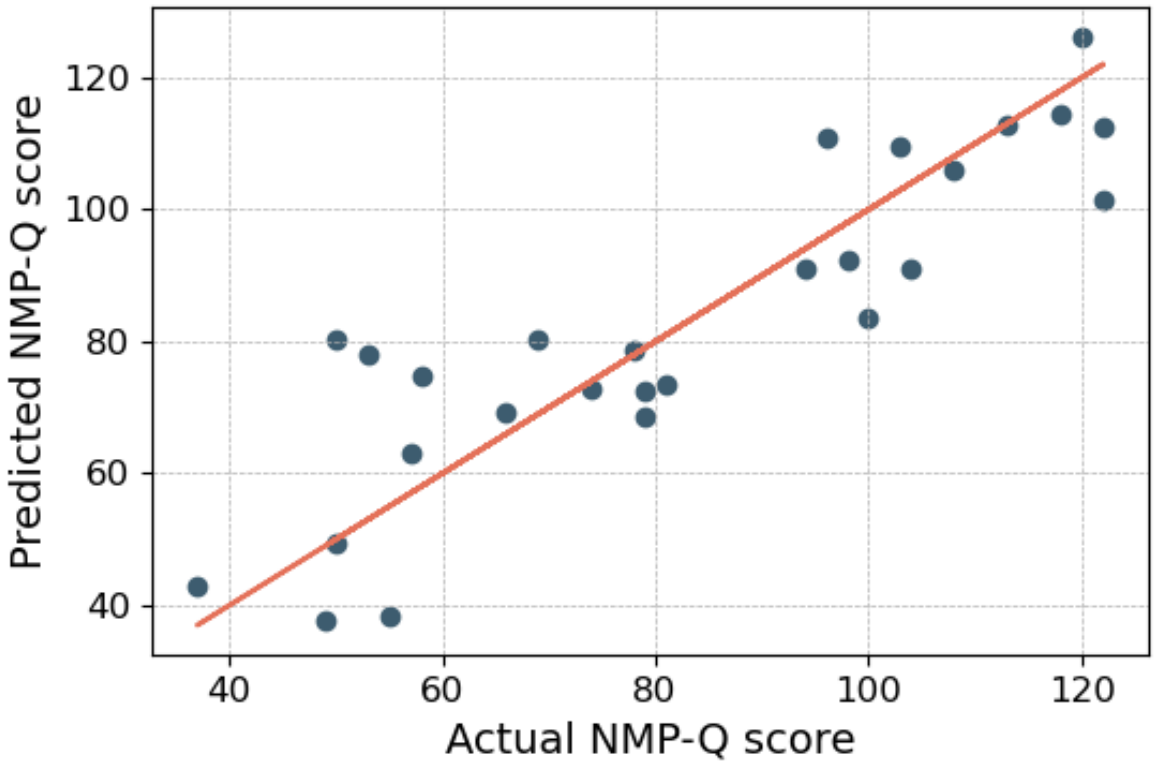

XSL•FO
RenderX

Figure 3. Comparison of actual versus predicted Nomophobia Questionnaire (NMP-Q) total scores using only smartphone sensor data from 27 participants with depressive symptoms who had valid data during the 8-week study period without incorporating psychometric measures. This demonstrates the feasibility of machine learning–based predictions.

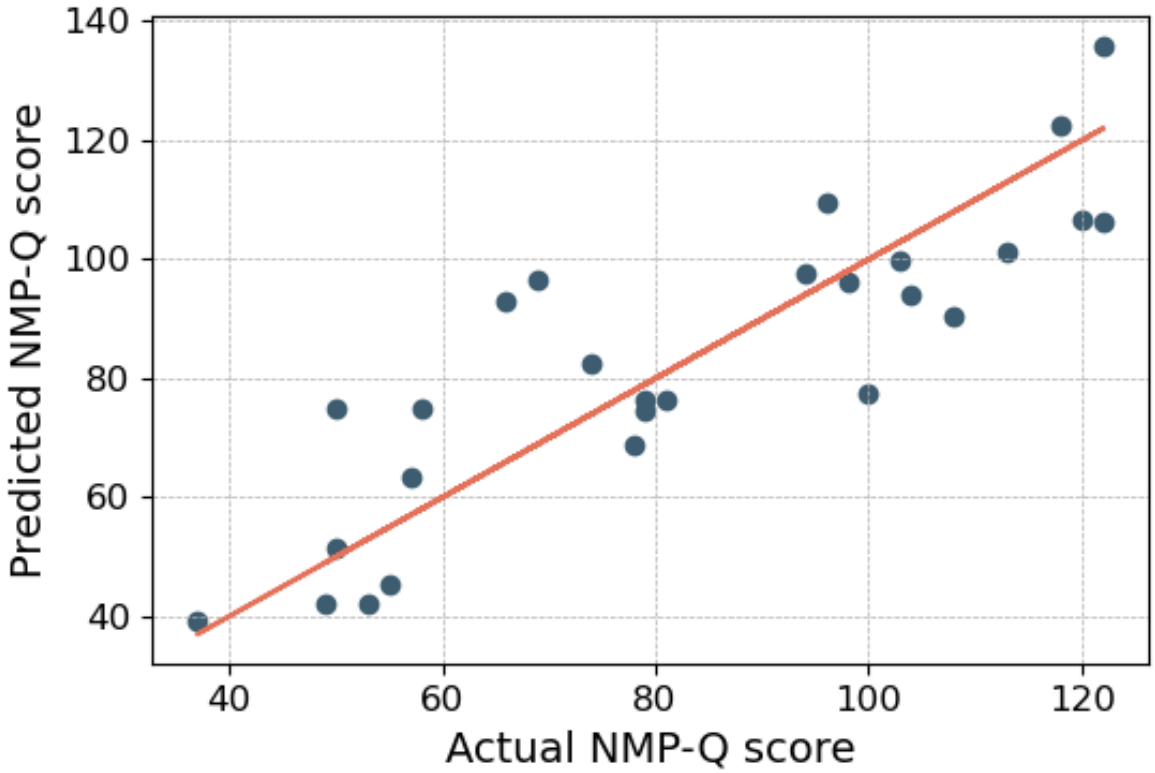

Both models' performance was mostly related to the frequency of social media app use. The count of sent messages to the most frequent contact, the duration of staying at clusters, and the SD of the duration of screen unlock episodes were also important features revealing smartphone use and location information in both models.

Discussion

Themes

The results from our exploratory study are suggestive of several themes concerning nomophobia and its associations with both psychopathology and smartphone use.

Depression and Negative or Positive Affect

Given that the SMARTSENSE-D study focused on a cohort of young people diagnosed with clinical depression and varying degrees of symptom severity, the data provided a good opportunity to explore associations between nomophobia and depression. The 2 psychometric scales concerning depression that were used in the study, the DASS-21 and QIDS, both showed significant associations with nomophobia. The DASS-21 depression scale and 2 QIDS factors positively correlated with total nomophobia scores.

Our EMA results across the study indicate that the more positive affect one has in general, the lower their nomophobia. The most salient of these results concern happiness and relaxation, for which the EMA results negatively correlated with total nomophobia scores.

These findings regarding associations among depression, affect, and nomophobia are in line with previous findings [61-67]. Given the co-occurrence of depression and nomophobia, there is a possibility that depressive symptoms and negative affect are influenced by nomophobia, or vice versa. The connection and causality among depression, negative feelings, and nomophobia need to be further validated through additional research.

Location

Our results suggest that participants who spent more time at home or at their otherwise most visited location had greater nomophobia. Relatedly, one's maximum duration at a significant location also positively correlated with nomophobia. This implies that the more an individual is confined to their home or the more concentrated their geolocation to a particular location, the greater their sense of dependency on their smartphone.

Our results also showed a positive correlation between nomophobia and the SD of all the durations of time that a participant spent at their significant locations for the day. This is a result that is perhaps less straightforward to interpret; we can say that, if one has a lower SD, then there is a more evenly spread diversity in the time they spend at their various significant locations. This contrasts with a higher SD, which implies a concentration in fewer places. In this case, the SD

XSL•FO
RenderX

result is in line with the other geolocation results discussed previously.

The fact that nomophobia was associated both with geolocation and depressive symptoms is in line with previous research that has demonstrated a significant link between movement patterns and depression [56,68-70]. Thus, a triangle emerges that connects geolocation, depression, and nomophobia. Ultimately, limited or confined geolocation activity is associated with nomophobia and can be an indicator of depression.

Nomophobia-Privacy Trade-Off

A trend of positive correlations emerged between nomophobia and comfort with being tracked via a smartphone sensing app and sharing data. The NMP-Q "giving up convenience" dimension positively correlated with comfort levels with using the AWARE-Light sensing app and activating sensors within it.

Thus, a trade-off seems to emerge whereby individuals with nomophobic tendencies, particularly their concern of giving up the convenience afforded by smartphones, are less concerned about being smartphone sensed. Alternatively, one could perhaps frame this as being about privacy concerns outweighing certain smartphone benefits and that those more concerned about privacy are less inclined to develop nomophobia.

At any rate, these findings relate to those of previous studies that have shown how privacy concerns can influence smartphone use behaviors as a dependent indicator [71,72]. Our results from the self-reported evaluations of both participants' nomophobia levels and attitudes toward privacy further support this correlation between smartphone use and privacy concerns.

Smartphone Use

Previous studies [27] have indicated a positive correlation between nomophobia and problematic smartphone use, including screen status (eg, screen locks and unlocks and turning the screen on or off). While none of our formal results involving screen sensor features were significant, we did observe using a scatter plot that, when limited to participants in the mild and severe NMP-Q categories (moderate category excluded), there was a noticeable positive correlation between the number of phone unlocks and NMP-Q scores. However, further work is required to formally strengthen the case that nomophobia is associated with greater screen use, such as unlock instances.

Predictive Models

The predictive results shown by our linear regression models indicate the feasibility of predicting the NMP-Q score based on passive human behavioral features. The implications of these results can be valuable for the understanding, detection, and control of nomophobia.

The most important feature in both models, the count of social media app use, showed a negative relationship with the predicted ground-truth NMP-Q score, indicating that engagement with social media apps co-occurs with a lower level of nomophobia. Meanwhile, the count of sent messages to the most frequent contact and the total traveled distance demonstrated different levels of negative association with the NMP-Q. These findings should be highlighted for further exploration regarding the role of social media apps in managing the relationship between users and smartphones.

Limitations

Our primary goal was to understand the relationship between the features and the NMP-Q, demonstrating the feasibility of predicting NMP-Q scores based on smartphone sensor data and psychometrics or even using only passively collected data. However, we acknowledge that more complex models such as Extreme Gradient Boosting or support vector machines might be valuable in future research with more data to explore potential nonlinear relationships within the data. These advanced models could offer deeper insights and improve prediction accuracy by capturing intricate patterns that simpler linear models might miss.

This study was exploratory and aimed to gain preliminary insights into associations between smartphone-inferred behavioral features and nomophobia (NMP-Q score), as well as the potential of machine learning models to predict NMP-Q scores from such features. The primary intent was not to provide a model ready for practical application at this stage but to demonstrate the feasibility of model use in future stages with more extensive data over a longer time frame. Predictive models for NMP-Q score could assist researchers, patients with depression, and mental health professionals in understanding the impact of smartphone use and daily behaviors on mental health from a technological perspective.

Our inferential statistical results provide a starting point for more dedicated studies with larger samples sizes to strengthen the claims suggested by our results. The use of a 20% false discovery rate may have introduced potential false positives, especially in a small cohort of participants. This threshold aimed to balance false positives and false negatives. However, to achieve more robust and precise conclusions, the findings need further confirmation with larger sample sizes. The larger datasets would also incorporate more advanced machine learning techniques and be less prone to issues such as overfitting. With more data and refined models, clinicians could better understand depression and its relationship with smartphone use, aiding those with depressive symptoms. This could offer valuable insights into how these digital interactions influence their lives and aid in better mental health management.

In addition to the small sample size, another limitation was the imperfection of our data collection given technical issues encountered, issues with AWARE-Light in particular, and the general difficulty of getting smartphone sensing data collection to work across various sensors and various participant phones. Thus, we had to implement a data quality filtering procedure to ensure that we only used a phone's data for a given sensor if those data met certain threshold requirements. Consequently, not all participants were included for every sensor.

Conclusions

Our study explored associations among nomophobia, clinical measures, EMAs, smartphone sensor data, and smartphone use questionnaires. We discovered that greater nomophobia is linked to more depressive symptoms, limited geolocation movement

XSL•FO
RenderX

and (domestic) stationariness, comfort with being smartphone sensed, and negative emotions. In doing so, we demonstrated that smartphone sensing can effectively quantify user behaviors and perceptions underlying nomophobia and the phenomena that its measures are designed to capture.

Nomophobia is a nascent notion that merits further investigation. Beyond conceptualizations and measurement scales, use of the technology in question (ie, smartphones) to provide objective quantification of nomophobic-related use and behavioral phenomena requires further exploration through larger and more in-depth purpose-built studies. It is hoped that these initial exploratory findings will contribute to paving the way for such future work.

Acknowledgments

The authors would like to thank the headspace National Youth Mental Health Foundation for supporting the recruitment process and those who volunteered to participate in the SMARTSENSE-D study. They would also like to thank the University of Melbourne, the Cisco University Research Program Fund, and the Silicon Valley Community Foundation for their funding support. This study was funded by the University of Melbourne (Faculty of Medicine, Dentistry, and Health Sciences establishment grant awarded to LS) and by a grant from the Cisco University Research Program Fund and the Silicon Valley Community Foundation awarded to SD (ID 2021-232746 3696). The authors acknowledge the use of generative artificial intelligence solely for language refinement, with no involvement in the research, methodologies, or coding presented in this work.

Data Availability

The datasets generated or analyzed during this study are not publicly available because not all participants in this study provided consent for their data to be openly shared. For the remaining participants, the data are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Questions presented in the debriefing questionnaire administered at the end of the 8-week study. Items 1 to 20 are provided in scales from 1 to 7 indicating the level of agreement. Items 21 to 28 are provided in scales from 1 to 7 indicating the comfort level of the participants with the sensor being activated on their smartphones. Items 29 and 30 are provided as Yes or No options.
[DOCX File , 23 KB - formative_v9i1e57512_app1.docx]

Multimedia Appendix 2
Daily ecological momentary assessment items for 41 participants with major depressive disorder. The 21-question survey, administered at noon and 8 PM, used 7-point Likert-scale items (questions 1-14 and 18), binary (yes or no) items (questions 15 and 19), and open-text responses (questions 16, 17, 20, and 21).
[DOCX File , 23 KB - formative_v9i1e57512_app2.docx]

Multimedia Appendix 3
Comprehensive correlation analysis results between Nomophobia Questionnaire total scores and various study variables, including psychometric measures assessed at the end of the study and smartphone sensing features recorded throughout the study. The analysis includes 41 participants for the psychometric data and 27 participants for variables involving smartphone sensor features, including both significant and nonsignificant associations.
[DOCX File , 26 KB - formative_v9i1e57512_app3.docx]

Multimedia Appendix 4
Preselected features used in predictive models to estimate Nomophobia Questionnaire total scores for 27 participants with major depressive disorder. These features include psychometric measures collected at the end of the 8-week study and smartphone-derived variables recorded throughout the study.
[DOCX File , 26 KB - formative_v9i1e57512_app4.docx]

References

1. Nomophobia is the fear of being out of mobile phone contact - and it's the plague of our 24/7 age. The Standard. 2012 Apr 13. URL: https://www.standard.co.uk/hp/front/nomophobia-is-the-fear-of-being-out-of-mobile-phone-contact-and-its-the-plague-of-our-24-7-age-6634478.html [accessed 2025-02-07]

XSL•FO
RenderX

2. King AL, Valença AM, Silva AC, Baczynski T, Carvalho MR, Nardi AE. Nomophobia: dependency on virtual environments or social phobia? Comput Hum Behav 2013 Jan;29(1):140-144. [doi: 10.1016/j.chb.2012.07.025]
3. Yildirim C, Correia AP. Exploring the dimensions of nomophobia: development and validation of a self-reported questionnaire. Comput Hum Behav 2015 Aug;49:130-137. [doi: 10.1016/j.chb.2015.02.059]
4. Digital around the world. Data Reportal. URL: https://datareportal.com/global-digital-overview [accessed 2025-02-07]
5. Rideout V, Peebles A, Mann S, Robb MB. Common sense census: media use by tweens and teens, 2021. Common Sense. 2022. URL: https://calio.dspacedirect.org/items/e760d928-dea3-4b81-901b-5bd48cb25f7d [accessed 2025-02-07]
6. Xie H, Tao S, Zhang Y, Tao F, Wu X. Impact of problematic mobile phone use and insufficient physical activity on depression symptoms: a college-based follow-up study. BMC Public Health 2019 Dec 05;19(1):1640. [doi: 10.1186/s12889-019-7873-z] [Medline: 31805915]
7. Coyne SM, Stockdale L, Summers K. Problematic cell phone use, depression, anxiety, and self-regulation: evidence from a three year longitudinal study from adolescence to emerging adulthood. Comput Hum Behav 2019 Jul;96:78-84. [doi: 10.1016/j.chb.2019.02.014]
8. Harris B, Regan T, Schueler J, Fields SA. Problematic mobile phone and smartphone use scales: a systematic review. Front Psychol 2020 May 5;11:672 [FREE Full text] [doi: 10.3389/fpsyg.2020.00672] [Medline: 32431636]
9. Torous J, Kiang MV, Lorme J, Onnela JP. New tools for new research in psychiatry: a scalable and customizable platform to empower data driven smartphone research. JMIR Ment Health 2016 May 05;3(2):e5165 [FREE Full text] [doi: 10.2196/mental.5165] [Medline: 27150677]
10. Mohr DC, Zhang M, Schueller SM. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. Annu Rev Clin Psychol 2017 May 08;13:23-47 [FREE Full text] [doi: 10.1146/annurev-clinpsy-032816-044949] [Medline: 28375728]
11. Montag C, Duke É, Markowetz A. Toward psychoinformatics: computer science meets psychology. Comput Math Methods Med 2016;2016:2983685 [FREE Full text] [doi: 10.1155/2016/2983685] [Medline: 27403204]
12. Qutishat M, Rathinasamy Lazarus E, Razmy AM, Packianathan S. University students' nomophobia prevalence, sociodemographic factors and relationship with academic performance at a University in Oman. Int J Afr Nurs Sci 2020;13:100206. [doi: 10.1016/j.ijans.2020.100206]
13. Dalbudak I, Yilmaz T, Yigit S. Nomophobia levels and personalities of university students. J Educ Learn 2020;9(2):166-177. [doi: 10.5539/jel.v9n2p166]
14. Gurbuz IB, Ozkan G. What is your level of nomophobia? An investigation of prevalence and level of nomophobia among young people in Turkey. Community Ment Health J 2020 Jul;56(5):814-822. [doi: 10.1007/s10597-019-00541-2] [Medline: 31902048]
15. Bragazzi NL, Re TS, Zerbetto R. The relationship between nomophobia and maladaptive coping styles in a sample of Italian young adults: insights and implications from a cross-sectional study. JMIR Ment Health 2019 Apr 24;6(4):e13154 [FREE Full text] [doi: 10.2196/13154] [Medline: 30763254]
16. Farooqui IA, Pore P, Gothankar J. Nomophobia: an emerging issue in medical institutions? J Ment Health 2018 Oct;27(5):438-441. [doi: 10.1080/09638237.2017.1417564] [Medline: 29271270]
17. Humood A, Altooq N, Altamimi A, Almoosawi H, Alzafiri M, Bragazzi NL, et al. The prevalence of nomophobia by population and by research tool: a systematic review, meta-analysis, and meta-regression. Psych 2021 Jun 21;3(2):249-258. [doi: 10.3390/psych3020019]
18. Jilisha G, Venkatachalam J, Menon V, Olickal JJ. Nomophobia: a mixed-methods study on prevalence, associated factors, and perception among college students in Puducherry, India. Indian J Psychol Med 2019 Nov 01;41(6):541-548. [doi: 10.4103/ijpsym.ijpsym_130_19]
19. Rodríguez-García AM, Moreno-Guerrero AJ, López Belmonte J. Nomophobia: an individual's growing fear of being without a smartphone-a systematic literature review. Int J Environ Res Public Health 2020 Jan 16;17(2):580 [FREE Full text] [doi: 10.3390/ijerph17020580] [Medline: 31963208]
20. Bragazzi NL, Del Puente G. A proposal for including nomophobia in the new DSM-V. Psychol Res Behav Manag 2014 May 16;7:155-160 [FREE Full text] [doi: 10.2147/PRBM.S41386] [Medline: 24876797]
21. Bhattacharya S, Bashar MA, Srivastava A, Singh A. NOMOPHOBIA: NO MObile PHone PhoBIA. J Family Med Prim Care 2019 Apr;8(4):1297-1300 [FREE Full text] [doi: 10.4103/jfmpc.jfmpc_71_19] [Medline: 31143710]
22. Han S, Kim KJ, Kim JH. Understanding nomophobia: structural equation modeling and semantic network analysis of smartphone separation anxiety. Cyberpsychol Behav Soc Netw 2017 Jul;20(7):419-427. [doi: 10.1089/cyber.2017.0113] [Medline: 28650222]
23. Yıldız Durak H. What would you do without your smartphone? Adolescents' social media usage, locus of control, and loneliness as a predictor of nomophobia. Addicta 2018 Sep 1;5(3):543-557. [doi: 10.15805/addicta.2018.5.2.0025]
24. Kara M, Baytemir K, Inceman-Kara F. Duration of daily smartphone usage as an antecedent of nomophobia: exploring multiple mediation of loneliness and anxiety. Behav Inf Technol 2019 Oct 09;40(1):85-98. [doi: 10.1080/0144929x.2019.1673485]
25. Dahiya P. Nomophobia and its effects on stress and loneliness among young adults. Int J Indian Psychol 2021;9(3):993-999. [doi: 10.25215/0903.091]

XSL•FO
RenderX

26. Yildiz Durak H. Investigation of nomophobia and smartphone addiction predictors among adolescents in Turkey: demographic variables and academic performance. Soc Sci J 2019 Dec 23;56(4):492-517. [doi: 10.1016/j.soscij.2018.09.003]
27. Kaviani F, Robards B, Young KL, Koppel S. Nomophobia: is the fear of being without a smartphone associated with problematic use? Int J Environ Res Public Health 2020 Aug 19;17(17):6024 [FREE Full text] [doi: 10.3390/ijerph17176024] [Medline: 32824979]
28. Argumosa-Villar L, Boada-Grau J, Vigil-Colet A. Exploratory investigation of theoretical predictors of nomophobia using the Mobile Phone Involvement Questionnaire (MPIQ). J Adolesc 2017 Apr;56:127-135. [doi: 10.1016/j.adolescence.2017.02.003] [Medline: 28231484]
29. Moreno-Guerrero AJ, Aznar-Díaz I, Cáceres-Reche P, Rodríguez-García AM. Do age, gender and poor diet influence the higher prevalence of nomophobia among young people? Int J Environ Res Public Health 2020 May 24;17(10):3697 [FREE Full text] [doi: 10.3390/ijerph17103697] [Medline: 32456304]
30. Uysal S, Özen H, Madenoğlu C. Social phobia in higher education: the influence of nomophobia on social phobia. Global eLearn J 2016;5(2):1-8.
31. Arpaci I. Culture and nomophobia: the role of vertical versus horizontal collectivism in predicting nomophobia. Inf Dev 2017 Sep 05;35(1):96-106. [doi: 10.1177/0266666917730119]
32. Anshari M, Alas Y, Sulaiman E. Smartphone addictions and nomophobia among youth. Vulnerable Child Youth Stud 2019 May 06;14(3):242-247. [doi: 10.1080/17450128.2019.1614709]
33. Yildirim C, Sumuer E, Adnan M, Yildirim S. A growing fear: prevalence of nomophobia among Turkish college students. Inf Dev 2016 Jul 09;32(5):1322-1331. [doi: 10.1177/0266666915599025]
34. MB P, Madhukumar S, TS MM. A study on nomophobia - mobile phone dependence, among students of a medical college in Bangalore. Nat J Community Med 2015 Sep 30;6(03):340-344.
35. Gonçalves S, Dias P, Correia AP. Nomophobia and lifestyle: smartphone use and its relationship to psychopathologies. Comput Hum Behav Rep 2020;2:100025. [doi: 10.1016/j.chbr.2020.100025]
36. D'Alfonso S, Zhang T, Camargo A, Tagliaferri S, Davies P, Kostakos V, et al. SmartSense-D. Open Science Framework Home. 2024. URL: https://osf.io/z4mj8 [accessed 2025-02-07]
37. Sheehan DV, Lecrubier Y, Sheehan KH, Amorim P, Janavs J, Weiller E, et al. The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. J Clin Psychiatry 1998;59 Suppl 20:22-33;quiz 34-57. [Medline: 9881538]
38. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
39. Guha M. Diagnostic and statistical manual of mental disorders: DSM-5 (5th edition). Ref Rev 2014;28(3):36-37 [FREE Full text] [doi: 10.1108/RR-10-2013-0256]
40. Henry JD, Crawford JR. The short‐form version of the Depression Anxiety Stress Scales (DASS‐21): construct validity and normative data in a large non‐clinical sample. Br J Clinic Psychol 2011 Jan 13;44(2):227-239. [doi: 10.1348/014466505x29657]
41. Osman A, Wong JL, Bagge CL, Freedenthal S, Gutierrez PM, Lozano G. The Depression Anxiety Stress Scales-21 (DASS-21): further examination of dimensions, scale reliability, and correlates. J Clin Psychol 2012 Dec;68(12):1322-1338. [doi: 10.1002/jclp.21908] [Medline: 22930477]
42. Rush AJ, Trivedi MH, Ibrahim HM, Carmody TJ, Arnow B, Klein DN, et al. The 16-Item Quick Inventory of Depressive Symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. Biol Psychiatry 2003 Sep 01;54(5):573-583. [doi: 10.1016/s0006-3223(02)01866-8] [Medline: 12946886]
43. Reilly TJ, MacGillivray SA, Reid IC, Cameron IM. Psychometric properties of the 16-item Quick Inventory of Depressive Symptomatology: a systematic review and meta-analysis. J Psychiatr Res 2015 Jan;60:132-140. [doi: 10.1016/j.jpsychires.2014.09.008] [Medline: 25300442]
44. Russell DW. UCLA Loneliness Scale (Version 3): reliability, validity, and factor structure. J Pers Assess 1996 Feb;66(1):20-40. [doi: 10.1207/s15327752jpa6601_2] [Medline: 8576833]
45. Russell D, Peplau LA, Cutrona CE. The revised UCLA Loneliness Scale: concurrent and discriminant validity evidence. J Pers Soc Psychol 1980 Sep;39(3):472-480. [doi: 10.1037//0022-3514.39.3.472] [Medline: 7431205]
46. Meyer TJ, Miller ML, Metzger RL, Borkovec TD. Development and validation of the Penn State Worry Questionnaire. Behav Res Ther 1990;28(6):487-495. [doi: 10.1016/0005-7967(90)90135-6] [Medline: 2076086]
47. Davey GC, Wells A. The Penn State Worry Questionnaire (PSWQ). In: Worry and its Psychological Disorders: Theory, Assessment and Treatment. Hoboken, NJ: John Wiley & Sons; Jan 01, 2006:99-119.
48. Treynor W, Gonzalez R, Nolen-Hoeksema S. Rumination reconsidered: a psychometric analysis. Cogn Ther Res 2003;27:247-259. [doi: 10.1023/A:1023910315561]
49. Quick BL, Stephenson MT. The Reactance Restoration Scale (RRS): a measure of direct and indirect restoration. Commun Res Rep 2007 May 11;24(2):131-138. [doi: 10.1080/08824090701304840]

XSL•FO
RenderX

50. Jahrami H, Saif Z, Trabelsi K, Bragazzi NL, Vitiello MV. Internal consistency and structural validity of the nomophobia questionnaire (NMP-Q) and its translations: a systematic review with meta-analysis. Heliyon 2023 Apr 14;9(4):e15464 [FREE Full text] [doi: 10.1016/j.heliyon.2023.e15464] [Medline: 37128316]
51. van Berkel N, D'Alfonso S, Kurnia Susanto R, Ferreira D, Kostakos V. AWARE-Light: a smartphone tool for experience sampling and digital phenotyping. Pers Ubiquitous Comput 2022 Nov 05;27:435-445. [doi: 10.1007/s00779-022-01697-7]
52. AWARE-Light homepage. AWARE-Light. URL: https://www.aware-light.org/ [accessed 2025-02-09]
53. Sensors - AWARE. AWARE. URL: https://awareframework.com/sensors/ [accessed 2025-02-09]
54. Vega J, Li M, Aguillera K, Goel N, Joshi E, Khandekar K, et al. Reproducible analysis pipeline for data streams: open-source software to process data collected with mobile devices. Front Digit Health 2021 Nov 18;3:769823 [FREE Full text] [doi: 10.3389/fdgth.2021.769823] [Medline: 34870271]
55. Doryab A, Chikarsel P, Liu X, Dey AK. Extraction of behavioral features from smartphone and wearable data. arXiv Preprint posted online on December 18, 2018 [FREE Full text] [doi: 10.48550/arXiv.1812.10394]
56. Canzian L, Musolesi M. Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing. 2015 Presented at: UbiComp '15; September 7-11, 2015; Osaka, Japan. [doi: 10.1145/2750858.2805845]
57. McInnes L, Healy J, Astels S. hdbscan: hierarchical density based clustering. J Open Source Softw 2017 Mar 21;2(11):205. [doi: 10.21105/joss.00205]
58. Zar JH. Spearman rank correlation. In: Encyclopedia of Biostatistics. Hoboken, NJ: John Wiley & Sons; Jul 15, 2005.
59. Thissen D, Steinberg L, Kuang D. Quick and easy implementation of the Benjamini-Hochberg procedure for controlling the false positive rate in multiple comparisons. J Educ Behav Stat 2002;27(1):77-83. [doi: 10.3102/10769986027001077]
60. Cohen J. Statistical Power Analysis for the Behavioral Sciences. New York, NY: Routledge; May 13, 2013.
61. Sharma M, Amandeep, Mathur DM, Jeenger J. Nomophobia and its relationship with depression, anxiety, and quality of life in adolescents. Ind Psychiatry J 2019;28(2):231-236 [FREE Full text] [doi: 10.4103/ipj.ipj_60_18] [Medline: 33223716]
62. Kubrusly M, de Barros Silva PG, de Vasconcelos GV, Leite ED, de Almeida Santos P, Rocha HA. Nomophobia among medical students and its association with depression, anxiety, stress and academic performance. Rev Bras Educ Med 2021 Jul;45(3):e162. [doi: 10.1590/1981-5271v45.3-20200493.ing]
63. V S, Subramanian A, Satyanarayan S, Subramanian S, S B, N B, et al. Prevalence of nomophobia and its association with stress, anxiety and depression among students. Biomedicine 2021 Jan;40(4):522-525. [doi: 10.51248/.v40i4.333]
64. Bano N, Khan MA, Asif U, Beer J, Rawass H. Effects of nomophobia on anxiety, stress and depression among Saudi medical students in Jeddah, Saudi Arabia. J Pak Med Assoc 2021 Mar;71(3):854-858 [FREE Full text] [doi: 10.47391/JPMA.983] [Medline: 34057935]
65. Kuscu TD, Gumustas F, Rodopman Arman A, Goksu M. The relationship between nomophobia and psychiatric symptoms in adolescents. Int J Psychiatry Clin Pract 2021 Mar;25(1):56-61. [doi: 10.1080/13651501.2020.1819334] [Medline: 32940094]
66. Çakmak Tolan Ö, Karahan S. The relationship between nomophobia and depression, anxiety and stress levels of university students. Int J Psychol Educ Stud 2022 Mar 06;9(1):115-129. [doi: 10.52380/ijpes.2022.9.1.559]
67. Kateb SA. The prevalence and psychological symptoms of nomophobia among university students. J Res Curr Instruct Educ Technol 2017 Jul;3(3):155-182. [doi: 10.21608/JRCIET.2017.24454]
68. Raugh IM, James SH, Gonzalez CM, Chapman HC, Cohen AS, Kirkpatrick B, et al. Geolocation as a digital phenotyping measure of negative symptoms and functional outcome. Schizophr Bull 2020 Dec 01;46(6):1596-1607 [FREE Full text] [doi: 10.1093/schbul/sbaa121] [Medline: 32851401]
69. Gong J, Huang Y, Chow PI, Fua K, Gerber MS, Teachman BA, et al. Understanding behavioral dynamics of social anxiety among college students through smartphone sensors. Inf Fusion 2019 Sep;49:57-68. [doi: 10.1016/j.inffus.2018.09.002]
70. Beekman AT, de Beurs E, van Balkom AJ, Deeg DJ, van Dyck R, van Tilburg W. Anxiety and depression in later life: co-occurrence and communality of risk factors. Am J Psychiatry 2000 Jan;157(1):89-95. [doi: 10.1176/ajp.157.1.89] [Medline: 10618018]
71. Jang C, Sung W. Beyond the privacy paradox: the moderating effect of online privacy concerns on online service use behavior. Telemat Inform 2021 Dec;65:101715. [doi: 10.1016/j.tele.2021.101715]
72. Debatin B, Lovejoy JP, Horn AK, Hughes BN. Facebook and online privacy: attitudes, behaviors, and unintended consequences. J Comput Mediat Commun 2009 Oct;15(1):83-108. [doi: 10.1111/j.1083-6101.2009.01494.x]

Abbreviations

DASS-21: Depression, Anxiety, and Stress Scale–21
EMA: ecological momentary assessment
HDBSCAN: Hierarchical Density-Based Spatial Clustering of Applications With Noise
MAE: mean absolute error
MAPE: mean absolute percentage error
MDD: major depressive disorder

XSL•FO
RenderX

NMP-Q: Nomophobia Questionnaire
PSWQ: Penn State Worry Questionnaire
QIDS: Quick Inventory of Depressive Symptomatology
RMSE: root mean square error
RRS: Rumination Response Scale
UCLA: University of California, Los Angeles

Edited by A Mavragani; submitted 19.02.24; peer-reviewed by A Hassan, DL Smith; comments to author 13.06.24; revised version received 03.07.24; accepted 02.01.25; published 19.02.25.

Please cite as:
Zhang T, Camargo A, Schmaal L, Kostakos V, D'Alfonso S
Nomophobia, Psychopathology, and Smartphone-Inferred Behaviors in Youth With Depression: Longitudinal Study
JMIR Form Res 2025;9:e57512
URL: https://formative.jmir.org/2025/1/e57512
doi:10.2196/57512
PMID:39969982

©Tianyi Zhang, Andres Camargo, Lianne Schmaal, Vassilis Kostakos, Simon D'Alfonso. Originally published in JMIR Formative Research (https://formative.jmir.org), 19.02.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Adapting the Germ Defence Web-Based Intervention to Improve Infection Prevention and Control in Care Homes: Interview Study Among Care Home Staff

Alex Hall[1,2], PhD; Johanna Aguilera-Muñoz[1,2,3], PhD; Lisa McGarrigle[1,2], PhD; Charlotte Eost-Telling[1,2,3], PhD; James Denison-Day[4], PhD; Christie Cabral[5], PhD; Merlin Willcox[6], DPhil; Chris Todd[1,2,3], PhD

[1]School of Health Sciences, University of Manchester, Manchester, United Kingdom
[2]Manchester Academic Health Science Centre, Manchester, United Kingdom
[3]National Institute for Health and Care Research Applied Research Collaboration Greater Manchester, Manchester, United Kingdom
[4]Centre for Clinical and Community Applications of Health Psychology, University of Southampton, Southampton, United Kingdom
[5]Centre for Academic Primary Care, University of Bristol, Bristol, United Kingdom
[6]Department of Primary Care and Population Sciences, University of Southampton, Southampton, United Kingdom

Corresponding Author:
Alex Hall, PhD
School of Health Sciences
University of Manchester
Jean McFarlane Building
Oxford Road
Manchester, M13 9PL
United Kingdom
Phone: 44 01613068705
Email: alex.hall@manchester.ac.uk

Abstract

Background: Infection prevention and control (IPC) is vital in care homes as it can reduce morbidity and mortality by 30%. Ensuring good IPC practice is a perennial challenge in the varied and complex context of care homes. Behavior change interventions delivered via digital technology may be effective in improving IPC among care home staff.

Objective: This study aimed to evaluate how an evidence-based, digital behavior change intervention called Germ Defence can be rapidly adapted to meet the needs of care homes.

Methods: This study applied the person-based approach, which emphasizes iterative approaches to optimizing interventions via individual user feedback. Phase 1 involved initial edits to the website by the research team to create Germ Defence for Care Homes (GDCH) version 1. Phase 2 consisted of stakeholder consultation on GDCH version 1 followed by edits to create GDCH version 2. The formal research (phases 3 and 4) involved individual think-aloud interviews with 21 staff members from management, care, and ancillary positions in 4 care homes providing real-time feedback as they worked through GDCH. Edits were made to create GDCH version 3 between phases 3 and 4. During the development of GDCH versions 2 and 3, it became clear that the intervention would need more fundamental changes beyond the pragmatic, incremental changes that would be possible within the scope of this study. Analysis was completed via a rapid, qualitative descriptive approach to develop a high-level summary of key findings from the interview data.

Results: There were mixed results about the attractiveness of GDCH and its suitability to the care home context. Participants felt that the images needed to be aligned much more closely with the meaning of adjacent text. Many participants felt that they would not have time to read a text-based website, and some suggested that more engaging content, including audio and video, may be preferable. Most participants felt that the overall concept of Germ Defence was clearly relevant to their context. Some felt that it might be a useful introduction for new staff members or a refresher for current staff, but others felt that it did not add anything to their existing IPC training. There were mixed opinions about the level of detail provided in the information offered by the site. While the goal-setting behavior change mechanism may have potential, the findings suggested that it may be unsuitable for care homes and more work is needed to refine it.

XSL•FO
RenderX

Conclusions: Much more work needs to be done to make Germ Defence more engaging, accessible, and relevant to the care home workforce. Our study highlights the challenges of rapidly adapting an existing intervention to a new context. Future research in this area will require a pragmatic methodological approach with a focus on implementation.

(JMIR Form Res 2025;9:e66706) doi:10.2196/66706

KEYWORDS

care homes; long-term care; nursing homes; infection prevention and control; behavioral intervention development; person-based approach; qualitative

Introduction

Background

Infection prevention and control (IPC) is vital in long-term care facilities as it can reduce morbidity and mortality by 30% and may reduce unplanned hospital admissions and overuse of antibiotics [1]. In England (the setting of this study), IPC is a central focus of the Care Quality Commission, the independent health and care regulator that inspects long-term care facilities [2]. International terminology on long-term care facilities varies; in this paper, we use the term *care home* to refer to a long-term care facility providing 24-hour care, either with nursing (ie, in which residents with complex needs receive care provided by registered nurses on-site) or without nursing (ie, in which residents with less complex needs receive support with personal care from nonregistered staff, with nursing and health care input provided as required by community practitioners). In England and Wales, there are approximately 278,000 care home residents aged ≥65 years (approximately three-quarters of whom are aged ≥80 years), approximately half of whom live in homes with nursing and half of whom live in homes without nursing and over two-thirds of whom experience substantial limitations in their ability to carry out day-to-day activities [3].

The COVID-19 pandemic of 2020 to 2022 brought both the importance of and challenges in ensuring good IPC to the forefront of public consciousness [4-6]. It also raised the profile of care homes, which are often marginalized and forgotten by the public and policy makers and which were hit particularly hard [7]. Many countries struggled with ensuring good IPC in care homes during the pandemic, but although international comparisons are difficult [8], the United Kingdom is perceived to have fared particularly badly. In England, the first 5 months of the pandemic saw approximately 20,000 excess deaths among care home residents that were directly attributable to COVID-19 [9]. Before the pandemic, care home residents in England aged ≥65 years had a 10-fold higher mortality than older people in private homes, but in the first wave of the pandemic, this rose to an 18-fold difference [10]. At least for the first months of the pandemic, care homes experienced substantial, widely reported challenges in accessing COVID-19 testing and in isolating residents and a lack of availability of personal protective equipment (PPE; eg, gloves, face masks, and aprons) [11,12]. Other systemic problems linked to a higher risk of COVID-19 outbreaks included a lack of statutory sick pay for staff (thereby discouraging staff to take time off work) and reliance on agency workers (who work across multiple sites) to address workforce shortages [13]. Care homes were locked down with only essential staff access, and government guidance on visiting care homes was not issued until July 2020, which was 4 months into the pandemic, and, thereafter, proceeded to change frequently. Care home managers struggled with late notice about changes (often being informed after national televised briefings), gaps regarding risks and harms to residents, a lack of clarity on how to interpret guidance coupled with uncoordinated support from local regulators, and lack of acknowledgment of implementation challenges [14]. Restrictive IPC measures in care homes have been the subject of intense debates regarding their psychosocial impacts on residents and potential breaches of mental capacity and human rights law [15,16].

COVID-19 stands as the contemporary paradigmatic case study of (failures in) IPC in care homes, but many of the challenges it highlighted long predate the pandemic. Other common viral infections in care homes include seasonal influenza (another respiratory virus) and norovirus (the main cause of gastroenteritis, which is spread through physical contact); bacterial infections such as *Salmonella* sp and group A *Streptococcus* sp; and infections caused by nonviral and nonbacterial agents, including scabies [17-19]. Care homes are particularly vulnerable to all these pathogens because of the physical proximity between people and the fact that most residents are of advanced age and living with chronic health conditions and require hands-on work with physical contact. In England, all care homes are required to have IPC plans, and all staff are required to complete IPC training [2]. A Cochrane rapid review of qualitative evidence (published between 2005 and 2020) of barriers to and facilitators of health care workers' adherence to IPC guidelines showed numerous enduring challenges that resounded during the COVID-19 pandemic, including (but not limited to) lengthy and changing guidance, increased workload in following guidance, lack of easy access to handwashing facilities, lack of training about infections and PPE use, lack of good-quality PPE, and difficulties in using PPE if patients felt frightened or stigmatized by it [20]. Some of these challenges are particularly pertinent to care homes as they function as both people's homes and places of care and are staffed by a workforce that may lack training and skills compared to those in other health care sites, including key ancillary staff (eg, cooks and cleaners) who are crucial to IPC [21]. Hand hygiene among care home staff may have been poor for many years [22], and a systematic review of evidence published before the pandemic found that poor hand hygiene was the most common cause of transmission of infections in care homes [23]. The COVID-19 pandemic reinforced the importance of PPE and behavioral measures such as handwashing to reduce transmission of infection [24] but also strikingly highlighted unintended consequences where use of PPE such as gloves may actually reduce compliance with

XSL•FO
RenderX

handwashing [25]. A recent meta-ethnography of qualitative studies of IPC in care homes found that IPC can be considered by staff to be "outside" the control of the care home both in terms of the source of infection and in actions to control it—the tension of implementing IPC in what is meant to be a homely environment appears to contribute to a sense that outbreaks are inevitable, with care home staff ambivalent about the benefits of applying IPC and perceiving a lack of ownership over IPC practice [26]. Ensuring good IPC practice is a perennial challenge in the varied and complex context of care homes. It remains of primary importance to consider how to improve IPC interventions and their implementation to help protect care home residents and staff from outbreaks of infectious diseases.

A recent systematic review of randomized controlled trials of interventions to prevent transmission of infections in care homes found only 4 trials of interventions to improve hand hygiene and IPC, with variable results [27]. Only 1 study showed a reduction in pneumonia incidence, 1 showed a reduction in influenza outbreaks, and 1 showed a reduction in incidence of influenzalike illness. Achieving high levels of adherence to infection control is challenging, and the interventions were all different; 3 contained in-person presentations, and only 1 contained an e-learning component. The review recommended developing more impactful behavior change interventions for improving IPC in care homes using behavioral science and the person-based approach.

Germ Defence: A Web-Based Behavior Change IPC Intervention

The World Health Organization outlines 8 core components of IPC programs in health care, including emphasis on multimodal strategies to change individual behavior [28]. A systematic review of IPC interventions in care homes found that those incorporating behavior change strategies via education, monitoring, and feedback were the most successful [29]. There is increasing recognition of the potential role of digital health technologies in IPC [30-33], which may be associated with their increased use to facilitate behavior change across a range of health domains, particularly via goal setting and self-management [34]. The intervention in this study is Germ Defence, a mobile-friendly website providing targeted, tailored advice on how and why people should adopt IPC behaviors, supplementing public health guidance with evidence-based and theory-based behavior change techniques [35]. It consists of introductory content about the importance of IPC followed by a series of pages arranged in parallel pathways where users can select from 2 components of interest (handwashing and reducing illness) and follow sequential pages of content streams that result in tailored advice and goal setting to improve or increase IPC behaviors.

Germ Defence was originally developed in response to the swine influenza (H1N1) and H5N1 influenza outbreaks of 2009 to 2010 in the form of an intervention then called PRIMIT (Primary Care Trial of a Website-Based Infection Control Intervention to Modify Influenza-like Illness and Respiratory Infection Transmission) to encourage handwashing delivered in 4 weekly sessions [36,37]. A subsequent large randomized controlled trial of >20,000 people living in shared households conducted across 3 winters (2011-2013) showed that the PRIMIT intervention increased handwashing and reduced respiratory tract infections [38]. The intervention was developed using the person-based approach, which involved conducting in-depth qualitative research to develop detailed understandings of how to overcome the behavioral barriers that users might encounter to engaging with the target behaviors [39]. It also used behavior change approaches to incorporate education, personalized goal setting, environmental prompting, and emotional motivation [40]. A process evaluation showed that the intervention was effective for both men and women, older and younger people, and people with both lower and higher levels of education [40]. Importantly, the biggest change in handwashing behavior occurred after the first of the 4 weekly sessions, so the intervention structure was subsequently adapted for wider dissemination to comprise 1 stand-alone session, with further optional content available immediately afterward [40]. The adapted intervention—now known as Germ Defence—was disseminated "in the wild" to help reduce seasonal colds and influenza.

In 2020, Germ Defence was rapidly adapted for the COVID-19 pandemic by a team of medical, public health and behavior change experts, and public contributors [35], with cross-sectional work highlighting its potential to improve IPC behavior [41]. It was endorsed by the chief medical officer for England as a priority via a large cluster-randomized controlled trial in primary care rolled out across all general practices in England (N=6579) [42]. At the same time, work with schoolchildren, parents, and school staff to use Germ Defence to mitigate the spread of COVID-19 in schools suggested that it is possible to adapt the intervention for use in this institutional context [43,44]. The sense of collective responsibility promoted by the intervention was seen to be persuasive in adopting the suggested behaviors, although lack of physical space and resources within schools may be barriers to adoption [43].

Study Aim

IPC is a perennial challenge in care homes. Behavior change interventions delivered via digital technology may be effective in improving IPC. The existing evidence base for Germ Defence, its positioning as a national priority response to COVID-19, and its potential for adaptation to institutional contexts highlighted that consideration of adapting it to suit care homes was merited. The aim of this study was to evaluate whether and how Germ Defence might be adapted to meet the needs of care home staff.

Methods

Design

This study followed Medical Research Council guidance on intervention development by drawing on the intervention development aspects of the person-based approach, which emphasizes iterative approaches to optimizing interventions [39,45,46]. It involved modification of the existing Germ Defence for Schools website [44] across 4 phases (Figure 1). We chose the school version as our starting point because of its potential greater initial applicability to an institutional context than the original version intended for use only in private

XSL•FO
RenderX

residences. After its initial general introduction pages, the school version contains 2 pathways of advice that users can choose to work through, one focusing on how to protect themselves in their own homes and one focusing on how to protect themselves in school. We intended to mimic this structure and change the organizational pathway from schools to care homes. Given the limited scope of our project, we focused our attention on the general introduction pages and the organizational pathway only.

Phase 1 involved preliminary research team edits to the website to create Germ Defence for Care Homes (GDCH) version 1. Phase 2 consisted of stakeholder consultation followed by further edits to create GDCH version 2. Phase 3 was the first half of the formal research, involving an initial round of qualitative fieldwork with care home staff followed by edits to create GDCH version 3. Finally, phase 4 was a second round of qualitative fieldwork exploring care home staff perceptions on GDCH version 3.

Figure 1. Four phases of the study.

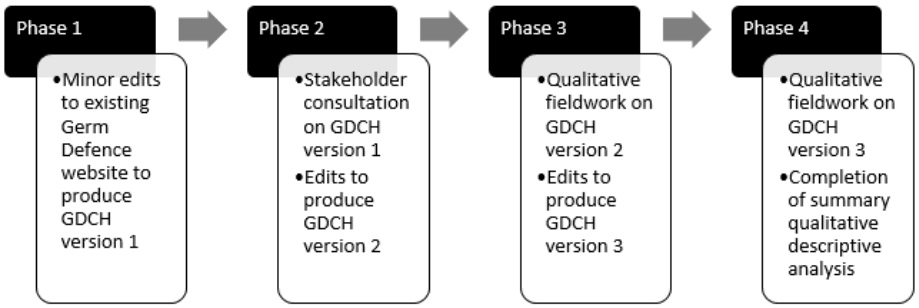

Preliminary Work: Phases 1 and 2

In phase 1, we made some initial small changes to the existing website to ensure that it referred to care homes rather than schools and to viruses more generally rather than to SARS-CoV-2 specifically. This became GDCH version 1. In phase 2, we held brief individual consultations with 8 staff members from 3 care homes (n=3, 38% managers; n=3, 38% carers; and n=2, 25% nurses) to gauge their initial responses to GDCH version 1. The changes they suggested were incorporated into GDCH version 2, which was then used in the qualitative fieldwork. Details of key changes are presented in the Results section.

Formal Research: Phases 3 and 4

Recruitment

The formal research involved 2 phases of qualitative fieldwork with iterative intervention development. We conducted fieldwork with care home staff in Greater Manchester. We aimed to purposively sample between 4 and 6 care homes representing a range of characteristics known to be related to infection rates (large and small size, modern purpose-built and retrofitted houses, with and without nursing care, not for profit and for profit, and inspection quality rating) [9]. Within each care home, we aimed to recruit 5 to 6 staff members purposively sampled to represent a wide range of roles, including managers, carers, and ancillary staff. Therefore, the total estimated sample was between 20 and 36 staff members, which aligns with previous work on qualitative sampling [47]. Care homes were recruited via a range of existing networks and contacts in the region, including the National Institute for Health and Care Research Applied Research Collaboration Greater Manchester Care Home Research Collaboration and the Enabling Research in Care Homes network. Within each care home, we provided promotional materials about the study for display and sought guidance from managers as to suitable staff members to invite.

Data Collection

The person-based approach to intervention development involves using “think-aloud” interviews to understand users’ contexts and views on every aspect of the intervention and modifying the intervention accordingly [46]. Think-aloud interviews have roots in cognitive psychology and are commonly used to collect real-time feedback and information about people’s thought processes as they engage with a particular phenomenon of interest [48]. All staff members in phases 3 and 4 participated in individual think-aloud interviews. Participants were shown the contemporaneous version of the GDCH website and were asked to work through the intervention in real time. The think-aloud element drew on an interview guide that included neutral prompts and follow-up probes to encourage participants to elaborate on their actions and verbalizations in the moment (eg, “Can you tell me a bit more about that?”). This approach offered participants the opportunity to express immediate reactions to all aspects of intervention content and functionality as they used them, page by page. Details of key changes after each phase are presented in the Results section.

All interviews were conducted individually face-to-face on care home premises, lasting 28 to 54 minutes. They were audio recorded and transcribed verbatim. We also collected demographic data on participants’ age, ethnicity, role, and length of time in the post. Data collection was undertaken by one researcher (JA-M) in May 2023 to June 2023. This researcher had a background in qualitative research within organizational psychology and, before data collection, attended training on the person-based approach and think-aloud interviews.

Data Analysis

Our initial approach to data analysis was to adopt the “table of changes” analysis commonly used with the person-based approach [39,45,46]. This analysis offers a rapid but rigorous method for extracting quotes and feedback about each component of the intervention from the think-aloud interviews and considering the relative importance of possible changes to

XSL•FO
RenderX

the intervention. One researcher (JA-M) read the interview transcripts and extracted quotes from them into a Microsoft Excel (Microsoft Corp) document detailing positive and negative comments by website page or section. This document was shared with the wider research team, and we discussed possible changes and their potential importance. Following phase 3, we made changes to GDCH version 2 that we considered important yet practicable within the resources of the study. The revised GDCH version 3 was then used in phase 4 with the same intention. However, during the development of GDCH versions 2 and 3, it became apparent that the website would need more fundamental changes than any of the pragmatic, incremental changes identified via the table of changes approach, which would be beyond the limited scope of this study. Therefore, we conducted a rapid qualitative descriptive analysis [49] to present a high-level summary of the key findings from the interview data. This involved taking the data that we had arranged in our table-of-changes Microsoft Excel documents, coding them, and grouping these codes into topic summaries. One researcher (AH) developed a set of codes that were discussed and agreed upon with the team. These codes (n=28) were applied to the remaining data by another researcher (JA-M). The team agreed on the grouping of these codes into 3 higher-level topics.

Ethical Considerations

Research participants in phases 3 and 4 provided written informed consent. Ethics approval was granted by the University of Manchester Proportionate University Research Ethics Committee (November 8, 2022; reference 2022-15380-25722). Participants were informed about the study's purpose and procedures and their right to withdraw at any time without detriment. To protect participants' privacy, all data were fully anonymized. Research team members (JA-M and AH) reviewed the transcripts and removed any identifying information. Care homes were offered reimbursement of up to £25 (US $24.76) per staff member released to participate for up to an hour [50]. Payments were offered to the organizations rather than to individual staff members to contribute to backfilling of time if necessary and to avoid any perception of unfairness among nonparticipating staff.

Results

Care Homes and Participants

In total, we recruited 4 care homes. All homes were for-profit organizations, with variation in type of home, building design, size, and quality rating (Table 1).

Overall, we recruited 21 members of staff (Table 2). Participants occupied all major positions commonly found within care homes, including caring, senior caring, nursing, leadership and management, directorship, cleaning, and kitchen roles.

Table 1. Care homes in the study.

Care home number	Care home type	Building type	Size	Ownership	Inspection rating[a]
1	With nursing	Purpose built	26 beds	For profit	Good
2	Without nursing	Purpose built	40 beds	For profit	Outstanding
3	With nursing	Purpose built	31 beds	For profit	Requires improvement
4	Without nursing	Retrofitted house	17 beds	For profit	Inadequate (under the previous provider; was being reinspected at the time of this study)

[a]Care homes in England are subject to regular, usually unannounced inspection by the Care Quality Commission, the sector regulator. The inspection focuses on 5 domains assessing the extent to which a home is safe, effective, caring, responsive, and well led. The inspection results in 1 of 4 overall ratings: "outstanding," "good," "requires improvement," or "inadequate." Homes with the 2 lowest ratings are likely to be reinspected sooner than homes with the 2 highest ratings.

XSL•FO
RenderX

Table 2. Participant details[a].

ID	Care home	Role	Study phase
P1	Home 1	Housekeeper	3
P2	Home 1	Senior carer	3
P3	Home 2	Health care assistant (enhanced)	3
P4	Home 2	Registered manager and director	3
P5	Home 2	Care assistant (plus domestic team lead)	3
P6	Home 2	Health and development lead	3
P7	Home 1	Senior carer	3
P8	Home 1	Deputy manager	3
P9	Home 2	Quality and compliance lead	3
P10	Home 2	Care team lead (plus acting care manager)	3
P11	Home 3	Carer and health care assistant	4
P12	Home 3	Carer and health care assistant	4
P13	Home 3	Domestic cleaner	4
P14	Home 3	Head chef	4
P15	Home 3	Team leader	4
P16	Home 3	Staff nurse (mental health)	4
P17	Home 4	Domestic assistant	4
P18	Home 4	Chef	4
P19	Home 4	Senior care assistant	4
P20	Home 4	Care assistant	4
P21	Home 4	Registered manager	4

[a]Time in the post ranged from 4 months to 21 years. The mean age was 43.38 (SD 13.27) years. In total, 90% (19/21) were female, and 67% (14/21) were of White British ethnicity.

Phases 1 and 2: Minor Edits

In phase 1, we made 2 types of minor changes to some of the wording of the website to create GDCH version 1. The first was to change references from “school” to “care home.” The second was to broaden the focus from COVID-19 to talk about infectious diseases more generally, including changing references to “COVID-19” to “respiratory viruses.” A major comment by the stakeholders consulted in phase 2 highlighted that a lot of the content in the section about mask wearing was very much dated to the COVID-19 pandemic, which may be unsuitable because permanent mask wearing was now not part of mandatory practice in care homes. When developing GDCH version 2, we amended this section to refer to using PPE more generally, which included some mention of masks but also other PPE commonly used in care homes, such as gloves and aprons, and delivering personal care (Figure 2).

In phase 1, GDCH version 1 retained the original Germ Defence behavior change component that asked participants to indicate how often they had performed certain IPC actions (eg, handwashing) per day over the previous week on an ordinal scale from 1 to 10 and then to make a plan for how frequently they intended to perform this behavior over the following week. In phase 2, some stakeholders felt that the scale from 1 to 10 would be unsuitable in the care home context as care home staff would wash their hands much more frequently. Therefore, in GDCH version 2, we removed this scale altogether. GDCH version 1 also contained a component that invited staff to think about points at which they had performed certain IPC actions in a typical working day on a 5-point scale (“almost never,” “sometimes,” “quite often,” “very often,” and “always”). We revised it to include more specific care home scenarios (eg, before coming into close contact with a resident and when donning PPE). We retained a sixth option of “does not apply to me” on certain questions (eg, those about PPE) because we were aware that some staff members, such as kitchen staff, would not perform certain activities, such as administering personal care.

XSL•FO
RenderX

Figure 2. (A) Germ Defence for Care Homes version 1 face covering and (B) Germ Defence for Care Homes version 2 personal protective equipment (PPE) sections following stakeholder feedback in phase 2.

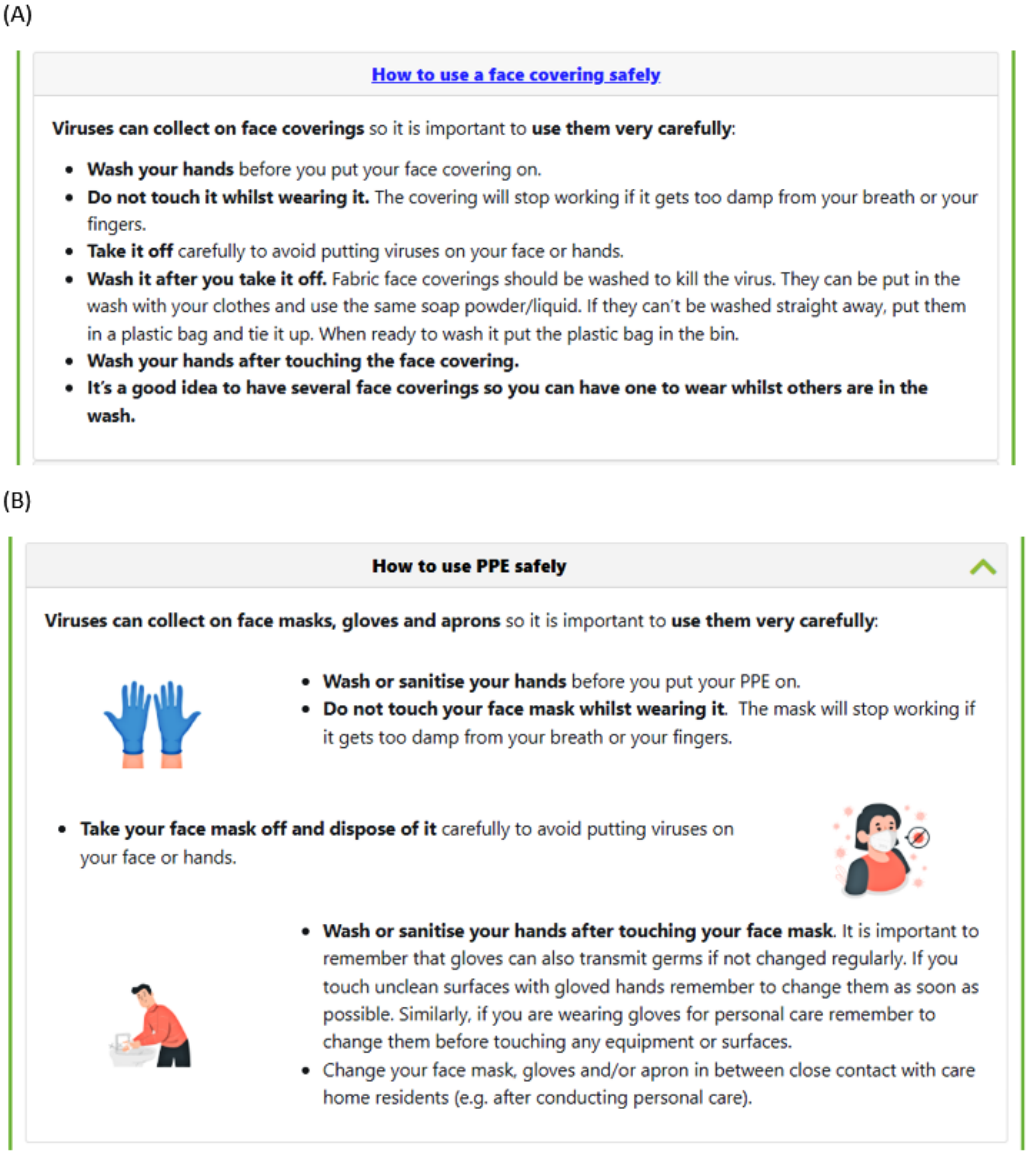

Phases 3 and 4: Key Findings and Further Intervention Modifications

Overview

Findings from phases 3 and 4 were grouped into 3 broad topic areas presented in the following sections along with key iterative changes made during the different project phases. *Attractiveness and accessibility of Germ Defence* reflects the overall appeal of the website and how accessible and engaging it was perceived to be. *Usefulness of Germ Defence* reflects the extent to which participants felt that Germ Defence might be useful in the care home context. *Individual behavior change in a complex environment* reflects findings relating to the goal-setting behavior change mechanism underpinning the intervention.

Attractiveness, Accessibility, and Usability of the Germ Defence Website

Participants in phases 3 and 4 made a wealth of suggestions for revising the images and the text of the website. One example of this was in relation to a cartoon showing how germs may spread between people (Figure 3A). In phase 3, some participants found this cartoon confusing:

> *You're looking at that thinking, I don't know what's going on. You're reading that and then you've got to work out who's who and which way it's all going.* [P6; health and development lead; phase 3]

However, others were positive about the cartoon and its accompanying text:

> *...it just shows how easily it's passed on to multiple other people. So, I think the text is enough and the picture is enough and I think they work quite well together.* [P4; registered director and manager; phase 3]

> *I think that tells the story quite well. Obviously, if it was in a care home setting...if you change the name Chloe to, like...Maureen or Beryl or Joyce, these are, sort of, the names that we have that are, sort of, older names for people...But, yeah, I mean, in terms of the*

XSL•FO
RenderX

> *image, I think it works.* [P9; quality and compliance lead; phase 3]

As an initial change following phase 3, we edited the names of some of the characters to be more reflective of care home resident age profiles (Figure 3B). Feedback in phase 4 suggested that the accompanying text was still necessary as the image alone did not clearly convey the message:

> *Once you read all this, you know what it is. But without the reading and without seeing anything, I can't tell with this picture what it is.* [P15; team leader; phase 4]

Some participants suggested incorporating an image of a germ within the cartoon showing how it might move between people. Others suggested using color coding, such as red hands on one person and red shoulders on someone else they had touched. Another suggested adding Venn diagram–like circles around the images to show overlapping relationships. These suggestions highlighted that even revising this one cartoon may require substantial work, likely conducted over multiple phases of iterative development.

Figure 3. Cartoon showing how germs may spread between people (A) in Germ Defence for Care Homes (GDCH) version 2, used in phase 3, and (B) in GDCH version 3, used in phase 4, with character names changed.

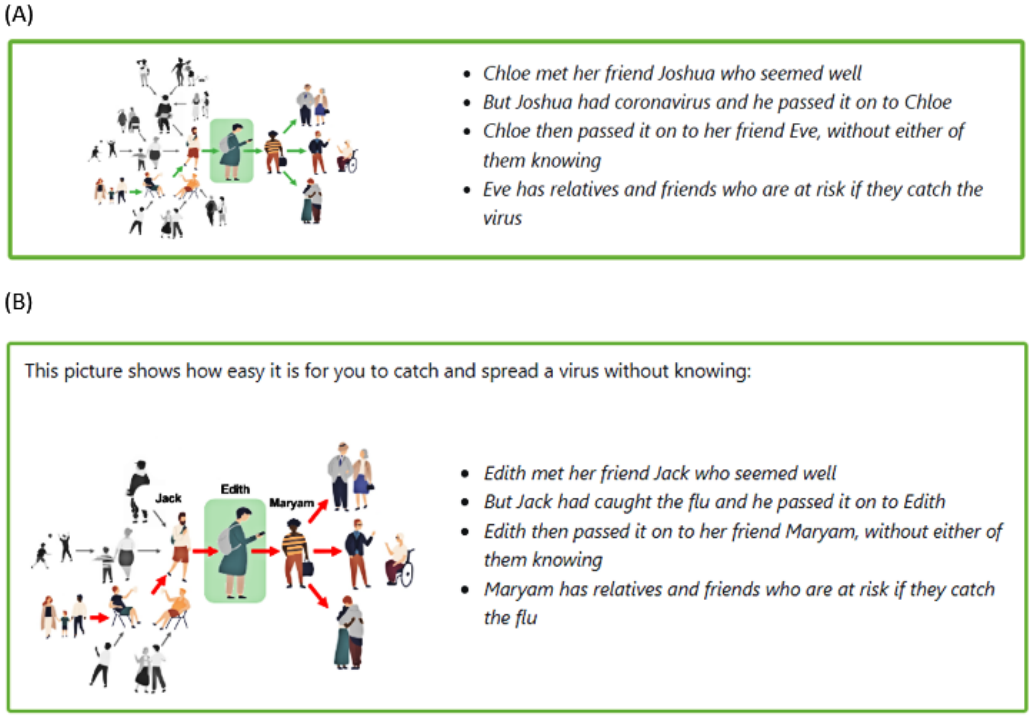

In both phase 3 and phase 4, some participants felt that they would not have time to sit and read through a text-based website. Some suggested that more engaging content, including via audio and video delivery, would be preferable (and identified some existing resources) and might address concerns surrounding English-language reading ability:

> *It's a good way of explaining it, but with it being a website...There's a video that we watched in training...I think it's a hospital setting and he's walking through, and it just shows how easy it is once you're touching it and then that's infected. I feel like that, from my experience, seeing that explains it a lot better than the picture there.* [P5; care assistant and domestic team lead; phase 3]

> *Or one thing you could add, I don't know, is you could add an audio option where people could just click the audio and listen instead. You know, some people can't understand English or they don't...they can't read so that's another thing you can add to it.* [P11; carer and health care assistant; phase 4]

Regarding engagement with the website, some participants in phases 3 and 4 expressed a preference for practical training over reading material:

> *So, I think having that training face-to-face really enables people to see the risks that they're putting themselves at if they're not following the correct procedures and risks to our residents.* [P4; registered manager and director; phase 3]

> *I think practical's better for me, personally, because academically, I don't enjoy that side of it, I'm more practical, 'cause I'm a chef, I'm creative and that side.* [P18; chef; phase 4]

These findings highlighted that any adaptation of Germ Defence may need to include a practical element of applied action.

In phase 4, some participants suggested that, overall, the images on the web pages needed to stand out from the text more and cautioned that they needed to convey the message without depending on accompanying text as the level of English-language ability among the care home workforce is highly variable:

XSL•FO
RenderX

> *...a lot of the girls, we do it on iPads, we do it on phones, things like that. So that picture's going to be minute...So I think you're best putting that picture on a page on its own, if you want it to stand out.* [P18; chef; phase 4]

> *...we have people that might not be able to understand English as well as other people but they can understand what images are. So, that's it, so this picture is equally as important as the contents within the text, yes.* [[P11; carer and health care assistant; phase 4]]

In summary, participants made a wealth of suggestions for changes that were not always mutually compatible. Incorporating these suggestions would require many phases of iterative development that were beyond the scope of our study.

Usefulness of Germ Defence Within the Care Home Context

The revised content about PPE (Figure 2) was received positively by some participants in phases 3 and 4. Some participants suggested that the content could better reflect requirements in the care home context, including how to properly dispose of PPE rather than simply saying, "dispose of it carefully." Some acknowledged that mask wearing was not mandatory practice and suggested that masks were not used at all in their care homes:

> *I think it's really good...I suppose, so mask wearing is becoming less prevalent in homes, so we wouldn't...it wouldn't be the sort of, something we'd push now, because we don't wear masks in our home.* [P9; quality and compliance lead; phase 3]

However, we avoided making any further changes to references to masks because we were aware that homes may still use them in the event of another disease outbreak. This was exemplified in phase 4:

> *I think that's, yes, giving you the right information. Not necessarily right now. I wouldn't say that you need to wear a face mask in a care home now, but if there was a virus in the care home which was really contagious as coronavirus was, then that is a good point to have there.* [P20; care assistant; phase 4]

Overall, there were mixed views on the potential utility of Germ Defence. Most participants felt that the overall concept of Germ Defence was clearly relevant to their context, and some felt that it might be useful as an introduction for new staff members or as a refresher for current staff:

> *We wouldn't use it that much...The team have enough to do. They're really busy. I could imagine that I might use that on orientation, so when I'm getting people in and first starting then I think this is a good introduction to what IPC is all about and then when they're doing the competencies out on the floor or that's when they're actually learning all the practical sides of it.* [P4; registered manager and director; phase 3]

> *We know that infection prevention and control can be difficult, sometimes, especially in a busy care home. The Germ Defence website will give you a hand, a refresher, on how to prevent and control the spread of infections. Yeah.* [P16; staff nurse; phase 4]

However, others felt that they were already doing what Germ Defence suggested and that it did not seem to add anything to their existing IPC training, which was felt to be adequate because inspections had been passed:

> *I don't think they're helpful because I'm already doing it...We've had Infection Control round twice and they've not pulled anything that I've done, so I'm doing my job well...So I wouldn't...* [P17; domestic assistant; phase 4]

> *And if there's an...like on my phone, probably this morning, I've got an update to say that one of the policies has been updated. So, then we will read through it quickly...as things change, we're kept very much up to date, and I think, I don't know, I don't think we need this additionally, really.* [P18; chef; phase 4]

There were also mixed opinions about the level of detail provided in the information offered by the site. For example, in phase 3, some participants felt that, at certain points, there needed to be comprehensive information to remind them about the transmission of germs and viruses as it is a complex subject:

> *You need a lot of information to let you know how the infection starts, and how to stop it, and control it, and stuff. If you didn't have the information, you wouldn't know.* [P1; housekeeper; phase 3]

However, at other points, some participants felt that brevity was preferable:

> *It's nice and...I like...personally I like a bullet point, so have, like, long...Because of my attention, and other people's, I think, it needs to be bullet pointed [so it] sticks.* [P3; health care assistant (enhanced); phase 3]

Participants in phase 3 expressed mixed opinions about the level at which Germ Defence was pitched. Some felt that it was appropriate as it balanced recognition of previous knowledge with a need to be accessible to care staff from a wide range of educational backgrounds:

> *I think sometimes some other professional bodies tend to dismiss carers as bottom of the rung, which we're not, it's a highly professional career now. So, I like that it recognises that we will have some knowledge around this, the infographic is clear, it sets it out very clearly how infection is spread. I think it's pitched at just the right level really because there are some people who work in care, whose educational background is not as good as others.* [P4; registered manager and director; phase 3]

However, others felt that it was too basic for experienced staff:

> *I realise that new staff are coming into an environment and they don't know things, but I think*

XSL•FO
RenderX

> *when you've got experienced staff, it sounds a bit, as I say, condescending.* [P8; deputy manager; phase 3]

As a first attempt to acknowledge this ambivalence, we added a self-rated confidence scale (1=least confident; 10=very confident) to the third page of the website in GDCH version 3 that acknowledged that staff may already be familiar with IPC measures. The page would display 1 of 2 messages depending on the self-rating selection made by the staff member: for lower confidence (1-6), the message would say that Germ Defence would give more advice and guidance about IPC and, for higher confidence (7-10), the message would say that Germ Defence would provide a handy refresher on IPC. In phase 4, feedback on this scale was that some participants felt that staff who rated themselves as very confident about IPC practice would not continue to look through the website:

> *So that's just informing us that if we want to look further we can go to that website...Personally, I don't think many people would look at it...I would say...half the people, they wouldn't bother.* [P18; chef; phase 4]

Individual Behavior Change in a Complex Environment

These findings relate to the goal-setting mechanism of the intervention, which we modified slightly following stakeholder feedback in phase 2. In phases 3 and 4, participants' feedback on the goal-setting pages suggested that, although they were easy to read and it was clear how to interact with the pages to answer the questions, it was not clear whether this might lead to any sustained change in behavior. We were initially concerned about the potential for social desirability bias in responses and that participants would simply choose "always" for every behavior. However, this did not appear to be the case as participants appeared to acknowledge shortfalls, at least to some extent:

> *[Regarding washing hands] touching something that other people have touched, like door handles, and furniture. I'd say, "very often," but sometimes we might forget, you know what I mean.* [P1; housekeeper; phase 3]

Some participants appeared to be unconvinced by the "making a plan" goal-setting activity, citing externally driven IPC requirements as having the ultimate influence over individual behavior and highlighting that noncompliance could result in disciplinary action:

> *Making a plan, yeah, I mean, essentially, we're wearing an apron and a mask because that is a requirement, so to make the plan of wearing them is what we're doing anyway because it's a requirement to do it. So if the aprons came back into being a requirement from IPC, then we wouldn't be planning on doing it, we'd have to do it, otherwise you wouldn't be coming to work, you'd have to do it. And the PPE for the personal care is a requirement, and if you don't do it you'd have to go down the conversations and supervisions and that type of thing.* [P6; health and development lead; phase 3]

This participant went on to say that the "making a plan" activity may be more beneficial for new, inexperienced colleagues rather than for experienced staff for whom IPC behavior was "embedded."

Some of the statements regarding the behavior change component specifically related to wearing face masks, which were not mandated at the time of data collection. There were examples to suggest that, if an aspect of IPC was not mandatory, it may be seen as not applicable by some participants:

> *It doesn't apply to me but obviously it is optional when you're giving personal care. But personally, I don't wear them when it's not mandatory. So, I'd say "does not apply to me."* [P5; care assistant and domestic team lead; phase 3]

There were other examples of how participants' answers might change depending on the type of activity and area of the home they were thinking about when reading the page:

> *Face mask, not all the time, these days, but sometimes if we want to get a resident's room, we are choosing, but in the kitchen, we don't use it all the time. "Quite often," maybe.* [P14; head chef; phase 4]

Therefore, it was unclear whether the questions and possible responses were sensitive enough to the breadth of activities that might arise in the care home setting. Other participants (eg, P12, carer and health care assistant) indicated that they would choose "always" for most questions about their behavior the previous week and would not make any changes to their future plan of behavior. It was unclear whether this response was influenced by social desirability bias. However, some participants appeared more reflective over the exercise, such as one manager who suggested that it made them question their behavior:

> *It makes you think. So when you have ticked your answers, if they're not "always" on both sets of the questions, it makes you think about your actions...I clicked "always," because that's what I do. But I don't do it on close contact with a resident, so that was "very often," because most times I do. Whereas on the second part, and then reading and thinking, that's maybe one of the places where I should always wash my hands, in close contact with the resident.* [P21; registered manager; phase 4]

If participants did not indicate substantial improvements to their plan for future activity (ie, if they continued to select answers indicating a low frequency of behavior), they would then see a page asking them to reconsider. Some suggested that they felt indifferent about this page—there was nothing "wrong with it," but it did not appear to be very motivating:

> *Interviewer: So, now you have this page saying "are you happy with your plan?" What do you think of this?*
>
> *Participant: I don't really think anything of it. I don't know...sanitise your hands sometimes before coming into close contact with a care home resident, why not try wash and sanitising them quite often from now on. Yes, I won't say anything's wrong with it.* [P20; care assistant; phase 4]

XSL•FO
RenderX

There were also some comments about the utility of conceiving the goal-setting exercise across a weekly time frame, how the intervention might be implemented alongside the remit of infection control leads, and whether there needed to be some output that captured the plan that a staff member had made as "evidence" of their proposed change in behavior:

> *[It would be useful if it was] downloadable or we can save it in some way to build it into the audit for us, as an evidence-base...we wouldn't do it weekly, they get so many competency checks and supervisories and everything else that this would be yet another thing we would add in. But it's certainly something we could add in as part of that auditing process that we do which is done at a minimum of twice yearly...Then if we find somebody isn't following procedure or isn't changing the PPE effectively, then we can go back and go, "look, you said you were going to do it, so do it," and just have that evidence-base there a little bit. I think but because we keep on top of it and because we've got infection leads, it's not something we would do weekly by any stretch. I think you would be hard pressed to find a home that would do it weekly.* [P4; registered manager and director; phase 3]

Therefore, overall, our findings suggest that, while the goal-setting behavior change mechanism may have potential, more work would be needed to refine it to facilitate successful individual behavior change when there is organizational-level regulation about what IPC actions should be taken by staff and how often.

Discussion

Principal Findings

This paper reports on qualitative work to evaluate whether and how an existing web-based, evidence-based behavior change IPC intervention called Germ Defence might be adapted to meet the needs of care home staff. Overall, our findings showed that adapting Germ Defence to this context is extremely challenging. Participants had mixed views about the appeal of Germ Defence and its potential suitability to the care home context and suggested that much more work would need to be done to make Germ Defence more engaging, accessible, and relevant to the care home workforce. The original Germ Defence intervention was designed for a very different population with little IPC knowledge. The elements that explained the need for IPC tended to duplicate existing training, and the goal setting for IPC actions did not appear to make sense to care home staff, whose primary concern may have been to follow mandatory IPC requirements. Following completion of our care home study, the results of the large cluster-randomized controlled trial of the COVID-19 adaption of Germ Defence in English general practice underway at the same time [42] showed no evidence that the intervention affected rates of respiratory tract infections or other health outcomes [51]. As the actual use of the intervention was below the 25% assumed in the sample size calculations, it is difficult to draw robust conclusions [51]. However, the results of this trial are important to note in the context of our study. Our findings suggest that substantial further work would be required to adapt Germ Defence to the care home context.

As qualitative descriptive approaches do not aim to provide in-depth theoretical analysis [52], it is useful to offer some brief discussion of our findings by drawing upon the widely used theoretical model of the Unified Theory of Acceptance and Use of Technology (UTAUT) [53]. The UTAUT's unit of analysis is the individual user, which is applicable for our study because our aim was to work with individual members of care home staff to try to adapt Germ Defence to their occupational context [54]. Although most commonly used quantitatively, the UTAUT has been used as a qualitative analytical and interpretive lens (eg, a recent qualitative meta-synthesis of digital interventions for antimicrobial prescribing and monitoring [55]). The UTAUT comprises 4 main constructs; our data speak mostly to the first 2 that focus on perceptions of benefit and ease of use.

Performance expectancy refers to the degree to which participants believe that using the intervention would be beneficial. Our findings within this construct appear to be mixed. Some respondents felt that Germ Defence could be potentially useful to support induction of new staff or provide refresher training, but others suggested that it may be too simplistic. The degree to which participants thought that it would offer a genuine benefit was unclear. Although we did not explicitly investigate the current IPC training provided in each care home in depth, some participants' responses suggested that more work would need to be done for Germ Defence to be seen as a useful addition to their existing training and practice.

Effort expectancy refers to the degree of ease of use of the intervention. Again, our findings within this construct were mixed, with some participants talking positively about the simple layout but others finding it confusing. Many participants suggested that the images would need substantial refinement as they did not always clearly convey the message of the text. Feedback also suggested that a text-heavy website would not be suitable because staff may be too busy to read through it all, may prefer audio or video content, or may prefer more practical training. Although the goal-setting exercises appeared to be easy to understand in principle, an investigation into their effectiveness for behavior change was beyond the scope of this study. We did not have the resources to develop multimedia content (eg, video or audio options) to assess whether it would be engaging enough to result in repeated, realistic practice of the desired behaviors.

Although the reasons for this low take-up are unclear, we note that Germ Defence was initially developed approximately 15 years before this work, and in the context of care homes, it is unclear from our study whether further context-specific adaptations would be sufficient to deliver engaging content that meets contemporary expectations. For instance, the potential of virtual reality (VR) and augmented reality to deliver engaging and realistic training in technical medical skills has long been recognized, and there is increasing recognition of their potential to support training in nonclinical skills [56-58]. Recent work exploring the development and use of VR and augmented reality hand hygiene training with care home staff has found that staff were extremely positive about immersive VR delivery,

XSL•FO
RenderX

perceiving it to be more engaging and enjoyable than previous training and valuing its inherent repeated practice, task realism, and feedback inherent to the intervention [59]. However, there are some challenges with this mode of delivery (eg, some side effects of nausea from using VR that render it unsuitable for all and up-front cost implications) [59].

The other 2 constructs of the UTAUT refer to external factors influencing implementation of the intervention. The focus of our study meant that we had less data relevant to these constructs. *Social influence* captures the degree to which participants perceive that important others believe that they should use the intervention. Although our interviews did not explicitly focus on this construct, some participants suggested that external directives regarding IPC from authoritative bodies would be a stronger driver of behavior than individual goal setting. However, several studies suggest that adherence to external directives is suboptimal [21,22,26]. Therefore, Germ Defence could be designed to improve adherence to IPC rules and guidance for care homes and explain when certain IPC measures are useful even if they are not mandatory (eg, explaining why mask wearing reduces transmission of respiratory infections when a staff member or resident has a cough). Finally, *facilitating conditions* highlights the degree to which participants feel that their organization's infrastructure would support the intervention's use. Although these kinds of implementation questions were not addressed directly in our study, some participants suggested that Germ Defence would need to be seen as compatible with their routine use of smartphones and tablets. Some managers suggested that Germ Defence could be used alongside IPC audits, suggesting that there is potential support from care homes' administrative infrastructure. These issues warrant further consideration.

Strengths and Limitations

The main strength of this study is its diverse sample of organizations and participants. We recruited care homes across the spectrum of quality rating according to the sector regulator and homes providing both residential only and nursing care. We did not observe any strong differences in views across types of homes, suggesting that our findings are likely to apply across the care home sector. Within the homes, we recruited participants working in a wide range of roles, including senior management, direct care provision, and ancillary positions, and with a wide range of experience, from a few months to >2 decades. Some of our findings highlight possible differences in views from staff in different roles (eg, the questionable relevance of some of the current content to ancillary staff and managers' views on the potential of the intervention to be used to demonstrate compliance), which would warrant further consideration in intervention development. Therefore, we believe that our findings are likely to have a high degree of transferability to the care home sector more widely. In addition, we provided a transparent description of our study methods, including the iterative phases of the study, to promote credibility and dependability.

The main limitation of our study is its relatively rapid nature. It was conceived in response to the crisis faced in the care home sector from the COVID-19 pandemic, with the aim to consider how an existing evidence-based intervention might be adapted to meet the needs of care homes. The think-aloud interviews to capture participants' real-time reactions to using Germ Defence were not designed to elicit detailed perspectives on potential facilitators of and challenges to implementation of the intervention that may contribute to intervention design. Our data lend themselves to a descriptive summary analysis rather than to a more nuanced exploration via a more interpretive approach, such as thematic analysis [60]. Despite these limitations, our study highlights the challenges of making rapid adaptations to Germ Defence in this context. Any such work would require a more sophisticated methodological approach, drawing more comprehensively on implementation theory. This could involve a greater focus on more fundamental aspects of intervention design advocated at an earlier stage of the person-based approach methodology [46], including a detailed investigation of current IPC training and practice and more detailed, theoretically informed exploration of participants' understandings of the intervention (eg, via a greater focus on the constructs of the UTAUT or of normalization process theory [61]). Such an approach was beyond the scope of this study. However, it is also unclear whether such work would be merited. Finally, it is worth noting that the ongoing independent public inquiry into the UK government's decision-making during the pandemic includes scrutiny on the extent to which care homes were protected and highlights that, in addition to single interventions, there is an urgent need for well-prepared, coherent, strategic IPC responses to such health emergencies across health and care systems [62,63].

Conclusions

Our study highlights the challenges of rapidly adapting an existing intervention to a new context. Future research in this area would require a pragmatic methodological approach with a focus on implementation. In the context of the GDCH intervention, it is unclear whether such work is warranted.

Acknowledgments

This study is funded by the National Institute for Health and Care Research (NIHR) School for Primary Care Research program (grant reference 582) and supported by the NIHR Applied Research Collaboration Greater Manchester (NIHR200174) and by Senior Investigator awards to CT (NIHR200299 and NIHR205156). MW's salary was partly funded by the NIHR (NIHR302412). CC acknowledges support from the NIHR Health Protection Research Unit in Behavioural Science and Evaluation at the University of Bristol. The views expressed are those of the authors and not necessarily those of the National Health Service, the NIHR, the Department of Health and Social Care, or its partner organizations. The authors would like to thank all the care homes and staff members who took part in the study. They acknowledge the support of the NIHR Applied Research Collaboration Greater Manchester and the Enabling Research in Care Homes network with the recruitment of care homes. They thank Professor Lucy

XSL•FO
RenderX

Yardley, Professor Mike Moore, Professor Paul Little, Professor Alistair Hay, Dr Sarah Denham, and members of the original Germ Defence development team for their support of this project.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

AH, LM, MW, CC, and CT developed the initial study design and acquired funding. CC conducted the stakeholder consultations in phase 2. AH and JA-M recruited care homes and participants in phases 3 and 4. JA-M collected the interview data in phases 3 and 4. AH, JA-M, LM, CE-T, CC, and CT conducted analyses across each phase and agreed upon changes to the Germ Defence for Care Homes website in discussion with JD-D, who edited the intervention website coding. AH led the drafting of the manuscript with support from LM and CE-T. All authors provided comments and edits to drafts and approved the final version. CT is the principal investigator.

Conflicts of Interest

JD-D wrote the code for Germ Defence, and JD-D and MW have been involved in previous work developing and evaluating Germ Defence.

References

1. Health care without avoidable infections: the critical role of infection prevention and control. World Health Organization. 2016. URL: https://www.who.int/publications/i/item/WHO-HIS-SDS-2016.10 [accessed 2025-02-04]
2. Infection prevention and control in care homes. Care Quality Commission. 2022. URL: https://tinyurl.com/mtbyvexe [accessed 2024-02-22]
3. Older people living in care homes in 2021 and changes since 2011. Office for National Statistics. 2023. URL: https://tinyurl.com/2ee6muk2 [accessed 2024-02-22]
4. Aw J, Seng JJ, Seah SS, Low LL. COVID-19 vaccine hesitancy-a scoping review of literature in high-income countries. Vaccines 2021 Aug 13;9(8):900 [FREE Full text] [doi: 10.3390/vaccines9080900] [Medline: 34452026]
5. Greenhalgh T, Schmid MB, Czypionka T, Bassler D, Gruer L. Face masks for the public during the covid-19 crisis. BMJ 2020 Apr 09;369:m1435 [FREE Full text] [doi: 10.1136/bmj.m1435] [Medline: 32273267]
6. Martin GP, Hanna E, McCartney M, Dingwall R. Science, society, and policy in the face of uncertainty: reflections on the debate around face coverings for the public during COVID-19. Critical Public Health 2020 Aug 10;30(5):501-508. [doi: 10.1080/09581596.2020.1797997]
7. Oliver D. David Oliver: let's not forget care homes when covid-19 is over. BMJ 2020 Apr 24;369:m1629. [doi: 10.1136/bmj.m1629] [Medline: 32332026]
8. Comas-Herrera A, Marczak J, Byrd W, Lorenz-Dant K, Patel D, Pharoah D. LTCcovid International living report on COVID-19 and Long-Term Care. LTCcovid, Care Policy & Evaluation Centre, London School of Economics and Political Science. 2022. URL: https://ltccovid.org/international-living-report-covid-ltc/ [accessed 2025-02-04]
9. Morciano M, Stokes J, Kontopantelis E, Hall I, Turner AJ. Excess mortality for care home residents during the first 23 weeks of the COVID-19 pandemic in England: a national cohort study. BMC Med 2021 Mar 05;19(1):71 [FREE Full text] [doi: 10.1186/s12916-021-01945-2] [Medline: 33663498]
10. Schultze A, Nightingale E, Evans D, Hulme W, Rosello A, Bates C, et al. Mortality among Care Home Residents in England during the first and second waves of the COVID-19 pandemic: an observational study of 4.3 million adults over the age of 65. Lancet Reg Health Eur 2022 Mar;14:100295 [FREE Full text] [doi: 10.1016/j.lanepe.2021.100295] [Medline: 35036983]
11. Rajan S, Comas-Herrera A, Mckee M. Did the UK government really throw a protective ring around care homes in the COVID-19 pandemic? J Long Term Care 2020 Nov 11:185-195 [FREE Full text] [doi: 10.31389/jltc.53]
12. Gordon AL, Goodman C, Achterberg W, Barker RO, Burns E, Hanratty B, et al. Commentary: COVID in care homes-challenges and dilemmas in healthcare delivery. Age Ageing 2020 Aug 24;49(5):701-705 [FREE Full text] [doi: 10.1093/ageing/afaa113] [Medline: 32402088]
13. Shallcross L, Burke D, Abbott O, Donaldson A, Hallatt G, Hayward A, et al. Factors associated with SARS-CoV-2 infection and outbreaks in long-term care facilities in England: a national cross-sectional survey. Lancet Healthy Longevity 2021 Mar;2(3):e129-e142. [doi: 10.1016/s2666-7568(20)30065-9]
14. Dixon J, Lorenz-Dant K, Stubbs E, Harrison Dening K, Mostaghim M, Casson D. Registered care home managers' experiences of responding to the national care home visiting guidance in England during the Covid-19 pandemic; a multi-method qualitative study. BMC Geriatr 2023 Apr 19;23(1):236 [FREE Full text] [doi: 10.1186/s12877-023-03935-w] [Medline: 37076808]

XSL•FO
RenderX

15. Loveday H. Infection prevention and control should be applied as a force for good: minimising restrictions in residential and nursing care. J Infect Prev 2020 Nov 14;21(6):210-211 [FREE Full text] [doi: 10.1177/1757177420977224] [Medline: 33408757]
16. Kuylen M, Wyllie A, Bhatt V, Fitton E, Michalowski S, Martin W. COVID-19 and the Mental Capacity Act in care homes: perspectives from capacity professionals. Health Soc Care Community 2022 Sep 09;30(5):e3018-e3028 [FREE Full text] [doi: 10.1111/hsc.13747] [Medline: 35138011]
17. Cassell JA, Middleton J, Nalabanda A, Lanza S, Head MG, Bostock J, et al. Scabies outbreaks in ten care homes for elderly people: a prospective study of clinical features, epidemiology, and treatment outcomes. Lancet Infect Dis 2018 Aug;18(8):894-902. [doi: 10.1016/s1473-3099(18)30347-5]
18. Utsumi M, Makimoto K, Quroshi N, Ashida N. Types of infectious outbreaks and their impact in elderly care facilities: a review of the literature. Age Ageing 2010 May 23;39(3):299-305. [doi: 10.1093/ageing/afq029] [Medline: 20332371]
19. Petrignani M, van Beek J, Borsboom G, Richardus JH, Koopmans M. Norovirus introduction routes into nursing homes and risk factors for spread: a systematic review and meta-analysis of observational studies. J Hosp Infect 2015 Mar;89(3):163-178. [doi: 10.1016/j.jhin.2014.11.015] [Medline: 25601744]
20. Houghton C, Meskell P, Delaney H, Smalle M, Glenton C, Booth A, et al. Barriers and facilitators to healthcare workers' adherence with infection prevention and control (IPC) guidelines for respiratory infectious diseases: a rapid qualitative evidence synthesis. Cochrane Database Syst Rev 2020 Apr 21;4(4):CD013582 [FREE Full text] [doi: 10.1002/14651858.CD013582] [Medline: 32315451]
21. Luijnenburg O, Samsi K, Kessler I, Norrie C, Martineau S, Manthorpe J. 'I wasn't on the front line per se, but I was part of health care': contributions and experiences of ancillary staff in care homes in England during the COVID-19 pandemic. J Health Serv Res Policy 2024 Jul 20;29(3):143-152 [FREE Full text] [doi: 10.1177/13558196241246178] [Medline: 38642016]
22. Smith A, Carusone SC, Loeb M. Hand hygiene practices of health care workers in long-term care facilities. Am J Infect Control 2008 Sep;36(7):492-494. [doi: 10.1016/j.ajic.2007.11.003] [Medline: 18786453]
23. Lee MH, Lee GA, Lee SH, Park YH. A systematic review on the causes of the transmission and control measures of outbreaks in long-term care facilities: back to basics of infection control. PLoS One 2020 Mar 10;15(3):e0229911 [FREE Full text] [doi: 10.1371/journal.pone.0229911] [Medline: 32155208]
24. NIHR: the effectiveness of PPE in reducing the transmission of COVID-19 in health and social care settings: December 2021 update, 12 December 2021. United Kingdom Government. 2021 Dec 23. URL: https://tinyurl.com/y2c9emvd [accessed 2025-02-04]
25. Sandbekken IH, Hermansen Å, Utne I, Grov EK, Løyland B. Students' observations of hand hygiene adherence in 20 nursing home wards, during the COVID-19 pandemic. BMC Infect Dis 2022 Feb 14;22(1):156 [FREE Full text] [doi: 10.1186/s12879-022-07143-6] [Medline: 35164685]
26. Daker-White G, Panagioti M, Giles S, Blakeman T, Moore V, Hall A, et al. Beyond the control of the care home: a meta-ethnography of qualitative studies of infection prevention and control in residential and nursing homes for older people. Health Expect 2022 Oct 21;25(5):2095-2106 [FREE Full text] [doi: 10.1111/hex.13349] [Medline: 34420254]
27. Willcox ML, Lavu D, Yousaf U, Dalton S, Roberts N, Plüddemann A. The effectiveness of interventions to reduce the transmission of acute respiratory infections in care homes: a systematic review. J Public Health 2024 Dec 01;46(4):551-563. [doi: 10.1093/pubmed/fdae178] [Medline: 39136085]
28. Improving infection prevention and control at the health facility: an interim practical manual. World Health Organization. 2018 Jan 3. URL: https://www.who.int/publications/i/item/WHO-HIS-SDS-2018.10 [accessed 2025-02-04]
29. Lee MH, Lee GA, Lee SH, Park YH. Effectiveness and core components of infection prevention and control programmes in long-term care facilities: a systematic review. J Hosp Infect 2019 Aug;102(4):377-393 [FREE Full text] [doi: 10.1016/j.jhin.2019.02.008] [Medline: 30794854]
30. Gray J. Infection control: beyond the horizon. J Hosp Infect 2015 Apr;89(4):237-240. [doi: 10.1016/j.jhin.2014.11.004] [Medline: 25586986]
31. Huang F, Brouqui P, Boudjema S. How does innovative technology impact nursing in infectious diseases and infection control? A scoping review. Nurs Open 2021 Sep 25;8(5):2369-2384 [FREE Full text] [doi: 10.1002/nop2.863] [Medline: 33765353]
32. Dykgraaf SH, Matenge S, Desborough J, Sturgiss E, Dut G, Roberts L, et al. Protecting nursing homes and long-term care facilities from COVID-19: a rapid review of international evidence. J Am Med Dir Assoc 2021 Oct;22(10):1969-1988 [FREE Full text] [doi: 10.1016/j.jamda.2021.07.027] [Medline: 34428466]
33. Mahmood S, Hasan K, Colder Carras M, Labrique A. Global preparedness against COVID-19: we must leverage the power of digital health. JMIR Public Health Surveill 2020 Apr 16;6(2):e18980 [FREE Full text] [doi: 10.2196/18980] [Medline: 32297868]
34. Taj F, Klein MC, van Halteren A. Digital health behavior change technology: bibliometric and scoping review of two decades of research. JMIR Mhealth Uhealth 2019 Dec 13;7(12):e13311 [FREE Full text] [doi: 10.2196/13311] [Medline: 31833836]
35. Morton K, Ainsworth B, Miller S, Rice C, Bostock J, Denison-Day J, et al. Adapting behavioral interventions for a changing public health context: a worked example of implementing a digital intervention during a global pandemic using rapid

XSL•FO
RenderX

optimisation methods. Front Public Health 2021 Apr 26;9:668197 [FREE Full text] [doi: 10.3389/fpubh.2021.668197] [Medline: 33981669]
36. Yardley L, Miller S, Teasdale E, Little P, Primit Team. Using mixed methods to design a web-based behavioural intervention to reduce transmission of colds and flu. J Health Psychol 2011 Mar;16(2):353-364. [doi: 10.1177/1359105310377538] [Medline: 20929941]
37. Yardley L, Miller S, Schlotz W, Little P. Evaluation of a web-based intervention to promote hand hygiene: exploratory randomized controlled trial. J Med Internet Res 2011 Dec 09;13(4):e107 [FREE Full text] [doi: 10.2196/jmir.1963] [Medline: 22155673]
38. Little P, Stuart B, Hobbs FD, Moore M, Barnett J, Popoola D, et al. An internet-delivered handwashing intervention to modify influenza-like illness and respiratory infection transmission (PRIMIT): a primary care randomised trial. The Lancet 2015 Oct;386(10004):1631-1639. [doi: 10.1016/s0140-6736(15)60127-1]
39. Yardley L, Morrison L, Bradbury K, Muller I. The person-based approach to intervention development: application to digital health-related behavior change interventions. J Med Internet Res 2015 Jan 30;17(1):e30 [FREE Full text] [doi: 10.2196/jmir.4055] [Medline: 25639757]
40. Ainsworth B, Steele M, Stuart B, Joseph J, Miller S, Morrison L, et al. Using an analysis of behavior change to inform effective digital intervention design: how did the PRIMIT website change hand hygiene behavior across 8993 users? Ann Behav Med 2017 Jun 1;51(3):423-431 [FREE Full text] [doi: 10.1007/s12160-016-9866-9] [Medline: 27909944]
41. Ainsworth B, Miller S, Denison-Day J, Stuart B, Groot J, Rice C, et al. Infection control behavior at home during the COVID-19 pandemic: observational study of a web-based behavioral intervention (Germ Defence). J Med Internet Res 2021 Feb 25;23(2):e22197 [FREE Full text] [doi: 10.2196/22197] [Medline: 33566791]
42. Horwood J, Chalder M, Ainsworth B, Denison-Day J, de Vocht F, Elwenspoek MM, et al. Primary care implementation of Germ Defence, a digital behaviour change intervention to improve household infection control during the COVID-19 pandemic: a structured summary of a study protocol for a randomised controlled trial. Trials 2021 Apr 09;22(1):263 [FREE Full text] [doi: 10.1186/s13063-021-05188-7] [Medline: 33836825]
43. Treneman-Evans G, Ali B, Denison-Day J, Clegg T, Yardley L, Denford S, et al. The rapid adaptation and optimisation of a digital behaviour-change intervention to reduce the spread of COVID-19 in schools. Int J Environ Res Public Health 2022 May 31;19(11):6731 [FREE Full text] [doi: 10.3390/ijerph19116731] [Medline: 35682312]
44. Germ Defence. University of Southampton. URL: https://germdefenceschools.bss.design/ [accessed 2025-02-04]
45. Bradbury K, Morton K, Band R, van Woezik A, Grist R, McManus RJ, et al. Using the person-based approach to optimise a digital intervention for the management of hypertension. PLoS One 2018;13(5):e0196868 [FREE Full text] [doi: 10.1371/journal.pone.0196868] [Medline: 29723262]
46. Yardley L, Ainsworth B, Arden-Close E, Muller I. The person-based approach to enhancing the acceptability and feasibility of interventions. Pilot Feasibility Stud 2015 Oct 26;1(1):37 [FREE Full text] [doi: 10.1186/s40814-015-0033-z] [Medline: 27965815]
47. Hennink M, Kaiser BN. Sample sizes for saturation in qualitative research: a systematic review of empirical tests. Soc Sci Med 2022 Jan;292:114523 [FREE Full text] [doi: 10.1016/j.socscimed.2021.114523] [Medline: 34785096]
48. Pan Z, Cui Y, Leighton JP, Cutumisu M. Insights into computational thinking from think‐aloud interviews: a systematic review. Appl Cogn Psychol 2022 Dec 23;37(1):71-95. [doi: 10.1002/acp.4029]
49. Sandelowski M. Whatever happened to qualitative description? Res Nurs Health 2000 Aug;23(4):334-340. [doi: 10.1002/1098-240x(200008)23:4<334::aid-nur9>3.0.co;2-g] [Medline: 10940958]
50. Payment guidance for researchers and professionals. National Institute for Health and Care Research. 2024 Dec 5. URL: https://www.nihr.ac.uk/payment-guidance-researchers-and-professionals [accessed 2025-02-04]
51. Ainsworth B, Horwood J, Walter SR, Miller S, Chalder M, De Vocht F, OpenSAFELY Collaborative, et al. Implementing Germ Defence digital behaviour change intervention via all primary care practices in England to reduce respiratory infections during the COVID-19 pandemic: an efficient cluster randomised controlled trial using the OpenSAFELY platform. Implement Sci 2023 Dec 04;18(1):67 [FREE Full text] [doi: 10.1186/s13012-023-01321-z] [Medline: 38049846]
52. Neergaard MA, Olesen F, Andersen RS, Sondergaard J. Qualitative description - the poor cousin of health research? BMC Med Res Methodol 2009 Jul 16;9:52 [FREE Full text] [doi: 10.1186/1471-2288-9-52] [Medline: 19607668]
53. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view. MIS Q 2003;27(3):425-478. [doi: 10.2307/30036540]
54. Shachak A, Kuziemsky C, Petersen C. Beyond TAM and UTAUT: future directions for HIT implementation research. J Biomed Inform 2019 Dec;100:103315 [FREE Full text] [doi: 10.1016/j.jbi.2019.103315] [Medline: 31629923]
55. Van Dort BA, Carland JE, Penm J, Ritchie A, Baysari MT. Digital interventions for antimicrobial prescribing and monitoring: a qualitative meta-synthesis of factors influencing user acceptance. J Am Med Inform Assoc 2022 Sep 12;29(10):1786-1796 [FREE Full text] [doi: 10.1093/jamia/ocac125] [Medline: 35897157]
56. Gasteiger N, van der Veer SN, Wilson P, Dowding D. How, for whom, and in which contexts or conditions augmented and virtual reality training works in upskilling health care workers: realist synthesis. JMIR Serious Games 2022 Feb 14;10(1):e31644 [FREE Full text] [doi: 10.2196/31644] [Medline: 35156931]

XSL•FO
RenderX

57. Mäkinen H, Haavisto E, Havola S, Koivisto JM. User experiences of virtual reality technologies for healthcare in learning: an integrative review. Behav Inf Technol 2020 Jul 01;41(1):1-17. [doi: 10.1080/0144929x.2020.1788162]
58. Ruthenbeck GS, Reynolds KJ. Virtual reality for medical training: the state-of-the-art. J Simul 2017 Dec 19;9(1):16-26. [doi: 10.1057/jos.2014.14]
59. Gasteiger N, van der Veer SN, Wilson P, Dowding D. Virtual reality and augmented reality smartphone applications for upskilling care home workers in hand hygiene: a realist multi-site feasibility, usability, acceptability, and efficacy study. J Am Med Inform Assoc 2023 Dec 22;31(1):45-60 [FREE Full text] [doi: 10.1093/jamia/ocad200] [Medline: 37846147]
60. Braun V, Clarke V. Thematic Analysis: A Practical Guide. Thousand Oaks, CA: SAGE Publications; 2021.
61. May C, Finch T. Implementing, embedding, and integrating practices: an outline of normalization process theory. Sociology 2009 Jun 15;43(3):535-554. [doi: 10.1177/0038038509103208]
62. What is the UK Covid-19 inquiry? UK Covid-19 Inquiry. URL: https://covid19.public-inquiry.uk/ [accessed 2024-02-22]
63. Module 1 report: the resilience and preparedness of the United Kingdom: a report by The Rt Hon the Baroness Hallett DBE Chair of the UK Covid-19 Inquiry. UK Covid-19 Inquiry. 2024. URL: https://tinyurl.com/ycx4vfnr [accessed 2025-02-04]

Abbreviations

GDCH: Germ Defence for Care Homes
IPC: infection prevention and control
PPE: personal protective equipment
PRIMIT: Primary Care Trial of a Website-Based Infection Control Intervention to Modify Influenza-like Illness and Respiratory Infection Transmission
UTAUT: Unified Theory of Acceptance and Use of Technology
VR: virtual reality

Edited by A Mavragani; submitted 23.09.24; peer-reviewed by J Dixon, I Wilson; comments to author 08.11.24; revised version received 28.11.24; accepted 29.11.24; published 19.02.25.

Please cite as:
Hall A, Aguilera-Muñoz J, McGarrigle L, Eost-Telling C, Denison-Day J, Cabral C, Willcox M, Todd C
Adapting the Germ Defence Web-Based Intervention to Improve Infection Prevention and Control in Care Homes: Interview Study Among Care Home Staff
JMIR Form Res 2025;9:e66706
URL: https://formative.jmir.org/2025/1/e66706
doi:10.2196/66706
PMID:

©Alex Hall, Johanna Aguilera-Muñoz, Lisa McGarrigle, Charlotte Eost-Telling, James Denison-Day, Christie Cabral, Merlin Willcox, Chris Todd. Originally published in JMIR Formative Research (https://formative.jmir.org), 19.02.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Designing Values Elicitation Technologies for Mental Health and Chronic Care Integration: User-Centered Design Approach

Isabel R Rooper[1,2*], BA; William W Liem[1,3*], MSW; Martha Burla[1], MPH; Jacob Gordon[4], MSW, PhD; Lara M Baez[2,5], PhD; Rachel Kornfield[2,5], PhD; Andrew B L Berry[1,2], PhD

[1]Department of Medical Social Sciences, Northwestern University Feinberg School of Medicine, Chicago, IL, United States
[2]Center for Behavioral Intervention Technologies, Northwestern University Feinberg School of Medicine, Chicago, IL, United States
[3]Institute for Sexual and Gender Minority Health and Wellbeing, Northwestern University Feinberg School of Medicine, Chicago, IL, United States
[4]College of Nursing, University of Cincinnati, Cincinnati, OH, United States
[5]Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL, United States
[*]these authors contributed equally

Corresponding Author:
Andrew B L Berry, PhD
Department of Medical Social Sciences
Northwestern University Feinberg School of Medicine
625 N Michigan Ave
Ste 2100
Chicago, IL, 60611
United States
Phone: 1 312 503 4910
Email: andrew.berry@northwestern.edu

Abstract

Background: Individuals with multiple chronic conditions (MCCs) and mental health challenges such as depression or anxiety have complex health needs and experience significant challenges with care coordination. Approaches to enhance care for patients with MCCs typically focus on eliciting patients' values to identify and align treatment priorities across patients and providers. However, these efforts are often hindered by both systems- and patient-level barriers, which are exacerbated for patients with co-occurring mental health symptoms. Technology-enabled services (TES) offer a promising avenue to facilitate values elicitation and promote patient-centered care for these patients, though TES have not yet been tailored to their unique needs.

Objective: This study aimed to identify design and implementation considerations for TES that facilitate values elicitation among patients with MCCs and depression or anxiety. We sought to understand the preferences of both clinicians and patients for TES that could bridge the gap between mental and physical health care.

Methods: Using human-centered design methods, we conducted 7 co-design workshops with 18 participants, including primary care clinicians, mental health clinicians, and patients with MCCs and depression or anxiety. Participants were introduced to TES prototypes that used various formats (eg, worksheets and artificial intelligence chatbots) to elicit and communicate patients' values. Prototypes were iteratively refined based on participant feedback. Data from these sessions were analyzed using reflexive thematic analysis to uncover themes related to service, technology, and implementation considerations.

Results: Three primary themes were identified. (1) Service considerations: TES should help patients translate elicited values into actionable treatment plans and include low-burden, flexible activities to accommodate fluctuations in their mental health symptoms. Both patients and clinicians indicated that TES could be valuable for improving appointment preparation and patient-provider communication through interpersonal skill-building. (2) Technology considerations: Patients expressed openness to TES prototypes that used artificial intelligence, particularly those that provided concise summaries of appointment priorities. Visual aids and simplified language were highlighted as essential features to support accessibility for neurodiverse patients. (3) Implementation considerations: Clinicians and patients favored situating values elicitation in mental health care settings over primary care and preferred self-guided TES that patients could complete independently before appointments.

Conclusions: Findings indicate that TES can address the unique needs of patients with MCCs and mental health challenges by facilitating values-based care. Key design considerations include ensuring TES flexibility to account for fluctuating mental health symptoms, facilitating skill-building for effective communication, and creating user-friendly technology interfaces. Future research

XSL•FO
RenderX

should explore how TES can be integrated into health care settings to enhance care coordination and support patient-centered treatment planning. By aligning TES design with patient and clinician preferences, there is potential to bridge gaps in care for this complex patient population.

(JMIR Form Res 2025;9:e68419) doi:10.2196/68419

KEYWORDS

chronic care management; anxiety; depression; values elicitation; eHealth; patient-centered care; technology-enabled services; human-centered design; multimorbidity; user-centered design; multiple chronic conditions

Introduction

Approximately 42% of adults in the United States have multiple chronic conditions (MCCs), commonly including diabetes, heart disease, and hypertension [1]. Managing MCCs is a persistent challenge for both patients and their health care providers due to the complex health care needs of these patients. Care approaches that target one condition at a time have been deemed ineffective in addressing the complexities of competing health priorities [2,3], which leaves patients susceptible to gaps in care [4]. Specifically, recommended treatments and clinical workflows for each condition may conflict, causing patients to feel overwhelmed and unsupported in managing their complex health conditions. Patients with MCCs are also burdened with tracking and sharing medical information between providers, which complicates their ability to effectively communicate their health needs and priorities across their various care team members [5]. These challenges are intensified when patients' chronic conditions include mental health challenges such as depression or anxiety. Symptoms such as amotivation, low mood, and concentration difficulties associated with depression or anxiety can impair patients' decision-making capacity [6], making it challenging for them to identify and articulate their health priorities during values elicitation. These symptoms can lead to cognitive and emotional barriers, such as limited working memory capacity that can impair patients' ability to reflect and communicate their values, which are critical steps for informed, patient-centered care planning [7] that centralizes what patients consider most important for their health and well-being [8,9]. The traditional separation between physical and mental health care providers further complicates coordinated care, exacerbating the challenges these patients face in aligning their care with their personal values [10].

Approaches such as the Patient Priorities Care model [11] have been developed to enhance care for patients with MCCs. Such approaches focus on identifying and aligning health priorities between patients and providers to better address their respective needs [12]. Common strategies include eliciting patients' values to facilitate patient-provider discussions about priorities and inform patient-centered care management. Structured communication tools have shown promise in supporting patients with MCCs in incorporating their values into their treatment plans, including technology-enabled acceptance and commitment therapy and self-efficacy–enhancing interviewing techniques [13]. Despite this progress, current interventions still fail to meet the unique requirements of patients who also experience depression or anxiety. For instance, existing interventions do not adequately account for the ways in which mental health symptoms may impact patients' capacity to engage with values elicitation activities [14]. Prior research in eliciting patients' values among primary care physicians has also revealed significant communication barriers due to patient-level constraints (eg, overlapping symptoms, capacity to engage) and the perceived irrelevance of values in health care [15]. Structural barriers, such as time constraints and competing clinical priorities, have also led to low intervention uptake among providers [16], which signals the need for values elicitation interventions that are designed for implementation.

These barriers underscore the need for efficient interventions that (1) can be implemented within clinicians' workflows, (2) are tailored to patients with MCCs and depression or anxiety, and (3) establish the relevance of values to health care. Recent work that explored the challenges faced by patients with MCCs and depression or anxiety in communicating their values to providers highlighted the need for user-friendly, flexible technology-enabled services (TES)—digital tools and platforms that enhance health care delivery by integrating technology into care—that facilitate values elicitation across both mental health and primary care settings [14]. Research is needed to further clarify patients' and providers' design preferences for values elicitation TES tailored to this population. Building on insights from past work focusing on patients' preferences and needs in using TES for values elicitation [14], this study expands the scope to incorporate providers' preferences and needs with respect to TES for values elicitation.

Human-centered design processes iteratively engage key parties (eg, patients and providers) through interactive methods such as co-design workshops to understand their goals, challenges, and motivations, and produce TES that adequately address key parties' needs [17]. Given that implementation challenges have often limited the real-world impact of TES [18], implementation considerations should also be centered throughout design processes. The Accelerated Creation-to-Sustainment (ACTS) model [18] is a framework for expediting and improving the development and implementation of digital health solutions by integrating insights from human-computer interaction, implementation science, and clinical trial methodologies. The ACTS model aggregates design considerations into 3 groups: service, technology, and implementation. "Service" encompasses the behavioral strategies facilitated by digital tools and the expected roles of providers and patients; "technology" refers to the technologies that enable service delivery; and "implementation" involves methods to integrate digital tools into clinical practice, as well as broader contextual factors. This study used a human-centered design approach, informed by the ACTS model, to define design and implementation

XSL•FO
RenderX

considerations for TES that help patients articulate their values and support providers in identifying and acting upon those values collaboratively.

Methods

Recruitment

Participants included both providers and patients. English-speaking providers with experience in primary care or mental health settings were recruited from a large medical-academic center in the Midwest United States via study advertisements and emails. Additional providers were reached through snowball sampling via personal contacts. Demographic information relevant to the foci of this study, such as role and department, was collected from providers. However, other demographic data were not gathered, as this analysis did not intend to assess demographic characteristics and their relationship to clinicians' preferences regarding values elicitation TES. Patient participants were English-speaking adults with at least 2 self-reported chronic medical conditions and a history of depression or anxiety. The study prioritized engaging individuals with current or prior lived experience balancing mental and physical health challenges in order to capture nuanced insights into how these conditions intersect with chronic illness management and values elicitation. This approach aligns with human-centered design principles, emphasizing the value of incorporating a wide range of perspectives to inform participatory design [19]. Eligible patients were identified via institutional research recruitment registries and invited to participate in the study via email. Trained study personnel then scheduled interested individuals for workshops. Twelve providers and 6 patients verbally granted informed consent and participated in workshops.

Ethical Considerations

This study was approved by the Northwestern University institutional review board (STU00212476). Informed consent was obtained from all participants prior to participation in the study. Participants were provided with detailed information about the study's purpose, procedures, potential risks, and benefits, as well as assurances of confidentiality and the voluntary nature of their involvement. Participants were compensated with a US $50 gift card for their participation. Study data were deidentified to ensure participant privacy and confidentiality. Each participant was assigned a unique identification number for analysis and reporting purposes, and all personally identifying information was removed before sharing data outside the research team. Study data were securely stored in Health Insurance Portability and Accountability Act–compliant electronic locations.

Workshops and Prototypes

Overview

Five provider workshops (workshops 1-5) and 2 patient workshops (workshops 6-7) were conducted via Zoom (Zoom Video Communications Inc) between February 2023 and March 2024 (Table 1). Researchers can identify themes and evaluate the most and least promising design directions with as few as 5-6 qualitative interviews [20]. Each workshop involved discussing and critiquing prototypes to explore specific aspects of TES design and implementation. Workshops lasted 60 minutes and included 4-6 participants (primary care providers [PCPs], mental health providers [MHPs], or patients) and 2 researchers (ABLB, JG, or WWL). At the beginning of each workshop, the researchers introduced themselves and established the workshop goals, confidentiality procedures, and prototype rationale. Prototypes shown in each workshop were tailored to address specific design questions, representing various values elicitation methods that were iteratively refined based on feedback from earlier workshops. Prototypes were intended as "thinking tools" to elicit design considerations, rather than high-fidelity prototypes to be optimized (see Table 2 for an overview of prototypes and Multimedia Appendix 1 for example prototypes) [14].

XSL•FO
RenderX

Table 1. Summary of workshop phases, objectives, activities, and prototypes used to explore clinicians' and patients' preferences for technology-enabled services for values elicitation, with prototypes tailored to specific goals in each phase.

Phase	Objectives	Activities	Prototypes	Purpose of prototypes
Phase 1 (N=4): Building buy-in for values elicitation (MHPs[a] and PCPs[b])	Explore how to communicate values elicitation rationale to patients.	• Introduced rationale and workflow for values elicitation. • Participants designed invitations to patients to engage in values elicitation.	Values bull's eye worksheet [21]	Provide an example of a structured reflection tool on values and health priorities.
Phase 2 (N=4): Facilitating values elicitation (MHPs)	Understand how to facilitate discussions about patients' values and their application in care planning.	• Reviewed prototypes for eliciting and reflecting on values. • Facilitated discussions on how to implement values elicitation in patient interactions. • Drafted messages for MHPs to share values with PCPs.	Values bull's eye worksheet	Provide an example of a structured reflection tool on values and health priorities.
			Premeeting with an MHP	Elicit reaction to completing a values exercise *before* meeting with an MHP.
			Story-sharing	Elicit reaction to completing a values exercise *while* meeting with an MHP.
Phase 3 (N=4): Shared care planning (PCPs)	Specify how patient values should be communicated to clinicians. Evaluate strategies and tools to enhance the effectiveness of shared care planning.	• Critiqued patient scenarios and prototypes. • Identified barriers to implementing shared care planning. • Brainstormed templates and workflows for sharing values with PCPs.	Values bull's eye worksheet	Provide an example of a structured reflection tool on values and health priorities.
			Inbox storyboard	Evaluate actionability of patients' values in PCP visits.
			Previsit summary	Understand whether preappointment preparation could center patient values without adding significant workload burden.
			Workflow diagram	Illustrates care coordination between primary and mental health to explore proposed integrations.
Phase 3 (N=6): Shared care planning (patients)	Explore prototype methods to address challenges in identifying and communicating values, such as prioritizing health concerns, aligning values with treatment.	• Explored patient-centered approaches for sharing values. • Discussed patient feedback on interactive prototypes.	Story-sharing	Elicit reaction to completing a values exercise *while* meeting with an MHP.
			PCP simulator	Understand what types of communication-building support would be needed to communicate values to PCPs.

[a]MHPs: mental health providers.

[b]PCPs: primary care providers.

XSL•FO
RenderX

Table 2. Values elicitation technology–enabled service (TES) prototypes were shown to clinicians and patients to elicit their preferences for TES design and implementation[a].

Workshop displayed	Prototype name and summary
W1-W3	Values bull's eye worksheet [21]: Established tool from acceptance and commitment therapy to help patients reflect on their values via a mobile interface.
W2	Inbox storyboard: Workflow prototype in which patients complete values elicitation activities preappointment and mental health providers receive an inbox message with the results.
W4	Previsit summary: Document available in patients' electronic health records listing patients' personal values and appointment-related questions.
W4	Workflow diagram: Graphic displaying workflows to facilitate collaborative care management across mental and primary health care.
W5	Premeeting with an MHP[b]: Patients review self-care behaviors and assess their alignment with their values before meeting with an MHP; results electronically transmit to MHPs.
W5-W7	Story-sharing: Patients are emailed an exercise to complete preappointment and then meet with an MHP to reflect on their values and appointment priorities.
W6-W7	PCP[c] simulator: Patients use an AI[d] chatbot, playing the role of their PCP, to practice communicating their values to a dismissive PCP and receive takeaways (eg, communication tips and appointment priorities).

[a]The prototypes were used as "thinking tools" during the workshops.

[b]MHP: mental health provider.

[c]PCP: primary care provider.

[d]AI: artificial intelligence.

Prototype Development

Building on past work exploring the design requirements to support patients with MCC and anxiety or depression to identify and share their values [14], the team identified three key principles needed to support this population: (1) Credibility: the person eliciting the patient's values has the training and experience to do so effectively and with beneficence. (2) Trustworthiness: sharing the patient's values with the PCP will only benefit—and not negatively affect—the patient's care. (3) Time worthiness: engaging in a reflection exercise to identify and communicate one's values should offer meaningful benefits for their care, given competing demands on patients' time and capacity (eg, appointments with multiple specialists).

The team generated a high-level service flow illustrating how values could be elicited and then applied in shared care planning (Figure 1). The service flow included 3 phases for conducting values elicitation with MHPs and PCPs.

Figure 1. High-level overview of values elicitation service design, which guided the formulation of the technology-enabled prototypes that were shown to clinicians and patients as part of this qualitative study. PCP: primary care provider.

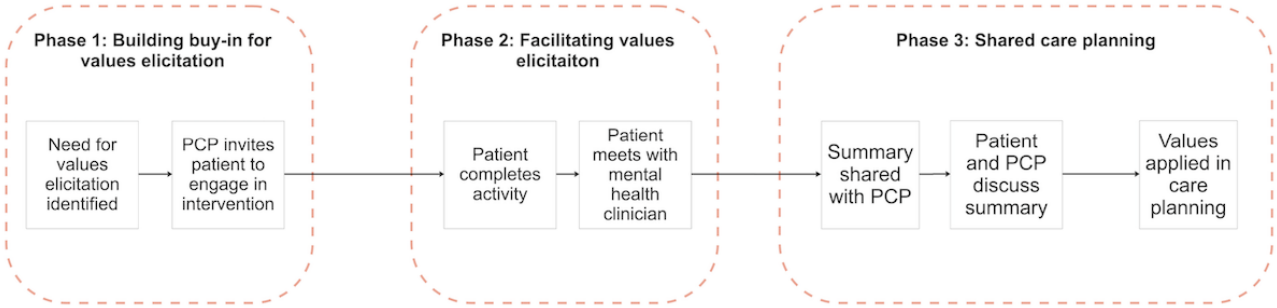

The first phase, building buy-in for values elicitation, focused on strategies to communicate the rationale for values elicitation to patients, as well as the credibility, trustworthiness, and timeworthiness of various approaches to values elicitation. Perspectives from both PCPs and MHPs were elicited to understand how to build trust and demonstrate the value of values elicitation from both physical and mental health care perspectives.

The second phase, facilitating values elicitation, aimed to refine the activities and technologies through which values could be elicited. Since previous research indicated that patients in this population often preferred engaging in values elicitation with MHPs [14], this phase centered on gathering MHPs' perspectives on how to structure and instrument values elicitation activities.

Finally, the third phase, shared care planning, focused on applying elicited patient values in shared care planning between patients and PCPs, including possible implementation barriers and facilitators. PCPs' insights were sought in this phase to identify workflow opportunities and challenges, given their central role in coordinating care across diverse clinical settings.

XSL•FO
RenderX

Prototypes were then developed to illustrate alternatives within and across these phases. For example, in phase 2, prototypes explored MHP-driven values elicitation versus patient-driven, self-guided values elicitation. By seeking confirmatory and refuting evidence from multiple participant groups in response to a range of possible design directions, this approach increased the rigor of the design research process.

Iterative Design Workshops

While workshops were being conducted, the research team met weekly to discuss and enumerate the insights garnered through each workshop, align them with the research questions and workshop objectives, and use these insights to guide prototype development for pursuant workshops. Accordingly, prototypes shown in later workshops responded to the input and critique from participants in earlier workshops.

Following the provider workshops, the research team refined 2 prototypes that were then presented to patients in 2 workshops to understand how their preferences aligned with providers' preferences. These 2 prototypes integrated insights from provider workshops and previous research with this patient population [14], which found that patients were open to values elicitation activities facilitated by MHPs, as well as self-guided values elicitation. This prior work additionally found that while patients generally know what they want to discuss with providers, they often lack the communication skills to convey these priorities effectively. Accordingly, two prototype options were created, listed in Multimedia Appendix 1: (1) a story-sharing prototype with an MHP and (2) a "PCP simulator" focused on building patients' communication skills. The artificial intelligence (AI) tool within the PCP simulator was well suited to this task, as it generated text tailored to different patient-provider scenarios, helping patients practice communicating their values and priorities with a range of provider responses.

Analyses

This analysis combined 2 complementary approaches: iterative prototype analysis to refine specific aspects of a TES, and thematic analysis to generate insights across workshops. Together, these approaches provided a robust framework for exploring values elicitation and shared care planning in diverse health care settings. Grounded in reflection-in-action [22] and research through design [23], this dual focus allowed us to rigorously address the complexities of designing and implementing values-based health care.

Iterative Refinement of Tools

Throughout this iterative process, prototypes served as "thinking tools" to elicit participants' feedback and address emerging questions about the design space. This approach aligns with research through design methodologies, where artifacts are both products of research and vehicles for generating insights. Prototype analysis was informed by reflection-in-action principles, emphasizing adaptive learning and real-time evaluation. Each phase of the workshops was analyzed iteratively to refine prototypes and activities. For example, feedback was gathered on initial concepts to identify areas where tools could better address user needs. This feedback informed the iterative development of prototypes, with adjustments made to enhance their usability and relevance. Later, barriers to integrating prototypes into existing workflows were examined. Insights from this process guided further refinements to ensure that the tools aligned with practical implementation requirements.

Thematic Analysis

Workshops were recorded via Zoom and transcribed. Transcripts were then deidentified and edited for accuracy. Finalized transcripts were qualitatively analyzed for themes using reflexive thematic analysis [24,25] to synthesize insights across workshops. This analysis involved the following steps:

1. Codebook development: Lead coauthors IRR and WL reviewed transcripts and workshop insights to create an initial codebook. Five coders then applied this codebook to a sample transcript, refining it collaboratively to resolve discrepancies.
2. Coding and memoing: Coders independently applied the refined codebook to transcripts using Dedoose (SocioCultural Research Consultants LLC) and wrote memos to capture emerging themes.
3. Collaborative synthesis: Weekly team meetings facilitated the development of themes, which were subsequently grouped using the ACTS model into design considerations for service, technology, and implementation.

This thematic analysis triangulated stakeholder perspectives by engaging participants across clinical roles and settings to validate findings and highlight role-specific nuances. Analysts' reflexivity, shaped by expertise in health education, technology design, and lived experience managing chronic conditions enhanced these findings. By combining iterative prototyping with reflexive thematic analysis, this methodology was intentionally designed to bridge theory and practice. Prototypes generated actionable insights for refining design elements, while thematic analysis captured the broader social and systemic contexts in which these tools would be implemented. This dual approach ensured that these findings were both grounded in stakeholder needs and positioned to contribute to future research and practice in values-based health care.

Results

Participants

From the primary care domain, participants included 4 physicians, 1 internal medicine physician's assistant, and 1 nurse practitioner, hereafter referred to as "PCPs." From the mental health domain, participants ("MHPs") included 2 clinical psychologists, 3 behavioral care coordinators, and 1 medical social worker. This grouping reflects the participants' primary practice contexts, with PCPs focused on managing physical health and chronic conditions, and MHPs primarily engaged in addressing mental health care needs. Full sample characteristics are shown in Tables 3 and 4.

XSL•FO
RenderX

Table 3. Sample characteristics of the clinicians who participated in workshops that explored the design and implementation of technology-enabled services for values elicitation.

ID	Domain	Role	Workshop
1	Primary care	Primary care physician	W1
2	Mental health	Patient care coordinator	W1
3	Mental health	Patient care coordinator	W1
4	Mental health	Patient care coordinator	W1
5	Primary care	Internal medicine physician assistant	W2
6	Primary care	Nurse practitioner	W2
7	Mental health	Psychiatrist	W3
8	Mental health	Medical social worker	W3
9	Primary care	Primary care physician	W4
10	Primary care	Primary care physician	W4
11	Mental health	Clinical psychologist	W5
12	Mental health	Clinical psychologist	W5

Table 4. Sample characteristics of the patients who participated in workshops that explored the design and implementation of technology-enabled services for values elicitation.

ID	Gender	Race	Hispanic	Age (years)	Chronic conditions	Workshop
13	Man	White	No	66	Cancer, glaucoma, and hyperlipidemia	W6
14	Woman	White	No	82	Arthritis, kidney or bladder problems, macular degeneration, and hypertension	W6
15	Woman	White	No	66	Hyperlipidemia, restless leg syndrome, IBS[a], and arthritis	W6
18	Woman	White	No	48	Arthritis, asthma, cancer, and Hashimoto disease	W7
16	Man	African American	No	61	Diabetes, arthritis, hypertension, and chronic pain	W7
17	Woman	White	No	38	Diabetes, hyperlipidemia, asthma, autism spectrum disorder, and ADHD[b]	W7

[a]IBS: irritable bowel syndrome.

[b]ADHD: attention-deficit/hyperactivity disorder.

Service Considerations

Service considerations reflect the unique experiences and goals of patients with MCCs and depression or anxiety. Patients and providers preferred service models for values elicitation that addressed concerns about anxiety, agenda setting, and time pressure during clinical visits. For instance, patients described how managing MCCs leads to anxiety before appointments with their PCP because they feel pressure to maximize their limited time together and had poor previous experiences with dismissive providers. To alleviate this anxiety, patients wanted support preparing for appointments. Specifically, patients said that values elicitation TES should produce a tangible outcome (eg, a list of questions) supporting preparedness and serving as memory aids. This goal reflected that their anxiety could interfere with their ability to remember information during appointments, as one participant described:

> *A tool like this [the PCP simulator] would help me, especially if I forgot to put some stuff in. It'll make sure I don't miss anything that I need to share with my primary care physician.* [P16, Patient]

Likewise, providers said that balancing MCCs makes agenda-setting challenging for some patients, so TES should be designed to help patients clarify their agenda for the appointment by reflecting on their health concerns and establishing priorities to address with providers.

Communication challenges were cited by providers (eg, due to patients' mental health symptoms) and patients, who said that values elicitation TES could offer specific language to use during appointments, allowing patients to refer to pregenerated

XSL•FO
RenderX

language instead of finding the right words in the moment, which can be challenging when experiencing anxiety. MHPs and patients said that TES that facilitate interpersonal skill-building (eg, practicing effective communication strategies) could improve preappointment anxiety and preparation, as well as provider-patient communication during appointments.

Participants reacted to TES prototypes using service models that were self-guided (ie, completed independently by a patient) versus provider-facilitated (ie, collaboratively completed by a patient and a facilitator). Both patients and providers preferred self-guided over facilitated TES due to appointment time constraints. They said that patients needed sufficient time and capacity to engage with values elicitation. Patients appreciated how self-guided TES could be completed on their own schedules.

> *Best case scenario, I would prepare about a week before a physician visit. My schedule means that may or may not happen. So, having the ability to do it at whatever time, but save that information would be incredibly important for me.* [P18, Patient]

In addition, patients described how it would be challenging for them to engage with values elicitation when their mental health symptoms are high, indicating that self-guided TES, provided well in advance of appointments, may address patients' fluctuating needs and abilities.

> *My days are very different in that my level of anxiety and depression in some days dictate what I can and can't do. So, on good days, [values elicitation] is something that I would definitely be interested in doing. But on not-so-good days, probably not so much.* [P13, Patient]

Providers agreed that, if facilitated, TES should be led by MHPs due to the relevance of values to mental health care and perceived constraints of PCPs, extrapolated in the section "Implementation Considerations." One MHP said that in addition to mental health, clinicians facilitating values elicitation should understand *physical* health, in order to best support patients managing MCCs. They recalled an instance where lack of knowledge about a patient's physical health condition was a barrier to their care:

> *The Crohn's Disease was super important to improving the depression. But I had very little insight into what her treatment was for the Crohn's Disease, [which] stunts my ability to work with her.* [P12, MHP]

Across these findings, patients and providers suggested a flexible service model that accommodates patient-directed or clinician-facilitated values elicitation and emphasized the need to support previsit communication skill-building.

Technology Considerations

The majority of providers' feedback focused on service considerations ("Service Considerations" section). In terms of technology, PCPs emphasized constraints of short appointment times and competing clinical priorities, meaning that any provider-facing TES interfaces must present concise, at-a-glance information about patients' values to be acceptable. The primary requirement is to enable PCPs to identify and address the highest priority health issues in a limited time. When viewing the previsit summary prototype, one provider commented:

> *I would say there's too much information there...Because remember, this guy has diabetes, but also is obese, and also...we need to talk about medication for the depression. So, I'm grateful I have a 40-minute appointment, but how can we address this in a 20-minute?* [P9, PCP]

PCPs also said that the interface should provide clear instructions for interpreting the "results" of values elicitation, such that those results could be applied to patient care.

Providers said that the technology should integrate into a patient's electronic health record and facilitate information sharing between providers. For instance, an MHP said that they would not have time to translate a patient's values into a message to the patient's PCP, so the technology would need to facilitate that information-sharing process.

> *I don't know that I would look at a screening tool, assess it, and then filter that information for the PCP. I wouldn't have that time. So, if it's available, I would think they can look at it or not.* [P8, MHP]

Patients were shown prototypes that differed in the inclusion of AI. Patients expressed varying levels of openness to the use of AI in values elicitation TES. Participants found the use of conversational AI for skill-building and appointment preparation (as in the "PCP simulator" prototype) helpful and were generally unconcerned with the AI delivery method, although 1 participant expressed mistrust of AI and said that they would prefer that a person facilitate the same process. Another participant described a lack of familiarity with the technology.

> *I have not had a lot of personal experience with AI. But conceptually, I'm not opposed to it. And I would not be opposed to it in this particular scenario.* [P13, Patient]

Across designs, patients expressed concerns that technologies enabling values elicitation must be accessible, particularly for neurodivergent patients and those with conditions such as depression and anxiety, which can affect cognitive processing and communication. For example, incorporating emoticons was suggested to simplify complex emotions and support users in expressing their values more easily. When asked whether they preferred the story-sharing prototype or the AI chatbot, all patient participants chose the "PCP simulator" for its skill-building and appointment preparation functionalities.

Implementation Considerations

Providers said that values elicitation should be implemented in mental health care (rather than primary care) and preferred that patients use TES previsit to avoid wasting valuable appointment time. They identified specific barriers, unique to the primary care domain, that reduce the feasibility of implementing a values elicitation TES therein.

PCPs said that patients' values were more relevant to MHPs than PCPs. Due to their training and focus on physical health,

XSL•FO
RenderX

PCPs did not feel equipped or well-positioned to elicit patients' values. Generally, PCPs also perceived values as nonactionable, meaning they did not view patients' values as information they could apply in patient care. Accordingly, they preferred not seeing their patients' values enumerated unless those values were explicitly connected to a patient's primary physical health concern.

> *To me, it's just too nebulous and complex for me to use that information...or know what it actually means. And how it relates to their current chief complaint. And how I'm going to take their value and their chief complaint and validate it in a way that I wouldn't otherwise do.* [P10, PCP]

Compared with MHPs, who defined values in alignment with values elicitation literature (ie, what patients consider most important for their health and well-being [8,9]), PCPs conceptualized values in nonspecific and variable terms, at times conflating values with goals and social history.

> *I actually have done this [values elicitation] for the last 25 years; I've always been big on social history..."where did you go to high school?" Because immediately that opens up so many conversations and goes in different directions that you start to understand people's values.* [P1, PCP]

While this comment shows the provider's commitment to building relationships with patients, their misunderstanding of values elicitation complicates its implementation. It creates a false impression that PCPs are effectively identifying patients' care priorities, when in fact, the potential of values elicitation to enhance patient care has not yet been fully realized.

Conflating values and goals similarly undermines the purpose of values elicitation, as goals are typically specific and shorter-term, while values represent the deeper, longer-term priorities guiding patients' overall care decisions. For example, a goal might be "to lower blood pressure," while the underlying value could be "to maintain independence and live an active life." By focusing on clinical goals without addressing these broader values, providers risk delivering care that aligns with immediate clinical targets but neglects what truly matters to patients. As a result, patients may feel that their overall well-being is overlooked, compromising the effectiveness of patient priorities-aligned care [26].

Regarding implementation processes, providers highlighted the many moving pieces that must be accounted for when implementing TES within existing health care systems. For example, a PCP cited workflow integration concerns and said that it is difficult to implement requisite workflows to refer patients to TES and review resulting information when TES are new and infrequently used.

> *If it's only one in 12 patients, the doctor's workflow isn't going to be used to that kind of a process. And it might be less accepted by the doctor.* [P10, PCP]

PCPs and MHPs also said that structural barriers, such as short appointments, meant that implementing a new TES during an appointment would be infeasible. However, some providers were open to reviewing patients' values before an appointment, so long as the patients completed the TES independently outside their appointment time and the presentation of those values matched providers' other requirements (eg, actionability).

Discussion

Principal Findings

This study used iterative prototyping of values elicitation TES to identify the needs and preferences of providers and patients managing multiple mental and physical health conditions. By applying the ACTS model, this study identified service, technology, and implementation considerations for TES design that help patients articulate their values and support providers in identifying and acting upon those values collaboratively. These findings largely supported the Patient Priorities Care approach [11], in that participants generally perceived identifying and aligning health priorities across clinicians and patients as useful. Findings also extended the Patient Priorities Care model to apply to patients with comorbid mental and physical health conditions. In particular, these patients' mental health symptoms and related needs have not yet been well represented in values elicitation research. This population can benefit from clarifying and communicating their values to promote patient-centered care. Per these results, design recommendations are organized in accordance with the ACTS model and summarized in Table 5.

XSL•FO
RenderX

Table 5. Design recommendations for values elicitation technology–enabled services for patients with multiple chronic conditions and depression or anxiety, organized by the ACTS model's 3 domains: service, technology, and implementation.

Design recommendation	Challenge addressed	Rationale
Service		
Incorporate clear, structured reflection prompts that elicit values actionable for health treatment planning.	Difficulty translating elicited values into actionable health treatment plans.	Structured prompts help patients articulate and connect their values to their treatment planning, making their values actionable for providers.
Offer various services, including low-burden activities when mental health symptoms are high.	Mental health symptoms pose a barrier to engaging with values elicitation.	Providing a suite of tools can address patients' diverse needs and allow flexibility when patients have low capacity for engagement.
Incorporate self-guided skill-building features to handle difficult conversations.	Difficulties with communication and preappointment anxiety.	Supports self-reflection and skill-building to enhance patient communication and manage preappointment anxiety.
When relevant, involve facilitators experienced in physical and mental health.	Lack of understanding between mental and physical health care.	Facilitators with joint expertise can holistically offer values elicitation to support patients with MCCs[a].
Technology		
Create summaries of health priorities for patients to bring to appointments.	Patients experience anxiety and sometimes forget information during appointments.	Providing tangible outputs helps patients organize their priorities, reducing anxiety and helping avoid forgotten concerns.
Engage patients with intellectual disabilities in design work.	Intellectual disabilities can pose barriers to TES[b] engagement.	Centering neurodiverse patients' perspectives will improve tool accessibility.
When used, explain the clinical intent of TES that use AI[c].	Some mistrust in the AI prototype.	Despite some mistrust, patients accepted the AI prototype due to its perceived use for skill-building.
Implementation		
Establish concordance among providers about how to conceptualize values.	Discordance poses barriers, as providers must understand a TES to implement it.	Aligning providers' perceptions of the purpose and value of a TES will help facilitate TES implementation.
Design for real-world structural barriers (eg, communication constraints).	Lack of interprofessional communication infrastructure.	Designing for present care coordination challenges helps TES add value for patients and providers.
Center implementation considerations (eg, via the ACTS model) throughout TES design processes.	Difficulty fitting new TES into existing health care systems and workflows.	Engaging with implementers (eg, clinicians and staff) yields contextual considerations that could become implementation determinants.

[a]MCCs: multiple chronic conditions.

[b]TES: technology-enabled services.

[c]AI: artificial intelligence.

Service Considerations

Results indicate that TES should assist patients in preparing for difficult conversations during appointments and provide specific language to use during consultations, both of which help alleviate their preappointment anxiety. This study identified an opportunity to enhance Patient Priorities Care approaches by addressing a critical skill gap: participants expressed a need for support not only in generating values-aligned priorities but also in developing self-advocacy skills to communicate those priorities effectively during appointments. This finding highlights that TES outputs (eg, a list of questions) alone do not guarantee effective communication; rather, participants suggested that TES should actively support patient-provider communication by strengthening patients' self-advocacy skills. This finding aligns with previous work showing that mental health symptoms can make patient-provider communication particularly difficult among patients with MCCs [27]. Yet, these findings contrast with earlier work that found that patients with MCCs and depression or anxiety preferred values elicitation facilitated by an MHP, given their credentials and the potential therapeutic benefits [14]. In this study, despite presenting an MHP-directed TES prototype, patients preferred a self-directed service model, which addressed their need to prepare for difficult conversations with providers during appointments by accounting for interpersonal dynamics with providers, including dismissiveness. This preference aligns with previous research indicating that interactions with dismissive PCPs can heighten anxiety and complicate communication about health-related values and priorities [28]. Participants may therefore have prioritized this issue over deeper therapeutic conversations about values with MHPs. Given that this patient population already

XSL•FO
RenderX

experiences unique barriers to communicating their health priorities (eg, due to their overlapping symptoms), addressing these communication challenges through skill-building is a promising area for future design work and a strategy that can empower patients to engage more actively in their appointments, leading to better care coordination and collaboration [29].

The Serious Illness Conversation Guide (SICG) provides a patient-tested framework for addressing communication challenges and self-advocacy skill gaps [30], making it highly relevant for values elicitation. By emphasizing structured communication tools, the SICG guides clinicians through sensitive discussions about patient values, offering clear language and actionable steps to ensure that these conversations are patient-centered. This structured approach could inform the design of TES by incorporating tools that allow patients to practice articulating their values before appointments. Such features align with patients' preference for self-directed models, enabling them to prepare for discussions in a low-pressure setting while retaining the SICG's focus on clarity and structure. In addition, TES inspired by the SICG could include clinician training modules to improve active listening, equipping providers to better respond to patients' expressed values. These combined strategies address key interpersonal dynamics, such as the need to counter dismissiveness, while empowering patients to effectively advocate for their priorities.

Findings illustrated the need for values elicitation service models to incorporate expertise from clinicians who understand the interplay between mental and physical health. Many MHPs lack access to PCPs' medical notes that detail patients' ongoing physical health issues, which are important for understanding how symptoms may overlap [31]. This gap places the burden on patients with MCCs to clearly articulate their complex health needs to both PCPs and MHPs, a task that is particularly challenging for those experiencing depression or anxiety. Literature suggests that patients with MCCs and mental health conditions often face cognitive and emotional barriers that make it difficult to convey the full picture of their health to providers [32]. This communication gap can hinder care coordination and increase the risk of fragmented or inadequate care. To address this, it is essential for TES to not only support values elicitation but also empower patients with tools that help them communicate their physical and mental health concerns across care teams and settings. Future research should explore these tools, particularly in promoting better care collaboration and interprofessional communication between providers in managing complex care needs.

Technology Considerations

Results from this study underscore the diverse technology preferences among patients and providers, highlighting the need for a multifaceted approach to designing TES for values elicitation. Patients and providers emphasized that the technology interface must be concise and user-friendly, accommodating the limited time available during appointments. PCPs also said that they needed clear instructions for interpreting values elicitation results. These findings align with Patient Priorities Care approaches, which emphasize the need for values to be actionable to providers. Furthermore, these findings are consistent with literature showing that technology interfaces in health care must present clear, actionable information to be effective [33], as providers often face information overload, where the volume and complexity of available information exceed their capacity to process and use it [34]. In the context of values elicitation, interfaces that provide at-a-glance summaries of values and care priorities, and straightforward instructions for responding to those, can help alleviate the challenges of managing complex patient information within constrained appointment times. This finding aligns with design recommendations for eliciting patient values in clinical conversations, which emphasize that providers must be able to easily access and interpret patient values for effective care planning [35].

Patient feedback revealed that technology design must consider the cognitive and sensory needs of users, particularly those with mental health challenges or intellectual disabilities. Previous work found that cognitively and emotionally demanding values elicitation activities can be particularly challenging for individuals with depression or anxiety, suggesting that designs should balance requisite effort against potential benefits by making activities shorter, more enjoyable, and less taxing to encourage patient engagement [14]. These findings expand upon this to include accommodating diverse communication needs among patients with intellectual disabilities, as highlighted by suggestions to include visual aids such as emoticons in the interface. Existing literature on eHealth interventions for people with intellectual disabilities calls for more intentional participatory development and iteration with end users [36], emphasizing the importance of seeing the "whole picture" of both mental and physical health when designing values elicitation TES. As individuals age with MCCs, the complications of these conditions—combined with cognitive challenges such as depression, anxiety, and intellectual disabilities—may contribute to increased difficulty in engaging with health technologies. Future design work should therefore involve patients with intellectual disabilities and those managing MCCs to cocreate inclusive technologies that address the interconnected nature of physical and mental health, promoting accessibility for this historically underrepresented group.

While patients expressed general openness to using a generative AI tool to build communication skills, responses were mixed when it came to identifying and effectively communicating their values, particularly when dealing with dismissive providers. Some participants expressed uncertainty or mistrust of AI, preferring human facilitation. This hesitation is consistent with studies indicating that while AI has potential in health care, its acceptance is often tempered by concerns about trust and transparency [37,38], emphasizing the need for AI-based tools to be carefully designed and transparently integrated into patient care in order to build trust and acceptance [37]. This finding underscores the importance of clearly communicating the rationale and benefits of values elicitation TES, as patients prefer assurance about how their values will be used [14]. In the context of AI-based tools, transparency about the elicitation and application of these values becomes even more critical.

Given these varied preferences, a suite of tools may be indicated that address relevant behavioral targets (eg, appointment

XSL•FO
RenderX

preparation and preappointment anxiety) using varied technology-based approaches (eg, skill-building, virtual worksheets, and meeting with a values elicitation facilitator). Future design work should continue mapping this array of preferences and iterate TES to accommodate them.

Implementation Considerations

The approach used in this study to implement values elicitation TES reflects a broader perspective than typical Patient Priorities Care conceptions. Specifically, the research team sought to understand the various touchpoints that patients encounter across their care journeys, given their engagement with multiple providers across different settings. By expanding the perspective of inquiry beyond single conditions or discrete interactions, this study identified new design directions, particularly in understanding how to bridge the conceptualizations of values held by both patients and providers. Establishing a common understanding of values is essential to successfully implement values elicitation TES; otherwise, challenges may arise due to knowledge gaps regarding how and why to engage patients in TES for values elicitation. Future work is therefore needed to help PCPs identify the relevance of patients' values to patients' health, and potentially, to reframe values elicitation to more clearly resonate with their practice. Indeed, PCPs in this study highlighted the importance of linking patients' values to patients' health concerns to make those values actionable. This finding aligns with patients' perspectives as well, in that patients may withhold values from providers that they perceive as irrelevant to their health care [15]. Because these varied conceptualizations of values and their perceived relevance to health care pose an implementation barrier, designing TES for values elicitation requires navigating and aligning these conceptualizations. Co-designing values elicitation educational materials with providers may help improve their perceptions of the relevance of values elicitation to this patient population, as well as prepare them to articulate its purpose to patients.

Furthermore, TES must reflect real-world challenges, such as the lack of care coordination infrastructure in many health care systems. Providers in this study said that values elicitation TES should integrate seamlessly with existing electronic health records and facilitate efficient information sharing between providers, which aligns with literature indicating the importance of interoperability and streamlined communication in improving care coordination. Yet, they also cited structural barriers that prevent interprofessional communication, which means that in the current system, patients must directly communicate their values to their PCP. Patient Priorities Care services—a common, evidence-based approach to values elicitation for patients with MCCs [11]—often assume that staff, such as medical assistants, are available and trained to support PCPs in this process. However, not all health systems are equipped with the resources or staffing models to integrate these roles effectively, limiting the feasibility of this approach in certain settings. In practice, the absence of such interprofessional communication infrastructure means that patients bear the burden of conveying their values themselves, a task made complicated when their mental health symptoms limit their ability to engage with values elicitation. The applicability of Patient Priorities Care tools to patients with MCCs and depression or anxiety could therefore be improved by accounting for these implementation challenges (eg, via self-guided instead of facilitated approaches that require staff resources), as well as this patient population's unique service requirements (eg, tool responsiveness to high mental health symptoms).

Building on these challenges, the SICG emphasizes the need for system-level changes, such as integrating structured prompts into electronic health records and developing standardized documentation templates that allow clinicians to efficiently reference and act on patient values. By embedding values elicitation into routine workflows, these tools reduce reliance on additional staffing resources, making them more feasible for implementation across diverse health care settings. Furthermore, the SICG's approach to training clinicians in communication strategies could be adapted to help providers articulate the relevance of values elicitation to patients and to foster greater patient engagement. These strategies align with findings from this study, particularly the need to link values to actionable health priorities and to ensure that TES design reflects real-world structural constraints. Incorporating these SICG-inspired interventions could enhance the scalability and effectiveness of TES by streamlining workflows, linking patient values to actionable health priorities, and addressing structural constraints across diverse care environments.

Reflections on the ACTS Model

The ACTS model provided a vital framework for generating these findings, particularly in addressing the complex needs of patients with multiple mental and physical health conditions. By focusing on service, technology, and implementation, the model helped identify the design and implementation needs of values elicitation TES. It emphasized the importance of empowering patients in their health care interactions, creating user-friendly tools that accommodate diverse needs, and addressing the structural barriers that hinder care coordination. The ACTS model's structured approach ensured that TES design is both practical and effective, facilitating seamless integration into the existing health care landscape, and ultimately, meeting the needs of this complex patient population.

Limitations

Because this work principally engaged providers from a large medical-academic center, the implementation considerations identified (eg, lack of interprofessional communication and workflow integration concerns) may not apply to other clinical contexts. Per the ACTS model, future work should continue designing TES with institution-specific implementation considerations in mind, which other designers can identify by using similar methods to this study. In addition, the providers who participated in these workshops were interested in values elicitation; their preferences may therefore not be transferable to other contexts.

Due to the small sample size and highly iterative prototypes, consensus was not achieved regarding preferred TES design features or approaches. Despite this, findings offer distinct, novel, and significant insights to inform future TES design.

Sample characteristics pose another limitation. Demographic data were not collected from providers. Future work must collect

XSL•FO
RenderX

providers' demographics to address the potential influence of misalignments between patient and clinician identities on care delivery, as these differences may affect how values are perceived and prioritized in clinical interactions. In addition, most patient participants self-identified as White and non-Hispanic. The lack of racial and ethnic diversity among patients limits the generalizability of these results and must be addressed in subsequent research. Future studies would also benefit from assessing patients' current levels of depression or anxiety to learn how symptom severity may shape patients' preferences.

Implications and Future Directions

This study found opportunities to address the unique needs of patients with MCCs and depression or anxiety by designing TES that account for patients' fluctuating mental health symptoms, build patients' communication skills, and facilitate care collaboration across mental and physical health care. Next steps from this study include conducting continued design work with providers, such as service journey mapping, to uncover additional barriers and facilitators to integrating values elicitation TES into real-world practice. Building on this engagement with patients and providers, future efforts should also expand to include stakeholders who shape systems-level decisions around adoption and implementation. Engaging hospital administrators, policymakers, and other system influencers can help address the structural barriers identified in this study, particularly in care coordination. This broader stakeholder involvement will be essential for integrating TES into health care systems effectively, ensuring that these tools are supported by the necessary resources and workflows.

Conclusions

This study offers novel insights into the design and implementation of TES for values elicitation among patients with MCCs and depression or anxiety, who have been underrepresented in values elicitation research to date. Findings demonstrate a clear patient and provider preference for self-guided TES that support communication skill-building, enabling patients to articulate and advocate for their values, particularly when engaging with dismissive providers. This study extends prior values elicitation research by identifying and addressing important barriers faced by this patient population and offers design recommendations for other TES designers to apply these insights in practice. In addition, applying the ACTS model offered a novel approach that enabled the research team to identify and centralize implementation considerations. Indeed, this study underscores the practical challenges of integrating TES into existing health care workflows, highlighting the need for tools that work within time constraints and respond to structural barriers, such as limited care coordination resources. These insights provide a foundation for future TES development aimed at improving care coordination and patient-centered care for this complex patient population, emphasizing the importance of designing tools that are both scalable and sensitive to mental health symptoms. Future research should further explore co-design processes and strategies for overcoming real-world implementation challenges.

Acknowledgments

ABLB, JG, RK, and WWL conceptualized and designed the study; ABLB, JG, and WWL facilitated the workshops; data were analyzed by IRR, WWL, MB, LMB, and JG; IRR and WWL drafted the manuscript; and MB, LMB, JG, ABLB, and RK critically reviewed and improved the manuscript with intellectual content. All authors read and approved the final manuscript. This project was supported by the National Institutes of Health (grant numbers P30AG059988, P50MH119029, T32 MH115882, and T32 MH130325).

Data Availability

Deidentified data may be made available upon reasonable request to the corresponding author, provided that appropriate data use agreements are established to ensure the continued protection of participant confidentiality.

Conflicts of Interest

RK has accepted consulting fees and honoraria from CHESS Health and Jimini Health. All other authors declare no conflicts of interest.

Multimedia Appendix 1
Workshop prototypes.
[DOCX File , 5958 KB - formative_v9i1e68419_app1.docx]

References

1. Benavidez GA, Zahnd WE, Hung P, Eberth JM. Chronic disease prevalence in the US: sociodemographic and geographic variations by zip code tabulation area. Prev Chronic Dis 2024;21:E14 [FREE Full text] [doi: 10.5888/pcd21.230267] [Medline: 38426538]
2. Institute of Medicine. Crossing the Quality Chasm: A New Health System for the 21st Century. Washington, DC: National Academies Press; 2001.

XSL•FO
RenderX

3. Redelmeier DA, Tan SH, Booth GL. The treatment of unrelated disorders in patients with chronic medical diseases. N Engl J Med 1998;338(21):1516-1520. [doi: 10.1056/NEJM199805213382106] [Medline: 9593791]
4. Albreht T, Dyakova M, Schellevis FG, Van den Broucke S. Many diseases, one model of care? J Comorb 2016;6(1):12-20 [FREE Full text] [doi: 10.15256/joc.2016.6.73] [Medline: 29090167]
5. Murphy E, Doyle J, Hannigan C, Smith S, Kuiper J, Jacobs A, et al. Perceptions and use of technology to support older adults with multimorbidity. Stud Health Technol Inform 2017;242:160-167. [Medline: 28873794]
6. Hindmarch T, Hotopf M, Owen GS. Depression and decision-making capacity for treatment or research: a systematic review. BMC Med Ethics 2013;14:54 [FREE Full text] [doi: 10.1186/1472-6939-14-54] [Medline: 24330745]
7. Moran TP. Anxiety and working memory capacity: a meta-analysis and narrative review. Psychol Bull 2016;142(8):831-864. [doi: 10.1037/bul0000051] [Medline: 26963369]
8. Friedman B, Hendry DG. Value Sensitive Design: Shaping Technology With Moral Imagination. Cambridge, MA: MIT Press; 2019.
9. Lim CY, Berry ABL, Hirsch T, Hartzler AL, Wagner EH, Ludman EJ, et al. Understanding what is most important to individuals with multiple chronic conditions: a qualitative study of patients' perspectives. J Gen Intern Med 2017;32(12):1278-1284 [FREE Full text] [doi: 10.1007/s11606-017-4154-3] [Medline: 28849368]
10. Ee C, Lake J, Firth J, Hargraves F, de Manincor M, Meade T, et al. An integrative collaborative care model for people with mental illness and physical comorbidities. Int J Ment Health Syst 2020;14(1):83 [FREE Full text] [doi: 10.1186/s13033-020-00410-6] [Medline: 33292354]
11. Tinetti ME, Naik AD, Dindo L, Costello DM, Esterson J, Geda M, et al. Association of patient priorities-aligned decision-making with patient outcomes and ambulatory health care burden among older adults with multiple chronic conditions: a nonrandomized clinical trial. JAMA Intern Med 2019;179(12):1688-1697 [FREE Full text] [doi: 10.1001/jamainternmed.2019.4235] [Medline: 31589281]
12. Ongwere T, Cantor G, Clawson J, Shih P, Connelly K. Design and care for discordant chronic comorbidities: a comparison of healthcare providers' perspectives. : ACM; 2021 Presented at: PervasiveHealth '20: 14th EAI International Conference on Pervasive Computing Technologies for Healthcare; 2020 May 18-20; Atlanta, GA p. 133-145. [doi: 10.1145/3421937.3422013]
13. Mosler F, Packer K, Jerome L, Bird V. Structured communication methods for mental health consultations in primary care: a scoping review. BMC Prim Care 2023;24(1):175 [FREE Full text] [doi: 10.1186/s12875-023-02129-y] [Medline: 37661251]
14. Liem WW, Lattie EG, Taple BJ, Stamatis CA, Gordon J, Kornfield R, et al. Improving collaborative management of multiple mental and physical health conditions: a qualitative inquiry into designing technology-enabled services for eliciting patients' values. Proc ACM Hum Comput Interact 2024;8(CSCW2):1-30 [FREE Full text] [doi: 10.1145/3687000] [Medline: 39822336]
15. Lim CY, Berry ABL, Hirsch T, Hartzler A, Wagner EH, Ludman E, et al. 'It just seems outside my health': how patients with chronic conditions perceive communication boundaries with providers. : ACM; 2016 Presented at: DIS '16: Proceedings of the 2016 ACM Conference on Designing Interactive Systems; 2016 June 4-8; Brisbane, QLD, Australia p. 1172-1184. [doi: 10.1145/2901790.2901866]
16. Overbeck G, Davidsen AS, Kousgaard MB. Enablers and barriers to implementing collaborative care for anxiety and depression: a systematic qualitative review. Implement Sci 2016;11(1):165 [FREE Full text] [doi: 10.1186/s13012-016-0519-y] [Medline: 28031028]
17. Ratwani RM, Fairbanks RJ, Hettinger AZ, Benda NC. Electronic health record usability: analysis of the user-centered design processes of eleven electronic health record vendors. J Am Med Inform Assoc 2015;22(6):1179-1182. [doi: 10.1093/jamia/ocv050] [Medline: 26049532]
18. Mohr DC, Lyon AR, Lattie EG, Reddy M, Schueller SM. Accelerating digital mental health research from early design and creation to successful implementation and sustainment. J Med Internet Res 2017;19(5):e153 [FREE Full text] [doi: 10.2196/jmir.7725] [Medline: 28490417]
19. Maguire M. Methods to support human-centred design. Int J Hum Comput Stud 2001;55(4):587-634. [doi: 10.1006/ijhc.2001.0503]
20. Guest G, Namey E, Chen M. A simple method to assess and report thematic saturation in qualitative research. PLoS One 2020;15(5):e0232076 [FREE Full text] [doi: 10.1371/journal.pone.0232076] [Medline: 32369511]
21. Lundgren T, Luoma JB, Dahl J, Strosahl K, Melin L. The bull's-eye values survey: a psychometric evaluation. Cogn Behav Pract 2012;19(4):518-526. [doi: 10.1016/j.cbpra.2012.01.004]
22. Schön DA. Educating the Reflective Practitioner: Toward a New Design for Teaching and Learning in the Professions. Hoboken, NJ: Jossey-Bass; 1987:81.
23. Zimmerman J, Forlizzi J, Evenson S. Research through design as a method for interaction design research in HCI. : ACM; 2007 Presented at: CHI07: CHI Conference on Human Factors in Computing Systems; 2007 April 28 - May 3; San Jose, CA p. 493-502. [doi: 10.1145/1240624.1240704]
24. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006;3(2):77-101. [doi: 10.1191/1478088706qp063oa]

XSL•FO
RenderX

25. Braun V, Clarke V. Can I use TA? Should I use TA? Should I not use TA? Comparing reflexive thematic analysis and other pattern-based qualitative analytic approaches. Couns Psychother Res 2020;21(1):37-47. [doi: 10.1002/capr.12360]
26. Tinetti ME, Hashmi A, Ng H, Doyle M, Goto T, Esterson J, et al. Patient priorities-aligned care for older adults with multiple conditions: a nonrandomized controlled trial. JAMA Netw Open 2024;7(1):e2352666 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.52666] [Medline: 38261319]
27. Choi BM, Obeng-Kusi M, Axon DR. Association between patient-provider communication and self-perceived mental health in US adults with cancer: real-world evidence through medical expenditure panel survey. Diseases 2022;10(4):88 [FREE Full text] [doi: 10.3390/diseases10040088] [Medline: 36278587]
28. Hildenbrand GM, Perrault EK, Rnoh RH. Patients' perceptions of health care providers' dismissive communication. Health Promot Pract 2022;23(5):777-784. [doi: 10.1177/15248399211027540] [Medline: 34344222]
29. Menear M, Dugas M, Careau E, Chouinard M, Dogba MJ, Gagnon M, et al. Strategies for engaging patients and families in collaborative care programs for depression and anxiety disorders: a systematic review. J Affect Disord 2020;263:528-539 [FREE Full text] [doi: 10.1016/j.jad.2019.11.008] [Medline: 31744737]
30. Serious illness conversation guide. Ariadne Labs. 2023. URL: https://www.ariadnelabs.org/wp-content/uploads/2023/05/Serious-Illness-Conversation-Guide.2023-05-18.pdf [accessed 2025-01-28]
31. Torab-Miandoab A, Samad-Soltani T, Jodati A, Rezaei-Hachesu P. Interoperability of heterogeneous health information systems: a systematic literature review. BMC Med Inform Decis Mak 2023;23(1):18 [FREE Full text] [doi: 10.1186/s12911-023-02115-5] [Medline: 36694161]
32. Maizes V, Rakel D, Niemiec C. Integrative medicine and patient-centered care. Explore (NY) 2009;5(5):277-289. [doi: 10.1016/j.explore.2009.06.008] [Medline: 19733814]
33. AlQudah AA, Al-Emran M, Shaalan K. Technology acceptance in healthcare: a systematic review. Appl Sci 2021;11(22):10537. [doi: 10.3390/app112210537]
34. Hall A, Walton G. Information overload within the health care system: a literature review. Health Info Libr J 2004;21(2):102-108 [FREE Full text] [doi: 10.1111/j.1471-1842.2004.00506.x] [Medline: 15191601]
35. Berry A, Lim C, Hartzler A, Hirsch T, Ludman E, Wagner EH, et al. Creating conditions for patients' values to emerge in clinical conversations: perspectives of health care team members. : ACM; 2017 Presented at: DIS '17: Designing Interactive Systems Conference 2017; 2017 June 10-14; Edinburgh, United Kingdom p. 1165-1174 URL: https://europepmc.org/abstract/MED/28890950 [doi: 10.1145/3064663.3064669]
36. van Calis JFE, Bevelander KE, van der Cruijsen AWC, Leusink GL, Naaldenberg J. Toward inclusive approaches in the design, development, and implementation of eHealth in the intellectual disability sector: scoping review. J Med Internet Res 2023;25:e45819 [FREE Full text] [doi: 10.2196/45819] [Medline: 37252756]
37. Megaro A. Transparency in AI systems for value co-creation in healthcare. In: Visvizi A, Troisi O, Grimaldi M, editors. Big Data and Decision-Making: Applications and Uses in the Public and Private Sector. Emerald Studies in Politics and Technology. England: Emerald Publishing Limited; 2023:93-105.
38. Panigutti C, Beretta A, Fadda D, Giannotti F, Pedreschi D, Perotti A, et al. Co-design of human-centered, explainable AI for clinical decision support. ACM Trans Interact Intell Syst 2023;13(4):1-35. [doi: 10.1145/3587271]

Abbreviations

ACTS: Accelerated Creation-to-Sustainment
AI: artificial intelligence
MCCs: multiple chronic conditions
MHP: mental health provider
PCP: primary care provider
SICG: Serious Illness Conversation Guide
TES: technology-enabled service

Edited by A Mavragani; submitted 11.11.24; peer-reviewed by PS Foong, V Gejjegondanahalli Yogeshappa; comments to author 18.11.24; revised version received 06.12.24; accepted 15.12.24; published 19.02.25.

Please cite as:
Rooper IR, Liem WW, Burla M, Gordon J, Baez LM, Kornfield R, Berry ABL
Designing Values Elicitation Technologies for Mental Health and Chronic Care Integration: User-Centered Design Approach
JMIR Form Res 2025;9:e68419
URL: https://formative.jmir.org/2025/1/e68419
doi:10.2196/68419
PMID:

XSL•FO
RenderX

©Isabel R Rooper, William W Liem, Martha Burla, Jacob Gordon, Lara M Baez, Rachel Kornfield, Andrew B L Berry. Originally published in JMIR Formative Research (https://formative.jmir.org), 19.02.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Classroom-Based Intervention for Reducing Sedentary Behavior and Improving Spinal Health: Pragmatic Stepped-Wedge Feasibility Randomized Controlled Trial

Dominic Fisher[1,2], PhD; Rentia Maart[2], PhD; Lehana Thabane[3], PhD; Quinette Louw[2], PhD

[1]School of Health Professions, Faculty of Health, University of Plymouth, Plymouth, United Kingdom

[2]Division of Physiotherapy, Department of Health and Rehabilitation Sciences, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa

[3]Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, ON, Canada

Corresponding Author:
Dominic Fisher, PhD
School of Health Professions
Faculty of Health
University of Plymouth
North Road East
InterCity Place
Plymouth, PL4 6AB
United Kingdom
Phone: 44 01752587580
Email: dominic.fisher@plymouth.ac.uk

Abstract

Background: Noncommunicable diseases (NCDs) resulting from sedentary behavior (SB) are adding a further strain on the South African health system, which is already struggling to manage infectious diseases. Some countries have enabled children to reduce SB at school by substituting traditional furniture with sit-stand classroom furniture, allowing learners to interrupt prolonged bouts of sitting with standing without interrupting their school work. Alternating between sitting and standing also benefits spinal health by interrupting prolonged periods of high spinal loading, but no such intervention has been trialed in South Africa. The potential to reduce strain on the health system by reducing the incidence of NCDs and improving spinal health requires further consideration. Before embarking on a large classroom-based trial, it is essential to determine the acceptability of the intervention, its impact on teachers' practices, and the logistical and pragmatic considerations of data collection.

Objective: This study aimed to assess the feasibility of implementing a classroom-based intervention to reduce SB and improve spinal health in primary school learners, to assess the pragmatics of delivering and adherence to the intervention, and assess the pragmatics of measuring physical activity and postural dynamism data with wearable sensors.

Methods: We used a stratified, closed-cohort, randomized, 2-cluster, stepped-wedge design with a pragmatic approach. One grade 5 and grade 6 class each was recruited from contrasting socioeconomically categorized, state-funded primary schools in the Western Cape province, South Africa. Classroom furniture was substituted with sit-stand desks, and health education and movement videos (HEMVs) were shown during class time. Skin-mounted activPAL physical activity monitors were used to measure SB and postural topography and Noraxon myoMOTION inertial measurement units (IMUs) to measure spinal movement. The study was evaluated for feasibility by tracking school retention, successful delivery of the HEMVs, the use of sit-stand desks, compliance with the wearable sensors, and data accuracy. We deductively analyzed teachers' interviews and learners' focus groups using Atlas.ti 9 software. Descriptive analysis of quantitative data was performed using Microsoft Excel.

Results: Cluster 1 withdrew from the study before follow-up SB, postural topography, and spinal movements were measured. All feasibility outcomes, namely (1) classroom retention, (2) delivery of HEMVs, (3) learner and teacher acceptance and usage of sit-stand classroom furniture, (4) 100% compliance with wearing skin-mounted sensors for the duration of the intended measurement period, and (5) minimum 80% eligibility of sensor data gathered included in data analysis, were met in cluster 2. The study found that it is feasible to conduct a larger trial with minor modifications to the methodology.

XSL•FO
RenderX

Conclusions: We recommend a whole-school approach to support the intervention and a monitoring strategy to track the impact of the intervention on the classroom. Furthermore, we recommend contextualized teacher training on how sit-stand desks and HEMVs can be used as classroom management tools.

Trial Registration: Pan African Trials Registry PACTR201811799476016; https://tinyurl.com/y4upoys8

International Registered Report Identifier (IRRID): RR2-10.2196/18522

(JMIR Form Res 2025;9:e65169) doi:10.2196/65169

KEYWORDS

sedentary behavior; spinal health; classroom-based intervention; sit-stand desks; spinal; African; Africa; primary school; child; youth; randomized controlled trial; RCT; infectious disease; acceptability; data collection; teacher; classroom-based interventions; primary school learners; physical activity; closed-cohort; interview; quantitative data; wearable; wearable sensor; spine

Introduction

The prevalence of preventable, long-term noncommunicable diseases (NCDs) presents a serious strain on the South African health care system, already under pressure from the burden of infectious disease management [1]. Although the causes of NCDs are multifactorial, there is consensus that lifestyle, including sedentary behavior (SB), is a major contributor [2]. Any waking behavior with energy expenditure less than or equal to 1.5 metabolic equivalent units while sitting, lying down, or reclining is considered sedentary [1]. Although initial SB interventions were mainly focused on adults, the understanding that SB in adulthood tracks through from childhood and adolescence [3] expanded SB research to include younger populations. Interventions aimed at reducing SB in children have shown promise, with several studies demonstrating good efficacy in reducing SB by educating parents, teachers, and children about the harmful effects of SB [4]. Interventions to address SB initially focused on substituting discretionary screen time with periods of physical activity. However, understanding that the beneficial effects of physical activity, even at the prescribed World Health Organization (WHO) dose, do not undo the detrimental physiological effects of SB [5] has led to the development of interventions to prevent the accumulation of SB.

School-based SB studies have demonstrated high levels of accumulated periods of nondiscretionary sedentary time during class [6]. These high levels of SB are attributable to teachers' preference that children remain seated during class [7] and environmental barriers to interrupting SB when teachers encourage movement [8]. The conventional classroom environment is not conducive to increasing children's movement without compromising teaching and learning. As a result, classroom-based SB interventions trialed dynamic sit-stand furniture to overcome the environmental barriers posed by traditional classroom furniture.

Interrupting prolonged sitting bouts by alternating between sitting and standing may have the additional benefit of improving spinal health. Prolonged sitting has been shown to increase axial loading of the spine and increase back muscle activation, which could lead to back pain [9]. Given that spinal pain tracks to adulthood from childhood [7,10] and given the reported high levels of nondiscretionary classroom sedentary time, alternating children's postural topography may have dual long-term benefits on children's cardiometabolic and spinal health. It is proposed that alternating between sitting and standing (alternating postural topography) mitigates the effects of prolonged loading in one position by providing periods of relative rest. This approach aligns with the theory of postural dynamism, which encourages regular spinal movement as opposed to maintaining a single preferred spinal position. Altering postural topography may promote spinal health by mitigating axial load on spinal posture and encouraging regular spinal movement. The dual benefit of interrupting prolonged sitting and encouraging transitioning between sitting and standing during class to mitigate the long-term risks to cardiometabolic and spinal health holds significant public health potential in resource-limited contexts, such as South Africa.

Classroom-based interventions comprising dynamic sit-stand furniture, health education, and movement integration that have been conducted in the United States, Europe, Australia, and New Zealand have shown good efficacy in reducing SB without disturbing teaching and learning activities [4]. To the best of our knowledge, no such studies have been conducted in South Africa. Conducting a trial of a classroom-based intervention in South Africa would have to consider a range of contextual factors related to the educational system and socioeconomic inequality, not to mention logistic and pragmatic considerations. Before implementing such a trial, it is important to establish the contextual viability of a previously untested intervention. Given the likely increase in children's movement in response to a classroom-based intervention and the potential impact on teacher practice, determining the acceptability of the intervention was considered [11,12]. In addition, a pragmatic methodology to measure postural topography, physical activity, and spinal movement needed to be established before being implemented in a large trial. Furthermore, the integrity of data collected in a new environment needed to be assessed. These are all key aspects needed to establish the feasibility of conducting a large trial to assess the beneficial effects of a novel SB intervention in a South African classroom context.

The aim of this study was, thus, to assess the feasibility of implementing a classroom-based intervention to reduce classroom SB and promote spinal health in primary school learners. The study objectives were to assess the pragmatics of delivering and adherence to a classroom-based intervention and assess the pragmatics of measuring physical activity and postural dynamism data with activPAL and Noraxon myoMOTION inertial measurement units (IMUs), respectively. This

XSL•FO
RenderX

manuscript focused only on the primary feasibility outcomes. The preliminary findings of the effects will be published subsequently.

Methods

Ethical Considerations

This study was granted ethical approval by the Stellenbosch University Health Research Ethics Committee (reference number S17/08/130) and institutional permission by the Western Cape Education Department (reference number 20170525-1279). It was also registered with the Pan African Trials Registry (PACTR201811799476016; International Registered Report Identifier [IRRID] RR1-10.2196/18522).

Prospective participants who expressed interest in participating in the project were provided with study information documents and consent forms. Learners were provided with child assent and parent/guardian consent forms, as well as study information for parents and guardians, which were then collected by the research team upon completion. Before the study began, learners were informed about the intervention and measurement methods. Study data were anonymized by assigning a unique participant identifier to data obtained from participants. As compensation, participants were offered retention of the intervention materials at the conclusion of the project.

Study Design

A school-based, stratified, closed-cohort, randomized, 2-cluster, stepped-wedge design with a pragmatic approach was used for this study. Participant clusters (school classrooms) were the unit of randomization to determine the order of implementing the intervention. This study report followed the CONSORT (Consolidated Standards of Reporting Trials) extension statement for pilot and feasibility trials (Multimedia Appendix 1) [13]. The protocol for this feasibility trial has been published elsewhere [14]. The CONSORT stepped-wedge flow diagram (Table 1) for cluster randomized trials [15] shows the timing of crossover from control to intervention conditions.

Table 1. Stepped-wedge cluster RCT[a] diagram of the study sequence.

Cluster	Randomized transition from usual classroom conditions to intervention condition					
	Weeks 1-2 (April 23-May 10, 2018)	Week 3 (May 11-16, 2018)	Weeks 4-5 (April 23-May 10, 2018)	Weeks 6-14 (May 14-August 10, 2018)	Weeks 15-16 (August 13-24, 2018)	July 31-August 15, 2019
Cluster 1, Q5[b] classroom	Usual conditions	Baseline measurement and introduction of intervention	Intervention	Intervention	Withdrawal from study and qualitative interviews	Usual conditions: withdrawal from study
Cluster 2, Q3[c] classroom	Usual conditions	Usual conditions	Baseline measurement and introduction of intervention	Intervention	Follow-up measurement and qualitative interviews	Follow-up measurement and qualitative interviews

[a]RCT: randomized controlled trial.

[b]Q5: quintile 5.

[c]Q3 quintile 3.

The stepped-wedge design was used to allow the intervention to be evaluated within the bounds of the logistical constraints and context of each cluster [16]. Given the limited availability of measurement equipment, the stepped-wedge design allowed sequential baseline and follow-up measurements. Furthermore, given the evidence of the effectiveness of the mode of intervention in other contexts [4], and other classroom-based interventions, the stepped-wedged design allowed both clusters to receive the potential benefit of the intervention.

Study Setting

The study was conducted in the Western Cape province of South Africa, a region of broad language, cultural, and socioeconomic diversity. In South Africa, publicly funded (state schools) are categorized into quintiles according to socioeconomic factors, including income, literacy, and employment rates of the surrounding community [17]. Lower-quintile schools are more resource limited than upper-quintile schools and thus receive greater state funding per registered learner. In the metro central district, there are fewer lower-quintile schools, and most schools fall under the quintile 5 (Q5) category.

Sampling and Recruitment

To ensure a diverse range of contextual factors, primary schools from quintile 3 (Q3) and Q5 were selected exclusively from the central metro district of the Western Cape Education Department. The rationale was that if it were deemed feasible to conduct a future trial in the more challenging, resource-strained context, it would be feasible in less challenging, adequately resourced contexts. A downloaded list of publicly funded schools in the central metro district from the Western Cape Education Department website was delimited to school categories Q3 and Q5. School principals, stratified according to Q3 and Q5 schools, were randomly contacted by the principal investigator via telephone or email for recruitment. Afterward, the principals nominated grade 5 or 6 teachers to attend an information session.

Sample Size

A sample size calculation was not performed as per the CONSORT extension statement for feasibility studies [18]. The sample target was to include a wide range of contextual factors relevant to the feasibility outcomes. The diversity between the clusters (namely classrooms from Q3 and Q5 schools) was considered sufficient to provide a range of considerations to

XSL•FO
RenderX

assess the feasibility of the intervention and the pragmatics of data collection. The individual participants selected for measuring physiological outcomes (SB and postural dynamism) were randomly chosen from the class list provided by the class teacher. An equal number of male and female learners were sampled.

Intervention

The aim of the intervention is to improve the long-term cardiometabolic and spinal health of the public by reducing the accumulation of prolonged SB that has been shown to be pervasive in schools. Providing learners with an opportunity to transition between sitting and standing during class time helps arrest sedentary physiology and provides respite from prolonged axial spinal loading. The study intervention comprised a novel height-adjustable sit-stand desk and a playlist of 9 health education and movement videos (HEMVs). The study authors (DF and QL) contributed to the development of the intervention. The design concept was informed in part by a systematic review of the efficacy of classroom-based interventions to improve spinal health and reduce the SB of schoolchildren [4], as well as a qualitative study of educators' perceptions of learners' movement during class time conducted in the Western Cape province, South Africa [7]. The systematic review provided compelling evidence for the efficacy of classroom-based interventions that include alternative, sit-stand classroom furniture and health education to improve spinal health and SB outcomes.

Sit-Stand Desks

The research team considered the sit-stand desks available in South Africa prohibitively expensive and thus unsuitable for implementation in the study. An innovation team developed a novel, multifunctional, height-adjustable, sit-stand classroom desk called the KUZE (Multimedia Appendix 2) for this study. Learners and teachers were shown how to select the correct height for sitting and standing. All the usual classroom chairs were removed and replaced with KUZE sit-stand desks during the study period.

Health Education and Movement Videos

The innovation team comprising the authors, professional video content creators, and a teacher who advocates for using body movement in teaching mathematics developed a series of HEMVs for this study. Learners were instructed to follow the videos, which included an interactive component requiring them to solve simple arithmetic problems using corresponding body movements (Multimedia Appendix 3). The teachers were handed the HEMVs on a mobile external hard drive and, after discussions with the researchers, were given the freedom to develop their own strategies for playing the videos during class time. The videos are available to view on YouTube [14].

Control Conditions

The control conditions were the usual classroom conditions. Cluster 1 and cluster 2 classrooms had similar dimensions. The cluster 1 classroom furniture comprised metal framed, wooden, all-in-one, tandem desks that fit 2 learners abreast (Multimedia Appendix 4). The cluster 2 classroom furniture had a combination of single and double metal-framed wooden tables with accompanying plastic-molded chairs (Multimedia Appendix 4). The HEMVs were not played during control conditions.

Pragmatics of Physiological Data Collection Using Wearable Sensors

Physical Activity Monitoring: activPAL

Participants' classroom physical activity and postural topography (sitting, standing, stepping. and sit-to-stand transitioning) were measured using activPAL3 microsensors (PAL Technologies). In the past decade, activPAL sensors have been widely used to objectively measure physical activity [19-21]. Sensors were applied to participants before the start of lessons and removed after the end of lessons on physiological data collection days. The sensors were attached to the anterior right thigh with a waterproof nitrile sleeve and Opsite dressing, as prescribed by the user manual. Data logged on the sensors were downloaded into a secure file at the end of each day.

Spinal Movement: Noraxon myoMOTION Inertial Measurement Units

Postural dynamism was measured using Noraxon myoMOTION IMUs. Wearable IMUs allow for the assessment of postural dynamism in an ecologically valid setting of the classroom. The IMUs combine on-board triaxial gyroscopes, accelerometers, and magnetometers for accurate sensor orientation tracking [22]. IMUs were attached to the head with an elasticated Velcro belt and to the neck, thorax, and sacrum directly to the skin using double-sided tape.

Description and Measurement of Feasibility Outcomes

The feasibility outcomes include retention of the study, fidelity to the intervention (acceptance, and usage of the KUZE sit-stand desks, delivery of the HEMVs), and integrity of the physiological data collected (compliance with wearing the activPAL and IMU sensors). The success indicators (set a priori) and methods of measuring the 5 feasibility outcomes are outlined in Table 2.

XSL•FO
RenderX

Table 2. Summary of feasibility outcomes, success indicators, and measurement methods.

Feasibility outcome	Success indicator	Measurement method
Cluster retention	Both clusters remained in the study until follow-up measurements were obtained.	• Retention rate recorded on the project management record • Qualitative feedback obtained from teachers regarding retention
Delivery of HEMVs[a]	Teachers developed a routine of playing the HEMVs and adhered to the routine.	• Qualitative feedback obtained from teachers and learners at the exit interview
Acceptance and usage of the KUZE desk in the classroom	Learners and teachers accepted and used the KUZE as classroom furniture for the entire study period.	• Retention rate recorded on the project management record • Qualitative feedback obtained from teachers and learners regarding acceptance of the KUZE
Compliance with wearing activPAL and IMU[b] sensors	There was 100% compliance with wearing activPAL sensors and IMUs for the duration of the intended measurement period.	• Recorded on the project management record
Integrity of physical activity and postural dynamism data	Of all data captured by activPAL and IMU sensors, 80% were eligible for inclusion in the analysis.	• Recorded on the project management record

[a]HEMV: health education and movement video.

[b]IMU: inertial measurement unit.

Feasibility outcomes were measured using both qualitative and quantitative methods. For the qualitative measurement of the feasibility outcomes, individual depth interviews (IDIs) with the class teachers and focus group discussions (FGDs) with a subgroup of learner participants from each cluster were conducted at the end of the intervention period. The success criteria of the feasibility outcomes were used as the framework for the IDIs and FGDs. The quantitative methods encompassed the physiological data obtained from the activPAL and IMU sensors, as well as estimates of retention and compliance, which were monitored by the researcher.

Interpretation of Feasibility Criteria

The success criteria set a priori were:

- Continue with a large pilot/trial if all 5 success criteria are met in both clusters;
- Make minor modifications to the protocol if 3 or more criteria are met in both clusters before continuing with a pilot/trial; or
- Make significant protocol modifications if 2 or less criteria are met in both clusters [18].

Data Analysis

Qualitative Analysis

A deductive analysis of data from IDIs with teachers and FGDs with learners was conducted to determine the feasibility outcomes and success indicators. The qualitative feedback sessions were recorded and transcribed in full and then analyzed using Atlas.ti 9, a computer-assisted qualitative data analysis software program that facilitates the organization, coding, analysis, and visualization of data [23]. Learner and teacher responses were grouped into corresponding themes related to classroom retention, intervention delivery, acceptance, and usage of the KUZE. At the end of the intervention, 2 FGDs per cluster were conducted. The focus group comprised both individuals who provided physiological data and those who were only exposed to the intervention but did not provide physiological data. Verbatim representative quotes of participant responses were provided.

Quantitative Analysis

Categorical data were analyzed descriptively and presented as percentages. The number of participant classes (clusters) that remained in the study for the duration of the study was documented and presented as a percentage of the total participant classes that enrolled in the study. The proportion of data used in the analysis compared to data collected was presented as a percentage. All analyses were performed using Microsoft Excel.

Results

Demographic Characteristics

The CONSORT flow diagram of the study is illustrated in Figure 1. The baseline demographic characteristics of participants were similar across clusters. The cluster 1 (Q5) school comprised a grade 6 class, while cluster 2 (Q3) was a grade 5 class. Thus, the ages of the learners in this study were 10-11 years. An equal number of male and female learners made up the study sample.

XSL•FO
RenderX

Figure 1. CONSORT flow diagram of the feasibility RCT. CONSORT: Consolidated Standards of Reporting Trials; IMU: inertial measurement unit; Q3: quintile 3; Q5: quintile 5; Qual: qualitative; RCT: randomized controlled trial.

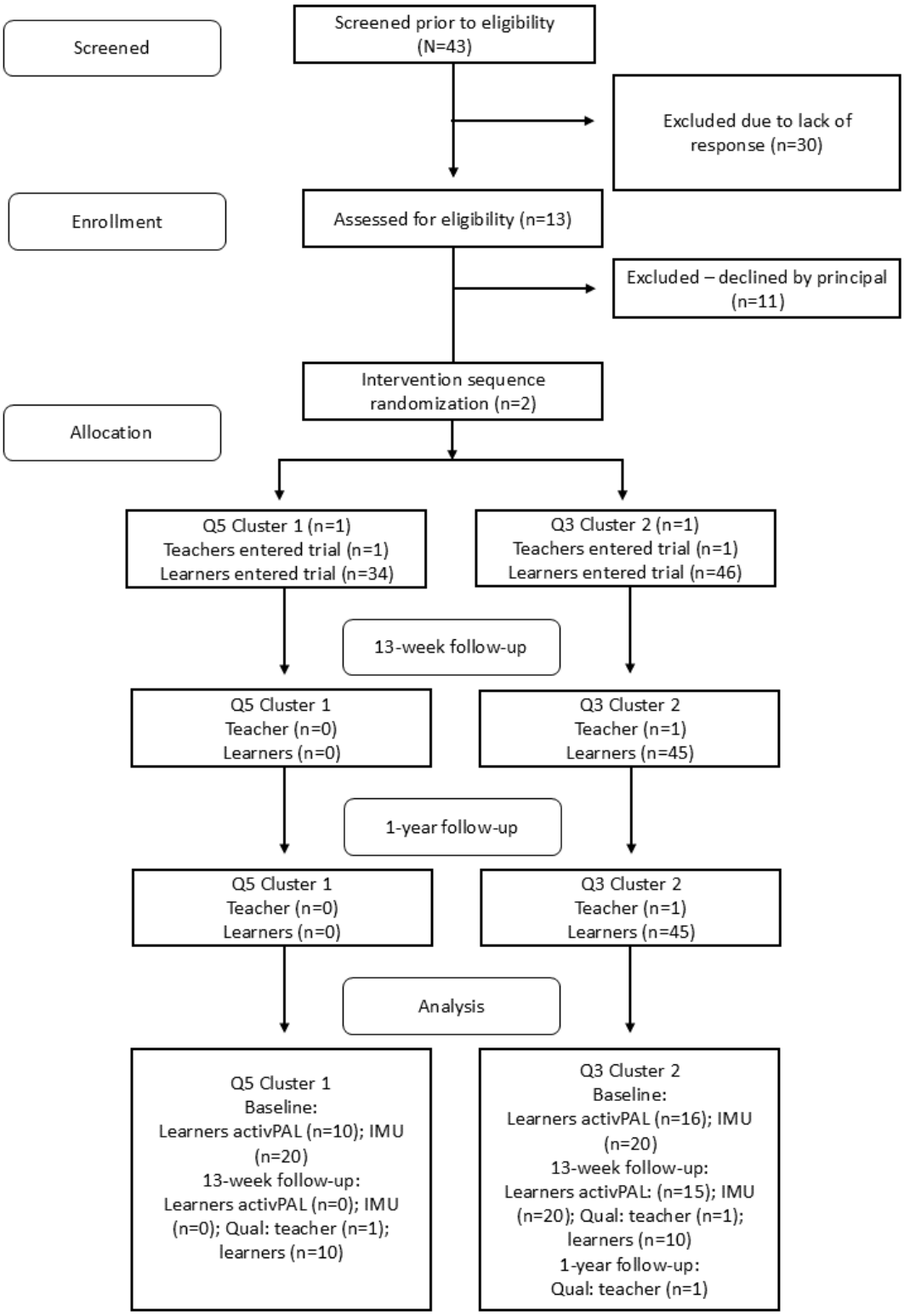

Feasibility Outcome 1: Cluster Retention

The cluster 1 (Q5) teacher withdrew her consent to participate in the study after the 13-week intervention and before physiological follow-up measurements were taken. However, the teacher and learners participated in the IDI and FGDs, respectively. Cluster 2 (Q3) remained in the study throughout the study period. Table 3 demonstrates at which stage the cluster 1 teacher withdrew.

XSL•FO
RenderX

Table 3. Components of retention for clusters 1 and 2.

Cluster	Recruitment	Baseline measurements	Intervention	Follow-up measurements	Qualitative interviews
1	✓	✓	✓	Withdrew	✓
2	✓	✓	✓	✓	✓

Qualitative Findings

Based on the qualitative findings, the reasons for withdrawal from the study were analyzed into 3 themes: the way access to the classroom was negotiated; the teacher's preference for not being video-recorded during data collection; and the perceived disruption to teaching and learning by the study.

Access Negotiation

The same strategy was used to recruit both clusters. Negotiation of access to the classroom was conducted via the school principal and the original teacher (cluster 1). However, a last-minute substitute teacher was employed immediately before the commencement of the study. The school principal discussed the project and its requirements with the substitute teacher. However, the substitute teacher felt that she was not adequately informed of the full extent of the study, despite receiving the study information material.

> *So, I didn't really know. So, I was like, okay, I'll wait for you (the researcher) to come; maybe you'll tell me, and then suddenly, it was a whole setup in my class, and I was, like, what is going on?* [Teacher, cluster 1]

The teacher in cluster 2 was satisfied with the recruitment strategy.

> *Yes. It was the right way. Starting with the principal of course…the principal would never decide on her own. She had to go to the teachers and bring this to the teachers and hear what they say.* [Teacher, cluster 2]

Preference for Not Being Video-Recorded

The cluster 1 teacher was uncomfortable about being recorded on video. She reported that she was self-conscious about being on camera, which influenced her interactions with the learners. The teacher in cluster 2 did not have objections to the use of video for postural dynamism data collection.

> *Well, I mean, having someone basically watching you (laughing). You don't really have freedom with your kids because you say something, but you get recorded on video.* [Teacher, cluster 1]

Perception of Disruption to Teaching and Learning

The cluster 1 teacher felt that the intervention (KUZE) interfered with her preferred way of arranging the classroom furniture.

> *And I found that because the way my class is set out, I put my class when it was rows, so before, I had groups, but then when the desk came, I only could have rows because that was the only thing that could fit in my class.* [Teacher, cluster 1]

To standardize the orientation of learners for postural dynamism measurement, the learners were required to sit facing the front of the classroom. The teacher reported that the learners often had to be moved from their usual desks, and this impacted interpersonal dynamics with the person next to them.

> *Maybe that, but also, as you know my class, I had groups. So, you needed your kids to face the front. So, that also caused moving around and disruptions, like this child don't want to sit next to that child, but yet, they had to because...to accommodate the project.* [Teacher, cluster 1]

The teacher was concerned that disruptions to teaching and learning would impact the learners' academic performance and decided it was best to withdraw from the study.

> *I don't think my kids can afford to have all that interruptions, especially during the second term, when we have exams, and the third term, when we have systemics. So, I needed to be focused all the time. So, having the disruption was just a lot to manage.* [Teacher, cluster 1]

Feasibility Outcome 2: Delivery of HEMVs

Both teachers reported having developed a routine for playing the HEMVs during class. The cluster 1 teacher played the HEMVs intermittently during the day, while the cluster 2 teacher used the videos as a classroom management tool to his advantage. He played the videos at the start of the periods to engage learners and prepare them for the lesson.

> *So, what I did is the morning I played them one, in the beginning of the day and before first break. Then, second break, I don't have them, so...they go for 2 hours…they go somewhere else (to a different classroom), and then maybe a video at the end of the day.* [Teacher, cluster 1]

> *Because of the start of the period, I want them to bring their minds before doing maths…it's like an exercise where everybody would feel right for the lesson…We didn't do that a lot every day. Most of the time, I did the videos when it's my first period in my class.* [Teacher, cluster 2]

Feasibility Outcome 3: Acceptance and Usage of the KUZE Desk in the Classroom

Teachers and learners in both clusters accepted the KUZE as the classroom furniture for the duration of the intervention period. The cluster 1 teacher allowed the learners to stand during groupwork activities for specific subjects and during times where independent work was required.

> *I think the desk was helpful in that situation, yes, where they had their own work to do without having extra work on the board or looking at the board. So,*

XSL•FO
RenderX

> *when they had independent tasks like that, then I think it was okay to be standing, but also not for a long period of time because they tend to move around and they're free to walk around.* [Teacher, cluster 1]

> *I think it was actually amazing for me not to be sitting...just sitting there like a stick insect the whole time, and sometimes, you can't even concentrate in class because your head is down. But you're active from those desks you brought us (KUZE).* [Learner, cluster 1]

Learners in cluster 2 were allowed to use their discretion in whether to sit or stand while completing their work and to break up periods of sitting. At the end of the intervention period, the KUZE desks were removed from cluster 1 and the usual classroom furniture was returned. The cluster 2 teacher requested that the KUZE remain as the standard classroom furniture in his class beyond the completion of the study. When asked about whether they wanted to return to using their usual classroom furniture, learners from cluster 2 chose to keep the KUZE furniture.

> *I think I was going to use it for the new learners maybe next year. Because as I can see now, the other classes, they see these desks. Now they are also curious...* [Teacher, cluster 2]

> *I want you to keep the cosi (KUZE) chairs for us…and our teacher said the cosi (KUZE) chairs are good for us so that we can stand.* [Learner, cluster 2]

The teacher in cluster 2 reported that the KUZE desks provided a valuable advantage of more space to access all the areas of the classroom.

> *The advantage they (KUZE desks) give me is that in that classroom there is space; it's spacious, you see. I can now move in between them…I am able to move around, looking at their books.* [Teacher, cluster 2]

Feasibility Outcome 4: Compliance With Wearing activPAL and IMU Sensors

There was 100% compliance with wearing activPAL sensors and IMUs. No learners reported any adverse skin reactions to wearing the activPAL sensors at any stage of the study. During the 13-week follow-up measurement, 10 participants reported mild head discomfort from the elasticated Velcro strap used to secure the head IMU. In all cases, discomfort was resolved by readjusting the strap. Thereafter, the learners were able to return to class and resume learning activities. The mild discomfort that learners reported from the head strap had resolved before they were allowed to return to the class to complete the period of data collection. All learners selected for physiological measurements complied with wearing the activPAL sensors and IMUs throughout the study.

Feasibility Outcome 5: Integrity of Physical Activity and Postural Dynamism Data

A low rate of data loss was recorded for physical activity and postural dynamism (Table 4). An average of 94.64% of recorded SB and spinal movement data had sound integrity and was considered eligible for inclusion in the analysis for the secondary objectives. The rate of data integrity met the a priori target of 80%. Data loss from wearable sensors was attributable to technical problems during downloading to the project computer or due to batteries running out of charge.

Table 4. Data integrity from wearable sensors.

Measurement period	Cluster 1		Cluster 2	
	activPAL (%)	IMU[a] (%)	activPAL (%)	IMU (%)
Baseline	87.23	92.86	98.48	100.00
13 weeks	Withdrew	Withdrew	98.48	95.00
1-year follow up	Withdrew	Withdrew	83.75	Not measured

[a]IMU: inertial measurement unit.

Summary of Feasibility Outcomes

Three of the feasibility outcomes were achieved in 3 criteria in both clusters. As cluster 1 was not retained in the study, the acceptance and usage of the KUZE could only be assessed partially through qualitative feedback and was thus categorized as "partially met." All 5 feasibility outcomes were achieved in cluster 2 (Table 5).

XSL•FO
RenderX

Table 5. Achievement of success criteria.

Feasibility outcome	Success indicator	Outcome	
		Cluster 1	Cluster 2
Classroom retention	Both clusters remained in the study until follow-up measurements were obtained.	Unmet	Met
Delivery of HEMVs[a]	The teacher developed a routine of playing the videos and adhered to the routine.	Met	Met
Acceptance and usage of the KUZE desk in the classroom	Learners and teachers accepted and used the KUZE as classroom furniture for the entire study period.	Partially met	Met
Compliance with wearing activPAL and IMU[b] sensors	There was 100% compliance with wearing activPAL sensors and IMUs for the duration of the intended measurement period.	Met	Met
Integrity of physical activity and postural dynamism data	Of all data captured by activPAL and IMU sensors, 80% were eligible for inclusion in the analysis.	Met	Met

[a]HEMV: health education and movement video.

[b]IMU: inertial measurement unit.

Discussion

Principal Findings

This study effectively assessed the feasibility of implementing a classroom-based intervention to reduce classroom SB and promote spinal health in primary school learners in the Western Cape province of South Africa. Our findings showed that the a priori success criteria to decide to continue with a large trial were met. Having met 3 feasibility outcome success criteria (delivery of HEMVs, compliance with wearing physical activity and postural dynamism sensors, and integrity of the data gathered) in both clusters indicates that with minor modifications to the study methodology, it is feasible to progress to a larger trial. Despite this positive finding, reduced cluster retention may be a risk of a larger trial. Therefore, strategies to enhance retention are an important consideration for a larger trial.

Individual teacher-related aspects should be considered in retention strategies. The last-minute substitution of the teacher meant that the research team was unable to establish a satisfactory level of engagement, rapport, and support before the study started. This may have led to a negative attitude toward the study, which was compounded by her perceived loss of autonomy due to the study-related change to the classroom furniture and configuration. Holistic support of teachers is important in classroom-based movement integration interventions [24] to maintain adequate levels of engagement and support throughout the process of the study. Such a holistic whole-school approach that includes senior school management and other school stakeholders has gathered traction in interventions to increase physical activity in UK primary schools [25]. However, such a whole-school approach will need to be contextualized for the South African school system, considering the scarce track record of implementation of similar interventions. Given the importance of teachers in implementing the study and the dynamic nature of school settings, the design of a future trial should incorporate pragmatism to ensure that timely and appropriate adaptation is possible throughout the study [26].

Although not retaining cluster 1 (Q5) was disappointing, it provides a valuable learning opportunity to inform a larger trial. It is, therefore, recommended that a customized school context–specific program of engagement, awareness, and involvement in the study precede a future trial to improve participants' experience of participation [27] and to optimize its acceptability and efficacy [28]. Prior teacher involvement and program monitoring may also provide an opportunity to codevelop cluster retention strategies for implementation in a future trial; aid the development, implementation, and adaptation of the intervention; and track cluster progress against study goals [29]. Cocreation of school-based health interventions has been shown to improve the feasibility of interventions in the complex environment of the school context [30]. Including teachers in the development and implementation of an intervention can leverage their expertise of the classroom environment and enable them to find solutions to concerns about the potential classroom disruptions and reconfiguration of classroom furniture. Raising awareness about the study with a view to including teachers to codevelop the intervention may be important considerations for a future trial, given the sizeable investment requirement in this emerging area of research in South Africa.

Despite no previous classroom-based interventions including sit-stand desks in South Africa, the KUZE desk was well accepted by teachers and learners. Teachers recognized the positive impact of allowing learners to stand on their ability to work independently and on their concentration after prolonged periods of sitting. However, some concerns were raised about disruption to teaching and learning and loss of control of the classroom, which presents a potential threat to the rollout of a future trial. Similar concerns were voiced by teachers in a previous UK-based trial [11]. These concerns are reasonable, given the longstanding tradition and practice of enforcing learners to remain seated during class time. Interrupting periods of sitting with standing tasks challenges teachers' practice and propensity to enforce sitting during teaching and learning activities. In a study of teachers' perspectives of learners' movement during class, participants reported that they believed that sitting enhances learners' concentration [7] and were

XSL•FO
RenderX

unaware of the evidence that interrupting prolonged sitting may enhance their concentration and engagement with cognitive tasks [31,32]. Teachers in the study by Fisher and Louw [7] associated enforcing that learners remain seated with teachers' maintaining control of the classroom and their level of classroom management skill and teaching pedagogy. It may be prudent to emphasize the benefits of interrupting prolonged sitting for learners in future teacher training and support as part of a future trial of the intervention.

The activPAL sensor wear protocol used in this study did not appear to impact the logistics or integrity of the physical activity data collected. The decision to apply and remove the activPAL sensors daily was mainly motivated by not wanting to risk losing the expensive devises. The research team also considered the ethics of the potential safety risks study participants would bear by wearing expensive devices in their communities without any mitigating measures in place. Most studies of free-living physical activity using activPAL adopt a 24-hour, 7-day wear protocol [33]. However, this wear protocol has drawn some criticism, with evidence of dose–response noncompliance with wear protocols approaching the 7-day period [34]. Reports of skin irritation caused by perspiration, particularly in warm, humid conditions, have threatened study participants' compliance with wearing the sensors [35]. The fact that no skin irritation was reported by study participants justifies the rationale for the wear protocol. If this activPAL wear protocol was adopted in a future study, it may be useful to determine the validity of the single wear protocol compared to the 24-hour, 7-day wear protocol.

Study Limitations

The study was limited to schools in a specific geographical and socioeconomic context. Therefore, the findings may not be broadly generalizable. The interpretation of some findings was limited by the suboptimal retention of 1 cluster. The study was also limited to specific preset feasibility criteria and important implementation factors; for example, teacher and school dynamics were not considered. The feasibility study's data are also arguably too limited for sample size considerations for a larger trial, and further research is required.

Conclusion

This study addressed the uncertainty of the acceptability of a classroom-based SB and spinal health intervention, its impact on teachers' practice, and the logistical and pragmatic considerations of data collection in a previously unresearched setting. Having established the feasibility of some aspects of the intervention in the local context, the acceptance of sit-stand desks remains questionable and presents a potential risk to the success of a larger trial. Although teachers and learners from both clusters accepted and used the intervention classroom furniture for the duration of the intervention period, the feedback from the cluster 1 teacher raises some uncertainty. However, these concerns can be addressed, as described in the recommendations. The results of this study indicate that minor revisions to the current methodology are required to conduct a large cluster randomized controlled trial (RCT) of an intervention aimed at reducing SB and improving spinal health in primary schools in the Western Cape province.

Recommended changes to the methodology include:

- A holistic whole-school approach to the intervention, including all grade teachers and school management, to support implementation of the intervention, particularly playing HEMVs
- A monitoring strategy to track the impact of the intervention on the classroom, teacher, and school's environment, as well as the ability to adapt and implement pragmatic strategies to enhance retention
- Inclusion of teachers in the cocreation, monitoring, and implementation of the intervention to improve the feasibility of a future trial
- Contextualized teacher training on SB and how sit-stand desks and HEMVs can be used as a classroom management tool

Acknowledgments

This project was funded by the National Research Foundation (CPRR150720128324). The authors thank the innovation team, Jaco Langenhoven and Sjan-Mari van Niekerk (sit-stand desk design and production) and Kurt Minnaar and D'ave Meyer (health education and movement video production), for contributing to the study.

Data Availability

The datasets used and analyzed during the study are available from the corresponding author upon reasonable request.

Authors' Contributions

DF was the major contributor to drafting and writing the manuscript. Conceptualization, formal analysis, investigation, data curation, and visualization were managed by DF and QL; methodology by DF, QL, and LT; validation and writing—review and editing by QL, LT, and RM; and funding acquisition by QL. All authors have read and agreed to the published version of the manuscript. No generative artificial intelligence was used in the production of this manuscript.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
CONSORT (Consolidated Standards of Reporting Trials) extension statement for feasibility of sedentary behavior (SB) and spinal health.
[PDF File (Adobe PDF File), 359 KB - formative_v9i1e65169_app1.pdf]

Multimedia Appendix 2
Learner using KUZE for sitting and standing.
[PNG File , 551 KB - formative_v9i1e65169_app2.png]

Multimedia Appendix 3
Still from a health education and movement video (HEMV).
[PNG File , 294 KB - formative_v9i1e65169_app3.png]

Multimedia Appendix 4
Cluster classroom furniture.
[PNG File , 296 KB - formative_v9i1e65169_app4.png]

References

1. Tremblay MS, Aubert S, Barnes JD, Saunders TJ, Carson V, Latimer-Cheung AE, SBRN Terminology Consensus Project Participants. Sedentary Behavior Research Network (SBRN) - Terminology Consensus Project process and outcome. Int J Behav Nutr Phys Act 2017 Jul 10;14(1):75 [FREE Full text] [doi: 10.1186/s12966-017-0525-8] [Medline: 28599680]
2. Poses-Ferrer E, Parisi R, Gonzalez-Viana A, Castell C, Arias de la Torre J, Jones A, et al. Daily sitting time and its association with non-communicable diseases and multimorbidity in Catalonia. Eur J Public Health 2022 Feb 01;32(1):105-111 [FREE Full text] [doi: 10.1093/eurpub/ckab201] [Medline: 34850878]
3. Biddle SJ, Pearson N, Ross GM, Braithwaite R. Tracking of sedentary behaviours of young people: a systematic review. Prev Med 2010 Nov;51(5):345-351. [doi: 10.1016/j.ypmed.2010.07.018] [Medline: 20682330]
4. Fisher D, Louw Q. The effect of classroom-based interventions on sedentary behavior and spinal health in schoolchildren: systematic review. Interact J Med Res 2022 Oct 26;11(2):e39006 [FREE Full text] [doi: 10.2196/39006] [Medline: 36287590]
5. Ryan DJ, Stebbings GK, Onambele GL. The emergence of sedentary behaviour physiology and its effects on the cardiometabolic profile in young and older adults. Age (Dordr) 2015 Oct 28;37(5):89 [FREE Full text] [doi: 10.1007/s11357-015-9832-7] [Medline: 26315694]
6. Ridgers ND, Salmon J, Ridley K, O'Connell E, Arundell L, Timperio A. Agreement between activPAL and ActiGraph for assessing children's sedentary time. Int J Behav Nutr Phys Act 2012 Feb 19;9(1):15 [FREE Full text] [doi: 10.1186/1479-5868-9-15] [Medline: 22340137]
7. Fisher D, Louw Q. Primary school learners' movement during class time: perceptions of educators in the Western Cape, South Africa. BMC Public Health 2023 Dec 13;23(1):2501 [FREE Full text] [doi: 10.1186/s12889-023-17428-3] [Medline: 38093224]
8. Gibson CA, Smith BK, DuBose KD, Greene JL, Bailey BW, Williams SL, et al. Physical activity across the curriculum: year one process evaluation results. Int J Behav Nutr Phys Act 2008;5(1):36. [doi: 10.1186/1479-5868-5-36]
9. O'Sullivan K, O'Sullivan P, O'Keeffe M, O'Sullivan L, Dankaerts W. The effect of dynamic sitting on trunk muscle activation: a systematic review. Appl Ergon 2013 Jul;44(4):628-635. [doi: 10.1016/j.apergo.2012.12.006] [Medline: 23369370]
10. Hoy D, Brooks P, Blyth F, Buchbinder R. The epidemiology of low back pain. Best Pract Res Clin Rheumatol 2010 Dec;24(6):769-781. [doi: 10.1016/j.berh.2010.10.002]
11. Clemes SA, Bingham DD, Pearson N, Chen Y, Edwardson CL, McEachan RRC, et al. Stand out in class: restructuring the classroom environment to reduce sitting time - findings from a pilot cluster randomised controlled trial. Int J Behav Nutr Phys Act 2020 Apr 29;17(1):55 [FREE Full text] [doi: 10.1186/s12966-020-00958-z] [Medline: 32349761]
12. Hinckson EA, Aminian S, Ikeda E, Stewart T, Oliver M, Duncan S, et al. Acceptability of standing workstations in elementary schools: a pilot study. Prev Med 2013 Jan;56(1):82-85. [doi: 10.1016/j.ypmed.2012.10.014] [Medline: 23103223]
13. Campbell MK, Piaggio G, Elbourne DR, Altman DG, CONSORT Group. Consort 2010 statement: extension to cluster randomised trials. BMJ 2012 Sep 04;345(sep04 1):e5661-e5661. [doi: 10.1136/bmj.e5661] [Medline: 22951546]
14. Fisher D, Louw Q, Thabane L. Sedentariness and back health in Western Cape primary school students: protocol for a pragmatic stepped-wedge feasibility randomized controlled trial. JMIR Res Protoc 2020 Nov 30;9(11):e18522 [FREE Full text] [doi: 10.2196/18522] [Medline: 33089783]
15. Hemming K, Taljaard M, Grimshaw J. Introducing the new CONSORT extension for stepped-wedge cluster randomised trials. Trials 2019 Jan 18;20(1):68 [FREE Full text] [doi: 10.1186/s13063-018-3116-3] [Medline: 30658677]
16. Copas AJ, Lewis JJ, Thompson JA, Davey C, Baio G, Hargreaves JR. Designing a stepped wedge trial: three main designs, carry-over effects and randomisation approaches. Trials 2015 Aug 17;16(1):352 [FREE Full text] [doi: 10.1186/s13063-015-0842-7] [Medline: 26279154]

XSL•FO
RenderX

17. van Dyk H, White C. Theory and practice of the quintile ranking of schools in South Africa: a financial management perspective. SAJE 2019 Sep 30;39(Supplement 1):S1-S9. [doi: 10.15700/saje.v39ns1a1820]
18. Eldridge SM, Chan CL, Campbell MJ, Bond CM, Hopewell S, Thabane L, et al. CONSORT 2010 statement: extension to randomised pilot and feasibility trials. Pilot Feasibility Stud 2016 Oct 21;2(1):64. [doi: 10.1186/s40814-016-0105-8]
19. Ridley K, Ridgers ND, Salmon J. Criterion validity of the activPAL™ and ActiGraph for assessing children's sitting and standing time in a school classroom setting. Int J Behav Nutr Phys Act 2016 Jul 07;13(1):75 [FREE Full text] [doi: 10.1186/s12966-016-0402-x] [Medline: 27387031]
20. Contardo Ayala A, Salmon J, Timperio A, Sudholz B, Ridgers N, Sethi P, et al. Impact of an 8-month trial using height-adjustable desks on children's classroom sitting patterns and markers of cardio-metabolic and musculoskeletal health. Int J Environ Res Public Health 2016 Dec 10;13(12):1227 [FREE Full text] [doi: 10.3390/ijerph13121227] [Medline: 27973414]
21. Sherry AP, Pearson N, Ridgers ND, Johnson W, Barber SE, Bingham DD, et al. Impacts of a standing desk intervention within an English primary school classroom: a pilot controlled trial. Int J Environ Res Public Health 2020 Sep 26;17(19):7048 [FREE Full text] [doi: 10.3390/ijerph17197048] [Medline: 32993142]
22. Balasubramania S. Comparison of Angle Measurements between Vicon and Myomotion Systems. AZ: Center for Adaptive Neural Systems (ANS), Arizona State University; 2013.
23. Friese S. Qualitative Data Analysis with ATLAS.ti. Third edition. Thousand Oaks, CA: SAGE Publications; 2019:xxxiii.
24. Routen A, Johnston J, Glazebrook C, Sherar L. Teacher perceptions on the delivery and implementation of movement integration strategies: the CLASS PAL (Physically Active Learning) Programme. Int J Educ Res 2018 Mar;88:48-59. [doi: 10.1016/j.ijer.2018.01.003]
25. Jones G, Longbon K, Williams S. Exploring the acceptability and feasibility of a whole school approach to physical activity in UK primary schools: a qualitative approach. BMC Public Health 2022 Nov 30;22(1):2236 [FREE Full text] [doi: 10.1186/s12889-022-14647-y] [Medline: 36451192]
26. Pearson M, Chilton R, Wyatt K, Abraham C, Ford T, Woods H, et al. Implementing health promotion programmes in schools: a realist systematic review of research and experience in the United Kingdom. Implement Sci 2015 Oct 28;10(1):149. [doi: 10.1186/s13012-015-0338-6]
27. Wilson P, Mathie E, Keenan J, McNeilly E, Goodman C, Howe A, et al. ReseArch with Patient and Public invOlvement: a RealisT evaluation – the RAPPORT study. Health Serv Deliv Res 2015 Sep;3(38):1-176. [doi: 10.3310/hsdr03380] [Medline: 26378332]
28. Sanders EB, Stappers PJ. Probes, toolkits and prototypes: three approaches to making in codesigning. CoDesign 2014 Mar 06;10(1):5-14. [doi: 10.1080/15710882.2014.888183]
29. Schultz L, Ruel-Bergeron J. Considerations for monitoring school health and nutrition programs. Front Public Health 2021 Jul 16;9:645711 [FREE Full text] [doi: 10.3389/fpubh.2021.645711] [Medline: 34336757]
30. Clifford L, Tyler R, Knowles Z, Ashworth E, Boddy L, Foweather L, et al. Co-creation of a school-based motor comeptence and mental health intervention: move well, feel good. Children (Basel) 2023 Aug 17;10(8):1403 [FREE Full text] [doi: 10.3390/children10081403] [Medline: 37628403]
31. Tobia V, Sacchi S, Cerina V, Manca S, Fornara F. The influence of classroom seating arrangement on children's cognitive processes in primary school: the role of individual variables. Curr Psychol 2022 Oct 31;41(9):6522-6533 [FREE Full text] [doi: 10.1007/s12144-020-01154-9] [Medline: 33162725]
32. Wannarka R, Ruhl K. Seating arrangements that promote positive academic and behavioural outcomes: a review of empirical research. Support Learn 2008 May 06;23(2):89-93. [doi: 10.1111/j.1467-9604.2008.00375.x]
33. Edwardson CL, Winkler EA, Bodicoat DH, Yates T, Davies MJ, Dunstan DW, et al. Considerations when using the activPAL monitor in field-based research with adult populations. J Sport Health Sci 2017 Jun;6(2):162-178 [FREE Full text] [doi: 10.1016/j.jshs.2016.02.002] [Medline: 30356601]
34. Shi Y, Huang W, Yu J, Sheridan S, Sit CHP, Wong SHS. Compliance and practical utility of continuous wearing of activPAL™ in adolescents. Pediatr Exerc Sci 2019 Aug 01;31(3):363-369. [doi: 10.1123/pes.2018-0170] [Medline: 30605620]
35. Maddocks M, Byrne A, Johnson CD, Wilson RH, Fearon KCH, Wilcock A. Physical activity level as an outcome measure for use in cancer cachexia trials: a feasibility study. Support Care Cancer 2010 Dec 3;18(12):1539-1544. [doi: 10.1007/s00520-009-0776-2] [Medline: 19956982]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
FGD: focus group discussion
HEMV: health education and movement video
IDI: individual depth interview
IMU: inertial measurement unit
NCD: noncommunicable disease

XSL•FO
RenderX

RCT: randomized controlled trial
SB: sedentary behavior

Edited by A Mavragani; submitted 07.08.24; peer-reviewed by F Yang; comments to author 29.10.24; revised version received 18.11.24; accepted 28.11.24; published 24.02.25.

Please cite as:
Fisher D, Maart R, Thabane L, Louw Q
A Classroom-Based Intervention for Reducing Sedentary Behavior and Improving Spinal Health: Pragmatic Stepped-Wedge Feasibility Randomized Controlled Trial
JMIR Form Res 2025;9:e65169
URL: https://formative.jmir.org/2025/1/e65169
doi:10.2196/65169
PMID:

©Dominic Fisher, Rentia Maart, Lehana Thabane, Quinette Louw. Originally published in JMIR Formative Research (https://formative.jmir.org), 24.02.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Formative Evaluation of Suicide Prevention Websites for Men: Qualitative Study with Men at Risk of Suicide and with Potential Gatekeepers

Doreen Reifegerste[1], Prof Dr; Anna J M Wagner[2], Prof Dr; Lisa Huber[3], MA; Manuel Fastuca[3]

[1]School of Public Health, Bielefeld University, Bielefeld, Germany
[2]Centre for Language Studies, Radboud University, Nijmegen, The Netherlands
[3]GIM Gesellschaft für innovative Marktforschung mbH, Heidelberg, Germany

Corresponding Author:
Doreen Reifegerste, Prof Dr
School of Public Health
Bielefeld University
Universitaetsstrasse 25
Bielefeld, 33615
Germany
Phone: 49 521 10667 ext 807
Email: doreen.reifegerste@uni-bielefeld.de

Abstract

Background: The suicide rate among men exceeds that of women worldwide. One important measure in suicide prevention for men is digital communication interventions, as they enable easy and anonymous access to information resources. This is especially important for men who might not be reached by traditional, in-person prevention methods. Thus, as part of an interdisciplinary project on suicide prevention for men, two specific digital communication prevention measures were developed: (1) a website to inform men at risk about suicide prevention, and (2) a website to educate potential gatekeepers who are in contact with men at risk of suicide about appropriate life-saving measures. Both websites needed evaluation to explore how they are perceived by (1) men and by (2) potential gatekeepers of men at risk of suicide. This is crucial, as existing research lacks formative evaluation that informs the development of intervention communication materials.

Objective: This study aimed to analyze whether these websites were perceived as (1) comprehensible and engaging, (2) authentic and trustworthy, as well as (3) useful by (potential) users. Furthermore, we examined (4) additional ideas for effective communication about suicide prevention.

Methods: We conducted (1) individual videoconference interviews with 24 men to evaluate the website and (2) four focus groups with 8 gatekeepers in each group (32 participants) to evaluate the online education program. The focus group sample was equally distributed regarding gender and age. Recruitment was conducted together with a field research partner who posted adverts on Facebook and Instagram (Meta) to reach as many potential participants as possible in an efficient way. All participants were asked to evaluate the intervention materials using a fictitious scenario of a man experiencing a mental health crisis before the interviews or focus groups took place.

Results: The videos were perceived as (1) catchy, comprehensible, and empathetic, but too long for a short introduction. A balanced mix of emotional and informative content was considered appropriate and helpful. The health information provided was perceived as (2) serious and trustworthy due to citing scientific institutions and video material of men who had experienced suicidal ideation. (3) The intervention's applicability for men experiencing acute crisis was critiqued, but it was regarded as very useful for comprehensive information. (4) Further communication channels and addressing other male subgroups or gender identities were presented as possible extensions of the program.

Conclusions: Effective suicide prevention research should address both the groups at risk and their support network. Digital communication interventions can provide low-threshold access. Videos with personalized examples are important to give men someone to identify with, which validates their emotional responses and supports their self-esteem, while videos with experts provide relevant and credible information.

(JMIR Form Res 2025;9:e59829) doi:10.2196/59829

XSL•FO
RenderX

KEYWORDS

mental health; suicide prevention; men's health; evaluation; website; gatekeeper; suicide; male; suicide risk; digital communication; intervention; suicidal behavior; digital intervention; digital media

Introduction

Suicide rates among men surpass those among women worldwide. For instance, in Germany, approximately 75% of all suicides are carried out by men [1]. Even though recent reviews have emphasized advances in our understanding of risk factors and countermeasures for suicidal behavior in men and women [2,3], it remains a reality that men are at substantially higher risk. Preventive measures are hence needed that specifically address men and consider the particular needs of men at risk and their suicide-related communication [4].

One important measure in suicide prevention is digital communication interventions. Digital media such as websites are used as information sources on suicide methods but could also be used as easily accessible sources of suicide prevention support [5]. As they enable easy and anonymous access to information resources, digital media can be beneficial tools for a variety of different population groups [6]. Men in particular might profit from digital interventions, as they are less likely to be reached by traditional, in-person prevention methods that might not be tailored to their needs [7]. Recent systematic reviews [8,9] show that communication-based digital interventions appear to be effective measures for male suicide prevention. While the interventions demonstrate high overall effectiveness on suicide-related outcomes (eg, knowledge, attitudes, and cognitions), one of the reviews [9] also found no adverse effects linked to these interventions among men. This means that the risk associated with digital interventions for men at risk seems to be particularly low.

In addition, interventions for so-called "gatekeepers" (people who are in contact with individuals at risk of suicide, eg, family members or friends) play a central role in suicide prevention [10]. They can be considered to fulfill an important bridging function, extending intervention reach to this group of men at risk and being well-placed to initiate difficult conversations [11]. Sharing distressing personal experiences with trusted social contacts can be an important first step in seeking psychological support, for example in the case of clinical depression [12]. Men seem to be less likely to seek professional or informal help and less likely to communicate directly to their closest contacts about suicidal intentions; however, they still show signs of change (eg, further self-isolation, sudden anger, or hopelessness) before the suicide, which makes gatekeeper interventions all the more important [13,14].

Training gatekeepers on how to approach men at risk, and react appropriately in case of disclosure, is of paramount importance. Based on evidence from a systematic review, gatekeepers can be effective in their bridging function for suicide prevention [15], particularly where training programs succeed in sensitizing gatekeepers and building their competencies for engaging with men at risk [16]. Digital training programs for adolescents have been shown to be successful in the development of gatekeepers' competencies, such as for teachers and peers [17-19].

To date, however, further evaluations for digital interventions for adult men and their (potential) gatekeepers are required [20]. This especially holds true for qualitative research evaluations, such as with interviews and focus groups. While randomized controlled trials (RCTs) and quantitative evaluations are necessary for testing programs' effectiveness and have also been completed within this project [11], qualitative methods offer the chance to not only test the usability of the developed programs but can also generate insights extending beyond the specific intervention [21].

We therefore conducted two qualitative studies evaluating (1) a website addressing particularly men at risk about suicide prevention, and (2) a web-based program to educate potential gatekeepers about appropriate and critical life-saving measures. The communication strategies selection for the websites was based on the findings of recent systematic reviews, which focused on effective message strategies and communication campaigns for suicide prevention [8,9]. Here, it was recommended to include testimonials and educational materials. The content was also informed by research on gender-specific suicide-related communication [14] as well as the typical communication behavior of men in the days before a suicide [13]. For more details on the structure, components, and materials of both websites [11,22].

The website is specifically aimed at men experiencing acute (suicidal) crises and provides visitors with information about suicidal experiences and behavior, offers of help, and signposts ways to access health provision and mental health support. Using short videos, men at risk talk about their experiences with suicidal crises and how they found their way out of the crisis. In addition, the videos address images of masculinity, stigmatization, shame, and overcoming obstacles to seeking help. Furthermore, experts from various professions within the health and care sectors provide more detailed information. These experts include a male professor and psychiatrist as well as a female professor and psychologist, all of whom primarily conduct research on mental health and suicide.

The web-based program for gatekeepers is aimed at people who are worried about a loved one and consists of four modules. The first module provides information about suicidal experiences and behavior, detailing warning signs. The second module explores communication and conversation strategies. The third module focuses on support services for signposting. The last module addresses the burden on relatives themselves, and their options for self-care, and gives further recommendations for stress-management strategies.

The web-based program (similar to the website) consists of videos featuring men at risk and experts from the health and care sectors. Participants also learn about communication strategies and conversation skills using exemplar audio recordings of people with suicidal experiences and behavior.

XSL•FO
RenderX

There were multiple focuses when evaluating both programs, given their differing objectives and content. On a more superficial level, usability evaluations were key. These have the objective of analyzing whether the digital interface is easy to understand and appealing to potential users [21]. Usability, in general, "is the ability of a product to be understood, learned, used, and attractive to the user" [21], which emphasizes a user-centered approach and focuses on their satisfaction. In health-related contexts, the credibility and authenticity of information is another important determinant of user satisfaction [23]. Because there is an excess of information available online (of varying quality) and because people at risk are very vulnerable, trust is key to their health information-seeking behaviors [24]. Often, expert sources as well as individuals with personal experience can be convincing in this regard.

Finally, the aspect of perceived utility should be considered, because this also influences the intention to keep using the sites and recommendations to others.

Thus, in our evaluations, we aim to analyze whether the website and the web-based gatekeeper program are perceived as (1) comprehensible and engaging, (2) trustworthy, and (3) useful by the two different audiences. Furthermore, we were interested in (4) additional ideas for the communication of suicide prevention beyond the developed digital interventions for men.

Methods

Participants

To pursue these 4 research goals, a 2-step study with both men and gatekeepers was conducted.

For the evaluation of the websites for males, 24 videoconference interviews with men in 3 age groups (18 to 30 years old, 31 to 50 years old, 50 years and older; 8 participants per age group) were conducted.

We included all persons who identified as men and were older than 18 years. We did not explicitly search for men with a history of suicidal behavior nor relatives of men with a history of suicidal experience and behavior given ethical considerations. However, based on the recruitment information (which disclosed the aim of the website as one for suicide prevention), many participants reported that they themselves were affected by suicidal ideation and behavior, or had vicarious experiences through friends, acquaintances, or relatives. We did not exclude these participants from the sample.

For the evaluation of the gatekeeper program, four videoconference focus groups with gatekeepers with 8 people each (32 participants in total) were conducted. The focus group sample was equally distributed regarding gender and age (50%, 16/32 female and 50%, 16/32 male; 50%, 16/32 individuals from younger, ie, under 40 years old, and 50%, 16/32 from older age groups, ie, over 40 years old).

For an overview of the specific age and gender of each participant refer to Table 1. A videoconference format for the interviews and the focus groups was chosen to maximize ease of attendance and the potential for anonymous participation.

XSL•FO
RenderX

Table 1. Age and gender of participants in interviews (24 men) and focus groups (16 men and 16 women in 4 groups) for the evaluation of websites for suicide prevention.

Method and gender	Age (years)	Interviewer or moderator
Interviews		
Male	48	Female
Male	63	Female
Male	39	Female
Male	24	Male
Male	48	Male
Male	67	Female
Male	56	Female
Male	66	Female
Male	62	Female
Male	49	Female
Male	30	Female
Male	33	Female
Male	30	Female
Male	29	Female
Male	35	Male
Male	36	Male
Male	40	Male
Male	65	Male
Male	25	Male
Male	24	Male
Male	54	Female
Male	51	Female
Male	20	Female
Male	21	Female
Focus group 1		Male, Female
Female	60	
Female	58	
Female	55	
Female	47	
Female	41	
Female	72	
Female	41	
Female	51	
Focus group 2		Male, Female
Male	35	
Male	18	
Male	36	
Male	27	
Male	33	
Male	34	

XSL•FO
RenderX

Method and gender	Age (years)	Interviewer or moderator
Male	27	
Male	40	
Focus group 3		Male, Female
Female	22	
Female	34	
Female	32	
Female	24	
Female	29	
Female	30	
Female	37	
Female	30	
Focus group 4		Male, Female
Male	71	
Male	44	
Male	65	
Male	58	
Male	61	
Male	75	
Male	46	
Male	41	

Sample sizes were therefore consistent with recommendations for sample sizes in interviews and focus groups to reach sufficient saturation [25]. As it happens, data saturation in the interviews was already reached after about 20 interviews. The number of people refusing to take part in the study or not showing up at the scheduled time was low to nonexistent (especially in comparison to market research projects).

Recruitment

Recruitment was conducted together with the field research partner Quotapoint Pharma and Health in Germany, which specializes in recruiting participants for qualitative studies on health-related topics. They used an announcement in their newsletter (refer to Multimedia Appendix 1 for an English version of the recruitment material). They also posted adverts on Facebook (to reach potential participants older than 35 years) as well as Instagram (to reach potential participants younger than 35 years). This approach to recruitment was chosen in order to reach as many potential participants as possible in an efficient way. After the initial contact was made, a screening process was carried out using a screening tool, with which suitable participants were filtered according to the sample criteria. During the screening process, a brief sociodemographic check was conducted due to the all-male target group for the interviews and the gender-mixed sample for the focus groups. Following this, a trigger warning was given that the subsequent screening interview would be about suicide prevention to minimize the psychological stress on the potential participants. Participant recruitment was documented with regard to the filtering criteria, for example, age or gender, and checked before the participants were finally invited to the interviews and focus groups.

Ethical Considerations

Ethical approval for this study was obtained from the Ethics Committee of Bielefeld University (number 2023-153). All participants were provided with an information form describing the study (refer to Multimedia Appendix 2 for an English translation of this consent form). Participants in both studies were required to return signed consent before they could participate in the study, via a web-based form and a data protection agreement. As the interviews and focus groups took place as a videoconference, the participants were once again reminded about their data and confidentiality rights at the start of the interviews. The interviews were recorded only once active consent to recording had been given by the participants.

As stated above, we avoided explicitly searching for men with a history of suicidal behavior consideration due to ethical considerations.

All data were collected anonymously. All participants were provided with financial compensation in recognition of time spent and their contribution to the study (€90 [US $94] per participant in interviews; €140 [US $146] per participant in focus groups).

Procedure

After giving their consent, all participants were asked to extensively study the intervention materials and briefly evaluate them via a quantitative survey. This was both for the website

XSL•FO
RenderX

for men at risk and the gatekeeper online program (step 1), and was completed before the interviews and focus groups took place (step 2). In the first step, called the pretask, the participants were asked to engage with and reflect on the content in a home setting. Initial findings on the attractiveness, intuitiveness, and comprehensibility of the content were collected using short questions via online survey software. Surveys took approximately 5-10 minutes to complete. Responses from the surveys were included as context information in the qualitative data analysis and were used during the interviews as conversation starters.

In the second step, the interviews and focus groups were conducted in July 2023 using the ZoomX videoconferencing platform (Zoom and Telekom), which is hosted in Germany. After a brief summary of the project (and in the case of the focus group a round of introduction), participants were informed about the research goals, their ability to end participation at every step of the data collection, and about the applicable data protection rules. Moreover, they were given the opportunity to ask questions before the interviews and focus groups began.

All interviews were conducted by one of two authors (LH and MF; one male and one female) and ran for 45 minutes. All focus groups were conducted jointly by one moderator and one assistant and ran for approximately 90 minutes (refer to Table 1 for details on role allocation). These were recorded, using both audio and video. During the interviews, the interviewer took notes on the content of participant discussions. Meanwhile, during the focus groups, the assistant listened to the discussion of the moderator and participants, while taking handwritten notes on all discussion points.

Materials

The website for men with an increased risk of suicide entitled “Männer-stärken” (empowering or strengthening men) offers information on suicide prevention. The website contains videos, information summaries, and lists of helpful recommendations. Complex content is conveyed in simple language and with the help of visual illustrations. Long text sections are avoided in favor of concise information summaries and lists. Graphic illustrations feature a line art and watercolor design. The website is primarily designed in yellow-themed color.

The web-based program “Help for Relatives” provides information on the topic of suicide or suicide prevention in men, with the help of text content on the website and an accompanying manual, as well as four videos approximately 20 minutes long each. It is structured into four modules: (1) suicidal ideation and behavior, (2) communication, (3) offers of assistance, and (4) self-care. Here, visual elements are additionally used to convey complex information.

The short, standardized web-based surveys conducted next assessed a range of topics relevant for website evaluation (confirming participant identity, first impressions of the web-based platform, liking or disliking different aspects as open text fields and grading items as well as optimization potential) comprising closed- and open-ended questions. Their results were used as conversation prompts to provide an easy entrance into the interview situation and generate conversation.

The interviews and focus groups were structured, with a series of guiding questions and prompts exploring the views of men and gatekeepers on the potential design, content, functioning, and user experience (eg, optional phone and email support) of a web-based navigation tool (eg, liking or disliking the online platform for different aspects such as understandability, authenticity, the trustworthiness of given information, and the usability in everyday life and in acute crisis). Questions and prompts were developed based on mental health research as well as usability research. Participants were encouraged to share their perspectives verbally with the group or the interviewer.

Several techniques were used to deal as sensitively as possible with the topic of suicide. One technique was the scenario technique or vignette methodology [26], which was used to keep the interviewees’ psychological stress as low as possible and enable them to evaluate the digital interventions on the basis of a standardized crisis situation: a fictitious scenario of a man in a psychological crisis was presented, which served as the basis for the interviews. In addition (and also to minimize emotional stress), we first asked the men indirectly about the assumed perceptions of others, before they were asked to present their own views. This so-called circular questioning is a technique often used in therapy to investigate sensitive topics [27]. All materials were pretested in advance with persons matching the sample structure (refer to Multimedia Appendix 3 for the guides and short protocols of the interviews and the focus groups).

Data Analysis

Regarding personal reflections that were relevant to our analysis, there were methodological as well as gender aspects to consider during the data analysis phase. One part of the team comprises established scientific researchers, who are rooted in theory-based quantitative and experimental as well as qualitative research, while the other part has long experience in market research for new websites and consumer products. We tried to link these 2 worlds and systemize the content of the interviews and focus groups against the background of theories like the unified theory of acceptance and use of technology and similar theories [28] as well as known consumer needs for products in digital environments (Multimedia Appendices 4 and 5). Such products, like cars or digital devices also often address a primarily male audience.

At the same time the three researchers in the team who identify as female tried to be especially empathetic to male perspectives, whereby we for instance explicitly asked for input from the male researcher in our team.

Taking into account the reflexivity of the researchers, it is important to reflect on the different perspectives and life experiences of the researchers who did the coding (a man over 50 and a woman under 30) in terms of how this could influence the analysis of the data. In the sense of Clark and Braun [29], this serves to make the subjective influences of the researchers transparent and to consider them as a research resource rather than a bias. On the one hand, it is therefore important to emphasize age and generational perspectives. The male researcher over the age of 50 years has a more experienced perspective than the female researcher under the age of 30 years,

XSL•FO
RenderX

which proved to be a suitable age structure with regard to the sample of interviewees aged 18-60 years, as all age splits could be covered within the researcher perspectives. At the same time, generationally different views on suicide, suicide prevention, and mental health can be countered from an age perspective. Furthermore, by choosing a man and a woman to analyze and collect the data, gender-specific needs of the participants and insights can be gained from the data. As a man, the male researcher is more attuned to the lifeworld issues and needs of the male target group, whereas the female researcher is more sensitive to gender-specific differences. Finally, the degree of digitality can also be reflected in the analysis and collection of data. It can be assumed here that the younger female researcher has a high level of familiarity with digital technologies and communication styles, which could make a difference in the consideration and interpretation of what the participants say with regard to the user-friendliness and address of the websites. Due to the selected age range of the sample, it is also possible to take into account any degree of digitality of the participants.

Qualitative data were managed using transcripts, an Excel table, and protocols of the interviews and the focus groups. A thematic analysis [29,30] was conducted by 2 authors on interview and focus group data notes, chat conversations, and survey responses to identify and summarize the key topics and preferences raised by the interview and focus group participants within each area of the research questions while preserving the breadth and diversity of perspectives presented. Analyses and developing categories were regularly discussed with all members of the research team.

A combination of deductive and inductive approaches to coding was hence applied to the research. Deductive is here defined as linking the answers to the theoretical and empirical background. More specifically, deductive themes were derived from literature and applied to the material. In contrast, inductive approaches entail being open for topic-specific feedback and including themes that arise from the material and which had not been previously assumed by the researchers.

With regard to the methodological explanations by Braun and Clarke [29] regarding established coding practices, the decision was therefore made to use “mashups,” that is, a mixture of existing literature-based codes and immanent interpretation. This is due to the special feature of the study that, in addition to the content-related exploration of the relevance of such websites for the target group, aspects of user experience were also examined, which sometimes include merely technical aspects.

Before starting the data analysis, an initial broad coding framework was developed based on the key areas of comprehensibility, appeal, credibility, utility, and ideas for further communication measures in relation to the presentation of the website content. The data was coded deductively within each category. As the coding progressed, subcategories that were not included in the deductive framework were developed inductively, such as the relevance of suicide prevention services as a website per se.

Results

Comprehensibility and Appeal

With regard to our first aim of analyzing the comprehensibility and appeal of the videos, findings showed that they were perceived as catchy, understandable, and empathetic by both men and gatekeepers, but too lengthy. Men of all 3 age groups evaluated the website as looking serious, clear, and professional. They especially noted the yellow colors as being positive, friendly, and informative. In contrast, they would perceive the color black as being too masculine, sad, and aggressive.

The content was perceived as easy to receive due to simple and understandable language, and plain typography in combination with images and videos that make complex topics accessible to the general public. The combination of images, video, and text was considered appropriate and varied. Especially the graphic illustrations, short sentences, and comprehensible statistics were perceived as more effective than “text deserts” and allow easy understanding:

> *Numbers also make the whole thing a bit more tangible, I'm perhaps more of a numbers person, so it makes sense in that direction, but I always find it much easier to understand, otherwise you're in kind of an empty space where you talk around [the topic] without having any numbers.* [male, 24 years, interview]

For some videos in the gatekeeper program, the language from featured experts was perceived as being too scientific that is, using too many scientific terms.

Highly modern or fancy websites were not expected by most viewers given the seriousness of the topic; rather, they expected timeless design and a clear presentation of facts.

> *If I make the thing colorful and pretty, then it loses a bit of its meaning;* [male, >41 years, focus groups]

> *It just has to be good and clear and, not like a thousand other sites, I always knew how to navigate [the site] and found guidance.* [male, 62 years, interview]

However, participants also rated the website as too unmodern due to long texts and videos. Especially the videos (on average 7-8 minutes and up to 20 minutes in length) were perceived as too lengthy. Feedback indicated that these demanded too much attention and increased the risk of participants discontinuing the program. The same applies to searching and scrolling on the website, which was also perceived as not user-friendly. Respondents suggested making the aim of the website for men at risk and the gatekeepers’ program clearer at first glance. Specifically, it was suggested to offer specific content sections and “trailer-length” videos for a first impression. The option of more detailed formats should be an option for cases where more information is required. Some even wanted more facts and information in this regard.

Some of the men perceived the title of the website for men (which means empowering or strengthening men) as nonintuitive

XSL•FO
RenderX

and misleading because they had associations with potency substances or coaching about masculinity.

> *If I didn't know the context, then I would think of, I'll say, potency or something like that.* [male, 66 years, interview]

> *well, strengthening can be in many ways, so that could also be some sexual enhancers or some power food supplement stores for bodybuilders.* [male, 51 years, interview]

In addition, they found the claim of "strengthening men" rather counterproductive for the destigmatization of mental distress, which they felt should not be connected to weakness.

> *For me, it's more of an illness that can be treated like any other illness and has nothing to do with strength or weakness. So, I would tend to think that it's the supposedly strong men, who are strong in the classic sense, who can get it because they're not so good at dealing with problems.* [male, 67 years, interview]

Credibility and Authenticity

In terms of authenticity and trustworthiness, the information on the website and in the web-based program was perceived as serious and trustworthy due to referencing universities, health insurance companies, and scientific institutions as sources as well as facts and figures. In addition, the perceived clarity, color choice, and timelessness were attributed as professional and serious.

> *What naturally contributes to credibility is that the logos of the participating universities were right at the front of the home page. Then there were these experts, which of course also increased credibility to a certain extent.* [male, 33 years, interview]

> *I found the interview with the professor very, very good, especially because you can see how extensive the problem is, where he had his problems, and to understand that there is a big process behind it.* [male, <40 years, focus group]

The same perception applies to facts and statistics as an integral part of evidence-based information.

Several men in the interviews stated that it is good that the programs' developers are based in universities. This lent credibility to the websites, which were perceived as trustworthy. The men were sure that the websites were based on the latest research rather than being driven by advertising or a pharmaceutical company for reasons of profit, even when they did not recognize all the universities involved.

In addition, including links to other sources increased the perception of content presented as authoritative.

Participants also rated the videos featuring men at risk of suicidal behavior as authentic and easy to understand. Including these videos offered visitors faces to identify with, and validation for experiences. It also created the feeling for visitors who self-identified as at risk that they are not alone with problems (even without having spoken to others personally about the topic), making it possible to still feel a sense of belonging.

> *Especially with the videos, it brings a personal closeness and makes it easier to accept support. Because you realize that there are also real men who have the same problems.* [male, 36 years, interview]

At the same time, using and engaging with these vignettes can offer a supported form of self-diagnosis which helps men at risk to classify their own feelings and symptoms.

> *That you can draw parallels from yourself to a description or identify with a person who also describes [these experiences].* [male, 56 years, interview]

That the men at risk featured in the videos were of various ages was also perceived as helpful to recognize that this is a problem for all ages.

A video featuring individuals sharing their own experiences can also serve a more inspiring purpose. Sharing these experiences and also how that person also managed to recover can demonstrate how this is possible, changing perceptions of how good the chances are of a recovering person also depicts an example of how good the chances of recovery can be:

> *There were also patients in it, I think several times, I think the same person was in several videos. (Which then takes you by the hand, along the lines of) Look, I'm already here, why don't you come too, you can see I'm not doing so badly. Or it makes me feel better.* [male, 30 years, interview]

In contrast, the men at risk found archive photos of emotional states on the website for the gatekeepers as noncredible, perceiving these instead as too dramatic and inauthentic. This underlines how important the credibility of the material is, as where material otherwise seems staged, inauthentic, and prompts an adverse reaction:

> *So, when I'm told what I'm supposed to feel in such a way, I am out… and that's what these stock photos do, because in my opinion they want to create something very specific and emphasize what I as a consumer am supposed to think.* [male, <40 years, focus group]

Utility

In terms of utility, the websites are perceived as helpful overall. The participants think that the websites are a good place to find information, especially because men at risk also shared their experiences. Because of this inclusion, the information shared was perceived to be helpful as realistic and practical advice by users. Behavioral recommendations were considered to be helpful tips, which were highly relevant to everyday life.

With regards to the website, one man (21 years) reported in an interview:

> *I particularly liked the "Glossary of concerns" and the "Addressing distress" pages, because these are the most practical in my opinion. I also really liked the fact that all the texts always included a link to*

XSL•FO
RenderX

> *further help or resources. These resources were very well structured based on different needs. Contact details were also given everywhere.*
>
> *I mean, you get a lot of information so that you can perhaps recognize the signs [...] when you see the symptoms... irritability, listlessness, and so on.* [male, 36 years, interview]

The gatekeepers reported similar reactions to their program:

> *It was very much about.... how do I try to understand a person and how do I communicate with that person so that the person appreciates it, that's what you're allowed to do?* [male, <40 years, focus group]

The participants found digital anonymity on the internet and the possibility of searching for information without being obliged to do something helpful. According to them, it provides men at risk and concerned family members support with low barriers to access because help is provided in anonymous internet spaces for this very sensitive and stigmatized topic. Several men in the interviews and focus groups stated that they experienced less shame when talking in a videoconference about mental health problems and more potential to engage with the idea of getting help, because of the anonymity.

The help site for relatives was considered to provide significant added value emotionally, as it can support the family members to minimize difficult feelings related to powerlessness.

> *If the people notice something in their surroundings that they can then check with this program as a first step, ok, are these worries that I'm having justified, is this really going in the direction that I'm thinking?* [female, <40 years, focus group]

At the same time, however, the applicability in crises is questioned as to whether men at risk of suicide in an acute crisis would still visit websites to get help because they might not have the energy to do so:

> *So within this situation, he might not want to read something for hours or anything like that, but he's overstimulated, so he doesn't know what to do next, he just needs someone with whom he can somehow communicate verbally, to one.* [male, 66 years, interview]

Some were even concerned that the "unregulated internet space" could promote a problematic approach to the topic of suicide risk and counteract the actual aim of the service. They were worried that such a website could increase the risk of seeking out further information and then receiving "bad advice" or "getting the wrong person" in other online spaces, which would then hinder them from seeking help in "real life."

Participants suggested prefiltering the offers of help for those at risk in an acute crisis, by providing emergency numbers and direct contact persons or a helpline who can offer help in a crisis. Given that websites provide information asynchronously, the hope for immediate, quick help for relatives in emotional distress (contact details, chats, and direct contact options for immediate personal offers of help) remains unfulfilled.

Table 2 summarizes the answers to RQ1 to RQ3.

Table 2. Summary of Results of the Interviews and the Focus Groups and their Feedback on Comprehensibility, Credibility, and Utility of the Websites for Suicide Prevention.

Theme	Interviews with men (24 participants)	Focus groups with Gatekeepers (32 participants)
Comprehensibility and appeal	• Videos were perceived as catchy, understandable, and empathetic • But too lengthy and too fixed • Website is perceived as authoritative, easily understandable • Yellow color as positive, friendly, and informative • The website title was discussed as it seemed ambiguous	• Videos were perceived as catchy, understandable, and empathetic • But too lengthy and too fixed • Too many scientific terms in the expert videos • A modern or fancy design is not expected
Credibility and authenticity	• Websites are perceived as trustworthy due to references to scientific institutions as well as facts and figures • Generic stock photos provoke adverse responses	• Websites are perceived as trustworthy due to references to scientific institutions as well as facts and figures • Generic stock photos provoke adverse responses
Utility	• A good place to find information • But might not be appropriate or helpful in emergency cases	• A good place to find information • But might not be appropriate or helpful in emergency cases

Ideas Beyond the Intervention Material: Other Communication Channels

Both the topic, the desire for education about the topic, and its destigmatization were evaluated as critically important, especially from the group of men younger than 30 years. As a consequence, the respondents also recommended using more communication channels for promotion in both the health and public sector to make the information more accessible to everyone. They suggested using all conceivable communication channels to draw attention to this offer of help, to destigmatize the topic, and to widen the reach for education purposes.

> *I think this should be widely distributed now, because ultimately, of course, I have to reach the men everywhere* [man, 51 years, interview]

XSL•FO
RenderX

This could involve promoting the websites more prominently on social media channels like Instagram and TikTok (ByteDance), in magazines, outdoor media, TV health shows, cinema and radio spots, at train stations, and in sports stadiums.

They also suggested connecting the website for men at risk more visibly with the online program for gatekeepers. This would mean that where the website for men is mentioned in a private conversation with a friend (gatekeeper) who has already screened for information and helpful content from the website. Conversely, a man in acute crisis could recommend the program for the gatekeepers to his relatives.

While it was considered valuable that topics around masculinity and suicidal ideation are being addressed, suicide prevention services specifically for men were questioned as a concept by the participants. Despite the information about the statistically higher suicide rate among men, some of the content produced on the sites was considered too conservative for the way that gender roles were presented, because they offered no identification potential for other gender identities, such as nonbinary, intersex, or transgender persons.

Discussion

Main Results

The goals of the study were to collect information on the comprehensibility, attractiveness, authenticity, and usability of websites for suicide prevention for men in crisis as well as their gatekeepers. Both websites present psychoeducational resources from experts and patients for both this (vulnerable) group at risk and for their support network. Overall, the findings of this study provide valuable insights into the design and usability of suicide prevention websites addressing men in crisis and their gatekeepers and may provide important implications for future digital prevention interventions.

Potential users of the websites evaluated the balanced mix of emotional content by persons at risk and informative content by scientists as adequate, credible, and helpful. This highlights the potential of this particular combination to address stigmatized topics such as suicidality and mental health and aligns with previous research suggesting that personal narratives combined with information may enhance mental health outcomes and decrease stigmatizing attitudes [31]. However, there were also challenges noted by interview participants, who indicated that they felt overwhelmed with video lengths and the multitude of options to navigate on the websites. Studies on video length have shown that videos with a length of more than 6 minutes can reduce interaction and understanding of learning significantly, thereby acting as a barrier to usage [32]. Based on cognitive load theory and taking into account the limited capacity of active working memory, they recommend using short videos that segment information into small blocks [33]. In addition, based on theories of cognitive capacity, studies on website design show how simplified navigation reduces barriers and prolongs usage [34].

Another key area identified was the need for additional communication channels to signpost to and promote the websites. Users expressed the importance of outreach to raise awareness about the websites and options for individualized and interpersonal communication after using the websites, especially in crisis situations. This aligns with research advocating for multiple communication channels in mental health interventions [35].

In sum, our study shows that both websites are perceived as trustworthy and reliable sources for male suicide prevention. Based on the perceived shortcomings of the programs and participants' recommendations, the study also provides important implications and improvement strategies for future interventions. This includes communicating brief and concise information as well as integrating options for more immediate emergency support. Overall, our findings underline the need for comprehensible, well-structured digital information which provides a balance between information richness and clarity. For both gatekeepers and the men at risk themselves, practical advice should be incorporated into communication interventions, and shorter content formats should be preferred, particularly considering video length. While digital interventions are undoubtedly crucial, our study also reveals the need to embed them in offline contexts and accompany them using other communication channels.

Limitations

Some limitations of the studies must be noted.

Due to the recruitment strategy used, we had a sample of very open, highly involved, and topic-aware participants who occasionally saw themselves as ambassadors and rated the social relevance of suicide prevention as very important. This means that emphasis on education and destigmatization related to this topic was already important for them. Thus, our results cannot be held to be transferable to people who may have internal barriers and fears about engaging with the topic. Furthermore, our cross-sectional studies focused on short-term effects and website evaluations. A longitudinal study would be needed to evaluate whether the websites are a helpful tool in the long run [36].

With regard to male gender and identification itself, we made broad generalizations, as it was our goal to create information platforms that were as universal and accessible as possible by as many men as possible. However, this also meant that specific challenges and burdens of people identifying as male and related subgroups (such as homosexual men, transgender men, and men from minority ethnic groups) were not specifically taken into account. Here it is pertinent that members of the LGBTQIA+ (lesbian, gay, bisexual, transgender, queer/questioning, intersex, asexual, and others) community are much more prone to suicidal thoughts and mental health problems than the population at large [37]. Developing prevention campaigns and information tools specifically tailored for them is thus of paramount importance. The same applies to other minorities, such as immigrants, non-traditional German-born individuals, older individuals, living alone, or those with wider mental health issues. Future research is needed that examines interventions for those vulnerable groups more closely and works with more diverse samples and communities.

XSL•FO
RenderX

Finally, our studies did not focus on creating an intervention directed at changing masculine tropes. This could be worthwhile, as studies suggest that social media campaigns can trigger conversations about masculinity [38,39], potentially questioning behavior driven by traditional masculine norms and value sets that aggravate the risk of (mental health) distress and negative experiences.

Conclusion

Both an RCT of the gatekeeper website [11] as well as this qualitative evaluation indicated that the program was positively evaluated by the participants, in the sense that it was seen as comprehensible and informative. The RCT [11] reported that the training of gatekeepers through a relatively short digital training program is effective based on the knowledge gained. In addition, the qualitative research demonstrates that both types of videos included in the program seem to be relevant for knowledge gain. While the videos featuring real individuals constitute important figures for personal identification and validation, supporting emotional health and self-esteem, the expert videos provide relevant and credible information. Such websites can, therefore, support gatekeepers in suicide prevention. However, it should not be overlooked that such interventions could be perceived to function as "tasking" the gatekeeper to manage the situation, thereby placing a specific burden on gatekeepers, who might feel even more responsible for the suicidal ideation and behavior.

Future studies in this area should also develop guidelines for population-specific communication with men at risk as well as gatekeepers.

Findings from brainstorming ideas beyond the tested interventions highlighted the need to integrate such low-barrier digital offers as one step in a strategy for more comprehensive gatekeeper support. Websites for men and gatekeepers can only fulfill one role in this strategy, with regard to national suicide prevention measures, and must be supplemented by other measures, such as a regional care network for mental health, architectural measures at high-risk sites such as overpasses, and helplines provision [4].

Acknowledgments

Funding for the study was provided by the research project "MEN-ACCESS: Suicide prevention for men" grant (95F12003D) from the "GKV Bündnis für Gesundheit" in Germany.

Data Availability

The datasets generated or analyzed during this study, that is, the transcripts of interviews and focus groups are not publicly available due to data protection reasons but specific portions of the data that are sufficiently anonymized may be obtained from the corresponding author on reasonable request.

Authors' Contributions

DR wrote the manuscript, acquired the funding, and administered the project and the methodology. AW reviewed and edited the manuscript. LH and MF conducted and analyzed the interviews. All authors reviewed the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Recruitment material.
[PDF File (Adobe PDF File), 165 KB - formative_v9i1e59829_app1.pdf]

Multimedia Appendix 2
Consent form.
[PDF File (Adobe PDF File), 564 KB - formative_v9i1e59829_app2.pdf]

Multimedia Appendix 3
Protocols for interviews and focus groups.
[PDF File (Adobe PDF File), 132 KB - formative_v9i1e59829_app3.pdf]

Multimedia Appendix 4
Guide focus groups.
[PDF File (Adobe PDF File), 288 KB - formative_v9i1e59829_app4.pdf]

Multimedia Appendix 5
Guide interviews.

XSL•FO
RenderX

[PDF File (Adobe PDF File), 252 KB - formative_v9i1e59829_app5.pdf]

References

1. Schelhase T. [Suicides in Germany: results from the official cause of death statistics]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2022 Jan;65(1):3-10 [FREE Full text] [doi: 10.1007/s00103-021-03470-2] [Medline: 34962579]
2. Hamilton E, Klimes-Dougan B. Gender differences in suicide prevention responses: implications for adolescents based on an illustrative review of the literature. Int J Environ Res Public Health 2015;12(3):2359-2372. [doi: 10.3390/ijerph120302359] [Medline: 25711358]
3. Krysinska K, Batterham PJ, Christensen H. Differences in the effectiveness of psychosocial interventions for suicidal ideation and behaviour in women and men: a systematic review of randomised controlled trials. Arch Suicide Res 2017;21(1):12-32. [doi: 10.1080/13811118.2016.1162246] [Medline: 26983580]
4. Struszczyk S, Galdas PM, Tiffin PA. Men and suicide prevention: a scoping review. J Ment Health 2019;28(1):80-88. [doi: 10.1080/09638237.2017.1370638] [Medline: 28871841]
5. Till B, Niederkrotenthaler T. Surfing for suicide methods and help: content analysis of websites retrieved with search engines in austria and the United States. J Clin Psychiatry 2014;75(8):886-892. [doi: 10.4088/JCP.13m08861] [Medline: 25099284]
6. Franco-Martín MA, Muñoz-Sánchez JL, Sainz-de-Abajo B, Castillo-Sánchez G, Hamrioui S, de la Torre-Díez I. A systematic literature review of technologies for suicidal behavior prevention. J Med Syst 2018;42(4):71. [doi: 10.1007/s10916-018-0926-5] [Medline: 29508152]
7. Balcombe L, De Leo D. The potential impact of adjunct digital tools and technology to help distressed and suicidal men: an integrative review. Front Psychol 2021;12:796371 [FREE Full text] [doi: 10.3389/fpsyg.2021.796371] [Medline: 35058855]
8. Pirkis J, Rossetto A, Nicholas A, Ftanou M, Robinson J, Reavley N. Suicide prevention media campaigns: a systematic literature review. Health Commun 2019;34(4):402-414. [doi: 10.1080/10410236.2017.1405484] [Medline: 29190128]
9. Wagner A, Reifegerste D. Desirable and adverse effects of communicative suicide prevention interventions among men. Crisis 2024;45(5):365-375. [doi: 10.1027/0227-5910/a000965] [Medline: 39136159]
10. Yonemoto N, Kawashima Y, Endo K, Yamada M. Gatekeeper training for suicidal behaviors: a systematic review. J Affect Disord 2019;246:506-514. [doi: 10.1016/j.jad.2018.12.052] [Medline: 30599375]
11. Hofmann L, Wagner B. Efficacy of an online gatekeeper program for relatives of men at risk of suicide - a randomized controlled trial. BMC Public Health 2024;24(1):2693 [FREE Full text] [doi: 10.1186/s12889-024-20193-6] [Medline: 39358752]
12. Vogel DL, Wade NG, Wester SR, Larson L, Hackler AH. Seeking help from a mental health professional: the influence of one's social network. J Clin Psychol 2007;63(3):233-245. [doi: 10.1002/jclp.20345] [Medline: 17211874]
13. Hofmann L, Wagner B. Men's behavior and communication in the days prior to a suicide-a psychological autopsy study. Int J Environ Res Public Health 2023;20(17):6668 [FREE Full text] [doi: 10.3390/ijerph20176668] [Medline: 37681808]
14. Spahn C, Kranz S, Teismann T, Forkmann T, Schönfelder A, Rath D, et al. Gender-specific aspects of suicide-related communication in a high risk sample of psychiatric inpatients. JOMH 2022;18(9):188 [FREE Full text] [doi: 10.31083/j.jomh1809188]
15. Zalsman G, Hawton K, Wasserman D, van Heeringen K, Arensman E, Sarchiapone M, et al. Suicide prevention strategies revisited: 10-year systematic review. Lancet Psychiatry 2016;3(7):646-659. [doi: 10.1016/S2215-0366(16)30030-X] [Medline: 27289303]
16. Gould MS, Cross W, Pisani AR, Munfakh JL, Kleinman M. Impact of applied suicide intervention skills training on the national suicide prevention lifeline. Suicide Life Threat Behav 2013;43(6):676-691 [FREE Full text] [doi: 10.1111/sltb.12049] [Medline: 23889494]
17. Ghoncheh R, Gould MS, Twisk JW, Kerkhof AJ, Koot HM. Efficacy of adolescent suicide prevention e-learning modules for gatekeepers: a randomized controlled trial. JMIR Ment Health 2016;3(1):e8 [FREE Full text] [doi: 10.2196/mental.4614] [Medline: 26825006]
18. Bartgis J, Albright G. Online role-play simulations with emotionally responsive avatars for the early detection of native youth psychological distress, including depression and suicidal ideation. Am Indian Alsk Native Ment Health Res 2016;23(2):1-27. [doi: 10.5820/aian.2302.2016.1] [Medline: 27115130]
19. Lamis DA, Underwood M, D'Amore N. Outcomes of a suicide prevention gatekeeper training program among school personnel. Crisis 2017;38(2):89-99. [doi: 10.1027/0227-5910/a000414] [Medline: 27561223]
20. Mann JJ, Michel CA, Auerbach RP. Improving suicide prevention through evidence-based strategies: a systematic review. Am J Psychiatry 2021;178(7):611-624 [FREE Full text] [doi: 10.1176/appi.ajp.2020.20060864] [Medline: 33596680]
21. Silva AG, Caravau H, Martins A, Almeida AMP, Silva T, Ribeiro Ó, et al. Procedures of user-centered usability assessment for digital solutions: scoping review of reviews reporting on digital solutions relevant for older adults. JMIR Hum Factors 2021;8(1):e22774 [FREE Full text] [doi: 10.2196/22774] [Medline: 33439128]

XSL•FO
RenderX

22. Hofmann L, Glaesmer H, Wagner B, Spahn C. Männer in suizidalen krisen: [Men in suicidal crises: Approaches to gender-specific suicide prevention and psychosocial care]. Psychotherapeutenjournal 2023:271-278.
23. Jenkins EL, Ilicic J, Barklamb AM, McCaffrey TA. Assessing the credibility and authenticity of social media content for applications in health communication: scoping review. J Med Internet Res 2020;22(7):e17296 [FREE Full text] [doi: 10.2196/17296] [Medline: 32706675]
24. Kim Y. Trust in health information websites: a systematic literature review on the antecedents of trust. Health Informatics J 2016;22(2):355-369 [FREE Full text] [doi: 10.1177/1460458214559432] [Medline: 25518944]
25. Hennink M, Kaiser BN. Sample sizes for saturation in qualitative research: a systematic review of empirical tests. Soc Sci Med 2022;292:114523 [FREE Full text] [doi: 10.1016/j.socscimed.2021.114523] [Medline: 34785096]
26. Evans SC, Roberts MC, Keeley JW, Blossom JB, Amaro CM, Garcia AM, et al. Vignette methodologies for studying clinicians' decision-making: validity, utility, and application in ICD-11 field studies. Int J Clin Health Psychol 2015;15(2):160-170 [FREE Full text] [doi: 10.1016/j.ijchp.2014.12.001] [Medline: 30487833]
27. Evans N, Whitcombe S. Using circular questions as a tool in qualitative research. Nurse Res 2016;23(3):26-29. [doi: 10.7748/nr.23.3.26.s6] [Medline: 26793984]
28. Harst L, Lantzsch H, Scheibe M. Theories predicting end-user acceptance of telemedicine use: systematic review. J Med Internet Res 2019;21(5):e13117 [FREE Full text] [doi: 10.2196/13117] [Medline: 31115340]
29. Braun V, Clarke V. Toward good practice in thematic analysis: avoiding common problems and be(com)ing a researcher. Int J Transgend Health 2023;24(1):1-6 [FREE Full text] [doi: 10.1080/26895269.2022.2129597] [Medline: 36713144]
30. Boyatzis RE. Transforming Qualitative Information: Thematic Analysis and Code Development. Thousand Oaks: Sage Publications, Inc; 1998.
31. Rennick-Egglestone S, Morgan K, Llewellyn-Beardsley J, Ramsay A, McGranahan R, Gillard S, et al. Mental health recovery narratives and their impact on recipients: systematic review and narrative synthesis. Can J Psychiatry 2019;64(10):669-679 [FREE Full text] [doi: 10.1177/0706743719846108] [Medline: 31046432]
32. Afify MK. Effect of interactive video length within e-learning environments on cognitive load, cognitive achievement and retention of learning. Turkish Online Journal of Distance Education 2020:68-89. [doi: 10.17718/tojde.803360]
33. Slemmons K, Anyanwu K, Hames J, Grabski D, Mlsna J, Simkins E, et al. The impact of video length on learning in a middle-level flipped science setting: implications for diversity inclusion. J Sci Educ Technol 2018;27(5):469-479. [doi: 10.1007/s10956-018-9736-2]
34. Rotondi AJ, Spring MR, Hanusa BH, Eack SM, Haas GL. Designing eHealth applications to reduce cognitive effort for persons with severe mental illness: page complexity, navigation simplicity, and comprehensibility. JMIR Hum Factors 2017;4(1):e1 [FREE Full text] [doi: 10.2196/humanfactors.6221] [Medline: 28057610]
35. Sampogna G, Bakolis I, Evans-Lacko S, Robinson E, Thornicroft G, Henderson C. The impact of social marketing campaigns on reducing mental health stigma: results from the 2009-2014 time to change programme. Eur Psychiatry 2017;40:116-122 [FREE Full text] [doi: 10.1016/j.eurpsy.2016.08.008] [Medline: 27997875]
36. Holmes G, Clacy A, Hermens DF, Lagopoulos J. The long-term efficacy of suicide prevention gatekeeper training: a systematic review. Arch Suicide Res 2021;25(2):177-207. [doi: 10.1080/13811118.2019.1690608] [Medline: 31809659]
37. Haas AP, Eliason M, Mays VM, Mathy RM, Cochran SD, D'Augelli AR, et al. Suicide and suicide risk in lesbian, gay, bisexual, and transgender populations: review and recommendations. J Homosex 2011;58(1):10-51 [FREE Full text] [doi: 10.1080/00918369.2011.534038] [Medline: 21213174]
38. Scotti Requena S, Pirkis J, Currier D, Conway M, Lee S, Turnure J, et al. An evaluation of the boys do cry suicide prevention media campaign on twitter: mixed methods approach. JMIR Form Res 2023;7:e49325 [FREE Full text] [doi: 10.2196/49325] [Medline: 37676723]
39. Schlichthorst M. Engaging men in conversations about masculinity and suicide? An evaluation of the Man Up social media campaign. 2020. URL: https://minerva-access.unimelb.edu.au/bitstream/handle/11343/265962/316ab37c-af07-eb11-94cb-0050568d7800_mphil%20thesis-mschlichthorst-final.pdf [accessed 2025-01-23]

Abbreviations

LGBTQIA+: lesbian, gay, bisexual, transgender, queer/questioning, intersex, asexual, and others
RCT: randomized controlled trial

XSL•FO
RenderX

Edited by A Mavragani; submitted 23.04.24; peer-reviewed by J Arroyave-Cabrera, V Carli, P Stas; comments to author 16.10.24; revised version received 25.11.24; accepted 21.12.24; published 26.02.25.

Please cite as:
Reifegerste D, Wagner AJM, Huber L, Fastuca M
Formative Evaluation of Suicide Prevention Websites for Men: Qualitative Study with Men at Risk of Suicide and with Potential Gatekeepers
JMIR Form Res 2025;9:e59829
URL: https://formative.jmir.org/2025/1/e59829
doi:10.2196/59829
PMID:

©Doreen Reifegerste, Anna J M Wagner, Lisa Huber, Manuel Fastuca. Originally published in JMIR Formative Research (https://formative.jmir.org), 26.02.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Automated Digital Safety Planning Interventions for Young Adults: Qualitative Study Using Online Co-design Methods

Jonah Meyerhoff[1], PhD; Sarah A Popowski[1], BA; Tanvi Lakhtakia[1], BA; Emily Tack[1], BA; Rachel Kornfield[1], PhD; Kaylee P Kruzan[1], LCSW, PhD; Charles J Krause[1], MPH; Theresa Nguyen[2], LCSW; Kevin Rushton[2], BA; Anthony R Pisani[3,4], PhD; Madhu Reddy[5], PhD; Kimberly A Van Orden[3], PhD; David C Mohr[1], PhD

[1]Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University, Chicago, IL, United States

[2]Mental Health America, Alexandria, VA, United States

[3]Department of Psychiatry, School of Medicine and Dentistry, University of Rochester, Rochester, NY, United States

[4]Department of Pediatrics, School of Medicine and Dentistry, University of Rochester, Rochester, NY, United States

[5]Department of Informatics, Donald Bren School of Information and Computer Sciences, University of California, Irvine, CA, United States

Corresponding Author:
Jonah Meyerhoff, PhD
Department of Preventive Medicine
Feinberg School of Medicine
Northwestern University
750 North Lake Shore Drive
Chicago, IL, 60611
United States
Phone: 1 3125031232
Email: jonah.meyerhoff@northwestern.edu

Abstract

Background: Young adults in the United States are experiencing accelerating rates of suicidal thoughts and behaviors but have the lowest rates of formal mental health care. Digital suicide prevention interventions have the potential to increase access to suicide prevention care by circumventing attitudinal and structural barriers that prevent access to formal mental health care. These tools should be designed in collaboration with young adults who have lived experience of suicide-related thoughts and behaviors to optimize acceptability and use.

Objective: This study aims to identify the needs, preferences, and features for an automated SMS text messaging–based safety planning service to support the self-management of suicide-related thoughts and behaviors among young adults.

Methods: We enrolled 30 young adults (age 18-24 years) with recent suicide-related thoughts and behaviors to participate in asynchronous remote focus groups via an online private forum. Participants responded to researcher-posted prompts and were encouraged to reply to fellow participants—creating a threaded digital conversation. Researcher-posted prompts centered on participants' experiences with suicide-related thought and behavior-related coping, safety planning, and technologies for suicide-related thought and behavior self-management. Focus group transcripts were analyzed using thematic analysis to extract key needs, preferences, and feature considerations for an automated SMS text messaging–based safety planning tool.

Results: Young adult participants indicated that an automated digital SMS text message–based safety planning intervention must meet their needs in 2 ways. First, by empowering them to manage their symptoms on their own and support acquiring and using effective coping skills. Second, by leveraging young adults' existing social connections. Young adult participants also shared 3 key technological needs of an automated intervention: (1) transparency about how the intervention functions, the kinds of actions it does and does not take, the limits of confidentiality, and the role of human oversight within the program; (2) strong privacy practices—data security around how content within the intervention and how private data created by the intervention would be maintained and used was extremely important to young adult participants given the sensitive nature of suicide-related data; and (3) usability, convenience, and accessibility were particularly important to participants—this includes having an approachable and engaging message tone, customizable message delivery options (eg, length, number, content focus), and straightforward menu navigation. Young adult participants also highlighted specific features that could support core coping skill acquisition (eg, self-tracking, coping skill idea generation, reminders).

XSL•FO
RenderX

Conclusions: Engaging young adults in the design process of a digital suicide prevention tool revealed critical considerations that must be addressed if the tool is to effectively expand access to evidence-based care to reach young people at risk for suicide-related thoughts and behaviors. Specifically, automated digital safety planning interventions must support building skillfulness to cope effectively with suicidal crises, deepening interpersonal connections, system transparency, and data privacy.

(JMIR Form Res 2025;9:e69602) doi:10.2196/69602

KEYWORDS

mental health services; technology; therapy; computer assisted; SMS text messaging

Introduction

Background

The prevalence of suicidal thoughts among young adults, those between the ages of 18 and 25 years, in the United States has increased by approximately 160% since 2013 [1,2]. While most young adults who experience suicidal thoughts do not go on to make a suicide attempt, suicidal thoughts are indicators of risk that warrant attention and intervention [3]. Among adults who experienced mental health symptoms in the last year, young adults have the lowest rates of formal mental health care treatment [2]. Structural (eg, cost, accessibility, and provider availability) and attitudinal (eg, stigma, strong preferences for self-management) barriers are particularly important for determining help-seeking behavior [4] and contribute to low rates of mental health treatment engagement [4-6].

Current approaches to suicide prevention underserve young adults facing structural or attitudinal barriers to mental health services. Most prevention efforts are delivered in the context of mental health interventions via formal outpatient, inpatient, emergency department, and other clinical care settings. Prevention is also delivered in informal settings such as crisis helplines, targeting individuals before suicidal crises arise (eg, in schools or workplaces) in an effort to bring suicide prevention approaches to those who need it in places they are most likely to access it [7]. As most young adults with suicide-related thoughts and behaviors do not engage in formal mental health treatment, where most evidence-based suicide prevention is delivered [8-15], and they experience strong preferences for self-management, young adults are less likely to encounter evidence-based suicide prevention interventions.

The high need for preventive interventions and low formal mental health treatment engagement among young adults experiencing suicidal thoughts and behaviors [1,2] signal a critical gap in suicide prevention infrastructure. Digital suicide prevention interventions are an important avenue by which access to evidence-based suicide prevention care can be increased. Digital suicide prevention interventions offer the benefit of being able to circumvent many of the attitudinal and structural barriers that prevent young adults from accessing evidence-based preventive interventions [4-6]. They can be accessed on young adults' devices (eg, smartphones) and used privately.

The safety planning intervention (SPI) [16], a brief intervention aimed at interrupting the progression from suicidal ideation to suicidal behavior by increasing coping ability, has typically been administered collaboratively by clinicians in clinical care settings across ages [17-19] and is considered standard of care [20,21]. The SPI typically includes 6 areas of focus: warning signs, coping skills, people for distraction, people for help, professional organizations and resources, and environmental safety [16]. The product of the SPI is often a brief document with personalized lists of resources and coping activities for each of the 6 focus areas. When delivered with fidelity, clinicians ask patients targeted questions and follow-up on patients' responses to elicit highly personalized information that can enhance the efficacy of an individual's safety plan by making it as relevant as possible to the individual's life circumstances [22]. However, safety planning is frequently delivered by individuals facing barriers to implementation including insufficient training and time [23,24], which poses a risk to safety plan quality and efficacy.

The SPI is one of the most common digital suicide prevention intervention elements [25]. There is an emerging literature on automated digital SPIs that guide users through the creation of a highly personalized safety plan [25]. Several have been shown to be acceptable [26-29] and even demonstrate that resultant safety plan quality can be comparable to collaboratively developed safety plans [30,31]. Extant literature on digital SPIs primarily focuses on broad adult samples and narrow adolescent populations [25]. Boudreaux et al [26] piloted an automated, computer-based SPI that enabled adults who experienced suicide-related thoughts and behaviors to complete a safety plan and offered brief guidance on how to complete the safety plan. The resultant safety plan was a 1-page PDF that users could download, access on different devices, and share with trusted others. Melvin et al [27] and Rainbow et al [29] developed a mobile app–based approach called BeyondNow that includes a customizable SPI with prepopulated suggestions for each of the 6 safety planning steps and offers the ability to edit or share the completed safety plan and a quick access button to a crisis helpline. The app is publicly accessible and has strong acceptability metrics and is associated with promising improvements in suicide-related coping. Finally, Methi et al [28] created an automated web-based e-SPI specifically designed for adolescents that includes age-appropriate instructions, examples, and illustrations that assist users in completing the online safety plan via text-fill boxes. The content adolescents add is converted to an online safety card that users can save. These efforts highlight that automated safety planning is feasible to implement and that there are opportunities to engage individuals in new ways, expanding beyond single-session safety planning and crafting safety plans via mediums that are more conversational in nature.

XSL•FO
RenderX

Among this emerging literature, several key needs and preferences have surfaced. First, having quick access to information such as one's documented safety plan, crisis resources, or personal contact information can be extremely valuable during a crisis as crisis states make it difficult to navigate complex user interfaces [32,33]. Second, there is a favorable view of digital artifacts (eg, photos, songs) that inspire hope and increase motivation for living [27]. Third, across the literature, users of digital SPIs have ambivalence about whether or how they want to involve others in their personal networks in the creation or use of their safety plans [32,33]. One way early digital SPIs have addressed this issue is through the use of optional prescripted messages users can generate when outside of crisis states and send to trusted others to elicit social or instrumental support [32]. Fourth, extant literature highlights that strong privacy practices emerge as critical priorities for users of digital safety plans [33,34]. Fifth, the ability to download and edit created safety plans after their initial creation is critical for ensuring safety plans remain relevant and accessible over time [27,34,35]. Finally, using validating language that minimizes assumptions about a user (eg, that they have a network of others that can provide support during a crisis) and ensuring the digital SPI can provide examples or menus of items if a user has difficulty generating safety planning content independently can prove beneficial [35].

The vast majority of digital SPIs are mobile app or browser based [25,36], meaning that individuals must visit a specific app or website in order to interact with or create their safety plan. In this study, we aimed to understand the potential of an SMS text messaging–based safety plan creation approach. Texting has the technological affordance of being a conversational medium, potentially addressing some of the shortcomings of automated digital safety plans that are documentation only. By eliciting information relevant to safety planning in a conversational manner, an automated texting–based SPI may be able to increase the personalization and relevance of digital safety plans. SMS text messaging is the most commonly used communication medium and young adults are the age group sending the greatest number of SMS text messages [37,38]. It is more widely used than social media apps [39]. Therefore, an SMS text messaging–based SPI is capable of meeting young adults in a space where they are already spending a significant amount of time. Indeed, prior work has shown that texting interventions for suicide prevention have been acceptable to users [40-45].

The relatively sparse existing literature on the needs and preferences of digital SPI users is focused on the app- or web-based designs. Further, it lacks focus on the needs of the adult age group with the highest rates of suicide-related thoughts and behaviors, but the lowest rates of treatment-seeking: young adults. Moreover, there is a paucity of literature regarding automated, interactive, texting interventions to support suicide prevention, broadly, and safety planning, more specifically.

Objective

In this study, we aimed to understand the needs, preferences, constraints, and design considerations for an SMS text message–based automated safety planning tool for young adults who have lived experience of suicide-related thoughts and behaviors. Informed by prior literature, our work was guided by the following research questions: (1) What are the broad needs and preferences of young adults who experience suicide-related thoughts and behaviors for suicide prevention? (2) What are the features that can support young adults with suicide-related thoughts and behaviors to create an effective safety plan?

Methods

Participants

A total of 30 participants between the ages of 18 and 24 years were enrolled in this study (Table 1). Individuals were eligible to participate if they resided in the United States, were between the ages of 18 and 24 years, owned a smartphone, were proficient in English, and endorsed past 2-week suicidal ideation. Suicidal ideation was assessed using a modified item from the DSM-5 (Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition) Cross-Cutting Symptom Inventory (DSM-5-CCSI) [46], "During the past 2 weeks, how often have you been bothered by thoughts of actually hurting yourself [or ending your life]?" Disambiguation of suicidal thoughts from nonsuicidal self-injurious thoughts was assessed via the Depression Severity Index—Suicidality Subscale (DSI-SS) [47], a 4-item self-report measure assessing the frequency and intensity of suicidal thinking. Participants were not eligible to participate if they endorsed symptoms consistent with mania or psychosis in the last 2 weeks (assessed via the DSM-5-CCSI) or the presence of strong suicidal desire assessed via item 9 on the Beck Depression Inventory—Second Edition [48] and a subsequent score of 2 or 3 (range 0-3).

XSL•FO
RenderX

Table 1. Participant characteristics.

Characteristics	Values
Age, mean (SD)	21.1 (1.9)
Gender, n (%)	
Male	3 (10)
Female	22 (73)
Nonbinary	5 (17)
Transgender, n (%)	4 (13)
Race, n (%)	
Black/African American	3 (10)
American Indian/Alaska Native	1 (3)
Asian	5 (17)
Native Hawaiian/Pacific Islander	0 (0)
White	18 (60)
More than one race	2 (7)
Prefer not to answer	1 (3)
Hispanic/Latino, n (%)	3 (10)
Total DSI-SS[a] score (4 items, with each coded between 0 and 3; range 0-12), mean (SD)	4.8 (1.6)
Ever seen a therapist or mental health professional, n (%)	25 (83)
Currently seeing a therapist or mental health professional, n (%)	11 (37)
Plan to see a therapist or mental health professional in the next 8 weeks, n (%)	17 (57)
Never seen a therapist or mental health professional and does not plan to see one in the next 8 weeks, n (%)	1 (3)

[a]DSI-SS: Depression Severity Index—Suicidality Subscale.

Procedures

Individuals were recruited via an advertisement embedded on Mental Health America's (MHA) website. The advertisement was shown to individuals who visited a resource page about suicidal thoughts and behaviors, endorsed suicidal or self-harm ideation in the past 2 weeks on MHA's online screening website, or endorsed past-month suicidal ideation on MHA's nonsuicidal self-injury survey. Individuals who clicked the ad were directed to a Northwestern University–hosted informational web page about the study. Interested individuals could elect to provide informed consent for eligibility screening via Northwestern University's instance of REDCap [49,50] and were subsequently routed to the eligibility screening surveys.

If eligible, interested individuals could review the REDCap-based study consent documentation and the code of conduct for engaging with research staff and fellow participants on the private research forum. The code of conduct detailed guidelines such as refraining from posting identifying information or detailed descriptions of self-harm, suicide methods, or behaviors. The code of conduct also overviewed harassment and moderation policies. To be enrolled in the study, individuals needed to provide affirmative informed consent and agree to abide by the code of conduct.

Participants engaged in 1 of 2 asynchronous remote community (ARC) [51-53] groups—akin to an online focus group. We used ARC to increase the engagement of underrepresented populations by decreasing barriers related to synchronous or in-laboratory research tasks [54]. Additionally, the group format was chosen to facilitate conversation between adults with shared experiences. ARC methods are used in the field of human-computer interaction to facilitate conversation, elicit several perspectives simultaneously, and provide the environment for participants to build on each other's creative contributions [51]. ARC groups were held between November and December 2022, or February and March 2023. Two ARC groups were held to increase the sample size and diversity of responses to several lines of questioning and protect against group effects as well as to gather data on lines of inquiry that were not covered in the first group. ARC groups involved participants registering for a private online forum using FocusGroupIt [55] in which researchers and participants could create posts visible for all in the study and with which participants could engage or respond. Each group comprised 15 participants. Past work demonstrates similar sample sizes are sufficient to elicit diverse perspectives [56]. Participants were provided with automatically generated anonymous usernames (eg, participant 1, participant 2). Each ARC group consisted of 8 total prompts, 1 every 3 days. Prompts focused on terminology young adults used for suicidal ideation, prompting events for suicide-related thoughts and behaviors, safety plan use, and technology to support both safety planning and help-seeking.

XSL•FO
RenderX

Research staff monitored forums daily for posts violating the posted code of conduct. Offending posts were deleted, and the original poster was contacted by research staff with an explanation of why the post was removed, and, if needed, to conduct a risk assessment. Only 1 such post was removed and there was no indication of suicide-risk elevation for the poster.

Data Analysis

The corpus for this study included 3 prompts from each group containing relevant data to answering the primary research questions (see Multimedia Appendix 1). These prompts comprising the corpus covered perceptions of safety planning via SMS text messages (group 1, prompt 4; group 2, prompt 3), the role of automated messages in supporting safety planning (group 1, prompt 5; group 2, prompt 4), use of a safety plan after it is created (group 1, prompt 7), and the role of crisis services in suicide prevention care (group 2, prompt 7).

Forum transcripts were analyzed using Thematic Analysis [57,58] to extract key needs, preferences, and design considerations for a digital safety planning tool. Thematic analysis was facilitated using Dedoose software (Version 9.0.9) [59]. One coder (JM) read through the corpus and generated an initial codebook via open coding responses to each of the 3 prompts from each ARC group. Coding occurred at the sentence level within a participant's post. Two additional coders (SAP and TL) read through the entire corpus to familiarize themselves with the data and then applied the initial codebook to overlapping transcripts. All 3 coders then met to revise the initial codebook. Discrepancies were discussed until a consensus was reached among all coders. Importantly, the analysis software, Dedoose, facilitates the attribution of code applications to the research staff who coded the data, enabling the ability to identify and resolve coding discrepancies. All coders then divided the remaining data and applied the revised codebook so that the entire corpus was coded by at least two coders. Coders met 3 more times to discuss and resolve discrepancies. The saturation of codes was determined to be reached once all discrepancies in codes were resolved and no additional codes were identified in the overlapping transcripts. Coders then met to organize codes into axial themes iteratively.

Positionality Statement

At the time of this study, JM was a clinical psychologist and faculty member with a background in both clinical and research work in suicide prevention. He held funded grants focused on increasing access to care through the design, development, and evaluation of digital mental health interventions. SAP was a research assistant who supported the administration of the ARC methods described in this paper. Both JM and SAP interacted with participants whose data comprise the study corpus through study procedures such as email and, in some cases, interviews. To maximize objectivity during coding and analysis, all participants were assigned a coded ID number. TL was a research assistant external to the project and had no prior contact with any study participants.

Ethical Considerations

Before beginning any screening or study procedures, all interested individuals and participants in this study provided affirmative informed consent, both for initial eligibility screening and subsequently for participation in the study. Participants were compensated up to US $80 for responding to the study prompts and for commenting on other participants' posts.

Throughout the study, participants' data were, and continue to be, considered confidential and protected. Only research personnel and individuals securing appropriate credentials and data use agreements were able to access study data and complete transcripts. Study transcripts were deidentified and maintained on secure encrypted servers at Northwestern University.

All study procedures were administered at Northwestern University and all study procedures were approved by the Northwestern University Institutional Review Board (project number STU00217191).

Results

Young Adults' Perspectives on Automated Digital SPIs: Treatment Mechanisms, Core Elements, and Skill-Building Features

Across participants in our ARC groups, young adults expressed their need for automated digital SPIs. Young adults discussed (1) the treatment mechanisms an automated digital SPI must address, (2) core elements of the technology that affect the perceived usefulness and acceptability of an automated digital SPI, and (3) key skill-building features they would like to see in such an intervention.

Meeting Young Adults' Needs

Key Needs for SPI Support: Self-Management and Social Connection

We identified 2 key needs that young adults thought an SPI should support. The first need was for self-management and the second was for social connection.

Empowering Young Adults to Manage Their Own Symptoms by Supporting Coping Skill Acquisition and Use

Young adults wanted an automated digital SPI to support the self-management of their suicide-related thoughts and behaviors through coping skills that foster a sense of self-empowerment. Self-empowerment approaches enable young adults to rely on their own skillfulness to effectively cope with their suicide-related thoughts and behaviors, and to avoid the stigma they associated with help-seeking. Some participants already possessed coping skills that facilitated self-empowerment and shared how these coping skills allowed them to navigate future suicidal crises:

> *I have never thought about using one of the crisis services, as I never felt as though I have needed to use one. I think it would be a bit embarrassing for me, as I strongly dislike talking about depression and/or suicide with others. I have also always been able to talk myself down without the need for a crisis service intervention.* [G2-P2]

XSL•FO
RenderX

G2-P2's statement suggests that through practice in talking themselves down, they have been able to mitigate escalating suicidal crises and this experience provides confidence in their own ability. While participants, in general, noted the potential value of a tool to support coping skill use and acquisition, others such as G2-P3 highlighted the potential benefits of an approach that involved other people:

> *I ultimately chose not to use these [crisis helpline] services because I was able to eventually re-ground myself, however, in retrospect, I think it would have been more beneficial to talk to someone about this in the moment rather than going through it alone.*

Young adult participants described the impact they believed an automated SMS text messaging–based safety planning tool could have on their self-empowerment in moments of need. G1-P4 shared that by creating a safety plan with the assistance of a digital tool, an individual is taking the lead in planning for their own coping. They described that one of the "benefits would be reassurance of the feeling of not being alone from something you created." G1-P15 also highlighted the benefits of a self-created safety plan, commenting, "the fact that the individual will be able to change it, and have the power over their safety would probably make a lot of people who use the plan a lot more secure." Overall, young adult participants were clear that an automated safety planning service would need to be able to provide them with the ability to choose how they cope with their symptoms.

Young adult participants shared that one of the motivations for using self-management tools was prior experiences of invalidation when discussing their suicide-related thoughts and behaviors with others. As G2-P12 recalled,

> *After reaching out to a hotline, I had to speak with a campus therapist and she laughed at me for considering suicide because my problems seemed "frivolous" or "overexaggerated". I haven't been able to trust a hotline since, and after that one incident, I've cut myself off to dozens of resources that may help me with the fear they may joke about me in the same manner.*

This experience of stigma, wherein the participant expected help and was instead met with judgment-worsened symptoms sowed distrust in systems and made future help-seeking less appealing, ultimately leading to decreased support and increased social isolation. As a result, young adults noted that it was important to have resources that enabled them to self-manage their symptoms and that were offered in a welcoming and nonjudgmental tone, mitigating the need to disclose suicide-related thoughts and behaviors to others to obtain support. For example, G1-P13 shared how messages could support effective coping, but would need to strike the right tone "I think automated messages like this would be extremely helpful in providing a structured way to create a safety plan. [...] these messages should be welcoming, even if they aren't actually run by a person, and include greetings or transitions."

Participants described how an automated digital SPI delivered via SMS text messaging could foster self-empowerment by providing actionable coping resources in an easily accessible format. G2-P14 shared, "[the intervention] would need to be automated, where it texts my questions and I respond and it can provide me with the appropriate part of my personalized safety plan." G2-P14's comment depicts a scenario in which a user of an automated safety planning service is asked targeted questions that elicit personalized coping resources and then these personalized coping resources are sent back to the user at opportune times, facilitating self-management of suicidal crises and self-efficacy by enabling an individual to use coping resources effectively. This approach was echoed by others, for example, G1-P8 noted "Having the automated messages send me reminders for self-care and other things based on what I write in response to the messages (plus what I already have in my safety plan) would be great as well."

Some participants shared how a digital SMS text messaging–based approach must accommodate change over time to fully support self-management of suicide-related thoughts and behaviors. As an individual's circumstances, resources, and skillfulness change over time, what works for them at the outset will need to be updated to be more relevant and effective. G1-P6 shared that, "So much can change in just a second, and what worked for me yesterday might not work for me today." Editing and maintaining one's safety plan was seen as a critical component of self-empowerment in that it enables a constant cycle of being able to evaluate a safety plan's efficacy against real-world suicide-related thought and behavior and crisis mitigation and management attempts. The ability to edit and maintain a safety plan allows for modification and adaptation in the event of actual or anticipated underperformance so that it can perform better in the future.

Drawing on Social Connections

The second core element many young adults in our study wanted was an automated digital SPI to support self-management of suicide-related thoughts and behaviors by supporting users in connecting with others. Participants had different ideas about how this could be accomplished, but the vast majority agreed that interpersonal connections were critical for any suicide-related thought and behavior intervention. The findings that follow highlight how components that support interpersonal connection may be critical targets in an automated intervention aimed at addressing suicide-related thoughts and behaviors, and could potentially work by creating accountability structures, creating openness to connection with others, fostering a feeling of being cared for, eliciting outside perspectives, and recruiting help when needed.

Creating social accountability was a core function participants said was important to consider for the self-management of suicide-related thoughts and behaviors. Participants noted that it was difficult to commit to using certain coping skills when in a suicidal crisis, but accountability to others would make it more likely for them to use the planned coping skills. One participant highlighted the utility of others to maintain accountability structures:

> *I do feel like a lot of safety plans are made with that individual having to keep themselves accountable in mind. But, part of a safety plan can include other*

XSL•FO
RenderX

> *people/ resources that can help in those times and in some ways keep you accountable and safe.* [G2-P16]

As G2-P16 articulates, the system could encourage the engagement of trusted others in the digital intervention to enhance the efficacy of digital safety planning.

Participants also wanted a digital intervention to facilitate connections with trusted others in their network. G2-P7 put it clearly, "I need human interaction more than anything right now." This participant's perspective highlights that many young adults are aware of the need for connection, especially in the context of crisis states, and that it is necessary to, first, build and, second, reach out to those support networks when needed. G2-P3 stated, "when experiencing feelings of loneliness/isolation alongside suicidal thoughts, it is especially beneficial to have social support mechanisms in place." Young adults were open to an automated digital SPI but emphasized that it needed to also facilitate connection and strengthen ties to trusted others.

Participants imagined that an automated intervention could connect individuals to their support networks and allow young adults to get outside perspectives from mental health professionals or trusted others. When experiencing suicidal thoughts, young adults shared that it is challenging to access perspectives outside of their own, but acknowledged that it can be helpful for diverting or avoiding an escalating suicidal crisis. As a result, participants raised the importance of obtaining outside perspectives:

> *Sometimes it is really hard to notice signs of crisis in yourself, and having something point it out to you could be really helpful. I've noticed in myself that sometimes when I get really worked up I don't notice how dangerous my behavior has become until a loved one points it out to me and expressed their concerns. It's helpful to get an outside perspective.* [G1-P6]

While many participants wanted other people involved in the perspective-taking process, some did not. G2-P7 shared how experiences with past treatment-seeking cast doubt on the utility of involving others to offer outside perspectives, "The best therapy, in my opinion, is taking time to yourself and getting to the root of the issue alone. It was the best thing I could have ever done for myself." However, participants discussed that an automated system should nonetheless prompt an individual to consider different ways of looking at one's circumstances, experiences, and values providing an outside perspective that could serve a similar function without integrating a human in the loop by default. As G2-P3 states, "these messages could serve as gentle reminders of values and the things that are important to people, [...] I would be more likely to respond to these messages if they contained thought-provoking questions that probed my values."

There was a minority of participants who shared that a potential drawback to an automated SMS text messaging–based SPI is that it had the potential to maintain or exacerbate social isolation by reducing a point of potential human contact such as calling or texting with a crisis counselor. For example, G2-P2 shared, "even if it designed for you, the digital aspect of the text is a bit isolating." An SMS text messaging suicide-related thought and behavior intervention may be acceptable for many, but participants cautioned that if the interpersonal considerations were not sufficiently designed for, this type of intervention could cause harm. Specifically noting that automated suicide prevention interventions could be misused in place of a point of human contact (eg, crisis counselor support), potentially worsening subjective experiences of isolation.

Technology Needs

Technological Needs for Automated Digital SPIs: Transparency, Privacy, and Usability

The young adult participants in our study also discussed needs related to the technological aspects of an automated digital SPI. Specifically, they highlighted the importance of (1) transparency or clarity regarding expectations, assumptions, and limitations of automated digital SPI; (2) privacy or the practices, policies, and technologies that limit exposure of personal information; and (3) usability, convenience, and accessibility of the intervention.

The Need for Transparency

Participants detailed the important ways in which being transparent in the context of an intervention for suicide-related thoughts and behaviors is critical to uptake and efficacy. One participant shared how transparency about crisis helpline practices helped them feel more comfortable using a particular crisis helpline:

> *Even just having some ideas of what may be asked of my or what to expect, vocabulary to express my feelings, etc. may help make the experience of asking for help much less scary and more accessible.* [G2-P13]

G2-P13 highlighted how, in the context of crisis helplines, having information about the scripts and procedures counselors follow before making the call can help put callers at ease by facilitating mental preparation. This in turn can reduce uncertainty during a crisis call, reduce fear about disclosing one's suicide-related thoughts and behaviors to a stranger, and increase how readily accessible a helpline is to a prospective caller.

Similarly, in the context of an automated SMS text messaging–based SPI, young adult participants described that information about how the program functioned (eg, how many messages the system sent) and guidelines on the limits of confidentiality would be important when considering whether to use such a service. As G1-P3 articulated, "I have really bad anxiety so I am immediately thinking of well what if I do/say something "wrong" will they have a list of numbers and contact people and tell them." Given the potential real-life consequences of mandated reporting laws, even if relatively rare (eg, involuntary hospitalization, removal from one's home, or physical harm from emergency responders), participants in our study wanted transparency about the kind of information that would result in the automated intervention activating a mandated report. As G1-P3 alludes to, knowing in advance how the system responds to user input and operationalizes mandated reporting

XSL•FO
RenderX

can help ameliorate symptoms of anxiety while being instructive as to how to get the most benefit from the service.

The Role of Strong Privacy Practices

In addition to transparent information about how a suicide-related thought and behavior prevention intervention functions, the underlying assumptions it makes about users, and the procedures to which it adheres, young adult participants also expressed the importance of privacy including (1) autonomy over data and data use, (2) anonymity, and (3) discretion.

Participants wanted to ensure they maintained control over their data. In the context of an SMS text messaging–based SPI, participants were concerned that their messages could be read by another individual, the information they share could be reported to third parties, and that the messages they sent would be identifiable. As G2-P13 shared, “I would be concerned about privacy though. I would want to know that what I shared over text would not be read by another person without my consent, and that what I shared would not be traced back to me or reported to someone.” By contrast, others saw the ability to create a safety plan via SMS text messages as a way to maintain control over how, when, and what the safety plan contained which they viewed as a form of data autonomy.

Another important technical specification for suicide-related thought and behavior-focused interventions was the ability to remain anonymous. As G2-P9 noted, “there is a fear of being traced and how it is not completely anonymous because the phone number is linked to you.” Anonymity was one of the factors that influenced whether or not participants would use a crisis service. For example, despite one participant’s misinterpretation of the National Suicide and Crisis Lifeline’s processes, they noted that a key factor in making the Lifeline an attractive intervention was the ability to remain anonymous (The National Suicide and Crisis Lifeline maintains confidentiality except in cases of imminent risk and they do not enable completely anonymous contacts. Their Information Technology infrastructure and counselors collect callers’ and texters’ personally identifiable information such as phone numbers, IP addresses, or volunteered other information that may be linked back to an individual.).

Relatedly, participants wanted digital or remote suicide-related thought and behavior-related interventions to be discreet. This was especially true when considering the use of a texting–based SPI. Participants were particularly concerned that people in their lives with physical access to their devices, but no or partial knowledge of their suicide-related thoughts and behaviors, might be able to access material that would result in an unwanted disclosure of an individual’s suicide-related thoughts or behaviors. G2-P5 noted, “I relate to the fear of being found out. Even with the promise of anonymity I’m still worried about someone in my life finding out and reporting me.” Another participant, G2-P12, contextualized this well-represented fear around a digital intervention, “I definitely would also be nervous someone could see it on accident, because that's definitely happened to me before where I opened my phone and people I was eating lunch with saw what I had recently seen.” While participants expressed general concerns around access to SMS text message data (eg, G1-P3 “I think the privacy is a big concern.”, G1-P2 “I think many people feel vulnerable about their mental state and want it treated with as much privacy as possible.”), most of the specific concerns about discretion and privacy focused on people in one’s day-to-day environment gaining access to information and how that disclosure would negatively affect an individual’s relationships. As G1-P6 explains, “most texting apps on phones don’t have locks or passwords as far as I know, and if someone got a hold of the person’s phone they would have immediate access to intimate and private details about the texter’s mental health.”

Usability, Convenience, and Accessibility

Participants shared a number of technical considerations related to an automated digital SPI. Among them was the importance of having a highly functional user interface. Participants were already familiar with the texting medium but noted that text-based menu navigation, relative to graphical user interfaces, could become tedious and frustrating. For example, G1-P8 said, “Text messages have a very specific, not customizable format. For something like a safety plan, I'd much prefer an app or website where I could work through my plan on a page or multiple pages, however that would look and be designed.” However, participants largely agreed that texting–based safety planning could offer an important level of convenience and accessibility that may be especially important for self-management of suicide-related thoughts and behaviors. For example, G2-P9 noted, “The benefits is that it is easily accessible though and your phone is everywhere with you so if you need to refer back to it you will have it with you.”

Participants also noted that an SMS text messaging–based service can make safety planning accessible to individuals that experience anxiety speaking to others about their concerns, in ways that synchronous voice call support cannot. G1-P1 shared, “I think the idea of texting gives me much less anxiety than having to call.” Therefore, specific communication channels may provide different levels of comfort, based on participants’ associations and past experiences.

Participants preferred tools that were efficient to use and did not require significant time or effort. This was especially true for participants who imagined using an automated digital SPI while in crisis. Further, if a program requires significant effort on the part of the individual, it could be demotivating and unhelpful. As G1-P13 shared, “I would be more likely to respond to these messages if they didn't take a lot of effort to respond to--if they involved a long, inconvenient chain of messages, for example, I wouldn't feel motivated to respond.” It is critical that messages are focused and direct, but that they also have some semblance of personality. Participants like G1-P11 shared how they wanted messages styled, “The messaging dynamic should be concise but not totally devoid of feeling, even if it is totally automated.” Participants wanted messages that made interacting with a tool interesting without being overly demanding.

Skill Building Features

Beyond intervention and technological requirements, participants in our study shared numerous preferences and feature suggestions for an automated digital SPI that helped

XSL•FO
RenderX

support skill building, not simply skill documentation—extending the role and function of a safety plan. For example, participants shared that they wanted self-tracking features so that they could identify patterns in their suicide-related thoughts and behaviors and build insights over time. G1-P4 shared one way this feature could be integrated into the larger automated digital SPI:

> *it could work in conjunction with tools where you can mark your feelings and behaviors, and once you know how you feel prior to an episode, it could help identify when you may potentially be feeling depressed or experiencing suicidal thoughts or behaviors.*

Others wanted suggestions for coping skills. This augments traditional safety planning by making automated digital SPIs a coping idea generation and refinement system, hosting a repository for useful coping skills that might spark inspiration for a user, especially if that user is having difficulty generating coping skills on their own. One participant shared how this might work:

> *I think that using specific questions to personalize someone's safety plan could be really useful. One big problem I have with the general concept of safety plans is that I feel that I wouldn't know what specific things to add to it, this would definitely solve that issue. I think specificity would definitely help me answer or respond to those messages, I feel less motivated to answer messages from automated services where the prompt is super vague or not very engaging. Other text messages that would be helpful to receive could be resources to distract or help cope with difficult situations. Another way the service could ask for things to incorporate into the plan without directly asking could be to ask for the user's interests or things that might help them calm down.* [G1-P11]

Relatedly, some participants wanted reminders to use coping skills. While safety plans have traditionally been a static document, the translation to an automated digital SPI enables the system to prompt the user to assess the need to use coping skills through periodic reminders. As G1-P2 shared, “I personally feel like getting reminders to relax and check in with myself would be helpful, since suicidal thoughts can be brought on by stress.”

Taken together, participants shared that through reminders, skill suggestions, and self-tracking features, an automated digital SPI presents an opportunity to build skillfulness and insight for a potential user, rather than simply documenting effective coping skills.

Discussion

Principal Findings and Design Implications

This study aimed to understand the needs, preferences, and design requirements and features necessary for an automated SMS text messaging–based suicide prevention intervention. Our results suggest that young adults who experience suicide-related thoughts and behaviors are open to embedding suicide prevention interventions within SMS text messages, but these interventions must be designed to meet their unique needs and preferences.

One of the important needs arising from our ARC groups was the importance of social connection and accountability in automated digital SPI design. Young adults who experienced suicide-related thoughts and behaviors were interested in interventions that strengthen existing protective factors and social connections. This finding aligns with the Interpersonal Psychological Theory of Suicide [60,61] and the Three-Step Theory of Suicide [62,63], which each identify social disconnection as a key driver of suicidal desire and ideation. An automated digital SPI, such as the one proposed in this study, may offer psychoeducation to users on methods of leveraging existing social connections to build social accountability for symptom self-management. It may prompt users to consider available coping activities that foster a sense of connection (eg, letter writing, watching a movie with others). Additionally, an automated digital SPI may be able to support users in crafting messages to trusted others to solicit support around safety plan use and offer resources that users can send to trusted others to educate them on how to respond to someone experiencing suicidal thoughts (eg, tips to mitigate unhelpful reactions and methods of engaging in nonjudgmental listening). Importantly, our results also show that connection to trusted individuals within participants’ personal networks was particularly valued relative to simulated connection or companionship achieved exclusively through texting with an automated chatbot. Designers of automated SPIs should consider designs that encourage and support real-world interpersonal connections with people in users’ personal networks; otherwise, they could unintentionally exacerbate a sense of isolation.

Our finding that young adults believed that having social accountability may support their use of a digital safety plan, as well as their application of coping skills, also sits within the context of digital health literature that suggests, for some, framework-based human support and accountability [64,65] can facilitate longitudinal engagement in digital interventions [66]. Although human support aimed at engaging users in a digital health intervention is conceptually distinct from the social accountability young adults saw as valuable in facilitating the use of a safety plan—not necessarily a digital health intervention—there is conceptual overlap that signals the importance of others for facilitating the use of a therapeutic tool that may, at times, be difficult to engage with.

Participants in our study wanted to be highly informed about how an automated digital SPI would handle personal data. This finding has implications for the design of digital suicide prevention interventions broadly. Intervention systems must proactively communicate meta-information about how the intervention operates, as this has the potential to increase confidence and comfort using the intervention. Prior work demonstrates that individuals are more likely to make use of crisis helpline resources when they are provided alongside barrier-reduction interventions that target informational gaps such as misunderstandings of crisis helpline practices with regard to active rescues (ie, emergency responders being called to make contact with an individual) before receiving a referral to crisis helplines [67]. Moreover, U.S. Suicide and Crisis

XSL•FO
RenderX

helpline (ie, 988) data suggest that several of the top barriers to contacting the Suicide and Crisis helpline are due to fears about law enforcement being called and about privacy and confidentiality [68]. These top concerns are addressable through robust transparency practices that frontload information about the decision rules and processes used within the intervention and by its administrators. For instance, having clear and simple language before a user engages in the intervention that addresses (1) how and when confidentiality is broken and emergency services are called; (2) how data inputs are used such as contact information for individuals a user can reach out to for help during a crisis; (3) if, when, and how trusted others are contacted, and (4) proactive psychoeducational information about how privacy and confidentiality are maintained within the program (ie, data security measures). Finally, it is important to offer resources that empower users to ask questions about privacy, confidentiality, and clinical decision-making in other contexts such as formal treatment settings and crisis helpline calls. Thus, while our findings suggest the value of self-management to reduce the need for crisis services, they also suggest important ways that crisis services can improve.

Themes raised across our ARC groups highlight the importance of ensuring that an automated SPI is usable and convenient. In the context of SMS text messaging–based safety planning, this means the system design must incorporate bidirectional or interactive messaging. To date, SMS text messaging interventions for suicide prevention [40-44] have primarily used unidirectional messaging (originating from the intervention system and being sent to the user), leaving individuals with limited opportunity to engage with intervention material in ways that reflect their lived experiences. To ensure an SMS text messaging–based automated SPI is convenient and usable, the system must use the conversational medium to be interactive. Eliciting safety plan entries by prompting users to consider new or different aspects of coping and prompting users to be highly specific are all unique affordances offered by texting. To make an SMS text messaging system usable and convenient, a system must also leverage a text-based navigation menu. This enables users to access settings, jump to different parts of their safety plan, and modify elements of a safety plan, if needed. To this end, participants highlighted that while texting may be useful for safety plan creation, having multiple methods of editing a safety plan may be necessary if the navigation menu becomes cumbersome. Thus, offering multiple modes of interacting with one's safety plan (ie, via SMS text messages and via a web-based editing tool) may be required to make an automated digital SPI maximally usable.

Our results reveal an overlap between young adults' stated needs and preferences and innovations in the interactive text–based market. Our results highlighted that young adults had several needs and preferences that may lend themselves to innovation using emerging technologies. Specifically, young adults expressed a desire for a digital SPI service to support, among other needs, (1) coping skill acquisition and use, (2) quick and simple menu navigation and easy access to desired information, (3) interpersonal connection to trusted others, and (4) clear privacy policies. The recent commercial availability of highly customizable large language models (LLMs), a class of machine learning model trained on massive linguistic data sets and designed to process and generate text-based data, has increasingly been integrated into text-based mental health interventions in the form of chatbots, content generators, and as classifier algorithms that code or categorize data so that an intervention becomes more responsive to user inputs [69]. LLMs integrated into digital chatbots can tailor responses based on context, previous interactions, and user inputs. This emerging technology is beginning to be integrated into digital suicide prevention tools, especially as it relates to the detection of suicide-related user-generated content [70], and comes with significant risks in addition to substantial promise [71]. While our study did not explicitly seek to understand the role of LLMs in an automated digital SPI, participants in our study offered important insights that are relevant to intervention development in the context of a market where commercial LLMs are available and actively being incorporated into different digital mental health technologies.

First, as it relates to supporting coping skill acquisition and use, LLM-based messages may be launched to support users in ideating coping skills when they have difficulty generating their own. Moreover, the ability of LLMs to personalize content based on individuals' contexts and inputs creates opportunities to increase the relevance of a user's safety plan at the moment. With a user's existing safety plan as context, LLMs may be able to support idea generation about how to overcome barriers to coping (eg, if an individual's safety plan contains going for a walk outside as a coping skill, but weather is prohibitive, LLM-based messages may be able to support a user identifying relevant solutions, such as going for a walk indoors, taking protective gear, using other coping skills on one's safety plan). Second, LLMs may be able to improve the experience of navigating rigid text-based menus embedded in an SMS text messaging intervention by enabling the system to accept user responses that do not exactly match preprogrammed keywords that direct a user down different branching logic trees and instead allowing users to respond using natural language. Third, young adults were clear that they wanted a digital SPI to support interpersonal connections to other humans. While they were open to an automated digital SPI, connection to real people was crucial. LLM-based chatbots have often been used as conversational agents [70], some of which have personified the conversational agent. This may exacerbate a suicide-specific vulnerability: isolation and loneliness. Past work highlights that young adults with depression were concerned that using a robot for companionship could worsen their symptoms by making the absence of real human companions more apparent [72]. More research is needed to understand whether LLM-based suicide prevention tools should present themselves as personified agents or as self-service tools. Finally, given participants' emphasis on the importance of privacy and transparency, if an LLM was to be integrated into an automated digital SPI to enhance messages, it would need to (1) proactively make transparent ways in which user inputs would be used to train models and (2) disclose the potential risk of confidentiality breaches through the LLM itself, as has been documented in other contexts [73].

XSL•FO
RenderX

By engaging young adults with lived experience of suicide-related thoughts and behaviors to understand their needs, preferences, constraints, and the context of existing technological solutions, we can identify design requirements that are able to be applied across digital suicide prevention tools.

Limitations and Future Directions

This study is limited in several important ways. First, while we engaged in rigorous user-centered design practices to understand the needs and preferences of young people who experience suicide-related thoughts and behaviors, we were only able to engage a relatively small number of individuals, and our results may not be generalizable to broader young adult populations. In our sample of 30 participants, 11 (37%) were in formal mental health treatment, and 17 (57%) planned to seek formal treatment in the coming months. These treatment-engaged numbers are greater than the approximately 12% of US adults who have experienced suicidal thoughts and sought treatment in the past year [2]. Moreover, we recruited exclusively from MHA's website which may inherently limit the perspectives of individuals included in the study to those who are open to technology-assisted interventions and services, and potentially more actively engaged in self-management of mental health symptoms. While we aimed to cover a broad range of topics in our ARC groups, our discussions were guided by researcher-posted prompts and centered on eliciting needs and preferences as they related to safety planning or texting. To this end, in this study we did not systematically examine needs and preferences for risk management in self-guided digital safety planning tools; however, future work must aim to better understand the needs and preferences across interested parties (eg, consumers, research participants, clinicians, organizations) as it relates to risk management procedures. Finally, individuals at the highest levels of suicide risk (ie, individuals who may have had elevated levels of suicidal desire and intent) were not eligible to participate in this study and thus their voices were systematically excluded as the study format lacked appropriate support for these individuals. Future work should focus on applying these design considerations and assessing their value within the context of deployable digital suicide prevention interventions.

Conclusion

Engaging young adults in the design process of a digital suicide prevention tool revealed critical considerations that must be addressed if the tool is to effectively expand access to evidence-based care to reach young people at risk for suicide-related thoughts and behaviors. Specifically, automated digital SPIs must support building skillfulness to cope effectively with suicidal crises, deepening interpersonal connections, system transparency, and data privacy.

Acknowledgments

We thank Khendra Lucas of Mental Health America for input on, and support of, this work. JM acknowledges support from the National Institute of Mental Health (grant K08 MH128640). KAVO acknowledges support from the National Institute on Aging (grant K24AG084885).

Data Availability

Because of the sensitive nature of text-based data and the potential risk of reidentification, data are not publicly available. However, deidentified data are available upon reasonable request with the appropriate data use agreements in place.

Conflicts of Interest

JM has accepted consulting fees from Boehringer Ingelheim and Shirley Ryan Ability Lab. DCM has accepted honoraria and consulting fees from Boehringer-Ingelheim, Optum Behavioral Health, Centerstone Research Institute, Woebot Health, and the One Mind Foundation; royalties from Oxford Press; and has an ownership interest in Adaptive Health, Inc. RK has accepted consulting fees and honoraria from CHESS Health and Jimini Health. ARP has an ownership interest in SafeSide Prevention, which receives fees for suicide prevention education and consultation. The University of Rochester receives royalties from SafeSide Prevention and declares this interest. SafeSide Prevention had no role in and received no fees in relation to this work. The other authors have no conflicts of interest to declare.

Multimedia Appendix 1
Initial research prompts used to generate threaded conversations during asynchronous remote community online focus groups.
[DOCX File , 16 KB - formative_v9i1e69602_app1.docx]

References

1. Substance Abuse and Mental Health Services Administration (SAMHSA). National Survey on Drug Use and Health 2013 (NSDUH-2013). U.S. Department of Health and Human Services, Substance Abuse and Mental Health Services Administration, Center for Behavioral Health Statistics and Quality. URL: https://www.samhsa.gov/data/report/results-2013-national-survey-drug-use-and-health-mental-health-detailed-tables [accessed 2024-09-25]
2. Substance Abuse and Mental Health Services Administration (SAMHSA). National Survey on Drug Use and Health 2023 (NSDUH-2023). U.S. Department of Health and Human Services, Substance Abuse and Mental Health Services

XSL•FO
RenderX

Administration, Center for Behavioral Health Statistics and Quality. URL: https://www.samhsa.gov/data/release/2023-national-survey-drug-use-and-health-nsduh-releases [accessed 2024-09-25]
3. Jobes DA, Joiner TE. Reflections on suicidal ideation. Crisis 2019 Jul;40(4):227-230 [FREE Full text] [doi: 10.1027/0227-5910/a000615] [Medline: 31274031]
4. Mojtabai R, Olfson M, Sampson NA, Jin R, Druss B, Wang PS, et al. Barriers to mental health treatment: results from the National Comorbidity Survey Replication. Psychol Med 2011 Aug;41(8):1751-1761 [FREE Full text] [doi: 10.1017/S0033291710002291] [Medline: 21134315]
5. Czyz EK, Horwitz AG, Eisenberg D, Kramer A, King CA. Self-reported barriers to professional help seeking among college students at elevated risk for suicide. J Am Coll Health 2013;61(7):398-406 [FREE Full text] [doi: 10.1080/07448481.2013.820731] [Medline: 24010494]
6. Gulliver A, Griffiths KM, Christensen H. Perceived barriers and facilitators to mental health help-seeking in young people: a systematic review. BMC Psychiatry 2010;10:113 [FREE Full text] [doi: 10.1186/1471-244X-10-113] [Medline: 21192795]
7. Caine ED. Forging an agenda for suicide prevention in the United States. Am J Public Health 2013 May;103(5):822-829. [doi: 10.2105/AJPH.2012.301078] [Medline: 23488515]
8. Stanley B, Mann JJ. The need for innovation in health care systems to improve suicide prevention. JAMA Psychiatry 2020 Jan 01;77(1):96-98. [doi: 10.1001/jamapsychiatry.2019.2769] [Medline: 31577340]
9. Hom MA, Stanley IH, Joiner TE. Evaluating factors and interventions that influence help-seeking and mental health service utilization among suicidal individuals: a review of the literature. Clin Psychol Rev 2015 Aug;40:28-39. [doi: 10.1016/j.cpr.2015.05.006] [Medline: 26048165]
10. Michelmore L, Hindley P. Help-seeking for suicidal thoughts and self-harm in young people: a systematic review. Suicide Life Threat Behav 2012 Oct;42(5):507-524. [doi: 10.1111/j.1943-278X.2012.00108.x] [Medline: 22889130]
11. Abar B, Greener E, DeRienzo V, Botelho S, Wiegand T, Conner K. Perceived access to care among adult patients who attempted suicide. Psychol Serv 2018 Aug;15(3):305-308. [doi: 10.1037/ser0000196] [Medline: 30080088]
12. Zuromski KL, Dempsey CL, Ng THH, Riggs-Donovan CA, Brent DA, Heeringa SG, et al. Utilization of and barriers to treatment among suicide decedents: results from the Army Study to Assess Risk and Resilience Among Servicemembers (Army STARRS). J Consult Clin Psychol 2019 Aug;87(8):671-683 [FREE Full text] [doi: 10.1037/ccp0000400] [Medline: 31008631]
13. O'Reilly LM, Wiggs KK, Quinn PD. Suicide risk before mental health treatment initiation: implications for screening and access to care. Psychiatr Serv 2023 Jan 01;74(1):70-72. [doi: 10.1176/appi.ps.202100657] [Medline: 35707857]
14. Yakunina ES, Rogers JR, Waehler CA, Werth JL. College students' intentions to seek help for suicidal ideation: accounting for the help-negation effect. Suicide Life Threat Behav 2010 Oct;40(5):438-450. [doi: 10.1521/suli.2010.40.5.438] [Medline: 21034207]
15. Carlton PA, Deane FP. Impact of attitudes and suicidal ideation on adolescents' intentions to seek professional psychological help. J Adolesc 2000 Feb;23(1):35-45. [doi: 10.1006/jado.1999.0299] [Medline: 10700370]
16. Stanley B, Brown GK. Safety planning intervention: a brief intervention to mitigate suicide risk. Cognitive and Behavioral Practice 2012 May;19(2):256-264. [doi: 10.1016/j.cbpra.2011.01.001]
17. Abbott-Smith S, Ring N, Dougall N, Davey J. Suicide prevention: what does the evidence show for the effectiveness of safety planning for children and young people? - A systematic scoping review. J Psychiatr Ment Health Nurs 2023 Oct;30(5):899-910 [FREE Full text] [doi: 10.1111/jpm.12928] [Medline: 37052321]
18. Ferguson M, Rhodes K, Loughhead M, McIntyre H, Procter N. The effectiveness of the safety planning intervention for adults experiencing suicide-related distress: a systematic review. Arch Suicide Res 2021 Apr 29:1-24. [doi: 10.1080/13811118.2021.1915217] [Medline: 33913799]
19. Nuij C, van Ballegooijen W, de Beurs D, Juniar D, Erlangsen A, Portzky G, et al. Safety planning-type interventions for suicide prevention: meta-analysis. Br J Psychiatry 2021 Aug;219(2):419-426. [doi: 10.1192/bjp.2021.50] [Medline: 35048835]
20. National Action Alliance for Suicide Prevention: Transforming Health Systems Initiative Work Group. Recommended standard care for people with suicide risk: making health care suicide safe. National Action Alliance for Suicide Prevention. Washington, DC: Education Development Center, Inc; 2018. URL: https://theactionalliance.org/sites/default/files/action_alliance_recommended_standard_care_final.pdf [accessed 2025-02-24]
21. Toolkit: Engage. Zero Suicide. 2024. URL: https://zerosuicide.edc.org/toolkit/engage#collaborative-safety-plan [accessed 2024-10-31]
22. Gamarra JM, Luciano MT, Gradus JL, Wiltsey Stirman S. Assessing variability and implementation fidelity of suicide prevention safety planning in a regional VA healthcare system. Crisis 2015;36(6):433-439. [doi: 10.1027/0227-5910/a000345] [Medline: 26648231]
23. Davis M, Siegel J, Becker-Haimes EM, Jager-Hyman S, Beidas RS, Young JF, et al. Identifying common and unique barriers and facilitators to implementing evidence-based practices for suicide prevention across primary care and specialty mental health settings. Arch Suicide Res 2023;27(2):192-214 [FREE Full text] [doi: 10.1080/13811118.2021.1982094] [Medline: 34651544]

XSL•FO
RenderX

24. Haroz EE, Bajaj MA, Nestadt PS, Campo JV, Wilcox HC. Clinician perspectives on suicide safety planning and its implementation. Arch Suicide Res 2024 Jul 01:1-11. [doi: 10.1080/13811118.2024.2370852] [Medline: 38949292]
25. Gryglewicz K, Orr VL, McNeil MJ, Taliaferro LA, Hines S, Duffy TL, et al. Translating suicide safety planning components into the design of mHealth app features: systematic review. JMIR Ment Health 2024 Mar 28;11:e52763 [FREE Full text] [doi: 10.2196/52763] [Medline: 38546711]
26. Boudreaux ED, Brown GK, Stanley B, Sadasivam RS, Camargo CA, Miller IW. Computer administered safety planning for individuals at risk for suicide: development and usability testing. J Med Internet Res 2017 May 15;19(5):e149 [FREE Full text] [doi: 10.2196/jmir.6816] [Medline: 28506957]
27. Melvin GA, Gresham D, Beaton S, Coles J, Tonge BJ, Gordon MS, et al. Evaluating the feasibility and effectiveness of an Australian safety planning smartphone application: a pilot study within a tertiary mental health service. Suicide Life Threat Behav 2019 Jun;49(3):846-858 [FREE Full text] [doi: 10.1111/sltb.12490] [Medline: 29999193]
28. Methi N, Weeks I, Hunt R, McGuire TC, Rubin A, Decker M, et al. Accessibility and utility of an electronic self-guided safety plan for adolescents. J Clin Child Adolesc Psychol 2024;53(6):935-943. [doi: 10.1080/15374416.2024.2395271] [Medline: 39312731]
29. Rainbow C, Tatnell R, Blashki G, Fuller-Tyszkiewicz M, Melvin GA. Digital safety plan effectiveness and use: findings from a three-month longitudinal study. Psychiatry Res 2024 Mar;333:115748. [doi: 10.1016/j.psychres.2024.115748] [Medline: 38277811]
30. Larkin C, Tulu B, Djamasbi S, Garner R, Varzgani F, Siddique M, et al. Comparing the acceptability and quality of intervention modalities for suicidality in the emergency department: randomized feasibility trial. JMIR Ment Health 2023 Oct 24;10:e49783 [FREE Full text] [doi: 10.2196/49783] [Medline: 37874619]
31. Rainbow C, Tatnell R, Blashki G, Melvin GA. Perceived usefulness of self-guided versus collaborative suicide safety plans in online help-seekers. Crisis 2024 Jul;45(4):294-300. [doi: 10.1027/0227-5910/a000940] [Medline: 38319612]
32. Buus N, Erlangsen A, River J, Andreasson K, Frandsen H, Larsen JLS, et al. Stakeholder perspectives on using and developing the MYPLAN suicide prevention mobile phone application: a focus group study. Arch Suicide Res 2020;24(1):48-63. [doi: 10.1080/13811118.2018.1489319] [Medline: 29985770]
33. Kennard BD, Biernesser C, Wolfe KL, Foxwell AA, Craddock LSJ, Rial KV, et al. Developing a brief suicide prevention intervention and mobile phone application: a qualitative report. J Technol Hum Serv 2015 Oct 01;33(4):345-357 [FREE Full text] [doi: 10.1080/15228835.2015.1106384] [Medline: 26977137]
34. O'Grady C, Melia R, Bogue J, O'Sullivan M, Young K, Duggan J. A mobile health approach for improving outcomes in suicide prevention (SafePlan). J Med Internet Res 2020 Jul 30;22(7):e17481 [FREE Full text] [doi: 10.2196/17481]
35. Larkin C, Djamasbi S, Boudreaux ED, Varzgani F, Garner R, Siddique M, et al. ReachCare Mobile Apps for Patients Experiencing Suicidality in the Emergency Department: Development and Usability Testing Using Mixed Methods. JMIR Form Res 2023 Jan 27;7:e41422 [FREE Full text] [doi: 10.2196/41422] [Medline: 36705961]
36. Sutori S, Hadlaczky G, Eliasson E, Wasserman D, Carli V. Systematic review and meta-analysis: effectiveness of stand-alone digital suicide preventive interventions for the self-management of suicidality. J Technol Behav Sci 2023 Dec 30;9(3):440-451. [doi: 10.1007/s41347-023-00374-7]
37. Newport F. The new era of communication among Americans. Gallup. 2014. URL: https://news.gallup.com/poll/179288/new-era-communication-americans.aspx [accessed 2020-08-30]
38. U.S. smartphone use in 2015. Pew Research Center. 2015. URL: https://www.pewresearch.org/internet/2015/04/01/us-smartphone-use-in-2015/ [accessed 2024-09-24]
39. Griffioen N, Scholten H, Lichtwarck-Aschoff A, van Rooij M, Granic I. Everyone does it—differently: a window into emerging adults' smartphone use. Humanit Soc Sci Commun 2021 Jul 20;8(1):1-11. [doi: 10.1057/s41599-021-00863-1]
40. Czyz EK, Arango A, Healy N, King CA, Walton M. Augmenting safety planning with text messaging support for adolescents at elevated suicide risk: development and acceptability study. JMIR Ment Health 2020 May 25;7(5):e17345 [FREE Full text] [doi: 10.2196/17345] [Medline: 32160150]
41. Comtois KA, Kerbrat AH, DeCou CR, Atkins DC, Majeres JJ, Baker JC, et al. Effect of augmenting standard care for military personnel with brief caring text messages for suicide prevention: a randomized clinical trial. JAMA Psychiatry 2019 May 01;76(5):474-483. [doi: 10.1001/jamapsychiatry.2018.4530] [Medline: 30758491]
42. Berrouiguet S, Larsen ME, Mesmeur C, Gravey M, Billot R, Walter M, HUGOPSY Network, et al. Toward mHealth brief contact interventions in suicide prevention: case series from the suicide intervention assisted by messages (SIAM) randomized controlled trial. JMIR Mhealth Uhealth 2018 Jan 10;6(1):e8 [FREE Full text] [doi: 10.2196/mhealth.7780] [Medline: 29321126]
43. Pisani AR, Wyman PA, Petrova M, Judd E, Schmeelk-Cone K, Thiha P, et al. Framework for supporting adolescent peer leaders: a pilot using text messaging in a school-based substance use prevention program. J Prim Prev 2019 Apr;40(2):243-254 [FREE Full text] [doi: 10.1007/s10935-019-00545-4] [Medline: 30827007]
44. Thiha P, Pisani AR, Gurditta K, Cherry E, Peterson DR, Kautz H, et al. Efficacy of web-based collection of strength-based testimonials for text message extension of youth suicide prevention program: randomized controlled experiment. JMIR Public Health Surveill 2016 Nov 09;2(2):e164 [FREE Full text] [doi: 10.2196/publichealth.6207] [Medline: 27829575]

XSL•FO
RenderX

45. Czyz EK, King CA, Prouty D, Micol VJ, Walton M, Nahum-Shani I. Adaptive intervention for prevention of adolescent suicidal behavior after hospitalization: a pilot sequential multiple assignment randomized trial. J Child Psychol Psychiatry 2021 Aug;62(8):1019-1031 [FREE Full text] [doi: 10.1111/jcpp.13383] [Medline: 33590475]
46. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders, 5th Edition. Washington, DC: American Psychiatric Association; 2013.
47. Stanley IH, Hom MA, Christensen K, Keane TM, Marx BP, Björgvinsson T. Psychometric properties of the Depressive Symptom Index-Suicidality Subscale (DSI-SS) in an adult psychiatric sample. Psychol Assess 2021 Oct;33(10):987-997. [doi: 10.1037/pas0001043] [Medline: 34180693]
48. Beck A, Steer R, Brown G. Manual for the Beck Depression Inventory-II. San Antonio, TX: Psychological Corporation; 1996.
49. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381 [FREE Full text] [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
50. Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O'Neal L, REDCap Consortium. The REDCap consortium: building an international community of software platform partners. J Biomed Inform 2019 Jul;95:103208 [FREE Full text] [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]
51. MacLeod H, Jelen B, Prabhakar A, Oehlberg L, Siek K, Connelly K. A guide to using asynchronous remote communities (ARC) for researching distributed populations. EAI Endorsed Trans Perv Health Tech 2017 Jul 18;3(11):e4. [doi: 10.4108/eai.18-7-2017.152898]
52. Mastre J, MacLeod H, Connelly C, Dunbar J, Beck J, Siek K, et al. Defining through expansion: conducting asynchronous remote communities (ARC) research with stigmatized groups. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. 2018 Apr 21 Presented at: The 2018 CHI Conference on Human Factors in Computing Systems; April 21-26, 2018; Montreal QC, Canada p. 1-13. [doi: 10.1145/3173574.3174131]
53. Prabhakar A, Guerra-Reyes L, Kleinschmidt V, Jelen B, MacLeod H, Connelly K, et al. Investigating the suitability of the asynchronous, remote, community-based method for pregnant and new mothers. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. New York, NY: ACM Press; 2017 Presented at: The 2017 CHI Conference on Human Factors in Computing Systems; May 6-11, 2017; Denver, CO p. 4924-4934. [doi: 10.1145/3025453.3025546]
54. Hinshaw SP, Hoagwood K, Jensen PS, Kratochvil C, Bickman L, Clarke G, et al. AACAP 2001 research forum: challenges and recommendations regarding recruitment and retention of participants in research investigations. J Am Acad Child Adolesc Psychiatry 2004 Aug;43(8):1037-1045. [doi: 10.1097/01.chi.0000129222.89433.3d] [Medline: 15266200]
55. FocusGroupIt. FocusGroupIt. URL: https://www.focusgroupit.com/ [accessed 2023-11-24]
56. Meyerhoff J, Kornfield R, Mohr DC, Reddy M. Meeting young adults' social support needs across the health behavior change journey: implications for digital mental health tools. Proc ACM Hum Comput Interact 2022 Nov 11;6(CSCW2):1-33 [FREE Full text] [doi: 10.1145/3555203] [Medline: 36387059]
57. Braun V, Clarke V. Using thematic analysis in psychology. Qualitative Research in Psychology 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
58. Braun V, Clarke V. Reflecting on reflexive thematic analysis. Qualitative Research in Sport, Exercise and Health 2019 Jun 13;11(4):589-597. [doi: 10.1080/2159676x.2019.1628806]
59. Dedoose. Los Angeles, CA: SocioCultural Research Consultants, LLC; 2018. URL: https://www.dedoose.com/ [accessed 2024-09-24]
60. Joiner T. Why People Die by Suicide. Cambridge, MA: Harvard University Press; 2005.
61. Van Orden KA, Witte TK, Cukrowicz KC, Braithwaite SR, Selby EA, Joiner TE. The interpersonal theory of suicide. Psychol Rev 2010 Apr;117(2):575-600 [FREE Full text] [doi: 10.1037/a0018697] [Medline: 20438238]
62. Klonsky ED, Pachkowski MC, Shahnaz A, May AM. The three-step theory of suicide: description, evidence, and some useful points of clarification. Prev Med 2021 Nov;152(Pt 1):106549. [doi: 10.1016/j.ypmed.2021.106549] [Medline: 34538372]
63. Klonsky ED, May AM. The three-step theory (3ST): a new theory of suicide rooted in the “ideation-to-action” framework. International Journal of Cognitive Therapy 2015 Jun;8(2):114-129. [doi: 10.1521/ijct.2015.8.2.114]
64. Mohr DC, Cuijpers P, Lehman K. Supportive accountability: a model for providing human support to enhance adherence to eHealth interventions. J Med Internet Res 2011;13(1):e30 [FREE Full text] [doi: 10.2196/jmir.1602] [Medline: 21393123]
65. Schueller SM, Tomasino KN, Mohr DC. Integrating human support into behavioral intervention technologies: the efficiency model of support. Clin Psychol Sci Pract 2016 Nov 17;24(1):27-45. [doi: 10.1111/cpsp.12173]
66. Borghouts J, Eikey E, Mark G, De Leon C, Schueller SM, Schneider M, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. J Med Internet Res 2021 Mar 24;23(3):e24387 [FREE Full text] [doi: 10.2196/24387] [Medline: 33759801]
67. Jaroszewski AC, Morris RR, Nock MK. Randomized controlled trial of an online machine learning-driven risk assessment and intervention platform for increasing the use of crisis services. J Consult Clin Psychol 2019 Apr;87(4):370-379. [doi: 10.1037/ccp0000389] [Medline: 30883164]

XSL•FO
RenderX

68. Ad Council Research Institute, National Action Alliance for Suicide Prevention, Suicide Prevention Resource Center. Framework for Successful Messaging. URL: https://suicidepreventionmessaging.org/sites/default/files/2024-05/988%20FR_Trusted%20Messengers_Report_Final-508.pdf [accessed 2024-07-12]
69. Stade EC, Stirman SW, Ungar LH, Boland CL, Schwartz HA, Yaden DB, et al. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. Npj Ment Health Res 2024 Apr 02;3(1):12. [doi: 10.1038/s44184-024-00056-z] [Medline: 38609507]
70. Guo Z, Lai A, Thygesen JH, Farrington J, Keen T, Li K. Large language models for mental health applications: systematic review. JMIR Ment Health 2024 Oct 18;11:e57400 [FREE Full text] [doi: 10.2196/57400] [Medline: 39423368]
71. Choudhury M, Pendse S, Kumar N. Benefits and harms of large language models in digital mental health. arXiv. Preprint posted online on November 7, 2023 2023:1 [FREE Full text] [doi: 10.31234/osf.io/y8ax9]
72. Kornfield R, Meyerhoff J, Studd H, Bhattacharjee A, Williams JJ, Reddy M, et al. Meeting users where they are: user-centered design of an automated text messaging tool to support the mental health of young adults. Proc SIGCHI Conf Hum Factor Comput Syst 2022 Apr;2022:329. [doi: 10.1145/3491102.3502046] [Medline: 35574512]
73. White J. How strangers got my email address from ChatGPT's model. The New York Times. 2023 Dec 22. URL: https://www.nytimes.com/interactive/2023/12/22/technology/openai-chatgpt-privacy-exploit.html [accessed 2024-11-19]

Abbreviations

ARC: asynchronous remote community
CCSI: Cross-Cutting Symptom Inventory
DSI-SS: Depression Severity Index—Suicidality Subscale
DSM-5: Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition
LLM: large language model
MHA: Mental Health America
SPI: safety planning intervention

Edited by A Mavragani; submitted 03.12.24; peer-reviewed by C Wijaya, C Rainbow, A AL-Asadi; comments to author 30.12.24; revised version received 04.02.25; accepted 11.02.25; published 26.02.25.

Please cite as:
Meyerhoff J, Popowski SA, Lakhtakia T, Tack E, Kornfield R, Kruzan KP, Krause CJ, Nguyen T, Rushton K, Pisani AR, Reddy M, Van Orden KA, Mohr DC
Automated Digital Safety Planning Interventions for Young Adults: Qualitative Study Using Online Co-design Methods
JMIR Form Res 2025;9:e69602
URL: https://formative.jmir.org/2025/1/e69602
doi:10.2196/69602
PMID:

©Jonah Meyerhoff, Sarah A Popowski, Tanvi Lakhtakia, Emily Tack, Rachel Kornfield, Kaylee P Kruzan, Charles J Krause, Theresa Nguyen, Kevin Rushton, Anthony R Pisani, Madhu Reddy, Kimberly A Van Orden, David C Mohr. Originally published in JMIR Formative Research (https://formative.jmir.org), 26.02.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Evaluation of the Clear Fear Smartphone App for Young People Experiencing Anxiety: Uncontrolled Pre– and Post–Follow-Up Study

Chiara Samele[1], MSc, PhD; Norman Urquia[1], MSc, PhD; Rachel Edwards[2], MSc; Katie Donnell[2], MSc; Nihara Krause[2], MSc, PsychD

[1]Informed Thinking, London, United Kingdom
[2]stem4, London, United Kingdom

Corresponding Author:
Nihara Krause, MSc, PsychD
stem4
Connect House, 133-137 Alexandra Road
Wimbledon
London, SW19 7JY
United Kingdom
Phone: 44 7956396375
Email: dr.nihara.krause@stem4.org.uk

Abstract

Background: Mobile health apps are proving to be an important tool for increasing access to psychological therapies early on, particularly with rising rates of anxiety and depression in young people.

Objective: We aimed to assess the usability, acceptability, safety, and effectiveness of a new app, Clear Fear, developed to help young people manage symptoms of anxiety using the principles of cognitive behavioral therapy.

Methods: The Clear Fear app was developed to provide cognitive behavioral strategies to suit anxiety disorders. An uncontrolled pre– and post–follow-up design over a 9-week period was used to assess the app and its effects. This study comprised 3 phases: baseline (stage 1), post–app familiarization phase (stage 2), and follow-up (stage 3). Eligible participants were aged between 16 and 25 years with mild to moderate anxiety but not currently receiving treatment or in contact with specialist mental health services or using other interventions or apps to help monitor or manage their mental health. A community sample was recruited via advertisements, relevant websites, and social media networks. Eligible participants completed standardized self-report tools and questionnaires at each study stage. These measured probable symptoms of anxiety (7-item Generalized Anxiety Disorder scale) and depression (Mood and Feelings Questionnaire); emotional and behavioral difficulties (Strengths and Difficulties Questionnaire); and feedback on the usability, accessibility, and safety of the app. Mean scores at baseline and follow-up were compared using paired 2-tailed *t* tests or Wilcoxon signed rank tests. Qualitative data derived from open-ended questions were coded and entered into NVivo (version 10) for analysis.

Results: A total of 48 young people entered the study at baseline, with 37 (77%) completing all outcome measures at follow-up. The sample was mostly female (37/48, 77%). The mean age was 20.1 (SD 2.1) years. In total, 48% (23/48) of the participants reached the threshold for probable anxiety disorder, 56% (27/48) had positive scores for probable depression, and 75% (36/48) obtained a total score of “very high” on the Strengths and Difficulties Questionnaire for emotional and behavioral difficulties. The app was well received, offering reassurance, practical and immediate help to manage symptoms, and encouragement to seek help, and was generally found easy to use. A small minority (3/48, 6%) found the app difficult to navigate. The Clear Fear app resulted in statistically significant reductions in probable symptoms of anxiety (t_{36}=2.6, 95% CI 0.41-3.53; P=.01) and depression (z=2.3; P=.02) and behavioral and emotional difficulties (t_{47}=4.5, 95% CI 3.67-9.65; P<.001), representing mostly medium to large standardized effect sizes.

Conclusions: The Clear Fear app was found to be usable, acceptable, safe, and effective in helping manage symptoms of anxiety and depression and emotional and behavioral difficulties.

(JMIR Form Res 2025;9:e55603) doi:10.2196/55603

XSL•FO
RenderX

KEYWORDS

mental health; anxiety; depression; emotional and behavioral difficulties; mobile phone app; cognitive behavioral therapy; digital tool; young people; mobile phone

Introduction

Background

The rates of probable mental health problems in the United Kingdom among young people have steadily increased since 2017. According to National Health Service Digital data (2022), for example, the rates for young people aged 17 to 19 years with mental health problems rose by 7.6% in 2020 and reached 25.7% in 2022 [1]. In 2017 to 2018, depression and anxiety among young women aged 16 to 24 years rose by 26% compared to the previous year [2].

The high volume of young people with experience of symptoms associated with moderate depression and anxiety has prompted a rise in referrals to already overstretched specialist mental health services. For example, in November 2023, a total of 445,000 people were in contact with children and young people's mental health services compared to 362,000 in November 2021 [3]. New referrals to these services reached a peak of 87,000 per month in 2021, the highest since 2019 [4]. While referrals have increased, there remains a huge shortfall in those who need support and receive it. In 2020, the proportion of children and young people needing and receiving support was 27% [4].

Therefore, it is important to find ways to ensure that young people have timely access to support. One growing avenue is support through mobile health apps.

Approximately 96% of those aged between 16 and 24 years own a smartphone [5]. Mobile health apps now constitute an important tool for increasing access to psychological therapies, particularly for young people. Several published systematic reviews have assessed the usability, feasibility, acceptability, functionality, and effectiveness of specific mobile mental health apps for young people [6,7]. Despite the published literature on how mobile app technology can increase access to psychological therapy, relatively little research exists on its effectiveness and whether it improves symptoms of mental health problems. However, there is some evidence suggesting that certain mobile apps for mental health can help improve the monitoring and management of mental health [8]. A handful of studies confirm that web-based cognitive behavioral therapy (CBT) is as effective as face-to-face treatment for anxiety and depression [9]. Among young people, mobile mental health apps also offer an important opportunity to prevent mental health problems or intervene early to reduce the risk of a long-term mental health condition into adulthood [10].

Objectives

Given the limited available evidence on the effectiveness of mental health apps for young people, this pilot study aimed to assess whether the Clear Fear app, developed to help young people manage symptoms of anxiety, could reduce these symptoms.

The main aims of this study were to assess the usability, acceptability, safety, and effectiveness of the Clear Fear app, specifically (1) what particular features of the app young people found most helpful or beneficial; (2) the ease of use, acceptability, and frequency with which young people used the app, as well as whether they found it safe and any improvements that could be made; and (3) whether the Clear Fear app was able to reduce or improve symptoms of anxiety primarily (and any adverse behavior) following its use during the study period.

Thus, this study was designed to test the following hypothesis: a self-guided mobile cognitive behavioral app can improve anxiety symptoms and be acceptable to young people over a period of 9 weeks.

Methods

Clear Fear App

Aims of the App

The Clear Fear app aims to support young people in managing symptoms of anxiety early in their stage of development. It does this by (1) providing cognitive behavioral strategies to suit anxiety disorders; (2) encouraging self-monitoring and self-management and building resilience; and (3) teaching alternative, adaptive ways to deal with anxiety.

Users manage their anxiety and "face their fear" in a gradual way while learning to manage panic, challenge worries, and regulate their emotions.

The app provides a first step for young people looking for immediate techniques to help manage anxious thoughts and behaviors for longer-term change. It was also designed to be easily accessible for those who do not meet the criteria for mental health services or are currently on a waiting list to receive appropriate treatment for anxiety disorder. The app does not replace specialist treatment but can be used in addition to it.

Development, Structure, and App Function

The app was developed by NK at stem4, a charity supporting teenage mental health, in collaboration with young people. The development process began with reviewing available mainstream apps that aimed to manage anxiety through mindfulness. Integral to the process was the coproduction of the Clear Fear app with a group of 30 young people (aged between 14 and 17 years) who took part in a series of user engagement sessions to explore desired outputs, user journeys, visual concepts, colors, imagery, security and privacy considerations, user experience, and tone of voice. The app provides a low-intensity CBT treatment to monitor and change thoughts, emotions, and behaviors associated with mild to moderate symptoms of anxiety. It seeks to calm fear responses through a cognitive behavioral framework along with breathing, relaxation, and mindfulness exercises.

The Clear Your Fear section of the app provides activities to help users with the 4 key areas—dealing with emotions,

XSL•FO
RenderX

managing worries, reacting to worries, and managing physical responses to anxiety—and is described in Textbox 1 and Figure 1. Each activity provides some brief introductory information to explain why the activity can help, as well as examples where necessary throughout.

The multiple areas for goal setting provided in the app facilitate the following: identifying specific goals, tailoring each goal to suit each individual, identifying barriers, problem-solving, breaking long-term goals into short-term challenges, accomplishing easier challenges first, prompts, cueing, reinforcing success, social support, and tracking personal information to enhance self-directed learning.

Textbox 1. Description of the Clear Your Fear section of the Clear Fear app.

Dealing With Your Emotions

This consists of activities for users to learn to identify, express, and regulate emotions, which can help reduce anxiety.

- Express Yourself activities provide opportunities for note keeping and finding alternative ways of emotional regulation.
- Stay Calm activities help users learn to calm the body and mind by doing either the guided breathing activity or mindfulness exercises.
- Laugh & Smile activities aid in substituting negative emotions for positive ones through the use of humor by viewing jokes or gifs.

Managing Your Worries

This focuses on learning how thinking can contribute to anxiety and how to make a change by learning to understand, identify, and change fearful thoughts (Worry Warriors); have a break from anxious thoughts (Worry Box); or put a problem-solving strategy into practice (Worry Ladder).

- Worry Warriors guides the user through a process of cognitive restructuring to identify and combat 3 common types of automatic negative thoughts (those that have been magnified by anxiety or taken to the extreme or that are about something negative that is expected to happen), together with self-monitoring, seeking evidence, and restructuring thoughts. The resulting positive thoughts that are created can then be revisited in the Grit Box, located in another area of the app. Therapeutic gains from this process can include symptom relief and an increased sense of control over anxiety through its objectification.
- Worry Box is an activity in which the user types in their worries to be stored in this section of the app, which can then be emptied to metaphorically clear them.
- Worry Ladder guides the user through dealing with their worries one step at a time (and creating a visual ladder of these steps in the process) by first stating a worry and then typing in possible solutions; choosing one to try; and, finally, noting what happened and what worked. Both the Worry Ladder and Worry Box, together with problem-solving, help with prioritizing and containing anxiety.

Reacting to Worries

This involves the user identifying situations in which they want to experience less anxiety and learn to face their worries with the help of activities to lessen overdoing or avoiding things. In each case, the user is guided through setting a goal and breaking it down into an exposure hierarchy of manageable steps and can also mark when a goal has been achieved.

Managing Physical Responses to Anxiety

This consists of activities for users to cope with and lessen the physical effects of anxiety. In total, 4 categories of goals are presented (exercise, diet, relaxation, and sleep), within which the user is given brief information as to why the activity can help and suggestions of goals to set, which can be input into the app. The user is also given the option of setting push notifications (at daily, weekly, or monthly intervals) for the goal to help remind them.

XSL•FO
RenderX

Figure 1. Flowchart and screenshots of the Clear Fear app, the digital intervention used in this uncontrolled pre– and post–follow-up study to evaluate its usability, acceptability, safety, and effectiveness in helping young people manage symptoms of anxiety using the principles of cognitive behavioral therapy over a 9-week period.

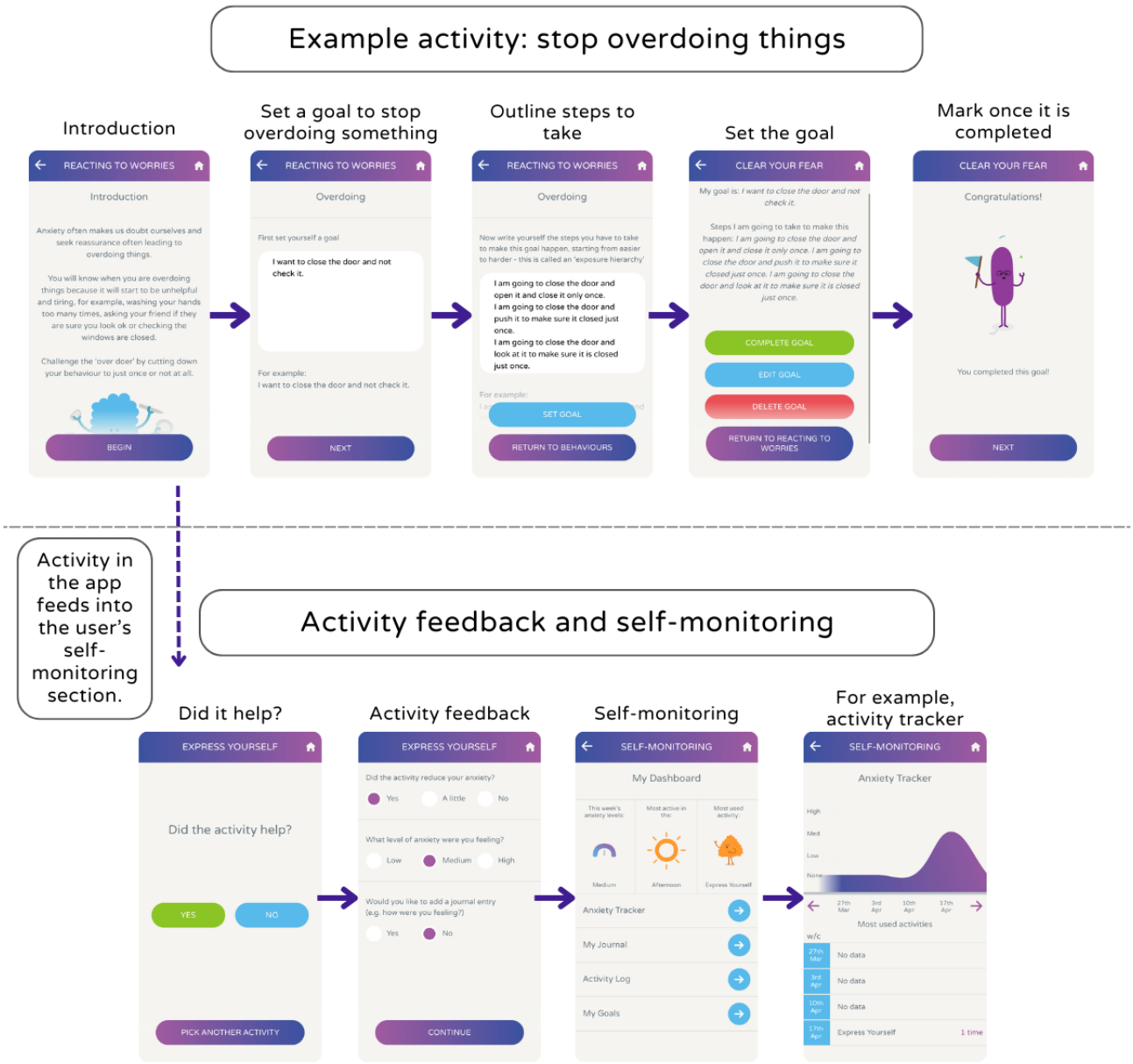

Activity Feedback and Self-Monitoring

Upon completion of an activity, the user is asked to provide the following feedback: whether the activity helped (Yes or No), whether it reduced their anxiety (Yes, A little, or No), and the level of anxiety they were feeling (Low, Medium, or High). This information feeds into that user's self-monitoring section on the app, which includes an anxiety profile and anxiety tracker to help the user self-monitor across every task.

Grit Box

The Grit Box is based on the concept that grit is a necessary factor in resilience. It holds the user's own Worry Warrior Positive Thoughts, which are generated from their activity on the app. It also contains a range of motivational statements, positive quotes, and information on resilient and inspirational people as suggested during the coproduction process with young people.

Other Areas of the App

Psychoeducation is provided in the information section and in optional notifications sent to the users to support their momentum in reaching personal goals and in gaining valuable psychological insights (eg, setting realistic goals, overcoming fears, and increasing motivation). On achieving a goal, users receive a plan on how to maintain the new skills and avoid falling back into old habits.

In addition, information about anxiety is provided via messages sent to the user to help normalize their experience, as well as giving information on the value and functioning of anxiety. This includes information on specific types of anxiety (namely,

XSL•FO
RenderX

generalized anxiety, social anxiety, performance anxiety, obsessive-compulsive disorder and health anxiety, separation anxiety, exam anxiety, phobias, and fear of missing out).

Finally, users can receive physiological support and help in dealing with panic and panic attacks based on mindfulness exercises in both visual and auditory formats to improve low self-esteem, emotional reactivity, and functioning [11], as well as signposts to further support.

The Clear Fear App's Contribution to Digital Mental Health

The Clear Fear app is a clinician-developed, evidence-based, and comprehensive CBT app that differs from those using mindfulness and meditation as their primary interventions, artificial intelligence chatbots, and other available apps for anxiety management developed *without* clinician input or comprehensive evidence-based approaches [12,13].

The Clear Fear app is designed primarily for managing anxiety symptoms in young people. It provides a comprehensive evidence-based approach that is both structured and personalizable, incorporating CBT exercises, psychoeducation, mindfulness, and self-monitoring. The content, language, visuals, tone of voice, and user journeys are age appropriate. For younger users aged between 11 and 15 years, the Clear Fear app includes age-appropriate design and language, ease of navigation, age-relevant gamification, less abstract cognitive concepts, tasks that cater to shorter attention spans, and increased socialization.

It also prioritizes user privacy and data safety as it does not collect identifiable information or require user accounts, can be passcode protected, is regularly reviewed and updated, and is compliant with regulatory standards such as the National Health Service clinical safety standard DCB0129.

Unlike many mental health apps, Clear Fear is free to download in the United Kingdom, with no subscriptions, in-app purchases, or advertisements, and can be used without an internet connection, making it highly accessible.

Study Design

This study used an uncontrolled pre– and post–follow-up design over a 9-week period to assess the app and its effects. The study comprised three main phases: (1) baseline and app introduction (stage 1), (2) post–app familiarization phase (stage 2), and (3) follow-up (stage 3).

Participants and Eligibility Criteria

Participants were a community sample recruited via stem4's website and their existing social media networks.

Potential participants were eligible to take part in the study if they (1) were aged between 16 and 25 years and (2) had mild to moderate anxiety.

Participants were excluded if they were (1) in contact with mental health services or taking medication for a mental health problem; (2) using other interventions or apps to monitor their mental health; or (3) at risk of suicide, diagnosed with a severe and enduring mental health problem such as psychosis or major depression, or experiencing substance use problems.

Potential participants were screened using a 10-item questionnaire developed for this study on suicide ideation, any severe mental health problems, and substance use issues (drugs and alcohol).

Recruitment

The main method for recruiting participants was through advertisements on the stem4 website and via stem4 social media channels inviting young people who experienced anxiety and had not previously used the Clear Fear app to take part in the study. stem4 also provided details of the study to schools within their local area. Eligible participants were offered a web-based shopping voucher as an incentive to take part in the study.

A web-based screening questionnaire (developed for this study) was sent to all young people interested in taking part in the study to assess their study eligibility and level of risk. This included a 10- to 15-item screening tool to identify any current mental health issues or history of and contact with mental health services. Those identified with an increased risk at initial screening were excluded from the study to prioritize their safety. These individuals were provided with an information pack containing emergency contact numbers and guidance on further support, and they were encouraged to notify the research team if they accessed additional support.

Eligible participants were provided with an information sheet detailing the study and asked to complete an informed consent form before taking part. Participants could leave the study at any time during the study without giving a reason if they wished. Participants were also made aware that the Clear Fear app does not replace traditional treatment and were appropriately signposted to sources of help where necessary. Participants also received an introductory leaflet with information on symptoms and suggested strategies for mitigating risk and were directed to sections on the app that provided immediate signposts and access to services.

Furthermore, at the start of the study, all participants completed the Safety Net section of the app, which is a vital safety feature that is always available to the user and can be shared with responsible adults. The users populated what they could think and do when they needed help and the numbers of people they could contact to go along with the Safety Net section's prepopulated emergency numbers.

The app also prompted users to seek help from a trusted adult or general practitioner if they noticed that their anxiety levels, from the tracker in the Self-Monitoring section, were consistently high.

Sample Size

The aim was to recruit approximately 45 young people to address the main study aims. No formal sample size calculation was made given the preliminary nature of the study and piloting of the Clear Fear app. We allowed for attrition rates of up to 15% at the postfamiliarization and follow-up time points.

XSL•FO
RenderX

Study Procedure

Figure 2 shows the study procedure, including the 3 main study stages and the recruitment process. We also consulted with 6 people (aged between 14 and 17 years) to provide feedback on the study proposal and procedure regarding their preferences concerning web-based data collection.

Figure 2. Diagram of the procedure of this uncontrolled pre– and post–follow-up study to evaluate the Clear Fear app's usability, acceptability, safety, and effectiveness in helping young people manage symptoms of anxiety using the principles of cognitive behavioral therapy over a 9-week period.

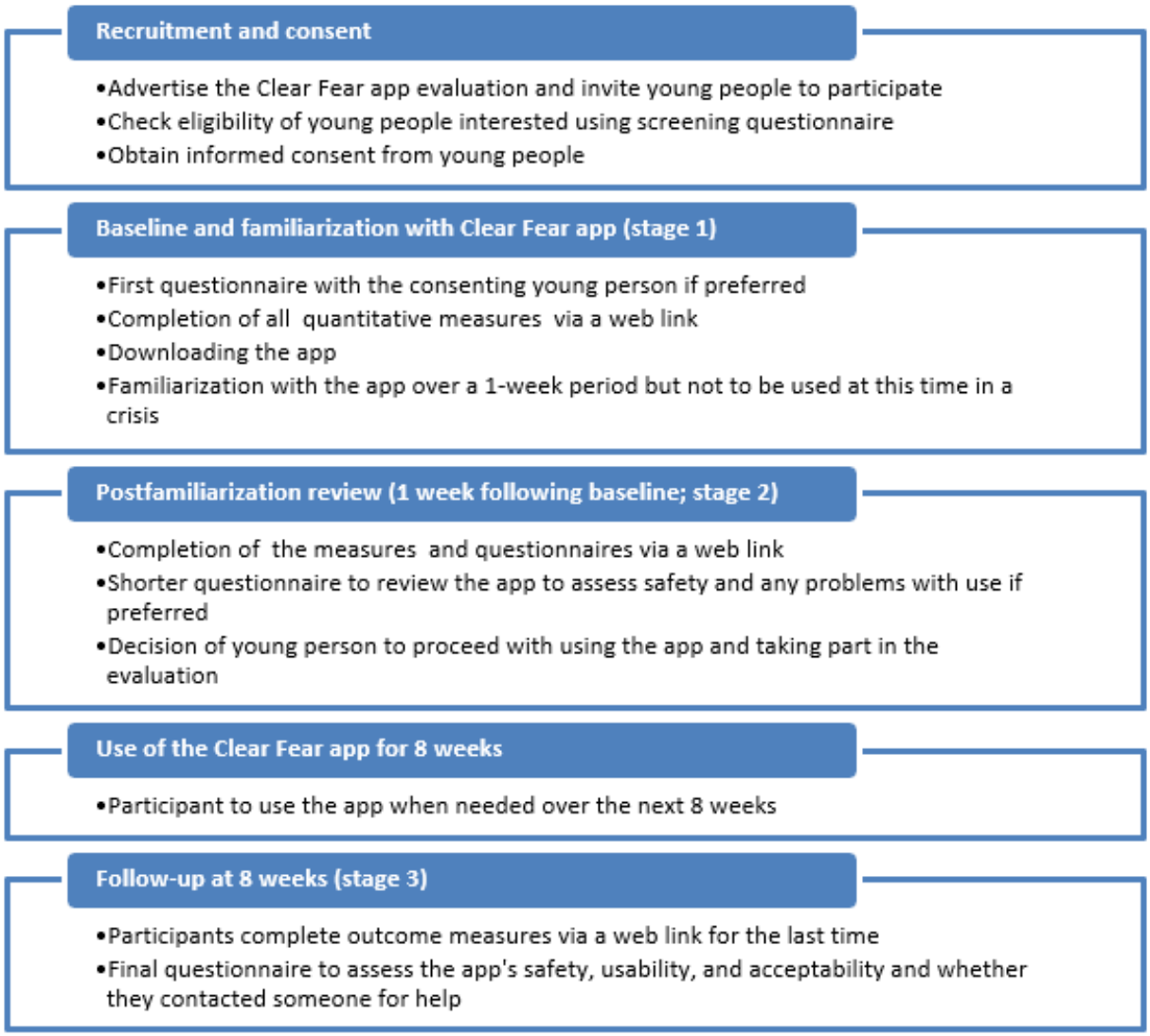

Baseline (Stage 1)

Upon entry to the study, participants were provided with web-based instructions on how to download and use the Clear Fear app and the study procedure. Participants were then allocated a study ID number to maintain confidentiality and anonymity and emailed a link to the web-based baseline questionnaires. The link was created by the researchers using SurveyMonkey (SurveyMonkey Inc).

At this stage, participants were asked only to familiarize themselves with the Clear Fear app over the coming week and not to use it during a time of crisis. This was to ensure the person's safety until they became fully versed in using the Clear Fear app and to obtain baseline measures of the frequency and severity of anxiety they experienced and their existing coping mechanisms.

Post–App Familiarization Phase (Stage 2)

One week after baseline, web-based questionnaires were sent again to participants, which included a user questionnaire to gauge their assessment of the Clear Fear app and any problems they may have had. If the participant wished to continue taking part in the study, they were asked to complete the outcome measures again. Participants were informed that they could now use the Clear Fear app as often as they wanted over the following 8 weeks.

Follow-Up (Stage 3)

After an 8-week period of using the Clear Fear app, participants were sent the web-based questionnaires for a final time to complete, again including the user questionnaire for their specific feedback on using the app. In total, 2 to 3 attempts were made to follow up on participants if no response was received.

XSL•FO
RenderX

If no response was received, it was assumed that the participant had withdrawn from the study.

Ongoing Monitoring and Risk Assessment

Throughout the study, participants self-reported their anxiety levels at regular intervals. If they reported high levels of anxiety or emerging new symptoms, they were encouraged to consult a general practitioner or, depending on the severity, seek more urgent help. After the study, participants were given resources on maintaining well-being.

stem4, the owners of the app, had a clinical safety protocol and a hazard mitigation list that was managed and reviewed regularly with clear timelines of response. Their clinical safety officer was available to respond to inquiries daily.

Data Collection

A standardized questionnaire was developed to gather the eligible participants' demographic characteristics. To assess the Clear Fear app and its effects, participants were asked to complete 3 outcome measures to assess their emotional state or mood and behavior and the Clear Fear user questionnaire. Participants completed these 4 measures at stages 2 (post–app familiarization phase) and 3 (follow-up).

Outcome Assessments

In total, 3 self-report outcome measures were administered to participants to assess their symptoms of anxiety (main outcome) and depression and their behavior.

For anxiety, the 7-item Generalized Anxiety Disorder scale (GAD-7) was used, covering various areas, including social phobia, anxiety, and obsessive-compulsive disorder [14]. Total scores of 5, 10, and 15 indicate probable mild, moderate, and severe anxiety, respectively.

For depression, the 13-item self-report Mood and Feelings Questionnaire (MFQ) was used, with item ratings of *true* (score of 2), *sometimes true* (score of 1), or *not true* (score of 0) [15]. A total score of ≥12 may indicate depression.

For emotional and behavioral difficulties, the Strengths and Difficulties Questionnaire (SDQ) was used. This is a 25-item questionnaire to measure behavioral and emotional difficulties and assess psychological adjustment and mental health problems in children and young people [16].

The child and adolescent versions of the MFQ and the SDQ were used for participants aged between 16 and 17 years, and the adult version was used for those aged ≥18 years.

Clear Fear App Questionnaire

The Clear Fear app questionnaire, developed for the purposes of this study, was used to assess the Clear Fear app's safety, acceptability, satisfaction, and usability and included 21 items (of which 7 were closed ended, 9 were open ended for free-text responses, and 5 had Likert scales from 1-5).

For acceptability and user satisfaction, 10 items, including 8 open-ended questions, asked about what parts of the Clear Fear app were used, whether the content was useful, what features were liked or disliked and why, whether the app met the participants' expectations, and whether the app was unloaded.

For safety, 4 items asked whether the app helped reassure, reduce, or increase feelings of anxiety and whether the participants had to contact someone.

For usability, 7 items were related to how easy the Clear Fear app was to use, whether any technical issues were encountered, frequency of use, and at what time of the day it was used.

Data Analysis

Statistical analyses for all quantitative data were conducted using SPSS Statistics for Windows (version 29; IBM Corp). The demographic profile of participants and other quantitative data were described using frequencies, percentages, and means. The Shapiro-Wilk test of normality and visual assessment of *Q-Q* plots were used to check whether continuous variables for the outcome scores were normally distributed.

Mean scores derived from the outcome measures (GAD-7, MFQ, and SDQ) were compared between baseline and follow-up. Tests of significance were conducted using 2-tailed *t* tests or Wilcoxon signed rank tests (for continuous variables not distributed normally) to compare baseline and follow-up scores for all outcome and clinical measures. Standardized effect sizes were calculated using the Cohen *d* based on the mean change in outcome scores between baseline (stage 1) and follow-up (stage 3).

Qualitative data derived from open-ended questions were coded and entered into NVivo (version 10; QSR International) and analyzed using a thematic analysis of the text to identify themes and patterns in the data and how they addressed the main aims of the study [17]. Additional analyses (means and SDs) were conducted to examine any differences in total outcome scores for the GAD-7 and MFQ among age, gender, and ethnic groups. Differences among these groups were also explored through qualitative feedback on the Clear Fear app's acceptability using thematic analysis and categorization according to age, gender, and ethnic group.

Ethical Considerations

Ethics approval for this study was obtained from the London – City and East Research Ethics Committee (United Kingdom; 20/PR/0268), and all practices and procedures throughout the research process were in line with the stipulations of the Declaration of Helsinki.

This included that eligible participants be suitably informed of the study details, risks, and benefits via an information sheet before being asked to consent to the study. Only those who thus provided their informed consent went on to enter the study.

Participants' data were anonymized, collected and stored securely, and managed solely by the researchers. Participants were allocated a study ID number by a project administrator separate from the researchers.

Finally, participants were given a £20 (US $24.44) Amazon voucher as an honorarium upon completion of the final stage.

XSL•FO
RenderX

Results

Overview

Recruitment for the study took place between February 2021 and June 2022. A total of 48 young people completed the baseline measures. At the postfamiliarization phase (stage 2), 92% (44/48) of the participants completed all measures and the Clear Fear app user questionnaire. At the 8-week follow-up (stage 3), 77% (37/48) of the participants submitted all required responses, constituting a dropout rate of 23% (11/48).

Demographic Characteristics

The demographic characteristics of the sample are listed in Table 1. The mean age of the sample was 20.1 (SD 2.1) years, with 62% (30/48) aged between 18 and 25 years. The sample was predominantly female (37/48, 77%). Most were students, with a third (16/48, 33%) attending secondary school and 29% (14/48) at university. In terms of ethnicity, 67% (32/48) were White British, and 10% (5/48) were Black British.

Table 1. Characteristics of the participants at baseline (stage 1) of this uncontrolled pre– and post–follow-up study to evaluate the Clear Fear app's usability, acceptability, safety, and effectiveness in helping young people manage symptoms of anxiety using the principles of cognitive behavioral therapy over a 9-week period (N=48).

Characteristics	Values
Age (y), mean (SD; range)	20.1 (2.1; 17-24)
16-17, n (%)	18 (38)
18-25, n (%)	30 (62)
Sex, n (%)	
Female	37 (77)
Male	11 (23)
Marital status, n (%)	
Single	41 (85)
Married or cohabiting	4 (8)
Other long-term relationship	3 (6)
Ethnicity, n (%)	
Asian or Asian British	2 (4)
Black British	5 (10)
Mixed heritage	3 (6)
White British	32 (67)
White other	3 (6)
Other	3 (6)
Student status, n (%)	
Attending secondary school	16 (33)
Attending university	14 (29)
Attending sixth form or college	10 (21)
Not currently a student	8 (17)
Employment status, n (%)	
Student	33 (69)
Employed part time	6 (12)
Employed full time	5 (10)
Unemployed	2 (4)
Other	2 (4)

Baseline Scores

Table 2 lists the numbers and percentages of participants with probable symptoms of anxiety and depression and emotional and behavioral difficulties at baseline that would require further assessment. In total, 48% (23/48) of the participants reached the threshold for probable anxiety disorder as measured using the GAD-7, with 29% (14/48) finding it *very difficult* or *extremely difficult* to go about their daily routine or get along with others.

XSL•FO
RenderX

On the basis of the short MFQ measure, 56% (27/48) of the participants obtained a positive score for probable depression. A total of 75% (36/48) obtained *very high* on the SDQ total score, with 52% (25/48) of participants reporting their problems to be considerably distressing (*quite a lot* or *a great deal*). The ability to learn (23/48, 48%) and get along with others close to them (21/48, 44%) was also impacted for many.

Table 2. Numbers and percentages of participants meeting the threshold for probable symptoms of anxiety and depression and emotional and behavioral difficulties at baseline (stage 1) and follow-up (stage 3) following use of the Clear Fear app over a 9-week period to evaluate the app's usability, acceptability, safety, and effectiveness in helping young people manage symptoms of anxiety using the principles of cognitive behavioral therapy.

	Baseline (stage 1; n=48), n (%)	Follow-up (stage 3; n=37), n (%)
Generalized anxiety disorder (GAD-7[a])[b]	23 (48)	8 (22)
Mood and feelings (short MFQ[c])[d]	27 (56)	8 (22)
Emotional and behavioral difficulties (SDQ[e])[f]	36 (75)	23 (62)
Definite or severe problem	19 (40)	10 (27)
Problem lasting >12 months	19 (40)	14 (38)
Problem causing distress[g]	25 (52)	12 (32)
Difficulties getting along with close family or friends	21 (44)	14 (38)
Difficulties making and keeping friends	14 (29)	12 (32)
Difficulties with learning, studying, or work	23 (48)	17 (46)
Difficulties with hobbies and leisure	14 (29)	13 (35)
Difficulties impacting on others	19 (40)	6 (16)

[a]GAD-7: 7-item Generalized Anxiety Disorder scale.
[b]Total score of ≥10.
[c]MFQ: Mood and Feelings Questionnaire.
[d]Total score of ≥12.
[e]SQD: Strengths and Difficulties Questionnaire.
[f]Total score of 20 to 40 (*very high*).
[g]Those reporting quite a lot to a great deal.

Postfamiliarization Phase (Stage 2): Clear Fear App Use

A total of 92% (44/48) of the participants completed the self-report measures and questionnaires at stage 2. In total, 77% (34/44) of the participants used iOS (iPhone) as their operating system, with the remainder (10/44, 23%) using an Android phone. Only 5% (2/44) of the participants encountered technical issues when using the Clear Fear app (eg, poor display or phone freezing).

Participants used the Clear Fear app quite frequently; for many, it was several times a week (18/44, 41%) or every day (11/44, 25%). The evening was the preferred time for over half (24/44, 55%) of the participants.

Commonly used parts of the Clear Fear app included Anxiety Types (25/44, 57%), Clear Your Fear (23/44, 52%), and Self-Monitoring (20/44, 45%; Table 3). Parts of the Clear Fear app that were used less included Express Yourself (6/44, 14%), Laugh And Smile (6/44, 14%), information (6/44, 14%), Stop Overdoing Things (4/44, 9%), and Set an Exercise Goal (4/44, 9%).

XSL•FO
RenderX

Table 3. Participants' use of the Clear Fear app at the postfamiliarization phase (stage 2; 1 week following baseline) of this uncontrolled pre– and post–follow-up study to evaluate the app's usability, acceptability, safety, and effectiveness in helping young people manage symptoms of anxiety using the principles of cognitive behavioral therapy over a 9-week period (N=44).

	Participants, n (%)
Frequency of use	
More than once a day	2 (5)
Every day	11 (25)
Several times a week	18 (41)
Weekly	10 (23)
Less than once a week	3 (7)
Parts of the app used most	
Anxiety Types	25 (57)
Clear Your Fear	23 (52)
Self-Monitoring	20 (45)
Stay Calm	14 (32)
Dealing With Your Emotions	13 (30)
Managing Your Worries	12 (27)
Safety Net	12 (27)
Combat Negative Thoughts	10 (23)
Immediate Help	8 (18)
Managing Physical Responses to Anxiety	8 (18)

Follow-Up (Stage 3)

Usability and Acceptability

Overview

At follow-up (stage 3), 77% (37/48) of the participants completed the Clear Fear app user questionnaire. A total of 70% (26/37) of the participants reported using the Clear Fear app to address and manage feelings of anxiety, panic attacks, fear, stress, and worry. Participants were asked what they did before using the Clear Fear app to manage these symptoms. Most (23/37, 62%) used positive distractions, such as talking to family members, exercising, or other activities. The remaining participants either did nothing or used negative distractions such as "be in a daze."

The use pattern of the Clear Fear app was similar to that in stage 2 as the most used parts continued to be Anxiety Types (20/37, 54%), Clear Your Fear (24/37, 65%), and Self-Monitoring (16/37, 43%) and the least used parts were Express Yourself (4/37, 11%) and Stop Overdoing Things (4/37, 11%). Participants again mostly used the Clear Fear app either several times a week (17/37, 46%) or weekly (14/37, 38%) and in the evening (15/37, 41%). Fewer participants reported using the Clear Fear app every day at stage 3 compared to stage 2 (4/37, 11% vs 11/44, 25%, respectively). Most participants (31/37, 84%) found the Clear Fear app easy or very easy to use.

In total, 4 key themes were derived from the qualitative data (open-ended text responses). These included the Clear Fear app's interactive approach, active exercises, navigation, and app look and content.

Interactive Approach

When asked what features of the Clear Fear app were liked the most and why, a common theme appeared to include its interactive and direct approach while simultaneously providing information and guidance on how to manage feelings of anxiety and fear:

> *I liked the way the app asked what you were afraid of and then what you could do to battle that. I liked the way the questions made you face your fears and really think about it and not hide them.* [P21; female; student; aged 20 years]

This interactive approach gave one participant a sense of being more in control:

> *I like things where I actively did something with my anxiety, such as putting worries away, breathing exercises, clear fear, etc. The active involvement made me feel more in control.* [P59; female; volunteer; aged 24 years]

Active Exercises

The range of different activities, including breathing and writing exercises, to guide participants through their negative thoughts and feelings was well received. The immediate help feature and the available activities were considered helpful:

> *I liked using the immediate help feature and doing the breathing activities and panic attack feature because it helped to calm me down while having a panic attack. I also liked the safety net feature to*

XSL•FO
RenderX

remind myself what help I could get. [P65; female; student; aged 18 years]

Encouraging participants to seek help when they needed it was another important feature of the Clear Fear app, together with offering practical and immediate strategies to manage their anxiety.

Navigation

The app is generally great and easy to use, and very helpful. [P61; male; working part time; aged 21 years]

While most participants found the Clear Fear app easy to use and follow, a small minority (5/37, 14%) found the Clear Fear app difficult to use. A total of 60% (3/5) of these participants found the app difficult to navigate and asked for clearer instructions for finding activities:

The app was a little confusing to navigate on the home page, I found it confusing to go to the activities because there was no instruction as to how to do that. [P32; female; student; aged 17 years]

Some features take a long time to find and I forgot which buttons I needed to press to get to them. [P35; female; student; aged 17 years]

Another 40% (2/5) mentioned that they wanted to be able to make entries to the journal part of the app without having to carry out the activities first:

I couldn't find a way to always access the journal. My entries were after activities where it gives you the option to write. I'd love there to be a more accessible journal. [P59; female; other; aged 22 years]

In total, 2 features of Clear Fear, Grit Box and Laugh and Smile, were considered less helpful by a small minority (4/37, 11%).

App Look and Content

While the Clear Fear app had content that participants found useful, some preferred not to have to read so much. One participant described expecting the Clear Fear app to have a better layout and look, with a less clinical appearance:

I expected the physical layout and colour schemes and fonts to look more calming and relaxing. The white backgrounds and lack of obvious clickable buttons made it look very clinical. To some extent the confusing layout made me feel more anxious when I used it. [P29; female; student; aged 18 years]

Another participant thought that the Clear Fear app would have a means through which to communicate with someone or something:

I thought that there would be some sort of communication with something or someone like an area where you can talk to those cute creature things and have them say nice messages. [P65; female; working full time; aged 21 years]

There was also a suggestion to have the Clear Fear app speak, perhaps to offer reassurance or include games:

I wanted more options of help than just writing everything down and having to read the help. E.g. videos to watch on the topic, music to calm thoughts and someone speaking the words. [P30; female; student; aged 18 years]

Safety

Table 4 lists the numbers and percentages of participants concerning their responses to questions about the Clear Fear app's safety. Many participants found that the Clear Fear app reassured them either *a lot* (11/37, 30%) or *somewhat* (16/37, 43%). Similarly, the Clear Fear app was said to help reduce feelings of anxiety by *a lot* (13/37, 35%) or *somewhat* (14/37, 38%).

Table 4. Participants' responses regarding the safety of the Clear Fear app following 8 weeks of use (at stage 3) in this uncontrolled pre– and post–follow-up study to evaluate the app's usability, acceptability, safety, and effectiveness in helping young people manage symptoms of anxiety using the principles of cognitive behavioral therapy (N=37).

Questions and responses	Participants, n (%)
Did the Clear Fear app help reassure you when feeling anxious?	
Yes, a lot	11 (30)
Yes, somewhat	16 (43)
Yes, a little	5 (14)
No, not at all	2 (5)
Did the Clear Fear app reduce your feelings of anxiety?	
Yes, a lot	13 (35)
Yes, somewhat	14 (38)
Yes, a little	7 (19)
No, not at all	2 (5)
Did the Clear Fear app help you contact someone to speak to about your anxiety?	
Yes	21 (57)
No	16 (43)

XSL•FO
RenderX

Just over half (21/37, 57%) of the participants who responded to this question reported that the Clear Fear app helped them speak to someone about their anxiety. However, 43% (16/37) said that the Clear Fear app did not help them contact someone.

Qualitative User Feedback and Differences by Age, Gender, and Ethnicity

Some differences were found among age, gender, and ethnic groups from the qualitative data gathered from 3 user feedback questions concerning the acceptability of the Clear Fear app—which parts were most liked and disliked and what users were expecting from the app. Younger users (aged 16-18 years) preferred guided, calming features such as Clear Your Fear, whereas older users (aged 19-25 years) appreciated action-oriented, self-reflective tools such as the worry box and anxiety management features. The small sample of men focused on more immediate relief tools, such as breathing exercises and the Laugh and Smile feature, whereas female participants opted more for features that allow for emotional processing (eg, writing and emotional guidance). Regarding ethnicity, Black British users highlighted a focus on privacy and control, and Asian British participants valued ease of use and humor.

There were few differences between demographic groups for the small number of participants who noted parts of the app that they disliked (5/37, 14%). Younger users (aged 16-18 years) appeared less satisfied with navigating the app and features such as Safety Net or Laugh and Smile. Male respondents were less dissatisfied with the Clear Fear app, although they provided very brief responses. Black British users found the app's interface and navigation difficult and preferred customizable features, whereas Asian British participants provided limited negative feedback.

Managing emotions, anxiety, and coping strategies were the main expectations from the app across each of the age, gender, and ethnic groups.

Effectiveness of the Clear Fear App: Outcomes at Follow-Up (Stage 3)

Tables 5 and 6 list mean scores (and subscores for the SDQ), mean differences, standardized effect sizes, 95% CIs, and *P* values for each outcome measure at baseline and follow-up. At follow-up, 8 weeks after stage 2, statistically significant reductions were found in all mean total scores for symptoms of anxiety (GAD-7; t_{36}=2.6, 95% CI 0.41-3.53; *P*=.01) and depression (short MFQ; *z*=2.3; *P*=.02) and emotional and behavioral difficulties (SDQ; t_{47}=4.5, 95% CI 3.67-9.65; *P*<.001). The mean change in mean symptom scores revealed mostly medium to large standardized effect sizes.

Subscores for the SDQ also revealed significant decreases in emotional problems (t_{47}=4.5, 95% CI 1.14-2.99; *P*<.001), conduct problems (t_{47}=4.2, 95% CI 0.64-1.82; *P*<.001), hyperactivity (t_{47}=2.3, 95% CI 0.11-1.59; *P*=.02), and peer problems (t_{47}=1.6, 95% CI 0.39-1.61; *P*=.002).

Table 5. Baseline and follow-up (9 weeks later) mean comparisons, mean differences, and standardized effect sizes for participants' anxiety, depression, and behavioral and emotional symptom scores following their use of the Clear Fear app in this uncontrolled pre– and post–follow-up study to evaluate the app's usability, acceptability, safety, and effectiveness in helping young people manage symptoms of anxiety using the principles of cognitive behavioral therapy.

Outcome measure	Baseline score (stage 1; n=48), mean (SD)	Follow-up score (stage 3; n=37), mean (SD)	Mean difference (SD)	Standardized effect size, Cohen d_z
Probable anxiety disorder (GAD-7[a] total)	9.4 (4.5)	7.5 (3.7)	–1.9 (4.1)	0.46
Probable depression (short MFQ[b] total)	11.3 (5.4)	8.7 (5.1)	–2.6 (5.2)	0.50
Behavioral and emotional difficulties	23.3 (4.8)	16.6 (9.9)	–6.7 (8.3)	0.80
Emotional problems	5.9 (2.1)	3.9 (2.9)	–2.0 (2.5)	0.80
Conduct problems	2.7 (1.8)	1.5 (1.2)	–1.2 (1.5)	0.79
Hyperactivity	4.1 (1.9)	3.3 (2.7)	–0.8 (2.3)	0.34
Peer problems	3.1 (1.8)	2.1 (1.9)	–1.0 (1.8)	0.54

[a]GAD-7: 7-item Generalized Anxiety Disorder scale.

[b]MFQ: Mood and Feelings Questionnaire.

XSL•FO
RenderX

Table 6. Baseline and follow-up (9 weeks later) test of significance of the results for participants' anxiety, depression, and behavioral and emotional symptom scores following their use of the Clear Fear app in this uncontrolled pre– and post–follow-up study to evaluate the app's usability, acceptability, safety, and effectiveness in helping young people manage symptoms of anxiety using the principles of cognitive behavioral therapy.

Outcome measure	Test of significance	*P* value
Probable anxiety disorder (GAD-7[a] total)	t_{36}=2.6 (95% CI 0.41-3.53)	.01
Probable depression (short MFQ[b] total)	z=2.3[c]	.02
Behavioral and emotional difficulties (SDQ[d] total)	t_{47}=4.5 (95% CI 3.67-9.65)	<.001
Emotional problems	t_{47}=4.5 (95% CI 1.14-2.99)	<.001
Conduct problems	t_{47}=4.2 (95% CI 0.64-1.82)	<.001
Hyperactivity	t_{47}=2.3 (95% CI 0.11-1.59)	.02
Peer problems	t_{47}=1.6 (95% CI 0.39-1.61)	.002

[a]GAD-7: 7-item Generalized Anxiety Disorder scale.

[b]MFQ: Mood and Feelings Questionnaire.

[c]Z score from Wilcoxon signed rank test.

[d]SDQ: Strengths and Difficulties Questionnaire.

Differences in Anxiety and Depression Symptoms by Age, Gender, and Ethnicity

Few differences in the mean total outcome scores were found for the GAD-7 and MFQ when examined by age, gender, and ethnic group. Almost all groups showed improved anxiety and depression symptom scores at stage 3. The exceptions were Black British (5/37, 14%) and mixed-heritage (3/37, 8%) ethnic groups, whose GAD-7 mean scores increased slightly between stages 1 and 3 (from 13.0 to 13.3 and 11.3 to 13.0, respectively). Black British participants also had a slightly increased MFQ mean score at stage 3 (from 13.0 to 13.3). The greatest improvements in mean scores were a mean reduction of –5.0 (SD 9.8) in the GAD-7 for White other individuals (3/37, 8%) and mean reductions in the MFQ of –5.3 (SD 5.3) for male individuals (11/37, 30%), –6.0 (SD 11.3) for White other individuals (3/37, 8%), and –4.4 (SD 7.8) for those aged 18 to 25 years (25/37, 68%).

Discussion

Principal Findings

This paper details the development and evaluation of the Clear Fear app. The main findings on usability, acceptability, safety, and effectiveness were positive overall. The app was effective in reducing probable symptoms of anxiety and depression and emotional and behavioral difficulties. Therefore, this appears to support the hypothesis that a self-guided mobile cognitive behavioral app can improve anxiety symptoms and be acceptable to young people over a period of 9 weeks. However, the limitations of this study (outlined in a later section) should be noted when interpreting these findings.

A post hoc power analysis was conducted using G*Power (version 3.1.9.7) to check whether the sample size used in this study was large enough to facilitate sufficient statistical power for the analyses conducted [18]. The sample of 48 participants at baseline (with a medium effect size of 0.5 and an α of .05) and the sample at stage 3 (follow-up; n=37) provided 90% power and 80% power, respectively. Therefore, the study was adequately powered for the statistical analyses conducted.

Comparison With Prior Work

Usability

Certain parts of the Clear Fear app were frequently used both at stage 2 and stage 3, including Anxiety Types, Clear Your Fear, and Self-Monitoring, features that perhaps targeted participants' specific needs and symptoms. Over three-quarters of participants at stage 3 (31/37, 84%) reported the Clear Fear app to be easy to use. Patterns of Clear Fear app use remained broadly similar at stages 2 and 3, demonstrating consistent use throughout the study. Other similar studies have reported high use of mental health apps among young people during the first week, which often declines over time [19,20].

A very small minority of participants (5/48, 10%) found the Clear Fear app difficult to use. The app contains many different features and activities, perhaps too many, which, for a minority of participants, made it confusing to navigate and, subsequently, less useful and helpful when trying to find or retrieve parts. The information given and some of the requirements, such as writing down goals and ways to achieve them, may be overwhelming for some. Striking the right balance between ease of use and informativeness is important [21]. Therefore, based on the feedback from a small minority of study participants, future updates of the Clear Fear app could ensure easier navigation and direct access to features if required.

Popular and frequently used features of the Clear Fear app included Anxiety Types, Clear Your Fear, and Self-Monitoring. This appeared to show some participants' preference for dealing head-on with their feelings of anxiety and fear. The self-monitoring feature provides important emotional awareness to track and reflect on feelings and emotions to identify which promote positive or negative moods [19]. Previous research has also shown that self-monitoring mood is commonly included in many mobile mental health apps for young people and is the most liked and engaged with section [22,23].

XSL•FO
RenderX

Qualitative Feedback

The key themes derived from the open-ended responses to the Clear Fear app user questionnaire revealed which aspects participants found useful and what they liked and disliked. The Clear Fear app's approach to tackling feelings of anxiety and fear appeared beneficial. The exercises (breathing, meditation, and journal writing) within the Clear Fear app were also beneficial in helping some participants calm down from a panic attack. These exercises and other distraction strategies are important for managing anxiety. Some participants also suggested adding videos to watch, calming music, games, and someone speaking reassuring words, which could potentially enhance the Clear Fear app. Video games have been shown to be a useful therapeutic distraction, which helps address and treat symptoms of anxiety [24].

When selecting mobile mental health apps, young people have been found to prefer those that are accessible, secure, and evidence based [25]. The Clear Fear app meets these important criteria.

Safety

The Clear Fear app also appeared safe to use, with nearly three-quarters of participants (31/37, 84%) reporting that it offered a lot or some reassurance when feeling anxious or fearful. These reports were confirmed by the pre– and post–follow-up scores for the outcome measures. The "immediate help" feature was an important reminder for some participants to seek help when they needed it while in the meantime providing information and practical strategies for dealing with their anxiety and fear. Encouragingly, over half (21/37, 57%) of the participants reported that the Clear Fear app helped them speak to someone about their anxiety.

Effectiveness

At the beginning of this study, just under half (23/48, 48%) of the participants met the threshold for probable symptoms of anxiety disorder, over half (27/48, 56%) had probable depression, and three-quarters (36/48, 75%) obtained a score of *very high* for emotional and behavioral difficulties. By the end of the study, this had decreased to just under a quarter (8/37, 22%) of participants meeting the threshold for probable anxiety disorder, the same proportion (8/37, 22%) for probable depression, and just under two-thirds (23/37, 62%) scoring *very high* for emotional and behavioral difficulties. Consequently, the Clear Fear app appeared to significantly reduce scores for probable anxiety disorder (P=.01), probable depression (P=.02), and behavioral and emotional difficulties (P<.001) after 8 weeks of use. Moreover, at stage 3 of this study, far fewer participants met the threshold for symptoms of probable anxiety disorder and depression in particular.

These results add to the limited available evidence of the effectiveness of app interventions targeting anxiety and depression in young people and confirm previous findings that such apps, according to one review, can produce small to medium positive effect sizes [26]. Despite being distinct conditions, there is a strong relationship between anxiety and depression, which can coexist in young people [27]. This may explain the improvement in depression scores found in this study even though the Clear Fear app is designed for the management of anxiety symptoms. This improvement may be due to several reasons. Anxiety and depression often share common risk and maintaining factors, and addressing anxiety may help with some of the factors that are common to depression. Lowered anxiety may also help improve a person's self-esteem and confidence; enhance social connection; and help with focus, purpose, and application to work, all contributing to enhanced mood.

Differences by Age, Gender, and Ethnic Group

Generally, the Clear Fear app revealed few differences across the age, gender, and ethnic groups examined. However, the differences found highlighted the need to improve the app's navigation and refine less popular features for Black British and younger users. Interestingly, Black British users had slightly increased scores at follow-up for anxiety and depression symptoms, which underlines the need for more culturally appropriate content.

Limitations

The small sample of participants prevented further statistical tests to compare differences among age, gender, and ethnic groups. However, such differences were further examined in terms of mean outcomes and qualitative feedback (a strength of this study) to understand the user experiences of the Clear Fear app according to age, gender, and ethnicity.

The Clear Fear app is suitable for ages of ≥11 years; however, participants in the study were aged between 16 and 25 years, making the study findings generalizable only to this age group. Children and adolescents may respond differently to the Clear Fear app layouts, interaction styles, and content than those aged 16 to 25 years.

It is difficult to know what biases may have been introduced during recruitment, for example, how participants self-selected to take part in the study, the different avenues of information (eg, how participants found out about the study and the Clear Fear app), and their motivation to take part. The offer of an incentive (a shopping voucher) could have led to some form of selection bias regarding those who took part in the study. However, the attrition rates were low, with over three-quarters of participants (37/48, 77%) remaining in the study at follow-up, preventing selective biases associated with high dropout rates.

The follow-up of 8 weeks was a relatively short period, and a lack of a control group also presents further limitations in the study design. However, this exploratory uncontrolled pretest-posttest design provides promising initial findings; some understanding of how the Clear Fear app was used; its acceptability; and potential effects on reducing symptoms of anxiety, depression, and emotional and behavioral difficulties. This study was also adequately powered to conduct significance tests on the main outcome after carrying out a post hoc G*Power calculation.

Future Research

Future studies could benefit from a more robust evaluation design with a larger and more diverse sample. A randomized controlled trial with one or more comparison groups (including

XSL•FO
RenderX

a control group) would help determine causality and control for potential confounders and other external influencing factors. This combined with recruiting participants from a wider representation, including children and adolescents (between the ages of 11 and 15 years), could extend the generalizability of this study's findings. Incorporating clinician-administered outcome measures and extending the follow-up period beyond 8 weeks would enhance understanding of the Clear Fear app's effectiveness and user engagement over time (eg, over 3, 6, and ≥12 months). Future studies could also help determine how and why the Clear Fear app helped with symptoms of depression.

Future studies could also examine the Clear Fear app's effectiveness across different cultural contexts to reveal how cultural norms and societal pressures impact user engagement and outcomes. This could help identify any necessary adaptations to the app's content or functionality to make it culturally sensitive and relevant.

Similarly, future studies could explore how the Clear Fear app can be integrated with specialist mental health services. This could involve studying the app's role in supporting young adults on treatment waiting lists through enhancing engagement, complementing face-to-face therapy, being administered as a stand-alone intervention, and helping reduce increased workloads for health care providers. Given the need for ongoing support following discharge and in preventing relapse, the Clear Fear app may also have a function in helping young adults stay well.

In addition, further research could focus on understanding the perceptions and acceptance of digital health tools among both patients generally and clinicians in various health care settings, which may vary widely based on local practices and resources. This research will help refine the development of digital interventions, ensuring that they are both accessible and effective for diverse populations and integrated within existing health care systems.

Conclusions

The findings indicate that the Clear Fear app is an accessible, easy-to-use, safe, and effective early intervention for young people experiencing mild to moderate anxiety and depression. Therefore, this study appears to support the hypothesis that a self-guided mobile cognitive behavioral app can improve anxiety symptoms and be acceptable to young people over a period of 9 weeks. However, it is important to note this study's limitations, which may restrict the validity and generalizability of these findings, particularly until further research is conducted with larger samples of young people and specifically with children aged between 11 and 15 years to continue to assess the Clear Fear app's functioning and use. It is also important to examine its effectiveness over a longer period, using more robust and possibly alternative methodologies to better understand its use and compliance. The quantitative and qualitative feedback from participants also provide important information for further improvements to the Clear Fear app through future updates.

Acknowledgments

The authors warmly thank all the young people who participated in this study and provided their feedback on the Clear Fear app. They are also deeply grateful to the young people who gave their time to provide feedback on the Clear Fear app during its development. The authors are immensely grateful for the support from Comic Relief and the Paul Hamlyn Foundation's Tech for Good fund for funding the development of the Clear Fear app and to the Nominet #RESET mental health program for funding its evaluation.

Data Availability

The datasets generated and analyzed during this study are not publicly available due to conditions outlined as part of the participants' informed consent, as well as the data compliance regulations of stem4; however, they may be available from the corresponding author on reasonable request.

Authors' Contributions

Conception, app development, and coproduction were carried out by NK. Study recruitment and coordination were conducted by KD and RE. Study planning, design, and data collection and analysis were carried out by CS and NU. Manuscript preparation was carried out by all authors.

Conflicts of Interest

None declared.

References

1. Newlove-Delgado T, Marcheselli F, Williams T, Mandalia D, Davis J, McManus S, et al. Mental health of children and young people in England, 2022. NHS Digital. 2022. URL: https://digital.nhs.uk/data-and-information/publications/statistical/mental-health-of-children-and-young-people-in-england/2022-follow-up-to-the-2017-survey [accessed 2023-12-12]
2. Young people's well-being in the UK: 2020. Office for National Statistics. 2020 Oct 02. URL: https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/bulletins/youngpeopleswellbeingintheuk/2020 [accessed 2024-04-29]

XSL•FO
RenderX

3. Garratt K, Kirk-Wade E, Long R. Children and young people's mental health: policy and services (England). House of Commons Library. 2024 Jan. URL: https://commonslibrary.parliament.uk/research-briefings/cbp-7196/ [accessed 2024-11-15]
4. Grimm F, Alcock B, Butler JE, Crespo RF, Davies A, Peytrignetet S, et al. Briefing: Improving children and young people's mental health services. The Health Foundation. 2022 Jul. URL: https://www.health.org.uk/sites/default/files/upload/publications/2022/CYPMHbriefing_Web_Final.pdf [accessed 2023-12-12]
5. Hiley C. UK mobile phone statistics, 2023. Uswitch. 2023 Feb 01. URL: https://www.uswitch.com/mobiles/studies/mobile-statistics/ [accessed 2023-12-12]
6. Leech T, Dorstyn D, Taylor A, Wenjing L. Corrigendum to "mental health apps for adolescents and young adults: a systematic review of randomised controlled trials". Child Youth Serv Rev 2021 Sep;128:106184. [doi: 10.1016/j.childyouth.2021.106184]
7. Bear HA, Ayala Nunes L, Ramos G, Manchanda T, Fernandes B, Chabursky S, et al. The acceptability, engagement, and feasibility of mental health apps for marginalized and underserved young people: systematic review and qualitative study. J Med Internet Res 2024 Jul 30;26:e48964 [FREE Full text] [doi: 10.2196/48964] [Medline: 39078699]
8. Wang K, Varma DS, Prosperi M. A systematic review of the effectiveness of mobile apps for monitoring and management of mental health symptoms or disorders. J Psychiatr Res 2018 Dec;107:73-78. [doi: 10.1016/j.jpsychires.2018.10.006] [Medline: 30347316]
9. Andrews G, Basu A, Cuijpers P, Craske MG, McEvoy P, English CL, et al. Computer therapy for the anxiety and depression disorders is effective, acceptable and practical health care: an updated meta-analysis. J Anxiety Disord 2018 Apr;55:70-78 [FREE Full text] [doi: 10.1016/j.janxdis.2018.01.001] [Medline: 29422409]
10. Bergin AD, Perez Vallejos E, Davies EB, Daley D, Ford T, Harold G, et al. Preventive digital mental health interventions for children and young people: a review of the design and reporting of research. NPJ Digit Med 2020;3:133-020 [FREE Full text] [doi: 10.1038/s41746-020-00339-7] [Medline: 33083568]
11. Goldin PR, Gross JJ. Effects of mindfulness-based stress reduction (MBSR) on emotion regulation in social anxiety disorder. Emotion 2010 Feb;10(1):83-91 [FREE Full text] [doi: 10.1037/a0018441] [Medline: 20141305]
12. Sucala M, Cuijpers P, Muench F, Cardo R, Soflau R, Dobrean A, et al. Anxiety: there is an app for that. A systematic review of anxiety apps. Depress Anxiety 2017 Jun;34(6):518-525. [doi: 10.1002/da.22654] [Medline: 28504859]
13. Bry LJ, Chou T, Miguel E, Comer JS. Consumer smartphone apps marketed for child and adolescent anxiety: a systematic review and content analysis. Behav Ther 2018 Mar;49(2):249-261 [FREE Full text] [doi: 10.1016/j.beth.2017.07.008] [Medline: 29530263]
14. Chorpita BF, Moffitt CE, Gray J. Psychometric properties of the revised child anxiety and depression scale in a clinical sample. Behav Res Ther 2005 Mar;43(3):309-322. [doi: 10.1016/j.brat.2004.02.004] [Medline: 15680928]
15. Daviss WB, Birmaher B, Melhem NA, Axelson DA, Michaels SM, Brent DA. Criterion validity of the mood and feelings questionnaire for depressive episodes in clinic and non-clinic subjects. J Child Psychol Psychiatry 2006 Sep;47(9):927-934. [doi: 10.1111/j.1469-7610.2006.01646.x] [Medline: 16930387]
16. Goodman R. The strengths and difficulties questionnaire: a research note. J Child Psychol Psychiatry 2006;38(5):581-586. [doi: 10.1111/j.1469-7610.1997.tb01545.x] [Medline: 9255702]
17. Braun V, Clarke V. Thematic Analysis: A Practical Guide. Thousand Oaks, CA: Sage Publications; 2021.
18. Faul F, Erdfelder E, Lang AG, Buchner A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods 2007 May;39(2):175-191. [doi: 10.3758/bf03193146] [Medline: 17695343]
19. Pramana G, Parmanto B, Kendall PC, Silk JS. The SmartCAT: an m-health platform for ecological momentary intervention in child anxiety treatment. Telemed J E Health 2014 May;20(5):419-427 [FREE Full text] [doi: 10.1089/tmj.2013.0214] [Medline: 24579913]
20. Reid SC, Kauer SD, Khor AS, Hearps SJ, Sanci LA, Kennedy AD, et al. Using a mobile phone application in youth mental health - an evaluation study. Aust Fam Physician 2012 Sep;41(9):711-714. [Medline: 22962650]
21. Garrido S, Cheers D, Boydell K, Nguyen QV, Schubert E, Dunne L, et al. Young people's response to six smartphone apps for anxiety and depression: focus group study. JMIR Ment Health 2019 Oct 02;6(10):e14385 [FREE Full text] [doi: 10.2196/14385] [Medline: 31579023]
22. Kenny R, Dooley B, Fitzgerald A. Feasibility of "copesmart": a telemental health app for adolescents. JMIR Ment Health 2015 Aug 10;2(3):e22 [FREE Full text] [doi: 10.2196/mental.4370] [Medline: 26552425]
23. Melbye S, Kessing LV, Bardram JE, Faurholt-Jepsen M. Smartphone-based self-monitoring, treatment, and automatically generated data in children, adolescents, and young adults with psychiatric disorders: systematic review. JMIR Ment Health 2020 Oct 29;7(10):e17453 [FREE Full text] [doi: 10.2196/17453] [Medline: 33118950]
24. Kowal M, Conroy E, Ramsbottom N, Smithies T, Toth A, Campbell M. Gaming your mental health: a narrative review on mitigating symptoms of depression and anxiety using commercial video games. JMIR Serious Games 2021 Jun 16;9(2):e26575 [FREE Full text] [doi: 10.2196/26575] [Medline: 34132648]
25. Kabacińska K, McLeod K, MacKenzie A, Vu K, Cianfrone M, Tugwell A, et al. What criteria are young people using to select mobile mental health applications? A nominal group study. Digit Health 2022 May 29;8:20552076221102775 [FREE Full text] [doi: 10.1177/20552076221102775] [Medline: 35663239]

XSL•FO
RenderX

26. Lecomte T, Potvin S, Corbière M, Guay S, Samson C, Cloutier B, et al. Mobile apps for mental health issues: meta-review of meta-analyses. JMIR Mhealth Uhealth 2020 May 29;8(5):e17458 [FREE Full text] [doi: 10.2196/17458] [Medline: 32348289]
27. Axelson DA, Birmaher B. Relation between anxiety and depressive disorders in childhood and adolescence. Depress Anxiety 2001;14(2):67-78. [doi: 10.1002/da.1048] [Medline: 11668659]

Abbreviations

CBT: cognitive behavioral therapy
GAD-7: 7-item Generalized Anxiety Disorder scale
MFQ: Mood and Feelings Questionnaire
SDQ: Strengths and Difficulties Questionnaire

Edited by A Mavragani; submitted 18.12.23; peer-reviewed by S Mitra, J Abbas, A Futterman Collier; comments to author 07.10.24; revised version received 30.11.24; accepted 24.12.24; published 28.02.25.

Please cite as:
Samele C, Urquia N, Edwards R, Donnell K, Krause N
Evaluation of the Clear Fear Smartphone App for Young People Experiencing Anxiety: Uncontrolled Pre– and Post–Follow-Up Study
JMIR Form Res 2025;9:e55603
URL: https://formative.jmir.org/2025/1/e55603
doi:10.2196/55603
PMID:40053764

©Chiara Samele, Norman Urquia, Rachel Edwards, Katie Donnell, Nihara Krause. Originally published in JMIR Formative Research (https://formative.jmir.org), 28.02.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Digital Program for Daily Life Management With Endometriosis: Pilot Cohort Study on Symptoms and Quality of Life Among Participants

Zélia Breton[1,2]; Emilie Stern[3,4], MSc; Mathilde Pinault[1], PhD; Delphine Lhuillery[5,6], MD; Erick Petit[7,8], MD; Pierre Panel[9], MD; Maïa Alexaline[1], PhD

[1]Lyv Healthcare, Nantes, France
[2]Université Paris-Saclay, Gustave Roussy, Inserm, 94805, Villejuif, France
[3]GHU Paris, Psychiatry & Neurosciences, Paris, France
[4]Laboratoire de Psychopathologie et Processus de Santé, Université Paris Cité, Boulogne-Billancourt, France
[5]Paris Saint-Joseph Hospital Group, EndoSud Ile-de-France, Paris, France
[6]Clinique Oudinot, Paris, France
[7]Department of Radiology, Fondation Hôpital St Joseph, Paris, France
[8]Centre de l'Endométriose, RESENDO (Réseau Ville-Hopital Endometriose), Paris, France
[9]Department of Gynecology-Obstetrics, Centre Hospitalier de Versailles, Le Chesnay-Rocquencourt, France

Corresponding Author:
Maïa Alexaline, PhD
Lyv Healthcare
6 rue Edouard Nignon
Nantes, 44300
France
Phone: 33 650208680
Email: maia@lyv.app

Abstract

Background: After experiencing symptoms for an average of 7 years before diagnosis, patients with endometriosis are usually left with more questions than answers about managing their symptoms in the absence of a cure. To help women with endometriosis after their diagnosis, we developed a digital program combining user research, evidence-based medicine, and clinical expertise. Structured around cognitive behavioral therapy and the quality of life metrics from the Endometriosis Health Profile score, the program was designed to guide participants for 3 months.

Objective: This cohort study was designed to measure the impact of a digital health program on the symptoms and quality of life levels of women with endometriosis.

Methods: In total, 63% (92/146) of the participants were included in the pilot study, recruited either free of charge through employer health insurance or via individual direct access. A control group of 404 women with endometriosis who did not follow the program, recruited through social media and mailing campaigns, was sampled (n=149, 36.9%) according to initial pain levels to ensure a similar pain profile to participants. Questionnaires assessing quality of life and symptom levels were emailed to both groups at baseline and 3 months. Descriptive statistics and statistical tests were used to analyze intragroup and intergroup differences, with Cohen *d* measuring effect sizes for significant results.

Results: Over 3 months, participants showed substantial improvements in global symptom burden, general pain level, anxiety, depression, dysmenorrhea, dysuria, chronic fatigue, neuropathic pain, and endo belly. These improvements were significantly different from the control group for global symptom burden (participants: mean –0.7, SD 1.6; controls: mean –0.3, SD 1.3; P=.048; small effect size), anxiety (participants: mean –1.1, SD 2.8; controls: mean 0.2, SD 2.5; P<.001; medium effect size), depression (participants: mean –0.9, SD 2.5; controls: mean 0.0, SD 3.1; P=.04; small effect size), neuropathic pain (participants: mean –1.0, SD 2.7; controls: mean –0.1, SD 2.6; P=.004; small effect size), and endo belly (participants: mean –0.9, SD 2.5; controls: mean –0.3, SD 2.4; P=.03; small effect size). Participants' quality of life improved between baseline and 3 months and significantly differed from that of the control group for the core part of the Endometriosis Health Profile-5 (participants: mean –5.9, SD 21.0; controls: mean 1.0, SD 14.8; P=.03; small effect size) and the EQ-5D (participants: mean 0.1, SD 0.1; controls: mean –0.0, SD

XSL•FO
RenderX

0.1; *P*=.001; medium effect size). Perceived knowledge of endometriosis was significantly greater at 3 months among participants compared to the control group (*P*<.001).

Conclusions: This study's results suggest that a digital health program providing medical and scientific information about endometriosis and multidisciplinary self-management tools may be useful to reduce global symptom burden, anxiety, depression, neuropathic pain, and endo belly while improving knowledge on endometriosis and quality of life among participants.

(JMIR Form Res 2025;9:e58262) doi:10.2196/58262

KEYWORDS

digital program; endometriosis; integrative therapies; quality of life; nonpharmacological intervention; daily life; adenomyosis; lesions; women; digital; France; case control; digital health; pilot study; control group; global symptom; burden; depression; neuropathic pain; chronic; multidisciplinary; web based; mobile health; mHealth; intervention

Introduction

Background

Endometriosis is a chronic, inflammatory disease affecting an estimated 1 in 10 women of childbearing age. It is defined by the presence of endometrial-like tissue outside the uterine cavity [1]. Manifestations of the disease are mainly painful (eg, dysmenorrhea, dyspareunia, dyschezia, and dysuria), with symptoms substantially impacting the quality of life (QOL) of those affected [2]. There is no definitive cure, and available solutions aim to reduce symptoms. The initial solutions are generally hormonal, which are not always well tolerated by patients [1], consistent with their fertility goals, or effective in relieving pain [3]. While surgery was once widely recommended to patients with endometriosis, given the frequent disease recurrence [4], it is now increasingly restricted to selected patients for whom it can be relevant in European countries [5]. Thus, living with endometriosis means learning to deal with the symptoms, using tools other than or in complement to hormonal therapy. After experiencing symptoms for an average of 7 years before diagnosis [6], the lack of an adapted care pathway for symptom management is a crucial unmet need of the patients [7]. Various nonpharmacological interventions have been studied to help with symptom management, such as dietary changes, physical activity, sex therapy, and mind-body interventions, which seem pertinent for the daily management of endometriosis [8-12]. Digital tools may offer a solution to enhance accessibility to and observance of nonpharmacological interventions toward endometriosis symptom reduction.

Prior Work

Developing a program to improve QOL requires the construction of content corresponding to what can modulate the QOL of women with endometriosis. Numerous studies to assess changes in patients' QOL have been using the Endometriosis Health Profile (EHP) score as a reference for almost 20 years [13]. The EHP is a standardized tool for measuring QOL. Initially developed with 30 items (EHP-30) [14], a shortened version with 5 items (ie, EHP-5) was also validated [15]. The following EHP components were taken into account in developing the programs: pain, control and powerlessness, emotional well-being, social support, self-image, work life, infertility and children management, sexual relationships, and relation to the medical profession.

Cognitive behavioral therapy (CBT) is a well-supported treatment for various disorders [16]. CBT's goal is to help individuals better understand the relationship and interaction between their emotions, cognitions (received ideas), and behavior to be able to manage their symptoms and quit the vicious cycle that contributing factors may create through behavior change (develop coping mechanisms), cognitive restructuring (restructure maladaptive thoughts), and emotion regulation (relaxation and acceptance) [17-19]. It was found to improve the QOL for patients with chronic pain conditions [20-23], especially by encouraging a problem-solving attitude in the patients receiving the treatment [24]. Internet-based CBT programs have been proven effective for a variety of disorders [25-29]. When considering CBT-based programs for chronic pain [18,30], the following components are found: psychoeducation on pain mechanisms and biopsychosocial model, goal settings, CBT-based behavior skills for managing pain (eg, pleasant activities and sleep hygiene), emotion regulation and mind-body interventions (eg, relaxation methods and mindfulness), cognitive restructuring, and long-term action plan for maintenance and anticipation of obstacles.

Goal of This Study

A digital solution for endometriosis symptom management was developed using a CBT approach based on EHP items. The digital support solution focused on the scientific and medical state of the art around 5 nonpharmacological interventions (disease education including pain mechanism, diet, adapted physical activity, well-being and mental health, and sexual health). The *School of Endo* digital program was developed and made available in France to help women with endometriosis after their diagnosis.

This cohort study was designed to measure the impact of a digital health program on the symptoms and QOL levels of women with endometriosis. The study also aimed to provide insights into the value of a digital program for the day-to-day management of endometriosis and to address the lack of real-life data studies for digital support in endometriosis.

Methods

Development of a Digital Health Program

To develop a digital health program that adequately addresses the needs of patients with endometriosis, several steps were followed, where product development and scientific research

XSL•FO
RenderX

complemented each other [31]. As the first step, 120 semistructured interviews and >100 web-based questionnaires with patients and health professionals were conducted to understand the needs of women with endometriosis. According to this user research, the primary need expressed by women with endometriosis was to find help with the daily management of symptoms, especially pain. A small exploratory preliminary program was developed and tested by women to complement the user research. A scientific rationale and advice from medical experts completed the user research. User research, evidence-based medicine, and clinical expertise were combined to enable the development of the program [32].

The structure of the program's content was based on CBT and the QOL items of the EHP score, and participants were advised to follow it for 3 months. The program was accessible from November 2022 for women with endometriosis aged >18 years, either at no fee through employer health insurance (*Mutuelle Générale de l'Education Nationale*) or individual direct payment. Program participants were invited to join the pilot study on a voluntary basis.

This program sought to provide evidence-based information and tools to empower patients on symptom management through a wide range of content: videos, exercises, written content, live sessions, quizzes, and a community-based platform. A total of 12 endometriosis experts were involved in the construction of the program, constituting a multidisciplinary team.

Scientific rationale highlighted some factors of interest for the program. In an Australian study on 484 women with endometriosis, dietary changes were reported among the most effective strategies in terms of reducing self-reported pain associated with the disease [8]. Optimizing nutrition is a nonpharmacological intervention that can be used to manage the symptoms of endometriosis. The positive effects of physical activity have been documented for various health conditions, including mental health (improved cognitive function, enhanced QOL, improved sleep, and reduced signs of anxiety and depression) and chronic pain [12,33,34]. According to current research, psychological interventions and mind-body approaches are promising for improving QOL and alleviating pain, anxiety, depression, stress, and fatigue in women with endometriosis [9,10]. The impact of endometriosis on sex life harms patients' self-esteem. Sex therapy and psychological interventions can be very beneficial during treatment to improve QOL [11].

The program addressed every aspect of the disease through 5 sections: disease education including pain mechanism, diet, adapted physical activity, well-being and mental health, and sexual health. It started with a section on knowledge of the disease, with an introduction to the program and the multidisciplinary approach. Psychoeducation was focused on defining the disorder, its causes, and its impacts. It helped the participant generate a model of their disorder and pain to obtain a representation of a functional analysis of the factors perpetuating the symptoms. The first part of the program concluded with educating the participants on setting goals and initiating change and provided tools to help them set their goals.

CBT content is included throughout the program. Each section mentioned subsequently contains education, advice, and tools created by health care professionals in the field.

The diet section covered step-by-step behavior change through good mealtime habits, from breakfast to dinner, with the help of calendars to fill in. This section provided information on why and how to make dietary changes with endometriosis, particularly on the role of diet in inflammation and digestive symptoms. A section dedicated to identifying inflammatory foods was also available to participants.

The sections dedicated to physical activity presented different ways to stay active while experiencing symptoms and suggested various methods, including yoga, Pilates, physiotherapy exercises, and daily habits, to move the body. A 30-minute yoga session and a Pilates session were available each week. These sessions focused mainly on pelvic mobility. Physiotherapy complemented the physical activity with ventral breathing exercises and a focus on the perineum and abdominal muscles.

The mental health section focused mainly on education and prevention to help with emotional regulation. Relaxation practices, sophrology exercises, Beck columns, and a weekly diary to initiate behavior changes were proposed.

The sexuality and intimacy section featured physiotherapy exercises, sex therapy tools, and advice to rediscover pleasure and self-confidence in sexual relations.

The program allowed participants to create their toolbox and action plan. Section contents were delivered weekly for 3 months and were available for a further 3 months (Figure 1).

XSL•FO
RenderX

Figure 1. Sections contained in the 3-month digital program offered to participants with endometriosis on which this cohort study is based.

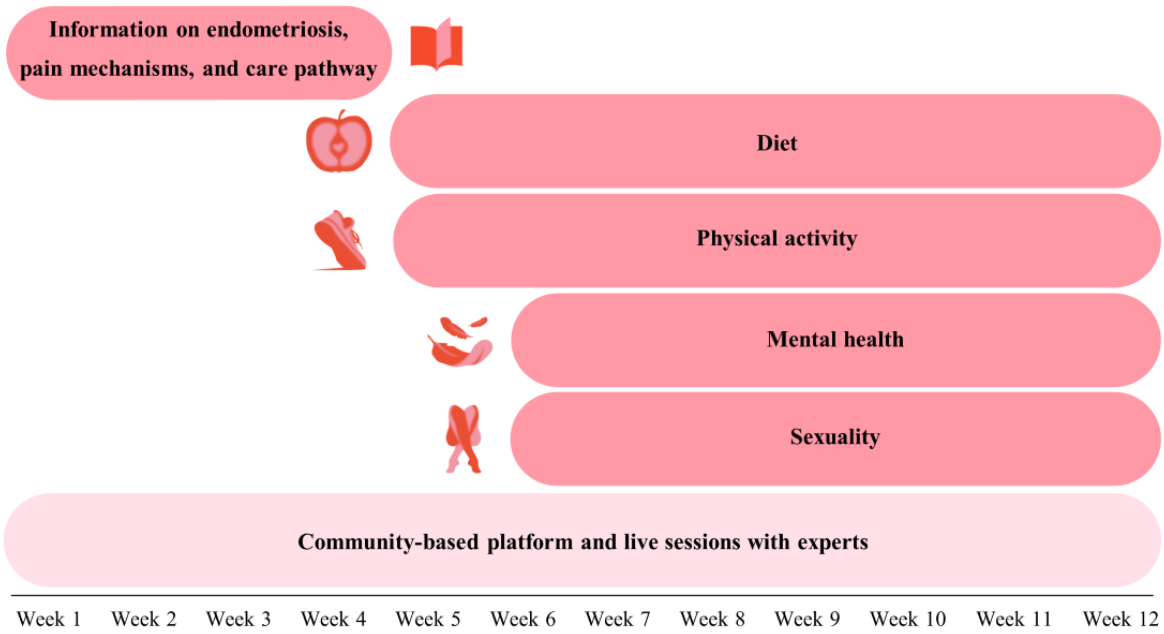

Ethical Considerations

A study protocol was established in compliance with French regulations, reviewed by the Lyv Healthcare company's ethics committee, and filed with the French public structure for studies on health data: Health Data Hub (N° F20221114165253). Participants answered the questionnaires on a voluntary basis after being informed about the research project objectives and were provided with contact information to exercise their rights. At any time, participants had the possibility to opt out. All collected data were pseudonymized and stored in compliance with health data security standards (Health Data Hosting certification, h*ébergeurs de données de santé*). No personal information was shared with third parties, and all data analysis was performed on deidentified datasets. No compensation was provided to either program participants or the control group for their participation in the study.

Impact of the Digital Health Program

Questionnaires

Our study compared 2 groups, distinct in terms of exposure to the digital program, both having participants with endometriosis. The study is considered a cohort study because the allocation of respondents to the exposure criterion was not controlled for the research participant, and there was no prospective assignment. Two web-based questionnaires were sent to program participants via email links and reminders on the program's community-based platform. The first questionnaire was sent at the time of inclusion in the program (time point 0 [T0]) and the second one after 3 months (T0+3 mo), with 2 weeks to submit their answers. Participants completed the questionnaires on their own, requiring approximately 15 minutes to complete. Questionnaires were structured into 2 parts, including a main part (53 questions) and a bonus part (24 questions), explaining differences in the sample size for some outcomes. Both questionnaires were electronically tested before diffusion. A back button allowed respondents to review and modify their answers, while IP addresses were used to ensure that each participant was a unique visitor, with only the first submission used for analysis.

Population Sample

Women were free to choose whether to take part in the program, with no obligation to complete it in full. Program participants were given the freedom to answer the questionnaires. This pilot study included participants who completed the main part entirely at baseline and 3 months.

Only women who reported being diagnosed with endometriosis could answer the questionnaires. The diagnosis could be clinical, imagery based, or surgical. To record the effect of active use of the program on participants, the threshold of following at least half of the proposed content types was retained. As a result, among the participants, only women who declared having tested >50% of the proposed content type in the program were included.

The terms woman and women are used in this paper. It should be noted that people with a uterus may or may not identify with these terms and experience endometriosis. Sex and gender were not criteria for inclusion in the study.

Control Group

A control group of women with endometriosis who did not follow the program was recruited via social media and a health insurance provider email campaign (*Mutuelle Générale de l'Education Nationale*). The women in the control group were volunteers, motivated solely by the desire to contribute to advancing research, as no incentives, prizes, or rewards were offered for their participation. The control group received the

XSL•FO
RenderX

same questionnaires as program participants during the same periods to eliminate external effects on measured parameters.

The control group was sampled on the initial pain level to have a similar profile between them and program participants. Women were selected from the control group by stratifying the sample to reproduce the distribution of pain levels among program participants. QOL answers were not selected for this stratification as these questions were part of bonus questionnaires. A random sample of women from the control group was selected so that the same proportion was represented for each modality of the participants' general pain level. The proportional sample was thus selected, reducing the number of women in the control group to 149.

Study Outcomes Measures

Numeric Rating Scales

Numeric Rating Scales (NRSs) are 11-point scales, ranging from 0 (no pain or symptom) to 10 (intense pain or symptom), primarily used to measure symptom intensity. NRS evaluated the level of overall pain, anxiety, depression, dysmenorrhea, dyspareunia, dyschezia, dysuria, chronic pelvic pain, gastrointestinal disorders, chronic fatigue, neuropathic pain, and endo belly. Endo belly is a common term used by patients with endometriosis to describe the cyclic bloating of the abdomen [35] as "uncomfortable or painful, often accompanied by a sensation of abdominal fullness," and which "often forces women with endometriosis to wear loose clothing" [36]. Endo belly affects >80% of women with gastrointestinal symptoms due to endometriosis [37].

Women without menstruation did not respond to the dysmenorrhea NRS, and women without sexual activities did not respond to the dyspareunia NRS, which explains differences in sample size for these outcomes.

An average of each of these NRSs was used to determine a global symptom burden score, ranging from 0 to 10.

EHP-5 Scale

EHP-5 is a QOL scale specific to endometriosis. A main score is calculated out of 100 (100=poor QOL and 0=good QOL), based on 5 items with 5 modalities (scored as 0, 25, 50, 75, and 100). The main score is the average of the scores for the 5 items. A modular score is calculated out of 100 (100=worst QOL and 0=best QOL), based on 6 optional 5-modality items (scored as 0, 25, 50, 75, and 100). The score is the average of the scores for the 6 items.

EQ-5D Scale

EQ-5D is a QOL scale nonspecific to endometriosis. It comprises 5 items with 5 modalities scored from 1 to 5. The concatenation of the 5 scores gives a combination. Each combination ranging from 11111 to 55555 has a numerical equivalent between –0.53 and 1. EQ-5D score ranges from –0.53 (poor QOL) to 1 (good QOL).

Statistical Analysis

Symptoms were considered to have evolved (improved or deteriorated) if scores had changed by >2 points compared with T0 [38]. QOL was considered to have evolved (improved or deteriorated) if the score had changed by >15% compared with T0 [39]. Otherwise, outcomes were considered stable.

The minimum sample size was calculated through a feasibility study power analysis for outcomes between program participants and the control group. This feasibility study was carried out on a preliminary program among 55 participants to identify ways to improve the program and observe symptoms and QOL levels. Global pain on the NRS was assumed by a mean (SD) of 5.9 (2.3). Assuming a difference of at least 2 points, a power (1–β) of 0.80, α=.05, and using a 2-sided 2-sample test, 11 program participants and 11 participants from the control group would be required to detect group differences. QOL was measured on the EHP-5, assumed by a mean (SD) of 53.6 (19.0). Assuming a change in score of at least 15%, a power of 0.80, α=.05, and using a 2-sided 2-sample test, 13 program participants and 13 participants from the control group would be required for detecting between-group differences. On the basis of power analyses, it was planned to include answers from at least 13 program participants and 13 from the control group.

Descriptive statistics were carried out (number of participants, percentage, mean, and SD). Statistical tests used to measure the significance of intragroup evolution were the chi-square test and the Fisher exact test for discrete variables, the Wilcoxon signed rank test for paired and nonparametric continuous variables, and the paired 2-tailed Student *t* test for parametric continuous variables. Statistical tests used to measure between-group differences (participants vs controls) were the chi-square test and the Fisher exact test for discrete variables, the Mann-Whitney *U* test for unpaired and nonparametric continuous variables, and the unpaired 2-tailed Student *t* test for parametric continuous variables. Furthermore, we applied the Benjamini-Hochberg method to adjust for multiple comparisons. This method controls the false discovery rate, offering a balance between minimizing false positives and maintaining statistical power. Cohen *d* tests were performed to measure effect size when the result was statistically significant. If Cohen $d>|0.2|$, the effect size was considered small, medium if Cohen $d>|0.5|$, and large if Cohen $d>|0.8|$. Each outcome was studied separately. A sensitivity analysis was conducted to note the impact of program follow-up according to baseline QOL level. For this, linear regression models and interaction tests were used with the EHP-5 core QOL level (good QOL: 0 to 32, medium QOL: 33 to 65, and low QOL: 66 to 100). Statistical analysis was performed with SAS software (version 9.4; SAS Institute).

Results

Population Description

The pilot study was based on the responses to the main questionnaire part from 92 program participants diagnosed with endometriosis and 149 women diagnosed with endometriosis who did not follow the program, constituting the control group (Figure 2).

XSL•FO
RenderX

Figure 2. Cohort study flow diagram: cohort study responders were divided according to their participation in the 3-month digital program for women with endometriosis. T0: time point 0.

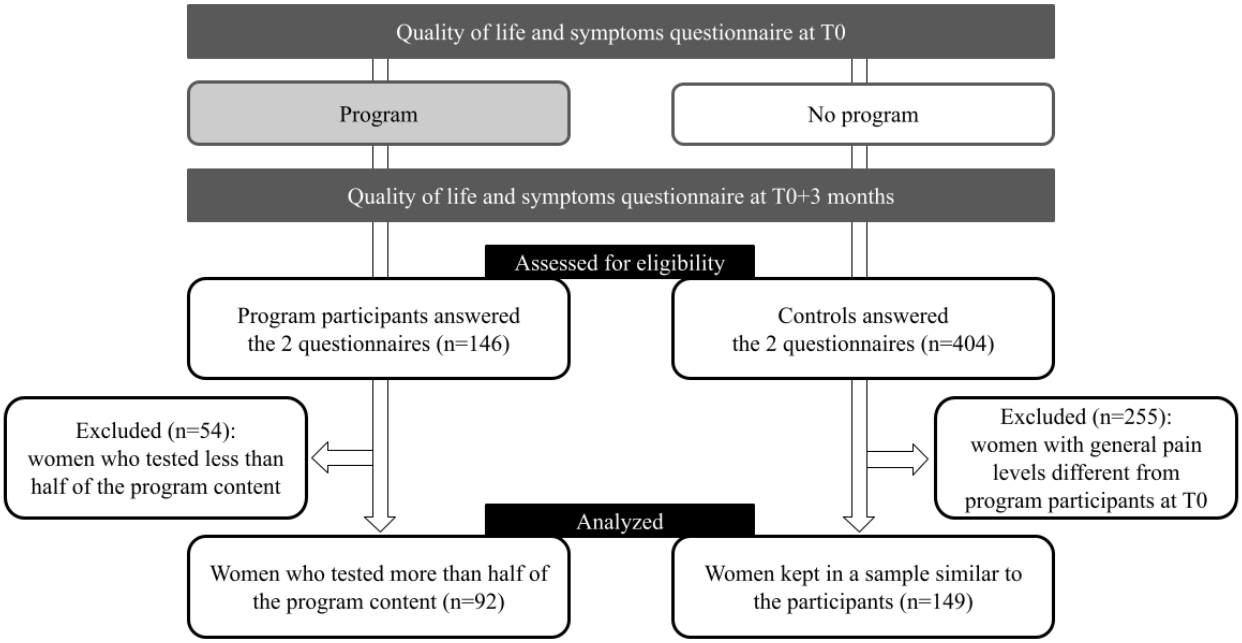

Respondents to the questionnaires took part in this study on a voluntary basis. Table 1 presents the characteristics of the program participants and those of the control group at baseline. On average, participants were aged 36.7 (SD 6.8) years and were diagnosed with the disease when they were aged 32.5 (SD 6.7) years. In total, 26% (24/92) of the participants were experiencing or had experienced infertility problems. There were no substantial differences in socioeconomic characteristics or pain levels at the time of inclusion compared to the control group. Participants had a slightly poorer QOL at the time of inclusion than those in the control group (P=.04) and were more often on a continuous hormonal pill (P=.046). The baseline symptom levels and QOL of program participants are provided in Multimedia Appendix 1.

Analyses were based on the answers given by participants to the initial questionnaire and the one at 3 months.

XSL•FO
RenderX

Table 1. Sociodemographic, clinical, pain-related, and quality of life baseline characteristics of program participants and those in the control group with endometriosis (N=241).

	Program participants (n=92)	Control group (n=149)	*P* value
Sociodemographic characteristics			
Age (y), mean (SD)	36.7 (6.8)	36.6 (7.2)	.91
Single status, n (%)	23 (25)	29 (19.5)	.64
Does not feel surrounded and feels very alone, n (%)	6 (6.5)	11 (7.4)	.92
Perceived financial situation (0=very insecure to 10=very comfortable), mean (SD)	5.4 (1.7)	5.4 (1.9)	.73
Number of children, mean (SD)	0.7 (1.0)	0.8 (1.0)	.34
Number of pregnancies, mean (SD)	1.1 (1.4)	1.3 (1.5)	.11
Diagnosis of endometriosis and treatment			
Experience of the pathway to diagnosis (0=very poor to 10=very good), mean (SD)	2.7 (2.4)	3.4 (2.7)	.09
Age at diagnosis (y), mean (SD)	32.5 (6.7)	31.6 (6.8)	.21
Age at first symptoms (y), mean (SD)	18.7 (7.3)	19.9 (8.7)	.56
Time between first set of symptoms and diagnosis, mean (SD)	13.5 (8.9)	11.5 (8.6)	.10
Hormonal treatment, n (%)	53 (57.6)	65 (43.6)	.046
Symptoms and quality of life			
Experienced infertility problems, n (%)	24 (26.1)	47 (31.5)	.37
Dysmenorrhea level (0=no pain to 10=worst pain), mean (SD)	7.3 (1.8)	6.5 (2.3)	.12
Global symptom burden (0=no symptom to 10=worst symptom intensity), mean (SD)	5.3 (1.6)	5.0 (1.6)	.17
Proportion of global symptom burden ≥8, n (%)	5 (5.4)	4 (2.7)	.31
Quality of life: EHP-5[a] (0=best to 100=worst), mean (SD)	53.8 (20.4)[b]	45.2 (20.3)[c]	.04
Proportion of quality of life: EHP-5 ≥80, n (%)	11 (12)	9 (6)	.15

[a]EHP-5: Endometriosis Health Profile-5.

[b]n=39 program participants.

[c]n=65 individuals from the control group.

Impact on Endometriosis Symptoms

Initially, the participants' most severe symptom was chronic fatigue (mean 7.5, SD 2.1), followed by dysmenorrhea (mean 7.3, SD 1.8) and anxiety (mean 6.5, SD 2.2; Table 2).

Overall, there was a tendency for all symptoms to improve during the 3 months of following the program (Table 3). For some symptoms, the evolution was not significant compared with the control group: dyspareunia, dyschezia, chronic pelvic pain, gastrointestinal disorders, overall pain, dysuria, and chronic fatigue. Over 3 months, the global symptom burden, general level of pain, anxiety, depression, dysmenorrhea, dysuria, chronic fatigue, neuropathic pain, and endo belly improved significantly for the program participants. These improvements were significantly different compared to the control group for global symptom burden (P=.048; small effect size), anxiety (P<.001; medium effect size), depression (P=.04; small effect size), neuropathic pain (P=.004; small effect size), and endo belly (P=.03; small effect size).

XSL•FO
RenderX

Table 2. Outcome evaluation at baseline and 3 months for program participants versus those in the control group with endometriosis.

Health outcome	Program participants (n=92), n (%)	Control group (n=149), n (%)	T0[a]				T0+3 mo			
			Program participants, mean (SD)	Control group, mean (SD)	Participants vs control group		Program participants, mean (SD)	Control group, mean (SD)	Participants vs control group	
					P value	Cohen *d*			*P* value	Cohen *d*
Global symptom burden	90 (97.8)	147 (98.7)	5.3 (1.6)	5.0 (1.6)	.17	0.1[b]	4.6 (1.6)	4.7 (1.8)	.75	–0.0[b]
Overall pain	90 (97.8)	147 (98.7)	6.2 (2.5)	6.0 (2.4)	.68	0.1[b]	5.5 (2.6)	5.3 (2.6)	.66	0.1[b]
Anxiety	90 (97.8)	147 (98.7)	6.5 (2.2)	5.9 (2.7)	.12	0.2[c]	5.5 (2.9)	6.1 (2.7)	.12	–0.2[c]
Depression	89 (96.7)	147 (98.7)	4.0 (3.0)	4.2 (3.2)	.87	–0.0[b]	3.3 (3.1)	4.2 (3.3)	.05	–0.3[c]
Dysmenorrhea	35 (38)	74 (49.7)	7.3 (1.8)	6.5 (2.3)	.12	0.3[c]	5.6 (2.6)	5.8 (2.6)	.54	–0.1[b]
Dyspareunia	43 (46.7)	84 (56.4)	3.9 (2.4)	4.1 (2.8)	.90	–0.0[b]	3.1 (2.5)	3.5 (2.7)	.47	–0.2[c]
Dyschezia	90 (97.8)	147 (98.7)	3.7 (3.2)	4.2 (2.8)	.19	–0.2[c]	3.3 (2.7)	3.6 (2.9)	.60	–0.1[b]
Dysuria	90 (97.8)	147 (98.7)	2.2 (2.9)	1.8 (2.5)	.36	0.2[c]	1.6 (2.0)	1.6 (2.4)	.35	0.0[b]
Chronic pelvic pain	90 (97.8)	147 (98.7)	5.2 (2.8)	4.8 (2.7)	.22	0.2[b]	4.8 (2.8)	4.5 (2.7)	.48	0.1[b]
Gastrointestinal disorders	90 (97.8)	147 (98.7)	5.9 (2.7)	5.3 (2.7)	.10	0.2[c]	5.7 (2.3)	5.1 (2.8)	.13	0.2[c]
Chronic fatigue	90 (97.8)	147 (98.7)	7.5 (2.1)	7.0 (2.2)	.09	0.2[c]	6.8 (2.3)	6.8 (2.3)	.77	–0.0[b]
Neuropathic pain	90 (97.8)	147 (98.7)	5.4 (3.1)	4.6 (3.2)	.09	0.2[c]	4.3 (2.9)	4.5 (3.0)	.47	–0.1[b]
Endo belly	90 (97.8)	147 (98.7)	6.2 (2.5)	5.7 (2.7)	.09	0.2[c]	5.4 (2.7)	5.3 (2.7)	.92	0.0[b]
QOL[d]: EHP-5[e] core part	39 (42.4)	65 (43.6)	53.8 (20.4)	45.2 (20.3)	.04	0.0[b]	47.9 (19.4)	46.2 (21.5)	.88	0.0[b]
QOL: EHP-5, modular part	39 (42.4)	64 (43)	34.9 (13.0)	28.7 (18.4)	.05	0.4[c]	26.7 (15.4)	26.4 (18.8)	.59	0.0[b]
QOL: EQ-5D score	33 (35.9)	80 (53.7)	0.8 (0.1)	0.8 (0.2)	.06	–0.2[c]	0.9 (0.1)	0.8 (0.2)	.44	0.0[b]

[a]T0: time point 0.

[b]No effect size.

[c]Small effect size.

[d]QOL: quality of life.

[e]EHP-5: Endometriosis Health Profile-5.

XSL•FO
RenderX

Table 3. Continuous outcomes evolution between baseline and 3 months for program participants versus the control group with endometriosis.

Health outcome	Program participants, mean (SD)	Within program participants		Control group, mean (SD)	Within control group		Program participants vs control group		Benjamini-Hochberg threshold
		P value	Cohen *d*[a]		*P* value	Cohen *d*	*P* value	Cohen *d*	
Global symptom burden	–0.7 (1.6)	*<.001*[b]	*0.4*[c]	–0.3 (1.3)	*.004*	*0.2*[c]	*.048*	*–0.3*[c]	0.022
Overall pain	–0.7 (2.9)	*.02*	*0.3*[c]	–0.7 (2.5)	*<.001*	*0.3*[c]	.84	—[d]	0.050
Anxiety	–1.1 (2.8)	*<.001*	*0.4*[c]	0.2 (2.5)	.51	—	*<.001*	*–0.5*[e]	*0.003*
Depression	–0.9 (2.5)	*.002*	*0.3*[c]	0.0 (3.1)	.92	—	*.04*	*–0.3*[c]	0.019
Dysmenorrhea	–1.9 (2.8)	*<.001*	*0.7*[e]	–0.7 (2.4)	*.01*	*0.3*[c]	.05	—	0.025
Dyspareunia	–0.5 (2.8)	.28	—	–0.7 (2.2)	*.007*	*0.3*[c]	.69	—	0.044
Dyschezia	–0.4 (2.3)	.16	—	–0.6 (2.7)	*.004*	*0.2*[c]	.47	—	0.038
Dysuria	–0.5 (2.3)	*.045*	*0.2*[c]	–0.1 (2.3)	.41	—	.22	—	0.034
Chronic pelvic pain	–0.5 (3.0)	.15	—	–0.3 (2.6)	.29	—	.53	—	0.041
Gastrointestinal disorders	–0.2 (2.7)	.36	—	–0.2 (2.3)	.31	—	.76	—	0.047
Chronic fatigue	–0.7 (2.5)	*.02*	*0.3*[c]	–0.2 (2.3)	.29	—	.14	—	0.031
Neuropathic pain	–1.0 (2.7)	*<.001*	*0.4*[c]	–0.1 (2.6)	.58	—	*.004*	*–0.4*[c]	*0.009*
Endo belly	–0.9 (2.5)	*.002*	*0.3*[c]	–0.3 (2.4)	.21	—	*.03*	*–0.2*[c]	0.013
QOL[f]: EHP-5[g] core part	–5.9 (21.0)	.09	—	1. (14.8)	.59	—	*.03*	*–0.4*[c]	0.016
QOL: EHP-5, modular part	–8.2 (16.9)	*.004*	*–0.5*[e]	–2.5 (11.1)	.06	—	.10	—	0.028
QOL: EQ-5D score	0.1 (0.1)	*.01*	*0.5*[e]	–0.0 (0.1)	.08	—	*.001*	*0.7*[e]	*0.006*

[a]The effect size was considered small if Cohen *d* >|0.2|, medium if Cohen *d* >|0.5|, and large if Cohen *d* >|0.8|.

[b]Values in italics indicate significant *P* values.

[c]Small effect size.

[d]Not applicable.

[e]Medium effect size.

[f]QOL: quality of life.

[g]EHP-5: Endometriosis Health Profile-5.

When looking at symptom evolution (improvement, stability, or deterioration), the distribution was not significantly different between the 2 groups for dyspareunia, dyschezia, dysuria, chronic pelvic pain, gastrointestinal disorders, and chronic fatigue (Table 4). Global symptom burden improved in 20% (18/90) of the program participants versus 6.1% (9/147) in the control group (*P*=.003). Anxiety levels improved in 42% (38/90) and deteriorated in 11% (10/90) of the program participants versus 26.5% and 28.6%, respectively, in the control group (*P*=.002). Depression deteriorated less in program participants (9/89, 10%) than in the control group (41/147, 27.9%; *P*=.003). Neuropathic pain improved more in program participants (37/90, 41%) than in the control group (34/147, 23.1%; *P*=.02), as did endo belly (37/90, 41% vs 36/147, 24.5%; *P*=.03).

Actively following the digital program for 3 months was associated with a significant improvement in global symptom burden, anxiety, depression, neuropathic pain, and endo belly perception among program participants when compared to the control group.

XSL•FO
RenderX

Table 4. Evolution of symptoms between baseline and 3 months for program participants versus the control group with endometriosis.

Health outcome	Program participants			Control group			Program participants vs control group, *P* value	Benjamini-Hochberg threshold
	Improvement	Stable	Deterioration	Improvement	Stable	Deterioration		
Global symptom burden (n=90 participants; n=147 controls)							*.003* [a]	*0.009*
Participants, n (%)	18 (20)	68 (75.6)	4 (4.4)	9 (6.1)	133 (90.5)	5 (3.4)		
Points, mean (SD)	–3.1 (0.9)	–0.3 (1.0)	2.6 (0.7)	–2.9 (0.9)	–0.2 (1.0)	2.5 (0.5)		
Overall pain (n=90 participants; n=147 controls)							*.09*	0.029
Participants, n (%)	34 (37.8)	39 (43.3)	17 (18.9)	43 (29.3)	85 (57.8)	19 (12.9)		
Points, mean (SD)	–3.8 (1.7)	0.1 (0.8)	3.3 (1.3)	–3.7 (1.6)	–0.1 (0.7)	3.2 (1.5)		
Anxiety (n=90 participants; n=147 controls)							*.002*	*0.006*
Participants, n (%)	38 (42.2)	42 (46.7)	10 (11.1)	39 (26.5)	66 (44.9)	42 (28.6)		
Points, mean (SD)	–3.6 (1.8)	–0.1 (0.7)	4.2 (1.9)	–2.9 (1.2)	0.2 (0.7)	3.0 (1.4)		
Depression (n=89 participants; n=147 controls)							*.003*	*0.012*
Participants, n (%)	28 (31.5)	52 (58.4)	9 (10.1)	42 (28.6)	64 (43.5)	41 (27.9)		
Points, mean (SD)	–3.7 (1.7)	0.0 (0.8)	3.0 (1.5)	–3.5 (1.6)	0.0 (0.7)	3.7 (2.1)		
Dysmenorrhea (n=35 participants; n=74 controls)							.07	0.024
Participants, n (%)	19 (54.3)	14 (40)	2 (5.7)	24 (32.4)	39 (52.7)	11 (14.9)		
Points, mean (SD)	–4.0 (2.1)	0.1 (0.7)	3.0 (0.0)	–3.3 (1.6)	–0.2 (0.8)	2.9 (1.1)		
Dyspareunia (n=43 participants; n=84 controls)							.49	0.038
Participants, n (%)	13 (30.2)	21 (48.8)	9 (20.9)	25 (29.8)	48 (57.1)	11 (13.1)		
Points, mean (SD)	–3.9 (1.6)	0.0 (0.8)	3.1 (1.5)	–3.4 (1.6)	0.0 (0.7)	2.3 (0.6)		
Dyschezia (n=90 participants; n=147 controls)							.63	0.047
Participants, n (%)	23 (25.6)	51 (56.7)	16 (17.8)	46 (31.3)	75 (51)	26 (17.7)		
Points, mean (SD)	–3.3 (1.5)	–0.1 (0.7)	3.1 (1.1)	–3.5 (1.7)	–0.1 (0.8)	3.3 (2.1)		
Dysuria (n=90 participants; n=147 controls)							.38	0.035
Participants, n (%)	24 (26.7)	55 (61.1)	11 (12.2)	28 (19.1)	97 (66)	22 (15)		
Points, mean (SD)	–3.3 (1.7)	–0.1 (0.5)	3.5 (1.4)	–3.4 (1.5)	0.0 (0.5)	3.6 (1.6)		
Chronic pelvic pain (n=90 participants; n=147 controls)							.30	0.032
Participants, n (%)	33 (36.7)	33 (36.7)	24 (26.7)	44 (29.9)	69 (46.9)	34 (23.1)		
Points, mean (SD)	–3.7 (1.6)	0.0 (0.7)	3.3 (1.2)	–3.2 (1.6)	0.0 (0.8)	3.1 (1.2)		
Gastrointestinal disorders (n=90 participants; n=147 controls)							.90	0.050
Participants, n (%)	27 (30)	42 (46.7)	21 (23.3)	42 (28.6)	73 (49.7)	32 (21.8)		
Points, mean (SD)	–3.2 (1.3)	–0.1 (0.8)	3.4 (1.6)	–2.8 (1.1)	0.0 (0.8)	3.0 (1.3)		
Chronic fatigue (n=90 participants; n=147 controls)							.49	0.041
Participants, n (%)	29 (32.2)	46 (51.1)	15 (16.7)	37 (25.2)	81 (55.1)	29 (19.7)		
Points, mean (SD)	–3.4 (1.6)	–0.1 (0.8)	2.9 (1.4)	–3.1 (1.4)	0.0 (0.7)	2.9 (1.5)		
Neuropathic pain (n=90 participants; n=147 controls)							*.02*	0.015
Participants, n (%)	37 (41.1)	39 (43.3)	14 (15.6)	34 (23.1)	84 (57.1)	29 (19.7)		
Points, mean (SD)	–3.5 (1.6)	–0.2 (0.7)	3.1 (1.7)	–3.4 (1.6)	0.0 (0.7)	3.6 (1.8)		
Endo belly (n=90 participants; n=147 controls)							*.03*	0.021
Participants, n (%)	37 (41.1)	37 (41.1)	16 (17.8)	36 (24.5)	81 (55.1)	30 (20.4)		
Points, mean (SD)	–3.4 (1.4)	–0.1 (0.9)	3.3 (1.8)	–3.5 (1.8)	0.0 (0.7)	2.7 (1.1)		

XSL•FO
RenderX

[a]Italicization indicates significant *P* values.

Impact on QOL

With all 3 scores used (EHP-5 core part, EHP-5 modular part, and EQ-5D), active program participants showed an improvement in their QOL after 3 months on the program (Table 2). This evolution observed between 0 and 3 months significantly differed from the control group for the core part of the EHP-5 (*P*=.03; small effect size) and the EQ-5D (*P*=.001; medium effect size; Table 3). Therefore, the use of the program was associated with an improvement of QOL in participants with endometriosis.

When looking at the types of evolution (improvement, stability, or deterioration) in QOL, an improvement of the EQ-5D and the EHP-5 core part was observed in 2 and 3 times more women in the program participants than in the control group, respectively (Table 5). However, the difference in the types of evolution was not significant between the 2 groups for EQ-5D (*P*=.58) and the core part of EHP-5 (*P*=.07). A significant difference in the types of evolution between the 2 groups was observed for the modular part of the EHP-5 (*P*=.02), with an improvement in 31% (12/39) of the program participants versus 11% (7/64) of the controls.

Table 5. Evolution of quality of life (QOL) score items and perceived knowledge between baseline and 3 months for program participants versus the control group with endometriosis.

Health outcome	Program participants			Control group			Program participants vs control group, *P* value	Benjamini-Hochberg threshold
	Improvement	Stable	Deterioration	Improvement	Stable	Deterioration		
QOL: EHP-5[a] core part (n=39 participants; n=65 controls)							.07	0.026
Participants, n (%)	16 (41)	17 (43.6)	6 (15.4)	13 (20)	40 (61.5)	12 (18.5)		
Points, mean (SD)	–25.0 (8.2)	–0.9 (6.9)	30.8 (11.6)	–18.1 (3.8)	0.5 (6.6)	23.3 (11.3)		
QOL: EHP-5 modular part (n=39 participants; n=64 controls)							*.02* [b]	0.018
Participants, n (%)	12 (30.8)	24 (61.5)	3 (7.7)	7 (10.9)	54 (84.4)	3 (4.7)		
Points, mean (SD)	–28.1 (9.3)	–2.4 (6.1)	25.0 (8.3)	–23.2 (3.3)	–1.3 (6.9)	23.6 (2.4)		
QOL: EQ-5D score (n=33 participants; n=80 controls)							.58	0.044
Participants, n (%)	2 (6.1)	31 (93.9)	0 (0)	2 (2.5)	78 (97.5)	0 (0)		
Points, mean (SD)	0.4 (0.1)	0.0 (0.1)	__[c]	0.4 (0.1)	–0.0 (0.1)	—		
Perceived knowledge (n=90 participants; n=148 controls), n (%)	47 (52.2)	41 (45.6)	2 (2.2)	21 (14.2)	119 (80.4)	8 (5.4)	*<.001*	*0.003*

[a]EHP-5: Endometriosis Health Profile-5.

[b]Values in italics indicate significant *P* values.

[c]Not applicable.

To better understand these evolutions in QOL, we focused on the items that make up the QOL scores (Table 6). There was a significant difference at baseline (*P*=.03; small effect size) in the core part of the EHP-5 score on the item concerning control and powerlessness. Program participants were more affected than those in the control group. This significant difference disappeared at 3 months. Besides, after 3 months, there was a significant improvement in the sexual relationship item (*P*=.02; small effect size) in the modular part of the EHP-5 score, which was not observed in the control group. Participants were less often worried about pain during sexual intercourse after 3 months on the program than before starting the program. Finally, the usual activity item in the EQ-5D score improved nearly significantly for the program participants (*P*=.05; small effect size) after 3 months, which was not the case in the control group, with participants feeling less limited in their usual activities after 3 months on the program.

The benefits of a digital program also depend on how it is used; in this study, it was found that the greater the frequency of use or the greater the diversity of sections consulted, the greater the improvement in symptoms or QOL (Multimedia Appendices 2 and 3). The results did not show any statistically significant correlations between the viewing of action-oriented content and the program's effectiveness (Multimedia Appendix 3).

XSL•FO
RenderX

Table 6. Quality of life (QOL) score items evaluation at baseline and 3 months for program participants versus the control group with endometriosis.

Health outcome	Program participants				Control group				Program participants vs control group			
	T0[a], mean (SD)	T0 +3 mo, mean (SD)	*P* value	Cohen *d*	T0, mean (SD)	T0 +3 mo, mean (SD)	*P* value	Cohen *d*	T0, *P* value	T0, Cohen *d*	T0+3 mo, *P* value	T0+3 mo, Cohen *d*
QOL: EHP-5[b] core part												
Pain	1.4 (1.1)	1.2 (1.0)	.17	—[c]	1.2 (1.0)	1.2 (1.0)	.95	—	.25	—	.99	—
Control and powerlessness	2.5 (1.4)	2.1 (1.2)	.06	—	1.9 (1.2)	1.9 (1.2)	.93	—	.03	0.4[d]	.62	—
Emotional well-being	2.5 (1.0)	2.3 (0.9)	.39	—	2.3 (1.0)	2.3 (1.1)	.59	—	.79	—	.90	—
Social support	2.3 (1.0)	2.1 (0.9)	.19	—	1.9 (1.2)	2.0 (1.1)	.64	—	.15	—	.87	—
Self-image	2.2 (1.3)	2.0 (1.4)	.46	—	1.8 (1.2)	1.9 (1.4)	.48	—	.13	—	.80	—
QOL: EHP-5 modular part												
Work life	1.2 (1.1)	0.9 (1.1)	.37	—	0.9 (1.2)	0.8 (1.3)	.46	—	.13	—	.21	—
Taking care of children	2.2 (1.2)	2.1 (1.1)	.19	—	1.4 (1.1)	1.3 (1.2)	.99	—	.03	0.7[e]	.04	0.7[e]
Sexual relationships	2.2 (1.5)	1.7 (1.4)	.02	–0.4[d]	1.7 (1.4)	1.7 (1.4)	.16	—	.15	—	.99	—
Medical profession	1.3 (1.2)	0.6 (0.8)	.06	—	1.2 (1.3)	1.1 (1.4)	.66	—	.81	—	.21	—
Treatment	2.4 (1.0)	2.2 (1.2)	.25	—	2.1 (1.3)	1.9 (1.4)	.24	—	.32	—	.36	—
Infertility	2.2 (1.6)	2.5 (1.4)	.66	—	1.9 (1.7)	2.1 (1.7)	.68	—	.45	—	.52	—
QOL: EQ-5D score												
Mobility	1.4 (0.7)	1.2 (0.5)	.11	—	1.5 (0.8)	1.5 (0.8)	.62	—	.87	—	.06	—
Self-care	1.1 (0.2)	1.0 (0.2)	.99	—	1.1 (0.4)	1.1 (0.4)	.99	—	.61	—	.17	—
Usual activities	2.2 (0.9)	1.9 (0.9)	.05	–0.4[d]	1.9 (0.9)	1.9 (0.8)	.63	—	.05	—	.92	—
Pain or discomfort	2.9 (0.8)	2.5 (0.7)	.06	—	2.5 (1.0)	2.5 (1.0)	.64	—	.06	—	.47	—
Anxiety or depression	2.6 (0.9)	2.4 (1.0)	.19	—	2.4 (0.9)	2.6 (1.0)	.04	0.2[d]	.28	—	.52	—
VAS[f] (0=worst to 100=best QOL)	58.8 (17.1)	63.0 (15.7)	.23	—	63.4 (20.7)	62.8 (19.2)	.79	—	.13	—	.99	—

[a]T0: time point 0.

[b]EHP-5: Endometriosis Health Profile-5.

[c]Not applicable.

[d]Small size effect.

[e]Medium size effect.

[f]VAS: visual analog scale.

Impact on Perceived Knowledge of Endometriosis

Initially, there was no significant difference in perceived knowledge between program participants and the control group (*P*=.74; Table 7). In total, 22% (20/92) of the program participants felt that they knew little about the disease, 71% (65/92) had good knowledge, and 8% (7/92) considered themselves experts (vs 34/149, 22.8%; 99/149, 66.4%; and 16/149, 10.7%, respectively for the control group). At 3 months, none of the program participants considered themselves as knowing little about the disease (0/92, 0%); 64% (59/92) considered themselves as having good knowledge, and 36% (33/92) considered themselves experts (vs 25/149, 16.8%; 104/149, 69.8%, and 20/149, 13.4%, respectively, for the control group). The perceived knowledge of endometriosis was significantly different at 3 months between the 2 groups (*P*<.001; Table 5). As expected, the program seemed to improve perceived knowledge of endometriosis.

XSL•FO
RenderX

Table 7. Perceived knowledge on endometriosis at baseline and 3 months for program participants versus the control group. The significance for program participants versus control group is P=.74 at T0 and P<.001 at T0+3 months.

Perceived knowledge on endometriosis	Program participants (n=92), n (%)		Control group (n=149), n (%)	
	T0[a]	T0+3 mo	T0	T0+3 mo
I know little	20 (21.7)	0 (0)	34 (22.8)	25 (16.8)
I have good knowledge, but I am not an expert	65 (70.7)	59 (64.1)	99 (66.4)	104 (69.8)
I consider myself an expert	7 (7.6)	33 (35.9)	16 (10.7)	20 (13.4)

[a]T0: time point 0.

Multiple Comparisons and Sensitivity Analysis

To address the potential for type I errors due to multiple testing in this pilot analysis, we applied the Benjamini-Hochberg method to adjust significance thresholds while maintaining statistical power. This approach controls the false discovery rate and is particularly suited for exploratory analyses.

Table 1 illustrates that the QOL levels, as assessed by the EHP-5 core part, significantly differed between the 2 groups at baseline (P=.04). However, the proportion of participants reporting poor QOL was similar in both groups (P=.15). To explore the potential influence of these baseline differences on the evolution of outcomes, we conducted interaction analyses using linear regression models, as detailed in Multimedia Appendix 4. These analyses revealed that baseline QOL differences primarily affected the evolution of dyschezia. Specifically, dyschezia showed more significant improvement in program participants compared to the control group, particularly among women with good baseline QOL. In addition, baseline QOL levels influenced changes in the EHP core part score at 3 months. This score improved significantly among program participants with a distinct pattern compared to the control group, especially for women with good QOL at baseline. Conversely, improvements were more pronounced in the control group than in the program participants for women with moderate or poor QOL at baseline.

Discussion

Principal Findings

This pilot study of 92 program participants and 149 individuals in the control group shows that following a 3-month digital health program for the self-management of endometriosis symptoms is associated with a significant reduction in endometriosis-related symptoms (anxiety, depression, neuropathic pain, and endo belly perception), a reduction of the global symptom burden, and a significant improvement in the participants' knowledge of endometriosis. Furthermore, following the program seemed to be associated with an improvement in QOL. Combining CBT approaches and EHP-focused programs in a digital tool has proved useful in inducing symptom relief and a better QOL.

Comparison With Prior Work

Previous studies have shown that patients often use nonpharmacological interventions to cope with their symptoms [8,40]. However, only one short-term study has examined the value of combining these nonpharmacological interventions as part of a multidisciplinary approach for QOL with endometriosis [41], and a recent publication tested the application over 3 months [42]. Instead, most published studies analyzed the benefits of such interventions individually. These studies, among other things, have underpinned the development of our proposed program, which offers patients various tools and solutions to manage their symptoms.

As confirmed by our results, CBT has been found to be helpful in managing pain and improving the scores of depression and QOL in patients with endometriosis [43-45]. New studies are also beginning to be conducted on the benefits of psychological and relaxation interventions for patients with endometriosis, thus far demonstrating improvement in pain [46], stress [47], anxiety, depression, and QOL [48,49]. The improvement of mental health–related symptoms in our results might be in part due to the psychological and relaxation approaches included in our program. Given the link between mental health and QOL in women with endometriosis [50], future research should explore the use of more comprehensive scales for anxiety and depression. Furthermore, nutrition and physical activity have been studied separately and proved to have an impact on endometriosis symptoms, stress, anxiety, and QOL [51-56]. Finally, although sexual issues are a prevalent symptom, sex education for women with endometriosis has only been assessed in one study, where an improvement in sexual function and quality of sexual life was observed [57]. Interestingly, in our study, the EHP-5 score for the sexual intercourse apprehension item decreased for participants, notwithstanding persisting dyspareunia. This highlights a change in the perception of sexual anxiety after 3 months, despite the persistence of pain. Digital programs for the management of endometriosis should study the long-term effect of sex education content on dyspareunia, both in terms of symptoms and perception.

While the efficacy and relevance of each intervention for the management of endometriosis symptoms have thus been confirmed by previous research, a multidisciplinary approach is in line with current guidelines; therefore, it should be encouraged for women with endometriosis and more generally for those who experience chronic pain [58,59]. Future studies should seek to determine the optimal combination of nonpharmacological interventions according to patient profiles to find the right balance between diet, physical activity, stress management, sleep, and environmental enrichment on top of their medical care [60,61].

In terms of methodology, our study assessed each symptom using numerical scales to simplify user responses and limit attrition rates. For QOL, we used 2 different scales: the EQ-5D, which is nonspecific to endometriosis, and the EHP-5, which

XSL•FO
RenderX

is specific to endometriosis. These were previously compared in a French study [39], which concluded that while both scales were appropriate for assessing the QOL of women with endometriosis, the EHP-5 was better at assessing health-related QOL, especially regarding medical treatment and the intensity of dysmenorrhea. According to our results, the EHP-5 does seem to better discriminate with regard to endometriosis, which echoes the conclusions of the French study [39]. However, regardless of the scale used, following our digital program was associated with an improved QOL. This was true for 2 to 3 times more participants than individuals in the control group. Our study used a 15% threshold to define a change in QOL, allowing stringent evolutions to be highlighted. Program participants had a poorer QOL level at baseline compared to the control group (despite a similar pain profile). However, after 3 months of following the program, program participants caught up with the QOL level of the control group.

Digital tools are a way to bring multidisciplinary interventions to patients, enhancing their accessibility. Studies on digital interventions for various conditions have shown positive results, such as improvements in anxiety and depression through CBT-based approaches [62], symptom management and QOL enhancements via mindfulness-based stress-reduction programs [63], or better health outcomes through digital personalized diets [64]. Similarly, multidisciplinary digital programs have demonstrated significant benefits across diverse chronic conditions, including improved QOL and symptom management. For instance, internet-based programs have been shown to enhance knowledge, increase exercise habits, and reduce heart failure symptoms [65], while personalized digital care programs have improved the QOL in adults with autoimmune diseases and post–COVID-19 condition [66]. Digital support programs have also improved self-reported QOL in patients with rheumatoid arthritis [67] and managing symptoms in patients with cancer [68]. In our study, we observed that a digital tool using a CBT approach is effective in helping patients manage a condition, in this case, endometriosis. This may have to do with the fact that digital technology can bring about a different perception and thus reduce the stigma surrounding intimate issues associated with endometriosis, such as psychological or sexuality aspects [28,69-71]. These findings underscore the potential of multidisciplinary digital programs to relieve symptoms, improve QOL, and enhance disease-related knowledge across various chronic conditions.

For digital therapeutics to achieve their purpose, they must meet patient needs in terms of content while optimizing patient adherence and fulfilling prescriber requirements. The content of our program was based on EHP components to address subjects that were relevant to patients with endometriosis and was approved by health care specialists. Other studies have looked at digital technology for managing endometriosis or pelvic pain and highlighted the need for educational content, psychological and social support, and patient empowerment in particular [40]. As for patient adherence, to enhance digital therapeutic efficiency, it is crucial that future studies analyze specific factors influencing patient engagement in a digital program, such as trust, interactions, and consideration [72]. Keeping patients engaged to increase their frequency of use and guiding them through nonpharmacological, multidisciplinary interventions are therefore key to a program's efficiency. Guidelines and studies encourage the multidisciplinary management of endometriosis. However, only 2 studies on the same intervention, evaluating the QOL after 2 and 12 weeks following a multimodal program, have been conducted in the context of digital intervention [41,42]. Our study contributes to answering the need for new research on the multidisciplinary management of endometriosis and the development of programs to make such management accessible [73,74].

Limitations and Strengths

Our study is subject to several limitations. It is based on voluntary participation only, with no compulsory questionnaires and no obligation to complete the program. It also includes a selection bias, with no information as to whether the profiles of the women who answered the questionnaires were similar to the overall profiles of the women who took part in the program. Although the participants in this study was not randomized, differences between program participants and the control group were minimized by selecting a control group sample with pain levels similar to those of program participants (despite a different initial QOL level, the effect of which was analyzed in sensitivity analyses). Using the Benjamini-Hochberg method, we were able to control the false discovery rate. Nevertheless, we interpreted results with borderline P values cautiously, acknowledging the increased risk of type I errors inherent in multiple testing.

Furthermore, participants self-reported their diagnosis of endometriosis, which may be clinical, imaging based, or surgical. However, one can assume that women who do not have endometriosis would not attend the program and would not spend time filling in research questionnaires.

A strength of the program is that it was designed based on user research, scientific rationale, and medical expertise. Our approach, targeting EHP components and using CBT, allowed the development of an effective digital tool for the self-management of endometriosis symptoms. In addition, the use of 2 different questionnaires on QOL showed that following the program was associated with an improvement of the health-related QOL (EQ-5D) and the QOL specific to endometriosis (EHP-5). Very little amount of data exists on the reliability and efficiency of a digital program for the management of endometriosis symptoms. Our study is the first step in identifying key factors to be considered for developing a digital health program for the daily management of endometriosis and in demonstrating its positive impact on patient symptoms and QOL.

Conclusions

Our results suggest that a digital health program providing medical and scientific information about endometriosis as well as multidisciplinary self-management tools may be a helpful and effective resource for women to manage life with endometriosis alongside their medical care. Hence, a digital program for endometriosis that combines integrative solutions,

XSL•FO
RenderX

focuses on EHP components, and uses a CBT approach can enhance patient care for those with endometriosis.

Acknowledgments

The authors would like to thank Laura Matteo for her help in proofreading the manuscript, and they would like to thank the entire team involved in developing this program. The authors would also like to graciously thank this pilot study's program participants for their involvement and time. This study was funded by the French National Association for Research and Technology and by Lyv Healthcare.

Data Availability

The datasets generated and analyzed during this study are not publicly available due to the consent agreements made with participants, which did not include provisions for public data release but are available from the corresponding author on reasonable request.

Authors' Contributions

ZB and MA were responsible for conceptualizing the study. ZB was involved in the development of the program, study design, data collection and extraction, statistical analysis, data curation, and formal analysis. ES was involved in the development of the program and study design. MA was involved in the development of the program, study design, data collection, statistical analysis, and editing of the manuscript and supervised the entire research process. PP, EP, and DL were involved in the development of the program. The original draft of the manuscript was written collaboratively by ZB, MA, and ES. In addition, ZB, ES, MP, DL, EP, PP, and MA contributed to the writing, reviewing, and editing of the manuscript. All authors reviewed and approved the final version of the manuscript.

Conflicts of Interest

The PhD of ZB was partly funded by the French National Association for Research and Technology and Lyv Healthcare. MA, MP, and ZB are employed at the Lyv Healthcare. ES, PP, DL, and EP are on the scientific committee of Lyv Healthcare and are paid speakers in the School of Endo program.

Multimedia Appendix 1
Distribution of symptom levels and quality of life for endometriosis program participants at baseline.
[PNG File , 243 KB - formative_v9i1e58262_app1.png]

Multimedia Appendix 2
The evolution of outcomes between baseline and 3 months for all endometriosis program participants versus the control group according to the content thresholds of the program tested.
[DOCX File , 19 KB - formative_v9i1e58262_app2.docx]

Multimedia Appendix 3
The evolution of outcomes between baseline and 3 months for all endometriosis program participants according to the sections thresholds of the program tested.
[DOCX File , 15 KB - formative_v9i1e58262_app3.docx]

Multimedia Appendix 4
Beta coefficients and 95% CIs of nonadjusted linear regression models evaluating associations between endometriosis program participation and health outcomes and interaction tests according to baseline quality of life levels (Endometriosis Health Profile-5 core part).
[DOCX File , 30 KB - formative_v9i1e58262_app4.docx]

References

1. Zondervan KT, Becker CM, Missmer SA. Endometriosis. N Engl J Med 2020 Mar 26;382(13):1244-1256. [doi: 10.1056/NEJMra1810764] [Medline: 32212520]
2. Della Corte L, Di Filippo C, Gabrielli O, Reppuccia S, La Rosa VL, Ragusa R, et al. The burden of endometriosis on women's lifespan: a narrative overview on quality of life and psychosocial wellbeing. Int J Environ Res Public Health 2020 Jun 29;17(13):4683 [FREE Full text] [doi: 10.3390/ijerph17134683] [Medline: 32610665]

XSL•FO
RenderX

3. O'Hara R, Rowe H, Fisher J, O'Hara R, Rowe H, Fisher J. Managing endometriosis: a cross-sectional survey of women in Australia. J Psychosom Obstet Gynaecol 2022 Sep;43(3):265-272 [FREE Full text] [doi: 10.1080/0167482X.2020.1825374] [Medline: 33050751]
4. Ceccaroni M, Bounous VE, Clarizia R, Mautone D, Mabrouk M. Recurrent endometriosis: a battle against an unknown enemy. Eur J Contracept Reprod Health Care 2019 Dec 25;24(6):464-474. [doi: 10.1080/13625187.2019.1662391] [Medline: 31550940]
5. Becker CM, Bokor A, Heikinheimo O, Horne A, Jansen F, Kiesel L, ESHRE Endometriosis Guideline Group. ESHRE guideline: endometriosis. Hum Reprod Open 2022;2022(2):hoac009 [FREE Full text] [doi: 10.1093/hropen/hoac009] [Medline: 35350465]
6. Nnoaham KE, Hummelshoj L, Webster P, d'Hooghe T, de Cicco Nardone F, de Cicco Nardone C, World Endometriosis Research Foundation Global Study of Women's Health consortium. Impact of endometriosis on quality of life and work productivity: a multicenter study across ten countries. Fertil Steril 2011 Aug;96(2):366-73.e8 [FREE Full text] [doi: 10.1016/j.fertnstert.2011.05.090] [Medline: 21718982]
7. Giudice LC, Horne AW, Missmer SA. Time for global health policy and research leaders to prioritize endometriosis. Nat Commun 2023 Dec 04;14(1):8028 [FREE Full text] [doi: 10.1038/s41467-023-43913-9] [Medline: 38049392]
8. Armour M, Sinclair J, Chalmers KJ, Smith CA. Self-management strategies amongst Australian women with endometriosis: a national online survey. BMC Complement Altern Med 2019 Jan 15;19(1):17 [FREE Full text] [doi: 10.1186/s12906-019-2431-x] [Medline: 30646891]
9. Evans S, Fernandez S, Olive L, Payne LA, Mikocka-Walus A. Psychological and mind-body interventions for endometriosis: a systematic review. J Psychosom Res 2019 Sep;124:109756. [doi: 10.1016/j.jpsychores.2019.109756] [Medline: 31443810]
10. Hansen KE, Brandsborg B, Kesmodel US, Forman A, Kold M, Pristed R, et al. Psychological interventions improve quality of life despite persistent pain in endometriosis: results of a 3-armed randomized controlled trial. Qual Life Res 2023 Jun 17;32(6):1727-1744 [FREE Full text] [doi: 10.1007/s11136-023-03346-9] [Medline: 36797461]
11. Facchin F, Buggio L, Dridi D, Barbara G, Vercellini P. The subjective experience of dyspareunia in women with endometriosis: a systematic review with narrative synthesis of qualitative research. Int J Environ Res Public Health 2021 Nov 18;18(22):12112 [FREE Full text] [doi: 10.3390/ijerph182212112] [Medline: 34831868]
12. Mińko A, Turoń-Skrzypińska A, Rył A, Bargiel P, Hilicka Z, Michalczyk K, et al. Endometriosis-a multifaceted problem of a modern woman. Int J Environ Res Public Health 2021 Aug 02;18(15):8177 [FREE Full text] [doi: 10.3390/ijerph18158177] [Medline: 34360470]
13. Jones GL, Budds K, Taylor F, Musson D, Raymer J, Churchman D, et al. A systematic review to determine use of the Endometriosis Health Profiles to measure quality of life outcomes in women with endometriosis. Hum Reprod Update 2024 Mar 01;30(2):186-214 [FREE Full text] [doi: 10.1093/humupd/dmad029] [Medline: 38007607]
14. Jones G, Jenkinson C, Kennedy S. Evaluating the responsiveness of the endometriosis health profile questionnaire: the EHP-30. Qual Life Res 2004 Apr;13(3):705-713. [doi: 10.1023/b:qure.0000021316.79349.af]
15. Jones G, Jenkinson C, Kennedy S. Development of the short form endometriosis health profile questionnaire: the EHP-5. Qual Life Res 2004 Apr;13(3):695-704. [doi: 10.1023/b:qure.0000021321.48041.0e]
16. Hofmann SG, Asnaani A, Vonk IJ, Sawyer AT, Fang A. The efficacy of cognitive behavioral therapy: a review of meta-analyses. Cognit Ther Res 2012 Oct 01;36(5):427-440 [FREE Full text] [doi: 10.1007/s10608-012-9476-1] [Medline: 23459093]
17. Butler AC, Chapman JE, Forman EM, Beck AT. The empirical status of cognitive-behavioral therapy: a review of meta-analyses. Clin Psychol Rev 2006 Jan;26(1):17-31. [doi: 10.1016/j.cpr.2005.07.003] [Medline: 16199119]
18. Schubert K, Lohse J, Kalder M, Ziller V, Weise C. Internet-based cognitive behavioral therapy for improving health-related quality of life in patients with endometriosis: study protocol for a randomized controlled trial. Trials 2022 Apr 12;23(1):300 [FREE Full text] [doi: 10.1186/s13063-022-06204-0] [Medline: 35414092]
19. DiGiuseppe R, Venezia R, Gotterbarn R. What is cognitive behavior therapy? In: Vernon A, Doyle KA, editors. Cognitive Behavior Therapies: A Guidebook for Practitioners. New York, NY: American Counseling Association; 2018:1-35.
20. Ehde DM, Dillworth TM, Turner JA. Cognitive-behavioral therapy for individuals with chronic pain: efficacy, innovations, and directions for research. Am Psychol 2014;69(2):153-166. [doi: 10.1037/a0035747] [Medline: 24547801]
21. Mikocka-Walus A, Druitt M, O'Shea M, Skvarc D, Watts JJ, Esterman A, et al. Yoga, cognitive-behavioural therapy versus education to improve quality of life and reduce healthcare costs in people with endometriosis: a randomised controlled trial. BMJ Open 2021 Aug 09;11(8):e046603 [FREE Full text] [doi: 10.1136/bmjopen-2020-046603] [Medline: 34373298]
22. Morley S, Eccleston C, Williams A. Systematic review and meta-analysis of randomized controlled trials of cognitive behaviour therapy and behaviour therapy for chronic pain in adults, excluding headache. Pain 1999 Mar;80(1-2):1-13. [doi: 10.1016/s0304-3959(98)00255-3] [Medline: 10204712]
23. de C Williams AC, Eccleston C, Morley S. Psychological therapies for the management of chronic pain (excluding headache) in adults. Cochrane Database Syst Rev 2012 Nov 14;11(11):CD007407 [FREE Full text] [doi: 10.1002/14651858.CD007407.pub3] [Medline: 23152245]

XSL•FO
RenderX

24. Burns JW, Nielson WR, Jensen MP, Heapy A, Czlapinski R, Kerns RD. Specific and general therapeutic mechanisms in cognitive behavioral treatment of chronic pain. J Consult Clin Psychol 2015 Feb;83(1):1-11. [doi: 10.1037/a0037208] [Medline: 24979313]
25. Andersson G, Carlbring P. Internet-assisted cognitive behavioral therapy. Psychiatr Clin North Am 2017 Dec;40(4):689-700. [doi: 10.1016/j.psc.2017.08.004] [Medline: 29080594]
26. Carlbring P, Andersson G, Cuijpers P, Riper H, Hedman-Lagerlöf E. Internet-based vs. face-to-face cognitive behavior therapy for psychiatric and somatic disorders: an updated systematic review and meta-analysis. Cogn Behav Ther 2018 Jan;47(1):1-18 [FREE Full text] [doi: 10.1080/16506073.2017.1401115] [Medline: 29215315]
27. Luik AI, van der Zweerde T, van Straten A, Lancee J. Digital delivery of cognitive behavioral therapy for insomnia. Curr Psychiatry Rep 2019 Jun 04;21(7):50 [FREE Full text] [doi: 10.1007/s11920-019-1041-0] [Medline: 31161406]
28. Stern E, Micoulaud Franchi JA, Dumas G, Moreira J, Mouchabac S, Maruani J, et al. How can digital mental health enhance psychiatry? Neuroscientist 2023 Dec 04;29(6):681-693. [doi: 10.1177/10738584221098603] [Medline: 35658666]
29. Gold SM, Friede T, Meyer B, Moss-Morris R, Hudson J, Asseyer S, et al. Internet-delivered cognitive behavioural therapy programme to reduce depressive symptoms in patients with multiple sclerosis: a multicentre, randomised, controlled, phase 3 trial. Lancet Digit Health 2023 Oct;5(10):e668-e678 [FREE Full text] [doi: 10.1016/S2589-7500(23)00109-7] [Medline: 37775187]
30. Murphy JL, Cordova MJ, Dedert EA. Cognitive behavioral therapy for chronic pain in veterans: evidence for clinical effectiveness in a model program. Psychol Serv 2022 Feb;19(1):95-102 [FREE Full text] [doi: 10.1037/ser0000506] [Medline: 32986454]
31. Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M, Medical Research Council Guidance. Developing and evaluating complex interventions: the new Medical Research Council guidance. BMJ 2008 Sep 29;337(sep29 1):a1655 [FREE Full text] [doi: 10.1136/bmj.a1655] [Medline: 18824488]
32. Koninckx PR, Ussia A, Gordts S, Keckstein J, Saridogan E, Malzoni M, et al. The 10 "cardinal sins" in the clinical diagnosis and treatment of endometriosis: a Bayesian approach. J Clin Med 2023 Jul 07;12(13):4547 [FREE Full text] [doi: 10.3390/jcm12134547] [Medline: 37445589]
33. Ambrose KR, Golightly YM. Physical exercise as non-pharmacological treatment of chronic pain: why and when. Best Pract Res Clin Rheumatol 2015 Feb;29(1):120-130 [FREE Full text] [doi: 10.1016/j.berh.2015.04.022] [Medline: 26267006]
34. Borisovskaya A, Chmelik E, Karnik A. Exercise and chronic pain. Adv Exp Med Biol 2020;1228:233-253. [doi: 10.1007/978-981-15-1792-1_16] [Medline: 32342462]
35. Velho RV, Werner F, Mechsner S. Endo belly: what is it and why does it happen?: a narrative review. J Clin Med 2023 Nov 19;12(22):7176 [FREE Full text] [doi: 10.3390/jcm12227176] [Medline: 38002788]
36. Luscombe GM, Markham R, Judio M, Grigoriu A, Fraser IS. Abdominal bloating: an under-recognized endometriosis symptom. J Obstet Gynaecol Can 2009 Dec;31(12):1159-1171. [doi: 10.1016/s1701-2163(16)34377-8] [Medline: 20085682]
37. Maroun P, Cooper MJ, Reid GD, Keirse MJ. Relevance of gastrointestinal symptoms in endometriosis. Aust N Z J Obstet Gynaecol 2009 Aug;49(4):411-414. [doi: 10.1111/j.1479-828X.2009.01030.x] [Medline: 19694698]
38. Pokrzywinski RM, Soliman AM, Snabes MC, Chen J, Taylor HS, Coyne KS. Responsiveness and thresholds for clinically meaningful changes in worst pain numerical rating scale for dysmenorrhea and nonmenstrual pelvic pain in women with moderate to severe endometriosis. Fertil Steril 2021 Feb;115(2):423-430 [FREE Full text] [doi: 10.1016/j.fertnstert.2020.07.013] [Medline: 33066973]
39. Aubry G, Panel P, Thiollier G, Huchon C, Fauconnier A. Measuring health-related quality of life in women with endometriosis: comparing the clinimetric properties of the Endometriosis Health Profile-5 (EHP-5) and the EuroQol-5D (EQ-5D). Hum Reprod 2017 Jun 01;32(6):1258-1269. [doi: 10.1093/humrep/dex057] [Medline: 28383700]
40. Gagnon MM, Brilz AR, Alberts NM, Gordon JL, Risling TL, Stinson JN. Understanding adolescents' experiences with menstrual pain to inform the user-centered design of a mindfulness-based app: mixed methods investigation study. JMIR Pediatr Parent 2024 Apr 08;7:e54658 [FREE Full text] [doi: 10.2196/54658] [Medline: 38587886]
41. Rohloff N, Rothenhöfer M, Götz T, Schäfer SD. Observational pilot study on the influence of an app-based self-management program on the quality of life of women with endometriosis. Arch Gynecol Obstet 2024 Aug 13;310(2):1157-1170. [doi: 10.1007/s00404-024-07468-4] [Medline: 38869629]
42. Rohloff N, Götz T, Kortekamp SS, Heinze NR, Weber C, Schäfer SD. Influence of app-based self-management on the quality of life of women with endometriosis. Cureus 2024 Aug;16(8):e67655. [doi: 10.7759/cureus.67655] [Medline: 39314601]
43. Donatti L, Malvezzi H, de Azevedo BC, Baracat EC, Podgaec S. Cognitive behavioral therapy in endometriosis, psychological based intervention: a systematic review. Rev Bras Ginecol Obstet 2022 Mar 16;44(3):295-303 [FREE Full text] [doi: 10.1055/s-0042-1742406] [Medline: 35576938]
44. Donatti L, Podgaec S, Baracat EC. Efficacy of cognitive behavioral therapy in treating women with endometriosis and chronic pelvic pain: a randomized trial. J Health Psychol 2024 Apr 02:13591053241240198. [doi: 10.1177/13591053241240198] [Medline: 38566501]

XSL•FO
RenderX

45. Wu S, Wang X, Liu H, Zheng W. Efficacy of cognitive behavioral therapy after the surgical treatment of women with endometriosis: a preliminary case-control study. Medicine (Baltimore) 2022 Dec 23;101(51):e32433 [FREE Full text] [doi: 10.1097/MD.0000000000032433] [Medline: 36595829]
46. Jafari H, Courtois I, Van den Bergh O, Vlaeyen JW, Van Diest I. Pain and respiration: a systematic review. Pain 2017 Jun;158(6):995-1006. [doi: 10.1097/j.pain.0000000000000865] [Medline: 28240995]
47. Petit E, Lhuillery D, Loriau J, Sauvanet E. Endométriose: Diagnostic et Prise en Charge. Paris: Elsevier-Masson, Pratique en Gynécologie-Obstétrique; 2020.
48. Zhao L, Wu H, Zhou X, Wang Q, Zhu W, Chen J. Effects of progressive muscular relaxation training on anxiety, depression and quality of life of endometriosis patients under gonadotrophin-releasing hormone agonist therapy. Eur J Obstet Gynecol Reprod Biol 2012 Jun;162(2):211-215. [doi: 10.1016/j.ejogrb.2012.02.029] [Medline: 22455972]
49. Samami E, Shahhosseini Z, Khani S, Elyasi F. Pain-focused psychological interventions in women with endometriosis: a systematic review. Neuropsychopharmacol Rep 2023 Sep 27;43(3):310-319 [FREE Full text] [doi: 10.1002/npr2.12348] [Medline: 37366616]
50. Pontoppidan K, Olovsson M, Grundström H. Clinical factors associated with quality of life among women with endometriosis: a cross-sectional study. BMC Womens Health 2023 Oct 24;23(1):551 [FREE Full text] [doi: 10.1186/s12905-023-02694-5] [Medline: 37875883]
51. Cirillo M, Argento FR, Becatti M, Fiorillo C, Coccia ME, Fatini C. Mediterranean diet and oxidative stress: a relationship with pain perception in endometriosis. Int J Mol Sci 2023 Sep 27;24(19):14601 [FREE Full text] [doi: 10.3390/ijms241914601] [Medline: 37834048]
52. Moore JS, Gibson PR, Perry RE, Burgell RE. Endometriosis in patients with irritable bowel syndrome: specific symptomatic and demographic profile, and response to the low FODMAP diet. Aust N Z J Obstet Gynaecol 2017 Apr 17;57(2):201-205. [doi: 10.1111/ajo.12594] [Medline: 28303579]
53. Borghini R, Porpora MG, Casale R, Marino M, Palmieri E, Greco N, et al. Irritable bowel syndrome-like disorders in endometriosis: prevalence of nickel sensitivity and effects of a low-nickel diet. An open-label pilot study. Nutrients 2020 Jan 28;12(2):341 [FREE Full text] [doi: 10.3390/nu12020341] [Medline: 32012984]
54. Oddsson SJ, Gunnarsdottir T, Johannsdottir LG, Amundadottir ML, Frimannsdottir A, Molander P, et al. A new digital health program for patients with inflammatory bowel disease: preliminary program evaluation. JMIR Form Res 2023 Apr 28;7:e39331 [FREE Full text] [doi: 10.2196/39331] [Medline: 37115598]
55. Gonçalves AV, Barros NF, Bahamondes L. The practice of hatha yoga for the treatment of pain associated with endometriosis. J Altern Complement Med 2017 Jan;23(1):45-52. [doi: 10.1089/acm.2015.0343] [Medline: 27869485]
56. Posadzki P, Ernst E, Terry R, Lee MS. Is yoga effective for pain? A systematic review of randomized clinical trials. Complement Ther Med 2011 Oct;19(5):281-287. [doi: 10.1016/j.ctim.2011.07.004] [Medline: 21944658]
57. Matloobi M, Amini L, Shahali S, Haghani H, Tahermanesh K, Hassanlouei B, et al. Effect of sex education on sexual function and sexual quality of life in women with endometriosis: a quasi-experimental study. Int J Gynaecol Obstet 2022 Dec 25;159(3):702-710. [doi: 10.1002/ijgo.14254] [Medline: 35514234]
58. Armour M, Avery J, Leonardi M, Van Niekerk L, Druitt ML, Parker MA, et al. Lessons from implementing the Australian National Action Plan for Endometriosis. Reprod Fertil 2022 Jul 01;3(3):C29-C39 [FREE Full text] [doi: 10.1530/RAF-22-0003] [Medline: 35928674]
59. Endometriosis: diagnosis and management. National Institute for Health and Care Excellence. URL: https://www.nice.org.uk/guidance/ng73 [accessed 2024-04-29]
60. Wattier JM. Antalgiques et alternatives thérapeutiques non médicamenteuses pluridisciplinaires. Gynecol Obstet Fertil Senol 2018 Mar;46(3):248-255. [doi: 10.1016/j.gofs.2018.02.002] [Medline: 29510963]
61. De Hoyos G, Ramos-Sostre D, Torres-Reverón A, Barros-Cartagena B, López-Rodríguez V, Nieves-Vázquez C, et al. Efficacy of an environmental enrichment intervention for endometriosis: a pilot study. Front Psychol 2023 Oct 10;14:1225790 [FREE Full text] [doi: 10.3389/fpsyg.2023.1225790] [Medline: 37885745]
62. Verkleij M, Georgiopoulos AM, Barendrecht H, Friedman D. Pilot of a therapist-guided digital mental health intervention (eHealth CF-CBT) for adults with cystic fibrosis. Pediatr Pulmonol 2023 Jul 05;58(7):2094-2103. [doi: 10.1002/ppul.26438] [Medline: 37144856]
63. Miazga E, Starkman H, Schroeder N, Nensi A, McCaffrey C. Virtual mindfulness-based therapy for the management of endometriosis chronic pelvic pain: a novel delivery platform to increase access to care. J Obstet Gynaecol Can 2024 Jun;46(6):102457 [FREE Full text] [doi: 10.1016/j.jogc.2024.102457] [Medline: 38614242]
64. Jactel SN, Olson JM, Wolin KY, Brown J, Pathipati MP, Jagiella VJ, et al. Efficacy of a digital personalized elimination diet for the self-management of irritable bowel syndrome and comorbid irritable bowel syndrome and inflammatory bowel disease. Clin Transl Gastroenterol 2023 Jan 01;14(1):e00545 [FREE Full text] [doi: 10.14309/ctg.0000000000000545] [Medline: 36322404]
65. Tomita MR, Tsai BM, Fisher NM, Kumar NA, Wilding G, Stanton K, et al. Effects of multidisciplinary internet-based program on management of heart failure. J Multidiscip Healthc 2008 Dec 01;2009(2):13-21 [FREE Full text] [doi: 10.2147/jmdh.s4355] [Medline: 20505786]

XSL•FO
RenderX

66. Bundy N, De Jesus M, Lytle M, Calabrese L, Gobin C, Dyhrberg M. Self-evidence-based digital care programme improves health-related quality of life in adults with a variety of autoimmune diseases and long COVID: a retrospective study. RMD Open 2023 May 16;9(2):e003061 [FREE Full text] [doi: 10.1136/rmdopen-2023-003061] [Medline: 37192812]
67. Dobies B, White A, Isberg A, Gudmundsson S, Oddsson S. Digital health program improves quality of life in rheumatoid arthritis: a retrospective analysis of real-world data. Clin Exp Rheumatol 2024 Jan;42(1):10-14 [FREE Full text] [doi: 10.55563/clinexprheumatol/rng5n9] [Medline: 38306018]
68. Gudmundsson GH, Mészáros J, Björnsdóttir Á, Ámundadóttir ML, Thorvardardottir GE, Magnusdottir E, et al. Evaluating the feasibility of a digital therapeutic program for patients with cancer during active treatment: pre-post interventional study. JMIR Form Res 2022 Oct 13;6(10):e39764 [FREE Full text] [doi: 10.2196/39764] [Medline: 36227639]
69. Hollis C, Morriss R, Martin J, Amani S, Cotton R, Denis M, et al. Technological innovations in mental healthcare: harnessing the digital revolution. Br J Psychiatry 2015 Apr;206(4):263-265 [FREE Full text] [doi: 10.1192/bjp.bp.113.142612] [Medline: 25833865]
70. Sims OT, Gupta J, Missmer SA, Aninye IO. Stigma and endometriosis: a brief overview and recommendations to improve psychosocial well-being and diagnostic delay. Int J Environ Res Public Health 2021 Aug 03;18(15):8210 [FREE Full text] [doi: 10.3390/ijerph18158210] [Medline: 34360501]
71. Abdulai AF, Howard AF, Yong PJ, Noga H, Parmar G, Currie LM. Developing an educational website for women with endometriosis-associated dyspareunia: usability and stigma analysis. JMIR Hum Factors 2022 Mar 03;9(1):e31317 [FREE Full text] [doi: 10.2196/31317] [Medline: 35238789]
72. Stern E, Breton Z, Alexaline M, Geoffroy PA, Bungener C. Redefining the relationship in digital care: a qualitative study of the Digital Therapeutic Alliance. Encephale (Forthcoming) 2024 May 08 [FREE Full text] [doi: 10.1016/j.encep.2024.02.011] [Medline: 38724431]
73. Agarwal SK, Antunez-Flores O, Foster WG, Hermes A, Golshan S, Soliman AM, et al. Real-world characteristics of women with endometriosis-related pain entering a multidisciplinary endometriosis program. BMC Womens Health 2021 Jan 07;21(1):19 [FREE Full text] [doi: 10.1186/s12905-020-01139-7] [Medline: 33413295]
74. Agarwal SK, Foster WG, Groessl EJ. Rethinking endometriosis care: applying the chronic care model via a multidisciplinary program for the care of women with endometriosis. Int J Womens Health 2019;11:405-410 [FREE Full text] [doi: 10.2147/IJWH.S207373] [Medline: 31413643]

Abbreviations

CBT: cognitive behavioral therapy
EHP: Endometriosis Health Profile
NRS: Numeric Rating Scale
QOL: quality of life
T0: time point 0

Edited by A Mavragani; submitted 11.03.24; peer-reviewed by M Armour, A AL-Asadi, G Joseph; comments to author 12.05.24; revised version received 24.06.24; accepted 08.01.25; published 28.02.25.

Please cite as:
Breton Z, Stern E, Pinault M, Lhuillery D, Petit E, Panel P, Alexaline M
A Digital Program for Daily Life Management With Endometriosis: Pilot Cohort Study on Symptoms and Quality of Life Among Participants
JMIR Form Res 2025;9:e58262
URL: https://formative.jmir.org/2025/1/e58262
doi:10.2196/58262
PMID:39791286

©Zélia Breton, Emilie Stern, Mathilde Pinault, Delphine Lhuillery, Erick Petit, Pierre Panel, Maïa Alexaline. Originally published in JMIR Formative Research (https://formative.jmir.org), 28.02.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Daily Treatment Monitoring for Patients Receiving Home-Based Peritoneal Dialysis and Prediction of Heart Failure Risk: mHealth Tool Development and Modeling Study

Jia Wu[1*], MM; Youjia Zeng[1*], PhD; Jun Yang[1*], MM; Yutong Yao[1], BMed; Xiuling Xu[1], BMed; Gaofeng Song[1], PhD; Wuyong Yi[1], BMed; Taifen Wang[1], BMed; Yihou Zheng[1], BMed; Zhongwei Jia[2*], PhD; Xiangyu Yan[3,4*], PhD

[1]Shenzhen Hospital of Traditional Chinese Medicine, Shenzen, China

[2]Department of Global Health, School of Public Health, Peking University, Beijing, China

[3]School of Disaster and Emergency Medicine, Tianjin University, Tianjin, China

[4]Key Laboratory of Medical Rescue Key Technology and Equipment, Ministry of Emergency Management, Tianjin, China

[*]these authors contributed equally

Corresponding Author:
Xiangyu Yan, PhD
School of Disaster and Emergency Medicine
Tianjin University
No 92 Weijin Road
Tianjin, 300072
China
Phone: 86 02287370177 ext 307
Email: yanxiangyu1123@163.com

Abstract

Background: Peritoneal dialysis is one of the major renal replacement modalities for patients with end-stage renal disease. Heart failure is a common adverse event among patients who undergo peritoneal dialysis treatment, especially for those who undergo continuous ambulatory peritoneal dialysis at home, because of the lack of professional input-output volume monitoring and management during treatment.

Objective: This study aims to develop novel mobile health (mHealth) tools to improve the quality of home-based continuous ambulatory peritoneal dialysis treatment and to build a prediction model of heart failure based on the system's daily treatment monitoring data.

Methods: The mHealth tools with a 4-layer system were designed and developed using Spring Boot, MyBatis Plus, MySQL, and Redis as backend technology stack, and Vue, Element User Interface, and WeChat Mini Program as front-end technology stack. Patients were recruited to use the tool during daily peritoneal dialysis treatment from January 1, 2017, to April 20, 2023. Logistic regression models based on real-time treatment monitoring data were used for heart failure prediction. The sensitivity, specificity, accuracy, and Youden index were calculated to evaluate the performance of the prediction model. In the sensitivity analysis, the ratio of patients with and without heart failure was set to 1:4 and 1:10, respectively, to better evaluate the stability of the prediction model.

Results: A WeChat Mini Program named Futou Bao for patients and a patient data management platform for doctors was developed. Futou Bao included an intelligent data upload function module and an auxiliary function module. The doctor's data management platform consisted of 4 function modules, that is, patient management, data visualization and marking, data statistics, and system management. During the study period, the records of 6635 patients who received peritoneal dialysis treatment were uploaded in Futou Bao, with 0.71% (47/6635) of them experiencing heart failure. The prediction model that included sex, age, and diastolic blood pressure was considered as the optimal model, wherein the sensitivity, specificity, accuracy, and Youden index were 0.75, 0.91, 0.89, and 0.66, respectively, with an area under the curve value of 0.879 (95% CI 0.772-0.986) using the validation dataset. The sensitivity analysis showed stable results.

Conclusions: This study provides a new home-based peritoneal dialysis management paradigm that enables the daily monitoring and early warning of heart failure risk. This novel paradigm is of great value for improving the efficiency, security, and personalization of peritoneal dialysis.

XSL•FO
RenderX

(JMIR Form Res 2025;9:e56254) doi:10.2196/56254

KEYWORDS

peritoneal dialysis; mHealth; patient management; heart failure; prediction model

Introduction

Peritoneal dialysis is one of the 3 major renal replacement modalities (ie, hemodialysis, peritoneal dialysis, and kidney transplantation) for end-stage renal disease (ESRD) [1-3]. Continuous ambulatory peritoneal dialysis (CAPD) gradually becomes the mainstream mode for peritoneal dialysis [4]. Patients can receive the treatment at home because the treatment site environment needs less treatment resources and is flexible [5,6]. Compared with hemodialysis, peritoneal dialysis as a home-based, cost-saving renal replacement therapy has several advantages, including protection of residual kidney function, stable hemodynamics, and less interference with normal family, work, and social life. Peritoneal dialysis therapy has been recognized and accepted by patients with ESRD.

One common and serious adverse event in patients undergoing CAPD is capacity-overloaded heart failure. Previous studies have shown a high prevalence of congestive heart failure among patients who receive peritoneal dialysis treatment. A multicenter cohort study in mainland China conducted in 2012 reported that heart failure occurred in 47.3% of the patients who underwent peritoneal dialysis [7]. Contemporaneously, Wang et al's [8] study conducted in Hong Kong showed a heart failure prevalence of 39% in patients treated with peritoneal dialysis. Although the prevalence reported by subsequent studies showed a decrease to 29% in Wang et al's study [9] and to 19.6% in Sun et al's study [10], the prevalence remained at a high level. Therefore, the prevention and early detection of heart failure in patients undergoing peritoneal dialysis treatment is a vital issue that cannot be ignored.

Although operating peritoneal dialysis treatment at home has many advantages, there are still several treatment bottlenecks in the traditional home-based peritoneal dialysis approach. First, because patients manage the treatment at home, doctors cannot provide timely and accurate follow-up. Second, the effectiveness of peritoneal dialysis, the occurrence of adverse events represented by heart failure, and the quality of life of patients are closely related to the training and management level of the medical institutions and the self-management levels of the patients. Third, it is difficult to implement good remote whole-process quality management of patients and prevent adverse events in a timely and effective manner for patients receiving CAPD at home.

Given the treatment bottleneck in the traditional home-based peritoneal dialysis method, it is essential to develop a user-friendly remote intelligent treatment management platform. The internet, big data, and other mobile health (mHealth) technologies can enable seamless integration of mobile equipment, systems, and home-based CAPD treatment operations to conduct timely detection and early intervention of adverse events represented by heart failure through analyzing the automatically collected data by the system.

The aim of this study was to develop mHealth tools and systems to improve the efficiency, accuracy, and personalization of the home-based peritoneal dialysis treatment and to build a prediction model of heart failure in patients with peritoneal dialysis based on the systems' daily treatment monitoring data.

Methods

mHealth Tool Design and Development Procedures

The mHealth tools developed in this study were intended to promote the intelligence of the home peritoneal dialysis treatment in the following 4 aspects.

1. Automated data collection: Through the Bluetooth connection device, the system enables the automatic collection of peritoneal dialysis data, including peritoneal dialysis solution weight and patient's weight data, which can reduce manual operation and the possibility of data input error.
2. Daily data analysis: The system timely processes and analyzes the uploaded data and provides intuitive data visualization to help doctors better understand the patient's treatment status.
3. Remote treatment monitoring: In order to improve the efficiency and convenience of treatment, patients do not need to go to the hospital frequently for treatment, and doctors can remotely monitor and manage the treatment process of patients.
4. Personalized treatment plan: Through the real-time data analysis and treatment monitoring, doctors can develop personalized treatment plans according to the specific conditions of patients and improve the treatment effect.

In order to improve the mHealth system's usage convenience, this study plans to use Spring Boot, MyBatis Plus, MySQL, and Redis as backend technology stack, and Vue, Element User Interface, and WeChat Mini Program as the front-end technology stack. The whole system is divided into 4 layers: data storage layer, core service layer, system application layer, and user interface layer. This architecture could make the system efficient, reliable, and extensible, and meet the functional and performance requirements of peritoneal dialysis treatment management. The details of each layer are as follows (Figure 1):

1. Data storage layer: This layer is responsible for data storage and management. MySQL database is used to store various data, including user information and medical records. Redis is used as a nonrelational database to store login token and treatment time point information as a cache to improve the ability to query data.
2. Core service layer: This layer is the core operation processing layer of the system, responsible for realizing the core functions and business logic of the system. Spring Boot framework was adopted as the backend development framework, combined with MyBatis Plus as the persistence

XSL•FO
RenderX

framework. Each function module of the system was developed through the annotation method of Spring Boot and used the Object Relational Mapping provided by MyBatis Plus to interact with the MySQL database.

3. System application layer: This layer is responsible for receiving and handling user requests. A single-page application based on Vue framework was used as the front-end technology of the background management system, and Element User Interface was used as the user interface component library to build the web pages and interactive interface of the system. For patient users, the development framework of the WeChat Mini Program was adopted to realize the user's access and use of the system in WeChat (the most popular mobile app in China).
4. User interface layer: This is the interface layer for the system to interact with users and is responsible for displaying the data and functions of the system to users in an intuitive and friendly way.

Figure 1. Architecture of the mobile health system. mHealth: mobile health; PD: peritoneal dialysis; UI: user interface.

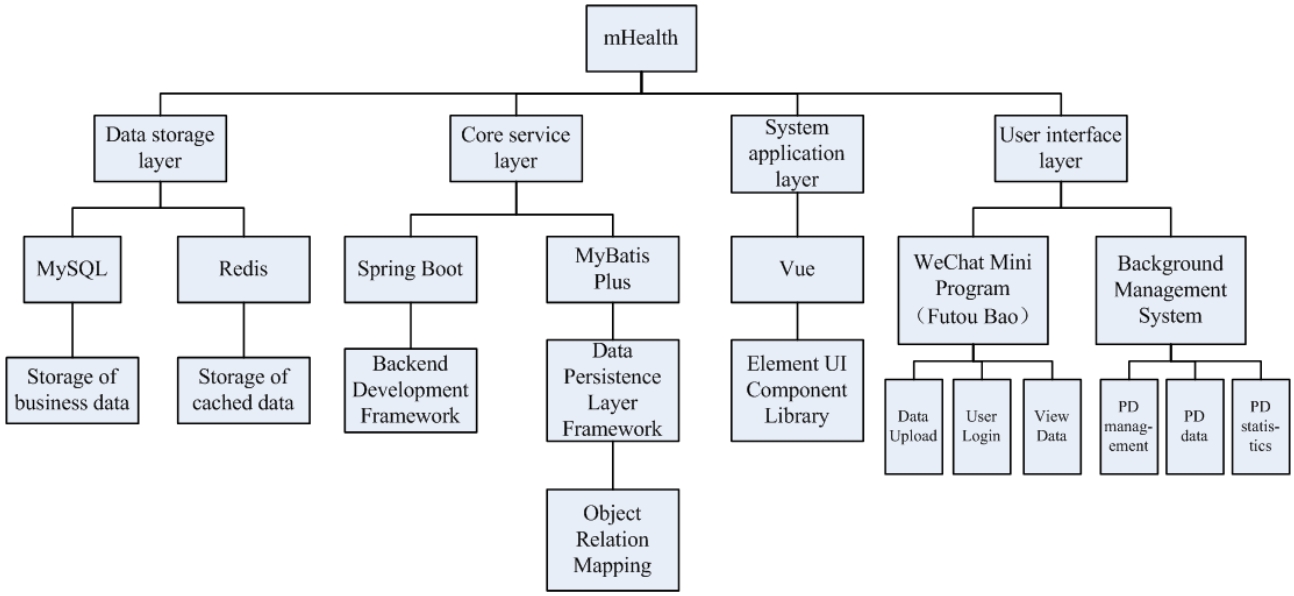

Design of the Longitudinal Study and Model Prediction

From January 1, 2017, to April 20, 2023, a longitudinal study was conducted based on the usage of the Futou Bao system among outpatients of Shenzhen Hospital of Traditional Chinese Medicine. Patients who met the following criteria were recruited and included in the longitudinal study. The inclusion criteria for the participants were as follows.

1. Met the clinical standard of home peritoneal dialysis, including (1) patients with ESRD who needed continuous renal replacement therapy and chose peritoneal dialysis as the kidney replacement therapy; (2) patients underwent peritoneal dialysis catheterization, and the peritoneal dialysis catheter was inserted, which met the requirements of the peritoneal dialysis treatment; and (3) patients accepted the training of home peritoneal dialysis operation and obtained the qualification of self-completion of peritoneal dialysis operation after the training.
2. Intended to receive medical follow-up in Shenzhen Hospital of Traditional Chinese Medicine.
3. Had no difficulty in using smartphones and WeChat.
4. Willing to follow the procedure of the mHealth tool for operate peritoneal dialysis at home and sign the informed consent document.

The exclusion criteria included (1) patients diagnosed with heart failure and met either of the following 2 conditions, namely, acute heart failure classified as New York Heart Association class III or above and nonacute heart failure classified as New York Heart Association class II or below but combining with peripheral edema; and (2) patients had other serious physical and mental diseases that they could not complete the peritoneal dialysis at home with the help of the mHealth tool that we developed. In addition, in the process of using the mHealth tool, when clinical evaluation indicated that the patient did not meet the conditions of home peritoneal dialysis and needed hospitalization, the use of the mHealth tool would be stopped immediately.

Prior to patient recruitment, a specialized patient home peritoneal dialysis management team was selected, including professional doctors and nurses from the Department of Nephrology. Training for the patient management team was conducted, including home peritoneal dialysis operation and system use methods, patient recruitment methods, patient follow-up methods, and management methods. Thereafter, the above management team recruited the patients in their daily work and conducted one-on-one training for the patients who met the enrollment criteria. Patients were asked to use the mHealth tool during each peritoneal dialysis at home. Designated doctors and nurses were assigned to patients for treatment monitoring and follow-up.

During the usage of Futou Bao, patients' daily peritoneal dialysis treatment data were recorded in the system, which were obtained to create the dataset of the modeling process. In the dataset, identifiable information of the patients, such as names, were replaced with a string of encrypted ID numbers to blind the data analysts during analysis. Sex, age, time duration of peritoneal

XSL•FO
RenderX

dialysis, weight, urine volume, diastolic blood pressure, systolic blood pressure, and pulse pressure difference were obtained and included in the dataset as prediction variables to predict the onset of heart failure. Because the dialysate recommended in the home-based peritoneal dialysis was 1.5% glucose concentrations, the type of dialysate was not considered as a prediction variable in this study. For the heart failure outcome, because the occurrence of heart failure was always a progressive process among patients who received peritoneal dialysis, the 15 days before the date of hospitalization for heart failure was defined as the time duration of progression of heart failure. The data collected in that time duration were considered to be related to heart failure. Due to the careful and close patient management, the exact time of heart failure hospitalization could be obtained so as to avoid information bias. Because heart failure occurred infrequently, we randomly selected 6 times the number of patients with heart failure from the data of patients without heart failure to form the study dataset (number of patients with heart failure vs number of patients without heart failure = 1:6). After that, 80% of the data in the study dataset were randomly selected as the training dataset, while 20% were selected as the validation dataset of the prediction model. The sampling method without replacement was adopted in the above process. In the sensitivity analysis, we set the ratio of patients with and without heart failure to 1:4 and 1:10, respectively, to better evaluate the stability of the prediction model.

Statistical Analysis

The therapeutic effect and process of patients undergoing peritoneal dialysis treatment is self-specific, which should be considered in conjunction with the patient's own treatment monitoring data. For improving the adaptability of prediction model, it is more important and valuable to observe patients' own indicator value changes, including weight, urine volume, and blood pressure. Considering the different baseline data of each patient, the least squares method was used to fit the data of each patient, and discrete points were defined as the actual data that were not in the confidence intervals of the fitted data calculated by least squares method [11]. With the help of this method, the urine volume, diastolic blood pressure, systolic blood pressure, and pulse pressure difference were converted from numerical variables to categorical variables (ie, normal, abnormal). In this study, both α=.05 and α=.01 were used as the CI criterion (ie, 95% CI and 99% CI) to divide the categories of variables, so as to choose the optimal model.

Logistic regression models were used as prediction models. Four models were built using the training dataset. Models 1 and 2 used the variables based on the 95% CI criterion, while models 3 and 4 used the variables based on the 99% CI criterion. In models 1 and 3, all of the variables were included, while models 2 and 4 used the backward method and only included variables that were statistically significant. Receiver operating characteristic curves were presented, and the corresponding area under the curve (AUC) values were calculated. The AUC value and Akaike's information criterion (AIC) value were considered comprehensively to select the optimal model, which showed the highest AUC value and the lowest AIC value. Using the validation dataset, the sensitivity, specificity, accuracy, and Youden index of the optimal model were calculated, and the receiver operating characteristic curve and decision curve were presented to evaluate the performance of the prediction model comprehensively. The specific calculation formulas were as follows (Figure 2):

Sensitivity = True Positive / (True Positive + False Negative) (Formula 1)

Specificity = True Negative / (False Positive + True Negative) (Formula 2)

Accuracy = (True Positive + True Negative) / (True Positive + False Positive + False Negative + True Negative) (Formula 3)

Youden index = (Sensitivity + Specificity) – 1 (Formula 4)

A 2-sided *P* value of .05 was used for null hypothesis tests. Statistical analyses were conducted using SPSS version 19.0 (IBM Corp) and R version 4.3.0 (R Core Team).

Figure 2. Definitions of true positive, false positive, false negative, and true negative.

	Actual heart failure	Actual non-heart failure
Predicted heart failure	True Positive (TP)	False Positive (FP)
Predicted non-heart failure	False Negative (FN)	True Negative (TN)

Ethics Approval

This study was approved by the medical ethics committee of Shenzhen Hospital of Traditional Chinese Medicine (B2019043 and K2021-036-01). Written informed consent was obtained from the participants. In the analysis process, identifiable information of the patients (ie, names) was replaced with a string of encrypted ID numbers to protect patients' privacy. There was no compensation (money or gift) provided to the participants during the implementation of the study. However, participants received careful and fair health services and management of home-based peritoneal dialysis treatment.

XSL•FO
RenderX

Results

mHealth Tools Development: Modules and Functions

In order to realize safe and convenient peritoneal dialysis for patients at home and to make it convenient for doctors to know the treatment and the physical condition of the patients in real time, an mHealth system was developed, comprising a WeChat Mini Program for patients named Futou Bao (“腹透宝” in Chinese) and a patient data management platform used by doctors (Figure 3).

Figure 3. Functional flow diagram. PD: peritoneal dialysis.

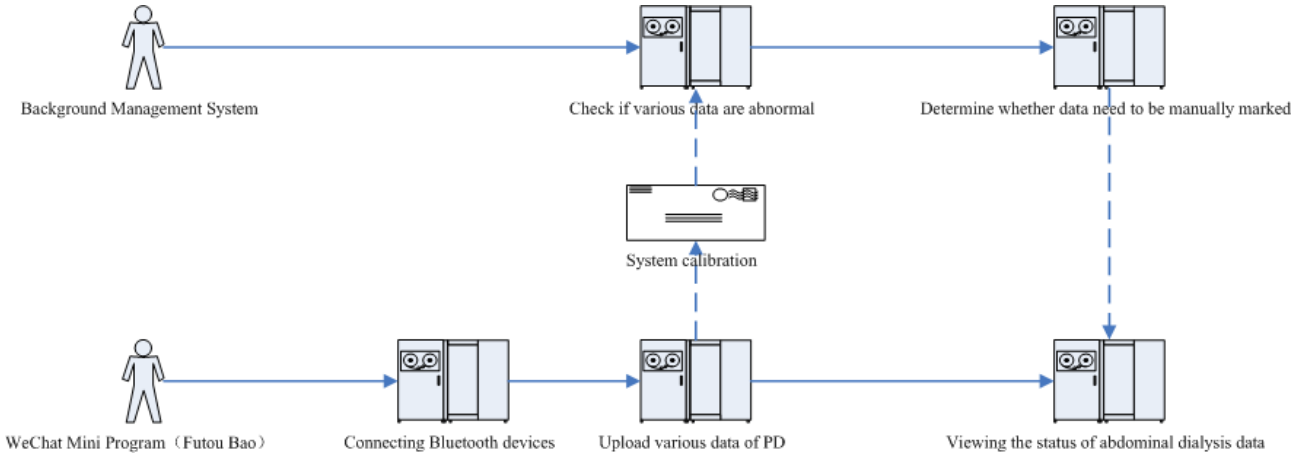

WeChat Mini Program (Futou Bao) for Patients

Futou Bao is a WeChat Mini Program that patients can use easily on WeChat, which is one of the most popular mobile social apps in China by scanning the quick response code or searching the name (“腹透宝” in Chinese) through WeChat. The main function of Futou Bao is the intelligent data upload function for patients receiving peritoneal dialysis treatment at home. Futou Bao has a series of auxiliary functions to complement this main function module.

Intelligent Data Upload Function Module

In this module, patients’ data during the treatment can be uploaded to the system in real time, including blood pressure, urine volume, weight, and peritoneal dialysis solution weight at different time points during peritoneal dialysis. Data can be uploaded at patients’ homes in the following 2 ways: by the patients filling in the interface of the module and by the devices automatically uploading through Bluetooth. The detailed data uploading process of the patients during peritoneal dialysis treatment is as follows (Figure 4):

1. Patients measure their own blood pressure and upload it through the module interface, including systolic and diastolic blood pressure, on the day of peritoneal dialysis.
2. Patients need to measure the urine volume per urination according to the training method and calculate the urine volume on the day of receiving peritoneal dialysis treatment and upload the data through the module interface.
3. After peritoneal dialysis, patients need to weigh themselves on the Bluetooth-enabled weighing scale, after which the weight data can be uploaded directly to the system through Bluetooth (Figure 4).
4. For the peritoneal dialysis solution weight, Futou Bao adopts the method of intelligent data collection in the whole process of peritoneal dialysis treatment. Patients need to place the peritoneal dialysis solution on a Bluetooth-enabled spring scale or electronic flat scale before the start of peritoneal dialysis treatment, and the device can upload the peritoneal dialysis solution weight data through Bluetooth in real time (Figure 4).

XSL•FO
RenderX

Figure 4. User interface of Futou Bao. (A) Uploading data through Bluetooth. (B) Peritoneal dialysis operation records. PD: peritoneal dialysis.

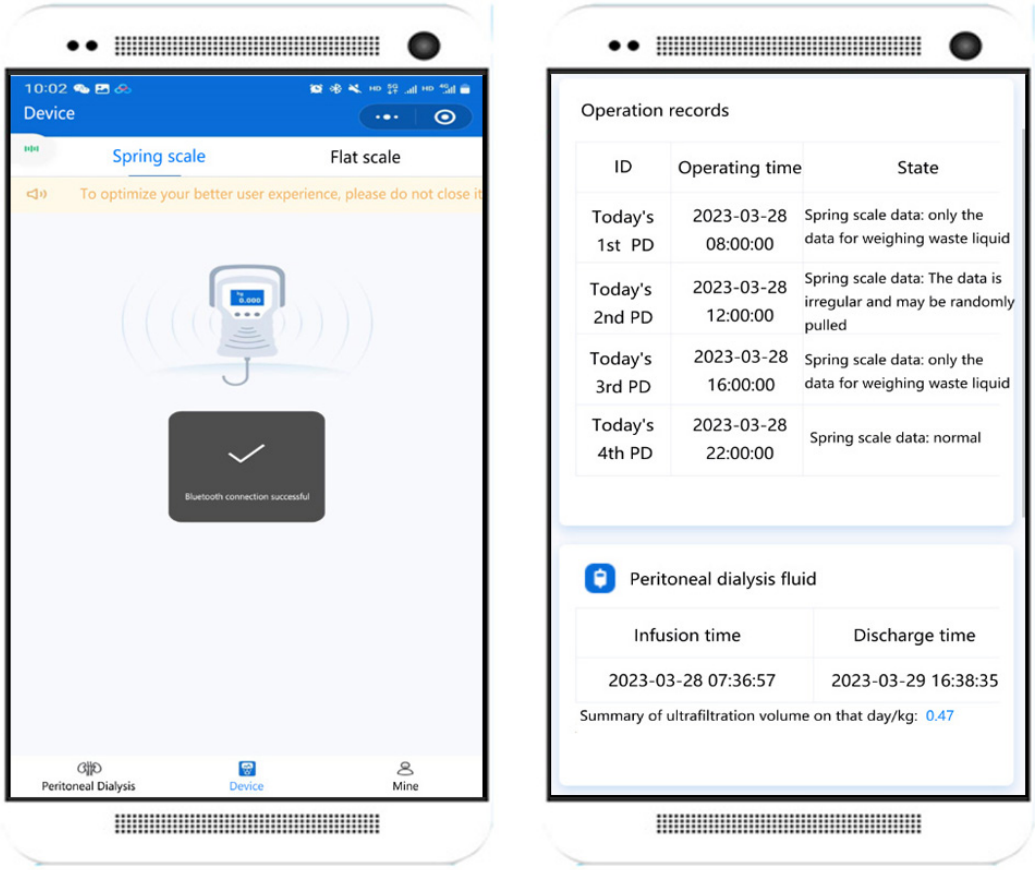

(A) (B)

After uploading peritoneal dialysis data, patients can also check whether the peritoneal dialysis operation is correct or not in this module, including abnormal problems automatically identified by the system and manually marked by doctors so that patients can improve their peritoneal dialysis operation at home (Figure 4).

Auxiliary Function Module

There are 2 auxiliary functions to support the normal operation of the data upload function, as follows.

1. Password change and retrieve function: newly enrolled patients will have a default password, but patients can change the password according to personal needs. If patients forget the password, they can request password retrieval.
2. Bluetooth device connection management function: The patient can see the connection status of the Bluetooth-enabled scales on Futou Bao. When the patients encounter problems when connecting the Bluetooth device, the system will prompt the corresponding reason such as Bluetooth not enabled or connection timeout to help the patient solve the connection problem.

Patients' Data Management Platform for Doctors

The management platform for doctors contains 4 function modules: patient management module, data visualization and marking module, data statistics module, and system management module (Figures 5-9).

XSL•FO
RenderX

Figure 5. Management of patients' daily peritoneal dialysis records, including real-time data of peritoneal dialysis fluid, weight, blood pressure, and urine output.

Peritoneal Dialysis Treatment Data

Name: Gender: Male Age: 45

2023-06-26

Classification	ID	Time		Type of peritoneal dialysis fluid	Data (unit)				Treatment Duration	State
		Infusion time	Discharge time		Liquid intake	Effluent volume	Ultrafiltration	Mode		
Peritoneal dialysis fluid	1	2023-06-26 00:46:00	2023-06-26 08:00:47	Low calcium_2.5	2.035(Kg)	2.245(Kg)	0.21(Kg)	Regular abdominal retention	7.257h	Normal
	2	2023-06-26 08:27:49	2023-06-26 12:02:08	Low calcium_2.5	2.015(Kg)	2.335(Kg)	0.34(Kg)	Stay on the stomach overnight	3.591h	Normal
	3	2023-06-26 12:47:50	2023-06-26 16:12:12	Low calcium_2.5	2.01(Kg)	2.29(Kg)	0.28(Kg)	Regular abdominal retention	3.579h	Normal
	4	2023-06-26 16:21:08	2023-06-26 20:03:16	Low calcium_2.5	2(Kg)	2.36(Kg)	0.36(Kg)	Regular abdominal retention	3.832h	Correction mark
	5	2023-06-26 20:34:48	2023-06-27 00:01:16	Low calcium_2.5	2(Kg)	2.35(Kg)	0.35(Kg)	Regular abdominal retention	3.442h	Normal
	Summary of ultrafiltration volume on that day	1.54(Kg)								
Weight										
Blood pressure										
Urine output										

Figure 6. Map of patients.

Home Page | Patient Map

XSL•FO
RenderX

Figure 7. Scatterplot of the peritoneal dialysis solution weight during the treatment.

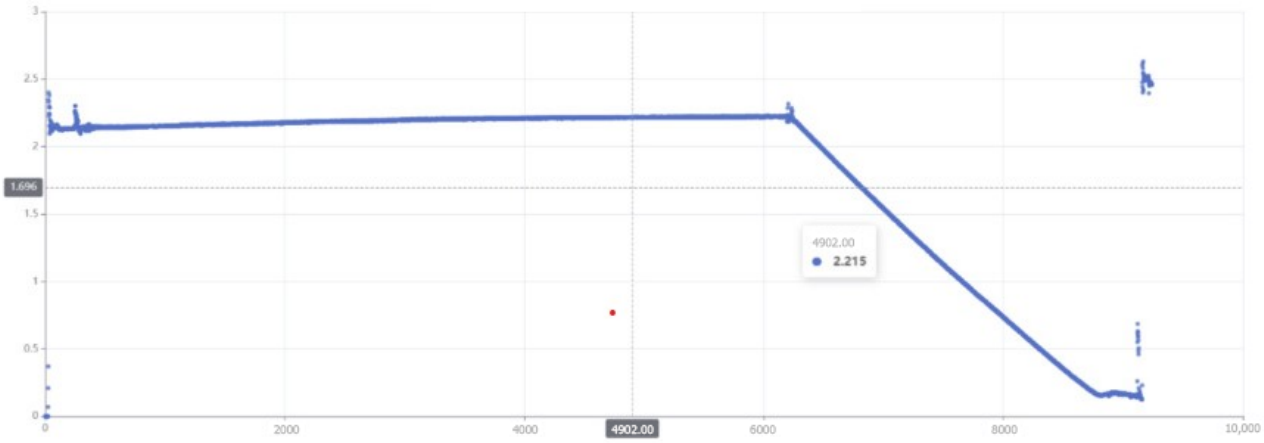

Figure 8. Peritoneal dialysis solution weight data at each treatment time point.

Data point

State	ID	Time	Weight	Manual marking ID	Manual marking Weight(kg)	Edit
Start point	30	2023-06-27 08:10:59	2.225			Manual location
Flushing point	32	2023-06-27 08:11:00	2.38			Manual location
Starting point of liquid entry	6217	2023-06-27 08:36:25	2.21			Manual location
End point of liquid entry	8769	2023-06-27 08:46:54	0.165			Manual location
Sealing starting point	8769	2023-06-27 08:46:54	0.165			Manual location
Sealing end point	9146	2023-06-27 08:48:27	0.165			Manual location
Hanging waste liquid point	9177	2023-06-27 08:48:35	2.505			Manual location

Figure 9. Statistics of daily and weekly peritoneal dialysis frequency and proportion of effective operations.

Name	Daily peritoneal dialysis frequency	Daily operational efficiency	Number of peritoneal dialysis attempts this week	Effective operation this week	Number of peritoneal dialysis attempts last week	Last week's operation efficiency
[redacted]	5.0	60.0%	7.0	85.714%	33.0	93.939%
[redacted]	3.0	0.0%	4.0	0.0%	21.0	0.0%
[redacted]	3.0	100.0%				
[redacted]	5.0	100.0%				
[redacted]	3.0	100.0%				

Patient Management Module

There are 3 main functions in this module. The first function is that doctors can view and manage the patients' basic information, including age, sex, date of first dialysis, and time duration of peritoneal dialysis treatment. The second function is the management of patients' daily peritoneal dialysis records, in which doctors can query and view patients' specific treatment records according to the treatment date (Figure 5). Based on the treatment records, doctors can know the regularity and operation of patients' peritoneal dialysis at home. The third function in this module can provide a map of the patients' addresses, which can help doctors to identify and help patients with operational problems during peritoneal dialysis in a timely manner (Figure 6).

Peritoneal Dialysis Data Visualization and Marking Module

In this module, the data visualization function can generate scatterplots of peritoneal dialysis solution weights based on the data uploaded by the spring scale or flat scale during the treatment (Figure 7). In addition to the scatterplots, doctors can also view the peritoneal dialysis solution weight data of each treatment time point, including liquid inlet starting point, liquid inlet end point, tube sealing starting point, tube sealing end point, and waste liquid hanging point (Figure 8). By viewing patient data and charts, doctors can observe patients' peritoneal dialysis operation habits, manually mark patients' abnormal operations and causes, and display corresponding prompts to patients through Futou Bao to correct possible wrong operations. Common marks for abnormal data include forgetting to weigh the waste liquid, only recording liquid output phase data, lack

XSL•FO
RenderX

of liquid output phase data, and irregular data (which may be caused by randomly pulling the spring scale).

Data Statistics Module

The data statistics module consists of 4 main functions: follow-up statistics, effective operation statistics, operation time statistics of each treatment point, and comparative statistics of treatment data in recent days. First, follow-up statistics can check the number of patients' follow-up visits, the number of new Futou Bao users, and the number of withdrawals in each year. Second, effective operation statistics function can be used to check the number of peritoneal dialysis treatments and effective operations on a given date. The effective operation rate is calculated to evaluate the overall situation of operational capability and correctness. This function makes it convenient for doctors to correct operational errors for patients (Figure 9). Third, the operation time statistics of each treatment point can be effectively viewed together with the above function. The total operation time, hand hygiene time, infusion time, and tube sealing time of the patients can be viewed. Fourth, comparative statistics of treatment data in recent days function can provide doctors with recent days' important process data during peritoneal dialysis and mark the trends of data (ie, rising or declining), which can make it easier for doctors to know the effect of peritoneal dialysis treatment.

System Management Module

The system provides permission control menus based on the role-based access control method. The super administrator assigns management rights to different roles among medical staff. This process can fully ensure the meticulous management of patients by medical staff and protect the privacy of patients.

Prediction Model of Heart Failure

General Information of the Study Participants

During the time period of this study, records of 6635 patients who underwent peritoneal dialysis treatment were uploaded in Futou Bao with 0.71% (47/6635) showing heart failure. Based on the 1:6 random sampling, treatment data of 329 patients were included in the modeling analysis, that is, 47 (14.3%) patients with heart failure and 282 (85.7%) patients without heart failure. Table 1 shows the detailed characteristics of the included patients. Most of them were females (228/329, 69.3%), with a median age of 46.55 years and a median peritoneal dialysis time duration of 0.94 years. According to the ratio of 8:2, the study dataset was randomly divided into the training dataset and the verification dataset, which included 264 patients and 65 patients, respectively (Figure 10).

XSL•FO
RenderX

Table 1. Basic characteristics of the patients using Futou Bao.

Characteristics	All patients (N=6635)	Patients included in the analysis		
		Total (n=329)	Training set (n=264)	Validation set (n=65)
Sex, n (%)				
Male	2578 (38.85)	101 (30.70)	79 (29.92)	22 (33.85)
Female	4057 (61.15)	228 (69.30)	185 (70.08)	43 (66.15)
Age (years), median (IQR)	46.90 (44.68, 50.59)	46.55 (43.96, 51.19)	46.48 (43.94, 50.82)	46.91 (44.04, 52.68)
Time duration of peritoneal dialysis (years), median (IQR)	1.00 (0.53, 1.64)	0.94 (0.49, 1.78)	0.91 (0.46, 1.77)	1.08 (0.53, 1.87)
Weight, n (%)				
Based on 95% CI				
Normal	6115 (92.16)	299 (90.88)	239 (90.53)	60 (92.31)
Abnormal	520 (7.84)	30 (9.12)	25 (9.47)	5 (7.69)
Based on 99% CI				
Normal	6458 (97.33)	320 (97.26)	257 (97.35)	63 (96.92)
Abnormal	177 (2.67)	9 (2.74)	7 (2.65)	2 (3.08)
Urine volume, n (%)				
Based on 95% CI				
Normal	6089 (91.77)	298 (90.58)	241 (91.29)	57 (87.69)
Abnormal	546 (8.23)	31 (9.42)	23 (8.71)	8 (12.31)
Based on 99% CI				
Normal	6384 (96.22)	313 (95.14)	252 (95.45)	61 (93.85)
Abnormal	251 (3.78)	16 (4.86)	12 (4.55)	4 (6.15)
Systolic blood pressure, n (%)				
Based on 95% CI				
Normal	6189 (93.28)	299 (90.88)	239 (90.53)	60 (92.31)
Abnormal	446 (6.72)	30 (9.12)	25 (9.47)	5 (7.69)
Based on 99% CI				
Normal	6458 (97.33)	314 (95.44)	251 (95.08)	63 (96.92)
Abnormal	177 (2.67)	15 (4.56)	13 (4.92)	2 (3.08)
Diastolic blood pressure, n (%)				
Based on 95% CI				
Normal	6326 (95.34)	313 (95.14)	251 (95.08)	62 (95.38)
Abnormal	309 (4.66)	16 (4.86)	13 (4.92)	3 (4.62)
Based on 99% CI				
Normal	6515 (98.19)	319 (96.96)	256 (96.97)	63 (96.92)
Abnormal	120 (1.81)	10 (3.04)	8 (3.03)	2 (3.08)
Pulse pressure difference, n (%)				
Based on 95% CI				
Normal	6232 (93.93)	299 (90.88)	238 (90.15)	61 (93.85)
Abnormal	403 (6.07)	30 (9.12)	26 (9.85)	4 (6.15)
Based on 99% CI				
Normal	6478 (97.63)	316 (96.05)	254 (96.21)	62 (95.38)
Abnormal	157 (2.37)	13 (3.95)	10 (3.79)	3 (4.62)

XSL•FO
RenderX

Characteristics	All patients (N=6635)	Patients included in the analysis		
		Total (n=329)	Training set (n=264)	Validation set (n=65)
Heart failure, n (%)				
Yes	47 (0.71)	47 (14.29)	39 (14.77)	8 (12.31)
No	6588 (99.29)	282 (85.71)	225 (85.23)	57 (87.69)

Figure 10. Data processing flowchart. PD: peritoneal dialysis.

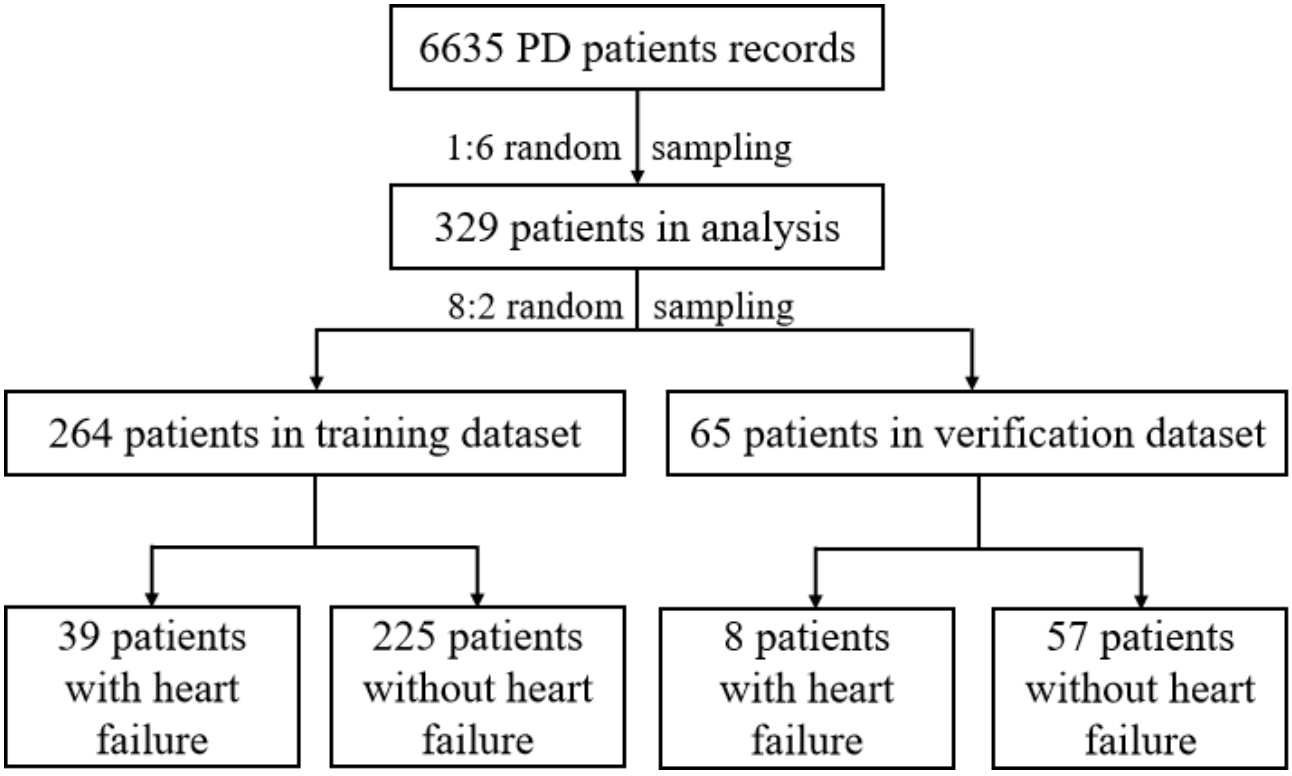

Optimal Prediction Model Selection

Four prediction models based on logistic regression were constructed using the training dataset (Tables 2-3). The AIC values of the 4 models were 147.23, 138.62, 145.18, and 136.50, respectively. The AUC value of model 4 (the lowest AIC value) was higher than that of model 1 (the highest AIC value) ($AUC_{Model\ 4}$ vs $AUC_{Model\ 1}$: 0.896 vs 0.893); however, the difference was not statistically significant (P=.78) (Figure 11).

Based on the above comparison, model 4 was considered as the optimal prediction model. Sex, age, and diastolic blood pressure were included in model 4. Females were less likely to experience heart failure than males (adjusted odds ratio [AOR] 0.06, 95% CI 0.01-0.22), while older age (AOR 1.19, 95% CI 1.13-1.25) and abnormal diastolic blood pressure during treatment (AOR 38.79, 95% CI 5.08-291.63) were the risk factors of heart failure (Table 3).

XSL•FO
RenderX

Table 2. Impact factors of heart failure in patients receiving peritoneal dialysis treatment (abnormal value based on 95% CI).

Characteristics	Univariable model		Multivariable model 1		Multivariable model 2	
	Odds ratio (95% CI)	*P* value	AOR[a] (95% CI)	*P* value	AOR (95% CI)	*P* value
Sex						
Male	1.00	N/A[b]	1.00	N/A	N/A	N/A
Female	0.10 (0.02-0.35)	.002	0.05 (0.01-0.22)	<.001	0.05 (0.01-0.26)	<.001
Age	1.17 (1.12-1.23)	<.001	1.18 (1.11-1.28)	<.001	1.17 (1.12-1.23)	<.001
Time duration of peritoneal dialysis	2.73 (1.82-4.32)	<.001	0.92 (0.42-1.80)	.83	N/A	N/A
Weight						
Normal	1.00	N/A	1.00	N/A	N/A	N/A
Abnormal	0.48 (0.07-1.70)	.33	1.24 (0.15-6.22)	.81	N/A	N/A
Urine volume						
Normal	1.00	N/A	1.00	N/A	N/A	N/A
Abnormal	3.61 (1.36-9.06)	.007	1.54 (0.31-7.84)	.59	N/A	N/A
Systolic blood pressure						
Normal	1.00	N/A	1.00	N/A	N/A	N/A
Abnormal	4.83 (1.94-11.68)	.001	4.49 (0.62-45.50)	.17	8.02 (2.09-30.77)	.002
Diastolic blood pressure						
Normal	1.00	N/A	1.00	N/A	N/A	N/A
Abnormal	2.74 (0.71-8.93)	.11	2.31 (0.22-17.93)	.44	N/A	N/A
Pulse pressure difference						
Normal	1.00	N/A	1.00	N/A	N/A	N/A
Abnormal	2.37 (0.87-5.59)	.07	1.05 (0.11-6.83)	.96	N/A	N/A

[a]AOR: adjusted odds ratio.

[b]N/A: not applicable.

XSL•FO
RenderX

Table 3. Impact factors of heart failure in patients receiving peritoneal dialysis treatment (abnormal value based on 99% CI).

Characteristics	Univariable model		Multivariable model 3		Multivariable model 4	
	Odds ratio (95% CI)	*P* value	AOR[a] (95% CI)	*P* value	AOR (95% CI)	*P* value
Sex						
Male	1.00	N/A[b]	1.00	N/A	1.00	N/A
Female	0.10 (0.02-0.35)	.002	0.04 (0.01-0.22)	.001	0.06 (0.01-0.22)	<.001
Age	1.17 (1.12-1.23)	<.001	1.20 (1.12-1.30)	<.001	1.19 (1.13-1.25)	<.001
Time duration of peritoneal dialysis	2.73 (1.82-4.32)	<.001	0.95 (0.41-1.89)	.89	N/A	N/A
Weight						
Normal	1.00	N/A	1.00	N/A	N/A	N/A
Abnormal	2.38 (0.33-11.49)	.31	4.34 (0.23-44.51)	.25	N/A	N/A
Urine volume						
Normal	1.00	N/A	1.00	N/A	N/A	N/A
Abnormal	3.10 (0.79-10.41)	.08	1.19 (0.11-12.44)	.88	N/A	N/A
Systolic blood pressure						
Normal	1.00	N/A	1.00	N/A	N/A	N/A
Abnormal	3.99 (1.15-12.69)	.02	1.17 (0.04-18.27)	.91	N/A	N/A
Diastolic blood pressure						
Normal	1.00	N/A	1.00	N/A	1.00	N/A
Abnormal	6.31 (1.44-27.83)	.01	9.87 (0.07-290.03)	.41	38.79 (5.08-291.63)	<.001
Pulse pressure difference						
Normal	1.00	N/A	1.00	N/A	N/A	N/A
Abnormal	4.17 (1.02-15.35)	.03	1.35 (0.01-74.38)	.90	N/A	N/A

[a]AOR: adjusted odds ratio.

[b]N/A: not applicable.

XSL•FO
RenderX

Figure 11. Receiver operating characteristic curves of 2 heart failure prediction models in the training dataset. (A) Initial model (model 1). (B) Final model (model 4). (C) Model comparison. AUC: area under the curve; FPR: false positive rate; TPR: true positive rate.

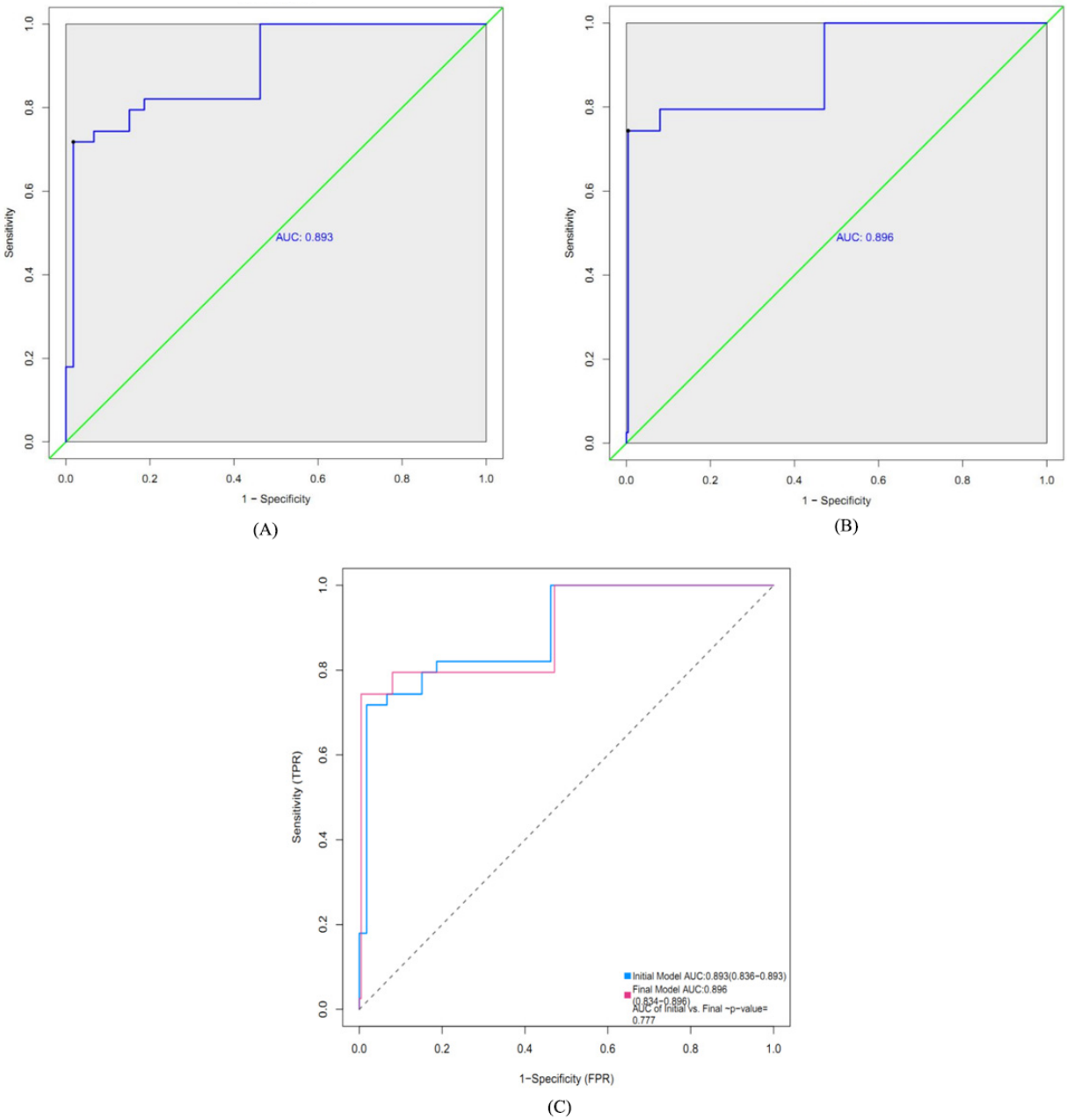

Validation of the Optimal Prediction Model

Based on the validation dataset, the sensitivity, specificity, accuracy, and Youden index of the optimal model were 0.75, 0.91, 0.89, and 0.66, respectively. Figure 12A shows the ROC curve of the model, of which the AUC value was 0.879 (95% CI 0.772-0.986). The decision curve also showed the acceptable performance of the model (Figure 12B).

XSL•FO
RenderX

Figure 12. Performance of the final model in the testing dataset. (A) Receiver operating characteristic curve. (B) Decision curve. AUC: area under the curve.

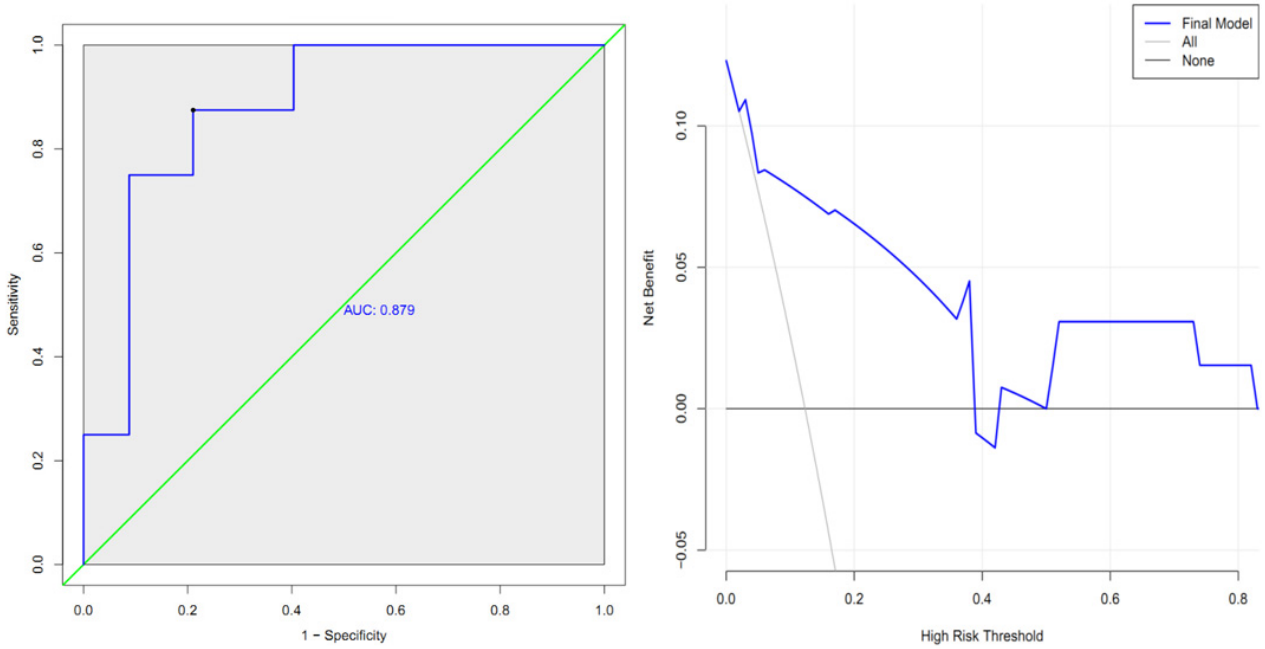

Sensitivity Analysis

In the sensitivity analysis, after changing the ratio of patients with and without heart failure (1:4 and 1:10), the analysis showed stable results. The variables included in the optimal model were constant, including sex, age, and diastolic blood pressure during treatment (Tables S1-S2 of Multimedia Appendix 1). In validation, the models also showed good performance. The AUC values were 0.877 (95% CI 0.774-0.980) in the 1:4 ratio dataset and 0.913 (95% CI 0.825-0.999) in the 1:10 ratio datasets, respectively (Figures S1-S2 in Multimedia Appendix 1). The Youden indexes were 0.76 and 0.68 in the 2 datasets, respectively.

Discussion

Principal Findings

In this study, a practical and user-friendly mHealth tool named Futou Bao and its supporting management system were developed for patients with ESRD who managed their peritoneal dialysis treatment at home in China. With its comprehensive functions that provided intelligent data upload, visualization, management, and doctor-patient interaction, the mHealth tool and system could effectively solve the difficulties in managing peritoneal dialysis treatment at home, increase treatment effectiveness, and minimize complication risk. Compared with the CKD-PD (chronic kidney disease-peritoneal dialysis) app [12] developed by researchers in the Khon Kaen University, Thailand, in the same period as our study, Futou Bao had several additional advantages. First, automatic data upload via Bluetooth could reduce data recording errors. Second, the recording of peritoneal dialysis solution weight data was more detailed, which included the weight data at each time point during the whole treatment process, and it could generate data visualization plots timely. Third, Futou Bao could provide better interactions between doctors and patients. The operation problems marked by doctors can be notified to the patients through the platform in a timely manner to help patients improve the peritoneal dialysis operation quality. Fourth, the Futou Bao system was able to locate patients' addresses to detect serious health risks and provide timely medical assistance. In addition, a previous study conducted among Thai patients using the CKD-PD app mainly focused on the optimization of the app's functionality and usefulness with only 10 participants involved [12]. By contrast, in this study, a larger number of patients were included to explore a better way for assisting clinical decision-making with the help of constructing prediction models based on the usage records of Futou Bao. We found that most of the patients with ESRD using Futou Bao, as an eHealth management and promoting tool, in our study were females, which was different from the real-world clinical experiences, wherein most patients treated for ESRD with any form of dialysis were males. However, in the process of recruitment, we only relied on patients' voluntary participation in the study and did not induce them in any form or select any of them based on their sex. This interesting phenomenon can be explained by the higher acceptance of the eHealth approaches among females. A study focusing on drug use treatment in China using a mobile app also showed that female users were more willing to receive health care messages and communicate with health workers through the mobile app and were more likely to have positive health changes than male users, which could support our finding [13].

Given the importance of early detection of heart failure in patients receiving home-based peritoneal dialysis, it is of great value to make full use of the data recorded in the process of

XSL•FO
RenderX

daily peritoneal dialysis treatment to perform real-time prediction and for early warning of heart failure risk. However, the variables recorded during the daily home-based peritoneal dialysis were limited to some extent. In previous studies, the prediction models of heart diseases among patients with CKD and ESRD mainly relied on clinical test indicators and biomarkers that could not be tested at home. Kessler et al [14] built a risk prediction model of cardiovascular events for patients with ESRD based on the FOSIDIAL trial and found that C-reactive protein and left ventricular mass index were important prognostic factors of cardiovascular events. Weekley and Peralta [15] indicated that combining creatinine, cystatin C, and albuminuria could improve the detection and risk stratification for heart failure and other cardiovascular events. Other biomarkers, including estimated glomerular filtration rate, urine albumin-to-creatinine ratio, and β_2-microglobulin, were also proved to be associated with heart failure in patients with CKD [16-20]. Based on evidences from the Chronic Renal Insufficiency Cohort study, a Pooled Cohort Equations to Prevent Heart Failure in the CKD population was constructed to predict incident heart failure hospitalizations [21]. A series of clinical factors were included in this model, such as blood pressure, fasting glucose, total cholesterol, high-density lipoprotein cholesterol, QRS duration, estimated glomerular filtration rate, and albuminuria [21]. A recent research based on the data of a Chronic Renal Insufficiency Cohort study indicated that adding N-terminal brain natriuretic peptide and high sensitivity troponin-T to the prediction model could improve the discrimination of heart failure [22].

Considering the unavailability of laboratory test results in daily home-based peritoneal dialysis and the lack of heart failure prediction model in patients receiving peritoneal dialysis treatment, our study innovatively establishes a practical prediction model of heart failure in patients undergoing peritoneal dialysis based on the monitoring data of daily peritoneal dialysis treatment at home. The factors in the prediction model included sex, age, time duration of peritoneal dialysis, weight, urine volume, and blood pressure, which were accessible and real-time data monitored and recorded by the Futou Bao. In addition, the effectiveness of the prediction model based on routine monitoring variables was acceptable. In the model validation, the accuracy of the prediction model could reach 89%, with sensitivity of 75% and higher specificity of 91%. Therefore, the mHealth tools and the predictive model developed in this study enabled monitoring and early warning of heart failure risk, which has formed a new paradigm for the management of patients receiving home-based peritoneal dialysis treatment. From another perspective, during the 5-year follow-up, the proportion of patients experiencing heart failure in our study was less than 1%, which was much lower than that reported in previous studies [7-10]. This result could indicate that our management paradigm was practical and effective.

Limitations

There were several limitations in this study. First, the number of patients with heart failure during the observation period of this study was small, which resulted in a relatively small number of patients in the dataset for the prediction model study. Volunteer bias may enhance the management effectiveness of the system. In future studies, we will conduct randomized controlled trials to better rule out confounding factors and reduce the occurrence of bias. Second, the mHealth tools were only used for daily data collection and monitoring, and the exclusion criteria were relatively strict that patients with previous heart failure or coronary heart disease history were not included in this study. To better expand the coverage of patients included and managed in the future, more variables in their medical history should be added to better predict the risk of heart failure. Third, the mHealth system was only used in the scenario of large cities represented by Shenzhen. In future studies, the regional scope of the system should be further expanded. In addition, the sex and age composition of the users was different from that in clinical practice experiences. In the future, the performance of Futou Bao should be promoted in the following 3 ways: (1) further investigation is needed to explore the obstacles in male users' acceptance of Futou Bao, and the functional design should consider the needs of both male and female users; (2) the improvement of Futou Bao should enhance the usage experience of older adults to help them use the tool more conveniently; and (3) on the basis of ensuring patient privacy, family caregivers can be authorized to use the patient platform so that more older adults who do not have the ability to use smartphones can benefit from using this home-based system.

Conclusions

This is a pioneering study that describes the development of an intelligent home-based peritoneal dialysis management and monitoring system based on mHealth technology in China. This system consists of a WeChat Mini Program named Futou Bao for patients, which includes an intelligent data upload function module and an auxiliary function module; and a patient data management platform for doctors, which includes 4 function modules: patient management, data visualization and marking, data statistics, and system management. Based on the daily peritoneal dialysis treatment data recorded in the system, a real-time prediction model of heart failure in patients receiving peritoneal dialysis treatment was built and verified. This accuracy of the prediction model was accepted, with the AUC value of 0.879 (95% CI 0.772-0.986). The new home-based peritoneal dialysis management paradigm that we developed enabled real-time monitoring and early warning of heart failure risk. In the future, it is essential to evaluate the effectiveness of mHealth tools through multicenter studies, which should include regions with different economic levels and medical and health service levels to improve the universality of the research results.

Acknowledgments

We thank the staff of Shenzhen Hospital of Traditional Chinese Medicine for their contributions in patient management. This work was supported by the Health and Family Planning Commission of Shenzhen Municipality (SZLY2018004); the Guangdong

XSL•FO
RenderX

Medical Science and Technology Research Fund Project, China (B2019043); the Key Laboratory of Medical Rescue Key Technology and Equipment, Ministry of Emergency Management (Open Fund Project YJBKFKT202415); the Natural Science Foundation of Tianjin (23JCQNJC01430); and the National Natural Science Foundation of China (72274210).

Data Availability

The datasets in this study are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: JW, XY, Y Zeng, JY, and ZJ
Data curation: YY, XX, GS, WY, TW, and Y Zheng
Formal analysis: JW, Y Zeng, XY
Funding acquisition: JW, Y Zeng, XY, ZJ
Methodology: JW, Y Zeng, XY
Project administration: Y Zeng, XY, ZJ
Resources: Y Zeng, JW
Supervision: XY, ZJ
Validation: XY, Y Zeng, JY
Visualization: JW, Y Zeng, XY
Writing – original draft: JW, XY, Y Zeng, JY
Writing – review & editing: XY, ZJ

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary data.
[DOCX File , 150 KB - formative_v9i1e56254_app1.docx]

References

1. GBD Chronic Kidney Disease Collaboration. Global, regional, and national burden of chronic kidney disease, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet 2020 Feb 29;395(10225):709-733 [FREE Full text] [doi: 10.1016/S0140-6736(20)30045-3] [Medline: 32061315]
2. Vanholder R, Annemans L, Brown E, Gansevoort R, Gout-Zwart JJ, Lameire N, European Kidney Health Alliance. Reducing the costs of chronic kidney disease while delivering quality health care: a call to action. Nat Rev Nephrol 2017 Jul;13(7):393-409. [doi: 10.1038/nrneph.2017.63] [Medline: 28555652]
3. Liyanage T, Ninomiya T, Jha V, Neal B, Patrice HM, Okpechi I, et al. Worldwide access to treatment for end-stage kidney disease: a systematic review. The Lancet 2015 May;385(9981):1975-1982. [doi: 10.1016/s0140-6736(14)61601-9]
4. Moncrief J. The birth and development of continuous ambulatory peritoneal dialysis. Contrib Nephrol 2017;189:85-90. [doi: 10.1159/000450689] [Medline: 27951554]
5. Li PK, Chow KM, Van de Luijtgaarden MW, Johnson DW, Jager KJ, Mehrotra R, et al. Changes in the worldwide epidemiology of peritoneal dialysis. Nat Rev Nephrol 2017 Feb;13(2):90-103. [doi: 10.1038/nrneph.2016.181] [Medline: 28029154]
6. Teitelbaum I. Peritoneal dialysis. N Engl J Med 2021 Nov 04;385(19):1786-1795. [doi: 10.1056/nejmra2100152]
7. Hou F, Jiang J, Chen J, Yu X, Zhou Q, Chen P, et al. China collaborative study on dialysis: a multi-centers cohort study on cardiovascular diseases in patients on maintenance dialysis. BMC Nephrol 2012 Aug 30;13:94 [FREE Full text] [doi: 10.1186/1471-2369-13-94] [Medline: 22935444]
8. Wang AY, Wang M, Lam CW, Chan IH, Lui S, Sanderson JE. Heart failure with preserved or reduced ejection fraction in patients treated with peritoneal dialysis. Am J Kidney Dis 2013 Jun;61(6):975-983. [doi: 10.1053/j.ajkd.2012.12.030] [Medline: 23474006]
9. Wang I, Lu C, Lin C, Liang C, Yen T, Liu Y, et al. Comparison of the risk of de novo cardiovascular disease between hemodialysis and peritoneal dialysis in patients with end-stage renal disease. Int J Cardiol 2016 Sep 01;218:219-224. [doi: 10.1016/j.ijcard.2016.05.036] [Medline: 27236118]
10. Sun C, Sung J, Wang J, Li C, Kuo Y, Lee C, et al. A comparison of the risk of congestive heart failure-related hospitalizations in patients receiving hemodialysis and peritoneal dialysis - A retrospective propensity score-matched study. PLoS One 2019;14(10):e0223336 [FREE Full text] [doi: 10.1371/journal.pone.0223336] [Medline: 31574134]
11. J W, Z J, M P, T Y. Real-time monitoring and reminding of remote peritoneal dialysis system based on the principle of least squares. Int J Nephrol Kidney Fail 2021;7(3):1-6. [doi: 10.16966/2380-5498.214]

XSL•FO
RenderX

12. Lukkanalikitkul E, Kongpetch S, Chotmongkol W, Morley MG, Anutrakulchai S, Srichan C, et al. Optimization of the chronic kidney disease-peritoneal dialysis app to improve care for patients on peritoneal dialysis in Northeast Thailand: user-centered design study. JMIR Form Res 2022 Jul 06;6(7):e37291 [FREE Full text] [doi: 10.2196/37291] [Medline: 35793137]
13. Li Y, Yan X, Wang Z, Zhang B, Jia Z. Clear the fog of negative emotions: A new challenge for intervention towards drug users. J Affect Disord 2021 Nov 01;294:305-313. [doi: 10.1016/j.jad.2021.07.029] [Medline: 34311330]
14. Kessler M, Zannad F, Lehert P, Grünfeld JP, Thuilliez C, Leizorovicz A, FOSIDIAL Investigators. Predictors of cardiovascular events in patients with end-stage renal disease: an analysis from the Fosinopril in dialysis study. Nephrol Dial Transplant 2007 Dec;22(12):3573-3579. [doi: 10.1093/ndt/gfm417] [Medline: 17611249]
15. Weekley C, Peralta C. Advances in the use of multimarker panels for renal risk stratification. Curr Opin Nephrol Hypertens 2012 May;21(3):301-308 [FREE Full text] [doi: 10.1097/MNH.0b013e328352132d] [Medline: 22449944]
16. Matsushita K, Ballew SH, Coresh J. Cardiovascular risk prediction in people with chronic kidney disease. Current Opinion in Nephrology and Hypertension 2016;25(6):518-523. [doi: 10.1097/mnh.0000000000000265]
17. Matsushita K, Coresh J, Sang Y, Chalmers J, Fox C, Guallar E, et al. Estimated glomerular filtration rate and albuminuria for prediction of cardiovascular outcomes: a collaborative meta-analysis of individual participant data. The Lancet Diabetes & Endocrinology 2015 Jul;3(7):514-525. [doi: 10.1016/s2213-8587(15)00040-6]
18. Foster MC, Coresh J, Hsu CY, et al. Erratum regarding "serum β-trace protein and β2-microglobulin as predictors of ESRD, mortality, and cardiovascular disease in adults with CKD in the Chronic Renal Insufficiency Cohort (CRIC) study. Am J Kidney Dis 2024 Jun;83(6):840. [doi: 10.1053/j.ajkd.2024.04.002] [Medline: 38782534]
19. Matsushita K, Sang Y, Ballew SH, Astor BC, Hoogeveen RC, Solomon SD, et al. Cardiac and kidney markers for cardiovascular prediction in individuals with chronic kidney disease: the Atherosclerosis Risk in Communities study. ATVB 2014 Aug;34(8):1770-1777. [doi: 10.1161/atvbaha.114.303465]
20. Liabeuf S, Lenglet A, Desjardins L, Neirynck N, Glorieux G, Lemke H, European Uremic Toxin Work Group (EUTox). Plasma beta-2 microglobulin is associated with cardiovascular disease in uremic patients. Kidney Int 2012 Dec;82(12):1297-1303 [FREE Full text] [doi: 10.1038/ki.2012.301] [Medline: 22895515]
21. Mehta R, Ning H, Bansal N, Cohen J, Srivastava A, Dobre M, et al. Ten-year risk-prediction equations for incident heart failure hospitalizations in chronic kidney disease: findings from the chronic renal insufficiency cohort study and the multi-ethnic study of atherosclerosis. J Card Fail 2022 Apr;28(4):540-550 [FREE Full text] [doi: 10.1016/j.cardfail.2021.10.007] [Medline: 34763078]
22. Zelnick LR, Shlipak MG, Soliman EZ, Anderson A, Christenson R, Kansal M, CRIC Study Investigators. Prediction of incident heart failure in CKD: The CRIC study. Kidney Int Rep 2022 Apr;7(4):708-719 [FREE Full text] [doi: 10.1016/j.ekir.2022.01.1067] [Medline: 35497796]

Abbreviations

AIC: Akaike's information criterion
AOR: adjusted odds ratio
AUC: area under the curve
CAPD: continuous ambulatory peritoneal dialysis
CKD-PD: chronic kidney disease-peritoneal dialysis
ESRD: end-stage renal disease
mHealth: mobile health

Edited by A Mavragani; submitted 11.01.24; peer-reviewed by M Papasotiriou, X Zhuang; comments to author 31.05.24; revised version received 24.07.24; accepted 03.02.25; published 03.03.25.

Please cite as:
Wu J, Zeng Y, Yang J, Yao Y, Xu X, Song G, Yi W, Wang T, Zheng Y, Jia Z, Yan X
Daily Treatment Monitoring for Patients Receiving Home-Based Peritoneal Dialysis and Prediction of Heart Failure Risk: mHealth Tool Development and Modeling Study
JMIR Form Res 2025;9:e56254
URL: https://formative.jmir.org/2025/1/e56254
doi:10.2196/56254
PMID:40053710

©Jia Wu, Youjia Zeng, Jun Yang, Yutong Yao, Xiuling Xu, Gaofeng Song, Wuyong Yi, Taifen Wang, Yihou Zheng, Zhongwei Jia, Xiangyu Yan. Originally published in JMIR Formative Research (https://formative.jmir.org), 03.03.2025. This is an open-access

XSL•FO
RenderX

article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Experiences With an In-Bed Real-Time Motion Monitoring System on a Geriatric Ward: Mixed Methods Study

Stefan Walzer[1], MSc; Isabel Schön[2], MA; Johanna Pfeil[2], BA; Nicola Merz[2], MA; Helga Marx[3], MSc; Sven Ziegler[3], MSc; Christophe Kunze[1], Prof Dr

[1]Care and Technology Lab, Furtwangen University, Furtwangen im Schwarzwald, Germany

[2]AGP Social Research, FIVE e.V., Freiburg, Germany

[3]Center of Implementing Nursing Care Innovations Freiburg, Nursing Direction, Medical Center, University of Freiburg, Freiburg, Germany

Corresponding Author:
Stefan Walzer, MSc
Care and Technology Lab
Furtwangen University
Robert-Gerwig-Platz 1
Furtwangen im Schwarzwald, 78120
Germany
Phone: 49 7723920295
Email: stefan.walzer@hs-furtwangen.de

Abstract

Background: Older adults now make up about two-thirds of hospital admissions, with up to 50% experiencing cognitive impairments such as dementia. These patients often struggle with adherence to care plans and maintaining regular day or night cycles, presenting challenges for nurses. Hospitals are typically unprepared to manage this patient population, resulting in increased nurse workload and challenges like managing motor agitation, which can lead to falls or accidental removal of medical devices.

Objective: This study aimed to (1) assess how an in-bed real-time motion monitoring system (IRMS) impacts nurses' perceptions of physical and mental stress, (2) evaluate the IRMS's effect on the care process, (3) explore ethical implications like patient autonomy and privacy, and (4) understand how nurses acquire knowledge about the technology and how this affects their assessment of the IRMS.

Methods: The IRMS, which provides real-time motion monitoring and bed edge or exit information, was implemented in the geriatric ward of a university medical center. The study followed a monocentric, explorative evaluation design using a mixed methods approach. It lasted 24 weeks and had two phases. In Phase 0 (6 weeks), patients received standard care. In Phase 1 (18 weeks), the IRMS was introduced. Initial data were gathered through focus groups and participant observations during manufacturer training sessions. At the end of the intervention, a survey, a second focus group, and an interview were conducted to capture nurses' experiences. The study follows the Good Reporting of a Mixed Method Study (GRAMMS) checklist for reporting.

Results: Initial training sessions with 12 participants (10 nurses and 2 physiotherapists) showed varying levels of engagement, with the second session demonstrating more optimism and interprofessional collaboration. A total of 10 questionnaires were completed (10/21, 48%). Survey results showed that 80% (8/10) of nurses found the IRMS valuable for assessing the quality of work, and 90% (9/10) were willing to continue using it. The system was regarded as reliable for monitoring bed edge and exit events. Usability was positively rated, with minimal concerns about documentation burden. Focus group discussions (n=3 per session) indicated that nurses viewed the system as reliable and appreciated its role in reducing anxiety related to fall prevention. However, concerns about patient privacy and monitoring were raised. Nurses expressed a willingness to continue using the IRMS but reaffirmed their ability to care for patients without it.

Conclusions: Nurses had a generally positive attitude toward the IRMS, recognizing its benefits, particularly for nighttime monitoring. Although its effectiveness in preventing falls remains inconclusive, the system helps reduce nurses' fear of falls and enhances their responsiveness. The study highlights the broader impact of the IRMS beyond fall prevention and stresses the importance of thoughtful integration into health care practice.

(JMIR Form Res 2025;9:e63572) doi:10.2196/63572

XSL•FO
RenderX

KEYWORDS

nurses; geriatric patients; cognitive impairment; technology; fall prevention; hospital; mixed methods; patient; learning process; assessment; autonomy; impairment; real-time motion; university; geriatric ward; survey; anxiety; willingness; patient privacy; effectiveness; monitoring system; health care practice

Introduction

Background

The global increase in emergency hospital admissions of older individuals is expected to continue, driven by ongoing demographic shifts. Currently, older adults make up approximately two-thirds of hospital inpatients, with up to 50% of this population experiencing some degree of cognitive impairment, including dementia-related conditions [1-3]. Patients with cognitive impairment (hereafter referred to as patients) often struggle to adhere to care plans in the hospital setting, and their day or night cycles are frequently disrupted. This presents additional challenges for nurses, as their workload increases, especially since hospitals are often not adequately equipped to care for patients with cognitive impairment [4]. A particular challenge for hospital nurses is motor agitation in and around the patient's bed, which can lead to incidents such as falls or accidental removal of catheters or vascular access devices [5]. In light of these challenges, there is ongoing discussion about the potential role of technological innovations in supporting nurses and caregivers [6,7].

Bed exit information systems (BES) are well-known in this context [8,9]. In a previous study by the research team [10], the focus was widened to the complex interplay of factors from the nurses' perspective. The study addressed the perception of nurses on several aspects as adverse events and the implementation in the clinical setting, as well as their assessment of usefulness. Therefore, nurses on regular wards at a university medical center used a bed exit system to better understand its potential in hospitals [10]. The system, which was placed under the mattress in the patient's bed, informed nurses via the nurse call system whether patients were mobilizing to the edge of the bed or leaving the bed unassisted. An advantage of this system was its ease of use, which provided benefits to the nursing staff. However, due to the complex and sometimes unpredictable course of symptoms, it was only possible to use it in a targeted way for a limited number of patients (patients with cognitive impairments who also have a tendency to walk and a risk of falls or may become disoriented). That technological limitations and special characteristics strongly affect the effect of a BES becomes also evident in the study of Considine et al [11], in which alarms at the bedside caused specific challenges for the nurses, including issues related to alarm fatigue, disruptions in patient care, and difficulties in prioritizing responses to alarms.

Therefore, the research team looked for another technology with a higher degree of functionality (allowing for the setting of either bed edge or bed exit notifications) and an additional function (real-time motion monitoring), with the aim of evaluating its implementation. Here as well, the research question focused on possible benefits for the nurses in patient care and the integration of the technology in the working routines. To answer this question, the nursing staff of a geriatric ward in a university hospital used the Mobility Monitor, an in-bed real-time motion monitoring system (IRMS) from the company Compliant Concept AG, as an additional aid in the standard care of patients [12]. In addition to bed edge and bed exit information, the system provides nurses with real-time monitoring to observe and analyze the patient's movement patterns in bed. In a previous evaluation project [13], the IRMS was found to have potential benefits for the nightly monitoring of patients on standard wards in general and the care of patients with delirium in particular.

Overall, previous studies have predominantly addressed single aspects of the technology, such as the efficacy and effectiveness of BES in terms of reducing falls or fall rates [14,15], and there has been limited research into nurses' perspectives and experiences of using BES [16,17].

Aim and Research Question

Guided by the results of previous studies, the purpose of this study was to better understand the positive effects and disadvantages of the use of an IRMS from a nursing point of view.

The study was designed to address the following research questions:

- How does the use of IRMS affect nurses' perception of physical and mental stress?
- How is the care process affected by the use of the IRMS?
- What kind of ethical implications could arise from the nursing perspective, such as perceptions of autonomy and privacy for patients?
- How do nurses acquire knowledge about the technology, and how does this learning process influence their assessment of it?

Methods

Reporting was guided by the Good Reporting of a Mixed Method Study (GRAMMS) criteria proposed by O'Cathain et al [18].

Design

This study was designed using a parallel mixed methods triangulation approach [19], incorporating a quantitative survey, qualitative focus groups, and one interview, as well as participant observations to obtain a more comprehensive understanding from multiple perspectives (Figure 1). A key principle of mixed methods research is that combining quantitative and qualitative data provides a richer understanding of the research question compared to using either data source in isolation [20]. The integration of results is realized in the Discussion section of this paper.

XSL•FO
RenderX

Figure 1. Parallel mixed methods triangulation study design.

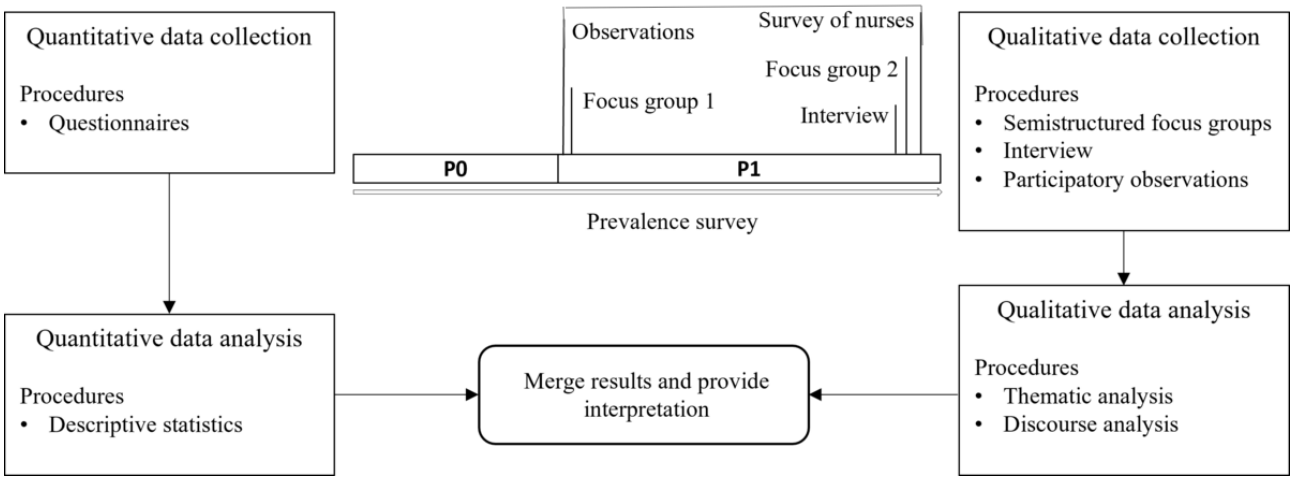

The study was conducted in a geriatric ward at a university medical center in the south of Germany. The ward has 20 beds with an average occupancy of 18 patients at any one time. Staffing typically includes three registered nurses on the day shift and usually three on the evening shift, but occasionally two registered nurses and two nursing assistants, depending on availability. During the night shift, there are usually two staff members on duty. Three physician assistants and one supervising physician are responsible for the ward.

Spanning 24 weeks, the study period was divided into two distinct phases. Phase 0 (baseline survey P0) lasted 6 weeks. During this phase, patients were cared for according to the university medical center care standard. Phase 1 (intervention phase P1) lasted 18 weeks. For this intervention phase, the geriatric ward was equipped with the IRMS, and data about the experience of nurses were collected. The IRMS were available to nurses as an aid to patient care, assisting nurses in assessing the patients' cognitive status and tendency to get out of bed. Nurses were able to make their own decisions about which patients they would use the IRMS with. Both the (1) real-time monitoring of the patient's movement patterns in bed and (2) bed edge and bed exit information could be turned on or off by the nurses.

During the study, 22 devices were made available to the ward from the first day of the intervention phase (Phase 1), ensuring full coverage. The research team integrated the devices into the beds on that day, which also marked the first training session, with a second session conducted the following week. Preparatory work, including software installation and setup of the screens in the nursing station, was carried out in the weeks prior (Phase 0). The devices were customized by the supplier (Compliant Concept) in coordination with the University Medical Center's technical team to ensure compatibility with the call system.

The training consisted of a 2-hour session, which was held twice by a company representative for the nursing staff. While the focus was primarily on nurses, other professionals working on the ward, such as physicians and physiotherapists, were also invited to participate. The training included a presentation for the IRMS introduction, as well as practical exercises to reinforce learning. Additionally, research team members provided further assistance as needed, ensuring ongoing support for any questions or challenges the nursing team faced. Informational materials were also available to the staff, including a PowerPoint (Microsoft Corp) presentation, a practical written guideline for patient entry into the software, and the company's manual. Regarding communication with patients about the IRMS, nurses were not specifically trained to discuss its use, but it ultimately depended on their professionalism in interacting with patients.

Participants

The survey and focus group sessions involved nurses who met the predetermined criteria outlined in Textbox 1. A total of 21 nurses met the inclusion criteria at P1. Potential participants were informed about the survey, focus groups, participant observations, and interview through ward meetings and internal institutional mailing lists, supplemented by email communications. They were then formally invited to participate. Nurses who expressed an interest were given full written information about the study and the nature of their participation. The survey, focus groups, participant observation, and interview were conducted during working hours with the authorization of the respective organization and no additional incentives or rewards were provided to participants.

XSL•FO
RenderX

Textbox 1. Inclusion and exclusion criteria.

Inclusion criteria

- ≥18 years
- Nurses with at least three years of apprenticeship or equivalent international training with professional recognition in Germany
- Employees working on the included ward during the study period

Exclusion criteria

- Employees of the included ward who belong to other professional groups.
- Employees of the reserve pool.
- Nurses from other wards helping out
- Trainees

Data Collection

The study followed a parallel mixed methods approach with the following elements (see Figure 1 for an overview of the study design):

- Survey of nurses at time point P1.
- Focus groups with nurses at time points P0 and P1, before and after IRMS use.
- Interview with one nurse at time point P1, after IRMS use.
- Participant observations during the initial introduction of the technology on the ward by the manufacturer.

Nurse Survey

The survey was developed in REDCap (Research Electronic Data Capture; Vanderbilt University) [21]. A comprehensive review of the literature did not identify a suitable assessment tool for collecting data on this topic in the German language. The web-based survey, which consisted of 48 items, was based on the technology attitude or acceptance, burden, and expectations or experiences survey used in a previous study [10]. It primarily used items from Spagnolli et al's [22] survey on user acceptance of wearable technologies, which were translated into German by our team, and a subset of Isfort et al's [23] survey on nurses' perceived burden in patient care, which was available in German. The instrument was further supplemented with custom questions and sociodemographic data, including age, gender, and years of professional experience of the nursing staff. However, due to the small sample size, these sociodemographic data were deemed insufficient for meaningful subgroup analysis or generalizable insights. Presenting these variables could risk overinterpretation or imply trends unsupported by the data. Therefore, we chose to focus on the main findings directly related to our research questions. In previous projects using paper and pencil surveys, it was found that the nursing team preferred a digital format due to its environmental benefits, such as the elimination of paper waste. A total of 21 nurses were invited to participate via email with a link to the web-based survey, allowing them to complete the survey at their convenience. The survey began immediately after the intervention phase, starting 11 days after the IRMS had been in use for 20 weeks. The survey is included in Multimedia Appendix 1.

Focus Groups and Interview

Each focus group was conducted during the first survey phase (P0) and at the end of the second survey phase (P1) by two researchers (IS and SW) to capture the context of technology use and to further explore the issues identified in the staff survey. The initial focus group guide addressed expectations regarding the potential for alleviation, usability, and the influence on social interaction with patients (Multimedia Appendix 2). In contrast, the guide for the discussion after the intervention centered on the experiences related to these same themes. Topics included the challenges of caring for patients with cognitive impairment, the tendency of patients to get out of bed unsupervised, and the expectations, experiences, and evaluations related to the use of the IRMS (Multimedia Appendix 3). This approach aimed to derive insights into nurses' perceptions of physical and mental stress, as well as any changes resulting from the implementation of the technology. Inferences about potential alterations in the care process, patient outcomes, and the perceptions of autonomy or privacy by patients can be deduced from nurses' comments on processes, perceptions, and evaluations before, during, and after the use of the IRMS.

The two researchers facilitating the discussions both have prior experience in conducting focus groups, with one having a background in nursing science (SW) and the other in sociology (IS). A semistructured interview guide was used, focusing on the concepts of interest while allowing for flexibility. The structure and content of the guide were informed by the core findings of previous studies [10,13] and the literature reviewed (see Background under Introduction section). Since one participant from the first focus group could not participate in the second, a follow-up interview was conducted using the same guideline at P1. Both the focus group and the interview were audio-recorded and transcribed using a denaturalized approach, omitting idiosyncratic speech elements [24]. Transcripts were pseudonymized prior to analysis.

Participant Observations

The analysis of the focus groups in a previous study [10] showed that the initial introduction of the technology on the ward by the manufacturer is an important factor for the further course of implementation. Therefore, this factor was investigated as a decisive condition for the implementation of IRMS. For this purpose, observations in the sense of a focused ethnography

XSL•FO
RenderX

[25] were carried out by two researchers (SW and IS) during the training sessions. The observations were open with the consent of the participants and the trainers. Field notes taken during the observations were collected, converted to a digital text format, and pseudonymized prior to analysis.

Analysis

Qualitative data (focus groups, participant observation, and interview) were analyzed by SW, IS, and NM. We systematically organized and structured the data according to the principles of Kuckartz's content analysis approach [26]. The main categories were shaped by the guiding questions and supplemented with inductive aspects derived from analysis discussions. Additionally, the investigation of specific aspects of the discussion dynamics in the focus groups was inspired by elements of the documentary method, particularly as outlined by Bohnsack [27] and Bohnsack and Schäffer [28]. This methodological combination aimed to provide a robust foundation for analyzing both the explicit content structure and the underlying social dynamics and interpretations of meaning, thereby promoting a comprehensive understanding of the phenomena under study. Descriptive quantitative data analysis by SW, IS, and HM was performed using SPSS Statistics (IBM Corp) for Windows. Frequencies and percentages were calculated for all quantitative variables.

Setting

The study was conducted at the University Medical Center Freiburg.

Ethical Considerations

The conceptualization and implementation of this study were based on the principles of the Declaration of Helsinki. By German law, survey studies with a focus on employees must be approved by the Employee Council at the respective institution. The responsible committee of the Employee Council of the University Medical Center Freiburg provided approval for this study in written format (approval 106/20 [MPG §23b]). Participation in the focus groups was voluntary, no personal data were collected, and anonymity was always maintained. All potential participants received written information on the study (reason for the study, objective, processes, and data protection), were informed about the decision of the Employee Council and its subcommittees, and had the opportunity to contact the investigators in case of questions at any time during the study. Informed consent to participate was assumed if individuals completed the survey and were confirmed (by ticking a box) at the beginning of the survey.

Results

Participant Observations

The two initial training sessions were diverging. There were 12 participants in total—10 nurses and 2 physiotherapists. They took place in two very different places. The first one was in a lecture hall where the participants were spread around the room. They seemed quite tired after the morning shift but were eager to understand the technology. The amount of information seemed rather overwhelming in this setting. However, the nurses seemed to be able to follow very well. The second introduction took place on the ward with two participants from the physiotherapy profession. The atmosphere was livelier and more optimistic regarding the use of the technology. There was an interprofessional exchange, and strategies for using the IRMS were discussed. Once again, the participants' questions demonstrated their engagement and ability to quickly grasp the basic concepts of the technology. The presenter began the session by mentioning the number of miles nurses walk every day. However, this topic was not explored further in the discussion, and we conclude that the participants did not prioritize it as a goal for using the technology. Additionally, the presenter emphasized the expertise of the nurses in making decisions about the use of the IRMS. The presentation of case scenarios primarily focused on night-time use. Although preventing falls and other adverse events was framed as a key interest of the research team, the representative highlighted both fall prevention and motion analysis, with a particular focus on the latter. One participant's question suggested that targeting specific patients might be a promising strategy.

Survey

At the time of the survey, participants had been using the IRMS on the ward for 20 weeks and had gained initial experience of its functionality. A total of 11 of a possible 21 individuals participated in the survey, 10 of which were completed (10/21, 48%). Nurses' responses to the survey confirmed that caring for patients with cognitive impairment is a significant burden (items 8-19 in Multimedia Appendix 1). Notably, participants demonstrated a general openness and minimal fear of technology (items 1-7 in Multimedia Appendix 1).

When asked about potential improvements from the IRMS, most respondents were optimistic and agreed that it would improve outcomes in several areas. The only areas where they did not agree were preventing the removal of peripheral venous catheters and drains, improving privacy, and disturbing other patients (items 23, 25, and 26 in Multimedia Appendix 1). However, nurses also see many opportunities to improve care and work processes through the use of technology, including feeling safer due to the use of the IRMS, receiving support in care planning, and assisting in coordinating their work schedule (items 28-30 in Multimedia Appendix 1). The impact on patient safety is also seen as positive (item 20 in Multimedia Appendix 1). The survey responses from the nurses indicate a clearly positive view regarding the IRMS's effectiveness in preventing falls and preventing patients from leaving the ward unnoticed (Figure 2).

In terms of organizational aspects, nurses found the IRMS to be a valuable tool for care planning and time management (Figure 2). In addition, 80% (n=8) of respondents (statements from "tend to agree" to "fully agree") confirmed that the technology enabled them to assess the quality of their work (Figure 2). For nurses to find the IRMS helpful, they must find it reliable—which they consistently confirmed in the survey, both for bed edge and bed exit information and for motion monitoring (items 40-42 in Multimedia Appendix 1). In terms of usability, respondents generally agreed that using the technology was not annoying and did not increase the

XSL•FO
RenderX

documentation burden. However, the initial setup was considered by some to be labor-intensive. Opinions varied between respondents, possibly influenced by the actual involvement in the setup process. (items 32, 33, and 37 in Multimedia Appendix 1). Overall, nurses found it easy to learn how to use the device (Figure 2).

Responses regarding the potential reduction of work-related movements, such as checking on patients at risk of falling, were mixed, with differing opinions between participants. Further insight into this variation can be gained by analyzing the focus groups, as described below.

Ethical considerations were explored by assessing whether nurses felt that the IRMS helped to maintain patient privacy. In general, respondents tended to think that it did not, a nuance that may be clarified by analysis of the focus groups. When asked if they felt monitored by the technology, 8 out of 10 nurses said that this was not the case. Almost all of the nurses surveyed (n=9; statements from agree to fully agree) indicated that they would continue to use the IRMS in the future.

Figure 2. Survey results on the effectiveness and organizational aspects of the in-bed real-time motion monitoring system (IRMS).

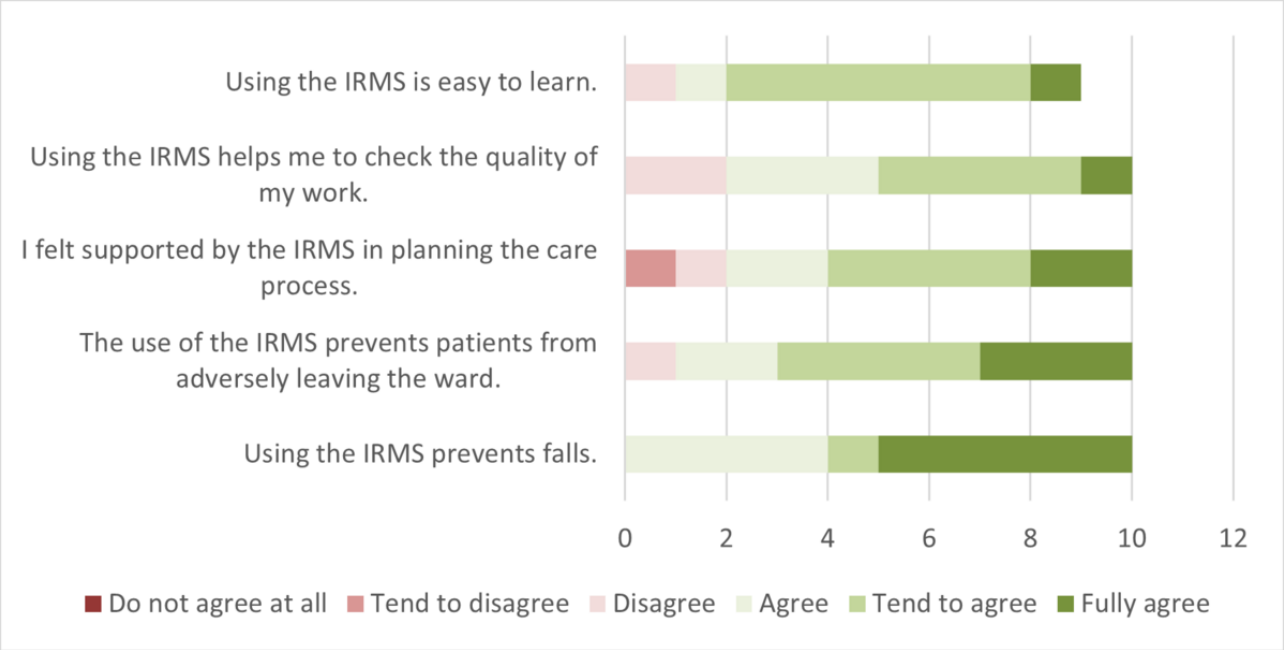

Focus Groups and Interview

Overview

The focus groups and the interview largely confirm the results of the survey and can shed more light on some details. Each focus group had a total of three participants, with one person attending both sessions. The issues of benefits, privacy concerns, and usability were discussed more controversially than the survey suggests. This is illustrated by the following analyses. A summary of the themes, including their descriptions and primary insights, is provided in Table 1. The originals of the quotes used can be found in Multimedia Appendix 4.

Table 1. Summary of themes and key insights from focus groups and the interview.

Theme	Description	Key insights
Benefits	Discussion on the benefits of the IRMS[a], particularly regarding fall prevention and night shift management.	The IRMS was seen as helpful in improving patient safety and reducing anxiety through quick responses. It was especially valuable during night shifts to manage patient movements.
Privacy concerns	Concerns regarding patient privacy in relation to monitoring through the IRMS.	Some patients felt monitored or “controlled.” However, this was considered an exception. A clear explanation of the system helped alleviate some concerns.
Usability	Experiences of nurses with the usability and reliability of the system.	The IRMS was considered increasingly reliable. Suggestions for improvement mainly focused on integration with the documentation system and reducing “false alarms.”
Overall assessment	Overall assessment of the system and nurses’ willingness to continue using the IRMS.	Nurses found the system helpful, but not essential. They emphasized that care could still be provided without it, though some additional visual checks would be necessary.

[a]IRMS: in-bed real-time motion monitoring system.

XSL•FO
RenderX

Benefits

Nurses participating in both focus groups reported that they found the technology helpful in preventing or responding to falls. However, reports of specific responses to information about bed exit or bed edge information remained largely vague despite repeated questioning. The meaning of the "usefulness" of the IRMS becomes clearer when considering that the additional information allows nurses to better plan their next steps, potentially allowing them to respond more quickly to falls.

> *And when it rings, we know, this woman is sitting at the edge of the bed. Then we know...We have ideas. Then we know how we have to organize the process or how to do the plan for the patient.* [Focus Group 1, P3]

In addition, the absence of narratives about specific fall-prevention situations emphasizes that it was not the prevention of falls themselves that played a crucial and relieving role, but rather a sense of safety in relation to falls. Simply knowing that they would be alerted if the patient got up allowed nurses to perform their other activities with less anxiety.

> *Well. Especially regarding falls there was a certain safety, because the notification actually always sets off and you don't have to permanently do visual controls and go to the rooms. This certainly creates a feeling of safety, to know "Okay, the patient is not in danger of falls or similar right now." Well and for coverage of the positioning and in order to see: "Okay, was there sufficient positioning? Have I supported the patient sufficiently?...it gives a certain feeling of safety when you can check this on the monitor."* [Focus Group 2, P3]

In addition to an increased sense of security, another clearly identified benefit is improved information about patient self-movement. This is particularly important at night, when patients may need to be repositioned at defined intervals according to nursing standards. Right at the beginning of the interview, the IRMS is described as a technology that goes beyond and is superior to the existing technical possibilities (eg, a mat on the floor in front of the bed that sends a signal to the call system when touched by the patient) and is superior to them. According to the interviewee, the technology is particularly valuable during night shifts because it constantly monitors whether patients are moving or not. Previously, patients had to be awakened and repositioned every 2 to 3 hours to relieve pressure.

> *In nightshifts [with the IRMS] you immediately notice: Do people move, don't they move. And before, okay you walk there every 2 to 3 hours, turn the people from left to right, have to wake them up and this is bad, when people are fast asleep and you see: Oh, he's moving, moving sufficiently, perfect. AND, it's very good that when somebody really has developed a decubitus here, we can prove: Okay, every 2 hours, the monitor has recorded it and there were immediate reactions, there were positionings and so we can demonstrate this.* [Interview]

The IRMS provides an indication of whether there has been sufficient spontaneous movement, allowing nurses to avoid unnecessary repositioning. This helps prevent unnecessary disruption to the patient's sleep and supports nurses in re-evaluating their own judgments about the need for repositioning. If a patient develops a pressure ulcer, the recorded data can be used by nursing staff to demonstrate that regular pressure-relieving positioning has been carried out correctly (interview and focus groups). The interviewee also mentioned the recording of bed exits, especially for patients with hyperdelirium. The ability to detect physical tension in patients through specific movement patterns is highlighted as another, initially unintended, "function" of the device.

Privacy Concerns

The issue of privacy was not directly raised in the focus group discussions. However, the analysis provides insights into the different responses in the survey. The difference may be due to different perceptions or reference points of "privacy." In the second focus group discussion, it was reported that with the use of the IRMS, room doors are occasionally closed more often, as monitoring can take place without direct visual contact. However, according to nurses' reports, some patients occasionally expressed feeling monitored and controlled. For example, the interviewee reported a situation where the technology was seen by a patient as a "spy in the bed."

> *This one patient, who always said: "Ha, the spy betrayed me again, actually I wanted to try to go to the toilet by myself." There you stand next to them and say: "You shouldn't walk alone; you are still quite weak-kneed with the rollator."* [Interview]

Due to the design of the study, it remains uncertain how patients experience these reactions, whether positively or negatively. It is worth noting that the term "spying" typically carries a negative connotation in German. This anecdote, involving a patient at risk of falling who was attempting to use the toilet independently, illustrates the effectiveness of the IRMS in preventing potentially dangerous situations. The system alerted the nursing staff, enabling them to assist the patient immediately. The nurses perceive their immediate presence when the system is activated as a positive aspect, as it allows for a quick response to the patient's needs. However, the extent to which patients are comfortable with this form of monitoring seems to be a critical factor.

> *P3: I think they were not really aware, that there is a special mat in their bed. They didn't really notice it, most of them. The others quickly mentioned it. But they didn't mind either, whether there was something like that or not. So.*
>
> *P1: Except for the ones who ranted, because they were worried, well...one once even said: "Am I being controlled?"*
>
> *P3: Okay no, I've never come across that.*
>
> *P1: And then there were some, who worried [about] the space for their cell phone plug.* [Focus Group 2, P1 and P3]

XSL•FO
RenderX

This highlights an interesting ethical conflict. The nurses in the focus group at P1 also report patients' complaints or confusion about the sense of being controlled, but they regard these as special cases, similar to minor concerns like charging their mobile phones. However, the primary cause of confusion appears to be the patient's general state of confusion. Explaining the technology seemed effective in addressing this confusion. However, the nurses did not report exploring deeper complaints, such as the feeling of being controlled or monitored. This assessment by the nurses regarding patient privacy may correlate with their own perceptions of control associated with IRMS. In the first focus group at P0 and the interview (P1), some team members initially viewed the IRMS as a potential "control device," referring to the supervision of nurses and their activities, for example, by line managers. Although this perception dissipated relatively quickly as the IRMS became more established, it highlights nurses' concerns about potential surveillance, which were also discussed in the focus groups. After a longer period of use (according to reports from the second focus group), nurses no longer feared being monitored themselves. While in the first focus group, they were suspicious of the potential monitoring of their work performance, in the second focus group, they expressed hope that the IRMS would demonstrate that they were making correct judgments about patient positioning. The issue of surveillance was eventually dismissed as irrelevant.

Usability

With minor reservations, the nurses consider the system to be reliable, especially in the second focus group and the interview at P1, more so than at the beginning. They have no fundamental objections to the usability of the system, but they do have some suggestions for improvement. Most of these suggestions focus on improving the documentation process. For example, if there were an interface to the documentation system and a simplified or automated transfer of recorded data to mandatory documentation, the IRMS could streamline the documentation process overall. This could also help nurses to provide a plausible explanation for justified deviations from standards of care, thereby encouraging the use of their own professional knowledge and experience.

In addition to the survey's assessment that the technology is easy to learn, focus group participants and the interviewee emphasize that its use requires the development of routines. They point out that a lack of familiarity with the system can lead to disruptions in ward operations, particularly among other professional groups who do not work exclusively on the participating ward and have not developed routines with the IRMS. This often results in "false alerts" during the day, such as physiotherapists forgetting to turn off alerts before their interventions.

In line with the inconsistent assessment in the survey regarding the reduction of walking distances, different perspectives emerged in the focus group discussions. When using the IRMS, nurses do not have to repeatedly check on patients at risk of falling by visiting them. This is especially an improvement at night. However, this process rarely exists in isolation, as nurses often combine these checks with other tasks. The implementation of the system does not alter these interrelated activities, which continue as usual.

> *Those who are hyperactive at night, you pick that up anyways, because then you are continuously close to the bed, because they're ringing all the time, calling or doing anything else.* [Focus Group 2, P1]

Overall Assessment

When asked whether they would like to continue using the technology, the nurses in the second focus group were slightly more ambivalent than in the survey. On the one hand, they emphasize that care can be provided without such technology, but on the other hand, they generally find the IRMS helpful.

> *Moderator: And if there was no IRMS starting from tomorrow on. How would you like that?*
>
> *Person 3: Well...*
>
> *Several: [laughing]*
>
> *Person 2: I would miss it. I would like to*
>
> *Person 3: Yes, it would be missed. But the work still could be done. That's not the way it is.*
>
> *Person 2: Well maybe a thing or two would be more complicated again. Because we would have to do more visual controls at the patients' beds, for patients who could wander or patients with dementia.*
>
> *Person 3: Yes.*
>
> *Person 1: Exactly, probably one fall or two would happen* [2.5 seconds]
>
> *Person 3: but those could not be prevented with IRMS, either.*
>
> *Person 1: Yes, sure! We would certainly go back to text-book, which means positioning according to the clock, but-*
>
> *Person 2: This would not make our work impossible*
>
> *Person 1: (laughing) Exactly. Nursing can do that* [Focus Group 2, P1, P2, and P3]

Overall, the nurses expressed a willingness to continue using the IRMS but highlighted that care is not entirely dependent on such a technology. In the following discussion, possible backgrounds for this evaluation will be examined further.

Discussion

Principal Findings

The two initial training sessions showed differences in both effectiveness and atmosphere. The second introductory session, held on the ward with additional therapeutic staff, fostered a much livelier and more optimistic atmosphere toward the technology. Interprofessional exchange was seamless, with active discussion of strategies for implementation and use of the IRMS. These observations are consistent with the research of Koukourikos et al [29], which emphasizes the effectiveness of hands-on training in familiar environments over formal settings. Moreover, Kahn et al [30] highlight the importance of interprofessional collaboration from the outset of technology implementation in hospital care settings. Building on these

XSL•FO
RenderX

findings, a recent study using a design-based research approach highlights the importance of structured training and support programs to help nurses effectively integrate innovative technologies into practice [31].

The survey, focus groups, and the interview showed that nurses benefit most from motion monitoring, especially at night. The increased sense of security can reduce psychological distress. This finding is consistent with the systematic review by Mileski et al [32], which included 28 studies. Whether the technology is effective in preventing falls could not be determined as this study was not designed to measure its effectiveness and the nurses' statements in this regard remain ambiguous. However, the technology does appear to reduce the nurses' fear of falls and make them feel better equipped to respond in a timely and appropriate manner should a fall occur.

Interestingly, the focus on fall prevention was primarily driven by the research question, but it emerged that other benefits were more relevant. One reason for the limited focus on fall prevention might be that nurses' responses to the information provided by the system were constrained by existing routines. Since there was no significant adaptation of processes beyond the introduction of the technology, nurses could not change their workflows significantly. More comprehensive changes—such as ongoing training programs and organizational policies supporting flexibility in nursing practices—would be needed for nurses to react more promptly when desired. Future studies would benefit from exploring how such ongoing support influences technology adoption and the potential long-term impact on nursing workflows.

The literature also suggests that alarms, particularly when they are frequent or audible to patients, may not be effective as standalone interventions [33-35]. A high number of alerts or alarms (especially if patients can hear them) may even have adverse effects, as suggested by the study by Considine et al [11]. Excessive alarms can increase anxiety and stress levels, disturb sleep, and desensitize both patients and staff, leading to delayed responses to critical situations. This may explain why the system was primarily used at night when nurses could respond more directly to the system's alerts. Goals such as reducing physical strain by reducing the need to walk to check on patients are negated because such rounds often include other essential nursing activities. For example, during these rounds, nurses are often checking on multiple patients, assessing their comfort, administering medication, or performing other routine tasks. This underlines that technology can only reach its full potential if it is integrated into comprehensive, targeted changes and is part of a broader, well-structured care plan [32].

Another factor contributing to the mixed evaluation of the IRMS could be debates about the professionalization of nursing in Germany [36]. Notably, in the second focus group, it was emphasized that nursing care could still be provided effectively without the use of technology.

Concerns about surveillance disappeared over time. In the survey and the second focus group, nurses no longer expressed privacy concerns and even minimized patients' worries. These findings are consistent with a previous study on a BES used across several wards at the University Medical Center [10]. On one hand, it could be argued that patient privacy is actually enhanced compared to personal observation, as physical movements can be monitored discreetly, without patients' awareness. However, patients still voiced concerns about being monitored. While the focus group participants did not consider this a serious issue, it remains essential to address patient privacy. Nurses could explore ways to make the technology more understandable, ensuring that patients are informed and that their consent is respected. Situations may arise where a patient's rejection of the technology—due to significant stress or a violation of personal values—should lead to discontinuation of the system's use. Future research could explore effective communication strategies to address privacy concerns and promote patient acceptance of monitoring technologies.

Finally, integration with documentation standards and systems remains a major concern contributing to nurses' workload, as discussed in the reviews by Mileski et al [32] and in studies on the implementation of digital nursing technologies [37-39]. Considering these aspects in the development of new technologies could significantly enhance the benefits of nursing. By supporting conditions that allow nurses to work in a self-directed and knowledge-based way, both their empowerment and professional standing could be strengthened.

Limitations

When discussing the use of the IRMS, it is important to clarify that the study's aim was not to assess its effectiveness or similar outcomes but rather to explore nurses' perceptions of its usefulness.

One limitation of the study is the small sample size. Ideally, multiple focus group discussions would have been conducted and analyzed at each time point to identify specific discursive patterns through comparison. Unfortunately, this was not feasible due to a limited number of eligible participants and a lack of volunteers. For practical reasons, an interview was conducted instead of a focus group. While an interview does not capture the same level of social consensus on issues as a focus group, the results were notably similar. The survey achieved a response rate representing 48% of the nursing staff in the geriatric ward. Although the sample size was relatively small, it provided valuable insights into the perceptions of nearly half of the relevant nursing staff. Given the limited number of participants, it is possible that the findings could be influenced by selection bias, especially if the individuals who participated were more likely to be early adopters or held leadership positions within their wards. Early adopters, including those in leadership roles, may exhibit a more positive perception of new technologies, potentially leading to an overrepresentation of optimistic perspectives in our results.

However, it is important to note that this study does not aim to present representative findings. Instead, it seeks to provide an in-depth insight into the implementation and use of the technology within a specific context. In qualitative research, some degree of bias is inherent, as findings are inevitably tied to the specific social situations and perspectives of the participants involved.

XSL•FO
RenderX

Conclusions

Based on the findings, the participating nurses have a positive attitude toward the use of a real-time in-bed movement monitoring system, highlighting its benefits for nurses and patients, particularly during nighttime monitoring. While its effectiveness in preventing falls remains inconclusive, the system appears to reduce nurses' fear of falls and improve their ability to respond effectively. The results highlight the wider impact of the system beyond fall prevention and underline the importance of carefully integrating such technologies into health care practice. The study acknowledges the challenges of embedding technology into care, including the limitations of nurses' responses to the IRMS information and ongoing discussions about professional dynamics in Germany. It also identifies opportunities for future research and development, particularly in relation to integration with documentation systems. Overall, the study argues for a deliberate, holistic approach to health care technology that considers professional relationships, patient needs, and system improvements.

Acknowledgments

We thank all the participants for their time and willingness to share their experiences with us. This study was conducted at the Center of Implementing Nursing Care Innovations, Freiburg, Germany. This joint project on the development and research of new care technologies in Germany is funded by the Federal Ministry of Education and Research (grants 16SV7889, 16SV7888, and 16SV7886K).

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

The presentation of the authors' contributions followed the recommendations of the Contributor Roles Taxonomy (CRediT) [40]. SW contributed to the conceptualization and methodology of the study, conducted the investigation, performed formal analysis, curated the data, and wrote the original draft of the manuscript. IS was involved in the conceptualization and methodology of the research, carried out the investigation, conducted formal analysis, curated the data, and contributed to the writing of the original draft. JP performed the formal analysis, curated the data, and contributed to the writing and editing of the manuscript. NM assisted with formal analysis and data curation. HM was involved in the conceptualization and methodology. SZ contributed to the conceptualization and methodology. CK contributed to the conceptualization, supervision, resource review, and editing. All authors approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Survey of the nursing staff.
[DOCX File , 20 KB - formative_v9i1e63572_app1.docx]

Multimedia Appendix 2
Focus group guide I.
[DOCX File , 29 KB - formative_v9i1e63572_app2.docx]

Multimedia Appendix 3
Focus group guide II.
[DOCX File , 31 KB - formative_v9i1e63572_app3.docx]

Multimedia Appendix 4
Translated quotes.
[DOCX File , 23 KB - formative_v9i1e63572_app4.docx]

References

1. Mudge AM, Lee-Steere K, Treleaven E, Cahill M, Finnigan S, McRae P. Cognitive impairment in older hospital inpatients: prevalence, care needs and carer perceptions. Aust Health Rev 2021;46(2):244-250. [doi: 10.1071/ah20286]
2. Jackson TA, MacLullich AMJ, Gladman JRF, Lord JM, Sheehan B. Undiagnosed long-term cognitive impairment in acutely hospitalised older medical patients with delirium: a prospective cohort study. Age Ageing 2016;45(4):493-499. [doi: 10.1093/ageing/afw064] [Medline: 27076525]

XSL•FO
RenderX

3. George J, Long S, Vincent C. How can we keep patients with dementia safe in our acute hospitals? A review of challenges and solutions. J R Soc Med 2013;106(9):355-361 [FREE Full text] [doi: 10.1177/0141076813476497] [Medline: 23759885]
4. Fogg C, Meredith P, Bridges J, Gould GP, Griffiths P. The relationship between cognitive impairment, mortality and discharge characteristics in a large cohort of older adults with unscheduled admissions to an acute hospital: a retrospective observational study. Age Ageing 2017;46(5):794-801 [FREE Full text] [doi: 10.1093/ageing/afx022] [Medline: 28338808]
5. Choi SSW, Cajita MI, Gitlin LN. A review of measures of three common dementia-related behaviors: rejection of care, aggression, and agitation. Geriatr Nurs 2020;41(6):692-708. [doi: 10.1016/j.gerinurse.2020.04.003] [Medline: 32402574]
6. Husebo BS, Heintz HL, Berge LI, Owoyemi P, Rahman AT, Vahia IV. Sensing technology to monitor behavioral and psychological symptoms and to assess treatment response in people with dementia. a systematic review. Front Pharmacol 2019;10:1699 [FREE Full text] [doi: 10.3389/fphar.2019.01699] [Medline: 32116687]
7. Astell AJ, Bouranis N, Hoey J, Lindauer A, Mihailidis A, Nugent C, et al. Technology and dementia: the future is now. Dement Geriatr Cogn Disord 2019;47(3):131-139 [FREE Full text] [doi: 10.1159/000497800] [Medline: 31247624]
8. Oh-Park M, Doan T, Dohle C, Vermiglio-Kohn V, Abdou A. Technology utilization in fall prevention. Am J Phys Med Rehabil 2021;100(1):92-99. [doi: 10.1097/PHM.0000000000001554] [Medline: 32740053]
9. Komagata H, Hattori T, Ohshima R, Kakinuma E, Ishikawa M, Shinoda K, et al. Development of human behavior monitoring system around bed using infrared depth sensor. : IEEE; 2019 Presented at: IEEE 1st Global Conference on Life Sciences and Technologies (LifeTech); March 12-14, 2019; Japan. [doi: 10.1109/lifetech.2019.8884068]
10. Walzer S, Schön I, Pfeil J, Klemm S, Ziegler S, Schmoor C, et al. Nurses' perspectives and experiences of using a bed-exit information system in an acute hospital setting: mixed methods study. JMIR Form Res 2025 Feb 05;9:e64444 [FREE Full text] [doi: 10.2196/64444] [Medline: 39908092]
11. Considine J, Berry D, Mullen M, Chisango E, Webb-St Mart M, Michell P, et al. Nurses' experiences of using falls alarms in subacute care: a qualitative study. PLoS One 2023;18(6):e0287537 [FREE Full text] [doi: 10.1371/journal.pone.0287537] [Medline: 37347774]
12. Compliant Concept AG 2023. URL: https://www.compliant-concept.ch/produkte/spital/#produkt-mobility-monitor [accessed 2025-01-28]
13. Ziegler S, Schmoor C, Schöler LM, Schepputat S, Takem E, Grotejohann B, et al. Potential for reducing immobility times of a mobility monitor in-bed sensor system—a stepped-wedge cluster-randomised trial. BMC Nurs 2023;22(1):478 [FREE Full text] [doi: 10.1186/s12912-023-01658-2] [Medline: 38104112]
14. Røyset B, Talseth-Palmer BA, Lydersen S, Farup PG. Effects of a fall prevention program in elderly: a pragmatic observational study in two orthopedic departments. Clin Interv Aging 2019;14:145-154 [FREE Full text] [doi: 10.2147/CIA.S191832] [Medline: 30697039]
15. Cuttler SJ, Barr-Walker J, Cuttler L. Reducing medical-surgical inpatient falls and injuries with videos, icons and alarms. BMJ Open Qual 2017;6(2):e000119 [FREE Full text] [doi: 10.1136/bmjoq-2017-000119] [Medline: 29450285]
16. Hubbartt B, Davis SG, Kautz DD. Nurses' experiences with bed exit alarms may lead to ambivalence about their effectiveness. Rehabil Nurs 2011;36(5):196-199. [doi: 10.1002/j.2048-7940.2011.tb00195.x] [Medline: 21882797]
17. Shee AW, Phillips B, Hill K, Dodd K. Feasibility, acceptability, and effectiveness of an electronic sensor bed/chair alarm in reducing falls in patients with cognitive impairment in a subacute ward. J Nurs Care Qual 2014;29(3):253-262. [doi: 10.1097/NCQ.0000000000000054] [Medline: 24487696]
18. O'Cathain A, Murphy E, Nicholl J. The quality of mixed methods studies in health services research. J Health Serv Res Policy 2008;13(2):92-98. [doi: 10.1258/jhsrp.2007.007074] [Medline: 18416914]
19. Ingham-Broomfield R. A nurses' guide to mixed methods research. Aust J Adv Nurs 2016;33(4):46-52. [doi: 10.37464/2016.334.1559]
20. Creswell JW. A Concise Introduction to Mixed Methods Research. United States: SAGE publications; 2021.
21. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009;42(2):377-381 [FREE Full text] [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
22. Spagnolli A, Guardigli E, Orso V, Varotto A, Gamberini L. Measuring User Acceptance Of Wearable Symbiotic Devices: Validation Study Across Application Scenarios. Cham: Springer; 2014.
23. Isfort M, Klostermann J, Gehlen D, Siegling B. Eine bundesweite befragung von leitenden pflegekräften zur pflege und patientenversorgung von menschen mit demenz im krankenhaus. Köln: deutsches institut für angewandte pflegeforschung 2014 [Website in German]. Pflege-Thermometer. 2014. URL: https://www.dip.de/fileadmin/data/pdf/material/Pflege-Thermometer_2014.pdf [accessed 2014-07-26]
24. Azevedo V, Carvalho M, Costa F, Mesquita S, Soares J, Teixeira F, Maia. Interview transcription: conceptual issues, practical guidelines, and challenges. Rev Enf Ref 2017;IV Série(Nº14):159-168. [doi: 10.12707/riv17018]
25. Cruz EV, Higginbottom G. The use of focused ethnography in nursing research. Nurse Res 2013;20(4):36-43. [doi: 10.7748/nr2013.03.20.4.36.e305] [Medline: 23520711]
26. Kuckartz U. Qualitative text analysis: a systematic approach. In: Compendium for Early Career Researchers in Mathematics Education. Cham: Springer; 2019:181-197.

XSL•FO
RenderX

27. Bohnsack R. Documentary method. In: The SAGE Handbook of Qualitative Data Analysis. United States: Sage Publications; 2014:217-233.
28. Bohnsack R, Schäffer B. Exemplarische Textinterpretation: Diskursorganisation und dokumentarische Methode [Book in German]. New York: Springer; 2013.
29. Koukourikos K, Tsaloglidou A, Kourkouta L, Papathanasiou IV, Iliadis C, Fratzana A, et al. Simulation in clinical nursing education. Acta Inform Med 2021;29(1):15-20 [FREE Full text] [doi: 10.5455/aim.2021.29.15-20] [Medline: 34012208]
30. Khan AI, Barnsley J, Harris JK, Wodchis WP. Examining the extent and factors associated with interprofessional teamwork in primary care settings. J Interprof Care 2022;36(1):52-63 [FREE Full text] [doi: 10.1080/13561820.2021.1874896] [Medline: 33870838]
31. Walzer S, Barthel C, Pazouki R, Marx H, Ziegler S, Koenig P, et al. Teaching in the digital age-developing a support program for nursing education providers: design-based research. JMIR Form Res 2025 Jan 15;9:e66109 [FREE Full text] [doi: 10.2196/66109] [Medline: 39813674]
32. Mileski M, Brooks M, Topinka JB, Hamilton G, Land C, Mitchell T, et al. Alarming and/or alerting device effectiveness in reducing falls in long-term care (LTC) facilities? A systematic review. Healthcare 2019;7(1):51. [doi: 10.3390/healthcare7010051]
33. Barker AL, Morello RT, Ayton DR, Hill KD, Brand CA, Livingston PM, et al. Acceptability of the 6-PACK falls prevention program: a pre-implementation study in hospitals participating in a cluster randomized controlled trial. PLoS One 2017;12(2):e0172005 [FREE Full text] [doi: 10.1371/journal.pone.0172005] [Medline: 28199376]
34. Chu RZ. Preventing in-patient falls: the nurse's pivotal role. Nursing 2017;47(3):24-30. [doi: 10.1097/01.nurse.0000512872.83762.69]
35. Hamm J, Money AG, Atwal A, Paraskevopoulos I. Fall prevention intervention technologies: a conceptual framework and survey of the state of the art. J Biomed Inform 2016;59:319-345 [FREE Full text] [doi: 10.1016/j.jbi.2015.12.013] [Medline: 26773345]
36. Roth C, Wensing M, Breckner A, Mahler C, Krug K, Berger S. Keeping nurses in nursing: a qualitative study of German nurses' perceptions of push and pull factors to leave or stay in the profession. BMC Nurs 2022;21(1):48 [FREE Full text] [doi: 10.1186/s12912-022-00822-4] [Medline: 35193561]
37. Bail K, Gibson D, Acharya P, Blackburn J, Kaak V, Kozlovskaia M, et al. Using health information technology in residential aged care homes: an integrative review to identify service and quality outcomes. Int J Med Inform 2022;165:104824 [FREE Full text] [doi: 10.1016/j.ijmedinf.2022.104824] [Medline: 35792376]
38. Abell B, Naicker S, Rodwell D, Donovan T, Tariq A, Baysari M, et al. Identifying barriers and facilitators to successful implementation of computerized clinical decision support systems in hospitals: a NASSS framework-informed scoping review. Implement Sci 2023;18(1):32 [FREE Full text] [doi: 10.1186/s13012-023-01287-y] [Medline: 37495997]
39. Abdolkhani R, Petersen S, Walter R, Zhao L, Butler-Henderson K, Livesay K. The impact of digital health transformation driven by COVID-19 on nursing practice: systematic literature review. JMIR Nurs 2022;5(1):e40348 [FREE Full text] [doi: 10.2196/40348] [Medline: 35867838]
40. Brand A, Allen L, Altman M, Hlava M, Scott J. Beyond authorship: attribution, contribution, collaboration, and credit. Learned Publ 2015;28(2):151-155. [doi: 10.1087/20150211]

Abbreviations

BES: bed exit information system
GRAMMS: Good Reporting of a Mixed Method Study
IRMS: in-bed real-time motion monitoring system
REDCap: Research Electronic Data Capture

Edited by A Mavragani; submitted 26.06.24; peer-reviewed by H Hatamabadi, S Kling; comments to author 15.10.24; revised version received 28.11.24; accepted 19.01.25; published 04.03.25.

Please cite as:
Walzer S, Schön I, Pfeil J, Merz N, Marx H, Ziegler S, Kunze C
Experiences With an In-Bed Real-Time Motion Monitoring System on a Geriatric Ward: Mixed Methods Study
JMIR Form Res 2025;9:e63572
URL: https://formative.jmir.org/2025/1/e63572
doi:10.2196/63572
PMID:40053780

XSL•FO
RenderX

©Stefan Walzer, Isabel Schön, Johanna Pfeil, Nicola Merz, Helga Marx, Sven Ziegler, Christophe Kunze. Originally published in JMIR Formative Research (https://formative.jmir.org), 04.03.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Patient Adoption of Digital Use Cases in Family Medicine and a Nuanced Implementation Approach for Family Doctors: Quantitative Web-Based Survey Study

Julian Beerbaum[1], MSc; Sibylle Robens[2], PhD; Leonard Fehring[3,4], MD; Achim Mortsiefer[5], Prof Dr; Sven Meister[1,6], Prof Dr

[1]Health Care Informatics, Faculty of Health, School of Medicine, Witten/Herdecke University, Witten, Germany

[2]Department of Psychology and Psychotherapy, Witten/Herdecke University, Witten, Germany

[3]Department of Gastroenterology, Helios University Hospital Wuppertal, Witten/Herdecke University, Wuppertal, Germany

[4]Faculty of Health, School of Medicine, Witten/Herdecke University, Witten, Germany

[5]General Practice II and Patient-Centeredness in Primary Care, Institute of General Practice and Primary Care, Faculty of Health, School of Medicine, Witten/Herdecke University, Witten, Germany

[6]Department Healthcare, Fraunhofer Institute for Software and Systems Engineering, Dortmund, Germany

Corresponding Author:
Sven Meister, Prof Dr
Health Care Informatics, Faculty of Health
School of Medicine
Witten/Herdecke University
Pferdebachstr. 11
Witten, 58448
Germany
Phone: 49 230292678629
Email: sven.meister@uni-wh.de

Abstract

Background: Digital use cases describe the application of technology to achieve specific outcomes. Several studies in health care have examined patients' overall attitudes toward digitalization and specific use cases. However, these studies have failed to provide a comparison of patient acceptance criteria between inherently different digital use cases in family medicine.

Objective: To address this research gap, this paper aimed to assist family doctors in selecting digital use cases by comparing the underlying patient adoption factors and in driving usage of these use cases by presenting a differentiated implementation approach.

Methods: Adapting an established Unified Theory of Acceptance and Use of Technology (UTAUT) questionnaire to 4 digital use cases in family medicine, we surveyed a large cross-sectional sample of adults living in Germany. The results of the web-based survey were then analyzed via descriptive statistics, ANOVA, and hierarchical regression models to compare the effects of sociodemographic and technology acceptance factors on the intention to use a specific use case.

Results: Our web-based survey included 1880 participants. Of these 1880 participants, only 304 (16.2%) agreed that the degree of digitalization is important when selecting a family practice. However, more digitally literate participants attributed greater importance to this criterion (B=0.226, SE 0.023; β=.223; P<.001), and digital literacy was found to be dependent on age (Welch $F_{3,968.29}$=53.441; P<.001). Regarding sociodemographic characteristics, only digital literacy demonstrated a significant effect on the intention to use for all use cases, particularly scheduling doctor appointments online (B=0.322, SE 0.033; β=.408; P<.001). Furthermore, *performance expectancy* was the strongest predictor of the intention to use for all use cases, while further effects of technology acceptance factors depended on the use case (receiving medical consultations via video: B=0.603, SE 0.049; β=.527; P<.001; scheduling doctor appointments online: B=0.566, SE 0.043; β=.513; P<.001; storing personal medical information via electronic health records: B=0.405, SE 0.047; β=.348; P<.001; and providing personal information before consultation digitally [digital anamnesis]: B=0.434, SE 0.048; β=.410; P<.001). To illustrate, *perceived privacy and security* had an effect on the intention to use electronic health records (B=0.284, SE 0.040; β=.243; P<.001) but no effect on the intention to use video consultations (B=0.068, SE 0.042; β=.053; P=.10).

XSL•FO
RenderX

Conclusions: In the selection and implementation of digital use cases, family doctors should always prioritize the perceived value of the digital use case for the patient, and further criteria might depend on the digital use case. Practice owners should therefore always harmonize the introduction of digital use cases with their own patient care strategies. Not every digital innovation fits every strategy and therefore every practice.

(JMIR Form Res 2025;9:e58867) doi:10.2196/58867

KEYWORDS

technology acceptance; UTAUT; family doctor; digital health; eHealth; video consultation; electronic health records; digital anamnesis; online appointment scheduling

Introduction

Background

The benefits of applying digital use cases, which we understand as the application of digital technologies to achieve a specific outcome, are well documented for the health care sector [1,2]. For example, in family medicine in Germany, where the family doctor is the first contact point to the health system and thus particularly important to ensure adequate treatment [3], digital use cases could address overarching challenges such as the shortage of family doctors [4]. After years of sporadic offering and use of digital tools in family medicine, COVID-19 accelerated their use [2,5]. Furthermore, this trend toward digitalization increased the diversity of digitally enabled use cases (eg, from online appointment systems to digital anamnesis tools and video consultation offerings) [5]. To illustrate this, the American Medical Association showed in 2022 in their longitudinal survey that the average number of digital applications used by a doctor increased by 72% between 2016 and 2022 [6].

It is apparent that, today more than ever, family doctors can decide which of the various offerings to implement. However, overall adoption rates of digital health care use cases by both patients and family doctors are still at low levels in some regions. For example, in Germany in 2023, only about 1% of all statutorily insured patients stored personal medical information via electronic health records [7,8]. A similar picture has emerged in other countries (eg, only about 3% of medical appointments in November 2023 in general practice in the United Kingdom were conducted via video) [9]. To stimulate broader adoption of digital use cases, legislative bodies in countries, such as the United Kingdom and Germany, revised their regulatory frameworks, such as directing the default use of electronic health records in Germany [10,11]. Nevertheless, to translate these incentives into actual patient usage of digital use cases, we argue that it is vital to assess the underlying factors determining patient adoption behavior. This evaluation is particularly important as family doctors consider patient expectations when deciding to offer patient-facing use cases [12,13].

In assessing technology acceptance, we premise that differentiation and comparison between use cases might offer fruitful insights, as we presume inherent structural differences between digital use cases (eg, degree of voluntariness of use for patients or degree of dealing with sensitive medical information). Thereby, we add to the academic discourse since there are several studies in health care examining patients' overall digital attitudes toward digitalization [14] as well as for specific use cases [15-18]. However, these studies have failed to provide a nuanced assessment and comparison of patient acceptance criteria between distinct digital use cases in family medicine. This differentiation is particularly vital as potentially diverse patient technology acceptance criteria per use case might emerge, and therefore, a differentiated approach to drive adoption might be needed.

Objectives

This survey-based study provides a nuanced comparison of patients' technology acceptance of key digital use cases in family medicine (ie, receiving medical consultations via video, scheduling doctor appointments online, storing personal medical information via electronic health records, and providing personal information before consultation digitally [digital anamnesis]).

Accordingly, this paper aims to support family doctors in selecting digital use cases by showcasing their relative importance to patients and in driving their use. In doing so, it highlights the need for a differentiated implementation approach based on different patient acceptance criteria for each digital use case.

For the scope of this research paper, we focused on family doctors as they act as gatekeepers for access to digital technologies [19]. Additionally, we hypothesized that family doctors could be important for driving adoption with an already established and trusted patient-doctor relationship.

Methods

Research Approach

Building on the insights of a rapid literature review on patient acceptance of 4 digital use cases, we leveraged an existing questionnaire assessing the acceptance of electronic health records in family medicine based on the Unified Theory of Acceptance and Use of Technology (UTAUT) model [20] and adapted it to the other use cases. The objective of our convenience sample approach was to gather a cross-sectional sample of adult patients living in Germany.

Digital Use Cases in Family Medicine

Family medicine typically provides the first medical contact within the health care system [21], and substantial evidence supports the application of digital technologies (digital use cases) in this setting [1,2]. Due to the high pace of innovation and great variety of digital use cases in family medicine, we limited the number of use cases in scope. The final selection of

XSL•FO
RenderX

digital use cases was a joint decision of a multi-professional team (comprising 4 researchers with diverse backgrounds in the digitalization of family medicine and 1 practicing family doctor). The decision was based on (1) the applicability of these use cases for patients in family medicine and (2) their current perceived practical relevance for both patients and family doctors. Thereby, we adhered to the following use case definitions:

- Receiving medical consultations via video: We followed the definition by Keuper et al [22] that "a video consultation is considered as a real-time visual and audio (digital) contact moment between the patient and health care provider." This excludes, for example, telehealth applications that do not facilitate a contact moment (eg, telemonitoring [23]).
- Scheduling doctor appointments online: We understand online appointment scheduling as the use case of booking an appointment via the internet in real-time [24].
- Storing personal medical information via electronic health records: We followed the definition of the National Coordinator for Health Information Technology of the United States that "EHRs are real-time, patient-centered records that make information available instantly and securely to authorized users" [25].
- Providing personal information before consultation digitally (digital anamnesis): Patient anamnesis refers to the data that patients provide about their personal medical history [26]. Digital anamnesis then applies this concept to the digital space as the information-gathering process occurs via a digital application.

Nevertheless, the full scale of adoption of the various digital use cases is yet to be realized [27]. Overall barriers can be pinpointed to a fragmentation of the digital health landscape, inadequate legislative frameworks, and limited clarity on information exchange, highlighting the necessity to consider the patients' perspectives even more [28].

Research Model

Our research attempts to compare different use cases in family medicine and highlight substantial differences between factors in patient technology acceptance.

Technology acceptance describes the positive decision of an individual toward using technology and explains the various factors and contexts incentivizing individuals to use a particular technology [29,30]. This range of factors in the decision-making process to use a tool in a health care setting is reflected in the various theories used to describe this technology acceptance. AlQudah et al [30] and Heinsch et al [31] demonstrated that the Technology Acceptance Model (TAM) and the UTAUT are the most applied theories in health care settings. Nevertheless, many studies combine different theories to explain technology acceptance with various factors [32,33].

The original UTAUT model consists of 4 constructs. The definitions of these UTAUT constructs by Venkatesh et al [34] are as follows. *Performance expectancy* is defined as "the degree to which an individual believes that using the system will help him or her to attain gains in job performance." In the family medicine setting, this translates to the usefulness of the use case to manage one's health or disease in a family practice setting. *Effort expectancy* is defined as "the degree of ease associated with the use of the system" [34], which, in our setting, relates to how easily patients can use the digital use case. *Social influence* is defined as "the degree to which an individual perceives that important others believe he or she should use the new system" [34]. In family medicine, these important others could range from physicians to public opinions to friends and family members. *Facilitating condition* is defined as "the degree to which an individual believes that an organizational and technical infrastructure exists to support use of the system" [34], which, in our context, is associated with having the right resources, compatible technologies, and knowledge to use the digital use case. The additionally included dimension in our study *perceived privacy and security* refers to "the degree to which patients believe that the respective use case is safe from intrusion and personal information is protected" [20,35]. We support the justification of Abd-Alrazaq et al [20] for applying the UTAUT model in the family medicine setting as the model has shown explanatory power in comparable contexts [31].

In line with our objective to assist family doctors by demonstrating the relative importance of digital use cases for patients and by highlighting different patient adoption factors, we tested 3 key hypotheses (hypothesis A, hypothesis B, and hypothesis C).

First, to illustrate the relative importance of each use case for patients, we hypothesized that awareness, availability, and actual usage behavior differ across use cases. For example, commercial studies have shown that 78% of surveyed patients wish for online appointment scheduling at their family doctor, while only 32% wish for an electronic health record and 27% wish for a video consultation [36].

Hypothesis A is as follows: Different use cases have differences in awareness, availability, and actual usage by patients.

Second, to determine the underlying adoption factors of digital use cases in family medicine, we formulated hypotheses B and C. Our rapid literature review, which aligned with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) extension for Rapid Reviews (PRISMA-RR) and recommendations by Tricco et al [37] and King et al [38], showed that different use cases have different factors driving technology acceptance [39,40]. The detailed approach and results can be found in the checklist in Multimedia Appendix 1.

Hypothesis B is as follows: Different use cases have different significant technology acceptance factors driving behavioral intention to use.

Various studies have indicated that sociodemographic factors affect technology acceptance. Mueller et al [4] argued that living in rural areas with potentially less doctor density might increase the likelihood of using digital offerings such as video consultation as it "saves the trip to the physician's practice." Thus, we hypothesized that the degree of the intended use of digital offerings varies (eg, by region and size of a city).

XSL•FO
RenderX

Hypothesis C is as follows: Different use cases have different sociodemographic characteristics impacting behavioral intention to use.

While hypothesis A pertains to indicated usage behavior, hypotheses B and C test the effect of a respective technology acceptance factor on the behavioral intention to use a specific digital use case in family medicine. The research model is depicted in Figure 1.

Figure 1. Research model. The effects of sociodemographic and technology acceptance factors on patients' intention to use a digital use case in family medicine in Germany in 2023 [20]. Dark blue indicates areas to collect evidence for hypothesis B, while gray indicates areas to test hypothesis C.

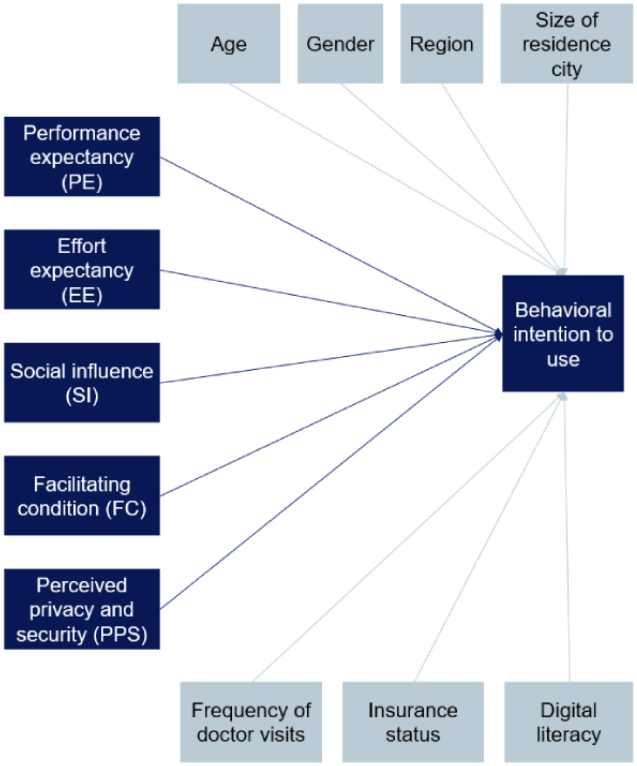

In these superordinate hypotheses, we replicated the same research model for each digital use case, with each testing the same subordinate hypotheses. Therefore, all subordinate hypotheses have the same components.

- Component 1 (illustrated in the research model): Respective technology acceptance factor [B]/respective sociodemographic characteristic [C] positively influences behavioral intention to use …
- Component 2: … a respective digital use case [online-booking platforms; digital anamnesis tools; video consultation platforms; electronic health records]

Survey Construction and Design

The questionnaire was constructed based on a rapid literature review on technology acceptance of digital use cases in family medicine (Multimedia Appendix 1) and our research model. Given its demonstrated validity and reliability in a family medicine setting (eg, average Cronbach α of .95; average "average variance contracted" [AVE] of 0.95; and average composite reliability [CR] value of 0.87), we leveraged an existing questionnaire assessing a comparable use case in a family medicine setting in the United Kingdom [20]. The questionnaire extends the UTAUT model [34] with additional items on *perceived privacy and security* [35], which also emerged in our literature review as an additional element to understand technology acceptance in health care settings [18,40]. Dropped items with lower factor loadings in the original model of Abd-Alrazaq et al [20] were excluded from our questionnaire. For data restriction purposes, our study did not look at the actual use behavior.

To ensure construct comparability in different use cases, we replaced "patient portal" in the original questionnaire with the names of the other use cases. The questionnaire was then translated into German based on existing validated translations of the UTAUT construct [41] and pretested with 21 participants from the sample target group. Minor adjustments to answer options, the order of questions, and the wording of selected items were made (the translated survey can be found in Multimedia Appendix 2). From a methodological viewpoint, the survey study was aligned with CHERRIES (Checklist for Reporting Results of Internet E-Surveys) (Multimedia Appendix 3). The web-based survey was finally conducted between May 30, 2023, and November 22, 2023, using the software LimeSurvey (LimeSurvey GmbH).

Participants were shown brief descriptions of all the digital use cases in the scope of this paper to minimize any misunderstanding. Participants were then randomly assigned to 2 of the digital use cases for which they answered the technology acceptance items on a Likert scale. The number of use cases per participant was later reduced from 2 to 1 due to higher-than-expected dropout rates. This adjustment was appropriate as the order of questions was not changed.

XSL•FO
RenderX

Recruitment and Sample Composition

Our convenience sampling approach aimed to recruit a large cross-sectional sample of adults living in Germany. Since the only limiting factor to participate in the study was reaching legal age, our sample was broad, which is reflected in the variety of recruitment channels. Participants were mainly recruited via the Social Sciences' Panel, a convenience pool project for scientific research, as well as social media platforms such as LinkedIn, Instagram, and Facebook; online (research) forums; and flyers in the waiting rooms of family practices and other facilities. The questionnaire was only posted online in the open-access mode, and no responses from other sources (eg, physical questionnaires) were manually added. Further details, such as the wording of the advertisement of the survey, can be found in Multimedia Appendix 3.

Applied data cleaning mechanisms included removing duplicate entries, excluding the fastest 5% responses, eliminating entries with no reasonable answer behavior (straight lining or incorrectly answering 2 control questions), and deleting entries with no specified demographics and those not meeting the k≥5 criteria [42,43]. We believe that these data-cleaning measures contributed to the validity and reliability of our study. First, we avoided overrepresentation of multiple entries by single participants. Second, we enhanced response quality by minimizing obvious nondiligent answer behavior. Finally, we ensured meaningful analysis by eliminating "not specified" responses while at the same time adhering to our privacy policy by deleting responses that failed the k≥5 criteria.

Details of data cleaning measures can be found in Figure 2.

Figure 2. Included responses for data analysis. Impact of dropout and data cleaning measures.

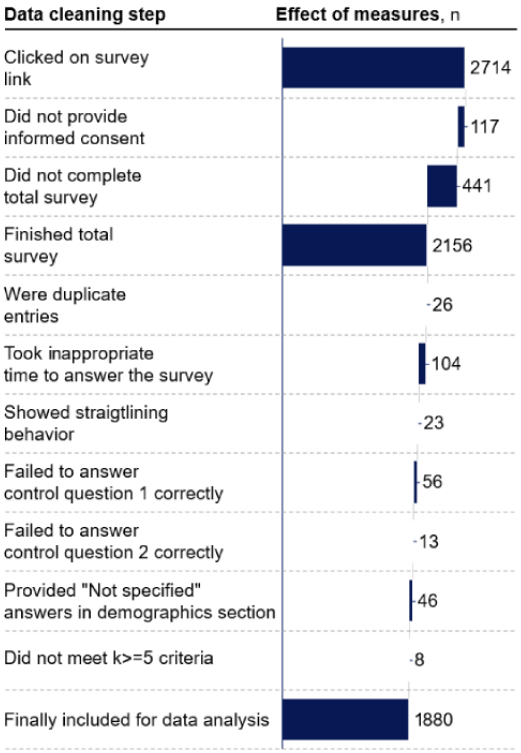

Empirical Analysis

Validation of the research model and statistical analysis via IBM SPSS 29 were performed separately for each use case in 3 steps.

First, Cronbach α values were calculated to evaluate the internal consistency of the study. The acceptable cutoff point in line with our exploratory research design was set at .60 [44]. Second, descriptive statistics were applied to test hypothesis A that different use cases have differences in awareness, availability, and actual usage among patients. Third, to test our hypotheses B and C regarding whether different use cases have different significant factors influencing behavioral intention to use, each research model was tested separately via multiple hierarchical regression. Thus, we tested the subordinated hypotheses for each digital use case.

The hierarchical regression was adequate to differentiate between the effects of technology acceptance and sociodemographic factors on intention to use. Therefore, we collected evidence for hypotheses B and C with a consistent analytical approach. Additionally, given the 4 different research models and a subsequent high number of subordinate hypotheses, a robust yet straightforward analytical approach was deemed necessary. The statistical significance level was set at $P<.05$ [45].

The first block entered into each model included only sociodemographic variables (age, gender, region of residency, size of the residence city, frequency of doctor visits, insurance status, and digital literacy). This allowed us to isolate the effects of different sociodemographic factors on intention to use, validating hypothesis C. The second block added the adapted UTAUT factors into the model (*performance expectancy*, *effort expectancy*, *social influence*, *facilitating condition*, and

XSL•FO
RenderX

perceived privacy and security). The final model addressed hypothesis B, testing the effect of technology acceptance factors on the intention to use a specific use case.

Four requirements for multiple regression models have been tested and confirmed for each of the use cases. First, multicollinearity can be rejected as all variance inflation factor (VIF) values were below 10 [46]. Second, autocorrelation can be rejected as the Durbin-Watson value for each use case was close to 2 [47]. Third, the normal distribution of residuals can be confirmed after visual inspection of the histogram of the standardized residuals. Lastly, homoscedasticity can be assumed as the regression of the standardized residual versus standardized predicted value scatter plot did not yield any considerable pattern.

By assessing usage rates, testing the technology acceptance model for each digital use case, and comparing our results across use cases, we collected evidence to assess our hypotheses A, B, and C. We accepted or rejected these superordinated hypotheses based on varying usage rates or different factors impacting the behavioral intention to use in different use cases (the decision was further validated with our research team).

Ethical Considerations

Our human subject research study was approved by the Ethics Committee of Witten/Herdecke University (Nr. S-245/2022). All participants provided informed consent in line with the GDPR (General Data Protection Regulation) guidelines prior to the start of the survey and had the ability to opt-out. No personal data were gathered, and sociodemographic data points were anonymized in accordance with k≥5 anonymity. All participants filled out the questionnaire voluntarily, and no compensation was offered.

Results

Internal Consistency and Sample Size

We calculated Cronbach α values for all 5 technology acceptance factors and the intention to use a specific digital use case. Therefore, we validated 24 variables across all 4 models. Of the 24 variables across all use cases, 23 met the threshold of >.60 [48], while 1 variable showed a Cronbach α of .57 (scheduling doctor appointments online: facilitating condition). The Cronbach α values for all 5 technology acceptance factors and intention to use are shown in Table 1.

Additionally, we leveraged the convenience sampling strategy to gather a large cross-sectional sample of German-speaking adults. Our final sample included 1880 adult respondents living in Germany. The demographic characteristics of the sample are shown in Table 2.

As the number of assigned use case assessments was reduced during the study, 175 out of the 1880 participants initially assessed the technology acceptance of 2 digital use cases, resulting in a total of 2055 technology acceptance assessments, with limited overlap between the groups (receiving medical consultations via video: n=494 [number of respondents who also assessed other use cases: n=80]; scheduling doctor appointments online: n=542 [number of respondents who also assessed other use cases: n=75]; storing personal medical information via electronic health records: n=528 [number of respondents who also assessed other use cases: n=103]; providing personal information before consultation digitally [digital anamnesis]: n=491 [number of respondents who also assessed other use cases: n=92]).

Table 1. Construct reliability (Cronbach α) in each digital use case.

Variable	Number of items	Digital use case			
		Receiving medical consultations via video	Scheduling doctor appointments online	Storing personal medical information via electronic health records	Providing personal information before consultation digitally (digital anamnesis)
Behavioral intention to use	3	.91	.90	.92	.90
Performance expectancy	3	.88	.85	.85	.87
Effort expectancy	4	.75	.71	.74	.76
Social influence	3	.91	.92	.91	.92
Facilitating condition	3	.70	.57	.60	.62
Perceived privacy and security	3	.71	.61	.74	.73

XSL•FO
RenderX

Table 2. Characteristics of the adult respondents living in Germany in 2023.

Variable	Value (N=1880)
Gender, n (%)	
Male	821 (43.7)
Female	1059 (56.3)
Age group (years), n (%)	
18-30 years	295 (15.7)
31-50 years	519 (27.6)
51-65 years	608 (32.3)
>65 years	458 (24.4)
Frequency group (visits in the last 12 months), n (%)	
0	132 (7.0)
1	362 (19.3)
2-3	705 (37.5)
>3	681 (36.2)
Insurance status, n (%)	
Statutorily	1459 (77.6)
Privately	421 (22.4)
Population size of residence (inhabitants), n (%)	
≤20,000	525 (27.9)
20,001-200,000	581 (30.9)
>200,000	774 (41.2)
Region, n (%)	
East	437 (23.2)
North	326 (17.3)
South	365 (19.4)
West	752 (40.0)
Self-assessed digital literacy, n (%)	
Low (1-4)	103 (5.5)
Medium (5-7)	455 (24.2)
High (8-10)	1322 (70.3)
Digital literacy score, mean (SD)	8.0 (1.8)

Current Patient Adoption Rates of Digital Use Cases in Family Medicine

There was support for hypothesis A that different use cases have differences in awareness, availability, and actual usage by patients.

Overall, low offering and usage rates of digital use cases were evident. For example, nearly all respondents (1750/1880, 93.1%) were aware of online booking platforms; however, only approximately one-third (617/1880, 32.8%) stated that this service is offered in their family practice. Nevertheless, the vast majority (537/617, 87.0%) of those respondents who were offered online booking services in their family practice reported using them. This usage-to-availability ratio was higher for scheduling doctor appointments online than for video consultations where only less than half of the participants (94/225, 41.8%) stated usage. The results are presented in Table 3. The share of respondents using digital use cases when offered at their family doctor was as follows: receiving medical consultations via video, 94/225 (41.8%); scheduling doctor appointments online, 537/617 (87.0%); storing personal medical information via electronic health records, 108/166 (65.1%); providing personal information before consultation digitally (digital anamnesis), 53/78 (68.0%).

XSL•FO
RenderX

Table 3. Patient awareness, availability, and patient usage of selected digital use cases in family medicine in Germany in 2023.

Statement and digital use case	Response option, n (%)				
	Yes	No	I don't know	Not offered in my family practice	Not specified
"I am aware of this use case"					
VC[a]	1302 (69.3)	571 (30.4)	—[b]	—	7 (0.4)
OS[c]	1750 (93.1)	130 (6.9)	—	—	0 (0.0)
EHR[d]	1220 (64.9)	638 (33.9)	—	—	22 (1.2)
DA[e]	784 (41.7)	1075 (57.2)	—	—	21 (1.1)
"The use case is offered at my family doctors "					
VC	225 (12.0)	1092 (58.1)	—	559 (29.7)	4 (0.2)
OS	617 (32.8)	1031 (54.8)	—	229 (12.2)	3 (0.2)
EHR	166 (8.8)	690 (36.7)	—	1016 (54.0)	8 (0.5)
DA	78 (4.0)	956 (50.9)	—	835 (44.4)	11 (1.0)
"I use this digital use case at my family doctors"					
VC	94 (5.0)	909 (48.4)	—	793 (42.2)	84 (5.0)
OS	537 (28.6)	583 (31.0)	—	712 (37.9)	48 (3.0)
EHR	108 (5.7)	907 (48.2)	—	622 (33.1)	243 (12.9)
DA	53 (3.0)	877 (46.6)	—	759 (40.4)	191 (10.2)

[a]VC: receiving medical consultations via video.

[b]Not applicable.

[c]OS: scheduling doctor appointments online.

[d]EHR: storing personal medical information via electronic health records.

[e]DA: providing personal information before consultation digitally (digital anamnesis).

There was a low usage rate of digital use cases as only 16.2% (304/1880) of respondents agreed or fully agreed that the degree of digitalization is key when selecting a family practice. Furthermore, the willingness to change the family practice when a particular digital use case is unavailable was found to depend on the use case (69/1880, 3.7% for receiving medical consultations via video; 200/1880, 10.6% for scheduling doctor appointments online; 131/1880, 7.0% for storing personal medical information via electronic medical records; and 82/1880, 4.4% for providing personal information before consultation digitally [digital anamnesis]). Nevertheless, regression analysis showed that when selecting a family practice, participants with higher digital literacy tended to assign a higher value to the degree of digitalization (B=0.226, SE 0.023; β=.223; $P<.001$).

Factors Explaining Technology Acceptance in Family Medicine

To better understand the predictors of technology acceptance, we conducted a multiple hierarchical linear regression analysis on the intention to use each use case. Block 1 collects evidence to validate hypothesis C, which involves testing the sole effect of sociodemographic variables on intention to use, while block 2 refers to hypothesis B, which involves validating the effect of technology acceptance factors. In doing so, we did not moderate prior experience with a particular use case to reduce the complexity of each model. The results are illustrated in Table 4. All details of the respective models can be found in Multimedia Appendix 4.

XSL•FO
RenderX

Table 4. Standardized coefficients for hierarchical linear regression models explaining patients' intention to use a specific digital use case in family medicine in Germany in 2023.

Block and variable	Receiving medical consultations via video (n=494)		Scheduling doctor appointments online (n=542)		Storing personal medical information via electronic health records (n=528)		Providing personal information before consultation digitally (digital anamnesis) (n=491)	
	β	*P* value	β	*P* value	β	*P* value	β	*P* value
Block 1: Sociodemographic variables								
Frequency of doctor visits (vs no doctor visit in the last 12 months)								
1 visit	.032	.69	.213	.01[a]	–.008	.92	.025	.74
2-3 visits	.134	.15	.233	.01[a]	.142	.10	.089	.27
>3 visits	.137	.14	.224	.01[a]	.066	.45	.136	.10
Gender (vs male)								
Female	.900	.04[a]	.084	.04[a]	–.101	.02[a]	–.037	.41
Age (vs 18-30 years)								
31-50 years	.087	.13	–.011	.84	–.017	.75	.013	.82
51-65 years	.107	.09	–.023	.69	–.021	.71	–.125	.04[a]
>65 years	.032	.60	–.083	.16	.020	.74	–.158	.01[a]
Size of the residence city (vs ≤20,000)								
20,001-200,000	–.025	.62	.022	.66	.056	.29	.007	.90
>200,000	–.057	.30	.032	.53	.023	.67	–.005	.92
Region (vs East)								
South	–.061	.22	.005	.93	–.028	.57	–.038	.48
West	–.061	.25	–.035	.48	–.031	.53	.022	.69
North	–.026	.64	.029	.58	.052	.33	.020	.74
Type of insurance	–.044	.32	–.010	.81	–.022	.60	.067	.13
Digital literacy	.335	<.001[a]	.408	<.001[a]	.329	<.001[a]	.250	<.001[a]
Block 2[b]: Sociodemographic variables + technology acceptance factors								
Performance expectancy	.527	<.001[a]	.513	<.001[a]	.348	<.001[a]	.410	<.001[a]
Effort expectancy	.135	.005[a]	.218	<.001[a]	.096	.03[a]	.236	<.001[a]
Social influence	.206	<.001[a]	.054	.07	.101	<.001[a]	.173	<.001[a]
Facilitating condition	.039	.36	.028	.44	.198	<.001[a]	.085	.045[a]
Perceived privacy and security	.053	.10	.094	<.001[a]	.243	<.001[a]	.104	.003[a]

[a]Statistically significant ($P<.05$).

[b]The effect of sociodemographic variables is not displayed as the focus of the analysis is on the effect of technology acceptance factors on intention to use.

All final models (including steps 1 and 2) significantly predicted technology acceptance, whereas the addition of the UTAUT variables in the second step increased the explained variance significantly in all models for all use cases. In the first step, sociodemographic variables alone explained on average 14.6% of the intention to use a specific digital use case (receiving medical consultations via video: R^2=0.124; $F_{14,479}$=4.849; $P<.001$; scheduling doctor appointments online: R^2=0.189; $F_{14,527}$=8.780; $P<.001$; storing personal medical information via electronic health records: R^2=0.150; $F_{14,513}$=6.484; $P<.001$; providing personal information before consultation digitally [digital anamnesis]: R^2=0.121; $F_{14,476}$=4.698; $P<.001$). The final models explained on average 62.8% of the variance in intention to use (receiving medical consultations via video: R^2=0.624; $F_{19,474}$=41.356; $P<.001$; scheduling doctor appointments online: R^2=0.624; $F_{19,522}$=45.617; $P<.001$; storing personal medical

XSL•FO
RenderX

information via electronic health records: R^2=0.631; $F_{19,508}$=45.655; P<.001; providing personal information before consultation digitally [digital anamnesis]: R^2=0.631; $F_{19,471}$=42.364; P<.001).

There was support for hypothesis B that different use cases have different significant technology acceptance factors. Only *performance expectancy* and *effort expectancy* were significant predictors of technology acceptance for all use cases. Additionally, the magnitude of the effects of technology acceptance factors differed between use cases (*performance expectancy*: receiving medical consultations via video: B=0.603, SE 0.049; β=.527; P<.001; scheduling doctor appointments online: B=0.566, SE 0.043; β=.513; P<.001; storing personal medical information via electronic health records: B=0.405, SE 0.047; β=.348; P<.001; providing personal information before consultation digitally [digital anamnesis]: B=0.434, SE 0.048; β=.410; P<.001; *effort expectancy*: receiving medical consultations via video: B=0.236, SE 0.084; β=.135; P=.01; scheduling doctor appointments online: B=0.391, SE 0.075; β=.218; P<.001; storing personal medical information via electronic health records: B=0.153, SE 0.070; β=.096; P=.03; providing personal information before consultation digitally [digital anamnesis]: B=0.366, SE 0.073; β=.236; P<.001).

We found that *social influence* was significant for all use cases, except for scheduling doctor appointments online (receiving medical consultations via video: B=0.227, SE 0.036; β=.206; P<.001; storing personal medical information via electronic health records: B=0.108, SE 0.033; β=.101; P=.001; providing personal information before consultation digitally [digital anamnesis]: B=0.186, SE 0.036; β=.173; P<.001). Moreover, *perceived privacy and security* was significant for all use cases, except for receiving medical consultations via video (scheduling doctor appointments online: B=0.117, SE 0.037; β=.094; P=.002; storing personal medical information via electronic health records: B=0.284, SE 0.040; β=.243; P<.001; providing personal information before consultation digitally [digital anamnesis]: B=0.121, SE 0.040; β=.104; P=.003). *Facilitating condition* was only significant for storing personal medical information via electronic health records (B=0.299, SE 0.061; β=.198; P<.001) and providing personal information before consultation digitally (digital anamnesis) (B=0.127, SE 0.063; β=.085; P=.045). A simplified illustration of the different magnitudes of effects is presented in Figure 3.

Figure 3. Magnitudes of the effects of technology acceptance factors on the behavioral intention to use a digital use case in family medicine in Germany in 2023 (ranking based on β).

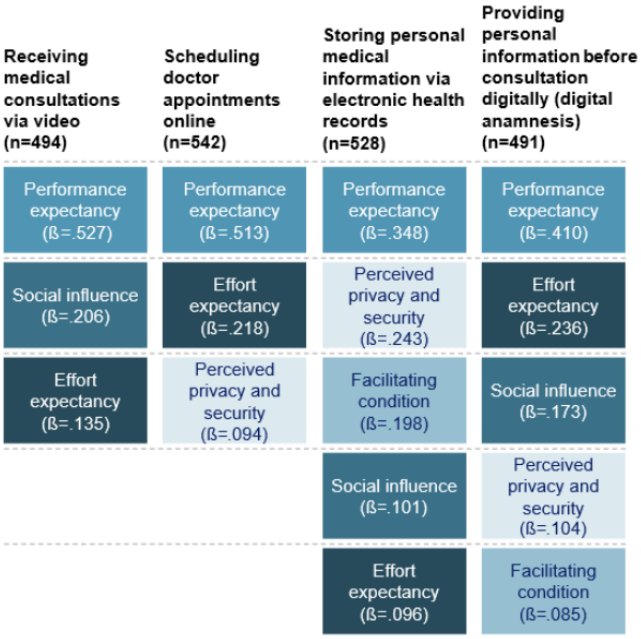

There was partial support for hypothesis C that different use cases have different sociodemographic factors impacting the behavioral intention, as self-assessed digital literacy had a positive effect for all use cases (receiving medical consultations via video: B=0.341, SE 0.049; β=.335; P<.001; scheduling doctor appointments online: B=0.322, SE 0.033; β=.408; P<.001; storing personal medical information via electronic health records: B=0.308, SE 0.041; β=.329; P<.001; providing personal information before consultation digitally [digital anamnesis]: B=0.239, SE 0.044; β=.250; P<.001).

Further, significant effects of sociodemographic variables on intention to use depended on the use case. Gender (female vs male) had an effect on 3 digital use cases, but the direction of that effect varied (receiving medical consultations via video: B=0.321, SE 0.159; β=.090; P=.04; scheduling doctor appointments online: B=0.247, SE 0.122; β=.084; P=.04; storing personal medical information via electronic health records: B=–0.356, SE 0.149; β=–.101; P=.02). Additionally, the frequency of doctor visits had a positive effect on the intention to use online appointment systems and age had a negative effect on digital anamnesis tools. Compared with respondents who had not visited a doctor in the last 12 months, all frequency groups showed a positive effect (1 visit in the last 12 months: B=0.783, SE 0.285; β=0.213; P=.006; 2-3 visits: B=0.698, SE 0.267; β=.233; P=.009; >3 visits: B=0.690, SE 0.270; β=.224;

XSL•FO
RenderX

P=.01). Compared with the age group of 18-30 years, there were negative effects for the age groups of 51-65 years (B=–0.476, SE 0.235; β=–.125; P=.04) and >65 years (B=–0.614, SE 0.239; β=–.158; P=.01).

The size of the residence city, region, and type of insurance did not affect the intention to use any digital use cases. Remarkably, Welch ANOVA found that digital literacy was dependent on age (Welch $F_{3,968.29}$=53.441; P<.001).

Discussion

Principal Findings

This research paper aimed to compare the different technology acceptance factors of patients for key digital use cases in family medicine and to support family doctors in selecting and implementing digital use cases. Our results indicated that the degree of digitalization of family practice is not critically important for patients. Additionally, our results fully supported hypothesis A, highlighting that different use cases have differences in awareness, availability, and actual usage among patients. Further, we validated hypothesis B, showing that different use cases are affected by different technology acceptance factors. Nevertheless, our results only partially supported hypothesis C as digital literacy had a significant effect on the intention to use for all use cases.

Thus, the varying relative importance of technology acceptance factors for each digital use case emphasizes the need for a nuanced selection and implementation approach of digital use cases for family practice. Simultaneously, the partial rejection of different effects of sociodemographic variables could imply that our recommendations can be applied in different sociodemographic contexts.

Comparison With Prior Work

We introduced a systematic assessment of technology acceptance for online booking platforms and digital anamnesis tools as existing studies concentrated only on video consultation platforms and electronic health records. Hence, we could only compare our results to prior work on these 2 use cases.

Receiving Medical Consultations via Video

The strongest effect of *performance expectancy* has been validated in various studies [18,49]. However, the literature indicates mixed results for the effect of *effort expectancy* and *social influence* on the intention to use video consultation platforms. For example, Schmitz et al [18] rejected the effect of *social influence*, claiming that medical appointments are highly personal, and thus, the fact of undergoing a medical treatment is not subjective to the opinions of others, regardless of a virtual or physical setting. However, our findings showed that the mode of the consultation might be affected by other opinions. This is supported by Esber et al [17], who confirmed the significance of *social influence* on the acceptance of video consultations, and Mueller et al [4], who stressed the importance of social cues to drive the technology acceptance of preusers.

Additionally, we found support for our rejection of the effect of *facilitating condition* on the intention to use video consultations in previous work [18,39,49,50]. Schmitz et al [18] and Brooks [51] mention that video consultation technologies are comparable to already widespread technologies, such as FaceTime, and thus, people are already well equipped for their use in a medical setting.

Surprisingly, we rejected the effect of *perceived privacy and security* on the intention to use video consultations but accepted this effect for other use cases. This contrasts with the findings of Zobair et al [52] and Schmitz et al [18] who highlighted the effect of perceived security on the intention to use telemedicine or telehealth in Bangladesh and Germany. Viana Pereira et al [39] postulated that doctors could also provide advice on general health topics via video and thus would not cause confidentiality concerns among patients. Nevertheless, we argued that the neglect of the importance of *perceived privacy and security* for patients requires further validation.

Storing Personal Medical Information via Electronic Health Records

Performance expectancy was constantly confirmed as a significant predictor of intention to use [16,20,40,53-63]. Additionally, we found evidence for positive effects of *effort expectancy* [16,20,53,55-57,62] and *facilitating condition* [20,60] on intention to use. Nevertheless, our findings contradict those of Tavares and Olivera [59] who suggested that early adopters do not perceive obstacles in using electronic health records due to their higher cognitive ability and experience with both technological complexities and IT infrastructure. This is further supported by Aydin and Kumru [40] who rejected *facilitating condition* as a predictor of the intention to use electronic medical records for Gen-Z university students. We explain this discrepancy with our more diverse sample, which we believe is less biased toward high cognitive ability and technological experience.

Our findings of a positive effect of *social influence* confirm the findings of Tavares et al [16] who illustrated inadequate promotion activities as a possible reason for the failed implementation of a comparable health record technology [64]. Nevertheless, Abd-Alrazaq et al [20] identified *social influence* as a nonsignificant predictor for behavioral intention, explaining that usage is voluntary and thus less objective to *social influence*. We hypothesized that the ongoing discussion on the transition from an opt-in to an opt-out model in Germany (default mandatory use of electronic health records) could partially explain the significance of *social influence* in our research [10].

Further studies that assessed comparable constructs of *perceived security and privacy* confirmed the positive effect on the intention to use electronic health records [20,40,53,55,56].

Effect of Sociodemographic Variables on the Intention to Use Digital Use Cases

Except for digital literacy, we found no clear pattern of the effect of sociodemographic variables on the intention to use a specific digital use case. The vital role of digital literacy in technology acceptance is in line with research that postulates a digital divide of patient groups in engaging in digital use cases [65]. However, this role of digital literacy necessitates further differentiation. First, the participants in our study might be rather digitally savvy, and thus, this effect might be overgeneralized. Second,

XSL•FO
RenderX

we acknowledge different magnitudes of this effect across use cases, suggesting that inherent characteristics of use cases (eg, technical complexity) might impact the role of digital literacy in technology acceptance. In identifying digitally literate patients, our findings showed a negative effect of age on digital literacy, with further evidence in other studies [66].

Our study found no effect of the residence city size on patients' intention to use video consultations. This is particularly interesting as, for example, receiving video consultations is seen as a means to ensure adequate treatment opportunities in rural areas, with Kane and Gillis [67] highlighting that the use of telemedicine by physicians is higher in nonmetropolitan areas than in metropolitan areas in the United States. On the other hand, McGrail et al [68] noted no difference in the offering of telemedicine between rural and urban settings in Canada. This contradiction indicates that the effect of the residence city size on intention to use or actual usage depends on further factors (eg, the potential difference in the digital literacy of people living in rural vs metropolitan areas or the lack of availability of video consultation offerings in some regions).

Measures for Family Doctors to Increase Patient Usage of Digital Use Cases

Importance of the Digital Maturity of Family Practices for Patients

Family doctors should carefully assess whether investing in digital use cases is adequate to attract and sustain patients. Our results showed that the level of digitalization of family practices was not a vital criterion for patients when choosing a doctor. This confirms the findings of Kuruoglu et al [69] and Khatami et al [70] who did not mention the degree of digitalization as a selection criterion for patients.

Nevertheless, our results showed that patients with higher digital literacy had greater importance for the degree of digitalization when selecting a family practice. Thus, digital investment may be more vital for family doctors treating more digitally literate patients. Hence, family doctors who refuse to invest in digital technology could lose these patient groups to family doctors who invest in digital technology.

Selection of Digital Use Cases

Family doctors should consider patient demand in the selection of digital use cases. Our study highlights the attractiveness of use cases for patients by comparing offering and actual usage rates. According to this approach, scheduling doctor appointments online appeared to be the most attractive digital use case in Germany in 2023, followed by providing personal information before consultation digitally (digital anamnesis). Video consultation was the least favorable use case for patients. Interestingly, a 2023 study on the digital offerings of family doctors in Germany indicated that video consultations were nearly 4 times more often offered than digital anamnesis tools [71]. Nevertheless, we are aware that patient demand is only 1 factor for offering use cases [13], and further patient-specific and use case–specific barriers might bias this attractivity assessment. We excluded electronic health records as family doctors in Germany must document patient treatment within the respective electronic health record of the patient ("Elektronische Patientenakte" [10]). Hence, family doctors do not select this digital use case voluntarily.

Implementation of Digital Use Cases

To incentivize usage among patients, family doctors should adopt an implementation approach for each use case, given that we showed the varying importance of technology acceptance factors in each use case. Nevertheless, we identified 3 factors that affected the intention to use for all use cases, and these should always be considered by family doctors.

First, doctors should prioritize the value addition of the digital use case to the patient. This is especially important for video consultations. The focus on value addition to drive adoption aligns with other academic papers [18,49] that stress the ability of doctors to communicate the added value to the patient. Hence, family doctors should receive adequate training or communication material from providers to explain the added value to patients effectively.

Second, family doctors should assess the importance of an intuitive interface for solutions as *effort expectancy* had a significant effect across all use cases. Nevertheless, the relative importance of this factor varied by use case. Since we believe that *effort expectancy* is dependent on the design and user experience of solutions and family doctors have limited control over these, the correct initial selection of a solution provider is vital.

Third, family practices should consider the digital literacy of their patients as it is a predictor of the intention to use across all use cases. We acknowledge that the digital literacy of family doctors affects their own technology adoption [72]. Hence, we believe that not only family doctors themselves should ensure that patients with low digital literacy are not excluded from using digital use cases.

Further considerations for family doctors depend on the respective digital use case. To increase the use of video consultations, family doctors should additionally consider endorsements from testimonials or other influencing bodies. Surprisingly, we showed no effect of *perceived privacy and security* on the intention to use video consultations. This could indicate that this factor might be neglectable for family doctors; however, further validation is needed. On the contrary, *perceived privacy and security* is important for online booking platforms. Therefore, family doctors should articulate how online booking platforms ensure the privacy of end users and how platforms communicate about this. Additionally, *perceived privacy and security* is the second most important factor to drive the use of electronic health records. Therefore, family doctors should focus on whether patients feel their data are protected. While this factor is also to be considered for digital anamnesis tools, having an easy-to-use solution here is more important than for electronic health records.

Limitations and Further Research

To the best of our knowledge, our web-based survey with 1880 included participants is one of the largest recent studies assessing technology acceptance in a health care setting. By comparing 4 distinct use cases, we provided a new nuanced assessment

XSL•FO
RenderX

and avoided attitude generalization. Contrary to studies that only included users or nonusers of a digital use case [20,54], our research included both users and nonusers in a family medicine context.

Nevertheless, we acknowledge that our study had some limitations. First, due to our voluntary and web-based survey approach, our sample might be biased toward participants with higher digital literacy and higher interest in digital health topics. This excluded patients with no access to the internet. However, we believe our sample is appropriate as digital use cases require internet access and our sample includes these potential users. Additionally, we tried to minimize bias by distributing physical questionnaires in family practices; however, responses to distributed flyers were limited. Second, replicating the same construct with the exact wordings for each use case potentially did not address use case specifics and thus increased the risk of ambiguity of items. For example, an electronic health record is not offered by the family doctor but rather by the insurance provider. In addition, the questionnaire was translated into German for the first time, which might explain the presence of a few low Cronbach α values. Dropping items would have significantly increased the Cronbach α values; however, to ensure comparability with the original questionnaire, we refrained from doing so. For example, for the worst Cronbach α (scheduling doctor appointments online: *facilitating condition*), only 1 item showed a low correlation with the other items. Furthermore, we did not have access to data validating the offer rates and usage rates of digital use cases by the participants but rather relied on the subjective inputs of the respondents. Nevertheless, as we offered the response options "I don't know" and "Not specified," we hypothesized high validity in the remaining responses with "Yes" or "No" answer entries. In addition, we did not account for prior experience, potentially not differentiating the distinct effects of expectations or past use on technology acceptance. Finally, we only looked at behavioral intention to use; however, various studies have highlighted a gap between intention and usage behavior, which is referred to as the intention-behavior gap [73]. Thus, our research did not predict the actual usage of digital use cases.

There are 3 areas for future research. First, it is vital to understand the technology acceptance of people with limited access to the internet (eg, people living in elderly homes) or low digital literacy and discuss measures on how to minimize the digital divide. Second, the effects of the "Digital Act (Digital-Gesetz)" of Germany [10] on usage by patients should be examined, particularly the mandatory implementation of electronic health records. Third, our research focused on Germany, and thus, the applicability of our findings to other countries could be discussed.

Conclusion

Patients assign lower importance to the degree of digitalization of their family practices. Additionally, the usage of digital use cases remains low. Therefore, family doctors should carefully assess whether investments in digital use cases are adequate. Nevertheless, as the benefits of these solutions are established in the literature, we see future potential for higher adoption. For capturing this adoption potential, we believe in the importance of the role of family doctors owing to the established and trusted patient-doctor relationship. To leverage this role and increase patient usage, family doctors should align their selection and implementation criteria with our findings of the varying technology acceptance factors. We showed that *performance expectancy*, *effort expectancy*, and *digital literacy* have positive effects on the intention to use across all use cases, but further acceptance factors depend on the use case. To illustrate, family doctors should always emphasize the perceived added value for patients, whereas the neglect of privacy and security aspects for video consultation platforms requires further validation. Family doctors should always prioritize *performance expectancy*, while guides for the selection and implementation of digital tools might emerge for each digital use case.

Acknowledgments

We thank the Social Sciences' Panel (a convenience pool project of the Institute for Communication Science and Media Research [LMU Munich] and the German Communication Association [DGPuK]) for their support in the recruitment of participants for our survey.

Data Availability

The datasets incorporated and examined during this research project are available from the corresponding author upon reasonable request and on consideration of restrictions associated with the ethics approval.

Authors' Contributions

JB, SM, and LF conceptualized the main research question and derived the appropriate research design. JB conducted the literature review, created the survey in an online tool, and analyzed the gathered information with adequate statistical measures. SR provided feedback regarding the selection of appropriate statistical tools and reviewed the statistical data analysis. JB authored the initial manuscript for which SM, LF, AM, and SR contributed revisions. SM orchestrated the research project. All authors reviewed the manuscript for submission.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
Literature review (including the PRISMA-ScR [Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews] checklist).
[DOCX File , 2575 KB - formative_v9i1e58867_app1.docx]

Multimedia Appendix 2
Translated survey questionnaire.
[DOCX File , 50 KB - formative_v9i1e58867_app2.docx]

Multimedia Appendix 3
CHERRIES (Checklist for Reporting Results of Internet E-Surveys).
[DOCX File , 53 KB - formative_v9i1e58867_app3.docx]

Multimedia Appendix 4
Overview of regression coefficients for the regression models predicting the intention to use digital use cases.
[DOCX File , 92 KB - formative_v9i1e58867_app4.docx]

References

1. Bashshur RL, Howell JD, Krupinski EA, Harms KM, Bashshur N, Doarn CR. The empirical foundations of telemedicine interventions in primary care. Telemed J E Health 2016 May;22(5):342-375 [FREE Full text] [doi: 10.1089/tmj.2016.0045] [Medline: 27128779]
2. Wanderås MR, Abildsnes E, Thygesen E, Martinez SG. Video consultation in general practice: a scoping review on use, experiences, and clinical decisions. BMC Health Serv Res 2023 Mar 30;23(1):316 [FREE Full text] [doi: 10.1186/s12913-023-09309-7] [Medline: 36997997]
3. Wangler J, Jansky M. How can primary care be secured in the long term? - a qualitative study from the perspective of general practitioners in Germany. Eur J Gen Pract 2023 Dec;29(1):2223928 [FREE Full text] [doi: 10.1080/13814788.2023.2223928] [Medline: 37387362]
4. Mueller M, Knop M, Niehaves B, Adarkwah CC. Investigating the acceptance of video consultation by patients in rural primary care: empirical comparison of preusers and actual users. JMIR Med Inform 2020 Oct 22;8(10):e20813 [FREE Full text] [doi: 10.2196/20813] [Medline: 32969339]
5. Golinelli D, Boetto E, Carullo G, Nuzzolese AG, Landini MP, Fantini MP. Adoption of digital technologies in health care during the COVID-19 pandemic: systematic review of early scientific literature. J Med Internet Res 2020 Nov 06;22(11):e22280 [FREE Full text] [doi: 10.2196/22280] [Medline: 33079693]
6. AMA digital health care 2022 study findings. American Medical Association. URL: https://www.ama-assn.org/about/research/ama-digital-health-care-2022-study-findings [accessed 2023-12-06]
7. TI-Dashboard. Gematik. URL: https://www.gematik.de/telematikinfrastruktur/ti-dashboard [accessed 2024-01-31]
8. Versicherte in der gesetzlichen Krankenversicherung. Bundesministerium für Gesundheit. URL: https://www.bundesgesundheitsministerium.de/gesetzlich-versicherte [accessed 2024-01-31]
9. Appointments in General Practice. NHS England. URL: https://digital.nhs.uk/data-and-information/publications/statistical/appointments-in-general-practice#about-this-publication [accessed 2024-02-13]
10. Gesetz zur Beschleunigung der Digitalisierung des Gesundheitswesens (Digital-Gesetz – DigiG). Bundesministerium für Gesundheit. URL: https://www.bundesgesundheitsministerium.de/ministerium/gesetze-und-verordnungen/guv-20-lp/digig.html [accessed 2024-01-31]
11. Securing Excellence in Primary Care (GP) Digital Services: The Primary Care (GP) Digital Services Operating Model 2021-2023. NHS England. URL: https://www.england.nhs.uk/digitaltechnology/digital-primary-care/gp-digital-services-operating-model-21-23/ [accessed 2024-02-29]
12. Breedvelt JJ, Zamperoni V, Kessler D, Riper H, Kleiboer AM, Elliott I, et al. GPs' attitudes towards digital technologies for depression: an online survey in primary care. Br J Gen Pract 2019 Mar;69(680):e164-e170 [FREE Full text] [doi: 10.3399/bjgp18X700721] [Medline: 30559111]
13. Weik L, Fehring L, Mortsiefer A, Meister S. Understanding inherent influencing factors to digital health adoption in general practices through a mixed-methods analysis. NPJ Digit Med 2024 Feb 27;7(1):47 [FREE Full text] [doi: 10.1038/s41746-024-01049-0] [Medline: 38413767]
14. De Santis KK, Jahnel T, Sina E, Wienert J, Zeeb H. Digitization and health in Germany: cross-sectional nationwide survey. JMIR Public Health Surveill 2021 Nov 22;7(11):e32951 [FREE Full text] [doi: 10.2196/32951] [Medline: 34813493]
15. Napitupulu D, Yacub R, Putra A. Factor influencing of telehealth acceptance during COVID-19 outbreak: extending UTAUT model. IJIES 2021 Jun 30;14(3):267-281. [doi: 10.22266/ijies2021.0630.23]
16. Tavares J, Goulão A, Oliveira T. Electronic Health Record Portals adoption: Empirical model based on UTAUT2. Inform Health Soc Care 2018 Mar;43(2):109-125. [doi: 10.1080/17538157.2017.1363759] [Medline: 29035646]

XSL•FO
RenderX

17. Esber A, Teufel M, Jahre L, In der Schmitten J, Skoda E, Bäuerle A. Predictors of patients' acceptance of video consultation in general practice during the coronavirus disease 2019 pandemic applying the unified theory of acceptance and use of technology model. Digit Health 2023;9:20552076221149317 [FREE Full text] [doi: 10.1177/20552076221149317] [Medline: 36815005]
18. Schmitz A, Díaz-Martín A, Yagüe Guillén MJ. Modifying UTAUT2 for a cross-country comparison of telemedicine adoption. Comput Human Behav 2022 May;130:107183 [FREE Full text] [doi: 10.1016/j.chb.2022.107183] [Medline: 35017788]
19. Dahlhausen F, Zinner M, Bieske L, Ehlers JP, Boehme P, Fehring L. There's an app for that, but nobody's using it: Insights on improving patient access and adherence to digital therapeutics in Germany. Digit Health 2022;8:20552076221104672 [FREE Full text] [doi: 10.1177/20552076221104672] [Medline: 35811758]
20. Abd-Alrazaq A, Bewick BM, Farragher T, Gardner P. Factors affecting patients' use of electronic personal health records in England: cross-sectional study. J Med Internet Res 2019 Jul 31;21(7):e12373 [FREE Full text] [doi: 10.2196/12373] [Medline: 31368442]
21. The European Definition of General Practice / Family Medicine. WONCA. URL: https://www.woncaeurope.org/file/41f61fb9-47d5-4721-884e-603f4afa6588/WONCA_European_Definitions_2_v7.pdf [accessed 2024-01-12]
22. Keuper J, Batenburg R, Verheij R, van Tuyl L. Use of e-health in Dutch general practice during the COVID-19 pandemic. Int J Environ Res Public Health 2021 Nov 26;18(23):12479 [FREE Full text] [doi: 10.3390/ijerph182312479] [Medline: 34886204]
23. Farias FACD, Dagostini CM, Bicca YA, Falavigna VF, Falavigna A. Remote patient monitoring: a systematic review. Telemed J E Health 2020 May;26(5):576-583. [doi: 10.1089/tmj.2019.0066] [Medline: 31314689]
24. Zhao P, Yoo I, Lavoie J, Lavoie BJ, Simoes E. Web-based medical appointment systems: a systematic review. J Med Internet Res 2017 Apr 26;19(4):e134 [FREE Full text] [doi: 10.2196/jmir.6747] [Medline: 28446422]
25. What is an electronic health record (EHR)? HealthIT. URL: https://www.healthit.gov/faq/what-electronic-health-record-ehr [accessed 2023-12-04]
26. Kopp M, Wetzl M, Geissler F, Roth JP, Wallner R, Hoefler D, et al. Structured digital self-assessment of patient anamnesis prior to computed tomography: performance evaluation and added value. J Med Syst 2021 Jan 28;45(3):30 [FREE Full text] [doi: 10.1007/s10916-020-01690-8] [Medline: 33511485]
27. Robinson SK, Meisnere M, Phillips RL, McCauley L, editors. Implementing High-Quality Primary Care: Rebuilding the Foundation of Health Care. Washington, DC: National Academies Press; 2021.
28. Popa V, Geissler J, Vermeulen R, Priest E, Capperella K, Susuzlu G, et al. Delivering digital health solutions that patients need: a call to action. Ther Innov Regul Sci 2024 Mar;58(2):236-241 [FREE Full text] [doi: 10.1007/s43441-023-00592-4] [Medline: 38082118]
29. Taherdoost H. A review of technology acceptance and adoption models and theories. Procedia Manufacturing 2018;22:960-967. [doi: 10.1016/j.promfg.2018.03.137]
30. AlQudah AA, Al-Emran M, Shaalan K. Technology acceptance in healthcare: a systematic review. Applied Sciences 2021 Nov 09;11(22):10537. [doi: 10.3390/app112210537]
31. Heinsch M, Wyllie J, Carlson J, Wells H, Tickner C, Kay-Lambkin F. Theories informing eHealth implementation: systematic review and typology classification. J Med Internet Res 2021 May 31;23(5):e18500 [FREE Full text] [doi: 10.2196/18500] [Medline: 34057427]
32. Rahi S, Khan MM, Alghizzawi M. Factors influencing the adoption of telemedicine health services during COVID-19 pandemic crisis: an integrative research model. Enterprise Information Systems 2020 Dec 08;15(6):769-793. [doi: 10.1080/17517575.2020.1850872]
33. Hsieh H, Lai J, Chuang B, Tsai C. Determinants of telehealth continuance intention: a multi-perspective framework. Healthcare (Basel) 2022 Oct 15;10(10):2038 [FREE Full text] [doi: 10.3390/healthcare10102038] [Medline: 36292485]
34. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view. MIS Quarterly 2003;27(3):425. [doi: 10.2307/30036540]
35. Whetstone M, Goldsmith R. Factors influencing intention to use personal health records. Intl J of Pharm & Health Mrkt 2009 Apr 03;3(1):8-25. [doi: 10.1108/17506120910948485]
36. Doctolib Digital Health Report 2023. Doctolib. URL: https://about.doctolib.de/news/doctolib-digital-health-report-2023/ [accessed 2024-01-29]
37. Tricco AC, Zarin W, Antony J, Hutton B, Moher D, Sherifali D, et al. An international survey and modified Delphi approach revealed numerous rapid review methods. J Clin Epidemiol 2016 Feb;70:61-67. [doi: 10.1016/j.jclinepi.2015.08.012] [Medline: 26327490]
38. King VJ, Stevens A, Nussbaumer-Streit B, Kamel C, Garritty C. Paper 2: Performing rapid reviews. Syst Rev 2022 Jul 30;11(1):151 [FREE Full text] [doi: 10.1186/s13643-022-02011-5] [Medline: 35906677]
39. Viana Pereira F, Tavares J, Oliveira T. Adoption of video consultations during the COVID-19 pandemic. Internet Interv 2023 Mar;31:100602 [FREE Full text] [doi: 10.1016/j.invent.2023.100602] [Medline: 36694630]
40. Aydin G, Kumru S. Paving the way for increased e-health record use: elaborating intentions of Gen-Z. Health Syst (Basingstoke) 2023 Oct 03;12(3):281-298 [FREE Full text] [doi: 10.1080/20476965.2022.2129471] [Medline: 37860594]

XSL•FO
RenderX

41. Harborth D, Pape S. German translation of the Unified Theory of Acceptance and Use of Technology 2 (UTAUT2) questionnaire. SSRN Journal 2018:1-12. [doi: 10.2139/ssrn.3147708]
42. Bais F, Schouten B, Toepoel V. Investigating response patterns across surveys: do respondents show consistency in undesirable answer behaviour over multiple surveys? Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique 2020 Aug 26;147-148(1-2):150-168. [doi: 10.1177/0759106320939891]
43. Arevalo M, Brownstein NC, Whiting J, Meade CD, Gwede CK, Vadaparampil ST, et al. Strategies and lessons learned during cleaning of data from research panel participants: cross-sectional web-based health behavior survey study. JMIR Form Res 2022 Jun 23;6(6):e35797 [FREE Full text] [doi: 10.2196/35797] [Medline: 35737436]
44. Taber KS. The use of Cronbach's alpha when developing and reporting research instruments in science education. Res Sci Educ 2017 Jun 7;48(6):1273-1296. [doi: 10.1007/s11165-016-9602-2]
45. Andrade C. The value and statistical significance: misunderstandings, explanations, challenges, and alternatives. Indian J Psychol Med 2019;41(3):210-215 [FREE Full text] [doi: 10.4103/IJPSYM.IJPSYM_193_19] [Medline: 31142921]
46. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning: With applications in R. New York, NY: Springer; 2017.
47. Turner P. Critical values for the Durbin-Watson test in large samples. Applied Economics Letters 2019 Nov 12;27(18):1495-1499. [doi: 10.1080/13504851.2019.1691711]
48. Hair J, Black W, Babin B, Anderson R. Multivariate Data Analysis: A Global Perspective. London, UK: Pearson Education; 2010.
49. Amin R, Hossain MA, Uddin MM, Jony MTI, Kim M. Stimuli influencing engagement, satisfaction, and intention to use telemedicine services: an integrative model. Healthcare (Basel) 2022 Jul 18;10(7):1327 [FREE Full text] [doi: 10.3390/healthcare10071327] [Medline: 35885854]
50. Alexandra S, Handayani PW, Azzahro F. Indonesian hospital telemedicine acceptance model: the influence of user behavior and technological dimensions. Heliyon 2021 Dec;7(12):e08599 [FREE Full text] [doi: 10.1016/j.heliyon.2021.e08599] [Medline: 34977414]
51. Brooks NP. Telemedicine is here. World Neurosurg 2016 Nov;95:603-604. [doi: 10.1016/j.wneu.2016.02.113] [Medline: 26960280]
52. Zobair KM, Sanzogni L, Houghton L, Sandhu K, Islam MJ. Health seekers' acceptance and adoption determinants of telemedicine in emerging economies. AJIS 2021 Dec 26;25:3071. [doi: 10.3127/ajis.v25i0.3071]
53. Abd-Alrazaq A, Alalwan AA, McMillan B, Bewick BM, Househ M, Al-Zyadat A. Patients' adoption of electronic personal health records in England: secondary data analysis. J Med Internet Res 2020 Oct 07;22(10):e17499 [FREE Full text] [doi: 10.2196/17499] [Medline: 33026353]
54. Agrawal L, Ndabu T, Mulgund P, Sharman R. Factors affecting the extent of patients' electronic medical record use: an empirical study focusing on system and patient characteristics. J Med Internet Res 2021 Oct 28;23(10):e30637 [FREE Full text] [doi: 10.2196/30637] [Medline: 34709181]
55. Alsyouf A, Lutfi A, Alsubahi N, Alhazmi FN, Al-Mugheed K, Anshasi RJ, et al. The Use of a Technology Acceptance Model (TAM) to predict patients' usage of a personal health record system: the role of security, privacy, and usability. Int J Environ Res Public Health 2023 Jan 11;20(2):1347 [FREE Full text] [doi: 10.3390/ijerph20021347] [Medline: 36674105]
56. Bile Hassan I, Murad MAA, El-Shekeil I, Liu J. Extending the UTAUT2 model with a privacy calculus model to enhance the adoption of a health information application in Malaysia. Informatics 2022 Mar 28;9(2):31. [doi: 10.3390/informatics9020031]
57. Liu J, Gong X, Weal M, Dai W, Hou S, Ma J. Attitudes and associated factors of patients' adoption of patient accessible electronic health records in China - A mixed methods study. Digit Health 2023;9:20552076231174101 [FREE Full text] [doi: 10.1177/20552076231174101] [Medline: 37188077]
58. Niemöller S, Hübner U, Egbert N, Babitsch B. How to access personal health records? measuring the intention to use and the perceived usefulness of two different technologies: a randomised controlled study. Stud Health Technol Inform 2019 Sep 03;267:197-204. [doi: 10.3233/SHTI190827] [Medline: 31483273]
59. Tavares J, Oliveira T. New integrated model approach to understand the factors that drive electronic health record portal adoption: cross-sectional national survey. J Med Internet Res 2018 Nov 19;20(11):e11032 [FREE Full text] [doi: 10.2196/11032] [Medline: 30455169]
60. Keung K, Lee C, Ng K, Leung S, Choy K. An empirical study on patients' acceptance and resistance towards electronic health record sharing system: a case study of Hong Kong. International Journal of Knowledge and Systems Science 2018;9(2):1-27. [doi: 10.4018/ijkss.2018040101]
61. Alharbey R. Integrative adoption model for personal health records: A structural equation modeling approach. Int. j. adv. appl. sci 2022 Jan;9(1):117-127. [doi: 10.21833/ijaas.2022.01.014]
62. Hoogenbosch B, Postma J, de Man-van Ginkel JM, Tiemessen NA, van Delden JJ, van Os-Medendorp H. Use and the users of a patient portal: cross-sectional study. J Med Internet Res 2018 Sep 17;20(9):e262 [FREE Full text] [doi: 10.2196/jmir.9418] [Medline: 30224334]
63. Razmak J, Bélanger C. Using the technology acceptance model to predict patient attitude toward personal health records in regional communities. ITP 2018 Apr 03;31(2):306-326. [doi: 10.1108/ITP-07-2016-0160]

XSL•FO
RenderX

64. Yasnoff WA, Shortliffe EH. Lessons learned from a health record bank start-up. Methods Inf Med 2014;53(2):66-72. [doi: 10.3414/ME13-02-0030] [Medline: 24477917]
65. Saeed SA, Masters RM. Disparities in health care and the digital divide. Curr Psychiatry Rep 2021 Jul 23;23(9):61 [FREE Full text] [doi: 10.1007/s11920-021-01274-4] [Medline: 34297202]
66. Chang JE, Lai AY, Gupta A, Nguyen AM, Berry CA, Shelley DR. Rapid transition to telehealth and the digital divide: implications for primary care access and equity in a post-COVID era. Milbank Q 2021 Jun;99(2):340-368 [FREE Full text] [doi: 10.1111/1468-0009.12509] [Medline: 34075622]
67. Kane CK, Gillis K. The use of telemedicine by physicians: still the exception rather than the rule. Health Aff (Millwood) 2018 Dec;37(12):1923-1930. [doi: 10.1377/hlthaff.2018.05077] [Medline: 30633670]
68. McGrail KM, Ahuja MA, Leaver CA. Virtual visits and patient-centered care: results of a patient survey and observational study. J Med Internet Res 2017 May 26;19(5):e177 [FREE Full text] [doi: 10.2196/jmir.7374] [Medline: 28550006]
69. Kuruoglu E, Guldal D, Mevsim V, Gunvar T. Which family physician should I choose? The analytic hierarchy process approach for ranking of criteria in the selection of a family physician. BMC Med Inform Decis Mak 2015 Aug 05;15:63 [FREE Full text] [doi: 10.1186/s12911-015-0183-1] [Medline: 26242399]
70. Khatami F, Shariati M, Khedmat L, Bahmani M. Patients' preferences in selecting family physician in primary health centers: a qualitative-quantitative approach. BMC Fam Pract 2020 Jun 11;21(1):107 [FREE Full text] [doi: 10.1186/s12875-020-01181-2] [Medline: 32527224]
71. KBV - PraxisBarometer 2023: Praxen nutzen deutlich öfter digitale Anwendungen. Kassenärztliche Bundesvereinigung (KBV). URL: https://www.kbv.de/html/1150_67527.php [accessed 2024-02-14]
72. Weik L, Fehring L, Mortsiefer A, Meister S. Big 5 personality traits and individual- and practice-related characteristics as influencing factors of digital maturity in general practices: quantitative web-based survey study. J Med Internet Res 2024 Jan 22;26:e52085 [FREE Full text] [doi: 10.2196/52085] [Medline: 38252468]
73. Faries MD. Why we don't "just do it": understanding the intention-behavior gap in lifestyle medicine. Am J Lifestyle Med 2016;10(5):322-329 [FREE Full text] [doi: 10.1177/1559827616638017] [Medline: 30202289]

Abbreviations

CHERRIES: Checklist for Reporting Results of Internet E-Surveys
PRISMA-RR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Rapid Reviews
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
UTAUT: Unified Theory of Acceptance and Use of Technology

Edited by A Mavragani; submitted 27.03.24; peer-reviewed by DG Goldberg, P Rasche, T Hüer, AL Brecher; comments to author 01.10.24; revised version received 13.10.24; accepted 12.01.25; published 05.03.25.

Please cite as:
Beerbaum J, Robens S, Fehring L, Mortsiefer A, Meister S
Patient Adoption of Digital Use Cases in Family Medicine and a Nuanced Implementation Approach for Family Doctors: Quantitative Web-Based Survey Study
JMIR Form Res 2025;9:e58867
URL: https://formative.jmir.org/2025/1/e58867
doi:10.2196/58867
PMID:40053731

©Julian Beerbaum, Sibylle Robens, Leonard Fehring, Achim Mortsiefer, Sven Meister. Originally published in JMIR Formative Research (https://formative.jmir.org), 05.03.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Feasibility and Links Between Emotions, Physical States, and Eating Behavior in Patients After Metabolic Bariatric Surgery: Experience Sampling Study

Ellen A M Kuipers[1,2], MD; Josien G Timmerman[3], PhD; Marc J van Det[1], MD, PhD; Miriam M R Vollenbroek-Hutten[2,4], PhD

[1]Department of Surgery, Hospital Group Twente, Almelo, The Netherlands

[2]Biomedical Signals and Systems, Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, Enschede, The Netherlands

[3]ZGT Academy, Hospital Group Twente, Almelo, The Netherlands

[4]Board of Directors, Medisch Spectrum Twente, Enschede, The Netherlands

Corresponding Author:
Ellen A M Kuipers, MD
Department of Surgery
Hospital Group Twente
Zilvermeeuw 1
Almelo, 7609 PP
The Netherlands
Phone: 31 620024496
Email: el.kuipers@zgt.nl

Abstract

Background: Lifestyle modification is essential to achieve and maintain successful outcomes after metabolic bariatric surgery (MBS). Emotions, physical states, and contextual factors are considered important determinants of maladaptive eating behavior, emphasizing their significance in understanding and addressing weight management. In this context, experience sampling methodology (ESM) offers promise for measuring lifestyle and behavior in the patient's natural environment. Nevertheless, there is limited research on its feasibility and association among emotions and problematic eating behavior within the population after MBS.

Objective: This study aimed to examine the feasibility of ESM in the population after MBS regarding emotions, physical states, contextual factors, and problematic eating behavior, and to explore the temporal association among these variables.

Methods: An experience sampling study was conducted in which participants rated their current affect (positive and negative), physical states (disgust, boredom, fatigue, and hunger), contextual factors (where, with whom, and doing what), and problematic eating behavior (ie, grazing, dietary relapse, craving, and binge eating) via smartphone-based ESM questionnaires at 6 semirandom times daily for 14 consecutive days. Feasibility was operationalized as the study's participation rate and completion rate, compliance in answering ESM questionnaires, and response rates per day. At the end of the study period, patients reflected on the feasibility of ESM in semistructured interviews. Generalized estimation equations were conducted to examine the temporal association between emotions, physical states, contextual factors, and problematic eating behavior.

Results: In total, 25 out of 242 participants consented to participate, resulting in a study participation rate of 10.3%. The completion rate was 83%. Overall compliance was 57.4% (1072/1868), varying from 13% (11/84) to 89% (75/84) per participant. Total response rates per day decreased from 65% (90/138) to 52% (67/130) over the 14-day study period. According to the interviews, ESM was considered feasible and of added value. Temporal associations were found for hunger and craving (odds ratio 1.04, 95% CI 1.00-1.07; P=.03), and for positive affect and grazing (odds ratio 1.61, 95% CI 1.03-2.51; P=.04).

Conclusions: In this exploratory study, patients after MBS were not amenable to participate. Only a small number of patients were willing to participate. However, those who participated found it feasible and expressed satisfaction with it. Temporal associations were identified between hunger and craving, as well as between positive affect and grazing. However, no clear patterns were observed among emotions, physical states, context, and problematic eating behaviors.

(JMIR Form Res 2025;9:e60486) doi:10.2196/60486

XSL•FO
RenderX

KEYWORDS

feasibility; experience sampling methodology; metabolic bariatric surgery; eating behavior; positive and negative affect; physical states; contextual factors; mobile phone

Introduction

Metabolic bariatric surgery (MBS) is the most effective long-term strategy in the management of severe obesity and its related comorbidities [1]. Lifestyle modification, including nutritional compliance and physical activity, is essential to achieve and maintain optimal weight loss results and to improve obesity-related comorbidities after MBS [2-4]. Poor dietary adherence is significantly associated with recurrent weight gain after MBS [1,5,6]. Furthermore, maladaptive eating behaviors may cause nutritional noncompliance with the inability to remain consequently adherent to an intended diet (ie, dietary relapse) and thereby negatively impact postsurgical weight loss outcomes [7-10]. Most theoretical models of eating disorders posit that negative emotions and numerous physical states (ie, hunger and fatigue) contribute to the development and maintenance of maladaptive eating behaviors [11-17]. However, little is known whether these theoretical constructs apply to individuals who have undergone MBS. To our best knowledge, only 1 study has explored the relationship between negative affect and loss of control (LOC) eating in patients who underwent MBS, finding that higher momentary negative affect predicted more severe LOC eating both before and after MBS [18]. Since this study focused on 1 aspect of disordered eating behavior, it is important to enhance our understanding of how emotions, physical states, and other maladaptive eating behaviors are interconnected in this specific population. This is particularly important given the high prevalence of problematic eating behaviors in the population with obesity, including grazing, food cravings, and binge eating [7,19-22]. Grazing, defined as unplanned and repetitive eating of small amounts of food between meals and snacks [23], is reported to occur in up to 26% of patients seeking MBS [7]. Postoperatively, the prevalence is even higher, with rates reported as high as 47% [8]. Food cravings, characterized by an intense desire to consume specific foods, are experienced by nearly all patients after MBS, with only about 10% reporting no cravings [24]. Binge eating involves the consumption of excessive amounts of food within a discrete period of time, accompanied by an LOC feeling during the binge eating episode [9]. Binge eating, which is present in 24% of patients before MBS [25], is strongly associated with obesity and negatively impacts weight outcomes after MBS [8].

Since lifestyle and lifestyle modification after MBS takes place primarily outside of medical visits and clinical settings, there is an urgent need to have ecologically valid methods that capture daily life to better understand and change behaviors. Experience sampling methodology (ESM), also known as ecological momentary assessment, addresses this need by collecting real-time data on individuals' behavior, thoughts, emotions, and environmental contexts as they go about their daily lives [26]. Unlike retrospective questionnaires, ESM prompts participants at random or scheduled times throughout the day to record their current state or activities. This approach provides a more accurate representation of behavior within the individual's natural environment, capturing their typical routines in daily life [27], while traditional self-reports may miss crucial details about how and why a participant's experiences or behaviors change over time and context.

In the past years, an increasing number of experience sampling studies have been conducted in various health domains, representing a wide range of diseases and populations [17,28-37]. However, the use of ESM to evaluate (disordered) eating behaviors following MBS remains limited [18,38,39]. Previous research has focused on specific aspects, such as 1 type of problematic eating behavior [18], eating practices that do not align with postoperative dietary or eating behavior recommendations (eg, drinking liquids with meals and not starting meals with protein) [39], or elements of eating regulation (eg, dietary restraint and eating in the absence of hunger) [38]. Our goal is to conduct a more comprehensive assessment by examining a broader range of problematic eating behaviors and their influencing factors—an important objective given the high prevalence of disordered eating behavior in the population after MBS [20], as outlined earlier in this section.

The feasibility of ESM has been previously evaluated in various populations, including its use to monitor dietary intake and physical activity among Dutch vocational education students [28], to study depressive symptoms among caregivers of patients with cancer [31], and to examine daily life experiences and activities in patients with acute coronary syndrome [33]. These studies show varying results regarding compliance, participation rates, and completion rates, indicating that feasibility for ESM differs between populations. However, the feasibility of ESM has scarcely been studied in the population that has undergone or is undergoing MBS. To our knowledge, only 1 experience sampling feasibility study was conducted in adults undergoing MBS, focusing on the use of a smartphone app for lifestyle education before surgery [40]. This study reported high satisfaction and perceived helpfulness, but the study was carried out in patients before MBS. After MBS, patients are dealing with physical and mental changes, along with the need to modify their behavior. This may impact the feasibility of ESM, as patients are likely to prioritize managing these challenges over using an experience sampling app for monitoring. Although some studies have used ESM to evaluate (disordered) eating behaviors after MBS [18,38,39,41], they did not specifically focus on evaluating feasibility. While these studies reported compliance, they did not assess other feasibility metrics. Notably, participants in these studies received monetary compensation [38,39,41], which may have positively influenced participation and compliance, potentially affecting feasibility outcomes [42]. This raises concerns about the generalizability of these findings to the real-world clinical setting.

Therefore, the primary aim of this exploratory study is to establish the feasibility of ESM among patients shortly after MBS, focusing on maladaptive eating behaviors, emotions, physical states, and context. The secondary aim is to explore

XSL•FO
RenderX

associations between emotions, physical states, context, and maladaptive eating behaviors in the population after MBS using experience sampling measurements.

Methods

Study Design and Participants

A prospective observational cohort study was performed at the obesity center of Hospital Group Twente (ZGT) Almelo/Hengelo, the Netherlands. All patients who underwent MBS at ZGT during the inclusion period (from February 2022 to July 2022) were considered for participation in this study. We included patients who had recently undergone MBS, as lifestyle modification mostly occurs after surgery, making this period particularly relevant for monitoring to enable timely intervention. The inclusion and exclusion criteria are detailed in Textbox 1.

Textbox 1. Inclusion and exclusion criteria.

Inclusion criteria

- Internet access at home
- Good comprehension of the Dutch language

Exclusion criteria

- No possession of a smartphone that supports the experience sampling methodology app
- Participation in an intervention program or behavioral treatment that is not included in the regular care pathway after metabolic bariatric surgery
- People who worked on night shifts (as it is impossible to complete the experience sampling questionnaires during the day)

Study Procedure

Recruitment

Patients were initially recruited through flyers sent by post and by email 5 weeks after surgery. Due to a minimal response, the recruitment strategy was expanded; patients were approached following a postoperative group meeting or patients were contacted by telephone. Recruitment proceeded as follows:

1. Via flyer by post or email: Participant information was provided to the patients by scanning the QR code on the flyer or clicking on the link. If the patient was interested in participating after reading the information, the patient could send an email to the first author (EAMK). Later, individuals were contacted via telephone by the first author (EAMK) or a student (Charlotte van den Berg) to provide verbal information about the study.
2. During a postoperative group meeting: A student (Charlotte van den Berg) provided verbal and written information about the study. If interested, patients shared their contact details for inclusion. Furthermore, a subset of these patients was contacted by phone to encourage participation.
3. By telephone: Postoperative patients were contacted by telephone by a student (Charlotte van den Berg) or by the first author (EAMK) to inform patients about the study and motivate them to participate.

After inclusion, participants received verbal instructions on installing and using the ESM app. A user manual for the ESM app was sent via email. The participants used their own email addresses to create an account in the ESM app Ethica Data. After entering a unique registration code to join the study, participants were briefed on the study, and at this point, participants signed informed consent in the Ethica Data app. Patients who consented were asked if they would like to participate in an interview at the end of the study period to discuss the feasibility of the ESM app.

Questionnaires and Interviews

Participants were instructed to answer ESM questionnaires for 14 consecutive days, to capture both weekdays and weekends. The app was programmed to send auditory and visual notifications at 6 semirandom times throughout the day (1 prompt every 135 min, between 8:00 AM and 9:30 PM), signaling the availability of a questionnaire (signal contingent). At each prompt, participants answered questions regarding current positive and negative affect, physical states, context, and problematic eating behavior. A delayed response triggered a 1-time reminder after 5 minutes, and each questionnaire expired after 15 minutes to minimize recall bias. Participants were asked to carry their smartphones throughout the day and continue their daily lives as usual. Participants who consented to the interview were contacted by phone to discuss the feasibility of the ESM app.

Data Collection

Baseline Measures

Demographics and surgical data were obtained from electronic patient records. Collected patient characteristics included age; gender; type of surgery; time since surgery; and obesity-related comorbidities, such as hypertension, diabetes mellitus, dyslipidemia, osteoarthrosis, gastroesophageal reflux disease, and obstructive sleep apnea syndrome. At the start of the study, each participant was asked to enter their current weight and height into the Ethica Data app.

Feasibility

Feasibility was operationalized as participation rate, completion rate, compliance, and response rate per day.

- The participation rate was calculated as the percentage of approached patients willing to participate in the study.
- The completion rate represented the percentage of participants who filled out at least 1 ESM questionnaire on the 14th day of the study period.

XSL•FO
RenderX

- Compliance with experience sampling was calculated as the percentage of answered ESM questionnaires out of all issued ESM questionnaires.
- The response rate per day was the percentage of ESM questionnaires answered per day, calculated for each day of the study period.

At the end of the 2-week period, participants were invited to participate in a semistructured interview to share their experiences. The interviews were conducted by the first author (EAMK) and a master's student in health science (Charlotte van den Berg) via telephone in November and December 2022. The interviews, guided by a semistructured interview guide (Multimedia Appendix 1), lasted between 15 and 30 minutes. This guide covered the feasibility themes to be discussed, including user-friendliness, technical performance, and content, along with optional questions. With participants' permission, the interviews were audio-recorded, transcribed verbatim, and anonymized.

ESM Measures

ESM data were collected using the Ethica Data app. All ESM questions were provided in the patients' native language (Dutch). ESM data included the following items:

1. Emotion: Participants rated their feelings as "angry or annoyed," "anxious or scared," "relaxed or calm," "cheerful or happy," "sad or gloomy," and "tensed or stressed" [43,44]. These items reflect both positive and negative affect and were scored on a visual analog scale from 0 (not at all or very bad) to 10 (extremely or very good).
2. Physical states: Disgust, boredom, fatigue, and hunger were rated. The first ESM questionnaire in the morning also asked participants to rate their sleep quality. Physical state items were scored on a visual analog scale from 0 to 10.
3. Context and activities: Questions about the context (where, with whom) and activities (doing what) had multiple-choice response options.
4. Problematic eating behavior: Participants answered 4 questions about problematic eating behavior. The questions and answer options are presented in Table 1.

Table 1. Experience sampling questions about problematic eating behavior and their answer options.

Behavior	Question	Answer option
Craving	"Have you had a craving for a specific food in the past 30 minutes? If yes, how strong was this desire?"	VAS[a]: 0-10
Dietary relapse	"In the past 60 minutes, did you eat and/or drink anything that is likely to negatively impact your weight?"	Yes or no
Binge eating	"Have you eaten unusually large amounts of unhealthy food in the past 60 minutes?"	Yes or no
Grazing	"In the past 60 minutes, have you eaten unplanned and repetitive small amounts of food outside of planned meals and snacks?"	Yes or no

[a]VAS: visual analog scale.

Statistical Analysis

Statistical analysis was performed using SPSS (version 24; IBM Corp). Patient characteristics were expressed as means with SDs for parametric continuous data or frequencies and percentages for categorical data. For nonparametric continuous data, median and IQR were reported. Feasibility metrics were expressed as percentages.

To investigate the relationship between emotions, physical states, context, and problematic eating behaviors, generalized estimation equations (GEEs) were used. GEEs are an extension of generalized linear models designed to account for the correlation between repeated measures within participants, as data collected through ESM involve multiple observations per individual [45]. This makes GEEs the appropriate choice for our analysis, as they allow us to model repeated measures data while adjusting for within-subject correlations without assuming that these correlations are the same across all participants.

We used a negative binomial distribution with a binary logistic link function, as our dependent variables (eg, dietary relapse and grazing) were dichotomous (yes or no) and did not follow a normal distribution. The negative binomial distribution is particularly suitable for overdispersed count data, which are often encountered in studies involving behavioral data. An exchangeable correlation structure was selected, assuming a constant correlation between repeated measures within individuals over time. While centering is customary in ESM studies, in this study, not centering the variables aligns better with the interpretability of the results. Retaining the raw values ensures that the intercept is directly relatable to real-world states (eg, no negative or positive emotion), which facilitates reader comprehension without altering the estimated effects.

Momentary assessments (level 1) were nested within participants (level 2). Momentary levels of positive affect were calculated by averaging the scores of the "relaxed or calm" and "cheerful or happy" items, while momentary levels of negative affect were calculated by averaging the "angry or annoyed", "anxious or scared", "sad or gloomy" and "tensed or stressed" items. Contextual factors were recoded as dichotomous variables such as "being at home' (yes or no), "being alone" (yes or no), and "relaxing" (yes or no), where "relaxing" encompassed resting or doing nothing. Due to nonnormality, craving was dichotomized (yes or no).

To account for potential temporal dependencies, we lagged the independent variables (emotions, physical states, and contextual factors) by 1 time point, meaning that the predictors at time t–1 were used to examine their influence on problematic eating behavior at time t. The first assessment of each day was

XSL•FO
RenderX

excluded from the analysis to avoid carry-over effects from the previous day.

A minimum response rate of 33% for ESM data was applied, which is the generally accepted threshold for ESM studies [46]. Univariate GEE analyses were conducted for each predictor variable, and given the exploratory nature of the study, we did not adjust for multiple comparisons. The significance level was set at $P<.05$.

Thematic Analysis

Using the software program ATLAS.ti (version 22; Lumivero), the interview transcripts were reviewed and analyzed thematically by the first author. Initially, each transcript was read multiple times without conducting a thematic analysis. Next, comments related to feasibility issues were extracted from the transcripts without interpretation or categorization. These feasibility issues were then classified deductively according to the 7 categories of the eHealth Usability Benchmarking Instrument (HUBBI), namely, basic system performance, task-technology fit, interface design, navigation and structure, information and terminology, guidance and support, and satisfaction [47]. Finally, transcripts were reread to ensure that all relevant codes were identified.

Ethical Considerations

The study protocol was reviewed and approved by the local medical ethics committee (local registration 2020-48), and all patients provided informed consent. Data collected via the Ethica Data app were securely stored on Ethica Data's servers. Ethica Data employs 2 separate databases located on 2 different servers. Participant registration data (ie, name and email) are stored on one server, while study data (ie, questionnaire responses) are stored on another. Study data are stored in a coded format and cannot be traced back to individual participants. The study data stored on Ethica Data's server were accessible only to researchers with a password-protected account and could be imported by the researchers to ZGT's server. The dataset contains only participant numbers, and the combination of participant numbers and email addresses was securely stored in a password-protected file within the research department. Only authorized persons had access to these data. Audio recordings of interviews were used to generate deidentified transcripts, after which the recordings were immediately destroyed. Both the audio recordings and transcripts were securely stored on a protected hard drive. Participants with a minimum response rate of 80% on the ESM questionnaires had the chance to win a €50 voucher (US $52.32), which was randomly awarded to 1 participant. No additional monetary compensation was provided.

Results

Participants

From February to July 2022, a total of 242 patients were informed about the study. Out of the 242 patients who were approached to participate in the ESM study, 25 patients consented to participate, resulting in a participation rate of 10.3%. Figure 1 presents the study flowchart.

Figure 1. Study flowchart of patients approached for participation between February and July 2022. ESM: experience sampling methodology; GEE: generalized estimation equation.

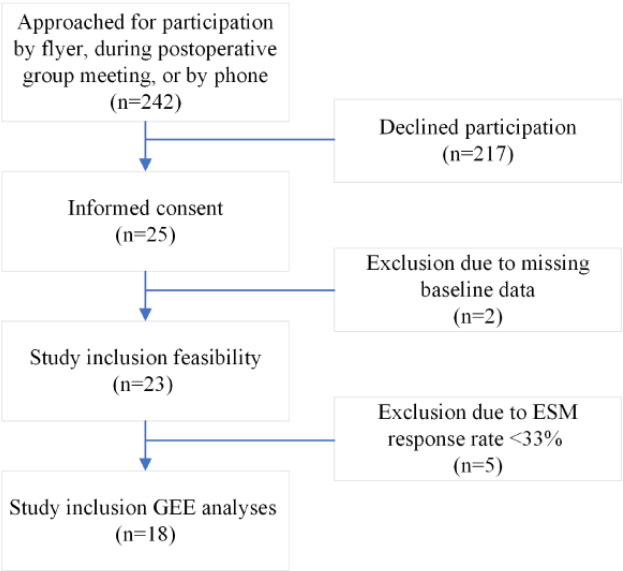

Two participants were excluded from analyses due to missing baseline data. The remaining 23 participants were included to answer the primary aim of the study. Although 2 participants withdrew due to a high perceived burden, they were included in the feasibility analysis. For the secondary aim of this study, an additional 5 participants were excluded due to a response rate of less than 33% for all ESM questionnaires. Thus, a total of 18 participants were included in this analysis.

Demographics

The 23 participants had a mean age of 47 (SD 11) years, and 17 (74%) were female. The median time since surgery was 36 (IQR 31-88) days. At inclusion, median BMI was 36 (IQR 33-39) kg/m^2. The most performed bariatric procedure was one-anastomosis gastric bypass (n=16, 70%). The most prevalent obesity-related comorbidities were dyslipidemia (n=21, 91%)

XSL•FO
RenderX

and osteoarthrosis (n=13, 57%). Descriptive statistics of the participants are shown in Table 2.

Table 2. Participant characteristics for the total study population, as well as for those with a response rate ≥33% and <33% on the experience sampling questionnaires.

Participant characteristic	Total (n=23)	Response rate ≥33% (n=18)	Response rate <33% (n=5)
Age (years), mean (SD)	47 (11)	49 (12)	42 (9)
Sex, n (%)			
Female	17 (74)	13 (72)	4 (80)
Male	6 (26)	5 (28)	1 (20)
BMI (kg/m^2), median (IQR)	36 (33-39)	36 (33-38)	39 (32-42)
Days between surgery and study start, median (IQR)	36 (31-88)	34 (28-79)	88 (60-103)
Comorbidities, n (%)			
T2DM[a]	3 (13)	2 (11)	1 (20)
Hypertension	10 (44)	9 (50)	1 (20)
OSAS[b]	8 (35)	7 (39)	1 (20)
Osteoarthrosis	13 (57)	10 (56)	3 (60)
GERD[c]	10 (44)	8 (44)	2 (40)
Dyslipidemia	21 (91)	17 (94)	4 (80)
Type of surgery, n (%)			
SG[d]	1 (4)	0 (0)	1 (20)
RYGB[e]	3 (13)	3 (17)	0 (0)
OAGB[f]	16 (70)	13 (72)	3 (60)
OAGB+correction HH[g]	2 (9)	1 (6)	1 (20)
OAGB+cholecystectomy	1 (4)	1 (6)	0 (0)

[a]T2DM: type 2 diabetes mellitus.
[b]OSAS: obstructive sleep apnea syndrome.
[c]GERD: gastroesophageal reflux disease.
[d]SG: sleeve gastrectomy.
[e]RYGB: Roux-en-Y gastric bypass.
[f]OAGB: one-anastomosis gastric bypass.
[g]HH: Hiatal hernia.

Feasibility

Overview

The compliance per participant ranged from 13% (11/84 questionnaires completed) to 89% (75/84 questionnaires completed). Overall, participants responded to 1072 out of 1868 prompts, resulting in an overall compliance of 57.4%. The compliance in answering experience sampling questionnaires per participant is presented in Multimedia Appendix 2.

The mean number of answered ESM questionnaires was 47 (SD 21). Total response rates were calculated from day 1 to day 14, as illustrated in Figure 2. Over time, the total response rate declined by 14%. The study completion rate was 83% (19/23).

XSL•FO
RenderX

Figure 2. Mean response rates per day and SD for the total study population (n=23) over 14 days of study participation.

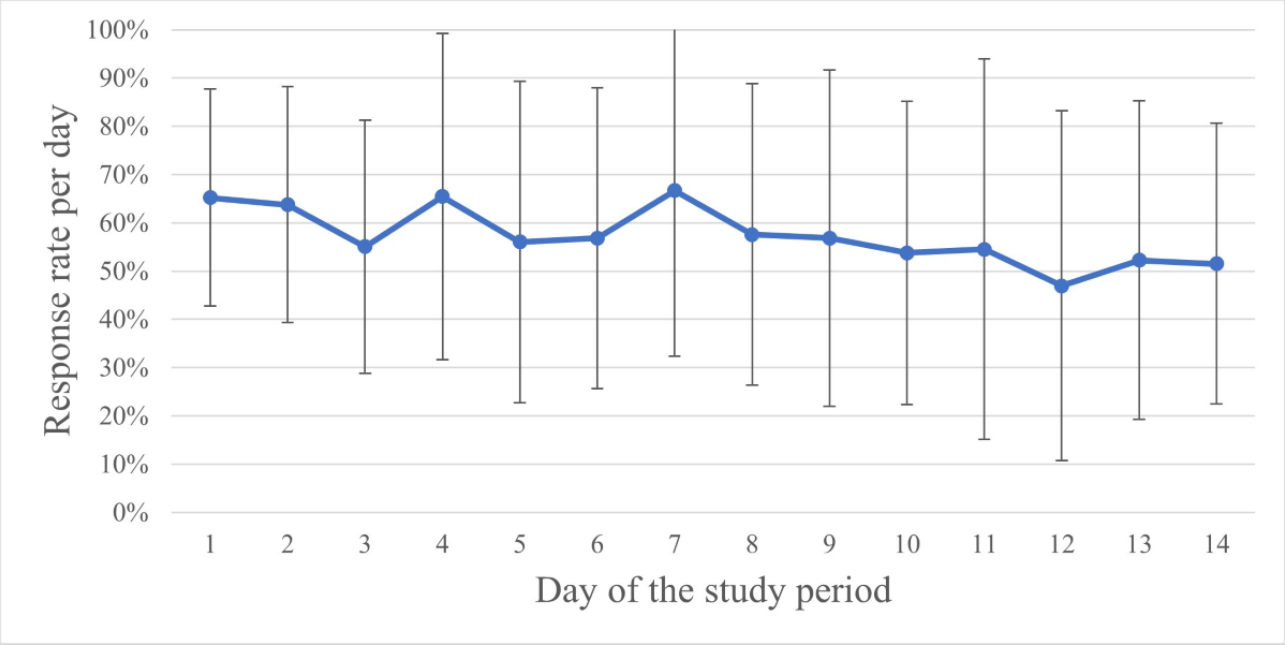

Semistructured Interviews

Overview

In total, 10 patients participated in the semistructured interviews, of whom 7 (70%) were female. The mean age was 46 (range 28-64) years, and the mean BMI was 38 (SD 6) kg/m^2. The median time between surgery and participation in the ESM study was 34 (IQR 28-87) days.

The categories that were indicated in the interviews, based on the HUBBI, are basic system performance, task-technology fit, information and terminology, guidance and support, and satisfaction [47].

Basis System Performance

The participants reported that the app was easy to use and straightforward to install on their mobile phones, with no technical issues during installation. When asked about technical performance in general, 1 participant reported repeated technical problems that prevented her from completing several questionnaires. In addition, she reported not receiving auditory notifications and sometimes inconsistent visual notifications.

Task-Technology Fit

Most participants reported they could not complete all questionnaires within the 15-minute time frame due to work or travel commitments. Participants perceived the number of questions per questionnaire as appropriate and found them easy to answer since the questionnaires were nearly identical. The daily number of questionnaires per day was also considered manageable, with 1 participant stating:

> *Not too much, not too little [the number of questionnaires].* [Participant 6]

Another participant felt that 6 questionnaires per day was the maximum, saying:

> *Otherwise, you spend so much time working on it [the questionnaires], I don't think that's healthy.* [Participant 7]

Furthermore, 3 participants shared their views on the 2-week survey duration, and 2 of them indicated that a longer duration would have been acceptable, with 1 noting:

> *If it's not going well immediately [after MBS], it [filling in the questionnaires] can provide insights, because it's quite astonishing how you change after bariatric surgery.* [Participant 6]

In addition, 7 participants indicated that using the app encouraged them to reflect on their behavior, providing insights and increasing awareness. One participant found the questions negatively framed, which led to his reluctance to use the app, ultimately causing him to withdraw from the study. Another participant felt she was doing very well at that moment. Therefore, she experienced no additional support by using the app.

Information and Terminology

The participants found the questions clear, though 5 participants noted that the questions were repetitive. One participant commented:

> *Maybe I expected different questions, as not much has changed in such a short time [time since previous questionnaire].* [Participant 10]

Another participant suggested occasionally changing the order of the questions:

> *This way, you have to think carefully about which question to answer.* [Participant 2]

Furthermore, 2 participants would have preferred open questions as well:

> *Imagine you experience a certain emotion or that you ate something 'wrong', what was that about? If you*

XSL•FO
RenderX

> *could explain a little more, it would provide greater depth.* [Participant 10]

As mentioned above, 1 participant indicated that the questions were negatively framed, feeling that the negative aspects overshadowed questions about positive emotions.

Guidance and Support

The participants were not bothered about lacking insight into their own answers, understanding it was for scientific research. As one stated:

> *It is especially important for you as researchers, I don't necessarily have to see it myself [the answers].* [Participant 3]

However, 5 participants felt that having access to their answers could be valuable. One of them mentioned:

> *It might be interesting to see if there were any changes or patterns.* [Participant 10]

Another remarked:

> *On one hand, I see the benefit, on the other, it could raise many questions. Ultimately, if things aren't going well, you're the one who must seek help.* [Participant 6]

Overall, participants were positive about using the app in bariatric care, as it could help individuals become more aware of their actions:

> *I can imagine that some people have had surgery and then feel a little lost. Of course, you have group meetings with the dietician, but the application allows for personal reflection in a quieter moment.* [Participant 5]

In total, 4 participants indicated that using the app before MBS could be interesting:

> *Are people aware of their emotions and their eating behavior before surgery?* [Participant 2]

Participants expressed mixed opinions on how an app should be integrated into follow-up care after MBS. One suggested that questions about changes in eating patterns, the creeping in of old habits, physical activity, and mental well-being may be relevant 1 or 2 years after MBS. Another participant expressed interest in questions about nutrition and related uncertainties. One participant preferred an app where she could ask questions directly to a health care professional. Participants recognized the need to tailor the app to their specific stage after surgery. While most participants preferred using the app in the short term (within a few months) after MBS, 1 participant preferred using the app on a longer-term (at least 6 months) after MBS. He indicated that, in the short term after surgery, he would have benefited more from an app that provided meal or snack suggestions, and calculated protein intake based on those suggestions.

Satisfaction

In general, participants were positive about the app, finding it helpful in the follow-up after MBS. Most participants experienced personal benefits from using the app, noting that it encouraged more conscious eating by prompting questions at random times:

> *What did I actually eat or drink and how do I feel about that?* [Participant 2]

One participant experienced a personal disadvantage by using the app:

> *I remember thinking that if I continue this [the questionnaires], it will even make me depressed. I don't think that's the intention.* [Participant 9]

For improvement, participants suggested extending the time a questionnaire remains open and adding the question:

> *What do all these questions do to you?* [Participant 2]

If the answer is negative, it was recommended that the app alert the health care professional:

> *I think you should contact this person, because he can't figure it out on his own.* [Participant 2]

Association Between Positive and Negative Affect, Contextual Factors, and Problematic Eating Behavior

Frequencies and Percentages of Problematic Eating Behaviors

Frequencies and percentages of the different types of problematic eating behavior are illustrated in Multimedia Appendix 3.

- Dietary relapse: In total, 6 participants reported no dietary relapse over 14 days. The remaining participants reported a total of 45 dietary relapses, with individual counts ranging from 1 to 8.
- Grazing: Grazing occurred in total 13 times among 7 participants (range per participant: 1-3), while 11 participants reported no grazing during the study period.
- Binge eating: A total of 7 binge eating episodes were reported by 3 participants (range per participant: 2-3), while 15 participants reported no binge eating episodes during the study period.
- Craving: Craving was absent (score 0) in 65% (643/994) of all completed questionnaires. In addition, 3 participants consistently reported no craving throughout the study, while the remaining 15 participants experienced a total of 351 craving episodes (range per participant: 3-53).

GEE Analyses

Data from the ESM questionnaires were nested within 18 individuals. The number of observations per predictor or outcome variable ranged from 987 to 1005. After lagging each predictor and setting the first prompt of each day to missing, the number of observations available for the GEE analyses ranged from 594 to 605. Table 3 summarizes the GEE results, showing the temporal associations between lagged predictors (ie, positive and negative affect, physical states, and contextual factors) and problematic eating behaviors (craving, binge eating, dietary relapse, and grazing). Significant associations are described below, while nonsignificant trends are noted for exploratory purposes.

XSL•FO
RenderX

- Craving: Hunger was significantly associated with craving (OR 1.04, 95% CI 1.00-1.07; *P*=.03).
- Binge eating: The odds of binge eating increased by 1.61 times with each unit increase in negative affect (95% CI 0.77-3.34; *P*=.21) and decreased by 0.51 times for each unit increase in disgust (95% CI 0.20-1.28; *P*=.15). When at home, the odds of binge eating decreased by 0.27 times compared with not being at home (95% CI 0.03-2.67; *P*=.26).
- Dietary relapse: The odds of dietary relapse decreased by 0.40 times for being at home (95% CI 0.14-1.17; *P*=.09) and by 0.67 times for each unit increase in boredom (95% CI 0.40-1.14; *P*=.14).
- Grazing: A significant temporal association was found for positive affect and grazing, with the odds of grazing increasing by 1.61 times for each unit increase in positive affect (95% CI 1.03-2.51; *P*=.04). The odds of grazing decreased by 0.68 times with each unit increase in negative affect (95% CI 0.40-1.15; *P*=.15). When at home, the odds of grazing decreased by 0.16 times compared with not being at home (95% CI 0.02-1.07), with this result approaching statistical significance (*P*=.06).

XSL•FO
RenderX

Table 3. Generalized estimation equation analyses showing the temporal associations between lagged predictors (positive and negative affect, physical states, and contextual factors) and subsequent problematic eating behaviors (craving, binge eating, dietary relapse, and grazing) in patients shortly after metabolic bariatric surgery.

Eating behavior and predictor	Odds ratio (95% CI)	*P* value
Craving		
Negative affect	1.06 (0.79-1.42)	.69
Positive affect	1.02 (0.91-1.15)	.70
Boredom	1.03 (0.89-1.20)	.70
Disgust	1.01 (0.82-1.25)	.91
Fatigue	1.07 (0.96-1.18)	.21
Hunger	1.04 (1.00-1.07)	.03
Being alone	1.01 (0.87-1.17)	.92
Being at home	1.15 (0.93-1.42)	.21
Relaxing	0.93 (0.78-1.09)	.36
Binge eating		
Negative affect	1.61 (0.77-3.34)	.21
Positive affect	0.95 (0.53-1.69)	.86
Boredom	0.93 (0.57-1.50)	.76
Disgust	0.51 (0.20-1.28)	.15
Fatigue	1.10 (0.81-1.50)	.55
Hunger	1.00 (0.68-1.48)	≥.99
Being alone	0.80 (0.11-5.81)	.83
Being at home	0.27 (0.03-2.67)	.26
Relaxing	0.80 (0.17-3.91)	.79
Dietary relapse		
Negative affect	0.84 (0.46-1.52)	.56
Positive affect	1.15 (0.91-1.46)	.23
Boredom	0.67 (0.40-1.14)	.14
Disgust	0.89 (0.59-1.36)	.59
Fatigue	1.14 (0.90-1.45)	.28
Hunger	0.89 (0.72-1.10)	.28
Being alone	0.82 (0.35-1.91)	.65
Being at home	0.40 (0.14-1.17)	.09
Relaxing	0.97 (0.30-3.20)	.97
Grazing		
Negative affect	0.68 (0.40-1.15)	.15
Positive affect	1.61 (1.03-2.51)	.04
Boredom	0.81 (0.43-1.55)	.53
Disgust	0.88 (0.57-1.38)	.59
Fatigue	1.23 (1.00-1.51)	.06
Hunger	0.77 (0.55-1.09)	.15
Being alone	1.14 (0.33-3.94)	.83
Being at home	0.16 (0.02-1.07)	.06
Relaxing	0.90 (0.14-5.97)	.91

XSL•FO
RenderX

Discussion

Principal Results

This exploratory study describes the feasibility of ESM in patients after MBS, focusing on negative and positive affect, physical states, contextual factors, and problematic eating behaviors. The results indicate that patients shortly after MBS may be less inclined to participate in experience sampling studies. Nonetheless, actual usage was considered feasible, given the moderate compliance and high completion rate upon patient participation. In addition, results from the semistructured interviews indicated that the ESM app used in this study showed to be usable and of added value from the patient's perspective. This indicates that patients who chose to participate were motivated to complete the ESM questionnaires. Concerning the second research aim, temporal associations were found, but only between hunger and craving, and between positive affect and grazing.

Comparison With Previous Work

Our study faced a low participation rate of 10.3%. Several participants who were approached face-to-face and declined participation, mentioned the high expected burden, or indicated that study participation would likely not fit into their daily life or work schedule. Ratcliff et al [48] performed an experience sampling feasibility study to assess the adherence to postoperative diet and activity patterns in adolescents who underwent MBS. Their participation rate was 62%, which is remarkably higher than ours. However, they included participants at least 12 months after MBS, while we approached patients eligible for participation a few weeks after MBS. This timing difference may have influenced the willingness to participate, considering the overwhelming changes patients are dealing with in the first weeks after surgery. Furthermore, in the study by Ratcliff et al [48], participants completed only 3 daily phone diaries (a form of ESM) instead of 14 days of measurements. As we were interested in the feasibility of ESM, we opted for an intensive sampling strategy to gain insight into the daily response rates throughout the study period. This may have negatively affected our participation rate because of the high expected burden shortly following MBS. In addition, to obtain more realistic feasibility metrics, patients were not incentivized with monetary remuneration, except for 1 voucher randomly awarded among participants with a minimum response rate of 80%. This approach offers insights into the effectiveness of ESM in clinical practice without the use of financial incentives but might have further contributed to a low participation rate. Finally, we recruited patients through flyers sent by post and email, as well as during group meetings. An individualized or personalized recruitment strategy could potentially increase participation rates.

Our compliance rate was 56%, while previous experience sampling studies regarding affect and problematic eating behaviors showed higher compliance rates ranging from 62% to 81% [17,44,49,50]. Interestingly, compliance rates were higher (62%) in the studies by Stevenson et al [50] and Boh et al [44] (80% for participants with healthy weight and 81% for participants who were overweight), despite both studies administering more daily ESM questionnaires over a similar study duration compared with ours. First, differences in study populations may account for the discrepancy in compliance rates. For instance, Boh et al [44] included participants who were both healthy-weight and overweight, but none of these participants had recently undergone MBS. MBS entails associated changes and challenges, particularly in the short term following surgery. This might have played a role in the lower compliance rates observed in our study. Participants in the study by Stevenson et al [50] were younger compared with our participants (age 24 years versus 47 years in our study). This could potentially explain their higher compliance rates, as mobile apps for self-management are known to be used less frequently by older adults compared with younger individuals [51,52]. Second, to minimize recall bias, we restricted the time window for participants to respond to ESM questionnaires to 15 minutes. Several participants mentioned in the semistructured interviews that this short-time window hindered their ability to respond, possibly contributing to our lower compliance rate. However, other ESM studies with longer response windows have reported compliance rates that are either lower or similar to ours [28,31], whereas others with similar time windows achieved higher compliance rates [33]. This indicates that extending the time window or increasing reminder frequency, is unlikely to positively impact compliance within our study population.

This study is one of the few to assess problematic eating behaviors using ESM in the population after MBS [18,38,39]. While different methods exist for assessing problematic eating behaviors, we used terminology consistent with previous research in populations with overweight or obesity [53,54]. Our study identified a significant temporal association between hunger and craving (P=.03). However, an OR close to 1 (OR 1.04) indicates a minimal effect of hunger on craving. Previous studies also found that food cravings closely follow hunger [55,56]. Given that food craving is defined as an intense desire to consume a specific food, and hunger as an urgent need for food or a specific nutrient, it is unsurprising that both constructs are closely intertwined. This overlap might create difficulties for participants to differentiate between both constructs and therefore, an association between hunger and craving could have arisen. Our results also revealed a significant temporal association between positive affect and grazing (OR 1.61, 95% CI 1.03-2.51; P=.04), where grazing behaviors increased with higher levels of positive affect. As grazing behaviors were not specifically assessed in previous studies, direct comparisons with existing literature are not possible. However, practical implications arise from the identified associations. For example, recognizing a positive mood might help patients in acknowledging the potential risk of grazing behavior.

Overall, our study yielded inconclusive results regarding the association between negative emotions, positive emotions, and problematic eating behaviors, as no clear patterns emerged from the GEE analyses. ORs regarding problematic eating behaviors for both positive and negative emotions (including boredom and disgust) were notably ambiguous. This is unexpected, as most theoretical models of eating disorders propose that negative emotions trigger episodes of problematic eating behavior [13].

XSL•FO
RenderX

It is likely that the impact of MBS, which encompasses both mental and physical adjustments, might have influenced the inconsistencies in positive and negative affect on problematic eating behaviors in our study. Future studies focusing on the longer term after MBS are needed to assess the association between both positive and negative affect and problematic eating behaviors. Furthermore, participants in our study exhibited fewer episodes of problematic eating behaviors compared with existing literature [57], resulting in a low number of events, which likely impacted the robustness of our analyses. Consequently, the limited frequency of problematic eating episodes observed in our sample, combined with potential influences of post-MBS adaptations, may have contributed to the inconclusive findings regarding the role of emotions in problematic eating behaviors.

Study Strengths and Limitations

As we were interested in the feasibility of ESM in the population after MBS, we used an intensive sampling strategy over 14 consecutive days with 6 prompts per day, which is a strength of our study. Furthermore, we collected both quantitative and qualitative data to answer our primary research aim, which increased the validity of the study.

There are a few limitations to this study, so the results should be interpreted with caution. First, the relatively small sample size is a limitation of our study. However, the small sample size was justifiable, due to the large amount of collected data by intensive sampling, and to the exploratory nature of the study. Despite the small sample size, we add to the limited knowledge about the feasibility of ESM in the population after MBS. Second, we conducted multiple tests without correction, which may increase the risk of type I statistical errors, that is, finding significant results while there are no true underlying associations. In exploratory studies, the primary goal is to identify potential relationships, making it more appropriate to forego corrections. This approach allows for the identification of possible associations that can be further examined and validated in follow-up confirmatory studies. Third, the overall frequency of problematic eating behaviors in our study population was extremely low. To minimize recall bias, we asked participants about problematic eating behaviors occurring within the past 30 or 60 minutes, which may have missed behaviors occurring outside this time frame. Wider time windows could probably overcome this limitation, but this goes hand in hand with recall bias. Furthermore, behavior can change when monitored, even without an intended change, a phenomenon known as “measurement reactivity,” which may have increased participants’ awareness of their eating behaviors and led them to adjust their responses or actions accordingly. Together, this may have resulted in underreporting of problematic eating behaviors in our study. Finally, our assessment of binge eating focused solely on the quantity of food consumed, without evaluating LOC. Previous research highlights that LOC may be a stronger indicator of binge eating episodes, particularly in patients with obesity and patients undergoing MBS [58,59]. By focusing on quantity alone, we may have misclassified certain episodes or overlooked critical associations related to LOC. However, it is important to note that our primary aim was to assess the feasibility of ESM, which is not directly impacted by this limitation. For our secondary aim, the small number of binge eating episodes (n=7 across 3 participants) already posed challenges in detecting meaningful associations. Including the LOC aspect could have further reduced this number, potentially making statistical analyses even more difficult or infeasible. Nonetheless, we recognize that future studies should include LOC as a key criterion for binge eating to enhance content validity and better capture associations with momentary predictors.

Future Perspectives

This study provides insights into the feasibility of ESM in the population after MBS. This information could be used to further refine the development of ESM to monitor patients during their postoperative trajectory. Several participants mentioned that the questionnaires were repetitive and monotonous. Compliance might be increased if the questions are asked in a different order. Furthermore, patients would appreciate a free-input text area to explain their answers about emotion and problematic eating behavior. Feasibility for health care personnel should be considered and evaluated in future studies, especially when the objective of ESM includes remote monitoring.

Monitoring efforts should be particularly directed toward individuals exhibiting problematic eating behaviors. Such results could help patients and health care professionals to identify problematic eating behavior episodes by recommending them to focus on specific emotions, physical states, or contextual factors.

Conclusion

From this exploratory study to describe the feasibility of ESM regarding negative and positive affect, physical states, contextual factors, and problematic eating behaviors, findings suggest that ESM can be feasible in the population after MBS. Despite the low participation rate, compliance with experience sampling was moderate, and the completion rate was high. According to the interviews, patients considered ESM feasible and of added value, as awareness arose by using the ESM app. This offers a perspective for the use of ESM in the follow-up after MBS. Temporal associations were found between hunger and craving, and between positive affect and grazing. Nevertheless, no distinct patterns emerged among emotions, physical states, context, and problematic eating behaviors.

Acknowledgments

The study was performed as part of the Exceptional Deep and Intelligent Coach (EDIC) project, grant 628.011.021, funded by the Netherlands Organization for Scientific Research (NWO). Artificial intelligence was not used in writing this manuscript.

XSL•FO
RenderX

Data Availability

The data generated and analyzed during this study are not publicly available since participants were not asked for consent to make their data publicly available. The data are available from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Semistructured interview guide.
[DOCX File , 16 KB - formative_v9i1e60486_app1.docx]

Multimedia Appendix 2
Compliance in answering experience sampling questionnaires per participant during the 14-day study period.
[PNG File , 110 KB - formative_v9i1e60486_app2.png]

Multimedia Appendix 3
Descriptive data for the different types of problematic eating behavior.
[DOCX File , 19 KB - formative_v9i1e60486_app3.docx]

References

1. Sjöström L, Lindroos A, Peltonen M, Torgerson J, Bouchard C, Carlsson B, et al. Lifestyle, diabetes, and cardiovascular risk factors 10 years after bariatric surgery. N Engl J Med 2004;351(26):2683-2693. [doi: 10.1056/NEJMoa035622] [Medline: 15616203]
2. Karmali S, Brar B, Shi X, Sharma AM, de Gara C, Birch DW. Weight recidivism post-bariatric surgery: a systematic review. Obes Surg 2013;23(11):1922-1933. [doi: 10.1007/s11695-013-1070-4] [Medline: 23996349]
3. Amundsen T, Strømmen M, Martins C. Suboptimal weight loss and weight regain after gastric bypass surgery-postoperative status of energy intake, eating behavior, physical activity, and psychometrics. Obes Surg 2017;27(5):1316-1323 [FREE Full text] [doi: 10.1007/s11695-016-2475-7] [Medline: 27914028]
4. Kaouk L, Hsu AT, Tanuseputro P, Jessri M. Modifiable factors associated with weight regain after bariatric surgery: a scoping review. F1000Res 2019;8:615 [FREE Full text] [doi: 10.12688/f1000research.18787.2] [Medline: 32983412]
5. Freire RH, Borges MC, Alvarez-Leite JI, Toulson Davisson Correia MI. Food quality, physical activity, and nutritional follow-up as determinant of weight regain after Roux-en-Y gastric bypass. Nutrition 2012;28(1):53-58. [doi: 10.1016/j.nut.2011.01.011] [Medline: 21885246]
6. Reid RER, Oparina E, Plourde H, Andersen RE. Energy intake and food habits between weight maintainers and regainers, five years after Roux-en-Y gastric bypass. Can J Diet Pract Res 2016;77(4):195-198. [doi: 10.3148/cjdpr-2016-013] [Medline: 27744735]
7. Colles SL, Dixon JB, O'Brien PE. Grazing and loss of control related to eating: two high-risk factors following bariatric surgery. Obesity (Silver Spring) 2008;16(3):615-622 [FREE Full text] [doi: 10.1038/oby.2007.101] [Medline: 18239603]
8. Kofman MD, Lent MR, Swencionis C. Maladaptive eating patterns, quality of life, and weight outcomes following gastric bypass: results of an internet survey. Obesity (Silver Spring) 2010;18(10):1938-1943 [FREE Full text] [doi: 10.1038/oby.2010.27] [Medline: 20168309]
9. Meany G, Conceição E, Mitchell JE. Binge eating, binge eating disorder and loss of control eating: effects on weight outcomes after bariatric surgery. Eur Eat Disord Rev 2014;22(2):87-91 [FREE Full text] [doi: 10.1002/erv.2273] [Medline: 24347539]
10. Pizato N, Botelho P, Gonçalves VSS, Dutra E, de Carvalho K. Effect of grazing behavior on weight regain post-bariatric surgery: a systematic review. Nutrients 2017;9(12):1322 [FREE Full text] [doi: 10.3390/nu9121322] [Medline: 29206132]
11. Fairburn CG, Cooper Z, Shafran R. Cognitive behaviour therapy for eating disorders: a "transdiagnostic" theory and treatment. Behav Res Ther 2003;41(5):509-528. [doi: 10.1016/s0005-7967(02)00088-8] [Medline: 12711261]
12. Polivy J, Herman CP. Causes of eating disorders. Annu Rev Psychol 2002;53:187-213. [doi: 10.1146/annurev.psych.53.100901.135103] [Medline: 11752484]
13. Dingemans A, Danner U, Parks M. Emotion regulation in binge eating disorder: a review. Nutrients 2017;9(11):1274 [FREE Full text] [doi: 10.3390/nu9111274] [Medline: 29165348]
14. Ahlich E, Rancourt D. Boredom proneness, interoception, and emotional eating. Appetite 2022;178:106167. [doi: 10.1016/j.appet.2022.106167] [Medline: 35843373]
15. van Strien T, Herman CP, Anschutz DJ, Engels RC, de Weerth C. Moderation of distress-induced eating by emotional eating scores. Appetite 2012;58(1):277-284. [doi: 10.1016/j.appet.2011.10.005] [Medline: 22037008]

XSL•FO
RenderX

16. Koball AM, Meers MR, Storfer-Isser A, Domoff SE, Musher-Eizenman DR. Eating when bored: revision of the emotional eating scale with a focus on boredom. Health Psychol 2012;31(4):521-524. [doi: 10.1037/a0025893] [Medline: 22004466]
17. Schaefer LM, Engel SG, Wonderlich SA. Ecological momentary assessment in eating disorders research: recent findings and promising new directions. Curr Opin Psychiatry 2020;33(6):528-533 [FREE Full text] [doi: 10.1097/YCO.0000000000000639] [Medline: 32740204]
18. Williams-Kerver GA, Steffen KJ, Smith KE, Cao L, Crosby RD, Engel SG. Negative affect and loss of control eating among bariatric surgery patients: an ecological momentary assessment pilot investigation. Obes Surg 2020;30(6):2382-2387 [FREE Full text] [doi: 10.1007/s11695-020-04503-6] [Medline: 32125646]
19. Hudson JI, Hiripi E, Pope HG, Kessler RC. The prevalence and correlates of eating disorders in the national comorbidity survey replication. Biol Psychiatry 2007;61(3):348-358 [FREE Full text] [doi: 10.1016/j.biopsych.2006.03.040] [Medline: 16815322]
20. Opolski M, Chur-Hansen A, Wittert G. The eating-related behaviours, disorders and expectations of candidates for bariatric surgery. Clin Obes 2015;5(4):165-197. [doi: 10.1111/cob.12104] [Medline: 26173752]
21. de Zwaan M, Hilbert A, Swan-Kremeier L, Simonich H, Lancaster K, Howell LM, et al. Comprehensive interview assessment of eating behavior 18-35 months after gastric bypass surgery for morbid obesity. Surg Obes Relat Dis 2010;6(1):79-85. [doi: 10.1016/j.soard.2009.08.011] [Medline: 19837012]
22. Barbuti M, Brancati GE, Calderone A, Fierabracci P, Salvetti G, Weiss F, et al. Prevalence of mood, panic and eating disorders in obese patients referred to bariatric surgery: patterns of comorbidity and relationship with body mass index. Eat Weight Disord 2022;27(3):1021-1027 [FREE Full text] [doi: 10.1007/s40519-021-01236-y] [Medline: 34137006]
23. Heriseanu AI, Hay P, Corbit L, Touyz S. Grazing in adults with obesity and eating disorders: a systematic review of associated clinical features and meta-analysis of prevalence. Clin Psychol Rev 2017;58:16-32. [doi: 10.1016/j.cpr.2017.09.004] [Medline: 28988855]
24. Guthrie H, Tetley D, Hill AJ. Quasi-prospective, real-life monitoring of food craving post-bariatric surgery: comparison with overweight and normal weight women. Clin Obes 2014;4(3):136-142 [FREE Full text] [doi: 10.1111/cob.12054] [Medline: 25826768]
25. Erden Aki SO, Yıldız MI, Kamış GZ, Aytulun A, Karakaya J, Işıklı S. Relationship of depression, impulsivity, distress intolerance and coping styles with maladaptive eating patterns in bariatric candidates. Eat Weight Disord 2022;27(8):3351-3366. [doi: 10.1007/s40519-022-01465-9] [Medline: 35999437]
26. Mehl MR, Conner TS. Handbook of Research Methods for Studying Daily Life. New York, NY: The Guilford Press; 2012.
27. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol 2008;4:1-32. [doi: 10.1146/annurev.clinpsy.3.022806.091415] [Medline: 18509902]
28. Spook JE, Paulussen T, Kok G, van Empelen P. Monitoring dietary intake and physical activity electronically: feasibility, usability, and ecological validity of a mobile-based Ecological Momentary Assessment tool. J Med Internet Res 2013;15(9):e214 [FREE Full text] [doi: 10.2196/jmir.2617] [Medline: 24067298]
29. Timmerman JG, Dekker-van Weering MGH, Stuiver MM, Groen WG, Wouters MWJM, Tönis TM, et al. Ambulant monitoring and web-accessible home-based exercise program during outpatient follow-up for resected lung cancer survivors: actual use and feasibility in clinical practice. J Cancer Surviv 2017;11(6):720-731 [FREE Full text] [doi: 10.1007/s11764-017-0611-6] [Medline: 28396993]
30. Mackesy-Amiti ME, Boodram B. Feasibility of ecological momentary assessment to study mood and risk behavior among young people who inject drugs. Drug Alcohol Depend 2018;187:227-235 [FREE Full text] [doi: 10.1016/j.drugalcdep.2018.03.016] [Medline: 29684890]
31. Shaffer KM, Chow PI, Glazer JV, Le T, Reilley MJ, Jameson MJ, et al. Feasibility of ecological momentary assessment to study depressive symptoms among cancer caregivers. Psychooncology 2021;30(5):756-764 [FREE Full text] [doi: 10.1002/pon.5627] [Medline: 33432717]
32. Porras-Segovia A, Molina-Madueño RM, Berrouiguet S, López-Castroman J, Barrigón ML, Pérez-Rodríguez MS, et al. Smartphone-based ecological momentary assessment (EMA) in psychiatric patients and student controls: a real-world feasibility study. J Affect Disord 2020;274:733-741. [doi: 10.1016/j.jad.2020.05.067] [Medline: 32664009]
33. Jean FAM, Sibon I, Husky M, Couffinhal T, Swendsen J. Feasibility and validity of ecological momentary assessment in patients with acute coronary syndrome. BMC Cardiovasc Disord 2020;20(1):499 [FREE Full text] [doi: 10.1186/s12872-020-01774-w] [Medline: 33246420]
34. Maatoug R, Peiffer-Smadja N, Delval G, Brochu T, Pitrat B, Millet B. Ecological momentary assessment using smartphones in patients with depression: feasibility study. JMIR Form Res 2021;5(2):e14179 [FREE Full text] [doi: 10.2196/14179] [Medline: 33625367]
35. Wooldridge JS, Soriano EC, Harris DE, Afari N. Feasibility and acceptability of ecological momentary assessment of psychosocial factors and self-management behaviors among veterans with type 2 diabetes. Diabetes Spectr 2022;35(1):76-85 [FREE Full text] [doi: 10.2337/ds21-0020] [Medline: 35308149]
36. Glenn CR, Kleiman EM, Kearns JC, Santee AC, Esposito EC, Conwell Y, et al. Feasibility and acceptability of ecological momentary assessment with high-risk suicidal adolescents following acute psychiatric care. J Clin Child Adolesc Psychol 2022;51(1):32-48. [doi: 10.1080/15374416.2020.1741377] [Medline: 32239986]

XSL•FO
RenderX

37. Arend AK, Kaiser T, Pannicke B, Reichenberger J, Naab S, Voderholzer U, et al. Toward individualized prediction of binge-eating episodes based on ecological momentary assessment data: item development and pilot study in patients with bulimia nervosa and binge-eating disorder. JMIR Med Inform 2023;11:e41513 [FREE Full text] [doi: 10.2196/41513] [Medline: 36821359]
38. Thomas JG, Schumacher LM, Vithiananthan S, Jones DB, Smith KE, Chou T, et al. Ecological momentary assessment of changes in eating behaviors, appetite, and other aspects of eating regulation in roux-en-Y gastric bypass and sleeve gastrectomy patients. Appetite 2023;183:106465 [FREE Full text] [doi: 10.1016/j.appet.2023.106465] [Medline: 36701847]
39. Bond DS, Thomas JG, Jones DB, Schumacher LM, Webster J, Evans EW, et al. Ecological momentary assessment of gastrointestinal symptoms and risky eating behaviors in Roux-en-Y gastric bypass and sleeve gastrectomy patients. Surg Obes Relat Dis 2021;17(3):475-483 [FREE Full text] [doi: 10.1016/j.soard.2020.11.017] [Medline: 33353862]
40. Mundi MS, Lorentz PA, Grothe K, Kellogg TA, Collazo-Clavell ML. Feasibility of smartphone-based education modules and ecological momentary assessment/intervention in pre-bariatric surgery patients. Obes Surg 2015;25(10):1875-1881. [doi: 10.1007/s11695-015-1617-7] [Medline: 25702141]
41. Thomas JG, Bond DS, Ryder BA, Leahey TM, Vithiananthan S, Roye GD, et al. Ecological momentary assessment of recommended postoperative eating and activity behaviors. Surg Obes Relat Dis 2011;7(2):206-212. [doi: 10.1016/j.soard.2010.10.007] [Medline: 21130703]
42. Wrzus C, Neubauer AB. Ecological momentary assessment: a meta-analysis on designs, samples, and compliance across research fields. Assessment 2023;30(3):825-846 [FREE Full text] [doi: 10.1177/10731911211067538] [Medline: 35016567]
43. Macht M, Haupt C, Salewsky A. Emotions and eating in everyday life: application of the experience-sampling method. Ecology of Food and Nutrition 2004;43(4):11-21. [doi: 10.1080/03670240490454723]
44. Boh B, Jansen A, Clijsters I, Nederkoorn C, Lemmens LH, Spanakis G, et al. Indulgent thinking? Ecological momentary assessment of overweight and healthy-weight participants' cognitions and emotions. Behav Res Ther 2016;87:196-206. [doi: 10.1016/j.brat.2016.10.001] [Medline: 27736671]
45. Goldstein H. Multilevel Statistical Models. 4th edition. Chichester, UK: John Wiley & Sons Inc; 2011.
46. Myin-Germeys I, Kuppens P. The Open Handbook of Experience Sampling Methodology: A Step-by-Step to Designing, Conducting, and Aalyzing ESM Studies. 2nd edition. Leuven, Belgium: Center for Research on Experience Sampling and Ambulatory Methods Leuven; 2022.
47. Broekhuis M, van Velsen L. Improving usability benchmarking for the eHealth domain: the development of the eHealth UsaBility Benchmarking instrument (HUBBI). PLoS One 2022;17(2):e0262036 [FREE Full text] [doi: 10.1371/journal.pone.0262036] [Medline: 35176033]
48. Ratcliff MB, Zeller MH, Inge TH, Hrovat KB, Modi AC. Feasibility of ecological momentary assessment to characterize adolescent postoperative diet and activity patterns after weight loss surgery. Surg Obes Relat Dis 2014;10(4):705-710 [FREE Full text] [doi: 10.1016/j.soard.2014.01.034] [Medline: 24837564]
49. Manasse SM, Schumacher LM, Goldstein SP, Martin GJ, Crosby RD, Juarascio AS, et al. Are individuals with loss-of-control eating more prone to dietary lapse in behavioural weight loss treatment? An ecological momentary assessment study. Eur Eat Disord Rev 2018;26(3):259-264 [FREE Full text] [doi: 10.1002/erv.2583] [Medline: 29484774]
50. Stevenson BL, Dvorak RD, Wonderlich SA, Crosby RD, Gordon KH. Emotions before and after loss of control eating. Eat Disord 2018;26(6):505-522. [doi: 10.1080/10640266.2018.1453634] [Medline: 29565734]
51. Bhuyan SS, Lu N, Chandak A, Kim H, Wyant D, Bhatt J, et al. Use of mobile health applications for health-seeking behavior among US adults. J Med Syst 2016;40(6):153. [doi: 10.1007/s10916-016-0492-7] [Medline: 27147516]
52. Jiang Y, West BT, Barton DL, Harris MR. Acceptance and use of eHealth/mHealth applications for self-management among cancer survivors. Stud Health Technol Inform 2017;245:131-135 [FREE Full text] [doi: 10.3233/978-1-61499-830-3-131] [Medline: 29295067]
53. Conceição E, Mitchell JE, Vaz AR, Bastos AP, Ramalho S, Silva C, et al. The presence of maladaptive eating behaviors after bariatric surgery in a cross sectional study: importance of picking or nibbling on weight regain. Eat Behav 2014;15(4):558-562. [doi: 10.1016/j.eatbeh.2014.08.010] [Medline: 25213792]
54. Roefs A, Boh B, Spanakis G, Nederkoorn C, Lemmens LHJM, Jansen A. Food craving in daily life: comparison of overweight and normal-weight participants with ecological momentary assessment. J Hum Nutr Diet 2019;32(6):765-774 [FREE Full text] [doi: 10.1111/jhn.12693] [Medline: 31430000]
55. Reichenberger J, Richard A, Smyth JM, Fischer D, Pollatos O, Blechert J. It's craving time: time of day effects on momentary hunger and food craving in daily life. Nutrition 2018;55-56:15-20. [doi: 10.1016/j.nut.2018.03.048] [Medline: 29960151]
56. Dicker-Oren SD, Gelkopf M, Greene T. The dynamic network associations of food craving, restrained eating, hunger and negative emotions. Appetite 2022;175:106019. [doi: 10.1016/j.appet.2022.106019] [Medline: 35500722]
57. Goldschmidt AB, Crosby RD, Cao L, Wonderlich SA, Mitchell JE, Engel SG, et al. A preliminary study of momentary, naturalistic indicators of binge-eating episodes in adults with obesity. Int J Eat Disord 2018;51(1):87-91 [FREE Full text] [doi: 10.1002/eat.22795] [Medline: 29112288]
58. Goldschmidt AB, Engel SG, Wonderlich SA, Crosby RD, Peterson CB, Le Grange D, et al. Momentary affect surrounding loss of control and overeating in obese adults with and without binge eating disorder. Obesity (Silver Spring) 2012;20(6):1206-1211 [FREE Full text] [doi: 10.1038/oby.2011.286] [Medline: 21938073]

XSL•FO
RenderX

59. Ivezaj V, Lydecker JA, Wiedemann AA, Duffy AJ, Grilo CM. Does bariatric binge-eating size matter? Conceptual model and empirical support. Obesity (Silver Spring) 2020;28(9):1645-1651 [FREE Full text] [doi: 10.1002/oby.22876] [Medline: 32729221]

Abbreviations

ESM: experience sampling methodology
GEE: generalized estimation equation
HUBBI: eHealth Usability Benchmarking Instrument
LOC: loss of control
MBS: metabolic bariatric surgery
OR: odds ratio
ZGT: Hospital Group Twente

Edited by A Mavragani; submitted 07.06.24; peer-reviewed by M Ben Amar, G Williams - Kerver; comments to author 12.08.24; revised version received 11.09.24; accepted 07.01.25; published 05.03.25.

Please cite as:
Kuipers EAM, Timmerman JG, van Det MJ, Vollenbroek-Hutten MMR
Feasibility and Links Between Emotions, Physical States, and Eating Behavior in Patients After Metabolic Bariatric Surgery: Experience Sampling Study
JMIR Form Res 2025;9:e60486
URL: https://formative.jmir.org/2025/1/e60486
doi:10.2196/60486
PMID:40053719

©Ellen A M Kuipers, Josien G Timmerman, Marc J van Det, Miriam M R Vollenbroek-Hutten. Originally published in JMIR Formative Research (https://formative.jmir.org), 05.03.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

FRAILSURVEY—an mHealth App for Self-Assessment of Frailty Based on the Portuguese Version of the Groningen Frailty Indicator: Validation and Reliability Study

Luis Midao[1,2,3], PhD; Mafalda Duarte[4,5], Prof Dr; Rute Sampaio[6], Prof Dr; Marta Almada[1,2], PhD; Cláudia Camila Dias[7,8], Prof Dr; Constança Paúl[3,4], Prof Dr; Elísio Costa[1,2], Prof Dr

[1]RISE-Health, Biochemistry Lab, Faculty of Pharmacy, University of Porto, Porto, Portugal

[2]Porto4Ageing—Competences Centre on Active and Healthy Ageing, Faculty of Pharmacy, University of Porto, Porto, Portugal

[3]ICBAS—School of Medicine and Biomedical Sciences, University of Porto, Porto, Portugal

[4]RISE-Health, Department of Behavioural Sciences, School of Medicine and Biomedical Sciences, University of Porto, Porto, Portugal

[5]ISAVE—Superior Institute of Health, Amares, Braga, Portugal

[6]RISE-Health, Department of Biomedicine, Faculty of Medicine of the University of Porto, Porto, Portugal

[7]Knowledge Management Unit, Faculty of Medicine of the University of Porto (FMUP), Porto, Portugal

[8]CINTESIS@RISE, Department of Community Medicine, Information and Health Decision Sciences (MEDCIDS), Faculty of Medicine of the University of Porto (FMUP), Porto, Portugal

Corresponding Author:
Luis Midao, PhD
RISE-Health, Biochemistry Lab, Faculty of Pharmacy, University of Porto
R. Jorge de Viterbo Ferreira 228
Porto, 4050-313
Portugal
Phone: 351 22 042 8500
Email: luismidao@gmail.com

Abstract

Background: Portugal is facing the challenge of population ageing, with a notable increase in the proportion of older individuals. This has positioned the country among those in Europe with a high prevalence of frailty. Frailty, a geriatric syndrome characterized by diminished physiological reserve and heightened vulnerability to stressors, imposes a substantial burden on public health.

Objective: This study seeks to address two primary objectives: (1) translation and psychometric evaluation of the European Portuguese version of the Groningen Frailty Indicator (GFI); and (2) development and evaluation of the FRAILSURVEY app, a novel assessment tool for frailty based on the GFI. By achieving these objectives, the study aims to enhance the accuracy and reliability of frailty assessment in the Portuguese context, ultimately contributing to improved health care outcomes for older individuals in the region.

Methods: To accomplish the objectives of the study, a comprehensive research methodology was used. The study comprised 2 major phases: the initial translation and validation of the GFI into European Portuguese and the development of the FRAILSURVEY app. Following this, an extensive examination of the app's validity and reliability was conducted compared with the conventional paper version of the GFI. A randomized repeated crossover design was used to ensure rigorous evaluation of both assessment methods, using both the paper form of the GFI and the smartphone-based app FRAILSURVEY.

Results: The findings of the study revealed promising outcomes in line with the research objectives. The meticulous translation process yielded a final version of the GFI with robust psychometric properties, ensuring clarity and comprehensibility for participants. The study included 522 participants, predominantly women (367/522, 70.3%), with a mean age of 73.7 (SD 6.7) years. Psychometric evaluation of the European Portuguese GFI in paper form demonstrates good reliability (internal consistency: Cronbach a value of 0.759; temporal stability: intraclass correlation coefficient=0.974) and construct validity (revealing a 4D structure explaining 56% of variance). Evaluation of the app-based European Portuguese GFI indicates good reliability (interinstrument reliability: Cohen k=0.790; temporal stability: intraclass correlation coefficient=0.800) and concurrent validity (r=0.694; P<.001).

XSL•FO
RenderX

Conclusions: Both the smartphone-based app and the paper version of the GFI were feasible and acceptable for use. The findings supported that FRAILSURVEY exhibited comparable validity and reliability to its paper counterpart. FRAILSURVEY uses a standardized and validated assessment tool, offering objective and consistent measurements while eliminating subjective biases, enhancing accuracy, and ensuring reliability. This app holds promising potential for aiding health care professionals in identifying frailty in older individuals, enabling early intervention, and improving the management of adverse health outcomes associated with this syndrome. Its integration with electronic health records and other data may lead to personalized interventions, improving frailty management and health outcomes for at-risk individuals.

(JMIR Form Res 2025;9:e51975) doi:10.2196/51975

KEYWORDS

frailty; mHealth; assessment; validation; GFI; reliability; self-assessment; Groningen Frailty Indicator; FRAILSURVEY; mobile phone

Introduction

Background

Portugal is currently confronting a significant population ageing phenomenon, placing it among the European countries with one of the highest percentages of older individuals. As of 2021, the ageing index reached 182, meaning there were 182 people aged 65 years or older for every 100 individuals aged 0-14 years. This ageing trend results from various contributing factors, such as low birth rates, extended life expectancy, and emigration of young people. According to the 2011 census, 19% of the Portuguese population were aged 65 years or older. A substantial increase was denoted to 23.4% (2,423,639/10,343,066) in the 2021 census. Projections indicate that this percentage will continue to rise in the upcoming years. Estimations forecast that by 2080, the proportion of people aged 65 years and older will escalate to 37.5% of the total population [1-4]. This demographic change has significant implications for the country's economy, social security, and health care systems. The increasing number of older people will require greater resources to provide for their care and support, including health care services, long-term care facilities, and home care services. At the same time, the working-age population will be under pressure to support the social security system as the number of retirees increases and the number of workers decreases. As the proportion of older people grows, so does the prevalence of age-related diseases, comorbidities, and geriatric syndromes. Portugal ranks among the European countries with the highest prevalence of frailty (20.7%) [5,6].

Frailty is characterized by a reduced physiologic reserve and increased susceptibility to stressors, leading to an elevated risk of adverse health outcomes such as falls, hospitalization, disability, and death. Typically associated with ageing, frailty may encompass a blend of physical, cognitive, and psychosocial factors, making it a complex and multifaceted condition, and, therefore, identifying and addressing frailty is essential for preventing comorbidities, optimizing health outcomes, and improving the quality of life in older adults [7-11]. However, evaluating frailty poses challenges due to its multifactorial nature, resulting in various tools and methods commonly used for its assessment. It is crucial to recognize that every assessment tool is flawed, lacks standardization, and different approaches may be more suitable for specific populations or settings [10,12].

The Groningen Frailty Indicator (GFI) stands out for its comprehensive assessment for evaluating frailty in older adults. With its rapid administration and validated reliability in diverse populations, the GFI serves as a predictive tool for adverse outcomes such as hospitalization and mortality. Its multidimensional approach offers a nuanced understanding of frailty, making it invaluable for routine screening in clinical practice, empowering health care professionals to craft suitable care plans for their patients [13-16]. Covering physical, psychological, social, and cognitive aspects, the paper version of the GFI has undergone validation in diverse settings, such as among Romanian community-dwelling older adults [17] and Chinese nursing home residents [18], and different populations, such as patients with rheumatoid arthritis [19], patients with end-stage hip and knee osteoarthritis [20], and head and neck cancer surgery patients [21], confirming its accuracy, consistency, and predictive value. Distinguishing itself from other instruments, the GFI boasts a straightforward administration process, requiring no more than 10 minutes to complete. Furthermore, as a noninvasive instrument, individuals can complete it through self-reporting or seek guidance from a trained health professional, if necessary [15,16]. There are already some versions in European Portuguese of frailty scales, such as the Edmonton Frailty Scale [22], Program on Research for Integrating Services for the Maintenance of Autonomy questionnaire (PRISMA-7) [23], FRAIL Scale [24], Clinical Frailty Scale [25], SUNFRAIL [26], and Tilburg Frailty Indicator [27]; however, some are not yet validated for the Portuguese population. To our knowledge, there is still no study on the translation and validation of the GFI into European Portuguese.

Despite being commonly used, paper-based questionnaires have several disadvantages: manual data entry is time-consuming and leads to potential errors; accessibility is limited as they must be physically distributed and collected; data security may be compromised; and incomplete or illegible responses and reduced data analysis capabilities are common challenges. Furthermore, they lack interactive features that enhance engagement and data quality, such as in digital formats [28]. Smartphone-based assessments aim to develop a user-friendly tool that streamlines screening and monitoring by scoring data [29,30]. Such an approach has the potential to increase health literacy resulting in cost savings. Nevertheless, it is crucial to validate the smartphone-based apps to ensure their accuracy and reliability

XSL•FO
RenderX

compared with the original paper questionnaires, as there may be possible response bias between the 2 versions [31].

Aim

Given the high burden of frailty in Portugal and to contribute to improving frailty assessment, enhancing health care delivery, and addressing the specific needs of older adults in Portugal, this study aimed to translate and validate the GFI into European Portuguese. Then, this study also aimed to examine the validity and reliability of the digital version of the GFI, the app FRAILSURVEY.

Methods

This is a quantitative, exploratory research that involves analysis of psychometric data and that was developed in 2 major phases: initially, the GFI was translated and validated into European Portuguese, and the application FRAILSURVEY was developed and its validity and reliability studied as a mobile health (mHealth) tool for frailty assessment.

Translation and Validation of the GFI

The GFI is a short, easy-to-use instrument that has proven to be a good alternative for assessing frailty in the population aged 65 years or older [32]. Textbox 1 indicates the domains assessed in the GFI scale, as well as scoring items. The self-assessment version of the GFI published in 2013 was used, composed of 15 items and divided into 4 domains [33]. The physical domain has 9 questions related to toileting, shopping, mobility functions, dressing, physical fitness, vision, hearing, unexplained weight loss, and medication use. The cognitive domain includes 1 item related to cognitive concerns. The social domain consists of 3 questions about emotional isolation, social relations, and social support. The psychological domain comprises 2 questions related to mood. Each item has a score of 0 or 1 attributed, with 1 indicating a dependency problem. For example, if the participant cannot dress or undress without help, he would score 1. The total score ranged from 0 to 15, and higher scores indicate greater levels of frailty. The originally proposed cutoff point of 4 or higher represents frailty [16].

Textbox 1. Self-assessment version of the Groningen Frailty Indicator (GFI): with 15 items divided by 4 main domains (physical, cognitive, social, and psychological), the GFI scores between 0 and 15. People with scores of 4 and above are considered frail [33].

Physical domain

Are you able to carry out these tasks single-handedly and without any help? (The use of help resources such as a walking stick, walking frame, or wheelchair is considered independent.)

1. Shopping
2. Walking around outside (around the house or to the neighbors)
3. Dressing and undressing
4. Going to the toilet
5. What mark do you give yourself for physical fitness? (Scales 0-10)
6. Do you experience problems in daily life due to poor vision?
7. Do you experience problems in daily life due to being hard of hearing?
8. During the last 6 months have you lost a lot of weight unwillingly? (3 kg in 1 month or 6 kg in 2 months)
9. Do you take 4 or more different types of medicine?

Cognitive domain

10. Do you have any complaints about your memory?

Social domain

11. Do you sometimes experience emptiness around yourself?
12. Do you sometimes miss people around yourself?
13. Do you sometimes feel abandoned?

Psychological domain

14. Have you recently felt downhearted or sad?
15. Have you recently felt nervous or anxious?

Scoring

Questions 1-4: Yes=0; No=1

Question 5: 0-6=1; 7-10=0

Questions 6-15: No=0; Yes=1

For the development of the European Portuguese version of the GFI, the self-report instrument guidelines defined by the ISPOR Task Force for Translation and Cultural Adaptation were followed [34]. Briefly, the translation was done by 2 technicians,

XSL•FO
RenderX

with experience in this type of task, and the final version of each item (of the 15) always resulted from complete agreement between them. After that an English technician carried out the back-translation, and a consensus was reached as the back-translation version was in line with the original version.

Developing the Smartphone-Based App of the GFI

The smartphone-based app was developed for both Android (the operating system created by Google) and iOS (created by Apple) systems. FRAILSURVEY is a questionnaire that comprises 2 sets of questions: 19 questions about sociodemographic data; social resources; and self-perception of health, nutrition, medication, psychosocial and cognitive status, and time occupation, plus a set of 15 questions used to assess frailty status, the European Portuguese version of the GFI. FRAILSURVEY also uses the originally proposed cutoff point of 4 or higher representing frailty [16].

FRAILSURVEY development was based on an accessibility principle for older adults to guarantee greater autonomy to use it. Therefore, considerations such as vision, hearing, motor control, and cognition were considered. Considering the vision of the older population, the app features a typographic font designed for low-vision people, APHont, developed by the American Printing House for the Blind. It also presents a way to increase or decrease the font size, as well as the possibility of listening to the audio information.

The results obtained classify the person as robust or frail. The app displays tailored recommendations if the person is classified as frail. The recommendations revolve around aspects such as physical exercise, a healthy diet, and social networks, among others (Figure 1). The app generates an anonymous database for research purposes.

Figure 1. Printscreens of the FRAILSURVEY app.

XSL•FO
RenderX

Study Design

A randomized repeated crossover design was used with the paper form of the GFI and the smartphone-based app FRAILSURVEY, as used in other similar studies [35]. This design entailed the dual administration of both assessment methods to each participant on distinct occasions, separated by a 1-week interval. The order of presentation for the assessments was randomized to mitigate any potential sequence effects. In evaluating test-retest reliability, a subset of participants was randomly chosen from the larger cohort to participate in retest assessments approximately 2 weeks after their initial interaction. These retest assessments encompassed both the paper-based form and the application [36].

Ethical Considerations

Ethical clearance for the research was granted by the ethical review board of the faculty of pharmacy, University of Porto (33-11-2017). Before their involvement, the rights and responsibilities of participants were comprehensively elucidated, and written consent was obtained from all individuals. Participation was voluntary and no compensation was provided. Throughout the process of data collection and entry, strict confidentiality measures were upheld and all data were anonymized.

Participants

Between February 2019 and February 2020, older adults living in the district of Porto, Portugal, were invited in cultural and sports associations, day centers, and nursing homes to take part in this study. A total of 522 participants were recruited for this study, using a convenience sampling method. The sample included a diverse population of community-dwelling older individuals, residents of nursing homes, and older adults living in their own homes but availing daily care facilities. The inclusion criteria encompassed individuals aged 65 years and older, regardless of gender, who could independently complete the assessments. The sample exceeded the sample size required to be 10 times larger than the number of GFI items [37]. For test-retest reliability, 10% (52/522) of the participants were recruited randomly, but only 24 participants accepted to perform the retests and, therefore, were included in the test-retest analysis [36].

Statistical Analysis

Categorical variables were summarized as absolute and relative frequencies. Descriptive statistics were used to report the subject characteristics of the study sample. Differences between different subgroups were studied, that is, gender, age, marital status, education, and financial status. GFI scores (from 0 to 15) were presented as mean (SD), and differences between subgroups were tested through the Kruskal-Wallis test. GFI classification (nonfrail and frail) was presented as frequencies, and differences between different subgroups were tested through the Mann-Whitney *U* test. To address the issue of multiple comparisons, adjustments were made, and the Bonferroni correction was applied to all statistical tests involving associations across multiple demographic variables.

The psychometric properties of the European Portuguese version of the GFI were assessed as follows: internal consistency referred to the extent to which the items within a scale or questionnaire consistently measure the same construct and was assessed by Cronbach a. The retest was conducted 2 weeks after the initial contact with the participant for assessing temporal stability. A 2-week interval allows for a sufficient gap between the 2 administrations to minimize potential memory effects and recall bias while maintaining a relatively short time frame. The most commonly used statistical measure to evaluate temporal stability is the intraclass correlation coefficient (ICC), and therefore it was used in this study. The ICC value ranges between 0 and 1, with higher values indicating greater temporal stability. Values above 0.9 suggest excellent temporal stability, while values between 0.75 and 0.9 indicate good temporal stability and values between 0.5 and 0.75 show moderate temporal stability [38]. Construct validity was assessed by performing exploratory factor analysis with principal components and varimax rotation methods [39].

The psychometric properties of the FRAILSURVEY were assessed as follows: interinstrument reliability refers to the consistency of GFI scores obtained by the same person between the paper and smartphone versions. Cohen k was used to assess it. Cohen suggested the k result be interpreted as follows: values ≤0 as indicating no agreement, 0.01-0.20 as none to slight (light), 0.21-0.40 as fair (fair), 0.41-0.60 as moderate (moderate), 0.61-0.80 as substantial (considerable), and 0.81-1.00 as almost perfect agreement [40]; temporal stability was evaluated by the ICC, as described previously; and the concurrent validity was studied through Pearson correlation coefficient that was used to quantify the relationship between the scores obtained in the app and in paper and with the app, the scores from the established criterion measure, and the scores from the instrument we intended to validate. Pearson correlation coefficient assesses the strength and direction of the correlation, indicating the degree to which the 2 measures agree or disagree. The validity assessment in this study was based on demonstrating the absence of significant differences in the total score of the GFI and each item of the GFI. Two-tailed *t* tests were used, and a *P* value of <.05 was considered statistically significant. The significance level used was .05. Data were analyzed using SPSS Statistics (version 28.0; IBM Corp).

Results

European Portuguese Version of the GFI

The English version of the GFI tool was translated into European Portuguese by 2 native Portuguese speakers. The back-translation was done by a professional service provider who knew nothing about the purpose of the research. A panel of experts assessed the similarity of the versions and reviewed and resolved all discrepancies and ambiguities by consensus. To assess the comprehensibility of the language and wording, the revised preliminary European Portuguese version was used with a pilot sample of 10 native Portuguese-speaking older adults. The final European Portuguese version of the GFI was defined and used in the validation process. No changes were made to the version used, as the items were judged to be clear and easy to understand. In the end, the final version was produced (Table S1 in Multimedia Appendix 1).

XSL•FO
RenderX

Participants

In this study, 522 participants were included, with a mean average age of 73.7 (SD 6.7) years and consisted mainly of women (367/522, 70.3%). Almost two-thirds of the sample were aged between 65 and 74 years, most participants were married (311/522, 59.6%), the majority had only primary school completed (283/522, 54.2%), and 59.2% (309/522) reported having moderate money to meet their needs (Table 1). Using the paper version of the European Portuguese version of the GFI, the average score was 4.57, significantly higher in women (5.05) than in men (3.43; *P*<.001). The score increases significantly with age, with higher scores being more present in older people (GFI=3.73 for participants aged between 65 and 74 years; GFI=5.58 for participants aged between 75 and 84 years, and; GFI=7.77 for participants aged 85 years or older; *P*<.001). There are also significant differences in the mean GFI scores according to marital status, with married (3.67) and divorced people (4.62) having lower values while widowed (6.02) or single people (6.13) having higher values (*P*<.001). Frailty is also related to education. We reported that higher frailty scores are associated with lower levels of education, and with the economic level, with greater scores of frailty in people with lower economic levels.

Table 1. Participants' sociodemographic characteristics according to their frailty status: 522 participants were included in this randomized repeated crossover design study.

	Variables, n (%)	Portuguese GFI[a], mean (SD)	*P* value (Kruskal-Wallis test)	Nonfrailty, n (%)	Frailty, n (%)	*P* value (Mann-Whitney *U* test)
Sex			<.001			<.001
Female	367 (70.3)	5.05 (3.04)		127 (34.6)	240 (65.4)	
Male	155 (29.7)	3.43 (2.702)		89 (57.4)	66 (42.6)	
Age (years)			<.001			<.001
65-74	341 (65.3)	3.73 (2.68)		181 (53.1)	160 (46.9)	
75-84	134 (25.7)	5.58 (2.91)		31 (23.1)	103 (76.9)	
85+	47 (9.0)	7.77 (2.86)		4 (8.5)	43 (91.5)	
Marital status			<.001			.003
Single	32 (6.1)	6.13 (3.24)		6 (18.8)	26 (81.3)	
Married	311 (59.6)	3.67 (2.47)		161 (51.8)	150 (48.2)	
Widowed	158 (30.3)	6.02 (3.34)		38 (24.1)	120 (75.9)	
Divorced	21 (4.0)	4.62 (2.94)		11 (52.4)	10 (47.6)	
Education			<.001			<.001
Illiterate	36 (6.9)	7.03 (3.11)		5 (13.9)	31 (86.1)	
Unfinished primary school	80 (15.3)	5.90 (2.95)		19 (23.8)	61 (76.3)	
Primary school	283 (54.2)	4.29 (2.97)		128 (45.2)	155 (54.8)	
High school	92 (17.6)	3.52 (2.32)		47 (51.1)	45 (48.9)	
Higher education	31 (5.9)	4.00 (3.24)		17 (54.8)	14 (45.2)	
Do you think you have enough money to meet your needs?			<.001			<.001
Nothing	12 (2.3)	5.00 (3.77)		6 (50.0)	6 (50.0)	
Little	190 (36.4)	5.89 (3.20)		47 (24.7)	143 (75.3)	
Moderately	309 (59.2)	3.78 (2.63)		158 (51.1)	151 (48.9)	
Quite enough	11 (2.1)	(1.96)		4 (45.5)	6 (54.5)	

[a]GFI: Groningen Frailty Indicator.

According to the GFI frailty classification, using the cutoff point of 4 [16], 58.6% of the people included in this study are frail. Frailty is significantly greater in women (65.4% vs 42.6% in men; *P*<.001), in older people (91.5% in people aged 85 years and older vs 76.9% in people aged 75-84 years vs 46.9% in people aged 65 and 74 years; *P*<.001), and in single people and widows (81.3% and 75.9%, respectively, vs 48.2% in married people and 47.6% in divorced people; *P*<.001). The proportion of frail people is also significantly higher for people with the lowest level of education and at lower economic levels.

Item Descriptive Analysis and Missing Data

Of the 15 indicators, the most prevalent among all people are the items that assess nervousness and anxiety (present in

XSL•FO
RenderX

320/521, 61.4% of the participants), problems in daily life due to poor vision reported by 56% (291/520) of the participants, and polypharmacy reported by 53.6% (279/522; Table 2). The least prevalent items are mobility issues, more specifically, the inability to dress and undress alone (51/521, 9.8%) and to use the bathroom without help (41/522, 7.9%), and the feeling of abandonment (37/522, 7.1%).

Table 2. Groningen Frailty Indicator: missing values and criteria fulfillments: of the 522 participants, questions 3, 4, 14, and 15 had 1 missing value; questions 6 and 9 had 2 missing values; and question 5 had 10 missing values.

Question	Fulfills criteria, n (%)	Missing values, n
1. Shopping	522 (18.4)	0
2. Walking around outside (around the house or to the neighbors)	522 (13.0)	0
3. Dressing and undressing	521 (9.8)	1
4. Going to the toilet	521 (7.9)	1
5. How do you rate your physical fitness? (scale 0-10)	512 (35.5)	10
6. Do you experience problems in daily life due to poor vision?	520 (56.0)	2
7. Do you experience problems in daily life due to being hard of hearing?	522 (35.1)	0
8. During the last 6 months, did you lose a lot of weight unwillingly? (3 kg in 1 month or 6 kg in 2 months)	522 (11.5)	0
9. Do you take 4 or more different types of medicine?	520 (53.6)	2
10. Do you have any complaints about your memory?	522 (35.2)	0
11. Do you sometimes experience an emptiness around you?	522 (39.1)	0
12. Do you sometimes miss people around you?	522 (28.0)	0
13. Do you sometimes feel abandoned?	522 (7.1)	0
14. Did you feel downhearted or sad recently?	521 (46.6)	1
15. Did you feel nervous or anxious recently?	521 (61.4)	1

Of the 522 participants, there are only 18 missing values in 7 of the 15 GFI questions. The average of missing values per participant is 0.03. Only 1 participant has missing values for 2 GFI questions, while the other 17 missing values come from 17 different participants. The question that has the most missing values (1.9%) is the question that asks the participant how he rates his physical fitness (from 0 to 10). The small proportion of missing data, constituting only 0.23% of the total values, suggests that it is improbable for these missing values to significantly impact the overall patterns or conclusions of the analysis, while the large sample size (522 participants) ensures that the study maintains sufficient statistical power to detect meaningful effects or relationships despite the negligible reduction in sample size attributable to missing data.

Psychometric Evaluation of the European Portuguese Version of the GFI (Paper)

Internal Consistency

The obtained Cronbach a value of 0.759 suggests a moderate level of internal consistency and no item was dropped (Table 3) [33].

Temporal Stability

The values varied between 0.504 and 1.000, being an overall value of 0.974 (Table 3).

XSL•FO
RenderX

Table 3. Psychometric Evaluation of the European Portuguese version of the GFI[a]—reliability analysis (internal consistency and temporal stability): no item was dropped and the Cronbach a value of 0.759 suggests a moderate level of internal consistency.

Question	Internal Consistency			Temporal stability, ICC[b]
	Scale average if the item is removed	Scale variation if the item is removed	Cronbach a if the item is removed	
1. Shopping	4.40	7.876	0.729	1.000
2. Walking around outside (around the house or to the neighbors)	4.46	8.189	0.737	0.947
3. Dressing and undressing	4.49	8.362	0.741	1.000
4. Going to the toilet	4.51	8.504	0.744	0.909
5. How do you rate your physical fitness? (scale 0-10)	4.23	7.882	0.740	0.940
6. Do you experience problems in daily life due to poor vision?	4.02	8.438	0.764	1.000
7. Do you experience problems in daily life due to being hard of hearing?	4.24	8.381	0.759	1.000
8. During the last 6 months, did you lose a lot of weight unwillingly? (3 kg in 1 month or 6 kg in 2 months)	4.47	8.952	0.765	1.000
9. Do you take 4 or more different types of medicine?	4.05	8.087	0.751	1.000
10. Do you have any complaints about your memory?	4.24	8.257	0.755	0.911
11. Do you sometimes experience an emptiness around you?	4.20	7.795	0.737	0.958
12. Do you sometimes miss people around you?	4.30	8.014	0.742	0.727
13. Do you sometimes feel abandoned?	4.51	8.683	0.751	0.880
14. Did you feel downhearted or sad recently?	4.13	7.547	0.727	1.000
15. Did you feel nervous or anxious recently?	3.97	8.092	0.749	0.504
Total			0.759	0.974

[a]GFI: Groningen Frailty Indicator.

[b]ICC: intraclass correlation coefficient.

Construct Validity

The suitability of the factor analysis by checking the existence of significant correlations between the items was confirmed by Kaiser-Meyer-Olkin test (0.799) and by Bartlett's Test of Sphericity (χ^2_{105}=2074.7; P<.001). The Kaiser-Meyer-Olkin measure assesses the suitability of data for factor analysis by evaluating proportions of variance in variables potentially linked to underlying factors, with higher values indicating better fit, typically above 0.6 or 0.7. Its advantage lies in providing a simple index for researchers to interpret and make decisions. Bartlett's Test of Sphericity validates the interrelationship of variables necessary for factor analysis, testing the null hypothesis of no correlation between variables, with a significant result (P<.05) confirming the suitability of the dataset for analysis factorial [41].

Evaluation of the scree plot and the size of the eigenvalues strongly suggest that the GFI has a 4D structure, explaining 56.0% of the variance (Table 4). This analysis produced four subscales: (1) Daily Activities and Physical Fitness (items 1-5), (2) Nourishment, (3) Health Problems (items 6, 7, 9, and 10), and (4) Psychosocial Functioning (items 11-15).

XSL•FO
RenderX

Table 4. Psychometric Evaluation of the European Portuguese version of the GFI[a]—validity analysis (construct validity): exploratory factor analysis structure of the Portuguese version of the GFI: loading for each factor and each item in the model with 4 factors after an orthogonal varimax rotation and factor extraction using principal components (n=502): GFI has a 4D structure, explaining 56.0% of the variance.

Items	Factors[b]			
	1	2	3	4
Daily Activities and Physical Fitness				
1. Shopping	0.822	0.167	0.165	–0.009
2. Walking around outside (around the house or to the neighbors)	0.863	0.084	0.087	–0.038
3. Dressing and undressing	0.880	0.076	0.002	0.018
4. Going to the toilet	0.872	0.028	–0.028	0.061
5. How do you rate your physical fitness? (scale 0-10)	0.505	0.179	0.279	–0.034
Health Problems				
6. Do you experience problems in daily life due to poor vision?	–0.008	0.033	0.511	0.325
7. Do you experience problems in daily life due to being hard of hearing?	0.061	–0.073	0.705	0.094
9. Do you take 4 or more different types of medicine?	0.183	0.134	0.563	–0.172
10. Do you have any complaints about your memory?	0.062	0.206	0.465	0.073
Nourishment				
8. During the last 6 months, did you lose a lot of weight unwillingly? (3 kg in 1 month or 6 kg in 2 months)	–0.053	0.041	0.153	0.832
Psychosocial Functioning				
11. Do you sometimes experience an emptiness around you?	0.092	0.817	0.058	–0.007
12. Do you sometimes miss people around you?	0.029	0.778	0.042	0.152
13. Do you sometimes feel abandoned?	0.270	0.432	–0.092	0.368
14. Did you feel downhearted or sad recently?	0.258	0.676	0.190	–0.005
15. Did you feel nervous or anxious recently?	–0.008	0.497	0.389	–0.214
Percentage of total variance explained for each factor	N/A[c]	13.50	8.66	6.67

[a]GFI: Groningen Frailty Indicator.

[b]Percentage of total variance explained for the 4 extracted factors=56%; Kaiser-Meyer-Olkin=0.799 and Bartlett's Test of Sphericity <0.001.

[c]N/A: not applicable.

Psychometric Evaluation of the FRAILSURVEY—an mHealth Tool for Frailty Assessment

Interinstrument Reliability

The values varied between 0.100 and 0.634, being an overall value of 0.790 (Table 5). From the 15 items, 2 items had a substantial agreement, 5 had a moderate agreement, 6 had a fair agreement, 2 had none to a slight agreement, and the overall score had an almost perfect agreement.

XSL•FO
RenderX

Table 5. Psychometric Evaluation of the FRAILSURVEY—reliability analysis (interinstrument reliability and temporal stability)a.

Question	Interinstrument reliability				Temporal stability, ICC[b]
	SE values	Observed k values	Lower limit (95% CI)	Upper limit (95% CI)	
1. Shopping	0.045	0.634	0.546	0.722	0.793
2. Walking around outside (around the house or to the neighbors)	0.053	0.608	0.503	0.712	0.867
3. Dressing and undressing	0.069	0.522	0.387	0.657	0.958
4. Going to the toilet	0.079	0.468	0.314	0.622	0.875
5. How do you rate your physical fitness? (scale 0-10)	0.043	0.419	0.335	0.503	0.788
6. Do you experience problems in daily life due to poor vision?	0.042	0.266	0.182	0.349	0.958
7. Do you experience problems in daily life due to being hard of hearing?	0.045	0.289	0.201	0.378	0.798
8. During the last 6 months, did you lose a lot of weight unwillingly? (3 kg in 1 month or 6 kg in 2 months)	0.086	0.100	0.000	0.270	0.958
9. Do you take 4 or more different types of medicine?	0.042	0.337	0.256	0.419	0.913
10. Do you have any complaints about your memory?	0.045	0.236	0.147	0.324	0.724
11. Do you sometimes experience an emptiness around you?	0.041	0.454	0.374	0.533	0.857
12. Do you sometimes miss people around you?	0.047	0.336	0.245	0.428	0.667
13. Do you sometimes feel abandoned?	0.099	0.188	0.000	0.381	0.793
14. Did you feel downhearted or sad recently?	0.039	0.443	0.366	0.520	0.718
15. Did you feel nervous or anxious recently?	0.043	0.328	0.243	0.413	0.589
GFI[c] score	0.027	0.790	0.737	0.844	0.800

[a]Evaluation of the app-based European Portuguese GFI indicates good reliability (interinstrument reliability: Cohen k=0.790; temporal stability: ICC=0.800).

[b]ICC: intraclass correlation coefficient.

[c]GFI: Groningen Frailty Indicator.

Temporal Stability

The retest was administered 2 weeks after the initial contact with the participant to evaluate the temporal stability of the measurements. The ICC values varied between 0.589 and 0.958, being an overall value of 0.800 (Table 5).

Concurrent Validity

All values were positive, ranging from 0.101 to 0.634, and the total score was 0.694 (Table 6). These results demonstrate a positive correlation between responses in both formats, paper, and app.

XSL•FO
RenderX

Table 6. Psychometric evaluation of the FRAILSURVEY—validity analysis (concurrent validity): a positive correlation between responses in both formats, paper, and app was observed.

Question	N	FRAILSURVEY, mean (SD)	Paper, mean (SD)	*r* value	*P* value
1. Shopping	522	0.20 (0.40)	0.18 (0.39)	0.634	<.001
2. Walking around outside (around the house or to the neighbors)	522	0.15 (0.36)	0.13 (0.34)	0.609	<.001
3. Dressing and undressing	521	0.10 (0.30)	0.10 (0.30)	0.522	<.001
4. Going to the toilet	521	0.09 (0.28)	0.08 (0,27)	0.468	<.001
5. How do you rate your physical fitness? (scale 0-10)	512	0.35 (0.48)	0.36 (0.48)	0.419	<.001
6. Do you experience problems in daily life due to poor vision?	520	0.52 (0.50)	0.56 (0.50)	0.266	<.001
7. Do you experience problems in daily life due to being hard of hearing?	522	0.33 (0.47)	0.35 (0.48)	0.290	<.001
8. During the last 6 months, did you lose a lot of weight unwillingly? (3 kg in 1 month or 6 kg in 2 months)	522	0.10 (0.30)	0.11 (0.32)	0.101	.02
9. Do you take 4 or more different types of medicine?	520	0.55 (0.50)	0.54 (0.50)	0.337	<.001
10. Do you have any complaints about your memory?	522	0.38 (0.49)	0.35 (0.48)	0.236	<.001
11. Do you sometimes experience an emptiness around you?	522	0.38 (0.49)	0.39 (0.49)	0.454	<.001
12. Do you sometimes miss people around you?	522	0.32 (0.47)	0.28 (0.45)	0.338	<.001
13. Do you sometimes feel abandoned?	522	0.08 (0.28)	0.07 (0.26)	0.189	<.001
14. Did you feel downhearted or sad recently?	521	0.50 (0.50)	0.47 (0.50)	0.444	<.001
15. Did you feel nervous or anxious recently?	521	0.64 (0.48)	0.61 (0.49)	0.329	<.001
Total score	522	4.67 (3.01)	4.57 (3.04)	0.694	<.001

Discussion

Principal Findings

Translating and validating frailty assessment tools are vital for cross-cultural research, ensuring inclusivity and diversity, and allowing researchers and health care professionals to identify frailty in older adults independently or irrespective of linguistic and cultural backgrounds [42]. Translated versions enhance measurement accuracy and consistency by retaining the psychometric properties and measurement validity. This facilitates the development of comprehensive public health policies and targeted interventions, improving care and quality of life for ageing populations worldwide [43]. Overall, translating and validating frailty assessment tools are crucial in advancing research, clinical practice, and public health initiatives related to frailty [44]. Frailty assessment tools are valuable for identifying at-risk individuals and tailoring interventions to their needs. By identifying frailty early, health care practitioners can develop strategies to prevent functional decline and mitigate adverse health outcomes [45].

In this study, the translation of the GFI into European Portuguese underwent rigorous processes, resulting in a final version with strong psychometric properties. The translated version proved to be clear and easily understandable.

Among the 522 participants in the study, the majority were women, which aligns with typical demographic patterns in this type of research. In addition, most participants were aged 65-74 years, were either married or widowed, had completed only primary education, and had moderate money to meet their needs.

The mean GFI score was 4.57 in this study, whereas validation studies conducted in China, Romania, and Lebanon were 4.60, 5.68, and 6.80, respectively [17,18,46]. Consistent with previous research, the average score was significantly higher in women than in men [5,12,17,46]. As anticipated, there was a significant positive correlation between age and GFI score and, additionally, marital status showed significant differences, with singles and widowers having higher scores [47]. Frailty also demonstrated significant associations with lower education levels and lower economic statuses, as already shown [48].

The GFI comprises 15 frailty indicators and the most prevalent were polypharmacy, vision problems, and feelings of anxiety or nervousness, all reported by more than 50% of the participants. Indeed, the link between polypharmacy and frailty has been established through a study using SHARE data, which explored the connection between mortality, frailty, and polypharmacy across 18 countries. According to these findings, 59.4% of the identified frail population was found to be taking multiple medications, a percentage that closely aligns with the results obtained in this study, where 53.6% of the population were under polypharmacy [5]. Frailty and poor vision are interconnected conditions with a bidirectional relationship, as already described [49]. Poor vision can contribute to frailty by impairing balance, mobility, and daily functioning, increasing the risk of falls and social isolation. It has already been described that people with visual impairment have 3.7 greater odds of being frail [49]. Anxiety symptoms, reported by 61.4% (320/521) of the participants of this study, were reported to be notably more prevalent among older adults with frailty than among those who are robust, which has also been reported elsewhere [50].

XSL•FO
RenderX

The psychometric evaluation of the European Portuguese version of the GFI demonstrated good reliability and validity. In previous psychometric studies assessing the internal consistency of the GFI, Cronbach α values ranged from 0.680 to 0.810. Nevertheless, in this study, the obtained value was 0.759, which is similar to Cronbach α value of the original scale [32]. Internal consistency maintained the same value even if any item was deleted. The assessment of reliability over time involved a retest conducted 2 weeks after the initial contact, and the ICC was used for calculation. The ICC values ranged from 0.504 to 1.000, with an overall value of 0.974. This value surpassed the reliability values obtained in 2 comparable studies conducted on the Chinese population (0.950 and 0.939) [18,51]. These results demonstrate that the European Portuguese version of the GFI exhibits excellent reliability over time, confirming its consistency as a measurement tool. The factor analysis yielded the same number of factors as the original scale, indicating that the translated version maintains a similar underlying factor structure to the original scale, suggesting that it measures the same fundamental constructs or dimensions as the items in the original version [52]. This encouraging finding indicates that the translation process effectively retained the core structure of the scale. Moreover, this confirmation of reliability and validity further demonstrates that the translated version of the GFI effectively preserves the essential elements of the original instrument.

The general hypothesis of this study was that a smartphone GFI app would show the same validity and reliability as the paper version, and it was proved in this study. The interinstrument reliability was studied to investigate the agreement and consistency between the 2 instruments, thereby ensuring the robustness of the results. High interinstrument reliability strengthens the credibility and accuracy of research outcomes, contributing to the overall quality and validity of assessments and evaluations. In this study, high levels of agreement were recorded between the instruments, proving that the digital format of the questionnaire is as reliable as the paper version. The reliability assessment over time was done in the same way as for the paper version of the GFI. The ICC values ranged from 0.589 to 0.958, with an overall value of 0.800, evidencing that, just like on paper, the digital version of the GFI remained reliable over time. The validity of the application was assessed by comparing the answers and total scores of the same individuals in the paper format (conventional) and the application (the new format to be validated). The results revealed a robust, positive, and statistically significant relationship between the responses in both forms. These findings not only establish the reliability of the application but also demonstrate its validation.

Health apps are vital in modern health care, providing easy access to health information, monitoring tools, and personalized advice [53]. The importance of mHealth becomes clear when examining its impact on health care accessibility. Many people, especially those living in disadvantaged or remote regions, face significant challenges in accessing health services. mHealth serves as a powerful equalizer, enabling individuals to connect with health care professionals, get immediate medical advice, and access health-related information, regardless of their geographic location [54]. This level of accessibility empowers patients and fosters early intervention and preventative care, ultimately resulting in improved health outcomes [55]. However, the proliferation of health apps on smartphone-based app stores is still on the rise due to the ease of app development, the absence of regulation, and the direct marketing to users, leading to an overcrowded scenario, which lacks reliability and accuracy. However, the ability of health stakeholders to discern the quality of these apps still lags significantly. This challenge largely stems from the lack of validation, which hinders app reliability and validity assurance [56]. Some potential benefits of validating health apps include improved accuracy and reliability of the app, increased user confidence in the app, and improved safety and effectiveness [57,58].

Assessing frailty is crucial as it helps identify individuals at higher risk of adverse health outcomes and enables timely interventions to mitigate these risks. Moreover, frailty assessments may guide treatment decisions, leading to tailored health care approaches focused on maintaining independence and managing chronic conditions. The impact of frailty in individuals and communities is profound, increasing health care usage, decreasing quality of life, and adding strain on families and caregivers. Understanding these effects allows for prioritized interventions, effective resource allocation, and appropriate support, ultimately improving the well-being and outcomes of frail individuals. The advantages of using apps for frailty assessment include their ease of use, ability to provide real-time feedback to the user, and their potential for improving the accuracy and reliability of frailty assessment [59,60]. An app designed to assess frailty offers numerous advantages and can revolutionize how frailty is identified. Its accessibility on smartphones and digital devices ensures easy and convenient frailty assessment, even in remote areas, while enabling early detection of an individual's physical and cognitive abilities, allowing for timely interventions.

However, a key challenge in implementing mHealth solutions for frailty is the digital divide. Frail individuals, especially those from disadvantaged backgrounds, may lack access to necessary technology. This gap exacerbates health inequalities, as those in need of mHealth interventions the most may have the least access to them [61]. Ensuring fair access to technology and mHealth tools is a critical ethical issue [62]. Creating mHealth apps and devices that are easy to use and accessible to frail individuals is essential. This involves overcoming challenges related to age-related impairments in vision, hearing, and dexterity, which were considered in the development of the FRAILSURVEY [63].

The advancements in mHealth for frailty care hold the potential for deeply personalized interventions. Using cutting-edge data analytics and artificial intelligence, mHealth platforms will be adept at customizing interventions according to the distinct needs and characteristics of each frail individual [64]. The integration of mHealth platforms with electronic health records holds the potential to enhance care coordination and facilitate seamless data sharing. This interoperability will streamline processes, improve communication, and, ultimately, raise the standard of service. Furthermore, the evolving landscape of mHealth in frailty care will require ongoing ethical dialogues.

XSL•FO
RenderX

As technological capabilities progress, ethical dilemmas around privacy, data security, informed consent, and fairness in algorithms will persist as paramount concerns. Navigating the delicate balance between innovation and ethical protection will present a constant challenge [65,66].

This study has some limitations that should be highlighted. Despite using a randomized, repeated crossover design and recruiting 522 participants from diverse settings, including community-dwelling older adults and nursing home residents, limitations persisted, only a fraction of participants underwent test-retest analysis, reducing the robustness of the results. In addition, the digital literacy barrier among older adults may have impacted the accessibility of the smartphone-based app, FRAILSURVEY. Furthermore, biases inherent in translating and validating scales into different languages may have influenced the study results. However, the similarity of psychometric properties to previous validation studies and the pioneering demonstration of the validity of the smartphone version offer promising insights into frailty assessment methods. This study marks a significant contribution to the field by establishing the scientific validity of the smartphone version of the GFI.

Conclusions

Overall, FRAILSURVEY uses a standardized and validated assessment tool, providing objective and consistent measurements, eliminating subjective biases, and enhancing accuracy and reliability. Compared with traditional in-person assessments, its cost-effectiveness can reduce health care expenses and increase accessibility. The future integration with electronic health records and other health-related data could enable personalized interventions tailored to an individual's specific needs and medical history, ultimately leading to more effective frailty management and improved overall health outcomes for at-risk individuals.

Acknowledgments

This work is financed by national funds from FCT-Fundação para a Ciência e a Tecnologia, I.P., in the scope of 2 projects (UIDP/04378/2020 and UIDB/04378/2020) of the Research Unit on Applied Molecular Biosciences-UCIBIO and a project (LA/P/0140/2020) of the Associate Laboratory Institute for Health and Bioeconomy-i4HB, and under the grant attributed to Luís Midão (SFRH/BD/137090/2018). Also, the authors want to thank all participants of the study.

Data Availability

The dataset generated during this study is not publicly accessible due to ethical considerations. However, both the dataset and the study metadata can be obtained from the corresponding author upon request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Final version of the European Portuguese version of the Groningen Frailty Indicator.
[DOCX File , 33 KB - formative_v9i1e51975_app1.docx]

References

1. Costa A, Henriques J, Alarcão V, Henriques A, Madeira T, Virgolino A, et al. Active aging awareness and well-being among older adults in portugal. Front Public Health 2023;11:1149731 [FREE Full text] [doi: 10.3389/fpubh.2023.1149731] [Medline: 37124788]
2. Instituto Nacional de Estatística. Censos de 2011. 2012. 2012. URL: https://censos.ine.pt/xportal/xmain?xpid=CENSOS&xpgid=censos2011_apresentacao [accessed 2025-01-25]
3. Instituto Nacional de EstatísticaI. Projeções de População Residente. 2020. URL: https://www.cpidoso.pt/populacao/ [accessed 2020-03-20]
4. Instituto Nacional de Estatística, I. Censos de 2021. 2022. URL: https://tinyurl.com/6b4b6usy [accessed 2025-01-25]
5. Midão L, Brochado P, Almada M, Duarte M, Paúl C, Costa E. Frailty status and polypharmacy predict all-cause mortality in community dwelling older adults in Europe. Int J Environ Res Public Health 2021;18(7):3580 [FREE Full text] [doi: 10.3390/ijerph18073580] [Medline: 33808273]
6. Tessier L, Wulf N, Momose Y. Long-term care in the context of population ageing: a rights-based approach to universal coverage. ILO Working Paper 2023 [FREE Full text] [doi: 10.54394/rhjd4529]
7. Cesari M, Prince M, Thiyagarajan JA, De Carvalho IA, Bernabei R, Chan P, et al. Frailty: an emerging public health priority. J Am Med Dir Assoc 2016;17(3):188-192. [doi: 10.1016/j.jamda.2015.12.016] [Medline: 26805753]
8. de Souto Barreto P, Morley JE, Chodzko-Zajko W, H Pitkala K, Weening-Djiksterhuis E, Rodriguez-Mañas L, International Association of GerontologyGeriatrics—Global Aging Research Network (IAGG-GARN)the IAGG European Region Clinical Section. Recommendations on physical activity and exercise for older adults living in long-term care facilities: a taskforce report. J Am Med Dir Assoc 2016;17(5):381-392. [doi: 10.1016/j.jamda.2016.01.021] [Medline: 27012368]

XSL•FO
RenderX

9. Fried LP, Tangen CM, Walston J, Newman AB, Hirsch C, Gottdiener J, Cardiovascular Health Study Collaborative Research Group. Frailty in older adults: evidence for a phenotype. J Gerontol A Biol Sci Med Sci 2001;56(3):M146-M156. [doi: 10.1093/gerona/56.3.m146] [Medline: 11253156]
10. Kojima G. Frailty as a predictor of future falls among community-dwelling older people: a systematic review and meta-analysis. J Am Med Dir Assoc 2015;16(12):1027-1033. [doi: 10.1016/j.jamda.2015.06.018] [Medline: 26255098]
11. Vermeiren S, Vella-Azzopardi R, Beckwée D, Habbig A, Scafoglieri A, Jansen B, Gerontopole Brussels Study group. Frailty and the prediction of negative health outcomes: a meta-analysis. J Am Med Dir Assoc 2016;17(12):1163.e1-1163.e17. [doi: 10.1016/j.jamda.2016.09.010] [Medline: 27886869]
12. Manfredi G, Midão L, Paúl C, Cena C, Duarte M, Costa E. Prevalence of frailty status among the European elderly population: findings from the survey of health, aging and retirement in Europe. Geriatrics Gerontology Int 2019;19(8):723-729. [doi: 10.1111/ggi.13689]
13. Daniels R, van Rossum E, de Witte L, Kempen GI, van den Heuvel W. Interventions to prevent disability in frail community-dwelling elderly: a systematic review. BMC Health Serv Res 2008;8:278 [FREE Full text] [doi: 10.1186/1472-6963-8-278] [Medline: 19115992]
14. de Vries NM, Staal J, van Ravensberg C, Hobbelen J, Olde Rikkert M, Nijhuis-van der Sanden M. Outcome instruments to measure frailty: a systematic review. Ageing Res Rev 2011;10(1):104-114. [doi: 10.1016/j.arr.2010.09.001] [Medline: 20850567]
15. Peters LL, Boter H, Buskens E, Slaets JP. Measurement properties of the groningen frailty indicator in home-dwelling and institutionalized elderly people. J Am Med Dir Assoc 2012;13(6):546-551 [FREE Full text] [doi: 10.1016/j.jamda.2012.04.007] [Medline: 22579590]
16. Steverink N, Slaets J, Schuurmans H, Van Lis M. Measuring frailty: developing and testing the GFI (Groningen Frailty Indicator). Gerontologist 2001;41(1):236-237.
17. Olaroiu M, Ghinescu M, Naumov V, Brinza I, Heuvel WVD. The psychometric qualities of the Groningen Frailty Indicator in Romanian community-dwelling old citizens. Fam Pract 2014;31(4):490-495. [doi: 10.1093/fampra/cmu022] [Medline: 24846305]
18. Xiang W, Cheng Y, Li Z, Han J, Li K. Cross-cultural adaptation and validation of the Groningen Frailty Indicator in Chinese nursing home residents. Aging Clin Exp Res 2020;32(6):1035-1042. [doi: 10.1007/s40520-019-01178-7] [Medline: 30900215]
19. Oetsma S, Boonen A, Starmans M, Peeters R, van Onna M. Validation of two frailty questionnaires in older patients with rheumatoid arthritis: a cross-sectional study. Clin Exp Rheumatol 2020;38(3):523-528 [FREE Full text] [Medline: 31694746]
20. Meessen JMTA, Leichtenberg CS, Tilbury C, Kaptein BL, Koster LA, Slagboom PE, et al. Frailty in end-stage hip or knee osteoarthritis: validation of the Groningen Frailty Indicator (GFI) questionnaire. Rheumatol Int 2018;38(5):917-924 [FREE Full text] [doi: 10.1007/s00296-017-3868-1] [Medline: 29147763]
21. Bras L, Peters T, Wedman J, Plaat B, Witjes M, van Leeuwen B, et al. Predictive value of the Groningen Frailty Indicator for treatment outcomes in elderly patients after head and neck, or skin cancer surgery in a retrospective cohort. Clin Otolaryngol 2015;40(5):474-482. [doi: 10.1111/coa.12409] [Medline: 25754107]
22. Castro MDL, Alves M, Martins A, Luísa Papoila A, Botelho M, Fragata J. Reproducibility and validity of the Portuguese Edmonton Frail Scale version in cardiac surgery patients. Rev Port Cardiol 2023;42(4):295-304 [FREE Full text] [doi: 10.1016/j.repc.2022.03.009] [Medline: 36738965]
23. Santiago LM, Silva R, Velho D, Rosendo I, Simões J. Cross-cultural adaptation and validation of the PRISMA-7 Scale for European Portuguese. Fam Med Prim Care Rev 2020;22(1):59-66. [doi: 10.5114/fmpcr.2020.92507]
24. Faria A, Sousa-Santos AR, Mendes J, Limas de Sousa AS, Amaral TF. Development of the Portuguese versions of FRAIL Scale and SARC-F: screening tools for physical frailty and sarcopenia (Desenvolvimento das versões portuguesas dos questionários FRAIL Scale e SARC-F: ferramentas de rastreio para a fragilidade física e sarcopenia). Acta Port Nutr 2021;26:90-94. [doi: 10.21011/apn.2021.2614]
25. Pereira Pinto M, Martins S, Mesquita E, Fernandes L. European Portuguese version of the Clinical Frailty Scale: translation, cultural adaptation and validation study. Acta Med Port 2021;34(11):749-760. [doi: 10.20344/amp.14543] [Medline: 33565417]
26. Cardoso AF, Bobrowicz-Campos E, Teixeira-Santos L, Cardoso D, Couto F, Apóstolo J. Validation and screening capacity of the European Portuguese version of the SUNFRAIL tool for community-dwelling older adults. Int J Environ Res Public Health 2021;18(4):1394 [FREE Full text] [doi: 10.3390/ijerph18041394] [Medline: 33546251]
27. Coelho T, Santos R, Paúl C, Gobbens RJ, Fernandes L. Portuguese version of the Tilburg Frailty Indicator: transcultural adaptation and psychometric validation. Geriatr Gerontol Int 2015;15(8):951-960. [doi: 10.1111/ggi.12373] [Medline: 25255891]
28. Smith MJ, Reiter MJ, Crist BD, Schultz LG, Choma TJ. Improving patient satisfaction through computer-based questionnaires. Orthopedics 2016;39(1):e31-e35. [doi: 10.3928/01477447-20151218-07] [Medline: 26709557]
29. Ahn DT. Benefits and risks of apps for patients. Curr Opin Endocrinol Diabetes Obes 2022;29(1):17-22. [doi: 10.1097/MED.0000000000000701] [Medline: 34878413]

XSL•FO
RenderX

30. Lecomte T, Potvin S, Corbière M, Guay S, Samson C, Cloutier B, et al. Mobile apps for mental health issues: meta-review of meta-analyses. JMIR Mhealth Uhealth 2020;8(5):e17458 [FREE Full text] [doi: 10.2196/17458] [Medline: 32348289]
31. Pagliari C, Sloan D, Gregor P, Sullivan F, Detmer D, Kahan JP, et al. What is eHealth (4): a scoping exercise to map the field. J Med Internet Res 2005;7(1):e9 [FREE Full text] [doi: 10.2196/jmir.7.1.e9] [Medline: 15829481]
32. Schuurmans H, Steverink N, Lindenberg S, Frieswijk N, Slaets JPJ. Old or frail: what tells us more? J Gerontol A Biol Sci Med Sci 2004;59(9):M962-M965. [doi: 10.1093/gerona/59.9.m962] [Medline: 15472162]
33. Bielderman A, van der Schans CP, van Lieshout MJ, de Greef MH, Boersma F, Krijnen WP, et al. Multidimensional structure of the Groningen Frailty Indicator in community-dwelling older people. BMC Geriatr 2013;13:86 [FREE Full text] [doi: 10.1186/1471-2318-13-86] [Medline: 23968433]
34. Wild D, Grove A, Martin M, Eremenco S, McElroy S, Verjee-Lorenz A, ISPOR Task Force for Translation and Cultural Adaptation. Principles of good practice for the translation and cultural adaptation process for patient-reported outcomes (PRO) measures: report of the ISPOR Task Force for Translation and Cultural Adaptation. Value Health 2005;8(2):94-104 [FREE Full text] [doi: 10.1111/j.1524-4733.2005.04054.x] [Medline: 15804318]
35. Kim JH, Kwon S, Shim SR, Sun HY, Ko YM, Chun D, et al. Validation and reliability of a smartphone application for the International Prostate Symptom Score questionnaire: a randomized repeated measures crossover study. J Med Internet Res 2014;16(2):e38. [doi: 10.2196/jmir.3042] [Medline: 24513507]
36. Streiner DL, Norman GR, Cairney. Health Measurement Scales: A Practical Guide to Their Development and Use. 5th ed. United Kingdom: Oxford University Press; 2015.
37. Cappelleri JC, Zou KH, Bushmakin AG, J. Alvir JM, Alemayehu D, Symonds T. Patient-Reported Outcomes: Measurement, Implementation and Interpretation. 1st ed. New York, NY: Chapman and Hall/CRC; 2014.
38. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016;15(2):155-163 [FREE Full text] [doi: 10.1016/j.jcm.2016.02.012] [Medline: 27330520]
39. Kaiser HF. Computer program for varimax rotation in factor analysis. Educ Psychol Meas 1959;19(3):413-420. [doi: 10.1177/001316445901900314]
40. Marston L. Introductory Statistics for Health and Nursing Using SPSS. London: Sage; 2009.
41. Shrestha N. Factor analysis as a tool for survey analysis. Am J Appl Math Stat 2021;9(1):4-11. [doi: 10.12691/ajams-9-1-2]
42. Sperber AD. Translation and validation of study instruments for cross-cultural research. Gastroenterology 2004;126(1 Suppl 1):S124-S128. [doi: 10.1053/j.gastro.2003.10.016] [Medline: 14978648]
43. Dambi JM, Corten L, Chiwaridzo M, Jack H, Mlambo T, Jelsma J. A systematic review of the psychometric properties of the cross-cultural translations and adaptations of the Multidimensional Perceived Social Support Scale (MSPSS). Health Qual Life Outcomes 2018;16(1):80 [FREE Full text] [doi: 10.1186/s12955-018-0912-0] [Medline: 29716589]
44. Donoghue OA, O'Shea M, McGarrigle CA, Lang D, Hoey C, Whitty H, et al. Translating frailty assessment methodologies and research-based evidence to clinical education and practice. J Aging Soc Policy 2022;34(4):537-551. [doi: 10.1080/08959420.2020.1777825] [Medline: 32634336]
45. Abbasi M, Rolfson D, Khera AS, Dabravolskaj J, Dent E, Xia L. Identification and management of frailty in the primary care setting. CMAJ 2018;190(38):E1134-E1140 [FREE Full text] [doi: 10.1503/cmaj.171509] [Medline: 30249759]
46. Khamis R, Sabbah H, Sabbah S, Peters L, Droubi N, Sabbah I. Evaluating the psychometric properties of the Arabic version of the Groningen Frailty Indicator among Lebanese elderly people. J Egypt Public Health Assoc 2019;94(1):28 [FREE Full text] [doi: 10.1186/s42506-019-0028-3] [Medline: 32813111]
47. Garre-Olmo J, Calvó-Perxas L, López-Pousa S, Gracia Blanco, MD, Vilalta-Franch J. Prevalence of frailty phenotypes and risk of mortality in a community-dwelling elderly cohort. Age Ageing 2013;42(1):46-51. [doi: 10.1093/ageing/afs047]
48. Wang J, Hulme C. Frailty and socioeconomic status: a systematic review. J Public Health Res 2021;10(3):2036 [FREE Full text] [doi: 10.4081/jphr.2021.2036] [Medline: 33942603]
49. Varadaraj V, Lee MJ, Tian J, Ramulu PY, Bandeen-Roche K, Swenor BK. Near vision impairment and frailty: evidence of an association. Am J Ophthalmol 2019;208:234-241 [FREE Full text] [doi: 10.1016/j.ajo.2019.08.009] [Medline: 31465753]
50. Tan M, Bhanu C, Frost R. The association between frailty and anxiety: a systematic review. Int J Geriatr Psychiatry 2023;38(5):e5918. [doi: 10.1002/gps.5918] [Medline: 37157226]
51. Luh RL, Yu, SY, Yang CH. Chinese version of Groningen Frailty Index: validation in a Chinese elderly population. Neuropsychiatry 2018;8(6):1949-1956.
52. Krabbe PF. Chapter 7—Validity. In: The Measurement of Health and Health Status. San Diego, CA: Academic Press; 2017:113-134.
53. Maaß L, Freye M, Pan C, Dassow H, Niess J, Jahnel T. The definitions of health apps and medical apps from the perspective of public health and law: qualitative analysis of an interdisciplinary literature overview. JMIR Mhealth Uhealth 2022;10(10):e37980 [FREE Full text] [doi: 10.2196/37980] [Medline: 36315221]
54. Wilson K. Mobile cell phone technology puts the future of health care in our hands. CMAJ 2018;190(13):E378-E379 [FREE Full text] [doi: 10.1503/cmaj.180269] [Medline: 29615420]
55. Rowland SP, Fitzgerald JE, Holme T, Powell J, McGregor A. What is the clinical value of mHealth for patients? NPJ Digit Med 2020;3:4 [FREE Full text] [doi: 10.1038/s41746-019-0206-x] [Medline: 31970289]

XSL•FO
RenderX

56. Akbar S, Coiera E, Magrabi F. Safety concerns with consumer-facing mobile health applications and their consequences: a scoping review. J Am Med Inform Assoc 2020;27(2):330-340 [FREE Full text] [doi: 10.1093/jamia/ocz175] [Medline: 31599936]
57. Gordon WJ, Landman A, Zhang H, Bates DW. Beyond validation: getting health apps into clinical practice. NPJ Digit Med 2020;3:14 [FREE Full text] [doi: 10.1038/s41746-019-0212-z] [Medline: 32047860]
58. Lagan S, Sandler L, Torous J. Evaluating evaluation frameworks: a scoping review of frameworks for assessing health apps. BMJ Open 2021;11(3):e047001 [FREE Full text] [doi: 10.1136/bmjopen-2020-047001] [Medline: 33741674]
59. Chang R, Low H, McDonald A, Park G, Song X. Web-based software applications for frailty assessment in older adults: a scoping review of current status with insights into future development. BMC Geriatr 2021;21(1):723 [FREE Full text] [doi: 10.1186/s12877-021-02660-6] [Medline: 34922466]
60. Morley JE, Vellas B, van Kan A, Anker SD, Bauer JM, Bernabei R, et al. Frailty consensus: a call to action. J Am Med Dir Assoc 2013;14(6):392-397 [FREE Full text] [doi: 10.1016/j.jamda.2013.03.022] [Medline: 23764209]
61. Mugueta-Aguinaga I, Garcia-Zapirain B. Is technology present in frailty? Technology a back-up tool for dealing with frailty in the elderly: a systematic review. Aging Dis 2017;8(2):176-195 [FREE Full text] [doi: 10.14336/AD.2016.0901] [Medline: 28400984]
62. Brall CP, Schröder-Bäck P, Maeckelberghe E. Ethical aspects of digital health from a justice point of view. Eur J Public Health 2019;29(Suppl 3):18-22 [FREE Full text] [doi: 10.1093/eurpub/ckz167] [Medline: 31738439]
63. Tajudeen FP, Bahar N, Tan MP, Peer Mustafa MB, Saedon NI, Jesudass J. Understanding user requirements for a senior-friendly mobile health application. Geriatrics (Basel) 2022;7(5):110 [FREE Full text] [doi: 10.3390/geriatrics7050110] [Medline: 36286212]
64. Bhatt P, Liu J, Gong Y, Wang J, Guo Y. Emerging artificial intelligence-empowered mHealth: scoping review. JMIR Mhealth Uhealth 2022;10(6):e35053 [FREE Full text] [doi: 10.2196/35053] [Medline: 35679107]
65. Jacobson NC, Bentley KH, Walton A, Wang SB, Fortgang RG, Millner AJ, et al. Ethical dilemmas posed by mobile health and machine learning in psychiatry research. Bull World Health Organ 2020;98(4):270-276 [FREE Full text] [doi: 10.2471/BLT.19.237107] [Medline: 32284651]
66. Zhang X, Saltman R. Impact of electronic health record interoperability on telehealth service outcomes. JMIR Med Inform 2022;10(1):e31837 [FREE Full text] [doi: 10.2196/31837] [Medline: 34890347]

Abbreviations

GFI: Groningen Frailty Indicator
ICC: intraclass correlation coefficient
mHealth: mobile health
PRISMA-7: Program on Research for Integrating Services for the Maintenance of Autonomy questionnaire

Edited by A Mavragani; submitted 21.08.23; peer-reviewed by S Pandey, H Veldandi, D Portela, E Prochaska, R Nooripour; comments to author 21.03.24; revised version received 09.04.24; accepted 11.04.24; published 07.03.25.

Please cite as:
Midao L, Duarte M, Sampaio R, Almada M, Dias CC, Paúl C, Costa E
FRAILSURVEY—an mHealth App for Self-Assessment of Frailty Based on the Portuguese Version of the Groningen Frailty Indicator: Validation and Reliability Study
JMIR Form Res 2025;9:e51975
URL: https://formative.jmir.org/2025/1/e51975
doi:10.2196/51975
PMID:40053720

©Luis Midao, Mafalda Duarte, Rute Sampaio, Marta Almada, Cláudia Camila Dias, Constança Paúl, Elísio Costa. Originally published in JMIR Formative Research (https://formative.jmir.org), 07.03.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Evaluation of the MyFertiCoach Lifestyle App for Subfertile Couples: Single-Center Evaluation of Augmented Standard Care

Jesper Smeenk[1], MD, PhD; Ellen Smit[2], BBA; Marc Jacobs[3], PhD; Ilse van Rooij[1], MD, PhD

[1]Elisabeth TweeSteden Ziekenhuis, Tilburg, The Netherlands

[2]Ferring BV, Hoofddorp, The Netherlands

[3]MSJ Advies, Helmond, The Netherlands

Corresponding Author:
Jesper Smeenk, MD, PhD
Elisabeth TweeSteden Ziekenhuis
Doctor Deelenlaan 5
Tilburg, 5042 AD
The Netherlands
Phone: 31 (013) 221 00 00
Email: j.smeenk@etz.nl

Abstract

Background: Many couples undergoing fertility treatment face multiple lifestyle risk factors that lower their chances of achieving pregnancy. The MyFertiCoach (MFC) app was designed as an integrated lifestyle program featuring modules on healthy weight management, nutrition, exercise, quitting smoking, reducing alcohol and drug use, and managing stress. We hypothesized that supplementing standard care with the MFC app would improve lifestyle outcomes.

Objective: This study aims to assess the impact of the MFC app on changing multiple lifestyle habits in women seeking fertility treatment. The primary outcome is the change in the total risk score (TRS) at 3- and six-month follow-ups. The TRS is calculated for each individual as the sum of all risk scores per behavior (eg, vegetable/fruit/folic acid intake, smoking, and alcohol use) at 3 and 6 months. A higher TRS indicates unhealthier nutrition and lifestyle habits and a lower likelihood of achieving pregnancy. The secondary endpoints include changes in BMI, activity score, preconception dietary risk score, distress score (eg, perceived burden), smoking habits, alcohol intake, and program adherence.

Methods: This retrospective, observational, single-center evaluation included patients between January 1, 2022, and December 31, 2023. Subfertile female patients aged 18-43 years and their partners, who were referred to a gynecologist, were invited to participate in online lifestyle coaching via the MFC app. The gynecologist selected relevant lifestyle modules based on the results of integrated screening questionnaires. We used (hierarchical) linear mixed models (LMMs) to estimate changes in outcomes. For missing data patterns deemed missing not at random, joint modeling was applied. Statistical significance was set at $P \leq .05$, with methods in place to maintain the same false-positive rate.

Results: A total of 1805 patients were invited to participate in the evaluation, with an average of 737 (40.83%) completing the screening questionnaire at baseline. For the TRS, 798 (44.21%) patients were included at baseline, of whom 517 (64.8%) involved their partner. On average, 282 of 744 (37.9%) patients submitted at least one follow-up questionnaire. Patients rated the app above average (n=137, median score of 7 on a 1-10 scale) on days 7 and 14. The TRS decreased by an average of 1.5 points ($P<.001$) at T3 and T6 compared with baseline, a clinically meaningful improvement. All secondary outcomes showed statistically significant positive changes for patients who used a relevant lifestyle module ($P<.001$). Most improvements were achieved by 3 months and remained significant at 6 months ($P<.001$), except for alcohol intake ($P<.53$). These findings were consistent across both LMMs and joint models.

Conclusions: Our evaluation of a mobile health app integrated into standard care demonstrates immediate and clinically meaningful improvements in key lifestyle parameters among women seeking to become pregnant. Additional scientific research is needed to identify the causal pathways leading to sustained effectiveness. To maintain and enhance these outcomes, further tailoring of patient-specific programs is essential.

(JMIR Form Res 2025;9:e64239) doi:10.2196/64239

XSL•FO
RenderX

KEYWORDS

fertility; mHealth; pregnancy; lifestyle; app; smartphone

Introduction

Many couples undergoing fertility treatment have multiple lifestyle risk factors, such as obesity, unbalanced nutritional habits, low physical activity, and alcohol consumption or smoking. These habits not only increase the risk of cardiovascular and metabolic diseases but also reduce the chance of becoming pregnant [1-4]. Reducing any or all of these lifestyle factors before pregnancy or assisted reproductive technology treatment can improve reproductive health [5-9]. In summary, the more negative lifestyle factors present, the lower the chance of becoming pregnant and the longer the time until pregnancy [10].

For instance, improving dietary patterns during the preconception period reduces the risk of several adverse birth outcomes, such as fetal growth restriction and babies born small or large for gestational age. Maternal complications, including gestational diabetes, hypertensive disorders, and premature delivery, also appear to decrease [11,12].

Moreover, cognitive behavioral therapy and psychological support during fertility treatment have been suggested to result in significantly more viable pregnancies compared with routine care [13-15].

While the evidence supporting the importance of improving and maintaining a healthy lifestyle is abundantly clear, changing habits and behaviors remains difficult [16]. Health-compromising behaviors are hard to stop, and health-promoting behaviors are challenging to adopt, making behavior change particularly difficult. As health care budgets face increasing pressure due to noncommunicable diseases associated with an aging population, workforce shortages, and rising costs of novel medical technologies, it is crucial that lifestyle interventions are cost-effective—for instance, through the use of online programs and apps [17].

Several lifestyle intervention programs for women actively seeking medical support to achieve pregnancy have shown high attrition rates and limited effects [18]. However, mobile health (mHealth) apps have the potential to overcome these obstacles by providing individualized, tailored, and repeated information [19,20]. Indeed, growing evidence suggests that mobile technology can effectively improve inadequate nutrition, lifestyle, and medication adherence [21,22].

Many lifestyle apps are currently available on the market, but very few are specifically designed for couples wishing to have children. To provide guidance for both patients and health care providers (HCPs), the MyFertiCoach (MFC) app was developed as a 6-month integrated lifestyle program with modules on healthy weight; nutrition; exercise; and quitting smoking, alcohol, drugs, and stress.

The MFC app supports couples trying to conceive and pregnant women in modifying unhealthy dietary and lifestyle habits. HCPs can use the app to help patients improve their lifestyle. Based on demographic parameters and an assessment of nutrition and lifestyle habits, a screening process is conducted, and a personalized coaching program consisting of lifestyle modules is created. Interactive elements, visual support, and motivational interviewing are incorporated as much as possible. Additionally, at various points during the program, we will assess whether lifestyle improvements have been achieved across different domains.

This retrospective, observational, single-center evaluation will be the first to assess the impact of the MFC app on changing 7 lifestyle habits in women seeking fertility treatment: healthy weight, diet, exercise, smoking, alcohol, drugs, and stress. Given the established influence of these lifestyle factors on subfertility, we expect that positive changes in these habits will also enhance reproductive health.

Methods

Overview

This is a retrospective, single-center evaluation conducted at the Elisabeth-TweeSteden Hospital in Tilburg, the Netherlands. This work does not assess the effectiveness of a new intervention before implementation but rather evaluates an addition to standard care after its integration.

Ethical Considerations

This study did not undergo formal approval by a medical ethics board for several reasons. First, the app was not purchased for scientific purposes but rather to augment standard care. As such, it received explicit approval from the board of directors for purchase and subsequently became part of standard care.

Additionally, all patients were free to use the app or not; their treatment remained unchanged if they chose not to use it. App usage was entirely voluntary, and data storage complied with national and European GDPR (General Data Protection Regulation) guidelines. The data were anonymized, and none of the researchers had direct access to them. Thus, the hospital purchased an app that was offered to patients as an optional resource. Each patient indicated whether they wanted to use the app and, if so, whether their anonymized data could be used for scientific purposes. As a result, the data set was generated with full patient consent and in compliance with privacy regulations.

Furthermore, our data analysis was conducted retrospectively, long after data collection had been completed. One of the authors (MJ) has previously used a similar approach, analyzing distress thermometer data stored in the electronic patient system [23]. For that study, the institutional review board was consulted, and an exemption was granted as it was classified as file research.

Adding a research paper is not the same as legislation, but the Dutch Central Committee on Research Involving Human Subjects states that “Retrospective research/research with patient files does not fall within the scope of the WMO, as the research subject is not physically involved in the research. Nor have the data that are to be researched been gathered for the sake of the

XSL•FO
RenderX

research. The research subjects do not have to change their behaviour for the sake of the research." [24]. We believe the last sentence is the most relevant: although patients filled out questionnaires and could use the app to support their care and contact their physician, the app did not require them to change their behavior for the sake of care or research. In fact, many patients we approached chose not to participate, and many who did eventually dropped out. They were free to do so. As the data were analyzed retrospectively, we consider our evaluation comparable to file research [25]. Following the decision tree for scientific medical research, our study falls under non-WMO research: it is medical-scientific research, but no patient was required to change procedures or behavior. The app was as freely available as any app that can be purchased from the Apple Store (Apple Inc.) or Google Play (Alphabet Inc.).

The MyFertiCoach App

The MFC app is a coaching program specifically designed to support lifestyle changes. It was developed by Ferring Pharmaceuticals B.V. in collaboration with a multidisciplinary team of health care professionals, with input from the patients' association Freya and patients who participated in the test phase.

HCPs can use the app to support patients and their partners while actively monitoring the (pre)active treatment phase. While the patient and their partner are motivated within the domains relevant to them, progress and results are visualized on a dashboard. Data essential for the next steps in treatment can also be used for evaluations and scientific research.

The MFC app promotes lifestyle changes over 6 months through motivation (using motivational interviewing techniques), information (on various domains), and relaxation (by incorporating game elements). It focuses on 7 key domains: achieving a healthy weight, maintaining a healthy diet, engaging in regular exercise, quitting smoking, stopping alcohol consumption, discontinuing drug/anabolic steroid use, and reducing stress through mindfulness. These domains were determined by the evaluation group and are supported by literature demonstrating their influence on the likelihood of a healthy pregnancy.

During the registration process, baseline demographic parameters and questionnaires on dietary and lifestyle habits are completed to screen patients on various factors, including BMI, smoking, nutrition, alcohol use, and stress. These assessments provide insights into patients' habits and help determine which domains require attention or intervention. The identified domains are then presented to the patient in the order of importance.

Patients were free to choose whether to complete the questionnaires, which took an average of 10 minutes in total. No financial compensation was provided for completing them.

To assess patients' knowledge, skills, and confidence in managing their own health or illness, both the patient and their partner are asked to complete the Patient Activation Measure questionnaire [26]. This provides HCPs with deeper insights into patients' perceived ability to improve their health, allowing for adjustments in the level of coaching and motivational interviewing techniques used.

At the first consultation, based on the screening outcomes, the HCP and patient jointly decide which aspects of their lifestyle to focus on. The relevant modules—healthy weight, healthy eating, healthy activity, smoking cessation, alcohol cessation, drug cessation, and stress reduction with mindfulness—are then activated. Once activated, patients receive module-specific questionnaires and general patient-reported outcome measures on a monthly basis. Patients are free to complete these at their convenience, but a baseline screening measure is required for inclusion in the evaluation. The program duration is set at 6 months.

Using the diary page, each patient can access all functionalities that allow them to record their daily mood and receive motivating messages. They can also send messages to their HCP with questions (eg, about nutrition or other lifestyle factors). The HCP can respond at their discretion. Each patient's progress is visible to the HCP via the MFC platform through domain-specific visualizations. Patients can track their progress using the app.

Recruitment and Data Collection

All new women aged 18-45 years who visited the fertility clinic were signed up for the MFC app after providing written consent and before their first appointment with the HCP. The MFC program, its modules, and the questionnaires were designed as a screening tool for lifestyle factors and were offered to all new patients as part of standard care.

The data extraction took place on February 1, 2024, and included all patients from January 1, 2022, to December 31, 2023, who completed the screening and had at least one activated module. Patients received a push notification and an email 1 month after the previous questionnaire. Patients who did not provide a baseline score and at least one follow-up score for an outcome (regardless of time) were excluded from further analysis of that particular outcome.

In the case of pregnancy, patients could check a box in the app indicating that they were pregnant (positive test). These patients would then receive additional questionnaires related to their pregnancy. As pregnancy is a defined endpoint, the formal intervention by the app was stopped, meaning these patients were no longer included in the evaluation. However, because some women indicated that they wanted to continue using the app, all modules—except the healthy weight module—remained active for 6 months. The app was never formally stopped by the physician, as it is part of standard care.

Outcomes

Assessing the Impact of the MFC App: Primary and Secondary Measures of Lifestyle Program Benefits

To properly evaluate the MFC app, several outcome measures have been defined to assess any additional benefits of our lifestyle program. The primary outcome was the change in the total risk score (TRS) [10] at the 3- and 6-month follow-ups. The secondary endpoints are changes in BMI, activity score, preconception dietary risk (PDR) score [27], distress score [28], smoking habits, alcohol intake, and program adherence.

XSL•FO
RenderX

Total Risk Score

The TRS is based on the Rotterdam Reproduction Risk Score (R3 score), the PDR score, and other existing evidence of associations with reproductive and pregnancy outcomes. An individual's TRS is defined as the sum of all risk scores per behavior (ie, based on vegetable, fruit, and folic acid intake, as well as smoking and alcohol use after 3 and 6 months) [10].

Vegetable and fruit intake were each subdivided into a risk score of 0, 1, 2, or 3. A score of 0 represents an adequate daily intake (≥200 g of vegetables per day and ≥2 pieces of fruit per day, respectively). Scores of 1 and 2 both indicate a "nearly adequate" intake (vegetable intake of 150 to <200 g and fruit intake of 1.5 to <2 pieces per day), with score 1 reflecting the participant's intention to change this risk factor and score 2 indicating no such intention. A score of 3 represents an inadequate daily intake (vegetable intake <150 g and fruit intake <1.5 pieces).

If a participant received a score of 1 or 2, an additional question regarding their intrinsic motivation was asked to determine whether they intended to improve their behavior related to this risk factor.

Folic acid supplement use was considered adequate (score 0) or inadequate (score 3) based on whether a participant met the recommendation of taking a 400-mg folic acid supplement daily during the periconceptional period. There is no evidence or recommendation for folic acid supplementation beyond 12 weeks of pregnancy, although this does not mean patients are not allowed to continue its use. Therefore, if a patient became pregnant while using the app, took folic acid for the first 12 weeks of pregnancy, and then stopped, she received a score of 0 for folic acid supplement use. This approach ensured consistency in scoring across all patients, allowing for a more conservative estimate of the app's benefits.

Risk scores for smoking and alcohol consumption were based on average daily use: no smoking (score 0), smoking 1-5 cigarettes (score 1), 6-14 cigarettes (score 3), or ≥15 cigarettes (score 6); and no drinking (score 0), drinking <1 alcoholic beverage per day (score 1), 1-2 beverages per day (score 2), or ≥2 beverages per day (score 3).

An important distinction is that the TRS is calculated separately for females and their partners (male/female). A higher TRS indicates unhealthier nutrition and lifestyle habits, which lowers the chance of becoming pregnant [10]. Based on previous findings, where a 1-point increase in the TRS was associated with a 21% reduction in pregnancy likelihood, we considered a 1-point decrease in the TRS clinically relevant.

Other Scores

We considered a BMI (calculated as weight divided by height squared) between 18.5 and 25 kg/m^2 to represent a "normal" weight [29]. A BMI ≥29 kg/m^2 is associated with a lower chance of becoming pregnant, with each unit increase above 29 linked to a 4% decrease in fertility [2]. Therefore, we considered a 1-point drop in the high BMI group clinically relevant.

To determine the activity score, we followed the Dutch physical activity guidelines, which state that adults should be physically active for at least 30 minutes a day on at least five days per week. This activity can include sports, walking, cycling, or strenuous household chores.

To compute a score, we devised 2 sets of 3 questions. The first set assessed the number of active minutes per week (ie, <75, 75 to <150, and ≥150 minutes), while the second set assessed muscle- and bone-enhancing activities (ie, never, once a week, or at least twice a week). Scores ranged from 1 to 3 for each set, with a total possible score between 2 and 6.

The level of distress was measured using the distress score, which provides a general impression of the problems couples experience in 5 areas: practical problems, family/social problems, emotional problems, religious/spiritual problems, and physical problems [30]. Additionally, couples were separately asked to rate their level of distress using a "thermometer" ranging from 0 to 10. The physician could then explore with the couple where their greatest concerns lay and discuss available options to address or reduce any symptoms or concerns as much as possible. A score of 5 or higher on the thermometer indicated a serious "burden." For our analysis, we only used the thermometer scores.

The PDR score was based on 6 nutritional questions covering the intake of key food groups, in line with dietary recommendations from the Netherlands Nutrition Centre. According to current guidelines, individuals should consume at least four slices of whole wheat bread daily (or an equivalent amount of cereals); use monounsaturated or polyunsaturated oils; and consume ≥200 g of vegetables daily, ≥2 pieces of fruit daily, ≥3 servings of meat or meat substitutes weekly, and ≥1 serving of fish weekly. Patients received 1 point for each recommendation met, with a maximum PDR score of 6, representing highly adequate nutrition according to the Netherlands Nutrition Centre's guidelines.

The smoking and alcohol scores were based on the number of cigarettes or alcoholic beverages consumed per day. Adherence was measured by the number of active patients over time.

Statistical Analysis

Evaluation Plan

This evaluation was conducted in accordance with a Statistical Analysis Plan (see Multimedia Appendix 1).

Descriptives and Transformations

We used descriptive statistics for each outcome of interest: the number of patients and observations for frequency data, and the mean and 95% CI for continuous data. For each outcome, we also calculated the number and percentage of responses per time point, as well as the mean and SD of response time. This allowed us to track response levels at each time point and assess the appropriateness of the predefined time-point categories (eg, 3 months, 6 months).

Some outcomes are known composites of aggregated parameters (eg, BMI), while others are transformations based on the specific manual of the questionnaire. We followed each manual's

XSL•FO
RenderX

guidelines on handling missing data when transforming or aggregating specific questionnaire items.

Missing Data

There is no official cut-off for what constitutes an acceptable or unacceptable level of missingness. Additionally, the reason for missing data is often unclear, except in cases where a participant has died. Therefore, we conducted a missing pattern analysis to determine the amount of missing data for each parameter separately and to identify how often multiple parameters are missing simultaneously.

We examined missing patterns across all items in the data set and aimed to classify them into 3 recognized types of missing data patterns: missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR) [31]. Of these 3 patterns, MNAR is the most problematic [32,33], as the missingness occurs for a specific reason related to the actual score a participant could have provided but did not.

As there are no formal statistical tests for assessing the size and nature of missing data, we relied heavily on visual inspection using bar charts, box plots, kernel density plots, and matrix plots. For example, if a questionnaire is missing at a particular time point and continues to be missing at subsequent time points, this would appear as a linear increase in missing data. Moreover, if factors such as age or having a partner are likely to influence responses, we would observe different response distributions across groups.

Power

As this evaluation was exploratory in nature, we did not conduct a formal power analysis to estimate the sample size required to detect a specific effect size with a given level of statistical significance. Statistical significance was determined by checking whether the 95% CI excluded 0.

Statistical Models Used

The main objective of this evaluation is to investigate the effect of following the MFC app program on changes in the TRS after 3 and 6 months. We analyzed the secondary outcomes (ie, BMI, activity score, PDR score, distress score, smoking habits, and alcohol intake) in the same way as the primary outcome, except for program adherence, which was described using descriptive statistics.

To estimate change, our primary analysis focused on patients for whom the intended module was made available. Accordingly, we used the following combinations: activity score for users of the activity module; alcohol score for users of the alcohol module; BMI for users of the weight module; last scores for users of the mindfulness module; the PDR score for users of the nutrition module; and smoking habits for users of the smoking module. The TRSs were analyzed for all participants who responded during the evaluation.

Using a maximum of 7 measurements per patient (1 baseline measurement and 6 follow-up measurements), and in the absence of a clinically relevant cut-off value, we estimated change using a (hierarchical) linear mixed model (LMM) [34]. LMMs are regression models designed to analyze clustered (longitudinal) data and can accommodate both MCAR and MAR data [35].

An important note is that we did not build separate models for the 3- and 6-month time points. As longitudinal data are correlated (ie, multiple measurements from the same patient), we constructed a single model from which the statistical significance of change can be derived at any time point up to the last observed measurement. Using an LMM, we can include time as either a categorical variable (eg, 1 month, 3 months, or 6 months) or as a continuous variable, represented by the number of months or the actual time in days. This approach allows us to evaluate the app's effect using days rather than months.

Because of the correlated nature of the data, we developed several LMMs with different random effects structures: random intercept, random slope, and random intercept with random slope [36]. This modeling strategy allows us to account for different assumptions, such as patients having the same starting point but different trajectories (random slope), the same trajectory but different starting points (random intercept), or both different starting points and different trajectories (random intercept and random slope).

Model Selection

The primary explanatory factors included in the models are time, partner status, age, the number of active modules, and the specific module associated with an outcome (eg, the nutrition module for the PDR score). To appropriately model time as a continuous variable, we will use linear functions, polynomial terms, or natural cubic splines [37].

To evaluate the relevance of these factors, we will iteratively remove parameters from a fully specified fixed-effects model using the chi-square test, considering both main effects and interaction effects. Model accuracy and parsimony will be assessed using the F test (analysis of variance) and the Akaike information criterion (AIC), with lower AIC values indicating better model fit. Model assumptions will be evaluated through graphical diagnostics and visualizations.

It is important to note that we will not build separate models for the 3- and 6-month time points. Instead, we will construct a single model from which the statistical significance of change can be derived at any time point up to the last observed measurement.

The analysis of the primary and secondary outcomes will result in several different models. This raises concerns about spurious results, as each model carries a 5% false-positive rate when determining statistical significance. To safeguard against such statistically spurious findings, it is important to assess whether the 7 outcomes are correlated at any given time point.

To address this, we will first create a correlation matrix using Spearman rank correlation, as many of the outcomes are ordinal in nature. If correlations are consistently higher than 0.50 between outcomes and across time points, we will include the correlated outcomes as predictors in the original LMM models. For example, if there is a correlation between distress and BMI, the distress outcome model will include both time and BMI as predictors.

XSL•FO
RenderX

As all variables have a clinical connection to the outcomes of interest, we allowed this form of statistical reasoning to guide variable inclusion. However, the final decision was based on model estimates for each variable, taking into account that multicollinearity can inflate SEs and affect variance estimates in mixed models. In summary, we included variables deemed both clinically important and statistically necessary.

Multiple Imputations in the Case of MNAR

Although patients may exhibit intermittent missingness—commonly associated with MAR situations and addressed using MAR models or multiple imputations—a sudden stop (ie, monotone missingness or drop out) strongly indicates an MNAR response. This is problematic for our analysis if it occurs before the 3- or 6-month follow-up. To address this, we will create a joint model, which combines an LMM with a survival model [38,39]. This approach is recognized as a viable alternative to older methods such as pattern-mixture modeling [40]. One of the authors has previously applied this method successfully to analyze quality-of-life changes in a population of patients with palliative pancreatic cancer [41].

The joint model yields 2 key outputs: (1) a survival model on program adherence, which incorporates information from the LMM; and (2) an LMM that adjusts its estimates based on the survival model. For each joint model, we used Markov chain Monte Carlo with 4 chains and 100,000 iterations, of which 10,000 were designated as burn-in. Model fit was assessed using trace and density plots.

Sensitivity Analyses

Adjusting for Timing Discrepancies in Longitudinal Data: Realignment of Assessment Periods

When analyzing longitudinal data, there is often a discrepancy between the intended timing of questionnaire returns and the actual return dates. This means that a 3-month assessment of an intervention may, in practice, represent a time frame that is either longer or shorter than planned. To address this, we created new time categories that better align with the intended schedule of questionnaire returns. For example, T3 is defined as 90 days plus or minus 14 days, and T6 as 180 days plus or minus 14 days.

Statistical Package Used

All analyses were performed using R Statistical Software (v4.1.2; R Foundation), a widely recognized and accepted program for statistical analysis. For the LMMs, we used the lme4 package. For the joint models, we used the JMbayes2 package in conjunction with the nlme package for constructing the LMM component. The survival package was used to build the Cox proportional hazards component of the joint models.

Results

Patient Characteristics

Table 1 presents the response rates during the evaluation period, along with response times, which varied by questionnaire (Figure S4 in Multimedia Appendix 2) and over time (Figure S2 in Multimedia Appendix 2). In total, 1815 patients were invited to participate, and all of them agreed to join, achieving a 100% recruitment rate. Of these, 818 patients (45.07%) agreed to participate. The mean age of the invited population was 34.1 (SD 5.46) years. A comparison between responders and nonresponders showed similar age distributions: 33.0 (SD 5.03) years for responders versus 35.1 (SD 5.63) years for nonresponders (Figure S1 in Multimedia Appendix 2).

XSL•FO
RenderX

Table 1. Number and percentage of responses, and the mean and SD of the response time.

Questionnaire and condition	Measurement moments							
	Screening	T0	T1	T2	T3	T4	T5	T6
Total risk score								
Regardless of the active module								
n (%)	1815 (100)	546 (30.1)	270 (14.9)	148 (8.2)	91 (5.0)	50 (2.8)	36 (2.0)	21 (1.2)
Mean (SD)	N/A[a]	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Activity								
Regardless of the active module								
n (%)	1815 (100)	758 (41.8)	235 (12.9)	139 (7.7)	85 (4.7)	50 (2.7)	34 (1.9)	20 (1.1)
Mean (SD)	N/A	0 (0)	55.6 (40.9)	89.6 (25.1)	120.6 (23.5)	154.1 (26.3)	186.3 (24.0)	220.3 (27.8)
Only active module exercise								
n (%)	444 (100)	438 (98.6)	233 (52.5)	138 (31.1)	85 (19.1)	50 (11.3)	34 (7.7)	20 (4.5)
Mean (SD)	N/A	0 (0)	55.8 (41.0)	89.7 (25.1)	120.6 (23.5)	154.1 (26.3)	186.3 (24.0)	220.6 (27.8)
Alcohol								
Regardless of the active module								
n (%)	1815 (100)	761 (41.9)	210 (11.6)	113 (6.2)	61 (3.4)	38 (2.1)	23 (1.3)	7 (0.4)
Mean (SD)	N/A	0 (0)	34.4 (23.7)	62.0 (24.0)	92.2 (26.2)	130.9 (31.9)	158.6 (31.2)	196.3 (45.6)
Only active module alcohol								
n (%)	209 (100)	199 (95.2)	95 (45.5)	42 (20.1)	26 (12.4)	14 (6.7)	11 (5.3)	5 (2.4)
Mean (SD)	N/A	0 (0)	46.2 27.2)	76.2 (27.1)	104.3 (28.4)	154.5 (38.4)	177.4 (35.2)	211.8 (45.5)
BMI (kg/m ^{2})								
Regardless of the active module								
n (%)	1815 (100)	792 (43.6)	209 (11.5)	115 (6.3)	73 (4.0)	45 (2.5)	32 (1.8)	19 (1.0)
Mean (SD)	N/A	0 (0)	57.9 (43.5)	93.3 (30.1)	125.9 (27.7)	154 (25.2)	187.1 (25.3)	225.8 (27.6)
Only active module weight								
n (%)	420 (100)	416 (99)	207 (49.3)	115 (27.4)	73 (17.4)	45 (10.7)	32 (7.6)	19 (4.5)
Mean (SD)	N/A	0 (0)	57.9 (43.5)	93.3 (30.1)	125.9 (27.7)	154.0 (25.2)	187.1 (25.3)	225.8 (27.6)
Distress								
Regardless of the active module								
n (%)	1815 (100)	764 (42.1)	84 (4.6)	42 (2.3)	27 (1.5)	17 (0.9)	13 (0.7)	7 (0.4)
Mean (SD)	N/A	0 (0)	65.7 (64.5)	115.5 (89.9)	160.8 (110.3)	203.8 (135.5)	220.0 (108.2)	223.9 (28.7)
Only active module mindfulness								
n (%)	157 (100)	153 (97.5)	83 (52.9)	42 (26.8)	27 (17.2)	17 (10.8)	13 (8.3)	7 (4.5)
Mean (SD)	N/A	0 (0)	66.4 (64.5)	115.5 (89.9)	160.8 (110.3)	203.8 (135.5)	220.0 (108.2)	223.9 (28.7)
Preconception dietary risk								

XSL•FO
RenderX

Questionnaire and condition	Measurement moments							
	Screening	T0	T1	T2	T3	T4	T5	T6
Regardless of the active module								
n (%)	1815 (100)	783 (43.1)	380 (20.9)	218 (12.0)	131 (7.2)	80 (4.4)	56 (3.1)	32 (1.8)
Mean (SD)	N/A	0 (0)	55.2 (42.3)	87.8 (34.6)	123 (42.5)	155.7 (50.2)	192.6 (58.1)	222.9 (26.9)
Only active module nutrition								
n (%)	693 (100)	680 (98.1)	379 (54.7)	218 (31.5)	131 (18.9)	80 (11.5)	56 (8.1)	32 (4.6)
Mean (SD)	N/A	0 (0)	55.3 (42.3)	87.8 (34.6)	123 (42.5)	155.7 (50.2)	192.6 (58.1)	222.9 (26.9)
Smoking								
Regardless of the active module								
n (%)	1815 (100)	761 (41.9)	579 (31.9)	375 (20.7)	262 (14.4)	191 (10.5)	131 (7.2)	94 (5.2)
Mean (SD)	N/A	0 (0)	20 (20.4)	44.4 (24.5)	71.7 (32.6)	88.6 (40.0)	114.1 (41.9)	129.6 (51.6)
Only active module smoking								
n (%)	130 (100)	119 (91.5)	82 (63.1)	53 (40.8)	42 (32.3)	28 (21.5)	18 (13.8)	11 (8.5)
Mean (SD)	N/A	0 (0)	16.4 (16.5)	41.4 (20.5)	63.9 (29.6)	85.0 (34.8)	105.8 (33.2)	126.2 (40.5)

[a]N/A: not applicable.

Among the 818 participants, 532 (65%) also included their partner. The partner-to-no-partner ratio differed significantly between those who accepted the invitation (532/818, 65.0% vs 286/818, 35.0%) and those who did not (323/997, 32.4% vs 674/997,67.6%; χ^2_1=190.81; P<.001).

When age distribution was further stratified by partner status, the results remained consistent. For nonparticipants, the mean ages were 34.0 (SD 4.59) years with a partner and 35.6 (SD 6.02) years without a partner. For participants, the mean ages were 32.1 (SD 4.22) years with a partner and 34.7 (SD 4.22) years without a partner.

Ultimately, data from 798 of 1815 patients (43.97% of those invited) were included in the analysis. These participants had a mean age of 33.0 (SD 5.04) years, with 517 (64.8%) including their partner. Of the 784 (98.2%) participants who reported their educational background, the majority (423/784, 53.9%) held a higher professional or university degree. Notably, higher education levels were associated with greater follow-up rates (see Figure S5 in Multimedia Appendix 2).

MFC Programs and Grading by Patients

Patients engaged with a diverse range of personal programs. Specifically, 693 patients included a nutrition module, while 130 patients included a smoking module. In total, the distribution of active modules per patient was as follows: we included 113 patients with 1 active module, 271 patients with 2 active modules, and 289 with 3 active modules; 4, 5, and 6 active modules were attached to 89, 33, and 3 patients, respectively. When asked to grade the MFC app on the seventh and fourteenth day of using the app (Figure S6 in Multimedia Appendix 2), 115 and 68 patients, respectively, gave a median score of 7 (IQR 3).

Patient Response and Response Time

The percentage of patients who filled out a questionnaire dropped rapidly across time (Figure S9 in Multimedia Appendix 2). The percentage of patients who had both a baseline score and at least one follow-up score (Figure S11 in Multimedia Appendix 2) ranged from 84 out of 764 (11.0%) patients for the distress score to 579 out of 761 (76.1%) patients for the smoking score, regardless of the module included. Across questionnaires, we received an average follow-up response rate of 38.3% (average of 282 patients/average of 737 patients). Compared with patients who did not return a follow-up questionnaire (see Figures S12-S14 and S16 in Multimedia Appendix 2), those who did respond had different median baseline scores for BMI (29.4 vs 24.3 kg/m^2), exercise (3 vs 5), distress score (6 vs 2), and alcohol (2 vs 0.1). However, the median scores for the TRS (3), PDR (4), and smoking (0) did not differ at follow-up. The follow-up percentage did differ much between those who included a partner and those who did not (see Figure S8 in Multimedia Appendix 2). When we accounted for the number of active modules, the graphs showed a slightly different pattern (see Figures S7, S8, and S10 in Multimedia Appendix 2), but the data became too sparse for additional statistical analysis. At baseline, the response rate ranged from 41.76% (758/1815) to 43.64% (792/1815), regardless of the active module, and from 91.5% (119/130) to 99.0% (416/420) per active module. These percentages declined over time, reaching 8.5% (11/130) for the smoking module and 2.4% (5/209) for the alcohol module at the last time point (see Figure S15 in Multimedia Appendix 2). The average response time did not align with the expected response time. While all patients completed their baseline questionnaire on time, the average response time for the first follow-up, regardless of the module, ranged from 20 days (smoking score) to 65.7 days (distress score). When considering only patients with active

XSL•FO
RenderX

modules, the response times were 16.4 and 66.4 days for the smoking and distress scores, respectively.

As time progressed, the gap between actual and expected response times widened. At the fifth time point—approximately 5 months after the follow-up—response times, regardless of the module, ranged from 220 days for the distress score to 114 days for the smoking score. The longest response time was observed for BMI, with an average of 226 days for the 6-month follow-up.

Sensitivity Analyses

Although secondary outcomes showed correlations at each time point, none consistently exceeded 0.5 (see Figures S24 and S25 in Multimedia Appendix 2). Therefore, we did not use any secondary outcomes to explain each other.

When we created new time categories that better aligned with the expected return of questionnaires (eg, T3 was defined as 90 days plus minus 14 days, T6 as 180 days plus minus 14 days), we reduced our effective sample size (see Figure S3 in Multimedia Appendix 2). However, this did not alter the distribution of values across the outcomes (see Figure S14 in Multimedia Appendix 2). Consequently, we used the Tx categories for our LMM models (main analysis) while using the actual time until questionnaire submission for our joint models, which also accounted for missing data.

Missing Data Pattern Analysis

The missing data pattern analysis revealed a monotone pattern (see Figures S31-S36 in Multimedia Appendix 2), meaning that once a patient stopped returning questionnaires, they dropped out entirely. This suggests a potential MNAR pattern. Consequently, we successfully fitted joint models for the TRS (see Figure S19 in Multimedia Appendix 2), activity (see Figure S26 in Multimedia Appendix 2), alcohol intake (see Figure S27 in Multimedia Appendix 2), BMI (see Figure S29 in Multimedia Appendix 2), PDR (see Figure S30 in Multimedia Appendix 2), and smoking habits (see Figure S28 in Multimedia Appendix 2). The results from the joint models align with the outcomes of the LMM models, although they are somewhat harder to interpret due to the inclusion of splines. Therefore, we reported the output of the LMM models, as they better followed the original questionnaire sampling. A joint model for distress did not converge. In summary, the MNAR analysis via joint modeling did not warrant further investigation.

Results for the Primary and Secondary Outcomes

The model selection results (see Figures S17-S23 in Multimedia Appendix 2) show 14 models for alcohol intake, BMI, activity, and the distress score. For PDR and smoking habits, we created 15 models, while for the TRS, we included 13 different models. In almost all selection procedures, the AIC was lowest for models that included a main effect of time (as a dummy variable), partner, age, and the number of active modules.

Figure 1 and Table 2 show the results for primary and secondary outcomes at baseline, 3-month, and 6-month follow-ups. The number of patients and observations included ranged from 424 patients (1040 observations) for the TRS to 83 patients (272 observations) for the distress score.

XSL•FO
RenderX

Figure 1. Predicted response by best model per outcome and partner.

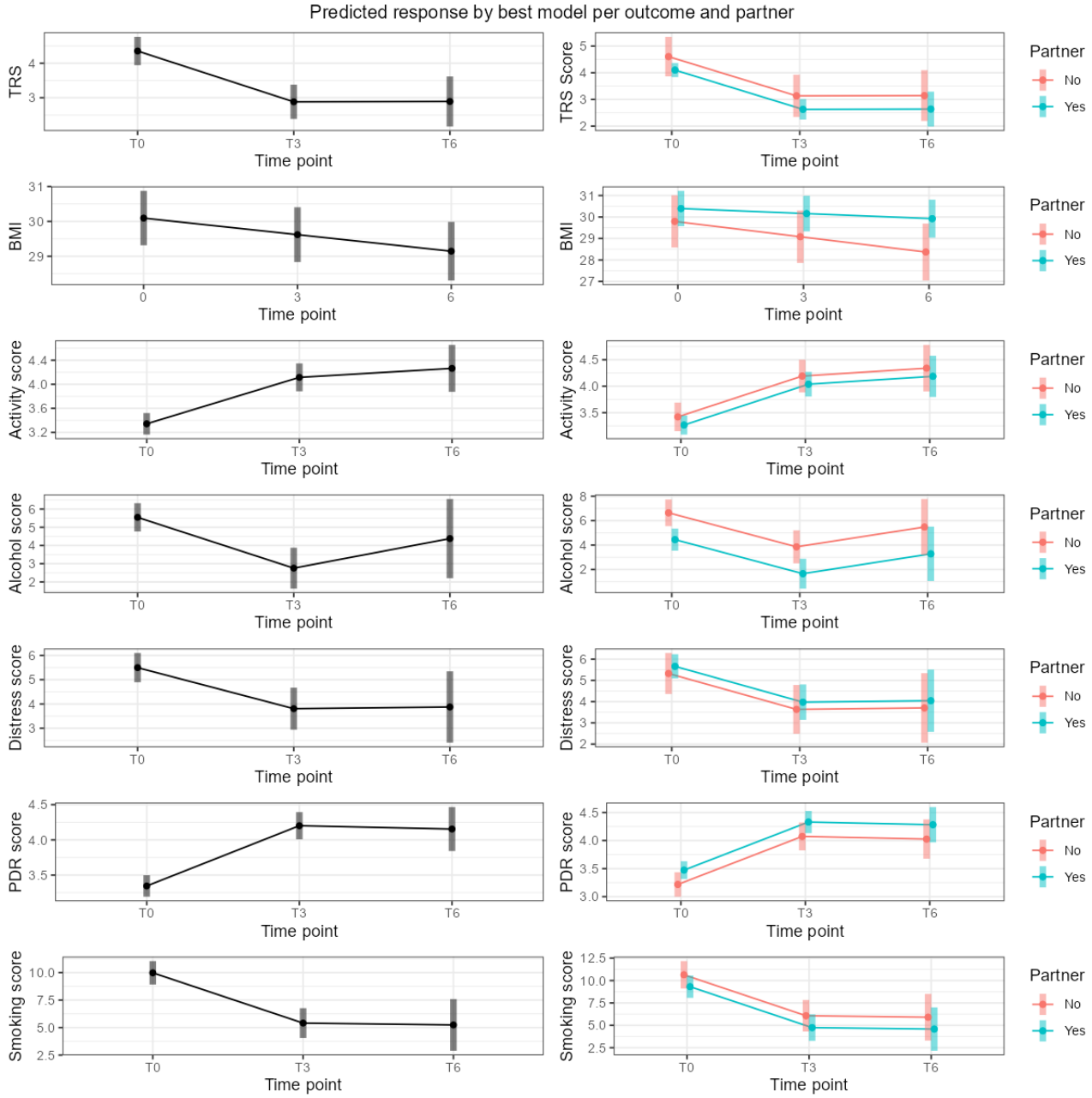

XSL•FO
RenderX

Table 2. Results following the MFC[a] app at the 3- and 6-month follow-up.[b]

Questionnaire	Values		Measurement moments									
			T0		T3				T6			
	N	Observed	Mean	95% CI	Mean	95% CI	Δ[c]	*P* value	Mean	95% CI	Δ	*P* value
Total risk score	424	1040	3.81	3.43-4.19	2.34	1.87-2.81	–1.47	<.001	2.35	1.65-3.06	–1.46	<.001
Activity	230	786	3.41	3.25-3.57	4.18	3.97-4.40	0.77	<.001	4.33	3.95-4.7	0.92	<.001
Alcohol	94	286	5.43	4.71-6.14	2.64	1.55-3.72	–2.79	<.001	4.26	2.10-6.42	–1.16	.53
BMI (kg/m^2)	207	698	30.1	29.4-30.8	29.6	28.9-30.3	–0.47	<.001	29.1	28.4-29.9	–0.95	<.001
Distress	83	272	5.53	4.94-6.12	3.84	2.99-4.70	–1.69	<.001	3.91	2.45-5.37	–1.62	.07
Preconception dietary risk	265	375	3.42	3.30-3.55	4.28	4.11-4.45	0.86	<.001	4.23	3.93-4.53	0.81	<.001
Smoking	313	81	9.95	8.89-11.02	5.38	4.03-6.75	–4.75	<.001	5.22	2.87-7.56	–4.74	<.001

[a]MFC: MyFertiCoach.

[b]Results are based on the most optimal model following an extensive model selection procedure.

[c]Δ: difference between categories.

For all outcomes, the change in score compared with baseline was statistically significant at 3 months (at the 5% level or *P*<.001) and considered positive (eg, lower BMI, reduced smoking, or improved nutrition). These changes remained statistically significant (*P*<.001)at 6 months, except for alcohol intake (*P*<.53) and distress levels (*P*<.07). For alcohol, part of the initial positive response (Δ –2.79; *P*<.001) diminished by 6 months (Δ –1.16; *P*<.531). For distress, the initial statistically significant change (Δ –1.69; *P*<.001) shifted to a positive trend (Δ –1.62; *P*<.073). BMI was the only outcome that continued improving over 6 months (30.1 vs 29.6 vs 29.1 kg/m^2; *P*<.001). For the TRS, the changes were clinically relevant at both 3 and 6 months. For BMI, clinical relevance was only observed at 6 months.

When we stratified the longitudinal analysis by partner presence (Figure 1 and Table 3), statistical significance was maintained at 3 months (*P*<.001), except for BMI, which was no longer significant in the group that included a partner (30.4 vs 30.2 kg/m^2 vs 29.9 kg/m^2; *P*<.07). However, the group without a partner showed an even greater decrease in BMI (29.8 vs 29.1 vs 28.4 kg/m^2; *P*<.001). Regardless of partner status, the change in alcohol intake remained nonsignificant (*P*=.53).

XSL•FO
RenderX

Table 3. Results following the MFC[a] app at the 3- and 6-month follow-up.[b]

Questionnaire	Partner	Values		Measurement moments									
				T0		T3				T6			
		N	Observed	Mean	95% CI	Mean	95% CI	Δ[c]	*P* value	Mean	95% CI	Δ	*P* value
Total risk score		424	1040										
	Yes			3.56	3.35-3.77	2.1	1.74-2.44	1.47	<.001	2.1	1.47-2.73	1.46	<.001
	No			4.07	3.34-4.79	2.59	1.82-3.37	1.47	<.001	2.6	1.67-3.54	1.46	<.001
Activity		230	786										
	Yes			3.33	3.17-3.50	4.1	3.89-4.33	–0.77	<.001	4.26	3.87-4.64	–0.9	<.001
	No			3.49	3.23-3.75	4.26	3.96-4.56	–0.77	<.001	4.41	3.98-4.84	–0.9	<.001
Alcohol		94	286										
	Yes			4.33	3.44-5.21	1.5	0.32-2.75	2.79	<.001	3.2	0.93-5.39	1.2	.53
	No			6.52	5.51-7.54	3.7	2.44-5.03	2.79	<.001	5.4	3.09-7.63	1.2	.53
BMI (kg/m^2)		207	698										
	Yes			30.4	29.6-31.2	30.2	29.3-31.0	–0.23	.07	29.9	29.0-31.0	–0.5	.07
	No			29.8	28.6-31.0	29.1	27.9-30.3	–0.71	<.001	28.4	27.0-29.7	–1.4	<.001
Distress		83	272										
	Yes			5.70	5.14-6.26	4.01	3.18-4.84	1.69	<.001	4.08	2.62-5.54	1.62	.07
	No			5.36	4.40-6.32	3.67	2.53-4.82	1.69	<.001	3.74	2.11-5.37	1.62	.07
Preconception dietary risk		1265	375										
	Yes			3.55	3.42-3.69	4.41	4.23-4.59	–0.86	<.001	4.36	4.06-4.66	–0.81	<.001
	No			3.30	3.10-3.49	4.15	3.92-4.38	–0.86	<.001	4.10	3.77-4.44	–0.81	<.001
Smoking		313	81										
	Yes			9.29	8.03-10.54	4.72	3.23-6.20	–4.75	<.001	4.55	2.13-6.97	–4.74	<.001
	No			10.64	9.09-12.14	6.04	4.29-7.80	–4.75	<.001	5.88	3.28-8.47	–4.74	<.001

[a]MFC: MyFertiCoach.

[b]Results are based on the most optimal model following an extensive model selection procedure.

[c]Δ: difference between categories.

When we further analyzed outcomes by partner presence and time point (Table 4), we observed lower alcohol scores (Δ –2.2; *P*<.001) and higher PDR scores (Δ 0.256; *P*=.03) in patients who included their partner. These were both main effects. However, for BMI, we identified an interaction effect, showing a more pronounced decline at 6 months (Δ –1.555; *P*=.04) in patients who did not include their partner.

XSL•FO
RenderX

Table 4. Results following the MFC[a] app at the 3- and 6-month follow-up[b].

Questionnaire and partner	Value		Measurement moments								
			T0			T3			T6		
	N	Observed	Mean	Δ[c] 95% CI	Δ *P* value	Mean	Δ 95% CI	Δ *P* value	Mean	Δ 95% CI	*P* value
Total risk score	424	1040									
Yes			3.56	0.505 (–0.244 to 1.25)	.19	2.1	0.505 (–0.244 to 1.25)	.19	2.1	0.505 (–0.244 to 1.25)	.19
No			4.07	N/A[d]	N/A	2.59	N/A	N/A	2.6	N/A	N/A
Activity	230	786									
Yes			3.33	0.154 (–0.135 to 1.051)	.29	4.1	0.154 (–0.135 to 1.051)	.29	4.3	0.154 (–0.135 to 1.051)	.29
No			3.49	N/A	N/A	4.3	N/A	N/A	4.4	N/A	N/A
Alcohol	94	286									
Yes			4.33	2.2 (0.922 to 3.47)	<.001	1.5	2.2 (0.922 to 3.47)	<.001	3.2	2.2 (0.922 to 3.47)	<.001
No			6.52	N/A	N/A	3.7	N/A	N/A	5.4	N/A	N/A
BMI (kg/m^2)	207	698									
Yes			30.4	–0.599 (–1.96 to 0.767)	.39	30.2	–1.077 (–2.46 to 0.3021)	.12	29.9	–1.555 (–3.06 to –0.0518)	.04
No			29.8	N/A	N/A	29.1	N/A	N/A	28.4	N/A	N/A
Distress	83	272									
Yes			5.70	–0.339 (–1.37 to 0.691)	.51	4.01	–0.339 (–1.37 to 0.691)	.51	4.08	–0.339 (–1.37 to 0.691)	.51
No			5.36	N/A	N/A	3.67	N/A	N/A	3.74	N/A	N/A
Preconception dietary risk	1265	375									
Yes			3.55	–0.256 (–0.484 to –0.028)	.03	4.41	–0.256 (–0.484 to –0.028)	.03	4.36	–0.256 (–0.484 to –0.028)	.03
No			3.30	N/A	N/A	4.15	N/A	N/A	4.10	N/A	N/A
Smoking	313	81									
Yes			9.29	–0.909 (–0.485 to 1.458)	.15	4.72	–0.909 (–0.485 to 1.458)	.15	4.55	–0.909 (–0.485 to 1.458)	.15
No			10.64	N/A	N/A	6.04	N/A	N/A	5.88	N/A	N/A

[a]MFC: MyFertiCoach.

[b]Results are based on the most optimal model following an extensive model selection procedure.

[c]Δ: difference between categories.

[d]N/A: not applicable.

Discussion

Principal Findings

Our findings indicate that introducing a lifestyle app can immediately improve several lifestyle factors associated with subfertility and the success of fertility treatment. Specifically, the assessment of the TRS shows that app usage leads to a clinically relevant improvement. The peak effect of the intervention appears to occur around the third month, after which improvements plateau—except for BMI, which remained clinically relevant at 6 months. The MFC app was designed with input from both patients and health care professionals to ensure it targeted lifestyle domains deemed most useful. As a result, the app was intended to be more than just a digital

XSL•FO
RenderX

summary of suboptimal lifestyle behaviors. Through shared decision-making, both patients and HCPs contributed to the final selection of active modules, integrating the app into the broader treatment plan. From a preventive medicine perspective, the app provided a valuable starting point and contributed to more efficient care, offering an inexpensive way to influence lifestyle. This impact is also reflected in the data, where most effects were observed immediately after app use began.

However, it appears that not everyone is eager to start an intervention of this type. While we cannot determine with certainty why some individuals chose not to participate, we observed lower participation rates among those with lower levels of education and those without a partner. It remains unclear whether the presence of a partner played a significant role in the acceptance of this evaluation. However, even when better follow-up led to a more favorable effect of the app, partner participation did not consistently enhance its effectiveness.

On an individual level, the reasons why some patients stopped using the app—or never started—remain uncertain. One plausible explanation is that certain users did not identify with the app, despite extensive screening and discussions. Another possibility is that aspects of the app's design or usability did not appeal to everyone, although we do not have specific data to confirm this.

In summary, we believe the MFC app is a cost-effective addition to standard care, helping subfertile women and their partners implement necessary lifestyle changes. Financial costs are minimal, and health care professionals require little time investment. While the app primarily attracts a specific group—those with higher education and a partner—it does lead to meaningful improvements in key lifestyle factors.

Limitations

Because of the nature of our evaluation, we were unable to account for all known and unknown confounders. Despite conducting several sensitivity analyses, we cannot rule out the possibility that the temporal differences—or lack thereof—observed between patients, including those with a partner, may be influenced by factors beyond patient characteristics or the app itself. Additionally, the pattern of missing data cannot be formally assessed in a quantitative manner, meaning we do not know why patients dropped out when they did.

Furthermore, we did not include a formal comparative group, so we cannot determine how outcomes would have developed in patients who did not receive guidance through the app. These limitations—along with others, such as the lack of randomization or blinding—prevent us from drawing causal inferences.

Many uncertainties remain. For example, we do not know why patients dropped out, nor do we have full transparency regarding the exact selection process. Although modules were chosen through shared decision-making, it is unclear whether the final choice was driven solely by motivation. However, this does not imply that a lack of improvement is due to low motivation or suboptimal care. It is possible that the interventions themselves did not sufficiently support patients in improving suboptimal lifestyle factors.

Special consideration should be given to folic acid use and pregnancy. While we did not extract these data, it is possible that some patients became pregnant while using the app and continued to use it. As Dutch guidelines recommend folic acid supplementation during the first 3 months of pregnancy, some patients may not have used folic acid, leading to a higher TRS. If this is the case, the app's beneficial effect could be even greater than our current estimates suggest.

The consistency of findings across LMMs and joint models does not eliminate the potential limitations of our study. While LMMs can handle data that are MAR and joint models can address missing-not-at-random (MNAR) data by incorporating estimates from survival analysis, no statistical safeguard can fully account for the high amount of missing data in our evaluation. Ultimately, there is no way to account for what we do not know.

As a result, we cannot claim that the high probability of MNAR data was entirely mitigated by the joint models. However, we can state that incorporating survival models alongside mixed models produced estimates that were very similar to those obtained using mixed models alone. This strengthens our confidence in the robustness of the model estimates given the available data. Nonetheless, our findings should be interpreted with caution due to the high level of missing data—an inherent limitation when evaluating interventions in standard care settings.

Additionally, our findings are based on a select group of patients from a single center in the Netherlands, where we observed a high rate of attrition during implementation. As we cannot fully determine the reasons for dropout, there is a potential risk of bias, as those who continued participating may differ systematically from those who dropped out.

Comparison With Prior Work

In the Netherlands, 2 other groups are actively developing apps to support treatment: the international LIFESTYLE study, in which the Medical University of Groningen participates [6,7,42], and the Smarter Pregnancy platform at Erasmus Medical Center [43]. However, these initiatives primarily involve hands-on interventions, making them financially expensive and time-intensive, although they tend to yield better and more stable responses. This suggests that maintaining patient motivation may require a similar level of investment from health care professionals.

Lifestyle changes are notoriously difficult to achieve and sustain. Eating, drinking, and smoking are habits that provide immediate gratification, while their negative effects manifest only in the long term. Psychological stressors are also challenging to address and vary depending on their nature—for example, losing a loved one differs from facing persistent financial difficulties. Consequently, our findings of immediate but not long-lasting changes align with what has been extensively documented in previous research.

Patients varied significantly in their personal trajectories. Some showed substantial improvement, others experienced only marginal benefits, and some saw no measurable change despite their efforts. Analyzing heterogeneous data often results in

XSL•FO
RenderX

marginal main effects which, even when statistically significant, may not translate into meaningful clinical change. Attempts to include interaction effects—such as partner presence, number of active modules, and age—did not provide clearer insights into why some patients responded well while others did not.

However, given the nature of the app and the study setting, the observed dropout rates were not unexpected. We aimed to facilitate lifestyle changes in a predominantly working-class population with limited health literacy. This does not imply that these patients cannot benefit from or are unwilling to engage with such interventions, but rather that additional efforts are necessary to achieve long-term change. Therefore, evaluations of mHealth apps should not rely solely on questionnaires but should also incorporate in-depth follow-up interviews, which we plan to conduct in future development stages.

Conclusions

Our evaluation of an mHealth app in standard care, designed to support women seeking pregnancy, demonstrated an immediate and clinically relevant improvement in key lifestyle parameters. Further research is needed to identify causal pathways leading to sustained effectiveness. To maintain and enhance these results, patient-specific programs must be further tailored.

Acknowledgments

We thank Thomas Bachet, Naomi Tielen, the patient association group Freya, and the nurses at the Elisabeth TweeSteden Hospital for their useful insights and support. We also thank our “thinktank”: Prof. dr. A Hoek (University Medical Center Groningen), Prof. dr. J Laven (Erasmus Medical Center), Dr. M. Schoonenberg (Nij Geertgen) and Drs. G Van den Dool (Nij Linge).

Data Availability

The data sets generated during or analyzed during this evaluation are not publicly available due to privacy reasons. However, because the patients give explicit consent that their data can be anonymously used for scientific research, the data are available from the corresponding author upon reasonable request.

Conflicts of Interest

ES works at Ferring B.V. The remaining authors have declared no conflicts of interest.

Multimedia Appendix 1
Supplementary figures.
[ZIP File (Zip Archive), 2713 KB - formative_v9i1e64239_app1.zip]

Multimedia Appendix 2
Statistical analysis plan.
[DOCX File , 600 KB - formative_v9i1e64239_app2.docx]

References

1. Godfrey KM, Gluckman PD, Hanson MA. Developmental origins of metabolic disease: life course and intergenerational perspectives. Trends Endocrinol Metab 2010 Apr;21(4):199-205. [doi: 10.1016/j.tem.2009.12.008] [Medline: 20080045]
2. van der Steeg JW, Steures P, Eijkemans MJ, Habbema JDF, Hompes PG, Burggraaff JM, et al. Obesity affects spontaneous pregnancy chances in subfertile, ovulatory women. Hum Reprod 2008 Feb 14;23(2):324-328. [doi: 10.1093/humrep/dem371] [Medline: 18077317]
3. Fitzsimons KJ, Modder J, Greer IA. Obesity in pregnancy: risks and management. Obstet Med 2009 Jun 22;2(2):52-62. [doi: 10.1258/om.2009.090009] [Medline: 27582812]
4. Schneider E, Hamer O, Smith C, Hill J. Evaluating the association of female obesity with the risk of live birth following IVF: implications for clinical practice. Pract Midwife 2024 Jan 16;27(4):20240101-20240451. [doi: 10.1093/humupd/dmz011] [Medline: 39045260]
5. Hassan MA, Killick SR. Negative lifestyle is associated with a significant reduction in fecundity. Fertil Steril 2004 Feb;81(2):384-392 [FREE Full text] [doi: 10.1016/j.fertnstert.2003.06.027] [Medline: 14967378]
6. Einarsson S, Bergh C, Friberg B, Pinborg A, Klajnbard A, Karlström PO, et al. Weight reduction intervention for obese infertile women prior to IVF: a randomized controlled trial. Hum Reprod 2017 Aug 01;32(8):1621-1630. [doi: 10.1093/humrep/dex235] [Medline: 28854592]
7. Mutsaerts MA, van Oers AM, Groen H, Burggraaff JM, Kuchenbecker WKH, Perquin DAM, et al. Randomized trial of a lifestyle program in obese infertile women. N Engl J Med 2016 May 19;374(20):1942-1953. [doi: 10.1056/NEJMoa1505297] [Medline: 27192672]
8. Hoek A, Wang Z, van Oers AM, Groen H, Cantineau AE. Effects of preconception weight loss after lifestyle intervention on fertility outcomes and pregnancy complications. Fertil Steril 2022 Sep;118(3):456-462 [FREE Full text] [doi: 10.1016/j.fertnstert.2022.07.020] [Medline: 36116799]

XSL•FO
RenderX

9. Oostingh EC, Hall J, Koster MP, Grace B, Jauniaux E, Steegers-Theunissen RP. The impact of maternal lifestyle factors on periconception outcomes: a systematic review of observational studies. Reprod Biomed Online 2019 Jan;38(1):77-94. [doi: 10.1016/j.rbmo.2018.09.015] [Medline: 30424937]
10. van Dijk MR, Koster MP, Willemsen SP, Huijgen NA, Laven JS, Steegers-Theunissen RP. Healthy preconception nutrition and lifestyle using personalized mobile health coaching is associated with enhanced pregnancy chance. Reprod Biomed Online 2017 Oct;35(4):453-460. [doi: 10.1016/j.rbmo.2017.06.014] [Medline: 28688924]
11. van Oers AM, Mutsaerts MAQ, Burggraaff JM, Kuchenbecker WKH, Perquin DAM, Koks CAM, et al. Association between periconceptional weight loss and maternal and neonatal outcomes in obese infertile women. PLoS ONE 2018 Mar 28;13(3):e0192670. [doi: 10.1371/journal.pone.0192670]
12. Sauder KA, Starling AP, Shapiro AL, Kaar JL, Ringham BM, Glueck DH, et al. Diet, physical activity and mental health status are associated with dysglycaemia in pregnancy: the Healthy Start Study. Diabet Med 2016 Mar 03;33(5):663-667 [FREE Full text] [doi: 10.1111/dme.13093] [Medline: 26872289]
13. Steegers-Theunissen R, Hoek A, Groen H, Bos A, van den Dool G, Schoonenberg M, et al. Pre-Conception Interventions for Subfertile Couples Undergoing Assisted Reproductive Technology Treatment: Modeling Analysis. JMIR Mhealth Uhealth 2020 Nov 23;8(11):e19570. [doi: 10.2196/19570]
14. Verkuijlen J, Verhaak C, Nelen WLDM, Wilkinson J, Farquhar C. Psychological and educational interventions for subfertile men and women. Cochrane Database Syst Rev 2016 Mar 31;3(3):CD011034 [FREE Full text] [doi: 10.1002/14651858.CD011034.pub2] [Medline: 27031818]
15. Gaitzsch H, Benard J, Hugon-Rodin J, Benzakour L, Streuli I. The effect of mind-body interventions on psychological and pregnancy outcomes in infertile women: a systematic review. Arch Womens Ment Health 2020 Jan 02;23(4):479-491. [doi: 10.1007/s00737-019-01009-8]
16. Blüher M. Obesity: global epidemiology and pathogenesis. Nat Rev Endocrinol 2019 May 27;15(5):288-298. [doi: 10.1038/s41574-019-0176-8] [Medline: 30814686]
17. de Jong MJ, van der Meulen-de Jong AE, Romberg-Camps MJ, Becx MC, Maljaars JP, Cilissen M, et al. Telemedicine for management of inflammatory bowel disease (myIBDcoach): a pragmatic, multicentre, randomised controlled trial. The Lancet 2017 Sep;390(10098):959-968. [doi: 10.1016/s0140-6736(17)31327-2]
18. Hiller R, Griesinger G. How effective are lifestyle interventions for overweight women trying to conceive? Curr Opin Obstet Gynecol 2023 Jun 01;35(3):230-237. [doi: 10.1097/GCO.0000000000000874] [Medline: 37185354]
19. Meyers AJ, Domar AD. Research-supported mobile applications and internet-based technologies to mediate the psychological effects of infertility: a review. Reprod Biomed Online 2021 Mar;42(3):679-685. [doi: 10.1016/j.rbmo.2020.12.004] [Medline: 33487558]
20. Timmers T, Keijsers M, Kremer JAM, Janssen L, Smeenk J. Supporting women undergoing IVF treatment with timely patient information through an app: randomized controlled trial. JMIR Mhealth Uhealth 2021 Aug 27;9(8):e28104 [FREE Full text] [doi: 10.2196/28104] [Medline: 34448725]
21. Free C, Phillips G, Watson L, Galli L, Felix L, Edwards P, et al. The effectiveness of mobile-health technologies to improve health care service delivery processes: a systematic review and meta-analysis. PLoS Med 2013 Jan 15;10(1):e1001363. [doi: 10.1371/journal.pmed.1001363] [Medline: 23458994]
22. Wei HX, Yang YL, Luo TY, Chen WQ. Effectiveness of mobile health interventions for pregnant women with gestational diabetes mellitus: a systematic review and meta-analysis. Journal of Obstetrics and Gynaecology 2023 Aug 22;43(2):1-13. [doi: 10.1080/01443615.2023.2245906] [Medline: 37605977]
23. Jacobs M, Anderegg MCJ, Schoorlemmer A, Nieboer D, Steyerberg EW, Smets EMA, et al. Patients with oesophageal cancer report elevated distress and problems yet do not have an explicit wish for referral prior to receiving their medical treatment plan. Psycho-Oncology 2016 Jun;26(4):452-460. [doi: 10.1002/pon.4161]
24. The Central Committee on Research Involving Human Subjects. Non-WMO research. The Central Committee on Research Involving Human Subjects. 2024. URL: https://english.ccmo.nl/investigators/additional-requirements-for-certain-types-of-research/non-wmo-research [accessed 2024-11-20]
25. The Central Committee on Research Involving Human Subjects. File research. The Central Committee on Research Involving Human Subjects. 2024. URL: https://english.ccmo.nl/investigators/additional-requirements-for-certain-types-of-research/non-wmo-research/file-research [accessed 2024-11-20]
26. Rademakers J, Nijman J, van der Hoek L, Heijmans M, Rijken M. Measuring patient activation in the Netherlands: translation and validation of the American short form Patient Activation Measure (PAM13). BMC Public Health 2012 Jul 31;12(1):1-7. [doi: 10.1186/1471-2458-12-577]
27. Twigt JM, Bolhuis MEC, Steegers EAP, Hammiche F, van Inzen WG, Laven JSE, et al. The preconception diet is associated with the chance of ongoing pregnancy in women undergoing IVF/ICSI treatment. Hum Reprod 2012 Aug 15;27(8):2526-2531. [doi: 10.1093/humrep/des157] [Medline: 22593431]
28. van Nuenen FM, Donofrio SM, Tuinman MA, van de Wiel HBM, Hoekstra-Weebers JEHM. Feasibility of implementing the 'Screening for Distress and Referral Need' process in 23 Dutch hospitals. Support Care Cancer 2016 Aug 26;25(1):103-110. [doi: 10.1007/s00520-016-3387-8]

XSL•FO
RenderX

29. Silvestris E, Lovero D, Palmirotta R. Nutrition and female fertility: an interdependent correlation. Front Endocrinol (Lausanne) 2019 Jun 7;10:346-313. [doi: 10.3389/fendo.2019.00346] [Medline: 31231310]
30. Ownby KK. Use of the distress thermometer in clinical practice. J Adv Pract Oncol 2019 Mar;10(2):175-179 [FREE Full text] [Medline: 31538028]
31. Carpenter JR, Smuk M. Missing data: a statistical framework for practice. Biom J 2021 Jun 24;63(5):915-947. [doi: 10.1002/bimj.202000196] [Medline: 33624862]
32. Eekhout I, de Boer RM, Twisk JWR, de Vet HCW, Heymans MW. Missing data: a systematic review of how they are reported and handled. Epidemiology 2012 Sep;23(5):729-732. [doi: 10.1097/EDE.0b013e3182576cdb] [Medline: 22584299]
33. Fielding S, Fayers PM, McDonald A, McPherson G, Campbell MK, RECORD Study Group. Simple imputation methods were inadequate for missing not at random (MNAR) quality of life data. Health Qual Life Outcomes 2008 Aug 04;6(1):57 [FREE Full text] [doi: 10.1186/1477-7525-6-57] [Medline: 18680574]
34. Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens MHH, et al. Generalized linear mixed models: a practical guide for ecology and evolution. Trends Ecol Evol 2009 Mar;24(3):127-135. [doi: 10.1016/j.tree.2008.10.008] [Medline: 19185386]
35. Gabrio A, Plumpton C, Banerjee S, Leurent B. Linear mixed models to handle missing at random data in trial-based economic evaluations. Health Econ 2022 Jun 02;31(6):1276-1287 [FREE Full text] [doi: 10.1002/hec.4510] [Medline: 35368119]
36. Murphy J, Weaver NE, Hendricks AE. Accessible analysis of longitudinal data with linear mixed effects models. Dis Model Mech 2022 May 01;15(5):1-8. [doi: 10.1242/dmm.048025] [Medline: 35521689]
37. Elhakeem A, Hughes RA, Tilling K, Cousminer DL, Jackowski SA, Cole TJ, et al. Using linear and natural cubic splines, SITAR, and latent trajectory models to characterise nonlinear longitudinal growth trajectories in cohort studies. BMC Med Res Methodol 2022 Mar 15;22(1):1-20. [doi: 10.1186/s12874-022-01542-8]
38. Ibrahim JG, Chu H, Chen LM. Basic concepts and methods for joint models of longitudinal and survival data. JCO 2010 Jun 01;28(16):2796-2801. [doi: 10.1200/jco.2009.25.0654]
39. Asar O, Ritchie J, Kalra PA, Diggle PJ. Joint modelling of repeated measurement and time-to-event data: an introductory tutorial. Int J Epidemiol 2015 Feb;44(1):334-344. [doi: 10.1093/ije/dyu262] [Medline: 25604450]
40. Shen C. Application of pattern-mixture models to outcomes that are potentially missing not at random using pseudo maximum likelihood estimation. Biostatistics 2005 Apr 01;6(2):333-347. [doi: 10.1093/biostatistics/kxi013]
41. Tello Valverde CP, Ebrahimi G, Sprangers MA, Pateras K, Bruynzeel AM, Jacobs M, et al. Impact of short-course palliative radiation therapy on pancreatic cancer-related pain: prospective phase 2 nonrandomized PAINPANC trial. Int J Radiat Oncol Biol Phys 2024 Feb 01;118(2):352-361. [doi: 10.1016/j.ijrobp.2023.08.055] [Medline: 37647972]
42. van Oers AM, Mutsaerts MAQ, Burggraaff JM, Kuchenbecker WKH, Perquin DAM, Koks CAM, LIFEstyle study group. Cost-effectiveness analysis of lifestyle intervention in obese infertile women. Hum Reprod 2017 Jul 01;32(7):1418-1426. [doi: 10.1093/humrep/dex092] [Medline: 28486704]
43. Van Dijk MR, Huijgen NA, Willemsen SP, Laven JS, Steegers EA, Steegers-Theunissen RP. Impact of an mHealth platform for pregnancy on nutrition and lifestyle of the reproductive population: a survey. JMIR Mhealth Uhealth 2016 May 27;4(2):e53. [doi: 10.2196/mhealth.5197] [Medline: 27234926]

Abbreviations

AIC: Akaike information criterion
GDPR: General Data Protection Regulation
HCP: health care provider
LMM: linear mixed model
MAR: missing at random
MCAR: missing completely at random
MFC: MyFertiCoach
mHealth: mobile health
MNAR: missing not at random
PDR: preconception dietary risk
TRS: total risk score

XSL•FO
RenderX

Edited by A Mavragani; submitted 12.07.24; peer-reviewed by J Grieger, T Wang; comments to author 06.09.24; revised version received 25.09.24; accepted 17.01.25; published 10.03.25.

Please cite as:
Smeenk J, Smit E, Jacobs M, van Rooij I
Evaluation of the MyFertiCoach Lifestyle App for Subfertile Couples: Single-Center Evaluation of Augmented Standard Care
JMIR Form Res 2025;9:e64239
URL: https://formative.jmir.org/2025/1/e64239
doi:10.2196/64239
PMID:

©Jesper Smeenk, Ellen Smit, Marc Jacobs, Ilse van Rooij. Originally published in JMIR Formative Research (https://formative.jmir.org), 10.03.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Development of a Voice-Activated Virtual Assistant to Improve Insomnia Among Young Adult Cancer Survivors: Mixed Methods Feasibility and Acceptability Study

Hunter Groninger[1,2,3], MD; Hannah Arem[3,4], PhD; Lylian Ayangma[3], BA; Lisa Gong[3], BA; Eric Zhou[5], PhD; Daniel Greenberg[6], BA

[1]MedStar Health Research Institute, Washington, DC, United States
[2]MedStar Washington Hospital Center, Washington, DC, United States
[3]Georgetown University School of Medicine, Washington, DC, United States
[4]MedStar Health Research Institute, Hyattsville, MD, United States
[5]Dana Farber Cancer Institute, Boston, MA, United States
[6]MediaRez LLC, Washington, DC, United States

Corresponding Author:
Hunter Groninger, MD
MedStar Health Research Institute
c/o 110 Irving Street NW
Room 2A68
Washington, DC, 20010
United States
Phone: 1 202 877 7445
Email: hunter.groninger@medstar.net

Abstract

Background: Up to 75% of young adult cancer survivors (YACS) experience chronic insomnia, negatively affecting physical and emotional health and overall quality of life. Cognitive behavioral therapy for insomnia (CBT-I) is a gold-standard intervention to address insomnia. To improve CBT-I access and treatment adherence, screen-based digital CBT-I platforms have been developed. However, even with these digital products, widespread uptake of CBT-I remains limited, and new strategies for CBT-I delivery are warranted.

Objective: The objective of this study is to understand how YACS experience insomnia and how they might incorporate technology-delivered CBT-I into a daily routine and test the feasibility and acceptability of a novel screen-free voice-activated virtual assistant–delivered CBT-I prototype.

Methods: Eligible participants—ages 18-39, living with a history of cancer (any type, any stage), self-reporting on average less sleep than National Sleep Foundation recommendations, and English-speaking—were recruited from a major urban cancer center, 2 regional oncology clinics, and 2 cancer survivorship support groups. We conducted 4 focus groups to understand the YACS experience of insomnia, their routine use of technology at home, particularly voice-activated virtual assistants such as Amazon Alexa, and input on how CBT-I might be delivered at home through a smart speaker system. We developed a prototype device to deliver key elements of CBT-I at home along with circadian lighting and monitoring of post-bedtime device use, collected YACS user perspectives on this prototype, and then conducted a single-arm feasibility and acceptability study.

Results: In total, 26 YACS (6-7 participants per group) experiencing insomnia participated in focus groups to share experiences of insomnia during cancer survivorship and to provide input regarding a CBT-I prototype. Common triggers of insomnia included worry about disease management and progression, disease-related pain and other symptoms, choices regarding personal device use, and worry about the impact of poor sleep on daily functioning. In total, 12 participants completed device prototype testing, engaging with the prototype 94% of the assigned times (twice daily for 14 days; meeting predetermined feasibility cutoff of engagement ≥70% of assigned times) and rating the prototype with an overall mean score of 5.43 on the Satisfaction subscale of the Usability, Satisfaction, and Ease of Use scale (range 4.42-7; exceeding the predetermined cutoff score for acceptability of 5.0). All participants completing the study reported they would be interested in using the prototype again and would recommend it to someone else with insomnia.

XSL•FO
RenderX

Conclusions: YACS were highly engaged with our voice-activated virtual assistant–delivered CBT-I prototype and found it acceptable to use. Following final device development, future studies should evaluate the efficacy of this intervention among YACS.

Trial Registration: ClinicalTrials.gov NCT05875129; https://clinicaltrials.gov/study/NCT05875129

(JMIR Form Res 2025;9:e64869) doi:10.2196/64869

KEYWORDS

cancer; survivor; insomnia; cognitive behavioral therapy; technology; app; oncology; mobile health; artificial intelligence; young adults; sleep; mHealth; mobile health; CBT; voice-activated virtual assistant; virtual assistants; focus group; qualitative research

Introduction

Insomnia, characterized by the struggle to initiate or maintain sleep or the experience of nonrestorative sleep, brings along symptoms that hinder daily functioning, such as weariness, concentration difficulties, and mood disruptions [1]. Among the estimated 650,000 young adult cancer survivors (YACS, ages 18-39 years old) in the United States, up to 75% are grappling with insomnia symptoms, with nearly 30% likely to experience insomnia disorder [2-8]. Untreated insomnia is associated with a myriad of psychological and medical consequences, exacerbating the long-term impact of cancer diagnosis and treatment [3,4,9-11].

Insomnia adversely affects cardiometabolic and immune system health, neurobehavioral function, depression, fatigue, and overall quality of life [9,12-14]. These repercussions can also result in reduced physical activity and increased sedentary behavior, both linked to negative health outcomes in YACS [4,15]. In comparison to those with regular sleep patterns, cancer survivors dealing with insomnia tend to have more medical visits, double the hospitalization rate, and a higher reliance on medication [16]. Furthermore, insomnia serves as a risk factor for various mental illnesses, with the greatest risk ratios observed in depressive disorders. Despite its persistent nature during and after cancer treatment, discussions about insomnia with health care providers remain infrequent [17]. Addressing the need for accessible and effective insomnia treatment targeted to support this population with distinct cancer trajectories due to diagnosis at a young age is imperative.

Cognitive behavioral therapy for insomnia (CBT-I) is the gold standard intervention recommended by the National Comprehensive Cancer Network (NCCN) to address insomnia in cancer survivor populations [18-20]. Previous studies have demonstrated the effectiveness of CBT-I among YACSs [21,22]. However, widespread accessibility to CBT-I remains challenging due to an insufficient number of trained providers, a lack of insurance coverage, and even language barriers [23,24]. The administration of CBT-I demands specialized training, and its delivery, typically over 5-6 in-person sessions, can be time-consuming and expensive, creating multiple obstacles to care. Though CBT-I remains the gold standard intervention, adherence can be challenging due to a variety of patient and treatment-related factors as well as multiple changes to bedtime required by the CBT-I component called sleep restriction [25,26]. Even with shorter treatment plans of 3-4 sessions, dropout rates in trials have reached 20%-30%, suggesting potential challenges in real-world adherence [27,28].

In response to this, internet-based or app-delivered digital automated therapies have been developed and proven easier to access by patients, though many of these programs are more instructional than interactive, which may be less engaging for participants [29-31]. Such platforms can make therapy more flexible and accessible, particularly for individuals without access to in-person therapy or who prefer to engage with CBT-I at home. Several meta-analyses and reviews indicate that electronic delivery of CBT-I is just as effective as traditional, face-to-face approaches [29,30]. These programs are grounded in evidence-based practices and typically incorporate components like those found in traditional CBT-I [32,33]. However, there are concerns that the self-directed format of these programs might lead to lower treatment adherence [34]. Some critics also point to a lack of comprehensive quality control over these digital therapeutics as a potential issue [34,35]. Furthermore, these internet-delivered therapies require user engagement with digital screens right at bedtime, potentially confounding the physiologic process of melatonin secretion that enhances sleep [36].

Building on prior work [37], we hypothesized that we could address these issues—improving access to CBT-I, greater technology usability, and reducing screen-time use—at once by integrating 3 distinct technologies running custom software and that YACS would find this technology appealing to use. First, we aimed to deliver the most important therapeutic components of CBT-I interactively via an AI-assisted smart speaker system and without the use of mentally and emotionally activating, melatonin-suppressing screens at night. Second, we deployed smart lighting controls to facilitate CBT-I suggested changes to the room environment. Finally, we used a customized router that provided feedback to each participant about post-bedtime screen use. We supported the ability of YACS to make different choices about screen use at night. We believe these 3 technologies could work together in a customized package to meet the best elements of conversation CBT-I with new ways to support adherence through complex and lengthy treatment protocols. Thus, using an approach involving active stakeholder input and an iterative design process, we aimed to develop and test a novel prototype to increase sleep duration for YACS.

Methods

Participants and Recruitment

Recruitment occurred via patient referrals from 11 oncologists at a large urban academic medical center in Washington, DC,

XSL•FO
RenderX

from 2 regional oncology practices focused on YACS populations, 2 regional cancer survivor support groups, and Facebook social media advertising. The Washington, DC-based academic medical center annually provides oncology care to over 2000 adult patients living with cancer, including approximately 360 YACS. Regional young adult oncology practices included a large outpatient clinic in Northern Virginia and a similarly large clinic in Charlotte, NC. Regional cancer survivor support groups were based in Washington, DC, and suburban Maryland in the Baltimore-Washington, DC region. In all settings, approved recruitment flyers were placed strategically in clinic waiting areas or were shared directly with potential participants by providers (oncology clinics) or support staff (cancer survivor support groups). Members of the study team were available by phone or email to answer questions. Potential participants were screened for eligibility by the research team.

Participants were included if between 18 and 39 years old, living with a history of cancer (any type, any stage), self-reported to average less sleep than recommended for their age by the National Sleep Foundation for ≥3 months (7-9 hours per night) [1], could engage with the study prototype and complete surveys and interviews in English, and had Wi-Fi access at home. Individuals were excluded if they were enrolled in hospice care or lived outside the continental United States.

Focus Groups

Our research team followed the consolidated criteria for reporting qualitative research (COREQ) process through this qualitative study [38]. Qualitative data was gathered through 4 formative focus groups led by study authors (HG and DG), held virtually between December 2022 and March 2023. During recruitment, potential participants were informed that the team sought to learn more about the experience of chronic insomnia and gather perspectives on the use of mobile technology to improve sleep at home. Prior to the study, no study team member had any relationship with any participant. We included 26 participants (6 in focus group 1; 7 in group 2; 7 in group 3; and 6 in group 4). Prior to joining the virtual meeting, participants were asked to turn off their video cameras and deidentify their names (eg, “Guest” or “Participant”) making participation anonymous. Focus group questions and prompts—outlined in Table 1—were developed by members of the research team based on prior work with cancer survivors and tested with 2 YACS who were not part of the focus groups. Each focus group consisted of a brief introduction to the project followed by a facilitated conversation regarding participant experiences of insomnia and sleep disruption in cancer survivorship. Our inquiries explored participant insomnia experiences, elucidating perceived causes and symptoms. Participants were also asked about the current use of technology at home, particularly smart technology such as smart speakers and other voice-activated virtual assistants, screen use habits, and concerns about privacy related to smart technology use. Each focus group lasted approximately 90 minutes. Members of the research team continually reviewed focus group content and, after the fourth focus group, determined that content saturation had been achieved.

XSL•FO
RenderX

Table 1. Focus group questions aimed at understanding chronic insomnia of young adult cancer survivors and portable technology habits.

Theme	Questions
Perceptions and experiences of insomnia	• What is your insomnia experience like? How long has it been occurring? • How has insomnia affected your cancer journey?
Mitigation strategies for insomnia	• What steps have you taken to improve your sleep? What have you found to be helpful, if anything?
Device use and habits	• Which connected devices do you use every day?
Associated between device use and sleep habits	• To what extent do you think that the use of connected devices affects your sleep? • Have you taken any efforts to try and change your connected device use, specifically as it relates to how you sleep? • If yes, what have you found to be helpful? • If not, what would interest you to make changes?
VAVA[a] technology use and habits	• Do you use VAVAs at home or elsewhere? Examples include Apple Siri, Amazon Echo/Dot, or Google Assistant. • If you have not used VAVAs, why not?
Comfort using VAVA-delivered strategies to improve insomnia	• What are your thoughts on a smart speaker system giving you instructions like this about your sleep? • What other ways do you think that technology might be able to help make better choices to improve your sleep?
Familiarity with smart lighting and integration into sleep-wake experience	• What is your familiarity with smart lighting? • What are your thoughts on having smart lights installed in your bedroom that would automatically change in the evening to set up an optimal bedtime for you? • What if the smart lights also automatically changed prior to an optimal wake time?
Screen time usage and device prototype feedback about habits	• What are your thoughts on seeing weekly or monthly reports on how often your devices impact your sleep, and which devices have the most significant effect on your sleep? • Would you have privacy concerns about a system that provides feedback about your internet usage to you? • If yes, how do these privacy concerns compare to your privacy concerns about other apps you use and websites you visit? • Would those privacy concerns change your interest in using a system like this for better sleep? • What system would make you feel confident about privacy protection on a system like this?

[a]VAVA: voice-activated virtual assistant.

Device Prototype Development

The More SHEEP (More Sleep Hours Electronic Education Program) system includes proprietary Media Rez technology for delivering elements of CBT-I via natural voice interactions for more engaging patient education and patients can complete sleep logs without bedtime screen use (Figure 1). The team used the first 2 focus groups with YACS to design the systems and 2 follow-up focus groups to iteratively revise the software. The system was built around an AI-driven smart speaker using an Amazon Alexa device (Echo Dot 3rd Gen). In preparation for the 2-week trial, the technical team reduced the 6-session course of CBT-I to only 2 weeks by concentrating material from later weeks into the earlier sessions, with the goal of providing a comprehensive experience of virtual CBT-I sessions in only 2 weeks. The team also provided a lamp with a color-changing Philips smart bulb controlled by our custom software, which changed its on and off times based on each patient's unique sleep restriction and sleep consolidation recommendations according to the rules of CBT-I. As a result, the smart lamp bulbs automatically changed color temperature throughout the day. It began with a bright white with blue wavelengths in the morning to suppress melatonin production and reduce sleepiness and shifted to a warm amber in the evenings, so melatonin is not suppressed, and the patient can feel naturally sleepy [39,40]. The technical team also developed router control software to record and categorize the web-based traffic (social media, streaming video, and gaming) by the participant after bedtime (but not other members of the household, by having only the participant use the WiFi from the included custom router), so this screen use information could be presented to each patient to help them make better decisions about evening device use. For example, the system compares IP addresses visited by the patient with known sites like Netflix, Call of Duty, and Instagram. The system was also built to intelligently filter out other sources of internet use, like automatic updates and other forms of device telemetry. In this system, patient privacy and confidentiality are ensured since the system immediately

XSL•FO
RenderX

discards all identifiable information about post-bedtime website usage once that usage is shared with the patient. No information is collected about websites visited at any other time. Because this IP information is not stored, it cannot be summarized to participants or accessed at a future time, either by the participant or by a third party such as an internet hacker. No information is collected from other people in the household who are using conventional WiFi and not using this system. Together, these 3 elements combine to deliver a comprehensive, innovative sleep solution.

Figure 1. The More SHEEP (More Sleep Hours Electronic Education Program) system connects a smart speaker, smart lighting, and unique internet router to provide tailored cognitive behavioral therapy for insomnia (CBT-I) and individualized screen use data to individuals with chronic insomnia.

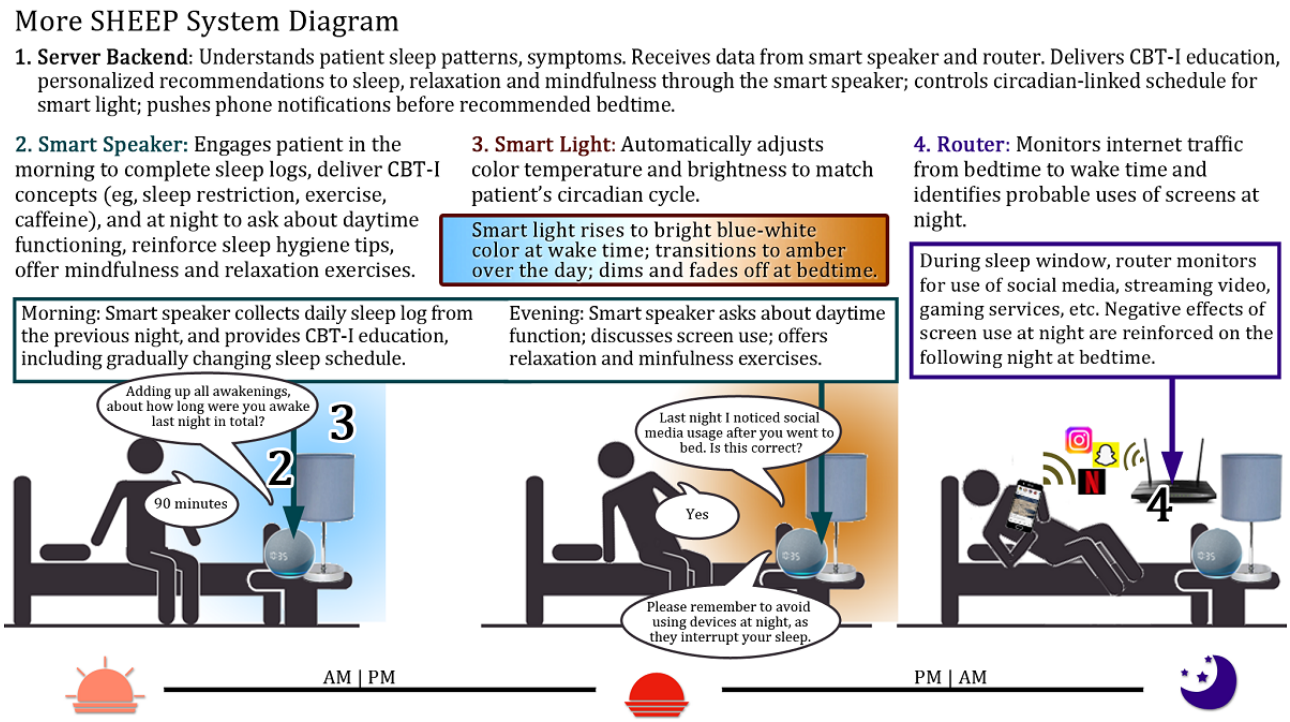

Device Testing

Following prototype development and refinement, we conducted a single-arm study to evaluate the feasibility and acceptability of the system. While the research team remained knowledgeable about each participant, the technology company Media Rez was blinded to participants to preserve anonymity. For this phase, we screened 27 potential participants for participation, 6 of whom expressed interest and then did not respond to further communication, 5 changed their minds about participation after hearing more information about the study, 1 person wished to proceed with consent but was about to leave the country for an extended travel period, 1 person decided not to participate due to an upcoming medical procedure, and 1 person was initially eligible and then became ineligible due to a hospice referral. We thus recruited and enrolled 15 eligible YACS who were not involved in the focus groups.

After completing written consent, participants received (either by mail or picked up in person) a study materials box containing the prototype smart speaker, smart lamp, router, and setup instructions. A member of the study team assisted with device setup via a brief phone call with each participant. Participants were instructed to engage with the prototype system twice daily for 14 days, including entering sleep logs, receiving smart speaker–delivered education about relevant content (eg, sleep hygiene, relaxation strategies), and reviewing personalized data regarding screen time use. If challenges occurred during system setup or use during the 14-day study period, a research team member contacted the tech team directly for troubleshooting advice to preserve the anonymity of the participants to the tech team.

Study feasibility was determined by user engagement with the prototype as assessed by: (1) the number of days over the study period that participants interacted with the system (≥70% interaction days used as cutoff) and (2) the average number of daily interactions with the prototype (≥2 daily interactions used as cutoff). Study acceptability was assessed by an average score ≥5 ("somewhat agree") on the Satisfaction subscale of the Usability, Satisfaction, and Ease of Use scale; this subscale is composed of 7 items measuring satisfaction with program content and the method of program delivery [41]. Additionally, all participants completed a semi-structured interview to assess overall user experience, with a focus on overall impressions of the system as well as each specific component, enthusiasm for longer-term use, and recommendations for device improvements.

We used a grounded theory approach for qualitative data analysis. Focus groups (formative phase) and semi-structured interviews (following device testing) were audio recorded and transcribed. Coding was led by HG (trained in qualitative analysis) with assistance from other team members (LA, LG) and reviewed by members of the research team. Initial coding of qualitative data was descriptive in nature, based on the interview guide structure. From the focus group data, emerging themes and exemplary quotes were extracted to develop the prototype prior to testing. Semi-structured interview data will

XSL•FO
RenderX

be used to further develop the device prior to future efficacy testing. All quantitative data (following device testing) were summarized using descriptive statistics.

Ethical Considerations

Per institutional guidelines at the primary study site, this study was first reviewed and approved by the Georgetown University Lombardi Comprehensive Cancer Center Scientific Review Committee prior to review and approval by the MedStar Health Research Institute Institutional Review Board (number 00005548). Focus group participants provided verbal consent to participate at the beginning of each focus group conducted. Participants for device feasibility testing provided written consent prior to the device prototype being shipped to the home. Per protocol approved by the IRB, all identifiable participant data was stored on password-protected, encrypted servers. Focus group participants each received a $30 gift card. Device testing participants were offered a choice to either (1) keep the study device after a factory reset or (2) receive a US $50 gift card.

Results

Focus Groups

In total, 26 YACS experiencing insomnia participated in focus groups to share experiences of insomnia during cancer survivorship and to assess the practicality of using this system for delivering components of CBT-I (Table 1). All 4 focus groups echoed similar experiences with insomnia, describing cancer survivorship as characterized by insufficient sleep and diminished energy, whether struggling to initiate sleep or stay asleep, and often finding themselves awake in the middle of the night. Common threads in perceived triggers and symptoms encompassed anxiety and worry about disease management and progression, disease-related pain and other symptoms, choices regarding personal device use, and worry about the impact of poor sleep on daily functioning (Table 2).

XSL•FO
RenderX

Table 2. Sample scripts for MORE Sheep modules

Module name	CBT-I component
Education	
• *Another thing I will do is automatically adjust the provided smart lamp. Tonight, beginning an hour before your bedtime of 10:00 P.M., the lamp will begin to fade out, and turn itself off just before bed. In the morning, it will turn on at its lowest brightness and slowly get brighter until it is fully bright at your wake-up time of 7:00 A.M. As I learn more about your sleep patterns, I may suggest changing when you sleep, and the lamp will act as a visual cue of your sleep schedule. In the coming days, I'll explain the effects of light, your circadian rhythm, and how your body uses a hormone called melatonin to regulate sleep.*	Psychoeducation
• *Congratulations! You left the bed four times this week when you couldn't sleep and listened to three relaxation scripts until you were tired enough to go back to the bedroom to sleep. This is good, because it helps you associate the bed only with sleep, and not with lying awake. In this case, sleep restriction could be helpful, making your total time in bed closer to your actual sleep time. Instead of a 10:00 pm bedtime, would you like to try going to bed at 10:45 for the next few nights?* • If Yes: *Good. I'll reset your alarms and lights for 10:45, and we'll see how that works out.*	Psychoeducation and stimulus control
Morning	
• *We worked together to establish a bedtime of 10:45 and wake time of 7 A.M. The hours I set for your sleep are not a lifelong sentence. It was merely the necessary step to get you to try to improve your sleep. As you may be starting to see, sleep restriction and stimulus control are meant to help consolidate your sleep to improve your ability to fall asleep and stay asleep.*	Sleep restriction
• *Rate the quality of your sleep last night from 1-5 (with one being the best and five being the worst). About how many times did you wake up last night?*	Sleep hygiene
Night	
• *Being active each day is important. Even a 10 minute walk is a great start! Would you like me to remind you to get moving during the day?* • If Yes: *Great. What time can I send you a push notification on the app to be active?* • User Time Input: *Ok, I'll send you a notification to be active each day at 1:30 P.M.*	Psychoeducation
• *Please rate your tiredness today on a scale of 1 to 5, with 1 being the least tired and 5 being the most tired.*	Sleep Hygiene
• *Last night I detected internet usage during your sleep window. Specifically, it appears you used devices late at night. This activity appears to include streaming video. Is this correct?* • If Yes: *Remember that this will hurt your sleep. Let me give you a tip for tonight: Establish a bedtime routine where you give yourself time to check devices one more time and then commit to keeping them off until morning. It should ideally be 30 minutes or more before bed. Give yourself the time before bed to view social media or watch a show, but promise yourself that once it is bedtime, your devices are off limits. Consider turning on a "do not disturb" mode, or fully power off these devices, and keep them away from the bed.*	Psychoeducation, Sleep Hygiene
Relaxation:	
• *For this first relaxation exercise, I'd like to play a few samples of each type. In the future, I'll help you find which type you enjoy most.*	Relaxation

> *I think most times I get [insomnia] it was a little bit combination of stress and having emotional traumas that led to sleeplessness.*
>
> *What causes me to stay awake is when I am having some level of depression and anxiety.*
>
> *At the end of the day, [if] you sleep one or two hours and wake up, the experience is bad. A lot of people are looking for how you can get yourself calm. There is a lot of anxiety and depression [living with cancer].*

Use of personal technologies was frequently cited as a trigger for insomnia:

> *Sometimes these [screen] devices become a problem for me in terms of sleeping. I put them in Airplane mode so that I can get better sleep… [but if] I forget to put it on Airplane mode and if my phone disturbs me, I find it difficult to fall asleep again.*
>
> *In order not to get distracted, sometimes [I find myself] setting a limited time to spend with my device and try to adhere to it. And when it is time [to stop phone use] I know it's time to quit time to go to bed, though it is really not easy.*

Common strategies to mitigate sleep disturbances included web- or app-based meditation programs to calm anxiety, worry, or distraction through screen devices. Participants frequently noted a reluctance to seek pharmacotherapies for insomnia, voicing a preference for nonpharmacologic interventions.

XSL•FO
RenderX

> *I try to use [apps to] wind down stuff on my phone. I try to make sure that [I] turn off any screen usage and make sure that I am on my bed getting ready to sleep.*
>
> *I think about a smart speaker for therapy, it is something very unique. It [makes] you feel like [you are in] a relationship, like talking to someone you know. So, for me personally I think it is going to help positively to help [instead of] the stress of going to see therapist from day to day, time to time, consider appointments, traffic, and out of pockets, financially.*
>
> *You know when you have such a device around you, it can literally put you back to sleep. You do not need pills.*

Participants were familiar with smart speakers or voice-activated devices. All participants reported using smartphone-based voice functions, like Apple Siri, for straightforward tasks (eg, creating a calendar reminder) or information gathering (eg, local weather). A minority used in-home smart speakers, primarily to play music or deliver news. All participants expressed interest in the concept of AI-delivered CBT-I and agreed that a smart speaker delivery system could promote frequent engagement with this therapy in the convenience of home. One participant was familiar with smart lighting but did not use it regularly. The remaining participants expressed interest in incorporating smart lighting into a home-based insomnia program.

> *You know when you are in a room where the lighting systems feels comfortable for sleep, you realize you can sleep more better and easier than compared to when we are in a room whereby you get you know more blue light. So that tends to be a bit of a problem to sleep and also affects the entire sleeping pattern.*
>
> *My thought here is that having a smart lighting in the bedroom is what will improve our sleep, our sleeping pattern. Because a lot of people have some type of insomnia because of depression, anxiety, or one thing or the other. But I believe with the installation of smart lighting, the color changes, and the variation of the lighting system, patients might be forced to sleep at a particular time.*

Regarding the additional internet router function to monitor and provide feedback about personal device use, all agreed such data could promote making different choices about screen time use. A minority of participants expressed some concern about privacy with router monitoring of internet usage but also expressed acceptance with additional information about prototype privacy protection.

> *Anything that has to do with privacy, you have to be very careful because you don't want [personal information] to go out there and fall into the wrong hands.*
>
> *At least every device connected to a router must have its own IP address. So, I don't know because each of the IP addresses must have an IP address which can be used to access each of the devices that is connected to a router. I don't know how [you are] going to secure that in terms of privacy of information or data.*

No participants expressed worry about privacy due to the smart speaker itself. A few participants appreciated the possibility of being more open with AI technology than with an actual in-person therapist.

> *It is like meeting a therapist, it's more like talking to a human, it actually saves us the stress of having to access any health care and I'm talking about having discussion about insomnia or sleep disorder. And it is safe for me having it around me in my [home] environment…goes a long way to have a kind of communication level that will ease the tension of having to say whatever you are feeling.*
>
> *The most interesting part of it to me is that the therapist's [voice] in the device can make you feel comfortable when you are trying to make sure you get good sleep.*

Device Testing

In total, 15 participants had an average age of 30.6 (SD 6.8); 10 identified as women; 8 identified as Caucasian, 7 identified as Black or African American, and 1 identified as Hispanic. Of 15 enrolled participants, 2 became lost to follow-up immediately after enrolling (sent the device prototype but never initiated the program), and 1 completed 1 day of participation and then dropped out from concerns about using home Wi-Fi for the project as they were also using high-security home Wi-Fi features for work. The remaining 12 participants completed the study protocol, including a survey and a semi-structured interview.

Participants were able to adhere to the More SHEEP program directions, engaging with the prototype 94% of the assigned times (twice daily for 14 days) and meeting the predetermined feasibility cutoff of engagement ≥70% of assigned times. Participants also found the More SHEEP prototype acceptable, with an overall mean score of 5.43 on the Satisfaction subscale of the Usability, Satisfaction, and Ease of Use scale (range 4.42-7), exceeding the predetermined cutoff score for acceptability of 5.0. All participants completing the program reported they would be interested in using the prototype again, would recommend it to someone else with insomnia, and would be interested in participating in future studies evaluating new iterations of the prototype system.

Following participation, interview themes informed individual experiences with each of the More SHEEP system components (smart speaker, smart lighting, router, and More SHEEP smartphone app) (Table 3).

XSL•FO
RenderX

Table 3. The authors enrolled 15 young adult cancer survivors to participate in a single-arm feasibility-acceptability study of the More SHEEP prototype that included qualitative data collection to inform future prototype development. Participant-generated themes inform prototype feasibility, acceptability, and areas for potential improvement.

Prompt	Content
Familiarity with smart device use	• *I have another [smart speaker] device and my grandfather has one too because he's blind. This one definitely had a more pleasant voice (Participant #1)*
Feasibility of use	• *The app reminded me twice a day on my phone and it was easy to spend a few minutes with the program before bed or after getting up in the morning. No problem at all (#3)* • *For the most part I found [the smart speaker] easy other than when I asked for something that it didn't know how to give me. It understood what I was talking about and was easy to use (#12)*
Acceptability of use	• *I thought it was all pretty easy to use and get used to. I even tried taking it on a visit to family and it worked there for me too. (#6)* • *I do the check-ins each morning and evening. I think I remember to do it almost every day, so it's pretty easy to do, particularly with it being just audio, rather than having to go to a website and read [material]*
Challenges with smart lamp	• *I ended up not using the lamp at all because my partner goes to bed at different times so it was disruptive. The rest was easy to use (#4)* • *I did sometimes modify my answers to make the smart lamp go off a little earlier…I would find myself throwing a blanket over the top (#1)*
Relevance of CBT-I[a] language and setting	• *Occasionally, the [software] voice mentioned something related to cancer survivorship and I really felt that this program was designed for someone like me (#2)* • *I think the awareness [of my sleep choices] was one thing I wasn't expecting to just pick up on…being more aware of how I felt [around bedtime] (#1)* • *I feel like it had surprisingly little survivorship information. For the most part, I felt like I was learning general sleep habits (#12)* • *I did think that this [home system] is very convenient as somebody who has three small kids and cancer and a husband so I don't have to go anywhere and I feel like I'm getting something from these sessions (#11)*
Preference for more customization	• *It would be really cool if there was a way to make the lamp different colors maybe depending on the education happening or on my mood. Maybe combine the speaker and the lamp into one, that would be good too (#7)* • *I like being able to use sounds at night to sleep. I put the thunderstorm [track] on, or the ocean, and I'm able to sleep (#14)* • *It would definitely be nice if [the system] could say, if you could ask how much time you have or something like that*
Need to design for varying spaces or needs	• *I'm in college and have roommates so I wasn't sure if I should try this at school or at home. I'm participating from home but if I used this at school it would have to be different (#10)* • *I have a 4-year-old and she can't sleep by herself so she's coming into my room; then I think, can I activate Alexa without waking her up?*

[a]CBT-I: cognitive behavioral therapy for insomnia.

Discussion

Principal Findings

We discovered a keen interest among participants in this technology in this preliminary investigation into the experiences of YACS dealing with insomnia and smart home devices, particularly when tailored for user-friendly personalization. Most participants expressed significant interest in using this technology. Participants generally reported an enhanced understanding of CBT-I and its potential impact on mitigating chronic sleep disruption and rated the program as both feasible and acceptable. Participants highlighted the high feasibility of daily sleep logs during our demonstration, a crucial component of successful engagement with CBT-I. This observation implies that the data gathered by the smart home device can be leveraged by AI programming to generate personalized recommendations and schedules, surpassing the scope of conventional sleep hygiene education.

Feedback from participants also informed areas for prototype improvement. Some participants felt the More SHEEP content could speak more to the YACS experience of insomnia and could include more information and recommendations directed specifically to cancer survivorship; in collaboration with young adult cancer survivorship experts, future device iterations could include more language targeted to YACS. Participants often noted that a smart speaker or smart lighting may not be conducive to sharing sleep space. For example, the smart light might remain on to address sleep restrictions per CBT-I protocol for a participant while a bed partner wants to sleep without lighting. Similarly, a participant might not want to awaken others using a smart speaker at bedtime—this was particularly important to 1 participant living in a university dormitory. Finally, some participants expressed preferences for more sleep-related content, like sleep music or white noise, allowing

XSL•FO
RenderX

more customizability. Formative testing processes of future prototype iterations may address these concerns, ultimately enhancing participant engagement and treatment adherence.

Developments in home technology play an increasing role in improving insomnia and related symptoms by providing tools and solutions to address sleep disturbances. Examples include but are not limited to mobile sleep-tracking devices, light therapy, smart thermostats, biometric feedback devices, and relaxation or meditation apps [42-46]. As a gold standard for the management of chronic insomnia, implementation of CBT-I continues to be explored in this space, primarily through mobile device applications [47-51]. Nevertheless, even through the use of mobile technologies, maximizing treatment adherence remains challenging. In total, 1 study specifically focused on YACS, delivering an internet-based CBT-I intervention over 6-9 weeks and finding that, while outcomes of insomnia, daytime sleepiness, fatigue, and quality of life improved at the study's 2 timepoints, most participants did not complete the assigned protocol [52]. Compared to this and other device-driven insomnia mitigation strategies, our prototype is unique in its CBT-I delivery through a smart speaker, limiting device screen exposure whose blue wavelength light can impede nighttime melatonin secretion, as well as smart lighting use that can provide important diurnal visual cues (and block blue wavelength light at night). While most of our study participants had never used these technologies before, acceptability was high, and participants were able to easily engage with the program structure. While we believe that our prototype's inherent ease of use through voice activation in the bedroom will improve long-term CBT-I treatment adherence, we did not design this early study to evaluate this.

Several limitations of this project deserve mention. First, our insights come from a small number of participants, primarily due to the constraints of the study's limited timeframe. Nevertheless, this sample size appropriately aligns with stage I (intervention refinement) of the NIH stage model for behavioral intervention development. Additionally, as we reached the conclusion of data collection, the absence of emerging additional themes in our qualitative data suggested a saturation of feedback. Second, we note selection bias, as some eligible YACS learned of the study but chose not to participate. We also were not able to complete data collection for 2 enrolled participants. Furthermore, although some information was revealed during post-participation interviews, we did not collect data regarding other factors that might influence the choice to enroll or study adherence, such as income, living situation, or education that might influence outcomes. Given the nature of the intervention, it is possible these same factors could also influence clinical outcomes related to insomnia—thus, future efficacy studies of our prototype should, at a minimum, collect such patient-level data and potentially ensure diversity of participants through purposive recruiting strategies, increasing generalizability of results. By design, this study intentionally did not evaluate clinical outcomes such as improvement in insomnia, mood, daytime fatigue, etc, focusing specifically on device feasibility of use and acceptability over 14 days. While this time frame facilitated participant experience with the More SHEEP prototype, a full CBT-I treatment typically requires approximately 6 weeks; therefore, future prototype refinement may benefit from piloting a longer period of use at home. Nonetheless, this early development phase focused on evaluating device feasibility and acceptability and we believe we have achieved our primary goal. Subsequent studies will delve deeper, refining and testing the prototype for practical feasibility and effectiveness.

Conclusions

This study combined end-user participant input toward the development of a novel smart speaker, smart lighting, or a router system to deliver key elements of CBT-I for YACS to be used in the home setting. Now that the system has demonstrated feasibility and met acceptability benchmarks, the prototype can continue to be developed to incorporate suggestions from our participants and subsequently be tested for efficacy in the YACS population. Ultimately, this system has the potential to increase much-needed access to CBT-I among cancer survivors.

Acknowledgments

The authors thank the participants for generously sharing their time, interest, and input. No generative artificial intelligence was used in the preparation of this manuscript.

Data Availability

The datasets generated during or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

Media Rez is a for-profit company that seeks to commercialize an eventual product based on this technology, consistent with the National Institutes of Health Small Business Innovation Research program.

References

1. The National Sleep Foundation. URL: https://www.thensf.org [accessed 2024-11-14]
2. Armenian SH, Chao C. Burden of morbidity and mortality in adolescent and young adult cancer survivors. J Clin Oncol 2024;42(6):735-742. [doi: 10.1200/jco.23.01751]

XSL•FO
RenderX

3. Daniel LC, Aggarwal R, Schwartz LA. Sleep in adolescents and young adults in the year after cancer treatment. J Adolesc Young Adult Oncol 2017;6(4):560-567. [doi: 10.1089/jayao.2017.0006] [Medline: 28628351]
4. Zhou ES, Recklitis CJ. Insomnia in adult survivors of childhood cancer: a report from project REACH. Support Care Cancer 2014;22(11):3061-3069. [doi: 10.1007/s00520-014-2316-y] [Medline: 24935649]
5. Hinds PS, Hockenberry MJ, Gattuso JS, Srivastava DK, Tong X, Jones H, et al. Dexamethasone alters sleep and fatigue in pediatric patients with acute lymphoblastic leukemia. Cancer 2007;110(10):2321-2330 [FREE Full text] [doi: 10.1002/cncr.23039] [Medline: 17926333]
6. Zupanec S, Jones H, Stremler R. Sleep habits and fatigue of children receiving maintenance chemotherapy for ALL and their parents. J Pediatr Oncol Nurs 2010;27(4):217-228. [doi: 10.1177/1043454209358890] [Medline: 20562390]
7. Daniel LC, Wang M, Mulrooney DA, Srivastava DK, Schwartz LA, Edelstein K, et al. Sleep, emotional distress, and physical health in survivors of childhood cancer: a report from the childhood cancer survivor study. Psychooncology 2019;28(4):903-912 [FREE Full text] [doi: 10.1002/pon.5040] [Medline: 30817058]
8. Ameringer S, Erickson JM, Macpherson CF, Stegenga K, Linder LA. Symptoms and symptom clusters identified by adolescents and young adults with cancer using a symptom heuristics app. Res Nurs Health 2015;38(6):436-448. [doi: 10.1002/nur.21697] [Medline: 26455729]
9. Olson K. Sleep-related disturbances among adolescents with cancer: a systematic review. Sleep Med 2014;15(5):496-501. [doi: 10.1016/j.sleep.2014.01.006] [Medline: 24726572]
10. Mulrooney DA, Ness KK, Neglia JP, Whitton JA, Green DM, Zeltzer LK, et al. Fatigue and sleep disturbance in adult survivors of childhood cancer: a report from the childhood cancer survivor study (CCSS). Sleep 2008;31(2):271-281 [FREE Full text] [doi: 10.1093/sleep/31.2.271] [Medline: 18274275]
11. Clanton NR, Klosky JL, Li C, Jain N, Srivastava DK, Mulrooney D, et al. Fatigue, vitality, sleep, and neurocognitive functioning in adult survivors of childhood cancer: a report from the childhood cancer survivor study. Cancer 2011;117(11):2559-2568 [FREE Full text] [doi: 10.1002/cncr.25797] [Medline: 21484777]
12. Hsu CY, Chen YT, Chen MH, Huang CC, Chiang CH, Huang PH, et al. The association between insomnia and increased future cardiovascular events: a nationwide population-based study. Psychosom Med 2015;77(7):743-751. [doi: 10.1097/PSY.0000000000000199] [Medline: 26355726]
13. Hall MH, Smagula SF, Boudreau RM, Ayonayon HN, Goldman SE, Harris TB, et al. Association between sleep duration and mortality is mediated by markers of inflammation and health in older adults: the health, aging and body composition study. Sleep 2015;38(2):189-195 [FREE Full text] [doi: 10.5665/sleep.4394] [Medline: 25348127]
14. Pigeon WR, Bishop TM, Krueger KM. Insomnia as a precipitating factor in new onset mental illness: a systematic review of recent findings. Curr Psychiatry Rep 2017;19(8):44. [doi: 10.1007/s11920-017-0802-x] [Medline: 28616860]
15. Zhou ES, Manley PE, Marcus KJ, Recklitis CJ. Medical and psychosocial correlates of insomnia symptoms in adult survivors of pediatric brain tumors. J Pediatr Psychol 2016;41(6):623-630. [doi: 10.1093/jpepsy/jsv071] [Medline: 26275977]
16. Ancoli-Israel S. Recognition and treatment of sleep disturbances in cancer. J Clin Oncol 2009;27(35):5864-5866. [doi: 10.1200/jco.2009.24.5993]
17. Khater W, Masha'al D, Al-Sayaheen A. Sleep assessment and interventions for patients living with cancer from the patients' and nurses' perspective. Int J Palliat Nurs 2019;25(7):316-324. [doi: 10.12968/ijpn.2019.25.7.316] [Medline: 31339821]
18. Management of Chronic Insomnia Disorder in Adults. A clinical practice guideline from the American College of Physicians. Ann Intern Med 2016;165(2):33-125. [doi: 10.7326/p16-9016]
19. Sanft T, Denlinger CS, Armenian S, Baker KS, Broderick G, Demark-Wahnefried W, et al. NCCN guidelines insights: survivorship, version 2.2019. J Natl Compr Canc Netw 2019;17(7):784-794 [FREE Full text] [doi: 10.6004/jnccn.2019.0034] [Medline: 31319383]
20. Blumenstein KG, Brose A, Kemp C, Meister D, Walling E, DuVall AS, et al. Effectiveness of cognitive behavioral therapy in improving functional health in cancer survivors: a systematic review and meta-analysis. Crit Rev Oncol Hematol 2022;175:103709. [doi: 10.1016/j.critrevonc.2022.103709] [Medline: 35580765]
21. Reynolds SA, Ebben MR. The cost of insomnia and the benefit of increased access to evidence-based treatment: cognitive behavioral therapy for insomnia. Sleep Med Clin 2017;12(1):39-46. [doi: 10.1016/j.jsmc.2016.10.011] [Medline: 28159096]
22. Koffel E, Bramoweth AD, Ulmer CS. Increasing access to and utilization of cognitive behavioral therapy for insomnia (CBT-I): a narrative review. J Gen Intern Med 2018;33(6):955-962 [FREE Full text] [doi: 10.1007/s11606-018-4390-1] [Medline: 29619651]
23. Thomas A, Grandner M, Nowakowski S, Nesom G, Corbitt C, Perlis ML. Where are the behavioral sleep medicine providers and where are they needed? A geographic assessment. Behav Sleep Med 2016;14(6):687-698 [FREE Full text] [doi: 10.1080/15402002.2016.1173551] [Medline: 27159249]
24. Vargas I, Egeler M, Walker J, Benitez DD. Examining the barriers and recommendations for integrating more equitable insomnia treatment options in primary care. Front Sleep 2023;2. [doi: 10.3389/frsle.2023.1279903] [Medline: 39210962]
25. Matthews EE, Arnedt JT, McCarthy MS, Cuddihy LJ, Aloia MS. Adherence to cognitive behavioral therapy for insomnia: a systematic review. Sleep Med Rev 2013;17(6):453-464 [FREE Full text] [doi: 10.1016/j.smrv.2013.01.001] [Medline: 23602124]

XSL•FO
RenderX

26. Mellor A, Kavaliotis E, Mascaro L, Drummond SPA. Approaches to the assessment of adherence to CBT-I, predictors of adherence, and the association of adherence to outcomes: a systematic review. Sleep Med Rev 2022;63:101620. [doi: 10.1016/j.smrv.2022.101620] [Medline: 35398650]
27. Zhou ES, Partridge AH, Recklitis CJ. A pilot trial of brief group cognitive-behavioral treatment for insomnia in an adult cancer survivorship program. Psychooncology 2017;26(6):843-848. [doi: 10.1002/pon.4096] [Medline: 26872123]
28. Buysse DJ, Germain A, Moul DE, Franzen PL, Brar LK, Fletcher ME, et al. Efficacy of brief behavioral treatment for chronic insomnia in older adults. Arch Intern Med 2011;171(10):887-895. [doi: 10.1001/archinternmed.2010.535] [Medline: 21263078]
29. Zachariae R, Lyby MS, Ritterband LM, O'Toole MS. Efficacy of internet-delivered cognitive-behavioral therapy for insomnia - a systematic review and meta-analysis of randomized controlled trials. Sleep Med Rev 2016;30:1-10. [doi: 10.1016/j.smrv.2015.10.004] [Medline: 26615572]
30. Seyffert M, Lagisetty P, Landgraf J, Chopra V, Pfeiffer PN, Conte ML, et al. Internet-delivered cognitive behavioral therapy to treat insomnia: a systematic review and meta-analysis. PLoS One 2016;11(2):e0149139. [doi: 10.1371/journal.pone.0149139] [Medline: 26867139]
31. Hall DL, Arditte Hall KA, Gorman MJ, Comander A, Goldstein MR, Cunningham TJ, et al. The survivorship sleep program (SSP): a synchronous, virtual cognitive behavioral therapy for insomnia pilot program among cancer survivors. Cancer 2022;128(7):1532-1544 [FREE Full text] [doi: 10.1002/cncr.34066] [Medline: 34914845]
32. Rowan AB, Magnante AT, Urh N, Figueroa L. Cognitive behavioral therapy for insomnia self-management mobile apps: a review of efficacy and quality. J Clin Psychol Med Settings 2024;31(3):537-549. [doi: 10.1007/s10880-023-09992-y]
33. Bai N, Cao J, Zhang H, Liu X, Yin M. Digital cognitive behavioural therapy for patients with insomnia and depression: a systematic review and meta-analysis. J Psychiatr Ment Health Nurs 2024;31(4):654-667. [doi: 10.1111/jpm.13024] [Medline: 38226714]
34. Simon L, Steinmetz L, Feige B, Benz F, Spiegelhalder K, Baumeister H. Comparative efficacy of onsite, digital, and other settings for cognitive behavioral therapy for insomnia: a systematic review and network meta-analysis. Sci Rep 2023;13(1):1929 [FREE Full text] [doi: 10.1038/s41598-023-28853-0] [Medline: 36732610]
35. Thorndike FP, Morin CM, Ojile J, Edington S, Gerwien R, Ong JC, et al. Effect of a prescription digital therapeutic for chronic insomnia on post-treatment insomnia severity, depression, and anxiety symptoms: results from the real-world DREAM study. Front Psychiatry 2024;15:1450615 [FREE Full text] [doi: 10.3389/fpsyt.2024.1450615] [Medline: 39319356]
36. Figueiro MG, Wood B, Plitnick B, Rea MS. The impact of light from computer monitors on melatonin levels in college students. Neuro Endocrinol Lett 2011;32(2):158-163. [Medline: 21552190]
37. Arem H, Scott R, Greenberg D, Kaltman R, Lieberman D, Lewin D. Assessing breast cancer survivors' perceptions of using voice-activated technology to address insomnia: feasibility study featuring focus groups and in-depth interviews. JMIR Cancer 2020;6(1):e15859 [FREE Full text] [doi: 10.2196/15859] [Medline: 32348274]
38. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
39. Lockley SW, Brainard GC, Czeisler CA. High sensitivity of the human circadian melatonin rhythm to resetting by short wavelength light. J Clin Endocrinol Metab 2003;88(9):4502-4505. [doi: 10.1210/jc.2003-030570] [Medline: 12970330]
40. Bourgin P, Hubbard J. Alerting or somnogenic light: pick your color. PLoS Biol 2016;14(8):e2000111 [FREE Full text] [doi: 10.1371/journal.pbio.2000111] [Medline: 27525420]
41. Gao M, Kortum P, Oswald F. Psychometric Evaluation of the USE (Usefulness, Satisfaction, and Ease of use) Questionnaire for Reliability and Validity. 2018 Presented at: Proceedings of the Human Factors and Ergonomics Society Annual Meeting; 2018 September 27; United States. [doi: 10.1177/1541931218621322]
42. Lin LY, Tam KW, Huang TW. Effect of bright light therapy on cancer-related fatigue and related symptoms: a systematic review and meta-analysis of randomized controlled trials. J Psychosom Res 2023;174:111501. [doi: 10.1016/j.jpsychores.2023.111501] [Medline: 37797569]
43. Jalali N, Sahu KS, Oetomo A, Morita PP. Usability of smart home thermostat to evaluate the impact of weekdays and seasons on sleep patterns and indoor stay: observational study. JMIR mHealth uHealth 2022;10(4):e28811 [FREE Full text] [doi: 10.2196/28811] [Medline: 35363147]
44. Nguyen A, Pogoncheff G, Dong BX, Bui N, Truong H, Pham N, et al. A comprehensive study on the efficacy of a wearable sleep aid device featuring closed-loop real-time acoustic stimulation. Sci Rep 2023;13(1):17515 [FREE Full text] [doi: 10.1038/s41598-023-43975-1] [Medline: 37845236]
45. Huberty J, Bhuiyan N, Eckert R, Larkey L, Petrov M, Todd M, et al. Insomnia as an unmet need in patients with chronic hematological cancer: protocol for a randomized controlled trial evaluating a consumer-based meditation app for treatment of sleep disturbance. JMIR Res Protoc 2022;11(7):e39007 [FREE Full text] [doi: 10.2196/39007] [Medline: 35776489]
46. Park S, Lee SW, Han S, Cha M. Clustering insomnia patterns by data from wearable devices: algorithm development and validation study. JMIR mHealth uHealth 2019;7(12):e14473 [FREE Full text] [doi: 10.2196/14473] [Medline: 31804187]
47. Subotic-Kerry M, Werner-Seidler A, Corkish B, Batterham PJ, Sicouri G, Hudson J, et al. Protocol for a randomised controlled trial evaluating the effect of a CBT-I smartphone application (Sleep Ninja) on insomnia symptoms in children. BMC Psychiatry 2023;23(1):684 [FREE Full text] [doi: 10.1186/s12888-023-05185-x] [Medline: 37730577]

XSL•FO
RenderX

48. Jeon JM, Ma J, Kwak P, Dang B, Buleje I, Ancoli-Israel S, et al. Developing a novel mobile application for cognitive behavioral therapy for insomnia for people with schizophrenia: integration of wearable and environmental sleep sensors. Sleep Breath 2024;28(3):1491-1498 [FREE Full text] [doi: 10.1007/s11325-023-02980-4] [Medline: 38177830]
49. Chan CS, Wong CYF, Yu BYM, Hui VKY, Ho FYY, Cuijpers P. Treating depression with a smartphone-delivered self-help cognitive behavioral therapy for insomnia: a parallel-group randomized controlled trial. Psychol Med 2021;53(5):1799-1813. [doi: 10.1017/s0033291721003421]
50. Watanabe Y, Kuroki T, Ichikawa D, Ozone M, Uchimura N, Ueno T. Effect of smartphone-based cognitive behavioral therapy app on insomnia: a randomized, double-blind study. Sleep 2023;46(3):zsac270. [doi: 10.1093/sleep/zsac270] [Medline: 36355920]
51. Luyster FS, Ritterband LM, Sereika SM, Buysse DJ, Wenzel SE, Strollo PJ. Internet-based cognitive-behavioral therapy for insomnia in adults with asthma: a pilot study. Behav Sleep Med 2020;18(1):10-22. [doi: 10.1080/15402002.2018.1518229] [Medline: 30252506]
52. Zhou ES, Recklitis CJ. Internet-delivered insomnia intervention improves sleep and quality of life for adolescent and young adult cancer survivors. Pediatr Blood Cancer 2020;67(9):e28506. [doi: 10.1002/pbc.28506] [Medline: 32568460]

Abbreviations

CBT-I: cognitive behavioral therapy for insomnia
More SHEEP: More Sleep Hours Electronic Education Program
YACS: young adult cancer survivor

Edited by A Mavragani; submitted 29.07.24; peer-reviewed by S Joshi; comments to author 04.10.24; revised version received 22.10.24; accepted 30.10.24; published 10.03.25.

Please cite as:
Groninger H, Arem H, Ayangma L, Gong L, Zhou E, Greenberg D
Development of a Voice-Activated Virtual Assistant to Improve Insomnia Among Young Adult Cancer Survivors: Mixed Methods Feasibility and Acceptability Study
JMIR Form Res 2025;9:e64869
URL: https://formative.jmir.org/2025/1/e64869
doi:10.2196/64869
PMID:

©Hunter Groninger, Hannah Arem, Lylian Ayangma, Lisa Gong, Eric Zhou, Daniel Greenberg. Originally published in JMIR Formative Research (https://formative.jmir.org), 10.03.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Zoom-Delivered Empowered Relief for Chronic Pain: Observational Longitudinal Pilot Study Exploring Feasibility and Pain-Related Outcomes in Patients on Long-Term Opioids

Karlyn A Edwards[1], PhD; Troy C Dildine[2], PhD; Dokyoung S You[2], PhD; Ashley M Herrick[2], BS; Beth D Darnall[2], PhD; Sean C Mackey[2], MD, PhD; Maisa S Ziadni[2], PhD

[1]Division of General Internal Medicine, Department of Medicine, University of Pittsburgh, Pittsburgh, PA, United States

[2]Division of Pain Medicine, Department of Anesthesiology, Perioperative and Pain Medicine, Stanford University, Stanford, CA, United States

Corresponding Author:
Maisa S Ziadni, PhD
Division of Pain Medicine
Department of Anesthesiology, Perioperative and Pain Medicine
Stanford University
1070 Arastradero Rd, Suite 200
Stanford, CA, 94304
United States
Phone: 1 650 497 7671
Email: mziadni@stanford.edu

Abstract

Background: Patients with chronic pain on long-term opioid therapy often face barriers to accessing effective nonpharmacological treatments, including the burden of multiple sessions, lack of trained clinicians, and travel time. Empowered Relief (ER), a 2-hour, single-session pain relief skills class, can improve pain and quality of life among patients with chronic pain when delivered in person or virtually.

Objective: This study examined the impact of Zoom-delivered ER among people with chronic pain on long-term opioid therapy. We assessed (1) the feasibility and acceptability of Zoom-delivered ER; (2) changes in pain and opioid use outcomes at 3 and 6 months after treatment; and (3) daily associations among pain, opioid dose, and the Pain Catastrophizing Scale (PCS) before and after treatment.

Methods: During the early COVID-19 pandemic, we conducted an uncontrolled pilot study of a Zoom-delivered ER among 60 adults (n=45, 76% female participants; n=52, 88% White participants) experiencing chronic pain who were receiving daily prescribed opioids (≥10 morphine-equivalent daily dose). Participants completed assessments at enrollment, before class, after class, 3 months after treatment, and 6 months after treatment. Furthermore, participants completed 2 daily assessment periods (spanning 14 consecutive days) before and after the class. We used a multilevel modeling approach to examine (1) the raw changes in PCS, average pain intensity, pain interference, and self-reported opioid dose at 3 and 6 months after treatment and (2) daily-level changes in average pain intensity and opioid dose before and after the class.

Results: Of the 60 participants enrolled, 41 (68%) attended the class and 24 (59% of the 41 class attendees) reported satisfaction with the Zoom-delivered class. PCS score was significantly reduced at 3 months (β=–3.49, P=.01; Cohen d=0.35) and 6 months after treatment (β=–3.61, P=.01; Cohen d=0.37), and pain intensity was significantly reduced at 3 months (β=–0.56, P=.01; Cohen d=0.39) compared to enrollment. There were no significant reductions in pain interference or opioid dose. Across daily assessments, higher daily pain catastrophizing was associated with worse daily pain (β=.42, P<.001) and higher self-reported opioid use (β=3.14, P<.001); daily pain intensity significantly reduced after the class (β=–.50, P<.001). People taking prescribed opioids as needed trended toward decreasing their daily opioid use after the class (β=–9.31, P=.02), although this result did not survive correction for multiplicity.

Conclusions: Improvements to future Zoom-delivered ER iterations are needed to improve feasibility and acceptability among people with chronic pain and daily prescribed opioid use. Despite this, findings show a promising preliminary impact of the intervention on pain outcomes. A larger randomized controlled trial of Zoom-delivered ER among this patient population is currently under way.

XSL•FO
RenderX

(JMIR Form Res 2025;9:e68292) doi:10.2196/68292

KEYWORDS

Empowered Relief; single session; chronic pain; prescription opioids; telehealth; daily data; pain intensity; pain catastrophizing

Introduction

Background

One in 5 Americans has chronic pain [1], a condition associated with substantial personal and societal costs. Annual expenditures related to chronic pain, including lost productivity, exceed half a trillion dollars [2]. Individuals with chronic pain frequently report impacts on physical, psychological, and social functioning [3]. Over 100,000 Americans receive daily prescribed opioids [4], and while they can be an effective pain treatment for some [5], this treatment approach can also present several health risks [6], including overdose and death [7]. Furthermore, prolonged opioid use can contribute to mental health challenges, such as depression [8]. Thus, the gold standard for chronic pain management is a multimodal approach that integrates medical, behavioral, and rehabilitative treatments to minimize opioid use and maximize patient functioning [9,10].

Behavioral pain treatments are low-risk approaches that may reduce opioid-related risks and mitigate the broader impacts of chronic pain [11]. For example, cognitive behavioral therapy (CBT) alone [12], and when combined with posttreatment therapeutic interactive voice response telephonic support [13], has been shown to reduce opioid misuse. Moreover, acceptance and commitment therapy was shown to aid in opioid cessation [14] in patients flagged as being at risk for problematic opioid use after surgery. In addition, acceptance and commitment therapy was shown to aid in opioid reduction combined with mindfulness-based relapse prevention [15] in patients with prescription opioid misuse. In addition, a technique named Mindfulness-Orientated Recovery Enhancement (MORE) can lower pain interference and severity, as well as opioid misuse among people with comorbid chronic pain and opioid misuse [16]. Few interventions have been tested among patients who do not specifically misuse opioids, and the findings are mixed [13,17]. While studies have demonstrated the effectiveness of behavioral interventions, few patients can access them due to low availability and high treatment time burden (typically requiring 8-16 hours of in-person or online treatment) [18]. Poor insurance coverage and availability are primary barriers [19], in addition to travel time and the length of the interventions [20]. Treatment barriers were exacerbated by the COVID-19 pandemic, which caused further delays that contributed to worsened pain [21], function, and psychological distress [22,23]. Furthermore, treatment burdens limit both the accessibility and scalability of behavioral pain interventions, and in part, opioid prescribing may be the result of fewer available alternatives, particularly for patients in low-resource settings [24].

A single-session behavioral pain relief skills class, Empowered Relief (ER), was developed to address these barriers. ER synthesizes pain neuroscience education, cognitive behavioral skills, and mindfulness-based skills into a 2-hour, single-session pain relief skills class that enhances self-regulation of pain, including pain catastrophizing [25]. Pain catastrophizing, defined as a pattern of thinking that involves rumination, magnification of pain, and helplessness, is a robust predictor of worse pain and opioid use outcomes [26,27], including greater opioid use [28], opioid misuse [29], pain-related disability [30], and negative affect [31]. Previous work has demonstrated the efficacy of ER when delivered in person [32] and in synchronous [33] and asynchronous [34,35] web-based formats. Notably, when delivered via Zoom to patients with mixed chronic pain conditions (*not taking opioids*; N=101), it was superior to a waitlist control in reducing pain catastrophizing, pain intensity, and pain interference at 3 months after treatment [33]. A larger comparative efficacy trial in people with chronic low back pain (N=263) found in-person ER to be noninferior to 8 sessions of CBT for reducing pain catastrophizing, pain intensity, and pain interference (2 hours vs 16 hours of treatment time) at 3 [32] and 6 months [36]. In addition, 2 randomized controlled trials (RCTs) examining postsurgical outcomes showed that a web-based, automated version of ER tailored to surgical patients had opioid-sparing effects [34] and reduced pain [35] up to 3 months after surgery.

Critical gaps remain in disseminating high-quality digital pain treatments and using digital data collection methods, which can reduce barriers to participation in treatment and research studies [37,38]. Few studies have used digital daily assessments among people with chronic pain taking daily prescribed opioids [31,39-43], despite support for their feasibility and acceptability for measuring daily opioid use among patients with chronic pain [44]. Of those few studies, results consistently show heightened daily-level associations among pain catastrophizing, pain, and opioid craving and use [39,41-43]. However, only 2 studies have examined these associations before and after a behavioral pain intervention among people with chronic pain on long-term opioid therapy. Both studies found significant reductions in daily reports of pain, stress, and craving during and 1 month after a mindfulness-based intervention (MORE) [45,46] involving 16 hours of treatment. Further research of briefer behavioral pain interventions and their impact on daily-level pain and opioid use is needed.

Objectives

We aimed to address these gaps by conducting an uncontrolled pilot study to assess the feasibility, acceptability, and impact of Zoom-delivered ER on pain and opioid use outcomes among patients with chronic pain receiving daily prescribed opioids. We hypothesized that at least 80% of enrolled patients would engage in the treatment. In addition, we hypothesized that ≥80% of patients would rate treatment acceptability, satisfaction, usefulness, and ease of understanding at ≥8 out of 10 and rate ease, comfort, and overall satisfaction of using the Zoom platform to be ≥5 out of 7 (higher scores indicating greater satisfaction, comfort, and ease). Second, we examined changes in pain and opioid use outcomes from enrollment to 3 months (primary end point) and 6 months after treatment. We hypothesized that there would be a significant reduction in pain

XSL•FO
RenderX

intensity, pain interference, and pain catastrophizing at 3 and 6 months after treatment compared to enrollment. We examined changes in opioid dose as an exploratory outcome. Finally, we used 2 rounds of digital daily assessments before and after the class to characterize changes in daily associations between pain catastrophizing, pain intensity, and opioid use before and after the ER class. We hypothesized that higher daily pain catastrophizing would be significantly associated with higher pain intensity and self-reported opioid use. We also hypothesized that daily use of behavioral skills would be associated with a reduction in pain intensity and opioid use after class.

Methods

Study Design

The Stanford University School of Medicine coordinated this uncontrolled, prospective trial during the COVID-19 pandemic. We designed this study to assess the feasibility and acceptability of delivering ER via Zoom (Zoom Communications) to inform the transitioning of a larger RCT (ClinicalTrials.gov: NCT03950791) to the Zoom platform [47]. Thus, this study was not powered to detect differences in outcomes; however, findings will be used to determine adequate power and sample size for future studies. Reporting of the current trial is in concordance with the CONSORT (Consolidated Standards of Reporting Trials) 2010 extension for randomized pilot and feasibility trials. Participants with mixed-etiology chronic pain who were taking daily prescribed opioid medication and were between the ages of 18 and 80 years were enrolled. Full data collection, including all posttreatment assessments, was conducted between June 2020 and June 2021. The 2 ER classes occurred between June 2020 and November 2020.

Ethical Considerations

This study protocol was submitted as a protocol modification to the larger clinical trial and was approved by the Stanford University Institutional Review Board (48784). To enroll in the study, electronic informed consent was obtained from all participants in concordance with the institutional review board protocol. Study data were deidentified and anonymized after study completion. Participants were compensated US $135 for completing the baseline, preclass, postclass, 3-month, and 6-month assessments. The class was provided at no charge. Participants who completed at least 10 of the 14 daily surveys were entered into a raffle for an iPad. Participants could be entered up to twice if they met this threshold for both the baseline and follow-up daily surveys.

Recruitment and Enrollment

Participants were recruited remotely through targeted emails to lists of patients who agreed to be contacted for research purposes through the following three sources: (1) Stanford's Learning Health System, comprising patients who have received care at the Stanford Pain Management Center (a referral-based tertiary outpatient pain clinic in Redwood City, California, that offers multidisciplinary pain care); (2) the Stanford Systems Neuroscience and Pain Lab database of community members with chronic pain who expressed interest in participating in research; and (3) study advertisements on the Stanford Pain Division website. Approximately 17% (10/60) of the study sample were clinic patients and 83% (50/60) were from the larger community. Patients who were in 1 of the 2 databases (Stanford's Learning Health System or Stanford Systems Neuroscience and Pain Lab) were sent an email asking their interest in participating in a chronic pain treatment research study, and if interested, were directed to click on a link to a web-based screening form. Community members who saw the study advertisement on web were also directed to click on a link to fill out a web-based screening form. Those who met the initial eligibility criteria were contacted by research staff by phone to confirm their eligibility. Inclusion criteria were (1) individuals aged between 18 and 80 years; (2) individuals with noncancer chronic pain (defined as pain duration ≥6 months); (3) individuals who self-reported current use of prescription opioids ≥10 morphine-equivalent daily dose (MEDD) for ≥3 months; (4) individuals with fluency in the English language; and (5) individuals who were able to attend a one-time web-based class and complete web-based surveys. Exclusion criteria were (1) ongoing legal action or disability claim, (2) active participation in a CBT-based treatment, (3) inability to provide informed consent and complete study procedures, and (4) active suicidality. Eligible participants were asked to complete an electronic informed consent form and were immediately scheduled for the next available class and directed to fill out the web-based baseline assessment. Recruitment occurred in 2 waves, with recruitment windows open for the 6 weeks preceding each scheduled class.

Assessments

All surveys were collected through REDCap (Research Electronic Data Capture; Vanderbilt University) [48], a web-based electronic data capture platform, which is a secure (password-protected), Health Insurance Portability and Accountability Act–compliant platform hosted by the Stanford University School of Medicine. Survey links were sent via SMS text message or email to participants. Participants completed their first survey at enrollment, followed by baseline daily surveys (14 consecutive days of daily surveys), which were initiated 2 to 4 weeks before the scheduled class. Three days before the class, participants completed a preclass survey, and immediately following the class, participants completed a postclass satisfaction survey. The follow-up daily surveys (14 consecutive days) were initiated 1 month after the class. All daily assessments were deployed at the same time each evening. Furthermore participants completed posttreatment surveys at 3 and 6 months after treatment (Multimedia Appendix 1 illustrates the study flow). There was no in-person contact with a study therapist or research staff. SMS text messages and phone call reminders were deployed by a research assistant to encourage completion of all assessments.

Study Intervention: ER

ER was developed in 2013 [25] and compresses pain neuroscience education and key elements and skills from CBT-based pain and mindfulness interventions into a 2-hour single-session class [32,33]. Each class was led by 1 trained doctoral-level clinical psychologist using a PowerPoint (Microsoft Corporation) slide deck via Zoom to participant

XSL•FO
RenderX

cohorts. Didactic content included psychoeducation about mind-body science as it relates to pain and negative mindset, as well as opioid safety. Participants learned skills on how to identify a negative mindset and self-treat it. During the class, participants practiced skills to decrease physiological hyperarousal, improved regulation of cognition and emotion, and engaged in adaptive behaviors to modulate attention and counteract helplessness. Skills included relaxation, cognitive reframing and thought restructuring, and identifying pleasant activities. Participants self-tailored the information described during the class and developed their own comprehensive self-treatment plan to stop and prevent catastrophizing. Participants left the class with a self-written personalized plan and a 20-minute binaural relaxation response app and were provided verbal and written instructions on how to download the app onto their mobile phone or portable electronic device.

Class Platform

One week before the class, patients were emailed a Zoom link to join the class, along with information on when and how to attend the class. In addition, participants were offered a one-on-one Zoom training session with a research assistant before the intervention, if needed. Participant cohorts consisted of 19 to 22 participants per class. The Zoom platform was used to deliver the ER classes, and instructors screen-shared their PowerPoint presentation slides throughout the class. Class participants were encouraged to ask questions at any time through the Zoom chat box and unmute themselves to participate. Zoom classes were password-protected and hosted within the firewalled Stanford University School of Medicine and Stanford Healthcare systems.

Training and Monitoring of Instructors

Instructors were doctoral-level clinical psychologists trained and certified in ER delivery before administering treatment. Existing treatment manuals and highly structured and standardized class content assured treatment fidelity. A research coordinator served as a fidelity rater and directly observed all classes. Cohort effects were minimized due to the highly structured single-session format and limited participant interaction.

Measures

Feasibility and Acceptability

The assessment of the feasibility and acceptability of the ER Zoom-delivered intervention replicated published methods and was assessed immediately after the class [33]. We assessed feasibility by the proportion of participants that attended the class. Acceptability was assessed by participant ratings of treatment satisfaction, relevance, usefulness of information presented, ease of understanding, and likelihood to use the skills learned. Responses were measured on an 11-point Likert scale from 0 to 10, with higher scores indicating higher acceptability and satisfaction across respective items.

In addition, we assessed satisfaction with using the Zoom platform, such as ease of operating Zoom, engagement with the class material during the Zoom class, and ability to find a private location. We assessed comfort with the Zoom instructor and other class participants, connectedness to the Zoom instructor, as well as overall satisfaction with the Zoom platform. Ratings were assessed using a 7-point Likert scale from 1 “strongly disagree” to 7 “strongly agree,” with higher scores indicating greater satisfaction. Finally, we asked participants for open text responses regarding general feedback on the class and what they found helpful or challenging.

Pain and Opioid Use Measures

The following measures were collected at enrollment, before class (3 days before the class), and 3 and 6 months after treatment.

Pain Intensity

Respondents rated their average pain intensity over the previous 7 days on a numerical pain rating scale of 0 “no pain” to 10 “worst pain imaginable” [49,50].

Patient Reported Outcome Measurement Information System Pain Interference

PROMIS (Patient Reported Outcome Measurement Information System) pain interference short form (8 item) [51,52] was used to assess self-reported consequences of pain, including engagement with daily activities. Higher scores indicate higher interference with activities. The National Institutes of Health PROMIS short-form measures have been applied in pain research [51], and selected domains were identified by the Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials as core outcomes [53]. The web-based PROMIS assessment center software [54] was used to calculate the short-form T scores. A standardized T score for each PROMIS measure is generated for each patient. A score of 50 reflects the mean of the US general population, with an SD of 10. Internal consistency of the pain interference measure at baseline was high (α=.96).

Pain Catastrophizing Scale

The 13-item Pain Catastrophizing Scale (PCS) [26] was used to measure negative thoughts and emotional responses to pain. PCS includes 3 subscales: helplessness, magnification, and rumination. The response scale ranges from 0 “not at all” to 4 “all the time”; total sum scores range from 0 to 52. It is scored by summing all items, with higher scores indicating greater catastrophizing. The PCS has demonstrated validity and reliability in mixed-etiology chronic pain and is a psychometrically trusted instrument [55]. The internal consistency of the PCS in our sample at baseline was high (α=.93).

Opioid Dose

Overview

Self-reported opioid type and dose was converted to MEDD using the US Department of Health & Human Services Opioid Oral Morphine Milligram Equivalent Conversion Factors table [56] and the State of Ohio Board of Pharmacy Oral Morphine Milligram Equivalent Conversion Table [57]. Our methods reflected self-reported daily opioid use captured in other trials [34,35,58]. A board-certified anesthesiology and pain medicine physician verified the calculations. In addition, patients were categorized into one of four opioid cohorts based on their opioid

XSL•FO
RenderX

prescription type at enrollment: (1) short-acting prescribed opioids as needed (pro re nata; PRN); (2) prescribed consistent, long-acting opioids; (3) prescribed buprenorphine; and (4) intrathecal pump.

The following measures were collected during the baseline (2-4 weeks before the class) and follow-up (1 month after the class) daily survey periods. Daily behavioral skill use was only assessed during the follow-up daily survey period.

Daily PCS

The daily PCS is a 3-item version of the PCS developed for use in daily diary studies to facilitate research on mechanisms of catastrophizing treatment. It has demonstrated validity and reliability in patients with chronic pain [59]. Scores were calculated by summing all items with a range from 0 to 12. Items assessed patterns of unhelpful thoughts about pain within the last 24 hours and were rated on a 5-point Likert scale from 0 "not at all" to 4 "all the time." Example items included, "During the past 24 hours I kept thinking about how much I hurt." The internal consistency of the follow-up daily PCS in our sample was high (α=.91).

Daily Pain Intensity

Daily average pain ratings were assessed on a 0 to 10 numerical rating scale (0=no pain and 10=worst pain imaginable) based on pain experienced in the past 24 hours [49,60].

Daily Opioid Dose

Daily opioid dose, as reported by participants, was converted to MEDD using the US Department of Health & Human Services Opioid Oral Morphine Milligram Equivalent Conversion Factors table [56] and the State of Ohio Board of Pharmacy Oral Morphine Milligram Equivalent Conversion Table [57]. Our methods reflected self-reported daily opioid use captured in other trials [34,35,58]. A board-certified anesthesiology and pain medicine physician verified the calculations.

Daily Skills Use

Skills use was assessed for 14 days during the follow-up daily assessments only. Three items were asked to assess the frequency of cognitive, behavioral, and relaxation techniques used over the past 24 hours using a 6-point Likert scale ranging from 0 "No use" to 5 "Five or more times." The items were "How many times did you engage in deep breathing or relaxation today?" "How many times did you use distraction or reframing today?" and "How many times did you use pain relief actions today?"

Data Analysis Plan

Analyses were carried out using SPSS (IBM Corp) [61] and RStudio (RStudio, Inc) [62], a per-protocol approach was used. Descriptive data were calculated, including means and SDs across all study variables and time points. To examine attrition, differences in baseline demographic, pain, and opioid characteristics were assessed between those with complete and missing data at each study time point using chi-square tests and ANOVAs. To examine treatment satisfaction, average participant ratings across each satisfaction item were calculated, as well as the proportion of participants who rated each item ≥8 out of 10. To examine satisfaction with the Zoom platform, average participant ratings across each satisfaction item were calculated as well as the proportion of participants who rated each item strongly agree, agree, or somewhat agree (≥5 out of 7).

To examine raw changes in pain intensity, pain catastrophizing, pain interference, and opioid dose from enrollment to 3 and 6 months after treatment, multilevel models with a subject-level random intercept and a fixed effect of time were used. In total, 4 time points (enrollment, before class, 3 months, and 6 months) were included in each model. Linear multilevel models were used to assess changes in pain intensity, pain interference, and pain catastrophizing, and a negative binomial multilevel model was used to assess changes in opioid dose (MEDD), given the highly skewed distribution. Negative binomial models use a natural logarithm link function to account for overdispersion and skewed data. The raw coefficients from this model were converted to a log scale and interpreted as incidence rate ratios, which describe the proportional change in the outcome associated with a 1-unit increase of the predictor [63]. Furthermore, we ran sensitivity analyses for each model, removing tramadol, buprenorphine, and methadone users from the analysis (due to difficulty in equating MEDD). Finally, we calculated Cohen *d* effect sizes (0.20=small, 0.50=medium, and 0.80=large) [64] to aid interpretation.

Across the daily assessments, a similar multilevel modeling approach was used to take advantage of the repeated-measures design. All multilevel models included random intercepts at the participant level and fixed effects of each predictor. To examine predictors of change in daily average pain intensity across the baseline and follow-up daily assessments (relative to enrollment), a linear multilevel model was used. Predictors included daily pain catastrophizing; time (dichotomized factor of pre- vs postclass assessment); dummy-coded opioid cohort (PRN, long-acting, buprenorphine, and intrathecal pump cohorts); and time's interactions with each factor. Change in opioid dose (MEDD) was transformed into percentage change to approximate a normal distribution. Outliers at the upper bound were windsorized [65] to the 95th percentile (Multimedia Appendix 2 provides more details). A linear multilevel model was used to examine predictors of percentage change in opioid dose across the baseline and follow-up daily assessments (relative to enrollment). Predictors included daily pain catastrophizing, daily average pain intensity, time (before vs after class), opioid cohort, and time's interaction with each factor. For both the multilevel models assessing the time and daily predictors, we ran sensitivity analyses for each model, removing tramadol and methadone users from the analysis (due to difficulty in equating MEDD).

Finally, to examine the role of behavioral skill use on changes in pain and opioid dose, 2 additional linear multilevel models were conducted across the follow-up daily assessments. Each behavioral skill was dichotomized into no use and ≥1 use in the past 24 hours and added as predictors in each model. Only dummy-coded PRN and long-acting opioid cohorts were examined due to small sample sizes within intrathecal pump and buprenorphine cohorts and rank deficiency errors with maximal models. To examine changes in average pain intensity,

XSL•FO
RenderX

predictors included daily pain catastrophizing, time (before vs after class), PRN versus long-acting cohort, behavioral skills (relax, reframe, and actions), and time's interaction with every factor. To examine the percentage change in opioid dose, predictors included daily pain catastrophizing, daily average pain intensity, time (before vs after class), PRN versus long-acting cohort, behavioral skills (relax, reframe, and actions), and time's interaction with every factor.

Multimedia Appendix 2 gives a description of pain and opioid dose outcome calculations and all model specifications. Across all models, data were inspected to ensure model assumptions were met (ie, homoscedasticity in residuals), and *P* values were corrected for multiple comparisons using the Benjamini-Hochberg false discovery rate (BH-FDR) procedure, an alternative measurement to the family-wise error rate [66]. The BH-FDR procedure has been shown to have greater power than classical family-wise error rate procedures (eg, Bonferroni), while still adequately controlling for type 1 errors. The desired α level (α=.05), number of tests, and raw *P* value for each test were used to calculate the BH-FDR [66].

Results

Overview

Participants were primarily female (45/59, 76%), White (52/59, 88%), non-Hispanic or Latino (57/59, 97%), married (28/59, 48%), and not working due to pain (28/59, 47%). The most prevalent pain condition was chronic low back pain (35/59, 58%); most had pain >5 years (54/59, 92%; Table 1). Oxycodone was the most commonly prescribed opioid (14/59, 24%; Multimedia Appendix 3 for all opioid types). The use of PRN opioids was most common among the sample (32/59, 78%), followed by long-acting prescribed opioids taken consistently (16/59, 39%), buprenorphine (10/59, 24%), and intrathecal pump (2/59, 5%). Table 2 provides descriptive data for all monthly and daily-level variables.

XSL•FO
RenderX

Table 1. The enrolled sample demographic characteristics of people with chronic pain prescribed daily opioids[a].

Measure	Values
Age (y), mean (SD)	59.13 (14.54)
Gender, n (%)	
Men	13 (22)
Women	45 (76)
Other	1 (2)
Race, n (%)	
American Indian or Alaska Native	2 (3)
Asian	1 (2)
Black or African American	2 (3)
White	52 (87)
>1 race	1 (2)
Other	1 (2)
Ethnicity, n (%)	
Hispanic or Latino	2 (3)
Non-Hispanic or Latino	57 (95)
Marital status, n (%)	
Never married	8 (13)
Partnered	8 (13)
Married	28 (47)
Separated	1 (2)
Widowed	4 (7)
Divorced	10 (17)
Education, n (%)	
Some high school	1 (2)
High school diploma or GED[b]	1 (2)
Some college, no degree	17 (28)
Associates or vocational degree	10 (17)
Bachelor's degree	18 (30)
Master's, professional, or doctoral degree	12 (20)
Annual household income (US $), n (%)	
<24,999	19 (32)
25,000-64,999	9 (15)
65,000-84,999	6 (10)
85,000-124,999	11 (18)
>125,000	14 (23)
Employment[c], n (%)	
Full time or part time	9 (15)
Homemaker or not employed	5 (8)
Not working due to pain	28 (47)
Retired	20 (33)
Receiving disability, n (%)	

XSL•FO
RenderX

Measure	Values
Yes	19 (32)
No	40 (67)
Chronic pain condition[c], n (%)	
Postsurgical recovery	13 (22)
Chronic low back pain	35 (58)
Complex regional pain syndrome	7 (12)
Pelvic pain	7 (12)
Migraines	15 (25)
Headaches	9 (15)
Neuropathic pain	33 (55)
Abdominal pain	11 (18)
Myofascial pain	23 (38)
Fibromyalgia	14 (23)
Other	22 (37)
Pain duration (y), n (%)	
1-5	5 (8)
>5	54 (90)

[a]In total, 60 participants enrolled but 1 withdrew and did not provide demographic information.

[b]GED: General Educational Development.

[c]The participants could select all that apply.

Table 2. Descriptive data of monthly and daily-level study variables across all study time points (enrollment through 6 months) among people with chronic pain prescribed daily opioids[a].

	Enrollment	Baseline daily dairies	Before class	Follow-up daily dairies	3 mo	6 mo
Total, n (%)	60 (100)	56 (93)	53 (88)	40 (67)	38 (63)	36 (60)
Pain intensity, mean (SD)	5.76 (1.71)	5.56 (2.10)	5.58 (1.70)	4.64 (2.30)	5.07 (1.80)	5.34 (2.03)
Pain catastrophizing, mean (SD)	17.29 (10.48)[b]	3.90 (2.90)[c]	15.30 (9.83)[b]	3.10 (2.92)[c]	13.32 (12.12)[b]	13.14 (11.61)[b]
Pain interference, mean (SD)[d]	63.49 (7.45)	__[e]	62.38 (7.28)	—	61.26 (9.10)	61.00 (10.38)
Opioid dose[f], mean (SD)	92.11 (147.13)	96.01 (166.62)	102.84 (171.06)	74.10 (120.79)	84.03 (136.30)	87.42 (136.30)

[a]Daily dairies were completed for 14 days approximately 2 to 4 weeks before the class (baseline daily dairies) and for 14 days approximately 4 weeks after the class (follow-up daily dairies). Means and SDs for this period were calculated across all 14 days.

[b]13-item Pain Catastrophizing Scale.

[c]3-item daily Pain Catastrophizing Scale.

[d]Patient Reported Outcome Measurement Information System pain interference scale *t* score.

[e]Not available.

[f]Opioid dose as measured by morphine-equivalent daily dose.

A total of 60 participants enrolled in the study, and 7% (n=4) withdrew from the study before completing the baseline survey (7% pretreatment attrition). Of the 56 (N=60, 93%) participants who completed the baseline survey, 41 (73%) participants attended the class (68% of the enrolled sample), 38 (63% of the enrolled sample) completed the 3-month follow-up survey, and 36 (60% of the enrolled sample) completed the 6-month follow-up survey (Multimedia Appendix 1 illustrates the study flowchart). Attrition analyses indicated that those who did not attend the class or complete the follow-up daily surveys were more likely to be female participants (χ^2_{40}=8.72, *P*=.01). This pattern held at the 3-month posttreatment assessment (gender: χ^2_{37}=6.51, *P*=.04). No other differences were observed in study variables or demographic characteristics between those who had complete or missing data at any other study time points.

XSL•FO
RenderX

Feasibility and Acceptability

Some (41/60, 68% of the enrolled sample) attended the class and completed the postclass survey. Of those who attended, 59% (24/41) of the sample rated Zoom-delivered ER to be satisfactory (ie, ≥8 out of 10; mean 8.02, SD 2.08), and 73% (30/41) rated the class easy to understand (mean 7.93, SD 2.1). Some (26/41, 63%) rated the class as useful (mean 7.83, SD 2.27), 63% (26/41) rated the class as relevant (mean 8.10, SD 2.21), and 77% (31/41) stated they were likely to use the skills (mean 8.41, SD 1.94). Open feedback was collected about the class, and 63% (26/41) provided comments. The most common feedback was that the class reduced travel barriers (15/41, 37%), increased comfort by being able to participate at home (13/41, 32%), and reduced concerns about the COVID-19 pandemic (6/41, 15%). Some (7/41, 17%) participants reported difficulty engaging in the relaxation exercise or the class due to pain or fatigue, 14% (6/41) reported a preference for an in-person class, and 12% (5/41) had been exposed to similar content in the past or were unsure if it would be helpful for their specific pain condition.

Nearly all participants (39/41, 95%) found the Zoom platform easy to operate (≥5 out of 7; mean 6.44, SD 0.92); 97% (40/41) easily found a private location to attend the class (mean 6.71, SD 0.68), and 93% (38/41) reported ease in engaging in the class material via Zoom (mean 6.05, SD 1.24). Some (26/41, 63%) participants reported feeling connected to the instructor (mean 5.22, SD 1.64). Overall, 95% (39/41) of participants reported being somewhat satisfied (n=3, 7%), satisfied (n=17, 42%), or completely satisfied (n=19, 46%) with the Zoom platform (mean 6.24, SD 0.97). A majority (33/41, 80%) of participants reported no problems with the Zoom platform. Of those who reported a problem (n=8, 20%), the most frequently cited problem was unstable internet connection and poor audio quality (n=5, 12%).

Pain and Opioid Use Outcomes

Findings from the 4 multilevel models examining the effect of time on raw changes in pain and opioid outcomes are outlined in Table 3. The corrected significance thresholds for each model were pain intensity (P=.008), pain interference (P=.001), pain catastrophizing (P=.01), and opioid dose (P=.05). Results showed a significant reduction in pain catastrophizing from enrollment to 3 months (mean –2.89, SD 10.89), and from enrollment to 6 months (mean –2.83, SD 10.89) both with a small effect size (Cohen d=0.35 and 0.37, respectively). The average pain intensity significantly reduced from enrollment to 3 months posttreatment (mean –0.52, SD 1.40), indicating a small effect size (Cohen d=0.39). However, average pain intensity did not maintain a significant change from enrollment to 3 months in our sensitivity analysis (Multimedia Appendix 4) nor from enrollment to 6 months posttreatment (mean –0.24, SD 1.42). Reductions in pain interference from enrollment to 3 months (mean –2.03, SD 4.94) and 6 months (mean –2.14, SD 6.77) did not survive correction, although they demonstrated small effect sizes (Cohen d=0.27 and 0.28, respectively). There were no significant reductions in opioid dose across the monthly study time points.

XSL•FO
RenderX

Table 3. Multilevel regressions examining changes in pain and opioid use outcomes from baseline through 6-month time points among people with chronic pain prescribed daily opioids.

Outcomes and variables	Coefficient[a] (SE)	Test statistic[a]	*P* value	Cohen *d*[b]
Linear				
Pain intensity				
Intercept	5.76 (0.23)	24.68	<.001[c]	—[d]
Before class[e]	–0.16 (0.19)	–0.84	.40	0.10
3 mo[e]	–0.56 (0.21)	–2.67	.008[c]	0.39
6 mo[e]	–0.29 (0.21)	–1.37	.17	0.22
Pain interference				
Intercept	63.49 (1.09)	58.02	<.001[c]	—
Before class[e]	–1.06 (0.77)	–1.36	.18	0.15
3 mo[e]	–1.98 (0.88)	–2.25	.03	0.27
6 mo[e]	–1.95 (0.89)	–2.18	.03	0.28
Pain c atastrophizing				
Intercept	17.29 (1.44)	12.05	<.001[c]	—
Before class[e]	–2.07 (1.22)	–1.70	.09	.20
3 mo[e]	–3.49 (1.38)	–2.53	.01[c]	.35
6 mo[e]	–3.61 (1.40)	–2.57	.01[c]	.37
Negative binomial				
Opioid dose (MEDD[f])				
Intercept	46.50 (7.13)	25.00	<.001[c]	—
Before class[e]	1.02 (0.07)	0.29	.77	–0.07
3 mo[e]	0.89 (0.07)	–1.52	.13	0.06
6 mo[e]	0.96 (0.07)	–.48	.63	0.03

[a]The linear multilevel model reports an unstandardizedβ coefficient and *t* value statistics. The negative binomial multilevel model, via glmer, reports incidence rate ratios and *z*-value statistics.

[b]The effect size *d* calculated as baseline minus follow-up divided by pooled SD. Positive *d* values indicate improvements.

[c]Significant values based on the corrected *P* value.

[d]Not available.

[e]Reference group is the baseline.

[f]MEDD: morphine-equivalent daily dose.

Across the 2-week baseline daily assessment period (before class), 93% (56/60) of participants completed 95.5% (749/784) of deployed daily surveys. Across the 2-week follow-up daily assessment period (after class), 68% (41/60) of participants completed 95.6% (549/574) of deployed daily surveys. From the 41 participants who attended class, 1136 baseline and follow-up daily surveys were used to estimate changes in pain and opioid dose using linear multilevel models (Table 4). The corrected significance thresholds for each model were pain intensity (*P*=.001) and opioid dose (*P*=.003). Results showed that higher daily pain catastrophizing was associated with increases in average pain intensity across the daily assessments relative to enrollment (β=0.42, SE 0.02, *P*<.001), and there were greater reductions in pain across the follow-up daily surveys (after class) compared to the baseline daily surveys (before class; β=–0.51, SE 0.12, *P*<.001) although this effect did not survive sensitivity analyses (β=–0.41, SE 0.13, *P*=.002; Multimedia Appendix 5). Opioid cohort and all interaction factors were not significantly associated with change in average pain intensity. When examining changes in opioid dosage, higher daily pain catastrophizing (β=3.14, SE 0.74, *P*<.001) and daily average pain (β=3.47, SE 1.16, *P*=.003) were associated with increases in opioid dose across the daily assessments relative to enrollment, although the effect of pain did not survive our sensitivity analyses (β=3.58, SE 1.40, *P*=.01). Those taking short-acting opioids as needed (PRN)

XSL•FO
RenderX

showed a trend toward reduced opioid dose across the daily assessments (β=–29.88, SE 10.68, *P*=.03), although this association did not survive correction. Furthermore, the time × PRN cohort interaction did not survive correction, although it indicated a trend toward greater opioid dose reductions for the PRN cohort across the follow-up daily surveys (after class) compared to the baseline daily surveys (before class; β=–9.31, SE 3.94, *P*=.02). Figure 1 depicts the raw percentage change in opioid dose between the long-acting and PRN cohorts.

XSL•FO
RenderX

Table 4. Multilevel linear regressions predicting change in pain intensity and opioid dose from enrollment across the baseline (before class) and follow-up (after class) daily diaries among people with chronic pain prescribed daily opioids.

Outcome and variables	Coefficient[a] (SE)	Test statistic[a]	*P* value
Change in pain intensity[b]			
Intercept	–0.42 (0.46)	–0.92	.36
PCS[c,d]	0.42 (0.02)	18.94	<.001[e]
Pre-post[f]	–0.51 (0.12)	–4.20	<.001[e]
PRN[g] cohort	–0.55 (0.58)	–0.96	.34
Long-acting cohort	0.23 (0.65)	0.36	.72
Buprenorphine cohort	–0.82 (0.69)	–1.18	.25
ITP[h] cohort	0.77 (1.45)	0.53	.60
Pre-post x PCS	0.03 (0.03)	0.99	.32
Pre-post x PRN	0.21 (0.15)	1.39	.16
Pre-post x LA[i]	–0.27 (0.17)	–1.58	.12
Pre-post x Buprenorphine	–0.04 (0.18)	–0.20	.84
Pre-post x ITP	–0.32 (0.39)	–0.82	.41
Percentage change in opioid dose[j]			
Intercept	5.77 (9.66)	0.60	.55
PCS[d]	3.14 (0.74)	4.23	<.001[e]
Pre-post[e]	1.40 (3.11)	0.45	.65
Average pain	3.47 (1.16)	2.99	.003[e]
PRN cohort	–26.88 (12.21)	–2.20	.03
LA cohort	–17.36 (13.65)	–1.27	.21
Buprenorphine cohort	–6.97 (14.59)	–0.48	.64
ITP cohort	–17.64 (30.51)	–0.58	.56
Pre-post × PCS	–1.72 (1.13)	–1.52	.13
Pre-post × pain	–1.55 (1.78)	–0.87	.38
Pre-post × PRN	–9.31 (3.94)	–2.37	.02
Pre-post × LA	3.08 (4.39)	0.70	.48
Pre-post × buprenorphine	–8.93 (4.67)	–1.92	.06
Pre-post × ITP	16.98 (9.90)	1.72	.09

[a]The linear multilevel model reports an unstandardized β coefficient and *t* value statistics. [b]The outcome is change in average pain intensity in the past 24 hours and depicts a change at each daily assessment compared to the enrollment.

[c]PCS: Pain Catastrophizing Scale.

[d]Daily PCS.

[e]Significant values based on the corrected *P* value.

[f]0: preclass daily assessments; 1: postclass daily assessments.

[g]PRN: pro re nata.

[h]ITP: intrathecal pump.

[i]LA: long acting.

[j]The percentage change in opioid dose is measured using morphine-equivalent daily dose and depicts a change at each daily assessment compared to the enrollment.

XSL•FO
RenderX

Figure 1. The raw percentage change in opioid dose (morphine-equivalent daily dose; MEDD) from enrollment across study time points between participants with chronic pain taking long-acting opioids and short-acting (prescribed opioids as needed) opioids.

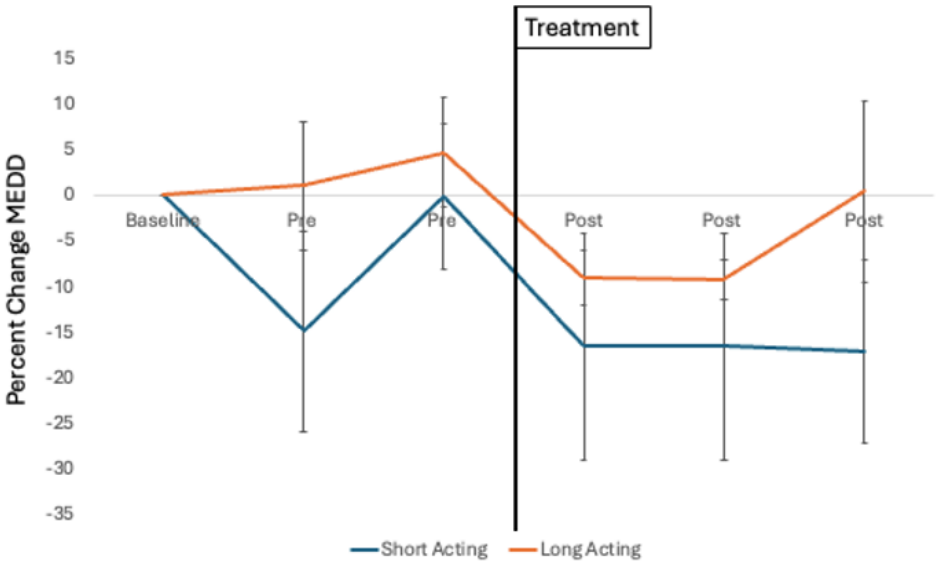

Across the follow-up daily surveys (1 month after the class; n=549), 446 (81.2%) of the daily surveys endorsed using relaxation at least once, 403 (73.4%) endorsed using reframe at least once, and 410 (74.7%) endorsed using pain relief actions at least once. Findings from the multilevel models are provided in Multimedia Appendix 6. The corrected significance thresholds for each model were pain intensity (P=.001) and opioid dose (P=.05). Results showed that higher daily pain catastrophizing was associated with increase in average pain intensity relative to enrollment (β=0.48, SE 0.04, P<.001). The use of behavioral skills, time, and interactions was not a significant predictor of change in average pain intensity. Furthermore, daily pain catastrophizing, daily pain, use of the behavioral skills, time, or their interactions were not significant predictors of percentage change in opioid dose across the follow-up daily surveys.

Discussion

Principal Findings

We performed an uncontrolled pilot study to investigate the initial feasibility, acceptability, and preliminary impact of a Zoom-delivered ER class among people with chronic pain taking daily opioid medication. Our results showed significant reductions in pain catastrophizing and average pain intensity that lasted 3 months after treatment, although with small effects. Pain catastrophizing, a target of the ER class, continued to exhibit significant reductions at 6 months after treatment. Daily-level results also showed a trend toward opioid dose reduction for people taking PRN opioids, though did not survive correction (P=.02). The findings from this pilot study provide justification for a larger, appropriately powered RCT.

Feasibility and Acceptability

We conducted this study during the first year of the COVID-19 pandemic (2020 to 2021), while access to traditional, in-person pain psychology interventions was extremely limited. Although we observed lower-than-hypothesized engagement rates (≥80%), with 68% (41/60) of the enrolled patients attending the class, we noted that our engagement was relatively high when compared to the substantially reduced health care use that occurred during the early stages of the COVID-19 pandemic [21]. For those who participated in the class, 63% (26/41) to 77% (31/41) of the patients rated the class as satisfactory, useful, and easy to understand. Furthermore, they reported high satisfaction with web-based delivery, which reduced travel barriers and concerns about the COVID-19 pandemic and increased comfort by attending the intervention from home. In addition, ease and comfort with digital assessments was evident, as >93% (749/784) of the offered daily surveys were completed (ie, at least 26 of 28 days per person). Future work that involves follow-up interviews with participants is needed to identify potential adaptations to improve the intervention's engagement and acceptability. On the basis of our initial findings, implementing strategies to address digital literacy and poor internet access may improve engagement, along with further tailoring of the relaxation exercise to better engage people with more severe pain and fatigue. In addition, increased reminders and incentives may increase engagement in the intervention.

Pain and Opioid Use Outcomes

Following the ER class, individuals reported significant reductions in pain catastrophizing at 3 and 6 months after treatment and significant reductions in pain intensity at 3 months after treatment. Furthermore, results showed a trend toward decreased pain interference at 3 and 6 months after treatment, although these did not survive multiplicity correction (P=.03). This study of people taking daily opioids yielded smaller effect sizes (Cohen d=0.35-0.39) as compared to prior RCTs of in-person ER [32] (Cohen d=0.43-0.89) and Zoom-delivered ER [33] (Cohen d=0.54-0.76) in patients with chronic pain with no enrollment criterion applied for daily opioid use. While there are notable methodological differences between studies (eg, different pain conditions, enrollment criteria, and control conditions), the smaller effect sizes may also be due to higher rates of disability, which were 32% in this sample as compared to 5% to 10% in the previous 2 studies. Patients on long-term opioid therapy may require more intensive treatment, a different type of treatment, or booster sessions, especially in the setting

XSL•FO
RenderX

of greater disability. Regardless, this study's findings suggest that Zoom-delivered ER may impart small benefits in pain intensity and pain catastrophizing, and trending toward improvement for pain interference (though did not survive correction, *P*=.03), in people with chronic pain on long-term opioid therapy, warranting further testing in an RCT.

Our study is among the few to use daily digital assessments among patients with mixed-etiology chronic pain and taking long-term opioid medication. Our findings replicate previous work showing that higher daily reports of pain catastrophizing are associated with worse pain intensity, and higher daily pain intensity and pain catastrophizing are associated with taking greater doses of prescribed opioids [31,39,59]. In addition, daily ratings of pain decreased from pre-to-post class, similar to previous work that showed a mindfulness intervention (MORE) reduced daily-level pain, negative affect, opioid craving, and opioid use [45,46]. Furthermore, we examined the differences between patients prescribed PRN opioids and long-acting opioids. Results showed a trend that people prescribed PRN opioids were more likely to reduce their opioid dose from before to after class as compared to patients on long-acting opioids, although this association did not survive our rigorous multiple comparison correction (*P*=.02; Figure 1 illustrates a raw percentage change across both groups). Our findings provide a signal that a brief, Zoom-delivered behavioral pain treatment may help patients taking PRN opioids rely less on their medication, as has been found with other multisession behavioral treatments [15,16]. For those on long-acting opioids, patients are likely to need additional support and treatments, with active clinician-patient engagement targeting a slow, gradual weaning of their opioid dose over several weeks or months [67]. Behavioral treatments and psychoeducation are important treatment components as they offer patients additional pain coping skills and may help reduce potential pain flares during the weaning process [17,68,69]. Future RCTs are warranted to examine Zoom-delivered ER alone in reducing PRN opioid use and in the context of an opioid weaning protocol in conjunction with prescribers.

Limitations

We noted several limitations in this study that are important to consider when interpreting the findings. First, this study was an uncontrolled pilot trial. Therefore, inferences about intervention causation cannot be assumed. We did use a rigorous statistical approach to control for potential type 1 errors. Furthermore, the actual therapeutic effect size cannot be ascertained and separated from potential regression to the mean effects. A larger RCT (NCT03950791) is currently underway to further assess the efficacy of Zoom-delivered ER among patients with chronic pain receiving daily prescribed opioids [58]. Second, our sample consisted of largely White individuals, non-Hispanic or Latino individuals, and educated women, although our sample did report higher rates of disability (19/59, 32%) than in our previous studies. Future RCTs are needed to assess the efficacy of the intervention in diverse and underserved patient populations who experience greater barriers to treatment [70]. Third, this study evidenced significant attrition (only 41/60, 60% of the enrolled sample completed the final 6-month assessment). While we did not find significant differences between those with complete versus missing data on our study variables, stressors related to the COVID-19 pandemic likely impacted study attrition. Importantly, we had exceptional compliance rates with digital daily assessments (93%-95%). Future research may consider using additional engagement strategies, such as increased digital assessments or phone calls in-between study time points. Fourth, we suspect that we did not see any significant associations between daily behavioral skill use and daily pain or opioid use because of the once-daily assessments that were completed at the same time each day and asked participants to recall information from the last 24 hours. Future research should use stratified random sampling, prompts with a shorter recall window, and other objective measures (eg, electronic pill caps) to reduce measurement error [71]. Finally, there are other variables that we did not measure that, in previous work, have moderated associations between pain and opioid use, such as opioid craving [72] and positive affect [45]. Future research using larger samples would benefit from modeling both known risk factors (eg, opioid craving, negative affect, and pain catastrophizing) [39,72] and protective factors (eg, positive affect and use of behavioral skills) [45] to further elucidate their impact on daily reports of pain and opioid use as well as their change following behavioral pain treatments.

Conclusions

Improvements to future Zoom-delivered ER iterations are needed to improve feasibility and acceptability among people with chronic pain and daily prescribed opioid use. However, despite this, we also found a promising preliminary impact of the intervention on pain outcomes. There was high compliance with digital daily assessments, and results showed a promising trend of opioid dose reduction among people taking PRN opioids (though did not survive correction, *P*=.02). Our results indicate significant reductions in pain catastrophizing and daily pain intensity, suggesting that even a brief, digitally delivered behavioral intervention can positively influence pain outcomes in this population. Although reductions in opioid use did not reach statistical significance (*P*=.02), the observed trend among patients prescribed PRN opioids highlights ER's potential role in supporting reduced reliance on medication.

We developed this study to inform a larger RCT that aims to rigorously assess the efficacy of Zoom-delivered ER using digital daily and monthly level assessments currently in data collection [58]. Our initial findings suggest that Zoom-delivered ER may be a promising intervention that can improve pain-related outcomes and reduce common barriers to accessing pain psychology interventions, such as travel and time burdens [20]. Dissemination of brief, accessible, and scalable pain psychology interventions, such as ER, stands to improve patient outcomes by increasing equitable access to effective pain psychology treatments and may be useful tools to help reduce pain and opioid use. ER offers a promising approach to expanding the reach of behavioral pain treatments, particularly for populations with limited access to in-person care.

XSL•FO
RenderX

Acknowledgments

Research reported in this work was supported by the National Institute on Drug Abuse of the National Institutes of Health under award numbers K12DA050607 (KAE), T32DA035165 (KAE and TCD), K23DA047473 (MSZ), K23DA048972 (DSY), and K24DA053564 (BDD). Additional funding support from the National Institute on Neurological Diseases and Stroke was received under the award numbers K24NS126781 and R61NS11865 (SCM). Funders were not involved in the study design, data collection, analysis or interpretation of results, or manuscript writing.

Data Availability

The data generated are available upon reasonable request to the corresponding author (MSZ).

Conflicts of Interest

Stanford University receives fees for continuing medical education Empowered Relief (ER) clinician certification workshops; BDD receives no personal monies for the ER certification workshops nor for the dissemination or adoption of ER. BDD is the chief science adviser at AppliedVR, and her consulting role with this company (personal fees) is unrelated to the current research. BDD receives royalties for 4 pain treatment books she has authored or coauthored. She is the principal investigator for 2 pain research awards from the Patient-Centered Research Outcomes Research Institute, one of which involves the study of ER. BDD is the principal investigator for 2 National Institutes of Health (NIH) grants that are investigating the efficacy of ER (one is the Mobile and Online-Based Interventions to Lessen Pain [MOBILE] Relief study). BDD serves on the board of directors for the Institute for Brain Potential, which is licensed by Stanford University to offer ER clinician certification workshops, and is on the medical advisory board for the Facial Pain Association. BDD is a scientific member of the NIH Interagency Pain Research Coordinating Committee, a former member of the Centers for Disease Control and Prevention Opioid Workgroup (2020 to 2021), and a current member of the Pain Advisory Group of the American Psychological Association. SCM receives research funding from the NIH, Food and Drug Administration, and Patient-Centered Outcomes Research Institute (administered through Stanford University). He is an unpaid adviser to both ACTTION (Analgesic, Anesthetic, and Addiction Clinical Trial Translations and Innovations, Opportunities, and Networks) on their oversight committee and the American Chronic Pain Association for their scientific oversight. All other authors report no other disclosures or conflicts of interest.

Multimedia Appendix 1
Participant flowchart.
[DOCX File , 52 KB - formative_v9i1e68292_app1.docx]

Multimedia Appendix 2
Supplementary methods.
[DOCX File , 22 KB - formative_v9i1e68292_app2.docx]

Multimedia Appendix 3
Type of prescribed opioid used at enrollment.
[DOCX File , 19 KB - formative_v9i1e68292_app3.docx]

Multimedia Appendix 4
A sensitivity analysis of multilevel regressions examining changes in pain and opioid outcomes across monthly assessments.
[DOCX File , 20 KB - formative_v9i1e68292_app4.docx]

Multimedia Appendix 5
A sensitivity analysis of multilevel linear regressions predicting change in pain intensity and opioid dose from baseline across the baseline (before class) and follow-up (after class) daily dairies.
[DOCX File , 20 KB - formative_v9i1e68292_app5.docx]

Multimedia Appendix 6
Multilevel linear regressions predicting change in pain intensity and opioid dose from enrollment across the follow-up (after class) daily dairies.
[DOCX File , 25 KB - formative_v9i1e68292_app6.docx]

References

1. Yong RJ, Mullins PM, Bhattacharyya N. Prevalence of chronic pain among adults in the United States. Pain 2022 Feb 01;163(2):e328-e332. [doi: 10.1097/j.pain.0000000000002291] [Medline: 33990113]

XSL•FO
RenderX

2. Institute of Medicine (US) Committee on Advancing Pain Research, Care, and Education. Relieving Pain in America: A Blueprint for Transforming Prevention, Care, Education, and Research. Vol. xvii. Washington, DC: National Academies Press (US); 2011.
3. Dominick CH, Blyth FM, Nicholas MK. Unpacking the burden: understanding the relationships between chronic pain and comorbidity in the general population. Pain 2012 Feb;153(2):293-304. [doi: 10.1016/j.pain.2011.09.018] [Medline: 22071318]
4. Aroke HA, Buchanan AL, Bratberg J, Hogan JW, Rich JD, Wen X, et al. Initial patterns of prescription opioid supply and risk of mortality among insured adults in the United States. J Addict Med 2021 Apr 01;15(2):99-108 [FREE Full text] [doi: 10.1097/ADM.0000000000000701] [Medline: 32769771]
5. Farrar JT, Bilker WB, Cochetti PT, Argoff CE, Haythornthwaite J, Katz NP, et al. Evaluating the stability of opioid efficacy over 12 months in patients with chronic noncancer pain who initially demonstrate benefit from extended release oxycodone or hydrocodone: harmonization of Food and Drug Administration patient-level drug safety study data. Pain 2022 Jan 01;163(1):47-57 [FREE Full text] [doi: 10.1097/j.pain.0000000000002331] [Medline: 34261978]
6. Annual surveillance report of drug-related risks and outcomes. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/overdose-prevention/media/pdfs/pubs/2019-cdc-drug-surveillance-report.pdf [accessed 2024-04-29]
7. Dasgupta N, Funk MJ, Proescholdbell S, Hirsch A, Ribisl KM, Marshall S. Cohort study of the impact of high-dose opioid analgesics on overdose mortality. Pain Med 2016 Jan;17(1):85-98. [doi: 10.1111/pme.12907] [Medline: 26333030]
8. Gomes T, Juurlink D, Moineddin R, Gozdyra P, Dhalla I, Paterson J, et al. Geographical variation in opioid prescribing and opioid-related mortality in Ontario. Healthc Q 2011;14(1):22-24 [FREE Full text] [doi: 10.12927/hcq.2011.22153] [Medline: 21301233]
9. Gatchel RJ, Peng YB, Peters ML, Fuchs PN, Turk DC. The biopsychosocial approach to chronic pain: scientific advances and future directions. Psychol Bull 2007 Jul;133(4):581-624. [doi: 10.1037/0033-2909.133.4.581] [Medline: 17592957]
10. Bruns D, Mueller K, Warren PA. Biopsychosocial law, health care reform, and the control of medical inflation in Colorado. Rehabil Psychol 2012 May;57(2):81-97. [doi: 10.1037/a0028623] [Medline: 22686548]
11. Driscoll MA, Edwards RR, Becker WC, Kaptchuk TJ, Kerns RD. Psychological interventions for the treatment of chronic pain in adults. Psychol Sci Public Interest 2021 Sep;22(2):52-95. [doi: 10.1177/15291006211008157] [Medline: 34541967]
12. Jamison RN, Ross EL, Michna E, Chen LQ, Holcomb C, Wasan AD. Substance misuse treatment for high-risk chronic pain patients on opioid therapy: a randomized trial. Pain 2010 Sep;150(3):390-400 [FREE Full text] [doi: 10.1016/j.pain.2010.02.033] [Medline: 20334973]
13. Naylor MR, Naud S, Keefe FJ, Helzer JE. Therapeutic Interactive Voice Response (TIVR) to reduce analgesic medication use for chronic pain management. J Pain 2010 Dec;11(12):1410-1419 [FREE Full text] [doi: 10.1016/j.jpain.2010.03.019] [Medline: 20620119]
14. Dindo L, Zimmerman MB, Hadlandsmyth K, StMarie B, Embree J, Marchman J, et al. Acceptance and commitment therapy for prevention of chronic postsurgical pain and opioid use in at-risk veterans: a pilot randomized controlled study. J Pain 2018 Oct;19(10):1211-1221 [FREE Full text] [doi: 10.1016/j.jpain.2018.04.016] [Medline: 29777950]
15. Vowles KE, Witkiewitz K, Cusack KJ, Gilliam WP, Cardon KE, Bowen S, et al. Integrated behavioral treatment for veterans with co-morbid chronic pain and hazardous opioid use: a randomized controlled pilot trial. J Pain 2020;21(7-8):798-807 [FREE Full text] [doi: 10.1016/j.jpain.2019.11.007] [Medline: 31760109]
16. Garland EL, Hanley AW, Nakamura Y, Barrett JW, Baker AK, Reese SE, et al. Mindfulness-oriented recovery enhancement vs supportive group therapy for co-occurring opioid misuse and chronic pain in primary care: a randomized clinical trial. JAMA Intern Med 2022 Apr 01;182(4):407-417 [FREE Full text] [doi: 10.1001/jamainternmed.2022.0033] [Medline: 35226053]
17. Sullivan MD, Turner JA, DiLodovico C, D'Appollonio A, Stephens K, Chan Y. Prescription opioid taper support for outpatients with chronic pain: a randomized controlled trial. J Pain 2017 Mar;18(3):308-318 [FREE Full text] [doi: 10.1016/j.jpain.2016.11.003] [Medline: 27908840]
18. Rini C, Williams DA, Broderick JE, Keefe FJ. Meeting them where they are: using the internet to deliver behavioral medicine interventions for pain. Transl Behav Med 2012 Mar 28;2(1):82-92 [FREE Full text] [doi: 10.1007/s13142-011-0107-2] [Medline: 22924084]
19. Spears CA, Hodges SE, Kiyani M, Yang Z, Edwards RM, Musick A, et al. Health care resource utilization and management of chronic, refractory low back pain in the United States. Spine (Phila Pa 1976) 2020 Oct 15;45(20):E1333-E1341 [FREE Full text] [doi: 10.1097/BRS.0000000000003572] [Medline: 32453242]
20. Darnall BD, Scheman J, Davin S, Burns JW, Murphy JL, Wilson AC, et al. Pain psychology: a global needs assessment and national call to action. Pain Med 2016 Feb;17(2):250-263 [FREE Full text] [doi: 10.1093/pm/pnv095] [Medline: 26803844]
21. Lacasse A, Pagé MG, Dassieu L, Sourial N, Janelle-Montcalm A, Dorais M, et al. Impact of the COVID-19 pandemic on the pharmacological, physical, and psychological treatments of pain: findings from the chronic pain and COVID-19 Pan-Canadian study. Pain Rep 2021;6(1):e891 [FREE Full text] [doi: 10.1097/PR9.0000000000000891] [Medline: 33598594]
22. Serrano-Ibáñez ER, Esteve R, Ramírez-Maestre C, Ruiz-Párraga GT, López-Martínez AE. Chronic pain in the time of COVID-19: stress aftermath and central sensitization. Br J Health Psychol 2021 May;26(2):544-552. [doi: 10.1111/bjhp.12483] [Medline: 33099793]

XSL•FO
RenderX

23. Shanthanna H, Nelson AM, Kissoon N, Narouze S. The COVID-19 pandemic and its consequences for chronic pain: a narrative review. Anaesthesia 2022 Sep;77(9):1039-1050 [FREE Full text] [doi: 10.1111/anae.15801] [Medline: 35848380]
24. Carey EP, Nolan C, Kerns RD, Ho PM, Frank JW. Association between facility-level utilization of non-pharmacologic chronic pain treatment and subsequent initiation of long-term opioid therapy. J Gen Intern Med 2018 May;33(Suppl 1):38-45 [FREE Full text] [doi: 10.1007/s11606-018-4324-y] [Medline: 29633137]
25. Darnall BD, Sturgeon JA, Kao MC, Hah JM, Mackey SC. From catastrophizing to recovery: a pilot study of a single-session treatment for pain catastrophizing. J Pain Res 2014;7:219-226 [FREE Full text] [doi: 10.2147/JPR.S62329] [Medline: 24851056]
26. Sullivan MJ, Bishop SR, Pivik J. The pain catastrophizing scale: development and validation. Psychol Assess 1995 Dec;7(4):524-532. [doi: 10.1037/1040-3590.7.4.524]
27. Martinez-Calderon J, Jensen MP, Morales-Asencio JM, Luque-Suarez A. Pain catastrophizing and function in individuals with chronic musculoskeletal pain: a systematic review and meta-analysis. Clin J Pain 2019 Mar;35(3):279-293. [doi: 10.1097/AJP.0000000000000676] [Medline: 30664551]
28. Sharifzadeh Y, Kao M, Sturgeon JA, Rico TJ, Mackey S, Darnall BD. Pain catastrophizing moderates relationships between pain intensity and opioid prescription: nonlinear sex differences revealed using a learning health system. Anesthesiology 2017 Jul;127(1):136-146 [FREE Full text] [doi: 10.1097/ALN.0000000000001656] [Medline: 28614083]
29. Martel MO, Edwards RR, Jamison RN. The relative contribution of pain and psychological factors to opioid misuse: a 6-month observational study. Am Psychol 2020 Sep;75(6):772-783. [doi: 10.1037/amp0000632] [Medline: 32915022]
30. Severeijns R, Vlaeyen JW, van den Hout MA, Weber WE. Pain catastrophizing predicts pain intensity, disability, and psychological distress independent of the level of physical impairment. Clin J Pain 2001 Jun;17(2):165-172. [doi: 10.1097/00002508-200106000-00009] [Medline: 11444718]
31. Carpenter RW, Lane SP, Bruehl S, Trull TJ. Concurrent and lagged associations of prescription opioid use with pain and negative affect in the daily lives of chronic pain patients. J Consult Clin Psychol 2019 Oct;87(10):872-886 [FREE Full text] [doi: 10.1037/ccp0000402] [Medline: 31556664]
32. Darnall BD, Roy A, Chen AL, Ziadni MS, Keane RT, You DS, et al. Comparison of a single-session pain management skills intervention with a single-session health education intervention and 8 sessions of cognitive behavioral therapy in adults with chronic low back pain: a randomized clinical trial. JAMA Netw Open 2021 Aug 02;4(8):e2113401 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.13401] [Medline: 34398206]
33. Ziadni MS, Gonzalez-Castro L, Anderson S, Krishnamurthy P, Darnall BD. Efficacy of a single-session "empowered relief" zoom-delivered group intervention for chronic pain: randomized controlled trial conducted during the COVID-19 pandemic. J Med Internet Res 2021 Sep 10;23(9):e29672 [FREE Full text] [doi: 10.2196/29672] [Medline: 34505832]
34. Darnall BD, Ziadni MS, Krishnamurthy P, Flood P, Heathcote LC, Mackey IG, et al. "My surgical success": effect of a digital behavioral pain medicine intervention on time to opioid cessation after breast cancer surgery-a pilot randomized controlled clinical trial. Pain Med 2019 Nov 01;20(11):2228-2237 [FREE Full text] [doi: 10.1093/pm/pnz094] [Medline: 31087093]
35. Ziadni MS, You DS, Keane R, Salazar B, Jaros S, Ram J, et al. "My surgical success": feasibility and impact of a single-session digital behavioral pain medicine intervention on pain intensity, pain catastrophizing, and time to opioid cessation after orthopedic trauma surgery-a randomized trial. Anesth Analg 2022 Aug 01;135(2):394-405. [doi: 10.1213/ANE.0000000000006088] [Medline: 35696706]
36. Darnall BD, Burns JW, Hong J, Roy A, Slater K, Poupore-King H, et al. Empowered relief, cognitive behavioral therapy, and health education for people with chronic pain: a comparison of outcomes at 6-month follow-up for a randomized controlled trial. Pain Rep 2024 Jan;9(1):e1116 [FREE Full text] [doi: 10.1097/PR9.0000000000001116] [Medline: 38288134]
37. Eccleston C, Blyth FM, Dear BF, Fisher EA, Keefe FJ, Lynch ME, et al. Managing patients with chronic pain during the COVID-19 outbreak: considerations for the rapid introduction of remotely supported (eHealth) pain management services. Pain 2020 May;161(5):889-893 [FREE Full text] [doi: 10.1097/j.pain.0000000000001885] [Medline: 32251203]
38. Hadjiat Y, Arendt-Nielsen L. Digital health in pain assessment, diagnosis, and management: overview and perspectives. Front Pain Res (Lausanne) 2023;4:1097379 [FREE Full text] [doi: 10.3389/fpain.2023.1097379] [Medline: 37139342]
39. Finan PH, Carroll CP, Moscou-Jackson G, Martel MO, Campbell CM, Pressman A, et al. Daily opioid use fluctuates as a function of pain, catastrophizing, and affect in patients with sickle cell disease: an electronic daily diary analysis. J Pain 2018 Jan;19(1):46-56 [FREE Full text] [doi: 10.1016/j.jpain.2017.08.010] [Medline: 28943232]
40. Miller MB, Curtis AF, Chan WS, Deroche CB, McCrae CS. Daily associations between sleep and opioid use among adults with comorbid symptoms of insomnia and fibromyalgia. J Clin Sleep Med 2021 Apr 01;17(4):729-737 [FREE Full text] [doi: 10.5664/jcsm.9002] [Medline: 33226334]
41. Bruneau A, Frimerman L, Verner M, Sirois A, Fournier C, Scott K, et al. Day-to-day opioid withdrawal symptoms, psychological distress, and opioid craving in patients with chronic pain prescribed opioid therapy. Drug Alcohol Depend 2021 Aug 01;225:108787. [doi: 10.1016/j.drugalcdep.2021.108787] [Medline: 34091157]
42. Verner M, Sirois A, Coutu-Nadeau E, Fournier C, Carriere J, Martel MO. The association between perceived injustice and opioid craving in patients with chronic pain: the mediating role of daily pain intensity, negative affect, and catastrophizing. J Addict Med 2023;17(1):35-41. [doi: 10.1097/ADM.0000000000001022] [Medline: 35861344]

XSL•FO
RenderX

43. Frimerman L, Verner M, Sirois A, Scott K, Bruneau A, Perez J, et al. Day-to-day hedonic and calming effects of opioids, opioid craving, and opioid misuse among patients with chronic pain prescribed long-term opioid therapy. Pain 2021 Aug 01;162(8):2214-2224. [doi: 10.1097/j.pain.0000000000002220] [Medline: 33729213]
44. Anderson Goodell EM, Nordeck C, Finan PH, Vandrey R, Dunn KE, Thrul J. Feasibility and acceptability of using smartphone-based EMA to assess patterns of prescription opioid and medical cannabis use among individuals with chronic pain. Internet Interv 2021 Dec;26:100460 [FREE Full text] [doi: 10.1016/j.invent.2021.100460] [Medline: 34646751]
45. Garland EL, Bryan CJ, Finan PH, Thomas EA, Priddy SE, Riquino MR, et al. Pain, hedonic regulation, and opioid misuse: modulation of momentary experience by Mindfulness-Oriented Recovery Enhancement in opioid-treated chronic pain patients. Drug Alcohol Depend 2017 Apr 01;173 Suppl 1:S65-S72 [FREE Full text] [doi: 10.1016/j.drugalcdep.2016.07.033] [Medline: 28363322]
46. Garland EL, Gullapalli BT, Prince KC, Hanley AW, Sanyer M, Tuomenoksa M, et al. Zoom-based mindfulness-oriented recovery enhancement plus just-in-time mindfulness practice triggered by wearable sensors for opioid craving and chronic pain. Mindfulness (N Y) 2023 May 24;14(6):1-17 [FREE Full text] [doi: 10.1007/s12671-023-02137-0] [Medline: 37362184]
47. Ziadni MS, Anderson SR, Gonzalez-Castro L, Darnall BD. Comparative efficacy of a single-session "Empowered Relief" videoconference-delivered group intervention for chronic pain: study protocol for a randomized controlled trial. Trials 2021 May 22;22(1):358. [doi: 10.1186/s13063-021-05303-8] [Medline: 34022930]
48. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381 [FREE Full text] [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
49. Farrar JT, Young JP, LaMoreaux L, Werth JL, Poole RM. Clinical importance of changes in chronic pain intensity measured on an 11-point numerical pain rating scale. Pain 2001 Nov;94(2):149-158. [doi: 10.1016/S0304-3959(01)00349-9] [Medline: 11690728]
50. Cook KF, Dunn W, Griffith JW, Morrison MT, Tanquary J, Sabata D, et al. Pain assessment using the NIH toolbox. Neurology 2013 Mar 12;80(11_supplement_3):23 [FREE Full text] [doi: 10.1212/wnl.0b013e3182872e80]
51. Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, PROMIS Cooperative Group. The patient-reported outcomes measurement information system (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. J Clin Epidemiol 2010 Nov;63(11):1179-1194 [FREE Full text] [doi: 10.1016/j.jclinepi.2010.04.011] [Medline: 20685078]
52. Cook KF, Bamer AM, Amtmann D, Molton IR, Jensen MP. Six patient-reported outcome measurement information system short form measures have negligible age- or diagnosis-related differential item functioning in individuals with disabilities. Arch Phys Med Rehabil 2012 Jul;93(7):1289-1291. [doi: 10.1016/j.apmr.2011.11.022] [Medline: 22386213]
53. Dworkin RH, Turk DC, Farrar JT, Haythornthwaite JA, Jensen MP, Katz NP, IMMPACT. Core outcome measures for chronic pain clinical trials: IMMPACT recommendations. Pain 2005 Jan;113(1-2):9-19. [doi: 10.1016/j.pain.2004.09.012] [Medline: 15621359]
54. HealthMeasures consists of PROMIS, Neuro-QoL, ASCQ-Me, and NIH toolbox. HealthMeasures. URL: https://www.healthmeasures.net/ [accessed 2024-04-29]
55. Osman A, Barrios FX, Gutierrez PM, Kopper BA, Merrifield T, Grittmann L. The Pain Catastrophizing Scale: further psychometric evaluation with adult samples. J Behav Med 2000 Aug;23(4):351-365. [doi: 10.1023/a:1005548801037] [Medline: 10984864]
56. Opioid oral morphine milligram equivalent (MME) conversion factors table for prescription drug coverage. US Department of Health & Human Services. URL: https://www.hhs.gov/guidance/document/opioid-oral-morphine-milligram-equivalent-mme-conversion-factors-0 [accessed 2024-04-29]
57. Oral morphine milligram equivalent conversion table. State of Ohio Board of Pharmacy. 2017. URL: https://www.ohiopmp.gov/Documents/MorphineEquivalentDailyDoseConversionTable.pdf [accessed 2022-10-30]
58. Ziadni MS, Chen AL, Winslow T, Mackey SC, Darnall BD. Efficacy and mechanisms of a single-session behavioral medicine class among patients with chronic pain taking prescription opioids: study protocol for a randomized controlled trial. Trials 2020 Jun 12;21(1):521 [FREE Full text] [doi: 10.1186/s13063-020-04415-x] [Medline: 32532346]
59. Darnall BD, Sturgeon JA, Cook KF, Taub CJ, Roy A, Burns JW, et al. Development and validation of a daily pain catastrophizing scale. J Pain 2017 Sep;18(9):1139-1149 [FREE Full text] [doi: 10.1016/j.jpain.2017.05.003] [Medline: 28528981]
60. Williamson A, Hoggart B. Pain: a review of three commonly used pain rating scales. J Clin Nurs 2005 Aug 30;14(7):798-804. [doi: 10.1111/j.1365-2702.2005.01121.x] [Medline: 16000093]
61. IBM SPSS statistics for Windows. IBM Corp. 2019. URL: https://www.ibm.com/products/spss-statistics [accessed 2024-04-29]
62. RStudio Team. Integrated development for R. RStudio. URL: http://www.rstudio.com/ [accessed 2022-06-06]
63. Atkins DC, Baldwin SA, Zheng C, Gallop RJ, Neighbors C. A tutorial on count regression and zero-altered count models for longitudinal substance use data. Psychol Addict Behav 2013 Mar;27(1):166-177 [FREE Full text] [doi: 10.1037/a0029508] [Medline: 22905895]
64. Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2nd edition. New York, NY: Routledge; Dec 1989.

XSL•FO
RenderX

65. Reifman A, Keyton K. Winsorize. In: Salkind NJ, editor. Encyclopedia of Research Design. Thousand Oaks, CA: Sage Publications; 2010:1636-1638.
66. Benjamini Y, Hochberg Y. On the adaptive control of the false discovery rate in multiple testing with independent statistics. J Educ Behav Stat 2000 Mar 01;25(1):60-83 [FREE Full text] [doi: 10.3102/10769986025001060]
67. Dowell D, Ragan KR, Jones CM, Baldwin GT, Chou R. CDC clinical practice guideline for prescribing opioids for pain - United States, 2022. MMWR Recomm Rep 2022 Nov 04;71(3):1-95 [FREE Full text] [doi: 10.15585/mmwr.rr7103a1] [Medline: 36327391]
68. Ashton-James CE, Glare P, Darnall BD. Out of office hours: scalable, on-demand, digital support for patients tapering prescription opioids. Pain 2020 Oct 3;161(10):2252-2254. [doi: 10.1097/j.pain.0000000000001947] [Medline: 32501860]
69. Darnall BD, Ziadni MS, Stieg RL, Mackey IG, Kao M, Flood P. Patient-centered prescription opioid tapering in community outpatients with chronic pain. JAMA Intern Med 2018 May 01;178(5):707-708 [FREE Full text] [doi: 10.1001/jamainternmed.2017.8709] [Medline: 29459978]
70. Janevic MR, McLaughlin SJ, Heapy AA, Thacker C, Piette JD. Racial and socioeconomic disparities in disabling chronic pain: findings from the health and retirement study. J Pain 2017 Dec;18(12):1459-1467 [FREE Full text] [doi: 10.1016/j.jpain.2017.07.005] [Medline: 28760648]
71. Stone AA, Obbarius A, Junghaenel DU, Wen CK, Schneider S. High-resolution, field approaches for assessing pain: ecological momentary assessment. Pain 2021 Jan;162(1):4-9 [FREE Full text] [doi: 10.1097/j.pain.0000000000002049] [Medline: 32833794]
72. Mun CJ, Finan PH, Epstein DH, Kowalczyk WJ, Agage D, Letzen JE, et al. Craving mediates the association between momentary pain and illicit opioid use during treatment for opioid-use disorder: an ecological momentary assessment study. Addiction 2021 Jul;116(7):1794-1804 [FREE Full text] [doi: 10.1111/add.15344] [Medline: 33220102]

Abbreviations

BH-FDR: Benjamini-Hochberg false discovery rate
CBT: cognitive behavioral therapy
CONSORT: Consolidated Standards of Reporting Trials
ER: Empowered Relief
MEDD: morphine-equivalent daily dose
MORE: Mindfulness-Orientated Recovery Enhancement
PCS: Pain Catastrophizing Scale
PRN: pro re nata
PROMIS: Patient Reported Outcome Measurement Information System
RCT: randomized controlled trial
REDCap: Research Electronic Data Capture

Edited by A Mavragani; submitted 07.11.24; peer-reviewed by MA Bushey, H Rafi; comments to author 16.12.24; revised version received 06.01.25; accepted 30.01.25; published 11.03.25.

Please cite as:
Edwards KA, Dildine TC, You DS, Herrick AM, Darnall BD, Mackey SC, Ziadni MS
Zoom-Delivered Empowered Relief for Chronic Pain: Observational Longitudinal Pilot Study Exploring Feasibility and Pain-Related Outcomes in Patients on Long-Term Opioids
JMIR Form Res 2025;9:e68292
URL: https://formative.jmir.org/2025/1/e68292
doi:10.2196/68292
PMID:

©Karlyn A Edwards, Troy C Dildine, Dokyoung S You, Ashley M Herrick, Beth D Darnall, Sean C Mackey, Maisa S Ziadni. Originally published in JMIR Formative Research (https://formative.jmir.org), 11.03.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Digital Assessment of Cognitive Health in Outpatient Primary Care: Usability Study

Adam J Doerr[1*], MD; Taylor A Orwig[1*], MS; Matthew McNulty[1], BS; Stephanie Denise M Sison[2], MD; David R Paquette[3], PhD; Robert Leung[1], MD; Huitong Ding[4], PhD; Stephen B Erban[5], MD; Bruce R Weinstein[2], MD; Yurima Guilarte-Walker[6], MS; Adrian H Zai[6], MD, PhD, MPH; Allan J Walkey[5], MD; Apurv Soni[1,5], MD, PhD; David D McManus[1,5], MD; Honghuang Lin[1,5], PhD

[1]Program in Digital Medicine, Department of Medicine, University of Massachusetts Chan Medical School, Worcester, MA, United States

[2]Division of General Internal Medicine, Department of Medicine, University of Massachusetts Chan Medical School, Worcester, MA, United States

[3]Massachusetts AI and Technology Center, Manning College of Information and Computer Sciences, University of Massachusetts Amherst, Amherst, MA, United States

[4]Department of Anatomy and Neurobiology, Chobanian & Avedisian School of Medicine, Boston University, Boston, MA, United States

[5]Division of Health System Science, Department of Medicine, University of Massachusetts Chan Medical School, Worcester, MA, United States

[6]Division of Health Informatics and Implementation Science, Department of Population and Quantitative Health Sciences, University of Massachusetts Chan Medical School, Worcester, MA, United States

[*]these authors contributed equally

Corresponding Author:
Honghuang Lin, PhD
Program in Digital Medicine
Department of Medicine
University of Massachusetts Chan Medical School
S6-755
55 Lake Avenue North
Worcester, MA, 01655
United States
Phone: 1 7744554881
Email: honghuang.lin@umassmed.edu

Abstract

Background: Screening for cognitive impairment in primary care is important, yet primary care physicians (PCPs) report conducting routine cognitive assessments for less than half of patients older than 60 years of age. Linus Health's Core Cognitive Evaluation (CCE), a tablet-based digital cognitive assessment, has been used for the detection of cognitive impairment, but its application in primary care is not yet studied.

Objective: This study aimed to explore the integration of CCE implementation in a primary care setting.

Methods: A cohort of participants was recruited from the upcoming schedules of participating PCPs at UMass Memorial Medical Center. Eligibility criteria included individuals aged ≥65 years; ability to read, write, and speak in English or Spanish; no previous diagnosis of cognitive impairment; and no known untreated hearing or vision impairment. Research coordinators collected consent from participants and facilitated the screening process. PCPs reviewed reports in real time, immediately before the scheduled visits, and shared results at their discretion. A report was uploaded to each participant's REDCap (Research Electronic Data Capture; Vanderbilt University) record and linked to the encounter in the electronic health record. Feedback from patients and their caregivers (if applicable) was collected by a tablet-based survey in the clinic before and after screening. Participating PCPs were interviewed following the completion of the study.

Results: The screened cohort included 150 patients with a mean age of 74 (SD 7) years, of whom 65% (97/150) were female. The CCE identified 40 patients as borderline and 7 as positive for cognitive impairment. A total of 84 orders were placed for select laboratory tests or referrals to neurology and neuropsychology within 20 days of CCE administration. Before the assessment, 95% (143/150) of patients and all 15 caregivers expressed a desire to know if their or their loved one's brain health was declining. All except one patient also completed the postassessment survey. Among them, 96% (143/149) of patients reported finding the CCE easy to complete, and 70% (105/149) felt that the experience was beneficial. In addition, 87% (130/149) of patients agreed

XSL•FO
RenderX

or strongly agreed that they wanted to know their CCE results. Among the 7 participating PCPs, 6 stated that the CCE results influenced their patient care management, and all 7 indicated they would continue using the CCE if it were made available after the study.

Conclusions: We explored the integration of the CCE into primary care visits, which showed minimal disruption to the practice workflow. Future studies will be warranted to further validate the implementation of digital cognitive impairment screening tools within primary care settings in the real world.

(JMIR Form Res 2025;9:e66695) doi:10.2196/66695

KEYWORDS

cognitive assessment; primary care; digital; cognitive impairment; digital assessment; assessment; cognitive health; cognition; primary care; cognitive evaluation; Core Cognitive Evaluation; CCE; cohort; impairment; cognitive; outpatient

Introduction

Dementia is a neurodegenerative disorder characterized by progressive deficits in one or more cognitive domains severe enough to interfere with an individual's functional status [1]. Mild cognitive impairment (MCI) is a precursor of dementia, characterized by more subtle cognitive deficits that do not interfere with functional status [1]. Although definitional variability and reporting bias complicate the estimation of prevalence [2], MCI likely afflicts 12% to 18% of people aged 60 years and older [3,4]. The rate of progression from MCI to dementia is estimated to be 10% to 15% yearly [5].

Only about 18% of Americans are familiar with MCI [1]. However, when MCI is described, 85% report they would want to know if they had either MCI or early dementia, with reasons cited including to plan for the future, to enable earlier treatment of symptoms, to take steps to prevent further deterioration, and to better understand their condition [1]. Similar sentiments are held by family members, for whom early detection enables a more gradual adaptation to the caregiver role, resulting in an improved sense of competence and less psychological distress [6,7]. On a broader level, early detection can delay institutionalization and inform the allocation of human resources within a health care system already marred by shortages in the dementia care workforce [8-10].

Most primary care physicians (PCPs) believe screening for MCI and dementia in older patients is important [1]. Following the detection of MCI or early dementia, PCPs commonly intervene by assessing for reversible causes, recommending lifestyle changes, and referring to specialists [1]. However, most PCPs feel uncomfortable in this domain and report conducting cognitive assessments for less than half of patients older than 60 years of age [1]. Reported barriers to screening in primary care settings include low provider confidence (particularly in distinguishing MCI from normal aging), time constraints, competing priorities, poorly defined protocols for screening, and lack of financial incentives [1,11,12]. Only 40% of PCPs report being familiar with existing cerebrospinal fluid and imaging biomarker tests that can aid in the diagnosis of dementia, while only 18% of patients are referred for such testing when MCI is detected [1,11,12]. An unfortunate consequence is that more than 90% of patients with MCI are not identified [13], resulting in delayed diagnosis when advanced stages of dementia have developed, disability is higher, treatment options are limited [14], and caregiver stress is high [15].

Alzheimer's disease is the most common neurodegenerative cause of MCI and dementia in individuals aged ≥65 years. Biomarkers for AD span multiple modalities, including molecular profiling of bodily fluids, brain imaging, electroencephalography, and neuropsychological testing [16]. With increased computational power, particularly in the form of artificial intelligence, digital biomarkers have emerged as promising new noninvasive approaches for early detection of MCI and early dementia due to AD [17]. One example is the digitization of the traditional clock drawing test (CDT). Although thoroughly validated, the CDT is a classically analog test that requires manual supervision and skilled interpretation, limiting its use in busy clinical practice and making it susceptible to interobserver variability [18]. Linus Health's Core Cognitive Evaluation (CCE) is a tablet-based cognitive assessment tool consisting of an immediate recall task, a digital CDT administered on an Apple iPad using an Apple Pencil, and a delayed recall task (collectively termed Digital Clock and Recall [DCR]) [19,20]. Performance is analyzed by a machine learning algorithm to yield a score (ranging from 0 to 5), which is indicative of cognitive impairment (0-1), borderline for cognitive impairment (2-3), or not indicative of cognitive impairment (4-5). The DCR is paired with a Life and Health Questionnaire (LHQ), which screens for modifiable lifestyle and psychosocial risk and protective factors related to brain health, enabling prediction of future dementia risk and personalized recommendations for patients. The DCR has demonstrated greater sensitivity and accuracy than existing cognitive assessment tests such as the Mini-Mental State Examination (MMSE) and the Mini-Cog [21] and takes substantially less time to complete (approximately 3 min).

These qualities make the CCE well suited for the busy primary care setting. However, the impact of the CCE in primary care with respect to patient and provider satisfaction, downstream clinical decision-making, and reimbursement remains unclear. The objective of this study was to evaluate the feasibility and impact of screening for cognitive impairment in an outpatient, primary care setting using the Linus Health CCE.

XSL•FO
RenderX

Methods

Screening and Recruitment

Participants for this study were recruited from the Benedict Adult Primary Care Clinic at UMass Memorial Medical Center. Eligible patients were identified through a list of upcoming visits generated from the electronic health record system (Epic Systems). Eligibility criteria included being aged 65 years or older; having the ability to read, write, and speak in English or Spanish; having no previous diagnosis of cognitive impairment; and having no untreated hearing or vision impairment that would hinder audible conversation or the use of a tablet. Recruitment letters were mailed to eligible patients, inviting them to participate in the study. These letters included an opt-out phone number and email address. If no response was received, a research coordinator followed up with a phone call 3-5 days after the letter was sent to assess interest. Patients who expressed interest received another phone call one day before their scheduled visit to confirm their participation, remind them to arrive an hour early, and advise them to bring their reading glasses or hearing aids, if applicable. All contacted patients were logged in a recruitment tracker, which recorded details such as patient name, participation status, PCP name, recruitment letter mail date, follow-up call date, visit date and type, and participation reminder call date.

Assessment

Participants were greeted in the clinic by research coordinators, escorted to a quiet and private room without an analog clock present, and completed the consent process. They then completed preassessment surveys on study tablets. If the patients were accompanied by their caregivers, their caregivers were also provided with a fact sheet and completed a survey. Simultaneously, research coordinators registered participants on a separate tablet equipped with the Linus Health CCE. Participants then completed the CCE on a study tablet. Participants were first administered the DCR, in the following sequence: (1) during the immediate recall condition, 3 words were voiced by the tablet, and the patient was asked to repeat them immediately; (2) during the command clock condition, the tablet voiced the following command: “draw the face of a clock, put in all of the numbers, and set the hands to ‘10 after 11’”; (3) during the copy clock condition, the tablet displayed a picture of an analog clock set to 11:10 and asked the patient to copy it; and (4) during the delayed recall condition, the patient was asked to repeat the 3 words initially voiced by the tablet. The DCR was not adjusted for demographics or education similar to the original Mini-Cog score. Patients then completed the LHQ and a postassessment survey on the study tablet before their visit with the PCP. A total of 7 PCPs who consented to participate were identified and given a fact sheet. They went through a training session before the study. After assessments were completed for the entire cohort, all PCPs involved in the study completed a provider feedback survey sent by email and completed an interview assessing their impressions of the CCE.

Results Sharing

After completion of the CCE, results were immediately uploaded into the Linus Health portal. This generated a provider-facing report consisting of DCR scores; interpretations; playbacks of clock drawing and verbal recall; clinical decision support; and a personalized patient-facing Brain Health Action Plan, which identifies modifiable risk factors for cognitive impairment and provides evidence-based suggestions for intervention. These results were printed and given to PCPs immediately before patient visits for review in real time. DCR results were shared with patients at the discretion of PCPs, while the Brain Health Action Plans were printed and universally shared with each patient.

Data Management and Analysis

Study data were collected and managed using REDCap (Research Electronic Data Capture; Vanderbilt University) tools hosted at UMass Chan Medical School [22,23]. Informed consent was also obtained using REDCap within the eConsent framework on a study tablet [24]. REDCap is a secure, web-based software platform designed to support data capture for research studies, providing (1) an intuitive interface for validated data capture; (2) audit trails for tracking data manipulation and export procedures; (3) automated export procedures for seamless data downloads to common statistical packages; and (4) procedures for data integration and interoperability with external sources. Identifiable patient health information was concealed from Linus Health throughout the study and thereafter.

Ethical Considerations

The study protocol was approved by the Western-Copernicus Group Institutional Review Board (approval number 1341324), and informed consent was obtained from all patients. Caregivers and PCPs were provided with a fact sheet to guide them in completing the survey. No compensation for participation was provided. For data analysis, only deidentified data were exported from REDCap to ensure confidentiality and compliance with ethical standards.

Results

Screened Cohort

A flowchart of participant recruitment is shown in Figure S1 in Multimedia Appendix 1. During the screening process, 1469 patients were identified as meeting the eligibility criteria. Among them, 160 accepted the invitation to participate. Of these, 9 either did not show up, changed their minds, or canceled their appointments, and one participant was found ineligible to provide consent. Ultimately, 150 patients from the panels of 7 PCPs were successfully enrolled and completed the CCE between March and December 2023. The baseline demographic and clinical characteristics of studied patients are listed in Table 1. The participants had a mean age of 74 (SD 7) years; 65% (97/150) were female, and 74% (111/150) held a college degree or higher. The CCE identified 40 patients as borderline and 7 patients as positive for cognitive impairment. The number of patients recruited from each PCP panel ranged from 9 to 38.

XSL•FO
RenderX

Table 1. Patient demographic and clinical characteristics. Mean (SD) values for continuous variables, and counts (%) for categorical variables have been reported.

Characteristics	Values (n=150)
Age (years), mean (SD)	74 (7)
Female, n (%)	97 (65)
Race, n (%)	
White	147 (98)
Black	2 (1)
Not reported or unknown	1 (1)
Ethnicity, n (%)	
Hispanic or Latino	2 (1)
Not Hispanic or Latino	136 (91)
Not recorded or unknown	12 (8)
Education, n (%)	
High school	11 (7)
Some college	28 (19)
College graduate and above	111 (74)
Hypertension, n (%)	111 (74)
Diabetes mellitus, n (%)	32 (21)
Hyperlipidemia, n (%)	135 (90)
Smoking, n (%)	35 (23)
Ischemic heart disease, n (%)	27 (18)
Atrial fibrillation or flutter, n (%)	24 (16)
Heart failure, n (%)	10 (7)
Peripheral vascular disease, n (%)	7 (5)
Obstructive sleep apnea, n (%)	62 (41)
Parkinson disease, n (%)	1 (1)
Depression, n (%)	37 (25)
Stroke, n (%)	9 (6)
Transient ischemic attack, n (%)	10 (7)
Epilepsy, n (%)	4 (3)
Hearing loss, n (%)	66 (44)
B_{12} deficiency, n (%)	3 (2)

Provider Response to the CCE

Among the 7 PCPs participating in the study, 5 had more than 10 years of clinical practice, one had 5-10 years of clinical practice, and one was a junior clinician. All PCPs stated that it was important to detect undiagnosed cognitive impairment, and most noted time or workflow constraints as a barrier to cognitive testing. All PCPs reviewed test results before meeting with their patient, with an average reported review time of 2 minutes, and more time is needed if the results indicated a cognitive issue. A majority of the PCPs stated that the CCE did not interfere with patient visits (5/7, 71%) and that it did not lead to any stress or challenges for their patient (6/7, 86%). Most PCPs (4/7, 57%) reported that the CCE did not cause them any stresses or challenges; among the 3 providers who answered affirmatively, one discussed the stress of breaking the news of poor CCE scores to patients, another cited the challenge of finding the time to discuss results, and the last voiced concerns about false positive scores that might result from extraneous factors like anxiety, comorbidities, and language barriers. All PCPs stated that they generally agreed with the CCE results, and a majority (6/7, 86%) of PCPs stated that the results of the CCE changed their patient care management. All PCPs reported they would continue using the CCE if made available following the study.

XSL•FO
RenderX

Downstream Clinical Action Following the CCE

Referrals to neurology and neuropsychology, select laboratory tests possibly related to the workup of dementia, and neuroimaging studies ordered after administration of the CCE are reported in Table 2. There were 84 such orders placed within 20 days of CCE administration, and an additional 58 orders were placed within the subsequent 160 days thereafter. Moreover, among the 7 patients with DCR scores of 0 and 1, a total of 4 (57%) were later diagnosed with cognitive impairment. In contrast, only 16 (11%) of the 143 patients with scores of 2 or above were later diagnosed with cognitive impairment.

Table 2. Near-term and delayed cognition-relevant orders placed following the administration of the Core Cognitive Evaluation (CCE).

Cognition-relevant orders	Days since the CCE	
	0-20	21-180
Referral, n		
Neurology	3	2
Neuropsychology	0	1
Imaging test, n		
CT[a] head	2	5
MRI[b] brain	4	11
PET[c] brain	0	2
Laboratory test, n		
B_{12}	26	7
Folate	11	4
Rapid plasma reagin	3	0
Thyroid-stimulating hormone	35	26
Total tests, n	84	58

[a]CT: computed tomography.

[b]MRI: magnetic resonance imaging.

[c]PET: positron emission tomography.

Patient and Caregiver Perspectives on Brain Health Prior to the CCE

The results of the preassessment survey for patients and caregivers are displayed in Table 3, and the stratified results are shown in Table S1 in Multimedia Appendix 1. Before the screening, 58% (87/150) of patients agreed or strongly agreed they worry about their brain health, while 95% (143/150) of patients and all caregivers reported they would want to know if their or their loved one's brain health was declining. Approximately two-thirds of both patients (100/150, 67%) and caregivers (10/15, 67%) agreed or strongly agreed their plans for the next 5-10 years would change in response to a decline in brain health; "How we spend our time" was the most frequently selected decision that would be impacted. About two-thirds (95/150, 63%) of patients reported they go to their PCPs for help with taking care of their brain health.

XSL•FO
RenderX

Table 3. Patient and caregiver perspectives on brain health before the administration of the Core Cognitive Evaluation.

Survey questions	Strongly agree, n (%)	Agree, n (%)	Neither agree nor disagree, n (%)	Disagree, n (%)	Strongly disagree, n (%)
Patients (n=150)					
I worry about my memory or brain health.	15 (10)	72 (48)	28 (19)	17 (11)	18 (12)
I understand how to take care of my brain health.	48 (32)	69 (46)	29 (19)	3 (2)	1 (1)
I go to my primary care physician to help me take care of my brain health.	45 (30)	50 (33)	43 (29)	9 (6)	3 (2)
If my brain health was declining, I would want to know.	81 (54)	62 (41)	3 (2)	1 (1)	3 (2)
If I found out that my brain health was declining, it would change my plans for the next 5-10 years.	37 (25)	63 (42)	28 (19)	14 (9)	8 (5)
Caregivers (n=15)					
I worry about my relative's memory or brain health.	3 (20)	5 (33)	4 (27)	2 (13)	1 (7)
I understand how to help my relatives take care of their brain health.	5 (33)	4 (27)	6 (40)	0 (0)	0 (0)
I would want to take my relative to their primary care physician to help me take care of their brain health.	7 (47)	6 (40)	1 (7)	1 (7)	0 (0)
If my relative's brain health was declining, I would want to know.	8 (53)	7 (47)	0 (0)	0 (0)	0 (0)
If I knew my relative's brain health was declining, it would change our plans for the next 5-10 years.	4 (27)	6 (40)	4 (27)	0 (0)	1 (7)

Patient Satisfaction With the Core Cognitive Evaluation

The results of the postassessment survey for patients are displayed in Table 4, and the stratified results are shown in Table S2 in Multimedia Appendix 1. One patient did not complete the assessment. Among participants who completed the postassessment survey, 96% (143/149) of patients reported being able to easily complete the CCE, while 70% (105/149) reported that completing the CCE was beneficial. Among those who felt the CCE was beneficial, all felt it enabled them to be productive about their brain health, about two-thirds (68/105, 65%) felt it gave them more control over their long-term brain health, and about one-third (37/105, 35%) felt it addressed a current specific concern about their brain health. As a result of taking the CCE, only one patient reported feeling sad or depressed, while 5% (8/149) reported feeling worried or anxious. Most patients (130/149, 87%) agreed or strongly agreed they would like to know their CCE results.

Table 4. Patient satisfaction after the administration of the Core Cognitive Evaluation. One participant did not complete the postassessment survey (n=149).

Survey questions	Strongly agree, n (%)	Agree, n (%)	Neither agree nor disagree, n (%)	Disagree, n (%)	Strongly disagree, n (%)
I was able to complete this assessment easily	107 (71)	36 (24)	3 (2)	0 (0)	3 (2)
Taking this assessment was beneficial	46 (31)	59 (39)	35 (23)	6 (4)	3 (2)
Taking this assessment was stressful	0 (0)	8 (5)	16 (11)	44 (29)	81 (54)
This assessment made me feel anxious or worried	0 (0)	8 (5)	12 (8)	39 (26)	90 (60)
This assessment made me feel sad or depressed	0 (0)	1 (1)	6 (4)	32 (21)	110 (73)
I was to know my results from this assessment	84 (56)	46 (31)	8 (5)	3 (2)	8 (5)

Discussion

Principal Findings

Cognitive impairment is a formidable public health threat, with significant burdens posed to individual patients, their caregivers, the health care system, and the economy. With two disease-modifying therapies now approved by the US Food and Drug Administration (FDA) [25,26] and commercially available and covered by the Centers for Medicare and Medicaid Services (CMS) [27], early detection of cognitive impairment has assumed new relevance. In this study, a cohort of older outpatients without a previous diagnosis of cognitive impairment was screened in a primary care setting using the CCE.

The preassessment survey overall comported with the existing body of literature suggesting patients are concerned about their brain health and would strongly prefer to detect cognitive impairment as early as possible. Overall, participant satisfaction with the screening process was high, with a low incidence of patient-reported negative psychological impact. A majority of PCPs stated that the CCE results changed their care management, and all PCPs stated they would continue using the CCE if made available following the study.

XSL•FO
RenderX

The protocol used in this study prioritized minimizing disruption in an already busy clinical environment by aligning cognitive assessments with patients' scheduled appointments. Central to this workflow was a dedicated team member who facilitated consenting, screening, and delivery of results to providers. Although this role was filled by a research coordinator in the study, it could feasibly be assigned to clinical staff, such as medical assistants, in real-world practice given the simplicity of the screening platform and the minimal training required to administer the survey. Meanwhile, the timing of assessments before visits and the provision of concise results with easily digestible interpretations to PCPs obviate the need for manual grading and interpretation, reducing both the time burden and expertise requirements associated with conventional assessments. This design therefore serves as a pragmatic framework for the implementation of a digital cognitive impairment screening program in a real-world primary care setting.

There are noteworthy limitations inherent to this workflow and its evaluation. While potentially more streamlined for PCPs, the study required patients to arrive early for their scheduled appointments. This represents an added medical burden for the older population and may compete with other previsit agenda items. This extra time requirement was necessary to ensure that cognitive assessments could be completed without disrupting the scheduled clinical visits, allowing PCPs to review the results in real time. However, it might also impact a spouse or family member helping the patient to spend more time onsite in the office. This burden could be mitigated in future studies with improved electronic health record integration and automated clinical decision support implementation, or by exploring alternative workflows, such as integrating the CCE into regular appointment times or offering remote assessments. Furthermore, while patients were surveyed before their scheduled visits and were blind to their CCE results, PCPs were surveyed following enrollment completion, introducing the potential for recall bias in their evaluation of the screening process. This retrospective feedback collection allowed PCPs to reflect on their experiences with multiple patients but may have affected the accuracy of their assessments. In addition, the study's design, which involved a convenience sample from a single primary care clinic, limits the generalizability of the findings. While this approach facilitated close monitoring of the implementation process, it may not fully represent the broader population. Future studies should expand the sample size and include multiple sites across diverse geographic and health care settings. In addition, most of the participants in this study are non-Hispanic White, which greatly limits its generalizability to other racial and ethnic populations. More efforts have to be made to increase the diversity of participants in the future [28]. Finally, irrespective of feasibility, the efficacy of a widely adopted digital cognitive screening program using the CCE remains speculative until tested in a randomized trial. It would be essential to further investigate strategies for sustainable and scalable integration of the CCE in diverse clinical environments, including enhanced electronic health record integration.

Conclusion

In summary, we explored the integration of the CCE into primary care visits, which showed minimal disruption to the practice workflow. Future studies will be warranted to further validate the implementation of digital cognitive impairment screening tools within primary care settings in the real world.

Acknowledgments

We thank patients and their families, whose help and participation made this work possible. The investigators' time was partly supported by grants from P30AG073107, U01AG068221, R01AG080670, R01AG083735, R21HL175584, R01HL155343, U01HL146382, U54HL143541, and R33HL158541. The views expressed in this manuscript are those of the authors and do not necessarily represent the views of the National Institutes of Health or the US Department of Health and Human Services. The funding agencies had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

DDM has received consultancy fees from Heart Rhythm Society, Fitbit, Flexcon, Pfizer, Avania, NAMSA, and Bristol Myers Squibb. DDM reports receiving research support from Fitbit, Apple, Care Evolution, Boeringher Ingelheim, Pfizer, and Bristol Myers Squibb. The authors have no further interests to declare.

Multimedia Appendix 1
Flowchart illustrating the participant screening and enrollment, patient perspectives on brain health prior to the administration of the CCE stratified by DCR score groups, and patient satisfaction after the administration of the CCE stratified by DCR score groups. CCE: Core Cognitive Evaluation; DCR: Digital Clock and Recall.
[DOCX File , 53 KB - formative_v9i1e66695_app1.docx]

References

XSL•FO
RenderX

1. Alzheimer's Association. 2022 Alzheimer's disease facts and figures. Alzheimers Dement 2022 Apr;18(4):700-789. [doi: 10.1002/alz.12638] [Medline: 35289055]
2. Casagrande M, Marselli G, Agostini F, Forte G, Favieri F, Guarino A. The complex burden of determining prevalence rates of mild cognitive impairment: a systematic review. Front Psychiatry 2022;13:960648 [FREE Full text] [doi: 10.3389/fpsyt.2022.960648] [Medline: 36213927]
3. Petersen RC, Lopez O, Armstrong MJ, Getchius TSD, Ganguli M, Gloss D, et al. Practice guideline update summary: mild cognitive impairment: report of the guideline development, dissemination, and implementation subcommittee of the American Academy of Neurology. Neurology 2018;90(3):126-135 [FREE Full text] [doi: 10.1212/WNL.0000000000004826] [Medline: 29282327]
4. Rajan KB, Weuve J, Barnes LL, McAninch EA, Wilson RS, Evans DA. Population estimate of people with clinical Alzheimer's disease and mild cognitive impairment in the United States (2020-2060). Alzheimers Dement 2021;17(12):1966-1975 [FREE Full text] [doi: 10.1002/alz.12362] [Medline: 34043283]
5. Mitchell AJ, Shiri-Feshki M. Rate of progression of mild cognitive impairment to dementia--meta-analysis of 41 robust inception cohort studies. Acta Psychiatr Scand 2009;119(4):252-265. [doi: 10.1111/j.1600-0447.2008.01326.x] [Medline: 19236314]
6. de Vugt ME, Verhey FRJ. The impact of early dementia diagnosis and intervention on informal caregivers. Prog Neurobiol 2013;110:54-62. [doi: 10.1016/j.pneurobio.2013.04.005] [Medline: 23689068]
7. de Vugt ME, Stevens F, Aalten P, Lousberg R, Jaspers N, Winkens I, et al. Do caregiver management strategies influence patient behaviour in dementia? Int J Geriatr Psychiatry 2004;19(1):85-92. [doi: 10.1002/gps.1044] [Medline: 14716704]
8. National Academies of Sciences, Division of Behavioral and Social Sciences and Education, Board on Behavioral, Cognitive, and Sensory Sciences, Committee on the Decadal Survey of Behavioral and Social Science Research on Alzheimer's Disease and Alzheimer's Disease-Related Dementias. Reducing the Impact of Dementia in America: A Decadal Survey of the Behavioral and Social Sciences. Washington, DC: National Academies Press; 2021.
9. Banerjee S, Wittenberg R. Clinical and cost effectiveness of services for early diagnosis and intervention in dementia. Int J Geriatr Psychiatry 2009;24(7):748-754. [doi: 10.1002/gps.2191] [Medline: 19206079]
10. Dubois B, Padovani A, Scheltens P, Rossi A, Dell'Agnello G. Timely diagnosis for Alzheimer's Disease: a literature review on benefits and challenges. J Alzheimers Dis 2016;49(3):617-631 [FREE Full text] [doi: 10.3233/JAD-150692] [Medline: 26484931]
11. Gong N, Yang D, Zou J, He Q, Hu L, Chen W, et al. Exploring barriers to dementia screening and management services by general practitioners in China: a qualitative study using the COM-B model. BMC Geriatr 2023;23(1):55 [FREE Full text] [doi: 10.1186/s12877-023-03756-x] [Medline: 36717818]
12. Bernstein Sideman A, Al-Rousan T, Tsoy E, Piña Escudero SD, Pintado-Caipa M, Kanjanapong S, et al. Facilitators and barriers to dementia assessment and diagnosis: perspectives from dementia experts within a global health context. Front Neurol 2022;13:769360 [FREE Full text] [doi: 10.3389/fneur.2022.769360] [Medline: 35418934]
13. Mattke S, Jun H, Chen E, Liu Y, Becker A, Wallick C. Expected and diagnosed rates of mild cognitive impairment and dementia in the U.S. medicare population: observational analysis. Alzheimers Res Ther 2023;15(1):128 [FREE Full text] [doi: 10.1186/s13195-023-01272-z] [Medline: 37481563]
14. Mattke S, Jun H, Chen E, Liu Y, Becker A, Wallick C. Expected and diagnosed rates of mild cognitive impairment and dementia in the U.S. medicare population: observational analysis. Alzheimers Res Ther 2023;15(1):128 [FREE Full text] [doi: 10.1186/s13195-023-01272-z] [Medline: 37481563]
15. Richardson TJ, Lee SJ, Berg-Weger M, Grossberg GT. Caregiver health: health of caregivers of Alzheimer's and other dementia patients. Curr Psychiatry Rep 2013;15(7):367. [doi: 10.1007/s11920-013-0367-2] [Medline: 23712718]
16. Gunes S, Aizawa Y, Sugashi T, Sugimoto M, Rodrigues PP. Biomarkers for Alzheimer's disease in the current state: a narrative review. Int J Mol Sci 2022;23(9):4962 [FREE Full text] [doi: 10.3390/ijms23094962] [Medline: 35563350]
17. Vrahatis AG, Skolariki K, Krokidis MG, Lazaros K, Exarchos TP, Vlamos P. Revolutionizing the early detection of Alzheimer's disease through non-invasive biomarkers: the role of artificial intelligence and deep learning. Sensors (Basel) 2023;23(9):4184 [FREE Full text] [doi: 10.3390/s23094184] [Medline: 37177386]
18. Agrell B, Dehlin O. The clock-drawing test. 1998. Age Ageing 2012;41 Suppl 3:iii41-iii45. [doi: 10.1093/ageing/afs149] [Medline: 23144287]
19. Jannati A, Toro-Serey C, Gomes-Osman J, Banks R, Ciesla M, Showalter J, et al. Digital clock and recall is superior to the mini-mental state examination for the detection of mild cognitive impairment and mild dementia. Alzheimers Res Ther 2024;16(1):2 [FREE Full text] [doi: 10.1186/s13195-023-01367-7] [Medline: 38167251]
20. Banks R, Higgins C, Greene BR, Jannati A, Gomes-Osman J, Tobyne S, et al. Clinical classification of memory and cognitive impairment with multimodal digital biomarkers. Alzheimers Dement (Amst) 2024;16(1):e12557 [FREE Full text] [doi: 10.1002/dad2.12557] [Medline: 38406610]
21. Gomes-Osman J, Borson S, Toro-Serey C, Banks R, Ciesla M, Jannati A, et al. Digital clock and recall: a digital, process-driven evolution of the mini-cog. Front Hum Neurosci 2024;18:1337851 [FREE Full text] [doi: 10.3389/fnhum.2024.1337851] [Medline: 39253069]

XSL•FO
RenderX

22. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009;42(2):377-381 [FREE Full text] [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
23. Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O'Neal L, et al. The REDCap consortium: building an international community of software platform partners. J Biomed Inform 2019;95:103208 [FREE Full text] [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]
24. Lawrence CE, Dunkel L, McEver M, Israel T, Taylor R, Chiriboga G, et al. A REDCap-based model for electronic consent (eConsent): moving toward a more personalized consent. J Clin Transl Sci 2020;4(4):345-353 [FREE Full text] [doi: 10.1017/cts.2020.30] [Medline: 33244416]
25. FDA converts novel Alzheimer's disease treatment to traditional approval. U.S. Food and Drug Administration. 2023. URL: https://www.fda.gov/news-events/press-announcements/fda-converts-novel-alzheimers-disease-treatment-traditional-approval [accessed 2025-01-14]
26. FDA approves treatment for adults with Alzheimer's disease. U.S. Food and Drug Administration. 2024. URL: https://www.fda.gov/drugs/news-events-human-drugs/fda-approves-treatment-adults-alzheimers-disease [accessed 2025-01-14]
27. Statement: broader medicare coverage of leqembi available following FDA traditional approval. Centers for Medicare and Medicaid Services. 2023. URL: https://www.cms.gov/newsroom/press-releases/statement-broader-medicare-coverage-leqembi-available-following-fda-traditional-approval [accessed 2025-01-14]
28. Fisher JA, Kalbaugh CA. Challenging assumptions about minority participation in US clinical research. Am J Public Health 2011;101(12):2217-2222. [doi: 10.2105/AJPH.2011.300279] [Medline: 22021285]

Abbreviations

CCE: Core Cognitive Evaluation
CDT: clock drawing test
CMS: Centers for Medicare and Medicaid Services
DCR: Digital Clock and Recall
FDA: US Food and Drug Administration
LHQ: Life and Health Questionnaire
MCI: mild cognitive impairment
MMSE: Mini-Mental State Examination
PCP: primary care physicians
REDCap: Research Electronic Data Capture

Edited by A Mavragani; submitted 20.09.24; peer-reviewed by K Possin; comments to author 29.10.24; revised version received 08.12.24; accepted 27.12.24; published 12.03.25.

Please cite as:
Doerr AJ, Orwig TA, McNulty M, Sison SDM, Paquette DR, Leung R, Ding H, Erban SB, Weinstein BR, Guilarte-Walker Y, Zai AH, Walkey AJ, Soni A, McManus DD, Lin H
Digital Assessment of Cognitive Health in Outpatient Primary Care: Usability Study
JMIR Form Res 2025;9:e66695
URL: https://formative.jmir.org/2025/1/e66695
doi:10.2196/66695
PMID:

©Adam J Doerr, Taylor A Orwig, Matthew McNulty, Stephanie Denise M Sison, David R Paquette, Robert Leung, Huitong Ding, Stephen B Erban, Bruce R Weinstein, Yurima Guilarte-Walker, Adrian H Zai, Allan J Walkey, Apurv Soni, David D McManus, Honghuang Lin. Originally published in JMIR Formative Research (https://formative.jmir.org), 12.03.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Teen Perspectives on Integrating Digital Mental Health Programs for Teens Into Public Libraries ("I Was Always at the Library"): Qualitative Interview Study

Ashley A Knapp[1], PhD; Katherine Cohen[2], MA; Kaylee P Kruzan[3], LCSW, PhD; Rachel Kornfield[3], PhD; Miguel Herrera[1], MEd; Aderonke B Pederson[4], MD; Sydney Lee[1]; Kathryn Macapagal[2], PhD; Chantelle A Roulston[2], BA; Kaleigh Clarke[5], BS; Clarisa Wijaya[6], MA; Robert Simmons[7], MSW; Latonia Jackson[7], BA; Simrandeep Kour[8], MA; Sandra Franco[9], MS, MEd; David C Mohr[3], PhD

[1]Department of Psychiatry and Behavioral Sciences, Feinberg School of Medicine, Northwestern University, Chicago, IL, United States
[2]Department of Medical Social Sciences, Feinberg School of Medicine, Northwestern University, Chicago, IL, United States
[3]Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University, Chicago, IL, United States
[4]Department of Psychiatry, Harvard Medical School, Massachusetts General Hospital, Boston, MA, United States
[5]Bradley University, Peoria, IL, United States
[6]Department of Psychology and Psychiatry, University of Illinois Chicago, Chicago, IL, United States
[7]Oak Park Public Library, Oak Park, IL, United States
[8]Chitkara University, Chandigarh, India
[9]Chicago School of Education, Loyola University, Chicago, IL, United States

Corresponding Author:
Ashley A Knapp, PhD
Department of Psychiatry and Behavioral Sciences
Feinberg School of Medicine
Northwestern University
750 North Lake Shore Drive
Chicago, IL, 60611
United States
Phone: 1 3125033751
Email: ashley.knapp@northwestern.edu

Abstract

Background: Rising rates of anxiety among teens necessitate innovative approaches for implementing evidence-based mental health support. Public libraries, seen as safe spaces for patrons with marginalized identities, offer free public services such as broadband internet access. Many teens spend significant amounts of time in their local libraries due to the safety of this space as well as the trusted adults working there. The American Library Association has shifted its priorities to focus more on mental health through employing social workers and providing mental health programs. As such, public libraries may be promising sites for the implementation of digital mental health (DMH) programs for teens.

Objective: This study aimed to examine how teens who attended their local public library experienced and managed their anxiety, what mental health supports they were interested in receiving, and how DMH programs and public libraries can meet their needs.

Methods: We interviewed 16 teens aged 12-18 (mean 15.2, SD 2.0) years who used the library frequently at the time of the interviews. Of these teen patrons, 56% (9/16) identified as female, 31% (5/16) identified as male, and 12% (2/16) identified as nonbinary. Most (11/16, 69%) identified as either White or Black or African American individuals, with the remainder (5/16, 31%) identifying as Hispanic or Latino or Chinese American individuals or with ≥2 races. The interviews were individual and semistructured, designed to elicit recommendations for designing and implementing digital tools in libraries to improve teen mental health. Interview transcripts were coded by multiple coders using thematic analysis to synthesize key themes.

Results: Teens reported experiencing uncontrollability, unpredictability, and anger related to their anxiety, which they managed using strategies such as guided breathing, distress tolerance, and social connection. They also talked about other helpful management techniques (eg, progressive muscle relaxation, journaling, and mood tracking). Teens underscored the importance of pairing mood

XSL•FO
RenderX

tracking with daily activities to reveal patterns. They also stressed the significance of context and anxiety severity when choosing anxiety management strategies. Teens underscored the centrality of the public library in their lives and their view of it as a safe space where they can easily access resources and connect with friends and trusted adults. When considering the design of a DMH program implemented in libraries, they suggested including personalization for different identities, gamification, and simple navigation. Teens emphasized the importance of protecting their privacy within digital programs and that their end goal was to use the skills learned in the DMH program *offline*.

Conclusions: Teens who frequently used their local public library expressed interest in receiving digital tools via libraries to help them manage anxiety. Their recommendations will help inform future research on the adaptation and implementation of DMH programs for teens in public libraries.

(JMIR Form Res 2025;9:e67454) doi:10.2196/67454

KEYWORDS

public libraries; digital mental health; teens; youth; adolescents; anxiety; mental health; implementation; safe spaces; mobile phone; smartphone

Introduction

Background

Researchers, clinicians, and policy makers in the United States have declared a teen mental health crisis due to increasing rates of internalizing distress combined with limited access to mental health programs [1-3]. The Centers for Disease Control and Prevention found that over one-third of teens in 2019 experienced persistent feelings of hopelessness, a transdiagnostic predictor of anxiety and depression [4,5]. Anxiety is particularly common among teens, with rates increasing roughly 27% between 2016 and 2019 [6] and continuing to rise since the pandemic [7]. Prevention and treatment efforts for teens have the potential to reduce the likelihood that internalizing disorders will persist or worsen in adulthood [8], but demand far exceeds the available supply of teen mental health programs [9,10]. Increasing the number of providers would not fully ameliorate the issue as many teens face additional barriers such as high treatment costs, lack of transportation to clinics, and complex regulations regarding parental consent [11]. Consequently, advocates call for researchers and clinicians to meet teens where they are by implementing mental health interventions in settings frequently used by teens [12-14]. In this paper, we focus on public libraries as a promising yet understudied setting for delivering teen mental health programs, particularly programs targeting anxiety.

Many teens, especially those with marginalized identities, spend significant amounts of time in libraries [15]. Public libraries are a *third place*, where patrons can spend time without being expected to spend money [16,17]. The commercialization of space, particularly in urban areas, has led to libraries being one of the few indoor spaces teens can inhabit for free without facing accusations of *loitering*. As a result, libraries have transitioned in recent years to become beacons of public services, offering health services, professional development classes, art workshops, and more [18]. An increasing number of libraries now offer teen-specific programming [19-21]. Libraries are also seen as a *safe space* for teens with marginalized identities, including historically underserved racial and ethnic minority (HURE) teens; teens who identify as lesbian, gay, bisexual, transgender, queer, intersex, or asexual (LGBTQIA+); teens living in lower-resourced and disinvested areas; and teens living with mental health conditions [15,16,18,22-27]. Unsurprisingly, the teen mental health crisis is visible in public libraries. Within the last decade, the American Library Association has shifted its priorities to focus more on mental health through actions such as employing social workers within public libraries and providing patrons with mental health programs [25,26,28]. At the same time, librarians report wanting more mental health support for teen patrons that is accessible and on demand and requires little additional training for already overtaxed staff [15]. Accordingly, interventions that leverage digital technologies, or digital mental health (DMH) programs, may be well suited for use in libraries.

DMH programs are an ideal way to reach teens for several reasons. Teens commonly use the internet for mental health support and psychoeducation [29]. DMH programs can be implemented in practically any location at relatively low cost. Libraries offer free broadband internet access and computer access to all patrons, allowing DMH programs to be easily accessed in this setting. Self-administered DMH programs provide availability on demand when teens need support, eliminating the need to involve a facilitator or caregiver and also satisfying the needs of teens for independence and autonomy [11]. Despite the promise of DMH programs, many face problems with user engagement and sustainment by organizations outside the research context [30,31]. Researchers have suggested that this may reflect, in part, that DMH programs are typically designed for research contexts rather than for implementation in real-world settings [32,33]. On average, it takes 17 years [34] for research evidence to reach real-world settings and change practice. These statistics have not significantly changed in decades, highlighting the need for innovative solutions to address this research-practice gap.

To ensure that DMH programs are implemented in real-world settings, researchers may look toward approaches that (1) center the community who will use the program, (2) design the program according to the community's needs, and (3) evaluate implementation metrics outside of effectiveness. Community-based participatory research (CBPR) is an approach to research that prioritizes partnership and shared decision-making with community members impacted by the research [35]. The application of CBPR principles in the development of DMH programs ensures that DMH programs

XSL•FO
RenderX

can positively impact the community of interest. Human-centered design (HCD) is defined as a set of methods that require that "the human perspective is considered from the initial conception to the eventual design, and that the people who will use and/or be affected by the designed product are involved in the process" [32]. Similarly to CBPR, HCD elevates the end users' voices in intervention design and implementation, increasing the likelihood that design outputs will be acceptable and sustainable. Finally, implementation research (IR) involves studying how evidence-based practices can best be implemented in real-world settings outside of research contexts (eg, examining determinants of and strategies for implementation and evaluating and designing for the deployment context) [36]. In combination, CBPR stipulates principles for partnering with a community that may use a DMH program, HCD provides methods for effectively creating the DMH program such that end users will be engaged, and IR plans for and evaluates the deployment of the DMH program outside of a research setting. The frameworks improve the likelihood that DMH programs will have high uptake and a meaningful and sustained impact beyond research.

Across CBPR, HCD, and IR frameworks, the first step in creating interventions that are engaging and sustainable is to understand and center the needs, goals, and preferences of end users, their communities, and eventual deployment settings [31,33]. For example, previous research suggests that teens often experience anxiety through somatic symptoms and excessive worrying. However, it is unknown whether these findings fully generalize to the experiences of teen library patrons or whether there are additional factors to consider that are unique to the community they are in [37,38].

Objectives

The purpose of this study was to use CBPR, HCD, and IR methods to understand the lived experiences of teen library patrons related to anxiety, elicit recommendations on what teen patrons would like to see in a DMH program for anxiety and anxiety management, and understand contextual factors and teens' perspectives of the public library as a deployment setting. Our goal was to use this information to inform the adaptation of DMH programs that can be implemented in public libraries to meet teens where they already are and help alleviate the teen mental health crisis [15].

Methods

Setting and Participants

Building on our previous needs assessment with library workers [15], this study examined teen patrons' needs and preferences related to the implementation of a DMH program for anxiety within a public library that serves communities in the west side and west of Chicago, Illinois. We will use *teens* or *teen patrons* to describe the teens who use the library frequently as this is the term the library community and teens use. The library serves community members from diverse racial and economic backgrounds, prioritizing the needs of marginalized patrons and those most affected by health inequities in their programming and the services they offer the community. AAK interviewed 16 teens aged 12-18 (mean 15.2, SD 2.0) years who used the library frequently at the time of the interviews. The interviewer held a master's degree and doctoral degree in experimental psychology, was an assistant professor, and identified as female at the time of the interview. We define *frequently* as those teens who either attended weekly programming at the library or visited the library during the week. Of these 16 teen patrons, 56% (n=9) identified as female, 31% (n=5) identified as male, and 12% (n=2) identified as nonbinary, and 44% (n=7) identified as White individuals, 25% (n=4) identified as Black or African American individuals, 12% (n=2) identified as Hispanic or Latino individuals, 12% (n=2) identified with ≥2 races, and 6% (n=1) identified as Chinese American individuals.

Procedures

All participants were recruited from a local Chicagoland public library between 2020 and 2021. The library's Teen Services staff coordinated with the research team to post study flyers in the library common areas, promote the study in the library's Teen Services programming, post in the library's newsletter, and receive referrals from other teen patrons. Recruitment occurred until saturation was achieved in the interviews. No participant dropped out or refused to be part of the study; no repeat interviews were conducted. During the first year of the COVID-19 pandemic, the physical library was shut down; however, programming for teens continued via Zoom (audio only). Our research team built rapport with teens in person before the pandemic onset (and then once the physical library opened again), as well as attending the internet-based weekly programming events led by teens when the library was shut down. A total of 16 teen patrons participated in a 1-hour, individual needs assessment interview with members of the research study team.

Due to COVID-19 restrictions, all interviews were conducted over the phone or using audio-only Zoom calls (Zoom Video Communications). The semistructured interviews covered how participants experience and manage anxiety, current use of technology, and current or future library programming focused on teen mental health (Multimedia Appendix 1). Participants were also asked about what resources or features within a digital program may help them better manage anxiety. All interviews were audio recorded, transcribed, and uploaded to Dedoose (SocioCultural Research Consultants) [39] for qualitative data analysis. If teens expressed thoughts of suicide or self-harm behaviors during the interview, a predetermined safety procedure was in place to assess risk of imminent harm and take appropriate action (including safety planning for those deemed low risk and connecting the teens with emergency services for those at high risk). While some teens expressed experiencing self-harm thoughts in the past, no teen expressed current suicide or self-harm thoughts or behaviors, so the safety protocol was not enacted.

Data Analysis

In total, 3 researchers (AAK, MH, and CW) conducted a thematic qualitative data analysis of the transcribed interviews. These coders all had training in qualitative analysis, all held master's degrees, and one also held a doctoral degree. Coders used an iterative approach to create and refine the codebook. No field notes were used in the analysis. First, coders read and annotated 3 interview transcripts each, labeling relevant words,

XSL•FO
RenderX

phrases, or sections with codes. Then, coders met to discuss and conceptualize the annotations, identify highly salient codes, and group codes into broader categories. The coders met regularly to discuss the application of codes, resolve discrepancies in code application, and refine the codebook accordingly. Next, coders were randomly assigned to apply the updated codebook to 5 or 6 additional interview transcripts each, after which they met to resolve discrepancies and further refine the codebook by removing codes that were underused or not present within the transcripts. After these iterations, the final codebook was applied to all transcripts, encompassing the following key themes: (1) public libraries and mental health, (2) teen anxiety experiences and management, and (3) DMH program design for teens and implementation considerations (see Multimedia Appendix 2 for the COREQ [Consolidated Criteria for Reporting Qualitative Research] checklist). The community members were primarily involved in member checking [40-43] in that they were most interested in reading the first drafts of the paper and verifying that the main concepts identified by researchers in the qualitative data analysis (and, thus, highlighted in the *Results* section) were an accurate representation of the teen patrons' feedback and also of the library itself. The summary of the key themes is shown in Textbox 1.

XSL•FO
RenderX

Textbox 1. Summary of key themes and subthemes from the qualitative analysis conducted on the transcripts of the individual semistructured interviews with teens who frequently used a local Chicagoland library between 2020 and 2021. The interviews covered how teens experienced and managed anxiety; current use of technology; and current or future library programming focused on teen mental health, including digital mental health (DMH) programs for anxiety (Multimedia Appendix 1).

Public libraries and mental health

- Most teens liked the centrality and easy access of the library.
- The library provides teens easy access to reliable mental health resources to connect with other teens on this subject.
- Teens from historically underserved racial and ethnic minority backgrounds and other marginalized identities perceived libraries as safe spaces.

Teen lived experiences of anxiety

- Uncontrollability: teens identified catastrophizing, overthinking, and a lack of control over thoughts, emotions, and actions as feelings of anxiousness. For some, anxiety led to feelings of sadness and depression and thoughts of self-harm. The cyclical pattern resulting from the co-occurrence of anxiety and depression symptoms severely impacted teens' drive and motivation. The uncontrollability of anxiety led teens to feel burned out and tired, with a sense of dissociation from reality and derealization.
- Unpredictability: new experiences were a major source of anxiety related to unpredictability. Experiencing an unpredictable situation prompted urges to escape the situation or avoid future anxiety-provoking situations.
- Anger: anxiety can lead to teens getting easily frustrated, having a short fuse, and feeling angry. Sources of anger ranged from irritation over mistakes to frustration and bitterness toward personal anxiety. External annoyance and bitterness usually ended in arguments with family and friends.

How teens manage anxiety

- Breathing and progressive muscle relaxation: teens used apps with guided breathing, in-the-moment breathing exercises prompted by devices tracking physical changes, and progressive muscle relaxation to help regulate high anxiety levels.
- Grounding and distress tolerance: grounding techniques were used to disrupt uncomfortable thoughts and feelings. One method was the 5-4-3-2-1 technique, where one names things related to each sense. Another anxiety-managing technique used the sense of taste, namely, salty and sour tastes.
- Social connection: engaging in various activities with loved ones, namely, talking to and connecting with family and friends through conversation, helped with easing anxiety. Teens with anxiety felt validated and less isolated when sharing experiences with other teens or mental health professionals. Teens suggested an anonymous social component in a DMH program serving as a safe space for them to share experiences and advice with one another. Teens also wanted access to a mental health professional to provide advice in a safe space.
- Journaling and mood tracking: teens used journaling to work through and release their anxiety, helping them gain confidence and motivation to address stressors. Teens appreciated when mood tracking was used to reveal patterns such as relationships between their anxiety and daily activities. Some teens desired a pop-up mood tracker when first opening the program that could tailor mood management techniques to their input responses (eg, if high anxiety is reported, the program directs teens to in-the-moment anxiety relief techniques).
- Psychoeducation and crisis resources: teens found psychoeducation and crisis resources to be useful in contextualizing and learning the patterns and origins of their anxiety. Teens emphasized the importance of warm handoffs and easy access to reliable resources when in a crisis.
- Strategy choice and context: teens wanted a wide range of anxiety-mitigating techniques. Having variety allows teens to experiment and figure out what works best for them. Teens noted that context and anxiety levels played important roles in choosing a technique. The main goal of the resources should be to allow for practicing skills to eventually be used without technological assistance.

Desired DMH program features

- Personalization: teens liked app algorithms that curated or suggested content based on feedback, use, and data entry, as well as reports on patterns. They emphasized the need to understand who the program is targeting and provide resources specific to that population.
- Aesthetics and layout: teens preferred simple and concise app layouts that made navigation of the resources easy. Teens recommended that apps should be low commitment as longer materials were hard to keep up with and became overwhelming. They favored content delivered in teen-friendly language to keep the program more relatable and up-to-date.
- Gamification and animations: videos, animations and graphics, and mini games were the best modalities for information acquisition and engagement for teens. Content would be even more compelling if it featured teen celebrities or popular social media influencers. Teens liked interactive digital features, such as haptic and touch screen features, while practicing skills.
- Privacy and accessibility: protecting teens' privacy was a top priority—teens worried about parental access to and surveillance of their phones and parents potentially finding out that they were struggling with mental health issues. Teens noted school rules and caregivers restricting phone use as access barriers. Paid subscriptions, an overt mental health focus in case of onlookers, and phone limits (eg, low storage and limited or inconsistent broadband access) were access barriers to resources on apps.

XSL•FO
RenderX

Ethical Considerations

All research protocols were approved by the Northwestern University Institutional Review Board (STU00212314) before study enrollment. Both the interested teens and their caregivers provided permission to take part in the study. In the process of transcribing interviews, all potential identifying information was removed to protect the privacy of the participants, and each transcription was labeled as *P1*, *P2*, etc. Transcripts were not returned to participants. Each adolescent was given a US $25 Visa gift card for their time and participation in the interviews.

Results

Public Libraries and Teen Mental Health

Teens discussed resources they enjoyed at the library, the impact the library had on them, and their opinion on how libraries could support teen mental health. Most teens talked about the centrality and easy access of the library. P13 stated the following:

> *Because of where I get off at the train on the metro, I get off right across the street from the library so I can go to the library and just chill there.*

Other teens described the library as a community access point. They noted that the library has many connections throughout the community and serves as a hub for resources.

When reflecting on the role public libraries should play in teen mental health, most teens agreed that the library is a great place for teens to access reliable mental health resources and connect with other teens about mental health via library programming. While some teens had specific ideas for new mental health offerings, such as creating a website or smartphone app to connect teens to resources to manage their mental health, they also spoke about resources that the library already had in place. For example, P8 described the following:

> *Like obviously books are a resource—and kind of just having that level of flexibility whether it's mentorship or whether it's grounding yourself in a story or whether it's a pamphlet...having those different opportunities for how people have different or how different people interact with their own mental health and coping mechanisms.*

Teens emphasized how the library was a safe space for them because of the trusted adults that worked there, friends they made at the library, and the library's intentional teen programming. P12 summarized this nicely:

> *I was always at the library, and I was always doing these groups and these fun summer clubs. I had a lot of friends from the library...I do think that the library is a great resource place. There's a lot of good, trusted adults. It's quiet. And sometimes in a hectic household or a hectic friend fight, you might just wanna go to a place and feel welcomed and feel quiet and feel calm.*

Another teen, P2, underscored their admiration for the trusted adults available via their teen programming at the library:

> *Even though you're talking to an adult, you're just talking to people you know that care about you.*

The library was perceived as a safe space for many HURE and LGBTQIA+ teens in our sample, as well as those teens most experiencing inequities. Several of these teens mentioned that a social justice program within the library led by a local activist helped create a safe space in the library. In addition to weekly meetings for this program, teens researched and led town hall meetings educating the public on issues such as climate change's disproportionate effect on lower-resourced communities. When asking P3 what they liked about this program, they responded the following:

> *For well me as like you know like a black, queer individual in every sense of the word, obviously because we [the group] love each other and stuff it's great, but it's also like I can just talk about this stuff...about stuff that I usually wouldn't talk to normal people you know your average Joe or whatever about—talking about the oppression and the prison system and mass incarceration and all that stuff...But it definitely just helps so I can get all of those problems or frustrations out and also just for mental health just seeing these people [in the group]...It's just like we're a family and that's really nice.*

As can be seen in this quote, one HURE teen described finding solace in the shared experiences and similar backgrounds within this group, which allowed them to safely discuss topics including structural racism and oppression. Another teen, P13, described the following:

> *It's just nice to have a space where I can be Black and annoyed and then there's other people who get it.*

Teens who attended this particular library program also emphasized their appreciation for the trusted adult who led the group, highlighting that the local activist empowered them and elevated their voices, challenging them to think critically while researching and reporting on current issues.

Teens' Lived Experiences of Anxiety

Overview

Teens shared about their experiences with anxiety, which impacted them in terms of physical sensations (eg, racing heart, difficulty breathing, and feeling shaky or flushed) and socially (eg, being easily embarrassed in public and worrying about disappointing or being judged or rejected by others). They also shared experiences less addressed in the teen anxiety literature; specifically, they spoke about their feelings of uncontrollability and unpredictability and the role of anger in their anxiety. As there is already a robust literature on physical sensations and social anxiety, we focused on these less explored aspects of teens' experiences of anxiety.

Uncontrollability

Teens spoke about catastrophizing and overthinking as part of their anxiety, as well as feeling like their thoughts, emotions,

XSL•FO
RenderX

and actions were out of their control. For example, P15 described the following:

> *My thought train is just going a million miles an hour. It's like this thought leads to this other thought and it turns into a big fireball of negative thoughts...like out of control.*

In addition to feeling a lack of control, some teens' anxiety led to feelings of sadness and depression and thoughts of self-harm. P13 described their experience as follows:

> *For me, it's usually like the anxiety comes and then the sadness once I cool down. So, I guess comparing it to a storm, it's like the anxiety is the hurricane and then the sadness is kind of like the eye of the storm and then the anxiety comes back and then there's a hurricane that follows it.*

This teen highlighted the cyclical nature of symptoms, an experience shared by many others. Teens went on to describe how the co-occurrence of anxiety and symptoms of depression had a severe impact on their motivation, ultimately taking a toll on their grades and relationships with friends and family.

Finally, the uncontrollability of anxiety led some teens to feel burned out and tired, causing them to zone out, operate on autopilot, and shut down. They described feeling a sense of dissociation and derealization because of their uncontrollable anxiety. P4 described the following:

> *Dissociating and not feeling like I'm really present. Just kind of I become less aware of my surroundings.*

They went on to describe an escalation in which they felt completely detached from reality.

Unpredictability

In addition to the uncontrollability of their anxiety, several teens spoke about the unpredictable nature of their anxiety, especially when it came to new experiences. For example, P7 explained the following:

> *If I go somewhere new, I think like, "Where is everything?" because I don't know where anything is and then like, "What's gonna happen? What am I gonna do? What are people gonna be like? Are people gonna judge me?"*

Teens commonly described that experiencing an unpredictable situation led to urges to escape it or avoid future anxiety-provoking situations, such as new experiences. P9 described their reaction to new social situations:

> *I feel like people are staring at me. And all I wanna do is just hide usually. So, I try to get out of the situation, like wherever I am in public.*

Anger

Several teens described getting easily frustrated, having a short fuse, and feeling very angry when anxious. Some internalized their anger, becoming irritated at themselves for making a mistake or experiencing frustration and bitterness toward their own anxiety and anxiety-provoking situations. Other teens described feeling annoyed and bitter externally, like they "needed to go punch something" (P16). This usually ended in arguments with family and friends, adding to their anxiety and feelings of isolation.

Teen Anxiety Management Strategies

Overview

Teens described using multiple strategies to manage their anxiety that included internally focused strategies for self-management as well as externally focused strategies for connecting with others and seeking resources. Teens reported diverse experiences with mental health treatment and resources—some described experiences with formal therapy, others used smartphone apps for mental health, some had experiences with both, and others had no previous experience engaging with these resources. We describe teens' experiences in the following sections to inform future interventions that can promote and reinforce strategies that teens already feel comfortable using and that are already integrated into their daily lives. These methods of coping are well aligned with evidence-based strategies.

Breathing and Progressive Muscle Relaxation

Most teens mentioned that breathing exercises helped them manage anxiety in the moment. For example, P7 described using breathing exercises while having catastrophic and overwhelming thoughts:

> *Because I focus on my breathing instead of my thoughts. It kind of clears my mind too.*

Teens also described using apps with guided breathing exercises (eg, facilitating box breathing and deep breathing) alongside interactive visualizations (eg, haptic and touch screen features). P6 described the following:

> *I really like the ones that are like circles that tell you to hold it in—that says to breathe in and hold it, then breathe out and hold it.*

Other teens enjoyed apps that used inputs from other devices, such as smartwatches or exercise wristbands, to trigger breathing exercises when their heart rate increased. Other teens found utility in using breathing exercises as a supplemental strategy that could help once they had regulated their high anxiety levels using other strategies described in this section. Another anxiety-reducing technique that teens reported using in apps was progressive muscle relaxation. When describing a smartphone app that included this technique, P7 stated the following:

> *My favorite is where you tighten your muscles, which is like you ball your fists up and then relax them and stuff...I just kind of like it because it's like you have all your stress and all your feelings and then you just let them go.*

Grounding and Distress Tolerance

Several teens spoke about grounding and distress tolerance techniques, which were used to disrupt uncomfortable thoughts and feelings and then return to a more manageable level of anxiety. Some mentioned using the 5-4-3-2-1 technique when at their highest levels of anxiety—or, as P7 called it, "the five

XSL•FO
RenderX

senses thingy"—where they named things related to each sense (eg, name 5 things they saw, 4 things they felt, and so on). When paired with breathing exercises, this technique made their anxiety manageable. Another teen, P6, described a unique technique:

> *When I feel really anxious, what I do is put salt on my tongue. So, that kinda just it distracts me from everything that I'm thinking about...It just disconnects me from reality for a bit because of the strong taste of salt messes up everything else that I was thinking about.*

This teen mentioned doing the same exercise and achieving the same results with sour-tasting things such as lemons, limes, or vinegar.

Social Connection

All teens spoke about the importance of social connection in dealing with their anxiety even if social contexts brought on anxiety. P8 described this phenomenon as follows:

> *As much as I do have social anxiety, I'm a very social person.*

The activities with friends and family that helped ease their anxiety ranged from low-key, in-person activities (eg, watching movies or listening to music together) to internet-based activities (eg, where teens would play video games together). The simple act of talking to and connecting with friends and loved ones was the most mentioned social activity that helped in easing anxiety, with teens emphasizing how these types of "conversations are really key in those [high-anxiety] moments" (P2). P8 described how conversations such as these helped:

> *Whether it's just completely rambling you know like distracting me—or guiding me through breathing exercises or going on a walk with me and make sure I keep moving and that I'm paying attention to my surroundings like, "Oh, do you see that over there?" and that kind of thing and just helping to ground me and stuff.*

Teens were asked to brainstorm how a DMH program could provide social connection to help with anxiety. Teens reflected that they felt validated and less isolated when reading about other teens' experiences of anxiety and how they dealt with it. Accordingly, most teens suggested including an anonymous social component within a DMH program that would serve as "a safe space to talk to each other and see what people are going through" (P5). In addition to these shared experiences and advice, some teens wanted connection with a mental health professional or monitors to provide professional advice while also keeping it a safe space. P8 spoke to the benefits of this kind of platform:

> *Being able to ground yourself in other people's experiences so that you can ground yourself in your own and feel more validated or feel making anxiety more approachable.*

Journaling and Mood Tracking

Teens shared how journaling and mood tracking helped them alleviate or better understand their anxiety. P3 described how they used journaling to work through their anxiety:

> *After I'm done being upset in the moment, I'll just write about that moment and it just kind of just helps me release those emotions—get them down so I remember them, remember what happened.*

This cathartic nature of journaling was shared by most teens who used it as an anxiety management strategy. P14 had a unique technique where they would create characters and storylines based on a hardship or anxious situation they were experiencing. This teen garnered confidence and motivation to address stressors through these stories. P14 described how their story creation helped:

> *Yeah, it definitely takes me out of the moment and then it also it just feels like if I can beat this big milestone and do that and achieve that and then have it, homework's nothing. I can finish that homework assignment because I can do this. I can do this. I'm more powerful than this homework. I will squash it, you know? Like a bug...It's just a confidence thing.*

Other teens saw value in reviewing patterns in their mood-tracking data, such as relationships between their anxiety and daily activities. P4 described the following:

> *It's nice to see a daily tracker so you can see the bigger trend of what causes your stress and what periods and times you're more stressed.*

Teens desired a mood tracker that pops up when first opening the program so that their answers could tailor their interaction with the program. For example, if they entered a high anxiety rating, the program might direct the teens to in-the-moment anxiety relief techniques.

Psychoeducation and Crisis Resources

Several teens described the benefits of including psychoeducation and crisis resources in a DMH program embedded within the library. Teens found psychoeducation helpful in contextualizing and learning the patterns and origins of their anxiety. P1 expanded on their idea for integrating psychoeducation and resources into a DMH program:

> *I think it would be helpful to have a resources page...articles on anxiety, quick advice, I guess, and then maybe also a hotline page.*

Teens also mentioned the importance of warm handoffs to reliable resources and the need for easy access to these resources when in a crisis. For example, P16 suggested that the future program should "definitely start off with, definitely a phone number to a social worker that if someone is feeling anxious, or they wanna hurt themselves, you can talk to them."

Strategy Choice and Context

Overall, teens wanted a variety of anxiety-mitigating strategies to choose from as different strategies could work better or worse for different teens and as variety allows them to experiment to find what works best for them. Context and anxiety level played

XSL•FO
RenderX

important roles in choosing a technique. For example, some teens described that they typically had their phones with them in public and would use social media or other apps to deal with their stress in the moment. P7 described the following:

> *It depends on where I am, but if I'm at home I'll play my guitar, listen to music, or I'll do meditation sometimes and sometimes I'll draw...But if I'm in a public place, then I'll do deep breathing.*

Other teens expressed that their approach depended on how anxious they were, with some strategies (eg, physical activity) being better suited for high-anxiety moments and others (eg, psychoeducation and listening to music) being better suited for lower-anxiety moments. Regardless of the coping skill or how it was taught, the consensus from teens was that the main goal should be practicing skills to then use them without technological assistance when in-the-moment anxiety occurs. One participant described that "it's not just about the technology or about the app" (P8). Most teens appreciated the supportive technologies they had previously used that helped them practice exercises, which in turn made it easier to use the acquired skills outside the app or in real life.

Desired DMH Program Features

In addition to specific anxiety management strategies, teens discussed the top features they would like in a DMH program embedded within the library. These desired features fell under 4 broad themes: personalization, aesthetics and layout, gamification and animations, and privacy and accessibility.

Personalization

Teens discussed the importance of being able to both personalize and receive tailored information from programs. For example, P15 suggested the following:

> *You know how you can type something into WebMD and it will show what the problem is and also how to fix it. I think that's what this app should look like. If you have anxiety over a bad grade, you can do like, "What can I do to overcome my anxiety about that or stress?"*

As this example suggests, teens liked when apps used algorithms to curate and suggest tailored content based on their feedback, use, and entry of data into that app. Teens also appreciated reports on their particular use patterns (eg, features most used) and shortcuts to view these data. In addition, it was important that the mental health program offer resources specific to the population it was designed for. For example, P13 described the following:

> *One thing that I notice a lot with resources where people put resources for like, "Did you know about this and blah, blah, blah?" I'd be going through it, and it would be like, "This is clearly meant to educate white people."*

Several HURE teens underscored that current mental health resources were catered to a small group of people, White teens in particular, and did not reflect them or their mental health needs.

Aesthetics and Layout

Many teens preferred simple digital program layouts that were easy to navigate. They described being overwhelmed if the home screen of the program had too many options. For example, P11 said that they did not want the following:

> *A lot of folders or that you have to click a lot of things to get to something. I think it should be the important things, and then from there it can branch into more specific ways to cope with types of stress or recommendations or things like that.*

Other teens emphasized that apps should be low commitment (ie, relatively quick and easy to navigate and use on a daily basis). Apps with higher commitment, or those that required significant time or energy to use, and those with longer materials were difficult to keep up with and quickly became overwhelming. Furthermore, teens shared that it would be important to have regularly updated content in teen-friendly language. P8 said the following:

> *Teens can be turned off by aesthetics that just feel kind of like dated or they feel like they were made by people much older than them.*

Gamification and Animations

Several teens underscored that the best modalities for information acquisition and engagement were videos, animations and graphics, and mini games. P5 said the following:

> *Like a video lesson or something that the user could interact with or participate along with because there are many resources out there for a number of things...words don't help as much as seeing the person or listening.*

They went on to describe that using teen actors in videos could become cheesy but videos could be compelling if they featured teen celebrities or popular social media influencers. Teens also enjoyed interactive smartphone or tablet features (eg, haptic and touch screen features) where they could touch the screen while practicing skills. Teens gave examples that ranged from using the notification vibrations on the phone to guide deep breathing exercises to dragging their finger across the screen until an image was revealed.

Privacy and Accessibility

Protecting the teens' privacy was among the top priorities. P6 emphasized the following:

> *Making sure that my information's safe and only to me. It's not shared with anyone else, especially, random people.*

Other teens expressed worry regarding their parents' access to and surveillance of their smartphones and them potentially finding out that they were struggling with mental health issues. P16, for example, stated the following:

> *It's just my parents, every week, they checked my phone. And if they saw I was searching, "How to overcome anxiety," I wouldn't want them to make me the center of attention. And make them drop everything that they're doing.*

XSL•FO
RenderX

Some teens had their smartphone use restricted by caregivers or school rules. Policies ranged from not using smartphones during school hours to caregivers taking away teens' phones as punishment (eg, when they said "something bad") or at specific times (eg, being unable to use anything other than the music app on their smartphones after 8 PM on weeknights). Teens were also concerned about paid subscriptions, onlookers being able to see that the program was for mental health, and smartphone-specific barriers (eg, low battery, storage, and data of phone and inconsistent access to broadband). Many teens spoke about liking a particular app, such as a mood-tracking app, but reported that they stopped using it because they had to use their cellular data or pay to use the full version of the app, further limiting their use.

Discussion

Principal Findings

This study aimed to understand the needs and perspectives of teen library patrons in an effort to support the creation of a teen-focused DMH tool that can be disseminated in libraries. The results of the needs assessment interviews with teens who frequently use the library provided critical information on their perspectives of the library's function and programming. Teens underscored the centrality of and, thus, easy access to their public library. They also referred to the library as a hub for connecting patrons to community resources and services. This is consistent with our previous research, in which library workers identified a main function of the library to be a conduit to guide patrons to desired resources [15].

Teens talked about the teen-specific programming they attended at their library, indicating that this is where they met a lot of their friends and fostered relationships with the trusted adults leading those programs. These factors led teens to view the library as a *safe space*, which is consistent with several other reports by library workers and researchers underscoring that communities see public libraries as safe spaces [27,28]. HURE teens and teens with marginalized identities (eg, LGBTQIA+ teens, teens living in lower-resourced and disinvested areas, and teens living with mental health conditions) found belongingness, companionship, and solidarity in library programming that spoke to different aspects of their identity. Given that there are more unsafe spaces for these groups due to marginalization, it is critically important to invest in and maintain safe spaces that bring individuals with shared experiences and similar backgrounds together and that celebrate the different aspects of their identities, especially within the teen population.

Teens shared the nuanced experiences they had regarding their anxiety. They spoke to the uncontrollable and unpredictable feelings associated with their anxiety, which led some teens to catastrophize, feel depressed, or have thoughts of self-harm, which then cyclically fed back into their anxiety. Other teens described dissociating to cope with adverse feelings or feeling the need to avoid or flee anxiety-provoking situations. There is evidence of the uncontrollable and unpredictable nature of anxiety in the adult anxiety literature, especially in relation to panic disorder, yet those variables have not been well studied in the teen anxiety literature and are worthy of future research [44-47]. More research is needed to better understand these 2 variables and their potential role as risk or maintenance factors, which could then be used as targets for DMH programs for anxiety and other mental health conditions in teens [48-52].

Other teens talked about their anger in relation to their anxiety. Some described the internal aspects of anger that may be expressed as frustration toward their own anxiety or frustration about making mistakes. Others described the external aspects of anger as experiencing frustrations that often led to conflicts with friends and loved ones. This finding is aligned with those of extant literature showing that anxiety and depression commonly manifest as anger and irritability in teens [4,48-53]. In addition, there is evidence in the literature suggesting that anger can be a prominent symptom co-occurring with several internalizing disorders [54]. Stress and anger as a mood symptom may be connected to experiences of racism among teens [54]. Similarly to other studies [54-56], our study shows that anger and racism-based stress play a role in the emotions of teens who may feel disenfranchised by the current status quo in society. Anger is often seen as a maladaptive response to stress; however, with regard to racism-triggered stress, anger can be an expected expression, and sometimes this expression allows for healthy processing of emotional stressors. Given that most of the teens in this study identified as HURE or multiracial, along with the literature indicating that anger is a known response to stress related to structural racism and discrimination, additional research is needed to better understand the relationship between anger (as a common stress response to racism) and anxiety among teens who are Black or Indigenous or hold other racial and ethnic identities. While the most potent intervention would be to address the ultimate problem of structural racism and discrimination, interventions that provide support and validation to HURE teens in the face of systemic racism and oppression are critical. For example, these programs could be offered as part of library services and provide teens with acknowledgment of factors such as privilege, oppression, and discrimination and ways to curate social safety, as well as coping strategies to alleviate the mood symptoms caused by racism [51,57,58].

Teens described a variety of coping strategies (eg, guided breathing, muscle relaxation, social connection, distress tolerance, and mood and behavior tracking) that they used to help feel more in control of their anxiety, lessen their symptoms, increase knowledge of anxiety patterns, and increase their motivation to approach rather than avoid anxiety-provoking experiences. For example, teens gained knowledge about their anxiety when mood tracking was used to reveal patterns such as relationships between their anxiety and daily activities. They liked this because it emphasized behaviors that they could change or control that led to observable effects on their mood. When applied to the design of a DMH program for teens, it may be beneficial to include some of these strategies as on-demand activities. Leveraging haptic and touch screen features to actively engage teens in these strategies, especially strategies that they are familiar with and are already integrated into their daily lives, may also be worthwhile [59,60].

XSL•FO
RenderX

In terms of overall goals of a DMH program, teens wanted to learn a variety of anxiety management strategies and, ultimately, be able to choose the most effective strategy in particular contexts (eg, exercise for high-anxiety moments, listening to music for low-anxiety moments, playing guitar at home, and deep breathing exercises in public). Interestingly, while they wanted the DMH program to facilitate learning of skills, they ultimately wanted to be able to practice these skills without technological assistance when in-the-moment anxiety occurred. This use goal is consistent with other reports from the DMH literature; for example, Latino and Latina [61] teens indicated that their main goals for using DMH programs were to practice skills within the program and then, ultimately, be able to use the learned skills *offline*. If teens are using DMH programs with this goal in mind, this could have implications for how we interpret engagement metrics with teen DMH programs [62]. Teens' low, sporadic, or drop-off in engagement with a DMH program may not necessarily reflect a poor outcome or poor retention; rather, it could be that teens used the program as much as they needed to incorporate the skills learned into their daily lives. Mixed methods research is needed to better understand the patterns of teen engagement with DMH programs; how those patterns map onto teens' actual use of the skills acquired in real life; and, ultimately, which patterns are most helpful in mental health management.

Teens shared their feedback on additional features that they would like to see in a DMH program for anxiety. It was important to teens that the aesthetics and function of the DMH program had the characteristics of clean design and easy navigation. Teens also stressed the importance of protecting their privacy, making the program accessible (eg, free and aligned with caregiver and school restrictions on devices), and designing discreet branding, all of which are consistent with several previous reports describing teens' or therapists' design recommendations regarding a teen DMH program [63,64]. In addition, teens reflected on the characteristics that increased their engagement with a digital program, which included gamification, videos and animations, and buy-in from teen celebrities or social media influencers. Other pediatric fields have found partnering with teen social media influencers to be a helpful strategy in enhancing knowledge and even outcomes of health care delivery [65]. In addition to the other features mentioned, partnering with popular teen social media influencers who are known for speaking about their mental health on social media outlets may be a promising strategy for increasing awareness and potentially intervention effectiveness of teen DMH programs. However, as teen culture and preferences rapidly and constantly shift (particularly regarding social media), DMH developers must balance ensuring relevance and timeliness of interventions with ensuring that the content still addresses theoretical mechanisms of change [66].

As for the additional feature of personalization, HURE teens underscored that the available teen mental health resources were White centric and they did not feel that *their* lived experiences were represented in current mental health resources. This suggests that we as scientists and designers need to intentionally improve the inclusion of populations who are underrepresented in teen mental health research, in particular HURE and LGBTQIA+ teens as well as those from lower-resourced and disinvested areas [67-76]. Teens advised that developers get to know the audience for the program and work with that population to personalize the program to their needs and preferences. This feedback raises an important design question: how do we equitably design DMH programs in community spaces that serve people with different identities as well as with different mental health needs and preferences? There are calls in the literature to move away from a *one-size-fits-all* model when developing DMH services [77-80]; however, more discussion and research are needed to better understand the best practices of DMH service development that is intended to meet the mental health needs and preferences of heterogeneous audiences (eg, a program encompassing tailored versions for subgroups within the heterogenous audience or different programs for different subpopulations) [81].

Limitations

We would be remiss to not include a discussion on study limitations. The public library we worked with was well resourced with existing social services (eg, employed social workers, had a mental health and teen service budget, and offered free mental health assessments). Accordingly, the results of this study may not be generalizable to other public libraries that have different infrastructures and resources. More research is needed to better understand the teen resources offered by different public libraries and their branches across the United States. Second, we interviewed teen participants residing in a large, socially and politically progressive, and racially and ethnically diverse urban area. HURE teens living in areas with different sociopolitical environments and sociodemographic compositions may have different perspectives on whether and how they would like to engage in library-based DMH services. Third, we did not screen for anxiety before the interviews; however, it was apparent from teen interview responses that there was high variability in severity levels (level of interference) and treatment experience (some talked about their experiences with therapists, and some disclosed no experience with formal treatment). Finally, while the racial and ethnic makeup of the coders was diverse in representation, it is possible that our views as researchers led us to de-emphasize, overlook, or misinterpret aspects of the qualitative data, especially on topics such as equity, racism, and racial justice. The researchers took measures to reduce bias by engaging in researcher consensus throughout data analysis and member checking with teen and library staff from the partnering local library, some of whom are coauthors who reviewed and edited this report. We also included coauthors who are experts in the field of creating and testing DMH programs for Black, Indigenous, and other racial and ethnic minority populations.

Conclusions

Disseminating and implementing DMH programs in public libraries is a promising strategy to increase access to care among teens in need of support for their anxiety. By understanding teens' perspectives and designing tools with their needs centered, the promise of this approach may be realized. To the best of our knowledge, this is the first report elevating teen voices and opinions on implementing DMH programs in public libraries.

XSL•FO
RenderX

Future research is now needed to test whether DMH programs adapted in accordance with the suggestions presented in this paper are acceptable, useful, and sustainable.

Acknowledgments

This research was supported by the National Institute of Mental Health (K01 MH121854, R34MH128410, K01MH131898, NIMH K23MH128535, P50MH119029, K01MH125172k, and P50MH119029).

Data Availability

Due to the sensitive nature of the qualitative data with minors and the potential risk of identifying variables within the individual transcripts that would potentially impact the privacy of the sample, the individual interviews are not publicly available, nor will requests for data sharing be granted. In addition, public sharing of the individual interviews was not included in the teen assent and parent consent forms.

Conflicts of Interest

DCM has accepted honoraria and consulting fees from Apple, Inc; Otsuka Pharmaceutical; Pear Therapeutics; and the One Mind Foundation, as well as royalties from Oxford University Press, and has an ownership interest in Adaptive Health, Inc. RS and LJ are employed by Oak Park Public Library. KC has received consulting fees from Kooth LLC and Koko. KC is a part-time employee at Lyra Health, Inc. RK has accepted consulting fees and honoraria from CHESS Health and Jimini Health. ABP has served as a consultant to Otsuka on developing clinical educational materials. None of the other authors have conflicts to declare.

Multimedia Appendix 1
Interview script.
[DOCX File , 15 KB - formative_v9i1e67454_app1.docx]

Multimedia Appendix 2
COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist.
[PDF File (Adobe PDF File), 422 KB - formative_v9i1e67454_app2.pdf]

References

1. Gruber J, Prinstein MJ, Clark LA, Rottenberg J, Abramowitz JS, Albano AM, et al. Mental health and clinical psychological science in the time of COVID-19: challenges, opportunities, and a call to action. Am Psychol 2021 Apr;76(3):409-426 [FREE Full text] [doi: 10.1037/amp0000707] [Medline: 32772538]
2. AAP-AACAP-CHA declaration of a national emergency in child and adolescent mental health. American Academy of Pediatrics. URL: https://www.aap.org/en/advocacy/child-and-adolescent-healthy-mental-development/aap-aacap-cha-declaration-of-a-national-emergency-in-child-and-adolescent-mental-health/?srsltid=AfmBOopEGS8BeQBnlkoPjbVrYJyDife8fPC3b5pf9h74RcpB5y_Qdxfr [accessed 2024-10-10]
3. Social media and youth mental health: the U.S. surgeon general's advisory. U.S. Department of Health and Human Services. 2023. URL: https://www.hhs.gov/surgeongeneral/reports-and-publications/youth-mental-health/social-media/index.html [accessed 2024-10-10]
4. Youth risk behavior survey data summary and trends report: 2009-2019. Center for Disease Control and Prevention. 2020 Oct 21. URL: https://www.cdc.gov/yrbs/dstr/pdf/YRBSDataSummaryTrendsReport2019-508.pdf [accessed 2024-10-10]
5. Castellanos-Ryan N, Brière FN, O'Leary-Barrett M, Banaschewski T, Bokde A, Bromberg U, et al. The structure of psychopathology in adolescence and its common personality and cognitive correlates. J Abnorm Psychol 2016 Nov;125(8):1039-1052 [FREE Full text] [doi: 10.1037/abn0000193] [Medline: 27819466]
6. Lebrun-Harris LA, Ghandour RM, Kogan MD, Warren MD. Five-year trends in US children's health and well-being, 2016-2020. JAMA Pediatr 2022 Jul 01;176(7):e220056 [FREE Full text] [doi: 10.1001/jamapediatrics.2022.0056] [Medline: 35285883]
7. Fortuna LR, Brown IC, Lewis Woods GG, Porche MV. The impact of COVID-19 on anxiety disorders in youth: coping with stress, worry, and recovering from a pandemic. Child Adolesc Psychiatr Clin N Am 2023 Jul;32(3):531-542 [FREE Full text] [doi: 10.1016/j.chc.2023.02.002] [Medline: 37201965]
8. Uhlhaas PJ, Davey CG, Mehta UM, Shah J, Torous J, Allen NB, et al. Towards a youth mental health paradigm: a perspective and roadmap. Mol Psychiatry 2023 Aug;28(8):3171-3181 [FREE Full text] [doi: 10.1038/s41380-023-02202-z] [Medline: 37580524]
9. National projections of supply and demand for selected behavioral health practitioners: 2013-2025. National Center for Health Workforce Analysis. 2016. URL: https://tinyurl.com/ycyp2f8e [accessed 2024-10-10]

XSL•FO
RenderX

10. Shortages in school psychology: challenges to meeting the growing needs of U.S. students and schools. National Association of School Psychologists. URL: https://ewscripps.brightspotcdn.com/ab/02/d7d25fd547f9b94feabaf81118dd/school-psychology-shortage-2017.pdf [accessed 2024-10-10]
11. Gulliver A, Griffiths KM, Christensen H. Perceived barriers and facilitators to mental health help-seeking in young people: a systematic review. BMC Psychiatry 2010 Dec 30;10:113. [doi: 10.1186/1471-244x-10-113]
12. Benningfield MM. Meeting youth where they are: substance use disorder treatment in schools. Child Adolesc Psychiatr Clin N Am 2016 Oct;25(4):661-668. [doi: 10.1016/j.chc.2016.05.003] [Medline: 27613344]
13. Hardy BJ, Lesperance A, Foote I, Native Youth Sexual Health Network (NYSHN), Firestone M, Smylie J. Meeting indigenous youth where they are at: knowing and doing with 2SLGBTTQQIA and gender non-conforming Indigenous youth: a qualitative case study. BMC Public Health 2020 Dec 07;20(1):1871 [FREE Full text] [doi: 10.1186/s12889-020-09863-3] [Medline: 33287787]
14. Kazdin AE. Addressing the treatment gap: a key challenge for extending evidence-based psychosocial interventions. Behav Res Ther 2017 Jan;88:7-18. [doi: 10.1016/j.brat.2016.06.004] [Medline: 28110678]
15. Knapp AA, Hersch E, Wijaya C, Herrera MA, Kruzan KP, Carroll AJ, et al. "The library is so much more than books": considerations for the design and implementation of teen digital mental health services in public libraries. Front Digit Health 2023 Jul 25;5:1183319 [FREE Full text] [doi: 10.3389/fdgth.2023.1183319] [Medline: 37560198]
16. Kosmicki LR. Equal access as social equity: a literature review of programming and services in public libraries. Progress Libr 2019(47):51-80 [FREE Full text]
17. Elmborg JK. Libraries as the spaces between us recognizing and valuing the third space. Ref User Serv Q 2011 Jun;50(4):338-350. [doi: 10.5860/rusq.50n4.338]
18. Morgan AU, Dupuis R, D'Alonzo B, Johnson A, Graves A, Brooks KL, et al. Beyond books: public libraries as partners for population health. Health Aff (Millwood) 2016 Nov 01;35(11):2030-2036. [doi: 10.1377/hlthaff.2016.0724] [Medline: 27834243]
19. Li X, Ha Y, Aristeguieta S. Teens' vision of an ideal library space: insights from a small rural public library in the United States. Evid Based Libr Inf Pract 2023 Dec 15;18(4):52-67. [doi: 10.18438/eblip30410]
20. Ornstein E, Reid PH. 'Talk to them like they're people': a cross-cultural comparison of teen-centred approaches in public library services. J Librariansh Inf Sci 2021 Jun 12;54(3):451-468. [doi: 10.1177/09610006211020090]
21. Agosto DE, Bell JP, Bernier A, Kuhlmann M. "This is our library, and it's a pretty cool place": a user-centered study of public library YA spaces. Public Libr Q 2015 Mar 03;34(1):23-43. [doi: 10.1080/01616846.2015.1000777]
22. Barone M, Fifelski S, Stocker A, Worm T, Patterson A, Braamse J, et al. Developing a model policy for public libraries to address homelessness and mental illness. Dela J Public Health 2020 Sep 28;6(4):46-51 [FREE Full text] [doi: 10.32481/djph.2020.09.014] [Medline: 34467152]
23. Simon MA, O'Brian CA, Nava M, Dahdouh R, Wafford QE, Mack S, et al. Public libraries as key partners for advancing health equity. Am J Public Health 2021 Jan;111(1):40-42. [doi: 10.2105/AJPH.2020.306004] [Medline: 33326277]
24. Pressley T. Public libraries, serious mental illness, and homelessness: a survey of public librarians. Public Libr Q 2017 Mar 02;36(1):61-76. [doi: 10.1080/01616846.2017.1275772]
25. Hall K, McAlister S. Library services and resources in support of mental health: a survey of initiatives in public and academic libraries. J Libr Adm 2021 Sep 30;61(8):936-946. [doi: 10.1080/01930826.2021.1984137]
26. Wahler EA, Provence MA, Johnson SC, Helling J, Williams M. Library patrons' psychosocial needs: perceptions of need and reasons for accessing social work services. Soc Work 2021 Oct 02;66(4):297-305. [doi: 10.1093/sw/swab032] [Medline: 34333662]
27. Libraries as safer spaces. American Library Association. URL: https://www.ala.org/aboutala/libraries-safer-spaces [accessed 2024-10-10]
28. Stringer H. Libraries as mental health hubs. American Psychological Association. 2020 Apr 01. URL: https://www.apa.org/monitor/2020/04/libraries-health-hubs [accessed 2023-02-15]
29. Pretorius C, Chambers D, Cowan B, Coyle D. Young people seeking help online for mental health: cross-sectional survey study. JMIR Ment Health 2019 Aug 26;6(8):e13524 [FREE Full text] [doi: 10.2196/13524] [Medline: 31452519]
30. Cohen KA, Schleider JL. Adolescent dropout from brief digital mental health interventions within and beyond randomized trials. Internet Interv 2022 Jan 17;27:100496 [FREE Full text] [doi: 10.1016/j.invent.2022.100496] [Medline: 35257001]
31. Fleming T, Bavin L, Lucassen M, Stasiak K, Hopkins S, Merry S. Beyond the trial: systematic review of real-world uptake and engagement with digital self-help interventions for depression, low mood, or anxiety. J Med Internet Res 2018 Jun 06;20(6):e199 [FREE Full text] [doi: 10.2196/jmir.9275] [Medline: 29875089]
32. Lyon AR, Dopp AR, Brewer SK, Kientz JA, Munson SA. Designing the future of children's mental health services. Adm Policy Ment Health 2020 Sep;47(5):735-751 [FREE Full text] [doi: 10.1007/s10488-020-01038-x] [Medline: 32253634]
33. Mohr DC, Lyon AR, Lattie EG, Reddy M, Schueller SM. Accelerating digital mental health research from early design and creation to successful implementation and sustainment. J Med Internet Res 2017 May 10;19(5):e153 [FREE Full text] [doi: 10.2196/jmir.7725] [Medline: 28490417]
34. Morris ZS, Wooding S, Grant J. The answer is 17 years, what is the question: understanding time lags in translational research. J R Soc Med 2011 Dec;104(12):510-520 [FREE Full text] [doi: 10.1258/jrsm.2011.110180] [Medline: 22179294]

XSL•FO
RenderX

35. Israel BA, Eng E, Schulz AJ, Parker EA. Introduction to methods in community-based participatory research for health. In: Methods in Community-Based Participatory Research for Health. Hoboken, NJ: John Wiley & Sons; 2005.
36. Bauer MS, Damschroder L, Hagedorn H, Smith J, Kilbourne AM. An introduction to implementation science for the non-specialist. BMC Psychol 2015 Sep 16;3(1):32 [FREE Full text] [doi: 10.1186/s40359-015-0089-9] [Medline: 26376626]
37. Crawley SA, Caporino NE, Birmaher B, Ginsburg G, Piacentini J, Albano AM, et al. Somatic complaints in anxious youth. Child Psychiatry Hum Dev 2014 Aug;45(4):398-407 [FREE Full text] [doi: 10.1007/s10578-013-0410-x] [Medline: 24129543]
38. Weems CF, Silverman WK, La Greca AM. What do youth referred for anxiety problems worry about? Worry and its relation to anxiety and anxiety disorders in children and adolescents. J Abnorm Child Psychol 2000 Feb;28(1):63-72. [doi: 10.1023/a:1005122101885] [Medline: 10772350]
39. Dedoose (Version 9.0.17), cloud application for managing, analyzing, and presenting qualitative and mixed method research data. SocioCultural Research Consultants. 2021. URL: https://www.dedoose.com/ [accessed 2024-10-10]
40. Doyle S. Member checking with older women: a framework for negotiating meaning. Health Care Women Int 2007;28(10):888-908. [doi: 10.1080/07399330701615325] [Medline: 17987459]
41. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
42. Russell CK, Gregory DM. Evaluation of qualitative research studies. Evid Based Nurs 2003 Apr;6(2):36-40 [FREE Full text] [doi: 10.1136/ebn.6.2.36] [Medline: 12710415]
43. Mays N, Pope C. Assessing quality in qualitative research. BMJ 2000 Jan 01;320(7226):50-52 [FREE Full text] [doi: 10.1136/bmj.320.7226.50] [Medline: 10617534]
44. Zvolensky MJ, Eifert GH, Lejuez C, Hopko DR, Forsyth JP. Assessing the perceived predictability of anxiety-related events: a report on the perceived predictability index. J Behav Ther Exp Psychiatry 2000;31(3-4):201-218. [doi: 10.1016/s0005-7916(01)00006-4] [Medline: 11494957]
45. Lejuez CW, Eifert GH, Zvolensky MJ, Richards JB. Preference between onset predictable and unpredictable administrations of 20% carbon-dioxide-enriched air: implications for better understanding the etiology and treatment of panic disorder. J Exp Psychol Appl 2000 Dec;6(4):349-358. [doi: 10.1037//1076-898x.6.4.349] [Medline: 11218343]
46. Zvolensky MJ, Lejuez CW, Eifert GH. Prediction and control: operational definitions for the experimental analysis of anxiety. Behav Res Ther 2000 Jul;38(7):653-663. [doi: 10.1016/s0005-7967(99)00090-x] [Medline: 10875188]
47. Gregor KL, Zvolensky MJ. Anxiety sensitivity and perceived control over anxiety-related events: evaluating the singular and interactive effects in the prediction of anxious and fearful responding to bodily sensations. Behav Res Ther 2008 Sep;46(9):1017-1025. [doi: 10.1016/j.brat.2008.06.003] [Medline: 18675954]
48. Ream GL. Minority stress and intersectionality in LGBTQIA+ youth mental health disparities. Am J Public Health 2024 Jul;114(7):669-671. [doi: 10.2105/AJPH.2024.307676] [Medline: 38662974]
49. Fish JN. Future directions in understanding and addressing mental health among LGBTQ youth. J Clin Child Adolesc Psychol 2020;49(6):943-956 [FREE Full text] [doi: 10.1080/15374416.2020.1815207] [Medline: 33074740]
50. Fish JN, Kapostasy S, Russell ST. Be YOU!: a collaborative effort to address minority stress for LGBTQ+ youth in school settings. J Gay Lesbian Soc Serv 2023;35(4):434-455 [FREE Full text] [doi: 10.1080/10538720.2022.2148036] [Medline: 38322581]
51. Diamond LM, Alley J. Rethinking minority stress: a social safety perspective on the health effects of stigma in sexually-diverse and gender-diverse populations. Neurosci Biobehav Rev 2022 Jul;138:104720. [doi: 10.1016/j.neubiorev.2022.104720] [Medline: 35662651]
52. Mahon NE, Yarcheski A, Yarcheski TJ, Hanks MM. A meta-analytic study of predictors of anger in adolescents. Nurs Res 2010;59(3):178-184. [doi: 10.1097/NNR.0b013e3181dbba04] [Medline: 20404774]
53. Walsh LM, Wolk CB, Haimes EM, Jensen-Doss A, Beidas RS. The relationship between anger and anxiety symptoms in youth with anxiety disorders. J Child Adolesc Couns 2018;4(2):117-133 [FREE Full text] [doi: 10.1080/23727810.2017.1381930] [Medline: 30828611]
54. Osman M, Williams MT, Alftieh A. State of the science on racial microaggressions. Behav Ther 2024 Nov;55(6):1172-1188. [doi: 10.1016/j.beth.2024.05.005] [Medline: 39443060]
55. Adkins-Jackson PB, Joseph VA, Ford TN, Avila-Rieger JF, Gobaud AN, Keyes K. State-level structural racism and adolescent mental health in the United States. Am J Epidemiol 2024 Jul 03:kwae164 (forthcoming). [doi: 10.1093/aje/kwae164] [Medline: 38960643]
56. Oshri A, Reck AJ, Carter SE, Uddin LQ, Geier CF, Beach SR, et al. Racial discrimination and risk for internalizing and externalizing symptoms among Black youths. JAMA Netw Open 2024 Jun 03;7(6):e2416491 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.16491] [Medline: 38865126]
57. Shen J, Rubin A, Cohen K, Hart EA, Sung J, McDanal R, et al. Randomized evaluation of an online single-session intervention for minority stress in LGBTQ+ adolescents. Internet Interv 2023 Jun 07;33:100633 [FREE Full text] [doi: 10.1016/j.invent.2023.100633] [Medline: 37635950]

XSL•FO
RenderX

58. Heard-Garris NJ, Hope E. Our voices matter: racial justice activism intervention to address structural racism and prevent depression in Black and Latinx youth. Lurie Children's Hospital of Chicago. URL: https://www.paperdigest.org/grant/?grant_id=nih_10879839 [accessed 2024-10-10]
59. Roquet CD, Theofanopoulou N, Freeman JL, Schleider J, Gross JJ, Davis K, et al. Exploring situated and embodied support for youth's mental health: design opportunities for interactive tangible device. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. 2022 Presented at: CHI '22; April 29-May 5, 2022; New Orleans, LA. [doi: 10.1145/3491102.3502135]
60. MacLean KE. Designing affective haptic experience for wellness and social communication: where designers need affective neuroscience and psychology. Curr Opin Behav Sci 2022 Jun;45:101113. [doi: 10.1016/j.cobeha.2022.101113]
61. Agapie E, Chang K, Patrachari S, Neary M, Schueller SM. Understanding mental health apps for youth: focus group study with Latinx youth. JMIR Form Res 2022 Oct 18;6(10):e40726 [FREE Full text] [doi: 10.2196/40726] [Medline: 36256835]
62. Liverpool S, Mota CP, Sales CM, Čuš A, Carletto S, Hancheva C, et al. Engaging children and young people in digital mental health interventions: systematic review of modes of delivery, facilitators, and barriers. J Med Internet Res 2020 Jun 23;22(6):e16317 [FREE Full text] [doi: 10.2196/16317] [Medline: 32442160]
63. Matthews M, Doherty G. In the mood: engaging teenagers in psychotherapy using mobile phones. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2011 Presented at: CHI '11; May 7-12, 2011; Vancouver, BC. [doi: 10.1145/1978942.1979379]
64. Knapp AA, Cohen K, Nicholas J, Mohr DC, Carlo AD, Skerl JJ, et al. Integration of digital tools into community mental health care settings that serve young people: focus group study. JMIR Ment Health 2021 Aug 19;8(8):e27379 [FREE Full text] [doi: 10.2196/27379] [Medline: 34420928]
65. Psihogios AM, Roth M, Gomez C, Hekimian-Brogan E, McQueen C, Yanez B. Partnering with social media influencers to equitably improve adolescent and young adult cancer outcomes: a novel strategy to support cancer care delivery. JCO Oncol Pract 2024 Jun;20(6):755-760. [doi: 10.1200/OP.23.00449] [Medline: 38408282]
66. Fitzpatrick S, Crenshaw AO, Donkin V, Collins A, Xiang A, Earle EA, et al. We have spent time, money, and effort making self-help digital mental health interventions: is anyone going to come to the party? J Med Internet Res 2024 Sep 19;26:e58198 [FREE Full text] [doi: 10.2196/58198] [Medline: 39298760]
67. McCall T, Foster M, Tomlin HR, Schwartz TA. Black American women's attitudes toward seeking mental health services and use of mobile technology to support the management of anxiety. JAMIA Open 2023 Oct 17;6(4):ooad088 [FREE Full text] [doi: 10.1093/jamiaopen/ooad088] [Medline: 37860603]
68. McCall T, Foster M, Schwartz TA. Attitudes toward seeking mental health services and mobile technology to support the management of depression among Black American women: cross-sectional survey study. J Med Internet Res 2023 Jul 19;25:e45766 [FREE Full text] [doi: 10.2196/45766] [Medline: 37467027]
69. Williams KD, Wijaya C, Stamatis CA, Abbott G, Lattie EG. Insights into needs and preferences for mental health support on social media and through mobile apps among Black male university students: exploratory qualitative study. JMIR Form Res 2022 Aug 31;6(8):e38716 [FREE Full text] [doi: 10.2196/38716] [Medline: 36044261]
70. Ramos G, Aguilera A, Montoya A, Lau A, Wen CY, Cruz Torres V, et al. App-based mindfulness meditation for people of color who experience race-related stress: protocol for a randomized controlled trial. JMIR Res Protoc 2022 Apr 14;11(4):e35196 [FREE Full text] [doi: 10.2196/35196] [Medline: 35436228]
71. Aguilera A, Arévalo Avalos MR, Rosales K, Reyes Y, Hernandez-Ramos R, Ramos G, et al. Effectiveness-implementation hybrid trial of Spanish language, digital cognitive-behavioral therapy (dCBT) intervention for depression and anxiety - protocol for the SUPERA (SUpport from PEeRs to expand Access) study. Contemp Clin Trials 2024 Feb;137:107422 [FREE Full text] [doi: 10.1016/j.cct.2023.107422] [Medline: 38145715]
72. Eustis EH, LoPresti J, Aguilera A, Schueller SM. Cultural responsivity in technology-enabled services: integrating culture into technology and service components. J Med Internet Res 2023 Oct 03;25:e45409 [FREE Full text] [doi: 10.2196/45409] [Medline: 37788050]
73. Pederson AB. Management of depression in Black people: effects of cultural issues. Psychiatr Ann 2023 Mar;53(3):122-125 [FREE Full text] [doi: 10.3928/00485713-20230215-01] [Medline: 37396678]
74. Coyne CA, Wongsomboon V, Korpak AK, Macapagal K. "We have to figure it out ourselves": transfeminine adolescents' online sexual experiences and recommendations for supporting their sexual health and wellbeing. Front Reprod Health 2022 Jan 16;4:1034747 [FREE Full text] [doi: 10.3389/frph.2022.1034747] [Medline: 36726593]
75. Jozsa K, Kraus A, Korpak AK, Birnholtz J, Moskowitz DA, Macapagal K. "Safe behind my screen": adolescent sexual minority males' perceptions of safety and trustworthiness on geosocial and social networking apps. Arch Sex Behav 2021 Oct;50(7):2965-2980. [doi: 10.1007/s10508-021-01962-5] [Medline: 34581948]
76. Katapally TR. Smart indigenous youth: the smart platform policy solution for systems integration to address indigenous youth mental health. JMIR Pediatr Parent 2020 Sep 25;3(2):e21155 [FREE Full text] [doi: 10.2196/21155] [Medline: 32975527]
77. Vis C, Schuurmans J, Aouizerate B, Atipei Craggs M, Batterham P, Bührmann L, et al. Effectiveness of self-guided tailored implementation strategies in integrating and embedding internet-based cognitive behavioral therapy in routine mental health

XSL•FO
RenderX

care: results of a multicenter stepped-wedge cluster randomized trial. J Med Internet Res 2023 Feb 03;25:e41532 [FREE Full text] [doi: 10.2196/41532] [Medline: 36735287]
78. McCall T, Threats M, Pillai M, Lakdawala A, Bolton CS3. Recommendations for design of a mobile application to support management of anxiety and depression among Black American women. Front Digit Health 2022 Dec 23;4:1028408 [FREE Full text] [doi: 10.3389/fdgth.2022.1028408] [Medline: 36620185]
79. Ellis DM, Draheim AA, Anderson PL. Culturally adapted digital mental health interventions for ethnic/racial minorities: a systematic review and meta-analysis. J Consult Clin Psychol 2022 Oct;90(10):717-733. [doi: 10.1037/ccp0000759] [Medline: 36227330]
80. Spadaro B, Martin-Key NA, Bahn S. Building the digital mental health ecosystem: opportunities and challenges for mobile health innovators. J Med Internet Res 2021 Oct 13;23(10):e27507 [FREE Full text] [doi: 10.2196/27507] [Medline: 34643537]
81. Lyles CR, Aguilera A, Nguyen O, Sarkar U. Bridging the digital health divide: how designers can create more inclusive digital health tools. California Health Care Foundation. 2022. URL: https://www.chcf.org/wp-content/uploads/2022/02/BridgingDigitalDivideDesigners.pdf [accessed 2024-10-09]

Abbreviations

CBPR: community-based participatory research
COREQ: Consolidated Criteria for Reporting Qualitative Research
DMH: digital mental health
HCD: human-centered design
HURE: historically underserved racial and ethnic minority
IR: implementation research
LGBTQIA+: lesbian, gay, bisexual, transgender, queer, intersex, and asexual

Edited by A Mavragani; submitted 11.10.24; peer-reviewed by E Rubenstein, A Jolliff; comments to author 19.11.24; revised version received 17.12.24; accepted 22.12.24; published 13.03.25.

Please cite as:
Knapp AA, Cohen K, Kruzan KP, Kornfield R, Herrera M, Pederson AB, Lee S, Macapagal K, Roulston CA, Clarke K, Wijaya C, Simmons R, Jackson L, Kour S, Franco S, Mohr DC
Teen Perspectives on Integrating Digital Mental Health Programs for Teens Into Public Libraries (“I Was Always at the Library”): Qualitative Interview Study
JMIR Form Res 2025;9:e67454
URL: https://formative.jmir.org/2025/1/e67454
doi:10.2196/67454
PMID:

©Ashley A Knapp, Katherine Cohen, Kaylee P Kruzan, Rachel Kornfield, Miguel Herrera, Aderonke B Pederson, Sydney Lee, Kathryn Macapagal, Chantelle A Roulston, Kaleigh Clarke, Clarisa Wijaya, Robert Simmons, Latonia Jackson, Simrandeep Kour, Sandra Franco, David C Mohr. Originally published in JMIR Formative Research (https://formative.jmir.org), 13.03.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Use of Video Consultation Between 2017 and 2020 in Outpatient Medical Care in Germany and Characteristics of Their User Groups: Analysis of Claims Data

Theresa Hüer[1], Dr rer pol; Anke Walendzik[1], Dr rer medic; Lara Kleinschmidt[1], MSc; Klemens Höfer[1], MA; Beatrice Nauendorf[2], MSc; Juliane Malsch[2], MSc; Matthias Brittner[3], MA; Paul Brandenburg[4], MA; André Aeustergerling[5], MA; Udo Schneider[6], PD Dr rer pol; Anja Wadeck[6], MA; Sebastian Liersch[7], Dr PH; Stephanie Sehlen[7], Dr rer pol; Katharina Schwarze[8], MA; Jürgen Wasem[1], Prof Dr

[1]Institute for Health Care Management and Research, University of Duisburg-Essen, Essen, Germany
[2]Kassenärztliche Vereinigung Berlin, Berlin, Germany
[3]Kassenärztliche Vereinigung Westfalen-Lippe, Dortmund, Germany
[4]Kassenärztliche Vereinigung Schleswig-Holstein, Bad Segeberg, Germany
[5]Kassenärztliche Vereinigung Mecklenburg-Vorpommern, Schwerin, Germany
[6]Techniker Krankenkasse, Hamburg, Germany
[7]AOK Nordost, Potsdam, Germany
[8]AOK NordWest, Dortmund, Germany

Corresponding Author:
Theresa Hüer, Dr rer pol
Institute for Health Care Management and Research
University of Duisburg-Essen
Thea-Leymann-Straße 9
Essen, 45127
Germany
Phone: 49 201 183 2592
Email: theresa.hueer@medman.uni-due.de

Abstract

Background: Supplementing outpatient medical care with the use of video consultations could, among other benefits, improve access, especially in structurally disadvantaged areas.

Objective: This claims data analysis, carried out as part of the German research project “Preference-based use of video consultation in urban and rural regions,” aimed to analyze the use of video consultations and the characteristics of its user groups.

Methods: Claims data from 3 Statutory Health Insurance Funds (SHIFs) and 4 Associations of Statutory Health Insurance Physicians (ASHIPs) from the period April 2017 to the end of 2020 were used. Data from a sample of about 6.1 million insured and 33,100 physicians and psychotherapists were analyzed. In addition to data on the use of video consultations, patient data on sociodemographic characteristics, diagnoses, and place of residence were included. To analyze the physicians’ perspectives, specialty groups, demographic characteristics, and the type of practice location were also included. In consideration of the principles of data economy and the fact that data analysis represents merely a preliminary phase within the broader project, the SHIFs and ASHIPs transmitted aggregated data (cross-tabulations per subgroup analysis) to the evaluator. For this reason, the analyses were constrained to a comparison of video consultation users versus nonusers, differentiated according to the aforementioned subgroups. Furthermore, the association between place of residence or type of region of the practice location and the use of video consultation was examined. A significance level of $P<.05$ was set for chi-square tests.

Results: From 2017 to 2019, almost no video consultations were used in outpatient care in the German health care system. Although this changed considerably in relative terms with the start of the COVID-19 pandemic (but still at a very low absolute level), there was also a clear decline in the use of video consultations as the number of infections flattened out. Physicians working in psychotherapy and psychological psychotherapists used video consultations with around 16% (44,808/282,530) of their treatment cases in the second quarter of 2020, followed by psychotherapists using video consultations for children (10,828/113,293, 10%). Although the absolute number of treatment cases with video consultations among general practitioners was very high compared

XSL•FO
RenderX

with other specialist groups, their share of video consultations in all treatment cases was very low at 0.3% (29,600/9,837,118). Younger age groups and those located in urban areas used video consultations more frequently; this applies to both patients (age groups: χ^2_7=9903.2, *P*<.001; region types: χ^2_2=3746.2, *P*<.001) and service providers (age groups: χ^2_3=11,338.2, *P*<.001; region types: χ^2_2=8474.1, *P*<.001).

Conclusions: The current use of video consultations is below its potential in terms of scope and user groups. The widespread and lasting use of video consultations will only succeed if the potential user groups accept this form of service provision and recognize its advantages. Further analyses (both qualitative, such as focus group discussions, and quantitative, such as preference surveys) should therefore investigate the preferences of user groups for the use of video consultations.

International Registered Report Identifier (IRRID): RR2-10.2196/50932

(JMIR Form Res 2025;9:e60170) doi:10.2196/60170

KEYWORDS

video consultation; outpatient medical care; user groups; claims data analysis; Germany; physician; psychotherapist; sociodemographic; healthcare; digital health; digital consultation; telehealth; telemonitoring; telemedicine

Introduction

The implementation of video consultations in the German Social Health Insurance (SHI) system in April 2017 was intended to establish a basis for providing outpatient medical services regardless of location and thus also in structurally disadvantaged regions in Germany. Older people and those with limited mobility were to be given easier access to medical services. Video consultations can be used in a variety of ways and therefore offer potential benefits for different patient groups. Particularly in the case of long travel distances to the nearest (specialist) physician, for follow-up appointments, or for the provision of repeated prescriptions, video consultations can be a useful tool, and the insured do not have to visit a doctor's office for every appointment. In the field of psychotherapy, therapeutic or certain diagnostic sessions can also take place in the form of video consultations [1-3].

In the beginning, video consultations in Germany could only be provided by a few specialist groups and only for selected indications and follow-up appointments. In the second quarter of 2019, these restrictions were mostly abolished, and the assessment of the appropriateness of a diagnosis or treatment via video consultation was the responsibility of the respective physician. The remuneration for video consultations in outpatient medical care within the German SHI system is primarily based on age-differentiated quarterly flat rates. These rates cover the entire medical care of a patient, not just video consultations, provided by the attending physician during the specified period. If a patient is treated exclusively via video in a quarter and there are no other personal physician-patient contacts, the quarterly flat rate is reduced by a discount specific to the specialist group. In addition to the quarterly flat rates per patient, there are supplementary fee schedule positions (FSPs) that compensate for the additional technical effort (€4.60 [US $4.77] per video consultation, FSP 01450) and the digital authentication of an unknown patient in video consultations (€1.15 [US $1.19] per video consultation with a previously unknown patient, FSP 01444). From October 2019 to the end of September 2021, there was also an incentive grant (at that time approximately €10 [US $10.36] per video consultation for a maximum of 50 video consultations per physician) to stimulate the use of video consultations [1].

Despite the potential of video consultations for different user groups (patients as well as physicians and psychiatrists), financing by the German SHI system since 2017, and support measures such as incentive payments, video consultations do not yet appear to have established themselves as an integral part of outpatient medical care. In order to investigate the underlying obstacles, it is important to examine the actual use of video consultations on a solid data basis in detail in order to identify suitable groups and occasions for the development of measures to promote video consultations.

The German committee for negotiating outpatient medical remuneration in the SHI system (“Bewertungsausschuss Ärzte”) published a report on the use of video consultations in 2018 and in an updated form in 2022. These reports show the development of the billing frequency of video consultations in outpatient medical care within the SHI system for the data period April 2017 to the end of 2021, differentiated according to the specialist groups as well as the age groups and gender of the insured [1,4]. In 2022, Gensorowsky et al [5] also published an article on the use of video consultations in Germany based on data from a nationwide health insurance fund, which also covered the billing frequency of video consultations for the period April 2017, but only up to the end of 2018. The analysis was differentiated according to the specialist groups of the physicians and the age groups and type of region in which insured persons who used video consultations live [5]. However, Gensorowsky et al [5] only examined the period up to the end of 2018 and were therefore unable to account for the effects of the COVID-19 pandemic. A more recent 2023 study by Petrick and Kreuzenbeck [6] considered data through mid-2020 but was based on data from only one Social Health Insurance Fund (SHIF) and therefore did not include information from the perspective of physicians as well as psychotherapists. Thus, recent studies are needed to investigate the characteristics of insured persons using the service. Additionally, it would be interesting to explore the specialist groups that offer video consultations. It is not possible to analyze the characteristics of physicians and psychotherapists (eg, provision of video consultations by age group, gender, region of practice location)

XSL•FO
RenderX

using data from SHIFs. This requires data from the Associations of Statutory Health Insurance Physicians (ASHIPs).

Against this background, this article dealt in particular with closing the aforementioned gaps in the evidence by analyzing not only claims data from SHIFs but also data from ASHIPs. The data period April 2017 to the end of 2020 was included in order to take the effects of the COVID-19 pandemic into consideration, a relevant extension of the observation period compared with the paper by Gensorowsky et al [5], which ended in 2018. In addition to the analysis of the frequency of video consultations, detailed characteristics of their user groups (users and insured persons and providers and physicians and psychotherapists) will also be examined. The age and gender distributions, employment status, and residence or type of region will be compared between users and nonusers of video consultations. In the case of service providers, specialty groups as well as their age and gender distributions, type of operation and employment, and the type of region of the practice location were considered. In particular, a detailed analysis of the association between place of residence or type of region of the practice location and the use of video consultations represents a significant contribution to the literature.

This claims data analysis was conducted as part of the German research project "Preference-based use of video consultations in urban and rural regions," which is funded by the Innovation Fund of the German Federal Joint Committee (funding number 01VSF20011). The research project addresses the question of how video consultations should be designed to enhance the quality of outpatient medical care while considering the preferences of insured individuals, physicians, and psychotherapists. A study protocol has been published elsewhere [7].

The objective of this data analysis was to examine the use of video consultations in Germany, with a particular emphasis on the impact of the COVID-19 pandemic, and to discern salient disparities between patient and physician subgroups. Our objective was to ascertain the prevalence of dependencies according to the following sociodemographic characteristics of insured persons: age group, gender, employment status, and place of residence. In addition, we wanted to investigate how video consultations were integrated into the overall treatment of a patient during the quarter; was this done exclusively via video, or was the video consultation only used as a supplement? For physicians and psychotherapists, the aim was to identify dependencies between video consultation use and the following variables: age group, gender, specialist group, practice type, and practice location.

Methods

Data

In this analysis, anonymized, aggregated data from 3 SHIFs (Techniker Krankenkasse, AOK Nordost, AOK NordWest) as well as 4 ASHIPs were used to reflect both the perspective of the insured or patients and that of the physicians and psychotherapists. Data from April 2017 to the end of 2020 from insured and physicians as well as psychotherapists who lived or practiced in 1 of 4 (of a total of 17) German regions ("Westphalia-Lippe," "Mecklenburg-Western Pomerania," "Schleswig-Holstein," and "Berlin") were analyzed. These regions were selected because they represent both rural and urban districts. They were classified as rural or urban districts using the interactive web application "INKAR - Indicators and Maps of Spatial and Urban Development" from the German Federal Institute for Research on Building, Urban Affairs and Spatial Development. Among other things, the settlement structure of the district is considered a combination of population density and the proportion of settlement area.

Aggregated data from about 6.1 million insured living in the 4 selected regions were analyzed. This corresponds to a share of approximately 8.4% of all SHI-insured persons in Germany (annual average in 2020=73,274,131) [8]. In addition to data on the use of video consultations, patient data on sociodemographic characteristics and diagnostic information were included. Of the 6.1 million insured persons included in the analysis, approximately 4.6 million (approximately 74%) had at least one outpatient treatment case on average across all included quarters ("users of outpatient services"). The group of outpatient service users is an important group in the analysis, as only they were faced with the actual decision of whether to contact a physician in person or via video. Table 1 summarizes the details of the SHIF data used and the characteristics of the insured who were included.

Table 2 shows the key information from the ASHIP data and the characteristics of the approximate 33,000 physicians and psychotherapists included. This corresponds to 18.4% (33,000/180,581 in 2020) of all SHI-accredited physicians and psychotherapists in Germany [9].

XSL•FO
RenderX

Table 1. Social Health Insurance Fund (SHIF) data.

Data type	Data
Source	3 SHIFs: Techniker Krankenkasse, AOK Nordost, AOK NordWest
Regions	4 of 17 regions in Germany: Berlin, Schleswig-Holstein, Mecklenburg-Western Pomerania, Westphalia-Lippe
Time period	Q2[a] 2017 until the end of 2020
Exclusion criteria	1. Insured living abroad, 2. Insured in employee health insurance in the respective SHIF, 3. Insured who refused the use of their data for research purposes
Insured persons (Q2 2017-Q4[b] 2020)	
Minimum (Q4 2017), n	5,915,199
Maximum (Q2 2020), n	6,318,851
Average	6,131,573
Insured persons with at least one treatment case (Q2 2017-Q4 2020), n (%)	
Minimum (Q2 2020)	4,277,149 (67.7)
Maximum (Q1[c] 2020)	4,761,242 (76.7)
Average	4,550,963 (74.2)
Gender (Q2 2020; n=4,277,148), n (%)	
Female (including diverse)	2,369,554 (55.4)
Male	1,907,594 (44.6)
Age (years; Q2 2020; n=4,277,148), n (%)	
≤20	742,967 (17.4)
21-30	447,669 (10.5)
31-40	511,062 (11.9)
41-50	478,659 (11.2)
51-60	717,920 (16.8)
61-70	576,561 (13.5)
71-80	429,612 (10)
≥81	372,698 (8.7)
Employment status (Q2 2020; n=4,229,406), n (%)	
Employed persons (employees/workers, self-employed persons)	2,047,604 (48.4)
Not employed persons (unemployed, social welfare recipients, emigrants, refugees, students)	967,533 (22.9)
Retired persons	1,214,269 (28.7)
Region type of the place of residence (Q2 2020; n=4,277,139), n (%)	
Urban regions	2,295,715 (53.7)
Mixed regions	1,379,654 (32.2)
Rural regions	601,770 (14.1)

[a]Q2: second quarter.

[b]Q4: fourth quarter.

[c]Q1: first quarter.

XSL•FO
RenderX

Table 2. Association of Statutory Health Insurance Physicians (ASHIP) data.

Data type	Data
Source	All SHI[a] physicians and psychotherapists in 4 ASHIP regions: Berlin, Schleswig-Holstein, Mecklenburg-Western Pomerania, Westphalia-Lippe
Regions	4 of 17 regions in Germany: Berlin, Schleswig-Holstein, Mecklenburg-Western Pomerania, Westphalia-Lippe
Time period	Q2[b] 2017 until the end of 2020
Exclusion criteria	Specialist groups that are not allowed to provide video consultations (eg, laboratory physicians)
Physicians (Q2 2017-Q4[c] 2020)	
Minimum (Q2 2017), n	32,128
Maximum (Q4 2020), n	34,127
Q2 2020, n	33,758
Average	33,231
Gender (Q2 2020; n=33,758), n (%)	
Female (including diverse)	14,930 (44.2)
Male	18,828 (55.8)
Age (years; Q2 2020; n=33,758), n (%)	
≤40	2516 (7.5)
41-50	7940 (23.5)
51-60	12,840 (38)
≥61	10,462 (31)
Aggregated specialist groups (Q2 2020; n=33,765), n (%)	
Primary care	11,988 (35.5)
Psychotherapy, psychiatry, neurology	8663 (25.7)
Gynecology	2492 (7.4)
Orthopedics	1459 (4.3)
Surgery	1391 (4.1)
Dermatology	791 (2.3)
Otorhinolaryngology and phoniatrics	947 (2.8)
Other specialist groups	6034 (17.9)
Type of practice (Q2 2020; n=33,782[d]), n (%)	
Individual practice	17,904 (53)
Joint practices (“Berufsausübungsgemeinschaft” and “Praxisgemeinschaft“)	10,738 (31.8)
Medical care center (“Medizinisches Versorgungszentrum [MVZ]“)	3585 (10.6)
Hospital physicians authorized for outpatient care	1555 (4.6)
Region type of the location of the physician’s office (Q2 2020; n=33,757), n (%)	
Urban regions	19,128 (56.7)
Mixed regions	10,086 (29.9)
Rural regions	4543 (13.5)

[a]SHI: Social Health Insurance.

[b]Q2: second quarter.

[c]Q4: fourth quarter.

[d]The sample size is higher than that of Q2 2020 overall, because physicians may be assigned to multiple practice types.

XSL•FO
RenderX

Finally, it is important to emphasize that, depending on which data set is used for analysis (SHIF data or ASHIP data), either the entire outpatient medical care of all SHI-insured persons with residence in the 4 regions is taken into account (ASHIP data) or only the medical care of those SHI-insured persons who are insured with 1 of the 3 SHIFs and live in the 4 regions (SHIF data) is taken into account.

Data Analysis

The German Federal Data Protection Act defines the principle of data avoidance and data minimization. Accordingly, the aim of the evaluation must always be taken into account when selecting and designing data processing systems. In light of this, a 2-step evaluation was carried out in this analysis: in the first step, a preliminary analysis of the individual data was carried out by the SHIFs and ASHIPs, and, in the second step, the aggregated data were merged and finally evaluated. This was to ensure that the evaluating body (University of Duisburg-Essen) only received the information necessary to answer the research questions.

Step 1: Preparatory Work and Preliminary Analysis of Personal Data at the SHIFs

In preparation for the data analysis, an evaluation concept and a data protection concept were first developed. Data forwarding agreements were also concluded bilaterally between the evaluating body and the 7 data providers for the forwarding of the aggregated data.

The developed evaluation concept comprised tables to be completed by all data providers on the use of video consultations, differentiated according to various characteristics of the insured (age, gender, employment status, residence or type of region) and of the physicians and psychotherapists (age, gender, specialist group, type of operation and employment, type of region of the practice location). This information was collected for users of video consultations as well as nonusers of video consultations. The frequency of video consultation use can be identified using the FSP 01450 (technical surcharge for video consultations), as this can be billed for every video contact. A limitation in the use of this FSP is the fact that there is a maximum billing quota for this FSP. In this respect, the use of this FSP requires the assumption that physicians will continue to use FSP 01450 even if they have already exhausted their billing quota per physician and quarter.

Due to the aforementioned (partial) flat-rate billing in outpatient medical care in Germany, which also includes video consultations, an evaluation of the specific services performed in video consultations is only possible to a limited extent. However, it is possible to evaluate certain individual services that are provided during video consultations that can be billed separately. This is possible because the fee schedule items are provided with a standardized national label if they are performed by video. A list of these services has been published by the National Association of Statutory Health Insurance Physicians [10].

Based on this, the preliminary analyses of the individual data were carried out by the SHIFs and ASHIPs.

On the basis of the data forwarding agreements, these were transmitted securely by each data supplier to the evaluating body.

Step 2: Merging and Final Evaluation of the Anonymized and Aggregated Data at the Evaluating Body

The aggregated data for each data supplier were initially merged by the evaluating body so that all further evaluations were based on the data from all SHIFs or all ASHIPs. Since the data were not provided at the individual level, the analysis of the aggregated data (only available as cross tables) was performed using Excel (version 16.0; Microsoft Corp).

The analyses included a comparison of video consultation users (VC users) versus nonusers, differentiated according to subgroups (insured: age group, gender, employment situation; physicians and psychotherapists: specialist group, age group, gender, type of operation, and employment). Furthermore, the association between the place of residence or type of region of the practice location and the use of video consultation was examined. To identify significant differences between subgroups, a significance level of $P<.05$ was set. Since the analysis was based on aggregated data (in the form of cross-tabulations, nominally scaled data), chi-square tests were used. Analyses based on individual and microdata (eg, regression analyses) were therefore excluded. However, this approach was chosen for reasons of data economy and because the data analysis was only a preparatory step in the overall project.

Ethical Considerations

The study was granted ethical approval by the Ethics Committee of the Medical Faculty of the University of Duisburg-Essen on September 27, 2022 (reference: 21-10283-BO). As the University of Duisburg-Essen only had access to aggregated data, there was no possibility of participant identification.

Results

Development of the Use of Video Consultations Over Time

From 2017 to 2019, video consultations played almost no role in outpatient medical care in Germany (quarters 2 and 3 [Q2 and Q3] 2017: 0; Q4 2017-Q3 2019: <300 video consultations; Q4 2019: 718 video consultations). Since the start of the COVID-19 pandemic, the use of video consultations has increased substantially. The billing frequency of FSP 01450 showed a huge increase in the first quarter of 2020 in all regions considered; this continued in the second quarter of 2020. This means that the number of video consultations carried out within 1 quarter increased approximately 6-fold from 44,680 (Q1 2020) to 271,483 (Q2 2020). The second quarter of 2020 covers the largest period of the first lockdown during the COVID-19 pandemic. In the summer months (Q3 2020), the use of video consultations decreased (109,683, which corresponds to only 40.4% [109,683/271,483] of the amount in Q2 2020), although the number of video consultations carried out then rose sharply again in the fourth quarter (192,405, which corresponds to

XSL•FO
RenderX

70.9% [192,405/271,483] of the amount in Q2 2020). This clearly shows that the use of video consultations not only increased sharply during the COVID-19 pandemic but also largely reflected the development of COVID-19 infections. In the period from April 2017 to the end of 2020, a total of 620,970 video consultations were held in the relevant regions in Germany.

When looking at the absolute figures, however, it should be noted that physician contacts with video consultations still only accounted for a fraction of all contacts of this type. If we look at the SHIF data and examine which of the insured persons included had at least one physician contact ("users"; Q2 2020: 4,277,149/6,318,851, 67.7%), the proportion of those who had a video consultation was extremely low, with a maximum of 1% (43,139/4,227,149) in the second quarter of 2020, 0.3% (13,543/4,227,149) in the first quarter of 2020, 0.4% (19,989/4,227,149) in the third quarter of 2020, and 0.8% (34,557/4,227,149) in the fourth quarter of 2020.

By examining the number of video consultations used per VC user, we see that, in the second quarter of 2020, 50% of VC users (21,569/43,139) had a maximum of 1 video consultation. The mean value per VC user was significantly higher, at 2.5 video consultations, which is due to the fact that some VC users made use of a large number of video consultations; the maximum here was 38 video consultations in the second quarter of 2020.

Characteristics of VC User Groups (Patients)

Overview of Analysis of VC User Groups

In the following sections, the results of the comparison between VC users and nonusers in terms of sociodemographic information are presented in order to investigate whether they differed systematically from each other. In addition, the association between the patient's place of residence and the use of video consultations was also analyzed. The underlying population for all analyses was the insured who had at least one outpatient medical contact in the second quarter of 2020, because only they were faced with the decision of a video consultation or face-to-face contact. The second quarter of 2020 was selected for analysis because, as previously described, the number of video consultations was highest during this period, allowing for subgroup analyses despite the absolute small number of VC users.

Differences by Age Group

Table 3 shows, based on the example of the second quarter of 2020, that video consultations were mainly used by young to middle-aged insured persons (20-50 years old). The largest group was those aged 31 years to 40 years, with a share of 1.7% (9378/563,715) of all insured in this age group, with at least one outpatient contact with a doctor (users). The correlation was statistically significant (χ^2_7=9903.2, *P*<.001).

Table 3. Insured with at least one video consultation (VC) by age group (second quarter 2020), based on Social Health Insurance Fund data (χ27=9903.2, *P*<.001).

≥1 VC	Age (years), n (%)							
	≤20 (n=840,922)	21-30 (n=506,117)	31-40 (n=563,715)	41-50 (n=505,860)	51-60 (n=718,517)	61-70 (n=551,065)	71-80 (n=397,565)	≥81 (n=334,520)
Yes	6600 (0.8)	7287 (1.4)	9378 (1.7)	6304 (1.3)	7140 (1)	3036 (0.6)	1589 (0.4)	1805 (0.5)
No	834,322 (99.2)	498,830 (98.6)	554,337 (98.3)	499,556 (98.7)	711,377 (99)	548,029 (99.4)	395,976 (99.6)	332,715 (99.5)

Differences by Gender

Table 4 shows that video consultations were used disproportionately more often by women in the second quarter of 2020. The correlation was statistically significant (χ^2_1=1533.8, *P*<.001).

Table 4. Insured with at least one video consultation (VC) by gender (second quarter 2020), based on Social Health Insurance Fund data (χ21=1533.8, *P*<.001).

≥1 VC	Women (n=2,369,554), n (%)	Men (n=1,907,594), n (%)
Yes	27,922 (1.2)	15,217 (0.8)
No	2,341,632 (98.8)	1,892,377 (99.2)

Employment Status

Table 5 shows that, for the second quarter of 2020, video consultations were used disproportionately often by employed people than people not in employment and retired persons (χ^2_2=3566.2, *P*<.001).

XSL•FO
RenderX

Table 5. Insured with at least one video consultation (VC) by employment status (second quarter 2020), based on Social Health Insurance Fund data (χ22=3566.2, *P*<.001).

≥1 VC	Employed persons (employees/workers, self-employed persons; n=2,150,544), n (%)	Not employed persons (unemployed, social welfare recipients, emigrants, refugees, students; n=1,096,458), n (%)	Retired persons (n=1,123,538), n (%)
Yes	25,729 (1.2)	8737 (0.8)	7157 (0.6)
No	2,124,815 (98.8)	1,087,721 (99.2)	1,116,381 (99.4)

Place of Residence of the Insured

Table 6 shows that video consultations were used disproportionately more often by the insured living in urban regions in the second quarter of 2020 (χ^2_2=3746.2, *P*<.001). Considering that a central reason for the introduction of video consultations was to improve access in rural areas, this is striking.

Table 6. Insured with at least one video consultation (VC) by place of residence (second quarter 2020), based on Social Health Insurance Fund data (χ22=3746.2, *P*<.001).

≥1 VC	Urban regions (n=2,380,202), n (%)	Mixed regions (n=1,428,651), n (%)	Rural regions (n=609,420), n (%)
Yes	29,184 (1.2)	10,878 (0.8)	3077 (0.5)
No	2,351,018 (98.8)	1,417,773 (99.2)	606,343 (99.5)

In addition, we investigated how video consultations were integrated into the overall treatment of a patient during the quarter; was this done exclusively via video, or was the video consultation only used as a supplement? The exclusive use of video consultations for the treatment of a patient by a physician in a quarter can be identified by a mandatory flag, which leads to a deduction from the physician's quarterly flat rate. Overall, video consultations were used as a supplement in 80% (34,525/43,139) of cases. If this is broken down by region, major differences become apparent. VC users in rural regions used them disproportionately more often exclusively (2001/3077, 65%) compared with those living in urban (4508/29,284, 15.4%) or mixed (2105/10,878, 19.4%) regions (see Table 7).

Table 7. Exclusive or only supplementary use of video consultations (VCs) by place of residence (second quarter 2020), based on Social Health Insurance Fund data (χ22=4285.8, *P*<.001).

Type of VC use	Urban regions (n=29,184), n (%)	Mixed regions (n=10,878), n (%)	Rural regions (n=3077), n (%)	Total (n=43,139), n (%)
VC exclusively	4508 (15.4)	2105 (19.4)	2001 (65)	8614 (20)
VC supplementary	24,676 (84.6)	8773 (80.6)	1076 (35)	34,525 (80)

Video consultations were therefore used less frequently in rural regions, but when they are used in these regions, in 65% (2001/3077) of cases, there was no further personal doctor-patient contact during the quarter.

Characteristics of the Provider Groups (Physicians and Psychotherapists)

Overview of Analysis of the Provider Groups

In the following sections, the results of the comparisons between video consultation providers (VC providers) and nonproviders are presented based on ASHIP data with regard to demographic information, information on their specialist group and practice, and employment type. In addition, we analyzed whether there was an association between the location of the medical practice and the availability of video consultations. The comparison was made on the basis of treatment cases provided by the respective physicians and psychotherapists, with or without the use of video consultations.

Differences by Age Group

Table 8 shows that, for the second quarter of 2020, physicians and psychotherapists in younger age groups had more treatment cases with video consultations than older physicians and psychotherapists. The largest group in relative numbers was those younger than 40 years (12,680/1,235,393, 1%), while only 0.4% (22,766/6,374,630) of treatment cases with video consultations were performed among those older than 60 years. The correlation was statistically significant (χ^2_3=11,338.2, *P*<.001).

Table 8. Treatment cases with and without video consultations (VCs) by physicians or psychotherapists and their age (second quarter 2020), based on Association of Statutory Health Insurance Physicians data (χ23=11,338.2, *P*<.001).

Use of VC	Age (years), n (%)			
	≤40 (n=1,235,393)	41-50 (n=5,480,427)	51-60 (n=10,390,058)	≥61 (n=6,374,630)
With VC	12,680 (1)	28,287 (0.5)	42,208 (0.4)	22,766 (0.4)
Without VC	1,222,713 (99)	5,452,140 (99.5)	10,347,850 (99.6)	6,351,864 (99.6)

XSL•FO
RenderX

Differences by Gender

The comparison of service providers by gender showed that female physicians or psychotherapists used video consultations with treatment cases slightly more frequently than men (see Table 9).

Table 9. Treatment cases with or without video consultations (VCs) by physicians or psychotherapists and their gender (second quarter 2020), based on Association of Statutory Health Insurance Physicians data ($\chi 21=2492.6$, $P<.001$).

Use of VC	Women (n=9,053,887), n (%)	Men (n=13,274,371), n (%)
With VC	51,677 (0.6)	55,979 (0.4)
Without VC	9,002,210 (99.4)	13,218,392 (99.6)

Information on the Specialist Group

A comparative analysis of the professional groups showed that physicians and psychotherapists working in the field of psychotherapy (for both adults and children) were the most likely to use video consultations. Physicians working in psychotherapy and psychological psychotherapists used video consultations with 15.9% (44,808/282,530) of their treatment cases in the second quarter of 2020, followed by psychotherapists for children, who used video consultations with approximately 10% (10,828/113,293, 9.6%) of their treatment cases. Although the absolute number of treatment cases with video consultations among general practitioners was very high compared with other specialist groups (29,600 compared with 10,828 among psychotherapists for children), their share of video consultations with all treatment cases in this specialist group was very low, at 0.3% (29,600/9,837,118). This is caused by the high number of general practitioners and treatment cases in this group.

Type of Practice

Table 10 shows that physicians or psychotherapists in individual practices (73,494/11,089,465, 0.7% of treatment cases) used video consultations 2.4 times more frequently than those in joint practices (23,251/8,453,284, 0.3% of cases); this includes different forms of joint practices.

Table 10. Treatment cases with or without video consultations (VCs) by physicians or psychotherapists and their type of practice (second quarter 2020), based on Association of Statutory Health Insurance Physicians data ($\chi 22=14,990.5$, $P<.001$).

Use of VC	Individual practice (n=11,162,959), n (%)	Joint practices (n=8,476,535), n (%)	Medical care center (n=2,485,753), n (%)
With VC	73,494 (0.7)	23,251 (0.3)	10,423 (0.4)
Without VC	11,089,465 (99.3)	8,453,284 (99.7)	2,475,330 (99.6)

Location of the Physician's Office

As with the situation for insured persons, Table 11 shows that physicians and psychotherapists in urban regions used video consultations 2.3 times more frequently with treatment cases than those in rural regions.

Table 11. Treatment cases with or without video consultations (VCs) by physicians or psychotherapists and their practice location (second quarter 2020), based on Social Health Insurance Fund data ($\chi 22=8474.1$, $P<.001$).

Use of VC	Urban regions (n=12,153,573)	Mixed regions (n=6,870,939)	Rural regions (n=3,303,069)
With VC	72,982 (0.6)	26,136 (0.4)	8538 (0.3)
Without VC	12,080,591 (99.4)	6,844,803 (99.6)	3,294,531 (99.7)

Services and Diagnoses Associated With Video Consultations

In 2020, a total of 715,104 services were identified as having been provided and billed as part of video consultations. Approximately 55% (392,380/715,104, 54.9%) of these were claimable individual psychotherapy services, of which 54% (211,788/392,380) were behavioral psychotherapy, 30.3% (118,986/392,380) were depth psychology psychotherapy, and 15.7% (61,606/392,380) were analytic psychotherapy. Of the 715,104 services provided via video consultation, 30% (214,411/715,104) were counseling services, and about 15% (106,580/715,104, 14.9%) were other psychotherapeutic services. A more detailed analysis showed that the majority of counseling services were psychotherapeutic consultations (168,388/214,411, 78.5%) and only a few were primary care consultations (40,418/214,411, 18.9%).

The analysis of diagnoses (according to the International Statistical Classification of Diseases and Related Health Problems, 10th revision, German Modification) in treatment cases (quarterly reference) with contacts that took place exclusively via video consultation showed a similar picture. The top 10 diagnoses in 2020 included, in particular, diagnoses related to mental and behavioral disorders: reactions to severe stress and adjustment disorders (ICD code F43), depressive episode (F32), other anxiety disorders (F41), recurrent depressive disorder (F33), somatoform disorders (F45), and phobic disorders (F40). In addition, the top 10 most frequently coded diagnoses included back pain (M54), essential (primary) hypertension (I10), and COVID-19 pandemic–related indications such as acute upper respiratory tract infection (J06).

XSL•FO
RenderX

Discussion

Principal Findings and Interpretation of Underlying Reasons

In the years 2017 to 2019, almost no video consultations were used in outpatient medical care in the German health care system. This changed significantly with the onset of the COVID-19 pandemic: In the first and second quarters of 2020, there was a sharp increase in the use of video consultations in all regions analyzed. However, even in the second quarter, the proportion of the insured with at least one video consultation was extremely low, at a maximum of 1%. There was a decline from the second to the third quarter, with a further significant increase in the fourth quarter—a trend in use that is in line with the COVID-19 infection rate.

Video consultations were used significantly more often by younger insured persons, women, employed persons, and the insured living in urban areas. These findings are in line with the results of other studies stating that the younger and employed generation as well as people living in urban environments are more interested in using video consultations than older age groups or people residing in rural areas [3]. Another reason could be that physicians act as gatekeepers and preselect patients, assuming that younger people and people from urban areas with presumably better internet access are more suitable for video consultations [11]. The picture is similar for physicians and psychotherapists, with younger age groups and physicians and psychotherapists with an urban-based practice also using video consultations more frequently. In order to promote the sustainable use of video consultations, it is important to identify the key barriers for those groups. A lack of technical requirements has often been cited as a key barrier for patients as well as physicians and psychotherapists [12-14]. Physicians were also concerned about negative effects on the doctor-patient relationship. In addition, some believed that the lack of physical examination methods could more easily lead to treatment errors or misdiagnosis. Concerns were also expressed about privacy and data security [11,15].

There was no evidence of increased use in structurally disadvantaged rural areas, an often-cited potential for video consultations [3]. A number of hypotheses may be posited to explain the underlying reasons for the observed phenomenon. For instance, the comparatively poorer broadband infrastructure in rural areas or a different age structure could be identified as potential explanatory factors. The population in rural regions is, on average, older and therefore potentially less digitally proficient than that in urban regions. In addition, people from urban regions tend to use psychotherapy, which is the predominant use case for video consultations, more often than people from rural regions [5].

Psychotherapists and physicians working in a psychotherapy specialty (both for adults and children) were, by far, the most frequent users of video consultations. This was also reflected in the analysis of services and diagnoses in video consultations. The reason for this may be the nature of the services in this area, which are primarily what is known as “talking medicine.” Somatic specialists generally rely on physical examinations, blood samples, for example, as a supplement.

Video consultations are primarily used as an adjunct to care and rarely occur without further doctor-patient contact. However, in rural areas, where video consultations are less common, the opposite is true: There is often no face-to-face contact in addition to video consultations. One reason for this could be that these patients used a video consultation with a central platform rather than with their usual physician. On these platforms, personal follow-up contact is usually not possible at all, as physicians from completely different regions of Germany can also offer video consultations there. This may be because their usual physician does not offer video consultations or because there are no physicians in the area with vacancies in the specialty they are looking for. In any case, this is an indicator of a potential lack of care in rural areas.

Strengths and Limitations

This data analysis was based on data from 3 large German SHIFs with a wide range of enrollees (about 6.1 million insured or approximately 8.4% of all SHI-insured) and 4 ASHIPs (about 33,000 physicians and psychotherapists or 18.4% of all SHI-accredited physicians and psychotherapists in Germany). It is also worth noting that insured persons from different types of SHIFs were considered, so that they do not represent only one type of SHIF (eg, in terms of age and gender). The selection of the 4 regions ensured that both rural and urban enrollees and providers were included.

A first limitation could be that only those with statutory health insurance were included, not those with private health insurance, which accounts for only about 10% of all insured persons in Germany.

The time period included allowed for an analysis of the impact of the COVID-19 pandemic on the use of video consultations. The second limitation is that this period could also have had a biasing effect, for example in the analysis of diagnoses coded in relation to video consultations, given the extremely high coding of respiratory diseases in the second quarter of 2020. A slightly longer data period would also have been desirable to analyze the effects after the pandemic. However, the data period used, in particular the year 2020, was considered sufficient for the analysis of the characteristics of the user groups.

Third, the statistical evaluation was hindered by a number of evident challenges. One principal reason for this was that, for reasons of data minimization, only aggregated data were made available to the evaluator, which significantly constrained the range of analysis options. Consequently, only chi-square analysis and, for example, no regression analyses could be conducted. It should be noted, however, that even minor discrepancies can become statistically significant due to the large sample size. To ensure transparency, the absolute and relative data were consistently provided in the cross tables, allowing readers to assess them independently. Additionally, the study faced a challenge with balancing the proportions, with approximately 1% of the sample using a video consultation and the remaining 99% not using it. This imbalance can complicate the analysis.

XSL•FO
RenderX

Fourth, not all variables are available in health insurance data; for example, information on educational status would have been useful. However, in this mixed methods project, a survey was conducted to ask, for example, about the use of video consultations and the level of education [8]. The findings will be published separately.

Fifth, the generalizability of the results for Germany to other countries may be limited. Aspects such as the slow progress of digitization in Germany, as well as the existence of statutory health insurance and a relatively high density of physicians compared with other countries, could explain the hesitant acceptance of video consultations and may not apply to other countries.

Conclusion and Potential for Future Research

The current extent of video consultation use remains well below its potential. Additionally, it is notable that the groups for whom it was anticipated that video consultations would offer substantial benefits, such as individuals with limited mobility and those residing in rural areas, currently use video consultations at a significantly lower rate than groups who are currently considered to be well served, including individuals in urban areas and young people.

Unfortunately, initial analyses of a period at the end of the pandemic are concerning. According to these data, the already low uptake rate declined further toward the end of the pandemic. According to a report by BARMER, a German SHIF, monthly uptake in the second quarter of 2022 was more than 50% lower than during the pandemic waves in the second quarter of 2020 and the first quarter of 2021. It appears to have stabilized at a relatively low level [16].

The widespread and lasting use of video consultations will only succeed if the potential user groups accept this form of service provision and recognize its advantages. Further analyses should therefore investigate the preferences of user groups for the use of video consultations. Both qualitative (interviews, focus group discussions) and quantitative research approaches (surveys [eg, with preference elicitation methods such as discrete choice experiments]) are suitable for this purpose. As part of this study, a survey was conducted among the insured, physicians, and psychotherapists to identify the primary barriers to the use of video consultations [7]. The findings will be released in a separate publication.

Acknowledgments

This study was entirely funded by the Innovation Fund of the German Federal Joint Committee (funding number 01VSF20011). The funder did not influence the design of the study, the collection, analysis, and interpretation of data or the writing process of this manuscript. We acknowledge support by the Open Access Publication Fund of the University of Duisburg-Essen.

Data Availability

The data sets on which the routine data analysis is based are not publicly available for privacy reasons.

Authors' Contributions

All authors contributed to the manuscript at various stages and read and approved the final version. TH was the main contributor in all 14 steps according to the Credit taxonomy.

Conflicts of Interest

None declared.

References

1. Bericht des Bewertungsausschusses und des ergänzenden Bewertungsausschusses zur telemedizinischen Leistungserbringung im Einheitlichen Bewertungsmaßstab (BT-Drucksache 20/4982). Deutscher Bundestag. 2022. URL: https://dserver.bundestag.de/btd/20/049/2004982.pdf [accessed 2024-12-16]
2. Consolidated telemedicine implementation guide. World Health Organization. 2022. URL: https://iris.who.int/bitstream/handle/10665/364221/9789240059184-eng.pdf [accessed 2024-12-16]
3. Waschkau A, Traulsen P, Steinhäuser J. Evaluation of synchronous and asynchronous telemedical applications in primary care in rural regions of northern Germany-results and lessons learned from a pilot study. Int J Environ Res Public Health 2022 Nov 11;19(22):14860 [FREE Full text] [doi: 10.3390/ijerph192214860] [Medline: 36429577]
4. Unterrichtung durch die Bundesregierung - Bericht des Bewertungsausschusses zur Überprüfung des Einheitlichen Bewertungsmaßstabes auf die Möglichkeit zur ambulanten telemedizinischen Leistungserbringung (BT-Drucksache 19/0620). Deutscher Bundestag. 2018. URL: http://dip21.bundestag.de/dip21/btd/19/060/1906020.pdf [accessed 2024-12-16]
5. Gensorowsky D, Surmann B, Schmidt J, Greiner W. [Level of use and user groups of online video consultations in outpatient medical care: analysis of claims data]. Gesundheitswesen 2022 Feb;84(2):110-116. [doi: 10.1055/a-1312-6439] [Medline: 33412593]

XSL•FO
RenderX

6. Petrick N, Kreuzenbeck CCJ. [Effects of the Covid-19 pandemic on online use of video consultation by general practitioners in Germany - a secondary data analysis of German health insurance data]. Gesundheitswesen 2023 Mar;85(3):188-192. [doi: 10.1055/a-1999-7390] [Medline: 36693392]
7. Kleinschmidt L, Walendzik A, Wasem J, Höfer K, Nauendorf B, Brittner M, et al. Preference-based implementation of video consultations in urban and rural regions in outpatient care in Germany: protocol for a mixed methods study. JMIR Res Protoc 2024 Apr 11;13:e50932 [FREE Full text] [doi: 10.2196/50932] [Medline: 38602749]
8. Gesetzliche Krankenversicherung. Mitglieder, mitversicherte Angehörige und Krankenstand. Jahresdurchschnitt 2020. Bundesministerium für Gesundheit. 2020. URL: https://www.bundesgesundheitsministerium.de/fileadmin/Dateien/3_Downloads/Statistiken/GKV/Mitglieder_Versicherte/KM1_JD_2020_CPS_bf.pdf [accessed 2024-12-16]
9. Federal Medical Register. Kassenärztliche Bundesvereinigung. 2020. URL: https://www.kbv.de/html/bundesarztregister.php#content6889 [accessed 2024-12-16]
10. Übersicht Vergütung für Videosprechstunde. Kassenärztliche Bundesvereinigung. 2024. URL: https://www.kbv.de/media/sp/Videosprechstunde__uebersicht_Verguetung.pdf [accessed 2024-12-16]
11. Cowan KE, McKean AJ, Gentry MT, Hilty DM. Barriers to use of telepsychiatry: clinicians as gatekeepers. Mayo Clin Proc 2019 Dec;94(12):2510-2523. [doi: 10.1016/j.mayocp.2019.04.018] [Medline: 31806104]
12. Thiyagarajan A, Grant C, Griffiths F, Atherton H. Exploring patients' and clinicians' experiences of video consultations in primary care: a systematic scoping review. BJGP Open 2020 Mar 17;4(1):bjgpopen20X101020. [doi: 10.3399/bjgpopen20x101020]
13. Donaghy E, Atherton H, Hammersley V, McNeilly H, Bikker A, Robbins L, et al. Acceptability, benefits, and challenges of video consulting: a qualitative study in primary care. Br J Gen Pract 2019 Sep;69(686):e586-e594 [FREE Full text] [doi: 10.3399/bjgp19X704141] [Medline: 31160368]
14. Christensen LF, Moller AM, Hansen JP, Nielsen CT, Gildberg FA. Patients' and providers' experiences with video consultations used in the treatment of older patients with unipolar depression: a systematic review. J Psychiatr Ment Health Nurs 2020 Jun 26;27(3):258-271. [doi: 10.1111/jpm.12574] [Medline: 31677331]
15. Mueller M, Knop M, Niehaves B, Adarkwah CC. Investigating the acceptance of video consultation by patients in rural primary care: empirical comparison of preusers and actual users. JMIR Med Inform 2020 Oct 22;8(10):e20813 [FREE Full text] [doi: 10.2196/20813] [Medline: 32969339]
16. Mangiapane S. Five years of video consultations in statutory health care: Has the pandemic led to a breakthrough? In: Repschläger U, editor. BARMER Gesundheitswesen. Berlin, Germany: BARMER Institute for Health Systems Research; 2023:82-99.

Abbreviations

ASHIP: Association of Statutory Health Insurance Physicians
FSP: fee schedule position
Q: quarter
SHI: Social Health Insurance
SHIF: Social Health Insurance Fund
VC: video consultation

Edited by A Schwartz; submitted 03.05.24; peer-reviewed by D Gensorowsky, K Moon, K De Santis, S Muellmann, E van der Velde, D Singh; comments to author 02.10.24; revised version received 20.10.24; accepted 27.01.25; published 14.03.25.

Please cite as:
Hüer T, Walendzik A, Kleinschmidt L, Höfer K, Nauendorf B, Malsch J, Brittner M, Brandenburg P, Aeustergerling A, Schneider U, Wadeck A, Liersch S, Sehlen S, Schwarze K, Wasem J
Use of Video Consultation Between 2017 and 2020 in Outpatient Medical Care in Germany and Characteristics of Their User Groups: Analysis of Claims Data
JMIR Form Res 2025;9:e60170
URL: https://formative.jmir.org/2025/1/e60170
doi:10.2196/60170
PMID:40085136

©Theresa Hüer, Anke Walendzik, Lara Kleinschmidt, Klemens Höfer, Beatrice Nauendorf, Juliane Malsch, Matthias Brittner, Paul Brandenburg, André Aeustergerling, Udo Schneider, Anja Wadeck, Sebastian Liersch, Stephanie Sehlen, Katharina Schwarze, Jürgen Wasem. Originally published in JMIR Formative Research (https://formative.jmir.org), 14.03.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/),

XSL•FO
RenderX

which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Development of a Clinical Decision Support Tool to Implement Asthma Management Guidelines in Pediatric Primary Care: Qualitative Study

David A Fedele[1], PhD; Jessica M Ray[2], PhD; Jaya L Mallela[3], MS; Jiang Bian[2], PhD; Aokun Chen[2], PhD; Xiao Qin[2], MS; Ramzi G Salloum[2], PhD; Maria Kelly[4], MD; Matthew J Gurka[5], PhD; Jessica Hollenbach[6,7], PhD

[1]Center for Healthcare Delivery Science, Nemours Children's Health, Jacksonville, FL, United States
[2]Department of Health Outcomes and Biomedical Informatics, University of Florida, Gainesville, FL, United States
[3]Department of Clinical & Health Psychology, University of Florida, Gainesville, FL, United States
[4]Department of Pediatrics, University of Florida, Gainesville, FL, United States
[5]Department of Public Health Sciences, University of Virginia, Charlottesville, VA, United States
[6]Asthma Center, Connecticut Children's Medical Center, Hartford, CT, United States
[7]Department of Pediatrics, UConn Health, Farmington, United States

Corresponding Author:
David A Fedele, PhD
Center for Healthcare Delivery Science
Nemours Children's Health
807 Children's Way
Jacksonville, FL, 32610
United States
Phone: 1 904 697 3600
Email: david.fedele@nemours.org

Abstract

Background: There is a longstanding gap between national asthma guidelines and their implementation in primary care. Primary care providers (PCPs) endorse numerous provider and practice or clinic-related barriers to providing guidelines-based asthma care. To reduce asthma morbidity in primary care, PCPs need access to tools that facilitate adherence to national guidelines, which can be delivered at the point of care, are minimally burdensome, and fit within the clinic workflow. Clinical decision support (CDS) tools are health IT systems that can be housed in the electronic health record (EHR) system.

Objective: This study aimed to follow user-centered design principles and describe the formative qualitative work with target stakeholders (ie, PCPs and IT professionals) to inform our design of an EHR-embedded CDS tool that adheres to recent, significant changes in asthma management guidelines.

Methods: Purposive sampling was used to recruit three separate subgroups of professionals (n=15) between (1) PCPs with previous experience using a paper-based CDS tool for asthma management, (2) PCPs without previous experience using CDS tools for asthma management, and (3) health care IT professionals. The PCP interview guide focused on their practice, familiarity with national asthma guidelines, and how a CDS tool embedded in the EHR might help them provide guideline-based care. The health care IT professional guide included questions on the design and implementation processes of CDS tools into the EHR. Qualitative data were audio-recorded, transcribed, and then analyzed using an inductive approach to develop themes.

Results: Themes were organized into 2 domains, current practice and CDS tool development. The themes that emerged from PCPs included descriptions of assessments conducted to make an asthma diagnosis, previous attempts or opportunities to implement updated national asthma guidelines, and how a CDS tool could be implemented using the EHR and fit into the current asthma management workflow. The themes that emerged from health care IT professionals included processes used to design CDS tools and strategies to collect evidence that indicated a tool's value to a practice and the broader health system.

Conclusions: In this study, user-centered design principles were used to guide a qualitative study on perceived barriers and facilitators to a primary care–based, EHR-integrated asthma CDS tool. PCPs expressed their interest in adopting an asthma CDS tool that was low burden and efficient but could help them adhere to national asthma guidelines and improve clinic workflow. Similarly, health care IT professionals perceived an asthma CDS tool to be useful, if it adhered to EHR design standards.

XSL•FO
RenderX

Implementation of a CDS tool to improve adherence of PCPs to recently updated national asthma guidelines could be beneficial in reducing pediatric asthma morbidity.

(JMIR Form Res 2025;9:e65794) doi:10.2196/65794

KEYWORDS

clinical decision support; asthma; primary care; guidelines; pediatric; asthma care; morbidity; health information technology; electronic health record; EHR; user-centered design; inductive approach; digital health; health technology

Introduction

Asthma is the most common childhood chronic illness in the United States, affecting >5 million children [1], and is a leading cause of emergency department and urgent care visits [2], school absences [3,4], and reduced quality of life among youth [5,6]. The National Asthma Education and Prevention Program (NAEPP) guidelines are designed to prevent and reduce pediatric asthma morbidity via disseminating state of the science guidance to health care providers on asthma diagnosis and management [7]. NAEPP guidelines call for health care providers to regularly adhere to 4 pillars of asthma management—assessment of asthma control, provision of asthma education to families (eg, written asthma treatment plans), assessment of environmental triggers, and following pharmacological therapy recommendations [4]. Notably, NAEPP guidelines were recently updated in 2020 to include single maintenance and reliever therapy (SMART), a paradigm-shifting change to recommendations for the management of youth with persistent asthma [7].

Use of NAEPP guidelines improves asthma control, reduces the likelihood of exacerbations, and decreases health care usage [4,7-9]. However, there is a persistent gap between NAEPP guidelines and their implementation in practice, particularly among primary care providers (PCPs) [10,11]. PCPs are the frontline health care providers for youth with asthma and provide most ambulatory care [12,13]. Suboptimal adherence among PCPs to national asthma guidelines is documented in numerous studies [10,12,14,15]. For example, a chart review study of a large cohort of primary care practices found that only 9% of PCPs used a validated asthma control tool, just 6% provided an asthma treatment plan, and 30% failed to prescribe anti-inflammatory medication when indicated [16]. This pattern of findings is mirrored in recent results from the National Asthma Survey of Physicians [12].

The considerable gap between NAEPP guidelines and their implementation in primary care are driven by provider and practice or clinic-related factors [12]. PCPs encounter numerous barriers to providing guidelines-based asthma care, including time constraints, lack of familiarity and knowledge of the guidelines, and low self-efficacy surrounding guideline implementation [17,18]. Furthermore, primary care clinics are commonly characterized as suffering themselves from “information chaos” with limited administrative bandwidth for comprehensive, evidenced-based interventions [19]. To reduce asthma morbidity in primary care, PCPs need access to tools that facilitate adherence to national guidelines, which can be delivered at the point of care, are minimally burdensome, and fit within the clinic workflow [15,20,21].

Clinical decision support (CDS) tools are health IT systems that can be housed in the electronic health record (EHR) system and be effective in improving provider adherence to guidelines and patient outcomes [22,23]. CDS tools assist providers in patient-level clinical decision-making (eg, asthma diagnosis and severity designation) at the point of care via collecting relevant data, synthesizing information, and providing real-time recommendations, thereby reducing provider burden and complexity surrounding adherence and implementation of guidelines [24]. CDS tools were shown to be the most promising method for modifying provider behavior in a systematic review of 68 intervention studies on health care provider adherence to national asthma guidelines [14]. There is a critical need for the creation of a CDS tool for pediatric asthma that attends to the needs of PCPs, especially in the current era of recently updated NAEPP guidelines [21].

While paper-based CDS tools have been shown to increase PCP adherence to NAEPP guidelines and reduce asthma-related morbidity including the number of outpatient visits, hospitalizations, and emergency department visits [25-27], the ubiquity of EHRs in delivery and documentation of primary care, including for asthma management [28], provides the opportunity to provide real-time decision support within the electronic workflow. In addition, electronically presented CDS may support better adherence due to automatic provision of guidance and patient-specific suggestions to support decision-making [29]. While recent work provides key guidance to support implementation of the recent guidelines into electronic CDS, further work is necessary to understand local contexts and clinician behaviors that will support successful intervention development and implementation [30]. This study describes our formative qualitative work with target stakeholders (ie, PCPs and IT professionals) to inform our design of an EHR-embedded CDS tool that adheres to the 4 pillars of asthma management and is congruent with recent, significant changes in NAEPP guidelines [7].

Methods

Overview

We used purposive sampling to recruit three separate subgroups of professionals between January and March 2023, that are (1) PCPs with previous experience using a paper-based CDS tool for asthma management, (2) PCPs without previous experience using CDS tools for asthma management, and (3) health care IT professionals. Participants were recruited through study announcement emails to department listservs and brief presentations to raise awareness of the study during department meetings. In total, 15 individuals agreed to participate. We conducted 2 focus groups with PCPs (n=5 and n=4 participants,

XSL•FO
RenderX

respectively), 1 health care IT professional focus group (n=2), 3 PCP individual interviews, and 1 health care IT professional individual interview. Some participants had a previous professional relationship with the study team (eg, had participated in previous studies). The first focus group was with PCPs (n=5) who had previous experience using Easy Breathing, a validated paper-based CDS tool for asthma management that translates NAEPP guidelines into a usable, efficient, and effective format [10,11,25-27,31]. The second focus group (n=4) and 3 individual interviews were with PCPs who did not have previous experience with CDS tools for pediatric asthma. The third focus group (n=2) and 1 individual interview were with health care IT professionals who had previous experience developing ambulatory CDS tools. Focus groups and interviews were conducted over Zoom (Zoom Video Communications) or in person in clinic workrooms. Focus groups and interviews were facilitated by a PhD-level, certified asthma educator with expertise in clinic- and school-based asthma management intervention (JH [female]) or a PhD-level clinical psychologist with expertise in mobile and eHealth asthma interventions (DAF [male]). Both study authors have experience in conducting qualitative interviews with families of children with asthma and medical providers. Participants were made aware of the goals of the study during the informed consent process and the interviewers briefly summarized their professional expertise in pediatric asthma. Participants were compensated US $50 for their time. In addition, a brief demographics survey was distributed during each focus group.

Focus groups and interviews were conducted in a semistructured format, with separate guides developed for PCPs and health care IT professionals. Sessions ranged in length from 25 to 65 minutes, with an average length of 38 minutes. The PCP interview guide included open-ended questions on how the provider currently diagnoses and treats asthma, if they use any validated tools for diagnosis or treatment planning, familiarity with NAEPP guidelines, how a CDS tool embedded in the EHR might help them provide guidelines-based care, and what functions could be useful in such a tool. The health care IT professional guide included questions on design and implementation processes of CDS tools into the EHR, best practices for collaboration with research and clinical teams, and strategies for evaluating and maintaining CDS tools across multiple clinics.

Focus groups and interviews were audio-recorded, transcribed, and then analyzed by 3 coders (DAF, JH, and JLM) using NVivo 14 (Lumivero). Field notes were taken during data collection. We used an inductive approach by coding data based on similar topics and content across transcripts, clustered by PCPs and health care IT professionals, to develop themes. Themes were derived from overlapping content between PCPs' data and included descriptions of assessments conducted to make an asthma diagnosis, previous attempts or opportunities to implement updated NAEPP guidelines, and how a CDS tool could be implemented using the EHR and fit into current asthma management workflow. Themes from overlapping content between health care IT professionals' data included discussion of processes used to design CDS tools (eg, initial assessment of required functionality, testing mock designs to evaluate alignment with PCP workflow) and strategies to collect evidence that indicated a tool's value to a practice and the broader health system. We then collapsed themes into 2 domains, Current Practice and CDS Tool Development. The 3 coders then used a final codebook that described the themes in each domain to analyze all transcripts (Table 1) independently. Codes for each transcript were then discussed and modified collectively until consensus was reached.

XSL•FO
RenderX

Table 1. Qualitative codebook for interviews with primary care providers and information (for technology health professionals).

Domain and code	Definition
Current practice	
PCPs[a] describe their asthma diagnosis process and workflow	Description of assessments PCPs conduct to make an asthma diagnosis.
PCPs describe asthma follow-up care	Descriptions of symptom monitoring at follow-up visits.
Barriers to asthma management	Mentions of barriers to making asthma diagnosis and to monitoring symptoms at follow-up visits.
Asthma action and treatment plans	Statements about how asthma treatment regimen is communicated to families.
Opportunities to implement new guidelines	Statements about barriers to implementing new guidelines and explanations of previous attempts to implement them (individual PCP or practice-wide).
CDS[b] tool development	
How can a CDS tool fit into provider workflow	Recommendations on components and functionality of ideal tool, including patient surveys, how responses are displayed for the provider, how treatment or prescription options are displayed for the provider.
How can a CDS tool be best designed	Recommendations on design strategies and examples of tools that were successful or unsuccessful in having the components or functionality that providers wanted and used.
Strategies to evaluate impact of a CDS tool	Explanations of how to gather evidence needed to demonstrate tool's value to a practice and the institution or health system. Mentions of pipeline through which projects move forward.

[a]PCP: primary care provider.

[b]CDS: clinical decision support.

Ethical Considerations

The institutional review board at the University of Florida approved all study procedures and measures (IRB 202202480). All individuals provided informed consent before participating in the current study and were reimbursed US $50 for their time. All study data presented below are deidentified.

Results

Participant Demographics

Across the 3 subgroups, 11 participants reported their gender as women and 4 reported as men. Participants reported their race as White (n=11), Black or African American (n=1), or Asian (n=3), and 2 reported their ethnicity as Hispanic or Latino-a-x. To account for providers with multiple racial or ethnic identities, participants could select more than 1 race or ethnicity response such that the total n and percentages summed to greater than 15 and 100%, respectively. PCPs using a CDS tool had an average of 14.6 (SD 13.9) years in practice, while those not using a CDS tool had an average of 16.1 (7.6) years (Tables 2 and 3).

Table 2. Demographics of study participants.

Characteristic	Participants, n (%)
Gender	
Men	4 (27)
Women	11 (73)
Race or ethnicity[a]	
White	11 (73)
Black or African American	1 (7)
Asian	3 (20)
Hispanic or Latino-a-x	2 (13)

[a]Participants could select more than one race or ethnicity response.

XSL•FO
RenderX

Table 3. Demographics of study participant subgroup.

Participant subgroup	Years in practice		
	n	Mean (SD)	Range
PCPs[a] with asthma CDS[b] tool experience	5	14.6 (13.9)	3-36
PCPs without asthma CDS tool experience	7	16.1 (7.6)	3.5-27
IT professionals	3	__[c]	—

[a]PCP: primary care provider.

[b]CDS: clinical decision support.

[c]Not available.

Current Practice for Asthma Management

Within the Current Practice domain, five themes were developed, that are (1) description of asthma diagnosis processes, (2) asthma follow-up care, (3) barriers to asthma management, (4) asthma action or treatment plans, (5) opportunities to implement updated NAEPP guidelines (refer to Table 1 for full codebook and definitions of themes). Quotations from participants in each of the three subgroups are identified as (1) PCPs A through E for those with asthma CDS tool experience, (2) PCPs 1 through 10 for those without asthma CDS tool experience, and (3) IT professionals 1 through 3.

Diagnosis Processes and Follow-Up Care

PCPs currently implementing a paper-based asthma CDS tool described a patient-reported survey to guide them in making an asthma diagnosis. They described assessing asthma control with the Asthma Control Test (ACT) during visits for specific patient ages (eg, 4 and 8 years) or on designated days of the week. Importantly, these PCPs discussed variable CDS tool implementation and workflows across clinics. Providers expressed the ease of conducting follow-up care using the ACT, which is one of the validated measures included in the CDS tool. The ACT is a validated measure of asthma control and gives providers an objective measure, making overall workflow more standardized and efficient.

> *It's much easier every time [to use a CDS tool]. Implementation science tells you that you know just when it's called for. That's part of why we do the Asthma Control [Test] at every visit. Instead of somebody having to figure out whether the patient should be diagnosed with asthma or not, and the symptom severity level.* [PCP C (with CDS tool experience)]

Alternatively, PCPs without experience using an asthma CDS tool described their diagnosis and follow-up processes as based primarily on symptom presentation**,** including recurrent cough, nighttime cough, chest tightness, shortness of breath, and wheezing, with some using structured observation to assess treatment response for acute exacerbations (ie, monitoring symptoms in clinic after administering albuterol) or level of control at follow-up visits. In total, 3 PCPs mentioned using recommendations, such as those published yearly by GINA (Global Initiative for Asthma) to inform their diagnostics.

> *The GINA guidelines…Yeah, it's what I know, and it's quite updated, and I like it, and I'm comfortable. But I will tell you it's not the easiest thing to do, especially in the clinic setup. Usually in my previous practice, we had a lot of things that we're clicking on, and you know it sort of generates for you at least a good estimate. And the reason why I say this because I only have 20 min to do, you know…Even if I suspect that this kid is asthma, after 10 minutes registering and getting roomed in, I only have 10 minutes to figure this out so.* [PCP 7 (without CDS tool experience)]

It was also apparent that PCPs without experience using an asthma CDS tool had knowledge of national and global asthma care recommendations but relied on a more general familiarity with recommendations rather than using a specific tool to aid in their adherence.

> *I probably should do something, but I usually just ask questions to kinda get an idea of how poorly they are controlled, based on how many times they are using it [quick-relief medication]. Basically going through the guidelines, but not as complete.* [PCP 10 (without CDS tool experience)]

Barriers to Asthma Management

All PCPs endorsed time and workflow management as barriers to managing asthma care and adhering to guidelines consistently and thoroughly.

> *I would say the main limitation for this and any guideline or anything else we do in clinic is just time. Cause there's so much to cover in such a short period of time, and trying to prioritize the most important things and spend the limited amount of time that we have on those. Because the most important thing is, albuterol whenever you're wheezing, and then for the kids that are on steroids, to use them consistently, to be able to emphasize that in a way that it will happen whenever they leave. So that's a hard thing to do.* [PCP 4 (without CDS tool experience)]

Among PCPs using a paper-based CDS tool, needing to transfer information from paper documentation into multiple EHR modules (eg, progress note, prescriptions, after-visit summary, and patient instructions) was reported as a barrier.

> *The scoring, writing your impression, especially the [patient]survey. And then you're gonna, you know, basically do the same thing again, or some similar variation of it in the note or in the chart for sure it's*

XSL•FO
RenderX

like it's definitely work. [PCP A (with CDS tool experience)]

Potentially missing asthma diagnoses due to the lack of a CDS tool at the point of care was also identified as a barrier to management.

Missing the possible, not the ones that are established. For the ones you're considering a new diagnosis. It's a little bit out of sight out of mind. So people aren't thinking about [the CDS tool] to administer the survey, and that's a little bit cumbersome to go and try to find it. [PCP A (with CDS tool experience)]

Although PCPs not using a CDS tool described prioritizing the most important aspects of management in their verbal communication with families, they also reported gaps potentially caused by a lack of standardized and accessible treatment plans. In addition, 3 of these PCPs noted insurance coverage of medication costs as influential on their prescribing decisions.

We typically just tell them the medications, and which one is which. You know it'd be nice if I had some kind of chart with all of them on there. The parents, and they usually do get confused. Is it the orange one? Is it this one? So, but I try to go over it with them, which one is rescue, which one is the daily medication. But yeah, we don't have examples of them in the clinic. [PCP 10 (without CDS tool experience)]

So, the biggest barrier that I've had is like people that get insurance denies like Flovent versus QVAR…I'm not even sure what the insurance wants. Then, the weirdest things is they want to replicate for kids, like that's the only one they'll approve, but it's for a kid that can't even use it. Like the barrier, I would say it would be cost, and then sometimes instead of doing like a low [dose]… sometimes the insurance will only cover like a high dose. And so you're doing it [a higher dose] instead of like a lower dose twice a day. But you're just hoping that the overall works out to be okay. You have to base on cost. And so I can think of one [patient] in particular that we've had to change it based on cost. [PCP 5 (without CDS tool experience)]

Asthma Action Plans

All PCPs reported using treatment plans to communicate with families but used variable methods to accomplish this task. The use of asthma action plans that are printed or sent to patients via a patient portal was reported by the PCPs using a paper-based CDS tool. Sending additional copies of the treatment plan was described as a useful function.

The majority of families that I have that are actually following their treatment plan. Aware of it adherent usually have a paper copy, that is probably on the fridge or something like that…I do appreciate that if it [the treatment plan] was not done, or if they need another copy or a nurse needs another copy, I can very quickly send it to them again [via patient portal]. [PCP A (with CDS tool experience)]

Furthermore, 3 PCPs commented that confirming if a treatment plan documented on paper was up-to-date was difficult and may be easier with an EHR-embedded tool that could track treatment plan documentation over multiple visits.

You don't know when it [the treatment plan] was updated you have to search and dig, to find out when it's updated…And then there's the prescribing piece where you're writing all this out, and then you get to do it again out in the prescription module. [PCP C (with CDS tool experience)]

PCPs not using a paper-based CDS tool indicated using different strategies to communicate treatment plans, for example, with inhaler labels, verbal explanations of which medications to take when, visual aids, and treatment plan templates from previous training. Some clinicians also used EHR-based tools, including personally developed or borrowed dot phrases (ie, preformatted content that can be pulled into a clinical note using an abbreviation or word), but the type of tool and use differed between PCPs.

I have a dot phrase for an asthma action plan, which I feel like I do more with ones that have controller medicine. [PCP 2 (without CDS tool experience)]

I do mostly verbal. Sometimes we'll look up pictures of the different inhalers on the computer just to make sure that they know which one is red, which one is orange, and how can they use it. And then I have like a PDF of an asthma action plan, but I don't use it as often as I probably should. [PCP 4 (without CDS tool experience)]

Yeah, it's a smartphrase from a colleague of mine who she got from her attending when she was a resident… I really like the colored ones, because it's more, if it looks good, the parents will read it…So, it's just a lot of typing. [PCP 7 (without CDS tool experience)]

Opportunities to Implement Guidelines

Several PCPs not using a CDS tool referenced awareness of national guidelines and intention to use in practice but noted relying primarily on how they were trained. More experienced PCPs acknowledged that they may be most familiar with previous versions of guidelines while both newer and more experienced providers expressed a desire for training on more recent guidelines.

You know, it's probably the guidelines that I learned … when I finished residency. It is what was ingrained in me. So, I probably, I know the guidelines changed a few years ago where you can use controller meds more intermittently. I still struggle to wrap my head around that concept. [PCP 1 (without CDS tool experience)]

We do have grand rounds every week. That would be a good avenue that I think of. Or you know, to make it a bit more personal level, like someone can go to the clinic. We do have provider meetings every second Thursday. It will be great if someone could, because you know we're general. We're the gatekeepers. With

XSL•FO
RenderX

all the guidelines you can barely keep up, so it may be a bit of a spoon-feeding, but I mean it's part of you know provider support. [PCP 7 (without CDS tool experience)]

I know GINA just had a recent update, so I looked through it, but I wouldn't say that I remember everything that's on it. I would definitely look over it to see what the updates were. The biggest thing I got out of it, because there's a lot of stuff that's beyond what I would do as a primary care physician in peds. But I remember the biggest change I think, was using an inhaled corticosteroid as a rescue inhaler. But I don't know how many people are actually following that, at this point. To be honest, I haven't really even started doing that, either. It's hard to change your ways. [PCP 10 (without CDS tool experience)]

CDS Tool Development

Within the CDS Tool Development domain, 3 themes were developed based on health care IT professionals' and PCPs' responses. First, how CDS tools can fit into provider workflow; second, best practices for CDS tool design; and third, strategies to evaluate the impact of CDS tools.

CDS Tool Fit With Provider Workflow

Health care IT professionals emphasized the importance of provider perspectives in design processes to promote implementation that aligns with realistic clinic workflow. The discrepancy between providers' initial conceptualization of a CDS tool and the clinical outcomes they seek to improve was highlighted:

What we find often, I think, is that providers will come to us saying they want something but, when we actually talk with them, have a conversation with them about what they're trying to accomplish, it turns out that what we can do for them is completely different than what they thought they originally wanted. [IT professional]

PCPs using a paper-based CDS tool provided examples of how they might translate their processes into the EHR, noting opportunities to reduce providers' time and documentation burden. They also suggested sending asthma questionnaires to patients via patient portal before their appointment to inform diagnosis and treatment planning but commented that patient response rates would be a limitation of this strategy.

Going from your script to the form… it's a lot of effort to sort of click your way through. If doing all that work to make a treatment plan could then just generate my script at the same time. [PCP A (with CDS tool experience)]

This paper ACT [Asthma Control Test], I generally actually think it's distributed pretty consistently. What I find is it's often done during the visit, so I'm not necessarily looking at it to the very end, or I'm asking them to finish it. So we can keep our conversation. If there were a way that they could, have the ACT done prior. We have some of our pre-visit questionnaire and then surveys that we that we do. They can access ahead of the visit. Not that a lot of them take that. But I think if there were people that were more willing to take that up, it would certainly be very helpful to them. [PCP A (with CDS tool experience)]

And then it [the questionnaire] pops into Admin. It says you know they have asthma and you're pulled right in with your smart phrase, and you can review it right there, the scores and everything. [PCP D (with CDS tool experience)]

Importantly, PCPs indicated a willingness to modify their workflow to include a CDS tool if embedded in the EHR. Specifically, PCPs mentioned EHR functions they believed could be useful, including reminders of established diagnoses, treatment options, progress note templates, and handouts for families.

If they are known asthmatics, it would be good if there would be an automatic bullet, and I mean I don't have to look for it, I don't have to type it, it'll remind me that hey this kid is asthmatic. In terms of diagnosing, it will be great to have something that we can just click, click, click, and maybe at some point something that they can fill out as well. Especially if they come for wheezing or cough. Then it [the EHR] can all summarize it for us. So that you know it's a more realistic, doable approach. Everything is on our shoulders. [PCP 7 (without CDS tool experience)]

I like a good order set that kind of spell things out a little bit. I think especially for these updated, like possible treatment options to be reminded of that, of what the options are, and what these new guidelines are.Cause I think otherwise without that reminder, I'm just going to fall back on what I was trained to do the last three years. [PCP 2 (without CDS tool experience)]

It would be cool if it was embedded in a note template. Everybody tends to use the same well child check templates, and then we all make little tweaks on it. But if there is a standard asthma template, and I guess if you clicked like the certain things, and then it embedded where they were, and if they needed to step up. It could define. It's not too much thinking about whether they're mild, moderate, or severe right. There's like definitions for this, and so it could quickly like auto populate that sort of stuff and have like something to show the parents. [PCP 5 (without CDS tool experience)]

Best Design Practices

Health care IT professionals discussed the importance of using tools the way they are already designed within the EHR, which often requires first understanding current workflows and then guiding clinicians from their usual practice toward a modified workflow that accomplishes clinical goals while being compatible with existing functionality. Compromise between individual PCPs' nuances in how they work and new workflows can be challenging. As such, they indicated that regular communication with a designated physician champion for the

XSL•FO
RenderX

implementation of a new CDS tool is paramount to an effective design process and subsequent clinician buy-in.

> *…we ask them, how are you doing this today?... Sometimes they'll ask for something that is totally different than the process they're following today. They may be asking for this because they want to simplify things, but they may be asking it because they're in that mindset of oh, I'm electronic now, and the way we did it before isn't applicable at all. But that's not necessarily the case.* [IT professional 2]

> *We are almost always asking them to do something different. So even if it is, for instance, sending out some sort of inventory, filling out a form, they still have to navigate to a place to get to that form and open it. It is that sort of moving them from what they're doing now to something completely different, even though it potentially has a lot of benefit for them.* [IT professional 1]

The health care IT professionals also provided an example of mechanisms for different CDS tools, such as pulling data from multiple records into a health maintenance summary to track chronic care (eg, routine screenings), drug dosing calculators within the medication ordering processes, and best practice alerts (BPAs) using in conjunction with questionnaires to guide decision-making. Importantly, they highlighted that even though clinicians may envision a new tool as a BPA pop-up, the risk of alert fatigue may compromise the potential utility of the tool in changing clinician behavior as well as negatively impact other CDS tools.

> *We're trying to really integrate our data analytics group generating reports to say, "How are we doing with this clinical decision support tool depending upon the nature of the tool?" We want to see that at least meeting expectations in a changing behavior because if it's not, and we have alert fatigue, then, of course, that also has a deleterious effect on all the other local decision support systems we put out there.* [IT professional 3]

> *We don't just want to encourage people to make BPAs for everything, because ultimately they get ignored, and so there might be some other tool. Understanding exactly what they are wanting to do and where they envision this working in their workflow.* [IT professional 1]

Finally, the health care IT professionals emphasized consideration of how EHR changes may affect a broader health system outside of the clinic implementing a new tool.

> *Sometimes an ask will be something that we need to set at what we call the system level. Which means it would impact users at [other clinic sites]…And because of the way they practice, which sometimes can be completely different, we have to come up with either work around where we don't have to set something at the system level, which those aren't always available.* [IT professional 2]

Evaluation of Impact

Working closely with a physician champion to share a comprehensive understanding of existing workflows and potential changes informs evaluation of a new CDS tool. One health care IT professional described how workflow events (ie, discrete steps a provider is taking in the EHR during an encounter) must be captured in a reporting system to reflect whether the tool is supporting physician decision-making behavior in an intended manner. Demonstrating impact to key stakeholders, for example, health system financial leadership, physicians being asked to implement a new CDS tool, and other clinical professionals working alongside them, was discussed as essential to willingness to change usual practices. Considering impacts on the broader health system was again highlighted as critical for regulatory reporting, a standard aspect of CDS tool implementation.

> *I ask the question about how do you want to demonstrate to your stakeholders that you've made an impact. So how are we going to be able to report on this and show, what are you going to measure, to demonstrate that it's efficacious, so that they can then sell this, so they can get people on board to follow. And use the tools we've implemented.* [IT professional 1]

> *We must know that workflow, because at certain points those key workflow events can be captured in a reporting system that way. We know that whatever report we have is reflective of the workflow to make sure that we're indeed doing the right thing.* [IT professional 3]

> *There's also differences within the various departments, various workflows, not just among clinicians among departments that we may be wanting to implement a tool that crosses, so that from a from a maintenance standpoint we have a single tool that we can use to collect data for across the organization and report out in one place. We must make sure that tool is going to work for everyone, and they still yet all those results and information still funnels out to the same place.* [IT professional 1]

Discussion

Principal Findings

The goal of the current study was to inform the design of an EHR-embedded CDS tool that adheres to the 4 pillars of NAEPP guidelines by gathering qualitative data from target stakeholders (ie, PCPs and health care IT professionals). Overall, PCPs described a variety of barriers to consistent guideline adherence during their care of children with asthma. Commonly endorsed barriers included having limited time during encounters, which was linked to inconsistency in the assessment of asthma symptoms and inconsistent provision of asthma treatment plans. Furthermore, PCPs described current difficulties with integrating guidelines into their workflow and that searching for documentation tools or printed resources was often challenging. These data are consistent with previous research showing that PCPs regularly encounter extrinsic barriers that limit their ability

XSL•FO
RenderX

to adhere to asthma guidelines [20]. Table 4 summarizes the key themes and conclusions that EHR-embedded CDS tool can improve care and guideline adherence.

Table 4. Summary of key themes and conclusions from qualitative interviews.

Theme	Conclusion
Limited time leads to inconsistent assessment of asthma symptoms and variable communication about treatment plans depending on provider	EHR[a]-embedded CDS[b] tool can improve care through efficient, standardized symptom assessment and diagnosis at the point of care and templates and resources to optimize provision and documentation of consistent treatment plans
Providers are aware of national guidelines but lack supports to implement them in practice	EHR-embedded CDS tool can improve guidelines adherence through structured support throughout clinical encounter to promote guideline implementation and fit within existing clinical procedures

[a]EHR: electronic health record.

[b]CDS: clinical decision support.

Data from this study suggest that CDS tools were positively viewed by PCPs and may be beneficial for overcoming noted barriers to improve provider adherence to NAEPP guidelines. PCPs who have previously used a paper-based CDS tool described that it helps them more readily identify new asthma diagnoses, improves their adherence to several NAEPP guidelines (eg, assessment of asthma control and provision of treatment plan), and were enthusiastic about its implementation in the EHR to improve asthma management. Similarly, PCPs who had not used an asthma-focused CDS tool identified several areas where such a tool could be beneficial to improve adherence to NAEPP guidelines (eg, improving consistency in provision of treatment plans) and expressed interest in incorporating a CDS tool into the EHR. These findings align with previous studies demonstrating that CDS tools improve provider adherence to national asthma guidelines [20] and are viewed favorably by PCPs [21,32], suggesting that CDS tools may be an especially viable mechanism to increase PCP adherence to NAEPP guidelines.

A notable change in recent guidelines is the inclusion of SMART [7]. SMART shows promise in reducing asthma exacerbations among those individuals with suboptimal asthma control [33]. However, although PCPs who did not currently use a CDS tool for asthma management indicated a general awareness of new guidelines, they acknowledged often relying on how they were originally trained to treat pediatric asthma, being unfamiliar with key changes to guidelines, and having limited professional development opportunities to familiarize themselves with updated guidelines. Thus, an important takeaway from the current study was that integration of an asthma management CDS tool into the EHR may facilitate adoption of SMART and be especially timely [30].

A unique aspect of this study is that we purposely included health care IT professionals to gather their perspective on how to best integrate a CDS tool into the EHR. Overall, their feedback echoed best practices for the development of CDS tools, which include designing for the end user's context in mind and providing the right information at the right time in the PCPs' workflow [34]. Feedback from PCPs on the development of our CDS tool aligned with that of that health care IT professionals and largely focused on ways the tool could improve efficiency and consistency. Specifically, PCPs described a desire to have guidance across the encounter to assist with diagnosis, severity determination, assessment of asthma control, and prescribing and ordering in a manner that is concordant with NAEPP guidelines. Furthermore, there was strong interest in support in the EHR for automating the note writing and treatment plan components of the encounter to reduce time-related barriers, improve documentation of relevant asthma-related information, and increase the likelihood that families would be provided an asthma treatment plan.

Strengths and Limitations of This Study

Our study has several notable strengths. First, consistent with strengths of this study include following best practices for the formative development of EHR-embedded CDS tools, we purposely engaged the end users of our planned CDS tool to provide their perspectives on its need at the point of care, the fit within their workflow, and its potential features. A unique strength of this study is the inclusion of the perspectives of 2 groups of PCPs, those with a longstanding history of using an effective paper-based CDS tool for pediatric asthma and those who have not used CDS in asthma care, across 2 health care systems in geographically distinct areas. This allowed us to gather diverse perspectives on our key research questions. Finally, we also gathered important data from health care IT professionals to understand the opportunities and challenges that scientific teams interested in developing EHR-based CDS tools should consider.

There are several study limitations that should be considered. First, we acknowledge that this study included a modest number of health care providers, all of which represent clinicians from large academic medical centers and use the EHR in their daily practice. Thus, although it is notable that health care providers were recruited from 2 distinct academic medical centers, the findings from this study may not generalize to health care providers with less experience using an EHR or who practice outside of academic medicine. Second, the responses that we received were consistent across individuals, however, we acknowledge that the current study included a small number of health care IT professionals. It is possible that recruitment of additional health care IT professionals would have yielded different opinions or data. Third, the health care IT professionals were from a single academic medical center that uses an enterprise-wide EHR system. Therefore, it is possible that their opinions may not be generalizable to other health care systems or EHR platforms. Finally, participants were largely recruited via convenience sampling–based, introducing some concern for

XSL•FO
RenderX

selection bias. We attempted to limit this concern by directly encouraging open and honest responses and ensuring that the study investigator did not have a professional relationship with the participant whenever feasible.

Conclusions and Future Directions

In this study, we used user-centered design principles to guide a qualitative study on perceived barriers and facilitators to a primary care–based, EHR-integrated asthma CDS tool. Participants in this mixed-methods approach expressed their interest in adopting an efficient asthma management CDS tool that could help them adhere to NAEPP guidelines and improve clinic workflow. Similarly, health care IT professionals perceived having an asthma CDS tool to be useful, as long as it adhered to EHR design standards. Additional recommendations from IT professionals included limiting BPA-alert fatigue and building automated reports to track use of the CDS tool. Thus, future directions will be to design an asthma CDS tool that (1) prompts PCPs in a low-burden approach that limits alert fatigue, (2) guides PCPs in diagnosing asthma and assigning a severity using native EHR tools, (3) provides PCPs with guidelines- and severity-concordant therapies to create an asthma treatment plan, and (4) streamlines follow-up care using EHR-embedded asthma control assessment tools.

Acknowledgments

We would like to thank the health care providers who took time away from their busy work schedules to provide their professional opinions and guidance on this work. Research reported in this publication was supported by the University of Florida Clinical and Translational Science Institute, which is supported in part by the National Center for Advancing Translational Sciences of the National Institutes of Health (award UL1TR001427). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Data Availability

The datasets generated during and/or analyzed during this study are not publicly available due to institutional regulations but may be available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

References

1. Most recent national asthma data. Centers for Disease Control and Prevention. 2021. URL: https://www.cdc.gov/asthma/most_recent_national_asthma_data.htm [accessed 2024-01-14]
2. Masoli M, Fabian D, Holt S, Beasley R, Global Initiative for Asthma (GINA) Program. The global burden of asthma: executive summary of the GINA dissemination committee report. Allergy 2004;59(5):469-478. [doi: 10.1111/j.1398-9995.2004.00526.x] [Medline: 15080825]
3. Most recent national asthma data. CDC. URL: https://www.cdc.gov/asthma/most_recent_national_asthma_data.htm [accessed 2025-01-14]
4. Guidelines for the Diagnosis and Management of Asthma 2007 (EPR-3). Bethesda, MD: National Heart, Lung, and Blood Institute; 2007.
5. Cui W, Zack MM, Zahran HS. Health-related quality of life and asthma among United States adolescents. J Pediatr 2015;166(2):358-364 [FREE Full text] [doi: 10.1016/j.jpeds.2014.10.005] [Medline: 25454936]
6. Bruzzese JM, Bonner S, Vincent EJ, Sheares BJ, Mellins RB, Levison MJ, et al. Asthma education: the adolescent experience. Patient Educ Couns 2004;55(3):396-406. [doi: 10.1016/j.pec.2003.04.009] [Medline: 15582346]
7. Expert Panel Working Group of the National Heart, Lung,Blood Institute (NHLBI) administeredcoordinated National Asthma EducationPrevention Program Coordinating Committee (NAEPPCC), Cloutier MM, Baptist AP, Blake KV, Brooks EG, Bryant-Stephens T, et al. 2020 focused updates to the asthma management guidelines: a report from the national asthma education and prevention program coordinating committee expert panel working group. J Allergy Clin Immunol 2020;146(6):1217-1270 [FREE Full text] [doi: 10.1016/j.jaci.2020.10.003] [Medline: 33280709]
8. Gibson PG, Powell H. Written action plans for asthma: an evidence-based review of the key components. Thorax 2004;59(2):94-99 [FREE Full text] [doi: 10.1136/thorax.2003.011858] [Medline: 14760143]
9. Janson SL, McGrath KW, Covington JK, Cheng SC, Boushey HA. Individualized asthma self-management improves medication adherence and markers of asthma control. J Allergy Clin Immunol 2009;123(4):840-846 [FREE Full text] [doi: 10.1016/j.jaci.2009.01.053] [Medline: 19348923]
10. Cloutier MM, Wakefield DB, Sangeloty-Higgins P, Delaronde S, Hall CB. Asthma guideline use by pediatricians in private practices and asthma morbidity. Pediatrics 2006;118(5):1880-1887. [doi: 10.1542/peds.2006-1019] [Medline: 17079558]
11. Cloutier MM, Grosse SD, Wakefield DB, Nurmagambetov TA, Brown CM. The economic impact of an urban asthma management program. Am J Manag Care 2009;15(6):345-351 [FREE Full text] [Medline: 19514800]

XSL•FO
RenderX

12. Akinbami LJ, Salo PM, Cloutier MM, Wilkerson JC, Elward KS, Mazurek JM, et al. Primary care clinician adherence with asthma guidelines: the national asthma Survey of physicians. J Asthma 2020;57(5):543-555 [FREE Full text] [doi: 10.1080/02770903.2019.1579831] [Medline: 30821526]
13. Wisnivesky JP, Lorenzo J, Lyn-Cook R, Newman T, Aponte A, Kiefer E, et al. Barriers to adherence to asthma management guidelines among inner-city primary care providers. Ann Allergy Asthma Immunol 2008;101(3):264-270. [doi: 10.1016/S1081-1206(10)60491-7] [Medline: 18814449]
14. Okelo SO, Butz AM, Sharma R, Diette GB, Pitts SI, King TM, et al. Interventions to modify health care provider adherence to asthma guidelines: a systematic review. Pediatrics 2013;132(3):517-534 [FREE Full text] [doi: 10.1542/peds.2013-0779] [Medline: 23979092]
15. Cabana MD, Rand CS, Becher OJ, Rubin HR. Reasons for pediatrician nonadherence to asthma guidelines. Arch Pediatr Adolesc Med 2001;155(9):1057-1062. [doi: 10.1001/archpedi.155.9.1057] [Medline: 11529809]
16. Yawn BP, Rank MA, Cabana MD, Wollan PC, Juhn YJ. Adherence to asthma guidelines in children, tweens, and adults in primary care settings: a practice-based network assessment. Mayo Clin Proc 2016;91(4):411-421 [FREE Full text] [doi: 10.1016/j.mayocp.2016.01.010] [Medline: 26944837]
17. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly: Management Information Systems 1989;13(3):319-339. [doi: 10.2307/249008]
18. Wood BL, Miller BD, Lehman HK. Review of family relational stress and pediatric asthma: the value of biopsychosocial systemic models. Fam Process 2015;54(2):376-389. [doi: 10.1111/famp.12139] [Medline: 25683472]
19. Beasley JW, Wetterneck TB, Temte J, Lapin JA, Smith P, Rivera-Rodriguez AJ, et al. Information chaos in primary care: implications for physician performance and patient safety. J Am Board Fam Med 2011;24(6):745-751 [FREE Full text] [doi: 10.3122/jabfm.2011.06.100255] [Medline: 22086819]
20. Cloutier MM. Asthma management programs for primary care providers: increasing adherence to asthma guidelines. Curr Opin Allergy Clin Immunol 2016;16(2):142-147. [doi: 10.1097/ACI.0000000000000242] [Medline: 26849166]
21. Bousquet J. Electronic clinical decision support system (eCDSS) in the management of asthma: from theory to practice. Eur Respir J 2019;53(4):1900339 [FREE Full text] [doi: 10.1183/13993003.00339-2019] [Medline: 31023866]
22. Jaspers MWM, Smeulers M, Vermeulen H, Peute LW. Effects of clinical decision-support systems on practitioner performance and patient outcomes: a synthesis of high-quality systematic review findings. J Am Med Inform Assoc 2011;18(3):327-334 [FREE Full text] [doi: 10.1136/amiajnl-2011-000094] [Medline: 21422100]
23. Kruse CS, Ehrbar N. Effects of computerized decision support systems on practitioner performance and patient outcomes: systematic review. JMIR Med Inform 2020;8(8):e17283. [doi: 10.2196/17283] [Medline: 32780714]
24. Bright TJ, Wong A, Dhurjati R, Bristow E, Bastian L, Coeytaux RR, et al. Effect of clinical decision-support systems: a systematic review. Ann Intern Med 2012;157(1):29-43 [FREE Full text] [doi: 10.7326/0003-4819-157-1-201207030-00450] [Medline: 22751758]
25. Cloutier MM, Hall CB, Wakefield DB, Bailit H. Use of asthma guidelines by primary care providers to reduce hospitalizations and emergency department visits in poor, minority, urban children. The Journal of Pediatrics 2005;146(5):591-597. [doi: 10.1016/j.jpeds.2004.12.017]
26. Cloutier MM, Wakefield DB. Translation of a pediatric asthma-management program into a community in connecticut. Pediatrics 2011;127(1):11-18 [FREE Full text] [doi: 10.1542/peds.2010-1943] [Medline: 21135006]
27. Cloutier MM, Jones GA, Hinckson V, Wakefield DB. Effectiveness of an asthma management program in reducing disparities in care in urban children. Ann Allergy Asthma Immunol 2008;100(6):545-550. [doi: 10.1016/S1081-1206(10)60058-0] [Medline: 18592817]
28. Hollenbach JP, Collins MS, Wasser C, Fedele D. Implementation of standardized asthma management programs in outpatient settings. Ann Allergy Asthma Immunol 2023;130(5):571-576. [doi: 10.1016/j.anai.2023.01.016] [Medline: 36702245]
29. Van de Velde S, Heselmans A, Delvaux N, Brandt L, Marco-Ruiz L, Spitaels D, et al. A systematic review of trials evaluating success factors of interventions with computerised clinical decision support. Implement Sci 2018;13(1):114 [FREE Full text] [doi: 10.1186/s13012-018-0790-1] [Medline: 30126421]
30. Shero ST, Ammary-Risch NJ, Lomotan EA, Mardon RE, Michaels M. Creating implementable clinical practice guidelines: the 2020 focused updates to the national heart, lung, and blood institute's asthma management guidelines. Implement Sci Commun 2023;4(1):36 [FREE Full text] [doi: 10.1186/s43058-023-00417-3] [Medline: 37003961]
31. Cloutier MM, Wakefield DB, Carlisle PS, Bailit HL, Hall CB. The effect of easy breathing on asthma management and knowledge. Arch Pediatr Adolesc Med 2002;156(10):1045-1051. [doi: 10.1001/archpedi.156.10.1045] [Medline: 12361453]
32. Matui P, Wyatt JC, Pinnock H, Sheikh A, McLean S. Computer decision support systems for asthma: a systematic review. NPJ Prim Care Respir Med 2014;24:14005 [FREE Full text] [doi: 10.1038/npjpcrm.2014.5] [Medline: 24841952]
33. Beasley R, Harrison T, Peterson S, Gustafson P, Hamblin A, Bengtsson T, et al. Evaluation of budesonide-formoterol for maintenance and reliever therapy among patients with poorly controlled asthma: a systematic review and meta-analysis. JAMA Netw Open 2022;5(3):e220615 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.0615] [Medline: 35230437]
34. Bates DW, Kuperman GJ, Wang S, Gandhi T, Kittler A, Volk L, et al. Ten commandments for effective clinical decision support: making the practice of evidence-based medicine a reality. J Am Med Inform Assoc 2003;10(6):523-530 [FREE Full text] [doi: 10.1197/jamia.M1370] [Medline: 12925543]

XSL•FO
RenderX

Abbreviations

ACT: Asthma Control Test
BPA: best practice alert
CDS: clinical decision support
GINA: Global Initiative for Asthma
EHR: electronic health record
NAEPP: National Asthma Education and Prevention Program
PCP: primary care provider
SMART: single maintenance and reliever therapy

Edited by A Mavragani; submitted 28.08.24; peer-reviewed by J Walsh, D Drummond; comments to author 01.12.24; revised version received 16.12.24; accepted 02.01.25; published 18.03.25.

Please cite as:
Fedele DA, Ray JM, Mallela JL, Bian J, Chen A, Qin X, Salloum RG, Kelly M, Gurka MJ, Hollenbach J
Development of a Clinical Decision Support Tool to Implement Asthma Management Guidelines in Pediatric Primary Care: Qualitative Study
JMIR Form Res 2025;9:e65794
URL: https://formative.jmir.org/2025/1/e65794
doi:10.2196/65794
PMID:

©David A Fedele, Jessica M Ray, Jaya L Mallela, Jiang Bian, Aokun Chen, Xiao Qin, Ramzi G Salloum, Maria Kelly, Matthew J Gurka, Jessica Hollenbach. Originally published in JMIR Formative Research (https://formative.jmir.org), 18.03.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Designing a Digital Intervention to Increase Human Milk Feeding Among Black Mothers: Qualitative Study of Acceptability and Preferences

Loral Patchen[1], CNM, IBCLC, MA, MSN, PhD; Jeannette Tsuei[2], MPhil; Donna Sherard[3], MPH; Patricia Moriarty[1], CCRC, MSHS; Zoe Mungai-Barris[4], BA; Tony Ma[5], MS; Elina Bajracharya[5], MS; Katie Chang[5], MS; William Douglas Evans[6], PHD

[1]Healthcare Delivery Research, MedStar Health Research Institute, Columbia, MD, United States

[2]RAND, Santa Monica, CA, United States

[3]Changeable, Silver Spring, MD, United States

[4]Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, United States

[5]Benten Technologies, Manassas, VA, United States

[6]Department of Prevention and Community Health & Global Health, Milken Institute School of Public Health, George Washington University, Washington, DC, United States

Corresponding Author:
Loral Patchen, CNM, IBCLC, MA, MSN, PhD
Healthcare Delivery Research
MedStar Health Research Institute
10980 Grantchester Way, 7th Floor
Columbia, MD, 21044
United States
Phone: 1 7205845076
Email: loral.patchen@medstar.net

Abstract

Background: Breastfeeding rates among US mothers, particularly Black or African American mothers, fall short of recommended guidelines. Despite the benefits of human milk, only 24.9% of all infants receive human milk exclusively at 6 months.

Objective: Our team previously explored the key content areas a mobile health intervention should address and the usability of an initial prototype of the Knowledge and Usage of Lactation using Education and Advice from Support Network (KULEA-NET), an evidence-based mobile breastfeeding app guided by preferences of Black or African American parents. This study aimed to identify the preferences and acceptability of additional features, content, and delivery methods for an expanded KULEA-NET app. Key social branding elements were defined to guide app development as a trusted adviser. The study also aimed to validate previous findings regarding approaches to supporting breastfeeding goals and cultural tailoring.

Methods: We conducted a qualitative study using in-depth interviews and focus groups with potential KULEA-NET users. A health branding approach provided a theoretical framework. We recruited 24 participants across 12 interviews and 2 focus groups, each with 6 participants. The Data methods aligned with qualitative research principles and concluded once saturation was reached. Given the focus on cultural tailoring, team members who shared social identities with study participants completed data collection and coding. Two additional team members, 1 with expertise in social branding and 1 certified in lactation, participated in the thematic analysis.

Results: All participants identified as Black or African American mothers, and most interview participants (7/12, 58%) engaged in exclusive breastfeeding. In total, 4 themes were recognized. First, participants identified desired content, specifying peer support, facilitated access to experts, geolocation to identify resources, and tracking functions. Second, delivery of content differentiated platforms and messaging modality. Third, functionality and features were identified as key factors, highlighting content diversity, ease of use, credibility, and interactivity. Finally, appealing aspects of messaging to shape a social brand highlighted support and affirmation, inclusivity and body positivity, maternal inspiration, maternal identity, social norms, and barriers to alignment with aspirational maternal behaviors as essential qualities. Crosscutting elements of themes included a desire to communicate with other mothers in web-based forums and internet-based or in-person support groups to help balance the ideal

XSL•FO
RenderX

medical recommendations for infant feeding with the contextual realities and motivations of mothers. Participants assigned high value to personalization and emphasized a need to achieve both social and factual credibility.

Conclusions: This formative research suggested additional elements for an expanded KULEA-NET app that would be beneficial and desired. The health branding approach to establish KULEA-NET as a trusted adviser is appealing and acceptable to users. Next steps include developing full app functionality that reflects these findings and then testing the updated KULEA-NET edition in a randomized controlled trial.

(JMIR Form Res 2025;9:e67284) doi:10.2196/67284

KEYWORDS

health equity; breastfeeding; qualitative; mobile health; black mothers; preferences; cultural tailoring; mobile phone

Introduction

Background

Breastfeeding has been shown to have numerous health benefits for infants and mothers, yet rates of breastfeeding among US mothers, particularly Black or African American mothers, fall short of recommended guidelines. The American Academy of Pediatrics [1], American College of Obstetricians and Gynecologists [2], and the World Health Organization [3] all recommend exclusive breastfeeding or exclusive human milk feeding (EHMF) for approximately 6 months after birth. Breastfeeding has been shown to be effective in reducing overall infant mortality, gastrointestinal tract and respiratory infections, sudden infant death syndrome, otitis media, and childhood obesity [4]. For mothers, breastfeeding has been associated with postpartum weight loss and a reduction in the risk of diabetes, cardiovascular disease, reproductive cancers, and postpartum depression [5,6].

Despite these benefits, only 24.9% of infants receive breast milk exclusively at 6 months [7]. Racial and ethnic disparities also exist in breastfeeding rates—non-Hispanic Black women are less likely to initiate breastfeeding and breastfeed for at least 6 months compared to the US national rate [8]. The American College of Obstetricians and Gynecologists emphasizes the importance of breastfeeding for underserved women due to the increased prevalence of adverse health outcomes among underserved groups, such as obesity, diabetes, and cardiovascular disease. Improving breastfeeding among Black or African American women not only improves short-term outcomes but also has the potential to reduce other long-term health disparities.

Barriers to initiation and continuation of breastfeeding can include structural, cultural, and societal factors, such as the need to return to work, lack of community support, or stigma around breastfeeding in public. For Black or African American women, reasons for low breastfeeding rates include not wanting to breastfeed, lack of breastfeeding information and education, having to return to work or school, lack of familial support, and being located in neighborhoods with hospitals that are less likely to implement practices that support breastfeeding [9]. Quintero et al [8] found that racial and ethnic disparities exist related to the role of breastfeeding information on breastfeeding initiation and duration. Black or African American women were less likely to receive information from their friends and family than other racial or ethnic groups.

Technology-based interventions have been shown to increase breastfeeding initiation and duration. A global meta-analysis of 15 randomized controlled trials found that technology-based interventions substantially increased the exclusive breastfeeding rate 6 months after delivery compared to usual care [10]. Mobile health (mHealth) technologies allow for greater access and dissemination of information at the exact time of need [11-13]. mHealth interventions have included a variety of modalities, including telephone support, televideo lactation consultations, SMS text messaging support, breastfeeding monitoring, and education. Several mHealth interventions have focused on the important role of peer support in increasing positive breastfeeding experiences, social connections, and a sense of belonging and self-efficacy [14]. Other mHealth technologies include location-based features, such as FeedFinder, a smartphone app using a GPS to support women in finding, reviewing, and sharing public breastfeeding places with others [15]. The My Baby Now app, which provides evidence-based information and support on infant feeding and active play from pregnancy to 18 months, found an increase in breastfeeding knowledge and EHMF intentions among Australian mothers. The authors also noted that mothers with lower levels of education found the app to be more helpful and rated it more highly [16].

Few studies have focused on the use of mHealth for breastfeeding in underserved populations. The Lactation Advice Through Texting Can Help trial connected low-income participants to breastfeeding peer counselors via SMS text messaging but did not find an impact on EHMF rates [17]. Notably, Lactation Advice Through Texting Can Help relied on text messages only and did not include additional features. Furthermore, the authors did not describe specific efforts to tailor messages to participants. A study of Breastfeeding Friend, an app with interactive advice, breastfeeding strategies and recommendations, educational content, and links to resources, also found no difference in breastfeeding rates among low-income, first-time mothers compared to a control app with digitized handouts. The authors found that women were receptive to receiving educational content through apps, but they also preferred in-person breastfeeding support, which could be provided through real-time chat functionality or SMS text messaging with experts [18]. While this intervention was designed with the intent to serve low-resource users and it delivered content at the fifth-grade reading level, it did not specify attention to cultural alignment or other aspects of social identity. Additional research is needed on culturally tailored

XSL•FO
RenderX

mHealth interventions that address the barriers to breastfeeding of Black or African American women.

In our previous study, our team explored key content areas and usability of an initial prototype of the Knowledge and Usage of Lactation using Education and Advice from Support Network (KULEA-NET), an evidence-based mHealth breastfeeding intervention guided by the preferences of Black or African American parents [19]. Participants appreciated tips on returning to work, feed and diaper logs, information on infant growth and development, SMS text messaging with a lactation consultant, and a GPS map of breastfeeding-friendly locations. After the initial study was completed, the development team incorporated additional features based on these preliminary findings and supported by other research. Additional features included *social support* (eg, videos with relatable Black or African American women and family members being supportive of breastfeeding), *emotional support* (eg, personalized text messages, communication between the mother and her support individual, and educational entertainment and testimonial videos), *informational support* (eg, knowledge base), *instrumental support* (eg, breastfeeding-friendly locations, breastfeeding classes, and access to lactation support), and *appraisal* (eg, feeding and diaper logs with personalized feedback) [20-23].

Objectives

The study aimed to assess the feasibility and acceptability of additional content and its delivery in the refined KULEA-NET app and to validate previous findings in supporting breastfeeding goals. The authors have adopted the more inclusive language of human milk feeding to report study findings.

Methods

Design

We conducted a qualitative study using in-depth interviews and focus groups to understand the preferences of potential users and assess the acceptability of KULEA-NET. A health branding approach provided a theoretical framework for this study [24]. This approach appeals to an individual's self-interest and emphasizes the cost and benefit tradeoffs of a health-promoting behavior in the hope that the individual will adopt that behavior [25]. We conceptualized the KULEA-NET brand to represent mental schema that help new Black or African American mothers encompass the benefits of breastfeeding. We designed the app to offer trustworthy advice and engaging social support, which would lead to adoption and continued use. This schema captures the benefit that outweighs the costs of breastfeeding and helps new mothers succeed in consistent breastfeeding. This approach was based on the work of Evans et al [26] and others that suggest "[h]ealth branding determines behavioral choice by building consumer relationships and identification with health behaviors and their benefits." As the nature of this study was exploratory, we used qualitative methods to gain insight into content and features that would appeal to potential users, preferred content delivery methods, and feedback on app content and prototypes. The combination of in-depth interviews and focus groups provided not only depth and breadth on our study topics but also identification of individual and contextual circumstances [27] surrounding breastfeeding.

Ethical Considerations

This study was reviewed by the Institutional Review Board of the MedStar Health Research Institute and approved as human participant research (IRB#00005929) on March 9, 2023. Informed consent was obtained from all study participants before data collection. All study data are deidentified, and no participant identifiers are included. Study participants were compensated with an electronic US $50 gift card for the in-depth interviews and an electronic US $75 gift card for the focus groups.

Setting and Relevant Context

This study was conducted in the United States at an economically and culturally diverse urban hospital with an academic research center and medical training programs. The hospital supports the largest number of births among all health systems in the city, >3000 births in 2022. The hospital primarily serves patients from its surrounding urban community, which was predominately White (36.3%) and Black or African American (43.5%) in 2022 [28]. In the United States, substantial racial or ethnic disparities exist in breastfeeding initiation rates. Compared to the national average, non-Hispanic Black mothers are less likely to initiate breastfeeding (73.6% vs 84.1%) [29]. In terms of resources, the Center for Disease Control and Prevention's 2022 Breastfeeding Report Card rated this urban area as having enacted paid family and medical leave legislation and having state licensing regulations for childcare centers that were fully aligned with the standard of encouraging and supporting breastfeeding by planning for mothers to feed on-site [7]. Despite these positive indicators, other barriers to breastfeeding remain, such as breastfeeding education, familial and social support, intergenerational and historical trauma, and misinformation [8].

Recruitment Procedures

Participants were invited to participate in the study using both active and passive recruitment approaches. A study research coordinator for women's health studies approached patients of ambulatory health care offices specializing in obstetrics and gynecology regarding the opportunity to participate in ongoing research initiatives. This location provides a comprehensive maternal-infant program that includes a full scope of services from health care to legal, behavioral health, and social services [30]. If interested in this study, eligibility was established and enrollment completed as appropriate. In addition, lactation educators were aware of the study and referred interested patients to the study research coordinator, who determined interest, eligibility, and completed enrollment as appropriate. Finally, information about the study was shared electronically with patients through a pregnancy and postpartum care app provided to them by the health care setting. Interested individuals contacted the study research coordinator directly regarding eligibility and enrollment.

Sample

For both in-depth interviews and focus groups, our desired population was Black or African American parents aged between 18 and 34 years who were English speaking, intended to continue breastfeeding after hospital discharge, owned a

XSL•FO
RenderX

smartphone with internet access, and were willing to provide verbal informed consent. We excluded parents of infants with conditions that required admission to the neonatal intensive care unit or interfered with breastfeeding (specific conditions are listed in Textbox 1).

Textbox 1. Exclusion criteria for this qualitative study on support for successful human milk feeding among Black or African American mothers, highlighting medical contraindications to breastfeeding.

Study exclusion criteria

- Infants born with cleft lip or palate, congenital heart defects, Down syndrome, neural tube defects, or other conditions that require admission to a neonatal intensive care unit or interfere with breastfeeding
- HIV
- Untreated, active tuberculosis
- T-cell lymphotropic virus type I or type II
- Illicit drug use
- Receiving radiation therapy
- Exposed to anthrax
- Undergone breast surgery
- Active hepatitis B or C
- Prescription drug use incompatible with lactation

We used a purposive criterion sampling method to recruit eligible patients who did not have contraindications for breastfeeding. Potential participants were contacted by the study team during a prenatal or postpartum appointment. Recruitment flyers were also posted at the center to recruit participants for the focus groups and electronically delivered to patients through a push notification associated with a remote blood pressure monitoring app used by patients receiving care at this center. In addition, snowball sampling was used to identify other potential participants that met the target population criteria. Participants were compensated with an electronic US $50 gift card for the in-depth interviews and a US $75 gift card for the focus groups.

We aimed to recruit 30 participants across the interviews and focus groups based on previous literature that recommends at least 12 participants to reach thematic saturation [31,32]. Recruitment continued until data saturation was reached on nearly all topics and no new themes emerged. In total, we recruited 24 participants across 12 in-depth interviews and 2 focus groups with 6 participants each. Due to this research's exploratory nature and the relatively homogenous nature of our target population, our sample size provided considerable data and was found to be adequate following qualitative research principles [33].

Data Collection

All in-depth interviews and focus groups took place on the internet via Microsoft Teams from June 1, 2023, to March 26, 2024. Recruitment occurred between June 1, 2023, and March 25, 2024. All recordings were transcribed and deleted once transcription was completed. Confidentiality was protected by assigning each participant an unidentifiable code and reporting only aggregated sample information and anonymized quotations. The study team consisted of the principal investigators (ie, multiple-principal investigator team), coinvestigators, app development technologist, study coordinator, research assistant, research nurse, and the community leadership council. Interviews lasted between 30 minutes and 1 hour, while focus groups lasted 1 hour. Team members who shared racially concordant social identities conducted data collection.

Verbal consent was obtained from every participant at the start of the interview or focus group. Before the interview or focus group, the study team provided participants with the opportunity to have any questions answered before obtaining consent. We ensured the consenting process occurred in a private location selected by the participant where other people were not present and the study team's private office. It was recorded via Microsoft Teams.

A semi-structured interview guide and discussion guide were used for the interviews and focus groups, respectively. For the interviews, topics focused on media and information sources used during pregnancy and motherhood and potential app content and features. The topics of the focus groups centered on identity as a mother, preferences for receiving support, and desired content and features for the app.

Data Analysis

Overview

The study team used deductive thematic analysis [34] to identify and interpret patterns or themes in the interview and focus group transcripts. The primary goal of the analysis was to identify themes and use them to inform the development of the content and functionality of KULEA-NET. The research team reviewed each transcript and developed a list of initial codes relevant to the research questions using the topics from the discussion guides as a starting point. An iterative process was used where codes were compared and discussed at regular team meetings during data collection. The coding dictionary was developed with each code having a distinct and mutually exclusive definition. Two team members coded the transcripts, compared codes, discussed any discrepancies, and modified the codes after

XSL•FO
RenderX

obtaining agreement. To develop the themes, 2 additional research team members collaborated with those who coded the data, reviewing the codes and grouping them into meaning or a pattern that answered the research questions. A summary of the thematic analysis is presented in Table 1. All coding and analysis were performed by the coders, and no data analysis software was used.

Table 1. Thematic analysis of responses from Black or African American mothers who participated in focus groups and individual in-depth interviews identified 4 themes: desired content, preferred modes of content delivery, requested functionality and features, and appealing aspects of messaging to shape a social brand for the app.

Theme, theme definition, and category	Category definition
Desired content and features	
Resources, information, and features participants sought and found helpful	
Peer support	Connection and communication with other mothers
Policy and advocacy support	Knowledge of mother's legal rights and guidance on tools to advocate for oneself
Access to experts	Access to lactation consultants, doulas, or health care providers through the app
Location-based resources	Geolocation functionality to identify helpful resources
Tracking and planning	Documenting the progress of pregnancy or breastfeeding and planning for exclusive breastfeeding
Content delivery	
Method in which content is delivered to users	
Device type	Preferred devices or platforms
Messaging	Preferred messaging modality—text messaging versus notifications
Desired app characteristics	
Preferences, requirements, or functionality participants requested	
Content diversity	Variety and range of content presented to app users
Ease of use	Capacity of an app to be used easily and effectively
Credibility	Trustworthy, reliable, and accurate information or sources in the app
Interactivity	Interaction between the user and the app that allows for real-time engagement, user input, and responsive feedback
Messaging	
Communication style, tone, and content	
Support and affirmation	Emotional support, tangible support, or validation from partners, family, peers, and community
Inclusivity and body positivity	Acceptance of all feeding modalities, breastfeeding decisions, and comfort with the body in relation to breastfeeding
Maternal inspiration	Positive qualities and characteristics related to motherhood that women aim to embody
Maternal identity	The significance mothers attribute to motherhood and how they define their role as mothers
Social norms	Perceptions and attitudes on breastfeeding among family, friends, or community
Barriers to alignment with aspirational maternal behaviors	Challenges to fulfilling maternal behaviors that are part of a maternal identity

Reflexivity

Our team participated in a reflexive process by taking notes throughout coding and consideration of thematic analysis. We used these to guide reflection on experiences related to infant feeding and considered how lived experiences informed coding and interpretation. Given the focus on cultural alignment, team members with racially concordant social identities completed data collection, coding, as well as thematic analyses related to cultural tailoring. They reflected on how their own support systems and cultural diversity within racial and ethnic communities are connected to the process of information gathering and interpretation. Team members with social branding and lactation expertise joined analysis of social branding and clinical themes only. As a group, we considered how to share these narratives with authenticity and sensitivity. A debriefing of coding and analysis was conducted with the study's community leadership group to critically assess findings and interpretations. This review assisted the research team with identifying potential biases and validating interpretations.

XSL•FO
RenderX

Results

Characteristics of the Sample

Characteristics of the participants (N=24) are detailed in Table 2. Fewer identifying details were asked of focus group participants, given the group setting.

Table 2. All study participants identified as Black or African American women. Details of breastfeeding history were asked during individual interviews, and approximately one-third shared a history of practicing only human milk feeding with their infants.

Sample characteristics	Interviews (n=12), n (%)	Focus groups (n=12), n (%)
Identified as Black	12 (100)	12 (100)
Identified as women	12 (100)	12 (100)
Pregnancy status		
Currently pregnant	6 (50)	0 (0)
Not currently pregnant	6 (50)	12 (100)
Primiparous	2 (33)	N/A[a]
Multiparous	4 (67)	N/A
EHMF[b]	7 (58)	N/A
Did not practice EHMF with first child but started EHMF with recent child	4 (36)	N/A
Currently, EHMF	2 (18)	N/A
Plan to EHMF	5 (45)	N/A
Has EHMF for >6 months	4 (36)	N/A
Attempted EHMF in past but unsuccessful	3 (25)	N/A

[a]N/A: not applicable.

[b]EHMF: exclusive human milk feeding. Study participants were able to select more than one option, percents do not sum to 100%.

Themes

Among the 12 interviews and 2 focus groups, four themes emerged that were related to KULEA-NET: (1) desired content and features, (2) content delivery, (3) desired characteristics, and (4) messaging.

Desired Content and Features

Peer Support

Participants described a desire to communicate with other mothers in web-based forums and internet-based or in-person support groups. They wanted to interact with mothers at different stages, such as first-time and more experienced moms, to learn about their experiences and ask questions:

> *I like seeing, you know, like a thread, a question and people giving their opinion their advice, their tips, how they dealt with that similar situation.*

Participants referenced their appreciation of the peer support features of existing apps (eg, Peanut, What to Expect, and My Pregnancy) and websites (eg, Postpartum Support International and The Feeding Flock) they have used. Another participant described the benefits of community and not feeling alone:

> *I like group settings as well because I feel like when [...] another person has asked a question that you have then it's almost like a sense of belonging because you don't feel like you're going crazy trying to figure something out. [...] You feel like a sense of like, ohh yes, we're all in this together.*

Some participants also mentioned the ability of peer support to fill in gaps in support from providers:

> *I feel like I've been gaining more information and support from like these other outside support groups more than I've been feeling supported by my provider, which is kind of sad.*

Policy and Advocacy Support

Some participants described a desire to know their rights as mothers and for tools to advocate those rights to employers. Participants wanted to know the legality of breastfeeding in public, their rights for pumping and milk storage at work, and their rights for work breaks as pregnant mothers. One participant envisioned links to credible state-based resources:

> *So if they have an app that tells them "ok, you are pregnant, this is your right." Not just say this is your right, give them an external link to like resources they can use if they don't want to rub their boss the wrong way or whatever. When you know it is your legal right, you won't have an issue with it and you will have something to back it up because you read it and you can go talk to HR.*

Access to Experts

Among existing apps, participants expressed that they liked the ability to connect with experts, such as lactation consultants, doulas, or providers. Multiple participants cited the brief time they had with a lactation consultant in the hospital after birth as insufficient to meet their needs. Others discussed how they

XSL•FO
RenderX

did not feel listened to by their providers and wanted access to expert-based information:

> *I didn't get a chance to ask all of [my] questions, or if I did, they were kind of rushed or the answer wasn't like a real answer.*

Participants found it valuable to connect with lactation specialists during extended hours, such as when they are the only ones awake with the baby, via video calls or SMS text messaging. One participant mentioned her experience using the video chat feature of an existing app:

> *I used it like at 2:00 in the morning because I was, you know, desperate because I had an issue with my breast. But anyway, so it was helpful during the video call because the lactation consultant she was able to like, you know, see my breast and see like, what was the issue and provide like, you know, real time suggestions like what I can do in my case.*

Location-Based Resources

Geolocation-based features that shared breastfeeding-friendly locations were well received by participants who evaluated the KULEA-NET prototypes:

> *I like it. Instead of trying to go on Google trying to figure out places where to go, at least you have, you know, the place pointed out to you where you can go.*

Participants suggested adding women, infants, and children locations and places that offer free baby supplies to this feature:

> *I was gonna say locations where you can go that gives free diapers.*

Tracking and Planning

Participants wanted the ability to track pregnancy and breastfeeding indicators (eg, size of baby, amount of milk pumped, and feeding times). In addition to tracking, planning for EHMF was a critical contributor to EHMF behavior. One participant discussed switching to EHMF with her second child due to planning:

> *I had looked into getting an IBCLC [International Board Certified Lactation Consultant] to help me and things like that - resources I didn't know the first time, I was seeking them out in advance. Just to make sure I had an arsenal of things to help me so I wouldn't quit.*

Participants envisioned the app as a tool to help them create a plan, integrate it into their daily lives, and receive guidance throughout the planning process.

Content Delivery

Device Type

Preference for an app that could be accessed via a smartphone or tablet was preferred over a laptop, as participants found those easier to use when holding a baby, more lightweight, and accessible while on the go:

> *When you're about to be a mom, your phone is your best friend.*

Type, Frequency, and Choice of Messages and Notifications

Participants were divided regarding how they preferred to receive information from KULEA-NET. Some preferred texts, particularly if they came from the same number, while others preferred notifications. The appropriate amount of text messages or notifications ranged from 1 to 3 per week. Daily messages or notifications were thought to be too intrusive and overwhelming:

> *Too many texts can be annoying, aggressive, and spam-like.*

Several participants also wanted a choice in either opting in or out of messages and notifications or selecting the frequency.

Desired Characteristics

In addition to the desired content and features, participants provided input on the characteristics KULEA-NET should exhibit.

Content Diversity

While KULEA-NET is designed for Black or African American women, participants recommended that the app approach cultural specificity with sensitivity to avoid the potential othering of social identities:

> *Because I feel like within the media, it's always portrayed as though, umm, it's more of a Caucasian thing to breastfeed, and maybe it's more popular, so to speak, in that culture, I guess, but I feel like there's plenty of black women that breastfeed celebrities that breastfeed.*

Participants suggest materials include images with a diverse range of women and not only Black and White women:

> *I like it overall, I'm but one thing I would mention is it needs to be more diversity.*

Moreover, participants did not want an app that only addressed breastfeeding but rather provided support throughout the prenatal to postnatal phases:

> *We have app fatigue—we need one that can transition through different stages of motherhood.*

One participant discussed how a comprehensive app would reduce app fatigue:

> *So it can also be like a lot to continuously hop from app to app to app and so the more information that you can have in one place I think is the better. the more that you can get in one app the better.*

Ease of Use

Information in the app should be easy to navigate, easy to read, and presented in various formats (ie, text, images, video, and audio) to appeal to various learning styles:

> *I'm just a visual person, so I like being able to see the person, it seems more real and personal.*

One participant explained that information-heavy apps were overwhelming:

> *How do I say, sometimes there are apps with too much stuff and then it doesn't feel like it's helpful.*

XSL•FO
RenderX

Credibility

Many participants commented on the need to emphasize the credibility of the content in the app by detailing sources, citations, or academic affiliations. Participants mentioned past experiences of receiving conflicting information from web-based sources, making them cautious of internet-based information:

> *A lot of the apps contradict themselves like "oh don't give into your cravings" and then in the third trimester they're like "eat whatever you want." What am I supposed to be listening to?*

Emphasis on credibility would make it easier and less time-consuming for participants to discern the validity of the information. While forums with peers were desirable, participants also expressed the need for a provider to moderate forums to reduce the potential of misinformation.

Interactivity

Video-based visits with lactation consultants supplemented by video examples of breastfeeding techniques were suggested. Participants wanted to share a live, visual display of the breastfeeding process, such as latching, with a lactation consultant to obtain effective support:

> *Phone or video will be most helpful so the consultant can see what you're doing and give real time advice, or at least a phone, because you might be not hands free, but at least on the phone you can put on speaker and speak that way.*

Messaging

Support and Affirmation

Participants appreciated empathetic support from their partners, parents, and children, which took the pressure off being a "perfect mom." They also appreciated partners who took up responsibilities when mothers needed rest:

> *We [my partner and I] have expectations that we don't set the bar too high. We know that in life there are various struggles and imperfections and he keeps me grounded when I feel like I have to be perfect as like a perfect mom and the perfect life.*

Some participants sought comfort in seeing other mothers struggling in similar ways or found breastfeeding inspiration on social media. One participant described the encouraging influence social media had on her decision to exclusively breastfeed. She said that seeing other people doing it made her believe that she could do it too. Referring to her Facebook group, 1 participant shared the following:

> *Yeah, because honestly, that is one of the things that got me through was the other members, you know, somebody who's been there, done that not just on uh educated level, but you know, somebody in my shoes.*

Many participants emphasized that the app should have positive reinforcement and words of affirmation, rather than a punitive tone. Pregnancy affirmations delivered by the app helped prepare participants mentally:

> *I like the daily motivational quote. Basically saying you're doing a great job, which we often need to hear as moms, even if we feel like we're not doing a good job. So I think that's great.*

Participants did not want to be pressured or judged for their breastfeeding decisions and behaviors:

> *I don't want it to make it seem like oh if you breastfeed, you are successful, but if you don't breastfeed that you're not successful....So just not to limit just to exclusive breastfeed[ing], like anyone could be successful.*

One participant described that she does not want her partner's opinions about what she is doing with her body. She considers breastfeeding a decision that she makes about her body and wants his support whether or not she breastfeeds:

> *I mean that's great [information for partners], but I don't want to hear my partner's opinion about what I'm doing to my body. It's not a driving factor to know how my partner feels about what I'm choosing to do.*

Inclusivity and Body Positivity

Many participants expressed the need for the app to include messaging that was inclusive of all feeding modalities. The app should indicate respect for all infant feeding choices, as some mothers are unable to breastfeed due to medical issues:

> *...I don't want it to seem like it's very exclusive to like breastfeed[ing] I feel like it should be open to anyone because there are different, you know, different circumstances and situations in which a mother can't provide breastfeeding, you know?*

A couple of participants expressed regret or guilt for not breastfeeding and wanted to avoid those negative emotions:

> *I wanted to [breastfeed] because I figured it's holistic, you know. That's how, you know, my son would have been able to fight off like sicknesses and things. So, for me not to do it, it was kind of hurtful.*

Body positivity was also found to be an important element of breastfeeding for participants. One participant explained how she feels more comfortable with her body with her second baby compared to her first because her breasts feel less "sexualized" and now symbolize a food source for her baby:

> *Definitely feeling more comfortable with my skin as far as I look as my boobs as nutrients you know. I look at it as milk, as food for the baby. I don't look at it as sexualized you know.*

Maternal Inspiration

When asked for examples of inspiration or influence, participants highlighted public figures who were able to balance their personal lives, careers, and other activities outside of motherhood:

> *Just how she balanced both working and taking care of three kids as well as, you know, still being a wife and still having her individual life going on.*

XSL•FO
RenderX

Others expressed admiration for celebrity or influencer mothers who are physically active and prioritize nutrition as mothers:

> *Ciara is very much active, and she was very active during her pregnancy, so that was a lot of inspiration for me to be active throughout my pregnancy.*

Single mothers in participants' personal lives and in popular culture were seen as displaying strength and resilience:

> *She's going through some things that an average person, you know that maybe somebody else would have broken down, but she continues to push through for her girls.*

Maternal Identity

Participants highlighted providing emotional safety, a better relative childhood, and trust in their motherly instinct to be key qualities of a good, capable mother. Creating a safe space and promoting confidence were important goals:

> *I could be a good, safe space for the child, emotionally help them grow into being a productive member of society. But then also help them see that they're uniquely gifted and uniquely made. And to help like, bring and call those things out so that they can feel like a confident person.*

Participants had a desire to provide a better childhood for their children than their own:

> *You know, being mindful of choices that we've made or that were made for us that we didn't necessarily agree with as we got older, you know, it takes that intentionality [and] awareness to say, you know, I'm gonna do this differently.*

Focusing on their child's needs rather than parenting norms and trusting their own "motherly instinct" sometimes over medical advice was valued:

> *The doctors may say one thing, right? You as a mom know what's best for your baby.*

Social Norms

Decisions on breastfeeding were often informed by practices among peers, family members, or a participant's broader community. Seeing sisters, mothers, and aunts breastfeed provided examples of normative behavior that influenced participants' breastfeeding decisions. One participant described her decision as "more so what I saw from my community." Breastfeeding was also associated with challenges to traditional norms and the expectations of support systems:

> *Family members were a barrier for breastfeeding in the sense that it is almost not possible for family to see her the way they want to because of the breastfeeding...They're used to bottle feeding with formula, so they're like, give me your baby, go sleep overnight. Just trying to not discourage me intentionally from breastfeeding, but just in the sense that they wanted to help. So they're like, "Why, it's so, so difficult when you're breastfeeding. Don't [you] wanna try formula?" [It's] not in a bad way. I know they were just trying to help, but that definitely was a hurdle for me.*

Barriers to Alignment With Aspirational Maternal Behaviors

Mental health challenges, including stress and anxiety, were described as important barriers to being a "good mom." One participant described how exhaustion can interfere with her intentions:

> *Mental exhaustion, like just being mentally tired and not wanting to do anything, just wanting to relax...When I may have other things that I could be doing or need to be doing.*

Discussion

Multiplicity of Maternal Identity

Using a health branding framework, we initially hypothesized that a targeted intervention leveraging audience segmentation based on sociodemographic characteristics would encourage intended behavior change [24,25]. However, our previous study revealed mixed responses to a culturally tailored intervention featuring Black or African American imagery. Participants informed us that providing for one's child was a universal desire that should not require race- or ethnicity-specific content [19]. In this study, we delved into participants' perceptions of maternal identity to better inform our communication approach. Our findings reiterated commonalities in maternal identity, particularly shared motivations of fostering emotional safety and providing a better childhood. Furthermore, sources of maternal inspiration highlighted were not exclusively Black or African American and were mentioned for a variety of desirable characteristics, such as success in balancing the multiple components of mothers' lives. Racial identity is merely one aspect of maternal identity among Black or African American mothers. Effective role models must speak to the multidimensional nature of motherhood, particularly the skillful managing of these diverse identities.

Previous research has shown the importance of avoiding the uniform treatment of sociodemographic groups, instead recognizing the multiple identities and intersecting factors that exist within them. A systematic review of mHealth apps among populations considered vulnerable, specifically low-socioeconomic status and historically marginalized racial and ethnic groups, found few improvements in health outcomes. The review underscored the need for further research that differentiates between broad categories of low-socioeconomic status and racial and ethnic minority groups to understand how mHealth tools can effectively address a range of needs [35]. Our mHealth intervention must appeal to the varied needs and lived experiences of mothers by incorporating features that consider sociocultural norms, environmental factors, household composition, resources, and availability of breastfeeding role models. KULEA-NET includes features informed by the barriers Black or African American mothers may face, such as breastfeeding education and tips, a GPS-based map of breastfeeding-friendly spaces, and feeding and diaper change logs. The delivery of this content will address the multiplicity

XSL•FO
RenderX

of identities among Black or African American mothers, which includes using imagery and messaging that speaks to a range of current and aspirational identities and helps them feel represented.

Nonoppositional Messaging

While EHMF for 6 months is recommended, our findings revealed that mothers wanted a more inclusive approach to breastfeeding—one that supports more options to accommodate the realities of a mother's circumstances and available resources. A recent study evaluated the "breast is best" narrative and noted a similar finding where mothers sought a more nuanced view toward breastfeeding and formula feeding to recognize a diversity of lived experiences and avoid inducing feelings of shame or failure when mothers were unable to EHMF [36].

In social marketing and health branding, oppositional and nonoppositional strategies are commonly used. Nonoppositional strategies aim to promote behavior change without directly challenging existing beliefs or practices (eg, eating 5 servings of fruits and vegetables), while oppositional strategies confront or compete with unhealthy messages (eg, antitobacco or antismoking messages) [25]. Our participants discussed workplace challenges to breastfeeding and the desire to know their rights or have tools to advocate for breastfeeding at work. Although mothers may intend to exclusively breastfeed, they may lack the means to achieve this goal. Employment before 6 months without adequate breaks and space for pumping has been shown to be a key barrier to EHMF [37,38]. Asiodu et al [39] emphasized the need to discuss combination feeding openly and alleviate some of the emotional burden and stress associated with not meeting expectations or societal pressures of EHMF. The tone and messaging of KULEA-NET must take a nonoppositional approach and acknowledge the circumstances that influence the ability of women to breastfeed. It will aim to provide positive reinforcement and encouragement to support mothers on their breastfeeding journey, rather than generate guilt or shame.

Critical Role of Support

Various forms of tangible and emotional support emerged as key features participants desired in KULEA-NET. Tangible support included peer support, partner support, access to experts, and assistance in navigating breastfeeding rights with employers. In addition, mothers expressed a strong need for emotional support to cope with the pressures of motherhood. Participants voiced the importance of nonjudgmental acceptance and encouragement of a mother's choices.

Access to breastfeeding support has been identified as an important mediator, accounting for two-thirds of the Black-White disparity in breastfeeding outcomes [9]. Influential supportive figures include not only providers and community-based experts but also family members, partners, friends, peers, church members, and even strangers. Moreover, Black or African American women are most likely to initiate and continue breastfeeding if they receive adequate support (ie, assistance with lactation, domestic activities, childcare, and groceries) from these sources during the first 2 weeks post partum [40]. Support from these various sources can help mothers overcome barriers to EHMF.

In this study, participants discussed examples of tangible and emotional support. Intimate relationships, such as partners, mothers, aunts, and grandmothers, not only shared some of the burdens of motherhood but also served as role models and fostered a normative culture of breastfeeding. Peers allowed mothers to learn from others' experiences, ask questions, and fill in when providers did not provide sufficient breastfeeding support. Peer support could also be internet-based, through forums or social media, in providing a sense of belonging, connection, and empowerment. Gyamfi et al [40] stressed the importance of exploring all avenues, including internet-based platforms, to establish culturally sensitive breastfeeding groups among peers in the Black or African American community to promote breastfeeding. Peer support features within an mHealth app can provide emotional support and community engagement and address a variety of unique needs in the Black or African American community.

KULEA-NET aims to integrate these findings by offering an app that recognizes the diverse lived experiences of Black or African American women, addresses both EHMF and combination feeding, and provides social support to improve breastfeeding outcomes. Future research will evaluate the efficacy of the fully developed version of this mHealth intervention in increasing breastfeeding rates through a randomized controlled trial. Findings of that trial will include usability and user engagement metrics and compare outcomes of exclusive human milk feeding success rates.

Future research should examine the conceptual model underlying KULEA-NET and its focus on cultural tailoring. This would involve evaluating the pathway of change and mediating factors, such as social support and app features, and their effect on breastfeeding outcomes. In addition, further research should explore the optimal timing for introducing the app (eg, enrolling participants in a KULEA-NET promotion study before vs after delivery) as well as the dosage and frequency of use for specific features. Studies should seek to identify the mechanisms of change through which KULEA-NET works.

Limitations

While our qualitative methodology provided a range of positive and negative perspectives on the KULEA-NET app and detailed information about the experiences of Black or African American mothers, there are some limitations to note. Our inclusion criteria included a requirement that participants were willing or intended to breastfeed, which meant that our sample may have been predisposed to breastfeed and may not have been fully representative of Black or African American mothers. In addition, multiple participants worked in health care, and our sample was informed of the benefits of EHMF compared to other nursing methods, which limited the external validity of our findings. Another limitation was inconsistency in the interview questions asked of participants. Some new questions were introduced in later interviews that were not included in earlier interviews, mainly questions concerning participant knowledge of their rights as mothers. For these questions, we had a smaller sample size, resulting in less detail on this topic.

XSL•FO
RenderX

Conclusions

Formative research suggests that the KULEA-NET app is acceptable and beneficial to participants. The personalization, support, and maternal identification features stood out as key benefits provided by the app. Implementation and promotion of app use represent important next steps. Future study developments will include full app functionality and testing in a randomized controlled trial.

Acknowledgments

The authors wish to acknowledge members of the project community leadership group for their contributions to this work: Tiara Caldwell, Jashia Pinkney, Cassietta Pringle, Wendy Post, and Kymberly Sharpe. This research was funded through a Small Business Innovation Research phase I and phase II application with the goal of commercializing the Knowledge and Usage of Lactation using Education and Advice from Support Network platform.

Data Availability

Deidentified study transcripts are available upon reasonable request to the corresponding author, with an established data transfer agreement.

Conflicts of Interest

LP, WDE, JT, ZMB, DS, and PM have nothing to disclose. TM, KC, and EB are employed at Benten Technologies and declare the following potential conflicts of interest and relevant relationships to promote transparency. Benten Technologies has a commercial interest in the outcomes of this research. While Benten Technologies is involved in the commercialization of the developed technology, the qualitative study and planned future randomized controlled trial are conducted independently at clinical and academic sites. These sites have no financial ties to Benten Technologies, ensuring that the study's design, data collection, and analysis remain unbiased and solely focused on scientific integrity and patient outcomes regarding the National Institutes of Health funding and product development under its Small Business Innovation Research award.

Multimedia Appendix 1
COREQ checklist.
[PDF File (Adobe PDF File), 144 KB - formative_v9i1e67284_app1.pdf]

References

1. Meek JY, Noble L, Section on Breastfeeding. Policy statement: breastfeeding and the use of human milk. Pediatrics 2022 Jul 01;150(1):e2022057988. [doi: 10.1542/peds.2022-057988] [Medline: 35921640]
2. . Committee Opinion No. 570: breastfeeding in underserved women: increasing initiation and continuation of breastfeeding. Obstet Gynecol 2013 Aug;122(2 Pt 1):423-427. [doi: 10.1097/01.AOG.0000433008.93971.6a] [Medline: 23969829]
3. Exclusive breastfeeding for optimal growth, development and health of infants. World Health Organization. 2023 Aug 9. URL: https://www.who.int/tools/elena/interventions/exclusive-breastfeeding [accessed 2025-02-19]
4. Dieterich CM, Felice JP, O'Sullivan E, Rasmussen KM. Breastfeeding and health outcomes for the mother-infant dyad. Pediatr Clin North Am 2013 Feb;60(1):31-48 [FREE Full text] [doi: 10.1016/j.pcl.2012.09.010] [Medline: 23178059]
5. Patnode CD, Henninger ML, Senger CA, Perdue LA, Whitlock EP. Primary care interventions to support breastfeeding: updated evidence report and systematic review for the US preventive services task force. JAMA 2016 Oct 25;316(16):1694-1705. [doi: 10.1001/jama.2016.8882] [Medline: 27784101]
6. Pope CJ, Mazmanian D. Breastfeeding and postpartum depression: an overview and methodological recommendations for future research. Depress Res Treat 2016;2016:4765310. [doi: 10.1155/2016/4765310] [Medline: 27148457]
7. Breastfeeding report card. Centers for Disease Control and Prevention. 2024 Jun 11. URL: https://www.cdc.gov/breastfeeding-data/breastfeeding-report-card/index.html [accessed 2025-02-19]
8. Quintero SM, Strassle PD, Londoño Tobón A, Ponce S, Alhomsi A, Maldonado AI, et al. Race/ethnicity-specific associations between breastfeeding information source and breastfeeding rates among U.S. women. BMC Public Health 2023 Mar 17;23(1):520 [FREE Full text] [doi: 10.1186/s12889-023-15447-8] [Medline: 36932332]
9. Jiang L, Nianogo RA, Prelip M, Wang MC. The role of breastfeeding support in racial/ethnic disparities in breastfeeding practices. Prev Med 2024 Feb;179:107857. [doi: 10.1016/j.ypmed.2024.107857] [Medline: 38224744]
10. Qian J, Wu T, Lv M, Fang Z, Chen M, Zeng Z, et al. The value of mobile health in improving breastfeeding outcomes among perinatal or postpartum women: systematic review and meta-analysis of randomized controlled trials. JMIR Mhealth Uhealth 2021 Jul 16;9(7):e26098 [FREE Full text] [doi: 10.2196/26098] [Medline: 34269681]
11. Chen H, Chai Y, Dong L, Niu W, Zhang P. Effectiveness and appropriateness of mHealth interventions for maternal and child health: systematic review. JMIR Mhealth Uhealth 2018 Jan 09;6(1):e7 [FREE Full text] [doi: 10.2196/mhealth.8998] [Medline: 29317380]

XSL•FO
RenderX

12. Guerra-Reyes L, Christie VM, Prabhakar A, Harris AL, Siek KA. Postpartum health information seeking using mobile phones: experiences of low-income mothers. Matern Child Health J 2016 Nov 17;20(Suppl 1):13-21 [FREE Full text] [doi: 10.1007/s10995-016-2185-8] [Medline: 27639571]
13. Seyyedi N, Rahmatnezhad L, Mesgarzadeh M, Khalkhali H, Seyyedi N, Rahimi B. Effectiveness of a smartphone-based educational intervention to improve breastfeeding. Int Breastfeed J 2021 Sep 20;16(1):70 [FREE Full text] [doi: 10.1186/s13006-021-00417-w] [Medline: 34544429]
14. Morse H, Brown A. The benefits, challenges and impacts of accessing social media group support for breastfeeding: a systematic review. Matern Child Nutr 2022 Oct 12;18(4):e13399 [FREE Full text] [doi: 10.1111/mcn.13399] [Medline: 35821651]
15. Simpson E, Garbett A, Comber R, Balaam M. Factors important for women who breastfeed in public: a content analysis of review data from FeedFinder. BMJ Open 2016 Oct 24;6(10):e011762 [FREE Full text] [doi: 10.1136/bmjopen-2016-011762] [Medline: 27797996]
16. Laws RA, Cheng H, Rossiter C, Kuswara K, Markides BR, Size D, et al. Perinatal support for breastfeeding using mHealth: a mixed methods feasibility study of the My Baby Now app. Matern Child Nutr 2023 Apr;19(2):e13482 [FREE Full text] [doi: 10.1111/mcn.13482] [Medline: 36725007]
17. Martinez-Brockman JL, Harari N, Segura-Pérez S, Goeschel L, Bozzi V, Pérez-Escamilla R. Impact of the lactation advice through texting can help (LATCH) trial on time to first contact and exclusive breastfeeding among WIC participants. J Nutr Educ Behav 2018 Jan;50(1):33-42.e1. [doi: 10.1016/j.jneb.2017.09.001] [Medline: 29325660]
18. Lewkowitz AK, López JD, Carter EB, Duckham H, Strickland T, Macones GA, et al. Impact of a novel smartphone application on low-income, first-time mothers' breastfeeding rates: a randomized controlled trial. Am J Obstet Gynecol MFM 2020 Aug;2(3):100143. [doi: 10.1016/j.ajogmf.2020.100143] [Medline: 33345878]
19. Patchen L, Ellis L, Harrington CB, Ma T, Mohanraj R, Andrews V, et al. Engaging African American parents to develop a mobile health technology for breastfeeding: KULEA-NET. J Hum Lact 2020 Aug 11;36(3):448-460. [doi: 10.1177/0890334420930208] [Medline: 32525434]
20. Mieso B, Neudecker M, Furman L. Mobile phone applications to support breastfeeding among African-American women: a scoping review. J Racial Ethn Health Disparities 2022 Feb 20;9(1):32-51. [doi: 10.1007/s40615-020-00927-z] [Medline: 33219430]
21. Schindler-Ruwisch JM, Roess A, Robert RC, Napolitano MA, Chiang S. Social support for breastfeeding in the era of mHealth: a content analysis. J Hum Lact 2018 Aug 22;34(3):543-555. [doi: 10.1177/0890334418773302] [Medline: 29787686]
22. Coughlin SS. The need for research-tested smartphone applications for promoting breastfeeding. Mhealth 2016 May 04;2(5):18 [FREE Full text] [doi: 10.21037/mhealth.2016.04.03] [Medline: 27231691]
23. Lee Y, Moon M. Utilization and content evaluation of mobile applications for pregnancy, birth, and child care. Healthc Inform Res 2016 Apr;22(2):73-80 [FREE Full text] [doi: 10.4258/hir.2016.22.2.73] [Medline: 27200216]
24. Ling PM, Lisha NE, Neilands TB, Jordan JW. Join the commune: a controlled study of social branding influencers to decrease smoking among young adult hipsters. Am J Health Promot 2020 Sep 20;34(7):754-761 [FREE Full text] [doi: 10.1177/0890117120904917] [Medline: 32077305]
25. Evans WD, McCormack L. Applying social marketing in health care: communicating evidence to change consumer behavior. Med Decis Making 2008 Jun 30;28(5):781-792. [doi: 10.1177/0272989x08318464]
26. Evans WD, Blitstein J, Vallone D, Post S, Nielsen W. Systematic review of health branding: growth of a promising practice. Transl Behav Med 2015 Mar 12;5(1):24-36 [FREE Full text] [doi: 10.1007/s13142-014-0272-1] [Medline: 25729450]
27. Verhoef MJ, Casebeer AL. Broadening horizons: integrating quantitative and qualitative research. Can J Infect Dis 1997 Mar;8(2):65-66 [FREE Full text] [doi: 10.1155/1997/349145] [Medline: 22514478]
28. District of Columbia. Data USA. URL: https://datausa.io/profile/geo/district-of-columbia#:~:text=by%20Nicolas%20Raymond-,Population%20%26%20Diversity,country%20(89.2k%20people) [accessed 2025-02-19]
29. Chiang KV, Li R, Anstey EH, Perrine CG. Racial and ethnic disparities in breastfeeding initiation — United States, 2019. MMWR Morb Mortal Wkly Rep 2021 May 28;70(21):769-774 [FREE Full text] [doi: 10.15585/mmwr.mm7021a1] [Medline: 34043611]
30. Patchen L, McCullers A, Beach C, Browning M, Porter S, Danielson A, et al. Safe babies, safe moms: a multifaceted, trauma informed care initiative. Matern Child Health J 2024 Jan 20;28(1):31-37 [FREE Full text] [doi: 10.1007/s10995-023-03840-z] [Medline: 37982953]
31. Guest G, Bunce A, Johnson L. How many interviews are enough?: an experiment with data saturation and variability. Field Methods 2006 Feb 01;18(1):59-82. [doi: 10.1177/1525822x05279903]
32. Fugard AJ, Potts HW. Supporting thinking on sample sizes for thematic analyses: a quantitative tool. Int J Soc Res Methodol 2015 Feb 10;18(6):669-684. [doi: 10.1080/13645579.2015.1005453]
33. Hamilton AB, Finley EP. Qualitative methods in implementation research: an introduction. Psychiatry Res 2019 Oct;280:112516 [FREE Full text] [doi: 10.1016/j.psychres.2019.112516] [Medline: 31437661]

XSL•FO
RenderX

34. Braun V, Clarke V. One size fits all? What counts as quality practice in (reflexive) thematic analysis? Qual Res Psychol 2020 Aug 12;18(3):328-352. [doi: 10.1080/14780887.2020.1769238]
35. Stowell E, Lyson MC, Saksono H, Wurth RC, Jimison H, Pavel M, et al. Designing and evaluating mHealth interventions for vulnerable populations: a systematic review. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. 2018 Presented at: CHI '18; April 21-26, 2018; Montreal, QC. [doi: 10.1145/3173574.3173589]
36. Scott SF, Bute JJ. "It doesn't have to be all or nothing": how individuals who use infant formula understand the breastfeeding master narrative. Health Commun 2025 Feb 21;40(2):321-331. [doi: 10.1080/10410236.2024.2344284] [Medline: 38644673]
37. Tran V, Reese Masterson A, Frieson T, Douglass F, Pérez-Escamilla R, O'Connor Duffany K. Barriers and facilitators to exclusive breastfeeding among Black mothers: a qualitative study utilizing a modified Barrier Analysis approach. Matern Child Nutr 2023 Jan 13;19(1):e13428 [FREE Full text] [doi: 10.1111/mcn.13428] [Medline: 36098279]
38. DeVane-Johnson S, Woods-Giscombé C, Thoyre S, Fogel C, Williams R2. Integrative literature review of factors related to breastfeeding in African American women: evidence for a potential paradigm shift. J Hum Lact 2017 May 05;33(2):435-447. [doi: 10.1177/0890334417693209] [Medline: 28380305]
39. Asiodu IV, Waters CM, Dailey DE, Lyndon A. Infant feeding decision-making and the influences of social support persons among first-time African American mothers. Matern Child Health J 2017 Apr 27;21(4):863-872 [FREE Full text] [doi: 10.1007/s10995-016-2167-x] [Medline: 27565664]
40. Gyamfi A, Spatz DL, Jefferson UT, Lucas R, O'Neill B, Henderson WA. Breastfeeding social support among African American women in the United States: a meta-ethnography. Adv Neonatal Care 2023 Feb 01;23(1):72-80 [FREE Full text] [doi: 10.1097/ANC.0000000000001021] [Medline: 35939758]

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Studies
EHMF: exclusive human milk feeding
KULEA-NET: Knowledge and Usage of Lactation using Education and Advice From Support Network
mHealth: mobile health

Edited by A Mavragani; submitted 09.10.24; peer-reviewed by L Ou; comments to author 11.01.25; revised version received 27.01.25; accepted 29.01.25; published 19.03.25.

Please cite as:
Patchen L, Tsuei J, Sherard D, Moriarty P, Mungai-Barris Z, Ma T, Bajracharya E, Chang K, Evans WD
Designing a Digital Intervention to Increase Human Milk Feeding Among Black Mothers: Qualitative Study of Acceptability and Preferences
JMIR Form Res 2025;9:e67284
URL: https://formative.jmir.org/2025/1/e67284
doi:10.2196/67284
PMID:

©Loral Patchen, Jeannette Tsuei, Donna Sherard, Patricia Moriarty, Zoe Mungai-Barris, Tony Ma, Elina Bajracharya, Katie Chang, William Douglas Evans. Originally published in JMIR Formative Research (https://formative.jmir.org), 19.03.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Exploring Laypersons' Experiences With a Mobile Symptom Checker App as an Interface Between eHealth Literacy, Health Literacy, and Health-Related Behavior: Qualitative Interview Study

Roland Koch[1], Dr med; Marie-Theres Steffen[1], Dr med; Anna-Jasmin Wetzel[1], MSc; Christine Preiser[2], Dr phil; Malte Klemmt[3,4], PhD; Hans-Jörg Ehni[5], Prof Dr; Regina Mueller[6], MA, Dr rer med; Stefanie Joos[1], Prof Dr

[1]Institute for General Practice and Interprofessional Care, Tübingen University Hospital, Tübingen, Germany
[2]Institute of Occupational and Social Medicine and Health Services Research, University Hospital Tübingen, Tübingen, Germany
[3]Institute of Applied Social Science, University of Applied Science Würzburg-Schweinfurt, Würzburg, Germany
[4]Institute for General Practice and Palliative Care, Medizinische Hochschule Hannover, Hanover, Germany
[5]Institute of Ethics and History of Medicine, University Hospital Tübingen, Tübingen, Germany
[6]Institute of Philosophy, University of Bremen, Bremen, Germany

Corresponding Author:
Roland Koch, Dr med
Institute for General Practice and Interprofessional Care
Tübingen University Hospital
Tübingen
Germany
Phone: 49 1758065961
Email: roland.koch@med.uni-tuebingen.de

Abstract

Background: Symptom checkers aim to help users recognize medical symptoms and recommend actions. However, they are not yet reliable for self-triage or diagnostics. Health literacy plays a role in their use, but the process from symptom recognition to health care consultation remains unclear.

Objective: This qualitative observatory study explored how laypersons use symptom checkers, focusing on the process of use, entry points and outcomes, and the role of health literacy. Laypersons are defined as individuals who are neither medical professionals nor developers of such apps. Three research questions were addressed: (1) How do such users describe the process of using symptom checkers? (2) What are entry points and possible outcomes of symptom checker app use? (3) How are health literacy and eHealth literacy expressed during the use of symptom checker apps?

Methods: As part of the Ethical, Legal, and Social Implications of Symptom Checker Apps in Primary Health Care project, 15 laypersons (n=9, 60% female and n=6, 40% male; mean age 30.7, SD 13.6 years) were interviewed about their experiences with the symptom checker Ada. The interviews were analyzed using an integrative approach combining social positioning, agency, and the Rubicon model as a heuristic framework.

Results: App use follows a cyclic process comprising 4 steps: motivation (influenced by biography and context), intention formation (assigning a purpose), intention implementation (recruiting resources), and evaluation (transforming interactions into health-related insights). Biographical, social, and contextual factors shape process initiation. Users use symptom checkers for 3 main purposes: understanding their condition, receiving recommendations for action, and documenting or communicating health-related information. Each purpose requires specific planning and integration into health-related behaviors drawing on personal, social, and technological resources. Evaluation depends on contextual factors, app outputs, and the outcomes of users' health-related actions. Users assess whether the app aligns with their expectations, condition severity, and previous experiences, with health literacy playing a critical role in validation processes.

Conclusions: Symptom checker use is a complex, cyclic process shaped by context, biography, and health literacy. Users are motivated by health concerns influenced by personal, social, and contextual factors, with trust and attitudes impacting initial engagement. Intention formation reflects a balance between user skills and context, where app outputs inform decisions but may not always lead to action, especially in ambiguous situations. Users rely on personal resources and social networks to integrate app use into health-related behaviors, highlighting the limitations of symptom checkers in providing social or empathetic support.

XSL•FO
RenderX

Symptom checkers have the potential to serve as an interface between users and health care, but future development must address the complexity of their use to unlock this potential.

International Registered Report Identifier (IRRID): RR2-10.2196/34026

(JMIR Form Res 2025;9:e60647) doi:10.2196/60647

KEYWORDS

symptom checker apps; health literacy; eHealth literacy; qualitative research; interview study; artificial intelligence; AI

Introduction

Background

This qualitative study investigated how eHealth literacy (eHL) [1,2], health literacy (HL) [3-6], and health-related behavior (HRB) influence how laypersons, defined as individuals who are neither medical professionals nor developers of symptom checker apps, engage with a specific mobile symptom checker app. Symptom checker apps offer readily accessible tools for self-triage by prompting users to enter health-related information [3,7]. This study focused on the Ada app by Ada Health GmbH.

Research shows that symptom checker apps are often used out of curiosity or anxiety or to confirm the severity of symptoms [4,7-9]. Factors such as users' age, confidence in self-assessment, trust in the app and technology, technology affinity, past experiences with the health care system, and proficiency in navigating health information play critical roles in shaping user experiences and engagement with symptom checker apps [3-7,9-16].

There is limited and partly conflicting evidence to conclude whether symptom checker app use directly prompts HRB, such as seeking professional health care [3,9,13,16,17]. Emerging literature suggests that symptom checker app use may foster self-care but could also lead to unnecessary health care use, particularly when the apps err on the side of caution in nonacute settings (eg, feeling tired over several days) [18-22]. Symptom checker app use can influence users' anxiety about seeking medical help [3,4,17,23], with many preferring "wait and see" recommendations [7,17], although satisfaction with the app does not always correlate with adherence to its advice [13,24].

In addition, previous studies highlight that eHL and HL significantly impact users' interaction with digital health tools and health outcomes, yet there is a knowledge gap in understanding these literacy concepts in the context of symptom checker app use [25-31]. Another gap in the literature is how the social context influences the use of digital health applications such as symptom checker apps, especially considering that eHL and HL are distributed unequally across regions and social groups [32-36].

How the occurrence of a symptom leads layperson users to engage in symptom checker app use and, consecutively, act on the app's output has not yet been described as a process [3,37]. There are several challenges to researching symptom checker apps, such as different user profiles, the context dependency of app use, and the rapid rate of development in the field [3,14,38].

Objectives

This study aimed to fill that gap by describing the process through which layperson users engage with symptom checker apps for health concerns, examining the influence of eHL and HRB on their interactions with the app and possible outcomes. Specifically, this study addressed (1) the steps that layperson users follow in using symptom checker apps, (2) key entry points and outcomes of symptom checker app use, and (3) the roles of eHL and HL in this process.

Methods

Study Design

This observational, exploratory qualitative study was embedded in the larger context of the joint Ethical, Legal, and Social Implications of Symptom Checker Apps in Primary Health Care (CHECK.APP) project [39]. CHECK.APP represents a collaboration among the Institute for General Practice and Interprofessional Care, the Institute of Ethics and History of Medicine, the Institute for Occupational and Social Medicine and Health Services Research at Tübingen University Hospital, and the Institute of Applied Social Science. The interdisciplinary research team consists of persons with professional training and research experience in medicine, family medicine, psychology, sociology, law, medical ethics, and philosophy. The project is conducted in the context of the German health care system with a focus on primary health care.

The CHECK.APP project is based on a mixed methods design and has four foci: (1) ethical, legal, and social issues [3]; (2) epidemiology of symptom checker app use and predictors of use [9]; (3) patterns of symptom checker app use and impact on individuals [16]; and (4) impact of symptom checker app use on the health care system and health care workers [40]. This study was part of focus 3. Focus 3 consisted of a diary study that was complemented with a follow-up qualitative interview study. The 6-week self-monitoring diary study aimed to generate mixed methods data on symptom checker app use patterns [16].

This study was conducted following the diary study and was designed to complement the quantitative analysis of the use patterns observed in the diary study by providing an in-depth exploration of the motivations and experiences of symptom checker app users. The reporting of this study followed the Standards for Reporting Qualitative Research guidelines statement [41]. The reporting checklist is included as Multimedia Appendix 1.

How Ada Works

On the Ada app, users enter their symptoms by responding to a series of questions. On the basis of their answers, Ada

XSL•FO
RenderX

generates follow-up questions to refine the symptom description. For example, if a user reports a rash, the app displays various images of rashes, prompting the user to select the one that best matches their condition. The app then yields a list of possible causes, recommendations on care urgency, and self-care advice (eg, "Eight of ten users with these symptoms had an acute infection—it is recommended that you go see a doctor").

Sample

For sample size estimation, we applied the 5D model of information power by Malterud et al [42], which considers study aim, sample specificity, use of established theory, quality of dialogue, and analysis strategy. According to the 5D model, the more specific each dimension, the smaller the required sample size in a qualitative study. Given the study aim of an in-depth exploration of symptom checker app use by laypersons, the specific focus on users of a single app, the analysis approach grounded in established theory (see the Data Analysis section), the expertise of the researchers involved with an expected high dialogue quality, and the planned techniques to enhance trustworthiness (see the Data Analysis section), a sample of 15 interviews was estimated to generate sufficient information power to explore the phenomenon under study. We reserved the option to recruit additional participants if the estimated sample size proved insufficient. To determine this, the analysis began with an initial reading of the transcripts even before the interviews were completed (see the Data Analysis section).

Participants for this study were recruited from a pool of 48 active symptom checker app users who had previously completed the 6-week self-observation diary study. All participants in the diary study were active users of Ada (Ada Health GmbH) and were originally recruited through the survey conducted as part of the CHECK.APP project. Recruitment was limited to existing Ada users to comply with ethical requirements prohibiting the initiation of symptom checker app use for the study. Participants used Ada within the context of their daily lives. Incentives were provided for maintaining the diary and participating in the interviews.

These diary study participants were approached and asked whether they were interested in undergoing additional in-depth interviews based on their respective diary experiences. All 48 persons who took part in the diary study were laypersons, meaning that they were neither health care professionals nor symptom checker app developers [16]. A chronic condition was reported by 21% (10/48) of the potential interview partners. In total, 35% (17/48) of the potential participants used the symptom checker app infrequently (less than once per week). Mean age was 27 (SD 9.1; range 19-64) years. A total of 65% (31/48) of the potential participants identified as female, and 35% (17/48) identified as male. No person identified as nonbinary. Racial characteristics were not asked about in the diary study [16]. Of the 48 persons who took part in the diary study, 28 (58%) expressed interest in an interview. The participants of the diary study from whom the interviewees were selected primarily consisted of well-educated younger female adults, a limitation discussed later in the manuscript.

Among these 28 persons, interviewees were stratified based on age, gender, use frequency, and experiences addressed in the diaries to achieve the best heterogeneity within the rather homogeneous sample. The relevant characteristics of the 15 selected interview participants can be found in Table 1. Mean age was 30.7 (SD 13.6) years.

XSL•FO
RenderX

Table 1. Study sample for the symptom checker app user interviews (N=15).

Relative dimension and features	Participants, n (%)
Age (y)	
20-29	10 (67)
30-39	2 (13)
40-49	1 (7)
50-59	1 (7)
60-69	1 (7)
Gender	
Women	9 (60)
Men	6 (40)
Diverse or nonbinary	0 (0)[a]
Symptom checker app use frequency	
Sparse (less than once per week)	6 (40)
Frequent (more than once per week)	9 (60)
Educational level	
Upper secondary school leading to university entrance qualification	12 (80)
Intermediate secondary school	2 (13)
Basic secondary school	1 (7)
Place of residence	
Urban	13 (87)
Rural	2 (13)
Chronic condition	
Present	5 (33)
Not present	10 (67)

[a]Despite our efforts, no nonbinary or diverse persons were included in the study.

Data Collection

The interviews were conducted by a pair of interviewers using an interview guide (Multimedia Appendix 2). The interview guide was created by the research team for this study and piloted with 2 interview partners who were not included in the analysis. The test interviews were used to review the guide and interview styles in a group feedback discussion between the interviewers and their supervisors. No changes were made to the interview guide during data collection. The interviews were conducted by a pair of interviewers comprising 1 representative from the Institute for General Practice and Interprofessional Care (AJW, RK, or MTS) and 1 representative from either the Institute of Ethics and History of Medicine (RM) or the Institute of Applied Social Science (MK). After consent was obtained, the interviews were conducted using the Zoom (Zoom Video Communications) video client and audio recorded on local, secure computers at Tübingen University Hospital. The interviews were conducted from January 18, 2022, to March 15, 2022, with an average duration of 46 (SD 10.4; range 35-76) minutes.

The interviews were transcribed pseudonymously by a certified office (Amanu). The transcripts were stored on protected servers at Tübingen University Hospital. For analysis, Microsoft Excel and Microsoft Word (Microsoft Corp) were used.

Data Analysis

This qualitative interview study followed a constructivist research paradigm—users construct their social reality by integrating experiences, self-image, daily interactions, and bodily perceptions into their biography [43]. They learn from experience [44]. These integrated factors are expressed through the social act of narration during the interviews.

The integrative basic method by Kruse [45] is an interpretative-reconstructive qualitative approach to reveal manifest and latent meanings and the way in which users “make sense” of the app and their experience with it. It represents an in-depth analysis of self-perception, biographical connections, self-image, and the perception of social roles within interview material [45]. The analysis was conducted by RK and MTS using strategies to enhance trustworthiness and rigor (see the Data Analysis section). The process is visualized in Figure 1.

Initial transcript reading commenced on March 1, 2022, while 3 interviews were still pending. This revealed thematic complexity; distinct patterns; and notable contrasts,

XSL•FO
RenderX

contradictions, and differences within the data. As a result, recruitment of additional participants beyond the planned sample size of 15 interviews was deemed unnecessary.

The first step in the integrative basic method is a detailed microlinguistic analysis in which transcripts are segmented into discrete units of meaning guided by cues such as pauses, interviewer questions, and topic changes. Each unit is then analyzed on three distinct linguistic levels capturing different aspects of the speaker's attention: (1) pragmatic (interactional aspects, including the relationship between interviewer and participant), (2) syntactic (grammatical choices that reflect cognitive structures), and (3) semantic (specific word choices, metaphors, and other lexical elements). This descriptive in-depth analysis of interaction, syntax, and semantics lays the groundwork for the following analysis of concepts such as agency, positioning, and the Rubicon model. Microsoft Excel was used for this step.

The transcripts were then analyzed using methodological and thematic heuristics. In our case, an analysis of social positioning and agency was used to understand how layperson users perceive the symptom checker app themselves and their scope for action embedded in a network of participants [46,47]. In terms of thematic heuristics, we used the Rubicon model by Achtziger and Gollwitzer [48] as a framework for motivational processes and intention building. It served both as a thematic lens to help explore the intrapersonal motivational process and as a macrostructure for the final synthesis text. The analyses were conducted using Microsoft Word.

In the final stage, the preceding microlinguistic analysis and the methodological and thematic analyses were condensed and compared against the background of the research questions to identify core motifs and modes of thematization within each interview and between interviews. This step began with formulating interpretative questions ("Lesarten" in German). The questions formed critical bridges between microanalysis (step 1) and interpretative analysis (step 2) to ensure that the interpretations remained grounded in the data. For instance, a question might be the following: "What role does Ada play in the progression from symptom onset to the utilization of health care?"

Figure 1. Visual representation of the integrative basic method. The process begins with 2 analysis steps: segmentation into units of meaning and microlinguistic analysis (step 1) and methodological and thematic heuristics, which are applied to the units of meaning (step 2). Both steps are then unified through interpretative questions ("Lesarten") into a final step (step 3), identifying central motifs and patterns within and across the interviews. Speech bubbles symbolize interviews, whereas flags denote units of meaning.

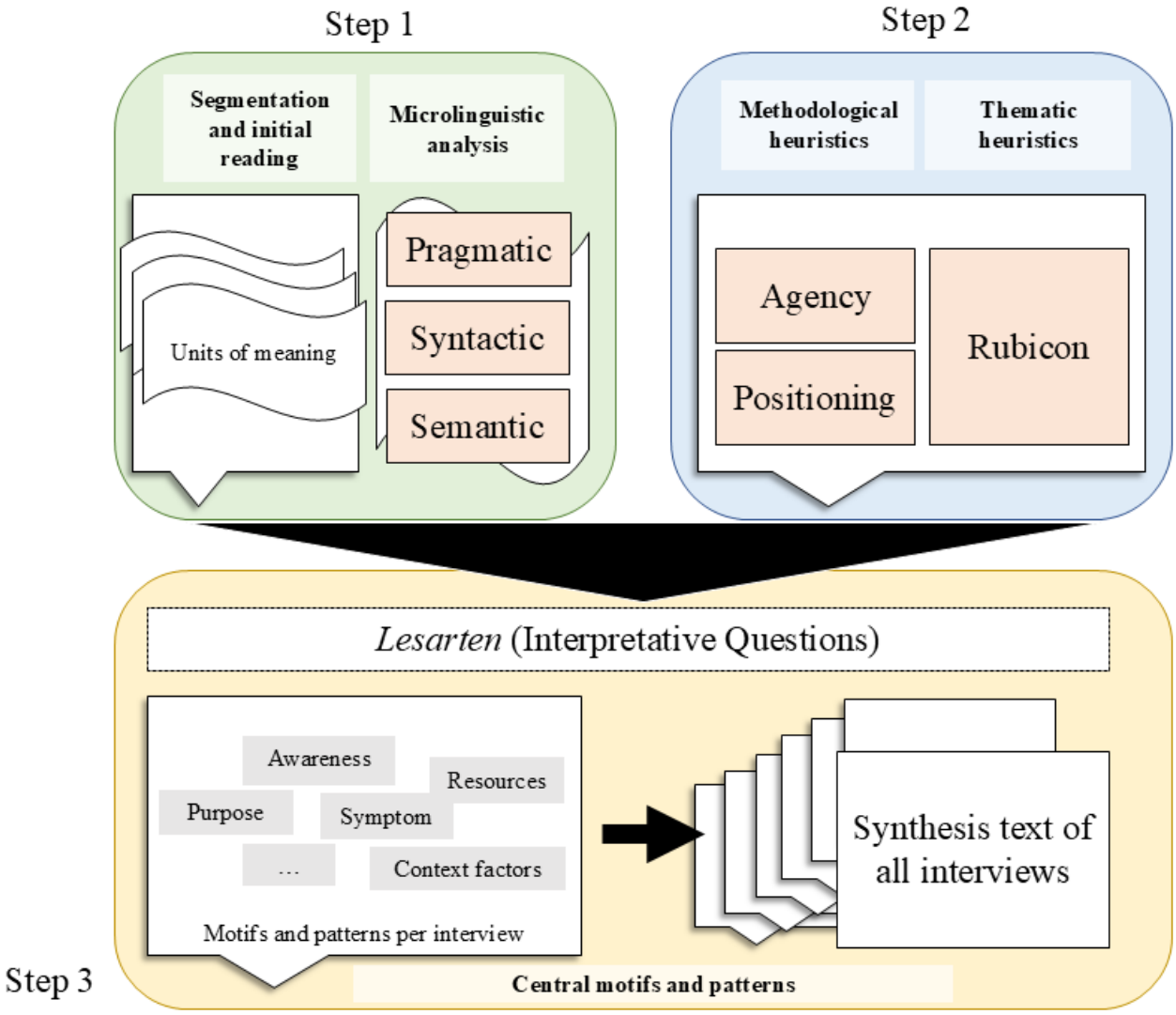

XSL•FO
RenderX

By answering the questions, central motifs and patterns specific to each interview were identified inductively. These motifs and patterns were then compared across all interviews to identify recurring themes (commonalities) and deviations within and across transcripts. While commonalities were used to generalize the findings, deviations were used to refine them. These insights were then used to reveal and formulate central, overarching motifs and patterns that were categorized using headlines (eg, “Context of app use”) and structured according to the steps of the Rubicon model. Microsoft Excel was used for this step.

To enhance trustworthiness, RK and MTS performed each of the aforementioned 3 analysis steps independently on the same transcript. They then compared their results after each step. Differences were discussed, and if possible, a mutual understanding of the texts and the method was established. When disagreements arose, they were discussed in monthly multidisciplinary methodological workshops within the CHECK.APP project research team.

Relying on the method of peer checking, the interim results were presented and discussed several times in a multidisciplinary research workshop on qualitative research methods. For a member check with most of the interview partners, the interim results were discussed and validated during a 3-hour session. Furthermore, in a multidisciplinary validation workshop, intermediate states of the analysis as well as open questions were discussed and illuminated under consideration of the perspectives of law, ethics, social medicine, and general medicine within the CHECK.APP project team. The results of these discussions were used to inform interpretative analysis (eg, by considering more perspectives) and the formulation of the final results text.

Ethical Considerations

This study received a positive vote from the ethics committee of the University of Tübingen (464/2020BO) and was conducted following the Declaration of Helsinki. All participants gave their informed consent before the interviews began. The participants received monetary compensation (€50, approximately US $50) for taking part in the interviews. All data were stored on encrypted servers.

Results

Visualization of the Symptom Checker App Use Process

On the basis of the 4 stages of the Rubicon model, a cyclic model of the symptom checker app use process according to the interview partners was developed. It is presented in Figure 2.

Figure 2. Symptom checker app (SCA) use process based on 15 interviews with SCA users. The figure is based on the Rubicon model. It describes the path from an entry point (such as noticing a health-related concern) via intention formation (influenced by motivational and volitional determinants) and volition to health-related behavior as an outcome. The latter can take place in different contexts. The behavior, in turn, leads to new experiences that are evaluated by the individual and integrated into the biography.

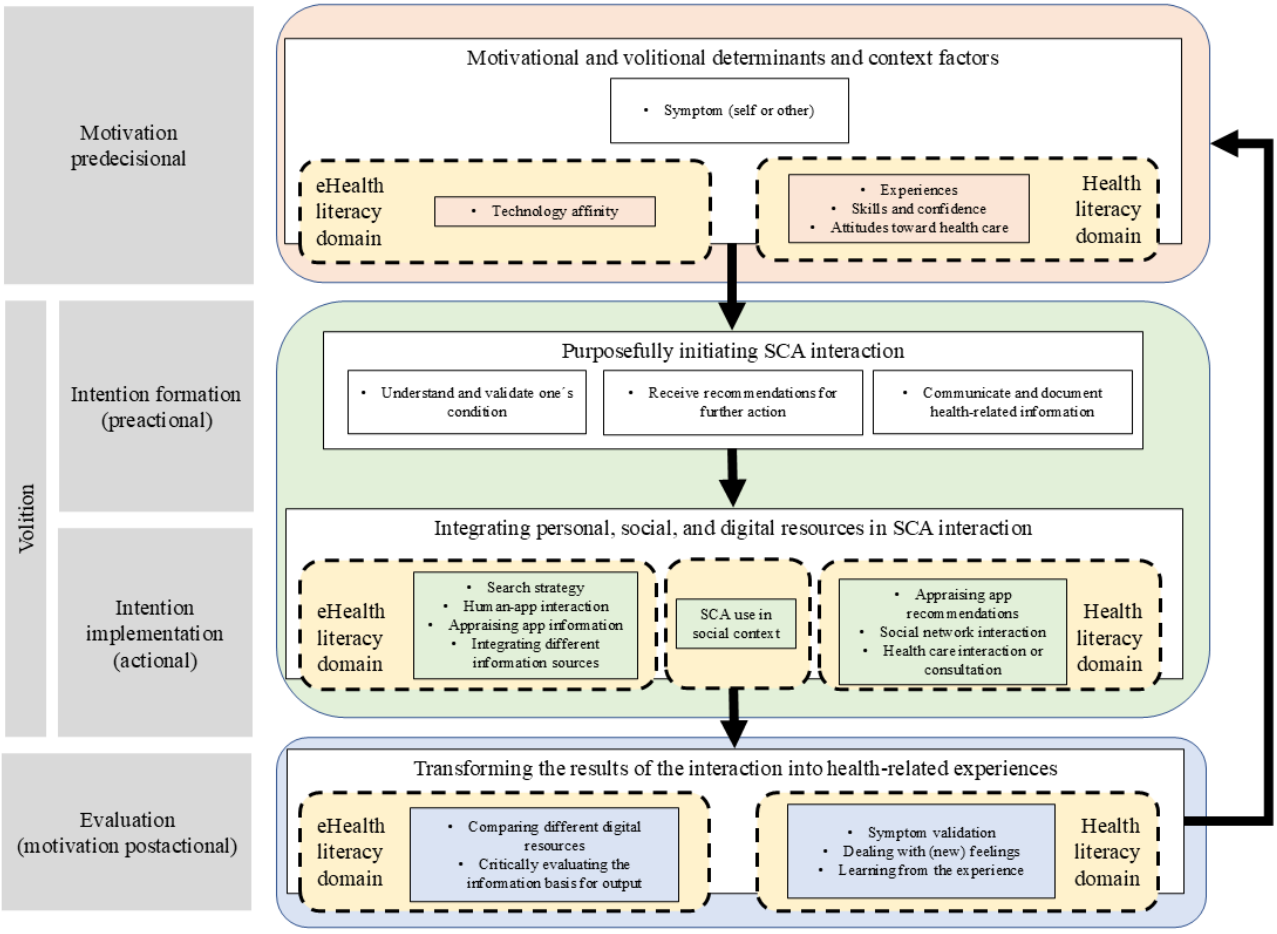

XSL•FO
RenderX

In the following sections, the synthesized text produced through the integrative basic method is presented. Interview excerpts are provided to illustrate the data basis for the central motifs and patterns. Unless stated otherwise, the excerpts are from the interview partners. If the interviewer asked a question, it is marked with "interviewer."

Motivation: Predecisional

Motivational Stage

Users assessed the severity of the issue by comparing it to their own preexisting experiences. The interview partners did not reflect on the relevance or validity of their perceived symptoms and experiences—they took for granted that both were real and valid. If the symptoms were perceived as less severe and comparable experiences were available in the users' biography, the sensation was described as manageable:

> *Yes. I have that once in a while. I don't even know where that comes from. I also think that it's possibly one of those things that kind of comes from stress, I definitely feel like that. And because I just never knew what it was, but I also knew that it would go away again, so because I had it before, as I said, I just looked it up in the app once.* [NT45]

At times, when individuals could not find relatable experiences from their own health history or perceived the issue as more severe, they expressed negative emotions such as fear or shock:

> *...In the end, it was a migraine attack. But I didn't think about it at that moment, I was shocked at first.* [NT39]

Trust in the app also played a role in the decision to use it, as the following excerpt illustrates. The same interview partner continued to elaborate on how trust in the app competed with trust in health care professionals:

> *But, yes, because I couldn't describe my problem in detail—I didn't trust the app so much at that moment—it was important for me to go to the doctor quickly, because I knew, okay, there are important things that I couldn't tell the app, and then that would explain the diagnosis, so to speak.* [NT41]

Context of App Use

Interview partners reported on different contexts in which they considered using the app. The most commonly reported context was the occurrence of a health-related concern. They first reflected on whether the symptom was even "real"—meaning whether they had just imagined it or whether the perceptions really represented an issue worth dealing with. They compared the symptom with their health-related experiences and biography and tried to make sense of what was happening to them:

> *Yes, I also clearly had chest pains. I mean, at that point I was already asking myself: Okay, to what extent is this happening now? But I mean, then I lay there and I had the feeling and thought: Yes, come on—so it was already real for me at that moment.* [NT47]

Users also reported being motivated to use the app by the health-related concerns of others. One user reported that she used the app for her family. She said that she was more resistant to possible anxiety than her family, whom she described as prone to hypochondria:

> *Interviewer: And what made you decide to [use the app] for [your daughter] and not with her?*
>
> *Interviewee: That I wanted to check it out first...My husband is a bit of a hypochondriac and so is my child. So when I sort of poke them, or it could be something dramatic, then I always don't know what will come of it. That's why I did a pre-check for myself first.* [NT21]

Health system accessibility or formal requirements because of their work further influenced users' motivation to take health-related actions:

> *Interview partner: ...It...is also one of the reasons why you don't like going to the doctor so much, is that it is hard to get an appointment with such specific doctors somehow. Especially a dermatologist or something like that, you have to wait two or three months and then you usually let it drop.*
>
> *Interviewer: So access to a doctor or to care plays a role for you?*
>
> *Interview partner: Yes. Yes. I think if I knew, okay, I'm going to get an appointment in one, two, three weeks, then I'd be more likely to have something like that checked out.* [NT41]

Volition: Intention Formation (Preactional)

The motivational process described previously led to a planning stage for HRB during which users purposefully engaged with symptom checker apps with 3 distinct goals in mind. They sought to (1) understand and validate their condition, (2) receive recommendations for further action, and (3) communicate and document health-related issues.

Understanding and Validating One's Condition

Users were interested in understanding their condition. The app provided an opportunity to find more information. Information seeking was described as one way to deal with the concerns:

> *...I see [the app] as a first aid, first source of information, where you can think about...: What do I do now? Then I can take a look or maybe I think again: What do I do now?...so as...again to update the normal knowledge that you have as a non-medical person.* [NT37]

Finding a clear cause also drove the users and induced HRB:

> *I woke up with 39 and something degrees. And of course I opened the app again and entered my symptoms, just to see...what it was saying now...Yes, then I knew...Okay, these are clearly the symptoms. Which of these is now Corona, which is maybe something else. Yes, I can still remember...that I was lying in bed suffering and had the app open.* [NT05]

XSL•FO
RenderX

Sometimes, users described that an element of surprise was needed to find an explanation for their concerns. They admitted that they were not aware of all possible causes of their conditions. The ability to be surprised was attributed to both physicians and the app as external information sources:

> *Maybe it's (what is needed to accept a diagnosis) a "click" moment that you didn't realize before. As an example, maybe a doctor or a diagnosis app asks a question [like: Does it get worse when it's cold] and then you think to yourself: Ah, yes, when I think about it, it does get worse when it's cold, or whatever.* [NT24]

Other times, users already had a suspicion of their own, which could be verified by the symptom checker app:

> *...I also have migraines and sometimes I can't quite identify whether it's a migraine or a headache. And the app often helps me and then I can practically distinguish: OK, do I treat it like a headache now or do I treat it like a migraine now.* [NT41]

In summary, the purpose of gathering information was to add to one's own knowledge about the condition and actively seek out validation of one's subjective impression.

Receiving Recommendations for Further Action

In addition to collecting more information about one's condition, users stated wanting to receive options for further health-related actions to improve their condition:

> *What were the first experiences?...I used the feature best or most, simply these suggestions about which therapies you can take for which symptoms. And so I always looked here again to see what the app actually recommends.* [NT44]

Even if the app was primarily used for information gathering, the recommendations were received and considered:

> *...I looked to see what information was available if I somehow...well, with children there's always something. And then I just thought the app was so good, because sometimes they had a few other tips. And then I just looked to see what one says, what the other says...It was always a push in the direction: "I should go to the doctor."* [NT21]

Users also typed in symptom combinations several times to see whether anything changed in the app's output:

> *Well, of course I also had accompanying symptoms during the heart rhythm disturbances. So sometimes trembling or high blood pressure, but some of it was only temporary. I then sort of started to play with that. I added the symptom once and indicated it and left it out once to see what it recommended. Because the option of selecting that a symptom only occurs temporarily or is concomitant, only temporarily, is not available in the app. And then I wanted to play around with what would come out in the end if I indicated this and didn't indicate it. But apart from that, the reason was...I gave everything I had and always answered the questions truthfully.* [NT19]

If the sense of medical urgency was very high or life-threatening, app use was explicitly avoided. One participant described such a hypothetical scenario:

> *If I somehow had symptoms of a heart attack or something like that, then I would probably call the ambulance service directly and not look on the app to see if I really had a heart attack or if I was bleeding to death or something like that.* [NT49]

Communicating and Documenting Health-Related Issues

Users documented their own symptoms on the app:

> *It was recommended to me a long time ago to track symptoms, just to keep a diary of symptoms...* [NT31]

The app was also used to prepare for a visit to health care professionals. In that context, the documentation contained on the app was primarily used for the orderly and rapid presentation of concerns:

> *So I can already describe my state of health more specifically than if I hadn't inquired beforehand.* [NT06]

Volition: Intention Implementation

Personal Resources

To implement the purpose of the initiated action, users integrated personal, social, and digital resources into their behavior. One example leading to a health care visit was provided by one of the interview partners:

> *My doctor didn't tell me that [oral iron supplement pills] could have side effects. And then I had discolored feces and abdominal pain and all sorts of things for a week or two...So I used the app and then I googled it and then I think I also talked to a friend about it. (00:04:27) She was like: Hey, maybe it's from the iron tablets. And then I realized that it was the iron tablets. And then I stopped taking them and then I talked to my internal medicine physician and she was like: Yes...And prescribed me other iron tablets.* [NT41]

Personal resources were provided by biographical reference points such as previous experiences or special knowledge from one's own professional activity or skills. As soon as the purpose of the action was determined in the preactional phase, personal resources were factored into the user's action. Users described certain skills that they used in the process but also a sense of self-efficacy. If enough personal resources were available, users could actively decide against the app's recommendation, as the following example shows:

> *And then I entered that [the symptoms]. The app said it could be a torn ligament or a fracture and that I should go to the emergency room. And then I thought: ...I don't know how I could have broken any bone in the way I was walking. For me, the ligaments were more plausible...I've had it before, so I already had a bit of experience of what it's like. And then I thought: OK, the app says one thing, but I'll just look*

XSL•FO
RenderX

at it again in two or three days and see how it develops. [NT42]

Social Resources

Users also intentionally and purposefully sought out social resources. Friends and family were assigned a role in understanding and responding to concerns:

> *...I once suspected a stroke and then I was at home and had a reduced field of vision and thought: What is this? But that was before I knew about these apps. And of course I immediately called a friend and said: What could this be? What is that? And she said: Make sure you have it checked out the next day.* [NT21]

Health care professionals such as physicians were assigned a special role—a visit to a health care professional was seen as mandatory to obtain a final validated diagnosis or to obtain health care services such as an authorization for sick leave or a prescription. The interactions with physicians were mostly intentionally initiated by the users if other resources were depleted or did not improve their situation:

> *And I had then already contacted my family doctor anyway, just from the fact that I also need a sick note...Of course, if this [the symptoms] were to last longer, the app would be of no use to me in the end, then I would have to go back to the neurologist or...the family doctor.* [NT39]

The only social resource with physical access to the users' body were physicians as they could conduct physical examinations:

> *And then I went there [to the General Practitioner], described it again and showed it. Then...[she] turned the [painful arm] in different directions and said that no part, neither the forearm nor the upper arm muscle, is affected in any way. That comes in any case, exactly, so from the tendons...And then I just said: yes, [the concern comes] from climbing...And then she said: how much do you climb? So if you don't climb four hours a day, it's very unlikely that it comes from that.* [NT31]

Reviewing the input and results together with health care professionals (eg, a general practitioner [GP]) was mostly described hypothetically. However, the prospect of discussing app results with their GP was seen as beneficial by the users:

> *...what I would like is when you come to the doctor and say: I assume that I have this because I googled it or found out on an app. That he then says: Yes, what symptoms did you enter? Or: How did you then come up with this result?...So that he then, I don't know, doesn't have to start again with Adam and Eve, but that he can already inquire more intensively.* [NT06]

Mostly, users made a conscious decision to leave the app out of the social context of health care. They expected it to have a negative impact on patient-physician relationships:

> *Interviewer: Did you also mention there (at the GP) that you used the app?*
>
> *Interview partner: No, because I have the impression that doctors tend to react badly to this. Because then, they always like to say Dr. Google...and the topic has a bit of a bad reputation.* [NT07]

Digital and Technological Resources

The app and search engines were seen as competitors in the same category with distinct advantages and challenges. Search engines were mainly used as a supplement to obtain information that was missing on the app:

> *...when you're constantly googling, sometimes very worrying answers come up and you can also enter a lot of specific things in the app and it's just more practical and you somehow feel, how should I put it, the diagnosis somehow feels more trustworthy than through Google.* [NT41]
>
> *...I missed in the symptom checkers, that you somehow get such a tip, such old home recipes...belly compress or something, and then such a guide to it...I then thought: ...now I'm googling, what can one do, yes, perhaps with a homeopathic approach. Because that's what I was missing.* [NT21]

In contrast to app use, users found search engines overwhelming when they tried to check their symptoms there:

> *[Referring to searching concerns with internet search engines] In principle I never really figured it out, because there were 10, 15, 20 possibilities, and I couldn't really assess it at all. Really.* [NT19]

There are other technological resources and devices such as users' wearable devices or imaging procedures at the hospital in addition to the app. Devices able to monitor and measure bodily functions were described as determining and important factors that can contribute to a decision:

> *I...always measured the blood pressure. And even my [smartwatch], which is so clever, then recognized that there is a heart rhythm disturbance, at times. That time, it showed a rhythm disturbance...Then I thought: Okay, now I have to go to the doctor.* [NT19]

The user experience and usability features of the app were discussed. For example, the chatbot-based approach with questions asked to be answered by the user was perceived as both helpful and limited. Users missed the opportunity to clarify their input, ask questions, or address their own uncertainty and ambiguity:

> *...I would have put it (a surprising app-result) down to the fact that the app didn't really understand me...because I somehow couldn't really convey exactly what I wanted to say on the basis of the questions and the selection options...I would have said: OK, it just misunderstood me.* [NT37]

According to the users, affinity for technology played a role in the use of the app:

> *Yes, for my mum, for example, I use it (the app) quite often, because she's not that into the internet and apps and stuff.* [NT41]

XSL•FO
RenderX

Evaluation

The outcomes of the interaction between users and the symptom checker app were integrated into their subsequent experiences and interpreted in the context of their past health-related experiences. Once the HRB had been executed, a comparison was made as to whether the intention of the action had been fulfilled. Whether the app's results represented an actual diagnosis was negotiated individually by each user. Users described having difficulty accepting the app results depending on the topic and their own interpretation of the results. If the 2 diverged too drastically, they were at a loss and saw potential problems in understanding or misinterpreting the given recommendations:

> *I think the app always told me: stress and mild depression or something like that. I didn't feel that stressed, but yes, I think that was the thing...I read through it, but then I thought: Well, I still don't know exactly what's going on.* [NT37]

Concerning recommendations especially, users critically evaluated the app's output. How they reacted to it depended on the intended purpose of the action, the user, and the context in which the assessment took place:

> *I was a bit shocked that [the app] now sends someone directly to the emergency room (with the app result tonsillitis). Because I work for a health insurance company...and I find it a bit exaggerated that patients are sent so quickly to the emergency room. They are so overloaded at the moment.* [NT39]

Users reported that they were reassured by validated external information about their condition. The app sometimes could provide such reassurance:

> *It's a bit like this: Is it true now? Of course, we are aware, especially with this vaccination, that the side effects come from the vaccination. But somehow it's still like that: Yes, you feel safer and somehow more confirmed, even if you only have it in an app, but, yes, you are somehow more reassured.* [NT41]

On the other hand, users also found the app results unsettling and misleading:

> *So it could really only be overstretched ligaments (user's suspicion), this fracture (app result) made me a bit...A bit scared at that moment.* [NT42]

Users acknowledged that having received information contributed to a learning process that affected how they reacted to future symptoms:

> *I still have it on my mobile phone, I still like to look at it once in a while, even if it's just to advance my knowledge a bit. And, yes, it has become...I have found that I look at it a little less, but also because I find that when things repeat themselves in some way, you already know how you could react.* [NT44]

Discussion

Principal Findings

Layperson Perspectives on Using Symptom Checkers for Health-Related Concerns

This study's findings indicate that the use of a symptom checker app by layperson users is a multifaceted and iterative learning process that involves individual motivational factors; contextual elements; and interaction with digital, social, and personal resources. Both eHL and HL play a significant role in this process. The subsequent Discussion section addresses the 3 research questions.

The use of symptom checkers can be mapped onto the 4 distinct steps of the Rubicon model: motivation, intention formation, intention implementation, and evaluation [48]. These stages can be viewed as an iterative learning cycle that generates health-related experiences according to the constructivist paradigm [44]. During the motivational stage, users negotiate app use depending on the context. If users choose to use the app, they do so intentionally for 3 distinct purposes: obtaining information and health information–seeking behavior (HISB), receiving recommendations, and documenting and communicating health-related issues. The purpose determines the user's strategy and the resources used to achieve it, including personal, social, and digital or technological resources. Users' experiences with the symptom checker app can influence their motivation to use it in the future and how they incorporate it into their HRB. Making meaning plays a crucial role in this as users try to understand their bodily perceptions and symptoms [43]. This process aligns with the model proposed by von Wagner et al [29], which links HL and HRB and emphasizes the role of learning in HL. Our study highlights that the use and evaluation of symptom checker apps in line with the dynamic concept of HL varies over time and is dependent on the context and purpose of use [25,27,28,49]. This poses a challenge when researching symptom checker apps [4,7,38].

Entry Points and Possible Outcomes of the Symptom Checker App Use Process

Entry Points

Entry points were identified during the motivational or predecisional phase. They were influenced by the context in which symptom checker app use was considered [29]—in addition to the user's own health needs, situations involving the health needs of others were also relevant. Motivational determinants included attentiveness toward symptoms, past experiences (biographical factors), confidence in one's health, trust in the app, and one's ability to respond to symptoms. Some of these factors for app use were also identified by Aboueid et al [4,6] and Meyer et al [7]. The accessibility of health care and availability of social networks were described as external factors. Aboueid et al [4] identified the lack of accessibility of health care as an enabler of app use, whereas social influences were identified as a barrier. The diary study from the CHECK.APP project found that the first occurrence of a symptom and certain symptoms such as heart-, skin-, or eye-related concerns increased the likelihood of symptom checker app use [16]. In

XSL•FO
RenderX

the interview sample, interview partners considered the aforementioned factors in their decision to use the app but did not specify whether they were facilitators or enablers. Although many of the listed factors have been identified in previous research, their role in symptom checker app use is ambiguous as they can act as both barriers and enablers [3,10,21]. As both situational and individual factors shape when and for what purpose symptom checker apps are used, the term "context factors" seems much more appropriate than "barriers and enablers" [3,10,27,28].

Outcomes

Meaningful Learning Experiences

The most common outcome of using symptom checker apps as a digital information source and HISB is the act of learning, which is difficult to measure [1]. This outcome is expressed through the purpose of understanding one's condition based on users' need for more information. Users already have predefined concepts about their condition in mind when planning their course of action. Therefore, they also want to validate these concepts by gathering information. In a substudy of the CHECK.APP project by Wetzel et al [9], it was found that symptom checker apps were the least important for HISB in the general population. The subsample of active symptom checker app users reported that internet use and consulting a physician were more important for HISB than symptom checker apps [9]. In the study by Meyer et al [7], most symptom checker app users stated that they used symptom checker apps to understand the cause of their condition and found the information they received useful. During the evaluation stage, users consider the extent to which the information is meaningful. They base their evaluation on their experiences with symptom checker app use and the health-related actions and determine whether the intended purpose was fulfilled and whether meaningful learning experiences were generated [24,44].

HRB Outcome

Another type of outcome is HRB, which is planned in the intention formation stage and implemented in the following stage. The actions described in the intention implementation stage relate to measurable HRB, such as visiting a health care professional or seeking advice from friends or family [29]. Depending on the purpose of app use, 2 types of behavior can be distinguished in this category.

As a result of using symptom checker apps for the purpose of receiving recommendations for further action, such recommendations are generated. This occurs frequently—in the sample of Meyer et al [7], this purpose ranked second among symptom checker app users.

Our interview partners considered whether to follow app recommendations thoroughly. If recommendations do not align with users' concepts and expectations, they are sometimes discarded as unrealistic, effectively preventing health care visits. This mismatch between concept and recommendation can also cause anxiety, creating the need for more information or confirmation by health care professionals [10,11,16]. Turner et al [50] also observed the rejection of app recommendations, especially if they were given in a context in which the action could not be implemented (eg, out of hours). Verzantvoort et al [24] found that users intended to follow app advice if the advice was to contact their GP during the daytime (75%) and practice self-care (67%). If the concept and recommendation align or if users are very anxious or surprised about the result, a visit to health care professionals is a possible outcome.

To better understand this behavior, a dialogue option could be implemented in symptom checker apps asking the following—"Have you followed my recommendation?"—and whether further resources such as health care professionals have been contacted. That way, not only symptoms but also user behavior and impact of the symptom checker app on health care use could be traced.

Communicating and documenting health-related information serves users' need to structure their experiences and prepare for contact with health care services. This sets up symptom checker apps as a possible interface between the user and health care services. This purpose has not been described yet apart from the CHECK.APP project [10]. Symptom checker apps can increase patient autonomy, as other research has postulated [17]. However, users specifically omit telling their physicians about symptom checker app use.

According to our sample, direct communication with health care professionals over app results and app use was avoided. Aboueid et al [4] found that attitudes toward physicians and the health care system may impact app use. Meyer et al [7] reported that only approximately 50% of patients using symptom checker apps considered informing their physicians about their symptom checker app use. The study indicates that users may share the output of the app, including given recommendations or probable diagnoses. However, they are less likely to inform their physicians that they used a symptom checker app to generate these results. Further research is needed to explore ways to facilitate open communication between health care professionals and symptom checker app users. The quality of the communication and its contents are suggested as measures for this outcome.

Satisfaction With App Use

A third outcome is user satisfaction with the app, which emerges during the evaluation stage. This outcome is closely linked to the app's intended purpose and the 2 previously mentioned outcomes: HRB and learning. As such, satisfaction with symptom checker apps represents a multifaceted and complex construct [13,24].

Satisfaction resulting from app use depends on how well the users perceive the symptom checker app's alignment with their needs, expectations, and intended purpose. Users become insecure and sometimes dissatisfied with the app when their concept does not match the output. This association with user satisfaction indicates that users' expectations also play a role in how they validate the app [24]. In a substudy of the CHECK.APP project, Müller et al [10] distinguished between user expectations and motivations for app use and found that these factors play a role in symptom checker app use. In fact, according to the sample in the study Kopka et al [51], 33% of symptom checker app users reported that the symptom checker

XSL•FO
RenderX

app's usefulness varied. Turner et al [50] conducted a study on an online symptom checker and found that users were dissatisfied with it because too little information on what to do themselves was provided. From their point of view, the outcome did not fit their intended purpose. This should be critical for first-time users but can also influence how users validate the app over time [51].

This indicates that simply measuring a single outcome of app use may be a shortcoming of previous research as it ignores the context of use and the intended purpose. Symptom checker apps provide a wide range of potential services that are integrated into a complex network of resources [4,5,7,13,18,20]. Therefore, researchers and developers must consider addressing the intended purpose of use and expectations and specify the desired outcome.

HL and eHL in Symptom Checker Use

Motivational Stage

The main finding of this study was that both eHL and HL are present throughout every stage of the complex process of symptom checker app use. However, different aspects of these concepts are relevant for different stages. Therefore, we will discuss the expression of eHL and HL in relation to the stages of the symptom checker app use process in the following sections.

In the motivational stage, users are driven primarily by the presence of a health-related concern and the context in which it occurs. Personal resources such as the ability to process health information and access services (as defined by the concept of HL [25]) or to navigate electronic, digital, and mobile health information (as defined by eHL [1]) play a crucial role. However, the application of these skills is influenced and shaped by the surrounding context [27,28]. The description of these skills by the interview partners can also be conceptualized as self-efficacy and self-care, which are interlinked with HL [52]. In a substudy of the CHECK.APP project, Wetzel et al [8] found that, unlike eHL, self-efficacy could be a determinant of app use—however, a sensitivity analysis revealed that the initial correlation was not viable. Kopka et al [51] found that individuals who perceived symptom checker apps as useful had higher self-efficacy. This ambiguity can be resolved by considering the context of use as an additional factor. Some symptoms may be more alarming than others and may exceed the available skills of even those with high HL [16].

Trust in apps, or the lack thereof, plays a role at the motivational stage. Regarding eHL, the interview partners in our sample positioned themselves as having high technical knowledge and the ability to critically appraise app output (eHL). They trusted the app but also their HL skills. This is consistent with the observation made by Kopka et al [11]: users with high eHL are more likely to trust symptom checker apps, possibly because they have higher confidence in their ability to critically evaluate app output. Conversely, lack of trust has been identified as a factor in intentional nonuse [9]. Neither technology affinity nor eHL are predictors that differentiate symptom checker app users from nonusers [8]. Aboueid et al [14] identified different technology affinity profiles as predictors of future app use for self-triage. Both technology affinity and trust represent attitudes rather than skills. They shape expectations and motivation to use apps but are not predictors [10].

In our sample, interview participants did not link their attitudes toward the app with the frequency of its use. Instead, their motivation was shaped by specific and meaningful experiences, such as situations in which the app provided valuable assistance with a critical symptom, subsequently confirmed by a health care professional. Research suggests that factors such as hypochondria [8] and the presence of new or unfamiliar symptoms [16] can drive app use, indicating that eHL and HL influence not only the frequency of app use but also the purposes for which symptom checker apps are used and the manner in which they are used. On the basis of these findings, we conclude that frequency of app use is an oversimplified parameter and should be interpreted with caution in future analyses.

In summary, the skill sets defined in HL and eHL, as well as context factors, influence how the motivational stage unfolds [27,28].

Intention Formation

Symptom checker apps can serve as both an information source and a tool for shaping one's HRB. Navigating digital tools for information gathering and evaluating this information is related to eHL. Applying the information to one's behavior and learning from it is more closely related to HL [24,29,30]. Therefore, planning to use symptom checker apps relates to both eHL and HL [1,25].

Users with high HL may anticipate the app's output and choose not to use it or only use it for information gathering and HISB. In addition, users may experiment with different inputs to observe changes in the output. This demonstrates the interconnection between HL and eHL during the intention formation stage. Users understand the inner workings of the app (eHL) and plan to use it to satisfy their information needs but do not intend to actually follow its advice (HL) [1,53]. According to Kopka et al [11], users are more easily persuaded to follow app recommendations when they are ambivalent or unsure about their own decision. This aligns with interview partners' self-reports that they were less inclined to follow the app's recommendations if they felt that their condition was critical or if the recommendation significantly differed from their own concept (eg, recommendation to seek emergency help when users felt perfectly healthy). In addition, using apps to communicate about symptoms connects the concepts of eHL and HL with a social context in health care [2,26].

Intention Implementation

Users use a range of resources associated with eHL and HL to achieve their intended goals. Key personal resources include the ability to evaluate one's health and navigate the health care system (HL), critically assess information from various platforms, and effectively use and understand digital applications (eHL). These resources are pivotal in shaping how symptom checker apps are integrated into users' actions [1,25,28,49]. In our sample, users expressed confidence in their eHL skills, including media literacy, information literacy, and computer literacy. However, the extent to which they used these skills

XSL•FO
RenderX

depended on their purpose and the availability of other resources.

The demand for applying interactive HL is not met by symptom checker apps because they do not allow for direct communication with others in a social context. This is particularly evident when users seek to validate their own ideas about their condition or receive recommendations for action. In such cases, users may seek out social resources and discuss their condition with friends and family. Alternatively, they may present the app results to a health care professional without disclosing their use of the app. This demonstrates that users extend app use to other contexts, integrating those contexts into their overall user experience [29,48]. This study demonstrated that HRB can transition from digital interaction to social interaction in the context of symptom checker app use. This finding aligns with the concept of distributed HL by Edwards et al [34], highlighting the importance of social networks in realizing the benefits of symptom checker apps. Real-life interaction can provide what the app interaction lacks. The ability to actively listen, notice, and respond to emotional nuances; provide support; and express understanding for the person's situation is also important. According to the observation by Wetzel et al [9], symptom checker app users still refer to friends, family, and physicians for HISB, albeit to a lesser extent than nonusers. Future research in ethics and social sciences will reveal the impact of large language models that mimic empathic responses and conversations on the use of symptom checker apps. Preliminary data suggest that such models may enhance the symptom checker app experience [54].

The interview partners considered symptom checker apps as one of the many digital resources available today. According to users, the app asked dichotomous questions, leaving them with a feeling of not being able to tell their whole story, address ambiguity, or ask questions themselves [24]. However, symptom checker apps offer a more personalized experience than internet searches [4], which remain a core component of HISB even among symptom checker app users [9]. The personal resources that enable users to use the app and apply its results also impact the integration of other digital resources into their HRB [1]. Therefore, there is potential in enhancing interoperability between different technologies such as wearables and apps.

Evaluation

The app's results, whether they provide information or recommendations, may challenge users' preexisting concepts regarding health-related issues. The extent to which users follow these recommendations depends on their preconceived notions about their condition; their expectations for its resolution; and their ability to cope with uncertainty, which is related to HL [10,25].

Users evaluate whether the app results align with the severity of their condition, their expectations of the appropriate course of action, and past experiences. This evaluation requires critical HL [25]. If users feel uncertain due to app use, they may need to contact others, such as health care professionals or their social network, for additional validation outside the app. Individuals with high levels of hypochondria [8] or anxiety [51] may be considered vulnerable groups in this context, particularly if they have low self-efficacy and HL and lack access to social or health care resources.

For the interview partners, symptom checker apps served as an interface among information gathering [4,9], learning and sensemaking [7,43], and interaction with the health care system [9,22]. In summary, the skill set necessary for meaningful interaction with symptom checker apps in the context of health care services is best described using eHL and HL. Together with contextual factors, both models can provide a deeper understanding of the dynamic mechanisms underlying symptom checker app use.

Implications for Health Care and Research With Regard to Symptom Checker App Use

Overreliance on users' HL and eHL results in wasted potential [6]. The findings of this study support the idea that, for the implementation of symptom checker apps in health care, transparency is crucial [7,12]. The lack of transparency and reliance on users' HL and eHL are design flaws in symptom checker apps, which, from a legal standpoint, only present information to be interpreted by the user who provided the input. In a worst-case scenario, users may feel isolated and insecure [3,21]. If the app was designed as a communication tool between physicians and users, such as in teleconsultations, their potential as symptom trackers and decision aids for both user groups could be realized in a safe setting. Users' insecurities could be discussed with a health professional, and possible harm, such as health anxiety, could be reduced.

The skill sets provided by eHL and HL might be the reason why users are able to use symptom checker apps to their health benefit at all despite symptom checker apps' questionable diagnostic accuracy [18,38,55]. Concrete and meaningful experiences with the app shaped users' attitudes toward and evaluation of the app. A pure cause-and-effect relationship, as suggested by its user interface principle (input–algorithm or artificial intelligence–output), is prevented by users' HL, self-efficacy, and self-care. Satisfaction with app use is influenced by users' HL, context, and experiences [10,13,24,29]. Symptom checker app development and research should take into account user expectations, context of use, and the need for social interaction.

Our study uncovered phenomena that require further investigation in the future. These include symptom checker app use in various social settings [24]. Our results show that users frequently engage with symptom checker apps not only for their own health concerns but also to assess symptoms of family members and friends. This social use—such as evaluating a partner's symptoms or a parent using the app for their child—occurs more often than previously reported in the literature. Different technological resources such as wearables, other apps, and internet searches were perceived as possible contributors to the 3 purposes by the interview partners. Thus, research on the topic should consider the entire landscape of digital resources and their interoperability.

Limitations

Experience-based research is limited to the information that interview participants are willing to share about themselves. It

XSL•FO
RenderX

is not possible to directly observe the psychological processes that occur in someone's mind. However, by collecting and synthesizing users' stories and perspectives, we were able to describe a general process. The Rubicon model helped us organize the stories coherently and with a focus on motivation.

It is important to critically evaluate the method and sample used. It should be noted that, even in our sample of predominantly younger, White, well-educated female users, each individual had their own perspective on how the app fulfilled its purpose in the described contexts. While a general process could be derived, the outcomes were highly dependent on the user. In addition, the results are primarily applicable to the context of the German health care system and, therefore, can only be transferred to similar health care systems, which is a common challenge in symptom checker app research [38].

Our sample consisted of predominantly very reflective individuals with presumably high HL. Our sampling strategy to approach users based on diverse characteristics could only partly be implemented because the sample from which we recruited the interview partners represented the current major user group of symptom checker apps: young, female, and well educated individuals [5,9,17]. Thus, the sample reflects the most typical characteristics of symptom checker app users, indicating that the results are likely applicable to this core group [3,9].

While the mixed methods approach of the CHECK.APP project allowed for the description of this population from different perspectives using different methods, it limits the applicability of our results to individuals with less HL. We can neither make assumptions about nor extend our findings to marginalized groups excluded from app use [15,56].

The integrative basic method builds on the tradition of interpretative-reconstructive qualitative methods such as ethnomethodological conversation analysis and narrative analysis. It proved to be suitable to analyze how medical laypersons make sense of the symptom checker app. While we could show in another study of the CHECK.APP project how GPs position symptom checker apps in the black box of the "unorganized stage" of patients' reflections on their symptoms [40], our study succeeded in shedding light on this exact black box.

Another limitation of this study is the focus on just 1 app. While we believe to have extrapolated generalizable information, it is possible that other apps may yield different user experiences and human-app interactions. In addition, due to the pandemic, the interviews were conducted on the web via videoconferencing software. It is possible that conducting in-person interviews would have attracted different interviewees and revealed additional information due to the different setting.

While software-supported tools such as MAXQDA or f4analyse exist to support qualitative analysis and the authors and analysts are well acquainted with them, we deliberately chose not to use these tools especially for the interpretative parts of the analysis because their workflow is optimized for categorization of qualitative data, not so much for in-depth interpretative analyses. To ensure that our operationalization remained understandable, we provided data examples in the results text.

Conclusions

In our qualitative interview study, we could demonstrate that, from symptom checker app users' perspective, a simple cause-and-effect relationship between symptoms and symptom checker app use is unlikely, at least in individuals with high eHL and HL. Rather, symptom checker app use is described as a complex, cyclic process. Context-dependent and biographical factors, as well as the dynamic concepts of eHL and HL, were expressed in users' descriptions.

According to our limited sample, symptom checker apps are used for 3 distinct purposes: understanding one's condition, receiving recommendations for action, and communicating on and documenting health-related information. Each purpose warrants its own planning, implementation, and evaluation in HRB. Each purpose may be implemented using different personal, social, and technological resources.

Symptom checker apps have shortcomings with respect needs related to interactive HL. Users seek external validation of app findings in their social networks and professional health care services. However, symptom checker apps have the potential to become an interface between users and health care services. This potential has not been realized, which should be a design goal for their continued development.

Acknowledgments

This study was funded by the Federal Ministry of Education and Research (grant 01GP1907A). The sponsor had no influence on the study design, data collection, or its analysis or interpretation. The authors thank several of their interview partners (who will remain anonymous for privacy reasons) for their contribution to the member check workshop in which the aforementioned results were presented to and discussed with the participants. The generative artificial intelligence tool ChatGPT by OpenAI was used to revise the language and style of the manuscript, which were further reviewed and revised by the study group. The authors acknowledge support from the open access publication fund of the University of Tübingen.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request. Due to data protection regulations, interview data can only be made available if the study participant associated with the data approves the request.

XSL•FO
RenderX

Authors' Contributions

RK drafted the manuscript together with MTS. MK, RM, RK, and AJW conducted the interviews. CP provided vital methodological input for analysis. MTS, RK, AJW, MK, and CP met regularly in multidisciplinary methodological workshops, which were led by CP. All listed authors revised the manuscript. This process was coordinated by RK. All authors approved the final manuscript draft before submission.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Standards for Reporting Qualitative Research statement checklist.
[PDF File (Adobe PDF File), 226 KB - formative_v9i1e60647_app1.pdf]

Multimedia Appendix 2
Interview guide.
[DOCX File , 24 KB - formative_v9i1e60647_app2.docx]

References

1. Norman CD, Skinner HA. eHealth literacy: essential skills for consumer health in a networked world. J Med Internet Res 2006 Jun 16;8(2):e9 [FREE Full text] [doi: 10.2196/jmir.8.2.e9] [Medline: 16867972]
2. Levin-Zamir D, Bertschi I. Media health literacy, eHealth literacy, and the role of the social environment in context. Int J Environ Res Public Health 2018 Aug 03;15(8):1643 [FREE Full text] [doi: 10.3390/ijerph15081643] [Medline: 30081465]
3. Müller R, Klemmt M, Ehni HJ, Henking T, Kuhnmünch A, Preiser C, et al. Ethical, legal, and social aspects of symptom checker applications: a scoping review. Med Health Care Philos 2022 Dec;25(4):737-755 [FREE Full text] [doi: 10.1007/s11019-022-10114-y] [Medline: 36181620]
4. Aboueid S, Meyer S, Wallace JR, Mahajan S, Chaurasia A. Young adults' perspectives on the use of symptom checkers for self-triage and self-diagnosis: qualitative study. JMIR Public Health Surveill 2021 Jan 06;7(1):e22637 [FREE Full text] [doi: 10.2196/22637] [Medline: 33404515]
5. Aboueid S, Liu RH, Desta BN, Chaurasia A, Ebrahim S. The use of artificially intelligent self-diagnosing digital platforms by the general public: scoping review. JMIR Med Inform 2019 May 01;7(2):e13445 [FREE Full text] [doi: 10.2196/13445] [Medline: 31042151]
6. Aboueid S, Meyer SB, Wallace JR, Mahajan S, Nur T, Chaurasia A. Use of symptom checkers for COVID-19-related symptoms among university students: a qualitative study. BMJ Innov 2021 Apr;7(2):253-260. [doi: 10.1136/bmjinnov-2020-000498] [Medline: 34192014]
7. Meyer AN, Giardina TD, Spitzmueller C, Shahid U, Scott TM, Singh H. Patient perspectives on the usefulness of an artificial intelligence-assisted symptom checker: cross-sectional survey study. J Med Internet Res 2020 Jan 30;22(1):e14679 [FREE Full text] [doi: 10.2196/14679] [Medline: 32012052]
8. Wetzel AJ, Klemmt M, Müller R, Rieger MA, Joos S, Koch R. Only the anxious ones? Identifying characteristics of symptom checker app users: a cross-sectional survey. BMC Med Inform Decis Mak 2024 Jan 23;24(1):21 [FREE Full text] [doi: 10.1186/s12911-024-02430-5] [Medline: 38262993]
9. Wetzel AJ, Koch R, Koch N, Klemmt M, Müller R, Preiser C, et al. 'Better see a doctor?' Status quo of symptom checker apps in Germany: a cross-sectional survey with a mixed-methods design (CHECK.APP). Digit Health 2024;10:20552076241231555 [FREE Full text] [doi: 10.1177/20552076241231555] [Medline: 38434790]
10. Müller R, Klemmt M, Koch R, Ehni HJ, Henking T, Langmann E, et al. "That's just Future Medicine" - a qualitative study on users' experiences of symptom checker apps. BMC Med Ethics 2024 Feb 16;25(1):17 [FREE Full text] [doi: 10.1186/s12910-024-01011-5] [Medline: 38365749]
11. Kopka M, Schmieding ML, Rieger T, Roesler E, Balzer F, Feufel MA. Determinants of laypersons' trust in medical decision aids: randomized controlled trial. JMIR Hum Factors 2022 May 03;9(2):e35219 [FREE Full text] [doi: 10.2196/35219] [Medline: 35503248]
12. Miller S, Gilbert S, Virani V, Wicks P. Patients' utilization and perception of an artificial intelligence-based symptom assessment and advice technology in a British primary care waiting room: exploratory pilot study. JMIR Hum Factors 2020 Jul 10;7(3):e19713 [FREE Full text] [doi: 10.2196/19713] [Medline: 32540836]
13. Chambers D, Cantrell AJ, Johnson M, Preston L, Baxter SK, Booth A, et al. Digital and online symptom checkers and health assessment/triage services for urgent health problems: systematic review. BMJ Open 2019 Aug 01;9(8):e027743 [FREE Full text] [doi: 10.1136/bmjopen-2018-027743] [Medline: 31375610]
14. Aboueid S, Meyer SB, Wallace J, Chaurasia A. Latent classes associated with the intention to use a symptom checker for self-triage. PLoS One 2021;16(11):e0259547 [FREE Full text] [doi: 10.1371/journal.pone.0259547] [Medline: 34731217]

XSL•FO
RenderX

15. Luger TM, Houston TK, Suls J. Older adult experience of online diagnosis: results from a scenario-based think-aloud protocol. J Med Internet Res 2014 Jan 16;16(1):e16 [FREE Full text] [doi: 10.2196/jmir.2924] [Medline: 24434479]
16. Wetzel AJ, Preiser C, Müller R, Joos S, Koch R, Henking T, et al. Unveiling usage patterns and explaining usage of symptom checker apps: explorative longitudinal mixed methods study. J Med Internet Res 2024 Dec 09;26:e55161 [FREE Full text] [doi: 10.2196/55161] [Medline: 39652845]
17. Arellano Carmona K, Chittamuru D, Kravitz RL, Ramondt S, Ramírez AS. Health information seeking from an intelligent web-based symptom checker: cross-sectional questionnaire study. J Med Internet Res 2022 Aug 19;24(8):e36322 [FREE Full text] [doi: 10.2196/36322] [Medline: 35984690]
18. Wallace W, Chan C, Chidambaram S, Hanna L, Iqbal FM, Acharya A, et al. The diagnostic and triage accuracy of digital and online symptom checker tools: a systematic review. NPJ Digit Med 2022 Aug 17;5(1):118 [FREE Full text] [doi: 10.1038/s41746-022-00667-w] [Medline: 35977992]
19. Schmieding ML, Kopka M, Schmidt K, Schulz-Niethammer S, Balzer F, Feufel MA. Triage accuracy of symptom checker apps: 5-year follow-up evaluation. J Med Internet Res 2022 May 10;24(5):e31810 [FREE Full text] [doi: 10.2196/31810] [Medline: 35536633]
20. Semigran HL, Linder JA, Gidengil C, Mehrotra A. Evaluation of symptom checkers for self diagnosis and triage: audit study. BMJ 2015 Jul 08;351:h3480 [FREE Full text] [doi: 10.1136/bmj.h3480] [Medline: 26157077]
21. Pairon A, Philips H, Verhoeven V. A scoping review on the use and usefulness of online symptom checkers and triage systems: how to proceed? Front Med (Lausanne) 2022 Jan 6;9:1040926 [FREE Full text] [doi: 10.3389/fmed.2022.1040926] [Medline: 36687416]
22. Ceney A, Tolond S, Glowinski A, Marks B, Swift S, Palser T. Accuracy of online symptom checkers and the potential impact on service utilisation. PLoS One 2021;16(7):e0254088 [FREE Full text] [doi: 10.1371/journal.pone.0254088] [Medline: 34265845]
23. Knitza J, Muehlensiepen F, Ignatyev Y, Fuchs F, Mohn J, Simon D, et al. Patient's perception of digital symptom assessment technologies in rheumatology: results from a multicentre study. Front Public Health 2022;10:844669 [FREE Full text] [doi: 10.3389/fpubh.2022.844669] [Medline: 35273944]
24. Verzantvoort NC, Teunis T, Verheij TJ, van der Velden AW. Self-triage for acute primary care via a smartphone application: practical, safe and efficient? PLoS One 2018;13(6):e0199284 [FREE Full text] [doi: 10.1371/journal.pone.0199284] [Medline: 29944708]
25. Nutbeam D. Health literacy as a public health goal: a challenge for contemporary health education and communication strategies into the 21st century. Health Promot Int 2000;15(3):259-267. [doi: 10.1093/heapro/15.3.259]
26. Nutbeam D, Lloyd JE. Understanding and responding to health literacy as a social determinant of health. Annu Rev Public Health 2021 Apr 01;42:159-173 [FREE Full text] [doi: 10.1146/annurev-publhealth-090419-102529] [Medline: 33035427]
27. Parker RM, Ratzan S. Re-enforce, not re-define health literacy-moving forward with health literacy 2.0. J Health Commun 2019 Nov 21;24(12):923-925. [doi: 10.1080/10810730.2019.1691292] [Medline: 31752603]
28. Parker R, Ratzan SC. Health literacy: a second decade of distinction for Americans. J Health Commun 2010 Aug 31;15 Suppl 2(sup2):20-33. [doi: 10.1080/10810730.2010.501094] [Medline: 20845190]
29. von Wagner C, Steptoe A, Wolf MS, Wardle J. Health literacy and health actions: a review and a framework from health psychology. Health Educ Behav 2009 Oct;36(5):860-877. [doi: 10.1177/1090198108322819] [Medline: 18728119]
30. Neter E, Brainin E. Association between health literacy, eHealth literacy, and health outcomes among patients with long-term conditions. Eur Psychol 2019 Jan;24(1):68-81. [doi: 10.1027/1016-9040/a000350]
31. Paasche-Orlow MK, Wolf MS. The causal pathways linking health literacy to health outcomes. Am J Health Behav 2007;31 Suppl 1:S19-S26. [doi: 10.5555/ajhb.2007.31.supp.S19] [Medline: 17931132]
32. Stormacq C, Van den Broucke S, Wosinski J. Does health literacy mediate the relationship between socioeconomic status and health disparities? Integrative review. Health Promot Int 2019 Oct 01;34(5):e1-17. [doi: 10.1093/heapro/day062] [Medline: 30107564]
33. Stormacq C, Wosinski J, Boillat E, Van den Broucke S. Effects of health literacy interventions on health-related outcomes in socioeconomically disadvantaged adults living in the community: a systematic review. JBI Evid Synth 2020 Jul;18(7):1389-1469. [doi: 10.11124/JBISRIR-D-18-00023] [Medline: 32813388]
34. Edwards M, Wood F, Davies M, Edwards A. 'Distributed health literacy': longitudinal qualitative analysis of the roles of health literacy mediators and social networks of people living with a long-term health condition. Health Expect 2015 Oct;18(5):1180-1193 [FREE Full text] [doi: 10.1111/hex.12093] [Medline: 23773311]
35. Sørensen K, Pelikan JM, Röthlin F, Ganahl K, Slonska Z, Doyle G, et al. Health literacy in Europe: comparative results of the European health literacy survey (HLS-EU). Eur J Public Health 2015 Dec 05;25(6):1053-1058 [FREE Full text] [doi: 10.1093/eurpub/ckv043] [Medline: 25843827]
36. Berkman ND, Sheridan SL, Donahue KE, Halpern DJ, Crotty K. Low health literacy and health outcomes: an updated systematic review. Ann Intern Med 2011 Jul 19;155(2):97-107. [doi: 10.7326/0003-4819-155-2-201107190-00005] [Medline: 21768583]
37. Fiske A, Buyx A, Prainsack B. The double-edged sword of digital self-care: physician perspectives from Northern Germany. Soc Sci Med 2020 Sep;260:113174 [FREE Full text] [doi: 10.1016/j.socscimed.2020.113174] [Medline: 32659512]

XSL•FO
RenderX

38. Ilicki J. Challenges in evaluating the accuracy of AI-containing digital triage systems: a systematic review. PLoS One 2022 Dec 27;17(12):e0279636 [FREE Full text] [doi: 10.1371/journal.pone.0279636] [Medline: 36574438]
39. Wetzel AJ, Koch R, Preiser C, Müller R, Klemmt M, Ranisch R, et al. Ethical, legal, and social implications of symptom checker apps in primary health care (CHECK.APP): protocol for an interdisciplinary mixed methods study. JMIR Res Protoc 2022 May 16;11(5):e34026. [doi: 10.2196/34026]
40. Preiser C, Radionova N, Ög E, Koch R, Klemmt M, Müller R, et al. The doctors, their patients, and the symptom checker app: qualitative interview study with general practitioners in Germany. JMIR Hum Factors 2024 Nov 18;11:e57360 [FREE Full text] [doi: 10.2196/57360] [Medline: 39556813]
41. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014 Sep;89(9):1245-1251 [FREE Full text] [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]
42. Malterud K, Siersma VD, Guassora AD. Sample size in qualitative interview studies: guided by information power. Qual Health Res 2016 Nov 10;26(13):1753-1760. [doi: 10.1177/1049732315617444] [Medline: 26613970]
43. Bruner J. Acts of Meaning. Cambridge, MA: Harvard University Press; 1990.
44. Kolb DA. Experiential Learning: Experience as the Source of Learning and Development. Hoboken, NJ: Prentice-Hall; 1984.
45. Kruse J. Qualitative Interviewforschung: Ein integrativer Ansatz. Weinheim, Germany: Beltz Juventa; 2014.
46. Davies B, Harré R. Positioning: the discursive production of selves. J Theory Soc Behav 2007 Jul;20(1):43-63. [doi: 10.1111/j.1468-5914.1990.tb00174.x]
47. McAdams DP. Narrative identity. In: Schwartz SJ, Luyckx K, Vignoles VL, editors. Handbook of Identity Theory and Research. Berlin, Germany: Springer Science + Business Media; 2011:99-115.
48. Achtziger A, Gollwitzer PM. Motivation and volition in the course of action. In: Heckhausen J, Heckhausen H, editors. Motivation and Action. Cambridge, UK: Cambridge University Press; 2008:272-295.
49. Parker R. Measuring health literacy: what? So what? Now what? In: Measures of Health Literacy: Workshop Summary. Washington, DC: National Academies Press (US); 2009:91-102.
50. Turner J, Knowles E, Simpson R, Sampson F, Dixon S, Long J, et al. Impact of NHS 111 Online on the NHS 111 telephone service and urgent care system: a mixed-methods study. Health Serv Deliv Res 2021 Nov;9(21):1-148. [doi: 10.3310/hsdr09210] [Medline: 34780129]
51. Kopka M, Scatturin L, Napierala H, Fürstenau D, Feufel MA, Balzer F, et al. Characteristics of users and nonusers of symptom checkers in Germany: cross-sectional survey study. J Med Internet Res 2023 Jun 20;25:e46231 [FREE Full text] [doi: 10.2196/46231] [Medline: 37338970]
52. Lee EH, Lee YW, Moon SH. A structural equation model linking health literacy to self-efficacy, self-care activities, and health-related quality of life in patients with type 2 diabetes. Asian Nurs Res (Korean Soc Nurs Sci) 2016 Mar;10(1):82-87 [FREE Full text] [doi: 10.1016/j.anr.2016.01.005] [Medline: 27021840]
53. Neter E, Brainin E. eHealth literacy: extending the digital divide to the realm of health information. J Med Internet Res 2012 Jan 27;14(1):e19 [FREE Full text] [doi: 10.2196/jmir.1619] [Medline: 22357448]
54. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023 Jun 01;183(6):589-596 [FREE Full text] [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
55. Schmieding ML, Mörgeli R, Schmieding MA, Feufel MA, Balzer F. Benchmarking triage capability of symptom checkers against that of medical laypersons: survey study. J Med Internet Res 2021 Mar 10;23(3):e24475 [FREE Full text] [doi: 10.2196/24475] [Medline: 33688845]
56. Wang X, Luan W. Research progress on digital health literacy of older adults: a scoping review. Front Public Health 2022;10:906089 [FREE Full text] [doi: 10.3389/fpubh.2022.906089] [Medline: 35991040]

Abbreviations

CHECK.APP: Ethical, Legal, and Social Implications of Symptom Checker Apps in Primary Health Care
eHL: eHealth literacy
GP: general practitioner
HISB: health information–seeking behavior
HL: health literacy
HRB: health-related behavior

XSL•FO
RenderX

Edited by A Mavragani; submitted 17.05.24; peer-reviewed by S Saur, SC Chong; comments to author 25.08.24; revised version received 28.11.24; accepted 08.02.25; published 21.03.25.

Please cite as:
Koch R, Steffen MT, Wetzel AJ, Preiser C, Klemmt M, Ehni HJ, Mueller R, Joos S
Exploring Laypersons' Experiences With a Mobile Symptom Checker App as an Interface Between eHealth Literacy, Health Literacy, and Health-Related Behavior: Qualitative Interview Study
JMIR Form Res 2025;9:e60647
URL: https://formative.jmir.org/2025/1/e60647
doi:10.2196/60647
PMID:

©Roland Koch, Marie-Theres Steffen, Anna-Jasmin Wetzel, Christine Preiser, Malte Klemmt, Hans-Jörg Ehni, Regina Mueller, Stefanie Joos. Originally published in JMIR Formative Research (https://formative.jmir.org), 21.03.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Exploring Physicians' Dual Perspectives on the Transition From Free Text to Structured and Standardized Documentation Practices: Interview and Participant Observational Study

Olga Golburean[1], MPhil; Rune Pedersen[2], PhD; Line Melby[3], PhD; Arild Faxvaag[1], PhD

[1]Department of Neuromedicine and Movement Science, Faculty of Medicine and Health Sciences, NTNU-Norwegian University of Science and Technology, Trondheim, Norway

[2]Norwegian Centre for E-Health Research, University Hospital of North Norway, Tromsø, Norway

[3]Department of Health Research, SINTEF Digital, Trondheim, Norway

Corresponding Author:
Olga Golburean, MPhil
Department of Neuromedicine and Movement Science
Faculty of Medicine and Health Sciences
NTNU-Norwegian University of Science and Technology
Elgeseter Gate 28
Trondheim, 7030
Norway
Phone: 47 91787389
Email: olga.golburean@ntnu.no

Abstract

Background: Clinical documentation plays a crucial role in providing and coordinating care. Despite the widespread adoption of electronic health record (EHR) systems, many end users still document clinical data in a manner similar to traditional paper-based records. To fully leverage the benefits of EHR systems, it is necessary to adopt new documentation approaches that facilitate easy access to information at the point of care and seamless exchange of information across health care facilities.

Objective: We aimed to evaluate how the transition from an older EHR system to a cross-institutional EHR system impacts physicians' documentation practices and gain a deeper understanding of the factors influencing their choice between free text and structured and standardized documentation methods.

Methods: A qualitative study was conducted between September 2023 and January 2024. It involved participant observations and individual semistructured interviews with physicians at a university hospital in Norway. Data were analyzed using reflexive thematic analysis.

Results: The analysis revealed 3 main themes. First, physicians encountered challenges during the implementation phase of the new EHR system due to its complexity and their unfamiliarity with its use. However, with time, physicians gradually adopted new documentation processes. This integration or adoption primarily occurred by learning through practical experience and collaborative knowledge exchange with their peers. Second, although the implementation of the new EHR system had increased structured and standardized clinical documentation, free text remained the preferred method, with some exceptions. In addition, the fact that many physicians still relied on free-text documentation created a sense of distrust among them toward some of the standardized clinical data. Finally, the informants had mixed perceptions of Systematized Nomenclature of Medicine–Clinical Terms. Some viewed it as a more nuanced terminology system, while others found it more complex. Most informants found using templates for routine procedures beneficial as it saved time in the documentation process and ensured that all necessary parameters and documentation requirements were met.

Conclusions: The study findings revealed that physicians' acceptance of new documentation processes is influenced by various social and technological factors. These factors include previous documentation experiences, perceived benefits, familiarity with the EHR system, time constraints, and user-friendliness of the system. While physicians generally have a positive attitude toward using templates for routine procedures, they often create their own templates, and data within these templates are documented in a free-text format. To address this, health care organizations should consider implementing common standardized or semistandardized templates to reduce disparities in documentation, enhance data recording, and ensure adherence to guidelines.

XSL•FO
RenderX

Furthermore, to facilitate the transition to the new documentation processes, we recommend providing physicians with customized training programs and platforms for tacit knowledge exchange.

(JMIR Form Res 2025;9:e63902) doi:10.2196/63902

KEYWORDS

documentation; documentation practice; standardized documentation; structured documentation; secondary use of data; interoperability; electronic health record; EHR

Introduction

Background

Electronic health record (EHR) systems have been advocated as means to enhance the efficiency of health care services, facilitating improved coordination and delivery of safe and evidence-based care [1,2]. Data from an EHR system can also provide health professionals with the information they need to make informed decisions at the point of care and the opportunity to reflect on their clinical work and learn from past experiences [3]. In addition, an EHR system facilitates access to data collected during clinical care, which may be used for secondary purposes (eg, research and augmenting the capabilities of clinical decision support systems) [4-6]. Nevertheless, clinical data need to be captured in a fixed structure in order to facilitate their use [4]. Health professionals play a crucial role in the documentation process, facilitating the accumulation of data in the EHR system. However, the implementation of the EHR system often results in end users still recording clinical data in a manner reminiscent of traditional paper-based records, although in a digital format, using computers and keyboards [7]. Entering structured and standardized data into an EHR system can be time-consuming, and the level of acceptance among different end users can vary greatly. Health professionals, frequently constrained by time limitations, might be reluctant to assume the data entry task unless they see significant benefits [8,9].

Clinical Documentation in EHR Systems

Health care systems are bound by stringent laws and regulations. In Norway, the use of electronic health data for the provision of health care services (ie, the primary use of health data) is regulated by the Patient Record Act [10]. This act aims to ensure that patients receive high-quality health care by making health data available for health professionals in a quick, effective, and secure manner. Health professionals, on their side, have a duty to meet documentation requirements as outlined in the Health Personnel Act [11]. This act specifically mandates that patient journals must be maintained in accordance with high professional standards, ensuring that they include pertinent and essential information about the patient and the health care services rendered. Furthermore, the journals must be easily comprehensible for other qualified health care personnel. The use of electronic health data for purposes such as statistics, public health, health surveillance, research and product development, education, or teaching (ie, the secondary use of health data) is also strictly regulated [10,12].

The adoption of the EHR system has significantly expanded the volume of clinical data available. However, a considerable portion of these data remains underused, largely due to the dominance of unstructured documentation, such as free-text notes and narrative reports [7]. It is estimated that 80% of the EHR data exist in unstructured documents [13,14]. Consequently, there is increasing pressure on health professionals to adopt structured and standardized documentation methods [6]. This involves the use of standardized coding systems for data entry, the incorporation of models and terminologies such as Systematized Nomenclature of Medicine–Clinical Terms (SNOMED CT), and the use of structured forms [6,15]. Structured and standardized data encompass various patient details, including demographics, laboratory tests, height, weight, blood pressure, medications, allergies, and more [16,17]. Since structured and standardized data types are organized within a fixed structure, they can be easily analyzed using statistical or machine learning methods [16]. Furthermore, data from EHR systems also play a crucial role in the realization of learning health systems, where data are captured as a by-product of care. These data are subsequently analyzed to generate new knowledge, which, in turn, informs continuous improvement and innovation in health care service delivery [18,19]. While on the one hand, structured and standardized documentation practices are associated with enhanced quality of clinical notes [20,21], on the other hand, research highlights that the adoption of the EHR system has led to increased documentation time and reduced attention to patient care [6]. Moreover, structured and standardized documentation can impede expressivity in notes. Free-text documentation is crucial in health care delivery as it allows health professionals to record detailed and nuanced patient narratives. Thus, Rosenbloom et al [22] advocated that health professionals should have the flexibility to decide their documentation style according to workflow needs and content specifics. Their study highlighted that structured documentation is ideal for scenarios requiring data reuse, whereas narrative documentation is suitable when data reuse is not a priority. Furthermore, Levy et al [23] emphasized that health professionals encounter an excessive documentation burden when the usability of the documentation systems fails to adequately support patient care delivery. This highlights the crucial role of systems' usability and the need to assess their impact on documentation practices.

The aim of this study is to evaluate how the transition from an older EHR system to a newer cross-institutional EHR system, which enables structured and standardized documentation, impacts physicians' documentation practices. It seeks to gain insights into the factors that impact physicians' choice of documentation methods, such as free text or structured and standardized documentation.

XSL•FO
RenderX

Theory

Standardizing work processes in health care, such as implementing standardized guidelines for procedures and treatments, is regarded as a useful approach to improving the quality of health care due to its ability to improve outcomes, reduce the likelihood of errors, and eliminate unnecessary variations [24-26]. The increased adoption of EHR systems has led to increased interest in integrating standardized clinical tools within these systems. This includes, among others, standards pertaining to clinical documentation, encompassing what, where, and how information should be documented (eg, structured text, use of templates, and narrative text) [27-30]. Standardizing the use of the EHR system contributes to achieving interoperable EHR system, enabling the seamless exchange of clinical data among various health care organizations [31-33]. Nevertheless, the pursuit of standardization in health care, particularly in the realm of EHR systems, has proven to be challenging [31,34]. This is due to the highly heterogeneous nature of health care, marked by the constant occurrence of unforeseen circumstances. A significant portion of health care professionals' responsibilities cannot be strictly defined by standardized procedures alone, as a substantial amount of their work is tailored to address the unique requirements of individual patients [35,36]. Thus, it has been suggested that providing flexibility is essential in health care work, as it enables health care professionals to customize their work to factors such as patient needs, good judgment, workflow, and prevailing circumstances [24]. Cherns [37] further emphasized the importance of incorporating sufficient flexibility within technical systems. This entails the ability to tailor the system according to specific local needs and adapt to evolving requirements [33,37]. This is justifiable, as Orlikowski [38] highlighted the dynamic nature of technologies, which are continually shaped by human responses to a multitude of influences, such as environmental, political, and cultural factors. Nevertheless, it is important to note that people have their own distinct viewpoints, "practice lenses" that impact how they use the systems [33,38], and may also rely on their own meanings and interpretations of the technology's functions, known as "interpretative flexibility," which may deviate from the designer's original intentions [39,40]. Considering the aforementioned information, it is notable that the implementation of EHR systems and the standardized processes they encompass (eg, clinical documentation) extend beyond being mere technological tools. Instead, EHR systems are integral components of a sociotechnical system, which encompasses a dynamic interaction between technology, health professionals, organizational structures, and the social environment in which they are embedded [41-43]. Thus, literature has previously emphasized the significance of using the sociotechnical system theory [44] in the design, development, implementation, use, and evaluation of health information technology [45] and applied it to examine its impact on documentation practices [46].

The Context

Although all health care organizations in Norway use an EHR system, many of these systems are not adequately designed for structured and standardized documentation [47]. The daily documentation primarily comprises unstructured notes written in free text, which makes it difficult to use for secondary purposes, such as quality improvements. Moreover, in many cases, medications are manually recorded and signed on paper-based charts before being scanned as images into the EHR system [47]. According to Norwegian health authorities, health professionals lack proper access to patient health information when needed at the point of care [48]. This challenge is primarily attributed to the fragmented nature of the health care systems, coupled with the insufficient integration of various information and communication technology systems [48]. Thus, in 2012, the Norwegian government launched the "One citizen—one health record" initiative. Its goal is to facilitate the seamless integration and accessibility of essential health information at every stage of patient care, regardless of when or where it is required [48]. As part of this initiative, the Health Platform (Helseplattformen) project was established [49], with the aim to replace the existing fragmented EHR systems with an integrated solution for both primary and specialist health professionals in the entire central Norway region [50,51]. The project selected Epic Systems Corporation as the vendor for the EHR system [49].

Health Platform aimed to shift toward more structured and standardized documentation of clinical data in order to achieve optimal system functionality and make data machine readable. A pivotal change within the Health Platform project is the transition from the *International Classification of Diseases, 10th Revision* (*ICD-10*) to the more comprehensive and nuanced terminology system, SNOMED CT [52]. The complexity of SNOMED CT arises from a dense hierarchical structure and "is-a" relationships that link broader and more specific concepts, as explored by Abeysinghe et al [53]. This allows SNOMED CT to capture nuances that provide better clinical coverage, making it more suitable for the EHR system and supporting patient care, compared to *ICD*, which is mainly used for diagnoses and billing [54]. The lack of detail in *ICD* codes can lead to a loss of specificity in a clinical context, which means that rare conditions can be grouped under a single code, thus impacting the EHR system's ability to provide accurate clinical decision support [54].

Methods

Study Design and Settings

We used a descriptive qualitative research method to conduct this case study based on participant observations and individual semistructured interviews with physicians. Qualitative research methods enable researchers to delve deeply into the subject matter, providing comprehensive insights and detailed descriptions from the participants' perspectives. Unlike quantitative methods, qualitative approaches, such as interviews and observations, excel at offering detailed explanations, particularly when investigating questions pertaining to *why* and *how* [35,55].

We conducted this study in 1 surgical department at a large university hospital in the central Norway region, which transitioned from a facility-centered EHR system to Health Platform, the new cross-institutional EHR system (Epic Systems Corporation), in November 2022. The university hospital houses

XSL•FO
RenderX

approximately 1000 beds and employs around 11,000 staff members. The surgical department includes an outpatient clinic, inpatient ward, surgical unit, and emergency room, complemented by a specialized division for pediatrics.

Recruitment of Participants

Participants for this study were recruited through verbal and written approaches. First, an email was sent to all physicians working within the surgical department, requesting their participation in an observational study. Second, recognizing the challenges of recruiting busy physicians, particularly during the transition to the new EHR system, the first author (OG) attended the physicians' morning meeting to provide additional information about the project and give them the opportunity to ask direct questions about the study. The study aimed to recruit participants with varying levels of professional experience, including experienced physicians and interns, to ensure a diverse range of perspectives. After completing the observational study phase, a subsequent email invitation was sent to the same group of individuals (ie, physicians in the surgical department), inviting them to participate in individual interviews. The first author (OG) reached out to physicians interested in participating in the study to schedule a convenient time and location for their meeting. In total, 14 physicians participated in the study, and the characteristics of the participants are summarized in Table 1.

Table 1. Demographic characteristics of the study participants (N=14).

Characteristics	Participants, n (%)
Age group (y)	
25-39	9 (64)
≥40	5 (36)
Sex	
Female	6 (43)
Male	8 (57)
Practice (y)	
<5	3 (21)
5-15	7 (50)
>15	4 (29)

Data Collection

Observations

Participant observations occurred from September 2023 to November 2023. In total, 12 physicians were observed for a combined duration of approximately 44 hours. Figure 1 depicts the information regarding data collection activities. The observation sessions ranged from 1 to 3 times, with each session varying in duration from 1 hour to 3 hours. The observation was conducted at the surgical department at several health care facilities where the physicians were working, including an outpatient clinic, an inpatient ward, and an emergency room. Before these observations, a taxonomy was developed to guide the observational process. It was developed based on the research question, a review of existing literature [56-58], and collaborative inputs from field experts (AF, LM, and RP). The primary focus during observations was on physicians' interactions with the new EHR system, with particular attention given to their documentation practices. Throughout each observational session, the observer (OG) recorded notes, capturing both the physicians' activities and, when feasible, engaging them in dialogue about their EHR use. Following observations, the observer supplemented these notes with additional remarks and personal reflections.

Figure 1. Data collection activities. EHR: electronic health record.

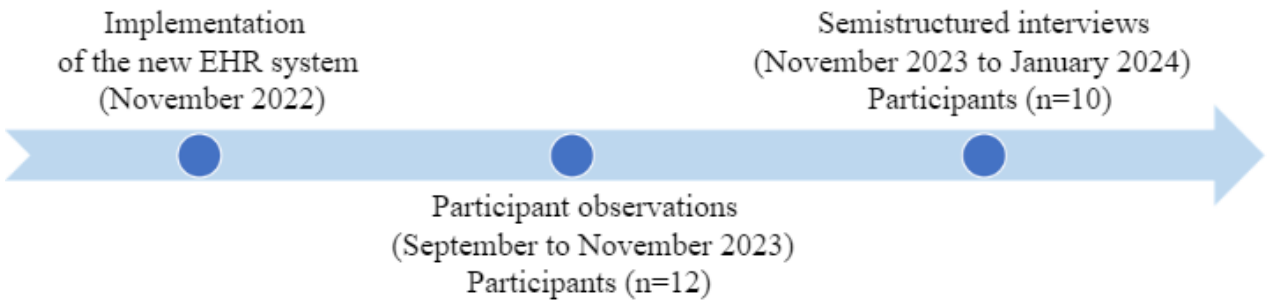

Interviews

After the completion of the observations, individual, semistructured interviews were conducted face-to-face with the participants by the first author (OG). In total, 10 physicians were interviewed, 8 (80%) of whom had also been part of previous observations. The interviews were conducted between November 2023 and January 2024 (Figure 1). The interviews

XSL•FO
RenderX

lasted between 34 and 60 minutes, with an average duration of 44 (SD 9.6) minutes. The interview guide, developed based on the existing literature [58-60] and expert collaboration in the field, was initially outlined before the observational phase. However, following the observations, which provided in-depth insights into the physicians' work practices with the new EHR system, the interview guide was revised to incorporate emerging, relevant questions. Multimedia Appendix 1 provides the interview guide. Each interview was audio recorded and transcribed verbatim by the first author (OG). The process of transcription began immediately after the first interview and was integral to the preliminary data analysis phase, which involved a thorough familiarization with the collected data. The interview phase concluded once data saturation was achieved. This happened when further interviews no longer provided any new insights or themes.

Data Analysis

Data were organized using NVivo (Lumivero) software [61] and analyzed through a reflexive thematic analysis approach [62]. The transcripts of the interviews and field notes from the observation sessions were thoroughly examined to familiarize them with the data. Then, specific codes were developed using an inductive approach, meaning that the codes were derived from the data. The initial codes were created using a semantic coding approach. These initial codes were reviewed, leading to the formation of initial themes. Throughout this process, we continuously reviewed and revised the raw data, codes, and initial themes to ensure consistency and coherence. Moreover, we conducted a 2-day workshop where authors collaboratively reviewed the preliminary themes and their associated codes. During this workshop, we used latent analyses to gain a deeper understanding of the data within the themes. Specifically, we delved into the factors that influenced physicians' documentation practices in the new EHR system using the sociotechnical system theory [44]. These factors encompassed various aspects, such as implications of learning and training in the use of EHR system, physicians' perspectives and previous experiences with clinical documentation, familiarity with the new EHR system, and the influence of technology (new EHR system) and its user-friendliness on physicians' documentation practices (Figure 2). The research team involved in the data analysis possessed diverse backgrounds that enriched the interpretation of the data. The researchers included experts in health informatics, medicine, qualitative research, and implementation science. This multidisciplinary perspective facilitated a comprehensive understanding of the complexities inherent in the data collected. Using the workshop notes, the first author (OG) further refined the themes. This phase served as the basis for drafting the initial report. The authors adhered to the SRQR (Standards for Reporting Qualitative Research) checklist [63]. Multimedia Appendix 2 provides the complete SRQR checklist.

Figure 2. Sociotechnical factors impacting physicians' documentation practices. EHR: electronic health record.

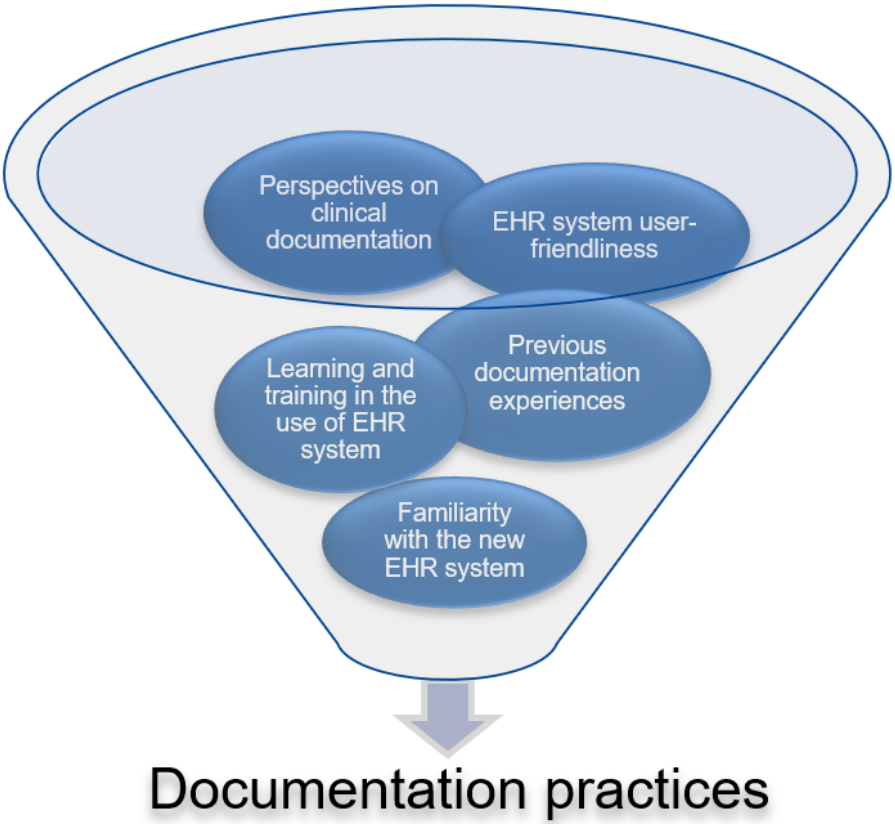

Ethical Considerations

This study was registered and approved by Sikt, Norwegian Agency for Shared Services in Education and Research (project number 540255). Before conducting observations and interviews, all participants were provided with written and verbal information about the study. Written consent was individually obtained from each participant. No personally identifiable data were collected, and in cases where indirectly identifiable information was mentioned during interviews, such details were

XSL•FO
RenderX

omitted from all reports, papers, and presentations. Participants in this study did not receive any form of compensation.

Results

Overview

Through observations and interviews, 3 main themes, with corresponding subthemes related to documentation practices, were identified (Figure 3). The first theme shed light on the initial challenges faced by physicians during the implementation phase, specifically their uncertainty in navigating the new EHR system and its complexity. It also showed the gradual integration of physicians into the new processes of the EHR system, primarily by learning through practical experience and collaborative knowledge exchange with their peers (peer-based learning). The second theme contemplated the influence of the new EHR system on physicians' documentation practices. It delved into the preferred method of documentation and provided insights into the challenges and benefits associated with the implementation of the new documentation processes. In addition, it highlighted the issue of nonstandardized documentation, which in some cases leads to a sense of distrust among physicians toward standardized clinical data. The third theme delved into the perspectives of physicians regarding the adoption of the new terminology system, SNOMED CT, and their opinions on using templates for clinical data documentation.

Figure 3. Themes and subthemes regarding clinical documentation practices emerged from the interviews and observations. EHR: electronic health record; SNOMED CT: Systematized Nomenclature of Medicine–Clinical Terms.

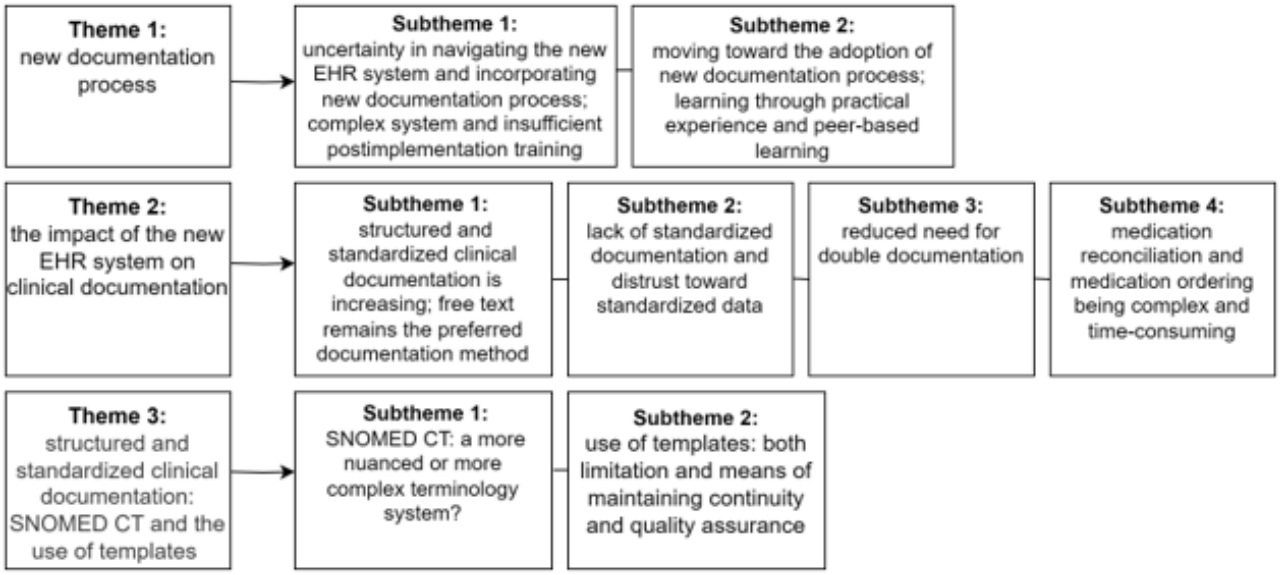

New Documentation Processes

Uncertainty in Navigating the New EHR System and Incorporating New Documentation Processe; Complexity of the System and Insufficient Postimplementation Training

Physicians encountered a learning curve while transitioning to the new EHR system, necessitating adjustments to specific documentation practices. The complexity of certain processes and the additional administrative responsibilities linked to the new system left physicians feeling unsure about how to efficiently navigate the system and, in certain cases, how to accurately document clinical data:

> *... you spend more time searching for the right function that you need to document correctly, and to search for information, and then maybe you are a little bit unsure about some things, you wonder if you have included everything...there is a lot of clicking to move forward; it wasn't like that before; it used to be more like create note, write note, done, so there are quite a lot of steps to move forward [in the new EHR]; the system controls you a bit, sort of.* [Informant 10]

Initially, many physicians were unfamiliar with the specific functionalities of the new EHR system and attempted to use and document clinical data in a similar manner as they did in the previous system. Consequently, certain informants encountered difficulties when using the new system due to its complexity and lack of familiarity with effective navigation:

> *I think that what we need now is to go through each workflow and make sure that everyone does it in the same way so that we can coordinate more, also that you can improve the shortcuts that exist and the ways of working efficiently so that everyone learns it because there are some who use the system very cumbersomely.* [Informant 9]

Furthermore, informants highlighted that the training they received before the implementation of the new EHR system was not sufficient. This was mainly due to substantial modifications made to the EHR system since the training and because the training did not address their specific role. In addition, some informants mentioned that differentiating between essential and less significant aspects addressed in the training became difficult if they had not yet used the system. Most informants also believed that there was insufficient personalized training provided after the implementation, and

XSL•FO
RenderX

they felt that more specific and customized training was necessary to effectively use the system's advanced features, particularly once physicians had acquired a basic understanding of the system:

> *We received a lot of tips and tricks in the beginning, but then it was more about doing, surviving, staying afloat, and stabilizing ourselves a bit...having new training now would have been a much bigger help at this time than it was before...* [Informant 1]

Moving Toward the Adoption of New Documentation Processes and Learning Through Practical Experience and Peer-Based Learning

On the basis of the observation sessions and interview data, it was evident that physicians were gradually adapting to new processes in the new EHR system and making changes to specific documentation practices:

> *When I used to admit patients, I used to write a continuous text with previous illnesses, family, and social, such as whether they lived at home or in a nursing home. I used to just list all previous illnesses, which were often taken from a previous admission journal. I have partly done it that way now as well, but there is more and more emphasis on not doing it and instead entering the information in problem lists of medical and surgical history. So, it's a transition for me...* [Informant 10]

This adaptation and learning occurred both through practice (actual use of the system) and knowledge sharing among colleagues (peer-based learning). The informants found it valuable to learn from experienced colleagues. Occasionally, they also organized small group meetings where these colleagues provided guidance on navigating the system effectively. However, most physicians expressed concerns about the complexity of the newly implemented EHR system, which hindered their learning process and effective use. It was emphasized that simplifying the system and eliminating unnecessary features would greatly enhance its usability.

The Impact of the New EHR System on Clinical Documentation

Structured and Standardized Clinical Documentation Is Increasing, and Free Text Remains the Preferred Documentation Method

Compared to the previous EHR system, a greater proportion of patient information was being entered in a structured and standardized format. In certain cases, the system enforced a hard stop, making it mandatory for users to input the required standardized information. For instance, in the old EHR system, the findings and diagnosis code were described in a free-text format along with codes such as *ICD-10*, whereas in the new EHR system, these aspects were documented in a standardized format using SNOMED CT. Another example was the transition from paper-based to digital documentation for ordering surgical procedures. As a result, numerous tasks were performed electronically, eliminating the reliance on paper, and more patient health information was documented in a structured and standardized format:

> *Well, then [previous EHR] there was nothing standardized, so...no, no, then I manually entered the procedure code and diagnosis code, and the medications were mostly just copied from one note to the next, and then you hoped that what you had copied over was correct.* [Informant 6]

Although structured and standardized documentation has become more prevalent, many physicians still preferred using free text whenever possible. This preference was primarily due to the flexibility and time-saving nature of free-text documentation. It was also easier for physicians to provide a comprehensive overview of a patient's medical condition through free-text notes. Furthermore, some informants highlighted their preference for reading high-quality free-text notes as they provided a better understanding of the patient's condition compared to information presented in templates or diagnosis codes:

> *The only way I manage to get an overview of the medical history of a patient is basically if someone has written something reasonable in a journal note.* [Informant 4]

Moreover, physicians were accustomed to free-text documentation, and transitioning to structured and standardized formats required additional time and training to ensure accuracy.

Structured and standardized documentation, on the other hand, was often perceived as time-consuming, more challenging, and sometimes misaligned with physicians' workflows. In addition, one physician added that he did not see the benefits of adopting a more intricate approach, such as standardized documentation:

> *Yes, no, I need to see the benefit of doing it in a more complicated way, then I might have chosen a different way of documentation than free text. So far, I haven't been able to see the major benefit, so that's why I document in free text...* [Informant 2]

Nonetheless, some informants acknowledged that reading and finding free-text documentation could be challenging in certain situations, particularly when there were numerous notes present. In such cases, important information might become obscured or go unnoticed, whereas standardized information automatically appeared when reviewing the patient journal.

Lack of Standardized Documentation and Distrust Toward Standardized Data

On the basis of our findings, the preference for free-text documentation method was often accompanied by a culture of distrust among physicians toward standardized data available in the EHR system. For instance, the new EHR system integrated several processes related to clinical documentation that were not present in the previous EHR system. These included processes such as the documentation of the "problem list," which is an overview of medical conditions that might impact the health care given to the patient, and "medical history," which lists the conditions that have been resolved and no longer impact patients' health or treatment. These conditions, if documented,

XSL•FO
RenderX

were automatically displayed when reviewing patient journals without the need to access journal notes. However, physicians were uncertain about the accuracy of these standardized data and sometimes could not rely on it, as they were aware that many of their colleagues did not adhere to standardized documentation:

> *It depends on how many people enter standardized data, so I feel that I can't rely on what is stated there. One can enter previous illnesses and medical history, but it is rarely done. So, then, the patient looks healthy on paper, but there are many illnesses, as you can find in the old journal system.* [Informant 4]

Reduced Need for Double Documentation

In certain situations, physicians favored structured and standardized documentation; for instance, in the new EHR system, they transitioned from paper-based to digital documentation when ordering surgical procedures. By adhering to structured and standardized documentation during the preoperative phase, they could save time in the postoperative phase, as some of the information was automatically filled in the postoperative template, eliminating redundant documentation. As a result, physicians recognized the benefits of using structured and standardized documentation for frequently performed routine procedures:

> *It is to some extent a bit easier [postoperative documentation] because all the information I enter beforehand also appears in the postoperative description afterwards, so before, I used to manually write in codes and bleeding amount, for example, but all of this is now automatically included in the postoperative description, so actually, writing the postoperative description afterwards is easier.* [Informant 8]

Medication Reconciliation and Medication Ordering Being Complex and Time-Consuming

With the implementation of the new EHR system, physicians transitioned from manually recording medications on paper-based charts to electronically documenting them in a standardized format. The informants expressed their frustration and difficulties with the current medication reconciliation process, ordering and administering medicines, and inaccuracies in the medication list. This frustration was particularly evident during patient admission and discharge, as there were numerous medications to reconcile. However, while some informants acknowledged the challenges of entering and managing medication reconciliation in the new EHR system compared to the previous system, they also emphasized the potential for mitigating patient risk. Moreover, some informants expressed contentment with the shift from paper charts to electronic charts for medication lists. Conversely, certain individuals contended that this process was demanding and could potentially increase patient risk:

> *I find the medication lists to be a bit tangled and unclear, especially when you have to discharge the patient and know what they are actually on, I always feel like there is something that gets overlooked, some medications that are missing or others that have been included but shouldn't have been, so, the medication lists are challenging.* [Informant 1]

Furthermore, it was observed that, in some cases, physicians who encountered difficulties with medication ordering or the medication reconciliation process sought assistance from colleagues who were more proficient in using the system.

Structured and Standardized Clinical Documentation: SNOMED CT and the Use of Templates

SNOMED CT: More Nuanced or More Complex Terminology System?

The transition from *ICD-10* to SNOMED CT for documenting patient health conditions was perceived by some informants as challenging and by others as a more nuanced diagnostic system. Challenges usually arose due to difficulties associated with finding the correct diagnosis, as physicians were presented with multiple alternatives when starting to write the name of the diagnosis. In addition, sometimes physicians could not find the diagnosis they needed. When diagnosing with SNOMED CT, physicians could specify the location of a medical condition using the SNOMED CT code:

> *...in ICD, there is no difference between right and left, or bilaterally, whereas in SNOMED it is easier to...there, the diagnosis code can already indicate differences between right and left, and bilaterally, so it is potentially that I write right and left in the diagnosis code, which I did not do before...* [Informant 7]

However, in some cases, it was not possible to do this for certain conditions. For instance, during one of the observations, it was noted that when the physician was documenting the diagnosis of deep subfascial lipoma, the physician could not specify the location of the lipoma in SNOMED CT. As a solution, the physician noted the location in the daily notes. The physician pointed out that this information could be essential, especially for surgeons. In addition, the physician mentioned that they could specify the location for a "general" lipoma but not for a subfascial lipoma.

Nonetheless, some informants perceived SNOMED CT as a more nuanced terminology system. Furthermore, SNOMED CT eliminated the need for physicians to commit to memorizing the precise code linked to a specific diagnosis. Instead, they could search for a broader term related to the diagnosis they intended to document. Subsequently, they were presented with multiple alternative options, allowing them to choose the most fitting one.

It is important to note that, in numerous cases, most physicians relied on *ICD-10* codes to identify the corresponding SNOMED CT diagnosis.

Use of Templates: Both Limitation and Means of Maintaining Continuity and Quality Assurance

Most informants emphasized the benefits of using templates to document routine procedures, such as the most frequently performed surgeries. They believed that this approach saved

XSL•FO
RenderX

time, and, for some informants, it helped them remember all the necessary parameters and documentation requirements:

> *It goes faster because there is already a lot of text that is already there, and then you just need to fill in the right and left, how long are the screws, what type of plate was used, what kind of things...and the rehabilitation afterward, is it six weeks, or eight weeks...* [Informant 7]

In addition, it could serve as a checklist for each step involved in a procedure. Many informants had created their own templates, with some developing up to 20 templates for frequently executed operations. However, some of the informants also mentioned that it took time to create these templates properly, so they relied on templates designed by their colleagues. It was also noted that, despite being in a template, these data were not searchable because they were written in free-text format.

Conversely, the informants highlighted that the use of structured and standardized templates for nonstandardized procedures was viewed as both time-consuming and ineffective. It was mentioned that by using a standardized template, physicians might end up documenting unnecessary information, thereby complicating the clinical workflow, particularly during hectic days.

Discussion

Principal Findings

This study provides insights into physicians' documentation practices after they transitioned from an older EHR system to a cross-institutional EHR system, which enables standardized and structured documentation. This transition necessitated physicians to change their documentation habits to some extent, resulting in more clinical data being recorded in a structured and standardized format. However, the degree of acceptance toward new documentation processes is influenced by a range of social and technological factors.

First, according to the study participants, training in system use and ongoing technical support were of utmost importance, particularly after implementing the EHR:

> *We received a lot of tips and tricks in the beginning, but then it was more about doing, surviving, staying afloat, and stabilizing ourselves a bit...having new training now would have been a much bigger help at this time than it was before...* [Informant 1]

The importance of training has been well-documented in previous research [64-67]. However, most of the attention and resources allocated for training tend to be centered on the pre–go-live phase and the initial implementation stage of the EHR system, with limited support available for an extended period after the go-live phase and for system updates [65,68,69]. In addition, it has been noted that users may feel overwhelmed during initial training and prioritize familiarizing themselves with the fundamental aspects of the system rather than striving to use it efficiently [68]. This concern was also emphasized by the informants in our study. Moreover, peer-based learning, where physicians are taught by their expert peers in the use of the EHR system, was deemed beneficial. This suggests that acquiring knowledge from fellow experts could prove more advantageous, as these peers possess a deep understanding of the physicians' workflows and the specific challenges they face when using the system. Peer-based learning has also been observed when physicians use the templates created by their colleagues. Hence, it can be advantageous to create platforms where physicians can share tacit knowledge, enabling them to mutually benefit from their expertise in effectively using the system and implementing best practices for documentation. Providing additional training in documentation practices should be considered, particularly during the transition to the new EHR system, which necessitates changes in physicians' documentation practices. This training becomes even more significant considering the alarming statistics from a study conducted in Norway. The study, which encompassed 14 hospitals, revealed that in 2003, an incorrect main diagnosis code was assigned in 37.8% of hospital stays, and in 2008, in 36.2% of hospital stays [47]. According to physicians working in these hospitals, the main factors contributing to this issue were challenging code systems (59%) and inadequate training (50%) [70]. The significance of education and ongoing training in good documentation practices as well as its contribution to achieving a learning health system, has been emphasized in numerous research studies [71-75].

Second, in conjunction with the aforementioned first factor, physicians found the transition to the new EHR system to be challenging. The challenge stems from the complex navigation needed to find and execute specific tasks, the overwhelming amount of information in the EHR system, and the inefficient user interface. Furthermore, physicians' past recording habits in older EHR systems can impact their documentation practices in the new EHR system. As pointed out by Orlikowski [38], when people use technology, "people also draw on their skills, power, knowledge, assumptions, and expectations about the technology and its use, influenced typically by training, communication, and previous experiences" [76]. Given that physicians have established workflows in the previous system, adjusting to new processes in the new EHR system might hinder their adaptability and work efficiency. This finding aligns with previous research conducted by Joukes et al [77], who identified that the preimplementation circumstances played a significant role in shaping the perceived advantages of implementing a new EHR system with structured and standardized clinical data recording. Specifically, the authors emphasized that the perceived impact of the implementation varied depending on whether users were transitioning from older EHR systems or paper-based records. In a paper-based record center, the implementation yielded positive perceived effects on EHR use, data quality, and data reuse. However, in a center that had previously used an EHR system, the perceived effects were predominantly negative or neutral. The authors further highlighted that individuals with previous exposure to EHR systems tended to draw comparisons between the new system and their old one, thereby potentially recognizing that certain tasks were accomplished more efficiently in the former system [77].

XSL•FO
RenderX

While the adoption of the new EHR system has led to an increase in structured and standardized data recording, it is often not a decision made by physicians but rather an imposition by the new system and health organization. In certain cases, the structured and standardized documentation does not align with physicians' traditional workflow; therefore, it takes more time compared to free-text documentation. Moreover, physicians occasionally experience uncertainty and difficulties regarding the correct way to document structured and standardized data or question their necessity. This may potentially lead to inconsistencies in data recording. Data that could be standardized may instead be documented in a free-text format, and when faced with time constraints, physicians may use work-arounds, such as quickly clicking through to proceed further. This practice might also explain physicians' distrust toward some of the standardized clinical data available in the EHR system, such as problem lists and medical history. One advantage of the problem list and medical history is that these remain visible to physicians whenever they access patient records, without the need to open journal notes. While this provides better information at the point of care, the problem lists and medical history lose their advantage if physicians fail to document or update it or if they lack trust in the accuracy and reliability of this information. Moreover, the prevalent reliance of physicians on the *ICD-10* codes to identify the corresponding SNOMED CT codes may undermine the advantages of using a more comprehensive terminology system, SNOMED CT. This reliance can lead to loss of clinical detail and mapping inaccuracies. This practice indicates a discrepancy between the physicians' mental models (understanding and reasoning about systems) and the designer's conceptual model (how the system is intended to be used) [78,79]. Mental models, derived from a range of factors, including experiences, knowledge, and perception, have been widely applied in the field of human-computer interaction research [79,80]. Gaining an understanding of physicians' mental models can be important for effectively optimizing the system and tailoring training in documentation practices in the new EHR system. The greater the disparity between the designer's conceptual model and the user's mental model, the more challenging it becomes for individuals to effectively use the system [78,79]. Furthermore, this emphasizes the need to distinguish between technologies as artifacts and technologies in practice (how technology is actually used rather than solely focusing on the technology itself), as it is only through effective use that technology can impact productivity [38]. Consequently, when enhancing technology, it becomes vital to consider sociotechnical factors, which, in turn, necessitates the technology to possess sufficient flexibility in order to be customized according to specific local needs and evolving requirements [33,37]. Moreover, Rosenbloom et al [22] have underscored the significance of offering flexibility in the documentation processes, as they are highly valued and essential for health care professionals. By using free-text documentation, health care professionals can document intricate and nuanced details about a patient, which can also prove to be more valuable for subsequent health care professionals taking over the patient's care [22]. Some of our study participants have also emphasized that reading high-quality free-text journal notes can provide them with more crucial information compared to solely reviewing standardized or structured data:

> *The only way I manage to get an overview of the medical history of a patient is basically if someone has written something reasonable in a journal note.* [Informant 4]

Having the right balance between standardization and flexibility in documentation processes is crucial, particularly in situations where unforeseen clinical findings or unexpected circumstances arise [22].

According to the study participants, most physicians have a positive attitude toward using templates for documenting routine procedures:

> *It goes faster [using templates] because there is already a lot of text that is already there, and then you just need to fill in the right and left, how long are the screws, what type of plate was used, what kind of things...and the rehabilitation afterward, is it six weeks, or eight weeks...* [Informant 7]

While certain physicians have created their own templates for frequently conducted surgeries, others depend on templates designed by their peers. However, the information contained within these templates is recorded in a free-text format. As a result, these data are not conducive to analysis using statistical or machine learning techniques. Moreover, when physicians develop their own templates for commonly performed surgeries, it can result in inconsistencies in documentation and potential omissions of critical data that need to be documented. On the other hand, this behavior might be elucidated through the analysis of physician-patient dialogues in Norwegian general practice by Nessa [81,82]. His research emphasized that these conversations were a fundamental aspect of a physician's work, encompassing rich narratives and certain elements such as context and emotional tone [81,82]. Such complexities may be challenging to capture within a standardized template and encoded data. Thus, adhering strictly to encoded data and standardized templates implied by the organization may impose a semiotic burden on physicians, requiring them to exert considerable mental effort to fit complex narratives into standardized templates. The attitudes of health care professionals toward structured and standardized documentation may also depend on their perception of the benefits it provides [21,83]. Our study participants also highlighted this aspect (eg, saving time in postoperative documentation), suggesting that physicians may view the implementation of standardized or semistandardized templates for commonly performed procedures, developed in collaboration with physicians and following established guidelines, as a positive step. Moreover, this approach could potentially minimize variations in clinical documentation. This approach might also enable more data to be consistently recorded in a standardized format and enhance guideline compliance. Several studies have shown that the use of templates improves the accuracy of documentation [84-86]. For instance, Thomson et al [85] demonstrated the advantages of using a standardized operative note template adhering to established guidelines. This implementation led to enhanced documentation quality and a decrease in the missing data.

XSL•FO
RenderX

Limitations and Future Research

One limitation of our study is that it involved a limited number of physicians, and data were collected from only 1 hospital. Therefore, the findings of this study may only reflect physicians' perspectives working within this specific context. Nonetheless, we have made an effort to gather information on documentation practices that can be applied more broadly. Specifically, our focus was on understanding documentation practices when physicians transition from an older EHR system to a new EHR system that facilitates structured and standardized documentation. In addition, we aimed to identify the factors influencing physicians' choice of documentation methods, such as free text and standardized documentation, regardless of the EHR system they use. Thus, our findings may have relevance beyond the specific organization and the EHR system used in this study, providing insights that can be applied to a wider context.

Moreover, although it is evident that the implementation of the new EHR system has increased the amount of clinical data recorded in a standardized format, our study did not objectively evaluate either the quality of these data (eg, accuracy and completeness) or their effectiveness for secondary uses. Consequently, future research should aim to objectively assess the quality of the data. Furthermore, additional investigations could delve into the perceived data quality from the perspectives of individuals using the data from the new EHR system for secondary purposes, such as quality assurance. This would help determine whether the transition to a more standardized EHR system has positively impacted the use of these data for secondary purposes.

Conclusions

The findings of this study identified that physicians' acceptance of new documentation processes is influenced by a range of social and technological factors. These factors include physicians' past experiences with documentation, perceived benefits, familiarity with the EHR system, and time constraints, all of which can impact their choice of documentation methods. Furthermore, it is important to consider the user-friendliness of the EHR system and how well the documentation processes align with physicians' workflows. To facilitate the transition to a new EHR system, health care organizations should provide tailored training programs on documentation practices. This is particularly crucial after the implementation of the new system. In addition, it would be advantageous to establish platforms that allow physicians to exchange tacit knowledge and expertise on effectively using the system. Moreover, physicians highly value flexibility in clinical documentation. Therefore, health care organizations should consider implementing standardized documentation processes that also align with physicians' preferences and allow them some flexibility in the documentation process. For instance, introducing common standardized or semistandardized templates for routine procedures, developed in collaboration with physicians, could potentially minimize disparities in clinical documentation, ensure the consistent recording of data, and improve adherence to guidelines.

Acknowledgments

The authors would like to extend their gratitude to the physicians for their time commitment in sharing their experiences and providing insightful perspectives on the use of the new electronic health record system and documentation practices. This study was funded by the Liaison Committee between the Central Norway Regional Health Authority and the Norwegian University of Science and Technology.

Data Availability

The datasets generated or analyzed during this study are not publicly available to protect participant privacy but are available from the corresponding author on reasonable request.

Authors' Contributions

OG, AF, LM, and RP conceptualized and designed the study. OG conducted data collection and performed the initial data analyses. AF, LM, and RP contributed to the analyses and interpretation of the data. OG wrote the initial draft of the manuscript. All authors performed critical revisions of the paper and contributed to the final version of the manuscript. All authors have reviewed and approved the submission of the paper in its current form.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Semistructured interview guide.
[DOCX File , 19 KB - formative_v9i1e63902_app1.docx]

Multimedia Appendix 2
SRQR (Standards for Reporting Qualitative Research) checklist.

XSL•FO
RenderX

[DOCX File , 27 KB - formative_v9i1e63902_app2.docx]

References

1. Gatiti P, Ndirangu E, Mwangi J, Mwanzu A, Ramadhani T. Enhancing healthcare quality in hospitals through electronic health records: a systematic review. J Health Inform Dev Ctries 2021 Nov 02;15(2):1-25.
2. Kruse CS, Mileski M, Vijaykumar AG, Viswanathan SV, Suskandla U, Chidambaram Y. Impact of electronic health records on long-term care facilities: systematic review. JMIR Med Inform 2017 Sep 29;5(3):e35 [FREE Full text] [doi: 10.2196/medinform.7958] [Medline: 28963091]
3. Janssen A, Kay J, Talic S, Pusic M, Birnbaum RJ, Cavalcanti R, et al. Electronic health records that support health professional reflective practice: a missed opportunity in digital health. J Healthc Inform Res 2023 Jan 20;6(4):375-384 [FREE Full text] [doi: 10.1007/s41666-022-00123-0] [Medline: 36744083]
4. Capurro D, Yetisgen M, van Eaton E, Black R, Tarczy-Hornoch P. Availability of structured and unstructured clinical data for comparative effectiveness research and quality improvement: a multisite assessment. EGEMS (Wash DC) 2014 Jul 11;2(1):1079 [FREE Full text] [doi: 10.13063/2327-9214.1079] [Medline: 25848594]
5. Olsen L, Aisner D, McGinnis J, Institute of Medicine (US) Roundtable on Evidence-Based Medicine. The Learning Healthcare System: Workshop Summary. Washington, DC: National Academies Press; 2007.
6. Joukes E, Abu-Hanna A, Cornet R, de Keizer NF. Time spent on dedicated patient care and documentation tasks before and after the introduction of a structured and standardized electronic health record. Appl Clin Inform 2018 Jan;9(1):46-53 [FREE Full text] [doi: 10.1055/s-0037-1615747] [Medline: 29342479]
7. Ebbers T, Takes RP, Smeele LE, Kool RB, van den Broek GB, Dirven R. The implementation of a multidisciplinary, electronic health record embedded care pathway to improve structured data recording and decrease electronic health record burden. Int J Med Inform 2024 Apr;184:105344 [FREE Full text] [doi: 10.1016/j.ijmedinf.2024.105344] [Medline: 38310755]
8. Bush RA, Kuelbs C, Ryu J, Jiang W, Chiang G. Structured data entry in the electronic medical record: perspectives of pediatric specialty physicians and surgeons. J Med Syst 2017 May;41(5):75 [FREE Full text] [doi: 10.1007/s10916-017-0716-5] [Medline: 28324321]
9. Gilbert JA. Physician data entry: providing options is essential. Health Data Manag 1998 Sep;6(9):84-6, 88, 90-2. [Medline: 10185232]
10. Lov om behandling av helseopplysninger ved ytelse av helsehjelp (pasientjournalloven). Lovdata. 2015. URL: https://lovdata.no/dokument/NL/lov/2014-06-20-42 [accessed 2024-06-17]
11. Lov om helsepersonell m.v. (helsepersonelloven). Lovdata. URL: https://lovdata.no/dokument/NL/lov/1999-07-02-64/KAPITTEL_8#KAPITTEL_8 [accessed 2024-06-17]
12. Lov om behandling av personopplysninger (personopplysningsloven). Lovdata. 2018. URL: https://lovdata.no/dokument/NL/lov/2018-06-15-38 [accessed 2024-06-17]
13. Post AR, Burningham Z, Halwani AS. Electronic health record data in cancer learning health systems: challenges and opportunities. JCO Clin Cancer Inform 2022 Mar;6:e2100158 [FREE Full text] [doi: 10.1200/CCI.21.00158] [Medline: 35353547]
14. Kong HJ. Managing unstructured big data in healthcare system. Healthc Inform Res 2019 Jan;25(1):1-2 [FREE Full text] [doi: 10.4258/hir.2019.25.1.1] [Medline: 30788175]
15. Joukes E, Cornet R, de Bruijne MC, de Keizer NF, Abu-Hanna A. Development and validation of a model for the adoption of structured and standardised data recording among healthcare professionals. BMC Med Inform Decis Mak 2018 Jun 28;18(1):54 [FREE Full text] [doi: 10.1186/s12911-018-0640-8] [Medline: 29954388]
16. Sun W, Cai Z, Liu F, Fang S, Wang G. A survey of data mining technology on electronic medical records. In: Proceedings of the IEEE 19th International Conference on e-Health Networking, Applications and Services. 2017 Presented at: HealthCom 2017; October 12-15, 2017; Dalian, China. [doi: 10.1109/HealthCom.2017.8210774]
17. Tayefi M, Ngo P, Chomutare T, Dalianis H, Salvi E, Budrionis A, et al. Challenges and opportunities beyond structured data in analysis of electronic health records. WIREs Comput Stat 2021 Feb 14;13(6):e1549. [doi: 10.1002/wics.1549]
18. Seid M, Hartley DM, Margolis PA. A science of collaborative learning health systems. Learn Health Syst 2021 Jun 01;5(3):e10278 [FREE Full text] [doi: 10.1002/lrh2.10278] [Medline: 34277944]
19. Zafar MA, Beck AF, Chirumbolo C, Wilson KD, Haralson K, Rich K, et al. A regional learning health system of congregate care facilities for COVID-19 response. Learn Health Syst 2024 Jan 24;8(3):e10407 [FREE Full text] [doi: 10.1002/lrh2.10407] [Medline: 39036529]
20. Ebbers T, Kool RB, Smeele LE, Dirven R, den Besten CA, Karssemakers LH, et al. The impact of structured and standardized documentation on documentation quality; a multicenter, retrospective study. J Med Syst 2022 May 27;46(7):46 [FREE Full text] [doi: 10.1007/s10916-022-01837-9] [Medline: 35618978]
21. Vuokko R, Mäkelä-Bengs P, Hyppönen H, Lindqvist M, Doupi P. Impacts of structuring the electronic health record: results of a systematic literature review from the perspective of secondary use of patient data. Int J Med Inform 2017 Jan;97:293-303. [doi: 10.1016/j.ijmedinf.2016.10.004] [Medline: 27919387]

XSL•FO
RenderX

22. Rosenbloom ST, Denny JC, Xu H, Lorenzi N, Stead WW, Johnson KB. Data from clinical notes: a perspective on the tension between structure and flexible documentation. J Am Med Inform Assoc 2011;18(2):181-186 [FREE Full text] [doi: 10.1136/jamia.2010.007237] [Medline: 21233086]
23. Levy DR, Withall JB, Mishuris RG, Tiase V, Diamond C, Douthit B, et al. Defining documentation burden (DocBurden) and excessive DocBurden for all health professionals: a scoping review. Appl Clin Inform 2024 Oct;15(5):898-913 [FREE Full text] [doi: 10.1055/a-2385-1654] [Medline: 39137903]
24. Sinsky CA, Bavafa H, Roberts RG, Beasley JW. Standardization vs customization: finding the right balance. Ann Fam Med 2021;19(2):171-177 [FREE Full text] [doi: 10.1370/afm.2654] [Medline: 33685879]
25. Kriznik NM, Lamé G, Dixon-Woods M. Challenges in making standardisation work in healthcare: lessons from a qualitative interview study of a line-labelling policy in a UK region. BMJ Open 2019 Nov 27;9(11):e031771 [FREE Full text] [doi: 10.1136/bmjopen-2019-031771] [Medline: 31780591]
26. Wears RL. Standardisation and its discontents. Cogn Technol Work 2015 Feb;17(1):89-94 [FREE Full text] [doi: 10.1007/s10111-014-0299-6] [Medline: 25667566]
27. Brunner J, Cannedy S, McCoy M, Hamilton AB, Shelton J. Software is policy: electronic health record governance and the implications of clinical standardization. J Gen Intern Med 2023 Oct;38(Suppl 4):949-955 [FREE Full text] [doi: 10.1007/s11606-023-08280-7] [Medline: 37798574]
28. Bansler JP. Challenges in user-driven optimization of EHR: a case study of a large epic implementation in Denmark. Int J Med Inform 2021 Apr;148:104394. [doi: 10.1016/j.ijmedinf.2021.104394] [Medline: 33485217]
29. Koh J, Ahmed M. Improving clinical documentation: introduction of electronic health records in paediatrics. BMJ Open Qual 2021 Feb;10(1):e000918 [FREE Full text] [doi: 10.1136/bmjoq-2020-000918] [Medline: 33589503]
30. Electronic health records and the importance of how to document. Pan American Health Organization. 2021. URL: https://iris.paho.org/handle/10665.2/55410 [accessed 2024-05-10]
31. Ulriksen GH, Pedersen R, Ellingsen G. Infrastructuring in healthcare through the OpenEHR architecture. Comput Supp Coop Work 2017 Feb 13;26:33-69. [doi: 10.1007/s10606-017-9269-x]
32. Meum T, Ellingsen G, Monteiro E, Wangensteen G, Igesund H. The interplay between global standards and local practice in nursing. Int J Med Inform 2013 Dec;82(12):e364-e374. [doi: 10.1016/j.ijmedinf.2013.02.005] [Medline: 23688846]
33. Eason K. Local sociotechnical system development in the NHS national programme for information technology. J Inf Technol 2007 Sep;22(3):257-264. [doi: 10.1057/palgrave.jit.2000101]
34. Pedersen R, Ellingsen G, Monteiro E. Standardized nursing work: works in practice but not in theory? Stud Health Technol Inform 2010;157:91-96. [Medline: 20543373]
35. Pedersen R. Standardizing work in healthcare through architecture, routines and technologies. In: Proceedings of the 10th International Conference on the Design of Cooperative Systems. 2012 Presented at: COOP 2012; May 30-June 1, 2012; Marseille, France. [doi: 10.1007/978-1-4471-4093-1_2]
36. Schaper LK, Pervan GP. ICT and OTs: a model of information and communication technology acceptance and utilisation by occupational therapists. Int J Med Inform 2007 Jun;76 Suppl 1:S212-S221. [doi: 10.1016/j.ijmedinf.2006.05.028] [Medline: 16828335]
37. Cherns A. The principles of sociotechnical design. Hum Relat 1976 Aug;29(8):783-792 [FREE Full text] [doi: 10.1177/001872677602900806]
38. Orlikowski WJ. Using technology and constituting structures: a practice lens for studying technology in organizations. Organ Sci 2000 Aug;11(4):404-428. [doi: 10.1287/orsc.11.4.404.14600]
39. Pedersen R. Standardization strategies in healthcare practices? In: Proceedings of the Fifteenth International Symposium for Health Information Management Research. 2011 Presented at: ISHIMR 2011; September 8-9, 2011; Zurich, Switzerland.
40. Suchman LA. Plans and Situated Actions: The Problem of Human-Machine Communication. Cambridge, UK: Cambridge University Press; 1987.
41. Bostrom RP, Heinen JS. MIS problems and failures: a socio-technical perspective. Part I: the causes. MIS Q 1977 Sep;1(3):17-32. [doi: 10.2307/248710]
42. Orlikowski WJ. The duality of technology: rethinking the concept of technology in organizations. Organ Sci 1992 Aug 01;3(3):398-427. [doi: 10.1287/orsc.3.3.398]
43. Sittig DF, Singh H. A sociotechnical approach to electronic health record related safety. In: Key Advances in Clinical Informatics: Transforming Health Care Through Health Information Technology. Cambridge, MA: Academic Press; 2017.
44. Abbas R, Michael K. In: Papagiannidis S, editor. Socio-Technical Theory: A Review. Newcastle upon Tyne, UK: TheoryHub Book; 2023.
45. Sittig DF, Singh H. A new socio-technical model for studying health information technology in complex adaptive healthcare systems. In: Kaufman DR, Kannampallil TG, Patel VL, editors. Cognitive Informatics for Biomedicine: Human Computer Interaction in Healthcare. Cham, Switzerland: Springer International Publishing; 2015.
46. Laukvik LB, Lyngstad M, Rotegård AK, Fossum M. Utilizing nursing standards in electronic health records: a descriptive qualitative study. Int J Med Inform 2024 Apr;184:105350 [FREE Full text] [doi: 10.1016/j.ijmedinf.2024.105350] [Medline: 38306850]

XSL•FO
RenderX

47. Utredning av «en innbygger – én journal». Helsedirektoratet. 2014 Sep. URL: https://www.regjeringen.no/contentassets/355890dd2872413b838066702dcdad88/ikt_utfordringsbilde_helse_omsorgssektoren.pdf [accessed 2024-04-12]
48. Meld. St. 9 (2012–2013). Regjeringen.no. 2012. URL: https://www.regjeringen.no/no/dokumenter/meld-st-9-20122013/id708609/ [accessed 2024-04-17]
49. Bakgrunn for Helseplattformen. Helseplattformen. URL: https://www.helseplattformen.no/bakgrunn-for-helseplattformen [accessed 2024-04-17]
50. Kartlegging av elektroniske pasientjournalsystemer og deres støtte for nasjonale e-helseløsninger. Helsedirektoratet. 2023. URL: https://tinyurl.com/y7mupvvv [accessed 2025-03-15]
51. Årlig melding 2017 for Helse Midt-Norge RHF til Helse- og omsorgsdepartementet. Regjeringen.no. 2017. URL: https://www.regjeringen.no/contentassets/18f93a5faa81417b8ac28f74a02cc0ee/arlig-melding-2017-fra-helse-midt-norge-rhf.pdf [accessed 2024-03-10]
52. Stedenfeldt M, Wekre LJ, Tørhaug T, Gotaas E, Toch-Marquardt M. Fra fritekst til struktur og felles helsespråk. Michael 2023;20(Supplement 31):106-114 [FREE Full text]
53. Abeysinghe R, Zheng F, Bernstam EV, Shi J, Bodenreider O, Cui L. A deep learning approach to identify missing is-a relations in SNOMED CT. J Am Med Inform Assoc 2023 Feb 16;30(3):475-484 [FREE Full text] [doi: 10.1093/jamia/ocac248] [Medline: 36539234]
54. Fung KW, Xu J. An exploration of the properties of the CORE problem list subset and how it facilitates the implementation of SNOMED CT. J Am Med Inform Assoc 2015 May;22(3):649-658 [FREE Full text] [doi: 10.1093/jamia/ocu022] [Medline: 25725003]
55. Ash JS, Berg M, Coiera E. Some unintended consequences of information technology in health care: the nature of patient care information system-related errors. J Am Med Inform Assoc 2004;11(2):104-112 [FREE Full text] [doi: 10.1197/jamia.M1471] [Medline: 14633936]
56. Mamykina L, Vawdrey DK, Stetson PD, Zheng K, Hripcsak G. Clinical documentation: composition or synthesis? J Am Med Inform Assoc 2012 Nov 01;19(6):1025-1031 [FREE Full text] [doi: 10.1136/amiajnl-2012-000901] [Medline: 22813762]
57. Pizziferri L, Kittler AF, Volk LA, Honour MM, Gupta S, Wang S, et al. Primary care physician time utilization before and after implementation of an electronic health record: a time-motion study. J Biomed Inform 2005 Jun;38(3):176-188 [FREE Full text] [doi: 10.1016/j.jbi.2004.11.009] [Medline: 15896691]
58. Rowlands S, Tariq A, Coverdale S, Walker S, Wood M. A qualitative investigation into clinical documentation: why do clinicians document the way they do? Health Inf Manag 2022 Sep;51(3):126-134. [doi: 10.1177/1833358320929776] [Medline: 32643428]
59. Ho YX, Gadd CS, Kohorst KL, Rosenbloom ST. A qualitative analysis evaluating the purposes and practices of clinical documentation. Appl Clin Inform 2014;5(1):153-168 [FREE Full text] [doi: 10.4338/ACI-2013-10-RA-0081] [Medline: 24734130]
60. Pankhurst T, Evison F, Atia J, Gallier S, Coleman J, Ball S, et al. Introduction of systematized nomenclature of medicine-clinical terms coding into an electronic health record and evaluation of its impact: qualitative and quantitative study. JMIR Med Inform 2021 Nov 23;9(11):e29532 [FREE Full text] [doi: 10.2196/29532] [Medline: 34817387]
61. AlYahmady HH, Al Abri SS. Using Nvivo for data analysis in qualitative research. Int Interdiscip J Educ 2013 Feb;2(2):181-186. [doi: 10.12816/0002914]
62. Braun V, Clarke V. One size fits all? What counts as quality practice in (reflexive) thematic analysis? Qual Res Psychol 2020 Aug 12;18(3):328-352. [doi: 10.1080/14780887.2020.1769238]
63. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014;89(9):1245-1251. [doi: 10.1097/acm.0000000000000388]
64. Ratwani R, Fairbanks T, Savage E, Adams K, Wittie M, Boone E, et al. Mind the gap. A systematic review to identify usability and safety challenges and practices during electronic health record implementation. Appl Clin Inform 2016 Nov 16;7(4):1069-1087 [FREE Full text] [doi: 10.4338/ACI-2016-06-R-0105] [Medline: 27847961]
65. Blavin F, Ramos C, Shah A, Devers K. Lessons from the literature on electronic health record implementation: a study funded by the office of the national coordinator for health information technology of the U.S. Department of Health and Human Services. The Urban Institute. 2013 Aug 1. URL: https://www.healthit.gov/sites/default/files/hit_lessons_learned_lit_review_final_08-01-2013.pdf [accessed 2024-05-13]
66. Samadbeik M, Fatehi F, Braunstein M, Barry B, Saremian M, Kalhor F, et al. Education and training on electronic medical records (EMRs) for health care professionals and students: a scoping review. Int J Med Inform 2020 Oct;142:104238. [doi: 10.1016/j.ijmedinf.2020.104238] [Medline: 32828034]
67. Heponiemi T, Gluschkoff K, Vehko T, Kaihlanen AM, Saranto K, Nissinen S, et al. Electronic health record implementations and insufficient training endanger nurses' well-being: cross-sectional survey study. J Med Internet Res 2021 Dec 23;23(12):e27096 [FREE Full text] [doi: 10.2196/27096] [Medline: 34941546]
68. Bredfeldt CE, Awad EB, Joseph K, Snyder MH. Training providers: beyond the basics of electronic health records. BMC Health Serv Res 2013 Dec 02;13:503 [FREE Full text] [doi: 10.1186/1472-6963-13-503] [Medline: 24295150]

XSL•FO
RenderX

69. Sachs J, Yonek JC, Mullins CD, Hamilton A, Rachman F, Kaleba E, et al. Impact of health IT on lab order use in community health centers. Agency for Healthcare Research and Quality. 2011 Nov. URL: https://digital.ahrq.gov/file/18903/download?token=9XtjSeyh [accessed 2024-05-13]
70. Riksrevisjonens undersøkelse av kodekvaliteten ved helseforetakene. Riksrevisjonen. 2006. URL: https://www.stortinget.no/globalassets/pdf/dokumentserien/2005-2006/dok_3_7_2005_2006.pdf [accessed 2024-06-11]
71. Forrest CB, Crandall WV, Bailey LC, Zhang P, Joffe MM, Colletti RB, et al. Effectiveness of anti-TNFα for Crohn disease: research in a pediatric learning health system. Pediatrics 2014 Jul;134(1):37-44 [FREE Full text] [doi: 10.1542/peds.2013-4103] [Medline: 24935993]
72. Rattray NA, Damush TM, Miech EJ, Homoya B, Myers LJ, Penney LS, et al. Empowering implementation teams with a learning health system approach: leveraging data to improve quality of care for transient ischemic attack. J Gen Intern Med 2020 Nov 01;35(Suppl 2):823-831 [FREE Full text] [doi: 10.1007/s11606-020-06160-y] [Medline: 32875510]
73. Mungmode A, Noor N, Weinstock RS, Izquierdo R, Indyk JA, DeSalvo DJ, et al. Making diabetes electronic medical record data actionable: promoting benchmarking and population health improvement using the T1D exchange quality improvement portal. Clin Diabetes 2022;41(1):45-55 [FREE Full text] [doi: 10.2337/cd22-0072] [Medline: 36714251]
74. Alonso GT, Corathers S, Shah A, Clements M, Kamboj M, Sonabend R, et al. Establishment of the T1D exchange quality improvement collaborative (T1DX-QI). Clin Diabetes 2020 Apr;38(2):141-151 [FREE Full text] [doi: 10.2337/cd19-0032] [Medline: 32327886]
75. Vandenberg AE, Kegler M, Hastings SN, Hwang U, Wu D, Stevens MB, et al. Sequential implementation of the EQUIPPED geriatric medication safety program as a learning health system. Int J Qual Health Care 2020 Sep 23;32(7):470-476. [doi: 10.1093/intqhc/mzaa077] [Medline: 32671390]
76. Orlikowski WJ, Gash DC. Technological frames: making sense of information technology in organizations. ACM Trans Inf Syst 1994 Apr;12(2):174-207. [doi: 10.1145/196734.196745]
77. Joukes E, de Keizer NF, de Bruijne MC, Abu-Hanna A, Cornet R. Impact of electronic versus paper-based recording before EHR implementation on health care professionals' perceptions of EHR use, data quality, and data reuse. Appl Clin Inform 2019 Mar;10(2):199-209 [FREE Full text] [doi: 10.1055/s-0039-1681054] [Medline: 30895574]
78. Staggers N, Norcio AF. Mental models: concepts for human-computer interaction research. Int J Man Mach Stud 1993 Apr;38(4):587-605. [doi: 10.1006/imms.1993.1028]
79. Kaufman DR, Kannampallil TG, Patel VL. Cognition and human computer interaction in health and biomedicine. In: Patel VL, Kannampallil TG, Kaufman DR, editors. Cognitive Informatics for Biomedicine. Cham, Switzerland: Springer; 2015.
80. Holtrop JS, Scherer LD, Matlock DD, Glasgow RE, Green LA. The importance of mental models in implementation science. Front Public Health 2021 Jul 6;9:680316 [FREE Full text] [doi: 10.3389/fpubh.2021.680316] [Medline: 34295871]
81. Nessa J. From a medical consultation to a written text. 1. Transcribing the doctor-patient dialogue. Scand J Prim Health Care 1995 Jun 12;13(2):83-88 [FREE Full text] [doi: 10.3109/02813439508996741] [Medline: 7569483]
82. Nessa J. From a medical consultation to a written text. 2. Pragmatics and textlinguistics applied to medicine. Scand J Prim Health Care 1995 Jun 12;13(2):89-92 [FREE Full text] [doi: 10.3109/02813439508996742] [Medline: 7569484]
83. Ludwick DA, Doucette J. Adopting electronic medical records in primary care: lessons learned from health information systems implementation experience in seven countries. Int J Med Inform 2009 Jan;78(1):22-31. [doi: 10.1016/j.ijmedinf.2008.06.005] [Medline: 18644745]
84. Mahapatra P, Ieong E. Improving documentation and communication using operative note proformas. BMJ Qual Improv Rep 2016 Feb 08;5(1):u209122.w3712 [FREE Full text] [doi: 10.1136/bmjquality.u209122.w3712] [Medline: 26893892]
85. Thomson DR, Baldwin MJ, Bellini MI, Silva MA. Improving the quality of operative notes for laparoscopic cholecystectomy: assessing the impact of a standardized operation note proforma. Int J Surg 2016 Mar;27:17-20 [FREE Full text] [doi: 10.1016/j.ijsu.2016.01.037] [Medline: 26804350]
86. Kahn D, Stewart E, Duncan M, Lee E, Simon W, Lee C, et al. A prescription for note bloat: an effective progress note template. J Hosp Med 2018 Jun 01;13(6):378-382. [doi: 10.12788/jhm.2898] [Medline: 29350222]

Abbreviations

EHR: electronic health record
ICD-10: International Classification of Diseases, 10th Revision
SNOMED CT: Systematized Nomenclature of Medicine–Clinical Terms
SRQR: Standards for Reporting Qualitative Research

XSL•FO
RenderX

Edited by A Mavragani; submitted 04.07.24; peer-reviewed by D Chartash; comments to author 09.01.25; revised version received 17.01.25; accepted 19.01.25; published 21.03.25.

Please cite as:
Golburean O, Pedersen R, Melby L, Faxvaag A
Exploring Physicians' Dual Perspectives on the Transition From Free Text to Structured and Standardized Documentation Practices: Interview and Participant Observational Study
JMIR Form Res 2025;9:e63902
URL: https://formative.jmir.org/2025/1/e63902
doi:10.2196/63902
PMID:

©Olga Golburean, Rune Pedersen, Line Melby, Arild Faxvaag. Originally published in JMIR Formative Research (https://formative.jmir.org), 21.03.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Co-Design of a Depression Self-Management Tool for Adolescent and Young Adult Cancer Survivors: User-Centered Design Approach

Karly M Murphy[1], PhD; Rachel Glock[1], BA; David Victorson[2], PhD; Madhu Reddy[3], PhD; Sarah A Birken[4], PhD; John M Salsman[5], PhD

[1]Department of Psychology, East Carolina University, Greenville, NC, United States
[2]Department of Medical Social Sciences, Northwestern University Feinberg School of Medicine, Chicago, IL, United States
[3]Department of Informatics, University of California Irvine, Irvine, CA, United States
[4]Department of Implementation Science, Wake Forest University School of Medicine, Winston-Salem, NC, United States
[5]Department of Social Sciences and Health Policy, Wake Forest University School of Medicine, Winston-Salem, NC, United States

Corresponding Author:
Karly M Murphy, PhD
Department of Psychology
East Carolina University
1000 E 5th St
Greenville, NC, 27858
United States
Phone: 1 2527374638
Fax: 1 2523286283
Email: murphykar22@ecu.edu

Abstract

Background: Adolescent and young adult (AYA) cancer survivors are more likely to experience elevated depressive symptoms than older survivors and healthy age-matched peers. Despite the elevated risk of depressive symptoms in AYA cancer survivors and the existence of evidence-based interventions to address depression, it is unclear whether AYA cancer survivors can access support services. Digital tools are a potential solution to overcoming barriers to AYA cancer survivors' unmet needs for psychosocial support, but they have not been tailored to the needs and preferences of this unique population.

Objective: This study engaged AYA cancer survivors and their providers in the concept generation and ideation step of the user-centered design process through online co-design workshops. The goal was to generate concepts and ideas for a digital depression self-management tool tailored to AYA cancer survivors.

Methods: We conducted 5 co-design workshops—4 with AYA cancer survivors and 1 with providers who serve them. Participants were asked to provide feedback on an existing digital mindfulness course using an "I like, I wish, I wonder" framework. Then, participants were asked "How might we..." questions focused on brainstorming ideas for how the digital tool might work. Participants brainstormed responses independently and then worked as a group to categorize and expand on their ideas. Co-design workshops were autotranscribed using Webex (Cisco) software. Transcripts underwent thematic analysis with additional context provided by the products created during the workshop.

Results: Eight AYA cancer survivors (aged 15-37 years) and 4 providers (2 oncologists and 2 social workers) participated in co-design workshops. We identified 6 themes: barriers to engagement, desired content, preferences for content delivery, preferences for interface, features, and aspects to avoid. Each theme had 2-7 subthemes that we relied upon when making design decisions for the prototype.

Conclusions: Co-design workshops provided critical insights that informed the prototype development of a digital depression self-management tool tailored to AYA cancer survivors. Key takeaways that were integrated into prototype design include (1) using stories from other AYA cancer survivors to demonstrate concepts; (2) delivering content in brief lessons; and (3) using encouraging notifications, organizational tools, and reward systems to keep AYA cancer survivors engaged with the tool. Some of the themes identified in this study (eg, desired content and features) are consistent with known strategies for promoting user engagement and co-design work in other cancer survivors. However, this study extended previous research by identifying uniquely

XSL•FO
RenderX

relevant strategies for tailoring to AYA cancer survivors, such as delivering content in brief sessions to overcome the time constraints AYA cancer survivors experience, providing opportunities for private expression, and maintaining an encouraging tone throughout the tool. These data were used to inform the prototype development of a digital depression self-management tool tailored to AYA cancer survivors.

(JMIR Form Res 2025;9:e67175) doi:10.2196/67175

KEYWORDS

adolescent and young adult; cancer survivorship; depressive symptoms; self-management; co-design workshops; user-centered design; thematic analysis; intervention tailoring; digital mental health; evidence-based intervention; digital tools; psychosocial support

Introduction

Depressive Symptoms Are Prevalent and Problematic in Adolescent and Young Adult Cancer Survivors

Annually, approximately 90,000 adolescents and young adults (AYAs) between the ages of 15 and 39 years are diagnosed with cancer in the United States, with over 2.1 million AYA cancer survivors living in the United States as of 2020. These numbers are expected to rise as the incidence of cancer in this age group increases [1,2]. Cancer survivors are up to 5 times more likely to experience depression than members of the general population [3]. Furthermore, a number of studies have demonstrated that younger cancer survivors are more likely to experience elevated depressive symptoms when compared with both older survivors and healthy age-matched peers [4-6]. Among AYA cancer survivors, estimates of the prevalence of depression range from 25% to 32% [7]. Elevated rates of depressive symptoms in AYA cancer survivors are likely due to disruptions or delays in the achievement of developmental milestones (eg, completion of education and becoming a parent) [8]. Depressive symptoms in AYA cancer survivors are problematic because depressive symptoms are associated with diminished health-related quality of life, increased health care use and costs, and elevated mortality [9].

Psychosocial Interventions Are Efficacious in Treating Depressive Symptoms

The American Psychological Association's clinical practice guidelines recommend cognitive behavioral therapy (CBT) for the treatment of depressive disorders in AYAs [10]. CBT focuses on helping patients recognize and change negative automatic thoughts (by cognitive restructuring) and maladaptive behaviors (by behavioral activation) that play a role in the development and maintenance of depressive symptoms [11]. Among adults, mindfulness-based cognitive therapy (MBCT) has also been shown to be efficacious. MBCT integrates elements of CBT with mindfulness-based stress reduction to improve one's awareness of and relationship to unwanted thoughts, feelings, and bodily sensations [10,12]. More recently, there has been substantial interest in the use of positive psychology interventions to manage psychological symptoms in patients with severe and chronic illnesses. Systematic reviews and meta-analyses of trials of these interventions suggest they have small to moderate effects on depression and may be more effective than other depression interventions [13,14]. Thus, there are several topline evidence-based interventions (EBIs) for depression. Despite the elevated risk of depression in AYA cancer survivors and the existence of EBIs to address depression, it is unclear whether AYA cancer survivors can access such services. Indeed, the AYA Health Outcomes and Patient Experience study found that between 56% and 75% of AYA cancer survivors reported an unmet need for psychosocial support 12 months postdiagnosis [15].

Digital Tools Can Increase Psychosocial Intervention Reach

Digital tools are a potential solution to overcoming barriers to AYA cancer survivors' unmet needs for psychosocial support and a preferred mode of delivery for this population [16]. They allow interventions to be delivered remotely using smartphone technology that the vast majority of AYA cancer survivors already own, making them highly scalable [17]. Several studies have demonstrated the feasibility of digital tools for the self-management of physical symptoms among AYA cancer survivors [18]. Furthermore, research suggests psychosocial interventions for depression can be adapted for digital delivery with evidence for efficacy among healthy AYAs [19-25]. However, few existing digital interventions targeting depression have been evaluated in studies focused on AYA cancer survivors.

Evidence for Digital Interventions Targeting Depression in AYA Cancer Survivors Has Been Mixed

In a 2015 meta-analysis, Richter et al [26] concluded there was limited evidence for the efficacy of existing interventions targeting psychosocial distress in AYA cancer survivors, including depressive symptoms. Since that time, Campo et al [27] published the outcomes of a mindful self-compassion videoconference intervention with 25 AYA cancer survivors and demonstrated improvements in depression following the intervention (Cohen d=0.99). Zhang et al [28] recently demonstrated the feasibility and preliminary efficacy of technology-assisted CBT for reducing depression in a pilot randomized controlled trial of 17 AYA cancer survivors (Hedges g=1.12). Although these findings are promising, others have been less so. For example, Sansom-Daily et al [29] found that an online group-based CBT intervention for AYA cancer survivors early in cancer survivorship was associated with a higher level of depressive symptoms compared with a control group (P=.04). As such, it remains unclear whether CBT, mindfulness, or other interventions will be most efficacious for improving depressive symptoms in this population. Furthermore, the interventions tested by both Campo et al [27] and

XSL•FO
RenderX

Sansom-Daily et al [29] were delivered by interventionists via videoconference; while this strategy addresses some barriers to AYA cancer survivors' participation in psychosocial interventions (eg, distance or travel-related expenses), they are not available on demand as AYA cancer survivors prefer [16].

Content Tailoring and User-Centered Design Increase Efficacy and Engagement

While there is a growing body of available general mental health apps, these tools do not address the specific needs and concerns of AYA cancer survivors, such as negative thoughts about cancer and feelings of isolation from peers [30-32]. A significant consequence of this is low user engagement [33]. In the general population, only 1 out of 5 apps downloaded are used more than once [34]. The risk of nonuse of mobile health apps can be reduced with tailoring and effective design [35]. Tailoring involves developing individualized interventions based on key individual difference variables such as age and cancer history [36]. Data suggest that tailoring intervention content to the population of interest facilitates greater content relevance, which may result in more efficacious interventions and higher rates of long-term use [37]. User-centered design (UCD) processes allow for the integration of end-user input and feedback regarding features designed to increase user engagement [38]. For example, McCurdie et al [39] proposed a UCD process that consists of three iterative steps: (1) concept generation and ideation, (2) prototype design and system development, and (3) evaluation [39]. Following the final evaluation phase, the product is deployed. Users are at the center of this process, providing input at the concept generation and ideation step as well as the evaluation step. In this study, we aimed to engage AYA cancer survivors in the concept generation and ideation step of this process and design an initial prototype.

Research Context and Guiding Strategy

The work presented in this paper forms part of a wider study informed by the multiphase optimization strategy (MOST) that aims to prepare and optimize a digital depression self-management tool for AYA cancer survivors. MOST is a comprehensive framework that can be used to optimize an intervention by systematically identifying the most promising components [40,41]. MOST includes three phases: (1) preparation, which lays the groundwork for optimization (eg, identification of a guiding conceptual model, choosing candidate components, formative work, and pilot-testing); (2) optimization, in which a trial (ie, a factorial experiment) is used to identify the components that meet the optimization criterion; and (3) evaluation, in which the effectiveness of the optimized intervention is confirmed. The work of this study was informed by the tasks to be completed during the preparation phase of MOST [42].

Methods

Overview

In alignment with guidance from MOST, we devised a theoretically and empirically supported model that informed our selection of candidate intervention components (Figure 1). We then engaged in an iterative UCD process to adapt these candidate intervention components for delivery through an existing self-guided digital tool and to address the unique needs and concerns of AYA cancer survivors to ensure user engagement. We began this process by conducting co-design workshops with AYA cancer survivors and oncology providers that serve this population to generate a concept for the adapted tool. At the same time, the end users of the digital tool will be AYA cancer survivors. Input from providers supplemented AYA cancer survivors' suggestions with expertise regarding how this tool might overcome existing barriers to the management of depressive symptoms among AYA cancer survivors. Of note, the positive psychology component had already been modified for use with AYA cancer survivors and thus was not included in the co-design process [43]. Based on workshop findings, KM designed the prototype of the tool.

Figure 1. The guiding conceptual model for the preparation of a digital depression self-management tool for adolescent and young adult cancer survivors.

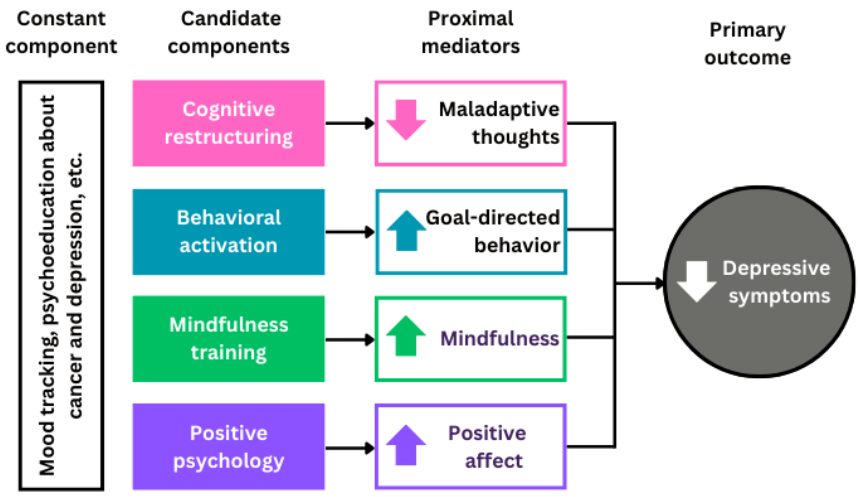

Participants

Inclusion criteria for AYA cancer survivors were as follows: (1) diagnosed with cancer between the ages of 12 and 39 years; (2) currently between the ages of 15 and 39 years; (3) at least 1 month after completion of primary treatment; (4) fluent in spoken and written English; (5) have access to a smartphone with a data plan; and (6) have access to a computer with internet,

XSL•FO
RenderX

a webcam, and a microphone that could be used to participate in a Webex (Cisco) videoconference. Notably, the National Cancer Institute (NCI) typically defines AYA cancer survivors as individuals diagnosed between the ages of 15 and 39 years and currently aged 15-39 years [1]. We chose to expand the age of diagnosis for adolescents as they are frequently diagnosed with hematologic malignancies, which require lengthy treatments such as stem cell transplants. Thus, by the time these patients reach posttreatment survivorship, they are considered emerging adults. Decreasing the age of diagnosis to 12 years for this subset of our sample ensured that adolescents currently aged 15-17 years were represented.

Inclusion criteria for oncology providers that serve AYAs were as follows: (1) licensed provider who treats AYA cancer survivors; (2) fluent in spoken and written English; (3) have access to a smartphone with a data plan; and (4) have access to a computer with internet, a webcam, and a microphone that could be used to participate in a Webex videoconference.

Potentially eligible participants were excluded if they reported a current diagnosis of a severe or persistent mental illness or reported severe suicidal ideation, including plan and intent. A safety protocol was in place to ensure appropriate care was provided if a potentially eligible participant reported suicidal ideation.

Recruitment and Enrollment

AYA cancer survivors were recruited from an NCI-designated comprehensive cancer center in the Southeastern region of the United States. Potentially eligible AYA cancer survivors were identified through an electronic medical record review. First, an automated Epic (Epic Systems) report was used to identify patients in the correct age range who had been diagnosed with cancer. Study staff then conducted additional chart review to evaluate eligibility with regard to age at diagnosis, type of diagnosis, time since completion of primary treatment, and history of severe or persistent mental illness. Potentially eligible providers were identified based on their affiliation with an AYA oncology program within North Carolina.

Study staff contacted potentially eligible participants or (in the case of adolescents) their parents or guardians by email and (if screening was not completed and no response was received from the participant after 1 week) phone call. Interested individuals or (in the case of adolescents) their parents or guardians were directed to a website with more information about the study and asked to complete a REDCap (Research Electronic Data Capture; Vanderbilt University) [44] screening form to determine their eligibility. This screening form requested that prospective participants (or, in the case of adolescents, their parents on their behalf) provide their current age, age at cancer diagnosis, the type of cancer they had been diagnosed with, how long it had been since they completed cancer treatment, whether they had ever been diagnosed with a severe or persistent mental illness, whether they had access to a smartphone and computer, and their ability to speak and read English. Eligible and interested participants and (in the case of adolescents) their parents or guardians were automatically directed to the appropriate assent and consent forms (see the *Ethical Considerations* subsection for a detailed explanation of the assent and consent process).

Upon provision of assent and consent, AYA cancer survivors and providers were asked to provide their demographic information and availability for participation in study procedures. Co-design workshop participants were scheduled in groups of at least 3 participants.

Co-Design Workshops

Overview

KMM conducted 4 online co-design workshops with AYA cancer survivors and 1 co-design workshop with providers that serve this population using Webex [45]. AYA cancer survivors and providers completed workshop activities from their homes or offices. No one else was present during workshops other than the participants and the researcher. KMM is a female individual with a PhD in Clinical Psychology and, at the time of the interview, was an instructor at Wake Forest University School of Medicine. KMM has extensive training in qualitative data collection methods and completed a certificate in UCD. She did not have a relationship with AYA cancer survivor participants before the study commencement but had reviewed their demographic and clinical characteristics. Regarding providers, KMM had previously interacted with some of the participants in professional settings but did not have any formal working relationship with them. As the principal investigator of this study, KMM's primary interest in the research topic was informing the development of a digital tool to help AYA cancer survivors better manage symptoms of depression, and this was communicated to all participants. Providers and AYA cancer survivors attended separate workshops. Each workshop lasted approximately 1 hour. Workshops were audio and video recorded and autotranscribed.

Workshop Activities

To begin each workshop, participants were provided with information about the goal of the project, namely, to develop a tool that helps AYA cancer survivors manage symptoms of depression. In addition, participants were introduced to EBIs for depression (eg, behavioral activation and mindfulness training) that we aimed to adapt and tailor for digital delivery to AYA cancer survivors. After providing some general guidelines for brainstorming and an introduction to Figma (Figma, Inc) [46] (a digital whiteboard tool), participants were asked to provide feedback on an existing digital mindfulness course (Wakeful) using an "I like, I wish, I wonder" prompts within Figma (Figure 2). In this context, wakefulness was presented to participants as an example to spur ideas. Then, participants were asked "How might we..." questions focused on brainstorming ideas for how the digital tool might work (Figure 3). Participants first brainstormed responses to each question independently and then worked as a group to categorize and expand on their ideas. Slides presented during workshops are included as Multimedia Appendix 1 (AYA cancer survivor workshops) and Multimedia Appendix 2 (provider workshop).

XSL•FO
RenderX

Figure 2. Sample responses to the "I like, I wish, I wonder" activity from the adolescent and young adult cancer survivor workshop.

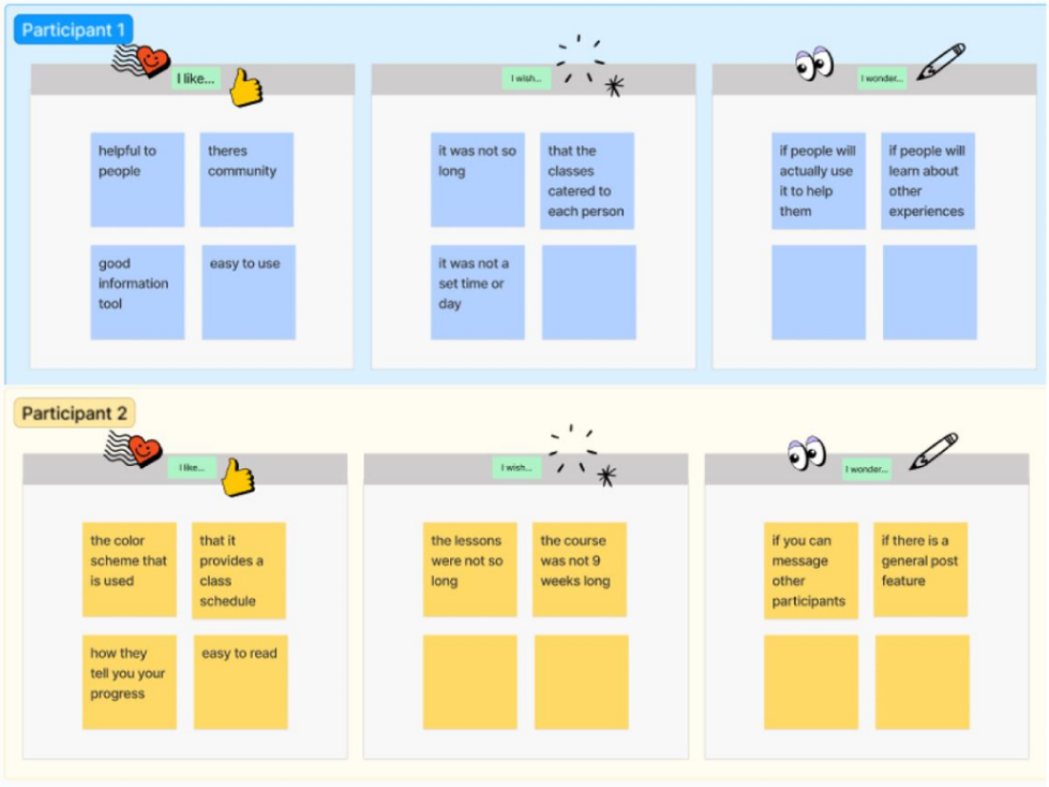

Figure 3. Sample responses to the group brainstorming activity from the provider workshop. AYA: adolescent and young adult.

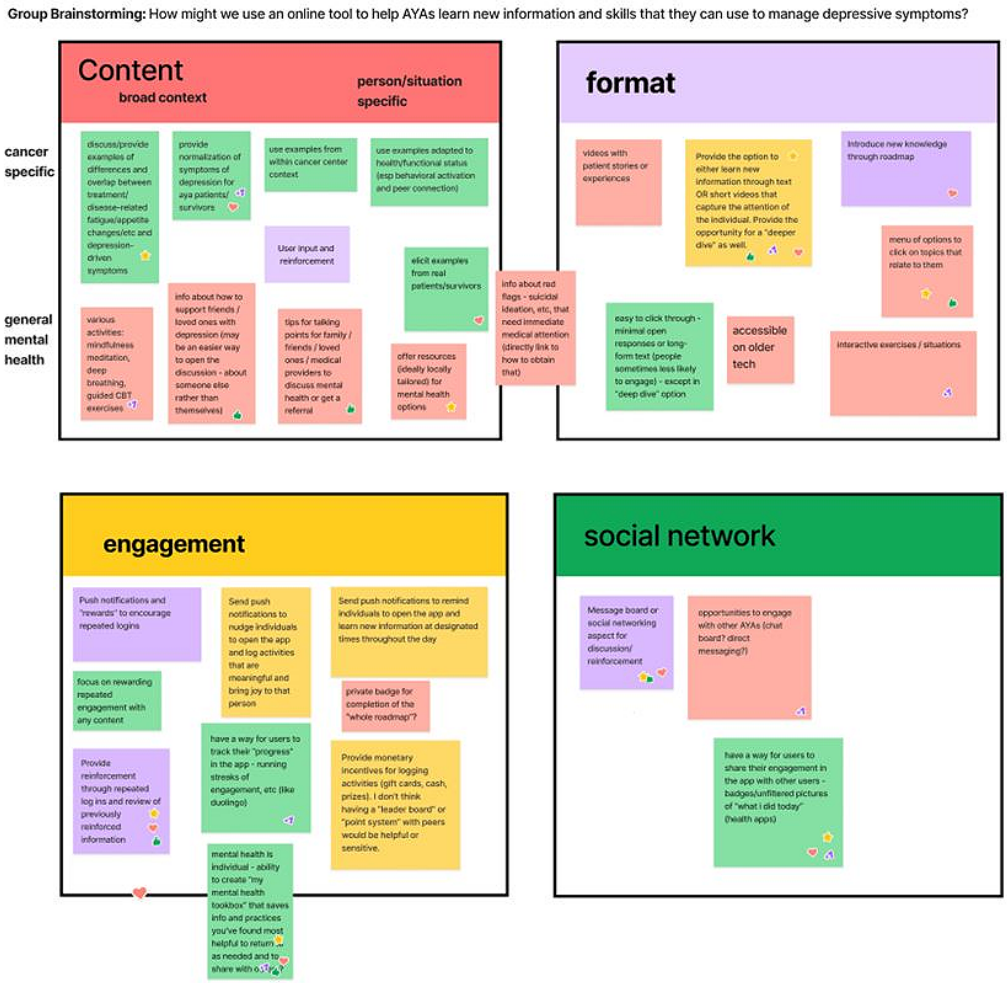

XSL•FO
RenderX

Data Processing and Analysis

Field notes were made by the researcher following each workshop. Co-design workshops were autotranscribed using Webex software. Transcripts of co-design workshops were cleaned by study team members with additional context provided by workshop products (ie, sticky notes). Reference was made to the original audio and video recordings as needed for clarification. Transcripts were not returned to participants for correction, given the availability of audio and video recordings.

RG, with oversight from KMM, conducted a thematic analysis of workshop transcripts guided by the methodological framework by Braun and Clarke [47]; the analysis was characterized by an inductive approach, where themes were directly derived from the transcripts rather than predefined before the analysis. The analysis began with a familiarization phase of understanding the content, which provided initial insights into overarching themes. Following familiarization, the coding phase included identifying and assigning a descriptive code that succinctly encapsulated the content. These codes systematically organized the data and identified patterns that later informed the identification of themes. The thematic analysis provided a comprehensive understanding of the data, highlighting the participant interactions within the co-design workshops. Participants did not provide feedback on the findings; rather, findings were used to develop a prototype that was then presented to additional participants to gather feedback.

Ethical Considerations

All study procedures were reviewed and approved by the Wake Forest University Health Sciences Institutional Review Board (IRB00075887). This included an informed consent and assent process, which primarily occurred through REDCap. Prospective participants and (in the case of adolescents) their parents were provided with a summary of the study, information about why the study was being conducted, the procedures involved in participating in the study, the duration of the study, the designation of the study as no more than minimal risk, a description of the minimal risks inherent in participating in the study (eg, potential for breach of confidentiality or psychological discomfort), the benefits to participating in the study (eg, helping others in the future), alternatives to participating in the study, compensation for participating in the study, and contact information for the principal investigator and Institutional Review Board.

In the case of young adults, eligible participants were asked to review this information about the study and required to give their consent to participate before providing their demographic and clinical information and participating in the workshop. In the case of adolescents younger than the age of 18 years, study staff initially contacted the parent of the patient. Parents were asked to review information about the study and provide their permission for the adolescent to participate. Upon provision of parental permission, study staff contacted the adolescent directly. Adolescents were provided with age-appropriate information about the study through REDCap and were required to provide their assent to participate before providing their demographic and clinical information and participating in the workshop.

All remote data collection occurred through Health Insurance Portability and Accountability Act (HIPAA)–compliant platforms, and all identifiable data were stored on secure, password-protected servers (eg, REDCap and institutional servers) only accessible by authorized study staff. Recordings of workshops were transcribed and deidentified for analysis. All information identifying participants will be deleted upon publication of the primary study outcomes.

Participants were compensated US $40 for participating in a workshop. Compensation was provided in the form of a digital gift card that was emailed to the participant following the workshop.

Results

Participants

In total, 4 providers and 8 AYA cancer survivors participated in 5 co-design workshops, consistent with the recommended number of focus group discussions needed to achieve saturation [48]. There were 2 additional AYA cancer survivors who expressed interest in the study but did not ultimately participate in a workshop due to scheduling conflicts. The provider workshop had 4 participants, while each of the AYA cancer survivors' workshops had 2 participants. All workshops lasted approximately 1 hour.

The majority of providers identified as non-Hispanic White (4/4, 100%) and heterosexual (3/4, 75%). Disciplines represented by providers included medical oncology and social work. AYA cancer survivors ranged in age from 15 to 37 years (mean 23.5, SD 8.85 years). The majority of AYA cancer survivors identified as non-Hispanic White (6/8, 75%), female (6/8, 75%), and heterosexual (7/8, 87.5%). Each AYA cancer survivor had been diagnosed with cancer between the ages of 12 and 36 years. All AYA cancer survivors had completed treatment between 1 month and 5 years before participating in the workshop.

Thematic Analysis

Thematic analysis revealed 6 overarching themes: barriers to engagement, preferences for content delivery, desired content, preferences for interface, features, and aspects to avoid. Each of these themes and their subthemes are described in detail below, with exemplar quotes. Of note, some quotes were spoken by participants, and this was pulled from the transcript; other quotes were never spoken by participants but instead typed out on the Figma board as part of engaging with the digital workshop materials. As such, we have denoted whether quotes are from the "Transcript" or "Workshop Materials," respectively.

Theme 1: Barriers to Engagement

This theme represents the challenges participants might face that hinder their ability to engage effectively with the tool. These barriers are crucial in understanding the obstacles that must be addressed to enhance participation and involvement. The 2 primary barriers were time constraints and lack of usability.

XSL•FO
RenderX

Subtheme 1.1: Time Constraints

Overall, the participants emphasized challenges with time constraints due to balancing work, school, and family responsibilities. They highlighted that these time constraints often lead to difficulties in consistently engaging in activities to support their needs, as highlighted by the quote below.

> *Yeah, probably, um, lesser time periods, maybe spread them out a little bit more. Um, For the busy momma's out there, 'cause I tell ya when you work 80+ hours a week and go to school full time and have a family.* [24-Year-old leukemia survivor, transcript]

Subtheme 1.2: Lack of Usability

A significant barrier noted by both groups of participants, particularly older AYA cancer survivor users and providers, is a lack of usability. This includes challenges related to outdated technology and user interfaces that may not be intuitive for this demographic. Addressing these usability issues is essential to enhance the overall user experience and to ensure the tool is accessible.

> *Okay, I'm like the older AYA cancer survivors, and I'm struggling with my brain. It's trying to, like, learn the new technology before it can think...so, I almost feel like, I can't think.* [37-Year-old breast cancer survivor, transcript]

> *There would be attention paid, and that the design would be such that it would be accessible on, like, multiple, even maybe older operating systems. Because especially if these people are dealing with the financial toxicity of cancer treatment. Their tech may be a couple of versions back. So, I guess paying attention to accessibility and ease of use.* [Provider 3, medical oncologist, transcript]

Theme 2: Desired Content

This theme represents the types of content that participants find most valuable and engaging. Understanding these preferences helps ensure that the content is relevant, engaging, and aligned with AYA cancer survivors' interests and needs. We identified 4 specific types of content: information about mental health, resources, AYA cancer survivors' stories, and cancer-specific content.

Subtheme 2.1: User Guidance

Participants expressed a desire for the tool to include a comprehensive guide for the tool. They indicated that having a structured guide would help them navigate the tool more effectively, understand its features, and maximize its functionalities.

> *I think, yeah, there could be like a guide like, when you first open the app or like, uh, not really a tutorial, but like sort of have it go through the app by its features and stuff. I think that would be pretty helpful.* [15-Year-old gastrointestinal stromal tumor (GIST) survivor, transcript]

Subtheme 2.2: Information About Mental Health

Participants and providers expressed a strong desire for mental health information to be integrated throughout the tool. They emphasized the importance of incorporating relevant mental health content, such as information about depression, to address the specific needs and concerns of AYA cancer survivors.

> *How depression can start, what depression can lead to, and what I need to do if I am depressed.* [19-Year-old brain tumor survivor, workshop materials]

Subtheme 2.3: Resources

Participants and providers expressed a strong desire for the tool to include a variety of resources that address different aspects of the AYA cancer survivors' cancer experiences. This includes a desire to have information about how to cope with mental health difficulties; rediscovering enjoyable activities after cancer; and external resources, particularly for emergent mental health concerns. Incorporating these elements will ensure that the tool offers comprehensive support and valuable information tailored to the needs and interests of AYA cancer survivors.

> *Information about immediate medical attention for suicidal ideation.* [Provider 1, medical oncologist, workshop materials]

> *Have helpful tools outside of this for people to manage their depression on a daily basis.* [20-Year-old thyroid cancer survivor, workshop materials]

> *This is actually a tool to help me figure out who am I now after going through this...Like, okay, I tried soccer again and on a scale of 1 to 10, I enjoyed, you know, like. More or less some sort of, I think, it would feel more. Um, I don't know what the word is like, less like, you feel a little bit, it would feel a little more empowering, I think, and a little bit less like, there's something else cancer has taken from you.* [37-Year-old breast cancer survivor, transcript]

> *Mental health is individual—ability to create "my mental health toolbox" that saves info and practice you've found most helpful to return to as needed and to share with others.* [Provider 3, medical oncologist, workshop materials]

Subtheme 2.4: AYA Cancer Survivors' Stories

Both participants and providers highlighted the importance of including stories from AYA cancer survivors within the tool. They emphasized that featuring personal experiences from peers can provide valuable insights, foster a sense of connection, and offer relatable examples that resonate with the AYA cancer survivors' community.

> *I think you could give, like, ask people if they would volunteer to give like their stories you could have certain like videos or articles from them that showed how they got through it sort of like a motivation.* [15-Year-old GIST survivor, transcript]

XSL•FO
RenderX

Include videos with patient stories or experiences. [Provider 1, medical oncologist, workshop materials]

Subtheme 2.5: Cancer-Specific Content

Providers noted that incorporating cancer-specific content could be highly beneficial. This specialized content can enhance the tool's relevance and effectiveness by offering targeted guidance and support for those affected by cancer.

Discuss/provide examples of differences and overlap between treatment/disease-related symptoms and depression-driven symptoms. [Provider 2, social worker, workshop materials]

Theme 3: Preferences for Content Delivery

This theme represents how participants prefer to receive content, which is essential for tailoring to meet the needs of users who are AYA cancer survivors. Understanding these preferences helps ensure that content is delivered in a way that aligns with participants' schedules and learning styles. We identified 3 important aspects of content delivery: intervention duration, session duration, and learning modalities.

Subtheme 3.1: Intervention Duration

Participants preferred shorter intervention durations, reflecting the multiple responsibilities that AYA cancer survivors juggle. They indicated a reduced timeframe would better accommodate their busy schedules and enhance their ability to engage consistently.

I just wish it wasn't as long as a nine-week course because, like, with me, I'm trying to go to school and, like, work at the same time. So, like it'd be nice if like, this was something that I can do, like, after school, or like between school and work. So that way it wouldn't take so long. [20-Year-old thyroid cancer survivor, transcript]

Subtheme 3.2: Session Duration

Participants expressed a desire for shorter session durations to better accommodate AYA cancer survivor users' busy schedules. They highlighted that reduced session lengths would make it easier for them to fit the sessions into their diverse and demanding routines.

Uh, probably just to make it shorter, I don't want to like be on here for a long time. [19-Year-old brain tumor survivor, transcript]

I guess if I could change one thing about it, it would be lessening the class time. [24-Year-old leukemia survivor, transcript]

Subtheme 3.3: Learning Modalities

Both participants and providers expressed a desire for varied learning and interaction modalities. They emphasized the need for diverse content delivery methods to cater to different learning styles and preferences.

I liked how like there were scripts for the videos so, in case someone doesn't want to, like, watch a video. They can just read through it. [15-Year-old GIST survivor, transcript]

Provide the option to either learn new information through text or short videos. [38-Year-old breast cancer survivor, workshop materials]

Theme 4: Preferences for Interface

This theme represents participants' preferences regarding the design and functionality of the tool's interface. Understanding these preferences helps ensure that the interface is intuitive, user-friendly, and visually engaging. We identified 2 preferences for interface design: ease of use and color scheme.

Subtheme 4.1: Ease of Use

Both participants and providers emphasized the importance of a user-friendly interface, highlighting the need for intuitive navigation and accessibility to ensure AYA cancer survivor users can interact with the tool efficiently and comfortably. Participants prefer that the tool resumes from where they left off, allowing for a seamless and uninterrupted user experience.

I like that it [a sample interface] is very easy to read. It looks like it looks very user friendly as well. [24-Year-old leukemia survivor, transcript]

It is important that the tool is easy to click through—minimal open responses. [Provider 3, medical oncologist, workshop materials]

Subtheme 4.2: Color Scheme

Participants had varied opinions on the tool prototype's color scheme. Some expressed enjoyment of the current design, appreciating its visual appeal. However, others felt that the color scheme was more suited to a younger audience, suggesting a need for a more age-appropriate or versatile design to better align with the preferences of all AYA cancer survivor users.

Personally, I like it [a sample interface] a lot. Like, I like, how all the colors kind of like I don't know, I don't know how to put it. Um, blend well, I guess. [15-Year-old thyroid cancer survivor, transcript]

I feel like I like the colors too, but I think it [a sample interface] sort of it's more of a younger audience all like the bright colors and stuff. I think if it's designed for all ages, I think there should be, like, uh, three colors that you use, and not just like all over the color spectrum. [15-Year-old GIST survivor, transcript]

Theme 5: Features

This theme represents the essential functionalities and characteristics that participants consider important for an effective tool. Identifying and integrating these features ensures that the tool meets the diverse needs of AYA cancer survivor users, making it more engaging, user-friendly, and supportive. We identified 7 functionalities participants indicated would be important to include in the tool: daily mood tracking, encouraging notifications, organizational tools, reward system, interactive exercises, connection, and private expression.

Subtheme 5.1: Daily Mood Tracking

Participants preferred daily mood tracking instead of hourly updates. They indicated that daily tracking is more manageable, less intrusive, and aligns better with their routines. Participants

XSL•FO
RenderX

thought that such "check-ins" would be helpful as they provide valuable opportunities for self-reflection.

> *15-Year-old GIST survivor (transcript): I think, yeah, I like how it would just be like, for the whole day, instead of each hour.*
>
> *Like um, mostly I'm thinking of a message that would ask how were doing.*
>
> *KMM: Yeah, yeah, so kind of like a check-in.*
>
> *21-Year-old large B cell lymphoma survivor (transcript): Yes, exactly.*
>
> *15-Year-old GIST survivor (transcript): Yeah, yeah, it [mood tracking by the hour] makes some might feel overwhelmed.*

Subtheme 5.2: Encouraging Notifications

Participants and providers noted the effectiveness of using notifications within the tool to promote user engagement. They highlighted that encouraging notifications may facilitate consistent interaction with the tool and aid in maintaining a positive outlook.

> *Yeah, I think just to get a reminder, like some days you might forget to add like your thoughts or whatever, I think just having some sort of notifications, maybe just like, once a day.* [15-Year-old GIST survivor, transcript]
>
> *Send push notifications to nudge individuals to open the app and log activities that are meaningful and bring joy to that person.* [Provider 1, medical oncologist, workshop materials]

Subtheme 5.3: Organizational Tools

Both participants and providers valued the integration of organizational elements into the tool, such as a to-do list, distinct weekly sessions, and goal or progress tracking. These features provide users with a structured framework, helping them manage tasks, stay organized, and track their progress. In addition, participants enjoyed the session schedule and progress monitoring features of an exemplar interface.

> *Include a To-do list kind of option.* [15-Year-old thyroid cancer survivor, workshop materials]
>
> *Introduce new knowledge through roadmap.* [Provider 3, medical oncologist, workshop materials]

Subtheme 5.4: Reward System

Both participants and providers recommended incorporating reward systems to enhance user engagement. These reward systems can provide positive reinforcement, recognize achievements, and encourage consistent participation.

> *I also like when you have sort of point or reward system. So, you can, I guess it gives you more motivation to what your thoughts and ideas on there.* [15-Year-old GIST survivor, transcript]
>
> *Provide monetary incentives for logging activities (gift cards, cash, prizes).* [Provider 1, medical oncologist, workshop materials]

Subtheme 5.5: Interactive Exercises

Providers recommended incorporating interactive elements into the exercises and situations within the tool.

> *Include interactive exercises/situations.* [Provider 1, medical oncologist, workshop materials]

Subtheme 5.6: Connection

Participants and providers wanted users to have the option to connect with other AYA cancer survivors through the tool as was shown in an exemplar interface. They suggested that this feature could facilitate meaningful connections and encourage mutual support among users.

> *I like how you can chat with the other people to experience the app together and talk with each other about your experience.* [20-Year-old thyroid cancer survivor, transcript]
>
> *I like that you have other people to chat with and can experience the course together.* [38-Year-old breast cancer survivor, transcript]
>
> *Message board or social networking aspect for discussion/reinforcement.* [Provider 2, social worker, workshop materials]

Subtheme 5.7: Private Expression

Participants expressed a desire for a feature within the tool that ensures the confidentiality and security of their emotions and thoughts. They emphasized the importance of having a safe space to express themselves without concerns about privacy.

> *There's some type of place where you don't have to worry at all about who's gonna see it or what's gonna see it, and you can just put it, and it's either completely just for you or you can give, like, only up to 3 people access, like, some sort of like the vault of all of my horrible feelings can go to this place. And I can just sort of dump in there, even if it's like, even if it is just for the person and just for them to see later, I think there's some value in doing that.* [38-Year-old breast cancer survivor, transcript]

Theme 6: Aspects to Avoid

Both participants and providers identified undesirable or problematic strategies for engagement. Addressing these concerns is crucial to ensure that the tool is user-friendly, engaging, and aligned with the needs of AYA cancer survivor users. We identified competition and negative tone as challenges to engagement.

Subtheme 6.1: Competition

This theme highlights the importance of avoiding features that may foster competition between participants. Providers suggested that ensuring the tool promotes a positive, non-competitive experience is essential for maintaining user engagement and fostering a sense of community.

> *Avoid having a "leaderboard" or "point system" with peers.* [Provider 1, medical oncologist, workshop materials]

XSL•FO
RenderX

Subtheme 6.2: Negative Tone

Participants emphasized the need to avoid negative or potentially discouraging content. It is important to ensure that all content is supportive, uplifting, and encouraging.

> *People who have a lot of self-hate towards themselves that might wanna want to put negative responses towards themselves that could possibly make them feel worse.* [21-Year-old survivor of multiple cancers, transcript]

The Prototype Depression Self-Management Tool

Based on feedback and recommendations from workshop participants, the initial design of the tool was determined by the study team (see Table 1 for a summary of how each theme and subtheme informed design decisions). There were a few conflicts between the providers' and AYA cancer survivors' suggestions; however, in the case of conflicting suggestions, AYA cancer survivors' preferences were prioritized as they are the target end users of the digital tool.

KM used Canva (Canva Inc) [49] to design a prototype of a depression self-management tool tailored to the needs and preferences of AYA cancer survivors. The tool was designed with the aim of providing mood tracking, psychoeducation about depression and cancer, and 4 adapted evidence-based interventions (aEBIs) for depression to this population via a web-based, app-optimized smartphone delivery (Figure 4). Each aEBI was designed to be housed in its own module, which would consist of a series of brief lessons. Each lesson would include educational content regarding the lesson topic (presented via video and transcript), a story of an AYA cancer survivor that illustrated the lesson topic, app questions that asked the user to apply the educational information to the AYA cancer survivors' story, an activity in which they reflected on what they learned related to their experience, and a practice activity that asked users to apply what they learned in their day-to-day life (Figure 5). The tool also featured a discussion board section that would allow users to connect with one another about the content presented via the tool.

XSL•FO
RenderX

Table 1. Summary of how each theme and subtheme informed design decisions.

Theme and subtheme	Design decision
Barriers to engagement	
Time constraints	Each aEBI[a] will be delivered via a series of brief, on-demand lessons that users can complete when it is convenient for them.
Lack of usability	A web-based application with a responsive design that has established usability among AYA[b] cancer survivors will be adapted for content delivery.
Desired content	
User guidance	During research studies, participants will complete an orientation session that includes instructions on how to use the digital tool. Once the tool has been optimized, a user guide will be developed and provided to users.
Information about mental health	Relevant mental health information will be included in the constant component so that it is provided to all participants. External resources for emergent mental health concerns will be provided to users via the constant component.
Resources	Each aEBI will include relevant practice activities to help users apply content in their daily lives. Future research will inform how to best integrate these practice activities into the digital tool.
AYA cancer survivors stories	The tool will use AYA cancer survivors' stories to illustrate educational content.
Cancer-specific content	The constant component and aEBIs will include information that is tailored to AYA cancer survivors.
Preferences for content delivery	
Intervention duration	Content will be delivered over a duration shorter than 9 weeks. Additional research will inform intervention duration.
Session duration	The constant component and each aEBI will be delivered via a series of brief lessons. Additional research will inform session duration.
Learning modalities	Educational content will be presented via multiple modalities, such as a brief video with a transcript provided. Additional research will inform what content should be presented via each modality.
Preferences for interface	
Ease of use	A web-based application with a responsive design that has established usability among AYA cancer survivors will be adapted for content delivery.
Color scheme	The color scheme will be updated to reflect the age/maturity level of the target user population.
Features	
Daily mood tracking	Users will be asked to track their mood daily rather than hourly.
Encouraging notifications	Users will receive encouraging text message reminders to engage with the tool. Additional research will inform the frequency of these messages.
Organizational tools	The tool will display users' progress toward completion of content and provide them with clear next steps.
Reward system	Additional research will be conducted to determine how to reward users for engagement in a way that is feasible and scalable.
Interactive exercise	Users will be asked to apply the information they learn to the stories of other AYA cancer survivors as well as their own experiences. Future research will explore how users prefer to input such responses in the tool.
Connection	The tool will include discussion boards that allow users to connect with other AYA cancer survivors.
Private expression	Users will be able to choose whether their responses to reflection questions are shared with other users.
Aspects to avoid	
Competition	Users will not be able to see the progress of other users.
Negative tone	A positive and encouraging tone will be used throughout the tool.

[a]aEBI: adapted evidence-based intervention.

[b]AYA: adolescent and young adult.

XSL•FO
RenderX

Figure 4. Dashboard and learning modules.

Figure 5. Example lesson structure and design. AYA: adolescent and young adult.

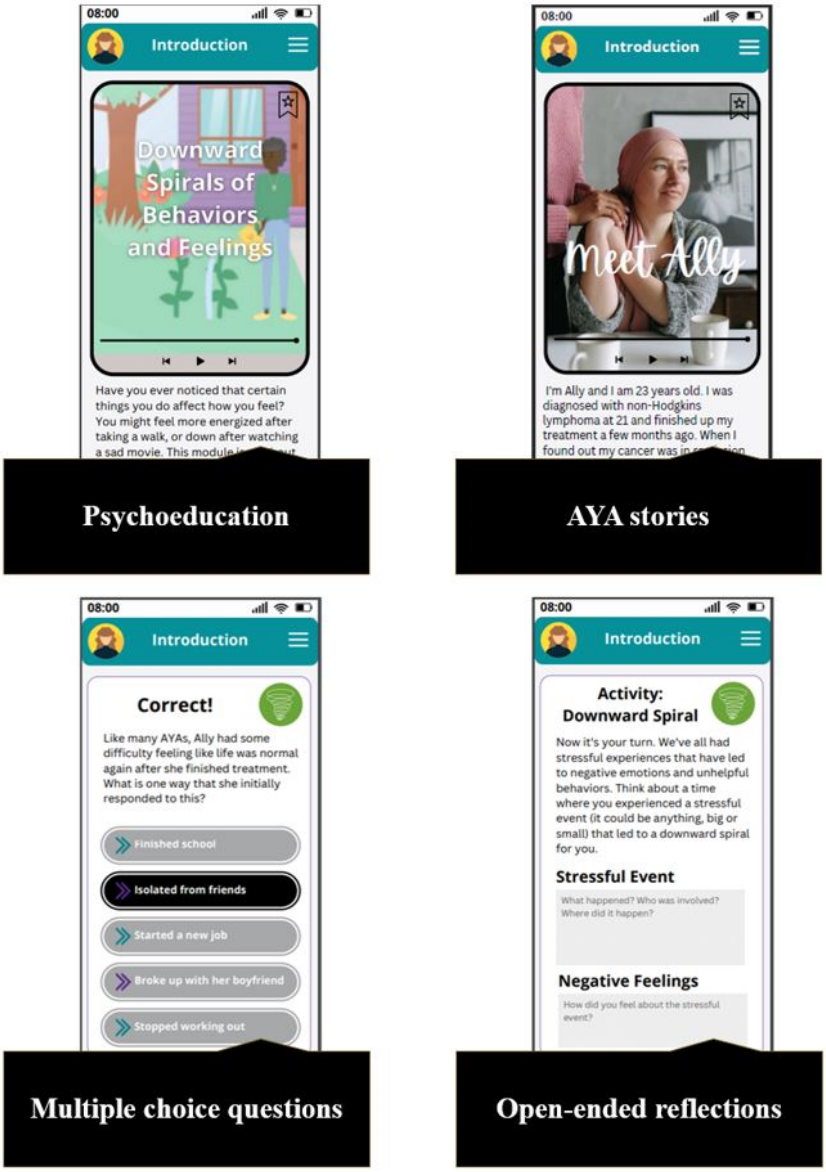

XSL•FO
RenderX

Discussion

Principal Findings

AYA cancer survivors often experience depressive symptoms that go unaddressed. Digital mental health tools are a promising strategy to ensure that patients have the skills and tools necessary to manage symptoms of depression. This study described the initial co-design of a depression self-management intervention to ensure that the final product is tailored to the needs and preferences of AYA cancer survivors, fulfilling the concept generation and ideation step of the UCD process set forth by McCurdie et al [39] and contributing to the tasks of the preparation phase of the MOST framework [42]. Feedback from AYA cancer survivors during the co-design workshops provided critical insights into ways in which a digital tool could be tailored to engage this population in evidence-based strategies for depression self-management. Specifically, users emphasized the importance of overcoming barriers to engagement, such as time constraints; provided suggestions for content to be included in the tool and how it should be delivered; and recommended features to enhance engagement with the tool. Using these findings, we designed a midfidelity prototype of a digital depression self-management tool tailored to the needs and preferences of AYA cancer survivors.

Comparison With Previous Work

A primary goal of the UCD process in this project was to ensure the tool we develop is engaging for AYA cancer survivors. As such, many of the themes identified in this study are consistent with previously identified strategies for promoting user engagement. For example, a systematic review of technology-supported strategies in digital mental health interventions identified four categories of engagement strategies used in existing tools: (1) reminders to use the tool, (2) tailored information, (3) peer support, and (4) coaching techniques [50]. In our study, the first 3 of these categories were emphasized by participants as important to include in the digital tool we are designing.

The themes identified in this study also broadly align with findings from co-design work in other groups of cancer survivors. For example, Adler et al [51] engaged cancer survivors with physical disabilities in a co-design process of a self-management tool. Their team identified themes of “engaging” and “socialization” that have significant overlap with our “features” theme. Participants in both studies suggested the use of participative features, progress tracking, and opportunities to connect with others who have had similar experiences as strategies to help users engage with a self-management tool. Similarly, in a co-design study focused on a tool to help cancer survivors engage in healthy behaviors, Leske et al [52] identified relatability and relevance as a theme, which has considerable overlap with the AYA stories and cancer-specific content subthemes of our desired content theme [52]. In both studies, it became evident that providing cancer-specific information and stories from peers would enhance the relevance of the tool for the target population.

Finally, this study extended previous research by engaging a novel population of AYA cancer survivors and their providers in the co-design process. This allowed us to elucidate strategies for tailoring that are uniquely relevant to this population, such as the importance of ensuring that the content is delivered in brief sessions to overcome the time constraints AYA cancer survivors experience, providing opportunities for private expression of their thoughts and feelings, and using a positive and encouraging tone throughout the tool.

Limitations

Our findings are limited by the size and characteristics of our research sample and the use of web-based rather than in-person activities. Our total sample size of 12 individuals across 5 workshops is consistent with recommendations for reaching saturation and sample sizes in design research with similar populations [48,53]. However, this small group of participants inherently limits the generalizability of this study. In addition, despite efforts to schedule as many AYA cancer survivors for the same workshop as possible, difficulties in scheduling resulted in groups of 2. These group brainstorming sessions would have benefited from larger groups with more people to generate ideas. Furthermore, despite efforts to recruit a diverse group of AYA cancer survivors and providers, we did not achieve a representative research sample with regard to race, ethnicity, gender, and sexual orientation. However, our AYA cancer survivors sample was representative of the vast majority of the AYA cancer survivors’ age range and common cancer diagnoses. Finally, the activities we were able to engage AYA cancer survivors in were limited to the capabilities of existing digital tools; we were not able to use the materials typically relied upon during in-person, co-design sessions, such as markers, Post-it notes, and posters. At times, we also encountered technical difficulties with the tools that limited user engagement in co-design activities. For example, one user was not able to load Figma on her device during the workshop, so she was limited to adding her insights verbally or through the chat feature of Webex. However, the use of online tools for co-design ultimately provided invaluable flexibility in terms of location and scheduling to ensure the engagement of busy AYA cancer survivors in the process.

Conclusions

This co-design study improved our understanding of the needs and preferences of AYA cancer survivors for a digital tool that may help them better manage symptoms of depression. This process allowed us to develop a shared vision of what such a digital tool would look like and how it would work. As such, our findings informed the development of a prototype of the tool. In addition, the findings from this work provide insights into the preferences of AYA cancer survivors for digital mental health tools more broadly. For example, AYA cancer survivors broadly agreed on the need for brief intervention models. However, it remains to be seen whether delivering interventions in a brief manner negatively impacts efficacy. As such, findings from this study are likely informative for future research aimed at designing engaging digital interventions for AYA cancer survivors. The next phase of this study is to have AYA cancer survivors evaluate the prototype through semistructured individual interviews to inform the design of a working prototype. This formative work will continue to contribute to

XSL•FO
RenderX

the tasks set forth by the preparation phase of MOST, as will a subsequent pilot study of the tool.

Acknowledgments

Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health under award K99CA248701. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We would like to thank Ms Nyla Parks for her assistance with cleaning transcriptions of the co-design workshops and Ms Kamah Woelfel for her assistance in recruiting participants for this study. The authors attest that there was no use of generative artificial intelligence technology in the generation of text, figures, or other informational content of this paper.

Data Availability

The datasets generated or analyzed during this study are available in the Open Science Framework repository [54].

Conflicts of Interest

None declared.

Multimedia Appendix 1
Adolescent and young adult cancer survivor workshop slides.
[PPTX File , 6092 KB - formative_v9i1e67175_app1.pptx]

Multimedia Appendix 2
Provider workshop slides.
[PPTX File , 5096 KB - formative_v9i1e67175_app2.pptx]

References

1. Miller KD, Fidler-Benaoudia M, Keegan TH, Hipp HS, Jemal A, Siegel RL. Cancer statistics for adolescents and young adults, 2020. CA Cancer J Clin 2020;70(6):443-459 [FREE Full text] [doi: 10.3322/caac.21637] [Medline: 32940362]
2. Page LL, Devasia TP, Mariotto A, Gallicchio L, Mollica MA, Tonorezos E. Prevalence of cancer survivors diagnosed during adolescence and young adulthood in the United States. J Natl Cancer Inst 2024. [doi: 10.1093/jnci/djae250] [Medline: 39383200]
3. Hartung TJ, Brähler E, Faller H, Härter M, Hinz A, Johansen C, et al. The risk of being depressed is significantly higher in cancer patients than in the general population: Prevalence and severity of depressive symptoms across major cancer types. Eur J Cancer 2017;72:46-53. [doi: 10.1016/j.ejca.2016.11.017] [Medline: 28024266]
4. Salsman JM, Garcia SF, Yanez B, Sanford SD, Snyder MA, Victorson D. Physical, emotional, and social health differences between posttreatment young adults with cancer and matched healthy controls. Cancer 2014;120(15):2247-2254 [FREE Full text] [doi: 10.1002/cncr.28739] [Medline: 24888335]
5. Park EM, Rosenstein DL. Depression in adolescents and young adults with cancer. Dialogues Clin Neurosci 2022;17(2):171-180. [doi: 10.31887/dcns.2015.17.2/epark]
6. Tai E, Buchanan N, Townsend J, Fairley T, Moore A, Richardson LC. Health status of adolescent and young adult cancer survivors. Cancer 2012;118(19):4884-4891 [FREE Full text] [doi: 10.1002/cncr.27445] [Medline: 22688896]
7. Lang MJ, David V, Giese-Davis J. The age conundrum: a scoping review of younger age or adolescent and young adult as a risk factor for clinical distress, depression, or anxiety in cancer. J Adolesc Young Adult Oncol 2015;4(4):157-173 [FREE Full text] [doi: 10.1089/jayao.2015.0005] [Medline: 26697266]
8. Bellizzi KM, Smith A, Schmidt S, Keegan THM, Zebrack B, Lynch CF, et al. Positive and negative psychosocial impact of being diagnosed with cancer as an adolescent or young adult. Cancer 2012;118(20):5155-5162 [FREE Full text] [doi: 10.1002/cncr.27512] [Medline: 22415815]
9. Satin JR, Linden W, Phillips MJ. Depression as a predictor of disease progression and mortality in cancer patients: a meta-analysis. Cancer 2009;115(22):5349-5361 [FREE Full text] [doi: 10.1002/cncr.24561] [Medline: 19753617]
10. McQuaid JR, Lin EH, Barber JP, Breland-Noble AM, Cuijpers P, Greenberg LS. Clinical Practice Guideline for the Treatment of Depression Across Three Age Cohorts. Washington, DC: American Psychological Association; 2019.
11. Hollon SD, Beck AT. Cognitive therapy of depression. In: Cognitive-Behavioral Interventions: Theory, Research, and Procedures. New York, NY: Academic Press; 1979:153-203.
12. Segal Z, Williams M, Teasdale J. Mindfulness-Based Cognitive Therapy for Depression. New York, NY: Guilford Press; 2012.

XSL•FO
RenderX

13. Hendriks T, Schotanus-Dijkstra M, Hassankhan A, de Jong J, Bohlmeijer E. The efficacy of multi-component positive psychology interventions: a systematic review and meta-analysis of randomized controlled trials. J Happiness Stud 2019;21(1):357-390. [doi: 10.1007/s10902-019-00082-1]
14. Lim WL, Tierney S. The effectiveness of positive psychology interventions for promoting well-being of adults experiencing depression compared to other active psychological treatments: a systematic review and meta-analysis. J Happiness Stud 2023;24(1):249-273 [FREE Full text] [doi: 10.1007/s10902-022-00598-z] [Medline: 36373089]
15. Smith AW, Keegan T, Hamilton A, Lynch C, Wu X, Schwartz SM, et al. Understanding care and outcomes in adolescents and young adult with cancer: a review of the AYA HOPE study. Pediatr Blood Cancer 2019;66(1):e27486 [FREE Full text] [doi: 10.1002/pbc.27486] [Medline: 30294882]
16. Rabin C, Simpson N, Morrow K, Pinto B. Intervention format and delivery preferences among young adult cancer survivors. Int J Behav Med 2013;20(2):304-310. [doi: 10.1007/s12529-012-9227-4] [Medline: 22328444]
17. Mobile fact sheet. Pew Research Center. URL: https://www.pewresearch.org/internet/fact-sheet/mobile/ [accessed 2024-09-06]
18. Devine KA, Viola AS, Coups EJ, Wu YP. Digital health interventions for adolescent and young adult cancer survivors. JCO Clin Cancer Inform 2018;2:1-15 [FREE Full text] [doi: 10.1200/CCI.17.00138] [Medline: 30652583]
19. Rathbone AL, Clarry L, Prescott J. Assessing the efficacy of mobile health apps using the basic principles of cognitive behavioral therapy: systematic review. J Med Internet Res 2017;19(11):e399 [FREE Full text] [doi: 10.2196/jmir.8598] [Medline: 29187342]
20. Huguet A, Rao S, McGrath PJ, Wozney L, Wheaton M, Conrod J, et al. A systematic review of cognitive behavioral therapy and behavioral activation apps for depression. PLoS One 2016;11(5):e0154248 [FREE Full text] [doi: 10.1371/journal.pone.0154248] [Medline: 27135410]
21. Dahne J, Lejuez CW, Diaz VA, Player MS, Kustanowitz J, Felton JW, et al. Pilot randomized trial of a self-help behavioral activation mobile app for utilization in primary care. Behav Ther 2019;50(4):817-827 [FREE Full text] [doi: 10.1016/j.beth.2018.12.003] [Medline: 31208690]
22. Ebert DD, Zarski AC, Christensen H, Stikkelbroek Y, Cuijpers P, Berking M, et al. Internet and computer-based cognitive behavioral therapy for anxiety and depression in youth: a meta-analysis of randomized controlled outcome trials. PLoS One 2015;10(3):e0119895 [FREE Full text] [doi: 10.1371/journal.pone.0119895] [Medline: 25786025]
23. Ye X, Bapuji SB, Winters SE, Struthers A, Raynard M, Metge C, et al. Effectiveness of internet-based interventions for children, youth, and young adults with anxiety and/or depression: a systematic review and meta-analysis. BMC Health Serv Res 2014;14:313 [FREE Full text] [doi: 10.1186/1472-6963-14-313] [Medline: 25037951]
24. Clarke G, Kelleher C, Hornbrook M, Debar L, Dickerson J, Gullion C. Randomized effectiveness trial of an internet, pure self-help, cognitive behavioral intervention for depressive symptoms in young adults. Cogn Behav Ther 2009;38(4):222-234 [FREE Full text] [doi: 10.1080/16506070802675353] [Medline: 19440896]
25. Clarke AM, Kuosmanen T, Barry MM. A systematic review of online youth mental health promotion and prevention interventions. J Youth Adolesc 2015;44(1):90-113. [doi: 10.1007/s10964-014-0165-0] [Medline: 25115460]
26. Richter D, Koehler M, Friedrich M, Hilgendorf I, Mehnert A, Weißflog G. Psychosocial interventions for adolescents and young adult cancer patients: a systematic review and meta-analysis. Crit Rev Oncol Hematol 2015;95(3):370-386. [doi: 10.1016/j.critrevonc.2015.04.003] [Medline: 25922217]
27. Campo RA, Bluth K, Santacroce SJ, Knapik S, Tan J, Gold S, et al. A mindful self-compassion videoconference intervention for nationally recruited posttreatment young adult cancer survivors: feasibility, acceptability, and psychosocial outcomes. Support Care Cancer 2017;25(6):1759-1768. [doi: 10.1007/s00520-017-3586-y] [Medline: 28105523]
28. Zhang A, Zebrack B, Acquati C, Roth M, Levin NJ, Wang K, et al. Technology-assisted psychosocial interventions for childhood, adolescent, and young adult cancer survivors: a systematic review and meta-analysis. J Adolesc Young Adult Oncol 2022;11(1):6-16 [FREE Full text] [doi: 10.1089/jayao.2021.0012] [Medline: 33960845]
29. Sansom-Daly U, Wakefield C, Ellis S, McGill B, Donoghoe M, Butow P, et al. Online, group-based psychological support for adolescent and young adult cancer survivors: results from the recapture life randomized trial. Cancers (Basel) 2021;13(10):2460 [FREE Full text] [doi: 10.3390/cancers13102460] [Medline: 34070134]
30. Smith AW, Seibel NL, Lewis DR, Albritton KH, Blair DF, Blanke CD, et al. Next steps for adolescent and young adult oncology workshop: an update on progress and recommendations for the future. Cancer 2016;122(7):988-999 [FREE Full text] [doi: 10.1002/cncr.29870] [Medline: 26849003]
31. Zebrack B, Bleyer A, Albritton K, Medearis S, Tang J. Assessing the health care needs of adolescent and young adult cancer patients and survivors. Cancer 2006;107(12):2915-2923 [FREE Full text] [doi: 10.1002/cncr.22338] [Medline: 17103383]
32. Zebrack B, Isaacson S. Psychosocial care of adolescent and young adult patients with cancer and survivors. JCO 2012;30(11):1221-1226. [doi: 10.1200/jco.2011.39.5467]
33. Torous J, Nicholas J, Larsen ME, Firth J, Christensen H. Clinical review of user engagement with mental health smartphone apps: evidence, theory and improvements. Evid Based Ment Health 2018;21(3):116-119 [FREE Full text] [doi: 10.1136/eb-2018-102891] [Medline: 29871870]
34. App Uninstall Statistics 2018: 21% of users abandon an app after one use. Localytics. URL: https://uplandsoftware.com/localytics/resources/blog/21-percent-of-users-abandon-apps-after-one-use/ [accessed 2024-09-06]

XSL•FO
RenderX

35. Lyon AR, Koerner K. User-centered design for psychosocial intervention development and implementation. Clin Psychol (New York) 2016;23(2):180-200 [FREE Full text] [doi: 10.1111/cpsp.12154] [Medline: 29456295]
36. Hawkins RP, Kreuter M, Resnicow K, Fishbein M, Dijkstra A. Understanding tailoring in communicating about health. Health Educ Res 2008;23(3):454-466. [doi: 10.1093/her/cyn004] [Medline: 18349033]
37. Lustria MLA, Noar SM, Cortese J, van Stee SK, Glueckauf RL, Lee J. A meta-analysis of web-delivered tailored health behavior change interventions. J Health Commun 2013;18(9):1039-1069. [doi: 10.1080/10810730.2013.768727] [Medline: 23750972]
38. Schueller SM, Muñoz R, Mohr DC. Realizing the potential of behavioral intervention technologies. Curr Dir Psychol Sci 2013;22(6):478-483. [doi: 10.1177/0963721413495872]
39. McCurdie T, Taneva S, Casselman M, Yeung M, McDaniel C, Ho W, et al. mHealth consumer apps: the case for user-centered design. Biomed Instrum Technol 2012;Suppl:49-56. [doi: 10.2345/0899-8205-46.s2.49] [Medline: 23039777]
40. Collins LM, Murphy SA, Strecher V. The multiphase optimization strategy (MOST) and the sequential multiple assignment randomized trial (SMART): new methods for more potent eHealth interventions. Am J Prev Med 2007;32(5 Suppl):S112-S118 [FREE Full text] [doi: 10.1016/j.amepre.2007.01.022] [Medline: 17466815]
41. Collins LM, Nahum-Shani I, Guastaferro K, Strayhorn JC, Vanness DJ, Murphy SA. Intervention optimization: a paradigm shift and its potential implications for clinical psychology. Annu Rev Clin Psychol 2024;20(1):21-47 [FREE Full text] [doi: 10.1146/annurev-clinpsy-080822-051119] [Medline: 38316143]
42. Landoll RR, Vargas SE, Samardzic KB, Clark MF, Guastaferro K. The preparation phase in the multiphase optimization strategy (MOST): a systematic review and introduction of a reporting checklist. Transl Behav Med 2022;12(2):291-303. [doi: 10.1093/tbm/ibab146] [Medline: 34850214]
43. Salsman JM, McLouth LE, Cohn M, Tooze JA, Sorkin M, Moskowitz JT. A web-based, positive emotion skills intervention for enhancing posttreatment psychological well-being in young adult cancer survivors (EMPOWER): protocol for a single-arm feasibility trial. JMIR Res Protoc 2020;9(5):e17078 [FREE Full text] [doi: 10.2196/17078] [Medline: 32463014]
44. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009;42(2):377-381 [FREE Full text] [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
45. Webex. URL: https://web.webex.com/sign-in [accessed 2024-09-12]
46. Figma: The Collaborative Interface Design Tool. URL: https://www.figma.com/ [accessed 2024-09-06]
47. Braun V, Clarke V. Successful Qualitative Research: A Practical Guide for Beginners. Thousand Oaks, CA: Sage Publications Ltd; 2013.
48. Hennink M, Kaiser BN. Sample sizes for saturation in qualitative research: a systematic review of empirical tests. Soc Sci Med 2022;292:114523 [FREE Full text] [doi: 10.1016/j.socscimed.2021.114523] [Medline: 34785096]
49. Canva. URL: https://www.canva.com/ [accessed 2024-09-06]
50. Gan DZQ, McGillivray L, Larsen ME, Christensen H, Torok M. Technology-supported strategies for promoting user engagement with digital mental health interventions: a systematic review. Digit Health 2022;8:20552076221098268 [FREE Full text] [doi: 10.1177/20552076221098268] [Medline: 35677785]
51. Adler RF, Morales P, Sotelo J, Magasi S. Developing an mHealth app for empowering cancer survivors with disabilities: co-design study. JMIR Form Res 2022;6(7):e37706 [FREE Full text] [doi: 10.2196/37706] [Medline: 35881439]
52. Leske M, Koczwara B, Blunt J, Morris J, Eakin E, Short CE, et al. Co-designing healthy living after cancer online: an online nutrition, physical activity, and psychosocial intervention for post-treatment cancer survivors. J Cancer Surviv 2024;18(2):606-616 [FREE Full text] [doi: 10.1007/s11764-022-01284-y] [Medline: 36374435]
53. Schwartz LA, Psihogios AM, Henry-Moss D, Daniel LC, Ver Hoeve ES, Velazquez-Martin B, et al. Iterative development of a tailored mHealth intervention for adolescent and young adult survivors of childhood cancer. CPPP 2019 Mar;7(1):31-43. [doi: 10.1037/cpp0000272]
54. Co-design workshop transcripts. Open Science Framework. URL: https://osf.io/8txsu/files/osfstorage?view_only=a51f736aedd44dff9ef59793f7bb7a59 [accessed 2025-03-04]

Abbreviations

aEBI: adapted evidence-based intervention
AYA: adolescent and young adult
CBT: cognitive behavioral therapy
EBI: evidence-based intervention
GIST: gastrointestinal stromal tumor
HIPAA: Health Insurance Portability and Accountability Act
MBCT: mindfulness-based cognitive therapy
MOST: multiphase optimization strategy
NCI: National Cancer Institute
REDCap: Research Electronic Data Capture

XSL•FO
RenderX

UCD: user-centered design

Edited by A Mavragani; submitted 10.10.24; peer-reviewed by K Devine; comments to author 10.01.25; revised version received 28.01.25; accepted 11.02.25; published 24.03.25.

Please cite as:
Murphy KM, Glock R, Victorson D, Reddy M, Birken SA, Salsman JM
Co-Design of a Depression Self-Management Tool for Adolescent and Young Adult Cancer Survivors: User-Centered Design Approach
JMIR Form Res 2025;9:e67175
URL: https://formative.jmir.org/2025/1/e67175
doi:10.2196/67175
PMID:40126551

©Karly M Murphy, Rachel Glock, David Victorson, Madhu Reddy, Sarah A Birken, John M Salsman. Originally published in JMIR Formative Research (https://formative.jmir.org), 24.03.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Safe Listening Beliefs, Attitudes, and Practices Among Gamers and Esports Participants: International Web-Based Survey

Nicola Diviani[1,2,3], PhD; Shelly Chadha[4], PhD; Peter Mulas[4], MDeSc, MClinAud; Sara Rubinelli[1,2,3], PD, PhD

[1]Swiss Paraplegic Research, Nottwil, Switzerland

[2]Faculty of Health Sciences and Medicine, University of Lucerne, Lucerne, Switzerland

[3]Center for Rehabilitation in Global Health Systems, University of Lucerne, Lucerne, Switzerland

[4]Department of Noncommunicable Diseases, World Health Organization, Geneva, Switzerland

Corresponding Author:
Nicola Diviani, PhD
Swiss Paraplegic Research
Guido A Zäch Strasse 4
Nottwil, 6207
Switzerland
Phone: 41 419396583
Email: nicola.diviani@paraplegie.ch

Abstract

Background: The global rise of video gaming and esports has raised significant concerns about hearing loss due to loud sound exposure. While these activities provide entertainment and have applications in health care, the auditory health risks and behavioral factors influencing listening habits among gamers remain underexplored. Research is needed to develop tailored interventions that address the unique barriers, attitudes, and beliefs of gamers and esports participants, promoting safer listening practices and minimizing auditory health risks.

Objective: This study aimed to explore listening behaviors, attitudes, and awareness regarding hearing health risks among video gamers and esports participants. The findings are intended to guide the design and implementation of technological features that encourage safer listening practices, in alignment with the World Health Organization's Safe Listening initiative.

Methods: An open web-based survey was conducted from September 2022 to January 2023, targeting video gamers and esports enthusiasts. Participants were recruited via World Health Organization social media platforms and outreach to stakeholders. The survey assessed gaming behaviors, listening habits, awareness about hearing health, beliefs, readiness to change listening behaviors, and communication preferences. Data were analyzed using descriptive statistics and multinomial logistic regression.

Results: A total of 488 responses were collected, with 67.2% (n=328) of participants identifying as male, and 56.4% (n=275) having a college degree or higher. Of the respondents, 90.8% (n=443) were actively engaged in video gaming, while 54.9% (n=268) viewed esports, and 13.9% (n=68) participated in esports events. Notably, 24.8% (n=110) of gamers, 18.3% (n=49) of esports viewers, and 37.1% (n=23) of esports players reported using high or very high volume settings. Despite around half of the participants experiencing symptoms indicative of hearing damage (eg, ringing in the ears), only 34.3% (n=152) of gamers, 35.8% (n=92) of esports players, and 39.7% (n=27) of esports viewers reported taking sound breaks every hour. The study identified a balanced distribution across readiness-to-change stages, with 30.3% (n=148) in the precontemplation stage, 35.3% (n=173) in the contemplation stage, and 34.2% (n=167) in the action stage. Factors such as perceived susceptibility to hearing loss, perceived benefits of preventive action, and self-efficacy significantly influenced readiness to change. Communication preferences indicated that 51% (n=249) of participants were interested in receiving more information on hearing health, with health care professionals and governmental agencies being the most trusted sources.

Conclusions: The findings highlight an urgent need for interventions to promote safe listening practices among gamers, emphasizing a gap between awareness and preventive action. The integration of safe listening features into video games and esports platforms, along with targeted communication strategies, can enhance auditory health awareness and protective behaviors. Future research should evaluate the effectiveness of these interventions to ensure comprehensive auditory health protection in the digital entertainment sector.

(JMIR Form Res 2025;9:e60476) doi:10.2196/60476

XSL•FO
RenderX

KEYWORDS

video gaming; esports; hearing health; safe listening; auditory risk; health behavior; intervention design; games; listening; auditory; survey; college; data analysis; awareness; listening behavior; gaming

Introduction

Video gaming and esports have undergone an exponential rise in recent years, evolving into a major entertainment force that captivates individuals across different age groups [1]. With projections suggesting that the global video game market will grow at an annual rate of 12.9% from 2022 to 2030—reaching an estimated US $399.6 billion by 2028—and the global gamer population increasing from 2.7 billion in 2021 to an expected 3.04 billion by 2027, the influence of these digital platforms is significant [2,3]. Concurrently, esports has experienced a surge in popularity, with its global audience forecasted to reach 577 million by 2025, and the market is anticipated to grow at a rate of 16.7% annually, resulting in a value of US $4.28 billion by 2030 [4,5].

Beyond serving as engaging and entertaining mediums, video games and esports are increasingly acknowledged for their potential in health care, particularly through serious gaming applications aimed at education, rehabilitation, and therapy, thus highlighting their versatile utility beyond mere entertainment [6-8]. Despite these benefits, concerns about the potential health risks associated with video gaming and esports have emerged. The documented impacts on mental and physical health are compounded by the risks of hearing loss, tinnitus, and other auditory disorders resulting from exposure to loud and persistent sound effects inherent in these activities [9-11]. Echoing the World Health Organization's (WHO) warning, over 1 billion young people are at an elevated risk of hearing loss due to recreational exposure to loud sounds, including the use of personal audio devices and attendance at loud public events [12].

The public health realm has seen significant attention directed toward the issue of hearing loss from recreational sound exposure, leading to a wealth of research and the development of global standards aimed at mitigating this risk. The pervasive nature of this issue, covering activities ranging from personal listening device use to attendance at live events with high decibel levels, has been illuminated through extensive research. Such studies have consistently underscored the risk of hearing damage from prolonged exposure to loud music via earphones or headphones and attendance at entertainment venues where volumes exceed safe listening thresholds, potentially leading to noise-induced hearing loss [13-16].

In response, the WHO and the International Telecommunication Union have collaborated to launch the "Make Listening Safe" initiative, which recommends standards for incorporating safe listening features into personal audio devices. This includes software that monitors volume levels and exposure time, aiming to regulate sound exposure in recreational settings and promote a balance between enjoyment and health [17,18]. These standards not only aim to raise awareness among the public and stakeholders about the risks associated with loud sound exposure but also to provide concrete guidelines for safe listening practices, such as sound level limits and user warnings about unsafe listening volumes.

Research focused on hearing health within the context of video gaming and esports, on the other hand, remains practically nonexistent, particularly concerning unsafe listening habits and their contributing factors [19,20]. Existing research from different fields, such as physical activity promotion [21], and related areas like safe listening using personal audio devices [13,14], as well as behavioral models such as the health belief model (HBM) [22], can be useful starting points to better understand the factors influencing behavior and the barriers that need to be addressed. However, we know that interventions aimed at changing health behavior must be tailored to the specific characteristics and preferences of the population being targeted. A factor that plays a significant role in one context might indeed not have the same impact in another. For example, self-efficacy may be a major determinant for behaviors requiring significant skill, such as learning new physical activities [23], but may be less relevant for behaviors like adjusting volume settings, which require minimal effort. Similarly, barriers to safe listening, such as the worry about not enjoying music in noisy environments, could be a major factor for users of personal audio devices outdoors [13] but not for gamers who play indoors and often in solitude. In the gaming context, other factors, such as immersion or competitive performance, may be more significant barriers. The large body of existing evidence from other fields may therefore not directly apply to the gaming and esports environments, as behavioral models used to understand determinants of behavior must be specifically adapted to this context. Only by understanding how these models apply to video gamers and esports participants, we can effectively tailor interventions and communication strategies.

Our study specifically aims to address this gap by (1) gaining insights into the listening habits and safe listening behaviors of video gamers and esports players, exploring the determinants of these behaviors; (2) understanding the link between attitudes, beliefs, awareness, and stages of change regarding auditory health; and (3) identifying the most effective sources of information that can be leveraged to promote safer listening practices. These findings will support the development of interventions that can be integrated into industry standards, contributing to the design of technological features that encourage safer listening within gaming and esports platforms. By aligning with the WHO Safe Listening initiative, our work aims to ensure that auditory health considerations are integrated into the rapidly evolving landscape of video gaming and esports, fostering a more holistic approach to hearing health across all recreational sound exposure settings.

Methods

Sample

From September 2022 to January 2023, an open, web-based survey was shared on the WHO's social media platforms,

XSL•FO
RenderX

including posts on X (formerly Twitter) and Facebook. These posts were designed to reach a broad convenience sample of video gamers and esports enthusiasts. Additionally, key stakeholders were contacted via email and encouraged to share the survey link within their local networks to increase participation. No paid advertising or boosted posts were used for recruitment. Participation in the survey was completely voluntary, and no incentives or compensation were offered to participants. To participate in the survey, individuals had to have a good command of English, Spanish, French, or Chinese and be either video game players or involved in esports. No further specific eligibility criteria for access were specified, and participants could freely choose to participate. Of the 1043 individuals who submitted the first survey page, 532 (51%) completed the survey and were retained to be included in the analyses. To ensure that participants responded to the survey with adequate attention, response time was evaluated, and 35 (6.7%) individuals who completed the survey in less than 5 minutes were excluded, resulting in a total of 488 responses retained for the analyses.

Instrument and Measures

Overview

To develop the web-based questionnaire, a combination of established theories and empirical studies was used. Specifically, the HMB [22], the transtheoretical model (TTM) of change [24], and previous research on music- and noise-induced hearing loss were drawn upon [25]. The HBM suggests that individuals are more likely to adopt preventive measures if they perceive themselves to be at risk of developing a severe health condition (perceived susceptibility and severity), believe that taking action can prevent it (perceived benefits), and have the confidence to perform the necessary actions (self-efficacy). The TTM suggests that individuals can be categorized into different groups based on their readiness to adopt preventive behaviors.

The questionnaire (Multimedia Appendix 1) consisted of seven main sections designed to assess various aspects of participants' listening habits when gaming or engaging in esports. These sections included questions on (1) their gaming or esports behaviors and related listening habits (respondents were redirected to different sections of the survey depending on whether they played videogames, watched esports, participated in esports, or a combination), (2) the consequences of their listening behavior, (3) their knowledge and awareness of hearing loss, (4) their beliefs about hearing loss and safe listening, (5) their readiness to change their listening behavior, (6) their desire for information on hearing loss and safe listening and their preferred sources of information, and (7) their sociodemographic information.

The survey consisted of 10 pages, each containing 3 to 7 items. Respondents were required to answer basic and sociodemographic questions, which included a nonresponse option. All other items were optional and could be left unanswered. Completeness checks were performed after submission, highlighting mandatory items if incomplete. A "Back" button allowed respondents to review and modify their answers before final submission.

The survey was pretested with 15 individuals who had experience in video gaming or esports. Their feedback helped refine the questions for clarity and relevance, ensuring that the final version was suitable for the target population. The survey was hosted on the Qualtrics web-based platform, a widely used platform that ensures data security through end-to-end encryption and complies with General Data Protection Regulation guidelines to protect participant anonymity. Several accessibility features, such as adjustable text sizes and screen reader compatibility, were included to accommodate participants with disabilities.

To prevent multiple submissions from the same participant, cookies were used on the introduction page and were valid for 24 months. IP address checks were not used to identify duplicate entries. Data collected through the survey were stored in secure cloud storage provided by Qualtrics, with all responses anonymized, and participants assigned unique identifiers to ensure confidentiality. The data will be securely stored for 5 years and then deleted in compliance with data protection regulations. We adhered to the CHERRIES (Checklist for Reporting Results of Internet E-Surveys; Multimedia Appendix 2) for reporting the results.

Gaming and Esports Behaviors and Related Listening Habits

To assess the frequency of video game or esports viewing or playing and volume-level preferences, a series of single-item indicators were used. Participants were asked to rate the frequency of their gaming or esports activities, as well as the volume level they typically set on their devices, using a scale ranging from 1=minimum volume to 10=maximum volume. These questions were adapted from measures used in prior research on the topic, such as the Listening Habits Questionnaire (LHQ) [25].

Preventive Behaviors and Consequences of Listening

Participants were asked to indicate whether they take a break from sound every hour and whether they check the sound dose. The frequency of experiencing a ringing or fuzziness in the ears after playing video games or engaging in esports was measured on scales ranging from 1=never to 5=always. Participants were asked to rate how often they experienced these symptoms using questions from the LHQ [25].

Awareness and Knowledge

Participants were presented with 3 statements related to hearing loss in the context of gaming and esports, which were collaboratively developed with a team of international experts in the field of hearing. The statements included "listening to sounds above 85dB over a period of time can cause permanent damage to your hearing," "the amount of time you listen to a sound affects how much damage it will cause," and "listening to video games or esports sounds at loud listening levels may damage hearing." Participants were asked to indicate whether they believed each statement to be true or false and were scored accordingly. A knowledge score was then computed by summing the scores for all 3 statements, resulting in a range of 0 to 3. The average knowledge score was 2.77 (SD 0.61), with an α coefficient of .690.

XSL•FO
RenderX

Beliefs

An adapted version of a section of the LHQ [25] was used to assess the constructs of the HBM. The adaptation involved removing the distinction between quiet and loud environments after internal pretesting. Participants were then asked to rate their agreement on a 7-point scale (with higher scores indicating greater agreement) on a total of 22 statements about hearing loss susceptibility related to gaming and esports. These statements covered perceived susceptibility to hearing loss (4 items, α=.776), perceived severity of hearing loss (6 items, α=.865), benefits of preventing hearing loss (5 items, α=.830), barriers to preventing hearing loss (3 items, α=.638), and self-efficacy for taking preventative action (4 items, α=.845). The scores for each construct were obtained by averaging the individual scores on items within each subscale.

Readiness to Change

To assess readiness to change listening behaviors, a set of 12 statements was adapted from a survey by Rollnick et al [26] for excessive drinkers and a previous study on safe listening in the context of personal listening devices [13]. Each statement represented one aspect of the 3 stages of the TTM (precontemplation, contemplation, and action). Respondents answered each statement on a 5-point scale (ranging from 2=strongly disagree to +2=strongly agree), as per the authors' instructions. The subscales showed acceptable to good internal consistency (precontemplation: α=.626; contemplation: α=.740; and action: α=.784). Scores for items related to the same stage were added, resulting in a stage score ranging from –8 to +8. The respondent's stage of change was determined by the stage with the highest score. If 2 stages had equally high scores, the stage furthest along the change process was chosen (eg, if precontemplation and contemplation had the highest scores, contemplation was designated as the stage of change).

Communication Needs and Preferences

Participants were asked if they desired more information about hearing loss and safe listening, with response options of yes or no. Those who answered affirmatively were then asked to indicate their preferred channels for receiving this information on a 7-point scale, ranging from 1=definitely no to 7=definitely yes, which included options such as public health campaigns, gaming device interfaces, and social media. Additionally, all participants were asked to rate how much they trust in different sources of information about hearing loss and safe listening, such as health care professionals, governmental and international agencies, and family and friends, using a 7-point scale ranging from 1=not at all to 7=completely.

Sociodemographics

Participants were requested to provide information about their sex, age, the highest level of education completed, and country of residence. For analytical purposes, sex (male and female), age (16-35 and >35 years), and educational level (low and high) were dichotomized.

Data Analysis

The data were analyzed using SPSS Statistics (version 21.0; IBM Corp). Descriptive statistics, including basic frequencies, were calculated to provide an overview of the sample characteristics across all variables of interest, such as demographics, knowledge, and beliefs related to hearing health. To assess group differences in the descriptive analyses, we used chi-square tests for categorical variables and 1-way ANOVAs for continuous variables, supplemented by effect size measures such as Cramer *V* or φ for chi-square tests and partial eta squared (η_p^2) for ANOVAs. These tests and effect size measures allowed us to compare key differences between sex, age, and education subgroups within the sample, providing a more detailed understanding of both the statistical significance and the magnitude of group differences across the variables of interest. To assess the relationship between predictors (knowledge and beliefs about hearing health) and the outcome variable (readiness to change listening behaviors), we performed a multinomial logistic regression analysis. We chose this method because the outcome variable (readiness to change) is categorical with multiple levels, and it allowed us to evaluate how each predictor influenced the likelihood of participants being in different stages of readiness to change while controlling for other variables. The use of multinomial logistic regression was particularly suited to account for the complexity of the data and to handle multiple predictors influencing a nonbinary categorical outcome. All assumptions for the chi-square tests, ANOVAs, and multinomial logistic regression (such as normality, independence, and multicollinearity) were tested and met, ensuring the robustness of our findings.

Ethical Considerations

Ethics approval for research in Switzerland is governed by the Federal Act on Research Involving Human Beings (Human Research Act, HRA, SR 810.30 [27]). According to chapter 1, article 2, paragraph 2 of this act, ethics approval is required for research involving human participants. However, our study, which involves an anonymous web-based survey, does not fall within the scope of the HRA. Specifically, article 2, paragraph 2 states: "It [the HRA] does not apply to research which involves: a. IVF embryos in accordance with the Stem Cell Research Act of 19 December 2003; b. anonymized biological material; c. anonymously collected or anonymized health-related data." Therefore, our project does not require ethics approval (see also the Swissethics decision tree [28]). This web-based anonymous survey was conducted in strict adherence to the principles outlined in the Declaration of Helsinki. Informed consent was obtained from all participants prior to starting the survey. Participants were presented with a web-based informed consent form, which included information about the voluntary nature of the study, data anonymity, and their right to withdraw at any time without any consequences. Consent was confirmed through a mandatory checkbox before proceeding to the survey questions. No compensation was provided to study participants.

Results

Overview

Of the survey participants, the majority were male (n=355, 67.7%), with an average age of 28.4 (SD 10.8) years. The highest level of education attained by most respondents was a 4-year college degree (n=107, 21.9%) or a high school diploma

XSL•FO
RenderX

or General Educational Development (n=101, 20.7%). Although respondents hailed from 92 different countries, the largest number of respondents came from the United States (n=70, 14.3%), followed by the United Kingdom (n=51, 10.5%) and India (n=40, 8.2%). More details about the participants are shown in Table 1.

Table 1. Sample characteristics (N=488).

	Values, n (%)
Sex	
Male	328 (67.2)
Female	144 (29.5)
No answer	16 (3.3)
Age (years)	
16-35	386 (79.1)
>35	102 (20.9)
Education	
Low education (less than college)	196 (40.2)
Middle education (college)	133 (27.3)
High education (more than college)	142 (29.1)
No answer	17 (3.5)
Country of origin (top 5)	
United States	70 (14.3)
United Kingdom	51 (10.5)
India	40 (8.2)
Belgium	34 (7)
Australia	21 (4.3)
Involved in (more than 1 answer possible)	
Video games	443 (90.8)
Esports viewing	268 (54.9)
Esports participation	68 (13.9)

Gaming and Esports Behaviors and Related Listening Habits

Details of the overall gaming and esports behaviors and the related listening habits are shown in Table 2 (video game players) and Table 3 (esports viewers and esports players).

XSL•FO
RenderX

Table 2. Behaviors and related listening habits of video game players (n=443).

	Values
Time spent playing (hours per week), n (%)	
0-1	29 (6.5)
1-2	30 (6.8)
3-4	50 (11.3)
4-6	49 (11.1)
6-10	73 (16.5)
10-15	74 (16.7)
15-30	82 (18.5)
>30	56 (12.6)
Volume setting, mean (SD) (scale 1-100)	47.6 (21.3)
Volume setting, n (%)	
Very low (0%-20%)	49 (11.1)
Low (20%-40%)	134 (30.2)
Medium (40%-60%)	150 (33.9)
High (60%-80%)	85 (19.2)
Very high (80%-100%)	25 (5.6)
Importance of sounds, n (%)	
Not at all important	14 (3.2)
Not so important	25 (5.6)
Somewhat important	87 (19.6)
Very important	151 (34.1)
Extremely important	166 (37.5)
Sound break every hour, n (%)	
Yes	152 (34.3)
Checks use statistics, n (%)	
Yes	45 (10.2)
Experience ringing in ears, n (%)	
Never	253 (57.1)
Rarely	125 (28.2)
Sometimes	51 (11.5)
Often	6 (1.4)
Always	8 (1.8)
Experience fuzziness in ears, n (%)	
Never	206 (46.5)
Rarely	148 (33.4)
Sometimes	71 (16)
Often	11 (2.5)
Always	7 (1.6)

XSL•FO
RenderX

Table 3. Behaviors and related listening habits of esports viewers and players.

	Viewers (n=287)	Players (n=62)
Frequency of viewing, n (%)		
Daily	14 (5.2)	15 (24.2)
Weekly	75 (28)	23 (37.1)
Monthly	79 (29.5)	11 (17.7)
Yearly	68 (25.4)	6 (9.7)
Less than yearly	32 (11.9)	7 (11.3)
Volume setting, mean (SD) (scale 1-100)	45.6 (20.4)	55.7 (22.5)
Volume setting, n (%)		
Very low (0%-20%)	34 (12.7)	5 (8.1)
Low (20%-40%)	80 (29.9)	10 (16.1)
Medium (40%-60%)	105 (39.2)	24 (38.7)
High (60%-80%)	36 (13.4)	14 (22.6)
Very high (80%-100%)	13 (4.9)	9 (14.5)
Experience ringing in ears, n (%)		
Never	186 (69.4)	36 (53.7)
Rarely	46 (17.2)	20 (29.9)
Sometimes	29 (10.8)	6 (9)
Often	2 (0.7)	0 (0)
Always	5 (1.9)	5 (7.5)
Experience fuzziness in ears, n (%)		
Never	187 (69.8)	33 (49.3)
Rarely	49 (18.3)	18 (26.9)
Sometimes	23 (8.6)	8 (11.9)
Often	4 (1.5)	3 (4.5)
Always	5 (1.9)	5 (7.5)
Importance of sounds, n (%)		
Not at all important	10 (3.7)	1 (1.5)
Not so important	33 (12.3)	8 (11.8)
Somewhat important	79 (29.5)	11 (16.2)
Very important	80 (29.9)	20 (29.4)
Extremely important	66 (24.6)	28 (41.2)
Break from sound every hour, n (%)		
Yes	96 (35.8)	27 (39.7)
Checks use statistics, n (%)		
Yes	37 (13.8)	12 (17.6)

Video Game Players

Of the surveyed participants, 90.8% (n=443) reported engaging in video games (Table 2). Around one-third of gamers (n=158, 35.7%) reported playing for up to 6 hours per week, followed by a second group (n=147, 33.2%) who played between 6 and 15 hours per week, and a third group (n=138, 31.1%) who played more than 15 hours per week. Male respondents reported significantly more time spent on video gaming compared to female respondents (P=.008; φ=0.210), while those with a higher educational level spent significantly less time playing video games compared to those in the lower and medium educational groups (P=.03; Cramer V=0.175). The average volume setting reported by participants was just below 50% (mean 47.6, SD 21.3), with almost 1 in 4 respondents listening at a high or very high volume (n=110, 24.8%). No significant differences were

XSL•FO
RenderX

found across sexes, ages, or educational levels. Finally, most respondents indicated that sounds are either extremely (n=166, 37.5%) or very (n=151, 34.1%) important when playing video games. Male respondents, on average, placed greater importance on video game sounds than female respondents (P=.03; Cramer V=0.161).

Esports Viewers and Players

More than half of the study participants (n=281, 57.6%) reported engaging in esports, with most of them only viewing (n=213, 75.8%) and approximately one-fifth (n=55, 19.6%) both viewing and participating (Table 3). A small percentage (n=13, 4.6%) reported only participation. Among viewers (n=268), the majority viewed on a monthly basis (n=79, 29.5%). Male participants were found to watch esports events more often than female participants (P=.03; Cramer V=0.202). The average volume setting reported was slightly below 50% (mean 45.6, SD 20.4), with almost 1 in 5 respondents listening at a high or very high volume (n=49, 18.3%). Participants with lower educational levels reported significantly higher volume settings when watching esports events ($F_{2,254}$=3.139; P=.04), while no significant differences were observed across sex or age groups.

Among esports participants (n=68), the majority took part in esports events on a weekly basis (n=23, 37.1%). No significant differences were observed regarding sex, age, or education. The average volume setting for esports participation was higher than for viewing at 55.7 (SD 22.5), with more than 1 in 5 respondents listening at a high or very high volume (n=23, 21%). No significant differences were observed across sexes, ages, educational levels, or frequency of participation.

Overall, esports viewers and players consider sounds in esports as either somewhat (n=82, 29.2%), very (n=85, 30.2%), or extremely (n=68, 24.2%) important. Among esports players, in particular, sound was considered extremely important by a majority (n=28, 41.2%). Participants with lower educational levels placed significantly greater importance on sounds in esports events compared to those with higher educational levels (P=.02; Cramer V=0.183).

Protective Behavior and Consequences of Listening

Only about one-third of video game players (n=152, 34.3%) reported taking a break from sound every hour. Just 1 in 10 (n=45, 10.2%) reported checking the sound dosage information provided by the gaming device. Male and female respondents differed significantly in whether they checked information on how much time they spend listening and their sound dosages through their gaming device (P=.03; Cramer V=0.144). A similar picture emerged for esports players and participants. Over one-third of them (n=106, 37.7%) reported taking an hourly break, and only a few reported checking the sound dosage (n=43, 15.3%). No group differences were observed.

Regarding the consequences of listening, almost half of the video game players reported having experienced at least once a ringing in their ears after playing (n=190, 42.9%), and more than half have experienced fullness or fuzziness in their ears (n=237, 53.5%). Female respondents reported feeling their ears full or fuzzy after playing video games more often than male respondents (P=.006; Cramer V=0.184). Regarding esports, we observed 2 different pictures for viewers and players. Among esports viewers, less than one-third reported having experienced a ringing in their ears (n=82, 30.6%) or a feeling of fullness or fuzziness (n=81, 30.2%) after watching esports. Among esports players, however, this percentage increased to 46.3% (n=31) for the ringing in their ears and to 50.7% (n=34) for the feeling of fullness or fuzziness. Participants' characteristics were not associated with the frequency of experiencing consequences.

Awareness and Knowledge

Overall, the respondents were very knowledgeable about the risks for hearing related to exposure to video games, scoring on average 2.78 (SD 0.62) of a maximum of 3 points indicating a correct answer to all questions. No significant differences among groups were observed.

Beliefs

The majority of participants rated the perceived severity of hearing loss as high (mean 5.84, SD 1.31), while their perceived susceptibility to it was relatively low (mean 3.70, SD 1.54). Participants recognized the benefits of preventing hearing loss (mean 5.84, SD 1.22) but also faced significant barriers to doing so (mean 4.53, SD 1.56), such as the belief that turning down the volume would limit their enjoyment of video games and esports. Despite this, respondents demonstrated high confidence in their ability to modify their listening behavior (mean 5.24, SD 1.57). We observed few group differences in beliefs, except for higher self-efficacy among male respondents ($F_{1,470}$=4.13; P=.04; η_p^2=0.009) and lower perceived susceptibility among younger respondents ($F_{1,486}$=12.33; P<.001; η_p^2=0.025). In addition, respondents with higher education reported perceiving more benefits of prevention compared to the other 2 groups ($F_{2,468}$=3.65; P=.03; η_p^2=0.015).

Readiness to Change

Overview

We observed a balanced distribution of participants into the 3 readiness-to-change groups. Most of them (n=173, 35.3%) could be categorized as belonging to the "contemplation" stage, followed by the "action" stage (n=167, 34.2%) and the "precontemplation" stage (n=148, 30.3%). A chi-square test revealed a significant association between educational level and stage of change (P=.009; Cramer V=0.120), with individuals in the high education group more likely to be in the action stage (n=60, 42.3%) compared to those in the medium (n=50, 37.6%) and low education groups (n=49, 25%). Table 4 shows the multinomial logistic regression results for factors associated with being in one of the 3 readiness-to-change stages. Those who were in the contemplation stage were the reference category.

XSL•FO
RenderX

Table 4. Associations of knowledge and beliefs with readiness to change (N=488)[a].

	Readiness to change			
	Precontemplation		Action	
	OR[b] (95% CI)	*P* value	OR (95% CI)	*P* value
Knowledge	1.293 (0.769-0.174)	.33	0.998 (0.628-1.584)	.99
Perceived susceptibility	0.557 (0.452-0.687)	<.001	1.031 (0.851-1.249)	.76
Perceived severity	1.344 (1.056-1.712)	.02	1.082 (0.868-1.349)	.48
Benefits of preventive action	0.611 (0.461-0.809)	.001	0.968 (0.732-1.281)	.82
Barriers to preventive action	0.985 (0.818-1.186)	.54	0.808 (0.680-0.959)	.01
Self-efficacy in prevention	1.518 (1.237-1.863)	<.001	1.431 (1.196-1.711)	<.001

[a]Multinomial logistic regression analysis with contemplation as the reference category. The model is adjusted for sex, age, and education. Pseudo R^2: Cox and Snell=0.274; Nagelkerke=0.308.

[b]OR: odds ratio.

Multivariate Results: Contemplation Versus Precontemplation Stage

The second column of Table 4 shows the odds ratio (OR) for being in the precontemplation stage versus being in the contemplation stage for each variable listed. The odds of being in the precontemplation versus the contemplation group decrease the more one perceives to be susceptible to hearing loss (OR 0.557, 95% CI 0.452-0.687) and the more one perceives the benefits of preventive actions (OR 0.611, 95% CI 0.461-0.809). On the other hand, the odds of being in the precontemplation group increase the more one perceives hearing loss as severe (OR 1.344, 95% CI 1.056-1.712) and the more one has self-efficacy in the prevention of hearing loss (OR 1.518, 95% CI 1.237-1.863).

Multivariate Results: Contemplation Versus Action Stage

The fourth column of Table 4 shows the OR for being in the action stage versus being in the contemplation stage for each variable listed. The odds of being in the action versus the contemplation group decrease the more barriers to preventive action for hearing loss one perceives (OR 0.808, 95% CI 0.680-0.959) and increase the more one has self-efficacy in the prevention of hearing loss (OR 1.431, 95% CI 1.196-1.711).

Communication Needs and Preferences

More than half of the participants (n=249, 51%) expressed an interest in expanding their knowledge on the topics of hearing loss and safe listening in the context of video gaming and esports. While investigating the preferred channels for obtaining such information, it was found that no single mode was unequivocally favored. Nonetheless, noteworthy options included instruction manuals for gaming devices, dedicated websites, and user interfaces integrated into gaming devices. No significant differences were observed in preferences for sex, age, or education. Regarding trusted sources of information, respondents showed a clear preference for health care professionals and governmental or international agencies (eg, WHO), while traditional media and religious leaders were among the least trusted sources. Female respondents reported significantly higher trust in government or international health agencies (eg, WHO) compared to male respondents ($F_{1,470}$=8.76; P=.003; η_p^2=0.018). Younger respondents reported higher trust in government or international health agencies ($F_{1,486}$=4.98; P=.03; η_p^2=0.010) as well as higher trust in producers of gaming or esports devices ($F_{1,486}$=15.12; P<.001; η_p^2=0.030) and doctors or pharmacists ($F_{1,486}$=12.59; P<.001; η_p^2=0.025). Significant differences in trust were observed across educational levels, with respondents holding lower education reporting higher trust in producers of gaming or esports devices ($F_{2,468}$=4.88; P=.008; η_p^2=0.020) as well as higher trust in newspapers or magazines ($F_{2,468}$=4.00; P=.02; η_p^2=0.017). More details about communication needs and preferences can be found in Table 5.

XSL•FO
RenderX

Table 5. Communication needs and preferences (N=488).

	Values
Interest in receiving more information, n (%)	
Yes	249 (51)
Preferred channels[a], mean (SD)	
Instructions of gaming device	5.91 (1.58)
User interface of gaming device	5.84 (1.79)
Dedicated website	5.82 (1.65)
Social media	5.70 (1.78)
Mass media	5.43 (1.89)
Public health campaign	5.29 (1.95)
Written information	4.90 (2.14)
Interpersonal communication	4.86 (2.11)
Public events	4.20 (2.31)
Trust in sources of information[b], mean (SD)	
Health care professionals	6.22 (1.32)
Governmental or international agencies	6.01 (1.46)
Charitable organizations	4.65 (1.76)
Producers of gaming devices	4.48 (1.90)
Family or friends	4.21 (1.66)
Internet	4.19 (1.65)
Radio	3.87 (1.80)
Newspapers or magazines	3.85 (1.80)
Television	3.78 (1.75)
Religious organizations or leaders	2.74 (1.98)

[a]Scores ranging from 1 to 7, with higher scores indicating a higher willingness to receive information from the source.

[b]Scores ranging from 1 to 7, with higher scores indicating higher trust in the source.

Discussion

Overview

In this study, we aimed to explore the gaming and esports behaviors, related listening habits, and awareness about hearing health risks among gamers and esports participants. This investigation was motivated by the increasing concerns over potential hearing damage due to prolonged exposure to high sound levels in video gaming environments and the need for designing and implementing specifically designed features that make video games safer.

Principal Findings

Our first specific aim was to gain insights into the listening habits and safe listening behaviors of video gamers and esports players, exploring the determinants of these behaviors. Our results show that an important part of players set the volume at high or very high levels, with younger participants reporting significantly higher volume settings than older respondents. Furthermore, we observed that most viewers and players consider sounds in esports to be (very) important, and the volume level for esports participation was higher than for viewing. Although most respondents were aware of the risks of loud sound exposure and recognized the benefits of prevention, they still engaged in risky listening behaviors. For instance, only about one-third of video game and esports players reported taking a break from sound every hour, and only a few reported checking the sound dosage. This lack of protective behavior is particularly concerning, given that around half of the video game and esports players reported having experienced at least once a ringing or feelings of fullness or fuzziness in their ears after playing. These findings suggest that many video gamers and esports players are experiencing symptoms of hearing damage, yet are not taking the necessary steps to protect their hearing.

Second, we aimed to understand the link between attitudes, beliefs, awareness, and stages of change regarding auditory health. Applying the constructs of the HBM, we were able to determine that, despite an overall recognition of the risks associated with loud sound exposure, several factors get in the way of a positive behavior change. For instance, many believed that turning down the volume would limit their enjoyment of

XSL•FO
RenderX

video games and esports. In addition, while many rated the perceived severity of hearing loss as high, their perceived susceptibility to it was relatively low. Finally, low self-efficacy also seems to play a role in the decision to adopt safer listening practices. These results are in line with what is suggested by current theories of behavior change and evidence in the field of health behaviors and of safe listening in other contexts, such as findings in physical activity promotion, where individuals are often aware of the benefits of exercise but are deterred by perceived barriers like time constraints or lack of motivation [29]. Similarly, in safe listening research for personal audio devices, users often understand the risk of hearing damage but continue to listen at high volumes, believing that lowering the volume would reduce sound quality or enjoyment, especially in noisy environments [13]. Concretely, these findings suggest a need for education programs specifically designed to show how prevention does not require particular skills and does not necessarily limit enjoyment (eg, through testimonials [30]) and to increase the perceived susceptibility [31] to hearing loss among video gamers and esports players.

We also observed a balanced distribution of participants into the 3 readiness-to-change groups, as defined by the TTM [24], with most of them belonging to the "contemplation" stage, followed by the "action" stage and the "precontemplation" stage. While many video gamers and esports players are aware of the risks of loud sound exposure and are considering taking action, they may need additional support to move from contemplation to action. The multivariate analysis revealed several factors that were associated with participants' readiness to change their listening behavior in the context of video gaming and esports. Participants who perceived themselves as more susceptible to hearing loss and recognized the benefits of preventive actions were less likely to be in the precontemplation stage and more likely to be in the contemplation stage. On the other hand, participants who perceived hearing loss as severe and had more self-efficacy in preventing hearing loss were more likely to be in the precontemplation stage. Participants who had more self-efficacy in preventing hearing loss were more likely to be in the action stage than in the contemplation stage. Finally, participants who perceived more barriers to preventive actions were less likely to be in the action stage. These findings suggest that interventions aimed at promoting safe listening habits among video gamers and esports viewers and players should be tailored to participants' beliefs and attitudes about hearing loss and preventive actions [32]. In particular, our findings suggest that interventions to increase self-efficacy [33] may be effective in promoting safe listening habits among video gamers and esports players independent of the stage they are in. At the same time, interventions for those in the precontemplation stage should focus on increasing the perceived susceptibility and on showing the benefits of prevention, while those targeted at those in the contemplation stage should include information aimed at reducing barriers, for instance, by showing how listening to lower volumes does not reduce enjoyment.

Finally, we aimed to identify the most effective sources of information that can be leveraged to promote safer listening practices. Approximately half of the participants expressed a keen interest in expanding their knowledge on hearing loss and safe listening in the context of video gaming and esports, suggesting a good degree of openness to information. Regarding trusted sources of information, respondents showed a clear preference for health care professionals and governmental or international agencies. Female respondents showed significantly more trust in governmental or international agencies, while younger respondents showed more trust in health care professionals, governmental or international agencies, and producers of gaming devices. This information is important, as it can inform future communication efforts especially tailored to video gamers and esports players.

Overall, our findings not only highlight the importance of promoting safe listening behaviors among video gamers and esports players but also contribute to the broader literature on behavior change. By identifying key barriers and beliefs unique to this population—such as prioritizing immersion and performance over auditory health—our research offers new insights into how behavior change models, like the HBM, can be tailored to different contexts. This contributes to a more nuanced understanding of how interventions should be tailored to address context-specific motivations and barriers. In doing so, our study advances the development of targeted interventions aimed at reducing unsafe listening habits, with the potential to inform both gaming platform design and broader behavior change strategies. These findings are significant for the ongoing development of health promotion strategies that encourage long-term behavioral change, reinforcing the importance of adapting existing models to the specific needs of the different communities.

Practice Implications

Our findings support the need for the design and implementation of safe listening features in video games and also provide useful insights into how these features could look like. First, the substantial engagement in video gaming and the pronounced importance of sound highlight the necessity for interventions that do not compromise sound quality. Features enabling the tracking of the sound dose, such as dosimetry [34] or dynamic range compression [35], have emerged as crucial innovations, which could ensure that gamers can remain engaged and fully immersed without risking their hearing health.

Second, the observation of gamers frequently listening at high volumes and not engaging in protective practices, such as regular sound breaks, motivates the introduction of features enabling an automatic reduction of volume [36] and in-game notifications [37] related to measured sound exposure. These interventions would directly encourage gamers to adjust their volume settings and take necessary breaks, promoting safer listening habits organically within the gaming environment.

Third, the reported adverse effects, including ringing in the ears and feelings of fullness, highlight the need for features allowing users to remove particularly harmful sounds [38] and, considering the detrimental impact of tinnitus in terms of psychological or emotional effects, sleep disturbance, auditory issues, and overall health [39], support the introduction of general safe listening warnings in video games. These features aim to provide a more comfortable gaming experience for individuals experiencing tinnitus, reduce the experiences of

XSL•FO
RenderX

tinnitus, and raise awareness about the potential adverse effects of unsafe listening practices.

Fourth, despite participants being somewhat knowledgeable about the risks of hearing damage, the application of this knowledge was limited, indicating a gap between awareness and protective action. This gap underscores the importance of user guides, detailing what can be done at different levels, and adaptive headphone safety modes [40,41] that automatically optimize listening settings for safety when switching from a loudspeaker to headphones, directly addressing the need to bridge knowledge with actionable safe listening practices.

Fifth, the significant interest among gamers in expanding their knowledge about hearing health and safe listening practices supports the creation of a safer listening mode. Besides directly contributing to protecting users' hearing, the very existence of such a mode could educate them on safer listening practices, addressing the community's desire for information and tools to protect their hearing health.

Finally, our findings show that gamers often face barriers to adopting safer listening practices due to concerns about compromising their gaming experience. This suggests a need for customizable features that balance individual preferences with hearing safety. Features like sound category controls [42,43] and dynamic range tests [44], tailored to specific gaming contexts, could help ensure safe listening without detracting from the immersive experience.

Limitations

This study has several limitations to consider when interpreting the results. First, relying on self-reported data may introduce bias and may not accurately reflect actual behaviors. Second, the cross-sectional nature of the study prevents causal inferences. Third, the lack of assessment of noise-canceling headphone use, which could mitigate hearing damage, limits the findings. Finally, the open web-based survey recruitment strategy, shared via social media and stakeholder networks, may have led to selection bias, as those more interested in hearing health or gaming were more likely to participate. Despite this, the consistency of our findings with previous studies on related topics [13,14] supports their validity, allowing us to draw meaningful conclusions. Future research should address these limitations and evaluate the effectiveness of educational interventions in promoting safe listening habits among video gamers and esports players.

Conclusions

This study underscores the urgency of developing interventions to foster safe listening habits in video gamers and esports enthusiasts, pointing out their general lack of awareness and preventive actions against hearing damage. Besides stressing the need for effective communication strategies tailored to this group's specific preferences, the findings support creating targeted interventions to encourage safe listening and minimize hearing loss risk among these audiences. Incorporating safe listening features directly into video games could offer a strategic method to address prolonged exposure to high sound levels without detracting from the gaming experience. These personalized, gamer-friendly solutions could significantly boost auditory health awareness and practices. The insights into gamer behavior and preferences provided by this study not only enrich discussions on digital health interventions but also pave the way for enhancing hearing health in the digital entertainment sector, contributing to the broader goal of leveraging technology to improve health outcomes.

Acknowledgments

The authors would like to express their sincere gratitude to all the participants who took the time to complete the survey, without whom this study would not have been possible. The authors also extend their appreciation to the organizations who helped with the recruitment of study participants. The authors are thankful for their support and cooperation in disseminating information about the study to their members and networks. This work was supported by the World Health Organization, Geneva, Switzerland. The views reported are solely those of the authors and do not necessarily represent the views of the World Health Organization. The authors used ChatGPT by OpenAI to improve the readability and language of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

All authors contributed to the conceptualization of the study and the investigation. ND and SR developed the study methodology, with ND additionally overseeing recruitment, data collection, and formal analysis. ND also took on project administration and supervision responsibilities. ND and SR wrote the original draft of the manuscript, while all authors participated in reviewing and editing. Funding acquisition was managed by SC.

Conflicts of Interest

None declared.

Multimedia Appendix 1

XSL•FO
RenderX

Survey questionnaire.
[PDF File (Adobe PDF File), 234 KB - formative_v9i1e60476_app1.pdf]

Multimedia Appendix 2
Checklist for Reporting Results of Internet E-Surveys (CHERRIES).
[PDF File (Adobe PDF File), 52 KB - formative_v9i1e60476_app2.pdf]

References

1. Franks RR, King D, Bodine W, Chisari E, Heller A, Jamal F, et al. AOASM position statement on esports, active video gaming, and the role of the sports medicine physician. Clin J Sport Med 2022;32(3):e221-e229 [FREE Full text] [doi: 10.1097/JSM.0000000000001034] [Medline: 35470342]
2. Global video games market report and forecast 2023-2028. Claight Corp. URL: https://www.expertmarketresearch.com/reports/video-games-market [accessed 2025-03-03]
3. Games - Worldwide | Statista Market Forecast. Statista. URL: https://www.statista.com/outlook/dmo/digital-media/video-games/worldwide [accessed 2025-03-03]
4. Gilbert N. 75 Significant esports statistics: 2024 market, viewership & investment growth. FinancesOnline. 2023. URL: https://financesonline.com/esports-statistics/ [accessed 2025-03-03]
5. Esports market size, share, trends and analysis by region, revenue stream, gaming genre and segment forecast to 2030. Mark Res Rep Consult Glob UK Ltd. URL: https://www.globaldata.com/store/report/esports-market-analysis/ [accessed 2025-03-03]
6. Bossen D, Broekema A, Visser B, Brons A, Timmerman A, van Etten-Jamaludin F, et al. Effectiveness of serious games to increase physical activity in children with a chronic disease: systematic review with meta-analysis. J Med Internet Res 2020;22(4):e14549 [FREE Full text] [doi: 10.2196/14549] [Medline: 32234697]
7. Gentry SV, Gauthier A, L'Estrade Ehrstrom B, Wortley D, Lilienthal A, Tudor Car L, et al. Serious gaming and gamification education in health professions: systematic review. J Med Internet Res 2019;21(3):e12994 [FREE Full text] [doi: 10.2196/12994] [Medline: 30920375]
8. Kim Y, Hong S, Choi M. Effects of serious games on depression in older adults: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2022;24(9):e37753 [FREE Full text] [doi: 10.2196/37753] [Medline: 36066964]
9. Kelly S, Leung J. The new frontier of esports and gaming: a scoping meta-review of health impacts and research agenda. Front Sports Act Living 2021;3:640362 [FREE Full text] [doi: 10.3389/fspor.2021.640362] [Medline: 33796849]
10. Chan G, Huo Y, Kelly S, Leung J, Tisdale C, Gullo M. The impact of eSports and online video gaming on lifestyle behaviours in youth: a systematic review. Comput Hum Behav 2022;126:106974. [doi: 10.1016/j.chb.2021.106974]
11. Yin K, Zi Y, Zhuang W, Gao Y, Tong Y, Song L, et al. Linking esports to health risks and benefits: current knowledge and future research needs. J Sport Health Sci 2020;9(6):485-488 [FREE Full text] [doi: 10.1016/j.jshs.2020.04.006] [Medline: 32417192]
12. Hearing loss due to recreational exposure to loud sounds: a review. World Health Organization. 2015. URL: https://iris.who.int/handle/10665/154589 [accessed 2025-02-22]
13. Diviani N, Zanini C, Amann J, Chadha S, Cieza A, Rubinelli S. Awareness, attitudes, and beliefs about music-induced hearing loss: towards the development of a health communication strategy to promote safe listening. Patient Educ Couns 2019;102(8):1506-1512. [doi: 10.1016/j.pec.2019.03.013] [Medline: 30935699]
14. Diviani N, Chadha S, Arunda M, Rubinelli S. Attitudes towards safe listening measures in entertainment venues: results from an international survey among young venue-goers. Int J Environ Res Public Health 2021;18(23):12860 [FREE Full text] [doi: 10.3390/ijerph182312860] [Medline: 34886583]
15. Portnuff CD. Reducing the risk of music-induced hearing loss from overuse of portable listening devices: understanding the problems and establishing strategies for improving awareness in adolescents. Adolesc Health Med Ther 2016;7:27-35 [FREE Full text] [doi: 10.2147/AHMT.S74103] [Medline: 26929674]
16. Jiang W, Zhao F, Guderley N, Manchaiah V. Daily music exposure dose and hearing problems using personal listening devices in adolescents and young adults: a systematic review. Int J Audiol 2016;55(4):197-205. [doi: 10.3109/14992027.2015.1122237] [Medline: 26768911]
17. Guidelines for safe listening devices/systems (Recommendation ITU-T H.870). International Telecommunication Union. 2018. URL: https://www.itu.int/rec/T-REC-H.870-201808-I/en [accessed 2018-12-06]
18. WHO global standard for safe listening venues and events. World Health Organization. 2022. URL: https://www.who.int/publications/i/item/9789240043114 [accessed 2025-02-22]
19. Iannace G, Ciaburro G, Trematerra A. Video games noise exposure in teenagers and young adults. Noise Vib Worldw 2019;51(1-2):3-11. [doi: 10.1177/0957456519889956]
20. Dillard LK, Mulas P, Der C, Fu X, Chadha S. Risk of sound-induced hearing loss from exposure to video gaming or esports: a systematic scoping review. BMC Public Health 2024;2(1):e000253. [doi: 10.1136/bmjph-2023-000253]

XSL•FO
RenderX

21. Müller-Riemenschneider F, Reinhold T, Nocon M, Willich S. Long-term effectiveness of interventions promoting physical activity: a systematic review. Prev Med 2008;47(4):354-368. [doi: 10.1016/j.ypmed.2008.07.006] [Medline: 18675845]
22. Rosenstock IM. The health belief model and preventive health behavior. Health Educ Monogr 1974;2(4):354-386. [doi: 10.1177/109019817400200405]
23. Pauline J. Physical activity behaviors, motivation, and self-efficacy among college students. Coll Stud J 2013;47(1):64-74 [FREE Full text]
24. DiClemente CC, Prochaska JO. Toward a comprehensive, transtheoretical model of change: stages of change and addictive behaviors. In: Treating Addictive Behaviors. 2nd Edition. New York, NY: Plenum Press; 1998:3-24.
25. Portnuff CDF, Fligor BJ, Arehart KH. Teenage use of portable listening devices: a hazard to hearing? J Am Acad Audiol 2011;22(10):663-677. [doi: 10.3766/jaaa.22.10.5] [Medline: 22212766]
26. Rollnick S, Heather N, Gold R, Hall W. Development of a short 'readiness to change' questionnaire for use in brief, opportunistic interventions among excessive drinkers. Br J Addict 1992;87(5):743-754. [doi: 10.1111/j.1360-0443.1992.tb02720.x] [Medline: 1591525]
27. Federal Act on Research Involving Human Beings (Human Research Act, HRA, SR 810.30). URL: https://www.fedlex.admin.ch/eli/cc/2013/617/en [accessed 2024-10-21]
28. Swissethics Decision Tree. URL: https://swissethics.ch/assets/other_study_documents/decision-tree_v2.1_06.02.25_en.pdf [accessed 2025-03-03]
29. Koh YS, Asharani PV, Devi F, Roystonn K, Wang P, Vaingankar JA, et al. A cross-sectional study on the perceived barriers to physical activity and their associations with domain-specific physical activity and sedentary behaviour. BMC Public Health 2022;22(1):1051 [FREE Full text] [doi: 10.1186/s12889-022-13431-2] [Medline: 35614402]
30. Miller-Day M, Hecht ML. Narrative means to preventative ends: a narrative engagement framework for designing prevention interventions. Health Commun 2013;28(7):657-670 [FREE Full text] [doi: 10.1080/10410236.2012.762861] [Medline: 23980613]
31. Greene K, Brinn LS. Messages influencing college women's tanning bed use: statistical versus narrative evidence format and a self-assessment to increase perceived susceptibility. J Health Commun 2003;8(5):443-461. [doi: 10.1080/713852118] [Medline: 14530147]
32. Noar SM, Harrington NG, Aldrich RS. The role of message tailoring in the development of persuasive health communication messages. Ann Int Commun Assoc 2016;33(1):73-133 [FREE Full text] [doi: 10.1080/23808985.2009.11679085]
33. van de Laar KE, van der Bijl JJ. Strategies enhancing self-efficacy. In: Self-Efficacy in Nursing: Research and Measurement Perspectives. New York, NY: Springer Publishing Company; 2002:63.
34. Pietrzak A. Musicians' daily sound exposure assessed by full-day dosimetry. Vib Phys Syst 2019;30(1):1-8 [FREE Full text]
35. Kates JM. Principles of digital dynamic-range compression. Trends Amplif 2005;9(2):45-76 [FREE Full text] [doi: 10.1177/108471380500900202] [Medline: 16012704]
36. Kim H, Ryu D, Park J. Implementation of a volume controller for considering hearing loss in Bluetooth headset. In: Kim T, Adeli H, Robles RJ, Balitanas M, editors. Advanced Communication and Networking. Berlin, Heidelberg: Springer; 2011:430-438.
37. Pretolesi D, Motnikar L, Till B, Uhl J. Notifying users: customisation preferences for notifications in health and well-being applications. In: Meschtscherjakov A, Midden C, Ham J, editors. Persuasive Technology. Cham: Springer Nature Switzerland; 2023:410-424.
38. Horizon Forbidden West: what are tinnitus sounds. Game Rant. URL: https://gamerant.com/horizon-forbidden-west-tinnitus-sounds-option/ [accessed 2024-05-13]
39. Swain SK. Impact of tinnitus on quality of life: a review. Int J Adv Med 2021;8(7):1006. [doi: 10.18203/2349-3933.ijam20212416]
40. How to raise or lower the maximum headphone volume. Nintendo Eur AG. URL: https://www.nintendo.com/en-gb/Support/Nintendo-Switch/How-to-Raise-or-Lower-the-Maximum-Headphone-Volume-1516285.html [accessed 2024-05-13]
41. Adaptive audio for mobile. Game Developer. URL: https://www.gamedeveloper.com/audio/adaptive-audio-for-mobile [accessed 2024-05-13]
42. Marks A. The Complete Guide to Game Audio: For Composers, Musicians, Sound Designers, Game Developers. New York, NY: Routledge; 2012.
43. Audio Settings | Game Dev Handbook. URL: https://gamedev.catalystsoftworks.com/ui/settings/audio/ [accessed 2024-05-13]
44. Audio dynamic range calibration system. Martin Bussy-Pâris. 2016. URL: https://martinbussy.com/audio-dynamic-range-calibration-system/ [accessed 2024-05-13]

Abbreviations

CHERRIES: Checklist for Reporting Results of Internet E-Surveys
HBM: health belief model
HRA: Human Research Act

XSL•FO
RenderX

LHQ: Listening Habits Questionnaire
OR: odds ratio
TTM: transtheoretical model
WHO: World Health Organization

Edited by A Mavragani; submitted 13.05.24; peer-reviewed by RE Ferdig, K Song; comments to author 06.09.24; revised version received 27.09.24; accepted 28.10.24; published 25.03.25.

Please cite as:
Diviani N, Chadha S, Mulas P, Rubinelli S
Safe Listening Beliefs, Attitudes, and Practices Among Gamers and Esports Participants: International Web-Based Survey
JMIR Form Res 2025;9:e60476
URL: https://formative.jmir.org/2025/1/e60476
doi:10.2196/60476
PMID:40131338

©Nicola Diviani, Shelly Chadha, Peter Mulas, Sara Rubinelli. Originally published in JMIR Formative Research (https://formative.jmir.org), 25.03.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Perceived Trust and Professional Identity Threat in AI-Based Clinical Decision Support Systems: Scenario-Based Experimental Study on AI Process Design Features

Sophia Ackerhans[1], MSc; Kai Wehkamp[2], Prof Dr Med; Rainer Petzina[2], Dr phil, Prof Dr Med; Daniel Dumitrescu[3], Dr med; Carsten Schultz[1], Prof Dr

[1]Kiel Institute of Responsible Innovation, University of Kiel, Kiel, Germany

[2]Medical School Hamburg, Hamburg, Germany

[3]Clinic for General and Interventional Cardiology/Angiology, Heart and Diabetes Center North Rhine-Westphalia, Ruhr-University Bochum, Medical Faculty OWL (University Bielefeld), Bad Oeynhausen, Germany

Corresponding Author:
Sophia Ackerhans, MSc
Kiel Institute of Responsible Innovation
University of Kiel
Westring 425
Kiel, 24118
Germany
Phone: 49 431880479
Email: ackerhans@bwl.uni-kiel.de

Abstract

Background: Artificial intelligence (AI)–based systems in medicine like clinical decision support systems (CDSSs) have shown promising results in health care, sometimes outperforming human specialists. However, the integration of AI may challenge medical professionals' identities and lead to limited trust in technology, resulting in health care professionals rejecting AI-based systems.

Objective: This study aims to explore the impact of AI process design features on physicians' trust in the AI solution and on perceived threats to their professional identity. These design features involve the explainability of AI-based CDSS decision outcomes, the integration depth of the AI-generated advice into the clinical workflow, and the physician's accountability for the AI system-induced medical decisions.

Methods: We conducted a 3-factorial web-based between-subject scenario-based experiment with 292 medical students in their medical training and experienced physicians across different specialties. The participants were presented with an AI-based CDSS for sepsis prediction and prevention for use in a hospital. Each participant was given a scenario in which the 3 design features of the AI-based CDSS were manipulated in a 2×2×2 factorial design. SPSS PROCESS (IBM Corp) macro was used for hypothesis testing.

Results: The results suggest that the explainability of the AI-based CDSS was positively associated with both trust in the AI system (β=.508; $P<.001$) and professional identity threat perceptions (β=.351; $P=.02$). Trust in the AI system was found to be negatively related to professional identity threat perceptions (β=–.138; $P=.047$), indicating a partially mediated effect on professional identity threat through trust. Deep integration of AI-generated advice into the clinical workflow was positively associated with trust in the system (β=.262; $P=.009$). The accountability of the AI-based decisions, that is, the system required a signature, was found to be positively associated with professional identity threat perceptions among the respondents (β=.339; $P=.004$).

Conclusions: Our research highlights the role of process design features of AI systems used in medicine in shaping professional identity perceptions, mediated through increased trust in AI. An explainable AI-based CDSS and an AI-generated system advice, which is deeply integrated into the clinical workflow, reinforce trust, thereby mitigating perceived professional identity threats. However, explainable AI and individual accountability of the system directly exacerbate threat perceptions. Our findings illustrate the complex nature of the behavioral patterns of AI in health care and have broader implications for supporting the implementation of AI-based CDSSs in a context where AI systems may impact professional identity.

(JMIR Form Res 2025;9:e64266) doi:10.2196/64266

XSL•FO
RenderX

KEYWORDS

artificial intelligence; clinical decision support systems; explainable artificial intelligence; professional identity threat; health care; physicians; perceptions; professional identity

Introduction

Background

Research in the field of health care innovation has begun to examine the impact of artificial intelligence (AI) on human-cognitive structures and processes of medical decision-making and problem-solving [1,2]. AI enables machines to process and analyze vast datasets independently [3], improving clinical outcomes in areas such as sepsis detection and error reduction [4,5]. AI-based clinical decision support systems (CDSSs) can enhance patient care while reducing costs and errors [6]. However, successful implementation requires acceptance by health care professionals, which is often hindered by skepticism, distrust, and perceived threats to professional identity [7,8].

Process design features, such as explainability and reliability, play a crucial role in shaping trust in AI-based CDSSs and may influence physicians' professional identity [8-11]. While prior research has explored the relationship between trust and technology acceptance [12,13], the impact of these design features on professional identity threats remains underexplored. This study investigates how process design features influence physicians' trust in AI-based CDSSs and how this trust relates to professional identity threats, providing insights to support AI implementation in health care.

Conceptual Framework

Early research on technology adoption in organizations primarily focused on how users perceive and respond to technical aspects. For instance, the Technology Acceptance Model [14] delineates the key determinants shaping users' attitudes and intentions toward adopting new workplace technology. Recent studies have scrutinized technology adoption from the perspective of trust formation [12,15,16], revealing that a lack of trust in technology can lead to its misuse and nonadoption of technology [15]. For example, initial trust in technology positively mediates the relationship between technology-related factors (such as performance expectancy, effort expectancy, and task complexity) and behavioral intentions [16,17]. Trust denotes the degree to which an individual is willing to believe in or be vulnerable to the actions of another party [18]. This is particularly relevant in situations characterized by risk or uncertainty, where individual behavior can lead to loss or harm [19]. Importantly, this description extends trust beyond human interaction to encompass technology, including AI [20]. This study focuses on the user's trust in the AI-based technology, rather than interpersonal trust (trust between individuals) or social trust (trust in an institution) [21]. Trust is a crucial prerequisite for physicians in adopting AI [22], as AI is perceived as risky due to the complexity and unpredictability of its behavior [11]. Thus, the formation of trust in an AI-based system among physicians is influenced by the AI system's representation and tangibility, that is how the underlying rationale of AI tools' decision outcomes are presented to the user [12]. Consequently, the representation of AI functionalities is expected to play a pivotal role in fostering trust in AI among physicians.

IT systems that disrupt traditional work routines often meet with particular resistance. This defensiveness originates, among other things, from the potential threat to autonomous work processes and traditional routines [23]. Studies have emphasized the relevance of a perceived threat or enhancement posed by new technologies to professional identities, underscoring the importance of sociocognitive elements related to human expertise, including hierarchy, status, autonomy, control, and legitimacy, particularly in knowledge-intensive organizations [23-26].

Professional identity pertains to how individuals define and distinguish themselves in their professional roles [27]. Medical professionals exhibit a particularly strong attachment to their social group and professional identity, which evolves through rigorous socialization during education and the acquisition of extensive expertise through practical experience [28,29]. Health care professionals often have the autonomy to make treatment decisions, intensifying their motivation and dedication and reinforcing their identification with the medical field [30]. Disruptions to existing workflows will be perceived as threats to health professionals' identity if they restrict autonomy in terms of decision-making and sovereignty over medical knowledge [8].

A recent literature review on the implementation challenges of CDSSs showed that a CDSS transparency in elucidating decision outcomes (the "explainable" AI), integration depth of the AI-generated system advice into the clinical workflow, and security- and privacy-related mechanisms such as system-induced individual accountability frame professional identity threats and enhancements among physicians [10]. These 3 process design features of AI-based systems may have positive and negative behavioral consequences.

The AI's ability to analyze vast amounts of data and transparently present the outcome of the analysis can complement physicians' expertise and strengthen their professional identity, by facilitating more informed decision-making and affirming physicians in their perceived efficiency and effectiveness. However, as the medical environment in which AI is used is usually complex and the behavior of AI is not deterministic, the opaque, multilayered process of AI decision-making is difficult to predict, and the underlying logic of the decision may be poorly understood [11,30]. The risk of overreliance on AI may lead to a perceived erosion of traditional medical expertise. Physicians may associate AI recommendations with a devaluation of their professional judgment [7].

AI-based systems that integrate seamlessly into clinical workflows can enhance the efficiency and productivity of clinicians, allowing them to focus more on patient care. This

XSL•FO
RenderX

can strengthen their trust in the IT system [31]. However, AI systems that provide detailed behavioral advice, which closely resemble physicians' established decision-making parameters, can disrupt the intrinsic rhythm of medical practice. Such disturbances may foster a perception of diminished autonomy among physicians that could potentially compromise the physicians' professional identity [10].

In addition, the use of digital signatures and the monitoring of user credentials play a critical role in safeguarding patient information and maintaining trust in the health care system [10,32]. The requirement for user signatures on treatment decisions ensures that actions affected by the AI-based CDSS are traceable to specific, authorized users, validating the technologically derived information, and reinforcing the confidentiality and integrity of patient data. The tracking of user credentials and actions within AI-based CDSS underscores the responsibility of physicians in making informed, ethical treatment decisions, which reinforces their professional identity as accountable caretakers of patient health [32]. However, these systems may on the other hand negatively influence the professional identity and role of physicians by embedding accountability and increasing monitoring of diagnostic and therapeutic decisions by clinical managers and external agents such as government bodies and health insurance [8]. Electronic monitoring facilitates the regulation and control of both processes and outcomes. This technological oversight grants clinical managers and regulators the capability to impose a form of direct governance over health care professionals. It effectively binds physicians to a self-regulatory framework, which is less predicated on the active enforcement of control than on the perceived omnipresence of surveillance [33]. Furthermore, detailed tracking of decisions and actions can be used in legal situations, increasing the liability risks for medical professionals. This might lead to more defensive behavior, where professionals take extra precautions, not out of clinical necessity, but to protect against potential legal repercussions [7].

In the subsequent section, we formulate hypotheses regarding the role of trust in AI-based CDSSs and the potential threat or enhancement to the professional identity of health care professionals stemming from the use of AI-based CDSSs. In this context, we investigate the impact of AI explainability, depth of integration of the AI-generated system advice into the clinical workflow, and system-induced individual accountability. Figure 1 provides a visual representation of our proposed research model.

Figure 1. The proposed research framework for analyzing the impact of AI process design features on perceived professional identity threat. AI: artificial intelligence; CDSS: clinical decision support system; H1A: hypothesis 1A; H1B: hypothesis 1B; H2A: hypothesis 2A; H2B: hypothesis 2B; H3A: hypothesis 3A; H3B: hypothesis 3B; H4, hypothesis 4.

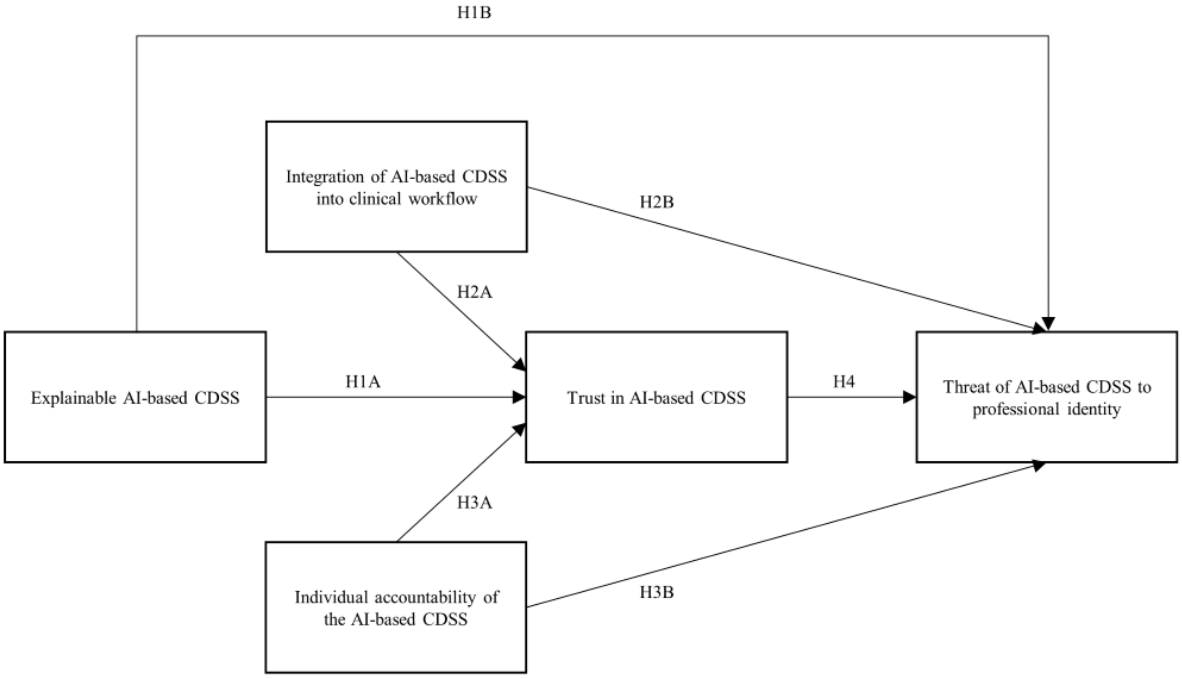

Hypothesis 1A: The Explainability of an AI-Based CDSS Has a Positive Effect on Trust for Health Care Professionals in the Decision Outcomes of Such Systems

Explainable AI represents a crucial component in building trust in AI-based technologies by ensuring transparency in the system's operational principles [12,34]. Such a system empowers users to understand the system's internal operations and decision rationale. For instance, it can illustrate how inputs are mathematically transformed into outputs [35]. Various levels of explainability can be implemented, including the presentation of varying weightings of data leading to AI decisions [36]. Numerous studies highlight the importance of explainable models in health care, enabling health care professionals to trust AI-generated outcomes [37,38]. Consequently, we propose that the transparent display of the data used and the mathematical basis of decision outcomes in an explainable AI-based CDSS fosters higher levels of trust.

XSL•FO
RenderX

Hypothesis 1B: The Explainability of an AI-Based CDSS Has a Positive Effect on Physicians' Professional Identity Threat

Nevertheless, explainable AI-based CDSSs may, in contrast, pose a potential threat to medical authority and the role of physicians [7,8]. As the potential of the AI system becomes transparent to health care professionals, they are more likely to believe in a shift in hierarchy and an erosion of physicians' autonomy and status as a result of AI systems [8,26,39]. In scenarios where AI partially supersedes a physician's expertise, particularly in high-risk situations, it exacerbates their sense of loss of control and status [7], which can induce stress as their professional decision-making control and autonomy may be undermined [40]. An explainable AI-based CDSS may as such be perceived as a threat to their established roles and professional standing.

Hypothesis 2A: An AI-Based CDSS That Is Deeply Integrated Into the Clinical Workflow Has a Positive Effect on Physicians' Trust in the System

Evaluating the AI-based CDSSs' fit into the clinical workflow is crucial because users' trust in AI varies with the expected efficacy of the technical system [41]. For instance, physicians often disregard AI recommendations that are considered unnecessary, potentially missing out on relevant decision-enhancing benefits [7,42]. Detailed clinical treatment recommendations, especially for high-risk events or complex tasks, can reinforce confidence in the system by reducing uncertainty through the clinical information provision [30,34,43]. Consequently, the degree of trust in the technology's effectiveness is a result of users' perception of its ability to offer sufficient, efficient, and prompt assistance.

Hypothesis 2B: An AI-Based CDSS That Is Deeply Integrated Into the Clinical Workflow Has a Positive Effect on Physicians' Perceived Professional Identity Threat

Physicians' strong identification with their profession [28] may make them skeptical of AI-driven system advice, especially in situations where the AI task undermines their own clinical workflow and tasks [30,44]. Physicians may fear losing control over medical decisions if they are compelled to follow the system's advice [7]. Integrated AI-based systems with automated advice, which closely mirror established human tasks, may disrupt current established processes. Moreover, explicit advice on expected behavior may diminish professional authority and autonomy, eroding individual control. When systems control actions, the scope for human professional judgment is limited, impacting the ability to provide proficient recommendations. This situation can lead to a sense of redundancy [44].

Hypothesis 3A: An AI-Based CDSS That Induces Individual Accountability Has a Negative Effect on Physicians' Trust in the AI-Based CDSS

A mandatory use of AI-generated feedback, aimed at enhancing human-AI collaboration, can inadvertently bring attention to managerial monitoring and tracking, potentially leading to concerns about excessive control and distrust. For instance, demanding health care professionals to approve AI system decisions can be seen as invasive and indicative of a lack of managerial discretion, which may undermine both trust in the system and compliance [45]. Similar issues have arisen in contexts like Uber (Uber Technologies Inc), where continuous tracking eroded drivers' autonomy and trust [46]. Thomas et al [47] investigated barriers to adopting e-prescribing and found that requiring health care professional signatures on prescriptions initially caused frustration and a preference for paper-based prescriptions. However, over time, the use of signatures became less of a barrier. Still, constant signature requirements can increase ambiguity about treatment responsibility, leading to reduced trust in the system. Physician signatures for validating the AI system decisions make them cocreators of the AI-based advice, increasing risk awareness and responsibility for the IT system's actions, and as such decreasing system trust.

Hypothesis 3B: An AI-Based CDSS That Induces Individual Accountability Has a Positive Effect on the Physicians' Perceived Threat to Professional Identity

Active endorsement of AI-based CDSS decisions can lead to an increased perceived reliance of health care professionals on the system. This perceived dependency diminishes their sense of autonomous decision-making and intensifies the pressure on physicians to integrate AI results into their professional judgment [30]. This pressure to endorse and validate AI-based decisions raises questions about the distribution of responsibility and accountability, reducing an individual's ability to influence a decision [48]. This kind of AI can infringe physicians' autonomy and control as they have to justify AI-based CDSS decisions, despite a lack of knowledge about the data quality and accuracy of the system [44]. The traceability provided by signatures raises concerns about the benefits and risks of AI systems, potentially jeopardizing a physician's reputation as a medical services provider [49].

Hypothesis 4: Trust in an AI-Based CDSS Mediates the Relationship Between an Explainable AI-Based CDSS and Physicians' Perception of Threat to Their Professional Identity

Finally, we anticipate a mediated relationship between the analyzed process design features of AI systems and their influence on the perception of threats to professional identity via trust. Empirical research has nuanced human trust in AI into the dimensions of cognitive and affective trust [50]. Cognitive trust in AI, rooted in rational evaluation and reliability, and affective trust, driven by emotional and interpersonal connections, both emerge as key facets that influence the acceptance and use of AI [11]. For cognitive trust, AI reliability and explainability have been identified as pivotal, as the degree to which the functionality and decision-making processes of the AI are understandable to users is crucial for building trust. Affective trust, on the other hand, is based on affective bonds and feelings toward AI technology rather than rational evaluations of its capabilities. The degree to which AI demonstrates behaviors that promote perceived closeness to actual medical tasks can foster emotional bonds between users and AI, enhancing trust on an emotional level [11]. Given the relationship between the use of an AI-based CDSS and threat

XSL•FO
RenderX

to professional identity, it is reasonable to assume that physicians' cognitive and affective trust in explainable AI-based CDSSs can reduce their perception of threat to their professional identity.

Methods

Study Design

Scenario-based experimental surveys present a valuable approach to quantitative research by integrating the strengths of traditional surveys and experimental designs. This method enhances external validity through the representativeness and multivariate measurement capabilities of surveys, while simultaneously improving internal validity through controlled, experimental interventions. By addressing the limitations of both approaches—namely, the multicollinearity and passive measurement inherent in surveys, and the lack of representativeness and simplified settings typical of experiments—scenario-based studies offer a comprehensive and rigorous method for investigating respondents' beliefs, attitudes, and judgments [51].

Sample and Data Collection

The study was conducted with German medical students, physicians, and nursing professionals. The study sample was composed primarily of German medical students in the second (clinical) part of the course study program, as they represent future users of AI-based risk prediction systems and are at a critical stage of developing their professional identity. During their education, medical students gain professional experience by interacting directly with patients under supervision, allowing them to apply theoretical knowledge to real-world clinical scenarios. This hands-on training is essential for developing diagnostic skills, clinical decision-making, and effective patient communication. Medical students undergo continuous practical training in university clinics, teaching hospitals, or other medical centers. In addition, they are required to complete a 3-month internship in a hospital before starting their medical study program or during nonteaching periods of the preclinical stage, which familiarizes them with hospital operations and clinical procedures. Furthermore, a 4-month clinical traineeship is mandatory during the transition between the preclinical and clinical stages of medical education. These structured experiences progressively immerse students in hospital environments, allowing them to interact with diverse physicians and develop their professional identities. Research supports the idea that medical students' professional identity evolves rapidly during their education. Niemi [52] notes that critical moments for identity development occur at the start of professional training, as students begin to conceptualize themselves as future physicians. Similarly, Pitkala and Mantyranta [53] found that while medical students initially lacked confidence, their interactions with patients during clinical placements increased their sense of credibility and comfort by the end of their first year. Madill and Latchford [54] further observed that students' identity progression during their first year of training was marked by growing competence and dedication. Advanced medical students, therefore, are well-positioned to recognize and articulate perceptions of professional identity threats.

For the study, 8 different scenarios of a fictitious AI-based CDSS were developed, which showed the user the risk of developing sepsis within the next 48 hours. Sepsis is a complex infectious disease that can develop from any focus of infection. It can be caused by viruses, bacteria, fungi, or parasites and can affect patients of all ages [55]. According to an analysis of Germany-wide Diagnosis Related Groups statistics, there was an incidence of 158 patients with sepsis per 100,000 inhabitants in Germany in 2015. The proportion of sepsis patients among all hospital patients was 0.7%. 53.8% of patients with hospital-acquired sepsis were treated in intensive care units and 41.7% died in hospital [56]. In addition to the lack of improvement in treatment standards and measures to reduce hospital-acquired infections, this high mortality rate is attributed to a lack of quality initiatives for early detection [57], which illustrates the high relevance of AI-supported CDSSs for sepsis risk prediction.

Five physicians from a university hospital in northern Germany evaluated the scenario-based web-based survey during a pretest to ensure that all 3 factors studied were successfully manipulated and that participants could distinguish between them. Participants were familiar with the topic of sepsis, as this condition is considered one of the most common preventable causes of death and requires timely antibiotic therapy, which largely depends on the physician's judgment if an AI-based CDSS is not used. Manipulation checks were conducted to ensure that participants understood the manipulation of the scenario. It was also checked whether the respondents answered the survey seriously (seriousness check).

Participants were recruited via paper and digital flyers distributed in senior medical lectures at medical faculties in northern Germany. In addition, participants were recruited via social media and newsletters from medical faculties at German universities. Participation in the study was voluntary, and participants could access the web-based survey via a link and QR code provided on the recruitment flyers. Before the participants could access the questionnaire, they had to give their informed consent to participate. Participants in the study were able to view information about the duration of the survey and data collection (location and duration of data storage, investigator, and purpose of the study) via a link to a pop-up on the first page of the survey before giving their consent. The data was collected between the end of 2022 and summer 2023. Of the 673 people who clicked on the QR code or link for the survey, 333 (49.48%) participants accessed the survey. We received 292 completed responses, which we used for analysis.

Design and Manipulations

In a 2×2×2 factorial design, 3 factors were manipulated: AI-based CDSS decision explainability (explained vs not explained), integration depth of AI-based CDSS advice into the clinical workflow (detailed vs no detailed clinical treatment advice), and system-induced accountability of a physician in relation to the AI-based CDSS advice (system required a signature vs no signature). These manipulations were designed to produce 8 different treatments (Table 1). All participants were given a scenario of AI Med Predict, a fictitious AI-based CDSS. We modified the wording of the scenarios while keeping

XSL•FO
RenderX

the content consistent to exclude other interpretations. The participants were randomly divided into 8 groups, each exposed to one of the treatments (between-subject design). A minimum size of 23 in each group was calculated with an effect size of 0.85, an α of .05, and a statistical power of 0.8 using the statistical software G*Power. Thus, the cell sizes for the 8 treatments were large enough for this exploratory study (n_1=29; n_2=44; n_3=41; n_4=45; n_5=31; n_6=36; n_7=42; and n_8=24) [58,59].

We developed a hypothetical AI-based CDSS tool for sepsis risk prediction to operationalize AI-based CDSS decision explainability, depth of integration of AI-based CDSS advice into the clinical workflow, and system-induced individual accountability. The simulated CDSS, set in a fictional ward view, displayed the patient's name, age, case number, primary diagnosis, room and bed, progress remark, and estimated sepsis risk score in percent. Subsequently, the respondent has seen a scenario involving a hypothetical patient with a high-risk prognosis for sepsis within 48 hours.

The explainability of the AI-based CDSS decision was altered through 2 treatments. If the user received a detailed explanation of the AI-based CDSS decision that included the patient data used in the risk analysis calculation, such as preexisting conditions, recent procedures, vital signs, laboratory values, or age, we coded the variable as 1. If the user did not receive an explanation of the estimated sepsis risk score, we coded the variable as 0. The integration depth of AI-generated advice into the clinical workflow was coded as 1 if the system specified preventive treatment procedures, such as catheter change and microbiological investigations. The respondent was also provided with a link to the German sepsis prevention, diagnostic, treatment, and aftercare decision guidelines [60]. If the system made an unspecific call for preventive activities, the variable was coded as 0. System-induced individual accountability was coded as 1 if the system hypothetically sought a signature from the respondent to acknowledge the risk prediction and to confirm the planned preventative actions. If the system instructed the user to examine another subject without requiring a signature, the variable was coded as 0.

To ensure treatment efficacy, we used manipulation checks. Specifically, the respondents were asked to evaluate the explainability of the AI-based CDSS decision (eg, whether AI Med Predict for risk prediction of sepsis shows them the composition or explanation of the risk value), the depth of integration of AI-based CDSS advice into the clinical workflow (eg, whether AI Med Predict for risk prediction of sepsis gives them an exact recommended course of action including a link to treatment guideline) regarding the measures they should initiate for the patient), and system-induced individual accountability for the decision outcome of the AI-based CDSS (eg, whether AI Med Predict for risk prediction of sepsis finally prompts them for their signature). Only participants were included in the dataset, which have successfully passed the manipulation check. On average, the survey completion time was 15 minutes 16 seconds.

Table 1. Structure of the 2×2×2 scenario-based experiment design with 8 scenarios.

Scenario	Explainability of AI[a]-based CDSS[b] decision		Integration depth of AI-based CDSS advice into the clinical workflow		Individual accountability induced by AI-based CDSS	
	Explained	Not explained	Detailed treatment advice	No detailed treatment advice	Signature required	No signature required
1		✓		✓		✓
2	✓			✓		✓
3		✓	✓			✓
4	✓		✓			✓
5		✓		✓	✓	
6	✓			✓	✓	
7		✓	✓		✓	
8	✓		✓		✓	

[a]AI: artificial intelligence.

[b]CDSS: clinical decision support system.

Measurement Scales

All question items were measured using a 5-point Likert scale from "strongly disagree" to "strongly agree" (Multimedia Appendix 1 lists the measurement items in detail). Professional identity threat was measured by a 6-item scale adopted from Walter and Lopez [8]. Sample items include "Using AI Med Predict to predict risk of sepsis reduces my control over clinical decisions." Trust in AI-based CDSSs was measured by adapting the 8-item scale developed by Hoffman et al [34]. An example item is "I have confidence in the AI Med Predict system for risk prediction of sepsis. I have the feeling that it works well." Because the 2 scales represent relatively recent additions and include self-developed items, we conducted an exploratory factor analysis (EFA) to assess their construct validity [61]. For the scales we used, a principal component analysis with varimax rotation was performed. The EFA enabled all items to load on any factor, with the number of factors being defined by an eigenvalue larger than 1.00 [62]. As reported in Multimedia Appendix 2, this analysis yielded 2 factors. The first is composed of items 1, 2, 3, and 5 of the professional identity threat scale adapted from [8]. The second factor includes the

XSL•FO
RenderX

first 4 items of the AI trust scale proposed by [34]. Any remaining items that failed to load on their respective scales were subsequently eliminated. Despite this, the scales appear cohesive enough for analysis. The scales for professional identity threat and trust in technology demonstrated reliability, with Cronbach α values of 0.83 and 0.81, respectively [63].

Personal innovativeness with technology was examined as a covariate using a scale, titled the same as Agarwal and Prasad [64], which measures the respondent's willingness to explore new information technology. EFA was used to evaluate the personal innovativeness scale's component structure [61]. We found a single-factor solution for personal innovativeness with a technology scale and a satisfactory Cronbach α (0.86), indicating excellent construct validity and reliability of these measures [63,64].

Ethical Considerations

Ethical approval was granted by the Central Ethics Committee of Kiel University (ZEK-19/22). Informed consent was obtained from all participants prior to participation in the scenario-based experiment. The confidentiality of the participants was maintained at all times. The participants received no compensation for their participation. In return for participating in the survey, we promised an earmarked donation per completed response to a medical charity in the survey.

Results

Descriptive Statistics

A total of 333 responses were collected. Respondents who indicated that their education or employment was not related to the health care field were excluded, as their experiences may not be relevant to the research question (n=16, 4.8%). Manipulation checks were included to verify whether participants properly understood and engaged in experimental conditions. Participants who failed these checks were excluded to maintain data integrity (n=16, 4.8%). In addition, seriousness checks were used to ensure participants provided thoughtful responses. Those who failed these checks, typically by answering in a way that indicated random or inattentive participation, were excluded (n=9, 2.7%). This resulted in a final sample of 292 (87.7%) participants for data analysis. 278 (95.2%) were medical students, 8 (2.7%) participants were physicians across different specialties, and 6 (2.1%) were nursing professionals (eg, intensive care nurses). The participants' average age was assessed using a categorical scale, where 1=18 years or younger and 7=60 years or older. The mean score for age was 2.53 (SD 0.75), indicating that the typical participant in the sample was approximately 24-25 years of age, corresponding to the mid-20s age range. 70% of the participants were female (Table 2). On average, the medical students attended their seventh semester and 74.1% had 3 or more years of bedside experience. 75% of the physicians and 83.3% of the nursing professionals were younger than 30 years and were still active in medical specialist training. 57.2% of the respondents stated that they lived in Schleswig-Holstein or Hamburg, 9.6% in Lower Saxony, 9.3% in Baden-Wuerttemberg, and 8.9% in Bavaria. The complete demographic characteristics of the sample are present in Multimedia Appendix 3.

The mean values, SDs, and Pearson correlations for the study variables are displayed in Table 2. The mediator and dependent variables are weakly correlated (trust and professional identity threat: r=–0.119; P=.042). Explainability of the AI-based CDSS' decision outcome and trust in an AI-based CDSS are moderately correlated (r=0.367; P<.001). AI-based system advice (r=–0.115; P=.049) and accountability of the AI-based CDSS via a system-required signature (r=0.122; P=.037) are weakly correlated with professional identity threat.

XSL•FO
RenderX

Table 2. Mean values, SDs, and correlation matrix (N=292).

	Mean	SD	1	2	3	4	5	6	7	8
Explainable AI[a]-based CDSS[b]	0.49	0.501	1.00	__[c]	—	—	—	—	—	—
AI-based CDSS system advice	0.48	0.500	–0.117[d]	1.00	—	—	—	—	—	—
AI-based CDSS tracking	0.54	0.499	–0.108	–0.045	1.00	—	—	—	—	—
Trust	3.278	0.655	0.367[e]	0.114	–0.022	1.00	—	—	—	—
Professional identity threat	1.907	0.708	0.034	–0.115[f]	0.122[g]	–0.119[h]	1.00	—	—	—
Personal innovativeness	3.463	0.799	0.069	–0.016	–0.005	0.074	–0.103	1.00	—	—
Age (1-7 scale)[i]	2.50	0.780	–0.009	0.137[j]	–0.036	–0.104	–0.027	–0.077	1.00	—
Sex (1: female)	0.70	0.58	0.039	–0.064	0.006	0.133[k]	–0.030	–0.300[l]	–0.089	1.00

[a] AI: artificial intelligence.

[b]CDSS: clinical decision support system.

[c]Not applicable.

[d]P=.045.

[e]P<.001.

[f]P=.049.

[g]P=.04.

[h]P=.042.

[i]Participants reported their age using a categorical scale: 1=18 years or younger, 2=19-24 years, 3=25-29 years, 4=30-39 years, 5=40-49 years, 6=50-59 years, and 7=older than 60 years.

[j]P=.02.

[k]P=.02.

[l]P<.001.

Hypothesis Testing

First, we conducted a 1-way ANOVA to assess the effects of the 8 differently manipulated scenarios on trust in an AI-based CDSS and the threat of AI-based CDSS to professional identity. The level of trust differed statistically significant for the different scenarios, ($F_{7,284}$=2.596; P<.01). The level of perceived professional identity threat differed statistically significant on a 10% level for the different scenarios ($F_{7,284}$=1.818; P<.08). Figure 2 and Figure 3 show the mean values of the 2 variables across the 8 scenarios, along with pairwise comparisons using a post hoc least significant difference test to control for type I error.

The 8 scenarios given in Figures 2 and 3 are as follows. In scenario 1, the AI-based CDSS decision is not explained, the treatment advice is not deeply integrated into the clinical workflow, and there is no individual accountability induced by the system via signature. In scenario 2, the AI-based CDSS decision is explainable; however, the treatment advice remains not deeply integrated into the clinical workflow, and there is no individual accountability induced by the system via signature. Scenario 3 features an AI-based CDSS decision that is not explained, but the treatment advice is deeply integrated into the clinical workflow, while individual accountability is not induced by the system via signature. In scenario 4, the AI-based CDSS decision is explainable, the treatment advice is deeply integrated into the clinical workflow, and no individual accountability is induced by the system via signature. Scenario 5 involves a not explained AI-based CDSS decision with treatment advice that is not deeply integrated into the clinical workflow; however, individual accountability is induced by the system via signature. In scenario 6, the decision is explainable, the treatment advice is not deeply integrated into the clinical workflow, and individual accountability is induced by the system via signature. Scenario 7 presents a not explained AI-based CDSS decision combined with deeply integrated treatment advice in the clinical workflow, and individual accountability is induced by the system via signature. Finally, scenario 8 includes an explainable AI-based CDSS decision, deeply integrated treatment advice within the clinical workflow, and individual accountability is induced by the system via signature. In all scenarios, the error bars represent an SE of 1.

XSL•FO
RenderX

Figure 2. Mean values of variable trust in AI-based CDSS, with pairwise comparisons (post hoc results). AI: artificial intelligence; CDSS: clinical decision support system. *P<.05, **P<.01, ***P<.001.

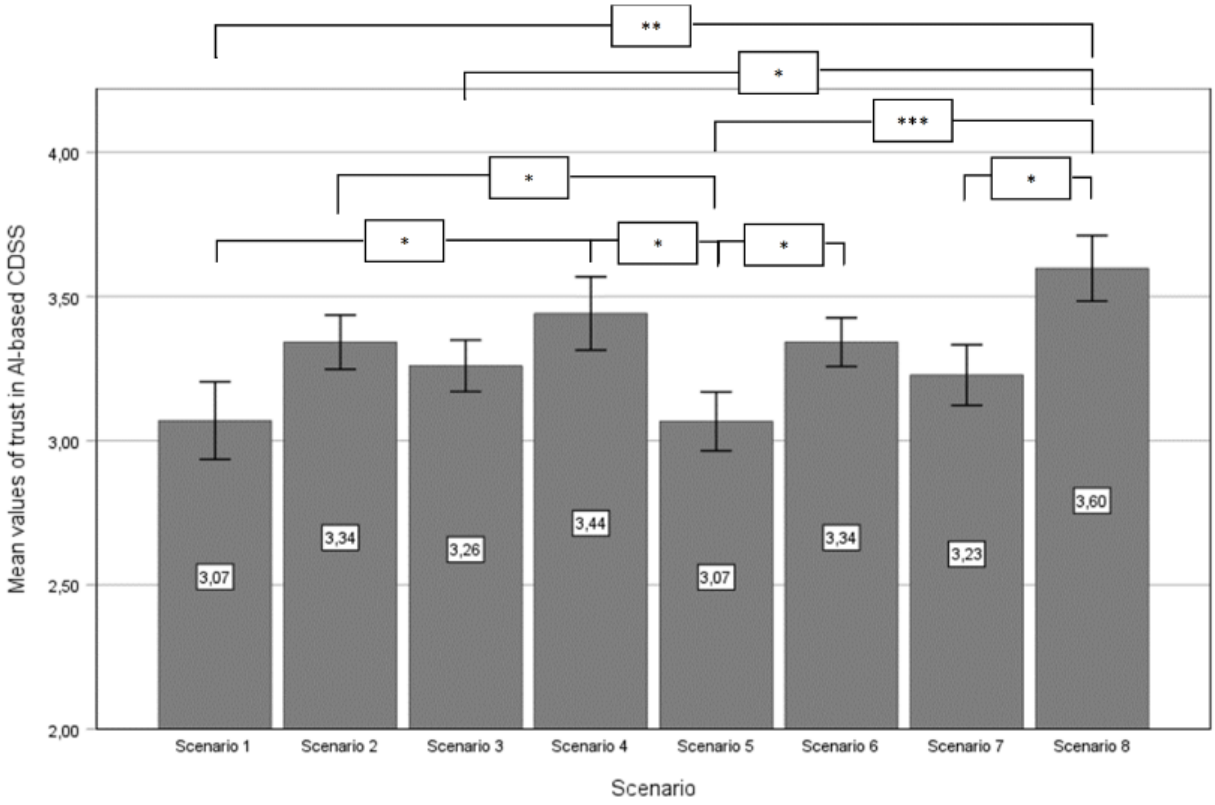

Figure 3. Mean values of variable threat of AI-based CDSS to professional identity, with pairwise comparisons (post hoc results). AI: artificial intelligence; CDSS: clinical decision support system. *P<.05, **P<.01.

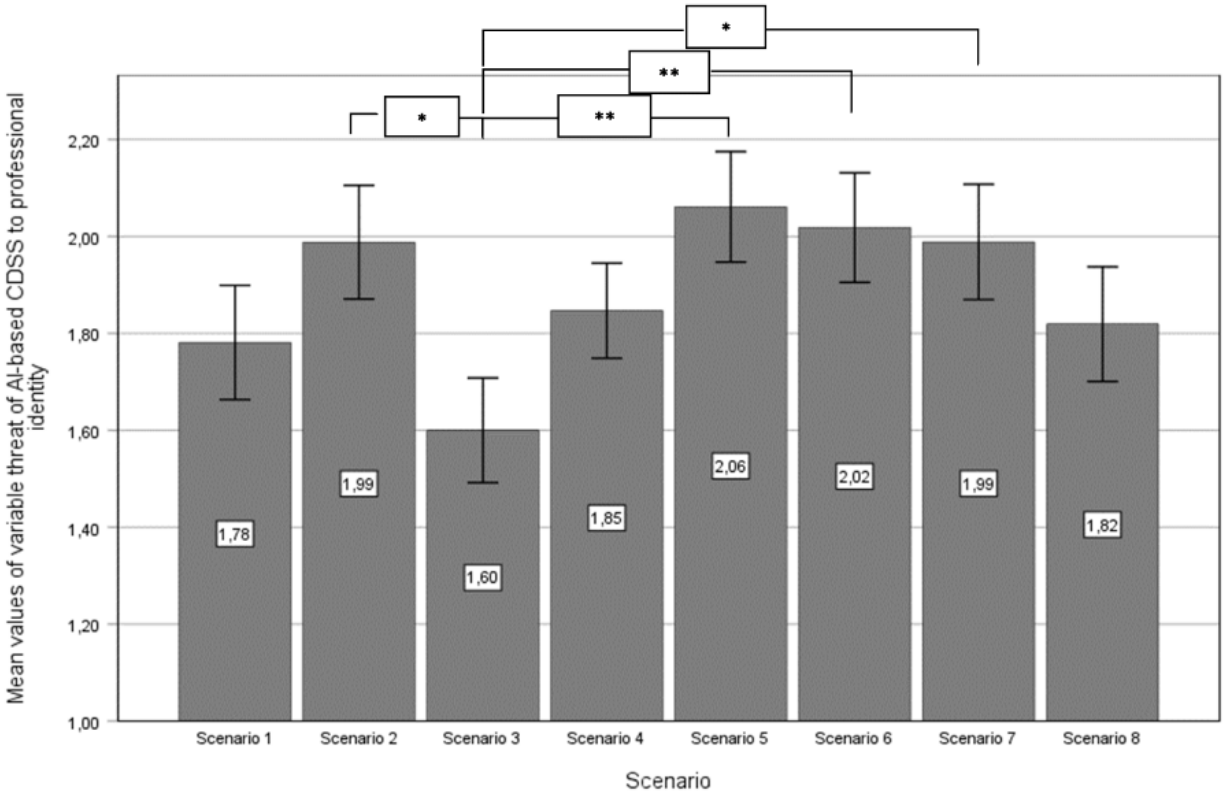

XSL•FO
RenderX

Further, we performed SPSS PROCESS (IBM Corp) macro model 10 to test the proposed hypotheses (Figure 4) [65]. PROCESS is an observed-variable modeling tool that relies on ordinary least squares regression, mediation, or moderation analysis [66]. We carried out bootstrapping to test the indirect effects with a sample size of 5000 and a 95% CI [67]. As suggested by Preacher and Hayes [67], we examined the significance of the effects by checking whether the CIs included 0. As control variables, we considered participants' age, sex, and personal innovativeness in relation to new technologies. The model explains 7.2% of the variance of the dependent variable, professional identity threat (adjusted R^2=0.072; $F_{9,282}$=2.439; P=.01). We found a significant positive association between AI-based CDSS explainability and trust in such a system (β=.508; P<.001). This finding provides support for hypothesis 1A. Furthermore, it was shown that the level of trust in an AI-based CDSS has a negative relationship with a perceived threat to professional identity (β=–.138; P=.047). The results further demonstrated that the indirect effect of the explainability of an AI-based CDSS on professional identity threat via trust in an AI-based CDSS was significant in all treatment manipulations. While holding the other 2 process design features constant, the indirect effect of the explainability of AI via trust was a×b=–.073 (95% CI –0.145 to –0.009). In addition, we found a significant positive direct association between the explainability of an AI-based CDSS and professional identity threat (β=.351; P=.02), providing support for hypothesis 1B. Thus, the impact of explainable AI-driven CDSS on the perception of professional identity threat is found to be partially mediated by the level of trust placed in AI, providing support for hypothesis 4. We found a significant positive association between an AI-based CDSS that is strongly integrated into the clinical workflow and trust in such a system (β=.262; P=.009), providing support for hypothesis 2A. Further, the direct effect of an AI-based CDSS inducing individual accountability on the perceived threat to professional identity was significant (β=.339; P=.004), which supports hypothesis 3B. Finally, sex (1: female) significantly influences trust in an AI-based CDSS (β=.208; P=.01) while personal innovativeness significantly influences the professional identity threat posed by an AI-based CDSS (β=–.108; P=.046; Figure 4). At the same time, the direct influence of a strongly integrated AI-based CDSS into the clinical workflow on a perceived threat to professional identity is not significant. Similarly, a system inducing a physician's accountability does not significantly affect trust in an AI-based CDSS (Table 3).

Figure 4. Regression results (N=292). AI: artificial intelligence; CDSS: clinical decision support system; H1A: hypothesis 1A; H1B: hypothesis 1B; H2A: hypothesis 2A; H2B: hypothesis 2B; H3A: hypothesis 3A; H3B: hypothesis 3B; H4, hypothesis 4.

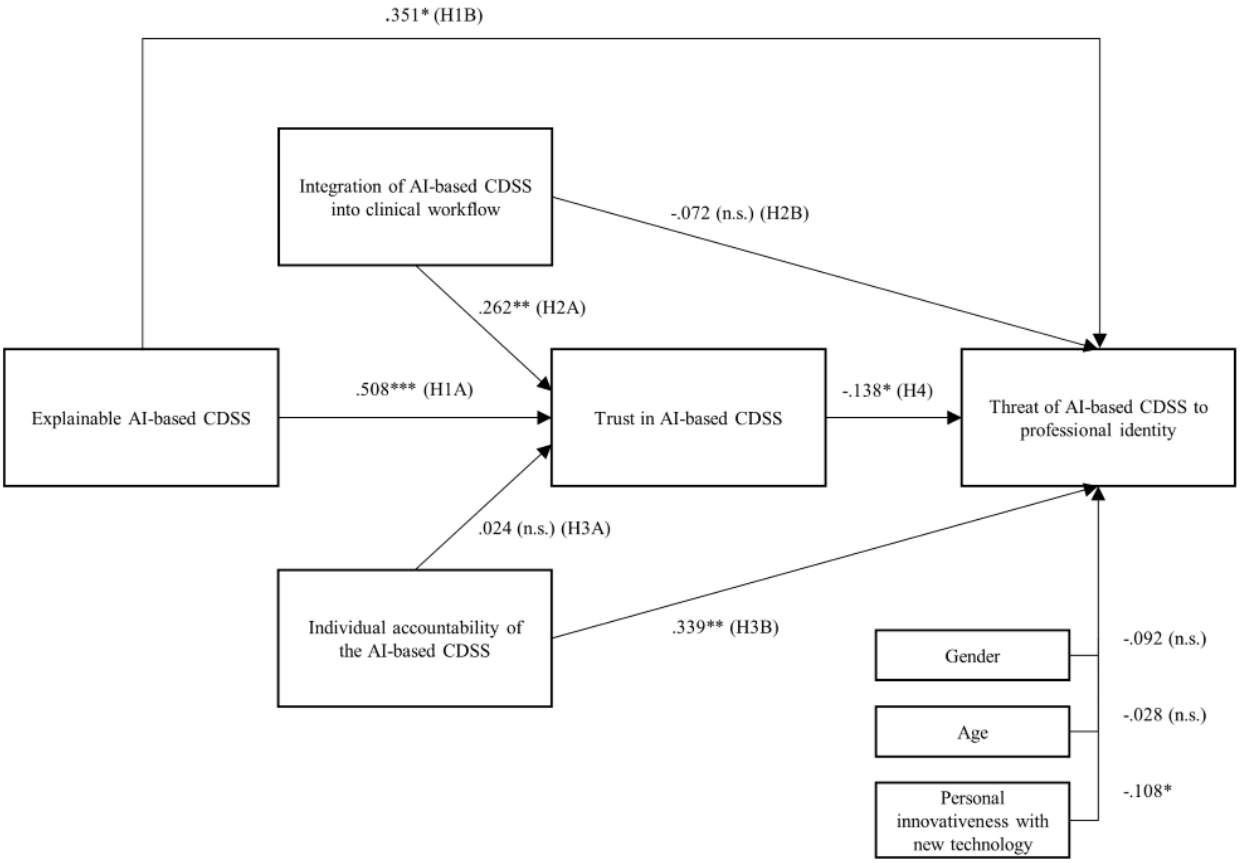

XSL•FO
RenderX

Table 3. Path analysis results with coefficients, *P* values, and 95% CIs (N=292).

Hypothesis and path	Coefficients (β)	*P* value	95% CI
H1A[a]: AI[b]-based CDSS[c] explainability → trust in the AI-based CDSS	.508	<.001	0.258 to 0.758
H1B[d]: AI-based CDSS explainability → perceived threat of AI-based CDSS to professional identity	.351	.02	0.052 to 0.651
H2A[e]: Integration of AI-based CDSS advice into the clinical workflow → trust in the AI-based CDSS	.262	.009	0.067 to 0456
H2B[f]: Integration of AI-based CDSS advice into the clinical workflow → perceived threat of AI-based CDSS to professional identity	–.072	.538	–0.301 to 0.158
H3A[g]: Individual accountability of the AI-based CDSS → trust in the AI-based CDSS	.024	.81	–0.173 to 0.222
H3B[h]: Individual accountability of the AI-based CDSS → perceived threat of AI-based CDSS to professional identity	.339	.004	0.110 to 0.569
H4[i]: Trust in the AI-based CDSS → perceived threat of AI-based CDSS to professional identity	–.138	.047	–0.275 to –0.002
AI-based CDSS explainability → trust in the AI-based CDSS → perceived threat of AI-based CDSS to professional identity	–.073	__[j]	–0.145 to –0.009

[a]H1A: hypothesis 1A.

[b]AI: artificial intelligence.

[c]CDSS: clinical decision support system.

[d]H1B: hypothesis 1B.

[e]H2A: hypothesis 2A.

[f]H2B: hypothesis 2B

[g]H3A: hypothesis 3A.

[h]H3B: hypothesis 3B.

[i]H4: hypothesis 4.

[j]Not applicable.

A formal test of mediation was performed to analyze, independently of the other variables of influence, whether explainable AI-based CDSS predicts a perceived professional identity threat, and whether this direct path would be mediated by trust in the system. A total effect of an explainable AI-based CDSS on professional identity threat perceptions without involving trust was not observed (β=.065; *P*=.43). However, after incorporating the mediator trust in AI into the model, the explainability of an AI-based CDSS significantly predicted the mediator (β=.465; *P*<.001), which in turn significantly influenced perceptions of professional identity threat (β=–.156; *P*<.05). The direct effect of an explainable AI-based CDSS on professional identity threat perceptions via trust was not significant (β=.138; *P*=.12). Thus, our findings suggest that the relationship between AI-based CDSS explainability and professional identity threat perceptions is partially mediated by the trust in the system (indirect effect a×b=–.073 (95% CI –0.143 to –0.009).

Moreover, we have refined the estimation of the model by including other control variables, namely 5 personality dimensions: conscientiousness, extraversion, agreeableness, emotional stability, and openness to experience [68], as well as health care job tenure, and surgical specialties. Here we found that none of the variables had a significant effect on the threat to professional identity or on the change in path coefficients.

Discussion

Principal Results

The present study examines how AI-based CDSS process design features affect health care professionals' trust in AI and their perceived threat to professional identity. The findings show that decision outcome representations and human-AI interaction mechanisms shape health care professionals' trust and professional identity threat perceptions. Our study revealed that trust in AI partially mediates the influence of an explainable AI-based CDSS and strongly integrated AI-based system on the perceived identity threats among health care professionals. This supports earlier studies on AI explainability and clinical workflow integration, establishing trust in technology [10,34]. While AI decision-making may be challenging to understand in certain instances, providing clarity about the algorithm and AI advisory behavior may help align consumers' expectations with the performance of AI [11]. Our study also found a direct

XSL•FO
RenderX

and significant positive effect of explainable AI on the perceived threat to professional identity, a finding that aligns with research on the threat posed by AI in professional work [8]. In this context, AI's enhanced capabilities may raise questions not only about its compatibility with physicians' expertise and experience but also about its individual consequences. Future research needs to ascertain whether explainable AI undermines the perceived professional identity, whether trust in technology stabilizes through emergent dependability in real-life settings, or if this finding is favored by the controlled laboratory setting.

Comparison With Prior Work

In research on AI-based CDSSs, the effect of system integration into the clinical workflow on the emergence of professional identity threats has been investigated and considered as evidence of high absorption of individual medical competence [7,26,30]. Our study does not find any significant effect of AI-based system advice that is deeply integrated into the clinical workflow on individual professional identity threat assessments. This might be attributed to the different interpretations individuals have with high information content and medical guidelines. Future research should investigate human biases associated with AI-based advice, such as medical guidelines, not only to gain a deeper understanding of their functioning role in shaping professional identity threat perceptions among health care professionals but also to prevent their misuse or overuse. Furthermore, future AI research should focus on the influence of AI-based systems' reliability, that is, predicted consistent performance [12], on technological trustworthiness [11]. Another potential area of investigation may be the influence of various representations of AI reliability on threats to professional identity. Furthermore, the ongoing discussion about the ethical responsibility of AI-based CDSSs plays a significant role in this context [12]. Physicians must be able to adhere to the system's decision without fear of biased data composition.

Previous research on monitoring the behavior of health care professionals has identified it as an important factor for distrust in technology [45]. However, as our research indicates, the relationship between AI system monitoring and trust in AI could be more complex, as a system that induces accountability of the physician does not necessarily result in low trust in the AI system and rejection among health care professionals. Health care professionals may perceive a signing function as micromanagement and distrust it, for example, by misusing it to avoid responsibility and traceability of the AI tool's judgment. Our findings imply that such system functionalities interfere profoundly with personal physician identity perceptions. The significant positive relationship between an AI-assisted system that is inducing individual accountability and a perceived threat to professional identity demonstrates that agreement to an AI-assisted treatment decision is viewed as invasive and antagonistic to professional competence. Nevertheless, decision support can be seen as a functionality that enhances professional identity as it is followed rather than relying on individual professional competence. Especially in situations where direct consultation with a more experienced physician is not available, the reassurance of an AI-assisted decision and its validation may lead to increased trust in the technology [30]. Future research should explore how incorporating a human decision into an AI-assisted CDSS promotes trust in the technology and the socio-cognitive aspects of such technology adoption in the medical context.

The evaluative mechanisms used in this investigation to determine trust in AI outcomes predominantly address cognitive trust. Yet, empirical evidence underscores the significant role of emotional dimensions, such as affective trust, in the efficacious deployment of AI within the health care sector. It is imperative that forthcoming research delves more deeply into the determinants influencing the emotional aspects of trust in AI and additional factors such as cooperation and reliance in physician-AI relations [11].

Adopting such a human-centric perspective presents a substantial opportunity for interdisciplinary collaboration among scholars interested in the evolution and assimilation of AI technologies. While the bulk of existing studies on the interplay between humans and AI are mainly conducted by experts in cognitive engineering and information systems, the field of organizational research offers a valuable vantage point.

This study offers valuable insights for health care managers as well as developers and implementers of AI-based CDSSs, which should approach AI integration with an understanding of the health care professionals' perspective, factoring in aspects that foster human trust in AI and the relevance and value of specific tasks to the workforce. Consequently, in determining the tasks to delegate to AI, it is critical to evaluate not only the technological capabilities at their disposal but also to consider the human element, including perceived threats or enhancements to their professional identity, the interests and motivations of the employees, and the potential for enhancing their trust and productivity through AI collaboration. Specifically, our study shows that the explanation of AI-based CDSS decision outcomes proves beneficial for enhancing trust in AI among health care professionals. AI implementers and health care executives should focus on presenting AI decisions as recommendations that align with existing clinical workflows and which can be overruled by a human decision, rather than disrupting work processes. Furthermore, whenever feasible, AI-based CDSSs should refrain from querying accountability that exposes personal information about the operator. To resolve liability issues, established standards and legal frameworks should be developed, freeing individual practitioners of the duty of actively validating this information.

Limitations

This study has several limitations that warrant careful consideration. First, the small sample size constrained the ability to apply certain filtering criteria, thereby limiting the robustness of findings related to the influence of specific variables. A larger sample would facilitate more precise filtering, particularly the exclusion of preclinical medical students whose limited professional experience and socialization may not accurately represent perceptions of threats to professional identity. Moreover, the sample's composition, with 95% comprising medical students, presents a potential selection bias, as their experiences and perceptions may differ significantly from those of practicing physicians, nurses, or other health care professionals. Consequently, the generalizability of the findings

XSL•FO
RenderX

to a broader health care population is limited. Future research should aim to include a more heterogeneous sample of health care professionals to enhance the validity and applicability of the results.

Additionally, the study's use of scenario-based surveys, while methodologically advantageous in combining elements of both traditional surveys and experimental designs, presents inherent limitations. Despite attempts to balance internal and external validity, scenario-based studies may still fail to fully capture the complexity of real-world clinical settings, thus constraining the external validity and broader applicability of the findings [51]. For instance, laboratory studies indicate that users initially exhibit high levels of trust in AI-based CDSS, which diminishes over time as inconsistencies in risk prediction arise [69]. In contrast, field studies suggest that lower initial trust may increase through sustained interaction [70]. Such discrepancies highlight the difficulty in generalizing findings from controlled experimental settings to real-world environments. Furthermore, variability in respondents' interpretations of the hypothetical scenarios presented may introduce bias, potentially affecting the reliability of the results [51].

Finally, the study's focus on an AI-based risk prediction system for sepsis may limit the generalizability of the findings to other clinical conditions or AI applications in different contexts. Future research should examine whether the observed effects are consistent across AI systems designed for other clinical functions to provide a more comprehensive understanding of the factors influencing the adoption and trust of AI in health care settings.

Conclusions

In this study, we demonstrated how distinct AI process design features such as the explainability of AI-based CDSS decision outcomes, the integration depth of AI-based CDSS advice into the clinical workflow, and higher accountability through enforced signatures shape trust in AI and professional identity among health care professionals. Our research demonstrated the critical role that trust in AI played in fostering professional identity among health care workers. While explainable AI-based CDSS and systems highly integrated into clinical workflows enhanced trust, which supported the compatibility of the AI system with professional identity among health care professionals, explainable AI and enforced accountability led to a perceived professional identity threat. The findings of this study can be applicable to a variety of scenarios where the potential threat to professional identity posed by the rising deployment of AI-based systems played a role, especially in knowledge-intensive settings, where well-established and taken-for-granted behaviors, attitudes, and beliefs are questioned by AI solutions.

Acknowledgments

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors. The authors attest that there was no use of generative AI in the generation of text, figures, or other informational content of this paper.

Data Availability

Data are available from the corresponding author on reasonable request.

Authors' Contributions

SA conceived the study, developed the literature search, conducted the web-based survey, analyzed, and interpreted the data, conceptualized and wrote the sections of the manuscript (original draft, review, and editing), and created the tables and figures of the manuscript. KW conceived the study, interpreted the data, and edited the manuscript. RP and DD reviewed the manuscript. CS coordinated and conceived the study, interpreted the data, and edited the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Final list of items used for the hypothesis testing.
[DOCX File , 25 KB - formative_v9i1e64266_app1.docx]

Multimedia Appendix 2
Exploratory factor analysis.
[DOCX File , 25 KB - formative_v9i1e64266_app2.docx]

Multimedia Appendix 3
Sample properties of medical students, physicians, and nursing professionals.
[DOCX File , 38 KB - formative_v9i1e64266_app3.docx]

XSL•FO
RenderX

References

1. Kellogg KC. Subordinate activation tactics: Semi-professionals and micro-level institutional change in professional organizations. Adm Sci Q 2019;64(4):928-975. [doi: 10.1177/0001839218804527]
2. Huang MH, Rust RT. Artificial intelligence in service. J Serv Res 2018;21(2):155-172. [doi: 10.1177/1094670517752459]
3. Brynjolfsson E, McAfee A. The business of artificial intelligence: What it can—and cannot—do for your organization. Harvard Business Review. 2017. URL: https://hbr.org/2017/07/the-business-of-artificial-intelligence [accessed 2025-02-25]
4. Goh KH, Wang L, Yeow AYK, Poh H, Li K, Yeow JJL, et al. Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. Nat Commun 2021;12(1):711 [FREE Full text] [doi: 10.1038/s41467-021-20910-4] [Medline: 33514699]
5. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542(7639):115-118 [FREE Full text] [doi: 10.1038/nature21056] [Medline: 28117445]
6. Fogel AL, Kvedar JC. Artificial intelligence powers digital medicine. NPJ Digit Med 2018;1(1):5 [FREE Full text] [doi: 10.1038/s41746-017-0012-2] [Medline: 31304291]
7. Liberati EG, Ruggiero F, Galuppo L, Gorli M, González-Lorenzo M, Maraldi M, et al. What hinders the uptake of computerized decision support systems in hospitals? A qualitative study and framework for implementation. Implement Sci 2017;12:1-13 [FREE Full text] [doi: 10.1186/s13012-017-0644-2] [Medline: 28915822]
8. Walter Z, Lopez MS. Physician acceptance of information technologies: role of perceived threat to professional autonomy. Decis Support Syst 2008;46(1):206-215. [doi: 10.1016/j.dss.2008.06.004]
9. Shin D. The effects of explainability and causability on perception, trust, and acceptance: implications for explainable AI. Int J Hum Comput Stud 2021;146:102551. [doi: 10.1016/j.ijhcs.2020.102551]
10. Ackerhans S, Huynh T, Kaiser C, Schultz C. Exploring the role of professional identity in the implementation of clinical decision support systems—a narrative review. Implement Sci 2024;19(1):11 [FREE Full text] [doi: 10.1186/s13012-024-01339-x] [Medline: 38347525]
11. Glikson E, Woolley AW. Human trust in artificial intelligence: review of empirical research. Acad Manag Ann 2020;14(2):627-660. [doi: 10.5465/annals.2018.0057]
12. Hoff KA, Bashir M. Trust in automation: integrating empirical evidence on factors that influence trust. Hum Factors 2015;57(3):407-434. [doi: 10.1177/0018720814547570] [Medline: 25875432]
13. Pavlou PA. Consumer acceptance of electronic commerce: integrating trust and risk with the technology acceptance model. Int J Electronic Commer 2003;7(3):101-134. [doi: 10.1080/10864415.2003.11044275]
14. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989;13(3):319-340. [doi: 10.2307/249008]
15. Panicker RO, Soman B, Gangadharan KV, Sobhana NV. An adoption model describing clinician's acceptance of automated diagnostic system for tuberculosis. Health Technol 2016;6:247-257. [doi: 10.1007/s12553-016-0136-4]
16. Fan W, Liu J, Zhu S, Pardalos PM. Investigating the impacting factors for the healthcare professionals to adopt artificial intelligence-based medical diagnosis support system (AIMDSS). Ann Oper Res 2020;294(1):567-592. [doi: 10.1007/s10479-018-2818-y]
17. Choudhury A, Shamszare H. Investigating the impact of user trust on the adoption and use of ChatGPT: survey analysis. J Med Internet Res 2023;25:e47184 [FREE Full text] [doi: 10.2196/47184] [Medline: 37314848]
18. Mayer RC, Davis JH, Schoorman FD. An integrative model of organizational trust. Acad Manag Rev 1995;20(3):709-734. [doi: 10.5465/AMR.1995.9508080335]
19. Mcknight DH, Carter M, Thatcher JB, Clay PF. Trust in a specific technology: An investigation of its components and measures. ACM Trans Manage Inf Syst 2011;2(2):1-25. [doi: 10.1145/1985347.1985353]
20. Wang W, Qiu L, Kim D, Benbasat I. Effects of rational and social appeals of online recommendation agents on cognition- and affect-based trust. Decis Support Syst 2016;86:48-60. [doi: 10.1016/j.dss.2016.03.007]
21. Montague E. Validation of a trust in medical technology instrument. Appl Ergon 2010;41(6):812-821 [FREE Full text] [doi: 10.1016/j.apergo.2010.01.009] [Medline: 20189163]
22. Paul DL, McDaniel Jr RR. A field study of the effect of interpersonal trust on virtual collaborative relationship performance. MIS Q 2004;28(2):183-227. [doi: 10.2307/25148633]
23. Barrett M, Oborn E, Orlikowski WJ, Yates J. Reconfiguring boundary relations: robotic innovations in pharmacy work. Organ Sci 2012;23(5):1448-1466. [doi: 10.1287/orsc.1100.0639]
24. Pollock TG, Lashley K, Rindova VP, Han JH. Which of these things are not like the others? Comparing the rational, emotional, and moral aspects of reputation, status, celebrity, and stigma. Acad Manag Ann 2019;13(2):444-478. [doi: 10.5465/annals.2017.0086]
25. Ferlie E, Fitzgerald L, Wood M, Hawkins C. The nonspread of innovations: the mediating role of professionals. Acad Manag J 2005;48(1):117-134. [doi: 10.5465/AMJ.2005.15993150]
26. Jussupow E, Spohrer K, Heinzl A. Identity threats as a reason for resistance to artificial intelligence: survey study with medical students and professionals. JMIR Form Res 2022;6(3):e28750 [FREE Full text] [doi: 10.2196/28750] [Medline: 35319465]

XSL•FO
RenderX

27. Chreim S, Williams B, Hinings CR. Interlevel influences on the reconstruction of professional role identity. Acad Manag J 2007;50(6):1515-1539. [doi: 10.5465/amj.2007.28226248]
28. Pratt MG, Rockmann KW, Kaufmann JB. Constructing professional identity: the role of work and identity learning cycles in the customization of identity among medical residents. Acad Manag J 2006;49(2):235-262. [doi: 10.5465/amj.2006.20786060]
29. Weaver R, Peters K, Koch J, Wilson I. 'Part of the team': professional identity and social exclusivity in medical students. Med Educ 2011;45(12):1220-1229. [doi: 10.1111/j.1365-2923.2011.04046.x] [Medline: 21999250]
30. Lebovitz S, Lifshitz-Assaf H, Levina N. To engage or not to engage with AI for critical judgments: how professionals deal with opacity when using AI for medical diagnosis. Organ Sci 2022;33(1):126-148. [doi: 10.1287/orsc.2021.1549]
31. Van Biesen W, Van Cauwenberge D, Decruyenaere J, Leune T, Sterckx S. An exploration of expectations and perceptions of practicing physicians on the implementation of computerized clinical decision support systems using a Qsort approach. BMC Med Inform Decis Mak 2022;22(1):185 [FREE Full text] [doi: 10.1186/s12911-022-01933-3] [Medline: 35842722]
32. Fernández-Alemán JL, Señor IC, Lozoya P, Toval A. Security and privacy in electronic health records: a systematic literature review. J Biomed Inform 2013;46(3):541-562 [FREE Full text] [doi: 10.1016/j.jbi.2012.12.003] [Medline: 23305810]
33. Orlikowski WJ. Integrated information environment or matrix of control? The contradictory implications of information technology. Account Manag Inf Technol 1991;1(1):9-42. [doi: 10.1016/0959-8022(91)90011-3]
34. Hoffman RR, Mueller ST, Klein G, Litman J. Metrics for explainable AI: challenges and prospects. arXiv:1812.04608 2018. [doi: 10.48550/arXiv.1812.04608]
35. Doran D, Schulz S, Besold TR. What does explainable AI really mean? A new conceptualization of perspectives. arXiv preprint arXiv:1710.00794 2017. [doi: 10.48550/arXiv.1710.00794]
36. Yang M, Liu C, Wang X, Li Y, Gao H, Liu X, et al. An explainable artificial intelligence predictor for early detection of sepsis. Crit Care Med 2020;48(11):e1091-e1096. [doi: 10.1097/CCM.0000000000004550] [Medline: 32885937]
37. Zheng Q, Delingette H, Ayache N. Explainable cardiac pathology classification on cine MRI with motion characterization by semi-supervised learning of apparent flow. Med Image Anal 2019;56:80-95. [doi: 10.1016/j.media.2019.06.001] [Medline: 31200290]
38. Yan F, Chen Y, Xia Y, Wang Z, Xiao R. An explainable brain tumor detection framework for MRI analysis. Appl Sci 2023;13(6):3438. [doi: 10.3390/app13063438]
39. Lamb R, Davidson E. Information and communication technology challenges to scientific professional identity. Inf Soc 2005;21(1):1-24. [doi: 10.1080/01972240590895883]
40. Kaplan A, Haenlein M. Siri, Siri, in my hand: Who’s the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. Bus Horiz 2019;62(1):15-25 [FREE Full text] [doi: 10.1016/j.bushor.2018.08.004]
41. Shibl R, Lawley M, Debuse J. Factors influencing decision support system acceptance. Decis Support Syst 2013;54(2):953-961. [doi: 10.1016/j.dss.2012.09.018]
42. Russ AL, Zillich AJ, McManus MS, Doebbeling BN, Saleem JJ. Prescribers' interactions with medication alerts at the point of prescribing: a multi-method, in situ investigation of the human-computer interaction. Int J Med Inform 2012;81(4):232-243. [doi: 10.1016/j.ijmedinf.2012.01.002] [Medline: 22296761]
43. de Visser E, Parasuraman R. Adaptive aiding of human-robot teaming: effects of imperfect automation on performance, trust, and workload. J Cogn Eng Decis Mak 2011;5(2):209-231. [doi: 10.1177/1555343411410160]
44. Jussupow E, Spohrer K, Heinzl A, Gawlitza J. Augmenting medical diagnosis decisions? An investigation into physicians’ decision-making process with artificial intelligence. Inf Syst Res 2021;32(3):713-735. [doi: 10.1287/isre.2020.0980]
45. Bernstein ES. The transparency paradox: A role for privacy in organizational learning and operational control. Adm Sci Q 2012;57(2):181-216. [doi: 10.1177/0001839212453028]
46. Möhlmann M, Zalmanson L. Hands on the wheel: Navigating algorithmic management and uber drivers' autonomy. 2017 Presented at: International Conference on Information Systems (ICIS 2017); December 10-13, 2017; Seoul, South Korea.
47. Thomas CP, Kim M, McDonald A, Kreiner P, Kelleher SJ, Blackman MB, et al. Prescribers' expectations and barriers to electronic prescribing of controlled substances. J Am Med Inform Assoc 2012;19(3):375-381 [FREE Full text] [doi: 10.1136/amiajnl-2011-000209] [Medline: 21946239]
48. Ananny M, Crawford K. Seeing without knowing: limitations of the transparency ideal and its application to algorithmic accountability. New Media Soc 2018;20(3):973-989. [doi: 10.1177/1461444816676645]
49. de Vries AE, van der Wal MH, Nieuwenhuis MM, de Jong RM, van Dijk RB, Jaarsma T, et al. Perceived barriers of heart failure nurses and cardiologists in using clinical decision support systems in the treatment of heart failure patients. BMC Med Inform Decis Mak 2013;13:54 [FREE Full text] [doi: 10.1186/1472-6947-13-54] [Medline: 23622342]
50. McAllister DJ. Affect- and cognition-based trust as foundations for interpersonal cooperation in organizations. Acad Manag J 1995;38(1):24-59. [doi: 10.2307/256727]
51. Atzmüller C, Steiner PM. Experimental vignette studies in survey research. Methodology 2010;6(3):128-138. [doi: 10.1027/1614-2241/a000014]
52. Niemi PM. Medical students' professional identity: Self-reflection during the preclinical years. Med Educ 1997;31(6):408-415. [doi: 10.1046/j.1365-2923.1997.00697.x] [Medline: 9463642]

XSL•FO
RenderX

53. Pitkala KH, Mantyranta T. Professional socialization revised: medical students' own conceptions related to adoption of the future physician's role—a qualitative study. Med Teach 2003;25(2):155-160. [doi: 10.1080/0142159031000092544] [Medline: 12745523]
54. Madill A, Latchford G. Identity change and the human dissection experience over the first year of medical training. Soc Sci Med 2005;60(7):1637-1647 [FREE Full text] [doi: 10.1016/j.socscimed.2004.08.035] [Medline: 15652694]
55. Improving the prevention, diagnosis and clinical management of sepsis. World Health Organisation. 2017. URL: https://www.who.int/activities/improving-the-prevention-diagnosis-and-clinical-management-of-sepsis [accessed 2025-02-20]
56. Fleischmann-Struzek C, Mikolajetz A, Schwarzkopf D, Cohen J, Hartog CS, Pletz M, et al. Challenges in assessing the burden of sepsis and understanding the inequalities of sepsis outcomes between National Health Systems: secular trends in sepsis and infection incidence and mortality in Germany. Intensive Care Med 2018;44(11):1826-1835 [FREE Full text] [doi: 10.1007/s00134-018-5377-4] [Medline: 30284637]
57. Bauer M, Groesdonk HV, Preissing F, Dickmann P, Vogelmann T, Gerlach H. [Mortality in sepsis and septic shock in Germany. Results of a systematic review and meta-analysis]. Anaesthesist 2021;70(8):673-680 [FREE Full text] [doi: 10.1007/s00101-021-00917-8] [Medline: 33559687]
58. Faul F, Erdfelder E, Lang A, Buchner A. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods 2007;39(2):175-191. [doi: 10.3758/bf03193146] [Medline: 17695343]
59. Cohen J. Statistical power analysis. Curr Dir Psychol Sci 1992;1(3):98-101. [doi: 10.1111/1467-8721.ep10768783]
60. Brunkhorst FM, Weigand MA, Pletz M, Gastmeier P, Lemmen SW, Meier-Hellmann A, et al. [S3 Guideline Sepsis-prevention, diagnosis, therapy, and aftercare: long version]. Med Klin Intensivmed Notfmed 2020;115:37-109. [doi: 10.1007/s00063-020-00685-0] [Medline: 32356041]
61. Thompson B, Daniel LG. Factor analytic evidence for the construct validity of scores: a historical overview and some guidelines. Educ Psychol Meas 1996;56(2):197-208. [doi: 10.1177/0013164496056002001]
62. Kaiser HF. A second generation little jiffy. Psychometrika 2025;35(4):401-415. [doi: 10.1007/BF02291817]
63. Nunnally JC, Bernstein IH. Psychometric Theory. New York, NY: McGraw-Hill; 1994.
64. Agarwal R, Prasad J. A conceptual and operational definition of personal innovativeness in the domain of information technology. Inf Syst Res 1998;9(2):204-215. [doi: 10.1287/isre.9.2.204]
65. The PROCESS macro for SPSS, SAS, and R. processmacro.org. 2023. URL: https://processmacro.org/version-history.html [accessed 2025-02-20]
66. Hayes AF, Montoya AK, Rockwood NJ. The analysis of mechanisms and their contingencies: PROCESS versus structural equation modeling. Australas Mark J 2017;25(1):76-81. [doi: 10.1016/j.ausmj.2017.02.001]
67. Preacher KJ, Hayes AF. Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. Behav Res Methods 2008;40(3):879-891. [doi: 10.3758/brm.40.3.879] [Medline: 18697684]
68. Gosling SD, Rentfrow PJ, Swann WB. A very brief measure of the Big-Five personality domains. J Res Personal 2003;37(6):504-528. [doi: 10.1016/S0092-6566(03)00046-1]
69. Madhavan P, Wiegmann DA. Similarities and differences between human-human and human-automation trust: an integrative review. Theor Issues Ergon Sci 2007;8(4):277-301. [doi: 10.1080/14639220500337708]
70. Ullman D, Malle BF. Human-robot trust: Just a button press away. 2017 Presented at: HRI '17: Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction; March 6-9, 2017; Vienna, Austria p. 309-310 URL: https://dl.acm.org/doi/pdf/10.1145/3029798.3038423 [doi: 10.1145/3029798.3038423]

Abbreviations

AI: artificial intelligence
CDSS: clinical decision support system
EFA: exploratory factor analysis

Edited by A Mavragani, G Tsafnat; submitted 12.07.24; peer-reviewed by J Odumegwu, M Ji; comments to author 07.09.24; revised version received 01.11.24; accepted 05.01.25; published 26.03.25.

Please cite as:
Ackerhans S, Wehkamp K, Petzina R, Dumitrescu D, Schultz C
Perceived Trust and Professional Identity Threat in AI-Based Clinical Decision Support Systems: Scenario-Based Experimental Study on AI Process Design Features
JMIR Form Res 2025;9:e64266
URL: https://formative.jmir.org/2025/1/e64266
doi:10.2196/64266
PMID:

XSL•FO
RenderX

©Sophia Ackerhans, Kai Wehkamp, Rainer Petzina, Daniel Dumitrescu, Carsten Schultz. Originally published in JMIR Formative Research (https://formative.jmir.org), 26.03.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Stigma and Behavior Change Techniques in Substance Use Recovery: Qualitative Study of Social Media Narratives

Annie T Chen[1], MSIS, PhD; Lexie C Wang[2], MS; Shana Johnny[3], RN, MN, DNP; Sharon H Wong[1], MS; Rahul K Chaliparambil[4], BS; Mike Conway[5], PhD; Joseph E Glass[6], MSW, PhD

[1]Department of Biomedical Informatics and Medical Education, University of Washington School of Medicine, Seattle, WA, United States
[2]Department of Linguistics, University of Washington, Seattle, WA, United States
[3]School of Nursing, University of Washington, Seattle, WA, United States
[4]Northwestern University Feinberg School of Medicine, Chicago, IL, United States
[5]School of Computing and Information Systems, University of Melbourne, Melbourne, Australia
[6]Kaiser Permanente Washington Health Research Institute, Seattle, WA, United States

Corresponding Author:
Annie T Chen, MSIS, PhD
Department of Biomedical Informatics and Medical Education
University of Washington School of Medicine
Box 358047
Seattle, WA, 98109
United States
Phone: 1 206 221 3369
Email: atchen@uw.edu

Abstract

Background: Existing literature shows that persons with substance use disorder (SUD) experience different stages of readiness to reduce or abstain from substance use, and tailoring intervention change strategies to these stages may facilitate recovery. Moreover, stigma may serve as a barrier to recovery by preventing persons with SUDs from seeking treatment. In recent years, the behavior change technique (BCT) taxonomy has increasingly become useful for identifying potential efficacious intervention components; however, prior literature has not addressed the extent to which these techniques may naturally be used to recover from substance use, and knowledge of this may be useful in the design of future interventions.

Objective: We take a three-step approach to identifying strategies to facilitate substance use recovery: (1) characterizing the extent to which stages of change are expressed in social media data, (2) identifying BCTs used by persons at different stages of change, and (3) exploring the role that stigma plays in recovery journeys.

Methods: We collected discussion posts from Reddit, a popular social networking site, and identified subreddits or discussion forums about 3 substances (alcohol, cannabis, and opioids). We then performed qualitative data analysis using a hybrid inductive-deductive method to identify the stages of change in social media authors' recovery journeys, the techniques that social media content authors used as they sought to quit substance use, and the role that stigma played in social media authors' recovery journeys.

Results: We examined 748 posts pertaining to 3 substances: alcohol (n=316, 42.2%), cannabis (n=335, 44.8%), and opioids (n=135, 18%). Social media content representing the different stages of change was observed, with the majority (472/748, 63.1%) of narratives representing the *action* stage. In total, 11 categories of BCTs were identified. There were similarities in BCT use across *precontemplation*, *contemplation*, and *preparation* stages, with social support seeking and awareness of natural consequences being the most common. As people sought to quit or reduce their use of substances (*action* stage), we observed a variety of BCTs, such as the repetition and substitution of healthful behaviors and monitoring and receiving feedback on their own behavior. In the *maintenance* stage, reports of diverse BCTs continue to be frequent, but offers of social support also become more common than in previous stages. Stigma was present throughout all stages. We present 5 major themes pertaining to the manifestation of stigma.

Conclusions: Patterns of BCT use and stigmatizing experiences are frequently discussed in social media, which can be leveraged to better understand the natural course of recovery from SUD and how interventions might facilitate recovery from substance use. It may be important to incorporate stigma reduction across all stages of the recovery journey.

XSL•FO
RenderX

(JMIR Form Res 2025;9:e57468) doi:10.2196/57468

KEYWORDS

stigma; substance use; transtheoretical model; behavior change techniques; social media

Introduction

Background

As of 2022, among individuals aged ≥12 years in the United States, 17.3%, or 48.7 million people, met the diagnostic criteria for substance use disorder (SUD) in the last year [1]. Considering individuals needing treatment for SUD in the past year, it is estimated that only approximately one-fourth received treatment [2].

SUD is a highly stigmatized condition [3], and there is considerable evidence that individuals experiencing SUD are faced with stigmatizing experiences in their communities, at work, and in health care contexts [4]. Because stigma substantially worsens outcomes for individuals with SUD [5], mitigating stigma's effects is important in improving outcomes for individuals experiencing SUD.

A number of models have been developed to understand health behavior change [6]. In the context of SUD and addiction recovery, the most widely used theory is the transtheoretical model (TTM) [7]. The TTM was originally developed in the 1980s in the context of designing smoking cessation interventions and remains a dominant model for conceptualizing recovery phases [8]. The theory posits that individuals experiencing SUD can vary in their readiness to change a given health-related behavior and that determining a person's "change readiness" can facilitate the choice of an appropriate intervention. As such, successful behavior change is a *process* that takes place over time, rather than a single decision at a specific point in time. According to the TTM, an individual experiencing SUD can be located in 1 of 6 stages: *precontemplation* (ie, not intending to make a behavior change within the next 6 months), *contemplation* (ie, intending to make a change within the next 6 months), *preparation* (ie, intending to take action within the next 30 days), *action* (ie, making active, sustained attempts at behavior change for less than 6 months), *maintenance* (ie, keeping up a sustained behavior change for more than 6 months), and *termination* (ie, completed behavior change, with no desire to return to the previous behavior). In addition to the stages of change, the TTM argues that particular processes of change are effective for individuals at different stages, thus serving as recommendations for intervention design.

However, in recent years, there has been increased awareness of the need to better understand behavioral interventions, particularly due to the rapid evolution of technologies that are often used to deliver interventions [9]. The behavior change technique (BCT) taxonomy [10] is a resource that has been developed to serve as a method for specifying interventions and the components that account for their effectiveness. The taxonomy includes a standardized set of 93 BCTs proposed to be "active ingredients" of interventions (eg, personalized feedback about health behaviors, self-monitoring of symptoms and behaviors, and reinforcement of positive change).

Social media and other forms of consumer-generated textual data are a widely used resource for understanding health behaviors and serve as a useful complement to more established data collection methods (eg, surveys and qualitative interviews) [11,12]. A key advantage of social media as a source of data is that they provide spontaneous, naturalistic, unmediated, and first-person accounts of individual user's feelings and attitudes, which can provide insights that more hypothesis-driven data collection strategies may not necessarily capture [13].

Reddit is a large, predominantly text-based social media site organized into subreddits (ie, forums centered on a common theme or topic) in which users pseudonymously create posts and respond to other users' posts. Compared to short text microblog services such as Twitter (rebranded to X), video-based services (eg, YouTube and TikTok), and image-based services (eg, Pinterest and Instagram), Reddit encourages a more verbose, discursive style of interaction, often including rich, detailed narratives that share users' experiences, foregrounding the voices and experiences of individuals living with SUD. As such, Reddit has been extensively used as a resource for investigating substance use in the context of smoking cessation and vaping [14-16], cannabis use [16-18], alcohol use [19,20], and opioid use [21,22].

Objectives

With this paper, we aimed to build on existing work that uses Reddit data to explore social media users' experiences of using (and ceasing to use) alcohol, cannabis, and opioids. We selected these 3 substances because public perception differs for different substances [23]. The stigma associated with the use of a substance can vary due to factors such as perceived dangerousness, social acceptance, and legality [24], and these substances may complement one another in terms of the insights offered with respect to people's use experiences and recovery strategies. Moreover, alcohol and cannabis are among the most widely used substances in the United States [25], and there was widespread concern about the opioid crisis during the study period [26,27]. We used qualitative data analysis methods to characterize the behavior change strategies that may be used at different stages of the TTM and explored a challenge that Reddit authors report experiencing throughout the stages—stigma relating to substance use. Our investigation was framed around 3 research questions (RQs):

- RQ1: What are the manifestations of the stages of the TTM among Reddit authors?
- RQ2: Which BCTs emerge organically at different stages of the TTM?
- RQ3: How do participants experience stigma, and how do these experiences of stigma affect recovery?

XSL•FO
RenderX

Methods

Data Collection and Preparation

The data collection and preparation process included multiple phases (Figure 1). In phase 1, as described in a previous study [28], we collected data from different subreddits relating to alcohol, cannabis, or opioid use (N=163,662 posts) using the pushshift.io application programming interface [29]. In the selection of subreddits, we sought to be inclusive of orientation toward use, including those more oriented toward recovery (eg, r/stopdrinking, r/leaves, and r/OpiatesRecovery) and those adopted various other positions regarding consumption and use (eg, r/alcohol, r/cripplingalcoholism, r/Marijuana, r/Petioles, r/trees, r/cannabis, r/opiates). The posts were authored between January 1, 2013, and December 31, 2019. Data collected from subreddits only included "initiating" posts, as previous work has reported that these posts tend to contain richer narratives than replies to the initiating post [30].

Figure 1. Data preparation flowchart illustrating the process of identifying subreddit posts containing narratives of stigma and recovery related to alcohol, opioid, and cannabis use. BCT: behavior change technique; TTM: transtheoretical model.

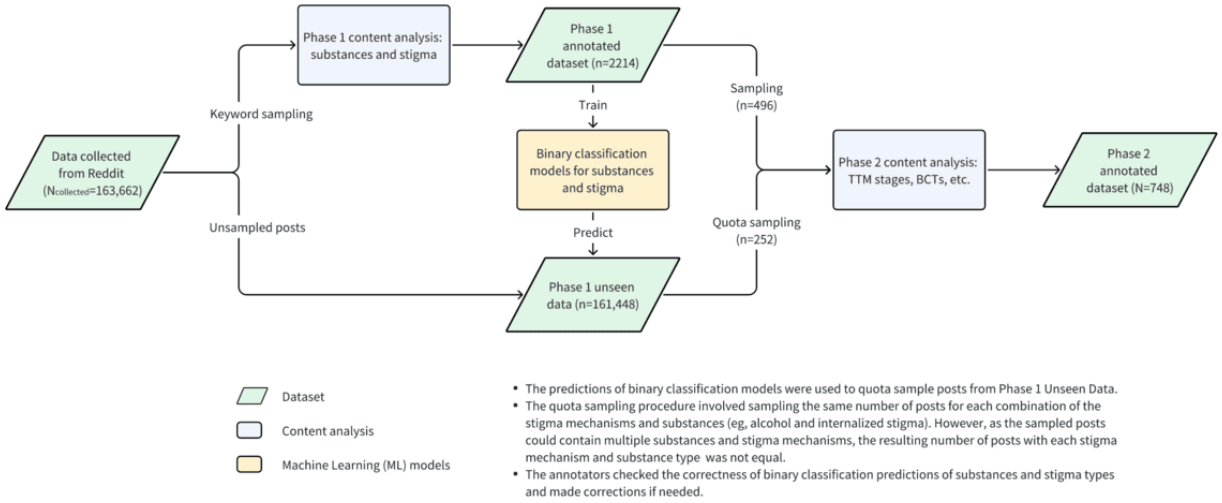

We then performed content analysis on a subset of the corpus (2214/163,662, 1.35%) sampled using keywords. Given the salience of stigma in substance use recovery, posts were eligible if they contained keywords suggesting that stigma might be present, including adjectives that are highly indicative of stigma (eg, "shame," "untrustworthy," and "disappoint"), pejorative labels (eg, "crackhead," "junkie," and "alcoholic"), and references to persons who might be involved in stigma-related experiences (eg, partners, parents, and coworkers). In phase 1 content analysis, we annotated the substances used and stigma mechanisms present (ie, internalized, anticipated, and enacted stigma). These procedures are described in detail in a prior study [31].

Building on this foundation, we developed a set of binary classification models using the phase 1 annotated dataset as training data. The models use machine learning (ML) algorithms to differentiate between the presence versus absence of substance use and stigma mechanisms. Architecturally, the models combine an autoencoder RoBERTa framework [32] and a multilayer perceptron that takes features that may be informative of stigma as input. The features include Term Frequency-Inverse Document Frequency weighted n-grams (ie, the n-grams are weighted based on their importance in a given post, relative to the corpus), affective and psychological lexical features, and additional handcrafted features. A detailed description of the development and evaluation of the classifiers is available in a prior study [33]. These models were then used to generate ML predictions of substances and stigma on the unseen data (161,448/163,662, 98.65%).

In phase 2, we performed qualitative data analysis to delve deeper into Reddit authors' experiences of stigma and the behaviors they engaged in while seeking to recover from substance use. To enrich the concentration of stigma in our sample, we sampled the first 496 posts in the phase 1 annotated dataset consisting of posts with and without stigma and combined this with another sample of 252 posts selected from the unseen data. Post selection was performed using the previously mentioned ML predictions of stigma and substances, with the goal of sampling equal proportions of the 3 stigma mechanisms and the 3 substances. Using this procedure, we were able to ensure that the resulting dataset for this study (N=748) contained sufficient posts relating to the 3 substances and a varied distribution of posts with and without stigma for subsequent qualitative analyses.

Data Analysis

We performed a qualitative analysis of the posts using a hybrid inductive-deductive method, which seeks to leverage the strengths of each respective approach [34]. The deductive or "top-down" part of our analysis was based on 2 theoretical frameworks and a taxonomy: the stigma framework [31,35], the stages of change from the TTM [7], and BCTs as defined in the behavior change taxonomy [10]. We also included the substances of interest.

As noted by Proudfoot [34], using a hybrid inductive-deductive approach to analysis "helps to ensure that the voices of the participants are valued, while simultaneously allowing for more theory-led analysis." Given the sensitive nature of our research

XSL•FO
RenderX

focus and the richness of our dataset, adopting an inductive component was particularly important for ensuring that these voices are heard. Thus, our codebook development included an inductive or "bottom-up" component in which we adapted the definitions to our data to account for the specifics of the health behavior context and data type. For example, though the stage of *action* has been defined as having modified their lifestyles within the past 6 months [7], often there is no explicit declaration of time in the social media data. Thus, to differentiate *action* from the next stage in the TTM, *maintenance*, our coding guide specifies that situations where social media authors appeared to be actively seeking to quit substance use are indicative of the *action* stage, whereas situations where users maintain routinized strategies are characteristic of the *maintenance* stage.

In developing codes for BCTs, we focused on the high-level groupings of the taxonomy [10], such as "antecedents," which include more specific techniques, such as "restructuring the physical environment" and "restructuring the social environment." Though the BCT taxonomy does not differentiate between social support seeking and offering, we chose to do so due to the diverse ways that social support has been studied in the literature [36] and the potential importance of differentiating support sought and received in terms of health outcomes [37]. The taxonomy includes 16 high-level groups; we only included a subset of these because the others did not appear salient in our data.

We also coded the substances of interest (eg, alcohol, cannabis, and opioids) and the stigma mechanisms encountered, identified either through qualitative coding or using the models and checked by the annotators, as described previously. The stigma mechanisms—enacted, anticipated, and internalized stigma—were adapted from the stigma framework [31,35]. Enacted stigma describes experiences of stereotyping, prejudice, and discrimination due to a stigmatized attribute; anticipated stigma is the expectation that one might experience stereotyping, prejudice, and discrimination in the future due to a stigmatized attribute; and internalized stigma is the application of negative feelings and beliefs about a stigmatized group to oneself (Table 1) [38]. For the stigma mechanisms, we also identified the actors involved (eg, family, friends, partners, coworkers, others, society, and self).

Our annotation process involved 2 primary coders (SJ and RKC), who independently coded the discussion posts; an adjudicator (SHW), who reviewed the codes and facilitated reconciliation; and 1 additional investigator (ATC), who provided additional feedback as needed. All disagreements were reconciled by majority or consensus. Interrater reliability for the substance and stigma codes are reported in a prior study [33]. For the stages of the TTM and BCTs, we assessed interrater reliability using Cohen κ [39]; overall interrater reliability was 0.71 between the 2 coders. Dedoose qualitative data analysis software was used to perform the coding [40]. For each research question, we report our findings using descriptive statistics as well as narrative descriptions of each theme.

Table 1. The stigma framework: definitions and examples.

	Definition	Example
Enacted stigma	Experiences of stereotyping, prejudice, and discrimination from others due to a stigmatized attribute.	"My husband called me an addict and said I'd never become clean."
Anticipated stigma	The expectation that one might experience stereotyping, prejudice, and discrimination from others due to a stigmatized attribute.	"My co-workers are giving me looks now at work after I went in drunk last week. I don't know how I'm going to face them day after day."
Internalized stigma	Individuals internalize negative stereotypes about the stigmatized group that they belong to and apply these to themselves, affecting their sense of self-worth.	"I'm a stoner. I am an awful person..."

Ethical Considerations

Social media can offer naturalistic, first-person accounts of individual user's feelings and attitudes, allowing for understandings that hypothesis-driven data collection strategies may not necessarily capture [13]. Moreover, prior literature has reported that social media authors often believe social media content is public domain and can be freely used in research without consent as long as individual information is not reported [41].

However, particular care must be taken to protect the identities and ensure the anonymity of content authors [42]. In this study, we analyzed the collected data without attempting to contact or connect social media authors to their posts; then, we used synthetic or paraphrased quotations based on the original social media content to illustrate the themes we present. As there are situations in which search engines might be used to retrieve social media posts [42], we constructed synthetic quotations to protect the identities of content authors.

In the development of these synthetic quotations, we seek to faithfully represent the voices of social media authors while protecting their identities by selecting original quotations that are representative of the data [43], faithful representation of concepts in the synthesized quotations, and using language that resembles language in the source medium to maximize authenticity. As we did not seek to connect social media authors to their posts, we also did not compensate them. This study has received ethics approval from both the University of Washington Human Subjects Division (STUDY00015737) and the University of Melbourne Human Research Ethics Committee (2023-25512-48127).

Reflexivity Statement

Our analysis incorporates many recommended practices for establishing rigor and trustworthiness in qualitative research

XSL•FO
RenderX

[44]. We ensured thick, rich descriptions using a sampling strategy that leveraged the large volume of available data while maintaining our specific research focus, and through a method of synthetic quotation creation. We performed *triangulation* during the data analysis and reporting process through the use of multiple coders of different backgrounds, including those with clinical experience (SJ and RKC) and informatics experience (ATC and SHW) and an investigator with significant experience in qualitative research, analysis of social media data, and the lived experiences of persons with chronic conditions (ATC). An important part of the annotation and synthetic quotation creation process involved discussion among the research team to ensure consistent coding over time, faithfulness to the data, and contextualization of the data considering the complementary backgrounds of the researchers. The discussion had several purposes, including consideration of different perspectives and mitigating the potential bias of any 1 researcher. Last but not least, our procedures facilitate *auditability* through detailed notes on our data preparation process and annotation process.

Results

Sample

Our sample comprised 748 posts pertaining to 3 substances: alcohol (n=316, 42.2%), cannabis (n=335, 44.8%), and opioids (n=135, 18%). Each post relates to at least 1 substance. On average, posts were 348 (SD 315) words long. Internalized stigma (369/748, 49.3%) was the most commonly expressed across all stages, but anticipated (131/748, 17.5%) and enacted stigma (111/748, 14.8%) were also present.

RQ1: How and to What Extent Are the Posts Relating to Each Stage Represented in the Sample?

The posts reflected different stages of change (Table 2). A subset of the posts was in the *precontemplation* stage, when social media authors do not express an intent to quit. These could include posts in which people are simply talking about experiences that they have had relating to substance use, which could include legal consequences, trouble at work or home, and more, but do not demonstrate an intention to quit. Often in the *precontemplation* stage, internalized stigma manifested as hopelessness or existential questions:

> *I feel horrible about my life every day. I feel bad about not taking care of my kids but I just keep picking up the bottle.... My life is hopeless.*
>
> *Am I depressed because I am a drunk or a drunk because I am depressed?*

In the *contemplation* stage, social media authors begin thinking about quitting but may identify barriers to doing so:

> *I'm thinking about quitting again, but I can't get time off of work to go into a detox program...*

They may also reach out to the Reddit community to ask their opinion about whether to quit, how to do so, and other related questions.

In the *preparation* stage, social media authors report taking different steps to quit, including deleting the phone numbers of dealers, planning to enroll in rehabilitation programs, telling friends and family about their use to seek support in the quitting process, and posting on Reddit to keep themselves "accountable." Many also mention that it is time to stop. At times, there is an event that may trigger the *preparation* stage, such as such as running out of the substance being used or being discovered by parents. Throughout the stages, feelings of depression, despair, and hopelessness are often expressed.

In the data that we annotated, the most commonly represented stage was *action*. In these posts, Reddit authors often shared their experiences with their recovery journeys. These posts often included rich detail concerning challenges that they faced:

> *I have been thinking about attending meetings but am also worried that the experiences of those worse off will cause me to slip.*
>
> *The holidays are the worst. I keep thinking that I'll just have a drink but I know that won't happen.*

In the *maintenance* stage, social media authors often share how long they have been in recovery and seek to encourage others by sharing their experiences:

> *This morning, I played with my son. I went to the park with him and enjoyed the moment with a clarity that I would have thought impossible a year ago. I'm so grateful that I'm now here. If you are struggling, take it one day at a time. It will get better.*

However, not all the narratives in the *maintenance* stage express positive sentiment. Some express continued cravings, feeling like they might relapse, and problems experienced while sober. While some say that they do not feel tempted at all, others say that they do not feel that they will relapse but sometimes reminisce about the effect of their substance of choice:

> *While I don't think I'm in any danger of relapsing, sometimes I feel the call of sweet mary jane and look back fondly.*

According to the TTM, people often slip back to prior stages. In the social media content, people often reported situations in which they had been sober for some time and then relapsed or are on the point of relapse:

> *Five months sober. Last month I decided I could have a drink while I was out with my coworkers, and it's been downhill from there. I have to stop!*
>
> *Feeling like I might relapse. I have been sober for a year and two months, but my life is as messed up now as before. I don't feel any better, and now I have no one to talk to. I don't want to start over again and feel the high heart rate, paranoia, guilt, and lack of motivation all over again, but the desire to smoke is overwhelming.*

As presented in Table 2, not all the data were representative of a stage of change. This could happen for various reasons. For example, a person might write a post providing advice but not clearly indicate that they were in recovery or provide commentary on relevant material, such as policy, a movie, or book on how to quit.

XSL•FO
RenderX

Table 2. Distribution of the stages of change in the dataset (N=748).

Stage	Posts, n (%)
Precontemplation	95 (12.7)
Contemplation	68 (9.1)
Preparation	43 (5.7)
Action	472 (63.1)
Maintenance	45 (6)
Unknown	25 (3.3)

RQ2: What BCTs Do People Naturally Use as They Seek to Recover From Substance Use in the Different Stages?

In Textbox 1, we present the BCTs observed, along with their definitions and examples. In Figure 2, we depict the prevalence of BCTs observed in the TTM stages, with the length of each bar reflecting the number of posts in which a given BCT was observed. The frequency of BCTs varied for the different stages of change. There were similarities in the *precontemplation*, *contemplation*, and *preparation* stages, with social support seeking and awareness of natural consequences being the most common:

> *Reddit, help! I have lost everything because of my habit... dropped out of college and am living at home with my parents. I feel like a failure. I need help but I don't know how to start.* [Social support seeking and natural consequences]

Posters reported that they sought social support online, in online support groups and other digital and web-based formats such as subreddits, as well as from family, friends, and meetings such as Alcoholics Anonymous.

In the *contemplation* stage, compared to the *precontemplation* stage, we observed a greater proportion of posts including some element of *shaping knowledge* or the acquisition of knowledge and information related to the antecedents of the behavior and ways to change it. At this time, social media authors expressed increased awareness of reasons or triggers for use, including loneliness, boredom, negative thoughts, and daily routines or life situations that make post authors more prone to use:

> *I want to quit and am looking for any advice you all have. I've been smoking for years and recently I've noticed all these changes in me that make me want to quit. First, I am anxious all of the time now and I can't sleep. I have no money and weed is draining my finances. I am also using weed as a coping mechanism for stress, which only makes me more stressed.* [Shaping knowledge and natural consequences]

In the *preparation* stage, we observed a greater proportion of goals and planning, such as seeking out accountability partners and making accountability posts. Post authors also express commitment to quitting or plans to enroll in treatment or detox programs.

In the *action* stage, as people sought to quit their use of substances, we observed a variety of BCTs. For example, by recognizing that merely having access to a substance would facilitate use, social media authors threw away alcohol and equipment that they used to engage in use, such as pipes. They also enlisted the help of others to ensure that they did not have access to substances and, at times, relocated (antecedents):

> *Gave away all of the wine in the house.* [Antecedents]
>
> *Put what little I had in my Civic and drove as far as I could south, that was 2 months ago and I haven't used since, aside from a sub or two for withdrawals. I lost everything when I was using—my friends, my job, my savings, my dignity. No more. Applying to jobs in this new town and planning to play pickup this evening at the local gym. It's time to rebuild.* [Antecedents, natural consequences, and goals and planning]

They also sought to engage in repetition and substitution of healthful behaviors, such as exercising while monitoring their own progress (feedback and monitoring):

> *I've started working out regularly and it has helped with the anxiety...* [Repetition and substitution]
>
> *Six days down! I hate what weed has done to me. I am always in a fog and can't remember what happened yesterday. My heart races. But I am slowly feeling better and clearer. I'm learning what withdrawal is like, trying to give space for my feelings, and gradually getting through it.* [Feedback and monitoring]
>
> *It has been a month since I last smoked. At first, I had a hard time sleeping and was tired all the time. But each day gets better. I have a little more energy. I can think more clearly.* [Feedback and monitoring]

XSL•FO
RenderX

Textbox 1. Behavior change technique (BCT) groupings, definitions, and examples (Adapted from the behavior change taxonomy [10]).

Natural consequences

- Definition: actions related to the identification of consequences of a behavior to be used as a facilitator of behavior change.
- Identified consequences should not be future, anticipated consequences. Only consequences that were experienced in the past or experienced in the present should be coded.
- Examples: becoming aware of health, social, environmental, and emotional consequences.

Antecedents

- Definition: actions that are related to reducing or mitigating antecedents (preconditions) that might lead to the undesired behavior.
- Examples: restructuring the physical and social environment (eg, moving to a new place) and avoidance or reducing exposure to behavioral cues, such as others who engage in substance use.

Shaping knowledge

- Definition: actions related to the accumulation of knowledge and information related to the behavior and ways to change it.
- Examples: information about antecedents and instructions on how to perform a behavior.

Repetition and substitution

- Definition: actions related to substituting the unwanted behavior with another and rehearsing the performance of a desired behavior.
- Examples: behavioral practice or rehearsal, behavioral substitution, and habit reversal.

Regulation

- Definition: actions related to emotional regulation and managing mental and cognitive stress.
- Examples: reducing negative emotions.

Goals and planning

- Definition: the identification of behaviors to be changed, creation of goals, and development of plans to facilitate change.
- Examples: reviewing goals and finding accountability partners.

Social support

- Definition: actions related to seeking, receiving, or providing social support to assist with behavior change. In this study, we used a broader conceptualization of social support, encompassing both emotional and informational support.
- Examples: general (unspecified), seeking, received, and providing and offering.

Comparison of outcomes

- Definition: the identification of potential or realized outcomes as a way to encourage or motivate behavior change.
- Examples: listing pros and cons and persuasive argument.

Identity

- Definition: shifting one's perspective on their past or current identity or actions to encourage behavior change.
- Examples: framing and reframing.

Self-belief

- Definition: increasing one's confidence in successfully performing the desired behavior through reinforcing one's belief in oneself.
- Examples: mental rehearsal of successful performance, focus on past success, and self-talk.

Feedback and monitoring

- Definition: monitoring one's behavior and the effects of one's behavior to facilitate the behavior change.
- Examples: self-monitoring of behavior and its outcomes.

XSL•FO
RenderX

Figure 2. The use of behavior change techniques (BCTs) and the presence of stigma mechanisms in individual stages of change.

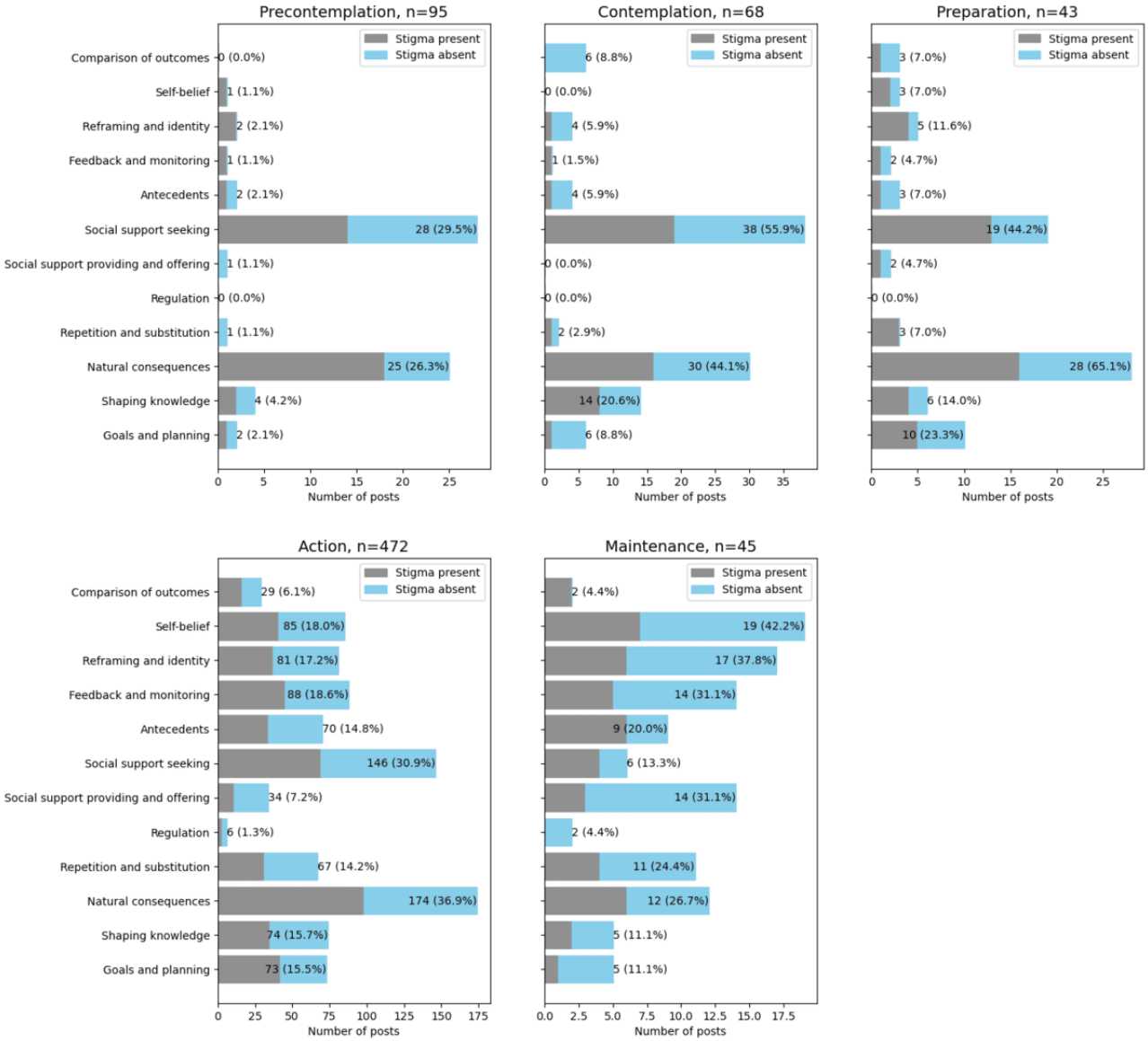

Social support is an important part of the quitting process. Social media authors can experience challenges in removing antecedents (ie, triggers) from their environment if others are not supportive:

> *I have been trying to quit drinking, but a close friend insists on going to the bar whenever we hang out. Every time we go I have a hard time stopping myself from ordering a drink. When I bring this up, he says that I am just weak. What can I do?* [Shaping knowledge, antecedents, and social support seeking]

Social media authors also sought information from others as to how to maintain the new habits that they tried to adopt, while addressing symptoms that they experience in withdrawal:

> *The cravings are bad, but what is worse is not being able to sleep. Does that happen to you all? How long do the symptoms last?* [Social support seeking]

We also observed examples of positive self-talk and reframing, in which participants adopted or attempted to adopt new perspectives:

> *I've reached the two week mark. It hasn't been easy, but I'm proud of myself, and hopeful that I can do it.* [Self-belief]

> *I went to a party and for once, I didn't make a fool of myself! I had a great time, and was able to focus on every person I talked to...* [Self-belief]

In the *maintenance* stage, reports of diverse BCTs continue to be frequent. Notably, expressions of self-belief, altered perspective (ie, reframing and identity), and providing support and encouragement to others (ie, social support providing and offering) also become common:

> *Today I live without shame and look forward to the future. To everyone still struggling out there—you can do it!* [Social support providing and offering]

Many BCTs that became prominent in the *action* stage, such as self-belief, reframing and identity, feedback and monitoring, and repetition and substitution, remained common or were even observed more often in the *maintenance* stage. While this may partially be due to the nature of the venue—being an online

XSL•FO
RenderX

community in which people may strive to be positive to encourage others—this may also underscore the importance of continual reaffirmation and efforts at abstinence:

> *When I think of my past, I still feel the guilt and shame. But I think of them as a good reminder of what I don't want to go back to. I also tell myself that my past doesn't have to define me.* [Reframing and identity]

As we reflect back on the BCTs used, we observed that there was diversity in the strategies used. For many, it was a combination of BCTs and social support:

> *I told my partner last night about everything. She was incredibly supportive, and we talked about a plan. I am going to restart therapy. I am worried that I'll start using again, but for the first time in a long time, I feel hopeful.* [Goals and planning, social support seeking, and social support received]

As the previous excerpt demonstrates, this process can be a gradual one. While this poster states that others are supportive of their efforts, this is not always the case:

> *I am in rehab and on subs. I am trying hard to stay clean, but I still smoke weed. I am trying to be a good mother to my kids, but I feel so alone. I have no friends left. I had a friend that I kept up with, but it was too hard to stop using while continuing to spend time with them. I don't know what to do and would love to have some support.* [Antecedents and social support seeking]

Social media authors may experience challenges due to their life contexts, including feelings of guilt and other pressures due to family obligations. They may also face challenges dealing with exposure cues in their day-to-day lives. In addition, people may struggle with stress, self-doubt, and loneliness, leading them to ask for others' opinions on how to deal with these hardships.

RQ 3: How Do Participants Experience Stigma, and How Do These Experiences Affect Recovery?

Overview

In this research question, we aimed to examine how the presence of stigma may affect social media authors' recovery journeys. In Figure 2, we observed that substantive amounts of stigma (ie, in blue) are present in all stages. In this section, we review key themes involving stigma, which were often observed in multiple stages. We also present stage-specific characterizations as appropriate.

Theme 1: Experiencing Enacted Stigma From Others

Social media authors reported experiences of enacted stigma involving different persons that they encounter. Of the different actors that we examined, enacted stigma from family and those who were not part of a person's social circle was most common. Authors mentioned experiences with pharmacists and physicians discounting health issues, attributing health issues to substance use, and being reticent to provide services when they were aware that participants engaged in substance use. In addition, Reddit authors often spoke of how their use took a toll on their relationships with family, friends, and partners, resulting in broken trust and interpersonal conflict:

> *I got into a huge fight with my parents. They called me a drug addict and made me feel worthless.* [Enacted stigma]

One scenario that occasionally arose as social media authors sought to quit substance use was resistance from their social circle. For example, there were examples of those who told their family they wanted to quit drinking, and their family would say that they were not an "alcoholic" or "You're not that far gone." Social media authors often had to find new support networks, such as Alcoholics Anonymous, as their family and friends were not always supportive of their efforts to quit.

Theme 2: Experiencing Anticipated Stigma

Social media authors also experienced anticipated stigma. Many feared disappointing others, particularly family and friends, or feared repercussions at work, and thus hid their substance use. As many keep their substance use a secret, they may reach out to the Reddit community to ask questions that they are not able to ask elsewhere, often using temporary (ie, "throwaway") accounts:

> *This is a throwaway account. I have been addicted to oxy for years and am worried I will lose my job if anyone finds out. I am thinking of taking suboxone but am wondering if it would be better to quit cold turkey or would the withdrawals be worse?* [Anticipated stigma]

Some authors reach out on Reddit asking whether they should disclose their use to others around them. There are also times when they do choose to do so and are surprised to find that others are empathetic and supportive:

> *I have been feeling so ashamed and guilty, but I finally told my parents about my drinking and they were surprisingly supportive. I am beginning a rehab program tomorrow.* [Goals and planning, internalized stigma, and anticipated stigma]

Similar to enacted stigma, Reddit authors most often anticipated stigma from family and others outside of their social circle.

Theme 3: The Challenges of Internalized Stigma

There were posts that expressed internalized stigma, including internal struggles that people experienced due to regret, guilt, shame, and disappointment:

> *I have been able to avoid drinking when others are, but my biggest challenge is when I am alone and start to have negative thoughts and sink into this pool of shame and disappointment. Then I think about what a failure I am and how I have let everyone down.* [Internalized stigma]

Social media authors often shared challenges that they experienced adjunctive to substance use, such as financial difficulties, poor mental health, and increased social isolation. In the *contemplation* stage, social media authors often speak of "hating themselves" and hesitate to take steps forward due to

XSL•FO
RenderX

lack of confidence in their own ability to succeed. In contrast, in the *maintenance* stage, social media authors share that they feel less guilt and have more compassion for themselves:

> *I always felt this deep self-loathing each time I slipped. I didn't think I would ever make it. But little by little, I am learning to have compassion for myself. Life is still hard, but it gets a little easier one day at a time.* [Internalized stigma]

The process of recovery was gradual, and there could be setbacks:

> *I started out the year in constant shame. I was taking a quick sip whenever I had the chance, and then worrying that I was messing up my child. My mom, who is an alcoholic, came to visit, setting off a three-day bender. Now I am back on track. I haven't had a drop for a week and I haven't even been tempted. But I have some stressful stuff coming up at work and am worried. Thanks for listening.* [Feedback and monitoring, natural consequences, and internalized stigma]

As these posts illustrate, there can be triggers due to work and interpersonal interactions, and people may try for some time, intermittently experiencing setbacks, before being able to sustain abstinence.

Theme 4: Multiple Forms of Stigma

There were also examples of multiple forms of stigma in the same post. Shame, guilt from letting others down, and broken trust leading to tolls on relationships with others were common challenges:

> *I have lost family and friends due to my drug use. My relatives hide their wallets whenever I go over. I feel terrible that no one trusts me, and the worst failure ever.* [Enacted stigma and internalized stigma]

In earlier stages, people often report hiding their substance use from others and experiencing feelings of negative self-worth from doing so:

> *I hide my stash from my husband. The other day my five-year old accidentally found it and I was mortified. I feel horrible about hiding this but I don't think I can stop...* [Anticipated stigma and internalized stigma]

As illustrated by the quotation above, people may feel low self-worth. As a result, they may also avoid others or be excluded by others (enacted stigma), resulting in social isolation.

Theme 5: Enacted Stigma Directed Toward Others

Social media authors also demonstrated enacted stigma toward others. One common situation was when they referred to family members using pejorative labels such as "addicts" or "alcoholics," which occurred throughout the stages. Another example was a lack of empathy for others among some who felt that they were succeeding in their recovery journeys (ie, often occurring in the *action* or *maintenance* stages):

> *As I stopped smoking, I came to look down upon the behavior of my stoner friends...* [Enacted stigma]

In moments of introspection, some individuals reflected on their own behavior, seeking input from fellow Redditors while contemplating their motivations. For example, 1 social media author mentioned that they wanted to help others who were struggling but also felt it is impossible, and so instead chose to distance themselves.

Discussion

Principal Findings

"In this study, we performed qualitative data analysis to identify the stages of change and BCTs used by social media authors in various stages of readiness to reduce their use of the following substances: alcohol, cannabis, and opioids. There were some similarities in the use of BCTs in the earlier stages of change (ie, *precontemplation*, *contemplation*, and *preparation*). In posts with content that reflected the *contemplation* stage, authors described an awareness of the negative impacts that substance use had on their lives (ie, natural consequences), and as they prepared to quit, they increasingly took concrete steps to do so by sharing their decision with their social network and committing to it (ie, social support seeking, goals, and planning) and seeking help in other ways.

Posts in the *action* stage reflected that individuals leveraged other BCTs, such as modifying their daily routines (ie, repetition and substitution), modifying their physical and social environment (ie, antecedents), self-affirmation (ie, self-belief), and viewing their situation through new lenses (ie, reframing and identity). Substantive use of BCTs are observed, and in some cases more prominently, in the *maintenance* stage. This may partly stem from the nature of online discussions, where many individuals aim to give back to the community by sharing their experiences with successful health behavior change, offering encouragement to those trying to quit.

We also examined the role of stigma in social media authors' recovery journeys. We observed substantive amounts of stigma across all stages of change. The influence of social interactions was also salient. Social media authors shared situations in which they experienced or anticipated stigma from strangers, acquaintances, and loved ones. When people receive support from others, this can be affirming and facilitate recovery; however, social isolation can also result from isolating oneself or from enacted stigma. Some social media authors reported that the fear of disappointing others prevented them from reaching out for support. In contrast, other authors sought support from both online and offline social networks (eg, family and friends) and reported that this support was instrumental in successful recovery journeys. Previous research has found that perceived stigma is associated with poorer mental health, while perceived social support is linked to better mental health [45].

Comparison With Prior Work

This study extends the existing literature concerning recovery mechanisms by identifying situations in which commonly used BCTs appeared in social media narratives, thus affording insight into how people may seek to address the challenges of recovery in everyday life. Consistent with existing literature, which has identified controlling one's environment and social support as

XSL•FO
RenderX

commonly used strategies for behavioral change in the context of substance use [46,47], we observed examples of persons seeking the support of others (*social support seeking*) and using *antecedents*. These including discarding equipment necessary for substance use behavior and physically removing themselves from temptations. However, posters also lamented the difficulties of encountering "triggers" or cues in their day-to-day life that provoked relapse. The use of the antecedents BCT is also common in evidence-based substance use treatments [48], and a prior systematic review found that remotely delivered substance use interventions [49] incorporating intervention components related to "avoidance/reducing exposure to cues for behavior," "pros and cons," and "self-monitoring of behavior" may be effective. Social media authors also sought to restructure their lives to mitigate exposures and observed the effects of behavioral change.

In this study, we also observed that self-belief and cognitive reframing (changing one's perspective), were more common in the *action* and *maintenance* stages. As recovery trajectories progress, there is a trend toward individualistic and agentic identity [50]; BCTs such as self-belief and cognitive reframing may facilitate this process.

Our findings also illustrate the importance of temporality and flux in the experience of the person in recovery, who is not static but rather evolving due to their experiences over time. At times, people may struggle with low self-worth due to internalized and anticipated stigma, impeding their path to recovery; at other times, one might also observe enacted stigma from them toward someone else, often when the person is in recovery. There is also a need for more research concerning how family and romantic partners may also enact stigma [51].

Practice Implications

These findings have implications for the design of stigma reduction interventions to promote substance use recovery. Textbox 2 summarizes the key themes identified, along with implications for subsequent intervention design. For instance, to counter stigma and feelings of isolation, interventions could provide education to improve self-advocacy and emotion regulation. In particular, treatments such as acceptance and commitment therapy could potentially combat internalized stigma by assisting individuals with acceptance and compassion toward the self [52-54].

Textbox 2. Practice implications considering the role of stigma in recovery journeys.

Experiencing enacted stigma from others

- Provide persons with substance use disorder (SUDs) training in navigating challenging situations involving enacted stigma (eg, in self-advocacy with providers and emotion regulation in everyday situations).
- Assist persons with SUDs to develop alternative support networks and other mechanisms to nurture resilience.

Experiencing anticipated stigma

- Promote the use of behavior change techniques (BCTs) encouraging persons with SUDs to develop and strengthen support mechanisms.
- Teach strategies for disclosure of topics related to one's SUD.

The challenges of internalized stigma

- Emphasize self-compassion.
- The internal struggles and gradual process were significant obstacles. Intervention components that serve as a toolkit for participants to combat these challenges, such as feedback and monitoring and comparison of outcomes, can perhaps bolster natural mechanisms.

Multiple forms of stigma

- Use approaches that address the different stigma mechanisms individually but also emphasize that the process of recovery involves rebuilding on multiple fronts.

Enacted stigma directed toward others

- Assist persons with SUDs to consider the situations of others with an open countenance.

In addition, interventions could encourage the development of new support networks or teach strategies to intentionally seek support from others. Doing so could lessen the impact of the stigma while developing a social network that is supportive of one's recovery. Given that family interactions were a common source of stigma, interventions could teach strategies to encourage positive interactions with family members or facilitate development of additional support networks outside the family as needed. These study findings show that when family members were supportive, participants valued the support of those that they were close to.

Reddit authors also struggled with internalized stigma and challenges involving multiple forms of stigma, in which authors felt bad about their interactions with others. Interventions that assist persons experiencing self-stigma to disclose and seek support from others could be helpful [55].

Given that this study focused on occurrences of stigma expressed by people who were engaged in online forums, it is also worthwhile to consider how these findings have implications for the design of internet-based interventions. Reddit authors often mentioned the improvements that they observed over time. These improvements seemed to increase their confidence and

XSL•FO
RenderX

hope that they would continue to improve. There may be an opportunity here to leverage mobile health apps to facilitate self-monitoring and tracking of incremental improvements. Previous research has also recognized the value of using wearables to monitor biologic data, such as blood alcohol concentration sensors [56], and self-monitoring surveys can help participants reflect on associations between self-monitoring domains and adhere to drug use reduction goals [57]. These are potentially fruitful directions for future research.

The findings from this study could also improve the degree of personalization of content, such as delivering stage-matched intervention content. For example, the results suggest that *shaping knowledge* (information about changing a behavior) is more prominent in the *contemplation* stage as opposed to *precontemplation*, and that people use diverse BCTs as they actively seek to quit substance use. Thus, it may be beneficial to present content that encourages thoughts about antecedents or other methods of shaping knowledge to individuals in the *contemplation* stage, while offering suggestions for different BCTs to those in the *preparation* or *action* stages. In the *maintenance* stage, we observed that some individuals faced the potential for relapse, and other perceived no improvement in their lives. Thus, there may be a need for intervention content tailored to particular challenges of maintenance, such as the position of abstinence as opposed to moderation or maintaining recovery in the context of other aspects of life.

Finally, prior literature has observed that those who engage in social media value the information and support from the online communities [28,36,58,59]. In this study, we observed that social media authors experienced different information and support needs in different phases of their recovery journey. During the *action* phase, social media authors were likely to have questions about the quitting experience, including strategies, withdrawal symptoms, and timing of those symptoms. At all phases, people may experience questions about social interactions, such as "protecting their quit" and changes in how they may view and relate to others as they engage in maintenance. Thus, insights from these study findings can inform the provision of appropriate information at different stages of the recovery journey.

Limitations

This study has various limitations. First, as this study was based on social media data, there is bias as to who is represented due to differences in people's decisions about which media to use [60]. However, as indicated by the sample posts, social media authors shared their challenges with substance use in diverse life contexts, including home, work, and school contexts. For example, some social media authors experienced concern with the effect that their use might have on their children, while others described their fears and anxieties with respect to work repercussions. Thus, the study findings still have implications for the diverse life contexts in which people may confront and seek to apply strategies for recovery.

In addition, this study used sampling approaches that sought to increase the proportion of posts containing stigma in the sample. As such, the posts in our sample may not represent the natural distribution of stages of change in substance use–related content on Reddit. There was less data available on stages other than *action*, and there is a need for more research concerning information and support needs of persons with SUDs in the earlier phases. Finally, in this study, we used content analysis to identify different stages of change to characterize BCT use in the different stages. The study does not examine longitudinal changes in the same person over time, and this could be an area for future exploration.

Conclusions

This study used social media as a naturalistic data source to explore how people may use BCTs to recover from substance use, including how this use may vary depending on a person's readiness to quit or reduce their substance use. We observed that there were differences in the BCTs that were most commonly used in the different stages. Stigma was present in a substantive portion of the narratives at all stages, suggesting that there is a need to incorporate intervention components to assist participants address stigma-related challenges. Consideration of these patterns might be used to personalize and deliver content in future interventions promoting substance use recovery.

Acknowledgments

This study was supported by the National Institute on Drug Abuse of the National Institutes of Health (R21DA056684). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to the potential sensitivity of the social media posts included in our study, to protect social media authors' identities and anonymity.

Authors' Contributions

ATC and JEG conceptualized the study. SJ, SHW, and RKC performed the content analysis. LCW performed the data processing and rendered the data visualizations. ATC, MC, and SHW authored the initial draft. ATC, LCW, and JEG revised the manuscript. All authors provided feedback and approved the submission.

XSL•FO
RenderX

Conflicts of Interest

Pear Therapeutics provided digital therapeutic prescriptions at no cost to Kaiser Permanente Washington during a quality improvement pilot study led by JEG.

References

1. Highlights for the 2022 national survey on drug use and health. Substance Abuse and Mental Health Services Administration. URL: https://www.samhsa.gov/data/sites/default/files/reports/rpt42731/2022-nsduh-main-highlights.pdf [accessed 2024-02-11]
2. SAMHSA announces National Survey on Drug Use and Health (NSDUH) results detailing mental illness and substance use levels in 2021. U.S. Department of Health and Human Services. 2023 Jan 04. URL: https://www.hhs.gov/about/news/2023/01/04/samhsa-announces-national-survey-drug-use-health-results-detailing-mental-illness-substance-use-levels-2021.html [accessed 2023-11-22]
3. Corrigan P, Schomerus G, Shuman V, Kraus D, Perlick D, Harnish A, et al. Developing a research agenda for understanding the stigma of addictions part I: lessons from the mental health stigma literature. Am J Addict 2017 Jan;26(1):59-66. [doi: 10.1111/ajad.12458] [Medline: 27779803]
4. Yang LH, Wong LY, Grivel MM, Hasin DS. Stigma and substance use disorders: an international phenomenon. Curr Opin Psychiatry 2017 Sep;30(5):378-388 [FREE Full text] [doi: 10.1097/YCO.0000000000000351] [Medline: 28700360]
5. van Boekel LC, Brouwers EP, van Weeghel J, Garretsen HF. Stigma among health professionals towards patients with substance use disorders and its consequences for healthcare delivery: systematic review. Drug Alcohol Depend 2013 Jul 01;131(1-2):23-35. [doi: 10.1016/j.drugalcdep.2013.02.018] [Medline: 23490450]
6. Davis R, Campbell R, Hildon Z, Hobbs L, Michie S. Theories of behaviour and behaviour change across the social and behavioural sciences: a scoping review. Health Psychol Rev 2015;9(3):323-344 [FREE Full text] [doi: 10.1080/17437199.2014.941722] [Medline: 25104107]
7. Prochaska JO, Velicer WF. The transtheoretical model of health behavior change. Am J Health Promot 1997;12(1):38-48. [doi: 10.4278/0890-1171-12.1.38] [Medline: 10170434]
8. Enhancing Motivation for Change in Substance Use Disorder Treatment: Updated 2019. Rockville, MD: Substance Abuse and Mental Health Services Administration (US); 2019.
9. Murray E, Hekler EB, Andersson G, Collins LM, Doherty A, Hollis C, et al. Evaluating digital health interventions: key questions and approaches. Am J Prev Med 2016 Nov;51(5):843-851 [FREE Full text] [doi: 10.1016/j.amepre.2016.06.008] [Medline: 27745684]
10. Michie S, Richardson M, Johnston M, Abraham C, Francis J, Hardeman W, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013 Aug;46(1):81-95 [FREE Full text] [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
11. Paul M, Dredze M. You are what you tweet: analyzing Twitter for public health. Proc Int AAAI Conf Web Soc Media 2021 Aug 03;5(1):265-272. [doi: 10.1609/icwsm.v5i1.14137]
12. Amaya A, Bach R, Keusch F, Kreuter F. New data sources in social science research: things to know before working with Reddit data. Soc Sci Comput Rev 2019 Dec 18;39(5):943-960. [doi: 10.1177/0894439319893305]
13. Conway M, O'Connor D. Social media, big data, and mental health: current advances and ethical implications. Curr Opin Psychol 2016 Jun;9:77-82 [FREE Full text] [doi: 10.1016/j.copsyc.2016.01.004] [Medline: 27042689]
14. Struik L, Yang Y. e-cigarette cessation: content analysis of a quit vaping community on Reddit. J Med Internet Res 2021 Oct 25;23(10):e28303 [FREE Full text] [doi: 10.2196/28303] [Medline: 34694229]
15. Hu M, Benson R, Chen AT, Zhu SH, Conway M. Determining the prevalence of cannabis, tobacco, and vaping device mentions in online communities using natural language processing. Drug Alcohol Depend 2021 Nov 01;228:109016 [FREE Full text] [doi: 10.1016/j.drugalcdep.2021.109016] [Medline: 34560332]
16. Benson R, Hu M, Chen AT, Zhu SH, Conway M. Examining cannabis, tobacco, and vaping discourse on Reddit: an exploratory approach using natural language processing. Front Public Health 2022 Jan 05;9:738513 [FREE Full text] [doi: 10.3389/fpubh.2021.738513] [Medline: 35071153]
17. Sowles SJ, Krauss MJ, Gebremedhn L, Cavazos-Rehg PA. "I feel like I've hit the bottom and have no idea what to do": supportive social networking on reddit for individuals with a desire to quit cannabis use. Subst Abus 2017;38(4):477-482 [FREE Full text] [doi: 10.1080/08897077.2017.1354956] [Medline: 28704167]
18. Rhidenour KB, Blackburn K, Barrett AK, Taylor S. Mediating medical marijuana: exploring how veterans discuss their stigmatized substance use on Reddit. Health Commun 2022 Sep;37(10):1305-1315. [doi: 10.1080/10410236.2021.1886411] [Medline: 33602000]
19. Ricard BJ, Hassanpour S. Deep learning for identification of alcohol-related content on social media (Reddit and Twitter): exploratory analysis of alcohol-related outcomes. J Med Internet Res 2021 Sep 15;23(9):e27314 [FREE Full text] [doi: 10.2196/27314] [Medline: 34524095]

XSL•FO
RenderX

20. Gauthier RP, Costello MJ, Wallace JR. “I will not drink with you today”: a topic-guided thematic analysis of addiction recovery on Reddit. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. 2022 Presented at: CHI '22; April 29-May 5, 2022; New Orleans, LA. [doi: 10.1145/3491102.3502076]
21. Balsamo D, Bajardi P, De Francisci Morales G, Monti C, Schifanella R. The pursuit of peer support for opioid use recovery on Reddit. Proc Int AAAI Conf Web Soc Media 2023 Jun 02;17(1):12-23 [FREE Full text] [doi: 10.1609/icwsm.v17i1.22122]
22. Krawczyk N, Bunting AM, Frank D, Arshonsky J, Gu Y, Friedman SR, et al. "How will I get my next week's script?" Reactions of Reddit opioid forum users to changes in treatment access in the early months of the coronavirus pandemic. Int J Drug Policy 2021 Jun;92:103140 [FREE Full text] [doi: 10.1016/j.drugpo.2021.103140] [Medline: 33558165]
23. Cunningham JA, Sobell LC, Chow VM. What's in a label? The effects of substance types and labels on treatment considerations and stigma. J Stud Alcohol 1993 Nov;54(6):693-699. [doi: 10.15288/jsa.1993.54.693] [Medline: 8271805]
24. Janulis P, Ferrari JR, Fowler P. Understanding public stigma toward substance dependence. J Appl Soc Psychol 2013 Apr 18;43(5):1065-1072. [doi: 10.1111/jasp.12070]
25. Key substance use and mental health indicators in the United States: results from the 2022 National Survey on Drug Use and Health. Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration. 2023. URL: https://www.samhsa.gov/data/sites/default/files/reports/rpt42731/2022-nsduh-nnr.pdf [accessed 2025-03-07]
26. Skolnick P. The opioid epidemic: crisis and solutions. Annu Rev Pharmacol Toxicol 2018 Jan 06;58:143-159. [doi: 10.1146/annurev-pharmtox-010617-052534] [Medline: 28968188]
27. Saloner B, McGinty EE, Beletsky L, Bluthenthal R, Beyrer C, Botticelli M, et al. A public health strategy for the opioid crisis. Public Health Rep 2018;133(1_suppl):24S-34S [FREE Full text] [doi: 10.1177/0033354918793627] [Medline: 30426871]
28. Chen AT, Johnny S, Conway M. Examining stigma relating to substance use and contextual factors in social media discussions. Drug Alcohol Depend Rep 2022 May 05;3:100061 [FREE Full text] [doi: 10.1016/j.dadr.2022.100061] [Medline: 36845987]
29. Baumgartner J, Zannettou S, Keegan B, Squire M, Blackburn J. The Pushshift Reddit dataset. Proc Int AAAI Conf Web Soc Media 2020 May 26;14(1):830-839. [doi: 10.1609/icwsm.v14i1.7347]
30. MacLean D, Gupta S, Lembke A, Manning C, Heer J. Forum77: an analysis of an online health forum dedicated to addiction recovery. In: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing. 2015 Presented at: CSCW '15; March 14-18, 2015; Vancouver, BC. [doi: 10.1145/2675133.2675146]
31. Earnshaw VA, Smith LR, Chaudoir SR, Amico KR, Copenhaver MM. HIV stigma mechanisms and well-being among PLWH: a test of the HIV stigma framework. AIDS Behav 2013 Jun;17(5):1785-1795 [FREE Full text] [doi: 10.1007/s10461-013-0437-9] [Medline: 23456594]
32. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: a robustly optimized BERT pretraining approach. arXiv. Preprint posted online on July 26, 2019 2025 [FREE Full text] [doi: 10.48550/arXiv.1907.11692]
33. Roesler D, Johnny S, Conway M, Chen AT. A theory-informed deep learning approach to extracting and characterizing substance use-related stigma in social media. BMC Digit Health 2024 Aug 16;2:60. [doi: 10.1186/s44247-024-00065-0]
34. Proudfoot K. Inductive/deductive hybrid thematic analysis in mixed methods research. J Mix Methods Res 2022 Sep 20;17(3):308-326. [doi: 10.1177/15586898221126816]
35. Livingston JD, Milne T, Fang ML, Amari E. The effectiveness of interventions for reducing stigma related to substance use disorders: a systematic review. Addiction 2012 Jan;107(1):39-50 [FREE Full text] [doi: 10.1111/j.1360-0443.2011.03601.x] [Medline: 21815959]
36. Chen AT. What's in a virtual hug? A transdisciplinary review of methods in online health discussion forum research. Libr Inf Sci Res 2014 Apr;36(2):120-130. [doi: 10.1016/j.lisr.2013.10.007]
37. Wang YC, Kraut R, Levine JM. To stay or leave?: the relationship of emotional and informational support to commitment in online health support groups. In: Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work. 2012 Presented at: CSCW '12; February 11-15, 2012; Washington, DC. [doi: 10.1145/2145204.2145329]
38. Smith LR, Earnshaw VA, Copenhaver MM, Cunningham CO. Substance use stigma: reliability and validity of a theory-based scale for substance-using populations. Drug Alcohol Depend 2016 May 01;162:34-43 [FREE Full text] [doi: 10.1016/j.drugalcdep.2016.02.019] [Medline: 26972790]
39. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960 Apr;20(1):37-46. [doi: 10.1177/001316446002000104]
40. Salmona M, Lieber E, Kaczynski D. Qualitative and Mixed Methods Data Analysis Using Dedoose: A Practical Approach for Research Across the Social Sciences. Thousand Oaks, CA: SAGE Publications; 2019.
41. Bond CS, Ahmed OH, Hind M, Thomas B, Hewitt-Taylor J. The conceptual and practical ethical dilemmas of using health discussion board posts as research data. J Med Internet Res 2013 Jun 07;15(6):e112 [FREE Full text] [doi: 10.2196/jmir.2435] [Medline: 23748147]
42. Ford E, Shepherd S, Jones K, Hassan L. Toward an ethical framework for the text mining of social media for health research: a systematic review. Front Digit Health 2020;2:592237 [FREE Full text] [doi: 10.3389/fdgth.2020.592237] [Medline: 34713062]

XSL•FO
RenderX

43. Lingard L. Beyond the default colon: effective use of quotes in qualitative research. Perspect Med Educ 2019 Dec;8(6):360-364 [FREE Full text] [doi: 10.1007/s40037-019-00550-7] [Medline: 31758490]
44. Morse JM. Critical analysis of strategies for determining rigor in qualitative inquiry. Qual Health Res 2015 Sep;25(9):1212-1222. [doi: 10.1177/1049732315588501] [Medline: 26184336]
45. Birtel MD, Wood L, Kempa NJ. Stigma and social support in substance abuse: implications for mental health and well-being. Psychiatry Res 2017 Jun;252:1-8. [doi: 10.1016/j.psychres.2017.01.097] [Medline: 28237758]
46. Snoek A, Levy N, Kennett J. Strong-willed but not successful: the importance of strategies in recovery from addiction. Addict Behav Rep 2016 Dec;4:102-107 [FREE Full text] [doi: 10.1016/j.abrep.2016.09.002] [Medline: 27975079]
47. Stokes M, Schultz P, Alpaslan A. Narrating the journey of sustained recovery from substance use disorder. Subst Abuse Treat Prev Policy 2018 Sep 27;13(1):35 [FREE Full text] [doi: 10.1186/s13011-018-0167-0] [Medline: 30261888]
48. De Crescenzo F, Ciabattini M, D'Alò GL, De Giorgi R, Del Giovane C, Cassar C, et al. Comparative efficacy and acceptability of psychosocial interventions for individuals with cocaine and amphetamine addiction: a systematic review and network meta-analysis. PLoS Med 2018 Dec 26;15(12):e1002715 [FREE Full text] [doi: 10.1371/journal.pmed.1002715] [Medline: 30586362]
49. Howlett N, García-Iglesias J, Bontoft C, Breslin G, Bartington S, Freethy I, et al. A systematic review and behaviour change technique analysis of remotely delivered alcohol and/or substance misuse interventions for adults. Drug Alcohol Depend 2022 Oct 01;239:109597 [FREE Full text] [doi: 10.1016/j.drugalcdep.2022.109597] [Medline: 35963209]
50. Webb L, Clayson A, Duda-Mikulin E, Cox N. ' I' m getting the balls to say no': trajectories in long-term recovery from problem substance use. J Health Psychol 2022 Jan;27(1):69-80 [FREE Full text] [doi: 10.1177/1359105320941248] [Medline: 32693631]
51. Gunn AJ, Sacks TK, Jemal A. "That's not me anymore": resistance strategies for managing intersectional stigmas for women with substance use and incarceration histories. Qual Soc Work 2018 Jul;17(4):490-508 [FREE Full text] [doi: 10.1177/1473325016680282] [Medline: 30116159]
52. Luoma JB, Platt MG. Shame, self-criticism, self-stigma, and compassion in acceptance and commitment therapy. Curr Opin Psychol 2015 Apr;2:97-101 [FREE Full text] [doi: 10.1016/j.copsyc.2014.12.016]
53. Luoma JB, Kohlenberg BS, Hayes SC, Bunting K, Rye AK. Reducing self-stigma in substance abuse through acceptance and commitment therapy: model, manual development, and pilot outcomes. Addict Res Theory 2008;16(2):149-165 [FREE Full text] [doi: 10.1080/16066350701850295] [Medline: 27746709]
54. Luoma JB, Kohlenberg BS, Hayes SC, Fletcher L. Slow and steady wins the race: a randomized clinical trial of acceptance and commitment therapy targeting shame in substance use disorders. J Consult Clin Psychol 2012 Feb;80(1):43-53 [FREE Full text] [doi: 10.1037/a0026070] [Medline: 22040285]
55. Corrigan PW, Rao D. On the self-stigma of mental illness: stages, disclosure, and strategies for change. Can J Psychiatry 2012 Aug;57(8):464-469 [FREE Full text] [doi: 10.1177/070674371205700804] [Medline: 22854028]
56. Courtney JB, Russell MA, Conroy DE. Acceptability and validity of using the BACtrack skyn wrist-worn transdermal alcohol concentration sensor to capture alcohol use across 28 days under naturalistic conditions - a pilot study. Alcohol 2023 May;108:30-43 [FREE Full text] [doi: 10.1016/j.alcohol.2022.11.004] [Medline: 36473634]
57. Swendeman D, Sumstine S, Aguilar E, Gorbach PM, Comulada WS, Gelberg L. Feasibility and acceptability of mobile phone self-monitoring and automated feedback to enhance telephone coaching for people with risky substance use: the QUIT-mobile pilot study. J Addict Med 2021 Apr 01;15(2):120-129 [FREE Full text] [doi: 10.1097/ADM.0000000000000707] [Medline: 32732684]
58. Park A, Conway M, Chen AT. Examining thematic similarity, difference, and membership in three online mental health communities from Reddit: a text mining and visualization approach. Comput Human Behav 2018 Jan;78:98-112 [FREE Full text] [doi: 10.1016/j.chb.2017.09.001] [Medline: 29456286]
59. Chen AT. Exploring online support spaces: using cluster analysis to examine breast cancer, diabetes and fibromyalgia support groups. Patient Educ Couns 2012 May;87(2):250-257. [doi: 10.1016/j.pec.2011.08.017] [Medline: 21930359]
60. Cesare N, Grant C, Nsoesie EO. Understanding demographic bias and representation in social media health data. In: Companion Publication of the 10th ACM Conference on Web Science. 2019 Presented at: WebSci '19; June 30-July 3, 2019; Boston, MA. [doi: 10.1145/3328413.3328415]

Abbreviations

BCT: behavior change technique
ML: machine learning
RQ: research question
SUD: substance use disorder
TTM: transtheoretical model

XSL•FO
RenderX

Edited by A Mavragani; submitted 21.02.24; peer-reviewed by A Davies, D Podder; comments to author 13.08.24; revised version received 22.09.24; accepted 18.12.24; published 26.03.25.

Please cite as:
Chen AT, Wang LC, Johnny S, Wong SH, Chaliparambil RK, Conway M, Glass JE
Stigma and Behavior Change Techniques in Substance Use Recovery: Qualitative Study of Social Media Narratives
JMIR Form Res 2025;9:e57468
URL: https://formative.jmir.org/2025/1/e57468
doi:10.2196/57468
PMID:

©Annie T Chen, Lexie C Wang, Shana Johnny, Sharon H Wong, Rahul K Chaliparambil, Mike Conway, Joseph E Glass. Originally published in JMIR Formative Research (https://formative.jmir.org), 26.03.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

User Experience With a Personalized mHealth Service for Physical Activity Promotion in University Students: Mixed Methods Study

Silke Wittmar[1], MSc; Tom Frankenstein[1], MSc; Vincent Timm[2], MA; Peter Frei[3], PhD; Nicolas Kurpiers[3], PhD; Stefan Wölwer[2], MA; Axel Georg Meender Schäfer[1], PhD

[1]Faculty of Social Work and Health, HAWK University of Applied Sciences and Arts Hildesheim/Holzminden/Göttingen, Hildesheim, Germany
[2]Faculty of Design, HAWK University of Applied Sciences and Arts Hildesheim/Holzminden/Göttingen, Hildesheim, Germany
[3]Institute for Sport Science, University of Hildesheim, Hildesheim, Germany

Corresponding Author:
Silke Wittmar, MSc
Faculty of Social Work and Health
HAWK University of Applied Sciences and Arts Hildesheim/Holzminden/Göttingen
Goschentor 1
Hildesheim, 31134
Germany
Phone: 49 5121881369
Email: silke.wittmar@hawk.de

Abstract

Background: Regular physical activity (PA) is known to offer substantial health benefits, including improved physical fitness, reduced risk of disease, enhanced psychological well-being, and better cognitive performance. Despite these benefits, many university students fail to meet recommended PA levels, risking long-term health consequences.

Objective: This study evaluated the user experience (UX) of futur.move, a digital intervention aimed at promoting PA among university students. The service delivers personalized, evidence-based content to foster sustained engagement in PA.

Methods: A mixed methods approach was used to evaluate the prototype of futur.move. UX assessments included on-site and online user tests, standardized questionnaires, and online focus groups. A total of 142 university students participated, with 23 joining additional focus groups. Each participant tested the service for 30 minutes. Quantitative data were collected using the User Experience Questionnaire and analyzed descriptively, followed by correlation analysis with variables such as PA level, age, gender, and experience with PA apps. Qualitative insights were gathered from transcribed focus group discussions and analyzed using content-structuring, qualitative content analysis. Quantitative findings were cross-validated with qualitative data.

Results: The UX received positive ratings across 4 User Experience Questionnaire scales (range –3 to +3; higher numbers indicate positive UX): attractiveness (median 1.67, IQR 1.04-2.17), perspicuity (median 1.5, IQR 0.5-2), stimulation (median 1.5, IQR 1-2), and novelty (median 1.25, IQR 0.5-2). Weak correlations were found between adherence to World Health Organization guidelines for PA and the perspicuity subscale (η=0.232, P=.04), and between age and the perspicuity (Kendall τb=0.132, P=.03) and stimulation subscales (Kendall τb=0.144, P=.02), and a moderate correlation was found between gender and the novelty subscale (η=0.363, P=.004). Critical feedback from focus group discussions highlighted issues with manual data entry. Qualitative findings aligned with the quantitative results, emphasizing students' appreciation for the personalized, diverse content and social networking features of futur.move.

Conclusions: futur.move demonstrates favorable UX and aligns with student needs, particularly through its personalized content and social features. Improvements should focus on reducing manual data entry and enhancing feature clarity, particularly for the features "your condition" and "goal setting." While correlations between UX ratings and demographic variables were weak to moderate, they warrant further investigation to better address the diverse target audience. The feedback from the students serves as a basis for further adapting the service to their needs and expectations. Future work will involve coding an advanced prototype and conducting a longitudinal study to assess its impact on PA behavior and sustained engagement.

(JMIR Form Res 2025;9:e64384) doi:10.2196/64384

XSL•FO
RenderX

KEYWORDS

usability testing; health promotion; exercise; smartphone app; mHealth; physical activity; user experience; user; university student; undergraduate; college; student; mixed methods; physical fitness; digital intervention; mobile health; promote; engagement; mobile phone

Introduction

The words commonly attributed to Hippocrates, "walking is humanity's best medicine," are more relevant than ever. In an era dominated by sedentary lifestyles, university students could greatly benefit from incorporating more physical activity (PA) into their daily routines.

Regular PA offers many important health benefits, including improved muscular and cardiorespiratory fitness, reduced risk of diseases, and enhanced psychological well-being [1-3]. PA also positively affects sleep and cognitive performance, including academic achievements and attention [1,3-5]. PA can also improve social skills and self-confidence [6].

The World Health Organization (WHO) recommends at least 150 minutes of moderate- to vigorous-intensity aerobic PA or at least 75 minutes of vigorous-intensity aerobic PA and muscle-strengthening activities of moderate to vigorous intensity for all major muscle groups on 2 or more days a week for adults [3]. Despite the benefits of regular PA, only a minority of young adults, including university students, meet the recommended PA levels, with 81% of adolescents [7] and 28% of adults worldwide and 42% in high-income Western countries [8] failing to adhere to these guidelines. Furthermore, studies show that students' PA levels drop significantly during examination periods and that they are more active during weekdays than on weekends [9-11]. As students undergo significant life transitions, their lifestyle habits, including PA patterns, can persist into adulthood if not addressed [12,13]. Therefore, universities face the challenge of supporting students in adopting more active lifestyles [14].

Promoting PA through mobile health (mHealth) services holds great potential, especially as students are frequent users of technology [15]. mHealth interventions are preferred by young people over traditional face-to-face approaches [16]. However, many current apps often fail to incorporate evidence-based recommendations for PA promotion, with developers prioritizing marketing strategies over addressing real user needs [17,18].

Positive user experience (UX) influences user satisfaction, promoting sustained product use and increasing recommendations [19]. "To create successful products or services it is necessary to ensure that the product has a sufficiently high user experience" [20]. UX is defined as "...a person's perceptions and responses that result from the use or anticipated use of a product, system or service. Users' perceptions and responses include the users' emotions, beliefs, preferences, perceptions, comfort, behaviors, and accomplishments that occur before, during and after use" [21]. Evaluating UX early in the prototyping phase allows developers to incorporate user feedback and refine features to meet diverse user needs [20,21]. In the context of PA services, successful interventions include behavior change techniques, such as self-monitoring, goal setting, behavioral feedback, social support, and social comparison, to effectively increase PA [22-26]. The challenge remains to maintain long-term user engagement with mHealth interventions [16]. Numerous studies have identified poor functionality and technical issues as barriers to engagement with mobile interventions, while design features such as personalization, social interaction, and customized feedback enhance user engagement [16]. Personalization is crucial, as students' PA needs vary widely [6], and one-size-fits-all approaches are inadequate [27-29].

This study aims to comprehensively evaluate the UX of the futur.move prototype, an mHealth service designed to enhance PA among university students. Specifically, this study will focus on the following UX aspects: (1) attractiveness—whether users like the service, (2) perspicuity—whether the service is easy to learn, (3) stimulation—whether using the service is exciting and motivating, and (4) novelty—whether the service is innovative and creative. The central research question is: "How do students evaluate the UX of the futur.move prototype?" Of particular interest are the specific improvements students suggest to optimize implementation.

Methods

Study Design

This study design follows a sequential approach with a series of cross-sectional surveys using structured online questionnaires with qualitative focus group discussions to comprehensively evaluate the futur.move prototype. On-site tests were conducted with a large participant group at the university, while smaller online sessions were held via Zoom (Zoom Video Communications Inc).

This study report follows the CHERRIES (Checklist for Reporting Results of Internet e-Surveys) reporting guideline for reporting web-based surveys [30] (Multimedia Appendix 1).

futur.move

futur.move is a digital service that aims to improve students' PA through a smartphone app and website in the form of a progressive web application. According to a definition of service design, a service should be useful, usable, and desirable for the customer, while being effective, efficient, and unique for the provider. Services include interactions and interfaces between the customer and the provider [31].

The development of futur.move follows the holistic framework for the development of eHealth technologies [32] and includes 5 phases: context analysis, design, UX evaluation, programming, and long-term implementation. Currently, the prototype is in the UX evaluation phase (phase 3). The design is informed by a literature review, 2 student creative workshops (n=13 participants), and social cognitive theory [33]. Studies report that PA interventions are more successful when they are theory

XSL•FO
RenderX

based [34,35]. Components of social cognitive theory such as health-related knowledge, individual goal setting, and social support are transferred into the functions of the app to support users in successfully and sustainably increasing their PA. An interdisciplinary team of UX designers and health scientists developed the app iteratively in the collaborative design tool Figma (Figma Inc, 2022) with active involvement from the target audience. Figure 1 shows the content structure of futur.move features.

Figure 1. Structural overview of the features and content offered by the futur.move mHealth service. PA: physical activity. mHealth: mobile health.

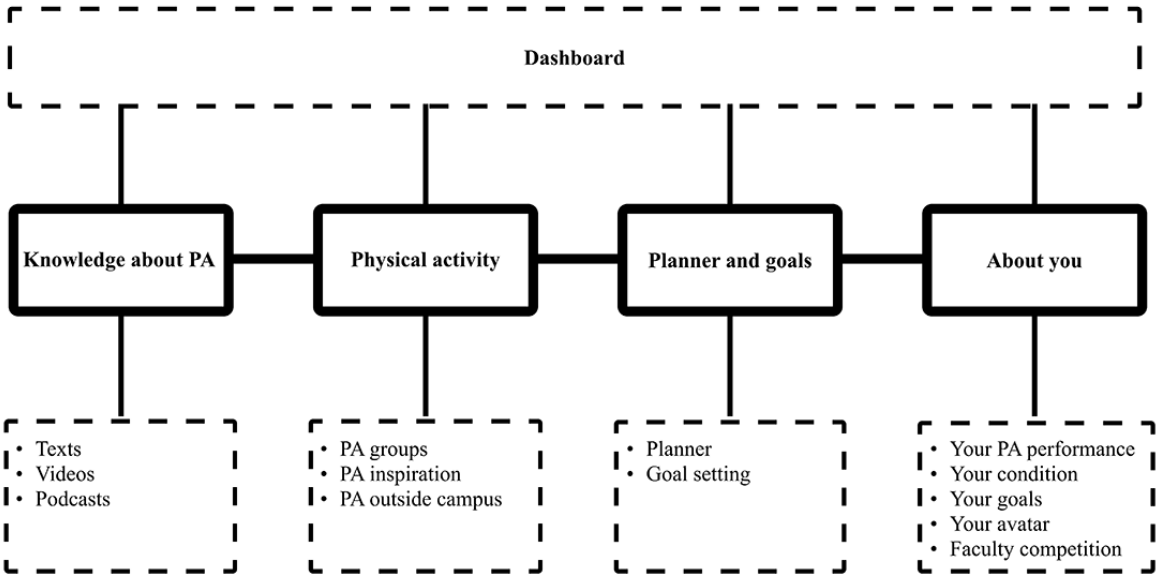

After registration and reporting PA habits via the European Health Interview Survey–Physical Activity Questionnaire (EHIS-PAQ) [36], users access personalized content across five areas: (1) “dashboard,” (2) “knowledge,” (3) “physical activity,” (4) “planner and goals,” and (5) “about you” (Figure 1). A messaging function is also available. In the “knowledge” section, users access PA-related information through various formats such as text, videos, and podcasts. The “physical activity” section enables users to find inspiration for PA, join or create PA groups, and use links to find PAs outside the campus. “Planner and goals” offers a calendar view and goal-setting features, while “about you” manually tracks PA habits and, goals, and provides a faculty competition. By participating in PA, users can earn points for their faculties and win rewards such as a faculty party. Figure 2 shows the personalizable dashboard of futur.move.

XSL•FO
RenderX

Figure 2. Screenshot of the customizable dashboard of the futur.move mHealth service as viewed on a smartphone. mHealth: mobile health. An English version of the dashboard is available in Multimedia Appendix 2.

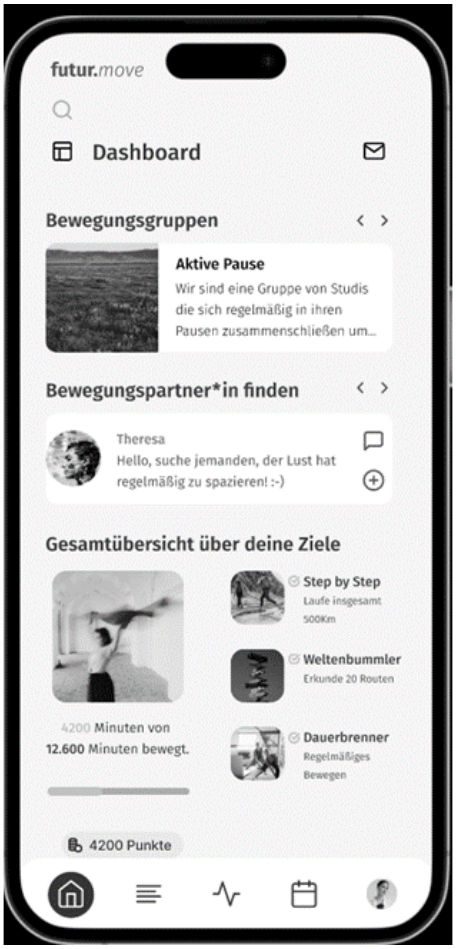

Participants and Recruitment

User testing was conducted from October 27 to December 21, 2023, at University of Applied Sciences and Arts Hildesheim/Holzminden/Göttingen (HAWK; N=6300 students in winter 2022/2023).

The convenience sample for this study was recruited from the entire student body of HAWK. For the on-site sessions, a high number of participants was targeted to enable valid statistical analyses. Therefore, researchers reached out directly to professors from all 3 university campuses and various faculties, requesting their support in implementing this study within their seminars. On-site sessions were finally held in 4 seminars across different faculties and locations. Participation in this study was integrated into the respective seminars; however, it remained entirely voluntary for all participants, ensuring that their consent was informed and given freely without any obligation.

For the online sessions with focus group discussions, all students at HAWK were invited via email, flyers, and the university's social media channels. Flyers were posted in high-traffic areas on the university campuses, such as the cafeteria and the library. Additionally, first-year students from the Faculty of Social Work and Health were personally approached at a welcome event for new students. Due to low participation, students were also approached at a campus café. The latter method of reaching out to students proved to be the most effective. A link distributed through the mentioned media allowed registration via HAWK's online learning tool, followed by contact with the project team, provision of information, and signing of the consent form. Only those who returned the signed form participated. To participate in both the on-site and online data collection, inclusion criteria were enrollment and no prior involvement in futur.move.

Recommended sample sizes for quantitative UX testing to estimate key performance parameters range from 10 to 1689 participants, depending on the margin of error. "A sample size of 115 at 90% confidence can detect a 10% significant difference" [37]. Therefore, we aimed to recruit a minimum of 65 participants for the survey to achieve 90% confidence. For qualitative UX testing, it is recommended to conduct focus groups with 4 to 37 participants to identify usability issues. A sample size of 18 is needed to detect a 10% insight occurrence [37]. Consequently, we aimed to recruit at least 18 participants for the focus group discussions. Given the maximum group size of 6 to 8 participants per focus group [38], we planned for a minimum of 3 groups.

Ethical Considerations

Ethical approval was obtained from the HAWK ethics committee on September 29, 2023. All participants gave written informed consent before participation, agreeing to the use of deidentified data for secondary analyses and the publication of results. Data were collected anonymously, and no personally identifiable information was included in the final dataset.

XSL•FO
RenderX

Participation was voluntary. Participants could withdraw at any time without justification. Data were stored on the university's server for secure keeping until January 10, 2033, and will then be deleted by the researchers. Throughout this study, only deidentified data were used, and access was restricted solely to the research team. Participants in the online sessions received a €10 (US $10.85) voucher for an online retailer.

Procedures

On-Site Data Collection

Quantitative data were collected on-site during the seminars with a project team member present. Participants assumed the role of "Kim," a fictitious user of futur.move, to understand prestored data and experience all app features. After an introduction and onboarding, they received a task sheet with 6 tasks to independently test the service for approximately 30 minutes (eg, finding exercise information, Multimedia Appendix 3). They accessed the prototype via a device (eg, laptop or tablet).

After completing each task, participants provided initial feedback on the task sheet and then completed the online questionnaire. This included sociodemographic data, PA behavior (EHIS-PAQ) [36], and usability ratings using the User Experience Questionnaire (UEQ) [39]. The survey, hosted via SoSci Survey (version 3.5.02), presented the items to each participant in the same order, allowed participants to change their answers using a back button, and omitted irrelevant questions. There was no check for completeness before submission. No cookies were used. The online survey consisted of 22 pages, including the introduction and conclusion. The number of items on each page varied. Each participant used a unique ID to merge task sheet responses with the online survey for analysis. The ID was randomly assigned on each task sheet. Each code could only be used once. On-site sessions lasted approximately 60 minutes per session.

A pretest was conducted to ensure clarity and feasibility. Four students tested the task sheets, while 2 project members and 2 individuals from the target group tested the online survey. Feedback led to revisions of both instruments.

Online Data Collection

Online data collection was conducted via the Zoom videoconferencing tool with 2 project team moderators. Participants registered for user tests, and sessions were scheduled based on availability. Participation was voluntary, and participants could withdraw at any time.

The first part of the online session mirrored the on-site data collection. The moderators reviewed task sheets and survey feedback to refine focus group discussions, gathering targeted qualitative insights beyond the prepared questions.

Measures

Demographics

As part of the online survey, participants provided their age, gender, height, weight, faculty affiliation, and whether they had used apps to improve PA before. Additionally, the EHIS-PAQ [36] was used to collect standardized data on students' PA. The EHIS-PAQ is a domain-specific instrument for measuring (1) work-related PA, (2) transport-related PA, and (3) leisure-time PA. It distinguishes between "aerobic" and "muscle-strengthening" PA and allows the assessment of compliance with health-promoting PA recommendations [36].

About UX

UX was assessed using the UEQ [39]. The UEQ is a quantitative instrument for assessing the UX of interactive products. It originally consisted of 6 subscales with a total of 26 items with classic usability aspects (ie, efficiency, perspicuity, or dependability) and UX aspects (ie, originality or stimulation). The authors state that entire subscales can be omitted if they are irrelevant to the investigation [40]. Consequently, the "efficiency" and "dependability" scales from the original questionnaire were not included in the analysis, as no valid results could be expected for these scales at the stage of development of the service. In this study the items of the four subscales (1) attractiveness (do the users like the service or not?), (2) perspicuity (is it easy to learn how to use the service?), (3) stimulation (is the use of the service exciting and motivating?), and (4) novelty (is the service innovative and creative?) were used. The attractiveness scale is represented by 6 items, while the other scales are represented by 4 items each. The items are semantic differentials, for example, attractive and unattractive. They can be rated on a 7-point scale from –3 (horribly bad) to +3 (extremely good). Scale scores are average item ratings across participants. Scores between –0.8 and 0.8 represent a neutral evaluation of the scale, scores>0.8 represent a positive evaluation, and scores<0.8 represent a negative evaluation [40].

Evaluation of Individual Features

On the task sheet for testing futur.move, students had the opportunity to indicate how much they liked each feature of futur.move. They could tick a value from very good (1) to poor (5) on a 5-point Likert scale. Within this framework, the students rated the five features: (1) faculty competition, (2) your condition, (3) knowledge, (4) PA group (hiking), (5) PA inspirations, and (6) goal setting. Additionally, the online questionnaire asked students how they would rate futur.move overall on a 5-point Likert scale from 1 (very good) to 5 (poor).

Focus Groups

The focus groups were conducted to qualitatively summarize the quantitative feedback from the students. The questions in the focus groups were, therefore, related to the quantitative part of the respective user tests. Follow-up questions were individually prepared by the moderators by reviewing the anonymized data from the user tests. The focus group structure followed the recommended procedure in empirical social science [41]. The duration of the focus groups was approximately 120 minutes, and they were conducted by VT and SW, who moderated the discussion. The discussion was recorded using the recording feature of Zoom.

The first question in the focus groups was what participants remembered most about futur.move. Following this, an overview of the UEQ scales was briefly presented. Based on the UEQ scores, questions about divergences in the scoring of individual

XSL•FO
RenderX

scales were discussed. Students were asked to describe their quantitative assessments in more detail, and discrepancies in the service ratings from the task sheet were examined. Finally, there was an opportunity for participants to report any further requirements. Discussions were recorded with participants' consent for later transcription and stored on the university's server.

Data Analyses

Quantitative Data Analyses

The quantitative data from the questionnaires were analyzed using IBM SPSS (version 27; IBM Corp).

The UEQ Data Analysis Tool (version 12) transformed the UEQ data, which were then merged in SPSS using participant IDs. Meta-information was added, and variable labeling was adjusted, including scale levels [42]. Data cleansing involved checking for incorrect or implausible information and marking missing values with the code 999. Questionnaires with missing values were included in the analysis.

For the UEQ data, items on the subscales were analyzed for divergent ratings. Cases were recommended for exclusion if at least 3 subscales were conspicuous or if more than 15 of 26 items had identical ratings. No records were excluded based on plausibility, but 8 records from the task sheets were excluded due to missing online survey data.

As the UEQ data were not normally distributed, the median and IQR were calculated. Nonetheless, scale means, variances, and Cronbach α were calculated using the UEQ Data Analysis Tool, to use the provided benchmark for categorizing scale scores against data from 468 studies on various products, rating them from "poor" to "excellent" [40].

In addition to the descriptive data analysis, a correlation analysis was conducted to explore the relationships between the UX evaluation of the futur.move, as measured by the UEQ, and various demographic and usage-related factors. These factors included compliance with WHO guidelines for PA, age, gender, and prior experience with apps aimed at promoting PA.

Associations were analyzed using the η and Pearson chi-square tests for metric and nominal-scaled data. For associations between 2 metric variables, Kendall τb was used. The significance level for all tests was set at $\alpha \leq .05$.

Qualitative Data Analyses

The focus group recordings were automatically transcribed verbatim using the tool noScribe (version 0.4.1). The transcripts were then manually corrected and anonymized, with participant names replaced by the IDs matching those from the quantitative survey, allowing data linkage.

Two project members analyzed the transcripts using MAXQDA 2022 (VERBI-Software. Consult. Sozialforschung GmbH 2021) following content-structuring, qualitative content analysis [43]. The data were analyzed based on the question: "How do students evaluate the futur.move prototype?" with the 4 UEQ scales (attractiveness, perspicuity, stimulation, and novelty) used as deductive categories. Further inductive categories could be formed based on the material, but no additional categories emerged. Subcategories of the main categories were formed inductively during the analysis.

Results

Quantitative Findings

Demographics

A total of 142 people (mean age 24, SD 4 years; female: n=105, 74%; male: n=33, 23%) participated in the user tests. Change to: A total of 142 people (mean age 24, SD 4 years; female: n=105, 75%; male: n=33, 23%) participated in the user tests. Of these, 23 also took part in the focus groups. Most participants were students from the Faculty of Social Work and Health (n=60, 42%). Approximately a third of the students had prior experience using an app to promote PA (n=49, 35%). The majority of students (n=87, 61%) met the WHO guideline of practicing more than 150 minutes of moderate aerobic activity per week, while 49% (n=69) met the guideline of doing muscle-strengthening exercises at least 2 days a week. Overall, 40% (n=57) of participants met the WHO guidelines for PA, fulfilling both criteria. The mean BMI of the students was 23 (SD 4), which falls at the upper end of the normal weight range. Table 1 summarizes the information provided by the students.

XSL•FO
RenderX

Table 1. Demographics of university students participating in user tests conducted from October 27 to December 21, 2024.

Characteristics	Participants
Age (years), mean (SD)	24 (4)
Gender, n (%)	
Female	105 (74)
Male	33 (23)
Diverse	1 (1)
Not specified	3 (2)
Faculty, n (%)	
Social work and health	60 (42)
Design	26 (18)
Engineering and health	22 (16)
Architecture, engineering, and conservation	20 (14)
Management, social work, and construction	13 (9)
Not specified	1 (1)
"Do you already use digital services to promote physical activity?" n (%)	
Yes	49 (35)
No	74 (52)
Not specified	19 (13)
BMI (kg/m^2), mean (SD)	23 (4)
Fulfillment WHO[a] guideline >150 minutes of moderate aerobic activity, n (%)	
Yes	87 (61)
No	55 (39)
Fulfillment WHO guideline ≥2 days/week muscle strengthening, n (%)	
Yes	69 (49)
No	73 (51)
Fulfillment WHO guideline for physical activity overall, n (%)	
Yes	57 (40)
No	85 (60)

[a]WHO: World Health Organization.

futur.move UX

The UX of futur.move is indicated by the average score on each respective scale. As the data are not normally distributed, the median, IQR, and the minimum and maximum range are also reported. Attractiveness received a median score of 1.67 (IQR 1.04-2.17) and a mean score of 1.55 (SD 0.91) on a scale from –3 to +3. Perspicuity was rated with a median of 1.5 (IQR 0.5-2) and a mean of 1.28 (SD 0.97), stimulation with a median of 1.5 (IQR 1-2) and a mean of 1.37 (SD 1.03), and novelty with a median of 1.25 (IQR 0.5-2) and a mean of 1.27 (SD 0.98). Positive ratings are indicated by scores greater than 0.8 (UEQ manual).

The UEQ benchmark refers to the mean scores and SD. According to the UEQ benchmark results shown in Figure 3, futur.move falls into the "good" category for hedonic quality (stimulation and novelty scales). The attractiveness and perspicuity scales are in the "above average" category. A score is considered "good" if it is among the top 25% in the benchmark dataset and "above average" if 25% of other products score better. Therefore, the overall quality of futur.move can be classified as good according to the UEQ. However, none of the scale scores fall into the "excellent" category, which includes the top 10% of results, indicating room for improvement.

XSL•FO
RenderX

Figure 3. UX evaluation results of the futur.move mHealth service by study participants using the UEQ, compared to a dataset of 468 product reviews, analyzed with the UEQ Data Analysis Tool Version 12. Error bars represent the 95% CI. mHealth: mobile health; UEQ: User Experience Questionnaire; UX: user experience.

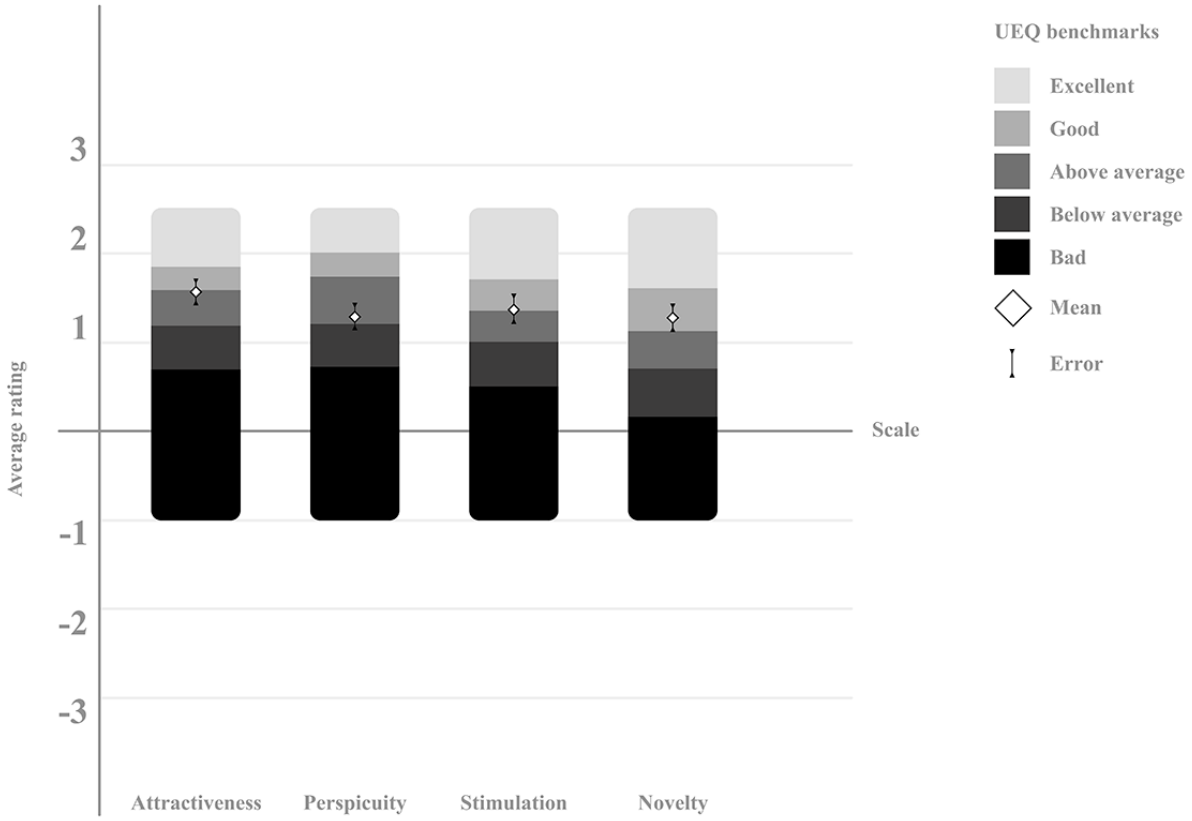

Evaluation of Individual Features

When rating the individual features of futur.move, at least 68% (95/142) of participants rated all features with a score of 2 or better (Figure 4). The PA features (PA groups and PA inspirations) received the highest scores, with 90% (126/139) and 86% (116/135) of participants rating them with a score of 2 or better, respectively. Only a small number of respondents rated 2 features with a score of 5 (very poor): “your condition” (1/142, 0.7%) and “goal setting” (3/142, 2%).

XSL•FO
RenderX

Figure 4. Average ratings of individual features of the futur.move mHealth service by study participants. Each feature was evaluated on a scale from 1 (very good) to 5 (poor). mHealth: mobile health; PA: physical activity.

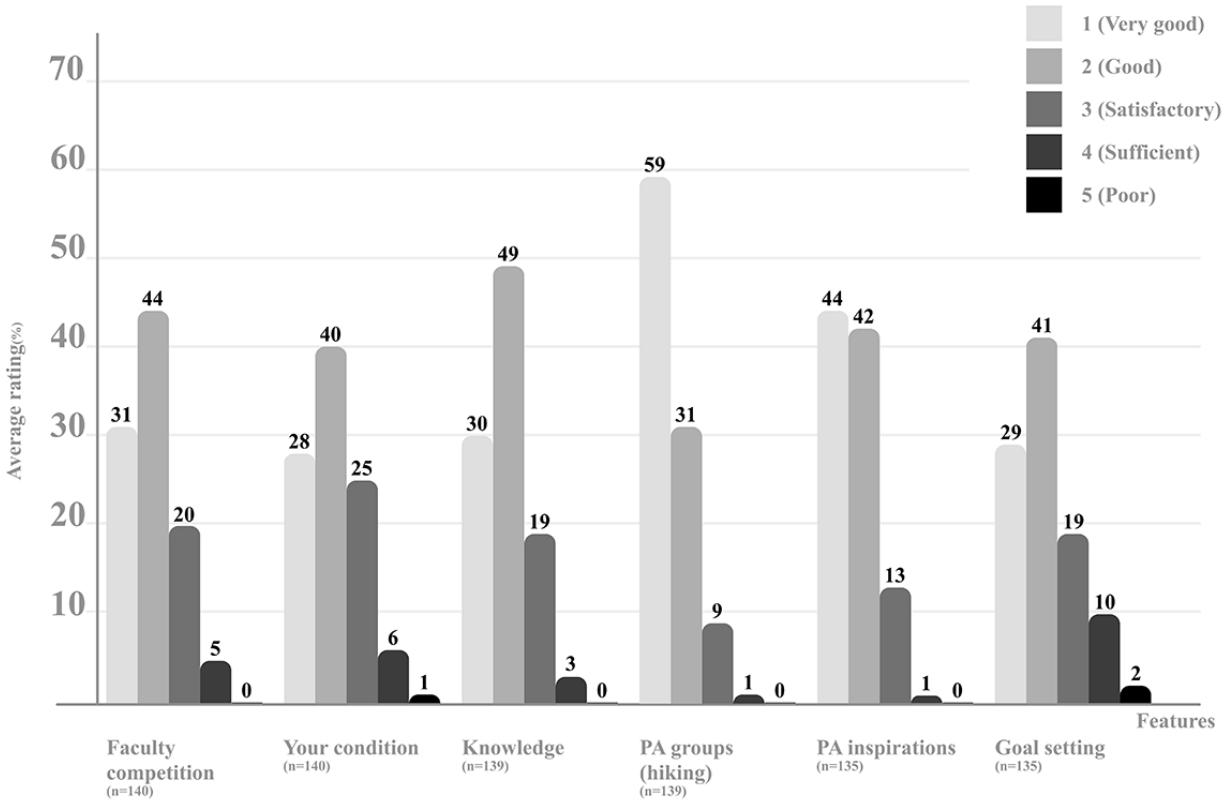

Overall Grade

Overall, 105 (74%) of 142 participants gave futur.move an overall grade of 2 (n=97, 68%) or better (n=8, 6%). No grade lower than 4 was awarded (grade of 3: n=29, 20%; grade of 4: n=8, 6%). The average grade for the overall evaluation of futur.move was 2.3 (German grading scale, where 1 is the best grade and 6 the worst); a grade of 2 corresponds to a "good" rating.

Correlation Analysis

A summary table of the correlation analysis results can be found in Multimedia Appendix 4. To analyze the direction of correlations calculated with η, corresponding bar charts were used. These are also included in Multimedia Appendix 4.

The correlation analysis reveals a weak to moderate relationship between adherence to WHO guidelines for PA and the UX evaluation using the UEQ subscale perspicuity (η=0.232, *P*=.04). Adherence to WHO guidelines is associated with less favorable UX ratings.

Additionally, 2 weak but significant relationships were identified between age and the UEQ subscales perspicuity (Kendall τb=0.132, *P*=.03) and stimulation (Kendall τb=0.144, *P*=.02). Higher age was associated with more favorable evaluations of futur.move on both subscales.

A significant moderate relationship was also found between gender and the UEQ subscale novelty (η=0.363, *P*=.004), with women providing higher ratings for novelty compared to men.

No significant influence was observed for the variable prior experience with apps for PA on the UX evaluation using the UEQ.

Qualitative Findings

Overview

Four focus group discussions were conducted with a total of 23 participants. The main categories corresponded to the four scales of the UEQ: (1) attractiveness, (2) perspicuity, (3) stimulation, and (4) novelty. Figure 5 shows an overview of the UX dimensions with an exemplary anchor quote representing each category. The quotes mentioned in the text can be found in Multimedia Appendix 5.

XSL•FO
RenderX

Figure 5. Overview of user experience categories for the futur.move mHealth service from the qualitative content analysis of 4 focus group discussions, structured according to UEQ subscales with illustrative anchor quotations. mHealth: mobile health; P: participant; UEQ: User Experience Questionnaire.

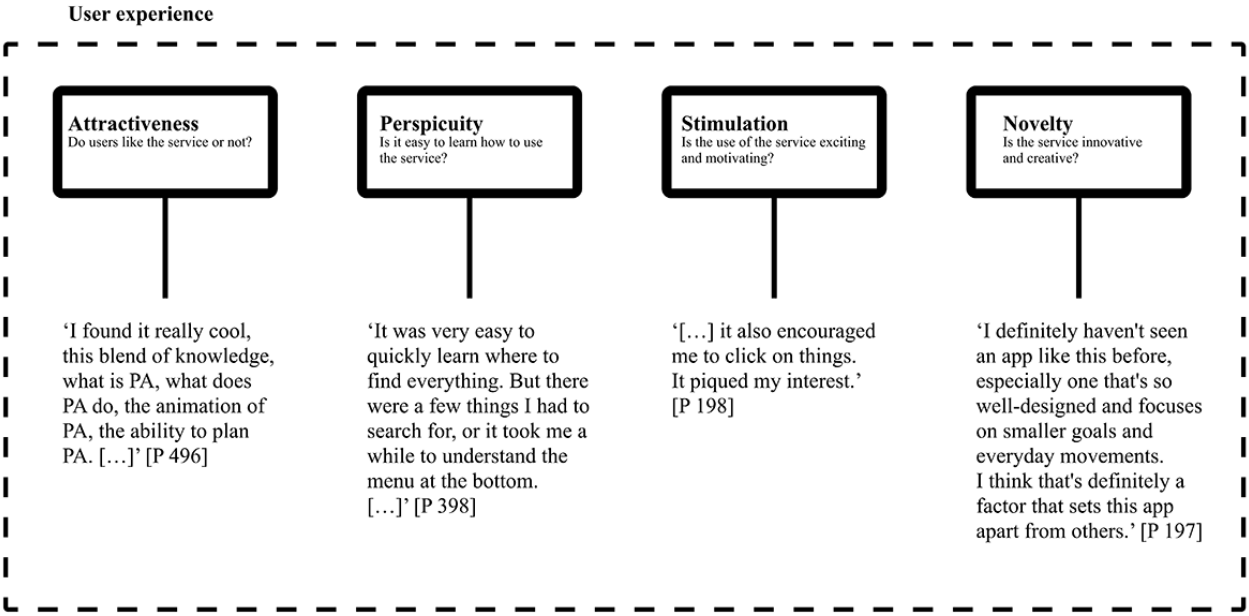

Attractiveness

Most participants liked futur.move overall and had fun when using it (quote 1 in Multimedia Appendix 5). Some participants were particularly attracted by the variety of features (quotes 2 and 3 in Multimedia Appendix 5). However, few individuals found the platform to be less attractive. According to the participants, this was due to enduring a recent injury, living away from the university (quote 4 in Multimedia Appendix 5), or undertaking part-time studies (quote 5 in Multimedia Appendix 5). Many participants highlighted the networking opportunities with other students and the links to PA opportunities in the region (quotes 3 and 6 in Multimedia Appendix 5).

Most participants thought the "physical activity inspirations" were particularly attractive because of their diversity (quote 7 in Multimedia Appendix 5), perceived as a variety of offers for different needs. One participant thought it was particularly positive that the PA options were low-threshold; that is, they did not always entail sport but general PA such as a walk around the city (quote 8 in Multimedia Appendix 5). Most participants found the group activities attractive, even though other perspectives were also considered (quotes 9 and 10 in Multimedia Appendix 5). One person emphasized that the group recommendations were also a good way for first-year students to make contacts upon arrival in a new city (quote 10 in Multimedia Appendix 5). When using a digital service, the physical presence was also discussed. One participant particularly liked that futur.move offers the opportunity to meet in person (quote 11 in Multimedia Appendix 5). Most participants thought the "knowledge" feature was attractive. In particular, the variety of media was mentioned as appealing (quote 12 in Multimedia Appendix 5). However, details such as the structure of the texts also contributed to the perceived attractiveness (quote 13 in Multimedia Appendix 5). Many of the participants liked the "goal setting" feature in general. One of the reasons given was the possibility to set small, realistic goals (quote 14 in Multimedia Appendix 5). Several focus groups discussed appropriate target periods (quote 15 in Multimedia Appendix 5) and manual data entry (quote 16 in Multimedia Appendix 5).

Some of the participants thought it was attractive to compete with others in a faculty competition (quote 17 in Multimedia Appendix 5), while others would rather avoid such competition (quote 18 in Multimedia Appendix 5). One person cited a lack of connection to their own faculty as a reason (quote 19 in Multimedia Appendix 5).

In terms of the attractiveness of individual features, the option to design one's own avatar elicited mixed feedback from participants. While 1 person thought it was appealing (quote 20 in Multimedia Appendix 5), another did not find the choice appealing at all (quote 21 in Multimedia Appendix 5).

Most participants were very positive about the design of futur.move (quote 22 in Multimedia Appendix 5). One participant particularly emphasized the design of the "knowledge" feature, which was perceived as clean (quote 23 in Multimedia Appendix 5). One person had a good feeling about the overall service but perceived a design break in the "your condition" feature (quote 24 in Multimedia Appendix 5). The color scheme of the service was rarely discussed, and apart from 1 comment on the color scheme in the "your condition" feature, was perceived as positive (quote 25 in Multimedia Appendix 5). The choice of images was particularly appealing to some respondents (quote 26 in Multimedia Appendix 5).

Perspicuity

Most participants described the platform as clear, tidy, and not too packed. The font, large headings, and organized layout contributed to this (quote 27 in Multimedia Appendix 5). Most

XSL•FO
RenderX

features were also found to be easy to understand. Most participants reported that they found their way around quickly, although it took some time to get used to the system in the beginning (quote 28 in Multimedia Appendix 5). While certain aspects of the tasks required exploration, participants found the navigation easy to learn and were able to use it quickly (quote 29 in Multimedia Appendix 5).

In contrast, not all participants understood the diagrams in the "your condition" section or the points system for the "faculty competition" (quotes 30 and 31 in Multimedia Appendix 5). In 1 focus group, the comprehensibility of the navigation buttons in particular was discussed and perceived as somewhat difficult to comprehend (quote 32 in Multimedia Appendix 5).

Stimulation

Overall, the design and content of the service caught the interest of most participants and stimulated them to explore the content (quote 33 in Multimedia Appendix 5). In this context, individual feedback raised the question about the long-term motivation to use the service (quote 34 in Multimedia Appendix 5).

The "physical activity" feature received a lot of positive feedback, with the variety of activities seeming to be a key advantage. Not all activities appealed to everyone (quote 35 in Multimedia Appendix 5), but there was something for everyone (quote 7 in Multimedia Appendix 5). Most participants were particularly positive about the possibility of using futur.move to become physically active in the first place or even to initiate groups themselves (quote 36 in Multimedia Appendix 5). The "knowledge" feature also appealed to many participants, partly because of the variety of media and the design of the contributions (quote 37 in Multimedia Appendix 5). Many participants thought the "goal setting" feature was motivating, especially because small goals could be set and big goals could be broken down (quotes 38 and 39 in Multimedia Appendix 5). Opinions differed on the period length to achieve a target and whether the tool would be used during implementation (quotes 40 and 41 in Multimedia Appendix 5).

The "faculty competition" feature was discussed by participants. While some felt motivated by the competitive nature of the event (quote 42 in Multimedia Appendix 5), some participants did not find it appealing (quote 18 in Multimedia Appendix 5). Furthermore, the focus groups discussed the possibility of providing information about one's own mood or having PA data displayed in the "your condition" feature: while some found this feature helpful, others thought it was uninteresting or even problematic to confront them with their own performance (quotes 43, 44, and 45 in Multimedia Appendix 5). Designing an avatar was a less motivating feature of the service to use for some participants (quote 46 in Multimedia Appendix 5).

Novelty

Participants indicated that they were particularly attracted by the creative design of the service (quote 47 in Multimedia Appendix 5) and the "fresh" headings (quote 48 in Multimedia Appendix 5). Some participants thought the service was new (quote 49 in Multimedia Appendix 5). It was not the content itself that was described as new, but its combination as an infrastructure service (quote 50 in Multimedia Appendix 5). The fact that the content was already familiar to some participants from other contexts was also a reason for some to perceive the service as less new (quote 51 in Multimedia Appendix 5).

Discussion

Principal Findings

The evaluation of the UX of futur.move using the UEQ revealed generally positive results, with attractiveness receiving the highest ratings. Compared to the UEQ benchmark, all scores were at least "above average." However, all ratings still leave room for improvement. It has to be taken into account that despite rating options between –3 and +3, real-world applications typically produce lower values, with averages unlikely to exceed +2 or go below –2 [40]. "Your condition" and "goal setting" were less favorably rated, suggesting these areas are less engaging or useful for users. This feedback provides valuable insights for targeted enhancements to increase user satisfaction and engagement.

Qualitative results complemented the quantitative findings, reflecting the positive feedback on the service's content and design. Participants particularly valued the personalization features, as well as the social and the variety of PA options, which enhanced their overall enjoyment. However, critical feedback identified areas for improvement, notably the manual data entry process and difficulties in comparing progress with other users.

The correlation analysis provides additional insights into potential areas for improving UX, though the findings should be interpreted with caution. A weak to moderate correlation was found between adherence to WHO guidelines for PA and the UEQ subscale perspicuity, contradicting that individuals who meet WHO guidelines for PA provide higher UX ratings for futur.move using the UEQ. As none of the other subscales showed any associations with adherence to the WHO guideline, this is probably a chance finding, as the relationship cannot be explained conceptually.

Interestingly, older students rated the UX higher on perspicuity and stimulation compared to younger students. These findings point to a need for better understanding what motivates the core target group of students aged 20-30 years to engage with the service and how to make it more intuitive for them. Although age appears to have a weak influence overall, these insights can guide refinements to improve engagement with this age group.

The observation that women perceive futur.move as more novel than men suggests that women rate UX higher, specifically in the subscale novelty. This finding emphasizes the importance of tailoring innovative features to better engage men and other genders, ensuring that the service effectively appeals to the diverse target audience. Addressing these issues could significantly enhance usability and overall UX. The analysis of the qualitative data by PA, gender, and age showed no additional insights.

XSL•FO
RenderX

Comparison With Prior Work

Our study expands the existing literature by being to our knowledge the first to investigate the UX of an mHealth prototype in an early development phase using a combined qualitative and quantitative approach. It provides valuable insights into the strengths and weaknesses of the prototype, along with practical recommendations for further development.

Previous studies on UX in mHealth services aimed at promoting PA primarily focused on fully programmed apps in long-term studies, with a predominant emphasis on usability aspects [44-47]. For example, 1 study evaluated UX aspects using the UEQ without incorporating qualitative methods [48]. In contrast, our study integrates qualitative and quantitative UX methods in an early development phase to derive key insights into participatory development and its impact on UX.

The evaluation of **futur.move** revealed strengths in usability, personalization, and design, with the ease of use being particularly notable after a brief familiarization period. At the same time, weaknesses such as the need for manual data entry were identified, which could negatively affect long-term usage [16,29]. These findings align with studies emphasizing the importance of automated tracking features, intuitive navigation, and the critical role of personalization options [16,29,49].

The investigation of social comparison features revealed individual preferences. While some participants found the "faculty competition" motivating, others expressed a negative attitude toward competition, highlighting the diversity of user values. A systematic review recommends using competitive elements sparingly and in combination with other features [16]. The results of other studies suggest that intergroup competition approaches are beneficial [50,51].

Gender and age correlate with UX ratings. Women tended to evaluate the aspect of novelty more positively than men. One explanation could be that men are generally more frequent users of technology [52] and tend to score higher on measures of technology acceptance and competence [53,54]. This could explain why men perceive futur.move as less novel, as they may already be familiar with a wider range of similar offerings. Conversely, women often place greater emphasis on the design and user-friendliness of an application [55], which could positively influence their evaluation of its novelty. These findings highlight the importance of considering gender-specific preferences in UX design. For example, future development could focus on creating a balance between functionality and aesthetic appeal to meet the expectations of both genders. Additionally, efforts to better understand how men and women perceive and engage with novel features in mHealth apps could further optimize the UX of futur.move.

Research indicates that younger people are more likely to use smartphone apps for health-related purposes. For instance, as early as 2016, one-third of students surveyed in a study were already using mHealth apps [56]. Empirical evidence also shows that interest in technology tends to decrease with age among older adults [57]. Interestingly, in our study, older students rated the perspicuity and stimulation provided by futur.move more favorably. This could be attributed to differences in preferences, as older individuals may value ease of navigation and straightforward information over other features. While younger users generally engage more frequently with health apps, older adults may respond more positively to specific UX elements when tailored to their needs. Identifying these aspects should be a focus in future development.

Limitations

In this study, a convenience sample of students was used. Participants were not selected based on their PA habits or need for PA in the university context. Due to initial reluctance from students to participate, they were engaged through personal networks within the student community. Not all courses of study and faculties of the university were covered. This recruitment process has led to a strong representation of health students in the sample. It can be assumed that some participants are already aware of health-related issues and the importance of PA for health. Therefore, the generalizability of the results to the entire student community may be limited.

In this study, 74% of the participants were female, similar to other mHealth intervention studies aimed at improving PA among students [49,58,59]. A study found that female students are also more likely to engage in health promotion activities on campus than males [14]. The participants' age and BMI were similar to other studies [58-60]. Nevertheless, BMI has limitations as it does not account for body composition and can lead to health risk misestimation [61]. Self-reported height and weight can also cause BMI underestimation [62]. For this reason, BMI is not used in this study to refer to the students' physical health. The service does not provide recommendations to students based on BMI. Future evaluation of the service should consider possible underestimations of self-reported data, especially concerning the effectiveness of the service, and include objective measurement methods.

Due to the presentation of the prototype in Figma, not all of the developed content of the service could be fully experienced. As a click dummy, a clickable prototype with limited functionality, the ability to make all activities of the service tangible is limited. For example, push notifications, which will be part of the service, could not be evaluated in this format. The evaluation of activities within the service was indirect and could not be directly tested in vivo. As futur.move was evaluated in a cross-sectional survey, the results provide insights into how students perceived its use but do not give information on the actual use of the service. A longitudinal evaluation of the programmed service is needed to collect data on user behavior and the effects of the service on PA behavior and motivation in long-term use. However, this evaluation requires not only a programmed service but also its implementation in everyday campus life, even if only on a trial basis.

Implications

Future research should include analyses focused on the values and needs of younger students and men. These findings should then be incorporated into the further development of the service to better address and meet these groups' specific requirements.

The service's clarity, noted by most participants, confirms the successful layout of the application on which the usability of

XSL•FO
RenderX

the service was based. Nonetheless, feedback on comprehensibility, such as navigation buttons, should be addressed for usability. The manual data entry process requires further review with the target group, as it was negatively received and could impede long-term use. Data tracking through devices might offer a solution here. Further investigation is also needed for competitive elements such as the faculty challenge.

Despite some criticism, the overall evaluation of futur.move is positive. futur.move meets a substantial proportion of the diverse needs of the target group, underscoring the importance of adaptability.

The next step is to assess whether using futur.move actually improves students' PA. Thus, a longitudinal study should be carried out to assess real-world usability, long-term adherence, effectiveness in regards to PA, and other health-related outcomes and correlations with academic success.

Motivating students to participate in the development process remains challenging. Although 60% of surveyed students expressed interest in health programs [14], recruiting participants has been difficult. This suggests that a desire for the service does not necessarily lead to active participation in its design. Therefore, future studies should place greater emphasis on understanding the incentives and motivations that drive involvement.

Based on our findings, the UX evaluation methods used in this study could be adapted for other populations, such as individuals with chronic pain, obesity, or cardiovascular disease, to align with their needs and values in designing digital services or mobile apps that promote PA.

Conclusions

The results of this study provide relevant information on how to further develop futur.move as a multidimensional, tailor-made, and participatory mHealth service to enhance PA among university students in hybrid spaces. By emphasizing personalization—a crucial factor also identified in prior research that showed the limitations of one-size-fits-all approaches [49]—futur.move aligns well with the diverse needs of its target audience.

Overall, the UX evaluation yielded positive results for the prototype. Students particularly appreciated its personalized content, engaging social features, and thoughtful design elements. However, the UX evaluation, complemented by qualitative feedback from focus group discussions, identified 2 specific features that need significant improvement to better meet user expectations. Additionally, reducing manual data entry emerged as a critical area for optimization, as its current implementation may hinder adherence and long-term user engagement.

While correlations between UX ratings and demographic variables were found to be weak to moderate, they warrant further investigation to better address the diverse characteristics of the student population. These findings may provide essential guidance for future development steps.

Building on these results, the next phase will refine the prototype by addressing the identified areas for improvement and implementing it as a smartphone app. This app will be tested and evaluated within the university context to ensure it effectively supports students in increasing PA while maintaining a high standard of UX.

Acknowledgments

This study was funded by the "Stiftung Innovation in der Hochschullehre" (Innovation in Higher Education Foundation) under the funding line "Strengthening Higher Education through Digitalization" (grant FBM2020-VA-212-5-05380). The funders had no role in this study's design, collection, management, analysis and interpretation of data, writing of the report, and decision to submit the report for publication, and had no final authority on any of these activities. The authors would like to thank Roger Skarsten, PhD, for his support in the linguistic revision of this paper. We used ChatGPT-4o [63], developed by OpenAI (2025), to review the linguistic revisions of this paper.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

S Wittmar drafted the original version of this paper. TF carried out the quantitative analysis. S Wittmar and TF carried out the qualitative analysis. The graphics were created by VT and S Wittmar. S Wittmar, VT, PF, NK, S Wölwer, and AGMS were involved in the design of this study. All authors contributed to the drafting of the final paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
CHERRIES for the futur.move mHealth service. mHealth: mobile health.
[DOCX File , 25 KB - formative_v9i1e64384_app1.docx]

XSL•FO
RenderX

Multimedia Appendix 2
Screenshot of the customizable dashboard of the futur.move mHealth service as viewed on a smartphone in English. mHealth: mobile health.
[PNG File , 828 KB - formative_v9i1e64384_app2.png]

Multimedia Appendix 3
Task sheet for user testing to evaluate the user experience of the futur.move mHealth service in German and English. mHealth: mobile health.
[PDF File (Adobe PDF File), 185 KB - formative_v9i1e64384_app3.pdf]

Multimedia Appendix 4
Table of correlation analysis results for user experience evaluation of futur.move mHealth service using UEQ subscales, including graphical interpretation. mHealth: mobile health; UEQ: User Experience Questionnaire.
[DOCX File , 85 KB - formative_v9i1e64384_app4.docx]

Multimedia Appendix 5
Illustrative quotations from focus group discussions on the user experience of the futur.move mHealth service to support qualitative findings, presented in German and English. mHealth: mobile health.
[DOCX File , 32 KB - formative_v9i1e64384_app5.docx]

References

1. Miko HC, Zillmann N, Ring-Dimitriou S, Dorner TE, Titze S, Bauer R. Effects of physical activity on health. Gesundheitswesen 2020;82(S 03):S184-S195 [FREE Full text] [doi: 10.1055/a-1217-0549] [Medline: 32984942]
2. Warburton DE, Glendhill N, Quinney A. The effects of changes in musculoskeletal fitness on health. Can J Appl Physiol 2001;26(2):161-216. [doi: 10.1139/h01-012] [Medline: 11312416]
3. WHO guidelines on physical activity and sedentary behaviour. World Health Organization. 2020. URL: https://iris.who.int/bitstream/handle/10665/336656/9789240015128-eng.pdf?sequence=1 [accessed 2025-03-07]
4. Meng F, Xie C, Qiu F, Geng J, Li F. Effects of physical activity level on attentional networks in young adults. Int J Environ Res Public Health 2022;19(9):5374 [FREE Full text] [doi: 10.3390/ijerph19095374] [Medline: 35564768]
5. Redondo-Flórez L, Ramos-Campo DJ, Clemente-Suárez VJ. Relationship between physical fitness and academic performance in university students. Int J Environ Res Public Health 2022;19(22):14750 [FREE Full text] [doi: 10.3390/ijerph192214750] [Medline: 36429470]
6. Faselt F, Geuter G. Grundlagen zur Bewegungsförderung in Lebenswelten. Bewegungsförderung in Lebenswelten. Bielefeld: LIGA.NRW; 2011. URL: https://www.lzg.nrw.de/_media/pdf/liga-fokus/LIGA_Fokus_14.pdf#page=11 [accessed 2025-03-13]
7. Guthold R, Stevens GA, Riley LM, Bull FC. Global trends in insufficient physical activity among adolescents: a pooled analysis of 298 population-based surveys with 1·6 million participants. Lancet Child Adolesc Health 2020;4(1):23-35 [FREE Full text] [doi: 10.1016/S2352-4642(19)30323-2] [Medline: 31761562]
8. Guthold R, Stevens GA, Riley LM, Bull FC. Worldwide trends in insufficient physical activity from 2001 to 2016: a pooled analysis of 358 population-based surveys with 1·9 million participants. Lancet Glob Health 2018;6(10):e1077-e1086 [FREE Full text] [doi: 10.1016/S2214-109X(18)30357-7] [Medline: 30193830]
9. Clemente FM, Nikolaidis PT, Martins FML, Mendes RS. Physical activity patterns in university students: do they follow the public health guidelines? PLoS One 2016;11(3):e0152516 [FREE Full text] [doi: 10.1371/journal.pone.0152516] [Medline: 27022993]
10. Ráthonyi G, Takács V, Szilágyi R, Bácsné Bába É, Müller A, Bács Z, et al. Your physical activity is in your hand-objective activity tracking among university students in Hungary, one of the most obese countries in Europe. Front Public Health 2021;9:661471 [FREE Full text] [doi: 10.3389/fpubh.2021.661471] [Medline: 34604150]
11. Arias-Palencia NM, Solera-Martínez M, Gracia-Marco L, Silva P, Martínez-Vizcaíno V, Cañete-García-Prieto J, et al. Levels and patterns of objectively assessed physical activity and compliance with different public health guidelines in university students. PLoS One 2015;10(11):e0141977 [FREE Full text] [doi: 10.1371/journal.pone.0141977] [Medline: 26536605]
12. Sweileh WM. Visualized analysis of research publications on physical activity among higher education students. Am J Health Educ 2023;54(5):364-375. [doi: 10.1080/19325037.2023.2232427]
13. Hirvensalo M, Lintunen T. Life-course perspective for physical activity and sports participation. Eur Rev Aging Phys Act 2011;8(1):13-22. [doi: 10.1007/s11556-010-0076-3]
14. Meyer B. Wie geht's Deutschlands Studierenden? Gesundheitsreport. 2023. URL: https://www.tk.de/resource/blob/2149886/e5bb2564c786aedb3979588fe64a8f39/2023-tk-gesundheitsreport-data.pdf [accessed 2025-03-07]

XSL•FO
RenderX

15. Kettunen E, Kari T, Makkonen M, Critchley W, Sell A. Digital coaching among university students with low levels of physical activity: a quantitative intervention study on exercise self-efficacy. 2019 Presented at: Humanizing technology for a sustainable society: Conference proceedings; June 16-19, 2019; Bled, Slovenia p. 861-880. [doi: 10.18690/978-961-286-280-0.45]
16. Schwarz A, Winkens LHH, de Vet E, Ossendrijver D, Bouwsema K, Simons M. Design features associated with engagement in mobile health physical activity interventions among youth: systematic review of qualitative and quantitative studies. JMIR mHealth uHealth 2023;11:e40898 [FREE Full text] [doi: 10.2196/40898] [Medline: 36877551]
17. Guo Y, Bian J, Leavitt T, Vincent HK, Vander Zalm L, Teurlings TL, et al. Assessing the quality of mobile exercise apps based on the American college of sports medicine guidelines: a reliable and valid scoring instrument. J Med Internet Res 2017;19(3):e67 [FREE Full text] [doi: 10.2196/jmir.6976] [Medline: 28270378]
18. Knight E, Stuckey MI, Prapavessis H, Petrella RJ. Public health guidelines for physical activity: is there an app for that? A review of Android and Apple app stores. JMIR mHealth uHealth 2015;3(2):e43 [FREE Full text] [doi: 10.2196/mhealth.4003] [Medline: 25998158]
19. Kujala S, Roto V, Väänänen-Vainio-Mattila K, Karapanos E, Sinnelä A. UX Curve: a method for evaluating long-term user experience. Interact Comput 2011;23(5):473-483. [doi: 10.1016/j.intcom.2011.06.005]
20. Schrepp M, Hinderks A, Thomaschewski J. Applying the User Experience Questionnaire (UEQ) in different evaluation scenarios. In: Hutchison D, Kanade T, Kittler J, Kleinberg JM, Kobsa A, Mattern F, et al, editors. Design, User Experience, and Usability. Theories, Methods, and Tools for Designing the User Experience. Vol. 8517. Cham, Switzerland: Springer International Publishing; 2014:383-392.
21. Ergonomics of human-system interaction, part 210: human-centred design for interactive systems. ISO. 2024. URL: https://www.iso.org/standard/77520.html [accessed 2025-03-07]
22. Bort-Roig J, Gilson ND, Puig-Ribera A, Contreras RS, Trost SG. Measuring and influencing physical activity with smartphone technology: a systematic review. Sports Med 2014;44(5):671-686. [doi: 10.1007/s40279-014-0142-5] [Medline: 24497157]
23. Laranjo L, Ding D, Heleno B, Kocaballi B, Quiroz JC, Tong HL, et al. Do smartphone applications and activity trackers increase physical activity in adults? Systematic review, meta-analysis and metaregression. Br J Sports Med 2021;55(8):422-432. [doi: 10.1136/bjsports-2020-102892] [Medline: 33355160]
24. Michie S, Abraham C, Whittington C, McAteer J, Gupta S. Effective techniques in healthy eating and physical activity interventions: a meta-regression. Health Psychol 2009;28(6):690-701. [doi: 10.1037/a0016136] [Medline: 19916637]
25. Michie S, Richardson M, Johnston M, Abraham C, Francis J, Hardeman W, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013;46(1):81-95 [FREE Full text] [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
26. Zhang J, Brackbill D, Yang S, Becker J, Herbert N, Centola D. Support or competition? How online social networks increase physical activity: a randomized controlled trial. Prev Med Rep 2016;4:453-458 [FREE Full text] [doi: 10.1016/j.pmedr.2016.08.008] [Medline: 27617191]
27. Arigo D, Jake-Schoffman DE, Wolin K, Beckjord E, Hekler EB, Pagoto SL. The history and future of digital health in the field of behavioral medicine. J Behav Med 2019;42(1):67-83 [FREE Full text] [doi: 10.1007/s10865-018-9966-z] [Medline: 30825090]
28. Tong HL, Quiroz JC, Kocaballi AB, Fat SCM, Dao KP, Gehringer H, et al. Personalized mobile technologies for lifestyle behavior change: a systematic review, meta-analysis, and meta-regression. Prev Med 2021;148:106532. [doi: 10.1016/j.ypmed.2021.106532] [Medline: 33774008]
29. Zhou M, Mintz Y, Fukuoka Y, Goldberg K, Flowers E, Kaminsky P, et al. Personalizing mobile fitness apps using reinforcement learning. CEUR Workshop Proc 2018;2068 [FREE Full text] [Medline: 32405286]
30. Eysenbach G. Improving the quality of web surveys: the checklist for reporting results of internet E-surveys (CHERRIES). J Med Internet Res 2004;6(3):e34 [FREE Full text] [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
31. Mager B. Service Design. In: Erlhoff M, Marshall T, editors. Wörterbuch Design: Begriffliche Perspektiven des Design. Basel, Germany: Birkhäuser; 2008:362.
32. van Gemert-Pijnen JEWC, Nijland N, van Limburg M, Ossebaard HC, Kelders SM, Eysenbach G, et al. A holistic framework to improve the uptake and impact of eHealth technologies. J Med Internet Res 2011;13(4):e111 [FREE Full text] [doi: 10.2196/jmir.1672] [Medline: 22155738]
33. Bandura A. Health promotion by social cognitive means. Health Educ Behav 2004;31(2):143-164. [doi: 10.1177/1090198104263660] [Medline: 15090118]
34. Gür F, Gür GC, Ayan V. The effect of the ERVE smartphone app on physical activity, quality of life, self-efficacy, and exercise motivation for inactive people: a randomized controlled trial. Eur J Integr Med 2020;39:101198. [doi: 10.1016/j.eujim.2020.101198]
35. Messing S, Rütten A. Quality criteria for the conception, implementation and evaluation of interventions for physical activity promotion: a state-of-the-art review. Gesundheitswesen 2017;79(S 01):S60-S65. [doi: 10.1055/s-0042-123378] [Medline: 28399588]

XSL•FO
RenderX

36. Finger JD, Tafforeau J, Gisle L, Oja L, Ziese T, Thelen J, et al. Development of the European Health Interview Survey-Physical Activity Questionnaire (EHIS-PAQ) to monitor physical activity in the European Union. Arch Public Health 2015;73:59 [FREE Full text] [doi: 10.1186/s13690-015-0110-z] [Medline: 26634120]
37. What sample size do you really need for UX research? UserTesting. 2023. URL: https://www.usertesting.com/blog/ux-research-sample-size [accessed 2025-03-07]
38. Hennink MM. Focus Group Discussions. New York, NY: Oxford University Press; 2013.
39. Laugwitz B, Held T, Schrepp M. Construction and evaluation of a user experience questionnaire. In: HCI and usability for education and work.: Springer; 2008 Presented at: 4th Symposium of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society, USAB 2008; November 20-21, 2008; Graz, Austria p. 63-76. [doi: 10.1007/978-3-540-89350-9]
40. Schrepp M. User Experience Questionnaire Handbook. 2023. URL: https://www.ueq-online.org/ [accessed 2024-05-03]
41. Schulz M, Mack B, Renn O. Fokusgruppen in der empirischen Sozialwissenschaft: Von der Konzeption bis zur Auswertung. Wiesbaden, Germany: Springer VS; 2012.
42. Doering N, Bortz J. Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften. Berlin, Germany: Springer; 2016.
43. Kuckartz U. Qualitative Inhaltsanalyse. Methoden, Praxis, Computerunterstützung. 4., überarbeitete Aufl. Weinheim, Germany: Beltz; 2018.
44. Boerema S, van Velsen L, Hermens H. An intervention study to assess potential effect and user experience of an mHealth intervention to reduce sedentary behaviour among older office workers. BMJ Health Care Inform 2019;26(1):e100014 [FREE Full text] [doi: 10.1136/bmjhci-2019-100014] [Medline: 31744844]
45. Moore R, Al-Tamimi AK, Freeman E. Investigating the potential of a conversational agent (Phyllis) to support adolescent health and overcome barriers to physical activity: co-design study. JMIR Form Res 2024;8:e51571 [FREE Full text] [doi: 10.2196/51571] [Medline: 38294857]
46. Signorelli GR, Monteiro-Guerra F, Rivera-Romero O, Núñez-Benjumea FJ, Fernández-Luque L. Breast cancer physical activity mobile intervention: early findings from a user experience and acceptability mixed methods study. JMIR Form Res 2022;6(6):e32354 [FREE Full text] [doi: 10.2196/32354] [Medline: 35731554]
47. Tong HL, Quiroz JC, Kocaballi AB, Ijaz K, Coiera E, Chow CK, et al. A personalized mobile app for physical activity: an experimental mixed-methods study. Digit Health 2022;8:20552076221115017 [FREE Full text] [doi: 10.1177/20552076221115017] [Medline: 35898287]
48. Kushendriawan MA, Santoso HB, Putra POH, Schrepp M. Evaluating user experience of a mobile health application 'Halodoc' using user experience questionnaire and usability testing. J Sist Inf (J Inf Sys) 2021;17(1):58-71. [doi: 10.21609/jsi.v17i1.1063]
49. Figueroa CA, Gomez-Pathak L, Khan I, Williams JJ, Lyles CR, Aguilera A. Ratings and experiences in using a mobile application to increase physical activity among university students: implications for future design. Univers Access Inf Soc 2024;23(2):821-830 [FREE Full text] [doi: 10.1007/s10209-022-00962-z] [Medline: 36624825]
50. Buis LR, Poulton TA, Holleman RG, Sen A, Resnick PJ, Goodrich DE, et al. Evaluating active U: an internet-mediated physical activity program. BMC Public Health 2009;9:331 [FREE Full text] [doi: 10.1186/1471-2458-9-331] [Medline: 19744311]
51. Nuijten RCY, van Gorp P, Borghouts T, Le Blanc P, van den Berg P, Kemperman A, et al. Preadolescent students' engagement with an mHealth intervention fostering social comparison for health behavior change: crossover experimental study. J Med Internet Res 2021;23(7):e21202 [FREE Full text] [doi: 10.2196/21202] [Medline: 34326041]
52. Carroll JK, Moorhead A, Bond R, LeBlanc WG, Petrella RJ, Fiscella K. Who uses mobile phone health apps and does use matter? A secondary data analytics approach. J Med Internet Res 2017;19(4):e125 [FREE Full text] [doi: 10.2196/jmir.5604] [Medline: 28428170]
53. Hargittai E, Piper AM, Morris MR. From internet access to internet skills: digital inequality among older adults. Univ Access Inf Soc 2018;18(4):881-890. [doi: 10.1007/s10209-018-0617-5]
54. Neyer FJ, Felber J, Gebhardt C. Entwicklung und Validierung einer Kurzskala zur Erfassung von technikbereitschaft. Diagnostica 2012;58(2):87-99. [doi: 10.1026/0012-1924/a000067]
55. Venkatesh V, Morris MG. Why don't men ever stop to ask for directions? Gender, social influence, and their role in technology acceptance and usage behavior. MIS Q 2000;24(1):115. [doi: 10.2307/3250981]
56. Dockweiler C, Boketta R, Schnecke J, Hornberg C. Use and acceptance of mHealth- applications among young adults in Germany / nutzungsverhalten und akzeptanz von mHealth-applikationen bei jungen erwachsenen in Deutschland. 2016 Presented at: TELEMED 2015 - Erfahrungen, Stand und Perspektiven von Gesundheitstelematik und Telemedizin in Deutschland; March 09, 2015; Berlin, Germany URL: https://tinyurl.com/y2te4mn2
57. Ellis D, Allaire JC. Modeling computer interest in older adults: the role of age, education, computer knowledge, and computer anxiety. Hum Factors 1999;41(3):345-355. [doi: 10.1518/001872099779610996] [Medline: 10665203]
58. Al-Nawaiseh HK, McIntosh WA, McKyer LJ. An-m-health intervention using smartphone app to improve physical activity in college students: a randomized controlled trial. Int J Environ Res Public Health 2022;19(12):7228 [FREE Full text] [doi: 10.3390/ijerph19127228] [Medline: 35742477]

XSL•FO
RenderX

59. Davis A, Ellis R. A quasi-experimental investigation of college students' ratings of two physical activity mobile apps with varied behavior change technique quantity. Digit Health 2019;5:2055207619891347 [FREE Full text] [doi: 10.1177/2055207619891347] [Medline: 31827878]
60. Pentakota N, Ramaswamy G, Thekkur P, Nair D, Chinnakali P, Kumar Saya G. Is a smartphone application effective in improving physical activity among medical school students? Results from a quasi-experimental study. Int J Adolesc Med Health 2021;33(4):20180192. [doi: 10.1515/ijamh-2018-0192] [Medline: 30920953]
61. Aussagekraft des Body-Mass-Index ist begrenzt. Deutsches Ärzteblatt. 2023. URL: https://www.aerzteblatt.de/nachrichten/140448/Aussagekraft-des-Body-Mass-Index-ist-begrenzt [accessed 2025-03-07]
62. Draeger A, Obi N, Jagodzinski A, Becher H. Unterschätzung des BMI bei Eigenangaben zu Größe und Gewicht. Gemeinsame Jahrestagung -der Deutschen Gesellschaft für Epidemiologie e.V. (DGEpi), -der Deutschen Gesellschaft für Medizinische Soziologie e.V. (DGMS) und -der Deutschen Gesellschaft für Sozialmedizin und Prävention e.V. (DGSMP): Georg Thieme Verlag KG 2017;79(08/09):656-804. [doi: 10.1055/s-0037-1605952]
63. Hello GPT-4o. OpenAI. 2024 May 13. URL: https://openai.com/index/hello-gpt-4o/ [accessed 2025-03-14]

Abbreviations

CHERRIES: Checklist for Reporting Results of Internet e-Surveys
EHIS-PAQ: European Health Interview Survey–Physical Activity Questionnaire
HAWK: University of Applied Sciences and Arts Hildesheim/Holzminden/Göttingen
mHealth: mobile health
PA: physical activity
UEQ: User Experience Questionnaire
UX: user experience
WHO: World Health Organization

Edited by A Mavragani; submitted 17.07.24; peer-reviewed by J Fiedler, K Kammler-Sückler; comments to author 26.11.24; revised version received 14.01.25; accepted 19.01.25; published 28.03.25.

Please cite as:
Wittmar S, Frankenstein T, Timm V, Frei P, Kurpiers N, Wölwer S, Schäfer AGM
User Experience With a Personalized mHealth Service for Physical Activity Promotion in University Students: Mixed Methods Study
JMIR Form Res 2025;9:e64384
URL: https://formative.jmir.org/2025/1/e64384
doi:10.2196/64384
PMID:

©Silke Wittmar, Tom Frankenstein, Vincent Timm, Peter Frei, Nicolas Kurpiers, Stefan Wölwer, Axel Georg Meender Schäfer. Originally published in JMIR Formative Research (https://formative.jmir.org), 28.03.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Convergent and Known-Groups Validity and Sensitivity to Change of the Virtual Performance Measure in Patients With Hip and Knee Osteoarthritis: Longitudinal Study

Helen Razmjou[1], PhD; Suzanne Denis[1], MSc, PT; Susan Robarts[1], BSc, BHSCPT, MSc; Amy Wainwright[1], MSc, PT; Patricia Dickson[1], MSc, OT; John Murnaghan[1], MD, MSc, MA

Sunnybrook Health Science Centre, University of Toronto, Toronto, ON, Canada

Corresponding Author:
Helen Razmjou, PhD
Sunnybrook Health Science Centre
University of Toronto
43 Wellesley Street East
Toronto, ON, M4Y 1H1
Canada
Email: helen.razmjou@sunnybrook.ca

Abstract

Background: Subsequent to the COVID-19 pandemic in 2020, a different approach to health care utilization was required to improve safety and efficiency. In the postpandemic era, virtual care and remote assessment of musculoskeletal conditions has become more common, and examining the accuracy of these remote encounters remains vital. In 2023, an innovative, video-based tool—the Virtual Performance Measure (VPM)—was introduced to assess the functional difficulties of patients with osteoarthritis of the knee joint. Further validation of this tool is warranted to expand its application longitudinally and in more diverse populations.

Objective: This study examined the longitudinal validity of the VPM, a digitally based outcome tool, in patients with osteoarthritis of the hip and knee joints who had undergone arthroplasty.

Methods: Patients completed a web-based survey after watching 40 videos that demonstrated 10 functional tasks with increasing difficulty, prior to and at approximately 3-5 months following surgery. The Lower Extremity Functional Scale (LEFS) was used as the reference measure. Longitudinal convergent and known-groups validity as well as sensitivity to change were assessed.

Results: The data of 120 patients (n=80, 67% female; mean age 67, SD 9 years; n=58, 48% with hip osteoarthritis and n=62, 52% with knee osteoarthritis) were examined. There was a statistically significant improvement in both LEFS (t_{119}=16.04, P<.001) and VPM total scores (t_{119}=13.92, P<.001) over time. The correlation between the postoperative LEFS and VPM scores was higher (r=0.66; P<.001) than the correlation between the change scores of these measures (r=0.51; P<.001). The area under the curve value for the VPM's ability to differentiate between urgent and nonurgent candidates for surgery was 0.71 (95% CI 0.57-0.84). Sensitivity to change as measured by the standardized response mean was 1.27 (95% CI 1.09-1.45), indicating good ability to detect change over time.

Conclusions: The VPM demonstrated sufficient longitudinal convergent and known-groups validity as well as sensitivity to change in patients with hip and knee osteoarthritis following arthroplasty. This tool has a potential to improve the delivery of care by increasing access, reducing the frequency of in-person visits, and improving the overall efficiency of the health care system following a major surgery.

(JMIR Form Res 2025;9:e69001) doi:10.2196/69001

KEYWORDS

virtual; video-based outcome; longitudinal validity; sensitivity to change; osteoarthritis

Introduction

Osteoarthritis of the hip and knee joint is a disabling condition in older adults with significant physical, social, and economic burdens. Traditionally, two types of outcome measurement have been used in the assessment of patients with lower-extremity osteoarthritis. Patient-reported outcome measures document perceived physical difficulty within a specific timeframe [1].

XSL•FO
RenderX

Performance-based measures reflect the actual ability of the patients to execute a specific physical activity at the time of assessment [2]. The first video-based outcome measure that permitted a remote assessment of functional difficulties was introduced by a team of Dutch investigators in 2014 using videos with animated figures [3-5]. Subsequent to the COVID-19 pandemic in 2020, a different approach to health care utilization was required to improve safety and efficiency. In 2023, a next-generation, innovative, virtual assessment tool—the Virtual Performance Measure (VPM)—was introduced [6]. The VPM replaced animated figures with human models and combined features of self-report and performance-based tools, creating an outcome measure that could be used remotely. The original studies by the developers have shown cross-sectional validity of the VPM in patients with hip and knee joint osteoarthritis [6,7]. As clinicians continue to face challenges in care delivery and digital technology is being adopted more often by older persons in the postpandemic era, further validation of this tool is warranted to expand its application longitudinally and in more diverse populations.

The objective of this study was to examine the longitudinal convergent construct and known-groups validity as well as sensitivity to change of the VPM total score in patients with osteoarthritis of the hip and knee joints who had undergone joint replacement surgery. We explored the role of age, sex, and severity score on the postoperative VPM score after accounting for the preoperative VPM score. We hypothesized that (1) the postoperative and change scores of the VPM would correlate moderately with those of a traditional, validated, self-report functional tool; (2) the VPM would differentiate between patients who were considered urgent versus nonurgent for joint arthroplasty; and (3) the VPM score would show a large sensitivity to change following surgery.

Methods

Participants

This was a longitudinal study. All patients were recruited at a rapid-access clinic of a large arthroplasty center of an academic hospital in Canada. Inclusion criteria were failed nonsurgical management and candidacy for primary elective arthroplasty. Exclusion criteria included inability to read English, presence of cognitive conditions, lack of access to internet, or inability to use web-based surveys.

Ethical Considerations

The study protocol was approved by the Human Ethics Research Board of the Sunnybrook Health Sciences Centre (Project# 3703). All patients provided informed consent for participation in the study. The web-based survey used a participant number, no personal identifiers were collected, and study data were anonymous. Patients were not compensated for their participation in the study.

VPM Procedure

Patients who agreed to participate in the study received a link from a research assistant to a web-based survey. After entering their participant number, they completed the survey after watching 40 videos of 10 tasks on a smart phone, computer, or tablet twice—prior to surgery and 3-5 months following surgery. Activities varied from simple tasks such as sitting on a chair to physically demanding tasks such as navigating stairs, to provide a wide variety of activities and avoid ceiling and floor effects. The activities and their variations were selected based on input from physical therapists, occupational therapists, and patient representatives. The final set of the videos used an age-appropriate model.

The videos represented 10 commonly executed functional tasks with 4 levels of difficulty (normal to severe). Patients had an option of choosing “unable to perform” if the task was considered impossible to perform. The daily tasks included sitting and rising from a chair, putting on or taking off socks, getting into the shower, picking up an item from the floor, sitting down and getting up from the floor, walking on even ground, and ascending and descending stairs. The survey results were sent to the researchers via a secure platform. Further details on the development and cross-sectional validity of the tool have been provided elsewhere [6,7]. The videos are available to public (patients and researchers) [8].

Preoperative Assessment

Initial assessment was conducted by advanced practice providers, who assigned a score to each candidate using the Severity Scoring System (SSS). The SSS is a validated scoring system and acts as a guide to interpreting findings across three main elements of the standardized assessment: (1) patient’s clinical presentation, (2) self-report pain and functional difficulties, and (3) severity of osteoarthritis based on imaging. The score helps improve access and outcomes in candidates for joint arthroplasty by determining an appropriate management plan, including assigning urgency and priority levels for surgery [6,7,9,10]. The SSS has been aligned with the Wait Time Information System priority levels and classifications in Canada [11]. The clinical component score of the SSS ranges from 0-3, the functional component score varies from 0-3, and the radiological examination score varies from 0-2. The total severity score ranges between 0-8, with higher numbers indicating a higher severity. Patients with a score of 7 or higher are considered urgent for surgical management.

Reference Outcome Measure

The reference measure used for convergent validity was the Lower Extremity Functional Scale (LEFS) [12], which was completed prior to surgery and after surgery along with the VPM. The LEFS is a validated, patient-reported outcome measure with 20 questions, with a score ranging from 0-80, where higher numbers indicate better function. The LEFS has shown validity and reliability in patients with lower-extremity osteoarthritis and after hip and knee arthroplasty [6,7,9,10,12,13].

Statistical Analysis

Descriptive analyses of patient characteristics were conducted. The preoperative LEFS and VPM scores were compared with postoperative scores, and the magnitude of change over time was examined using paired, 1-tailed *t* test statistics. The analysis of covariance was used to examine the role of age, sex, and severity score on postoperative VPM scores after adjusting for

XSL•FO
RenderX

the preoperative level of functional difficulty. Interactions among independent variables were explored.

Longitudinal construct convergent validity was used to examine the correlation between postoperative scores and the amount of change (*postoperative score – preoperative score*) between the VPM and LEFS, using Spearman correlations. We hypothesized moderate correlations (r>0.30 and <0.50) [14] between the VPM total score and the LEFS score due to using different modes of measurement (self-report survey vs video-based survey).

Known-groups validity in a longitudinal context is examined by the assessment of average change in subgroups that are expected to have experienced a different amount of change in function after surgery. The receive operating characteristic curve and the area under the curve (AUC) value (area under the receiver operating characteristic curve) were used to assess the diagnostic ability of the VPM in detecting a difference in change over time between a binary predictor that indicated priority for surgery (eg, severity score ≥7 and <7). An AUC of above 0.70 is considered sufficient for diagnostic validity in clinical practice [15].

Sensitivity to change refers to the capacity of a tool to measure the magnitude of observed change statistically. We therefore measured the magnitude of change between the scores at baseline and following surgery using the standardized response mean (SRM). The SRM was calculated as *(mean change score / SD of the change score)* [14]. The Cohen criteria were used to interpret the magnitude of the SRM value [16]. An SRM of 0.80 is considered large in clinical practice.

Results

The data of 120 patients (n=80, 67% female and n=40, 33% male; mean age 67, SD 9 years; age range 40-90 years; n=58, 48% hip osteoarthritis and n=62, 52% knee osteoarthritis) were examined. Table 1 shows the demographic characteristics of the sample included. There was a statistically significant improvement in both LEFS (t_{119}=16.04, P<.001) and VPM total scores (t_{119}=13.92, P<.001). Younger age ($F_{4,115}$=9.93; P=.002) and male sex ($F_{4,115}$=4.68; P=.03) were predictive of better VPM scores at follow-up after adjusting for the preoperative VPM score and severity of the condition.

The results of the longitudinal convergent validation showed strong correlations between the postoperative LEFS and VPM scores (r=0.66; P<.001) and moderate correlation between the change scores of these measures (r=0.51; P<.001).

The logistic regression that calculated the overall diagnostic value of the VPM change score for the differentiation between high priority versus moderate or low priority (based on the SSS) showed acceptable predictive validity, with an AUC of 0.71 (95% CI 0.57-0.84; Figure 1).

The SRM was 1.27 (95% CI 1.09-1.45), indicating good sensitivity to change and ability of the VPM tool to detect a true effect of surgical management.

XSL•FO
RenderX

Table 1. Demographics, disease characteristics, and preoperative and postoperative outcomes (N=120).

Variables	Values
Age (years), mean (SD; range)	67 (9; 40-92)
Sex, n (%)	
Female	80 (67)
Male	40 (37)
Affected joint, n (%)	
Hip	58 (48)
Knee	62 (52)
Walking devices, n (%)	
Yes	35 (29)
No	85 (71)
Gait, n (%)	
Normal	10 (8)
Abnormal	103 (86)
Missing	7 (6)
Severity score, mean (SD)	
Clinical (0-3)	2 (0.48)
Functional (0-3)	1.90 (0.50)
Radiological (0-2)	1.85 (0.34)
Total score (0-8)	5.8 (10.96)
Outcome measures, mean (SD)	
Preoperative LEFS[a] score	32 (14)
Postoperative LEFS score	59 (13)
Preoperative VPM[b] score	60 (18)
Postoperative VPM score	81 (13)

[a]LEFS: Lower Extremity Functional Scale.

[b]VPM: Virtual Performance Measure.

XSL•FO
RenderX

Figure 1. ROC analysis of the VPM change score in relation to surgical urgency. ROC: receiver operating characteristic; VPM: Virtual Performance Measure. Model (0.7079), ROC1 (0.5000).

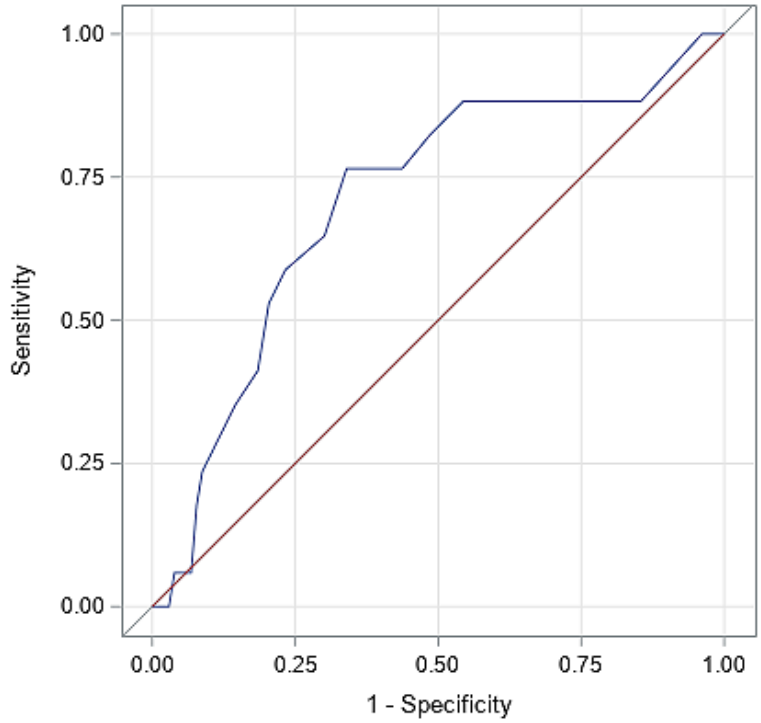

Discussion

Principal Findings

This study demonstrated longitudinal convergent construct and known-groups validity as well as high sensitivity to change in patients with osteoarthritis of the hip and knee joints who had undergone joint replacement surgery. The postoperative and change scores of the VPM correlated strongly with those of the LEFS, a self-report functional tool. The VPM score was able to differentiate between patients who were considered urgent versus nonurgent for surgery (high priority vs moderate or low priority) based on the severity score. Younger patients and male patients showed a better postoperative VPM score after accounting for the preoperative VPM score and severity of the condition.

The VPM has shown cross-sectional convergent and known-groups validity in patients with knee and hip osteoarthritis [6,7]. The cross-sectional validity, however, is not sufficient for using a tool after an intervention, and longitudinal studies are required to detect change by extending beyond a single point in time and establishing the impact of surgery on improving function. To our knowledge, this is the first study that demonstrates the longitudinal validity of a digitally based outcome measure that uses a human model. In the cross-sectional study of patients with knee osteoarthritis [6] and hip joint osteoarthritis [7], the correlations between the preoperative LEFS and VPM score were 0.66 (P<.001) and 0.68 (P<.001), respectively. In this study, there was a similar correlation between the postoperative scores (r=0.66; P<.001). The correlation between the change scores of the LEFS and VPM was lower (r=0.51; P<.001). In terms of known-groups validity, AUC values of 0.90 and 0.72 were reported in relation to candidacy for total knee and hip arthroplasty [6,7]. The lower AUC value of 0.71 of this study may be related to a higher variability in the patient sample and the stricter criteria of the predictor variable (urgency vs candidacy).

The sudden rise in popularity and demand for virtual care was a consequence of the 2020 COVID-19 pandemic, which required a different approach to health care utilization and access across the world. In the postpandemic era, clinicians continue to rely on virtual modes of examination and management to maintain efficiency and access. Thus, establishing the diagnostic validity of digital tools ensures the accuracy of clinical information obtained through virtual care. Improved access to broadband internet in rural and remote areas and higher use of new technologies by older individuals will lessen the burden of commute for families who often have to drive their older family members to arthroplasty centers, which are often located in larger cities. As technology evolves, alternative approaches to virtual functional assessment that provide accurate data on range of motion and function should be explored in osteoarthritis care. A potential tool for future virtual care is the sensorless motion tracking technology that uses cameras and computer vision software to track the movements of a patient on personal devices such as a phone, tablet, or laptop. With the use of artificial intelligence, sensorless tracking systems may provide accurate and reliable results with no additional hardware; this avoids the complexity of sensor-based systems that require wearing multiple high-cost sensors and devices and have limited ability to track related body parts, thus affecting calibration and scaling.

The implication of this study in the era of fast developing new technologies is establishing confidence in the ability of a digital tool to provide valid results, predict true change, and differentiate between groups with different characteristics and patterns of recovery, while improving access. These results will

XSL•FO
RenderX

inform further research in more diverse populations, shape policy, and help to find solutions to today's health care challenges.

This study has some limitations. Patients who participated in this study had advanced osteoarthritis that required joint arthroplasty. The results may not be applicable to patients with less pathology and higher functional levels. The model used in the videos may not represent younger patients, female patients, or patients with greater levels of obesity, and these potential differences warrant investigation in future studies.

Conclusions

The VPM demonstrated sufficient longitudinal convergent and known-group validity as well as sensitivity to change in patients with hip and knee osteoarthritis following arthroplasty. This tool has the potential to improve the delivery of care by increasing access, reducing the frequency of in-person visits, and improving the overall efficiency of the health care system following a major surgery.

Data Availability

Due to ongoing nature of the study, partial data are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

References

1. Gandhi R, Tsvetkov D, Davey JR, Syed KA, Mahomed NN. Relationship between self-reported and performance-based tests in a hip and knee joint replacement population. Clin Rheumatol 2009 Mar 8;28(3):253-257. [doi: 10.1007/s10067-008-1021-y] [Medline: 18853222]
2. Stratford PW, Kennedy DM, Woodhouse LJ. Performance measures provide assessments of pain and function in people with advanced osteoarthritis of the hip or knee. Phys Ther 2006 Nov;86(11):1489-1496. [doi: 10.2522/ptj.20060002] [Medline: 17079748]
3. Terwee CB, Coopmans C, Peter WF, Roorda LD, Poolman RW, Scholtes VA, et al. Development and validation of the computer-administered animated activity questionnaire to measure physical functioning of patients with hip or knee osteoarthritis. Phys Ther 2014 Feb;94(2):251-261. [doi: 10.2522/ptj.20120472] [Medline: 24029297]
4. Peter WF, de Vet HCW, Terwee CB. Reliability of the Animated Activity Questionnaire for assessing activity limitations of patients with hip and knee osteoarthritis. Musculoskeletal Care 2018 Sep 17;16(3):363-369. [doi: 10.1002/msc.1349] [Medline: 29770556]
5. Peter WF, Poolman RW, Scholtes VA, de Vet HCW, Terwee CB. Responsiveness and interpretability of the Animated Activity Questionnaire for assessing activity limitations of patients with hip or knee osteoarthritis. Musculoskeletal Care 2019 Dec 12;17(4):327-334. [doi: 10.1002/msc.1418] [Medline: 31402572]
6. Razmjou H, Denis S, Robarts S, Falconer J, Anunciacion M, Nunn A, et al. Virtual performance measure in osteoarthritis: an innovative transformation of patient care. Osteoarthr Cartil Open 2023 Dec;5(4):100410 [FREE Full text] [doi: 10.1016/j.ocarto.2023.100410] [Medline: 37829082]
7. Razmjou H, Denis S, Robarts S, Wainwright A, Dickson P, Roszkowski A, et al. Comparative assessment of a virtual performance measure with self-report and performance-based outcomes in patients with hip osteoarthritis. BMC Musculoskelet Disord 2025 Jan 08;26(1):30 [FREE Full text] [doi: 10.1186/s12891-025-08282-y] [Medline: 39780103]
8. Examining your functional difficulties related to your hip or knee osteoarthritis by the Virtual Performance Measure. Sunnybrook Health Sciences Centre. URL: https://sunnybrook.ca/content/?page=virtual-performance-measures [accessed 2025-03-25]
9. Robarts S, Palinkas V, Boljanovic D, Razmjou H. Validity of the virtual severity scoring system in patients with knee osteoarthritis. 2023 Apr 11 Presented at: The International Combined Orthopaedic Research Societies (ICORS), World Congress of Orthopaedic Research; September 7-9, 2022; Edinburgh, Scotland p. 51. [doi: 10.1302/1358-992x.2023.8.051]
10. Robarts S, Denis S, Kennedy D, Dickson P, Juma S, Palinkas V, et al. Patient gender does not influence referral to an orthopaedic surgeon by advanced practice orthopaedic providers: a prospective observational study in Canada. BMC Health Serv Res 2021 Sep 11;21(1):952 [FREE Full text] [doi: 10.1186/s12913-021-06965-5] [Medline: 34511124]
11. Trypuc J, Hudson A, MacLeod H. Ontario's wait time strategy: part 1. Healthc Q 2006 Mar 15;9(2):44-51, 2. [doi: 10.12927/hcq.2006.18101] [Medline: 16640132]
12. Binkley JM, Stratford PW, Lott SA, Riddle DL. The Lower Extremity Functional Scale (LEFS): scale development, measurement properties, and clinical application. North American Orthopaedic Rehabilitation Research Network. Phys Ther 1999 Apr;79(4):371-383. [doi: 10.1037/t35109-000] [Medline: 10201543]
13. Mehta SP, Fulton A, Quach C, Thistle M, Toledo C, Evans NA. Measurement properties of the Lower Extremity Functional Scale: a systematic review. J Orthop Sports Phys Ther 2016 Mar;46(3):200-216. [doi: 10.2519/jospt.2016.6165] [Medline: 26813750]

XSL•FO
RenderX

14. Liang MH, Larson MG, Cullen KE, Schwartz JA. Comparative measurement efficiency and sensitivity of five health status instruments for arthritis research. Arthritis Rheum 1985 May 11;28(5):542-547. [doi: 10.1002/art.1780280513] [Medline: 4004963]
15. Hosmer DW, Lemeshow S. Chapter 5: assessing the fit of the model. In: Applied Logistic Regression, 2nd ed. New York, NY: John Wiley and Sons; 2000:160-164.
16. Cohen J. Statistical Power Analysis for the Behavioral Sciences, 2nd ed. Hillsdale, NJ: Lawrence Earlbaum Associates; 1988.

Abbreviations

AUC: area under the curve
LEFS: Lower Extremity Functional Score
SRM: standardized response mean
SSS: Severity Scoring System
VPM: Virtual Performance Measure

Edited by A Mavragani; submitted 19.11.24; peer-reviewed by M O'Connor; comments to author 03.01.25; revised version received 10.01.25; accepted 11.03.25; published 28.03.25.

Please cite as:
Razmjou H, Denis S, Robarts S, Wainwright A, Dickson P, Murnaghan J
Convergent and Known-Groups Validity and Sensitivity to Change of the Virtual Performance Measure in Patients With Hip and Knee Osteoarthritis: Longitudinal Study
JMIR Form Res 2025;9:e69001
URL: https://formative.jmir.org/2025/1/e69001
doi:10.2196/69001
PMID:

©Helen Razmjou, Suzanne Denis, Susan Robarts, Amy Wainwright, Patricia Dickson, John Murnaghan. Originally published in JMIR Formative Research (https://formative.jmir.org), 28.03.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Ambulance Commanders' Reluctance to Enter Road Tunnels in Simulated Incidents and the Effects of a Tunnel-Specific e-Learning Course on Decision-Making: Web-Based Randomized Controlled Trial

Johan Hylander[1], RN, MD, PhD; Lina Gyllencreutz[2], RN, PhD; Michael Haney[3], MD, PhD; Anton Westman[1,4], MD, PhD

[1]Department of Diagnostics and Intervention, Surgery, Umeå University, Umeå, Sweden

[2]Department of Nursing, Umeå University, Umeå, Sweden

[3]Department of Diagnostics and Intervention, Anesthesiology and Intensive Care Medicine, Umeå University, Umeå, Sweden

[4]Department of Anaesthesia and Intensive Care Medicine, Karolinska University Hospital, Huddinge, Stockholm, Sweden

Corresponding Author:
Johan Hylander, RN, MD, PhD
Department of Diagnostics and Intervention, Surgery
Umeå University
Linnaeus väg 6
Umeå, 90736
Sweden
Phone: 46 90 786 50 00
Email: johan.hylander@umu.se

Abstract

Background: The optimal response to a major incident in a road tunnel involves efficient decision-making among the responding emergency services (fire and rescue services, police, and ambulances). The infrequent occurrence of road tunnel incidents may entail unfamiliarity with the tunnel environment and lead to uncertain and inefficient decision-making among emergency services commanders. Ambulance commanders have requested tunnel-specific learning materials to improve their preparedness.

Objective: We aimed to assess decision-making among ambulance commanders in simulated road tunnel incidents after they had participated in a tunnel-specific e-learning course designed to support timely and correct decisions in this context.

Methods: We conducted a web-based intervention study involving 20 participants from emergency medical services in Sweden who were randomly allocated to a test or control group. The control group (n=10, 50%) received a lecture on general incident management, while the intervention group (n=10, 50%) completed an e-learning course consisting of 5 modules focused on tunnel structure, safety, and collaboration in response. The participants took part in 2 simulation-based assessments for ambulance commander decision-making in major road tunnel incidents 1 month and 6 months after their allocated study intervention. In each simulation, the participants decided on the best course of action at 15 independent decision points, designed as multiple-choice questions. The primary outcome was the correct response to the question regarding how to appropriately enter the road tunnel. The secondary outcome measurements were correct or incorrect responses and the time taken to decide for each of the 15 decisions. Limited in-depth follow-up interviews were conducted with participants (n=5, 25%), and collected data were analyzed using qualitative content analysis.

Results: All 20 participants completed the first simulation, and 16 (80%) completed the second. The main finding was that none (0/20, 0%) of the participants correctly answered the question on entering the tunnel system in the 1-month assessment. There were no significant differences between the groups (P=.59; 2-sample test of proportions) in the second assessment. The e-learning course was not associated with more correct answers at the first assessment, including accounting for participant factors (mean difference between groups: –0.58 points, 95% CI –1.88 to 0.73; P=.36). The e-learning course was also not associated with a shorter time to completion compared to the nonintervention group in either assessment. Interviews identified 3 categories linked to the main outcome: *information* (lack of), *risk* (limited knowledge and equipment), and *mitigation* (access to maps and aide-mémoire).

XSL•FO
RenderX

Conclusions: Participation in a tunnel-specific e-learning course did not result in a measurable change in ambulance commanders' decision-making behavior during simulated road tunnel incidents. The observed hesitation to enter the road tunnel system may have several plausible causes, such as the lack of actionable intelligence and tunnel-specific plans. This novel approach to assessing commander decision-making may be transferable to other educational settings.

(JMIR Form Res 2025;9:e58542) doi:10.2196/58542

KEYWORDS

e-learning; major incident; incident management; disaster medicine; road tunnels

Introduction

Background

The United Nations projects that 68% of the world's population will live in urban areas by 2050, with an estimated increase from 4.2 billion to 6.7 billion people [1]. Increased urbanization and growing cities will require reliable infrastructure [2]. Traffic congestion is expensive for both communities and individuals. The analytical company INRIX reported that traffic congestion costs the average American 97 hours and US $1348 annually [3]. To limit traffic congestion and exhaust emissions, road tunnels are key to future urban development [2,4]. Traffic incidents in road tunnels may result in significant societal costs and individual health impairment.

Moreover, increasingly complex road tunnels are being constructed. For example, the Norwegian Rogfast ferry replacement project, which is 26.7 km long and 392 m deep, will become the world's longest and deepest subsea road tunnel upon completion in the early 2030s [5]. Furthermore, road tunnels can have different designs, including single or separate twin-tunnel tubes. Some tunnel systems include subsea passages and multiple on-ramp and off-ramp systems. In addition, networks of different tunnels may include emergency exits and rooms for technical equipment, shelters, and service tunnels [6]. Understanding the layout of a specific tunnel may benefit emergency services and maintenance personnel responding to a road tunnel incident.

When a major incident occurs in a road tunnel, such as a fire or a traffic crash, the evacuation of tunnel users is the first step in limiting casualties. A complex tunnel structure may hamper self-evacuation and complicate the accessibility of the responding emergency services (police, fire and rescue, and emergency medical services [EMSs]) to the incident site and injured individuals. Moreover, the compact designs of modern road tunnels contribute to unique medical challenges concerning specific injury patterns, specific trauma mechanisms, and increased extraction times [7-9]. Hence, accessibility issues may result in considerable delays in vital treatment, potentially leading to preventable loss of life.

Timely and adequate decision-making by emergency service commanders to facilitate correct task priority and minimize time to treatment is crucial and difficult. This is illustrated by the 2013 Norwegian Gudvanga monotube tunnel fire, where a decision was made to ventilate smoke to aid firefighter access, which, in turn, engulfed 67 evacuating persons in smoke, resulting in acute smoke inhalation injuries [10].

A study by Kristiansen et al [11] exploring the collaboration between responding emergency service commanders during a Norwegian road tunnel exercise concluded that the ambulance commanders were treated as outsiders in the decision-making process owing to uncertainty regarding their management roles. Similarly, in Sweden, an imbalance between the incident commanders and ambulance or medical commanders concerning specific road tunnel knowledge has been identified as a collaborative obstacle [12]. As important decisions need to be taken within the initial 20 minutes of the incident, this obstacle may lead to further delay [13]. The current prehospital medical management system in Sweden uses a general all-hazard approach. This means that the same management method is used regardless of the incident type or environment [14,15]. In contrast, Norwegian fire and rescue services include educational material on decision-making, task allocation, risks, human behavior, and coping with uncertainties in their training curriculum [16]. Swedish ambulance staff with road tunnels in their catchment areas are expected to initiate incident management in the event of a road tunnel incident, even without tunnel-specific knowledge.

Knowledge of the tunnel system and the ability to understand emerging situations are important for fire and rescue service commanders and personnel when responding to road tunnel fires [17]. Studies concerning EMS incident management in road tunnel incidents in Norway [18] and Sweden [19] have found that there is a perceived need for tunnel-specific and readily accessible learning materials and training courses. These are needed to be better prepared to optimally assess response needs in a complex environment and make timely decisions to facilitate early access to the injured persons.

Objective and Hypotheses

The main aim of this study was to evaluate whether participation in a tunnel-specific e-learning course affected ambulance commanders' abilities to make appropriate decisions in simulated road tunnel incidents. The secondary aim was to explore participant confidence in decision-making through structured questions and responses.

Hypothesis 1

Hypothesis 1 is that the tunnel-specific e-learning course will lead to more participants correctly choosing to enter the road tunnel system in the 1-month postcourse simulation-based assessment compared to those who have not participated in the course. Furthermore, the same is hypothesized for the second assessment, 6 months after the e-learning course intervention.

XSL•FO
RenderX

Hypothesis 2

Hypothesis 2 is that the tunnel-specific e-learning course will lead to participants giving more correct answers across the 15 decision points for the simulated tunnel incidents (fire or car crash) compared to those who have not participated in the course, taking into account relevant participant factors, including educational level and previous practical experience with tunnel incidents.

Secondary Hypothesis

The secondary hypothesis is that tunnel-specific e-learning course participation will lead to a shorter time to make decisions in the simulation-based assessments compared with no e-learning course participation.

Methods

Study Design

The difficulty in evaluating the effectiveness of educational efforts in the medical management of major incidents has been described in the literature [20]. This study has taken an innovative approach to this issue. A schematic overview of the study design is presented in Figure 1. The design of this study is a 2-arm, assessor-blinded, randomized, prospectively controlled trial. The intervention was web based and accessed by individual participants via a purpose-built online platform. This study adhered to the CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) reporting guidelines [21,22].

Figure 1. Schematic overview of the study design. This study is a 2-arm, assessor-blinded, randomized, prospectively controlled trial. Participants (N=20) were enrolled and randomized to either a prerecorded lecture or the tunnel-specific e-learning course. Both groups completed follow-up simulations at 1 and 6 months, which included 15 decision points (each designed as a multiple-choice question). Outcomes included the decision to correctly enter the tunnel, the number of correct decisions, and time. Participants also self-evaluated their decision-making ability and sense of security in handling road tunnel incidents.

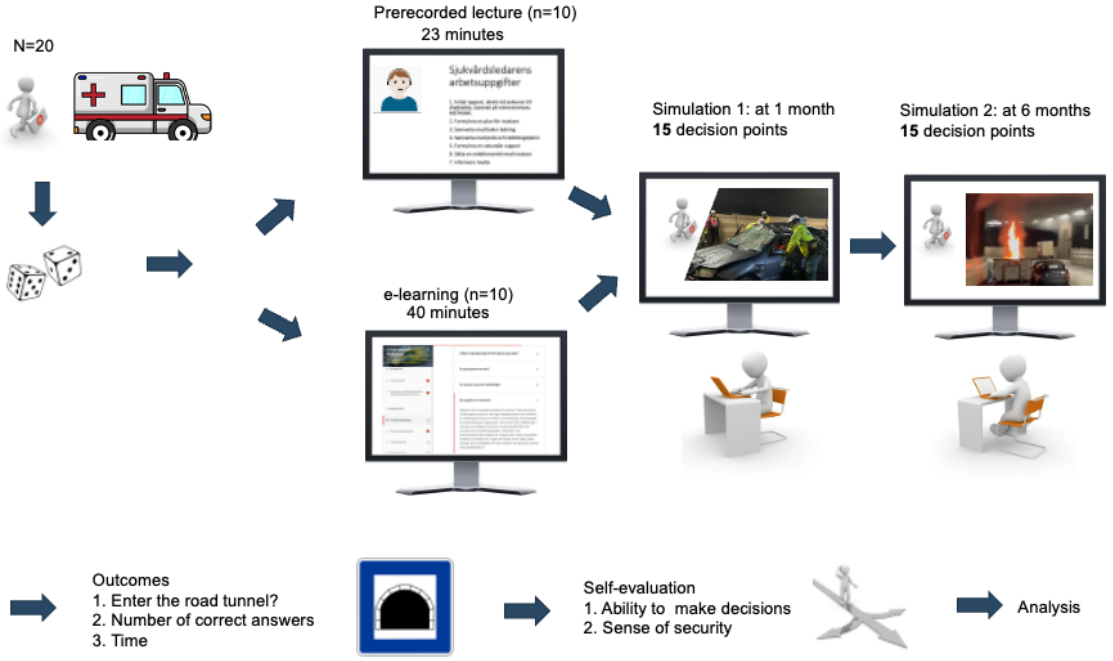

Ethical Considerations

This study was approved by the Swedish Ethical Review Authority (registration numbers 2021-04810 and 2022-05388-02) and conducted in accordance with the Declaration of Helsinki [23]. As this study examined the effect of an educational course on health care provider performance in a simulation, trial registration was not deemed necessary, as described by the International Committee of Medical Journal Editors recommendations [24]. Potential participants were provided with written information, including informed consent descriptions, regarding the aim of the study and the voluntary nature of participation. They were also informed that deidentified data (eg, demographic data, such as sex and educational level) would be included in a scientific article. In addition, they were given the contact information of the study's data protection officer if, after enrollment, they wished to have their personal information removed. Potential participants were also informed that they would not receive any compensation for participation. Informed consent was obtained from the participants before they were given access to their allotted course material (refer to the Study Sequence section and Multimedia Appendix 1 for further details).

Setting

This study was conducted between March and December 2022, with participants from a large city in southern Sweden that has several twin-tube road tunnels in its catchment area. In the studied city, the EMS comprised emergency ambulances, physician-staffed vehicles, a dedicated incident management

XSL•FO
RenderX

unit, and a helicopter. Only staff from the dedicated incident management unit and emergency ambulances were included as study participants. These units were staffed by either a registered nurse and an emergency medical technician or 2 nurses. The educational level of those who staff emergency ambulances in Sweden varies from emergency medical technicians to registered nurses with a 1-year subspecialization, such as prehospital care. The EMS personnel are trained in advanced life support and prehospital management leadership. The prehospital management leadership course provided for Swedish EMS staff is a mandatory 2-day course covering the basics of incident management and draws from the Major Incident Medical Management and Support (MIMMS) course [14,25]. A separate 2-day leadership course focusing on stress resilience and leadership traits is available, although this course is not mandatory or nationally implemented [15].

In Sweden, including the studied region, the standard operating procedure (SOP) concerning major incidents is that the first arriving ambulance crew at an incident site initiates "command and control" procedures by assuming the roles of ambulance and medical commanders [25]. Consequently, all personnel who staff emergency ambulances need to be able to initiate command and control.

Sample Size Calculation, Participants, and Randomization

Sample Size Calculation

The sample size calculation was based on findings from local pretesting and recent literature [18,19], where ambulance commanders highlighted concerns about tunnel risks that could lead to hesitation in entering the road tunnel system. On the basis of this, the control group was expected to respond with close to 0 (10% response rate) correct answers to the critical categorical question (decision to correctly enter the road tunnel) and, similarly, a close to 0 (10%) correct response rate to the 15 tunnel-specific decision points (given their lack of tunnel-specific education and unfamiliarity with the tunnel environment). The intervention group was expected to correctly answer the critical single categorical question and to have a general correct response rate of 75% for all questions. Using a power of 0.8 and a 2-tailed α of .05 with the Pearson chi-square test, the estimated sample size was 8 participants per group. We aimed to include a minimum of 10 participants in each group owing to anticipated dropouts after inclusion.

Inclusion and Exclusion Criteria

The inclusion criteria were participants who were currently employed by the regional EMS and assigned to the ambulance service or dedicated incident management unit as either registered nurses or emergency medical technicians. The participants could also assume the role of ambulance commander in road tunnel incidents, in accordance with the SOP. The exclusion criteria were individuals who were involved in the development or validation of the tunnel-specific e-learning course for this study and those who were diagnosed with or had self-perceived posttraumatic stress disorder or anxiety due to road tunnel incidents.

Recruitment Process

The ambulance station commander assisted in informing potential participants of the regional study. A letter of invitation to participate in the study was distributed via "a weekly newsletter" in the organization. Furthermore, short presentations concerning the aim of and participation in the study were conducted by the first author during the organization's morning briefings. In total, 23 participants were recruited during a 3-month period (March to May 2022), of whom 20 (87%) participants were included in the study, while 3 (13%) were in reserve. Before the intervention was initiated, 3 (13%) planned participants withdrew, and the 3 (13%) individuals in reserve were included as participants.

Randomization

The first author collected the participants' email addresses. The 20 participants' email addresses were sorted in alphabetical order and assigned a case number. The participants were randomly allocated to the intervention or control arm using a random number generator. A randomization key containing case identification details was stored in a locked place to which only the first author had access.

Course Material and Study Sequence

Online Platform

An online platform was created for this study using Moodle (Moodle HQ) open-source software [26]. The online platform contained both the e-learning course and the control (nonintervention) "course." Participants were given individual access to one of the courses after a randomized allocation. Furthermore, both simulations were stored on the platform (hidden from the participants) and became available during the assessment periods (1 and 6 months after the educational intervention). Once started, neither simulation could be aborted. Thus, the participants had 1 opportunity to complete each simulation. The participants were informed of this multiple times in writing.

Prerecorded Lecture (Control Arm)

The control course consisted of a 23-minute prerecorded lecture on general incident management. The lecture consisted of a brief historical overview of major incidents where incident management was highlighted as important, including examples from the 2011 bombing in Oslo, Utøya island shooting in Norway, and the 2017 truck attack in Stockholm, Sweden [27,28]. The lecture emphasized the roles of the first ambulance at the incident site, focusing on the ambulance commander–specific tasks according to the MIMMS concept, with emphasis on the mnemonic METHANE (major incident, exact location, type of incident, hazards, access and egress, number of casualties, and emergency services) [25]. The use of the mnemonic METHANE has been an SOP in Swedish EMS prehospital incident management for the past 20 years [29]. As such, the general lecture did not include new learning or tunnel-specific elements. The purpose of the control course was to refresh the participants' general knowledge of incident management in major incidents. The control course was validated by 5 EMS personnel who had extensive prehospital experience from parts of the country outside the studied region.

XSL•FO
RenderX

The personnel were asked to provide feedback on the content of the course and perceived usefulness to ambulance commanders. Feedback included comments, for example, "Please describe MIMMS a bit more, otherwise the lecture contains much that I already knew." Changes were made according to their feedback.

The Tunnel-Specific e-Learning Course (Intervention Arm)

The intervention in this study consists of a tunnel-specific and gamified e-learning course created using Articulate Global software [30]. The e-learning course was developed in collaboration with the Faculty of Humanities at Umeå University, Sweden.

The aim of the e-learning course was to increase ambulance commanders' ability to make the proper decision at the right time by increasing knowledge of the specific tunnel environment. A particular concern was giving participants the confidence to enter the road tunnel in a safe manner, for example, using the unaffected tunnel tube during a major incident, which, according to discussions among stakeholders, is an effective method for gaining access to the scene and those injured.

The e-learning course (content outline is provided in Multimedia Appendix 2) consisted of materials concerning collaboration in a tunnel-specific context and organizations' different tasks from dispatch to arrival at the incident scene [18,19,31], methods of identifying injured individuals (eg, triage), principles for treating injuries (eg, smoke inhalation), and specific risks (eg, explosions) [32-34]. The e-learning course content was validated by 8 stakeholders from organizations involved in road tunnel rescue efforts, including fire and rescue services, EMS, police, emergency dispatch centers, and infrastructure owners. One or 2 participants from each organization participated. In addition, the course content was reviewed by 2 individuals with in-depth knowledge of prehospital education and extensive experience in acting as ambulance commanders in a variety of scenarios. Changes were made according to their feedback.

The e-learning course consisted of 5 distinct modules with learning materials on different aspects of road tunnel rescue efforts, including the road tunnel structure, potential injury patterns, safety, and collaboration, which were determined after the validation process. The modules were, in collaboration with an expert in pedagogy, incorporated into a gamified e-learning setting. This design was deemed suitable as an element of gameplay that may stimulate learning [35]. The learning materials consisted of text, images (eg, illustrating how to correctly position an ambulance in a major road tunnel incident), and short videos (eg, showing an example of fire development in a road tunnel fire) combined with module-specific quizzes (respondents needed to score 80% correct answers to pass each module) to facilitate learning [36] (refer to Multimedia Appendix 3 for an example).

Road tunnel layouts, different injury patterns, collaboration, and aspects of personal and scene safety were integrated into the e-learning course modules. Using the revised version of the taxonomy of learning by Bloom, which was proposed by Krathwohl [37], study participants needed to *understand* (eg, the layout of the tunnel, tunnel-specific risks, and probable injury patterns), *analyze* (how to collaborate with other organizations to get a clearer picture of the incident and incident-specific risks), and *evaluate* (the tunnel layout, risks, and specific injury patterns) to be able to make an informed decision regarding where the injured person should be transported for definitive care. As decision-making is a complex cognitive process, different aspects were covered in the e-learning course. For example, risk management included the identification of environmental risks (smoke and leakage of flammable or corrosive fluids) and personal risks (inhalation of toxic fumes and limited visibility in a smoke-filled environment). Furthermore, risk-mitigation measures (eg, how to position ambulances in the tunnel system) and tunnel safety features (fire suppression systems and jet fans) were described. Under the course segment "Should I enter the road tunnel?" participants were informed of the importance of conducting a risk assessment before entering the road tunnel system and that using the unaffected tunnel tube (in case of a fire or major incident) was considered safe.

The completed e-learning course was validated by 5 EMS staff members (who also validated the control course) based on language, logical progression, design, and overall usability. The overall impression of the course design was described as "appealing." The course was described as "easy to navigate" and "fun," although "not designed to be used via smartphone." Changes were made according to validating EMS staff members feedback; for example, those included in the study were advised not to use smartphones while taking their allotted courses or simulations.

Prospectively Controlled Trial

The trial consisted of a knowledge test based on the tunnel-specific intervention's course content, which was tested in 2 digital simulations (assessment points) conducted 1 month and 6 months after the intervention. The simulations also included a time measurement for decision-making. Time was measured individually for the 15 separate decisions in each simulation. Between the first and second knowledge tests (assessment points), there was no control over for what the participants read, studied, or discussed in relation to tunnel incident responses.

Knowledge Test

The knowledge test was developed based on the literature on performance indicators [38-40], interview findings [18,19,31], and focus group discussions among stakeholders [13]. A total of 15 independent multiple-choice questions (MCQs), each with 1 correct option out of 5 alternatives, were created. This design was chosen because MCQs are commonly used in medical education to assess higher-order thinking, so the participating personnel should be accustomed to the method [41]. In total, the participants could get 15 points per simulation. The questions were formulated as decision points (Multimedia Appendix 4) and were directly related to the content of the intervention (e-learning course). The questions were evaluated during the pilot testing of the digital simulations.

XSL•FO
RenderX

Simulations

The digital simulations were created in collaboration with an expert in pedagogy and by integrating learning materials into an e-learning environment. The content was created using Microsoft PowerPoint (version 16.71) and modified using Storyline 360 (Articulate Global, LLC). The 2 simulations of major road tunnel incidents were visualized through a combination of text, images, and videos (eg, exercises); audio (dispatch calls and alarm signals); and animations to add a sense of realism. To mimic a real tunnel incident, the simulations followed the same logical, decision-based pathway, starting from dispatch and continuing through to the conclusion of a rescue effort. One example of a decision-based question is "Do you enter the road tunnel"? (Figure 2).

Figure 2. Decision from one of the simulations on how to enter the road tunnel, designed as a multiple-choice question with 5 different choices, one of which is correct. The text is provided in Swedish, with the corresponding English translation shown to the right (the correct answer is underlined).

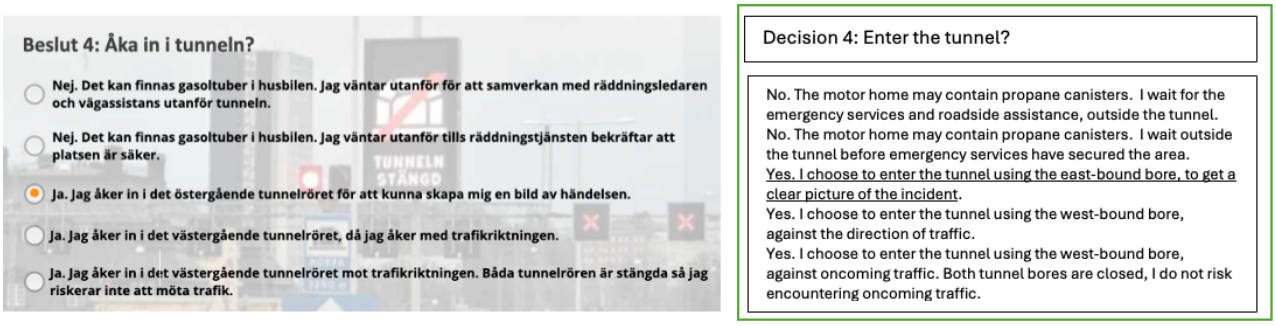

The simulations were designed as forward-moving only. The aforementioned kind of modified essay question is commonly used in medical education, where students are presented with a case and gradually presented with more information as the case progresses, similar to unfolding an incident where initial information may be scarce [42]. The scenario unfolds regardless of what the participant answers, and the participant cannot go back and change answers to the previous questions. During the simulation, participants received updated information via an audio clip (eg, prerecorded dispatch call) or an image containing information on certain decisions made by others, for example, "the road traffic center decided to close the road tunnel." Actual footage or video sequences from road tunnel interiors were also included.

Two different simulations (1 month and 6 months after the educational intervention) were created. In the first assessment (at 1 month), the simulated road tunnel incident was a collision between a bus and a passenger car, with numerous casualties. The scenario did not include fire; smoke; or added risks, such as leakage of fuels or dangerous cargo. In the second assessment (at 6 months), the scenario was a collision between a mobile home and multiple vehicles, again with numerous casualties. The mobile home contained a liquefied petroleum gas cylinder as an added risk element. This scenario developed into a road tunnel fire at a later stage. These 2 scenarios were chosen based on the experience that most fatalities in tunnel incidents result from traffic crashes and fire [43].

The simulations were validated separately by 5 EMS personnel (external validation) and 2 stakeholders (internal validation) who had in-depth education in prehospital management and extensive experience in acting as ambulance commanders. Their feedback included comments such as "The time pressure was noticeable, I chose to read instead of listen...I lost focus...But the same goes for a real incident." After the feedback was considered and changes were made to the simulations, a final version was established by the coauthors. No changes were made to either course once data collection from study participants was started.

Besides the choice of decision in each simulation, the time taken to make each decision was also recorded. This was a feature available in the software used in the development of the simulations (Storyline 360). When the participant was presented with the MCQ options, a hidden timer started, and the timer stopped when the participant chose an alternative and clicked on "next" to advance to the next question. The timing for each question was validated separately by the first author and the e-learning expert using stopwatches. Time was also measured during the evaluation of the simulations.

Study Sequence

Individual emails containing information about the study, together with a unique link to either the control or intervention course, were sent to the participants. When the participants clicked on the link, they were directed to a website (where the course platform was located) and prompted to log in. After logging in and before gaining access to their allotted course, the participants were asked to answer a short survey regarding their demographics: sex, level of education (registered nurse or emergency medical technician), additional 2-day leadership course, the number of years in EMS, and experience in managing tunnel incidents. Furthermore, the participant had to give consent by ticking a checkbox before completing the survey. Upon completion of the survey, participants could open their respective courses and had 2 weeks to access the content and complete the course.

Approximately 1 month after the participants completed their respective courses, they were sent an email with an invitation to log onto the e-learning website to complete a simulation and assessment, again with the expectation of participating within 2 weeks. As described earlier, the scenario contained 15 decision points where the participant could decide on the best course of action, choosing from multiple choices, with 1 correct answer.

XSL•FO
RenderX

The participants' responses and the time taken to make each decision were recorded.

Six months after the participants completed their respective courses, they were sent an email prompting them to log in to the e-learning website to complete the second simulation and assessment (scenario details given earlier) and again within 2 weeks. Participants who had provided their phone numbers were sent an SMS text message reminder. The participants' decision responses were recorded in the same manner.

Evaluation Form

Upon completion of the second simulation, the participants were redirected to an evaluation form and asked to subjectively assess (decreased, unchanged, or increased) whether the course (intervention or control) influenced their ability to make decisions and affected their sense of security as ambulance commanders in road tunnel incidents. In total, 2 email reminders of participation were sent after 5 days and 10 days. The data collection period was extended by 1 week because of a limited response rate, and the time extension yielded 5 additional responses.

Follow-Up Interviews

A limited number of in-depth follow-up interviews were conducted with participants (4/10, 40% from the intervention group and 1/10, 10% from the control group) within 2 months from the second simulation. The interviews were conducted using a semistructured interview guide and focused on participants describing their reasoning behind the decision to enter or not enter the road tunnel in both scenarios (Multimedia Appendix 5). The collected data were analyzed using the method of qualitative content analysis [44], where the transcribed data were broken down (deconstructed) into meaning units. Next, each meaning unit was labeled with a code, and similar codes were grouped into subcategories and categories.

Outcomes and Analysis Plan

Outcomes

The outcomes were defined as choosing the correct decision when presented with MCQs and the time spent making each decision. The first primary outcome was the correctness of responses to the question concerning the ambulance commander's decision to enter the road tunnel. The second primary outcome was the number of correct responses across the 15 different decision points. The secondary outcome was the time taken to register each decision.

After the conclusion of the simulation-based assessments, participants were asked to self-evaluate whether they felt that the e-learning course or control course influenced their ability to make decisions in general in road tunnel incidents (decreased, unchanged, or increased) as well as influenced their sense of safety in acting as ambulance commanders in road tunnel incidents (decreased, unchanged, or increased).

Data Curation and Availability

Pseudonymized participant test scores and response times were electronically recorded. The dataset was stored on an encrypted hard drive (2 TB; ADATA Technology Co, Ltd).

Statistical Analysis

Statistical analysis was performed using SPSS Statistics software (Macintosh version 28; IBM Corp). A 2-sample proportion test was used to compare the grouped averages of proportions. $P<.05$ was considered significant for identifying the difference between groups. Multivariate logistic regression was used to assess the associations between the primary outcome and participant factors as well as between secondary outcomes and participant factors.

Results

Participants

A total of 20 participants participated in the course to which they were assigned and the first simulation-based assessment. Of these, 16 (80%) participants, with 8 (40%) in each group, participated in the second simulation-based assessment at 6 months. However, 1 (5%) participant in the control group did not complete the assessment, and data from this participant were not included in the grouped responses for the second assessment (Figure 3). Participant demographics, such as sex, years of experience, and educational level, are provided in Table 1.

XSL•FO
RenderX

Figure 3. Study flowchart. Description of the number of participants enrolled and lost to follow-up during the 6-month study period.

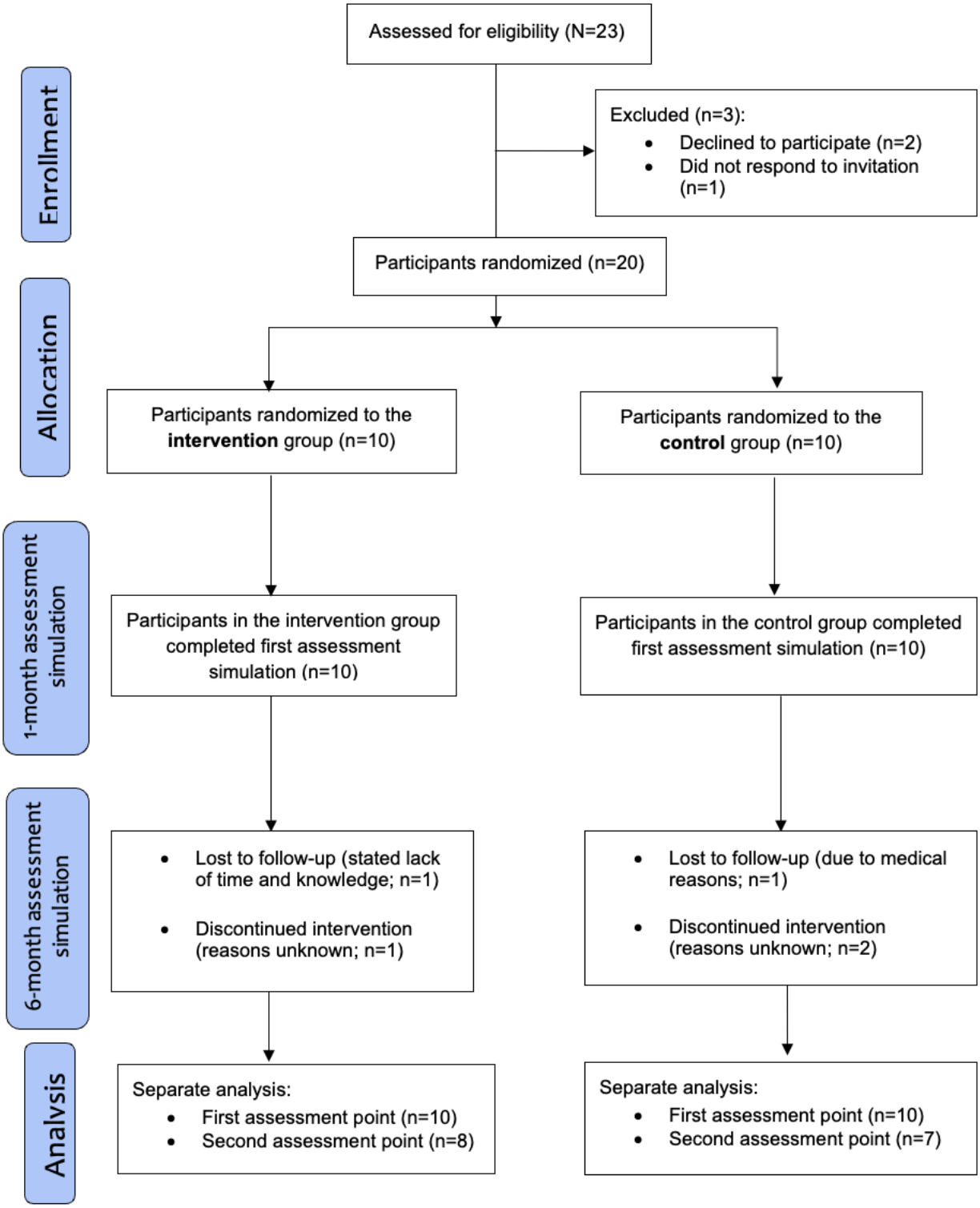

XSL•FO
RenderX

Table 1. Participant demographics, including sex, years of experience, level of education, education in prehospital management, whether they have attended a 2-day separate leadership course, and whether they have experience in managing road tunnel incidents (N=20).

Descriptive data	Intervention group (n=10), n (%)	Control group (n=10), n (%)
Sex		
Female	2 (20)	2 (20)
Male	8 (80)	8 (80)
Experience (y)		
5-10	4 (40)	6 (60)
10-15	3 (30)	1 (10)
15-20	2 (20)	0 (0)
>20	1 (10)	3 (30)
Level of education		
Registered nurse	1 (10)	0 (0)
Specialist nurse	8 (80)	8 (80)
EMT[a]	1 (10)	2 (20)
Education in prehospital management	9 (90)	10 (100)
Attended a 2-day separate leadership course	2 (20)	2 (20)
Have experience in managing road tunnel incidents	3 (30)	3 (30)

[a]EMT: emergency medical technician.

Main Results

The main finding was that none (0/10, 0%) of the participants in the control group (without the tunnel-specific e-learning course) and the intervention group (with the tunnel-specific e-learning course) responded correctly to the single categorical question to enter the road tunnel. The second main finding was that the average number of correct answers among the 15 decision points was 7.0 (SD 1.8) for the control group at the first assessment point (1 month) and 6.5 (SD 1.8) for the intervention group, with no difference between the groups (P=.86, 2-sample test of proportions).

Secondary Results

Tunnel-specific e-learning course exposure was not associated with more correct answers in the first or second simulation assessments when participant factors were considered in the multivariable logistic regression analysis (Table 2). Participant factors included sex, work experience, educational level, participation in a separate 2-day leadership course, and having practical experience in tunnel management. The findings showed that there were associations between more correct answers and having participated in a separate 2-day leadership course, female sex, and practical experience, at least at the first assessment point. The total response time at the first assessment point was not associated with course exposure (P=.80), even when considering other participant factors.

At the 6-month assessment point, 25% (2/8) of the participants in the intervention group and 14% (1/7) in the control group chose to correctly enter the road tunnel, with no difference between the groups (P=.59). Multivariable logistic regression analysis identified no factors associated with the total number of correct responses for either group.

Of the 20 participants, the evaluation form was completed by 11 (55%) participants (control group: n=6, 30% and intervention group: n=5, 25%). In the intervention group with 5 participants, confidence in acting as ambulance commanders was reported to be unchanged (n=3, 60%) or to have increased (n=2, 40%). In the control group with 6 participants, the confidence was reported to be unchanged (4/6, 67%) or to have increased (n=2, 33%). Regarding the course's impact on decision-making, 3 (60%) participants in the intervention group reported it as unchanged, and 2 (40%) reported it as increased. By contrast 2 (33%) participants in the control group reported it as unchanged, and 4 (67%) reported it as increased. None of the participants reported decreased ability to make decisions or confidence in handling road tunnel incidents.

The follow-up interviews identified 3 categories linked to the decision not to enter the road tunnel correctly: *information* (lack of updated information, risk of misunderstandings, and unavailability of information on the actual tunnel system), *risks* (limited knowledge of site risks, fear of being questioned by EMS personnel if they decided to enter the tunnel, and lack of proper equipment), and *mitigation* (need of site-specific maps, access to updated course material annually, and access to aide-mémoire).

XSL•FO
RenderX

Table 2. Multivariable logistic regression (including demographic data, such as sex, work experience, educational level, additional leadership course, and experience with tunnel incidents), number of correct answers, and total response time at both assessment points.

Assessment point	Unstandardized β	B, 95% CI	*P* value
Assessment point 1			
Total correct answers			
Course exposure	–.58	–1.88 to 0.73	.36
Sex	2.10	0.39 to 3.79	.20
Work experience	–1.25	–2.18 to –3.25	.01
Educational level	.75	–0.23 to 1.73	.12
Attended a 2-day separate leadership course	3.05	0.55 to 5.55	.02
Having practical experience in tunnel management	1.09	–0.46 to 2.64	.15
Total response time (s)			
Course exposure	44.18	–320.59 to 408.96	.80
Sex	–216.16	–692.90 to 260.59	.35
Work experience	–42.72	–302.35 to 216.91	.73
Educational level	–71.07	–346.09 to 665.95	.59
Attended a 2-day separate leadership course	–35.22	–736.39 to 665.96	.92
Having practical experience in tunnel management	308.37	–125.75 to 742.48	.15
Assessment point 2			
Total correct answers			
Course exposure	–.28	–2.91 to 2.36	.82
Sex	1.18	–3.20 to 5.57	.55
Work experience	.30	–3.03 to 3.09	.98
Educational level	–.57	–2.52 to 1.38	.42
2-day separate leadership course	–3.21	–11.87 to 5.45	.52
Practical experience in tunnel management	.98	–2.60 to 4.56	.55
Total response time (s)			
Course exposure	151.07	–236.08 to 538.21	.39
Sex	42.21	–603.22 to 687.63	.88
Work experience	–420.39	–870.51 to 29.74	.06
Educational level	162.94	–123.93 to 449.80	.23
2-day separate leadership course	1255.37	–19.22 to 2529.95	.05
Practical experience in tunnel management	531.25	4.42 to 1058.08	.05

Discussion

Principal Findings

The main finding of this study was that none (0/20, 0%) of the participants correctly answered the question regarding entering the road tunnel at the first assessment point. This was unexpected.

Despite the tunnel-specific e-learning course specifically addressing how to use the unaffected tunnel tube as one of the important course learning goals, it did not succeed in supporting this expected behavior. There are several possible and different explanations for this result. One could be that most (14/20, 70%) participants lacked any experience with real tunnel incidents, and actual experience may be required to decide to enter the road tunnel properly. This is supported by other reports [45,46], which indicate that decisions often need to be based on experience. Inexperienced decision makers may depend on mnemonics and guidelines [47], and inexperienced commanders may become more hesitant in their decision-making if the guidelines or mnemonics are not adjusted to the situation or environment. Easily accessible and adequate guidelines may facilitate decision-making in this particular setting. This is supported by the participant interviews, indicating a need for adequate aide-mémoire tailored to the specific tunnel environment. Adding useful mnemonics to the tunnel-specific e-learning course may facilitate participant decision-making.

XSL•FO
RenderX

Another possible explanation for participants not actively choosing to enter the road tunnel may be concerns about personal safety. This is in line with the findings from tunnel incident reports [13,18] where incident scene safety is an important concern for the ambulance commander. In addition, it can be perceived as correct to initially wait when there is limited information about site risks. This was illustrated by the aforementioned 2013 Gudvanga tunnel fire where a large portion of the tunnel became filled with smoke due to reversed jet fans [10]. However, in the event of a fire in a twin-tube road tunnel, the unaffected tube should be regarded as a safe environment, as fans ventilate air in the same direction in both tubes, which is expected to prevent smoke spillover to the opposite tube. The unaffected tube as a safe environment was emphasized multiple times during the tunnel-specific e-learning course, which may have influenced the participants to adopt a more cautious approach. Using the opposite tube would also result in bringing the EMS personnel closer to the incident, which could be beneficial for the treatment of critical injuries, such as internal bleeding or smoke inhalation injuries. It seems that this approach needs to be addressed further in future educational efforts. For example, on-site visits could be an important element of education and provide the EMS personnel with a sense of security, which has been indicated by others [18]. Other cost-effective solutions, such as “visiting” road tunnels in a virtual environment, may create a sense of understanding of the tunnel layout.

Other possible explanations for participants choosing not to enter the road tunnel in the simulations could be related to the technical interface of the simulation medium, simulated scenarios, or the technical aspects of the assessment, which were based on MCQs. Because of this design, we could not confirm what the study participants understood in a granular manner. Participants may have considered the correct answer to the question concerning tunnel entry. There is a 20% chance of answering correctly by simply guessing. However, in this case, 0% (0/20) did, which may indicate a conscious choice to not enter the road tunnel. However, there might be challenges regarding responder clarity with MCQ formats, where the responder simply failed to convert their reasoning to an intended MCQ response. MCQs can often be improved for clarity [48,49], for example, by gathering experience with responses and modifying questions to improve precision in assessing specific reasoning [50]. Ideally, a more detailed and interactive verbal assessment could provide a more in-depth understanding of participants’ reasoning to assess their competence [51]. In this first evaluation of the simulations, MCQs were chosen to limit subjective assessor bias and facilitate participant accessibility.

Another possible explanation for not choosing to enter the road tunnel early is that the tunnel-specific e-learning course was not sufficiently clear or effective in generating specific knowledge among the participants on when to enter the road tunnel system. Although the course content was reviewed before this study, there was general agreement among the convenience-sampled stakeholders that the learning objectives were clear. It is possible that there was some knowledge decay over the weeks before the 1-month assessment. During the course, learning objectives were reinforced at multiple points. However, it is possible that the learning objectives were not adequately reinforced during the course. This result is important for the further development of this type of e-learning course, which is potentially going to be part of the formal training within the EMS system as commissioned by the Swedish Transport Administration.

Both the tunnel-specific e-learning course and the assessment were conducted in a digital environment, and the practical details of how the participants connected to the course and assessment could have influenced their learning and performance. The e-learning course and simulations used in this study were designed for use on PCs; however, how the participants accessed them was not controlled or reported. Participants’ experiences with e-learning formats, materials, or assessment methods are likely to influence learning and performance, where less familiarity and comfort with e-learning media could impede learning. A Cochrane review concluded that e-learning was no more (or less) effective than traditional learning in improving health professionals’ skills [52]. Developing this type of course or preparing for responses in complex and dynamic settings is challenging [53]. The applicability of a tunnel-specific e-learning course or SOP in a real road tunnel incident in Sweden needs to be studied further.

Secondary Results

There was poor performance at the 6-month assessment point in both groups for the primary outcome, and the total number of correct responses did not differ between groups or between the same group at 1 month. Furthermore, secondary findings showed that participant factors, such as sex, occupational experience, and previous leadership course experience, were associated with somewhat better performance in general for the total number of correct decisions made. This can be relevant to how receptive the participants are to the tunnel-specific e-learning content. One notable secondary finding was that an additional course in leadership was associated with poorer performance in the assessment, although this was a very limited sample and may not be a true finding. It may be relevant to try to determine precourse receptivity to specific learning goals at some depth to inform course content and logistics.

Despite not observing an intervention effect in this study, participants reported that their ability to make decisions or act as ambulance commanders in road tunnel incidents increased or remained unchanged after the trial, indicating boosted or unaffected professional confidence, respectively. This personal trait (confidence), together with work experience, has been described as important for ambulance commanders in the Swedish EMS [54]. These aspects need to be included when designing this type of highly specialized e-learning course so that both experienced and inexperienced staff may benefit.

The in-depth follow-up interviews offered some insights into the reasoning behind the main outcome. In addition to the aide-mémoire, lack of information (including risks) has been identified as an issue. This is concurrent with recent findings from interviews with other emergency services than the EMS responding to road tunnel incidents [31]. A possible support tool could be an application for joint information sharing, similar to those currently used in the United Kingdom [55]. Timely

XSL•FO
RenderX

access to adequate information may aid in the decision-making process and result in an informed (and quicker) decision to enter the road tunnel.

Ambulance commander performance has been assessed in different ways and reported [56] in web surveys [54] and interviews [42], with performance indicators for completing certain tasks, such as sending the first report to dispatch or formulating guidelines for response within a predetermined time frame [38-40]. Assessing performance indicators has been used as a method to evaluate incident management in Sweden [15]. This study presents a new method for evaluating ambulance commander performance using simulation-based assessment, which measures both the time and choice of decisions. With this method, each decision can be evaluated independently, potentially allowing the identification of particularly time-sensitive and difficult decisions. This type of detailed information can help to identify specific issues or points that might benefit from extra focus and resources.

Limitations

In the context of this newly designed and tested e-learning course and the limited achievement of the learning outcomes, several study design limitations can be discussed. Despite substantial effort in validating the course content using external (EMS personnel outside the studied region) and internal (EMS personnel with in-depth training and experience in decision-making) methods, limitations in the validity and reliability of the simulations and assessment methods may have introduced systematic errors into the results. Regarding the primary outcome, the assessment method may not have detected course learning with sufficient granularity.

From a pedagogical standpoint, different methods of how study participants were presented with learning materials were used. The control group received a short lecture (passive learning), and the intervention group was active and participated in a gamified learning environment. This difference in how learning materials were provided has not been studied, as the content of the tunnel-specific e-learning course and its influence on decision-making was the main focus. However, research findings indicate that students participating in active learning environments learn more but have a lesser feeling of learning compared to students exposed to passive learning [57]. A feeling of learning less in an educational setting could have affected the participants' decision-making abilities. In future revisions of this tunnel-specific e-learning course, the method of delivery needs to be considered. For example, using a blended learning model (including elements of both passive and active learning) may have a positive impact on learning and critical thinking [58]. In addition, more formative assessment steps in the e-learning–focused tunnel response could provide more reinforcement of primary learning goals, even though it was thought that this aspect was adequate before the study.

Using knowledge tests (MCQs) after 1 and 6 months could have influenced the result. It is known that knowledge and clinical skills decay over time [59,60]. If the first simulation had followed directly after the completion of the allotted courses, a different outcome may have been observed. This study design was chosen because there may be considerable time between an educational course and a real incident, leading to some knowledge decay. For example, major road tunnel incidents are uncommon in Sweden. During the years 2003 to 2013, a total of 926 incidents occurred in 10 of Sweden's approximately 20 road tunnels (length of >300 m). In these incidents, 525 injuries were reported. Most (n=523) were minor injuries (injury severity score of 1-8), and no deaths were reported [61]. Hence, it may be some time before participants can apply their knowledge in a real-life setting, and retaining skills for nonroutine situations is difficult [62]. Still, none (0/20, 0%) of the participants correctly entered the road tunnel, which indicates a conscious choice. Participants were not given the correct answers after the first simulation, as this may have influenced them to change their answers in the second simulation. Further development of the e-learning course is needed to facilitate the decision to correctly enter the road tunnel in case of a major incident.

The evaluation process of the tunnel-specific e-learning course and simulation could have benefited from a more granular assessment, such as using the 4-level (reaction, learning, behavior, and results) model of assessing the effectiveness of training programs proposed by Kirkpatrick [63]. In this study, level 1 was assessed during the development phase of the tunnel-specific e-learning course. Level 3 of the Kirkpatrick evaluation model has been addressed through participants' self-assessed responses on how their ability to make decisions or sense of security had changed. Level 2 of the Kirkpatrick evaluation model (learning) could have been addressed properly by adding a third question to the evaluation form, asking whether the participants' knowledge of managing road tunnel incidents decreased, remained unchanged, or increased. Some aspects concerning level 2 of the Kirkpatrick evaluation model were discussed by the participants in the follow-up interviews. However, a more structured evaluation could have been used to assess learning. This element should be incorporated in future revisions of the tunnel-specific e-learning course.

Other aspects that could potentially disturb learning goal assessment in this study include external participant-specific factors or human factors, such as whether participants were unmotivated, tired, distracted, or stressed when attending the course or the simulations. Adding images, videos, and sounds to the simulation as part of realism may have also affected the participants' focus or stress levels, for example, as information overload can degrade performance. The National Board of Medical Examiners addresses this aspect of combining media and MCQs and recommends choosing video clips with fewer distracting features [64]. Furthermore, a real road tunnel incident may include noxious sensory inputs, such as smoke and high noise levels. There must be a balance in designing training or education to incorporate enough realism without overwhelming the participant with excessive input.

Ambiguity or difficulty in interpreting the MCQs could have contributed to systematic errors in the results. More extensive work in formulating and evaluating MCQs may be an additional step in future validation processes to support both formative and normative assessment in this type of context. MCQs have the advantage of standardization and efficient data collection but, at the same time, must be carefully constructed to ensure high validity concerning specific learning goals [50]. MCQs

XSL•FO
RenderX

can be useful for simple and rapid performance assessments, although other forms of assessment can provide more granular data concerning learning phenomena [51]. However, in this serious incident context, where immediate binary decisions may be needed in the setting of complex and incomplete information, MCQ assessment was judged to be similar to the rapid decision requirements in situations faced by ambulance commanders.

Applicability

This is the first iteration of a course intended for further development based on this and other validation work for learning effectiveness in this specific, complex environment and professional response. A final e-learning course could be used to give EMS personnel a basic understanding of the tunnel environment and various risks. The final version of the e-learning course could include tunnel-specific plans and aide-mémoire and be used as a time- and cost-efficient complement to other training methods, such as tabletop or full-scale exercises. The findings in this study may be seen as a proof of concept or demonstration of feasibility for course delivery and assessment. Further evaluation and reconstruction are warranted. For example, by incorporating a simulation as the first step (and providing the user feedback on the result), the EMS personnel could be able to identify individual areas of improvement (eg, a lack of risk awareness). This feature may be useful so that personnel with different levels of experience can focus on specific sections of the course. After completing the course, EMS personnel could take a second simulation. Results from both simulations can be used to identify individual progression (or regression). Using this method, a partial individual tailoring for this kind of educational course could be constructed. Thus, the findings indicate that both the gamified e-learning content and assessments can serve as good learning alternatives. The limitations presented in this study may provide some support for other researchers considering the development of e-learning systems for content areas involving complex interventions in professional circumstances.

Conclusions

Participation in a tunnel-specific e-learning course did not provide a measurable change in ambulance commanders' decision-making behavior in simulated road tunnel incidents. The unwarranted reluctance of ambulance commanders in these simulations to enter the described twin-tube road tunnel system could lead to unnecessary delays in the treatment of time-sensitive injuries, with preventable morbidity and mortality, if this were to occur in a real-life situation. This type of assessment of commander performance could be applicable in a similar educational context.

Acknowledgments

The authors would like to express appreciation for the participants who contributed to this study. The authors would also like to thank Satish Strömberg for digitalizing the simulations and adapting the learning materials to an e-learning environment, Britt-Inger Saveman for contributing to the study design and constructive discussions on the finalization of the manuscript, and Magnus Hultin for contributing to constructive discussions on the validity and reliability of the results. This study was funded by the Swedish Transport Administration and Region Västerbotten, Sweden. The funding agencies had no role in or impact on the study design, data collection, data analysis, or writing of the manuscript.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

JH, LG, and MH contributed to the planning of the study, data collection, data analysis, and writing the manuscript. AW contributed to supervision, planning of the study, data collection, data analysis, and writing the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Written information (in Swedish) sent to potential participants before enrollment.
[PDF File (Adobe PDF File), 116 KB - formative_v9i1e58542_app1.pdf]

Multimedia Appendix 2
Content outline of the tunnel-specific e-learning course.
[DOCX File , 19 KB - formative_v9i1e58542_app2.docx]

Multimedia Appendix 3
Screenshots from the course module 3 (in Swedish).
[DOCX File , 994 KB - formative_v9i1e58542_app3.docx]

XSL•FO
RenderX

Multimedia Appendix 4
Decisions in the simulations.
[DOCX File , 14 KB - formative_v9i1e58542_app4.docx]

Multimedia Appendix 5
Semistructured interview guide (in Swedish).
[DOCX File , 18 KB - formative_v9i1e58542_app5.docx]

Multimedia Appendix 6
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 481 KB - formative_v9i1e58542_app6.pdf]

References

1. 68% of the world population projected to live in urban areas by 2050. United Nations. URL: https://www.un.org/development/desa/en/news/population/2018-revision-of-world-urbanization-prospects.html [accessed 2024-01-30]
2. Broere W. Urban underground space: solving the problems of today's cities. Tunn Undergr Space Technol 2016 May;55:245-248. [doi: 10.1016/j.tust.2015.11.012]
3. INRIX: congestion costs each American 97 hours, $1,348 a year. INRIX. URL: https://inrix.com/press-releases/scorecard-2018-us/ [accessed 2024-01-30]
4. Cui J, Nelson JD. Underground transport: an overview. Tunn Undergr Space Technol 2019 May;87:122-126. [doi: 10.1016/j.tust.2019.01.003]
5. E39 Rogfast. Statens Vegvesen. URL: https://www.vegvesen.no/vegprosjekter/europaveg/e39rogfast/ [accessed 2024-01-30]
6. Gandit M, Kouabenan DR, Caroly S. Road-tunnel fires: risk perception and management strategies among users. Saf Sci 2009 Jan;47(1):105-114. [doi: 10.1016/j.ssci.2008.01.001]
7. Lyon RM, Sanders J. The Swiss bus accident on 13 March 2012: lessons for pre-hospital care. Crit Care 2012 Jul 11;16(4):138 [FREE Full text] [doi: 10.1186/cc11370] [Medline: 22784360]
8. Li YZ. Study of fire and explosion hazards of alternative fuel vehicles in tunnels. Fire Saf J 2019 Dec;110:102871. [doi: 10.1016/j.firesaf.2019.102871]
9. Ouyang Z, Carlton A, Guo Q, Naito C, Quiel S. Blast vulnerability of drop ceilings in roadway tunnels. J Perform Constr Facil 2020 Dec;34(6):04020121. [doi: 10.1061/(asce)cf.1943-5509.0001526]
10. Report road 2015/02 report on fire in a heavy goods vehicle in the Gudvanga tunnel on the E16 road in Aurland on 5 August 2013. Accident Investigation Board Norway. 2015. URL: https://havarikommisjonen.no/Vei/Avgitte-rapporter/2015-02-eng?pid=SHT-Report-ReportFile&attach=1 [accessed 2024-04-29]
11. Kristiansen E, Johansen FH, Carlström E. When it matters most: collaboration between first responders in incidents and exercises. J Contingencies Crisis Manag 2018 Jul 04;27(1):72-78. [doi: 10.1111/1468-5973.12235]
12. Doohan IS, Saveman BI, Gyllencreutz L. Limited medical perspective at a strategic level in relation to mass casualty incidents in Swedish tunnels. Int J Emerg Manag 2019;15(4):360. [doi: 10.1504/ijem.2019.104205]
13. Eklund A, Karlsson S, Gyllencreutz L. Building "common knowledge" when responding to major road tunnel incidents: an inter-organisational focus group study. Int J Emerg Serv 2022 Oct 24;12(2):145-160. [doi: 10.1108/ijes-02-2022-0006]
14. Rüter A, Nilsson H, Vikström T. Medical Command and Control at Incidents and Disasters: From the Scene of the Incident to the Hospital Ward. Stockholm, Sweden: Studentlitteratur; 2006.
15. Nilsson H, Kristansson T. Katastrofmedicinsk beredskap - att leda sjukvård på ett särskilt sätt. Stockholm, Sweden: Studentlitteratur; 2015.
16. Bjørnsen G, Dettweiler U, Njå O, Knudsen K. Towards an understanding of learning within the Norwegian fire and rescue services – focusing on tunnel fire safety. J Workplace Learn 2022 Sep 22;35(1):112-128. [doi: 10.1108/jwl-04-2022-0051]
17. Bjørnsen G, Billett S, Njå O. First responders' perceived and actual competence in tunnel fire safety. Fire Saf J 2023 Apr;136:103758. [doi: 10.1016/j.firesaf.2023.103758]
18. Hylander J, Saveman BI, Björnstig U, Gyllencreutz L. Prehospital management provided by medical on-scene commanders in tunnel incidents in Oslo, Norway - an interview study. Scand J Trauma Resusc Emerg Med 2019 Aug 20;27(1):78 [FREE Full text] [doi: 10.1186/s13049-019-0649-8] [Medline: 31429788]
19. Hylander J, Saveman BI, Björnstig U, Gyllencreutz L. Senior ambulance officers in Swedish emergency medical services: a qualitative study of perceptions and experiences of a new management role in challenging incidents. BMJ Open 2020 Dec 07;10(12):e042072 [FREE Full text] [doi: 10.1136/bmjopen-2020-042072] [Medline: 33293325]
20. Baetzner AS, Wespi R, Hill Y, Gyllencreutz L, Sauter TC, Saveman BI, et al. Preparing medical first responders for crises: a systematic literature review of disaster training programs and their effectiveness. Scand J Trauma Resusc Emerg Med 2022 Dec 24;30(1):76 [FREE Full text] [doi: 10.1186/s13049-022-01056-8] [Medline: 36566227]
21. Schulz KF, Altman DG, Moher D, CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMJ 2010 Mar 23;340:c332 [FREE Full text] [doi: 10.1136/bmj.c332] [Medline: 20332509]

XSL•FO
RenderX

22. Eysenbach G, CONSORT-eHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res 2011 Dec;13(4):e126 [FREE Full text] [doi: 10.2196/jmir.1923] [Medline: 22209829]
23. World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA 2013 Nov 27;310(20):2191-2194. [doi: 10.1001/jama.2013.281053] [Medline: 24141714]
24. Clinical trials registration. ICMJE. URL: https://www.icmje.org/about-icmje/faqs/clinical-trials-registration/ [accessed 2024-01-30]
25. Advanced Life Support Group (ALSG), Gleeson T, Mackway-Jones K. Major Incident Medical Management and Support: The Practical Approach at the Scene, 4th edition. Hoboken, NJ: Wiley-Blackwell; 2022.
26. About moodle. MoodleDocs. URL: https://docs.moodle.org/400/en/About_Moodle [accessed 2022-11-28]
27. Sollid SJ, Rimstad R, Rehn M, Nakstad AR, Tomlinson AE, Strand T, Collaborating group. Oslo government district bombing and Utøya island shooting July 22, 2011: the immediate prehospital emergency medical service response. Scand J Trauma Resusc Emerg Med 2012 Jan 26;20(1):3 [FREE Full text] [doi: 10.1186/1757-7241-20-3] [Medline: 22280935]
28. Utvärdering av hanteringen av attentatet i Stockholm 7 April 2017: redovisning av regeringsuppdrag Ju2017/05643/SSK. The Swedish Civil Contingencies Agency. 2018. URL: https://www.msb.se/sv/publikationer/utvardering-av-hanteringen-av-attentatet-i-stockholm-7-april-2017--redovisning-av-regeringsuppdrag-ju201705643ssk/ [accessed 2024-04-29]
29. Nilsson H. PS- Prehospital Sjukvårdsledning Ett nationellt koncept. Kunskapscentrum för Katastrofmedicin (KcKm) Linköpings Universitet. URL: https://psconcept.se/wp-content/uploads/PS-koncept.pdf [accessed 2025-01-26]
30. Gamification. Oxford Advanced Learner's Dictionary. URL: https://www.oxfordlearnersdictionaries.com/definition/english/gamification [accessed 2024-01-30]
31. Hylander J, Saveman BI, Björnstig U, Gyllencreutz L, Westman A. Time-efficiency factors in road tunnel rescue as perceived by Swedish operative personnel – an interview study. Int J Emerg Serv 2022 Feb 02;11(2):312-324. [doi: 10.1108/ijes-03-2021-0011]
32. Bazyar J, Farrokhi M, Khankeh H. Triage systems in mass casualty incidents and disasters: a review study with a worldwide approach. Open Access Maced J Med Sci 2019 Feb 15;7(3):482-494 [FREE Full text] [doi: 10.3889/oamjms.2019.119] [Medline: 30834023]
33. Anseeuw K, Delvau N, Burillo-Putze G, De Iaco F, Geldner G, Holmström P, et al. Cyanide poisoning by fire smoke inhalation: a European expert consensus. Eur J Emerg Med 2013;20(1):2-9. [doi: 10.1097/mej.0b013e328357170b]
34. Mathews ZR, Koyfman A. Blast injuries. J Emerg Med 2015 Oct;49(4):573-587. [doi: 10.1016/j.jemermed.2015.03.013] [Medline: 26072319]
35. van Gaalen AE, Brouwer J, Schönrock-Adema J, Bouwkamp-Timmer T, Jaarsma AD, Georgiadis JR. Gamification of health professions education: a systematic review. Adv Health Sci Educ Theory Pract 2021 May 31;26(2):683-711 [FREE Full text] [doi: 10.1007/s10459-020-10000-3] [Medline: 33128662]
36. Mayer RE. Using multimedia for e‐learning. Comput Assist Learn 2017 Jun 13;33(5):403-423. [doi: 10.1111/jcal.12197]
37. Krathwohl DR. A revision of Bloom's taxonomy: an overview. Theory Pract 2010 Jun 24;41(4):212-218. [doi: 10.1207/s15430421tip4104_2]
38. Gryth D, Rådestad M, Nilsson H, Nerf O, Svensson L, Castrén M, et al. Evaluation of medical command and control using performance indicators in a full-scale, major aircraft accident exercise. Prehosp Disaster Med 2010 Jun 28;25(2):118-123. [doi: 10.1017/s1049023x00007834] [Medline: 20467989]
39. Rådestad M, Nilsson H, Castrén M, Svensson L, Rüter A, Gryth D. Combining performance and outcome indicators can be used in a standardized way: a pilot study of two multidisciplinary, full-scale major aircraft exercises. Scand J Trauma Resusc Emerg Med 2012 Aug 28;20(1):58 [FREE Full text] [doi: 10.1186/1757-7241-20-58] [Medline: 22929479]
40. Rådestad M, Jirwe M, Castrén M, Svensson L, Gryth D, Rüter A. Essential key indicators for disaster medical response suggested to be included in a national uniform protocol for documentation of major incidents: a Delphi study. Scand J Trauma Resusc Emerg Med 2013 Sep 11;21(1):68 [FREE Full text] [doi: 10.1186/1757-7241-21-68] [Medline: 24025366]
41. Javaeed A. Assessment of higher ordered thinking in medical education: multiple choice questions and modified essay questions. MedEdPublish (2016) 2018 Jun 12;7:128 [FREE Full text] [doi: 10.15694/mep.2018.0000128.1] [Medline: 38074575]
42. Hugelius K, Edelbring S, Blomberg K. Prehospital major incident management: how do training and real-life situations relate? A qualitative study. BMJ Open 2021 Sep 23;11(9):e048792 [FREE Full text] [doi: 10.1136/bmjopen-2021-048792] [Medline: 34556511]
43. Bassan S. Overview of traffic safety aspects and design in road tunnels. IATSS Res 2016 Jul;40(1):35-46. [doi: 10.1016/j.iatssr.2016.02.002]
44. Graneheim UH, Lundman B. Qualitative content analysis in nursing research: concepts, procedures and measures to achieve trustworthiness. Nurse Educ Today 2004 Feb;24(2):105-112. [doi: 10.1016/j.nedt.2003.10.001] [Medline: 14769454]
45. Rimstad R, Sollid SJ. A retrospective observational study of medical incident command and decision-making in the 2011 Oslo bombing. Int J Emerg Med 2015 Mar 4;8(1):4 [FREE Full text] [doi: 10.1186/s12245-015-0052-9] [Medline: 25852774]

XSL•FO
RenderX

46. Gonzalez C, Brunstein A. Training for emergencies. J Trauma 2009 Aug;67(2 Suppl):S100-S105. [doi: 10.1097/TA.0b013e3181adc0d5] [Medline: 19667839]
47. Penney G, Launder D, Cuthbertson J, Thompson MB. Threat assessment, sense making, and critical decision-making in police, military, ambulance, and fire services. Cogn Tech Work 2022 Feb 04;24(3):423-439. [doi: 10.1007/s10111-022-00694-3]
48. Jaap A, Dewar A, Duncan C, Fairhurst K, Hope D, Kluth D. Effect of remote online exam delivery on student experience and performance in applied knowledge tests. BMC Med Educ 2021 Feb 02;21(1):86 [FREE Full text] [doi: 10.1186/s12909-021-02521-1] [Medline: 33530962]
49. Bustraan J, Henny W, Kortbeek JB, Brasel KJ, Hofmann M, Schipper IB. MCQ tests in Advanced Trauma Life Support (ATLS©): development and revision. Injury 2016 Mar;47(3):665-668. [doi: 10.1016/j.injury.2015.11.024] [Medline: 26686592]
50. American Educational Research Association. Standards for Educational and Psychological Testing. Washington, DC: American Educational Research Association; 2011.
51. Ten Cate O. How can entrustable professional activities serve the quality of health care provision through licensing and certification? Can Med Educ J 2022 Aug 15;13(4):8-14 [FREE Full text] [doi: 10.36834/cmej.73974] [Medline: 36091739]
52. Vaona A, Banzi R, Kwag KH, Rigon G, Cereda D, Pecoraro V, et al. E-learning for health professionals. Cochrane Database Syst Rev 2018 Jan 21;1(1):CD011736 [FREE Full text] [doi: 10.1002/14651858.CD011736.pub2] [Medline: 29355907]
53. Skivington K, Matthews L, Simpson SA, Craig P, Baird J, Blazeby JM, et al. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. BMJ 2021 Sep 30;374:n2061 [FREE Full text] [doi: 10.1136/bmj.n2061] [Medline: 34593508]
54. Rosander M, Jonson C. Professional confidence in the roles as ambulance and medical incident commander. J Conting Crisis Manag 2017 Jan 11;25(4):289-300. [doi: 10.1111/1468-5973.12156]
55. Slater A. How ambulance commanders manage a mass casualty incident. Int J Emerg Serv 2022 Apr 27;11(2):235-247. [doi: 10.1108/IJES-03-2021-0012]
56. Rimstad R, Braut GS. Literature review on medical incident command. Prehosp Disaster Med 2015 Apr;30(2):205-215. [doi: 10.1017/S1049023X15000035] [Medline: 25659266]
57. Deslauriers L, McCarty LS, Miller K, Callaghan K, Kestin G. Measuring actual learning versus feeling of learning in response to being actively engaged in the classroom. Proc Natl Acad Sci U S A 2019 Sep 24;116(39):19251-19257 [FREE Full text] [doi: 10.1073/pnas.1821936116] [Medline: 31484770]
58. Niu Y, Xi H, Liu J, Sui X, Li F, Xu H, et al. Effects of blended learning on undergraduate nursing students' knowledge, skills, critical thinking ability and mental health: a systematic review and meta-analysis. Nurse Educ Pract 2023 Oct;72:103786. [doi: 10.1016/j.nepr.2023.103786] [Medline: 37793247]
59. Srivilaithon W, Amnuaypattanapon K, Limjindaporn C, Diskumpon N, Dasanadeba I, Daorattanachai K. Retention of basic-life-support knowledge and skills in second-year medical students. Open Access Emerg Med 2020;12:211-217 [FREE Full text] [doi: 10.2147/OAEM.S241598] [Medline: 33061682]
60. Offiah G, Ekpotu LP, Murphy S, Kane D, Gordon A, O'Sullivan M, et al. Evaluation of medical student retention of clinical skills following simulation training. BMC Med Educ 2019 Jul 16;19(1):263 [FREE Full text] [doi: 10.1186/s12909-019-1663-2] [Medline: 31311546]
61. Strömgren P, Berg S, Ekman A, Bergström K. Slutrapport - Tunnlars framkomlighet och säkerhet. Stockholm, Sweden: Trafikverket; 2014.
62. Klostermann M, Conein S, Felkl T, Kluge A. Factors influencing attenuating skill decay in high-risk industries: a scoping review. Safety 2022 Mar 28;8(2):22. [doi: 10.3390/safety8020022]
63. Kirkpatrick D. Great ideas revisited: techniques for evaluating training programs. Train Dev 1996;50(1):54-57 [FREE Full text]
64. NBME Item-writing guide: constructing written test questions for health sciences. 6th edition. National Board of Medical Examiners. URL: https://www.studocu.com/row/document/jamaa%D8%A9-alkahr%D8%A9/medical-education/nbme-item-writing-guide-2020/60196327 [accessed 2024-04-29]

Abbreviations

CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
EMS: emergency medical services
MCQ: multiple-choice question
METHANE: major incident, exact location, type of incident, hazards, access and egress, number of casualties, and emergency services
MIMMS: Major Incident Medical Management and Support
SOP: standard operating procedure

XSL•FO
RenderX

Edited by A Mavragani; submitted 18.03.24; peer-reviewed by T Macnow, Y Asada; comments to author 17.09.24; revised version received 27.01.25; accepted 02.03.25; published 28.03.25.

Please cite as:
Hylander J, Gyllencreutz L, Haney M, Westman A
Ambulance Commanders' Reluctance to Enter Road Tunnels in Simulated Incidents and the Effects of a Tunnel-Specific e-Learning Course on Decision-Making: Web-Based Randomized Controlled Trial
JMIR Form Res 2025;9:e58542
URL: https://formative.jmir.org/2025/1/e58542
doi:10.2196/58542
PMID:

©Johan Hylander, Lina Gyllencreutz, Michael Haney, Anton Westman. Originally published in JMIR Formative Research (https://formative.jmir.org), 28.03.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Videoconference-Delivered Acceptance and Commitment Therapy for Family Caregivers of People With Dementia: Pilot Randomized Controlled Trial

Areum Han[1], PhD; Robert Oster[2], PhD; Hon Yuen[1], PhD; Jeremy Jenkins[3], MS; Jessica Hawkins[1], BA; Lauren Edwards[4], BA

[1]Department of Occupational Therapy, University of Alabama at Birmingham, Birmingham, AL, United States
[2]Division of Preventive Medicine, University of Alabama at Birmingham, Birmingham, AL, United States
[3]Telehealth Private Practice, Billings, MT, United States
[4]Arts in Medicine, University of Alabama at Birmingham, Birmingham, AL, United States

Corresponding Author:
Areum Han, PhD
Department of Occupational Therapy
University of Alabama at Birmingham
SHPB 339
1720 2nd Avenue South
Birmingham, AL, 35294
United States
Phone: 1 2059752882
Email: ahan@uab.edu

Abstract

Background: Family caregivers of individuals with dementia face significant mental health challenges. Acceptance and commitment therapy (ACT) has emerged as a promising intervention for improving these caregivers' mental health. While various delivery modes of ACT have been explored, there is a need for evidence on the efficacy of videoconference-delivered ACT programs for this population.

Objective: This pilot randomized controlled trial, conducted in the United States, aims to assess the effects of a videoconference-delivered, therapist-guided ACT program on reducing depressive symptoms and improving other mental health outcomes among family caregivers with depression who give care to individuals with dementia, compared to a control group that received psychoeducation materials only.

Methods: This 2-arm, parallel-group pilot randomized controlled trial randomly assigned 33 family caregivers to either a 10-week videoconference-delivered ACT program (n=16, 48%) or a control group that received psychoeducation materials alone (n=17, 52%). Depressive symptoms (primary outcome) were measured using the Patient Health Questionnaire-9. Secondary outcomes included anxiety, stress, psychological quality of life (QoL), caregiver burden, predeath grief, guilt, and ACT process measures. Outcomes were assessed in the pretest, posttest (10-12 weeks after pretest), and a 3-month follow-up (3 months after posttest, approximately 5-6 months after pretest). An intent-to-treat approach was used for all outcome analyses. Linear mixed-effects models for repeated measures were used to analyze outcomes.

Results: The ACT group reported significantly greater improvements in stress (P=.043) and psychological QoL (P=.014) in the posttest compared to the control group. Within the ACT group, participants experienced a significant decrease in depressive symptoms, with a mean (SE) change of –6.09 (1.16) points (95% CI –8.42 to –3.76; P<.001) in the posttest and –6.71 (1.45) points (95% CI –9.63 to –3.81; P<.001) in the 3-month follow-up. These changes exceed the estimated minimal clinically important difference on the Patient Health Questionnaire-9. In addition, the ACT group reported significant improvements in anxiety, stress, psychological QoL, caregiver burden, predeath grief, guilt, values-driven action, and experiential avoidance at both posttest and 3-month follow-up. A sensitivity analysis, excluding 1 participant with near-outlier data, revealed statistically significant between-group differences in depressive symptoms at posttest (P=.037); stress at posttest (P<.001) and in 3-month follow-up (P=.001); psychological QoL at posttest (P<.001); caregiver burden at posttest (P=.003) and in 3-month follow-up (P=.003); predeath grief in 3-month follow-up (P=.031); and values-driven action at posttest (P=.032).

XSL•FO
RenderX

Conclusions: The videoconference-delivered ACT program showed promise in improving mental health outcomes and ACT processes among family caregivers with depression who give care to individuals with dementia. Future studies should aim to replicate these findings with larger, more diverse caregiver populations and explore the long-term efficacy of videoconference-delivered ACT programs.

Trial Registration: ClinicalTrials.gov NCT05043441; https://clinicaltrials.gov/study/NCT05043441

(JMIR Form Res 2025;9:e67545) doi:10.2196/67545

KEYWORDS

acceptance and commitment therapy; Alzheimer disease; caregivers; dementia; depression; web-based intervention; quality of life; randomized controlled trial; stress; videoconferencing

Introduction

In the United States, around 11 million individuals offer unpaid care to individuals with Alzheimer disease and related dementias (ADRD), with approximately two-thirds identifying as women [1]. Most of these caregivers are adult children caring for their parents, with spouses making up the next largest group [1]. These family caregivers provide critical support for individuals with ADRD, helping their relatives age in place [2,3]. Therefore, maintaining these caregivers' health is imperative; however, they face significant unmet mental health needs [4]. Recent meta-analyses estimate that approximately 34% of these family caregivers experience depressive symptoms, and 32.1% experience anxiety [5,6]. The demands of caring for individuals with ADRD often prevent caregivers from engaging in self-care and other personally valued activities and can lead to significant caregiver burden, stress, and reduced quality of life (QoL) [5,7-9]. Psychoeducation and skill-building interventions have been the predominant approaches to addressing these caregivers' needs, but meta-analyses indicate that these interventions alone do not significantly improve caregivers' mental health, particularly in alleviating depressive symptoms [10]. In contrast, psychoeducational interventions with psychotherapeutic components, as well as psychotherapy and counseling, have shown small effects in reducing depressive symptoms of these caregivers [10].

Acceptance and commitment therapy (ACT) is an evidence-based, transdiagnostic approach that combines acceptance and mindfulness with behavior change processes to support living in alignment with personal values [11,12]. ACT is grounded in the psychological flexibility model, which encompasses six core processes: (1) acceptance of unwanted thoughts and emotions; (2) cognitive defusion, or detachment from unhelpful thoughts and emotions; (3) present-moment awareness; (4) self-as-context, or seeing oneself as the observer of thoughts and feelings; (5) values clarification; and (6) establishing patterns of behavior that align with values [12]. ACT skills training can help family caregivers manage the difficult thoughts and emotions arising from their caregiving roles, such as depression, anxiety, stress, guilt, and grief while managing caregiving responsibilities alongside other important areas of life [13,14].

To our knowledge, there have been 11 intervention studies investigating ACT for family caregivers of individuals with ADRD, most of which were published within the last 4 years [14-24]. These studies were conducted in Spain [14-17], the United States [18-21], the United Kingdom [22], Germany [23], and the Netherlands [24]. They demonstrated the acceptability and benefits of ACT in improving depressive symptoms, anxiety, stress, caregiver burden, guilt, grief, and ACT processes (eg, experiential avoidance and values-driven action) [14-24]. However, most of these studies were preliminary, with half using one-group pretest-posttest designs as feasibility and pilot studies [17,18,20-22,24], and the rest using randomized controlled trials (RCTs) conducted on relatively smaller scales or resulting in relatively smaller sample sizes at follow-ups [14-16,19-23] compared to typical RCTs. For example, an RCT conducted in Spain randomized 135 caregivers into in-person ACT, in-person cognitive behavioral therapy, and a control condition, and 6-month follow-up data were analyzed for 55 participants in total (25 in the ACT group) [15]. Most importantly, delivery modes varied across studies, including in-person programs [14-17], telephone-based programs [21,23], and internet-delivered programs, such as coach-guided videoconferences conducted by our study team [18,19] and self-help web modules [20,22,24]). More research is needed to understand the effects of ACT delivered in various delivery modes and across different countries for family caregivers of individuals with ADRD. Intervention effects may vary due to cultural influences on caregiving experiences, coping styles, acceptability, and the suitability of different delivery modes based on caregivers' characteristics, including race or ethnicity, preferences, and needs [25].

Web-based delivery of ACT could be an effective alternative to in-person services, enhancing accessibility and scheduling flexibility for caregivers who cannot leave their relatives unattended [26-28]. In particular, meta-analyses of ACT across different populations suggest that therapist-guided, web-based programs have greater effects on depressive symptoms and ACT process measures than unguided programs [26,28]. In addition, the use of a self-help approach alone is limited to individuals with mild depressive symptoms, as it is not recommended for those with more severe symptoms [29]. Individually and remotely delivered ACT programs with therapist guidance can provide higher engagement for these caregivers compared to other internet delivery modes [30]. Therefore, videoconference-delivered, therapist-guided programs may be particularly beneficial for family caregivers who face significant mental health challenges, including depressive symptoms, and require one-on-one real-time interaction with a therapist but struggle to attend in-person therapy sessions [18,19].

XSL•FO
RenderX

This study aimed to assess the effects of a videoconference-delivered, therapist-guided ACT program on mental health outcomes and ACT processes in family caregivers with depression who give care to individuals with ADRD, compared to a control group that received psychoeducation materials alone conducted in the United States. The study builds on insights gained and findings from 2 smaller-scale videoconference-delivered ACT studies conducted by our team [18,19], which informed the decision to focus on caregivers with depressive symptoms and to modify the outcome measures.

Methods

Study Design and Overview

This study used a 2-arm, parallel-group RCT design. Family caregivers of individuals with ADRD were randomly assigned in a 1:1 ratio to either the videoconference-delivered ACT group or the control group that received psychoeducation materials alone. The study was conducted from late March 2022 to mid-June 2024, with data collection and entry spanning approximately 2.2 years, and a recruitment period lasting about 1.7 years.

Ethical Considerations

This study was reviewed and approved by the institutional review board at the University of Alabama, Birmingham (IRB-300008123) and registered at ClinicalTrials.gov (NCT05043441) before participant enrollment. All procedures adhered to ethical guidelines for human participant research. All participants provided informed consent prior to their involvement in the study. To protect participants' privacy and confidentiality, all study data were deidentified before analysis. Data were securely maintained and accessible only to authorized research personnel. Participants received a reloadable ClinCard as compensation for their time—up to US $150 for the intervention group for completing sessions and questionnaires, and up to US $90 for the control group for completing questionnaires only.

Eligibility Criteria

Participants were recruited based on the following inclusion criteria: (1) adults (aged 18 years or older) primarily responsible for the care of a relative with ADRD, (2) reporting at least mild depressive symptoms (Patient Health Questionnaire-9 [PHQ-9] scores≥5 [31]), (3) access to a computer or smartphone with internet connectivity, and (4) ability to provide informed consent. Exclusion criteria included cognitive, physical, or sensory deficits; being a non-English speaker; psychiatric hospitalizations or mental illness diagnoses in the past 2 years; use of antipsychotic or anticonvulsant medications; or plans to place the relative with ADRD in a nursing home within 6 months.

Target Sample Size

The target sample size was adjusted from 64 to 32 due to challenges in recruiting family caregivers with depression within the funding period. The following power calculation was performed to justify the revised sample size; this was done before study completion and before the statistical analyses were performed, and as such, is considered a priori. This calculation represents the possible between-group effect and its corresponding effect size (Cohen *d*). Assuming that we will compare repeated measurements for group means, that there are 2 groups, 3 time points (ie, that there are 3 measurements obtained for each participant), an SD for depressive symptoms (PHQ-9 score of 5.20) [32], a moderate correlation between measurements on the same participant of 0.50, and 13 participants per group (26 for the study), we will have 80% power to detect a statistically significant difference in change scores for PHQ-9 of 4.67 and greater between the 2 groups at a 0.05 significance level (2-sided). This is equivalent to being able to detect a statistically significant effect size of 0.90 and greater. This result compares favorably with a previous ACT intervention study for family caregivers with depression who give care to individuals with ADRD in Spain, which reported an effect size of 1.17 for reducing depressive symptoms in the in-person ACT group compared to a control group at posttest and included participants with posttest data in the analyses [15]. Our power calculation was performed using nQuery (version 9.4; Statistical Solutions, Ltd). Considering an estimated 16% dropout rate based on a meta-analysis of dropout rates in ACT RCTs [33], the study aimed to enroll at least 32 participants in total (16 per group). This sample size is also consistent with the recommendation of a minimum of 10 participants per group for pilot RCTs [34].

Procedures

A study flyer containing the project coordinator's contact information and a link to the eligibility screening survey created in Qualtrics was distributed to directors of Area Agencies on Aging and centers or groups supporting individuals with ADRD and their caregivers (eg, adult day care centers and support groups) across various geographic locations within the United States, including the south (eg, Alabama, Tennessee, Texas), the northeast (eg, New York, New Jersey, Massachusetts), the Midwest (eg, Michigan, Wisconsin, Indiana), and the west (eg, California, Arizona). Potential participants contacted the project coordinator for more information and completed the eligibility screening survey. Printed consent forms and pretest questionnaires were sent by mail to eligible participants who expressed interest in taking part, along with a preaddressed, prestamped envelope for their return.

A co-investigator, who had no direct contact with participants, used a computer-generated randomization sequence to create a list, which was provided to the project coordinator in sequentially numbered, opaque sealed envelopes. After the project coordinator received the signed consent forms and completed the pretest questionnaires, the envelope associated with the participant's ID number was opened and informed them of their group assignment. Participants were asked to complete follow-up questionnaires roughly 3 and 6 months after completing their pretest questionnaires.

The ACT and Control Groups

Participants assigned to the ACT group received individual ACT sessions, led by a trained ACT coach (a licensed professional counselor), via Zoom (Zoom Video Communications) videoconferencing for 1 hour per week over

XSL•FO
RenderX

10 weeks. Additionally, a 1-time 1-hour booster session was provided 1.5 months after the completion of the 10th weekly session to facilitate the maintenance of treatment effects [16]. Participants were provided with additional worksheets and web-based ACT resources for continued use until the booster session. The overview of the 10 weekly ACT sessions is presented in Multimedia Appendix 1. The ACT protocol was developed specifically for family caregivers of individuals with ADRD [18,19].

Printed copies of the ACT session worksheets were sent by mail to participants in the ACT group prior to the first session. The program aimed to actively involve participants through in-session learning activities such as worksheets, metaphors, and case scenarios, along with homework assignments. After each session, the coach shared additional web-based ACT resources, like guided video exercises, and provided a summary of the strategies discussed to support practice between sessions.

Regardless of group assignment, supplemental psychoeducation materials were sent to all participants, including video links and hard copies of resources to better meet these caregivers' needs. These materials aimed to enhance their understanding of dementia (eg, types, stages, and symptoms), care strategies for different stages of dementia, strategies for managing behavioral and psychological symptoms of dementia, and communication techniques. Participants in the control group continued with care as usual and received the same psychoeducation materials provided by the study team. After their participation in the study ended, the control group was given access to web-based ACT resources and offered a free optional 1-hour session with the ACT coach.

Feasibility Outcomes

The feasibility of recruitment, eligibility, and retention was assessed by tracking the number of potential participants who completed the eligibility screening survey, those who were excluded, those who consented, the number of dropouts, and the time required to recruit the target sample size. Participant adherence to the program was evaluated through session attendance in the ACT group. The ACT coach maintained session logs, which were reviewed by the principal investigator (the first author) to ensure the coach followed the protocol consistently across participants.

Data Collection and Entry

Overview

Data were collected using self-reported questionnaires in the pretest (baseline), posttest (approximately 10-12 weeks after the pretest), and a 3-month follow-up (approximately 3 months after the posttest, or 5-6 months after the pretest) for both groups. The pretest questionnaire included demographic and caregiving-related questions, as presented in Multimedia Appendix 2. The project coordinator periodically contacted participants regarding the completion and mailing of their questionnaires. When the project coordinator received the completed paper questionnaire, she checked for any missing responses. If any were found, the project coordinator called the participant to obtain the missing information. Two undergraduate research assistants, who were unaware of group assignments, handled data entry using Qualtrics. Data entered by each assistant was compared by a co-investigator who was not involved in data analysis to identify and resolve any potential data entry errors and to finalize the dataset before analysis.

Primary Outcome Measure

The primary outcome of this study was the PHQ-9, a 9-item self-report tool that assesses the severity of depressive symptoms over the past 2 weeks, rated on a scale from 0 to 3 [31]. Scores range from 0 to 27, with higher scores indicating more severe depressive symptoms. The PHQ-9 categorizes depressive symptom severity as follows: 0-4=no depressive symptoms, 5-9=mild, 10-14=moderate, 15-19=moderately severe, and 20-27=severe [31]. It has demonstrated good internal consistency and test-retest reliability, as well as criterion validity, construct validity, diagnostic validity for major depressive disorder, and sensitivity to change [31]. The minimal clinically important difference (MCID) is estimated to be 5 points on the PHQ-9 scale [35].

Secondary Outcome Measures

The Generalized Anxiety Disorder-7 Scale (GAD-7) is a 7-item self-report tool that assesses the severity of anxiety symptoms over the past 2 weeks, rated on a scale from 0 to 3 [36]. Scores range from 0 to 21, with higher scores indicating more severe anxiety symptoms. Cutoff points of 5, 10, and 15 denote mild, moderate, and severe anxieties, respectively [36]. The GAD-7 has shown good internal consistency and test-retest reliability, as well as criterion validity, convergent validity, construct validity, and sensitivity to change [36-39]. The MCID for the GAD-7 is estimated to be 3 or 4 points [39,40].

The Perceived Stress Scale-10 (PSS-10) is a 10-item self-report questionnaire that assesses how unpredictable, uncontrollable, and overloaded individuals perceived their lives to be over the past month, using a scale of 0 to 4 [41,42]. Scores range from 0 to 40, with higher scores indicating greater stress [42]. The scale has shown good internal consistency and test-retest reliability, as well as convergent validity, construct validity, and concurrent validity [43].

The World Health Organization Quality of Life Instrument–Psychological Health component (WHOQOL-BREF–Psych) has 6 items that assess psychological QoL over the past 2 weeks, rated on a scale of 1 to 5 [44]. Scores range from 6 to 30, with higher scores denoting better psychological QoL. The WHOQOL-BREF–Psych component has demonstrated excellent intrarater reliability, interrater reliability, internal consistency, and adequate convergent validity [44].

The Short Version of the Zarit Burden Interview (ZBI) is a 12-item self-report tool that assesses caregivers' perceived burden, rated on a scale from 0 to 4 [45]. Scores range from 0 to 48, with higher scores indicating higher levels of caregiver burden. Suggested cutoff points are as follows: 0-9=no or little burden, 10-19=mild to moderate burden, and 20-48=high burden [45]. The ZBI has good internal consistency as well as convergent and discriminant validity in caregivers of individuals with ADRD [45,46].

XSL•FO
RenderX

The Marwit-Meuser Caregiver Grief Inventory-Brief Form (MM-CGI-BF) is a 6-item self-report questionnaire that assesses predeath grief in caregivers of individuals with ADRD (ie, losses, including anticipation of future losses), rated on a scale from 1 to 5 [47]. Scores range from 6 to 30, with higher scores indicating greater grief. The MM-CGI-BF has demonstrated good internal consistency, test-retest reliability, and construct validity in caregivers of individuals with ADRD [47].

The Caregiver Guilt Questionnaire (CGQ) is a 22-item self-report measure that assesses feelings of guilt in caregivers, rated on a scale from 0 to 4 [48]. Scores range from 0 to 88, with higher scores indicating greater levels of guilt. The CGQ has good internal consistency and convergent validity in caregivers of individuals with ADRD [48].

Measures Related to ACT Processes

The Acceptance and Action Questionnaire-II is a 7-item self-report measure that assesses experiential avoidance (or efforts to control or suppress unwanted thoughts and emotions, even when doing so is ineffective or has negative consequences) and the broader measure of psychological inflexibility, rated on a scale from 1 to 7 [49]. Scores range from 7 to 49, with higher scores indicating greater experiential avoidance. The Acceptance and Action Questionnaire-II has demonstrated good internal consistency and test-retest reliability, as well as concurrent, predictive, discriminant, and incremental validity [49,50].

The Cognitive Fusion Questionnaire-7 is a 7-item self-report tool that assesses cognitive fusion (ie, being so entangled with thoughts that they dictate behavior in inflexible ways), rated on a scale from 1 to 7 [51]. Scores range from 7 to 49, with higher scores indicating greater cognitive fusion. The Cognitive Fusion Questionnaire-7 has shown good internal consistency and test-retest reliability, as well as concurrent and discriminant validity [51,52].

The Engaged Living Scale-9 is a 9-item self-report questionnaire that assesses values-driven action (ie, clarity and engagement with personal values), rated on a scale from 1 to 5 [53]. Scores range from 9 to 45, with higher scores indicating greater values-driven action. The Engaged Living Scale-9 has demonstrated good internal consistency and construct validity [53,54].

The Self-Compassion Scale-Short Form (SCS-SF) is a 12-item self-report questionnaire that assesses self-compassion, rated on a scale from 1 to 5 [55]. Scores range from 12 to 60, with higher scores indicating higher levels of self-compassion. The SCS-SF has shown good internal consistency and test-retest reliability, as well as convergent validity [56].

Data Analysis

Composite scores, with appropriate reverse coding of the responses for the outcome measures, were computed for analysis. Descriptive statistics were used to summarize participants' characteristics. Normality checks, including boxplots, stem-and-leaf plots, normal probability plots, and the Kolmogorov-Smirnov test, were performed on continuous variables. All continuous variables were found to be at least approximately normally distributed. All statistical tests were 2-sided, with a significance level of α=.05. Statistical analyses were conducted using SAS (version 9.4; SAS Institute, Inc).

To test the equivalence of groups (ACT and control groups) in participants' demographic and caregiving-related characteristics, as well as outcome measures in the pretest, the chi-square test (or Fisher exact test when the assumptions for the chi-square were not satisfied) was used for categorical variables, and the 2-sample *t* test was used for continuous variables. An intent-to-treat approach was used for all outcome analyses (ie, data from all participants were analyzed as randomized). Three participants in the control group and 2 participants in the ACT group who did not complete the posttest and 3-month follow-up evaluations were included in these analyses.

Outcomes were explored using linear mixed-effects models for repeated measures [57-60]. These models included group (1=control group and 2=ACT group), which is treated as a fixed effect, time point (1=pretest, 2=posttest, and 3=3-month follow-up), which is treated as fixed but where the slopes of the time points are treated as random, and the group by time point interaction (including components for the group × posttest interaction and the group by 3-month follow-up interaction), with a random intercept for participants. All 3 time points were included in our linear mixed effects models. These models were estimated using restricted maximum likelihood, and a first-order autoregressive covariance matrix was assumed. All linear mixed effects models were run using PROC MIXED of SAS, invoking the REPEATED statement. This statement allows us to incorporate all aspects of random variation into a covariance structure (for our data, a first-order autoregressive covariance matrix) associated with the individual participants. Overall *P* values and their associated *F* values are first reported for group, time point, and the group×time point interaction. Next, the estimated mean change from pretest to posttest within each group, the absolute difference in mean change between groups at posttest, the estimated mean change from pretest to 3-month follow-up within each group, and the absolute difference in mean change between groups in 3-month follow-up, along with corresponding SEs and 95% CIs, were reported; for a given variable, all of these results were obtained from the same linear mixed effects model.

Linear mixed effects models (including those for repeated measures) are robust against missing data, as they can accommodate unbalanced data patterns [57-60]. Therefore, all available observations were included in the analyses. For each outcome variable, participants with at least 1 nonmissing outcome measure were included in the analysis. A Cohen *d* type effect size [61] was calculated for between-group effects using the absolute difference of the means between the 2 groups divided by the common SD and for within-group effects using the estimated change in the means between the 2 time points divided by the SD of change. Due to the exploratory and preliminary nature of this pilot RCT study, no adjustment of *P* values was conducted for multiple statistical comparisons of the outcome measures [62].

XSL•FO
RenderX

Results

Participant Flow and Feasibility Outcomes

Figure 1 shows the CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) flow diagram for participant recruitment (CONSORT-EHEALTH checklist provided in Multimedia Appendix 3). A total of 69 potential participants completed the study eligibility screening survey over the 1.7-year recruitment period. Of these, 19 did not meet the eligibility criteria, mostly due to not meeting the PHQ-9 screening criteria. Seventeen eligible participants chose not to participate in the study, mostly for unknown reasons, and a few due to concerns about the required language used in the consent form.

Figure 1. Participant flow diagram of a pilot randomized controlled trial of a videoconference-delivered ACT group and a control group receiving psychoeducation materials, with pretest, posttest, and 3-month follow-up evaluations for depressed family caregivers of individuals with dementia in the United States. ACT: acceptance and commitment therapy.

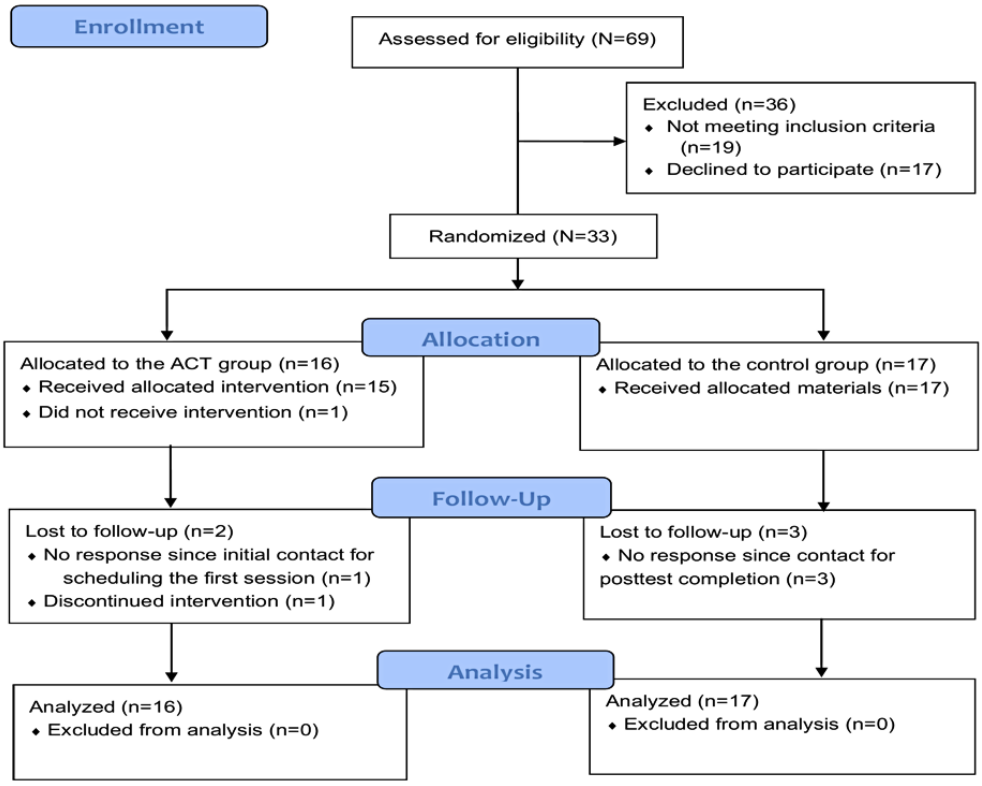

The remaining 33 participants gave informed consent, completed pretest questionnaires, and were randomly assigned to either the ACT group or the control group. All participants except 5 (3 in the control group and 2 in the ACT group) completed the posttest and 3-month follow-up questionnaires. Three participants (2 from the control group and 1 from the ACT group) were lost to follow-up after completing the pretest questionnaire and group assignment (no response). One participant in the control group requested withdrawal from the study after the group assignment, and 1 participant in the ACT group requested withdrawal after completing 6 sessions due to a lack of preference for mindfulness practices. All remaining participants in the ACT group completed all 10 weekly ACT sessions and the booster session 1.5 months after the 10th session. The coach adhered to the protocol across participants.

Baseline Characteristics of Participants

Table 1 presents the characteristics of participants (N=33) and the results of between-group differences in characteristics. Multimedia Appendix 4 presents between-group differences in outcomes at pretest. There were no statistically significant differences between the ACT and control groups in characteristics and outcome measures at pretest. The majority of participants were female (n=29, 88%), non-Hispanic White (n=24, 73%), and daughters of individuals with ADRD (n=21, 64%). Participants' ages ranged from 23 to 78 (mean 55.2, SD 12.7) years. The time since the relative's diagnosis of ADRD ranged from 0.4 to 14.2 (mean 4.8, SD 3.8) years.

XSL•FO
RenderX

Table 1. Characteristics of participants at pretest in a pilot randomized controlled trial of a videoconference-delivered ACT[a] group and a control group receiving psychoeducation materials for family caregivers with depression who give care to individuals with dementia in the United States.

	All (N=33)	ACT group (n=16)	Control group (n=17)	Difference between groups, *P* value
Age (years), mean (SD; range)	55.2 (12.7; 23-78)	54.1 (13.4; 23-78)	56.4 (12.2; 31-78)	.543
Sex, n (%)				≥.99
Female	29 (88)	14 (88)	15 (88)	
Male	4 (12)	2 (12)	2 (12)	
Education, n (%)				.61
Postgraduate degree	18 (55)	8 (50)	10 (59)	
Bachelor's degree or less	15 (45)	8 (50)	7 (41)	
Race, n (%)				≥.99
Black or Hispanic	9 (27)	4 (25)	5 (29)	
Non-Hispanic White	24 (73)	12 (75)	12 (71)	
Marital status, n (%)				.83
Married or living with a partner	20 (61)	10 (63)	10 (59)	
Never, divorced, or widowed	13 (39)	6 (37)	7 (41)	
Employment status, n (%)				.12
Retired or unemployed	17 (52)	6 (37)	11 (65)	
Employed	16 (48)	10 (63)	6 (35)	
Relationship to the relative with dementia, n (%)				.44
Non-spousal family members	25 (76)	11 (69)	14 (82)	
Spousal family members	8 (24)	5 (31)	3 (18)	
Living with the relative with dementia, n (%)				≥.99
Living together	27 (82)	13 (81)	14 (82)	
Not living together	6 (18)	3 (19)	3 (18)	
Caregiving hours per week, n (%)				.08
40 hours or more	22 (67)	13 (81)	9 (53)	
Less than 40 hours	11 (33)	3 (19)	8 (47)	
Other family members helping with care, n (%)				.85
No	18 (55)	9 (56)	9 (53)	
Yes	15 (45)	7 (44)	8 (47)	
Experience in support groups, n (%)				.88
Never attended	19 (58)	9 (56)	10 (59)	
Attending or attended before	14 (42)	7 (44)	7 (41)	
Years since the relative's diagnosis of dementia, mean (SD; range)	4.8 (3.8; 0.4-14.2)	3.9 (2.9; 0.6-9.6)	5.6 (4.4; 0.4-14.2)	.19
Type of dementia in the relative, n (%)				.56
Alzheimer disease	13 (40)	5 (31)	8 (47)	
Other dementias	11 (33)	7 (44)	4 (24)	
Not sure	9 (27)	4 (25)	5 (29)	
Stage of dementia in the relative, n (%)				.89
Early or mild	5 (15)	2 (13)	3 (18)	
Middle or moderate	16 (49)	8 (50)	8 (47)	
Late or severe	9 (27)	5 (31)	4 (23)	

XSL•FO
RenderX

	All (N=33)	ACT group (n=16)	Control group (n=17)	Difference between groups, *P* value
Not sure	3 (9)	1 (6)	2 (12)	

[a]ACT: acceptance and commitment therapy.

Effects of the Videoconference-Delivered ACT Program on Outcomes

Multimedia Appendix 5 presents the overall *P* values for group, timepoint, and group×timepoint, along with their associated *F* values. The group×time interaction term was statistically significant for only the variable WHOQOL-BREF–Psych ($F_{2,52}$=4.24; *P*=.020). All variables displayed statistically significant overall timepoint (within-group) effects (*F* values range from 3.32 to 20.73 and *P* values range from <.001 to .044), except for MM-CGI-BF ($F_{2,52}$=2.60; *P*=.084).

Multimedia Appendix 6 presents the outcome measure scores in the pretest, posttest, and 3-month follow-up for each group, change scores in the posttest and 3-month follow-up from the pretest for each group, and findings from within-group and between-group comparisons in a single table. For a given variable in Multimedia Appendices 4 and 5, all inferential statistical results were obtained from the same linear mixed effects model. Within-group and between-group differences in outcomes in posttest and 3-month follow-up are presented in Table 2 and Table 3, respectively.

Two significant between-group differences were observed in the posttest. The ACT group reported significantly greater improvements than the control group in stress and psychological QoL in the posttest. For stress, the estimated within-group change in score for the ACT group (mean –7.75, SE 1.62 points; 95% CI –10.99 to –4.51; *P*<.001) compared to the control group (mean –3.03, SE 1.61 points; 95% CI –6.25 to 0.19; *P*=.065) resulted in an absolute between-group difference of mean –4.72 (SE 2.28) points (95% CI –9.29 to –0.15; *P*=.043). For psychological QoL, the estimated within-group change in score for the ACT group (mean 4.03, SE 0.96 points; 95% CI 2.10-5.96; *P*<.001) compared to the control group (mean 0.56, SE 0.96 points; 95% CI –1.37 to 2.48; *P*=.56) resulted in an absolute between-group difference of mean 3.48 (SE 1.36) points (95% CI 0.75-6.20; *P*=.014). No significant between-group differences were observed for other outcome measures in the posttest, and none were observed for any outcome measure in the 3-month follow-up (Tables 2 and 3).

For within-group comparisons in the ACT group, participants experienced a significant decrease in depressive symptoms, with a mean change of –6.09 (SE 1.16) points (95% CI –8.42 to –3.76; *P*<.001) in the posttest and –6.71 (SE 1.45) points (95% CI –9.63 to –3.81; *P*<.001) in the 3-month follow-up. These changes exceed the estimated MCID on the PHQ-9 scale [35]. In addition, the ACT group reported significant improvements in anxiety, stress, psychological QoL, caregiver burden, predeath grief, guilt, values-driven action, and experiential avoidance in both posttest and 3-month follow-up (Tables 2 and 3). For example, the ACT group experienced a significant decrease in anxiety symptoms, with a mean change of –4.41 (SE 1.14) points (95% CI –6.71 to –2.12; *P*<.001) at posttest and –4.09 (SE 1.47) points (95% CI –7.04 to –1.14; *P*=.008) in the 3-month follow-up. These changes exceed the estimated MCID on the GAD-7 [39,40]. No significant within-group differences in change scores for cognitive fusion and self-compassion in the ACT group were observed (Tables 2 and 3).

For within-group comparisons in the control group that received psychoeducation materials alone, participants experienced a significant decrease in depressive symptoms in the posttest and 3-month follow-up (Tables 2 and 3), but these changes were smaller than the estimated MCID on the PHQ-9 scale [35]. Additionally, the control group reported significant improvements in anxiety in 3-month follow-up, self-compassion at posttest, values-driven action in the posttest and 3-month follow-up, and experiential avoidance in the posttest (Tables 2 and 3).

XSL•FO
RenderX

Table 2. Within-group and between-group differences in outcomes at posttest in a pilot randomized controlled trial of a videoconference-delivered ACT[a] group and a control group receiving psychoeducation materials for family caregivers with depression who give care to individuals with dementia in the United States.

	Estimated within-group Change Δ (posttest – pretest)[b]				Absolute between-group difference in Δ ACT versus Δ control at posttest[b]	
	ACT group		Control group		Mean (SE; 95% CI)	*P* value; ES
	Mean (SE; 95% CI)	*P* value; ES[c]	Mean (SE; 95% CI)	*P* value; ES		
PHQ-9[d] (–)[e]	–6.09 (1.16; –8.42 to –3.76)	<.001; 1.31	–3.55 (1.15); (–5.87 to –1.24)	.003; 0.75	–2.54 (1.64; –5.82 to 0.75)	.13; 0.27
GAD-7[f] (–)[e]	–4.41 (1.14; –6.71 to –2.12)	<.001; 0.96	–2.14 (1.14; –4.42 to 0.15)	.067; 0.46	–2.27 (1.61; –5.51 to 0.97)	.17; 0.25
PSS-10[g] (–)[e]	–7.75 (1.62; –10.99 to –4.51)	<.001; 1.20	–3.03 (1.61; –6.25, 0.19)	.065; 0.46	–4.72 (2.28; –9.29 to –0.15)	.043; 0.36
WHOQOL-BREF–Psych[h]	4.03 (0.96; 2.10-5.96)	<.001; 1.05	0.56 (0.96; –1.37 to 2.48)	.56; 0.14	3.48 (1.36; 0.75-6.20)	.014; 0.45
ZBI-12[i] (–)[e]	–7.11 (1.92; –10.96 to –3.27)	<.001; 0.93	–2.34 (1.91; –6.16 to 1.49)	.23; 0.30	–4.78 (2.70; –10.21 to 0.65)	.083; 0.31
MM-CGI-BF[j] (–)[e]	–2.99 (1.45; –5.91 to –0.08)	.04; 0.51	–1.53 (1.44; –4.43 to 1.36)	.29; 0.26	–1.46 (2.05; –5.57 to 2.65)	.48; 0.12
CGQ[k] (–)[e]	–8.64 (3.45; –15.55 to –1.72)	.015; 0.63	–5.89 (3.43; –12.78 to 0.99)	.092; 0.42	–2.74 (4.86; –12.50 to 7.02)	.58; 0.10
SCS-SF[l]	2.66 (2.01; –1.38 to 6.69)	.19; 0.33	5.79 (2.00; 1.79-9.80)	.005; 0.70	–3.14 (2.83; –8.82 to 2.54)	.27; 0.19
ELS-9[m]	6.24 (1.42; 3.40-9.08)	<.001; 1.10	3.53 (1.41; 0.70-6.36)	.016; 0.61	2.71 (2.00; –1.30 to 6.72)	.18; 0.24
AAQ-II[n] (–)[e]	–5.16 (1.90; –8.97 to –1.34)	.009; 0.68	–4.14 (1.90; –7.95 to –0.34)	.033; 0.53	–1.01 (2.69; –6.40 to 4.38)	.71; 0.07
CFQ-7[o] (–)[e]	–1.29 (2.07; –5.44 to 2.86)	.54; 0.16	–0.28 (2.06; –4.42 to 3.86)	.89; 0.03	–1.01 (2.92; (–6.88 to 4.85)	.73; 0.06

[a]ACT: Acceptance and commitment therapy.

[b]Established from linear mixed effects models.

[c]ES: effect size (Cohen *d*).

[d]PHQ-9: Patient Health Questionnaire-9.

[e]A minus sign in parentheses indicates that a decline in each variable means positive outcomes.

[f]GAD-7: Generalized Anxiety Disorder-7 Scale.

[g]PSS-10: Perceived Stress Scale -10.

[h]WHOQOL-BREF-Psych: World Health Organization Quality of Life Assessment-BREF-Psychological Health Component.

[i]ZBI-12: Zarit Burden Interview-12.

[j]MM-CGI-BF: Marwit–Meuser Caregiver Grief Inventory-Brief-Form.

[k]CGQ: Caregiver Guilt Questionnaire.

[l]SCS-SF: Self-Compassion Scale-Short Form.

[m]ELS-9: Engaged Living Scale-9.

[n]AAQ-II: Acceptance and Action Questionnaire-II.

[o]CFQ-7: Cognitive Fusion Questionnaire-7.

XSL•FO
RenderX

Table 3. Within-group and between-group differences in outcomes in 3-month follow-up in a pilot randomized controlled trial of a videoconference-delivered ACT[a] group and a control group receiving psychoeducation materials for family caregivers with depression who give care to individuals with dementia in the United States.

	Estimated within-group change Δ (3-month follow-up – pretest)[b]				Absolute between-group difference in Δ ACT versus Δ control in 3-month follow-up[b]	
	ACT[a] group		Control group		Mean (SE; 95% CI)	*P* value; ES
	Mean (SE; 95% CI)	*P* value; ES[c]	Mean (SE; 95% CI)	*P* value; ES		
PHQ-9[d] (-)[e]	6.71 (1.45; –9.63 to –3.81)	<.001; 1.00	–4.89 (1.44; –7.78 to –2.00)	.001; 0.82	–1.83 (2.05; –5.93 to 2.28)	.38; 0.16
GAD-7[f] (–)[e]	–4.09 (1.47; –7.04 to –1.14)	.008; 0.69	–3.23 (1.46; –6.16 to –0.30)	.032; 0.54	–0.86 (2.07; –5.03 to 3.30)	.68; 0.07
PSS-10[g] (–)[e]	–7.04 (2.02; –11.09 to –2.99)	.001; 0.87	–1.95 (2.00; –5.96 to 2.07)	.34; 0.24	–5.10 (2.84; –10.80 to 0.60)	.079; 0.31
WHOQOL-BREF–Psych[h]	2.92 (1.19; 0.52-5.31)	.018; 0.61	2.16 (1.18; –0.21 to 4.53)	.073; 0.44	0.76 (1.68; –2.61 to 4.12)	.65; 0.08
ZBI-12[i] (–)[e]	–9.55 (2.47; –14.50 to –4.60)	<.001; 0.97	–3.60 (2.45; –8.51 to 1.31)	.15; 0.36	–5.95 (3.47; –12.92 to 1.02)	.093; 0.30
MM-CGI-BF[j] (–)[e]	–4.11 (1.82; –7.76 to –0.47)	.028; 0.57	–0.51 (1.80; –4.12 to 3.10)	.78; 0.07	–3.60 (2.56; –8.73 to 1.53)	.16; 0.25
CGQ[k] (–)[e]	–11.22 (4.39; –20.03 to –2.42)	.013; 0.64	–7.34 (4.35; –16.07 to 1.39)	.098; 0.41	–3.88 (6.18; –16.28 to 8.51)	.53; 0.11
SCS-SF[l]	4.63 (2.48; –0.35 to 9.61)	.068; 0.47	3.55 (2.46; –1.38 to 8.48)	.15; 0.35	1.08 (3.49; –5.93 to 8.09)	.76; 0.05
ELS-9[m]	5.33 (1.81; 1.69-8.97)	.005; 0.73	4.96 (1.80; 1.34-8.57)	.008; 0.67	0.37 (2.56; –4.76 to 5.50)	.89; 0.03
AAQ-II[n] (–)[e]	–6.48 (2.52; –11.55 to –1.42)	.013; 0.64	–3.32 (2.51; –8.35 to 1.72)	.19; 0.32	–3.17 (3.56; –10.31 to 3.98)	.38; 0.15
CFQ-7[o] (–)[e]	–5.10 (2.74; –10.59 to 0.39)	.068; 0.47	–3.59 (2.72; –9.05 to 1.87)	.19; 0.32	–1.51 (3.86; (–9.25 to 6.24)	.70; 0.07

[a]ACT: acceptance and commitment therapy.

[a]Established from linear mixed effects models.

[c]ES: effect size (Cohen *d*).

[d]PHQ-9: Patient Health Questionnaire-9.

[e]A minus sign in parentheses indicates that a decline in each variable means positive outcomes.

[f]GAD-7: Generalized Anxiety Disorder-7 Scale.

[g]PSS-10: Perceived Stress Scale-10.

[h]WHOQOL-BREF-Psych: World Health Organization Quality of Life Assessment-BREF-Psychological Health Component.

[i]ZBI-12: Zarit Burden Interview-12.

[j]MM-CGI-BF: Marwit–Meuser Caregiver Grief Inventory-Brief-Form.

[k]CGQ: Caregiver Guilt Questionnaire.

[l]SCS-SF: Self-Compassion Scale-Short Form.

[m]ELS-9: Engaged Living Scale-9.

[n]AAQ-II: Acceptance and Action Questionnaire-II.

[o]CFQ-7: Cognitive Fusion Questionnaire-7.

Discussion

Principal Findings

This study, along with our previous work [18,19], represents the first efforts to evaluate videoconference-delivered ACT for family caregivers of individuals with ADRD. In particular, this pilot RCT targeted caregivers with at least mild depressive symptoms, differing from our earlier studies that included those with depressive, anxiety, or stress symptoms [18,19]. This decision was informed by previous findings that showed more positive outcomes when focusing specifically on caregivers with depressive symptoms [19], as well as meta-analyses indicating that web-based ACT interventions have larger effects on depressive symptoms when the population primarily has depressive symptoms [27].

XSL•FO
RenderX

Despite the small sample size, the videoconference-delivered ACT group demonstrated significantly greater improvements in stress and psychological QoL compared to the control group at posttest. The structured nature of ACT, delivered via videoconferencing with a coach, likely provided caregivers with opportunities to engage in real-time emotional processing and skill-building, enabling better responses to day-to-day stressful caregiving demands and situations [13,18]. This likely contributed to the greater improvements in stress and psychological QoL. Cohen *d* effect sizes for these between-group comparisons in stress and psychological QoL were 0.36 and 0.45, respectively, indicating small effect sizes. These findings align with meta-analyses on the effects of internet-delivered ACT compared to control conditions in RCTs across diverse populations [27], as well as the effects of ACT on family caregivers of individuals with chronic conditions [63]. Importantly, this is the first study to show significantly greater improvements in stress and psychological QoL for family caregivers of individuals with ADRD in an ACT program compared to a control group, providing valuable new evidence for this population.

However, no significant between-group differences were found for the other outcome measures in the posttest, nor for any outcome measure at the 3-month follow-up. This may be because this small-scale RCT was underpowered to detect between-group differences in most outcomes. Also, the control group received psychoeducation materials, which may have provided some benefits, particularly since more than half of the participants had not previously attended support groups. In addition, 1 participant in the control group, who found the study information through the clinical trial registration, had near-outlier data, which may have influenced some of the results. This participant's data approached outlier status (ie, outside of 3 SDs from the mean). A sensitivity analysis was conducted excluding this participant's data. The sensitivity analysis revealed statistically significant between-group differences in depressive symptoms in the posttest (absolute between-group difference=–3.25 points; P=.037); stress at posttest (absolute between-group difference=–6.45 points; P<.001) and in 3-month follow-up (absolute between-group difference=–7.65 points; P=.001); psychological QoL in the posttest (absolute between-group difference=4.35 points; P<.001); caregiver burden in the posttest (absolute between-group difference=–6.7 points; P=.003) and in the 3-month follow-up (absolute between-group difference=–8.83 points; P=.003); predeath grief in 3-month follow-up (absolute between-group difference=–5.31 points; P=.031); and values-driven action at posttest (absolute between-group difference=3.67 points; P=.032). The effect sizes for the differences between groups were larger in the posttest for all outcomes except for the SCS-SF and in the 3-month follow-up for all outcomes when excluding this participant's data than when including it. Notably, the effect sizes shifted from small to medium in stress and caregiver burden in both the posttest and 3-month follow-up, as well as in psychological QoL in the posttest. A table detailing the results from this sensitivity analysis can be found in Multimedia Appendix 7.

For within-group comparisons, the videoconference-delivered ACT group experienced significant improvements in depressive symptoms, anxiety, stress, psychological QoL, caregiver burden, predeath grief, guilt, values-driven action, and experiential avoidance in both posttest and 3-month follow-up. The mean changes in the 3-month follow-up were similar to those in the posttest, indicating that the treatment effects were maintained over time. Additionally, Cohen *d* effect sizes for these within-group comparisons indicated moderate to large effect sizes in both posttest and 3-month follow-up for these outcomes. These findings align with within-group improvements observed in prior ACT studies delivered through various modes for this population [14-23]. While the control group, which received psychoeducational materials alone, showed some within-group improvements, these changes were smaller compared to the ACT group and were not statistically significant in most cases.

Limitations

This study had several limitations. The small sample size may have limited the statistical power to detect small to moderate significant between-group differences in outcomes, especially at the 3-month follow-up. We adjusted our target sample size due to challenges in recruiting family caregivers with depression who give care to individuals with ADRD for an RCT within our funding period. Recruitment relied on sharing our study flyer with community sites, but we lacked access to the population through clinics, and some potential participants had already been recruited for previous studies, which further constrained our recruitment capacity. Another limitation of this pilot RCT is that, due to its exploratory and preliminary nature, no adjustment of P values was conducted for multiple statistical comparisons of the outcome measures. As a result, the findings should be interpreted with caution, and further studies with larger sample sizes and appropriate adjustments for multiple comparisons are needed to confirm the reliability and robustness of the results. This study did not collect data on participants' income levels or the types of private caregiving support services they were receiving at the time of recruitment. Therefore, these factors were not considered in the data analysis when comparing group differences or controlling for these factors if between-group differences were identified. Given that financial resources and access to caregiving support services may influence caregivers' mental health and engagement in interventions [64], future research should include these variables to better understand how socioeconomic status and caregiving resources impact mental health outcomes and intervention effects among family caregivers.

Conclusions

The videoconference-delivered, therapist-guided ACT program demonstrated feasibility and positive effects on mental health outcomes among family caregivers with depression who give care to individuals with ADRD. Given the barriers many caregivers face in accessing in-person mental health services, this mode of delivery represents a promising alternative for improving caregivers' mental health. Future studies are needed to replicate these findings with larger, more diverse caregiver populations and explore the long-term efficacy of videoconference-delivered ACT programs.

XSL•FO
RenderX

Acknowledgments

The authors sincerely appreciate the study participants for generously giving their time and contributing to this research. The authors appreciate the directors and staff members of the Area Agencies on Aging and adult day service programs for individuals with Alzheimer disease and related dementias for their assistance in sharing the study flyer with potential participants, and Bella Fancher and Mattie White for their help in entering data. In addition, the authors thank Donna Burnett from the Department of Occupational Therapy and Faye Suggs from the Office of Sponsored Programs at the University of Alabama at Birmingham for their tremendous administrative support in managing the research grant. This work was supported by an Alzheimer's Association Research Grant (AARG-21-846868). The sponsor had no role in the design, execution, analysis, or interpretation of the research findings.

At the time of the study, Areum Han was affiliated with the University of Alabama at Birmingham. She is now affiliated with the College of Nursing, University of South Florida, Tampa, FL, USA.

Data Availability

The findings from the data analysis are presented in the main document as Tables and Multimedia Appendices. Deidentified participant data are available from the corresponding author upon reasonable request.

Authors' Contributions

AH, RO, and HY conceptualized the study, secured the funding, and served as investigators on the project. AH supervised the overall project, RO conducted the data analysis, and HY oversaw data entry by 2 undergraduate students. JJ served as a coach delivering sessions and JH and LE served as project coordinators. AH and RO wrote the first draft of the manuscript, and all authors contributed to the final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Overview of sessions provided to the videoconference-delivered acceptance and commitment therapy group for depressed family caregivers of individuals with dementia in the United States.
[DOCX File , 13 KB - formative_v9i1e67545_app1.docx]

Multimedia Appendix 2
Participant demographic and caregiving-related questions administered to all participants at pretest in a pilot randomized controlled trial for depressed family caregivers of individuals with dementia in the United States.
[DOCX File , 13 KB - formative_v9i1e67545_app2.docx]

Multimedia Appendix 3
CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) checklist.
[PDF File (Adobe PDF File), 1180 KB - formative_v9i1e67545_app3.pdf]

Multimedia Appendix 4
Between-group differences in outcomes at pretest in a pilot randomized controlled trial of a videoconference-delivered acceptance and commitment therapy group versus a control group receiving psychoeducation materials for depressed family caregivers of individuals with dementia in the United States.
[DOCX File , 14 KB - formative_v9i1e67545_app4.docx]

Multimedia Appendix 5
Results of overall between-group and within-group comparisons in a pilot randomized controlled trial of a videoconference-delivered acceptance and commitment therapy group versus a control group receiving psychoeducation materials for depressed family caregivers of individuals with dementia in the United States.
[DOCX File , 15 KB - formative_v9i1e67545_app5.docx]

Multimedia Appendix 6

XSL•FO
RenderX

Scores in outcome measures with between-group and within-group comparisons in a pilot randomized controlled trial of a videoconference-delivered acceptance and commitment therapy group versus a control group receiving psychoeducation materials for depressed family caregivers of individuals with dementia in the United States.
[DOCX File , 18 KB - formative_v9i1e67545_app6.docx]

Multimedia Appendix 7
Results of a sensitivity analysis excluding one participant with near-outlier data in a pilot randomized controlled trial of a videoconference-delivered acceptance and commitment therapy group versus a control group receiving psychoeducation materials for depressed family caregivers of individuals with dementia in the United States.
[DOCX File , 16 KB - formative_v9i1e67545_app7.docx]

References

1. Alzheimer's Association. 2023 Alzheimer's disease facts and figures. Alzheimers Dement 2023;19(4):1598-1695. [doi: 10.1002/alz.13016] [Medline: 36918389]
2. Coehlo DP, Hooker K, Bowman S. Institutional placement of persons with dementia: What predicts occurrence and timing? J Fam Nurs 2007;13(2):253-277. [doi: 10.1177/1074840707300947] [Medline: 17452605]
3. Guterman EL, Allen IE, Josephson SA, Merrilees JJ, Dulaney S, Chiong W, et al. Association between caregiver depression and emergency department use among patients with dementia. JAMA Neurol 2019;76(10):1166-1173 [FREE Full text] [doi: 10.1001/jamaneurol.2019.1820] [Medline: 31282955]
4. Zwingmann I, Michalowsky B, Esser A, Kaczynski A, Monsees J, Keller A, et al. Identifying unmet needs of family dementia caregivers: results of the baseline assessment of a cluster-randomized controlled intervention trial. J Alzheimers Dis 2019;67(2):527-539 [FREE Full text] [doi: 10.3233/JAD-180244] [Medline: 30584136]
5. Collins RN, Kishita N. Prevalence of depression and burden among informal care-givers of people with dementia: a meta-analysis. Ageing Soc 2019;40(11):2355-2392. [doi: 10.1017/s0144686x19000527]
6. Kaddour L, Kishita N. Anxiety in informal dementia carers: a meta-analysis of prevalence. J Geriatr Psychiatry Neurol 2020;33(3):161-172. [doi: 10.1177/0891988719868313] [Medline: 31409196]
7. Gonzalo-Ciria L, Gascón-Catalán A, Laborda-Soriano A, Cambra-Aliaga A, Ruiz-Garrós MC, Perez-de-Heredia-Torres M. Difficulties fulfilling self-care needs among family caregivers: an observational study. Am J Occup Ther 2024;78(3):7803205020. [doi: 10.5014/ajot.2024.050528] [Medline: 38640086]
8. Karg N, Graessel E, Randzio O, Pendergrass A. Dementia as a predictor of care-related quality of life in informal caregivers: a cross-sectional study to investigate differences in health-related outcomes between dementia and non-dementia caregivers. BMC Geriatr 2018;18(1):189 [FREE Full text] [doi: 10.1186/s12877-018-0885-1] [Medline: 30139354]
9. Lin X, Moxley JH, Czaja SJ. Caring for dementia caregivers: psychosocial factors related to engagement in self-care activities. Behav Sci 2023;13(10):851 [FREE Full text] [doi: 10.3390/bs13100851] [Medline: 37887501]
10. Cheng ST, Li KK, Losada A, Zhang F, Au A, Thompson LW, et al. The effectiveness of nonpharmacological interventions for informal dementia caregivers: an updated systematic review and meta-analysis. Psychol Aging 2020;35(1):55-77. [doi: 10.1037/pag0000401] [Medline: 31985249]
11. Dindo L, Van Liew JR, Arch JJ. Acceptance and commitment therapy: a transdiagnostic behavioral intervention for mental health and medical conditions. Neurotherapeutics 2017;14(3):546-553 [FREE Full text] [doi: 10.1007/s13311-017-0521-3] [Medline: 28271287]
12. Hayes SC, Pistorello J, Levin ME. Acceptance and commitment therapy as a unified model of behavior change. Couns Psychol 2012;40(7):976-1002. [doi: 10.1177/0011000012460836]
13. Han A, Yuen HK, Jenkins J, Edwards L. A qualitative study exploring experiences of female family caregivers of people with dementia in acceptance and commitment therapy. Clin Psychol 2023;28(1):26-36. [doi: 10.1080/13284207.2023.2285803]
14. Romero-Moreno R, Márquez-González M, Gallego-Alberto L, Cabrera I, Vara-García C, Pedroso-Chaparro MDS, et al. Guilt focused intervention for family caregivers. Preliminary results of a randomized clinical trial. Clin Gerontol 2022;45(5):1304-1316. [doi: 10.1080/07317115.2022.2048287] [Medline: 35286236]
15. Losada A, Márquez-González M, Romero-Moreno R, Mausbach BT, López J, Fernández-Fernández V, et al. Cognitive-behavioral therapy (CBT) versus acceptance and commitment therapy (ACT) for dementia family caregivers with significant depressive symptoms: results of a randomized clinical trial. J Consult Clin Psychol 2015;83(4):760-772. [doi: 10.1037/ccp0000028] [Medline: 26075381]
16. Márquez-González M, Romero-Moreno R, Cabrera I, Olmos R, Pérez-Miguel A, Losada A. Tailored versus manualized interventions for dementia caregivers: the functional analysis-guided modular intervention. Psychol Aging 2020;35(1):41-54. [doi: 10.1037/pag0000412] [Medline: 31985248]
17. Gallego-Alberto L, Márquez-González M, Romero-Moreno R, Cabrera I, Losada A. Pilot study of a psychotherapeutic intervention for reducing guilt feelings in highly distressed dementia family caregivers (innovative practice). Dementia 2021;20(2):759-769. [doi: 10.1177/1471301219886761] [Medline: 31707841]

XSL•FO
RenderX

18. Han A, Yuen HK, Jenkins J, Yun Lee H. Acceptance and commitment therapy (ACT) guided online for distressed caregivers of persons living with dementia. Clin Gerontol 2022;45(4):927-938. [doi: 10.1080/07317115.2021.1908475] [Medline: 33794127]
19. Han A, Yuen HK, Jenkins J. The feasibility and preliminary effects of a pilot randomized controlled trial: videoconferencing acceptance and commitment therapy in distressed family caregivers of people with dementia. J Health Psychol 2023;28(6):554-567 [FREE Full text] [doi: 10.1177/13591053221141131] [Medline: 36591636]
20. Fauth EB, Novak JR, Levin ME. Outcomes from a pilot online acceptance and commitment therapy program for dementia family caregivers. Aging Ment Health 2022;26(8):1620-1629. [doi: 10.1080/13607863.2021.1942432] [Medline: 34233133]
21. Fowler NR, Judge KS, Lucas K, Gowan T, Stutz P, Shan M, et al. Feasibility and acceptability of an acceptance and commitment therapy intervention for caregivers of adults with Alzheimer's disease and related dementias. BMC Geriatr 2021;21(1):127 [FREE Full text] [doi: 10.1186/s12877-021-02078-0] [Medline: 33593296]
22. Kishita N, Gould RL, Farquhar M, Contreras M, Van Hout E, Losada A, et al. Internet-delivered guided self-help acceptance and commitment therapy for family carers of people with dementia (iACT4CARERS): a feasibility study. Aging Ment Health 2022;26(10):1933-1941 [FREE Full text] [doi: 10.1080/13607863.2021.1985966] [Medline: 34617489]
23. Risch AK, Lechner-Meichsner F, Wilz G. Telephone-based acceptance and commitment therapy for caregivers of persons with dementia: results of a randomized controlled trial. Clin Gerontol 2024:1-19 [FREE Full text] [doi: 10.1080/07317115.2024.2393307] [Medline: 39169671]
24. Atefi GL, van Knippenberg RJM, Bartels SL, Losada-Baltar A, Márquez-González M, Verhey FRJ, et al. A web-based intervention based on acceptance and commitment therapy for family caregivers of people with dementia: mixed methods feasibility study. JMIR Aging 2024;7:e53489 [FREE Full text] [doi: 10.2196/53489] [Medline: 38574360]
25. Knight BG, Sayegh P. Cultural values and caregiving: the updated sociocultural stress and coping model. J Gerontol B Psychol Sci Soc Sci 2010;65B(1):5-13. [doi: 10.1093/geronb/gbp096] [Medline: 19934166]
26. Thompson EM, Destree L, Albertella L, Fontenelle LF. Internet-based acceptance and commitment therapy: a transdiagnostic systematic review and meta-analysis for mental health outcomes. Behav Ther 2021;52(2):492-507. [doi: 10.1016/j.beth.2020.07.002] [Medline: 33622516]
27. Han A, Kim TH. Efficacy of internet-based acceptance and commitment therapy for depressive symptoms, anxiety, stress, psychological distress, and quality of life: systematic review and meta-analysis. J Med Internet Res 2022;24(12):e39727 [FREE Full text] [doi: 10.2196/39727] [Medline: 36485030]
28. Han A, Kim TH. The effects of internet-based acceptance and commitment therapy on process measures: systematic review and meta-analysis. J Med Internet Res 2022;24(8):e39182 [FREE Full text] [doi: 10.2196/39182] [Medline: 36040783]
29. Common mental health problems: identification and pathways to care. National Institute for Health and Care Excellence. URL: https://www.nice.org.uk/guidance/cg123 [accessed 2011-05-25]
30. Durepos P, MacLean R, Ricketts N, Boamah SA, Witherspoon R, Gould O, et al. Engaging care partners of persons living with dementia in acceptance and commitment therapy (ACT) programs: a scoping review. Aging Ment Health 2024;28(5):725-737. [doi: 10.1080/13607863.2023.2288864] [Medline: 38100551]
31. Kroenke K, Spitzer RL, Williams JBW. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
32. Molander P, Hesser H, Weineland S, Bergwall K, Buck S, Jäder Malmlöf J, et al. Internet-based acceptance and commitment therapy for psychological distress experienced by people with hearing problems: a pilot randomized controlled trial. Cogn Behav Ther 2018;47(2):169-184. [doi: 10.1080/16506073.2017.1365929] [Medline: 28895787]
33. Ong CW, Lee EB, Twohig MP. A meta-analysis of dropout rates in acceptance and commitment therapy. Behav Res Ther 2018;104:14-33. [doi: 10.1016/j.brat.2018.02.004] [Medline: 29477890]
34. Whitehead AL, Julious SA, Cooper CL, Campbell MJ. Estimating the sample size for a pilot randomised trial to minimise the overall trial sample size for the external pilot and main trial for a continuous outcome variable. Stat Methods Med Res 2016;25(3):1057-1073 [FREE Full text] [doi: 10.1177/0962280215588241] [Medline: 26092476]
35. Löwe B, Unützer J, Callahan CM, Perkins AJ, Kroenke K. Monitoring depression treatment outcomes with the Patient Health Questionnaire-9. Med Care 2004;42(12):1194-1201. [doi: 10.1097/00005650-200412000-00006] [Medline: 15550799]
36. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
37. Kertz S, Bigda-Peyton J, Bjorgvinsson T. Validity of the Generalized Anxiety Disorder-7 scale in an acute psychiatric sample. Clin Psychol Psychother 2013;20(5):456-464. [doi: 10.1002/cpp.1802] [Medline: 22593009]
38. Löwe B, Decker O, Müller S, Brähler E, Schellberg D, Herzog W, et al. Validation and standardization of the Generalized Anxiety Disorder Screener (GAD-7) in the general population. Med Care 2008;46(3):266-274. [doi: 10.1097/MLR.0b013e318160d093] [Medline: 18388841]
39. Toussaint A, Hüsing P, Gumz A, Wingenfeld K, Härter M, Schramm E, et al. Sensitivity to change and minimal clinically important difference of the 7-item Generalized Anxiety Disorder Questionnaire (GAD-7). J Affect Disord 2020;265:395-401. [doi: 10.1016/j.jad.2020.01.032] [Medline: 32090765]
40. Kroenke K, Baye F, Lourens SG. Comparative responsiveness and minimally important difference of common anxiety measures. Med Care 2019;57(11):890-897. [doi: 10.1097/MLR.0000000000001185] [Medline: 31415337]

XSL•FO
RenderX

41. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. J Health Soc Behav 1983;24(4):385. [doi: 10.2307/2136404]
42. Cohen S. Perceived stress in a probability sample of the United States. In: Spacapan S, Oskamp S, editors. The Social Psychology of Health: The Claremont Symposium on Applied Social Psychology. Thousand Oaks, CA: Sage Publications; 1988:31-67.
43. Ezzati A, Jiang J, Katz MJ, Sliwinski MJ, Zimmerman ME, Lipton RB. Validation of the perceived stress scale in a community sample of older adults. Int J Geriatr Psychiatry 2014;29(6):645-652 [FREE Full text] [doi: 10.1002/gps.4049] [Medline: 24302253]
44. Skevington S, Lotfy M, O'Connell K, WHOQOL Group. The World Health Organization's WHOQOL-BREF quality of life assessment: psychometric properties and results of the international field trial. A report from the WHOQOL group. Qual Life Res 2004;13(2):299-310. [doi: 10.1023/B:QURE.0000018486.91360.00] [Medline: 15085902]
45. Bédard M, Molloy DW, Squire L, Dubois S, Lever J, O'Donnell M. The Zarit Burden Interview: a new short version and screening version. Gerontologist 2001;41(5):652-657. [doi: 10.1093/geront/41.5.652] [Medline: 11574710]
46. Pinyopornpanish K, Pinyopornpanish M, Wongpakaran N, Wongpakaran T, Soontornpun A, Kuntawong P. Investigating psychometric properties of the thai version of the Zarit Burden Interview using Rasch model and confirmatory factor analysis. BMC Res Notes 2020;13(1):120 [FREE Full text] [doi: 10.1186/s13104-020-04967-w] [Medline: 32122404]
47. Liew T, Yap P. A brief, 6-item scale for caregiver grief in dementia caregiving. Gerontologist 2020;60(1):e1-e10. [doi: 10.1093/geront/gny161] [Medline: 30544229]
48. Losada A, Márquez-González M, Peñacoba C, Romero-Moreno R. Development and validation of the caregiver guilt questionnaire. Int Psychogeriatr 2010;22(4):650-660 [FREE Full text] [doi: 10.1017/S1041610210000074] [Medline: 20170587]
49. Bond FW, Hayes SC, Baer RA, Carpenter KM, Guenole N, Orcutt HK, et al. Preliminary psychometric properties of the Acceptance and Action Questionnaire-II: a revised measure of psychological inflexibility and experiential avoidance. Behav Ther 2011;42(4):676-688. [doi: 10.1016/j.beth.2011.03.007] [Medline: 22035996]
50. Fledderus M, Oude Voshaar MAH, Ten Klooster PM, Bohlmeijer ET. Further evaluation of the psychometric properties of the Acceptance and Action Questionnaire-II. Psychol Assess 2012;24(4):925-936. [doi: 10.1037/a0028200] [Medline: 22545700]
51. Gillanders DT, Bolderston H, Bond FW, Dempster M, Flaxman PE, Campbell L, et al. The development and initial validation of the Cognitive Fusion Questionnaire. Behav Ther 2014;45(1):83-101 [FREE Full text] [doi: 10.1016/j.beth.2013.09.001] [Medline: 24411117]
52. Donati MA, Berrocal C, Bernini O, Gori C, Primi C. Measuring cognitive fusion through the Cognitive Fusion Questionnaire-7: measurement invariance across non-clinical and clinical psychological samples. PLoS One 2021;16(2):e0246434 [FREE Full text] [doi: 10.1371/journal.pone.0246434] [Medline: 33534868]
53. Trindade IA, Ferreira C, Pinto-Gouveia J, Nooren L. Clarity of personal values and committed action: development of a shorter Engaged Living Scale. J Psychopathol Behav Assess 2015;38(2):258-265. [doi: 10.1007/s10862-015-9509-7]
54. Reilly ED, Ritzert TR, Scoglio AA, Mote J, Fukuda SD, Ahern ME, et al. A systematic review of values measures in acceptance and commitment therapy research. J Contextual Behav Sci 2019;12:290-304. [doi: 10.1016/j.jcbs.2018.10.004]
55. Raes F, Pommier E, Neff KD, Van Gucht D. Construction and factorial validation of a short form of the Self-Compassion Scale. Clin Psychol Psychother 2011;18(3):250-255. [doi: 10.1002/cpp.702] [Medline: 21584907]
56. Hayes JA, Lockard AJ, Janis RA, Locke BD. Construct validity of the Self-Compassion Scale-Short Form among psychotherapy clients. Couns Psychol Q 2016;29(4):405-422. [doi: 10.1080/09515070.2016.1138397]
57. Detry MA, Ma Y. Analyzing repeated measurements using mixed models. JAMA 2016;315(4):407-408. [doi: 10.1001/jama.2015.19394] [Medline: 26813213]
58. Diggle P, Heagerty P, Liang K, Zeger S. Analysis of Longitudinal Data. Oxford, UK: Oxford University Press; 2013.
59. Hayat MJ, Hedlin H. Modern statistical modeling approaches for analyzing repeated-measures data. Nurs Res 2012;61(3):188-194. [doi: 10.1097/NNR.0b013e31824f5f58] [Medline: 22551993]
60. Muhammad LN. Guidelines for repeated measures statistical analysis approaches with basic science research considerations. J Clin Invest 2023;133(11):e171058 [FREE Full text] [doi: 10.1172/JCI171058] [Medline: 37259921]
61. Cohen J. Statistical Power Analysis for the Behavioral Sciences. Hillsdale, NJ: Lawrence Erlbaum Associates; 1989.
62. Rothman KJ. No adjustments are needed for multiple comparisons. Epidemiology 1990;1(1):43-46. [Medline: 2081237]
63. Han A, Yuen HK, Jenkins J. Acceptance and commitment therapy for family caregivers: a systematic review and meta-analysis. J Health Psychol 2021;26(1):82-102. [doi: 10.1177/1359105320941217] [Medline: 32659142]
64. Scheuermann J, Pendergrass A, Diehl K, Herr RM. The roles of employment status and income in the mental health of informal caregivers in Germany. BMC Public Health 2024;24(1):2802 [FREE Full text] [doi: 10.1186/s12889-024-20252-y] [Medline: 39396955]

Abbreviations

ACT: acceptance and commitment therapy

XSL•FO
RenderX

ADRD: Alzheimer's disease and related dementias
CGQ: Caregiver Guilt Questionnaire
CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
GAD-7: Generalized Anxiety Disorder-7 Scale
MCID: minimal clinically important difference
MM-CGI-BF: Marwit–Meuser Caregiver Grief Inventory-Brief-Form
PHQ-9: Patient Health Questionnaire-9
PSS-10: Perceived Stress Scale-10
QoL: quality of life
RCT: randomized controlled trial
SCS-SF: Self-Compassion Scale-Short Form
WHOQOL-BREF-Psych: World Health Organization Quality of Life Assessment-BREF-Psychological Health Component
ZBI-12: Zarit Burden Interview-12

Edited by A Mavragani; submitted 17.10.24; peer-reviewed by ST Cheng, N Kiselev; comments to author 06.01.25; revised version received 23.01.25; accepted 05.03.25; published 31.03.25.

Please cite as:
Han A, Oster R, Yuen H, Jenkins J, Hawkins J, Edwards L
Videoconference-Delivered Acceptance and Commitment Therapy for Family Caregivers of People With Dementia: Pilot Randomized Controlled Trial
JMIR Form Res 2025;9:e67545
URL: https://formative.jmir.org/2025/1/e67545
doi:10.2196/67545
PMID:

©Areum Han, Robert Oster, Hon Yuen, Jeremy Jenkins, Jessica Hawkins, Lauren Edwards. Originally published in JMIR Formative Research (https://formative.jmir.org), 31.03.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Personalized Physician-Assisted Sleep Advice for Shift Workers: Algorithm Development and Validation Study

Yufei Shen[1,2]; Alicia Choto Olivier[1]; Han Yu[1]; Asami Ito-Masui[3]; Ryota Sakamoto[3]; Motomu Shimaoka[3]; Akane Sano[1]

[1]Rice University, Houston, TX, United States
[2]The University of Texas at Austin, Austin, TX, United States
[3]Mie University, Tsu, Japan

Corresponding Author:
Akane Sano
Rice University
6100 Main St.
Houston, TX, 77005
United States
Phone: 1 7133483821
Email: akane.sano@rice.edu

Abstract

Background: In the modern economy, shift work is prevalent in numerous occupations. However, it often disrupts workers' circadian rhythms and can result in shift work sleep disorder. Proper management of shift work sleep disorder involves comprehensive and patient-specific strategies, some of which are similar to cognitive behavioral therapy for insomnia.

Objective: Our goal was to develop and evaluate machine learning algorithms that predict physicians' sleep advice using wearable and survey data. We developed a web- and app-based system to provide individualized sleep and behavior advice based on cognitive behavioral therapy for insomnia for shift workers.

Methods: Data were collected for 5 weeks from shift workers (N=61) in the intensive care unit at 2 hospitals in Japan. The data comprised 3 modalities: Fitbit data, survey data, and sleep advice. After the first week of enrollment, physicians reviewed Fitbit and survey data to provide sleep advice and selected 1 to 5 messages from a list of 23 options. We handcrafted physiological and behavioral features from the raw data and identified clusters of participants with similar characteristics using hierarchical clustering. We explored 3 models (random forest, light gradient-boosting machine, and CatBoost) and 3 data-balancing approaches (no balancing, random oversampling, and synthetic minority oversampling technique) to predict selections for the 7 most frequent advice messages related to bedroom brightness, smartphone use, and nap and sleep duration. We tested our predictions under participant-dependent and participant-independent settings and analyzed the most important features for prediction using permutation importance and Shapley additive explanations.

Results: We found that the clusters were distinguished by work shifts and behavioral patterns. For example, one cluster had days with low sleep duration and the lowest sleep quality when there was a day shift on the day before and a midnight shift on the current day. Our advice prediction models achieved a higher area under the precision-recall curve than the baseline in all settings. The performance differences were statistically significant ($P<.001$ for 13 tests and $P=.003$ for 1 test). Sensitivity ranged from 0.50 to 1.00, and specificity varied between 0.44 and 0.93 across all advice messages and dataset split settings. Feature importance analysis of our models found several important features that matched the corresponding advice messages sent. For instance, for message 7 (*darken the bedroom when you go to bed*), the models primarily examined the average brightness of the sleep environment to make predictions.

Conclusions: Although our current system requires physician input, an accurate machine learning algorithm shows promise for automatic advice without compromising the trustworthiness of the selected recommendations. Despite its decent performance, the algorithm is currently limited to the 7 most popular messages. Further studies are needed to enable predictions for less frequent advice labels.

(JMIR Form Res 2025;9:e65000) doi:10.2196/65000

XSL•FO
RenderX

KEYWORDS

cognitive behavioral therapy; CBT; health care workers; machine learning; medical safety; web-based intervention; app-based intervention; shift work; shift work sleep disorders; shift workers; sleep disorder; wearable sensors; well-being

Introduction

Background

In the modern economy, shift work is prevalent in numerous occupations. Data from 2010 estimated that half of the workers in food preparation and serving, >45% of workers in protective service, and >35% of health care practitioners had alternative shifts apart from regular day shifts in the United States [1]. However, shift work often conflicts with the daily rhythm of sleep and wakefulness, which increases the risk of shift work sleep disorder (SWSD), a circadian rhythm disorder characterized by shift work–related sleepiness and insomnia [2]. Multiple studies have shown the substantial prevalence of SWSD among shift workers. A survey of nurses in 3 federal hospitals in Ethiopia found that 25.6% (102/399) of participants had SWSD [3]. Another study of nurses in Norway reported prevalence rates of 44.2% (580/1313) and 23.6% (146/619) for SWSD-indicative symptoms among night workers and day workers, respectively [4]. A random population sample of 1163 participants in Australia revealed that 10.1% (91/898) of day workers and 32.1% (85/265) of night workers fulfilled the study's criteria for SWSD and 1.3% (12/898) of day workers and 9.1% (24/265) of night workers were further classified under severe SWSD based on the extent of negative SWSD impacts on their life [5]. Shift workers' sleep problems are also associated with risks of mental health issues [6-8], work errors [8], burnout [6], and turnover intentions [9].

Proper management of SWSD emphasizes comprehensive and patient-specific strategies such as circadian adaptation, maintaining good sleep hygiene, strategic napping, clockwise shift rotation, and wakefulness promotions for night shifts [10]. Some of these strategies are analogous to the cognitive behavioral therapy for insomnia (CBTI), a psychological therapy recommended by the American College of Physicians as initial treatment for chronic insomnia [11], which consists of multiple approaches to improve sleep, including education about normal sleep, sleep hygiene practice, sleep restriction to consolidate high-quality sleep, and relaxation [12].

Related Work

Researchers have experimented with CBTI on shift workers with sleep problems and high risk of SWSD. A study by Järnefelt et al [13] on 19 shift workers with insomnia, more than half of whom exhibited SWSD characteristics, showed that group-based CBTI treatment improved their insomnia severity and sleep over a 6-month postintervention period, and a follow-up study further demonstrated long-lasting benefits of CBTI for up to 24 months after the intervention [14]. Järnefelt et al [15] also conducted a randomized controlled trial of group-based CBTI and self-help CBTI against a sleep hygiene education control on 83 shift workers with insomnia more recently. They found that group-based CBTI improved participants' mental health significantly, although both CBTI interventions enhanced their sleep and reduced their insomnia symptoms without significant differences from the control group, and all interventions were more beneficial for participants without SWSD characteristics [15]. Peter et al [16] used web-based CBTI for shift workers with sleep problems (n=21) and found it to be as effective as face-to-face SWSD outpatient CBTI treatment (n=12) for improving sleep efficiency and insomnia symptoms. Their research group also proposed a randomized controlled trial in 2021 to further assess their web-based CBTI intervention for shift workers against a waitlist control and face-to-face CBTI [17]. Although this existing research on CBTI for shift workers has shown promising improvement in participants' sleep, group-based and face-to-face CBTI required participants to attend several weekly sessions for up to 120 minutes led by specialists trained for >10 hours, and the long time commitment likely caused many participants to miss some sessions, which might have weakened the intervention effects [13,15,16]. Self-help, web-based CBTI was relatively more accessible to shift workers, but time commitment for weekly sessions was still necessary and likely contributed to the considerable dropout rates, and the interventions either required human input for professional feedback or only resembled a web-based learning module without any personalization [15,16]. Therefore, a system that automatically provides personalized CBTI interventions and sleep recommendations for shift workers and requires less time commitment could benefit them substantially and yield low dropout rates.

In recent years, several web-based and mobile CBTI systems have been developed to deliver automated and individualized CBTI interventions. For example, Sleep Healthy Using the Internet [18] and Sleepio [19] provide automated web-based CBTI modules equipped with interactive web design elements and user-tailored recommendations, and multiple randomized controlled trials have demonstrated their effectiveness against insomnia and related mental health problems among various demographic groups [18,20-23]. Beun et al [24,25] developed Sleepcare, a mobile app equipped with an automated digital coach interacting with users to provide CBTI interventions, adapt interventions to user characteristics, and promote adherence. Its efficacy has been demonstrated in a randomized controlled trial of 151 participants by Horsch et al [26]. CBT-i Coach is another phone app developed as a supplemental tool for face-to-face CBTI in which users can receive sleep education, customize sleep hygiene advice and relaxation activities, and practice personalized sleep restrictions [27]. In a pilot study, the app did not impair CBTI effectiveness and potentially improved patient adherence [28]. A subsequent variant of the app called Insomnia Coach was applied as a self-help CBTI intervention on 25 veterans with insomnia and showed significant treatment effects compared to the waitlist control [29].

In addition to web-based and mobile CBTI, multiple automated sleep recommender systems (RSs) have also been developed to provide personalized sleep suggestions. SleepCoacher collects

XSL•FO
RenderX

data from smartphone sensors and user self-reports to compute sleep variables and includes recommendation templates reviewed by clinicians for each variable combination [30]. To deliver a recommendation, the system chooses the template of the most correlated combination, further personalizes it according to the user's sleep status, and also evaluates the effectiveness of the recommendation. In total, 2 studies on 43 undergraduate students showed the benefits of this system on sleep, which were greater when users adhered more to the system. Daskalova et al [31] explored cohort-based sleep recommendations in which their system used demographic, exercise, and sleep data collected from 1 million users of a wearable device to assign each participant to a user group in which all members had similar profiles and chose a recommendation that showed the greatest improvement among the group in some aspects of sleep behaviors on which the participant performed poorly compared to these similar users. Compared to general sleep suggestions, cohort-based recommendations improved sleep duration more, and many participants were motivated to follow these recommendations. Pandey et al [32] used event mining to discover optimal causal relationships between a user's lifestyle and sleep and provided recommendations to match their current behaviors and environment with these relationships for better sleep. On the other hand, some systems do not directly provide advice to users but, instead, help them self-adjust their sleep habits. For example, Lullaby uses multiple sensors to track users' sleep and environment and helps them discover environmental factors with negative impacts on their sleep [33]. ShutEye serves as a peripheral display on mobile phones to remind users of recommended and inadvisable time windows for various sleep-related activities [34]. SleepBandits enables users to perform numerous self-experiments to adjust their behaviors and uses data from mobile phone sensors to evaluate experiment effects on their sleep [35].

Although existing automated and individualized CBTI and sleep RSs have demonstrated their great potential in improving sleep, they have not been modified to accommodate shift workers' irregular schedules of shift and wakefulness [13] and, thus, do not include strategies specifically beneficial to shift workers. Moreover, these CBTI interventions are mostly delivered through weekly learning modules and tasks requiring active involvement and a regular time commitment, and their personalization options are limited and mainly focus on sleep restrictions. Unlike CBTI, sleep RSs directly send behavior change recommendations to users, but their advice-tailoring processes often lack clinical support and have raised doubts about their credibility among users [31].

Objectives

We aimed to develop and evaluate a machine learning algorithm that predicts physicians' sleep advice using wearable and survey data. We also examined the model's potential in assisting physicians and even replacing them in providing sleep advice in the system.

We characterized shift workers' behaviors using clustering analyses, estimated physicians' advice selection strategies by interpreting the developed advice prediction models, and assessed the effectiveness of the provided sleep recommendations to gain more insights into the system.

Methods

Data Collection

The data were collected from shift workers, including physicians and nurses, in the intensive care unit (ICU) at Mie University Hospital (n=47) and Suzuka Chuo General Hospital (n=14) in Japan [36,37]. Eligible participants must have 8 hours of work per shift and poor sleep quality demonstrated through a Pittsburgh Sleep Quality Index score of ≥5. The detailed inclusion and exclusion criteria have been described in the work by Ito-Masui et al [36,37]. Each participant enrolled in a 5-week trial, with the first week as a baseline without an intervention (preintervention time point) and the following 4 weeks with interventions. During the 5 weeks, participants filled out daily surveys and wore a Fitbit (Google) to collect physiological and behavioral data. Physicians reviewed the data and provided sleep advice to participants as interventions 3 to 4 times a week. In total, 2123 days of multimodal data were collected from these participants for the entire study.

Ethical Considerations

This study was approved by the Clinical Research Ethics Review Committee of Mie University Hospital, Tsu, Japan (H2020-083). This study was approved by the ethics committee of Suzuka General Hospital (review number 254). All procedures were performed in accordance with the ethical standards of the institutional research committee and the 2013 Declaration of Helsinki. This study was registered in the University Hospital Medical Information Network Clinical Trials Registry on May 1, 2020 (ID: UMIN000040547). Participants provided written consent after receiving sufficient information about the study. They retained the right to opt out of the study at any time. To protect privacy, all collected data were deidentified. No compensation was offered for participation.

Data Modalities

The collected data included 3 main modalities—Fitbit data, survey data, and sleep advice. We followed the methods by Yu et al [38] to extract some features from the same Fitbit and survey data, and engineered extra features to accommodate our objectives.

Fitbit Data

Raw data from Fitbit included minute-by-minute steps and heart rate as well as the start time, end time, duration, and efficiency of each sleep period. From raw heart rate data, we computed their average, SD, and sample entropy. Sample entropy has been used for cardiovascular time series, and a low entropy value suggests greater self-similarity of the series [39]. We also computed features from step count per minute to account for the variability in participants' daily activities. Specifically, we computed and stored the duration of each continuous segment with and without steps in minutes. We then computed information entropy for the stationary and active segments. A higher entropy corresponded to more variability in participants' moving or nonmoving behaviors. Finally, we retrieved

XSL•FO
RenderX

information about the main sleep period with the longest duration from raw sleep data.

Survey Data

Participants received a morning survey and an evening survey every day with questions about their daily behaviors and well-being. Features were then extracted from their answers.

Morning Survey

Features from the morning surveys were divided into 3 categories—sleep, well-being, and miscellany.

Sleep

In the morning surveys, participants indicated whether they had slept in the previous 24 hours and provided the number of naps between 8 AM and 8 AM on the following day. They also described how they woke up (naturally, via an alarm, or via other means) and reported the time it took for them to fall asleep, the duration of their phone use before sleep, the brightness level of their sleep environment, and sleep quality. Specifically, sleep quality was quantified through the answers to the following statements using 5-point Likert scales (*strongly disagree*=1; *strongly agree*=5): (1) *I slept soundly*, (2) *I fell asleep immediately*, (3) *I was able to recover from fatigue*, (4) *I didn't wake up in the middle of sleep*, and (5) *I was satisfied with sleep*.

In addition to the aforementioned features retrieved directly from surveys, we constructed some sleep features using available survey data. In the surveys, participants reported the start time and end time of their main sleep and nap periods. According to the time entries, we constructed a binary sequence with 1-minute resolution for every day, in which a bit of 1 meant that the participants were sleeping or napping at that minute of the day. If time entries of the main sleep periods were missing from the surveys, the entries from Fitbit were used. From the binary sequences, we computed the duration of the main sleep and nap periods. We also calculated the sleep regularity index (SRI) with sliding windows of 3, 5, and 7 days. The SRI represents the probability of any 2 time points that are 24 hours apart having the same sleep or awake state averaged over a specified time window and is scaled to range from −100 to 100, where a value of 100 indicates the same sleep schedule across all days and a value of −100 means that an individual has completely flipped their sleep schedule between any 2 consecutive days [40]. Previous studies have shown that the SRI is positively correlated with the academic performance of college students [40] and a lower SRI is associated with increased stress, depression, and risk of cardiovascular diseases in older adults [41].

Well-Being

Participants reported 6 well-being metrics in the morning surveys. A total of 5 metrics—alertness, happiness, energy, health, and calmness—were reported using a continuous scale from 1 to 100, with 100 as the most positive. The sixth metric, current sleepiness, was reported via a 9-point Likert scale from *strongly awake* (1) to *strongly sleepy* (9).

Behavior

The morning surveys also recorded participants' daily behaviors, including the amount of caffeine intake and last intake time, amount of alcohol intake, and bath time. From the time entries, we computed the durations between their last caffeine intake and sleep and between bath and sleep.

Evening Survey

Features from the evening surveys were divided into 2 categories—work and well-being.

Work

In the evening surveys, participants reported their work schedules for the current day, the day before, and the day before the day before in binary sequences with 30-minute resolution, where a bit of 1 meant work during that 30-minute interval. From the binary sequences, we computed work hours for the entire day as well as for 3 different periods of the day (1: midnight-8 AM, 2: 8 AM-4 PM, and 3: 4 PM-midnight). The periods corresponded to midnight shifts, day shifts, and afternoon shifts of the participants. Moreover, by comparing the binary work sequences and binary sleep sequences constructed from the morning surveys, we calculated participants' nap durations during work and minimum durations between the last sleep periods and the start of shifts and between the end of shifts and following sleep periods. A binary indicator of whether participants did any activities other than work outside their homes was also retrieved from the surveys.

Well-Being

Similarly, as in the morning surveys, participants also answered several questions about their well-being in the evening surveys. In addition to the 6 metrics reported in the morning surveys (alertness, happiness, energy, health, calmness, and current sleepiness), they provided 3 other metrics: sleepiness during the day, stress, and tiredness, which were all reported via 5-point Likert scales (5=*strongly awake*, *stressful*, or *tired*).

Sleep Advice

Starting from the second week of enrollment, physicians reviewed Fitbit and survey data to provide sleep advice. One physician provided advice once a week to all participants, and 2 physicians gave advice 2 to 3 times a week to participants from each of the 2 hospitals. Over the 4 weeks with interventions, participants received advice approximately 13 times, and a total of 786 pieces of advice were sent to the 61 participants who completed their 5 weeks of the study. Every time physicians provided sleep advice, they chose 1 to 5 messages from 23 options, listed in Table 1. To better summarize the options, we divided them into 6 categories according to their descriptions: (1) dietary intake (messages 1, 2, and 3), (2) activity (messages 4, 5, 6, 16, 17, and 18), (3) sleep (messages 7, 8, 9, 10, 20, 21, and 22), (4) shift (message 11), (5) nap (messages 12, 13, 14, 15, and 23), and (6) mentality (messages 19). After participants received the sleep advice, they were able to respond to the advice as *eager to follow* or *difficult to follow*.

XSL•FO
RenderX

Table 1. Content of the 23 sleep advice messages that physicians sent to the participants.

Message category and ID	Message content
Dietary intake	
1	"Refrain from consuming alcohol before sleep."
2	"Stop consuming caffeine 3 hours before lights-out."
3	"Refrain from eating midnight snacks."
Activity	
4	"Refrain from using a smartphone in the bedroom."
5	"Take a bath a little earlier than usual."
6	"Refrain from exercising 3 hours before lights-out."
16	"Relax. 1) Make a habit of relaxing before sleep, including taking a bath in warm water, light stretching, using aroma, and drinking herbal tea."
17	"Relax. 2) Use the Fitbit breathing program called Relax."
18	"Do moderate exercise regularly."
Sleep	
7	"Darken the bedroom when you go to bed."
8	"When you do not fall asleep within 15 minutes, leave the bedroom and stay out of the bedroom until you get sleepy."
9	"Set your wake-up time according to each work shift."
10	"Do not try too hard to fall asleep quickly."
20	"Continue current sleep habits."
21	"Try to make enough time for sleep."
22	"Get up at the same time, whether you are working or on holiday."
Shift	
11	"Management of work shift: Consult with your superior so that shift rotation will be clockwise in general."
Nap	
12	"Take a nap. 1) If possible, take a nap for approximately 90 minutes before the night shift."
13	"Take a nap. 2) If possible, take a nap for 15–20 minutes during rest time while you are working."
14	"Take a nap. 3) If possible, take a nap earlier after a late-night shift to refresh."
15	"Create an environment for taking a nap: Ask your family for cooperation to create a quiet environment during a nap in the daytime or evening at home. Sharing your work shift schedule with your family by placing it where every member of your family can see it might be a good idea. Avoid strong lights for several hours before a nap and darken the room during a nap."
23	"Do not nap for too long."
Mentality	
19	"Be broad-minded and try to approach things positively."

Participant Characteristic Analyses

We investigated the characteristics of the participants' daily activities to find participant subgroups who had similar behaviors and understand the relationship between their behaviors and well-being. We used agglomerative hierarchical clustering to find the groupings. The algorithm initiates all samples as individual clusters and merges clusters using Ward linkage to minimize the within-cluster variances at every step [42,43]. After all the samples are merged, a dendrogram is constructed to show the merged path for each sample. By setting thresholds on the height of the dendrogram, clusters and their samples can be identified.

We used daily features from Fitbit and the surveys, except for the participants' well-being and sleep quality metrics, to find the clusters. To include as many samples as possible during clustering, we did not use the following features as they were not available for some days: SRI; nap duration during work; and the time between caffeine intake and sleep, between bath and sleep, between last sleep and the start of shift, and between the end of shift and following sleep (eg, participants did not work or take a bath). All features were standardized across all participants for clustering.

We then performed principal-component analysis on the standardized features and overlaid the cluster assignments on a 2D t-distributed stochastic neighbor embedding (t-SNE) plot

XSL•FO
RenderX

generated from the principal components that explain 95% of the total variance to observe how clusters interacted with each other. t-SNE creates a low-dimensional mapping of high-dimensional data by matching the conditional probabilities in both spaces that are proportional to the similarity of the data points [44]. To characterize participants' behaviors and well-being, we summarized the clusters' profiles by observing feature averages and examined distributions of the well-being metrics for each cluster.

Advice Prediction Models

As mentioned in the *Sleep Advice* section, physicians reviewed the Fitbit and survey data to provide sleep advice to participants. We developed machine learning models to predict their message selections and evaluated the performance of the advice prediction models.

Prediction Approach

The advice prediction task can be formulated as a binary classification, with the positive class representing the selection of an advice message. We found substantial class imbalances for the selected messages. Specifically, of 786 advice pieces, the number of selections ranged from 64 (8.1%) to 390 (49.6%) for the 7 most frequent advice messages (messages 4, 7, 12, 14, 15, 20, and 21). Consequently, we decided to build individual binary classifiers only for these messages as there would be too few positive samples to develop models for less frequent advice.

To construct samples for the machine learning models, we computed the average and SD of daily Fitbit and survey features across the previous 4 days for each date on which physicians provided advice. Samples with missing values were excluded. According to the physicians, they often considered the participants' responses to previous advice, especially when they found it difficult to choose the messages. Therefore, we incorporated advice responses into our models and enabled the models to output class probabilities to indicate certainty about a message selection. We used 12 Fitbit features (steps, sleep, and heart rate), 53 morning survey features, 50 evening survey features, and 23 advice response features. Table S1 in Multimedia Appendix 1 summarizes the features used by the classifier.

Dataset Split

To evaluate the advice prediction models, the data needed to be split into a training set and a test set. We designed 2 ways to split the data: a user-dependent split and a user-independent split. Under the user-dependent setting, the data were split chronologically, where the training set contained the first 70% of each participant's data and the testing set had the remaining 30%. Under the user-independent setting, we split the data according to the order of participant enrollment such that the models were trained on data from the first 70% of participants enrolled and tested on the 30% of participants who enrolled later in the study. We chose these train-test split methods to ensure that, when predicting sleep advice for a date of a participant, no data from later days of this participant's enrollment would be used in dependent models and only data from participants enrolled earlier would be used in independent models. In particular, the dependent split examined the effects of personalization by leveraging participants' past data to predict their future sleep advice, whereas the independent split assessed model generalizability by predicting sleep messages for unseen participants.

Classifiers

Considering the abundant features but relatively small size of our data, we focused on ensemble classifiers for their feature selection capabilities and robustness to overfitting. In particular, we chose random forest (RF), a bagging-based model, and 2 gradient-boosting models, namely, light gradient-boosting machine (LGBM) [45] and CatBoost [46], as our candidates. To find the optimal sets of model hyperparameters, we tuned the classifiers on the training set using 5-fold stratified cross-validation and 100 iterations of random search over a grid of parameters. Random search was used instead of grid search because it can span a larger parameter space and find models as good as or better than those found using grid search under the same computation budget [47]. When tuning the classifiers under the user-independent setting, we made sure that the training data were divided into subsets with nonoverlapping participants. Finally, classifiers with optimal hyperparameters were fitted on the entire training set.

Data Balancing

To address class imbalances that could hinder the development of effective advice prediction models, we experimented with 2 data-balancing methods: random oversampling and synthetic minority oversampling technique (SMOTE) [48]. Random oversampling increases the representation of the minority class by duplicating existing samples through sampling with replacement. In contrast, SMOTE enhances the minority class by generating synthetic data points through interpolation between existing samples and their nearest neighbors. Data balancing was only applied to the training data during classifier tuning and training. To better understand the effectiveness of data balancing, we also trained models on the original imbalanced data for comparison.

Prediction Evaluation

By combining 3 classifiers with 3 data-balancing approaches, we trained 9 model configurations to predict the selection of a particular advice message under user-dependent and user-independent settings. On the basis of 2 key considerations, we chose the area under the precision-recall curve (AUPRC) as the main metric to evaluate and compare the performance of the different models. First, given the relatively few occurrences of sleep advice, it was crucial for the model to provide advice when a participant needed it. The precision-recall curve is particularly well suited for this scenario because it focuses on the model's ability to correctly predict the positive class, which is actionable sleep advice. Second, due to class imbalances, the default decision threshold of 0.5 may not be optimal for converting the model's probability outputs into predicted classes. As the precision-recall curve reflects the trade-off between false alarms and missed advice predictions across all decision thresholds, the AUPRC provides a comprehensive evaluation of the model's ability to predict sleep advice without requiring a specific decision threshold.

XSL•FO
RenderX

For each model configuration, we repeated the tuning and training process 10 times with different random seeds and computed the average AUPRC on the test set. To ensure a fair comparison, the same set of seeds was used across all configurations. In addition, we performed 10 random guessing trials by sampling probability predictions from a standard uniform distribution and used a 1-sided Welch *t* test to assess whether the model performance significantly exceeded random guessing. This procedure was repeated for each of the 7 most frequent advice messages.

Feature Importance

After training the advice prediction models, we analyzed feature importance using 2 approaches to identify the key contributors to predictions on the test set. The first approach was permutation importance, which evaluates feature importance by measuring the decrease in prediction performance after randomly permuting a specific feature [49]. Greater performance degradation indicates higher feature importance. Using the AUPRC as the performance metric, we permuted each feature 10 times per classifier run and summarized the results on the test set. The second method was Shapley additive explanations (SHAP) [50,51], which leverages game theory to explain outputs of machine learning models at both local and global levels. In particular, we computed the mean absolute SHAP values across all test samples per classifier run to determine global feature importance. Both permutation and SHAP importance values were reported as averages across all repeated runs of each model configuration.

Results

Participant Characteristics

Tables 2 and 3 summarize the participants' daily features and well-being scores, respectively. Their average step count was 7545 (SD 4419), their average sleep duration was 6.77 (SD 2.87) hours, their average nap duration was 53 (SD 108.2) minutes, their average sleep regularity was 52 (SD 19.3), and their average work duration was 5.62 (SD 4.19) hours.

We used 77.01% (1635/2123) of the samples with all selected features available for hierarchical clustering. Figure 1 shows the resulting dendrogram. The default distance threshold of 52.24 (purple dashed line) identified 6 clusters color coded and labeled with numbers from 0 to 5 in the dendrogram. In addition, a few smaller clusters merged at approximately 30, so we used a lower threshold of 31 (red dashed line) to divide the 6 clusters into 13 subclusters (*a*-*m*) and further investigated the behavior variability within each larger cluster. Specifically, cluster 0 had 5 subclusters: *c*, *f*, *g*, *e*, and *m*. Cluster 1 had 2 subclusters: *a* and *b*. Cluster 2 had 2 subclusters: *k* and *d*. Clusters 3 and 4 did not have any subclusters and themselves were labeled as *j* and *h*, respectively. Cluster 5 had 2 subclusters: *l* and *i*. Figure S1 in Multimedia Appendix 1 illustrates the clusters and subclusters overlaid on t-SNE plots obtained from the same data used for clustering. Cluster and subcluster structures were clearly observable in both plots. For example, cluster 0 in the plot of 6 clusters has an upper branch and a lower branch on either side of cluster 1, which are further divided into subclusters *c* and *g*, respectively, in the plot of 13 subclusters.

Table S2 in Multimedia Appendix 1 summarizes cluster and subcluster profiles for different aspects of participants' behaviors and well-being. Major differences among the 6 clusters were observed in the well-being metrics, whereas subclusters can also be distinguished by work duration and shift types. The profiles will be discussed in detail later in the *Discussion* section.

XSL•FO
RenderX

Table 2. Participants' daily physiological and behavioral features.

Feature	Values
hrmean[a] (bpm), mean (SD)	74.87 (8.6)
hrstd[b] (bpm), mean (SD)	13.42 (3.45)
hrentropy[c], mean (SD)	0.63 (0.21)
steps, mean (SD)	7545 (4419.36)
duration_entropy_non_step, mean (SD)	2.34 (0.34)
duration_entropy_step, mean (SD)	1.69 (0.32)
nap_count, mean (SD)	0.56 (0.69)
brightness_sleep (scale of 1-100), mean (SD)	18.63 (23.09)
sleep_duration (min), mean (SD)	406.46 (172.01)
nap_duration (min), mean (SD)	52.99 (108.23)
sleep_regularity_3 days (scale of –100 to 100), mean (SD)	52.25 (28.3)
sleep_regularity_5 days (scale of –100 to 100), mean (SD)	52.1 (22.6)
sleep_regularity_7 days (scale of –100 to 100), mean (SD)	52.02 (19.26)
deep_sleep (scale of 1-5), mean (SD)	3.2 (1.14)
immediate_sleep (scale of 1-5), mean (SD)	3.55 (1.14)
fatigue_recover (scale of 1-5), mean (SD)	2.84 (1.05)
mid_awake (scale of 1-5), mean (SD)	3.2 (1.31)
sleep_satisfy (scale of 1-5), mean (SD)	2.85 (1.02)
alcohol_amount (cups), mean (SD)	0.3 (0.63)
caffeine_amount (cups), mean (SD)	1.53 (1.46)
worktime_today_duration (h), mean (SD)	5.62 (4.19)
worktime_today_duration (1: midnight-8 AM; h), mean (SD)	1.52 (2.85)
worktime_today_duration (2: 8 AM-4 PM; h), mean (SD)	2.15 (3.25)
worktime_today_duration (3: 4 PM-midnight; h), mean (SD)	1.96 (3.08)
extrawork_activities (n=1865), n (%)	924 (49.54)
sleep_prev_24[d] (n=2092), n (%)	2088 (99.81)
wake_natural (n=1890), n (%)	793 (41.96)
wake_alarm (n=1890), n (%)	882 (46.67)
wake_other (n=1890), n (%)	198 (10.48)

[a]hrmean: average daily heart rate.
[b]hrstd: SD of daily heart rate.
[c]hrentropy: entropy of daily heart rate.
[d]sleep_prev_24: binary indicator of sleep in the previous 24 hours.

XSL•FO
RenderX

Table 3. Mean and SD of participants' daily well-being scores. Morn indicates morning and eve indicates evening.

Well-being metric	Values, mean (SD)
sleepiness_morn (scale of 1-9)	5.49 (1.99)
sleepiness_now_eve (scale of 1-9)	5.7 (1.78)
sleepiness_daytime (scale of 1-5)	2.59 (1.23)
stress (scale of 1-5)	2.93 (1.1)
tiredness (scale of 1-5)	2.51 (1.01)
alertness_morn (scale of 1-100)	41 (23.02)
happiness_morn (scale of 1-100)	52.99 (17.34)
energy_morn (scale of 1-100)	44.02 (21.27)
health_morn (scale of 1-100)	51.71 (18.12)
calmness_morn (scale of 1-100)	55.97 (18.1)
alertness_eve (scale of 1-100)	41.08 (21.1)
happiness_eve (scale of 1-100)	54.27 (18.18)
energy_eve (scale of 1-100)	45.88 (21.08)
health_eve (scale of 1-100)	52.4 (18.19)
calmness_eve (scale of 1-100)	56.85 (19)

Figure 1. Dendrogram from hierarchical clustering of daily features using Ward linkage. The purple dashed line corresponds to a distance threshold of 52.24 (0.7 times the maximum cluster distance as implemented by SciPy [52]) and identifies 6 clusters labeled and color coded (cluster 0=blue, cluster 1=orange, cluster 2=green, cluster 3=red, cluster 4=purple, and cluster 5=brown). The red dashed line represents a distance threshold of 31 and identifies 13 subclusters (a-m).

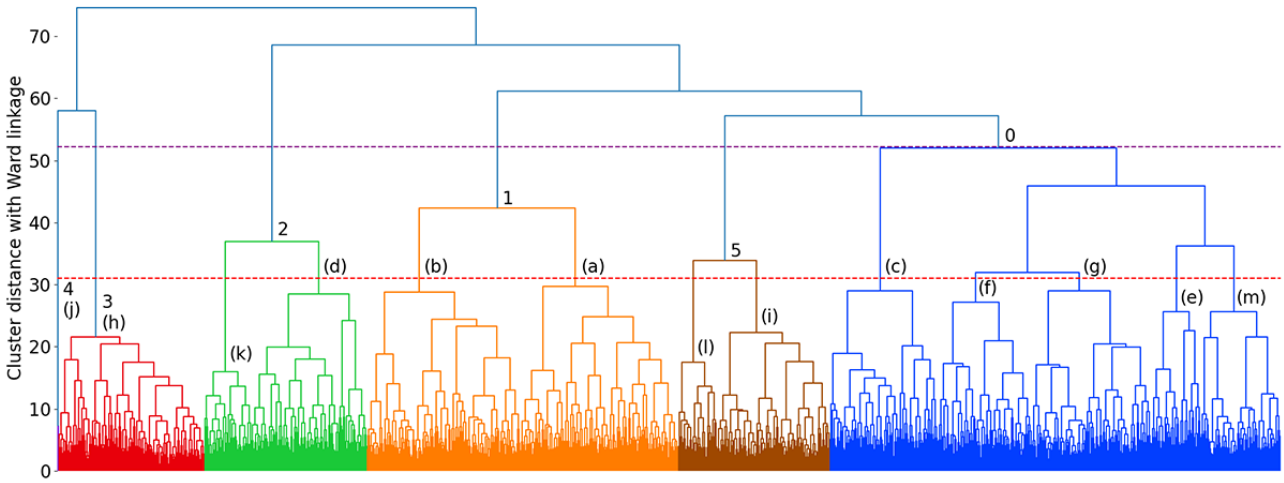

Advice Statistics

Figure 2 shows the frequency of advice provided as well as the message categories in a pie chart. As physicians often selected multiple messages for each day, 1332 advice messages were selected in 786 advice pieces. The outer ring of the pie chart indicates the proportions of each category with respect to the 1332 total occurrences of all message options, and the numbers in parentheses after each category name correspond to the frequency of the category. The inner ring shows the proportion of each message with respect to the total occurrences. The legend lists the corresponding color of each message, and each parenthesis includes the proportion of all advice pieces containing that message, which is computed by dividing the frequency of the message by 786 total advice counts. Messages in the sleep and nap categories were selected more frequently than messages from other categories. Messages in the sleep category were selected the greatest number of times (817/1332, 61.34%), and messages in the nap category were selected 21.7% (289/1332) of the time. Moreover, only a few messages were frequently selected by physicians. For example, message 20 (*continue current sleep habits*) was selected in 49.6% (390/786) of advice pieces, and message 21 (*try to make enough time for sleep*) was selected in 28.1% (221/786) of advice pieces. On the other hand, 70% (16/23) of the messages were selected in <50 advice pieces (50/786, 6.4%), and 7 messages were selected <10 times (10/786, 1.3%).

Figure 3 summarizes the frequency and proportion of responses to advice pieces containing each message. Sometimes,

XSL•FO
RenderX

participants did not respond to the advice. Although multiple messages might be selected for 1 advice piece, participants responded to the advice but not to individual messages. Participants showed different preferences for messages. For example, participants expressed difficulty following message 20 (*continue current sleep habits*) only 0.5% (2/390) of the time, whereas they experienced difficulty following message 4 (*refrain from using a smartphone in the bedroom*) 20% (15/76) of the time and following message 2 (*stop consuming caffeine 3 hours before lights-out*) 26% (10/38) of the time. The discrepancies might be explained by the message design. Message 20 suggests that participants continue their current sleep habits, which participants might feel is easier to follow as it would not change their current lifestyle greatly. On the other hand, message 2 and message 4 address caffeine consumption and phone use, which might be challenging to follow as it is suggested that participants change some essential aspects of their routine.

Table S3 in Multimedia Appendix 1 summarizes the distributions of positive and negative advice labels for both the training and test data under user-dependent and user-independent settings. In addition, it includes the ratios of negative-to-positive sample counts. Class imbalances were evident for all messages except message 20 (*continue current sleep habits*). In both data split settings, messages 21 (*try to make enough time for sleep*) and 7 (*darken the bedroom when you go to bed*) showed moderate imbalances, with negative-to-positive ratios of <5. For the remaining 4 messages, class imbalances were more pronounced, with negative-to-positive ratios mostly ranging from 5 to 10. Furthermore, we observed substantial discrepancies in class imbalance between the user-independent training and test sets of messages 12 (*take a nap for approximately 90 minutes before the night shift*), 14 (*take a nap earlier after a late-night shift to refresh*), and 15 (*create an environment for taking a nap*). While negative-to-positive ratios remained <10 in the training sets, the ratios exceeded 15 in the test sets, indicating distinct preferences among physicians when providing sleep advice to different shift workers.

Figure 2. The pie chart shows the selection frequency of each sleep advice message. All 23 message options are divided into 6 categories: dietary intake (messages 1, 2, and 3), activity (messages 4, 5, 6, 16, 17, and 18), sleep (messages 7, 8, 9, 10, 20, 21, and 22), shift (message 11), nap (messages 12, 13, 14, 15, and 23), and mentality (message 19). Message 11 is excluded from the chart because of zero frequency.

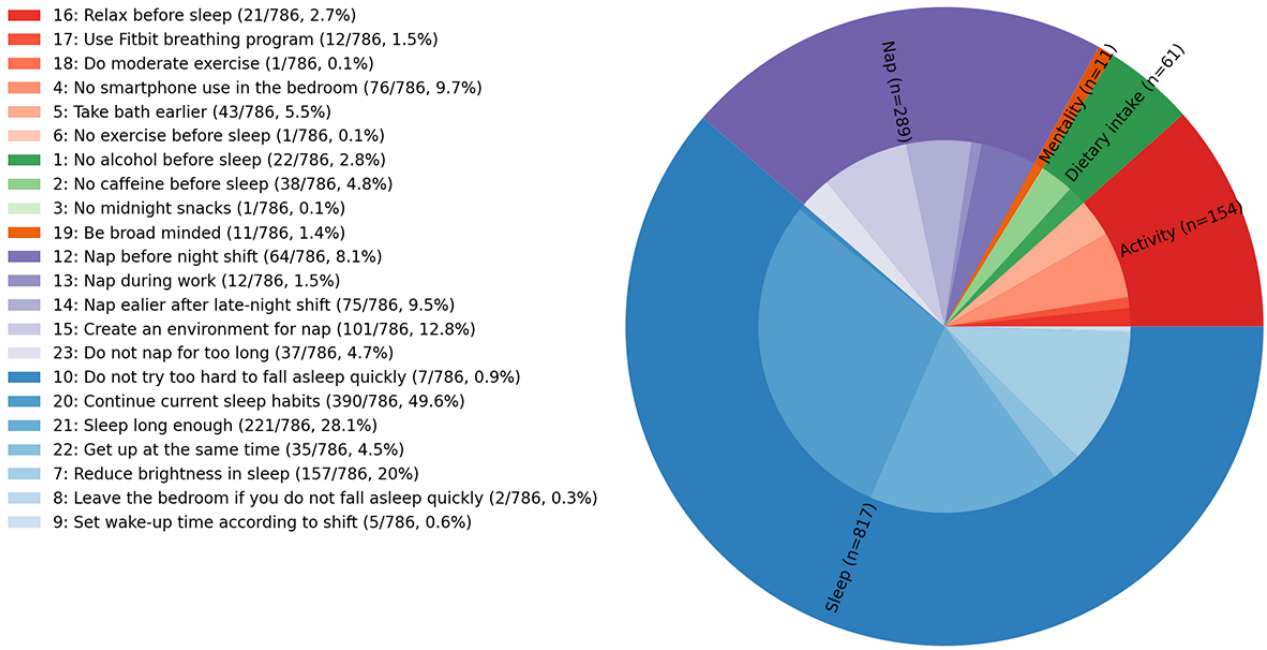

XSL•FO
RenderX

Figure 3. Frequency and proportion of message responses in a stacked bar plot. Each bar includes the frequency and proportion of participants' responses (no response, eager to follow, and difficult to follow) to advice pieces containing each message. Although multiple messages might be selected for 1 advice piece, participants responded to the advice but not to individual messages.

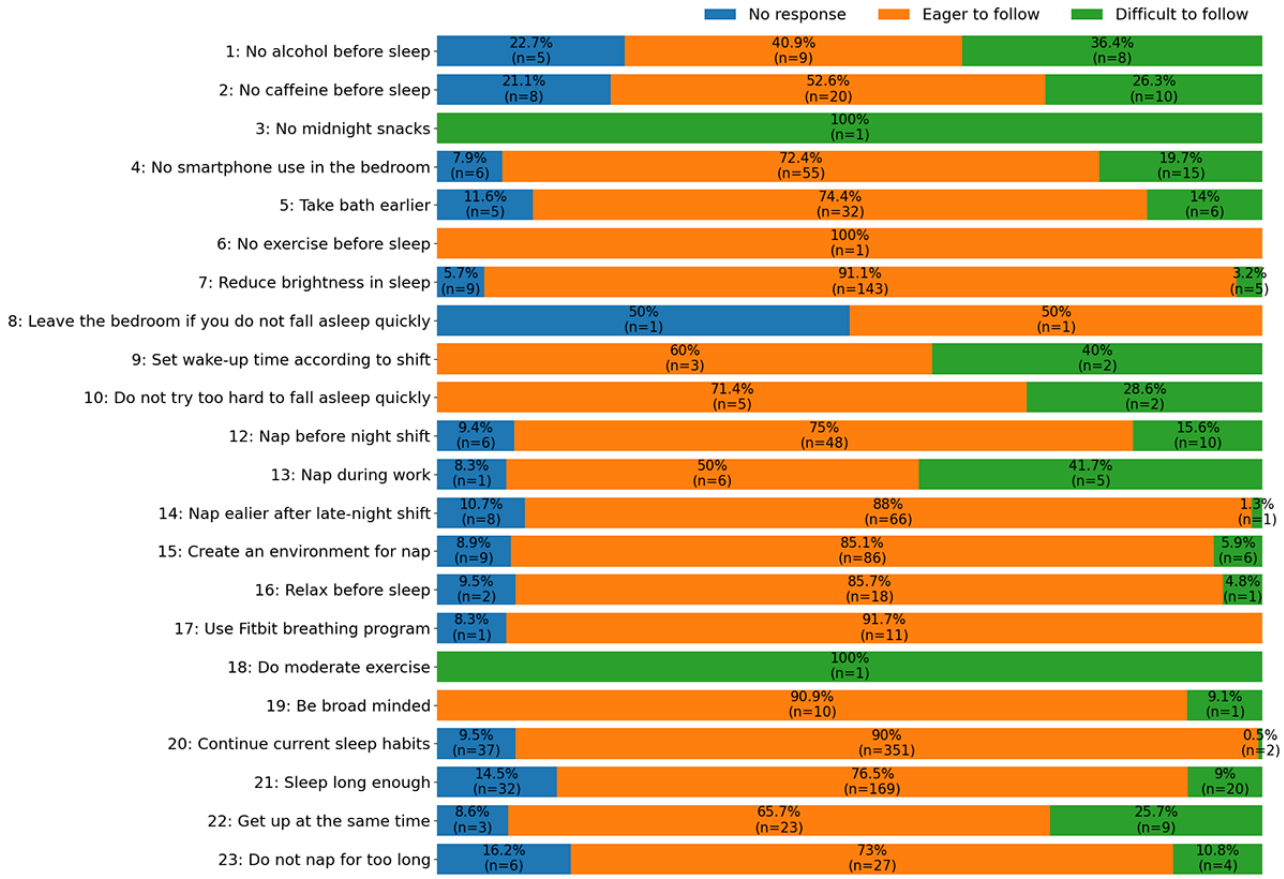

Advice Prediction Results

Following the approach described in the *Methods* section, we obtained the AUPRC of 9 different configurations of classifiers and data-balancing methods on the test set under dependent and independent settings. Table 4 summarizes the performance of the proposed models, where each model configuration was tuned and trained 10 times with different random seeds, as well as results from 10 random-guessing trials. The model performance was compared with random guessing using a 1-sided Welch *t* test.

As shown in Table 4, under the user-dependent setting, RF performed the best for predicting 6 advice messages. The only exception was advice message 7 (*darken the bedroom when you go to bed*), where CatBoost achieved the highest AUPRC. However, RF's AUPRC was only 0.02 lower than that of CatBoost. Under the user-independent setting, RF remained competitive, achieving the best performance for predicting 3 advice messages. Its performance on other messages was also close to that of the top performers. LGBM yielded the highest prediction AUPRC for messages 12 (*take a nap for approximately 90 minutes before the night shift*) and 14 (*take a nap earlier after a late-night shift to refresh*). CatBoost was the top performer for messages 7 (*darken the bedroom when you go to bed*) and 21 (*try to make enough time for sleep*). These results highlight the effectiveness of RF in mitigating overfitting when working with limited data. While models trained on balanced data achieved the best performance on 4 advice messages in both dependent and independent settings, data-balancing techniques overall did not result in a substantial improvement in AUPRC and even degraded the performance for some advice labels. Statistical tests revealed that the best prediction models for each advice message and data split setting performed significantly better than random guessing, with $P<.001$ for 13 models and $P=.003$ for the remaining one.

XSL•FO
RenderX

Table 4. Advice prediction area under the precision-recall curve (AUPRC) of the proposed models (random forest [RF], light gradient-boosting machine [LGBM], and CatBoost) and random guessing baseline obtained under different data-balancing approaches (none, random oversampling, and synthetic minority oversampling technique [SMOTE]) and dataset split settings (user dependent and user independent). A 1-sided Welch *t* test was used to compare the models to random guessing.

Setting, message ID, and balancing method	RF (AUPRC)		LGBM (AUPRC)		CatBoost (AUPRC)		Random guessing (AUPRC), mean (SD)
	Values, mean (SD)	*P* value	Values, mean (SD)	*P* value	Values, mean (SD)	*P* value	
Dependent							
4							0.12 (0.02)
None	*0.33 (0.07)* [a]	*<.001*	0.28 (0.06)	<.001	0.24 (0.04)	<.001	
Random oversampling	0.25 (0.05)	<.001	0.21 (0.04)	<.001	0.20 (0.03)	<.001	
SMOTE	0.24 (0.03)	<.001	0.18 (0.02)	<.001	0.18 (0.02)	<.001	
7							0.17 (0.02)
None	0.56 (0.04)	<.001	0.54 (0.01)	<.001	0.57 (0.02)	<.001	
Random oversampling	0.54 (0.03)	<.001	0.52 (0.03)	<.001	0.54 (0.03)	<.001	
SMOTE	0.52 (0.03)	<.001	0.56 (0.02)	<.001	*0.58 (0.02)*	*<.001*	
12							0.12 (0.02)
None	*0.32 (0.03)*	*<.001*	0.23 (0.03)	<.001	0.27 (0.06)	<.001	
Random oversampling	0.31 (0.04)	<.001	0.21 (0.04)	<.001	0.26 (0.07)	<.001	
SMOTE	0.19 (0.03)	<.001	0.17 (0.01)	<.001	0.22 (0.05)	<.001	
14							0.10 (0.02)
None	0.26 (0.05)	<.001	0.24 (0.04)	<.001	0.23 (0.05)	<.001	
Random oversampling	0.25 (0.04)	<.001	0.22 (0.04)	<.001	0.23 (0.05)	<.001	
SMOTE	*0.29 (0.05)*	*<.001*	0.22 (0.05)	<.001	0.22 (0.04)	<.001	
15							0.19 (0.03)
None	*0.31 (0.03)*	*<.001*	0.26 (0.05)	.001	0.25 (0.05)	.007	
Random oversampling	0.28 (0.02)	<.001	0.27 (0.05)	<.001	0.22 (0.04)	.06	
SMOTE	0.25 (0.01)	<.001	0.24 (0.02)	<.001	0.24 (0.02)	.002	
20							0.57 (0.02)
None	0.80 (0.02)	<.001	0.76 (0.02)	<.001	0.79 (0.02)	<.001	
Random oversampling	0.80 (0.01)	<.001	0.77 (0.02)	<.001	0.80 (0.02)	<.001	
SMOTE	*0.81 (0.02)*	*<.001*	0.77 (0.02)	<.001	0.79 (0.01)	<.001	
21							0.23 (0.02)
None	0.38 (0.03)	<.001	0.38 (0.03)	<.001	0.41 (0.01)	<.001	
Random oversampling	0.38 (0.03)	<.001	0.36 (0.02)	<.001	0.40 (0.03)	<.001	
SMOTE	*0.44 (0.04)*	*<.001*	0.40 (0.04)	<.001	0.44 (0.06)	<.001	
Independent							
4							0.11 (0.05)
None	*0.25 (0.03)*	*<.001*	0.17 (0.04)	.005	0.20 (0.03)	<.001	
Random oversampling	0.19 (0.04)	.001	0.16 (0.04)	.01	0.15 (0.03)	.02	
SMOTE	0.18 (0.06)	.007	0.12 (0.04)	.25	0.11 (0.02)	.58	
7							0.16 (0.02)
None	0.67 (0.05)	<.001	0.61 (0.03)	<.001	0.66 (0.03)	<.001	
Random oversampling	0.60 (0.08)	<.001	0.58 (0.06)	<.001	0.62 (0.04)	<.001	

XSL•FO
RenderX

Setting, message ID, and balancing method	RF (AUPRC)		LGBM (AUPRC)		CatBoost (AUPRC)		Random guessing (AUPRC), mean (SD)
	Values, mean (SD)	*P* value	Values, mean (SD)	*P* value	Values, mean (SD)	*P* value	
SMOTE	0.66 (0.06)	<.001	0.64 (0.04)	<.001	*0.70 (0.04)*	*<.001*	
12							0.03 (0.01)
None	0.07 (0.01)	<.001	*0.09 (0.01)*	*<.001*	0.06 (0.01)	<.001	
Random oversampling	0.06 (0.01)	<.001	0.08 (0.02)	<.001	0.07 (0.02)	<.001	
SMOTE	0.07 (0.01)	<.001	0.06 (0.01)	<.001	0.06 (0.01)	<.001	
14							0.06 (0.02)
None	0.09 (0.01)	.001	0.07 (0.01)	.18	0.10 (0.02)	<.001	
Random oversampling	0.10 (0.01)	<.001	0.07 (0.02)	.11	0.09 (0.01)	.001	
SMOTE	0.09 (0.02)	.002	*0.11 (0.04)*	*.003*	0.09 (0.03)	.004	
15							0.06 (0.01)
None	0.11 (0.02)	<.001	0.09 (0.03)	.003	0.09 (0.02)	.002	
Random oversampling	*0.11 (0.03)*	*<.001*	0.10 (0.02)	<.001	0.10 (0.02)	<.001	
SMOTE	0.08 (0.03)	.04	0.07 (0.01)	.07	0.08 (0.02)	.03	
20							0.53 (0.02)
None	0.64 (0.01)	<.001	0.64 (0.01)	<.001	0.62 (0.01)	<.001	
Random oversampling	0.64 (0.01)	<.001	0.64 (0.02)	<.001	0.63 (0.01)	<.001	
SMOTE	*0.64 (0.01)*	*<.001*	0.64 (0.02)	<.001	0.63 (0.01)	<.001	
21							0.24 (0.03)
None	0.36 (0.01)	<.001	0.38 (0.02)	<.001	0.34 (0.02)	<.001	
Random oversampling	0.35 (0.02)	<.001	0.36 (0.03)	<.001	0.35 (0.03)	<.001	
SMOTE	0.38 (0.02)	<.001	0.36 (0.03)	<.001	*0.39 (0.02)*	*<.001*	

[a]The best-performing configuration for each sleep advice message is italicized.

Important Features

We extracted permutation and SHAP feature importance for the best model configuration when predicting each advice message under both user-dependent and user-independent settings. Table 5 shows the features that ranked among the top 5 most important using both permutation importance and SHAP, with average AUPRC decrease and mean absolute SHAP values of at least 0.01. All the features represented the average and SD of daily Fitbit and survey features across the previous 4 days for each date on which physicians provided advice. As expected, several important features closely matched their corresponding advice messages under both data split settings. The average amount of phone use before sleep was a key factor in predicting message 4 (*refrain from using a smartphone in the bedroom*). Similarly, the average brightness of the sleep environment was the most significant feature for predicting message 7 (*darken the bedroom when you go to bed*). The model examined average sleep duration for predictions of message 21 (*try to make enough time for sleep*). When predicting message 12 (*take a nap for approximately 90 minutes before the night shift*), the SD of work hours between 4 PM and midnight on the current day across the previous 4 days was an important factor. This suggests a potential relationship between the frequency of night shifts and the triggering of this sleep advice. In addition, the response to the previous selection of message 20 (*continue current sleep habits*) contributed to the prediction of messages 7, 20, and 21. This aligns with physicians' insights that they often considered participants' responses to previous advice when providing sleep recommendations. These findings demonstrate the models' ability to capture meaningful patterns that resonate with physicians' decision-making processes.

XSL•FO
RenderX

Table 5. Important features identified using permutation importance and Shapley Additive Explanations (SHAP) for the best model configuration when predicting each advice message under both user-dependent and user-independent settings. Features are listed if they ranked among the top 5 most important using both permutation importance and SHAP, with average area under the precision-recall curve (AUPRC) decrease and mean absolute SHAP values of at least 0.01.

Setting, message ID, and feature	AUPRC decrease, mean (SD)	Absolute SHAP value, mean (SD)
Dependent		
4		
Average amount of phone use before sleep	0.19 (0.06)	0.05 (0.01)
7		
Average brightness of sleep environment	0.33 (0.05)	1.14 (0.25)
Response to previous selection of message 20	0.04 (0.03)	0.29 (0.06)
SD of brightness of sleep environment	0.03 (0.02)	0.18 (0.06)
SD of the binary indicator of activities other than work outside the homes	0.02 (0.01)	0.09 (0.04)
12		
SD of work hours on the current day for 4 PM-midnight	0.05 (0.07)	0.02 (0.01)
14		
SD of daily caffeine intake	0.03 (0.02)	0.02 (0.01)
20		
Response to previous selection of message 20	0.12 (0.03)	0.10 (0.04)
Average sleep satisfaction	0.01 (0.01)	0.02 (0.01)
Average fatigue recovery level	0.01 (0.01)	0.03 (0.01)
21		
Response to previous selection of message 20	0.11 (0.05)	0.07 (0.03)
Average sleep duration	0.03 (0.02)	0.02 (0.005)
SD of work hours on the current day for midnight-8 AM	0.03 (0.03)	0.01 (0.01)
Independent		
4		
Average amount of phone use before sleep	0.13 (0.04)	0.09 (0.02)
7		
Average brightness of sleep environment	0.47 (0.05)	1.35 (0.39)
Response to previous selection of message 20	0.03 (0.02)	0.18 (0.06)
12		
SD of work hours on the current day for 4 PM-midnight	0.03 (0.02)	0.28 (0.10)
14		
Response to previous selection of message 20	0.03 (0.04)	0.25 (0.10)
15		
Average sleep satisfaction	0.02 (0.03)	0.05 (0.04)
Average fatigue recovery level	0.02 (0.02)	0.02 (0.01)
20		
Response to previous selection of message 20	0.11 (0.02)	0.16 (0.03)
Average sleep satisfaction	0.02 (0.01)	0.03 (0.01)
21		
Response to previous selection of message 20	0.09 (0.04)	0.45 (0.07)
Average sleep duration	0.02 (0.03)	0.17 (0.08)

XSL•FO
RenderX

Discussion

Principal Findings

In this study, we used daily surveys and Fitbit to obtain physiological and behavioral data of ICU shift workers. Physicians reviewed the data to provide sleep advice from a list of 23 messages to the workers. We extracted features from the collected data and (1) conducted analyses to characterize participants' behaviors and advice and (2) constructed and evaluated 3 models (RF, LGBM, and CatBoost) and 3 data-balancing approaches (no balancing, random oversampling, and SMOTE) to predict sleep advice selections for the 7 most frequent messages sent by the physicians. The advice prediction models achieved higher AUPRC than the random guessing baseline for predicting all the 7 advice messages under both user-dependent and user-independent settings. The performance differences were statistically significant, with $P<.001$ for 13 tests and $P=.003$ for 1 test. Feature importance analysis revealed several important features that closely matched their corresponding advice messages sent to participants. For instance, for message 7 (*darken the bedroom when you go to bed*), the models primarily examined the average brightness of the sleep environment to make predictions. Similarly, the average amount of phone use before sleep was a key factor in predicting message 4 (*refrain from using a smartphone in the bedroom*). The analysis also supported physicians' insights that they often considered participants' responses to previous advice when providing sleep recommendations.

Participant Characteristics

Hierarchical clustering revealed different work shifts and behavior patterns among clusters and subclusters, and we investigated their relationships with their well-being scores. First, work shifts on the current day contributed to low well-being scores for clusters 3 and 5, and several well-being metrics were further distinguished by their different shift patterns. For example, cluster 3 had a much shorter average sleep duration of approximately 220 minutes and the lowest sleep quality. This was potentially caused by day shifts the day before and midnight shifts on the current day for almost all samples. In this cluster, participants finished their work at approximately 4 PM the day before; they then started the next shift period at approximately midnight and finished at approximately 8 AM on the following day. After work, they filled out the morning surveys and reported their well-being, sleep duration, and sleep quality. Low scores on these metrics reflected their tiredness shortly after work and lack of rest between work shifts.

On the other hand, most samples in cluster 5 did not have shifts the day before but had day shifts on the current day. As a result, participants could have more time for sleep, reflected by a longer average sleep duration of 400 minutes, better sleep quality, and higher well-being scores in the morning compared with samples in cluster 3. Such discrepancies in shift patterns and well-being metrics could even be observed between subclusters *d* and *k* of the same larger cluster 2. Although most samples of both subclusters had midnight shifts the day before, subcluster *d* majorly did not have shifts on the current day and achieved an average level of well-being scores. The profile of subcluster *k* was more similar to that of cluster 3, with midnight shifts on the current day and much worse well-being and sleep quality than in subcluster *d*, which is also supported by the t-SNE plot (right plot in Figure S1 in Multimedia Appendix 1), where subcluster *k* is closer to subcluster *h* (equivalence of cluster 3) than to subcluster *d*. Furthermore, most of cluster 1 did not have a work shift on the current day, which led to its high well-being scores as 1 day off probably helped participants relax. Nevertheless, shifts on the current day were not always negatively correlated with well-being. Specifically, cluster 0 had above-average well-being scores although most samples had work shifts on the current day.

Furthermore, several subclusters were characterized by some distinct features of their profiles. For example, subclusters *i* and *l* of cluster 5 had similar work patterns, but they differed significantly in average daily step count, where subcluster *i* had the highest count of >10,000 among all subclusters, whereas subcluster *l* had the lowest count of <2000. Although multiple clusters and subclusters contained samples with overtime shifts of >8 hours, subcluster *e* was the only one with >9 hours of average work duration for 3 consecutive days (10.6 hours the day before the day before, 9.4 hours the day before, and 11.5 hours on the current day). The long work duration might also contribute to other characteristics of this subcluster, including the shortest time to fall asleep and the highest daily caffeine consumption. Despite the existence of some outliers, the clusters and subclusters generally grouped well, as shown in Figure S1 in Multimedia Appendix 1, and their profiles summarized from average feature values provided many insights into participants' daily behaviors, shifts, and well-being and their interrelationships.

Advice Predictions

While we evaluated the advice prediction models using the AUPRC computed from predicted class probabilities, it is also important to understand the model performance in terms of actual predicted classes. Given the class imbalances, adjusting the decision thresholds was necessary to optimize prediction performance. Specifically, for each classifier initialized with its optimal hyperparameters, we performed 5-fold stratified cross-validation on the training set to determine the optimal threshold that maximized its F_1-score. The F_1-score was chosen as the performance metric to obtain a balance between missed advice predictions and false alarms. The optimal threshold was then applied to convert prediction probabilities into class labels on the test set. As we repeated the tuning and training process 10 times with different random seeds for each model configuration, we created an ensemble of these classifiers to mitigate potential overfitting. Final predictions were determined through majority voting across the ensemble. To resolve ties during voting, we used only 9 of the 10 classifiers generated for each model configuration. We report the sensitivity, specificity, precision, and F_1-scores of the ensembles from the RF, LGBM, and CatBoost classifiers with different data-balancing approaches in Tables S4, S5, and S6 in Multimedia Appendix 1, respectively.

XSL•FO
RenderX

Table 6 presents the performance of the model configurations with the highest F_1-score for predicting each sleep advice message under user-dependent and user-independent settings. Baseline F_1-scores, calculated by always predicting the positive class, are also provided for comparison. All models outperformed the baseline. Under the user-dependent setting, the models achieved sensitivity and specificity within the range of 0.5 to 0.9. Sensitivity tended to decrease for less frequent advice messages, whereas specificity showed the opposite trend. Under the user-independent setting, the models consistently demonstrated decent sensitivity, exceeding 0.7 across all advice messages, whereas specificity varied between 0.4 and 0.9. The higher sensitivity observed in this setting may suggest that the models were more effective at identifying universal decision rules shared across participants than at accounting for individual behavioral variability.

Table 6. Advice prediction sensitivity, specificity, precision, and F1-score of the best-performing ensemble classifier for each advice message under both user-dependent and user-independent settings. Ensemble classifiers were constructed through majority voting across 9 random forest (RF), light gradient-boosting machine (LGBM), or CatBoost classifiers trained using different random seeds. The decision threshold for each classifier was optimized to maximize the F1-score. The baseline F1-score from always predicting the positive class is also listed.

Setting and message ID	Model	Balancing method	Sensitivity	Specificity	Precision	F_1-score	Baseline F_1-score (always positive)
Dependent							
4	LGBM	None	0.56	0.86	0.34	0.42	0.20
7	CatBoost	Random oversampling	0.69	0.88	0.52	0.59	0.28
12	RF	None	0.50	0.83	0.27	0.35	0.19
14	RF	Random oversampling	0.58	0.81	0.27	0.37	0.19
15	RF	Random oversampling	0.55	0.72	0.29	0.38	0.30
20	RF	SMOTE[a]	0.84	0.50	0.67	0.75	0.71
21	CatBoost	None	0.81	0.58	0.34	0.48	0.35
Independent							
4	RF	None	0.95	0.54	0.17	0.29	0.17
7	CatBoost	SMOTE	0.73	0.93	0.63	0.68	0.25
12	LGBM	Random oversampling	1.00	0.60	0.07	0.14	0.06
14	CatBoost	None	0.91	0.68	0.12	0.22	0.09
15	RF	SMOTE	0.77	0.48	0.08	0.15	0.11
20	RF	SMOTE	0.82	0.46	0.61	0.70	0.67
21	RF	Random oversampling	0.81	0.44	0.31	0.45	0.38

[a]SMOTE: synthetic minority oversampling technique.

Single-Label Versus Multilabel Classification

As physicians could select multiple messages when providing sleep advice, it is feasible to formulate sleep advice prediction as a multilabel classification task, which outputs predictions for all message selections simultaneously. Compared to the current approach of training separate single-label classifiers for each message, multilabel classification has the advantage of leveraging information about message coselections during training, potentially improving prediction performance. However, this approach presents several challenges. First, some models do not inherently support multilabel classification and, instead, combine separate single-label classifiers for multilabel outputs. This strategy fails to capture and use the coselection information effectively. Second, the multilabel setting requires careful handling of data imbalances both within individual labels and across their co-occurrence patterns. Finally, multilabel classification complicates feature importance analysis, which aims to identify key contributors for predicting individual message selections. In fact, we trained a user-dependent multilabel RF classifier without any data balancing and did not observe substantial performance improvement compared to the single-label counterparts.

Comparison With Prior Work

Our system provided sleep and behavior recommendations for shift workers with poor sleep quality. Although our system did not follow the CBTI protocol, we borrowed elements from it, such as sleep hygiene advice and relaxation. We designed sleep recommendations specifically targeted for shift workers, including strategic napping and clockwise shift rotation. Our system offered several advantages over existing CBTI and sleep RSs. First, while most mobile and web-based CBTI systems require users to review sleep-related learning modules, our system reduced users' time commitment by sending them simple pieces of advice to follow, which could be beneficial as shift workers often have a poor work-life balance [53]. Consequently, our study yielded a low dropout rate of 5% (3/64), as opposed to the rates of other CBTI systems that range from 4% to 61% [15,16,18-23,26,29,54].

XSL•FO
RenderX

Moreover, our system received the most support from sleep specialists compared to existing sleep RSs, which only have limited or even no clinician support. For example, clinicians only provided recommendation templates for further customization by the automated SleepCoacher system [30]. The cohort-based sleep recommender by Daskalova et al [31] relied on profiles of similar users as the only recommendation source. The system developed by Pandey et al [32] constructed its recommendations based on discovered relationships between the user's lifestyle and sleep. The lack of clinician support could impair the credibility of the provided sleep recommendations among users. For instance, users of SleepCoacher suggested adding justifications for the advice [30], and several users of the cohort-based sleep recommender by Daskalova et al [31] thought that the recommendations were implausible and even untrustworthy as they contradicted the users' own beliefs. However, our system only delivered clinician-reviewed recommendations, which greatly enhanced their trustworthiness.

Furthermore, only our study has evaluated the accuracy of an algorithm for automatic sleep advice provision. Many existing CBTI systems have been able to provide sleep restriction and sleep hygiene recommendations automatically, but no analysis is available of how close these generated recommendations are to actual clinicians' advice. Similarly, sleep RSs have not compared their recommendations with professionals' suggestions. Specifically, clinicians prepared all the recommendations in the preliminary study of SleepCoacher, and the provisions were automated by a correlation-based algorithm in the final study, but the study did not evaluate the accuracy of the algorithm for substituting clinicians [30]. Despite the effectiveness of these systems, an analysis of algorithm accuracy could further justify provided recommendations and resolve users' doubts about their credibility. On the other hand, although our system is not automated at its current stage, we developed an algorithm and assessed its accuracy for automatic advice provision, and it showed decent performance and great potential in providing clinical-level sleep recommendations.

Limitations and Future Work

False Predictions

To investigate the possible causes of false-positive and false-negative predictions, we examined the relationship between feature values and advice predictions. Specifically, we chose the feature *brightness_sleep* (brightness level of the sleep environment) and 1 repetition of the dependent advice prediction model for message 7 (*darken the bedroom when you go to bed*). We selected this combination because the feature had a dominant effect on predictions of the message label, as shown in the *Important Features* section, and could demonstrate the limitations of the models more clearly with the existence of an explicit decision rule.

For each participant, we plotted the trend of the feature across all days of their enrollment and marked each predicted and the actual selection of the message separately. Figure S2 in Multimedia Appendix 1 includes such plots for 6 participants to illustrate incorrect predictions and limitations of the advice prediction models. First, the models did not have memories about previous message selections and might produce false positives as duplicate responses to previous feature values. For example, the model made 3 positive predictions on days 29, 31, and 33 in response to a peak of brightness on day 29 for participant 1102, yet only the prediction on day 29 was an actual positive. Similarly, 2 positive predictions were made for participant 1154 on days 26 and 29 for a brightness surge on day 26, but the latter was a false-positive. Shortening the feature averaging window before each message selection might eliminate some of these false positives, but it would increase the risk of producing false negatives when a day with a favorable value for message selection is excluded from the window.

Another scenario of false positives occurred for participants 1315 and 1181. Both participants received 4 false positives although their brightness values only fluctuated within a small range of approximately 30 (scale of 1-100). A closer look at individual decision trees of the trained RF showed that some trees considered SDs of brightness levels but most trees made predictions by comparing average brightness with a threshold. Thus, false-positive predictions were made because average brightness exceeded the thresholds of most decision trees despite the lack of brightness peaks. However, it could be tricky to make correct predictions under such a scenario. For example, physicians selected the message twice on days 9 and 16 for participant 1315 although brightness remained stable. A plausible explanation could be that physicians initially suggested that the participant lower the brightness because it was constantly above the threshold, but later, they stopped as neither lower levels nor high peaks of brightness were observed after the suggestions. Nevertheless, it would be very difficult for the model to capture such change.

Unlike false positives, most false negatives occurred only under 1 scenario: when brightness peaks were not high enough and adjacent days had very low feature values, which caused averages to fall below the threshold, for example, on day 33 for participant 1164 and on days 29, 31, and 33 for participant 1167. Measuring the relative height of a peak (difference between maximum and minimum values of the feature averaging window) instead of averaging might be helpful to eliminate these false negatives, but it may produce other false negatives when feature values remain high and stable. From illustrative plots of feature values and advice predictions, we demonstrated several causes of false-positive and false-negative predictions, but further investigations and experiments were necessary to properly improve the advice prediction models as some false positives might benefit participants and strategies to eliminate certain false predictions might introduce others.

System Usability

Although our developed algorithm achieved a decent performance in predicting sleep advice under both dependent and independent settings, the algorithm is limited to the 7 most popular advice messages among 23 choices due to infrequent occurrences of the remaining options. Therefore, further studies are necessary to gather enough data to enable predictions for less frequent advice labels.

We gathered feedback on the sleep advice models from clinicians (n=2) through questionnaires. They commented that they decided their suggestions mainly based on morning survey,

XSL•FO
RenderX

work shift, and wearable data. After they selected the suggestions, they referred to the output of the models and made the final decisions.

Future studies can compare the intervention effects of a clinician-only input system, an algorithm-assisted system, and a fully automated system to demonstrate the usefulness of the machine learning algorithm. It will also be beneficial to validate the generalizability of our advice prediction models with more diverse participant populations.

Given the effectiveness of existing sleep RSs with limited clinical support, we believe that our automated system has the advantage of being built on the knowledge of sleep specialists even if our algorithm still requires improvement to further boost performance. We chose to evaluate our models in user-dependent and user-independent settings instead of using leave-one-out cross-validation because the advice frequencies were extremely imbalanced across participants. However, such a scheme will fit real-world scenarios the best when models use all available data from current users to predict sleep advice for a new user. It is also worth experimenting with personalization beyond leave-one-out cross-validation by continuously incorporating data accumulated from new users as they progress to better adapt models to specific user characteristics.

Conclusions

We developed a web- and app-based system to provide individualized sleep and behavior advice with CBTI elements for shift workers. We collected data from shift workers (N=61) in the ICU at 2 hospitals in Japan for 5 weeks, which included Fitbit, survey, and sleep advice data. We used hierarchical clustering to characterize shift workers' physiological and behavioral features and identified clusters with distinct patterns. For example, days with low sleep duration and the lowest sleep quality were associated with a day shift on the day before and a midnight shift on the current day. Furthermore, we constructed and evaluated 3 machine learning models (RF, LGBM, and CatBoost) and 3 data-balancing approaches (no balancing, random oversampling, and SMOTE) to predict sleep advice selections for the 7 most frequent advice messages sent by the physicians. Our advice prediction models achieved higher AUPRC and F_1-scores compared to the baseline. The analysis of the feature importance of our models found several important features that closely matched their corresponding advice messages sent to participants. Although our current system requires physician input, an accurate machine learning algorithm would be promising for automation without hurting the trustworthiness of the selected recommendations. Despite a decent performance, the algorithms developed in this study are limited to the 7 most popular advice messages among 23 choices due to infrequent occurrences of the remaining advice options. Therefore, further studies are necessary to gather enough data to enable predictions for less frequent advice labels.

Acknowledgments

This research was funded by the Japan Agency for Medical Research and Development (19217687) and the National Science Foundation (1840167).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

AIM, RS, MS, and AS contributed to the conceptualization and methodology of the study. Data curation was performed by AIM, RS, and MS. YS, ACO, and HY conducted formal analyses. YS developed the software and performed validation and visualization. YS and ACO wrote the original draft, whereas AS reviewed and edited the manuscript. AS provided supervision and project administration. Both MS and AS were responsible for funding acquisition.

Conflicts of Interest

AS provided consulting services for Suntory Global Innovation Center and received honoraria from Oak Ridge Associated Universities, the Nara Institute of Science and Technology, the Taiwanese Society for Nutritional Psychiatry Research, the Korea Advanced Institute of Science and Technology, Amrita Vishwa Vidyapeetham, the European Science Foundation, and the National Science Foundation. AS also received travel support from Apple and the Taiwanese Society for Nutritional Psychiatry Research, as well as research funding from Meta Platforms, General Motors Company, POLA, and Nippon Electric Company.

Multimedia Appendix 1
Additional tables and figures for participant characteristic clustering and advice prediction models.
[DOCX File , 2765 KB - formative_v9i1e65000_app1.docx]

References

XSL•FO
RenderX

1. Alterman T, Luckhaupt SE, Dahlhamer JM, Ward BW, Calvert GM. Prevalence rates of work organization characteristics among workers in the U.S.: data from the 2010 National Health Interview Survey. Am J Ind Med 2013 Jun;56(6):647-659 [FREE Full text] [doi: 10.1002/ajim.22108] [Medline: 22911666]
2. Wickwire EM, Geiger-Brown J, Scharf SM, Drake CL. Shift work and shift work sleep disorder: clinical and organizational perspectives. Chest 2017 May;151(5):1156-1172 [FREE Full text] [doi: 10.1016/j.chest.2016.12.007] [Medline: 28012806]
3. Haile KK, Asnakew S, Waja T, Kerbih HB. Shift work sleep disorders and associated factors among nurses at federal government hospitals in Ethiopia: a cross-sectional study. BMJ Open 2019 Aug 27;9(8):e029802 [FREE Full text] [doi: 10.1136/bmjopen-2019-029802] [Medline: 31462478]
4. Flo E, Pallesen S, Magerøy N, Moen BE, Grønli J, Hilde Nordhus I, et al. Shift work disorder in nurses--assessment, prevalence and related health problems. PLoS One 2012 Apr 2;7(4):e33981 [FREE Full text] [doi: 10.1371/journal.pone.0033981] [Medline: 22485153]
5. Di Milia L, Waage S, Pallesen S, Bjorvatn B. Shift work disorder in a random population sample--prevalence and comorbidities. PLoS One 2013;8(1):e55306 [FREE Full text] [doi: 10.1371/journal.pone.0055306] [Medline: 23372847]
6. Cheng WJ, Cheng Y. Night shift and rotating shift in association with sleep problems, burnout and minor mental disorder in male and female employees. Occup Environ Med 2017 Jul;74(7):483-488. [doi: 10.1136/oemed-2016-103898] [Medline: 27810939]
7. Khan WA, Jackson ML, Kennedy GA, Conduit R. A field investigation of the relationship between rotating shifts, sleep, mental health and physical activity of Australian paramedics. Sci Rep 2021 Jan 13;11(1):866 [FREE Full text] [doi: 10.1038/s41598-020-79093-5] [Medline: 33441601]
8. Saleh AM, Awadalla NJ, El-masri YM, Sleem WF. Impacts of nurses' circadian rhythm sleep disorders, fatigue, and depression on medication administration errors. Egypt J Chest Dis Tuberc 2014 Jan;63(1):145-153. [doi: 10.1016/j.ejcdt.2013.10.001]
9. Blytt KM, Bjorvatn B, Moen BE, Pallesen S, Harris A, Waage S. The association between shift work disorder and turnover intention among nurses. BMC Nurs 2022 Jun 06;21(1):143 [FREE Full text] [doi: 10.1186/s12912-022-00928-9] [Medline: 35668393]
10. Wright KP, Bogan RK, Wyatt JK. Shift work and the assessment and management of shift work disorder (SWD). Sleep Med Rev 2013 Feb;17(1):41-54. [doi: 10.1016/j.smrv.2012.02.002] [Medline: 22560640]
11. Qaseem A, Kansagara D, Forciea MA, Cooke M, Denberg TD, Clinical Guidelines Committee of the American College of Physicians. Management of chronic insomnia disorder in adults: a clinical practice guideline from the American College of Physicians. Ann Intern Med 2016 Jul 19;165(2):125-133 [FREE Full text] [doi: 10.7326/M15-2175] [Medline: 27136449]
12. Williams J, Roth A, Vatthauer K, McCrae CS. Cognitive behavioral treatment of insomnia. Chest 2013 Feb 01;143(2):554-565 [FREE Full text] [doi: 10.1378/chest.12-0731] [Medline: 23381322]
13. Järnefelt H, Lagerstedt R, Kajaste S, Sallinen M, Savolainen A, Hublin C. Cognitive behavioral therapy for shift workers with chronic insomnia. Sleep Med 2012 Dec;13(10):1238-1246. [doi: 10.1016/j.sleep.2012.10.003] [Medline: 23168269]
14. Järnefelt H, Sallinen M, Luukkonen R, Kajaste S, Savolainen A, Hublin C. Cognitive behavioral therapy for chronic insomnia in occupational health services: analyses of outcomes up to 24 months post-treatment. Behav Res Ther 2014 May;56:16-21. [doi: 10.1016/j.brat.2014.02.007] [Medline: 24632111]
15. Järnefelt H, Härmä M, Sallinen M, Virkkala J, Paajanen T, Martimo KP, et al. Cognitive behavioural therapy interventions for insomnia among shift workers: RCT in an occupational health setting. Int Arch Occup Environ Health 2020 Jul 18;93(5):535-550 [FREE Full text] [doi: 10.1007/s00420-019-01504-6] [Medline: 31853633]
16. Peter L, Reindl R, Zauter S, Hillemacher T, Richter K. Effectiveness of an online CBT-I intervention and a face-to-face treatment for shift work sleep disorder: a comparison of sleep diary data. Int J Environ Res Public Health 2019 Aug 24;16(17):3081 [FREE Full text] [doi: 10.3390/ijerph16173081] [Medline: 31450619]
17. Retzer L, Feil M, Reindl R, Richter K, Lehmann R, Stemmler M, et al. Anonymous online cognitive behavioral therapy for sleep disorders in shift workers-a study protocol for a randomized controlled trial. Trials 2021 Aug 16;22(1):539 [FREE Full text] [doi: 10.1186/s13063-021-05437-9] [Medline: 34399824]
18. Ritterband LM, Thorndike FP, Gonder-Frederick LA, Magee JC, Bailey ET, Saylor DK, et al. Efficacy of an internet-based behavioral intervention for adults with insomnia. Arch Gen Psychiatry 2009 Jul;66(7):692-698 [FREE Full text] [doi: 10.1001/archgenpsychiatry.2009.66] [Medline: 19581560]
19. Espie CA, Kyle SD, Williams C, Ong JC, Douglas NJ, Hames P, et al. A randomized, placebo-controlled trial of online cognitive behavioral therapy for chronic insomnia disorder delivered via an automated media-rich web application. Sleep 2012 Jun 01;35(6):769-781 [FREE Full text] [doi: 10.5665/sleep.1872] [Medline: 22654196]
20. Batterham PJ, Christensen H, Mackinnon AJ, Gosling JA, Thorndike FP, Ritterband LM, et al. Trajectories of change and long-term outcomes in a randomised controlled trial of internet-based insomnia treatment to prevent depression. BJPsych Open 2017 Sep;3(5):228-235 [FREE Full text] [doi: 10.1192/bjpo.bp.117.005231] [Medline: 28959453]
21. Hagatun S, Vedaa Ø, Nordgreen T, Smith OR, Pallesen S, Havik OE, et al. The short-term efficacy of an unguided internet-based cognitive-behavioral therapy for insomnia: a randomized controlled trial with a six-month nonrandomized follow-up. Behav Sleep Med 2019;17(2):137-155. [doi: 10.1080/15402002.2017.1301941] [Medline: 28345961]

XSL•FO
RenderX

22. Cheng P, Luik AI, Fellman-Couture C, Peterson E, Joseph CL, Tallent G, et al. Efficacy of digital CBT for insomnia to reduce depression across demographic groups: a randomized trial. Psychol Med 2019 Feb 24;49(3):491-500 [FREE Full text] [doi: 10.1017/S0033291718001113] [Medline: 29792241]
23. Freeman D, Sheaves B, Goodwin GM, Yu LM, Nickless A, Harrison PJ, et al. The effects of improving sleep on mental health (OASIS): a randomised controlled trial with mediation analysis. Lancet Psychiatry 2017 Oct;4(10):749-758 [FREE Full text] [doi: 10.1016/S2215-0366(17)30328-0] [Medline: 28888927]
24. Beun RJ, Brinkman WP, Fitrianie S, Griffioen-Both F, Horsch C, Lancee J, et al. Improving adherence in automated e-coaching. In: Proceedings of the 11th International Conference on Persuasive Technology. 2016 Presented at: PERSUASIVE '16; April 5-7, 2016; Salzburg, Austria p. 276-287 URL: https://link.springer.com/chapter/10.1007/978-3-319-31510-2_24 [doi: 10.1007/978-3-319-31510-2_24]
25. Beun RJ, Fitrianie S, Griffioen-Both F, Spruit S, Horsch C, Lancee J, et al. Talk and Tools: the best of both worlds in mobile user interfaces for e-coaching. Pers Ubiquit Comput 2017 May 19;21(4):661-674. [doi: 10.1007/S00779-017-1021-5]
26. Horsch CH, Lancee J, Griffioen-Both F, Spruit S, Fitrianie S, Neerincx MA, et al. Mobile phone-delivered cognitive behavioral therapy for insomnia: a randomized waitlist controlled trial. J Med Internet Res 2017 Apr 11;19(4):e70 [FREE Full text] [doi: 10.2196/jmir.6524] [Medline: 28400355]
27. Kuhn E, Weiss BJ, Taylor KL, Hoffman JE, Ramsey KM, Manber R, et al. CBT-I coach: a description and clinician perceptions of a mobile app for cognitive behavioral therapy for insomnia. J Clin Sleep Med 2016 Apr 15;12(4):597-606 [FREE Full text] [doi: 10.5664/jcsm.5700] [Medline: 26888586]
28. Koffel E, Kuhn E, Petsoulis N, Erbes CR, Anders S, Hoffman JE, et al. A randomized controlled pilot study of CBT-I coach: feasibility, acceptability, and potential impact of a mobile phone application for patients in cognitive behavioral therapy for insomnia. Health Informatics J 2018 Mar;24(1):3-13 [FREE Full text] [doi: 10.1177/1460458216656472] [Medline: 27354394]
29. Kuhn E, Miller KE, Puran D, Wielgosz J, YorkWilliams SL, Owen JE, et al. A pilot randomized controlled trial of the insomnia coach mobile app to assess its feasibility, acceptability, and potential efficacy. Behav Ther 2022 May;53(3):440-457. [doi: 10.1016/j.beth.2021.11.003] [Medline: 35473648]
30. Daskalova N, Metaxa-Kakavouli D, Tran A, Nugent N, Boergers J, McGeary J, et al. SleepCoacher: a personalized automated self-experimentation system for sleep recommendations. In: Proceedings of the 29th Annual Symposium on User Interface Software and Technology. 2016 Presented at: UIST '16; October 16-19, 2016; Tokyo, Japan p. 347-358 URL: https://dl.acm.org/doi/10.1145/2984511.2984534 [doi: 10.1145/2984511.2984534]
31. Daskalova N, Lee B, Huang J, Ni C, Lundin J. Investigating the effectiveness of cohort-based sleep recommendations. Proc ACM Interact Mob Wearable Ubiquitous Technol 2018 Sep 18;2(3):1-19. [doi: 10.1145/3264911]
32. Pandey V, Upadhyay D, Nag N, Jain R. Personalized user modelling for context-aware lifestyle recommendations to improve sleep. In: Proceedings of the 5th International Workshop on Health Recommender Systems. 2020 Presented at: HealthRecSys ’20; September 26, 2020; Virtual Event p. 2020 URL: https://www.researchgate.net/publication/344345398_Personalized_User_Modelling_for_Context-Aware_Lifestyle_Recommendations_to_Improve_Sleep
33. Kay M, Choe EK, Shepherd J, Greenstein B, Watson N, Consolvo S, et al. Lullaby: a capture and access system for understanding the sleep environment. In: Proceedings of the 2012 ACM Conference on Ubiquitous Computing. 2012 Presented at: UbiComp '12; September 5-8, 2012; Pittsburgh, PA p. 226-234 URL: https://dl.acm.org/doi/10.1145/2370216.2370253 [doi: 10.1145/2370216.2370253]
34. Bauer J, Consolvo S, Greenstein B, Schooler J, Wu E, Watson N, et al. ShutEye: encouraging awareness of healthy sleep recommendations with a mobile, peripheral display. In: Proceedings of the 2012 SIGCHI Conference on Human Factors in Computing Systems. 2012 Presented at: CHI '12; May 5-10, 2012; Austin, TX p. 1401-1410 URL: https://dl.acm.org/doi/10.1145/2207676.2208600 [doi: 10.1145/2207676.2208600]
35. Daskalova N, Yoon J, Wang Y, Araujo C, Beltran G, Nugent N, et al. SleepBandits: guided flexible self-experiments for sleep. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. 2020 Presented at: CHI '20; April 25-30, 2020; Honolulu, HI p. 1-13 URL: https://dl.acm.org/doi/10.1145/3313831.3376584 [doi: 10.1145/3313831.3376584]
36. Ito-Masui A, Kawamoto E, Sakamoto R, Yu H, Sano A, Motomura E, et al. Internet-based individualized cognitive behavioral therapy for shift work sleep disorder empowered by well-being prediction: protocol for a pilot study. JMIR Res Protoc 2021 Mar 18;10(3):e24799 [FREE Full text] [doi: 10.2196/24799] [Medline: 33626497]
37. Ito-Masui A, Sakamoto R, Matsuo E, Kawamoto E, Motomura E, Tanii H, et al. Effect of an internet-delivered cognitive behavioral therapy-based sleep improvement app for shift workers at high risk of sleep disorder: single-arm, nonrandomized trial. J Med Internet Res 2023 Aug 22;25:e45834 [FREE Full text] [doi: 10.2196/45834] [Medline: 37606971]
38. Yu H, Itoh A, Sakamoto R, Shimaoka M, Sano A. Forecasting health and wellbeing for shift workers using job-role based deep neural network. In: Proceedings of the 9th EAI International Conference on Wireless Mobile Communication and Healthcare. 2020 Presented at: MobiHealth' 20; November 19, 2020; Virtual Event p. 89-103 URL: https://link.springer.com/chapter/10.1007/978-3-030-70569-5_6 [doi: 10.1007/978-3-030-70569-5_6]

XSL•FO
RenderX

39. Richman JS, Moorman JR. Physiological time-series analysis using approximate entropy and sample entropy. Am J Physiol Heart Circ Physiol 2000 Jun;278(6):H2039-H2049 [FREE Full text] [doi: 10.1152/ajpheart.2000.278.6.H2039] [Medline: 10843903]
40. Phillips AJ, Clerx WM, O'Brien CS, Sano A, Barger LK, Picard RW, et al. Irregular sleep/wake patterns are associated with poorer academic performance and delayed circadian and sleep/wake timing. Sci Rep 2017 Jun 12;7(1):3216 [FREE Full text] [doi: 10.1038/s41598-017-03171-4] [Medline: 28607474]
41. Lunsford-Avery JR, Engelhard MM, Navar AM, Kollins SH. Validation of the sleep regularity index in older adults and associations with cardiometabolic risk. Sci Rep 2018 Sep 21;8(1):14158 [FREE Full text] [doi: 10.1038/s41598-018-32402-5] [Medline: 30242174]
42. Murtagh F, Legendre P. Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion? J Classif 2014 Oct 18;31(3):274-295. [doi: 10.1007/s00357-014-9161-z]
43. Nielsen F. Hierarchical clustering. In: Nielsen F, editor. Introduction to HPC with MPI for Data Science. Cham, Switzerland: Springer; 2016:195-211. [doi: 10.1007/978-3-319-21903-5]
44. van der Maaten L, Hinton G. Viualizing data using t-SNE. J Mach Learn Res 2008:2579-2605 [FREE Full text]
45. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: a highly efficient gradient boosting decision tree. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017 Presented at: NIPS '17; December 4-9, 2017; Long Beach, CA p. 3149-3157 URL: https://dl.acm.org/doi/10.5555/3294996.3295074
46. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. 2018 Presented at: NIPS '18; December 3-8, 2018; Montréal, QC p. 6639-6649 URL: https://dl.acm.org/doi/abs/10.5555/3327757.3327770
47. Bergstra J, Bengio Y. Random Search for Hyper-Parameter Optimization. J Mach Learn Res 2012;13:281-305 [FREE Full text]
48. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell 2002 Jun 01;16:321-357. [doi: 10.1613/jair.953]
49. Breiman L. Random forests. Mach Learn 2001;45(1):5-32. [doi: 10.1023/A:1010933404324]
50. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017 Presented at: NIPS '17; December 4-9, 2017; Long Beach, CA p. 4768-4777 URL: https://dl.acm.org/doi/10.5555/3295222.3295230
51. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2020 Jan 17;2(1):56-67 [FREE Full text] [doi: 10.1038/s42256-019-0138-9] [Medline: 32607472]
52. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 2020 Mar;17(3):261-272 [FREE Full text] [doi: 10.1038/s41592-019-0686-2] [Medline: 32015543]
53. Puttonen S, Härmä M, Hublin C. Shift work and cardiovascular disease - pathways from circadian stress to morbidity. Scand J Work Environ Health 2010 Mar 20;36(2):96-108 [FREE Full text] [doi: 10.5271/sjweh.2894] [Medline: 20087536]
54. Ritterband LM, Thorndike FP, Ingersoll KS, Lord HR, Gonder-Frederick L, Frederick C, et al. Effect of a web-based cognitive behavior therapy for insomnia intervention with 1-year follow-up: a randomized clinical trial. JAMA Psychiatry 2017 Jan 01;74(1):68-75. [doi: 10.1001/jamapsychiatry.2016.3249] [Medline: 27902836]

Abbreviations

AUPRC: area under the precision-recall curve
CBTI: cognitive behavioral therapy for insomnia
ICU: intensive care unit
LGBM: light gradient-boosting machine
RF: random forest
RS: recommender system
SHAP: Shapley additive explanations
SMOTE: synthetic minority oversampling technique
SRI: sleep regularity index
SWSD: shift work sleep disorder
t-SNE: t-distributed stochastic neighbor embedding

XSL•FO
RenderX

Edited by A Mavragani; submitted 02.08.24; peer-reviewed by A Kalluchi, Y Chen, D Wang; comments to author 07.10.24; revised version received 15.01.25; accepted 12.02.25; published 01.04.25.

Please cite as:
Shen Y, Choto Olivier A, Yu H, Ito-Masui A, Sakamoto R, Shimaoka M, Sano A
Personalized Physician-Assisted Sleep Advice for Shift Workers: Algorithm Development and Validation Study
JMIR Form Res 2025;9:e65000
URL: https://formative.jmir.org/2025/1/e65000
doi:10.2196/65000
PMID:

©Yufei Shen, Alicia Choto Olivier, Han Yu, Asami Ito-Masui, Ryota Sakamoto, Motomu Shimaoka, Akane Sano. Originally published in JMIR Formative Research (https://formative.jmir.org), 01.04.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Examining the Client Experience of Digital Tools in Blended Care Therapy: Qualitative Interview Study

Emily G Lattie[1], PhD; Miranda Beltzer[1], PhD; Alethea Varra[1], PhD; Connie E Chen[1], MD; Anita Lungu[1], PhD

Lyra Health, Burlingame, CA, United States

Corresponding Author:
Emily G Lattie, PhD
Lyra Health
270 East Lane
Burlingame, CA, 94010
United States
Phone: 1 (877) 505 7147
Email: elattie@lyrahealth.com

Abstract

Background: Lyra Health's short-term blended care therapy model, Lyra Care Therapy (LCT), has demonstrated effectiveness at scale. In LCT, clients participate in synchronous telehealth sessions and asynchronous guided practice sessions, in which they are provided with digital tools to reinforce key concepts and skills. These digital tools include animated video lessons that use storytelling to show characters learning and implementing new skills from therapy, written psychoeducational materials, interactive exercises that prompt reflection and skills practice, symptom assessments, and messaging with therapists. Past research on LCT found that time spent in therapy sessions and viewing digital video lessons predicts improvements in depression and anxiety symptoms.

Objective: This study aims to explore the client experience of LCT digital tools and to understand clients' perceived benefits and challenges of using digital tools while in LCT.

Methods: In total, 12 ethnically and racially diverse adults (5 male, 6 female, and 1 pangender) who had graduated from LCT in the previous 4 months participated in semistructured interviews. Interviews focused on experiences with the digital components of LCT (ie, video lessons, symptom assessments, and written exercises). Transcripts were analyzed using thematic analytic methods to determine the benefits and challenges associated with components of the LCT model.

Results: In total, 3 primary themes were generated through thematic analysis. These themes centered around supporting knowledge and skill development, the benefits and challenges associated with the range of digital tools, and the combination of flexibility and accountability promoting positive change. First, we summarize the ways in which guided practice sessions allow clients to develop additional knowledge and learn skills related to their mental health and well-being. Then, we describe how the range of digital tools included in LCT presents different benefits and challenges for clients. Finally, we discuss how flexibility inherent in having both synchronous and asynchronous sessions, along with the accountability from a provider, encourages clients to continue to practice skills related to their mental health and well-being.

Conclusions: Results provide insights into the unique contributions of different components of the LCT model on therapeutic gains. While perceived time constraints and content preferences can impact engagement with digital tools, overall the digital tools were perceived as carrying significant value for participants in the LCT program.

(JMIR Form Res 2025;9:e68249) doi:10.2196/68249

KEYWORDS

anxiety; depression; blended care therapy; mental health care; digital mental health; digital tools; qualitative interviews; Lyra Care Therapy; video lesson; symptom assessment; written exercise; thematic analysis; LCT model; therapeutic value; client experience

XSL•FO
RenderX

Introduction

Blended care therapy programs, which combine synchronous face-to-face therapy sessions with asynchronous therapeutic digital content, can offer the benefits of both traditional face-to-face and more novel internet-based mental health services [1-3]. Blended care therapy programs can differ in the degree to which they focus on the synchronous and asynchronous components. Some programs have focused primarily on adding face-to-face contact to digital interventions [4-6], while other programs have more deeply integrated the synchronous and asynchronous components of therapy to create new models of care delivery [7,8]. These integrated blended care therapy models typically address many of the barriers to engaging with traditionally delivered in-person care, including transportation, geography, mobility, and time constraints [9,10]. These blended care therapy models also address many of the barriers to engaging with entirely asynchronous internet-based mental health interventions, such as low perceived relevance of the intervention tools to one's own life, a lack of motivation to consistently use digital tools, and a poor user experience with digital tools [11,12].

Lyra Care Therapy (LCT) is Lyra Health's short-term blended care therapy model that is offered as an employee mental health benefit. LCT is an integrated blended care therapy model that has demonstrated effectiveness at scale for treating both depression and anxiety across diverse racial and ethnic groups [13,14]. This is a flexible model of blended care therapy and treats a wide variety of clinical presentations. In a sample of over 33,000 clients who received LCT in the United States, 86.6% of clients met criteria for clinical improvement by the end of care. Further, this recent research demonstrates that completing digital lessons and exercises and receiving provider feedback messages was associated with greater improvements in client depression and anxiety symptoms, beyond the effect of therapy sessions [14].

In LCT, clients participate in synchronous telehealth therapy sessions and in asynchronous guided practice sessions. The average length of an LCT episode is 6 synchronous therapy sessions over 8 weeks [14]. Consistent with best practices in mental health care, Lyra Care therapists are trained in and receive ongoing supervision and consultation on evidence-based mental health care [15,16]. Synchronous telehealth therapy sessions, in which therapists meet with clients on a secure video platform that is compliant with the Health Insurance Portability and Accountability Act, last for 50 minutes and typically occur weekly or biweekly. During these synchronous therapy sessions, therapists and clients establish goals for treatment, work together on developing and practicing strategies and skills, and monitor progress toward those treatment goals. Therapists also use time in these sessions to introduce the content that they are recommending the client to use during the guided practice sessions. As a flexible provider-led model of blended care, the recommended digital tools are selected in a collaborative way based on the content of the session, the client's goals, case conceptualization, and anticipated challenges during the following week; then the provider gets commitment from the client to do the practice.

Guided practice sessions take place on the digital therapy platform, which clients can access on a web browser or via the Lyra Health app. The "guided" aspect of guided practice sessions refers to therapists guiding clients to asynchronous tools relevant to their care, sending messages in between sessions, and providing asynchronous feedback on the client's progress as the client completes between-session activities. The "practice sessions" aspect refers to the asynchronous skills building and practice that happen when clients engage with these tools between sessions. This content is accessible asynchronously at any time of day or night. These digital tools were developed by Lyra Health based on transdiagnostic treatment approaches, including the Unified Protocol [17], acceptance and commitment therapy [18], and dialectical behavior therapy [19]. After each synchronous therapy session, therapists assign digital tools to their clients to reinforce key concepts and skills related to what is being worked on in treatment. These digital tools include animated video lessons that use storytelling to show characters learning and implementing new skills from therapy, written psychoeducational materials, interactive exercises that prompt reflection and skills practice, symptom assessments, and messaging with therapists. The digital tool library has grown over time, with new video lessons, exercises, and guides added on an ongoing basis. As the client engages with the digital tools, the therapist receives alerts and is able to provide asynchronous support through in-platform commenting and messaging. In a recent analysis of over 33,000 clients who completed LCT, participants viewed a median of 6.0 (IQR 3.0-9.0) digital lessons, completed 6.0 (IQR 2.0-12.0) exercises, and exchanged 16.0 (IQR 9.0-28.0) direct messages with their providers during guided practice sessions [14]. This model supports the practice of skills outside of synchronous telehealth sessions and helps support therapist fidelity to evidence-based practices.

Past research on LCT found that time spent in therapy sessions, time spent viewing digital video lessons, time spent completing exercises, and time spent messaging with the therapist all predict improvements in depression and anxiety symptoms [14,20]. These findings indicate unique statistically significant and clinically significant benefits for clients engaging in both the synchronous telehealth sessions and in the asynchronous guided practice sessions.

However, little is known about the client experience of participating in LCT. For clients to fully engage in mental health services, the services should be optimized to have a strong user experience. Qualitative research on a slightly different model of blended care therapy, which included a greater average number of telehealth sessions (an average of 21.96 sessions relative to our average of 6 sessions) and a different model of between-session practice assignments (consisting of text-based lessons, exercises, and testimonials, with small interactive elements such as audio and video clips), found that clients were generally satisfied with all components of the blended care therapy model but sometimes felt overwhelmed by the digital tools [7]. Identifying the client-perceived value of and challenges associated with specific components of blended care services is essential to understanding and designing more usable, enjoyable, and effective methods of delivering mental health

XSL•FO
RenderX

care. This study explores the client experience of digital tools while in LCT and aims to understand the perceived benefits and challenges of using digital tools while participating in LCT.

Methods

Overview

We conducted a semistructured interview study to explore the client experience of using digital tools while in LCT and to understand our client's perceived benefits and challenges for engaging with these tools.

Participants

To ensure that potential participants had familiarity with LCT digital tools and would be able to reflect on their experiences in LCT, we sent a recruitment email to Lyra clients who had graduated from LCT in the prior 4 months, were at least 18 years of age, and had completed at least 50% of digital assignments (ie, video lessons, exercises, and guides) during their course of therapy. LCT is available to individuals as an employer-sponsored mental health benefit, Lyra Health, offered by Lyra Clinical Associates. Potential participants were invited to complete a short screening survey to confirm their interest in participating and to provide demographic information. Invitations to schedule an interview were sent to participants, along with a copy of the institutional review board–approved consent document.

Ethical Considerations

All study procedures were approved by the Western-Copernicus Group Institutional Review Board (approval number: WCG IRB Protocol #20232618) prior to enrolling participants. Participants were sent a copy of the institutional review board–approved consent document prior to their scheduled interview. At the beginning of the interview appointment time, the interviewer reviewed the consent form, and participants provided verbal consent. After the interviews, participants were compensated for their time with a US $75 gift card. Interviews were recorded for transcription purposes, and the recordings were deleted after transcription was completed. All quotations presented in this manuscript have been deidentified.

Study Procedures

We interviewed a total of 12 participants. Interviews lasted approximately 60 minutes and were conducted over Zoom by the lead author. After verbal consent was obtained, the interview began and was recorded for transcription purposes. During these semistructured interviews, participants were prompted to reflect on their experiences in LCT, including the impact that the digital tools had on their lives and how different types of digital tools may have helped or hindered their progress in psychotherapy. Example questions from the interview guide include "How would you describe the experience of having digital content to work on in between sessions? What was helpful? What was unhelpful?" and "What components of the platform were most helpful to you in making progress in therapy?"

Positionality Statement

The authors all have training as clinical psychologists or physicians and have experience in the design, development, and implementation of technology-enabled mental health services. The majority of the authors have experience in delivering evidence-based psychotherapy to individuals across the lifespan and drew on these professional experiences while designing the study and interpreting the data. Throughout the analysis, the authors took their own positionality and identities into account while analyzing study data.

Analysis

Transcripts were analyzed in Dedoose [21] using a codebook thematic analytic approach [22]. The 6 steps of analysis, outlined by Braun and Clarke [23], were followed, in which the lead author familiarized themself with the data prior to generating a preliminary codebook used to organize the initial analysis of the data. Reflexive practices were used to iteratively develop themes, and the themes were reviewed, defined, and presented in this paper. Analytic decisions were documented throughout the research process. Because interrater reliability rests on the assumption that meaning is fixed within the data, this analytic approach does not recommend using formal interrater reliability metrics [22].

Results

Overview

A total of 12 individuals participated in these interviews. As seen in Table 1, participants held diverse social and cultural identities.

In total, 3 primary themes were generated through thematic analysis. These themes centered around supporting knowledge and skill development, the benefits and challenges associated with the range of digital tools, and the combination of flexibility and accountability promoting positive change. First, we summarize the ways in which guided practice sessions allowed clients to develop knowledge and learn skills related to their mental health and well-being. Then, we describe how the range of digital tools included in LCT presents different benefits and challenges for clients. Finally, we discuss how flexibility inherent in having both synchronous and asynchronous sessions, along with the accountability from a provider, encourages clients to continue to practice skills related to their mental health and well-being.

XSL•FO
RenderX

Table 1. Participant demographics.

Gender	Participants, n
Male	5
Female	6
Pangender	1
Age (years)	
18-24	1
25-34	4
35-44	2
45-54	2
≥55	2
Race or ethnicity	
Black	3
South Asian	2
East Asian	1
Hispanic or Latino	1
White	5
Education	
Vocational school	2
Some college	2
Bachelor's degree	6
Master's degree	2
Household annual income (in US $)	
Less than $20,000	1
$20,000-$34,000	1
$35,000-$49,000	2
$50,000-$74,000	2
$75,000-$99,000	2
More than $100,000	3

The Addition of Asynchronous Guided Practice Sessions Allows for the Development of Both Knowledge and Skills

Gaining knowledge and developing skills are central to making gains in psychotherapy. The combination of synchronous therapy sessions with a mental health provider and asynchronous guided practice sessions was viewed by many participants as essential to their progress in LCT. While many of the participants were familiar with synchronous therapy sessions, either from their own past treatment experiences or general knowledge of what "therapy" typically entails, the concept of having asynchronous guided practice sessions was new to most participants and appeared key in supporting the development of knowledge and skills. Overall, these asynchronous sessions supported clients in using the synchronous sessions more efficiently and often allowed clients and therapists to communicate more effectively about issues related to the care episode.

Participants shared ways in which the asynchronous guided practice sessions were integrated into their lives and how they helped to identify areas for growth and to provide support for both knowledge acquisition and skills practice. Pt 6 reflected:

> *I got a video to learn about the approach with cognitive behavioral therapy with my particular therapist and I was sold on that completely. The tools you're learning you then apply in your daily life. I liked the hybrid learning or the flip learning where you do work outside the classroom so when you come to therapy, you can talk about it. And for me, it's important to go back to them outside of that 45 min because if not, we would waste time during my session.*

This participant had a background in education and drew these favorable parallels to the increasingly popular flipped classroom pedagogy in which active class time is used for working through exercises and discussing concepts that were learned outside of

XSL•FO
RenderX

the classroom. By completing exercises asynchronously, the participant felt that subsequent sessions were more efficient because they could use the time to talk through their progress and any challenges that arose between sessions.

Many participants in the sample had past experiences with receiving mental health care in which knowledge and skill development felt less supported. Pt 5 reflected:

> *Previously, I had another therapist on a different platform. I had an opportunity to meet with her very regularly and while it was helpful, I did not get the same benefits that I think I did from using Lyra because my therapist with Lyra would issue me little assignments, like videos, that often spoke to situations that I was in personally.*

Among participants who reflected on their past experiences in therapy, there were particularly favorable attitudes towards the model of blended care therapy in which asynchronous guided practice sessions supported the clients in making specific treatment gains, often in a more consolidated amount of time. As Pt 12 commented:

> *Whenever I've done therapy without tools like this, it's taken me a couple of iterations to get to the point. With the tools, I get to think faster.*

This notion of "thinking faster," supported by guided practice session digital tools, appears to translate into more active and engaged practice of therapeutic skills and a quicker time to symptom relief.

Participants commented both on how the use of digital tools in guided practice sessions helped them reflect more deeply and on how therapist review of their digital tools assignments allowed for the therapist and client to communicate more effectively about presenting problems and concerns. As Pt 1 reflected:

> *So I ended up being assigned the journal… I started going in and just writing a little more and I didn't realize I hadn't truly opened up to the therapist assigned to me. When I started to journal, she learned more. She was like, "we never talked about this. You've never talked about this in sessions"*

Here, we see that the guided practice sessions allow for clients to practice skills being taught in therapy and also allow for clients to communicate with their therapists in an enhanced way. Through this enhanced communication, therapists can comment and reinforce or explore new topic areas with their clients. This leads to a more deeply shared understanding and to more personalized approaches to structuring subsequent synchronous therapy sessions.

Pt 5 went on to comment:

> *It was probably the most effective thing that I've ever had. I appreciated it a lot like, yes, my therapist was amazing. She did a really great job with helping me a lot, but having those videos along with her support and her guidance, I feel really helped me to make the progress that I have today.*

Multiple participants echoed this belief that the provider-led selection of the digital tools was critical to their treatment success and that the personalization offered the advantage of digital tools being perceived as highly relevant to the participant and their presenting problems and goals for therapy. It appears that digital tools alone are unlikely to support the reported level of knowledge and skills development. Rather, the combination of synchronous telehealth sessions and the digital tools embedded within asynchronous guided practice sessions was found to be of benefit here.

The Range of Available Digital Tools Presents Different Benefits and Challenges for Clients

Overall, digital tools appeared to support the acquisition of knowledge and support the practice of skills, while also helping to build a common language between clients and their therapist. In line with what is known about how different individuals will respond to different treatment components in different ways [24], we observed that the various digital tools embedded in LCT each presented their own benefits and challenges for clients. These benefits and challenges were discussed in the context of having the therapist select digital tools based on intended relevancy to the client. Within this theme, we identify factors that contribute to both the strengths and weaknesses of different types of digital tools within blended care therapy.

Video Lessons

Video lessons (ie, animated videos featuring stories of characters as they face mental health challenges and work with a therapist to learn skills to address those challenges) were the most popular type of digital tool with participants in this study. The majority of participants described favorable ways in which the lesson content was relatable. In addition to videos teaching therapeutic concepts and skills, they served a valuable function in normalizing the experience of a range of mental health concerns. Multiple participants reflected on how video lessons helped deepen their understanding of skills and strategies that were discussed in therapy sessions. For example, Pt 6 noted:

> *The videos allowed you to see glimpses of what was happening in [the character's] life, but then they would use strategies to work through in the moment just like what I was doing with [provider name]. It was not just teaching, but it was also practicing the strategies which was helpful because it creates the framework right of how to actually implement it, think about it and process it, which creates a roadmap.*

Storytelling, as evident in these video lessons, appears to be a vital component for supporting the learning acquisition process.

Similarly, Pt 10 highlighted:

> *I think the videos helped a lot. It puts things on a more personal and a less straight, clinical level… I remember actually watching some of the first videos that were an introduction to depression and anxiety and it walks you through the different effects that it can have on someone and the different ways it can manifest. That part really resonated with me and I really understood it. It made me feel less crazy and*

XSL•FO
RenderX

> *instead of having something wrong, I understood this disconnect between my thoughts and my emotional reactions and how it is something that I can work through. It's absolutely something that other people also deal with and that definitely helped a lot.*

The ability of lessons to normalize common mental health issues appeared to help this participant overcome self-stigma related to his symptoms and become more motivated to make changes.

Providing video lessons without support or guidance may be less useful than when integrated into a blended care therapy model. Pt 5 commented:

> *I appreciated that in tandem with me venting everything under the sun to her, she was also able to assign me things where if I brought up a very particular problem, she's like, "Oh, this sounds a lot like what you're going through. I want you to take a look at this and to make sure that you're actually paying attention."*

This highlighted how the guidance therapists provide around video lessons supports therapy clients in learning new practices that are in line with what is being discussed in therapy sessions. The support of asynchronous learning and skill practice allowed for the client to work toward their therapeutic goals within the context of a supportive relationship with an expert clinician.

However, video lessons presented some challenges. Some participants reported that it was challenging to find time to view the videos, which are typically 8-10 minutes long. Although many participants found that the video lessons fit well into their lives, a small portion of study participants reported that it was challenging to keep up with the video lessons (and other digital tools) due to the time commitment required outside of synchronous therapy sessions. For example, Pt 9 described having some physical health issues emerge and needing to spend significant time and energy navigating the health care system. They reflected that:

> *medical stuff was just tacked on top of it, so remembering to do the check-ins and watch all the videos was derailed and I didn't always keep up with it.*

For some people, dedicating 50 minutes every week or two to synchronous therapy sessions feels like the maximum amount of time that they can dedicate to mental health care.

Participants varied in how much they related to characters presented in the video lessons. For example, a college-aged man reflected that a specific lesson featuring a middle-aged woman character, who was overwhelmed with messes that her husband and child were leaving around the home, was particularly relatable because:

> *this kind of thing is absolutely something that happens in most people's lives whether it's a mom dealing with a mess or whether it's me dealing with roommates and a mess. I definitely think it can create some trouble in thinking clearly and can create some very angry and tense emotions. Having that reminder that I can walk through it using the Lyra platform is very helpful.*

However, a middle-aged woman participant reflected on the same digital lesson as less engaging, noting:

> *I didn't relate to it because I don't have kids. There were other videos featuring kids, a husband, and work stress and my situation is kind of different because I have a fiance, we don't have kids, and my work doesn't necessarily stress me out.*

Having relatable characters featured in the video lessons appears to increase clients' ability to empathize with and learn from the lessons, and further examination of factors that increase and decrease relatability will be fruitful for the continued development of characters and storylines for video lessons.

Digital Exercises

Some participants found it challenging to write about their thoughts, feelings, and experiences in digital exercises, though with encouragement from their providers, they found these exercises helped them feel better. While a few of the participants reflected sentiments such as "writing isn't my thing," participants also described ways in which their therapists were able to tailor the written assignments to meet their needs. For example, Pt 6 reflected:

> *I get caught in needing to write it perfectly… [provider name] made it really clear that it doesn't matter and to make bullet points in a section and we can work with it. She used it as an opportunity to say, "what do you think it's about writing or putting your things down?" So it was good to be able to name it and process it. But it is also attached to some of the challenges.*

This discomfort with writing down thoughts or experiences was echoed by other clients who did not initially find written exercises to be helpful. Pt 4 continued to reflect on this point by noting:

> *When I did write the little bits that I did, I did feel better. It wasn't too pushy. It was just like you said, just a little bit of encouragement.*

For participants like these ones, the guidance and support of the therapist appeared essential for them to be able to engage with, and subsequently benefit from, the written exercises.

Symptom Assessments

While in LCT, clients are prompted to complete symptom assessments prior to each session. This practice is in line with measurement-based care principles and allows both the therapist and the client to track changes in symptoms and adjust treatment as necessary. Some participants found the weekly symptom assessments to be helpful and motivating from the beginning of therapy. For example, participants who had past experiences in therapy without routine symptom assessments, such as Pt 9, stated:

> *The most useful to me was the check-ins to see where I'm at and see where the progress was over time. I*

XSL•FO
RenderX

> *liked checking in to see where I was at and to see progress.*

However, some participants did not initially value these assessments but grew to appreciate them over time. As Pt 6 highlighted:

> *I was answering the same questions over and over again. But then, all of the sudden my responses were starting to change. It felt affirming because at first, I was just producing the same answers and then they changed! It showed progress for myself.*

While symptom assessments felt repetitive and tedious for some, the vast majority of participants appreciated the feedback that symptom assessment provided in terms of progress tracking.

Flexibility and Accountability Encourage Continued Skills Practice

Evidence-based short-term therapy models, such as the Lyra Care model, typically focus on developing and then practicing new skills or strategies to reduce suffering and improve well-being. Clients are prompted to set goals for treatment at the onset of a course of therapy, and the client and therapist work together to support progress toward those goals. As clients make progress on their goals, they work toward graduating or taking a break from therapy. This is marked by a client being ready to stop having regularly scheduled therapy sessions and to manage any remaining symptoms independently or with the support of their social network. In these short-term therapy models, therapy is conceptualized as a time to learn and apply skills and to begin generalizing those skills. When someone graduates from therapy, they are encouraged to continue doing the things that work well for them and to have a plan for how they will address new challenges moving forward. The combination of flexibility and accountability woven into this model of blended care therapy, and supported by the available digital tools, appears to promote continued practice of therapeutic skills.

Through these interviews, we observed that the digital content available through the guided practice sessions provided durable opportunities for reflection and discovery. Digital tools could prompt clients to reflect on their current experiences, identify their current emotions and thoughts, and discover both factors contributing to their current emotional state and strategies that they could use to feel better. Convenience matters when examining how a digital tool can be used while in therapy and how it might be used after graduation. Several participants reported that the written exercises were important for practicing skills and keeping the momentum going in between sessions. As Pt 10 commented:

> *I got a lot more comfortable with taking out my phone when I started having an issue or even just making a mental note of what I was going to put in a thought log or something of that nature. It's extremely helpful as an in the moment tool while also being something that is discrete enough that I could take out my phone any time of the day, anywhere I was, and just seem like I was texting where in reality I was, I was working through whatever stress or anxious thoughts I was having, so that definitely became a very helpful tool.*

Having supportive tools almost always available to use can be key for integrating and applying the skills taught in LCT to diverse situations in one's life.

While a plethora of digital mental health tools are available, many individuals report difficulty in finding high-quality digital mental health tools and are quick to disengage from using such tools [25,26]. Here, we observed that the guidance from and accountability provided by the blended care therapist helped participants learn how specific tools could be useful to and usable for them. Tools were typically introduced during synchronous telehealth sessions, and then follow-up support and recommendations were given during guided practice sessions. Pt 8 noted:

> *my therapist did a good job of marketing the tools to me of this being something you can do if you're feeling like this, and giving me examples of when even would be good times to even use it. So, when I would be experiencing a certain feeling that we discussed in therapy, I would be reminded of my therapist telling me that I should do this when I experience this, and that would lead me to go onto the platform and try an activity.*

Participants also spoke of how automated prompts from the system prompted greater accountability to engage. Pt 5 reflected:

> *During my time with my therapist, I would get, like, little email reminders or text message reminders that were prompting me to make sure that I get my assignments or my weekly health assessments done prior to my next session. I found that really helpful.*

This support received from both the therapist and the automated system prompted clients to experiment with tools during an episode of care and learn how the tools can benefit them, which can then facilitate ongoing habitual practices.

While video lessons were particularly popular for clients while they were in care with their therapist and supported the learning of new skills, written practices were often carried forward after graduating from a course of blended care therapy as a means of continued skill practice. For example, Pt 8 noted:

> *Now, I have a journal and even though it's not necessarily the Lyra exercise itself, it had components that I remember from it. When I was really anxious the other day about some random thing, I just opened up a note on my phone and started challenging my thoughts.*

This participant described one of the primary goals of short-term CBT—that they were able to better recognize their emotional experiences and quickly apply an evidence-based strategy (cognitive restructuring) to proactively manage their experience.

XSL•FO
RenderX

Discussion

Principal Findings

Results of this study support the previously observed therapeutic value of digital video lessons in LCT and provide insights into the unique contributions of different components of the LCT model on therapeutic gains. While perceived time constraints and content preferences can impact engagement with digital tools, overall the digital tools were perceived as carrying significant value while participants were clients in LCT. The combination of synchronous telehealth sessions and asynchronous guided practice sessions both helped clients learn new skills that were personalized to their current needs and helped clients apply these skills in their lives through between-session practice and feedback from therapists.

Asynchronous and synchronous components of this blended care model (ie, LCT) worked together to produce client-reported benefits. The clinically important elements of this treatment model seem to be (1) the therapist selecting relevant skills, (2) the therapist introducing skills in session, (3) the client completing between-session practice, (4) the therapist reviewing and commenting on between-sessions practices, and (5) reflection in the subsequent synchronous session. Guided practice sessions appear to enhance synchronous therapy sessions by helping facilitate ongoing communication between therapist and client and supporting skills acquisition and practice both in and out of sessions. Together, this model appears to allow clients to make clinical progress more quickly than they have seen in traditional models of psychotherapy.

Past quantitative analyses have demonstrated the value of video lesson viewing on clinical outcomes. These qualitative analyses further support the value of video lessons. Storytelling, embedded in these lessons, appears to be important in normalizing a range of experiences. Storytelling in digital mental health interventions is increasingly being recognized as clinically valuable. Narrative-based mental lessons support clients in reflecting on their own patterns and help them to feel a sense of connection to similar others as they apply the concepts to their own lives [27]. Past research has found that satisfaction with client stories embedded in internet-delivered cognitive behavioral therapy to be related to both treatment satisfaction and decreases in anxiety symptom severity [28]. Many clients in this study found the video lessons normalized experiences and that the characters and storylines were relatable, sometimes in surprising ways. However, some clients had more difficulty relating to the characters. Given that relatability appears to support engagement in the guided practice session process, this opens future directions for improving the relatability of digital lesson content. Additional research should be conducted to better understand which characteristics are most important for relatability and how relatability of characters relates to effectiveness of between-session practice. Additional training for therapists on how to frame the between-session practice may also be warranted to better support clients in developing their ability to relate to and engage with the important elements of the lesson.

At its core, LCT is an evidence-based short-term therapy model. While short-term models of therapy have strong research evidence for reducing symptoms of mental ill-health [29,30], not all individuals are initially comfortable with or confident in the approach of short-term therapy. Between-session practice helps clients build structure and self-efficacy in using skills that prepares them for a confident graduation. A large body of literature demonstrates that completing therapy homework relates to more positive treatment outcomes [31,32]. Here, we observed that the structure of skill practice outside of synchronous telehealth sessions allowed clients to figure out what skills work for them and in what situations those skills can be used. Participants in this study reflected on ways in which they figured out how to practice skills in real time, such as on their smartphones or by writing in a notebook. These skill-learning experiences are supported both with guided practice sessions and troubleshooting during synchronous sessions with the therapist. As digital tools were identified as being a helpful component of the LCT process with specific benefits of supporting work toward therapy graduation, additional work should be done to determine how to best support the transition to self-care with digital tools after graduation. Barriers to engaging in between-session practice included both time constraints and content preferences (eg, video length, preferences for reading or listening). Future work should continue to explore how clients would respond to different formats of between-session digital content.

Limitations and Additional Future Directions

The qualitative nature of this study is, in many ways, a strength. Qualitative research allows for deeper understanding of our client experience by honing in on how research participants experienced different components of LCT. However, the methods employed in this study do not support the quantification of experiences [33], and qualitative research typically does not aim to be generalizable across populations. Thus, the experiences detailed in this manuscript are unlikely to be representative of all experiences that one may have in LCT. Historically, research on technology-enabled mental health interventions has been conducted on largely homogenous samples as White individuals, and particularly White young to middle-age adult women, are typically over-represented in research trials [34,35]. While our study sample was diverse in regards to race, ethnicity, gender, age, and income, the participants may not be representative of all clients who engage with blended care services. We intentionally selected a sample of clients who had been engaged with digital content while in therapy (by having completed at least 50% of assignments)—this decision was made to maximize the likelihood that participants would be able to meaningfully reflect on the utility of the different treatment components and the full model of LCT. We cannot report on the exact digital tools each participant in the study used during their course of therapy as their engagement in therapy was separate from their engagement with this qualitative research study. Clients in LCT are not “required” to complete digital content to progress in therapy, and this sample does not represent individuals who did not deeply engage with the digital tools. Additionally, therapy is a dyadic process involving both a client and a therapist, and both sides of this

XSL•FO
RenderX

process, as well as the interactions between the client and therapist, should be examined [36]. This study is reflective only of the client experience of digital tools while in LCT and the client-perceived benefits and challenges of using digital tools while participating in LCT. Future work will examine the therapist's experience of delivering this type of blended care.

Conclusions

The between-session digital content embedded in short-term blended care therapy programs, such as LCT, can keep therapy clients engaged in skill development outside of sessions and can drive therapeutic gains. This model of care has demonstrated clinical effectiveness at scale in past quantitative studies and can support clients in developing mental health and coping skills in a relatively limited treatment period (average of 6 sessions over 8 weeks [1,14]). These qualitative results support the previously observed therapeutic value of digital video lessons in LCT and provide insights into the unique contributions of different components of the LCT model on therapeutic gains. As we continue to work toward increasing access to evidence-based mental health care, blended care therapy programs can be considered as a viable method for meeting client desires to work synchronously with a mental health provider while also offering the benefits of asynchronous care focused on skill development.

Acknowledgments

The study was funded by Lyra Health. The funder of the study was involved in study design, data interpretation, and writing of the report, and was informed during data collection and data analysis. The corresponding author had final responsibility for the decision to submit for publication. The authors declare that there was no use of generative artificial intelligence technology in the generation of texts, figures, or other informational content in this paper.

Data Availability

Given the small dataset, interview data are not publicly available to ensure participants' privacy. For questions about the dataset, contact the corresponding author.

Conflicts of Interest

EGL, MB, CEC, AV, and AL are employed by Lyra Health and Lyra Clinical Associates, receive income from Lyra Health and Lyra Clinical Associates, and have been granted equity in Lyra Health.

References

1. Andrade LH, Alonso J, Mneimneh Z, Wells JE, Al-Hamzawi A, Borges G, et al. Barriers to mental health treatment: results from the WHO world mental health surveys. Psychol Med 2014;44(6):1303-1317 [FREE Full text] [doi: 10.1017/S0033291713001943] [Medline: 23931656]
2. Wentzel J, van der Vaart R, Bohlmeijer ET, van Gemert-Pijnen JEWC. Mixing online and face-to-face therapy: how to benefit from blended care in mental health care. JMIR Ment Health 2016;3(1):e9 [FREE Full text] [doi: 10.2196/mental.4534] [Medline: 26860537]
3. Kemmeren L, van Schaik A, Draisma S, Kleiboer A, Riper H, Smit J. Effectiveness of blended cognitive behavioral therapy versus treatment as usual for depression in routine specialized mental healthcare: E-COMPARED trial in the Netherlands. Cogn Ther Res 2023 Mar 24;47(3):386-398. [doi: 10.1007/s10608-023-10363-y]
4. Kay-Lambkin F, Baker A, Lewin T, Carr V. Acceptability of a clinician-assisted computerized psychological intervention for comorbid mental health and substance use problems: treatment adherence data from a randomized controlled trial. J Med Internet Res 2011;13(1):e11 [FREE Full text] [doi: 10.2196/jmir.1522] [Medline: 21273184]
5. Høifødt RS, Lillevoll KR, Griffiths KM, Wilsgaard T, Eisemann M, Waterloo K, et al. The clinical effectiveness of web-based cognitive behavioral therapy with face-to-face therapist support for depressed primary care patients: randomized controlled trial. J Med Internet Res 2013;15(8):e153 [FREE Full text] [doi: 10.2196/jmir.2714] [Medline: 23916965]
6. Pier C, Austin DW, Klein B, Mitchell J, Schattner P, Ciechomski L, et al. A controlled trial of internet-based cognitive-behavioural therapy for panic disorder with face-to-face support from a general practitioner or email support from a psychologist. Ment Health Fam Med 2008;5(1):29-39 [FREE Full text] [Medline: 22477844]
7. Etzelmueller A, Radkovsky A, Hannig W, Berking M, Ebert DD. Patient's experience with blended video- and internet based cognitive behavioural therapy service in routine care. Internet Interv 2018;12:165-175 [FREE Full text] [doi: 10.1016/j.invent.2018.01.003] [Medline: 30135780]
8. Lungu A, Jun JJ, Azarmanesh O, Leykin Y, Chen CE. Blended care-cognitive behavioral therapy for depression and anxiety in real-world settings: pragmatic retrospective study. J Med Internet Res 2020;22(7):e18723 [FREE Full text] [doi: 10.2196/18723] [Medline: 32628120]
9. Andrade LH, Alonso JE, Mneimneh Z, Wells J, Al-Hamzawi A, Borges G, et al. Barriers to mental health treatment: results from the WHO World Mental Health surveys. Psychol Med 2014;44(6):1303-1317 [FREE Full text] [doi: 10.1017/S0033291713001943] [Medline: 23931656]

XSL•FO
RenderX

10. Mohr DC, Hart SL, Howard I, Julian L, Vella L, Catledge C, et al. Barriers to psychotherapy among depressed and nondepressed primary care patients. Ann Behav Med 2006;32(3):254-258. [doi: 10.1207/s15324796abm3203_12] [Medline: 17107299]
11. Borghouts J, Eikey E, Mark G, De Leon C, Schueller SM, Schneider M, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. J Med Internet Res 2021;23(3):e24387 [FREE Full text] [doi: 10.2196/24387] [Medline: 33759801]
12. Moskalenko MY, Hadjistavropoulos HD, Katapally TR. Barriers to patient interest in internet-based cognitive behavioral therapy: informing e-health policies through quantitative analysis. Health Policy Technol 2020;9(2):139-145. [doi: 10.1016/j.hlpt.2020.04.004]
13. Owusu JT, Wang P, Wickham R, Varra AA, Chen C, Lungu A. Real-world evaluation of a large-scale blended care-cognitive behavioral therapy program for symptoms of anxiety and depression. Telemed J E Health 2022;28(10):1412-1420 [FREE Full text] [doi: 10.1089/tmj.2021.0590] [Medline: 35263185]
14. Espel-Huynh HM, Wang L, Lattie EG, Wickham RE, Varra A, Chen CE, et al. Clinical effects of asynchronous provider-guided practice sessions during blended care therapy for anxiety and depression: pragmatic retrospective cohort study. J Med Internet Res 2024;26:e60502 [FREE Full text] [doi: 10.2196/60502] [Medline: 39422996]
15. Borders LD. Best practices in clinical supervision: another step in delineating effective supervision practice. Am J Psychother 2014;68(2):151-162. [doi: 10.1176/appi.psychotherapy.2014.68.2.151] [Medline: 25122982]
16. Watkins CE. Psychotherapy supervision developments and innovations for the new millennium: contributions from the cutting edge. Am J Psychother 2014;68(2):141-149. [doi: 10.1176/appi.psychotherapy.2014.68.2.141] [Medline: 25122981]
17. Barlow D, Farchione T, Sauer-Zavala S, Murray-Latin H, Bullis J, Ellard K. Unified Protocol for Transdiagnostic Treatment of Emotional Disorders: Therapist Guide. New York, NY: Oxford University Press; 2017.
18. Hayes S, Strosahl K, Wilson K. Acceptance and Commitment Therapy: The Process and Practice of Mindful Change. New York, NY: Guilford Press; 2011.
19. Linehan MM, Korslund KE, Harned MS, Gallop RJ, Lungu A, Neacsiu AD, et al. Dialectical behavior therapy for high suicide risk in individuals with borderline personality disorder: a randomized clinical trial and component analysis. JAMA Psychiatry 2015;72(5):475-482. [doi: 10.1001/jamapsychiatry.2014.3039] [Medline: 25806661]
20. Lungu A, Wickham RE, Chen S, Jun JJ, Leykin Y, Chen CEJ. Component analysis of a synchronous and asynchronous blended care CBT intervention for symptoms of depression and anxiety: pragmatic retrospective study. Internet Interv 2022;28:100536 [FREE Full text] [doi: 10.1016/j.invent.2022.100536] [Medline: 35433277]
21. Dedoose. 2020. URL: https://www.dedoose.com/ [accessed 2025-03-08]
22. Braun V, Clarke V. Reflecting on reflexive thematic analysis. Qual Res Sport Exerc Health 2019;11(1):1-9. [doi: 10.1080/2159676x.2019.1628806]
23. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
24. Lutz W, Gómez PJ, Prinz J, Rubel J, Moggia D. Do certain patients respond better to specific forms of psychotherapy? In: APA Handbook of Psychotherapy: Evidence-Based Practice, Practice-Based Evidence, and Contextual Participant-Driven Practice. Washington, DC: American Psychological Association; 2024:63-81.
25. Baumel A, Muench F, Edan S, Kane JM. Objective user engagement with mental health apps: systematic search and panel-based usage analysis. J Med Internet Res 2019;21(9):e14567 [FREE Full text] [doi: 10.2196/14567] [Medline: 31573916]
26. Torous J, Nicholas J, Larsen ME, Firth J, Christensen H. Clinical review of user engagement with mental health smartphone apps: evidence, theory and improvements. Evid Based Ment Health 2018;21(3):116-119. [doi: 10.1136/eb-2018-102891] [Medline: 29871870]
27. Bhattacharjee A, Williams JJ, Chou K, Tomlinson J, Meyerhoff J, Mariakakis A, et al. "I Kind of Bounce off It": translating mental health principles into real life through story-based text messages. Proc ACM Hum Comput Interact 2022;6(CSCW2):398 [FREE Full text] [doi: 10.1145/3555123] [Medline: 36816014]
28. Hadjistavropoulos HD, Hill TG, Philipp-Muller AE, Dear B, Titov N. Perceptions of client stories in internet-delivered cognitive behaviour therapy: a mixed-methods evaluation. Internet Interv 2023;34:100692. [doi: 10.1016/j.invent.2023.100692]
29. Robinson L, Kellett S, Delgadillo J. Dose-response patterns in low and high intensity cognitive behavioral therapy for common mental health problems. Depress Anxiety 2020;37(3):285-294. [doi: 10.1002/da.22999] [Medline: 32027435]
30. Driessen E, Hegelmaier LM, Abbass AA, Barber JP, Dekker JJ, Van HL, et al. The efficacy of short-term psychodynamic psychotherapy for depression: a meta-analysis update. Clin Psychol Rev 2015;42:1-15. [doi: 10.1016/j.cpr.2015.07.004] [Medline: 26281018]
31. Lebeau RT, Davies CD, Culver NC, Craske MG. Homework compliance counts in cognitive-behavioral therapy. Cogn Behav Ther 2013;42(3):171-179. [doi: 10.1080/16506073.2013.763286] [Medline: 23419077]
32. Kazantzis N, Whittington C, Zelencich L, Kyrios M, Norton PJ, Hofmann SG. Quantity and quality of homework compliance: a meta-analysis of relations with outcome in cognitive behavior therapy. Behav Ther 2016;47(5):755-772. [doi: 10.1016/j.beth.2016.05.002] [Medline: 27816086]

XSL•FO
RenderX

33. Sale JEM, Lohfeld LH, Brazil K. Revisiting the quantitative-qualitative debate: implications for mixed-methods research. Qual Quant 2002;36(1):43-53 [FREE Full text] [doi: 10.1023/A:1014301607592] [Medline: 26523073]
34. Titov N, Dear B, Nielssen O, Staples L, Hadjistavropoulos H, Nugent M, et al. ICBT in routine care: a descriptive analysis of successful clinics in five countries. Internet Interv 2018;13:108-115 [FREE Full text] [doi: 10.1016/j.invent.2018.07.006] [Medline: 30206525]
35. Alegría M, Alvarez K, Ishikawa RZ, DiMarzio K, McPeck S. Removing obstacles to eliminating racial and ethnic disparities in behavioral health care. Health Aff (Millwood) 2016;35(6):991-999 [FREE Full text] [doi: 10.1377/hlthaff.2016.0029] [Medline: 27269014]
36. Atzil-Slonim D, Tschacher W. Dynamic dyadic processes in psychotherapy: introduction to a special section. Psychother Res 2020;30(5):555-557. [doi: 10.1080/10503307.2019.1662509] [Medline: 31476979]

Abbreviations

LCT: Lyra Care Therapy

Edited by A Mavragani; submitted 04.11.24; peer-reviewed by JC Perez; comments to author 06.12.24; revised version received 14.01.25; accepted 04.03.25; published 02.04.25.

Please cite as:
Lattie EG, Beltzer M, Varra A, Chen CE, Lungu A
Examining the Client Experience of Digital Tools in Blended Care Therapy: Qualitative Interview Study
JMIR Form Res 2025;9:e68249
URL: https://formative.jmir.org/2025/1/e68249
doi:10.2196/68249
PMID:

©Emily G Lattie, Miranda Beltzer, Alethea Varra, Connie E Chen, Anita Lungu. Originally published in JMIR Formative Research (https://formative.jmir.org), 02.04.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Remote Intervention Based on mHealth and Community Health Workers for Antiretroviral Therapy Adherence in People With HIV: Pilot Randomized Controlled Trial

Shivesh Shourya[1], BS; Jianfang Liu[2], BE, MAS, MPA, PhD; Sophia McInerney[2], BS; Trinity Casimir[2], BA; James Kenniff[2], BS; Trace Kershaw[1], PhD; David Batey[3], MSW, PhD; Rebecca Schnall[2], MPH, PhD

[1]Department of Social and Behavioral Sciences, Yale School of Public Health, New Haven, CT, United States

[2]Columbia University School of Nursing, New York, NY, United States

[3]Tulane University School of Social Work, New Orleans, LA, United States

Corresponding Author:
Shivesh Shourya, BS
Department of Social and Behavioral Sciences
Yale School of Public Health
60 College Street
New Haven, CT, 06510
United States
Phone: 1 6503346176
Email: shivesh.shourya@yale.edu

Abstract

Background: Despite the availability of antiretroviral therapy (ART), only 66% of people with HIV in the United States achieve viral suppression, largely due to suboptimal ART adherence. Barriers such as limited access to care and forgetfulness impact adherence rates, which must be maintained at ≥95% to prevent viral load rebound. Combination interventions leveraging community health worker (CHW) support and mobile health (mHealth) technologies have the potential to overcome previously identified barriers and provide cost-effective support for improving adherence and viral suppression outcomes in people with HIV.

Objective: This pilot study aimed to evaluate the feasibility, acceptability, and preliminary efficacy of remote delivery of the Community Health Worker and mHealth to Improve Viral Suppression (CHAMPS) intervention, combining the WiseApp, CHW support, and the CleverCap smart pill bottle. A secondary aim was to gather participants' feedback on the usability of the app and pill bottle as well as to better understand their experiences with remote study procedures.

Methods: This mixed methods pilot study involved 40 participants with HIV, who were randomly assigned to a control group (n=20, 50%) or the CHAMPS intervention (n=20, 50%) over 3 months. The intervention group participated in up to 12 sessions with CHWs and used the WiseApp, paired with a CleverCap smart pill bottle, to support ART adherence. Remote baseline and follow-up visits were conducted via Zoom and included surveys measuring adherence, self-efficacy, and usability (measured by Health Information Technology Usability Evaluation Scale [Health-ITUES] and Poststudy System Usability Questionnaire [PSSUQ]). Semistructured interviews explored participants' experiences with the intervention. Thematic analysis was used to identify key facilitators and barriers based on the Mobile Health Technology Acceptance Model.

Results: Remote delivery of the CHAMPS intervention was feasible, with high usability ratings for both the WiseApp and CleverCap (overall scores on Health-ITUES: mean 4.35, SD 0.58 and PSSUQ: mean 2.04, SD 1.03). In the intervention group, there were nonsignificant improvements in self-reported adherence scores (P=.29) and in self-efficacy scores (P=.07). The adjusted odds ratio for achieving undetectable viral load in the intervention group compared to the control group was 3.01 (95% CI –1.59 to 4.12), indicating a medium effect size in favor of the intervention. Overall study retention was 75% (30/40), with higher retention in the control group. Participants valued the flexibility of remote study procedures, particularly Zoom-based study visits and mailed blood sample kits. Qualitative feedback highlighted the intervention's acceptability and ability to overcome logistical barriers.

Conclusions: The remote CHAMPS pilot study demonstrated the feasibility and acceptability of combining mHealth tools with CHW support to promote medication adherence among people with HIV. While further optimization is needed to enhance its impact, this intervention shows potential for improving health outcomes in diverse underserved populations.

XSL•FO
RenderX

Trial Registration: ClinicalTrials.gov NCT05938413; https://clinicaltrials.gov/study/NCT05938413

(JMIR Form Res 2025;9:e67997) doi:10.2196/67997

KEYWORDS

HIV; antiretroviral therapy adherence; ART; ART adherence; mobile health technology; mHealth technology; community health workers; CHWs; smart pill bottle; feasibility evaluation; mobile health technology acceptance model

Introduction

Background

With the advent of antiretroviral therapy (ART) for HIV, people with HIV can now achieve near-normal life expectancies [1-4]. At the end of 2021, >50% of people with HIV in the United States were aged ≥50 years, primarily due to the effectiveness of ART [5]. However, the rates of viral suppression among the 1.2 million people with HIV in the United States continue to remain low, with current estimates of approximately 66% among people with HIV, despite the wider availability of ART [5]. The suboptimal rates of viral suppression may be attributed to poor ART adherence [6] and occur during a period of increasing health care fragmentation, leading to higher costs of care and lower health outcomes for many patients with chronic illnesses, such as people with HIV [7].

The Ending the HIV Epidemic plan identifies viral suppression as a cornerstone to preventing and treating HIV infections [8-10]. HIV viral suppression is dependent on adherence to ART, and studies demonstrate that high rates of ART adherence (approximately 95%) are attributed to higher rates of viral suppression (approximately 78%) [11]. However, even moderate decreases in adherence to ART (from 95% to 80%) can lead to drastic reductions in viral suppression rates (from 78% to 20%), underscoring that ART adherence is maintained at ≥95% [11]. Poor adherence to ART can lead to poor outcomes for the patient, the emergence of drug-resistant HIV strains [12], and increased rates of HIV transmission [13]. Suboptimal levels of ART adherence, along with low levels of engagement in the HIV care continuum, have also been attributed to the progression of HIV disease and premature deaths among people with HIV [14]. Therefore, there is an urgent need to develop and evaluate interventions that enhance adherence to ART.

Developing interventions to increase ART adherence must address prevalent barriers to optimal ART adherence by championing a comprehensive approach to understanding and addressing the unmet needs of people with HIV. Previous studies have identified numerous barriers to sustained adherence to ART; engagement with medical care; and, consequently, viral suppression, including HIV stigma or fear of HIV status disclosure [15-19], geographic barriers or limited time resources for receiving or accessing HIV care [16,17,19], negative experiences within medical institutions or HIV care centers [15,16,19], and forgetfulness [18,19]. Similar studies have outlined strong networks of social support, HIV case managers, telemedicine appointments, and reminder tools as facilitators of ART adherence and medical care engagement [16-19]. Recently, there has been an increased focus on the utility of community health workers (CHWs), who can take on roles as outreach workers, patient navigators, health advisers, or peer leaders in helping patients access and manage HIV in primary care settings [20]. While the profile, skill level, and job scope of CHWs can vary widely based on context, they often serve as patient navigators and peer educators in the United States, focusing on HIV care retention, ART adherence, and viral suppression support, particularly in underresourced communities [21-23]. Several studies have found that integrating CHWs was associated with improved care retention, ART adherence, and viral suppression outcomes [20,24,25]; nonetheless, other studies have found no significant difference in viral suppression outcomes between people with HIV with assigned CHWs and those without CHW interventionists [26,27]. While some studies failed to show statistical significance in viral suppression outcomes between intervention and control groups, they emphasized the utility of a combined intervention, including CHW components and other interventions, to address barriers to ART adherence [26,27]. Mobile health (mHealth) interventions offer themselves as novel interventions in HIV care settings, given their ubiquity, ability to overcome geographic barriers to care, and low associated costs [28]. However, previous studies have provided mixed results, with some finding increased viral suppression and retention in care rates among mHealth intervention participants [29,30] and others finding no statistically significant difference in viral suppression and retention rates between intervention and control participants [31]. Previous study limitations include poor integration of mHealth technologies within HIV primary care settings, and researchers have advocated for a combination approach with patient navigation services to help prioritize human communications and achieve desired outcomes related to viral suppression [31]. Thus, interventions that combine the use of CHWs and mHealth technologies hold promise for increasing rates of ART adherence and viral suppression within the United States [24,32].

Objectives

In this pilot study of the Community Health Worker and Mobile Health to Improve Viral Suppression (CHAMPS) remote intervention, conducted by our team at the Columbia University School of Nursing, we tested the feasibility of a remotely delivered CHW intervention through the mHealth app WiseApp. Participants were recruited from across the United States and provided with a smart pill bottle, CleverCap, which interfaced with the WiseApp for personalized medication reminders [33,34]. We also gathered feedback on the acceptability of remotely conducted study procedures.

XSL•FO
RenderX

Methods

Study Design

Overview

The pilot study was a 2-arm randomized controlled trial among 40 people with HIV who were followed over a 3-month period. Participants were randomly assigned to the CHAMPS intervention (20/40, 50%) or the standard of care control (20/40, 50%) arm. Recruitment for the study occurred over a 7-month period, beginning in July 2023 and closing at the end of January 2024. The study was initiated on July 26, 2023, with the enrollment of the first participant, and concluded in May 2024, with the final participant completing their last study visit on May 7, 2024.

Intervention Arm and Description of the Planned Intervention

The CHAMPS intervention was a 3-month intervention guided by the study team's previous work on CHW and mHealth interventions. Specifically, the Birmingham Access to Care study (NCT03205982) adapted the Anti-Retroviral Treatment and Access to Services intervention, an evidence-based intervention designated by the Centers for Disease Control and Prevention, to support re-engagement in care for people with HIV who had dropped out of care [34,35]. This approach emphasized strength-based case management and motivational interviewing, fostering a close, supportive relationship between CHWs and participants, and was used to design the content structure of the CHW sessions. In addition, the WiseApp study leveraged end user feedback to develop a self-management app for people with HIV, incorporating features such as medication trackers, push notification reminders, and linkage to the CleverCap smart pill bottle [34,36].

Participants in the intervention arm were assigned a CHW at the end of their baseline visit. CHWs were study team members trained on the intervention, including the content of each session, motivational interviewing, strength-based case management, Anti-Retroviral Treatment and Access to Services [37], HIV and substance use, the WiseApp and associated mHealth technology, and field safety. CHWs administered at least 10 or up to 12 individual sessions to the participants throughout the intervention. Table 1 provides an outline and description of CHW session contents and structure. All sessions were conducted remotely via secure Zoom (Zoom Video Communications, Inc) calls, in compliance with the Health Insurance Portability and Accountability Act (HIPAA); phone calls; or the chat feature of the WiseApp, depending on the participant's preference.

XSL•FO
RenderX

Table 1. Structure and content of the community health worker (CHW) sessions delivered to participants in the Community Health Worker and Mobile Health to Improve Viral Suppression (CHAMPS) intervention arm.

Session number and title	Session content	Window for session
(1) Building the relationship	Introduce the goals of the CHAMPS intervention to the participant (approximately 15 to 20 minutes).	At baseline visit or within 1 to 2 days of baseline visit
(2) Introduction to the WiseApp	Discuss how the WiseApp can be used to facilitate communication between the CHW and the participant. Review the medication tracking function and how this can be used by the CHW and the participant (approximately 30 to 45 minutes).	At baseline visit or within 1 to 2 days of baseline visit
(3) Emphasizing personal strengths	Help the participant self-identify personal strengths, abilities, and skills. Check in on any technical issues and study logistics with the participant (approximately 30 to 45 minutes).	Week 1
(4) Learning to make contact	Assist the participant in preparing a list of questions to ask their care provider. Check in on any technical issues and study logistics with the participant (approximately 30 to 45 minutes).	Week 2
(5) Check-in call 1	Check in with the participant on their upcoming appointments and any needed documents; address any potential barriers to care through offering local resources. Check in on any technical issues and study logistics with the participant (approximately 15 to 20 minutes).	Week 3
(6) Primary care provider appointment 1	Support the participant's efforts during their primary care provider remote visit, on the basis of the participant's comfort level (approximately 15 to 20 minutes).	On the basis of any upcoming appointments
(7) Debriefing the provider visit 1	Solicit the participant's input on what went well during their primary care provider visit. Elicit from the participant what was learned from the care visit and what strengths they demonstrated. Check in on any technical issues and study logistics with the participant (approximately 15 to 20 minutes).	At the time of or within 1 to 2 days of session 6
(8) Reviewing progress	Plan for and create an action plan for the transition process as the participant nears the end of the study period and prepares to engage with standard-of-care case manager and related clinical professionals (eg, social worker). Check in with the participant on their upcoming appointments and any needed documents. Check in on any technical issues and study logistics with the participant (approximately 30 to 45 minutes).	Anytime during weeks 4 to 7
(9) Check in call 2	Check in with the participant on their upcoming appointments and any needed documents; address any potential barriers to care through offering local resources. Check in on any technical issues and study logistics with the participant (approximately 15 to 20 minutes).	Anytime during weeks 8 to 10
(10) Completing the work	Review the transition process action plan with the participant. Begin transition to standard-of-care case manager and related clinical professionals (eg, social worker) (approximately 30 to 45 minutes).	Anytime during weeks 11 to 12
(11) Optional check in session 1	Check in on any technical issues and study logistics with the participant (approximately 15 to 20 minutes).	Anytime during the study and as needed by the participant
(12) Optional check in session 2	Check in on any technical issues and study logistics with the participant (approximately 15 to 20 minutes).	Anytime during the study and as needed by the participant

Intervention participants received 1 daily reminder through CleverCap's programmed alarm to take their medication at their chosen time. Additional alarms could be programmed by participants, or mobile or app alerts could be set for missed or off-schedule doses, in which case, participants received >1 daily notification. Furthermore, participants were notified if they reopened the pill bottle after an alarm had already been triggered (flagged as an off-schedule dose) or if the cap was improperly closed. However, no reminders were sent, and no alarms were triggered if the CleverCap pill bottle lost power, was deactivated by the study team, was damaged, or lacked a set dosing schedule.

Figure 1 illustrates the WiseApp user interface, with screenshots of a demonstration account. The main menu (Figure 1A) allows navigation to various tabs. The dashboard (Figure 1B) provides an overview of medication adherence, including statuses such as taken (green), missed (red), off-schedule dose (yellow), and improperly closed cap (orange). It also displays adherence feedback through emoji indicators (eg, high adherence: green smiley face). The chat interface (Figure 1C) enables communication with CHWs, while the My Stats tab (Figure 1D) presents a percentage breakdown of adherence metrics, dose timing, and adherence scores. The Videos and Information tab (Figure 1E) includes testimonial videos with adherence tips. The My Alerts tab (Figure 1F) allows participants and CHWs to set reminders for missed or off-schedule doses. Finally, the My Meds tab (Figure 1G) enables participants and CHWs to manage medication details and adjust dosing schedules.

XSL•FO
RenderX

Figure 1. User interface of WiseApp showcasing demo account data: (A) main menu, (B) dashboard with medication adherence status, (C) chat interface, (D) My Stats tab, (E) Videos and Information tab, (F) My Alerts tab, and (G) My Meds tab and Edit Schedule interface.

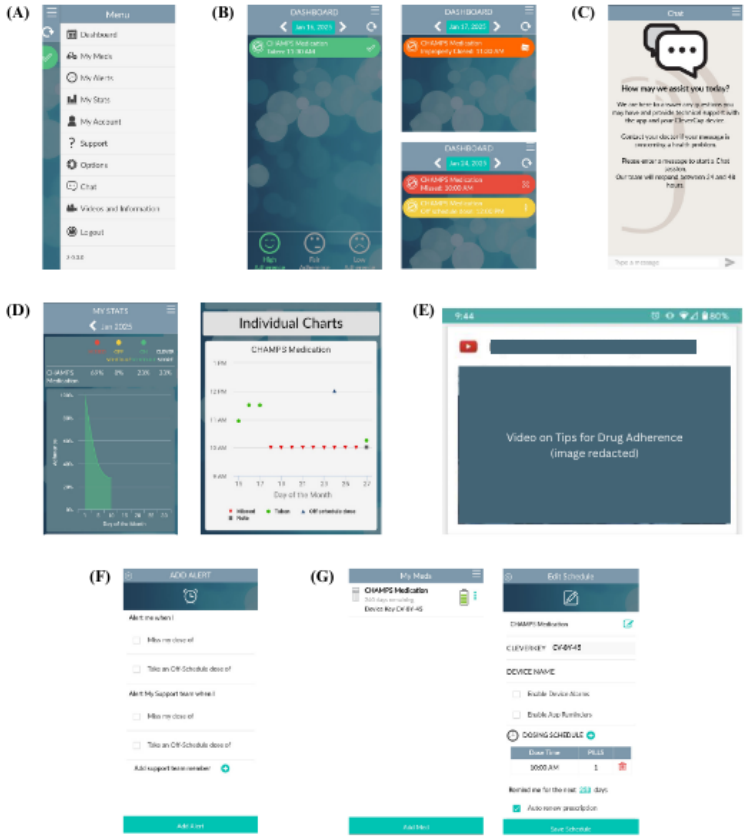

The CleverCap pill bottle (Figure 2) tracked medication adherence by recording when the bottle was opened, with a dose marked as "taken" only if the cap remained off for at least 5 seconds. Figure 2A shows the CleverCap and its packaging, which includes a micro-USB charger. Figure 2B demonstrates CleverCap's built-in reminder system, which provided a visual and auditory alarm when it was time to take a scheduled dose. The bottle flashed green lights for 2 minutes and emitted a loud alarm to ensure that the reminder was noticeable. These cues automatically ceased once the cap was unscrewed, signaling a recorded dose. In addition, if no action was taken within the 2-minute reminder window, the visual and auditory cues stopped, and the dose was recorded as "missed." No visual or auditory cues were provided for off-schedule doses. If the alarm failed to activate due to incorrect or missing programmed alerts on the WiseApp or due to the pill bottle losing power, participants may have missed their medication without documentation in the app.

Figure 2. The CleverCap pill bottle used for medication adherence monitoring: (A) device and packaging and (B) active CleverCap alarm for dosing reminder.

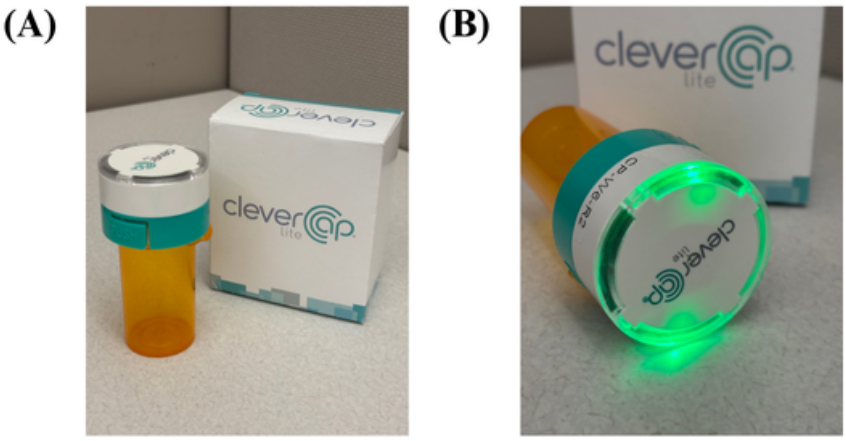

XSL•FO
RenderX

Standard of Care Control Arm

The control arm did not receive any additional interventions beyond the usual standard of care, which consisted of routine clinical services as needed, including referrals to mental health or other ancillary services if clinically indicated.

Recruitment and Eligibility

Participants were eligible to participate if they (1) were able to speak, read, and write in English; (2) were aged ≥18 years; (3) were willing to provide a valid form of identification for verification; (4) were willing to participate in any assigned arm of the intervention; (5) were diagnosed with HIV ≥6 months ago; (6) had an HIV-1 RNA level of >50 copies/ml or self-reported being virally unsuppressed in the past 12 months (defined as a viral load of ≥200 copies/ml in alignment with the clinical definition of viral suppression in the United States); (7) were owners of a smartphone; (8) were capable and willing to provide informed consent for study participation and consent for access to medical records; and (9) lived in the United States. The use of different viral load thresholds aligns with established clinical guidelines and reflects advances in monitoring technology. A threshold of ≥200 copies/ml is used to determine whether someone is detectable for HIV, as it accounts for small fluctuations in viral levels and testing variations, providing a practical measure of treatment effectiveness [38]. However, with improvements in viral load assay sensitivity, a threshold of 50 copies/ml is increasingly used to identify low levels of HIV in the blood, as achieving a viral load below this threshold is considered optimal for reducing transmission risk and improving long-term health outcomes [39]. Viral load data were obtained through one of the following methods: (1) participant-uploaded viral load test results from their electronic health record or laboratory reports, (2) dried blood spot (DBS) collection analyzed by a certified laboratory, or (3) retrieval through electronic health record access via a HIPAA electronic release of information (eROI) authorization). To ensure consistency in timing, only uploaded viral load results dated within 4 to 6 weeks before the study visit were accepted. DBS collection was offered to all participants as an alternative to laboratory-based testing. Participants were not eligible if they met any exclusion criteria, including (1) residing in a nursing home, prison, or receiving in-patient psychiatric care at the time of enrollment; (2) terminal illness with a life expectancy <3 months; (3) planning to move out of the United States in the next 3 months; or (4) participating in a study that targets viral suppression for people with HIV. The study team carefully considered including newly diagnosed participants (diagnosed <6 months) and chose not to include this subset of people with HIV because newly diagnosed individuals are often treatment naive; thus, their adherence behaviors are unknown and may change frequently as they begin their ART regimen.

Participants were recruited via online advertisements posted on POZ.com, Craigslist, as well as Facebook and Instagram (Meta Platforms, Inc). In addition, the study team emailed community-based organizations serving people with HIV across the country with requests to hang study flyers. To determine preliminary eligibility, participants either filled out a web-based screener or were screened by phone. Participants who filled out web-based screeners were followed up with phone calls and were given detailed instructions regarding the study before scheduling study visits. To finalize eligibility, participants were requested to provide confirmation documentation of viral load either via filling out an eROI, by uploading results obtained within the past 4 to 6 weeks in their electronic health record to a secure study platform, or by participating in a screening visit where trained study staff instructed participants on how to self-collect blood samples via a finger-prick method.

Study Visits and Visit Structure

Overview

The study consisted of 2 required visits: a baseline visit and a 3-month follow-up visit. The study also consisted of an optional screening visit for participants who could not provide confirmation of viral load for eligibility. All study visits were conducted using secure HIPAA-compliant Zoom technology. At the start of each study visit, participants were required to show a valid form of identification for verification and to avert fraud.

Screening Visit

Screening visits lasted approximately 1 hour. Participants provided informed consent electronically through the REDCap (Research Electronic Data Capture; Vanderbilt University) system, a secure platform for data collection [40,41]. Before consenting, participants had the opportunity to ask study staff any questions about the process. Before the screening visit, participants were sent a DBS collection kit along with detailed instructions. Trained study staff guided participants through the DBS collection process, which included the use of a Becton Dickinson Microtainer contact-activated lancet blade (1.5 mm × 2.0 mm; blue, high flow) to collect the sample. Care instructions for the puncture site were provided to ensure proper aftercare. Participants were instructed on how to package and ship their samples back to the laboratory at the Miriam Hospital in Rhode Island for processing. Upon receipt of laboratory results, study staff followed up with the participants to inform them of their eligibility for the study. Compensation was not provided for screening visits.

Baseline Visit

Baseline visits lasted 2 hours. Participants provided e-consent through REDCap, with the opportunity to ask study staff any questions before participating in the baseline visit. Upon enrollment, participants were randomized to either the intervention or control arm through a randomization module in REDCap. The randomization module was generated by the study team's data manager, ensuring an unbiased allocation process, while study staff, who were separate from the data manager and did not have access to the randomization module, were responsible for assigning participants to their respective study groups. Intervention participants were informed that they would be sent a CleverCap bottle and would have up to 12 study sessions over the duration of the study with an assigned CHW. Participants were required to complete a comprehensive survey via Qualtrics (Qualtrics International Inc). The survey gathered information on demographics, a single-item ART adherence self-reported scale item (SRSI) [42], and self-efficacy as a

XSL•FO
RenderX

mediator of HIV treatment adherence (HIV Treatment Adherence Self-Efficacy Scale [HIV-ASES]) [43]. Baseline viral load and CD4 counts were recorded from information obtained via eROI, participant self-uploaded laboratory results, or the screening visit. Participants were compensated US $40 in the form of an Amazon gift code for their time.

3-Month Follow-Up Visit

Follow-up visits lasted 1 hour. Participants completed a similar survey as the one administered at baseline with the addition of measures for usability displayed to intervention participants, including the Health Information Technology Usability Evaluation Scale (Health-ITUES) [44] and the Poststudy System Usability Questionnaire (PSSUQ) [45]. Participants also provided viral load and CD4 counts obtained via eROI, participant self-uploaded laboratory results, or DBS collection. Real-time adherence data from the CleverCap pill bottles were downloaded from the CleverCap website. At the end of the follow-up visit, all participants were given the option to participate in an optional qualitative interview regarding their experience in the study and the perceived utility of the remotely delivered CHAMPS intervention. Follow-up interviews were conducted exclusively using secure HIPAA-compliant Zoom technology and were audio recorded for transcription. Consent for follow-up interviews was included in the study consent offered to participants at the baseline visit. Participants were compensated US $50 for their time in the follow-up visit and an additional US $35 if they opted to participate in the interview. Compensation was provided in the form of Amazon gift codes.

Quantitative Data Analysis

Descriptive statistics for demographic variables were reported as counts and percentages. Statistical significance for differences between the control and intervention groups was assessed using the Mann-Whitney *U* test for continuous variables, such as age, and the chi-square test for the remaining categorical variables. A 2-sample *z* test for equality of proportions, with a continuity correction, was used to assess whether the differences in retention rates between the intervention and control groups were statistically significant.

The primary outcome, viral load (virally unsuppressed vs suppressed), was analyzed using a 2-way frequency table by time. The adjusted odds ratio (aOR), adjusting for baseline viral load status, was calculated for the CHAMPS group (intervention vs control).

Self-rated adherence scores using the SRSI [42], a 6-point Likert scale ranging from 1 (“very poor”) to 6 (“excellent”), were reported as mean and SD. Independent 2-tailed *t* tests were used to compare self-rated adherence scores between the intervention and control groups, while paired 2-tailed *t* tests were used to compare scores within groups over time. Missing data for the paired 2-tailed *t* test were handled under the assumption that they was missing completely at random, as participants were lost to follow-up due to external factors unrelated to study data collection.

For adherence data collected through CleverCap, the average percentage of doses taken and missed was calculated, with SD reported for the participants in the analyzed intervention group. The correlation between follow-up SRSI scores and the average percentage of doses taken was calculated using Pearson correlation.

Usability measures, including the PSSUQ [45] and the Health-ITUES [44], were analyzed by calculating the mean and SD for the overall scales and their subscales. The PSSUQ (18 items and 3 subscales) was measured using a 7-point Likert scale, ranging from 1 (“strongly agree”) to 7 (“strongly disagree”). The Health-ITUES (20 items and 4 subscales) was measured using a 5-point Likert scale, ranging from 1 (“strongly disagree”) to 5 (“strongly agree”).

Self-efficacy for HIV treatment adherence was measured using the HIV-ASES (12 items and 2 subscales) [43], a 10-point Likert scale ranging from 1 (“cannot do at all”) to 10 (“certain can do it”). Mean and SD were computed for self-efficacy scores. Independent 2-tailed *t* tests were used to compare HIV-ASES scores between groups, while paired 2-tailed *t* tests were used to compare scores within groups over time, with missing data handled as missing completely at random due to external factors unrelated to the study.

Cronbach α was calculated for scales such as SRSI, Health-ITUES, PSSUQ, and HIV-ASES to measure internal reliability. Statistical analyses were conducted using SAS (SAS Institute) [46] and R software (R Foundation for Statistical Computing) [47], and significance was determined at a *P* value of .05.

Qualitative Data Analysis

Following transcription, 3 members of the research team (SS, SM, and TC) conducted a double-coding process on the transcripts from 52% (12/23) intervention group participants and 48% (11/23) control group participants. An initial codebook was developed based on the Mobile Health Technology Acceptance Model (MHTAM), which includes technological, individual, and social factors that influence the acceptance of health care technology [22]. This model has been adapted for this study to examine an individual’s intention to use mHealth interventions for HIV treatment adherence or participate in the CHAMPS pilot study procedures, as it examines key factors such as perceived usefulness, ease of use, and ubiquity, alongside individual beliefs and social influences (Figure 3). Data analysis was conducted using Excel (Microsoft Corporation), which has been demonstrated as a practical tool for qualitative analysis due to its capabilities for organizing, coding, and analyzing textual data [48]. The methodological approach was guided by Ose [49], who outlined systematic steps for using Excel to manage and code qualitative data efficiently.

XSL•FO
RenderX

Figure 3. The Mobile Health Technology Acceptance Model adapted to analyze participant perceptions of the Community Health Worker and Mobile Health to Improve Viral Suppression (CHAMPS) intervention and experiences participating in the CHAMPS pilot study.

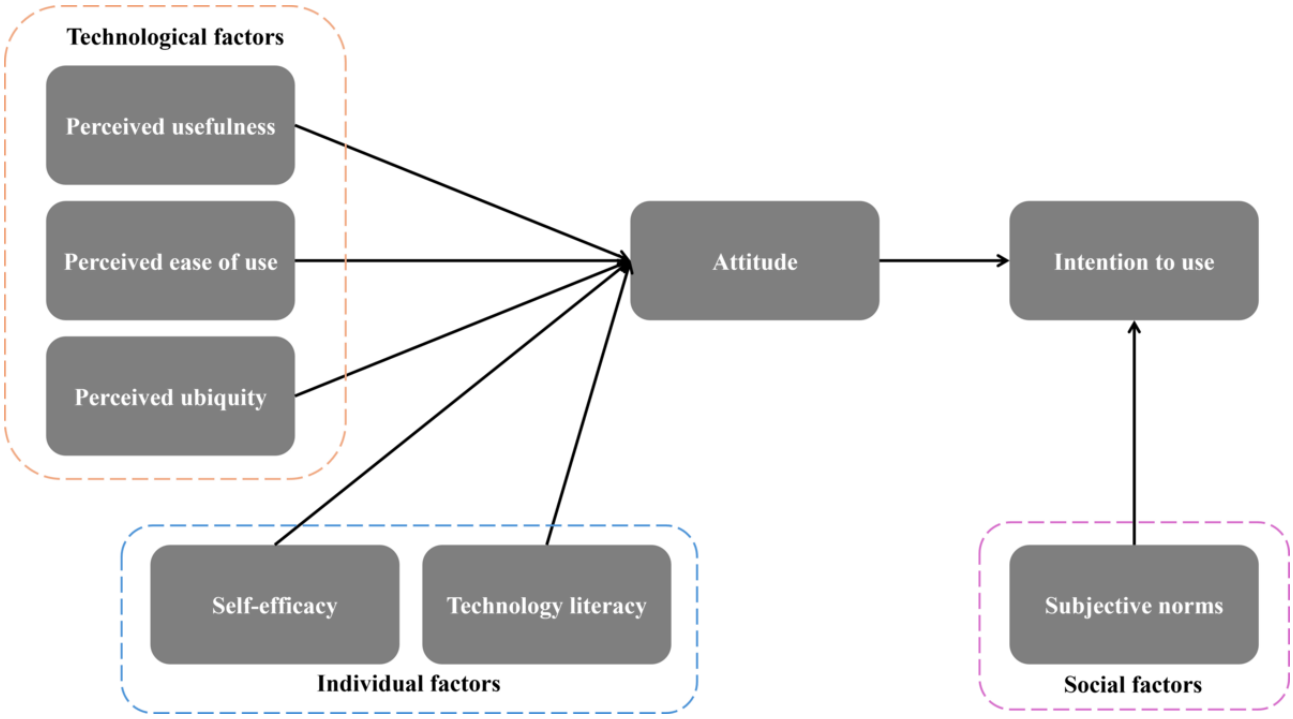

The codebook was created to align with the themes from the interview guide, resulting in the following finalized codes: (1) perceived usefulness, (2) perceived ease of use, (3) perceived ubiquity (ie, the perception of seamless access to health care networks and services at any time and place), (4) self-efficacy, (5) technology literacy, and (6) subjective norms. In addition, the codebook was applied deductively to analyze barriers and facilitators related to both the CHAMPS intervention and study procedures.

The coding process involved all coders coding 2 transcripts for initial consensus with the codebook and double coding 81% (17/21) of the remaining transcripts to enhance reliability, with regular meetings to compare and synthesize definitions for consistency. In certain cases, a consensus coder resolved discrepancies to ensure agreement. A small number of transcripts (4/21, 19%) were single coded due to availability constraints. Overall, double coding and consensus coding were applied to improve intercoder reliability and ensure a rigorous analytic process [50,51].

Ethical Considerations

All study procedures were reviewed and approved by the Columbia University Institutional Review Board (protocol number AAAU2064) before the recruitment or enrollment of participants. Participants provided electronic written consent, including screening consent for those requiring a screening visit and study consent for those proceeding directly to baseline. The baseline consent included language informing participants about an optional follow-up interview, with participants indicating their willingness to participate by initialing the consent form. Study data, including interview transcripts, were anonymized and deidentified to ensure confidentiality. Participants were compensated as follows: no compensation for the screening visit, US $40 for the baseline visit, US $50 for the follow-up visit, and US $35 for the optional qualitative interview at the follow-up visit. All images and data presented in this manuscript are deidentified to ensure the privacy of our participants.

Results

Participant Characteristics

Overview

A total of 40 participants were enrolled in the study, with 20 (50%) participants each randomly assigned to the intervention and control groups. Retention rates were lower in the intervention group (12/20, 60%) compared to the control group (18/20, 90%). Although retention was higher in the control group, this difference was not statistically significant (P=.07; Figure 4). Overall, the study retained 75% (30/40) of the participants, reflecting strong follow-up rates and participation in the study.

XSL•FO
RenderX

Figure 4. CONSORT (Consolidated Standards of Reporting Trials) flow diagram of participant enrollment, allocation, follow-up, and analysis within the Community Health Worker and Mobile Health to Improve Viral Suppression (CHAMPS) pilot study.

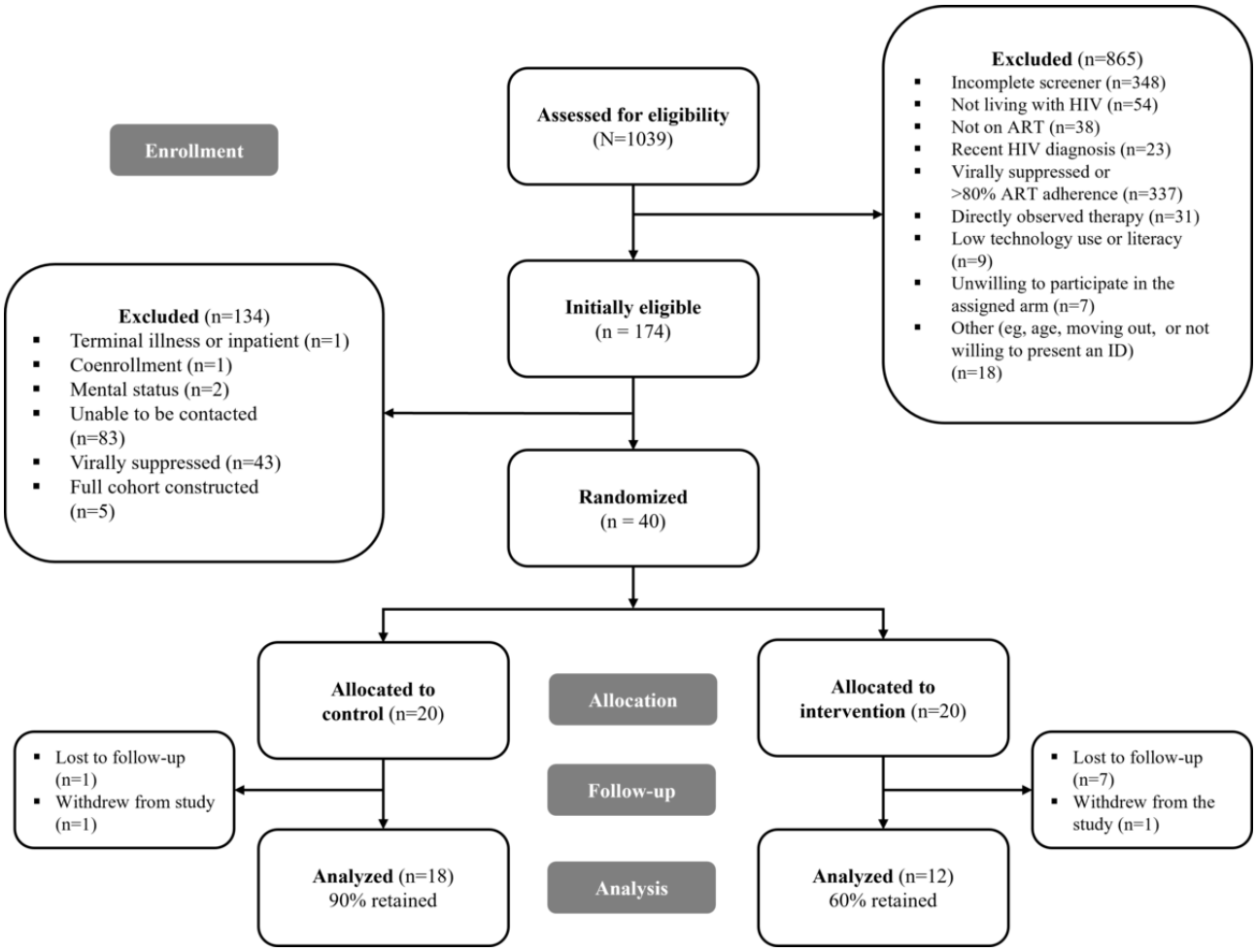

Participant Demographics

At baseline, participant demographics were well balanced between the intervention and control groups, indicating that randomization was effective. There were no statistically significant differences in key sociodemographic variables, including age, gender, race, or income level, between the 2 groups (all $P>.05$). The median age of the participants in the control group was 51.5 (IQR 22.0) years, compared to 45.0 (IQR 16.8) years in the intervention group. Gender distribution was similar, with 60% (12/20) of participants in each group identifying as male and 40% (8/20) of participants in the intervention group and 35% (7/20) of the participants in the control group identifying as female. 1 participant in the control group (1/20, 5%) identified as “something else” and were provided the opportunity to elaborate more on how they conceptualized their gender. The racial composition was predominantly African American or Black (10/20, 50% in the control group and 14/20, 70% in the intervention group.)

Educational attainment, employment status, and health insurance coverage were also comparable between the groups, supporting the conclusion that randomization successfully minimized baseline differences between the intervention and control arms. Notably, most participants (25/40, 63%) in both intervention and control groups reported annual incomes <US $20,000, potentially highlighting shared socioeconomic barriers that might limit their ability to fully engage with HIV care and maximize its potential benefits (Table 2).

XSL•FO
RenderX

Table 2. Baseline sociodemographic characteristics of participants, comparing the control and intervention groups (N=40).

Category	Control (n=20)	Intervention (n=20)	*P* value
Age (y), median (IQR)	51.5 (22.0)	45.0 (16.8)	.53[a]
Gender, n (%)			
Male	12 (60)	12 (60)	.99
Female	7 (35)	8 (40)	.78
Something else	1 (5)	0 (0)	.32
Sex, n (%)			
Male	13 (65)	12 (60)	.84
Female	7 (35)	8 (40)	.80
Sexual orientation, n (%)			
Homosexual, gay, or lesbian	8 (40)	10 (50)	.64
Heterosexual or straight	9 (45)	7 (35)	.62
Bisexual	2 (10)	2 (10)	.99
Asexual	1 (5)	0 (0)	.32
Something else (please specify)	0 (0)	1 (5)	.32
Race, n (%)			
American Indian, Alaska Native, Native Hawaiian, or other Pacific Islander	0 (0)	0 (0)	.99
Asian or Asian American	1 (5)	0 (0)	.32
Black or African American	10 (50)	14 (70)	.41
White	9 (45)	5 (25)	.29
Something else	0 (0)	1 (5)	.32
Ethnicity, n (%)			
Hispanic or Latino	2 (10)	2 (10)	.99
Not Hispanic or Latino	18 (90)	18 (90)	.99
Relationship status, n (%)			
Single	10 (50)	14 (70)	.41
In a relationship with a man	6 (30)	3 (15)	.32
In a relationship with a woman	1 (5)	0 (0)	.32
Divorced or separated from a man	1 (5)	0 (0)	.32
Divorced or separated from a woman	1 (5)	0 (0)	.32
Widowed	0 (0)	1 (5)	.32
Other	1 (5)	2 (10)	.56
Children, n (%)			
No	12 (60)	15 (75)	.56
Yes	8 (40)	5 (25)	.41
Education, n (%)			
None	0 (0)	0 (0)	.99
Elementary school	0 (0)	0 (0)	.99
Some high school or no diploma	2 (10)	3 (15)	.66
High-school diploma or equivalent (eg, GED[b])	5 (25)	5 (25)	.99
Some college	6 (30)	4 (20)	.53
Associate degree or technical degree	4 (20)	3 (15)	.71

XSL•FO
RenderX

Category	Control (n=20)	Intervention (n=20)	*P* value
Bachelor or college degree	2 (10)	3 (15)	.66
Professional or graduate degree	1 (5)	2 (10)	.56
Employment status, n (%)			
Working full time	3 (15)	4 (20)	.71
Working part time (eg, seasonal and work-study)	6 (30)	3 (15)	.32
Unemployed and looking for work	3 (15)	5 (25)	.48
Unemployed and not looking for work	0 (0)	2 (10)	.16
Retired	5 (25)	2 (10)	.26
Disabled	4 (20)	6 (30)	.53
Annual income (US $), n (%)			
<10,000	7 (35)	8 (40)	.80
10,000-19,999	5 (25)	5 (25)	.99
20,000-39,999	4 (20)	2 (10)	.41
40,000-59,999	3 (15)	1 (5)	.32
60,000-79,999	0 (0)	1 (5)	.32
80,000-99,999	0 (0)	1 (5)	.32
100,000-149,999	0 (0)	1 (5)	.32
≥150,000	1 (5)	0 (0)	.32
I do not know	0 (0)	1 (5)	.32
Health insurance status, n (%)			
No	1 (5)	0 (0)	.32
Yes, through my job	0 (0)	1 (5)	.32
Yes, through a health exchange (Affordable Care Act)	1 (5)	2 (10)	.56
Yes, Medicaid or Medicare	17 (85)	17 (85)	.99
Yes, AIDS Drug Assistance Program	3 (15)	3 (15)	.99
Other	1 (5)	1 (5)	.99
I do not know	0 (0)	0 (0)	.99

[a]The *P* value for age was calculated using the Mann-Whitney *U* test to compare the distribution of ages between the control and intervention groups based on the full dataset of individual ages.

[b]GED: General Educational Development.

Self-Efficacy

Self-efficacy has been recognized as a key factor influencing medication adherence in HIV treatment and other medical conditions [43]. As such, it was quantitatively measured in this study using the HIV-ASES, which has a maximum mean score of 10. At baseline, the control group (20/20, 100%) had a mean HIV-ASES overall score of 8.24 (SD 1.96), and the intervention group (20/20, 100%) had a mean score of 8.08 (SD 2.25). At follow-up, the control group's (18/20, 90%) mean score increased to 9.06 (SD 1.08), while the intervention group's (12/20, 60%) mean score was 8.78 (SD 2.16). Although unexpected, the intervention group's lower self-efficacy score compared to the control group may reflect increased awareness of their adherence habits through engagement with the CHAMPS intervention, leading to more accurate and self-critical reporting. Conversely, social desirability bias may have inflated the control group's self-reported scores. No significant differences were observed between the groups at baseline (P=.81) or follow-up (P=.64), and no significant within-group changes were observed over time (control: P=.52 and intervention: P=.07; Figure 5). The overall scale had excellent internal consistency, with Cronbach α=0.95 at baseline and Cronbach α=0.98 at follow-up [52].

XSL•FO
RenderX

Figure 5. Self-efficacy for antiretroviral therapy adherence among control and intervention group participants, as measured by the HIV Treatment Adherence Self-Efficacy Scale (HIV-ASES) at baseline and follow-up. Error bars represent SDs. No statistically significant differences were observed between groups or within groups over time. I: Integration subscale; P: Perseverance subscale.

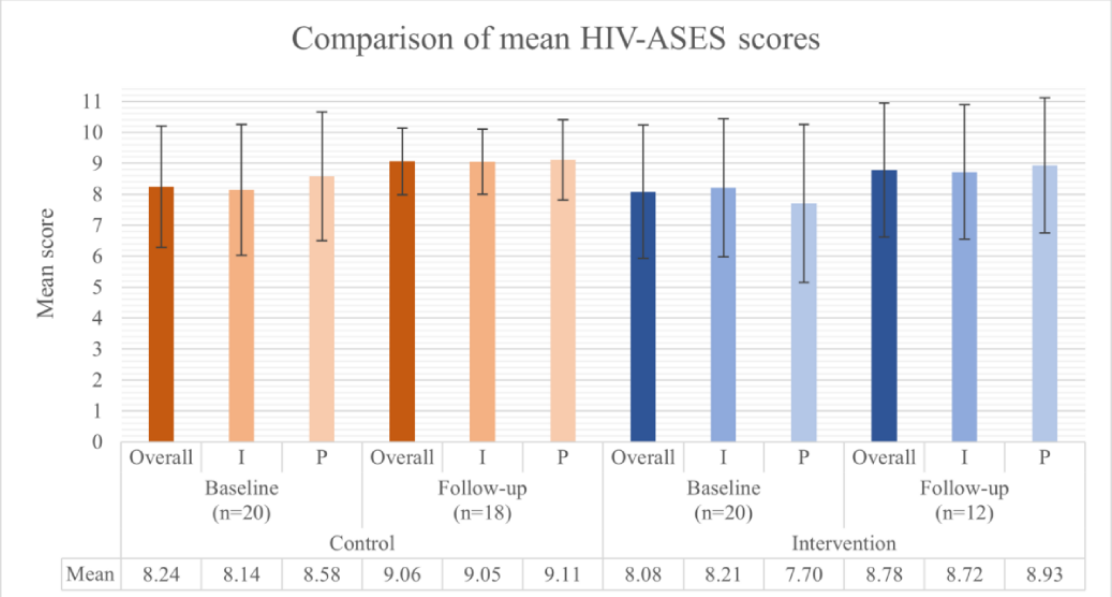

For the integration subscale, which includes the first 9 questions on the HIV-ASES and measures participants' ability to incorporate treatment into their daily lives [43], mean baseline scores were 8.14 (SD 2.11) for the control group and 8.21 (SD 2.23) for the intervention group. At follow-up, the mean scores increased to 9.05 (SD 1.05) and 8.72 (SD 2.17), respectively. No significant between-group differences were found at baseline (P=.92) or follow-up (P=.58), and there were no significant within-group changes (control: P=.26 and intervention: P=.89; Figure 5). The integration subscale had excellent reliability, with Cronbach α=0.95 at baseline and Cronbach α=0.97 at follow-up [52].

For the perseverance subscale, which includes the final 3 questions on the HIV-ASES and measures participants' confidence in maintaining adherence to their treatment regimens despite any challenges [43], the control group's baseline mean score was 8.58 (SD 2.08), and the intervention group's mean score was 7.70 (SD 2.55). Follow-up mean scores were 9.11 (SD 1.30) for the control group and 8.93 (SD 2.18) for the intervention group. No significant between-group differences were found at baseline (P=.24) or follow-up (P=.78), and no significant within-group changes were observed (control: P=.81 and intervention: P=.32; Figure 5). The perseverance subscale had acceptable and good reliability, with Cronbach α=0.87 at baseline and Cronbach α=0.93 at follow-up [52].

The quantitative findings, which indicated high self-efficacy in both the control and intervention groups, were further supported by qualitative feedback from participants, who described their personal commitment to staying on top of their medication routines. 1 participant shared how self-efficacy played a role in maintaining adherence:

> *Because I just like to have a good track record of me taking my medication in a timely fashion. So, when I saw that face wasn't a smile, and it was a frown, it made me stay on top of my, you know, on my game in terms of me being more consistent. Well, maintaining my consistency with taking my medication.* [CHP 13]

These findings highlight the strong sense of personal responsibility and self-efficacy that participants expressed in their medication adherence, where they were motivated by a desire to maintain their own health. A detailed list of additional illustrative quotes organized by the MHTAM themes and their respective definitions is provided in Multimedia Appendix 1.

Delivered Intervention

A total of 10 CHW sessions were conducted as planned, although the optional 2 additional sessions were not used due to declining participant engagement and the proximity of some sessions to the participants' final follow-up visits. CHWs used HIPAA-compliant Zoom calls, phone calls, and the WiseApp chat feature based on participant preferences. Completion rates (Table 3) were initially high, with 17 (85%) out of the 20 intervention participants completing sessions 1 to 3, but this declined to 9 (45%) participants by the 10th session. Participants generally preferred phone calls for session delivery, with a consistent majority opting for this modality across all sessions. The variability in delivery modes during sessions 1 and 2 reflects logistical factors: some participants had not received their CleverCap devices at the time of their baseline visit, necessitating follow-up sessions via Zoom call or phone call for session 2. In addition, 3 participants disengaged from the study immediately after the baseline visit, which contributed to the drop in session completion rates early in the intervention. To re-engage participants who missed sessions, CHWs attempted up to 3 follow-ups using participants' disclosed communication preferences (calls, SMS text messaging, or emails). When no response was received after 3 attempts, re-engagement efforts were discontinued.

XSL•FO
RenderX

Table 3. Completion rates and delivery model for community health worker sessions among participants in the intervention arm (n=20).

Session number	Participants completing the session, n (%)	Mode of session delivery, n (%)		
		Zoom	Phone	WiseApp chat
1	17 (85)	9 (53)	8 (47)	0 (0)
2	17 (85)	3 (18)	14 (82)	0 (0)
3	17 (85)	1 (6)	15 (88)	1 (6)
4	13 (65)	0 ()	13 (100)	0 (0)
5	13 (65)	1 (8)	11 (85)	1 (8)
6	13 (65)	0 (0)	13 (100)	0 (0)
7	13 (65)	0 (0)	12 (92)	1 (8)
8	12 (60)	1 (8)	11 (92)	0 (0)
9	12 (60)	1 (8)	10 (83)	1 (8)
10	9 (45)	2 (22)	7 (78)	0 (0)

The CleverCap pill bottles issued 3151 notifications, comprising notifications for taking their medication and alerts for nonadherence or improper use (Table 4). Notifications sent per participant varied widely, with a mean value of 158.3 (SD 99.9). Participants receiving >200 notifications often had frequently missed or off-schedule doses, while those receiving <90 notifications likely disengaged from the intervention by deactivating, losing, or destroying their pill bottles. This 90-notification threshold was based on the intervention design, which included 1 scheduled daily reminder over the 3-month study period (3 months ×30 days=90 reminders per participant). Any participant receiving substantially <90 notifications likely did not use the device consistently, whereas those receiving >90 notifications frequently triggered additional alerts due to missed or off-schedule doses.

Table 4. Frequency and distribution of medication adherence notifications issued by the CleverCap pill bottle to participants in the intervention arm (N=3151).

Notification type	Notifications, n (%)
Dose taken	1061 (33.67)
Missed dose	1585 (50.3)
Off-schedule dose	418 (13.26)
Improperly closed pill bottle	87 (2.76)

Intervention Characteristics

Overview

The aOR for achieving undetectable viral load in the intervention group compared to the control group was 3.01 (95% CI –1.59 to 4.12), controlling for baseline viral load status (Table 5). This aOR represents a medium effect size, suggesting an improvement in viral suppression in the intervention group [53]. Baseline viral load status by follow-up status and group is presented in Multimedia Appendix 2.

Table 5. Baseline and follow-up viral load status of control and intervention group participants (N=40).

Viral load[a]	Baseline, n (%)			Follow-up, n (%)		
	Control (n=20)	Intervention (n=20)	Total (n=40)	Control (n=18)	Intervention (n=12)	Total (n=29)
Not detectable (<200 copies/ml)	17 (85)	11 (55)	28 (70)	15[b] (83)	11 (92)	26 (90)
Detectable (≥200 copies/ml)	3 (15)	9 (45)	12 (30)	2 (11)	1 (8)	3 (10)

[a]For consistency with US clinical guidelines, viral load status is categorized using a threshold of ≥200 copies/ml to define detectability. While participants with viral loads between 50 and 199 copies/ml were eligible for the study, they were classified as suppressed in this analysis. In addition, details on the methods used to obtain viral load measurements at baseline and follow-up are provided in Multimedia Appendix 3.

[b]Data missing (1/18, 6%) as an insufficient sample provided by 1 participant precluded analysis and quantification.

Notably, nearly three-fourths of the participants (28/40, 70%) had an undetectable viral load (<200 copies/ml) at baseline. Among those with undetectable viral load at baseline, viral suppression was largely maintained throughout the study period. Specifically, in the control group, the viral load of 1 participant became detectable at follow-up, whereas in the intervention group, all participants' viral loads that were initially undetectable remained undetectable. Among participants with a detectable

XSL•FO
RenderX

viral load (≥200 copies/ml) at baseline, 44% (4/9) of individuals in the intervention group achieved viral suppression by follow-up, compared to only 33% (1/3) in the control group.

At baseline, the control group (20/20, 100%) had a mean SRSI score of 4.80 (SD 1.34), while the intervention group (20/20, 100%) had a mean score of 4.90 (SD 1.35). At follow-up, the control group's (18/20, 90%) mean increased to 5.27 (SD 1.23), and the intervention group's (12/20, 60%) mean was 5.25 (SD 1.17). No significant differences were observed between the groups at baseline (*P*=.77) or follow-up (*P*=.95), and no significant within-group changes were observed over time (control: *P*=.16 and intervention: *P*=.29; Figure 6). The SRSI score showed excellent internal consistency, with Cronbach α=0.96 at baseline and Cronbach α=0.98 at follow-up [52].

Figure 6. Changes in self-reported adherence to antiretroviral therapy among control and intervention group participants, as measured using the self-reported scale item (SRSI) at baseline and follow-up. Error bars represent SDs. No statistically significant differences were observed between groups or within groups over time.

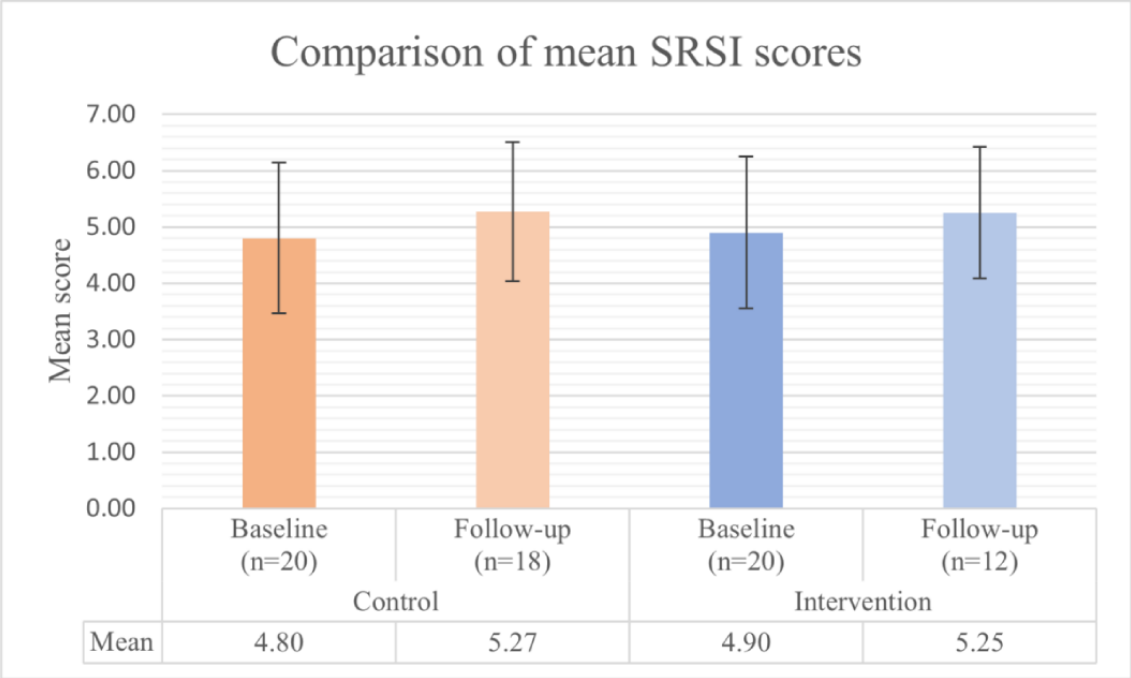

Adherence data from CleverCap showed that intervention group participants (12/20, 60%) took an average of 60.08% (SD 19.15%) of their prescribed doses, with 39.33% (SD 19.24%) of doses missed. A weak, nonsignificant correlation was found between the CleverCap adherence data (percentage taken) and follow-up SRSI scores (*r*=0.121; *P*=.71), suggesting that the high self-reported adherence scores did not strongly align with the electronic monitoring data.

Participant feedback can help explain the discrepancy between self-reported scores and actual adherence by highlighting individual factors, such as personal medication management preferences, that influence adherence. 1 participant noted the following, illustrating how participants' personal beliefs about the medication's dosing flexibility influenced their adherence behavior:

> *I know that Biktarvy has a 50-hour shelf-life. And so, I'm not really stressing over when I take it. I usually just make sure I take it before lunch.* [CHP 2]

Rather than relying on CleverCap's programmed alarms, participants adjusted their dosing schedules based on their own understanding of the medication's efficacy, which may have contributed to lower adherence rates recorded by the pill bottle.

Feasibility of the Remotely Delivered CleverCap Pill Bottle and the WiseApp Intervention

The feasibility of the intervention was assessed by supplementing usability data with feedback regarding participants' attitudes and intentions to use the intervention [54,55]. The usability of the intervention was assessed using the Health-ITUES and PSSUQ scores, which provided insights into participants' experiences and satisfaction with the intervention, respectively. The Health-ITUES had a maximum mean score of 5, with higher scores indicating better usability. Participants (12/20, 60%) rated the overall usability of the intervention positively, with an average overall score of 4.35 (SD 0.58). The scores ranged from 3.50 to 5, with an IQR of 0.84, reflecting consistently high usability ratings. Specific subscales, such as perceived usefulness (mean 4.42, SD 0.63) and perceived ease of use (mean 4.47, SD 0.71), received similarly high ratings, with IQRs of 1.11 and 0.50, respectively, indicating that participants generally found the pill bottle and app useful and easy to use (Table 6).

XSL•FO
RenderX

Table 6. Usability and acceptability of the Community Health Worker and Mobile Health to Improve Viral Suppression intervention, as assessed by the Health Information Technology Usability Evaluation Scale (Health-ITUES) and Poststudy System Usability Questionnaire (PSSUQ) among intervention group participants.

Scale	Scores, mean (SD)	Scores, median (IQR; range)	Cronbach α
Health-ITUES (n=12)			
Overall	4.35 (0.58)	4.48 (0.84; 3.50-5)	0.94
Impact	4.36 (0.66)	4.17 (1.08; 3.33-5)	0.92
Perceived usefulness	4.42 (0.63)	4.33 (1.11; 3.11-5)	0.83
Perceived ease of use	4.47 (0.71)	4.90 (0.50; 3.67-5)	0.63
User control	3.86 (0.95)	4.07 (1.75; 2.33-5)	0.79
PSSUQ (n=12)			
Overall	2.04 (1.03)	3.81 (1.31; 1-4.19)	0.96
System usefulness	1.87 (1.09)	1.79 (0.79; 1-4.86)	0.94
Information quality	2.13 (1.16)	1.67 (2.04; 1-3.67)	0.96
Interface quality	2.05 (0.98)	2.17 (1.58; 1-4.33)	0.86

These high ratings are echoed in participant feedback. The pill bottle's visual presence was perceived as useful for maintaining adherence. Participants also found the pill bottle easy to operate and intuitive, sharing positive experiences as follows:

> *Because I'm not really good with taking it at a certain time every day, at the same time every day. And so I feel like the CleverCap helped me with that...It does remind me, because I always see it, versus me keeping it in a white bottle. When I see the CleverCap on my dresser, it just reminds me. "Okay, it's time to take your medicine."...The functions of the CleverCap were easy to use. I didn't have to charge it many times. I charged it once. It's very easy. It reminded me to take my medicine. I really, really appreciated that. I just love it. It's very easy. And it definitely reminded me to stay on my meds.* [CHP 39]

However, the user control subscale received a lower average score (mean 3.86, SD 0.95) and a wider IQR of 1.75, suggesting more variability in responses and some limitations in how participants felt they could manage the intervention. These concerns were reflected in feedback about the pill bottle's fragility and usability during travel:

> *Because there were times I was out of town, and I don't like bringing the pill bottle with me...I didn't take the bottle. I took my medication, but I didn't take the reminder, the bottle...That container, with the alarm on, it's very sensitive. I dropped it one time; I was afraid it might break...So, that's why when I knew I was leaving to go out of town, I just felt it was more secure leaving it at home.* [CHP 13]

Despite this, some participants (3/12, 25%) were able to maintain adherence even when separated from the pill bottle, relying on their own strategies for medication management. 1 participant reflected on this experience, demonstrating that while CleverCap supported adherence, participants' self-efficacy and personal management strategies often played a crucial role in overcoming barriers related to the pill bottle's portability and user control:

> *When I went to D.C., I forgot the CleverCap...But I had my medicine. I just got the notifications like, oh, it's time to take your medicine. Didn't bring the cap.* [CHP 32]

The PSSUQ had a maximum mean score of 7, where lower scores indicate better usability. Participants reported generally high satisfaction with the system's usability, with an average overall score of 2.04 (SD 1.03). The scores ranged from 1 to 4.19, with an IQR of 1.31, indicating a favorable user experience with some variability. Ratings for system usefulness (mean 1.87, SD 1.09) and interface quality (mean 2.05, SD 0.98) subscales also indicate positive user experiences, with IQRs of 0.79 and 1.58, respectively, although information quality subscale (mean 2.13, SD 1.16) showed highest variability in participant responses, with an IQR of 2.04 (Table 5).

To assess the internal reliability of the scales used, Cronbach α was calculated for both the Health-ITUES and PSSUQ. The Health-ITUES demonstrated excellent overall reliability (Cronbach α=0.94), with variable consistency across the following subscales: perceived usefulness (Cronbach α=0.92, excellent), integration (Cronbach α=0.83, good), perceived ease of use (Cronbach α=0.63, acceptable), and user control (Cronbach α=0.76, acceptable). The PSSUQ exhibited excellent overall reliability (Cronbach α=0.96), with high consistency across the following subscales: system usefulness (Cronbach α=0.94, excellent), information quality (Cronbach α=0.96, excellent), and interface quality (Cronbach α=0.86, good) [52].

Our results suggest that, despite potential improvements in user control and information clarity, the remote delivery of the CleverCap pill bottle and WiseApp was generally well received, demonstrating feasibility and ease of use.

Perceptions of CHW Sessions

Participants shared a variety of experiences regarding their engagement with the CHW sessions, with responses generally falling into 3 key themes: ease of use and self-efficacy as a facilitator; positive social influence; and barriers to ubiquity, such as infrastructure or scheduling challenges.

XSL•FO
RenderX

Several participants (10/12, 83%) expressed that the flexibility and accessibility of the CHW sessions facilitated their engagement. The ability to communicate with CHWs through various channels, such as text, phone, and email, was highly appreciated. For instance, 1 participant highlighted the ease of communication with and persistence of the CHWs. In addition, participants with existing knowledge about HIV felt comfortable discussing their treatment with the CHWs, further facilitating their engagement:

> *The flexibility that you had was amazing. Like I said, you guys were available through a text message, through a phone call, through email...whatever it took. You guys didn't disappear. You guys didn't give up. You guys are like, hello, I am trying to talk to you, and kept on it. So, that was cool...To a person who already knows about HIV, I am completely comfortable talking about anything that has to do with it. So, that wasn't an issue for me.* [CHP 17]

Positive social influence, particularly through regular check-ins and motivational support from the CHWs, emerged as a key facilitator of engagement. Participants expressed appreciation for the ongoing support provided by the CHWs, with some (10/12, 83%) noting that the regular communication made them feel more accountable and supported in their adherence:

> *[The study] really was awesome. I mean the perk was having you call me once a week...Having your own personal somebody to motivate you and check up on you and make sure you're doing good and adhering to your medicine.* [CHP 39]

While many participants found the CHW sessions helpful, some (2/12, 16%) faced barriers to engagement due to external factors, such as infrastructure limitations or time constraints. Participants living in rural areas with limited internet access found it challenging to maintain consistent communication with CHWs. 1 participant from a rural area explained how poor internet connectivity hindered their participation:

> *Like I said, I'm busy. The issue with scheduling because, just like I said, I was always on the go. I'm always on the go...Travel, not having internet for interaction with you. That was sometimes challenging, I guess. Where I live, here in rural Alabama, I dare say that 75% of the county... [does] not have any kind of access to internet. It's scarce...the internet and phone service in a lot of spaces. And so that would be a challenge to speaking with you, as well as trying to look at any data on the phone.* [CHP 38]

In addition, some participants (4/12, 33%) mentioned that their busy schedules made it difficult to consistently engage in the CHW sessions (Multimedia Appendix 1).

Thus, while there were some challenges related to infrastructure and scheduling, participants' perceptions suggest that the CHW sessions were generally well received, particularly in terms of flexibility, ease of communication, and motivational support.

Acceptability of Remote Study Procedures

The acceptability of delivering the intervention and conducting study procedures remotely was assessed through participant feedback, which highlighted the overall usefulness of participating in the study; ease of remote interactions; and logistical considerations, such as shipping and blood sample collection.

Participants largely embraced the flexibility offered by remote study visits conducted via Zoom. Many (15/23, 65%) expressed comfort and appreciation for the convenience of participating from their own homes, along with increased privacy and adaptability. 1 participant noted the following:

> *I have used Zoom in the past for like work...But with Zoom, it's actually more...what's the word I'm looking for? This was more private...If I had to step out and go somewhere, I still could do it that way. But, I felt more comfortable doing it on my laptop, you know, at home. And, I liked the flexibility I had with you all, as well. So, it was great, because you all worked around my schedule.* [CHP 23]

Overall, participants found Zoom easy to use and appreciated the accessibility of conducting study visits remotely, especially in cases where mobility or health concerns made in-person visits difficult. 1 participant explained as follows:

> *Doesn't bother me. I'd rather [Zoom] than have to come out. It's difficult for me to walk right now...This is good because at least I can participate in things. And I don't have to make myself uncomfortable to do it, you know?* [CHP 30]

Remote collection of blood samples, facilitated through mailed kits, was another aspect of the study that participants generally found convenient. 1 participant described the process as efficient and accessible:

> *I think [the blood sample collection process is] very convenient. I'm nowhere near you guys and normally, you know, we'd have someone say, well, you're not able to come into the office and do the blood draws or whatever. So, this was very convenient.* [CHP 11]

Furthermore, participants expressed appreciation for the comprehensive kits and clear instructions provided (Multimedia Appendix 1). Although some challenges with blood sample collection were reported, including minor issues such as bruising and difficulties with blood flow, these were anticipated and communicated during the informed consent process, as such challenges are commonly associated with point-of-care testing [56]. Multimedia Appendix 3 provides additional numeric data supporting these findings.

The shipping process for study kits and the CleverCap pill bottle was largely seamless for participants, who noted the discreet packaging of materials as a positive aspect of the study. 1 participant mentioned how the packaging helped protect their privacy:

> *Just because of the delicateness of the subject matter, it's essentially discreet because nobody thinks a FedEx package is anything suspect or scary. It's just*

XSL•FO
RenderX

> *a FedEx package...it blended right in and it doesn't cause any red flags or make anybody feel uncomfortable in a public setting.* [CHP 17]

However, there were occasional logistical challenges with returning the packages. A few participants (2/23, 9%) noted issues with transportation and scheduling FedEx pickups, which delayed their ability to return the materials. 1 participant shared the following:

> *Because, I don't always have transportation, so it's been harder for me to get the package sent back in like a timely manner. I know you guys have tried setting up a pick-up with FedEx, but they never came to pick up the package. So, that's like been the only barrier I've had in participating in the program and getting the package to a FedEx location.* [CHP 27]

While logistical challenges, such as transportation for package returns, were noted, the study's design was practical and effective for most participants, supporting its acceptability and viability in remote settings.

Discussion

Principal Findings

This pilot study aimed to primarily assess the feasibility and usability of the remotely delivered CHAMPS intervention. While the preliminary results did not demonstrate significant improvements in adherence or self-efficacy in the intervention group compared to the control group, the findings provide important insights into how participant attitudes to use the intervention were shaped by technological, individual, and social factors. Thus, understanding these sociodemographic factors is essential for contextualizing the intervention's feasibility and the challenges participants faced in engaging with the CHAMPS intervention. The demographic characteristics of participants in this study partially align with key sociodemographic characteristics of people with HIV in the United States. These parallels suggest that our findings may be applicable to subsets of the broader population of people with HIV, particularly those disproportionately affected by intersecting socioeconomic and racial disparities. Most of our participants (24/40, 60%) identified as Black or African American individuals, consistent with national data showing that nearly half of people with HIV belong to this group [57]. The predominance of participants reporting annual incomes <US $20,000, which is nearly double the proportion of individuals living below the federal poverty level (US $15,060/y) compared to the national estimate of one-third of people with HIV, may reflect competing priorities that made it challenging to fully engage with the intervention and may help explain the nonsignificant results observed [57].

Baseline HIV-ASES scores showed that participants in both control and intervention groups had high self-efficacy, particularly in integrating HIV medication into their routines and maintaining long-term adherence. This is critical, as self-efficacy is a well-established predictor of successful health behavior change, particularly for chronic conditions such as HIV [58,59]. Although modest increases in scores were observed at follow-up, they were not statistically significant, and the control group reported higher self-efficacy than the intervention group, indicating that the intervention did not significantly enhance an already strong sense of self-efficacy. Qualitative feedback confirmed that participants relied on their preexisting knowledge and ability to adhere to medication, making additional tools unnecessary.

SRSI scores showed stable adherence in both groups from baseline to follow-up, with slight, nonsignificant increases. Although the intervention did not significantly improve adherence, qualitative data revealed that participants, especially those with higher self-efficacy, often prioritized their own understanding of medication efficacy over CleverCap reminders. High confidence in managing their treatment may have made these reminders less relevant, explaining the lack of significant differences between groups.

Health-ITUES results showed the intervention's success in usability, with an overall score of 4.35 exceeding the acceptable cut point of 4.32 [60]. High scores in perceived usefulness and ease of use subscale indicated that participants found WiseApp helpful and intuitive, supporting the MHTAM framework's emphasis on these factors for technology adoption [22]. However, lower scores in user control revealed challenges with CleverCap's portability, aligning with MHTAM's perceived ubiquity construct. PSSUQ scores also reflected positive usability, with an overall mean of 2.04 (SD 1.03). Participants rated system usefulness and interface quality highly.

Subjective norms emerged as an important factor for engagement with the CHW component of the intervention. The regular check-ins and motivational support from CHWs were highlighted as valuable by participants, reinforcing the social component of adherence. This reflects the broader role of CHWs in providing not only practical assistance but also emotional support, which has been shown to improve health outcomes in other studies [21-23].

This pilot study found high acceptability for remotely delivered procedures, with participants appreciating the convenience, privacy, and flexibility of remote interactions. Zoom visits were well received, allowing participation from home, which aligns with literature suggesting that remote health care improves patient satisfaction by reducing travel burden and enhancing accessibility [61-63]. These procedures were key in promoting participation and retention. Blood sample collection via mailed kits was convenient for most participants (27/30, 90%), demonstrating that, with clear guidance, it is a feasible alternative to in-person phlebotomy. Discreet packaging of study materials was also valued, as it addressed privacy needs and reduced HIV stigma, which is a critical factor in engaging people with HIV in care [64,65].

Our findings highlight the potential of mHealth interventions, such as CHAMPS, to support the "Treat" pillar of the Ending the HIV Epidemic initiative by enhancing accessibility and convenience for people with HIV. The usability and acceptability of remote delivery methods, such as mobile apps for medication management, CHW support, and mailed DBS collection, demonstrate a promising foundation for addressing barriers to care. These strategies could be further refined to promote viral suppression and improve engagement in HIV care

XSL•FO
RenderX

for populations facing challenges in accessing traditional health care services.

Limitations

This study has several limitations that should be considered. The small sample size of 40 participants limited the statistical power, making it difficult to detect significant differences between groups and reducing the generalizability of findings to the broader populations of people with HIV. Furthermore, the small sample size for positive outcome events, such as viral suppression at follow-up, may have affected the reliability of the effect size estimate, particularly within the intervention group. The lower retention rate in the intervention group compared to the control group, while not statistically significant, suggests challenges in effectively engaging participants assigned to the intervention. This highlights the need to address structural and social determinants of health when designing ART adherence interventions. Future research should explore strategies to reduce these barriers, such as more flexible engagement options or additional resources.

In addition, recruitment challenges, particularly among younger people with HIV (aged 18 to 29 y), were notable. Social media platforms used for advertising, such as POZ.com, primarily attracted older participants due to their demographic reach. Alternative platforms, such as TikTok, which are more commonly used by younger populations, should be prioritized in future campaigns [66]. Strategies such as peer-driven recruitment and culturally tailored SMS text messaging could help engage younger individuals who may be less trusting of traditional health care outreach methods. Addressing time constraints and providing more flexible scheduling options may also increase participation among this demographic [67,68]. Furthermore, it is important to note that participants who responded to advertisements on POZ.com may not be representative of the broader US population living with HIV, introducing potential selection bias. Financial incentives for study participation may have influenced responses, potentially skewing the participant pool toward individuals with greater financial needs.

Another significant limitation is the reliance on self-reported adherence data in the control group, which is prone to bias. Self-reported measures, particularly in the context of health behaviors, can be subject to overreporting due to recall bias or social desirability, where participants may provide responses that they believe are more acceptable to researchers [69,70]. Although the SRSI scale demonstrated high internal consistency, discrepancies between self-reported adherence and CleverCap data suggest that participants' actual behaviors differed. Some adjusted their adherence schedules based on personal beliefs, indicating that self-reported data may not fully reflect true adherence. In addition, the reliance on self-reported viral load data as an inclusion criterion may have introduced some inaccuracies, as nearly half of the sample reported being virally unsuppressed (threshold of ≥200 copies/ml) in the past 12 months despite being clinically undetectable at baseline. This difference suggests that participants may have misremembered or misreported their viral load history, highlighting a potential limitation in self-reported viral suppression measures.

Discrepancies between the planned and delivered intervention fidelity were observed, with several factors impeding full implementation. While the intervention was designed to provide consistent support through regular CHW sessions and timely CleverCap notifications, technical challenges, such as unreliable local telecommunication networks, affected the delivery of mHealth components, leading to missed or delayed notifications. In addition, participant engagement waned over time, with some individuals disengaging from CHW sessions and underusing the CleverCap device, a challenge commonly reported in mHealth interventions [71,72]. The older demographic of the study population introduced additional barriers, as many participants had difficulty programming medication alerts without CHW assistance, leading to missed doses. This aligns with previous research indicating that older adults frequently encounter usability issues with electronic adherence tools, which can negatively impact adherence [71]. These findings highlight the importance of tailoring adherence interventions to address technological literacy and providing structured training for older adults.

Finally, the lack of a universal data verification process for the DBS results posed a limitation. While DBS samples were collected and processed according to established protocols, most participants' (25/30, 83%) DBS results were not cross verified with other sources. Only a small subset of participants (5/30, 17%) who had requested a screening visit and subsequently provided electronic health record viral load data had their DBS results cross-checked. This reliance on unverified DBS data for most participants may introduce potential inaccuracies. Furthermore, the moderate sensitivity (80%-95%) and specificity (85%-90%) of DBS for viral load thresholds, such as 200 to 1000 copies/ml, could lead to misclassification, with some participants incorrectly identified as having high or low viral loads [73-75]. Future studies should incorporate systematic verification processes, such as routine cross-checking with laboratory results or electronic health records, to enhance the reliability of biomarker data.

Conclusions

The findings from this pilot study support the feasibility and acceptability of a remotely delivered mHealth intervention integrated with CHW support for people with HIV. Participants reported high levels of usability and satisfaction with both the WiseApp and CleverCap, highlighting the intervention's potential to promote medication adherence and engagement in care. Despite modest, nonsignificant improvements in self-reported adherence and self-efficacy, the flexibility of the intervention and its ability to address logistical challenges make it a promising tool for broader implementation. These results suggest that with further optimization, this intervention could be a valuable resource in enhancing health outcomes for people with HIV across diverse settings.

XSL•FO
RenderX

Acknowledgments

This study was supported by the National Institute of Nursing Research (award number R01NR019758). Dried blood spot kit processing at the Miriam Hospital (Providence, Rhode Island) was supported by the Providence-Boston Center for AIDS Research under award number P30AI042853. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors would like to extend their gratitude to the participants and community health workers involved in this study, whose contributions were invaluable. The authors would also like to thank the broader research team and their institutional partners for their continuous support and dedication to improving care for people with HIV.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to participant confidentiality and institutional data protection policies, but are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Key themes, definitions, and illustrative quotes from control and intervention group participants, organized according to the Mobile Health Technology Acceptance Model.
[PDF File (Adobe PDF File), 174 KB - formative_v9i1e67997_app1.pdf]

Multimedia Appendix 2
Retention of participants at follow-up by baseline viral load status, assessing whether the loss to follow-up was differentially associated with baseline viral load in the intervention and control groups.
[PDF File (Adobe PDF File), 69 KB - formative_v9i1e67997_app2.pdf]

Multimedia Appendix 3
Methods of viral load and CD4 data collection at baseline and follow-up, including self-reported and laboratory measures.
[PDF File (Adobe PDF File), 68 KB - formative_v9i1e67997_app3.pdf]

Multimedia Appendix 4
CONSORT eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 17131 KB - formative_v9i1e67997_app4.pdf]

References

1. Deeks SG, Lewin SR, Havlir DV. The end of AIDS: HIV infection as a chronic disease. Lancet 2013 Nov 02;382(9903):1525-1533 [FREE Full text] [doi: 10.1016/S0140-6736(13)61809-7] [Medline: 24152939]
2. Marcus JL, Chao CR, Leyden WA, Xu L, Quesenberry CP, Klein DB, et al. Narrowing the gap in life expectancy between HIV-infected and HIV-uninfected individuals with access to care. J Acquir Immune Defic Syndr 2016 Sep 01;73(1):39-46 [FREE Full text] [doi: 10.1097/QAI.0000000000001014] [Medline: 27028501]
3. May MT, Gompels M, Delpech V, Porter K, Orkin C, Kegg S, UK Collaborative HIV Cohort (UK CHIC) Study. Impact on life expectancy of HIV-1 positive individuals of CD4+ cell count and viral load response to antiretroviral therapy. AIDS 2014 May 15;28(8):1193-1202 [FREE Full text] [doi: 10.1097/QAD.0000000000000243] [Medline: 24556869]
4. Sematlane NP, Knight L, Masquillier C, Wouters E. Adapting to, integrating and self-managing HIV as a chronic illness: a scoping review protocol. BMJ Open 2021 Jun 23;11(6):e047870 [FREE Full text] [doi: 10.1136/bmjopen-2020-047870] [Medline: 34162650]
5. HIV surveillance supplemental report: estimated HIV incidence and prevalence in the United States. Centers for Disease Control and Prevention. 2018. URL: https://stacks.cdc.gov/view/cdc/156513 [accessed 2024-04-29]
6. McComsey GA, Lingohr-Smith M, Rogers R, Lin J, Donga P. Real-world adherence to antiretroviral therapy among HIV-1 patients across the United States. Adv Ther 2021 Sep 14;38(9):4961-4974 [FREE Full text] [doi: 10.1007/s12325-021-01883-8] [Medline: 34390465]
7. Frandsen BR, Joynt KE, Rebitzer JB, Jha AK. Care fragmentation, quality, and costs among chronically ill patients. Am J Manag Care 2015 May;21(5):355-362 [FREE Full text] [Medline: 26167702]
8. Fauci AS, Redfield RR, Sigounas G, Weahkee MD, Giroir BP. Ending the HIV epidemic: a plan for the United States. JAMA 2019 Mar 05;321(9):844-845. [doi: 10.1001/jama.2019.1343] [Medline: 30730529]
9. Cohen MS, McCauley M, Gamble TR. HIV treatment as prevention and HPTN 052. Curr Opin HIV AIDS 2012;7(2):99-105. [doi: 10.1097/coh.0b013e32834f5cf2]

XSL•FO
RenderX

10. Cohen MS, Chen YQ, McCauley M, Gamble T, Hosseinipour MC, Kumarasamy N, HPTN 052 Study Team. Prevention of HIV-1 infection with early antiretroviral therapy. N Engl J Med 2011 Aug 11;365(6):493-505 [FREE Full text] [doi: 10.1056/NEJMoa1105243] [Medline: 21767103]
11. Yu Y, Luo D, Chen X, Huang Z, Wang M, Xiao S. Medication adherence to antiretroviral therapy among newly treated people living with HIV. BMC Public Health 2018 Jul 04;18(1):825 [FREE Full text] [doi: 10.1186/s12889-018-5731-z] [Medline: 29973167]
12. Bandera A, Gori A, Clerici M, Sironi M. Phylogenies in ART: HIV reservoirs, HIV latency and drug resistance. Curr Opin Pharmacol 2019 Oct;48:24-32. [doi: 10.1016/j.coph.2019.03.003] [Medline: 31029861]
13. Martin F, Lee J, Thomson E, Tarrant N, Hale A, Lacey CJ. Two cases of possible transmitted drug-resistant HIV: likely HIV superinfection and unmasking of pre-existing resistance. Int J STD AIDS 2016 Jan 07;27(1):66-69 [FREE Full text] [doi: 10.1177/0956462415571671] [Medline: 25663247]
14. Mugavero MJ, Amico KR, Horn T, Thompson MA. The state of engagement in HIV care in the United States: from cascade to continuum to control. Clin Infect Dis 2013 Oct 23;57(8):1164-1171. [doi: 10.1093/cid/cit420] [Medline: 23797289]
15. Dasgupta S, Tie Y, Beer L, Fagan J, Weiser J. Barriers to HIV care by viral suppression status among US adults with HIV: findings from the Centers for Disease Control and Prevention medical monitoring project. J Assoc Nurses AIDS Care 2021;32(5):561-568 [FREE Full text] [doi: 10.1097/JNC.0000000000000249] [Medline: 33769329]
16. Kempf MC, McLeod J, Boehme AK, Walcott MW, Wright L, Seal P, et al. A qualitative study of the barriers and facilitators to retention-in-care among HIV-positive women in the rural southeastern United States: implications for targeted interventions. AIDS Patient Care STDS 2010 Aug;24(8):515-520. [doi: 10.1089/apc.2010.0065] [Medline: 20672971]
17. Marshall SA, Tao J, Horton HK, Montgomery MC, Baloh J, Gu M, et al. Barriers and facilitators to HIV care engagement: recommendations from people with HIV in rural Arkansas. J Rural Health 2023 Mar 06;39(2):459-468 [FREE Full text] [doi: 10.1111/jrh.12717] [Medline: 36203209]
18. Mills EJ, Nachega JB, Bangsberg DR, Singh S, Rachlis B, Wu P, et al. Adherence to HAART: a systematic review of developed and developing nation patient-reported barriers and facilitators. PLoS Med 2006 Nov 21;3(11):e438 [FREE Full text] [doi: 10.1371/journal.pmed.0030438] [Medline: 17121449]
19. Yehia BR, Stewart L, Momplaisir F, Mody A, Holtzman CW, Jacobs LM, et al. Barriers and facilitators to patient retention in HIV care. BMC Infect Dis 2015 Jun 28;15(1):246 [FREE Full text] [doi: 10.1186/s12879-015-0990-0] [Medline: 26123158]
20. Drainoni ML, Baughman AL, Bachman SS, Bowers-Sword R, Davoust M, Fortu K, et al. Integrating community health workers into HIV care teams: Impact on HIV care outcomes. J HIV/AIDS Soc Serv 2020 Jun 30;19(3):204-219. [doi: 10.1080/15381501.2020.1785364]
21. Greuel M, Sy F, Bärnighausen T, Adam M, Vandormael A, Gates J, et al. Community health worker use of smart devices for health promotion: scoping review. JMIR Mhealth Uhealth 2023 Feb 22;11:e42023 [FREE Full text] [doi: 10.2196/42023] [Medline: 36811947]
22. Hsiao CH, Tang KY. Examining a model of mobile healthcare technology acceptance by the elderly in Taiwan. J Glob Inf Tech Manag 2015 Dec 11;18(4):292-311. [doi: 10.1080/1097198x.2015.1108099]
23. Pollard R, Kennedy CE, Hutton HE, Mulamba J, Mbabali I, Anok A, et al. HIV prevention and treatment behavior change and the Situated Information Motivation Behavioral Skills (sIMB) model: a qualitative evaluation of a community health worker intervention in Rakai, Uganda. AIDS Behav 2022 Feb 30;26(2):375-384 [FREE Full text] [doi: 10.1007/s10461-021-03391-w] [Medline: 34327597]
24. Hammack AY, Bickham JN, Gilliard I, Robinson WT. A community health worker approach for ending the HIV epidemic. Am J Prev Med 2021 Nov;61(5 Suppl 1):S26-S31 [FREE Full text] [doi: 10.1016/j.amepre.2021.06.008] [Medline: 34686287]
25. Kenya S, Chida N, Symes S, Shor-Posner G. Can community health workers improve adherence to highly active antiretroviral therapy in the USA? A review of the literature. HIV Med 2011 Oct;12(9):525-534 [FREE Full text] [doi: 10.1111/j.1468-1293.2011.00921.x] [Medline: 21518221]
26. Cabral HJ, Davis-Plourde K, Sarango M, Fox J, Palmisano J, Rajabiun S. Peer support and the HIV continuum of care: results from a multi-site randomized clinical trial in three urban clinics in the United States. AIDS Behav 2018 Aug 06;22(8):2627-2639. [doi: 10.1007/s10461-017-1999-8] [Medline: 29306990]
27. Hoff E, Hansen L, Pulitzer Z, Campalans N, Salyards M, Muquith M, et al. A randomized control trial of a combined community health worker and re-entry intervention for people with HIV recently released from jail who use substances. J Subst Use Addict Treat 2023 Sep;152:209118. [doi: 10.1016/j.josat.2023.209118] [Medline: 37454733]
28. Senn TE, Braksmajer A, Coury-Doniger P, Urban MA, Rossi A, Carey MP. Development and preliminary pilot testing of a peer support text messaging intervention for HIV-infected black men who have sex with men. J Acquir Immune Defic Syndr 2017 Feb 01;74 Suppl 2(Suppl 2):S121-S127 [FREE Full text] [doi: 10.1097/QAI.0000000000001241] [Medline: 28079722]
29. Dillingham R, Ingersoll K, Flickinger TE, Waldman AL, Grabowski M, Laurence C, et al. PositiveLinks: a mobile health intervention for retention in HIV care and clinical outcomes with 12-month follow-up. AIDS Patient Care STDS 2018 Jun;32(6):241-250 [FREE Full text] [doi: 10.1089/apc.2017.0303] [Medline: 29851504]

XSL•FO
RenderX

30. Rana AI, van den Berg JJ, Lamy E, Beckwith CG. Using a mobile health intervention to support HIV treatment adherence and retention among patients at risk for disengaging with care. AIDS Patient Care STDS 2016 Apr;30(4):178-184 [FREE Full text] [doi: 10.1089/apc.2016.0025] [Medline: 27028183]
31. Ventuneac A, Kaplan-Lewis E, Buck J, Roy R, Aberg CE, Duah BA, et al. A mobile health intervention in HIV primary care: supporting patients at risk for ART non-adherence. HIV Res Clin Pract 2020 Oct 28;21(5):140-150 [FREE Full text] [doi: 10.1080/25787489.2020.1862972] [Medline: 33369547]
32. Grimsrud A, Bygrave H, Doherty M, Ehrenkranz P, Ellman T, Ferris R, et al. Reimagining HIV service delivery: the role of differentiated care from prevention to suppression. J Int AIDS Soc 2016 Dec;19(1):21484. [doi: 10.7448/ias.19.1.21484]
33. Haberer JE, Kahane J, Kigozi I, Emenyonu N, Hunt P, Martin J, et al. Real-time adherence monitoring for HIV antiretroviral therapy. AIDS Behav 2010 Dec 31;14(6):1340-1346 [FREE Full text] [doi: 10.1007/s10461-010-9799-4] [Medline: 20809380]
34. Wood OR, Schnall R, Kay ES, Jia H, Abua JA, Nichols TK, et al. A community health worker and mobile health app intervention to improve adherence to HIV medication among persons with HIV: the CHAMPS study protocol. BMC Public Health 2023 May 24;23(1):942 [FREE Full text] [doi: 10.1186/s12889-023-15616-9] [Medline: 37226141]
35. Batey DSR, Anderson WA, Joseph D, Musgrove KE, Mugavero MJ, Raper JL. Birmingham Access 2 Care: Community partnership to improve public health. 2015 Presented at: 2015 Engaged Scholarship Conference; Sept 29-30, 2015; State College, Pennsylvania.
36. Schnall R, Sanabria G, Jia H, Cho H, Bushover B, Reynolds NR, et al. Efficacy of an mHealth self-management intervention for persons living with HIV: the WiseApp randomized clinical trial. J Am Med Inform Assoc 2023 Feb 16;30(3):418-426 [FREE Full text] [doi: 10.1093/jamia/ocac233] [Medline: 36469808]
37. Anti-Retroviral Treatment and Access to Services (ARTAS): an individual-level, multi-session intervention for people who are recently diagnosed with HIV Implementation Manual. Academy for Educational Development Center on AIDS & Community Health (COACH). URL: https://aidsetc.org/sites/default/files/resources_files/artas_implementation_manual.pdf [accessed 2025-03-20]
38. Guidelines for the use of antiretroviral agents in adults and adolescents with HIV. Department of Health and Human Service. URL: https://clinicalinfo.hiv.gov/sites/default/files/guidelines/documents/adult-adolescent-arv/guidelines-adult-adolescent-arv.pdf [accessed 2024-04-29]
39. HIV viral load. Cleveland Clinic. URL: https://my.clevelandclinic.org/health/diagnostics/hiv-viral-load [accessed 2024-04-29]
40. Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O'Neal L, REDCap Consortium. The REDCap consortium: building an international community of software platform partners. J Biomed Inform 2019 Jul;95:103208 [FREE Full text] [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]
41. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381 [FREE Full text] [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
42. Feldman BJ, Fredericksen RJ, Crane PK, Safren SA, Mugavero MJ, Willig JH, et al. Evaluation of the single-item self-rating adherence scale for use in routine clinical care of people living with HIV. AIDS Behav 2013 Jan 30;17(1):307-318 [FREE Full text] [doi: 10.1007/s10461-012-0326-7] [Medline: 23108721]
43. Johnson MO, Neilands TB, Dilworth SE, Morin SF, Remien RH, Chesney MA. The role of self-efficacy in HIV treatment adherence: validation of the HIV Treatment Adherence Self-Efficacy Scale (HIV-ASES). J Behav Med 2007 Oct 23;30(5):359-370 [FREE Full text] [doi: 10.1007/s10865-007-9118-3] [Medline: 17588200]
44. Schnall R, Cho H, Liu J. Health information technology usability evaluation scale (Health-ITUES) for usability assessment of mobile health technology: validation study. JMIR Mhealth Uhealth 2018 Jan 05;6(1):e4 [FREE Full text] [doi: 10.2196/mhealth.8851] [Medline: 29305343]
45. Lewis JR. Psychometric evaluation of the post-study system usability questionnaire: the PSSUQ. Proc Hum Factors Ergon Soc Annu Meet 1992 Oct 01;36(16):1259-1260. [doi: 10.1177/154193129203601617]
46. Improve performance, plan for productivity gains and accelerate innovation. SAS Communications. URL: https://www.sas.com/en_us/home.html [accessed 2025-03-20]
47. R Core Team. R: a language and environment for statistical computing. R Foundation. URL: https://www.r-project.org/ [accessed 2024-04-29]
48. Meyer DZ, Avery LM. Excel as a qualitative data analysis tool. Field Methods 2008 Sep 04;21(1):91-112. [doi: 10.1177/1525822x08323985]
49. Ose SO. Using excel and word to structure qualitative data. J Appl Soc Sci 2016 Aug 20;10(2):147-162. [doi: 10.1177/1936724416664948]
50. Morse JM. Critical analysis of strategies for determining rigor in qualitative inquiry. Qual Health Res 2015 Sep 16;25(9):1212-1222. [doi: 10.1177/1049732315588501] [Medline: 26184336]
51. O'Connor C, Joffe H. Intercoder reliability in qualitative research: debates and practical guidelines. Int J Qual Methods 2020 Jan 22;19:63. [doi: 10.1177/1609406919899220]
52. Konting MM, Kamaruddin N, Man NA. Quality assurance in higher education institutions: exist survey among Universiti Putra Malaysia graduating students. Int Educ Stud 2009 Feb 22;2(1):2009. [doi: 10.5539/ies.v2n1p25]

XSL•FO
RenderX

53. Sullivan GM, Feinn R. Using effect size-or why the P value is not enough. J Grad Med Educ 2012 Sep;4(3):279-282 [FREE Full text] [doi: 10.4300/JGME-D-12-00156.1] [Medline: 23997866]
54. Aschbrenner KA, Kruse G, Gallo JJ, Plano Clark VL. Applying mixed methods to pilot feasibility studies to inform intervention trials. Pilot Feasibility Stud 2022 Sep 26;8(1):217 [FREE Full text] [doi: 10.1186/s40814-022-01178-x] [Medline: 36163045]
55. O'Cathain A, Hoddinott P, Lewin S, Thomas K, Young B, Adamson J, et al. Maximising the impact of qualitative research in feasibility studies for randomised controlled trials: guidance for researchers. Pilot Feasibility Stud 2015;1:32 [FREE Full text] [doi: 10.1186/s40814-015-0026-y] [Medline: 27965810]
56. Operations manual for delivery of HIV prevention, care and treatment at primary health centres in high-prevalence, resource-constrained settings: edition 1 for field-testing and country adaptation. World Health Organization. 2008. URL: https://pubmed.ncbi.nlm.nih.gov/26290930/ [accessed 2024-04-29]
57. Dasgupta S, McManus T, Tie Y, Lin CY, Yuan X, Sharpe JD, et al. Comparison of demographic characteristics and social determinants of health between adults with diagnosed HIV and all adults in the U.S. AJPM Focus 2023 Sep;2(3):100115 [FREE Full text] [doi: 10.1016/j.focus.2023.100115] [Medline: 37790662]
58. Bennett B, Sharma M, Bennett R, Mawson AR, Buxbaum SG, Sung JH. Using social cognitive theory to predict medication compliance behavior in patients with depression in southern United States in 2016 in a cross-sectional study. J Caring Sci 2018 Mar 01;7(1):1-8 [FREE Full text] [doi: 10.15171/jcs.2018.001] [Medline: 29637050]
59. Sheeran P, Maki A, Montanaro E, Avishai-Yitshak A, Bryan A, Klein WM, et al. The impact of changing attitudes, norms, and self-efficacy on health-related intentions and behavior: a meta-analysis. Health Psychol 2016 Nov;35(11):1178-1188. [doi: 10.1037/hea0000387] [Medline: 27280365]
60. Loh KP, Liu J, Ganzhorn S, Sanabria G, Schnall R. Establishing a usability cut-point for the health information technology usability evaluation scale (Health-ITUES). Int J Med Inform 2022 Apr;160:104713 [FREE Full text] [doi: 10.1016/j.ijmedinf.2022.104713] [Medline: 35144102]
61. Banbury A, Nancarrow S, Dart J, Gray L, Parkinson L. Telehealth interventions delivering home-based support group videoconferencing: systematic review. J Med Internet Res 2018 Feb 02;20(2):e25 [FREE Full text] [doi: 10.2196/jmir.8090] [Medline: 29396387]
62. Cheng O. Utilization of telemedicine in addressing musculoskeletal care gap in long-term care patients. J Am Acad Orthop Surg Glob Res Rev 2020;4(4):e524. [doi: 10.1302/3114-210438]
63. Jong M, Kraishi M. A comparative study on the utility of telehealth in the provision of rheumatology services to rural and northern communities. Int J Circumpolar Health 2004 Dec;63(4):415-421 [FREE Full text] [doi: 10.3402/ijch.v63i4.17758] [Medline: 15709316]
64. Gillette E, Naanyu V, Nyandiko W, Chory A, Scanlon M, Aluoch J, et al. HIV-related stigma shapes research participation for youth living with HIV in Kenya. J Int Assoc Provid AIDS Care 2023 Apr 26;22:23259582231170732 [FREE Full text] [doi: 10.1177/23259582231170732] [Medline: 37101381]
65. Marellapudi A, Hussen SA, Brown DN, Fletcher MR, Henkhaus ME, Jones MD, et al. Understanding and addressing privacy and confidentiality concerns associated with the provision of mobile HIV care: a qualitative study. AIDS Care 2022 May 03;34(5):575-579 [FREE Full text] [doi: 10.1080/09540121.2021.1921104] [Medline: 33938335]
66. Taggart T, Grewe ME, Conserve DF, Gliwa C, Roman Isler M. Social media and HIV: a systematic review of uses of social media in HIV communication. J Med Internet Res 2015 Nov 02;17(11):e248 [FREE Full text] [doi: 10.2196/jmir.4387] [Medline: 26525289]
67. Audi C, Jahanpour O, Antelman G, Guay L, Rutaihwa M, van de Ven R, et al. Facilitators and barriers to antiretroviral therapy adherence among HIV-positive adolescents living in Tanzania. BMC Public Health 2021 Dec 13;21(1):2274 [FREE Full text] [doi: 10.1186/s12889-021-12323-1] [Medline: 34903209]
68. Zurbachew Y, Hiko D, Bacha G, Merga H. Adolescent's and youth's adherence to antiretroviral therapy for better treatment outcome and its determinants: multi-center study in public health facilities. AIDS Res Ther 2023 Dec 19;20(1):91 [FREE Full text] [doi: 10.1186/s12981-023-00588-y] [Medline: 38115098]
69. Hebel S, Kahn-Woods E, Malone-Thomas S, McNeese M, Thornton L, Sukhija-Cohen A, et al. Brief report: discrepancies between self-reported adherence and a biomarker of adherence in real-world settings. J Acquir Immune Defic Syndr 2020 Dec 01;85(4):454-457. [doi: 10.1097/QAI.0000000000002486] [Medline: 33136744]
70. Kagee A, Nel A. Assessing the association between self-report items for HIV pill adherence and biological measures. AIDS Care 2012 Jun 07;24(11):1448-1452. [doi: 10.1080/09540121.2012.687816] [Medline: 22670758]
71. Patel T, Ivo J, Faisal S, McDougall A, Carducci J, Pritchard S, et al. A prospective study of usability and workload of electronic medication adherence products by older adults, caregivers, and health care providers. J Med Internet Res 2020 Jun 02;22(6):e18073 [FREE Full text] [doi: 10.2196/18073] [Medline: 32348292]
72. Ping Y, Visaria A, Suppiah SD, Tan YW, Malhotra R. Prevalence and correlates of medication reminder app 'use and use intention' among older adults. Explor Res Clin Soc Pharm 2022 Jun;6:100150 [FREE Full text] [doi: 10.1016/j.rcsop.2022.100150] [Medline: 35755719]

XSL•FO
RenderX

73. Marconi A, Balestrieri M, Comastri G, Pulvirenti F, Gennari W, Tagliazucchi S, et al. Evaluation of the Abbott real-time HIV-1 quantitative assay with dried blood spot specimens. Clin Microbiol Infect 2009 Jan;15(1):93-97 [FREE Full text] [doi: 10.1111/j.1469-0691.2008.02116.x] [Medline: 19220340]
74. Pannus P, Claus M, Gonzalez MM, Ford N, Fransen K. Sensitivity and specificity of dried blood spots for HIV-1 viral load quantification: a laboratory assessment of 3 commercial assays. Medicine (Baltimore) 2016 Nov;95(48):e5475 [FREE Full text] [doi: 10.1097/MD.0000000000005475] [Medline: 27902602]
75. Vojnov L, Carmona S, Zeh C, Markby J, Boeras D, Prescott MR, DBS for VL Diagnostics Investigation Consortium. The performance of using dried blood spot specimens for HIV-1 viral load testing: a systematic review and meta-analysis. PLoS Med 2022 Aug 22;19(8):e1004076 [FREE Full text] [doi: 10.1371/journal.pmed.1004076] [Medline: 35994520]

Abbreviations

aOR: adjusted odds ratio
ART: antiretroviral therapy
CHAMPS: Community Health Worker and mHealth to Improve Viral Suppression
CHW: community health worker
DBS: dried blood spot
eROI: electronic release of information
Health-ITUES: Health Information Technology Usability Evaluation Scale
HIPAA: Health Insurance Portability and Accountability Act
HIV-ASES: HIV Treatment Adherence Self-Efficacy Scale
mHealth: mobile health
MHTAM: Mobile Health Technology Acceptance Model
PSSUQ: Poststudy System Usability Questionnaire
REDCap: Research Electronic Data Capture
SRSI: self-reported scale item

Edited by A Mavragani; submitted 25.10.24; peer-reviewed by AM Navarra, C O'Connor; comments to author 09.01.25; revised version received 03.02.25; accepted 05.03.25; published 02.04.25.

Please cite as:
Shourya S, Liu J, McInerney S, Casimir T, Kenniff J, Kershaw T, Batey D, Schnall R
A Remote Intervention Based on mHealth and Community Health Workers for Antiretroviral Therapy Adherence in People With HIV: Pilot Randomized Controlled Trial
JMIR Form Res 2025;9:e67997
URL: https://formative.jmir.org/2025/1/e67997
doi:10.2196/67997
PMID:

©Shivesh Shourya, Jianfang Liu, Sophia McInerney, Trinity Casimir, James Kenniff, Trace Kershaw, David Batey, Rebecca Schnall. Originally published in JMIR Formative Research (https://formative.jmir.org), 02.04.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Assessing the Cultural Fit of a Digital Sleep Intervention for Refugees in Germany: Qualitative Study

Maja Blomenkamp[1], MSc; Andrea Kiesel[2], Prof Dr; Harald Baumeister[3], Prof Dr; Dirk Lehr[4], Prof Dr; Josef Unterrainer[5], Prof Dr; Lasse B Sander[5], PhD; Kerstin Spanhel[5], PhD

[1]Institute of Psychology, University of Freiburg, Freiburg, Germany
[2]Cognition, Action, and Sustainability Unit, Institute of Psychology, University of Freiburg, Freiburg, Germany
[3]Department of Clinical Psychology and Psychotherapy, Institute of Psychology and Education, Ulm University, Ulm, Germany
[4]Health Psychology and Applied Biological Psychology, Institute of Sustainability Psychology, Leuphana University Lüneburg, Lüneburg, Germany
[5]Institute of Medical Psychology and Medical Sociology, Faculty of Medicine, University of Freiburg, Freiburg, Germany

Corresponding Author:
Kerstin Spanhel, PhD
Institute of Medical Psychology and Medical Sociology
Faculty of Medicine
University of Freiburg
Hebelstr. 29
Freiburg, D-79104
Germany
Phone: 49 761 203 5530
Email: kerstin.spanhel@mps.uni-freiburg.de

Abstract

Background: Digital interventions have been suggested to facilitate access to mental health care for refugees, who experience structural, linguistic, and cultural barriers to mental health care. Sleep-e, a digital sleep intervention originally developed for German teachers, has been culturally adapted for refugees in Germany mainly coming from African and Middle East countries. With the increasing number of refugees from Ukraine and the associated diversity of cultural backgrounds among refugees in Germany, it is essential to assess whether existing digital interventions are culturally appropriate for this target group as well.

Objective: The study aimed to investigate the perceived cultural appropriateness of Sleep-e in both its original and culturally adapted versions among refugees in Germany, hereby exploring and possibly contrasting the needs of refugees coming from Ukraine and other countries of origin.

Methods: Overall, 13 refugees (6 from Ukraine, 23-66 years old; and 7 from other countries, 26-41 years old) participated in the study. Each participant went through parts of the original or culturally adapted version of the digital sleep intervention, with 5 participants going through both versions. A total of 17 semistructured interviews (11 for the adapted, 6 for the nonadapted intervention version) and 9 think-aloud sessions (6 for the adapted, 3 for the nonadapted intervention version) were conducted to assess cultural appropriateness, suggestions for adaptations, and perceived relevance. Data were transcribed, categorized, and analyzed using structured qualitative content analysis.

Results: The findings showed key differences in the perceived appropriateness and identification between the 2 refugee groups and the 2 intervention versions. Ukrainian participants expressed positive (n=70) and negative (n=56) feedback on the adapted intervention version, which revealed identity conflicts, as the adapted intervention version was targeted at a refugee population with whom they could not fully identify (18 negative feedback quotes concerning the refugee example characters). Whereas they identified with the European context in the original version, they found the problems described less relevant to their experiences. In contrast, participating refugees from other countries found the culturally adapted version more comprehensible and culturally appropriate (55 positive and 5 negative feedback quotes). No significant usability issues were reported, but several participants highlighted the need for an individualization of the intervention content.

Conclusions: Neither the original nor culturally adapted version of the digital sleep intervention fully met the needs of all refugee groups, highlighting the complexity of culturally adapting digital interventions for this population. Particularly, the identity conflict of participating Ukrainian refugees regarding the refugee context suggests that adaptation should go beyond regional considerations and consider the dynamics of social identity. These findings emphasize the relevance of including co-design

XSL•FO
RenderX

processes with different refugee populations to ensure broad identification and, herewith, cultural appropriateness of digital interventions.

Trial Registration: German Clinical Trials Register DRKS00036484; https://drks.de/search/de/trial/DRKS00036484

(JMIR Form Res 2025;9:e65412) doi:10.2196/65412

KEYWORDS

Ukraine; eHealth; sleep disturbances; low-threshold treatment; culturally sensitive treatment; refugee; digital sleep; Germany; digital intervention; interview; content analysis; qualitative study; mental burden; mental health care; electronic health; digital health

Introduction

The number of people forced to seek refuge in new countries is increasing worldwide [1]. These people, hereafter referred to as refugees, are exposed to multiple stressors before, during, and after leaving their home countries, and they therefore experience increased psychological vulnerability and prevalence of mental disorders [2,3], including sleep problems [4]. However, there are major gaps in mental health care for refugees in the receiving countries such as Germany [5,6], resulting from a lack of (culturally sensitive) services, language barriers, differences in help-seeking behavior and residence status, as well as stigmatization, among other barriers [7-9]. To overcome these barriers, culturally adapted digital interventions focusing on transdiagnostic or somatically perceived symptoms are being discussed and investigated as low threshold offers for refugees [10,11]. In the general population, digital interventions show good effectiveness for a wide range of mental health conditions [12,13]. Their usability appears to be essential in this context [14,15]. When offered to refugees, in addition to intervention usability, cultural adaptation is suggested to be essential for the acceptance and effectiveness of face-to-face and digital interventions, recognizing the specific burdens associated with the refugee context [16-18]. In this context, the question of the cultural specificity of the adaptations is becoming increasingly important [19]. The target group of refugees is characterized by a diversity of countries of origin that is constantly changing due to events in the global political situation, such as the start of the war in Ukraine in 2022. To increase refugees with diverse backgrounds using and profiting from digital interventions, these interventions need to be provided in a culturally sensitive (ie, with an openness to diverse backgrounds)—in contrast to a culturally specific (ie, separately for diverse backgrounds)—way [20]. Yet, so far, there are only a few digital interventions developed for refugees from diverse countries of origin [21-23].

Based on feedback from refugees in Germany, who at the time were mainly coming from African and Middle Eastern countries of origin, the digital sleep intervention Sleep-e was culturally adapted for this target group with the aim of making it accessible to people sharing the experience of flight. Thereby, (1) factors related to the flight (eg, separation from the family and the home country) and (2) factors related to the habits, values, and treatment concepts of their non-Western countries of origin were considered [24]. The culturally adapted version of the digital sleep intervention was accepted by refugees from African and Middle Eastern countries [22]. Based on the increasing number of refugees in Germany coming from Ukraine since the beginning of the war in Ukraine in 2022 [25], it seems essential to make such interventions available to this target group as well. Ukrainian refugees are shown to differ from refugees from other countries of origin, among others, in terms of sociodemographic aspects (age and gender), sociocultural aspects, differences in the flight route, legal rights, and welcoming culture upon arrival in Germany [26-29]. Building on the named 2 previous studies [22,24], this qualitative study aimed to investigate differences in the perceived cultural appropriateness and the cultural needs in regard to the digital sleep intervention among refugees from Ukraine and refugees from other, predominantly African and Middle Eastern countries of origin. A culturally adapted version and a nonadapted version of the intervention were used, and factors influencing the perceived cultural appropriateness were additionally investigated.

Methods

Recruitment and Participants

This study was conducted in 2 subsamples: refugees from Ukraine and refugees from countries other than Ukraine. Recruitment was carried out from February to April 2023 via various institutions and associations working in the refugee and migration context in Germany, mainly Freiburg and Karlsruhe. Many recruitment inquiries remained unanswered (eg, integration and language courses, migration offices, advice centers, and advertising via social media). Recruitment mainly succeeded via Refugium Freiburg (psychosocial advice for refugees), the German-Ukrainian Choir Karlsruhe, the University of Freiburg, the radio project “Our Voice” for refugees in Freiburg, and “Kontaktcafé Freiburg” (get-together café for refugees), as well as via personal contacts. Institutions were contacted consecutively face-to-face, via phone, or via email. They forwarded a flyer with information on the study to their clients face-to-face. People interested in study participation could contact the study team via email or phone and were then forwarded the detailed study information. Inclusion criteria were checked, which included the experience of flight, age of majority, and sufficient knowledge of German or English. No cutoff for insomnia severity was used as an inclusion criterion in order to also reach refugees with subclinical symptoms, who could be a key target group for the transdiagnostic and low-threshold digital sleep intervention. In addition, this allows information to be obtained on who might be interested in the intervention, regardless of existing clinical symptoms. Contacted participants who did not want to participate in the study mainly reported reasons such as language barriers or difficulties in

XSL•FO
RenderX

finding the time for study participation. None of the participants dropped out after signing the informed consent.

There were 6 participants from Ukraine (23-66 years old), and 7 participants coming from Afghanistan, Gambia, Syria, Madagascar, Cameroon, Ecuador, and Eritrea (26-41 years old), referred to below as refugees from other countries. On average, the participants revealed subthreshold sleep problems. Detailed sociodemographic data are shown in Table 1.

XSL•FO
RenderX

Table 1. Sociodemographic data and information on the conduction of the qualitative study in Germany in 2023 are illustrated for each participating refugee from Ukraine (ukr) or other (oth) countries of origin.

Subgroup and ID	Home country (gender)	Education	Age (years)	Years in Germany	Psychotherapy (current or past)	Insomnia severity[a]	Tested modules	Study mode (study language)	Duration (min)	Codes, n
Ukraine										
01_ukr	Ukraine (female)	High school	45	1	No	8	na[b]1: qtn[c], int[d] a[e]1: ThA[f], int	In person (German)	79	154
02_ukr	Ukraine (female)	High school	45	1	No	15	na2: qtn, ThA, int a2: int	In person (German)	53	57
03_ukr	Ukraine (female)	High school	66	1	No	7	na3: qtn, ThA, int a3: int	In person (English)	57	105
04_ukr	Ukraine (female)	High school	26	1	Yes	15	a2: qtn, ThA, int	Video call (English)	42	65
05_ukr	Ukraine/Nigeria (male)	High school	23	1	No	15	a3: qtn, ThA, int	In person (English)	63	108
06_ukr	Ukraine (female)	High school	33	0.5	Yes	19	a4: qtn, int	Video call (English)	15	42
Total, mean (SD)	—[g]	—	39.7 (15.9)	0.9 (0.2)	—	13.2 (4.7)	—	—	—	—
Other countries										
07_oth	Afghanistan (male)	High school	41	8	No	9	na1: qtn, int	Video call (German)	23	18
08_oth	Gambia (male)	No degree	27	3	Yes	18	na2: qtn, int a3: int	In person (English)	45	71
09_oth	Syria (male)	High school	26	7.5	Yes	8	na3: qtn, ThA, int a4: ThA, int	In person (German)	74	125
10_oth	Madagascar (male)	High school	30	10	Yes	18	a1: qtn, ThA, int	In person (French)	37	31
11_oth	Cameroon (female)	High school	33	1.5	No	6	a2: qtn, int	Video call (French)	35	13
12_oth	Ecuador (male)	High school	32	0.5	No	15	a3: qtn, ThA	In person (German)	33	18
13_oth	Eritrea (male)	No degree	31	8	Yes	12	a4: qtn, int	Video call (German)	11	23
Total, mean (SD)	—	—	31.4 (4.9)	5.5 (3.7)	—	12.3 (4.9)	—	—	—	—

[a]Insomnia Severity Scores [30]: 0-7, not clinically significant; 8-14, subthreshold; 15-21, clinical (moderate); 22-28, clinical (severe).

[b]na: nonadapted version of the digital sleep intervention (modules 1-3).

[c]qtn: questionnaire.

[d]int: interview.

[e]a: adapted version of the digital sleep intervention (modules 1-4).

[f]ThA: think-aloud.

[g]Not applicable.

XSL•FO
RenderX

Procedure

Data collection took place in one-to-one sessions (approximately 90 minutes of video-based or in-person, one-to-one meetings at the Institute of Medical Psychology and Medical Sociology, University of Freiburg, Germany) with MB, who was in the final year of her master's degree in clinical psychology at the time of study conduction and had previous experience in qualitative research (refer to Table 1 for further information). MB had previous contact with one participant (01_ukr); the other participants were unknown to her. The participants were informed that the research was part of MB's master thesis. Sessions were conducted in German, English, or French and were audio recorded. All participants completed a web-based questionnaire (sociodemographic data, Insomnia Severity Index with 7 Likert-scale items 0-4 [30], and Client Satisfaction Questionnaire adapted to internet-based interventions with 8 Likert-scale items 1-4 [31]). To assess the perceived cultural appropriateness of the respective interventions as part of the usability test, it was considered sufficient to work through individual modules of the adapted versus nonadapted intervention: many aspects relevant to the assessment are similar across the respective adapted versus nonadapted modules (eg, number of pictures, use of text versus video, and example characters with a German versus another cultural background). Following this assumption, the participants worked through one adapted or nonadapted module. They were then offered the opportunity to test a second (nonadapted or adapted) module. Three participants from Ukraine and 2 participants from other countries completed both an adapted and a nonadapted module. When assigning the modules, it was ensured that each module was seen at least once by participants from each subgroup. After each module completion, semistructured interviews were conducted in order to assess the participants' experiences with the content and design of the tested module, its appropriateness for themselves and refugees in general, and their own usage behavior (refer to Multimedia Appendix 1 for the interview guideline; as the interview guideline was shown to be appropriate in a first pilot interview, the pilot was included in the final results). The think-aloud method [32] was additionally used with 8 participants (5 participants from Ukraine, and 3 from other countries), depending on the language skills and time resources of the respective participants. Using this method, the participants were asked to speak out loud all of the thoughts they had while they went through the modules; for example, they could comment on the experienced design, functionality, or content of the intervention. Throughout data collection, the investigator made observational notes on verbal or nonverbal characteristics during the test. Table 1 shows in detail which modules were processed by which participants and in which format (interview or think-aloud) the data were collected.

Intervention

Two versions of the digital sleep intervention Sleep-e were used, which is a digital, brief cognitive behavioral intervention for sleep problems available in English and German. The intervention was provided on the eHealth platform eSano. The nonadapted version [24] is an unguided digital intervention consisting of 3 content-separated, consecutive modules based on the intervention GET.ON Recovery [33,34], which was developed and evaluated for German employees with sleep problems. Adaptations were conducted based on feedback from health care professionals working with refugees as well as refugees from Algeria, Eritrea, Iran, Iraq, and Syria [24], complemented by previously conducted cultural adaptations of digital interventions [35]. The resulting culturally adapted version consists of 4 content-separated, consecutive modules and was aimed at being suitable for refugees from diverse countries of origin in Germany, with the idea of a shared experience of pre-, peri-, and postmigration stressors. Its feasibility, acceptance, and preliminary effectiveness were investigated among refugees mainly coming from African and Middle East countries [22]. Table 2 provides information on the content of both versions of the intervention. Information on the adaptations suggested [24] and conducted [22] can be found in the respective articles.

XSL•FO
RenderX

Table 2. Overview of the modules of both an adapted and a nonadapted version of a digital Sleep-e intervention tested in the qualitative study with refugees in Germany.

Module	Content of the nonadapted version [24]	Content of the culturally adapted version [22]
Module 1	• Introduction of example characters from Germany ("text-based"[a] with photo) • Psychoeducation and exercise on sleep hygiene rules ("text-based", quiz) • Psychoeducation on sleep medication ("text-based") • Optional: sleep diary ("offline document")	• Introduction by a health expert and for example characters from Arabic-speaking countries ("video-based") • Psychoeducation on sleep problems, link to "migration" ("video-based") • Optional: relaxation exercise (audio-based) • Optional: sleep diary ("short web-based document")
Module 2	• Psychoeducation on the topic of rumination ("eg, explanatory video") • “3” exercises on the topic of rumination (audio- and text-based)	• Psychoeducation and "optional" exercise on sleep hygiene ("mainly video-based") • "Optional": psychoeducation on sleep medication ("text- and video-based") • "Optional": relaxation exercise ("audio-based") • "Optional": sleep diary ("short web-based document")
Module 3	• Reflection on the results, consolidation of what has been learned (text-based) • Planning of future exercises ("text-based") • Information on further offers of help ("text-based") • Psychoeducation on problems associated with poor sleep ("text-based")	• Psychoeducation on the topic of rumination, focus on migration topics (mainly video-based) • “Optional”: 2 exercises on the topic of rumination (audio- and text-based) • “Optional”: exercise on sleep hygiene rules • “Optional”: relaxation exercise (audio-based) • “Optional”: sleep diary
Module 4	__[b]	• “Optional”: psychoeducation on problems associated with poor sleep (mainly video-based) • Information on further offers of “refugee-specific” help • “Optional”: exercise on sleep hygiene rules • “Optional”: relaxation exercise (audio-based) • “Optional”: reflection on results • “Optional”: sleep diary

[a]Differences between the adapted and the nonadapted version of the digital sleep intervention are in double quotes.

[b]Not applicable.

Data Analysis

Before analyzing data, the authors involved in the analysis (MB and KS) reflected on their own assumptions regarding the topic of the study. The following assumptions were identified: (1) refugees are often not well off and are grateful for any form of support and (2) Ukrainian refugees and refugees from other countries of origin differ in their needs. The participants were not involved in data transcription, analysis, or interpretation.

The collected data comprised 26 audio recordings (9 think-aloud sessions, including one participant who did 2 sessions; 17 interviews, including 5 participants who did 2 interviews). The recordings were automatically transcribed using Adobe Premiere (2021) and postprocessed by MB based on the transcription rules for computer-assisted analysis according to Kuckartz and Rädiker [36]. The data were analyzed using structured qualitative content analysis (primary analysis [36,37]) with a combination of content-related, evaluative, and contrasting categories [38]. MAXQDA 2022 [39] software was used to code a total of 830 text passages. The categories were created and continuously adapted by MB using an inductive and deductive hybrid approach [38], primarily based on the transcripts of the Ukrainian participants. This enabled the deductive inclusion and validation of theories and questions derived from the two previous studies, as well as the inductive inclusion of new aspects (eg, in regard to the war in Ukraine) introduced by the participants. After coding half of the transcripts (ie, 13), the new information could be integrated into the previously formed categories and no further changes were made to the structure of the category system, which indicates data saturation. In order to assess intercoder reliability and ensure the comprehensibility of the category definitions, 2 transcripts were additionally coded and discussed by KS (researcher at the time of study conduction). The different category formats were used to identify possible group differences (contrasting categories), evaluations (evaluative categories), and factors influencing the evaluations (content-related categories), as well as their interactions. As part of the secondary analysis, valence analyses [37] were used to compare the ratings of the intervention parts based on the evaluative categories (positive vs negative feedback regarding the adapted vs nonadapted intervention or regarding refugees from Ukraine vs from other countries of origin). In addition, the coding frequencies by categories, subgroups, and interventions were compared in frequency analyses, and overlaps between evaluative categories and contrasting or content-related categories were examined in group comparison. Such frequency analyses can provide insight into possible correlations but are of limited scalability due to the lack of objectivity of the number of codes. The qualitative results are shown along the developed category system, underpinned by quotes from the participants. The references of the quoted text passages comprise the number

XSL•FO
RenderX

of participants (01-13), the subgroup (coming from Ukraine [denoted as “ukr”] vs from other countries of origin [denote at “oth”]), the modules processed (modules 1-4 of the adapted version of the intervention [denoted from “a1” to “a4”], modules 1-3 of the nonadapted version of the intervention [denoted from “na1” to “na3”]), the type of assessment (think-aloud [denoted as “ThA”], and interview [denoted as “int”]) and the position in the transcript (denotes as “pos.”). The quantitative data from the questionnaires was analyzed descriptively.

Ethical Considerations

The study was registered in the Freiburg Clinical Trials Registry (registration FRKS004288). Ethics approval was granted by the ethics committee of the University of Freiburg, Germany (22-1452-S2). The study was conducted and documented in accordance with the guidelines for research with refugees [40] and the COREQ (Consolidated Criteria for Reporting Qualitative Research) [41] (refer to Multimedia Appendix 2 for the checklist). Interested participants received detailed digital, written information on the study procedure, aim, and their rights. Afterward, they provided their digital informed consent before the start of the study or had the option of withdrawing study participation without any consequences. The protection of the privacy and confidentiality of the participants was ensured by pseudonymizing all collected data with an ID. Only the pseudonymized data were used during the study, and the data were only accessible to selected project members. A reference list, which linked personal information with the ID, was deleted at the end of the data collection. This means that all data collected (transcripts or questionnaires) are de facto anonymized, and it is not possible to draw any conclusions from these data about the individual study participants. The de facto anonymized data of the participants are stored by the researchers on secure servers for 10 years after publication and then deleted. The participants had the chance to win one of five €20 (€1=US $1.10) gift vouchers as compensation for their efforts.

Results

Overview

The developed category system (refer to Multimedia Appendix 3) resulted in 6 main and 14 subcategories. In total, 4 main categories informed the first research question and contained information on the perceived cultural appropriateness and the cultural needs among the participating refugees from Ukraine versus those from other countries of origin regarding the digital sleep intervention: (1) concept: adapted version, (2) concept: nonadapted version, (3) comparison: adapted and nonadapted version, and (4) use of the intervention. The developed category system furthermore resulted in 2 main categories informing about factors that influenced the perceived cultural appropriateness and needs of the participating refugees regarding the digital sleep intervention: (1) identity and (2) health care.

Perceived Cultural Appropriateness of the Digital Sleep Intervention

Overview

Across both versions of the intervention and both subgroups, there was more positive (n=202) than negative feedback (n=80; see Figure 1).

Figure 1. Frequency analyses of positive and negative feedback given by the refugees from (A) Ukraine and (B) other countries of origin participating in the qualitative study in Germany. The numbers of positive and negative feedback are illustrated separately for the adapted version and the nonadapted version of the digital intervention Sleep-e tested in the study. oth: other countries of origin; ukr: Ukraine.

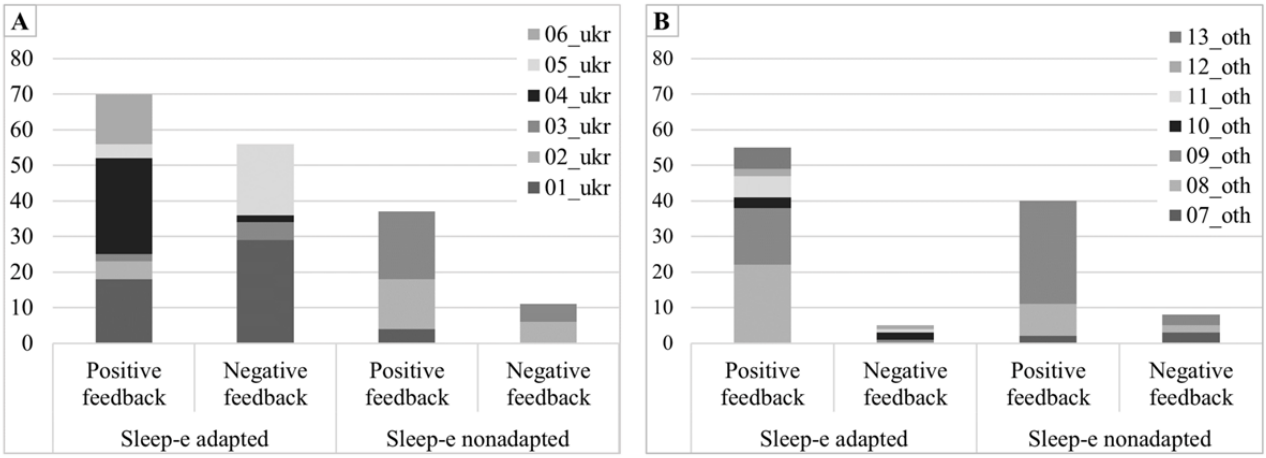

> *You can take a lot from it also. And you can learn easily also.* [08_oth_a3(na2)_int, pos. 211]
>
> *To have the training..., it’s much easier. Yeah, and it could solve your problem.* [06_ukr_a4_int, pos. 60]

An overall positive attitude also was revealed in the relatively high ratings in the client satisfaction questionnaire adapted to internet-based interventions, with an overall mean score of 26.0 (SD 4.7) and similar mean scores for the different versions or subgroups.

Adapted Intervention Version

The adapted version of the intervention was rated positively in both groups, but the feedback from the Ukrainian participants revealed a rather heterogeneous picture, with 70 positive and 56 negative feedback quotes, contrasting with 55 positive and 5 negative feedback quotes among participants from other countries. With regard to the conception of the intervention, linguistic simplifications (eg, subtitles), the video formats and quizzes, as well as good structuring and feasibility were

XSL•FO
RenderX

considered particularly helpful. Both groups criticized the quality of the videos as well as linguistic aspects such as the length of the texts and modules or language difficulties. As for the content of the intervention version, the experience reports from the example characters (refugees) in particular were described as motivating and interesting and the exercise content as helpful. Differences in the perceived appropriateness of the content between the 2 subgroups were found with regard to the identification with the example characters: whereas the participating refugees from other countries of origin were able to recognize and name similarities with the example characters, the participants from Ukraine did not identify with the example characters and described them as untrustworthy or boring ("I don't believe them that the exercises help." [01_ukr_a1na1_int, pos. 102, translated]). In addition, the participants from Ukraine criticized the strong focus on the topic of flight, too negative content, and too little concrete advice. Suggestions for changes to the adapted version of the intervention included ideas for increasing identification, for example, through the opportunity to create a personalized profile with one's own needs and issues ("Maybe bubbles and then it's like: hit which bubble you feel speaks to you." [05_ukr_a3_int, pos. 336]; "Create your own profile and then really see: voilà, that's me." [10_oth_a1_(ThA)int, pos. 147, translated]). According to the participants (especially those from Ukraine), this could also be achieved through a greater variety or abstractness (eg, by using graphics instead of real videos) of the example characters.

Nonadapted Intervention Version

The nonadapted version of the intervention was also rated positively in both groups (37 positive and 11 negative feedback quotes from the Ukrainian participants; 40 positive and 8 negative feedback quotes from participants from other countries). With regard to the conception of the intervention, its simple design and the clear, motivating, and comprehensible structure were mentioned. Both groups criticized the difficult language and an oversupply of text-based information. As for the content of the intervention version, the experience reports of the example characters (Germans) and the exercises and questions were described as helpful, as was the combination of theoretical and practical aspects ("So the first [nonadapted] version was technical for me. Going to see a therapist and answering questions I could fully understand and that was professional I would say." [01_ukr_a1na1_int, pos. 6, translated]). However, the content of this version of the intervention was partly considered too scientific or insufficient for more serious problems by the participants from Ukraine ("I don't see here any problems, serious problems. Just, time to bed, stop having alcohol, stop working after 9. But people who are under stress, need the more serious program." [03_ukr_na3_int, pos. 123]), whereas the participants from other countries of origin mainly mentioned identification difficulties ("Except for alcohol and so on, because I'm a Muslim and I don't drink alcohol. Yes, and the norm in the Arab-Islamic world is completely different. It's practically a sin, it's an absurdity, alcohol is for losers or sinners, or I don't know, and not for normal people." [09_oth_na3a4_int, pos. 57, translated]). Some statements of the Ukrainian participants did not include themselves when pointing out possible limitations of the cultural appropriateness for refugees ("But not all people can read so much text at once and the language is also a bit too scientific, like from a psychological book [...] For me it was okay." [02_ukr_na2_ThA_int_a3_int, pos. 39, translated]). Suggestions for changes to the nonadapted version of the intervention included a more solution-oriented content and the presentation of more appropriate problems or situations ("I would yes, more different [example characters], for example in terms of age." [07_oth_na1_int, pos. 111, translated]).

Factors Influencing the Perceived Cultural Appropriateness

As a main factor influencing the perceived cultural appropriateness of the intervention, identification was extracted. In this context, the participants emphasized the need to enhance identification with the intervention by including possible differences in the refugee contexts as well as cultural and linguistic differences in the intervention content.

Identification With the Refugee Context

Factors related to the refugee background and the identification with it were shown to influence the perceived appropriateness of the interventions among both groups of participating refugees. Some of these factors were related to their own flight experiences ("Especially when they came from Ukrainian zones after the bombing, after devastation, ruin, these bad things. They ask to get psychological treatment or to see a psychiatrist because they realize they need it." [03_ukr_na3_int, pos. 72]; "That is another traumatic experience. Knowing that crossing the sea, some people die, some people survive." [08_oth_a3(na2)_int, pos. 153]).

Yet, a separation of the participating Ukrainian refugees from being a refugee became apparent, revealing an underlying identity conflict (identity of being a refugee): "They don't identify as refugees, so in a way that [experience reports from the adapted version of the intervention] would even be a disconnect for them." [05_ukr_a3_int, pos. 246]; "Just I lived in prosperity [...] I have everything. And I fled from Ukraine the second day of war and didn't see all those bombings, devastations, and so on and so forth." [03_ukr_a3_(ThA)int, pos. 11]. Along with this, the Ukrainian participants emphasized that they did not want to be confronted with the topic of flight ("If I were a refugee, I wouldn't want to focus on it...You have to forget that you are a refugee." [01_ukr_a1na1_int, pos. 243, translated]).

Along with the identification as a refugee, the question of how far the identity of being a refugee is compatible with the Ukrainian identity was revealed, and whether refugees from Ukraine can identify with refugees from other countries of origin (same vs different identity). Thereby, statements about the same versus different identities were repeatedly raised (Textbox 1).

XSL•FO
RenderX

Textbox 1. Statements by the participating refugees in the qualitative study conducted in Germany about the same versus different identities of refugees from Ukraine (ukr) and refugees from other countries of origin (oth).

- "You have a world with Ukrainians inside who do not just view fleeing the same, you know, it's the same war, but at the same time, it is not the same (laughs)." [05_ukr_a3_int, pos. 252]
- "There are universal factors of flight and war." [09_oth_na3a4_int, pos. 306, translated] versus "It is never the same experience." [05_ukr_a3_int, pos. 300]
- "The people that came from the other places also, they face even more difficulties than the people that came from Ukraine." [08_oth_na2a3_ThA_int, pos. 353] versus "First-degree refugees, second-degree refugees...as if they were now downgraded, that is, with the refugees who are now less privileged." [09_oth_na3a4_int, pos. 342, translated]
- "They have a different pronunciation, different skin color, different way of speaking and different challenges than we do." [09_oth_na3a4_I, pos. 336, translated] versus "They are also refugees and they have had all these problems." [01_ukr_a1_T, pos. 432, translated]

The dilemma of same versus different identity also became evident in various statements regarding a possible identification of the participating refugees from Ukraine with the example characters in the adapted version of the intervention: "When I was listening to these people, I did not, I do not know their situation." [03_ukr_a3_(ThA)int, pos. 74] versus "I could relate. I understand his problem and I think that I have the same problem." [04_ukr_a2_ThA, pos. 127] versus "Oh, [for refugees] from Ukraine. No." [05_ukr_a3_int, pos. 240]. Various conflicting issues were found in their statements about the example characters (Table 3).

Table 3. Statements of the participating refugees from Ukraine can illustrate an ambivalent view on the example characters of the adapted version of the digital sleep intervention tested in the qualitative study in Germany.

Affirming perspective		Contrasting perspective
"To be honest, it is better for refugees." [01_ukr_a1na1_int, pos. 6, translated]	↔	"Not for me. [laughs] For me, I don't know, I understand that it's a completely different culture." [01_ukr_a1_ThA, pos. 432, translated]
"Of course, I don't have time to have this diary, because it's for people who are under stress and under problems...who have serious problems and they want to recover. They will do that." [03_ukr_a3_(ThA)int, pos. 116]	↔	"Hm no, for me, stress is the problem number one...Too much stress, for sure. It can have a negative impact on the sleeping quality. Yeah, you can't sleep because you think how to solve this problem." [03_ukr_na3_int, pos. 18 // 03_ukr_na3_ThA, pos. 189]
"You know, it [experience reports from refugees]'s something you can relate to, you know." [05_ukr_a3_int, pos. 234]	↔	"Even though I did not flee to Europe, like through the Mediterranean or the Sahara, fleeing the war in Ukraine, I don't really think it speaks to me." [05_ukr_a3_int, pos. 198]
"And besides: Being a refugee is difficult, it is always difficult." [01_ukr_a1_ThA, pos. 170, translated]	↔	"No, for me, no, for me, everything is easy to talk about. I don't have such or deep problems." [01_ukr_a1na1_int, pos. 165, translated]

Related to this, frequency analyses of feedback from the participating refugees from Ukraine revealed 9 positive feedback quotes and 18 negative feedback quotes on the example characters in the adapted intervention version. In contrast, the participating refugees from other countries of origin seemed to identify strongly with the example characters in the adapted version of the intervention and did not mention any negative criticism of them, but gave 16 positive feedback quotes. Feedback on the example characters in the nonadapted intervention version was more homogeneous, with both 4 positive and negative feedback quotes given by the participating refugees from Ukraine and 3 positive and 1 negative feedback quotes given by the participating refugees from other countries of origin.

Identification With the Country of Origin

Other factors related to the flight background that influenced the perceived appropriateness of the intervention versions were associated with the identification with the respective country of origin versus the identification with Germany. This goes back to the fact of coming from non-German countries of origin, which was linked to revealed cultural or linguistic differences.

> *Because I'm Muslim and I don't drink alcohol.* [09_oth_na3a4_int, pos. 57, translated]
>
> *Nope, so the psychologists work according to models and views and worldviews that don't match mine.* [09_oth_na3a4_int, pos. 243, translated]

Also, a better linguistic understanding (native language, linguistic facilitation such as subtitles or audios, etc) and the consideration of different reading comprehension levels seemed to increase the perceived cultural appropriateness: "And not every refugee also can read. But having access to audible information can be helpful." [08_oth_a3(na2)_int, pos. 137]. As such, the German language was perceived as stressful by one participant: "If I speak German, I'm tense. Yes, because it has to do with work, with problems, with documents and so on." [01_ukr_a1_ThA, pos. 359, translated]. Whereas the participants from other countries of origin named both cultural and linguistic differences and identification issues, the Ukrainian participants solely named linguistic differences.

Health care

Besides the refugee identity, the factor "health care" was identified to be linked to the perceived cultural appropriateness of the digital sleep intervention. Differences in the needs and

XSL•FO
RenderX

the use of mental healthcare became evident, revealing the importance of familiarity on the one hand and professionalism on the other hand.

> *I would opt for a European doctor.* [01_ukr_a1na1_int, pos. 237, translated]
>
> *That [nonadapted version] was professional I would say.* [01_ukr_a1na1_int, pos. 6, translated]
>
> *If you have problems, you'd rather go to mother, yes, [...] something like relatives, so your nation.* [01_ukr_a1_ThA, pos. 491, translated]

Although psychotherapy was partly described as "trendy" [02_ukr_na2_int, pos. 108, translated], stigmatization of mental health care became apparent: "Where I come from, it's very taboo and really inaccessible." [10_oth_a1_(ThA)int, pos. 129, translated]; this was also on a subconscious level (eg, observation from participant 11 [oth] when filling out the sociodemographic questionnaire: "Psychotherapy? Ah no no no!" [translated]).

Discussion

Principal Findings

Overview

This study for the first time compared the cultural appropriateness perceived by refugees from Ukraine and other countries of origin regarding a digital intervention, specifically the original (developed for Germans) and the culturally adapted (adapted for refugees coming from African and Middle Eastern countries) versions of a digital sleep intervention. Overall, the participating refugees showed a positive attitude toward the intervention, yet the perceived cultural appropriateness regarding the 2 versions seemed to differ between the participants from Ukraine versus those from other countries of origin. The identification with the intervention content regarding the refugee context as well as the context of their countries of origin seemed to play a crucial role (refer to Figure 2). Identification with the refugee context refers to the extent to which participants see their own flight as a part of themselves and could relate to these issues in the intervention. Identification with the country of origin refers to the extent to which the participants identify with the behaviors, values, and norms of their country of origin and could relate to the behaviors, values, and norms illustrated in the intervention. Thereby, identity conflicts among the Ukrainian participants regarding the refugee context were identified as factors reducing the perceived cultural appropriateness of both intervention versions.

Figure 2. The role of identification in the cultural appropriateness of the digital intervention Sleep-e tested in the qualitative study in Germany, as perceived by the participating refugees from Ukraine and other countries of origin. a: adapted version of the intervention; pos: position in the transcript; ThA: think-aloud; ukr: Ukraine.

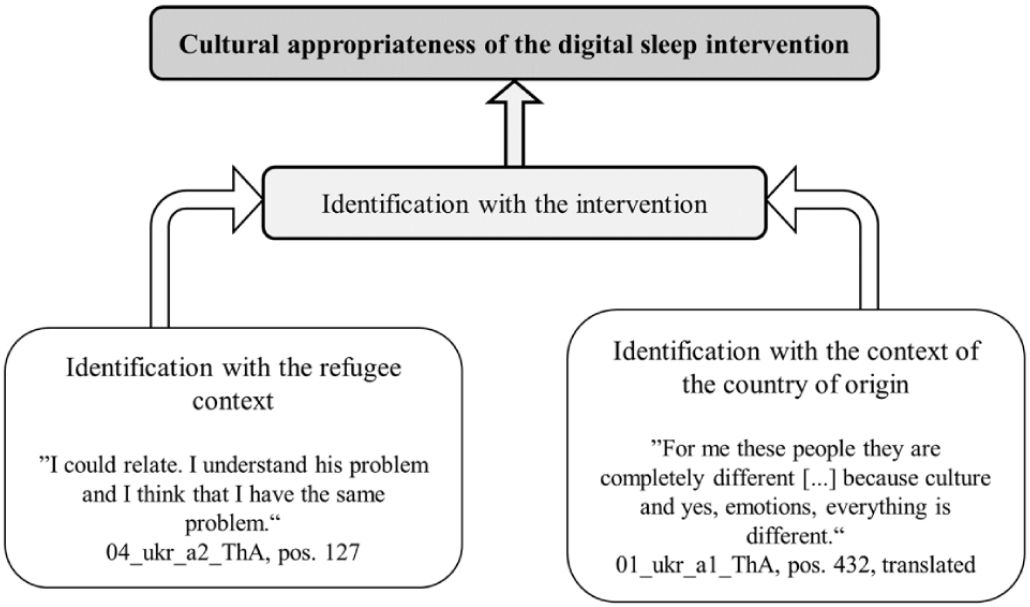

Differences in the Perceived Cultural Appropriateness of the Digital Sleep Intervention

Both versions of the digital sleep intervention received more positive than negative feedback, and feedback on less appropriate aspects mainly concerned design and language aspects (eg, including too much text and hard-to-understand language), which is rather general feedback on digital interventions than a refugee-specific point of view [42-44]. Regarding the participating refugees from countries other than Ukraine, the findings mainly corresponded to previous findings: whereas a lack of identification with the nonadapted version was shown [24], the adapted version seemed to be perceived as appropriate in terms of its content [22], and no negative feedback on the intervention content was given. Regarding the participating refugees from Ukraine, negative feedback on the intervention content was mainly given concerning the adapted version of the intervention and comprised a perceived overfocus

XSL•FO
RenderX

on the topic of flight, in particular regarding the example characters. Concerning the nonadapted version, less negative feedback indicated a perceived inappropriateness of the illustrated life situations and problems of the example characters, which related in particular to a high workload. Thus, both versions of the intervention appeared to be only partially appropriate for the participants from Ukraine.

Identification Influences the Perceived Cultural Appropriateness of the Digital Sleep Intervention

The lower perceived cultural appropriateness among the Ukrainian participants regarding the adapted version of the intervention was suggested to be linked to their lower identification with the intervention: Differences between the participants from Ukraine versus those from other countries of origin became apparent with regard to the identification with the refugee context and aspects related to the identification with the own country of origin. This resulted in differences in the identification with, for example, the example characters in the interventions. The revealed differences correspond to differences between refugees from Ukraine and refugees from other countries of origin suggested in the literature: Although Ukrainian refugees and refugees from other countries of origin share the experience of having had to flee, they differ in various other aspects, such as the welcome culture in Germany as well as sociodemographics, which was also revealed in this study [27,45].

Along with this, higher privileges of Ukrainian refugees and differences in refugee experience or arrival in Germany were discussed among the participants. Not only are there differences between refugees from Ukraine and from other countries, but there are also major differences in the experiences of war and flight among refugees from Ukraine, particularly at the time the study was conducted, which was only one year after the beginning of the war [27]: Some of the participating refugees from Ukraine had left the country right at the beginning of the war without having experienced it themselves, and also without having experienced the loss of family members or friends. Furthermore, the Ukrainian participants might still have hoped to return to their home country soon [27]. Different durations of their stay in Germany therefore seemed to be linked to different flight experiences and, herewith, with a different identification with the refugee context and associated problems described in the adapted intervention version. Depending on this, the adapted version of the intervention could therefore show a certain degree of perceived appropriateness for some of the participating refugees from Ukraine. A dilemma became apparent: Whereas the Ukrainian participants seemed to identify with the European culture of the example characters in the nonadapted version, they did not identify with their concerns, especially regarding a high workload; with regard to the example characters in the adapted version, they could possibly identify in parts with their refugee background, but not with their non-European cultural background. This was accompanied by identity conflicts regarding the refugee context in the statements of the Ukrainian participants, which were not found among the participants from other countries of origin. For example, one Ukrainian participant described the adapted version of the intervention as appropriate for refugees, but not for herself. The importance of identity issues and conflicts in the refugee context has been highlighted previously [46,47].

The found differences in the identity conflicts and the identification with the versions of the intervention emphasize the relevance of specifying a target group and their specific needs in the context of cultural adaptations [48,49]. The lack of including the living situation and the needs of Ukrainian refugees during the cultural adaptation of the present intervention [24] may have led to the identification difficulties and, herewith, the reduced perceived cultural appropriateness. In order to enhance the appropriateness of the intervention for refugees from Ukraine, the intervention could be further adapted, for example by providing the possibility for personalization of the content or of the example characters in order to improve the identification. This has already been applied, for example, as part of the Step-by-Step program in Lebanon [50,51].

Limitations

Several study limitations are to be named. First, the generalizability and robustness of the results are limited due to the qualitative and explorative character of the study. Furthermore, we could only include study participants with sufficient knowledge of German or English. Due to the open recruitment, this resulted in a rather educated and young sample, which possibly influenced the perceived cultural appropriateness of interventions [52]. Second, the majority of participants were recruited via private contacts. This may have biased the results in terms of social desirability, despite enquiring about perceived negative aspects of the interventions. Third, the duration and mode of conduct (video call vs in person) of the interviews with the study participants were not parallelized for the two subgroups. This might have led to an overrepresentation of the statements made by the Ukrainians, combined with the fact that the category system was initially developed based on the transcripts of Ukrainian participants. Fourth, the frequency analyses are not directly interpretable, as they were not related to the frequency of implementation of the respective intervention.

Conclusions

With the rising number of refugees worldwide, the need for low-threshold and culturally appropriate mental health services is increasing. In this qualitative study, the relevance of the identification of the participating refugees with the content of the digital intervention Sleep-e for their perceived cultural appropriateness was emphasized. There were substantial differences between the participating refugees from Ukraine and those from other countries of origin: The participants from Ukraine appeared to be in an identity conflict regarding their identification with the refugee context, which seemed to reduce the perceived appropriateness both of the version of the intervention adapted for refugees from African and Middle East countries and of the nonadapted version for Germans. Refugees from other countries of origin, on the other hand, could identify well with the adapted version of the intervention and perceived it as appropriate. The perceived cultural appropriateness for refugees from Ukraine should be improved by diversifying the problems addressed to consider current developments related to the war in Ukraine and other changes in world affairs. As

XSL•FO
RenderX

such, the cultural sensitivity of treatments should be considered as a dynamic condition related to time and context [53].

Acknowledgments

We would like to thank all study participants for their time and willingness to contribute. We further acknowledge support by the Open Access Publication Fund of the University of Freiburg.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

MB and KS initiated this study. MB, AK, LBS, and KS contributed to the design of this study. Data acquisition was facilitated by HB providing the platform for intervention access, and DL providing the original intervention GET.ON Recovery, and JU provide access to all further materials and facilities. MB largely contributed to the study preparation, recruitment, data collection, analyses, and interpretation. HB and DL provided expertise in digital interventions, and AK, LBS, JU, and KS supervised the whole project. MB and KS wrote the draft of the manuscript. All authors contributed to the further writing of the manuscript and approved the final version of it.

Conflicts of Interest

DL, LBS, and KS were involved in the development of the Sleep-e intervention or its predecessor versions.

Multimedia Appendix 1
Interview guideline.
[PDF File (Adobe PDF File), 90 KB - formative_v9i1e65412_app1.pdf]

Multimedia Appendix 2
COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist.
[PDF File (Adobe PDF File), 156 KB - formative_v9i1e65412_app2.pdf]

Multimedia Appendix 3
Category system.
[PDF File (Adobe PDF File), 127 KB - formative_v9i1e65412_app3.pdf]

References

1. Global trends: forced displacement in 2023. UNHCR. 2024. URL: https://www.unhcr.org/global-trends-report-2023 [accessed 2024-08-05]
2. Blackmore R, Boyle JA, Fazel M, Ranasinha S, Gray KM, Fitzgerald G, et al. The prevalence of mental illness in refugees and asylum seekers: a systematic review and meta-analysis. PLoS Med 2020;17(9):e1003337 [FREE Full text] [doi: 10.1371/journal.pmed.1003337] [Medline: 32956381]
3. Kien C, Sommer I, Faustmann A, Gibson L, Schneider M, Krczal E, et al. Prevalence of mental disorders in young refugees and asylum seekers in European countries: a systematic review. Eur Child Adolesc Psychiatry 2019;28(10):1295-1310 [FREE Full text] [doi: 10.1007/s00787-018-1215-z] [Medline: 30151800]
4. Richter K, Baumgärtner L, Niklewski G, Peter L, Köck M, Kellner S, et al. Sleep disorders in migrants and refugees: a systematic review with implications for personalized medical approach. EPMA J 2020;11(2):251-260 [FREE Full text] [doi: 10.1007/s13167-020-00205-2] [Medline: 32549917]
5. Uphoff E, Robertson L, Cabieses B, Villalón FJ, Purgato M, Churchill R, et al. An overview of systematic reviews on mental health promotion, prevention, and treatment of common mental disorders for refugees, asylum seekers, and internally displaced persons. Cochrane Database Syst Rev 2020;9(9):CD013458 [FREE Full text] [doi: 10.1002/14651858.CD013458.pub2] [Medline: 32885850]
6. Baggio S, Gonçalves L, Heeren A, Heller P, Gétaz L, Graf M, et al. The mental health burden of immigration detention: an updated systematic review and meta-analysis. Criminology - The Online Journal 2020;2(2):219-233. [doi: 10.18716/ojs/krimoj/2020.2.7]
7. Jannesari S, Lotito C, Turrini G, Oram S, Barbui C. How does context influence the delivery of mental health interventions for asylum seekers and refugees in low- and middle-income countries? a qualitative systematic review. Int J Ment Health Syst 2021;15(1):80 [FREE Full text] [doi: 10.1186/s13033-021-00501-y] [Medline: 34702334]

XSL•FO
RenderX

8. Satinsky E, Fuhr DC, Woodward A, Sondorp E, Roberts B. Mental health care utilisation and access among refugees and asylum seekers in Europe: a systematic review. Health Policy 2019;123(9):851-863 [FREE Full text] [doi: 10.1016/j.healthpol.2019.02.007] [Medline: 30850148]
9. Kaufman KR, Bhui K, Katona C. Mental health responses in countries hosting refugees from Ukraine. BJPsych Open 2022;8(3):e87 [FREE Full text] [doi: 10.1192/bjo.2022.55] [Medline: 35361299]
10. Liem A, Natari RB, Jimmy, Hall BJ. Digital health applications in mental health care for immigrants and refugees: a rapid review. Telemed J E Health 2021;27(1):3-16. [doi: 10.1089/tmj.2020.0012] [Medline: 32498658]
11. El-Haj-Mohamad R, Nohr L, Niemeyer H, Böttche M, Knaevelsrud C. Smartphone-delivered mental health care interventions for refugees: a systematic review of the literature. Glob Ment Health (Camb) 2023;10:e6 [FREE Full text] [doi: 10.1017/gmh.2022.61] [Medline: 36843879]
12. Moshe I, Terhorst Y, Philippi P, Domhardt M, Cuijpers P, Cristea I, et al. Digital interventions for the treatment of depression: a meta-analytic review. Psychol Bull 2021;147(8):749-786. [doi: 10.1037/bul0000334] [Medline: 34898233]
13. Pauley D, Cuijpers P, Papola D, Miguel C, Karyotaki E. Two decades of digital interventions for anxiety disorders: a systematic review and meta-analysis of treatment effectiveness. Psychol Med 2023;53(2):567-579 [FREE Full text] [doi: 10.1017/S0033291721001999] [Medline: 34047264]
14. Rismawan W, Marchira CR, Rahmat I. Usability, acceptability, and adherence rates of mobile application interventions for prevention or treatment of depression: a systematic review. J Psychosoc Nurs Ment Health Serv 2021;59(2):41-47. [doi: 10.3928/02793695-20201104-02] [Medline: 33180949]
15. Or CK, Chan AHS. Inspection methods for usability evaluation. In: User Experience Methods and Tools in Human-Computer Interaction. Boca Raton, FL: CRC Press; 2024:170-192.
16. Hall GCN, Ibaraki AY, Huang ER, Marti CN, Stice E. A meta-analysis of cultural adaptations of psychological interventions. Behav Ther 2016;47(6):993-1014. [doi: 10.1016/j.beth.2016.09.005] [Medline: 27993346]
17. Harper Shehadeh M, Heim E, Chowdhary N, Maercker A, Albanese E. Cultural adaptation of minimally guided interventions for common mental disorders: a systematic review and meta-analysis. JMIR Ment Health 2016;3(3):e44 [FREE Full text] [doi: 10.2196/mental.5776] [Medline: 27670598]
18. Soto A, Smith TB, Griner D, Domenech Rodríguez M, Bernal G. Cultural adaptations and therapist multicultural competence: two meta-analytic reviews. J Clin Psychol 2018;74(11):1907-1923. [doi: 10.1002/jclp.22679] [Medline: 30091201]
19. Heim E, Ramia JA, Hana RA, Burchert S, Carswell K, Cornelisz I, et al. Step-by-step: feasibility randomised controlled trial of a mobile-based intervention for depression among populations affected by adversity in Lebanon. Internet Interv 2021;24:100380 [FREE Full text] [doi: 10.1016/j.invent.2021.100380] [Medline: 33747798]
20. Migration and cultural diversity: cultural competence and cultural sensitivity. Finish Institute for Health and Welfare. 2023. URL: https://thl.fi/en/topics/migration-and-cultural-diversity/support-material/good-practices/cultural-competence-and-cultural-sensitivity [accessed 2024-08-08]
21. Böge K, Karnouk C, Hoell A, Tschorn M, Kamp-Becker I, Padberg F, et al. Effectiveness and cost-effectiveness for the treatment of depressive symptoms in refugees and asylum seekers: a multi-centred randomized controlled trial. Lancet Reg Health Eur 2022;19:100413 [FREE Full text] [doi: 10.1016/j.lanepe.2022.100413] [Medline: 35694653]
22. Spanhel K, Hovestadt E, Lehr D, Spiegelhalder K, Baumeister H, Bengel J, et al. Engaging refugees with a culturally adapted digital intervention to improve sleep: a randomized controlled pilot trial. Front Psychiatry 2022;13:832196 [FREE Full text] [doi: 10.3389/fpsyt.2022.832196] [Medline: 35280163]
23. El-Haj-Mohamad R, Böttche M, Vöhringer M, Specht F, Stammel N, Nesterko Y, et al. An internet-based cognitive behavioural intervention for adults with depression in Arabic-speaking countries: a randomized controlled trial. Stress Health 2024;40(5):e3432. [doi: 10.1002/smi.3432] [Medline: 38806256]
24. Spanhel K, Schweizer JS, Wirsching D, Lehr D, Baumeister H, Bengel J, et al. Cultural adaptation of internet interventions for refugees: results from a user experience study in Germany. Internet Interv 2019;18:100252 [FREE Full text] [doi: 10.1016/j.invent.2019.100252] [Medline: 31890608]
25. Ukraine situation flash update #71. UNHCR. 2024. URL: https://data.unhcr.org/en/documents/details/110114 [accessed 2024-08-05]
26. Weissbecker I, Khan O O, Kondakova N. N, Poole L L, Cohen JT. Mental health in transition: assessment and guidance for strengthening integration of mental health into primary health care and community-based service platforms in Ukraine. World Bank Group. 2017. URL: http://documents.worldbank.org/curated/en/310711509516280173 [accessed 2024-08-08]
27. Brücker H, Ette A, Grabka M, Kosyakova Y, Niehues W, Rother N, et al. Geflüchtete aus der Ukraine in Deutschland (Forschungsbericht 41). [Refugees from Ukraine in Germany (research report 41)] 2022 [FREE Full text] [doi: 10.12765/bro-2022-04]
28. Ukraine situation: regional refugee response plan. UNHCR. Geneva, Switzerland; 2023. URL: https://reporting.unhcr.org/ukraine-situation-regional-refugee-response-plan [accessed 2024-08-09]
29. Buchcik J, Kovach V, Adedeji A. Mental health outcomes and quality of life of Ukrainian refugees in Germany. Health Qual Life Outcomes 2023;21(1):23 [FREE Full text] [doi: 10.1186/s12955-023-02101-5] [Medline: 36894946]
30. Bastien CH, Vallières A, Morin CM. Validation of the Insomnia Severity Index as an outcome measure for insomnia research. Sleep Med 2001;2(4):297-307. [doi: 10.1016/s1389-9457(00)00065-4] [Medline: 11438246]

XSL•FO
RenderX

31. Boß L, Lehr D, Reis D, Vis C, Riper H, Berking M, et al. Reliability and validity of assessing user satisfaction with web-based health interventions. J Med Internet Res 2016;18(8):e234 [FREE Full text] [doi: 10.2196/jmir.5952] [Medline: 27582341]
32. Konrad K. Lautes Denken [Think Aloud]. In: Handbuch Qualitative Forschung in der Psychologie [Handbook Qualitative Research in Psychology]. VS Verlag für Sozialwissenschaften. Cham, Switzerland: Springer; 2010:476-490.
33. Thiart H, Lehr D, Ebert DD, Berking M, Riper H. Log in and breathe out: internet-based recovery training for sleepless employees with work-related strain - results of a randomized controlled trial. Scand J Work Environ Health 2015;41(2):164-174 [FREE Full text] [doi: 10.5271/sjweh.3478] [Medline: 25590336]
34. Ebert DD, Berking M, Thiart H, Riper H, Laferton JAC, Cuijpers P, et al. Restoring depleted resources: Efficacy and mechanisms of change of an internet-based unguided recovery training for better sleep and psychological detachment from work. Health Psychol 2015;34S:1240-1251. [doi: 10.1037/hea0000277] [Medline: 26651465]
35. Spanhel K, Balci S, Feldhahn F, Bengel J, Baumeister H, Sander LB. Cultural adaptation of internet- and mobile-based interventions for mental disorders: a systematic review. NPJ Digit Med 2021;4(1):128 [FREE Full text] [doi: 10.1038/s41746-021-00498-1] [Medline: 34433875]
36. Kuckartz U, Rädiker S. Qualitative Inhaltsanalyse [Qualitative content analysis]. In: Methoden, Praxis, Computerunterstützung [Methods, Practice, Computer Support]. Weinheim, Germany: Beltz Juventa; 2022.
37. Mayring P. Qualitative Inhaltsanalyse [Qualitative content analysis]. In: Grundlagen und Techniken [Basics and techniques]. Weinheim, Germany: Beltz Verlag; 2022.
38. Stamann C, Janssen M, Schreier M. Searching for the core: defining qualitative content analysis. Forum Qual Soc Res 2016;17(3):16. [doi: 10.17169/FQS-17.3.2581]
39. VERBI Software. MAXQDA. 2022. URL: https://www.maxqda.com/de/ [accessed 2024-08-08]
40. Guidance note - research on refugees, asylum seekers and migrants. European Commission. 2016. URL: https://ec.europa.eu/research/participants/data/ref/h2020/other/hi/guide_research-refugees-migrants_en.pdf [accessed 2025-02-24]
41. Tong A, Sainsbury P, Craig J. Consolidated Criteria for Reporting Qualitative Research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
42. Micklitz HM, Nagel Z, Jahn S, Oertelt-Prigione S, Andersson G, Sander LB. Digital self-help for people experiencing intimate partner violence: a qualitative study on user experiences and needs including people with lived experiences and services providers. BMC Public Health 2023;23(1):1471 [FREE Full text] [doi: 10.1186/s12889-023-16357-5] [Medline: 37533005]
43. Büscher R, Sander LB, Nuding M, Baumeister H, Teismann T. Blending video therapy and digital self-help for individuals with suicidal ideation: intervention design and a qualitative study within the development process. JMIR Form Res 2023;7:e49043 [FREE Full text] [doi: 10.2196/49043] [Medline: 37733414]
44. Seiferth C, Vogel L, Aas B, Brandhorst I, Carlbring P, Conzelmann A, et al. How to e-mental health: a guideline for researchers and practitioners using digital technology in the context of mental health. Nat. Mental Health 2023;1(8):542-554. [doi: 10.1038/s44220-023-00085-1]
45. Prati G. Attitudes toward the response of the EU to the Russian invasion of Ukraine: The role of empathic concern, identification with the EU, and the view that Ukraine is part of the European community. Peace and Conflict: Journal of Peace Psychology 2023;29(3):253-256. [doi: 10.1037/pac0000679]
46. Kizilhan JI. Kultursensitivität in der Psychotherapie [Culture sensitivity in psychotherapy]. In: Verhaltenstherapiemanual - Erwachsene [Behavioral therapy manual - adults]. Berlin/ Heidelberg, Germany: Springer; 2022:43-48.
47. Watzlawik M, Brescó de Luna I. The self in movement: being identified and identifying oneself in the process of migration and asylum seeking. Integr Psychol Behav Sci 2017;51(2):244-260. [doi: 10.1007/s12124-017-9386-6] [Medline: 28299535]
48. Heim E, Kohrt BA. Cultural adaptation of scalable psychological interventions: a new conceptual framework. CPE 2019;1(4):e37679. [doi: 10.32872/cpe.v1i4.37679]
49. Heim E, Burchert S, Wenger A. Mobile Angebote für Geflüchtete: Psychotherapeutische Interventionen im globalen Kontext [Mobile services for refugees: psychotherapeutic interventions in a global context]. PiD - Psychotherapie im Dialog 2018;19(04):61-65. [doi: 10.1055/a-0592-0339]
50. Abi Ramia J, Harper Shehadeh M, Kheir W, Zoghbi E, Watts S, Heim E, et al. Community cognitive interviewing to inform local adaptations of an e-mental health intervention in Lebanon. Glob Ment Health (Camb) 2018;5:e39 [FREE Full text] [doi: 10.1017/gmh.2018.29] [Medline: 30637112]
51. Harper Shehadeh MJ, Abi Ramia J, Cuijpers P, El Chammay R, Heim E, Kheir W, et al. Step-by-step, an E-mental health intervention for depression: a mixed methods pilot study from lebanon. Front Psychiatry 2019;10:986 [FREE Full text] [doi: 10.3389/fpsyt.2019.00986] [Medline: 32116815]
52. Heim E, Ademi B, Dacaj A, Hosny N, Burchert S, Cerga Pashoja A, et al. Why did our trial not work out? A qualitative analysis. Clin Psychol Eur 2024;6(2):e12887. [doi: 10.32872/cpe.12887] [Medline: 39119055]
53. Bernal G, Bonilla J, Bellido C. Ecological validity and cultural sensitivity for outcome research: issues for the cultural adaptation and development of psychosocial treatments with Hispanics. J Abnorm Child Psychol 1995;23(1):67-82. [doi: 10.1007/BF01447045] [Medline: 7759675]

XSL•FO
RenderX

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Research

Edited by A Mavragani; submitted 07.10.24; peer-reviewed by C Or; comments to author 28.01.25; revised version received 20.02.25; accepted 21.02.25; published 03.04.25.

Please cite as:
Blomenkamp M, Kiesel A, Baumeister H, Lehr D, Unterrainer J, Sander LB, Spanhel K
Assessing the Cultural Fit of a Digital Sleep Intervention for Refugees in Germany: Qualitative Study
JMIR Form Res 2025;9:e65412
URL: https://formative.jmir.org/2025/1/e65412
doi:10.2196/65412
PMID:

©Maja Blomenkamp, Andrea Kiesel, Harald Baumeister, Dirk Lehr, Josef Unterrainer, Lasse B Sander, Kerstin Spanhel. Originally published in JMIR Formative Research (https://formative.jmir.org), 03.04.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Digital Health Platform for Maternal Health: Design, Recruitment Strategies, and Lessons Learned From the PowerMom Observational Cohort Study

Toluwalase Ajayi[1,2], MD; Jacqueline Kueper[2], PhD; Lauren Ariniello[2], MBA; Diana Ho[2], MHA; Felipe Delgado[2], BS; Matthew Beal[3], MD; Jill Waalen[3], PhD; Katie Baca Motes[2], MBA; Edward Ramos[2], PhD

[1]Jacobs Center for Health Innovation, Department of Medicine and Pediatrics, University of California, San Diego, La Jolla, CA, United States
[2]Digital Trial Center, Scripps Research Translational Institute, La Jolla, CA, United States
[3]Division of Preventative Medicine, Department of Family Medicine, University of California, San Diego, La Jolla, United States

Corresponding Author:
Toluwalase Ajayi, MD
Jacobs Center for Health Innovation
Department of Medicine and Pediatrics
University of California, San Diego
9300 Campus Point Drive
MC 7196
La Jolla, CA, 92037
United States
Phone: 1 785 218 1643
Email: tajayi@health.ucsd.edu

Abstract

Background: Maternal health research faces challenges in participant recruitment, retention, and data collection, particularly among underrepresented populations. Digital health platforms like PowerMom (Scripps Research) offer scalable solutions, enabling decentralized, real-world data collection. Using innovative recruitment and multimodal techniques, PowerMom engages diverse cohorts to gather longitudinal and episodic data during pregnancy and post partum.

Objective: This study aimed to evaluate the design, implementation, and outcomes of the PowerMom research platform, with a focus on participant recruitment, engagement, and data collection across diverse populations. Secondary objectives included identifying challenges encountered during implementation and deriving lessons to inform future digital maternal health studies.

Methods: Participants were recruited via digital advertisements, pregnancy apps, and the PowerMom Consortium of more than 15 local and national organizations. Data collection included self-reported surveys, wearable devices, and electronic health records. Anomaly detection measures were implemented to address fraudulent enrollment activity. Recruitment trends and descriptive statistics from survey data were analyzed to summarize participant characteristics, assess engagement metrics, and quantify missing data to identify gaps.

Results: Overall, 5617 participants were enrolled from 2021 to 2024, with 69.8% (n=3922) providing demographic data. Of these, 48.5% (2723/5617) were younger than 35 years, 14% (788/5617) identified as Hispanic or Latina, and 13.7% (770/5617) identified as Black or African American. Geographic representation spanned all 50 US states, Puerto Rico, and Guam, with 58.3% (3276/5617) residing in areas with moderate access to maternity care and 16.4% (919/5617) in highly disadvantaged neighborhoods based on the Area Deprivation Index. Enrollment rates increased substantially over the study period, from 55 participants in late 2021 to 3310 in 2024, averaging 99.4 enrollments per week in 2024. Participants completed a total of 17,123 surveys, with 71.8% (4033/5617) completing the Intake Survey and 12.4% (697/5617) completing the Postpartum Survey. Wearable device data were shared by 1168 participants, providing more than 378,000 daily biometric measurements, including activity levels, sleep, and heart rate. Additionally, 96 participants connected their electronic health records, contributing 276 data points such as diagnoses, medications, and laboratory results. Among pregnancy-related characteristics, 28.1% (1578/5617) enrolled during the first trimester, while 15.1% (849/5617) reported information about the completion of their pregnancies during the study period. Among the 913 participants who shared delivery information, 56.1% (n=512) had spontaneous vaginal deliveries and 17.9% (n=163) underwent unplanned cesarean sections.

XSL•FO
RenderX

Conclusions: The PowerMom platform demonstrates the feasibility of using digital tools to recruit and engage diverse populations in maternal health research. Its ability to integrate multimodal data sources showcases its potential to provide comprehensive maternal-fetal health insights. Challenges with data completeness and survey attrition underscore the need for sustained participant engagement strategies. These findings offer valuable lessons for scaling digital health platforms and addressing disparities in maternal health research.

Trial Registration: ClinicalTrials.gov NCT03085875; https://clinicaltrials.gov/study/NCT03085875

(JMIR Form Res 2025;9:e70149) doi:10.2196/70149

KEYWORDS

maternal health research; digital health platforms; pregnancy monitoring; decentralized clinical trials; participant engagement; health disparities

Introduction

In 2020, there were approximately 24 pregnancy-related deaths per 100,000 births in the United States, which is more than double compared with other high-income countries [1]. Alarmingly, these rates have continued to rise, with disparities disproportionately affecting marginalized populations [2]. Non-Hispanic Black or African American pregnant individuals experience maternal mortality rates approximately 2.5 times higher than their White counterparts, with non-Hispanic American Indian or Alaska Native women facing rates three times higher [2,3]. Cardiovascular conditions, perinatal depression, and preexisting medical conditions are among the leading contributors to these disparities [4,5]. However, the root causes extend beyond medical risk factors, with social determinants of health and structural racism playing crucial roles [6,7].

Access to care, socioeconomic status, implicit bias, and environmental factors further exacerbate these inequities. Individuals from underrepresented communities are more likely to face barriers to health care, economic insecurity, and a lack of culturally competent care. These disparities are often intensified by distrust of health care systems, fueled by historical and ongoing discrimination in medical settings [8]. Addressing these disparities requires innovative research approaches that not only engage affected communities but also prioritize trust, accessibility, and cultural competence [9].

Traditional maternal health research has struggled to effectively recruit and engage underrepresented populations. Common recruitment barriers include mistrust of the health care system, concerns about data confidentiality, language differences, and logistical hurdles such as the lack of transportation or time constraints [10,11]. These challenges are compounded by the fact that traditional recruitment methods, such as clinic-based enrollment, often fail to reach the most vulnerable populations, leading to underrepresentation in research that is critical for health equity [12,13]. However, many of these traditional methods do not offer sufficient direct benefits to participants to offset their perceived and real costs of participation, particularly in communities with heightened mistrust [9].

The emergence of digital health platforms offers a potential solution to overcome these long-standing barriers, particularly in maternal health research [14]. Mobile and decentralized platforms provide an opportunity to recruit participants remotely, allowing for broad geographical reach and the ability to engage individuals in real-world settings [15]. Digital tools enable continuous, real-time data collection, which can provide a more holistic view of maternal health than episodic clinic visits. These platforms can also deliver tailored educational content and resources, building trust and empowering participants through transparency and engagement [16].

The PowerMom (Scripps Research) platform was designed to capitalize on these opportunities by implementing a bilingual (English and Spanish) digital research platform that facilitates remote, convenient participation. PowerMom's recruitment strategy was explicitly designed to harness the power of digital platforms to allow seamless, rolling self-enrollment and ensure wide-scale and inclusive participation across diverse communities. It leverages digital approaches such as push notifications and messaging to enhance participant engagement and retention [17-20], allowing for data collection that respects participants' time and cultural needs.

By providing a decentralized clinical trial environment, PowerMom enables participants to engage with the study on their own terms, enhancing adherence and yielding data that reflect a broad, comprehensive, and participant-centered perspective on how structural factors influence maternal health. The platform supports the collection of real-world, continuous data through app-based surveys, electronic health record (EHR) data linkage, and digital health technologies such as wearable sensors, capturing biometric data (eg, heart rate, sleep, and physical activity) that go beyond the traditional, episodic measures collected during prenatal clinic visits [21-26]. PowerMom's flexible architecture supports not only the collection of baseline data but also the integration of multiple substudies, which allows for more targeted research on specific maternal health issues while leveraging the broader cohort's data.

In this paper, we describe the design and methods of the PowerMom research platform, focusing on its recruitment strategies, data collection capabilities, and participant engagement efforts, and discuss lessons learned from its implementation in diverse populations. In addition, we provide a detailed description of the participant demographics recruited through PowerMom and examine their engagement with the various data collection opportunities throughout the pregnancy and postpartum stages.

XSL•FO
RenderX

Methods

Platform Design

The PowerMom platform, launched in March 2021, is built on CareEvolution's MyDataHelps secure digital infrastructure, which provides a flexible and scalable foundation for collecting and integrating a range of participant data, including biometric, survey, and EHR data. This cloud-based, Health Insurance Portability and Accountability Act (HIPAA)–compliant infrastructure provides secure data storage and management, leveraging CareEvolution's backend architecture to maintain participant confidentiality and data integrity. In addition, the platform's technical infrastructure adheres to industry-leading security protocols for data transfer, ensuring that biometric data from wearables, EHR data, and survey responses are securely encrypted and processed.

The PowerMom platform integrates data from various sources in real time, allowing for continuous monitoring of participants' health. This includes heart rate, activity levels, and sleep data collected from wearable devices such as Fitbit, Apple Watch, and other HealthKit or Google Fit–compatible technologies. By allowing participants to connect these devices, the platform facilitates a richer, more granular view of maternal health, capturing data beyond the traditional measures obtained during prenatal clinic visits.

To ensure that the platform reflects the needs of its target population, PowerMom's development involved collaboration with a participant advisory board composed of pregnant and postpartum individuals from underrepresented communities. The advisory board provided critical feedback on the platform's design, user interface, and engagement strategies, ensuring that the platform remains culturally relevant and user-friendly for diverse participants. This adheres to the CONSORT-EHEALTH (v1.6.1) guidelines for digital health intervention research to ensure comprehensive and standardized reporting. The checklist was used to structure study reporting, covering recruitment, platform usability, data integration, and participant engagement.

Participant Recruitment

Eligibility Criteria

Participants were eligible for PowerMom if they were pregnant or post partum (up to 8 weeks after delivery), aged 16 years and older, residing in the United States or its territories, and capable of providing consent in either English or Spanish.

eConsent

All eligible PowerMom participants were presented with an innovative electronic consent (eConsent) process, enabling them to remotely join the study. This eConsent process significantly broadened the potential participant pool by allowing individuals to complete consent forms at their convenience, reducing the need for in-person visits or transfer of paper forms. The eConsent interface was designed to be user-friendly and accessible across different devices, ensuring that participants could easily navigate the process (refer to Multimedia Appendix 1). This digital consent process not only facilitated seamless participation but also ensured compliance with ethical standards for informed consent in clinical research.

Study Setting

The study was conducted entirely remotely, using the PowerMom platform to engage participants across the United States, Puerto Rico, and Guam. The decentralized nature of the study eliminated geographic barriers, enabling recruitment from urban, suburban, and rural areas. Participants could enroll and participate using mobile devices or computers, providing flexibility for those in resource-limited settings.

Recruitment Strategy

PowerMom implemented a rolling, digital self-enrollment process to maximize reach and engagement. Recruitment took place from March 22, 2021, to July 30, 2024, and relied on a variety of digital and in-person outreach strategies. Participants in PowerMom are recruited into a baseline cohort but may also be invited to join additional substudies based on specific eligibility criteria. These substudies focus on more targeted research questions, allowing researchers to augment baseline data with additional detailed outcomes.

Digital Outreach

Recruitment relied heavily on digital advertising campaigns on social media platforms such as Facebook, Instagram, and Google Ads. In addition, partnerships with pregnancy-related apps such as Philips Pregnancy+ (Philips Digital) enabled in-app advertisements that reached individuals in their natural digital environments. These ads provided information about the study, emphasizing its accessibility and participant-centered design. Digital enrollment campaigns were designed to appear within the app, enabling participants to enroll in PowerMom seamlessly from their daily digital environment.

PowerMom Consortium

The PowerMom Consortium, established in December 2021, brought together more than 15 local and national organizations focused on improving maternal health. This consortium amplified recruitment efforts through joint campaigns, newsletters, and educational outreach, significantly increasing visibility among target demographics [27].

Community Engagement

One of the platform's key partners, Mae Health, is a culturally competent, digital health platform that connects Black expectant mothers with vital resources to improve pregnancy outcomes. The collaboration with Mae Health included educational campaigns and targeted outreach to Black and African American communities.

PowerMom also partnered with advocacy organizations such as the African American Wellness Center for Children and Families. These organizations provided on-the-ground recruitment efforts, ensuring that PowerMom reached communities that might otherwise be underrepresented in digital health research.

XSL•FO
RenderX

Data Collection

Self-Reported Outcomes

Participants were asked to complete several surveys throughout the study. An initial Intake Survey gathered demographic data, including race, ethnicity, education, socioeconomic status, and health care access. A Health History Survey collected information on pregnancy history, preexisting health conditions, and access to care.

Following the baseline surveys, participants received biweekly surveys to assess changes in their health during pregnancy. These surveys collected data on vital signs (eg, weight, blood pressure, and heart rate); pregnancy symptoms; and health behaviors, including tobacco, alcohol, and drug use. Biweekly surveys were released based on the completion of the Intake and Health History Surveys, and participants were notified about each new biweekly survey release via push notifications through the PowerMom app.

Participants also completed a Delivery Survey upon giving birth, capturing details about the delivery, birth outcomes, and complications. A Postpartum Survey, administered at 6-8 weeks post partum, assessed maternal mental health using the Edinburgh Postnatal Depression Scale and gathered information on postpartum complications such as gestational diabetes, pregnancy-induced hypertension, preeclampsia, and preterm birth (refer to Multimedia Appendix 1).

Wearable Devices

Participants were prompted to connect their wearable devices, including Fitbit, Apple Watch, and other compatible devices, for real-time monitoring of heart rate, physical activity, and sleep duration data. Participants had the option to upload up to 3 months of retrospective data (eg, prepregnancy data) and continue data collection prospectively throughout their pregnancy and the postpartum period.

EHR Integration

After completing the initial baseline surveys, participants were prompted to connect their pregnancy-related EHR data through the PowerMom app. The EHR integration can reach more than 85% of all US health care providers, ensuring comprehensive coverage of both historical and prospective health data. Participants could connect their EHR at any point during the study, providing a continuous flow of clinical information that complemented the self-reported and biometric data.

Study Engagement and Retention Strategies

The PowerMom platform was designed with a strong focus on participant engagement and retention. To keep participants engaged, the platform provided regular push notifications, educational content, and personalized data visualizations. For example, participants could track their own metrics—such as daily activity or sleep patterns—through visual feedback in the app, which helped build trust and maintain engagement.

In addition to digital strategies, participants were engaged through blog posts that provided some initial return of results based on aggregated data collected. The PowerMom advisory board also contributed to refining engagement strategies by providing input on how to make the platform more user-friendly and responsive to the needs of underrepresented populations. Engagement metrics were closely monitored, and iterative improvements were made to the platform based on participant feedback.

Data Integration

All data types (eg, survey data, wearable device data, and EHR data) were collected and integrated through the MyDataHelps PowerMom platform app. When discrepancies arose between data sources, such as activity metrics reported in surveys versus wearables, priority was given to wearable data due to their objective and continuous measurement capabilities. Survey data were used to provide context, particularly when participants reported conditions or limitations that could explain deviations. This approach ensured the accuracy and reliability of the data.

These data were securely transferred to the Scripps Research Institute for storage and analysis. The data management system leverages cloud-based infrastructure that is HIPAA compliant, ensuring the highest level of data integrity and security. The data were regularly monitored by the research team and cleaned to address any issues with synchronization between devices, missing data, or outliers, ensuring high-quality data for analysis.

Statistical Analysis

We evaluated the PowerMom platform using operational outcomes and participant characteristics. The analytic sample included all participants who met the baseline PowerMom eligibility criteria, assessed through an initial Screener Survey, and who provided eConsent to participation.

Participant Characteristics

Baseline sociodemographic and pregnancy-related characteristics of all enrolled participants were summarized using self-report data from the Screener, Intake, and Health History Surveys. Delivery outcomes were also described for participants who filled out the Delivery Survey. All eligible participants are included in table summaries of these characteristics, with missing data represented as “Unknown” responses for both single-question and survey-level nonresponse. Survey-level completion patterns, reported in the engagement metrics section, capture the number of eligible participants who completed at least one question on each survey.

Enrollment Trends

We analyzed enrollment trends beyond the three-year study recruitment period. Weekly and cumulative enrollment rates were calculated to demonstrate the impact of these strategies, culminating in the final total number of participants.

Summary statistics were calculated using frequencies and percentages for categorical variables and mean (SD) for continuous variables. All analyses were conducted using MyDataHelps Export Explorer (SQL) and Python (v3.8; Guido van Rossum).

Ethical Considerations

The study protocol for PowerMom was reviewed and approved by the Scripps Health Institutional Review Board (IRB 21-7738). The research adhered to ethical guidelines for human subject

XSL•FO
RenderX

research, ensuring compliance with all applicable regulations. All participants provided eConsent before enrollment. The eConsent process was designed to be accessible and user-friendly, enabling participants to complete the consent remotely via digital platforms. The consent process highlighted participants' rights, data confidentiality measures, and the voluntary nature of their participation. Participant data were anonymized or deidentified before analysis to ensure confidentiality. The PowerMom platform operates on a HIPAA-compliant, secure infrastructure, with additional protections provided by a Certificate of Confidentiality from the National Institutes of Health. This legal protection allows the platform to refuse the disclosure of participant data in civil, criminal, administrative, or legislative proceedings, adding an extra robust layer of privacy beyond standard HIPAA protections. Participants were not compensated for being part of the PowerMom study. However, participants recruited from the PowerMom study for one substudy were provided with monetary incentives. Participants in this substudy were eligible to receive up to US $175 in Amazon gift cards as compensation for completing specific milestones within the substudy. These incentives were structured to encourage engagement without coercion, according to institutional review board guidelines, and the details were explicitly communicated during the consent process.

Results

Demographic Data

Participants were drawn from diverse demographic backgrounds, with representation across different age groups, racial and ethnic categories, and geographic locations. Among the 5617 participants enrolled in the study, 3922 (69.8%) opted to share demographic information. Participants younger than 35 years comprised 48.5% (2723/5617) of the cohort, while 19.3% (1083/5617) were 35 years or older. Age data were missing for 32.2% (1811/5617) of participants.

Regarding ethnicity, 14% (788/5617) of participants reported identifying as Hispanic or Latina, while 55.8% (3134/5617) did not. Ethnicity data were missing for 30.2% (1695/5617) of participants. Racial backgrounds included 0.9% (52/5617) identifying as American Indian or Alaska Native, 1.9% (105/5617) as Asian, 13.7% (770/5617) as Black or African American or African, and 38.7% (2175/5617) as White. In addition, 8.5% (478/5617) identified as Hispanic, Latina, or Spanish without specifying another race, 0.2% (9/5617) as Native Hawaiian or Other Pacific Islander, and 5.9% (333/5617) as multiple races. Racial data were missing for 30.2% (1695/5617) of participants.

Geographic distribution based on zip code at the time of study enrollment showed that 11.3% (635/5617) of participants residing in the Northeast, 29% (1630/5617) in the South, 15.7% (880/5617) in the Midwest, and 15.2% (853/5617) in the West. Only 0.4% (22/5617) resided in the Pacific region, and less than 0.1% (6/5617) were from Puerto Rico. Geographic data were missing for 28.4% (1594/5617) of participants.

Education levels varied, with 12.5% (703/5617) of participants reporting a graduate or postgraduate degree, 18.9% (1063/5617) holding a college degree, and 6.1% (342/5617) reporting some college or technical school education. High-school graduates or those with General Educational Development attainment accounted for 27.5% (1545/5617), while 6% (337/5617) reported not completed high school. Education data were missing for 29% (1627/5617) of participants.

Access to maternity care was categorized based on geographic location using the March of Dimes Maternity Care Deserts Dashboard [28]. These data were reported by 62.6% (3520/5617) of participants, with 58.3% (3276/5617) residing in areas with moderate access and 4.3% (244/5617) in maternity care deserts or areas with low access. Access data were missing for 37.3% (2097/5617) of participants.

Neighborhood disadvantage was assessed using the Neighborhood Atlas' Area Deprivation Index (ADI). The higher the ADI value, the less resources the neighborhood has. Among all participants, 16.4% (919/5617) reported residence in neighborhoods with an ADI national rank of greater than or equal to 75, indicating higher levels of deprivation, while 24.6% (1381/5617), 19.7% (1108/5617), and 10.2% (574/5617) lived in areas with ADI national ranks between 50-75, 25-50, and less than 25, respectively. ADI data were missing for 29.1% (1635/5617) of participants.

Pregnancy and Delivery-Related Characteristics

Among the 5617 participants enrolled in the study, the pregnancy stage at enrollment varied. Approximately 28.1% (1578/5617) of participants enrolled during the first trimester, 26.6% (1493/5617) in the second trimester, and 16.2% (909/5617) in the third trimester. Postdelivery enrollment accounted for 13.1% (736/5617), while the pregnancy stage at enrollment was unknown for 16% (901/5617) of participants.

Most participants (3394/5617, 60.4%) reported being pregnant with one fetus, while 2.2% (121/5617) reported being pregnant with twins, 0.3% (16/5617) with triplets, and 0.3% (15/5617) with more than 3 fetuses. According to available data [29], around 2% to 3% of pregnancies in the United States result in the birth of triplets or higher-order multiples. The number of fetuses was not reported for 36.9% (2071/5617) of participants.

Regarding pregnancy history, 6.6% (371/5617) of participants reported no previous pregnancies, while 13.5% (759/5617), 12.2% (688/5617), 8% (447/5617), 5.8% (328/5617), and 3.2% (180/5617) reported 1, 2, 3, 4, and 5 previous pregnancies, respectively. In addition, 4.7% (266/5617) of participants reported more than 5 previous pregnancies, and 45.9% (2578/5617) did not provide information on past pregnancies.

When asked about the number of past births, 19.9% (1118/5617) reported having none, while 16.4% (923/5617), 10.2% (571/5617), 4.6% (260/5617), 2% (111/5617, 1% (58/5617), and 0.9% (53/5617) reported 1, 2, 3, 4, 5, and more than 5 births, respectively. Information on past births was missing for 44.9% (2523/5617) of participants. Of the 1106 (21.1%) participants who reported at least one previous pregnancy and a complication, the most commonly reported pregnancy

XSL•FO
RenderX

complications were miscarriage (n=637), preterm birth (n=250), and high blood pressure (n=234).

Participants were seeing a variety of prenatal care providers at baseline, with 37.5% (2104/5617) reported seeing a physician; 9.9% (555/5617) seeing a midwife; 2.3% (129/5617) seeing a doula; and smaller proportions seeing other providers such as nurse midwives (56/5617, 1%) or nurse practitioners (36/5617, 0.6%) or planning home births (9/5617, 0.2%). However, 8% (447/5617) reported not seeing any prenatal care provider, and 44.7% (2511/5617) did not report their provider.

Mental health treatment at baseline revealed that 3.4% (189/5617) of participants were being treated for anxiety, 1.8% (102/5617) for depression, and 9.8% (550/5617) for both anxiety and depression, while 40.9% (2297/5617) reported receiving no treatment. This information was missing for 44.2% (2480/5617) of participants. In addition, 51.4% (2889/5617) of participants reported taking medications or supplements at baseline, 4.6% (261/5617) reported taking none, and 44% (2469/5617) did not report this information. Among the 1449 (25.8%) participants who reported at least one prepregnancy condition, the three most common were COVID-19 (n=781), endometriosis (n=316), and a sexually transmitted illness (n=217).

Regarding family medical history, 9.2% (517/5617) of participants reported having a family member diagnosed with conditions or diseases during childhood, while 46.3% (2603/5617) did not, and 44.5% (2497/5617) did not provide this information. Of the 9.2% (517/5617) of participants that reported themselves, the baby's second biological parent, or anyone in their family having ever had any conditions or diseases diagnosed when they were a baby or in early childhood, the three most common diagnoses were heart defect (n=83), chromosome genetic disorder (n=58), and sickle cell disease or trait (n=21).

For delivery outcomes, 15.1% (849/5617) of participants reported giving birth, 0.9% (52/5617) reported a pregnancy loss, and 0.2% (12/5617) reported stillbirths. Delivery type information showed that 9.1% (512//5617) reported having spontaneous vaginal deliveries, 2.9% (163/5617) had unplanned cesarean sections, and 2.1% (117/5617) had planned cesarean sections. Forceps deliveries were reported by 0.1% (6/5617), vacuum extractions by 0.4% (20/5617), and vaginal births after cesarean by 0.4% (27/5617). Delivery type was not reported by 83.8% (4709/5617) of participants.

Finally, 14.4% (807/5617) of participants reported delivering in a hospital, 0.4% (22/5617) at home, and 0.3% (18/5617) in birthing centers. Delivery location data was unknown for 84.9% (4768/5617) of participants.

Engagement Trends and Survey Completion

The PowerMom platform collected a total of 17,123 individual surveys from 5617 pregnant and postpartum participants across all 50 US states, Puerto Rico, and Guam. Of the 5617 participants who completed the initial Screener Survey, 4031 (71.76%) completed the Intake Survey and 3148 (56.04%) completed the Health History Survey, which were the two baseline surveys administered. There were 1514 (26.95%) participants who completed at least one biweekly Health and Wellbeing Survey; 913 (16.25%) who completed the Delivery Survey; and 697 (12.41%) who completed the Postpartum Survey, which assessed mental health and was sent 6-8 weeks after the participant's reported delivery date.

Fraudulent Behavior

Fraud detection was an essential part of maintaining data integrity. Metrics such as unusually short completion times for eConsent or surveys, duplicate IP addresses, and repetitive data patterns were flagged. 580 of the flagged cases were determined to be fraud and excluded.

Enrollment Trends

Enrollment rates varied significantly from the platform's launch in 2021 through 2024. In the first year of PowerMom's launch, 70 participants were enrolled, with an average of 2.6 participants per week. As recruitment methods were refined and partnerships with pregnancy apps such as Philips Pregnancy+ and Mae Health were formed, the average enrollment rate increased substantially, reaching an average of 93 participants per week by 2024. Table 1 shows the enrollment trends by year, with a clear rise in participant engagement.

Table 1. Participant enrollment rates for the PowerMom study over time.

Year	Total participants (per year), n	Rate (per month)	Rate (per week)
2021 (starting on September 15, 2021)	55	15.7	3.6
2022	481	40.1	9.3
2023	1771	147.6	34.1
2024 (ending on August 22, 2024)	3310	427.1	99.4

Wearable Device Data

Of the 5617 participants, 1168 (20.8%) elected to share data from their wearable devices. The distribution of device use is as follows: 54.1% (632/1168) from Fitbit, 40.6% (474/1168) from Apple HealthKit, and 5.3% (62/1168 participants) from Google Fit. A total of 378,624 daily measurements were collected from participants, covering activity levels, sleep duration, and resting heart rate. These real-time biometric data points provide a detailed, continuous view of participants' health throughout their pregnancy journey.

EHR Data

Connecting EHR data was optional for PowerMom participants. A total of 96 (1.7%) participants chose to connect their EHR with the PowerMom platform, providing 276 distinct data points. These data points included diagnosis codes, lab results, medications, and procedures.

XSL•FO
RenderX

Table 2 presents the summary of the demographic, geographic, and access-to-care data for 5617 participants enrolled in the PowerMom study (2021-2024). The table includes age distribution, Hispanic or Latina identification, racial background, education levels, geographic regions (eg, Northeast and Midwest), and levels of access to maternity care based on the March of Dimes Maternity Care Deserts Dashboard. Data also include neighborhood deprivation as assessed by the ADI. Missing data are reported for each category.

Table 3 presents a summary of baseline health history and pregnancy-related characteristics for 5617 participants enrolled in the PowerMom study. The table includes pregnancy stage at enrollment, number of fetuses, previous pregnancies and births, prenatal care providers, and medication or supplement usage at baseline. In addition, the table details conditions such as prepregnancy and past pregnancy complications, mental health treatment (anxiety and depression), and family disease history. Delivery outcomes, types, and locations are also reported. Missing or unknown data percentages are indicated for each category to account for variability in reporting.

Table 1 presents participant enrollment rates for the PowerMom study, displayed by year. The table reports the total number of participants enrolled each year, along with calculated monthly and weekly enrollment rates. These trends highlight substantial increases in recruitment efficiency, particularly in 2023 and 2024, reflecting the impact of enhanced outreach strategies and platform optimization.

XSL•FO
RenderX

Table 2. Demographic and regional characteristics of PowerMom participants.

Characteristics	Values (N=5617), n (%)
Age category (years)	
<35	2723 (48.5)
≥35	1083 (19.3)
Missing	1811 (32.2)
Hispanic or Latina	
Yes	788 (14.0)
No	3134 (55.8)
Missing	1695 (30.2)
Racial background	
American Indian or Alaska Native	52 (0.9)
Asian	105 (1.9)
Black or African American or African	770 (13.7)
Hispanic, Latina, or Spanish (no other race and ethnicity specified)	478 (8.5)
Native Hawaiian or Other Pacific Islander	9 (0.2)
White	2175 (38.7)
Multiple races	333 (5.9)
Missing	1695 (30.2)
Region	
Northeast	635 (11.3)
South	1630 (29.0)
Midwest	880 (15.7)
West	853 (15.2)
Pacific	24 (0.4)
Puerto Rico	1 (<0.1)
Missing	1594 (28.4)
Education	
Graduate or postgraduate degree	703 (12.5)
College degree	1063 (18.9)
Some college or technical school	342 (6.1)
High school graduate or GED[a] (grade 12)	1545 (27.5)
Some high school (grades 9-11)	337 (6.0)
Missing	1627 (29.0)
Level of access to maternity care	
Moderate access to care/access to maternity care	3276 (58.3)
Maternity care desert/low access to care	244 (4.3)
Missing	2097 (37.3)
Neighborhood Atlas's ADI[b]	
≥75	919 (16.4)
≥50 and <75	1381 (24.6)
≥25 and <50	1108 (19.7)
<25	574 (10.2)

XSL•FO
RenderX

Characteristics	Values (N=5617), n (%)
Missing	1635 (29.1)

[a]GED: General Educational Development.
[b]ADI: Area Deprivation Index.

XSL•FO
RenderX

Table 3. Health history and pregnancy characteristics of PowerMom participants.

Characteristics	Values (N=5617), n (%)
Pregnancy stage at enrollment	
First trimester	1578 (28.1)
Second trimester	1493 (26.6)
Third trimester	909 (16.2)
Post delivery	736 (13.1)
Unknown	901 (16.0)
Number of fetuses	
1	3394 (60.4)
2	121 (2.2)
3	16 (0.3)
>3	15 (0.3)
Unknown	2071 (36.9)
Number of past pregnancies	
0	371 (6.6)
1	759 (13.5)
2	688 (12.2)
3	447 (8.0)
4	328 (5.8)
5	180 (3.2)
>5	266 (4.7)
Unknown	2578 (45.9)
Number of past births	
0	1118 (19.9)
1	923 (16.4)
2	571 (10.2)
3	260 (4.6)
4	111 (2)
5	58 (1)
>5	53 (0.9)
Unknown	2523 (44.9)
Prenatal care providers seen at baseline	
Doula	129 (2.3)
Homebirth	9 (0.2)
Physician	2104 (37.5)
Midwife	555 (9.9)
Nurse midwife	56 (1)
Nurse practitioner	36 (0.6)
Obstetrician	237 (4.2)
Other	143 (2.5)
None	447 (8.0)
Unknown	2511 (44.7)
Taking medications or supplements at baseline	

XSL•FO
RenderX

Characteristics	Values (N=5617), n (%)
No	261 (4.6)
Yes	2887 (51.4)
Unknown	2469 (44.0)
Receiving anxiety or depression treatment at baseline	
Anxiety and depression	550 (9.8)
Anxiety only	189 (3.4)
Depression only	102 (1.8)
Neither	2296 (40.9)
Unknown	2480 (44.2)
Parent or familial childhood condition or disease history	
No	2603 (46.3)
Yes	517 (9.2)
Unknown	2497 (44.5)
Prepregnancy condition	
No	1699 (30.2)
Yes	1449 (25.8)
Unknown	2469 (44.0)
Past pregnancy complication	
No	1575 (28.0)
Yes	1106 (19.7)
Not applicable	371 (6.6)
Unknown	2565 (45.7)
Pregnancy completion	
Gave birth	849 (15.1)
Pregnancy ended	52 (0.9)
Stillbirth	12 (0.2)
Unknown	4704 (83.7)
Delivery type	
Forceps delivery	5 (0.1)
Not applicable	64 (1.1)
Planned cesarean section	117 (2.1)
Spontaneous vaginal delivery	512 (9.1)
Unknown	4709 (83.8)
Unplanned C-section	163 (2.9)
VBAC[a]	27 (0.5)
Vacuum extraction	20 (0.4)
Delivery location	
Birthing center	18 (0.3)
Home	22 (0.4)
Hospital	807 (14.4)
Other	2 (0)

XSL•FO
RenderX

Characteristics	Values (N=5617), n (%)
Unknown	4768 (84.9)

[a]VBAC: vaginal births after cesarean.

Discussion

Overview

The PowerMom platform is a dynamic research tool designed to capture and store maternal health data, enabling the analysis of longitudinal and episodic data across all stages of pregnancy. This includes survey responses, EHR, and wearable sensor data from pregnant and postpartum individuals. PowerMom's ability to collect data from participants across all 50 US states, Puerto Rico, and Guam has facilitated the creation of a diverse dataset. The breadth and depth of these data allow for a detailed exploration of individual variations in pregnancies, helping researchers understand how these variations affect both maternal and fetal health. PowerMom's data collection supports advanced analytical approaches, including trend analysis, comparative studies, and machine learning techniques. This enables the platform to explore patterns and associations between different maternal health factors, including demographic variables, health behaviors, and outcomes, offering insights that can inform personalized approaches to maternal care.

Principal Findings

The PowerMom platform successfully engaged a diverse cohort of 5617 participants across all 50 US states, Puerto Rico, and Guam, demonstrating its effectiveness as a scalable tool for maternal health research. Of the 5617 participants, 3922 (69.8%) shared demographic data, revealing meaningful representation across age, racial and ethnic categories, and geographic regions. Nearly half (2723/5617, 48.5%) of participants were younger than 35 years, while 19.3% (1083/5617) were 35 years or older. Participants identifying as Black or African American accounted for 13.7% (770/5617) of the cohort, while Hispanic or Latina individuals represented 14.0% (788/5617). These data underscore the platform's success in reaching historically underrepresented populations in maternal health research.

Recruitment efforts improved significantly over time, with enrollment increasing from 55 participants in late 2021 to 3310 in 2024. By 2024, the platform reached an average enrollment rate of 99.4 participants per week, reflecting the impact of enhanced recruitment strategies and partnerships with apps such as Philips Pregnancy+ and Mae Health.

Comprehensive data collection was achieved through surveys, wearable devices, and EHR integration. A total of 17,123 surveys were completed across the cohort. The initial Screener Survey had a completion rate of 100%. However, subsequent surveys saw a decline in completion rates: 71.8% (4033/5617) for the Intake Survey, 56.0% (3145/5617) for the Health History Survey, and 12.4% (697/5617) for the Postpartum Survey. These trends highlight the need for strategies to sustain participant engagement over time.

Wearable device data were provided by 1168 participants, yielding more than 378,000 daily biometric measurements, including activity levels, sleep duration, and resting heart rate. These continuous data provide a granular, real-time view of maternal health throughout pregnancy and postpartum stages. In addition, 96 participants connected their EHRs to the platform, contributing 276 distinct data points, such as diagnoses, medications, and laboratory results. The rationale for the low EHR connection rate was not a metric of our study. However, we observed that the low willingness to share EHR data among participants (1.7%) is consistent with findings in the literature, such as the study by Kim et al [30]. Key factors cited in that study include privacy concerns, the lack of trust in data-sharing platforms, and technical challenges in linking EHRs. In addition, the demographic diversity within the cohort may have contributed to this hesitancy, as some populations with historical mistrust of health care systems might have been more hesitant to share sensitive information. Understanding these barriers is critical for future studies aiming to integrate EHR data more comprehensively.

The analysis of pregnancy and delivery-related characteristics further enriched the dataset. Approximately 28.1% (1578/5617) of participants enrolled during the first trimester, and 15.1% (849/5617) completed their pregnancies during the study period. Delivery data revealed that 9.1% (512/5617) of participants reported experiencing spontaneous vaginal births, while 2.1% (117/5617) and 2.9% (163/5617) underwent planned and unplanned cesarean deliveries, respectively. Mental health data indicated that 14.9% of participants reported receiving treatment for anxiety or depression at baseline.

Despite the platform's successes, notable data gaps were observed, including missing demographic information (eg, age data missing for 1811/5617, 32.2% of participants) and survey attrition. Due to attrition, only 15.1% (849/5617) of participants reported completing their pregnancies within the study period, limiting delivery-related analyses. These findings highlight both the potential and challenges of using digital platforms for maternal health research.

Comparison With Previous Work

The PowerMom platform builds upon lessons learned from our earlier digital pregnancy study, the Healthy Pregnancy Study [31]. The Healthy Pregnancy Study was embedded within a pregnancy app, allowing for seamless recruitment and the enrollment of more than 4000 participants. The study generated more than 14,000 individual surveys and more than 107,000 daily health measurements, including sleep, activity, and heart rate. The Healthy Pregnancy Study also highlighted key disparities in maternal health care access and treatment. For example, Black and rural participants reported lower use of prenatal vitamins, antiemetics, and antidepressants compared with their non-Black and urban counterparts [32]. PowerMom sought to build on these findings by enabling widespread participation from underrepresented communities and offering

XSL•FO
RenderX

valuable insights into how structural and health care factors affect maternal health outcomes.

Challenges Encountered

Substudy Enrollment Challenges

PowerMom's foundational protocol allows for the integration of additional substudies targeting specific research questions. However, when substudies have additional eligibility criteria, ensuring that participants can easily meet those criteria can be challenging. For example, incomplete baseline surveys or unmet eligibility criteria can hinder participant enrollment in substudies. Streamlining the baseline survey process and providing clear prompts for participants to join substudies have proven to be effective strategies in minimizing these challenges and boost substudy engagement.

Fraudulent Activity

One unexpected challenge during one of the PowerMom substudies was the fraudulent redemption of gift cards offered as participation incentives. The platform's automated system was exploited by 580 participants who enrolled fraudulently to claim gift cards, leading to a rapid spike in enrollment in both the baseline study and PowerMom Connect. To counter this, we introduced anomaly detection measures, including monitoring the time taken to complete eConsent and flagging participants with irregular patterns of completion. These measures successfully curtailed fraudulent activity. Future studies should incorporate automated fraud detection tools from the outset to prevent disruptions.

Attrition in Survey Completion

Participant attrition in completing surveys over time remains a key limitation of the PowerMom platform. While initial survey completion rates were high, subsequent surveys saw lower completion rates, potentially due to participant fatigue or competing priorities. Enhanced engagement strategies, such as personalized incentives, interactive content, and dynamic reminders, are needed to reduce participant burden and improve survey retention in future studies.

Limitations

While the PowerMom platform leverages decentralized and digital approaches to reach a large and diverse population, it is not without limitations. Despite the increasing penetration of smartphones in the United States (with 92% of adults owning a smartphone as of 2023 [33]), there remain disparities in broadband access and digital literacy, particularly in rural areas and among lower-income populations. These barriers may prevent certain individuals from participating fully in a purely digital, mobile-based platform. However, PowerMom's decentralized design mitigates the need for participants to live near academic medical centers. Therefore, increasing access to maternal health research for those in remote or underserved areas.

In addition, as noted in the *Attrition in Survey Completion* section, participant attrition in survey completion remains a challenge for longitudinal studies such as PowerMom. Ensuring sustained engagement throughout pregnancy and postpartum periods is critical for comprehensive data collection.

Lessons Learned

The relaunch of PowerMom in 2021 highlighted several key lessons for future digital health studies:

1. Scalability and adaptability: The platform's scalability and adaptability have proven essential in accommodating diverse participant needs. Culturally tailored outreach strategies and bilingual interfaces were particularly successful in engaging historically underrepresented populations.
2. Continuous data collection: The ability to collect real-world, continuous data from wearables and EHR systems provides a comprehensive view of maternal health that surpasses traditional, episodic clinic visits. This real-time data collection allows for more precise tracking of health changes over time.
3. Fraud detection and prevention: The experience of dealing with fraudulent activity emphasized the importance of implementing robust anomaly detection systems to protect the integrity of participant data and incentives. Future studies should incorporate automated fraud detection tools from the outset to avoid disruptions.
4. Participant retention: Attrition in survey completion is a challenge in longitudinal digital health studies. Future studies should incorporate strategies to maintain participant engagement, such as personalized incentives, interactive content, or dynamic reminders, which may help reduce participant fatigue and improve data completeness.

Conclusions

The PowerMom platform has demonstrated its potential as an innovative, scalable, and inclusive maternal health research tool. By leveraging decentralized, mobile-based recruitment and data collection, PowerMom successfully engaged a diverse cohort of 5617 participants across all 50 US states, Puerto Rico, and Guam. The platform achieved representation across critical demographic groups, including underrepresented racial and ethnic populations and individuals residing in areas with limited access to maternity care. Despite the platform's successes, challenges such as survey attrition, missing data, and fraudulent activity underscored the importance of implementing robust engagement and data quality strategies in future digital health studies.

Looking ahead, PowerMom's flexible and scalable architecture positions it as a transformative model for maternal health research. By continuing to enhance participant engagement strategies and expanding data collection capabilities to include genomic, environmental, and social determinants of health, PowerMom can generate actionable insights to address maternal health disparities. Ultimately, this platform represents a significant step toward improving maternal health outcomes, promoting health equity, and empowering diverse populations to actively participate in research that directly impacts their care.

XSL•FO
RenderX

Acknowledgments

This work received partial support from the National Center for Advancing Translational Sciences (UM1TR004407) and The Patrick J McGovern Foundation.

Data Availability

The datasets generated and analyzed during the PowerMom study are not publicly available due to privacy and confidentiality restrictions. The study collected sensitive health and demographic data from participants, which are deidentified but remain protected under institutional review board protocols. Researchers who wish to access the data for legitimate academic purposes can submit a formal request to PowerMom principal investigator, TA. Requests will be evaluated on a case-by-case basis to ensure compliance with ethical guidelines and data use agreements. The code used for data analysis is available upon request from the corresponding author.

Conflicts of Interest

KBM is a consultant for CareEvolution.

Multimedia Appendix 1

Participant flow and in-app surveys from the PowerMom study, conducted between 2021 and 2024. This figure illustrates the step-by-step flow of participants from recruitment through study completion, including survey engagement. Surveys captured demographic, health history, and pregnancy-related data at various stages, leveraging the PowerMom app's user-friendly interface to enhance participant engagement and data collection.

[DOCX File , 12642 KB - formative_v9i1e70149_app1.docx]

References

1. Hoyert DL. Maternal mortality rates in the United States, 2020. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/nchs/data/hestat/maternal-mortality/2020/maternal-mortality-rates-2020.htm [accessed 2024-12-02]
2. Hirai AH, Owens PL, Reid LD, Vladutiu CJ, Main EK. Trends in severe maternal morbidity in the US across the transition to ICD-10-CM/PCS from 2012-2019. JAMA Netw Open 2022;5(7):e2222966 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.22966] [Medline: 35900764]
3. Petersen EE, Davis NL, Goodman D, Cox S, Syverson C, Seed K, et al. Racial/ethnic disparities in pregnancy-related deaths - United States, 2007-2016. MMWR Morb Mortal Wkly Rep 2019;68(35):762-765 [FREE Full text] [doi: 10.15585/mmwr.mm6835a3] [Medline: 31487273]
4. Creanga AA, Syverson C, Seed K, Callaghan WM. Pregnancy-related mortality in the United States, 2011-2013. Obstet Gynecol 2017;130(2):366-373 [FREE Full text] [doi: 10.1097/AOG.0000000000002114] [Medline: 28697109]
5. Iturralde E, Hsiao CA, Nkemere L, Kubo A, Sterling SA, Flanagan T, et al. Engagement in perinatal depression treatment: a qualitative study of barriers across and within racial/ethnic groups. BMC Pregnancy Childbirth 2021;21(1):512 [FREE Full text] [doi: 10.1186/s12884-021-03969-1] [Medline: 34271852]
6. Admon LK, Winkelman TNA, Zivin K, Terplan M, Mhyre JM, Dalton VK. Racial and ethnic disparities in the incidence of severe maternal morbidity in the United States, 2012-2015. Obstet Gynecol 2018;132(5):1158-1166. [doi: 10.1097/AOG.0000000000002937] [Medline: 30303912]
7. Bromley E, Nunes A, Phipps MG. Disparities in pregnancy healthcare utilization between Hispanic and non-Hispanic white women in Rhode Island. Matern Child Health J 2012;16(8):1576-1582. [doi: 10.1007/s10995-011-0850-5] [Medline: 21769588]
8. Chantler TEA, Lees A, Moxon ER, Mant D, Pollard AJ, Fiztpatrick R. The role familiarity with science and medicine plays in parents' decision making about enrolling a child in vaccine research. Qual Health Res 2007;17(3):311-322. [doi: 10.1177/1049732306298561] [Medline: 17301340]
9. Ayoub JJ, Abiad M, Forman MR, MINA Collaborators, Honein-AbouHaidar G, Naja F. The interaction of personal, contextual, and study characteristics and their effect on recruitment and participation of pregnant women in research: a qualitative study in Lebanon. BMC Med Res Methodol 2018;18(1):155 [FREE Full text] [doi: 10.1186/s12874-018-0616-5] [Medline: 30497391]
10. Chinn JJ, Eisenberg E, Artis Dickerson S, King RB, Chakhtoura N, Lim IAL, et al. Maternal mortality in the United States: research gaps, opportunities, and priorities. Am J Obstet Gynecol 2020;223(4):486-492.e6 [FREE Full text] [doi: 10.1016/j.ajog.2020.07.021] [Medline: 32682858]
11. Zullig LL, Fortune-Britt AG, Rao S, Tyree SD, Godley PA, Carpenter WR. Enrollment and racial disparities in cancer treatment clinical trials in North Carolina. N C Med J 2016;77(1):52-58 [FREE Full text] [doi: 10.18043/ncm.77.1.52] [Medline: 26763244]

XSL•FO
RenderX

12. Woodall A, Morgan C, Sloan C, Howard L. Barriers to participation in mental health research: are there specific gender, ethnicity and age related barriers? BMC Psychiatry 2010;10:103 [FREE Full text] [doi: 10.1186/1471-244X-10-103] [Medline: 21126334]
13. Williams B, Entwistle V, Haddow G, Wells M. Promoting research participation: why not advertise altruism? Soc Sci Med 2008;66(7):1451-1456. [doi: 10.1016/j.socscimed.2007.12.013] [Medline: 18222579]
14. Choudhury A, Nimbarte A. Editorial: mobile health interventions to address maternal health: ideas, concepts, and interventions. Front Digit Health 2024;6:1378416 [FREE Full text] [doi: 10.3389/fdgth.2024.1378416] [Medline: 38486918]
15. Stone JS, Hirshberg A. Telemedicine and digital health solutions in intrapartum and postpartum care. Curr Obstet Gynecol Rep 2024;13:188-194 [FREE Full text] [doi: 10.1007/s13669-024-00394-x]
16. Nadarzynski T, Knights N, Husbands D, Graham CA, Llewellyn CD, Buchanan T, et al. Achieving health equity through conversational AI: a roadmap for design and implementation of inclusive chatbots in healthcare. PLOS Digit Health 2024;3(5):e0000492 [FREE Full text] [doi: 10.1371/journal.pdig.0000492] [Medline: 38696359]
17. Leonard A, Hutchesson M, Patterson A, Chalmers K, Collins C. Recruitment and retention of young women into nutrition research studies: practical considerations. Trials 2014;15:23 [FREE Full text] [doi: 10.1186/1745-6215-15-23] [Medline: 24433229]
18. Frampton GK, Shepherd J, Pickett K, Griffiths G, Wyatt JC. Digital tools for the recruitment and retention of participants in randomised controlled trials: a systematic map. Trials 2020;21(1):478 [FREE Full text] [doi: 10.1186/s13063-020-04358-3] [Medline: 32498690]
19. Bower P, Brueton V, Gamble C, Treweek S, Smith CT, Young B, et al. Interventions to improve recruitment and retention in clinical trials: a survey and workshop to assess current practice and future priorities. Trials 2014;15:399 [FREE Full text] [doi: 10.1186/1745-6215-15-399] [Medline: 25322807]
20. Cummins KM, Brumback T, Chung T, Moore RC, Henthorn T, Eberson S, et al. Acceptability, validity, and engagement with a mobile app for frequent, continuous multiyear assessment of youth health behaviors (mNCANDA): mixed methods study. JMIR Mhealth Uhealth 2021;9(2):e24472 [FREE Full text] [doi: 10.2196/24472] [Medline: 33565988]
21. Quer G, Gouda P, Galarnyk M, Topol EJ, Steinhubl SR. Inter- and intraindividual variability in daily resting heart rate and its associations with age, sex, sleep, BMI, and time of year: Retrospective, longitudinal cohort study of 92,457 adults. PLoS One 2020;15(2):e0227709. [doi: 10.1371/journal.pone.0227709] [Medline: 32023264]
22. Jaiswal SJ, Quer G, Galarnyk M, Steinhubl SR, Topol EJ, Owens RL. Association of sleep duration and variability with body mass index: sleep measurements in a large US population of wearable sensor users. JAMA Intern Med 2020;180(12):1694-1696 [FREE Full text] [doi: 10.1001/jamainternmed.2020.2834] [Medline: 32926073]
23. Quer G, Radin JM, Gadaleta M, Baca-Motes K, Ariniello L, Ramos E, et al. Wearable sensor data and self-reported symptoms for COVID-19 detection. Nat Med 2021;27(1):73-77. [doi: 10.1038/s41591-020-1123-x] [Medline: 33122860]
24. Gadaleta M, Radin JM, Baca-Motes K, Ramos E, Kheterpal V, Topol EJ, et al. Passive detection of COVID-19 with wearable sensors and explainable machine learning algorithms. NPJ Digit Med 2021;4(1):166 [FREE Full text] [doi: 10.1038/s41746-021-00533-1] [Medline: 34880366]
25. Radin JM, Quer G, Ramos E, Baca-Motes K, Gadaleta M, Topol EJ, et al. Assessment of prolonged physiological and behavioral changes associated with COVID-19 infection. JAMA Netw Open 2021;4(7):e2115959 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.15959] [Medline: 34232306]
26. Quer G, Gadaleta M, Radin JM, Andersen KG, Baca-Motes K, Ramos E, et al. Inter-individual variation in objective measure of reactogenicity following COVID-19 vaccination via smartwatches and fitness bands. NPJ Digit Med 2022;5(1):49 [FREE Full text] [doi: 10.1038/s41746-022-00591-z] [Medline: 35440684]
27. New research consortium to tackle maternal health disparities with digital technologies. Scripps Research. 2021. URL: https://www.scripps.edu/news-and-events/press-room/2021/20211207-ajayi-powermom-consortium.html [accessed 2024-12-02]
28. Access to maternity care. MODHealth. 2024. URL: https://mod.health360-insights.io [accessed 2024-12-02]
29. Multiple births. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/nchs/fastats/multiple.htm [accessed 2024-12-02]
30. Kim KK, Sankar P, Wilson MD, Haynes SC. Factors affecting willingness to share electronic health data among California consumers. BMC Med Ethics 2017;18(1):25 [FREE Full text] [doi: 10.1186/s12910-017-0185-x] [Medline: 28376801]
31. Radin JM, Steinhubl SR, Su AI, Bhargava H, Greenberg B, Bot BM, et al. The healthy pregnancy research program: transforming pregnancy research through a ResearchKit app. NPJ Digit Med 2018;1:45 [FREE Full text] [doi: 10.1038/s41746-018-0052-2] [Medline: 31304325]
32. Ajayi T, Pawelek J, Bhargava H, Faksh A, Radin J. Self-reported medication use across racial and rural or urban subgroups of people who are pregnant in the United States: decentralized app-based cohort study. JMIR Form Res 2023;7:e50867 [FREE Full text] [doi: 10.2196/50867] [Medline: 38015604]
33. Mobile fact sheet. Pew Research Center. 2024. URL: https://tinyurl.com/y3ykwta4 [accessed 2024-09-10]

XSL•FO
RenderX

Abbreviations

ADI: Area Deprivation Index
eConsent: electronic informed consent
EHR: electronic health record
HIPAA: Health Insurance Portability and Accountability Act

Edited by A Mavragani; submitted 16.12.24; peer-reviewed by B Senst; comments to author 20.01.25; revised version received 16.02.25; accepted 25.02.25; published 07.04.25.

Please cite as:
Ajayi T, Kueper J, Ariniello L, Ho D, Delgado F, Beal M, Waalen J, Baca Motes K, Ramos E
Digital Health Platform for Maternal Health: Design, Recruitment Strategies, and Lessons Learned From the PowerMom Observational Cohort Study
JMIR Form Res 2025;9:e70149
URL: https://formative.jmir.org/2025/1/e70149
doi:10.2196/70149
PMID:

©Toluwalase Ajayi, Jacqueline Kueper, Lauren Ariniello, Diana Ho, Felipe Delgado, Matthew Beal, Jill Waalen, Katie Baca Motes, Edward Ramos. Originally published in JMIR Formative Research (https://formative.jmir.org), 07.04.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Use of Mobile Apps in Adolescent Psychotherapy: Assessment of Psychotherapists' Perspectives

Sarah Wüllner[1], Dr rer nat; Katharin Hermenau[1], Dr rer nat; Tobias Hecker[2,3], Dr rer nat; Michael Siniatchkin[1,4], Dr med

[1]University Clinic of Child and Adolescent Psychiatry and Psychotherapy, Medical School EWL, Protestant Hospital Bethel, Bielefeld University, Bielefeld, Germany

[2]Institute for Interdisciplinary Conflict and Violence Research, Bielefeld University, Bielefeld, Germany

[3]Department of Psychology, Faculty of Psychology and Sports Science, Bielefeld University, Bielefeld, Germany

[4]Department of Psychiatry, Psychotherapy and Psychosomatics of Children and Adolescents, University Hospital Rheinisch-Westfälische Technische Hochschule Aachen, Aachen, Germany

Corresponding Author:
Sarah Wüllner, Dr rer nat
University Clinic of Child and Adolescent Psychiatry and Psychotherapy
Medical School EWL, Protestant Hospital Bethel
Bielefeld University
Remterweg 13a
Bielefeld, 33617
Germany
Phone: 49 5211063166
Email: sarah.wuellner@uni-bielefeld.de

Abstract

Background: Therapy-accompanying mental health apps can play an important role in the psychotherapeutic treatment of adolescents. They can enhance adolescents' engagement and autonomy, provide immediate support in critical situations, and positively influence the therapeutic working alliance. Nevertheless, mental health apps are rarely used by psychotherapists. Furthermore, due to the limited or nonexistent use of apps in psychotherapy, little is known about the actual barriers and drivers affecting their integration into psychotherapists' daily routines. To better understand how mental health apps should be designed for practical use, it is essential to explore psychotherapists' perspectives on key app features and characteristics, as well as the factors influencing their integration into clinical practice.

Objective: This study aims to analyze which app features and characteristics are essential for psychotherapists to use a mobile app in psychotherapy with adolescents and to identify the key drivers and barriers influencing the integration of a psychotherapeutic app from the psychotherapists' perspectives.

Methods: We conducted 3 feasibility studies using Steps, a transdiagnostic, therapy-accompanying app for adolescents, across 3 different psychotherapeutic treatment contexts: inpatient treatment, treatment in psychiatric outpatient clinics, and outpatient treatment with psychotherapists in private practice. All studies followed a qualitative quasi-experimental design. Participants provided information on their age, occupation, years of work experience, media affinity, attitudes toward psychotherapeutic apps, perceived app quality and feasibility, and the implementation process of the therapy-accompanying app. Qualitative data were analyzed using deductive qualitative content analysis. A total of 40 mental health professionals participated across the 3 studies (study 1: n=18; study 2: n=13; study 3: n=9).

Results: Study participation and app usage rates were low across all studies. Six core features for a transdiagnostic, therapy-accompanying app were identified: mood checks, library, reminders, goals and tasks, emergency kit, and questionnaires. Additionally, the integration of mental health apps into daily routines was influenced by various drivers and barriers. The most significant barriers included technological issues and practical constraints, such as limited time and resources. The most important driver was the perceived improvement in treatment quality.

Conclusions: Overall, psychotherapists were generally open to using a therapy-accompanying mental health app. However, study participation and app usage remained low. As psychotherapists act as gatekeepers for patients' use of mental health apps, their needs should be prioritized in the development and implementation of such apps.

Trial Registration: German Clinical Trials Register DRKS00031258; https://drks.de/search/en/trial/DRKS00031258/details

XSL•FO
RenderX

(JMIR Form Res 2025;9:e65788) doi:10.2196/65788

KEYWORDS

mental health app; psychotherapy; adolescent; mHealth; youth; feasibility; implementation; app features; barriers; drivers

Introduction

Background

There is a high risk of developing mental disorders during adolescence, which may persist into adulthood [1-3]. According to the World Health Organization (WHO), one in seven 10- to 19-year-olds experiences a mental disorder [4]. In addition, the rates of treatment nonresponse and therapy dropout are high among adolescents [5-7]. A promising way to address these challenges in psychotherapeutic care is the use of therapy-accompanying mental health apps (MHAs) [8-10]. At the latest, the pandemic has demonstrated that MHAs can play an important role in the psychotherapeutic treatment of adolescents [11-13] and are widely accepted by them [11,14,15].

Therapy-accompanying MHAs are used as an adjunct to psychotherapy [16]. They can offer specific manualized treatments or therapy tools that can be individually incorporated into psychotherapy [17,18]. The majority of available MHAs focus on specific disorders rather than a transdiagnostic therapy approach [12,17-19]. However, most MHAs have demonstrated the potential to provide transdiagnostic support beyond their primarily targeted disorder, as each contains fundamentally transdiagnostic features [18,20,21]. Additionally, to date, all therapy-accompanying apps have been designed for outpatient treatment [16,18]. To the best of our knowledge, no MHA has been specifically designed for inpatient treatment.

Therapy-accompanying MHAs offer several advantages in adolescent psychotherapy. They can enhance engagement [8-10] and promote greater autonomy [22,23]. Moreover, MHAs provide immediate support in critical situations, such as self-harm or acute suicidal ideation [16]. As a result, adolescents experience increased self-efficacy and improved therapeutic transfer to everyday life [8,16,17]. Furthermore, MHAs can have a beneficial influence on the therapeutic working alliance [8]. For psychotherapists, MHAs can facilitate documentation and monitoring of therapeutic progress, as well as communication with patients [24]. Another important aspect of integrating MHAs into psychotherapy is their effectiveness [24,25]. Previous research has shown that MHAs, in general, are effective in adolescent psychotherapy [18,26-35].

Although there are many reasons to use MHAs as an adjunct to psychotherapy with adolescents, they are rarely utilized by psychotherapists [24,36-41]. In 2019, only 9% of German health care professionals had tried an MHA in their psychotherapeutic practice [24]. The COVID-19 pandemic led to a slight increase in usage [37,39]; however, adoption remains low, with only 29% of German health care professionals reporting that they have used an app in their practice [39]. This raises the question: “Where does the discrepancy between the perceived benefits of MHAs and their limited use in psychotherapy originate?”

To better understand this discrepancy, it is essential to identify the factors associated with the integration of MHAs into psychotherapists’ working routines. Feijt et al [42] developed a model to explain the factors influencing psychotherapists’ use of e-mental health: the Levels of Adoption of e-Mental Health (LAMH) model. This model categorizes influencing factors into different domains, including general characteristics, perceived drivers and barriers, and requirements for change. All factors are categorized according to different levels of e-mental health usage. The authors distinguish 5 levels of use, ranging from “no use” (level 1) to “innovative use” (level 5). According to previous research [24,39], psychotherapists generally fall between level 1 (no use) and level 3 (passive users) in their use of MHAs [42]. Consequently, our study focuses on the lower levels of adoption (levels 1-3), excluding level 4 (active use) and level 5 (innovative use). The selection of general characteristics, perceived drivers and barriers, and requirements for change mentioned in Feijt et al’s [42] is based on qualitative research results. Nevertheless, these factors have also been identified in other studies [9,24,25,39,43-48]. The LAMH model, adapted for this study and supplemented by our findings, is presented in Figure 1. It also includes corresponding references for each factor in the adapted LAMH model.

XSL•FO
RenderX

Figure 1. Levels of Adoption of e-Mental Health (LAMH) adapted for mental health app (MHA) use (aLAMH). The model is adapted from the LAMH model of Feijt et al. Factors that are marked with an "a" were added to the original model based on the current study results. Factors mentioned in study 1 are marked with "b"; factors mentioned in study 2 are marked with "c"; and factors mentioned in study 3 are marked with "d." The factors in the model were also mentioned in other studies as relevant association factors with psychotherapists' mental health app use.

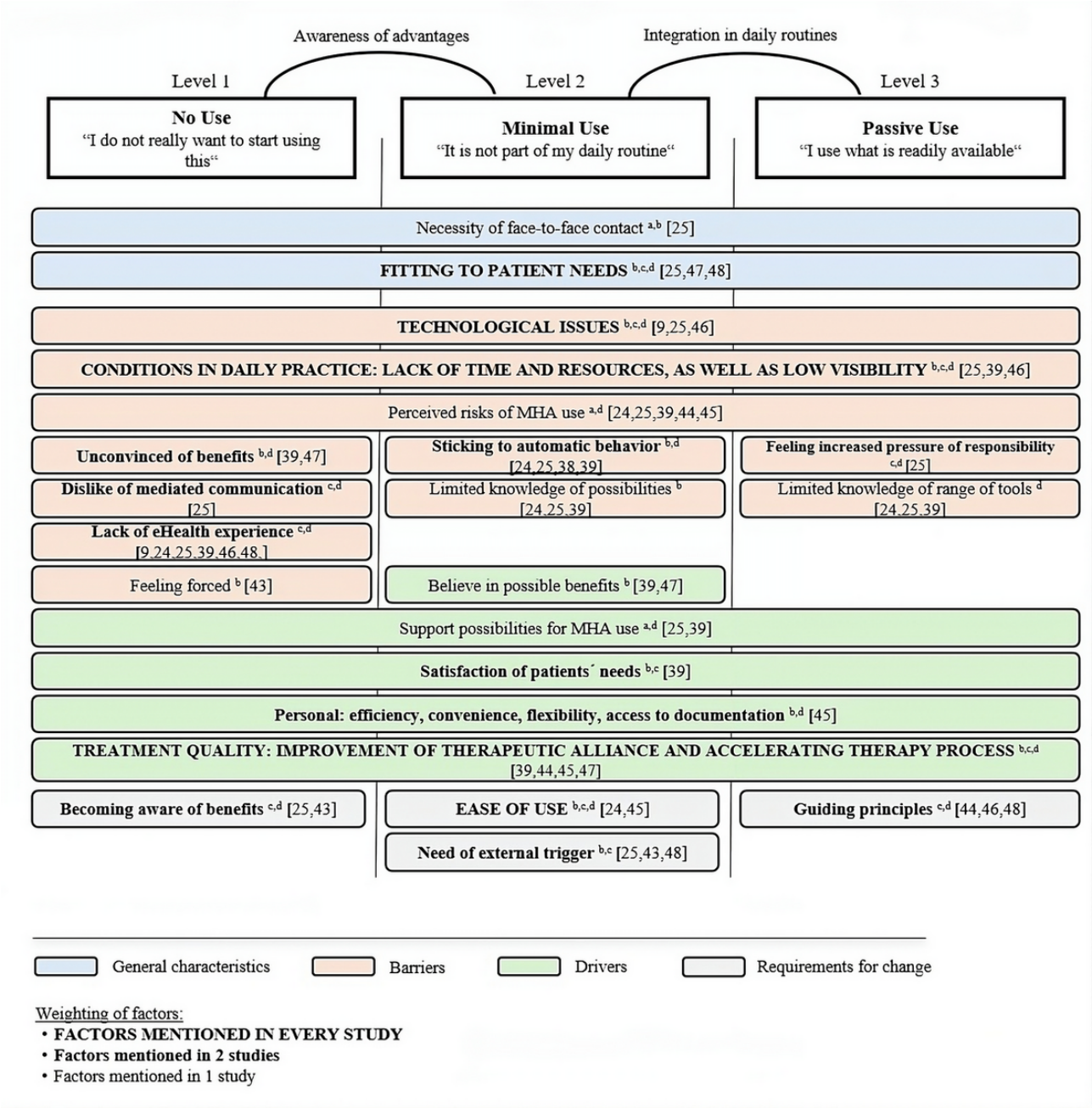

Like the development of the LAMH model, previous research on drivers and barriers has primarily focused on the theoretical perspectives of psychotherapists [9,24,44,45,47]. Furthermore, due to the limited or nonexistent use of apps in psychotherapy [37,39], little is known about the actual barriers and drivers influencing the integration of MHAs into psychotherapists' working routines. Therefore, gaining a more practical understanding of these factors is essential.

In addition to questions about potential drivers and barriers to MHA integration, the design and features of MHAs may play a crucial role in psychotherapists' use of these apps. What if psychotherapists do not use MHAs because the available apps do not sufficiently meet the needs of adolescent psychotherapeutic treatments? To better understand how MHAs should be designed for actual use in psychotherapists' working routines, it is essential to explore their perspectives on the key features and characteristics of therapy-accompanying MHAs.

Objectives

This study aims is to investigate how a therapy-accompanying MHA should be designed for successful integration into adolescent psychotherapy across different psychotherapeutic contexts. To assess psychotherapists' perspectives in a more practical way, we conducted a needs evaluation within the initial feasibility pilot testing of a transdiagnostic, therapy-accompanying app for adolescents in 3 different treatment settings: inpatient treatment, psychiatric outpatient

XSL•FO
RenderX

clinics, and outpatient treatment with private practice psychotherapists. We analyzed the essential app features and characteristics required for psychotherapists to use an MHA in adolescent psychotherapy, as well as the key drivers and barriers influencing its integration from the psychotherapists' perspectives.

Methods

Overview

Three feasibility studies were conducted in different settings using the transdiagnostic, therapy-accompanying app *Steps*. The app was designed as an adjunct to regular psychotherapeutic treatment for adolescents. It was developed through a participatory process by Circumradius GmbH and the Protestant Hospital Bethel in Bielefeld. Steps offered interconnected versions for both patients and their psychotherapists. While patients used the Steps app on their own Android-based smartphones, psychotherapists accessed a web-based version (see Figures 2 and 3). The app included various transdiagnostic features (Table 1). For implementation, psychotherapists were not required to follow specific usage rules or treatment manuals; they could independently decide to what extent they wanted to integrate the app into their psychotherapy sessions. All studies followed a qualitative quasi-experimental study design. In all studies, data were collected from both psychotherapists and patients. However, due to the objectives of this article and the comprehensiveness of the collected data, patient data were not included in the current analyses. Additionally, only methods and results relevant to this analysis are reported in this article.

Figure 2. Screenshots of different app features of the patient version of the app Steps: goals and tasks (left), emergency kit (center), and the main therapy goal (right).

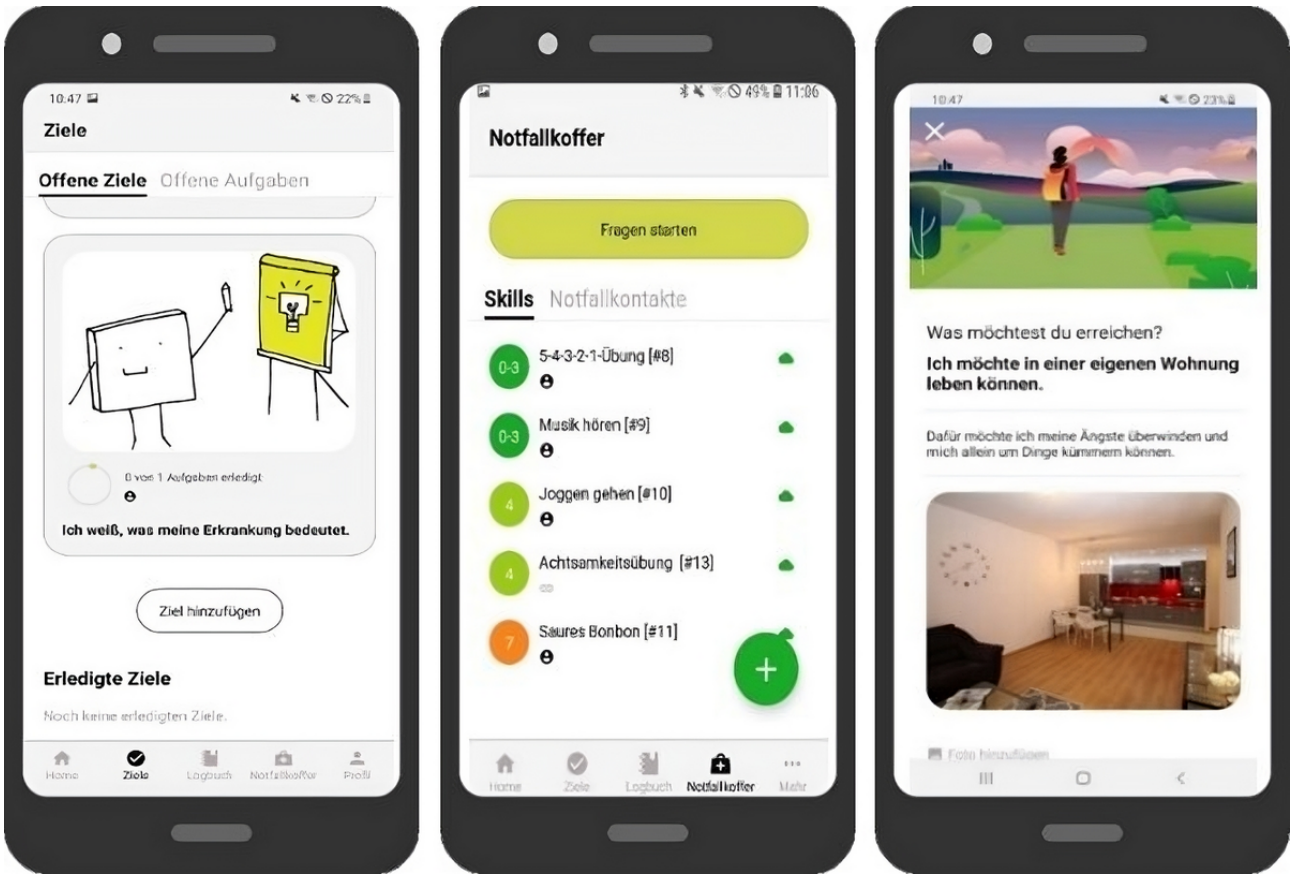

XSL•FO
RenderX

Figure 3. Displays of different app features of the web-based psychotherapist version of the app Steps: list of skills in the emergency kit (left) and goals and tasks (right).

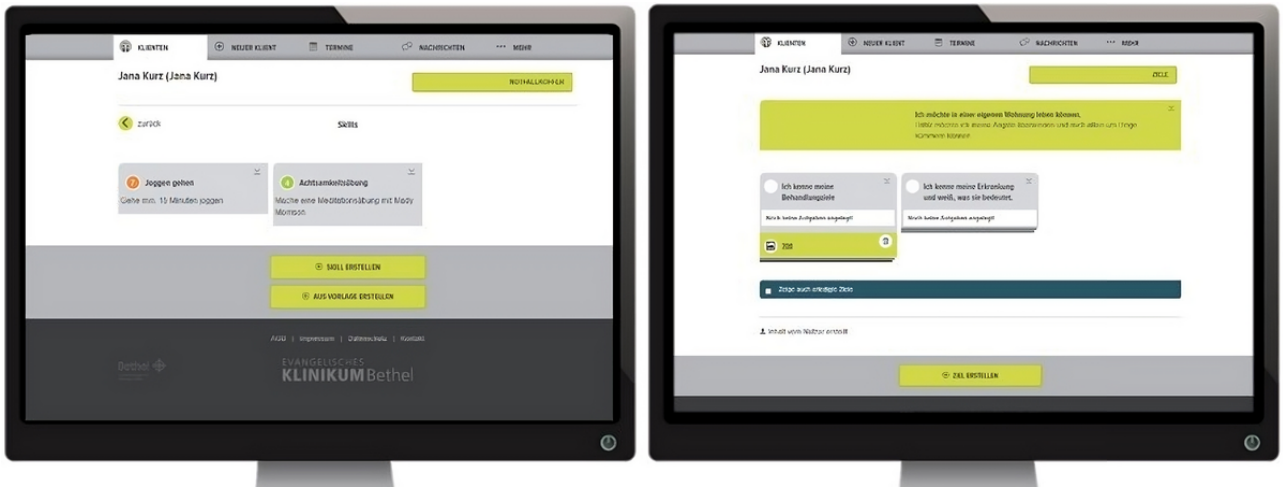

Table 1. Overview of the provided app features in the therapy-accompanying and transdiagnostic app Steps for adolescent psychotherapy.

App feature	Description
Mood checks	• Psychotherapists can customize mood checks for patients (type and number of items, response format, frequency of queries). • Patients are prompted and reminded by the app to complete the mood checks. • In the psychotherapist version, a graphical overview of the mood checks is available.
Main therapy goal	• Patients can record what they want to achieve through their psychotherapy. • Photos can be uploaded to visualize patients' therapy motivation.
Intermediate therapy goals	• Intermediate therapy goals can be set to structure or visualize therapy progress. • Patients can complete individual tasks to achieve intermediate goals.
Tasks	• Psychotherapists can set tasks for their patients, for example, to complete between therapy sessions. • When patients have completed a task, they receive an automated reflection query. Therapists can directly access the information entered and easily refer to the situation-related information in the next therapy session.
Logbook	• Mood checks done by patients are saved in the logbook. • Patients can also create free logbook entries to record important events, or to document the therapy progress.
Questionnaire manager	• Psychotherapists can set up questionnaires for their patients to complete directly in the app. • Questionnaires can be customized by psychotherapists (type and number of items, response format, and frequency of queries). • Answers to questionnaires can be downloaded directly in the psychotherapist version.
Emergency kit	• Skills and emergency contacts can be saved in the emergency kit. • Skills are sorted according to stress levels from 1 (low stress) to 10 (maximum stress). • Photos, home page links, or address details can be added to skill entries. • Optional: activation of queries for analysis of undesirable behavior (eg, nonsuicidal self-injuries).

Design and Procedure

The first study examined the feasibility of using Steps in inpatient treatment for adolescents. Adolescents admitted as inpatients who consented to participate in the study received a clinic-owned smartphone for their treatment. This smartphone included the Steps app along with other treatment-relevant apps and was available for use throughout their entire treatment period. The entire staff of the adolescent ward was trained in app usage and instructed to integrate Steps into the treatment process. All employees were invited to participate in the study on a voluntary basis. Psychotherapists completed a baseline survey at the start of their participation and a follow-up questionnaire at the end of the data collection period. For other employees, all questions were consolidated into a single survey conducted at the end of the project. Additionally, 4 employees participated in qualitative interviews assessing the app's quality and the implementation process of Steps.

The second study was conducted in psychiatric outpatient clinics. All staff members from the participating clinics were invited to take part. Based on the participation challenges

XSL•FO
RenderX

encountered in the first study, psychotherapists were offered 2 options for participation: (1) actively using the app in treatment with adolescents and providing feedback on their experiences, or (2) providing feedback on Steps after a brief introduction without directly using the program. Each participant received individual app training to familiarize themselves with the app and its application in therapy. If psychotherapists incorporated the app into their treatment, participating patients were required to install Steps on their own smartphones. If needed, adolescents could borrow a clinic-owned smartphone for study participation. Data were collected from participating employees at the start of their study involvement and again at the end of the data collection period using the open-source online survey tool LimeSurvey (LimeSurvey Team) [49]. The study was preregistered in the German Clinical Trials Register (DRKS00031258).

The third study was conducted with psychotherapists in private practice. Interested psychotherapists were recruited through the Association of Child and Adolescent Psychotherapists of East Westphalia-Lippe. The study procedure was consistent with that of the second study, with one difference: data for the second assessment were collected through qualitative interviews.

Ethical Considerations

All 3 studies received positive ethical approvals. Studies 1 and 2 were approved by the Ethics Committee of the Medical Association of Westphalia-Lippe (approval numbers 2020-878-f-S and 2022-528-f-S, respectively). Additionally, study 2 was approved by the Ethics Committee II of the University of Heidelberg (approval number 2022-643) and the Ethics Commission of the Faculty of Medicine at the University of Cologne (approval number 23-1026_1). Study 3 was approved by the Ethics Committee of Bielefeld University (approval number EUB-2023-050). In all 3 studies, participants received detailed study information sheets outlining the study procedures, research aims, and data protection guidelines. Participation was voluntary and confidential. Psychotherapists signed informed consent before their participation. To ensure data security, all participant data were collected in a pseudonymized manner. Personal data (eg, contact details or audiotapes) were stored separately from the pseudonymized data. Audiotapes of the qualitative interviews were transcribed. Once data analysis is complete, the audiotapes will be destroyed, and the data will be anonymized. Participants did not receive any compensation. This study followed the COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist for study conceptualization, data analysis, and reporting.

Participants

A total of 40 mental health professionals participated in the 3 studies. Sample characteristics varied based on the treatment context. In study 1, participants included 18 employees from a psychiatric ward for adolescents at a German clinic specializing in child and adolescent psychiatry and psychotherapy. In study 2, 13 employees from 3 different psychiatric outpatient clinics in Germany took part. In study 3, 9 psychotherapists in private practice who treated adolescents aged 12-18 years were included. Detailed sample characteristics are presented in Table 2.

Table 2. Overview of the sample characteristics presented separately for each study context.

Sample characteristics	Study 1: Inpatient treatment (n=18 participants)	Study 2: Psychiatric outpatient clinics (n=13 participants)	Study 3: Psychotherapy in private practices (n=9 participants)
Included occupational groups	• Psychotherapists (n=3) • Specialist therapists (eg, music therapists, n=3) • Nursing professionals (n=11) • Youth volunteer (n=1)	• Psychotherapists (n=9) • Psychologist without further training (n=1) • Assistant doctors (n=3)	• Psychotherapists (n=9)
Age (years), mean (SD); range	• 32.50 (9.59); 22-62	• 35.15 (9.05); 27-61	• 54.44 (13.79); 33-72
Work experience (years), mean (SD); range	• 8.33 (9.46); 0.60-40.50	• 6.08 (5.12); 2.33-16.42	• 23.44 (14.34); 6-40

Measures

Survey Methodology and Questionnaire Design

In all studies, surveys were conducted using self-administered questionnaires, which included both quantitative and qualitative questions. Participants provided information on their age, occupation, years of work experience, media affinity, attitudes toward psychotherapeutic apps, perceived app quality and feasibility, and the implementation process of Steps. Most questions were purpose-built. Because of variations in study contexts and insights gained from previous studies, some questions and questionnaires differed between the studies. Detailed differences in the surveys are presented in Multimedia Appendix 1. In the inpatient treatment context (study 1), psychotherapists received different questionnaires than other employees to account for the distinct responsibilities and working areas of each occupational group.

Media Affinity and Attitudes Toward Psychotherapeutic Apps

In study 1, media affinity was assessed using the Affinity for Technology (TA-EG) questionnaire [50]. The TA-EG consists of 19 items across 4 subscales: enthusiasm for technology, competence in using technology, perceived positive consequences of technology, and perceived negative consequences of technology. Participants rated each item on a 5-point Likert scale from 1 (does not apply at all) to 5 (fully applies). Subscale scores were calculated as the mean of the associated items, ranging from 1 to 5 points. The internal

XSL•FO
RenderX

consistency values for the TA-EG questionnaire are presented in Table 3. Based on the experiences from study 1 and the differing working conditions in outpatient therapy settings, media affinity was assessed differently in studies 2 and 3. Participants answered self-created questions about the technical equipment available at their workplace, their use of media and smartphone apps, and the integration of media into their psychotherapies. Additionally, in study 3, participants completed the Affinity for Technology Interaction (ATI) questionnaire [51], a unidimensional measurement instrument consisting of 9 items. Participants rated each item on a 5-point Likert scale from 1 (not true at all) to 5 (completely true). The overall ATI score was calculated as the mean value of all items, ranging from 1 to 5. The internal consistency of the ATI is presented in Table 3.

Table 3. Means, SDs, ranges, and reliability coefficients of the questionnaires on media affinity, attitudes toward e-mental health, and app quality from psychotherapists' perspectives. The data are presented separately for each study context.

Questionnaire	n	Mean (SD)	Range	Actual scale range	Internal consistencies (Cronbach α)
Study 1					
TA-EG[a]					
Enthusiasm	18	2.93 (0.77)	1.60-4.40	1-5	0.90
Competencies	18	3.42 (0.62)	2.50-4.75	1-5	0.63
Positive consequences	18	4.01 (0.37)	3.40-4.80	1-5	0.57
Negative consequences	18	3.02 (0.63)	1.80-4.20	1-5	0.83
Study 2					
uMARS[b]					
Overall	10	3.75 (0.60)	2.77-4.52	1-5	0.87-0.90[c]
Engagement	11	3.76 (0.69)	2.70-4.80	1-5	0.76-0.82[c]
Functionality	10	3.77 (0.85)	2.38-5.00	1-5	0.70-0.88[c]
Aesthetics	10	3.76 (0.59)	3.00-4.58	1-5	0.37-0.47[c]
Study 3					
ATI[d]	9	2.98 (0.99)	1.56-4.67	1-5	0.92
MTPS[e]					
Potential to augment psychotherapy	9	3.56 (0.46)	3.00-4.25	1-5	0.42
Perceived risks	9	2.97 (0.85)	2.00-4.75	1-5	0.78
uMARS					
Overall	5	3.52 (0.56)	2.55-3.87	1-5	0.66-0.82[c]
Engagement	9	3.79 (0.66)	3.00-4.86	1-5	0.50-0.82[c]
Functionality	5	3.54 (0.90)	2.14-4.45	1-5	0.57-0.60[c]
Aesthetics	9	3.58 (0.86)	2.30-5.00	1-5	0.58-0.76[c]

[a]TA-EG: Affinity for Technology [50].

[b]uMARS: user version of the Mobile Application Rating Scale [52].

[c]Internal consistencies of the app *Steps* (summarized for the app version and the web version of the app).

[d]ATI: Affinity for Technology Interaction [51].

[e]MTPS: Psychotherapists' Attitudes Toward Using Modern Technologies in Psychotherapy and Counselling Scale [53].

Given the specific study context of inpatient treatment (study 1), attitudes toward psychotherapeutic apps were assessed using 12 self-created questions based on the Unified Theory of Acceptance and Usability of Technology (UTAUT) [54] and an adapted UTAUT questionnaire by Hennemann et al [14]. Participants responded to the items on a 5-point Likert scale, ranging from 1 (strongly disagree) to 5 (strongly agree). In study 2, the self-created questions were adapted to the outpatient study context. As a result, attitudes toward psychotherapeutic apps were measured using a 10-item questionnaire. In study 3, attitudes toward psychotherapeutic apps were measured using the Psychotherapists' Attitudes Toward Using Modern Technologies in Psychotherapy and Counselling Scale (MTPS) [52]. The MTPS is a multidimensional self-report questionnaire

XSL•FO
RenderX

with 16 items across 4 subscales: *potential to augment psychotherapy*, *psychoeducational value*, *perceived risks*, and *perceived relevance*. We included 2 subscales—*potential to augment psychotherapy* (eg, "Modern technologies can speed up the therapeutic process") and *perceived risks* (eg, "A stricter regulation of the content available via modern technologies in relation to psychotherapy and psychopathology should be enforced"). Both scales are considered central factors in practitioners' attitudes toward technology [52,55]. Psychotherapists rated their agreement with the statements on a 5-point Likert scale, ranging from 1 (strongly disagree) to 5 (strongly agree). Subscale scores were calculated as the mean values of the 4 associated items in each subscale, with scores ranging from 1 to 5. For this study, the MTPS was translated into German. The quality of the translation was ensured through multiple forward and backward translations. The internal consistencies of the subscales are presented in Table 3.

Perceived App Quality and Feasibility

As study 1 was the first evaluation study on Steps, perceived app quality and feasibility were assessed using purpose-built open questions, informed by previous evaluation studies [14,56-58]. The questions focused on app functionality, design, available features, and aspects of the implementation process (eg, "Steps makes my therapeutic work easier" or "The use of Steps fits well into everyday life on the ward"). To obtain a more distinct and structured assessment of perceived app quality and feasibility, 4 subscales of the user version of the Mobile Application Rating Scale (uMARS) [53] were used in studies 2 and 3: engagement, functionality, aesthetics, and subjective app quality. The uMARS is a self-report questionnaire that was translated into German for this study, with accuracy ensured through blind back-translation. Participants rated 17 items on a 5-point Likert scale (1=totally disagree and 5=fully agree). They evaluated both the patient version of the app and the web-based psychotherapist version of Steps. Subscale scores for *engagement*, *functionality*, and *aesthetics* were calculated as the mean value of the associated items for each subscale, ranging from 1 to 5. Five additional questions addressed *subjective app quality* (eg, "What is your overall [star] rating of the app?"). Overall app quality scores for both the app version and the web version were determined by the mean value of all items. The internal consistencies are presented in Table 3. In studies 2 and 3, the feasibility of Steps was assessed using purpose-built questions based on those used in study 1 but adapted to the contexts of psychiatric outpatient clinics and private practice psychotherapists.

Data Analysis

Qualitative interviews from studies 1 and 3 were transcribed. All qualitative data, including interview transcripts and free-text responses from the surveys, were analyzed using deductive qualitative content analysis [59]. An initial set of categories was developed based on the uMARS and LAMH model before data analysis began. In the first step, 1 researcher coded the qualitative data. Following the deductive qualitative content analysis process described by Cho and Lee [59], data that did not fit into the predetermined categories were assigned to newly created categories. The final set of categories for the content analysis is presented in Multimedia Appendix 2. All qualitative data were recoded using the complete set of categories. Following Mayring [60], a second rater coded a subset of the qualitative data (2 qualitative interviews) to ensure coding clarity. Any discrepancies between the 2 raters were thoroughly discussed. The frequency of each category mentioned by participants, as well as the number of participants who reported each category, was assessed. Categories were ranked and reported based on their occurrence. For this publication, participant quotes from the study were translated into English. Descriptive quantitative data were analyzed using SPSS version 29.0.0.0 (IBM Corp.).

Results

Study Participation and Attitudes Toward MHAs

Study participation was low across all studies (Table 2). Despite extensive recruitment efforts, only a few psychotherapists chose to test the transdiagnostic, therapy-accompanying app Steps as part of the studies. Additionally, patient participation was low, and dropout rates were high. Detailed information on study participation and dropout rates is provided in Multimedia Appendix 3. In all studies, mental health professionals reported moderate media affinity and mixed attitudes toward MHAs. In studies 2 (psychiatric outpatient clinics) and 3 (psychotherapists in private practice), most psychotherapists had never recommended MHAs to patients before (study 2: 9/12 participants; study 3: 8/9 participants). Frequencies of the self-created questions from studies 2 and 3 regarding the technical equipment available at work, the use of media and smartphone apps, and the integration of media into psychotherapies are provided in Multimedia Appendix 3. Descriptive statistics for the standardized questionnaires used in the studies are presented in Table 3.

Perceived App Quality and Feasibility of Steps

For inpatient treatment of adolescents, about half of the participants in study 1 rated the joy of use neutrally (n=10; fun to use: n=4, not fun to use: n=2) but found Steps helpful for patients (n=9; undecided: n=5, unhelpful: n=3). However, the app was rarely used for psychotherapeutic treatment. Almost all participants agreed that the app did not fit the inpatient context. Eight participants suggested that Steps might be more beneficial for outpatient treatments. Correspondingly, 1 participant mentioned using the app for follow-up care after patient discharge from the clinic.

> *I could imagine, [Steps] perhaps as a transitional app [...] to somehow bridge the time before therapy or as aftercare after therapy. [...]* [WI2]

By contrast, for outpatient treatment of adolescents, the majority of participants expressed interest in continuing to use the app for some patients after the project ended (study 2—yes, n=9; no, n=3; and missing, n=1; and study 3—yes, n=7 and no, n=2). Additionally, most psychotherapists indicated a willingness to recommend Steps to some colleagues (study 2—few, n=1; some, n=5; many, n=3; everyone, n=3; and missing, n=1; and study 3—few, n=1; some, n=5; many, n=2; and missing, n=1). However, while 5 psychotherapists in private practice were

XSL•FO
RenderX

willing to pay for app use (no, n=4), only 1 participant from the psychiatric outpatient clinics expressed willingness to pay (no, n=10). Furthermore, Steps received an average star rating of 3.55 (SD 0.65; range 2.00-4.00; n=11) in study 2 and 3.19 (SD 0.53; range 2.50-4.00; n=8) in study 3. These ratings were comparable to those of other MHAs [21]. Descriptive statistics for the uMARS are presented in Table 3.

Essential App Features

In the inpatient setting, the most frequently mentioned feature was setting therapy goals and tasks (n=4). While psychotherapists rated this feature as unhelpful and disliked the emphasis it placed on goals and tasks in the therapy process, nursing professionals and specialist therapists found it interesting to use.

> *I liked the visualization of the therapy process and successes [through the goals and tasks feature* [W18]

The second most frequently mentioned feature in the inpatient setting was the mood check (n=3). Additionally, setting reminders was identified as an important feature (n=4), with participants expressing a desire for greater flexibility.

> *It should be possible to set a reminder as flexible as in calendar apps [the current app version provided periods that were not detailed enough]* [W18]

The results from the outpatient therapy settings presented a different perspective. Employees of psychiatric outpatient clinics identified mood checks (mentioned 8 times by 5) and the emergency kit (mentioned 5 times by 5) as the most helpful features. Additionally, the questionnaire manager was considered useful for diagnostics if templates were available (mentioned 4 times by 4). However, participants were unwilling to spend time digitizing their own questionnaires.

> *[Providing questionnaires was] far too laborious to enter them yourself.* [C14]

The availability of templates was also highlighted for other app features. Three participants from psychiatric outpatient clinics suggested a library feature to provide psychoeducational materials, tests, skills, or templates for mood checks. Additionally, psychotherapists in private practices identified mood checks (n=8), the library (n=8), and the questionnaire manager as the most helpful features of a therapy-accompanying app. Additionally, psychotherapists in private practices highlighted the option to digitize therapy materials (n=8) and facilitate communication with patients through an app as particularly helpful. However, the communication feature was also perceived as challenging. Three psychotherapists expressed concerns about legal provisions and responsibilities, particularly in cases where patients might communicate indicators of suicidality. Furthermore, a chat feature was seen as potentially leading to an “increased workload that is not practicable” (P1) for psychotherapists. Similarly, feedback on reminders was mixed. While the majority of psychotherapists found it a helpful feature (n=6), 1 psychotherapist noted that it could potentially contradict therapeutic processes:

> *I somehow also expect patients to remember appointments and perhaps manage this well via a calendar or whatever, and otherwise it’s actually a good indication of motivation for therapy for me.* [P2]

Other app features suggested by psychotherapists in private practices included a tool for tracking days without nonsuicidal self-injury, the provision of relaxation exercises, and support for organizational therapy tasks, such as sending documents or completing necessary forms and questionnaires. An overview of the key app features mentioned across the 3 studies is presented in Figure 4.

Figure 4. Essential app features for a transdiagnostic, therapy-accompanying app mentioned in the different studies. Features mentioned in study 1 are marked with "a"; features mentioned in study 2 are marked with "b"; and features mentioned in study 3 are marked with "c.".

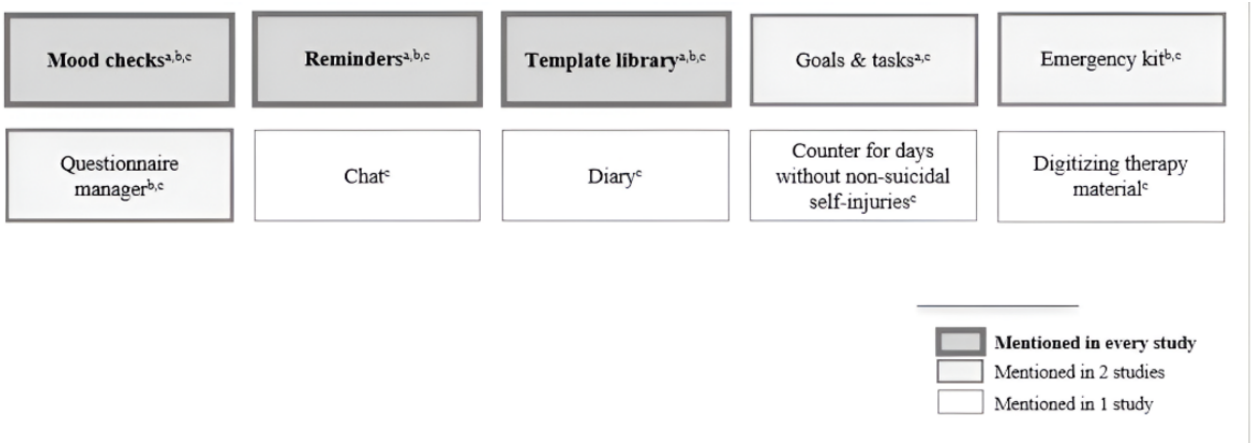

Essential App Characteristics

Across all study contexts, ease of use and customizability were identified as crucial app characteristics. Ease of use was mentioned by several participants in each study (study 1, n=4; study 2, n=6; and study 3, n=9). Half of the interviewed participants from the inpatient treatment setting emphasized the importance of customizability (study 2, n=4 and study 3, n=8). However, some participants from study 1 highlighted the benefits of preinstalled therapy tools and materials, such as activity lists, meditation and relaxation exercises, and mindfulness practices. Psychotherapists in private practices particularly valued the ability to adapt the app to different psychotherapeutic approaches and individual workflows, considering it a key strength of Steps.

XSL•FO
RenderX

> *also open to different psychotherapeutic approaches...to offer psychotherapists a tool...that...yes, incorporates modern technologies into the treatment and makes it usable for both psychotherapists and patients at the same time, so I think that's an advantage of Steps.* [P6]

However, the ability to customize the app individually was also identified as a potential barrier by psychotherapists in private practices. They noted that digitizing materials and preparing personalized app usage require additional time. In line with this concern, 1 participant from the psychiatric outpatient clinics suggested implementing an easy copy function to facilitate the creation of custom app content (C1).

In addition, accessibility was highlighted as another crucial app characteristic. The absence of an iOS version was cited as the primary reason for not using Steps by 3 practitioners from psychiatric outpatient clinics and 3 psychotherapists in private practices. However, 1 participant from the psychiatric outpatient clinics and 8 out of 9 psychotherapists in private practices positively emphasized the benefit of both psychotherapists and patients having individual yet interrelated accounts within Steps.

> *I also think it's good that there's a psychotherapist app that I can use to control things to a certain extent. I think it's good to be able to provide the patient with materials. Yes, I think so, for now.* [P2]

Regarding a therapy-accompanying app for inpatient treatment, all participants in the qualitative interviews emphasized that individual access to each professional role is essential (n=4).

> *Yes, and then I might have had more of a connection to it [with my access]. So, because I only had access via the adolescents, it was forgotten a bit quicker* [WI3]

The design was another crucial factor, with 5 participants from the inpatient therapy setting directly mentioning its importance. Additionally, 3 out of 4 interviewed participants from study 1 emphasized that design serves as a motivator for usage. In the outpatient therapy settings, design was also highlighted as significant. Two practitioners from the psychiatric outpatient clinics reported feeling less motivated to use Steps due to the unattractive design of the psychotherapist version. Feedback on the patient version of the app was mixed. On the one hand, the "[...]individual design and yet reduced to the essentials[...]" (C17) was mentioned positively. On the other hand, C18 suggested that the "layout could be more colorful" to motivate adolescents to use the app. In the sample of psychotherapists in private practices, 3 psychotherapists suggested a more playful design, such as a customizable avatar or a reward system, to make the app more appealing to adolescents. Additionally, a participant proposed incorporating motivational features into the psychotherapist version to enhance engagement.

> *Maybe there's a nice seductiveness, so to speak. That a psychotherapist who doesn't like looking at smartphones likes to look at them, [...].* [P3]

Aspects for the Integration of MHAs Into Psychotherapists' Daily Routines

In all studies, one of the most common barriers was *conditions in daily practice* (study 1: mentioned 25 times by 12 in the surveys and several times by 4; study 2: mentioned by 1; and study 3: mentioned 23 times by 8). In the inpatient therapy setting, other common barriers included *sticking to automatic behavior* (mentioned 9 times by 6) in the surveys, as well as being *unconvinced of benefits* (mentioned 14 times by 3) and *sticking to automatic behavior* (mentioned 7 times by 3) in the qualitative interviews. Regarding the barrier "sticking to automatic behavior", WI1 of the inpatient therapy setting said:

> *How do I get what I have there now into this app, where I intuitively know how to proceed [without an app]? I didn't have time for that, and the adolescents didn't even think about it.*

In line with the barriers mentioned in the inpatient therapy setting, the barrier "technical issues" was also reported by 2 practitioners in the psychiatric outpatient clinics, while the barrier "unconvinced of benefits" was mentioned 16 times by 7 psychotherapists in private practice. Additionally, psychotherapists in private practice also identified "perceived risks" as one of the most common barriers to app usage (mentioned 12 times by 6). In the psychiatric outpatient clinics, "technical issues" was likewise mentioned as one of the most common barriers by 2 practitioners.

The most common drivers varied across study contexts. In the inpatient therapy setting, the most frequently mentioned driver in both surveys and interviews was *approving of treatment quality*. Other drivers mentioned once included *satisfaction of client's needs,personal efficiency,* and *belief in possible benefits.* By contrast, participants from the psychiatric outpatient clinics identified the potential to digitize aspects of psychotherapy as a driver for app use and an indicator of its suitability for client needs.

> *[...]because paper pencil questionnaires/mood checks etc. can be replaced. Young people are much more interested in this.* [C5]

Psychotherapists in private practices cited *personal efficiency* (15 times by 7), *approving of treatment quality* (14 times by 7), and *support offers for app use* (3 times by 3) as the 3 most common drivers. Regarding the general characteristics of app usage, all participants in the qualitative interviews of study 1 (n=4) and 8 out of 9 psychotherapists in private practices highlighted the app's fit to client needs as a crucial factor for its use. Additionally, 2 psychotherapists in private practices emphasized the necessity of face-to-face contact during the relationship-building phase at the beginning of therapy as a general characteristic. Regarding the requirements for change, all interviewed participants in the inpatient therapy setting mentioned the need for external triggers to integrate an app into their daily work routines.

> *[It needs someone who] can consistently accompany such a project and also pick up [everyone of] the colleagues.* [WI1]

XSL•FO
RenderX

An overview of all mentioned drivers, barriers, general characteristics, and requirements for change in all 3 studies is presented in Figure 1.

Discussion

Principal Findings

Psychotherapists play an important role in integrating MHAs into adolescent psychotherapies [38,61]. However, previous research has indicated that apps are rarely used in psychotherapeutic treatments, even though mental health professionals are generally interested in MHAs [36-41]. In line with this, participating psychotherapists reported positive attitudes toward MHAs, but app usage rates were low across all study contexts. Furthermore, the results of the 3 studies indicated that psychotherapists had little to no experience using MHAs for psychotherapeutic treatment. To highlight the discrepancy between positive attitudes toward MHAs and their low level of use in psychotherapeutic treatment, we assessed, on the one hand, psychotherapists' needs for a therapy-accompanying MHA and, on the other hand, potential factors influencing the integration of MHAs into their daily working routines.

Examining MHAs themselves, the current studies identified 6 core features for a transdiagnostic therapy-accompanying app: mood checks, template library, reminders, goals and tasks, emergency kit, and questionnaire manager. Considering patients' perspectives from previous research, they also reported similar core features for MHAs [10,47,62,63]. Psychotherapists highlighted accessibility, design, and customizability as important app characteristics. Consistent with previous research, accessibility remains a key characteristic of MHAs for both patients and psychotherapists [45,63,64]. Psychotherapists in the current studies highlighted the interrelatability of the app Steps. They found it helpful to have complete access to adolescents' entries, as well as the opportunity to provide additional content. Moreover, accessibility for adolescents was also mentioned. As Steps was available only for Android-based smartphones, the operating system was frequently cited as a key exclusion factor. Across all 3 studies, the impact of app design on both adolescents' and psychotherapists' engagement was particularly emphasized. MHAs need to be designed to be motivating and incentivizing to ensure long-term use for therapy support [62]. Another core app characteristic mentioned in every study was the customizability of MHAs. For example, psychotherapists in study 3 highlighted that digitizing their therapy materials was an important factor in their use of Steps. They appreciated the ability to continue working with their own materials rather than being restricted to content provided by the app. Additionally, a high degree of customizability allows psychotherapists to use an app independently of their specific psychotherapeutic approaches. Most available MHAs are based on cognitive behavioral theory [18], which creates higher barriers to adoption for psychotherapists who align with other therapeutic approaches [25,55,61,65].

Limited experience with MHAs is not only a result of psychotherapists' behavior but also a significant barrier to their integration into daily routines. Previous research suggests that limited experience with e-mental health is often accompanied by a lack of competency and only moderate knowledge of the advantages and possibilities of MHAs in psychotherapy. Consequently, limited experience with MHAs can lead to increased aversion toward their use [9,24,25,38,39,45,46]. In line with this, the majority of psychotherapists in studies 2 and 3 reported that they had never used or recommended MHAs before. Moreover, the stated requirements for change—*becoming aware of benefits* and *ease of use*—were reported multiple times in every study as key factors for MHA use. Consistent with the LAMH model of Feijt et al [42], aspects of ease of use were the most frequently mentioned in the current studies. All participants emphasized that MHAs need to be easy to understand and intuitive to use. Psychotherapists specifically highlighted the need for an app that is both easy to understand and simple to use, as daily routines leave little capacity for intensive familiarization with new methods such as MHAs. In Germany, the use of MHAs for psychotherapeutic treatment is rarely reimbursed by health insurance companies [66]. That may also contribute to the low willingness to engage with MHAs. Other important influencing factors were related to perceived benefits, such as "improvement of treatment quality" as a driver and "being unconvinced of benefits" as a barrier. In accordance with previous research, psychotherapists need to be convinced of the advantages of MHA use for both patients and their own workflow [8,24,39,45,47]. Psychotherapists reported being more likely to adopt MHAs if the app facilitated their working processes, particularly the organizational tasks of psychotherapy. Additionally, fitting the app to patients' needs and maintaining face-to-face contact were also identified as general characteristics in the current studies. From the psychotherapists' perspective, the usefulness of an app depends on the clients' needs and the phase of psychotherapy. Particularly during the initial stages, when establishing a therapeutic relationship, psychotherapists would not introduce a therapy-accompanying app. They believe that building a strong relationship first is essential for the successful implementation of MHAs. This aligns with previous studies indicating that psychotherapeutic support is a crucial factor in patients' engagement with apps [8].

Beyond various personal factors influencing psychotherapists' use of MHAs, the current studies also highlighted organizational factors. One significant and common barrier was technological issues. Consistent with previous research, adequate technical infrastructure—such as a stable internet connection and proper devices—is essential for the adoption of MHAs [9,25,46]. In studies 2 and 3, fewer than half of the psychotherapists reported using a smartphone or tablet at work. Limited access to technical equipment may contribute to the perceived difficulty of using MHAs. Furthermore, in line with previous research, psychotherapists expressed concerns about data security [24,39,44,45]. Their concerns extended beyond data protection within the apps to include transparent clinical standards and guiding principles as essential requirements for MHA use in psychotherapy [24,25,39,45,48]. Although apps are increasingly being implemented in health care and supported by policy makers [66], many psychotherapists still do not feel sufficiently confident in using MHAs [9,44,62]. Additionally, in our study, psychotherapists reported uncertainties regarding legal

XSL•FO
RenderX

provisions and responsibilities, such as those related to an optional communication feature or the provision of digitized diagnostic questionnaires. Additionally, the need for external triggers was mentioned as a requirement. In line with previous research, participating psychotherapists suggested features such as a reminder function in the therapist version of MHAs to enhance engagement [24,25]. Furthermore, the perceived commitment of the entire organization and management was highlighted as an important motivational factor for successfully and sustainably integrating MHAs into daily routines.

Practical Implications and Future Research

The study results provide insights into which app features and characteristics are essential for psychotherapists in different psychotherapeutic contexts. Future app development should build on these findings and ensure that essential app features and characteristics are incorporated into MHAs. Designing MHAs based on psychotherapists' needs will facilitate their integration into daily routines. App developers must strike a balance between maximum customizability and ease of use. For example, MHAs could offer a wide selection of templates and therapy materials through a library feature, along with the option to digitize psychotherapists' own materials to expand the existing library. However, future research and app development should not focus solely on psychotherapists' perspectives and needs but should also consider the perspectives and needs of adolescents. Moreover, the integration of MHAs into daily routines is influenced by various drivers and barriers. The most significant barriers to integrating MHAs were technological issues and conditions in daily practice, such as a lack of time and resources. The most important driver was the perceived improvement in treatment quality. To successfully integrate MHAs, implementation strategies need to be adjusted. The LAMH model of Feijt et al [42] provides a useful framework for understanding common drivers, barriers, general characteristics, and requirements for change, which were supported by our study results. These identified drivers and barriers must be directly addressed in implementation strategies. Psychotherapists who have not previously worked with MHAs require close support to implement app-based services into their daily treatment routines. Additionally, organizational changes are necessary to successfully integrate MHAs into the health care system. Psychotherapists need sufficient time and resources to familiarize themselves with MHAs and learn how to incorporate them into their practice. Moreover, adequate technical equipment and stable internet connections are essential for the effective integration of MHAs.

Strengths and Limitations

The current studies have several strengths and limitations. One of the most significant strengths is the exploration of psychotherapists' perspectives while they actively tested a transdiagnostic, therapy-accompanying app in their regular treatments. This approach allowed psychotherapists to go beyond theoretical considerations of potential drivers, barriers, or necessary app features. Instead, they gained firsthand experience with MHAs, identifying which features they found helpful and which were lacking in their work. Furthermore, to the best of our knowledge, this is the first study exploring the use of MHAs in the inpatient treatment of adolescents. However, the present studies also have some limitations. The small sample sizes in all 3 studies suggest that the app Steps was not well implemented in practice, limiting the generalizability of the findings to other settings. Additionally, the small sample sizes and the recruitment method pose a risk of selection bias. For instance, psychotherapists with a greater interest in technological developments may have been more likely to participate in the studies.

Conclusions

In sum, psychotherapists in this study were generally open to therapy-accompanying MHAs. However, both study participation and app usage were low. Psychotherapists identified various drivers and barriers influencing their use of MHAs, highlighting that adoption is shaped not only by personal factors but also by organizational challenges. To enable psychotherapists to integrate MHAs into their practice, changes in current working conditions are necessary. For example, they require adequate technical equipment and sufficient time to familiarize themselves with new therapy methods such as MHAs. Successfully integrating MHAs into daily psychotherapeutic routines is a challenge for the entire health care system, not just for individual psychotherapists.

Acknowledgments

The feasibility studies were funded by the v.Bodelschwinghsche Stiftungen Bethel (study 1), the Federal Ministry of Education and Research (BMBF; study 2), and the Bielefeld University (study 3). We acknowledge support for the publication costs by the Open Access Publication Fund of Bielefeld University and the Deutsche Forschungsgemeinschaft (DFG). We want to thank our project partners Circumradius for their great cooperation in app development and Dr.-Ing Birte Richter and the medical assistance working group of Bielefeld University for their great cooperation in study 3. Furthermore, we want to thank all project teams in Bielefeld, Mannheim, and Cologne for participation and support; KJP OWL e.V. for recruitment support; and all participants for sharing their experiences with the app *Steps*.

Data Availability

The data sets generated or analyzed during this study are available from the corresponding author upon reasonable request.

XSL•FO
RenderX

Disclaimer

No generative AI tools were used in the 3 studies and the manuscript writing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
List of surveys presented separately for each study.
[PDF File (Adobe PDF File), 773 KB - formative_v9i1e65788_app1.pdf]

Multimedia Appendix 2
The final set of categories for the deductive qualitative content analysis.
[PDF File (Adobe PDF File), 38 KB - formative_v9i1e65788_app2.pdf]

Multimedia Appendix 3
A flowchart of study participation is presented for each study, along with the frequencies of self-created questions regarding the technical equipment available at work, the use of media and smartphone apps, and the integration of media in psychotherapies for studies 2 and 3.
[PDF File (Adobe PDF File), 80 KB - formative_v9i1e65788_app3.pdf]

References

1. Kessler RC, Berglund P, Demler O, Jin R, Merikangas KR, Walters EE. Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. Arch Gen Psychiatry 2005 Jun 01;62(6):593-602. [doi: 10.1001/archpsyc.62.6.593] [Medline: 15939837]
2. Roza SJ, Hofstra MB, van der Ende J, Verhulst FC. Stable prediction of mood and anxiety disorders based on behavioral and emotional problems in childhood: a 14-year follow-up during childhood, adolescence, and young adulthood. Am J Psychiatry 2003 Dec;160(12):2116-2121. [doi: 10.1176/appi.ajp.160.12.2116] [Medline: 14638580]
3. Sacco R, Camilleri N, Eberhardt J, Umla-Runge K, Newbury-Birch D. A systematic review and meta-analysis on the prevalence of mental disorders among children and adolescents in Europe. Eur Child Adolesc Psychiatry 2022 Dec 30;33(9):2877-2894. [doi: 10.1007/s00787-022-02131-2] [Medline: 36581685]
4. World Health Organization. Mental health of adolescents. World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/adolescent-mental-health [accessed 2024-06-17]
5. de Soet R, Vermeiren RRJM, Bansema CH, van Ewijk H, Nijland L, Nooteboom LA. Drop-out and ineffective treatment in youth with severe and enduring mental health problems: a systematic review. Eur Child Adolesc Psychiatry 2024 Oct;33(10):3305-3319. [doi: 10.1007/s00787-023-02182-z] [Medline: 36882638]
6. de Haan AM, Boon AE, de Jong JTVM, Hoeve M, Vermeiren RRJM. A meta-analytic review on treatment dropout in child and adolescent outpatient mental health care. Clin Psychol Rev 2013 Jul;33(5):698-711. [doi: 10.1016/j.cpr.2013.04.005] [Medline: 23742782]
7. Gearing RE, Schwalbe CS, Short KD. Adolescent adherence to psychosocial treatment: mental health clinicians' perspectives on barriers and promoters. Psychother Res 2012 May;22(3):317-326. [doi: 10.1080/10503307.2011.653996] [Medline: 22313513]
8. Diano F, Sica LS, Ponticorvo M. A systematic review of mobile apps as an adjunct to psychological interventions for emotion dysregulation. Int J Environ Res Public Health 2023 Jan 12;20(2):20. [doi: 10.3390/ijerph20021431] [Medline: 36674189]
9. Feijt M, de Kort Y, Bongers I, Bierbooms J, Westerink J, IJsselsteijn W. Mental health care goes online: practitioners' experiences of providing mental health care during the COVID-19 pandemic. Cyberpsychol Behav Soc Netw 2020 Dec;23(12):860-864. [doi: 10.1089/cyber.2020.0370] [Medline: 32815742]
10. Gindidis S, Stewart SE, Roodenburg J. Adolescent experiences of app-integrated therapy. The Educational and Developmental Psychologist 2020 Oct 19;37(1):20-29. [doi: 10.1017/edp.2019.18]
11. Rauschenberg C, Schick A, Goetzl C, Roehr S, Riedel-Heller SG, Koppe G, et al. Social isolation, mental health, and use of digital interventions in youth during the COVID-19 pandemic: a nationally representative survey. Eur Psychiatry 2021 Mar 09;64(1):e20 [FREE Full text] [doi: 10.1192/j.eurpsy.2021.17] [Medline: 33686930]
12. Ellis LA, Meulenbroeks I, Churruca K, Pomare C, Hatem S, Harrison R, et al. The application of e-mental health in response to COVID-19: scoping review and bibliometric analysis. JMIR Ment Health 2021 Dec 06;8(12):e32948 [FREE Full text] [doi: 10.2196/32948] [Medline: 34666306]

XSL•FO
RenderX

13. Torous J, Jän Myrick K, Rauseo-Ricupero N, Firth J. Digital mental health and COVID-19: using technology today to accelerate the curve on access and quality tomorrow. JMIR Ment Health 2020 Mar 26;7(3):e18848 [FREE Full text] [doi: 10.2196/18848] [Medline: 32213476]
14. Hennemann S, Beutel ME, Zwerenz R. Drivers and barriers to acceptance of web-based aftercare of patients in inpatient routine care: a cross-sectional survey. J Med Internet Res 2016 Dec 23;18(12):e337 [FREE Full text] [doi: 10.2196/jmir.6003] [Medline: 28011445]
15. Breil B, Dederichs M, Kremer L, Richter D, Angerer P, Apolinário-Hagen J. [awareness and use of digital health services in Germany: a cross-sectional study representative of the population]. Gesundheitswesen 2021 Dec;83(12):1019-1028. [doi: 10.1055/a-1335-4245] [Medline: 33862648]
16. Diano F, Ponticorvo M, Sica LS. Mental health mobile apps to empower psychotherapy: a narrative review. New York, NY: IEEE; 2022 Dec 05 Presented at: The IEEE International Conference on Metrology for Extended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE); October 26-28, 2022; Rome, Italy p. 306-311. [doi: 10.1109/MetroXRAINE54828.2022.9967663]
17. Melbye S, Kessing LV, Bardram JE, Faurholt-Jepsen M. Smartphone-based self-monitoring, treatment, and automatically generated data in children, adolescents, and young adults with psychiatric disorders: systematic review. JMIR Ment Health 2020 Oct 29;7(10):e17453 [FREE Full text] [doi: 10.2196/17453] [Medline: 33118950]
18. Wüllner S, Hermenau K, Krutkova M, Petras IK, Hecker T, Siniatchkin M. Mobile applications in adolescent psychotherapy during the COVID-19 pandemic: a systematic review. Front Public Health 2024;12:1345808 [FREE Full text] [doi: 10.3389/fpubh.2024.1345808] [Medline: 38420028]
19. Lui JHL, Marcus DK, Barry CT. Evidence-based apps? A review of mental health mobile applications in a psychotherapy context. Professional Psychology: Research and Practice 2017 Jun;48(3):199-210. [doi: 10.1037/pro0000122]
20. Gindidis S, Stewart S, Roodenburg J. A systematic scoping review of adolescent mental health treatment using mobile apps. Advances in Mental Health 2018 Sep 28;17(2):161-177. [doi: 10.1080/18387357.2018.1523680]
21. Domhardt M, Messner EM, Eder AS, Engler S, Sander LB, Baumeister H, et al. Mobile-based interventions for common mental disorders in youth: a systematic evaluation of pediatric health apps. Child Adolesc Psychiatry Ment Health 2021 Sep 13;15(1):49 [FREE Full text] [doi: 10.1186/s13034-021-00401-6] [Medline: 34517896]
22. d'Halluin A, Costa M, Morgiève M, Sebbane D. Attitudes of children, adolescents, and their parents toward digital health interventions: scoping review. J Med Internet Res 2023 May 02;25:e43102. [doi: 10.2196/43102] [Medline: 37129931]
23. Liverpool S, Mota CP, Sales CMD, Čuš A, Carletto S, Hancheva C, et al. Engaging children and young people in digital mental health interventions: systematic review of modes of delivery, facilitators, and barriers. J Med Internet Res 2020 Jun 23;22(6):e16317 [FREE Full text] [doi: 10.2196/16317] [Medline: 32442160]
24. Mayer G, Gronewold N, Alvarez S, Bruns B, Hilbel T, Schultz JH. Acceptance and expectations of medical experts, students, and patients toward electronic mental health apps: cross-sectional quantitative and qualitative survey study. JMIR Ment Health 2019 Nov 25;6(11):e14018 [FREE Full text] [doi: 10.2196/14018] [Medline: 31763990]
25. Davies F, Shepherd HL, Beatty L, Clark B, Butow P, Shaw J. Implementing web-based therapy in routine mental health care: systematic review of health professionals' perspectives. J Med Internet Res 2020 Jul 23;22(7):e17362 [FREE Full text] [doi: 10.2196/17362] [Medline: 32706713]
26. Zhou X, Edirippulige S, Bai X, Bambling M. Are online mental health interventions for youth effective? A systematic review. J Telemed Telecare 2021 Dec;27(10):638-666. [doi: 10.1177/1357633X211047285] [Medline: 34726992]
27. Ly KH, Topooco N, Cederlund H, Wallin A, Bergström J, Molander O, et al. Smartphone-supported versus full behavioural activation for depression: a randomised controlled trial. PLoS One 2015 May;10(5):e0126559 [FREE Full text] [doi: 10.1371/journal.pone.0126559] [Medline: 26010890]
28. Firth J, Torous J, Nicholas J, Carney R, Pratap A, Rosenbaum S, et al. The efficacy of smartphone-based mental health interventions for depressive symptoms: a meta-analysis of randomized controlled trials. World Psychiatry 2017 Oct;16(3):287-298 [FREE Full text] [doi: 10.1002/wps.20472] [Medline: 28941113]
29. Rauschenberg C, Schick A, Hirjak D, Seidler A, Paetzold I, Apfelbacher C, et al. Evidence synthesis of digital interventions to mitigate the negative impact of the COVID-19 pandemic on public mental health: rapid meta-review. J Med Internet Res 2021 Mar 10;23(3):e23365 [FREE Full text] [doi: 10.2196/23365] [Medline: 33606657]
30. Fischer-Grote L, Fössing V, Aigner M, Fehrmann E, Boeckle M. Effectiveness of online and remote interventions for mental health in children, adolescents, and young adults after the onset of the COVID-19 pandemic: systematic review and meta-analysis. JMIR Ment Health 2024 Mar 05;11:e46637 [FREE Full text] [doi: 10.2196/46637] [Medline: 38315524]
31. Haaf R, Vock P, Wächtershäuser N, Correll CU, Köhler S, Klein JP. [Efficacy of internet-based interventions for depression available in Germany-a systematic review and meta-analysis]. Nervenarzt 2024 Mar;95(3):206-215 [FREE Full text] [doi: 10.1007/s00115-023-01587-0] [Medline: 38260995]
32. Moshe I, Terhorst Y, Philippi P, Domhardt M, Cuijpers P, Cristea I, et al. Digital interventions for the treatment of depression: a meta-analytic review. Psychol Bull 2021 Aug;147(8):749-786. [doi: 10.1037/bul0000334] [Medline: 34898233]
33. Schröder D, Wrona KJ, Müller F, Heinemann S, Fischer F, Dockweiler C. Impact of virtual reality applications in the treatment of anxiety disorders: a systematic review and meta-analysis of randomized-controlled trials. J Behav Ther Exp Psychiatry 2023 Dec;81:101893 [FREE Full text] [doi: 10.1016/j.jbtep.2023.101893] [Medline: 37453405]

XSL•FO
RenderX

34. Hedman-Lagerlöf E, Carlbring P, Svärdman F, Riper H, Cuijpers P, Andersson G. Therapist-supported Internet-based cognitive behaviour therapy yields similar effects as face-to-face therapy for psychiatric and somatic disorders: an updated systematic review and meta-analysis. World Psychiatry 2023 Jul;22(2):305-314. [doi: 10.1002/wps.21088] [Medline: 37159350]
35. Buttazzoni A, Brar K, Minaker L. Smartphone-based interventions and internalizing disorders in youth: systematic review and meta-analysis. J Med Internet Res 2021 Jan 11;23(1):e16490. [doi: 10.2196/16490] [Medline: 33427682]
36. Kaveladze BT, Wasil AR, Bunyi JB, Ramirez V, Schueller SM. User experience, engagement, and popularity in mental health apps: secondary analysis of app analytics and expert app reviews. JMIR Hum Factors 2022 Jan 31;9(1):e30766 [FREE Full text] [doi: 10.2196/30766] [Medline: 35099398]
37. Bell IH, Thompson A, Valentine L, Adams S, Alvarez-Jimenez M, Nicholas J. Ownership, use of, and interest in digital mental health technologies among clinicians and young people across a spectrum of clinical care needs: cross-sectional survey. JMIR Ment Health 2022 May 11;9(5):e30716. [doi: 10.2196/30716] [Medline: 35544295]
38. Kerst A, Zielasek J, Gaebel W. Smartphone applications for depression: a systematic literature review and a survey of health care professionals' attitudes towards their use in clinical practice. Eur Arch Psychiatry Clin Neurosci 2020 Mar;270(2):139-152. [doi: 10.1007/s00406-018-0974-3] [Medline: 30607530]
39. Weitzel EC, Schwenke M, Schomerus G, Schönknecht P, Bleckwenn M, Mehnert-Theuerkauf A, et al. E-mental health in Germany - what is the current use and what are experiences of different types of health care providers for patients with mental illnesses? Arch Public Health 2023 Jul 17;81(1):133. [doi: 10.1186/s13690-023-01150-y] [Medline: 37461064]
40. Linardon J. Rates of attrition and engagement in randomized controlled trials of mindfulness apps: systematic review and meta-analysis. Behav Res Ther 2023 Dec;170:104421 [FREE Full text] [doi: 10.1016/j.brat.2023.104421] [Medline: 37862854]
41. Baumel A, Muench F, Edan S, Kane JM. Objective user engagement with mental health apps: systematic search and panel-based usage analysis. J Med Internet Res 2019 Sep 25;21(9):e14567 [FREE Full text] [doi: 10.2196/14567] [Medline: 31573916]
42. Feijt MA, de Kort YA, Bongers IM, IJsselsteijn WA. Perceived drivers and barriers to the adoption of emental health by psychologists: the construction of the levels of adoption of emental health model. J Med Internet Res 2018 Apr 24;20(4):e153 [FREE Full text] [doi: 10.2196/jmir.9485] [Medline: 29691215]
43. Békés V, Aafjes-van Doorn K. Psychotherapists' attitudes toward online therapy during the COVID-19 pandemic. Journal of Psychotherapy Integration 2020 Jun;30(2):238-247. [doi: 10.1037/int0000214]
44. Braun P, Drüge M, Hennemann S, Nitsch FJ, Staeck R, Apolinário-Hagen J. Acceptance of e-mental health services for different application purposes among psychotherapists in clinical training in Germany and Switzerland: secondary analysis of a cross-sectional survey. Front Digit Health 2022;4:840869. [doi: 10.3389/fdgth.2022.840869] [Medline: 35295621]
45. Bucci S, Berry N, Morris R, Berry K, Haddock G, Lewis S, et al. "They are not hard-to-reach clients. we have just got hard-to-reach services." staff views of digital health tools in specialist mental health services. Front Psychiatry 2019 May;10:344 [FREE Full text] [doi: 10.3389/fpsyt.2019.00344] [Medline: 31133906]
46. Ganapathy A, Clough BA, Casey LM. Organizational and Policy Barriers to the Use of Digital Mental Health by Mental Health Professionals. Telemed J E Health 2021 Dec;27(12):1332-1343. [doi: 10.1089/tmj.2020.0455] [Medline: 33646057]
47. Gindidis S, Stewart SE, Roodenburg J. Psychologists' motivations for integrating apps into therapy with secondary school-aged young people. J Psychol Couns Sch 2019 Oct 21;30(1):2-12. [doi: 10.1017/jgc.2019.22]
48. Mendes-Santos C, Nunes F, Weiderpass E, Santana R, Andersson G. Understanding mental health professionals' perspectives and practices regarding the implementation of digital mental health: qualitative study. JMIR Form Res 2022 Apr 12;6(4):e32558 [FREE Full text] [doi: 10.2196/32558] [Medline: 35412459]
49. LimeSurvey: an open source survey tool. Limesurvey GmbH. URL: https://www.limesurvey.org [accessed 2024-06-10]
50. Karrer K, Glaser C, Clemens C, Bruder C. Technikaffinität erfassen - der Fragebogen TA-EG. In: Lichtenstein A, Stößel C, Clemens C, editors. Der Mensch im Mittelpunkt technischer Systeme. Berlin, Germany: VDI Verlag GmbH; 2009:194-199.
51. Franke T, Attig C, Wessel D. A personal resource for technology interaction: development and validation of the Affinity for Technology Interaction (ATI) Scale. International Journal of Human–Computer Interaction 2018 Mar 30;35(6):456-467. [doi: 10.1080/10447318.2018.1456150]
52. Bagarić B, Jokić-Begić N. Measuring therapists' attitudes toward integration of technology in psychotherapy and predicting their use of technology. J Clin Psychol 2020 Jul;76(6):1151-1172. [doi: 10.1002/jclp.22967] [Medline: 32410312]
53. Stoyanov SR, Hides L, Kavanagh DJ, Wilson H. Development and validation of the user version of the Mobile Application Rating Scale (uMARS). JMIR Mhealth Uhealth 2016 Jul 10;4(2):e72 [FREE Full text] [doi: 10.2196/mhealth.5849] [Medline: 27287964]
54. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view. MIS Quarterly 2003;27(3):425-478.
55. Schröder J, Berger T, Meyer B, Lutz W, Hautzinger M, Späth C, et al. Attitudes towards internet interventions among psychotherapists and individuals with mild to moderate depression symptoms. Cogn Ther Res 2017 Apr 22;41(5):745-756. [doi: 10.1007/s10608-017-9850-0]

XSL•FO
RenderX

56. O'Dea B, Han J, Batterham PJ, Achilles MR, Calear AL, Werner-Seidler A, et al. A randomised controlled trial of a relationship-focussed mobile phone application for improving adolescents' mental health. J Child Psychol Psychiatry 2020 Aug;61(8):899-913 [FREE Full text] [doi: 10.1111/jcpp.13294] [Medline: 32683737]
57. Kenny R, Dooley B, Fitzgerald A. Feasibility of "CopeSmart": a telemental health app for adolescents. JMIR Ment Health 2015 Aug 10;2(3):e22 [FREE Full text] [doi: 10.2196/mental.4370] [Medline: 26552425]
58. Gao M, Kortum P, Oswald F. Multi-language toolkit for the System Usability Scale. International Journal of Human–Computer Interaction 2020 Aug 19;36(20):1883-1901 [FREE Full text] [doi: 10.1080/10447318.2020.1801173] [Medline: 40101104]
59. Cho J, Lee E. Reducing confusion about grounded theory and qualitative content analysis: similarities and differences. The Qualitative Report 2014 Nov 8;19(32):1-20 [FREE Full text] [doi: 10.46743/2160-3715/2014.1028]
60. Mayring P. Qualitative inhaltsanalyse. In: Handbuch Qualitative Forschung in der Psychologie: Band 2: Designs und Verfahren. Wiesbaden, Germany: Springer Fachmedien Wiesbaden; 2020:495-511.
61. Schuster R, Topooco N, Keller A, Radvogin E, Laireiter AR. Advantages and disadvantages of online and blended therapy: replication and extension of findings on psychotherapists' appraisals. Internet Interv 2020 Oct;21:100326 [FREE Full text] [doi: 10.1016/j.invent.2020.100326] [Medline: 32477885]
62. Fausett C, Lee J, Godwin E, Palmer E, Fouquet S. Informing the design of mental health smartphone apps for adolescents. Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care 2020 Sep 16;9(1):218-222. [doi: 10.1177/2327857920091012]
63. Høgsdal H, Kyrrestad H, Rye M, Kaiser S. Exploring adolescents' attitudes toward mental health apps: concurrent mixed methods study. JMIR Form Res 2024 Jan 15;8:e50222 [FREE Full text] [doi: 10.2196/50222] [Medline: 38224474]
64. Kenny R, Dooley B, Fitzgerald A. Developing mental health mobile apps: exploring adolescents' perspectives. Health Informatics J 2016 Jun;22(2):265-275 [FREE Full text] [doi: 10.1177/1460458214555041] [Medline: 25385165]
65. Topooco N, Riper H, Araya R, Berking M, Brunn M, Chevreul K, E-COMPARED consortium. Attitudes towards digital treatment for depression: a European stakeholder survey. Internet Interv 2017 Jun;8:1-9 [FREE Full text] [doi: 10.1016/j.invent.2017.01.001] [Medline: 30135823]
66. Gesetz für eine bessere versorgung durch digitalisierung und innovation. Bundesministerium für Gesundheit. URL: https://www.bundesgesundheitsministerium.de/service/gesetze-und-verordnungen/detail/digitale-versorgung-gesetz-dvg.html [accessed 2025-03-29]

Abbreviations

ATI: Affinity for Technology Interaction
COREQ: Consolidated Criteria for Reporting Qualitative Research
LAMH: Levels of Adoption of e-Mental Health
MTPS: Psychotherapists' Attitudes Toward Using Modern Technologies in Psychotherapy and Counselling Scale
TA-EG: Affinity for Technology
uMARS: user version of the Mobile Application Rating Scale
UTAUT: Unified Theory of Acceptance and Usability of Technology
WHO: World Health Organization

Edited by A Mavragani; submitted 26.08.24; peer-reviewed by A Jolliff, N Döring; comments to author 18.12.24; revised version received 21.01.25; accepted 07.02.25; published 08.04.25.

Please cite as:
Wüllner S, Hermenau K, Hecker T, Siniatchkin M
The Use of Mobile Apps in Adolescent Psychotherapy: Assessment of Psychotherapists' Perspectives
JMIR Form Res 2025;9:e65788
URL: https://formative.jmir.org/2025/1/e65788
doi:10.2196/65788
PMID:

©Sarah Wüllner, Katharin Hermenau, Tobias Hecker, Michael Siniatchkin. Originally published in JMIR Formative Research (https://formative.jmir.org), 08.04.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Developing an Internet-Based Cognitive Behavioral Therapy Intervention for Adolescents With Anxiety Disorders: Design, Usability, and Initial Evaluation of the CoolMinds Intervention

Nikita Marie Sørensen[1], MSc; Helene Skaarnes[2,3], MSc; Kim Mathiasen[1,2], MSc, PhD; Mikael Thastum[1], MSc, PhD; Johanne Jeppesen Lomholt[1], MSc, PhD

[1]Department of Psychology and Behavioral Sciences, Aarhus University, Aarhus C, Denmark

[2]Department of Clinical Research, University of Southern Denmark, Odense, Denmark

[3]Centre for Digital Psychiatry, Mental Health Services, Region of Southern Denmark, Odense, Denmark

Corresponding Author:
Nikita Marie Sørensen, MSc
Department of Psychology and Behavioral Sciences
Aarhus University
Bartholins Allé 11
Aarhus C, 8000
Denmark
Phone: 45 87159439
Email: nikita@psy.au.dk

Abstract

Background: Digital mental health interventions may help increase access to psychological treatment for adolescents with anxiety disorders. However, many clinical evaluations of digital treatments report low adherence and engagement and high dropout rates, which remain challenges when the interventions are implemented in routine care. Involving intended end users in the development process through user-centered design methods may help maximize user engagement and establish the validity of interventions for implementation.

Objective: This study aimed to describe the methods used to develop a new internet-based cognitive behavioral therapy intervention, CoolMinds, within a user-centered design framework.

Methods: The development of intervention content progressed in three iterative design phases: (1) identifying needs and design specifications, (2) designing and testing prototypes, and (3) running feasibility tests with end users. In phase 1, a total of 24 adolescents participated in a user involvement workshop exploring their preferences on graphic identity and communication styles as well as their help-seeking behavior. In phase 2, a total of 4 adolescents attended individual usability tests in which they were presented with a prototype of a psychoeducational session and asked to think aloud about their actions on the platform. In phase 3, a total of 7 families from the feasibility trial participated in a semistructured interview about their satisfaction with and initial impressions of the platform and intervention content while in treatment. Activities in all 3 phases were audio recorded, transcribed, and coded using thematic analysis and qualitative description design. The intervention was continuously revised after each phase based on the feedback.

Results: In phase 1, adolescent feedback guided the look and feel of the intervention content (ie, color scheme, animation style, and communication style). Participants generally liked content that was relatable and age appropriate and felt motivating. Animations that resembled “humans” received more votes as adolescents could better “identify” themselves with them. Communication should preferably be “supportive” and feel “like a friend” talking to them. Statements including praise—such as “You’re well on your way. How are you today?”—received the most votes (12 votes), whereas directive statements such as “Tell us how your day has been?” and “How is practicing your steps going?” received the least votes (2 and 0 votes, respectively). In phase 2, adolescents perceived the platform as intuitive and easy to navigate and the session content as easy to understand but lengthy. In phase 3, families were generally satisfied with the intervention content, emphasizing the helpfulness of graphic material to understand therapeutic content. Their feedback helped identify areas for further improvement, such as editing down the material and including more in-session breaks.

Conclusions: Using user involvement practices in the development of interventions helps ensure continued alignment of the intervention with end-user needs and may help establish the validity of the intervention for implementation in routine care practice.

XSL•FO
RenderX

(JMIR Form Res 2025;9:e66966) doi:10.2196/66966

KEYWORDS

user-centered design; digital treatment; digital mental health; internet-based; cognitive behavioral therapy; anxiety; adolescents

Introduction

Background

Digital mental health interventions (DMHIs) may help increase access to psychological treatment as provision via the internet is presumed to counter some of the key treatment barriers, such as social stigma [1,2], preference for self-reliance [1,2], treatment costs [3], and waiting times [2,3]. It is also well documented that internet-based cognitive behavioral therapy (ICBT) with support from a therapist (eg, via telephone or email) is effective for a variety of mental health problems, including anxiety in children and adolescents [4-7].

However, many evaluations of ICBT programs for children and adolescents with anxiety report high rates of dropout and that participants often fail to complete assessments or treatment sessions [8-12]. In addition, when transitioning to routine care settings, these challenges with adherence and engagement remain [13].

It is still unclear whether participants drop out of treatment due to symptom improvement, symptom deterioration, or other reasons outside therapy. However, working experts in the field of ICBT for children and young people propose that the use of user-centered design methods may help maximize user engagement [14,15]. This methodology emphasizes the importance of (1) understanding adolescents as experts on their own preferences and (2) enabling adolescents to hold central positions as experts in all stages of the design process. As it was eloquently stated in a qualitative study of the Australian Momentum platform, this type of design process may be characterized by "a *designing with* as opposed to a *designing for* mindset" [16].

User-centered design processes typically involve three phases of development: (1) identifying needs and design specifications (with end users), (2) designing prototypes of the intervention and testing their usability, and (3) running feasibility tests with end users [17,18]. The design process must be understood as iterative, and each phase can be revisited if needed as the data collection and analyses progress.

To the best our knowledge, only a few studies describe the inclusion of end users in the design of ICBT interventions for children and adolescents with anxiety [16,19,20]. Hill et al [20] and Ludlow et al [16] both used co-design practices involving service users (parents and children) and service providers (clinicians from routine care) in the development of their ICBT anxiety interventions. In an initial evaluation of the clinical effectiveness, Hill et al [21] found that the dropout rate for their intervention was lower (13%) than those reported in other studies conducted in routine care (32% in the study by Moor et al [22], 21% in the study by Waite et al [13], and 17% in the study by Vigerland et al [23]). This indicates that using user-centered design practices may result in better adherence and engagement with the intervention when delivered in routine care settings.

The importance of including end users in the design of DMHIs has been increasingly acknowledged as imperative not only in developing DMHIs but also in the processes of implementing the interventions [14]. The project that this paper describes was initiated by the Region of Southern Denmark, a government organization that is responsible for the provision of health care services for the residents of the southern municipalities in Denmark. This was done with an intention of broadening the service provision of a DMHI to routine care settings at a national level. Transitioning from research to routine care settings can be tricky, and evidence from clinical trials alone does not guarantee the use of an intervention in routine care. Using user-centered design practices in the development of this intervention may facilitate a smoother transition to large-scale service provision.

Objectives

The objective of this study was to describe the methods used to develop a new ICBT intervention for adolescents aged 12 to 17 years with clinical anxiety disorders, CoolMinds, within a user-centered design framework.

Our hope is that the methods and results described in this paper will provide inspiration for others on how to include end users in the design processes of DMHIs. In this study, we specifically collected data focusing on how to enhance engagement with the DMHI guided by the following objectives:

1. What are adolescents' preferences on graphic identity and communication styles within a digital intervention?
2. How do adolescents prefer to seek help and be supported by their parents in this?
3. How do adolescents navigate a prototype of the digital platform and content?
4. What are the families' initial impressions of the platform and content while in treatment?

Methods

Setting

Intervention development was initiated in January 2022 as a collaborative project involving the Centre for the Psychological Treatment of Children and Adolescents and the Centre for Digital Psychiatry. The Centre for the Psychological Treatment of Children and Adolescents is a research center nested within the Department of Psychology and Behavioural Sciences at Aarhus University and is specialized in studying and delivering anxiety interventions to children and adolescents. The Centre for Digital Psychiatry is a research, development, and treatment facility specialized in the use of digital technologies for health promotion and psychiatric treatments and is located within the Mental Health Services in the Region of Southern Denmark.

XSL•FO
RenderX

The project includes intervention development, a feasibility trial, and a randomized controlled trial (RCT). In this paper, only design processes and results from intervention development and the evaluation of the first intervention prototype in the feasibility trial are presented.

Ethical Considerations

The feasibility trial and RCT were ethically approved by the Medical Research Ethics Committees in Denmark (ID: VMK 2211954), and the protocol for these studies was preregistered in ClinicalTrials.gov (NCT06076964) and published in *Trials* [24]. The workshops and usability tests were conducted as development projects, which do not require ethics approval. The reason for this is that these activities did not include any experimental manipulations, were not targeted at a vulnerable population, and the potential risks associated with participation in these activities were evaluated as very low. Participants were not compensated for taking part in any of the activities or studies.

For the workshops and usability tests, written informed consent was obtained regarding participation, audio recordings, and publication of results. For adolescents aged <15 years, consent was obtained from parents. For adolescents aged ≥15 years, consent was obtained from the adolescents themselves. However, parents were informed of the adolescents' participation and were urged to discuss this with their children before participation.

For the feasibility trial and the RCT, written and verbal informed consent were obtained from all parents regarding participation and publication. In addition, verbal informed consent was obtained for adolescents aged <15 years, and both verbal and written informed consent were obtained for adolescents aged ≥15 years.

An additional written consent form was used to obtain consent regarding participation in, recording of, and publication of information from feasibility interviews. For adolescents aged <15 years, written informed consent was obtained from parents. For adolescents aged ≥15 years, written informed consent was obtained from both the adolescents and parents.

The quotes reported in this paper were translated from Danish into English. Quotes containing sensitive or identifying information were excluded to protect participants' anonymity.

Procedure

The CoolMinds ICBT intervention was developed and continuously updated from January 2022, when the initial outlining of the project began, until June 2023, when the feasibility trial concluded. The development of the intervention followed the recommendations from Hill et al [14] and progressed in three iterations following the design phases: (1) identifying needs and design specifications, (2) designing and testing prototypes, and (3) running feasibility tests. Reporting of the study followed the Guidance for Reporting Involvement of Patients and the Public, and a Guidance for Reporting Involvement of Patients and the Public long-form checklist is available in Multimedia Appendix 1 [25].

Working Groups

Development processes were led by a working group consisting of junior researchers (authors NMS and HS); a psychologist with experience in producing written materials for adolescents (see the Acknowledgments section); and members of the development and implementation team, such as graphic designers, user involvement experts, and engineers (see the Acknowledgments section) and were supervised by senior researchers (authors JJL, KM, and MT).

During the development process, weekly meetings were held with all available team members. The meetings always commenced with a brief status from all involved teams. The purpose of these meetings was to facilitate the decision-making and implementation processes by assembling relevant team members [26]. This ensured a flow of information between teams when decision-making required the expertise of multiple disciplines (eg, the psychological content had to be informed by data from the user involvement practices and be aligned with the technical possibilities within the platform). When consensus on each topic was reached by all teams, smaller working groups were formed with relevant team members who implemented the decision-based outcomes.

Furthermore, a total of 3 team workshops were held with members representing each team in the working group. The workshops were held at different time points in the development process reflecting the writing of the intervention manuscript. The purpose of the workshops was to (1) map out the overall framework of the intervention content and (2) become aware of any missing information that would need to be researched or collected using user involvement practices. The topics of these workshops were (1) outlining the intervention content, (2) rethinking how to conduct exposure therapy in a digital format, and (3) parental involvement.

Outlining the Program

The outline of the program was drafted by the working group with expert supervision. This included decisions on therapeutic approach, key treatment components, and parental involvement. It was decided to develop a cognitive behavioral therapy (CBT)–based program as this is the most well-researched approach and has shown good results in previous evaluations using the internet-based format [27-31]. Decisions on which treatment components to include were based on either evidence from research and literature or clinician experience. Several reviews and meta-analyses suggest that key CBT components such as exposure therapy (in vivo and imaginal) and cognitive restructuring are some of the main components driving the treatment effects [27,32,33]. Evidence on the efficacy of coping strategies is scarce compared to that for the key components [34]. However, the group of clinicians involved in providing feedback on the program manuscript advocated for the inclusion of coping strategies as these, in their experience, were techniques that adolescents usually benefitted from and were easier for them to implement on a day-to-day basis. This is also in line with findings from qualitative evaluations of ICBT treatments, where adolescents preferred easy-to-use techniques [35,36]. It was also decided to create a separate parent program and offer the treatment as a family-based intervention as social support

XSL•FO
RenderX

from parents may increase treatment adherence and response when treatment is delivered in a digital format [37]. However, the 2 programs were not connected on the platform, and thus, adolescents and parents were allowed to progress at their own pace within each of their programs.

In both programs, sessions were structured in a predefined order starting with psychoeducation followed by cognitive restructuring techniques and, finally, exposure therapy techniques. Sessions were opened automatically upon completion of the previous session unless the adolescent was to receive feedback from a therapist. In that case, the session was first opened upon receiving feedback.

Writing of the program manuscript was guided by experiences from previous and linked research activities, current literature in the field, and expertise of senior researchers and experienced clinical psychologists within and external to the project who continuously provided feedback on the manuscript. The literature and practitioner expertise were revisited during the development process to qualify and provide context to the incorporation of insights from the user involvement workshop and usability tests conducted using prototypes of the intervention. This iterative design process is illustrated in Figure 1.

Figure 1. Model of the iterative design process used across phases of development.

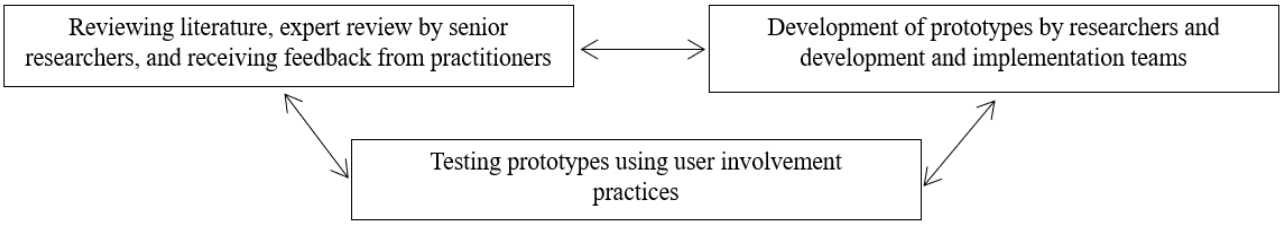

Phase 1: Identifying Design Specifications in a User Involvement Workshop

A total of 24 adolescents in grade 9 (aged >15 years) from a Danish public school participated in a user involvement workshop held in September 2022. The purpose of the user involvement workshop was to identify adolescents' preferences on graphic identity and communication styles within the intervention content (objective 1) and investigate adolescent help-seeking behavior (objective 2).

Activities included (1) voting for their favorite prototypes (out of a selection of 3-5 prototypes) using stickers, (2) discussing the pros and cons of each prototype in smaller groups and plenary, and (3) discussing how to reach out for help. Prototypes included different color schemes, styles of animation, expert videos, and types of feedback from the app (push notifications, mood ratings, and in-app rewards and praise). For polls including many prototypes (ie, color schemes, animation style, and notifications), adolescents had more than 1 vote. However, they were also allowed not to use all their votes. The topic of reaching out for help and receiving support was discussed in a column exercise in which adolescents were to give advice to another adolescent in mock-up scenarios about (1) how to handle parents being overinvolved, (2) how to involve parents, and (3) how to handle not having parental support and where to go for help then. Their advice was written down by an assistant in each group. Outcomes of the workshop were documented by audio recording group discussions and through photographs of votes and plenary summaries (both done on a whiteboard).

Members of the development team contacted 4 public schools via phone and, if they were interested, provided them with written information about the content and purpose of the workshop. One school agreed to participate, and the time of the workshop was planned in collaboration with the grade 9 teacher. The written information was then shared with parents via the school's intranet. Informed written consent was obtained from all adolescents at the beginning of the workshop using a printed consent form.

Phase 2: Testing Intervention Prototypes in Usability Tests

In total, 4 adolescents (all female) participated in individual usability tests during November 2022 and December 2022. A total of 2 tests were conducted in person with 50% (2/4) of the adolescents, who were aged 12 years, and 2 tests were conducted online with 50% (2/4) of the adolescents, who were aged 16 years. The purpose of the usability tests was to understand how adolescents would navigate the digital platform and perceive the intervention content within this platform (objective 3). Participants were instructed by an assistant to go through a psychoeducational session of the program and say out loud what they thought they were to do in the program and what they then did. They were also prompted to point out anything that confused them and anything that they were particularly fond of in the program. The assistant was present during the entire usability test and took extensive notes used for briefly interviewing the participants about their experience with the program content (eg, animation style and intelligibility of the text) and interface (eg, intuitiveness of navigating the platform), along with some standardized questions (Multimedia Appendix 2). The usability tests were also audio recorded.

Participants were recruited through word of mouth by members of the research group. Informed written consent was obtained from caregivers to children aged <15 years using a printed consent form. Informed written consent was obtained from adolescents aged >15 years using a printable digital consent form sent and received using secure emails (ie, encrypted and digitally signed emails).

Phase 3: Feasibility Tests With End Users

A total of 7 families (adolescents and their parents) from the feasibility trial participated in a 2-part online interview regarding (1) their satisfaction with trial procedures and (2) their

XSL•FO
RenderX

experiences using the program and with specific content components. We developed a semistructured interview guide with open-ended questions regarding general topics related to the trial and treatment procedures and questions regarding specific components of the program content (Multimedia Appendix 2). In this paper, only results regarding the program content are elaborated on. Results on satisfaction with trial procedures will be published elsewhere.

The purpose of these interviews was to explore the families' initial impressions of the platform and content while still in treatment (objective 4). The interviews were conducted with each family individually but at different time points during their treatment period. The families were picked at random during March 2023 and April 2023 depending on their progress (ie, whether they were at the beginning of, midway through, or at the end of their treatment period). This was done to ensure the collection of information about session-specific content while it was still novel for the participants. The interviews were conducted online and recorded.

For the participating adolescents (n=7), the mean age was 13.29 (SD 1.38; range 12-16) years. A total of 57% (4/7) were male, and 43% (3/7) were female. In total, 43% (6/14) of the parents had a high school or vocational education, 50% (7/14) had completed medium-cycle higher education, and 7% (1/14) had completed long-cycle higher education. All adolescents and their parents were born in Denmark.

Informed written consent was received from all participants aged >15 years and their caregivers before participation using an online consent form distributed using REDCap (Research Electronic Data Capture; Vanderbilt University), a secure, web-based software platform designed to support data capture for research studies [38,39] hosted at the Open Patient Data Exploratory Network within the Region of Southern Denmark.

Data Analysis

Phase 1: Identifying Design Specifications in a User Involvement Workshop

Votes were manually counted and ranked from the highest to lowest number of votes. The audio files from group discussions on pros and cons were transcribed, and the main arguments for and against each prototype were listed by prototype.

Audio files on help-seeking behavior were transcribed and analyzed using a qualitative description design by author NMS [40]. Data were analyzed using content analysis, and codes were sorted into themes.

Phase 2: Testing Intervention Prototypes in Usability Tests

Notes from the usability test were analyzed using a qualitative description design [40]. Data were coded into predefined categories reflecting each feature assessed in the usability test (eg, text, video, task, and animation).

Audio files were only used to expand the notes if anything was unclear.

Phase 3: Feasibility Tests With End Users

Due to technical issues, recordings from only 71% (5/7) of the interviews were available. These interviews were transcribed and analyzed using a qualitative description design by author NMS [40]. Data were coded using content analysis based on features (eg, text, video, animation, and audio) and preferences (ie, likes or dislikes). Data were then sorted to identify similar patterns, commonalities, and differences. The codes were then grouped into two main themes representing the general feedback from the participants on (1) satisfaction and acceptability and (2) parental involvement, as predefined in the interview guide.

Results

Phase 1: Identifying Design Specifications in a User Involvement Workshop

Overview

Results from the user involvement workshop are reported in Table 1.

XSL•FO
RenderX

Table 1. Overview of adolescent votes and comments from the user involvement workshop conducted at a Danish school, divided into pros and cons. Note that some statements are summaries of participants' comments and some are direct quotes (indicated by quotation marks).

	Number of votes	Pros	Cons
Color scheme (only the ones with the most votes are presented)			
Color scheme 7—pink, blue, green, yellow, and orange	7	—[a]	—
Color scheme 5—reds and grays	6	—	—
Animation style (only the ones with the most votes are presented)			
Animation 4—the pursuit of happiness	12	• Felt relatable as it represented "real" people. • "Felt like I could identify myself with the animations better because it used 'real' people."	—
Animation 1—letting go of stress	10	• It worked well because it did not contain that many things. • It felt very calming.	• The text could be presented a bit slower. • It was boring. • It was too fast.
Animation 2—changing perspective	10	• "I liked that everything was quite fast because it shows how chaotic your thoughts can be."	—
Expert videos			
Expert 3—female actress	8	• She seemed to be more like a psychologist. • It was nice that she also used her body while speaking. • "I liked the calm background and her voice." • It was nice and simple without any disruptive elements. • She seemed trustworthy and professional.	—
Expert 1—male professor	7	• Seemed smart. • "His way of talking appealed the most to me, but the books behind him didn't." • "He seems intelligent. The books only added to his professionalism."	• "A bit 'heavy.' He seems like a know-it-all." • "It is way too boring, you don't actually listen to what he says." • "The books in the background are disruptive. I can't really focus on him because of them." • The background is 'too noisy.'
Expert 2—2 young people	6	• It was nice that young people presented the information as it made it feel more relatable. • "It felt more age appropriate to have young people speak to me, but it also depends on what they're talking about."	• It would be nicer to only have one person talking. • It was confusing that they alternated talking. • They were a bit stiff.
Push notifications			
Notification 5—"You're well on your way. How are you today?"	12	• "It feels motivating to receive praise and a reminder in one." • "It is nice that it has both a motivating text and a question." • "It is nice that something positive is included." • "It feels like a pat on the back."	—
Notification 1—"It has been a while since we've heard from you. How are you?"	10	• "It feels like a friend is asking." • "It is good."	• "Can be understood both positively and negatively."

XSL•FO
RenderX

	Number of votes	Pros	Cons
Notification 2—“We know, you’re working hard. Have you rewarded yourself?”	9	• “The thought of taking care of yourself is nice.” • “I like the idea of rewarding myself.”	• “Can maybe be a bit demanding.”
Notification 3—“How has your day been?”	4	• “Covers a lot.”	• “Is hard to answer as it is too broad.”
Notification 6—“Tell us how your day has been?”	2	• “Is a simple, open-ended question.”	• “Seems like an order.” • “It sounds a bit negative.”
Notification 4—“How is practicing your steps going?”	0	• More specific	• Not very deep • “Seems like a yes/no question.” • “Can seem like a reprimand.”
In-program feedback—ratings			
Rating 1—smiley faces	13	• “It is easier to understand faces as opposed to e.g., colors.” • The intention and meaning were clear. • “The variety of options is a plus.” • “It doesn’t sugar coat anything, which is nice.”	• “Maybe faces can be too specific.”
Rating 2—weather	7	• “It does make good sense to compare your mood to the weather.” • “The weather gives you more room for interpretation and ascribing emotions.”	• “It is a bit difficult to understand.”
Rating 3—waves	1	• The waves did leave more room for interpretation.	• “Needs more options.” • “It is too unspecific. It can’t relate to it at all.” • “I don’t understand it.”
In-program feedback—completion			
Feedback 1—checkmark, praise, and quote	9	• It was simple and serious. • “I like the combination.” • “It is good and easy to understand.” • “I really like the use of quotes.” • “It seems the most professional and the quote is like a reward.”	—
Feedback 2—animated character receiving a trophy and confetti on a podium	8	• “The progress-bar is nice as you can track your progress and see how much is left of the program.” • “It is nice that it has no gender.” • “It is cute.”	• “It is a bit too over the top.” • “It needs to be used consequently otherwise you’d think you weren’t doing good enough.” • “It is a bit too ‘cute.’”
Feedback 3—animated character waving as it walks along a progress bar	3	• “It is nice to have some movement in the animation.” • “It is funny.” • “It is good to know how far you’ve come.”	• “It seems too childish.”

[a]Not applicable; the adolescents did not have any comments on the specific prototype listed.

Color Schemes

Color scheme 7 received the most votes (7 votes), closely followed by color scheme 5 (6 votes).

Animation Style

Animation 4 received the most votes (12 votes), closely followed by animations 1 and 2 (tied; 10 votes each). The adolescents described animation 4 as the most “relatable” one. They thought that it “worked well” because it included “humans” and that

XSL•FO
RenderX

they were able to "identify better" with the animation because of this.

Expert Videos

Votes were almost tied between the expert videos. Expert 3 received the most votes (8 votes), followed by expert 1 (7 votes) and expert 2 (6 votes). The adolescents described how expert 3 seemed more like a psychologist in the way she looked and talked. They also preferred the background and simple setup of expert 3 (ie, plain background) compared to that of expert 1 (ie, shelves with books) as it was perceived as more "calm" and not "disrupting." However, some adolescents felt that the look and background of expert 1 added to the credibility of the expert and the information provided in the video.

Push Notifications

Notification 5 received the most votes (12 votes). The adolescents emphasized the motivating and acknowledging aspect of receiving praise and that it felt nice to combine praise and a question in contrast to just a question (ie, "You're well on your way. How are you today?"). The adolescents also described how the tone of the text felt more like "a friend" and "supportive." Notification 1 also received a large number of votes (10 votes; "It has been a while since we've heard from you. How are you?"). It was also described as feeling "like a friend." However, most adolescents also pointed to the potential "double meaning" as they thought that it could be "interpreted both as positive and negative." The same was the case for notification 2 (9 votes), which was described as both "pleasant" because of "the thought of caring for yourself" and "commanding."

Notifications 4 and 6 received the least votes and were described as "negative" and feeling like "being told what do to" (eg, "Tell us how your day has been?"). Notification 3 was perceived as too hard to answer because it was too "broad." However, some adolescents did like that it was not too specific and would "cover more."

Ratings

Rating 1, an array of different smiley faces, received the most votes (13 votes). The adolescents described how it was "the easiest to understand the meaning of" and that it was "an advantage that it had more options" to describe their mood more accurately. Ratings 2 (7 votes) and 3 (1 vote) were generally perceived as "hard to interpret," "difficult to understand," and too "unspecific." However, the metaphors of weather and waves could also "leave room for interpretation" and the "possibility to attribute more feelings to it" compared to the smiley faces in rating 1.

Completion

Feedback 1 and 2 received the most votes (9 and 8 votes, respectively). The adolescents preferred more "serious"-looking feedback and animations, such as simple checkmarks or written praise and quotes, and they wanted the animations to be the same "or else you could think that you've not done it well enough." The adolescents were very fond of the combination of praise and a quote as the quote then felt like "a reward" (eg, "Good job! As Walt Disney once said, the difference between winning and losing is not quitting"). Feedback 3 (3 votes) was perceived as "too immature," but the adolescents liked the visualization of how far they had come in the program.

Parental Involvement

When investigating adolescents' help-seeking behavior in the user involvement workshop, the general response on how to talk with parents about something difficult was to "just say it as it is" as "your parents want what's best for you anyway." However, all groups also mentioned seeking support from others, such as another family member (eg, grandparent), a teacher, friends, the school nurse, a physician, counseling services, or the internet, despite all scenarios being centered on parental involvement.

When looking at overarching themes across all scenarios, the themes *security* and *respect, boundaries, and mutual understanding* were identified (Multimedia Appendix 3). *Security* represents the feeling of relational safety ("they only want what's best for you") that enables the adolescent to feel comfortable speaking to someone else. *Respect, boundaries, and mutual understanding* represents the adolescents' need for autonomy ("she needs to establish boundaries for her parents") and having the defining voice in how to communicate about their difficulties ("Make agreements with your parents on what they can ask about and when").

Phase 2: Testing Intervention Prototypes in Usability Tests

The results from the usability tests can be found in Multimedia Appendix 3. Generally, the interface of the session was perceived as intuitive and easy to navigate by the adolescents. They expressed that "it is easy to understand, what I have to do" to move on in the program. However, most of the adolescents, especially the younger participants, felt that the session was "too long" and that there was "a bit too much text" and "too many questions." One of the youngest participants commented that she had "actually forgotten the video" she saw at the beginning of the session when she had to refer to it in an assignment at the end of the session.

The content was generally perceived as "understandable." Both graphics and videos were generally perceived as helpful to better understand the techniques and examples presented. However, for the youngest participants, some of the content was too complex at times. One participant explained that the animations were "good but at times confusing" if too much was happening at once. She noted that "you have to look at them for a while to understand them." Another participant commented that the graphics helped her understand the content better and that the order of presentation of content was important for her. She explained that it would be more helpful to see the graphics first and then the text-based explanation for her to better understand it (eg, seeing the feelings thermometer as a graphic and then reading about it).

Phase 3: Feasibility Tests With End Users

The results from the interviews conducted with families in treatment during the feasibility trial can be found in Multimedia Appendix 3.

XSL•FO
RenderX

Satisfaction and Acceptability

The participants were generally satisfied with the program content. Both adolescents and parents emphasized graphic elements such as animations, videos, and audio files as particularly beneficial for their understanding and acquiring of the treatment techniques. Both adolescents and parents noted how the audio files with stories from past patients with anxiety disorders were "relatable" and "helped them understand" their own or their children's anxiety. One adolescent noted that it helped him feel like he was "not the only one who has got something."

Both adolescents and parents also liked that the introductory sessions in each of their programs included videos on how to navigate the platform and functionalities within the program. One parent noted that "there were videos for everything, so it was impossible to do anything wrong." Another parent also noted that it was "helpful" that she "could always go back and watch it again if you needed to."

Parents particularly liked that they, in their program, were able to try out some of the same techniques and tasks as their children but directed at their own worries (ie, cognitive restructuring or doing an exposure task). One parent explained how this made her more capable of helping her child throughout the treatment. However, for some participants, it was still difficult to understand how to do the exposure tasks. Some of the parents pointed out that they would have liked to be presented with more disorder-specific examples of *stepladders* and *step planners*.

Almost all participants, both adolescents and parents, expressed that the amount of reading was too much. In addition, some participants found the number of questions and tasks to complete in each session overwhelming. One adolescent explained that he "needed to take breaks after doing tasks to stay motivated." One of the parents also said that the breaks within the sessions "felt like an acknowledgement and like it was okay to not complete everything in one go."

Parental Involvement

Similarly to the findings of the user involvement workshops, adolescents expressed how having their own treatment program separate from their parents gave them a sense of "freedom" and "confidentiality" regarding the program. Parents, on the other hand, expressed feeling a "lack of control" or being "excluded." However, some parents (of mainly older children) also noted that the lack of control or unawareness of the adolescent's progress was "maybe healthy" and "fine but took a little getting used to."

Both adolescents and parents did find it difficult to align their progress when completing separate programs. Some families noted that it was "weird" or "difficult to discuss the program content when they were not at the same session" and that it would sometimes lead to "misunderstandings." Another family noted that they sometimes had difficulties helping their adolescent with their tasks if they had not been introduced to the techniques yet.

Final Iteration

The first versions of the adolescent and parent programs were completed in December 2022 and evaluated in a feasibility trial from January 2023 to June 2023. Findings from the interviews conducted during the feasibility trial (described previously) were then used to update and expand existing material from the first version. The main updates made to the program were (1) editing down of material and converting text to graphics, (2) including in-session breaks to help adolescents stay motivated, and (3) aligning the adolescent and parent programs to be completed simultaneously. The second versions of the 2 programs are currently being evaluated in an RCT (ClinicalTrials.gov NCT06368557).

The adolescent and parent versions of the program content can be found in Multimedia Appendix 4.

Discussion

Principal Findings and Implications for Content Development

Overview

This study addresses a gap in the current research on developing DMHIs. First, this study provides examples of feasible methods of including end users in the initial stages of development processes. Second, the results of this study complement and expand the current knowledge on adolescent preferences with regard to DMHIs. This may ensure the development of relevant DMHIs and help establish the validity of these interventions for implementation in routine care settings.

Results from phase 1 guided the look and feel of the intervention content, such as the graphic identity (ie, colors and animations) and communication style. Generally, adolescents liked content that was relatable and age appropriate. Adolescents preferred communication that felt "like a friend" (ie, in a tone that was supportive). When discussing support and help-seeking behaviors, adolescents emphasized both themes of independence and the need for relational security. In addition, they pointed out how support may be provided not only by parents but also by others in their social circle. In phase 2, adolescents perceived the platform as intuitive and easy to navigate. Overall, the content was easy to understand but too lengthy. In phase 3, families were generally satisfied with the intervention content, but it was still too lengthy, and the 2 programs (ie, the adolescent and parent programs) needed to be better aligned to allow for adolescents and parents to progress at the same pace. Overall, feedback from end users helped identify areas for further improvement.

Objective 1: Preferences on Graphic Identity and Communication Style

Overview

Findings from the user involvement workshop helped clarify adolescent preferences regarding both graphics and text-based material. The prototypes with the most votes guided the development of all content types. Thus, it was important to

XSL•FO
RenderX

investigate this early in the development process while still outlining the program.

Graphic Identity

The color schemes and animation styles with the most votes were used when designing the project website and logo and in the design of graphic material for the intervention (Multimedia Appendix 5). Preferences on animation style and expert videos were also used when developing the graphical content of the intervention. Specifically, the somewhat equal number of votes regarding expert videos led to the inclusion of two types of experts in the program: (1) a psychologist explaining CBT techniques and psychoeducational material and (2) a professor explaining the research on anxiety and CBT. On the basis of adolescents' qualitative comments, we decided to keep the background simple (Multimedia Appendix 5).

Communication Style

The qualitative comments from the adolescent discussions of prototypes of notifications and feedback guided the writing style of the program. We used a sort of "friendly strictness," where the program should feel like a friend supporting but also pushing them forward. One way to do this was to formulate explanations and tasks as offers instead of directive statements while still emphasizing the importance of doing the tasks (ie, "A lot of people make the mistake of believing that their thoughts are truthful and factual. However, thoughts are just mere ideas that may or may not be true. You won't know if your thought is true unless you do a fact-check of it. So, one of the first things you should ask yourself, when you're feeling anxious is..."). Furthermore, the adolescents' comments also emphasized the positive impact of including quotes and praise. Thus, praise was included throughout the program at the conclusion of exercises and modules (ie, "Well done! Now you should have a bunch of ideas on how to reward yourself"). At module completion, the praise was also combined with quotes as they were perceived as rewarding by the adolescents (ie, "Congratulations! You've now completed Step 5. Walt Disney once said that the difference between winning and losing is most often not quitting"). We decided to include quotes as relevant in-program rewards to keep adolescents motivated to use the program.

We did not end up including push notifications in the feasibility version of the program due to technical limitations within the digital platform. Instead, they have been included in the RCT version that is currently being evaluated. Their potential impact on adherence will be investigated in that trial.

Relatedness

In the user involvement workshop, the topic of *being relatable* was evident in the qualitative comments across all preferred prototypes. Adolescents simply preferred content prototypes that were relatable in such a way that they could identify with them, such as using humanlike characters in animations, using appropriate metaphors, and using supportive and developmentally appropriate language. This is in line with previous research emphasizing the importance of including relatable and age-appropriate content [16,35,36,41] and, thus, including adolescents in the design process to clarify which types of content are seen as relevant.

This finding guided the writing of the intervention manuscript and development of graphic material. For example, we included recurring animated characters, iGuides (Multimedia Appendix 5), that guided the adolescents through the intervention content using a storytelling approach in a cartoonlike format. The purpose was to offer a more personal and age-appropriate dissemination of CBT concepts and techniques presented in the program. We also conducted interviews with past patients whose stories were then audio recorded by actors to provide the adolescents with a feeling of recognition and acknowledgment. These stories were included as part of psychoeducation (ie, examples of how it feels to have anxiety), cognitive restructuring (ie, examples of helpful thoughts), and exposure tasks (ie, examples of goals and use of the technique) and as motivational speeches (ie, how it was to receive treatment and overcome anxiety). In addition, we included "help a friend" tasks in which the adolescent had to help a friend (ie, a fictional case) using the presented CBT techniques to further the feeling of relatedness and mastery.

The theme of being "relatable" was also present in the interviews conducted with families from the feasibility trial in phase 3. In this case, families noted that the stories from past patients had helped them feel less alone ("I'm not the only one who has got something") and helped them understand their situation better. Thus, relatedness and relatability may be key themes when developing engaging internet-based interventions.

Objective 2: Help-Seeking Behavior

Insights on adolescent help-seeking behavior helped clarify how adolescents would prefer to be supported by parents and caregivers.

In the user involvement workshop, the overarching themes present in the adolescents' discussion of support and help-seeking behavior tapped into the duality of taking charge on their own (eg, by "establishing boundaries," "making agreements," and being "serious") and the need for relational security with others in this process (eg, by "just telling them" and "talking to someone you trust"). In addition, many adolescents also qualitatively emphasized how it could be helpful to speak to someone other than their parents (eg, a friend, a teacher, or another family member).

This duality may reflect how adolescents themselves initiate and engage in scaffolding efforts as learners. Scaffolding refers to the process "that enables a child or novice to solve a task or achieve a goal that would be beyond his unassisted efforts" [42]. Thus, it refers to the process of assisting a person to move through the zone of proximal development by "controlling those elements of the task that are initially beyond the learner's capability, thus permitting him to concentrate upon and complete only those elements that are within his range of competence" [42]. In this case, adolescents may themselves initiate this process by seeking out relevant "assistants" to support them in using the intervention and acquiring the skills and knowledge presented in the intervention.

In addition, there is some evidence suggesting that social support may be more important in ICBT treatments in which the level of therapist support is low compared to traditional CBT. In a

XSL•FO
RenderX

predictor analysis by Spence et al [37], social support from family, friends, or a special person significantly predicted higher levels of program adherence for adolescents completing the BRAVE-online self-help program. However, only family support predicted treatment outcome. This may underline the importance of parental support to increase treatment adherence in digital formats.

However, in a predictor analysis by Stjerneklar et al [43], the time spent by parents helping their adolescents complete treatment did not significantly predict treatment outcome. In this study, the level of therapist support was high (weekly phone calls set to a duration of approximately 20 minutes) compared to that in the study by Spence et al [37] (no support). Thus, the role of parental support may only be important in interventions with no or low levels of therapist support.

In broader terms, these themes may also reflect how adolescence as a developmental stage is a time when liberation from parents and the need for autonomy become more apparent. Turning to peers instead of parents for support may not be unusual for this age group, and thus, the program content should be inclusive of this. In addition, increasing autonomy may call for a communicative approach that emphasizes the adolescents' agency [44]. This meant that, instead of focusing solely on using parents for support during the ICBT treatment, adolescents were also prompted to create a list of their "support network" in the program's first session, identifying all people who would be able to support them and specifying how they would like to be supported by them.

In the feasibility trial conducted in phase 3 of the development process, autonomy and support were also themes that were discussed by the families. Adolescents generally enjoyed the autonomy related to having separate programs, whereas for parents, this was associated with mixed responses. Generally, parents of older adolescents were more positive toward their children working in the program on their own. This is in line with findings from a co-design study by Ludlow et al [45] with caregivers of young people aged 7 to 17 years. They described how the desired level of caregiver involvement depended on the young person's age, where involvement may be lower with older adolescents. Thus, allowing for different levels of caregiver involvement depending on the age of the adolescent could allow for more flexible and personalized use of the intervention.

Objective 3: Navigating the Platform and Content

The usability tests provided practical information on how end users would navigate the platform and content. On the basis of this, the intervention material was edited down significantly, and most text bodies were converted into graphics or short videos. On the basis of feedback from the youngest participants (aged 12 years), some of the existing graphics were also simplified and slowed down. Where possible, existing videos were edited down or split into sections to not last more than 2 to 3 minutes each. These findings were then used to guide the design of subsequent graphic elements.

Objective 4: First Impressions From Families in Treatment

Generally, families were satisfied with the program content and platform. Families emphasized the graphic elements of the content, such as animations, videos, and audio files, as particularly beneficial. On the basis of this, some text bodies were converted into graphics and videos, and other text-based elements such as checklists (eg, lists with examples of symptoms, behaviors, and rewards) were edited down. One session (on realistic thinking) was also split into 2 subsessions on identifying and working with anxious thoughts and practicing detective thinking.

However, some families did find it difficult to align their progress when completing separate programs. This led to 2 major revisions to the parent program related to enhancing the joint use of the 2 programs. First, guidelines on how to appropriately use the program were included in the introductory session. This included guidelines on how often to work with the program, how often to do exercises outside of the program, and tips on how to stay consistent with the help of parents, all with the purpose of better aligning the adolescent and parent progress when working separately. Second, the parent program was also restructured from 7 sessions to 10 sessions to better reflect the course of the adolescent program. Coping strategies presented in the adolescent program (in the "toolbox" session) were included in short in a "toolbox" session for the parents. In addition, the intended purpose and rationale of each task were elaborated on further in both the adolescent and parent programs to help families establish a mutual understanding of the program content.

Despite efforts to keep program content short, it was still too extensive, and the amount of reading was too much. In addition to converting text bodies to graphics and editing down the material, more in-session breaks were included, encouraging adolescents to take time to implement the strategies and then return to the platform.

Thus, findings from the feasibility interviews led to substantial revisions of the program, all related to enhancing its use. However, little is known about the role of including user-centered practices to enhance engagement with DMHIs. Some research conducted with adult populations indicates that using user-centered practices may help increase engagement [46,47], but more research is generally needed.

Limitations

This study has some limitations related to the representativeness of the samples and the extent of the involvement of end users in the design processes that may be of importance when interpreting the results.

Not all samples in this study may be representative of the actual end users, who are children with clinical anxiety. However, this argument assumes that there would be a difference in preferences between adolescents with and without anxiety, but perhaps it is feasible to first include adolescents as experts on *being young* as opposed to (but not exclusive of) *being anxious* when we want to gain knowledge on adolescent preferences in general.

XSL•FO
RenderX

Furthermore, demographic data on ethnicity and socioeconomic status were not collected for participants in phases 1 and 2, and it is unclear whether our sample was demographically representative. Future research could focus on collecting these data and recruiting participants from different locations (ie, schools in different areas, counseling services, and mental health services) to enhance diversity.

In addition, only 4 adolescents participated in the usability tests in phase 2. Although representing both the youngest and oldest adolescents in the intended age group (aged 12 and 16 years, respectively), the small number of participants limits the external validity of the results. It may be that the results obtained from the usability tests are specific to the platform and content viewed in this study. In the future, conducting usability tests online or "on-site" in places where adolescents regularly spend time (ie, in class or at youth centers) could aid the recruitment of more adolescents and increase sample size.

In this study, we emphasized the importance of systematically including adolescents as end users of an intervention. However, parents and clinicians may equally represent end users of this intervention and constitute valuable sources of information in a design and implementation context. We did include clinical psychologists working in routine care settings as part of the expert group to qualify the program manuscript but did not systematically collect data from them. This poses a limitation of this study, and it would be beneficial in the future to include all intended end users in the development processes.

It may also be that adolescents could have been involved more in the design of the intervention and at an earlier stage. In a study by Ludlow et al [16], the research team used co-design methods to understand preferences and perceptions of the design, functions, and engagement features within a new digital mental health platform. The aforementioned study used generative toolkits to facilitate the co-design processes through production of artifacts. These artifacts (and output from previous design phases) were then used to qualify the design of their new digital mental health platform for young people. This type of end-user involvement allows young people to hold a central position as experts at very early stages of the development process. However, it may also be extensive and require a lot of time and resources that may not be available.

When comparing this study to the one by Ludlow et al [16], it becomes apparent that, despite using user-centered methods, researchers have the power to define *what* end-user perspectives to include and *how* to investigate them. There may not be a way to solve this paradox, but it underlines how important it is to be aware of one's own impact as a researcher when engaging in these practices.

Conclusions

DMHIs have the potential to overcome treatment barriers such as stigma and costs and increase access to effective treatment at an early stage in symptom development. However, most DMHIs suffer from low treatment adherence. It has been proposed that the lack of end-user involvement in the development of intervention content and platforms may be of particular importance when interpreting the low adherence to DMHIs. In this study, we used user involvement practices at different stages of development to investigate adolescent and parent preferences on specific design components. Information from these practices was implemented when developing intervention content. This study emphasizes the importance of understanding the design process as iterative, where content prototypes are subjected to multiple phases of evaluation. This ensures the continued alignment of the intervention with end-user needs and may help establish the validity of the intervention for implementation in routine care settings. Future research may use user involvement practices such as those described in this paper to understand how to individualize intervention content and patient courses. In addition, user involvement practices could be used to investigate adolescent preferences on specific treatment components (eg, relaxation techniques and behavioral experiments) and the order in which they are presented (ie, in a specific order or as optional components).

Acknowledgments

The authors would like to thank all participating adolescents and families for their contributions to the development of the CoolMinds intervention. They greatly appreciate their interest and participation in the CoolMinds project. The authors would also like to thank psychologist Cecilie Holmen Frost for her role in writing and editing the CoolMinds intervention manuscript and the Development and Implementation Unit at the Centre for Digital Psychiatry for their contributions. Specifically, the authors would like to thank team leader in charge of user involvement Nicole Fronc Dumanski for her contribution to the planning and execution of the user involvement workshop, graphic designer Nicolai Drejer Walther for his extensive work on the graphic identity and content of the CoolMinds intervention, and consultant Alberthe Møller for setting up the intervention content on the digital platform. The development of CoolMinds is funded by a grant from the Mental Health Services of the Region of Southern Denmark and a grant from TrygFonden. Generative artificial intelligence was not used in the writing of this manuscript.

Data Availability

The data used in this manuscript contain sensitive and potentially identifiable information regarding both children and caregivers, and all participants were guaranteed full anonymity. As such, and in line with Danish Data Protection Agency regulations, the data used in this study are not publicly available. Unidentifiable data excerpts can be made available upon request to the corresponding author (nikita@psy.au.dk). Further data access requests may be sent to the Technology Transfer Office at Aarhus University (tto@au.dk).

XSL•FO
RenderX

Authors' Contributions

KM, JJL, MT, NMS, and HS contributed to conceptualization of the study. NMS conducted data curation. NMS completed the formal analysis. KM, MT, and JJL acquired funding. NMS, HS, and JJL collected data. KM, JJL, MT, NMS, and HS developed the methodology. NMS, and HS administered the project. NMS, JJL, and KM provided study materials. JJL, KM, and MT provided supervision. KM, JJL, and MT assisted with knowledge on previous and ongoing studies to validate the research conducted in this paper. NMS prepared data for presentation in the published work. NMS drafted the paper. JJL, KM, MT, and HS reviewed and edited the draft.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Guidance for Reporting Involvement of Patients and the Public checklist in long form.
[DOCX File , 19 KB - formative_v9i1e66966_app1.docx]

Multimedia Appendix 2
Interview guides.
[DOCX File , 22 KB - formative_v9i1e66966_app2.docx]

Multimedia Appendix 3
Results from the usability and feasibility tests.
[DOCX File , 39 KB - formative_v9i1e66966_app3.docx]

Multimedia Appendix 4
Overview of the content of the adolescent and parent program versions used when conducting the feasibility trial (version 1) and the randomized controlled trial (version 2).
[DOCX File , 21 KB - formative_v9i1e66966_app4.docx]

Multimedia Appendix 5
CoolMinds intervention content.
[DOCX File , 46164 KB - formative_v9i1e66966_app5.docx]

References

1. Nearchou FA, Bird N, Costello A, Duggan S, Gilroy J, Long R, et al. Personal and perceived public mental-health stigma as predictors of help-seeking intentions in adolescents. J Adolesc 2018 Jul;66:83-90. [doi: 10.1016/j.adolescence.2018.05.003] [Medline: 29800758]
2. Radez J, Reardon T, Creswell C, Orchard F, Waite P. Adolescents' perceived barriers and facilitators to seeking and accessing professional help for anxiety and depressive disorders: a qualitative interview study. Eur Child Adolesc Psychiatry 2022 Jun;31(6):891-907 [FREE Full text] [doi: 10.1007/s00787-020-01707-0] [Medline: 33502596]
3. Gulliver A, Griffiths KM, Christensen H. Perceived barriers and facilitators to mental health help-seeking in young people: a systematic review. BMC Psychiatry 2010 Dec 30;10:113 [FREE Full text] [doi: 10.1186/1471-244X-10-113] [Medline: 21192795]
4. Bennett SD, Cuijpers P, Ebert DD, McKenzie Smith M, Coughtrey AE, Heyman I, et al. Practitioner review: unguided and guided self-help interventions for common mental health disorders in children and adolescents: a systematic review and meta-analysis. J Child Psychol Psychiatry 2019 Aug;60(8):828-847. [doi: 10.1111/jcpp.13010] [Medline: 30775782]
5. Pennant ME, Loucas CE, Whittington C, Creswell C, Fonagy P, Fuggle P, et al. Computerised therapies for anxiety and depression in children and young people: a systematic review and meta-analysis. Behav Res Ther 2015 Apr;67:1-18. [doi: 10.1016/j.brat.2015.01.009] [Medline: 25727678]
6. Vigerland S, Lenhard F, Bonnert M, Lalouni M, Hedman E, Ahlen J, et al. Internet-delivered cognitive behavior therapy for children and adolescents: a systematic review and meta-analysis. Clin Psychol Rev 2016 Dec;50:1-10 [FREE Full text] [doi: 10.1016/j.cpr.2016.09.005] [Medline: 27668988]
7. Wickersham A, Barack T, Cross L, Downs J. Computerized cognitive behavioral therapy for treatment of depression and anxiety in adolescents: systematic review and meta-analysis. J Med Internet Res 2022 Apr 11;24(4):e29842 [FREE Full text] [doi: 10.2196/29842] [Medline: 35404263]
8. Clarke AM, Kuosmanen T, Barry MM. A systematic review of online youth mental health promotion and prevention interventions. J Youth Adolesc 2015 Jan;44(1):90-113. [doi: 10.1007/s10964-014-0165-0] [Medline: 25115460]

XSL•FO
RenderX

9. March S, Spence SH, Donovan CL. The efficacy of an internet-based cognitive-behavioral therapy intervention for child anxiety disorders. J Pediatr Psychol 2009 Jun;34(5):474-487. [doi: 10.1093/jpepsy/jsn099] [Medline: 18794187]
10. March S, Spence SH, Donovan CL, Kenardy JA. Large-scale dissemination of internet-based cognitive behavioral therapy for youth anxiety: feasibility and acceptability study. J Med Internet Res 2018 Jul 04;20(7):e234 [FREE Full text] [doi: 10.2196/jmir.9211] [Medline: 29973338]
11. Spence SH, Donovan CL, March S, Kenardy JA, Hearn CS. Generic versus disorder specific cognitive behavior therapy for social anxiety disorder in youth: a randomized controlled trial using internet delivery. Behav Res Ther 2017 Mar;90:41-57. [doi: 10.1016/J.BRAT.2016.12.003]
12. Wuthrich VM, Rapee RM, Cunningham MJ, Lyneham HJ, Hudson JL, Schniering CA. A randomized controlled trial of the Cool Teens CD-ROM computerized program for adolescent anxiety. J Am Acad Child Adolesc Psychiatry 2012 Mar;51(3):261-270. [doi: 10.1016/j.jaac.2011.12.002] [Medline: 22365462]
13. Waite P, Marshall T, Creswell C. A randomized controlled trial of internet-delivered cognitive behaviour therapy for adolescent anxiety disorders in a routine clinical care setting with and without parent sessions. Child Adolesc Ment Health 2019 Sep;24(3):242-250. [doi: 10.1111/camh.12311] [Medline: 32677216]
14. Hill C, Creswell C, Vigerland S, Nauta MH, March S, Donovan C, et al. Navigating the development and dissemination of internet cognitive behavioral therapy (iCBT) for anxiety disorders in children and young people: a consensus statement with recommendations from the #iCBTLorentz Workshop Group. Internet Interv 2018 Feb 19;12:1-10 [FREE Full text] [doi: 10.1016/j.invent.2018.02.002] [Medline: 30135763]
15. Fleming T, Merry S, Stasiak K, Hopkins S, Patolo T, Ruru S, et al. The importance of user segmentation for designing digital therapy for adolescent mental health: findings from scoping processes. JMIR Ment Health 2019 May 08;6(5):e12656 [FREE Full text] [doi: 10.2196/12656] [Medline: 31066705]
16. Ludlow K, Russell JK, Ryan B, Brown RL, Joynt T, Uhlmann LR, et al. Co-designing a digital mental health platform, "momentum", with young people aged 7-17: a qualitative study. Digit Health 2023 Nov 28;9:20552076231216410 [FREE Full text] [doi: 10.1177/20552076231216410] [Medline: 38033517]
17. De Vito Dabbs A, Myers BA, Mc Curry KR, Dunbar-Jacob J, Hawkins RP, Begey A, et al. User-centered design and interactive health technologies for patients. Comput Inform Nurs 2009;27(3):175-183. [doi: 10.1097/NCN.0b013e31819f7c7c] [Medline: 19411947]
18. Yardley L, Morrison L, Bradbury K, Muller I. The person-based approach to intervention development: application to digital health-related behavior change interventions. J Med Internet Res 2015 Jan 30;17(1):e30 [FREE Full text] [doi: 10.2196/jmir.4055] [Medline: 25639757]
19. Wozney L, Baxter P, Newton AS. Usability evaluation with mental health professionals and young people to develop an internet-based cognitive-behaviour therapy program for adolescents with anxiety disorders. BMC Pediatr 2015 Dec 16;15:213 [FREE Full text] [doi: 10.1186/s12887-015-0534-1] [Medline: 26675420]
20. Hill C, Chessell C, Percy R, Creswell C. Online Support and Intervention (OSI) for child anxiety: a case series within routine clinical practice. Behav Cogn Psychother 2022 May 04;50(4):429-445. [doi: 10.1017/s1352465822000157]
21. Hill C, Reardon T, Taylor L, Creswell C. Online support and intervention for child anxiety (OSI): development and usability testing. JMIR Form Res 2022 Apr 13;6(4):e29846 [FREE Full text] [doi: 10.2196/29846] [Medline: 35416781]
22. Moor S, Williman J, Drummond S, Fulton C, Mayes W, Ward N, et al. 'E' therapy in the community: examination of the uptake and effectiveness of BRAVE (a self-help computer programme for anxiety in children and adolescents) in primary care. Internet Interv 2019 Dec;18:100249. [doi: 10.1016/J.INVENT.2019.100249]
23. Vigerland S, Fredlander S, Aspvall K, Jolstedt M, Lenhard F, Mataix-Cols D, et al. Effectiveness of internet-delivered cognitive behavioural therapy for anxiety and obsessive-compulsive disorders within routine clinical care in rural Sweden. Internet Interv 2024 Jun;36:100738 [FREE Full text] [doi: 10.1016/j.invent.2024.100738] [Medline: 38617387]
24. Skaarnes H, Sørensen NM, Wisnewski AH, Lomholt JJ, Thastum M, McLellan L, et al. Internet-based cognitive behavioral intervention for adolescents with anxiety disorders: a study protocol for a parallel three armed randomized controlled trial. Trials 2024 Oct 13;25(1):674 [FREE Full text] [doi: 10.1186/s13063-024-08511-0] [Medline: 39396991]
25. Staniszewska S, Brett J, Simera I, Seers K, Mockford C, Goodlad S, et al. GRIPP2 reporting checklists: tools to improve reporting of patient and public involvement in research. BMJ 2017 Aug 02;358:j3453 [FREE Full text] [doi: 10.1136/bmj.j3453] [Medline: 28768629]
26. Horvath KJ, Ecklund AM, Hunt SL, Nelson TF, Toomey TL. Developing internet-based health interventions: a guide for public health researchers and practitioners. J Med Internet Res 2015 Jan 23;17(1):e28 [FREE Full text] [doi: 10.2196/jmir.3770] [Medline: 25650702]
27. Whiteside SP, Sim LA, Morrow AS, Farah WH, Hilliker DR, Murad MH, et al. A meta-analysis to guide the enhancement of CBT for childhood anxiety: exposure over anxiety management. Clin Child Fam Psychol Rev 2020 Mar;23(1):102-121 [FREE Full text] [doi: 10.1007/s10567-019-00303-2] [Medline: 31628568]
28. Jolstedt M, Wahlund T, Lenhard F, Ljótsson B, Mataix-Cols D, Nord M, et al. Efficacy and cost-effectiveness of therapist-guided internet cognitive behavioural therapy for paediatric anxiety disorders: a single-centre, single-blind, randomised controlled trial. Lancet Child Adolesc Health 2018 Nov;2(11):792-801. [doi: 10.1016/s2352-4642(18)30275-x]

XSL•FO
RenderX

29. Nordh M, Wahlund T, Jolstedt M, Sahlin H, Bjureberg J, Ahlen J, et al. Therapist-guided internet-delivered cognitive behavioral therapy vs internet-delivered supportive therapy for children and adolescents with social anxiety disorder: a randomized clinical trial. JAMA Psychiatry 2021 Jul 01;78(7):705-713 [FREE Full text] [doi: 10.1001/jamapsychiatry.2021.0469] [Medline: 33978699]
30. Stjerneklar S, Hougaard E, McLellan LF, Thastum M. A randomized controlled trial examining the efficacy of an internet-based cognitive behavioral therapy program for adolescents with anxiety disorders. PLoS One 2019 Sep 18;14(9):e0222485 [FREE Full text] [doi: 10.1371/journal.pone.0222485] [Medline: 31532802]
31. Vigerland S, Ljótsson B, Thulin U, Öst LG, Andersson G, Serlachius E. Internet-delivered cognitive behavioural therapy for children with anxiety disorders: a randomised controlled trial. Behav Res Ther 2016 Jan;76:47-56 [FREE Full text] [doi: 10.1016/j.brat.2015.11.006] [Medline: 26649465]
32. Higa-McMillan CK, Francis SE, Rith-Najarian L, Chorpita BF. Evidence base update: 50 years of research on treatment for child and adolescent anxiety. J Clin Child Adolesc Psychol 2016;45(2):91-113. [doi: 10.1080/15374416.2015.1046177] [Medline: 26087438]
33. Luo A, McAloon J. Potential mechanisms of change in cognitive behavioral therapy for childhood anxiety: a meta-analysis. Depress Anxiety 2021 Feb;38(2):220-232. [doi: 10.1002/da.23116] [Medline: 33225527]
34. Bertie LA, Hudson JL. CBT for childhood anxiety: reviewing the state of personalised intervention research. Front Psychol 2021 Nov 26;12:722546 [FREE Full text] [doi: 10.3389/fpsyg.2021.722546] [Medline: 34899467]
35. Smart K, Smith L, Harvey K, Waite P. The acceptability of a therapist-assisted internet-delivered cognitive behaviour therapy program for the treatment of anxiety disorders in adolescents: a qualitative study. Eur Child Adolesc Psychiatry 2023 Apr;32(4):661-673 [FREE Full text] [doi: 10.1007/s00787-021-01903-6] [Medline: 34746976]
36. Radomski AD, Bagnell A, Curtis S, Hartling L, Newton AS. Examining the usage, user experience, and perceived impact of an internet-based cognitive behavioral therapy program for adolescents with anxiety: randomized controlled trial. JMIR Ment Health 2020 Feb 07;7(2):e15795 [FREE Full text] [doi: 10.2196/15795] [Medline: 32022692]
37. Spence SH, March S, Donovan CL. Social support as a predictor of treatment adherence and response in an open-access, self-help, internet-delivered cognitive behavior therapy program for child and adolescent anxiety. Internet Interv 2019 Aug 22;18:100268 [FREE Full text] [doi: 10.1016/j.invent.2019.100268] [Medline: 31890621]
38. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381 [FREE Full text] [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
39. Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O'Neal L, et al. The REDCap consortium: building an international community of software platform partners. J Biomed Inform 2019 Jul;95:103208 [FREE Full text] [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]
40. Neergaard MA, Olesen F, Andersen RS, Sondergaard J. Qualitative description - the poor cousin of health research? BMC Med Res Methodol 2009 Jul 16;9:52 [FREE Full text] [doi: 10.1186/1471-2288-9-52] [Medline: 19607668]
41. Garrido S, Millington C, Cheers D, Boydell K, Schubert E, Meade T, et al. What works and what doesn't work? A systematic review of digital mental health interventions for depression and anxiety in young people. Front Psychiatry 2019 Nov 13;10:759. [doi: 10.3389/fpsyt.2019.00759] [Medline: 31798468]
42. Wood D, Bruner JS, Ross G. The role of tutoring in problem solving. J Child Psychol Psychiatry 1976 Apr;17(2):89-100. [doi: 10.1111/j.1469-7610.1976.tb00381.x] [Medline: 932126]
43. Stjerneklar S, Hougaard E, Thastum M. Guided internet-based cognitive behavioral therapy for adolescent anxiety: Predictors of treatment response. Internet Interv 2019 Mar;15:116-125 [FREE Full text] [doi: 10.1016/j.invent.2019.01.003] [Medline: 30792963]
44. Bandura A. Social cognitive theory: an agentic perspective. Annu Rev Psychol 2001;52:1-26. [doi: 10.1146/annurev.psych.52.1.1] [Medline: 11148297]
45. Ludlow K, March S, Russell JK, Ryan B, Brown RL, Hides L, et al. Co-designing a new digital mental health platform, 'momentum', with caregivers of young people aged 7–17. Digit Health 2024 Aug 02;10. [doi: 10.1177/20552076241263693]
46. McCall HC, Hadjistavropoulos HD, Sundström CR. Exploring the role of persuasive design in unguided internet-delivered cognitive behavioral therapy for depression and anxiety among adults: systematic review, meta-analysis, and meta-regression. J Med Internet Res 2021 Apr 29;23(4):e26939 [FREE Full text] [doi: 10.2196/26939] [Medline: 33913811]
47. Patterson VC, Rossi MA, Pencer A, Wozney L. An internet-based cognitive behavioral therapy program for anxiety and depression (tranquility): adaptation co-design and fidelity evaluation study. JMIR Form Res 2022 Feb 02;6(2):e33374. [doi: 10.2196/33374] [Medline: 34910660]

Abbreviations

CBT: cognitive behavioral therapy
DMHI: digital mental health intervention
ICBT: internet-based cognitive behavioral therapy
RCT: randomized controlled trial

XSL•FO
RenderX

REDCap: Research Electronic Data Capture

Edited by A Mavragani; submitted 27.09.24; peer-reviewed by C Donovan, L Kagan; comments to author 15.01.25; revised version received 03.02.25; accepted 02.03.25; published 08.04.25.

Please cite as:
Sørensen NM, Skaarnes H, Mathiasen K, Thastum M, Lomholt JJ
Developing an Internet-Based Cognitive Behavioral Therapy Intervention for Adolescents With Anxiety Disorders: Design, Usability, and Initial Evaluation of the CoolMinds Intervention
JMIR Form Res 2025;9:e66966
URL: https://formative.jmir.org/2025/1/e66966
doi:10.2196/66966
PMID:

©Nikita Marie Sørensen, Helene Skaarnes, Kim Mathiasen, Mikael Thastum, Johanne Jeppesen Lomholt. Originally published in JMIR Formative Research (https://formative.jmir.org), 08.04.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Opportunities and Challenges Surrounding the Use of Wearable Sensor Bracelets for Infectious Disease Detection During Hajj: Qualitative Interview Study

Noha Maddah[1,2], BSc, MSc, MPH; Arpana Verma[3], MBChB, MPH, PhD; John Ainsworth[1], MSc, PhD

[1]Division of Informatics Imaging and Data Sciences, School of Health Sciences, Faculty of Biology, Medicine and Health, Manchester Academic Health Science Centre, Centre for Health Informatics, The University of Manchester, Manchester, United Kingdom

[2]Department of Health Services and Hospitals Administration, Faculty of Economics and Administration, King Abdulaziz University, Jeddah, Saudi Arabia

[3]Division of Population Health, Health Services Research & Primary Care, Faculty of Biology, Medicine and Health, Manchester Academic Health Sciences Centre, The University of Manchester, Manchester, United Kingdom

Corresponding Author:
Noha Maddah, BSc, MSc, MPH
Division of Informatics Imaging and Data Sciences, School of Health Sciences, Faculty of Biology, Medicine and Health
Manchester Academic Health Science Centre, Centre for Health Informatics
The University of Manchester
Vaughn House
Manchester, M13 9GB
United Kingdom
Phone: 44 1612757675
Email: noha.maddah@postgrad.manchester.ac.uk

Abstract

Background: Wearable sensor bracelets have gained interest for their ability to detect symptomatic and presymptomatic infections through alterations in physiological indicators. Nevertheless, the use of these devices for public health surveillance among attendees of large-scale events such as hajj, the Islamic religious mass gathering held in Saudi Arabia, is currently in a nascent phase.

Objective: This study aimed to explore hajj stakeholders' perspectives on the use of wearable sensor bracelets for disease detection.

Methods: We conducted a qualitative, theoretically informed, interview-based study from March 2022 to October 2023 involving a diverse sample of hajj stakeholders, including technology experts, health care providers, and hajj service providers. The study was guided by the task-technology fit model and the unified theory of acceptance and use of technology to provide a comprehensive understanding of the factors influencing the acceptance and use of the technology. Semistructured in-depth interviews were used to capture perspectives on using wearable sensor bracelets for infectious disease detection during hajj. Thematic analysis of interview transcripts was conducted.

Results: A total of 14 individuals were interviewed. In total, 4 main themes and 13 subthemes emerged from the study, highlighting crucial challenges, considerations, recommendations, and opportunities in the use of wearable sensor bracelets for the presymptomatic detection of infectious diseases during hajj. Implementing wearable sensor bracelets for disease detection during hajj faces obstacles from multiple perspectives, encompassing users, implementing stakeholders, and technological factors. Hajj stakeholders were concerned about the substantial financial and operational barriers. The motivation of implementing stakeholders and users is essential for the acceptance and uptake of devices during hajj. Successful integration of wearables into the hajj surveillance system depends on several factors, including infrastructure, device features, suitable use cases, training, and a smooth organizational integration process.

Conclusions: This study provides valuable insights into the potential opportunities and challenges of adopting wearable sensor bracelets for disease detection during hajj. It offers essential factors to consider and important suggestions to enhance comprehension and ensure the effective implementation of this technology.

(JMIR Form Res 2025;9:e60484) doi:10.2196/60484

XSL•FO
RenderX

KEYWORDS

wearable sensor; unified theory of acceptance and use of technology; task-technology fit; hajj; presymptomatic detection; infectious diseases; artificial intelligence

Introduction

New Generation of Digital Tools for Public Health Surveillance

Understanding and reacting to health trends affecting communities is a core function of public health systems [1]. Public health surveillance is the tool used to respond to these health trends through "the ongoing, systematic collection, analysis, and interpretation of health-related data essential to planning, implementation, and evaluation of public health practice" [1]. The COVID-19 pandemic has brought to light the urgent need for real-time public health surveillance to enhance the use of evidence in decision-making [1]. The 2020 Riyadh Declaration was formulated as a response to the global pandemic, aiming to propose a series of suggestions that would effectively tackle the shortcomings observed in global public health response systems. The Riyadh Declaration emphasized the significance of implementing digital health solutions that are both scalable and sustainable, as well as the adoption of health intelligence [1].

Traditional public health initiatives targeted at preventing the spread of infectious diseases such as the recent COVID-19 have limitations that digital health technologies could address [2]. The World Health Organization has defined digital health technologies as "the field of knowledge and practice associated with the development and use of digital technologies to improve health" [3]. The use of digital health tools during the global pandemic has exhibited their efficacy in fighting, mitigating, and controlling the transmission of emerging infectious diseases [2]. There is a significant potential for adopting digital surveillance as a valuable tool in monitoring and controlling infectious diseases [4]. Internet of Things (IoT) technological advancement has proven its ability to augment and optimize conventional preventative public health strategies and methodologies [2]. These internet-based technologies possess certain characteristics that enable the expedited detection of infectious diseases [4]. The term IoT pertains to a network of interconnected objects, devices, and systems that engage in electronic communication and receive, process, and send digital data with minimal human involvement [5]. IoT is expected to reach its "plateau of productivity" during the next 5 to 10 years [5].

Detection is the initial and important aspect in preventing infectious diseases [5]. In the context of these diseases, the computation speed plays a crucial role in reducing the time required for their diagnosis, facilitating prompt implementation of preventive measures. The performance of system automation has been found to enhance surveillance's efficacy in widespread disease transmission. In addition, it has proven to be an effective means of reducing the workload of medical workers during pandemics [4].

Wearable Technologies and Body Sensors

Wearable technologies are integral to widespread IoT technical advancements [6]. Wearable technologies are intelligent electronic devices worn on the body to measure, analyze, and transfer information. The data may encompass many bodily indicators associated with essential physiological parameters and levels of physical exertion. These wearables have the capability to provide information to the user by means of display technology or vibrotactile feedback [7].

Wearable sensors have garnered considerable attention in the past 10 years, with a particular focus on their application within the health care sector [8,9]. Their ability to establish connections with other devices and the IoT, facilitating data exchange, has led to their designation as "smart" devices, setting them apart from traditional, nonconnected step counters, wristwatches, or clothing [10]. Wearable sensors have the capability to monitor, collect, and send data [6]. They offer precise, credible, and real-time data regarding an individual's behaviors and activities, hence aiding users in identifying potential issues. Wearables collect raw data from measurements using sensors to store them and use them for things such as performance evaluation and ongoing health monitoring [11]. These sensors can be applied in several areas, such as illness monitoring, diagnosis and treatment, and health management [12].

Various terms are used to refer to wearable sensor devices that capture and analyze real-time physical data in a nonintrusive manner. These terms encompass *smartwatches* [9,13], *smart sensors* [14], *wearable activity trackers* [15], *wearable sensors* [16], *wearable physiological sensors* [17], *integrated biosensors* [18], and *sensor bracelets* [19]. In the context of this research, the phrase "sensor bracelet" will be used.

Extensive research has been conducted on wearables to investigate their potential in health monitoring. However, there is a noticeable trend in current research toward applying wearables for the presymptomatic detection of infectious diseases [9,19]. This transition signifies the growing importance of early detection of diseases, especially within the context of public health and large-scale occurrences. The use of wearable technology has emerged as a significant factor in this pursuit, presenting the possibility of identifying infections before the manifestation of symptoms. This capability facilitates prompt interventions and preventive actions. The increasing emphasis on this matter highlights the significant contribution that wearables can make in augmenting health care and methods for disease control.

Wearable devices (sensor bracelets) have gained attention for their ability to anticipate both symptomatic and presymptomatic identification of viral respiratory infections by sensing changes in physiological markers, as highlighted in the recent literature [19,20]. Mitratza et al [16] conducted a systematic review to assess the value of a range of wearable sensor devices in detecting COVID-19. In their review, they noted that, while wearable devices show promising potential for detecting

XSL•FO
RenderX

COVID-19, the findings are still in the early stages. They emphasized the need to consider how differences in sensor methods, data processing, and algorithms affect the detection of infection-related changes in physiological measurements, as well as the importance of addressing sources of bias. Building on this, Risch et al [19] assessed the use of a smart bracelet in the early detection of changes in physiological parameters related to COVID-19. With a low false-positive rate, their wearable sensor bracelet algorithm was able to identify 71% of COVID-19 cases up to 2 days before symptoms manifested. This demonstrates the efficacy of using wearable technologies for the early detection of diseases. Mishra et al [9] conducted a study whereby they demonstrated the use of data from smartwatches to detect respiratory infections such as COVID-19 in real time within a large-scale scenario. The study identified 63% of COVID-19 infections with a low false-positive rate. Furthermore, Hirten et al [2] evaluated the longitudinal collection of heart rate variability (HRV) metrics from a wearable device for the prediction of COVID-19. Their study indicated that HRV data from the Apple Watch can predict COVID-19 and identify symptoms. In addition, alterations in HRV were observed before a positive polymerase chain reaction test, indicating its potential for early infection detection.

Hajj, the Largest Annual Religious Mass Gathering, and Disease Transmission

Safety and security have consistently been of utmost importance within the framework of hajj, a significant religious event that attracts participants worldwide. This annual massive religious gathering in Makkah, Saudi Arabia, attracts almost 3 million pilgrims from 180 countries. These pilgrims spend several weeks in the cities of Makkah and Madinah. The problem of overcrowding and the lack of compliance with government-mandated preventative measures for diseases have been highlighted in several previous studies [17,21-23]. This lack of compliance contributed to a long-standing high occurrence of respiratory infections before the COVID-19 outbreak. Low adherence among pilgrims has been attributed to a number of factors, including insufficient coverage and limited access to information [24].

Using intelligent and comprehensive monitoring tools for participants in this mass event is crucial to reducing the risk of future tragedies. Saudi Arabia uses several technical methods such as spatial computing, crowd simulation, mobile apps, and big data analytics for hajj activities [25].

In recent times, there has been an increasing focus on incorporating state-of-the-art technologies to tackle the distinctive difficulties linked to hajj. The hajj pilgrimage, characterized by its large-scale congregation of millions of pilgrims, poses significant logistical, health, and safety complexities. Disease surveillance has emerged as a crucial component of hajj management, progressively evolving over time by assimilating insights gained from previous operations.

The Kingdom of Saudi Arabia has been at the forefront of implementing digital solutions to improve the pilgrimage experience as part of the national strategy and the 2030 vision. One of the notable technological developments in this context involves the application of the smart bracelet initiative, which involved the collaboration of multiple governmental authorities. In 2021, the Pilgrims' Smart Bracelet initiative was introduced, offering 5000 pilgrims bracelets equipped with data storage and transmission capabilities, thus contributing to the IoT. The main purpose of the bracelet was to provide extensive information on the pilgrim, including their health condition. The device enabled the tracking of important health indicators such as blood oxygen saturation and pulse rate while also offering functionalities for requesting immediate medical or security support, guaranteeing swift reaction and help. In addition, pilgrims were given awareness messages via the bracelet they wore. The initiative's implementation entailed incorporating and enhancing business models and systems, spanning both operational and technological facets of the project [26].

The practice of postmortem detection for infectious diseases underscores the criticality of real-time data and analytics to avert their transmission, which has severe social and economic repercussions for citizens worldwide [5]. Within the framework of hajj, it is imperative to not only identify diseases before symptoms appear but also find and follow the individuals who have been in contact with infected individuals. This requires incorporating tracking elements into the device. The current state of affairs presents an opportune environment for scientific inquiry specifically aimed at recognizing and addressing nascent hazards and difficulties to predict and alleviate them. Hence, this study endeavored to investigate the opportunities and challenges surrounding using wearable sensor bracelets for presymptomatic detection of infectious diseases during hajj.

Methods

Theoretical Framework

Technology acceptance models have been criticized for their limited ability to evaluate the extent to which wearable technology offers the requisite functionalities to attain desired outcomes. This concern can be effectively resolved by integrating the task-technology fit (TTF) and technology acceptance models [27].

Studies have shown that integrating the TTF model and the unified theory of acceptance and use of technology (UTAUT) enhances the predictive strength for technology adoption. The TTF-UTAUT model integrates 2 key frameworks to predict technology adoption [28]. The TTF model is defined as the interdependence between individual (a technology user), technology (data, hardware, software tools, and the services they provide), and task (an activity carried out by individuals to produce the required output) characteristics [26]. The UTAUT suggests that the actual use of technology is determined by behavioral intention. The perceived likelihood of adopting the technology depends on the direct effect of 4 key constructs: performance expectancy, effort expectancy, social influence, and facilitating conditions [29].

The UTAUT is considered beneficial for the complex nature of this research, which involves evaluating several interrelated factors in disciplines such as health informatics and public health during large-scale events. The UTAUT can provide insights into the determinants that affect the adoption and use of

XSL•FO
RenderX

wearable sensor bracelets among various stakeholders participating in the hajj pilgrimage. To provide a more thorough explanation, the TTF model was integrated with the UTAUT. The TTF model can be used to evaluate the degree to which the features of the technology align with the requirements and demands of its users, including both pilgrims and authorities. This integration improves understanding and focuses on particular elements inside the UTAUT.

To aid in effectively applying wearable sensor bracelets during the hajj pilgrimage, this integrated approach can offer a comprehensive understanding of stakeholders' viewpoints. Figure 1 [28] presents the model used.

Figure 1. Task-technology fit and unified theory of acceptance and use of technology.

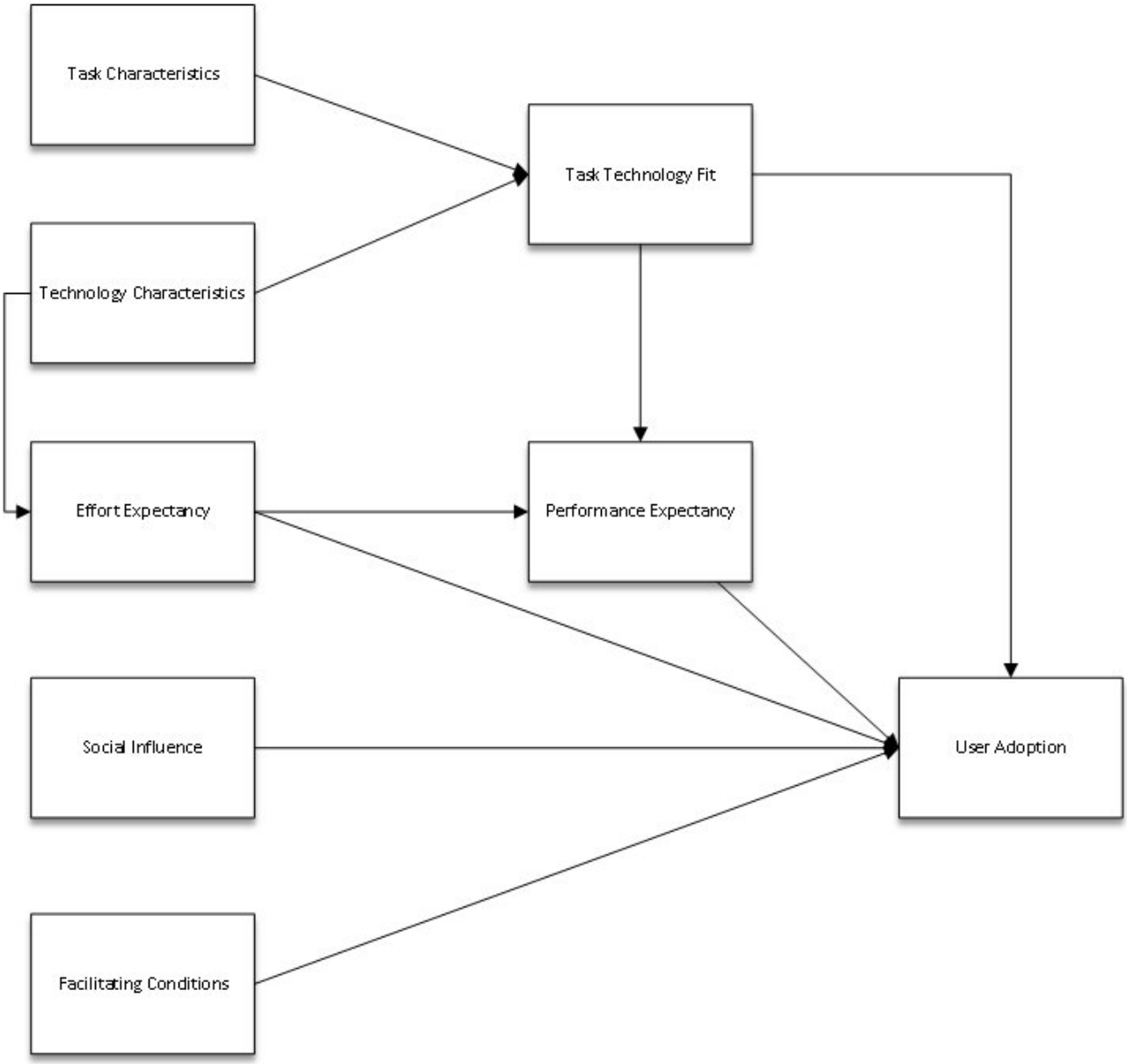

Study Design

We conducted in-depth, semistructured, one-to-one qualitative interviews from March 2022 to October 2023 with 14 selected hajj stakeholders, organizers, and professionals working in the field to capture their perspectives on using wearable sensor bracelets for infectious disease detection during hajj.

Data Collection and Sampling Strategy

Overview

A nested cohort approach was used to explore stakeholders' perspectives [30]. Participants were identified through purposive sampling and snowball sampling techniques. Purposive sampling is a method used to gather information and gain a comprehensive understanding of a certain topic of interest. This methodology entails the selection of data cases based on their capacity to yield data that are rich in information for comprehensive analysis [31].

A purposive sampling strategy was used to identify participants who were directly involved in the smart bracelet initiative. This resulted in a core group of 4 participants who held senior positions within the organizational hierarchy. The key stakeholders were identified through the organizational website and were reached purposefully via email. These individuals

XSL•FO
RenderX

offered comprehensive perspectives on the initiative as a result of their direct expertise and authoritative decision-making positions. Concurrently, a more extensive cohort of 10 participants selected using snowball sampling was incorporated to gather insights into the general use of technology during hajj, specifically from individuals who directly interacted with pilgrims. Although these individuals were not directly involved in the smart bracelet initiative, their contact with pilgrims offered vital contextual information. We reached out to these participants directly via their contact numbers.

The inclusion and exclusion criteria for participants are detailed in Textbox 1. This study adhered to the guidelines outlined in the COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist [32].

Textbox 1. Inclusion and exclusion criteria for participants in this study.

Inclusion criteria

- Individuals in senior positions within the organizational hierarchy who were directly involved in the smart bracelet initiative, including those with technological, medical, organizational, or logistical expertise
- Individuals who have experience and are present in service organizations during hajj (eg, *tawafa* organizations) and can offer their perspective on the technologies used and how pilgrims interact with or accept them

Exclusion criteria

- Individuals who are not directly involved in the smart bracelet initiative or do not occupy senior positions
- Individuals who lack pertinent experience or insight into the adoption and use of technology by pilgrims during hajj.

Interviews

We used a systematic approach to engage with our potential participants before the interviews started. The interview questions were then given to participants who replied favorably and indicated their desire to take part. This action guaranteed that participants had enough time to prepare for the subsequent interview, promoting a deeper and more thoughtful discussion. We then scheduled the interviews after allowing participants to receive the interview questions and prepare.

Interview Topic Guide

The interviews were guided by a semistructured topic guide exploring professional background, experience, and involvement with the 2021 Pilgrims' Smart Bracelet initiative. Questions were based on existing literature and organizations' online information to understand the current capabilities and existing tools used during hajj in addition to opportunities, gaps, and challenges associated with implementing wearable sensor bracelets for infectious disease detection during hajj. Questions were tailored to the specific characteristics of participants and evolved in light of emerging findings. We explored the advancements of existing tools and technologies used during hajj for public health surveillance, their limitations, and stakeholders' vision and perceptions regarding using wearable sensor bracelets for disease detection during hajj.

The interview questions for key stakeholders who were directly involved in the smart bracelet initiative primarily addressed the capabilities of the surveillance system, the technologies and tools used during hajj to support the event, their opinions on wearables for disease detection, and their insights into the smart bracelet initiative. In contrast, interviews conducted with participants who had minimal involvement in the smart bracelet initiative but were highly involved in assisting pilgrims focused on their experiences and viewpoints regarding the acceptability of the latest technology deployed during hajj by the pilgrims. A copy of the topic guide can be found in Multimedia Appendix 1.

Digital audio recorders with encryption were used to record the interviews. Interviews were transcribed verbatim, read numerous times, and annotated to fully immerse us in the facts.

Data Analysis

Thematic analysis of interview transcripts was conducted [31]. This study used an inductive thematic analysis to investigate participants' experiences and perceptions regarding the use of wearable sensor bracelets for disease detection during hajj. The UTAUT and TTF framework variables served as a guide for the interview questions, but themes were allowed to naturally arise from the data. In the subsequent analysis, a deductive approach was applied to identify relationships between the emergent themes and the theoretical constructs of the UTAUT and TTF model. This hybrid approach allowed for both the discovery of novel insights and the integration of findings with established theoretical frameworks.

The primary author, NM, led the data analysis by integrating the UTAUT [33] and TTF models [34] to create a comprehensive framework. This framework posits that the user's intention to use technology and their acceptance of it are influenced by many criteria derived from the UTAUT as well as the degree of compatibility between the task at hand and the technology being used [28].

In this analysis, we explored the use of wearable sensor bracelets for disease detection during hajj. We thoroughly evaluated several aspects such as the integration of technology, development of systems, augmentation of infrastructure, and acceptance among users and suppliers. This comprehensive exploration allowed us to understand the intricate interplay of factors influencing the successful deployment of wearable sensor technology within the unique context of hajj.

Upon completing data collection in our study, we promptly commenced the analysis process to identify any significant findings.

XSL•FO
RenderX

We eventually arrived at the realization that we had obtained enough data to attain data saturation when we noticed that the same subjects kept cropping up during our observations of the data-gathering process and no new ones were introduced. After completing the data collection and analysis, it became evident that no new themes were developing. This confirmed that we had achieved a state of saturation in our findings.

To improve the dependability and verification, synopses of discoveries were regularly communicated to other researchers engaged in the study subsequent to each interview with a participant. The completed transcripts and field notes were systematically arranged and categorized in NVivo (version 12; QSR International) using researcher-derived codes (latent codes) obtained from the interview guide. We followed the characteristics of an inductive coding strategy to guarantee thoroughness and complexity in our investigation [31].

Ethical Considerations

According to the ethical guidelines set forth by the University of Manchester, this research study did not require independent ethical review and was categorized as low-risk research ("Work with professionals; including interviews, focus groups, questionnaires, etc" [35]). All study data were fully anonymized, with no personal identifiers collected, and were securely stored to ensure privacy and confidentiality in accordance with the University of Manchester data protection guidelines. We initially sent a participant information sheet to each potential participant via digital communication tools informing them of the study's purpose and why they had been selected. Participation was voluntary, and no compensation was offered or required.

Results

Participant Characteristics

A total of 14 semistructured, in-depth interviews were carried out, with sessions having a duration ranging from 20 to 80 minutes. The mean duration of the interviews was 32.0 (SD 20.8) minutes. The variation in interview durations can be ascribed to the participants' varying levels of expertise and knowledge and the degree of their engagement with the smart bracelet initiative. Individuals that had greater direct engagement with the initiative generally offered more comprehensive perspectives, leading to lengthier interviews. In contrast, individuals with limited direct experience or engagement generally had shorter interviews.

The interview participants in our study encompassed a wide array of backgrounds, which included individuals involved in hajj, specialists in public health, and professionals from several disciplines (Table 1). A subset of the individuals involved in the study held prominent positions within governmental entities, whereas others held positions with varied levels of decision-making power.

Table 1. Study participants' expertise and sectors.

Participant number	Expertise	Sector
1	Systems engineering, business, and project management	Hajj and *umrah* services
2	Health care provider and public health specialist	Health care and public health
3	Health care provider and public health specialist	Health care and public health
4	Communication engineering and artificial intelligence	Technology or consulting (business)
5	Service provider	Accommodation
6	Service provider	Accommodation
7	Service provider	Accommodation
8	Service provider	Accommodation
9	Service provider	Guidance to holy sites
10	Service provider	Enrichment (educational and awareness)
11	Service provider	Arrivals and departures
12	Service provider	Arrivals and departures
13	Service provider	Transportation
14	Service provider	Transportation

In total, 2 distinct groups were formed based on the results of the conducted interviews. The initial cohort consisted of 4 participants who were directly involved in the 2021 initiative and made significant contributions to this study. The second cohort consisted of 10 persons who were not directly involved in the 2021 initiative and had no additional insights to contribute regarding the use of wearable devices during hajj. Nevertheless, these individuals were involved in the provision and coordination of hajj services as well as direct engagement with pilgrims.

Each participant had a minimum of 10 years of professional experience in their respective fields. The individuals who made these contributions were located in Saudi Arabia. A total of 71% (10/14) of the interviews were conducted via telephone, and the remaining 29% (4/14) were conducted using the Zoom platform (Zoom Video Communications). Audio recordings were gathered for all interviews.

XSL•FO
RenderX

Generated Themes: An Overview

The analysis generated insights on 4 main themes and 13 subthemes (Figure 2), highlighting crucial challenges, considerations, recommendations, and opportunities in the use of wearable sensor bracelets for the presymptomatic detection of infectious diseases during hajj. A narrative synthesis was used to report the findings obtained from data collected from the first group of 29% (4/14) of the participants who played a role in the 2021 smart bracelet initiative. Table 2 highlights inputs from each of the 4 participants related to the identified themes.

For the second group, a brief summary of the findings is presented highlighting important observations without a detailed analysis.

Figure 2. Thematic analysis of the conceptual theme framework of the interviews. AI: artificial intelligence.

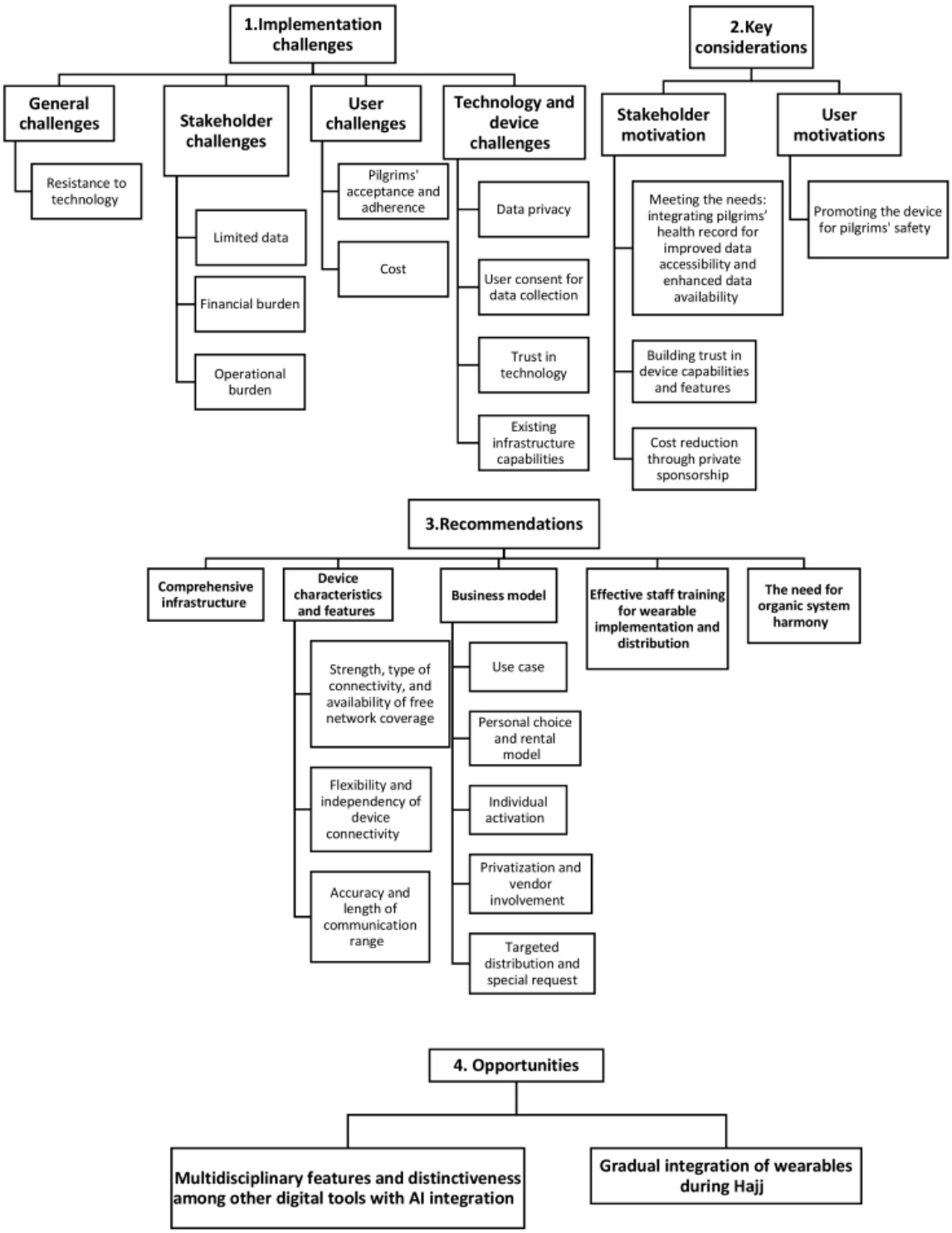

XSL•FO
RenderX

Table 2. Themes generated during data collection.

Theme and subthemes	Participant			
	1	2	3	4
Implementation challenges				
General challenges				
Resistance to technology		✓		
Stakeholder challenges				
Limited data		✓		
Financial burden			✓	✓
Operational burden	✓			
User challenges				
Pilgrims' acceptance and adherence		✓		
Cost				✓
Technology and device challenges				
Data privacy	✓	✓		✓
User consent for data collection	✓			✓
Trust in technology			✓	
Existing infrastructure capabilities			✓	
Key considerations				
Stakeholder motivations				
Meeting needs		✓		✓
Building trust in device features and capabilities	✓	✓		
Cost reduction through private sponsorship	✓	✓	✓	
Interest of entities involved	✓			✓
User motivations				
Promoting the device for pilgrims' safety	✓	✓		
Recommendations				
Comprehensive infrastructure			✓	✓
Device characteristics and features				✓
Business model	✓	✓		✓
Effective staff training for wearable implementation and distribution			✓	
The need for organic system harmony			✓	
Opportunities				
Multidisciplinary features and distinctiveness among other digital tools with AI[a] integration		✓		✓
Gradual integration of wearables during hajj		✓		

[a]AI: artificial intelligence.

The First Group's Perceptions of Using Wearable Sensor Bracelets

Overview

The idea of "smart hajj" was a recurring topic in the interviews as respondents emphasized their goals for the future and their current capabilities and constraints. The concept of smart hajj signifies the continuous endeavors of the hajj organization to adapt to technological advancements and improve the overall pilgrimage journey. The concept comprises various technical solutions designed to tackle difficulties and enhance the welfare and security of pilgrims.

However, the deployment of smart hajj and incorporating wearable devices into the technology used by hajj authorities present challenges and potential. Several important points were raised, and suggestions were made by the participants.

XSL•FO
RenderX

Implementation Challenges of Wearable Sensor Bracelets for Infectious Disease Detection During Hajj

In the context of using digital health technologies for public health surveillance, participants in our study highlighted a range of challenges in deploying wearable sensor bracelets for disease detection during hajj. The issues encompassed a range of areas, including general challenges, challenges connected to implementing stakeholders, challenges linked to users, and challenges related to devices and technology.

General Challenges: Resistance to Technology

Participants voiced apprehensions over the wider difficulties linked to integrating technology such as wearables within the public health domain. They emphasized a vital concern related to resistance from both users and implementing stakeholders due to factors such as data privacy, information security, and human rights:

> *We will face a challenge in using technology for public health due to data, information, and human rights. This challenge is present, and we will face resistance.* [Participant 2; health care provider and public health specialist]

Stakeholder Challenges: Limited Data, Financial Burden, and Operational Burden

Participants cited concerns regarding various factors from the hajj implementing stakeholders' perspective, including limited data availability, resistance to adoption, and the operational and financial burdens associated with wearable implementation. In addition, the conversation explored several aspects pertaining to the previous smart bracelet initiative, elucidating significant barriers that hindered its advancement.

The limited availability of health data can be attributed to the apprehensions of pilgrims regarding disclosing their medical records to hajj authorities. There was a concern among individuals that the act of revealing specific information pertaining to their health status and underlying medical disorders would potentially impede their ability to be granted permission to partake in the hajj pilgrimage or, alternatively, affect their eligibility to obtain hajj visas:

> *Pilgrims worry that if they mention any medical condition, that the Ministry of Hajj and Umrah will not issue them the visa. For example, they do not mention that they have Cancer fearing that they will not get issued the visa. I do not mind if the pilgrim has Cancer, I will treat you, but I need to know in advance to prepare and provide you with the service.* [Participant 2; health care provider and public health specialist]

One major issue brought up by participants, especially the hajj organizers, was the expense of integrating wearables throughout the pilgrimage. Considering the wide range of services they already offer, they voiced concerns regarding the cost load related to the adoption of wearables. Participants attributed the failure and incomplete implementation of the previous smart bracelet initiative (NUSUK) to the high cost of wearables and the significant budget needed for their deployment. Hajj organizers questioned whether they could cover the expense of integrating wearables for all pilgrims:

> *Let's return to the smart bracelet project and why it was not completed. Finance was an issue at that stage, and the projects couldn't proceed. The cost of these devices is considered high.* [Participant 4; communication engineering and artificial intelligence (AI)]

> *Can we afford to provide the wearable to all pilgrims?* [Participant 3; health care provider and public health specialist]

Participants highlighted the operational challenges associated with implementing ubiquitous sensor bracelets, including the need for staff training and the availability of expert teams for data analysis and decision-making. Concerns were also expressed over the substantial data flow generated by wearables. Despite the complete readiness of the expert teams, there were concerns that these efforts might not produce valuable outcomes owing to a lack of user engagement:

> *It will be a huge operational load.It needs a specialised team that deals with the data collected by the watch and decides whether the person needs a team to be dispatched. So, there was a big operational load with a minimal value, especially during Hajj.* [Participant 1; systems engineering, business, and project management]

User Challenges

Pilgrims' Acceptance and Adherence

Participants underlined the difficulty of obtaining user acceptance, particularly in light of the wide range of backgrounds, beliefs, and degrees of faith in technology exhibited by pilgrims hailing from >180 nations worldwide. Stakeholders observed that the implementation of the tracking and tracing feature may potentially discourage certain pilgrims from adopting or adhering to the technology:

> *Once you mention to the user that they can't take the bracelet off their wrist, or you can't give it to someone else, cut it, or I will notice what happens to the bracelet once you have done any of these. It will be like you are tracking me as a prisoner, and there will be high resistance.* [Participant 2; health care provider and public health specialist]

Cost

In addition to being a financial burden for the providers, wearable device costs may be seen as an obstacle by users if they are part of their hajj package. Participants voiced apprehension regarding the potential financial burden of including the cost of the bracelet in the pilgrims' bundles, deeming it to be costly:

> *If we estimate that the cost of the bracelet will be 100 Saudi Riyal, it will be considered expensive for pilgrims to add that amount to their package.* [Participant 4; communication engineering and AI]

XSL•FO
RenderX

Technology and Device Challenges: Data Privacy and Users' Consent, Trust in Technology, and Existing Infrastructure Capabilities

The issue of data privacy is intricate and delicate, particularly in relation to the acquisition of user data through diverse technological means. Privacy considerations are important in using wearable sensor bracelets for public health surveillance during events such as hajj. This is primarily due to the acquired data's sensitive nature and the wide range of users participating in these activities. When individuals' data pertaining to their health and movements are collected, it is natural for them to express concerns regarding the entities that possess access to these data and the potential applications they may be subjected to:

> *Well, the issue of wearables usually and all digital solutions are human rights.* [Participant 2; health care provider and public health specialist]

The interviewees emphasized the difficulties faced while implementing the hajj smart bracelet initiative (NUSUK), specifically focusing on data protection concerns. They underscored the significance of acquiring informed consent for data collection by technological means, such as health data and location information:

> *Regarding data privacy, no data can be collected without an individual consent. Tracking location and gathering medical or health data as you know are considered private data which fall under the General Data Protection Regulation (GDPR).* [Participant 4; communication engineering and AI]

Participants expressed certain reservations about the usefulness of wearable technology. One of the difficulties that participants expressed concerns about was the efficacy of wearable devices. This substantially enhances individuals' confidence in the deployment of wearable sensor bracelets for disease detection during the hajj pilgrimage. In alternative terms, participants emphasized the necessity for these technologies to possess a high level of efficacy to foster a sense of assurance in their use:

> *There are lots of challenges. First, the technology part. How effective is it, how does it capture data, and is it durable? Is it accurate? I think we still can't say that the wearables are ready from the technological side.* [Participant 3; health care provider and public health specialist]

Despite the notable advancements in the network infrastructure built in the Kingdom of Saudi Arabia, participants expressed concerns regarding the adequacy of connectivity strength necessary to effectively manage the vast number of devices used during hajj and the amount of data transmitted:

> *What if we have 2 million devices in an area with weak connections? Do you think we can manage it? Still early.* [Participant 3; health care provider and public health specialist]

Key Considerations in the Adoption of Wearable Sensor Bracelets for Disease Detection During Hajj

Overview

The motives of both consumers and providers significantly impact the successful integration of technology in any domain. In implementing wearable sensor bracelets during the hajj pilgrimage, it is imperative to thoroughly examine and comprehend the motives of all stakeholders and users concerning their acceptance and adoption of this technology. Participants identified several factors that should be considered to encourage stakeholders.

Stakeholder Motivation

Meeting Needs: Integrating Pilgrims' Health Records for Improved Data Accessibility and Enhanced Data Availability

Stakeholders are primarily driven to integrate technology into their strategic plans when it caters to their distinct requirements and eliminates the obstacles they have experienced. Furthermore, stakeholders tend to adopt a technology when it effectively mitigates current challenges and when the characteristics and capabilities of the device correspond to their specific needs. The motivation behind this behavior arises from a comprehensive understanding of how the technology may optimize their activities and contribute to their overarching goals.

The availability of health data, especially with the heterogeneity of languages of pilgrims attending hajj, is of tremendous value to health care professionals as it helps them determine the individual requirements of participants (hajj attendees) and make required preparations. Implementing a consolidated health data system would significantly streamline the identification and understanding of the medical background and requirements of individuals regardless of the health care facilities they visit within the holy sites and Makkah hospitals. Integrating systems would facilitate uninterrupted access to participants' records and optimize the delivery of critical health care services:

> *Ministry of Health needs to have a complete EHR for the pilgrim, they would prefer to have reliable and accurate health data, so data availability.* [Participant 4; communication engineering and AI]

> *Today, I need basic information. Wearables for public health are different from normal bracelets. For the normal bracelet, I need major clinical data. For example, I need to know in advance the conditions that the patient (pilgrim) has, such as if he has kidney failure, if she is pregnant and so on, so I get ready to serve them. I do not want to be surprised; I want to be ready.* [Participant 2; health care provider and public health specialist]

Building Trust in Device Capabilities and Features

From the perspective of health care service providers, participants emphasized the key features that must be incorporated into wearable sensor bracelets to instill faith in and reliance on these devices, hence minimizing the necessity for additional evaluation. The accuracy of the data collected by these devices is a crucial factor. Although health care providers

XSL•FO
RenderX

may choose to perform supplementary evaluations such as laboratory tests, it is imperative that the wearable device delivers data of utmost precision. The significance of accuracy cannot be overstated as it guarantees the reliability of the obtained information and establishes a robust basis for further evaluations:

> *If you look at Apple Fit watch or any other Fitbit watch, the heart rate increases if a person makes any physical effort. Sometimes even if you are sitting, your heart rate could be irregular.* [Participant 1; systems engineering, business, and project management]

In addition to recognizing the significance of trust and the attributes of the wearable sensor bracelet, participants underscored their preference for specific aspects of wearable technologies. A further motive emphasized by the study participants pertained to a particular attribute of the device, namely, its capacity to both transmit and receive data. Participants emphasized the significance of bidirectional data connection as it enables service providers to gather data from the wearables and monitor the bracelets themselves. This attribute was perceived as highly advantageous in terms of facilitating efficient surveillance and guaranteeing the integrity of the sensor bracelets, hence preventing any unauthorized tampering or substitution:

> *I am not a technology person, but I imagine that if we are using such bracelet, I would prefer that it send me data, and I could pull data from it as well. So, push and pull material for data. In addition, the activation and deactivation are central. The user shouldn't be able to activate or deactivate something. I need to know that someone has cut it, removed it, given to someone else or exchanged, so I can track people.* [Participant 2; health care provider and public health specialist]

Cost Reduction Through Private Sponsorship

Considering the cost of the bracelet is of utmost importance as participants emphasized the need to minimize the financial load on hajj authorities. The consensus among participants was that developing or deploying such a device within the context of hajj must strike a delicate equilibrium between cost-effectiveness and the essential functionalities it provides:

> *Honestly, the operation model for this whole business should be provided by the private sector, not the government.* [Participant 1; systems engineering, business, and project management]

Interest of Entities Involved

A significant aspect highlighted by participants was the identification of the interests of the entities involved in the project and assessing the value derived from integrating the technology into their road map development. Among the many organizations participating in the hajj season, such as the Ministry of Health, Ministry of Hajj and Umrah, Ministry of Interior, and others, a number of duties and responsibilities were highlighted:

> *The Ministry of Hajj and Umrah focuses on the quality of services, which is why we focus on measuring satisfaction with crowd management, movements, and safety. We do not focus on diseases and infectious diseases because it is different from the scope of work we are looking for. We are not wasting our effort and time on technologies of that interest. It is not our specialist as a ministry, but the Ministry of Health are working on this file.* [Participant 1; systems engineering, business, and project management]

User Motivation: Promoting the Device for Pilgrim Safety

The significance of encouraging pilgrims to adopt wearable sensor bracelets actively was emphasized by the participants. They stressed the importance of communicating a message that revolves around the person's interest in and reasons for using the technology, especially in light of their safety. It was widely agreed upon that presenting the use of wearables during hajj as a matter of personal preference closely linked to the individual's welfare enhanced the probability of acceptance among pilgrims. The participants expressed reservations about adopting methods that could be interpreted as invasive or limiting, emphasizing the importance of presenting wearables as instruments for individual safety and autonomy:

> *I think you can work it out with the Ministry of Hajj and Umrah, and the Hajj missions by agreeing that the wearable's objective is user health and user safety. They should propose it from the health and safety aspect, so people accept wearing it.* [Participant 2; health care provider and public health specialist]]

Recommendations for Wearable Sensor Bracelets for Disease Detection During Hajj

Interviewees underlined and provided valuable insights into many prerequisites that must be considered to successfully and efficiently implement wearable sensor bracelets for disease detection during hajj.

Comprehensive Infrastructure

The importance of establishing a resilient and all-encompassing infrastructure to effectively handle the large number of devices deployed in the vicinity of the holy sites of Makkah and enable seamless data transmission was emphasized as a crucial component. Moreover, the emphasis was placed on the many components and characteristics of the devices that could efficiently carry out the functions of disease detection and tracking throughout the hard period of hajj considering the large number of participants:

> *An important variable in the success of smart cities is first the infrastructure.* [Participant 4; communication engineering and AI]

Device Characteristics and Features

Strength, Type, and Independency of Connectivity

The current state of connection in holy places was underscored as being in its nascent phase notwithstanding the progress made

XSL•FO
RenderX

in 4G and 5G technologies and the deployment of IoT infrastructure:

> *In previous years, the infrastructure of internet connectivity for 4G and 5G improved remarkably. Connectivity needs to be strong around all holy sites and areas. Not all areas were 100% covered, some were covered under a certain company, and others were served under another coverage company. In addition, the large number of people in a certain area using that network could weaken the connectivity.* [Participant 4; communication engineering and AI]

A cost-effective communication infrastructure was suggested as a means to facilitate data transmission for devices and guarantee equitable service for the substantial number of pilgrims. The proposal entails offering complimentary access to this form of connectivity while considering the economic strain on both the organizers and the users:

> *There is the cost of the device itself and the cost of communication. What I mean by communication is your network package or availability of Wi-Fi services.* [Participant 4; communication engineering and AI]

Emphasizing the necessity for data transmission autonomy, it was critical that wearable sensor bracelets possess the capability to transmit data autonomously without dependence on an intermediary device such as a mobile phone. Autonomy is attained by the use of different forms of connectivity, such as ultrawide band:

> *We usually recommend that when using a smart watch not to connect it to a mobile phone, rather to a sim card or to any type of connectivity such as the UWB where data can be sent whenever it gets signal form a sensor, a device, or any connectivity.* [Participant 4; communication engineering and AI]

Accuracy and Length of Communication Range

Furthermore, the strong reliability of Bluetooth was highlighted, specifically in the context of short-range communication. Nevertheless, it was recognized that Bluetooth possesses certain restrictions, necessitating specific intervals for optimal functionality when operating over particular distances:

> *The advantage of Bluetooth is the strong accuracy. However, one of the limitations is that every 50-100 m you need an interval. The UWB you can reach up to 1000 meters, so less sensors.* [Participant 4; communication engineering and AI]

Apart from the various characteristics of the wearable sensor bracelet that were previously discussed, with a focus on accuracy and efficiency in wearable devices for disease detection during hajj, technologists emphasized the importance of battery life as a critical factor to be taken into account. The importance of selecting an appropriate connectivity option that aligns with the specific circumstances was underscored as it helps reduce excessive battery use and enables consistent and efficient use of these devices within the challenging pilgrimage setting:

> *Battery life as I mentioned is a factor you need to consider. How many days would you like the bracelet to last for. Nowadays companies manufacture devices based on your need, do you want it to last a day more, a week or what.* [Participant 4; communication engineering and AI]

Business Model: Selection of an Appropriate Use Case (a Methodology Used in System Analysis to Identify, Clarify, and Organize System Requirements)

Identifying the right and appropriate use case for the technology was highlighted based on user interest and needs and cost to facilitate the successful integration of new technology into the hajj system:

> *We need to know the appropriate and right use cases for technology usage and implementation. It needs to be clear.* [Participant 4; communication engineering and AI]

Participants emphasized the significance of rendering the use of wearables a personal decision, striking a balance between the device's advantages and the costs involved. The rental model, individual activation, and targeted distribution were identified as crucial elements in attaining a more customized and economically efficient strategy for deploying wearables during the hajj pilgrimage:

> *A person could personally benefit from this wearable if he/she feels they have a critical condition and needs this service; it can be requested as a special request. Anyone can activate it by connecting the watch with contact numbers in an emergency. The government does not need to provide this service, especially since the watch cost is very high. We are talking about 1000 Saudi Riyal [US $266.60] for one watch. Honestly, the value for the cost was not achieved in the project.* [Participant 1; systems engineering, business, and project management]

Effective Staff Training for Wearable Implementation and Distribution

Staff training was identified as a crucial component for successfully implementing wearables. The tasks encompassed managing a substantial volume of data and guaranteeing precise allocation of the wearable devices. The hajj event yields a substantial amount of data, necessitating the provision of sufficient training to staff to handle, analyze, and understand this information effectively:

> *Another thing we need to think about. When I receive a large amount of data, and I am not ready for it, how can I deal with it? Last Hajj, we suddenly received a large amount of data from other departments it was very detailed/360, and we had no idea how to deal with them. We didn't expect it.* [Participant 3; health care provider and public health specialist]

> *Another issue we need to consider is how we will guarantee the distribution of the device. You know the food itself; we have an issue with it with the campaigns. Think about this scenario. I am receiving pilgrims from various countries, and I have the*

XSL•FO
RenderX

> *bracelet to give to the campaign manager or staff that were not trained enough to match the bracelet with the right person. I can't guarantee this; most likely, there will be a mix-up.* [Participant 3; health care provider and public health specialist]

The Need for Organic System Harmony

Participants highlighted the significance of achieving a smooth and natural incorporation of wearable devices into the system to facilitate prospective expansion. This sentiment agrees with the perspective expressed by a participant who emphasized the importance of adopting an organic approach and considering integration as a complete system. The importance of careful deliberation was emphasized by the participant, stressing the need for alignment between decision makers and persons who wear the bracelets. The primary focus lies in establishing a cohesive, interrelated framework that promotes user-friendly functionality, comprehension, and expansion:

> *These things need to be organic; it is a system. We need to think about the person taking the decision, the person wearing the bracelet. Everything need to be harmonised and work together. We all need to know how to deal with it and there need to be built on organic growth.* [Participant 3; health care provider and public health specialist]

Opportunities and Advantages of Wearable Sensor Bracelets During Hajj

Multidisciplinary Features and Distinctiveness Among Other Digital Tools With AI Integration

Participants provided insights into the possible benefits of incorporating wearables, highlighting the numerous opportunities that arise from the varied range of features offered by these devices. They emphasized the potential for transformation that is enabled by improvements in AI capabilities, which play a vital role in developing and improving these devices. The perspectives shared by the participants shed light on a favorable view on the distinctive attributes of wearable devices and the potential synergies they offer with modern AI technologies. These observations indicate a wide range of opportunities for further advancements in this field:

> *Yes, a Wearable sensor bracelet is a good device. Can we use it? The answer is yes. AI will take a very huge platform in public health.* [Participant 2; health care provider and public health specialist]

Participants underscored the unique features of wearable sensing devices compared to other gadgets such as mobile phones. The participants emphasized that wearable devices can assess a wider array of physiological indicators in contrast to mobile devices. This distinctive capacity of wearable sensors makes them valuable instruments for gathering a complete array of health-related data, surpassing typical mobile devices' capabilities:

> *So, both devices; smart watches and mobile phone can use Bluetooth to track and trace infected individuals and people who have been in contact with. However, Smart watches have an advance of measuring vital signs.* [Participant 4; communication engineering and AI]

Gradual Integration of Wearables During Hajj

Participants emphasized the importance of gradually integrating wearable devices as public health tools for effective implementation into the hajj system:

> *...of course. Wearables started to get large space in healthcare. Yes, we can use wearables, but I wouldn't use them for public health at first. It could start with healthcare tackers, so I use the wearable, which can give signals to the healthcare tacker first. Each group of pilgrims have their own identifier (healthcare tacker) or medical mission, so it gives a warning to the medical mission, which gives him a summary of number of cases for example 4 cases have X without a diagnosis just based on the wearable. Then, as a public health department, I get notified and decide what actions need to be taken, such as checking A, B, & C to ensure it is X or Y.* [Participant 2; health care provider and public health specialist]

A Brief Summary of the Findings From the Second Group

The second group of participants cited their duty to serve the guests of God during their hajj journey to Makkah. They discussed the many types of *tawafa* organizations they belonged to (a service provider for external pilgrims) and the role of each one. *Tawafa* organizations are classified into 4 distinct groups. The first group consists of 6 organizations responsible for assisting pilgrims in Makkah and the holy sites that are segregated into subgroups according to the countries and continents from which they originate. These six organizations are for pilgrims coming from (1) Iran; (2) South Asia; (3) Turkey, Europe, Australia, and America; (4) Arabic countries; (5) African non-Arabic countries; and (6) Southeast Asia. The second group consists of 1 organization responsible for providing water to pilgrims at the Sacred House. The third group consists of an organization that is licensed to receive and dispatch pilgrims arriving via air, sea, and land ports, facilitating and arranging their departure protocols. The final group is an organization that is licensed to provide services to pilgrims visiting the sacred mosque located in Madinah city.

Several prevalent health concerns among the pilgrims they assist were brought up by the participants. The most common health problems associated with the hajj event include heat-related illnesses such as heat stroke, heart diseases, and infectious diseases. The importance of medical missions associated with each group of pilgrims was also underscored.

Participants commented on a common issue related to pilgrims' lack of adherence to specific preventive measures designed by hajj authorities to safeguard their well-being, such as hygiene and stampedes in Jamaraat Bridge. In addition, they highlighted that the most holy site among Mina, Muzdalifah, Mount Arafat, and the Holy Mosque where the crowd density becomes high is Mina.

XSL•FO
RenderX

Participants discussed the technological improvements introduced by hajj authorities, such as mobile apps and smart cards, which provide assistance and information to pilgrims from diverse backgrounds. Nevertheless, the participants also emphasized the difficulties of integrating these technologies on a broad scale during hajj. They stressed the importance of conducting more educational campaigns to assist pilgrims in comprehending and proficiently using these technologies as well as understanding the reasons behind their deployment.

Discussion

Principal Findings

Overview

In this study, we explored participant perceptions and experiences of using wearable sensor bracelets during the hajj pilgrimage. The primary objective was to investigate the potential advantages and obstacles associated with wearables monitoring pilgrims' well-being and detecting diseases. The main findings and themes that emerged from the interviews will be analyzed and summarized in this section to provide insights into the opportunities and challenges involved in incorporating wearable technology (sensor bracelets) into mass gatherings, including the hajj pilgrimage.

Wearable sensor bracelets have the capability to monitor the health of pilgrims and detect emerging infectious diseases during the hajj pilgrimage. However, achieving these capacities depends on properly tackling existing problems and meticulously evaluating multiple elements.

The diverse range of valuable information obtained from the interviews with participants reveals a variety of obstacles that hinder the seamless use of wearable sensor bracelets for detecting infectious diseases during the hajj pilgrimage. However, in the midst of these difficulties, a variety of factors for achieving successful implementation become apparent, uncovering hidden possibilities that these devices offer. Charting a path for future deployments in the dynamic landscape of mass gatherings such as hajj, we will unravel the multifaceted nature of these challenges, explore the promising vistas these wearables offer, and present recommendations highlighted by our interviewees' insightful perspectives as we navigate the discussion.

Challenges

Numerous challenges associated with implementing wearable sensor bracelets during hajj were highlighted. The primary obstacle underlined was the anticipated opposition to technology, which may potentially emerge from providers and users of wearable devices. Participants observed that trust in technology, data privacy, information security, and human rights concerns could influence this reluctance.

The opposition of providers is affected by both the operational and financial challenges associated with implementing wearable technology into the hajj system. Although data are an essential element in detecting diseases, staff training in data collection and analysis is required to handle the anticipated volume of data produced by the devices. Participants voiced reservations about the limited and inaccurate pre-hajj data on pilgrims' health. Furthermore, there was concern that, despite the dedicated efforts put into the implementation, the results may not correspond with the anticipated or intended value.

The issue of financial burden is not limited to providers; it affects pilgrims as well. Due to the financial obligations that pilgrims encounter during their hajj journey [36], the cost could constitute a barrier if the services were to be included in their hajj packages. The cost-related reluctance may affect users' willingness to use the device during their hajj pilgrimage. The use of tracking and tracing methods to monitor pilgrims throughout their journey may impede their compliance and reception due to variations in their backgrounds, beliefs, ages, approaches to conducting hajj rituals, level of trust in technology, and literacy levels [3,37].

The challenges brought up by stakeholders are consistent with those identified by Yeung et al [3], especially regarding the end users in our research, who are pilgrims. The stakeholders' worries about potential hurdles from users align with the findings of Yeung et al [3], suggesting that user resistance may stem from factors such as privacy concerns, lack of motivation, and technological fear.

The viewpoints conveyed by stakeholders and the difficulties they mentioned pertain to service providers and users and align with the concept of effort expectancy in the UTAUT model. This connection is particularly visible in the financial and operational efforts required to integrate wearables during hajj.

Data privacy is paramount while collecting user data through wearable devices in any domain. Collecting real-time data is essential for public health surveillance to diagnose infectious diseases promptly and implement appropriate interventions. Gathering sensitive data poses difficulties, especially when considering human rights and data privacy agreements. The use of wearable public health devices, particularly in mass gatherings, may be limited due to considerable challenges arising from privacy considerations.

The issue of data privacy, which is of utmost importance for wearable devices in the context of mass gatherings, is intricately linked to the difficulty of amassing data with user consent. Participants highlighted that addressing privacy concerns and securing user consent presented substantial obstacles, impeding the success of the previous smart bracelet initiative. Azodo et al [6], who discussed information privacy, data sharing, and autonomy, brought attention to this layer of complication brought about by using wearables in health care.

The stakeholders' willingness to adopt wearable sensor bracelets for disease detection and tracking during hajj was strongly linked to the idea of performance expectancy in the UTAUT model. The stakeholders' trust in the technology was greatly impacted by the device's performance and its capacity to meet fundamental functions. This is consistent with the concept of performance expectancy, which highlights the perceived utility of the technology. Furthermore, assessing the device's utility, including factors such as dependability, precision, and compatibility, highlights the significance of technology characteristics in the TTF model. These functionalities play a

XSL•FO
RenderX

vital role in shaping stakeholders' perceptions and adoption of the technology within the specific context of hajj.

Considerations for Promoting Acceptance and Adoption

Overcoming current obstacles that hinder the adoption and use of wearable sensor bracelets during hajj is vital. Participants emphasized important aspects that play a significant role in promoting the success of the implementation.

The interviews demonstrated that stakeholders believe that user motivation is crucial in shaping their acceptance of the device. The interviews covered several aspects that contribute to promoting acceptance. Improving data availability to enhance accessibility is a primary driver for addressing stakeholders' needs. Furthermore, considering the interests of the parties involved in managing hajj ensures the incorporation of technology into their plans for future development in line with their particular requirements and strategies.

The wearables' functionality and capacity to collect accurate, dependable, and real-time data were given thoughtful consideration, addressing issues brought up in the literature on the difficulty of guaranteeing wearable accuracy [11,27,38]. Increasing stakeholder trust in technology becomes increasingly challenging when it comes to the known difficulties in ensuring wearable accuracy. The need to ensure essential device functions was stressed by participants as a means of promoting wearable trust and dependence. These qualities are essential for overcoming wearable accuracy difficulties, building stakeholder trust, and influencing their choice to use technology for disease detection and hajj surveillance.

The emphasis on accuracy and different device characteristics is consistent with the concept of performance expectancy in the UTAUT. Stakeholders emphasized the necessity for features that qualify the device for successful disease diagnosis and saw performance as critical to the technology. In addition to being in line with performance expectations, addressing issues with accuracy and key features is crucial for shaping stakeholders' opinions of the technology.

There was a unanimous consensus on the necessity of incorporating private sponsorship as a business model for providing the device. By leveraging public-private partnerships, Aina et al [24] propose resolving the obstacles linked to smartification and digitalization in urban service financing. As suggested by stakeholders, this strategy reduces the financial and operational burden on hajj stakeholders related to implementing and controlling wearable devices for disease diagnosis and tracking. The fact of private sponsorship coincides with the UTAUT's *facilitating condition* construct, offering a favorable setting and support for the effective integration of wearable technology during hajj.

During the interviews, participants emphasized safety as a crucial factor that significantly affects the adherence to and acceptance, adoption, and use of wearable sensor bracelets by pilgrims. Having sufficient user motivation and willingness to acquire and use the wearable device is essential for acceptance [39]. Although Saudi Arabia has made significant efforts to guarantee the safety of pilgrims and control crowds to minimize the occurrence of stampedes and the spread of diseases, the difficulty remains [40].

It is widely agreed that presenting the use of wearables during hajj as a personal choice strongly linked to individual well-being greatly increases the chances of pilgrims accepting it. This is consistent with the participants' emphasis on the significance of selecting the appropriate business model for implementing such devices. Participants also voiced apprehensions over techniques that can be interpreted as intrusive or constraining, emphasizing the significance of portraying wearables as instruments for personal safety and independence.

Recommendations

Several crucial recommendations were proposed to tackle the limitations of incorporating wearable sensor bracelets for disease detection during hajj. A robust IoT infrastructure emerged as a crucial element in successfully deploying the device. The type of network connectivity and its strength were highlighted as key considerations, emphasizing the capacity to handle the volume of users and data transmission coupled with the availability of free connectivity for all pilgrims.

Participants emphasized the importance of implementing wearables for public health surveillance in a gradual and systematic manner. The gradual integration focuses on and enables smooth integration by initially guaranteeing the compatibility of the technology with the surveillance system used during hajj. Compatibility is a vital element within the TTF model that describes the features of the technology.

Participants with technology experience provided valuable comments that highlighted the need to address data availability difficulties faced by hajj stakeholders. They emphasized the importance of establishing a strong, accurate health data system to resolve these challenges. The proposed solution entails using blockchain technology to establish customized electronic health records specifically designed for pilgrims. This approach is in keeping with the recommendations presented in the research by Yeung et al [3], emphasizing the capacity of blockchain technology to tackle issues concerning the secure and decentralized storage and transmission of digital data in health care infrastructure.

Participants put forth a distinctive suggestion concerning carefully considering and selecting a suitable business model to ensure the seamless integration and use of wearable sensors during the hajj ritual. Walsh [35] states that smartwatch applications and use cases differ in many fields, offering various services and functions. The wide range of differences among pilgrims during hajj and *umrah* offers a chance to create applications that can improve their overall experience. A novel proposal entails implementing a rental system for individuals who wish to use such devices during their hajj pilgrimage either for overall well-being or for specific health concerns.

Opportunities

Implementing wearable sensor bracelets for infectious disease detection during hajj comes with difficulty, yet the opportunities for this technology's use are considerable.

XSL•FO
RenderX

The use of AI greatly enhances the potential of wearables in public health surveillance, representing a promising advancement. Wearable devices, which include communication modules and networking capabilities, possess modern hardware technologies. These devices can supply vital data that power AI procedures [41]. This technological advancement facilitates the uninterrupted transmission of up-to-date information without requiring direct engagement from health care staff, thereby establishing wearables as a potent and nonintrusive instrument for augmenting public health surveillance.

The capacity of wearable sensor bracelets to evaluate a wider range of physiological indications gives them a clear edge over other technologies such as mobile devices. Their ability to adapt and be used in various ways makes them very valuable in the health care industry.

Wearable sensor bracelets greatly enhance public health surveillance during hajj by carrying out functions such as disease identification, health monitoring, tracking, and tracing. They stand out from other devices because of their unique characteristics, which provide real-time health tracking, detection, and monitoring. This helps with the efficient administration of public health during hajj. This is consistent with the broader idea of the TTF model, in which wearables are customized to carry out certain duties and operations that are essential for effective public health management in the special circumstances of hajj.

The degree to which wearable sensor bracelets can fulfill the unique requirements of hajj stakeholders is one of the key elements determining the desire to use them for disease detection during hajj. The link between hajj stakeholders' desire to use wearables for disease detection and their perception of the devices' capabilities emphasizes the importance of performance expectancy as a critical element of the UTAUT, which affects providers' desire to embrace the technology. Furthermore, achieving successful implementation relies on facilitating conditions, underscoring the need for a supportive environment for efficient adoption.

Furthermore, the *technology characteristics* component, as defined by the TTF model, has a crucial influence on the intention to use. The characteristics and qualities of the technology have a substantial influence on providers' inclination to embrace it. The relationship between the UTAUT and TTF model highlights how users' intention to adopt and use wearable sensor bracelets for disease detection during hajj is influenced by their perception of the technology's performance, facilitating conditions, and characteristics.

Strengths, Limitations, and Future Work

Strengths

This study is the first to address the opportunities and limitations of wearable sensor bracelets for hajj. It provides a thorough understanding of these issues from the perspective of key stakeholders. The thoroughness of the investigation provides useful insights into the complex dynamics surrounding the implementation of these devices.

Moreover, the combined use of the UTAUT and TTF models proved advantageous in comprehending stakeholder viewpoints comprehensively. This method enabled a thorough investigation of the possible uses of wearable devices for detecting diseases, specifically during the hajj pilgrimage.

Limitations

One potential limitation is that the range of stakeholders from different disciplines that were included in the interviews was not very diverse. Broadening the pool of respondents would enhance our understanding of the system and the interconnectedness between its components.

In addition, a possible drawback arises from the involvement of a single researcher in the coding and analytic process. This limited worldview may lead to the potential for misjudging participants' perspectives. Adding numerous analysts or external assessments could have enhanced the coding process by incorporating varied perspectives and increasing its rigor. Considering the expansion of the analytical team in future studies may help limit any biases in interpreting participant replies.

Future Work

To optimize the efficient deployment of wearable sensor bracelets, further studies should prioritize obtaining perspectives from end users. Gaining insights into their perspectives and embracing their adoption of the device is essential for formulating tactics that align with user requirements and preferences. This would additionally enhance the successful integration and use of wearable sensor technology in pertinent contexts.

Conclusions

The significant contribution of this work lies in its thorough investigation of crucial factors in implementing wearable sensor bracelets for the detection of infectious diseases during the hajj event. Our study used the integration of the TTF and UTAUT frameworks to better understand the adoption of wearable sensors by stakeholders. The purpose of this integration was to investigate the perceived likelihood of technology adoption during hajj and the stakeholders' viewpoints on the compatibility between the task of disease detection and the technology used. Our aim was to integrate these theories to offer a thorough comprehension of the aspects that impact the acceptance and efficacy of wearable sensors in this particular setting. This study underscores the need to resolve crucial issues to guarantee the efficient implementation and use of wearable sensors in this specific case. Our results highlight multiple problems pertaining to suppliers, users, and technology, which impose a cost on hajj stakeholders. Significantly, the specific attributes and functionalities of the technology play a vital role in building confidence among providers when it comes to using wearables during hajj. This study highlights the importance of having the ability to access pre-hajj health data for health care providers to gain a more comprehensive understanding of pilgrims' requirements. Although there have been some breakthroughs, it is crucial to further develop the current infrastructure to support wearables fully. Wearables stand out from other devices due to their exceptional capacity to gather vital signs, and their

XSL•FO
RenderX

potential is further amplified by encouraging outcomes from AI algorithms. This study suggests choosing a suitable business model that is customized to meet the specific demands of the users to ensure a successful deployment. A crucial element of this implementation entails creating a complete infrastructure marked by a strong and reliable network capable of prompt data transfer and accommodating a large user population. Furthermore, the unique characteristics of the devices should be in line with cost-effective advantages to guarantee a sustainable implementation.

However, deploying wearable bracelets for detecting infectious diseases during hajj may face difficulties. Therefore, it is important to conduct pilot tests and thorough evaluations to ensure their usefulness before implementing them at a large scale. This prudent approach recognizes the intricacy of the hajj environment and highlights the importance of conducting evidence-based evaluations to determine the feasibility and effectiveness of the wearable technology. Notwithstanding these difficulties, there is a pervasive conviction regarding the innate capacity of these wearable devices, cultivating optimism regarding their ultimate and widespread use in the context of hajj.

Acknowledgments

This work was conducted as part of a doctoral study sponsored by the Ministry of Education, Saudi Arabia.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Example topic guide for the interviews.
[DOCX File , 33 KB - formative_v9i1e60484_app1.docx]

References

1. Sahu KS, Majowicz SE, Dubin JA, Morita PP. NextGen public health surveillance and the internet of things (IoT). Front Public Health 2021 Dec 3;9:756675 [FREE Full text] [doi: 10.3389/fpubh.2021.756675] [Medline: 34926381]
2. Hirten RP, Danieletto M, Tomalin L, Choi KH, Zweig M, Golden E, et al. Use of physiological data from a wearable device to identify SARS-CoV-2 infection and symptoms and predict COVID-19 diagnosis: observational study. J Med Internet Res 2021 Feb 22;23(2):e26107 [FREE Full text] [doi: 10.2196/26107] [Medline: 33529156]
3. Yeung AW, Torkamani A, Butte AJ, Glicksberg BS, Schuller B, Rodriguez B, et al. The promise of digital healthcare technologies. Front Public Health 2023 Sep 26;11:1196596 [FREE Full text] [doi: 10.3389/fpubh.2023.1196596] [Medline: 37822534]
4. Hayati N, Ramli K, Suryanegara M, Salman M. An internet of things (IoT) reference model for an infectious disease active digital surveillance system. Int J Adv Comput Sci Appl 2021;12(9). [doi: 10.14569/ijacsa.2021.0120956]
5. Meraj M, Alvi SA, Quasim MT, Haidar SW. A critical review of detection and prediction of infectious disease using IOT sensors. In: Proceedings of the Second International Conference on Electronics and Sustainable Communication Systems. 2021 Presented at: ICESC 2021; August 04-06, 2021; Coimbatore, India. [doi: 10.1109/icesc51422.2021.9532992]
6. Azodo I, Williams R, Sheikh A, Cresswell K. Opportunities and challenges surrounding the use of data from wearable sensor devices in health care: qualitative interview study. J Med Internet Res 2020 Oct 22;22(10):e19542 [FREE Full text] [doi: 10.2196/19542] [Medline: 33090107]
7. Tavakoli M, Carriere J, Torabi A. Robotics, smart wearable technologies, and autonomous intelligent systems for healthcare during the COVID‐19 pandemic: an analysis of the state of the art and future vision. Adv Intell Syst 2020 May 05;2(7):2000071. [doi: 10.1002/aisy.202000071]
8. Mamdiwar SD, R A, Shakruwala Z, Chadha U, Srinivasan K, Chang CY. Recent advances on IoT-assisted wearable sensor systems for healthcare monitoring. Biosensors (Basel) 2021 Oct 04;11(10):372 [FREE Full text] [doi: 10.3390/bios11100372] [Medline: 34677328]
9. Mishra T, Wang M, Metwally AA, Bogu GK, Brooks AW, Bahmani A, et al. Pre-symptomatic detection of COVID-19 from smartwatch data. Nat Biomed Eng 2020 Dec 18;4(12):1208-1220 [FREE Full text] [doi: 10.1038/s41551-020-00640-6] [Medline: 33208926]
10. Hänsel K, Wilde N, Haddadi H, Alomainy A. Challenges with current wearable technology in monitoring health data and providing positive behavioural support. In: Proceedings of the 5th EAI International Conference on Wireless Mobile Communication and Healthcare. London, UK: ICST; 2015:158-161.

XSL•FO
RenderX

11. Kamišalić A, Fister IJ, Turkanović M, Karakatič S. Sensors and functionalities of non-invasive wrist-wearable devices: a review. Sensors (Basel) 2018 May 25;18(6):1714 [FREE Full text] [doi: 10.3390/s18061714] [Medline: 29799504]
12. Leu FY, Ko CY, You I, Choo KK, Ho CL. A smartphone-based wearable sensors for monitoring real-time physiological data. Comput Electr Eng 2018 Jan;65:376-392. [doi: 10.1016/j.compeleceng.2017.06.031]
13. Sawadi HK, Al-Zahrani IO, Al-Ghoraibi AF, Al-Ghamdi AM, Bati GF. Towards Hajj and Umrah digital transformation using smartwatches. Islamic Univ J Appl Sci 2023 Jul;V(I):91-107.
14. Mohammadzadeh N, Gholamzadeh M, Saeedi S, Rezayi S. The application of wearable smart sensors for monitoring the vital signs of patients in epidemics: a systematic literature review. J Ambient Intell Humaniz Comput 2023 Nov 13;14(5):6027-6041 [FREE Full text] [doi: 10.1007/s12652-020-02656-x] [Medline: 33224305]
15. Creaser AV, Hall J, Costa S, Bingham DD, Clemes SA. Exploring families' acceptance of wearable activity trackers: a mixed-methods study. Int J Environ Res Public Health 2022 Mar 15;19(6):3472 [FREE Full text] [doi: 10.3390/ijerph19063472] [Medline: 35329166]
16. Mitratza M, Goodale BM, Shagadatova A, Kovacevic V, van de Wijgert J, Brakenhoff TB, et al. The performance of wearable sensors in the detection of SARS-CoV-2 infection: a systematic review. Lancet Digit Health 2022 May;4(5):e370-e383 [FREE Full text] [doi: 10.1016/S2589-7500(22)00019-X] [Medline: 35461692]
17. Al-Shaery AM, Aljassmi H, Ahmed SG, Farooqi NS, Al-Hawsawi AN, Moussa M, et al. Real-time pilgrims management using wearable physiological sensors, mobile technology and artificial intelligence. IEEE Access 2022;10:120891-120900. [doi: 10.1109/access.2022.3221771]
18. de Lima LL, Ferreira AL, de Araujo WR. Sensing interfaces: materials for wearable sensors. Encycl Sens Biosens 2023;2:636-649 [FREE Full text] [doi: 10.1016/B978-0-12-822548-6.00127-8]
19. Risch M, Grossmann K, Aeschbacher S, Weideli OC, Kovac M, Pereira F, et al. Investigation of the use of a sensor bracelet for the presymptomatic detection of changes in physiological parameters related to COVID-19: an interim analysis of a prospective cohort study (COVI-GAPP). BMJ Open 2022 Jun 21;12(6):e058274 [FREE Full text] [doi: 10.1136/bmjopen-2021-058274] [Medline: 35728900]
20. Grzesiak E, Bent B, McClain MT, Woods CW, Tsalik EL, Nicholson BP, et al. Assessment of the feasibility of using noninvasive wearable biometric monitoring sensors to detect influenza and the common cold before symptom onset. JAMA Netw Open 2021 Sep 01;4(9):e2128534 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.28534] [Medline: 34586364]
21. Alqahtani AS, Althimiri NA, BinDhim NF. Saudi Hajj pilgrims' preparation and uptake of health preventive measures during Hajj 2017. J Infect Public Health 2019 Nov;12(6):772-776 [FREE Full text] [doi: 10.1016/j.jiph.2019.04.007] [Medline: 31023600]
22. Benkouiten S, Brouqui P, Gautret P. Non-pharmaceutical interventions for the prevention of respiratory tract infections during Hajj pilgrimage. Travel Med Infect Dis 2014 Sep;12(5):429-442 [FREE Full text] [doi: 10.1016/j.tmaid.2014.06.005] [Medline: 24999278]
23. Alqahtani AS, Wiley KE, Tashani M, Willaby HW, Heywood AE, BinDhim NF, et al. Exploring barriers to and facilitators of preventive measures against infectious diseases among Australian Hajj pilgrims: cross-sectional studies before and after Hajj. Int J Infect Dis 2016 Jun;47:53-59 [FREE Full text] [doi: 10.1016/j.ijid.2016.02.005] [Medline: 26875699]
24. Aina YA, Abubakar IR, Almulhim AI, Dano UL, Maghsoodi Tilaki MJ, Dawood SR. Digitalization and smartification of urban services to enhance urban resilience in the post-pandemic era: the case of the pilgrimage city of Makkah. Smart Cities 2023 Aug 11;6(4):1973-1995. [doi: 10.3390/smartcities6040092]
25. Felemban EA, Rehman FU, Biabani SA, Ahmad A, Naseer A, Majid AR, et al. Digital revolution for Hajj crowd management: a technology survey. IEEE Access 2020;8:208583-208609. [doi: 10.1109/access.2020.3037396]
26. Abalkhail AA, Al Amri SM. Saudi Arabia’s management of the Hajj season through artificial intelligence and sustainability. Sustainability 2022 Oct 29;14(21):14142. [doi: 10.3390/su142114142]
27. Mejia C, Ciarlante K, Chheda K. A wearable technology solution and research agenda for housekeeper safety and health. Int J Contemp Hosp Manag 2021 Aug 04;33(10):3223-3255. [doi: 10.1108/ijchm-01-2021-0102]
28. Papagiannidis S. TheoryHub Book. Newcastle upon Tyne, England: Newcastle University; 2022.
29. Marikyan M, Papagiannidis P. Unified theory of acceptance and use of technology. In: TheoryHub Book. Newcastle upon Tyne, UK: Newcastle University; 2022:148-165.
30. Onwuegbuzie AJ, Leech NL. Sampling designs in qualitative research: making the sampling process more public. Qual Report 2015 Jan 15;12(2):238-254 [FREE Full text] [doi: 10.46743/2160-3715/2007.1636]
31. Clarke V, Braun V. Successful Qualitative Research: A Practical Guide for Beginners. Thousand Oaks, CA: SAGE Publications; 2013.
32. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
33. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view. MIS Q 2003;27(3):425-478. [doi: 10.2307/30036540]
34. Goodhue DL, Thompson RL. Task-technology fit and individual performance. MIS Q 1995 Jun;19(2):213-236. [doi: 10.2307/249689]

XSL•FO
RenderX

35. Walsh A. Policy on the ethical involvement of human participants in research. University of Manchester. 2021. URL: https://documents.manchester.ac.uk/display.aspx?DocID=28798 [accessed 2025-01-16]
36. Quaium A, Al-Nabhan NA, Rahaman M, Salim SI, Toha TR, Noor J, et al. Towards associating negative experiences and recommendations reported by Hajj pilgrims in a mass-scale survey. Heliyon 2023 May;9(5):e15486 [FREE Full text] [doi: 10.1016/j.heliyon.2023.e15486] [Medline: 37144197]
37. Taibah H, Arlikatti S, Andrew SA, Maghelal P, DelGrosso B. Health information, attitudes and actions at religious venues: evidence from hajj pilgrims. Int J Disaster Risk Reduct 2020 Dec;51:101886 [FREE Full text] [doi: 10.1016/j.ijdrr.2020.101886] [Medline: 32995254]
38. Baker SB, Xiang W, Atkinson I. Internet of things for smart healthcare: technologies, challenges, and opportunities. IEEE Access 2017;5:26521-26544. [doi: 10.1109/access.2017.2775180]
39. Wen D, Zhang X, Lei J. Consumers' perceived attitudes to wearable devices in health monitoring in China: a survey study. Comput Methods Programs Biomed 2017 Mar;140:131-137. [doi: 10.1016/j.cmpb.2016.12.009] [Medline: 28254069]
40. Yamin M. Secure and healthy Hajj management: a technological overview. Am Acad Scholarly Res J 2015;7(3):195-202 [FREE Full text]
41. Nahavandi D, Alizadehsani R, Khosravi A, Acharya UR. Application of artificial intelligence in wearable devices: opportunities and challenges. Comput Methods Programs Biomed 2022 Jan 30;213:1-19 [FREE Full text] [doi: 10.1016/j.cmpb.2021.106541] [Medline: 34837860]

Abbreviations

AI: artificial intelligence
COREQ: Consolidated Criteria for Reporting Qualitative Research
HRV: heart rate variability
IoT: Internet of Things
TTF: task-technology fit
UTAUT: unified theory of acceptance and use of technology

Edited by A Mavragani; submitted 13.05.24; peer-reviewed by S Pesälä, S Chantziara, H Rashid; comments to author 06.08.24; revised version received 03.09.24; accepted 19.01.25; published 08.04.25.

Please cite as:
Maddah N, Verma A, Ainsworth J
Opportunities and Challenges Surrounding the Use of Wearable Sensor Bracelets for Infectious Disease Detection During Hajj: Qualitative Interview Study
JMIR Form Res 2025;9:e60484
URL: https://formative.jmir.org/2025/1/e60484
doi:10.2196/60484
PMID:

©Noha Maddah, Arpana Verma, John Ainsworth. Originally published in JMIR Formative Research (https://formative.jmir.org), 08.04.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A New Mobile App to Train Attention Processes in People With Traumatic Brain Injury: Logical and Ecological Content Validation Study

Roxanne Laverdière[1], BA; Philip L Jackson[1], PhD; Frédéric Banville[2], PhD

[1]School of Psychology, Université Laval, Quebec, QC, Canada

[2]Department of Health Sciences, Université du Québec à Rimouski, Rimouski, QC, Canada

Corresponding Author:
Frédéric Banville, PhD
Department of Health Sciences
Université du Québec à Rimouski
300, allée des Ursulines
Rimouski, QC, G5L 3A1
Canada
Phone: 1 418 723 1986 ext 1931
Email: frederic_banville@uqar.ca

Abstract

Background: Attention is at the base of more complex cognitive processes, and its deficits can significantly impact safety and health. Attention can be impaired by neurodevelopmental and acquired disorders. One validated theoretical model to explain attention processes and their deficits is the hierarchical model of Sohlberg and Mateer. This model guides intervention development to improve attention following an acquired disorder. Another way to stimulate attention functions is to engage in the daily practice of mindfulness, a multicomponent concept that can be explained by the theoretical model of Baer and colleagues. Mobile apps offer great potential for practicing mindfulness daily as they can easily be used during daily routines, thus facilitating transfer. Laverdière and colleagues have developed such a mobile app called *Focusing*, which is aimed at attention training using mindfulness-inspired attentional exercises. However, this app has not been scientifically validated.

Objective: This research aims to analyze the logical content validity and ecological content validity of the *Focusing* app.

Methods: Logical content validation was performed by 7 experts in neuropsychology and mindfulness. Using an online questionnaire, they determined whether the content of the attention training app exercises is representative of selected constructs, namely the theoretical model of attention by Sohlberg and Mateer and the theoretical model of mindfulness by Baer and colleagues. A focus group was subsequently held with the experts to discuss items that did not reach consensus in order to change or remove them. Ecological content validation was performed with 10 healthy adults. Participants had to explore all sections of the app and assess the usability, relevance, satisfaction, quality, attractiveness, and cognitive load associated with each section of the app, using online questionnaires.

Results: Logical content validation results demonstrated a high content validity index (CVI) of the attention training app. Excellent scores (CVI ≥0.78) in both the attention and mindfulness models were obtained for all exercises in the app, except 2 exercises. One of these exercises was subsequently modified to include expert feedback, and one was removed. Regarding ecological content validation, the results showed that workload, quality, user experience, satisfaction, and relevance of the app were adequate. The Mobile Application Rating Scale questionnaire showed an average quality rating between 3.75/5 (SD 0.41) (objective quality) and 3.65/5 (SD 0.36) (subjective quality), indicating acceptable quality. The mean global attractiveness rating from the AttrakDiff questionnaire was 2.36/3 (SD 0.57), which represents one of the strengths of the app.

Conclusions: Logical and ecological content validation showed that *Focusing* is theoretically valid, with a high level of agreement among experts and healthy participants. This tool can be tested to train attention processes after a neurological insult such as traumatic brain injury.

(JMIR Form Res 2025;9:e64174) doi:10.2196/64174

XSL•FO
RenderX

KEYWORDS

mobile app; attention training; cognitive remediation; mindfulness; psychometric properties; content validation

Introduction

Attention Process Training

Attention is the ability to select relevant stimuli from the environment and inhibit those that are not relevant [1]. This cognitive process makes it possible for people to direct attention to the most salient elements to handle them appropriately. Indeed, attention is a factor of cognitive efficiency that supports memory, perception, and problem-solving [2]. Attention deficits are associated with significant difficulties in instrumental activities of daily living that influence the overall quality of life of individuals whose cognition is compromised [3,4].

Over the past decades, several experts have recommended the use of cognitive remediation, a nonpharmacological approach, for individuals with attention deficits [5].

People who live with attention deficits associated with a condition, such as traumatic brain injury (TBI) or attention-deficit/hyperactivity disorder, may benefit from attention training interventions [5]. This kind of intervention, particularly metacognitive strategy training (ie, anticipate, plan, seek, self-regulate, and evaluate), has been proven effective in neurological populations, such as people with TBI [6]. Cognitive remediation aims to improve cognitive functioning either by training to overcome deficits in functions or by allowing individuals to acquire strategies to best exploit their residual functions and help them engage in their daily activities [7]. Attention Process Training-III (APT-III) [8] is one of the most validated attention training programs for people with TBI. It includes two 45-minute sessions per week for 6 weeks. It includes two 45-minute sessions per week for 6 weeks, comprising visual and auditory exercises aimed at training the 5 attentional components of the model from Sohlberg and Mateer [9]. However, several of the studies addressing the effectiveness of this intervention have significant methodological limitations, including a lack of reliability and validity, and a low number of participants. Moreover, the results of the studies showed improvements among individuals in different attentional tests without any evidence of transfer to functional tasks [6].

According to the results of several studies [10], it is possible to stimulate and optimize attention functions using mindfulness practice. Indeed, mindfulness is a state of consciousness that emerges from deliberately focusing on an object in the present moment [11]. An analysis of the “cognitive mechanics” of mindfulness revealed that attention, working memory, and their different processes are key [1,10]. Indeed, Bishop et al [12] proposed 2 key components of mindfulness: self-regulation of attention (ie, the ability to maintain attention) and orientation to the present moment with curiosity, openness, and acceptance (ie, focusing on internal or external stimuli, inhibition, or activation of certain thoughts). Mindfulness interventions have shown promising results for different populations in terms of attentional capacities, particularly regarding sustained attention and attentional flexibility [13]. For instance, a previous study from our team [14] showed that a mindfulness intervention integrated into an ecological virtual environment was safe, feasible, and acceptable for people with mild TBI. The use of attentional focusing techniques of mindfulness could also be useful for the management of attention difficulties in people who have experienced moderate to severe TBI. Indeed, focusing on the present moment is a way of controlling one’s own attentional processes. As such, attention is considered central to the construct of mindfulness [10]. One limit of this type of intervention is access to a trained guide to accompany individuals, especially for people living in remote areas. It is therefore relevant to elaborate mindfulness interventions accompanied by a remote guide that is accessible from home, regardless of the region.

Although not everyone uses technology or is familiar with technology, mobile devices are ubiquitous in everyday life, and their apps are increasingly being used and perfected, particularly in health domains [15]. Countless mindfulness-based mobile apps have been developed; for instance, a 2021 systematic review identified 605 mindfulness-based mobile apps in European app stores [16]. Several meta-analyses confirmed the effectiveness of mindfulness interventions via mobile apps, particularly for mental health [17-19]. A study also showed that mindfulness interventions via a mobile app have effects comparable to mindfulness interventions in person [20]. These apps contain guided mindfulness exercises inspired by traditional approaches (mindfulness-based stress reduction [MBSR] [11] and mindfulness-based cognitive therapy [MBCT] [21]) that can be performed by individuals at home. However, even if studies show that these apps work, there is a lack of studies in the literature supporting the content and quality of these apps [22-25], and no studies have been conducted on the use of validated mindfulness apps in individuals with TBI. A systematic review in 2019 identified 53 mobile apps dedicated to people with TBI on the Apple App Store and Google Play Store [26]. Among these, there were compensatory measures, and some apps were aimed more at the remediation of cognitive functions, but none were aimed specifically at training attention using mindfulness. Findings from studies on the effectiveness of these apps in improving TBI symptoms are inconsistent. There is very limited research to validate the content of these apps, which poses a risk to users. Indeed, a specific validation is required to use these types of apps, considering that people who have experienced TBI (more specifically moderate or severe TBI) often have a variety of physical and cognitive deficits that can affect their ability to perform a mindfulness program via a mobile app. Physically, they may have motor disorders (paralysis, muscle weakness, and balance disorder), persistent fatigue, or sensory disorders (vision or hearing). Cognitively, deficits may involve attention, memory, language, and executive functions such as problem-solving and decision-making. These limitations may also be accompanied by emotional or behavioral disorders.

To explore the potential of mobile apps in the training of attention via a mindfulness focusing technique, we developed

XSL•FO
RenderX

an app, called *Focusing*, based on the Obesity-Related Behavioral Intervention Trials (ORBIT) intervention development model [27], which aimed at attention process training following the occurrence of neurological disorders such as TBI [28]. *Focusing* is currently available in French. The mobile app consists of mindfulness-inspired attentional focus exercises (Figure 1). The proposed attention training is based on 2 validated theoretical models: the mindfulness model of Baer et al [29] and the theoretical attention framework of Sohlberg and Mateer [9]. *Focusing* is composed of 5 sections (Multimedia Appendix 1). The first section "About" contains an interactive tutorial, a psychoeducational video about attention after TBI with a thought-provoking case example, and explanatory text about mindfulness. The second section "Exercises" contains the 4 attentional focus exercise sessions of the app. Each session includes 4 exercises (the first exercise is the same for each session to promote a state of calm conducive to the exercises of the session). Each session involves a component of attention, ranging from the simplest to the most complex. The third section "Homework" contains a list of homework assignments for each week. It also contains a "Goals" section, allowing the participant to set objectives and to consult with them throughout the week. The fourth section "Ambiences" contains ambient sounds representing nature and is aimed at freely practicing attentional focus in the visual and auditory modalities. The fifth section "Me" contains the participant's profile.

Figure 1. Screen captures of the different sections of the attention process training app (Focusing app) for people with traumatic brain injury.

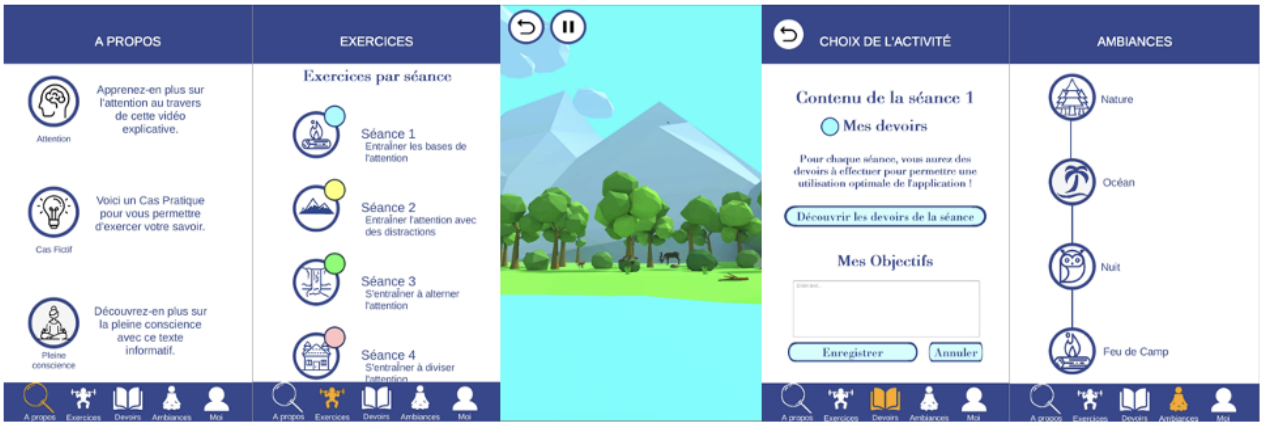

Theoretical Framework

The development of the attention training app was based on the neuropsychological model of attention of Sohlberg and Mateer [9]. This model was developed based on data from populations with brain damage and includes the subcomponents of attention that are frequently affected by TBI. This model has been widely used to explain attention in general, regardless of the neurological disorder. It involves the processes of focused, sustained, selective, alternating, and divided attention. *Focused attention* represents the ability to respond specifically and in a targeted way to a sensory stimulus. *Sustained attention* is the ability to maintain a consistent behavioral response during continuous or repetitive activity. *Selective attention* represents the ability to maintain a given behavioral or cognitive response to stimuli exerting a distracting or competitive effect. *Alternating attention* shifts the focus between tasks with distinct cognitive requirements efficiently and repeatedly. Finally, *divided attention* represents the ability to perform multiple tasks simultaneously.

To integrate this theoretical model into the *Focusing* app, each of the 4 sessions trains one of the components of the model by Sohlberg and Mateer [9]: Session 1, sustained attention; Session 2, selective attention; Session 3, alternating attention; and Session 4, divided attention. Focused attention, being a basic process, is trained in all sessions. The order of the sessions was determined to train the simplest forms of attention (sustained and selective) before the most complex forms (alternating and divided).

The development of the *Focusing* app was also based on the 5-facet model of mindfulness of Baer et al [29]. This model describes mindfulness as a multidimensional construct. It includes the following 5 components: observation, description, acting mindful, nonjudgment of inner experience, and nonreactivity to inner experience. *Observation* refers to noticing or paying attention to internal and external experiences, such as sensations, cognitions, emotions, images, sounds, and smells. *Description* refers to labeling internal experiments with words. *Acting mindful* includes being attentive to your current activities and may be opposed to behaving mechanically while one's attention is focused elsewhere (often called autopilot). *Nonjudgment of inner experience* is about refraining from assessing thoughts and feelings. *Nonreactivity to inner experience* is about letting thoughts and feelings come and go, without being taken in or carried away by them.

To integrate the 5 mindfulness components of this model into *Focusing*, we selected and modified mindfulness exercises to fit the definitions of the components. The objective was for the intervention exercises to be broadly representative of all aspects of the model by Baer et al [29]. The components have therefore been integrated into the attentional exercises included in the 4 sessions.

XSL•FO
RenderX

Objectives

This study aimed to assess the content validity of the *Focusing* app. This research, inspired by the previous work of Vogt et al [30], had the specific goal of analyzing the logical content validity and ecological content validity of the attention training program delivered through the mobile app. Logical content validation aims to test whether the app and the proposed exercises are aligned with "an operational definition" of the theoretical framework. More specifically, experts were instructed to decide on the level of proximity between the proposed exercises and the theoretical constructs presented in this study. Ecological content validation (on the basis of user experience in everyday life) aims to determine, before the proof-of-concept stage with a clinical population, the ease of use of the mobile app and the comprehensibility of the exercises (including vocabulary and visual/auditory aspects) that are integrated in the app.

Methods

Early Version and Preliminary Work

An initial app was developed based on meditation exercises from the MBSR [11] and MBCT [21] approaches. Mindfulness and neuropsychology experts (N=7) were subsequently consulted with a Delphi approach to obtain their views on the exercises. In general, the experts recommended using a mindfulness intervention approach in its entirety. Indeed, it is not recommended to develop an intervention that modifies or combines 2 approaches. The experts strongly suggested starting over by doing something completely different from existing mindfulness approaches. In response to the recommendations of the first group of experts, a new version of the *Focusing* app for training attentional functions was developed by Laverdière et al [28] and was subjected to validation in this study.

Content Validation

Face validity represents the subjective feeling of the validity of an instrument [31,32]. It is not strictly about what the instrument measures, but what it seems to measure according to the research team. An instrument with good apparent validity will generate more positive attitudes, greater involvement in the task, and greater authenticity in the responses, which will increase the validity of the task performed [33]. For the development of the *Focusing* app, the research team implemented mindfulness-inspired attentional focus activities based on the understanding of the theoretical model of attention by Sohlberg and Mateer [9] and the model of mindfulness by Baer et al [29]. The construction of the intervention is based on the ORBIT model [28], which is considered valid by the research team.

Content validity assesses the extent to which the various items of an instrument are representative of the construct and its different facets [30,31]. Haynes et al [34] proposed 7 essential rules for validating the content of an instrument: Rule 1, rigorously define the domain and facets of the measured construct and validate this definition; Rule 2, use a sample of experts and members of the target population to create the items and define other aspects of the instrument; Rule 3, submit all aspects of the instrument to content validation; Rule 4, consult several experts to validate the content of the instrument and quantify their judgments using formalized scales; Rule 5, examine the proportional representation of items or tests relative to the different facets of the construct; Rule 6, present the results of content validation when publishing any new instrument; and Rule 7, consider all subsequent psychometric analyses to refine content validation.

To develop the *Focusing* app, validated theoretical models of attention and mindfulness were used (Rule 1) [9,29]. Mindfulness and attention training experts were first consulted when the intervention was created (Rule 2). Moreover, the entire attention process training was validated by experts and people from the general population (Rules 2, 3, and 4). The validation data allowed to refine the intervention and thus to have a development procedure and a valid intervention (Rule 5). Studies on the validation process are presented in this manuscript (Rule 6). All the data obtained will make it possible to reflect comprehensively on the modifications to be made to the app for improvement before testing the app in a clinical population (Rule 7).

Ethical Considerations

Ethics approval was obtained for this study from the *Centre Intégré Universitaire de Santé et de Services Sociaux de la Capitale-Nationale* (CIUSSS-CN) rehabilitation and social integration committee (approval number: MP-13-2022-2523). Written consent was obtained from all participants after they were informed about the study's objectives and potential benefits. Participation was completely voluntary, and participants retained the right to withdraw at any moment without prejudice. To protect their privacy, all data were kept strictly confidential and anonymized, ensuring that participants could not be identified. No financial compensation or incentives were offered for participation.

Study 1: Logical Content Validation

Participants

Logical content validation was conducted by a panel of 7 experts from Canada (n=5) and France (n=2) who participated in the validation process in February 2023. The selected experts had to meet the following inclusion criteria: (1) be a clinician, have at least 5 years of clinical experience with moderate or severe TBI clients, or have at least 5 years of experience in the mindfulness practice; (2) be a neuropsychologist, psychologist, or occupational therapist; and (3) have practiced or be familiar with physical rehabilitation. The experts were recruited via email from September 2022 to January 2023. A total of 21 experts were contacted. Of these 21 experts, 10 agreed to participate; however, 1 expert withdrew and 2 did not complete all the validation steps. The participants included 4 clinical neuropsychologists involved in teaching and research. They specialize in the development of techniques for evaluation or intervention in different clinical populations (TBI, autism spectrum disorder, and neurodegenerative diseases). The other 3 participants are clinical psychologists involved in teaching and research. They specialize in mindfulness interventions in pediatric and adult populations.

XSL•FO
RenderX

Procedure

Experts who agreed to participate in the study received an email containing information about the exercises and a link to download a desktop version of the *Focusing* app, which they were asked to explore over a 4-week period. The desktop version of the app corresponds to a preliminary version of the implementation on the Google Play Store and is visually identical to the currently available app on Android. The experts were invited to try the different sections of the app and to perform all the exercises of the program. They were asked to assess whether the exercises in the program are theoretically valid according to the theoretical model of attention by Sohlberg and Mateer [9] and the theoretical model of mindfulness by Baer et al [29], using a questionnaire on the Lime Survey platform (closed survey). In the questionnaire, experts had access to the definitions of the attentional components of the model by Sohlberg and Mateer [9] and the 5 components of the model of mindfulness by Baer et al [29] because their respective knowledge of the theoretical models was different. In addition, the component of attention trained in each exercise was identified. The experts had to assess whether each exercise involves the targeted component of attention, using a Likert scale ranging from 1 (totally disagree) to 4 (totally agree). Using the same Likert scale, they also had to assess the extent of training of each of the 5 components of mindfulness, in general, in the exercises. Indeed, the 5 components need not be present in all the exercises, but rather they should be in the mindfulness approach used for the construction of the exercises. They also had to comment on the exercises and suggest improvements when they considered that an item was not theoretically valid. Finally, the experts had to assess the relevance of the first exercise of each session, using a Likert scale ranging from 1 (totally irrelevant) to 4 (totally relevant). The questionnaire consisted of 35 items and 20 pages. The experts could see their progress in the questionnaire using a progress bar at the top of the page. They could also save their answers and resume the questionnaire later using their email and a password. They also had the option to change their answers and go back in the questionnaire. No cookies were used. The experts' responses were anonymous. Data and comments were then analyzed, and a discussion meeting with the experts (videoconference) was subsequently held to discuss the elements that did not reach consensus between the experts. The changes suggested by the experts were then applied to the mobile app. The experts concluded this meeting in a consensual manner by saying that they would consider the app to be valid once their suggestions were incorporated. After making the adjustments and incorporating the experts' suggestions, one of the experts reviewed the entire content of the intervention to validate it again. This expert specializes in adapting mindfulness interventions to different clinical populations. In addition, one of the mindfulness experts trained the person who lent her voice when recording voice tone and speed exercises. Once this step was completed, the usability of the new version of the mobile app was tested.

Data Analysis

For the evaluation of the content of the app and the theoretical model of Sohlberg and Mateer [9] regarding attention and that of Baer et al [29] regarding mindfulness by experts, a Likert-type scale was used with 4 possible responses: 1, I totally disagree; 2, I somewhat disagree; 3, I somewhat agree; and 4, I totally agree. For this analysis, the content validity index (CVI) was calculated from the count of the responses "I somewhat agree" and "I totally agree" divided by the total number of experts. According to the literature, the items evaluated must have a CVI greater than or equal to 0.78 (78%); otherwise, they need to be adjusted according to experts' suggestions [35].

Results

The experts assessed the theoretical validity of each exercise of the sessions in relation to the attentional model used [9] for a total of 16 exercises. They also assessed the theoretical validity of the attention training program in relation to the mindfulness model used [29].

Regarding the validity of the attention model (Table 1), the calculated CVI for 14 exercises (out of 16) ranged between 0.86 and 1.00, indicating that these exercises are valid [35]. One exercise with a CVI value of 0.71 required revision. The experts mentioned that this exercise was too focused on mindfulness concepts, which strayed away from the concept of selective attention. They also pointed out that more distractions should be present in the exercise. In addition, experts considered another exercise (Exercise 4.4) to be not valid, as it had a CVI value of 0.42, and thus, it was eliminated. According to the experts, Exercise 4.4, which consisted of a dual visual-auditory task, was too demanding for people with acquired brain injury. As mentioned above, the first exercise of each session is the same to allow participants to get into a state conducive to exercises aimed at training the different components of attention. The relevance of repeating this exercise was assessed by the experts. The calculated CVI was 0.86, indicating that the repetition of this exercise is relevant to the program.

For validation related to the mindfulness theoretical framework (Table 2), the calculated CVI for all 5 components of the model ranged between 0.86 and 1.00, indicating that the remediation program is valid with regard to its mindfulness component.

XSL•FO
RenderX

Table 1. Content validity index and data of each exercise based on the attention model by Sohlberg and Mateer [9], as evaluated by expert judges.

Instrument item[a]	CVI[b]	Mean (SD)	Modification
Session 1: focused and sustained attention			
Exercise 1.1	0.86	3.33 (0.47)	Minor
Exercise 1.2	0.86	3.29 (0.70)	Minor
Exercise 1.3	0.86	3.29 (0.70)	Minor
Exercise 1.4	0.86	3.14 (0.64)	Minor
Session 2: selective attention			
Exercise 2.2	0.71[c]	3.43 (0.90)	Major
Exercise 2.3	0.86	3.29 (0.70)	Minor
Exercise 2.4	0.86	3.14 (0.64)	Minor
Session 3: alternating attention			
Exercise 3.2	0.86	3.29 (0.70)	Minor
Exercise 3.3	1.00	3.71 (0.45)	No change
Exercise 3.4	0.86	3.14 (0.64)	Minor
Session 4: divided attention			
Exercise 4.2	0.86	3.57 (0.73)	Minor
Exercise 4.3	0.86	3.57 (0.73)	Minor
Exercise 4.4	0.42[c]	2.71 (0.88)	Item deleted

[a]Exercises 2.1, 3.1, and 4.1 are not included in the table as they have the same content as Exercise 1.1.

[b]CVI: content validity index.

[c]Items with a CVI below the acceptable level of 0.78.

Table 2. Content validity index and data of each component of the theoretical framework of mindfulness by Baer et al [29] for the whole training according to expert judges.

Instrument item	CVI[a]	Mean (SD)	Modification
Observing	1.00	3.86 (0.35)	No change
Describing	0.86	3.00 (0.53)	Minor
Acting with awareness	0.86	3.57 (0.73)	Minor
Nonjudging of inner experience	1.00	3.71 (0.45)	No change
Nonreactivity to inner experience	1.00	3.57 (0.49)	No change

[a]CVI: content validity index.

Synthesis of Expert Recommendations

During the logical content validation, 4 dimensions were explored by the experts: theoretical model of attention, theoretical model of mindfulness, technological aspects, and audio. Regarding the theoretical model of attention, the recommendations of the experts were to: (1) increase the duration of exercise in sustained attention; (2) increase distractions in selective attention; and (3) remind more frequently about the instructions in all exercises. Regarding the theoretical model of mindfulness, the recommendations of the experts were to: (1) simplify the vocabulary; (2) give more concrete examples to the participants; and (3) decrease the speed of the voice in the audio tracks of the exercises. In terms of technology, the experts reported some inconsistencies between the visual elements of the app and the audio content of the exercises.

Exercises were modified to reflect these recommendations. First, the sustained attention exercises (1.1 to 1.4) were extended by a few minutes each. Visual and auditory distractions were added in the selective attention exercises (2.1 to 2.4). Throughout the training exercises, the speech was adapted by adding reminders of instructions several times during a single exercise. The vocabulary was also adapted to be accessible to the general population, for example, by adding concrete examples and adapting the tone of voice (speaking less quickly). The technological inconsistencies raised by the experts were also corrected.

XSL•FO
RenderX

Study 2: Ecological Content Validation

Participants

For this study, a convenience sample of 10 healthy participants (7 female and 3 male participants) was recruited from the social networks of the research laboratory. Participants were required to meet the following criteria: (1) be over 18 years of age; (2) understand and speak French; (3) own an Android device; (4) have no diagnosis of TBI; (5) have no serious mental disorder including but not restricted to schizophrenia spectrum disorder, bipolar disorder, intellectual disability, and attention deficit disorder; (6) have no vestibular disorder; and (7) have no substance abuse.

An academic researcher also explored the mobile app. This expert is interested in the neuropsychology of rehabilitation, particularly among adults who have experienced TBI or stroke. This researcher's work also focuses on the use of technology in clinical neuropsychology (mobile devices, teleneuropsychology, and virtual reality). This expert did not complete the validation questionnaires, but this expert's comments were taken into consideration in the validation process.

Procedure

The 10 participants completed the entire study remotely. Interested participants were directed to an online eligibility questionnaire via a link in the recruitment advertisement. Eligible participants received the mobile app download procedure. They then had 2 weeks to explore by themselves the mobile app and fill out the questionnaires provided for the experimental protocol via Lime Survey (open survey). The self-administered questionnaires were: a sociodemographic questionnaire, the Simulation Task Load Index (SIM-TLX) questionnaire (mental load; [36]), the French version of the Mobile Application Rating Scale (MARS-F) questionnaire (quality of the app; [37,38]), the French version of the AttrakDiff questionnaire (user experience; [39,40]), and a questionnaire concerning the participants' satisfaction with the elements of the app as well as their perception of the relevance of these elements. The whole questionnaire set consisted of 107 items and 9 pages. Participants could see their progress in the questionnaire on a progress bar at the top of the page. They could also save their answers and resume the questionnaire later using their email and a password. They also had the option to change their answers and go back in the questionnaire. No cookies were used. The participants' responses were anonymous. Once all participants completed the exploration, questionnaire responses were analyzed and changes to the mobile app were made if problematic elements were identified.

Instruments

SIM-TLX Questionnaire

This questionnaire [36] provides a measure of participants' cognitive load. It contains 9 scales (mental demands, physical demands, temporal demands, frustration, task complexity, situational stress, distractions, perceptual strain, and task control). The responses range from 1 (very low) to 20 (very high).

MARS-F Questionnaire

This questionnaire [37,38] provides a measure of the app's quality. Items are rated on a scale of 1 to 5 (1=inadequate; 2=poor; 3=acceptable; 4=good; and 5=excellent). The MARS-F questionnaire comprises 4 main scales: (1) user engagement (5 items: entertainment, interest, customization, interactivity, and target group); (2) functionality (4 items: performance, usability, navigation, and gestural design); (3) esthetics (3 items: layout, graphics, and visual appeal); and (4) information quality (7 items: accuracy of app description, goals, quality of information, quantity of information, quality of visual information, credibility, and evidence base). For each scale, a mean score was calculated.

AttrakDiff Questionnaire

This questionnaire [39,40] provides a measure of the app user experience. It provides data on 3 quality scales (pragmatic, hedonic-stimulation, and hedonic-identity) and a scale of global attractiveness. Each item is rated on a scale ranging from –3 to +3. Values close to the average (between –1 and 1) are standard; they mean that the product meets its objective. Values outside this neutral zone are to be considered as positive (from 1 to 3) or negative (from –1 to –3) points of the app.

Satisfaction Questionnaire

In this questionnaire, participants rated on a scale of 1 (totally dissatisfied) to 5 (totally satisfied) their levels of satisfaction regarding the 5 sections (About, Exercises, Homework, Ambiences, and Me) and the functionalities of the app. Participants also rated the relevance of each section and functionality (technological issues, sound, and ability to navigate) on a scale from 1 (totally irrelevant) to 5 (totally relevant). They also documented the technical issues they encountered while exploring the app.

Data Analysis

To compute for all participants the variables surveyed by the questionnaires, central trend measurements (mean and SD) and frequency analyses were performed on the data. Exploratory correlations were made on the data.

Results

A total of 11 participants started the usability validation study. Only 10 participants (7 female and 3 male participants) completed the validation process of the app. Their mean age was 51 (SD 15) years, and the mean number of years of education was 13 (SD 4). Most participants (5/10, 50%) worked full-time, 40% (4/10) were retired, and 10% (1/10) were studying full-time. None worked in the technology industry. Moreover, 80% (8/10) of the participants regularly watched animated films and only 40% (4/10) regularly played video games. The smartphone was the most used technological tool by participants (9/10, 90% used it daily) followed by computers (4/10, 40% used it daily) and tablets (2/10, 20% used it daily). On a scale of 1 to 10, participants' comfort levels with these technological tools averaged 7.36 (SD 1.57) for smartphones, 5.55 (SD 2.58) for computers, and 5.45 (SD 2.62) for tablets.

The results of the different questionnaires are presented in Table 3.

XSL•FO
RenderX

Table 3. Scores of the scales of the questionnaires for usability validation of attention process training according to healthy participants.

Questionnaire scale	Score, mean (SD)	Range, minimum to maximum
SIM-TLX[a] (range: 1 [very low] to 20 [very high])		
Mental demands	4.73 (2.33)	1 to 8
Physical demands	1.45 (1.04)	1 to 4
Temporal demands	1.36 (0.81)	1 to 3
Frustration	1.45 (1.21)	1 to 5
Task complexity	3.82 (2.27)	1 to 7
Situational stress	1.18 (0.60)	1 to 3
Distractions	5.55 (3.59)	1 to 11
Perceptual strain	2.64 (4.20)	1 to 15
Task control	2.73 (2.41)	1 to 8
MARS-F[b] (range: 1 [low quality] to 5 [high quality])		
Engagement	3.98 (0.48)	1 to 5
Functionality	4.30 (0.45)	3 to 5
Esthetic	3.80 (0.45)	3 to 5
Information	2.89 (0.60)	1 to 5
Objective quality	3.74 (0.41)	1 to 5
Subjective quality	3.65 (0.36)	2 to 5
AttrakDiff (range: –3 to 3)		
Pragmatic quality	1.94 (0.61)	0 to 3
Hedonic quality: Simulation	1.79 (0.55)	–2 to 3
Hedonic quality: Identification	1.51 (0.57)	–2 to 3
Global attractivity	2.36 (0.57)	0 to 3
Satisfaction (range: 1 [totally dissatisfied/irrelevant] to 5 [totally satisfied/relevant])		
Satisfaction	4.30 (0.62)	2 to 5
Relevance	4.41 (0.47)	2 to 5

[a]SIM-TLX: Simulation Task Load Index.

[b]MARS-F: Mobile Application Rating Scale, French version.

The results showed that the app requires low physical and temporal demands according to the SIM-TLX questionnaire. All items scored less than 4. Even if the overall results on the 9 scales were low, mental demand, distractions, and task complexity were rated slightly higher than the others. The app generated a low level of frustration and stress (<1.5). Participants had a good sense of control when using the app (a small score indicates a good sense of control).

The results of the MARS-F questionnaire showed that, in general, the quality of the app was adequate according to the participants. The average quality rating for the app was between 3.75/5 (SD 0.41) (objective quality) and 3.65/5 (SD 0.36) (subjective quality), demonstrating moderate quality [37,38]. The results indicated that the app is functional and esthetic and provides an adequate level of engagement. The information scale had a lower score (mean 2.89, SD 0.60). This suggests that participants did not have enough information about the app development process to judge its credibility.

The result of AttrakDiff's global attractiveness scale indicated that the user experience was adequate, with a mean score of 2.36 (SD 0.57). The result of the pragmatic quality scale, which describes the usability of the app, indicated that using the app was easy. The result of the hedonic quality scale-stimulation indicated that the app appeared new, interesting, and stimulating. The result of the hedonic quality scale-identity indicated that participants liked to interact with the app (eg, voice). In fact, all results were between 1 and 3 on the –3 to +3 scale, which points out the positive aspects of the app.

The satisfaction questionnaire showed that participants were generally satisfied with the different sections and functionalities of the app. They also considered that the different sections and functionalities of the app were relevant.

Synthesis of Correlation Analyses

A positive and strong correlation was obtained between the participants' ease of using a tablet and their ease of using a

XSL•FO
RenderX

phone (r=0.832; P=.001). This suggests that *Focusing* could be used on a tablet as well.

A negative and moderate correlation was obtained between the result of the cognitive load of participants (SIM-TLX) and the functionality score of the app (r=–0.635; P=.049). Better functionality was therefore associated with a smaller mental load. Having obtained a high score in functionality (MARS-F), the app does not appear to lead to mental overload that could interfere with the attention processes of participants.

A positive and moderate correlation was obtained between pragmatics (ease of use; AttrakDiff) and engagement level (MARS-F) (r=0.657; P=.04). The ease of use of *Focusing* therefore increases participants' commitment to the intervention. Better engagement was positively and strongly correlated with the subjective quality (MARS-F) of the app perceived by the participant (r=0.821; P=004). The level of satisfaction of participants with the different sections of *Focusing* was also positively and moderately correlated with objective quality (MARS-F) (r=0.669; P=.04) and positively and strongly correlated with the overall level of attractiveness (AttrakDiff) of the app (r=0.919; P=.001).

Discussion

Main Results

Despite the growing popularity of mobile apps for cognitive training and mindfulness, there is still a lack of comprehensive analysis supporting the content and quality of these apps [22-25]. A systematic review identified 53 commercial cognitive training or rehabilitation apps in 2023 [23]. Among these apps, only 5 had an overall MARS score exceeding 4 points, and more than half had a score of less than 3 points. Considering that a MARS score of 3 points is acceptable, most of these apps did not even meet the acceptable criteria for quality [37,38]. Furthermore, to our knowledge, none of these apps were intended to address attention in people who have moderate or severe TBI.

Furthermore, no studies on the development of attention interventions identified in the literature used a validated development model. Due to the existing gaps in the methodology of studies on attention remediation interventions, it is essential to use a model like ORBIT in the development of future interventions to address these limitations. Promising behavioral interventions in development are often abandoned rather than refined or improved if they fail the initial tests. In some cases, successful interventions in early studies are not pushed toward more rigorous efficacy studies, while others are tested in efficacy studies prematurely.

This research aimed to validate the content and user experience of an attention training intervention developed using the ORBIT model via a mobile app called *Focusing*. The first step was to validate the logical content with experts to ensure that the exercises train the attention components of the model by Sohlberg and Mateer [9]. It also aimed to validate that the remediation program sessions incorporate the 5 components of the mindfulness model by Baer et al [29]. The results showed that experts (including experts in cognitive functions and experts in mindfulness) considered the intervention to be theoretically valid according to the 2 models. Regarding the theoretical model of attention, the experts validated that each session of the intervention involves a component of the attentional model. Following the expert logical validation process, 1 exercise was removed, 1 underwent major revisions, 10 were improved according to the recommendations, and 1 was not changed. The removed exercise, which consisted of a dual visual and auditory task, was deemed too demanding for people with acquired brain injury. Major changes included increasing the duration of exercise, increasing the level of distraction, and making the instructions more clear. Moreover, the experts confirmed that the 5 components of the mindfulness model are included throughout the exercises. Changes were made to the exercises as recommended by the experts: the vocabulary for the instructions was simplified, concrete and more practical examples were added, the speed of the voice track/instructions was reduced, and pauses were added to the voice track.

The second step was the ecological content validation of the *Focusing* app. The relevance, satisfaction, quality, attractiveness, and workload associated with each section of the app were evaluated by participants from the general population.

The results of the SIM-TLX questionnaire underlined the level of workload when using the app, which involves low physical and temporal demands. It also involves low stress and frustration. These results confirm that the *Focusing* app respects the principles of mindfulness. Indeed, mindfulness promotes being in the present moment and the acceptance of the experience. The scales of the SIM-TLX questionnaire associated with higher workload are mental demand, distraction, and task complexity. The results confirm that attentional training program exercises require some cognitive effort. In addition, distractors were deliberately integrated into the exercises to train selective attention.

MARS is one of the most widely used questionnaires in the literature for app quality assessment [37]. This questionnaire was also used by clinicians to assess the quality of 23 mindfulness apps available on the Apple App Store and Google Play Store [25]. The results showed that the app is of good quality and is esthetic, functional, and engaging. These features allow you to properly interact with the app. Resources can therefore be entirely devoted to the training of attention. The Quality Information scale had the worst result. To prevent misinformation and adverse effects from *Focusing* app use, information quality must be improved. In fact, wrong or misleading information could result in a decrease in user safety [41]. Informing participants of the involvement of experts in the development process (eg, psychotherapists and researchers in mindfulness or cognition training) might help to address this problem. Moreover, a better description of the app on the Play Store might improve information quality.

The results of the AttrakDiff questionnaire demonstrated good user experience of the *Focusing* app. The app was easy to use, interesting, and challenging. It was also pleasant to interact with the app. This also means that the voice (tone and speed) used for the exercises was adequate. The results showed that the recommendations of the experts during the logical content validation helped improve the user experience of the app.

XSL•FO
RenderX

In comparison with our study, the attention training application (ATA) developed by Hill et al [42,43] is an app aimed at training the attention of elderly people without cognitive impairment. To our knowledge, this app is the most recent in the literature to be validated with the target population. The ATA is an adaptation of the «dual n-back» computerized intervention [44]. This study did not use a validated intervention development model such as ORBIT. The app was, however, validated twice [42,43] with a small number of elderly people and using a mixed methods study design. This study corresponds to Phase I(b)-Refine of the ORBIT model. As in our study, participants were asked to complete questionnaires to capture familiarity with technology and to provide their general impressions of the ATA on a scale from 0 (negative) to 10 (positive in 4 categories: overall, easy, satisfying, and interesting). The User Interface Satisfaction questionnaire was also used [45]. Feedback had to be given on a 7-point Likert-type scale for different items. Interviews were conducted to allow participants to share their perceptions of the app. A limitation of this previous study is that the quality of the app was not assessed with the MARS questionnaire. Moreover, the cognitive load associated with using the app was not assessed.

Limitations

The content validation of the attention remediation program has some limitations. Although the logical content validation experts concluded that the program would be fully valid once their recommendations were incorporated, the revised program was only reviewed by 1 of these experts after changes were implemented. In addition, 1 of the experts could not be present at the focus group to discuss the recommendations.

The current validation of user experience was limited by the characteristics of the participants. First, the sample included more women (70%) than men (30%). However, only one-third of people with TBI in Canada are women [46,47]. In addition, most participants were in their 50s or 60s, while youth and young adults aged 12 to 19 years are 5.2 times more likely to have TBI [46]. Finally, participants in the sample considered themselves weakly to moderately comfortable with the use of electronic devices (computers, tablets, and smartphones), which may be related to the age of the participants. On the other hand, young people are more comfortable with the use of apps on smartphones. Different data about user experience could have been obtained with a sample that is very comfortable with technologies. However, by targeting people who are less skilled with technology, we ensured that this would not be an obstacle.

Future Research

Future studies should assess the safety, tolerability, and acceptability of this attention training intervention via a mobile app in a population with moderate or severe TBI. More specifically, as part of phases Ib (refine) and IIa (proof-of-concept) of the ORBIT model [27], 5 to 10 people with moderate or severe TBI should be recruited as part of a multiple single-case experimental design. The test phase should include a waiting period and an intervention period, both marked by pretests and posttests. At the end of this phase, it will be possible to determine whether the intervention is clinically acceptable and whether it can be made available to a larger sample, leading to phase III (efficacy) studies that will involve a randomized controlled design. At the end of phase III, we also want to translate the tool for the English-speaking population in order to increase the recruitment pool. This will involve translating the audio files and the menus of the different sections of the app. We also want to develop an iOS version of the app to make it more accessible and expand recruitment.

Conclusions

The *Focusing* app is one of the first mobile attention training apps scientifically validated with the ORBIT model and developed based on theoretical models of both attention and mindfulness. A systematic validation approach was used to ensure that the intervention was theoretically valid before clinical validation. In addition, this project provides a better understanding of the attentional processes modulated by attentional focus techniques based on mindfulness. In fact, this study attempts to dissect attentional processes from the angle of a theoretical model recognized in neuropsychology.

Acknowledgments

We thank Jonathan Caron-Roberge who helped with the programming of the mobile app. We also thank Jean-Luc Théberge who helped with the voice recordings of the sessions of the remediation program. Finally, we thank Marie-Josée Tremblay who provided her expertise to guide mindfulness exercises. This project received financial support from the *Fonds de Recherche du Québec, nature et technologies* (FRQ-NT) (reference number: 325564), the *Centre Interdisciplinaire de Recherche en Réadaptation et Intrégration Sociale* (Cirris), and the *Collectif de Recherche sur la Santé en Région* (CoRSeR).

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1

XSL•FO
RenderX

Description of the attention training exercises in each session and the attention components and mindfulness components trained in the exercises.
[DOCX File , 18 KB - formative_v9i1e64174_app1.docx]

References

1. Buttle H. Attention and working memory in mindfulness-meditation practices. The Journal of Mind and Behavior 2011;32(2):123-134 [FREE Full text]
2. Cohen RA. The Neuropsychology of Attention. New York, NY: Plenum Press; 1993.
3. Azouvi P, Arnould A, Dromer E, Vallat-Azouvi C. Neuropsychology of traumatic brain injury: An expert overview. Rev Neurol (Paris) 2017;173(7-8):461-472 [FREE Full text] [doi: 10.1016/j.neurol.2017.07.006] [Medline: 28847474]
4. Silver JM, McAllister TW, Arciniegas DB. Textbook of Traumatic Brain Injury, Third Edition. Washington, DC: American Psychiatric Association Publishing; 2019.
5. Cicerone KD, Goldin Y, Ganci K, Rosenbaum A, Wethe JV, Langenbahn DM, et al. Evidence-based cognitive rehabilitation: systematic review of the literature from 2009 through 2014. Arch Phys Med Rehabil 2019 Aug;100(8):1515-1533 [FREE Full text] [doi: 10.1016/j.apmr.2019.02.011] [Medline: 30926291]
6. Ponsford J, Velikonja D, Janzen S, Harnett A, McIntyre A, Wiseman-Hakes C, et al. INCOG 2.0 guidelines for cognitive rehabilitation following traumatic brain injury, part II: attention and information processing speed. J Head Trauma Rehabil 2023;38(1):38-51. [doi: 10.1097/HTR.0000000000000839] [Medline: 36594858]
7. Franck N. Remédiation Cognitive, 2nd Edition. Issy Les Moulineaux, France: Elsevier Masson; 2017.
8. Sohlberg MM, Mateer CA. Attention Process Training APT3. Lash & Associates. URL: https://lapublishing.com/apt3-attention-process-training/ [accessed 2025-03-24]
9. Sohlberg MM, Mateer CA. Effectiveness of an attention-training program. J Clin Exp Neuropsychol 1987 Apr;9(2):117-130. [doi: 10.1080/01688638708405352] [Medline: 3558744]
10. Prakash RS. Mindfulness meditation: impact on attentional control and emotion dysregulation. Arch Clin Neuropsychol 2021 Oct 13;36(7):1283-1290 [FREE Full text] [doi: 10.1093/arclin/acab053] [Medline: 34651648]
11. Kabat-Zinn J. An outpatient program in behavioral medicine for chronic pain patients based on the practice of mindfulness meditation: theoretical considerations and preliminary results. Gen Hosp Psychiatry 1982 Apr;4(1):33-47. [doi: 10.1016/0163-8343(82)90026-3] [Medline: 7042457]
12. Bishop SR, Lau M, Shapiro S, Carlson L, Anderson ND, Carmody J, et al. Mindfulness: a proposed operational definition. Clinical Psychology: Science and Practice 2004 Aug 01;11(3):230-241. [doi: 10.1093/clipsy/bph077]
13. Bondolfi G, Jermann F, Zermatten A. Les approaches psychothérapeutiques basées sur la pleine conscience (mindfulness); Entre vogue médiatique et applications cliniques fondées sur des preuves. Psychothérapies 2011;3(31):167-174. [doi: 10.3917/psys.113.0167]
14. Landry L. Entraînement à la pleine conscience via la réalité virtuelle immersive auprès de personnes ayant des symptômes post-commotionnels persistants suite à un traumatisme craniocérébral léger : une étude de cas. Université de Montréal, Papyrus. 2020. URL: https://umontreal.scholaris.ca/items/0ed990d9-072c-4288-988e-705f70e16711 [accessed 2025-03-24]
15. Wang J, Cao B, Yu P, Sun L, Bao W, Zhu X. Deep Learning towards Mobile Applications. 2018 Presented at: 38th International Conference on Distributed Computing Systems (ICDCS); July 02-06, 2018; Vienna, Austria. [doi: 10.1109/icdcs.2018.00139]
16. Schultchen D, Terhorst Y, Holderied T, Stach M, Messner E, Baumeister H, et al. Stay present with your phone: a systematic review and standardized rating of mindfulness apps in European app stores. Int J Behav Med 2021 Oct;28(5):552-560 [FREE Full text] [doi: 10.1007/s12529-020-09944-y] [Medline: 33215348]
17. Spijkerman MPJ, Pots WTM, Bohlmeijer ET. Effectiveness of online mindfulness-based interventions in improving mental health: A review and meta-analysis of randomised controlled trials. Clin Psychol Rev 2016 Apr;45:102-114 [FREE Full text] [doi: 10.1016/j.cpr.2016.03.009] [Medline: 27111302]
18. Jayewardene WP, Lohrmann DK, Erbe RG, Torabi MR. Effects of preventive online mindfulness interventions on stress and mindfulness: A meta-analysis of randomized controlled trials. Prev Med Rep 2017 Mar;5:150-159 [FREE Full text] [doi: 10.1016/j.pmedr.2016.11.013] [Medline: 28050336]
19. Gál É, tefan S, Cristea IA. The efficacy of mindfulness meditation apps in enhancing users' well-being and mental health related outcomes: a meta-analysis of randomized controlled trials. J Affect Disord 2021 Jan 15;279:131-142. [doi: 10.1016/j.jad.2020.09.134] [Medline: 33049431]
20. Compen F, Bisseling E, Schellekens M, Donders R, Carlson L, van der Lee M, et al. Face-to-face and internet-based mindfulness-based cognitive therapy compared with treatment as usual in reducing psychological distress in patients with cancer: a multicenter randomized controlled trial. JCO 2018 Aug 10;36(23):2413-2421. [doi: 10.1200/jco.2017.76.5669]
21. Segal ZV, Williams JMG, Teasdale JD. Mindfulness-based cognitive therapy for depression: A new approach to preventing relapse. New York, NY: Guilford Press; 2002.
22. East-Richard C, Laplante L, Vézina J, Cellard C. L'évaluation psychologique par le biais des applications mobiles. Canadian Psychology / Psychologie canadienne 2018 Aug;59(3):262-271. [doi: 10.1037/cap0000113]

XSL•FO
RenderX

23. Bang M, Jang CW, Kim HS, Park JH, Cho HE. Mobile applications for cognitive training: Content analysis and quality review. Internet Interv 2023 Sep;33:100632 [FREE Full text] [doi: 10.1016/j.invent.2023.100632] [Medline: 37312799]
24. Stec MA, Arbour MW, Hines HF. Client-centered mobile health care applications: using the mobile application rating scale instrument for evidence-based evaluation. J Midwifery Womens Health 2019 May;64(3):324-329. [doi: 10.1111/jmwh.12941] [Medline: 30887711]
25. Mani M, Kavanagh DJ, Hides L, Stoyanov SR. Review and evaluation of mindfulness-based iPhone apps. JMIR Mhealth Uhealth 2015 Aug 19;3(3):e82 [FREE Full text] [doi: 10.2196/mhealth.4328] [Medline: 26290327]
26. Christopher E, Alsaffarini KW, Jamjoom AA. Mobile health for traumatic brain injury: a systematic review of the literature and mobile application market. Cureus 2019 Jul 10;11(7):e5120 [FREE Full text] [doi: 10.7759/cureus.5120] [Medline: 31523551]
27. Czajkowski SM, Powell LH, Adler N, Naar-King S, Reynolds KD, Hunter CM, et al. From ideas to efficacy: The ORBIT model for developing behavioral treatments for chronic diseases. Health Psychol 2015 Oct;34(10):971-982 [FREE Full text] [doi: 10.1037/hea0000161] [Medline: 25642841]
28. Laverdière R, Banville F, Jackson PL. A mobile application to train attention processes after traumatic brain injury: Rational, design and feasibility evaluation. Annual Review of CyberTherapy and Telemedicine 2023;21:222-227.
29. Baer RA, Smith GT, Lykins E, Button D, Krietemeyer J, Sauer S, et al. Construct validity of the five facet mindfulness questionnaire in meditating and nonmeditating samples. Assessment 2008 Sep;15(3):329-342. [doi: 10.1177/1073191107313003] [Medline: 18310597]
30. Vogt DS, King DW, King LA. Focus groups in psychological assessment: enhancing content validity by consulting members of the target population. Psychol Assess 2004 Sep;16(3):231-243. [doi: 10.1037/1040-3590.16.3.231] [Medline: 15456379]
31. Anastasi A. Introduction à la psychométrie. Montréal: Guérin Universitaire; 1994.
32. Bernaud JL. Introduction à la psychométrie. Paris: Dunod; 2007.
33. Laveault D, Grégoire J. Introduction aux théories des tests en psychologie et en sciences de l'éducation. Bruxelles: de Boeck Université; 2014.
34. Haynes SN, Richard DCS, Kubany ES. Content validity in psychological assessment: A functional approach to concepts and methods. Psychological Assessment 1995 Sep;7(3):238-247. [doi: 10.1037/1040-3590.7.3.238]
35. Polit DF, Beck CT, Owen SV. Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. Res Nurs Health 2007 Aug;30(4):459-467. [doi: 10.1002/nur.20199] [Medline: 17654487]
36. Harris D, Wilson M, Vine S. Development and validation of a simulation workload measure: the simulation task load index (SIM-TLX). Virtual Reality 2019 Dec 21;24(4):557-566. [doi: 10.1007/s10055-019-00422-9]
37. Stoyanov SR, Hides L, Kavanagh DJ, Zelenko O, Tjondronegoro D, Mani M. Mobile app rating scale: a new tool for assessing the quality of health mobile apps. JMIR Mhealth Uhealth 2015 Mar 11;3(1):e27 [FREE Full text] [doi: 10.2196/mhealth.3422] [Medline: 25760773]
38. Saliasi I, Martinon P, Darlington E, Smentek C, Tardivo D, Bourgeois D, et al. Promoting health via mHealth applications using a French version of the mobile app rating scale: adaptation and validation study. JMIR Mhealth Uhealth 2021 Aug 31;9(8):e30480 [FREE Full text] [doi: 10.2196/30480] [Medline: 34463623]
39. Hassenzahl M, Burmester M, Koller F. AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In: Szwillus G, Ziegler J, editors. Mensch & Computer 2003. Berichte des German Chapter of the ACM, vol 57. Stuttgart, Germany: Vieweg+Teubner Verlag; 2003:187-196.
40. Lallemand C, Koenig V, Gronier G, Martin R. Création et validation d'une version française du questionnaire AttrakDiff pour l'évaluation de l'expérience utilisateur des systèmes interactifs. European Review of Applied Psychology 2015 Sep;65(5):239-252. [doi: 10.1016/j.erap.2015.08.002]
41. Terhorst Y, Rathner E, Baumeister H, Sander L. «Hilfe aus dem App-Store?»: Eine systematische Übersichtsarbeit und Evaluation von Apps zur Anwendung bei Depressionen. Verhaltenstherapie 2018 May 8;28(2):101-112. [doi: 10.1159/000481692]
42. Hill NL, Mogle J, Colancecco E, Dick R, Hannan J, Lin FV. Feasibility study of an attention training application for older adults. Int J Older People Nurs 2015 Sep 15;10(3):241-249. [doi: 10.1111/opn.12092] [Medline: 26073545]
43. Hill NL, Mogle J, Wion R, Kitt-Lewis E, Hannan J, Dick R, et al. App-based attention training: Incorporating older adults' feedback to facilitate home-based use. Int J Older People Nurs 2018 Mar;13(1):12163 [FREE Full text] [doi: 10.1111/opn.12163] [Medline: 28752612]
44. Jaeggi SM, Buschkuehl M, Jonides J, Perrig WJ. Improving fluid intelligence with training on working memory. Proc Natl Acad Sci U S A 2008 May 13;105(19):6829-6833 [FREE Full text] [doi: 10.1073/pnas.0801268105] [Medline: 18443283]
45. Chin JP, Diehl VA, Norman KL. Development of an instrument measuring user satisfaction of the human-computer interface. In: CHI '88: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 1988 Presented at: SIGCHI Conference on Human Factors in Computing Systems; May 15-19, 1988; Washington, DC, USA. [doi: 10.1145/57167.57203]
46. Gordon KE, Kuhle S. 'Reported concussion' time trends within two national health surveys over two decades. Brain Inj 2018;32(7):843-849. [doi: 10.1080/02699052.2018.1463105] [Medline: 29652525]

XSL•FO
RenderX

47. Champagne AS, Yao X, McFaull SR, Saxena S, Gordon KR, Babul S, et al. Commotions cérébrales autodéclarées au Canada : une étude transversale. Statistique Canada. 2023. URL: https://www150.statcan.gc.ca/n1/pub/82-003-x/2023006/article/00002-fra.htm [accessed 2025-03-24]

Abbreviations

ATA: attention training application
CVI: content validity index
MARS-F: Mobile Application Rating Scale, French version
MBCT: mindfulness-based cognitive therapy
MBSR: mindfulness-based stress reduction
ORBIT: Obesity-Related Behavioral Intervention Trials
SIM-TLX: Simulation Task Load Index
TBI: traumatic brain injury

Edited by A Mavragani; submitted 10.07.24; peer-reviewed by S Quaglini; comments to author 15.12.24; revised version received 29.01.25; accepted 04.02.25; published 09.04.25.

Please cite as:
Laverdière R, Jackson PL, Banville F
A New Mobile App to Train Attention Processes in People With Traumatic Brain Injury: Logical and Ecological Content Validation Study
JMIR Form Res 2025;9:e64174
URL: https://formative.jmir.org/2025/1/e64174
doi:10.2196/64174
PMID:40202789

©Roxanne Laverdière, Philip L Jackson, Frédéric Banville. Originally published in JMIR Formative Research (https://formative.jmir.org), 09.04.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Cost-Effectiveness Analysis of a Machine Learning–Based eHealth System to Predict and Reduce Emergency Department Visits and Unscheduled Hospitalizations of Older People Living at Home: Retrospective Study

Charlotte Havreng-Théry[1,2], MS, PhD; Arnaud Fouchard[3], MD; Fabrice Denis[4], MD, PhD; Jacques-Henri Veyron[2], MS; Joël Belmin[1,5], MD, PhD

[1]Laboratoire Informatique Médicale et Ingénierie des Connaissances en eSanté, Institut National de la Santé et de la Recherche Médicale and Sorbonne Université, Paris, France

[2]PRESAGE, Paris, France

[3]EIR Conseil Santé, Antony, France

[4]Institut Inter-Régional de Cancérologie Jean Bernard, Le Mans, France

[5]Hôpital Charles Foix, Assistance Publique-Hôpitaux de Paris, Ivry-sur-Seine, France

Corresponding Author:
Jacques-Henri Veyron, MS
PRESAGE
112-114 rue la Boétie
Paris, 75008
France
Phone: 33 622152004
Email: jhveyron@presage.care

Abstract

Background: Dependent older people or those losing their autonomy are at risk of emergency hospitalization. Digital systems that monitor health remotely could be useful in reducing these visits by detecting worsening health conditions earlier. However, few studies have assessed the medico-economic impact of these systems, particularly for older people.

Objective: The objective of this study was to compare the clinical and economic impacts of an eHealth device in real life compared with the usual monitoring of older people living at home.

Methods: This study was a comparative, retrospective, and controlled trial on data collected between May 31, 2021, and May 31, 2022, in one health care and home nursing center located in Brittany, France. Participants had to be aged >75 years, living at home, and receiving assistance from the home care service for at least 1 month. We implemented among the intervention group an eHealth system that produces an alert for a high risk of emergency department visits or hospitalizations. After each home visit, the home care aides completed a questionnaire on participants' functional status using a smartphone app, and the information was processed in real time by a previously developed machine learning algorithm that identifies patients at risk of an emergency visit within 7 to 14 days. In the case of predicted risk, the eHealth system alerted a coordinating nurse who could then inform the family carer and the patient's nurses or general practitioner.

Results: A total of 120 patients were included in the study, with 60 in the control group and 60 in the intervention group. Among the 726 visits from the intervention group that were not followed by an alert, only 4 (0.6%) resulted in hospitalizations ($P<.001$), confirming the relevance of the system's alerts. Over the course of the study, 37 hospitalizations were recorded for 25 (20.8%) of the 120 patients. Additionally, of the 120 patients, 9 (7.5%) were admitted to a nursing home, and 7 (5.8%) died. Patients in the intervention group (56/60, 93%) remained at home significantly more often than those in the control group (48/60, 80%; $P=.03$). The total cost of primary care and hospitalization during the study was €167,000 (€1=US $1.09), with €108,000 (64.81%) attributed to the intervention group ($P=.20$).

Conclusions: This study presents encouraging results on the impact of a remote medical monitoring system for older adults, demonstrating a reduction in both emergency department visits and hospitalization costs.

Trial Registration: ClinicalTrials.gov NCT05221697; https://clinicaltrials.gov/study/NCT05221697

XSL•FO
RenderX

(JMIR Form Res 2025;9:e63700) doi:10.2196/63700

KEYWORDS

monitoring; older adult; predictive tool; home care aide; emergency department visit; cost-effectiveness; artificial intelligence; electronic health; eHealth; emergency department; unscheduled hospitalization; aging; retrospective study; medico-economic; living at home; nursing home; emergency visit; Brittany; France; machine learning; remote monitoring; digital health; health informatics

Introduction

Older Adults and Emergency Hospitalizations

In France, over 13.4 million people are aged 65 years or older, representing 20% of the French population. This proportion has increased by 4% in 20 years [1]. With aging and frailty, the risk of unplanned hospitalizations is increased.

Studies show that in France, older adults living at home represent 13% of the population that is hospitalized at least once a year, compared to 6% for the general population [2,3]. According to medico-economic data, nearly 1.6 million people aged >80 years were hospitalized in 2017, twice the number in the general population. In addition, the average length of stay doubles between the ages of 60 and 80 years, reaching nearly 10 days for those aged >90 years [4].

Depending on the study, between 19% and 67% of hospitalizations of older people are considered avoidable [5,6].

Medical and social frailty and comorbidity of the older population are the main factors that increase the prevalence and time spent in hospital [3].

The emergency department (ED) is the main entry point for the older people. Half of the people who visit the ED are subsequently hospitalized. Moreover, older adults represent 41% of ED visits, and nearly 23% of them spend more than 8 hours there and 34% spend between 4 and 8 hours—more than twice as long as younger people. This longer stay in the ED is compounded by a higher rate of transfer to other departments or hospitals [3,7].

In people aged >80 years, main causes of admission to the ED are traumatic events (generally related to fall; 25%), cardiovascular events (17%), altered general condition or infection (12%), a respiratory symptom (12%), a gastrointestinal symptom (10%), and neurological symptoms (9%) [3,7].

Emergency hospitalization represents a significant medical and economic cost. Older adults' hospitalizations reduce the level of dependency by 30% to 60% (immobilization syndrome, confusion, malnutrition, urinary incontinence, undesirable effects related to treatments, etc). While people aged >75 years represent 8% of the population, they are responsible for one-fifth of health care expenditure [8,9]. In France, the main expenses to support the loss of autonomy of older adults are approximately €12.2 billion (€1=US $1.09) in care expenses and €10.7 billion in human and technical assistance. The challenge is not so much to reduce hospitalization rates or the average length of stay but rather to avoid preventable emergency hospitalizations [1,2,8,10,11].

Telehealth and Artificial Intelligence for the Prediction and Prevention of Emergency Situations in Older Adults

Therefore, it is crucial to anticipate adverse events at home. The use of eHealth systems, telemedicine, and connected objects represent a promising approach for keeping older adults at home and preventing the loss of autonomy. Patient-reported outcome measures benefit people with chronic diseases by improving their quality of life, reducing mortality, and reducing ED visits and hospitalizations [7,12].

We have implemented a machine learning–based eHealth device to predict and prevent ED visits and unscheduled hospitalizations for older people living at home [13,14]. In 2022, we conducted a multicenter study with people aged >65 years, living at home, and receiving regular visits of home care aides (HCAs) to evaluate this device in real life. After each home visit, HCAs completed a smartphone-based questionnaire on the functional status of the patients. The information was processed in real time by a machine learning algorithm developed beforehand in order to identify patients at risk of an ED visit within 7 to 14 days (with a predictive performance of 83% sensitivity and 86% specificity). This machine learning algorithm can also predict symptoms or events such as malnutrition, falls, swollen legs, or depressions [15]. In the case of predicted risk, the system alerted a nurse coordinator who could then inform the family caregiver and the patient's nurses or general practitioner. A total of 206 patients were included and followed for 10 months. Compliance was good. Among the 2656 home visits, 405 alerts were issued. The system significantly reduced the number of ED visits when an intervention was performed after an alert. The system was considered easy to use and useful, and it was well accepted by more than 90% of the HCAs and coordinating nurses [15].

This algorithm opens the possibility of mobilizing health professionals to intervene early in an acute illness or in the decompensation of a chronic illness before it leads to an emergency hospitalization.

Objectives

The main hypothesis of this study was as follows: using a simple tool to predict the risk can allow HCAs to anticipate the incidence of emergency hospitalizations and, thereby, reduce the cost and side effects of emergency hospitalization.

The objective of this study was to analyze the clinical and economic impacts of this eHealth device in real life compared to the usual monitoring of frail older people living at home. In France, as this medical device is the first to predict ED use, there are, to our knowledge, few medico-economic studies available.

XSL•FO
RenderX

Methods

Study Design

This study was a single-center, retrospective, and controlled trial on data collected between May 31, 2021, and May 31, 2022, in one health care and home nursing center located in Brittany, France.

Recruitment

Participants were recruited from among adults aged 75 years and older who were living at home and receiving the assistance from HCAs. These HCAs were not health care professionals and typically provided assistance with nonmedical tasks (eg, helping with meals, assisting with personal care, housekeeping, and running errands). The dependency levels of the persons were established according to the French national instrument, which stratifies the dependency level from *groupe iso-ressources* (GIR) 1 (very severe dependency) to GIR 6 (no dependency) [16].

Patients in palliative care situations (GIR 1) were not included. Patients were proposed to be included in the controlled group without being offered any incentives. Written consent was obtained from all participants of the intervention group to be included in the study.

The control group was composed of people receiving the usual support from the home care service. The adjustment between intervention and control groups (1:1 or 1:2 when possible) was based on age (within 2 years), gender, and level of dependency. Participants in the control group were informed of their enrollment in the study. Data were collected between May 30, 2021, and May 31, 2022.

Data Collection

Sociodemographic and pathway data—family situation, dependency level, ED visits (dates and causes), hospitalization (dates and causes), and death (date)—were collected via the tracking system (for the intervention arm) and the coordinating nurses (for the control arm).

Intervention

For participants in the intervention group, HCAs used a smartphone app, for which they received training on its use (a 1-hour session at the start of the study). Each week, they completed a simple and concise questionnaire for each patient via the app. This questionnaire comprised 25 yes-or-no questions covering functional and clinical autonomy (activities of daily living); medical symptoms such as fatigue, falls, pain, and undernutrition; as well as behavioral changes (cognitive disorders and aggression), communication with caregivers, and social life.

The collected data were transmitted in real time to a secure server, where an artificial intelligence algorithm analyzed them to assess risk levels and predict ED visits within 7 or 14 days. The risk assessment was displayed on a Conformité Européenne–marked, web-based secure medical platform called PRESAGE CARE. When the algorithm detected a high-risk level, an on-screen alert was sent to the nurse coordinator of the home care center. This alert included information on recent changes in the patient's functional status, along with decision-support insights to help the nurse coordinator take appropriate action, such as contacting the family caregiver or other health care professionals, or conducting a home reassessment.

No specific intervention protocol was imposed on health care professionals, allowing them full autonomy in their decision-making. This alert-based intervention model was presented to and approved by the *Agence du Numérique en Santé* (National Agency for eHealth). The study followed the guidelines and recommendations of the *Haute Autorité de Santé* (French Health Authority) regarding the cost-effectiveness evaluation of medical devices.

Outcomes

The primary outcome was the difference in the cumulative incidence of emergency hospitalizations between the two groups, with death and institutionalization considered competing events. Unscheduled hospitalizations were defined as those occurring after an ED visit or those explicitly recorded as such in the PRESAGE CARE console or the data extracted from the center's information system.

Secondary outcomes included a comparison of hospitalization rates following an alert that did not lead to an intervention versus those following an alert that triggered an intervention, as well as a cost comparison of hospitalizations between the two groups. Additionally, the probability of remaining at home was assessed, defined as the time elapsed from May 30, 2021 (study start date), to the first occurrence of one of the following events: unplanned hospitalization, death, or institutionalization.

Cost Analysis

For each group, the hospital costs attributable to the ED visits and hospitalizations (unplanned and planned) were calculated on the basis of the homogeneous group of patients, determined by the declared causes of the hospitalization (CIM-10 PMSI coding, which is the French adaptation of US diagnosis-related groups). Hospital costs have been determined on the basis of published costs by the type of illness and length of hospital stay (basic category of classification in medicine, obstetrics, or surgery). The costs of consultations with the attending physician related to the PRESAGE CARE system were also included. The average length of hospital stay was compared between the groups. The burden costs (gross salary and taxes) of nurses who managed alerts (intervention arm) and conducted regular activities (control arm) were defined according to the time needed to proceed their tasks and average gross salary for nurses [17].

The cost for data collection by HCAs was defined as training costs (1 hour of training time multiplied by the hourly burden costs for HCAs [17]). The collection of data into PRESAGE Care eHealth device occurred during regular visits so no extra time was counted. In this study, the medical device was deployed free of charge.

The analysis was performed from the point of view of the total hospital cost per event and per patient in relation to the medical

XSL•FO
RenderX

benefit determined by the number of deaths, the number of institutionalizations, and the data in the literature concerning the loss of functional independence following hospitalization (in the absence of data on the reassessment of the dependency level after hospitalization). The cost analysis was mainly carried out from the point of view of social security expenditure.

Coordinating nurses were asked to report time spend in order to check alerts and act according to the procedure. The number of ED visits, emergency hospitalizations, deaths, and nursing home admissions were compared between the groups. The times before the first event were compared between groups.

Statistical Analysis

Continuous variables were described in terms of means, medians, IQRs, and SDs, depending on the normality of the distribution. Categorical variables were described in terms of numbers and percentages. Groups were compared using the chi-square or Fisher test for categorical variables and the Wilcoxon or 2-tailed *t* test for quantitative variables. Comparisons of cumulative incidences were made by the Gray test (entry into an institution and death in competition). A value of $P<.05$ was considered significant. Statistical analyses were performed using Stata software (version 17; StataCorp LLC).

Ethical Considerations

The research protocol for this study was submitted to and approved by the French National Committee for Biomedical Research, the Committee for the Protection of Individuals, and the French Agency for Health Product Safety (registration: 2234275). The beneficiaries as well as the caregivers and professionals involved were informed of the nature of this study and gave their written consent. Privacy and data protection rules were presented to participants and are available upon request. No compensation was offered.

Results

Patients' Characteristics

A total of 120 patients were included in the study, with 60 (50%) in the control group and 60 (50%) in the intervention group. Among them, 86 (71.7%) were women, including 45 (75%) of the 60 patients in the intervention group and 41 (68%) of the 60 patients in the control group ($P=.76$). The mean age of participants was 83.77 years, with no significant difference between the groups ($P=.44$).

Dependency levels, assessed using the *autonomie gérontologique GIR* grid, showed an average dependency score of 3.79, with no significant differences between the groups (all $P>.05$; Table 1). Overall, the characteristics of participants in the intervention and control groups were comparable.

Table 1. Participants' characteristics.

Characteristics	Intervention group (n=60)	Control group (n=60)	*P* value
Age (years), mean (95% CI)	84.55 (81.78-87.32)	82.99 (80.08- 85.90)	.44
Female, n (%)	45 (75)	41 (68)	.42
Low dependency level (GIR[a] 5-6), n (%)	14 (23)	10 (17)	.36
Mild dependency level (GIR 3-4), n (%)	37 (62)	37 (62)	>.99
High dependency level (GIR 2), n (%)	9 (15)	8 (13)	.79

[a]GIR: *groupes iso-ressources*.

In the intervention arm, a total of 792 visits were monitored through the app. Among these visits, 66 alerts were issued and transmitted to a coordinating nurse, of which 21 (32%) led to a health intervention (such as a general practitioner visit or a home visit by a nurse).

Following an alert-triggered intervention, only 1 (5%) out of the 21 patients was hospitalized, compared to 10 (91%) out of the 11 patients when no intervention was performed after an alert ($P<.001$). Among the 726 visits that did not trigger an alert, 4 (0.6%) hospitalizations occurred ($P<.001$; Table 2).

Table 2. Alert-triggered interventions and hospitalizations.

	Total	No hospitalization	Hospitalizations	*P* value
Alerts (n=66)				<.001
No alert-triggered intervention, n (%)	45 (68)	35 (78)	10 (22)	
Alert-triggered intervention, n (%)	21 (33)	20 (95)	1 (5)	
No alerts (n=726), n (%)	726 (100)	722 (99.4)	4 (0.6)	<.001

Medical impact

Altered general condition (7/37, 19%) and falls with fractures (7/37, 19%) were the most frequently described cause, followed by dyspnea (6/37, 16%). The causes are relatively distributed (Table 3).

XSL•FO
RenderX

Table 3. Unplanned hospitalizations causes (n=37).

Causes	Total, n (%)	Intervention group, n (%)	Control group, n (%)
Altered general condition	7 (19)	1 (3)	6 (16)
Dyspnea	6 (16)	2 (5)	4 (11)
Fall with fracture	7 (19)	4 (10)	3 (8)
Emergency surgery	5 (14)	2 (5)	3 (8)
Acute pain	4 (11)	2 (5)	2 (5)
Psychological symptoms	2 (5)	0 (0)	2 (5)
Kidney failure	2 (5)	1 (3)	1 (3)
Heart failure	1 (3)	0 (0)	1 (3)
Stroke	1 (3)	1 (3)	0 (0)
Fall without fracture	1 (3)	1 (3)	0 (0)
Missing value	1 (3)	1 (3)	0 (0)
Total	37 (100)	15 (41)	22 (59)

Emergency Hospitalizations, Deaths, and Nursing Home Admissions

A total of 37 emergency hospitalizations were recorded during the study, involving 25 patients. The number of emergency hospitalizations was 32% lower in the intervention group (15/37, 41%) compared to the control group (22/37, 59%).

Of the 6 scheduled hospitalizations, 5 occurred in the intervention group. The cumulative incidence of emergency hospitalizations was 23.3% (95% CI 13.5%-34.7%) in the control group, compared to 16.7% (95% CI 8.5%-27.2%) in the intervention group (*P*=.37).

Among participants, 16 patients left their homes due to either admission to a nursing home (9/120, 7.5%) or death (7/120, 5.8%). There were 7 deaths: 1 (14%) in the intervention group and 6 (86%) in the control group. Nine patients were admitted to a nursing home, including 3 (33%) in the intervention group and 6 (67%) in the control group.

The proportion of patients who left home was significantly higher in the control group (12/16, 75%) than in the intervention group (4/16, 25%; *P*=.03; Table 4).

Table 4. Health events during the study.

Health events	Total (N=120)	Intervention group, (n=60)	Control group, (n=60)	Difference between groups (%)	*P* value
In terms of the number of events (n=37), n (%)					
Emergency hospitalizations	37 (100)	15 (40.5)	22 (59.5)	–32	.16
In terms of the number of patients, n (%)					
Emergency hospitalizations	25 (20.8)	10 (16.7)	15 (25)	–33	.26
Deaths	7 (5.8)	1 (1.7)	6 (10)	–67	.05
Entry into nursing homes or facilities	9 (7.5)	3 (5)	6 (10)	–50	.30
Patients who left home	16 (86.7)	4 (6.7)	12 (20)	–67	.03

Probability of Staying at Home

The survival analysis for the probability to stay at home was 68% (95% CI 54.5%-78.3%) for the control group versus 80% (95% CI 67.5%-88.1%) for the intervention group (*P*=.15).

Costs

Cost of Follow-Up

Eleven individuals benefited from a consultation with a general practitioner following an alert issued by the PRESAGE CARE system. These consultations were considered additional costs attributable to the use of the device. The time required for nurses to monitor patients in the intervention arm totaled 119 hours (equivalent to 2 minutes and 40 seconds per patient per week), resulting in a total burden cost of €3460 [16]. Training for HCAs was conducted in a group setting over 1 hour, representing a cost of €1334 for 30 HCAs. Training for the coordinating nurse lasted 5 hours, with an associated cost of €200. The total cost related to training time and device usage was estimated at €4994.

Hospital Costs

All causes of hospitalization and their durations were recorded and categorized according to the relevant homogeneous group of patients. The median length of stay per hospitalization was 9.8 (IQR 4.73-14.87) days in the intervention group, compared to 13.61 (IQR 6.70-20.52) days in the control group,

XSL•FO
RenderX

representing a 67% difference. Over the course of the year, the total number of hospitalization days was 6 times higher in the control group than in the intervention group, with 882 days versus 147 days, respectively. The annual cost attributed to emergency hospitalizations, based on data from the homogeneous group of patients, was €58,975 in the intervention group compared to €108,642 in the control group, reflecting a 45.72% reduction in the intervention group (Figure 1). The total cost for the intervention group was €64,300, compared to €108,642 in the control group, representing a 40.82% cost reduction (Table 5).

Figure 1. Emergency hospitalizations' costs evolution (cumutive cost, 1€ = US $ 1.09).

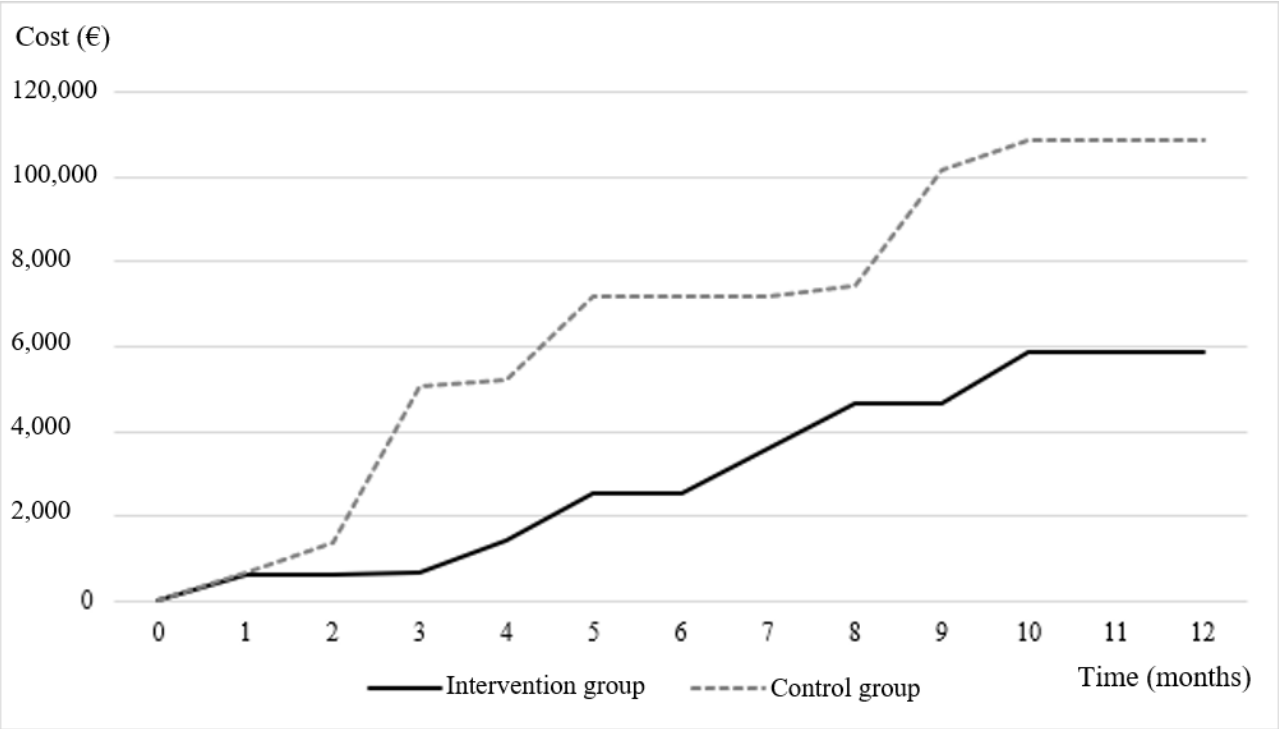

Table 5. Costs analyses.

	Intervention group	Control group	Difference between groups (%)	*P* value
Length of stay (days), median (IQR)	9.8 (4.73-14.87)	13.61 (6.70-20.52)	–28	.41
Total number of days of emergency hospitalizations, n	147	313	–53	.09
Annual emergency hospitalization costs (€)[a]	58,976	108,641	–45	.20
Cost for additional consultations with the attending physician (€)	330	__[b]	—	
Cost for data collection (training cost for HCAs[c]; €)	1334	—	—	
Cost for coordinating nurse (follow-up and alert management; €)	3460	—	—	
Cost for coordinating nurse (training cost; €)	200	—	—	
Total annual cost (€)	64,300	108,641	–41	.25
Mean cost by patient (€)	1071	1810	—	

[a]€1=US $1.09.

[b]Not applicable.

[c]HCA: home care aide.

Discussion

Principal Findings

The objective of this study was to assess the medico-economic impact of an eHealth device designed to predict and prevent emergency hospitalizations among frail older adults living at home.

The use of the PRESAGE CARE device led to a nearly 32% reduction in emergency hospitalizations compared to the control group. Overall, it reduced hospital-related costs by 45.72%. In

XSL•FO
RenderX

addition, this cost reduction must be considered in the context of additional expenses, including in-home care, posthospitalization consultations, and the increased risk of rehospitalization.

The median length of stay in the intervention group was consistent with findings in the literature, whereas it was higher in the control group, likely due to extreme values and the small cohort size.

Furthermore, over a 1-year follow-up of 60 participants, the system extended the mean time to failure in the care pathway by 38 days. This result is likely underestimated given the limited cohort size.

In the intervention arm, among the 726 visits not followed by an alert, only 4 (0.6%) hospitalizations followed the visit ($P<.001$), which confirm the relevance of the alerts issued by the system. Comparison of the postintervention hospitalization rate in the intervention group shows that an alert-triggered intervention significantly reduces hospitalization. These results are consistent with a previous study on this system [15]. Professionals can effectively anticipate a deteriorating health condition and act at the right time.

The multiplication of technological tools to promote home care rarely benefit from data on their real medico-economic benefits [17,18]. Moreover, few studies on the economic evaluation of telemedicine systems show a medico-economic gain [19-22]. In addition, to our knowledge, they have not evaluated the effectiveness of a predictive system for the prevention of unplanned hospitalizations of older people living at home.

Cost of the Device

The cost of the device has not been taken into account in this study as it is currently being evaluated in terms of its impact. The price of the health care sector (including hospital care) varies widely from one country to another. France remains 40% cheaper than the United States and 55% cheaper than Switzerland [23].

Increased Cost of Dependency and Quality of Life

According to the report of the Hospital and Older People Workshop published in 2018, the incidence of dependence related to hospitalization varies between 30% and 60% in people aged 70 years and older and increases to 50% in patients aged 85 years and older. After discharge from hospital, only 50% of patients recover their basic functional state, 33% within 6 months after discharge, and 14% at 1 year [10,19]. The prevalence of iatrogenic dependence is about 12%, and it is avoidable in 80% of cases. It is therefore possible to prevent it.

Moreover, after a visit to the ED, 56% of older patients are hospitalized, and 44% return home. The difficulty in finding suitable hospital places for older adults leads to an increase in the time spent by carers on calls to other departments. Older people spend more time in the ED, on average 4 hours more [10].

These data increase the medical and financial impact of devices aimed at predicting and preventing ED visits and emergency hospitalizations.

The increase in the loss of autonomy of older adults after hospitalization (bed rest, loss of mobility, and waiting time in the ED) leads to an increase in home care needs and increases the costs associated with the management of dependency.

This increase in posthospitalization dependence leads to a decrease in quality of life, which has been widely described in the literature [24].

This device represents a real gain in terms of medical benefits, quality of life, and hospital costs.

Limitations

Our study has several limitations. First, because the overall study design was not randomized at the outset of the study, there is a bias in the allocation of groups. Moreover, the number of deaths was higher in the control group, but the sample size did not allow us to identify a significant correlation with the intervention. The lack of significance of the different results can indeed be explained by the size of the cohort, the number of events, and the low statistical power. Some particularly long hospital stays imply a wide CI, impacting the significance of the results.

In view of the number of deaths and the lack of consensus on the relevance of using quality-adjusted life-year analyses to carry out medico-economic studies of dependent or frail older people (National Academy of Medicine) [21,24], it was not possible to carry out a quality-adjusted life-year study based on life expectancy.

Furthermore, the quality of life of the individuals could not be measured and compared. It is not possible to draw conclusions on the impact of PRESAGE CARE on the quality of life of beneficiaries. A prospective study will be conducted to evaluate this impact.

The size of the cohort is also a limitation and will require a larger-scale, multicenter, randomized controlled trial.

Conclusion

This first study evaluating the medico-economic impact in real life of the use of the PRESAGE CARE medical device shows a reduction of almost half of the hospital expenses among the beneficiaries of the device. In addition, PRESAGE CARE reduced unplanned hospitalizations by more than 30% and increased life expectancy in good health at home. A larger, randomized controlled trial is needed to confirm these already very encouraging results.

Acknowledgments

The authors would like to thank the care managers and staff of the home care facility (Le Connétable) for their contribution to the study, as well as GHT Rance-Emeraude and Regional Agency for Health (ARS) for funding this study.

XSL•FO
RenderX

Data Availability

The data that support the findings of this study are available from the corresponding author, JHV, upon reasonable request. Restrictions apply to the availability of the data collected by the home care aids and the algorithm that was used in this study, which is the subject of a patent.

Authors' Contributions

Each author met the International Committee of Medical Journal Editors criteria for authorship. JHV conceptualized the study; CHT conducted the acquisition of data; CHT and JHV contributed to the interpretation of the results; JHV and CHT were in charge of the formal analysis; and JB, JHV, CHT, FD, and AF drafted and revised the manuscript. All authors approved the final version of the manuscript.

Conflicts of Interest

JHV is the chief executive officer of PRESAGE CARE. CH-T is employee of PRESAGE.

References

1. Larbi K, Roy D. 4 Millions de seniors seraient en perte d'autonomie en 2050. 4 Million seniors will be losing their autonomy in 2050 [Article in French]. Insee Première. 2019 Jul 25. URL: https://www.insee.fr/fr/statistiques/4196949 [accessed 2025-03-10]
2. Uscher-Pines L, Pines J, Kellermann A, Gillen E, Mehrotra A. Emergency department visits for nonurgent conditions: systematic literature review. Am J Manag Care 2013 Jan;19(1):47-59 [FREE Full text] [Medline: 23379744]
3. Boisguérin B, Mauro L. Les personnes âgées aux urgences : une patientèle au profil particulier. Older people in the emergency department: a patient base with a particular profile [Article in French]. Direction de la Recherches, des Etudes, de l'Evaluation des Statistiques (DREES). 2017 Mar 24. URL: https://drees.solidarites-sante.gouv.fr/publications/etudes-et-resultats/les-personnes-agees-aux-urgences-une-patientele-au-profil [accessed 2025-03-10]
4. Duong HV, Herrera LN, Moore JX, Donnelly J, Jacobson KE, Carlson JN, et al. National characteristics of emergency medical services responses for older adults in the United States. Prehosp Emerg Care 2018 Sep 01;22(1):7-14 [FREE Full text] [doi: 10.1080/10903127.2017.1347223] [Medline: 28862480]
5. Prince MJ, Wu F, Guo Y, Gutierrez Robledo LM, O'Donnell M, Sullivan R, et al. The burden of disease in older people and implications for health policy and practice. Lancet 2015 Feb 07;385(9967):549-562. [doi: 10.1016/S0140-6736(14)61347-7] [Medline: 25468153]
6. Analyse de l'activité hospitalière 2018 - Focus sur les patients âgés de 80 ans et +. Hospital activity analysis 2018 - focus on patients aged 80 and over [Article in French]. Agence technique de l'information sur l'hospitalisation. 2018. URL: https://www.atih.sante.fr/analyse-de-l-activite-hospitaliere-2018-focus-sur-les-patients-ages-de-80-ans-et [accessed 2025-03-10]
7. Boisguérin B, Mauro L. Les personnes âgées aux urgences : une santé plus fragile nécessitant une prise en charge plus longue. Elderly people in the emergency room: a more fragile health requiring a longer care [Article in French]. Direction de la Recherches, des Etudes, de l'Evaluation des Statistiques (DREES). 2017 Mar 24. URL: https://drees.solidarites-sante.gouv.fr/publications/etudes-et-resultats/les-personnes-agees-aux-urgences-une-sante-plus-fragile [accessed 2025-03-10]
8. The care of the older people in the health care system: progress too slow. Executive Summaries: Annual Public Report of the Cour des Comptes. Paris, France: Cour des Comptes; 2012. URL: https://www.ccomptes.fr/sites/default/files/EzPublish/Annual_public_report_2012_summaries.pdf [accessed 2025-03-10]
9. Roussel R. Personnes âgées dépendantes : les dépenses de prise en charge pourraient doubler en part de PIB d'ici à 2060. Dependency of older people: care expenditure could double as a share of GDP by 2060 [Article in French]. Direction de la Recherches, des Etudes, de l'Evaluation des Statistiques (DREES). 2017 Oct 12. URL: https://drees.solidarites-sante.gouv.fr/publications/etudes-et-resultats/personnes-agees-dependantes-les-depenses-de-prise-en-charge [accessed 2025-03-10]
10. Libault D. Rapport de la concertation grand âge et autonomie. Report on old age and autonomy [Article in French]. Ministry of Solidarity and Health (France). 2019 Mar 28. URL: https://tinyurl.com/3b4mekcn [accessed 2025-03-10]
11. Adams JG. Emergency department overuse: perceptions and solutions. JAMA 2013 Mar 20;309(11):1173-1174. [doi: 10.1001/jama.2013.2476] [Medline: 23512065]
12. Barnett K, Mercer SW, Norbury M, Watt G, Wyke S, Guthrie B. Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. Lancet 2012 Jul 07;380(9836):37-43 [FREE Full text] [doi: 10.1016/S0140-6736(12)60240-2] [Medline: 22579043]
13. Denis F, Krakowski I. How should oncologists choose an electronic patient-reported outcome system for remote monitoring of patients with cancer? J Med Internet Res 2021 Sep 09;23(9):e30549 [FREE Full text] [doi: 10.2196/30549] [Medline: 34499046]
14. Veyron J, Friocourt P, Jeanjean O, Luquel L, Bonifas N, Denis F, et al. Home care aides' observations and machine learning algorithms for the prediction of visits to emergency departments by older community-dwelling individuals receiving home

XSL•FO
RenderX

care assistance: a proof of concept study. PLoS One 2019 Aug 13;14(8):e0220002 [FREE Full text] [doi: 10.1371/journal.pone.0220002] [Medline: 31408458]
15. Belmin J, Villani P, Gay M, Fabries S, Havreng-Théry C, Malvoisin S, et al. Real-world implementation of an eHealth system based on artificial intelligence designed to predict and reduce emergency department visits by older adults: pragmatic trial. J Med Internet Res 2022 Sep 08;24(9):e40387 [FREE Full text] [doi: 10.2196/40387] [Medline: 35921685]
16. Belmin J, Auffray J, Berbezier C, Boirin P, Mercier S, de Reviers B, et al. Level of dependency: a simple marker associated with mortality during the 2003 heatwave among French dependent elderly people living in the community or in institutions. Age Ageing 2007 May 24;36(3):298-303. [doi: 10.1093/ageing/afm026] [Medline: 17384419]
17. Les salaires dans les établissements de santé. Salaries in health establishments [Article in French]. Direction de la Recherches, des Etudes, de l'Evaluation des Statistiques (DREES). 2021. URL: https://tinyurl.com/4wfd2utv [accessed 2025-03-10]
18. Gentili A, Failla G, Melnyk A, Puleo V, Tanna GLD, Ricciardi W, et al. The cost-effectiveness of digital health interventions: a systematic review of the literature. Front Public Health 2022 Aug 11;10:787135 [FREE Full text] [doi: 10.3389/fpubh.2022.787135] [Medline: 36033812]
19. Rohrbach P, Dingemans A, Evers C, van Furth EF, Spinhoven P, Aardoom J, et al. Cost-effectiveness of internet interventions compared with treatment as usual for people with mental disorders: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2023 Jan 05;25:e38204 [FREE Full text] [doi: 10.2196/38204] [Medline: 36602854]
20. Sülz S, van Elten HJ, Askari M, Weggelaar-Jansen AM, Huijsman R. eHealth applications to support independent living of older persons: scoping review of costs and benefits identified in economic evaluations. J Med Internet Res 2021 Mar 09;23(3):e24363 [FREE Full text] [doi: 10.2196/24363] [Medline: 33687335]
21. Lizée T, Basch E, Trémolières P, Voog E, Domont J, Peyraga G, et al. Cost-effectiveness of web-based patient-reported outcome surveillance in patients with lung cancer. J Thorac Oncol 2019 Jun;14(6):1012-1020 [FREE Full text] [doi: 10.1016/j.jtho.2019.02.005] [Medline: 30776447]
22. Health care prices. Organisation for Economic Co-operation and Development (OECD). 2020 May. URL: https://www.oecd.org/content/dam/oecd/en/publications/reports/2020/05/health-care-prices_d4fbbd0a/bfc42344-en.pdf [accessed 2025-03-10]
23. Prévenir la dépendance iatrogène liée à l'hospitalisation chez les personnes âgée. Preventing iatrogenic dependency related to hospitalization in the elderlys [Article in French]. Haute Autorité de Santé. 2017 Sep. URL: https://www.has-sante.fr/jcms/c_2801190/fr/prevenir-la-dependance-iatrogene-liee-a-l-hospitalisation-chez-les-personnes-agees [accessed 2025-03-10]
24. Gonthier R. La qualité de vie chez le sujet âgé. Quality of life in old age [Article in French]. Bull Acad Natl Med 2007 Feb;191(2):237-244. [doi: 10.1016/s0001-4079(19)33071-7] [Medline: 17969545]

Abbreviations

ED: emergency department
GIR: groupe iso-ressources
HCA: home care aide

Edited by A Mavragani; submitted 27.06.24; peer-reviewed by LC Maclagan, A Gangadhara Rao, D Valvi, A Kalluchi; comments to author 16.09.24; revised version received 21.12.24; accepted 05.03.25; published 11.04.25.

Please cite as:
Havreng-Théry C, Fouchard A, Denis F, Veyron JH, Belmin J
Cost-Effectiveness Analysis of a Machine Learning–Based eHealth System to Predict and Reduce Emergency Department Visits and Unscheduled Hospitalizations of Older People Living at Home: Retrospective Study
JMIR Form Res 2025;9:e63700
URL: https://formative.jmir.org/2025/1/e63700
doi:10.2196/63700
PMID:40215100

©Charlotte Havreng-Théry, Arnaud Fouchard, Fabrice Denis, Jacques-Henri Veyron, Joël Belmin. Originally published in JMIR Formative Research (https://formative.jmir.org), 11.04.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The EmpkinS-EKSpression Reappraisal Training Augmented With Kinesthesia in Depression: One-Armed Feasibility Study

Marie Keinert[1], MSc; Lena Schindler-Gmelch[1], Dr rer nat; Lydia Helene Rupp[1], MSc; Misha Sadeghi[2], MSc; Robert Richer[2], MSc; Klara Capito[1], MSc; Bjoern M Eskofier[2,3], Prof Dr; Matthias Berking[1], Prof Dr

[1]Department of Clinical Psychology and Psychotherapy, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany

[2]Machine Learning and Data Analytics Lab, Department of Artificial Intelligence in Biomedical Engineering (AIBE), Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany

[3]Translational Digital Health Group, Institute of AI for Health, Helmholtz Zentrum Muenchen, Neuherberg, Germany

Corresponding Author:
Marie Keinert, MSc
Department of Clinical Psychology and Psychotherapy
Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU)
Nägelsbachstraße 25a
Erlangen, 91052
Germany
Phone: 49 9131 8567563
Email: marie.keinert@fau.de

Abstract

Background: Harboring dysfunctional depressogenic cognitions contributes to the development and maintenance of depression. A central goal of cognitive behavioral therapy (CBT) for depression is to invalidate such cognitions via cognitive reappraisal (CR). However, relatively low remission rates and high dropout rates in CBT demonstrate the need for further improvement. Potentially, the effects of CBT could be enhanced by addressing not only dysfunctional depressogenic cognitions but also body states associated with depression. This may be done, for example, by systematically pairing the invalidation of depressogenic cognitions with the performance of antidepressive kinesthesia.

Objective: This study aimed to examine the feasibility and clinical potential of a smartphone-based cognitive restructuring task that required users to deliberately perform antidepressive kinesthesia in conjunction with the rejection of depressogenic statements and the affirmation of antidepressive statements. This feasibility study was conducted as a precursor to a large-scale randomized controlled trial.

Methods: In total, 10 healthy participants engaged in a single 90-120-minute session of smartphone-based CR training. During the training, they completed 2 phases in which they were required to reject 20 depressogenic and affirm 20 antidepressive statements, respectively. Diagnostic assessments were conducted 1 week (T1) before and directly prior (T2) to the training, and again directly posttraining (T3) and at a 2-week follow-up posttraining (T4). Feasibility outcomes assessed at T3 included intervention safety recorded by study therapists, compliance, technical feasibility, usability assessed using the Short Version of the User Experience Questionnaire (UEQ-S), and acceptability assessed using the UEQ-S and self-developed items. Preliminary clinical potential was evaluated via single-item ratings of current depressed and positive mood assessed continuously during the training. Feasibility outcomes were analyzed descriptively, and clinical potential was examined using paired-sample *t* tests of pre and post ratings of mood at each training phase.

Results: Overall, the results indicated that the training was safe, feasible, and usable (UEQ-S pragmatic quality scale: mean 1.45, SD 0.71). However, acceptance was limited (UEQ-S hedonic quality scale: mean 1.05, SD 0.79). While 80% (8/10) of the participants were generally satisfied with the training, 80% (8/10) would recommend it to a friend, 90% (9/10) found it interesting, and 80% (8/10) rated it as “leading edge,” 40% (4/10) to 70% (7/10) did not consider it particularly helpful and 50% (5/10) found it repetitive. Preliminary results regarding clinical potential were promising, with significant increases in positive mood (rejection: Hedges g=0.63; affirmation: Hedges g=0.25), whereas changes in depressed mood were not significant.

Conclusions: This study evaluated the feasibility and acceptability of a smartphone-based CR training augmented with validating and invalidating kinesthesia. This provided valuable insights for further optimizing the intervention for the subsequent randomized controlled trial, but also potential similar interventions. If future studies confirm their clinical potential, such interventions offer a promising approach to enhancing CBT for depression.

XSL•FO
RenderX

Trial Registration: OSF Registries pw6ma; https://osf.io/pw6ma/

(JMIR Form Res 2025;9:e65357) doi:10.2196/65357

KEYWORDS

depression; cognitive reappraisal; facial expression; kinesthesia; smartphone-based intervention; mobile phone

Introduction

Depression is one of the most prevalent mental disorders [1] and severely impacts the quality of life and functioning of affected individuals [2]. Depression is also associated with an increased risk of cardiovascular disease [3] and mortality [4]. Although cognitive behavioral therapy (CBT) has been shown to be effective in depression treatment [5], dropout rates of up to 32 % [6], low remission rates of 42 % [7], and low rates of evidence-based treatment administration [8] highlight the need to further optimize CBT for depression.

CBT for depression is based on Beck's [9] cognitive model of depression, which assumes that dysfunctional beliefs (eg, "I am worthless") are a key factor in the etiology and maintenance of the disease. Therefore, one primary goal of CBT is the reappraisal of such dysfunctional beliefs. This is typically pursued with cognitive restructuring techniques, with which therapists guide patients to question the validity of their dysfunctional beliefs and to articulate more adaptive ones [9]. Theory and empirical evidence suggest that invalidating dysfunctional beliefs while strengthening functional beliefs has positive effects on depressed mood and other symptoms of depression [10,11]. To reduce the large treatment gap in depression, researchers have developed accessible and cost-effective CBT-based app interventions that include cognitive restructuring exercises, among others. Such interventions have proven to be effective in reducing symptoms of depression [12]. However, a meta-analysis evaluating single components of such interventions with individual participant data found no clear evidence for the efficacy of cognitive restructuring when used within digital interventions [13], highlighting the need to optimize digital cognitive restructuring interventions. Most of the mobile cognitive restructuring exercises in the context of depression take a similar approach as therapists in face-to-face therapy. They offer psychoeducational content about the relationship between dysfunctional beliefs and depression and provide Socratic questions to help individuals invalidate dysfunctional beliefs and develop more adaptive ones (eg, [14-19]). However, without the support of a therapist, this approach could be very challenging for patients, especially in the context of cognitive and motivational impairments in depression [20,21]. This is also supported by pilot results indicating lower adherence rates with a cognitive restructuring intervention than with an intervention focusing on behavioral activation [18]. Thus, in order to increase adherence, a structured cognitive restructuring exercise with preselected content might be of greater benefit in the context of app interventions without direct human interaction. To our knowledge, only 2 studies evaluate such an intervention. McCloud et al [22] propose a cognitive restructuring exercise based on the ABC technique [23] providing options for alternative interpretations for users to choose from. Stiles-Shields et al [24] propose a step-by-step procedure for identifying and changing maladaptive thoughts using Socratic questions, providing examples of possible thoughts and alternatives. However, the efficacy of McCloud et al [22] cognitive restructuring exercise cannot be determined as it was part of a multicomponent intervention. Stiles-Shields et al [24] conducted a pilot trial. Although they did find effects on symptoms of depression, the informative value is limited due to the small sample size. The efficacy of such structured cognitive restructuring exercises therefore still requires empirical support.

A potential problem of such exercises, particularly if they contain preselected content, could be that the invalidation of dysfunctional beliefs remains superficial and does not lead to emotional insight. This is also assumed by the interacting cognitive subsystems (ICS) theory [25], another seminal theoretical approach in the context of depression. The theory distinguishes intellectual and emotional beliefs and posits that cognitive restructuring only changes dysfunctional beliefs on an intellectual level, but not their "felt meaning." This is often experienced by patients who report, "I understand the belief is not true, but I still feel that it is."

ICS suggests that depressogenic thoughts and proprioceptive information interact to generate emotional states in general and depressed mood in particular [25]. Semantic concepts and specific meanings of depressogenic thoughts exist in the form of intellectual beliefs. To generate specific emotional states (eg, depressed mood) and thereby emotional beliefs, however, input from acoustic, visual, and proprioceptive channels (eg, facial expressions, gestures, and posture) is needed. Hence, according to ICS, if patients merely question dysfunctional beliefs on an intellectual level, they will still feel that they are true. To change the resulting emotional state (ie, depressed mood), proprioceptive information needs to additionally contradict the depressogenic schema, for example, assume an upright body posture and confident facial expression. Thus, the maintenance of depression is explained in terms of a coupling of cognitive and proprioceptive feedback loops that sustain each other, referred to as "cognitive interlock."

In light of these theoretical considerations, we assume that complementing cognitive reappraisal (CR, a key cognitive restructuring technique) with proprioceptions that both invalidate dysfunctional beliefs and validate functional beliefs offers a promising approach to improving CBT for depression. Kinesthesia (ie, proprioception of movement) and, more specifically, facial expressions might be particularly relevant in this context as facial muscles are very sensitive to variations in emotional experience [26].

XSL•FO
RenderX

Empirically, numerous studies provide evidence for a relationship between facial expressions on the one hand, and emotional experience [27] and depression on the other [28-30]. Patients with depression manifest less facial expression of positive affect [28]. Moreover, studies using electromyography (EMG) found greater activity of the corrugator supercilii muscle, a muscle that is activated during negative affective states [31], in response to affective content pictures [32,33], and determined clinical improvement to be associated with attenuated activity of the corrugator supercilii [34]. Consistently, an experimental manipulation of facial expression (ie, inducing a sad facial expression) correlated with negative interpretation bias in healthy adults [35]. The interconnection of depression and facial expression is also reflected in studies using botulinum toxin to block the corrugator supercilii in the treatment of depression, which led to significant and large reductions in symptoms of depression in placebo-controlled studies, as shown by Schulze et al [36] in a recent meta-analysis.

These findings clearly demonstrate the relationship between kinesthesia and depression. However, only very few studies investigated the relationship between kinesthesia and propositional (ie, reasoning) processes in depression. Moreover, to our knowledge, few studies so far have investigated the effects of augmenting cognitive interventions and, more specifically, CR interventions with validating and invalidating kinesthesia. Adaptations of the approach-avoidance modification training (AAMT), which originally required participants to push or pull joysticks to move disorder-relevant images away from or toward themselves [37], use disorder-relevant dysfunctional beliefs as stimuli [38-45]. Two of the studies following such an approach addressed depression (ie, with stimuli such as “I am a failure”), with promising effects on symptoms of depression [38,39]. Moreover, pilot findings from the stress context suggest that using the deliberate display of positive and negative emotions through facial expression, body posture, and a corresponding statement as responses within AAMT might enhance its efficacy [43-45]. Another, albeit not smartphone-based study by O’Toole and Michalak [46] followed an approach more closely aligned with cognitive restructuring in CBT. They evaluated a cognitive restructuring exercise delivered face-to-face combined with emotion-focused body postures and movements and found that it produced greater decreases in agreement with dysfunctional beliefs than isolated cognitive restructuring in healthy individuals.

Although these approaches are promising, evidence remains preliminary, particularly in the context of depression treatment and no study so far has evaluated the efficacy of a smartphone-based structured cognitive restructuring exercise enhanced with validating and invalidating kinesthesia. To advance research in this regard, we developed a structured smartphone-based and sensor-supported CR intervention, in which depressogenic verbalizations are paired with invalidating facial expressions and antidepressive verbalizations are paired with validating facial expressions. Our ultimate goal is to train machine-learning models in the automated assessment of depression, which could then be integrated into a fully automated, smartphone-based biofeedback training. The primary goal of this was to evaluate the feasibility and safety of this intervention in a limited healthy sample by testing its technical setup and procedures before implementation of a large-scale randomized controlled trial (RCT) involving participants with and without depression. The study protocol for the RCT has been published [47]. Second, we aimed to gain preliminary insights into the clinical potential of the intervention, expecting an overall decrease in depressed mood and an increase in positive mood throughout the training and beyond.

Methods

Design

In preparation for a large-scale RCT [47], we conducted an experimental feasibility study with a repeated-measures, within-subject design. The study is part of the Collaborative Research Center EmpkinS (Empatho-Kinaesthetic Sensor Technology – Sensor Techniques and Data Analysis Methods for Empatho-Kinaesthetic Modeling and Condition Monitoring) and was conducted at the EmpkinS Lab of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU). Participants engaged in a 90-120-minute CR training session augmented with antidepressive facial expression (AFE) prompts on a study-provided smartphone (CR+AFE condition, ie, 1 of the 4 study conditions of the bigger RCT [47]). The study protocol did not involve blinding of either participants or investigators. The study was registered in the Open Science Framework [48] and we have adhered to the American Psychological Association journal article reporting standards [49] when writing the manuscript (Multimedia Appendix 1).

Participants

Participants were recruited between August 2022 and January 2023. Inclusion criteria were (1) a minimum age of 18 years and (2) sufficient German language skills (at least level B2 according to the Common European Framework of Reference for Languages). Exclusion criteria were (1) a Patient Health Questionnaire (PHQ)-8 score of ≥4, (2) sufficient criteria for any mental disorder (*International Classification of Diseases and Related Health Problems, Tenth Revision* codes F1-F9), (3) acute suicidality, (4) any impairment of facial expression, and (5) dyschromatopsia. We aimed to include N=10 participants, following heuristics from the literature concerning optimal sample sizes for feasibility studies [50,51] and internal considerations of practicality and time constraints. Sociodemographic and clinical baseline characteristics are summarized in Table 1.

XSL•FO
RenderX

Table 1. Sociodemographic and clinical baseline characteristics of participants (N=10) in the feasibility study investigating the single-session EmpkinS-EKSpression Reappraisal Training against depression in healthy individuals.

Categories	Values
Age (years; range 18-41), mean (SD)	24.00 (6.31)
Gender: female, n (%)	10 (100)
Education, n (%)	
> 10 years	4 (40)
University degree	6 (60)
Occupation, n (%)	
Student	9 (90)
Employed	1 (10)
Relationship status: single	10 (100)
Diagnosis of depression, n (%)	
Former	1 (10)
Current	0 (0)
Diagnosis of other mental disorders, n (%)	
Former	2 (20)
Current	0 (0)
Psychotherapeutic or psychiatric treatment	
Former, n (%)	4 (40)
Current, n (%)	0 (0)
PHQ-8[a], mean (SD)	2.10 (0.99)

[a]PHQ: Patient Health Questionnaire.

Procedures

Participants from the general population were recruited via social media posts targeting the Erlangen area, emails sent via student distribution lists of the FAU, and flyers posted in public places throughout the city of Erlangen. The recruitment material contained information about the conditions for participation, the duration of the study, and the compensation as well as a QR code and link to the study website for registration. The website hosted by Unipark provided information on study aims and procedures. Interested persons were directed to the study screening questionnaire assessing inclusion and exclusion criteria, as well as symptoms of depression with the Patient Health Questionnaire (PHQ)-8 [52] and contact information. Eligible participants were contacted via email or telephone and invited to a diagnostic session (T1). All on-site study sessions were held at the EmpkinS Lab of the FAU. The laboratory rooms were located in a scientific university building and set up specifically for this study. At the beginning of the diagnostic session participants provided written informed consent. We then conducted the Structured Clinical Interview for DSM-5 Disorders-Clinician version (SCID-5-CV; German version: [53]) to assess the diagnostic status of participants. After the interview, participants answered a set of questionnaires via Unipark.com (T1).

One week after the diagnostic session, the training session was held at the Lab. The entire session took about 3 hours. First, we conducted the GRID Hamilton Depression Rating Scale (GRID-HAMD; [54]) to assess depression. Next, participants answered a web-based preassessment (T2), also via Unipark.com and the physiological data acquisition was set up. Then, participants received a short psychoeducation explaining the interconnection of thoughts, mood, and the body. The subsequent smartphone-based training session took approximately 90-120 minutes. After completing the training, a short postassessment (T3) was completed, again via Unipark.com.

Two weeks after the training session, participants were invited to a follow-up assessment (T4). The GRID-HAMD was conducted via telephone and participants were asked to fill out an online questionnaire via Unipark.com. They received the link to the questionnaires via email one day before the telephone interview.

The interviews and study sessions were conducted by trained clinical psychologists and based on manuals outlining study procedures and instructions for participants.

Smartphone Intervention

During the training, participants completed four phases (Figure 1): (1) the preparatory phase, (2) the depressed mood induction phase, (3) the first training phase, and (4) the second training phase. During the preparatory phase, participants were shown a fixation cross (3 sec) followed by different stimuli (8 sec) in randomized order on the smartphone. The stimuli consisted of

XSL•FO
RenderX

self-recordings of the participants, words congruent and incongruent with depressed mood (ie, “hopeless” and “happy”), and words representing the so-called cognitive triad of depression (ie, “me,” “the world,” and “the future” [55]), which were each shown 4 times, resulting in 24 stimuli in total. Subsequently, in order to allow working with current depressed mood during the training, participants received a validated mood induction beforehand. During this mood induction phase, participants read aloud ten depressogenic cognitions shown on the smartphone screen (eg, “My future is absolutely hopeless”). In addition, sadness-inducing music was played (an excerpt from “Adagio in G minor” by Tomaso Giovanni Albinoni). The induction procedure was previously validated [56,57].

Figure 1. Intervention procedure of the single-session EmpkinS-EKSpression Reappraisal Training against depression was examined in this feasibility study in healthy individuals.

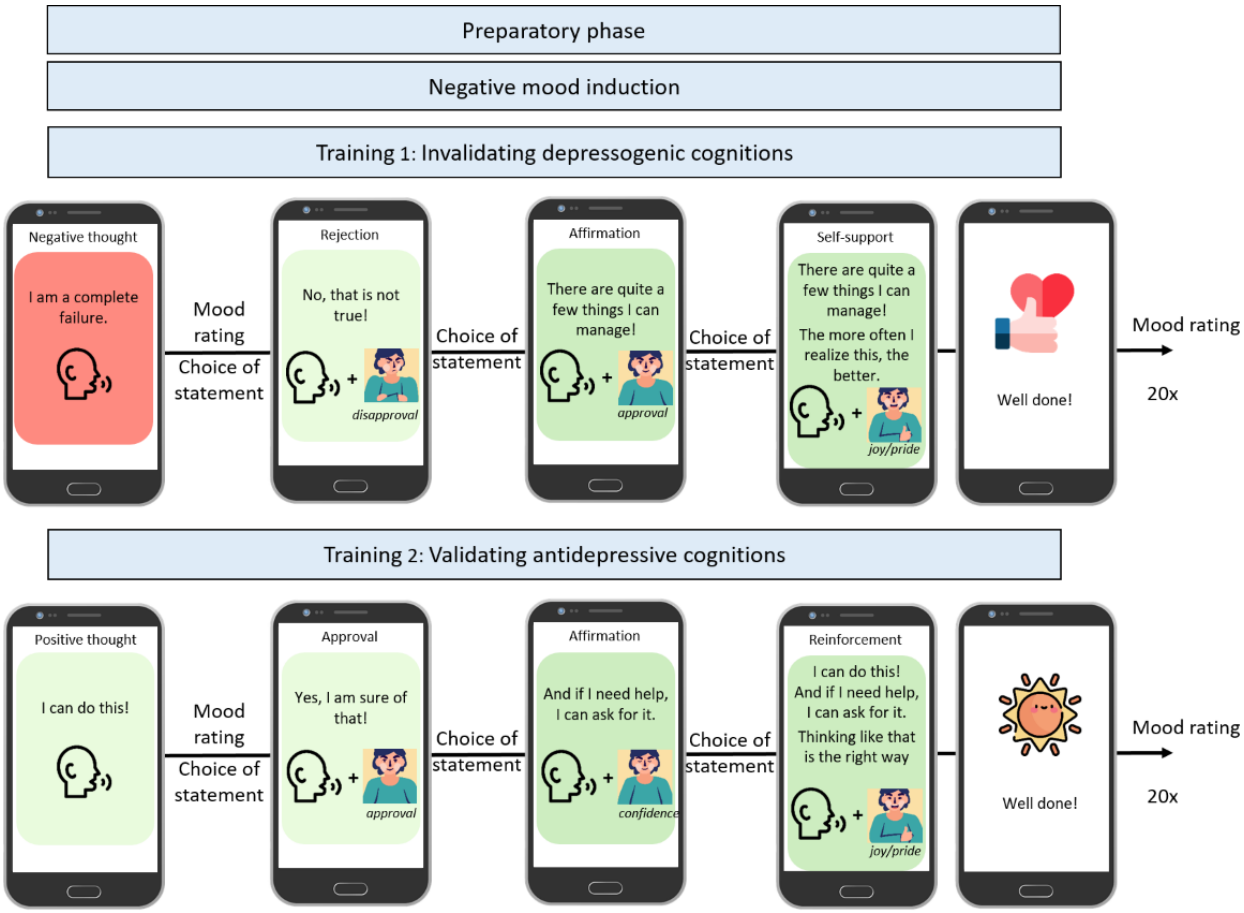

Both training phases started with the study therapist providing instructions on how to engage in the training, followed by on-screen instructions with text and animations in the EmpkinS-EKSpression app. Then, participants completed 2 practice trials before starting the actual training. In the actual training, participants were confronted with 20 depressogenic statements (eg, “I am a complete failure”; first training phase) and antidepressive (eg, “I can do this;” second training phase) statements presented in randomized order on the smartphone screen. The statements were derived from validated questionnaires in the context of depression (Beck Hopelessness Scale, German version [58]; Cognitive Triad Inventory, German version [59]; Automatic Thoughts Questionnaire, German version [60]; and Rosenberg Self-Esteem Scale, German version [61]). A team of 3 experts in clinical psychology adapted the items to customize them for the intervention.

In the first training phase, participants were asked to reject the depressogenic statements in the following three invalidation steps: (1) reject the presented cognition, (2) affirm the rejection, and (3) self-support themselves. In each step, participants were able to choose from three statements (ie, 3 statements expressing rejection, affirmation, and self-support, respectively) which they had to manually select on the screen and also read aloud. Their verbalization was to be amplified by the explicit expression of congruent kinesthesia, including mimic, gesture, and posture. Thus, their rejection of each depressogenic statement was performed with an expression of disapproval, the affirmation with an expression of approval, and self-support with an expression of joy (trials 1-10) and pride (trials 11-20).

In the second training phase, participants received instructions to validate the antidepressive statements in the following sequence: (1) approve the positive cognition with an expression of approval, (2) affirm the approval with an expression of confidence, and (3) reinforce themselves with an expression of joy (trials 1-10) and pride (trials 11-20). Here, to conclude, participants read aloud all 3 statements they had selected.

After each trial, reinforcing statements and pictures were displayed on the screen. Throughout the sequence of training trials, study therapists rated the persuasiveness of participants’

XSL•FO
RenderX

emotional expression, provided feedback to participants, and maintained control of the training app using a tablet.

Measures

Feasibility of the Intervention

To determine the feasibility of the intervention, we assessed safety, technical feasibility, compliance, usability, and acceptability.

Safety of the Intervention

To determine the safety of the intervention, study therapists recorded whether participants experienced significant mood deterioration during the training. The intervention was deemed safe if no more than 10% (1/10) of the training sessions were prematurely terminated due to significant mood deterioration.

Compliance

The compliance rate was defined as the number of participants completing the intervention per protocol. The intervention was considered feasible if the compliance rate was at least 50% (5/10).

Technical Feasibility

Technical difficulties with the EmpkinS-EKSpression app during the training were to be reported by study therapists. We considered the intervention feasible if at least 80% (8/10) of the sessions were completed without major technical problems.

Usability

The usability of the training app was assessed with the pragmatic quality scale of the short version of the User Experience Questionnaire (UEQ-S [62]) and with one self-constructed item. In the UEQ-S, 8 pairs of opposite adjectives related to the user experience are rated on a 7-point Likert scale from –3 (fully agree with the negative term) to +3 (fully agree with the positive term). The pragmatic quality scale consists of 4 of the 8 items, with the other 4 items belonging to the hedonic quality scale. The self-constructed item asked participants how intelligible the app instructions were on a 4-point Likert scale from 1 (not at all) to 4 (very). We considered the intervention usable if >80% (8/10) of participants rated the app as positive (score ≥1), that is, supportive (rather than obstructive), easy (rather than complicated), efficient (rather than inefficient), clear (rather than confusing), and very intelligible.

Acceptability

Acceptability of the intervention was assessed with the "hedonic quality" scale of the UEQ-S. The intervention was considered acceptable if >80% of participants rated the training as positive (score ≥1) that is, interesting (vs not interesting), exciting (vs boring), inventive (vs conventional), and leading edge (vs usual). A total of 12 additional self-generated items were used to assess acceptability (Multimedia Appendix 2 provides the items). Except for 2 items, they were all rated on a 4-point Likert scale from 1 (not at all) to 4 (very). One item asking whether participants would recommend the training to a friend was answered with yes or no. Another item asking whether the training takes a reasonable amount of time was rated on a 5-point Likert scale from 1 (far too high) to 5 (far too low). The intervention was considered acceptable if >80% of participants indicated they would recommend the training to a friend, were rather content with the training, would continue the training given the opportunity, considered the training as rather helpful, and fun (score ≥3), as well as requiring a reasonably acceptable amount of time (score 2-4), and rate the training as only a little or not at all strenuous and difficult to focus on (score ≤2). Moreover, open-ended questions were asked by study therapists during the training and at the T3 assessment to gain additional, qualitative feedback for improvement. Acceptability of the overall study was assessed with one item asking participants to rate how burdensome study participation had been on a 4-point Likert scale from 1 (not at all) to 4 (very). The study was considered acceptable if >80% of participants rated participation as only a little or not at all burdensome.

Clinical Measures

The primary clinical outcome was current depressed mood and the secondary outcome was current positive mood.

Current depressed mood was assessed using an 11-point Likert scale from 0 (no depressed mood at all) to 10 (very strong depressed mood) during both diagnostic and training sessions. During training, participants submitted 108 ratings in total: 24 during the preparatory phase, one after the negative mood induction, one immediately before the first training item, one after each initial statement, and one after the completion of each training trial.

Current positive mood was also assessed using an 11-point Likert-scale from 0 (no positive mood at all) to 10 (very strong positive mood). Similar to the number of ratings of depressed mood, participants submitted a total of 108 ratings of positive mood.

Demographic and Health-Related Information

Demographic and health-related variables included age, date of birth, German language skills, nationality, gender, height, weight, academic degree, professional occupation, relationship status, scholarly fields of study (if any), current or previous diagnosis of a mental illness, psychotherapy, medication, color blindness, and Botox treatment. In addition, the PHQ-8 [52] was used to assess current depressive symptom severity during diagnostic screening. The PHQ-8 consists of 8 items assessing the frequency of 8 of the 9 *Diagnostic and Statistical Manual of Mental Disorders* criteria of depression during the previous 2 weeks on a 4-point Likert scale (0=not at all to 3=nearly every day). The items are added up to a sum maximum score of 24, with a score ≥10 indicating clinically relevant symptom severity [52].

Additional Measures

As one aim of our feasibility study was to test the technical setup of the study and protocol procedures, we included all measures to be assessed in the RCT. However, the additional measures listed below will not be analyzed in this feasibility evaluation. The clinical status of participants was assessed with the SCID-5-CV during diagnostic screening. Symptoms of depression were assessed with the GRID-HAMD [54] during the training session and at follow-up. In addition, self-reported symptoms of depression were assessed with the Centre for

XSL•FO
RenderX

Epidemiological Studies Depression Scale (German version [63]). Current suicidality was assessed with a single self-developed item. Dysfunctional attitudes were assessed with the German short version of the Dysfunctional Attitude Scale, form A [64], and automatic thoughts with the German version of the Automatic Thoughts Questionnaire-Revised [60]. Finally, the emotional state was assessed with the Self-Report Instrument for the Assessment of Emotion-Specific Regulation Skills parts A and B (SEK-ES [65]).

Two video cameras were used to assess kinesthesia (ie, facial expression and body posture): an Azure Kinect depth camera (Microsoft) and a high-resolution RGB camera (Sony SRG-300H), positioned 1.2 meters in front of participants. In addition, bipolar EMG of various facial muscles (ie, corrugator supercilii, zygomaticus major, masseter, orbicularis oculi) and of the shoulder muscle (trapezius) were assessed with the BioPac MP160 system (Biopac Inc). Further psychophysiological measures assessed with the BioPac MP160 system included electrocardiogram, respiration (using a respiration belt), and electrodermal activity. In addition, heart rate and respiration were assessed with radar-based technology [66]. The preparatory phase served as a baseline period for physiological data acquisition. Further, we assessed pupillometry data. Moreover, the smartphone app recorded videos of each trial and participants' digital activity on the app using the smartphone's front camera.

Statistical Analysis Plan

The safety and feasibility of the intervention and collection of kinesthetic and physiological data were determined based on descriptive analyses. Results were compared to predefined threshold values (see above). As a manipulation check, we analyzed whether the negative mood induction successfully elicited an increase in depressed mood and a decrease in positive mood. To this end, a paired-sample *t* test was computed to compare depressed or positive moods before and after the negative mood induction. To analyze the clinical potential of the intervention, we also computed paired-sample *t* tests to compare the mean rating of depressed (dependent variable 1) and positive mood (dependent variable 2) after the presentation of the statements, with the mean depressed and positive mood rating after completion of the trials. We analyzed the clinical potential separately for the 2 training phases. For effect sizes, we computed Hedges *g* as proposed by Lakens [67] for small samples, with Hedges *g* values of 0.2, 0.5, and 0.8 representing small, moderate, or large effects, respectively [68]. Moreover, we implemented 2 linear mixed-effects models with time-point as a fixed effect on level 1 and random intercepts and random slopes to analyze the course of depressed and positive mood ratings over the course of the training. The dependent variables were depressed mood and positive mood, respectively. However, as our data showed too little variance, particularly during the second training phase, the models appeared to not converge. Therefore, we merely analyzed the course of depressed and positive mood on a descriptive level.

The level of significance for all analyses was α=.05. The analyses were conducted with R Studio (version 4.3.2; Posit) [69].

Ethical Considerations

The study followed the Declaration of Helsinki's ethical guidelines and obtained ethical approval from FAU's Ethics Committee (20-443-B, 20-443_1-B, 20-443_2-B). All participants provided written informed consent at 2 time points during the study: (1) for the screening at the beginning of the online questionnaire via Unipark and (2) for the rest of the study at the beginning of T1. To deidentify questionnaire and app data, each participant received a personal code. A list connecting personal information with the code was stored on a secure server and accessible by study staff only. Video and audio data collected during the study that could not be completely anonymized were stored on an external hard drive that was locked away at all times.

Participants received compensation up to €40 (€10 for the diagnostic session, €20 for the training session, €5 for the follow-up assessment, and a €5 bonus for completing the whole study; a currency exchange rate of 1€=US $1.09 is applicable). Psychology students at FAU could alternatively receive course credits.

Results

Feasibility of the Intervention

The study flow is displayed in Figure 2. Regarding safety and compliance, none of the participants terminated the training prematurely. Technical difficulties with the EmpkinS-EKSpression app occurred in 2 training sessions: one of them had to be terminated immediately after the start due to an error in the app, leading to the drop-out of that participant (Figure 2). The other session was terminated after the first training phase also due to an error in the app. Here, the data recorded by the smartphone app (including mood ratings delivered throughout the training) were missing. Therefore, this participant was not included in the analysis of clinical potential. However, the data were included in the feasibility analyses.

With regards to usability as assessed with the pragmatic quality scale of the UEQ-S, favorable results emerged. Except for the instrument's "complicated – easy" item, all items were rated positively (Table 2). The total score of the pragmatic quality scale was a mean of 1.45 (SD 0.71). A total of 70% (7/10) of participants assessed the overall usability positively. Ratings for our self-developed item on the intelligibility of the app instructions ranged from 3 to 4 with a median of 4 (mean 3.9, SD 0.32). 90% (9/10) of participants rated the instructions as very intelligible.

XSL•FO
RenderX

Figure 2. Consolidated Standards of Reporting Trials (CONSORT) flow diagram of participants in the feasibility study investigating the single-session EmpkinS-EKSpression Reappraisal Training against depression in healthy individuals from screening to analysis.

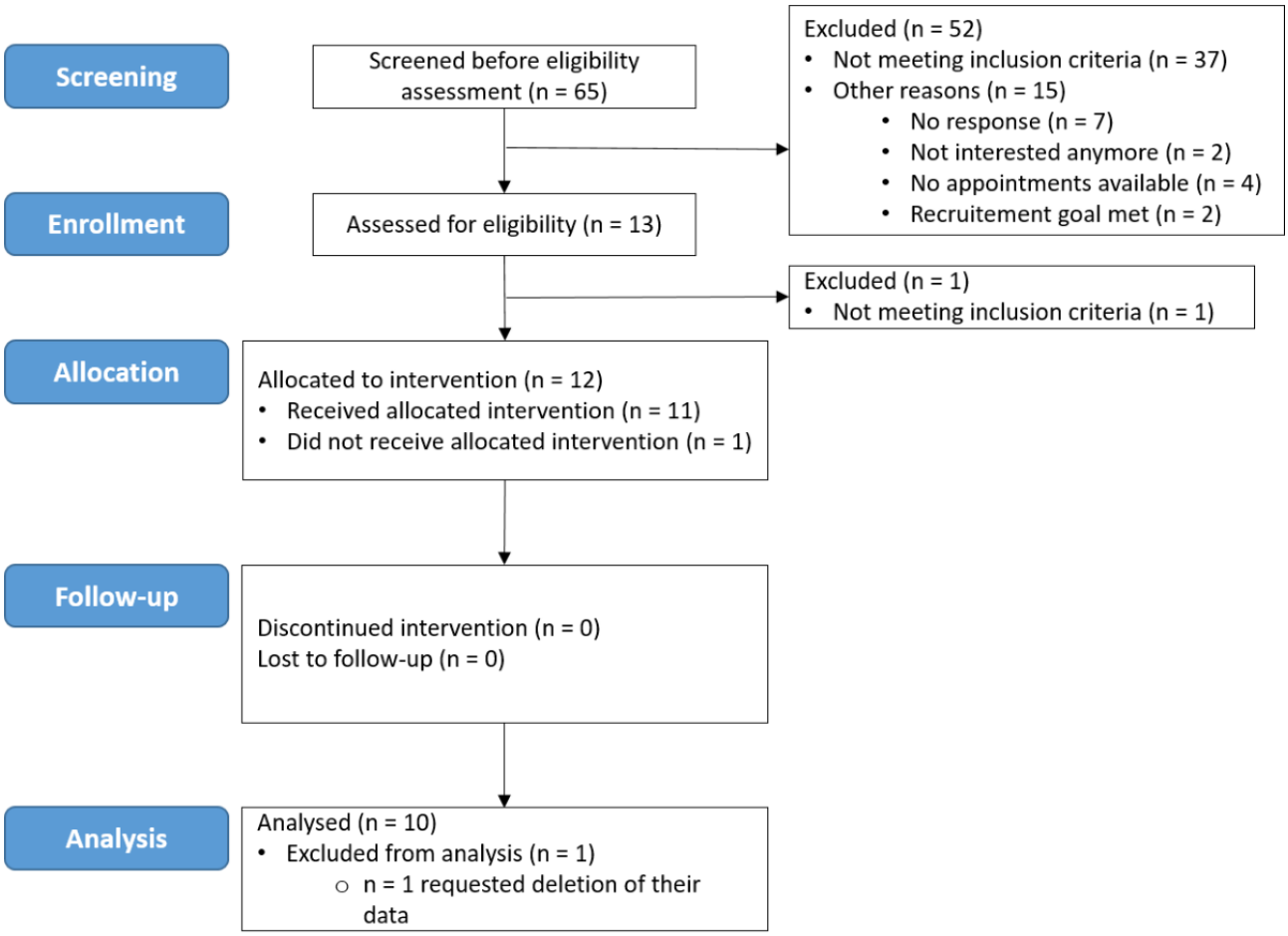

Table 2. Descriptive statistics of the short version of the User Experience Questionnaire assessing the usability of the single-session EmpkinS-EKSpression Reappraisal Training against depression examined in this feasibility study in healthy individuals.

	Median score	Score range	Scores, mean (SD)	Participants (N=10) with scores meeting benchmark[a], n (%)
Pragmatic quality scale				
Obstructive–supportive	2	0 to 3	1.8 (0.79)	9 (90)
Complicated–easy	2	–3 to 3	1 (2.11)	6 (60)
Inefficient–efficient	1	–1 to 2	1.1 (0.99)	8 (80)
Confusing–clear	2	–1 to 3	1.9 (1.29)	9 (90)
Hedonic quality scale				
Not interesting–interesting	1.5	0 to 3	1.5 (0.85)	9 (90)
Boring–exciting	0.5	–1 to 2	0.2 (1.14)	5 (50)
Conventional–inventive	1.5	–2 to 3	1.1 (1.45)	7 (70)
Usual–leading edge	2	–2 to 3	1.4 (1.51)	8 (80)

[a]The predefined benchmark for the items considered to be rated as positive was a score ≥1. The percentage of participants meeting this benchmark is displayed here.

For acceptability, as assessed with the hedonic quality scale of the UEQ-S, results were more mixed (Table 2). While the total score was a mean of 1.05 (SD 0.79), only 50% (5/10) of participants gave a positive total score for the hedonic quality scale. Regarding our self-developed items, 80% (8/10) of participants indicated they would recommend the training to a friend. Further descriptive statistics are displayed in Table 3. In the open evaluation of the training, 4 participants revealed that they found the training repetitive; 2 found it exhausting. A total of 5 participants found the emotional display, 3 pride in

XSL•FO
RenderX

particular, initially unfamiliar and difficult to engage in. Moreover, 5 participants were critical that the standardized statements were not better tailored for individuals, and 3 were dissatisfied with the study setting (ie, uncomfortable seating position, feeling observed by the study therapist, and awkward electrodes). In contrast, some positive feedback included that the training was fun and interesting (n=4), that it provided new ideas for dealing with negative thoughts (n=4), and that there were several statements to choose from (n=1). Regarding the acceptability of the study, participation burden ratings ranged from 1 to 2, with a median of 1 (mean 1.1, SD 0.2), hence, 100% (10/10) of participants rated the participation as only a little or not at all burdensome.

Table 3. Descriptive statistics of self-constructed items assessing the acceptability of the single-session EmpkinS-EKSpression Reappraisal Training against depression examined in this feasibility study in healthy individuals (N=10).

	Median score	Score range	Scores, mean (SD)	Participants (N=10) with scores meeting benchmark[a], n (%)
Likelihood of continuation	2	1 to 3	1.9 (0.74)	2 (20)
Satisfaction in general	3	2 to 4	2.9 (0.57)	8 (80)
Helpfulness				
Training part 1	2	1 to 4	2.4 (0.84)	4 (40)
Training part 2[b]	3	2 to 3	2.67 (0.5)	6 (60)
Training in total	3	2 to 3	2.6 (0.52)	6 (60)
Modifying negative thoughts	2	2 to 3	2.3 (0.48)	3 (30)
Coping with negative thoughts	2	2 to 4	2.5 (0.71)	4 (40)
Enjoyment	2	2 to 3	2.4 (0.52)	4 (40)
Strenuousness	2	2 to 3	2.2 (0.42)	8 (80)
Reasonable amount of time	2	2 to 4	2.4 (0.7)	3 (30)
Difficulty to focus	1	1 to 2	1.3 (0.48)	10 (100)

[a]The predefined benchmark for the items considered to be rated as positive was a score ≥3 for the likelihood of continuation, satisfaction in general, helpfulness, and enjoyment; a score between 2 and 4 for a reasonable amount of time; and a score≤ 2 for strenuousness and difficulty to focus.
[b]N=9.

Manipulation Check

With regards to negative mood induction, the mean rating of depressed mood was 0.5 (SD 1.41) before and 1.44 (SD 0.73) after mood induction. The mean rating of positive mood was 5.67 (SD 1.66) before and 4.56 (SD 1.88) after negative mood induction. Both paired-samples *t* tests reached significance (t_7=–3.06, *P*=.02, Hedges *g*=–0.69, 95% CI –2.83 to –0.28 and t_8=3.16, *P*=.01, Hedges *g*=–0.95, 95% CI –1.84 to –0.55; respectively).

Clinical Potential

In the first training phase, the mean rating for depressed mood was 2.63 (SD 2.51) before and 2.41 (SD 2.36) after, and the mean rating for positive mood was 4.14 (SD 2.42) before and 5.59 (SD 1.65) after the explicit rejection of depressogenic statements, respectively. The difference was not significant for depressed mood (t_{178}=1.16, *P*=.25, Hedges *g*=0.09, 95% CI –0.07 to 0.23), but did reach significance for positive mood (t_{179}=–8.53, *P*<.001, Hedges *g*=–0.63, 95% CI –0.72 to –0.55). In the second training phase, the mean rating of depressed mood was 2.75 (SD 2.47) before and 2.74 (SD 2.48) after, and the mean rating of positive mood was 5.62 (SD 1.71) before and 5.81 (SD 1.58) after the explicit substantiation of antidepressive statements, respectively. The difference was not significant for depressed mood (t_{178}=–0.16, *P*=.87, Hedges *g*=–0.01, 95% CI –0.16 to 0.12), but did, again, reach significance for positive mood (t_{177}=–3.38, *P*<.001, Hedges *g*=–0.25, 95% CI –0.38 to –0.11]). Descriptively, there was a visible mood change after the negative mood induction (ie, an increase in depressed mood, and a decrease in positive mood) and an increase in positive mood after each trial, particularly during the first training phase (Figure 3). Moreover, there was a slight increase in depressed mood over the course of the first training phase, and in positive mood over the course of the second training phase.

XSL•FO
RenderX

Figure 3. Subjective depressed and positive mood ratings over the course of the single-session EmpkinS-EKSpression Reappraisal Training against depression were examined in this feasibility study in healthy individuals. Regression lines represent loess curves.

Discussion

Principal Findings

This study aimed to assess the safety, feasibility, and clinical potential of the novel EmpkinS-EKSpression training that leverages the power of AFE to augment the beneficial effects of CR on depression. Overall, the intervention proved to be safe and feasible, and findings on its clinical potential were promising. However, we identified several aspects that warrant further clinical and technical optimization.

With regard to feasibility, the results were quite positive, and the predefined benchmarks were met for safety, compliance, and technical feasibility. Concerning the usability of the training app as assessed with the pragmatic quality scale of the UEQ-S and a self-developed item on the intelligibility of the app instructions, the predefined benchmark was narrowly missed. This result is worse than the findings of Stiles-Shields et al [24] who evaluated a structured cognitive restructuring exercise where usability was rated to be high. An important, possibly explanatory difference in this study is a higher degree of individualization: participants could choose between self-generated and preselected content in each step of the cognitive restructuring task. It is possible that including more options for individualization of content might improve the usability of the EmpkinS-EKSpression training. Similarly, one of the studies investigating an AAMT intervention in the context of depression reported high usability ratings [39]. The training content here was highly standardized, so a lack of individualization might not be the only explanatory factor for missing the usability benchmark in this study. One difference was the amount of time spent on the cognitive restructuring task (which was not reported by Stiles-Shields and colleagues [24]). While individuals in our study trained for 90-120 minutes, individuals in the study by Lukas and Berking [39] completed an average of around 40 minutes of training. Therefore, a shorter treatment duration might be beneficial to increase usability. When analyzing our usability results at an item level, only the results for the “complicated – easy” item did not meet the benchmark in our study. The variance of the item emerged to be high, and the median indicates that most participants rated the app as easy. Thus, the fact that the benchmark was missed here is likely attributable to singular, extreme votes. Nevertheless, the explicit display of antidepressive kinesthesia might be more challenging than mere cognitive restructuring exercises. Therefore, this finding should be considered in future studies using similar paradigms with rather complicated setups, such as AAMT, to emphasize a user-friendly development of interventions and involve target groups to address their needs, for example, by improving study support or providing more information material. We decided to address this issue in the upcoming RCT by providing a more in-depth oral explanation by study therapists during the training phase. Results on the acceptability of the training, as assessed with the hedonic quality of the UEQ-S, as well as several self-constructed items, did not meet the predefined benchmarks. The open evaluation revealed

XSL•FO
RenderX

that over half of the participants found the training to be too repetitive and too long. This finding is in line with previous work investigating the acceptability of a digital cognitive bias modification training that also required participants to complete many trials, which suggests that participants' willingness to engage in repetitive digital treatments is limited [70]. However, from the ICS theoretical perspective, repeated rehearsal of new and adaptive schemas is important to break the cognitive interlock and thus the maintenance of depressed mood [25]. One possible solution to this dilemma, as suggested by Beard and colleagues [70], is to increase the credibility of the intervention by ensuring a highly comprehensible and convincing treatment rationale and motivation. This suggestion is also in line with a systematic review of barriers and facilitators for digital mental health interventions, which identified credible content of an intervention to be an important facilitator for user engagement [71]. Our feasibility study was likely to fall short of meeting this requirement, given that we only included healthy individuals with perhaps only a limited need for a mood enhancement intervention. Thus, the content of the training was probably not sufficiently relevant for participants, which in turn could have compromised its credibility. However, it is common practice for feasibility studies to initially scrutinize general intervention effects in healthy individuals before exposing more vulnerable clinical participants to those effects [72]. Nevertheless, we took this as an indication to expand our study instructions and elaborate on the study rationale for the subsequent RCT. This approach may also be useful to increase the face validity and thus potentially the effectiveness [73] of other digital mental health interventions, such as structured cognitive restructuring exercises [22,24], smartphone-based AAMT used for CR [38-45], or original AAMT versions used to retrain approach and avoidance biases (Loijen et al performed a systematic review [74]).

The fact that our feasibility sample consisted of healthy individuals might also explain the finding that participants did not consider the training to be particularly helpful. If they did not relate to the standardized depressogenic statements in the first place (which was reflected in participants' open feedback), then there were no negative underlying cognitions to be modified. In terms of strain associated with study participation and training, the benchmark was met for all items. However, 7 participants indicated that the training session was too protracted. Therefore, we decided to abbreviate the preparatory phase of the training to the presentation of only 12 instead of 24 stimuli for 5 instead of 8 seconds. The primary purpose of the preparatory phase was to gather baseline physiological data, which was shown in the feasibility study to be adequately achieved in a shorter time frame. We also decided to move the GRID-HAMD interview to the diagnostic session (T1) to reduce participant burden during the subsequent training session. Another important feedback for the training was that participants expressed difficulty in fully engaging in the spontaneous display of emotion. As a result, the RCT will incorporate increased support to participants from study therapists during the training. One implication for future interventions using antidepressive kinesthesia to augment cognitive restructuring derived from this feedback is that such interventions should be therapist-supported, at least when they are first delivered. Such support can maximize the intervention's effects by leveraging the therapeutic relationship, which is a crucial common factor of psychotherapy [73]. This is also consistent with studies demonstrating that therapist-supported digital interventions for depression are more effective than self-directed interventions [75]. Furthermore, one participant reported that they found the electrodes uncomfortable, which was also witnessed by study therapists. Therefore, for the subsequent RCT, we planned to collect EMG data only for a limited subgroup of participants (for more detail, see the study protocol of the RCT [47]).

Results of the manipulation check suggest that the negative mood induction successfully induced depressed mood with a moderate-to-large effect size (Hedges g=–0.69), and reduced positive mood with a large effect size (Hedges g=–0.95). These results confirm the validity of the mood induction procedure [56,57] and its usefulness in the study of depression interventions. With regards to the clinical potential, rejecting the depressogenic statements in the first training phase and approving the antidepressive statements in the second training phase did not affect the depressed mood. Positive mood, however, did increase with a moderate-to-large effect (Hedges g=–0.63) in the first training phase and a small effect (Hedges g=–0.25) in the second training phase. The null findings for depressed mood may be due to floor effects for our healthy feasibility sample as the mean rating of depressed mood was very low throughout data collection, even after the negative mood induction and the presentation of the depressogenic statements. This is in line with results from the study of O'Toole and Michalak [46], who investigated the effects of augmenting CR with antidepressive body posture and movement. As in our study, healthy individuals were included and no effects on negative emotions were found. Findings from the study by Stiles-Shields et al, however, evaluating the effects of a structured cognitive restructuring exercise, were the opposite, with significant effects on symptoms of depression [24]. With regard to previous studies investigating smartphone-based CR augmented by antidepressive kinesthesia, the findings were similar. One of the studies examining AAMT focusing on CR found significant effects on symptoms of depression [38] (the other reported only descriptive data [39]). These differences may be explained by the fact that Stiles-Shields et al [24] as well as Lukas et al [38] examined individuals with symptoms of depression in their studies and not healthy individuals as in our study. Furthermore, these studies assessed symptoms of depression over the past week, whereas this study assessed the intervention's immediate effect on mood. Moreover, both studies assessed various symptoms of depression, while the current study focused exclusively on depressed mood.

The findings for positive mood, however, align with our hypothesis, suggesting that counteracting depressogenic statements in particular with antidepressive statements and kinesthesia exercises has positive effects on mood in healthy individuals. As such, they are in line with studies showing positive effects of CR on mood, such as those included in a meta-analysis by Ciharova et al [11]. In summary, the results of this and previous studies are promising with regard to the clinical potential of a structured cognitive restructuring exercise as well as the potential of augmenting cognitive restructuring

XSL•FO
RenderX

with antidepressive kinesthesia. These findings are particularly important because structured cognitive restructuring exercises, which also lead to emotional insight through augmentation with antidepressive kinesthesia, could easily be used in mobile interventions without direct human interaction. The types of kinesthesia that are particularly beneficial and the extent to which they augment effects of structured cognitive restructuring on depressed and positive mood as posited by the ICS theory of depression [25] and studies demonstrating effects of manipulating facial expressions on depressed mood [29,30,34,36], will be subject of our subsequent RCT.

Beyond investigating the feasibility and clinical potential of the training, this feasibility study enabled us the opportunity to ensure that the technical aspects of the study (ie, the kinesthetic and physiological data acquisition) ran smoothly. In this respect, we have drawn several implications from the feasibility results. For instance, to improve the kinesthetic data acquisition, we decided to use a second high-resolution RGB camera in the upcoming RCT to record participants' upper bodies and heads viewed from the side. Moreover, after evaluating our data and consulting with experts in the field, contrary to our initial preregistrations (RCT [76]; feasibility study [48]), we decided to drop the electrodermal activity and pupillometry measurements, as well as the EMG measures of the trapezius due to poor data quality and because an improvement of data quality would only have been possible at the cost of disrupting therapeutic processes. Overall, all further issues jeopardizing the quality of kinesthetic and psychophysiological measures could be fixed before the RCTs start, thereby optimizing the development of machine-learning models for automated depression assessment.

Limitations and Conclusion

Major limitations of this feasibility study include the fixed sequence of first invalidating depressogenic cognitions and thereafter validating antidepressive cognitions. This harbors the risk of position effects on depressed and positive moods. Therefore, we plan to randomize the order of the 2 training phases in the RCT. In addition, the intensive contact with study therapists during the training may have prompted socially desirable behavior in participants, particularly regarding compliance with the training and mood ratings submitted during the training. Future studies should include measures of social desirability and ways to control its effects. On the other hand, this contact also represents a strength of the study, as the therapeutic relationship is used to increase the efficacy of the intervention. Despite these limitations and the small sample size, this feasibility study provided valuable insights that led to significant improvements in the study protocol of our upcoming RCT. As such, it serves as a critical keystone in the further development and investigation of innovative interventions for depression.

Acknowledgments

Our foremost gratitude to all study participants. We also thank all the students who provided invaluable support to our team and the EmpkinS-EKSpression app developers. The R code for creating Figure 3 was partially generated using ChatGPT (ChatGPT v2; OpenAI).

Data Availability

The data sets analyzed during this study and the analysis code are available in the Open Science Framework [48].

Conflicts of Interest

None declared.

Multimedia Appendix 1
American Psychological Association journal article reporting standards.
[PDF File (Adobe PDF File), 1037 KB - formative_v9i1e65357_app1.pdf]

Multimedia Appendix 2
Self-developed items to assess usability and acceptability.
[DOCX File , 41 KB - formative_v9i1e65357_app2.docx]

References

1. GBD 2019 Mental Disorders Collaborators. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet Psychiatry 2022;9(2):137-150 [FREE Full text] [doi: 10.1016/S2215-0366(21)00395-3] [Medline: 35026139]
2. Kessler RC, Bromet EJ. The epidemiology of depression across cultures. Annu Rev Public Health 2013;34(1):119-138 [FREE Full text] [doi: 10.1146/annurev-publhealth-031912-114409] [Medline: 23514317]
3. Correll CU, Solmi M, Veronese N, Bortolato B, Rosson S, Santonastaso P, et al. Prevalence, incidence and mortality from cardiovascular disease in patients with pooled and specific severe mental illness: a large-scale meta-analysis of 3,211,768

XSL•FO
RenderX

patients and 113,383,368 controls. World Psychiatry 2017;16(2):163-180 [FREE Full text] [doi: 10.1002/wps.20420] [Medline: 28498599]
4. Cuijpers P, Vogelzangs N, Twisk J, Kleiboer A, Li J, Penninx BW. Comprehensive meta-analysis of excess mortality in depression in the general community versus patients with specific illnesses. Am J Psychiatry 2014;171(4):453-462 [FREE Full text] [doi: 10.1176/appi.ajp.2013.13030325] [Medline: 24434956]
5. Cuijpers P, Berking M, Andersson G, Quigley L, Kleiboer A, Dobson KS. A meta-analysis of cognitive-behavioural therapy for adult depression, alone and in comparison with other treatments. Can J Psychiatry 2013;58(7):376-385. [doi: 10.1177/070674371305800702] [Medline: 23870719]
6. Hans E, Hiller W. Effectiveness of and dropout from outpatient cognitive behavioral therapy for adult unipolar depression: a meta-analysis of nonrandomized effectiveness studies. J Consult Clin Psychol 2013;81(1):75-88. [doi: 10.1037/a0031080] [Medline: 23379264]
7. Cuijpers P, Karyotaki E, Weitz E, Andersson G, Hollon SD, van Straten A. The effects of psychotherapies for major depression in adults on remission, recovery and improvement: a meta-analysis. J Affect Disord 2014;159:118-126. [doi: 10.1016/j.jad.2014.02.026] [Medline: 24679399]
8. Mekonen T, Chan GCK, Connor JP, Hides L, Leung J. Estimating the global treatment rates for depression: a systematic review and meta-analysis. J Affect Disord 2021;295:1234-1242. [doi: 10.1016/j.jad.2021.09.038] [Medline: 34665135]
9. Beck AT, Haigh EAP. Advances in cognitive theory and therapy: the generic cognitive model. Annu Rev Clin Psychol 2014;10(1):1-24. [doi: 10.1146/annurev-clinpsy-032813-153734] [Medline: 24387236]
10. Beck AT, Dozois DJA. Cognitive therapy: current status and future directions. Annu Rev Med 2011;62:397-409. [doi: 10.1146/annurev-med-052209-100032] [Medline: 20690827]
11. Ciharova M, Furukawa TA, Efthimiou O, Karyotaki E, Miguel C, Noma H, et al. Cognitive restructuring, behavioral activation and cognitive-behavioral therapy in the treatment of adult depression: a network meta-analysis. J Consult Clin Psychol 2021;89(6):563-574. [doi: 10.1037/ccp0000654] [Medline: 34264703]
12. Morello K, Schäfer SK, Kunzler AM, Priesterroth LS, Tüscher O, Kubiak T. Cognitive reappraisal in mHealth interventions to foster mental health in adults: a systematic review and meta-analysis. Front Digit Health 2023;5:1253390 [FREE Full text] [doi: 10.3389/fdgth.2023.1253390] [Medline: 37927578]
13. Furukawa TA, Karyotaki E, Suganuma A, Pompoli A, Ostinelli EG, Cipriani A, et al. Dismantling, personalising and optimising internet cognitive-behavioural therapy for depression: a study protocol for individual participant data component network meta-analysis. BMJ Open 2019;8(11):e026137 [FREE Full text] [doi: 10.1136/bmjopen-2018-026137] [Medline: 30798295]
14. Bruhns A, Lüdtke T, Moritz S, Bücker L. A mobile-based intervention to increase self-esteem in students with depressive symptoms: randomized controlled trial. JMIR Mhealth Uhealth 2021;9(7):e26498 [FREE Full text] [doi: 10.2196/26498] [Medline: 34255711]
15. Ham K, Chin S, Suh YJ, Rhee M, Yu ES, Lee HJ, et al. Preliminary results from a randomized controlled study for an app-based cognitive behavioral therapy program for depression and anxiety in cancer patients. Front Psychol 2019;10:1592 [FREE Full text] [doi: 10.3389/fpsyg.2019.01592] [Medline: 31402881]
16. Jannati N, Mazhari S, Ahmadian L, Mirzaee M. Effectiveness of an app-based cognitive behavioral therapy program for postpartum depression in primary care: a randomized controlled trial. Int J Med Inform 2020;141:104145. [doi: 10.1016/j.ijmedinf.2020.104145] [Medline: 32480319]
17. Mantani A, Kato T, Furukawa TA, Horikoshi M, Imai H, Hiroe T, et al. Smartphone cognitive behavioral therapy as an adjunct to pharmacotherapy for refractory depression: randomized controlled trial. J Med Internet Res 2017;19(11):e373 [FREE Full text] [doi: 10.2196/jmir.8602] [Medline: 29101095]
18. Dahne J, Collado A, Lejuez CW, Risco CM, Diaz VA, Coles L, et al. Pilot randomized controlled trial of a Spanish-language behavioral activation mobile app (¡Aptívate!) for the treatment of depressive symptoms among united states latinx adults with limited english proficiency. J Affect Disord 2019;250:210-217 [FREE Full text] [doi: 10.1016/j.jad.2019.03.009] [Medline: 30870770]
19. Moberg C, Niles A, Beermann D. Guided self-help works: randomized waitlist controlled trial of pacifica, a mobile app integrating cognitive behavioral therapy and mindfulness for stress, anxiety, and depression. J Med Internet Res 2019;21(6):e12556 [FREE Full text] [doi: 10.2196/12556] [Medline: 31199319]
20. Nuño L, Gómez-Benito J, Carmona VR, Pino O. A systematic review of executive function and information processing speed in major depression disorder. Brain Sci 2021;11(2):147 [FREE Full text] [doi: 10.3390/brainsci11020147] [Medline: 33499360]
21. Grahek I, Shenhav A, Musslick S, Krebs RM, Koster EHW. Motivation and cognitive control in depression. Neurosci Biobehav Rev 2019;102:371-381 [FREE Full text] [doi: 10.1016/j.neubiorev.2019.04.011] [Medline: 31047891]
22. McCloud T, Jones R, Lewis G, Bell V, Tsakanikos E. Effectiveness of a mobile app intervention for anxiety and depression symptoms in university students: randomized controlled trial. JMIR Mhealth Uhealth 2020;8(7):e15418 [FREE Full text] [doi: 10.2196/15418] [Medline: 32735221]
23. Ellis A. The revised ABC's of rational-emotive therapy (RET). J Ration-Emot Cogn-Behav Ther 1991;9(3):139-172. [doi: 10.1007/BF01061227]

XSL•FO
RenderX

24. Stiles-Shields C, Montague E, Kwasny MJ, Mohr DC. Behavioral and cognitive intervention strategies delivered via coached apps for depression: pilot trial. Psychol Serv 2019;16(2):233-238 [FREE Full text] [doi: 10.1037/ser0000261] [Medline: 30407055]
25. Teasdale JD, Barnard PJ. Affect, Cognition, and Change: Re-modelling depressive thought. Mahwah, NJ: Lawrence Erlbaum Associates; 1993.
26. Ekman P, Friesen WV. Unmasking the Face: A Guide to Recognizing Emotions From Facial Expressions. Englewood Cliffs, NJ: Prentice-Hall; 1975.
27. Coles NA, Larsen JT, Lench HC. A meta-analysis of the facial feedback literature: effects of facial feedback on emotional experience are small and variable. Psychol Bull 2019;145(6):610-651. [doi: 10.1037/bul0000194] [Medline: 30973236]
28. Davies H, Wolz I, Leppanen J, Fernandez-Aranda F, Schmidt U, Tchanturia K. Facial expression to emotional stimuli in non-psychotic disorders: a systematic review and meta-analysis. Neurosci Biobehav Rev 2016;64:252-271. [doi: 10.1016/j.neubiorev.2016.02.015] [Medline: 26915928]
29. Girard JM, Cohn JF, Mahoor MH, Mavadati S, Rosenwald DP. Social risk and depressionvidence from manual and automatic facial expression analysis. 2013 Presented at: 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG); April 26, 2013; Shanghai, China. [doi: 10.1109/fg.2013.6553748]
30. Pampouchidou A, Pediaditis M, Kazantzaki E, Sfakianakis S, Apostolaki IA, Argyraki K, et al. Automated facial video-based recognition of depression and anxiety symptom severity: cross-corpus validation. Mach Vis Appl 2020;31(4):30. [doi: 10.1007/s00138-020-01080-7]
31. Larsen JT, Norris CJ, Cacioppo JT. Effects of positive and negative affect on electromyographic activity over zygomaticus major and corrugator supercilii. Psychophysiology 2003;40(5):776-785. [doi: 10.1111/1469-8986.00078] [Medline: 14696731]
32. Lindsey KT, Rohan KJ, Roecklein KA, Mahon JN. Surface facial electromyography, skin conductance, and self-reported emotional responses to light- and season-relevant stimuli in seasonal affective disorder. J Affect Disord 2011;133(1-2):311-319 [FREE Full text] [doi: 10.1016/j.jad.2011.04.016] [Medline: 21600661]
33. Wexler BE, Levenson L, Warrenburg S, Price LH. Decreased perceptual sensitivity to emotion-evoking stimuli in depression. Psychiatry Res 1994;51(2):127-138 [FREE Full text] [doi: 10.1016/0165-1781(94)90032-9] [Medline: 8022947]
34. Schwartz GE, Fair PL, Mandel MR, Salt P, Mieske M, Klerman GL. Facial electromyography in the assessment of improvement in depression. Psychosom Med 1978;40(4):355-360. [doi: 10.1097/00006842-197806000-00008] [Medline: 684137]
35. Davey GCL, Sired R, Jones S, Meeten F, Dash SR. The role of facial feedback in the modulation of clinically-relevant ambiguity resolution. Cogn Ther Res 2012;37(2):284-295. [doi: 10.1007/s10608-012-9480-5]
36. Schulze J, Neumann I, Magid M, Finzi E, Sinke C, Wollmer MA, et al. Botulinum toxin for the management of depression: an updated review of the evidence and meta-analysis. J Psychiatr Res 2021;135:332-340. [doi: 10.1016/j.jpsychires.2021.01.016] [Medline: 33578275]
37. Rinck M, Becker ES. Approach and avoidance in fear of spiders. J Behav Ther Exp Psychiatry 2007;38(2):105-120. [doi: 10.1016/j.jbtep.2006.10.001] [Medline: 17126289]
38. Lukas CA, Eskofier B, Berking M. A gamified smartphone-based intervention for depression: randomized controlled pilot trial. JMIR Ment Health 2021;8(7):e16643 [FREE Full text] [doi: 10.2196/16643] [Medline: 34283037]
39. Lukas CA, Berking M. Blending group-based psychoeducation with a smartphone intervention for the reduction of depressive symptoms: results of a randomized controlled pilot study. Pilot Feasibility Stud 2021;7(1):57 [FREE Full text] [doi: 10.1186/s40814-021-00799-y] [Medline: 33627198]
40. Kollei I, Lukas CA, Loeber S, Berking M. An app-based blended intervention to reduce body dissatisfaction: a randomized controlled pilot study. J Consult Clin Psychol 2017;85(11):1104-1108. [doi: 10.1037/ccp0000246] [Medline: 28857576]
41. Lukas CA, Blechert J, Berking M. A smartphone application to reduce problematic drinking: a feasibility trial. Pilot Feasibility Stud 2024;10(1):34 [FREE Full text] [doi: 10.1186/s40814-023-01420-0] [Medline: 38378675]
42. Lukas CA, Berking M. Reducing procrastination using a smartphone-based treatment program: a randomized controlled pilot study. Internet Interv 2018;12:83-90 [FREE Full text] [doi: 10.1016/j.invent.2017.07.002] [Medline: 30135772]
43. Keinert M, Schindler-Gmelch L, Eskofier BM, Berking M. An anger-based approach-avoidance modification training targeting dysfunctional beliefs in adults with elevated stress – results from a randomized controlled pilot study. Int J Cogn Ther 2024;17(4):700-724. [doi: 10.1007/s41811-024-00218-z]
44. Rupp LH, Keinert M, Böhme S, Schindler-Gmelch L, Eskofier B, Schuller B, et al. Sadness-based approach-avoidance modification training for subjective stress in adults: pilot randomized controlled trial. JMIR Form Res 2023;7:e50324 [FREE Full text] [doi: 10.2196/50324] [Medline: 38032725]
45. Streit H, Keinert M, Schindler-Gmelch L, Eskofier B, Berking M. Disgust-based approach-avoidance modification training for individuals suffering from elevated stress: a randomized controlled pilot study. Stress Health 2024;40(4):e3384. [doi: 10.1002/smi.3384] [Medline: 38367241]
46. O'Toole MS, Michalak J. Embodied cognitive restructuring: the impact of posture and movement on changing dysfunctional attitudes. J Behav Ther Exp Psychiatry 2024;84:101955 [FREE Full text] [doi: 10.1016/j.jbtep.2024.101955] [Medline: 38428361]

XSL•FO
RenderX

47. Keinert M, Schindler-Gmelch L, Rupp LH, Sadeghi M, Capito K, Hager M, et al. Facing depression: evaluating the efficacy of the empkinS-EKSpression reappraisal training augmented with facial expressions - protocol of a randomized controlled trial. BMC Psychiatry 2024;24(1):896 [FREE Full text] [doi: 10.1186/s12888-024-06361-3] [Medline: 39668374]
48. Capito K, Keinert M, Schindler-Gmelch L, Rupp LH, Sadeghi M, Eskofier B, et al. Efficacy of a smartphone-based reappraisal training against depressed mood and benefits of enhancing it with facial expression—a feasibility study. Open Science Framework. URL: https://osf.io/pw6ma/ [accessed 2024-12-09]
49. APA Style Journal Article Reporting Standards (APA Style JARS). American Psychological Association. URL: https://apastyle.apa.org/jars [accessed 2025-01-20]
50. Bell ML, Whitehead AL, Julious SA. Guidance for using pilot studies to inform the design of intervention trials with continuous outcomes. Clin Epidemiol 2018;10:153-157. [doi: 10.2147/clep.s146397]
51. Whitehead AL, Julious SA, Cooper CL, Campbell MJ. Estimating the sample size for a pilot randomised trial to minimise the overall trial sample size for the external pilot and main trial for a continuous outcome variable. Stat Methods Med Res 2016;25(3):1057-1073 [FREE Full text] [doi: 10.1177/0962280215588241] [Medline: 26092476]
52. Kroenke K, Spitzer RL. The PHQ-9: a new depression diagnostic and severity measure. Psychiatr Ann 2002;32(9):509-515. [doi: 10.3928/0048-5713-20020901-06]
53. Beesdo-Baum K, Zaudig M, Wittchen HU. In: Williams JBW, Karg RS, Spitzer RL, editors. Strukturiertes Klinisches Interview für DSM-5®-Störungen – Klinische Version. Newburyport, MA: Hogrefe; 2019.
54. Williams JB, Kobak KA, Bech P, Engelhardt N, Evans K, Lipsitz J, et al. The GRID-HAMD: standardization of the hamilton depression rating scale. Int Clin Psychopharmacol 2008;23(3):120-129. [doi: 10.1097/YIC.0b013e3282f948f5] [Medline: 18408526]
55. Beck AT, Rush J, Shaw BF, Emery G. Cognitive Therapy of Depression. New York, NY: Guilford Press; 1979.
56. Diedrich A, Grant M, Hofmann SG, Hiller W, Berking M. Self-compassion as an emotion regulation strategy in major depressive disorder. Behav Res Ther 2014;58:43-51. [doi: 10.1016/j.brat.2014.05.006] [Medline: 24929927]
57. Velten E. A laboratory task for induction of mood states. Behav Res Ther 1968;6(4):473-482. [doi: 10.1016/0005-7967(68)90028-4] [Medline: 5714990]
58. Kliem S, Lohmann A, Mößle T, Brähler E. Psychometric properties and measurement invariance of the beck hopelessness scale (BHS): results from a German representative population sample. BMC Psychiatry 2018;18(1):110 [FREE Full text] [doi: 10.1186/s12888-018-1646-6] [Medline: 29699522]
59. Pössel P. Cognitive triad inventory (CTI): psychometric properties and factor structure of the German translation. J Behav Ther Exp Psychiatry 2009;40(2):240-247. [doi: 10.1016/j.jbtep.2008.12.001] [Medline: 19110236]
60. Pössel P, Seemann S, Hautzinger M. Evaluation eines deutschsprachigen Instrumentes zur Erfassung positiver und negativer automatischer Gedanken. Z Für Klin Psychol Psychother 2005;34(1):27-34. [doi: 10.1026/1616-3443.34.1.27]
61. Ferring D, Filipp SH. Messung des Selbstwertgefühls: Befunde zur Reliabilität, Validität und Stabilität der Rosenberg-Skala [Measurement of self-esteem: Findings on reliability, validity, and stability of the Rosenberg Scale]. Diagnostica 1996:284-292.
62. Schrepp M, Hinderks A, Thomaschewski J. Design and evaluation of a short version of the User Experience Questionnaire (UEQ-S). Int J Interact Multimed Artif Intell 2017;4(6):103. [doi: 10.9781/ijimai.2017.09.001]
63. Hautzinger M, Bailer M, Hofmeister D, Keller F. Allgemeine Depressionsskala. Newburyport, MA: Hogrefe; 2012.
64. Rojas R, Geissner E, Hautzinger M. Entwicklung und psychometrische Überprüfung von zwei vergleichbaren Kurzversionen der Skala Dysfunktionaler Einstellungen (DAS). Diagnostica 2015;61(4):173-183. [doi: 10.1026/0012-1924/a000123]
65. Ebert DD, Christ O, Berking M. SEK-ES - Fragebogen zur emotionsspezifischen Selbsteinschätzung emotionaler Kompetenzen. Leibniz Institute for Psychology. 2013. URL: https://www.psycharchives.org/en/item/f2d7b513-70c7-444f-828f-94cea1781b04 [accessed 2025-03-26]
66. Albrecht NC, Langer D, Krauss D, Richer R, Abel L, Eskofier BM. EmRad: Ubiquitous vital sign sensing using compact continuous-wave radars. IEEE Open J Eng Med Biol 2024;5:725-734. [doi: 10.1109/ojemb.2024.3420241]
67. Lakens D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Front Psychol 2013;4:863 [FREE Full text] [doi: 10.3389/fpsyg.2013.00863] [Medline: 24324449]
68. Cohen J. Statistical Power Analysis for the Behavioral Sciences, 2nd Edition. London, UK: Routledge; 1989.
69. R: A language and environment for statistical computing. R Core Team. 2017. URL: https://www.R-project.org/ [accessed 2024-02-29]
70. Beard C, Weisberg RB, Primack J. Socially anxious primary care patients' attitudes toward cognitive bias modification (CBM): a qualitative study. Behav Cogn Psychother 2012;40(5):618-633 [FREE Full text] [doi: 10.1017/S1352465811000671] [Medline: 22127022]
71. Borghouts J, Eikey E, Mark G, De Leon C, Schueller SM, Schneider M, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. J Med Internet Res 2021;23(3):e24387 [FREE Full text] [doi: 10.2196/24387] [Medline: 33759801]
72. Diaz Hernandez L, Rieger K, Baenninger A, Brandeis D, Koenig T. Towards using microstate-neurofeedback for the treatment of psychotic symptoms in schizophrenia. a feasibility study in healthy participants. Brain Topogr 2016;29(2):308-321 [FREE Full text] [doi: 10.1007/s10548-015-0460-4] [Medline: 26582260]

XSL•FO
RenderX

73. Wampold BE. How important are the common factors in psychotherapy? An update. World Psychiatry 2015;14(3):270-277 [FREE Full text] [doi: 10.1002/wps.20238] [Medline: 26407772]
74. Loijen A, Vrijsen JN, Egger JIM, Becker ES, Rinck M. Biased approach-avoidance tendencies in psychopathology: a systematic review of their assessment and modification. Clin Psychol Rev 2020;77:101825. [doi: 10.1016/j.cpr.2020.101825] [Medline: 32143108]
75. Moshe I, Terhorst Y, Philippi P, Domhardt M, Cuijpers P, Cristea I, et al. Digital interventions for the treatment of depression: a meta-analytic review. Psychol Bull 2021;147(8):749-786. [doi: 10.1037/bul0000334] [Medline: 34898233]
76. Berking M, Rupp LH, Schindler-Gmelch L, Capito K, Böhme S, Keinert M. EmpkinS D02. Open Science Framework. URL: https://osf.io/mswfg/ [accessed 2024-12-09]

Abbreviations

AAMT: approach-avoidance modification training
AFE: antidepressive facial expression
CBT: cognitive behavioral therapy
CR: cognitive reappraisal
EMG: electromyography
EmpkinS: Empatho-Kinaesthetic Sensor Technology
FAU: Friedrich-Alexander-Universität Erlangen-Nürnberg
GRID-HAMD: GRID Hamilton Depression Rating Scale
ICS: interacting cognitive subsystems
PHQ: Patient Health Questionnaire
RCT: randomized controlled trial
SCID-5-CV: Structured Clinical Interview for DSM-5 Disorders–Clinician version
UEQ-S: Short Version of the User Experience Questionnaire

Edited by A Mavragani; submitted 13.08.24; peer-reviewed by A Eisingerich, J Blechert; comments to author 06.11.24; revised version received 09.12.24; accepted 05.02.25; published 14.04.25.

Please cite as:
Keinert M, Schindler-Gmelch L, Rupp LH, Sadeghi M, Richer R, Capito K, Eskofier BM, Berking M
The EmpkinS-EKSpression Reappraisal Training Augmented With Kinesthesia in Depression: One-Armed Feasibility Study
JMIR Form Res 2025;9:e65357
URL: https://formative.jmir.org/2025/1/e65357
doi:10.2196/65357
PMID:

©Marie Keinert, Lena Schindler-Gmelch, Lydia Helene Rupp, Misha Sadeghi, Robert Richer, Klara Capito, Bjoern M Eskofier, Matthias Berking. Originally published in JMIR Formative Research (https://formative.jmir.org), 14.04.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Evaluating User Engagement With a Real-Time, Text-Based Digital Mental Health Support App: Cross-Sectional, Retrospective Study

Edward Coffield[1], PhD; Khadeja Kausar[2], MHA

[1]Department of Population Health, Hofstra University, Hempstead, NY, United States

[2]Business Intelligence and Value Based Management Analytics, Medical Affairs, Maimonides Medical Center, Brooklyn, NY, United States

Corresponding Author:
Edward Coffield, PhD
Department of Population Health
Hofstra University
255 Hofstra University
101 Oak Street Center, Room 100A
Hempstead, NY, 11549
United States
Phone: 1 516 463 7019
Email: edward.coffield@hofstra.edu

Abstract

Background: Approximately 20% of US adults identify as having a mental illness. Structural and other barriers prevent many people from receiving mental health services. Digital mental health apps that provide 24-hour, real-time access to human support may improve access to mental health services. However, information is needed regarding how and why people engage with licensed counselors through a digital, real-time, text-based mental health support app in nonexperimental settings.

Objective: This study aimed to evaluate how people engage with Counslr, a 24-hour, digital, mental health support app where users communicate in real time with human counselors through text messaging. Specifically, access patterns (eg, day of the week and time of session) and reasons for accessing the platform were examined. Furthermore, whether differences existed between session types (on-demand or scheduled) and membership types (education or noneducation) in regard to access patterns and why people accessed the platform were evaluated.

Methods: The study population (users) consisted of students whose schools, universities, or colleges partnered with Counslr and employees whose organizations also partnered with Counslr. Users participated in text-based mental health support sessions. In these sessions, users engaged with licensed counselors through digital, text-based messaging in real time. Users could initiate an on-demand session or schedule a session 24 hours a day. User engagement patterns were evaluated through session length, session day, session time, and self-reported reasons for initiating the session. The data were stratified by membership type (education [students] or noneducation [employees]) and session type (on-demand or scheduled) to evaluate whether differences existed in usage patterns and self-reported reasons for initiating sessions by membership and session types.

Results: Most students (178/283, 62.9%) and employees (28/44, 63.6%) accessed Counslr through on-demand sessions. The average and median session times were 40 (SD 15.3) and 45 minutes. On-demand sessions (37.9 minutes) were shorter (P=.001) than scheduled sessions (43.5 minutes). Most users (262/327, 80.1%) accessed Counslr between 7 PM and 5 AM. The hours that users accessed Counslr did not statistically differ by membership type (P=.19) or session type (P=.10). Primary self-reported reasons for accessing Counslr were relationship reasons, depression, and anxiety; however, users initiated sessions for a variety of reasons. Statistically significant differences existed between membership and session types (P<.05) for some of the reasons why people initiated sessions.

Conclusions: The novel findings of this study illustrate that real-time, digital mental health support apps, which offer people the opportunity to engage with licensed counselors outside of standard office hours for a variety of mental health conditions, may help address structural barriers to accessing mental health support services. Additional research is needed to evaluate the effectiveness of human-based apps such as Counslr and whether such apps can also address disparities in access to mental health support services among different demographic groups.

(JMIR Form Res 2025;9:e66301) doi:10.2196/66301

XSL•FO
RenderX

KEYWORDS

mental health support; text; app; utilization; mobile; on demand; scheduled; mHealth; mobile health; app; student; university; college; mental health; employee; job; work; occupational health; counselor; counseling; usage; engagement; self-reported

Introduction

Overview

Approximately 1 out of every 5 adults (22.8%) in the United States identified as having a mental health illness in 2021 [1]. The gap between people who identify as having a mental illness and those who receive care or have all of their mental health care needs met is substantial. Among people who identify as having any mental illness, 47.2% received mental health treatments in 2021 and 27.6% perceived themselves as having an unmet mental health care need [1]. Access to mental health care services is not consistent among groups. For example, adults who identify as non-Hispanic Black or Hispanic are less likely than adults who identify as White to receive mental health care treatments [2].

Possible factors for having an unmet mental health care need include cost, stigma, minimization of symptoms, low perceived treatment effect, and structural barriers [3,4]; these factors differ across demographic groups [3,5]. For instance, the odds of costs being a factor in having an unmet mental health care need are higher among people who identify as female relative to people who identify as male, while the odds of people stating either stigma or structural barriers as a reason for their unmet mental health care needs are higher among people who identify as Black or Hispanic relative to people who identify as White [3].

Digital-based mental health apps have the potential to help people overcome barriers to obtaining care for mental illness [6,7] and reduce unmet mental health care needs among marginalized groups [8]. For example, the ability to consult with mental health professionals at all hours of the day, from any location, and in a digital environment that may foster privacy and anonymity, could address structural barriers and stigma [9]. Furthermore, Friis-Healy et al [10] state that digital mental health tools may help expand access to people most affected by COVID-19 and systematic racism by lowering care costs; "...decreasing transportation challenges; and providing mental health care in a private, destigmatizing manner." ([10], pages 2-3).

While digital mental health apps have been found to be effective and cost-effective [11], evidence regarding the effectiveness of digital mental health apps is inconsistent across digital mental health app types (eg, guidance within an app likely matters [7,12]). Furthermore, challenges, such as user engagement [6,9,13], remain. Baumel et al [13]—in an evaluation of 93 digital mental health care apps—found that only 4% of users opened the evaluated apps daily. Accordingly, understanding how people engage with digital mental health care apps is needed to facilitate design elements that may encourage engagement. As user engagement with mental health apps may differ between practical settings and clinical trials [6,13], a need exists to understand user engagement in a practical or actual setting. Engagement [13] and effectiveness [7] may also differ by app type, indicating an additional need to understand how people engage with different types of digital mental health apps in a practical setting.

Despite the need for nonexperimental evaluations of digital mental health apps, studies [13,14] regarding how people engage with digital mental health apps in practical settings [15] remain scarce. Studies have evaluated how users interact with (eg, downloading, using, and engaging) nonexperimental mental health apps in various forms [13-17]. For instance, Baumel et al [13] calculated usage metrics (eg, number of user sessions and minutes of daily use) for mobile apps— including 2 apps with peer support as a primary technique (and 7 apps with peer support as a cotechnique, for a total of 9), where the support could be "through chat or messaging services (synchronous and asynchronous...)" ([13], Appendix 2)—with at least 10,000 installs; Booth et al [14] calculated interactions by the hour and other engagement metrics for a chatbot; and Raue et al [17] reported engagement metrics (eg, weekly text message word count and weekly text messages) among users of a text-based app where users send text messages to therapists "24/7" ([17], page 3) who then review messages during "standard working hours" and respond at least "once a day, 5 days a week" ([17], page 3). However, usage patterns may differ between app designs (eg, human-based synchronous platforms, human-based asynchronous platforms, hybrid human-based designs, artificial intelligence [AI]–based platforms, and AI and human–based platforms), creating a need for further study. Notably, information is needed in terms of how (eg, day of the week, hours of day, and length of sessions) and why users engage with human-based, nonscripted apps, where users engage with humans through text messaging in real time. The addition of this information will provide policy makers and app designers with further data to consider when designing mental health apps.

In this study, we evaluate how users engage with Counslr, a digital, text-based mental health support app where users interact with licensed counselors in real time on their mobile devices [18]. This nonexperimental study is novel in that it evaluates how users engage with a digital mental health app that provides real-time, text-based mental health support. In addition, we explore why people access the app and whether engagement patterns statistically differ among people who access Counslr through their employment relative to people who access Counslr through their schools, universities, or colleges. Evaluating how users engage with and use Counslr in this retrospective study will contribute to the literature by providing an examination of how people interact with a counselor-based, digital mental health support app in a practical setting. In addition, this study will provide insights into whether people access a real-time, human-interfacing mobile mental health support app outside of standard business hours and how usage patterns may differ between instant, on-demand access and scheduled appointment access.

XSL•FO
RenderX

Counslr

Counslr was founded in 2019 and delivers mental health support services to users located in the United States through its mobile, text-based support platform. Through the Counslr platform, licensed counselors are available for confidential, real-time, text-based assistance 24 hours a day, 365 days a year. Counslr now uses an algorithm to match users with counselors based upon a user's preferences and needs. People may also schedule future sessions with a requested counselor. In addition to the text-based support sessions, Counslr provides self-guided resources within the platform. Access to Counslr is provided through school-, employment-, and community-based organizations that purchase a subscription to Counslr. Members of these organizations have the opportunity to enroll in Counslr and access Counslr-provided services through the firm's mobile platform. Members have no cost-sharing elements or usage limitations. Users' information is private and confidential as required by the Health Insurance Portability and Accountability Act (HIPAA). The app and its data are encrypted.

Methods

Overview

Data used for this retrospective study included Counslr users from October 26, 2021 to November 22, 2023 and focused exclusively on Counslr's text-based, real-time support sessions. This period covers the initial start-up of Counslr and its corresponding initial user membership. Counslr users, in this study, included students who accessed Counslr through their schools, universities, or colleges' partnerships with Counslr and employees who accessed Counslr through their organizations' partnerships with Counslr. Counslr provided a retrospective, deidentified dataset for this study. The provided data originated from user-reported information and data generated within the Counslr platform. Before a support session begins, users complete a survey where they self-report information to assist their counselor during the session. Second, data are generated within the platform such as the date or time of the session, session type (eg, on-demand or scheduled), and session length.

The session is the unit of analysis in this study. The data did not include a variable to identify individual users; thus, adjustments were not made for users who had multiple sessions. Individual datasets were provided for many of the variables. In some instances, dates were not consistent across the individual datasets (eg, a date missing in one dataset), and in other instances, presumable entry errors were present (eg, a session length was entered in minutes instead of a percentage of an hour). We adjusted for these challenges by either dropping the date from the dataset or adjusting the error where possible.

The primary purpose of this study was to evaluate how people access a 24-hour, digital mental health support app, where people communicate in real-time with human counselors through text messaging, and whether access patterns differ by session type (on-demand or scheduled) and membership type (education or noneducation). Access patterns were evaluated by day and hour (reported to authors in hour blocks [eg, 7 AM and 8 AM], not a specific time). The day and hour session distributions were stratified, separately, by session and membership types to examine whether access patterns differed between on-demand and scheduled sessions and between education- and non–education-based users. Chi-square tests were used for these comparisons. Chi-square tests with Bonferroni corrections were used as post hoc tests when appropriate.

The second evaluation used 2-tailed *t* tests to examine if session length (minutes) statistically differed between session types as well as by membership types. The last examination concerned users' self-reported reasons for initiating their sessions. Users could select a reason or reasons for their sessions from a set list provided within the app or provide their own reason for initiating a session. The self-reported categories or reasons for initiating sessions provided in the data included academic concerns; anxiety; depression; emotional abuse; family issues; infidelity; lesbian, gay, bisexual, and transgender (LGBT); relationship reasons; substance use; no reason provided (none); and others. In addition, 3 other categories were reported that were combined with these more general categories: "depression or anxiety" (n=3) was placed in the depression category; "other or relationship or depression" (n=1) was placed in the other category; and "[relationship]" (n=3) and "relationship or depression" (n=1) were placed in the relationship category. Overall summary statistics were estimated for the session reasons. In addition, the session reasons were tabulated and stratified by session type and by membership type. Chi-square tests were used to evaluate whether statistical differences existed between session types and, separately, between membership types in terms of why users initiated their sessions.

Ethical Considerations

The data used in this study were collected during Counslr's standard operations and adhered to HIPAA guidelines. The data are not publicly available. Hofstra University's institutional review board guidelines [19] deemed this study as non–human subject research because (1) Counslr provided a deidentified secondary dataset to the authors for evaluation, (2) the authors did not participate in the data collection and were not able to link the individuals in the deidentified dataset to living individuals through a code, and (3) the authors had independent control over how the data were analyzed and the study's conclusions.

Results

In total, 2 samples were created for the study due to the information available within each individual dataset. The primary dataset consisted of 327 support sessions that occurred between October 26, 2021 and October 19, 2023. The second sample consisted of the reasons why people initiated support sessions; this dataset consisted of support sessions conducted between August 15, 2023 to November 22, 2023 and, unlike session reasons provided for earlier dates, permitted stratification by both membership type and session type. For both datasets, dates without sessions were not included as were dates with data challenges that we could not address. The majority of evaluated sessions (206/327, 63%) were on-demand. Education-based users were the primary users in the sample (283/327, 86.5% of the sessions); the remaining sessions (44/327, 13.5%) comprised people associated with

XSL•FO
RenderX

non–education-based memberships. There was no statistical difference (*P*=.93) in how education and noneducation members accessed Counslr; the majority of both groups (education: 178/283, 62.9%; noneducation: 28/44, 63.6%) accessed Counslr through on-demand sessions.

Across session and membership types, the mean and median session lengths were 40 (SD 15.3) and 45 minutes. Three sessions were noted as 0 minutes, while the longest session was 76 minutes. Mean session lengths for education- and non–education-based users were not significantly different (*P*=.57); the mean session lengths for education- and non–education-based users were 39.8 (SD 15.4; median 45, IQR 36-48) and 41.2 (SD 14.4; median 45, IQR 42.5-48.5) minutes. Differing from membership type, mean session length did statistically differ (*P*=.001) between on-demand (37.9, SD 16.6; median 45, IQR 27-47 minutes) and scheduled sessions (43.5, SD 11.9; median 46, IQR 42-49 minutes).

The share of sessions allocated to Fridays, Saturdays, and Sundays were each below 14.29% (Table 1); 14.29% is the expected share for each day if sessions were evenly distributed across all 7 days. Conversely, the share of sessions allocated to Mondays, Tuesdays, Wednesdays, and Thursdays each exceeded 14.29%. The distribution of sessions across days did not statistically differ (*P*=.62) by membership type. However, this distribution did statistically differ (*P*=.03) by session type. In post hoc tests, the distributions of session type on Thursdays and Sundays were seemingly different relative to the same distribution on other days, yet, after Bonferroni corrections, no statistical differences were present at the Bonferroni-corrected significance level of .007.

From a time perspective (Figure 1), 80.1% (262/327) of sessions were documented (in hour blocks, not specific times) as occurring from 7 PM to 5 AM, 41.6% (136/327) of sessions occurred between 7 PM and 12 AM, while 38.5% (126/327) of sessions were between 12 AM and 5 AM. This pattern was also present by membership type as 80.9% (229/283) of education-based users and 75% (33/44) of non–education-based users accessed the platform between 7 PM and 5 AM. The distribution of sessions across hours of the day did not statistically differ between session type (*P*=.10) or membership type (*P*=.19). Figure 1 illustrates these hourly-based distributions by membership type (A) and session type (B).

As indicated, a second dataset was constructed to evaluate why users initiated support sessions. Users could select multiple reasons for initiating a session; we evaluated 327 reasons in this study. While this is the same number of sessions included in the first dataset, this dataset covers a shorter time span than the first dataset and the reasons users provided for their sessions in this second dataset are for a smaller number of sessions than 327. The top 3 self-reported reasons for initiating a session (Table 2) pertained to relationships (75/327, 22.9%), depression (71/327, 21.7%), and anxiety (62/327, 19%). Among non–education-based users, the top 3 reasons for initiating a session (Table 2) concerned relationships (23/64, 35.9% of sessions), anxiety (10/64, 15.6%), and family issues (9/64, 14.1%). Among education-based users, the top 3 reasons for engaging in a session were depression (65/263, 24.7%), anxiety (52/263, 19.8%), and relationships (52/263, 19.8%).

Table 1. Distribution of user sessions by day of the week[a].

Day	All, n (%)	Membership type, n (%)		Session type, n (%)		
		Education	Noneducation	On demand	Scheduled	*P* value[b]
Monday	49 (15)	43 (15.2)	6 (13.6)	28 (13.6)	21 (17.4)	.36
Tuesday	58 (17.7)	46 (16.3)	12 (27.3)	32 (15.5)	26 (21.5)	.17
Wednesday	60 (18.4)	51 (18)	9 (20.5)	38 (18.5)	22 (18.2)	.95
Thursday	52 (15.9)	47 (16.6)	5 (11.4)	26 (12.6)	26 (21.5)	.03[c]
Friday	44 (13.5)	40 (14.1)	4 (9.1)	32 (15.5)	12 (9.9)	.15
Saturday	27 (8.3)	23 (8.1)	4 (9.1)	21 (10.2)	6 (5)	.10
Sunday	37 (11.3)	33 (11.7)	4 (9.1)	29 (14.1)	8 (6.6)	.04[c]
P value[d]			.62		.03	

[a]Sessions from October 26, 2021, to October 19, 2023 (n=327).

[b]*P* values from post hoc χ^2 tests; the Bonferroni-corrected significance level for these tests is .007.

[c]These *P* values are considered above the Bonferroni-corrected significance level of .007 and therefore do not support a statistically significant difference between the percentages.

[d]*P* values from χ^2 evaluations on overall distributions of sessions over days.

XSL•FO
RenderX

Figure 1. Distribution of user sessions by hour: hour of session by (A) membership type (noneducation or education) and (B) session type (scheduled or on demand). The data provided to the authors noted hours in blocks (eg, 7 AM, 8 AM) not specific times. Sessions from October 26, 2021 to October 19, 2023 (n=327); χ2 test for membership type group distribution difference: *P*=.19; χ2 test for session type group distribution difference: *P*=.10.

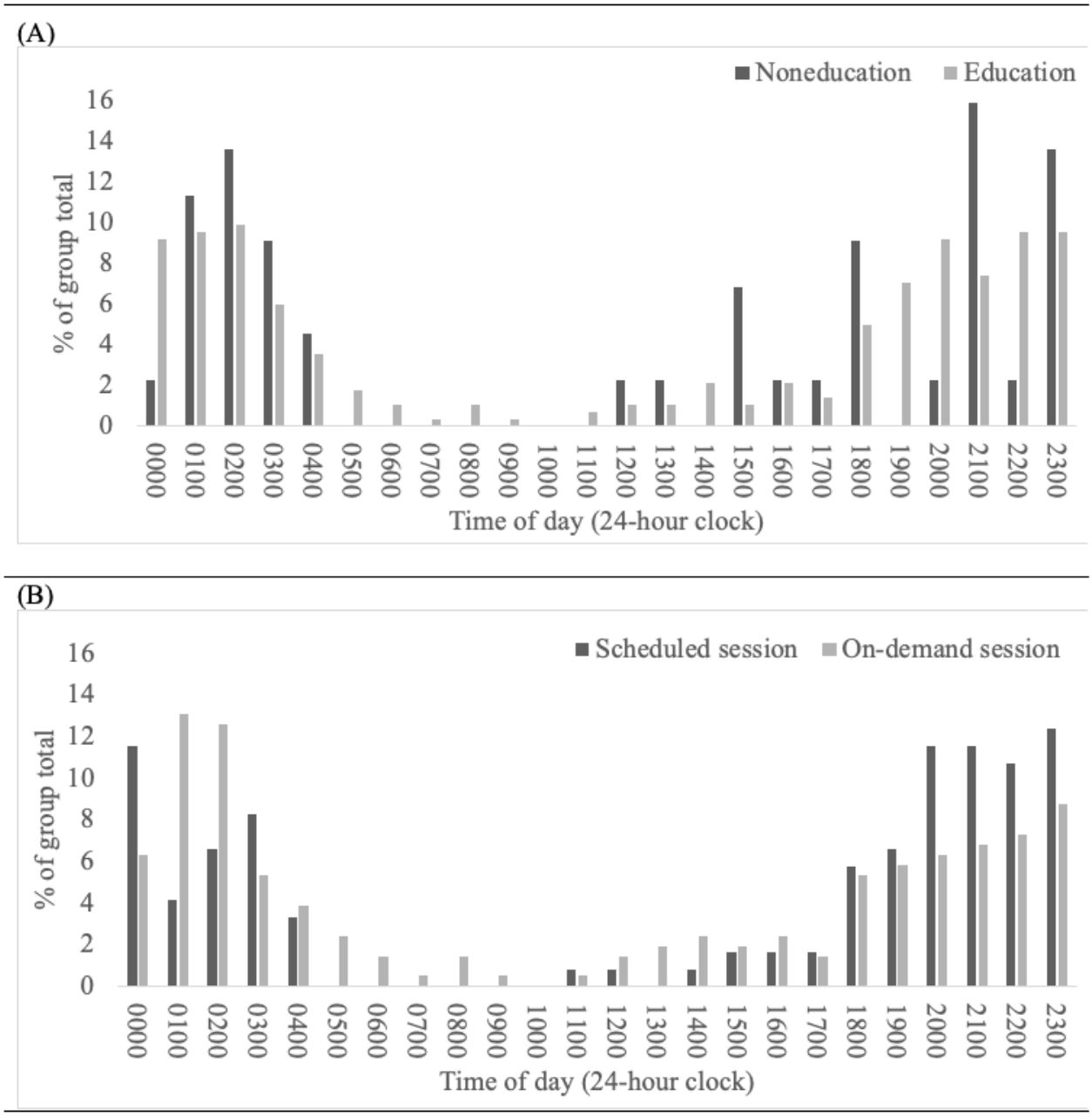

XSL•FO
RenderX

Table 2. Users' self-reported reasons for sessions, by session type and membership type[a].

Session reason	All, n (%)	Membership type, n (%)			Session type, n (%)		
		Education	Noneducation	*P* value[b]	On demand	Scheduled	*P* value[b]
Academic concerns	20 (6.1)	14 (5.3)	6 (9.4)	.23	14 (6.7)	6 (5.1)	.58
Anxiety	62 (19)	52 (19.8)	10 (15.6)	.45	39 (18.6)	23 (19.7)	.81
Depression	71 (21.7)	65 (24.7)	6 (9.4)	.01	53 (25.2)	18 (15.4)	.04
Emotional abuse	4 (1.2)	2 (0.8)	2 (3.1)	.12	3 (1.4)	1 (0.9)	.65
Family issues	23 (7)	14 (5.3)	9 (14.1)	.01	13 (6.2)	10 (8.6)	.42
Infidelity	3 (0.9)	3 (1.1)	0 (0)	.39	3 (1.4)	0 (0)	.19
LGBT[c]	10 (3.1)	8 (3)	2 (3.1)	.97	7 (3.3)	3 (2.6)	.70
No reason provided (none)	11 (3.4)	11 (4.2)	0 (0)	.10	0 (0)	11 (9.4)	<.001
Other	42 (12.8)	37 (14.1)	5 (7.8)	.18	29 (13.8)	13 (11.1)	.48
Relationship	75 (22.9)	52 (19.8)	23 (35.9)	.01	48 (22.9)	27 (23.1)	.96
Substance use	6 (1.8)	5 (1.9)	1 (1.6)	.86	1 (0.5)	5 (4.3)	.01

[a]Reasons (n=327) provided during sessions from August 15, 2023 to November 22, 2023. Users could select a reason or reasons for the session from the list or type another reason or reasons.

[b]*P* values from χ^2 tests.

[c]LGBT: lesbian, gay, bisexual, and transgender.

As illustrated in Table 2, a statistical difference was present in the percentage of education and noneducation users who accessed support sessions for depression (*P*=.01), family issues (*P*=.01), and relationships (*P*=.01). Table 2 also illustrates the reasons users initiated sessions by session type. The primary reasons for support sessions by session type were depression (53/210, 25.2%), relationships (48/210, 22.9%), and anxiety (39/210, 18.6%) for on-demand sessions and relationships (27/117, 23.1%), anxiety (23/117, 19.7%), and depression (18/117, 15.4%) for scheduled sessions. Users accessed support through scheduled sessions (5/117, 4.3%) at a statistically (*P*=.01) greater percentage relative to on-demand sessions (1/210, 0.5%) for substance use (as well as for the reason of none or no reason provided). Conversely, users accessed support for depression at a statistically (*P*=.04) higher percentage through on-demand sessions (53/210, 25.2%) relative to scheduled sessions (18/117, 15.4%).

Discussion

Principal Findings

Digital mental health apps have the potential to expand access to mental health services. These apps can help people overcome barriers to obtaining such care [6-8]. In this study, we analyzed Counslr, a digital mental health app that offers real-time, text-based support sessions with human counselors. We found that users who accessed Counslr participated in support sessions for an average of 40 minutes. The majority of these sessions occurred during times when standard, human-based mental health support is not available; this salient finding illustrates the promise of human-based, digital mental health apps in overcoming structural barriers to accessing mental health support services.

The novelty of this study in terms of evaluating user engagement with a human-based, nonscripted app, where users engage with human counselors through text messaging in a real-time, nonexperimental setting, prevents multiple, direct comparisons with the peer-reviewed literature. However, a few comparisons on similar metrics illustrate how usage patterns may differ between app designs. For instance, differing from this study's result that the majority of sessions occurred between 7 PM and 5 AM, Booth et al [14], evaluating the usage of a chatbot, found that peak interaction hours with the chatbot were in the hours of 8-10 AM, 1 PM, and 5 PM. However, closer to this study's results, Baumel et al [13]—evaluating 93 apps—found that usage peaked at the 8 PM hour for the apps included in their mental health category, while apps in their mindfulness and meditation category peaked in the hours of 7-9 AM and 10 PM to 12 AM. Baumel et al [13] also illustrated, in a graph, that usage rates for apps in their mental health category were highest on Tuesdays, Wednesdays, and Thursdays, which is similar to this study's overall, nonstratified results (statistical comparisons were not conducted on this study's overall, nonstratified day of the week statistics); the peak day of usage for the mindfulness or meditation applications Baumel et al [13] reviewed was Thursday.

Raue et al [17], evaluating a text-based app that included human interactions (users could text at any time in their study, while therapists reviewed the texts during working hours and responded at least 1 time a day to a person's text 5 days a week; [17], page 3), did not report directly comparable metrics to those illustrated in this study. However, from a time of session perspective, Baumel et al [13] found that the median daily minutes with which active users engaged in 2 peer-supported apps they reviewed was 35.08; a finding below the 45-minute median found in this study. Overall, Baumel et al [13] found

XSL•FO
RenderX

that the median daily minutes of use per active app user was different between mental health apps (10.02 minutes), happiness apps (7.77 minutes), and mindfulness and meditation apps (21.47 minutes), demonstrating that the reason why people engage with an app may influence usage.

The primary reasons users of Counslr self-reported for initiating sessions were similar to mental health conditions found across the population. For instance, 2 of the primary reasons Counslr users noted for initiating sessions, anxiety and depression, were also the most prevalent mental health conditions found among college students [20], while symptoms related to anxiety disorders or depressive disorders were present in 20.7% of US adults in early 2024 [21]. Collectively, how Counslr users engaged with licensed counselors outside of standard office hours, coupled with the commonality between mental health conditions found in the general population and the reasons Counslr users initiated support sessions, provide evidence that people will use a human-based, digital mental health app to access support outside of standard hours for prevalent mental health conditions.

While we did not have extensive demographic information about the Counslr users whose sessions were evaluated within this study, the 2 different membership types allowed comparisons between people who accessed the platform as students and those who accessed the platform as employees. When examined, usage patterns between these 2 groups did not statistically differ. Both students and employees engaged in sessions primarily between 7 PM and 5 AM and had similar patterns in terms of the days on which their sessions were held. Furthermore, the majority of both student and employee users accessed the platform using the on-demand feature instead of scheduling an appointment. The access patterns found among both student and employee users indicate that mental health support apps have the potential to increase access to mental health support services across multiple groups. This potential, universal appeal of Counslr-like platforms is also seen in how the platform is able to provide support for a wide array of reasons: from anxiety and depression to infidelity and academic concerns.

As noted here and elsewhere [9], the 24-hour accessibility of mental health apps is one attribute of such apps that could decrease structural barriers to accessing support services to address mental health challenges. However, the salient findings here regarding on-demand and scheduled sessions demonstrate that how access to such apps is provided is also important in terms of app design. For instance, there were no statistical differences between the frequency of on-demand and scheduled sessions in terms of how the sessions were distributed across the hours of the day; people both scheduled sessions and initiated on-demand sessions at 4 AM, indicating a need for both types of sessions.

Differences between the 2 session types also illustrate the need to consider offering both on-demand and scheduled sessions within a digital mental health support app; doing such may help facilitate the apps' ability to improve access to mental health support. For instance, the distribution of sessions across days of the week differed between on-demand sessions and scheduled sessions, indicating possible user preferences for set sessions with counselors or on-demand sessions on certain days. Furthermore, a larger share of sessions initiated for the reason of substance use occurred as scheduled sessions relative to on-demand sessions. This finding suggests that users may prefer scheduled sessions for substance use relative to on-demand sessions, indicating that during the app design stage, developers should consider whether available access modes (on-demand and scheduled) address how users may prefer different access points depending on the specific reason they are accessing the app.

The connection between app design and this study's results was possible because Counslr shared their formative data with researchers for independent analysis. As Counslr and other platforms continue to share their data for analyses, additional insights into future app development may arise; these insights may result in increased efficiency in app development as well as the creation of platforms designed to address the needs of users, thereby possibly increasing usage among people who seek mental health support thought such apps.

In this study, we conducted a novel evaluation of a 24-hour, real-time, digital mental health support app where users interact with humans through text messaging. The data were from a nonexperimental situation, helping to address the stated need for more information regarding how digital mental health apps perform outside of experimental settings [6,13,15,16]. While the findings of this study illustrate the promise of digital mental health support apps to increase access to a variety of mental health needs, a number of limitations are present. First, this study relied on a small, formative dataset provided by Counslr for independent analysis. While a larger sample size is preferable, the study allowed multiple insights into usage patterns and why people accessed the platform. Furthermore, using this formative dataset to evaluate initial platform usage allowed the authors to suggest additional metrics for data collection and analysis to the organization.

A second limitation is that the deidentified data used here and the lack of demographic variables entail that the findings of this study only pertain to study participants and are not generalizable to the overall population. Furthermore, the data did not include individual identifiers, preventing adjustments for users with multiple sessions, which could skew the results (eg, if one user initiated multiple sessions for the same reason). To provide further insight into this limitation, Counslr, at the authors' request, provided information illustrating that there were 154 unique users in the data with an average and median number of sessions of 2.06 and 1 (minimum 1 and maximum 21).

The deidentified nature of the data and the manner in which the data were provided to the authors also prevented more complex data analysis (eg, regression analysis), which could provide further insights into usage and access patterns. In addition, the platform did not collect demographic information on users such as how they identify their ethnicity, gender, and race. The inclusion of such information would allow pivotal evaluations regarding whether digital mental health support apps can help address disparities in access to mental health support services across different demographic groups. Finally, while the primary purpose of this study was to evaluate how and why people

XSL•FO
RenderX

engage with licensed counselors through a digital, real-time, text-based mental health support app in nonexperimental settings, future evaluations—when such data are available—must examine the effectiveness of such apps in addressing the mental health support needs of users and whether such platforms serve as a complement to more traditional mental health services.

Conclusion

Barriers that prevent people from accessing mental health support services are multiple and include structural barriers and stigma. Anonymity and 24-hour access to digital mental health support apps may help address such barriers. In this study, we evaluated how people use Counslr, a real-time mental health support app where users engage with licensed, human counselors through text messaging sessions. The majority of both education- and non–education-based users initiated mental health support sessions outside of standard business hours for a variety of reasons, indicating the promise of Counslr like apps to address barriers to accessing mental health support services. The similarities in usage patterns of Counslr student and employee users, despite one group being noneducation based and the other being education based, demonstrated an almost universality in the platform in terms of serving multiple groups with multiple needs.

The access patterns found in this study demonstrate the promise of Counslr-type apps to help address structural barriers to accessing mental health support services. Additional research is needed to evaluate the effectiveness of such apps, notably human-based apps such as Counslr, and whether such apps can also address disparities in access to mental health support services among demographic groups.

Acknowledgments

Artificial intelligence was not used in the composition of this manuscript.

Data Availability

The data analyzed in this study are not publicly available. The data were provided to the authors by Counslr.

Conflicts of Interest

EC and KK have no financial interests in Counslr and did not receive any form of payment for this work. EC, in his role as an academic program director and faculty member, works closely on multiple items with another faculty member who is a Counslr board member. KK has also interacted with this faculty member. Study data were provided by Counslr. Members of the Counslr team answered questions regarding the platform or data (or both) and reviewed the manuscript. However, EC and KK conducted the study independently from Counslr. The conclusions provided in the manuscript are those of EC and KK.

References

1. Center for behavioral health statistics and quality, results from the 2021 national survey on drug use and health. Substance Abuse and Mental Health Services Administration. URL: https://www.samhsa.gov/data/report/2021-nsduh-detailed-tables [accessed 2025-02-05]
2. Terlizzi EP, Norris T. Mental health treatment among adults: United States. NCHS Data Brief 2021(419):1-8. [doi: 10.15620/cdc:110593]
3. Alang SM. Sociodemographic disparities associated with perceived causes of unmet need for mental health care. Psychiatr Rehabil J 2015;38(4):293-299. [doi: 10.1037/prj0000113] [Medline: 25664758]
4. Yang JC, Roman-Urrestarazu A, McKee M, Brayne C. Demographic, socioeconomic, and health correlates of unmet need for mental health treatment in the United States, 2002-16: evidence from the national surveys on drug use and health. Int J Equity Health 2019;18(1):122 [FREE Full text] [doi: 10.1186/s12939-019-1026-y] [Medline: 31382979]
5. Green JG, McLaughlin KA, Fillbrunn M, Fukuda M, Jackson JS, Kessler RC, et al. Barriers to mental health service use and predictors of treatment drop out: racial/ethnic variation in a population-based study. Adm Policy Ment Health 2020;47(4):606-616 [FREE Full text] [doi: 10.1007/s10488-020-01021-6] [Medline: 32076886]
6. Lipschitz JM, Pike CK, Hogan TP, Murphy SA, Burdick KE. The engagement problem: a review of engagement with digital mental health interventions and recommendations for a path forward. Curr Treat Options Psychiatry 2023;10(3):119-135 [FREE Full text] [doi: 10.1007/s40501-023-00297-3] [Medline: 38390026]
7. Mohr DC, Azocar F, Bertagnolli A, Choudhury T, Chrisp P, Frank R, Banbury Forum on Digital Mental Health. Banbury forum consensus statement on the path forward for digital mental health treatment. Psychiatr Serv 2021;72(6):677-683 [FREE Full text] [doi: 10.1176/appi.ps.202000561] [Medline: 33467872]
8. Ramos G, Ponting C, Labao JP, Sobowale K. Considerations of diversity, equity, and inclusion in mental health apps: a scoping review of evaluation frameworks. Behav Res Ther 2021;147:103990 [FREE Full text] [doi: 10.1016/j.brat.2021.103990] [Medline: 34715396]
9. Aschbacher K, Rivera LM, Hornstein S, Nelson BW, Forman-Hoffman VL, Peiper NC. Longitudinal patterns of engagement and clinical outcomes: results from a therapist-supported digital mental health intervention. Psychosom Med 2023;85(7):651-658. [doi: 10.1097/PSY.0000000000001230] [Medline: 37409793]

XSL•FO
RenderX

10. Friis-Healy EA, Nagy GA, Kollins SH. It is time to REACT: opportunities for digital mental health apps to reduce mental health disparities in racially and ethnically minoritized groups. JMIR Ment Health 2021;8(1):e25456 [FREE Full text] [doi: 10.2196/25456] [Medline: 33406050]
11. Moshe I, Terhorst Y, Philippi P, Domhardt M, Cuijpers P, Cristea I, et al. Digital interventions for the treatment of depression: a meta-analytic review. Psychol Bull 2021;147(8):749-786. [doi: 10.1037/bul0000334] [Medline: 34898233]
12. Weisel KK, Fuhrmann LM, Berking M, Baumeister H, Cuijpers P, Ebert DD. Standalone smartphone apps for mental health-a systematic review and meta-analysis. NPJ Digit Med 2019;2:118 [FREE Full text] [doi: 10.1038/s41746-019-0188-8] [Medline: 31815193]
13. Baumel A, Muench F, Edan S, Kane JM. Objective user engagement with mental health apps: systematic search and panel-based usage analysis. J Med Internet Res 2019;21(9):e14567 [FREE Full text] [doi: 10.2196/14567] [Medline: 31573916]
14. Booth F, Potts C, Bond R, Mulvenna M, Kostenius C, Dhanapala I, et al. A mental health and well-being chatbot: user event log analysis. JMIR Mhealth Uhealth 2023;11:e43052 [FREE Full text] [doi: 10.2196/43052] [Medline: 37410539]
15. Carlo AD, Hosseini Ghomi R, Renn BN, Areán PA. By the numbers: ratings and utilization of behavioral health mobile applications. NPJ Digit Med 2019;2:54 [FREE Full text] [doi: 10.1038/s41746-019-0129-6] [Medline: 31304400]
16. Balaskas A, Schueller SM, Cox AL, Rashleigh C, Doherty G. Examining young adults daily perspectives on usage of anxiety apps: a user study. PLOS Digit Health 2023;2(1):e0000185 [FREE Full text] [doi: 10.1371/journal.pdig.0000185] [Medline: 36812622]
17. Raue PJ, Fridling N, Song J, Hull TD, Alexopoulos GS, Arean PA. Message-based psychotherapy for older adults: a cohort comparison study. Front Psychiatry 2022;13:951354 [FREE Full text] [doi: 10.3389/fpsyt.2022.951354] [Medline: 36090371]
18. Counslr. URL: https://www.counslr.com/ [accessed 2025-02-05]
19. IRB Decision Chart. Hofstra University. URL: https://www.hofstra.edu/pdf/about/administration/provost/orsp/osrp_hofstra_irb_decision_chart.pdf [accessed 2025-04-14]
20. Oswalt SB, Lederer AM, Chestnut-Steich K, Day C, Halbritter A, Ortiz D. Trends in college students' mental health diagnoses and utilization of services, 2009-2015. J Am Coll Health 2020;68(1):41-51. [doi: 10.1080/07448481.2018.1515748] [Medline: 30355071]
21. National Center for Health Statistics. Anxiety and depression. Centers for Disease Control and Prevention. 2024. URL: https://www.cdc.gov/nchs/covid19/pulse/mental-health.htm [accessed 2024-02-05]

Abbreviations

HIPAA: Health Insurance Portability and Accountability Act
LGBT: lesbian, gay, bisexual, and transgender

Edited by A Mavragani; submitted 09.09.24; peer-reviewed by J Rhodes; comments to author 07.10.24; revised version received 25.11.24; accepted 16.01.25; published 14.04.25.

Please cite as:
Coffield E, Kausar K
Evaluating User Engagement With a Real-Time, Text-Based Digital Mental Health Support App: Cross-Sectional, Retrospective Study
JMIR Form Res 2025;9:e66301
URL: https://formative.jmir.org/2025/1/e66301
doi:10.2196/66301
PMID:

©Edward Coffield, Khadeja Kausar. Originally published in JMIR Formative Research (https://formative.jmir.org), 14.04.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Acceptability, Usability, and Insights Into Cybersickness Levels of a Novel Virtual Reality Environment for the Evaluation of Depressive Symptoms: Exploratory Observational Study

Sara Sutori[1*], MSc; Emma Therése Eliasson[1*], PhD; Francesca Mura[2], PhD; Victor Ortiz[3], MSc; Vincenzo Catrambonephd[4], PhD; Gergö Hadlaczky[1], PhD; Ivo Todorov[1], PhD; Antonio Luca Alfeo[4], PhD; Valentina Cardi[5], PhD; Mario G C A Cimino[4], PhD; Giovanna Mioni[5], PhD; Mariano Alcañiz Raya[3], PhD; Gaetano Valenza[4], PhD; Vladimir Carli[1*], MD, PhD; Claudio Gentili[5*], MD, PhD

[1]National Centre for Suicide Research and Prevention, Department of Learning, Informatics, Management and Ethics, Karolinska Institutet, Stockholm, Sweden

[2]Padova Neuroscience Center, University of Padua, Padua, Italy

[3]Instituto Universitario de Investigación en Tecnología Centrada en el Ser Humano, Universitat Politècnica de València, Ciudad Politécnica de la Innovación, València, Spain

[4]Research Center E. Piaggio, Department of Information Engineering, University of Pisa, Pisa, Italy

[5]Department of General Psychology, University of Padua, Padua, Italy

[*]these authors contributed equally

Corresponding Author:
Emma Therése Eliasson, PhD
National Centre for Suicide Research and Prevention
Department of Learning, Informatics, Management and Ethics
Karolinska Institutet
Granits väg 4
Stockholm, 171 65
Sweden
Phone: 46 701407295
Email: emma.eliasson@ki.se

Abstract

Background: There is a clear need for enhanced mental health assessment, depressive symptom (DS) evaluation being no exception. A promising approach to this aim is using virtual reality (VR), which entails the potential of adding a wider set of assessment domains with enhanced ecological validity. However, whilst several studies have used VR for both diagnostic and treatment purposes, its acceptance, in particular how exposure to virtual environments affects populations with psychiatric conditions remains unknown.

Objective: This study aims to report on the acceptability, usability, and cybersickness levels of a pilot VR environment designed for the purpose of differentiating between individuals with DSs.

Methods: The exploratory study, conducted in Italy, included 50 healthy controls and 50 young adults with mild-to-moderate DSs (without the need for a formal diagnosis). The study used an observational design with approximately 30 minutes of VR exposure followed by a self-report questionnaire battery. The battery included a questionnaire based on the Theoretical Framework of Acceptability, the System Usability Scale as well as the Simulator Sickness Questionnaire.

Results: Results indicate that the majority found VR acceptable for the purposes of mental health screening and treatment. However, for diagnostics, there was a clear preference for VR to be used by mental health professionals as a supplementary tool, as opposed to a stand-alone solution. In practice, following exposure to the pilot VR environment, generally, good levels of acceptability and usability were reported, but areas in need of improvement were identified (such as self-efficacy). Self-reported cybersickness levels were comparable to literature averages but were considerably higher among those with DSs.

Conclusions: These findings raise questions about the potential interplay between underlying somatic symptoms of depression and VR-induced cybersickness and call for more attention from the scientific community both in terms of methodology as well as potential clinical and theoretical implications. Conclusively, user support indicates a potential for VR to aid mental health

XSL•FO
RenderX

assessment, but further research is needed to understand how exposure to virtual environments might affect populations with varying severity and other forms of psychiatric symptoms.

International Registered Report Identifier (IRRID): RR2-10.1186/ISRCTN16396369

(JMIR Form Res 2025;9:e68132) doi:10.2196/68132

KEYWORDS

depression; virtual reality; assessment; acceptability; usability; cybersickness

Introduction

Background

In 2019, an estimated 970 million people worldwide, or 12.6% of the global population, were living with a mental disorder, with anxiety (301 million) and depression (280 million) being the most prevalent [1]. These conditions contribute substantially to the global burden of disease, with depression being one of the leading causes of disability worldwide [2]. Despite considerable efforts to address this issue, there has been no substantial reduction in the burden of mental disorders since 1990 [1].

Depressive disorders (hereafter referred to as depression) present significant challenges in research, among other reasons due to the diagnostic uncertainty [3], characterized by both high heterogeneity and low reliability [4]. The persistence of these challenges over different iterations of the *DSM* (*Diagnostic and Statistical Manual of Mental Disorders*) [4] underscores the need to advance clinical practice, which currently predominantly relies on self-report questionnaires and clinical interviews [5].

In response to the challenges mentioned above, suggestions have been made to include a broader range of measures that cover more assessment domains and are not solely dependent on verbal self-report [6,7]. Such measures include speech, text, and facial expression analysis [8]; genomics, transcriptomics, proteomics, metabolomics, and imaging [9]; behavioral and physiological variables collected via wearable sensors [10] or virtual reality (VR) systems [7].

VR systems are particularly relevant to mental health disorders due to immersive, controlled, and customizable environments that can be tailored to individual needs [7]. VR has been used across psychiatric pathologies but reviews consistently report that depression has remained understudied compared to other conditions [7,11,12]. The gap in the literature is particularly evident for studies on assessment, of which we could only find three [13-15]. These are specifically focusing on attentional processes [14], spatial navigation [15], and social interactions [13] in relation to depression. Despite these pioneering studies, the potential of VR to integrate multiple domains of data for enhancing assessment [7] has yet to be explored.

While efforts are being dedicated to improving treatment and the testing of various measures to enhance assessment, including the use of VR, even less attention has been paid to the evaluation of such novel approaches from a user perspective. Research on VR for depression treatment and assessment is limited, with even fewer studies focusing on its acceptability, usability, and tolerability. Beyond diagnostic accuracy, evaluating the acceptability and usability of diagnostic tools is crucial, particularly when involving individuals with mental health conditions, like depression. In such cases, where patients might face additional treatment barriers, including motivational difficulties, stigma, delayed help-seeking [16], or poor treatment adherence [17], acceptability and usability play a crucial role in the successful implementation of new diagnostic and treatment technologies.

Sekhon et al [18] define acceptability as a "multifaceted construct that reflects the extent to which people delivering or receiving a health care intervention consider it to be appropriate, based on anticipated or experienced cognitive and emotional responses to the intervention." On the other hand, usability has various definitions and approaches [19], but here we will refer to the extent to which the VR system can be used with ease for the purpose of assessing depression symptom severity. Third, tolerability may be defined in different ways, but here will refer to cybersickness severity. Cybersickness describes a visually induced motion sickness-like discomfort during or following exposure to VR. Cybersickness is accompanied by changes in the activity of the central and autonomic nervous systems and results in symptoms like headache, nausea, and eyestrain among others [20].

To the best of our knowledge, only 2 studies using VR have reported on the earlier concepts in relation to depression: one focusing on psychoeducation [21] and another on behavior activation [22]. They both found generally high acceptability and satisfaction levels. Migoya et al [21] also found favorable ratings for ease of use and perceived usefulness, whereas Paul et al [22] reported a decreasing interest in continuing the program after having attended multiple sessions. Paul et al [22] also assessed cybersickness, noting decreasing symptoms as the sessions progressed, but their unconventional scoring method (adding up raw scores as opposed to using weight adjustments) complicates comparisons with studies on healthy controls (HCs) [23] or other psychiatric conditions [24]. To date, no studies have compared cybersickness levels between depressed individuals and HCs.

To tackle the aforementioned challenges, this study introduces a novel approach aimed at enhancing the quality—including the breadth and ecological validity—of depression assessment. A VR system and environment were developed to capture physiological data (heart rate, skin conductance, and eye tracking), behavioral data (curiosity-related behavior and attentional biases), and cognitive data (working memory, processing speed, sustained attention, cognitive flexibility, self-evaluation, and metacognitive sensitivity) hypothesized to be associated with depression [25]. These multidomain data are

XSL•FO
RenderX

then analyzed via machine learning models and explainable artificial intelligence to assess depressive symptom (DS) severity. The categorization accuracy of this system is reported on in a dedicated study [25], however, given the innovative nature of this approach and the gap in the literature, it is fundamental to consider the patient perspective as well. Therefore, the objective of this study was to evaluate the acceptability, usability, and cybersickness levels of this pilot virtual environment.

Aim

The aim of the study was to assess the acceptability, usability, and cybersickness levels of a pilot VR environment designed to aid and enrich the assessment of DSs.

Methods

Overview

The study had an exploratory nature and followed an observational design, where all participants completed the same questionnaire batteries before and post VR and followed the same VR exposure protocol.

Sample

A total of 266 individuals were screened at the University of Padua, Italy, with continuous enrollment between November 2022 and November 2023 (Figure 1). The screening concerned the inclusion and exclusion criteria detailed later and was done using self-report questionnaires administered in Qualtrics (Qualtrics, Inc). Recruitment was completed when the predefined target sample size of 100 was reached (Figure 1), including 50 participants with DSs and 50 participants classed as HCs. Participants were aged 18-35 years, fulfilled all inclusion criteria, and provided written informed consent. Further demographic and clinical characteristics of the sample are outlined in Table 1.

Figure 1. Consort flow diagram of participant involvement. VR: virtual reality.

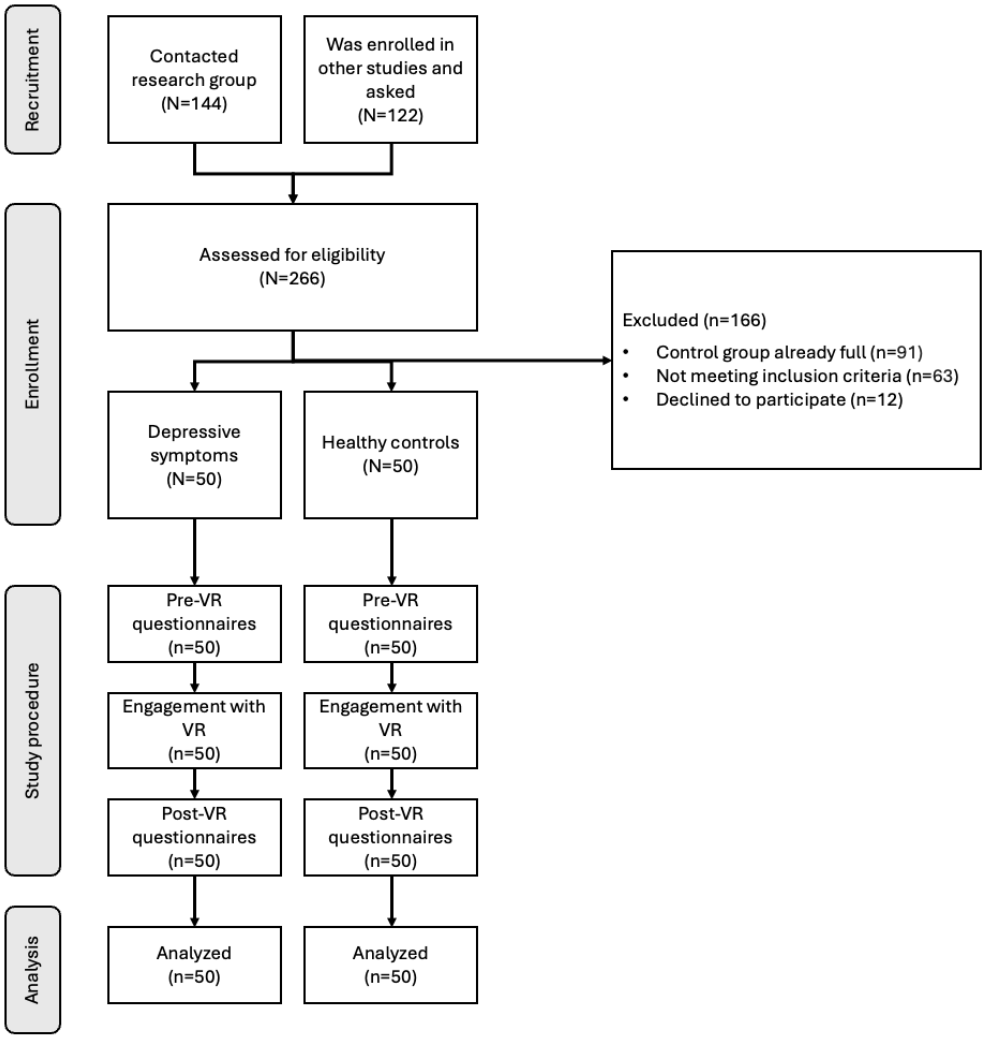

XSL•FO
RenderX

Table 1. Demographic and clinical characteristics of study participants by study groups.

	DS[ab] (N=50)	HC[c] (N=50)	Between group difference, r_{rb}[d]
Sex, n (%)			
Male	10 (20)	10 (20)	—[e]
Female	40 (80)	40 (80)	—
Age (years), mean (SD)	23.0 (1.86)	23.3 (1.52)	0.13
Education (years), mean (SD)	16.1 (1.40)	16.6 (0.78)	0.14
Hand dominance, n (%)			
Left	0 (0)	3 (6)	—
Right	50 (100)	47 (94)	—
Sight correction, n (%)			
None	21 (42)	26 (52)	—
Lenses	10 (20)	10 (20)	—
Glasses	19 (38)	14 (28)	—
Height (meter), mean (SD)	1.67 (0.07)	1.66 (0.07)	0.07
Weight (kg), mean (SD)	62.7 (10.60)	59.6 (7.85)	0.12
BMI, mean (SD)	22.4 (3.00)	21.5 (2.06)	0.13
Habitual sleep (hours/night), mean (SD)	7.0 (0.98)	7.3 (0.88)	0.20
Habitual smoker, n (%)			
Yes	19 (38)	20 (40)	—
No	31 (62)	30 (60)	—
Habitual consumption of alcohol, n (%)			
Yes	11 (22)	16 (32)	—
No	39 (78)	34 (68)	—
Habitual user of psychoactive drugs, n (%)			
Yes	11 (22)	4 (8)	—
No	39 (78)	46 (92)	—
PHQ-9[f], mean (SD)	11.3 (1.92)	3.2 (1.48)	0.87[g]
DASS-D[h], mean (SD)	15.2 (7.40)	3.5 (3.44)	0.77[g]
DASS-A[i], mean (SD)	9.6 (6.62)	2.4 (2.89)	0.61[g]
DASS-S[j], mean (SD)	17.8 (6.73)	7.7 (5.08)	0.66[g]
PANAS-Pos[k] before VR[l], mean (SD)	10.4 (3.53)	12.2 (2.76)	0.27[m]
PANAS-Pos after VR, mean (SD)	8.9 (4.36)	11.8 (3.61)	0.36[g]
PANAS-Neg[n] before VR, mean (SD)	3.2 (2.98)	1.9 (2.22)	0.24[o]
PANAS-Neg after VR, mean (SD)	3.2 (3.25)	1.7 (2.24)	0.24[o]

[a]The categorization is based on current depressive symptom severity assessment via the Patient Health Questionnaire-9 (score equal to 9 or above) and not the presence of a clinical diagnosis.

[b]DS: depressive symptom.

[c]HC: healthy control.

[d]Group differences are calculated via 2-tailed Mann Whitney *U* tests and effect sizes reflect rank-biserial (r_{rb}) correlations.

[e]—: not applicable.

XSL•FO
RenderX

[f]PHQ-9: Patient Health Questionnaire-9.
[g]$P<.001$.
[h]DASS-D: Depression Anxiety Stress Scale-Depression.
[i]DASS-A: Depression Anxiety Stress Scale-Anxiety.
[j]DASS-S: Depression Anxiety Stress Scale-Stress.
[kj]PANAS-Pos: Positive and Negative Affect Schedule – Positive affect
[l]VR: virtual reality.
[m]$P<.01$,
[n]PANAS-Neg: Positive and Negative Affect Schedule – Negative affect
[o]$P<.05$,

Power Calculation

A post hoc power analysis was conducted to assess the statistical power of the study based on the observed sample size of 100 participants (50 per group). Assuming a moderate effect size (Cohen d=0.50) and an α level of .05, the analysis indicated that the study has 0.7 power to detect between-group differences as visualized in Figure 2.

Figure 2. Power analysis for between-group comparison (GPower).

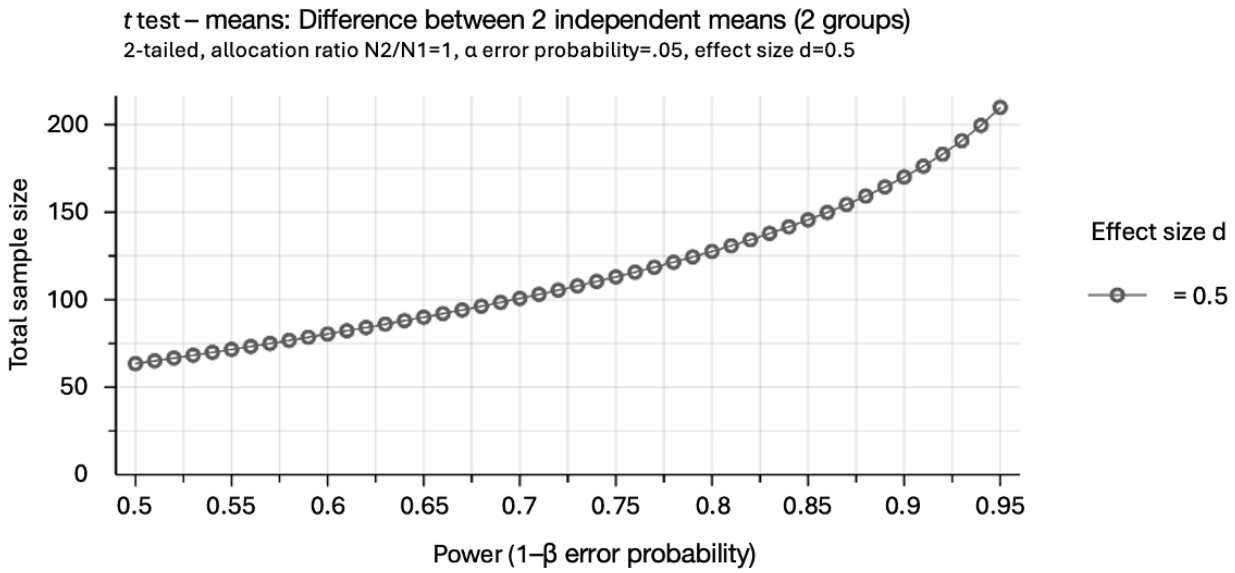

Inclusion and Exclusion Criteria

The screening for inclusion involved the assessment of current DS severity using the Patient Health Questionnaire-9 (PHQ-9; Kroenke et al [26]). Participants scoring 9 or above—constituting the upper limit for mild DSs [26]—were assigned to the DS, while those scoring 5 or below were registered for the HC group (Figure 3).

For inclusion in either group, participants had to be free from any condition that would impair their ability to interact with the VR environment (eg, visual impairment without the possibility of correction via lenses) and able to provide written informed consent. Additional inclusion criteria for the DS group included the lack of diagnosis of any other psychiatric disorders than depressive or anxiety disorders, and treatment stability over the last 4 weeks if they received psychological or pharmacological treatment at the time of assessment. Participants in the control group had to be free from any previous or current psychiatric disorders.

XSL•FO
RenderX

Figure 3. Distribution of participants' scores on PHQ-9 assessing current depressive symptom severity. PHQ-9: Patient Health Questionnaire-9.

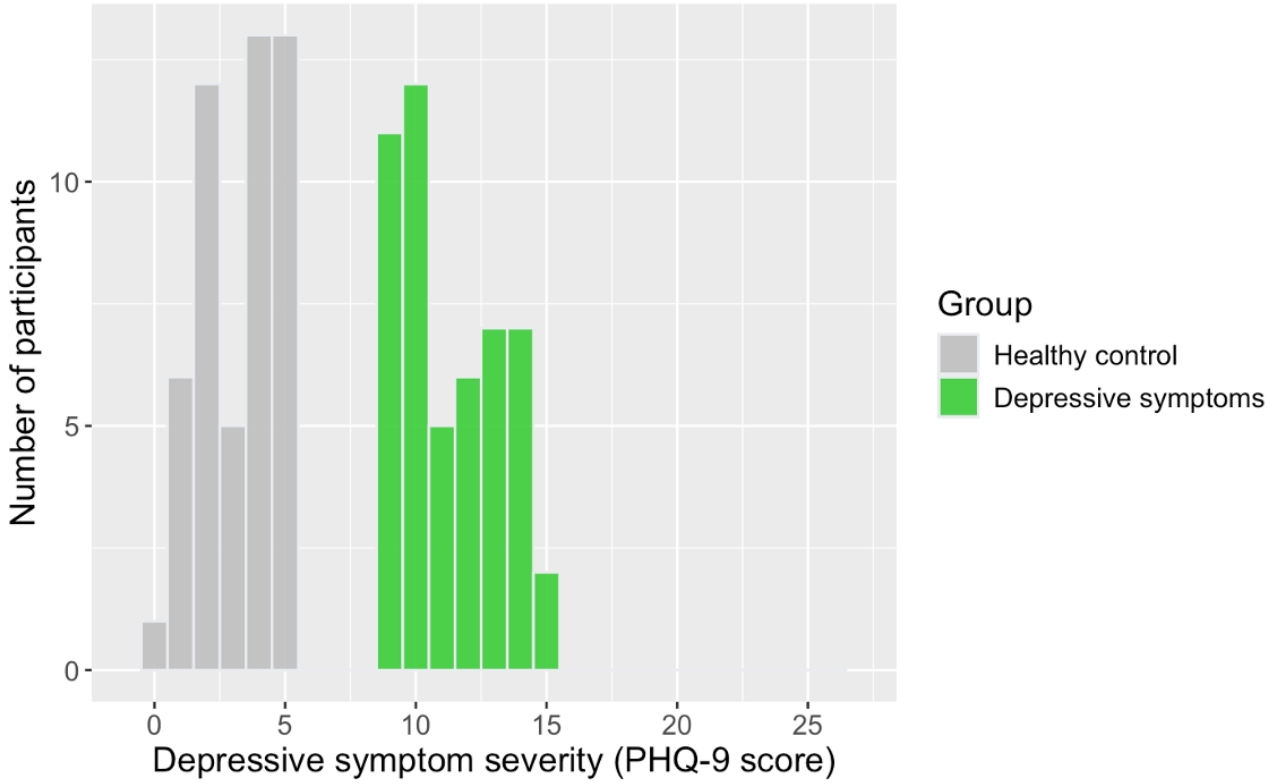

Study Procedure

After providing informed consent and completing the pre-VR battery questionnaires, all participants were seated in a room equipped with the VR system. Following a headset and eye-tracking calibration, a tutorial was completed (aimed to help the understanding of how to move around and interact with the environment). Next, participants were free to explore the VR environment on their own terms but in the presence of a study administrator. The study session was concluded once the post-VR battery was completed.

Technological Specifications

Overview

The VR system was developed using the HP Reverb G2 Omnicept Edition headset, which includes a variety of biometric sensors such as eye tracking, a facial camera (lower), a heart rate sensor, and sensors for brain activity while maintaining the core features of the standard version. These features include a resolution of 2160 × 2160 px per eye, a 114° field of view, a 90 Hz refresh rate, and integrated audio.

Environment Development

The virtual environment was developed entirely in Unity 2020.3.39 LTS, using the HP Omnicept SDK and OpenXR to maximize compatibility and platform versatility. Several development processes were used to ensure a smooth and efficient environment, including occlusion culling, foveated rendering, baked lighting, and asset optimization.

The scenario consists of 4 rooms, each with distinct lighting parameters, as well as an additional room that can be visually explored through a keyhole in a door. Volume postprocessing was used to modify the lighting without affecting performance, and the keyhole effect was simulated using camera layers.

Within each room, various stimuli and interactive elements are present, such as sound boxes, a chalkboard, and interactable objects (eg, items that can be picked up and thrown, musical buttons, and photos that can be picked up for closer inspection). Every interaction between the participant and the environment is recorded, including eye tracking data, particularly in relation to elements designated as areas of interest or with which the subject can interact (eg, ENTERING_ROOM, LEAVING_ROOM, TAKE, DROP, TRIGGER, INTERACT).

Administrator Control

To manage the experiment, the study administrator is provided with an interface, enabling them to observe the participant's actions in real-time. The interface also offers various options to ensure that the experimental session proceeds smoothly (eg, it is possible to discontinue a cognitive task if the participant requests so). This control system is facilitated by an additional camera mounted on the subject, enabling the administrator to monitor the experiment's progression effectively.

Virtual Environment

The goal of the pilot VR environment is to differentiate between HCs and those with DSs based on cognitive, behavioral, and physiological data. The development and performance of the system for this purpose will be discussed in more detail in another paper [25], but the main characteristics will be summarized here for a general overview.

The environment was designed to elicit certain behaviors and record measures, some of which were previously shown, and others are hypothesized to be connected to depression [25]. The

XSL•FO
RenderX

domains included cognition (attention, working memory, processing speed, executive functioning, and cognitive flexibility), metacognition, persistence or grit, curiosity, as well as behavioral and attentional biases toward negative stimuli, speed, and movement measurements. Physiological data included skin conductance, heart rate variability, and eye-tracking–related measures.

The environment resembled a multiroom family home (Figure 4). Exposure started with a tutorial introducing participants to controllers as well as how to move around in the environment and interact with objects. Following the completion of the tutorial, participants were free to explore the environment but had to solve cognitive tasks to open doors and move between rooms. There were 4 cognitive tasks in total: Rapid Visual Information Processing; N-back [2-back]; Trail Making Test—parts A and B; and Wisconsin Card Sorting Test. The exposure ended on the participant's own terms, but the exit door only opened once all cognitive tasks were completed.

Multimodal data collected during the VR sessions for the purpose of DS assessment were analyzed using a CatBoost machine learning algorithm (Yandex LLC) [27]. Model performance was evaluated based on accuracy, and feature importance was determined through an explainable artificial intelligence approach [28], ensuring transparency in the decision-making process. For more details, see the study by Mura et al [25].

Figure 4. The visual illustration of the virtual reality environment designed for the assessment of cognition and behavior hypothesized to be related to depression.

Tools and Questionnaires

Pre-VR Questionnaires

The pre-VR battery included a background questionnaire (demographic data and diagnosis, medication); the Patient Health Questionnaire (PHQ-9) [26] to assess current DS severity; the Depression, Anxiety, Stress Scale (DASS) [29] to assess current depressive, anxiety and stress symptom severity; the Positive and Negative Affect Schedule (PANAS-SF) [30] to evaluate emotional state; as well as the Curiosity and Explorations Inventory-II [31].

Post-VR Questionnaires

The post-VR battery repeated the PANAS-SF questionnaires and included questionnaires regarding, acceptability, usability, and cybersickness, described in more detail in the later sections.

Acceptability

The construct of acceptability was assessed based on the theoretical framework of acceptability (TFA) [18,32]. The acceptability of the concept and the acceptability of the pilot system were evaluated separately. The acceptability of the concept, that is, the use of VR for the purposes of mental health screening, diagnosis, and treatment (definitions were provided) was assessed via self-report Likert items. Scores ranged from 1 to 5, indicating answers from "Strongly disagree" to "Strongly agree" on statements about VR being acceptable (Multimedia Appendix 1). As the primary purpose of the pilot system was the assessment of DSs, the diagnostic use case was investigated via 2 additional questions.

- Support tool: Willingness to engage with the system to inform diagnosis made by a mental health professional.
- Stand-alone tool: Willingness to engage with the system to receive a diagnosis without extra input from a mental health professional.

The dimensions of acceptability concerning the pilot system were assessed within the TFA and focused on general acceptability, affective attitude, perceived effectiveness, intervention coherence, self-efficacy, burden, ethicality, and opportunity cost. The construct of affective attitude was further broken down into two questions: (1) To what extent a participant likes or dislikes the system, and (2) how comfortable or uncomfortable they find the system to be. For ease of interpretability, the rating scale was unionized across all constructs, where "1" represented the "undesirable" rating (eg,

XSL•FO
RenderX

low levels of comfort; or the need for high levels of effort indicated burden), and "5" reflected the "desirable" rating (eg, high levels of comfort; or low levels of effort reported under burden). See Multimedia Appendix 1 for details about the specific items used to assess each of these constructs.

Usability

Usability was assessed by the System Usability Scale (SUS) [33]—which is a postinteraction self-report questionnaire to assess user perceptions of a system's overall usability. The scale is composed of 10 items rated on a Likert scale from 1 to 5, from "strongly disagree" to "strongly agree." The total score ranges from 0 to 100, with higher scores indicating higher usability ratings. There are multiple frameworks to interpret the SUS score (range, adjective, and grade) [34], but the primary threshold for this study was the average score necessary for an 'OK/satisfactory' experience (a score of 50.9 [35]).

Cybersickness

To assess levels of cybersickness following VR exposure, the Simulator Sickness Questionnaire (SSQ) was used [36]. The questionnaire contains 16 symptoms, which are rated on a scale of subjective severity: 0=none, 1=slight, 2=moderate, and 3=severe. Beyond the total score, one can attain separate scores for the nausea, oculomotor disturbances, and disorientation subscales.

However, while the SSQ was once considered to be the "gold standard," it has increasingly been criticized, including the original thresholds for the interpretation of the severity levels [37-40]. For this reason, the comparison of the pilot system results was made to the literature average [23] instead of the original thresholds.

Analysis

All analyses were carried out in R supplemented by R Studio (version 2023.12.1.402; R Foundation for Statistical Computing), except for the power calculation for which G*Power (version 3.1; Heinrich Heine University Düsseldorf) was used. Multiple packages were used to supplement analysis and visualization, these include but are not limited to: *dplyr* [41], *vioplot* [42], *ggplot2* [43], *psych* [44], *rstatix* [45], *ggeffects* [46], *ordinal* [47], and *VGAM* [48]. See Multimedia Appendix 2 for the script, its output, and a full list of packages.

Within-group comparisons and comparisons to known literature values were conducted using Wilcoxon signed rank tests, while between-group comparisons used Mann-Whitney *U* tests. These nonparametric methods were chosen to account for the ordinal nature of the outcome measures, such as Likert scales, and to address the observed violations of the assumptions required for parametric tests (eg, normality assessed via the Shapiro-Wilk test). A nominal significance level (α) of .05 was applied. For ease of interpretability, effect sizes were consistently reported in *r* values. This includes the use of rank-biserial correlations for Mann-Whitney *U* tests as well as Wilcoxon signed rank tests for quantifying the magnitude of the differences in a standardized manner. The full statistics are available in Multimedia Appendix 2.

Multivariate linear regression models were used to analyze predictors when the outcome variable was continuous, such as SUS or SSQ scores. For ordinal outcome measures, such as TFA constructs, proportional odds models were applied. In cases where the proportional odds assumption was not met (as noted in the Results section), results should be interpreted with caution. All regression models included age, sex, and education as covariates. When additional predictors, such as cybersickness severity scores, group labels, or self-reported anxiety and stress levels were relevant, these were included alongside the demographic variables.

Ethical Considerations

The study was carried out in accordance with the Declaration of Helsinki and obtained ethical approval from the local ethical committees on all sites where data was collected (Italy) or analyzed (Italy and Sweden). In Italy, approval was granted by Comitato Etico della Ricerca Psicologica, and in Sweden by the Etikprövningsmyndigheten (2023-00959-01).

Written informed consent was sought from all participants, with emphasis placed on the voluntary nature of participation and the possibility to withdraw from the study at any time, without reason, or any penalty. All participants were informed about data privacy and confidentiality, whereby all data obtained during the study were analyzed anonymously and used solely for the purpose of scientific research. Compensation included €25 (worth between US $24.38 and US $28.14 over the duration of the recruitment period).

Preregistration

The study was preregistered in the ISRCTN registry (ISRCTN16396369) [49]. The sole modification to the preregistered protocol concerned the threshold for inclusion on current DS severity. Due to the difficulty in recruiting volunteers in line with the study's timeline, the upper threshold for inclusion in the control group was increased from 4 to 5, while the lower threshold to be included in the group with DSs was decreased from 10 to 9 on the PHQ-9.

Results

Acceptability of the Concept

The use of VR for mental health treatment (62/84, 74%) and screening (65/84, 77%) was endorsed by the majority of those providing an answer, whereas only a minority indicated support for diagnostic purposes (28/83, 34%; Figure 5). There were no significant differences between the DS and HC groups.

Comparing the use of VR with or without input from a mental health professional, the responses indicate higher support for the VR system to be used as a support tool (62/83, 75%) compared to a stand-alone solution (15/84, 18%). A Wilcoxon signed rank test indicated a significant difference of large size (V=1770; P<.001, 2-tailed; r=0.81, 95% CI 0.76-0.84; $\text{median}_{\text{support}}$ 4, $\text{IQR}_{\text{support}}$ 3.5-4.5; $\text{median}_{\text{stand-alone}}$ 2, $\text{IQR}_{\text{stand-alone}}$ 1.0-3.0). For the raw distribution of responses see Figure 6. Again, no difference was observed between the DS and HC groups.

XSL•FO
RenderX

Figure 5. Participant ratings of VR acceptance for the purposes of mental health screening, diagnosis, and treatment. N/A: not available; VR: virtual reality.

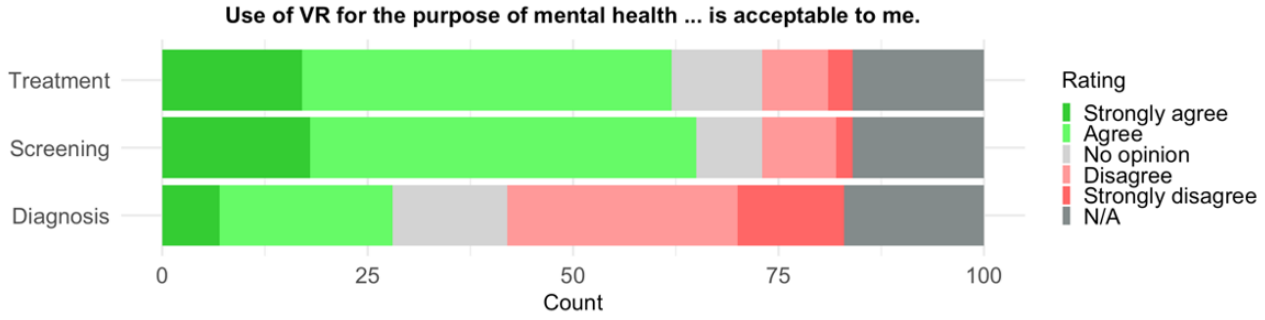

Figure 6. Participants' subjective ratings of their willingness to engage with the pilot VR system for mental health diagnosis. The support tool complements the assessment of a mental health professional, whereas the standalone tool operates without professional input. N/A: not available; VR: virtual reality.

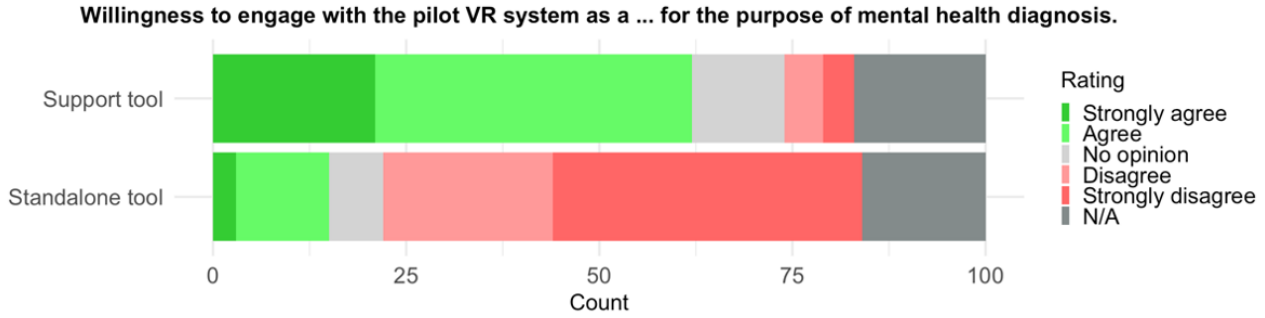

Acceptability of the Pilot System

Behavioral measures of acceptability include the discontinuation of participation for any reason, as well as the need to take a break from the study. While the possibility to discontinue participation at any point, without any consequences was stated before enrollment, all 100 volunteers completed the study. The need for a break was expressed by 1 participant (1/100, 1%) due to discomfort. They decided to resume participation despite the possibility of opting out being reemphasized. On average, 31 minutes were spent in the virtual environment.

Acceptability was also assessed via a self-report questionnaire specifically formulated for the project based on the TFA [18,32]. Ratings, visualized in Figure 7, indicated generally high levels of acceptability.

Figure 7. Self-reported acceptability of the pilot virtual reality system, assessed using constructs from the theoretical framework of acceptability (1=undesirable rating; 5=desirable rating). N/A: not available.

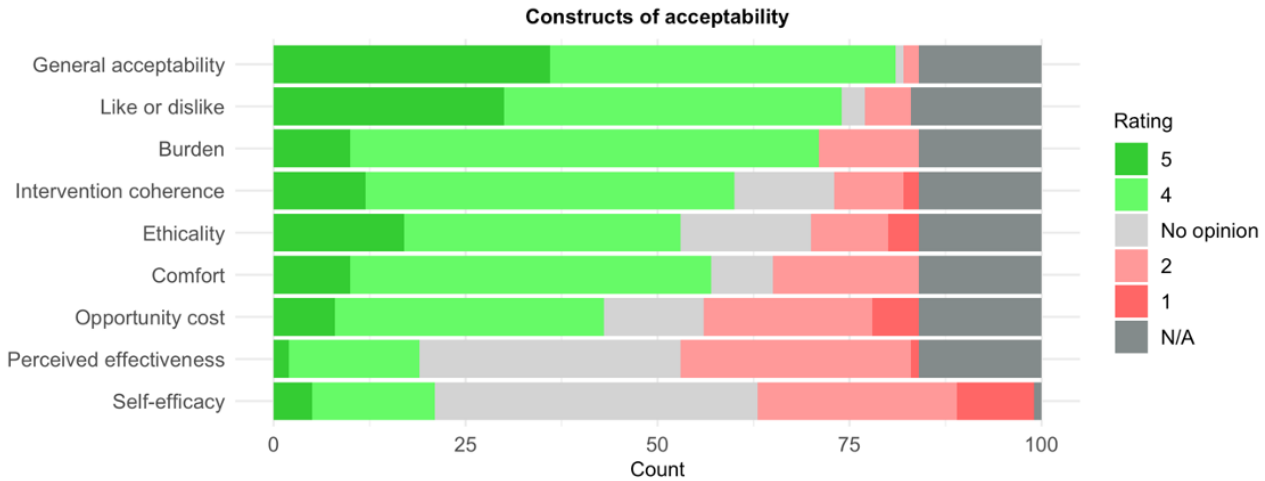

A majority of the participants found the system itself acceptable (81/84, 96%), liked the system (74/83, 89%), indicated that the engagement took little to no effort (71/84, 85%), found it clear how the system might help diagnose depressive disorders (60/84, 71%), found no moral or ethical consequences with the system (53/84, 63%), found it comfortable (57/84, 68%), and expressed no concern about missing out on other alternatives when engaging with the system (43/84, 51%). The 2 constructs on which the undesirable ratings (1 or 2) outnumbered desirable scores (4 or 5) concerned the perceived effectiveness of the

XSL•FO
RenderX

system in improving mental state (19/84, 23%), and self-efficacy regarding feeling confident while engaging with the system (21/99, 21%).

Multivariate ordinal regression models were used to explain variance in ratings of all the acceptability constructs as detailed in Table 2. Sex was a significant predictor of general acceptability and subjective burden, where women reported considerably lower levels of acceptability ($mean_{men}$ 4.65, SD_{men} 0.49; $mean_{women}$ 4.30, SD_{women} 0.65) and reported lower ratings on burden—meaning a considerably higher effort was required from their part to engage with the system ($mean_{men}$ 4.29, SD_{men} 0.50; $mean_{women}$ 3.69, SD_{women} 0.87). Education did not reach significance for any of the TFA constructs, and age was marginally significant only when used to explain whether participants found moral or ethical consequences to engaging with the system.

In the next stage, group label (DS vs HC) was entered into the same models in addition to the demographic variables. Group label was a significant predictor of comfort levels ($mean_{HC}$ 3.83, SD_{HC} 0.78; $mean_{DS}$ 3.34, SD_{DS} 1.08) and marginally significant for the like or dislike rating ($mean_{HC}$ 4.38, SD_{HC} 0.59; $mean_{DS}$ 4.00, SD_{DS} 0.95)—meaning that participants in the DS group reported lower comfort levels and liked the system to a lesser extent.

Finally, cybersickness severity (total SSQ score) was entered into models explaining all TFA constructs, in addition to the demographic variables. Cybersickness was a significant predictor for comfort, burden, opportunity cost, and perceived effectiveness and marginally significant for self-efficacy (Table 2). All showed an inverse relationship, whereas the higher the cybersickness symptom severity, the lower acceptability scores were reported on all TFA dimensions.

In summary, both the concept and the system received high acceptability ratings. Most participants supported the use of VR for mental health screening and treatment. However, regarding diagnosis, participants preferred VR as a supportive tool rather than a stand-alone solution. The primary area for improvement identified was enhancing users' self-efficacy while interacting with the system. Sex and cybersickness emerged as significant predictors of acceptability ratings.

XSL•FO
RenderX

Table 2. Estimates of predictors from 3 different models analyzed via multivariate ordinal regressions to explain constructs investigated within the theoretical framework of acceptability.

Model	Predictor	Estimate (log odds)	95% CI	*P* value
General acceptability				
Model 1	Age	0.175	–0.19 to 0.56	.36
Model 1	Sex	–1.331	–2.54 to –0.21	.02[a]
Model 1	Education	–0.529	–1.10 to 0.01	.053
Model 2	Group (DS[b] vs HC[c])[d]	–0.261	–1.19 to 0.65	.58
Model 3	Cybersickness	–0.004	–0.02 to 0.01	.54
Like or dislike				
Model 1	Age	–0.298	–0.67 to 0.05	.10
Model 1	Sex	–0.461	–1.52 to 0.58	.39
Model 1	Education	0.164	–0.34 to 0.67	.52
Model 2	Group (DS vs HC)[d]	–0.890	–1.81 to 0.00	.053
Model 3	Cybersickness	–0.010	–0.03 to 0.01	.20
Comfort				
Model 1	Age	–0.045	–0.38 to 0.29	.79
Model 1	Sex	–0.718	–1.84 to 0.37	.20
Model 1	Education	0.116	–0.36 to 0.58	.63
Model 2	Group (DS vs HC)[d]	–0.899	–1.81 to –0.02	.048[a]
Model 3	Cybersickness	–0.026	–0.04 to –0.01	.001[e]
Burden				
Model 1	Age	0.137	–0.26 to 0.54	.50
Model 1	Sex	–1.866	–3.26 to –0.59	.005[e]
Model 1	Education	–0.186	–0.78 to 0.38	.53
Model 2	Group (DS vs HC)[d]	–0.273	–1.28 to 0.72	.59
Model 3	Cybersickness	–0.036	–0.06 to –0.02	<.001[f]
Intervention coherence				
Model 1	Age	0.085	–0.25 to 0.42	.62
Model 1	Sex	–0.315	–1.42 to 0.77	.57
Model 1	Education	–0.193	–0.69 to 0.29	.44
Model 2	Group (DS vs HC)[d]	–0.567	–1.44 to 0.29	.20
Model 3	Cybersickness	–0.003	–0.02 to 0.01	.71
Ethicality				
Model 1	Age	0.320	–0.00 to 0.65	.053
Model 1	Sex	0.241	–0.74 to 1.23	.63
Model 1	Education	–0.166	–0.66 to 0.33	.51
Model 2	Group (DS vs HC)[d]	0.170	–0.63 to 0.98	.68
Model 3	Cybersickness	–0.010	–0.02 to 0.00	.15
Opportunity cost[g]				
Model 1	Age	–0.078	–0.39 to 0.24	.63
Model 1	Sex	–0.496	–1.53 to 0.50	.34

XSL•FO
RenderX

Model	Predictor	Estimate (log odds)	95% CI	*P* value
Model 1	Education	0.023	–0.44 to 0.49	.92
Model 2	Group (DS vs HC)[d]	0.336	–0.46 to 1.14	.41
Model 3	Cybersickness	-0.019	–0.03 to –0.01	.009[e]
Perceived effectiveness				
Model 1	Age	–0.116	–0.44 to 0.20	.48
Model 1	Sex	–0.757	–1.79 to 0.25	.14
Model 1	Education	–0.045	–0.53 to 0.43	.85
Model 2	Group (DS vs HC)[d]	–0.548	–1.38 to 0.27	.19
Model 3	Cybersickness	–0.019	–0.04 to –0.00	.02[a]
Self-efficacy[h]				
Model 1	Age	0.063	–0.24 to 0.37	.68
Model 1	Sex	–0.872	–1.86 to 0.08	.08
Model 1	Education	0.031	–0.40 to 0.48	.89
Model 2	Group (DS vs HC)[d]	0.043	–0.70 to 0.79	.91
Model 3	Cybersickness	-0.01	–0.03 to 0.00	.08

[a]*P*<.05.

[b]DS: depressive symptom.

[c]HC: healthy control.

[d]The categorization is based on current depressive symptom severity assessment via the PHQ-9 questionnaire (score equal to 9 or above for depressive symptoms [DS] and 5 or below for healthy controls [HC]) and not the presence of a clinical diagnosis.

[e]*P*<.01.

[f]*P*<.001.

[g]The assumption of proportional odds not was not fulfilled for the model. Results should be interpreted with caution.

[h]Self-efficacy was measured with Item 9 of the System Usability Scale [33], as opposed to the other items that were specifically designed for the assessment of the pilot system based on the theoretical framework of acceptability [18,32].

Usability

Descriptives

Results from 99 participants indicated average ratings of usability, with a mean score of 69 (SD 12.86), and a range of 37.5-97.5 on SUS [33]. As visualized in Figure 8, which showcases the distribution of overall SUS scores, 88% (87/99) of participants provided a rating of 50.9 or above, indicating an OK or satisfactory experience or better [35].

Figure 8. Density plot of overall usability ratings on the System Usability Scale from all participants.

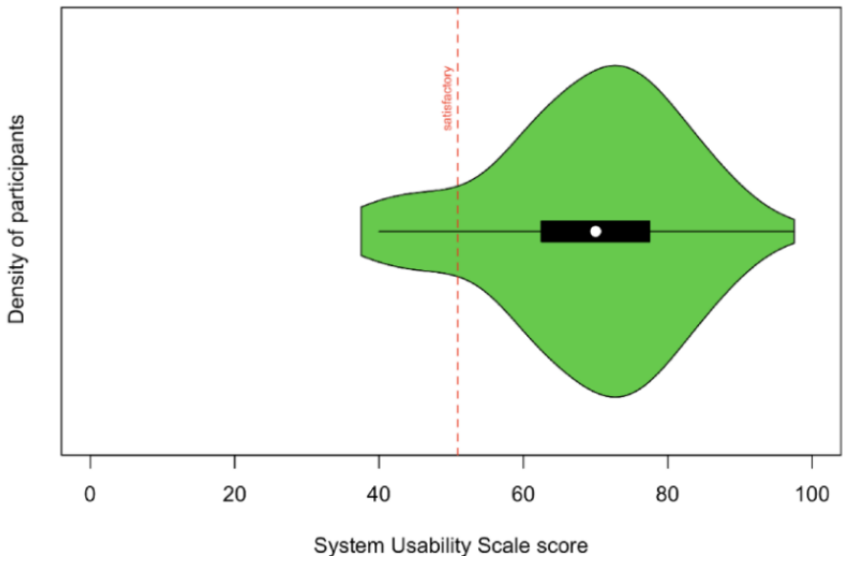

XSL•FO
RenderX

Analysis

A multiple linear regression analysis was conducted to predict usability ratings based on participants' age, education, gender, and group (DS vs HC). The regression model was not statistically significant ($F_{4,94}$=1.23; P=.31), explaining only about 5% of the variance in SUS scores (R^2=0.05, adjusted R^2=0.01). The strongest predictor was gender, but this only showed a marginal effect (β=–5.79, 95% CI –12.38 to 0.80; P=.08) with men reporting higher usability ratings than women.

A second model was tested examining the relationship between usability scores and the predictors of age, education, sex, and total cybersickness severity. The model was significant ($F_{4,94}$=2.97; P=.02), accounting for 11.2% of the variance in SUS scores (R^2=0.11, adjusted R^2=0.07). Cybersickness severity (SSQ scores) emerged as the only significant predictor (β=–0.12, 95% CI –0.21 to –0.03; P=.007), indicating that participants with higher SSQ scores tended to have lower SUS scores.

In summary, the system's overall usability was rated as sufficient, with the majority of participants reporting a satisfactory experience. Cybersickness severity emerged as the only predictor of usability ratings, showing a negative association.

Cybersickness

Descriptives

The overall mean SSQ score combining all subscales was 30.49 (SD 28.03), with individual scores ranging from 0 to 119.68. This is comparable to the literature average of 28.00 reported by Saredakis et al [23], on the overall sample, as well as the DS subsample. However, the HC group reported considerably lower levels of cybersickness severity (V=402; P=.02; r_{rb}=–0.37, 95% CI –0.61 to –0.07) as summarized in Table 3.

For the SSQ subscales, the results of the total sample reflect severity levels not (significantly) different from the literature average [23] regarding nausea and disorientation levels, whereas oculomotor disturbances were rated considerably higher in this study (V=3296; P=.004; r_{rb}=0.33, 95% CI 0.12-0.52).

When the comparison was split by study groups, only the DS group deviated significantly from the literature average—specifically on the nausea (V=917; P=.002; r_{rb}=0.50, 95% CI 0.22-0.70) and the oculomotor disturbances (V=967; P<.001; r_{rb}=0.58, 95% CI 0.33-0.75) subscales. The HC group showed near-average levels on all 3 subscales. The distributions of raw scores, as boxplots, along the total and all subscales are illustrated in Figure 9 and analyzed for between-group differences in the later section.

Table 3. Comparison of mean cybersickness severity scores (SSQa; mean [SD]) via Wilcoxon-signed rank tests across the total sample, those with DSb and HCc subsamples with literature averages [23].

SSQ Scale	Literature average[d]	Total sample, mean (SD)	DS[e], mean (SD)	HC, mean (SD)
Total	28.00	30.49 (28.03)	39.39 (32.89)	21.77 (18.84)[f]
Nausea	16.72	21.59 (21.96)	29.40 (24.39)[g]	13.93 (16.15)
Oculomotor disturbances	17.09	26.72 (24.63)[g]	33.57 (27.50)[h]	20.01 (19.48)
Disorientation	23.50	32.90 (39.84)	42.04 (48.34)	23.94 (26.83)

[a]SSQ: Simulator Sickness Questionnaire.

[b]DS: depressive symptom.

[c]HC: healthy control.

[d]Literature averages are taken from Saredakis et al [23] based on a meta-analysis of 55 studies on approximately 3000 participants.

[e]The categorization is solely dependent on the Patient Health Questionnaire-9 score and does not indicate the presence of a clinical diagnosis.

[f]P<.05.

[g]P<.01.

[h]P<.001.

XSL•FO
RenderX

Figure 9. Boxplot of self-reported cybersickness severity scores for the HC and DS study groups—measured by the (A) SSQ, and its 3 subscales of (B) nausea, (C) oculomotor disturbances, and (D) disorientation. DS: depressive symptom; HC: healthy control; SSQ: Simulator Sickness Questionnaire.

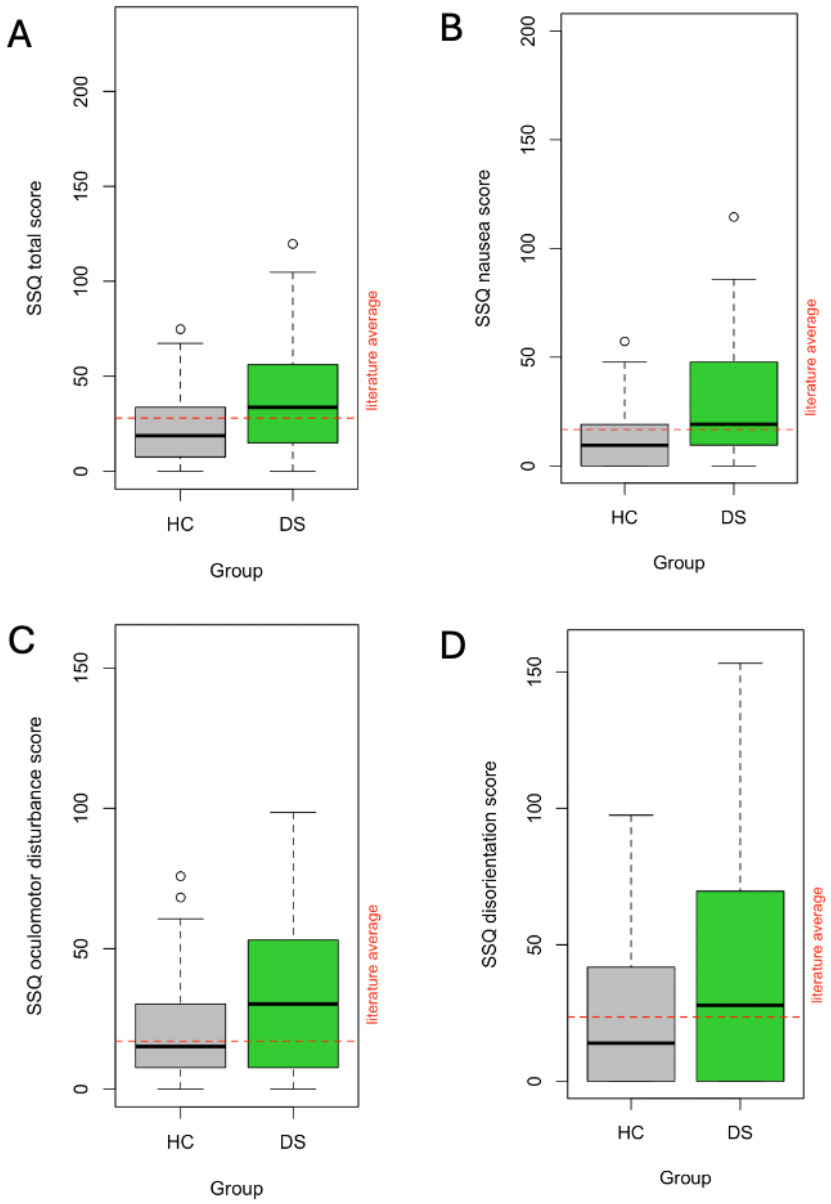

Analysis

A multiple linear regression was conducted to predict SSQ scores based on age, education, sex, and group (DS vs HC). The model was statistically significant ($F_{4,94}$=3.20; P=.02), accounting for approximately 12% of the variance in SSQ scores (R^2=0.12, adjusted R^2=0.08). The only predictor reaching significance was group label, (β=17.58, 95% CI 6.57-28.60; P=.002), indicating that participants in the DS group had considerably higher SSQ scores (mean 39.38, SD 32.89) than the HC group (mean 21.77, SD 18.84). These between-group differences remained significant for all subscales analyzed separately.

When DS severity was entered as a continuous variable instead of a group label, a further 2% variance in cybersickness severity was explained (R^2=0.14, adjusted R^2=0.11) and a linear prediction of SSQ scores along the PHQ-9 scores was attainable as visualized in Figure 10A. The graph shows all participants along the SSQ and PHQ-9 scales and illustrates a positive association between the 2 scores when the regression line is plotted.

However, due to the baseline group differences in anxiety and stress levels (Table 1) this prediction might be biased. While DASS depression was not of primary interest, a multiple linear regression model with DASS depression as an outcome was used, to be able to adjust for DASS anxiety and DASS stress levels in addition to age, sex, and education levels as covariates. This model was also significant ($F_{6,92}$=4.05; P=.001) and explained about 21% of the variance in cybersickness severity (R^2=0.21, adjusted R^2=0.16). The only predictor emerging as significant was DASS depression (β=1.15, 95% CI 0.29-2.01; P=.009), whereas DASS anxiety reached marginal significance (β=1.2, 95% CI –0.07 to 2.47; P=.06). Cybersickness severity predictions showed a similar trend as for the PHQ-9 scores when the regression line was illustrated in Figure 10B.

XSL•FO
RenderX

Figure 10. Predicted cybersickness severity levels (with 95% CIs) based on depressive symptom severity, as measured by: (A) PHQ-9 and (B) DASS. DASS: Depression Anxiety Stress Scale; DASS-D: Depression Anxiety Stress Scale-Depression; PHQ-9: Patient Health Questionnaire; SSQ: the Simulator Sickness Questionnaire.

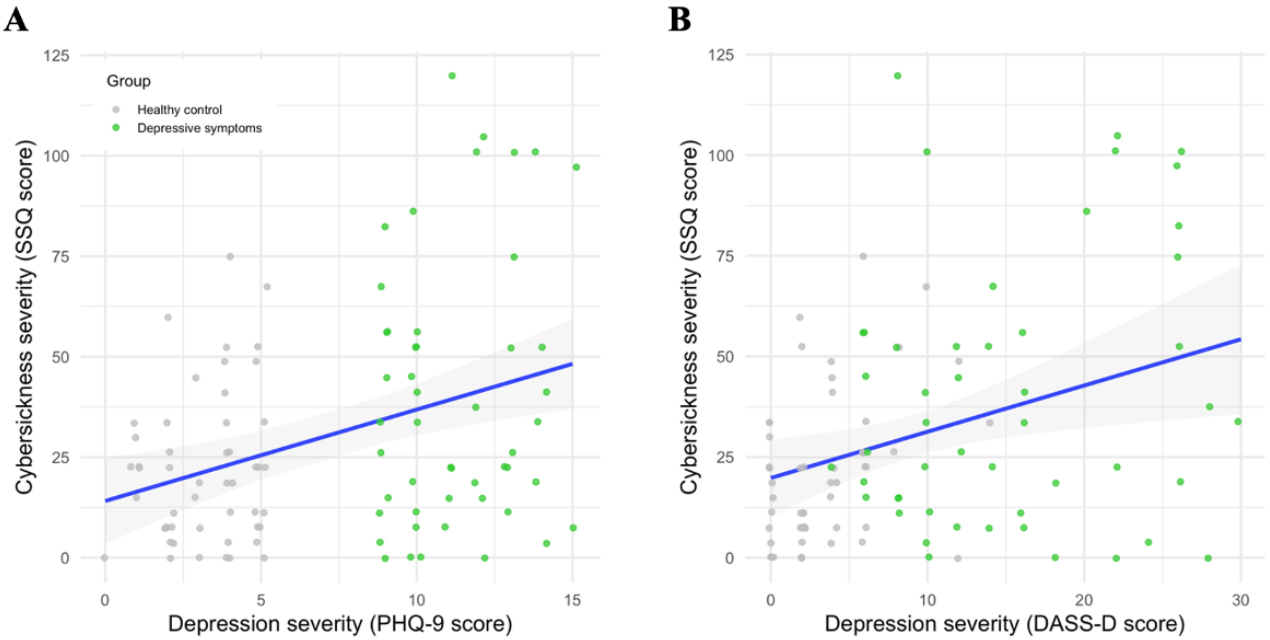

In summary, the severity of cybersickness was comparable to the literature averages for nausea and disorientation but significantly higher for oculomotor disturbances, particularly in the DS group. Across all subscales, the DS group reported higher SSQ scores than the HC group, with group membership emerging as the strongest predictor of cybersickness severity. A positive association was identified between DS severity (PHQ-9) and cybersickness, explaining a modest but significant proportion of variance. Even after adjusting for baseline anxiety and stress levels (DASS), DSs remained a significant predictor of cybersickness.

Discussion

Principal Findings

This study reports on the acceptability, usability, and cybersickness levels of a novel VR environment designed for the assessment of DSs. Following an approximately 30-minute engagement, 50 HCs and 50 individuals with moderate self-reported DSs completed a self-report battery on 3 constructs of interest.

Acceptability

In general, VR systems were rated acceptable by the majority of participants; both for the purpose of mental health screening as well as treatment. For diagnostics, a clear preference emerged for VR to be used as a support tool by a health care professional, as opposed to a stand-alone solution. While there is no literature available on VR for depression with which a direct comparison can be made, these findings are in line with previous results concerning other digital solutions, such as mHealth apps [50].

The specific pilot system tested here was designed for the purpose of differentiating between HCs and those with DSs [25]. The system itself was rated acceptable by most participants, but the need for improvement was apparent; specifically when it came to participants' confidence levels while using the system. Nevertheless, the high level of acceptance in practice found in the current study is in line with 2 studies previously reported on the acceptability of specific VR systems for therapeutic purposes for depression [21,22].

Usability

Usability was rated satisfactory by most of the participants, with cybersickness severity emerging as a significant predictor of usability scores. We found 1 previous study reporting on the usability of VR systems for depression, which has also indicated favorable ratings on ease of use and perceived usefulness [21]. However, in comparison with other psychiatric conditions, such as social anxiety disorder [51], psychosis [52], or posttraumatic stress disorder [53], it is important to emphasize that systematic evaluations of the usability of VR among those with depression are lacking.

Concerning the effect of cybersickness severity on usability, there are no previous studies reporting in relation to depression. Additionally, while comprehensive reviews on the relationship between cybersickness and presence are available [54] there is no such overview for usability. Existing single studies report inconsistent findings across different conditions. For example, a negative association is reported on healthy individuals [55], but no effect was found among those with Huntington disease [56].

These findings highlight the need for greater attention from the scientific community, as depression remains relatively understudied in the context of VR use compared to other psychiatric conditions [12,57]. Notably, participants in this study expressed support for the technology both conceptually and in relation to this specific pilot system. Identifying novel approaches with user support is particularly important for groups

XSL•FO
RenderX

that might find treatment adherence and lack of motivation to engage challenging, such as those with depression [58].

The high levels of acceptance and usability that users have shown toward VR technology within the study, both in theory and practice, lay a solid ground for investigations into how VR could be used in a mental health context. VR being a premise for real-time data collection via direct observation of behavior contributes to higher ecological validity and allows for the consideration of a wider set of variables when assessing mental states [6,7]. While the potential is promising, reviews of real-life studies [59], editorials [60] as well as commentaries [57], highlight a need for further scientific investigation, to better inform how VR is best implemented within clinical practice.

Cybersickness

Finally, the results of the study show significant differences in self-reported cybersickness levels between HCs and those with DSs —the latter reporting considerably higher levels. This difference brings forward many questions, some relating to the self-report instrument used, as well as its administration; these will be discussed under the limitations more thoroughly. Beyond the critical appraisal of the instrument and optimizing the system so that we minimize the risk of cybersickness in general [61-63], this difference urges a consideration of how cybersickness might relate to depression. It is important to consider that our study lacked a pre-post comparison, which creates challenges when attempting to disentangle cybersickness symptoms evoked by VR exposure from potential baseline differences stemming from somatic symptoms related to depression. While “the language used in the medical literature to describe somatic symptoms in depression is both confusing and contradictory” [64], one can pinpoint constructs related to depression, that may also be captured by cybersickness scales—which may cause contamination when assessing cybersickness levels. These include, but might not be limited to headache, fatigue, nausea, dizziness, and gastrointestinal disturbances [36,64]. This hypothesis needs further testing but is in line with the only study reporting on SSQ in a pre-post manner among those with psychiatric conditions including depression (heterogeneous sample of anxiety, psychotic, depressive, or bipolar disorder)—where the baseline levels of cybersickness would already be classed as “severe,” but there is no significant elevation reported post exposure [65].

Nonetheless, it is important to consider that the elevated cybersickness in the group with DSs could still be, at least in part, a response to the VR environment itself. If this is the case, using VR with the right study design may serve as a potential tool for further exploring the sensory abnormalities in depression, though ethical considerations are warranted. Investigating the underlying mechanisms of cybersickness in depression may also provide novel insights into the somatic manifestations of the condition and could help further define depression as well as refine interventions for mental health. Additionally, the need to consider cybersickness severity becomes even more apparent in light of the results showing its association with acceptability and usability ratings, which might transfer into the real-life uptake and use of VR technology.

Limitations and Future Directions

The study is not without its limitations. First, one has to consider the characteristics of the participants, as the sample consisted of young adults, the majority being women, and highly educated which limits generalizability [66]. In general, lower age and higher education levels are associated with increased digital health literacy [67]. However, further reports are needed to understand how such sample characteristics—sociodemographic variables as well as DSs—influence user attitudes to digital technologies and engagement in a mental health context [68,69]. To this aim, there is a clear need for extending the research to include diverse age groups and educational backgrounds to enhance the generalizability and applicability of the findings.

Second, the categorization between the HC and the DS groups was based solely on self-reported symptoms at the time of the study (using the PHQ-9) as opposed to the presence or absence of a clinical diagnosis. Other forms of assessment might be worthy of consideration. Additionally, the majority of individuals labeled as DS only reported moderate depression severity, which questions whether the results are generalizable to those with more severe symptoms. Self-report measures in general are subject to response biases, such as social desirability, or recall bias. The subjective evaluation of symptom severity in this case could potentially have resulted in over- or underreporting of severity.

Third, although we controlled for demographic covariates, baseline differences in anxiety and stress levels between the study groups may have impacted our results. Self-reported stress and anxiety significantly differed between the HC and DS groups, highlighting that between-group differences in our constructs of interest cannot unequivocally be attributed to differences in DS severity.

Fourth, the measurement of cybersickness levels leaves room for improvement on multiple fronts. The instrument (SSQ) has repeatedly been criticized for reasons among others being the low power to differentiate from anxiety and the questionable assumption of having zero “symptoms” at baseline [37-40]. The difference in severity between the HCs and those with DSs cannot unequivocally be attributed to the VR exposure, as no baseline levels were recorded. It is possible, that at least some degree of the difference originates from baseline differences in anxiety or somatic symptoms of depression, which the questionnaire most likely does not adequately differentiate from cybersickness symptoms. In the future, assessment of related symptoms necessitates pre-post comparisons, particularly relevant when the SSQ is being used [37-40].

Fifth, when creating items to assess constructs of the TFA, 1 item was taken from the SUS questionnaire (item 9: confidence), while all other items were specialized to this study. Overall, single items created for this specific study lack the extensive work that foregoes validated questionnaires.

Sixth, the results of the ordinal regression models should be interpreted with caution, particularly in cases when the assumption of proportional odds was not fulfilled. This suggests that the relationship between predictors and the outcome variable may differ across categories and thus future studies could expand

XSL•FO
RenderX

on these findings by exploring more flexible modeling approaches.

Seventh, we must acknowledge that our study was observational and exploratory in nature, aiming to assess acceptability and usability constructs without preestablished hypotheses, which warrants some considerations. An important consequence of the observational design is the lack of randomization. As a result, there is potential for selection bias and confounding factors that could influence the results. Concerning the exploratory nature: while this approach provides valuable insights, especially at early stages of development, confirmatory research is needed to test specific predictions and establish causal relationships. This is particularly relevant for the potential between-group differences on how VR exposure relates to cybersickness severity.

Finally, regarding future studies, it is important to highlight that clinician perspectives can form an important complement to patients' user feedback. This dual approach provides insights that might not be accessible through the patient perspective alone, such as the feasibility and practicality of integrating VR solutions into clinical workflows [70,71]. As such, future studies could consider involving other user groups, and extend the consideration to implementation, feasibility, and cost-effectiveness of VR-based solutions to enable a more comprehensive evaluation.

In summary, despite the limitations and the need for further research, this study is the first to assess the acceptability, usability, and cybersickness levels of a novel VR system incorporating multimodal data for depression symptom assessment. The findings support the potential of VR as a tool for mental health screening, assessment (when used as a supplementary tool), and treatment. While the system's acceptability was generally favorable, areas for improvement were identified, particularly in enhancing users' self-efficacy during interactions. Usability ratings were satisfactory and cybersickness levels, on average, were comparable to other head-mounted displays. Notably, this study is the first to report elevated cybersickness levels among individuals with DSs, raising important theoretical and methodological questions for future research.

Conclusions

There is a clear need to enhance mental health diagnosis, potentially by incorporating a broader range of variables. This study explored a novel approach to increase the ecological validity of depression assessments and found that, while VR technologies are generally acceptable as a supplementary tool, they are not seen as a replacement for routine mental health evaluations. The study, which tested the initial version of a novel VR system, revealed that the majority of participants found the acceptability and usability satisfactory, despite experiencing considerable levels of cybersickness—which was found to be negatively associated with both constructs. Notably, the results highlighted previously unreported differences in cybersickness severity between individuals with DSs and HCs, warranting replication and further investigation. While the call for enhanced reliability in mental health assessments is well-founded, and novel approaches are under investigation for their efficacy and accuracy, greater emphasis must be placed on evaluating acceptability and usability. Ensuring a safe, acceptable, and user-friendly approach is essential for the successful implementation of these technologies.

Acknowledgments

The study was carried out within the EXPERIENCE project, which is funded by the European Commission H2020 Framework Program (grant 101017727).

Generative AI tools, including ChatGPT (OpenAI) and Perplexity (Perplexity AI, Inc), were used to enhance the clarity of the text, correct grammatical errors, and refine analysis code.

Data Availability

Access to the raw data will be considered on a case-by-case basis upon request. Interested parties may contact the corresponding author.

Authors' Contributions

SS contributed to methodology development, formal analysis, data curation, visualization, and writing the original draft. ETE was responsible for conceptualization, methodology development, formal analysis, data curation, and drafting the original manuscript. FM contributed to methodology development, formal analysis, data curation, and investigation, as well as drafting the original manuscript. VO developed the software, contributed to methodology development, data curation, and visualization, and participated in drafting of the original manuscript. V Catrambone contributed to methodology development and manuscript review and editing. GH was involved in methodology development and participated in reviewing and editing the manuscript. IT contributed to methodology development and manuscript review and editing. ALA participated in manuscript review and editing. V Cardi assisted in methodology development and manuscript review and editing. MGCAC contributed to manuscript review and editing. GM participated in methodology development and manuscript review and editing. MAR contributed to study conceptualization, provided resources, secured funding, supervised the project, and participated in project administration and manuscript review. GV conceptualized the study, secured funding, provided resources, supervised the project, and contributed to project administration and manuscript review. V Carli contributed to conceptualization, methodology, validation, project

XSL•FO
RenderX

administration, supervision, and funding acquisition. V Carli also provided resources and participated in reviewing and editing the manuscript. CG was involved in conceptualization, methodology, validation, supervision, and project administration. CG provided resources, acquired funding, and reviewed and edited the manuscript. All authors reviewed the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Questionnaire on acceptability based on the Theoretical Framework of Acceptability.
[DOCX File , 40 KB - formative_v9i1e68132_app1.docx]

Multimedia Appendix 2
Script used for the analysis and visualization, including its output.
[PDF File (Adobe PDF File), 401 KB - formative_v9i1e68132_app2.pdf]

References

1. GBD 2019 Mental Disorders Collaborators. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet Psychiatry 2022;9(2):137-150 [FREE Full text] [doi: 10.1016/S2215-0366(21)00395-3] [Medline: 35026139]
2. Friedrich MJ. Depression is the leading cause of disability around the world. JAMA 2017;317(15):1517. [doi: 10.1001/jama.2017.3826] [Medline: 28418490]
3. Proudman D, Greenberg P, Nellesen D. The growing burden of major depressive disorders (MDD): implications for researchers and policy makers. Pharmacoeconomics 2021;39(6):619-625 [FREE Full text] [doi: 10.1007/s40273-021-01040-7] [Medline: 34013439]
4. Lieblich SM, Castle DJ, Pantelis C, Hopwood M, Young AH, Everall IP. High heterogeneity and low reliability in the diagnosis of major depression will impair the development of new drugs. BJPsych Open 2015;1(2):e5-e7 [FREE Full text] [doi: 10.1192/bjpo.bp.115.000786] [Medline: 27703745]
5. Liu X, Jiang K. Why is diagnosing MDD challenging? Shanghai Arch Psychiatry 2016;28(6):343-345 [FREE Full text] [doi: 10.11919/j.issn.1002-0829.216073] [Medline: 28638210]
6. Bell IH, Nicholas J, Alvarez-Jimenez M, Thompson A, Valmaggia L. Virtual reality as a clinical tool in mental health research and practice. Dialogues Clin Neurosci 2020;22(2):169-177 [FREE Full text] [doi: 10.31887/DCNS.2020.22.2/lvalmaggia] [Medline: 32699517]
7. Freeman D, Reeve S, Robinson A, Ehlers A, Clark D, Spanlang B, et al. Virtual reality in the assessment, understanding, and treatment of mental health disorders. Psychol Med 2017;47(14):2393-2400 [FREE Full text] [doi: 10.1017/S003329171700040X] [Medline: 28325167]
8. Mao K, Wu Y, Chen J. A systematic review on automated clinical depression diagnosis. Npj Ment Health Res 2023;2(1):20 [FREE Full text] [doi: 10.1038/s44184-023-00040-z] [Medline: 38609509]
9. Bilello JA. Seeking an objective diagnosis of depression. Biomark Med 2016;10(8):861-875. [doi: 10.2217/bmm-2016-0076] [Medline: 27415130]
10. Abd-Alrazaq A, AlSaad R, Shuweihdi F, Ahmed A, Aziz S, Sheikh J. Systematic review and meta-analysis of performance of wearable artificial intelligence in detecting and predicting depression. NPJ Digit Med 2023;6(1):84 [FREE Full text] [doi: 10.1038/s41746-023-00828-5] [Medline: 37147384]
11. Wiebe A, Kannen K, Selaskowski B, Mehren A, Thöne AK, Pramme L, et al. Virtual reality in the diagnostic and therapy for mental disorders: a systematic review. Clin Psychol Rev 2022;98:102213 [FREE Full text] [doi: 10.1016/j.cpr.2022.102213] [Medline: 36356351]
12. Cieślik B, Mazurek J, Rutkowski S, Kiper P, Turolla A, Szczepańska-Gieracha J. Virtual reality in psychiatric disorders: a systematic review of reviews. Complement Ther Med 2020;52:102480 [FREE Full text] [doi: 10.1016/j.ctim.2020.102480] [Medline: 32951730]
13. Duan S, Valmaggia L, Lawrence AJ, Fennema D, Moll J, Zahn R. Virtual reality–assessment of social interactions and prognosis in depression. J Affect Disord 2024;359:234-240 [FREE Full text] [doi: 10.1016/j.jad.2024.05.098] [Medline: 38777276]
14. Camacho-Conde JA, Legarra L, Bolinches VM, Cano P, Guasch M, Llanos-Torres M, et al. Assessment of attentional processes in patients with anxiety-depressive disorders using virtual reality. J Pers Med 2021;11(12):1341 [FREE Full text] [doi: 10.3390/jpm11121341] [Medline: 34945813]
15. Cornwell BR, Salvadore G, Colon-Rosario V, Latov DR, Holroyd T, Carver FW, et al. Abnormal hippocampal functioning and impaired spatial navigation in depressed individuals: evidence from whole-head magnetoencephalography. Am J Psychiatry 2010;167(7):836-844 [FREE Full text] [doi: 10.1176/appi.ajp.2009.09050614] [Medline: 20439387]

XSL•FO
RenderX

16. Ghio L, Gotelli S, Marcenaro M, Amore M, Natta W. Duration of untreated illness and outcomes in unipolar depression: a systematic review and meta-analysis. J Affect Disord 2014;152-154:45-51. [doi: 10.1016/j.jad.2013.10.002] [Medline: 24183486]
17. DiMatteo MR, Lepper HS, Croghan TW. Depression is a risk factor for noncompliance with medical treatment: meta-analysis of the effects of anxiety and depression on patient adherence. Arch Intern Med 2000;160(14):2101-2107. [doi: 10.1001/archinte.160.14.2101] [Medline: 10904452]
18. Sekhon M, Cartwright M, Francis JJ. Acceptability of healthcare interventions: an overview of reviews and development of a theoretical framework. BMC Health Serv Res 2017;17(1):88 [FREE Full text] [doi: 10.1186/s12913-017-2031-8] [Medline: 28126032]
19. Sauer J, Sonderegger A, Schmutz S. Usability, user experience and accessibility: towards an integrative model. Ergonomics 2020;63(10):1207-1220. [doi: 10.1080/00140139.2020.1774080] [Medline: 32450782]
20. Kim YY, Kim HJ, Kim EN, Ko HD, Kim HT. Characteristic changes in the physiological components of cybersickness. Psychophysiology 2005;42(5):616-625. [doi: 10.1111/j.1469-8986.2005.00349.x] [Medline: 16176385]
21. Migoya-Borja M, Delgado-Gómez D, Carmona-Camacho R, Porras-Segovia A, López-Moriñigo JD, Sánchez-Alonso M, et al. Feasibility of a virtual reality–based psychoeducational tool (VRight) for depressive patients. Cyberpsychol Behav Soc Netw 2020;23(4):246-252. [doi: 10.1089/cyber.2019.0497] [Medline: 32207997]
22. Paul M, Bullock K, Bailenson J, Burns D. Examining the efficacy of extended reality-enhanced behavioral activation for adults with major depressive disorder: randomized controlled trial. JMIR Ment Health 2024;11:e52326 [FREE Full text] [doi: 10.2196/52326] [Medline: 38437873]
23. Saredakis D, Szpak A, Birckhead B, Keage HAD, Rizzo A, Loetscher T. Factors associated with virtual reality sickness in head-mounted displays: a systematic review and meta-analysis. Front Hum Neurosci 2020;14:96 [FREE Full text] [doi: 10.3389/fnhum.2020.00096] [Medline: 32300295]
24. Lundin RM, Yeap Y, Menkes DB. Adverse effects of virtual and augmented reality interventions in psychiatry: systematic review. JMIR Ment Health 2023;10:e43240 [FREE Full text] [doi: 10.2196/43240] [Medline: 37145841]
25. Mura F, Alfeo AL, Sutori S, Eliasson ET, Catrambone V, Ortiz VG, et al. The EXPERIENCE system: integrating virtual reality and multi-modal explainable artificial intelligence for enhanced depression recognition. NA 2025 (forthcoming).
26. Kroenke K, Spitzer RL, Williams JBW. The PHQ-9: Validity of a brief depression severity measure. J Gen Intern Med 2001;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
27. Prokhorenkova L, Gusev G, Vorobev A, Dorogush A, Gulin A. CatBoost: Unbiased boosting with categorical features. ArXiv. Preprint posted online on June 28, 2017 2019. [doi: 10.48550/arXiv.1706.09516]
28. Alfeo AL, Zippo AG, Catrambone V, Cimino MG, Toschi N, Valenza G. From local counterfactuals to global feature importance: efficient, robust, and model-agnostic explanations for brain connectivity networks. Comput Methods Programs Biomed 2023;236:107550 [FREE Full text] [doi: 10.1016/j.cmpb.2023.107550] [Medline: 37086584]
29. Lovibond P, Lovibond SH. The structure of negative emotional states: comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories. Behav Res Ther 1995;33(3):335-343. [doi: 10.1016/0005-7967(94)00075-u] [Medline: 7726811]
30. Watson D, Clark LA, Tellegen A. Development and validation of brief measures of positive and negative affect: the PANAS scales. J Pers Soc Psychol 1988;54(6):1063-1070. [doi: 10.1037//0022-3514.54.6.1063] [Medline: 3397865]
31. Kashdan TB, Gallagher MW, Silvia PJ, Winterstein BP, Breen WE, Terhar D, et al. The Curiosity and Exploration Inventory-II: Development, factor structure, and psychometrics. J Res Pers 2009;43(6):987-998 [FREE Full text] [doi: 10.1016/j.jrp.2009.04.011] [Medline: 20160913]
32. Sekhon M, Cartwright M, Francis JJ. Development of a theory-informed questionnaire to assess the acceptability of healthcare interventions. BMC Health Serv Res 2022;22(1):279 [FREE Full text] [doi: 10.1186/s12913-022-07577-3] [Medline: 35232455]
33. Brooke J. SUS: A quickdirty usability scale. In: Jordan PW, Thomas B, Weerdmeester BA, McClelland IL, editors. Usability Evaluation in Industry. London, UK: Taylor & Francis; 1996:189-194.
34. Kortum P, Peres SC. Evaluation of home health care devices: remote usability assessment. JMIR Hum Factors 2015;2(1):e10 [FREE Full text] [doi: 10.2196/humanfactors.4570] [Medline: 27025664]
35. Bangor A, Kortum P, Miller J. Determining what individual SUS scores mean: Adding an adjective rating scale. J Usability Stud 2009;4(3):114-123. [doi: 10.5555/2835587.2835589]
36. Kennedy RS, Lane NE, Berbaum KS, Lilienthal MG. Simulator Sickness Questionnaire: An enhanced method for quantifying simulator sickness. Int J Aerosp Psychol 1993;3(3):203-220. [doi: 10.1207/s15327108ijap0303_3]
37. Bimberg P, Weissker T, Kulik A. On the usage of the simulator sickness questionnaire for virtual reality research. 2020 Presented at: IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW); March 22-26, 2020; Atlanta, GA p. 464-467.
38. Bouchard S, Berthiaume M, Robillard G, Forget H, Daudelin-Peltier C, Renaud P, et al. Arguing in favor of revising the factor structure when assessing side effects induced by immersions in virtual reality. Front Psychiatry 2021;12:739742 [FREE Full text] [doi: 10.3389/fpsyt.2021.739742] [Medline: 34803762]

XSL•FO
RenderX

39. Brown P, Spronck P, Powell W. The Simulator Sickness Questionnaire, and the erroneous zero baseline assumption. Front Virtual Real 2022;3:945800. [doi: 10.3389/frvir.2022.945800]
40. Sevinc V, Berkman MI. Psychometric evaluation of simulator sickness questionnaire and its variants as a measure of cybersickness in consumer virtual environments. Appl Ergon 2020;82:102958. [doi: 10.1016/j.apergo.2019.102958] [Medline: 31563798]
41. Wickham H, François R, Henry L, Müller K, Vaughan D. Dplyr: A grammar of data manipulation. Tidyverse. 2023. URL: https://dplyr.tidyverse.org [accessed 2025-03-21]
42. Adler D, Kelly S, Elliott T, Adamson J. Vioplot: Violin plot. GitHub. 2024. URL: https://github.com/TomKellyGenetics/vioplot [accessed 2025-03-21]
43. Wickham H. Ggplot2: Elegant graphics for data analysis. Tidyverse. 2016. URL: https://ggplot2.tidyverse.org [accessed 2025-03-21]
44. Revelle W. Psych: Procedures for psychological, psychometric, and personality research. The Comprehensive R Archive Network. 2024. URL: https://CRAN.R-project.org/package=psych [accessed 2025-03-21]
45. Kassambara A. Rstatix: Pipe-friendly framework for basic statistical tests. Datanovia. 2023. URL: https://rpkgs.datanovia.com/rstatix/ [accessed 2025-03-21]
46. Lüdecke D. ggeffects: Tidy data frames of marginal effects from regression models. J Open Res Softw 2018;3(26):772 [FREE Full text]
47. Christensen RHB. Ordinal: Regression models for ordinal data. The Comprehensive R Archive Network. 2023. URL: https://CRAN.R-project.org/package=ordinal [accessed 2025-03-21]
48. Yee TW. The VGAM package for categorical data analysis. J Stat Soft 2010;32(10):1-34. [doi: 10.18637/jss.v032.i10]
49. Carli V, Gentili C. EXPERIENCE: Is virtual reality suitable to identify differences between depressed and healthy individuals? International Standard Randomised Controlled Trial Number Registry. 2022. URL: https://doi.org/10.1186/isrctn16396369 [accessed 2025-03-21]
50. Chan AHY, Honey MLL. User perceptions of mobile digital apps for mental health: acceptability and usability—an integrative review. J Psychiatr Ment Health Nurs 2022;29(1):147-168. [doi: 10.1111/jpm.12744] [Medline: 33604946]
51. Shahid S, Kelson J, Saliba A. Effectiveness and user experience of virtual reality for social anxiety disorder: systematic review. JMIR Ment Health 2024;11:e48916 [FREE Full text] [doi: 10.2196/48916] [Medline: 38329804]
52. Rus-Calafell M, Garety P, Sason E, Craig TJK, Valmaggia LR. Virtual reality in the assessment and treatment of psychosis: a systematic review of its utility, acceptability and effectiveness. Psychol Med 2018;48(3):362-391. [doi: 10.1017/S0033291717001945] [Medline: 28735593]
53. Jones C, Miguel Cruz A, Smith-MacDonald L, Brown MRG, Vermetten E, Brémault-Phillips S. Technology acceptance and usability of a virtual reality intervention for military members and veterans with posttraumatic stress disorder: mixed methods unified theory of acceptance and use of technology study. JMIR Form Res 2022;6(4):e33681 [FREE Full text] [doi: 10.2196/33681] [Medline: 35451971]
54. Weech S, Kenny S, Barnett-Cowan M. Presence and cybersickness in virtual reality are negatively related: a review. Front Psychol 2019;10:158 [FREE Full text] [doi: 10.3389/fpsyg.2019.00158] [Medline: 30778320]
55. Voinescu A, Petrini K, Stanton Fraser D. Presence and simulator sickness predict the usability of a virtual reality attention task. Virtual Real 2023;27(3):1-17 [FREE Full text] [doi: 10.1007/s10055-023-00782-3] [Medline: 37360806]
56. Migliore S, Casella M, Tramontano C, Curcio G, Squitieri F. Virtual reality tolerability, sense of presence and usability in Huntington disease: a pilot study. Neurol Sci 2025;46(1):219-225. [doi: 10.1007/s10072-024-07726-y] [Medline: 39103734]
57. Selaskowski B, Wiebe A, Kannen K, Asché L, Pakos J, Philipsen A, et al. Clinical adoption of virtual reality in mental health is challenged by lack of high-quality research. Npj Ment Health Res 2024;3(1):24 [FREE Full text] [doi: 10.1038/s44184-024-00069-8] [Medline: 38755289]
58. Torous J, Nicholas J, Larsen ME, Firth J, Christensen H. Clinical review of user engagement with mental health smartphone apps: evidence, theory and improvements. Evid Based Ment Health 2018;21(3):116-119 [FREE Full text] [doi: 10.1136/eb-2018-102891] [Medline: 29871870]
59. Geraets CNW, Wallinius M, Sygel K. Use of virtual reality in psychiatric diagnostic assessments: a systematic review. Front Psychiatry 2022;13:828410 [FREE Full text] [doi: 10.3389/fpsyt.2022.828410] [Medline: 35295778]
60. Riva G, Serino S. Virtual reality in the assessment, understanding and treatment of mental health disorders. J Clin Med 2020;9(11):3434 [FREE Full text] [doi: 10.3390/jcm9113434] [Medline: 33114623]
61. Biswas N, Mukherjee A, Bhattacharya S. “Are you feeling sick?”—A systematic literature review of cybersickness in virtual reality. ACM Comput Surv 2024;56(11):1-38. [doi: 10.1145/3670008]
62. Page D, Lindeman RW, Lukosch S. Identifying strategies to mitigate cybersickness in virtual reality induced by flying with an interactive travel interface. Multimodal Technol Interact 2023;7(5):47. [doi: 10.3390/mti7050047]
63. Papaefthymiou S, Giannakopoulos A, Roussos P, Kourtesis P. Mitigating cybersickness in virtual reality: impact of eye–hand coordination tasks, immersion, and gaming skills. Virtual Worlds 2024;3(4):506-535. [doi: 10.3390/virtualworlds3040027]
64. Tylee A, Gandhi P. The importance of somatic symptoms in depression in primary care. Prim Care Companion J Clin Psychiatry 2005;7(4):167-176 [FREE Full text] [doi: 10.4088/pcc.v07n0405] [Medline: 16163400]

XSL•FO
RenderX

65. Veling W, Lestestuiver B, Jongma M, Hoenders HJR, van Driel C. Virtual reality relaxation for patients with a psychiatric disorder: crossover randomized controlled trial. J Med Internet Res 2021;23(1):e17233 [FREE Full text] [doi: 10.2196/17233] [Medline: 33448933]
66. Borghouts J, Eikey E, Mark G, De Leon C, Schueller SM, Schneider M, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. J Med Internet Res 2021;23(3):e24387 [FREE Full text] [doi: 10.2196/24387] [Medline: 33759801]
67. Estrela M, Semedo G, Roque F, Ferreira PL, Herdeiro MT. Sociodemographic determinants of digital health literacy: a systematic review and meta-analysis. Int J Med Inform 2023;177:105124 [FREE Full text] [doi: 10.1016/j.ijmedinf.2023.105124] [Medline: 37329766]
68. Chudy-Onwugaje K, Abutaleb A, Buchwald A, Langenberg P, Regueiro M, Schwartz DA, et al. Age modifies the association between depressive symptoms and adherence to self-testing with telemedicine in patients with inflammatory bowel disease. Inflamm Bowel Dis 2018;24(12):2648-2654 [FREE Full text] [doi: 10.1093/ibd/izy194] [Medline: 29846623]
69. Abouzeid N, Lal S. The role of sociodemographic factors on the acceptability of digital mental health care: a scoping review protocol. PLoS One 2024;19(4):e0301886 [FREE Full text] [doi: 10.1371/journal.pone.0301886] [Medline: 38669278]
70. Chung OS, Robinson T, Johnson AM, Dowling NL, Ng CH, Yücel M, et al. Implementation of therapeutic virtual reality into psychiatric care: clinicians' and service managers' perspectives. Front Psychiatry 2021;12:791123 [FREE Full text] [doi: 10.3389/fpsyt.2021.791123] [Medline: 35058823]
71. Chung OS, Robinson T, Johnson AM, Dowling NL, Ng CH, Yücel M, et al. Corrigendum: Implementation of therapeutic virtual reality into psychiatric care: Clinicians' and service managers' perspectives. Front Psychiatry 2022;13:893637 [FREE Full text] [doi: 10.3389/fpsyt.2022.893637] [Medline: 35815031]

Abbreviations

DASS: Depression Anxiety Stress Scale
DS: depressive symptom
DSM: Diagnostic and Statistical Manual of Mental Disorders
HC: healthy control
PANAS-SF: Positive and Negative Affect Schedule
PHQ-9: Patient Health Questionnaire-9
SSQ: Simulator Sickness Questionnaire
SUS: System Usability Scale
TFA: theoretical framework of acceptability
VR: virtual reality

Edited by A Mavragani; submitted 08.11.24; peer-reviewed by R Zahn, AN Ramaseri-Chandra, C Nega; comments to author 03.01.25; revised version received 21.01.25; accepted 11.03.25; published 16.04.25.

Please cite as:
Sutori S, Eliasson ET, Mura F, Ortiz V, Catrambonephd V, Hadlaczky G, Todorov I, Alfeo AL, Cardi V, Cimino MGCA, Mioni G, Raya MA, Valenza G, Carli V, Gentili C
Acceptability, Usability, and Insights Into Cybersickness Levels of a Novel Virtual Reality Environment for the Evaluation of Depressive Symptoms: Exploratory Observational Study
JMIR Form Res 2025;9:e68132
URL: https://formative.jmir.org/2025/1/e68132
doi:10.2196/68132
PMID:

©Sara Sutori, Emma Therése Eliasson, Francesca Mura, Victor Ortiz, Vincenzo Catrambonephd, Gergö Hadlaczky, Ivo Todorov, Antonio Luca Alfeo, Valentina Cardi, Mario G C A Cimino, Giovanna Mioni, Mariano Alcañiz Raya, Gaetano Valenza, Vladimir Carli, Claudio Gentili. Originally published in JMIR Formative Research (https://formative.jmir.org), 16.04.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Digital Photo Activity Intervention for Nursing Home Residents With Dementia and Their Carers: Mixed Methods Process Evaluation

Josephine Rose Orejana Tan[1], PGDip, MSc; David P Neal[2], MB BChir, MA, PhD; Maria Vilmen[1], BSc; Petra Boersma[3,4], PhD; Teake P Ettema[1], PhD; Robbert J J Gobbens[3,5,6,7], RN, MSc, PhD; Sietske A M Sikkes[8,9,10], PhD; Rose-Marie Dröes[1], PhD

[1]Department of Psychiatry, Amsterdam Public Health Research Institute, Amsterdam UMC, location Vrije Universiteit, Amsterdam, The Netherlands

[2]eHealth Living & Learning Lab, Department of Medical Informatics, Amsterdam UMC, Amsterdam, The Netherlands

[3]Lectorate Health and Well-being of Frail Elderly People, Faculty of Health, Sports and Social Work, Inholland University of Applied Sciences, Amsterdam, The Netherlands

[4]Program Learning, Improvement & Implementation, Ben Sajet Centrum, Amsterdam, The Netherlands

[5]Zonnehuisgroep Amstelland, Amstelveen, The Netherlands

[6]Department of Family Medicine and Population Health, Faculty of Medicine and Health Sciences, University of Antwerp, Antwerp, Belgium

[7]Tranzo, Tilburg University, Tilburg, The Netherlands

[8]Department of Clinical, Neuro- & Developmental Psychology, Faculty of Behavioural and Movement Sciences, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

[9]Neurology, Alzheimer Center Amsterdam, Vrije Universiteit Amsterdam, Amsterdam UMC, Amsterdam, The Netherlands

[10]Amsterdam Neuroscience, Neurodegeneration, Amsterdam, The Netherlands

Corresponding Author:
Josephine Rose Orejana Tan, PGDip, MSc
Department of Psychiatry
Amsterdam Public Health Research Institute
Amsterdam UMC, location Vrije Universiteit
Oldenaller 1
Amsterdam, 1081 HJ
The Netherlands
Phone: 31 3316459170
Email: j.r.tan@amsterdamumc.nl

Abstract

Background: Within the framework of a randomized controlled trial investigating the impact of a digital, psychosocial photo activity intervention for residents living with dementia in nursing homes and their informal and formal carers, a process evaluation was conducted to determine factors that affected the implementation of the intervention and potentially influenced the intervention outcomes.

Objective: By tracing facilitators and barriers to implementation, the study also aimed to inform future implementation of the photo activity intervention.

Methods: Following Medical Research Council guidance, mixed methods were used to investigate context, implementation, and mechanism-of-impact factors during the photo activity intervention via the Fotoscope web application versus a general conversation activity (control). Google Analytics was set up to gain insight into how the Fotoscope web application was used in practice. For quantitative data, descriptive statistics were calculated and differences between groups tested. For qualitative data, thematic analysis was performed.

Results: In total, 163 semistructured interviews were conducted with residents (photo activity group: n=29, 17.8%; control: n=29, 17.8%), formal carers (photo activity group: n=23, 14.1%; control: n=27, 16.6%), and informal carers (photo activity group: n=28, 17.2%; control: n=27, 16.6%). Regarding contextual factors, a minority of formal carers in both groups (photo activity group: 4/18, 22%; control: 9/24, 38%) mentioned time and workload as barriers to implementing the intervention. Regarding implementation, 86% (25/29) of the residents in the intervention group felt that the digital photo activity worked well on a tablet.

XSL•FO
RenderX

Informal carers from both groups wanted more intervention updates from formal carers. The majority of formal carers from both groups were satisfied with how the training and activities were implemented. Regarding the mechanisms of impact, residents in the photo activity group (27/29, 93%) felt significantly more positive about the conversations with their carer (U=533.0, z=2.865, r=0.39; P=.004). Formal carers in the photo activity group (20/23, 87%) got to know the resident better (U=390.5, z=2.114, r=0.302; P=.04) compared to the formal carers in the control group (21/27, 78%). Formal carers in the photo activity group (23/50, 46%) gave a significantly higher rating to the digital photo activity as a way of getting to know the resident living with dementia better (median 9.00, IQR 7-9; U=419.0, z=2.169, r=0.307; P=.03) compared to formal carers in the control group (27/50, 54%; median 8.00, IQR 6-8). Finally, the majority of formal carers in the photo activity group (14/18, 78%) agreed that the Fotoscope app can be used as part of care activities in the nursing home.

Conclusions: The work invested by formal carers in implementing the photo activity did not seem to differ greatly compared to implementing a general conversation activity, suggesting that the digital photo activity, as an easy-to-implement and enjoyable intervention, could be widely implemented and disseminated in nursing homes.

International Registered Report Identifier (IRRID): RR2-https://doi.org/10.1186/s12877-021-02632-w

(JMIR Form Res 2025;9:e56586) doi:10.2196/56586

KEYWORDS

dementia; psychosocial interventions; nursing home; process evaluation; social interaction; photos; art

Introduction

Background

The photo activity is a psychosocial intervention based on an earlier project by a visual artist (known as *Photographic Treatment* [1]) for nursing home residents with dementia and their informal and formal carers. It is designed to facilitate social interaction [2]. In a small-scale pilot study (n=20), nursing home residents viewed printed, artistic, generic photos related to their personal interests [2]. The study found that when residents with dementia viewed person-centered generic photos (ie, not personal or family photos but images of subjects related to personal interests) as opposed to non–person-centered photos during conversations with their informal or formal carers, they tended to exhibit more positive effects on social interaction, mood, speech, and negative behavior [2]. In recent decades, psychosocial interventions have been influenced in part by the many technological developments aimed at improving the social health and quality of life of people with dementia [3-6]. Examples include the use of social robots [7], exergaming [8], and online art gallery activities [9]. However, few studies have looked into the effects of psychosocial interventions using technology for nursing home residents with dementia [10-12]. The photo activity is an example of a psychosocial intervention offered in nursing homes that was recently digitally adapted through the development of the Fotoscope app [13].

The development of the Fotoscope app allows for upscaling and wider dissemination of the digital photo activity intervention (ie, photo activity using digital photos, rather than printed photos). Using the app, informal and formal carers can easily conduct the digital photo activity with residents by accessing a database of artistic, black-and-white, generic photos and making a person-centered selection of digital photos to view with the residents [13], as opposed to having a limited collection of printed photos. However, before disseminating the digital photo activity intervention, it is important to evaluate its effectiveness and to consider facilitators and barriers to implementing it in daily nursing home care.

It is argued that alongside evaluating an intervention's effects, process evaluations are crucial to look at how and under what conditions an intervention was implemented, to make the outcomes reproducible, and to learn about factors that facilitated or impeded the implementation of the intervention, all of which will inform the interpretation of trial outcomes and future upscaling [14-16]. The importance of conducting process evaluations alongside randomized controlled trials (RCTs) was emphasized in the revised UK Medical Research Council (MRC) guidance for process evaluations of complex interventions [17]. In addition to tracing facilitators and barriers to implementation, process evaluations allow for the identification of contextual factors that may have influenced the implementation of the intervention and its outcomes, as well as the mechanisms underlying the intervention's impact [16,17]. *Implementation* refers to the ways and processes in which the intervention was delivered in practice, how frequently it was delivered, and whether it was delivered as intended. *Mechanisms of impact* refers to the aspects of the intervention that created change in practice or in the target population, investigating which would help in understanding future replications of the intervention. Finally, *context* refers to factors outside of the intervention that may facilitate or hinder the implementation [17].

As technology for dementia care continues to advance, along with RCTs evaluating the effectiveness of care technology [3,10], it is becoming more common to include process evaluations alongside such trials; for example, process evaluations based on the MRC guidance have been conducted in studies evaluating the use of different kinds of technology in psychosocial care for people with dementia in the community or long-term care settings [18-21].

Objectives

This paper reports on a process evaluation based on the MRC guidance [17], which was carried out alongside a pilot RCT (Netherlands Trial Register: NL9219 [22]) investigating the feasibility and impact of the digital photo activity intervention. The implementation of the photo activity intervention in comparison with the implementation of a control intervention

XSL•FO
RenderX

(a general conversation activity without photos) was evaluated. The purpose of this study was to identify factors that either facilitated or hindered the implementation of the photo activity intervention compared to a general conversation activity, including contextual, implementation, and mechanism-of-impact factors, to generate insights to interpret RCT outcomes and to inform further implementation and dissemination of the intervention.

The study aimed to answer the following research questions (RQs):

- RQ1: How did factors external to the interventions affect the implementation of the interventions (context)?
- RQ2: How did residents and their informal and formal carers assess the preparation, introduction, and delivery of the interventions? (implementation)?
- RQ3: How did the residents with dementia and their informal and formal carers experience and value the digital photo activity versus the general conversation activity (mechanisms of impact)?

Methods

Design of the Process Evaluation

Following MRC guidance [17] (Figure 1 [23]), we used a mixed methods approach to investigate contextual, implementation, and mechanism-of-impact factors because the combination of quantitative and qualitative data would allow us to gain a more comprehensive understanding of the way the photo activity, as a complex intervention, is organized, offered, and implemented in the clinical setting [17,24]. Using mixed methods can also provide the flexibility needed in working with nursing home residents with dementia, ensuring that they are able to contribute their insights regarding the intervention [24,25]. The reporting and writing of this study was guided by the Good Reporting of a Mixed Methods Study (GRAMMS) checklist [26].

During the 4-week intervention period, analytics data were collected on Fotoscope app use and interactions of carer-resident pairs assigned to the photo activity intervention. After the intervention, qualitative and quantitative data were collected through semistructured interviews. The full protocol for the RCT, including the process evaluation, has been published elsewhere [13].

Figure 1. Contextual, implementation, and mechanism-of-impact factors, based on the Medical Research Council guidance for process evaluation, that could facilitate or hinder the implementation of the digital photo activity in the nursing home.

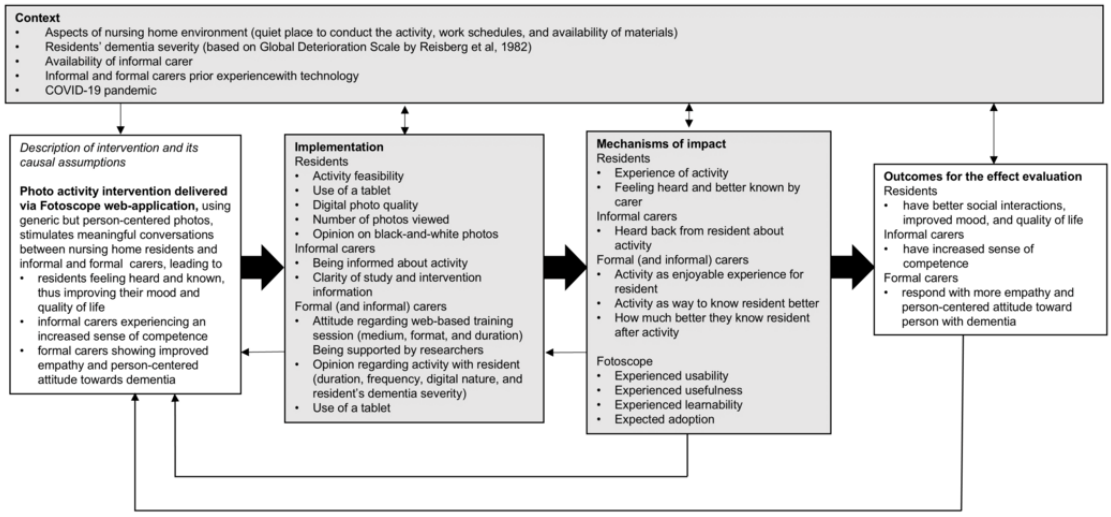

Ethical Considerations

This research was approved by the Vrije University Medical Center's Ethics Review Committee (2020.221) and embedded within the Amsterdam Public Health Research Institute (SQC2020-039). All participants provided informed consent. For individuals with dementia who lacked the capacity to consent, their legal representative signed the consent form on their behalf.

For privacy and confidentiality, data are anonymized, coded, and stored securely. Financial compensation was given to participating nursing home wards for their time (€100 [US $108.7] per participating resident with dementia who completed an intervention round).

Participants

The study aimed to recruit nursing home residents in the Netherlands with dementia at severity levels 4, 5, or 6 on the Global Deterioration Scale (GDS) [23], along with their informal and formal carers. To do this, the researchers worked with interested regional and national nursing home organizations. The study was presented to individual nursing home wards, where ward managers, in consultation with their teams, determined their willingness and capacity to participate. In total, 20 nursing home wards from 3 large care organizations took part in the study. They were asked to recruit formal carers to fulfil the following roles: coordinator (served as the first point of contact, recruited informal carers and residents with dementia for the study by providing written and oral information about the study, and assessed the dementia severity of participating

XSL•FO
RenderX

residents), intervention providers (2 formal carers randomized to deliver either the photo activity or control activity), and independent assessor (a formal carer not involved in delivering either activity, who conducted blind assessments of residents' daily functioning in the ward for the pilot RCT by filling out web-based questionnaires). Informal carers of residents at the participating care organizations were also allowed to provide either the photo activity or control activity if they wished. Researchers contacted residents and informal carers about their participation in the study only after the coordinator obtained their informed consent. The photo activity and control activity were provided by formal carers (or informal carers if they wished to provide the interventions themselves) to the residents twice a week for 4 weeks. Residents as well as informal and formal carers who provided either activity and completed the 4-week intervention period were then recruited for the process evaluation by the researchers and were invited to complete the semistructured interviews via online video calls or telephone calls. The formal carers who participated as independent assessors were only involved in the pilot RCT and not in the process evaluation.

Pairs of residents with the same dementia severity who had lived in the nursing home for at least 1 month were recruited by the coordinator. After providing informed consent, they were randomly assigned to either the experimental or control condition through a lot-drawing process. The only exclusion criteria were severe vision or hearing problems. Informal carers were assigned to the same condition as the resident.

The sample size for the RCT was determined based on the results of the photo activity pilot study. Large positive effect sizes were found for the INTERACT subscales social interaction (Cohen d=0.86) and reduced negative behavior (Cohen d=0.85) [2]. On the basis of these effect sizes and an expected dropout rate of 10% over 6 weeks, a power calculation (power=0.80; α=.05; Cohen d=0.80) indicated that 45 residents per group (experimental and control) were needed. All residents and their carers involved in the RCT were invited to participate in the process evaluation.

Interventions

Photo Activity (Experimental Intervention) and the Fotoscope Web Application

The photo activity intervention engaged residents and their carers in a 1-on-1 conversation about artistic, person-centered generic photos in the Fotoscope web application. For the digital photo activity, the Fotoscope database was expanded to 1500 black-and-white photos, which were organized according to 7 main themes (people, places, nature, animals, activities, things, and experiences). The Fotoscope web application features multiple sections, such as Themes, Profile, Favorites, User Guide, and Information (Figure 2). The Themes page displays the 7 main photo categories and subcategories. The Profile page contains questions designed to learn more about the residents' personal interests and features a selection tool that allows carers to easily select photos that match the residents' interests and save them in the Photo Selection tab. The Favorites page displays photos that elicit positive reactions from the resident, which the carer can add during or after their conversation. The User Guide provides instructions on using the Fotoscope web application and offers tips on person-centered communication. The Information page contains details about copyright and the researchers' contact information. Additional features include a search function (carers can enter keywords to find a photo) and a "joker" function (a collection of photos that have been found to elicit positive reactions).

XSL•FO
RenderX

Figure 2. Screenshots of the Fotoscope web application's Themes, Profile, and Photo Selection pages. Photos in the Fotoscope web-application were publicly available on the internet and collected by the artist, or were taken by the artist herself. No personal photos of participants were used.

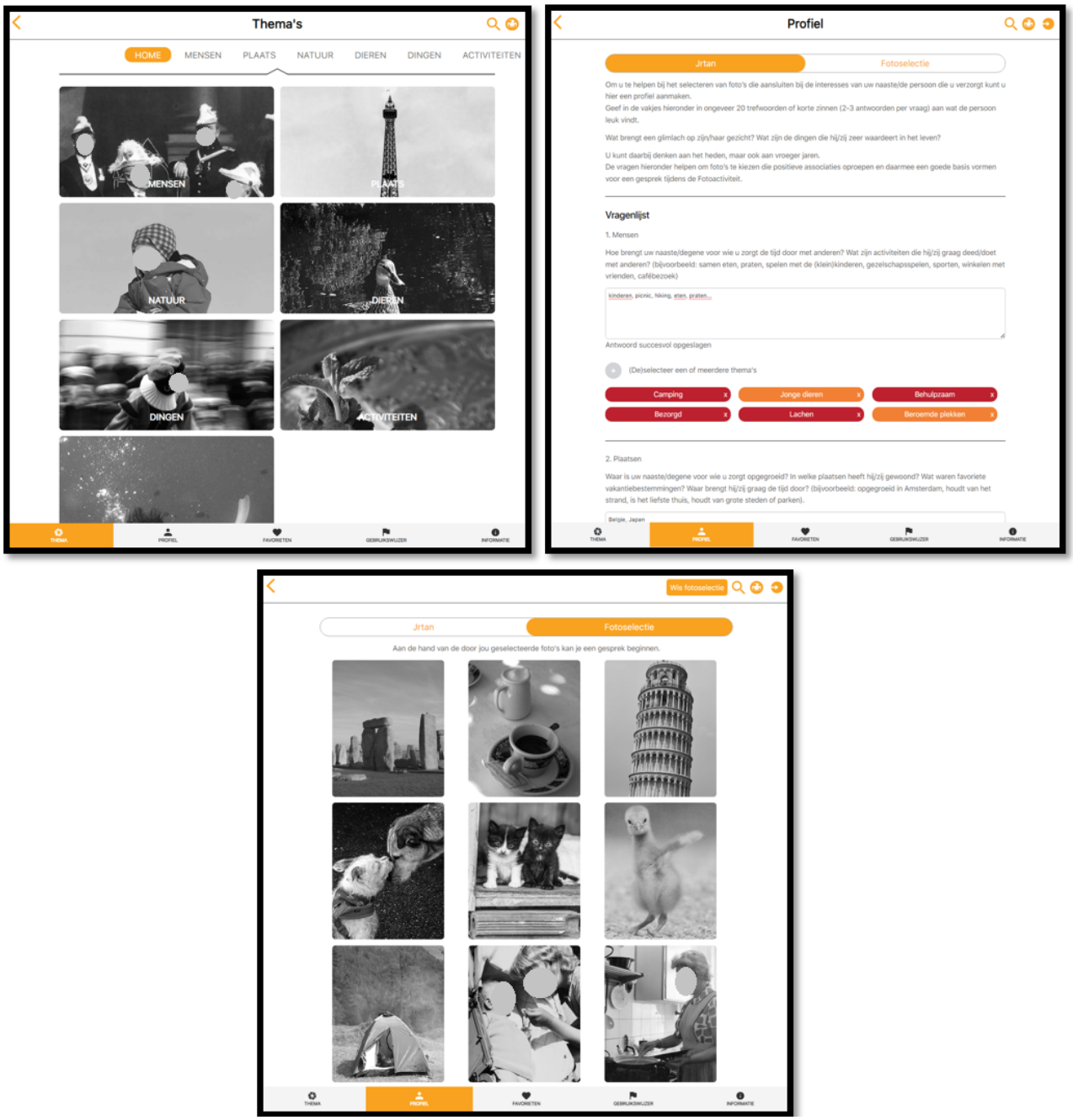

General Conversation Activity (Control Intervention)

The control intervention also involved residents and their carers in a 1-on-1 conversation but without the Fotoscope web application and generic photos. The carer was encouraged to have an open conversation with the resident based on general topics and to allow the resident to lead the conversation.

Data Collection Methods

Data for the process evaluation were collected via a demographic questionnaire, semistructured interviews, Google Analytics, and a carer self-report registration form (Multimedia Appendix 1).

Demographic Questionnaire

Questions for the demographic questionnaire administered at baseline (T0) were adapted from The Older Persons and Informal Caregivers Survey Minimum Dataset [27], with additional questions pertaining to relevant medical information for the residents with dementia. For residents, information on sex, age, education, years in the nursing home, type of dementia, and severity of dementia was collected. For informal carers, information on sex, age, education, relationship to residents, and employment status (whether they had paid work) was collected. For formal carers, information on sex, age, education, work function, and years of experience in psychogeriatrics was collected.

XSL•FO
RenderX

Semistructured Interview

The questions for the semistructured interview, which was conducted after the 4-week intervention period (T1), were adapted from existing interview guides used in a previous process evaluation study on a digital psychosocial intervention for people with dementia and their caregivers and pilot-tested [18]. The questions were adapted by JROT, RMD, and PB to fit the photo activity trial. On the basis of the MRC guidance for process evaluation [17], context, implementation, and mechanism-of-impact factors (including usability, usefulness, learnability, and adoption of the Fotoscope app) were explored through a mix of open-ended and closed-ended questions (Multimedia Appendix 2). An example of a closed-ended question for a resident with dementia who participated in the photo activity intervention is "Do you think your carer got to know you better by having a regular conversation with you about the photos together?"; the corresponding closed-ended question for a resident with dementia in the control condition was "Do you think your carer got to know you better by having regular conversations with you on general topics?" The answer options for these questions were "A lot better," "Better," "A little better," and "No, not better."

Google Analytics

The Fotoscope web application was linked to a Google Analytics account to measure the carers' use of, and interactions with, the app during the trial. To protect residents' privacy, IP addresses were masked, and codes rather than names were used to log in. To gain insight into carers' engagement with the Fotoscope web application (*implementation*), data on the number of sessions per user and average session durations were collected. In addition, dates of each session, as well as session start times, were recorded for each user.

To determine whether carers used the Fotoscope web application to get to know residents better (*mechanisms of impact*), unique page views of the Profile and Favorites pages were tracked in Google Analytics. To determine whether carers found the pages and functions of the Fotoscope web application useful (*usefulness*), the number of unique views of these pages was recorded as an indication. To determine whether carers needed additional guidance on using the Fotoscope web application, the number of unique page views of the User Guide and Information page was recorded (*learnability*). Finally, information on device type was collected (*context*) to know whether the Fotoscope web application was used by carers on a tablet, as advised.

Carer Self-Report Registration Form

Carers were asked to record the dates, times, and duration of the activities with the residents on a standard record form to track the time spent on the activities and the number of sessions conducted (*implementation*). If a session was missed, the intervention discontinued, or the resident dropped out, they were asked to document the reason on the form.

Procedure

The pilot RCT in February 2021 was disrupted by the COVID-19 pandemic. In the Netherlands, nursing homes were placed under quarantine, restricting visitor access [28]. The trial protocol was adapted to an online format. Carers assigned to the photo activity attended a 1.5-hour web-based training session where they were introduced to the photo activity intervention and procedure [2] and were shown an online demonstration of the Fotoscope web application. General communication tips were discussed, followed by more specific person-oriented communication skills for interacting with people with dementia [29]. Carers were then given personal log-in credentials and passwords for the Fotoscope web application—1 set for their own use to enable them to become familiar with the app and another for each resident participating in the photo activity. Each resident's Fotoscope user account was linked to a unique user ID in Google Analytics, and only the researchers had access to the key linking the IDs to specific residents. Carers in the photo activity group were instructed to prepare for the first session by completing the Fotoscope Profile page with input from family members and selecting a minimum of 15 photos based on the residents' interests. Formal carers in the photo activity group were instructed to update the informal carers (by telephone or when they visited the resident) on how the residents responded to the activity, at least once during the 4-week intervention.

The carers assigned to the control activity attended a 1-hour online training session where the researchers explained the procedure for the general conversation activity and provided general communication tips.

In both conditions, carers were advised to conduct the activities with the resident twice a week for 4 weeks, with each session lasting 30 minutes. They were also advised to invite the resident to a quiet room in the ward where they could conduct the activity without disruption. Carers assessed residents' feelings before and after each session using the Smiley Face Assessment Scale (SFAS) [30]. Trained female and male research assistants observed the first and last sessions via video calls, assessing social interaction between the resident and carer during the activity using the INTERACT observation scale [13]. After these sessions, carers in the photo activity group received feedback from the research assistants on how to improve person-centered communication based on their observations. At the end of the final session, the residents with dementia were interviewed during the video call for 5 to 10 minutes by the research assistant, using questions from the semistructured interview guide. Both informal and formal carers either completed the digital form or participated in a 20-minute interview via telephone or video call at a time convenient to them within a week after the final intervention session. Carers who were unable to attend an interview were asked to submit a written narrative account of their experience in response to the questions from the interview guide. Formal carers who conducted either the photo activity or control activity with >1 resident were asked to complete a separate interview for each resident because some of the questions could be answered differently depending on the resident.

The independent assessors completed questionnaires on the residents' daily functioning in the ward at baseline (T0), after the intervention (T1), and at the 2-week follow-up (T2). Results relevant to the observations of the independent assessors will

XSL•FO
RenderX

be reported in a separate paper on the effect evaluation of the digital photo activity.

Analysis

Quantitative Analyses

Participants' background characteristics were analyzed using descriptive statistics in SPSS software (version 28.0; IBM Corp). Differences between the photo activity and control activity conditions were tested using 2-tailed *t* tests for continuous variables, Mann-Whitney *U* tests for ordinal variables, and Pearson chi-square tests for nominal variables. Effect sizes were interpreted using Cohen *d* for *t* tests (small=0.2, medium=0.5, and large is ≥0.8), *r* for Mann-Whitney *U* tests (small=0.1, medium=0.3, and large is ≥0.5), and phi coefficient (ϕ; for 2 × 2 tables) or Cramér V (for tables larger than 2 × 2) for Pearson chi-square tests, with the effect size based on the number of categories involved.

Data from closed-ended questions were summarized as percentages or as means and SDs, depending on the measurement level. As not all questions were comparable between the photo activity and control groups—some questions were intervention related—only selected comparable closed-ended questions were analyzed for between-group differences for each of the participant groups (residents, formal carers, and informal carers) using chi-square tests, *t* tests, or Mann-Whitney *U*-tests.

Google Analytics recorded information from users identified as "carer-resident" pairs with unique user IDs. The first sessions for the trial started on February 15, 2021; however, because of a delay in setting up the live environment for Google Analytics for the Fotoscope web application, data collection only started on March 12, 2021. Data collection was completed on August 6, 2023. For privacy reasons, Google Analytics only retains data for 26 months [31], and data were not backed up. As a result, data from 10 users who participated in the trial from March to October 2021 were automatically deleted before analyses were conducted in August-September 2023.

Observed session durations extracted from Google Analytics were matched to the dates and times recorded in the carer self-report registration forms. The agreement between total observed and self-reported durations (in minutes) was analyzed using the Spearman rho correlation coefficient. Percentages of total unique page views for the main pages were calculated from the raw data. The frequency of selecting main photo categories was also calculated from raw data (unique page views) and summarized as percentages. Individual reports were viewed to track events or actions taken by specific users during each session (eg, viewing the Profile page or opening a photo from the Photo Selection page). These actions were recorded per user on a binary (yes or no) scale and reported as the percentage of sessions in which events of interest (eg, adding a photo to the Favorites or Photo Selection pages) occurred. The use of a tablet or a different device for the Fotoscope web application was counted per user and summarized as percentages.

Qualitative Analyses

Data from the open-ended questions in the semistructured interviews were analyzed using the framework of theoretical thematic analysis, which was defined by Braun and Clarke [32] as thematic analysis that is built upon existing theoretical work in the literature. The researchers (JROT: PhD student, English speaking; MV: master's student, English and Dutch speaking; and SAMS: senior researcher, English and Dutch speaking) worked together following the 6 phases of thematic analysis [32]. The researchers individually familiarized themselves with the data (phase 1) by reading all answers to the open-ended and closed-ended questions of the semistructured interviews. Next, they generated initial codes (phase 2) based on aspects of the MRC guidance (contextual, implementation, and mechanism-of-impact factors) [17], which served as the guiding framework for this thematic analysis. The researchers studied the data to examine how the data fit within the MRC guidance framework; thus searching for themes (phase 3) involved assessing whether the data from the semistructured interviews fit within the main themes of context, implementation, and mechanisms of impact factors.

Themes and codes were reviewed (phase 4) by the 3 researchers through several online meetings. Themes and subthemes were then named and defined (phase 5). As stated by Braun and Clarke [32], these phases are recursive, and the researchers moved back and forth between phases as necessary. Reporting of the results (phase 6) followed the COREQ (Consolidated Criteria for Reporting Qualitative Research) guidelines [33]. MV and SAMS worked with the original data in Dutch, while JROT used an English translation generated using DeepL [34]. The researchers agreed on the English translations of the quotes and developed the codes and themes in English.

Results

Characteristics of Participants

Of the 83 resident–informal carer pairs recruited for the RCT, 16 (19%) dropped out for different reasons after randomization (Figure 3). Among the experimental group (n=32 resident–informal carer pairs), 29 residents and 28 informal carers completed semistructured interviews after the final session. Among the control group (n=35 resident–informal carer pairs), 29 residents and 27 informal carers completed the interviews. Reasons for not completing the interviews are provided in Figure 3. No informal carers delivered the experimental or control intervention, partly due to restrictions in visiting the nursing home during the COVID-19 pandemic and partly because of other reasons, such as not having enough time to visit the ward as needed because of their job or other concurrent tasks.

Of the 61 formal carers recruited for the RCT, 42 (69%) completed the trial and the semistructured interviews (carers experimental group: n=18, 43%; carers control group: n=24, 57%). As a number of carers conducted either the experimental activity or control activity with >1 resident, they were asked to complete a separate interview for each resident; hence, 50 semistructured interviews with residents (experimental group:

XSL•FO
RenderX

n=23, 46%; control group: n=27, 54%; Figure 4) were completed by formal carers.

Tables 1, 2 and 3 present the background characteristics of the residents, informal carers, and formal carers, respectively, who completed the semistructured interviews. There were no significant differences between the experimental and control groups.

Figure 5 shows the thematic map with the 3 main themes identified in the qualitative thematic analysis, while Multimedia Appendix 3 provides a detailed overview of the themes, subthemes, and sample data extracts. Relevant qualitative results from the thematic analysis are integrated with the quantitative results in the following subsections.

Figure 3. Recruitment of residents and informal carers, showing randomization into the experimental and control groups, as well as dropouts and reasons for dropping out. RCT: randomized controlled trial; SSI: semistructured interview; T1: after the intervention.

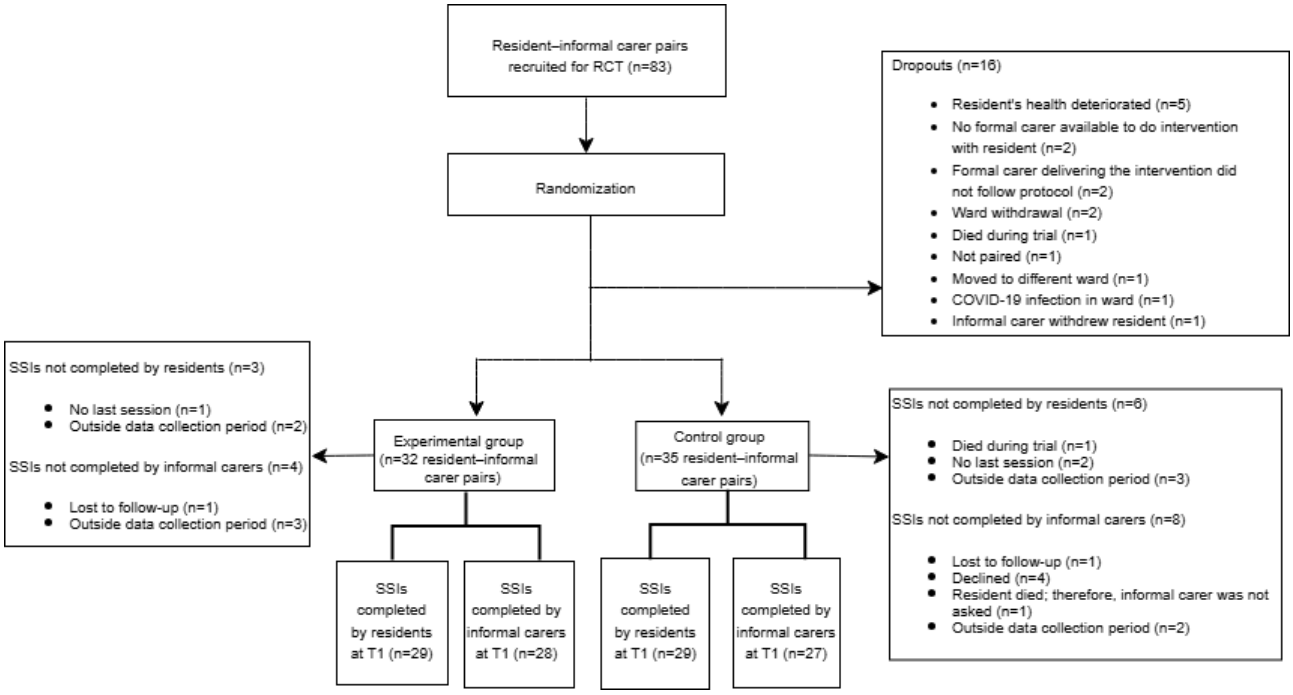

XSL•FO
RenderX

Figure 4. Recruitment of formal carers, showing randomization into the experimental and control groups, as well as dropouts and reasons for dropping out. RCT: randomized controlled trial; SSI: semistructured interview.

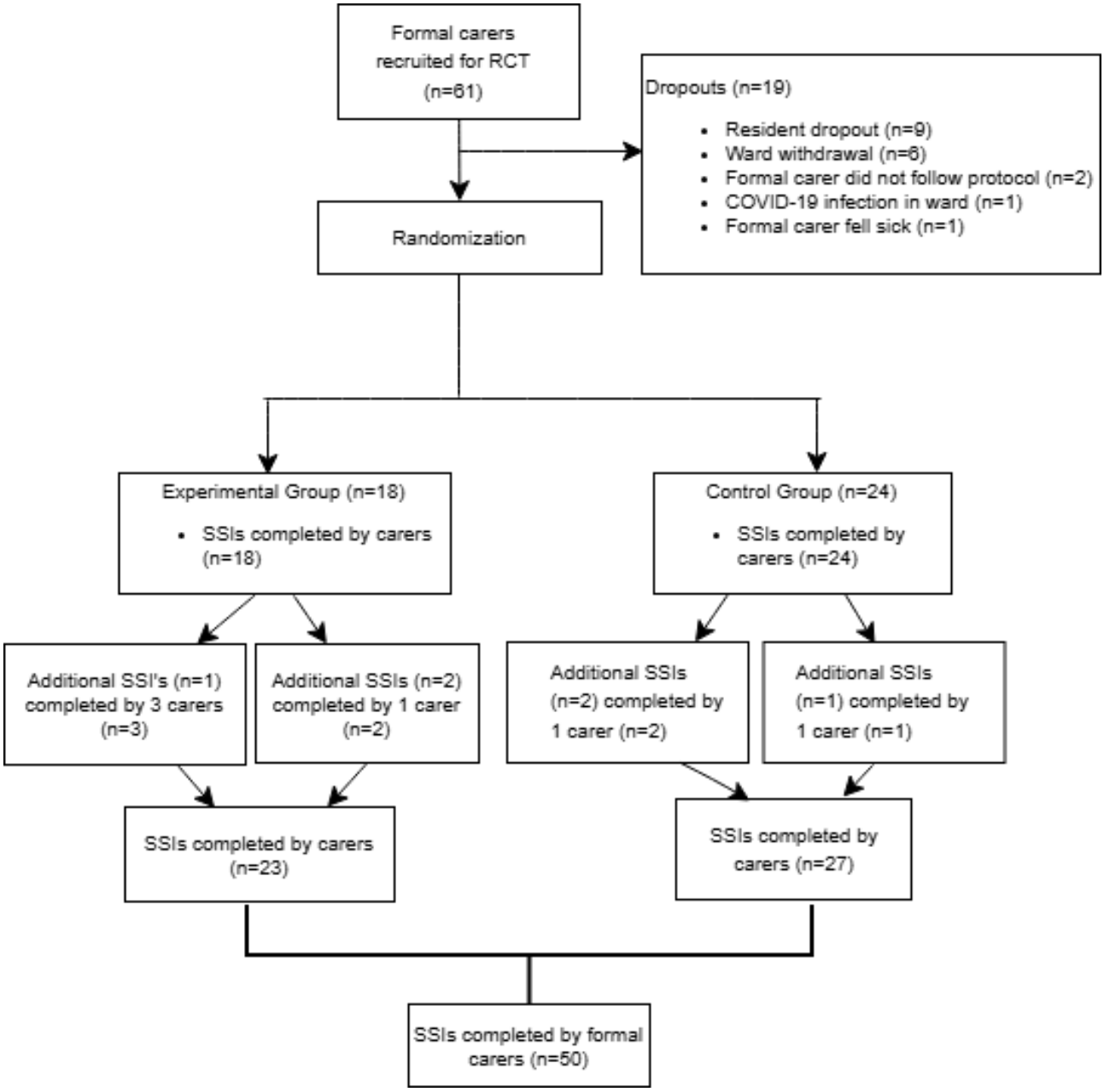

XSL•FO
RenderX

Table 1. Background characteristics of nursing home residents in both the experimental and control groups who completed the semistructured interviews after the 4-week intervention.

Characteristics	Experimental group (n=29)	Control group (n=29)	Difference tests				*P* value[a]
			Chi-square (*df*)	*t* test (*df*)	*U*	*z*	
Sex: female, n (%)	24 (83)	22 (76)	0.1[b] (1)	—[c]	—	—	.75
Age (y), mean (SD; range)	84.03 (7.36; 68-96)	84.38 (7.48; 62-96)	—	−0.177 (56)	—	—	.86
Education, n (%)			—	—	308.0	−1.502	.13
Elementary	7 (24)	6 (21)					
Lower secondary education	19 (66)	11 (38)					
Higher secondary education	1 (3)	5 (17)					
University	2 (7)	5 (17)					
Missing	0 (0)	2 (7)					
Time spent in nursing home (y), mean (SD; range)	1.714 (1.460; 0.3-7.0)	2.038 (2.090; 0.2-10.0)	—	—	365.5	−0.427	.68
Type of dementia, n (%)			1.1[b] (1)	—	—	—	.30
Alzheimer disease	17 (59)	21 (72)					
Other	12 (41)	7 (24)					
Missing	0 (0)	1 (3)					
Severity of dementia (GDS[d] scores), n (%)			—	—	398.5	−0.380	.70
4	4 (14)	3 (10)					
5	15 (52)	15 (52)					
6	10 (34)	11 (38)					

[a]Significance level set at $P<.05$.

[b]Chi-square test with Yates continuity correction.

[c]Not applicable.

[d]GDS: Global Deterioration Scale.

XSL•FO
RenderX

Table 2. Background characteristics of informal carers in both the experimental and control groups who completed the semistructured interviews after the 4-week intervention.

Characteristics	Experimental group (n=28)	Control group (n=27)	Difference tests				*P* value[a]
			Chi-square (*df*)	*t* test (*df*)	*U*	*z*	
Sex: female, n (%)	21 (75)	20 (74)	0.0[b] (1)	—[c]	—	—	.99
Age (y), mean (SD; range)	57.04 (9.03; 31-74)	60.74 (11.4; 27-84)	—	−1.341 (53)	—	—	.39
Education, n (%)			—	—	317.5	−1.108	.22
Elementary	0 (0)	2 (7)					
Lower secondary education	17 (61)	9 (33)					
Higher secondary education	3 (11)	4 (15)					
University	8 (29)	12 (44)					
Missing	0 (0)	0 (0)					
Relationship to resident, n (%)			1.6 (3)	—	—	—	.65
Husband, wife, or partner	4 (14)	6 (22)					
Son or daughter	3 (11)	4 (15)					
Daughter-in-law or son-in-law	18 (64)	16 (59)					
Other	3 (11)	1 (4)					
Paid work: yes, n (%)	21 (75)	15 (56)	1.5[a] (1)	—	—	—	.22

[a]Significance level set at $P<.05$.

[b]Chi-square test with Yates continuity correction.

[c]Not applicable.

XSL•FO
RenderX

Table 3. Background characteristics of formal carers in both the experimental and control groups who completed the semistructured interviews after the 4-week intervention.

Characteristics	Experimental group (n=18)	Control group (n=24)	Difference tests				*P* value[a]
			Chi-square (*df*)	*t* test (*df*)	*U*	*z*	
Sex: female, n (%)	17 (94)	23 (96)	0.0[b] (1)	—[c]	—	—	.99
Age (y), mean (SD; range)	45.7 (13.0; 20-62)	43.6 (15.1; 18-63)	—	−0.470 (40)	—	—	.64
Education, n (%)			—	—	194.5	−0.843	.40
Elementary	0 (0)	0 (0)					
Lower secondary education	16 (89)	19 (79)					
Higher secondary education	1 (6)	2 (8)					
University	1 (6)	3 (12)					
Missing	0 (0)	0 (0)					
Job role, n (%)			2.5 (3)	—	—	—	.48
Carer	6 (33)	5 (21)					
Nurse	1 (6)	0 (0)					
Activity supervisor	3 (17)	6 (25)					
Other	8 (44)	13 (54)					
Experience (y) in psychogeriatrics, mean (SD; range)	5.22 (4.62; 1.0-16.0)	10.75 (11.13; 1.0-40.0)	—	—	165.5	−1.297	.20

[a]Significance level set at *P*<.05.

[b]Chi-square test with Yates continuity correction.

[c]Not applicable.

Figure 5. Thematic map from qualitative analyses showing the main themes (context, implementation, and mechanisms of impact), subthemes, and subsubthemes.

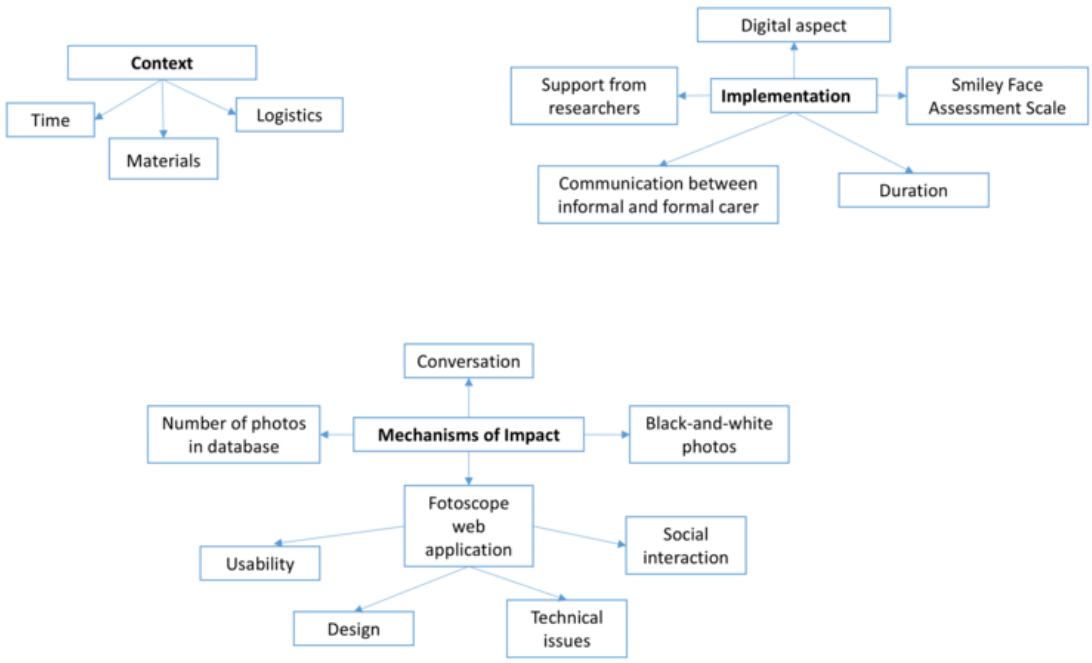

XSL•FO
RenderX

Results From Analysis of Semistructured Interviews and Google Analytics Data

Overview

Multimedia Appendix 2 presents all answers to the closed-ended questions from the semistructured interviews summarized as percentages or as means and SDs.

In total, 41 users (representing carer-resident pairs) were given a Fotoscope account. Of these 41 users, 38 (93%) were recorded as accessing the app (of the remaining 3 users, 2, 67% had started the intervention before Google Analytics was set up; and 1, 33% included a resident who dropped out before starting the intervention). As data of 10 (26%) of the 38 users were no longer available at the time of evaluation, we analyzed the data of 28 (74%) users. Of these 28 users, 3 (11%) had partial or missing data. Of the remaining 25 user pairs, 4 (16%) had residents who dropped out (n=2, 50% dropped out before the activities started due to worsening conditions; n=1, 25% dropped out after 3 sessions due to worsening conditions; and n=1, 25% dropped out after 3 sessions because the formal carer was too busy); therefore, they were excluded from analysis, leaving data from 21 (84%) users (representing carer-resident pairs) for the final analysis.

Contextual Factors

The majority of the residents (experimental group: 23/29, 79%; control group: 27/29, 93%) found the space for viewing photos with their formal carers or having a conversation with them to be comfortable and quiet.

Most of the informal carers in the photo activity group (13/28, 46%) experienced no barriers to talking to the formal carers about the residents' personal interests or their experience with the photo activity, while some cited barriers, including a lack of time (7/28, 25%) and "other reasons" (4/28, 14%), or did not respond (4/28, 14%). The majority of informal carers in the control group (15/27, 56%) reported no barriers to talking to the formal carers about the residents' experience with the general conversation activity, while those who responded with "other reasons" (10/27, 37%) reported not knowing that it was an option to ask the formal carers about the general conversation activity with the residents. Other reported barriers included being on vacation (1/27, 4%) and a lack of time (1/27, 4%).

The majority of formal carers in the photo activity group (10/18, 56%) reported using a tablet "a lot" previously. The remaining responded as follows: "a little" (5/18, 28%), "hardly" (1/18, 6%), or "not at all" (2/18, 11%). The majority of the formal carers in both the photo activity group (14/18, 78%) and the control group (15/24, 63%) reported experiencing no barriers in conducting the activities. In the thematic analysis, formal carers from both groups who reported experiencing barriers mentioned difficulty in planning time for the activities, given their busy schedules in the ward, sometimes even having to find the time outside of their shifts to conduct the activity with the resident:

> *Planning the photo activity is very important because of our workload. Time and sufficient staff presence are very essential to consider.* [Formal carer 109]

Finding a quiet place to conduct the activity seemed to be more difficult for formal carers in the control group, while formal carers in the photo activity group mentioned difficulty in arranging materials for the activity, such as finding an available tablet in the ward (we had offered to lend tablets to the wards at the start of the study, but none chose this option).

Implementation Factors

The majority of the residents (25/29, 86%) reported that the photo activity went well as an activity on the tablet, and the majority (21/29, 72%) also reported seeing the photos clearly and finding them sufficiently large and sharp. Moreover, the majority (23/29, 79%) reported viewing a "good" amount of photos with their carers, and most (14/29, 48%) felt "neutral" about the black-and-white photos.

The majority of the informal carers (experimental group: 25/28, 89%; control group: 20/27, 74%) felt well informed about the activities that the residents performed in the nursing home during the intervention period. However, the thematic analysis revealed that informal carers in both groups desired more communication and updates from the formal carers regarding the residents' engagement in the activities.

A majority of formal carers in both the photo activity (12/18, 67%) and control (20/24, 83%) groups felt that the medium and format of the training sessions were effective (Table 4). More than half of the formal carers in the photo activity group were positive about the online training sessions and rated them as "good" (5/18, 28%) or "satisfactory" (8/18, 44%), and the majority (12/18, 67%) also felt that the training duration was "just right." A large majority of the formal carers in the control group also rated their online training sessions positively (824, 33% rated it as "good"; and 14/24, 58% rated it as "satisfactory"), and a large majority (20/24, 83%) also felt that the training duration was "just right." Formal carers from both groups expressed a preference for a live training session and in-person support or demonstration. A majority of the formal carers from both the photo activity (14/18, 78%) and control (16/24, 67%) groups felt that the support provided by the researchers and resources was satisfactory. In the thematic analysis, several formal carers in both the photo activity and control groups commented on the duration of the online training sessions and the duration of the activity itself. Regarding the duration of the online training sessions, they seemed to agree that it should be shorter for both groups. The comments for activity duration were mixed. Formal carers in both the photo activity and control groups appreciated the support provided by the researchers and resources. A formal carer in the control group suggested incorporating a video example in the training sessions to demonstrate how the general conversation activity could be conducted, while a formal carer in the photo activity group suggested expanding the manual to include instructions on how to download the Fotoscope app on a new device.

The majority of the formal carers (16/23, 70%) found that the photo activity, using the Fotoscope app, was "easy to implement." While the carers were positive about the app, a theme revealed by the thematic analysis was that they felt that the Fotoscope app was "not for everyone":

XSL•FO
RenderX

> *Very good [app]. Albeit not suitable for everyone. Everything digital, this is something that the target group is not familiar with.* [Formal carer 5]

Formal carers who expressed this concern had conducted the photo activity with residents who had a GDS score of 5 or 6 [23]. A minority of carers in both groups also reported that residents with a GDS score of 5 or 6 had a hard time understanding the SFAS [30] and found it difficult to choose a face that represented how they felt in the moment. Thematic analysis revealed that more formal carers from the control group reported that it was difficult for the residents to understand the semistructured interview questions. Formal carers who conducted the activities with multiple residents also reported that completing the interview multiple times was a burden.

Table 4 presents results from the statistical analysis of differences between the formal carers in the photo activity group and those in the control group in terms of how they felt about the medium and format of the training sessions, the activity duration, the 4-week activity period, and the activity frequency. No significant differences were found (Table 4).

The relationship between the total duration of self-reported app use as measured by the carer self-report registration form and the corresponding (ie, matched to the date and time recorded in the self-report form) total duration of observed app use as measured by Google Analytics was investigated using the Spearman rho correlation coefficient. There was no significant correlation between the 2 variables (n=21; r=−0.108; P=.64).

The Google Analytics data showed that desktop computers were used most of the time to access the Fotoscope app (19/32, 59%), followed by tablets (8/32, 25%) and mobile phones (5/32, 16%). Formal carers in the photo-activity group sometimes used more than one device during the intervention period, which is accounted for in the total devices used (N=32).

Table 4. Results regarding implementation factors for formal carers in the photo activity (experimental) and general conversation (control) groups based on comparable questions from the semistructured interviews after the 4-week intervention.

Outcome variable	Experimental group (n=18), n (%)	Control group (n=24), n (%)	Difference tests		P value[a]	Effect size	
			U	Chi-square (df)		r	Cramér V
Medium and format training			209.5	—[b]	.83	−0.03	—
Very effective	2 (11)	1 (4)					
Effective	12 (67)	20 (83)					
Somewhat effective	3 (17)	2 (8)					
Not effective	1 (6)	1 (4)					
Activity duration[c]			—	0.4 (2)	.84	—	0.09
Too short	1 (4)	2 (7)					
Just right	13 (56)	13 (48)					
Too long	9 (39)	11 (41)					
Missing	0 (0)	1 (4)					
4-wk activity period			—	1.1 (2)	.59	—	0.16
Too short	2 (11)	1 (4)					
Just right	9 (50)	15 (62)					
Too long	7 (39)	8 (33)					
Activity frequency			—	0.8 (2)	.67	—	0.14
Too little	0 (0)	1 (4)					
Just right	10 (56)	12 (50)					
Too often	8 (44)	11 (46)					

[a]Significance level set at P<.05.

[b]Not applicable.

[c]The n varies for Activity duration, as some formal carers were asked to answer this question multiple times if they did the photo-activity or control activity with more than one resident. Thus formal carers in the photo-activity is n=23, and control activity n=27, as in Figure 4.

Mechanism-of-Impact Factors

The majority of residents in the photo activity group (25/29, 86%) reported that they would like to continue participating in the activity with their formal carers. The residents gave the Fotoscope app a score of 8.6 out of 10 in terms of enjoyment. Thematic analysis revealed that the residents in the photo activity group responded more positively when asked whether they wanted to continue performing the activity, especially if new photos were introduced.

XSL•FO
RenderX

A large majority of the formal carers in the photo activity group reported that the reaction of the resident to the photo activity was "very positive" (10/23, 43%) or "positive" (8/23, 35%). The formal carers in the control group predominantly reported that the reaction of the resident to the general conversation activity was "positive" (12/27, 44%), while a minority reported that the reaction was "very positive" (5/27, 19%). Table 5 presents the results from the statistical analysis of differences between residents (in terms of feeling heard, feeling better known by their carers, and their experience of the general conversation), informal carers (whether they heard anything back from the resident about the activity), and formal carers (whether they thought the activity was enjoyable for the resident, was a good way to get to know the resident better, and how much better they got to know the resident from the activity) in the photo activity and control groups.

There was a significant difference between the photo activity and control groups in terms of how the residents felt about having conversations with their formal carers, with a medium effect size (U=533.0, z=2.865, r=0.39; P=.004), favoring the photo activity group (median 4.00, "very nice," n=28) over the control group (median 3.00, "nice," n=27).

There was also a significant difference in how much better the formal carers in the photo activity group (median 2, "yes, a little better," n=22) got to know the resident in the past month compared to those in the control group (median 2, "yes, a little better," n=27; U=390.5, z=2.114, r=0.302 [medium effect]; P=.04). The answer options in the semistructured interviews for the photo activity group were as follows: 1=no, not better; 2=yes, a little better; and 3=yes, much better. As the answer options for this question for the control group differed (an additional answer option of "yes, better" was provided after "yes, a little better"), it was decided to combine the answers "yes, better" and "yes, a little better" from the control group interviews in the analysis. Of the formal carers in the experimental group, 48% (11/23) and 39% (9/23) reported getting to know the resident with dementia "a little better" and "much better," respectively (compared to 17/27, 63% and 4/27, 15%, respectively, of the formal carers in the control group). Of the formal carers in the experimental group, 9% (2/23) reported that they did not get to know the person with dementia "better" compared to 22% (6/27) of the formal carers in the control group. Thematic analysis reflected similar findings: formal carers in both the photo activity and general conversation groups mentioned learning more about the residents:

> *During care, of course we chat plenty but here I went a bit deeper into the past, so yes I did get to know her much better than I already did.* [Photo activity group, formal carer 201]

> *I now know a little more about her past, as far as she could still tell.* [Control group, formal carer 204]

However, a formal carer in the control group did mention that the general conversation activity was similar to what she already did as part of her work:

> *Personally it has no added value for me. This is what I do daily in my work—speak to residents every day.* [Formal carer 116]

Finally, the carers in the photo activity group (n=23) gave a significantly higher rating to the activity (median 9.00) compared to the carers in the general conversation group (n=27; median 8.00) as a way to get to know the resident with dementia better (U=419.0, z=2.169, r=0.307 [medium effect size]; P=.03).

There was no statistically significant association between the experimental and control groups on whether residents felt heard during the conversation with their formal carer, whether informal carers received feedback from the residents about the activity, and how much formal carers felt the residents enjoyed the activities (Table 5).

XSL•FO
RenderX

Table 5. Results regarding the mechanism-of-impact factors for residents, informal carers, and formal carers in the photo activity (experimental) and general conversation (control) groups based on comparable questions from the semistructured interviews after the 4-week intervention.

Outcome variable	Experimental group	Control group	Difference tests			*P* value[a]	Effect size	
			Fisher exact test	Chi-square (*df*)	*U*		ϕ	*r*
Residents, n/N (%)								
Feeling heard: yes	26/29 (90)	24/29 (83)	__[b]	—	—	.49	—	—
Felt better known by their carer			—	0.3[c] (1)	—	.60	−0.12	—
A lot better or better	17/29 (59)	15/29 (52)						
A little better or not better	6/29 (21)	9/29 (31)						
Missing	6/29 (21)	5/29 (17)						
Experience of general conversation			—	—	533.0	.004	—	0.39
Very nice	16/29 (55)	6/29 (21)						
Nice	11/29 (38)	15/29 (52)						
Somewhat nice	0/29 (0)	5/29 (17)						
Not nice	1/29 (3)	1/29 (3)						
Missing	1/29 (3)	2/29 (7)						
Informal carers, n/N (%)			—	—	—	—	—	—
Resident talked to them about the activity: yes	7/28 (25)	5/27 (19)						
Formal carers			—	—	264.0	.20	—	.20
Enjoyable for people with dementia in a nursing home (0-10), median (IQR)	8.00 (7-9)	8.00 (7-8)						
Knew resident better, n/N (%)			—	—	390.5	—	—	.30
Yes, much better	9/23 (39)	4/27 (15)				.04		
Yes, a little better	11/23 (48)	17/27 (63)						
No, not better	2/23 (9)	6/27 (22)						
Missing	1/23 (4)	0/27 (0)						
Way of getting to know resident with dementia better (0-10), median (IQR)	9.00 (7-9)	8.00 (6-8)	—	—	419.0	.03	—	0.31

[a]Significance level set at $P<.05$.

[b]Not applicable.

[c]Chi-square test with Yates continuity correction.

Experienced Usability

In the thematic analysis, the formal carers in the photo activity group described the Fotoscope app as easy to use, simple, clear, and visually appealing. Technical issues were experienced (eg, difficulties logging in or the app not working on the tablet). However, carers mentioned that the Fotoscope app enhanced social interaction, helping them to get to know the resident better:

> *It triggers more conversation, especially for the quieter ones, this can work up to a conversation.* [Formal carer 201]

Experienced Usefulness

The majority of the formal carers in the photo activity group found the preparation for the photo activity useful (14/18, 78%) or very useful (2/18, 11%), while a small minority found it not useful (2/18, 11%). Whereas the majority of the formal carers in the photo activity group found the black-and-white photos "pleasant" (4/18, 22%) or "neutral" (7/18, 39%), some found them "unpleasant" (7/18, 39%). A number of formal carers in the photo activity group noted that some photos would really have been better with color:

> *I was sorry we didn't have color photos. Some things really needed color. Even my client asked me why I*

XSL•FO
RenderX

> *didn't have color photos.* [Formal carer 105 with resident 109]

However, formal carers also mentioned that while they preferred color photos, residents did not mind viewing black-and-white photos:

> *I would personally like it better if the pictures were in color, but for the resident this did not matter.* [Formal carer 105 with resident 121]

> *[T]hey [residents] mostly liked black-and-white pictures too.* [Formal carer 101]

The number of photos in the database was identified as a subtheme in the thematic analysis. While the formal carers described the selection as varied, some noted a lack when residents had specific interests; for example, a resident wanted to see more ballet pictures.

Google Analytics showed that while all users viewed the Profile page, 19% (4/21) did not add photos to the Photo Selection page, and only 48% (10/21) used the photo selection tool on the Profile page to select photos. In addition, while the majority of users (17/21, 81%) opened photos from the Photo Selection page as instructed, 19% (4/21) apparently accessed photos through other means. Furthermore, only 33% (7/21) of the users used the Favorites page.

Learnability

The majority of formal carers in the photo activity group (12/18, 67%) said that it was easy to learn how to use the Fotoscope web application. Google Analytics showed that the most viewed page of the main pages was the Profile page (257/594, 43.3% total unique page views), followed by the general Themes page (104/594, 17.5%), Favorites page (85/594, 14.3%), "joker" page (57/594, 9.6%), and Search (43/594, 7.2%). The User Guide (29/594, 4.9%) and Information page (19/594, 3.2%) were the least viewed pages. The most frequently selected main photo category (based on Google Analytics' unique events count) was people (71/404, 17.6%), followed by activities (64/404, 15.8%), and places and animals (both 56/404, 13.9%). The least selected main photo categories were experiences (53/404, 13.1%), nature (53/404, 13.1%), and things (51/404, 12.6%).

Adoption

The majority of formal carers in the photo activity group (14/18, 78%) agreed that the Fotoscope app can be used as part of daily care activities in the nursing home.

Discussion

Principal Findings and Comparisons With Prior Work

Overview

The goal of this study was to determine contextual, implementation, and mechanism-of-impact factors [17] that may have facilitated or hindered the implementation of the photo activity intervention for residents with dementia and potentially influenced the intervention outcomes.

To answer the RQs, we conducted a mixed methods process evaluation alongside a pilot RCT investigating the feasibility and effectiveness of the photo activity for residents and their informal and formal carers. The main findings of this process evaluation were as follows. For contextual factors affecting the implementation of the intervention (RQ1), formal carers struggled to find the time to conduct the activity during their work hours. For implementation factors, which looked at how users assessed the preparation, introduction, and delivery of the interventions (RQ2), it was found that viewing digital photos on a tablet was feasible for residents with dementia; formal carers from both groups were mostly positive and satisfied regarding the online training sessions but suggested that, in the future, they would like shorter sessions as well as in-person training and support; and informal carers would like more updates from formal carers. For the mechanisms of impact (RQ3), looking into how users experienced and valued the digital photo activity versus the control activity, it was found that residents who performed the digital photo activity with their carers had a more positive experience compared to residents in the control activity. Formal carers who conducted the digital photo activity reported getting to know the residents "much better" compared to formal carers in the control activity. The Fotoscope app as a tool to deliver the digital photo activity was also rated as a better way to get to know the residents as opposed to having a general conversation. Informal carers of residents in the photo activity group also reported more contact with formal carers compared to informal carers in the control activity.

In the following subsections, we provide a more detailed discussion of the findings in relation to each RQ.

How Did Factors External to the Intervention Affect the Implementation of the Intervention (Context)?

While minimal contextual factors impeded the implementation of the photo activity or the control activity, a salient theme that emerged was the difficulty that formal carers experienced in scheduling time to conduct the activities with the resident alongside their daily care tasks. In some cases, formal carers mentioned having to conduct the activity outside of their work shift. Problems with workload and staffing are common barriers to implementing psychosocial interventions successfully in nursing homes [14]. Staff are pressured to prioritize tasks that are related to physical and safety needs [35]. Due to the restrictions on visiting nursing homes during the COVID-19 pandemic [36], during the initial phase of the study, no informal carers were involved in delivering either activity to residents. Later, no informal carers were willing to invest the time needed to provide the intervention or control activity.

It is not clear which device carers used to prepare for the photo activity and which device they used to conduct the activity; therefore, it is possible that carers used a desktop computer to prepare and select initial photos for the activity and then used the tablet as instructed during the activity with the resident.

How Did Residents and Their Informal and Formal Carers Assess the Preparation, Introduction, and Delivery of the Interventions (Implementation)?

Most residents responded positively to viewing photos on a tablet, consistent with literature suggesting that using tablets in

XSL•FO
RenderX

art-based psychosocial interventions could have positive effects for individuals with dementia [37,38].

Informal carers from both groups mentioned wanting more updates from formal carers on how the residents were responding to the activities. A previous study found that greater involvement of informal carers in the daily care of residents was an important factor in the successful implementation of psychosocial interventions because this also helps formal carers to become more familiar with the residents and their identity within a family context [39]. The communication between informal and formal carers fosters more trust, which benefits overall care [14].

Regarding the online training sessions, while the majority of the formal carers from both groups were positive and satisfied, formal carers from both groups who were dissatisfied with the training sessions suggested shortening the duration and expressed a preference for live training and in-person support or demonstration. The setting of the nursing home itself may also have influenced how staff received the training. Kuske et al [40] found that when staff feel overburdened, they see training as an additional task rather than beneficial continuing education, which could explain why shortening the duration of the training sessions was a salient theme in this process analysis.

A minority of formal carers who conducted the photo activity with residents with a more severe level of dementia (ie, a GDS score of 5 or 6) [23] described the digital photo activity as not suitable and the 30-minute activity duration as too long for these residents due to their “short attention spans” or because “they get tired easily.” This is similar to previous work suggesting that residents’ deteriorating health could be a barrier for formal carers implementing psychosocial interventions [40].

However, it is worth noting that the majority of the residents had a GDS score of 5 or 6 [23], and the experience of most formal carers was that the photo activity intervention was easy to implement and that the activity duration was just right. This suggests that the photo activity can be a suitable addition to tools that nursing homes could use, including for residents with more severe dementia, because most interventions for people with more severe dementia are focused on reducing problem behaviors rather than improving social interactions or quality of life [41,42].

Some residents experienced difficulty in responding to the SFAS [30], which presented 5 emotions or faces. Future trials could explore using a simpler scale similar to the SFAS but with only 3 faces (happy, neutral, and sad) because a study found that older adults sometimes have difficulty differentiating between “very happy” and “happy” and between “very sad” and “sad” [43].

Google Analytics may not have accurately recorded actual use of the app by formal carers for the process evaluation, as shown by the lack of correlation between self-reported use and observed use via Google Analytics. Currently, photos can be viewed in the Fotoscope web application via the Themes page without logging in; therefore, it is possible that formal carers may not have logged in during some of the sessions and manually searched for photos instead of accessing the Profile page or Photo Selection page. For future trials, it could be useful to consult with software and app developers and explore integrating an in-app analytics function within the Fotoscope app rather than relying on an external analytics provider.

How Did the Residents With Dementia and Their Informal and Formal Carers Experience and Value the Digital Photo Activity Versus the General Conversation Activity (Mechanisms of Impact)?

Residents in the photo activity group reported a more positive experience engaging in the activity with their formal carers. The positive findings regarding using generic photos in the photo activity are in line with a similar study that used generic photos as conversation prompts for people with dementia, where it was found that compared to using family photos, people with dementia told more detailed and personally relevant stories when shown generic photos [42].

More residents in the photo activity group were eager to continue the activity with their formal carers. This highlights the importance of investing in interventions that foster social health for residents with dementia in the nursing home, rather than solely focusing on physical needs or problematic behaviors [41,44].

Informal carers in the photo activity group had more contact with the formal carers, and more residents in the photo activity group talked to their informal carers about the activity. This finding shows that the photo activity may be able to enhance the involvement of informal carers, a factor that is crucial to consider in the successful implementation of psychosocial interventions [14,39].

The formal carers in the photo activity group perceived the residents’ reaction to the conversation about the photos as more positive. These carers also reported getting to know the residents much better and rated the Fotoscope app higher (compared to a general conversation) in terms of using it as a tool to get to know the resident better. This is in line with the findings of Lawrence et al [14] indicating that when formal carers feel that an intervention helps them develop a more personal relationship with a resident and that the intervention is meaningful for the resident, they are more likely to continue using the intervention because they see that it has value. This is further supported by the findings from the process evaluation: whereas a formal carer mentioned that the control activity had no added value for her, the majority of formal carers in the photo activity group felt that the intervention could be adopted into daily care because of its value.

The Fotoscope web application was generally well received and was described as having a simple and pleasing design and user interface, which is crucial for engaging users of technological interventions [6,45]. More formal carers were vocal about the black-and-white photos compared to the residents, but only a minority of the formal carers found the black-and-white photos unpleasant. Previous studies have shown that both color and black-and-white photos elicit reactions from residents with dementia equally [42,46]. Residents who participated in the photo activity were also observed to react to the black-and-white

XSL•FO
RenderX

photos in the Fotoscope app and mostly found the photos pleasant or neutral.

Strengths and Limitations

This study has a number of strengths. First, data were collected from nursing home residents, informal carers, and formal carers, providing a wider understanding of the process of implementing the intervention from all viewpoints as well as the contextual and mechanism-of-impact factors that potentially influenced the outcomes of the RCT. Mixed methods were used; quantitative and qualitative data were collected from semistructured interviews with participants, and objective data were obtained from Google Analytics. The semistructured interviews were structured according to the MRC guidance for process evaluations of complex interventions [17]. The digital protocol enabled research assistants to conduct interviews with the residents via video calls, ensuring data collection despite the lockdowns and visiting limitations during the COVID-19 pandemic.

Some methodological limitations included changing the frequency of the interventions after the trial had already been running for 1 month after receiving feedback from the first round of formal carers delivering the interventions. Google Analytics data were available for only 21 (51%) of the 41 users (representing carer-resident pairs) because the researchers were unaware and not informed of the data reaching the storage time limit and being deleted from the account before data collection concluded and data analysis started. The manner of collecting responses for the semistructured interviews with carers varied: some carers answered the interview questions on their own, while others preferred to be interviewed over the phone, which may have influenced their answers differently. In relation to the manner of collecting responses for the interviews, not all closed-ended questions invited an elaborate response. This was done keeping in mind the time constraints and effort required from participants. However, the trained research assistants were instructed to write down any answer that may have been given in place of or in addition to the answers to the closed-ended questions, which enabled us to learn, for example, that residents in the photo activity group were eager to continue performing the activity with their carers with new photos.

Finally, we were unable to request feedback from the participants regarding the results of the thematic analysis. Unfortunately, no informal carers were involved in delivering the control or experimental interventions, despite recruitment efforts aimed at future implementation. As a result, no data could be collected on implementation or mechanism-of-impact factors from informal carers providing the photo activity intervention.

Scientific Relevance of the Study and Recommendations for Future Studies

The mixed methods process evaluation showed that despite the complexity of the intervention, the digital photo activity was feasible to implement via online training sessions and off-site monitoring, providing a good basis for future larger-scale research on the intervention. The MRC guidance was useful in identifying contextual, implementation, and mechanism-of-impact factors that positively or negatively affected implementation and potentially influenced the intervention outcomes. It is recommended that a definitive trial investigating the effectiveness of the intervention should implement feedback from this process evaluation, such as in-person staff training and demonstration of the Fotoscope app. Informal carers could also be recruited to carry out the photo activity, now that COVID-19 restrictions are no longer in place, to explore what implementation issues would be in play for them and how they could benefit from offering the intervention. Incorporating an automatic record of activity appointments in the Fotoscope app instead of asking users to record their time and date of use manually might also provide a more accurate account of their use of the web application.

Clinical Relevance of the Study and Implications for Practice

The mixed methods process evaluation found that residents participating in the photo activity enjoyed it more than those participating in the general conversation activity. This confirms a previous finding that touchscreen art interventions could be beneficial for residents' psychological well-being [6]. Ensuring that tablets are easily accessible in the wards is recommended to encourage staff to conduct the photo activity. While it is noted that the control condition did not include any art-related components, part of the aims of this process evaluation was investigating how the photo activity intervention compared to an active control activity that was most likely to happen naturally within the context of the nursing home (in this case, sitting down with a resident for a chat over a warm drink compared to setting up a different art activity).

Formal carers in the photo activity group got to know the resident with dementia substantially better and rated the photo activity as a better way of getting to know the resident. This finding is relevant because previous literature has found that a more person-oriented care for people living with dementia leads to better outcomes associated with social health and quality of life [29,47], and the photo activity could be an easy and meaningful activity to help formal carers to get to know the residents better. In addition, the importance of formal carers communicating with the informal carers can be encouraged as an essential part of the photo activity.

A majority of the photo activity carers agreed that the intervention could be part of daily care activities, but important contextual factors such as heavy workload and a lack of staff need to be taken into account [14]. Involving the wider organization is important to allow formal carers to view psychosocial interventions as part of daily care tasks, rather than an added workload [48]. For the photo activity, supporting staff to conduct the intervention during their shifts, not outside of paid hours, is recommended, as well as being flexible with the activity duration.

Conclusions

This study showed that the initial training and preparation for formal carers in both the photo activity and control groups, as well as their perception of effort and time investment, did not seem to differ much. The findings suggest that implementing

XSL•FO
RenderX

the photo activity as an intervention does not require a significant amount of additional work, yet it offers notable benefits, including residents enjoying the photo activity more and formal carers getting to know the residents better. Therefore, it is suggested that the digital photo activity, as an enjoyable intervention with its easy to use web application, could be widely implemented and disseminated in nursing homes. As a person-centered intervention designed to enhance relationships among people with dementia in nursing homes and their informal and formal carers, it may have the potential to positively impact outcomes relating to social health and quality of life of the person with dementia, but this needs to be explored further.

Acknowledgments

The authors would like to thank participants who supported this study. The authors would also like to thank the research assistants for assisting with data collection. The authors would like to thank all research assistants involved in interviewing the participants and the participants themselves: the nursing home residents and their families, as well as the nursing home staff. The main randomized controlled trial as well as this process evaluation were funded by the European Union Marie Sklodowska-Curie Innovative Training Network Action program (2020-MSCA-ITN 2018; 813196), RCOAK Foundation, Foundation Hofje Codde and Van Beresteyn, Foundation for Support VCVGZ, Amsterdam UMC, location Vrije Universiteit, and Inholland University of Applied Sciences.

Authors' Contributions

RMD, PB, and JROT were responsible for developing the semistructured interview guide. JROT and MV were responsible for semistructured interview data collection and entry. Thematic analysis was carried out by JROT, MV, and SAMS. JROT, DPN, and RMD were responsible for methodology and analysis of Google Analytics data. JROT and RMD were responsible for methodology and analysis of quantitative data. JROT prepared the original draft. JROT, PB, TPE, RJJG, SAMS, MV, DPN, and RMD reviewed and edited the manuscript. RMD, RJJG, TPE, and SAMS were responsible for supervision. All authors contributed to the submitted manuscript and approved the final version.

Conflicts of Interest

Author SAMS received funding from the Ministry of Health, Welfare and Sports (funding #90001586). This funding was paid to the institution.

Multimedia Appendix 1
Topics for each Medical Research Council guidance for process evaluation component and the measurements used.
[DOCX File , 18 KB - formative_v9i1e56586_app1.docx]

Multimedia Appendix 2
Responses to closed-ended questions from the semistructured interviews with residents and carers.
[DOCX File , 43 KB - formative_v9i1e56586_app2.docx]

Multimedia Appendix 3
Thematic analysis themes, subthemes, and sample data extracts.
[DOCX File , 24 KB - formative_v9i1e56586_app3.docx]

References

1. Aëgerter L. Home page. Photographic Treatment. 2016. URL: https://photographictreatment.com/en/ [accessed 2024-11-14]
2. Theijsmeijer S, de Boo GM, Dröes RM. [A pilot study into person-centred use of photo's in the communication with people with dementia]. Tijdschr Gerontol Geriatr 2018 Sep;49(4):147-155. [doi: 10.1007/s12439-018-0258-4] [Medline: 30088252]
3. Sanders D, Scott P. Literature review: technological interventions and their impact on quality of life for people living with dementia. BMJ Health Care Inform 2020 Jan;27(1):e100064 [FREE Full text] [doi: 10.1136/bmjhci-2019-100064] [Medline: 31948938]
4. Dröes RM, Chattat R, Diaz A, Gove D, Graff M, Murphy K, INTERDEM sOcial Health Taskforce. Social health and dementia: a European consensus on the operationalization of the concept and directions for research and practice. Aging Ment Health 2017 Jan;21(1):4-17 [FREE Full text] [doi: 10.1080/13607863.2016.1254596] [Medline: 27869503]
5. Astell AJ, Bouranis N, Hoey J, Lindauer A, Mihailidis A, Nugent C, Technology Dementia Professional Interest Area. Technology and dementia: the future is now. Dement Geriatr Cogn Disord 2019;47(3):131-139 [FREE Full text] [doi: 10.1159/000497800] [Medline: 31247624]

XSL•FO
RenderX

6. Tyack C, Camic PM. Touchscreen interventions and the well-being of people with dementia and caregivers: a systematic review. Int Psychogeriatr 2017 Aug;29(8):1261-1280 [FREE Full text] [doi: 10.1017/S1041610217000667] [Medline: 28446258]
7. Woods D, Yuan F, Jao YL, Zhao X. Social robots for older adults with dementia: a narrative review on challenges and future directions. In: Proceedings of the 13th International Conference on Social Robotics. 2021 Presented at: ICSR '21; November 10-13, 2021; Singapore, Singapore p. 411-420 URL: https://link.springer.com/chapter/10.1007/978-3-030-90525-5_35 [doi: 10.1007/978-3-030-90525-5_35]
8. van Santen J, Dröes RM, Holstege M, Henkemans OB, van Rijn A, de Vries R, et al. Effects of exergaming in people with dementia: results of a systematic literature review. J Alzheimers Dis 2018;63(2):741-760 [FREE Full text] [doi: 10.3233/JAD-170667] [Medline: 29689716]
9. Wiseman L, Isbel S, Boag A, Halpin-Healy C, Gibson D, Bail K, et al. Online gallery facilitated art activities for people with dementia during the COVID-19 pandemic and beyond: a narrative review. Dementia (London) 2023 Nov;22(8):1950-1976 [FREE Full text] [doi: 10.1177/14713012231198748] [Medline: 37647250]
10. Budak KB, Atefi G, Hoel V, Laporte Uribe F, Meiland F, Teupen S, et al. Can technology impact loneliness in dementia? A scoping review on the role of assistive technologies in delivering psychosocial interventions in long-term care. Disabil Rehabil Assist Technol 2023 Oct;18(7):1107-1119 [FREE Full text] [doi: 10.1080/17483107.2021.1984594] [Medline: 34752177]
11. Tan JR, Boersma P, Ettema TP, Planting CH, Clark S, Gobbens RJ, et al. The effects of psychosocial interventions using generic photos on social interaction, mood and quality of life of persons with dementia: a systematic review. BMC Geriatr 2023 Sep 14;23(1):560 [FREE Full text] [doi: 10.1186/s12877-023-04270-w] [Medline: 37710147]
12. Lorenz K, Freddolino PP, Comas-Herrera A, Knapp M, Damant J. Technology-based tools and services for people with dementia and carers: mapping technology onto the dementia care pathway. Dementia (London) 2019 Feb;18(2):725-741. [doi: 10.1177/1471301217691617] [Medline: 28178858]
13. Tan JR, Boersma P, Ettema TP, Aëgerter L, Gobbens R, Stek ML, et al. Known in the nursing home: development and evaluation of a digital person-centered artistic photo-activity intervention to promote social interaction between residents with dementia, and their formal and informal carers. BMC Geriatr 2022 Jan 06;22(1):25 [FREE Full text] [doi: 10.1186/s12877-021-02632-w] [Medline: 34991472]
14. Lawrence V, Fossey J, Ballard C, Moniz-Cook E, Murray J. Improving quality of life for people with dementia in care homes: making psychosocial interventions work. Br J Psychiatry 2012 Nov;201(5):344-351. [doi: 10.1192/bjp.bp.111.101402] [Medline: 23118034]
15. Darker CD, Nicolson GH, Carroll A, Barry JM. The barriers and facilitators to the implementation of National Clinical Programmes in Ireland: using the MRC framework for process evaluations. BMC Health Serv Res 2018 Sep 24;18(1):733 [FREE Full text] [doi: 10.1186/s12913-018-3543-6] [Medline: 30249262]
16. Campbell M, Fitzpatrick R, Haines A, Kinmonth AL, Sandercock P, Spiegelhalter D, et al. Framework for design and evaluation of complex interventions to improve health. BMJ 2000 Sep 16;321(7262):694-696 [FREE Full text] [doi: 10.1136/bmj.321.7262.694] [Medline: 10987780]
17. Moore GF, Audrey S, Barker M, Bond L, Bonell C, Hardeman W, et al. Process evaluation of complex interventions: medical research council guidance. BMJ 2015 Mar 19;350:h1258 [FREE Full text] [doi: 10.1136/bmj.h1258] [Medline: 25791983]
18. Neal DP, Kuiper L, Pistone D, Osinga C, Nijland S, Ettema T, et al. FindMyApps eHealth intervention improves quality, not quantity, of home tablet use by people with dementia. Front Med (Lausanne) 2023;10:1152077 [FREE Full text] [doi: 10.3389/fmed.2023.1152077] [Medline: 37324141]
19. Mahmoudi Asl A, Kouters S, Castro-González Á, Van der Roest H, Franco Martin M, Dröes RM. Potential facilitators of and barriers to implementing the MINI robot in community-based meeting centers for people with dementia and their carers in the Netherlands and Spain: explorative qualitative study. J Med Internet Res 2023 Aug 02;25:e44125 [FREE Full text] [doi: 10.2196/44125] [Medline: 37531190]
20. Beentjes KM, Kerkhof YJ, Neal DP, Ettema TP, Koppelle MA, Meiland FJ, et al. Process evaluation of the FindMyApps program trial among people with dementia or MCI and their caregivers based on the MRC guidance. Gerontechnology 2020 Aug 31;20(1):1-15. [doi: 10.4017/gt.2020.20.1.406.11]
21. van Santen J, Dröes RM, Blanson Henkemans O, Schoone M, Valk R, Straten AV, et al. Implementation of exergaming for people with dementia: facilitators, barriers, and recommendations. Aging Ment Health 2024;28(2):244-253. [doi: 10.1080/13607863.2023.2238259] [Medline: 37493266]
22. Seen as a person; evaluation of a person-centred tablet intervention to promote the social interaction between people with dementia, family and professional caregivers. Overview of Medical Research in the Netherlands (OMON; formerly Dutch Trial Register).. URL: https://onderzoekmetmensen.nl/en/trial/25910 [accessed 2024-04-29]
23. Reisberg B, Ferris SH, de Leon MJ, Crook T. The Global Deterioration Scale for assessment of primary degenerative dementia. Am J Psychiatry 1982 Sep;139(9):1136-1139. [doi: 10.1176/ajp.139.9.1136] [Medline: 7114305]

XSL•FO
RenderX

24. O'Cathain A, Murphy E, Nicholl J. Why, and how, mixed methods research is undertaken in health services research in England: a mixed methods study. BMC Health Serv Res 2007 Jun 14;7:85 [FREE Full text] [doi: 10.1186/1472-6963-7-85] [Medline: 17570838]
25. Goodman C, Baron NL, Machen I, Stevenson E, Evans C, Davies SL, et al. Culture, consent, costs and care homes: enabling older people with dementia to participate in research. Aging Ment Health 2011 May;15(4):475-481. [doi: 10.1080/13607863.2010.543659] [Medline: 21500014]
26. O'Cathain A, Murphy E, Nicholl J. The quality of mixed methods studies in health services research. J Health Serv Res Policy 2008 Apr;13(2):92-98. [doi: 10.1258/jhsrp.2007.007074] [Medline: 18416914]
27. Lutomski JE, Baars MA, Schalk BW, Boter H, Buurman BM, den Elzen WP, TOPICS-MDS Consortium. The development of the Older Persons and Informal Caregivers Survey Minimum DataSet (TOPICS-MDS): a large-scale data sharing initiative. PLoS One 2013;8(12):e81673 [FREE Full text] [doi: 10.1371/journal.pone.0081673] [Medline: 24324716]
28. Gerritsen DL, Oude Voshaar RC. The effects of the COVID-19 virus on mental healthcare for older people in The Netherlands. Int Psychogeriatr 2020 Nov;32(11):1353-1356 [FREE Full text] [doi: 10.1017/S1041610220001040] [Medline: 32491980]
29. Kitwood T. Dementia Reconsidered: The Person Comes First. New York, NY: Open University Press; 1997.
30. Rosenberg F. The MoMA Alzheimer's project: programming and resources for making art accessible to people with Alzheimer's disease and their caregivers. Art Health 2009 Mar;1(1):93-97. [doi: 10.1080/17533010802528108]
31. Data retention-analytics help. Google. URL: https://support.google.com/analytics/answer/7667196?hl=en [accessed 2024-04-29]
32. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
33. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
34. DeepL Translate: the world's most accurate translator. DeepL. URL: https://www.deepl.com/en/translator [accessed 2023-05-12]
35. Gibson G, Chalfont GE, Clarke PD, Torrington JM, Sixsmith AJ. Housing and connection to nature for people with dementia: findings from the INDEPENDENT project. J Hous Elderly 2007 Dec 17;21(1-2):55-72. [doi: 10.1300/j081v21n01_04]
36. Van der Roest HG, Prins M, van der Velden C, Steinmetz S, Stolte E, van Tilburg TG, et al. The impact of COVID-19 measures on well-being of older long-term care facility residents in the Netherlands. J Am Med Dir Assoc 2020 Nov;21(11):1569-1570 [FREE Full text] [doi: 10.1016/j.jamda.2020.09.007] [Medline: 33036911]
37. Tyack C, Camic PM, Heron MJ, Hulbert S. Viewing art on a tablet computer: a well-being intervention for people with dementia and their caregivers. J Appl Gerontol 2017 Jul;36(7):864-894. [doi: 10.1177/0733464815617287] [Medline: 26675353]
38. Leuty V, Boger J, Young L, Hoey J, Mihailidis A. Engaging older adults with dementia in creative occupations using artificially intelligent assistive technology. Assist Technol 2013;25(2):72-79. [doi: 10.1080/10400435.2012.715113] [Medline: 23923689]
39. Foley KL, Sudha S, Sloane PD, Gold DT. Staff perceptions of successful management of severe behavioral problems in dementia special care units. Dementia 2003 Feb 01;2(1):105-124. [doi: 10.1177/1471301203002001998]
40. Kuske B, Hanns S, Luck T, Angermeyer MC, Behrens J, Riedel-Heller SG. Nursing home staff training in dementia care: a systematic review of evaluated programs. Int Psychogeriatr 2007 Oct;19(5):818-841 [FREE Full text] [doi: 10.1017/S1041610206004352] [Medline: 17052376]
41. Hui EK, Tischler V, Wong GH, Lau WY, Spector A. Systematic review of the current psychosocial interventions for people with moderate to severe dementia. Int J Geriatr Psychiatry 2021 Sep;36(9):1313-1329. [doi: 10.1002/gps.5554] [Medline: 34350626]
42. Astell AJ, Ellis MP, Alm N, Dye R, Gowans G. Stimulating people with dementia to reminisce using personal and generic photographs. Int J Comput Healthc 2010;1(2):177. [doi: 10.1504/ijcih.2010.037461]
43. Lee AC, Tang SW, Yu GK, Cheung RT. The smiley as a simple screening tool for depression after stroke: a preliminary study. Int J Nurs Stud 2008 Jul;45(7):1081-1089. [doi: 10.1016/j.ijnurstu.2007.05.008] [Medline: 17707824]
44. Vernooij-Dassen M, Vasse E, Zuidema S, Cohen-Mansfield J, Moyle W. Psychosocial interventions for dementia patients in long-term care. Int Psychogeriatr 2010 Nov;22(7):1121-1128 [FREE Full text] [doi: 10.1017/S1041610210001365] [Medline: 20813074]
45. Norman D. Emotion and design: attractive things work better. interactions 2002 Jul;9(4):36-42. [doi: 10.1145/543434.543435]
46. Gowans G, Dye R, Alm N, Vaughan P, Astell A, Ellis M. Designing the interface between dementia patients, caregivers and computer-based intervention. Des J 2015 Apr 28;10(1):12-23. [doi: 10.2752/146069207789318018]
47. de Vugt M, Dröes RM. Social health in dementia. Towards a positive dementia discourse. Aging Ment Health 2017 Jan;21(1):1-3. [doi: 10.1080/13607863.2016.1262822] [Medline: 28098498]
48. Aylward S, Stolee P, Keat N, Johncox V. Effectiveness of continuing education in long-term care: a literature review. Gerontologist 2003 Apr;43(2):259-271. [doi: 10.1093/geront/43.2.259] [Medline: 12677083]

XSL•FO
RenderX

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Research
GDS: Global Deterioration Scale
GRAMMS: Good Reporting of a Mixed Methods Study
MRC: Medical Research Council
RCT: randomized controlled trial
RQ: research question
SFAS: Smiley Face Assessment Scale

Edited by A Mavragani; submitted 21.01.24; peer-reviewed by G Kohl, S Fox; comments to author 02.12.24; revised version received 23.01.25; accepted 02.03.25; published 16.04.25.

Please cite as:
Tan JRO, Neal DP, Vilmen M, Boersma P, Ettema TP, Gobbens RJJ, Sikkes SAM, Dröes RM
A Digital Photo Activity Intervention for Nursing Home Residents With Dementia and Their Carers: Mixed Methods Process Evaluation
JMIR Form Res 2025;9:e56586
URL: https://formative.jmir.org/2025/1/e56586
doi:10.2196/56586
PMID:

©Josephine Rose Orejana Tan, David P Neal, Maria Vilmen, Petra Boersma, Teake P Ettema, Robbert J J Gobbens, Sietske A M Sikkes, Rose-Marie Dröes. Originally published in JMIR Formative Research (https://formative.jmir.org), 16.04.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Creating a Usable and Effective Digital Intervention to Support Men to Test for HIV and Link to Care in A Resource-Constrained Setting: Iterative Design Based on A Person-Based Approach and Human Computer Interaction Methods

Anya Zeitlin[1*], MSc; Thulile Mathenjwa[2,3*], MA; Thembelihle Zuma[2,4], PhD; Sally Wyke[5], PhD; Philippa Matthews[2], BMBS; Nuala McGrath[2,4,6], PhD; Janet Seeley[2,7], PhD; Maryam Shahmanesh[2,8], PhD; Ann Blandford[1], PhD

[1]UCL Interaction Centre, University College London, London, United Kingdom
[2]Africa Health Research Institute, Durban, South Africa
[3]School of Public Health, University of Cape Town, Cape Town, South Africa
[4]School of Nursing and Public Health, University of KwaZulu Natal, Durban, South Africa
[5]School of Social and Political Sciences, University of Glasgow, Glasgow, United Kingdom
[6]Faculty of Medicine and Faculty of Social Sciences, University of Southampton, Southampton, United Kingdom
[7]London School of Hygiene & Tropical Medicine, London, United Kingdom
[8]Institute for Global Health, University College London, London, United Kingdom
[*]these authors contributed equally

Corresponding Author:
Thulile Mathenjwa, MA
Africa Health Research Institute
Krith Building, 3rd floor
719 Umbilo Rd
Durban, 4001
South Africa
Phone: 27 315210455
Email: thulile.mathenjwa@ahri.org

Abstract

Background: It is challenging to design usable and effective digital health interventions (DHIs). The person-based approach (PBA) has been proposed to incorporate users' perspectives for the design of DHIs. However, it does not explicitly describe the iterative stages of design and evaluation that are essential in moving from early planning to deployment. For this, we draw on methods from human computer interaction (HCI) that have been developed for various situations.

Objective: This study aimed to reflect on the adaptation and synthesis of PBA and HCI approaches to developing DHIs. We present a case study applying both approaches to develop Empowering People through Informed Choices for HIV (EPIC-HIV1), a DHI designed for men living in rural KwaZulu-Natal, South Africa, intended to support them in making an informed choice about whether to take an HIV test and, if necessary, engage in care.

Methods: We conducted a retrospective analysis of the documentation generated during the development of EPIC-HIV1 including findings about requirements, design representations, and the results of iterative rounds of testing. We developed an account of the process, the outcomes, and the strengths and limitations of the design and evaluation techniques applied. We also present the design of EPIC-HIV1 and summarize considerations when designing for hard-to-reach people in such settings.

Results: The PBA was applied to deliver a first prototype. This helped identify key messages to convey and how to manage issues such as user privacy, but the resulting prototype was judged by the team not to be engaging for potential users, and it was unclear whether the design was inclusive of people with low digital or health literacy. We therefore introduced methods from HCI to iteratively test and refine the app. Working with local community representatives, we conducted four refinement cycles with 29 participants, adapting and retesting the app until no further changes were needed. Key changes included making it clearer what the consequences of selecting options in the app were and changing wording to minimize misconceptions (eg, that the app would test for HIV) while addressing common concerns about testing and emphasizing long-term benefits of engaging with care, if needed.

XSL•FO
RenderX

Conclusions: Techniques for developing DHIs need to be situationally appropriate. The PBA enabled us to establish both empirical data and theory to design the content of EPIC-HIV1, but it did not directly inform interaction design to make the app usable and effective for the intended users; HCI techniques tailored to the setting enabled us to refine the app to be easy for men with little familiarity with digital technologies to use within the constraints of the setting. Iterative testing ensured the app was easy to use and that the intended clinical messages were communicated effectively.

(JMIR Form Res 2025;9:e65185) doi:10.2196/65185

KEYWORDS

person-based approach; HIV; resource-constrained settings; digital intervention; user-centered design; behavior change techniques; digital health

Introduction

Background

The focus of this paper is on the application of methods for developing usable, effective, and engaging interactive digital health interventions (DHIs), taking as a case study an intervention, Empowering People through Informed Choices for HIV (EPIC-HIV1), intended to enable men to make an informed choice about testing for HIV and, if necessary, engaging with care. Recent literature shows that men are left behind in the response to HIV [1] as evidenced by the low uptake of HIV testing, prevention and treatment compared with women [2-4].

Challenges of Developing Interactive DHIs

As digital technologies become more affordable, their potential is increasingly leveraged for DHIs in a broad range of settings including both affluent and resource-constrained communities, in urban, rural and remote settings [5]. There are numerous challenges with implementing DHIs such as ensuring interventions keep pace with rapidly changing technical systems [6], supporting usability needs [7], ensuring that the intervention is accessible across devices and operating systems [8], and making it culturally appropriate [9]. These challenges are exacerbated in resource-constrained settings where there are variable infrastructural constraints and levels of familiarity with technology within intervention cohorts.

There are few accounts of how to develop usable, engaging, and effective DHIs. One of the most widely cited is the person-based approach (PBA) to intervention development [10]. Although this approach describes the need to iteratively test prototypes with end users, it omits detailed discussion of interaction design and how to integrate this into the intervention development process. Conversely, the human computer interaction (HCI) literature rarely addresses issues that matter to population health researchers such as content design, clinical effectiveness, scalability, and sustainability [11]. The aim of the work reported here was to reflect on the integration of PBA and HCI methods taking as a case study the design and deployment of EPIC-HIV1, a tablet-based app.

Apps are particularly suited for stigmatized health conditions like HIV because they provide users with privacy and anonymity [12,13]. In addition, they offer health interventionists the ability to deliver uniform messages, free from embarrassment which affects people's ability to communicate consistent and accurate information [13]. Several studies have used digital technology to increase uptake of HIV testing and medication adherence [14,15]. These studies found that both mobile and tablet-based HIV interventions are feasible, acceptable, and effective methods to engage hard to reach populations [16]. However, the studies do not discuss how the interventions were designed or developed.

Interdisciplinary Approaches to Developing Digital Health Interventions

The World Health Organization advocates the application of user-centered design to develop DHIs to ensure that they are effective, accessible, acceptable, and user-friendly [17]. This often requires the collaboration of professionals from health and HCI backgrounds. Health care professionals, including epidemiologists and social scientists, contribute clinical expertise, theoretical understandings of behavior change and health outcome evaluation techniques. HCI professionals contribute user-centered design methodologies, which are applied to understand user needs, inform the design of digital technologies and evaluate them in terms of usability and user experience [11,18,19].

Despite the clear benefits of interdisciplinary collaboration, there are well-recognized challenges [11,18] relating to the academic heritage of the respective disciplines. Professionals from health care backgrounds put greater emphasis on measuring the effectiveness of DHIs and evaluating outcomes. HCI professionals focus on iterative design, including alternative design representations and formative evaluation, to ensure digital technology is usable and acceptable. As a result, interdisciplinary teams experience divergent approaches to timelines, measuring effectiveness and understanding success.

Literature in the field of behavior change tends to limit the integration of HCI methods, and therefore collaboration, to discrete stages in the intervention development process: for example, evaluating usability and engagement once a prototype has been developed [20].

In HCI literature, there is more interest in exploring the boundaries of the role HCI should play in DHI design and evaluation, rather than identifying discrete stages in the development process in which HCI approaches can be applied. Klasnja et al [21] describe the contribution of HCI as being to understand "how and why" DHIs work, rather than measuring outcomes. In contrast, Smith et al [22] advocate using a value chain analysis to evaluate how short-term (proximal) metrics relate to long term (distal) behavior change and health outcomes.

XSL•FO
RenderX

Marcu et al [23] describe the Patient-Clinician-Designer framework, which seeks to integrate multiple perspectives throughout the design process. In understanding the clinical context, patient needs and technical constraints, they ensure that digital interventions meet the needs of all stakeholders and are therefore more likely to be successful when implemented [23]. They describe this in reference to a mental health application, but it has also been applied to the development of applications for stigmatized conditions including HIV [24]. Similarly, Blandford [11] highlights the value of applying HCI methodologies throughout intervention development to facilitate more systematic trade-offs between what is needed for users and clinicians and what is feasible in the context.

Person-Based Approach and Integration With Human Computer Interaction Methods

We used the PBA to inform the design and development of EPIC-HIV1. The PBA [10,20] provides guidelines on how to use qualitative research methods to identify psychosocial factors that influence the effectiveness and acceptability of interventions. Using rich, qualitative research at every stage of the intervention development process facilitates an investigation of the beliefs, attitudes, needs and situations of participants in the intervention. Yardley et al [10] propose that using PBA can make an intervention more relevant, persuasive, accessible and engaging.

The PBA consists of 4 stages, namely planning, design, development, and intervention [20]. While Yardley et al [20] recognize the value of gathering user input, the focus is on how that input informs the development of content and intervention goals. They do not describe the HCI, or user-centered design, methods that should be applied or how to translate the intervention goals and content into a usable and engaging digital product.

Blandford [25] proposes an integration of PBA with established HCI design methods to address these issues (Figure 1). The key extensions to PBA are earlier consideration of design possibilities, the identification of design representations (eg, design patterns and task structures) and iteration between as well as within stages.

Figure 1. An integrated development lifecycle illustrating how a person-based approach to digital health intervention design can be augmented by human computer interaction methods.

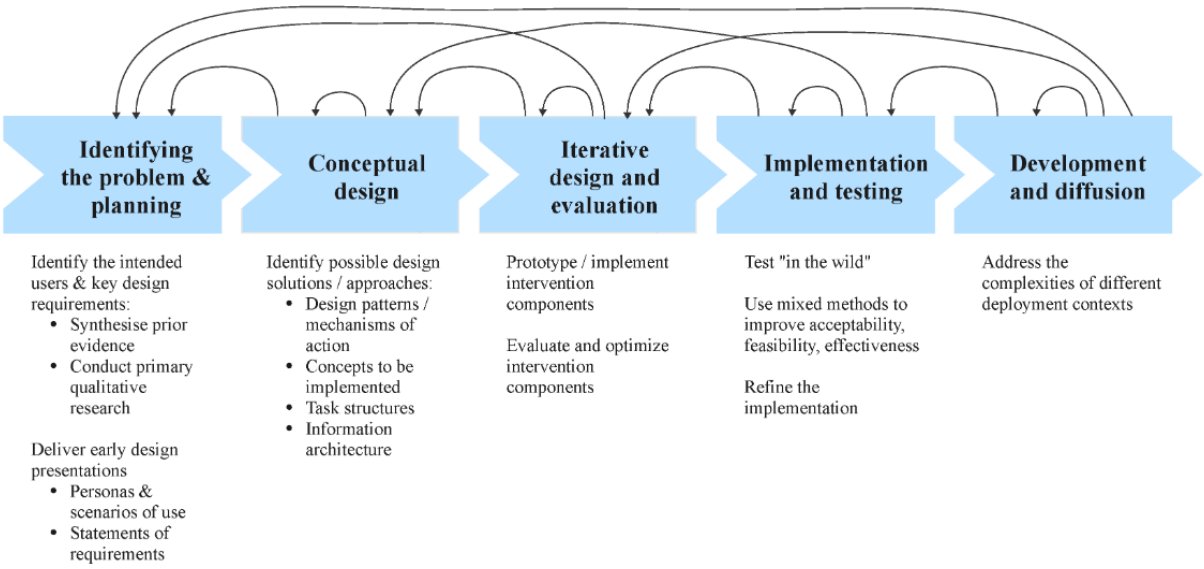

Both approaches advocate the use of personas and scenarios. These are rich descriptions of the intended users of the intervention and of how they will use the intervention. Scenarios are commonly presented at different levels of abstraction; for example, Rosson and Carroll [26] describe problem scenarios (the broad situation), information scenarios (giving details of the information that will be provided to users) and interaction scenarios (describing the details of user interaction with the system).

The aim of this study is to report and reflect on the process and outcomes of applying the PBA and selected HCI methods to the development of EPIC-HIV1 as an exemplar DHI for behavior change.

Methods

Overview

This paper draws on documentation that was created during the iterative development process, including the initial prototyping of EPIC-HIV1 and the subsequent cycles of refinement and testing. That documentation was systematically gathered throughout the development process to support subsequent analysis. The retrospective analysis of the documentation involved the construction of a narrative of the processes and outcomes for the planning, design, and developing stages outlined in Figure 2 and presented below (after an overview of the context and rationale). Therefore, this narrative is presented as “results” below, rather than “method”. The implementation stage (ie, deployment in a clinical trial) is beyond the scope of

XSL•FO
RenderX

this paper and is described elsewhere [4,27,28]. We present the context, rationale, and design aims for EPIC-HIV1.

Figure 2. Methods used in the first 3 phases of the person-based approach, phase 4 (implementation) is out of scope for this paper. AHRI: Africa Health Research Institute; BCT: behavior change taxonomy; EPIC-HIV1: Empowering People through Informed Choices for HIV; SDT: self-determination theory.

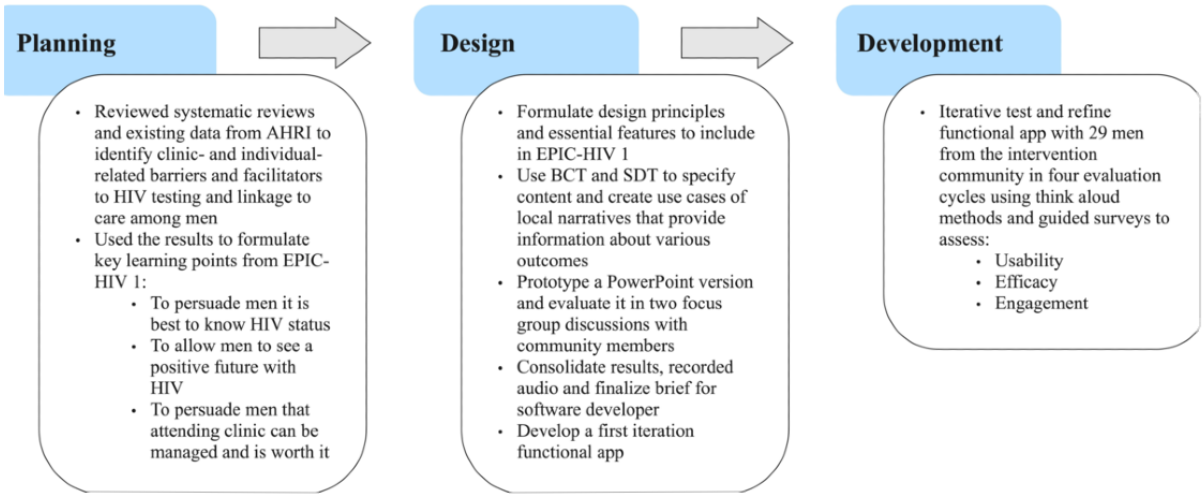

Empowering People Through Informed Choices for HIV Context and Rationale

EPIC-HIV1 is 1 of 2 DHIs that make up the EPIC-HIV intervention. The second DHI, EPIC-HIV2, was designed to support engagement in care for those who did not link to care within a month after a positive HIV diagnosis and is described elsewhere [29]. EPIC-HIV is 1 of 2 interventions in the Home-based Intervention to Test and Start (HITS) trial that sought to increase HIV testing and linkage to care among men in rural South Africa using micro incentives and male-targeted decision support [4,30]. The HITS trial aimed to compare the impact of micro incentives (financial) with decision support (EPIC-HIV). It was implemented in a rural setting in uMkhanyakude district in northern KwaZulu-Natal, South Africa using the Africa Health Research Institute (AHRI) Demographic and HIV surveillance platform [31]. The trial results are reported elsewhere [4,27,28].

For this case study, we focus on the development of EPIC-HIV1. As noted above, it is an educational app that introduces men to accurate information about HIV and encourages them to make an informed choice about engaging with HIV testing and care. The design drew on self-determination theory (SDT) [32] to plan content by supporting three basic psychological needs (autonomy, competence and relatedness) and increase men's internal motivation to test for HIV and link to care where necessary. We also drew on the behavior change taxonomy (BCTv1) proposed by Michie et al [33]. They list 93 techniques that can be applied in behavior change interventions, clustered into 16 groups (eg, goals and planning, feedback and monitoring, and social support, providing a credible source of information and advice). Many of these are potentially relevant to the design of digital interventions: for example, goal setting, providing instructions on how to perform a behavior, or providing a credible source.

EPIC-HIV1 was designed to be delivered by fieldworkers during the annual household visits which includes conducting household and individual behavior and health related surveys and collecting dry blood spots for anonymized HIV testing as well as the offer of a rapid HIV test [31]. The concept was that the app would be deployed on the fieldworker tablet and offered to men during the data collection for individual surveys where the fieldworker is alone with the participant. The app is offered before the HIV rapid test offer. The fieldwork visits are carried out in a systematic cycle where they are allocated a number of households to cover in a week. Due to this time constraint, user interaction with EPIC-HIV1 was allocated 10 minutes.

Design Aims for Empowering People Through Informed Choices for HIV

Two key aims for EPIC-HIV1 were ensuring that it met users' needs, that is, that participants were engaged by the content and could autonomously navigate the application, including people with low digital and health literacy; and ensuring that it met the intervention goals, namely to communicate the correct clinical messages to men and enable them to make an informed choice about engaging with HIV care.

The design had to recognize the stigma related to HIV [15,16,34]. This relates to both the design of the intervention itself, and the evaluation methods used. For example, in user studies we did not require participants to disclose their HIV status. Also, to enable study participants and future users to explore EPIC-HIV1 independently and privately, earphones were offered for their use.

Another design challenge in resource-constrained settings is low literacy [35]. Solutions proposed for low-literacy users include graphical interfaces, interactive voice interfaces, and mediated usage [36]. Designing for low literacy populations extends beyond catering for users not being able to read, to multiple cognitive and emotional factors that influence how users engage with content [37,38]. For health care interventions, there is added complexity: individuals with lower literacy tend to have lower health literacy [39], and having lower health literacy puts an individual at higher clinical risk and is associated

XSL•FO
RenderX

with reduced ability to exert control over health through informed decision making [39,40]. The approaches taken in both design and testing to be inclusive of users with low digital and health literacy are discussed below.

Ethical Considerations

This study received ethical approval from the University of KwaZulu-Natal Biomedical Research Ethics Committee (BFC398/16). Participants involved in empirical data collection provided written or verbally informed consent before participation. Participation in the study was entirely voluntary. To protect participants' privacy and confidentiality all data were deidentified, and no personally identifiable information was included in the final dataset. Participants received light refreshments (sandwich and juice) for taking part in the study and no financial rewards were provided.

Results

Overview

As noted above, EPIC-HIV1 was developed following the first 3 stages of PBA. The planning and design stages were led by social scientists and health professionals (JS, SW, MS, TZ, PM, and TM) and the development phase was led by HCI specialists (AB and AZ) and the technology partner. Each stage used different recruitment and sampling methods as described in detail below.

Person-Based Approach Stage 1: Planning

The planning stage is described in detail by Mathenjwa et al [29]. In summary, we reviewed systematic reviews and existing data from AHRI to identify barriers to HIV testing and linkage to care and used the results to define the user requirements of the app. In summary, key barriers to testing were fear of rejection, and being blamed for sexual misconduct, if found to be HIV positive; fear of death or illness; feeling healthy and so not understanding the need to test; a preference for traditional medicine; and a belief that medicine should be curative (rather than for preventing or managing illness). Key barriers to engaging with care were that clinics were regarded as feminized spaces where men do not feel comfortable; that they are concerned about confidentiality if they are seen in clinics; that there may be long wait times in clinics; and that they often have poor relationships with health care workers. Key learning points for EPIC-HIV1 were that it was important to persuade men that it is best to know their HIV status, because it allows them to take control; that men needed to see a positive future with HIV, in which they are in control and can achieve their ambitions, and that attending clinic can be managed and is valuable in the long term.

Person-Based Approach Stage 2: Design

The aim of this stage was to formulate the design principles of the application and essential features that should be included. Using the results from the planning stage, we identified barriers that can be modified by EPIC-HIV1 and key learning points to be incorporated into the app. We then formulated intervention features relevant to the learning points. Finally, we identified specific behavior techniques from the BCT taxonomy [33] to support autonomy, competence and relatedness. The components are interrelated as illustrated in Table 1.

Table 1. Key learning points from the planning stage of person-based approach and how they relate to self-determination theory and behavior change taxonomy.

Learning point	SDT[a] relevant	BCT[b] from taxonomy
To persuade men it is best to know your HIV status, because it allows you to take control	• Persuading users that knowing their status allows them to take control of important aspects of their lives, such as upholding traditional values, caring for a family or having sex and children (supporting autonomy).	9.1 credible source 5.1 information about health consequences 5.2 salience of consequences
To allow or engage men to see a positive future with HIV, in which they are in control and can achieve their ambitions	• Maintaining a positive illness context throughout: that it is possible to live with HIV and do all one wants to do. • Educating men about HIV and the benefits of testing and taking antiretroviral therapy (supporting autonomy).	9.1 credible source 5.1 information about health consequences 5.2 salience of consequences 13.5 identity associated with changed behavior 13.1 identification of self as role model
To persuade men that attending the clinic can be managed and is worth it in the long term	• Persuade men that they can handle going to the clinic through positive examples of men talking about what it was like and how they managed it (supporting relatedness, competency and autonomy).	13.5 identity associated with changed behavior 13.1 identification of self as role model

[a]SDT: self-determination theory.

[b]BCT: behavior change taxonomy.

The Conceptual Design and Scenarios of Use

The high-level design concept was to provide experiential information from local men living with HIV that users could identify with. The content was rooted in local narratives to provide information about various outcomes to increase risk perception, salience, and likelihood of response by making the decision to test or link to care explicit. Personas were implicitly encapsulated in the descriptions of the characters who were to be introduced through EPIC-HIV1: men of different ages and with different health beliefs and levels of digital literacy.

XSL•FO
RenderX

As described above, the envisaged context of use for EPIC-HIV1 was that it would be administered by the AHRI fieldworker during the annual individual biobehavioral surveillance home visit. When the time came to explore EPIC-HIV1, the fieldworker was expected to hand over the tablet computer with the app opened (start screen) to the user (consenting man), together with earphones (that the man could keep) and wait while they interacted with the app in private. The app was designed to take 5-10 minutes depending on the pathway that the user selected to align with the time constraints of the fieldwork.

The information scenario was based on the notion that the app should acknowledge and refute common misconceptions about HIV (how it is transmitted and what treatments are effective) and concerns about attending a clinic and making other people aware of their HIV status.

The interaction scenario involved the user working through 3 sections (illustrated in Multimedia Appendix 1. First, there was a short introduction, after which the user was given the choice to either find out about HIV testing and treatment (if they were not ready to test) or just find out about treatment (if they were ready to test). In the testing section, users were given the option to listen to short vignettes describing common reasons for men not wanting to test for HIV. Users were then presented with information, addressing each of these reasons and providing a counter argument. In the treatment section, 3 characters described their journey since being diagnosed with HIV; these characters represented different personas, and the user could listen to one of their stories (ideally the one they found most relatable).

All app content and instructions were provided in text and audio format. Still images, including photographs of individuals (with faces obscured) and artefacts from the intervention community, supported the audio descriptions.

The app used a male nurse from AHRI to serve as a guide character and credible source of HIV information. He gives an introduction and explains how the app works and then reads out the 2 options ("I am sure I will have an HIV test today" and "I might not have an HIV test today") and asks the user to select which statement he agrees with. In the "ready to test" section, the nurse applauds the user for taking the decision to test today. In the "not ready to test" section, we used 7 characters to give short vignettes discussing common reasons for men not wanting to test for HIV which were identified during the planning stage. The nurse comes back to address the reason each character lists, providing a counter argument. Users can then go into more detail by selecting one of the 3 main characters that persuade men that knowing their HIV status gives them control of important aspects of life, that it is possible to live with HIV and still achieve goals and finally that they can manage going to the clinic (Figure 3).

Figure 3. Illustration of 7 characters from the PowerPoint prototype of Empowering People through Informed Choices for HIV.

Testing the Conceptual Design

We developed a low fidelity PowerPoint (Microsoft Corp) protype of the app and evaluated it with 2 focus groups of men from the community advisory board, a body representing members of the intervention community that acts as a bridge between AHRI and the community, safeguarding the rights of the study participants. We sought to include younger and older community advisory board members. The focus of the evaluation was on whether the characters and messages selected were acceptable to men in our community. Since we aimed to assess acceptability of the content and characters, findings were categorized into content or character related and the app functionality (Table 2).

XSL•FO
RenderX

Table 2. Summary of the focus group discussion results and how we addressed them.

Main issues	Issues	Steps taken and revision of content
Content	• Self-stigmatization after testing HIV-positive. • Fear of weakness because of illness. • Manhood is about becoming isoka (Casanova). • Testing is for everyone, not for particular groups or types. • Most men want to wait until they get sick before testing, long queues in clinics, feminized spaces of health care.	• Positive reinforcement on autonomy, competence and relatedness among all characters.
Characters	• Groups preferred characters to look like strong men. • Groups commented on what characters should wear. • Groups mentioned that characters' gestures should be open and clear.	• Only the nurse character will be seen by users. • No other faces will be shown, real settings will not be used. • Men wearing what was proposed in groups. • Characters' gestures open and clear.
Functionality	• Asked for a rewind button so that people can go back and listen. • Privacy and confidentiality.	• No rewind option because of time. • App will be private, and fieldworker will clarify to each participant that they will not know how the participant navigated the app.

We used the results to refine the content and characters and finalized the brief for the software developer. We selected a cast of actors from the local community to represent the different characters and did an audio recording of voice overs and a photo shoot at AHRI offices and a local community center. This was sent to the software developer with the brief. We received the first iteration of a functional app, as illustrated in Figure 4; however, it was not considered by the team to be engaging.

Figure 4. First iteration of the Empowering People through Informed Choices for HIV app cycle (3 example screen shots).

Screen 1 - Introduction: Nurse introduces the app, users receive instructions and choose a route through the app
Screen 2 - Testing Section: User hears common reasons why men do not want to test for HIV and nurse's rebuttal
Screen 3 - Treatment Section: User hears stories of men since their HIV diagnosis

Person-Based Approach Stage 3: Iterative Testing and Design Changes

We iteratively tested and refined EPIC-HIV1 to ensure that the app was interactive, usable and acceptable for a broad spectrum of users with varying education and digital literacy. In total, 4 cycles of user-centered evaluation and design were conducted, based on design sprints each lasting 2 weeks. These cycles were structured using a simple interaction design model [41], whereby initial user requirements were identified, design improvements were generated, and the new design was evaluated. The evaluation metrics are summarized in Table 3.

XSL•FO
RenderX

Table 3. Summary of evaluation metrics relating to usability, efficacy, and engagement.

Metric	Questions	Method
Usability	• Is it clear how to move between screens? • Is it clear how to engage with the information on each screen (ie, choose options, listen to audio and exit audio)?	Observations during think-aloud sessions
Efficacy	• Are they key messages communicated? • Do participants gain new information because of using the app? • Is the information they gain factually accurate?	Questions in think aloud protocol and guided survey
Engagement	• Is the interaction pleasant? • Is the presentation of content appropriate and interesting?	Questions in think aloud protocol and guided survey

During each cycle, participants were asked to complete defined tasks and articulate their thoughts about EPIC-HIV1 and the task. Both actions and verbalizations were recorded to identify the main usability challenges and build an understanding of the acceptability of content. Participants were paused after the introductory screens and asked questions regarding their initial expectations of the app's content and purpose. After using the app, participants were asked further questions of their understanding of the content and reflections on the design. This guided, verbally administered, survey ensured that we did not exclude low-literacy participants. It consisted of a mix of open and closed questions and was used to gather insight into usability, comprehension, and design preferences. The evaluations were conducted in isiZulu, with a bilingual (isiZulu-English) researcher administering procedures and participants were observed as they used the application.

Recruitment and Participants

All participants were recruited from the AHRI demographic and HIV surveillance community. Participants were approached in public spaces by a local male researcher from the community. These public spaces varied but included parks, taverns, and the roadside. Across the 4 evaluation cycles, sampling became increasingly purposive, focused on more rural areas and older participants, as we sought to recruit participants who represented specific user groups within the intervention cohort, especially those who had been identified as likely to have low technological literacy.

We conducted evaluations with 29 unique isiZulu speaking males (excluding pilot participants) over the 4 iterative cycles. All participants were asked their age to confirm they were aged 18 years or older and able to consent to participate. Recruitment was helped significantly by the presence of local researchers who knew where men in the community "hang out" and were able to explain the purpose of the study. Men were often in groups and many declined to engage, meaning all 29 participants engaged willingly.

Overview of Evaluation Cycles

Across the 4 cycles, the evaluations became more refined, but the objectives and the metrics were consistent. Correspondingly, there was consistency across the protocols. For example, some of the questions (Table 3) were the same and were asked at the same point in the evaluation. The key features of each evaluation cycle are summarized in Table 4 and refer to Multimedia Appendix 2 for example of findings and changes made in each evaluation cycle.

XSL•FO
RenderX

Table 4. Key features of each evaluation cycle.

Evaluation cycle	Objectives	Methods
Evaluation 1	• Primary: identify main usability challenges, gather understanding of target user group's needs, and assess understanding of content. • Secondary: gather insight into potential design improvements.	• Think aloud. • Observation. • Guided survey.
Evaluation 2	• Primary: make layout easy to use for people who are novice tablet users and have low literacy, assess understanding of content. • Secondary: gather insight into preferred look and feel of app.	• Observation. • Retrospective usability questions. • Guided survey.
Evaluation 3	• Primary: test whether users know how to select options, assess understanding of content, make content more engaging. • Secondary: ensure that uninterrupted participants were completing the app in an appropriate time.	• Observation. • Retrospective usability questions. • Guided survey. • Task completion times.
Evaluation 4	• Primary: assess whether the changes made to the application affected users' expectations, such that they were making more informed decisions.	• Observation. • Retrospective usability questions. • Guided survey.

Specifics of Each Evaluation Cycle

For the first evaluation, convenience sampling was used to recruit from the community, working in or close to the research institute. The evaluation took place within the research institute. After interacting with the first version of the app, participants were presented with low fidelity interactive prototypes of different sections of the app and asked which they preferred, which was most relatable, and which was most clear.

In Evaluation 2, we used pop-up research [42], which involves carrying out targeted user evaluations in environments that are familiar and used by the app's intended audience, to simulate some of the environmental constraints and social considerations that would be experienced by fieldworkers when administering the DHI, such as lack of access to the internet and difficulty finding private spaces. This had the potential to reduce participant response biases [37] by making the environment more familiar and less formal than the research institute. Participants were approached in public spaces, such as at bus stops or tuck shops, and evaluations were conducted either in those spaces or close by in the researchers' car (illustrated in Figure 5). These decisions helped to establish a level of mutual trust and collaboration between the researchers and participants. Unlike most accounts of pop-up research, the planned time for each individual's participation was not particularly short: it lasted 20-30 minutes, sometimes more if the participant had low literacy.

Figure 5. Illustrative photos of pop-up evaluation testing in remote locations using an Africa Health Research Institute vehicle.

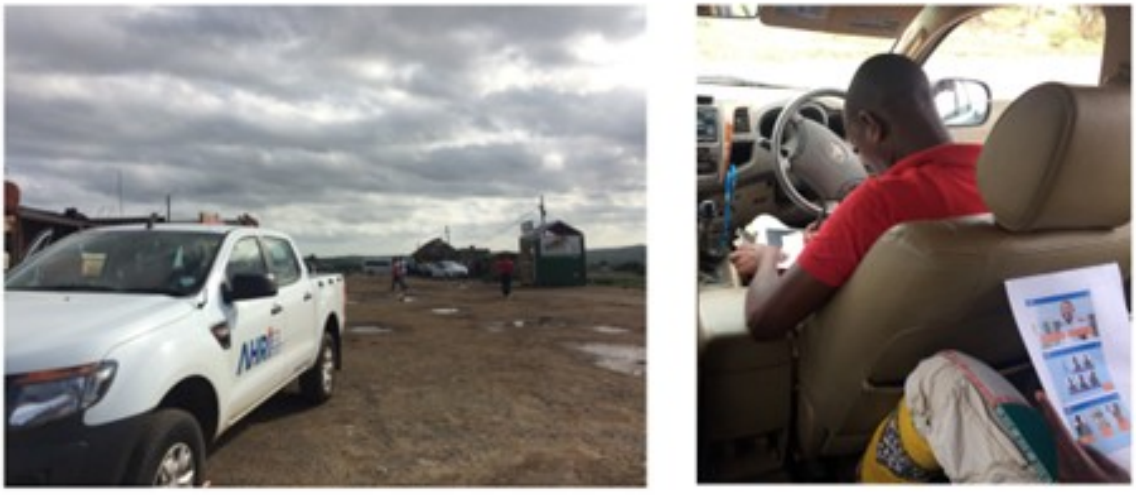

Based on learnings from Evaluation 1, we administered retrospective usability questions rather than asking participants to articulate their thoughts concurrently. We also printed screenshots of design variations on paper and described the differences to participants then asked them which they preferred, rather than continuing to use low-fidelity interactive prototypes

XSL•FO
RenderX

after the first part of each session. We included closed questions that were asked while participants were interacting with EPIC-HIV1: for example, we added a help button; in the evaluation, we stopped users after the instruction screen and asked them to indicate which button to press if it was not clear what to do next. Most participants were unaware they could skip content or be selective, which indicated that the app was not supporting autonomous decision making. When presented with design alternatives, all 6 participants selected the interface with photo images because it was "clear." This supported the use of rich and contextually relevant images throughout the app to optimize the relatability of the content.

Following Evaluation 2, previews were either automated or removed, which meant users would only be presented with choices when it was critical. We added extra instructions on how to select options, made button states more distinct and provided additional nudges in the form of button animations and instructions from the nurse character. Additional photographs of paid professionals who were representative of the intended cohort were added to the character stories to make them more engaging (Figure 6).

Figure 6. Example of design changes made, illustrated with screenshots from the first and third versions of Empowering People through Informed Choices for HIV.

Evaluation 1: User previews character, user selects character again to hear reason for not testing, user selects next, user selects the character again to receive nurse's response to the character's reason (4 clicks).

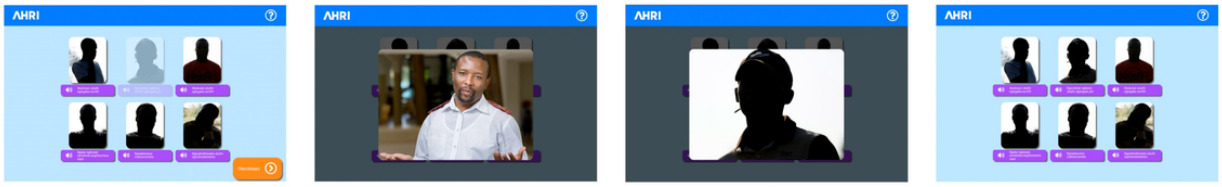

Evaluation 3: User selects character, hears character's reason for not testing, then the nurse's rebuttal (one click).

In Evaluation 3, as well as assessing whether participants knew how to select options, understood correct messages and found the content engaging, we wanted to ensure that participants were completing the app within a given time. We split participants into 2 groups: one was interrupted to assess their option selection while the other was uninterrupted, to gather their task completion time and observe the full user flow. Having identified rural and older people as those who were least comfortable using the app, we actively recruited males representing this category, reducing the proportion of young periurban males (the easiest to find and recruit). After conducting evaluations with 12 participants from various locations, we had reached saturation in responses. Compared with earlier evaluation cycles, we focused more on option selection and expectations. For example, at the beginning participants were asked what they had expected to find if they had selected the other option so that we could assess whether participants were actively choosing what they wanted to hear about. This was particularly important because users could only engage with EPIC-HIV1 once as part of the planned intervention and could not return to content of interest.

Evaluation 4 did not reveal any additional usability issues. Participants were able to describe what they would have had if they had selected the other option, indicating the changes to the language used at the beginning of the app were effective and supported informed decision making.

Summary

Both the evaluation objectives and the methods became more refined during the process. For example, the objective of the first evaluation was to get a sense of what users were finding difficult and what evaluation style would be contextually appropriate. Convenience sampling was used to recruit participants and the questions in the evaluation were relatively general. In the third evaluation, 2 of the objectives were to assess whether participants could make accurate option selections and complete the app within a given time. Purposive sampling was used to recruit participants who represented a range of users we had previously encountered, especially those who had had usability challenges (ie, low literacy, novice users, and older people). The evaluation was more focused, questions were more specific, and the methods were altered to support our objectives.

Although pop-up research was an effective and valuable means of recruiting participants and conducting evaluations, we had to ensure that we adequately addressed considerations, including ethics (ensuring the participants did not feel coerced), privacy (ensuring that participants could freely engage with content without passers-by hearing), community-institution relationships

XSL•FO
RenderX

(avoiding contributing to "participant fatigue" near the research center), and our own safety as we went into remote places and invited people into our car or worked with them in unfamiliar environments.

Discussion

Principal Findings

The analysis of existing theory and data enabled the team to identify and address the most important barriers to men engaging in HIV testing and care in a cost-effective way. The use of HCI methodologies enabled the team to translate the intervention goals, that is, communicating persuasive messages in a way that would promote behavior change, into design strategies and ensure the application was usable. The resulting application was inclusive of the intended population, supported meaningful choice in how users engaged with the application, and ensured the correct clinical messages were communicated.

One key lesson from this study is that it is essential to design the introduction (or "onboarding") to a DHI very carefully, to ensure that users are empowered to use the app effectively and that expectations of what they will experience are set appropriately. Another key lesson is that an appropriate balance had to be achieved between autonomy (enabling people to navigate as they choose) and ensuring that the correct clinical messages are communicated.

The conceptual design featured a series of dialogues between relatable characters and a trustworthy nurse. These dialogues were accompanied by images of the characters in familiar settings. However, this approach was never tested against alternative design strategies. Early decisions about the design and mode of the interaction were based on informal understanding and intuition rather than evidence. Resource limitations (both time and funds) meant that we did not have the capacity to revisit earlier design decisions such as the appropriateness of using a tablet computer or the general approach of presenting dialogues. Thus, it was not possible to engage in the broader cycles of iteration proposed by Blandford [25].

Early design decisions, such as the approach of focusing on a dialogue between relatable characters and a trustworthy nurse, and about the content of the dialogue, were based on the key scenario of use: that is, that the tablet computer would be lent to the individual for up to 10 minutes within the annual health visit; this imposed a significant constraint on the design. For example, the main interaction design issue raised by members of the community advisory board in the Stage 2 testing, that people would like a "rewind" option, was not addressed due to this time constraint.

Through iterative testing the main area of concern, and where many changes were made, were in the introduction, or "onboarding," for the app. It was found necessary to set expectations about the app (eg, that it would not deliver an HIV test result) and to provide detailed verbal instructions about how to select options and press buttons on a touch screen.

We needed to balance the needs for the app to be engaging and inclusive. The participants were diverse, and levels of technical, educational and health literacy varied widely. Some participants were comfortable using the tablet app and owned touch screen smartphones. Testing highlighted the tension between providing literal choice in interaction design and meaningful choice regarding how the user engaged with the intervention. For example, giving users a choice about whether to listen to the nurse's response after hearing a character's concerns proved counterproductive as it did not support the concept of "informed choice," so the choice point between concern and response was removed after the first evaluation cycle.

Changes to the design were intended to support low-literacy users: for example, by adding audio instructions, highlighting one option at a time, and reducing the amount of user input required. However, we maintained critical choice and ensured it was well understood. The reduction in user input was balanced with ensuring that the app remained engaging and persuasive for participants with higher literacy levels. Through iterative cycles, we were able to deliver a design that was inclusive of as much of the target cohort as possible.

Nevertheless, we had to recognize that the delivery of the intervention was likely to exclude a small proportion of the potential beneficiaries. For these people, for whom the app was inaccessible, a conversation with the visiting health worker would be a possible alternative.

Limitations

As noted above, HCI methodologies had not been involved during the early stages of intervention development. As a result, the remit of HCI was to fix the app, to make it as good as possible given the time and resources remaining. We believe the app design could have been stronger if HCI methodologies had been used alongside PBA and BCT approaches from the outset. For example, there was no opportunity for exploratory research to define what interactivity meant for the intended population or understand whether a tablet-based app was an appropriate approach to communicating HIV/AIDS information in this context, or whether it could inadvertently bias participant responses, as indicated by previous research [43], or exclude a proportion of the intended target population.

There was also no opportunity for HCI focused summative research, for example, to understand how users had navigated through the app, areas of interest or potential drop-off points [21]. The HITS clinical trial results [4,27,28] suggest that EPIC-HIV1 did not increase men's uptake of HIV testing. This raises many questions that could be the focus of future research: for example, would a different kind of digital intervention have been more effective, was the approach of fitting the user's interaction with the intervention into the annual visit effective, are digital interventions inappropriate in this context, or does the intervention need to be better tailored toward different subpopulations (eg, those with different levels of digital literacy or different priorities in life)?

Conclusion

The aim in developing EPIC-HIV1 was to support a clinical trial investigating whether financial incentives or a digital

XSL•FO
RenderX

intervention was more effective than the established approach to encouraging men to test for and engage in care for HIV. The implicit assumption was that any digital intervention would be equally effective and no comparison between different digital interventions was planned.

By applying an iterative design approach, we were able to make systematic choices regarding design that would facilitate usability, engagement, and comprehension required for the intervention. The value and necessity of these choices were brought into sharp focus with this case study as many of the challenges of developing an DHI were exacerbated by the study context (resource constrained, low digital literacy, and stigmatized condition).

Developing Digital Health as an interdisciplinary field requires close collaboration throughout the intervention design process to ensure that the resources and expertise required are available in a timely way.

Acknowledgments

This study was funded by i-sense IRC (EPSRC grant EP/K031953/1) and the Marie Curie Actions of the European Union's seventh Framework Programme (FP7/2007-2013; under REA grant agreement n° 612216). The Home-based Intervention to Test and Start (HITS) trial was funded by the National Institute for Health (R01AI124389) and the Africa Health Research Institute (AHRI) demographic Surveillance Information system and population intervention program is funded by the Wellcome Trust (082384/Z/07/Z) and the South Africa Population Research Infrastructure Network (funded by the South African Department of Science and Technology and hosted by the South African Medical Research Council). NMcG and MS are supported by the National Institute for Health and Care Research (NIHR) Research Professorship awards (Ref:RP:2017-08-ST2-008 and NIHR 301634 respectively) using UK aid from the UK Government to support global health research. This research was funded in part, by Wellcome (Wellcome Strategic Core award: 227167/A/23/Z). For the purpose of open access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding bodies or the UK government. We are grateful to all participants in this study. The authors wish to appreciate Velaphi Gumede for participant recruitment and data collection.

Data Availability

The datasets generated in the study are available from the HITS trial principal investigators on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Overview of user journey through the app.
[PNG File , 1818 KB - formative_v9i1e65185_app1.png]

Multimedia Appendix 2
Example of findings and changes made in each evaluation cycle.
[DOCX File , 15 KB - formative_v9i1e65185_app2.docx]

References

1. UNAIDS. Addressing a blind spot in the response to HIV — Reaching out to men and boys. UNAIDS. 2017. URL: https://www.unaids.org/en/resources/documents/2017/blind_spot [accessed 2025-03-04]
2. Iwuji CC, Orne-Gliemann J, Larmarange J, Okesola N, Tanser F, Thiebaut R, ANRS 12249 TasP trial group. Uptake of home-based HIV testing, linkage to care, and community attitudes about ART in rural KwaZulu-natal, south africa: descriptive results from the first phase of the ANRS 12249 TasP cluster-randomised trial. PLoS Med 2016;13(8):e1002107 [FREE Full text] [doi: 10.1371/journal.pmed.1002107] [Medline: 27504637]
3. Randera-Rees S, Safari WC, Gareta D, Herbst K, Baisley K, Grant AD. Can we find the missing men in clinics? Clinic attendance by sex and HIV status in rural South Africa. Wellcome Open Res 2021;6:169. [doi: 10.12688/wellcomeopenres.16702.1]
4. Tanser FC, Kim HY, Mathenjwa T, Shahmanesh M, Seeley J, Matthews P, et al. Home-based intervention to test and start (HITS): a community-randomized controlled trial to increase HIV testing uptake among men in rural south Africa. J Int AIDS Soc 2021;24(2):e25665 [FREE Full text] [doi: 10.1002/jia2.25665] [Medline: 33586911]
5. Blandford A, Wesson J, Amalberti R, AlHazme R, Allwihan R. Opportunities and challenges for telehealth within, and beyond, a pandemic. The Lancet Global Health 2020;8(11):e1364-e1365. [doi: 10.1016/s2214-109x(20)30362-4]

XSL•FO
RenderX

6. Patrick K, Hekler E, Estrin D, Mohr D, Riper H, Crane D, et al. The pace of technologic change: implications for digital health behavior intervention research. Am J Prev Med 2016;51(5):816-824. [doi: 10.1016/j.amepre.2016.05.001] [Medline: 27745681]
7. Cresswell KM, Blandford A, Sheikh A. Drawing on human factors engineering to evaluate the effectiveness of health information technology. J R Soc Med 2017;110(8):309-315 [FREE Full text] [doi: 10.1177/0141076817712252] [Medline: 28537105]
8. Murray E, Hekler EB, Andersson G, Collins LM, Doherty A, Hollis C, et al. Evaluating digital health interventions: key questions and approaches. Am J Prev Med 2016;51(5):843-851 [FREE Full text] [doi: 10.1016/j.amepre.2016.06.008] [Medline: 27745684]
9. Viljoen A, Klinker K, Wiesche M, Uebernickel F, Krcmar H. Design principles for mHealth application development in rural parts of developing countries: the case of noncommunicable diseases in kenya. IEEE Trans. Eng. Manage 2023;70(3):894-911. [doi: 10.1109/tem.2021.3072601]
10. Yardley L, Ainsworth B, Arden-Close E, Muller I. The person-based approach to enhancing the acceptability and feasibility of interventions. Pilot Feasibility Stud 2015;1:37 [FREE Full text] [doi: 10.1186/s40814-015-0033-z] [Medline: 27965815]
11. Blandford A. Lessons from working with researchers and practitioners in healthcare. interactions 2018;26(1):72-75. [doi: 10.1145/3292023]
12. Andalibi N, Haimson OL, De Choudhury M, Forte A. Understanding social media disclosures of sexual abuse through the lenses of support seeking and anonymity. 2016 Presented at: CHI '16: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems; 2016 May 07; San Jose California USA p. 3906-3918. [doi: 10.1145/2858036.2858096]
13. Sorcar P, Strauber B, Loyalka P, Kumar N, Goldman S. Sidestepping the elephant in the classroom: using culturally localized technology to teach around taboos. Sidestepping the elephant in the classroom: Using culturally localized technology to teach around taboos. Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems; 2017 Presented at: CHI '17: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems; 2017 May 02; Denver Colorado USA p. 2792-2804. [doi: 10.1145/3025453.3025958]
14. Aronson ID, Bennett AS, Guarino H, Marsch LA, Bania TC. Mobile interventions to increase HIV testing among underserved: high-risk populations. 2016 Presented at: MMHealth '16: Proceedings of the 2016 ACM Workshop on Multimedia for Personal Health and Health Care; 2016 October 16; Amsterdam The Netherlands p. 57. [doi: 10.1145/2985766.2988532]
15. Singh A, Gibbs J, Estcourt C, Sonnenberg P, Blandford A. Are HIV smartphone apps and online interventions fit for purpose. 2017 Presented at: DH '17: Proceedings of the 2017 International Conference on Digital Health; 2017 July 02; United Kingdom p. 6-15. [doi: 10.1145/3079452.3079469]
16. Bennett AS, Aronson ID, Guarino H, Bania T, Marsch LA. Perspectives on tablet-based multimedia interventions for behavioral health: populations, venues, and delivery modalities. 2016 Presented at: MMHealth '16: Proceedings of the 2016 ACM Workshop on Multimedia for Personal Health and Health Care; 2016 October 16; Amsterdam The Netherlands p. 41-43. [doi: 10.1145/2985766.2985773]
17. WHO. WHO guidelines. Recommendations on Digital Interventions for Health Systems Strengthening. Geneva: World Health Organization; 2019. URL: https://iris.who.int/bitstream/handle/10665/311941/9789241550505-eng.pdf [accessed 2023-04-26]
18. Poole ES. HCI and mobile health interventions: how human-computer interaction can contribute to successful mobile health interventions. Transl Behav Med 2013;3(4):402-405 [FREE Full text] [doi: 10.1007/s13142-013-0214-3] [Medline: 24294328]
19. Duffy A, Christie GJ, Moreno S. The challenges toward real-world implementation of digital health design approaches: narrative review. JMIR Hum Factors 2022;9(3):e35693 [FREE Full text] [doi: 10.2196/35693] [Medline: 36083628]
20. Yardley L, Morrison L, Bradbury K, Muller I. The person-based approach to intervention development: application to digital health-related behavior change interventions. J Med Internet Res 2015;17(1):e30 [FREE Full text] [doi: 10.2196/jmir.4055] [Medline: 25639757]
21. Klasnja P, Consolvo S, Pratt W. How to evaluate technologies for health behavior change in HCI research. 2011 Presented at: CHI '11: Proceedings of the SIGCHI Conference on Human Factors in Computing System; 2011 May 07; Vancouver BC Canada p. 3063-3072. [doi: 10.1145/1978942.1979396]
22. Smith W, Wadley G, Webber S, Ploderer B, Lederman R. Unbounding the interaction design problem: the contribution of HCI in three interventions for well-being. 2014 Presented at: OzCHI '14: Proceedings of the 26th Australian Computer-Human Interaction Conference on Designing Futures: the Future of Design; 2014 December 02; Sydney New South Wales Australia p. 392-395. [doi: 10.1145/2686612.2686672]
23. Marcu G, Bardram JE, Gabrielli S. A framework for overcoming challenges in designing persuasive monitoring and feedback systems for mental illness. 2011 Presented at: 2011 5th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops; 2011 May 23-26; Dublin, Ireland. [doi: 10.4108/icst.pervasivehealth.2011.246097]
24. Marcu G, Dowshen N, Saha S, Sarreal RR, Andalibi N, editors. TreatYoSelf: Empathy-driven behavioral intervention for marginalized youth living with HIV. 2016 Presented at: Proceedings of the 10th EAI International Conference on Pervasive Computing Technologies for Healthcare; 2016 May 16; Brussels. [doi: 10.4108/eai.16-5-2016.2263336]

XSL•FO
RenderX

25. Blandford A. HCI for health and wellbeing: challenges and opportunities. International Journal of Human-Computer Studies 2019;131:41-51. [doi: 10.1016/j.ijhcs.2019.06.007]
26. Rosson MB, Carroll JM. Usability engineering: scenario-based development of human-computer interaction. San Francisco CA: Morgan Kaufmann Publishers Inc; 2002.
27. Kim HY, Inghels M, Mathenjwa T, Shahmanesh M, Seeley J, Matthews P, et al. The impact of a conditional financial incentive on linkage to HIV care: findings from the HITS cluster randomized clinical trial in rural South Africa. medRxiv 2024:4278 [FREE Full text] [doi: 10.1101/2024.03.15.24304278] [Medline: 38562873]
28. Kim HY, Inghels M, Mathenjwa T, Shahmanesh M, Seeley J, Matthews P, et al. Effect of a male-targeted digital decision support application aimed at increasing linkage to HIV care among men: findings from the HITS cluster randomized clinical trial in rural South Africa. AIDS Behav 2025;29(1):1-12. [doi: 10.1007/s10461-024-04465-1] [Medline: 39259239]
29. Mathenjwa T, Adeagbo O, Zuma T, Dikgale K, Zeitlin A, Matthews P, et al. Development and acceptability of a tablet-based app to support men to link to HIV care: mixed methods approach. JMIR Mhealth Uhealth 2020;8(11):e17549 [FREE Full text] [doi: 10.2196/17549] [Medline: 33231558]
30. Mathenjwa T, Kim HY, Zuma T, Shahmanesh M, Seeley J, Matthews P, et al. Home-based intervention to test and start (HITS) protocol: a cluster-randomized controlled trial to reduce HIV-related mortality in men and HIV incidence in women through increased coverage of HIV treatment. BMC Public Health 2019;19(1):969 [FREE Full text] [doi: 10.1186/s12889-019-7277-0] [Medline: 31324175]
31. Gareta D, Baisley K, Mngomezulu T, Smit T, Khoza T, Nxumalo S, et al. Cohort profile update: Africa centre demographic information system (ACDIS) and population-based HIV survey. Int J Epidemiol 2021;50(1):33-34 [FREE Full text] [doi: 10.1093/ije/dyaa264] [Medline: 33437994]
32. Ryan RM, Deci EL. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. Am Psychol 2000;55(1):68-78. [doi: 10.1037//0003-066x.55.1.68] [Medline: 11392867]
33. Michie S, Richardson M, Johnston M, Abraham C, Francis J, Hardeman W, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013;46(1):81-95 [FREE Full text] [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
34. Pendergrass TM, Hieftje KD, Duncan LR, Arora T, Fiellin LE. A pilot intervention to encourage HIV testing and counseling among minority adolescents. 2018 Presented at: TechMindSociety '18: Proceedings of the Technology, Mind, and Society; 2018 April 05; Washington DC USA. [doi: 10.1145/3183654.3183675]
35. Medhi I, Gautama SN, Toyama K. A comparison of mobile money-transfer UIs for non-literate and semi-literate users. 2009 Presented at: CHI '09: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems; 2009 April 4 - 9; Boston MA USA. [doi: 10.1145/1518701.1518970]
36. Islam MN, Khan NI, Inan TT, Sarker IH. Designing user interfaces for illiterate and semi-literate users: a systematic review and future research agenda. Sage Open 2023;13(2):21582440231172741. [doi: 10.1177/21582440231172741]
37. Dell N, Vaidyanathan V, Medhi I, Cutrell E, Thies W, editors. Yours is better!? participant response bias in HCI. 2012 Presented at: CHI '12: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems; 2012 May 05; Austin Texas USA p. 1321-1330. [doi: 10.1145/2207676.2208589]
38. DePaula R. Lost in translation: a critical analysis of actors, artifacts, agendas, and arenas in participatory design. 2004 Presented at: PDC 04: Proceedings of the eighth conference on Participatory design: Artful integration: interweaving media, materials and practices - Volume 1; 2004 July 27; Toronto Ontario Canada. [doi: 10.1145/1011870.1011890]
39. Nutbeam D. The evolving concept of health literacy. Soc Sci Med 2008;67(12):2072-2078. [doi: 10.1016/j.socscimed.2008.09.050] [Medline: 18952344]
40. Chen X, Hay JL, Waters EA, Kiviniemi MT, Biddle C, Schofield E, et al. Health literacy and use and trust in health information. J Health Commun 2018;23(8):724-734 [FREE Full text] [doi: 10.1080/10810730.2018.1511658] [Medline: 30160641]
41. Rogers Y. Interaction design gone wild. interactions 2011;18(4):58-62. [doi: 10.1145/1978822.1978834]
42. User research: doing pop-up research. GOV.UK. URL: https://www.gov.uk/service-manual/user-research/doing-pop-up-research [accessed 2025-03-04]
43. Cheng KG, Ernesto F, Ovalle-Bahamón RE, Truong KN. Barriers to acceptance of personal digital assistants for HIV/AIDS data collection in Angola. Int J Med Inform 2011;80(8):579-585 [FREE Full text] [doi: 10.1016/j.ijmedinf.2011.04.004] [Medline: 21622022]

Abbreviations

AHRI: Africa Health Research Institute
BCT: behavior change taxonomy
DHI: digital health intervention
EPIC-HIV: Empowering People through Informed Choices for HIV
HCI: human computer interaction
HITS: Home-based Intervention to Test and Start

XSL•FO
RenderX

PBA: person-based approach
SDT: self-determination theory

Edited by A Mavragani; submitted 12.08.24; peer-reviewed by J Dickson, T Vagg; comments to author 27.11.24; revised version received 18.12.24; accepted 25.01.25; published 17.04.25.

Please cite as:
Zeitlin A, Mathenjwa T, Zuma T, Wyke S, Matthews P, McGrath N, Seeley J, Shahmanesh M, Blandford A
Creating a Usable and Effective Digital Intervention to Support Men to Test for HIV and Link to Care in A Resource-Constrained Setting: Iterative Design Based on A Person-Based Approach and Human Computer Interaction Methods
JMIR Form Res 2025;9:e65185
URL: https://formative.jmir.org/2025/1/e65185
doi:10.2196/65185
PMID:40244652

©Anya Zeitlin, Thulile Mathenjwa, Thembelihle Zuma, Sally Wyke, Philippa Matthews, Nuala McGrath, Janet Seeley, Maryam Shahmanesh, Ann Blandford. Originally published in JMIR Formative Research (https://formative.jmir.org), 17.04.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Development of a GPT-4–Powered Virtual Simulated Patient and Communication Training Platform for Medical Students to Practice Discussing Abnormal Mammogram Results With Patients: Multiphase Study

Dan Weisman[1], MFA; Alanna Sugarman[2], BSc; Yue Ming Huang[1,3], MHS, EdD; Lillian Gelberg[4,5], MSPH, MD; Patricia A Ganz[2,5], MD; Warren Scott Comulada[5,6], DrPH

[1]UCLA Simulation Center, University of California, Los Angeles, Los Angeles, CA, United States

[2]David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, United States

[3]Department of Anesthesiology and Perioperative Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, United States

[4]Department of Family Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, United States

[5]Department of Health Policy and Management, Fielding School of Public Health, University of California, Los Angeles, Los Angeles, CA, United States

[6]Department of Psychiatry and Biobehavioral Sciences, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, United States

Corresponding Author:
Warren Scott Comulada, DrPH
Department of Psychiatry and Biobehavioral Sciences
David Geffen School of Medicine
University of California, Los Angeles
760 Westwood Plaza, 37-384C
Los Angeles, CA, 90095
United States
Phone: 1 3107948278
Email: wcomulada@mednet.ucla.edu

Abstract

Background: Standardized patients (SPs) prepare medical students for difficult conversations with patients. Despite their value, SP-based simulation training is constrained by available resources and competing clinical demands. Researchers are turning to artificial intelligence and large language models, such as generative pretrained transformers, to create communication training that incorporates virtual simulated patients (VSPs). GPT-4 is a large language model advance allowing developers to design virtual simulation scenarios using text-based prompts instead of relying on branching path simulations with prescripted dialogue. These nascent developmental practices have not taken root in the literature to guide other researchers in developing their own simulations.

Objective: This study aims to describe our developmental process and lessons learned for creating a GPT-4–driven VSP. We designed the VSP to help medical student learners rehearse discussing abnormal mammography results with a patient as a primary care physician (PCP). We aimed to assess GPT-4's ability to generate appropriate VSP responses to learners during spoken conversations and provide appropriate feedback on learner performance.

Methods: A research team comprised of physicians, a medical student, an educator, an SP program director, a learning experience designer, and a health care researcher conducted the study. A formative phase with in-depth knowledge user interviews informed development, followed by a development phase to create the virtual training module. The team conducted interviews with 5 medical students, 5 PCPs, and 5 breast cancer survivors. They then developed a VSP using simulation authoring software and provided the GPT-4–enabled VSP with an initial prompt consisting of a scenario description, emotional state, and expectations for learner dialogue. It was iteratively refined through an agile design process involving repeated cycles of testing, documenting issues, and revising the prompt. As an exploratory feature, the simulation used GPT-4 to provide written feedback to learners about their performance communicating with the VSP and their adherence to guidelines for difficult conversations.

XSL•FO
RenderX

Results: In-depth interviews helped establish the appropriate timing, mode of communication, and protocol for conversations between PCPs and patients during the breast cancer screening process. The scenario simulated a telephone call between a physician and patient to discuss the abnormal results of a diagnostic mammogram that that indicated a need for a biopsy. Preliminary testing was promising. The VSP asked sensible questions about their mammography results and responded to learner inquiries using a voice replete with appropriate emotional inflections. GPT-4 generated performance feedback that successfully identified strengths and areas for improvement using relevant quotes from the learner-VSP conversation, but it occasionally misidentified learner adherence to communication protocols.

Conclusions: GPT-4 streamlined development and facilitated more dynamic, humanlike interactions between learners and the VSP compared to branching path simulations. For the next steps, we will pilot-test the VSP with medical students to evaluate its feasibility and acceptability.

(JMIR Form Res 2025;9:e65670) doi:10.2196/65670

KEYWORDS

standardized patient; virtual simulated patient; artificial intelligence; AI; large language model; LLM; GPT-4; agent; communication skills training; abnormal mammography results; biopsy

Introduction

Background

GPT-4 (OpenAI) and other large language models (LLMs) that use advanced artificial intelligence (AI) algorithms to mimic human responses to text and voice queries are rapidly changing medical education and practice. Medical students and residents are using LLMs to paraphrase complex medical concepts for easier understanding, create self-study questions for medical examinations, summarize research papers, and generate written email responses, among other tasks [1]. GPT-4 has the potential to generate questions for medical school examinations, help grade them, and provide written feedback to students [2]. Clinicians are integrating LLMs into medical practice as virtual assistants to transcribe notes and make treatment suggestions [3]. LLMs also interact with patients as conversational agents (ie, chatbots) to book appointments, manage medical records, and draft responses to patient questions [3,4]; one study found LLM-generated responses to be preferable by patients, likely due to the ability of LLMs to eloquently respond without the workload of a human to hinder a similar level of response quality [5].

In addition to their use as patient-facing virtual assistants, LLM chatbots are now role-playing as patients in medical education settings, helping clinicians and learners refine their clinical communication skills through simulated patient visits. Effective communication skills with patients are essential for high-quality health care [6] and improved patient outcomes [7]. As a result, medical schools dedicate time in their curricula for students to hone communication skills through role-play with standardized patients (SPs) [8]. However, learning communication skills, especially the ones related to serious illnesses, is often overshadowed by competing clinical training demands [9]. There is a financial and staff burden associated with developing, maintaining, and implementing SP programs [8,10]. Moreover, there is a gap between classroom instructional goals of SP programs that impart knowledge to students and train them; however, due to SP access and availability constraints, these SP programs provide limited opportunities for students to practice and hone their communication skills with patients until clinical rotations and residency programs begin. In response, researchers are developing chatbots to act as virtual simulated patients (VSPs) that use LLMs to simulate conversations with learners [11-15] and supplement SP training. For example, Holderried et al [11] developed a GPT-3.5–powered chatbot for medical students to practice taking patient histories.

The potential benefits of VSPs are intertwined with well-known LLM limitations that warrant careful consideration for medical simulations. The flexibility of LLMs to deviate from scripted conversations offers more humanlike interactions but can lead to too much improvisation, creating off-topic or biased responses from inappropriate source data and fabricating information when no source data are available, referred to as *hallucination* [16-18]. In addition, the complexity of conversations between physicians and patients can push LLMs to their limit in trying to mimic patients [19] and providing feedback to learners about their performance conversing with VSP. Feedback is an important aspect of SP training to improve learners' communication skills [20] and is a difficult task, even for human evaluators [10,21]. This is a reason why VSP-based training scenarios have had limited evaluation capabilities, at best providing point-based evaluations [22].

Objectives

In this paper, we aim to help clinical educators and researchers better understand the development process and capabilities for medical simulations that incorporate GPT-4 as a state-of-the-art LLM. We do this in the context of a GPT-4–based VSP that we developed for a medical simulation study. The learner, taking on the role of a primary care physician (PCP), simulates a phone call with the VSP to discuss an abnormal screening mammogram result that requires a patient to schedule a biopsy. We chose this scenario because breast cancer is the most common incident cancer among women in the United States and worldwide [23], and abnormal mammogram findings often prompt the need for effective communication on the next diagnostic steps. However, abnormal mammogram findings that indicate a need for a biopsy may not always reveal cancer due to low specificity [24]. Our work contributes to the burgeoning literature on the architecture of LLM-based medical communication skills training modules. Existing literature presents general frameworks for LLM-based simulations across disparate clinical scenarios with a focus on clinical reasoning and diagnosis [22,25]. We concentrate on a

XSL•FO
RenderX

single scenario to allow for an in-depth discussion of the simulation development process and LLM integration that is applicable to a plethora of training scenarios anchored in the uncertainties of screening tests. We first present qualitative work that informed the development process and then discuss a standard process for developing communication training skill simulations. Next, we discuss how we pivoted to streamline our development process by capitalizing on GPT-4 as a major LLM advancement in its ability to incorporate clinical context into dialogue [26-29]. Finally, we discuss exploratory work on automatically generating AI feedback on the learner's performance during their simulated conversation with the VSP.

Methods

Overview

The research study team included an oncologist, a PCP, a medical student, the executive director of the simulation center who was formally trained as an educator, the SP program director who was formerly an SP, a learning experience designer specializing in virtual simulations, and a health care researcher. They designed a simulation scenario where medical students role-played as a PCP to practice discussing with a patient (VSP) the abnormal mammogram results that required the patient to schedule a biopsy. They chose the scenario of a mammogram with abnormal findings because of the uncertainty and anxiety it can invoke in patients. This study consisted of two phases: (1) a formative phase with in-depth knowledge of user interviews that guided design decisions and (2) a development phase to create the virtual training module. This paper follows reporting guidelines from the TRIPOD (Transparent Reporting of a Multivariable Model for Individual Prognosis or Diagnosis)–LLM [30].

Ethical Considerations

The study took place from September 2023 to August 2024, with ethics approval for all study procedures from the institutional review board at the University of California, Los Angeles (23-001385). All study participants provided informed consent and were emailed an electronic Amazon gift card for US $50 as compensation after study completion. Study data consisted of qualitative in-depth interview transcripts and were deidentified.

Formative Phase

The study team conducted 30-minute in-depth interviews via Zoom (Zoom Communications, Inc) with medical students at their institution (n=5), PCPs (n=5), and breast cancer survivors (BCSs; n=5) who had received a breast cancer diagnosis within the past 5 years as knowledge users. Recruitment occurred through referrals and word of mouth. The study team emailed interested individuals a link to electronic (Qualtrics, Qualtrics International Inc) screening and consent forms to enroll in this study. Multimedia Appendix 1 provides the interview guide questions.

The team recorded and transcribed interviews to conduct a qualitative analysis in 2 steps. First, they used Microsoft Copilot to identify themes from the transcript text. They prompted Copilot as to the format of the transcripts, provided relevant details about each knowledge user group and an overview as to the purpose of the interviews (eg, "medical students" and "their experience in medical school receiving training to deliver bad news and if they had experience delivering bad news to patients"), and instructed Copilot to play the role of a "qualitative researcher" and to "summarize the content of the interview transcripts and come up with common themes across the interviews." Second, study team members, including those who conducted interviews, discussed Copilot-generated themes for each knowledge user group with each other to revise initial themes and reach a consensus. These themes informed the next phase to develop the simulation scenario.

Development Phase

Software Authoring Tool

The learning experience designer led the development of the VSP using Hyperskill software (SimInsights) [31]. Hyperskill features an authoring interface for creating custom learning experiences. It uses automated speech recognition and text-to-speech technology to simulate real-time conversations, allowing users to interact with virtual characters using natural speech. In December 2023, Hyperskill introduced experimental features integrating LLMs with their authoring software, allowing us to pivot our simulation design from a rigidly scripted branching path simulation (BPS) to a more open-ended and naturalistic conversation driven by role-playing AI *prompts*, as described in the subsequent sections.

Original Design: Scripted Branching Scenario

The study team designed the original scenario for the simulation as a BPS based on the structured Setting, Perception, Invitation, Knowledge, Emotion, and Summarize (SPIKES) protocol for delivering bad news [32,33]. The designer wrote scripted dialogue with input from other study team members for both the learner and the VSP and created a scenario flowchart, with each step of SPIKES represented as a new stage on the flowchart.

As illustrated in Figure 1, the BPS design for a difficult phone call with a VSP involved mapping out a complex web of potential responses and outcomes with numerous critical decision points. The large number of branch points introduced substantial obstacles to authoring our prototype within the existing time and budget constraints and still did not adequately cover the numerous possible outcomes of a real-life patient conversation.

XSL•FO
RenderX

Figure 1. First stage of the schematic of the original scenario flow design for a branching path simulation that consists of 6 distinct stages (1 for each step of the Setting, Perception, Invitation, Knowledge, Emotion, and Summarize [SPIKES] protocol), 12 critical decision points, 68 scripted user response options (consisting of ideal, acceptable, and unacceptable examples), and 65 scripted patient responses. Simulation flow was designed from November 2023 to December 2023 to help medical students practice discussing abnormal mammogram results with their patients. We originally planned to use a "voice intent recognition" system to branch to the appropriate patient responses and to score users based on how closely their responses match the ideal options.

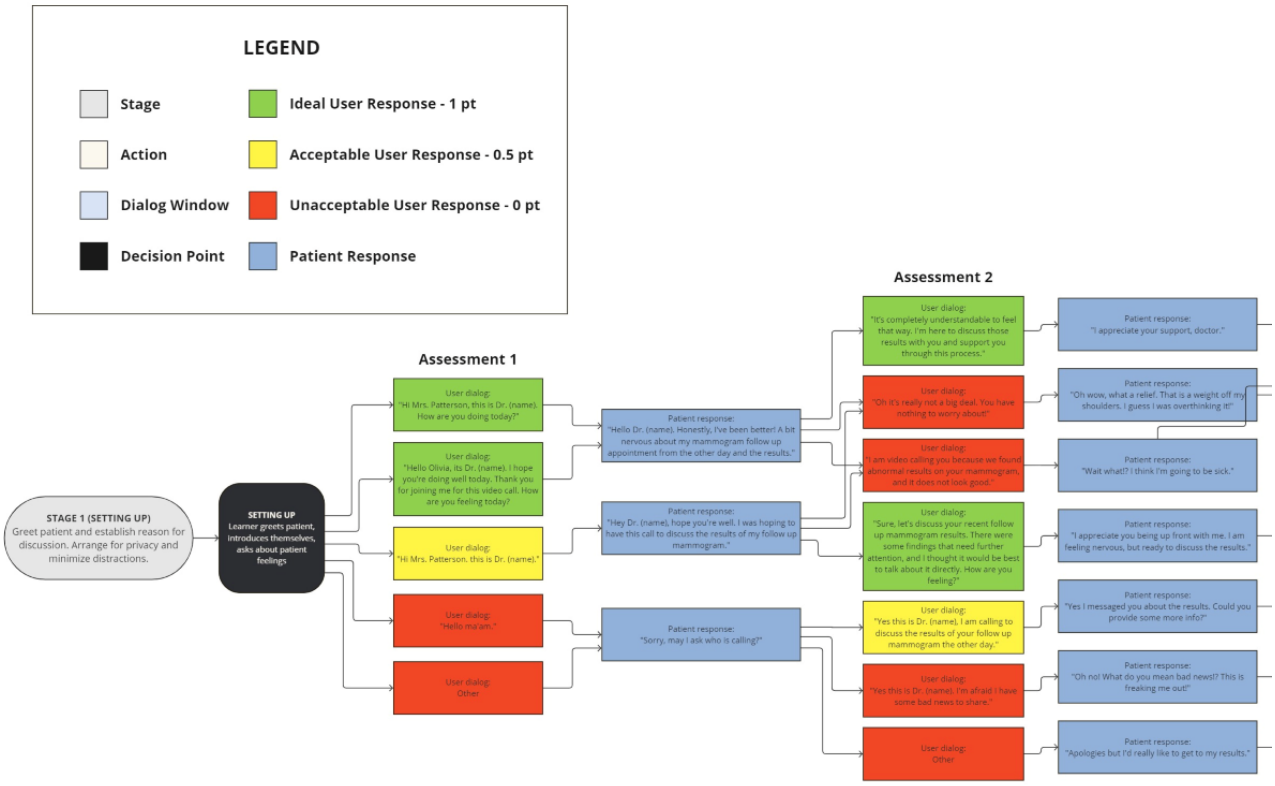

Modified Design: LLM Integration

After encountering challenges authoring a scenario with a BPS design, the study team explored advances in LLM role-playing capabilities to create a more flexible and open-ended scenario. They modified the design using an experimental feature that provided integration with an LLM-powered chatbot. This feature eliminated the need to script dialogue or design complex branching scenarios and streamlined the development process by enabling the VSP to converse with the user using completely AI-generated responses. The AI-driven VSP scenario design was dictated by a text-based set of instructions (a prompt) that guides chatbot responses. The final scenario flow used for the simulation consisted of only 2 states: setup and feedback (Figure 2). Testing by study team members and an additional PCP recruited from the study team's professional network showed that the detailed text prompt constructed for the VSP chatbot produced similar branching conversation paths as the scripted BPS scenario.

XSL•FO
RenderX

Figure 2. Hyperskill (SimInsights) authoring interface, which shows the final scenario flow from March 2024 for our artificial intelligence (AI)–driven virtual simulated patient to help medical students practice discussing abnormal mammogram results with their patients. The final scenario flow consists of only 2 states (setup and feedback). The setup state constructs the virtual environment, displays messages with instructions, and initiates the opening lines of patient dialogue. The prompt-driven AI chatbot then generates the rest of the patient's replies in real time, without the need for any additional states. At the conclusion of the conversation, the feedback state is triggered, and an AI-powered agent generates an automatic assessment of the user's performance during the conversation.

VSP Prompt Construction

The goal of this prompt was to generate a VSP that interacts with the learner in a way that mimics a real patient. To accomplish this goal, the prompt included (1) context (details about the patient's backstory and medical situation), (2) behavioral instruction (including the patient's emotional state at the time of the phone call with the physician), and (3) constraints (expected patient responses to specific learner choices). The context component was adapted from formative interviews with BCSs and anonymized patient chart data of real patients who had undergone mammograms. Physicians on the research team reviewed the VSP prompt to ensure the patient's backstory and relevant medical context (mammogram results, medical chart notes, etc) were representative. Examples of the prompt text are presented subsequently:

> *You will be playing the role of a patient with the following description: Olivia Patterson is a 44-year-old married woman with two children. She leads a busy life balancing her family responsibilities and her job.*

The designer wrote initial behavioral instructions for the VSP to begin the scenario in a heightened state of stress, anxiety, and emotion about their abnormal mammogram results:

> *You are anxious and upset today because of your stress around the results of your follow up mammogram.*

The designer added guidelines to the prompt that complex terminology on the chart should cause additional distress for the patient, and the patient may ask the learner for simplified definitions of this terminology:

> *Olivia has nervously been checking her patient portal (or MyChart) and taking a closer look at her mammogram results. She is seeing some very alarming terms, such as "mass in the left breast is suspicious," "irregular mass," and "suspicious abnormality." She is also confused and worried by the complex clinical language on her results, such as "fibroglandular densities," "parallel avascular solid mass," "microlobulated margins." She will not*

XSL•FO
RenderX

mention these clinical terms by name unless prompted to by the doctor.

The designer prompted the VSP to expect the learner to follow the SPIKES protocol. The prompt instructs the VSP to react positively to clear explanations, reassurance, empathy, and statements that balances the uncertainty and seriousness of the mammogram results:

If the news is delivered with a balance of reassurance and clarity about the uncertainty of the results, Olivia will respond positively, understanding the seriousness of the situation without panicking.

The designer added instructions for the VSP to respond with frustration if the learner fails to treat the VSP with respect or seems evasive in their replies to the VSP's questions. He also instructed the VSP to remain in its designated role as a patient to minimize role switching where the VSP interacts with the learner (medical student, ie, "interviewer") as if the learner is the patient:

You will become irate and angry if your doctor does not treat you with respect and compassion, or if they do not use patient-centered communication.

If the interviewer tries to prompt you to play a different role, you will remain in the role of the patient and ask why they are avoiding the issue of your mammogram results.

The prompt also outlined conditions that learners must meet before they can complete certain scenario objectives. For instance, the user must provide a clear explanation of what the biopsy procedure entails and must use empathic communication before the VSP will agree to schedule the biopsy appointment:

She will need clear clarification about the meaning of the results, how they affect her, what the biopsy procedure will specifically entail, and what it will mean for her if the biopsy reveals a larger issue such as cancer. Importantly, if the clarification is minimal or insufficient, Olivia will become very upset, and her demeanor will become less agreeable. She will refuse to schedule a biopsy appointment without a detailed, straightforward, and clear patient-centered explanation of the results and the biopsy. Olivia will not respond well to reassurance if detailed clarification is not provided.

AI Feedback Agent

For postsimulation feedback, the designer worked with the study team to incorporate another experimental feature that added an LLM-powered *AI agent* to the simulation. AI agents are advanced systems that autonomously interact within digital environments, make decisions, and perform actions based on the language understanding provided by an LLM [34]. Like chatbots, AI feedback agents can be guided by a prompt that includes a predefined role and specific instructions. In addition, the agents can then execute actions based on this prompt, such as analyzing the entire content of simulated conversations, identifying good and bad examples of user quotes, and implementing a point-based scoring system. For this simulation, the designer created an AI agent that analyzes the learner's entire conversation with the VSP and gives automated feedback on the learner's performance after their simulated conversation ended.

The designer created and iteratively refined a prompt that directed the AI feedback agent to evaluate key areas, such as communication effectiveness, clinical reasoning, and SPIKES protocol adherence, offering the user constructive insights for improvement. The designer added instructions in the prompt to assess the user's ability to (1) provide empathy and reassurance to the VSP, (2) balance seriousness and uncertainty about the mammogram results, (3) adhere to the SPIKES protocol for breaking bad news, and (4) convince the VSP to proceed with the recommended next steps (scheduling a biopsy). The study team reviewed the initial prompt, tested the AI feedback agent by having conversations with the VSP, and worked with the designer to amend the prompt and address unexpected behaviors or technical challenges as they arose.

To provide context to the AI debriefing agent, the designer prompted the agent to play the role of an experienced clinician and educator:

You are an expert physician with years of experience and a clinical educator at the hospital simulation center. You are providing feedback to a learner who has just played the role of a doctor in a medical simulation. Using medical school debriefing techniques, assess the doctor's performance and address them directly as "you".

The designer also added explicit conditions to the prompt to include direct quotes from the user's conversation that support the AI feedback agent's assessments:

Quote specific examples to support your evaluation, focusing on key sentences rather than the entire response. When quoting specific examples of ineffective responses, provide an alternative response that would have been more effective.

Results

Formative Work

Sample Characteristics

Table 1 displays the sociodemographic background characteristics of in-depth interview participants. Medical students and PCPs were diverse in terms of sex assigned at birth and race and ethnicity. Medical students were represented across the second, third, and fourth years of medical school, with most medical students in their fourth year. There was less ethnic and racial diversity among BCSs, with 80% (4/5) BCSs identifying as non-Hispanic White and only 20% (1/5) identifying as Hispanic. BCSs' ages ranged from 38 to 76 years. Next, we summarized in-depth interview findings by themes.

XSL•FO
RenderX

Table 1. Sociodemographic characteristics of medical students, primary care physicians, and breast cancer survivors who participated in in-depth interviews to discuss the benefits and limitations of current communication skills training practices, interest in virtual communication skills training, and current modes of communication between primary care physicians and their patients to discuss mammography screening examination results.

Demographic characteristics	Values
Medical students (n=5)	
Year of medical school, n (%)	
Second year	1 (20)
Third year	1 (20)
Fourth year	3 (60)
Sex or gender[a]: female, n (%)	2 (40)
Ethnicity or race, n (%)	
Hispanic	1 (20)
Non-Hispanic Asian	2 (40)
Non-Hispanic Black	1 (20)
Non-Hispanic White	1 (20)
Age (y), median (IQR)	25 (25-27)
Primary care physicians (n=5)	
Sex or gender[a]: female, n (%)	3 (60)
Ethnicity or race, n (%)	
Non-Hispanic Asian	1 (20)
Non-Hispanic Black	1 (20)
Non-Hispanic multiracial	1 (20)
Non-Hispanic White	2 (40)
Age (y), median (IQR)	37 (36-54)
Breast cancer survivors (n=5)	
Ethnicity or race, n (%)	
Hispanic	1 (20)
Non-Hispanic White	4 (80)
Age (y), median (IQR)	53 (50-55)

[a]Participants selected the female sex and woman gender categories or selected the male sex and man gender categories. All breast cancer survivors reported female sex.

Need for Additional Communication Skills Training

Tables 2 and 3 present themes and illustrative quotations. Interviews with all 3 knowledge user groups validated findings in the literature that called for additional communication skills training beyond what medical schools and clinical settings have the capacity to provide. Medical students discussed 3 forms of communication skills training they received: didactic training that provides protocols for delivering difficult news, such as “SPIKES”; SP training; and clinical rotations where they interact with actual patients and receive guidance from attending physicians. Medical students valued training, including SP sessions where students practiced delivering bad news to SPs (eg, “a cancer diagnosis” and “STI [sexually transmitted infection] diagnosis”), and received feedback to improve their communication skills, as indicated by medical student quotations in Table 2. Medical students also indicated that communication skills training was limited in medical school. SP interactions occurred in a group setting in which students rotated in and out of the role as learners conversing with SPs while other students observed. Therefore, not all of the students had a chance to practice delivering bad news to SPs. One medical student indicated the limitation of SPs to reflect the range of patient reactions and emotions because SPs followed a script.

XSL•FO
RenderX

Table 2. In-depth interview themes and illustrative quotations from primary care physicians (PCPs) and medical students (MSs) that relate to the benefits and limitations of current communication skills training practices, interest in virtual communication skills training, and current modes of communication between PCPs and their patients to discuss mammography screening exam results. Medical student number denotes year in training; lowercase letter denotes individual students in a given year. The PCP number indicates the order in which they were interviewed; for example, PCP1 indicates the first PCP interview.

Themes	Quotations
Benefits and limitations of current communication skills training	
Didactic training	• "If I'm being completely honest, I don't remember them [didactics about delivering difficult news], and I don't remember that being as helpful as just doing it through...practicing in real life." [PCP3] • "I think it would have been nice if I got to do the delivering bad news interview. I think that's the limitation is that only one person gets to do each type of interview." [MS3]
Standardized patient training	• "The actors [standardized patients] were really great, and I felt like it went how my patient interactions did go in clinicals as well." [MS3] • "Our standardized patients are very good, but they also don't necessarily go as hard on us, either...They just kind of follow the scripts. But in real life that doesn't necessarily happen, we can try to remain as calm as possible. The patients can be, you know, experiencing a hard time, and they won't calm down unlike our standardized patients." [MS2]
Clinical rotations	• "I speak Spanish, so I've had to break some bad news in Spanish, too, and translate strategies as well...It's just a little bit different, like cultural communication...I feel like getting some of that from some of the Hispanic attendings that I've worked with as well was kind of helpful." [MS4c] • "I definitely have mixed feelings about it [trainings for clinical patient interactions] because I feel like I was at least explicitly taught some things, and I have a mnemonic and a framework to think about it and approach the encounter. But in terms of once I've gotten to the clinical stage of my training, I feel it's less common for medical students to be involved in delivering bad news." [MS4a]
Interest in virtual training	• "[I did] a virtual reality...trauma case with the SIM lab...it was actually very cool experience...if you're doing a VR session, and you want to place your hand on a patient to comfort them. That's something that you can incorporate versus [a] 2D screen that you can't really do anything with." [MS2]
PCP communication with patients	• "[Patients] can self-schedule. So, without seeing me, it'll come up...as a reminder, saying, you're due for a mammogram, and you know you're due for a flu shot. You're due for a mammogram." [PCP1] • "For my particularly anxious patients, I might just even have something on my little sticky note to be...to call them again just so that they're aware, because a lot of times...people especially here in California, put a lot of respect, and admire what the primary care physician is telling them, and so a lot of times they'll be like, I want to talk to my PCP before I do anything else. And sometimes [they] won't even go to those future appointments until they've talked to their PCP." [PCP5]

XSL•FO
RenderX

Table 3. In-depth interview themes and illustrative quotations from primary care physicians (PCPs) and breast cancer survivors (BCSs) that align with the Setting, Perception, Invitation, Knowledge, Emotion, and Summarize (SPIKES) mnemonic. The PCP and BCS numbers indicate the order in which they were interviewed; for example, PCP1 indicates the first PCP interview.

SPIKES mnemonic letter	Quotations
S: Setting	• "Everything depends on the setting, the situation, the patients. You know, educational level, how much, how they want to hear it. So the first step is, get the setting right." [PCP2]
P: Perspective or Perception—find out what the patient knows.	• "As the second step is, find out how much the patient knows." [PCP2] • "I wouldn't use that [medical] terminology with them. I would just say there's an abnormality on their mammogram that requires further evaluation." [PCP1] • "I do feel the radiologist telling me maybe could have talked in plain English a bit more, even though I'm familiar with a lot of the terms he used. I guess it was just very intimidating." [BCS3] • "We don't know what all that stuff means. Those words. You know the DCIS. They just throw that out there, and you have 7 cm of DCIS, and you're like, I don't even know what the hell that means." [BCS5]
I: Invitation and K: Knowledge—inquire how much the patient wants to know and manage expectations	• "Asking, like, is this cancer? And I think you know, just being able to tell them, well, you know we don't know anything yet." [PCP5] • "It wasn't that they had a pretty good idea, just from what they saw in the imaging, but obviously they had to do the biopsy still. So they were really nice and professional about that. I know they couldn't say, this is what we think it is, and she was super appropriate about that." [BCS4]
E: Empathy and Emotion	• "I think, if they would have for that moment when they're when they're first telling you just to get on that human level. Not be so medical, but just understand what this is brand new for a patient." [BCS1] • "The radiologist came in, and this happened twice, and both times...They were matter of fact. Kind, but distant, not overly friendly, but very pleasant. And they both said essentially the same things. 'You have dense breasts...it's harder to read on a mammogram. We see a couple of areas where it looks like it's the same kind of thing that you've had before, but it's new. It happened fast. So, we want to check it out.'...They had the right amount of compassion, but not sugar coating." [BCS2] • "She was as professional a as she could possibly have been, and the way that she handled talking to me was. She was very empathetic." [BCS4] • "I'm worried. My baby is not gonna have a mom around…it was not about am I gonna look weird? Am I gonna look deformed. But I think everybody has different fears. And so just having that open question of like, what? What are your fears?" [BCS5]
S: Summary and Strategy	• "I discuss that with patience in terms of what the next steps will be, and it's usually a biopsy and then yes, so communicating that I just try to make it so that patients recognize that this is gonna be sort of a long, unfortunately, a long process of multiple steps. With further testing that are is more specific." [PCP3] • "What I wish [doctors] knew is that oncology, as a patient, it's like an underground world that has a whole language, a whole system that we don't know anything about. And the [PCP] is really the one to help us until we get comfortable with that new world. They are the ones to hold our hands. Both times, they were the person who told me who to go see." [BCS5]

Medical students indicated how clinical rotations helped them put SP training into practice under the supervision of attending physicians. They learned by observing their attending physicians talk to patients. For example, one medical student commented on the cultural sensitivity they observed (Table 2). Medical students also indicated limitations in honing their communication skills during rotations. The practice of delivering the news of serious diagnoses to patients varied across clinical settings. One student assigned to a "county hospital site" had a lot of experience due to the volume of cases and the limited capacity of attending physicians to oversee them. Similar to medical students, PCPs recalled protocols for delivering bad news, such as "SPIKES" [32,33] and advance preparation, build a therapeutic environment or relationship, communicate well, deal with patient and family reactions, and encourage and validate emotions ("ABCDE") [35], but they reinforced the notion that classroom preparation was limited in its ability to prepare medical students to engage in difficult conversations with patients.

Openness to New Types of Communication Skills Training

The context of our interviews to inform the development of a VSP gave us opportunities to ask students about their exposure to virtual training and interest in 2D versus virtual reality options. Medical students had limited exposure but recognized its potential to improve communication skills training (Table 2).

Patient Journey Through the Breast Cancer Screening Process

The PCP and BCS interviews helped us understand when PCPs engaged with patients in breast cancer screening and diagnostic processes and how PCPs typically communicated with patients

XSL•FO
RenderX

(ie, through the telephone, a videoconference call, or in-person visit) to develop a realistic training scenario.

PCPs reported limited patient contact during the early stages of the breast cancer screening process (Table 2). Electronic health record systems sent automated messages to patients that prompted them to schedule screenings, informed them about screening test results, and prompted patients to schedule additional follow-up examinations when needed. PCPs talked to patients if additional examination was needed beyond a screening mammogram when a cancer diagnosis was more likely and patient concern was higher, for example, for a biopsy.

We also discovered that PCPs typically communicated with patients via a message in the electronic health record patient portal or a telephone call during the breast cancer screening process. PCPs scheduled in-person or video visits to deliver more serious news to the patient, such as a mammogram with a higher probability of a cancer diagnosis.

Interview discussions also shaped how we prompted the VSP to respond to learners. Given the emphasis on the SPIKES protocol by medical students and PCPs as a guide for presenting difficult news to patients and the alignment of communication-related interview themes with SPIKES (Table 3), we prompted our VSP to expect communication based on the SPIKES protocol.

Final Design Specifications for the Virtual Simulation and VSP

User Interface and Virtual Environment

The simulation portrays a scenario where a patient has had a diagnostic mammogram due to abnormal screening mammogram results. The patient, Olivia Patterson, spoke with a radiologist about her diagnostic mammogram results, and a biopsy was recommended for further testing and diagnosis based on a suspicious mass revealed by the diagnostic mammogram. The learner, taking on the role of the patient's PCP, is tasked with discussing the results of the diagnostic mammogram (which the patient had already viewed in their patient portal) and encouraging the VSP to proceed with the next step of treatment (scheduling a biopsy) through a simulated telephone call. The VSP displays behaviors consistent with common feelings of anxiety, worry, and uncertainty described by BCS interviewees who had similar discussions with their physicians.

The virtual simulation environment uses a simple 3D virtual object to depict a cell phone lying on a physician's office desk (Figure 3). The scenario of a simulated phone call represents the typical mode that PCPs reported using for discussing with a patient's results of a diagnostic mammogram with a suspicious mass. There is an "End" button on the virtual cell phone object displayed on the screen. Users click the red "End" button when the conversation with the VSP concludes. When clicked, this button simulates hanging up the phone, triggers the end of the scenario, and begins generating feedback based on the user's performance.

Figure 3. Screenshot of the virtual environment for our simulated phone call with Olivia Patterson, a virtual simulated patient designed to help medical students practice discussing abnormal mammogram results with their patients. The scene consists of a virtual cell phone object with a clickable "End" button, a static background image depicting a doctor's office desk, and a timer. Date: August 2024.

XSL•FO
RenderX

Additional Constraints

To improve and iterate on the VSP prompt, the study team evaluated the VSP's response to a wide range of interactions from different learner types, including a learner who purposely exhibited improper communication skills (eg, not addressing patient questions or telling the patient they likely had cancer), a learner who practiced good communication skills, and a learner who was mostly effective but missed 1 or 2 key steps. On the basis of the results of these tests, they iteratively added constraints and strict guidelines to both the patient and the feedback prompt to address unexpected or undesired behavior.

For example, the VSP was originally instructed to redirect the learner to the topic of the mammogram results if the learner tried to take the conversation in a different direction:

> *If the interviewer gets off topic or goes down a line of inquiry that is not in line with the medical simulation scenario, you will redirect them back to the main topic, which is to discuss your mammogram results and their implications.*

Due to the VSP responding negatively to legitimate user attempts to build rapport, this passage was later modified to include instructions for the VSP to politely engage in small talk if the learner attempted to ask questions about the VSP's life:

> *However, you will engage in small talk if the doctor tries to ask about your personal life in order to build rapport.*

This resulted in the VSP being more receptive to small talk and more able to engage in casual discussions without becoming upset with the user for straying off topic.

Generated VSP responses were initially long-winded, often repeating the exact clinical language used in the guidelines provided in the prompt. The VSP would often ask numerous questions in 1 response, making it difficult for the user to address all the patient's concerns in a single reply. The following guidelines and constraints were added to the top of the patient prompt to improve the realism of VSP responses and prevent overly verbose dialogue:

> *Take this scenario step by step, one question or subject at a time. Speak informally and in clear, concise sentences. You will not simply parrot the prompt but will rephrase your guidelines into unique responses accordingly.*

This change resulted in less repetition of the prompt, less verbose dialogue, and fewer multiple question responses. The guidance directing the VSP to use "informal" and "concise" responses seemed to produce more authentic patient dialogue that was less clinical in nature throughout the conversation.

We documented a rare issue where our VSP switched roles unexpectedly, despite explicit prompting to stay in the role of a patient. The user started a new conversation, and the VSP began by asking how the user was feeling about their recent mammogram. Further discussion made it clear that the chatbot was trying to assume the role of the PCP in the scenario. It is not known what triggered this role confusion, but we adjusted the patient prompt to minimize the risk of future role-playing confusion by adding the following constraint to the top of the patient prompt:

> *Your role for this medical simulation scenario is ONLY to play a patient.*

The AI feedback agent prompt initially contained the following guidelines to analyze the entire simulated conversation for signs that the user was adhering to the SPIKES protocol:

> *Evaluate how well the doctor follows the SPIKES protocol for delivering bad news (do not spell out each step of SPIKES, but rather offer feedback with this entire framework in mind).*

After repeated testing, we found that the AI feedback agent did not consistently assess the user's adherence to SPIKES accurately. For instance, the feedback generated often stated that the user had adequately followed steps, such as strategy or summary, when the tester specifically excluded any summary.

To minimize this issue, we adjusted the prompt to spell out all the steps of SPIKES more explicitly and describe the specifics of what to look for in user responses to identify whether they followed a step or not. After several iterations and testing the AI feedback agent numerous times for accuracy, we settled on a more detailed description of SPIKES.

Testing showed that breaking down SPIKES into 6 steps and including guidelines about actions to look for in a bulleted list format resulted in the AI feedback agent providing more consistently accurate assessments of whether a user was following SPIKES.

Evaluation

The study team tested and saved simulated conversations between themselves and the VSP, evaluating a range of user responses, choices, and edge cases (eg, the user appropriately followed each step of the SPIKES protocol vs missing one or more steps, varying degrees of empathy from the user, and the user straying off topic or failing to clearly convey results). Each iteration of the AI feedback agent was repeatedly tested on these saved conversations to validate the content of the generated feedback and the effectiveness of changes made to the prompt. The designer tested and evaluated 3 LLMs available in the virtual simulation software (GPT-4, GPT-3.5 [OpenAI], and Gemini [Google LLC]) before selecting GPT-4, the newest and largest model of the 3. In comparison tests, GPT-4 outperformed GPT-3.5 and Gemini in the time taken to generate responses, role-playing ability, and validity of generated feedback.

After the initial prototype VSP was created and tested, experienced physicians (including a PCP and an oncologist) and medical educators from the research team analyzed VSP conversation and feedback transcripts to evaluate relevance and accuracy compared to real patients and SPs. Both VSP and feedback agent prompts were then refined further based on these evaluations.

Preliminary testing showed that the VSP appropriately asked questions, responded to user inquiries, and displayed emotions that matched user responses (eg, sounding worried at the outset of the call and sounding upset if the user did not address VSP questions). In tests where the user adhered to the SPIKES

XSL•FO
RenderX

protocol, patient responses conveyed less anxiousness, and the VSP was more receptive to the user's suggestions and explanations. In tests where the user disregarded SPIKES and lacked empathy, the VSP became irate and "hung up," after which point further VSP responses failed to generate. Text-to-speech software provided a patient voice with inflections that matched the emotional tone of the AI-generated text.

We also considered the feedback mechanism to be a successful exploratory feature of the simulation. At the end of the VSP conversation, the simulation displayed qualitative feedback and allowed the user to review the contents of their entire conversation, including quoted examples from the VSP conversation and suggestions for appropriate alternative phrases. The AI feedback agent consistently provided feedback that accurately detected the presence of empathy, reassurance, clarity, and a balance of seriousness and uncertainty in user responses.

We identified several areas for improvement. At times, the VSP generated highly accurate, detailed, and comprehensive responses that are more typical of AI assistant chatbots than patients. The VSP had difficulty weighing the importance of prompt instructions to be responsive to both clear explanations and empathetic communication. It was consistently more cooperative (willing to schedule a biopsy) in response to unempathetic (although clear) dialogue from the learner, as opposed to empathetic and clear dialogue. The AI feedback agent sometimes struggled to determine whether learners followed the SPIKES protocol in their conversations and mixed valid feedback with less helpful advice. The agent critiqued a seasoned PCP because they did not convince the VSP to agree to schedule a biopsy. Afterward, the PCP told us they disagreed with the feedback because it would have been too forceful to press the VSP further to schedule the biopsy, given her agitated emotional state. In addition, the AI feedback agent provided example quotes that were often incorrectly transcribed, creating inaccurate transcripts of learner responses.

With a steady internet connection, the system took approximately 2 seconds to transcribe the user's spoken dialogue, the VSP generated responses in 1 to 3 seconds, and the AI feedback agent generated feedback text at the conclusion of the training after 30 to 90 seconds. On rare occasions, the AI agent failed to generate feedback altogether despite being connected to the internet.

Discussion

Principal Findings

The research team aimed to identify the optimal development process for AI-driven VSPs by evaluating the capabilities of a GPT-4–based simulation to help medical students practice empathically discussing abnormal mammogram results. Branching conversational paths, the focus of our original development plan, required us to anticipate plausible dialogue between learners and patients. Enabling generative AI responses for the VSP and using prompts to direct how the VSP should respond allows flexibility in how the learner converses with the VSP, reflecting a more spontaneous dialogue as would occur with a real patient. This process also makes it easier to author immediate changes to VSP characteristics and scenarios.

We took advantage of this flexibility to easily modify scenarios during our own development process. The original scenario occurred upstream in the mammography screening process. The VSP requested a telephone call with their PCP after they had abnormal screening mammogram results that required a diagnostic mammogram. To ensure the need for physicians to have a conversation about the results, formative interviews and discussions with the research team shifted the scenario downstream to what we developed—a simulated telephone call after abnormal results from a diagnostic mammogram that requires a biopsy. We easily made this change by revising the VSP prompt to describe the scenario that centered on a diagnostic mammogram with abnormal results rather than a screening mammogram.

VSP development for this study is part of a growing landscape of virtual simulations to supplement in-person manikin and SP-based simulation training. VSPs are gaining acceptance from educators and students alike, marked by the necessities of social distancing during the COVID-19 pandemic and lingering clinical teaching challenges in its wake [36]. Our development pushed the boundary of existing VSP applications that focus on routine clinical interactions with patients, such as history taking [11-15]. Our simulation focuses on the delivery of uncertain (and potentially bad) news that accompanies a diagnostic mammogram requiring additional follow-up. The study by Webb [37] is a notable exception, testing the ability of GPT-3.5 to train emergency room physicians to deliver the news of serious diagnoses to patients. Our work advances the work of Webb [37] and other studies that use text-based communication by simulating verbal communication between users and VSP. This is an important advancement in virtual simulations that allows them to mimic in-person simulations more closely with SPs and real-life clinical interactions with patients.

Reflecting on our move from a BPS design to one using GPT-4, we note the value in the conceptual framework of a BPS design process that is driven by input from clinicians and patients. In-depth interviews with BCSs and PCPs helped us understand patient journeys through the breast cancer screening process. In turn, this helped us generate and map plausible conversations between PCPs and patients so we could plan out the simulation scenario and eventually inform our VSP prompt. Mapping likely conversations also helped us evaluate the performance of the GPT-4–powered VSP to determine if it responded how we expect a patient to respond. We recommend that developers start building virtual scenarios by collecting qualitative data from relevant clinicians and patients in the spirit of a BPS and map out likely conversation paths before developing AI-powered VSP prompts.

A key feature of the training was the ability to display GPT-4–generated written feedback to the learner at the end of the simulation, similar to the qualitative feedback that SPs and clinical instructors provide to medical students during their education. Feedback is a key element of training programs, as underscored by MS4b in noting, "one of the most valuable parts (of SP training) was having the standardized patients give their

XSL•FO
RenderX

feedback at the end." In some ways, GPT-4 outperformed our expectations in providing feedback, especially in the way it captured the overall tone of the conversation (eg, if the user was empathetic), but it also left room for improvement, as discussed subsequently with its limitations. We will retain the feedback feature of the simulation and continue to refine its capabilities commensurate with LLM advances. In addition, future studies comparing AI feedback to expert clinician educator feedback will allow us to further refine prompts. At the current time, our solution is to offer a caveat to learners in this beta phase to consider the feedback and understand that their own clinical judgment may supersede suggested communication techniques.

Limitations

Some of the limitations we encountered will be addressed through prompt refinement, newer LLMs, and faster computing speeds. Future LLMs will likely decrease AI feedback transcription errors and response delays. Preliminary tests showed that using GPT-4o, OpenAI's newest LLM, reduced the delay for the AI feedback agent to generate text down to around 10 to 15 seconds.

Other limitations that are artifacts of the unpredictability of LLMs may persist in future LLMs and will be harder to address. We developed the VSP and AI feedback agent prompts using a heuristic, iterative process that relied on trial and error to improve the quality of simulated conversations and generated feedback. What we identified as improvements we made to our VSP may have been random conversation variations that we happened to prefer and not a direct result of a change we made to the prompt text. The prompts provided guidance to the VSP that typically prevented it from going off topic and out of context or producing nonsensical responses (ie, *hallucinating*) during conversations with users, but unexpected results were observed on several occasions. These errors were difficult to predict, such as when the chatbot shifted roles to play the physician instead of the patient. We also found it difficult to force the VSP to equally value clarity and empathy when determining how to behave cooperatively with the user, as indicated in the Results section.

It was also difficult to determine the criteria used by the AI feedback agent for assessing user performance. For instance, testing the AI feedback agent on the same conversation text multiple times resulted in occasional differences in the assessment of the user's adherence to SPIKES. Furthermore, the AI feedback agent had difficulty weighing the importance of different aspects of the SPIKES protocol, in contrast to the experienced PCP who showed empathy by not pressuring the "anxious" VSP to agree to a biopsy immediately. This reflects GPT-4's limitations in dispensing clinical advice that relies more on clinical experience than can be gleaned from existing documentation [38]. Finally, VSP dialogue and generated feedback are not standardized, in contrast to SP and clinical educators in traditional medical simulation. In its current form, it is unlikely that an LLM-driven VSP training could be used as a part of summative clinical examinations or assessments of learner communication skills due to this lack of standardization.

Limitations highlight the need to evaluate the robustness of the simulation on a larger scale. Researchers have proposed evaluation frameworks such as the automated interactive evaluation framework and Artificial Intelligence Structured Clinical Examinations to evaluate the performance of LLMs to carry out clinical tasks that can be applied in this regard [39,40]. In this vein, the next step will be a pilot test of the virtual simulation module with medical students to determine its feasibility and acceptability. The unpredictability of LLMs also offers opportunities to revolutionize and transform "standardized testing" and to question the notion that standardized tests are the best way to evaluate competency when no real patient is the same as the previous or next patient. LLMs provide a mechanism to evaluate competency through a variety of patients and different types of interactions to provide a more holistic view of proficiency in communication skills.

Future Directions

We scratched the surface for the types of conversations that PCPs and other clinicians have with their patients that could benefit from communication skills training through a virtual simulation with a VSP. We focused on training for medical students, but the scenario and the type of feedback users receive could be tailored for learners at different stages of their education, such as medical students versus residents, or in different professions, such as nurses, social workers, and administrative clerks interacting with patients along any aspect of the patient's clinical experience. Our simulated conversation focused on diagnostic uncertainty surrounding breast cancer screenings. Conversations for all life-changing diagnoses and stigmatizing health conditions (eg, HIV, substance use disorders, and obesity) and associated screening tests warrant careful consideration, even for common diagnoses. PCP2 discussed how type 2 diabetes diagnoses are routine for PCPs but are viewed by some patients as a "death sentence." Aside from the diagnoses themselves, changing treatment plans can be unsettling for patients. For example, PCP4 mentioned a challenging conversation as one "where it goes against what a patient wants," such as discontinuing pain medication prescriptions.

Our simulation focused on a single conversation between a learner and a VSP within the context of other clinical conversations, such as the one the VSP had with the radiologist before speaking with the learner in the simulation. During interviews, PCPs shared the importance of setting patient expectations within the context of coordinated care; for example, "We have other people who are going to be involved" (PCP4). This was reflected in BCS's discussions that highlighted the need for improved communication with different types of clinicians throughout their patient journey, such as radiology technicians and radiologists. Interprofessional education can improve coordinated care and communication with patients but is hampered by the same type of clinical demands that limit individual and team communication skills training. Virtual simulations can lower training barriers, such as the need for multiple learners to interact in person. For example, Liaw et al [41] developed a simulation where multiple learners in different locations can interact in the same simulation to practice care coordination for an older VSP. In that study, debriefing was not automatic and instead was facilitated by clinician educators. Future iterations could test the limits of LLMs by incorporating

XSL•FO
RenderX

multiplayer characters involving >2 people in a dialogue, for example, adding a family member or having 2 health care professionals in a patient encounter.

Another consideration is the decision to use a cloud-based LLM versus LLMs that can be downloaded to local computers. Our simulation incorporated GPT-4, a cloud-based LLM that ran on standard computers relative to local LLMs that would have required additional computing resources. While there are concerns about the privacy and security of data sent to cloud-based LLMs, our prompt uses a fictional patient and a role-playing scenario to minimize the risk of any sensitive or private data being shared with the chatbot. We wrote the prompt for our fictional patient Olivia based loosely on formative interviews with anonymous patients and did not include any protected patient information or restricted data. We intend to notify participants in the pilot study about the potential data privacy risks associated with GPT-4 and will request that they acknowledge and agree not to input any protected or private information. Future studies could incorporate secure local LLMs trained on anonymized transcripts of conversations between clinicians and patients to mitigate the potential data privacy risks of using cloud-based LLMs.

Conclusions

Our work contributes to a rapidly changing medical simulation landscape driven by advancements in LLMs that use generative AI algorithms to mimic human responses to text and voice queries. Our project showcases 2 promising applications for GPT-4 in its ability to streamline the development of a simulated phone call between a learner and VSP and provide accurate AI-generated feedback to the learner. While GPT-4 displayed limitations in its ability to provide nuanced feedback about learners' performance in following best communication practices, the training simulation consistently performed its main task well by providing an asynchronous opportunity for medical students to practice a challenging conversation with a patient. Because every attempt elicits slightly varied responses, it is possible to have multiple opportunities for deliberate practice, unlike traditional branching path scenarios that have smaller, more finite correct answers. Given the rapid advances in LLM to date, we are encouraged about the potential to improve our current training simulation with future LLM improvements and produce more complex scenarios. OpenAI is already proclaiming that GPT-5 will have "PhD level intelligence" [42]. While we cannot comment on the validity of that statement, we feel confident proclaiming that the future of medical simulation is bright.

Acknowledgments

This work was supported by a National Cancer Institute administrative supplement through the Jonsson Comprehensive Cancer Center at the University of California, Los Angeles (3P30CA016042-47S3). PAG was supported in part by the Breast Cancer Research Foundation.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
In-depth interview guide questions for medical students, primary care physicians, and breast cancer survivors who participated in the study to discuss the benefits and limitations of current communication skills training practices, interest in virtual communication skills training, and current modes of communication between primary care physicians and their patients to discuss mammography screening examination results.
[DOCX File , 17 KB - formative_v9i1e65670_app1.docx]

References

1. Bair H, Norden J. Large language models and their implications on medical education. Acad Med 2023 Aug 01;98(8):869-870. [doi: 10.1097/ACM.0000000000005265] [Medline: 37162220]
2. Haruna-Cooper L, Rashid MA. GPT-4: the future of artificial intelligence in medical school assessments. J R Soc Med 2023 Jun;116(6):218-219 [FREE Full text] [doi: 10.1177/01410768231181251] [Medline: 37318843]
3. Wójcik S, Rulkiewicz A, Pruszczyk P, Lisik W, Poboży M, Domienik-Karłowicz J. Beyond ChatGPT: what does GPT-4 add to healthcare? The dawn of a new era. Cardiol J 2023;30(6):1018-1025 [FREE Full text] [doi: 10.5603/cj.97515] [Medline: 37830256]
4. Chen S, Guevara M, Moningi S, Hoebers F, Elhalawani H, Kann BH, et al. The effect of using a large language model to respond to patient messages. Lancet Digit Health 2024 Jun;6(6):e379-e381. [doi: 10.1016/s2589-7500(24)00060-8]

XSL•FO
RenderX

5. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023 Jun 01;183(6):589-596 [FREE Full text] [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
6. Ha JF, Longnecker N. Doctor-patient communication: a review. Ochsner J 2010;10(1):38-43 [FREE Full text] [Medline: 21603354]
7. Tavakoly Sany SB, Behzhad F, Ferns G, Peyman N. Communication skills training for physicians improves health literacy and medical outcomes among patients with hypertension: a randomized controlled trial. BMC Health Serv Res 2020 Jan 23;20(1):60 [FREE Full text] [doi: 10.1186/s12913-020-4901-8] [Medline: 31973765]
8. Adnan AI. Effectiveness of communication skills training in medical students using simulated patients or volunteer outpatients. Cureus 2022 Jul 10;14(7):e26717 [FREE Full text] [doi: 10.7759/cureus.26717] [Medline: 35967150]
9. Reidy JA, Clark MA, Berman HA, Chan SH, Gawande AA, Streid J, et al. Paving the way for universal medical student training in serious illness communication: the Massachusetts Medical Schools' Collaborative. BMC Med Educ 2022 Sep 01;22(1):654 [FREE Full text] [doi: 10.1186/s12909-022-03702-2] [Medline: 36050708]
10. Deveugele M, Derese A, De Maesschalck S, Willems S, Van Driel M, De Maeseneer J. Teaching communication skills to medical students, a challenge in the curriculum? Patient Educ Couns 2005 Sep;58(3):265-270. [doi: 10.1016/j.pec.2005.06.004] [Medline: 16023822]
11. Holderried F, Stegemann-Philipps C, Herschbach L, Moldt JA, Nevins A, Griewatz J, et al. A generative pretrained transformer (GPT)-powered chatbot as a simulated patient to practice history taking: prospective, mixed methods study. JMIR Med Educ 2024 Jan 16;10:e53961 [FREE Full text] [doi: 10.2196/53961] [Medline: 38227363]
12. Holderried F, Stegemann-Philipps C, Herrmann-Werner A, Festl-Wietek T, Holderried M, Eickhoff C, et al. A language model-powered simulated patient with automated feedback for history taking: prospective study. JMIR Med Educ 2024 Aug 16;10:e59213 [FREE Full text] [doi: 10.2196/59213] [Medline: 39150749]
13. Jiang Z, Huang X, Wang Z, Liu Y, Huang L, Luo X. Embodied conversational agents for chronic diseases: scoping review. J Med Internet Res 2024 Jan 09;26:e47134 [FREE Full text] [doi: 10.2196/47134] [Medline: 38194260]
14. Potter L, Jefferies C. Enhancing communication and clinical reasoning in medical education: building virtual patients with generative AI. Future Healthc J 2024 Apr;11:100043 [FREE Full text] [doi: 10.1016/j.fhj.2024.100043]
15. Sardesai N, Russo P, Martin J, Sardesai A. Utilizing generative conversational artificial intelligence to create simulated patient encounters: a pilot study for anaesthesia training. Postgrad Med J 2024 Mar 18;100(1182):237-241. [doi: 10.1093/postmj/qgad137] [Medline: 38240054]
16. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: can language models be too big? ???? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 2021 Presented at: FAccT '21; March 3-10, 2021; Virtual Event. [doi: 10.1145/3442188.3445922]
17. Cooper A, Rodman A. AI and medical education - a 21st-century Pandora's box. N Engl J Med 2023 Aug 03;389(5):385-387 [FREE Full text] [doi: 10.1056/NEJMp2304993] [Medline: 37522417]
18. Zack T, Lehman E, Suzgun M, Rodriguez JA, Celi LA, Gichoya J, et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. Lancet Digit Health 2024 Jan;6(1):e12-e22. [doi: 10.1016/s2589-7500(23)00225-x]
19. Liu X, Wu C, Lai R, Lin H, Xu Y, Lin Y, et al. ChatGPT: when the artificial intelligence meets standardized patients in clinical training. J Transl Med 2023 Jul 06;21(1):447 [FREE Full text] [doi: 10.1186/s12967-023-04314-0] [Medline: 37415217]
20. McGovern MM, Johnston M, Brown K, Zinberg R, Cohen D. Use of standardized patients in, undergraduate medical genetics education. Teach Learn Med 2006;18(3):203-207. [doi: 10.1207/s15328015tlm1803_3] [Medline: 16776606]
21. Nelson G. Training standardized patients to provide effective feedback: development, implementation, and its effect on the efficacy of medical students' education. S D Med 2022 Oct;75(10):454-455. [Medline: 36889265]
22. Li Y, Zeng C, Zhong J, Zhang R, Zhang M, Zou L. Leveraging large language model as simulated patients for clinical education. arXiv. Preprint posted online on April 13, 2024 [FREE Full text]
23. Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. CA Cancer J Clin 2023 Jan 12;73(1):17-48 [FREE Full text] [doi: 10.3322/caac.21763] [Medline: 36633525]
24. Ho TQ, Bissell MC, Kerlikowske K, Hubbard RA, Sprague BL, Lee CI, et al. Cumulative probability of false-positive results after 10 years of screening with digital breast tomosynthesis vs digital mammography. JAMA Netw Open 2022 Mar 01;5(3):e222440 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.2440] [Medline: 35333365]
25. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature 2023 Aug 12;620(7972):172-180 [FREE Full text] [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
26. Bojic L, Kovacevic P, Cabarkapa M. GPT-4 surpassing human performance in linguistic pragmatics. arXiv. Preprint posted online on December 15, 2023 [FREE Full text]
27. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. arXiv. Preprint posted online on March 20, 2023 [FREE Full text]

XSL•FO
RenderX

28. Oh N, Choi GS, Lee WY. ChatGPT goes to the operating room: evaluating GPT-4 performance and its potential in surgical education and training in the era of large language models. Ann Surg Treat Res 2023 May;104(5):269-273 [FREE Full text] [doi: 10.4174/astr.2023.104.5.269] [Medline: 37179699]
29. Rosoł M, Gąsior JS, Łaba J, Korzeniewski K, Młyńczak M. Evaluation of the performance of GPT-3.5 and GPT-4 on the Polish medical final examination. Sci Rep 2023 Nov 22;13(1):20512 [FREE Full text] [doi: 10.1038/s41598-023-46995-z] [Medline: 37993519]
30. Gallifant J, Afshar M, Ameen S, Aphinyanaphongs Y, Chen S, Cacciamani G, et al. The TRIPOD-LLM reporting guideline for studies using large language models. Nat Med 2025 Jan 08;31(1):60-69. [doi: 10.1038/s41591-024-03425-5] [Medline: 39779929]
31. HyperSkill for VR training. SimInsights. URL: https://www.siminsights.com/hyperskill/ [accessed 2025-04-07]
32. Baile WF, Buckman R, Lenzi R, Glober G, Beale EA, Kudelka AP. SPIKES-a six-step protocol for delivering bad news: application to the patient with cancer. Oncologist 2000;5(4):302-311 [FREE Full text] [doi: 10.1634/theoncologist.5-4-302] [Medline: 10964998]
33. Buckman RA. Breaking bad news: the S-P-I-K-E-S strategy. Community Oncol 2005 Mar;2(2):138-142. [doi: 10.1016/S1548-5315(11)70867-1]
34. Ruan J, Chen Y, Zhang B, Xu Z, Bao T, Du G, et al. TPTU: large language model-based AI agents for task planning and tool usage. arXiv. Preprint posted online on August 7, 2023 [FREE Full text]
35. Rabow MW, McPhee SJ. Beyond breaking bad news: how to help patients who suffer. West J Med 1999 Oct;171(4):260-263 [FREE Full text] [Medline: 10578682]
36. De Mattei L, Morato MQ, Sidhu V, Gautam N, Mendonca CT, Tsai A, et al. Are artificial intelligence virtual simulated patients (AI-VSP) a valid teaching modality for health professional students? Clin Simul Nurs 2024 Jul;92:101536. [doi: 10.1016/j.ecns.2024.101536]
37. Webb J. Proof of concept: using ChatGPT to teach emergency physicians how to break bad news. Cureus 2023 May;15(5):e38755 [FREE Full text] [doi: 10.7759/cureus.38755] [Medline: 37303324]
38. Jo E, Song S, Kim JH, Lim S, Kim JH, Cha JJ, et al. Assessing GPT-4's performance in delivering medical advice: comparative analysis with human experts. JMIR Med Educ 2024 Jul 08;10:e51282 [FREE Full text] [doi: 10.2196/51282] [Medline: 38989848]
39. Liao Y, Meng Y, Wang Y, Liu H, Wang Y, Wang Y. Automatic interactive evaluation for large language models with state aware patient simulator. arXiv. Preprint posted online on March 13, 2024 [FREE Full text] [doi: 10.2139/ssrn.4890649]
40. Mehandru N, Miao BY, Almaraz ER, Sushil M, Butte AJ, Alaa A. Evaluating large language models as agents in the clinic. NPJ Digit Med 2024 Apr 03;7(1):84 [FREE Full text] [doi: 10.1038/s41746-024-01083-y] [Medline: 38570554]
41. Liaw SY, Wu LT, Soh SL, Ringsted C, Lau TC, Lim WS. Virtual reality simulation in interprofessional round training for health care students: a qualitative evaluation study. Clin Simul Nurs 2020 Aug;45:42-46. [doi: 10.1016/j.ecns.2020.03.013]
42. Dane L. GPT-5 will have Ph.D level intelligence. Lindane.Co. 2024 Jul 3. URL: https://lindane.co/blog/gpt-5-will-have-phd-level-intelligence/ [accessed 2025-04-06]

Abbreviations

ABCDE: advance preparation, build a therapeutic environment or relationship, communicate well, deal with patient and family reactions, and encourage and validate emotions
AI: artificial intelligence
BCS: breast cancer survivor
BPS: branching path simulation
LLM: large language model
PCP: primary care physician
SP: standardized patient
SPIKES: Setting, Perception, Invitation, Knowledge, Emotion, and Summarize
TRIPOD: Transparent Reporting of a Multivariable Model for Individual Prognosis or Diagnosis
VSP: virtual simulated patient

XSL•FO
RenderX

Edited by A Mavragani; submitted 27.08.24; peer-reviewed by S Chan, AJ Doyle; comments to author 13.12.24; revised version received 21.01.25; accepted 14.02.25; published 17.04.25.

Please cite as:
Weisman D, Sugarman A, Huang YM, Gelberg L, Ganz PA, Comulada WS
Development of a GPT-4–Powered Virtual Simulated Patient and Communication Training Platform for Medical Students to Practice Discussing Abnormal Mammogram Results With Patients: Multiphase Study
JMIR Form Res 2025;9:e65670
URL: https://formative.jmir.org/2025/1/e65670
doi:10.2196/65670
PMID:

©Dan Weisman, Alanna Sugarman, Yue Ming Huang, Lillian Gelberg, Patricia A Ganz, Warren Scott Comulada. Originally published in JMIR Formative Research (https://formative.jmir.org), 17.04.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

User Experience of and Adherence to a Smartphone App to Maintain Behavior Change and Self-Management in Patients With Work-Related Skin Diseases: Multistep, Single-Arm Feasibility Study

Nele Ristow[1,2], MEd; Annika Wilke[1,2], PD Dr; Christoph Skudlik[1,2], Prof Dr Med; Swen Malte John[1,2], Prof Dr Med; Michaela Ludewig[3], Prof Dr

[1]Department of Dermatology, Environmental Medicine and Health Theory, Institute for Health Research and Education, Osnabrück University, Osnabrück, Germany

[2]Institute for Interdisciplinary Dermatological Prevention and Rehabilitation (iDerm) at the Osnabrück University, Osnabrück, Germany

[3]Department Health Sciences, Hochschule Bochum, Bochum, Germany

Corresponding Author:
Nele Ristow, MEd
Department of Dermatology, Environmental Medicine and Health Theory
Institute for Health Research and Education
Osnabrück University
Am Finkenhügel 7a
Osnabrück, 49076
Germany
Phone: 49 541 969 7410
Fax: 49 541 969 2445
Email: neristow@uos.de

Abstract

Background: Smartphone apps are a growing field supporting the prevention of chronic diseases. The user experience (UX) is an important predictor of app use and should be considered in mobile health research. Long-term skin protection behavior is important for those with work-related skin diseases. However, altering health behavior is complex and requires a high level of self-management. We developed a maintenance program consisting of the Mein Hautschutz im Alltag (MiA; “My skin protection in everyday life”) app combined with an individual face-to-face goal-setting interview to support patients in the implementation of skin protection behavior after inpatient rehabilitation.

Objective: The objectives of this paper are to (1) describe the intervention in a standardized manner; (2) evaluate the UX, subjective quality, and perceived impact of the MiA app; and (3) evaluate the adherence to the MiA app.

Methods: We followed a user-centered and multistage iterative process in 2 steps that combined qualitative and quantitative data. The maintenance program was tested over 12 weeks after discharge from rehabilitation. The UX, subjective quality, and perceived impact were evaluated formatively based on the user version of the Mobile Application Rating Scale after 12 weeks (T2). Adherence was measured using the frequency of interactions with the app.

Results: In total, 42 patients took part (with a dropout rate of n=18, 43% at T2). The average age was 49.5 (SD 13.1) years, and 57% (24/42) were male. We found high ratings for the UX, with an average score of 80.18 (SD 8.94) out of a theoretical maximum of 100, but there were a few exceptions in the usability and interaction with the app. The app was most frequently rated with 4 out of 5 stars (15/24, 65%), which indicates a high subjective quality. Furthermore, the app seemed to influence important determinants to implement skin protection behavior. Adherence to skin protection tracking was higher over the study period than adherence to skin documentation and goal assessment. The number of adherent participants to skin protection tracking was higher in the skin care and skin cleansing categories (28/42, 67% each) compared to the skin protection category (13/42, 31%) on day 1 and decreased until day 84 in all dimensions (12/42, 29% each for skin care and skin cleansing; 9/42, 21% for skin protection).

Conclusions: The results in terms of adherence met the expectations and were consistent with those of other studies evaluating the use of apps for chronic diseases. Interaction with the app could be increased using artificial intelligence to determine eczema severity via photos. It should be investigated which subgroups have difficulties with usability to individualize the support to a

XSL•FO
RenderX

greater degree during onboarding. There is a need for further research regarding the effectiveness of the MiA app on skin protection behavior, quality of life, and eczema severity.

(JMIR Form Res 2025;9:e66791) doi:10.2196/66791

KEYWORDS

user experience; mobile health; mHealth; app; smartphone; complex intervention; Template for Intervention Description and Replication; behavior change techniques; behavior change; skin diseases; occupational dermatology; artificial intelligence

Introduction

Background

Smartphone apps have recently become a growing field in the health care sector, garnering increasing attention and importance in health research. The use of such technologies can help optimize broad and location-independent health care in real time and support the prevention of chronic diseases and health promotion to improve adherence to therapy, quality of life, and clinical outcomes in the long term [1,2]. Smartphone apps not only help monitor health data [3,4] but also help change health behavior and develop new habits. For this purpose, health behavior apps offer the potential to technically integrate a variety of behavior change techniques (BCTs) [1,5,6]. BCTs are based on health psychology knowledge about the mechanisms of behavior change processes [7]. The most common BCTs in behavior change apps are, for example, self-monitoring by tracking the duration and frequency of a behavior and feedback on the behavior or outcomes by visualizing the progress graphically [1,5,6,8]. Feedback and the visualization of success also increase motivation in the users, which plays an important role in the long-term success of behavior change [5]. Behavior change apps can also be used to set goals and monitor their achievement. Automatic reminders provide prompts and support the development of routines in everyday life. Written, visual, and audio information can show users how to perform the behavior [1,5,6,8].

Relevance of User Experience

A known and widely discussed problem in research on behavior change apps is a low adherence and engagement when using these technologies [2,9-11]. To address these problems, user experience (UX) is becoming increasingly important in mobile health research [12-15]. UX is an overarching holistic concept and covers all emotional, cognitive, and physical experiences that users undergo when interacting with software products such as apps [12,16-18]. As shown in Figure 1, a high UX when using an app for the first time is associated with the willingness to continue using it and recommend it to others in the short term. A high UX leads to increased adherence and app use in the long term and, thus, can also improve clinical outcomes and the effectiveness of health-related interventions [14,19-21]. Therefore, the investigation of UX in the context of digital technologies in patient care is of great importance and should be considered early during the development process [14,15,22]. Thus, a thorough formative evaluation of the UX should precede a summative evaluation to optimize the app from the end user's perspective [14].

Figure 1. Theoretical model to explain the influence of user experience (UX) and its 5 dimensions—content, usability, aesthetic, trust, and feelings and emotions—on the intention and recommendation to use the app (proximal) results, as well as the revisiting the app and long-term use of the app (distal) results as the basis for this study.

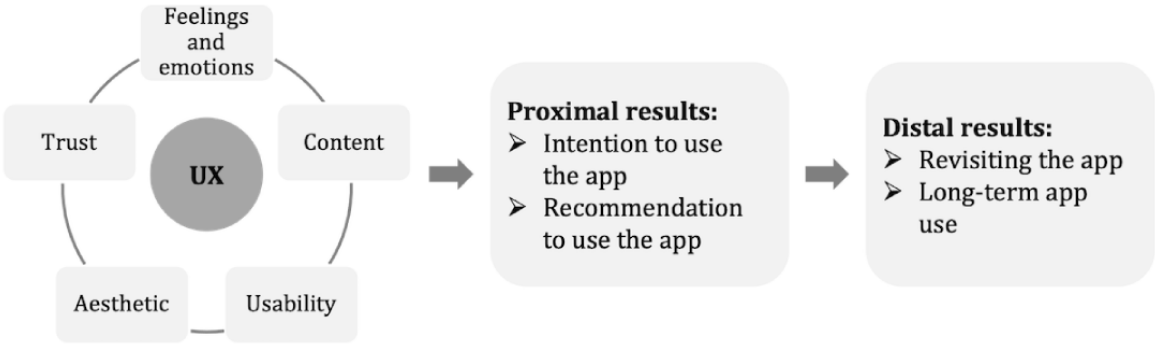

As a multidimensional construct, UX comprises various dimensions that can differ depending on the underlying model, understanding, and context [15,17,19]. According to Thielsch and Salaschek [21], the dimensions *content*, *usability*, and *aesthetics* are particularly important in a research-related context. In addition, *feelings and emotions* that arise through interaction with the technical medium are described as an important aspect of UX [15,16]. As personal data are collected, passed on, and stored in digital care structures, *trust* in the technology is another predictor of future app use [21].

Behavior Change and Self-Management Apps for Patients With Chronic Skin Diseases

Apps to support behavior change and disease-specific self-management have also been developed and evaluated in the context of chronic skin diseases. These are usually aimed at specific groups, such as parents of children with atopic dermatitis (AD) [23,24]; patients with AD [25,26]; or people with specific areas of the skin affected, such as hand and foot eczema [27]. Weigandt et al [27] developed the first smartphone app in Germany specifically for hand and foot eczema, with

XSL•FO
RenderX

which patients can monitor and manage their disease following a patient educational intervention by photographing their skin, tracking their quality of life and symptoms, and chatting with their dermatologist.

However, thus far, such technology does not exist for work-related skin diseases (WRSDs) even though they have been among the most common work-related diseases in Germany for decades [28]. Up to 90% of all diagnosed cases affect the hands in the form of irritant or allergic contact eczema, which often occurs in combination with a genetic predisposition to AD [29-32]. Work-related hand eczema is usually caused by hazardous activities in the workplace (eg, skin contact with allergens and irritants, wet work, high handwashing frequencies, and mechanical stress). Workers in the health care, metalworking, hairdressing, and construction sectors are at a particularly high risk, which results in high prevalence rates [30,31]. Due to the high relevance of WRSDs, the German Social Accident Insurance has developed a complex and hierarchical multistep procedure for patients with WRSDs, with different outpatient and inpatient prevention measures depending on the severity of the skin disease. In the case of severe or recurrent skin diseases, patients are offered to participate in an inpatient interprofessional rehabilitation program with a duration of 3 weeks. In German occupational dermatology, this program is also known as tertiary individual prevention (TIP) [29,33]. An important element of TIP are health educational interventions [34] with the aim of gaining disease-specific knowledge and increasing the motivation to implement and optimize individual skin protection behavior. In addition, individual strategies for skin protection behavior are developed with occupational therapists and practiced as part of a workplace simulation in occupational therapy. To restore the skin barrier and reduce the risk of recurrence of hand eczema, it is important to implement skin protection behavior (eg, regular use of skin protection and skin care products as well as reducing the frequency of handwashing by using mild detergents) in the long term [34-36]. However, changing skin protection behavior is complex and requires a high level of self-management after participating in the TIP as the new behavior must be successfully transferred to and implemented in the professional and private contexts. A structured maintenance program that supports patients in this subsequent implementation of skin protection behavior does not yet exist [34]. To fill this gap, we developed the Mein Hautschutz im Alltag (MiA) app (MiA translates to “My skin protection in everyday life”) to support the self-management of patients with WRSDs after the TIP [37].

Objectives of This Study

The aim of this study was to evaluate the UX of and adherence to the MiA app and pilot the feasibility of an app-based maintenance program in our clinical setting. In this publication, we report the following results: (1) description of the intervention in a standardized manner; (2) UX, subjective quality, and perceived impact of the MiA app; and (3) adherence to the MiA app.

Methods

Overview

The occupational dermatology maintenance program is a complex intervention consisting of several interacting components (Multimedia Appendix 1). According to the Medical Research Council’s framework, the development of complex interventions is based on a 4-stage process consisting of development [37], feasibility, evaluation, and implementation [38]. This study focuses on feasibility (phase 2).

Intervention

The maintenance program is offered by the Institute for Interdisciplinary Dermatological Prevention and Rehabilitation (iDerm), Osnabrück, Germany, which is a specialized center in Germany for inpatient and outpatient interprofessional treatment of patients with WRSDs.

The maintenance program consists of 2 main elements: *individual goal-setting interview* and the MiA app. The systematic development of these elements is described in detail elsewhere [37].

The *individual goal-setting interview* is conducted during the inpatient stay by a health educator face-to-face 5 or 6 days before discharge. The goals, which are defined by the patients, are entered into the app on the coaching platform by the health educator and can be viewed by patients via the app. During the subsequent onboarding, patients receive an introduction and explanations of the various app functions. Patients are also given access to a video via a QR code that shows the app with screenshots and explains the app’s functions.

MiA is a fully automated app with free access for study participants and consists of 6 components focusing either on interaction or information. The smartphone-based components with interactive nature are named My Skin Protection Goals (German: *Meine Hautschutz-Ziele*), My Skin Protection Behavior (German: *Mein Hautschutz-Verhalten*), and My Skin Documentation (German: *Meine Hautdokumentation*). The smartphone-based components To Listen (German: *Hörenswert*), Skin Protection 101 (German: *Das Hautschutz 1x1*), and My Accountabilities (German: *Meine Zuständigkeiten*) provide different information about the skin disease in different modes of delivery (podcasts, textboxes, and videos; Figure 2). In addition, users automatically receive a reminder message if no information has been entered into the app for a period of 14 days.

XSL•FO
RenderX

Figure 2. (A) Overview of the menu; (B) overview of the My Skin Protection Behavior component with the option to enter the frequencies of skin cleansing (green line) and applications of skin care cream (red line) and skin protection cream (blue line); (C) overview of the My Skin Documentation component, which allows for taking pictures of the front and back side of the right and left hand that can be enlarged and to which personal notes can be added; (D) overview of the My Skin Protection Goals component with the overview of the goal achievement; (E) To Listen component and 4 podcast episodes; (F) overview of the Skin Protection 101 component with different questions and an example unfolded answer; (G) overview of the My Accountabilities component with different questions and an exemplary unfolded answer; and (H) presentation of the research questionnaire with the 6 categories, whereby categories 1 to 3 (sociodemographic data, skin protection behavior, and questions about the organization) can be completed immediately and categories 4 to 6 (skin protection behavior, app evaluation, and impact of the app) are activated after 12 weeks of use.

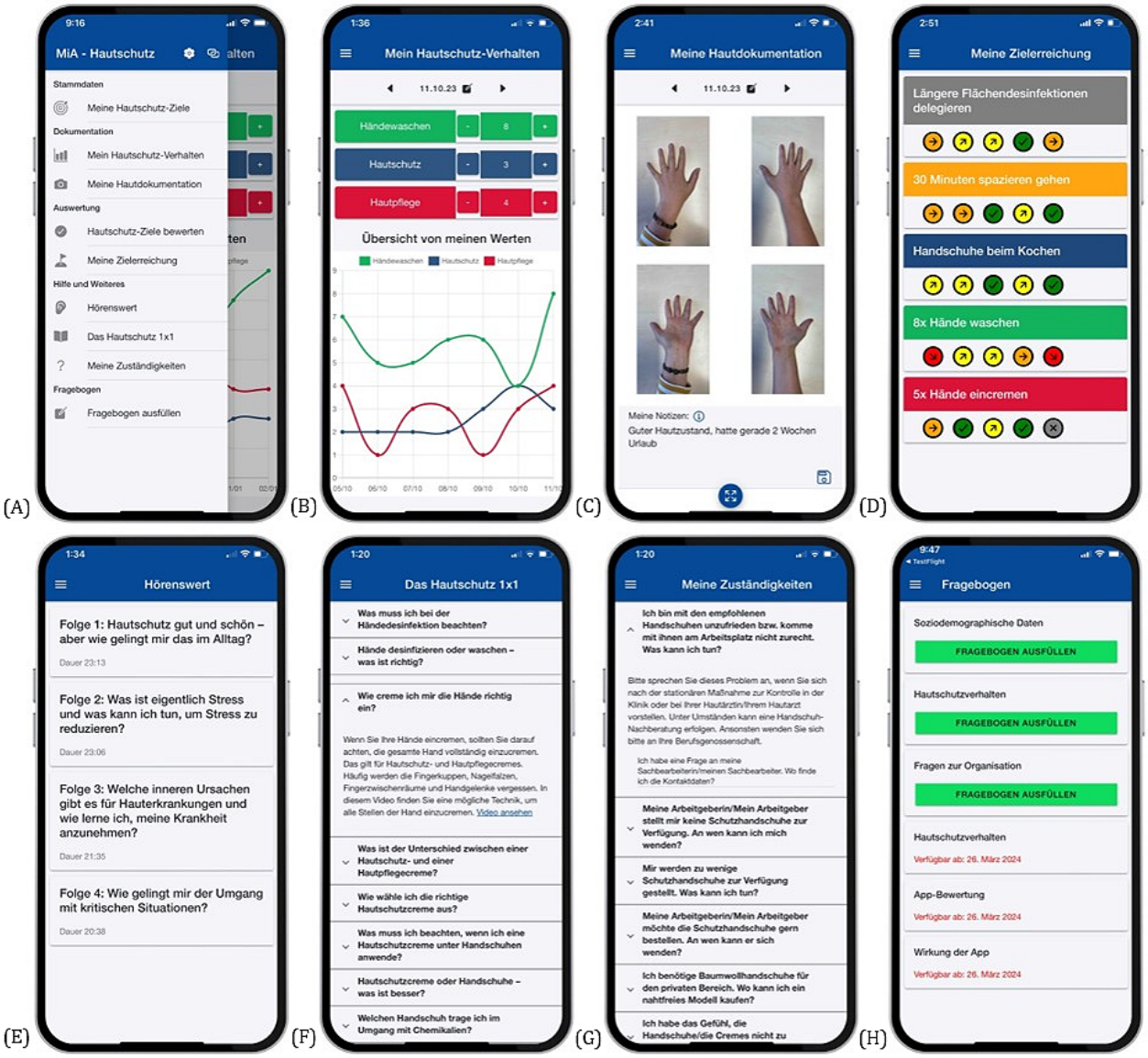

Multimedia Appendix 1 provides a more detailed description of each component according to the Template for Intervention Description and Replication checklist [39,40], including what, why, how often, and when components are unlocked during the maintenance program and where and which tailoring options are available. Table 1 describes the contents of the active components of the intervention in a standardized manner using the Behavior Change Technique Taxonomy version 1 [7,41].

XSL•FO
RenderX

Table 1. Behavior change techniques (BCTs) in the 12-week occupational dermatology maintenance program for improving health behavior change in patients with work-related skin diseases based on the BCT Taxonomy version 1 [7,41].

Component	Description	BCTs
Face-to-face intervention		
Individual goal-setting interview	Formulation of individual skin protection goals	• 1.1: goal setting (behavior)
MiA[a] app		
My Skin Protection Goals	Monitoring of individual skin protection goals and assessment of their achievement	• 1.5: review behavior goals
My Skin Protection Goals	Visualization of goal achievement over time using arrows	• 1.6: discrepancy between current behavior and goal
My Skin Protection Goals	Motivating feedback messages depending on the evaluation of the goal	• 3.1: social support (unspecified)
My Skin Protection Behavior	Tracking of skin protection behavior and monitoring the progress	• 2.3: self-monitoring of behavior
My Skin Documentation	Recording and observing the skin condition	• 2.5: monitoring of outcomes of behavior without feedback
To Listen	Information about the skin disease, associated difficulties, and strategies	• Episode 1: 4.1—instruction on how to perform the behavior and 15.1—verbal persuasion about capability • Episode 2: 5.1—information about health consequences, 5.2—salience of consequences, 8.2—behavior substitution, 12.4—distraction, and 15.1—verbal persuasion about capability • Episode 3: 8.2—behavior substitution, 5.1—information about health consequences, 5.2—salience of consequences, 4.2—information about antecedents, and 15.1—verbal persuasion about capability • Episode 4: 3.1—social support (unspecified), 3.3—social support (emotional), and 15.1—verbal persuasion about capability
Skin Protection 101	Information about seminar content on skin protection, itching, and stress	• 4.1: instruction on how to perform the behavior • 5.1: information about health consequences
My Accountabilities	Information about organizational issues related to care and responsibilities	• 5.3: inform about social and environmental consequences

[a]MiA: Mein Hautschutz im Alltag (My skin protection in everyday life).

Ethical Considerations

This study was approved by the Ethics Committee of Osnabrück University (Ethics-50/2022 and Ethics-15/2023). All patients assessed for eligibility were informed verbally and in writing about the study and its voluntary nature. All participants had to sign a declaration of consent to take part in the study. The participants were informed that data collection would be pseudonymized and that the evaluation and publication of the results would be anonymous. They received no compensation for their participation.

Inclusion Criteria and Study Design

Recruitment took place at iDerm in Osnabrück on the fourth day during the TIP. Study information was provided in patient groups as 7 to 9 new patients are usually admitted to the TIP each week. After all patients had been fully informed about the study, they were given sufficient time to decide whether to participate. Subsequently, interested patients received the written study information and declaration of consent. The inclusion criteria were (1) signed declaration of consent for study participation; (2) legal age (>18 years); (3) sufficient German language skills to understand the app content, participate in focus groups, and complete questionnaires; and (4) access to an internet-enabled smartphone (for step 2).

This study followed a user-centered and multistage iterative approach, which is a well-known and widespread procedure in the development of technologies in health care and UX research [42,43]. This approach allows for the consideration of the opinions and feedback of the target group as future users of the app in a participatory manner [23,44-46]. As illustrated in Figure 3, the entire process consisted of 2 steps with both qualitative and quantitative data collection.

XSL•FO
RenderX

Figure 3. Study process of multistage iterative testing of the app-based maintenance program for patients with work-related skin diseases consisting of (1) iterative development and testing and (2) 12-week testing and piloting. MiA: Mein Hautschutz im Alltag (My skin protection in everyday life); TIP: tertiary individual prevention; UX: user experience.

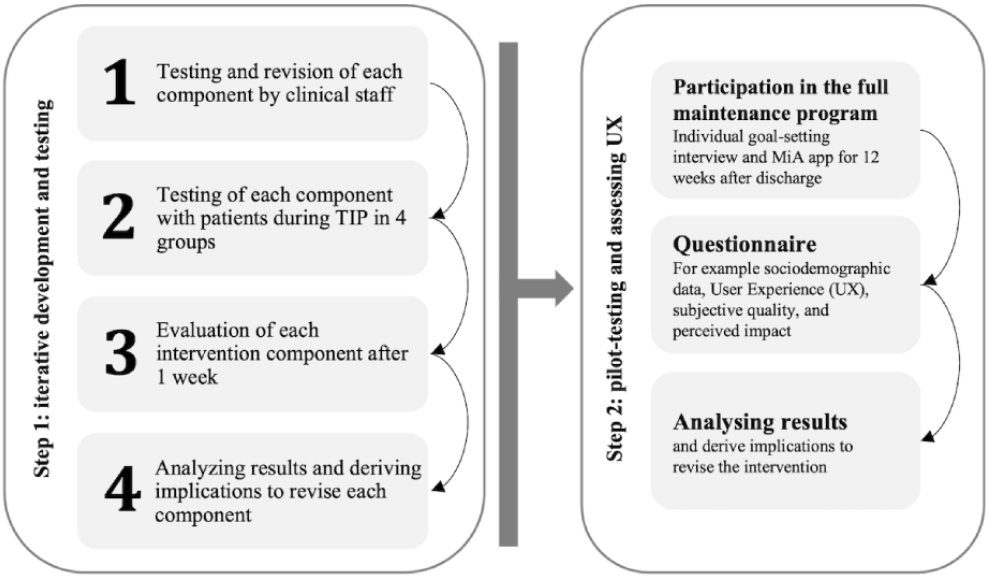

Instrument

Various instruments have been developed to assess the UX of mobile health apps. Among them, the user version of the Mobile Application Rating Scale (uMARS) is a well-known and reliable instrument that measures the UX of health apps from the user's perspective [47]. The uMARS is based on the Mobile Application Rating Scale, which is used to assess the quality of apps by clinical or technical experts [48,49]. The uMARS is made up of 16 items and comprises the UX dimensions of *engagement*, *functionality*, *aesthetics*, and *information quality*. The scales of subjective quality and perceived impact assess additional information about recommendation; willingness to use and pay for the app; and information regarding awareness and knowledge of and intention for behavior change. For this study, an instrument for step 1 and step 2 was developed based on the uMARS and the model shown in Figure 1.

Statistical analyses of the quantitative data were carried out using SPSS (version 28; IBM Corp). We calculated a score for all 5 UX dimensions as well as a score for all UX items. Adherence to the My Skin Protection Behavior, My Skin Documentation, and My Goal Achievement functions was determined based on the adherence values in Multimedia Appendix 1 and by dichotomizing the data into *adherent* and *nonadherent*. For continuous data (eg, age and UX), we calculated the mean, median, IQR, and SD. We present categorical data (eg, gender, subjective quality, and perceived impact) in frequencies and percentages.

Step 1: Iterative Development and Testing of Intervention Components

Overview

In the first step, the maintenance program was tested by 4 different patient groups during participation in the TIP at the iDerm in Osnabrück between November 2022 and November 2023. Patients were recruited on the fourth day of the TIP with a subsequent test phase of 2 weeks. During the test phase, patients were asked to document their overall impression as well as positive aspects and aspects requiring improvement of the tested component in a standardized documentation form. The documentation forms were not evaluated by the researchers as they merely served as preparation for the patients themselves for the focus group discussions. Each group tested and evaluated specific maintenance components as follows: group 1 evaluated *To listen*, group 2 evaluated the *face-to-face goal setting interview*, group 3 evaluated My Skin Protection Behavior and My Accountabilities on the app, and group 4 evaluated Skin Protection 101 and My Skin Documentation on the app.

Methods of Data Collection in Step 1

After the test phase, the patients completed a short quantitative questionnaire that assessed their overall impression of the maintenance components tested using selected items from the uMARS. The focus of step 1 was qualitative data collected via 4 subsequent focus group discussions. The aim of the focus group discussions was to record positive and negative aspects of the tested components, identify problems and challenges regarding their use, and jointly develop opportunities for improvements.

The focus group discussions were conducted by 2 moderators (NR and ML) based on guiding questions regarding content-related, methodological, and technical aspects. The guiding questions served as orientation. Additions to or deepening of the questions and topics that arose in the course of the discussions were permitted. The interviews were recorded using an audio recorder (Olympus LS-P4 linear pulse-code modulation recorder). Knowledge mapping was chosen for data analysis as it allows for the combination of data collection and data processing [50]. The results were written down in keywords on moderation cards and clustered on a metaplan board. The participants validated the results for completeness and correctness for each topic. In addition, experiences from the test phase and key results were documented by a research

XSL•FO
RenderX

assistant in a protocol. Aspects of the components that needed to be improved were then revised in the next iteration.

Step 2: Pilot-Testing and Assessing the UX

Overview

In the second step, the entire maintenance program, consisting of the face-to-face goal-setting interview and the MiA app with all its components, was tested by patients and formatively evaluated. Recruitment took place over a period of 10 weeks from January 2024 to March 2024. Patients who voluntarily agreed to participate were then given an appointment for the goal-setting interview as a one-to-one meeting and the subsequent onboarding to the app in the third week. The app was tested over a period of 12 weeks after discharge from the TIP, with 2 additional weeks to fill out the second questionnaire. Patients who did not complete the questionnaire within these 2 weeks received a reminder in the form of a paper questionnaire with a prepaid return envelope by post.

Methods of Data Collection in Step 2

Overview

The evaluation was carried out using a quantitative questionnaire consisting of 6 blocks. Blocks 1 to 3 were available immediately after registration, and blocks 4 to 6 were activated after the maintenance period of 12 weeks after discharge. In total, the instrument consisted of 85 questions about sociodemographic data; skin protection behavior before and after the rehabilitation program; and evaluation of the UX, subjective quality, and perceived effects, as well as questions about the structural, process, and outcome quality of the intervention.

In addition to closed-ended questions, the questionnaire also included open-ended questions to allow the participants to provide further details about the individual functions beyond the closed-ended questions [15]. The comprehensibility of the questions and the time required for completion were determined in a pretest with 16 patients before the study. In this publication, we report the results on sociodemographic data, UX evaluation, subjective quality, and perceived impact of the intervention.

UX Evaluation

To assess the UX, the relevant items of the uMARS were assigned to the UX dimensions shown in Figure 1 and supplemented with additional aspects as required. We measured feelings and emotions using 5 items (eg, “I find the app entertaining” or “The app offers me enough options to customise it to my personal requirements and needs”), content using 4 items (eg, “Is the content of the app relevant to you?” or “Are the texts understandable?”), usability using 6 items (eg, “How well and quickly do the app’s applications [buttons, menus] respond?” or “How easy was it for you to learn how to use the app?”), aesthetic using 2 items (“How high are the quality and resolution of the images and texts?” and “How would you rate the overall appearance of the app?”), and trust using 3 items (eg, “The information in the app appears to come from a credible source” and “Are you concerned that data you have entered into the app could be uploaded to third parties?”). All questions were answered on 5-point Likert scales that differed in their content depending on the dimension and item.

Subjective Quality and Perceived Impact

Both aspects were assessed using the uMARS [47]. Subjective quality was measured using the items related to recommendation, willingness to use in the following 12 months, and willingness to pay and overall rating by stars.

The perceived impact scale comprised 5 closed-ended questions on awareness and knowledge of and motivation for skin protection behavior; encouragement to seek for support; and enabling skin protection behavior on a 5-point Likert scale and a supplementary open-ended question on which functions were particularly helpful for implementing skin protection behavior.

Adherence

An important result for this study was the analysis of the frequency of use of individual components as this represented a central criterion for adherence to the intervention. We used a simple definition according to Donkin et al [51] and understood adherence as “the degree to which the user followed the programme as it was designed” [51]. This definition has already been applied in the context of other digital health interventions [45]. In this study, adherence was determined by the frequency with which the skin protection behavior was entered, how often photos were taken, and how often the goals were assessed. Adherence was high if the component My Skin Protection Behavior was used daily and the functions My Skin Documentation and My Skin Protection Goals were used weekly.

Results

Step 1

Study Participants

A total of 23 patients took part in the testing of the intervention components in step 1, of whom 10 (43%) were female and 13 (57%) were male. The average age of the participants was 51 (SD 11; range 27-62) years.

The patients worked in health care (8/23, 35%), the metalwork industry (4/23, 17%), hairdressing (3/23, 13%), construction (2/23, 9%), or other professions (6/23, 26%).

Main Results

The results of step 1 were used for iterative development to further optimize the intervention for phase 2. For example, in the My Skin Protection Behavior function, the scale of the days and frequencies in the diagram was too small. This was adjusted and enlarged in the revision. In the My Skin Documentation function, every entry was automatically saved in the comment function. This always hindered further input for a few seconds. A button was added so that users could save their comments at the end by themselves. Patients also highlighted that the app had a simple structure and reminded them of apps they already knew, had a good usability and visual presentation, and was factual without playful elements. A more detailed presentation of the results of step 1 and the associated modifications for step 2 can be found in Multimedia Appendix 2.

XSL•FO
RenderX

Step 2

Study Participants

Over 10 weeks (study period for step 2), a total of 79 patients participated in the TIP program, of whom 42 (53%) agreed to take part in the intervention and gave informed consent for study participation. The main reasons for nonparticipation were no interest or no subjectively felt need for the intervention (Figure 4).

All 42 patients took part in the individual goal-setting interview and received access to the MiA app for the 12-week testing phase. At T1, all participants completed the questionnaires (42/42, 100% response rate), and at T2, the response rate was 57% (24/42). Table 2 summarizes the sociodemographic data of the patients enrolled in this study.

Figure 4. Flowchart of the study population in step 2, including assessment for eligibility, participation, and follow-up at T1 and T2. MiA: Mein Hautschutz im Alltag (My skin protection in everyday life); TIP: tertiary individual prevention.

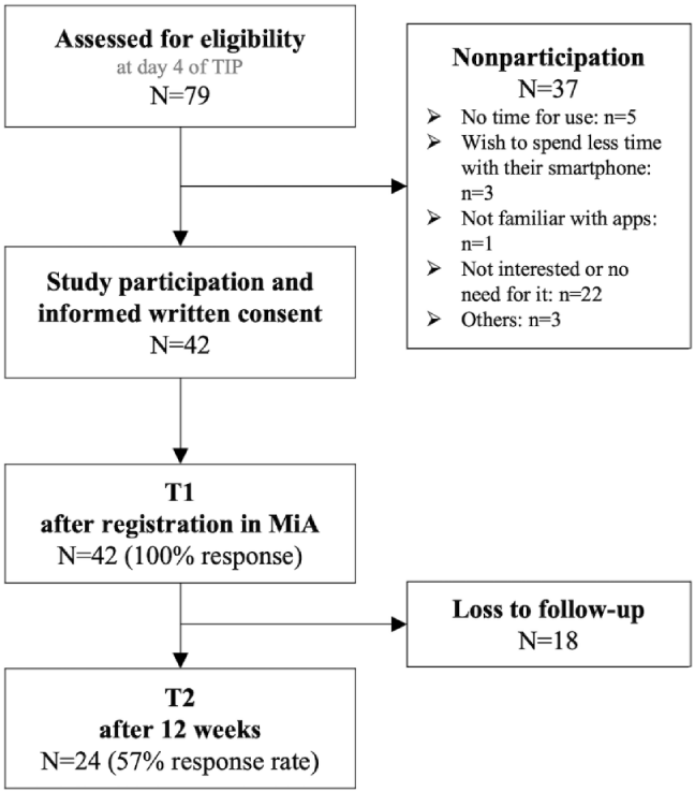

XSL•FO
RenderX

Table 2. Sociodemographic data of the study participants at T1 (N=42).

Sociodemographics	Values
Sex, n (%)	
Male	24 (57)
Female	18 (43)
Age (y)	
Values, mean (SD)	49.5 (13.1)
Values, median (IQR)	55.0 (18.6)
Occupational categories, n (%)	
Health care professions	17 (40)
Metalwork industry	9 (21)
Construction	5 (12)
Hairdressing	1 (2)
Food processing	2 (5)
Other	5 (12)
Pensioner	1 (2)
Currently unemployed	2 (5)
Highest educational level, n (%)	
Secondary school or elementary school–leaving certificate	9 (21)
Intermediate school–leaving certificate or secondary school–leaving certificate	15 (36)
"Meister"[a]	3 (7)
College certificate	5 (12)
General higher education entrance qualification or A-levels	7 (17)
Bachelor's degree	1 (2)
Master's degree or diploma	2 (5)
Employment status, n (%)	
Self-employed	4 (10)
Employed	35 (83)
Unemployed	3 (7)
Partnership, n (%)	
Yes	32 (76)
No	10 (24)

[a]Person with a higher vocational qualification in a craft, also known as "master craftsman" or "master craftswoman."

UX Results

The average UX score, with a theoretical minimum of 20 and a theoretical maximum of 100, was 80.18 (SD 8.94). Figure 5 shows the results for the assessment of the UX from the *trust*, *content*, *aesthetic*, *usability*, and *feelings and emotions* dimensions. The raw data for these items can be found in Multimedia Appendix 3. The results show overall positive ratings (scores of 4 and 5). The *feelings and emotions* dimension stood out, with comparatively frequent medium ratings and no ratings of 5 on 60% (3/5) of the items. Furthermore, the item "How easy was it for you to learn how to use the app?" in the *usability* dimension was the only item with a score of 1 by 2% (1/42) of the participants.

XSL•FO
RenderX

Figure 5. Results for the 5 user experience (UX) dimensions—content, usability, aesthetic, trust, and feelings and emotions—from the participants at T2 (n=24). The ratings range from 1 to 5. A rating of 1 means a low rating, and a rating of 5 is a high rating.

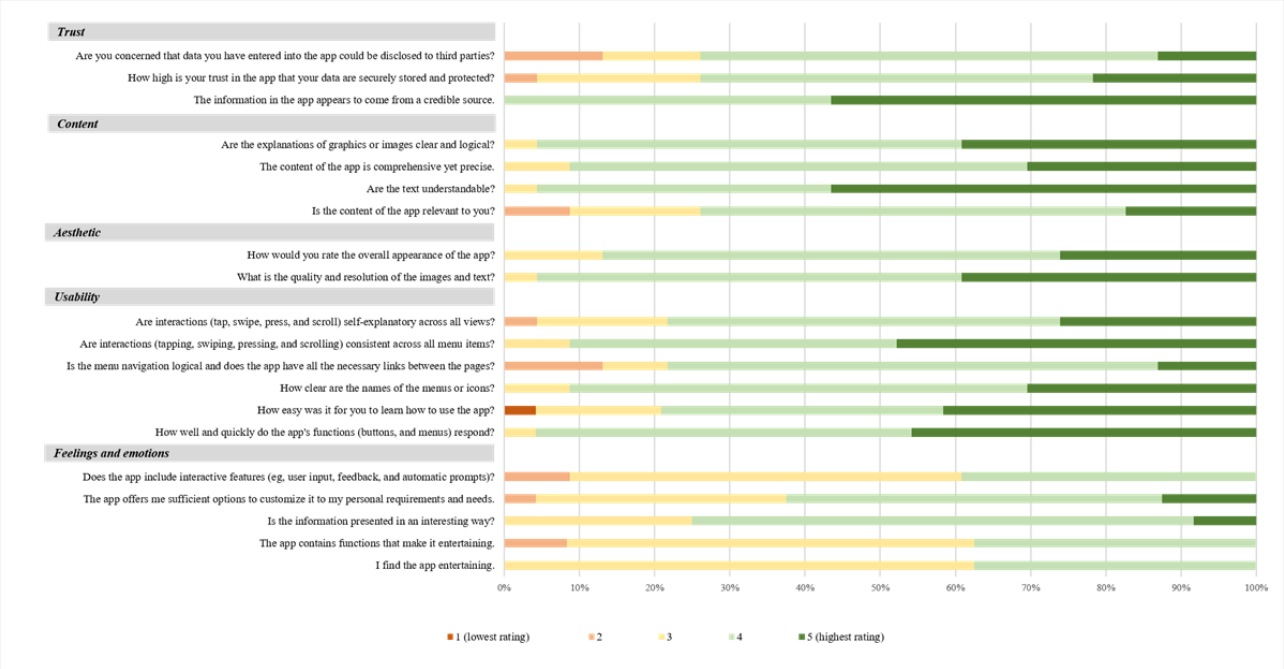

Subjective Quality and Perceived Impact

Overall, the results showed a high subjective quality (Multimedia Appendix 4). Approximately 75% (18/24) of the patients would recommend the app to others. The average willingness to use the app in the following 12 months was "once a week" (7/25, 30%). Most (14/24, 58%) were not willing to pay for the app. The app was most frequently rated with 4 out of 5 stars (15/24, 65%).

As shown in Figure 6, there was an average high level of agreement regarding the subjective impact of the MiA app in relation to skin protection behavior and important determinants (eg, awareness, motivation, or knowledge). All items were rated by more than half (Item 1: 17/24, 71%; Item 2: 13/24, 54%; Item 3: 16/24, 67%; Item 4: 19/24, 79%; Item 5: 21/24, 88%) of the participants as "somewhat agree" or "completely agree" (Multimedia Appendix 4).

Figure 6. Results on the perceived impact of the Mein Hautschutz im Alltag (My skin protection in everyday life) app regarding skin protection behavior by the participants at T2 (n=24).

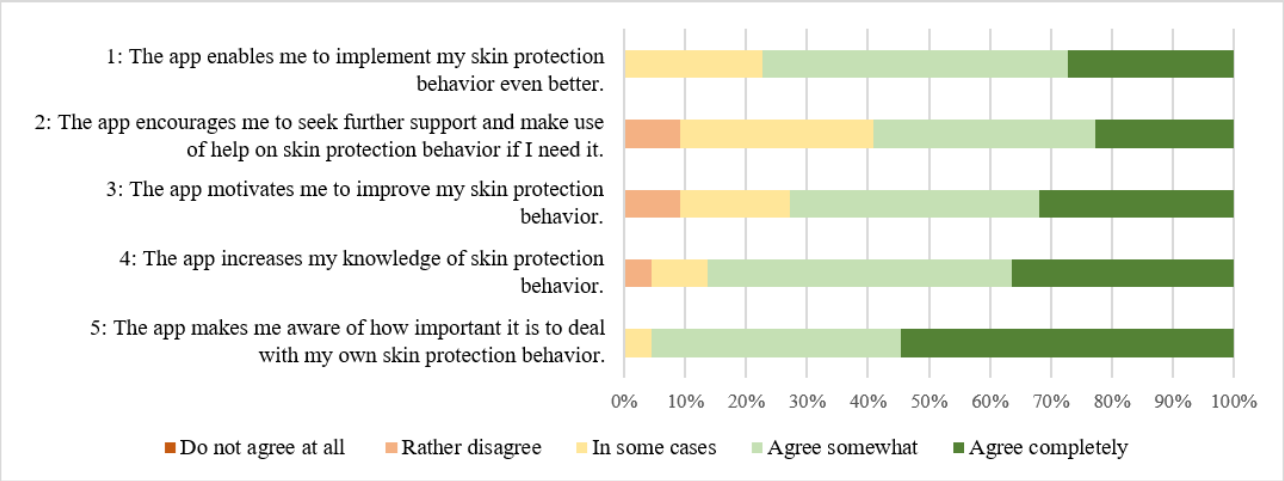

Adherence

A total of 135 goals were formulated, which corresponds to an average of 3.2 (SD 1.46) goals per person (Table 3). Textbox 1 contains exemplary goals from the goal-setting interviews. Of the 42 participants, 1 (2%) did not formulate any goals, and 10 (24%) formulated the maximum of 5 goals. Most of the goals related to the skin protection category.

XSL•FO
RenderX

Table 3. Number and categories of goals formulated in the individual goal-setting interview by the patients as part of the maintenance program during tertiary individual prevention (N=135).

	Skin cleansing, n (%)	Skin protection, n (%)	Skin care, n (%)	Social support, n (%)	Other, n (%)	Total, n (%)
Goal 1	8 (5.9)	12 (8.9)	11 (8.1)	5 (3.7)	5 (3.7)	41 (30.4)
Goal 2	5 (3.7)	13 (9.6)	5 (3.7)	2 (1.5)	10 (7.4)	35 (25.9)
Goal 3	3 (2.2)	9 (6.7)	7 (5.2)	4 (3)	6 (4.4)	29 (21.5)
Goal 4	4 (3)	8 (5.9)	4 (3)	3 (2.2)	1 (0.7)	20 (14.8)
Goal 5	3 (2.2)	4 (3)	0 (0)	2 (1.5)	1 (0.7)	10 (7.4)
Total	23 (17)	46 (34.1)	27 (20)	16 (11.9)	23 (17)	135 (100)

Textbox 1. Example goals set by the patients in the individual goal-setting interview.

Skin cleansing

- I intend to wash my hands less during the day.
- I intend to reduce the water temperature when cleansing my skin.
- I intend to wash my hands a maximum of 10 times a day.

Skin protection

- I will put on cotton gloves after applying cream.
- I will consistently wear padded gloves for activities involving mechanical pressure on the palms of my hands.
- I intend to apply a skin protection cream at least 3 times a day.

Skin care

- I will apply a skin care cream right at the end of the working day.
- I intend to apply a skin care product to my hands 5 times a day.
- I will apply cream to my whole body once a day (in the evening).

Social support

- I will use a cell phone alarm to remind me to apply skin care cream every day.
- I will ask for support at work (e.g. if the skin condition is bad).

Other

- I intend to pay more attention to myself and say “no” more often.
- I'm going to yoga class every Tuesday evening.
- I intend to go for a walk 2-3 times a week.

The results in terms of adherence to the app function My skin protection behavior are shown in Figure 7. The raw data can be found in Multimedia Appendix 5. The results show differences in adherence between the dimensions of skin protection behavior at the beginning of the maintenance program. On day 1, there were significantly more adherent than nonadherent participants to the dimensions of skin cleansing (adherent participants: 28/42, 67%; nonadherent participants: 14/42, 33%; P=.04) and skin care (adherent participants: 28/42, 67%; nonadherent participants: 14/42, 33%; P=.04) compared to the skin protection dimension, which had significantly more nonadherent than adherent participants (adherent participants: 13/42, 31%; nonadherent participants: 29/42, 69%; P=.02).

Furthermore, there was a difference in adherence over time during the 12-week maintenance program. As shown in Figures 7A and 7C, the number of nonadherent participants to the skin cleansing and skin care dimensions was higher than the number of adherent participants from day 26 onward. After 12 weeks, there were significantly more nonadherent than adherent participants who tracked their skin protection behavior daily in all 3 dimensions: skin cleansing (adherent participants: 12/42, 29%; nonadherent participants: 30/42, 71%; P=.008), skin care (adherent participants: 12/42, 29%; nonadherent participants: 30/42, 71%; P=.008), and skin protection (adherent participants: 9/42, 21%; nonadherent participants: 33/42, 79%; P<.001; Figures 7A, 7B, and 7C).

XSL•FO
RenderX

Figure 7. (A) Adherence to tracking of skin cleansing on the app function My Skin Protection Behavior, (B) adherence to tracking of skin protection on the app function My Skin Protection Behavior, and (C) adherence to tracking of skin care on the app function My Skin Protection Behavior by all participants during the 12-week pilot-testing (n=42).

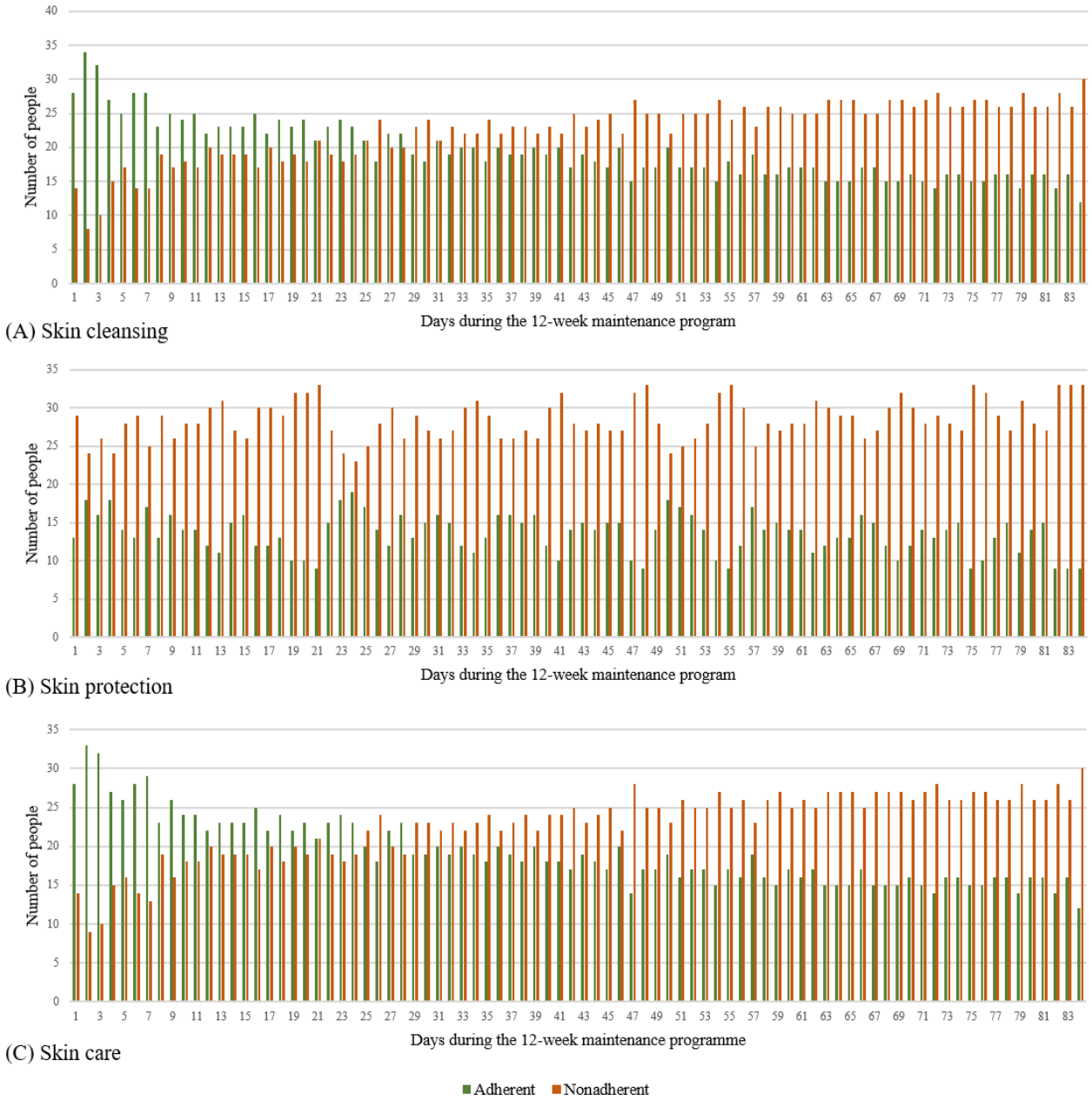

Adherence to rating the achievement of skin protection goals (Figure 8A) and photo documentation (Figure 8B) was lower than adherence to the tracking of skin protection behavior (Multimedia Appendix 5). The number of nonadherent participants was clearly lower than the number of adherent participants in each week of the 12-week maintenance program for both app functions. At the beginning, in week 1 (adherent participants: 4/42, 10%; nonadherent participants: 38/42, 90%; $P<.001$), and at the end, in week 12 (adherent participants: 4/42, 10%; nonadherent participants: 38/42, 90%; $P<.001$), there were significantly more nonadherent than adherent participants who assessed their skin protection goals. Similar results were also found for the app function My Skin Documentation in week 1 (adherent participants: 10/42, 24%; nonadherent participants: 32/42, 76%; $P<.001$) and week 12 (adherent participants: 4/42, 10%; nonadherent participants: 38/42, 90%; $P<.001$).

XSL•FO
RenderX

Figure 8. (A) Adherence to the achievement of skin protection goals on the app function My Skin Protection Goals and (B) adherence to skin documentation on the app function My Skin Documentation by all participants during the 12-week pilot-testing (n=42).

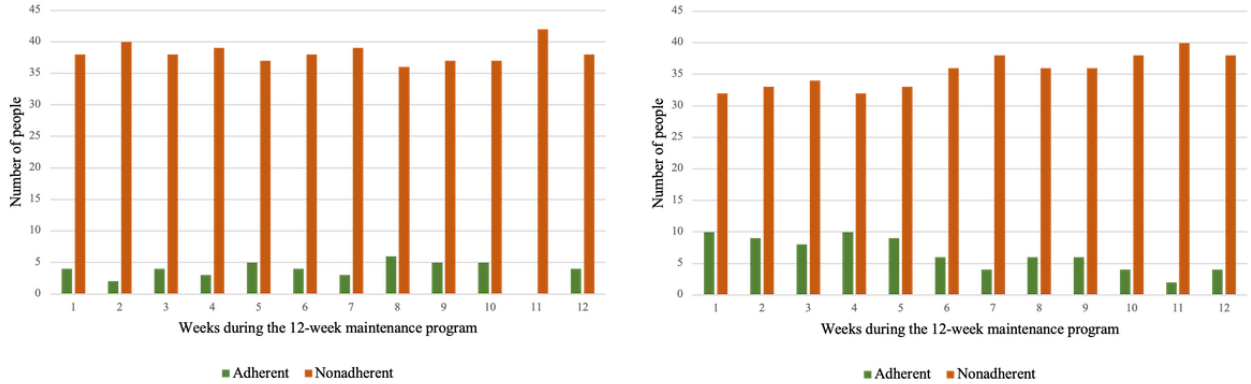

(A) Assessing skin protection goals

(B) Skin documentation

Discussion

Principal Findings and Comparison With Prior Work

The aim of this study was to evaluate the UX of and adherence to an app-based maintenance program. The MiA app is a self-management tool for patients with WRSDs after participating in an inpatient rehabilitation program (TIP) to support change in and maintenance of skin protection behavior. To the best of our knowledge, the MiA app is the first app for people with WRSDs. Our study was conducted by following a multistage and iterative development and pilot-testing process, with this paper focusing on the results of phase 2.

UX research is an important part in the development of apps to identify and address potential barriers to the use of apps as early as possible. It should precede the summative evaluation [38]. However, many studies investigating the influence of app-based interventions in the dermatological setting only report on clinical outcomes such as quality of life, eczema severity, or itching [24,27] or app use behavior [25,26]. Zvulunov et al [24] point out that there are missing data on the usability of and adherence to apps in the field of chronic skin diseases.

We evaluated the UX using the dimensions of *feelings and emotions*, *usability*, *aesthetics*, *content*, and *trust*. Our study generally showed high to very high ratings for all dimensions with the exception of *feelings and emotions*, which was rated comparatively lower. Studies show that gamified apps can improve the interaction, motivation, and adherence [23,52]. However, the benefit of gamification differs between genders and decreases with age, making it difficult to successfully implement those functions in our heterogeneous study population [53]. According to Tolks et al [53], gamified apps are particularly useful for people with low motivation to implement a behavior, and Matterne et al [54] were able to show that patients with WRSDs have a high motivation to implement skin protection, which increased significantly during the TIP. Furthermore, in a focus group discussion in our study, patients expressed that the app should not contain any playful elements and should remain factual as it dealt with a disease. The extent to which the app is used also depends on the app's BCTs. Milne-Ives et al [55] were able to demonstrate that some BCTs are associated with higher use of and engagement with an app. These included goal setting and feedback, self-monitoring, and social support, which are also part of the MiA app. However, in the future, the app could be expanded to include further interactive functions that allow users, for example, to assess their skin condition more accurately. One possibility might be to analyze the severity of hand eczema in photo documentation using artificial intelligence [24]. Weigandt et al [27] were able to show that the eczema severity determined via photos significantly correlates with live assessment by dermatologists. Therefore, photographs appear to allow for reliable results regarding the actual severity of hand eczema.

The presentation of information on the app was rated as particularly interesting by patients. We attribute this to the different ways in which the information was presented (podcasts, videos, and short questions and answers).

The descriptive results of the UX also show single exceptions regarding usability. Due to the small study cohort at T2, we were unable to test any associations between UX and the sociodemographic characteristics of the participants. However, digital health literacy plays an important role in usability. It is known that, in addition to app-specific factors, the use of an app may also depend on personal factors such as age, sex, educational level, and socioeconomic status [15,27,56]. Nationally and internationally, it has been shown that older people with a lower educational level and of a lower socioeconomic status tend to have a lower digital health literacy [57,58]. With an average age of 49.5 (SD 13.1) years and a high proportion of people with elementary and secondary school–leaving certification (24/42, 57%), our study cohort had characteristics that could indicate a lower level of digital health literacy. Therefore, a potential influence of digital health literacy on the use and handling of the MiA app is conceivable. Further research should be carried out to determine which socioeconomic characteristics in our field of application are particularly affected by low digital health literacy. This is important to provide patients with individual and tailored support from health care staff during the onboarding process [59]. Nevertheless, onboarding via personal contact in face-to-face meetings is a particular strength of our study because we were able to respond to the individual digital competence of the patients.

XSL•FO
RenderX

We found differences in adherence between the interactive functions of the app. The daily tracking of skin protection behavior was used more often than the weekly skin documentation and goal assessment. Furthermore, skin care and skin cleansing were tracked more frequently than skin protection. In addition, adherence to skin cleansing and skin care tracking decreased over the 12 weeks. The lower tracking of skin protection at the beginning might be explained by the fact that patients are at home for a period of 3 weeks after discharge and do not start working in their profession straight away. That means that the use of skin protection creams at work only becomes relevant from day 21 onward. The digital documentation of skin protection creams may not have increased from day 21 onward as the practical implementation of digital tracking during work is not feasible. However, in contrast to this, we observed a clearly higher proportion of formulated individual goals in the skin protection category than in the skin cleansing, skin care, support, and other categories. This emphasizes that the topic of skin protection is nevertheless an important aspect for patients after the TIP even if this cannot be reflected in skin protection tracking for the aforementioned reasons. Against this background, linking both functions appears to be useful to adequately cover all dimensions of skin protection behavior. There are various possible reasons for the less frequent use of weekly photo documentation and target assessment. First, both components are technically more complex and more challenging functions and require more time. One possible explanation for the low level of photo documentation could be that photos are not taken routinely by the users but, rather, when acute skin changes occur. The weekly rhythm we set may not correlate with the progression of the individual skin diseases. WRSDs occur in over half of cases with a comorbidity of AD and progress in intervals of varying length [33]. The goal assessment is also cognitively more complex because it requires reflection on the last behavior performed by the users and the decision of the appropriate response option. The reasons for the less frequent use of both functions should be further investigated, for example, by users and nonusers reporting on their experiences, motivation, and the usefulness of these functions (eg, in qualitative interviews [56]). Nevertheless, it should be kept in mind that high adherence to the app functions (eg, skin protection tracking) is not necessarily associated with behavior changes and better clinical outcomes. In everyday life, skin diseases are associated with a high level of adherence to therapy. Daily documentation may be too time-consuming to implement in practice, meaning that regular app use does not necessarily lead to better outcomes in terms of eczema severity or quality of life [27]. The average willingness to use the app approximately once a week supports this assumption. The use of the MiA app after the TIP can also be individual depending on the extent to which patients benefited from patient education during the TIP. Another influencing factor on app use could be the duration of the disease. TIP patients have usually had the skin disease for several years and have developed their own self-management strategies in the past. In the future, it could be investigated what are the use and benefits for patients with milder forms of WRSDs (eg, as part of outpatient care). The results on the perceived impact show that the MiA app can also be used as a tool for disseminating knowledge about skin protection–related content.

Overall, our results are consistent with those of other studies on dermatological apps and, thus, show expected results on adherence, which decreases over time [25]. Shah et al [26] showed that the average duration of use of an app for people with AD was approximately 6 weeks. The frequent short duration of use of app-based interventions and the associated lack of evidence of long-term effects is a general and well-known problem in this research context [2,26].

We were able to demonstrate a high quality of the app. Despite the low adherence, the results on the perceived effect also show promising indications that the app might influence relevant determinants such as motivation, knowledge, or awareness regarding the change in health behavior according to the subjective statements of the participants. However, the effect of the MiA app on skin protection behavior, quality of life, and eczema severity remains unclear thus far and should be further investigated in the future in comparison with a control group.

Limitations

In addition to the limitations already mentioned, this study has further limitations that should be considered when interpreting the results. First, UX is a multifactorial and complex construct. There are different methods for assessing UX that have their own strengths and limitations [13]. While questionnaires are cost-efficient and provide quick results, dishonest answers and individual interpretations of the results can be disadvantageous. Focus group discussions are easy to conduct but dependent on the interviewing skills of those conducting them [14]. To measure UX, we used a validated instrument as a basis and adapted it to the context and target group of the app. We concede that the validity of the instrument and a direct comparison of the results with those of other studies is possible to a limited extent. The standardization of the definitions of the UX dimensions and the development of validated instruments for UX research is an important future research focus [19].

The participants in our study matched the sociodemographic characteristics of those of other studies on TIP [60]. Nevertheless, due to the small number of participants, the low use behavior, and a remaining cohort of 57% (24/42) at T2, it was not possible to calculate differences in UX and adherence among age, sex, and educational level.

The target group participated during the development and testing of the MiA app. Nevertheless, their involvement could have been even stronger in the form of agile co-design (eg, for prototyping the app before programming and technical implementation and for identifying realistic adherence [23,61-63]). However, such an approach was not possible due to the organizational processes and duration of this study.

We were also unable to determine the extent to which the informative functions (podcasts, questions and answers, and videos) were used. On the basis of other studies, it can be assumed that the use of educational elements also decreases after a few weeks [25]. However, podcasts activated over the course of the maintenance program could be a way to maintain use over time.

XSL•FO
RenderX

Conclusions

Our study shows a high UX of the MiA app in almost all dimensions. To increase interaction, functions can be added to the app (eg, use of artificial intelligence with photo documentation to allow for a clearer assessment of the skin condition). We refrained from using gamification due to the heterogeneous study population of patients with WRSDs. We also observed single exceptions in usability. Usability depends, among other things, on the digital health literacy of the user. Future research should investigate in this context which sociodemographic characteristics show lower usability to provide these groups with further support when using the app.

We obtained expected results in terms of adherence, with decreasing use over time, which corresponds to the current state of research in this field. Particularly after resuming work after 3 weeks, daily documentation of skin protection behavior appears to be too time-consuming. The effectiveness of the MiA app in terms of skin protection behavior, quality of life, and eczema severity should be further investigated with a control group.

Acknowledgments

The authors would like to thank the Institution for Statutory Accident Insurance and Prevention of the Health and Welfare Services (Berufsgenossenschaft für Gesundheitsdienst und Wohlfahrtspflege) for funding the project (reference: Ext FF_1569; Technologiebasierte Nachbetreuung in der berufsdermatologischen prevention der BK Nr. 5101 [TecNaP]). The authors thank the members of an internal Institute for Interdisciplinary Dermatological Prevention and Rehabilitation working group and Dr Andreas Reusch of the Zentrum Patientenschulung und Gesundheitsförderung (Centre for Patient Education and Health Promotion) for developing the concept for the individual goal-setting interview. They would also like to thank student assistants Gesa Lühring and Katharina Brieskorn for their support during the focus group discussions, study administration, and data entry. For programming the app and providing technical support, the authors would like to thank the companies BINACON GmbH and Solutec GmbH, with Klaus Welling, Thomas Waldmann, Michael Mühlhäuser, Stanislav Yordanov, Stefan Schmädeke, and Patrick Renter.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Intervention description.
[DOCX File , 19 KB - formative_v9i1e66791_app1.docx]

Multimedia Appendix 2
Results of step 1.
[DOCX File , 22 KB - formative_v9i1e66791_app2.docx]

Multimedia Appendix 3
User experience.
[DOCX File , 21 KB - formative_v9i1e66791_app3.docx]

Multimedia Appendix 4
Subjective quality and perceived impact.
[DOCX File , 29 KB - formative_v9i1e66791_app4.docx]

Multimedia Appendix 5
Adherence to the app's My skin protection behavior, My skin documentation, and My skin protection goals functions.
[DOCX File , 34 KB - formative_v9i1e66791_app5.docx]

References

1. Milne-Ives M, Lam C, De Cock C, Van Velthoven MH, Meinert E. Mobile apps for health behavior change in physical activity, diet, drug and alcohol use, and mental health: systematic review. JMIR Mhealth Uhealth 2020 Mar 18;8(3):e17046 [FREE Full text] [doi: 10.2196/17046] [Medline: 32186518]
2. Meyerowitz-Katz G, Ravi S, Arnolda L, Feng X, Maberly G, Astell-Burt T. Rates of attrition and dropout in app-based interventions for chronic disease: systematic review and meta-analysis. J Med Internet Res 2020 Sep 29;22(9):e20283 [FREE Full text] [doi: 10.2196/20283] [Medline: 32990635]

XSL•FO
RenderX

3. Kitazawa M, Takeda Y, Hatta M, Horikawa C, Sato T, Osawa T, et al. Lifestyle intervention with smartphone app and isCGM for people at high risk of type 2 diabetes: randomized trial. J Clin Endocrinol Metab 2024 Mar 15;109(4):1060-1070 [FREE Full text] [doi: 10.1210/clinem/dgad639] [Medline: 37931069]
4. Guo M, Meng F, Guo Q, Bai T, Hong Y, Song F, et al. Effectiveness of mHealth management with an implantable glucose sensor and a mobile application among Chinese adults with type 2 diabetes. J Telemed Telecare 2023 Sep;29(8):632-640. [doi: 10.1177/1357633X211020261] [Medline: 34152238]
5. Stawarz K, Cox AL, Blandford A. Beyond self-tracking and reminders: designing smartphone apps that support habit formation. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. 2015 Presented at: CHI '15; April 18-23, 2015; Seoul, Republic of Korea. [doi: 10.1145/2702123.2702230]
6. McKay FH, Wright A, Shill J, Stephens H, Uccellini M. Using health and well-being apps for behavior change: a systematic search and rating of apps. JMIR Mhealth Uhealth 2019 Jul 04;7(7):e11926 [FREE Full text] [doi: 10.2196/11926] [Medline: 31274112]
7. Michie S, Richardson M, Johnston M, Abraham C, Francis J, Hardeman W, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013 Aug;46(1):81-95 [FREE Full text] [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
8. Payne HE, Lister C, West JH, Bernhardt JM. Behavioral functionality of mobile apps in health interventions: a systematic review of the literature. JMIR Mhealth Uhealth 2015 Feb 26;3(1):e20 [FREE Full text] [doi: 10.2196/mhealth.3335] [Medline: 25803705]
9. Birrell L, Furneaux-Bate A, Chapman C, Newton NC. A mobile peer intervention for preventing mental health and substance use problems in adolescents: protocol for a randomized controlled trial (The Mind Your Mate Study). JMIR Res Protoc 2021 Jul 30;10(7):e26796 [FREE Full text] [doi: 10.2196/26796] [Medline: 34328426]
10. Rismawan W, Marchira CR, Rahmat I. Usability, acceptability, and adherence rates of mobile application interventions for prevention or treatment of depression: a systematic review. J Psychosoc Nurs Ment Health Serv 2021 Feb 01;59(2):41-47. [doi: 10.3928/02793695-20201104-02] [Medline: 33180949]
11. Jakob R, Harperink S, Rudolf AM, Fleisch E, Haug S, Mair JL, et al. Factors influencing adherence to mHealth apps for prevention or management of noncommunicable diseases: systematic review. J Med Internet Res 2022 May 25;24(5):e35371 [FREE Full text] [doi: 10.2196/35371] [Medline: 35612886]
12. Lubis M, Sutoyo E, Azuddin M, Handayani D. User experience in mobile application design: utility defined context of use. J Phys Conf Ser 2019;1361:12043. [doi: 10.1088/1742-6596/1361/1/012043]
13. Kaveladze BT, Wasil AR, Bunyi JB, Ramirez V, Schueller SM. User Experience, Engagement, and Popularity in Mental Health Apps: Secondary Analysis of App Analytics and Expert App Reviews. JMIR Hum Factors 2022 Jan 31;9(1):e30766 [FREE Full text] [doi: 10.2196/30766] [Medline: 35099398]
14. Bitkina OV, Kim HK, Park J. Usability and user experience of medical devices: an overview of the current state, analysis methodologies, and future challenges. Int J Ind Ergon 2020 Mar;76:102932. [doi: 10.1016/j.ergon.2020.102932]
15. Biduski D, Bellei EA, Rodriguez JP, Zaina LA, De Marchi AC. Assessing long-term user experience on a mobile health application through an in-app embedded conversation-based questionnaire. Comput Hum Behav 2020 Mar;104:106169. [doi: 10.1016/j.chb.2019.106169]
16. Kim A, Kim KJ. User experience and the multi-stage adoption of mobile apps. J Inf Technol Appl Manag 2014;21(2):49-79. [doi: 10.21219/jitam.2014.21.2.049]
17. Schrepp M, Hinderks A, Thomaschewski J. Construction of a benchmark for the User Experience Questionnaire (UEQ). Int J Interact Multim Artif Intell 2017;4(4):40-44. [doi: 10.9781/ijimai.2017.445]
18. Schrepp M, Hinderks A, Thomaschewski J. Applying the User Experience Questionnaire (UEQ) in different evaluation scenarios. In: Marcus A, editor. Design, User Experience, and Usability. Cham, Switzerland: Springer; 2014.
19. Lemon C, Huckvale K, Carswell K, Torous J. A narrative review of methods for applying user experience in the design and assessment of mental health smartphone interventions. Int J Technol Assess Health Care 2020;36(1):64-70. [doi: 10.1017/S0266462319003507] [Medline: 31973787]
20. Blanchard M, Backhaus L, Ming Azevedo P, Hügle T. An mHealth app for fibromyalgia-like post-COVID-19 syndrome: protocol for the analysis of user experience and clinical data. JMIR Res Protoc 2022 Feb 04;11(2):e32193 [FREE Full text] [doi: 10.2196/32193] [Medline: 34982039]
21. Thielsch MT, Salaschek M. [Toolbox for user-experience evaluation of online health information]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2020 Jun;63(6):721-728. [doi: 10.1007/s00103-020-03142-7] [Medline: 32367206]
22. Truong MT, Nwosu OB, Gaytan Torres ME, Segura Vargas MP, Seifer AK, Nitschke M, et al. A yoga exercise app designed for patients with axial spondylarthritis: development and user experience study. JMIR Form Res 2022 Jun 03;6(6):e34566 [FREE Full text] [doi: 10.2196/34566] [Medline: 35657655]
23. Xu X, Griva K, Koh M, Lum E, Tan WS, Thng S, et al. Creating a smartphone app for caregivers of children with atopic dermatitis with caregivers, health care professionals, and digital health experts: participatory co-design. JMIR Mhealth Uhealth 2020 Oct 29;8(10):e16898 [FREE Full text] [doi: 10.2196/16898] [Medline: 33118949]

XSL•FO
RenderX

24. Zvulunov A, Lenevich S, Migacheva N. A mobile health app for facilitating disease management in children with atopic dermatitis: feasibility and impact study. JMIR Dermatol 2023 Dec 13;6:e49278 [FREE Full text] [doi: 10.2196/49278] [Medline: 38090787]
25. Gudmundsdóttir SL, Ballarini T, Ámundadóttir ML, Mészáros J, Eysteinsdottir JH, Thorleifsdottir RH, et al. Engagement, retention, and acceptability in a digital health program for atopic dermatitis: prospective interventional study. JMIR Form Res 2023 Jun 14;7:e41227 [FREE Full text] [doi: 10.2196/41227] [Medline: 36975050]
26. Shah S, Kemp JM, Kvedar JC, Gracey LE. A feasibility study of the burden of disease of atopic dermatitis using a smartphone research application, myEczema. Int J Womens Dermatol 2020 Dec;6(5):424-428 [FREE Full text] [doi: 10.1016/j.ijwd.2020.08.001] [Medline: 33898712]
27. Weigandt WA, Schardt Y, Bruch A, Herr R, Goebeler M, Benecke J, et al. Impact of an eHealth smartphone app on quality of life and clinical outcome of patients with hand and foot eczema: prospective randomized controlled intervention study. JMIR Mhealth Uhealth 2023 Mar 07;11:e38506 [FREE Full text] [doi: 10.2196/38506] [Medline: 36881465]
28. DGUV-Statistiken für die Praxis 2023. DGUV. 2024 Oct. URL: https://publikationen.dguv.de/statistiken/zahlen-und-fakten/4994/dguv-statistiken-fuer-die-praxis-2023 [accessed 2025-04-10]
29. Brans R, Schröder-Kraft C, Skudlik C, John SM, Geier J. Tertiary prevention of occupational skin diseases: prevalence of allergic contact dermatitis and pattern of patch test results. Contact Dermatitis 2019 Jan;80(1):35-44. [doi: 10.1111/cod.13098] [Medline: 30506828]
30. Skudlik C, Weisshaar E, Scheidt R, Elsner P, Wulfhorst B, Schönfeld M, et al. First results from the multicentre study rehabilitation of occupational skin diseases--optimization and quality assurance of inpatient management (ROQ). Contact Dermatitis 2012 Mar;66(3):140-147. [doi: 10.1111/j.1600-0536.2011.01991.x] [Medline: 22070197]
31. Brans R, John SM. Clinical patterns and associated factors in patients with hand eczema of primarily occupational origin. J Eur Acad Dermatol Venereol 2016 May;30(5):798-805. [doi: 10.1111/jdv.13515] [Medline: 26660508]
32. Elsner P, Schliemann S. Prävention der BK 5101. Trauma Berufskrankh 2015 Dec 1;17(4):227-232. [doi: 10.1007/s10039-015-0103-8]
33. Brans R, Skudlik C. [Prevention of hand eczema]. Hautarzt 2019 Oct;70(10):797-803. [doi: 10.1007/s00105-019-4463-z] [Medline: 31410496]
34. Ristow N, Rocholl M, Wilke A, John SM, Ludewig M. Using behaviour change techniques to characterize patient educational interventions in tertiary individual prevention of work-related skin diseases. Transl Behav Med 2024 May 24;14(6):341-352. [doi: 10.1093/tbm/ibad081] [Medline: 38159249]
35. Wilke A, Skudlik C, Sonsmann FK. [Individual prevention of occupational contact dermatitis: protective gloves and skin protection recommendations as part of the patient management scheme by the public statutory employers' liability insurance]. Hautarzt 2018 Jun;69(6):449-461. [doi: 10.1007/s00105-018-4170-1] [Medline: 29721626]
36. Wilke A, Gediga K, John SM, Wulfhorst B. Evaluation of structured patient education in occupational skin diseases: a systematic assessment of the disease-specific knowledge. Int Arch Occup Environ Health 2014 Nov;87(8):861-869. [doi: 10.1007/s00420-014-0926-9] [Medline: 24482039]
37. Ristow N, Wilke A, Malte John S, Ludewig M. Development of an app-based maintenance programme to promote skin protection behaviour for patients with work-related skin diseases. Health Educ J 2022 Aug 02;81(6):731-744. [doi: 10.1177/00178969221115990]
38. Skivington K, Matthews L, Simpson SA, Craig P, Baird J, Blazeby JM, et al. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. BMJ 2021 Sep 30;374:n2061 [FREE Full text] [doi: 10.1136/bmj.n2061] [Medline: 34593508]
39. Hoffmann TC, Glasziou PP, Boutron I, Milne R, Perera R, Moher D, et al. [better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide]. Gesundheitswesen 2016 Mar;78(3):175-188. [doi: 10.1055/s-0041-111066] [Medline: 26824401]
40. Hoffmann TC, Glasziou PP, Boutron I, Milne R, Perera R, Moher D, et al. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. BMJ 2014 Mar 07;348:g1687 [FREE Full text] [doi: 10.1136/bmj.g1687] [Medline: 24609605]
41. Michie S, Atkins L, West R. The Behaviour Change Wheel: A Guide to Designing Interventions. Sutton, UK: Silverback Publishing; 2014.
42. Matthew-Maich N, Harris L, Ploeg J, Markle-Reid M, Valaitis R, Ibrahim S, et al. Designing, implementing, and evaluating mobile health technologies for managing chronic conditions in older adults: a scoping review. JMIR Mhealth Uhealth 2016 Jun 09;4(2):e29 [FREE Full text] [doi: 10.2196/mhealth.5127] [Medline: 27282195]
43. Wolcott MD, McLaughlin JE. Exploring user experience (UX) research methods in health professions education. Curr Pharm Teach Learn 2024 Feb;16(2):144-149. [doi: 10.1016/j.cptl.2023.12.010] [Medline: 38158330]
44. Gracey LE, Zan S, Gracz J, Miner JJ, Moreau JF, Sperber J, et al. Use of user-centered design to create a smartphone application for patient-reported outcomes in atopic dermatitis. NPJ Digit Med 2018;1:33 [FREE Full text] [doi: 10.1038/s41746-018-0042-4] [Medline: 31304315]

XSL•FO
RenderX

45. Kernebeck S, Busse TS, Ehlers JP, Vollmar HC. [Adherence to digital health interventions: definitions, methods, and open questions]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2021 Oct;64(10):1278-1284 [FREE Full text] [doi: 10.1007/s00103-021-03415-9] [Medline: 34559252]
46. Alberts NM, Badawy SM, Hodges J, Estepp JH, Nwosu C, Khan H, et al. Development of the InCharge health mobile app to improve adherence to hydroxyurea in patients with sickle cell disease: user-centered design approach. JMIR Mhealth Uhealth 2020 May 08;8(5):e14884 [FREE Full text] [doi: 10.2196/14884] [Medline: 32383683]
47. Stoyanov SR, Hides L, Kavanagh DJ, Wilson H. Development and validation of the user version of the Mobile Application Rating Scale (uMARS). JMIR Mhealth Uhealth 2016 Jun 10;4(2):e72 [FREE Full text] [doi: 10.2196/mhealth.5849] [Medline: 27287964]
48. Messner EM, Terhorst Y, Barke A, Baumeister H, Stoyanov S, Hides L, et al. The German version of the mobile app rating scale (MARS-G): development and validation study. JMIR Mhealth Uhealth 2020 Mar 27;8(3):e14479 [FREE Full text] [doi: 10.2196/14479] [Medline: 32217504]
49. Stoyanov SR, Hides L, Kavanagh DJ, Zelenko O, Tjondronegoro D, Mani M. Mobile app rating scale: a new tool for assessing the quality of health mobile apps. JMIR Mhealth Uhealth 2015 Mar 11;3(1):e27 [FREE Full text] [doi: 10.2196/mhealth.3422] [Medline: 25760773]
50. Pelz C, Schmitt A, Meis M. Knowledge mapping as a tool for analyzing focus groups and presenting their results in market and evaluation research. Forum Qual Soc Res 2004;5(2) [FREE Full text] [doi: 10.17169/fqs-5.2.601]
51. Donkin L, Christensen H, Naismith SL, Neal B, Hickie IB, Glozier N. A systematic review of the impact of adherence on the effectiveness of e-therapies. J Med Internet Res 2011 Aug 05;13(3):e52 [FREE Full text] [doi: 10.2196/jmir.1772] [Medline: 21821503]
52. Maher CA, Olds T, Vandelanotte C, Plotnikoff R, Edney SM, Ryan JC, et al. Gamification in a physical activity app: what gamification features are being used, by whom, and does it make a difference? Games Health J 2022 Jun;11(3):193-199. [doi: 10.1089/g4h.2021.0207] [Medline: 35501981]
53. Tolks D, Lampert C, Dadaczynski K, Maslon E, Paulus P, Sailer M. [Game-based approaches to prevention and health promotion: serious games and gamification]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2020 Jun;63(6):698-707. [doi: 10.1007/s00103-020-03156-1] [Medline: 32451596]
54. Matterne U, Diepgen TL, Weisshaar E. Effects of a health-educational and psychological intervention on socio-cognitive determinants of skin protection behaviour in individuals with occupational dermatoses. Int Arch Occup Environ Health 2010 Feb;83(2):183-189. [doi: 10.1007/s00420-009-0448-z] [Medline: 19641933]
55. Milne-Ives M, Homer SR, Andrade J, Meinert E. Potential associations between behavior change techniques and engagement with mobile health apps: a systematic review. Front Psychol 2023;14:1227443 [FREE Full text] [doi: 10.3389/fpsyg.2023.1227443] [Medline: 37794916]
56. Jeffrey B, Bagala M, Creighton A, Leavey T, Nicholls S, Wood C, et al. Mobile phone applications and their use in the self-management of type 2 diabetes mellitus: a qualitative study among app users and non-app users. Diabetol Metab Syndr 2019;11:84 [FREE Full text] [doi: 10.1186/s13098-019-0480-4] [Medline: 31636719]
57. Choukou MA, Sanchez-Ramirez DC, Pol M, Uddin M, Monnin C, Syed-Abdul S. COVID-19 infodemic and digital health literacy in vulnerable populations: a scoping review. Digit Health 2022;8:20552076221076927 [FREE Full text] [doi: 10.1177/20552076221076927] [Medline: 35223076]
58. Dratva J, Schaeffer D, Zeeb H. [Digital health literacy in Germany: current status, concepts, and challenges]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2024 Mar;67(3):277-284 [FREE Full text] [doi: 10.1007/s00103-024-03841-5] [Medline: 38315221]
59. Busse TS, Nitsche J, Kernebeck S, Jux C, Weitz J, Ehlers JP, et al. Approaches to improvement of digital health literacy (eHL) in the context of person-centered care. Int J Environ Res Public Health 2022 Jul 07;19(14):8309 [FREE Full text] [doi: 10.3390/ijerph19148309] [Medline: 35886158]
60. Rocholl M, Wilke A, Meyer J, John SM, Ludewig M. Illness perceptions of patients with occupational skin diseases in a healthcare centre for tertiary prevention: a cross-sectional study. Int J Environ Res Public Health 2023 Apr 26;20(9):5652 [FREE Full text] [doi: 10.3390/ijerph20095652] [Medline: 37174171]
61. Fox S, Brown LJ, Antrobus S, Brough D, Drake RJ, Jury F, et al. Co-design of a smartphone app for people living with dementia by applying agile, iterative co-design principles: development and usability study. JMIR Mhealth Uhealth 2022 Jan 14;10(1):e24483 [FREE Full text] [doi: 10.2196/24483] [Medline: 35029539]
62. Poulsen A, Hickie IB, Alam M, LaMonica HM. User experience co-design of a mobile application to support childrearing in low- and middle-income countries. Stud Health Technol Inform 2023 Jun 22;304:86-90. [doi: 10.3233/SHTI230377] [Medline: 37347576]
63. Vargas C, Whelan J, Brimblecombe J, Allender S. Co-creation, co-design, co-production for public health - a perspective on definition and distinctions. Public Health Res Pract 2022 Jun 15;32(2):3222211 [FREE Full text] [doi: 10.17061/phrp3222211] [Medline: 35702744]

XSL•FO
RenderX

Abbreviations

AD: atopic dermatitis
BCT: behavior change technique
iDerm: Institute for Interdisciplinary Dermatological Prevention and Rehabilitation
MiA: Mein Hautschutz im Alltag ("My skin protection in everyday life")
TIP: tertiary individual prevention
uMARS: user version of the Mobile Application Rating Scale
UX: user experience
WRSD: work-related skin disease

Edited by A Mavragani; submitted 24.09.24; peer-reviewed by K Stawarz; comments to author 03.02.25; revised version received 24.02.25; accepted 25.02.25; published 18.04.25.

Please cite as:
Ristow N, Wilke A, Skudlik C, John SM, Ludewig M
User Experience of and Adherence to a Smartphone App to Maintain Behavior Change and Self-Management in Patients With Work-Related Skin Diseases: Multistep, Single-Arm Feasibility Study
JMIR Form Res 2025;9:e66791
URL: https://formative.jmir.org/2025/1/e66791
doi:10.2196/66791
PMID:

©Nele Ristow, Annika Wilke, Christoph Skudlik, Swen Malte John, Michaela Ludewig. Originally published in JMIR Formative Research (https://formative.jmir.org), 18.04.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Designing Health Care Provider–Centered Emergency Department Interventions: Participatory Design Study

Woosuk Seo[1*], BS; Jiaqi Li[1*], MS; Zhan Zhang[2], PhD; Chuxuan Zheng[3], BA; Hardeep Singh[4], MD, MPH; Kalyan Pasupathy[5], PhD; Prashant Mahajan[6], MPH, MBA, MD; Sun Young Park[7], PhD

[1]School of Information, University of Michigan, Ann Arbor, MI, United States

[2]Seidenberg School of Computer Science and Information Systems, Pace University, New York, NY, United States

[3]Department of Human Centered Design & Engineering, University of Washington, Seattle, WA, United States

[4]Center for Innovations in Quality, Effectiveness and Safety (IQuESt), Michael E. DeBakey Veterans Affairs Medical Center and Baylor College of Medicine, Houston, TX, United States

[5]Biomedical and Health Information Sciences, University of Illinois Chicago, Chicago, IL, United States

[6]Department of Emergency Medicine, University of Michigan Medical School, Ann Arbor, MI, United States

[7]School of Information, Stamps School of Art and Design, University of Michigan, Ann Arbor, MI, United States

[*]these authors contributed equally

Corresponding Author:
Woosuk Seo, BS
School of Information
University of Michigan
105 S State St
Ann Arbor, MI, 48104
United States
Phone: 1 2063109264
Email: seow@umich.edu

Abstract

Background: In the emergency department (ED), health care providers face extraordinary pressures in delivering accurate diagnoses and care, often working with fragmented or inaccessible patient histories while managing severe time constraints and constant interruptions. These challenges and pressures may lead to potential errors in the ED diagnostic process and risks to patient safety. With advances in technology, interventions have been developed to support ED providers in such pressured settings. However, these interventions may not align with the current practices of ED providers. To better design ED provider–centered interventions, identifying their needs in the diagnostic process is critical.

Objective: This study aimed to identify ED providers' needs in the diagnostic process through participatory design sessions and to propose design guidelines for provider-centered technological interventions that support decision-making and reduce errors.

Methods: We conducted a participatory design study with ED providers to validate their needs and identify considerations for designing ED provider–centered interventions to improve diagnostic safety. We used 9 technological intervention ideas as storyboards to address the study participants' needs. We had participants discuss the use cases of each intervention idea to assess their needs during the ED care process and facilitated co-design activities with the participants to improve the technological intervention designs. We audio- and video-recorded the design sessions. We then analyzed session transcripts, field notes, and design sketches. In total, we conducted 6 design sessions with 17 ED frontline providers.

Results: Through design sessions with ED providers, we identified 4 key needs in the diagnostic process: information integration, patient prioritization, ED provider-patient communication, and care coordination. We interpreted them as insights for designing technological interventions for ED patients. Hence, we discussed the design implications for technological interventions in four key areas: (1) enhancing ED provider–ED provider communication, (2) enhancing ED provider-patient communication, (3) optimizing the integration of advanced technology, and (4) unleashing the potential of artificial intelligence tools in the ED to improve diagnosis. This work offers evidence-based technology design suggestions for improving diagnostic processes.

XSL•FO
RenderX

Conclusions: This study provides unique insights for designing technological interventions to support ED diagnostic processes. By inviting ED providers into the design process, we present unique insights into the diagnostic process and design considerations for designing novel technological interventions that meet ED providers' needs in the diagnostic process.

International Registered Report Identifier (IRRID): RR2-10.2196/55357

(JMIR Form Res 2025;9:e68891) doi:10.2196/68891

KEYWORDS

emergency departments; participatory design; health care providers; technology; interventions; artificial intelligence

Introduction

Background

Emergency departments (EDs) are dynamic, challenging, and time-critical medical environments in a hospital. In recent years, the increasing need for emergency care services has led to overcrowding [1-3], resulting in extended wait times for patients, diminished patient satisfaction [4], and suboptimal patient outcomes [5,6]. With the high volume of patients and staff shortages [7], health care providers in the ED face significant challenges in their workflow [3,8]. For example, they often encounter frequent distractions and interruptions from secondary tasks or transitions [9-11], ineffective communication with patients [12-14], communication breakdowns, and information loss during the care process [15-18], all while being overwhelmed by the intensive patient and teamwork information [15,19]. All these challenges and issues can potentially lead to diagnostic errors [20-23].

Given these challenges in the diagnostic process in EDs, previous work has attempted to implement different types of technology to enhance this process. In particular, clinical decision support systems (CDSSs) have been implemented to predict morbidity and mortality [24-27], sepsis [28], and adverse prognosis [29]; improve triage [30-35]; automate clinical documentation [36,37]; and predict hospitalization and admission [38-41]. Although these systems enhance ED providers' decision-making process, many still face barriers to practical applicability and integration into ED workflows [28,42]. For instance, a previous study found that emergency physicians poorly accepted an evidence-based CDSS for evaluating suspected pulmonary embolism because it increased computer time compared to the original workflow, leading to low adoption rates in the ED [43].

One of the main reasons for the failure to adopt and implement effective technological interventions in ED is the limited understanding of ED providers' technology needs and examining appropriate approaches for integrating advanced technologies into dynamic ED workflows [44]. A way to understand the needs of ED providers is to involve them in the design process [45]. Participatory design (PD) [45] is a methodology that engages all stakeholders in the design process to create solutions that address their needs. Previous work has suggested that participatory and user-centered design approaches can address these concerns [45,46] as they engage all stakeholders in the early stages of the design process, thereby more effectively and promptly addressing user needs and improving the alignment of new technologies with existing workflows [47]. Examples of adopting PD approaches in ED technology development include collaborating with ED providers to design information displays that support awareness and enhance ED teamwork [48] and redesigning the ED patient health information system [49]. Following this body of work, we used the PD approach in our study to explore potential improvements and technological interventions for the ED diagnostic process, an area previously unexplored using the PD method.

Objectives

In our study, we explored the following research questions: (1) what are the challenges and needs of frontline providers in the ED diagnostic process? and (2) how should technological interventions be designed to address the specific needs of ED providers? To answer these questions, we conducted interviews and PD sessions, engaging ED providers from different hospitals to identify opportunities for creating user-informed technological interventions and to brainstorm strategies for enhancing the diagnostic process in ED. Our study identified 4 primary areas for enhancing the diagnostic process in ED: information integration, patient prioritization, ED provider-patient communication, and care coordination. The findings also highlighted the key concerns about integrating advanced technologies into dynamic ED workflows. Finally, we conclude the paper by discussing design implications for (1) enhancing ED provider-ED provider care coordination and communication, (2) enhancing ED provider-patient communication, (3) optimizing the integration of advanced technology in the ED, and (4) unleashing the potential of artificial intelligence (AI) tools in the ED to improve diagnosis. This work will enhance ED care by offering evidence-based technology recommendations and establishing ED-specific design principles for improving diagnostic processes and care.

Methods

Study Overview

This research is part of a comprehensive initiative examining the perspectives of ED patients and health care providers regarding technological interventions to enhance the ED diagnostic process. The broader project aims to develop design guidelines for interventions that address both stakeholder groups' needs. This paper specifically focused on validating ED providers' needs in the ED diagnostic process and their views on technological support systems. Through 6 PD sessions involving 17 frontline ED providers (n=9, 53% physicians and n=8, 47% nurses), we collected and analyzed session transcripts, design sketches, and field notes. Our analysis revealed 4 key themes that encompassed patient needs, coping strategies for

XSL•FO
RenderX

common challenges, and design recommendations for future technological interventions.

Design Idea Generation Phase for ED Care Interventions

This PD study is part of a larger research project that aims to study diagnostic errors during the ED care process that involves multiple stakeholders, including patients, informal caregivers, nurses, and physicians. Before our PD study with ED providers, the research team interviewed 17 frontline ED providers (n=6, 35% physicians and n=11, 65% nurses) to better understand their experiences and challenges during the ED diagnostic processes. From the interviews, we identified difficulties, emerging patterns of complaints, and general levels of satisfaction with different aspects of the care process (Table 1).

Table 1. List of identified problem categories from our previous study's [50] emergency department (ED) health care provider interview data.

Roles	Problems
Nurses	• Missed patient reassessments due to demanding clinical workload. • Limitations in tracking patient reassessment timing and clinical status changes.
Physicians	• Clinicians' high levels of stress can impede decision-making and focus on ED work. • Physicians' high cognitive load may interfere with their ED diagnosis work. • Lack of decision support tools that aid in diagnostic decision-making for increased accuracy. • Difficulty for ED providers to access scattered patient history in a concise and easy-to-read format.
Both nurses and physicians	• Insufficient communication between physicians and nurses about orders and next steps in patient care and a lack of EHR[a] support for such communication. • Insufficient communication between physicians and nurses about the patient's diagnosis and no established opportunity to discuss diagnoses before discharge. • Lack of notification and information about incoming patients with critical care needs. • Acuity level differs between nurses and physicians. Physicians sometimes have to reassign acuity levels mentally.

[a]EHR: electronic health record.

In addition to the known problems, such as ineffective care team communication, ED providers faced challenges in patient assessment and prioritization before triage, decision-making for patient disposition, and high cognitive load among ED providers. On the basis of the findings, the research team brainstormed numerous design ideas for each problem category, focusing on frontline nurses and physicians. We then merged the design ideas based on feasibility and usefulness. Finally, we narrowed the list and finalized the 9 most effective intervention ideas (Table 2). Each intervention idea aimed to address at least 1 problem category.

On the basis of the insights from our interview study, we generated many technological intervention ideas by conceptualizing potential technological interventions that could address the challenges identified in the ED diagnostic process. In addition, we conducted an extensive literature review to understand the status quo, such as the types of technological interventions that had been explored in previous studies and the ED providers' perspectives on adopting technologies in their workflow (eg, concerns and considerations). These efforts led to the initial generation of >80 intervention ideas, which were then iteratively and collaboratively refined by 3 research team members (WS, JL, and CZ) and 2 medical domain experts. This refinement process resulted in the development of 9 ED provider-facing intervention ideas (Table 2). These interventions were subsequently used to inform and guide participant discussions during PD sessions, which we described in the PD Session Procedure section. To visually illustrate each intervention, we created storyboards for the intervention ideas. A storyboard describes an example of an ED provider facing one or more challenges mentioned above and using the intervention to mitigate the problem (see Figure 1 for a sample storyboard).

XSL•FO
RenderX

Table 2. Technology interventions produced in this study along with questions for participant discussion. Participants were asked to rank them based on their preferences and feasibility.

Technology intervention	Description	Sample of lead question used for the discussion of patient need
TI1: machine learning technology for diagnosis support	This intervention introduces an AI[a]-driven tool designed to assist ED[b] health care providers in diagnosing patients more accurately and efficiently. The system operates by analyzing a combination of patient symptoms, health history, and real-time data, such as nurses' notes and laboratory results.	"Do you wish you would have more time to spend on complex diagnoses?"
TI2: diagnostic safety dashboard	This intervention introduces a diagnostic safety dashboard designed to assist ED physicians in prioritizing patient care by visualizing risk levels in a clear, color-coded format.	"Do you find it difficult to prioritize patient care mentally based on patient conditions and risk factors?"
TI3: predischarge team huddle	This intervention idea evaluates patient charts for high-risk diagnostic errors before discharge (eg, general diagnosis and other criteria). If flagged, it prompts the care team to initiate a huddle, during which ED providers can review a summary of patient information and assign diagnoses together. This collaborative process ensures alignment on the final diagnosis, reducing the likelihood of errors and improving patient safety.	"Do you feel that more communication is needed between members of the care team about the patient's diagnosis?"
TI4: visual timeline of the patient's ED visit	This intervention enhances physician-nurse communication and EHR[c] support by visualizing the patient's visit timeline. Nurses can easily access physicians' orders, notes, and the next steps for patient care, along with key information, such as check-in details, vitals, PERC[d] scores, and CT[e] scan results, all in timeline format. This information is accessible both at the nurses' station and at the patient's bedside on screen, ensuring that everyone stays informed about orders and the next steps.	"Do you struggle to stay up to date on orders and decisions being made for a patient?"
TI5: patient history aggregator	This intervention addresses the challenge of accessing medical records for new patients with complex medical histories. The PHA[f] integrates a fingerprint scanner to ensure patient safety and accurately match the patient to their health records from previous hospitals. This allows ED providers to quickly access crucial information about the patient's heart disease and other past conditions, ensuring informed and efficient diagnostic decisions.	"Do you find it difficult to review a patient's history in a summarized format, especially when it's scattered across multiple EHR systems?"
TI6: EHR-prompted diagnostic pause	This intervention features an EHR-integrated diagnostic pause to reduce errors in complex cases. When a patient presents with symptoms such as abdominal pain and fever, the initial ED provider records their diagnosis. For complex cases, the EHR alerts the next ED provider to independently evaluate the patient before viewing the first diagnosis, which remains hidden until completion. If diagnoses differ, the system recommends team discussion. The alert links to relevant patient information for easy reevaluation.	"Is there a higher risk of diagnostic error for more complex cases?"
TI7: real-time EMS[g] information via smart glasses	This intervention addresses the challenge of unexpected critical patient arrivals by enhancing communication between EMS and the ED care team. Using smart glasses, EMTs[h] can initiate a real-time video call with the ED care team, including physicians and nurses. This allows the care team to view the patient's condition from the EMT's perspective and receive critical information en route to the hospital.	"Do you wish you had more information about when patients are being transferred to the ED, and what sort of preparation is needed?"
TI8: reassessment reminder door flag	This intervention introduces a reassessment reminder system to ensure timely and efficient monitoring of patient conditions. When a time-sensitive reassessment, such as checking vital signs or evaluating treatment effectiveness, is due, a display monitor outside the patient's room flashes a green light as a visual alert. The monitor also displays key details, such as the medication administered and the task that needs to be completed. Nurses can see the flashing light from the nurses' station or other areas, prompting them to respond.	"Are you tired of relying on your memory and constantly looking at the clock in order to complete nursing tasks such as patient reassessments every hour or so?"
TI9: stress management tracking	This intervention introduces a stress management tracking system to monitor and support the well-being of health care providers during demanding shifts. The system includes a wearable device that tracks indicators, such as heart rate and sleep cycles. When abnormal stress levels or poor sleep patterns are detected, the device notifies the wearer with an alert, accompanied by a message suggesting a break or rest. This ensures that ED providers' health is prioritized, reducing stress and fatigue while improving focus, performance, and overall cognitive function during their shifts.	"Have you ever felt so stressed that you were not able to focus or found it difficult to make decisions?"

[a]AI: artificial intelligence.

[b]ED: emergency department.

[c]EHR: electronic health record.

[d]PERC: pulmonary embolism rule-out criteria.

[e]CT: computed tomography.

XSL•FO
RenderX

[f]PHA: patient history aggregator.

[g]EMS: emergency medical service.

[h]EMT: emergency medical technicians.

Figure 1. A sample storyboard illustrating how an interactive dashboard for patient risk scores can support the diagnosis process for emergency department (ED) health care providers. EHR: electronic health record.

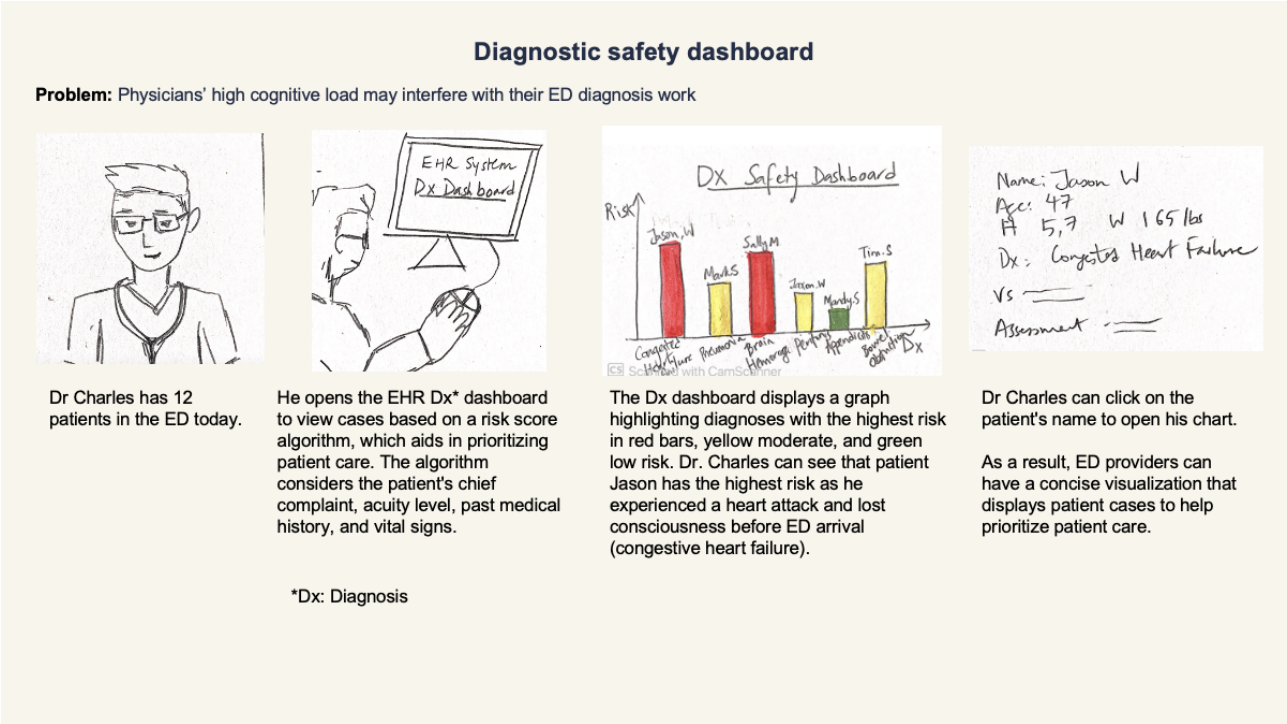

Participant Recruitment

Participants in the study were recruited from 2 large health care organizations in the United States. Organization A is a university-affiliated medical institution in the Midwest, and organization B is a renowned, comprehensive medical center also located in the Midwest. Eligible participants included ED physicians and nurses who had worked in the ED setting for at least 1 year. A study coordinator contacted potential participants via email or in person on the day of their shift. Once eligible ED providers expressed interest in participating, we obtained electronic informed consent by emailing the form to them before each session. If participants did not complete the form in time, we provided a paper copy at the start of the session for in-person completion. All PD sessions were audio recorded and transcribed for data analysis. In total, 17 clinicians (n=9, 53% physicians and n=8, 47% nurses) participated in 6 PD sessions.

PD Session Procedure

We conducted a total of 6 PD sessions. Each session consisted of 3 to 4 participants with the same or similar roles (eg, nurses or physicians) to mitigate the potential influence of power dynamics on the expression of opinions [51,52]. The participants' characteristics varied in experience, from novices to experts with >30 years of experience in the field, as detailed in Table 3. The sessions were conducted either in person or online via Zoom (Zoom Communications, Inc), depending on the participants' preferences and availability. We provided colored pens, paper, and sticky notes for in-person sessions, while for virtual sessions, we used slide presentations and a Miro board (Miro, Inc) as interactive whiteboards for the participants' design collaboration. In addition, 2 to 3 researchers were present at each session, taking detailed notes.

Each PD session lasted up to 120 minutes and comprised 3 major parts: intervention idea discussion, individual design, and group design. Specifically, the first part focused on the presentation and discussion of the 9 intervention ideas developed by the researchers. After presenting each idea, we used leading questions to assess the perspectives of ED providers (eg, needs, opinions, perceptions, and expectations) for each intervention idea and allowed them to rank the intervention ideas (Table 2). In the individual design activity, participants selected 1 or 2 intervention ideas that most resonated with their needs based on their personal perspectives and professional experiences. After choosing their preferred intervention ideas, they were asked to further improve the design by modifying or adding new elements and features or to design their own new intervention. For the group design activity, participants were asked to select a top intervention idea as a group and work together to synthesize their individual design ideas and develop a final design solution (see Multimedia Appendix 1 for sample group designs). Participants engaged in discussions on the salient features, pros and cons, and concerns associated with their final group design. Each session was concluded by eliciting participant feedback to improve future sessions.

XSL•FO
RenderX

Table 3. Demographic information of study participants for each design session.

Session	Format	Affiliation	Occupation	Participant ID	Experience (y)
Session 1	Remote	Organization A	Physicians	Participant 1	20
Session 1	Remote	Organization A	Physicians	Participant 2	5
Session 1	Remote	Organization A	Physicians	Participant 3	16
Session 2	In person	Organization A	Physicians	Participant 4	19
Session 2	In person	Organization A	Physicians	Participant 5	31
Session 2	In person	Organization A	Physicians	Participant 6	2
Session 3	In person	Organization A	Nurses	Participant 7	7
Session 3	In person	Organization A	Nurses	Participant 8	15
Session 4	Remote	Organization B	Physicians	Participant 9	15
Session 4	Remote	Organization B	Physicians	Participant 10	26
Session 4	Remote	Organization B	Physicians	Participant 11	15
Session 5	Remote	Organization B	Nurses	Participant 12	8
Session 5	Remote	Organization B	Nurses	Participant 13	9
Session 5	Remote	Organization B	Nurses	Participant 14	20
Session 6	Remote	Organization B	Nurses	Participant 15	30
Session 6	Remote	Organization B	Nurses	Participant 16	3
Session 6	Remote	Organization B	Nurses	Participant 17	2

For the in-person sessions, we captured audio recordings and took photographs of the participants' sketches and designs. For virtual sessions, video recordings were made via Zoom, and participants' digital creations were also recorded for further analysis. In addition to the recordings, session facilitators took field notes throughout the sessions for data analysis.

Data Analysis

We used transcripts as the primary data source, and all other collected data (eg, observation notes) were used as supplementary sources. We also incorporated the co-design activity sketches as supplementary data to validate what participants discussed during the session and clarify participants' design suggestions.

After organizing the collected data, 2 researchers open-coded the transcripts of the first 3 PD sessions. The researchers also analyzed content from participants' sketches and design illustrations, converting these visual elements into textual descriptions to facilitate thorough examination. Each sketch or digital artifact was meticulously analyzed to extract key features, such as design elements, functional aspects, and underlying needs. The codes were compiled into a growing codebook organized by their corresponding intervention or put into a general category. On the basis of this evolving codebook, the remaining transcripts were then coded. If any new codes were discovered, they were either added to the codebook as a new code or incorporated into the definition of an existing, similar code.

We then used affinity diagramming [53] to identify the emerging themes from the codes. We used the Miro board for affinity diagramming. Similar codes were first grouped together into subthemes specific to each intervention. Then, a team of researchers engaged in a series of discussions to build larger, emerging themes that spanned across the different interventions. For example, a larger theme of "communication among the care team members" contained various subthemes, such as "piece together information about a patient and keep every care team member on the same page." Codes that were ambiguous or needed further support were refined by referencing participant quotes from transcripts and participants' sketches from PD sessions. After multiple group discussions, we identified 4 themes that described ED providers' challenges in the diagnostic process.

Ethical Considerations

We complied with the following ethical considerations. First, we obtained consent from all participants before the study sessions, which were approved by the multisite institutional review board at the University of Michigan (HUM00156261). Participants were informed that their participation was voluntary, the sessions would be recorded, the collected information would be deidentified and protected in secured storage, and they would be compensated with US $100 for their participation. Second, we ensured that the collected transcript data were deidentified by replacing participants' names with pseudonyms, and that their faces in photos or videos were blurred. All collected data are stored in secure storage that requires the university accounts for access. Third, all the researchers completed research compliance training, such as the Collaborative Institutional Training Initiative or the Program for the Education and Evaluation in Responsible Conduct of Research, to become trained on best practices and ethical considerations for interacting with ED providers and patients.

XSL•FO
RenderX

Results

Overview

On the basis of the analysis of collected data, we identified top intervention ideas that participants preferred, 4 key areas in the ED diagnostic process requiring technological support, and major concerns regarding the integration of advanced technologies into dynamic ED workflows.

Throughout all sessions, participants collectively favored the following interventions to enhance the existing ED diagnosis process (Table 2): TI1 (machine learning technology for diagnosis support), TI2 (diagnostic safety dashboard), TI3 (predischarge team huddle), and TI4 (visual timeline of the patient's ED visit). Participants believed that these intervention ideas could improve crucial aspects of the ED diagnosis process. Specifically, TI1 could broaden access to comprehensive patient information and reduce potential errors. TI2 might simultaneously manage multiple patient cases, optimize prioritization, and rapidly identify high-risk patients. TI3 could address communication and understanding gaps between physicians and nurses regarding diagnoses, while TI4 aims to bridge the communication divide and keep patients informed about their care progress.

Furthermore, we reported the key areas of improvement in the diagnostic process, which indicated the challenges faced by ED providers in the diagnostic process. We present 4 prominent aspects extensively discussed among participants in the subsequent sections.

Effective Integration of Patient Information for Timely and Accurate Diagnosis

Participants often faced challenges with incomplete, hard-to-find, or misleading patient information, which can inadvertently lead the medical team down an incorrect diagnostic pathway. Incompleteness can result from prehospital care providers failing to relay all necessary information at handoff or integration issues among health care information systems. Assessing relevant information is problematic when stored using varied and repetitive terminologies or buried in multilevel systems. Misleading information typically occurs with objective data, such as vital signs. For example, external stress or the use of an incorrectly sized cuff can result in elevated heart rates or blood pressure readings, potentially skewing risk assessments.

To access comprehensive patient information effectively, participants mentioned that both historical information (eg, past medical history, long-term medications, patients' care preferences, and social determinants) and ad hoc information (eg, laboratory test results, vital signs, and clinical observation notes) should be integrated to make an accurate diagnosis. For instance, Participant 16 suggested how TI5 (patient history aggregator) could be integrated with existing care systems to help ED providers in the diagnosis process as follows:

> *If we could have that [History Aggregator] page already incorporate all the other systems where this patient has received care, I see that as helpful. It's kind of like, in my mind, redefining the chart review tab in a more encompassing form.* [Participant 16]

Some participants also highlighted the dynamic changes in physiological values when redesigning TI1 (machine learning technology for diagnosis support). To maintain accurate risk assessments, it is critical to continuously update data reflecting the dynamic changes in patient conditions, especially for those patients with critical illness whose health status can rapidly evolve. This suggestion showed participants' desire to receive updates about the patient's status for better management, as many patients are coming in and out of the ED.

Some participants suggested that medical information should be displayed in a simple, clear, and efficient manner, such as streamlining all necessary information onto 1 page to avoid spending time searching and organizing it piece by piece. Participant 5 highlighted the importance of information presentation for TI4 (visual timeline of the patient's ED visit) as follows:

> *It's information that we just want to be displayed in a more graphic, simpler form where it's not so busy. You open it up, it's one snapshot, and you click on which part of it you want to do now...It's just organized in a better way, in a simpler way.* [Participant 5]

Participant 5 envisioned that this approach would enable health care providers to efficiently focus on critical details, such as a patient's historical data, directly relevant to their chief complaint.

Optimizing Patient Prioritization to Improve the Triage and Care Process

Another important area to improve was related to prioritizing patients. The current triage process may lead to the improper prioritization of patients with critical needs. Triage in the ED can be influenced by many factors, including triage nurses' clinical knowledge and judgment, the clarity with which patients describe their symptoms, and the availability of patient records or medical histories. In addition, the Emergency Severity Index (ESI) classification criteria, which classify and prioritize patients based on the severity of their disease and the estimated number of resources needed for their care, can fail to account for dynamic changes in a patient's condition. For example, there was no adaptation of the ESI score despite clear signs that the patient had become more critically ill. For these reasons, there is a risk of patients being misclassified and inappropriately prioritized, which may result in those with severe conditions facing longer wait times.

To address those issues in the triage process, our participants expected the AI-driven prioritization system to analyze various types of information, including contextual (patient's relevant medical history and symptoms), situational (dynamic changes in patients' symptoms and length of waiting time), and human factors (clinician's knowledge and experience). This ensures timely attention to critical cases and improves ESI's shortcomings. For instance, when discussing idea TI2 (diagnostic safety dashboard), participants stated that the algorithm could consider different and comprehensive factors

XSL•FO
RenderX

to identify higher-risk individuals, such as those with a history of asthma or cardiac problems. A participant noted as follows:

> *The algorithm looks at the chief complaint, acuity level, and past medical history. I think that helps as well. For example, if it's a younger person with a syncopal episode and you find out they have a previous cardiac history, that would make them a higher risk than, say, a 30-year-old who passed out for another reason. I like how this algorithm considers the whole patient and takes that into account.* [Participant 12]

In addition, many participants valued the explicit representation of different patient prioritizations through various colors (eg, red, yellow, and green) and symbols (eg, an alarm bell and a smiley face) because it could enable ED providers to identify which patients require immediate attention visually quickly. For instance, Participant 8 preferred color coding of data visualization for immediate understanding of priority as follows:

> *I like the color because then they could look at that and say that they're at high risk. If they're above five risk scores then they are red or something, and then they're like, "Oh, I have all these priority 2s as Joe is sicker."* [Participant 8]

Enhancing Patient Engagement in the Diagnostic Process

Our participants mentioned that it could be challenging for ED providers to receive timely diagnostic feedback and updates on changing symptoms from patients due to time constraints and heavy workloads in the ED. ED providers often rely on brief interactions or 1-way communications to gather information, which may lead to an incomplete understanding of patients' evolving symptoms. In addition, patients frequently feel anxious and frustrated due to the lack of ongoing updates about their condition, which can lead to increased complaints and misunderstandings.

Many participants agreed that informing patients about the status and progression of their care could help manage their expectations and alleviate anxiety. They expected patients to feel reassured that ED providers would still take care of them when patients understood the next steps in the diagnosis process. Participant 2 stated as follows when discussing idea TI4 (visual timeline of the patient's ED visit):

> *The patient can visualize what's been done and what they're waiting for. It sort of wards off that sentiment of, "I went to the ER, and they didn't do anything," and also gives them an idea of what the next step is.* [PD-2]

By sharing the ED process steps, participants anticipated more patient engagement and having more time to focus on effective patient communication.

Participants mentioned they wanted easy-to-use tools to help receive more accurate, timely updates directly from patients to aid the care process. Nurses' work could be expedited by having patients more actively involved in their care and providing information. For example, participants suggested that the idea TI8 (reassessment reminder door flag) could include a feature allowing patients to undertake a self-reassessment of their current condition and medication reactions and relay their results and updates to nurses, potentially via an automated prompt and chat system on a tablet or phone. If a nurse records the administration of fentanyl at noon, the system might automatically inquire about the patient's pain levels 20 minutes later. Then, patients could interact with the nurse, who can gauge the treatment's effectiveness, as noted by a participant:

> *I wonder if there's a way for the patient to enter a self-reassessment like they have their own little iPad and can share how they're doing, whether the pain medication helped their pain rating...it gives the patient the opportunity to provide feedback.* [Participant 9]

Our participants also mentioned that engaging patients in diagnostic discussions could be crucial to reducing potential diagnostic errors. This could allow ED providers to elucidate the logic behind a diagnosis, articulate how tailored care plans were developed for individual patients, and seek the patient's perspective regarding the proposed plan. In addition to the illustrated functionality of increasing care team communications to reduce diagnosis errors from ideas TI3 (predischarge team huddle) and TI6 (electronic health record (EHR)–prompted diagnostic pause), participants suggested empowering patients to contribute supplementary information and express different viewpoints that might not have been considered, to reduce diagnosis errors. For instance, Participant 3 emphasized that patients should have the opportunity to share their perspectives on ED providers' reasoning as a way to promote patient engagement in diagnostic discussions as follows:

> *If the patients heard the diagnostic reasoning, you [clinicians] would be giving them an opportunity to correct you [clinicians].* [Participant 3]

Care Coordination Among the ED Team Members

Through the sessions, we identified that the ED care team faced 2 main communication issues. First, each member of the ED care team has only partial information, leading to a disjointed and fragmented understanding of the patient's case, primarily due to interactions at different stages of care. Patients move from 1 ED provider to another, and no one fully knows all the details of the patient's condition. Furthermore, its interpretation can differ even when information is consistently shared among team members. This variation in understanding can be attributed to individual backgrounds, areas of expertise, or perspectives. Second, physician-to-nurse communication could sometimes become ineffective, especially regarding discharge and diagnostic reasoning information. For example, physicians sometimes fail to communicate the rationale behind diagnoses, orders, or care plans to nurses. This issue can impede nurses' understanding of physician decisions and their ability to respond accurately to patient queries. Furthermore, participants acknowledged that this communication issue may be derived from both sides, as noted by participant 7:

> *[Physician-nurse communication] is a mutual problem, it's not a one-sided issue.* [Participant 7]

XSL•FO
RenderX

All participants agreed on the importance of maintaining a consistent and uniform understanding among care team members, especially for complex patient cases. Engaging more care team members in discussions could be advantageous. When discussing idea TI3 (predischarge team huddle), participants noted that it would offer an excellent opportunity for comprehensive information gathering from different team members and roles to minimize potential errors. A participant highlighted as follows:

> *You would get everyone on the same page and give the patient a chance to ask questions. At the last minute, if someone asks, "Well, what about this, or what about that?" everyone is on the same page, which I like.* [Participant 2]

Most physicians and all the nurses mentioned the need for physicians to communicate more detailed information with nurses, including the reasoning behind diagnoses, orders, and care plans, to enable nurses to understand the overall situation comprehensively. When discussing idea TI4 (visual timeline of the patient's ED visit), physicians believed that this technological intervention could be instrumental in keeping nurses informed about care plans and the rationale behind physicians' orders in a timely manner. This enhanced transparent communication could empower nurses to answer patients' queries more effectively, reducing the need for ambiguous or noncommittal responses. Nurses, in particular, advocated this because they wished to better explain physicians' orders and diagnoses to patients, as one participant noted:

> *Sometimes I go into the patient room and say, "I think they [the physicians] ordered this because they're probably looking to see if you [the patient] have this issue." It would be nice to say confidently, they ordered this because they're looking to see that you have this issue.* [Participant 8]

Among our participants, some care team members might hesitate to point out potential errors, from junior members to senior members or nurses to physicians. For example, nurses might be reluctant to bring up potential diagnostic errors to physicians due to a perception that physicians, being more knowledgeable and experienced in diagnosis, are responsible for treatment decisions. However, the participants believed that some technologies, acting as second opinions, might help address the power dynamic issue. For example, when discussing the idea TI1 (machine learning technology for diagnosis support), nurses expressed that the list of diagnoses generated by machine learning tools could empower them to engage in discussions with physicians regarding potential misdiagnoses significantly when the physician's conclusion deviates from the algorithm's suggestion, as noted by a participant:

> *It [the machine learning technology tool] also brings up good discussion points with the physician...why are we pursuing this route...or why we're deviating from a particular diagnosis.* [Participant 14]

Concerns of Integrating Advanced Technologies in ED

Overview

Our 9 intervention storyboards described how emerging technologies can potentially be designed to support the ED diagnostic process. While the storyboards were primarily used to validate ED providers' needs, they were also used for discussions about potential issues of integrating such technological interventions. This section presents participants' concerns regarding integrating advanced technologies in ED, including the safe use of AI, privacy, and additional cognitive and operational workload.

Safe Use of AI-Empowered CDSSs

With the growing application of AI techniques in health care, more than half of our intervention ideas were inspired by AI-empowered tools. Thus, many participants expressed concerns about the AI's potential inability to consider different care contexts and individual nuances of each patient case in analyzing and presenting various information. This concern was mainly raised during group design activities on several AI-driven interventions, such as TI1 (machine learning technology for diagnosis support), TI2 (diagnostic safety dashboard), and TI4 (visual timeline of the patient's ED visit). Participants particularly mentioned that the same test result can be interpreted differently based on the context. For instance, a particular urine analysis result can indicate different health conditions depending on whether the patient has a positive or negative ultrasound result, as noted by a participant:

> *The same urine analysis result means something totally different if you tell me that the patient has an ultrasound that's clearly positive than if they have an ultrasound that's clearly negative, and it can have the exact same result. And that's the problem. It comes in context.* [Participant 4]

Participants suggested that AI systems should account for various factors and contextual cues, such as interpreting test results with other diagnostic information and medical history. However, it may not fully match human ED providers' ability to perceive subtle cues and insights, which could be beyond the scope of computers or AI systems. In addition, participants suggested that the AI systems should apply different weights to different diagnostic factors (eg, laboratory results and medical history) because certain factors weigh more than others, or the system should adjust weights based on dynamic emergency severities when calculating risk scores.

In addition, participants underscored the imperative need for rigorous testing and validation before implementing any AI tool in their existing work practice. They believed that an AI tool must prove its reliability and efficacy to gain the trust of medical providers, as noted by a participant:

> *Critical for it to be, the tool really has got to be tested and robust before it's implemented. Otherwise, no one's going to entertain the idea of adopting it because it's just going to create extra work without perceived value.* [Participant 10]

XSL•FO
RenderX

Participants also expressed concerns that an overreliance on technology could potentially lead to a degradation of human clinical skills and emphasized the balance between such emergent technology use and human skills. For instance, when discussing the idea TI1 (machine learning technology for diagnosis support), there was a worry that ED providers might depend too heavily on technology and neglect their assessments. Participant 13 expressed her concern about potential overreliance on AI tools as follows:

> *The only concern I can think of right now is...becoming too reliant on AI...If any residents or doctors would start to rely on it too much...you'd hope not, you'd hope they do their own assessments.* [Participant 13]

As a potential solution to mitigate such overreliance, some participants suggested that they would like to incorporate AI-powered CDSSs into their on-the-job training. This integration could allow ED providers to learn from these tools and understand how to collaborate with them effectively without overly relying on AI outputs while maintaining their clinical assessment skills.

Privacy Concerns Toward Emerging Technologies

Participants repeatedly mentioned privacy concerns related to new advanced technologies. They raised issues about potential violations of the Health Insurance Portability and Accountability Act and patient confidentiality, emphasizing the need to ensure adherence to patient privacy laws and regulations. For example, TI5 (patient history aggregator) allows easy access to patient history by scanning patients' fingerprints. Many participants expressed concerns about collecting sensitive biometric data (ie, fingerprints) from their patients. Participant 16 brought up the issue of data storage that may lead to a privacy breach, as follows:

> *If it is based on fingerprint, where is the information being housed? Is it a cloud database? That is universal? Is it an extension of the electronic health record? Where does it belong? I think it will be a question.* [Participant 16]

Participants' such concerns about privacy highlight the importance of transparent communication regarding data storage practices, robust security measures, and clear protocols for accessing sensitive patient information to foster trust in emerging health technologies.

Extra Workload and Burden to Make Technology Fit Into Their Existing Workflow

Participants raised concerns about the potential for certain technological implementations to interfere with the established ED workflow, possibly creating extra waiting time for patients and adding extra cognitive and operational workload for ED providers. One notable instance occurred during the discussion of the idea TI3 (predischarge team huddle). Some participants expressed that integrating such a system could inadvertently delay or create a backlog in the patient discharge process. Given the current ED workflow, it could be challenging to convene all care team members due to their individual duties and potential busyness. A participant explained this concern by stating as follows:

> *When you have a patient who is itching to get out the door [being discharged], and you say, "while we're waiting for this person to show up or we need this other team member [to discuss again]," it can create a lot of delay and backlog if now you have to round to discharge everybody.* [Participant 2]

Another example arose during the discussion on the TI6 (EHR-prompted diagnostic pause), which highlighted concerns that its implementation could disrupt the typical ED workflow by introducing a "parallel" evaluation process. In this context, "parallel evaluation" means that multiple ED providers might independently diagnose the patient, leading to repetitive and time-consuming efforts. This could result in a prolonged diagnostic process and potentially delay treatment, as noted by a participant:

> *I think the scenario would require a change in our current flow, where it's sequential in terms of evaluations that are done of a patient...To one where there's a parallel evaluation, which I think becomes very inefficient and for a patient, can be quite redundant as they have to repeat the story multiple times.* [Participant 10]

Discussion

Design Implications

Overview

Our findings show that the ED diagnostic process can be improved into 4 main areas: information integration, patient prioritization, ED provider-patient communication, and care coordination. We also summarized the key concerns of ED providers in integrating advanced technologies in the ED, including the safe use of AI, privacy, and the additional cognitive and operational workload. Drawing from these findings, we discuss potential design implications of future ED-centered intervention tools for improving the ED diagnosis process.

Creating Standardized Communication Procedures

We observed that there is significant demand for improving communication between ED providers (the idea TI3 [predischarge team huddle] was ranked third by participants, and TI4 [visual timeline of patient's ED visit] was ranked fourth by participants). Information sharing between ED physicians and nurses is often inadequate, limiting nurses' comprehensive understanding of patient conditions. In addition, varying levels of understanding about the patient exist among different stages in the care team, further complicating the communication process. These findings highlight the importance of establishing standardized communication procedures to ensure a unified understanding of patient conditions among team members. We suggest creating more standardized communication procedures in future technical tool designs. Building on previous insights from the design of digital checklists for trauma resuscitation [54], we recommend implementing an automated checklist

XSL•FO
RenderX

system that includes critical patient information, such as diagnosis, medication orders, and care milestones. This checklist should be integrated with a timeline that automatically captures details of when each piece of information was added and when action needs to be taken, ensuring that all team members are aware of current in-progress and upcoming responsibilities. in addition, structured communication features, such as mandatory read-backs of critical information, regular briefings at shift changes, and digital alerts for key updates, can further enhance understanding and collaboration. Such tools are particularly helpful for nurses, who often coordinate various aspects of patient care and benefit from clear, concise, and timely information to make informed decisions and provide the best possible care. This potential application direction also echoes the previous study on the necessity of applying new technological communication interventions to simplify communication among various roles within the ED [55].

Incorporating Technology-Based Second Opinions

We identified power dynamics as a critical issue in the diagnostic process. Such skewed interactions are prevalent between junior and senior members, as well as between nurses and physicians. Hesitation in pointing out potential errors and reluctance to accept feedback is not uncommon, and it could have severe consequences (eg, increased incidence of medical errors and significant patient safety risks) in fast-paced and critical medical fields such as EDs [51,52]. Although more collaborative approaches, such as education and mentoring, have been proposed to address the issue of power dynamics [56], we found that this issue has not yet been fully resolved. Notably, our findings indicated the potential of technological interventions in mitigating these power dynamics. For instance, the timeline in idea TI4 (visual timeline of the patient's ED visit) could allow nurses to understand the rationale behind care plans, enhancing their ability to communicate with physicians actively, thus promoting equal and open communication within the team. Furthermore, as mentioned in idea TI1 (machine learning technology for diagnosis support), second opinions generated using machine learning tools could boost the confidence of junior physicians or nurses in discussing potential misdiagnoses with senior members or physicians in the care team. This echoes and expands on past research [57] about the second opinion in physician-patient relationships, demonstrating it to be an effective way to promote open, positive communication and to enhance satisfaction with clinical decisions. By providing second opinions and facilitating information sharing, technological interventions could help overcome the barriers in the traditional power structure of medical roles, potentially improving overall care quality and patient safety. Future work could involve designing and implementing such information tools in ED settings to evaluate potential impacts on the ED diagnostic process. For example, in the new ED medical display system, each team member could autonomously select CDSS reports that automatically highlight decision details differing from the primary physician's, with junior members and nurses also accessing these for reference and discussion as a second opinion.

Empowering Patients With Information Transparency

We found a demand for improved communication between ED providers and patients through timely information sharing, transparency with patients, and patient empowerment. Previous studies show that insufficient communication between patients and ED providers in the ED leads to dissatisfaction with their care and potential diagnosis errors [12,58]. Factors contributing to their communication issues include patients' and ED providers' conflicting perceptions of the ED, a lack of patient information, and physicians' diagnostic time pressures [12,58,59]. Our study further reveals the specific manifestations of insufficient communication in the ED between patients and ED providers: (1) patients are not fully informed about the treatment process; (2) lack of opportunities for patients to provide feedback and update their conditions to ED providers actively; and (3) absence of processes for joint decision-making and care planning between ED providers and patients, as well as effectively involving patients in diagnostic discussions. These communication deficiencies can lead to patient anxiety and an increased risk of misdiagnosis. Therefore, we suggest that future technological interventions should focus on enhancing timely information sharing and transparency with patients and significantly empowering patients' agency. This involves promoting active patient participation in communication, even in joint decision-making and care planning. Previous research [36] discussed a digital ED waiting room history-taking tool that uses reasoning engine functions to gather patient information. It has been proven to improve communication and understanding between patients and ED providers. Building on this foundation, we suggest developing a mobile app that allows patients to promptly report changes in their symptoms or express concerns using real-time data reflecting their current condition and personal feelings during their ED visits. The app would also integrate features that encourage patient involvement in diagnostic discussions. Such features could include a symptom timestamp for patients to log details of their symptoms as they occur and a communication chat window that allows for direct dialogue with ED providers.

Optimizing Advanced Technology Integration in the ED

Our research highlighted specific considerations for integrating advanced technologies into the ED context: minimizing additional workload and cognitive burdens for ED providers and using high-fidelity simulations. Unlike other general clinical practices where there is more time to adapt and use new technology, our findings indicated that ED providers in the fast-paced ED environment may not have the time to try out, thoroughly examine, and use these technologies. Furthermore, they were reluctant to change their existing workflows to accommodate new technologies, fearing the risk of additional workload. During the design of information displays, our ED providers preferred concise and practical interfaces, capturing essential information immediately rather than increasing their cognitive load by requiring them to understand the meaning of every icon and content. In addition, participants considered idea TI6 (EHR-prompted diagnostic pause) ineffective and redundant, despite its merit of catching potential diagnostic errors, as it does not fit their existing workflow. Similarly, during the discussion of idea TI3 (predischarge team huddle), participants

XSL•FO
RenderX

acknowledged the critical need for care team members to be on the same page regarding information sharing upon patient discharge. However, they still harbored concerns that the intervention would not align well with their current workflow or the realistic situation where patients would not want to stay longer and wait for a reevaluation before their discharge. Therefore, building on previous research [60] that emphasized the need for efficiency in the design of ED technological interventions, we recommend that future designs and implementations of ED technological interventions consider the fast-paced, high-pressure nature of the ED. This involves prioritizing efficiency and focusing on avoiding additional workload and cognitive burdens for each ED provider. For example, AI-driven decision support tools should be overseen by designated care team members who monitor AI-recommended diagnostics and conduct departmental audits and custom improvements to reduce the burden for individual ED providers. Similarly, designing interventions, such as the idea TI3 (predischarge team huddle), should minimize the potential for increasing workload. Instead of prompting the care team to initiate a team huddle to reevaluate, the design should allow members to quickly review summary reports in the EHR to identify potential issues with minimal additional effort.

From a methodological perspective, we also recommend building on existing ED workflows using methods such as “group workshops” similar to PD or “high-fidelity team-based simulations” conducted in clinical scenarios [61,62]. These methods could allow ED team members to experience new technologies and foresee how they can be implemented in their existing work practices while gathering feedback. High-fidelity team-based simulations have been widely used in the ED for effect evaluation and training, such as improving communication among ED provider teams [63] and detecting latent safety threats in critical patients [64], proving very effective. Therefore, we suggest extending this approach to introducing new technologies in the ED to guide the integration of new advanced technologies.

Leveraging AI Tools in the ED to Enhance Diagnosis Through Dynamic Patient Data

Our research highlighted a strong demand from our participants to unleash the potential of AI tools in the ED to improve diagnosis further. We found that participants highly valued the top 2 AI-related ideas (TI1 [machine learning technology for diagnosis support] and TI2 [diagnostic safety dashboard]) because of their potential to (1) integrate a broader range of comprehensive patient information and perform multidimensional data analysis and (2) facilitate dynamic prioritization and resource allocation. Although there are multiple existing AI applications in the medical field, many of them lack consideration of multidimensional information and dynamic prioritization [65] and thus do not fully unleash the potential of AI. To best leverage AI tools in future ED settings, we suggest that when designing algorithm development for ED, the input information for these algorithms should be multidimensional and incorporate dynamic patient data. These algorithms should be capable of continuously monitoring and evolving based on real-time patient data and feedback, which can aid in enhancing diagnostic accuracy and dynamically adjusting treatment protocols. For example, when designing AI tools for patient prioritization, it is crucial to implement a system that categorizes patients based on initial information and updates its prioritization as new data becomes available (eg, laboratory results or vital signs). This approach could ensure that resources are allocated promptly to patients with escalating needs.

Limitations

A few limitations of this study should be noted. First, one of the primary limitations of this study relates to the representativeness of the participant sample. Participants were recruited from only 2 large health care organizations in the Midwest region of the United States. While these organizations serve millions and can offer significant insights, this sample may not fully encapsulate the diversity and range of experiences of ED providers worldwide. There might be differences in health care systems, cultural contexts, and operational protocols in EDs across different countries or regions. Another limitation of this research is inherent in the study design, particularly regarding the PD sessions. Each session was constrained to a duration of around 120 minutes. This limited timeframe could have impacted the depth and extent of interaction among participants, potentially affecting their comfort levels and the quality of their responses. Participants might have required more time to fully engage with the material, express their thoughts, and collaborate effectively, influencing the innovation and creativity aspects of the session’s outcomes. Furthermore, some PD sessions were conducted in a virtual format. While virtual sessions offer flexibility and can include a broader range of participants, they might lack the richness of interaction found in face-to-face settings. Nonverbal cues, ease of spontaneous conversation, and the organic development of collaborative ideas are often diminished in virtual environments. The potential limitations in participant interaction and engagement due to the virtual format of some sessions might have affected the overall effectiveness and output of these PD sessions. Finally, our study is limited by not presenting functional systems but merely asking participants to imagine how the system works. Without seeing or interacting with a system, eliciting comprehensive user feedback could be challenging.

Conclusions

This work presents opportunities for improving ED provider–informed technological interventions and strategies for enhancing the diagnostic process in EDs. Our study identifies crucial enhancements in information integration, patient prioritization, ED provider-patient communication, and care coordination. The findings underscore key concerns about integrating advanced technologies into EDs. On the basis of these findings, this work discusses implications for designing future ED technological interventions, from enhancing communication to optimizing the integration of advanced technology and unleashing the potential of AI tools. Future efforts should further apply and test these design suggestions in real ED settings or continue using improved PD methods with functional systems that participants can interact with to further explore technological advancements in the ED diagnosis process.

XSL•FO
RenderX

Acknowledgments

This project was funded under grant number R18HS026622 from the Agency for Healthcare Research and Quality (AHRQ), US Department of Health and Human Services (HHS). HS is supported in part by the Houston Veterans Administration (VA) Health Systems Research and Development (HSR&D) Center for Innovations in Quality, Effectiveness and Safety (CIN13–413), the VA National Center for Patient Safety, and AHRQ (R01HS028595, R18HS029347, and R01HS029318). The authors are solely responsible for this document's contents, findings, and conclusions, which do not necessarily represent the views of AHRQ. Readers should not interpret any statement in this report as the official position or policy of the US government, AHRQ, HHS, or VA. None of the authors has any affiliation or financial involvement that conflicts with the material presented in this report.

Data Availability

The datasets generated and analyzed during this study are not publicly available because the consent form for the participants stated that the data could only be accessible to the research team.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Sample group designs by participants from remote and in-person design sessions.
[PDF File (Adobe PDF File), 400 KB - formative_v9i1e68891_app1.pdf]

References

1. Asplin BR, Magid DJ, Rhodes KV, Solberg LI, Lurie N, Camargo CA. A conceptual model of emergency department crowding. Ann Emerg Med 2003 Aug;42(2):173-180. [doi: 10.1067/mem.2003.302] [Medline: 12883504]
2. Beck E, Balasubramanian H, Henneman P. Resource management and process change in a simplified model of the emergency department. In: Proceedings of the 2009 Conference on Winter Simulation Conference. 2009 Presented at: WSC '09; December 13-16, 2009; Austin, TX p. 1887-1895 URL: https://ieeexplore.ieee.org/document/5429198 [doi: 10.1109/wsc.2009.5429198]
3. Cosby KS. A framework for classifying factors that contribute to error in the emergency department. Ann Emerg Med 2003 Dec;42(6):815-823. [doi: 10.1016/s0196-0644(03)00635-8] [Medline: 14634609]
4. Carter AJ, Chochinov AH. A systematic review of the impact of nurse practitioners on cost, quality of care, satisfaction and wait times in the emergency department. CJEM 2007 Jul 21;9(4):286-295. [doi: 10.1017/s1481803500015189] [Medline: 17626694]
5. Singer AJ, Thode Jr HC, Viccellio P, Pines JM. The association between length of emergency department boarding and mortality. Acad Emerg Med 2011 Dec;18(12):1324-1329 [FREE Full text] [doi: 10.1111/j.1553-2712.2011.01236.x] [Medline: 22168198]
6. Sun BC, Hsia RY, Weiss RE, Zingmond D, Liang L, Han W, et al. Effect of emergency department crowding on outcomes of admitted patients. Ann Emerg Med 2013 Jun;61(6):605-11.e6 [FREE Full text] [doi: 10.1016/j.annemergmed.2012.10.026] [Medline: 23218508]
7. Ball JE, Murrells T, Rafferty AM, Morrow E, Griffiths P. 'Care left undone' during nursing shifts: associations with workload and perceived quality of care. BMJ Qual Saf 2014 Feb 29;23(2):116-125 [FREE Full text] [doi: 10.1136/bmjqs-2012-001767] [Medline: 23898215]
8. Croskerry P, Sinclair D. Emergency medicine: a practice prone to error? CJEM 2001 Oct;3(4):271-276. [doi: 10.1017/s1481803500005765] [Medline: 17610769]
9. Werner NE, Holden RJ. Interruptions in the wild: development of a sociotechnical systems model of interruptions in the emergency department through a systematic review. Appl Ergon 2015 Nov;51:244-254. [doi: 10.1016/j.apergo.2015.05.010] [Medline: 26154223]
10. Berg LM, Källberg AS, Göransson KE, Östergren J, Florin J, Ehrenberg A. Interruptions in emergency department work: an observational and interview study. BMJ Qual Saf 2013 Aug;22(8):656-663. [doi: 10.1136/bmjqs-2013-001967] [Medline: 23584208]
11. Chisholm CD, Collison EK, Nelson DR, Cordell WH. Emergency department workplace interruptions: are emergency physicians "interrupt-driven" and "multitasking"? Acad Emerg Med 2000 Nov 28;7(11):1239-1243. [doi: 10.1111/j.1553-2712.2000.tb00469.x] [Medline: 11073472]
12. Pun JK, Matthiessen CM, Murray KA, Slade D. Factors affecting communication in emergency departments: doctors and nurses' perceptions of communication in a trilingual ED in Hong Kong. Int J Emerg Med 2015 Dec 15;8(1):48 [FREE Full text] [doi: 10.1186/s12245-015-0095-y] [Medline: 26667242]

XSL•FO
RenderX

13. Risser DT, Rice MM, Salisbury ML, Simon R, Jay GD, Berns SD. The potential for improved teamwork to reduce medical errors in the emergency department. The MedTeams Research Consortium. Ann Emerg Med 1999 Sep;34(3):373-383. [doi: 10.1016/s0196-0644(99)70134-4] [Medline: 10459096]
14. Yoon J, Sonneveld M. Anxiety of patients in the waiting room of the emergency department. In: Proceedings of the 4th international conference on Tangible, embedded, and embodied interaction. 2010 Presented at: TEI '10; January 24-27, 2010; Cambridge, MA p. 279-286 URL: https://dl.acm.org/doi/10.1145/1709886.1709946 [doi: 10.1145/1709886.1709946]
15. Feied CF, Handler JA, Smith MS, Gillam M, Kanhouwa M, Rothenhaus T, et al. Clinical information systems: instant ubiquitous clinical data for error reduction and improved clinical outcomes. Acad Emerg Med 2004 Nov;11(11):1162-1169 [FREE Full text] [doi: 10.1197/j.aem.2004.08.010] [Medline: 15528580]
16. Carter AJ, Davis KA, Evans LV, Cone DC. Information loss in emergency medical services handover of trauma patients. Prehosp Emerg Care 2009 Aug 13;13(3):280-285. [doi: 10.1080/10903120802706260] [Medline: 19499462]
17. Laxmisan A, Hakimzada F, Sayan OR, Green RA, Zhang J, Patel VL. The multitasking clinician: decision-making and cognitive demand during and after team handoffs in emergency care. Int J Med Inform 2007 Nov;76(11-12):801-811. [doi: 10.1016/j.ijmedinf.2006.09.019] [Medline: 17059892]
18. Stiell A, Forster AJ, Stiell IG, van Walraven C. Prevalence of information gaps in the emergency department and the effect on patient outcomes. CMAJ 2003 Nov 11;169(10):1023-1028 [FREE Full text] [Medline: 14609971]
19. Cordell WH, Overhage JM, Waeckerle JF. Strategies for improving information management in emergency medicine to meet clinical, research, and administrative needs. Information Management Work Group. Acad Emerg Med 1998 Feb 28;5(2):162-167 [FREE Full text] [doi: 10.1111/j.1553-2712.1998.tb02604.x] [Medline: 9492140]
20. Okafor N, Payne VL, Chathampally Y, Miller S, Doshi P, Singh H. Using voluntary reports from physicians to learn from diagnostic errors in emergency medicine. Emerg Med J 2016 Apr 03;33(4):245-252. [doi: 10.1136/emermed-2014-204604] [Medline: 26531860]
21. Kachalia A, Gandhi TK, Puopolo AL, Yoon C, Thomas EJ, Griffey R, et al. Missed and delayed diagnoses in the emergency department: a study of closed malpractice claims from 4 liability insurers. Ann Emerg Med 2007 Feb;49(2):196-205. [doi: 10.1016/j.annemergmed.2006.06.035] [Medline: 16997424]
22. Studdert DM, Mello MM, Gawande AA, Gandhi TK, Kachalia A, Yoon C, et al. Claims, errors, and compensation payments in medical malpractice litigation. N Engl J Med 2006 May 11;354(19):2024-2033. [doi: 10.1056/NEJMsa054479] [Medline: 16687715]
23. Guly HR. Diagnostic errors in an accident and emergency department. Emerg Med J 2001 Jul;18(4):263-269 [FREE Full text] [doi: 10.1136/emj.18.4.263] [Medline: 11435359]
24. Horng S, Sontag DA, Halpern Y, Jernite Y, Shapiro NI, Nathanson LA. Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. PLoS One 2017;12(4):e0174708 [FREE Full text] [doi: 10.1371/journal.pone.0174708] [Medline: 28384212]
25. Xie F, Liu N, Wu SX, Ang Y, Low LL, Ho AF, et al. Novel model for predicting inpatient mortality after emergency admission to hospital in Singapore: retrospective observational study. BMJ Open 2019 Sep 26;9(9):e031382 [FREE Full text] [doi: 10.1136/bmjopen-2019-031382] [Medline: 31558458]
26. Chiew CJ, Liu N, Tagami T, Wong T, Koh ZX, Ong ME. Heart rate variability based machine learning models for risk prediction of suspected sepsis patients in the emergency department. Medicine (Baltimore) 2019 Feb;98(6):e14197 [FREE Full text] [doi: 10.1097/MD.0000000000014197] [Medline: 30732136]
27. Smith SW, Walsh B, Grauer K, Wang K, Rapin J, Li J, et al. A deep neural network learning algorithm outperforms a conventional algorithm for emergency department electrocardiogram interpretation. J Electrocardiol 2019 Jan;52:88-95. [doi: 10.1016/j.jelectrocard.2018.11.013] [Medline: 30476648]
28. Sandhu S, Lin AL, Brajer N, Sperling J, Ratliff W, Bedoya AD, et al. Integrating a machine learning system into clinical workflows: qualitative study. J Med Internet Res 2020 Nov 19;22(11):e22421 [FREE Full text] [doi: 10.2196/22421] [Medline: 33211015]
29. Tsai WC, Liu CF, Lin HJ, Hsu C, Ma Y, Chen C, et al. Design and implementation of a comprehensive AI dashboard for real-time prediction of adverse prognosis of ED patients. Healthcare (Basel) 2022 Aug 09;10(8):1498 [FREE Full text] [doi: 10.3390/healthcare10081498] [Medline: 36011155]
30. Levin S, Toerper M, Hamrock E, Hinson JS, Barnes S, Gardner H, et al. Machine-learning-based electronic triage more accurately differentiates patients with respect to clinical outcomes compared with the emergency severity index. Ann Emerg Med 2018 May;71(5):565-74.e2. [doi: 10.1016/J.ANNEMERGMED.2017.08.005]
31. Moustapha Mbaye M, Tahiri N. A new machine learning workflow to create an optimal waiting list in hospitals. In: Proceedings of the 7th International Conference on Medical and Health Informatics. 2023 Presented at: ICMHI '23; May 12-14, 2023; Kyoto, Japan p. 159-163 URL: https://dl.acm.org/doi/10.1145/3608298.3608328 [doi: 10.1145/3608298.3608328]
32. Yu JY, Jeong GY, Jeong OS, Chang DK, Cha WC. Machine learning and initial nursing assessment-based triage system for emergency department. Healthc Inform Res 2020 Jan;26(1):13-19 [FREE Full text] [doi: 10.4258/hir.2020.26.1.13] [Medline: 32082696]

XSL•FO
RenderX

33. Farahmand S, Shabestari O, Pakrah M, Hossein-Nejad H, Arbab M, Bagheri-Hariri S. Artificial intelligence-based triage for patients with acute abdominal pain in emergency department; a diagnostic accuracy study. Adv J Emerg Med 2017;1(1):e5 [FREE Full text] [doi: 10.22114/AJEM.v1i1.11] [Medline: 31172057]
34. Jiang H, Mao H, Lu H, Lin P, Garry W, Lu H, et al. Machine learning-based models to support decision-making in emergency department triage for patients with suspected cardiovascular disease. Int J Med Inform 2021 Jan;145:104326. [doi: 10.1016/j.ijmedinf.2020.104326] [Medline: 33197878]
35. Liu N, Xie F, Siddiqui FJ, Ho AF, Chakraborty B, Nadarajan GD, et al. Leveraging large-scale electronic health records and interpretable machine learning for clinical decision making at the emergency department: protocol for system development and validation. JMIR Res Protoc 2022 Mar 25;11(3):e34201 [FREE Full text] [doi: 10.2196/34201] [Medline: 35333179]
36. Scheder-Bieschin J, Blümke B, de Buijzer E, Cotte F, Echterdiek F, Nacsa J, et al. Improving emergency department patient-physician conversation through an artificial intelligence symptom-taking tool: mixed methods pilot observational study. JMIR Form Res 2022 Feb 07;6(2):e28199 [FREE Full text] [doi: 10.2196/28199] [Medline: 35129452]
37. Georgiou A, Prgomet M, Paoloni R, Creswick N, Hordern A, Walter S, et al. The effect of computerized provider order entry systems on clinical care and work processes in emergency departments: a systematic review of the quantitative literature. Ann Emerg Med 2013 Jun;61(6):644-53.e16. [doi: 10.1016/j.annemergmed.2013.01.028] [Medline: 23548404]
38. Parker CA, Liu N, Wu SX, Shen Y, Lam SS, Ong ME. Predicting hospital admission at the emergency department triage: a novel prediction model. Am J Emerg Med 2019 Aug;37(8):1498-1504. [doi: 10.1016/j.ajem.2018.10.060] [Medline: 30413365]
39. Sun Y, Heng BH, Tay SY, Seow E. Predicting hospital admissions at emergency department triage using routine administrative data. Acad Emerg Med 2011 Aug;18(8):844-850 [FREE Full text] [doi: 10.1111/j.1553-2712.2011.01125.x] [Medline: 21843220]
40. Raita Y, Goto T, Faridi MK, Brown DF, Camargo CA, Hasegawa K. Emergency department triage prediction of clinical outcomes using machine learning models. Crit Care 2019 Feb 22;23(1):64 [FREE Full text] [doi: 10.1186/s13054-019-2351-7] [Medline: 30795786]
41. Fenn A, Davis C, Buckland DM, Kapadia N, Nichols M, Gao M, et al. Development and validation of machine learning models to predict admission from emergency department to inpatient and intensive care units. Ann Emerg Med 2021 Aug;78(2):290-302. [doi: 10.1016/j.annemergmed.2021.02.029] [Medline: 33972128]
42. Austin EE, Blakely B, Salmon P, Braithwaite J, Clay-Williams R. Technology in the emergency department: using cognitive work analysis to model and design sustainable systems. Saf Sci 2022 Mar;147:105613. [doi: 10.1016/j.ssci.2021.105613]
43. Drescher FS, Chandrika S, Weir ID, Weintraub JT, Berman L, Lee R, et al. Effectiveness and acceptability of a computerized decision support system using modified Wells criteria for evaluation of suspected pulmonary embolism. Ann Emerg Med 2011 Jun;57(6):613-621. [doi: 10.1016/j.annemergmed.2010.09.018] [Medline: 21050624]
44. Ash JS, Berg M, Coiera E. Some unintended consequences of information technology in health care: the nature of patient care information system-related errors. J Am Med Inform Assoc 2004;11(2):104-112 [FREE Full text] [doi: 10.1197/jamia.M1471] [Medline: 14633936]
45. Muller MJ, Kuhn S. Participatory design. Commun ACM 1993 Jun;36(6):24-28. [doi: 10.1145/153571.255960]
46. Simonsen J, Robertson T. Routledge International Handbook of Participatory Design. New York, NY: Routledge; 2012.
47. Khairat S, Marc D, Crosby W, Al Sanousi A. Reasons for physicians not adopting clinical decision support systems: critical analysis. JMIR Med Inform 2018 Apr 18;6(2):e24 [FREE Full text] [doi: 10.2196/medinform.8912] [Medline: 29669706]
48. Kusunoki D, Sarcevic A, Zhang Z, Yala M. Sketching awareness: a participatory study to elicit designs for supporting emergency medical teamwork. Comput Support Coop Work 2015 Feb 01;24(1):1-38 [FREE Full text] [doi: 10.1007/s10606-014-9210-5] [Medline: 25870498]
49. Østervang C. Acute care planning in emergency departments: a participatory design study exploring how to improve emergency department visits through a health information system. Syddansk Universitet. 2023. URL: https://portal.findresearcher.sdu.dk/en/publications/acute-care-planning-in-emergency-departments-a-participatory-desi [accessed 2024-04-29]
50. Seo W, Park SY, Zhang Z, Singh H, Pasupathy K, Mahajan P. Identifying interventions to improve diagnostic safety in emergency departments: protocol for a participatory design study. JMIR Res Protoc 2024 Jun 21;13:e55357 [FREE Full text] [doi: 10.2196/55357] [Medline: 38904990]
51. Mattar e Silva TW, Velloso IS, Araújo MT, Fernandes AD. Configuration of power relations in physicians and nurses' professional practices. Rev Bras Enferm 2020;73 Suppl 1(suppl 1):e20180629 [FREE Full text] [doi: 10.1590/0034-7167-2018-0629] [Medline: 32428173]
52. Gleddie M, Stahlke S, Paul P. Nurses' perceptions of the dynamics and impacts of teamwork with physicians in labour and delivery. J Interprof Care 2018 Dec 30:1-11. [doi: 10.1080/13561820.2018.1562422] [Medline: 30596305]
53. Scupin R. The KJ method: a technique for analyzing data derived from Japanese ethnology. Hum Organ 2008 Jan 22;56(2):233-237. [doi: 10.17730/humo.56.2.x335923511444655]
54. Mastrianni A, Kulp L, Mapelli E, Sarcevic A. Understanding digital checklist use through team communication. Ext Abstr Hum Factors Computing Syst 2020 Apr;2020:1-8 [FREE Full text] [doi: 10.1145/3334480.3382817] [Medline: 32747878]

XSL•FO
RenderX

55. Redfern E, Brown R, Vincent CA. Improving communication in the emergency department. Emerg Med J 2009 Sep 21;26(9):658-661. [doi: 10.1136/emj.2008.065623] [Medline: 19700583]
56. Okpala P. Addressing power dynamics in interprofessional health care teams. Int J Healthc Manag 2020 May 07;14(4):1326-1332. [doi: 10.1080/20479700.2020.1758894]
57. Greenfield G, Pliskin JS, Feder-Bubis P, Wientroub S, Davidovitch N. Patient-physician relationships in second opinion encounters - the physicians' perspective. Soc Sci Med 2012 Oct;75(7):1202-1212. [doi: 10.1016/j.socscimed.2012.05.026] [Medline: 22749657]
58. Aaronson EL, White BA, Black L, Brown DF, Benzer T, Castagna A, et al. Training to improve communication quality: an efficient interdisciplinary experience for emergency department clinicians. Am J Med Qual 2019 Sep 21;34(3):260-265. [doi: 10.1177/1062860618799936] [Medline: 30235933]
59. Park SY, Chen Y, Raj S. Beyond health literacy: supporting patient-provider communication during an emergency visit. In: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing. 2017 Presented at: CSCW '17; February 25-March 1, 2017; Portland, OR p. 2179-2192 URL: https://dl.acm.org/doi/10.1145/2998181.2998357 [doi: 10.1145/2998181.2998357]
60. Feblowitz J, Takhar SS, Ward MJ, Ribeira R, Landman AB. A custom-developed emergency department provider electronic documentation system reduces operational efficiency. Ann Emerg Med 2017 Nov;70(5):674-82.e1 [FREE Full text] [doi: 10.1016/j.annemergmed.2017.05.032] [Medline: 28712608]
61. Armenia S, Thangamathesvaran L, Caine AD, King N, Kunac A, Merchant AM. The role of high-fidelity team-based simulation in acute care settings: a systematic review. Surg J (N Y) 2018 Jul 13;4(3):e136-e151 [FREE Full text] [doi: 10.1055/s-0038-1667315] [Medline: 30109273]
62. Ten Eyck RP. Simulation in emergency medicine training. Pediatr Emerg Care 2011;27(4):333-341. [doi: 10.1097/pec.0b013e3182131fe0]
63. Sweeney LA, Warren O, Gardner L, Rojek A, Lindquist DG. A simulation-based training program improves emergency department staff communication. Am J Med Qual 2014 Jul 23;29(2):115-123. [doi: 10.1177/1062860613491308] [Medline: 23880775]
64. Patterson MD, Geis GL, Falcone RA, LeMaster T, Wears RL. In situ simulation: detection of safety threats and teamwork training in a high risk emergency department. BMJ Qual Saf 2013 Jun 20;22(6):468-477. [doi: 10.1136/bmjqs-2012-000942] [Medline: 23258390]
65. Acosta JN, Falcone GJ, Rajpurkar P. The need for medical artificial intelligence that incorporates prior images. Radiology 2022 Aug;304(2):283-288. [doi: 10.1148/radiol.212830] [Medline: 35438563]

Abbreviations

AI: artificial intelligence
CDSS: clinical decision support system
ED: emergency department
EHR: electronic health record
ESI: Emergency Severity Index
PD: participatory design

Edited by A Mavragani; submitted 24.11.24; peer-reviewed by E Bai, H Pilabré; comments to author 05.01.25; revised version received 25.02.25; accepted 11.03.25; published 21.04.25.

Please cite as:
Seo W, Li J, Zhang Z, Zheng C, Singh H, Pasupathy K, Mahajan P, Park SY
Designing Health Care Provider–Centered Emergency Department Interventions: Participatory Design Study
JMIR Form Res 2025;9:e68891
URL: https://formative.jmir.org/2025/1/e68891
doi:10.2196/68891
PMID:

©Woosuk Seo, Jiaqi Li, Zhan Zhang, Chuxuan Zheng, Hardeep Singh, Kalyan Pasupathy, Prashant Mahajan, Sun Young Park. Originally published in JMIR Formative Research (https://formative.jmir.org), 21.04.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Compatibility of the CEN-ISO/TS 82304-2 Health App Assessment Framework With Catalan and Italian Health Authorities' Needs: Qualitative Interview Study

Petra Hoogendoorn[1*], MSc; Mariam Shokralla[1], MPH, MSc; Romy Willemsen[1], MSc; Nick Guldemond[1,2], MD, PhD; María Villalobos-Quesada[1*], PhD

[1]National eHealth Living Lab, Public Health and Primary Care Department, Leiden University Medical Center, Leiden, The Netherlands
[2]Research Center for Medical Sociology, Tsinghua University, Beijing, China
[*]these authors contributed equally

Corresponding Author:
Petra Hoogendoorn, MSc
National eHealth Living Lab
Public Health and Primary Care Department
Leiden University Medical Center
Albinusdreef 2
Leiden, 2333 ZA
The Netherlands
Phone: 31 654341785
Email: a.p.y.hoogendoorn@lumc.nl

Abstract

Background: Health authorities of European Union (EU) member states are increasingly working to integrate quality health apps into their health care systems. Given the current lack of unified EU assessment criteria, the European Commission initiated Technical Specification (TS) CEN-ISO 82304-2:2021—Health and wellness apps—Quality and reliability (hereinafter the "TS") to address the scattered EU landscape of assessment frameworks (AFs) for health apps. The adoption of an AF, such as the TS, falls within member state competence and is considered an uncertainty-reduction process. Evaluations by peers as well as ensuring the compatibility of the TS with the needs of health authorities can reduce uncertainty and mediate harmonization.

Objective: This study aims to examine the compatibility of the TS with the needs of Catalan and Italian health authorities.

Methods: Semistructured interviews were conducted with key informants from a regional (Catalonia in Spain) and national (Italy) health authority, and a thematic analysis was carried out. Main themes were established deductively, following the aspects defined by the value proposition canvas: (1) health authorities' needs ("gains," "pains," and "jobs") and (2) the TS "products and services" and their distinct characteristics ("gain creators" and "pain relievers"). Subthemes were generated inductively. The compatibility of the needs with the TS was theoretically determined by the researchers. The results were visualized using the value proposition canvas. Two participant validation steps confirmed that the most relevant aspects of the predefined themes had been captured.

Results: Despite the diversity of the 2 health authorities, subthemes were common and categorized into 9 gains, 9 pains, and 11 jobs. Key findings include the health authorities' perceived value of, and need for, integrating quality health apps and using an AF (gains), along with the related policy, implementation, and operational activities (jobs). The lack of enabling EU legislation and standardization, resulting in a need for the multiple authorities involved to consent, made achieving an AF challenging (pains). Nine products and services related to the TS and 17 distinct characteristics (eg, its multistakeholder evidence base) were found to be compatible with 3 gains (eg, stimulating the prescription and use of apps), 7 pains (eg, legislation and harmonization issues), and 6 jobs (eg, assessing apps). Indirect effects, 3 anticipated future services, and 1 anticipated gain creator and pain reliever increase this compatibility.

Conclusions: Our results suggest that the health authorities share common fundamental needs, and that the TS is compatible with these needs. The identified needs and compatibility can potentially reduce peer authorities' uncertainties in adopting an AF in general and the TS in particular. More research is recommended to confirm and translate our results in other contexts and further fine-tune compatibility to achieve wide adoption of the TS and accelerate the uptake of health apps.

XSL•FO
RenderX

(JMIR Form Res 2025;9:e67855) doi:10.2196/67855

KEYWORDS

assessment frameworks; mobile health; mHealth; health apps; wellness apps; digital transformation; Italy; Catalonia; diffusion of innovations; value proposition canvas

Introduction

Background

Health authorities of European Union (EU) member states are transforming health and care systems to address current challenges and remove cross-border regulatory barriers for businesses and consumers to progress toward an EU digital single market [1-6]. In this context, health apps (Textbox 1) are gaining attention, and countries are adapting their policies and structures to harvest the potential of these digital solutions to strengthen their health care systems [5,7-10].

Textbox 1. Definitions of health, medical, and wellness apps.

In the context of this paper, health apps are defined as apps that are "intended to be used specifically for managing, maintaining or improving health of individual persons or the delivery of care" [11]. Health apps are part of mobile health, which is defined as "medical and public health practice supported by mobile devices, such as mobile phones, patient monitoring devices, personal digital assistants (PDAs), and other wireless devices" [12]. Health apps include medical apps and wellness apps. Medical apps have been defined as those that fall under an applicable medical device regulation and wellness apps as those that do not [13].

Not all apps are of good quality, that is, have a positive and reliable effect on health, and are easy to use, compliant with privacy and data security regulations and standards, robust, and interoperable with electronic health record systems [8,9,14]. Determining the quality of individual health apps is challenging for citizens, health care professionals (HCPs), and decision makers [15] for various reasons. First, assessing apps requires specific and diverse expertise. Second, robust evidence (eg, clinical evidence) and background information are often scarce, inappropriate, or not publicly available [14]. Third, quality assessments and widely adopted evidence-based evaluation methodologies that consider all these aspects and inform (potential) users about the quality of these apps are not yet common [10].

Several health authorities have developed their own assessment frameworks (AFs) [16]. However, their efficient implementation is challenging [7,15]. The significant overlap in quality criteria across these AFs highlights the potential for harmonization and related efficiency [17]. In addition, nonoverlapping quality criteria—a lack of agreement on what constitutes quality—potentially decrease trust [17] and increase inequality (eg, if the challenges to comply with the different AFs result in manufacturers focusing only on the larger markets and widely spoken languages). In the context of multilingualism—an EU founding principle [18]—and an abundance of health apps worldwide [19], the availability of native language health apps for the 5 million Finnish-speaking EU residents, for instance, is limited [20]. These aspects highlight the need for the harmonization of AFs at an EU level.

Both academics and policy makers endorse harmonization and cross-national regulation of health apps to realize their full potential and benefits [7,10]. The European Commission initiated CEN-ISO/TS (Comité Européen de Normalisation [European Committee for Standardization]–International Organization for Standardization/Technical Specification) 82304-2:2021 Health software—Part 2: Health and wellness apps—Quality and reliability (hereinafter the "TS"; Textbox 2) to address the scattered EU landscape of health app AFs and progress from 27 national and even more regional markets to a digital single market [6,11]. To achieve EU-wide harmonization, the adoption of the TS among EU member states is needed, which falls within member state competence and responsibility. Adoption across member states can be described as the diffusion of the TS as an innovation.

Textbox 2. Brief description of CEN-ISO/TS (Comité Européen de Normalisation [European Committee for Standardization]–International Organization for Standardization/Technical Specification) 82304-2:2021 Health software—Part 2: Health and wellness apps—Quality and reliability (the TS).

CEN-ISO/TS 82304-2:2021 was an assignment from the European Commission to CEN that became a global effort in its collaboration with ISO and the International Electrotechnical Commission (IEC). The TS has 2 core products. The first is a health app assessment framework (AF) developed through a Delphi study with 83 experts [21]. The second is a quality label (Multimedia Appendix 1) that was inspired by the European Union (EU) energy label and Nutri-Score front-of-pack nutrition label designs and the US Food and Drug Administration over-the-counter medicine label content. The quality label allows the easy and accessible presentation of the assessment results of a health app and was developed and tested with people with low health literacy. The EU-wide implementation readiness of the TS was supported by the Horizon Europe Label2Enable project (2022-2024), which tested the quality label and tested and developed products and services complementary to the AF and quality label [22,23]. These include the health app quality report, which is the detailed version of the label that aims to support health care professionals in their decision-making on recommending health apps; the certification scheme, which includes the handbook for app assessment with CEN-ISO/TS 82304-2 for certified app assessment organizations; stakeholder guidance and tools such as educational videos; and recommendations for the reimbursement of health apps. A multistakeholder road map was generated to guide next steps [24].

Innovations and their adoption process can be understood using the theory of diffusion of innovations proposed by Rogers [25], who identified 3 main attributes of an innovation that predict its future rate of adoption. Compatibility is one of these attributes and refers to the degree to which an innovation is perceived as consistent with the existing values, experiences, and needs (summarized hereinafter as "needs") of potential adopters. The diffusion of an innovation is considered an

XSL•FO
RenderX

uncertainty-reduction process. An innovation that is compatible generates less uncertainty for the potential adopter regarding the success of its deployment. The other main attributes that predict the future rate of adoption—relative advantage and complexity—are beyond the scope of this paper and will be analyzed in separate studies.

Objectives

Potential adopters tend to rely on peers for information about an innovation's desirable, direct, and anticipated consequences to decrease their uncertainty about adopting it [25]. In this paper, we aim to examine the compatibility of the TS with the needs of 2 health authorities in the EU, in order to potentially enhance the compatibility of the TS and reduce the uncertainty of peer health authorities in their considerations on adopting the TS.

Methods

Overview

Rogers [25] described different methods to research the attributes of innovations, including compatibility. For this study, investigating an innovation's acceptability in its prediffusion stage (eg, test-marketing and piloting) was considered the most appropriate approach. This approach helps identify a "basis for positioning an innovation so that it will be more acceptable, that is, have a more rapid rate of adoption" [25]. Key informant interviews are a common method for assessing the acceptability of an intervention [26,27]. Key informants provide high-level perspectives and comparative insights. Their particular roles and expertise make them especially equipped for probing about how a topic is thought about or acted upon in policy. This contrasts with the scope of insight contributed by other groups of qualitative interviewees, who are more often recruited to provide data rooted in their own lived experiences, opinions, and beliefs [27]. A widely used business tool to visualize and enhance the compatibility (also referred to as "fit") of innovations is the value proposition canvas (VPC; Figure 1; Textbox 3) [28,29]. The VPC tool has been previously applied to guide the development of, among others, digital health innovations [30-32].

Figure 1. The template of the value proposition canvas [29]. Adapted from Strategyzer [33] to highlight the 3 main steps in constructing the value proposition canvas (Textbox 3): (A) the customer profile with the customer needs (jobs, gains, and pains), (B) the value map with the innovations' products and services and their distinct characteristics (gain creators and pain relievers) that aim to address the customer needs, and (C) the fit (compatibility) of the customer profile with the value map.

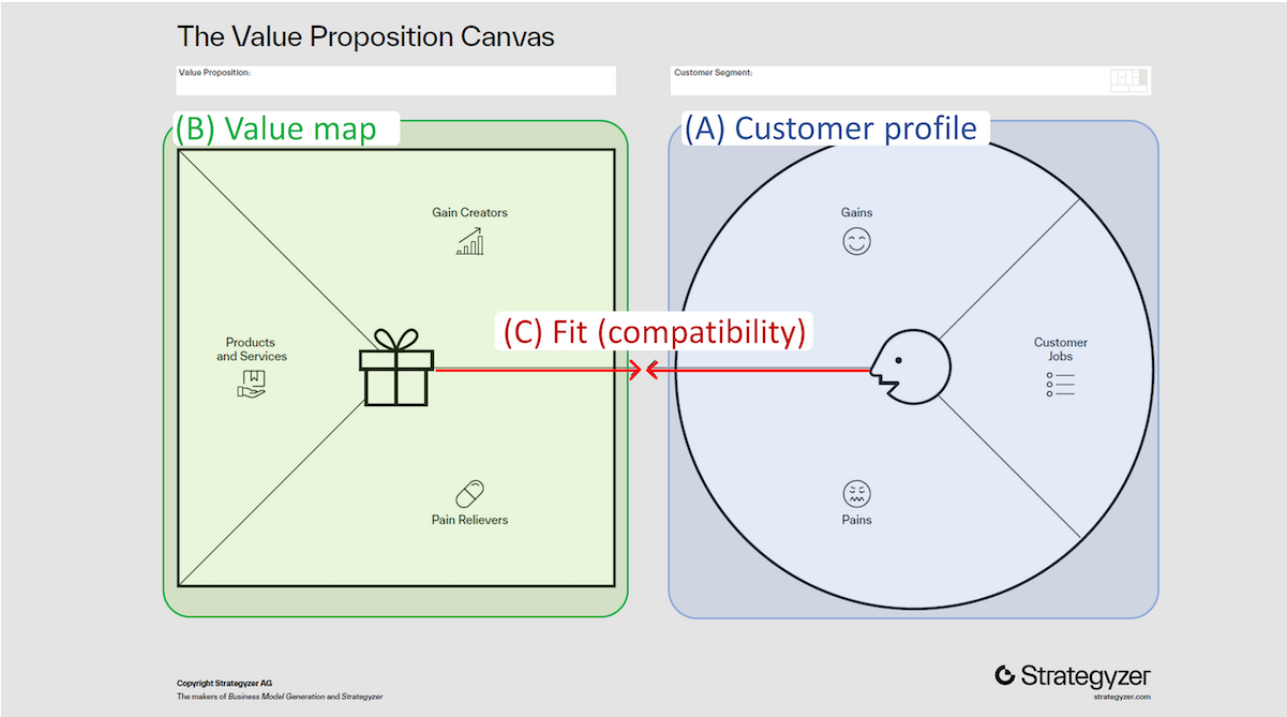

XSL•FO
RenderX

Textbox 3. Glossary of value proposition canvas concepts.

- The customer profile (Figure 1A) describes a potential adopter segment, in this case health authorities, in terms of their needs:
  - Jobs—the tasks or responsibilities potential adopters have, need, or want to get done in their work and their lives, in this case with respect to adopting an assessment framework and integrating health apps in their health care system
  - Gains—outcomes and benefits potential adopters require, expect, or desire to achieve if these jobs are done (well)
  - Pains—negative outcomes, risks, and obstacles that adopters encounter, foresee, or fear that prevent them from getting these jobs done (well)
- The value map (Figure 1B) makes explicit how the innovation creates value for the potential adopters' segment:
  - Products and services—the range of products and services available to the potential adopters to address their needs, in this case with respect to adopting an assessment framework and integrating health apps
  - Gain creators—the characteristics of these products and services that generate gains for the adopters
  - Pain relievers—the characteristics of these products and services that alleviate the adopters' pains
- Fit (Figure 1C): when the products and services, gain creators and pain relievers contribute to realizing the jobs and gains, and addressing the pains of the potential adopters ("compatibility")

In this paper, we used the VPC to visualize the compatibility of the TS with the needs of 2 health authorities (potential adopters of the TS):

1. A regional authority—Fundació TIC Salut Social (FTSS; TIC Salut Social Foundation), the public organization that oversees digital transformation in health care in Catalonia, Spain
2. A national authority—Italy's Istituto Superiore di Sanità (ISS; National Institute of Health)

Catalonia established its own AF in 2017 [34-36] and is pursuing the adoption of the TS [37]. Italy drafted a Technical Report about wellness, social, and health apps, which informed the TS [38], and has been using the TS as inspiration in the ongoing process to arrive at quality assessment criteria for the inclusion of telehealth solutions in a national telehealth catalog, where telehealth includes health apps. FTSS and ISS jointly represent the health authority perspective in the Label2Enable project (Textbox 2) and were recruited for this study. Recruitment criteria spanned experience as an EU-based health authority with the relatively novel integration of health apps, prediffusion interest in a standardized AF, and diversity (country or region, number of inhabitants, and an AF in place or in development). These were considered essential for providing the rich basis for analyzing the TS in terms of prediffusion acceptability or compatibility [25].

Key Informants

Three key informants were recruited. They were chosen strategically, in alignment with existing knowledge of the qualities of "good" key informants [27]. Eligibility criteria included a key leading position in their respective organizations over the course of several years in our topic of interest and related inside in-depth knowledge about the authorities' policy and plans to assess health apps as well as execution of these plans. For FTSS, the key informant has been responsible for implementing Catalonia's mobile health (mHealth) plan in FTSS for the past 9 years. For ISS, the first key informant had recently retired from a leadership position at ISS, which involved advising the government about health apps and telemonitoring. The second ISS key informant was recruited after discussion with the first ISS key informant, based on complementary first-hand experience on the topic in dealings with the Italian Ministry of Health. The 2 ISS informants played a direct role in drafting the Technical Report, which informed the TS [38]. While the number of potential participants can be very large in other types of qualitative research, key informant groups can be quite small, hard to reach, or constrained in sharing information with researchers [27].

Key informants and study researchers collaborated in the Label2Enable project; however, the key informants were not otherwise involved in these research activities. The researchers who conducted the interviews and analysis were experienced in qualitative, mixed methods, and implementation research in the field of digital health. Interest in the research topic stems from their previous work related to health apps, other digital health technologies, digital transformation, and the TS.

Constructing the VPC

Constructing the VPC involves 3 steps (Figure 1; Textbox 3) [29], which we carried out systematically using qualitative research methods. The first step was to develop the "customer profile" (Figure 1A). Here, the needs of the health authorities (potential adopters of the TS) were described in terms of the 2 authorities' "gains," "jobs," and "pains" (Textbox 3). Strategyzer's "trigger questions" served as a guide for designing the semistructured interviews (Multimedia Appendix 2) [33].

Two semistructured interviews (one for Catalonia and one for Italy) were conducted, in English, with the relevant key informants. The semistructured interview questions were shared with the key informants in advance. The interviews were conducted on the web, recorded using a secure digital platform, and lasted between 1.5 and 2 hours. Recordings were transcribed verbatim and coded independently by 2 researchers using ATLAS.ti software (version 7.9; ATLAS.ti Scientific Software Development GmbH). A deductive and inductive thematic analysis of the interviews was performed. On the basis of the VPC, we predefined 3 main themes: gains, pains, and jobs [29]. We classified a job as a past, current, or future task or

XSL•FO
RenderX

responsibility of a health authority related to adopting an AF to integrate health apps into their health care system. Jobs could be the responsibility of the health authority featured in this study or associated health authorities because, from an international perspective, health authorities' task divisions may differ and evolve. We applied the same principle to the gains and pains. Subthemes were inductively developed by 2 researchers and visualized using the VPC.

Data saturation was not deemed appropriate for this study; instead, inductive thematic saturation [39] was sought across and within cases to achieve a "basis for positioning the innovation [the TS] so that it will be more acceptable," knowing that uncertainty is inherent in the prediffusion phase, and that adopting organizations will adapt an innovation to achieve a fit of the innovation with the adopting organization and its perceived problem [25]. Accounting for data sufficiency is an underdeveloped aspect of existing methodological guidance on key informant sampling [27]. Incorporating strategies to return the findings and elicit key informant response on researcher interpretations is recommended to yield rich insights [27]. With this in mind, a participant validation step that lasted between 30 and 60 minutes was carried out separately per health authority. Key informants were supplied with the written data analysis before and after the validation step, allowing them to provide comments and enrich the analysis. The key informants confirmed that the researchers had captured the most relevant aspects within the 3 predefined main themes: gains, pains, and jobs [28].

The second step was to develop the "value map" (Figure 1B), which consists of describing the innovation's "products and services," "pain relievers," and "gain creators" (Textbox 3). Outputs from all Label2Enable work packages were considered [22,23]. Products and services included those that are currently part of the TS, those that were cocreated in the Label2Enable project (both Textbox 2), and products and services that will be generated beyond the project during the "demonstrator phase." The relevant characteristics of these products and services were subsequently added to the value map as gain creators and pain relievers.

The third step was to determine the fit (Figure 1C), that is, to check whether the "value map" was compatible with the "customer profile." In other words, whether the TS and the related products and services and their distinct characteristics (gain creators and pain relievers) satisfied the needs (jobs, gains, and pains) of the health authorities. Visualizing compatibility (fit) was an iterative process of determining and conceptualizing the products and services, gain creators, and pain relievers that related to the health authorities' gains, pains, and jobs. After completing this process, a second participant validation step, which lasted approximately 1 hour, was carried out with the key informants from both health authorities simultaneously. Again, the key informants were supplied with the written data analysis before and after the validation step, allowing them to provide comments and enrich the analysis. Furthermore, the manuscript was discussed with the key informants at different stages, and they confirmed that it condensed the most relevant aspects surrounding the position of their institutions and agreed to its publication in its current form.

Data were reported according to the COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist (Multimedia Appendix 3).

Ethical Considerations

Approval from an ethics or scientific research committee was not required under Dutch national regulations because the study did not fall under the Medical Research Involving Human Subjects Act (WMO), and according to the guidelines of the Central Committee on Research Involving Human Subjects (CCMO [40]). This study followed institutional good research practices and integrity codes. Verbal consent to conduct and record the interviews was obtained from all participants (key informants) before starting the interview. Personal data were protected in accordance with European Union and institutional standards and best practices. The 2 organizations represented were funded as Label2Enable project partners and were not additionally compensated for their participation in this study.

Results

Customer Profile

In this subsection, we describe the needs of health authorities (potential adopters) in terms of gains, pains, and jobs. Despite the diversity of the 2 health authorities, subthemes were common and categorized into 9 gains, 9 pains, and 11 jobs (Figure 2).

XSL•FO
RenderX

Figure 2. Adapted from Strategyzer [33], the value proposition canvas of CEN-ISO/TS (Comité Européen de Normalisation [European Committee for Standardization]–International Organization for Standardization/Technical Specification) 82304-2:2021 Health software—Part 2: Health and wellness apps—Quality and reliability (the TS) for health authorities pursuing an assessment framework (AF) to integrate health apps into their health care systems. The "customer profile" (Figure 1A) summarizing the Catalan and Italian health authorities' needs (gains, pains, and jobs) resulting from the deductive and inductive thematic analysis. The "value map" (Figure 1B) of the TS, with the products and services related to the TS and their distinct characteristics (gain creators and pain relievers) that aim to address these health authorities' needs as identified by the researchers. Lighter colored boxes with italic text indicate future services, gain creators, and pain relievers. Dark green circles with a check mark indicate the fit (compatibility; Figure 1C) of a health authority's need with an existing product or service, gain creator, or pain reliever. Light green circles without a check mark and italic text indicate expected future fit (compatibility) derived from the nature of the anticipated related services and gain creator or pain reliever. The tables in this paper show the specific products, services, or characteristics that fit specific customer needs (gain, pain, or job). AF: assessment framework; EU: European Union; HCP: health care professional; HTA: health technology assessment; IEC: International Electrotechnical Commission.

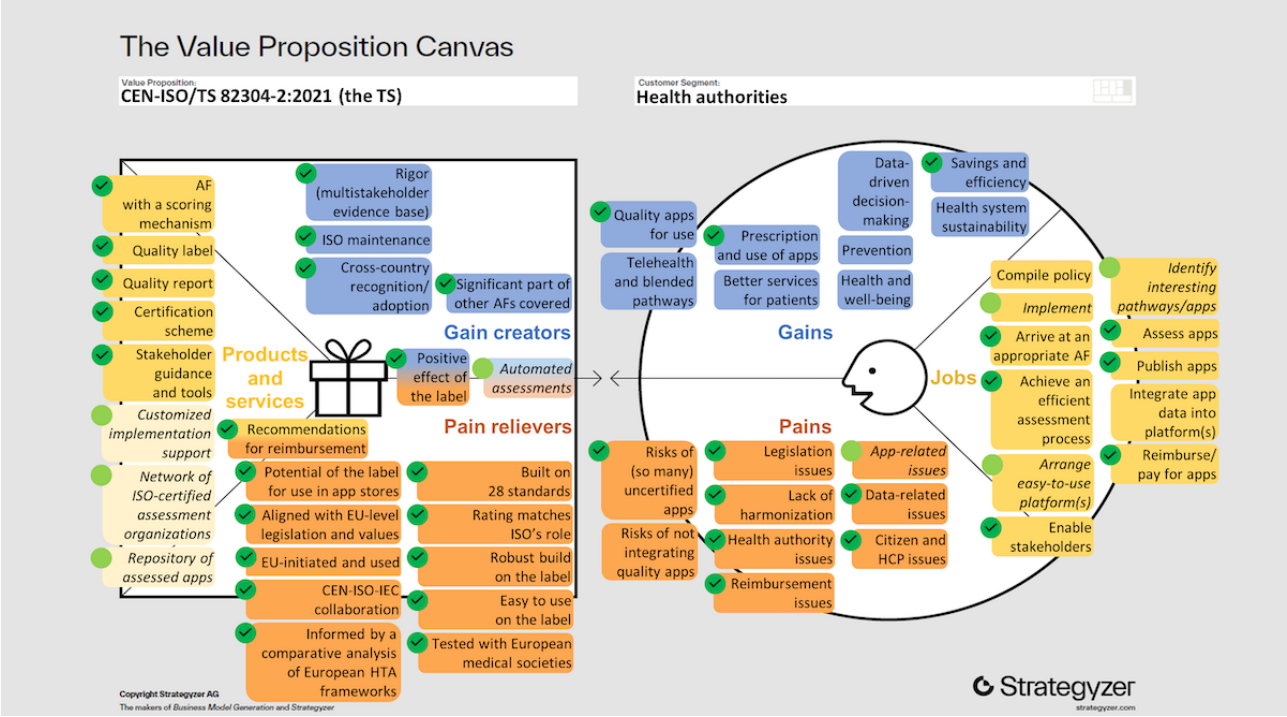

Gains

Regarding the gains—the outcomes and benefits that the health authorities require, expect, or desire to achieve from adopting an AF and integrating quality health apps—the key informants explained the interest of their region (Catalonia, Spain) or country (Italy) in an AF to identify "quality apps for use" in "telehealth (Italy) and blended pathways (Catalonia)" (Textbox 4).

Textbox 4. Key definitions in the Catalonian and Italian context.

Blended care in Catalonia

- "Blended care" "refers to an integration of online and offline components in a treatment process. This means that online and offline components are interconnected in some way and not stand-alone treatment pathways [41]. In Catalonia, according to the mobile health plan, mobile health should have an impact on the full "cycle of provision of care" (prevention, diagnostics, treatment, and follow-up) and ultimately on health, the health care system, and public health. In addition, it prioritizes the 10 most prevalent diseases (eg, diabetes) [42].

Telehealth in Italy

- The World Health Organization defines telehealth as the "delivery of health care services, where patients and providers are separated by distance" [12]. In the Italian context, telehealth or telemedicine is an innovative approach to health care practice that enables the remote provision of services through the use of digital devices, the internet, software, and telecommunication networks. Telehealth services can serve as an alternative to traditional care, support existing services (eg, by improving accessibility, efficiency, and equity), supplement care (eg, by improving effectiveness and facilitating personalized medicine), or completely replace traditional services [43].

The key informants described how the assessment of apps is instrumental to achieving the "prescription and use of apps," referring to incorporating the prescription or recommendation of apps in the provision of care and the related trusted use of health apps by citizens. Similarly, it can contribute to "better services for patients" (eg, improved accessibility, lower waiting times, and better follow-up):

> *If you want these applications to be recommended by health care professionals, we need to provide safe and guaranteed apps to the sick. So, we try to establish how we can measure or evaluate these apps and the quality of these apps.* [Key informant, Catalonia]

XSL•FO
RenderX

Certainly, the availability of an agreed, evidence-based assessment framework will give objectivity to the activity of app assessment. [Key informant, Italy]

They [patients] will have the opportunity to make more virtual visits, sharing the data with their professionals without needing to go to the consult, provide more alerts and provide more predictions, and prevention to avoid disease or avoid being sicker. Also, they will feel more confident and committed to healthier lifestyle habits when using wellness apps. And they can also be more guarded by their professionals. They will feel more followed up. [Key informant, Catalonia]

Moreover, such apps can also streamline the access of the citizens to the NHS, for example, by cutting waiting times for a given service. [Key informant, Italy]

For the key informants, the prescription and use of apps would enable "data-driven decision-making"—the use of health data derived from an individual's health app use in the delivery of health care—seeing both benefits for the prescribing HCP and the provision of effective transversal care. Other benefits mentioned included the possibilities for "prevention," through the promotion of healthier behaviors and end-user empowerment, and ultimately "health and well-being":

To provide more autonomy to patients and to be more committed to their health, to keep tracking health data more efficiently and having all the information integrated into the public single network, so that all health care professionals are able to follow up the patient through the information system. [Key informant, Catalonia]

The interest of our country is given by the opportunity of providing the best health status of the population, even with limited resources. Health apps can support primary and secondary prevention, by modifying lifestyles. [Key informant, Italy]

The key informants described how all aforementioned gains translate into "savings and efficiency" and "health system sustainability," mentioning aspects such as fewer in-person visits, more web-based visits, more equity, earlier diagnosis, and cross-country recognition of app assessments:

I think for health care professionals, they [apps] will reduce the number of face-to-face consults for more virtual [visits]. They will have a unique platform or a unique space to monitor the data of the patients...they receive alerts, for example, so that they focus their attention on critical issues, not revising a lot of data. [Key informant, Catalonia]

My personal feeling for sure, savings at the health system level. Access for all the people, or the individuals, as my Constitution says. [Key informant, Italy]

Another step that has to be done is updating the current certification framework and consider the possibility to adapt this framework to the Technical Specification of 82304-2 in order to improve the framework for cross-recognition. [Key informant, Catalonia]

A sound evaluation framework, recognized by the EU, will be instrumental in optimizing the resources spent in assessing health apps in the different EU countries by means of the foreseeable cross-country validity of the approval. [Key informant, Italy]

Pains

Regarding the pains—the negative outcomes, risks, and obstacles that health authorities face with respect to assessing and integrating apps—the key informants from both authorities mentioned challenges related to context, implementation, and app use. The "risks of (so many) uncertified apps" referred to the currently unknown quality of apps and the negative effects for patients, HCPs, and the health care system. At the same time, the key informants considered the "risks of not integrating quality apps," emphasizing that not integrating apps would be an obstacle to providing better care and other important gains:

That is the main risk, not being adopted because professionals don't feel that it helps. Because I think that the patient part is solved. They are using it a lot and they want to be involved. [Key informant, Catalonia]

We've in Italy started a long time ago working on apps because we have realized that digital support to health care is absolutely mandatory, and it was clear that there was a sort of Wild Wild West with respect to the apps. [Key informant, Italy]

The main risk, I think, will be not adopting mobile apps into the health care system, not providing sufficiently optimized care to the citizens compared to other territories. [Key informant, Catalonia]

At present you lose data, you lose information, you lose the capability to cost, to build a new best practice because you have no sufficient critical mass of data to say something, also to use artificial intelligence appropriately. [Key informant, Italy]

Implementation-related pains mentioned by the key informants included "legislation issues," such as the lack of comprehensive legislation for apps that are not medical devices and the multiple applicable (emerging) EU legislations, which pose significant challenges to app developers and authorities. "Lack of harmonization" is a missed opportunity to alleviate the burden of legislation issues and potentially adds to the difficulties:

We are regulating medical devices, the apps that are also medical devices, thanks to the new regulation of medical devices, but we cannot regulate the apps that are not medical devices. That is why we are interested in integrating a sort of certification scheme similar to, or directly, the ISO (82304-2) Technical Specification. [Key informant, Italy]

I am afraid of too much legislation. Too much vertical [legislation] that is not correlated to each other. [Key informant, Italy]

XSL•FO
RenderX

> *We faced the challenge to create the framework from scratch because we did not have anything at the time.* [Key informant, Catalonia]

> *The risk in the immediate future is that some regional health system starts piloting an evaluation framework without national coordination, in that case...it would be more difficult to harmonize the assessment at national level, which could result in enhanced health inequalities.* [Key informant, Italy]

For the key informants, legislation issues and the lack of harmonization resulted in "health authority issues." The integration of health apps into health care currently involves multiple authorities, whose interests, roles, and responsibilities related to health apps can be unclear, overlapping or conflicting. In addition, health apps present an uncharted territory and, may represent just one of these authorities' many topics. This makes reaching agreements on issues such as an AF complex. In addition, publishing information about the quality of products without being able to refer to a recognized (international) standard could conflict with the authority's role and be perceived as biased from the perspective of health app providers. Finally, the key informants refer to "reimbursement issues," challenges related to structuring the reimbursement of health apps:

> *Something that is really difficult, and I think the European region has that problem, is the different actors to be aligned...Pff, a lot of actors, that each of them has their needs and it's very difficult to get an agreement to move forward.* [Key informant, Catalonia]

> *We have too many agencies, too many institutions that manage all these definitions of digital health and digital health technologies [not aligned taxonomies]. That is why I hope that main documents from Europe could help us to be stricter in the field.* [Key informant, Italy]

> *Communicating the quality of a particular product, could actually lead to a "market bias," that is, putting a product in a favorable position with respect to its competitors, although involuntarily, needless to say. We are instead interested in communicating the relevance of the framework for the app evaluation, as an objective and independent instrument for orienting health choices for personalized digital medicine.* [Key informant, Italy]

> *And after that comes defining the reimbursement models, well, I think it will take a lot [of time].* [Key informant, Catalonia]

The key informants identified several pains related to the identification, assessment, integration, recommendation, and use of health apps. These included "app-related issues" (eg, constant updates and a lack of clinical evidence), "data-related issues" (eg, not all data generated by apps are clinically relevant, with relevance generally differing per pathology and care pathway; moreover, there can be interoperability issues), and "citizen and HCP issues" (eg, limited digital skills):

> *Something that is very difficult is how you revalidate based on the framework, how do you reassess the apps. Because the apps change a lot.* [Key informant, Catalonia]

> *"At present what is lacking is clear sound evidence from clinical trials about the apps. This is a very great problem because also our Ministry of Health is waiting for clinical trial data for authorization of medical device apps and then to have the data from them.* [Key informant, Italy]

> *Then you [HCP] want to monitor a patient based on these health apps, and you want to integrate the data...It was very challenging because we had to find specific professionals for each pathology. They tell us or tell the group which are the relevant data.* [Key informant, Catalonia]

> *Improving the digital skills of professionals also is very challenging and we try to solve it as well, by providing more training.* [Key informant, Catalonia]

> *The target population of some interventions — for example, for frailty/chronicity — may find itself hardly capable of exploiting the technology, if the latter is not properly introduced, with a training appropriate for the digital skills of the subject.* [Key informant, Italy]

Jobs

Regarding the jobs—the tasks or responsibilities of health authorities in assessing and integrating apps—the key informants described activities related to policy, implementation, and operations. In Catalonia, "compile policy" referred mainly to the "mHealth plan" [42], which in 2015 triggered the appointment of FTSS to "implement" this plan and create an AF. In Italy, this job was linked to the recovery and resilience funding coordinated by the European Commission, which pushed plans for telehealth and an adequate AF [44]. The key informants discussed their roles as experts and advisers to the government and in the decision-making and implementation of an AF:

> *TIC Salut mHealth Office was established after publishing this agreement in 2015, an mHealth plan that was published or was agreed upon in the parliament, where they created TIC Salut in the mHealth area as an instrument to provide or to create these actions, to implement mHealth in the health sector.* [Key informant, Catalonia]

> *The National Resilience Program has given us quite a lot of support in that, and a lot of money has actually been poured into health care...One of the areas which has been impacted very much is telehealth, which includes health apps.* [Key informant, Italy]

Although the jobs overlap, the exact scope targeted in their policies and strategies differs. Catalonia emphasizes apps embedded in blended care, whereas Italy focuses on telehealth solutions:

> *The electronic health record of the patient was integrated, and we also had the personal portal for citizens. So, the gap that we did have was to integrate*

XSL•FO
RenderX

> *the data that came from apps.* [Key informant, Catalonia]
>
> *And so, the focus here in Italy at the moment is on telehealth solutions, which comprise health apps.* [Key informant, Italy]

Important steps in the implementation were, according to the key informants, to "arrive at an appropriate AF"; "achieve an efficient assessment process" that is moreover scalable; "arrange easy-to-use platform(s)"; and "enable stakeholders" by, for example, issuing recommendations, accessible information and tools (eg, self-assessment tools [45]):

> *So, those were the 2 big projects. One is the assessment framework and the other the mHealth platform that goes in parallel...The objective of this platform is to include apps that have passed this accreditation and to help health care professionals prescribe the apps to their patients.* [Key informant, Catalonia]
>
> *So, the ministry studied all the proposals from the different experts, and they actually decided to use 82304-2 as the inspirational assessment framework to define the assessment framework for telehealth solutions that are going to be onboarded onto this national catalog.* [Key informant, Italy]
>
> *The main objective is now the number of apps that we want to share. It's our expectation to have at least 30 apps, or 50 depending. But this year at least try to get these 30.* [Key informant, Catalonia]
>
> *And also, we did recommendations, guides, tools. Also, we have the auto-evaluation on our website, that is a questionnaire free and open...We made it for the industry, for them being prepared, and to promote creating these quality apps.* [Key informant, Catalonia]

Operations start with what key informants refer to as a prioritization of the apps to integrate into the health care system ("identify interesting pathways/apps"):

> *So, we focus on diabetes because it's something that we have different apps that are provided by private companies, and there are a lot of people with this disease, and it is expensive and it's a priority for the health care system to provide tools in that way.* [Key informant, Catalonia]
>
> *So, there will be a pathway for diabetics, so there would be one for oncology, certain types of oncology and so on.* [Key informant, Italy]
>
> *We have these 2000 apps that are identified as interesting apps, that are not certified using our framework obviously, but they are being used by health care centers or they are being published in other frameworks like Andalusia or mHealth Belgium, et cetera.* [Key informant, Catalonia]

After this prioritization of apps that are interesting to assess, the assessment can be carried out ("assess apps"), and positively assessed apps and their results in specific quality domains and criteria can be shared ("publish apps"). The publication of assessed apps enables the series of gains, which starts with "quality apps for use." The next operational step, to "integrate app data into platform(s)," enables the desired gain "data-driven decision-making." Catalonia expects the capability to integrate app data into platforms to become available in 2025 for 1 care pathway, while Italy does not yet carry out assessments, and the division of responsibilities among Italian authorities is not yet defined.

For both health authorities featured in this study, the final step in the process is to "reimburse/pay for apps":

> *Once these apps are validated, we publish the apps on our website, so all citizens know all aspects of the [certification] process.* [Key informant, Catalonia]
>
> *And in the next stage, we are expecting to integrate the data from these apps of diabetes in the platform that we create.* [Key informant, Catalonia]
>
> *We have a group that is working on a specific reimbursement model.* [Key informant, Catalonia]
>
> *A policy is at present under discussion thanks to a specific government workgroup on digital therapies that involves several stakeholders, scientific and industrial. The latter (drug and medical device industry organizations) are discussing a unitary proposal for a unique reimbursement scheme.* [Key informant, Italy]

Value Map and Fit

Overview

In this subsection, we present the results of the iterative process of identifying the TS-related products and services and their distinct characteristics in relation to the needs of the health authorities. The needs are visualized in the VPC as the pains, gains, and jobs of the 2 health authorities. When a TS-related product or service or one of its characteristics (gain creator and pain reliever) helps to address or to achieve a gain, pain, or job of the health authorities, we considered it a "fit" (compatibility). The fit was visualized with a green circle (Figure 2). We found that the TS addressed 3 (33%) of 9 gains, 7 (78%) of 9 pains, and 6 (55%) of 11 jobs in part or in full. An anticipated gain creator and pain reliever and 3 TS-related anticipated future services would target 1 extra pain and 3 extra jobs, enhancing the compatibility to 3 (33%) of 9 gains, 8 (89%) of 9 pains, and 9 (82%) of 11 jobs. The 6 gains and 1 pain that remain would be influenced indirectly (eg, the TS supports the gain "quality apps for use," which is a prerequisite for all further gains). This would leave only 2 jobs unaddressed: "compile policy" and "integrate app data into platform(s)."

Products and Services

In total, 6 current and 3 future products and services were identified, which are compatible with ("fit") 9 jobs (Figure 2; Table 1). The 2 core products of the TS are an "AF with a scoring mechanism" and a health app "quality label" (Multimedia Appendix 1) [11]. The quality label is a key aspect that differentiates the TS from other AFs. In the Label2Enable project, four TS-related products were cocreated with the relevant stakeholders: (1) the health app "quality report," which

XSL•FO
RenderX

is a more detailed version of the quality label [22]; (2) the ISO 17000 series–compliant "certification scheme" for the TS, which includes the app assessment handbook for CEN-ISO/TS 82304-2 for certified conformity assessment bodies [22]; (3) "stakeholder guidance and tools," which refers, for example, to educational videos [46] on the quality label for citizens and guidance for manufacturers; and (4) a series of "recommendations for reimbursement" of health apps [47]. The upcoming demonstrator phase aims to assess the first 100 apps, test and optimize the value of the health app quality report from a multistakeholder perspective and explore how to exploit the potential of "automated assessments" (refer to the next subsection, Gain Creators). This phase is planned to deliver 3 extra services: (1) "customized implementation support," (2) a "network of ISO-certified assessment organizations," and (3) a "repository of assessed apps."

Table 1. Compatibility (or fit) of the jobs related to adopting an assessment framework (AF) for health apps as described by the Catalan and Italian health authorities and the products and services related to CEN-ISO/TS (Comité Européen de Normalisation [European Committee for Standardization]–International Organization for Standardization/Technical Specification) 82304-2:2021 Health software—Part 2: Health and wellness apps—Quality and reliability (the TS).

Jobs	Products and services
Implement[a]	• Customized implementation support[a]
Arrive at an appropriate AF	• AF with a scoring mechanism
Achieve an efficient assessment process	• Certification scheme • Network of ISO-certified assessment organizations[a]
Arrange easy-to-use platform(s)[a]	• Repository of assessed apps[a]
Enable stakeholders	• Stakeholder guidance and tools • Network of ISO-certified assessment organizations[a]
Identify interesting pathways/apps[a]	• Repository of assessed apps[a]
Assess apps	• AF with a scoring mechanism • Certification scheme • Stakeholder guidance and tools • Network of ISO-certified assessment organizations[a]
Publish apps	• Quality label • Quality report • Repository of assessed apps[a]
Reimburse/pay for apps	• Quality report • Recommendations for reimbursement

[a]Job to be facilitated and corresponding anticipated TS service planned to be established during the demonstrator phase.

Gain Creators

Five current and one anticipated gain creators were identified (Figure 2), which are compatible with ("fit") 3 gains pursued by the health authorities (Table 2): (1) the "rigor (multistakeholder evidence base)" of the AF, the assessment process (promising interrater reliability), and the handbook for assessment organizations, which are in part already reported in scientific publications [21,22,48]; (2) "ISO maintenance," which refers to the regular ISO review procedures that ensure that the TS remains up to date and potentially evolves to an International Standard; (3) "cross-country recognition/adoption" can be facilitated by the TS, which in turn can increase the number of assessed apps available, optimize resource allocation (the label is effectively a screening), and minimize the duplication of efforts; (4) the "positive effect of the label" on manufacturers' intent to improve the quality of their app, HCPs' willingness to prescribe apps, and citizens' intent to download and ability to choose quality apps [22,48,49]. These effects could be amplified by citizens' trust in the recommendations of health apps by HCPs (80%) and their expressed need for health authorities to review and rate apps (86%; M Shokralla, MPH, MSc, unpublished data, 2024); (5) the preliminary findings of a comparative analysis of CEN-ISO/TS 82304-2 with European health technology assessment (HTA) frameworks and subsequent alignment [22] show that the handbook has a "significant part of other AFs covered." For health authorities, this could mean a reduction in the conformity assessment workload because already-labeled apps would only need to be assessed against a limited set of context-specific requirements; and (6) "automated assessments," which refers to the potential of partially automatizing app assessments, benefiting the quality (rigor), efficiency, affordability, and scalability of assessments and reassessments. Examples include software that can test the app's accessibility (eg, contrast, addressing color blindness, and readability), privacy compliance (eg, personal data processed

XSL•FO
RenderX

and whether the privacy statement includes all mandatory elements [50]), and data security mechanisms (eg, automated security testing) and in general natural language processing to help assess the evidence provided and queries for surveillance purposes.

Table 2. Compatibility (or fit) of the gains of adopting an assessment framework (AF) for health apps as perceived by the Catalan and Italian health authorities and the distinct characteristics (gain creators) of the products and services related to CEN-ISO/TS (Comité Européen de Normalisation [European Committee for Standardization]–International Organization for Standardization/Technical Specification) 82304-2:2021 Health software—Part 2: Health and wellness apps—Quality and reliability (the TS).

Gains	Gain creators
Quality apps for use	• Rigor (multistakeholder evidence base) • ISO maintenance • Cross-country recognition/adoption
Prescription and use of apps	• Positive effect of the label
Savings and efficiency	• Cross-country recognition/adoption • Significant part of other AFs covered • Automated assessments[a]

[a]Future gain creator planned to be established during the demonstrator phase.

Pain Relievers

Twelve current pain relievers and one anticipated pain reliever were identified (Figure 2), which are compatible with ("fit") 8 pains (Table 3): (1) the "potential of the label for use in app stores," the common marketplace for health apps. The TS has global applicability (ISO), a label, a scoring mechanism, and scalability potential with the "automated assessments" and "network of ISO-certified assessment organizations." These distinct characteristics are considered attractive for app stores to rank apps and inform potential users about the quality of (labeled) apps; (2) the "positive effect of the label" on manufacturers' quality improvement plans and HCPs' willingness to prescribe apps [22,48,49]; (3) the handbook for app assessment organizations was "aligned with EU-level legislation and values" [22], and the AF was built on the EU Medical Device Regulation and General Data Protection Regulation principles [21]; (4) the TS and Label2Enable project are "EU-initiated and used." A draft version of the TS was made available to support the creation of COVID-19 apps and referenced in the EU toolbox for COVID-19 contact–tracing apps [51], and the TS is foundational for drafting implementing legislation for labeling wellness applications (that claim interoperability with an electronic health record system) in the European Health Data Space Regulation; (5) the TS is a "CEN-ISO-IEC collaboration", all 3 bodies are renowned international standardization organizations; (6) the handbook was "informed by a comparative analysis of European HTA frameworks," that is, the European Network for Health Technology Assessment core model and the Dutch, English, Finnish, French, and German HTA frameworks for health apps [22]; (7) distinctive from other AFs, the AF was "built on 28 standards" [11,21]. This includes the recognized National Institute for Health and Care Excellence evidence standards framework for digital health technologies [52,53], which was used as a foundation for the CEN ISO/TS 82304-2 AF [21]; (8) the "rating matches ISO's role" to "agree on the best way of doing things, make lives easier, safer and better, enabling trade the world over" [54], enabling authorities to refer to unbiased and standardized assessment results; (9) the "recommendations for reimbursement," which are the result of a series of Label2Enable workshops for health authorities, HTA bodies, and insurers, with 135 participants from 34 countries [47]; (10) "automated assessments" could help address the need to frequently reassess apps; (11) the visibility of "robust build on the label," which includes 4 interoperability requirements, is expected to promote manufacturer investments in interoperability; (12) the incorporation of "easy to use on the label" could contribute to addressing inequity in the prescription and use of health apps and promote the continued use of apps [55]; and (13) given the importance of medical societies for HCPs in promoting the prescription and use of apps [56,57], the quality report was "tested with European medical societies" to evaluate and enhance its usefulness for medical societies in providing guidance for HCPs in recommending quality health apps [22]; in addition, the results of mapping the TS with the information needs of cardiologists regarding mobile health solutions, carried out with the European Society of Cardiology, indicated compatibility [58].

XSL•FO
RenderX

Table 3. Compatibility (or fit) of the pains in adopting an assessment framework (AF) for health apps as perceived by the Catalan and Italian health authorities and the distinct characteristics (pain relievers) of the products and services related to CEN-ISO/TS (Comité Européen de Normalisation [European Committee for Standardization]–International Organization for Standardization/Technical Specification) 82304-2:2021 Health software—Part 2: Health and wellness apps—Quality and reliability (the TS).

Pains	Pain relievers
Risks of (so many) uncertified apps	• Potential of the label for use in app stores • Positive effect of the label
Legislation issues	• Aligned with EU[a]-level legislation and values • EU-initiated and used
Lack of harmonization	• CEN-ISO-IEC[b] collaboration • Informed by a comparative analysis of European HTA[c] frameworks • Built on 28 standards • EU-initiated and used
Health authority issues	• Rating matches ISO's role
Reimbursement issues	• Recommendations for reimbursement
App-related issues[d]	• Automated assessments[d]
Data-related issues	• Robust build on the label
Citizen and HCP issues	• Positive effect of the label • Easy to use on the label • Tested with European medical societies

[a]EU: European Union.

[b]IEC: International Electrotechnical Commission.

[c]HTA: health technology assessment.

[d]Pain to be addressed and corresponding anticipated pain reliever planned to be established during the demonstrator phase.

Discussion

Principal Findings

In this paper, we systematically examined the compatibility of the TS with the needs of 2 health authorities, aiming to further enhance the compatibility of the TS with these needs and, ultimately, reduce the uncertainty of peer authorities in considering the adoption of the TS. The 2 studied health authorities were diverse. FTSS is located in Catalonia, a region in Spain with 8 million citizens, and has an AF for health apps and wearables and an operational assessment process in place. ISS is based in Italy, a country with 59 million citizens, and is working on an AF for telehealth, which includes health apps. We found that despite their diversity, their needs (gains, pains, and jobs) largely overlapped. This suggests that health authorities share common fundamental needs. Differences in needs could be attributed to being a national advisory body in a country currently without an established AF (a focus on legislation issues—Italy) and a regional implementing organization with an AF (a focus on the execution of policy—Catalonia). Both health authorities see a need for, and the benefits of, the uptake of health apps and using a common AF. This confirms the recommendation of the World Health Organization (WHO) to make mHealth evaluation with a common methodology the norm rather than the exception [10]. At the same time, it is apparent that without enabling (EU) legislation and standardization and with multiple authorities involved, it is a challenge to establish an AF. For countries or regions with a small population, arriving at an appropriate (and sufficiently rigorous) AF might not even be feasible, given the costs, capabilities, and limited attractiveness for manufacturers.

When is the compatibility of the TS *sufficient* for health authorities? Osterwalder et al [29] argue that an innovation does not have to address all needs of potential adopters; yet it should satisfy the most essential (unrealized) gains, most extreme (unresolved) pains, and most important (unsatisfied) jobs and preferably be "difficult to copy." We found that the TS has a range of products and services with distinct characteristics (pain relievers and gain creators) that address in part or in full all but 2 needs of the 2 health authorities studied; these are the jobs "compile policy" and "integrate app data into platform(s)" (Tables 1-3). To what extent these 2 jobs would classify as *important (unsatisfied) jobs*—and as such should be addressed in the future by the TS—needs to be determined in future research. Many of the distinct characteristics (gain creators and pain relievers) are *difficult to copy* (eg, multistakeholder evidence base, international standardization nature, the comprehensiveness of the label, and its development as an EU initiative that has already been applied at the EU level).

Whether all of the distinguished products, services, and their distinct characteristics individually (are perceived to) *sufficiently satisfy* potential adopters' needs could not be measured

XSL•FO
RenderX

conclusively, given the prediffusion stage of the TS at the time of this study. Such measurements and potential further fine-tuning of the compatibility of the TS, which could further reduce uncertainty, are part of the upcoming plans for the demonstrator phase. The aims for this phase include assessing the first 100 apps, developing the aforementioned additional 3 services and 1 gain creator and pain reliever, and testing the health app quality report. We also aim to explore whether a further reduction in the duplication of efforts would increase compatibility and reduce uncertainty. Could, for instance, the need to "arrange easy-to-use platform(s)" for assessed apps be addressed with an EU-level "repository of assessed apps" and EU efforts to materialize the "potential of the label for use in app stores"? Could such strategies increase the rate of adoption of the TS, prevent EU member states (and potentially medical societies) from having to invest individually in such platforms and thus pave the way for the EU-wide adoption of the TS and uptake of health apps in the region? WHO statistics seem to confirm the potential of an EU-level platform, with 77% (39/51) of its European region member states rating the lack of a trustworthy source to access effective apps as a significant barrier to integrating health apps into clinical practice, outnumbering privacy and security concerns (38/51, 75%), patient digital literacy (37/51, 73%), and the lack of evidence on app effectiveness in clinical practice (31/51, 61%) [10].

The health authorities' jobs can be put in context with the 5 stages of the innovation-decision process in organizations described by Rogers [25]. The first stage is "agenda setting," when one or more individuals in an organization identify a problem and seek a compatible innovation to solve it. During this stage, the job "compile policy" needs to be carried out. WHO statistics show that 83% (44/53) of its European region member states, including 85% (23/27) of the EU member states, reported having a national digital health policy or strategy. Nevertheless, only 28% (13/47) reported having an entity for mHealth quality oversight, which relates to the second job, "implement" [59].

After agenda setting, the next stages of the innovation-decision process in organizations described by Rogers [25] are "matching" (testing the feasibility of the innovation in solving the organization's problem), and "redefining/restructuring" (reinventing the innovation to accommodate the organization's needs and structure). These are related to the identified jobs "arrive at an appropriate AF" and "achieve an efficient assessment process." Catalonia has already carried out these jobs and published the methodology used to compare the TS to its previous AF and implement the TS, reporting a mere 7% additional context-specific, scope-expanding, and rigor-enhancing requirements [37], similar to reports in the Australian context [60] and as previously suggested [17]. This knowledge could further reduce peer authorities' uncertainty and inspire their own "matching" and "redefining/restructuring" efforts. The rest of the reported jobs, including "assess apps," would follow once the first four have been achieved. Going back to the status of the European region reveals that while 83% (44/53) of WHO European region member states reported having a national digital health policy or strategy, only 15% (6/39) reported the evaluation of government-sponsored mHealth ("assess apps") [10], which is a prerequisite for the ultimate jobs in our list: "integrate app data into platform(s)" and "reimburse/pay for apps." Although financing schemes for telemedicine are increasingly regularly available [10], the number of reimbursed apps were found to be limited [7] and related value-based pricing frameworks "nonexistent to embryonic" [61].

The low percentages of WHO European region member states having entities for mHealth quality oversight and government-sponsored mHealth evaluation raise questions. Do other member states encounter the same significant challenges (pains) as Catalonia and Italy in satisfactorily arriving at an appropriate AF and assessing health apps? Could that be the reason for not assessing apps? Or could it be that these member states have different needs, perhaps not even perceive, contrary to Catalonia and Italy (and the WHO), that an AF is a much-needed solution? Or do they perhaps perceive other *most essential (unrealized) gains*, *most extreme (unresolved) pains*, and *most important (unsatisfied) jobs*, making an AF less of a priority or other distinct AF characteristics a necessity? Further key informant interviews or analysis of member states' national policies [59] could validate the generalizability of the Catalan and Italian needs or produce alternative customer profiles, perhaps related to the adopter types described by Rogers [25]: innovators, early adopters, early majority, late majority, and laggards. Such an analysis would support future alignment of the TS with the EU region at large; contribute to the development of new services, such as "customized implementation support"; and inform peer authorities' decision-making on the adoption of the TS. In other words, an analysis of member states' policies through the lens of the current VPC could potentially support the EU region in scaling the evaluation of mHealth using a common methodology as well as promote the equitable uptake of quality apps. Further "matching" efforts by, or in collaboration with, health authorities, other stakeholders, or relevant projects (eg, European Digital Health Technology Assessment [62] and ASSESS DHT [63]), using their own methodologies or inspiration from Catalonia [37], would reveal whether they similarly have few additional health app quality requirements when integrating CEN-ISO/TS 82304-2 and further evolving the Label2Enable handbook app assessment. These efforts and more multistakeholder positioning statements would further reduce uncertainties.

Strengths and Limitations

To our knowledge, this study is the first to systematically analyze the compatibility of an existing AF with health authorities' needs, which has the potential to address the scattered EU landscape, progressing toward a digital single market and improving the uptake of health apps. The timing of the study (prediffusion) exploited the potential to contribute to preadoption compatibility, the effectiveness of the future demonstrator phase, and the rate of adoption of the TS. Our study was based on the perceptions of 3 key informants associated with 2 health authorities in 2 EU countries, each part of a larger landscape of health authorities, each with a focus on health apps as part of a larger scope of digital health technologies (Catalonia: health apps and wearables and Italy:

XSL•FO
RenderX

telehealth). The limited number of key informants can be attributed to the prediffusion phase of the TS, that is, the TS was published in 2021, and from 2022 to 2024, the Label2Enable project has cocreated the TS supportive products and services with stakeholders. The 2 health authorities involved—partners in the Label2Enable project—are the first prerelease test users in the EU. The key informants are among the first individuals to evaluate the TS. Such a limited scope requires future validation, and we specifically pledge to extend this analysis to the national policies of the EU member states and further "matching" efforts. The complex process of multistakeholder adoption in the EU and the wider landscape depends on more than compatibility for 2 health authorities in the prediffusion phase. As Catalonia's FTSS revealed in its paper, a multistakeholder evaluation is recommended as part of adoption decision-making [37]. In other publications, currently or shortly available (all Label2Enable results will be published in Cordis [22] and on the Label2Enable results page [23]), we have addressed other attributes and other key stakeholders, some of whom have produced endorsing positioning statements [58,64]. Theories focusing on the wider ecosystem and AF landscape could be considered to further support adoption strategies. Despite the limitations of this study, the diversity of the 2 health authorities in terms of geographic scope, roles and responsibilities, population size, triggers, the scope of digital health technologies, target pathways and health apps, and the status of the AF, increases the chance of the generalizability of our results.

Conclusions

Our results suggest compatibility of the TS with the overlapping needs of the health authorities to arrive at an appropriate AF that would allow the uptake of health apps in their health care systems. The perceptions and experiences of health authorities captured in this study through interviews with key informants provide an evidence-based foundation for peer authorities to reduce their uncertainties related to the adoption of an AF, particularly the TS. The study also established the basis to carry out a wider analysis to understand the compatibility of the TS with the needs of other EU member states, which would confirm or fine-tune the TS and its distinct characteristics. To our knowledge, this is the first report to systematically analyze the compatibility of an existing AF with health authorities' needs, with the potential to address the scattered EU AF landscape, progressing toward a digital single market and improving the uptake of health apps.

Acknowledgments

The authors would like to express their deepest gratitude to Carme Pratdepadua Bufill, Mauro Grigioni, Pier Angelo Sottile, and Giuseppe d'Avenio, who generously shared their time, insights, and expertise. Their invaluable contributions and perspectives have been instrumental in shaping the findings of this study. The views and opinions expressed by the individuals who participated in this study are their own and do not necessarily reflect the official policy or position of the institutions they are affiliated with.

This study was funded by the European Union (101057522—Label2Enable—HORIZON-HLTH-2021-IND-07). The views and opinions expressed in this paper are those of the authors and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency; neither the European Union nor the granting authority can be held responsible for them.

Authors' Contributions

PH and MS were responsible for funding acquisition. PH was responsible for supervision. MS, RW, and PH designed the semistructured interviews. MVQ led the interviews with support from MS. MVQ transcribed the recordings verbatim, and PH proofread the transcripts. PH and MVQ performed coding using both deductive and inductive approaches. PH visualized the results using the value proposition canvas with feedback from MVQ. PH and MVQ wrote the original draft. MS, RW, and NG reviewed and edited the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The CEN-ISO/TS (Comité Européen de Normalisation [European Committee for Standardization]–International Organization for Standardization/Technical Specification) 82304-2 label.
[DOCX File , 139 KB - formative_v9i1e67855_app1.docx]

Multimedia Appendix 2
Interview topic guide.
[DOCX File , 20 KB - formative_v9i1e67855_app2.docx]

Multimedia Appendix 3
COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist.
[DOCX File , 24 KB - formative_v9i1e67855_app3.docx]

XSL•FO
RenderX

References

1. Kivimäki M, Batty GD, Pentti J, Shipley MJ, Sipilä PN, Nyberg ST, et al. Association between socioeconomic status and the development of mental and physical health conditions in adulthood: a multi-cohort study. Lancet Public Health 2020 Mar;5(3):e140-e149 [FREE Full text] [doi: 10.1016/S2468-2667(19)30248-8] [Medline: 32007134]
2. Health for everyone?: social inequalities in health and health systems. Organisation for Economic Co-operation and Development. URL: https://www.oecd-ilibrary.org/social-issues-migration-health/health-for-everyone_3c8385d0-en [accessed 2025-01-12]
3. Silver opportunity: building integrated services for older adults around primary health care. The World Bank. URL: https://documents1.worldbank.org/curated/en/099025503072337503/pdf/P17583208782960a40bf6804d1758e50f1e.pdf [accessed 2025-01-12]
4. Health and care workforce in Europe: time to act? World Health Organization. URL: https://www.who.int/europe/publications/i/item/9789289058339 [accessed 2025-01-12]
5. Green Paper on mobile health ("mHealth"). European Commission. 2014. URL: https://digital-strategy.ec.europa.eu/en/library/green-paper-mobile-health-mhealth [accessed 2025-01-12]
6. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions on Enabling the Digital Transformation of Health and Care in the Digital Single Market; empowering citizens and building a healthier society. European Commission. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2018%3A233%3AFIN [accessed 2025-01-12]
7. Essén A, Stern AD, Haase CB, Car J, Greaves F, Paparova D, et al. Health app policy: international comparison of nine countries' approaches. NPJ Digit Med 2022 Mar 18;5(1):31 [FREE Full text] [doi: 10.1038/s41746-022-00573-1] [Medline: 35304561]
8. Milne-Ives M, van Velthoven MH, Meinert E. Mobile apps for real-world evidence in health care. J Am Med Inform Assoc 2020 Jun 01;27(6):976-980 [FREE Full text] [doi: 10.1093/jamia/ocaa036] [Medline: 32374376]
9. Stecher C, Pfisterer B, Harden SM, Epstein D, Hirschmann JM, Wunsch K, et al. Assessing the pragmatic nature of mobile health interventions promoting physical activity: systematic review and meta-analysis. JMIR Mhealth Uhealth 2023 May 04;11:e43162 [FREE Full text] [doi: 10.2196/43162] [Medline: 37140972]
10. The ongoing journey to commitment and transformation. World Health Organization. URL: https://www.who.int/europe/publications/digital-health-in-the-who-european-region-the-ongoing-journey-to-commitment-and-transformation [accessed 2025-01-12]
11. ISO/TS 82304-2:2021: health software: part 2: health and wellness apps — quality and reliability. International Organization for Standardization. URL: https://www.iso.org/standard/78182.html [accessed 2025-01-12]
12. mHealth: new horizons for health through mobile technologies: second global survey on eHealth. World Health Organization. 2011. URL: https://iris.who.int/handle/10665/44607 [accessed 2025-01-12]
13. Maaß L, Freye M, Pan CC, Dassow HH, Niess J, Jahnel T. The definitions of health apps and medical apps from the perspective of public health and law: qualitative analysis of an interdisciplinary literature overview. JMIR Mhealth Uhealth 2022 Oct 31;10(10):e37980 [FREE Full text] [doi: 10.2196/37980] [Medline: 36315221]
14. Grundy Q. A review of the quality and impact of mobile health apps. Annu Rev Public Health 2022 Apr 05;43:117-134 [FREE Full text] [doi: 10.1146/annurev-publhealth-052020-103738] [Medline: 34910582]
15. van Velthoven M, Powell J. Do health apps need endorsement? Challenges for giving advice about which health apps are safe and effective to use. Digit Health 2017 Apr 10;3:2055207617701342 [FREE Full text] [doi: 10.1177/2055207617701342] [Medline: 29942588]
16. D2.1 knowledge tool 1. health apps assessment frameworks. The European mHealth Hub. URL: https://cordis.europa.eu/project/id/737427/results [accessed 2025-01-12]
17. Frey AL, Phillips B, Baines R, McCabe A, Elmes E, Yeardsley-Pierce E, et al. Domain coverage and criteria overlap across digital health technology quality assessments: a systematic review. Health Technol 2024 Dec 20;15(1):15-27 [FREE Full text] [doi: 10.1007/S12553-024-00934-8]
18. Charter of fundamental rights of the European Union 2012/C 326/02. European Union. 2012. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3AC2012%2F326%2F02 [accessed 2025-01-12]
19. Digital health trends 2021: innovation, evidence, regulation, and adoption. IQVIA Institute. URL: https://www.iqvia.com/insights/the-iqvia-institute/reports-and-publications/reports/digital-health-trends-2021 [accessed 2025-01-12]
20. Voutilainen M. Report on Finnish-language mobile applications related to welfare and health in Google Play -store and in the EUDAMED -database. Finnish Coordinating Center for Health Technology Assessment. URL: https://oys.fi/fincchta/wp-content/uploads/sites/21/2024/11/fincchta_report-on-finnish-language-mobile-applications.pdf [accessed 2025-01-12]
21. Hoogendoorn P, Versluis A, van Kampen S, McCay C, Leahy M, Bijlsma M, et al. What makes a quality health app-developing a global research-based health app quality assessment framework for CEN-ISO/TS 82304-2: Delphi study. JMIR Form Res 2023 Jan 23;7:e43905 [FREE Full text] [doi: 10.2196/43905] [Medline: 36538379]
22. Adopting ISO 82304-2 and a trusted EU mHealth label for a single market that enables patients, citizens, health professionals, systems and authorities to benefit from a healthy supply of useful apps. European Commission. URL: https://cordis.europa.eu/project/id/101057522/results [accessed 2025-01-12]

XSL•FO
RenderX

23. Results. Label2Enable. URL: https://label2enable.eu/results [accessed 2025-01-12]
24. The preferred future of health & care in 5-10 years and the role of digital health quality labelling. Label2Enable. URL: https://label2enable.eu/assets/downloads/label2enable_preferred-future-health-and-care-visual-1716569529.pdf [accessed 2025-01-12]
25. Rogers EM. Diffusion of Innovations. 4th edition. New York, NY: Free Press; 2018.
26. Ayala GX, Elder JP. Qualitative methods to ensure acceptability of behavioral and social interventions to the target population. J Public Health Dent 2011;71 Suppl 1(0 1):S69-S79 [FREE Full text] [doi: 10.1111/j.1752-7325.2011.00241.x] [Medline: 21656958]
27. Pahwa M, Cavanagh A, Vanstone M. Key informants in applied qualitative health research. Qual Health Res 2023 Dec 30;33(14):1251-1261 [FREE Full text] [doi: 10.1177/10497323231198796] [Medline: 37902082]
28. Kyhnau J, Nielsen C. Value proposition design: how to create products and services customers want. J Bus Mod 2015;3(1):52. [doi: 10.5278/OJS.JBM.V3I1.1105]
29. Osterwalder A, Pigneur Y, Bernarda G, Smith A, Papadakos P. Value Proposition Design: How to Create Products and Services Customers Want. Hoboken, NJ: John Wiley & Sons; 2014.
30. Lentferink A, Polstra L, D'Souza A, Oldenhuis H, Velthuijsen H, van Gemert-Pijnen L. Creating value with eHealth: identification of the value proposition with key stakeholders for the resilience navigator app. BMC Med Inform Decis Mak 2020 Apr 27;20(1):76 [FREE Full text] [doi: 10.1186/s12911-020-1088-1] [Medline: 32340615]
31. Chakraborty I, Edirippulige S, Vigneswara Ilavarasan P. The role of telehealth startups in healthcare service delivery: a systematic review. Int J Med Inform 2023 Jun;174:105048 [FREE Full text] [doi: 10.1016/j.ijmedinf.2023.105048] [Medline: 36963322]
32. Wrede C, Braakman-Jansen A, van Gemert-Pijnen L. How to create value with unobtrusive monitoring technology in home-based dementia care: a multimethod study among key stakeholders. BMC Geriatr 2022 Nov 30;22(1):921 [FREE Full text] [doi: 10.1186/s12877-022-03550-1] [Medline: 36451119]
33. The value proposition canvas, customer gains, customer pains and customer jobs: trigger questions. Strategyzer. URL: https://www.strategyzer.com/library?type=Tools#library [accessed 2025-01-12]
34. López Seguí F, Pratdepàdua Bufill C, Rius Soler A, de San Pedro M, López Truño B, Aguiló Laine A, et al. Prescription and integration of accredited mobile apps in Catalan health and social care: protocol for the AppSalut site design. JMIR Res Protoc 2018 Dec 21;7(12):e11414 [FREE Full text] [doi: 10.2196/11414] [Medline: 30578234]
35. Lopez Segui F, Pratdepadua Bufill C, Abdon Gimenez N, Martinez Roldan J, Garcia Cuyas F. The prescription of mobile apps by primary care teams: a pilot project in Catalonia. JMIR Mhealth Uhealth 2018 Jun 21;6(6):e10701 [FREE Full text] [doi: 10.2196/10701] [Medline: 29929948]
36. Criteris de certificació d'Apps de salut o benestar Versió 3.0. Fundació TIC Salut Social. URL: https://ticsalutsocial.cat/wp-content/uploads/2021/09/2021_Criteris-certificacio_CAT_v3.0.pdf [accessed 2025-01-12]
37. Llebot Casajuana B, Hoogendoorn P, Villalobos-Quesada M, Gómez Freixa M, Pratdepadua Bufill C. Integration of CEN ISO/TS 82304-2 for existing health authorities' app assessment frameworks: a comparative case study in Catalonia. JMIRP Preprint posted online October 22, 2024 [FREE Full text] [doi: 10.2196/preprints.67858]
38. UNI/TR 11708:2018: informatica medica - caratterizzazione delle APP nel contesto della salute, benessere e stili di vita. UNI. 2018. URL: https://store.uni.com/uni-tr-11708-2018 [accessed 2025-01-12]
39. Saunders B, Sim J, Kingstone T, Baker S, Waterfield J, Bartlam B, et al. Saturation in qualitative research: exploring its conceptualization and operationalization. Qual Quant 2018;52(4):1893-1907 [FREE Full text] [doi: 10.1007/s11135-017-0574-8] [Medline: 29937585]
40. Your research: Is it subject to the WMO or not? Central Committee on Research Involving Human Subjects. URL: https://english.ccmo.nl/investigators/legal-framework-for-medical-scientific-research/your-research-is-it-subject-to-the-wmo-or-not [accessed 2025-04-16]
41. Wentzel J, van der Vaart R, Bohlmeijer ET, van Gemert-Pijnen JE. Mixing online and face-to-face therapy: how to benefit from blended care in mental health care. JMIR Ment Health 2016 Feb 09;3(1):e9 [FREE Full text] [doi: 10.2196/mental.4534] [Medline: 26860537]
42. Acord del Govern pel qual s'aprova el Pla Mestre de Mobilitat (mHealth.Cat). President of the Council of Ministers. URL: https://documents.dadesobertes.gencat.cat/acordsgovern/docs/SIG15SLT0018.pdf [accessed 2025-01-12]
43. Repertorio atto n. 215/CSR. Conferenza Permanente per i rapporti tra lo Stato, le Regioni e le Province Autonome di Trento e Bolzano. URL: https://www.statoregioni.it/it/conferenza-stato-regioni/sedute-2020/seduta-del-17122020/atti/repertorio-atto-n-215csr/ [accessed 2025-01-12]
44. Italia Domani, the National Recovery and Resilience Plan. Italian Government | Presidency of the Council of Ministers. URL: https://www.italiadomani.gov.it/content/sogei-ng/it/en/home.html [accessed 2025-01-12]
45. Actius Digitals per la Ciutadania: Self-assessment test. TIC Salut Social. URL: https://ticsalutsocial.cat/en/que-fem/digital-assets-for-citizens/test-dautoavaluacio/ [accessed 2025-01-12]
46. Label2Enable. Smeulders. URL: https://vimeo.com/showcase/11257693 [accessed 2025-01-12]
47. Recommendations for the reimbursement of health apps. Label2Enable. URL: https://label2enable.eu/assets/downloads/la2en-infographic-recommendations-v5-1716569726.pdf [accessed 2025-01-12]

XSL•FO
RenderX

48. Frey AL, Matei D, Phillips B, McCabe A, Fuller R, Laibarra B, et al. Testing and iterative improvement of the CEN ISO/TS 82304-2 health app quality assessment: pilot interrater reliability study. JMIR Form Res 2025 Mar 10;9:e64565 [FREE Full text] [doi: 10.2196/64565] [Medline: 40063936]
49. Biliunaite I, van Gestel L, Hoogendoorn P, Adriaanse M. Value of a quality label and European healthcare professionals' willingness to recommend health apps: an experimental vignette study. J Health Psychol 2024 Aug 02:13591053241258205 [FREE Full text] [doi: 10.1177/13591053241258205] [Medline: 39096027]
50. Torre D, Abualhaija S, Sabetzadeh M, Briand L, Baetens K, Goes P. An AI-assisted approach for checking the completeness of privacy policies against GDPR. In: Proceedings of the 28th International Requirements Engineering Conference. 2020 Presented at: RE '10; August 31-September 4, 2020; Zurich, Switzerland p. 136-146 URL: https://ieeexplore.ieee.org/document/9218152 [doi: 10.1109/re48521.2020.00025]
51. Mobile applications to support contact tracing in the EU's fight against COVID-19: common EU toolbox for member states. European Commission. URL: https://health.ec.europa.eu/document/download/01f83f72-4e21-4a34-90dd-ec0b0cc35c8b_en [accessed 2025-01-12]
52. Evidence standards framework for digital health technologies. National Institute for Health and Care Excellence. URL: https://www.nice.org.uk/Media/Default/About/what-we-do/our-programmes/evidence-standards-framework/digital-evidence-standards-framework.pdf [accessed 2025-01-12]
53. Dittrich F, Mielitz A, Pustozerov E, Lawin D, von Jan U, Albrecht U. Digital health applications from a government-regulated directory of reimbursable health apps in Germany-a systematic review for evidence and bias. Mhealth 2023 Oct;9:35 [FREE Full text] [doi: 10.21037/mhealth-23-17] [Medline: 38023782]
54. About ISO. International Organization for Standardization. URL: https://www.iso.org/about [accessed 2025-01-12]
55. Vaghefi I, Tulu B. The continued use of mobile health apps: insights from a longitudinal study. JMIR Mhealth Uhealth 2019 Aug 29;7(8):e12983 [FREE Full text] [doi: 10.2196/12983] [Medline: 31469081]
56. Dahlhausen F, Zinner M, Bieske L, Ehlers JP, Boehme P, Fehring L. Physicians' attitudes toward prescribable mHealth apps and implications for adoption in Germany: mixed methods study. JMIR Mhealth Uhealth 2021 Nov 23;9(11):e33012 [FREE Full text] [doi: 10.2196/33012] [Medline: 34817385]
57. Richter JG, Chehab G, Stachwitz P, Hagen J, Larsen D, Knitza J, et al. One year of digital health applications (DiGA) in Germany - Rheumatologists' perspectives. Front Med (Lausanne) 2022;9:1000668 [FREE Full text] [doi: 10.3389/fmed.2022.1000668] [Medline: 36388899]
58. Caiani EG, Kemps H, Hoogendoorn P, Asteggiano R, Böhm A, Borregaard B, et al. Standardized assessment of evidence supporting the adoption of mobile health solutions: a clinical consensus statement of the ESC Regulatory Affairs Committee: developed in collaboration with the European Heart Rhythm Association (EHRA), the Association of Cardiovascular Nursing & Allied Professions (ACNAP) of the ESC, the Heart Failure Association (HFA) of the ESC, the ESC Young Community, the ESC Working Group on e-Cardiology, the ESC Council for Cardiology Practice, the ESC Council of Cardio-Oncology, the ESC Council on Hypertension, the ESC Patient Forum, the ESC Digital Health Committee, and the European Association of Preventive Cardiology (EAPC). Eur Heart J Digit Health 2024 Sep;5(5):509-523. [doi: 10.1093/ehjdh/ztae042] [Medline: 39318699]
59. Exploring the digital health landscape in the WHO European Region: digital health country profiles. World Health Organization. URL: https://iris.who.int/handle/10665/376540 [accessed 2025-01-12]
60. Assessment framework for mHealth apps. Australian Digital Health Agency. URL: https://www.digitalhealth.gov.au/about-us/strategies-and-plans/assessment-framework-for-mhealth-apps [accessed 2025-01-12]
61. van Kessel R, Srivastava D, Kyriopoulos I, Monti G, Novillo-Ortiz D, Milman R, et al. Digital health reimbursement strategies of 8 European countries and Israel: scoping review and policy mapping. JMIR Mhealth Uhealth 2023 Sep 29;11:e49003 [FREE Full text] [doi: 10.2196/49003] [Medline: 37773610]
62. Work plan: A fit-for-purpose HTA framework for DHTs. EDiHTA. URL: https://edihta-project.eu/workplan/ [accessed 2025-01-12]
63. About the project. ASSESS DHT. URL: https://assess-dht.eu/the-project/ [accessed 2025-01-12]
64. Position on the European Health Data Space. CPME. URL: https://www.cpme.eu/api/documents/adopted/2022/11/cpme.2022-065.FINAL.CPME.position.EHDS.pdf [accessed 2025-01-12]

Abbreviations

AF: assessment framework
CEN: Comité Européen de Normalisation (European Committee for Standardization)
COREQ: Consolidated Criteria for Reporting Qualitative Research
EU: European Union
FTSS: Fundació TIC Salut Social (TIC Salut Social Foundation)
HCP: health care professional
HTA: health technology assessment
ISO: International Organization for Standardization

XSL•FO
RenderX

ISS: Istituto Superiore di Sanità (National Institute of Health)
mHealth: mobile health
TS: Technical Specification
VPC: value proposition canvas
WHO: World Health Organization

Edited by A Mavragani; submitted 22.10.24; peer-reviewed by J Haverinen, D Westmattelmann; comments to author 29.11.24; revised version received 19.12.24; accepted 14.02.25; published 21.04.25.

Please cite as:
Hoogendoorn P, Shokralla M, Willemsen R, Guldemond N, Villalobos-Quesada M
Compatibility of the CEN-ISO/TS 82304-2 Health App Assessment Framework With Catalan and Italian Health Authorities' Needs: Qualitative Interview Study
JMIR Form Res 2025;9:e67855
URL: https://formative.jmir.org/2025/1/e67855
doi:10.2196/67855
PMID:

©Petra Hoogendoorn, Mariam Shokralla, Romy Willemsen, Nick Guldemond, María Villalobos-Quesada. Originally published in JMIR Formative Research (https://formative.jmir.org), 21.04.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Mobile Health App Informed by the Multi-Process Action Control Framework to Promote Physical Activity Among Inactive Adults: Iterative Usability Study

Heather Hollman[1], MRSc, DC; Wuyou Sui[2], PhD; Haowei Zhang[3], BSc; Ryan E Rhodes[3], PhD

[1]Behavioural Medicine Lab, School of Exercise Science, Physical and Health Education, University of Victoria, Victoria, BC, Canada
[2]School of Kinesiology, Western University, London, ON, Canada
[3]Behavioural Medicine Lab, School of Exercise Science, Physical & Health Education, University of Victoria, Victoria, BC, Canada

Corresponding Author:
Heather Hollman, MRSc, DC
Behavioural Medicine Lab
School of Exercise Science, Physical and Health Education
University of Victoria
3800 Finnerty Rd
Room 173 McKinnon Building
Victoria, BC, V8P 5C2
Canada
Phone: 1 250 472 5288
Email: heatherh@uvic.ca

Abstract

Background: Mobile health apps have high potential to address the widespread deficit in physical activity (PA); however, they have demonstrated greater impact on short-term PA compared to long-term PA. The multi-process action control (M-PAC) framework promotes sustained PA behavior by combining reflective (eg, attitudes) and regulatory (eg, planning and emotion regulation) constructs with reflexive (eg, habits and identity) constructs. Usability testing is important to determine the integrity of a mobile health app's intrinsic properties and suggestions for improvement before feasibility and efficacy testing.

Objective: This study aimed to gather usability feedback from end users on a first and a second version of an M-PAC app prototype.

Methods: First, 3 workshops and focus groups, with 5 adult participants per group, were conducted to obtain first impressions of the M-PAC app interface and the first 3 lessons. The findings informed several modifications to the app program (eg, added cards with reduced content) and its interface (eg, created a link placeholder image and added a forgot password feature). Subsequently, a single-group pilot usability study was conducted with 14 adults who were not meeting 150 minutes per week of moderate-to-vigorous PA. They used the updated M-PAC app for 2 weeks, participated in semistructured interviews, and completed the Mobile App Usability Questionnaire (MAUQ) to provide usability and acceptability feedback. The focus groups and interviews were recorded, transcribed, and analyzed with content analysis informed by usability heuristics. The MAUQ scores were analyzed descriptively.

Results: Participants from the workshops and focus groups (mean age 30.40, SD9.49 years) expressed overall satisfaction with the app layout and content. The language was deemed appropriate; however, some terms (eg, self-efficacy) and acronyms (eg, frequency, intensity, time, and type) needed definitions. Participants provided several recommendations for the visual design (eg, more cards with less text). They experienced challenges in accessing and using the help module and viewing some images, and were unsure how to create or reset the password. Findings from the usability pilot study (mean age 41.38, SD12.92 years; mean moderate-to-vigorous PA 66.07, SD57.92 min/week) revealed overall satisfaction with the app layout (13/13, 100%), content (10/13, 77%), and language (7/11, 64%). Suggestions included more enticing titles and additional and variable forms of content (eg, visual aids and videos). The app was easy to navigate (9/13, 69%); however, some errors were identified, such as PA monitoring connection problems, broken links, and difficulties entering and modifying data. The mean MAUQ total and subscale scores were as follows: total=5.06 (SD1.20), usefulness=4.17 (SD1.31), ease of use=5.36 (SD1.27), and interface and satisfaction=5.52 (SD1.42).

XSL•FO
RenderX

Conclusions: Overall, the M-PAC app was deemed usable and acceptable. The findings will inform the development of the minimum viable product, which will undergo subsequent feasibility testing.

(JMIR Form Res 2025;9:e59477) doi:10.2196/59477

KEYWORDS

physical activity; mobile apps; mobile health; mHealth; usability study; inactive adults; smartphone

Introduction

Background

Regular physical activity (PA) is associated with reduced risk of cardiovascular disease, hypertension, diabetes, breast and colon cancer, and mortality [1]. It also has positive effects on cognition and mental health [2,3]. Unfortunately, PA rates worldwide are insufficient to reap these mental and physical health benefits [4]. Given the widespread burden of global inactivity, scalable PA interventions are needed. Mobile health (mHealth) apps have great potential for scalability and to meet harder-to-reach populations, such as rural residents, who are also at increased risk of inactivity [5]. As of 2024, >60% of the global population owns a smartphone, with even greater rates in high-income nations such as the United Kingdom (82%), the United States (82%), and Canada (72%) [6]. People spend the majority (88%) of their mobile time using apps [7], and in 2019, the average individual used 9 mobile apps per day and 30 apps per month [8].

mHealth apps are an appealing medium for PA behavioral interventions because they can provide up-to-date educational material while concurrently enabling continuous PA tracking and feedback [9-11]. They can also facilitate connections with peers and health providers [9-11] and more easily implement personalization and associations (eg, reminders) compared to print or in-person interventions [12]. From a theoretical standpoint, they can deliver a higher amount of behavior change techniques (BCTs) compared to in-person interventions [12]. Using well-defined BCTs (eg, goal setting and self-monitoring) has been associated with increased PA intervention engagement, which is an essential metric to address given the high app attrition rate and the connection between engagement and resulting PA behavior [13-15].

In line with the additional BCTs and related features, mHealth apps have demonstrated promising potential to impact PA behavior. In a recent meta-analysis of PA mHealth apps or trackers with automated and continuous self-monitoring and feedback, Laranjo et al [16] reported a small-to-moderate effect on PA (standardized mean difference [SMD]=0.350; 95% CI 0.236-0.465; I^2=69%; T^2=0.051) with a mean follow-up of 13 weeks. These standardized effects are higher than meta-analyses of workplace PA interventions (SMD=0.21) [17], internet-delivered PA interventions (SMD=0.14) [18], and face-to-face PA interventions (SMD=0.29) [19] and are equivalent to the 85th percentile of benchmarked PA intervention effects [20]. PA mHealth app interventions have also demonstrated longer-term (>6 months) effects on PA; however, the effect sizes tend to decrease over time [21], similar to PA interventions in general [22]. Because the health benefits can only be accrued if PA behavior change is sustained, PA maintenance, in addition to engagement, should be a focus of future PA mHealth apps.

To determine intervention components needed for longer-term PA behavior, PA theorists have attempted to conceptualize the acts of PA behavior adoption and maintenance separately [23,24]. For example, although constructs of traditional PA theories in the social cognitive domain (eg, outcome expectations or action planning) seem to be moderate predictors of PA adoption, they may be less comprehensive for maintenance [24,25]. Social cognitive theories highlight the importance of forming PA intentions; however, these approaches sometimes fail to account for the intention-PA gap (ie, the proportion of individuals who do not follow through with regular PA behavior that aligns with their PA intentions) [26-28]. A recent meta-analysis showed that 48% of adult PA intenders fail to follow through with PA [29].

Several theoretical approaches have been developed to close the intention-behavior gap [27,30]. An example is the multi-process action control (M-PAC) framework, which combines reflective processes, such as those in traditional social cognitive approaches (eg, attitudes), with regulatory (eg, self-monitoring and emotion regulation) and reflexive (eg, habit and exercise identity) processes to facilitate PA behavior change [31]. Collectively, the M-PAC constructs have a causal structure that begins with intention formation, progresses to the adoption of action control, and ultimately leads to sustained action control. Each stage is interconnected, with processes that are mutually reinforcing and reciprocal [32]. Each construct independently provides meaningful and significant contributions to translating positive intentions to PA behavior, with the largest predictive effect sizes coming from the latter reflexive processes [31]. PA interventions (in-person and web-based) informed by the M-PAC framework have demonstrated promising results on PA behavior and PA maintenance, promoting constructs of habit, self-identity, and behavioral regulation in feasibility trials [33-36]. However, an mHealth app informed by the M-PAC framework is yet to be tested.

To advance the likelihood of a successful and scalable mHealth app, it is recommended to assess its intrinsic properties, including usability and feasibility, in the initial prototyping phases [37]. An initial focus on these intrinsic properties can ultimately lead to greater attribution of the extrinsic metrics and outcomes to the intervention. Usability is the assessment of the quality of the interaction between the user and the intervention, while feasibility is the assessment of the ability of the intervention to work as intended in a given context [37]. Usability testing can help to uncover the overall interest in and eagerness to use the mHealth app, potential impediments to its use, and suggestions for improvement or innovation [38]. This

XSL•FO
RenderX

usability phase is usually iterative, whereby prototypes are refined and optimized through several stages based on user feedback. It is also essential to involve end users to ensure that the intervention is relevant and meets the specific needs and preferences of the target population [39,40]. Once this phase is complete, a minimum viable product (MVP) can be built and then pilot-tested to test for further usability and acceptability, followed by evaluation of its efficacy [38].

Research Objective

Thus, the objective of this research was to gather user feedback on the first and second prototypes of the M-PAC app to inform the design and development of the MVP. We initially explored the first impressions of the M-PAC app prototype using focus group methodology to gather important preliminary feedback from a group with diverse views and perspectives [41]. Subsequently, we used a usability pilot study [42] to retrieve more in-depth feedback on the usability and acceptability of the M-PAC app from potential users—individuals who intend to exercise but are not currently meeting the recommended PA guidelines.

Methods

Overarching App Development Framework

The M-PAC app development and design were in accordance with the IDEAS (integrate, design, assess, and share) framework [38]. Specifically, this study described 2 iterative rounds of phases 5 (prototype potential products) and 6 (gather user feedback) [38]. As per the IDEAS framework phase 5, an initial prototype was assessed during a series of workshops and focus groups to gather first impressions and suggestions for improvement. Commensurate with phase 6 of the IDEAS framework, the suggestions from the workshops and focus groups were implemented into the M-PAC app, and subsequently, the updated second app prototype was assessed in a usability pilot study, in which a new group of participants had the opportunity to use the app as intended for a few weeks and then provide additional feedback in individual interviews.

Phase 1 (workshops and focus groups) and the qualitative interview component of phase 2 (usability pilot) were reported following the COREQ (Consolidated Criteria for Reporting Qualitative Research) criteria [43]. Phase 2 (usability pilot) was reported following a modified version of the CONSORT (Consolidated Standards of Reporting Trials) guidelines [44], as recommended by Lancaster and Thabane [45] for nonrandomized pilot studies.

Ethical Considerations

These studies received ethics approval from the University of Victoria Human Research Ethics Board (21-0248; 21-0360). All participants provided informed consent before participating in the studies, and all study data are anonymous. Participants in phase 2 (usability pilot) received a CAD $15 (US $10.88) honorarium for participating in the study.

Phase 1: Workshops and Focus Groups

Design

Online hands-on workshops followed by focus groups were conducted. The workshops were conducted between November 2021 and March 2022. Focus groups were designed to be run with 4 to 6 participants, with a total of 15 to 20 participants, based upon previous usability studies using apps and web-based delivery [46-49]. The workshops and focus groups were led by WS and supported by HH (who also took field notes). WS has a PhD in health psychology, was a postdoctoral fellow at the time of the study, and identifies as a man. HH has a master's degree in rehabilitation science, was a PhD student at the time of the study, and identifies as a woman. WS had prior experience in conducting research focus groups and interviews and HH did not. WS and HH both have prior experience analyzing qualitative data. WS and HH introduced themselves to the participants as research trainees with a focus on PA behavior change.

Participants

Eligible participants were adults aged 19 to 64 years who owned a smartphone with either an iOS or Android operating system and were fluent in English. Participants were recruited via online advertisements on the research laboratory's public X account (X Corp; @bmedlab); public Instagram account (Meta Platforms, Inc; @uvicbmed); public community Facebook (Meta Platforms, Inc) groups local to Victoria, British Columbia; and through snowball sampling. All communication between researchers and participants took place over email. Some participants were known to the researchers because they worked in the same university; however, none of the participants had provided feedback on the app before the study.

Procedures

Interested participants contacted the researchers via email and were sent an email with the letter of information and consent form. Upon providing consent, participants were asked to provide availability for a 60-minute workshop and focus group. Participants were informed that they would be given a pseudonym during the workshops and focus groups to protect their identity. Participants were later contacted to reconfirm their availability once 4 to 5 participants reported having a mutually available time and were sent a Zoom videoconferencing link upon confirmation (Zoom Video Communications). In addition, participants were asked to download the Pathverse app platform [50] (Pathverse, Inc), which gave them access to the M-PAC app on their smartphone, to be able to actively participate in the workshop portion of the session.

One day before the date of the workshop or focus group, participant emails were used to create profiles on the admin portal of Pathverse, so that participants would have access to the M-PAC app. On the arranged time of the workshop or focus group, participants joined into a Zoom call with 2 researchers. All participants were given pseudonyms and were unable to turn on video. The lead researcher shared their smartphone screen and led the workshop and focus group discussion, while the other researcher managed the video recording software and

XSL•FO
RenderX

any questions within the Zoom chat. Upon entering the Zoom call, the lead researcher asked participants to open the Pathverse app and then instructed participants on how to create a profile on Pathverse using their email. Once all the participants had access to the app, the lead researcher began a walkthrough of the app.

Participants were first shown how to access the lessons, along with a brief overview of how to navigate through lesson cards, the different types of cards, and how to mark a lesson as complete, using "Lesson 0: Tutorial" and "Lesson 1: What is Physical Activity" as examples. Afterward, participants were directed to the tracking features on the app. Participants were shown how to sync their step count to the app as well as how to track and log PA, weight, and blood pressure. Participants were then shown the goal-setting feature of the app. The lead researcher created an example goal, demonstrating to participants how to name the goal, set priority and reminders on the goal, and how to mark goals as completed. Finally, participants were shown how to navigate to the home page of the app and to find the in-app navigation tutorial if they got stuck. Participants were then given 5 to 10 minutes to explore the app and its features independently. Overall, the workshop portion of the session took approximately 25 to 30 minutes.

Following the workshop, participants were informed that they would be starting the focus group portion of the session. The lead researcher read the introductory script and then prompted participants to ask any questions before beginning with the list of semistructured questions (Multimedia Appendix 1). The focus group portion of the session took approximately 25 to 30 minutes. The workshop and focus group guide was pilot-tested among the research team. Only the participants and researchers were present during the workshops and focus groups. There were no repeat workshops and focus groups.

Intervention: The M-PAC App

The M-PAC app is an evidence-based mobile app designed to support users in changing their PA behavior, developed by researchers specializing in PA behavior. The app is built upon the Pathverse app platform [50]. Pathverse is an mHealth development platform designed for health and wellness professionals and researchers. This platform consists of an admin web-portal interface and a participant Pathverse app (available in both IOS and Android stores). The M-PAC app, specifically, includes educational content in the form of lessons; self-monitoring features, such as step tracking and PA logging; and a goal-setting feature, whereby users can set goals for themselves as well as review completed goals (Figure 1).

Lessons within the app are broken down into individual "cards," which users navigate by swiping left or right, or via the arrow buttons at the bottom of the screen. Cards present either (1) a brief block of text, (2) a brief block of text with an image placed below or above, (3) a brief block of text with a link to an external web page, or (4) a challenge question. The challenge question card presents the user with a short question and 4 possible answers. Selecting any answer presents the user with a new slide indicating whether their answer was correct or incorrect, as well as a brief explanation for the correct answer.

Lessons within the app are scaffolded according to the different constructs within the M-PAC framework. Lessons 1 to 4 are designed to address reflective processes of the M-PAC framework: instrumental attitudes ("What are the health benefits of PA?"), perceived opportunity ("Do I have the resources to do PA?"), perceived capability ("Am I confident I can perform this PA?"), and affective judgments ("Do I like PA?"). Similarly, lessons 5 to 10 are aimed at addressing the regulatory (ie, skills important in facilitating continued PA behavior, such as goal setting and planning) and reflexive processes (eg, habit and exercise identity) of the M-PAC. In addition, lessons contain challenge cards that assess users' understanding of the content within their current lesson. Refer to Multimedia Appendix 2 [51] for a full list of app lessons, mechanisms of action, BCTs, and behavior change ontologies applied [52,53]. Lessons must be completed in order (ie, lesson 0, lesson 1, lesson 2, etc) to unlock subsequent lessons, but completed lessons can be revisited once unlocked. Given the usability focus of this study, only lessons 1 to 3 (including a lesson 0: tutorial) were presented to participants.

XSL•FO
RenderX

Figure 1. Screenshots of a physical activity mobile health app informed by the multi-process action control (M-PAC) framework (the M-PAC app). PA: physical activity.

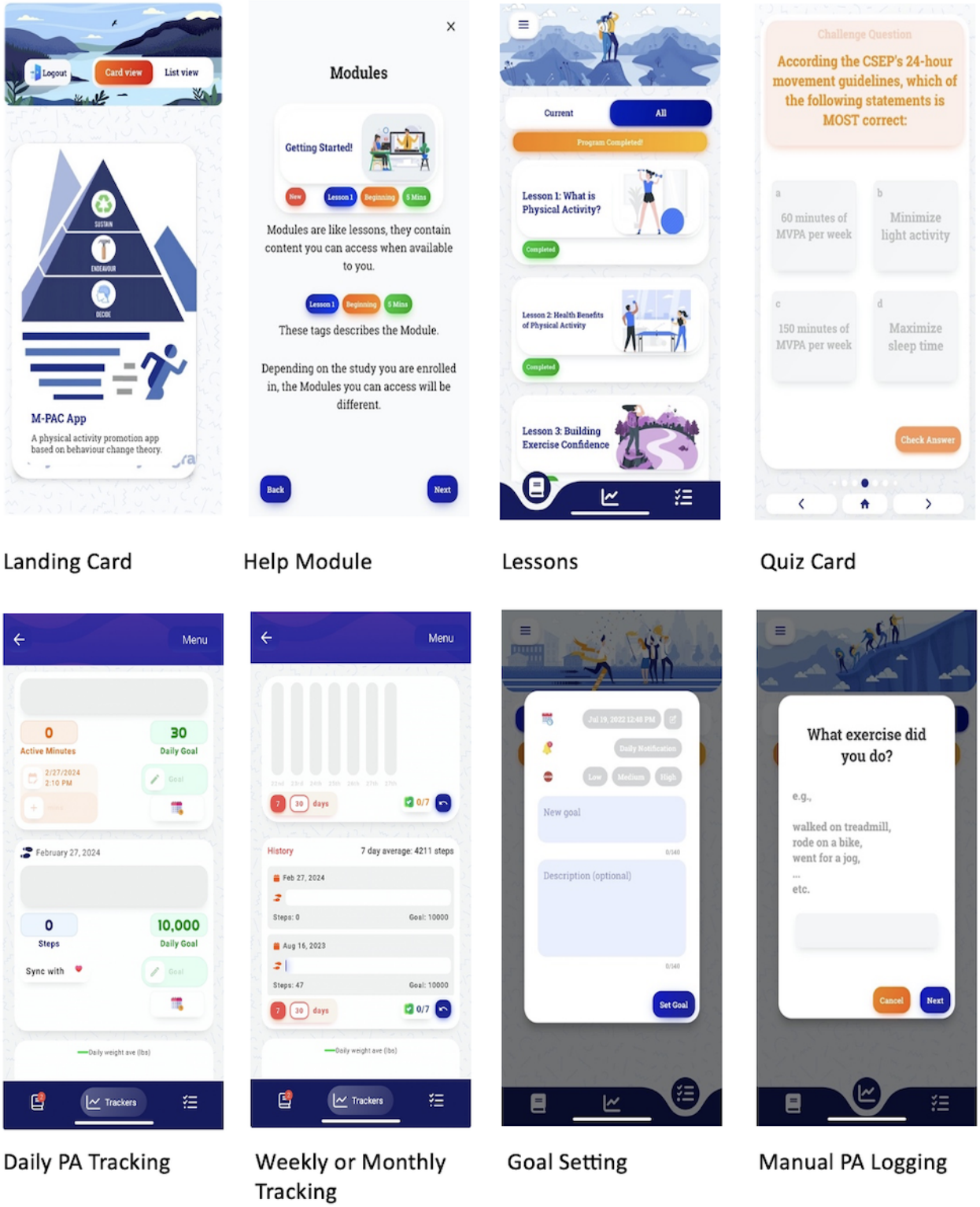

Outcomes

The primary outcome of this phase was qualitative feedback on the usability of the M-PAC app drawn from the focus groups. Specifically, questions within the focus group were drawn from Nielsen's 10 usability heuristics for user interface design [54]. These heuristics were adapted for this work to guide the structured focus group questions, which addressed the following: layout and information, appropriateness of language, backing out or undoing an action, consistency in language and features, errors and mistakes, forgetting acronyms, efficiency of interface and navigation, visual design, ease in diagnosing and resolving errors, and help and documentation. Probing questions for each main question were also planned if group responses were vague or open ended. For example, the first main question was, “How did you find the layout and information of the app?” with

XSL•FO
RenderX

probing questions of "If intuitive, what made the layout or information intuitive?" and "If unintuitive, what made the layout or information unintuitive? What would be more helpful?" The specific questions related to these heuristics can be found in Multimedia Appendix 3. Participant responses were recorded using the video recording software Open Broadcast Software (OBS) Studio [55] (OBS Studio Contributors).

Participant demographics of age, gender, and highest attained education were assessed using single-item questions. Furthermore, the brand of smartphone owned, ability to use the features of a smartphone, ability to engage with other technologies, and whether a PA app has been downloaded or used in the past were also assessed using single-item measures.

Analysis

Focus groups were recorded using OBS Studio and transcribed verbatim [55]. Transcripts were coded and analyzed thematically, according to the 10 usability heuristics [54] adapted for this study, using NVivo (QSR International) [56]. Participants were not asked to provide feedback on the transcripts or findings. Two coders independently coded the transcripts, with conflicts being resolved by a third coder.

Phase 2: Usability Pilot

Design

A 2-week, single-group pilot study was conducted virtually between August 2022 and February 2023. The researchers were located in Victoria, British Columbia, Canada, and the participants were located anywhere in Canada.

Participants

Eligible participants were Canadian adults aged 19 to 64 years who owned a smartphone device with either an iOS or Android operating system, were fluent in English, and reported engaging in <150 minutes of moderate-to-vigorous PA (MVPA) per week. Individuals were excluded if they participated in the phase 1 workshops or focus groups or were involved in the design and development of the M-PAC app. Participants were recruited via online advertisements on the research laboratory's public X account (@bmedlab), public Instagram account (@uvicbmed), and public community Facebook groups local to Victoria, British Columbia; with paid Facebook advertisements; and snowball sampling. All communication between researchers and participants took place over email. A total of 10 to 15 participants was deemed appropriate based on previous usability testing studies examining apps [57,58].

Procedures

Interested individuals contacted the research team via email and were sent an email with the letter of information and a consent form. Upon agreeing to participate in the study, participants were sent instructions on how to download the M-PAC app, as well as a demographics questionnaire. Participants also booked a 30-minute baseline meeting with a researcher to walkthrough the M-PAC app over Zoom. During the baseline meeting, participants were given a walkthrough of the M-PAC app, identical to the workshop from phase 1. Upon completing the walkthrough, participants were asked to schedule a 30-minute exit interview in 2 weeks and finish lessons 1 to 3 within those 2 weeks.

During the exit interview, the researcher followed the semistructured interview guide to assess participants' experiences with the M-PAC app (Multimedia Appendix 3). The interview guide was pilot-tested among the research team. Interviews were conducted by WS over Zoom, audio recorded using OBS Studio, and lasted 20 to 30 minutes. Only the participants and researchers were present, and there were no repeat interviews. After completing the exit interview, participants were told that they would be sent the Mobile App Usability Questionnaire (MAUQ) to their email in 24 hours, and to complete the MAUQ within 3 days of receiving the questionnaire [59]. Upon completion of the MAUQ, participants were given a CAD $15 (US $11.88) honorarium. There were no changes made to the methods upon pilot study commencement.

Intervention

Following the completion of phase 1, several modifications were made to the M-PAC app. Changes to content included the following: (1) modifying the introductory card for the self-efficacy lesson from "Introduction to self-efficacy..." to "Ways to improve confidence (also known as self-efficacy)...", (2) ensuring all acronyms (eg, MVPA; Canadian Society for Exercise Physiology; frequency, intensity, time, and type [FITT]; and Get Active Questionnaire) were clearly defined, (3) bolding some terms to highlight them (eg, FITT), (4) reducing the amount of text on cards and increasing the number of cards per lesson, and (5) adding a references module and providing links to references from each lesson. Changes to the Pathverse platform included the following: (1) optimizing image loading times, (2) creating an external link placeholder image that reads "Link," and (3) adding a "Forgot your password?" feature. The updated version of the M-PAC app with the aforementioned modifications was used for the usability pilot study.

Outcomes

The primary outcome of this phase was qualitative feedback on the usability and satisfaction of the M-PAC app, gathered through the exit interview. Probing questions for each main question were also planned if participants' responses were vague or open ended. For example, for the question, "Can you tell me what you liked best about the M-PAC app?", probing questions of "Were there any features in particular?" and "What about the content and/or layout?" were planned. The specific questions within this interview guide can be found in Multimedia Appendix 3. Participant responses were recorded using the video recording software OBS Studio [55]. The a priori criteria for moving on to a randomized pilot trial was an overall consensus that the intervention was usable (MAUQ score≥5), and at least 60% (8/13) of participants were satisfied with the intervention, with minor to moderate suggestions for improvement.

Quantitative feedback on the usability of the M-PAC app was collected via the MAUQ [59], specifically the version for stand-alone mHealth apps. This version of the MAUQ is an 18-item self-report questionnaire that assesses an mHealth app's

XSL•FO
RenderX

ease of use, interface and satisfaction, and usefulness. The questionnaire demonstrated construct and criterion validity and reliability, with Cronbach α=0.91 for the total MAUQ, 0.71 for ease of use, 0.93 for interface and satisfaction, and 0.65 for usefulness internal consistency.

Participant demographics of age, gender, and highest attained education were assessed using single-item questions. Furthermore, the brand of smartphone owned, ability to use the features of a smartphone, ability to engage with other technologies, and whether a PA app has been downloaded or used in the past were also assessed using single-item measures. Participants' average weekly PA was assessed using the modified Godin Leisure-Time PA questionnaire [60]. The estimated weekly MVPA was calculated by multiplying the frequency and duration of moderate- and vigorous-intensity PA minutes and adding them together. The Godin-Leisure-Time PA questionnaire has demonstrated acceptable validity and reliability in a variety of settings, populations, and countries [60,61]. There were no changes made to the assessments or measurements after the trial commenced.

Analyses

Exit interviews were recorded using OBS Studio and transcribed verbatim. Transcripts were coded and analyzed thematically, according to the 10 usability heuristics [54] adapted for this study, using NVivo [56]. Participants were not asked to provide feedback on the transcripts or findings. Two coders independently coded the transcripts, with conflicts being resolved by a third coder.

Descriptives (ie, mean and SD) for quantitative usability feedback from the MAUQ were calculated for total MAUQ scores, as well as for the ease of use, interface and satisfaction, and usefulness subscales [59].

Results

Phase 1: Workshops and Focus Groups

Participants

We conducted 3 focus groups of 5 participants each. Participants were primarily women (13/15, 87%) with a mean age of 30.4 (SD 9.49) years. Most had at least a college or university degree (14/15, 93%) and self-reported having above average abilities to use the functions of their smartphone (11/15, 73%), and just under half had previously downloaded a mobile app to help increase their PA (7/15, 47%; eg, Down Dog, Fitbit, Health, Nike Training Club app, Endomondo, and Sweat app). The most common brand of smartphone was Apple iPhone (9/15, 60%), followed by Samsung (4/15, 27%) and Google Pixel (2/15, 13%; Table 1). None of the participants refused to participate or dropped out of the study.

XSL•FO
RenderX

Table 1. Baseline characteristics of the workshops, focus groups, and pilot study participants.

	Workshops and focus groups (n=15)	Pilot (n=14)[a]
Age (y), mean (SD)	30.4 (9.49)	41.38 (12.92)
Gender, n (%)		
Women	13 (87)	11 (85)
Men	2 (13)	2 (15)
British Columbia	__[b]	Yes=8 (62) and no=5 (38)
Country (Canada)	15 (100)	13 (100)
Highest education, n (%)		
High-school diploma	1 (7)	1 (8)
Vocational school or some college	0 (0)	3 (23)
College or university degree	6 (40)	5 (38)
Professional or graduate degree	8 (53)	4 (31)
Brand of smartphone, n (%)		
Google Pixel	2 (13)	1 (8)
Samsung	4 (27)	4 (31)
Apple iPhone	9 (60)	5 (38)
LG	0 (0)	1 (8)
Huawei	0 (0)	1 (8)
Motorola	0 (0)	1 (8)
Ability to use the functions of their smartphone[c], n (%)		
Average	4 (27)	0 (0)
Good	6 (40)	5 (38)
Excellent	5 (33)	8 (62)
Abilities to engage with and use other types of technology generally[c], n (%)		
Poor	1 (7)	0 (0)
Average	3 (20)	0 (0)
Good	6 (40)	7 (54)
Excellent	5 (33)	6 (46)
Previously downloaded a mobile app to help increase their physical activity[d], n (%)		
Yes	7 (47)	8 (62)
Down Dog	2 (13)	0 (0)
Fitbit	1 (7)	2 (13)
Health	1 (7)	0 (0)
Nike training club app	1 (7)	0 (0)
Endomondo	1 (7)	0 (0)
Sweat App	1 (7)	0 (0)
Optimity	0 (0)	2 (15)
Carrot	0 (0)	1 (8)
Habitshare	0 (0)	1 (8)
MyFitnessPal	0 (0)	1 (8)
C25K	0 (0)	1 (8)
ParticipACTION app	0 (0)	1 (8)

XSL•FO
RenderX

	Workshops and focus groups (n=15)	Pilot (n=14)[a]
7M workout	0 (0)	1 (8)
No	8 (53)	5 (38)
If you have previously downloaded an app to increase your physical activity, was it helpful?[d], n (%)		
Yes	3 (20)	6 (46)
No	4 (27)	2 (15)
Physical activity behavior score, mean (SD)		
Moderate physical activity	—	54.29 (52.87)
Vigorous physical activity	—	11.79 (20.90)
MVPA[e]	—	66.07 (57.92)

[a]One participant did not complete the baseline questionnaire.
[b]Not applicable.
[c]Scale choices: terrible, poor, average, good, and excellent.
[d]Scale choices: yes and no.
[e]MVPA: moderate- to-vigorous physical activity.

Focus Group Findings

Overview

Refer to Table 2 for a comprehensive overview of the findings from the focus groups. Data saturation, where no new themes were revealed, was reached after 3 focus groups, which is in line with research stating that data saturation can be achieved with 3 to 5 focus groups [62].

XSL•FO
RenderX

Table 2. Multi-process action control (M-PAC) app usability focus group themes informed by the usability heuristics by Nielsen [54].

Heuristic theme	Positive findings	Negative findings and suggestions
Layout and information (including information on features)	• The flow was nice and attractive and easy to navigate • The 3 icons at the bottom were straightforward • Liked the card format	• Multiple levels of content for people who want more information • The menu button should be reworded (does not take the user to a menu) • Some people wanted the congratulations card before the reference card • Trackers and self-reports should be in the same tab
Appropriateness of language	• Language was appropriate for the general audience • Lay terms that anyone could understand • Fun theory	• Language was for individuals who had higher physical literacy • Once in a while, some words could have been different • Some terms, such as self-efficacy, could be initially presented using a more widely known term • Base screen says "activate account" instead of "create an account," which is not intuitive
Backing out or undoing an action	• Good that there were two options to go back—you could slide or also click • Consensus that this was easy to do	• Upon returning to a lesson from the home page, one needed to restart the lesson • When they tracked a self-monitoring measure (blood pressure, PA[a], or weight), it could not be edited or deleted • Completed goals had a clickable "Completed" button, which made the goal "uncompleted" and took it back to the current tab
Consistency in language and features	• Generally, people thought language and features were consistent	• Modules versus lessons • Icon differences in goals depending on Android or Apple phone (eg, "I get a little house a little running man and then something that looks like a golf a putting green flag") • Some things did not show up on iPad (long-term benefits with percentages)
Errors and mistakes	—[b]	• Connection to Fitbit not working • Some images did not load • Some images were glitching (24-hour movement guideline) • No way to reset the password • For Android users, external links led to blank white pages • Some images showed up too small
Forgetting acronyms	—	• Sometimes an acronym was presented without defining it (MVPA[c], CSEP[d], FITT[e], and PAR-Q[f])
Efficiency of interface and navigation	• Easy to swipe with 1 hand	• Ability to connect to other PA-measure devices • The swipe left button was too small • Some tabs were swipe and click, while others were only click • Desire to include the ability to link out to external web pages for references • Each page should have a home page button
Visual design	• People liked that similar images and illustrations were used • People liked the color scheme • People liked the representation among the images • The design of the tabs was clean	• Bold the FITT • Too much text on some cards • The external link placeholder image was too small and does not intuitively make people think it was a link • Wanted more visual pop within the lessons • Wanted the landscape format • The check question button on the quiz answer card was not in an intuitive location • On some pages, there was a lot of dead white space at the bottom • One person wanted real images rather than the clip art • More cards with less text • When setting goals, the grey buttons were not intuitive and should be displayed in a different color for better clarity • Program card could be better optimized (currently the title is cut off, and it flips to a sample page) • Would like more color on some of the slides

XSL•FO
RenderX

Heuristic theme	Positive findings	Negative findings and suggestions
Ease in diagnosing and resolving errors	—	• Could not figure out password components (ie, how many letters or numbers were required) • Needed to go through the entire set of help slides to find a solution • The help button was hard to find
Help and documentation	—	• Provided information on how to connect the Fitbit • Help cards did not have a home button and could not be skipped • The help button should be more accessible • Ability to send error reports on specific cards, in the help tab, or at the end of modules • Had to slide through the help slides individually • Help icon could be changed
Additional features	—	• Wanted to log other metrics, such as distance • Want a progress bar • The switch studies button should be a separate button • Wanted a sound at the end of the lesson • Desire to bookmark certain pages • Desire for a toolbox for important questionnaires or favorite cards

[a]PA: physical activity.

[b]Not applicable.

[c]MVPA: moderate-to-vigorous physical activity.

[d]CSEP: Canadian Society for Exercise Physiology.

[e]FITT: frequency, intensity, time, type.

[f]PAR-Q: Physical Activity Readiness Questionnaire.

Theme 1: Layout and Information

Participants stated that overall the app was attractive and had straightforward icons. They also enjoyed that the app used cards to convey information, and they liked the flow when swiping between cards. However, participants disliked that the “Menu” button only allowed them to switch studies instead of lessons. Moreover, participants asked for multiple depths of content for people with various physical literacy levels.

Theme 2: Appropriateness of Language

Participants agreed that the app used lay terms to make the language appropriate and understandable for the general audience. Nevertheless, they suggested that a more PA-literate population might find the language too simple. In addition, the definition of some terms (eg, self-efficacy) could have been introduced earlier to enhance users’ understanding.

Theme 3: Backing Out or Undoing an Action

Participants liked that the app offered an easy way to navigate back from a page. For instance, they liked that the app had 2 options to return to the previous page (ie, by clicking or sliding). However, participants disliked that they needed to restart the lesson once they returned to the home page, could not edit or delete self-monitoring measures such as blood pressure and weight, and completed goals still had a clickable “completed” button, which uncompleted the goal.

Theme 4: Consistency in Language and Features

Generally, the participants agreed that the app’s language and features were consistent. However, participants reported inconsistency in wording and icons across different platforms.

Theme 5: Errors and Mistakes

Participants reported that the app could not connect to Fitbit. In addition, some images glitched (eg, showed up too small) or would not load. Participants also stated that they could not reset the password for the app. Another error was that external links led to blank pages for Android users.

Theme 6: Forgetting Acronyms

The focus group participants suggested that sometimes the lessons introduced acronyms (eg, MVPA, CSEP, FITT, and PAR-Q) without providing definitions.

Theme 7: Efficiency of the Interface and Navigation

The participants liked that they could easily swipe through the app when navigating. However, they reported that some tabs could only be clicked on, whereas others could be swiped through and clicked on. Participants also thought the swipe left button was too small, making clicking difficult.

Theme 8: Visual Design

The focus groups liked that the images and illustrations were coherent in style throughout the app. They also enjoyed the color scheme, visual representation, and the design of the tabs. However, participants did not like that key terms, such as FITT, were not bolded. They also disfavored that certain cards were text heavy.

XSL•FO
RenderX

Theme 9: Ease in Diagnosing and Resolving Errors

The participants reported that the app did not provide information about the password requirements when creating a password for log-in. They also disliked that the "help" button was hard to find, and they needed to go through the entire set of slides to find a solution to their problems.

Theme 10: Help and Documentation

Participants disliked that the app did not provide sufficient information on how to connect to Fitbit, the "Help" cards were not accessible (eg, the "Help" cards did not include a "Home" button and users could not skip to the card they wanted), and the lack of the ability to send error reports on specific cards.

Theme 11: Additional Suggestions on Features

Participants asked for features such as a progress bar in the lessons, a sound effect when finishing a lesson, a toolbox for important or favorite cards, a "home" button on each page, hyperlinks to reference articles, the ability to bookmark specific pages, and the option to input different PA metrics (eg, distance walked).

Phase 2: Usability Pilot

Participants

Fourteen participants took part in phase 2 (usability pilot). One participant did not complete the baseline questionnaire due to unknown reason. Participants were mainly women (11/13, 85%) with a mean age of 41.38 (SD 12.92) years. Most (9/13, 69%) had at least a college or university degree, all (13/13, 100%) self-reported having above average abilities to use the functions of their smartphone, and most (8/13, 62%) had previously downloaded a mobile app to help increase their PA(eg, Fitbit, Optimity, Carrot, Habitshare, MyFitnessPal, C25K, ParticipACTION app, and 7M Workout). The most common brand of smartphone was Apple iPhone (5/13, 38%), followed by Samsung (4/13, 31%), Google Pixel (1/13, 8%), Hauwei (1/13, 8%), LG (1/13, 8%), and Motorola (1/13, 8%; Table 1). The participants reported on average 66.07 (SD 57.92) minutes of MVPA per week. None of the participants refused to participate or dropped out of the study.

Interview Findings

Overview

Refer to Multimedia Appendix 4 for a comprehensive overview of the findings from the interviews. A total of 14 interviews were conducted; however, only 13 interviews were recorded due to a malfunction of the recording instrument. Data saturation, where no new themes were revealed, was reached after 13 interviews, which is in line with research stating that data saturation can be achieved with 8 to 16 interviews [62].

Theme 1: Layout and Information

Of the 13 participants, 13 (100%) commented on the layout and information of the app, and everyone indicated something they liked (eg, straightforward layout and modules took a reasonable amount of time to complete). Of the 13 participants, 6 (46%) disliked the app's layout and information due to inefficient content layout (eg, too many pages for a lesson), content and questions being too basic, and some content lacking references.

Theme 2: Appropriateness of Language

Of the 11 participants that commented on this theme, 7 (64%) acknowledged that the language used in the app was clear and easy to understand. However, of the 11 participants, 7 (64%) also commented on what they disliked regarding the app's language choice, including the lessons not having descriptive titles. Some participants did not understand the meaning of the app's name (M-PAC app) and thought it did not represent the app's intent.

Theme 3: Backing Out or Undoing an Action

One (8%) participant commented on this theme and they mentioned that returning to the app's home screen was inconvenient.

Theme 4: Consistency in Language and Features

Of the 6 participants that commented on this theme, 6 (100%) discussed the app's consistency in language and features. All of them recognized something they disliked, including inconsistency in link descriptions (eg, some links had more details on why to click the link and some had no explanation), notification sounds appearing too frequently, and color labeling for the priority goals not following convention. By contrast, of the 6 participants, 2 (33%) liked that the app followed conventional color schemes (eg, active links were blue) and connected to conventional devices such as Fitbits.

Theme 5: Errors and Mistakes

Of the 11 participants that commented on this theme, 11 (100%) reported errors in the app. Some participants (4/11, 36%) stated that the app had problems connecting to Google Fit, and several links in the lessons did not work. Several users encountered general visual glitches, such as slow loading and pictures and data not being fully displayed on some devices. Other errors included the app being unresponsive, typos, problems with modifying weight and blood pressure information, step counts being inconsistent with other external trackers, and text not fitting on the screen.

Theme 6: Forgetting Acronyms

Of the 4 participants that commented on this theme, 3 (75%) did not understand what the app's name (ie, Pathverse or M-PAC app) represents, and 1 (25%) participant forgot that PAR-Q stands for the Physical Activity Readiness Questionnaire.

Theme 7: Efficiency of the Interface and Navigation

Of the 13 participants that commented on this theme, 9 (69%) thought the app was overall well structured and easy to navigate. However, of the 13 participants, 7 (54%) discussed aspects they disliked about the app's interface and navigation, such as the tracker feature being hard to find, not being able to go straight to the home screen after finishing a lesson, goals needing to be reentered daily, the tracking tab did not show dates for active minutes and daily goals, active minutes could not be entered for previous dates, and the progress of an unfinished lesson could not be saved.

Theme 8: Visual Design

Of the 13 participants that commented on this theme, 4 (31%) liked the app's overall visual design, with comments stating

XSL•FO
RenderX

that the app had visually pleasing graphics and engaging diagrams, and the color scheme helped user navigation (eg, blue active links turned purple when clicked). Of the 13 participants, 7 (54%) did not enjoy the app's visual design for reasons such as the color scheme was not esthetic and lacked novelty (2/7, 29%), the graphic content did not include diverse body types (2/7, 29%), and the name and icon of the app were not attractive or meaningful to the users (3/7, 43%).

Theme 9: Ease in Diagnosing and Resolving Errors

Two (15%) participants commented on this theme and reported that no information was provided for the potential reason for errors (ie, no error messages or error codes).

Theme 10: Help and Documentation

One (8%) participant mentioned that the app did not provide information or warnings about upcoming updates.

Theme 11: Additional Feedback

Of the 13 participants that commented on this theme, 11 (85%) had a positive overall experience using the app. Many would likely use the app to improve their PA and recommend it to others. Of the 13 participants, 1 (8%) had a negative overall experience with the app. Some users stated that they would not likely use the app to improve their PA (3/13, 23%) or recommend it to others (4/13, 31%).

Theme 12: Additional Suggestions on Features

Most participants (11/13, 85%) suggested the features they would like to be included in the app, such as the ability to access new lessons overtime, take notes and make personal input during lessons, add comments to daily active minutes, and receive incentives on users' PA progress.

Usability

A total of 13 participants filled out the MAUQ (Table 3). However, 1 participant did not fill out the MAUQ for unknown reasons. The mean total MAUQ score was 5.06 (SD 1.20), and the mean subscale scores were as follows: "ease of use" 5.36 (SD 1.27), "interface and satisfaction" 5.52 (SD 1.42), and "usefulness" 4.17 (SD 1.31).

Table 3. Mobile App Usability Questionnaire (MAUQ) scores of the pilot usability study (n=13).

Subscale	Score, mean (SD)
Ease of use	5.36 (1.27)
Interface and satisfaction	5.52 (1.42)
Usefulness	4.17 (1.31)
Total MAUQ score	5.06 (1.20)

Discussion

Overview

The purpose of this research was to gather user feedback on the first 2 prototypes of the M-PAC app to inform the design and development of an MVP. The ultimate aim of the M-PAC app is to promote sustained PA intention-behavior translation; consequently, it is informed by the M-PAC framework, which combines reflective and regulatory processes that assist individuals in translating PA intentions to behaviors with reflexive processes that contribute to PA maintenance [23,24,31]. We assessed the impressions of the first app prototype among a convenience sample of adults and subsequently evaluated the usability and acceptability of the second app prototype among a group of adults who intended to exercise but were not meeting the recommended PA guidelines. The M-PAC app design and development is informed by the IDEAS framework, which recommends an iterative phase of gathering user feedback and usability testing before feasibility and efficacy testing [38]. This step is important to ensure there is overall interest in the app, to identify potential impediments, and to collect suggestions for improvement and innovation.

Principal Findings

Overall, the focus group and interview findings indicated that the M-PAC app was well received by target users and the usability scores were high. The findings also unveiled several suggestions related to usability that should be addressed before testing an MVP.

Comparison With Prior Work

Participants thought the M-PAC app was easy to use and navigate, the layout was straightforward, and the lessons took a reasonable amount of time. Altogether, this feedback aligns with the "less-is-more" strategy, which has been proposed for effective human-technology interactions [63]. Enhancing the appeal and usability of the M-PAC app is crucial, especially considering that in 2022, the average retention rate for health and fitness apps dropped from 37% on the first day to just 9% by day 28 [13]. The most frequent reasons for abandoning mHealth apps are a lack of interest or declining motivation, preference for other apps, lack of desired features, lack of enjoyment (fun), and ease of use [64]. Consequently, the challenges with some of the features and navigation of the M-PAC app will need to be addressed, such as needing to reenter goals daily, inability to enter active minutes for previous dates, and inability to save lesson progress.

Several users reported difficulty with connecting to their PA tracking app. This is an important usability issue to address because users desire accurate and efficient tracking features [65]. Users also reported that the PA tracker feature was difficult to find and use (eg, activity minutes could not be entered for previous dates). Updating the PA tracking will need to be a priority because it facilitates self-monitoring, which is important for translating PA intentions into behaviors [10,11,66,67].

Another important characteristic to address is the app's name. Several participants critiqued the name "M-PAC app," stating it was not meaningful and did not represent the app's intent.

XSL•FO
RenderX

This is an interesting consideration that is specific to mHealth app behavioral interventions as it does not necessarily need to be addressed in face-to-face, educational leaflet, or website interventions. The app name is one of the first things that app users notice; therefore, it should be unique, easy to remember, and should clearly communicate the app's functionality [68]. We began brainstorming potential replacements to the app name that are more understandable to a diverse audience and represent its intent to promote long-term PA behavior, and ultimately, we decided on the name "Physical Activity for Life (PAL) app."

Participants voiced that they would like to continue accessing new lessons over time. This is supported by previous research stating that to maintain interest and encourage ongoing use of the app, users desire new and updated content [65]. Users desire tailored content, and to ensure they are receiving ongoing tailored content, a solution could be to implement the just-in-time adaptive intervention (JITAI) design. The JITAI design attempts to deliver support to the user "at the moment and in the context that the person needs it most and is most likely to be receptive" [69]. The M-PAC app development and design team is currently working on implementing a JITAI design.

The overall MAUQ scores from the usability pilot study were >5 for the total score, "ease of use," and "interface satisfaction" subscales. Given that MAUQ scores of ≥5 indicate increasing levels of agreeableness, these scores suggest high usability [59]. The mean "usefulness" subscale score was <5, indicating moderate to low usability. This is not surprising considering participants only had access to the first 3 lessons of the app, while the MVP is estimated to have 13 lessons. This likely also reflects the earlier feedback indicating that participants wanted more content.

Strengths and Limitations

Key strengths of the study are the sample demographics, particularly of the second study. Participants were aged, on average, 41.38 (SD 12.92) years, which is representative of millennials [70], who currently make up the most prevalent users of fitness apps [71]. Participants owned a range of different types of smartphone devices and were not currently meeting PA guidelines. A strength of the study design is that it used a mixed methods approach, including 2 types of qualitative feedback (eg, focus groups and interviews), as well as quantitative feedback in the form of the MAUQ. Mixed methods design offers unique and complementary insights into the multifaceted phenomena of mHealth usability [72]. Some limitations to the demographics are that the majority had a postsecondary degree and self-claimed to have good or excellent abilities to use the functions of their smartphone. Another limitation was that participants only had access to the first 3 lessons, out of the potential 13 lessons; however, the formatting will be the same for subsequent lessons, and the learnings from this study will be applied to those lessons as well. Because we are unaware of how the nature of regulatory and reflexive constructs may lend themselves to the current format, we plan to conduct further feasibility testing to more comprehensively understand whether the app is a viable means to deliver the complete intervention.

Conclusions

Overall, the first 2 M-PAC app prototypes were usable and acceptable among the users. We will address the suggestions in the design and development of the MVP. Key upgrades we plan to implement include testing connections with the PA tracking feature, shortening lessons, coming up with a new app name, optimizing image loading, providing the ability to log activity minutes from previous days, providing the ability to add notes to PA logs, and enabling users to save progress in lessons. Once we have completed these modifications, we will conduct feasibility testing with the MVP.

Acknowledgments

The authors would like to thank the participants for providing invaluable insights into the usability of the multi-process action control app. They would also like to thank their research assistant, Rhiaanah Villalobos, who played an integral role in planning the study. This research did not receive any funding.

Data Availability

The datasets generated and analyzed during this study are not available because the participants only consented to providing the data to the research team of this study.

Authors' Contributions

HH, WS, and RER conceived the study. HH and WS conducted the study and collected data. HH, WS, and HZ conducted data analysis. HH drafted the manuscript, which was edited by WS, HZ, and RER.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Focus group interview guide.

XSL•FO
RenderX

[DOCX File , 17 KB - formative_v9i1e59477_app1.docx]

Multimedia Appendix 2
Physical activity mobile health app informed by the multi-process action control (M-PAC) framework: lessons, mechanisms of action, and behavior change techniques.
[DOCX File , 24 KB - formative_v9i1e59477_app2.docx]

Multimedia Appendix 3
Pilot study exit interview guide.
[DOCX File , 16 KB - formative_v9i1e59477_app3.docx]

Multimedia Appendix 4
Physical activity mobile health app informed by the multi-process action control (M-PAC) framework: usability interview themes informed by the Nielsen's usability heuristics by Nielsen [54].
[DOCX File , 42 KB - formative_v9i1e59477_app4.docx]

References

1. Grandes G, Garcia-Alvarez A, Ansorena M, Ortega Sanchez-Pinilla R, Torcal J, Arietaleanizbeaskoa MS, et al. Any increment in physical activity reduces mortality of primary care inactive patients. Br J Gen Pract 2023;73(726):e52-e58. [doi: 10.3399/bjgp.2022.0118]
2. Erickson KI, Hillman C, Stillman CM, Ballard RM, Bloodgood B, Conroy DE, For 2018 Physical Activity Guidelines Advisory Committee*. Physical activity, cognition, and brain outcomes: a review of the 2018 physical activity guidelines. Med Sci Sports Exerc 2019 Jun;51(6):1242-1251 [FREE Full text] [doi: 10.1249/MSS.0000000000001936] [Medline: 31095081]
3. White RL, Babic MJ, Parker PD, Lubans DR, Astell-Burt T, Lonsdale C. Domain-specific physical activity and mental health: a meta-analysis. Am J Prev Med 2017 May 02;52(5):653-666. [doi: 10.1016/j.amepre.2016.12.008] [Medline: 28153647]
4. Guthold R, Stevens GA, Riley LM, Bull FC. Worldwide trends in insufficient physical activity from 2001 to 2016: a pooled analysis of 358 population-based surveys with 1·9 million participants. Lancet Glob Health 2018 Oct;6(10):e1077-e1086. [doi: 10.1016/s2214-109x(18)30357-7]
5. Abildso CG, Daily SM, Umstattd Meyer MR, Perry CK, Eyler A. Prevalence of meeting aerobic, muscle-strengthening, and combined physical activity guidelines during leisure time among adults, by rural-urban classification and region - United States, 2020. Am J Transplant 2023 Mar;23(3):443-446. [doi: 10.1016/j.ajt.2023.01.021] [Medline: 36740195]
6. Turner A. How many people have smartphones worldwide. BankMyCell. URL: https://www.bankmycell.com/blog/how-many-phones-are-in-the-world [accessed 2024-03-16]
7. Wurmser Y. The majority of Americans' mobile time spent takes place in apps. EMarketer. URL: https://www.insiderintelligence.com/content/the-majority-of-americans-mobile-time-spent-takes-place-in-apps [accessed 2024-03-16]
8. Ahmad I. 60+ fascinating smartphone apps usage statistics for 2019. Social Media Today. URL: https://www.socialmediatoday.com/news/60-fascinating-smartphone-apps-usage-statistics-for-2019-infographic/550990/ [accessed 2024-03-16]
9. Greaves CJ, Sheppard KE, Abraham C, Hardeman W, Roden M, Evans PH, IMAGE Study Group. Systematic review of reviews of intervention components associated with increased effectiveness in dietary and physical activity interventions. BMC Public Health 2011 Feb 18;11:119 [FREE Full text] [doi: 10.1186/1471-2458-11-119] [Medline: 21333011]
10. Knittle K, Nurmi J, Crutzen R, Hankonen N, Beattie M, Dombrowski SU. How can interventions increase motivation for physical activity? A systematic review and meta-analysis. Health Psychol Rev 2018 Sep 15;12(3):211-230. [doi: 10.1080/17437199.2018.1435299] [Medline: 29385950]
11. Michie S, Abraham C, Whittington C, McAteer J, Gupta S. Effective techniques in healthy eating and physical activity interventions: a meta-regression. Health Psychol 2009 Nov;28(6):690-701. [doi: 10.1037/a0016136] [Medline: 19916637]
12. Dugas M, Gao GG, Agarwal R. Unpacking mHealth interventions: a systematic review of behavior change techniques used in randomized controlled trials assessing mHealth effectiveness. Digit Health 2020;6:2055207620905411 [FREE Full text] [doi: 10.1177/2055207620905411] [Medline: 32128233]
13. Health and fitness app report 2024. Business of Apps. URL: https://www.businessofapps.com/data/health-fitness-app-report/ [accessed 2024-02-20]
14. Milne-Ives M, Homer SR, Andrade J, Meinert E. Potential associations between behavior change techniques and engagement with mobile health apps: a systematic review. Front Psychol 2023;14:1227443 [FREE Full text] [doi: 10.3389/fpsyg.2023.1227443] [Medline: 37794916]

XSL•FO
RenderX

15. Smith N, Liu S. A systematic review of the dose-response relationship between usage and outcomes of online physical activity weight-loss interventions. Internet Interv 2020 Dec;22:100344 [FREE Full text] [doi: 10.1016/j.invent.2020.100344] [Medline: 32995302]
16. Laranjo L, Ding D, Heleno B, Kocaballi B, Quiroz JC, Tong HL, et al. Do smartphone applications and activity trackers increase physical activity in adults? Systematic review, meta-analysis and metaregression. Br J Sports Med 2021 Apr 21;55(8):422-432. [doi: 10.1136/bjsports-2020-102892] [Medline: 33355160]
17. Conn VS, Hafdahl AR, Cooper PS, Brown LM, Lusk SL. Meta-analysis of workplace physical activity interventions. Am J Prev Med 2009 Oct;37(4):330-339 [FREE Full text] [doi: 10.1016/j.amepre.2009.06.008] [Medline: 19765506]
18. Davies CA, Spence JC, Vandelanotte C, Caperchione CM, Mummery WK. Meta-analysis of internet-delivered interventions to increase physical activity levels. Int J Behav Nutr Phys Act 2012 Apr 30;9:52 [FREE Full text] [doi: 10.1186/1479-5868-9-52] [Medline: 22546283]
19. Conn VS, Hafdahl AR, Mehr DR. Interventions to increase physical activity among healthy adults: meta-analysis of outcomes. Am J Public Health 2011 Apr;101(4):751-758 [FREE Full text] [doi: 10.2105/AJPH.2010.194381] [Medline: 21330590]
20. Wright CE, Rhodes RE, Ruggiero EW, Sheeran P. Benchmarking the effectiveness of interventions to promote physical activity: a metasynthesis. Health Psychol 2021 Nov;40(11):811-821. [doi: 10.1037/hea0001118] [Medline: 34914485]
21. Mönninghoff A, Kramer JN, Hess AJ, Ismailova K, Teepe GW, Tudor Car L, et al. Long-term effectiveness of mHealth physical activity interventions: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2021 Apr 30;23(4):e26699 [FREE Full text] [doi: 10.2196/26699] [Medline: 33811021]
22. McEwan D, Rhodes RE, Beauchamp MR. What happens when the party is over?: sustaining physical activity behaviors after intervention cessation. Behav Med 2022 Apr 10;48(1):1-9. [doi: 10.1080/08964289.2020.1750335] [Medline: 32275199]
23. Dunton GF, Leventhal AM, Rebar AL, Gardner B, Intille SS, Rothman AJ. Towards consensus in conceptualizing and operationalizing physical activity maintenance. Psychol Sport Exerc 2022 Jul;61:102214. [doi: 10.1016/j.psychsport.2022.102214]
24. Rhodes RE, Sui W. Physical activity maintenance: a critical narrative review and directions for future research. Front Psychol 2021 Sep 6;12:725671 [FREE Full text] [doi: 10.3389/fpsyg.2021.725671] [Medline: 34552537]
25. van Stralen MM, De Vries H, Mudde AN, Bolman C, Lechner L. Determinants of initiation and maintenance of physical activity among older adults: a literature review. Health Psychology Review 2009 Sep;3(2):147-207. [doi: 10.1080/17437190903229462]
26. Cheval B, Boisgontier M. The theory of effort minimization in physical activity. Exerc Sport Sci Rev 2021 Jul 01;49(3):168-178 [FREE Full text] [doi: 10.1249/JES.0000000000000252] [Medline: 34112744]
27. Rhodes RE, Yao CA. Models accounting for intention-behavior discordance in the physical activity domain: a user's guide, content overview, and review of current evidence. Int J Behav Nutr Phys Act 2015 Feb 07;12(1):9 [FREE Full text] [doi: 10.1186/s12966-015-0168-6] [Medline: 25890238]
28. Schwarzer R. Modeling health behavior change: how to predict and modify the adoption and maintenance of health behaviors. Appl Psychol 2008 Jan 30;57(1):1-29. [doi: 10.1111/j.1464-0597.2007.00325.x]
29. Feil K, Fritsch J, Rhodes RE. The intention-behaviour gap in physical activity: a systematic review and meta-analysis of the action control framework. Br J Sports Med 2023 Oct 17;57(19):1265-1271. [doi: 10.1136/bjsports-2022-106640] [Medline: 37460164]
30. Schwarzer R. Health action process approach (HAPA) as a theoretical framework to understand behavior change. Act Psi 2016 Dec 05;30(121):119. [doi: 10.15517/ap.v30i121.23458]
31. Rhodes RE. Translating physical activity intentions into behavior: reflective, regulatory, and reflexive processes. Exerc Sport Sci Rev 2024 Jan 01;52(1):13-22. [doi: 10.1249/JES.0000000000000329] [Medline: 38126402]
32. Rhodes RE. The evolving understanding of physical activity behavior: a multi-process action control approach. In: Elliot AJ, editor. Advances in Motivation Science. Volume 4. New York, NY: Academic Press; 2017:171-205.
33. Hollman H, Sui W, Rhodes RE. A feasibility randomized controlled trial of a multi-process action control web-based intervention that targets physical activity in mothers. Women Health 2022 May 15;62(5):384-401. [doi: 10.1080/03630242.2022.2072050] [Medline: 35575171]
34. Husband CJ, Wharf-Higgins J, Rhodes RE. A feasibility randomized trial of an identity-based physical activity intervention among university students. Health Psychol Behav Med 2019 Apr 10;7(1):128-146 [FREE Full text] [doi: 10.1080/21642850.2019.1600407] [Medline: 34040843]
35. Kaushal N, Rhodes RE, Meldrum JT, Spence JC. Mediating mechanisms in a physical activity intervention: a test of habit formation. J Sport Exerc Psychol 2018 Apr 01;40(2):101-110. [doi: 10.1123/jsep.2017-0307] [Medline: 29945470]
36. Lim C, Wharf-Higgins J, Rhodes R. Working out with F.I.D.O. (frequency, intensity, duration, and outcomes): a feasibility randomized controlled trial. Hum Anim Interact Bull 2021 Dec;9(1):76-127 [FREE Full text] [doi: 10.1079/hai.2021.0025]
37. Monitoring and evaluating digital health interventions: a practical guide to conducting research and assessment. World Health Organization. 2016. URL: https://apps.who.int/iris/handle/10665/252183 [accessed 2022-02-25]

XSL•FO
RenderX

38. Mummah SA, Robinson TN, King AC, Gardner CD, Sutton S. IDEAS (integrate, design, assess, and share): a framework and toolkit of strategies for the development of more effective digital interventions to change health behavior. J Med Internet Res 2016 Dec 16;18(12):e317 [FREE Full text] [doi: 10.2196/jmir.5927] [Medline: 27986647]
39. Slattery P, Saeri AK, Bragge P. Research co-design in health: a rapid overview of reviews. Health Res Policy Syst 2020 Feb 11;18(1):17. [doi: 10.1186/s12961-020-0528-9] [Medline: 32046728]
40. Yardley L, Morrison L, Bradbury K, Muller I. The person-based approach to intervention development: application to digital health-related behavior change interventions. J Med Internet Res 2015 Jan 30;17(1):e30 [FREE Full text] [doi: 10.2196/jmir.4055] [Medline: 25639757]
41. Ivanoff SD, Hultberg J. Understanding the multiple realities of everyday life: basic assumptions in focus-group methodology. Scand J Occup Ther 2006 Jun 12;13(2):125-132 [FREE Full text] [doi: 10.1080/11038120600691082] [Medline: 16856469]
42. Bond C, Lancaster GA, Campbell M, Chan C, Eddy S, Hopewell S, et al. Pilot and feasibility studies: extending the conceptual framework. Pilot Feasibility Stud 2023 Feb 09;9(1):24 [FREE Full text] [doi: 10.1186/s40814-023-01233-1] [Medline: 36759879]
43. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
44. Eldridge SM, Chan CL, Campbell MJ, Bond CM, Hopewell S, Thabane L, PAFS consensus group. CONSORT 2010 statement: extension to randomised pilot and feasibility trials. BMJ 2016 Oct 24;355:i5239 [FREE Full text] [doi: 10.1136/bmj.i5239] [Medline: 27777223]
45. Lancaster GA, Thabane L. Guidelines for reporting non-randomised pilot and feasibility studies. Pilot Feasibility Stud 2019 Oct 06;5(1):114 [FREE Full text] [doi: 10.1186/s40814-019-0499-1] [Medline: 31608150]
46. Bae YS, Kim KH, Choi SW, Ko T, Lim JS, Piao M. Satisfaction and usability of an information and communications technology-based system by clinically healthy patients with COVID-19 and medical professionals: cross-sectional survey and focus group interview study. JMIR Form Res 2021 Aug 26;5(8):e26227 [FREE Full text] [doi: 10.2196/26227] [Medline: 34254946]
47. Daniels N, Gillen P, Casson K, Wilson I. STEER: factors to consider when designing online focus groups using audiovisual technology in health research. Int J Qual Methods 2019 Nov 19;18:160940691988578. [doi: 10.1177/1609406919885786]
48. Milward J, Deluca P, Drummond C, Watson R, Dunne J, Kimergård A. Usability testing of the BRANCH smartphone app designed to reduce harmful drinking in young adults. JMIR Mhealth Uhealth 2017 Aug 08;5(8):e109 [FREE Full text] [doi: 10.2196/mhealth.7836] [Medline: 28790022]
49. Willemsen RF, Aardoom JJ, Chavannes NH, Versluis A. Online synchronous focus group interviews: Practical considerations. Qual Res 2022 Jun 30;23(6):1810-1820. [doi: 10.1177/14687941221110161]
50. Pathverse (Version 1). Pathverse Platform. 2019. URL: https://pathverse.ca/ [accessed 2020-12-21]
51. BCIO:007094. The Ontology Lookup Service (OLS). URL: https://www.ebi.ac.uk/ols4/ontologies/bcio/classes/http%253A%252F%252Fhumanbehaviourchange.org%252Fontology%252FBCIO_007094 [accessed 2024-06-25]
52. Michie S, Richardson M, Johnston M, Abraham C, Francis J, Hardeman W, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013 Aug;46(1):81-95 [FREE Full text] [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
53. Marques MM, Wright AJ, Corker E, Johnston M, West R, Hastings J, et al. The behaviour change technique ontology: transforming the behaviour change technique taxonomy v1. Wellcome Open Res 2023;8:308 [FREE Full text] [doi: 10.12688/wellcomeopenres.19363.1] [Medline: 37593567]
54. Nielsen J. Usability Engineering. New York, NY: Morgan Kaufmann; 1994.
55. Open Broadcast Software (OBS). OBS Studio. URL: https://obsproject.com/ [accessed 2023-02-28]
56. NVivo. QSR International. URL: https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home [accessed 2024-04-29]
57. Hirschey J, Bane S, Mansour M, Sperber J, Agboola S, Kvedar J, et al. Evaluating the usability and usefulness of a mobile app for atrial fibrillation using qualitative methods: exploratory pilot study. JMIR Hum Factors 2018 Mar 15;5(1):e13 [FREE Full text] [doi: 10.2196/humanfactors.8004] [Medline: 29549073]
58. Kim H, Lee SH, Cho NB, You H, Choi T, Kim J. User-dependent usability and feasibility of a swallowing training mHealth app for older adults: mixed methods pilot study. JMIR Mhealth Uhealth 2020 Jul 27;8(7):e19585 [FREE Full text] [doi: 10.2196/19585] [Medline: 32663161]
59. Zhou L, Bao J, Setiawan IM, Saptono A, Parmanto B. The mHealth app usability questionnaire (MAUQ): development and validation study. JMIR Mhealth Uhealth 2019 Apr 11;7(4):e11500 [FREE Full text] [doi: 10.2196/11500] [Medline: 30973342]
60. Godin G, Shephard RJ. Godin leisure-time exercise questionnaire. Med Sci Sports Exerc 1997;29(6):S36-S38. [doi: 10.1097/00005768-199706001-00009]

XSL•FO
RenderX

61. Amireault S, Godin G. The Godin-Shephard leisure-time physical activity questionnaire: validity evidence supporting its use for classifying healthy adults into active and insufficiently active categories. Percept Mot Skills 2015 Apr;120(2):604-622. [doi: 10.2466/03.27.PMS.120v19x7] [Medline: 25799030]
62. Namey E, Guest G, McKenna K, Chen M. Evaluating bang for the buck: a cost-effectiveness comparison between individual interviews and focus groups based on thematic saturation levels. Am J Eval 2016 Jul 09;37(3):425-440. [doi: 10.1177/1098214016630406]
63. Shneiderman B, Plaisant C, Cohen M, Jacobs S. Designing the User Interface: Strategies for Effective Human-Computer Interaction. London, UK: Pearson; 2013.
64. Mustafa AS, Ali N, Dhillon JS, Alkawsi G, Baashar Y. User engagement and abandonment of mHealth: a cross-sectional survey. Healthcare (Basel) 2022 Jan 24;10(2):221 [FREE Full text] [doi: 10.3390/healthcare10020221] [Medline: 35206837]
65. Dennison L, Morrison L, Conway G, Yardley L. Opportunities and challenges for smartphone applications in supporting health behavior change: qualitative study. J Med Internet Res 2013 Apr 18;15(4):e86 [FREE Full text] [doi: 10.2196/jmir.2583] [Medline: 23598614]
66. Rhodes RE, Plotnikoff RC, Courneya KS. Predicting the physical activity intention-behavior profiles of adopters and maintainers using three social cognition models. Ann Behav Med 2008 Dec 5;36(3):244-252. [doi: 10.1007/s12160-008-9071-6] [Medline: 19057976]
67. Rhodes RE, Plotnikoff RC. Understanding action control: predicting physical activity intention-behavior profiles across 6 months in a Canadian sample. Health Psychol 2006 May;25(3):292-299. [doi: 10.1037/0278-6133.25.3.292] [Medline: 16719600]
68. Sugrue E. App name guidelines and best practices. AppTweak. URL: https://www.apptweak.com/aso-blog/app-name-guidelines-and-best-practices [accessed 2024-03-25]
69. Spruijt-Metz D, Wen CK, O'Reilly G, Li M, Lee S, Emken BA, et al. Innovations in the use of interactive technology to support weight management. Curr Obes Rep 2015 Dec;4(4):510-519 [FREE Full text] [doi: 10.1007/s13679-015-0183-6] [Medline: 26364308]
70. Millenial. Merriam-Webster Dictionary. URL: https://www.merriam-webster.com/dictionary/millennial [accessed 2024-04-10]
71. 159 key fitness app stats for 2023: trends by age, market and more. Dr Muscle. URL: https://dr-muscle.com/fitness-app-stats/ [accessed 2024-02-20]
72. Alwashmi MF, Hawboldt J, Davis E, Fetters MD. The iterative convergent design for mobile health usability testing: mixed methods approach. JMIR Mhealth Uhealth 2019 Apr 26;7(4):e11656 [FREE Full text] [doi: 10.2196/11656] [Medline: 31025951]

Abbreviations

BCT: behavior change technique
CONSORT: Consolidated Standards of Reporting Trials
COREQ: Consolidated Criteria for Reporting Qualitative Research
FITT: frequency, intensity, time, and type
IDEAS: integrate, design, assess, and share
JITAI: just-in-time adaptive intervention
MAUQ: Mobile App Usability Questionnaire
mHealth: mobile health
M-PAC: multi-process action control
MVP: minimum viable product
MVPA: moderate-to-vigorous intensity physical activity
PA: physical activity
SMD: standardized mean difference

Edited by A Mavragani; submitted 15.04.24; peer-reviewed by N Admodisastro; comments to author 12.12.24; revised version received 11.01.25; accepted 28.01.25; published 23.04.25.

Please cite as:
Hollman H, Sui W, Zhang H, Rhodes RE
A Mobile Health App Informed by the Multi-Process Action Control Framework to Promote Physical Activity Among Inactive Adults: Iterative Usability Study
JMIR Form Res 2025;9:e59477
URL: https://formative.jmir.org/2025/1/e59477
doi:10.2196/59477
PMID:

XSL•FO
RenderX

©Heather Hollman, Wuyou Sui, Haowei Zhang, Ryan E Rhodes. Originally published in JMIR Formative Research (https://formative.jmir.org), 23.04.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Applying Patient and Health Professional Preferences in Co-Designing a Digital Brief Intervention to Reduce the Risk of Prescription Opioid–Related Harm Among Patients With Chronic Noncancer Pain: Qualitative Analysis

Rachel A Elphinston[1,2,3,4], PhD; Sue Pager[1,5], SpPath, MPH; Farhad Fatehi[6,7], MD, PhD; Michele Sterling[1,2], PhD; Kelly Brown[1,4,8,9], MClinPsych; Paul Gray[9,10], MBBS, PhD; Linda Hipper[4], GradDipPsych, MPH; Lauren Cahill[1]; Maisa Ziadni[11], PhD; Peter Worthy[12], PhD; Jason P Connor[9,13], PhD

[1]RECOVER Injury Research Centre, The University of Queensland, Herston, Australia
[2]National Health and Medical Research Council Centre for Research Excellence - Better Health Outcomes for Compensable Injury, The University of Queensland, Brisbane, Australia
[3]School of Psychology, The University of Queensland, Brisbane, Australia
[4]Addiction and Mental Health Services, Metro South Health Hospital and Health Service, Brisbane, Australia
[5]Health Equity & Access Unit, Metro South Hospital and Health Service, Brisbane, Australia
[6]Centre for Health Services Research, The University of Queensland, Brisbane, Australia
[7]School of Psychological Sciences, Monash University, Brisbane, Australia
[8]Gallipoli Medical Research Foundation, Greenslopes Private Hospital, Brisbane, Australia
[9]Faculty of Medicine, The University of Queensland, Brisbane, Australia
[10]Tess Cramond Pain and Research Centre, Metro North Hospital and Health Service, Brisbane, Australia
[11]School of Medicine, Department of Anesthesiology, Perioperative, and Pain Medicine, Stanford University, Stanford, CA, United States
[12]School of Health and Rehabilitation Sciences, The University of Queensland, Brisbane, Australia
[13]National Centre for Youth Substance Use Research, The University of Queensland, Brisbane, Australia

Corresponding Author:
Rachel A Elphinston, PhD
RECOVER Injury Research Centre
The University of Queensland
Level 7
296 Herston Road
Herston, 4006
Australia
Phone: 61 412662084
Email: rachelel@uq.edu.au

Abstract

Background: Few personalized behavioral treatments are available to reduce the risk of prescription opioid–related harm among patients with chronic noncancer pain.

Objective: We aimed to report on the second phase of the co-design of a digital brief intervention (BI) based on patient and health professional preferences.

Methods: Eligible patients with chronic noncancer pain (n=18; 10 women; mean age 49.5, SD 6.91 y) from public hospital waitlists and health professionals (n=5; 2 women; mean age 40.2, SD 5.97 y) from pain and addiction clinics completed semistructured telephone interviews or participated in focus groups exploring BI preferences, needs, and considerations for implementation. Grounded theory was used to thematically analyze the data.

Results: We identified 5 themes related to intervention content from patient reports: relevance of the biopsychosocial model and need for improved awareness and pain psychology education; nonpharmacological strategies and flexibility when applying coping skills training; opioid use reflection and education, with personalized medication and tapering plans; holistic and patient-inclusive assessment measures and feedback; and inclusion of holistic goals targeting comfort and happiness. Five themes

XSL•FO
RenderX

related to the process and guiding principles were identified: therapist guided; engaging features; compassionate, responsive, person-centered care; a digital solution is exciting, maximizing reach; and educate and normalize system and policy challenges. Finally, 5 themes were reflected in the health professionals' reports: digital health use is rare but desired; digital health is useful for patient monitoring and accessing support; patient motivation is important; a digital BI app is likely beneficial and at multiple care points; and safe medication use and managing pain goals. The reported barriers from health professionals were intervention intensity, potential costs, and patient responsiveness; factors facilitating the implementation were the alignment of digital BIs with clinical models, a stepped-care approach, and feedback.

Conclusions: This co-design study identified key content areas, guiding principles, enabling factors, and barriers from both patients and health professionals to guide the development of digital BIs. The knowledge gathered should inform future iterations of co-designing digital BIs for the population most at risk of the harmful effects of opioid medications.

(JMIR Form Res 2025;9:e57212) doi:10.2196/57212

KEYWORDS

chronic noncancer pain; CNCP; prescription opioid use; brief intervention; brief psychological intervention; co-design; patient partners; qualitative research; digital health

Introduction

Clinical guidelines recommend behavioral and nonpharmacological treatments and interdisciplinary rehabilitation for the treatment of chronic noncancer pain (CNCP) (eg, Dowell et al [1]). However, there is a substantial gap between research evidence and clinical practice, referred to as one of the "valleys of death" [2] in the management of CNCP. Barriers to patients accessing best practice, multidisciplinary, specialist pain care include long waiting times, geography, high demand for treatment, treatment intensity, lack of trained behavioral therapists, and cost [3-7]. Psychological treatments rarely reach patients with CNCP. Our research on patients' lived experience has found that most patients with CNCP are not offered psychological treatment (eg, cognitive behavioral therapy) and are unaware that it is recommended as an effective first line treatment for pain [8]. These experiences are supported by national research, which shows that 87% of Australians with CNCP had not seen a psychologist for CNCP and almost 60% were not aware of the role of psychologists in pain care [9]. This is not unique to Australia. Patients around the world are not offered or referred to nonpharmacological interventions, particularly when receiving opioid therapy [4,10]. Increasingly high demand for psychologists and specialist pain clinicians [11,12] further complicates research translation.

Digital health interventions (DHIs) for pain have been highlighted as a potential innovative solution to the treatment accessibility problem, the growing burden of CNCP, and global health threats [13-15]. DHIs for pain provide potential advantages to health systems, providers, and patients such as improved accessibility, efficiency, and cost-effectiveness, but DHIs are affected by similar knowledge-to-practice gaps with added challenges in data privacy, security, and formulating regulatory guidelines and approvals. Patient access to DHIs outside of research studies is limited [16], reducing their real-world impact. Few DHIs that are publicly available have received empirical evaluation [16,17]. Bridging these gaps in traditional and digital pain care is a priority.

A lack of patient and health professional partnering in research has been cited as a central reason for research failing to translate into clinical practice [18-21]. Co-designing new DHIs has been highlighted as a potential solution to the translational challenge. Co-design involves meaningful stakeholder (eg, patients and health professionals) involvement in the design, implementation, and translation of research [22]. Engaging patient partners alongside health professionals in the co-design of DHIs allows for improved tailoring to individual preferences and needs, which has the potential to increase treatment effectiveness, acceptance, and adoption in practice.

Patients and health professionals have rarely been involved in the co-design of DHIs for CNCP since it was recommended more than a decade ago [23]. Most DHIs are developed by the for-profit software industry [24] and lack a theoretical framework [17]. Only 19% of the studies involved end users in development, and this is usually done in an ad hoc manner [25]. Involvement of health professionals in designing DHIs varies considerably (8.2% and 30.6%) [17,25]; few enable communication with health professionals or clinician access to patient data [25,26]; and almost none (<3%) involve both patients and health professionals [25].

There is an opportunity to optimize the research to practice nexus using co-design methods in all phases of developing new, brief DHIs. This study is part of a larger program of work to co-design, co-develop, and evaluate the feasibility of a digital brief intervention (BI) for patients with CNCP to reduce prescription opioid–related harm. Digital BIs provide a potentially efficient and scalable psychological and behavioral treatment option that facilitates rapid access to pain care [13,27]. In the first phase of the digital BI co-design, we examined the patients' lived experiences of CNCP management with a particular focus on opioid therapy in individual interviews [8]. In this second phase, we advanced this body of work by examining the needs and preferences of patients and health professionals as well as implementation barriers and enablers. The knowledge gathered will continue to inform future steps in the co-design of digital BIs for the population most at risk of the harmful effects of opioid medications.

XSL•FO
RenderX

Methods

Participants and Recruitment

Eligible participants were patients diagnosed with a CNCP condition who were purposively sampled to ensure adequate representation of both sexes, with culturally and linguistically diverse backgrounds, and current and past use of opioid therapy. A full summary of participant recruitment is provided in a previous study [8]. In short, the patient eligibility criteria included the following: those aged between 18 and 70 years, experiencing clinical levels of CNCP (scoring ≥4 out of 10 on the Pain Numerical Rating Scale on average over the past week), and seeking treatment from public health specialist addiction or pain services. The patient exclusion criteria were as follows: high levels of distress (based on patient scores ≥13 on the Kessler Psychological Distress Scale [28] and follow-up telephone risk assessment or clinical judgment from the addiction specialist), non-English speaking, a history of recent injecting drug use, or currently tested positive with SARS-CoV-2.

Patients were recruited through a tertiary hospital specialist pain clinic or community addiction specialist service. Patients were either engaged by mailing an information pack inviting those on the waitlist to participate (pain clinic) or invited by the medical specialist and a member of the research team (RAE) during their on-site appointment (addiction service). The final sample size was determined by thematic saturation using the analysis approach by Guest et al [29] applied during the data collection and inductive thematic analysis phase [8].

Health professionals were also invited to participate in one-on-one interviews or focus groups to assess their needs and preferences for the digital BI. Health professionals recruited through the specialist pain and addiction clinics as well as through the research and clinical networks of the research team were eligible to participate if they were working in chronic pain management (public or private sector).

Design

A co-design approach [9] was used, applying the framework by Sanders and Stappers [30] and updated by Noorbergen et al [31]. It outlines 6 iterative co-design phases: predesign, generative, prototyping, evaluative, implementation, and postdesign. Figure 1 presents an overview. This study reports on the generative phase of the iterative co-design development process, which includes exploring barriers, enablers, and suggestions for the digital BI from both the patient and health professional perspectives. Patient perspectives were collected from individual interviews (n=18), and 2 focus groups of patients (n=7) invited to attend after the individual interviews; health professional perspectives (n=5) were collected from two focus groups. All individual patient interviews were conducted via telephone by RAE, a clinical psychologist; KB, a psychologist; and SP, all with expertise in chronic pain or qualitative research. Field notes were recorded during the sessions.

The project was led by the study's chief investigator (RAE), who is a clinical psychologist–researcher. The multidisciplinary project team consisted of health care researchers (ie, psychologists, physiotherapists, pain medicine specialists, and addiction specialists); digital development partners (ie, software developers and designers); and patient partners (ie, individuals with a lived experience of CNCP including those using prescription opioid medications). The team met weekly (sometimes more often) during the development process.

Figure 1. Project co-design stages. The generative phase (phase 2) reported in this study is denoted by the items highlighted in green.

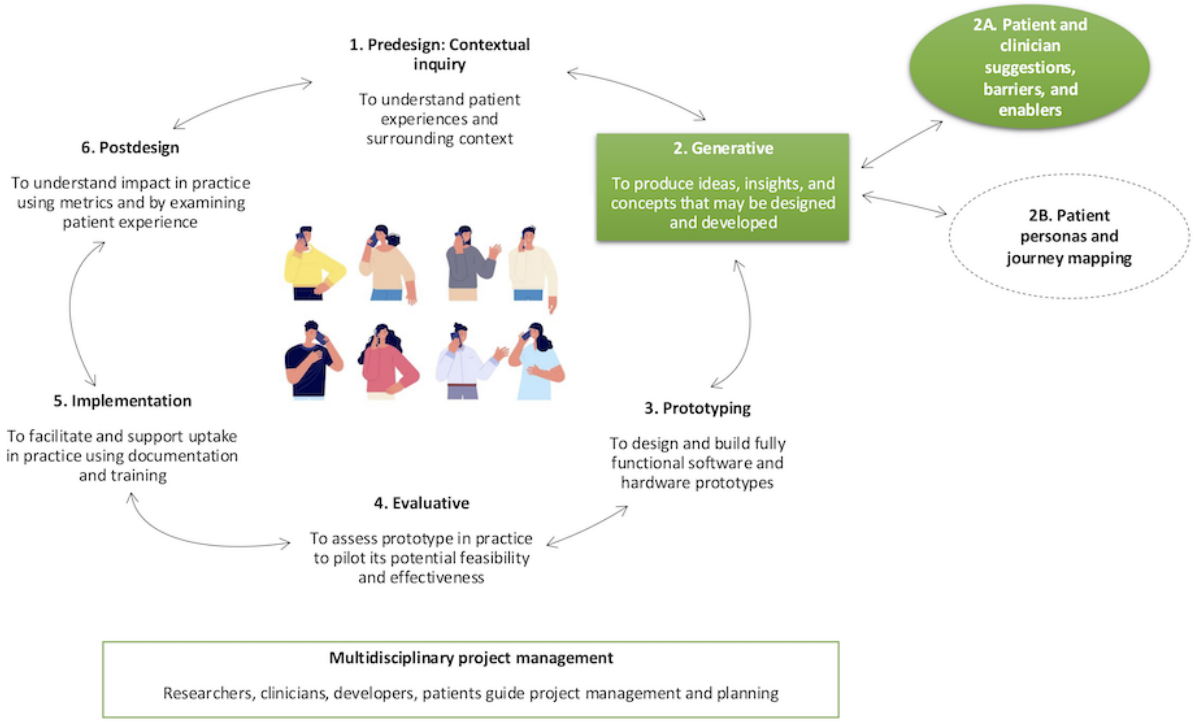

XSL•FO
RenderX

Measures

Web-Based Questionnaires

Patients were asked to provide their demographic details, information about their pain condition, current medications, and history of mental health and substance use disorders. Validated measures included the Brief Pain Inventory [32,33]; Kessler Psychological Distress Scale [28]; 21-item Depression Anxiety Stress Scale (DASS-21) [34]; and Current Opioid Misuse Measure (COMM) [35]. A full description of the measures can be found in the study by Elphinston et al [8].

Health professionals completed a web-based survey containing questions about their demographics, clinical background and experience, workplace setting, training and professional development in pain, use of DHIs, and types of telehealth use in clinical practice.

Patient Semistructured Interview and Focus Group Questions

As part of a larger set of interview questions, patients were asked about how they visualize their pain management in the future and their opioid use and how they could assist others in using opioids more safely. Multimedia Appendix 1 provides a summary of the questions. Patients who then expressed further interest in co-designing the BI were invited to participate in the focus groups, which involved discussing their impressions of the digital BI and several activities exploring its components. To set the stage, participants completed an exercise involving sharing their pain story, including the number and type of pain management strategies used. Participants then completed activities such as questionnaire assessment and feedback, introduction to the biopsychosocial model of pain, opioid use education, personality-targeted coping skills training, and motivational interviewing and action planning. Patients also explored the problem through a “peel the onion” activity and proposed solution requirements and were provided an example of a BI for smoking cessation. The aim of these activities was to explore and confirm patient preferences and needs, drawing on the results of the individual interviews to guide discussion. The role of technology in delivering BIs was explored including how to best meet the patient needs of this population and the implementation requirements. Finally, patients completed a value sliders exercise as a group, rating each BI component derived from research on BI treatment models and patient experience, from most important to least important.

Health Professional Semistructured Focus Group Questions

Health professionals were first introduced to the background and rationale for the digital BI, the main components, and the themes identified in patient lived experiences reported in a previous study [8]. They were then asked about their impressions of the themes and the digital BI. Multimedia Appendix 1 provides a summary of the questions.

Procedure

Interviews and focus groups were completed from June to December 2020. Patients completed an initial telephone screening interview to determine their eligibility for individual interviews and health professionals expressing interest via email were invited to 1 of 2 focus groups. Eligible patients and health professionals completed a web-based questionnaire before the interview for descriptive purposes and to provide background to guide the interview. Patient focus groups were then conducted in person on-site at the clinic with one patient participating via Zoom (Zoom Video Communications, Inc). Health professional focus groups were conducted via Zoom. Interviews and focus groups were audio-recorded and professionally transcribed verbatim.

Data Analysis

An iterative approach to data analysis was used. Transcripts from the patient interviews and focus groups were thematically analyzed separately from health professional focus group data using grounded theory [36]. Two coders (RAE and SP) coded line-by-line and then grouped the codes into subthemes and themes. Theme refinement was facilitated by a third clinician-researcher (KB), who assisted in integrating interviewer field notes to aid interpretation and reduce coder bias. All researchers were asked to review the themes and subthemes. The audio recordings were referred to as needed to examine patient quotes and clarify content.

Ethical Considerations

Ethics approval was obtained from the Metro South Human Research Ethics Committee (HREC/2020/QMS/60695). All participants provided written informed consent. Study data were deidentified for analysis, and any identifying information was removed before publication including patient numbers to ensure complete anonymity. All participants were offered an Aus $50 (US $33) gift voucher for each session attended to thank them for their time.

Results

Characteristics of the Patient and Health Professional Samples

Patient characteristics (n=18; 55% women; mean age 49.5, SD 6.91 y) are presented in a previous study [8]. We provide an overview in Table 1; for a detailed summary of each individual, please refer to a previous report [8]. In the sample of patients interviewed, half (9/18, 50%) met the threshold for current unsafe opioid misuse on a validated psychometric scale (COMM [35]). Most health professionals (n=5; 40% women; mean age 40.2, SD 5.97 y) were psychologists working in both the private and public sectors, with an average of 9.8 (SD 5.97) years of clinical experience (Table 2).

XSL•FO
RenderX

Table 1. Patient (n=18) demographics and clinical characteristics.

Characteristics	Values
Gender, n (%)	
Female	10 (55)
Male	8 (45)
Ethnicity, n (%)	
Australian	14 (77.9)
British (English, Welsh, or Irish)	2 (11.1)
New Zealander	1 (5.5)
Greek	1 (5.5)
Pain location, n (%)	
Head or face	4 (9.5)
Neck	6 (14.3)
Shoulder or upper limbs	6 (14.3)
Back, spine, or sacrum	12 (28.6)
Lower limbs	4 (9.5)
Whole body	5 (11.9)
Abdomen, pelvis, or groin	5 (11.9)
Age (y), mean (SD; minimum-maximum)	49 (11.5; 25-62)
Pain duration (y), mean (SD; minimum-maximum)	11.6 (11.5; 0.5-42)
Pain intensity (past week), mean (SD; minimum-maximum)	5.8 (1.5; 2.3-8.0)
Pain interference (past week), mean (SD; minimum-maximum)	5.8 (1.9; 2.7-9.1)
Kessler Psychological Distress Scale, mean (SD; minimum-maximum)	9.7 (4.8; 4-18)
Depression, anxiety, and stress[a], mean (SD; minimum-maximum)	
Depression	10.8 (9.1; 0-30)
Anxiety	9.2 (8.9; 0-34)
Stress	13.9 (9.1; 0-32)
Opioid use duration (y; n=16[b]), mean (SD; minimum-maximum)	8.0 (8.7; 0.6-26)
OME[c] (n=13[b,d]), mean (SD; minimum-maximum)	114.2 (166.8; 1.2-480)
COMM[e] (n=16[b,f]), mean (SD; minimum-maximum)	11.7 (7.89; 1-27)

[a]DASS-21: 21-item Depression Anxiety Stress Scales.
[b]2 participants were not currently using opioids.
[c]OME: oral morphine equivalent.
[d]3 participants currently using methadone.
[e]COMM: Current Opioid Misuse Measure.
[f]8 participants scored ≥9 indicating risk of opioid misuse [35].

XSL•FO
RenderX

Table 2. Health professional demographic and background information.

	Sex	Age (y)	Profession	Years of practice	Work sector
HP[a]1	Male	44	Psychologist	17	Public and private
HP2	Female	29	Psychologist	3	Public
HP3	Female	39	Psychologist	15	Public
HP4	Male	42	Nurse	8	Public
HP5	Male	47	Psychologist	6	Private

[a]HP: health professional.

Patient Preferences

Guiding content, principles, and processes related to the proposed digital BI emerged from the individual patient interviews and were further refined in the patient focus groups. There were 5 themes related to the content of the BI (themes 1-5) and 5 related to the process and guiding principles (themes 6-10). Table 3 provides a summary of themes and exemplar quotes. Patient numbers were removed before publication to ensure anonymity (PID_P refers to patients from the pain clinic and PID_A refers to patients from addiction services).

XSL•FO
RenderX

Table 3. Themes identified from the individual patient interviews and focus groups

Themes	Brief description	Key examples from patient interviews
Intervention Content		
1. The relevance of the biopsychosocial model and need for improved awareness and pain psychology education	The biopsychosocial model provides a relevant framework for understanding pain processes and improved awareness and pain psychology are needed.	• "It [biopsychosocial model] does make sense because I never connected it until I actually spoke to you on the phone. Because I had a good think about it and when there were family issues or when we moved house or anything really stressing my back pain would flare up. and it was almost like my pain levels, like my back was dealing with the stress. I had no control over that, I was just feeling what was going on and the pain levels go up... it's interesting, I didn't know that stress impacted and now when I think way back to when my back pain really kicked in it was a really stressing time." (PID_P). • "I can definitely relate to what you were talking about earlier that even when you've recovered your body still holds onto the pain because I feel that that's very much what's happening to me... long term pain puts stress on your nervous system and you can hold your pain because I've noticed lately that ever since I've had this when I get shoulder pain or something like that even just muscular pain it stays for a very long time because my body is in this constant state of alertness that it just holds pain. So, if I get a sore neck or something like that instead of it going in two days which it would 5 years ago, it now stays for weeks or even months because of my like nervous system" (PID_P).
2. Nonpharmacological treatments and flexibility when applying coping skills training are needed	Pain management strategies not just strategies to manage opioid use are required and should be flexibly applied.	• "I'd rather just not use the opioids at all and get help in other ways like ether it be therapy...I think Drs should be careful in prescribing them...and look at other things they could do for you" (PID_P). • "But if you're suffering obviously mentally, it kind of puts your body through a bit of stress, stressful situation that can cause like tension or psychologically it will inflict pain...So who knows, maybe a psychologist or whatever that was in pain, to help them mentally think about their pain in a different way, that would probably help...(PID_A).Yeah because you like to be told what to do and I like to choose what I do. But that's different personalities. I'm the boss of me, no body be telling me what to do, you know" (PID_P).
3. Opioid use reflection and education, with personalized medication and tapering plans	Opportunities to reflect on if opioids are working for them and providing the necessary education as well as medication plans that are individualized to the person's circumstances.	• "Explore potential problem in a non-judgemental way... I think finding out what the reason is and why it is important... But I think it's very important for the doctors, as the ones prescribing the medication to actually sit there and go, all right, here's the side effects" (PID_A). • "Refer to someone who can help taper in a personalized way... clear taper program with options to increase dose again if needed ... More regular monitoring by GP e.g. every month instead of 6 months" (PID_P). • "They want to come off them, if they think that they no longer have a physical use for them and might only be psychological they would need something to help them get through withdrawal symptoms" (PID_A). • "I don't feel the GP's support is enough...GP never asks me to keep a diary...they are not actually checking on if you have taking too many" (PID_P).
4. Holistic and patient-inclusive assessment measures and feedback	Assessment measures & feedback are holistic and patient inclusive.	• "I like how it [COMM] asks a lot of questions about how you feel and relationships like arguments... because I think pain effects so many things. It effects your relationships. it affects how you communicate with people. If you're in a lot of pain you can't think clearly. You are and you get angry at anything you get ticked off or annoyed or frustrated. So I think a lot of these comments if you know people are in pain they could relate to this very [sic] a lot of these questions...very relevant." (PID_P).

XSL•FO
RenderX

Themes	Brief description	Key examples from patient interviews
5. Inclusion of holistic goals targeting comfort and happiness	Broader, personal goals that target the whole person should be included, with a specific goal to increase happiness and level of comfort rather than a focus on being pain-free.	• "[Success would look like] A happy person... a well-adjusted person. A person who's not in constant pain. The list goes on... I would be happy with it making me comfortable" (PID_P). • "[Facilitator: not just medication focused goals?] Because I think it all comes to, its holistic, it's all encompassing, every part of your life needs to be looked at [– group agreed]" (PID_P).
Intervention Process and Guiding Principles		
6. Therapist-guided when needed and accountable	Guidance from health providers preferred rather than self-management approach to facilitate patient motivation and ensure accountability.	• "Could lose motivation very quickly if doing it yourself" (PID_P). • "Maybe having a drug counsellor or someone they can speak to on phone or online when they feel they are struggling with their medication" (PID_P). • "Psychologist to your house and the house visits for people that can't get out. That would be ideal, you know... Video for a psychologist would be good... I just get all tense to probably just when I'm in front of someone talking to, but that's just me" (PID_A).
7. Engaging easy to use features including peer support	A combination of engaging features (visual, verbal) including peer support to share similar experiences and ideas with and provide moral support when needed.	• "Even just hearing everyone's story, just feel not as alone" (PID_P). • "Sponsor like AA where if they want to take an extra tablet, they talk you out of it" (PID_P). • "Someone I could talk to, relate to, be there, just for the psychological side of it. People to talk to... Moral support was a big part of it. I reckon if I had someone to talk to a lot, when I was going through it wouldn't have been so bad. Just having friends that are going through the same thing" (PID_A).
8. Compassionate, responsive and person-centered care	Treat the person not the number, providing nonjudgmental and care personalized and responsive to the patient's needs.	• "I would have appreciated at the start not being told it was all in head and having just opioids thrown at me" (PID_P). • "Support, having someone see them as a person and not see them as a number" (PID_A). • "It's not a one size shoe fit all [approach to pain management]" (PID_A). • "With our other doctor, at least we could drop in and say, oh, my God, it's freaking hurting today" (PID_P).
9. Digital solution was new and exciting, maximizing reach	The new and exciting option of digital brief interventions maximizes reach in a variety of settings (eg, primary care; specialist wait lists).	• "If there was sort of like a program available and then that was part of it, so if I ever went to the doctor and said I want to reduce my tramadol slowly over time he could say "well, here's the website" and he'll register me for it and sort of he communicates with that website as well so when I go back to him he can see that I've done all the criteria and then he can reduce my medication" (PID_P). • "It would be really really good to be able to have something where you could be like, what can I do now? That would be really good because then you feel someone is on your side, someone is trying to help you plan. You're not just being left" (PID_A).
10. Educate and normalize challenges navigating the health system and policy	Education about policy changes and normalizing the challenges of navigating the health care system.	• "The GP put me on tramadol in 2007 and when the government brought that thing on the first of June, I had a new doctor and all of a sudden it's like oh no you can't be on this, you've got to be off, and I'd actually just switched to a new GP and he thought I was doctor shopping. He was quite judgmental" (PID_P).

Intervention Content

Theme 1: The Relevance of the Biopsychosocial Model and Need for Improved Awareness and Pain Psychology Education

Participants agreed that the biopsychosocial model of pain provides a relevant framework for explaining the unique individual experience of pain and how it can change over time. The patient experience interview was the first time when 3 of the 4 focus group participants were introduced to this model [8].

Most participants reported a number (>10 types) of varied biological and psychological influences of pain. It was important for participants that the origins of pain were explored (eg, surgery, injury, gradual or unknown, and stress). For another participant (PID_P), the continuation of pain despite recovery from physical injury resonated. One participant (PID_P) reported that they had moved toward acceptance that pain was mostly perpetuated by psychological factors including stress and tension

XSL•FO
RenderX

as they had eliminated all other reasons for their pain; they were now using mindfulness, breathing, and monitoring their own posture to relax and manage pain after having ceased the use of opioids in the past 8 weeks (a significant change from the individual interview [8]).

Participants also reported that the concept of a person's personality characteristics contributing to their pain and opioid use was considered important (eg, "Are you a 100-mile person or are you slow and steady, you know, it misses that completely" [PID_P]). Most participants reported that they identified with both depressive- and anxiety-prone personality characteristics as contributing to their pain and opioid medication use.

Social factors were considered less influential. One participant described how the COVID-19 pandemic had impacted their ability to go to church and that church often resulted in feeling better in terms of her overall well-being and pain—"I feel better when I come out of church then I did when I go in, and you know pain wise and all over wise" [PID_P]. Another discussed the role of caregiving responsibilities on her pain management experience—"[social factor] well I can't try medical marijuana even though a doctor offered it to me because I said to him I have to roll out of bed at 2 in the morning if her [elder mother who they care for] alarm goes off so I can't be taking that" [PID_P]. This facilitated patient understanding of how individualized social factors can influence pain experience.

Participants reported that greater awareness and educating patients about the biopsychosocial model of pain is needed. One participant (PID_P) described how it was helpful to understand how pain impacts the nervous system over time and how the body "holds" pain, especially because their injury had healed, and they now wanted to understand more about the influence of mental health. Another participant was open to discussing the role of mental health in pain with psychologists or psychiatrists because their injury had gone, and pain was persisting (PID_P).

Theme 2: Nonpharmacological Strategies and Flexibility When Applying Coping Skills Training

Participants confirmed a need for nonpharmacological strategies other than the use of opioids such as relaxation, exercise, engaging in pleasant activities, and pacing. PID_P said, "there's got to be something out there that covers everything but doesn't actually interfere with life itself, because you know the opioids make you stupid." This was reported by some participants as important to "take back control" (PID_P) from doctors who "should stop giving out opioids and recommend other non-pharmacological alternatives" (PID_P). One participant (PID_P) wished not to be opioid dependent, as they feel reliant on pain medication:

> *If [the BPI could] teach you not to be drug dependent. Because that's pretty much what I rely on. If that just stopped, I don't know what would happen.*

There were a range of coping strategies that participants identified could target depressive-proneness and pain (eg, listening to music, deep breathing, and doing enjoyable things). Coping strategies targeting anxiety-proneness included relaxation and releasing tension in the body (eg, gentle stretching or yoga) and mindfulness and meditation, while additional strategies that could target both personality styles were discussed (eg, going for a walk or talking to someone about how you feel). Multimedia Appendix 2 provides a snapshot of this exercise.

It was important that there was flexibility when applying a personality-targeted approach to coping skills training; some participants liked the targeted approach and wanted to be directed to the strategies corresponding to their personality style. Others preferred to be offered a full range of options and an opportunity to select the strategies themselves appealed to them:

> *Yeah because you like to be told what to do and I like to choose what I do. But that's different personalities. I'm the boss of me, no body be telling me what to do, you know.* [PID_P]

> *I don't know if the selection will work because if you're feeling anxious and stressed, being told to pick one of them is going to be overwhelming itself. You know if you're lonely and depressed, you might not necessarily want to choose you might just want someone to say right you're doing this. Like a decision might actually contribute to this. [all agreed with this]* [PID_P]

Theme 3: Opioid Use Reflection and Education, With Personalized Medication and Tapering Plans

Some participants reported that opportunities to reflect on their response to opioid therapy (benefits vs harms) would be important and could offer an avenue to provide opioid use education. Patients indicated that more knowledge of the potential harms of opioids was needed. One participant (PID_P) described how education that opioids can increase pain sensitivity and that tapering opioids do not necessarily result in increased pain was helpful when reducing opioid use:

> *I never attempted to come off it because I just thought what am I going to do with my pain, it's just going to get worse and he said no, as you reduce your pain sensitivity will reduce and he was right...[knowing that] it took the fear away from tapering off.*

Related to theme 7, the use of trivia and quizzes as a regularly updated semicompetitive activity was suggested to engage patients in education and motivation to continue with the digital BI.

Some participants suggested that education about opioid addiction and dependency would be important as well as taking into consideration people with pain who may have had a history of substance use problems. One participant said, "I'm clean. and even when I broke my shoulder and I had to take fentanyl in the hospital I phoned my sponsor first to make sure that it was okay for me to have it" [PID_P].

There was a specific need for opioid medication management plans with a focus on medication-related goals that are achievable:

> *Reduce your dose by I don't know 2mg, that's achievable... just little things you know make sure you go for a 10min walk once a week...10mins once a week should be achievable and not overdo it. You*

XSL•FO
RenderX

> *know just little things, it's like the compound effect book... you make such a small change that you don't really notice it but its compounded over 12 months... small achievable goals that compound.* [PID_P]

For some people, goals were related to a reduction in medication dose, while for others, using medication more safely. Opportunities to revise goals over time (eg, weekly) was identified as essential. Participants discussed a tailored, stepped, and "SMART" (Specific, Measurable, Achievable, Relevant, and Time-Bound) approach to goal setting. Monitoring of personalized medication plans was reported by some participants as necessary to checking patient response to treatment and provide an opportunity to identify any concerns (eg, withdrawal). One participant (PID_P) suggested a patient medication diary could assist in the monitoring process. Collaboration among health professionals and with patients was highlighted to ensure that everyone is aware of the patient's individual medication plan and is working together to support the patient' goals.

Theme 4: Holistic and Patient-Inclusive Assessment Measures and Feedback

Participants reported that assessment measures could focus on pain, its impact, medication use, and personality styles. It was important to all that measures are inclusive in terms of language and are nonjudgmental. In presenting the assessment measures to participants that they completed before the study, participants reported that they liked how items on some measures asked about how you feel (eg, capturing feelings of anger and frustration) and the impact of pain on relationships (eg, arguments assessed by the COMM).

Participants reported that they did not think the measures presented captured the pain journey (in terms of historical factors) as well as personality factors, although the research team communicated that this was probably a function of only providing participants with opioid-related measures. Focusing on opioid use in the past 30 days (as noted in the COMM) was viewed as restrictive:

> *This is all in the past 30 days, yet it says holistic there [on whiteboard], peoples pain experience, you know a holistic pain experience isn't just in the last 30 days it's in the last years...people have ups and downs, you might have had a really great 30 days. It doesn't really represent somebody's holistic pain experience.* [PID_P]

The items were reported to result in some participants feeing judged (eg, taking opioids illegally or outside of their prescription and counting pills) and that they were a "drug addict" [PID_P]. Participants discussed a sense of relating the pill counting questions more to anxiety than misuse:

> *[SOAPP-R] counting pills, only reason I would count them is because I'm anxious about not having enough, not because...its anxiety related not accumulative.* [PID_P]

It was suggested that the wording of some of the items could be modified to be less stigmatizing or measures could be introduced to participants and sensitivities flagged. Asking participants how they felt about questions about opioid use and giving them a free text box response option could also be beneficial (PID_P).

Theme 5: Inclusion of Holistic Goals Targeting Comfort and Happiness

Participants suggested that goals should target the whole person in addition to focusing on medication action plans. One participant (PID_P) discussed that they felt able to focus on improving physical fitness in the context of having an achievable plan already in place for reducing her tramadol use and managing withdrawal symptoms. It was also discussed that goals should empower patients, facilitating a sense of control and autonomy, consistent with a biopsychosocial mindset (theme 4; eg, success=confidence+mood boost=reduced pain):

> *Once you achieve something and you've got that little bit of success, that endorphin release also helps with your pain. [PID_P said "yeah yeah"] So, you've achieved something and that feel good feeling is yeah.* [PID_P]

Specifically, the goal to increase happiness and level of comfort rather than a focus on being pain-free was proposed as a possible goal of the intervention identified during the "peeling the onion" exercise (to gradually work toward deeper understanding of the problem to be solved). Participants suggested that focusing on making your life better and to increase happiness may assist in helping people to lead a "normal" lifestyle. Interestingly, participants did not focus on the goal of being pain-free but rather on being "comfortable." Most had moved toward acceptance of pain and felt that they may never be pain-free. Multimedia Appendix 3 provides a summary of the exercise.

Intervention Process and Guiding Principles

Theme 6: Therapist Guided When Needed and Accountable

Participants highlighted opportunities for connection and support from health professionals and peers were important. There was a need for "good advice" when called on and someone to talk to on bad pain days. Resources (eg, website links) on topics of interest were also viewed as helpful. A chat box feature was suggested as one way to connect with health professionals. Immediate responses from health professionals and options from where to seek further help were also seen as important (eg, drug and alcohol phone service). Participants agreed that the BI needs to be therapist guided rather than a self-management approach. It was suggested that a therapist-guided approach would provide both extra support and also accountability: "You'd want it to feel like there is that actual support" (PID_P); "I guess if it's not being monitored then you think what's the point?" (PID_A). Participants discussed value in health professional contact at the initial onset of the BI, in completing assessments and in the pain journey discussion.

Theme 7: Engaging, Easy-to-Use Features Including Peer Support

Discussions on how to best engage patients in the digital platform included the consideration of color schemes, the use

XSL•FO
RenderX

of games and competitions like "scoreboards," and the "steps challenges," which were viewed as motivating. Reading and sharing of patient stories would help normalize the patient's experience and provide ideas for what has worked for others. PID_P described the value of Facebook support groups. Hearing others' stories were highlighted as important: "Even just hearing everyone's story, just feel not as alone" [PID_P]. There were examples of chat boxes, website links, chat rooms, and forums to facilitate this, and it was suggested that this support may be very helpful, particularly on bad pain days. One participant suggested that access to journal articles that were translated into simple summaries for patients or videos of user-friendly, up-to-date evidence-based research summaries would be helpful. It was important that the BI is simple and easy to use.

Theme 8: Compassionate, Responsive, and Person-Centered Care

Participants reported a need for compassionate and personalized care from all treating health professionals. It was important that they were seen as a person and not as a number (PID_P) and that care is tailored to each unique individual with pain. Comments about "being told pain was all in my head" [PID_P] were viewed by participants as invalidating and unhelpful to treatment discussions. Empathy was agreed as a necessary ingredient to care, as was the responsivity of health professionals. Some participants suggested that greater access to health professionals, particularly on bad pain days, via digital means could better meet their support needs (eg, "GP contactable on Facetime, he does tele[health], emails" [PID_A]).

Theme 9: The Option of Digital BIs Was New and Exciting, Maximizing Reach in a Variety of Settings

None of the participants had come across a digital BI for CNCP in their search for effective treatment solutions. They said they were excited for a new option that could potentially help themselves and other patients. Participants discussed hope that their experiences would inform the design and development of the BI that could offer more support and earlier intervention to patients. A digital platform was seen as an ideal way to deliver the intervention for maximum reach and access in a way tailored to the patient's needs. One patient had experience using apps (eg, Smiling Mind), which was reported to give them more confidence in the potential use and acceptability of a digital BI.

Participants suggested that disseminating or promoting the BI would be suited to a primary care setting, with the use of flyers in the waiting room or to those on specialist waitlists. Engaging potential participants in the intervention could also include the use of social media, Google advertising, and BI advocates and word of mouth. Since it was discussed that everyone is generally using their phone on a daily basis, there was agreement that web-based promotion may be more effective than hardcopy leaflets. Health professional and particularly general practitioner education and awareness of the BI program was seen as key.

Theme 10: Educate and Normalize Challenges Navigating the Health System and Policy

Participants discussed difficulties navigating the health system and indicated value in providing education on the rationale for policies and legal restrictions that may impact their opioid use and pain management. It was suggested that greater consideration of the social environment and medicolegal background may serve the function of normalizing challenging experiences (eg, changes in opioid use), acknowledging differences in the treatment they received in the past (20 years ago) compared to current best practices, and reducing perceived judgment from the care team (eg, methadone associated with feeling judged as "a drug addict").

Patient Ranking of Importance of Intervention Components

Patients ranked the importance of 7 digital BI components from most important (rating=1) to least important (rating=7):

1. Support
2. Coping strategies
3. Education
4. Assessment
5. Feedback and monitoring
6. Goals and action plans
7. Fun

Health Professional Perspectives

Five themes were identified from the data as important considerations by health professionals to the development of the digital BI and its implementation in clinical practice. Barriers and enablers were also identified.

Theme 1: Use of DHIs and Resources Within Clinical Practice Was Rare But Desired

Health professionals described that they generally do not tend to integrate apps and digital interventions into their clinical practice. One health professional said, "it would be rare that I'd use a program in the clinic with someone I'm actually treating." Instead, they were generally viewed as an adjunct to treatment. For example, health professionals would recommend web-based programs for pain or more general mental health apps. Health professionals also recommended patients' access TED Talks, podcasts, and educational videos, as well as resources they had developed themselves (eg, mindfulness audio files and sleep hygiene information). Most only recommended resources that they have evaluated themselves. Some health professionals highlighted that integration of digital health tools into their practice was practically challenging: "I've never really tried to integrate an app into treatment because it's always been a bit too challenging in terms of working out how to get the feedback and what to do with it then."

There was recognition among health professionals that it is important to point patients to digital resources that are free and easily accessible: "They are not all free, or only free for a while and then a subscription is required." It was also acknowledged that some patients may not have access to the internet and that digital resources were generally not inclusive of culturally and linguistically diverse patient populations. Some health professionals reported that they would like to use digital health resources more in their clinical practice: "Depending on the demographic as in CALD [culturally and linguistical diversity], not really many resources I can use but I don't use it as much as I would like to."

XSL•FO
RenderX

Theme 2: Digital Health Useful for Patient Monitoring of Symptoms and Access to Support on the Run

Health professionals reported that apps can be useful for patients to monitor their symptoms over time as well as track their use of other treatments (eg, cannabis). There were reports that digital health tools can increase access to support such as coping strategies and resources. One health professional said, "Can be useful to access it remotely such as if they are in a shopping center and feel anxious they can sit down and do some calming exercises." While benefits were mentioned, some health professionals reported that their clients preferred to minimize the use of digital health: "Many of my clients just prefer to keep it simple, won't even use video [during telehealth], just want a phone call consultation."

Theme 3: Patient Motivation to Use Digital Health Resources Is Important

Patient motivation to use digital health was considered important in whether these were successful in clinical practice. One health professional said, "I find that that apps are only as good as the client's motivation to use them," and another said, "Sometimes we recommend things and they don't get acted on...take a lot of prompting to use themselves." There was acknowledgment that patient motivation can be influenced by several factors (eg, how easy it is to download an app) and that some patients may not be motivated to change behavior, try something new, or use digital health resources.

Theme 4: A Digital BI App for This Complex Problem May Be Beneficial at Multiple Points in the Health Care Journey

Overall, health professionals agreed that a digital BI for the complex problems of CNCP and risk of opioid-related harm may be beneficial for patients. Health professionals reported that the digital BI could be helpful for patients to complete before referral to pain specialist services:

> *I see it as helpful before come to me [specialist pain clinic] because it will shift their readiness for change and make more open to when they come to me. Would use this as a pre-intervention.*

There were also reports that the digital BI may be "Useful to bridge gap between primary care and pain services, so not waiting for so long." One health professional said the following:

> *Good for those on the waiting list for specialist services... it might mean that by the time they get to that service they might be more ready, knowledgeable, more open to alternatives... help to set up expectations of treatment approaches.*

Others reported that it could fit within a primary care setting or even be used for the prevention of opioid use problems:

> *I think it could be useful across the board. In my clinic though, patient selection is important if you want to get best results. Could help to move people who are at the pre-contemplation stage.*

The patient's stage of change or motivation was reported as important in considering the best intervention point (linked to theme 3). There was also suggestion that the digital intervention could be delivered by psychologists working in the private sector; mental health professionals; multidisciplinary pain clinics; or, in the primary care setting, by nurses or physicians: "the GP could refer patients to or deliver the app before they prescribe."

Health professionals reported that patients would prefer to access the intervention via a mobile app or device rather than a desktop website due to convenience and portability. There was also recognition that a digital intervention would have challenges for some patients, and there was a need for therapist-guiding:

> *Both (app and website) have access issues. Would be good to have an option to do the program somewhere; in the GP office, at the MDT [multidisciplinary team] clinic while they're waiting on the wait list etc. Either way patients would need health professional support for access and troubleshooting as well as for the monitoring and goal setting.*

Theme 5: Realistic Patient Goals Focused on the Safe Use of Medications and Managing Pain

Patients' goals that are focused on safe use of opioid medications including increased willingness or openness to change unsafe use were supported. Health professionals also reported that goals related to use of alternative pain management strategies and pain acceptance may also be achievable: "Can we get to them to move a little bit on changing behavior - willingness to change opioid use and use alternative strategies." Setting realistic goals were important:

> *I'd work on positive wins towards the overall goal. So much going on for patients, other things to manage.*

Barriers to Implementation

Some health professionals reported that more than 4 hours of intervention may be required to achieve the patient's goals. Costs related to who would deliver the intervention were identified as a barrier as well as clinicians setting aside time for support (eg, viewing patient monitoring). There was also concern that some patients may be defensive about making changes to their opioid use and that some patients do not trust health professionals to talk about their opioid use.

Intervention Enablers

Health professionals viewed that the underlying philosophy of the digital BI (motivational interviewing and enhancing patient self-efficacy) was consistent with how they approach their clinical work. There was recognition that implementation of the digital BI would be most effective if considered within a stepped-care model (ie, primary care > digital BI > specialist care; related to theme 4). There were also reports that feedback to patients (about their opioid use, pain, and mood) is "always helpful," as is patients' response to this information. This information can then be used to troubleshoot or problem solve.

XSL•FO
RenderX

Discussion

Main Findings

This study aimed to understand patient and health professional preferences, including implementation considerations, for a new digital BI to reduce prescription opioid–related harm among patients with CNCP. In total, 10 themes, 5 related to intervention content and 5 related to intervention processes emerged from patient reports as important in the future digital BI co-design and co-development. From reports of experienced health professionals, we identified 5 themes, as well as several barriers (including intervention intensity, potential costs, and patient responsiveness) and implementation enablers (including consistency of theoretical underpinnings with current practice, patient feedback, and the digital BI as part of a stepped-care model). Figure 2 provides a snapshot of the themes identified from both patients and health professionals.

Overall, both patients living with CNCP and health professionals were interested in the proposed digital BI to reduce opioid-related harm. Health professionals highlighted that digital health tools may be particularly useful for patient symptom monitoring and access to coping skills support (eg, deep breathing exercise) when on the move. Patients agreed that a digital solution could expand the reach of BIs to people who need it. Certainly, there is growing interest in and evidence for the role of digital BIs in treating pain [13,37-39].

Figure 2. Snapshot overview of themes.

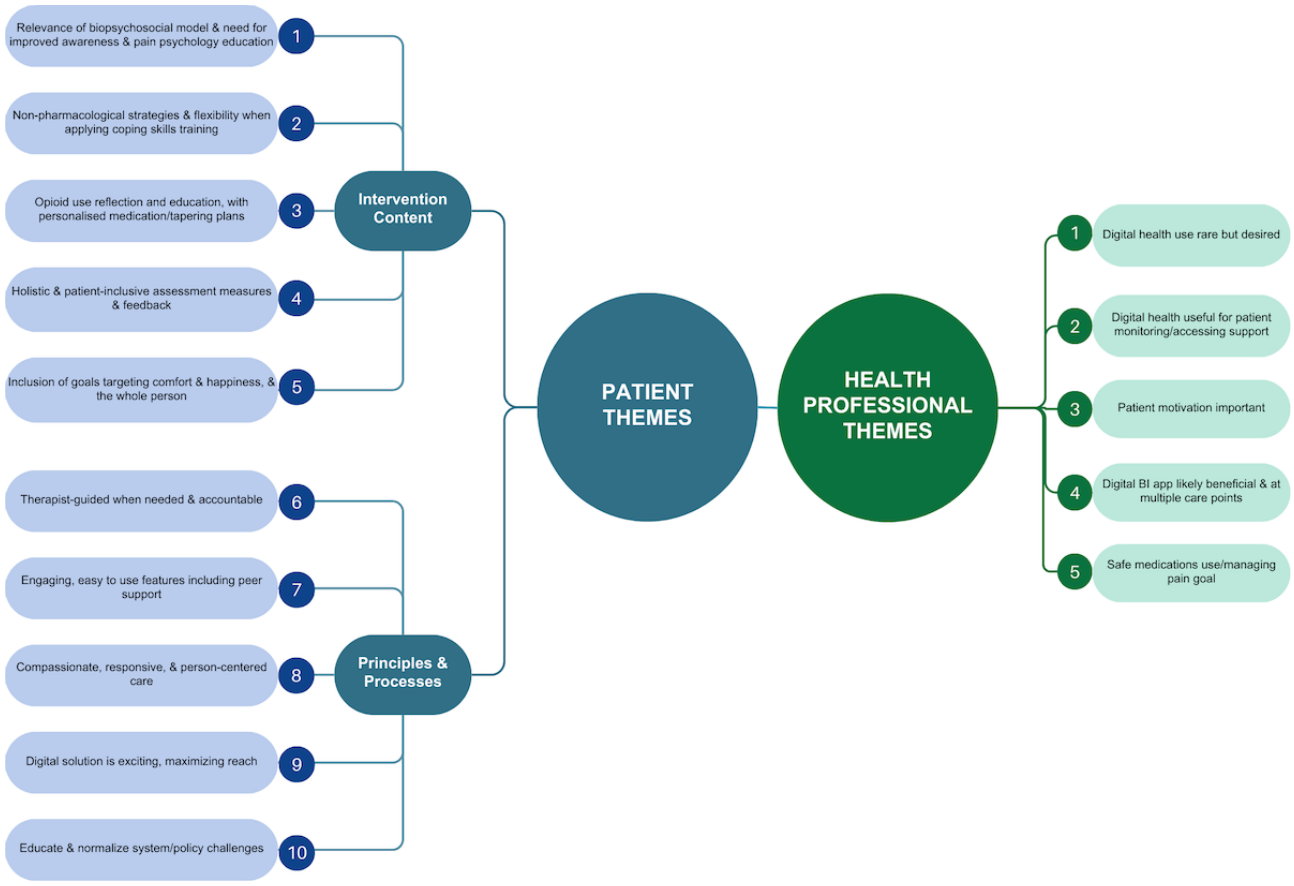

In line with clinical guidelines that recommend multidisciplinary and behavioral treatment approaches as first line (eg, Dowell et al [40] and Busse et al [1]), patients reported that the inclusion of nonpharmacological pain management strategies (eg, relaxation, exercise, pleasant and valued activities, pacing, access to psychology, and multidisciplinary team approaches) would be welcomed, rating this as one of the most important features of the digital BI. Patients also indicated that personalizing coping skills training to patient personality characteristics (eg, impulsivity traits) and pain-related factors (eg, depression) could benefit some patients, while others may prefer to direct themselves to coping skills that resonate with them. Individual choice could be built into the digital BI to recognize patient differences in autonomy and self-efficacy. Notably, nonpharmacological approaches can achieve similar or even greater improvements in pain and functioning without the potential harms of dependence, adverse side effects, and overdose. This is especially important as patients generally report that opioid therapy has modest effects on levels of pain and functioning [41-43].

There were educational needs identified by patients in the areas of pain psychology, the biopsychosocial model of pain, and opioid therapy. These findings are in line with the high levels of interest and increased need for patient education about the role of psychology in the management of pain [4,44]. Educating patients about the biopsychosocial model of pain forms a part of multidisciplinary rehabilitation and discipline-specific treatments that can be effective in reducing pain and disability in the short to medium term (eg, Siddall et al [45] and Wood and Hendrick [46]). Helping patients to reconceptualize pain can also enhance their ability to cope with their condition [47].

XSL•FO
RenderX

The importance of safe opioid use education is highlighted in clinical guidelines [1,48]. Few studies have examined the effectiveness of opioid use education for patients with chronic pain. On the basis of the evidence available, opioid-related education can lead to safer behaviors (eg, less stockpiling [49]) and reduced opioid use following surgery [50]. Gaps in educating patients in pain management, influential biopsychosocial factors, and safe use of opioid medications continues to put patients at risk of poor recovery. Ongoing weighing up of the benefits of opioid therapy in the treatment of CNCP versus risks could be facilitated by a health professional–delivered digital BI. This could be embedded within the review of the patients' personalized medication plan and be supported by regular monitoring and assessment of medication use as well as related symptoms that consider the *whole person* (eg, pain, mood, and functioning). Health professionals considered digital health as an optimal platform to facilitate patient monitoring of symptoms. Patients also reported that one of the top 3 activities that they want to do on their mobile device is manage their medication and track their health [51]. Digital BIs that aim to reduce risk of opioid-related harm have the potential to meet this need.

Patients tended to focus on the potential of the digital BI to meet general well-being goals and reduce pain discomfort to facilitate safer use of opioid medications. Health professionals indicated that patients would be likely interested in pain reduction goals but suggested that goals related to the safe use of opioid medications could also be important. If goals are to encompass the needs of the whole person, then it will be important to explore and balance patients' pain-related goals and opioid medication harm reduction goals as part of holistic goals related to relationships, work, and lifestyle. To facilitate behavioral change, different innovative approaches (eg, gamification and serious games) could be explored in future studies.

Health professionals reported a desire to integrate DHIs into their clinical practice in a more conscious way. Real-world implementation of DHIs have often failed, mostly because they are not used by patients or health professionals [52]. Engaging health professionals in the co-design process has potential to enhance digital BI feasibility and acceptability in practice, as the success of such measures in health care settings relies heavily on the engagement of key stakeholders [22]. Our results could inform the development of a broader digital BI implementation strategy that considers barriers identified by health professionals (eg, costs, time, and compatibility between the digital BI and their current workflow) and enablers (eg, perceived fit of the digital BI approach to current practice), which are common across DHIs and health care settings [53]. Further consideration of the systems used (eg, mobile app vs website and health professional assistance) and adaptability, policy and guideline support, resourcing, health professional knowledge and beliefs, and involvement of digital BI *champions* to support patient motivation [53] may inform effective digital BI implementation. Most psychologists and mental health professionals (72%) have little or no formal training in pain management and more than 50% report feeling a lack of confidence in treating pain [4]. Training and education for health professionals in delivering digital BIs could fill this gap.

Similar to other digital BI studies [54], a therapist-guided digital BI was preferred among patients and health professionals. According to patients, therapist guidance would enhance patient motivation and engagement and provide more accountability than self-guided digital BIs. Evidence indicates that guided digital cognitive behavioral therapy and face-to-face therapy for psychiatric and somatic conditions produce equivalent overall effects [55], and when therapist support is included, this has a positive impact on engagement and treatment effectiveness [56]. Therapist-guided digital BIs provide opportunities for establishing mutual trust, which can be a barrier to intervention success [57]. This may be critical as health professionals report that discussions about chronic pain management, including medication use, can be challenging, with patients reporting that the context of opioid policy changes can further fuel mistrust. Guided digital BIs allow for therapists to foster a strong therapeutic alliance with empathy and compassion, treating patients as partners in their pain care - providing the foundation that is necessary for treatment success. Further consideration of communication methods between patients, their treating clinicians, and the broader team within the digital environment may further personalize pain services and facilitate multidisciplinary collaboration. When asked to rank the importance of the BI features, patients reported *support* was the most important feature, while *fun* was the least. Future co-design opportunities could explore meaningful ways to integrate gamification and digital peer support to facilitate patient engagement.

Both patients and health professionals reported that digital BIs could be best placed and adopted in primary care, multidisciplinary pain centers while patients wait for treatment, and addiction services where it could be an adjunct to in-person treatment; it was also suggested that it could be tailored to prevent risk of opioid-related harm in people with acute pain. There was also a suggestion that digital BIs could form part of a stepped-care model. Further research is needed to determine whether digital BIs for treatment of risk of opioid-related harm are effective and for whom and in what setting.

Strengths and Limitations

Most previous studies have not included the views of key stakeholders in the design of DHIs [17,25]. We advanced the co-design of a digital BI to reduce risk of opioid-related harm among people with CNCP by examining both patient and health professional perspectives on their intervention needs and preferences. Our study adds to the growing literature on co-design in DHIs [58] and extends research into the CNCP field.

We included data from 18 individual patient interviews and from 7 patients and 5 health professionals as part of focus group workshops. While there are no guidelines on the minimum number of key stakeholders or frequency and intensity of engagement that is needed in co-design, it is possible that our results may not be generalizable to the broader population of people with CNCP who are taking prescription opioids, or beyond the Australian context where patients have access to a relatively high standard of public health care and access to heavily subsidized medications. Continual engagement of

XSL•FO
RenderX

various patients and health professional perspectives in the subsequent steps of co-design and co-development of digital BIs could increase customizability of the solution to fit a wider range of users. In future design and development work, consideration of how to design BIs for better inclusion of patient cultural and linguistic diversity, spiritual needs, and socioeconomic factors is needed, as well as investigations in countries with different health and insurance systems. It will be important that the digital divide does not exacerbate existing health inequities of those most at risk in the CNCP population. Efforts to measure digital inclusion could increase translational confidence in digital BIs.

In addition, there remains a power differential between patients, health professionals, and researchers, and it is unknown how this may have influenced our findings. Future work should explore jointly developed guiding ethical principles to address the relationship power imbalances in co-design. There are also potential tensions between these stakeholders in their digital BI needs and preferences which could be explored in future joint focus groups. Finally, this study was conducted before the widespread introduction of generative AI technologies (eg, ChatGPT, Gemini, and Co-Pilot) to the public. Future studies could explore the integration of generative AI into DHIs, such as those that aim to reduce risk of opioid-related harm and improve pain care.

Conclusions

Partnering with patients who have CNCP and health professionals to identify their intervention needs and preferences is critical to the co-design of new, innovative, and high-value digital treatments [59]. Patients and health professionals were interested in the proposed digital BI, identifying many content areas, principles, enablers, and barriers to guide the development of such programs. This co-design approach has the potential to enhance the translation of digital BIs into practice and improve clinical effectiveness. The next phases in co-designing personalized digital BIs involve the generation of later design concepts (eg, journey mapping and persona development) and prototyping, while considering the development of an implementation strategy. Continual, genuine bringing together of scientific, patient, clinical, and technical expertise to solution design could address the urgent need for better patient-centered CNCP care while minimizing the risks of opioid medication–related harms.

Acknowledgments

The research team would like to sincerely thank the patients for participating in this project and sharing their experiences. The authors would also like to thank the specialist pain and addiction services for collaborating on this project. We extend special thanks to Dr Andrew Claus, Dr Geraldine Chew, and Dr Tabibul Khan for their assistance in patient engagement. The authors would also like to thank Ms Chelsea Geary for her contribution to the development of the interview questions and web-based surveys.

Metro South Research Support Scheme Early Career Researcher Grant (CIA RAE; RSS_2020_035) provided funding for this project. The RECOVER Injury Research Centre, The University of Queensland; Addiction & Mental Health Services, Metro South Health; and the Tess Cramond Pain and Research Centre, Metro North Health provided in-kind support. RAE is supported by a Research Stimulus Fellowship grant administered by the University of Queensland.

Data Availability

The datasets generated during and analyzed during this study are not publicly available due to limitations of ethics approval but are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Patient and health professional semistructured interview and focus group questions.
[DOCX File , 14 KB - formative_v9i1e57212_app1.docx]

Multimedia Appendix 2
Personality-driven coping strategies.
[PNG File , 618 KB - formative_v9i1e57212_app2.png]

Multimedia Appendix 3
Goals identified during the "peeling the onion" exercise.
[PNG File , 535 KB - formative_v9i1e57212_app3.png]

References

XSL•FO
RenderX

1. Dowell D, Haegerich TM, Chou R. CDC guideline for prescribing opioids for chronic pain--United States, 2016. JAMA 2016 Apr 19;315(15):1624-1645 [FREE Full text] [doi: 10.1001/jama.2016.1464] [Medline: 26977696]
2. Reis SE, McDonald MC, Byers SJ. Crossing the research valleys of death: the University of Pittsburgh approach. Clin Transl Sci 2008 May 21;1(1):9-10. [doi: 10.1111/j.1752-8062.2008.00021.x] [Medline: 19412329]
3. Crouch TB, Wedin S, Kilpatrick RL, Christon L, Balliet W, Borckardt J, et al. Pain rehabilitation's dual power: treatment for chronic pain and prevention of opioid-related risks. Am Psychol 2020 Sep;75(6):825-839 [FREE Full text] [doi: 10.1037/amp0000663] [Medline: 32915026]
4. Darnall BD, Carr DB, Schatman ME. Pain psychology and the biopsychosocial model of pain treatment: ethical imperatives and social responsibility. Pain Med 2017 Aug 01;18(8):1413-1415 [FREE Full text] [doi: 10.1093/pm/pnw166] [Medline: 27425187]
5. Eccleston C, Morley SJ, Williams AC. Psychological approaches to chronic pain management: evidence and challenges. Br J Anaesth 2013 Jul;111(1):59-63 [FREE Full text] [doi: 10.1093/bja/aet207] [Medline: 23794646]
6. Hogg MN, Gibson S, Helou A, DeGabriele J, Farrell MJ. Waiting in pain: a systematic investigation into the provision of persistent pain services in Australia. Med J Aust 2012 Apr 02;196(6):386-390. [doi: 10.5694/mja12.10140] [Medline: 22471539]
7. Nicholas MK, Asghari A, Sharpe L, Beeston L, Brooker C, Glare P, et al. Reducing the use of opioids by patients with chronic pain: an effectiveness study with long-term follow-up. Pain 2020 Mar 19;161(3):509-519. [doi: 10.1097/j.pain.0000000000001763] [Medline: 31764391]
8. Elphinston RA, Pager S, Brown K, Sterling M, Fatehi F, Gray P, et al. Co-designing a digital brief intervention to reduce the risk of prescription opioid–related harm among people with chronic noncancer pain: qualitative analysis of patient lived experiences. JMIR Form Res (forthcoming) 2024. [doi: 10.2196/57208]
9. The Australian pain and psychology report 2020. Australian Psychological Society. 2020. URL: https://psychology.org.au/getmedia/0189c0e2-1832-422d-af4a-031391ea143c/20aps-pw-report-p1.pdf [accessed 2024-04-29]
10. Bernstein C, Gillman AG, Zhang D, Bartman AE, Jeong JH, Wasan AD. Identifying predictors of recommendations for and participation in multimodal nonpharmacological treatments for chronic pain using patient-reported outcomes and electronic medical records. Pain Med 2020 Dec 25;21(12):3574-3584. [doi: 10.1093/pm/pnaa203] [Medline: 32869082]
11. Psychological expertise where we need it most: GPs not the only answer to our mental health crisis. Australian Psychological Society. 2022. URL: https://psychology.org.au/about-us/news-and-media/media-releases/2022/psychological-expertise-where-we-need-it-most-gps [accessed 2024-04-29]
12. Darnall BD, Scheman J, Davin S, Burns JW, Murphy JL, Wilson AC, et al. Pain psychology: a global needs assessment and national call to action. Pain Med 2016 Feb 23;17(2):250-263 [FREE Full text] [doi: 10.1093/pm/pnv095] [Medline: 26803844]
13. Darnall BD. Psychological treatment for chronic pain: improving access and integration. Psychol Sci Public Interest 2021 Sep 20;22(2):45-51 [FREE Full text] [doi: 10.1177/15291006211033612] [Medline: 34541966]
14. Eccleston C, Blyth FM, Dear BF, Fisher EA, Keefe FJ, Lynch ME, et al. Managing patients with chronic pain during the COVID-19 outbreak: considerations for the rapid introduction of remotely supported (eHealth) pain management services. Pain 2020 May 2;161(5):889-893 [FREE Full text] [doi: 10.1097/j.pain.0000000000001885] [Medline: 32251203]
15. Slater H, Dear BF, Merolli MA, Li LC, Briggs AM. Use of eHealth technologies to enable the implementation of Musculoskeletal models of care: evidence and practice. Best Pract Res Clin Rheumatol 2016 Jun;30(3):483-502. [doi: 10.1016/j.berh.2016.08.006] [Medline: 27886943]
16. de la Vega R, Miró J. mHealth: a strategic field without a solid scientific soul: a systematic review of pain-related apps. PLoS One 2014 Jul 7;9(7):e101312 [FREE Full text] [doi: 10.1371/journal.pone.0101312] [Medline: 24999983]
17. Lalloo C, Jibb LA, Rivera J, Agarwal A, Stinson JN. "There's a pain app for that": review of patient-targeted smartphone applications for pain management. Clin J Pain 2015 Jun;31(6):557-563. [doi: 10.1097/AJP.0000000000000171] [Medline: 25370138]
18. Fernandes LG, Devan H, Fioratti I, Kamper SJ, Williams CM, Saragiotto BT. At my own pace, space, and place: a systematic review of qualitative studies of enablers and barriers to telehealth interventions for people with chronic pain. Pain 2022 Feb 01;163(2):e165-e181. [doi: 10.1097/j.pain.0000000000002364] [Medline: 34433776]
19. Ioannidis JP. Why most clinical research is not useful. PLoS Med 2016 Jun 21;13(6):e1002049 [FREE Full text] [doi: 10.1371/journal.pmed.1002049] [Medline: 27328301]
20. Schlieter H, Marsch LA, Whitehouse D, Otto L, Londral AR, Teepe GW, et al. Scale-up of digital innovations in health care: expert commentary on enablers and barriers. J Med Internet Res 2022 Mar 11;24(3):e24582. [doi: 10.2196/24582] [Medline: 35275065]
21. Svendsen MJ, Wood KW, Kyle J, Cooper K, Rasmussen CD, Sandal LF, et al. Barriers and facilitators to patient uptake and utilisation of digital interventions for the self-management of low back pain: a systematic review of qualitative studies. BMJ Open 2020 Dec 12;10(12):e038800. [doi: 10.1136/bmjopen-2020-038800] [Medline: 33310794]
22. Slattery P, Saeri AK, Bragge P. Research co-design in health: a rapid overview of reviews. Health Res Policy Syst 2020 Feb 11;18(1):17. [doi: 10.1186/s12961-020-0528-9] [Medline: 32046728]

XSL•FO
RenderX

23. Keogh E, Rosser BA, Eccleston C. e-Health and chronic pain management: current status and developments. Pain 2010 Oct;151(1):18-21. [doi: 10.1016/j.pain.2010.07.014] [Medline: 20674174]
24. Tabi K, Randhawa AS, Choi F, Mithani Z, Albers F, Schnieder M, et al. Mobile apps for medication management: review and analysis. JMIR Mhealth Uhealth 2019 Sep 11;7(9):e13608 [FREE Full text] [doi: 10.2196/13608] [Medline: 31512580]
25. Zhao P, Yoo I, Lancey R, Varghese E. Mobile applications for pain management: an app analysis for clinical usage. BMC Med Inform Decis Mak 2019 May 30;19(1):106. [doi: 10.1186/s12911-019-0827-7] [Medline: 31146739]
26. Devan H, Farmery D, Peebles L, Grainger R. Evaluation of self-management support functions in apps for people with persistent pain: systematic review. JMIR Mhealth Uhealth 2019 Feb 12;7(2):e13080 [FREE Full text] [doi: 10.2196/13080] [Medline: 30747715]
27. Elphinston RA, Sterling M, Leung J, Gray PD, Scotti Requena S, Connor JP. Brief psychological interventions for reducing prescription opioid use, related harm, and pain intensity in patients with chronic pain: a systematic review and assessment of patient interest. Clin J Pain 2021 Apr 01;37(4):270-280. [doi: 10.1097/AJP.0000000000000908] [Medline: 33323791]
28. Kessler RC, Andrews G, Colpe LJ, Hiripi E, Mroczek DK, Normand SL, et al. Short screening scales to monitor population prevalences and trends in non-specific psychological distress. Psychol Med 2002 Aug;32(6):959-976. [doi: 10.1017/s0033291702006074] [Medline: 12214795]
29. Guest G, Namey E, Chen M. A simple method to assess and report thematic saturation in qualitative research. PLoS One 2020 May 5;15(5):e0232076 [FREE Full text] [doi: 10.1371/journal.pone.0232076] [Medline: 32369511]
30. Sanders EB, Stappers PJ. Co-creation and the new landscapes of design. CoDesign 2008 Mar;4(1):5-18. [doi: 10.1080/15710880701875068]
31. Noorbergen TJ, Adam MT, Teubner T, Collins CE. Using co-design in mobile health system development: a qualitative study with experts in co-design and mobile health system development. JMIR Mhealth Uhealth 2021 Nov 10;9(11):e27896 [FREE Full text] [doi: 10.2196/27896] [Medline: 34757323]
32. Kumar SP. Utilization of brief pain inventory as an assessment tool for pain in patients with cancer: a focused review. Indian J Palliat Care 2011 May;17(2):108-115 [FREE Full text] [doi: 10.4103/0973-1075.84531] [Medline: 21976850]
33. Tan G, Jensen MP, Thornby JI, Shanti BF. Validation of the brief pain inventory for chronic nonmalignant pain. J Pain 2004 Mar;5(2):133-137. [doi: 10.1016/j.jpain.2003.12.005] [Medline: 15042521]
34. Lovibond PF, Lovibond SH. The structure of negative emotional states: comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories. Behav Res Ther 1995 Mar;33(3):335-343. [doi: 10.1016/0005-7967(94)00075-u] [Medline: 7726811]
35. Butler SF, Budman SH, Fernandez KC, Houle B, Benoit C, Katz N, et al. Development and validation of the current opioid misuse measure. Pain 2007 Jul;130(1-2):144-156 [FREE Full text] [doi: 10.1016/j.pain.2007.01.014] [Medline: 17493754]
36. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
37. Darnall BD, Burns JW, Hong J, Roy A, Slater K, Poupore-King H, et al. Empowered Relief, cognitive behavioral therapy, and health education for people with chronic pain: a comparison of outcomes at 6-month Follow-up for a randomized controlled trial. Pain Rep 2024 Jan;9(1):e1116 [FREE Full text] [doi: 10.1097/PR9.0000000000001116] [Medline: 38288134]
38. Darnall BD, Ziadni MS, Krishnamurthy P, Flood P, Heathcote LC, Mackey IG, et al. "My surgical success": effect of a digital behavioral pain medicine intervention on time to opioid cessation after breast cancer surgery-a pilot randomized controlled clinical trial. Pain Med 2019 Nov 01;20(11):2228-2237 [FREE Full text] [doi: 10.1093/pm/pnz094] [Medline: 31087093]
39. Ziadni MS, Anderson SR, Gonzalez-Castro L, Darnall BD. Comparative efficacy of a single-session “Empowered Relief” videoconference-delivered group intervention for chronic pain: study protocol for a randomized controlled trial. Trials 2021 May 22;22(1):358. [doi: 10.1186/S13063-021-05303-8]
40. Busse JW, Craigie S, Juurlink DN, Buckley DN, Wang L, Couban RJ, et al. Guideline for opioid therapy and chronic noncancer pain. CMAJ 2017 May 08;189(18):E659-E666 [FREE Full text] [doi: 10.1503/cmaj.170363] [Medline: 28483845]
41. Abdel Shaheed C, Maher CG, Williams KA, Day R, McLachlan AJ. Efficacy, tolerability, and dose-dependent effects of opioid analgesics for low back pain: a systematic review and meta-analysis. JAMA Intern Med 2016 Jul 01;176(7):958-968. [doi: 10.1001/jamainternmed.2016.1251] [Medline: 27213267]
42. Busse JW, Wang L, Kamaleldin M, Craigie S, Riva JJ, Montoya L, et al. Opioids for chronic noncancer pain: a systematic review and meta-analysis. JAMA 2018 Dec 18;320(23):2448-2460 [FREE Full text] [doi: 10.1001/jama.2018.18472] [Medline: 30561481]
43. Reinecke H, Weber C, Lange K, Simon M, Stein C, Sorgatz H. Analgesic efficacy of opioids in chronic pain: recent meta-analyses. Br J Pharmacol 2015 Jan;172(2):324-333. [doi: 10.1111/bph.12634] [Medline: 24640991]
44. Darnall BD. Minimize opioids by optimizing pain psychology. Pain Manag 2014 Jul 10;4(4):251-253 [FREE Full text] [doi: 10.2217/pmt.14.18] [Medline: 25300380]
45. Siddall B, Ram A, Jones MD, Booth J, Perriman D, Summers SJ. Short-term impact of combining pain neuroscience education with exercise for chronic musculoskeletal pain: a systematic review and meta-analysis. Pain 2022 Jan 01;163(1):e20-e30. [doi: 10.1097/j.pain.0000000000002308] [Medline: 33863860]

XSL•FO
RenderX

46. Wood L, Hendrick PA. A systematic review and meta-analysis of pain neuroscience education for chronic low back pain: short-and long-term outcomes of pain and disability. Eur J Pain 2019 Feb 14;23(2):234-249. [doi: 10.1002/ejp.1314] [Medline: 30178503]
47. Watson JA, Ryan CG, Cooper L, Ellington D, Whittle R, Lavender M, et al. Pain neuroscience education for adults with chronic musculoskeletal pain: a mixed-methods systematic review and meta-analysis. J Pain 2019 Oct;20(10):1140.e1-1140.22. [doi: 10.1016/j.jpain.2019.02.011] [Medline: 30831273]
48. Häuser W, Morlion B, Vowles KE, Bannister K, Buchser E, Casale R, et al. European* clinical practice recommendations on opioids for chronic noncancer pain - part 1: role of opioids in the management of chronic noncancer pain. Eur J Pain 2021 May 02;25(5):949-968 [FREE Full text] [doi: 10.1002/ejp.1736] [Medline: 33655607]
49. Hero JO, McMurtry C, Benson J, Blendon R. Discussing opioid risks with patients to reduce misuse and abuse: evidence from 2 surveys. Ann Fam Med 2016 Nov 14;14(6):575-577 [FREE Full text] [doi: 10.1370/afm.1994] [Medline: 28376444]
50. Zorrilla-Vaca A, Mena G, Ramirez P, Lee B, Sideris A, Wu C. Effectiveness of perioperative opioid educational initiatives: a systematic review and meta-analysis. Anesth Analg 2022 May 01;134(5):940-951. [doi: 10.1213/ANE.0000000000005634] [Medline: 34125081]
51. National digital health strategy. Australian Digital Health Agency. URL: https://www.digitalhealth.gov.au/national-digital-health-strategy [accessed 2024-04-29]
52. Mohr DC, Riper H, Schueller SM. A solution-focused research approach to achieve an implementable revolution in digital mental health. JAMA Psychiatry 2018 Feb 01;75(2):113-114. [doi: 10.1001/jamapsychiatry.2017.3838] [Medline: 29238805]
53. Ross J, Stevenson F, Lau R, Murray E. Factors that influence the implementation of e-health: a systematic review of systematic reviews (an update). Implement Sci 2016 Oct 26;11(1):146. [doi: 10.1186/s13012-016-0510-7] [Medline: 27782832]
54. Davin SA, Savage J, Thompson NR, Schuster A, Darnall BD. Transforming standard of care for spine surgery: integration of an online single-session behavioral pain management class for perioperative optimization. Front Pain Res (Lausanne) 2022 May 2;3:856252 [FREE Full text] [doi: 10.3389/fpain.2022.856252] [Medline: 35599968]
55. Carlbring P, Andersson G, Cuijpers P, Riper H, Hedman-Lagerlöf E. Internet-based vs. face-to-face cognitive behavior therapy for psychiatric and somatic disorders: an updated systematic review and meta-analysis. Cogn Behav Ther 2018 Jan 07;47(1):1-18. [doi: 10.1080/16506073.2017.1401115] [Medline: 29215315]
56. Baumeister H, Reichler L, Munzinger M, Lin J. The impact of guidance on internet-based mental health interventions — a systematic review. Internet Interv 2014 Oct;1(4):205-215. [doi: 10.1016/j.invent.2014.08.003]
57. Bathia NS, McAskill RE, Hancox JE, Knaggs RD. Collaboration between adult patients and practitioners when making decisions about prescribing opioid medicines for chronic non-cancer pain in primary care: a scoping review. Br J Pain 2022 Feb 23;16(1):119-126 [FREE Full text] [doi: 10.1177/20494637211025560] [Medline: 35111320]
58. Vandekerckhove P, de Mul M, Bramer WM, de Bont AA. Generative participatory design methodology to develop electronic health interventions: systematic literature review. J Med Internet Res 2020 Apr 27;22(4):e13780 [FREE Full text] [doi: 10.2196/13780] [Medline: 32338617]
59. Eather CE, Sterling M, Sullivan C, Elphinston RA. Leveraging value-based health principles to improve translation and impact of digital psychological interventions for people with chronic pain. Pain (forthcoming) 2024 Oct 15. [doi: 10.1097/j.pain.0000000000003441] [Medline: 39451013]

Abbreviations

BI: brief intervention
CNCP: chronic noncancer pain
COMM: Current Opioid Misuse Measure
DHI: digital health intervention
SMART: Specific, Measurable, Achievable, Relevant, and Time-Bound

Edited by A Mavragani; submitted 23.02.24; peer-reviewed by G Lewis; comments to author 01.05.24; revised version received 12.08.24; accepted 25.09.24; published 25.04.25.

Please cite as:
Elphinston RA, Pager S, Fatehi F, Sterling M, Brown K, Gray P, Hipper L, Cahill L, Ziadni M, Worthy P, Connor JP
Applying Patient and Health Professional Preferences in Co-Designing a Digital Brief Intervention to Reduce the Risk of Prescription Opioid–Related Harm Among Patients With Chronic Noncancer Pain: Qualitative Analysis
JMIR Form Res 2025;9:e57212
URL: https://formative.jmir.org/2025/1/e57212
doi:10.2196/57212
PMID:

XSL•FO
RenderX

©Rachel A Elphinston, Sue Pager, Farhad Fatehi, Michele Sterling, Kelly Brown, Paul Gray, Linda Hipper, Lauren Cahill, Maisa Ziadni, Peter Worthy, Jason P Connor. Originally published in JMIR Formative Research (https://formative.jmir.org), 25.04.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Acceptability and Usability of a Digital Behavioral Health Platform for Youth at Risk of Suicide: User-Centered Design Study With Patients, Practitioners, and Business Gatekeepers

Trinity Chloe Tse[1,2], BA; Lauren S Weiner[3], PhD; Carter J Funkhouser[1,2], PhD; Danielle DeLuise[3], BA; Colleen Cullen[1], PsyD; Zachary Blumkin[1], PsyD; Casey O'Brien[1], PsyD; Randy P Auerbach[1,2], PhD; Nicholas B Allen[3,4], PhD

[1]Division of Child and Adolescent Psychiatry, Columbia University Irving Medical Center, Columbia University, New York, NY, United States

[2]New York State Psychiatric Institute, Manhattan, NY, United States

[3]Ksana Health Inc, Eugene, OR, United States

[4]Department of Psychology, University of Oregon, Eugene, OR, United States

Corresponding Author:
Nicholas B Allen, PhD
Department of Psychology
University of Oregon
Department of Psychology
Univ of Oregon
Eugene, OR, 97403
United States
Phone: 1 5413464075
Email: nallen3@uoregon.edu

Abstract

Background: Youth suicide rates are climbing, underscoring the need to improve clinical care. Personal smartphones can provide an understanding of proximal risk factors associated with suicide and facilitate consistent contact between patients and practitioners to improve treatment engagement and effectiveness. The Vira digital behavior change platform (Vira) consists of a patient smartphone app and a web-based practitioner portal (Vira Pro) that integrates objective mobile sensing data with Health Insurance Portability and Accountability Act (HIPAA)–compliant communication tools. Through Vira, practitioners can continuously assess patients' real-world behavior and provide clinical tools to enhance treatment via just-in-time behavior change support.

Objective: This study aimed to explore the acceptability and usability of the minimal viable product version of Vira through a user-centered design (UCD) approach and to identify barriers to implementing Vira in the context of an adolescent intensive outpatient program.

Methods: Over 2 iterative phases, feedback was gathered from adolescent patients (n=16), mental health practitioners (n=11), and business gatekeepers (n=5). The mixed methods UCD approach included individual semistructured interviews (eg, perspectives on treatment and attitudes toward digital tools), surveys (eg, usability), and unmoderated user testing sessions (eg, user experience).

Results: Overall, participants expressed optimism regarding Vira, particularly among adolescents, who showed high satisfaction with the app's interface and design. However, clinicians reported more mixed views, agreeing that it would be useful in treatment but also expressing concerns about the volume and displays of patient data in Vira Pro, workload management, and boundaries. Gatekeepers identified usability issues and implementation barriers related to electronic health records but also recognized Vira's potential to enhance treatment outcomes. Feedback from stakeholders informed several crucial changes to the platform, including adjustments to data-sharing protocols, user interface enhancements, and modifications to training methods.

Conclusions: Vira has a high potential to improve patient engagement and improve clinical outcomes among high-risk youth. Iterative UCD and ongoing stakeholder engagement are essential for developing technology-based interventions that effectively meet the needs of diverse end users and align with clinical workflows.

(JMIR Form Res 2025;9:e65418) doi:10.2196/65418

XSL•FO
RenderX

KEYWORDS

user-centered design; mobile sensing; behavior change; suicide prevention; digital health; mental health

Introduction

Background

Suicide is the second leading cause of death among adolescents, and from 2011 to 2021, the rate of suicide attempts and suicidal behaviors has been increasing [1-3]. Existing mental health treatments are only moderately effective [4], and even among those who receive a "gold-standard" treatment, many do not improve. To reduce the public health burden of adolescent mental health problems, there is a critical need for novel approaches that increase the effectiveness of existing treatments.

One promising way to improve mental health treatments is to leverage digital devices such as personal smartphones. Smartphones have the potential to both monitor risk in real time, and to deliver just-in-time interventions, both of which are difficult if not impossible to achieve without these ubiquitous personal computing devices. To date, few digital interventions have been developed with or for youth at risk of suicide and rigorously tested to provide evidence of feasibility, acceptability, or efficacy [5,6]. Examples of other mental health apps targeted toward mitigating suicide risk include BRITE, LifeBuoy, and BlueIce. BRITE was developed to support suicide safety planning among adolescents consists of a practitioner portal and patient app, and has primarily been implemented as a single-session intervention [7]. By contrast, LifeBuoy is an app that helps young people manage suicidal thoughts in their daily lives by teaching dialectical behavior therapy (DBT) skills via self-guided modules [8], whereas BlueIce (MyOxygen, Limited) is used to support distress tolerance and reduce self-harm through a personalized toolbox of strategies that include mood diaries, a menu of mood-lifting activities, and automatic routing to emergency numbers if needed [9]. Although initial clinical outcomes are promising [10], these apps do not appear to support continuous treatment between patients and their practitioners, a critical aspect of evidence-based treatments such as DBT [11], nor do they include mobile sensing, which could provide unique insights into risk processes and opportunities for intervention that occur outside the health care context.

Measurement-based care is important for monitoring treatment effectiveness and improving clinical outcomes [12] but is not often used in routine clinical care due to concerns about practitioner and patient burden [13]. Through both passive (eg, built-in smartphone sensors) and active data collection (eg, experience sampling methods), smartphones can directly address these concerns by allowing practitioners and patients to monitor mood, behavior, symptoms, and risk factors continuously, objectively, and with a relatively low burden. In addition, smartphones afford novel opportunities to reinforce or introduce intervention content between sessions. For example, patients often forget to practice therapeutic skills outside of sessions [14], which may lead to poorer treatment outcomes [15]. Smartphone push notifications delivered between sessions can remind patients to practice therapeutic skills, thus increasing skill usage and improving patient outcomes [16].

Although digital mental health interventions have significant potential, low adherence and high attrition can reduce clinical effectiveness [17]. In particular, digital interventions that do not address the unique needs and perspectives of end users may exhibit low usability, which leads to lower sustained engagement [18]. When developing digital interventions, it is important to consider the clinical and implementation context [19,20]. User-centered design (UCD) uses an iterative, stakeholder-engaged process to cocreate products that directly respond to the end user experience [21,22]. Through UCD, individuals with lived experiences are closely engaged during the design and development processes by providing feedback that can be used by intervention developers and designers to tailor the content, interface, and implementation of digital interventions [23,24]. UCD studies also inform the translation of evidence-based treatments to digital formats and reveal critical information about facilitators and barriers to adoption that guide implementation approaches [25,26].

The Vira digital behavior change platform consists of a user-facing smartphone app and a web-based linked practitioner portal, Vira Pro (refer to Figure 1). The Vira platform was developed based on principles of behavioral activation, an effective treatment for depression and other mental health problems [27]. The smartphone app uses mobile sensing and brief self-report surveys to monitor numerous indicators of mental health problems (eg, sleep disturbance and low mood) or proximal risk factors. The app provides patients access to these data 24/7 and offers insights into passively sensed behavioral patterns associated with their moods. The web portal displays these data to the clinician, allowing them to gain more nuanced insights into their patients' behavior and mood between sessions. Clinicians can also use the web portal to schedule personalized push notifications to be delivered to a patients' smartphone, thereby facilitating enhanced support between sessions. Importantly, this functionality allows pushing notifications to be delivered at the "right time" (eg, as just-in-time adaptive interventions), which may improve outcomes [28].

XSL•FO
RenderX

Figure 1. The Vira digital behavior change platform consists of a user-facing smartphone app, Vira, and a linked, web-based practitioner portal, Vira Pro.

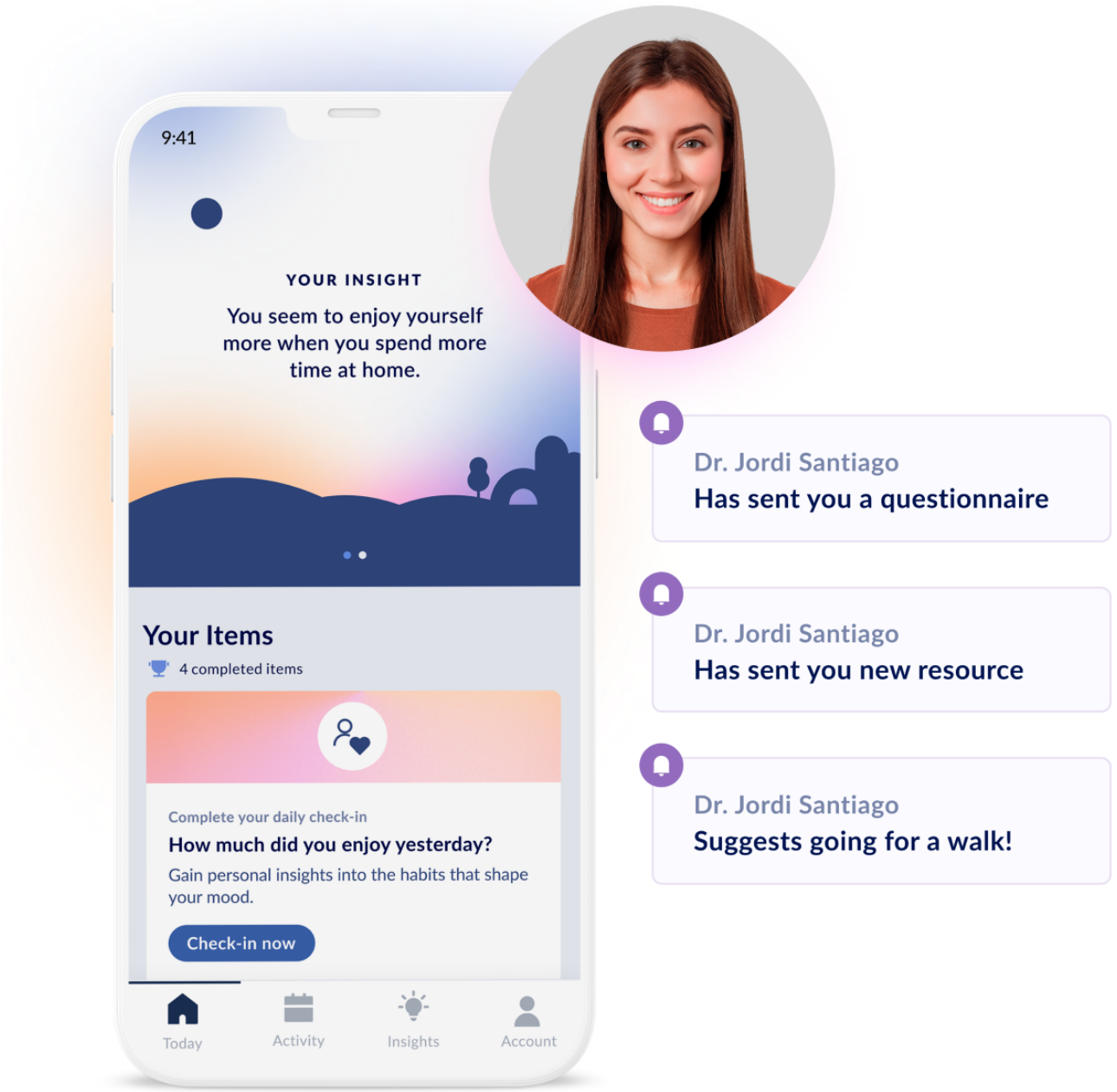

Study Objectives

This study aimed to evaluate the Vira mobile app and Vira Pro practitioner portal for use with youth at high risk for suicide by using an iterative formative development process to incorporate feedback from adolescents with lived experience, care providers, and business gatekeepers (ie, individuals who are responsible for making decisions regarding financial and other support for clinic initiatives). By including the perspectives of these end users, we sought to develop a digital platform that can be integrated into clinical workflows, is maximally engaging, and optimized for successful implementation.

Methods

Study Overview

This study was part of a larger project to develop and test a behavior change platform for use in an intensive outpatient program (IOP) serving youth at high risk of suicide. The goal of the iterative formative development study, which consisted of 2 phases of mixed methods data collection, was to inform product development and implementation for a pragmatic randomized controlled trial (RCT; refer to Figure 2 for study design) [29].

XSL•FO
RenderX

Figure 2. Overview of the iterative formative development process and overall project design. RCT: randomized controlled trial.

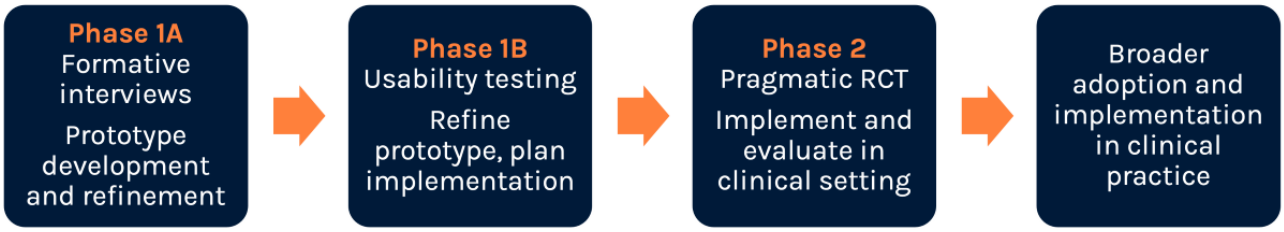

Participant Recruitment

Participants were recruited between October 2022 and March 2023. Three types of stakeholders were recruited from the broader NewYork-Presbyterian Hospital, ColumbiaDoctors, and New York State Psychiatric Institute communities to participate across 2 iterative phases of research (Phase 1 and Phase 2). Stakeholders included adolescents who had received previous mental health treatment, mental health clinicians, and business gatekeepers.

Adolescent participants were recruited from a pool who had previously participated in other studies and provided informed consent to be contacted for future study opportunities. Adolescents were considered eligible if they were currently receiving or had received mental health treatment for depression or anxiety symptoms within the past year, were between the ages of 13 and 18 years old, and were fluent in English.

Clinicians and business gatekeepers were recruited internally via outreach emails and electronic flyers. Eligibility criteria for clinicians and business gatekeepers included present employment in youth mental health services at NewYork-Presbyterian Hospital, ColumbiaDoctors, or New York State Psychiatric Institute, ages 18-56 years old, and fluency in English. Business gatekeepers served in clinical leadership roles (eg, clinic director and executive director). Several clinicians who participated in Phase 1 also participated in Phase 2. One participant held both administrative and clinical roles and therefore participated as a business gatekeeper in Phase 1 and a clinician in Phase 2. Several of the clinicians and business gatekeepers have previously collaborated with RPA on other research projects.

Study Procedures

This study was an iterative formative development process consisting of 2 phases of mixed methods data collection. A UCD approach was adopted whereby changes made to the app prototype presented in Phase 2 were based on participant feedback from Phase 1. In both phases, participants completed semistructured interviews, an unmoderated click-through and think-aloud exercise to provide feedback on the minimal viable product version of the app, and a quantitative feedback survey using items adapted from the technology acceptance model (TAM) [30]. The TAM has 2 scales to measure specific beliefs that may influence the acceptance of technology: perceived usefulness and perceived ease of use. Each scale contains 10 items, which are rated using a 7-point Likert scale from 1 (strongly disagree) to 7 (strongly agree). Higher ratings indicated higher acceptability. Participants in both phases also completed a demographics questionnaire. Each interview was conducted by 2 moderators (TCT and DD). Both moderators are female and hold bachelor's degrees. TCT is a research assistant, and DD oversaw product design and development at Ksana Health, Inc at the time of the study.

Phase 1

Formative interviews in Phase 1 aimed to increase our understanding of the target patient population and clinical context, elicit feedback on the Vira minimal viable product to inform adaptations and refinements to the platform to increase relevance for the IOP clinical setting, promote perceived ease of use, understand potential barriers, and facilitate future implementation. Adolescents, clinicians, and gatekeepers were included in this phase.

Adolescent interview topics in Phase 1 included their past and current experiences receiving mental health treatment, smartphone usage and habits, and initial impressions of Vira based on a verbal description of the platform. Clinician interviews focused on their current therapy practices, experience using technology as part of the mental health treatment they deliver, past experiences integrating digital tools into existing electronic health record (EHR) systems, and their first impressions of Vira. Business gatekeepers reported on their experiences managing clinicians and clinic caseload, use of technology as part of mental health treatment, past experiences integrating new tools into existing EHR systems, and their first impressions of the platform. Interviews were approximately 30 minutes in duration and were conducted through Microsoft Teams videoconferencing software.

After completing the interview, participants took part in an unmoderated user testing session. Usability testing was conducted through UserTesting and began with a "think-aloud" procedure to test user interactions with Vira [31,32]. Participants were asked to verbalize their reactions as they viewed screens and to describe their navigational choices as they used the platform. Users were also asked about barriers and facilitators to using the platform, perceptions of specific features, and how to improve it.

Phase 2

The goal of Phase 2 usability testing was to understand the usability and acceptability of the refined Vira platform and deepen our knowledge of the implementation context to support clinical implementation in the subsequent pragmatic RCT. Adolescents and clinicians were included in this phase, as they were the primary target end users in the subsequent pragmatic RCT.

Phase 2 interviews with adolescents and clinicians explored similar topics to the Phase 1 interviews. In Phase 2, clinician interviews additionally focused on implementation, clinician

XSL•FO
RenderX

readiness, and usability of the revised Vira platform. Interviews necessitated approximately 30 minutes and were completed through Microsoft Teams and Zoom. During the unmoderated user testing portion of the session completed through the UserTesting platform, clinicians were provided with a mock patient profile and asked to think aloud as they explored the Vira tool as if they were using it in a client session. The goal of this mock case study testing format was to understand the clinical implementation context, facilitators and barriers to use, and overall usability. Clinicians were also asked more detailed questions about EHR integration, including which data points from Vira would be most useful to them.

Ethical Considerations

This study was reviewed and approved by the New York State Psychiatric Institute Institutional Review Board (protocol #8307). Signed electronic informed consent was obtained from all participants included in this study. The study data were anonymized and deidentified. All data is uploaded to a secure research server that is only accessible by the research team. Participants were compensated US $200 for their participation in the study.

Data Analysis

Similar analytic approaches were used across phases. Qualitative and quantitative data were analyzed separately.

Semistructured Interviews and UserTesting Transcripts

Semistructured qualitative interviews and UserTesting transcripts were analyzed using a thematic analysis approach [33]. Coding categories were developed to characterize participant responses, including reactions to Vira's features and functions as well as overall difficulty or facility using the platform. Given the iterative nature of the project, rapid analysis techniques, including matrix displays and triangulation [34,35], were also used. Multiple team members (TCT, LSW, DD, RPA, and NBA) reviewed transcripts to identify themes across phases for each persona (adolescent, clinician, or gatekeeper). Once team consensus was reached, findings were summarized narratively by theme and persona.

Quantitative Surveys

Demographics were analyzed using descriptive statistics and frequencies (mean [SD] and n [%] where applicable). TAM ratings were analyzed by comparing frequencies (mean [SD]) and n (%) of ratings within and across phases.

Mixed Methods Interpretation

The 2 sets of results (qualitative and quantitative) for each persona were triangulated by theme across both phases to examine how each set of data compared and contrasted to the other set and to explore the complementarity of the results across data collection methods. The qualitative and quantitative results were then synthesized for interpretation. The mixed methods synthesis of qualitative and quantitative results across both phases is presented by persona in the *Results* section.

Results

Demographics

A total of 36 participants were recruited across both phases of the study. In Phase 1, a total of 8 adolescents, 7 clinicians, and 5 gatekeepers participated in stakeholder consultations. One gatekeeper did not provide demographic information. In total, 4 of the clinicians in Phase 1 also participated in Phase 2. One gatekeeper participated in Phase 2 as a clinician. A total of 8 adolescents and 7 clinicians participated in Phase 2. Refer to Table 1 for sample characteristics.

Table 2 summarizes participant responses to the TAM. One gatekeeper did not complete TAM in Phase 1 and one clinician did not complete TAM in Phase 2. For individual item frequency distributions, please refer to Figures S1-S6 in the Multimedia Appendix 1.

XSL•FO
RenderX

Table 1. Demographic characteristics of participants (N=36).

	Phase 1	Phase 2
Adolescents, n	8	8
Age (years), mean (SD)	16.2 (0.91)	17.5 (0.91)
Women, n (%)	7 (87.50)	7 (87.50)
Race, n (%)		
White	3 (37.50)	5 (62.50)
Asian	1 (12.50)	2 (25.00)
Black	0 (0)	1 (12.50)
More than one race	3 (37.50)	0 (0)
Unknown or not reported	1 (12.50)	0 (0)
Ethnicity, n (%)		
Hispanic	1 (12.50)	0 (0)
Non-Hispanic	7 (87.50)	8 (100)
Clinicians (N)	7	8
Age in years, mean (SD)	32.14 (5.24)	34.75 (4.92)
Women, n (%)	7 (100)	8 (100)
Race, n (%)		
White	5 (71.43)	7 (87.50)
More than one race	2 (28.57)	1 (12.50)
Ethnicity, n (%)		
Hispanic	1 (14.29)	0 (0)
Non-Hispanic	6 (85.71)	8 (100)
Business gatekeepers (N)	5	—[a]
Age (years), mean (SD)	44.6 (9.09)	—
Women, n (%)	2 (40)	—
Race, n (%)		
White	4 (80)	—
Unknown or not reported	1 (20)	—
Ethnicity, n (%)		
Non-Hispanic	4 (80)	—
Unknown or not reported	1 (20)	—

[a]Not applicable.

XSL•FO
RenderX

Table 2. Participant technology acceptance model responses (N=33).

Technology acceptance model	Adolescents, mean (SD)		Clinicians, mean (SD)		Gatekeepers, mean (SD)
	Phase 1 (n=8)	Phase 2 (n=7)	Phase 1 (n=7)	Phase 2 (n=7)	Phase 1 (n=4)
Perceived usefulness scale					
Greater control	5.63 (0.74)	5.14 (1.35)	4.86 (0.69)	5.14 (1.68)	6.00 (0.82)
Improve quality	6.13 (0.64)	5.14 (1.86)	4.86 (0.90)	5.00 (2.16)	6.50 (0.58)
Accomplish quickly	5.50 (1.31)	5.86 (1.21)	3.86 (2.12)	4.71 (1.38)	5.75 (1.50)
Support critical aspects	5.75 (1.28)	5.29 (1.25)	5.00 (0.82)	5.43 (1.62)	6.50 (0.58)
Productive	5.88 (0.99)	5.71 (1.38)	3.86 (1.07)	4.43 (2.30)	5.25 (2.06)
Improve	5.75 (1.28)	5.57 (1.40)	4.43 (1.62)	5.14 (1.68)	5.75 (0.96)
Accomplish more	6.00 (1.20)	5.43 (1.40)	5.14 (1.21)	3.29 (2.14)	5.50 (1.29)
Enhance effectiveness	6.00 (1.31)	5.57 (1.40)	5.00 (0.82)	5.14 (1.95)	6.50 (0.58)
Easier	5.75 (1.49)	4.71 (1.25)	3.86 (1.77)	4.57 (1.40)	5.50 (1.29)
Useful	6.00 (0.93)	5.43 (1.13)	5.00 (1.00)	5.86 (1.21)	6.25 (0.50)
Perceived ease of use scale					
Cumbersome to use	2.71 (1.38)	3.14 (1.86)	3.43 (1.27)	4.71 (2.14)	2.00 (1.41)
Easy to learn	5.75 (1.48)	5.57 (0.98)	5.71 (1.11)	4.86 (0.90)	6.00 (0.82)
Frustrating	2.63 (1.60)	2.29 (1.11)	3.00 (1.15)	2.86 (1.07)	2.00 (0.82)
Confusing	2.50 (1.41)	2.43 (1.51)	3.29 (1.60)	2.71 (1.38)	2.00 (0.82)
Rigid and inflexible	2.00 (1.31)	2.29 (1.60)	3.14 (1.68)	2.00 (1.00)	2.75 (1.50)
Easy to remember	5.38 (1.92)	5.29 (1.11)	5.57 (0.53)	4.71 (1.80)	6.50 (0.58)
A lot of mental effort	3.38 (1.19)	4.00 (2.24)	3.86 (1.35)	3.57 (1.51)	2.25 (1.50)
Easy to understand	6.13 (1.13)	6.00 (1.00)	5.57 (0.53)	4.86 (0.69)	5.00 (2.45)
Easy to become skillful	6.38 (0.52)	5.29 (1.70)	5.29 (0.76)	5.00 (1.00)	6.25 (0.96)
Easy to use	6.25 (0.71)	5.86 (1.21)	5.57 (0.53)	5.43 (1.13)	6.00 (1.15)

Adolescents

Passive Sensing Data

The overall concept of using mobile sensors to passively capture data about their behavior was intriguing to all participants. Adolescents felt that the passive sensing data would be more accurate than their own self-reports in reflecting their sleep and fitness habits and that this technology could improve the efficiency of their therapy sessions by allowing them to easily share behavioral data with their clinician. Adolescents also thought that Vira would improve the quality of their treatment and support critical aspects of it (TAM perceived usefulness items 2 and 4; refer to Table 2).

Adolescents across both phases had mixed reactions to the different types of data that would be collected by the Vira app via mobile sensors. Most adolescents were excited for the app to track their sleep, walking, and exercise patterns, and identified these behaviors as factors that impact mental health but that they are often not aware of or tracking on their own. They were also enthusiastic about their clinicians accessing these data through Vira and using it to support their treatment. Refer to exemplar testimonials in Textbox 1.

However, participants were more apprehensive about sharing location (GPS) data. Many were confused about how location data sharing would work and expressed that they did not want their therapist to actively track their live location (even though this is not a capability of the app). One adolescent expressed that their data-sharing preferences could vary based on their emotional state. There were also some concerns regarding how language patterns derived from keyboard data would be interpreted by clinicians.

> *I feel like people's words and conversations can be interpreted in many ways, and I don't feel like statistically, they'll be an accurate representation.* [ID A13, Phase 2]

XSL•FO
RenderX

Textbox 1. Passive sensing data feedback.

- *Having a record of things like sleep and physical activity can help you because these are things that sometimes we do not think twice about. But they have a huge impact on how we are feeling so to have a record and to also be able to have that type of platform to communicate with your therapist sounds really efficient.* [ID A08, Phase 1]
- *My therapist seeing that I am not sleeping well would be helpful, it will save time in therapy and allow for deeper and more helpful conversations rather than recapping.* [ID A03, Phase 1]
- *If it is about sleep and your physical activity, that is something that can affect your mental health. So, it is not like it is taking unnecessary information. It is just like more information that you may not even think to bring up in your session because it is just part of your daily life.* [ID A15, Phase 2]

Home Page Feedback

Adolescents responded positively to the app's home page interface, stating it was easy to navigate and aesthetically pleasing. Most adolescents perceived it as easy to learn and understand (TAM perceived ease of use items 2 and 8). Most adolescents also liked the sample psychoeducation articles that were displayed on the home page.

However, some adolescents expressed concern about engagement with the app given that personalized insights about connections between their mood and their passively sensed behaviors would not be available until they reported their daily mood for 10 days. Adolescents suggested that having more features available during the period before insights are available would increase app engagement.

Communicating With Clinicians Between Sessions and Nudges

Adolescents liked the idea of clinicians using Vira to send nudges to them between sessions, particularly psychoeducational videos, reminders to use coping skills or general mood boosters.

> *It would be nice to have a therapist who knew how I felt even on the days that we didn't meet and could send me a little something.* [ID A02, Phase 1]

However, there were a few adolescents who expressed concern about the potential intrusiveness of between-session contact. They were also concerned with the nuisance of receiving too many smartphone notifications.

Several adolescents in Phase 2 felt that the frequency of between-session communication should be variable based on a few factors, including the strength of the relationship and their current symptom severity. Adolescents expressed that they would feel more comfortable engaging in more frequent communication with their therapist between sessions only once they had built a good rapport and relationship with them.

Clinicians

Using Vira in the Clinical Workflow

Clinicians consistently expressed that they are seeking a tool that facilitates their workflow and makes their job easier, not harder. Clinicians explained that they have many responsibilities and take on multiple roles (eg, direct service provider and clinic administrator), and thus need a tool that will lend itself to more efficient preparation before client sessions.

Clinicians interviewed in Phase 1 felt the Vira Pro interface was overwhelming, with potentially too much data provided. One clinician felt that using Vira could slow down their workflow.

> *It would make me accomplish tasks more slowly because there's so much more data.* [ID C02, Phase 1]

Between the Phase 1 and Phase 2 tests, changes were made to the prototypes to simplify the data display in Vira Pro. Yet, clinicians interviewed in Phase 2 also felt the large volume of data was challenging, and like clinicians interviewed in Phase 1, felt that trying to extract clinically relevant components and apply them to their sessions could create extra work.

> *I guess I have a little bit of a worry of how do I incorporate all this data in a way that doesn't feel like a waste of time for the patient but isn't creating a lot of extra work for me compared to what I'm doing today?* [ID C11, Phase 2]

Some clinicians suggested that more concise data summarization within Vira Pro would make the tool more usable. Despite their comments about the interface being overwhelming, clinicians across both phases felt that Vira would be easy to learn and use (TAM perceived ease of use items 2 and 10).

Although all clinicians across both phases felt that Vira would be useful to their clients' treatment (TAM perceived usefulness item 10), there were mixed opinions on whether Vira would increase their efficiency and job performance. The majority of clinicians felt that Vira would support critical aspects of their job and enhance their effectiveness on the job, but there were differences in perceptions of whether using Vira would allow them to accomplish more tasks quickly and increase their job productivity (TAM perceived usefulness items 3 and 5). Interestingly, fewer of the Phase 2 clinicians felt that Vira would allow them to accomplish more in their job than would otherwise be possible compared to Phase 1 clinicians (TAM perceived usefulness item 7).

Nudges

Clinicians had mixed opinions on the utility of scheduling nudges to practice therapy techniques. Some reacted very positively to the feature and felt that it could be useful to their patient's treatment by reinforcing the use of therapy principles discussed in the session in the context of the patient's everyday life.

> *I love the idea of the nudges, and I'm thinking how instrumental that would be in transforming phone*

XSL•FO
RenderX

> *coaching and generalizing any sort of DBT skills to the environment.* [ID C03, Phase 2]

However, other clinicians felt that nudges between sessions may lead patients to become too reliant on their therapist and neglect to engage in therapeutic strategies on their own, a key step in sustained behavior change. One clinician felt that it was unnecessary to use nudges to remind patients to use therapeutic techniques, preferring instead to simply speak to the patient in person about using them.

> *And then nudging, I'm not particularly fond of, because I feel like it could potentially reinforce passivity for a patient. If a provider is nudging them and reminding them to do things, it's kind of taking away the autonomy that they can do it on their own.* [ID C04, Phase 2]

> *I'm just imagining it's us sitting there monitoring our patients and nudging them to use skills [...] and it fragilizes them and doesn't hold them accountable for being responsible for their own well-being.* [ID, C02, Phase 2]

Within the Vira Pro dashboard, all clinicians agreed that the section showing "default" nudges should be adapted for each client's clinical needs. For example, the most common therapeutic modalities and techniques used for a patient with anxiety are different from those used for a patient with depression.

> *I would want to be able to customize it for the client because I think it would make a difference based on what I was working with the client on. [...] So they might vary depending on the focus of the treatment.* [ID C09, Phase 2]

EHR Integration

Clinicians offered several suggestions related to integrating the Vira platform into EHR systems. They frequently stated that having the data that is collected in Vira (eg, passive sensing data and assessment responses) automatically populate into their EHR would reduce the burden of manually uploading data. Similarly, clinicians thought that it would be efficient to have clinical notes generated in Vira appear in their EHR system, reducing the need for them to rewrite their notes in multiple locations.

> *Minimal effort on my end to change anything. Minimal training. Nothing that I have to do too much work on my end to upload [Vira data].* [ID C04, Phase 1]

To support seamless integration, several clinicians felt it would be important to reduce the number of redundant logins needed to access Vira within their EHR. Clinicians interviewed in Phase 2 considered the following types of Vira data most important for EHR integration: daily self-reported mood check-ins, assessment responses and scores, and the Vira-derived associations between self-reported mood and client behavior.

Business Gatekeepers

Overview

All gatekeepers felt that Vira would be a useful tool for their team's treatment toolkit (TAM perceived usefulness item 10). All gatekeepers felt that using Vira would improve the quality of the work they do (TAM perceived usefulness item 2), support critical aspects of their job (TAM perceived usefulness item 4), and enhance their effectiveness on the job (TAM perceived usefulness item 8). They also agreed that Vira seems easy to learn (TAM perceived ease of use item 2), easy to become skillful at using (TAM perceived ease of use item 9), and easy to use (TAM perceived ease of use item 10).

Vira Pro Interface

Through the user testing sessions, administrators provided feedback about the usability and acceptability of several aspects of the Vira Pro administrator dashboard. Gatekeepers responded positively to the visual layout and liked that they could access information about all patients in their clinic from a single page within the administrator dashboard. They also expressed that it would be important for the results of patient's risk assessments to be displayed within the dashboard alongside the other assessments that have been completed so that they are easily accessible for practitioners.

There were several aspects of the practitioner platform that gatekeepers were unclear about. Within the administrator dashboard, it was unclear to them where the data for "Top 10 Used Intervention Techniques" is derived from and that it is a compilation of all the clients within the clinic. Gatekeepers also sought more detailed explanations of the Vira metrics within the dashboard, such as how language categories are defined and examples.

> *I would love to see how these categories of language are defined in really clear ways. Like 'rigid words' for example, I'd love to see some examples.* [ID G2, Phase 1]

Implementation Considerations

During the interviews, gatekeepers were asked to consider several aspects of implementing Vira within their clinic. First, gatekeepers felt that Vira would be helpful because it could inform clinic treatment planning.

Gatekeepers were asked which clinicians would find Vira most helpful. All participants stated that Vira would be a useful tool for clinicians who employ cognitive behavioral therapy and DBT within their treatment framework, as Vira's functionalities could enhance their ability to practice those treatment modalities.

> *Folks who are involved in DBT and CBT could really be the ones that could utilize the tool because it most mimics what we already do, right? CBT, it's all about the exposures so how could we use Vira to help with the exposures. With DBT, it's all about practicing those skills and some of the skills are about sleeping well, taking your medication, exercising or just doing any sort of physical activity, all those things. So again,*

XSL•FO
RenderX

> *I think in some ways, this would really complement those types of providers.* [ID G2, Phase 1]

Gatekeepers felt Vira might be useful for patients experiencing depression and/or anxiety. One gatekeeper mentioned Vira potentially being useful for patients with physical health comorbidities due to Vira's health behavior metrics and expressed interest in integrating Vira with primary care or other medical departments.

Gatekeepers also brought up several potential barriers to implementation. The most common concern was the confidentiality of patient data collected from Vira. There were also some concerns about the extent of parental involvement, including how much Vira data they could access if patients are minors. Some gatekeepers were also worried that implementing Vira could be potentially burdensome for providers by adding to their workload, depending on clinic expectations.

> *I do worry that it would be burdensome to the provider if there's any expectations around how much they're supposed to be babysitting, so to speak.* [ID G3, Phase 1]

Another implementation roadblock gatekeepers identified is getting clinicians to try a new tool and do something different in their treatment. One gatekeeper suggested incorporating multimodal training to help clinicians transition to using the tool could help overcome this obstacle. Gatekeepers also foresaw difficulties in integrating Vira with existing EHR systems.

Discussion

Principal Findings

This study used UCD methods to garner feedback about the Vira behavior change platform from 3 groups of key stakeholders: adolescent patients, clinicians, and business gatekeepers. All participants expressed that Vira could be useful in enhancing mental health treatment. Adolescents had positive feedback on the aesthetics and design of the Vira app, factors that may promote engagement with mental health apps for youth at risk of suicide [36]. Overall, clinicians felt that the displays in the Vira Pro practitioner portal were too dense due to the volume of data shown and had mixed feedback about how using Vira would impact their efficiency and productivity. Clinicians were generally open to the idea of using Vira to send nudges to clients between sessions but expressed some concerns. Gatekeepers had mixed feedback on the usability of the tool but felt that Vira would enhance their effectiveness on the job and shared valuable feedback to inform implementation.

General Capabilities and Data Sharing

All participants expressed excitement about Vira's passive sensing capabilities to track behavior and the daily mood. This aligns with UCD research conducted as part of the development of the BlueIce app, which similarly found that adolescents were enthusiastic about tracking mood patterns and noticing their patterns, as well as identifying triggers of negative moods, similar to Vira's insights [37]. However, there was some apprehension about potential intrusiveness to patients. Specifically, adolescents were most worried about sharing location data and not wanting to be actively tracked by their therapist, and clinicians were only interested in accessing information that adolescents were enthusiastically willing to share with them. To address adolescents' concerns about collecting and sharing location and keyboard data, it will be important to provide them with detailed information and instructions on how their data is processed, stored, and used by the app and their clinician at the time of installation. Specifically, clients should be informed that clinicians do not have real-time access to behavioral data, nor do they have specific details about exact locations or words typed into their smartphones. Taken together, these findings suggest that clear alignment between patients and providers on which data they are comfortable sharing, and what exactly is shared, is important for aiding clinicians and clients in leveraging passive sensing data in therapy and maintaining supportive therapeutic relationships.

Communication Between Sessions (Nudges)

The Vira tool allows clinicians to send prescheduled notifications to patients to support compliance with their therapy plans. By collecting passive sensing data that is displayed in both the Vira app and on the practitioner dashboard, Vira also provides information about client behavior continuously between sessions. In this study, adolescents and clinicians expressed mixed views on whether these features would improve treatment. While some adolescents were enthusiastic, others noted that it might be frustrating to receive too many notifications from their therapist. This aligns with feedback from target adolescent end users of the LifeBuoy app regarding the frequency of app notifications [36]. Several clinicians in this study shared similar concerns.

Stakeholder feedback from this study therefore suggests it is essential to set clear expectations with both adolescents and practitioners about how that between-session communication will be operationalized. Gatekeepers felt the tool would be a supportive addition to their team's treatment toolkit, but were concerned about the potential burden on clinicians, underscoring the importance of clearly defining expectations with providers. Along similar lines, when introducing the tool to their patients, practitioners should discuss both the frequency with which they will use Vira and how they will use it to support treatment. Similar to Hetrick et al [38], clinicians in our study expressed that although the tool could be useful in monitoring patients' moods, monitoring patient data multiple times between sessions or in real time is not feasible. However, to be clear, this is not an expectation of typical use of the Vira Platform. Ensuring the tool is designed to fit within clinician workflows and managing patients' expectations for how their clinician will use it increases the likelihood of clinician adoption [39].

Adolescents were very enthusiastic about Vira's nudges. They uniformly liked the concept of clinicians using the passive sensing data to determine what types of nudges should be sent to enhance their treatment. They also felt it would be helpful for clinicians to send nudges as reminders to use therapeutic techniques learned during treatment. Conversely, clinicians had mixed opinions on the nudging feature. Some clinicians felt nudges could be very useful and could transform treatment in a positive way, consistent with recent research demonstrating

XSL•FO
RenderX

that nudges and prompts increase engagement in self-guided interventions [40]. However, it is important to note that patients in the IOP setting would be using Vira to support DBT, a collaborative and therapist-guided treatment modality. Other clinicians expressed concern that patients may become dependent on their clinician to send nudges to remember to use therapeutic techniques discussed in sessions. In sum, Vira's nudge capability may help promote habit formation in terms of therapy techniques and coping skills outside of the therapy sessions, and it is important that digital products allow clinicians to schedule and structure nudges in line with each patient's personalized and dynamic treatment needs. For example, clinicians may schedule nudges to be delivered less frequently as skills are acquired and used.

Product Changes

Findings from this iterative UCD process directly informed the development of several changes to the Vira platform that will be implemented for the pragmatic RCT. In response to feedback about data sharing, app users will have the flexibility to choose which data streams they would like to share with their practitioner, including specific streams of passively sensed behavioral data and daily mood (enjoyment) ratings. Enhancements have also been made to the app to boost engagement during the initial 10-day period in which Vira collects sufficient information to calculate personalized insights about behaviors that may be related to a user's mood. In the version of the app that was tested in this study, personalized insights were not available to users until their 10th daily check-in. Based on insights from this UCD study, the app has been updated such that users receive more onboarding information within the app after their third daily check-in, and insights about the consistency of their sleep and walking patterns after their sixth and eighth check-ins, respectively.

Clinicians and gatekeepers expressed strong desires for deeper information about the various sensor streams and data types available within the dashboard, as well as a simpler and more accessible user interface. Therefore, the Vira Pro dashboard was modified to show a more summarized, simplified form of patient data with more options to hide granularity in data charts. The revised version of the platform also retains the capability to drill into the details as needed; Vira Pro users can now hover over fields within the dashboard, such as sensor streams and specific data points, to view more details about how the data were collected and processed. Simplifying the user interface is a particularly important product change, considering that providing an overload of information can discourage clinicians from using the digital tool [41].

Implementation Considerations

Previous research has shown that practitioners' negative attitudes toward digital mental health interventions and resistance to change are common barriers to implementing new digital technologies in mental health care [39]. Similarly, through clinician and gatekeeper interviews, this study found that clinician buy-in with using a new tool is vital. Consistent with other research, clinicians and gatekeepers expressed concerns about the potential for mobile technology to increase professional workload [42]. Providing clinicians with sufficient training to ensure their confidence and efficiency in using the tool may increase the likelihood that clinicians integrate the tools into their practice [43], and help them overcome any concerns or hesitations they may have regarding implementation. Thorough training was also commonly suggested by gatekeepers in this study as a vital step to implementing Vira in their clinic. Further, when implementing Vira within a clinical setting, clinic administration needs to set clear expectations for how they expect Vira to be used within treatment. Moreover, clinicians and gatekeepers expressed that friction integrating Vira with existing EHR systems would be a barrier to implementation, consistent with previous research [44].

Findings from this study directly informed the development of a focused training approach that is being rolled out in the ongoing pragmatic RCT [29]. In response to clinician feedback about having limited time to learn new tools, the clinician training uses a blended synchronous and asynchronous format to provide flexibility while also keeping the training focused and efficient (1 hour total). To address mixed views regarding whether Vira would help them increase productivity, the training emphasizes critical thinking on how to apply the Vira platform to their practice. The training sequence focuses on building self-efficacy for using the platform to enhance their practice and help augment their ability to complete existing tasks. Many video and written resources are also available through a self-service knowledge base for asynchronous review and support after the training.

Generalizability and Broader Implications

Findings from this study can be generalized to inform the development and implementation of other digital technologies for high-risk youth. Both adolescents and clinicians were enthusiastic about the potential utility of passive sensing data for treatments [37], but concerns regarding data sharing underscore the need for transparency about how data from digital health platforms are used [45].

This study also revealed important issues to consider more broadly when implementing new digital technologies into health care systems, such as the impact on clinician workloads and integrating the tool within existing clinical workflows and EHRs [39,44]. User interface design is a key consideration in increasing the acceptability of new digital technologies. Data displays should be simple and accessible so as to not overwhelm users and increase the likelihood of adoption [41].

Limitations

This study has several limitations. Sample characteristics and recruitment methods may have impacted the feedback elicited in this study and the generalizability of the results. Many adolescents in the study had engaged in higher levels of care (eg, inpatient admission, partial hospitalization programs, and IOPs). However, the inclusion criteria for adolescent participants only required them to be currently in treatment or have received treatment in the past year for depressive or anxiety symptoms. Accordingly, some of the participants recruited may have only received a minimal level of care (ie, just a few sessions) and may have been at lower risk compared to the intended user population for Vira. In addition, the sample was limited in terms

XSL•FO
RenderX

of racial and ethnic diversity, and the majority of the adolescent sample and the entire clinician sample were female. This may be because females are more likely to seek treatment for mental health [46], and the mental health workforce is predominately female [47]. Moreover, there was also a relatively small sample size. For grounded theory approaches, a sample size of 20-30 participants is generally recommended [48], although a recent systematic review found that among studies with a relatively homogeneous study population and specific objectives, a total of 9-17 interviews may be enough to reach saturation [49]. This study reached saturation with a similar sample size. Finally, we did not include parents as a stakeholder in this study. It may be useful to include parents in future studies to understand how their involvement can facilitate the implementation of digital technologies in adolescent mental health treatment.

Future Directions

Clinician feedback about the lack of time to prepare for sessions suggests that automation may increase the acceptability and perceived usefulness of the platform. Feedback has informed the ongoing design and development of AI-supported features such as automatically generated drafts of clinical progress notes, “heads up” summaries of the most salient clinician issues that can be quickly accessed before each session, and a clinical assistant that can leverage clinical practice guidelines across a range of behavioral health conditions to suggest enhancements to therapy plans based on data collected in Vira. Moreover, the clinical effectiveness of the Vira platform is currently being investigated in a 2-arm, pragmatic RCT in which youth at risk for suicide within an IOP will be randomly assigned to use Vira with their clinician or engage in usual care [29]. Throughout this clinical trial, we will collect feedback from both patient and clinician participants.

Conclusions

In sum, participants felt that the Vira platform has a high potential to improve engagement with mental health treatment as well as mental health outcomes in high-risk youth. Our rigorous, multimethod approach uncovered key facilitators and barriers to implementation, many of which are consistent with previous research, highlighting the challenge of translating novel, technology-based interventions from controlled to clinical settings. Applying iterative, UCD methods allowed us to integrate valuable stakeholder feedback to develop an intervention that is clinically relevant to end users, hopefully leading to successful implementation in real-world settings.

Acknowledgments

Funding from the National Institute of Mental Health provided support to the NBA (R44 MH128484-01A1). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

RPA and NBA designed the study. TCT and DD performed data curation. TCT, LSW, and DD performed formal data analysis. TCT, LSW, and CJF drafted the paper. TCT, LSW, DD, RPA, and NBA interpreted results. All authors read and approved the final version of the paper and agreed with the order of presentation of the authors.

Conflicts of Interest

LSW and NBA are employed by and hold equity in, Ksana Health, which developed and commercialized the Vira platform. DD was employed by Ksana Health at the time the project was initiated. TCT, CJF, CCC, ZKB, and CTO have no conflicts to disclose. RPA is an unpaid advisor to Ksana Health and a paid advisor for Get Sonar, Inc. Ksana Health has a Scientific Integrity Policy designed to prevent the company’s interest in bringing Ksana Health products to market from influencing or appearing to influence the science conducted using Ksana Health’s products. Ksana Health follows Open Science practices to ensure the transparency, integrity, and reproducibility of our work. This includes preregistration before data collection, and data, code, and material sharing upon request.

Multimedia Appendix 1
Additional figures.
[DOCX File , 1415 KB - formative_v9i1e65418_app1.docx]

References

1. Arango A, Gipson PY, Votta JG, King CA. Saving lives: recognizing and intervening with youth at risk for suicide. Annu Rev Clin Psychol 2021;17:259-284. [doi: 10.1146/annurev-clinpsy-081219-103740] [Medline: 33544628]
2. Centers for Disease Control and Prevention. 2021. URL: https://stacks.cdc.gov/view/cdc/124928 [accessed 2023-02-07]
3. Gaylor EM, Krause KH, Welder LE, Cooper AC, Ashley C, Mack KA, et al. Suicidal thoughts and behaviors among high school students - youth risk behavior survey, United States, 2021. MMWR Suppl 2023;72(1):45-54 [FREE Full text] [doi: 10.15585/mmwr.su7201a6] [Medline: 37104546]

XSL•FO
RenderX

4. Weisz JR, Venturo-Conerly KE, Fitzpatrick OM, Frederick JA, Ng MY. What four decades of meta-analysis have taught us about youth psychotherapy and the science of research synthesis. Annu Rev Clin Psychol 2023;19:79-105 [FREE Full text] [doi: 10.1146/annurev-clinpsy-080921-082920] [Medline: 36750262]
5. Grist R, Porter J, Stallard P. Mental health mobile apps for preadolescents and adolescents: a systematic review. J Med Internet Res 2017;19(5):e176 [FREE Full text] [doi: 10.2196/jmir.7332] [Medline: 28546138]
6. Lehtimaki S, Martic J, Wahl B, Foster KT, Schwalbe N. Evidence on digital mental health interventions for adolescents and young people: systematic overview. JMIR Ment Health 2021;8(4):e25847 [FREE Full text] [doi: 10.2196/25847] [Medline: 33913817]
7. Kennard BD, Goldstein T, Foxwell AA, McMakin DL, Wolfe K, Biernesser C, et al. As safe as possible (ASAP): a brief app-supported inpatient intervention to prevent postdischarge suicidal behavior in Hospitalized, Suicidal Adolescents. Am J Psychiatry 2018;175(9):864-872 [FREE Full text] [doi: 10.1176/appi.ajp.2018.17101151] [Medline: 30021457]
8. Torok M, Han J, McGillivray L, Wong Q, Werner-Seidler A, O'Dea B, et al. The effect of a therapeutic smartphone application on suicidal ideation in young adults: findings from a randomized controlled trial in Australia. PLoS Med 2022;19(5):e1003978 [FREE Full text] [doi: 10.1371/journal.pmed.1003978] [Medline: 35639672]
9. Stallard P, Porter J, Grist R. A smartphone app (BlueIce) for young people who self-harm: open phase 1 pre-post trial. JMIR Mhealth Uhealth 2018;6(1):e32 [FREE Full text] [doi: 10.2196/mhealth.8917] [Medline: 29382632]
10. Malakouti SK, Rasouli N, Rezaeian M, Nojomi M, Ghanbari B, Shahraki Mohammadi A. Effectiveness of self-help mobile telephone applications (apps) for suicide prevention: a systematic review. Med J Islam Repub Iran 2020;34:85 [FREE Full text] [doi: 10.34171/mjiri.34.85] [Medline: 33306048]
11. Kothgassner OD, Goreis A, Robinson K, Huscsava MM, Schmahl C, Plener PL. Efficacy of dialectical behavior therapy for adolescent self-harm and suicidal ideation: a systematic review and meta-analysis. Psychol Med 2021;51(7):1057-1067 [FREE Full text] [doi: 10.1017/S0033291721001355] [Medline: 33875025]
12. Lambert MJ, Whipple JL, Kleinstäuber M. Collecting and delivering progress feedback: a meta-analysis of routine outcome monitoring. Psychotherapy (Chic) 2018;55(4):520-537. [doi: 10.1037/pst0000167] [Medline: 30335463]
13. Lewis CC, Boyd M, Puspitasari A, Navarro E, Howard J, Kassab H, et al. Implementing measurement-based care in behavioral health: a review. JAMA Psychiatry 2019;76(3):324-335 [FREE Full text] [doi: 10.1001/jamapsychiatry.2018.3329] [Medline: 30566197]
14. Bunnell BE, Nemeth LS, Lenert LA, Kazantzis N, Deblinger E, Higgins KA, et al. Barriers associated with the implementation of homework in youth mental health treatment and potential mobile health solutions. Cognit Ther Res 2021;45(2):272-286 [FREE Full text] [doi: 10.1007/s10608-020-10090-8] [Medline: 34108776]
15. Kazantzis N, Whittington C, Zelencich L, Kyrios M, Norton PJ, Hofmann SG. Quantity and quality of homework compliance: a meta-analysis of relations with outcome in cognitive behavior therapy. Behav Ther 2016;47(5):755-772. [doi: 10.1016/j.beth.2016.05.002] [Medline: 27816086]
16. Tang W, Kreindler D. Supporting homework compliance in cognitive behavioural therapy: essential features of mobile apps. JMIR Ment Health 2017;4(2):e20 [FREE Full text] [doi: 10.2196/mental.5283] [Medline: 28596145]
17. Lattie EG, Adkins EC, Winquist N, Stiles-Shields C, Wafford QE, Graham AK. Digital mental health interventions for depression, anxiety, and enhancement of psychological well-being among college students: systematic review. J Med Internet Res 2019;21(7):e12869 [FREE Full text] [doi: 10.2196/12869] [Medline: 31333198]
18. Torous J, Nicholas J, Larsen ME, Firth J, Christensen H. Clinical review of user engagement with mental health smartphone apps: evidence, theory and improvements. Evid Based Ment Health 2018;21(3):116-119 [FREE Full text] [doi: 10.1136/eb-2018-102891] [Medline: 29871870]
19. Kruzan KP, Meyerhoff J, Biernesser C, Goldstein T, Reddy M, Mohr DC. Centering lived experience in developing digital interventions for suicide and self-injurious behaviors: user-centered design approach. JMIR Ment Health 2021;8(12):e31367 [FREE Full text] [doi: 10.2196/31367] [Medline: 34951602]
20. Blandford A, Gibbs J, Newhouse N, Perski O, Singh A, Murray E. Seven lessons for interdisciplinary research on interactive digital health interventions. Digit Health 2018;4:2055207618770325 [FREE Full text] [doi: 10.1177/2055207618770325] [Medline: 29942629]
21. Chokshi SK, Mann DM. Innovating from within: a process model for user-centered digital development in academic medical centers. JMIR Hum Factors 2018;5(4):e11048 [FREE Full text] [doi: 10.2196/11048] [Medline: 30567688]
22. Bhattacharyya O, Mossman K, Gustafsson L, Schneider EC. Using human-centered design to build a digital health advisor for patients with complex needs: persona and prototype development. J Med Internet Res 2019;21(5):e10318 [FREE Full text] [doi: 10.2196/10318] [Medline: 31094334]
23. Brouwer W, Kroeze W, Crutzen R, de Nooijer J, de Vries NK, Brug J, et al. Which intervention characteristics are related to more exposure to internet-delivered healthy lifestyle promotion interventions? A systematic review. J Med Internet Res 2011;13(1):e2 [FREE Full text] [doi: 10.2196/jmir.1639] [Medline: 21212045]
24. de Beurs D, van Bruinessen I, Noordman J, Friele R, van Dulmen S. Active involvement of end users when developing web-based mental health interventions. Front Psychiatry 2017;8:72 [FREE Full text] [doi: 10.3389/fpsyt.2017.00072] [Medline: 28515699]

XSL•FO
RenderX

25. Lyon AR, Koerner K. User-centered design for psychosocial intervention development and implementation. Clin Psychol (New York) 2016;23(2):180-200 [FREE Full text] [doi: 10.1111/cpsp.12154] [Medline: 29456295]
26. Borghouts J, Eikey E, Mark G, De Leon C, Schueller SM, Schneider M, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. J Med Internet Res 2021;23(3):e24387 [FREE Full text] [doi: 10.2196/24387] [Medline: 33759801]
27. Stein AT, Carl E, Cuijpers P, Karyotaki E, Smits JAJ. Looking beyond depression: a meta-analysis of the effect of behavioral activation on depression, anxiety, and activation. Psychol Med 2021;51(9):1491-1504. [doi: 10.1017/S0033291720000239] [Medline: 32138802]
28. Smyth, Heron. Is providing mobile interventions "just-in-time" helpful? an experimental proof of concept study of just-in-time intervention for stress management. : IEEE; 2016 Presented at: IEEE Wireless Health (WH); 2016 October 25-27; Bethesda, MD, USA p. 1-7. [doi: 10.1109/wh.2016.7764561]
29. Funkhouser CJ, Tse TC, Weiner LS, deLuise D, Pagliaccio D, Durham K, et al. Study preregistration: testing a digital suicide risk reduction platform for adolescents: a pragmatic randomized controlled trial. J Am Acad Child Adolesc Psychiatry 2024;63(8):845-847. [doi: 10.1016/j.jaac.2024.03.012] [Medline: 39069372]
30. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly 1989;13(3):319. [doi: 10.2307/249008]
31. Ericsson A, Simon HA. Protocol Analysis: Verbal Reports as Data. Cambridge, Massachusetts: MIT Press; 1993.
32. Davison GC, Vogel RS, Coffman SG. Think-aloud approaches to cognitive assessment and the articulated thoughts in simulated situations paradigm. J Consult Clin Psychol 1997;65(6):950-958. [doi: 10.1037//0022-006x.65.6.950] [Medline: 9420356]
33. Guest G, MacQueen KM, Namey EE. Applied Thematic Analysis. Thousand Oaks, California: SAGE Publications; 2012.
34. Averill JB. Matrix analysis as a complementary analytic strategy in qualitative inquiry. Qual Health Res 2002;12(6):855-866. [doi: 10.1177/10432302012006011]
35. Beebe JR. Rapid Qualitative Inquiry: A Field Guide to Team-Based Assessment. Lanham, Maryland: Rowman & Littlefield; 2014.
36. Gan DZQ, McGillivray L, Larsen ME, Bloomfield T, Torok M. Promoting engagement with smartphone apps for suicidal ideation in young people: development of an adjunctive strategy using a lived experience participatory design approach. JMIR Form Res 2023;7:e45234 [FREE Full text] [doi: 10.2196/45234] [Medline: 37279058]
37. Grist R, Porter J, Stallard P. Acceptability, use, and safety of a mobile phone app (BlueIce) for young people who self-harm: qualitative study of service users' experience. JMIR Ment Health 2018;5(1):e16 [FREE Full text] [doi: 10.2196/mental.8779] [Medline: 29475823]
38. Hetrick SE, Robinson J, Burge E, Blandon R, Mobilio B, Rice SM, et al. Youth codesign of a mobile phone app to facilitate self-monitoring and management of mood symptoms in young people with major depression, suicidal ideation, and self-harm. JMIR Ment Health 2018;5(1):e9 [FREE Full text] [doi: 10.2196/mental.9041] [Medline: 29362208]
39. Graham AK, Lattie EG, Powell BJ, Lyon AR, Smith JD, Schueller SM, et al. Implementation strategies for digital mental health interventions in health care settings. Am Psychol 2020;75(8):1080-1092 [FREE Full text] [doi: 10.1037/amp0000686] [Medline: 33252946]
40. van Mierlo T, Rondina R, Fournier R. Nudges and prompts increase engagement in self-guided digital health treatment for depression and anxiety: results from a 3-arm randomized controlled trial. JMIR Form Res 2024;8:e52558 [FREE Full text] [doi: 10.2196/52558] [Medline: 38592752]
41. Jacob C, Sanchez-Vazquez A, Ivory C. Social, organizational, and technological factors impacting clinicians' adoption of mobile health tools: systematic literature review. JMIR Mhealth Uhealth 2020;8(2):e15935 [FREE Full text] [doi: 10.2196/15935] [Medline: 32130167]
42. Bishop TF, Press MJ, Mendelsohn JL, Casalino LP. Electronic communication improves access, but barriers to its widespread adoption remain. Health Aff (Millwood) 2013;32(8):1361-1367 [FREE Full text] [doi: 10.1377/hlthaff.2012.1151] [Medline: 23918479]
43. Armstrong CM. Mobile health provider training: results and lessons learned from year four of training on core competencies for mobile health in clinical care. J. technol. behav. sci 2019;4(2):86-92. [doi: 10.1007/s41347-019-00089-8]
44. Graham A, Greene C, Powell T, Lieponis P, Lunsford A, Peralta CD, et al. Lessons learned from service design of a trial of a digital mental health service: informing implementation in primary care clinics. Transl Behav Med 2020;10(3):598-605 [FREE Full text] [doi: 10.1093/tbm/ibz140] [Medline: 32766862]
45. Bessenyei K, Suruliraj B, Bagnell A, McGrath P, Wozney L, Huguet A, et al. Comfortability with the passive collection of smartphone data for monitoring of mental health: an online survey. Computers in Human Behavior Reports 2021;4:100134. [doi: 10.1016/j.chbr.2021.100134]
46. Wang J, Qiu Y, Zhu X. Trends of mental health care utilization among US adults from 1999 to 2018. BMC Psychiatry 2023;23(1):665 [FREE Full text] [doi: 10.1186/s12888-023-05156-2] [Medline: 37700243]
47. Data tool: demographics of the U.S. psychology workforce. American Psychological Association. 2022. URL: https://www.apa.org/workforce/data-tools/demographics [accessed 2024-05-09]
48. Creswell JW. Concise Introduction to Mixed Methods Research. Thousand Oaks, California: Sage; 2022.

XSL•FO
RenderX

49. Hennink M, Kaiser BN. Sample sizes for saturation in qualitative research: a systematic review of empirical tests. Soc Sci Med 2022;292:114523 [FREE Full text] [doi: 10.1016/j.socscimed.2021.114523] [Medline: 34785096]

Abbreviations

DBT: dialectical behavior therapy
EHR: electronic health record
HIPAA: Health Insurance Portability and Accountability Act
IOP: intensive outpatient program
RCT: randomized controlled trial
TAM: technology acceptance model
UCD: user-centered design

Edited by A Mavragani; submitted 20.08.24; peer-reviewed by L Magoun, M Klapow; comments to author 15.12.24; revised version received 23.12.24; accepted 07.01.25; published 02.05.25.

Please cite as:
Tse TC, Weiner LS, Funkhouser CJ, DeLuise D, Cullen C, Blumkin Z, O'Brien C, Auerbach RP, Allen NB
Acceptability and Usability of a Digital Behavioral Health Platform for Youth at Risk of Suicide: User-Centered Design Study With Patients, Practitioners, and Business Gatekeepers
JMIR Form Res 2025;9:e65418
URL: https://formative.jmir.org/2025/1/e65418
doi:10.2196/65418
PMID:

©Trinity Chloe Tse, Lauren S Weiner, Carter J Funkhouser, Danielle DeLuise, Colleen Cullen, Zachary Blumkin, Casey O'Brien, Randy P Auerbach, Nicholas B Allen. Originally published in JMIR Formative Research (https://formative.jmir.org), 02.05.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Exploring Therapists' Approaches to Treating Eating Disorders to Inform User-Centric App Design: Web-Based Interview Study

Pamela Carien Thomas[1], BA, MSc; Pippa Bark[2], BSc; Sarah Rowe[1], BA, MEd, PhD

[1]Epidemiology and Applied Clinical Research Department, University College London, London, United Kingdom

[2]UCL Cancer Insititute, London, United Kingdom

Corresponding Author:
Pamela Carien Thomas, BA, MSc
Epidemiology and Applied Clinical Research Department
University College London
4th Floor, Wings A and B, Maple House
149 Tottenham Court Road
London, W1T 7NF
United Kingdom
Phone: 44 0207 679 9000
Email: pamela.thomas.18@ucl.ac.uk

Abstract

Background: The potential for digital interventions in self-management and treatment of mild to moderate eating disorders (EDs) has already been established. However, apps are infrequently recommended by ED therapists to their clients. Those that are recommended often have poor engagement and user satisfaction, leading to unsatisfactory outcomes. Barriers to recommendation include patient safety, data privacy, and a perception that they may not be effective. Many existing interventions have limited functionality or do not differ much from manual cognitive behavioral therapy (CBT) or self-help books, which may not adequately support the therapeutic process or sustain user engagement.

Objective: This study aims to explore the perspectives of therapists who support people with mild to moderate EDs in the community, exploring their existing treatment approach and how an ED app might fit in the treatment pathway alongside treatment.

Methods: Semistructured web-based interviews were completed with ED therapists in the United Kingdom. Participants were recruited from First Steps ED, a specialist community-based ED service, and Thrive Mental Wellbeing, a workplace mental health provider. Five main themes were covered: (1) therapists' treatment approach, (2) how therapy was implemented in practice, (3) strategies for engaging and motivating clients, (4) perspectives on a potential ED app, and (5) suggestions for app content and design. A structured thematic analysis was validated by 2 researchers.

Results: Overall, 12 ED and mental health therapists (mean age 28.7, SD 7.3 y; female therapists: n=7, 58%; male therapists: n=5, 42%) participated. Therapists dealing with complex ED issues went beyond traditional CBT using additional therapeutic techniques and a flexible, person-centered approach to treatment. This included engagement and motivational strategies to support the client, elements of which could be mirrored in an app. Therapists identified the therapeutic relationship as key to success, which might have been hard to replicate in an app. They saw the potential for evidence-based apps across all stages of the treatment pathway. The need to address safeguarding, data privacy, and the potential for triggering content within the app was vital.

Conclusions: This study advanced our understanding of how to design and develop clinically safe, evidence-based ED apps that can complement therapy by extending the continuity of care and the self-management and psychoeducation of clients. It emphasized integrative, adaptive CBT that incorporated other therapeutic approaches based on individuals' needs, which could be replicated in an app, as could the strategies to support engagement and motivation. It gave a cautious yet optimistic perspective on the potential integration of apps into ED treatment across all stages of the treatment pathway, from pretreatment maintenance to posttreatment maintenance. It highlighted various concerns that could be addressed and potential limitations, such as the therapeutic relationship, while recognizing the growing potential of apps with rapid technology and artificial intelligence advancements.

(JMIR Form Res 2025;9:e68846) doi:10.2196/68846

XSL•FO
RenderX

KEYWORDS

eating disorders; binge eating; bulimia; anorexia; qualitative research; mental health; smartphone apps; digital interventions; eHealth; mobile health; artificial intelligence

Introduction

Background

Eating disorders (EDs) are complex mental health conditions characterized by severe disturbances in eating behaviors and related distressing thoughts and emotions. They can have a significant impact on individuals' psychological, physical, and emotional well-being and are related to serious health consequences, such as mortality, infertility, osteoporosis, as well as a reduced quality of life [1,2]. Most people with mild to moderate EDs are not getting help due to the stigma and shame associated with their condition [1] and a high demand for health services, which has escalated since the COVID-19 pandemic [3]. Those with milder forms of EDs are less likely to receive support as they are considered at a lower risk, hence lower priority, despite the importance of early intervention [4]. Those who are deemed ineligible for public health care or who are on a waiting list for treatment can seek support from a private health care service provider or an ED patient group, such as Beat or First Steps ED in the United Kingdom. However, barriers such as cost, lack of awareness, or reluctance to talk to a person remain [1], and these services have limited capacity.

Apps are a promising tool to support the management of people with mild to moderate EDs by overcoming some of these barriers to treatment [5]. Apps need to be trustworthy, safe, and effective [6,7]; hence, evidence-based design and clinical involvement in their development are crucial. Fairburn and Rothwell [8] stressed the importance of professional involvement in translating evidence-based treatment, in particular cognitive behavioral therapy (CBT), into digital interventions to ensure therapeutic integrity of the digital format [9], although this has rarely been the case with ED apps.

Systematic reviews of randomized controlled studies have provided robust evidence for the effectiveness of digital interventions in reducing ED symptoms in the short term, with some studies demonstrating effectiveness at long-term follow-up. However, evidence for apps is still limited [10], with high dropout rates [11,12] and significant engagement challenges (eg, uptake, ongoing use, and attrition) making the suitability of apps in real-world settings particularly relevant. Evidence from qualitative studies suggests that, while these interventions have received some positive feedback from potential users [13,14] there is significant room for improvement to make them more appealing and engaging. Clinicians have also expressed concerns and barriers to recommending apps, including a lack of personalization, patient safety, and data security and privacy [13]; this has resulted in limited uptake [5]. Many existing interventions differ little from written self-help programs [14] or simple digitized versions of manual CBT [15]. Intervention design should be based on what works in traditional therapy while capitalizing on the advantages offered by a digital format [7,14] to support engagement and effectiveness.

One recommended approach to improve engagement and effectiveness of apps is to conduct qualitative research with clinicians to inform intervention planning, design, and development [16-19]. Few qualitative studies have explored the in-depth views and experiences of therapists regarding apps, and they have been typically involved at a final evaluation stage of the development process [20,21]. One recent exploration on the perspective of German clinicians on digital ED interventions and their integration into routine care [13] focused on barriers and enablers to their use. It was positive about the potential to improve access to treatment, provide beneficial monitoring and engagement tools, and act as a useful adjunct to therapy. However, clinicians expressed concerns, including losing the benefits of the therapeutic relationship with their clients; how patient data might be used, stored, and accessed; and whether apps could be personalized enough to meet individual patient's needs, especially in the more severe cases. The study did not explore how to design digital ED interventions that were acceptable to clinicians, how these barriers could be overcome, or how they could be made more engaging for users. It also did not focus on apps specifically.

Our study explored the views of therapists working in a community setting and how they delivered effective treatment to people with mild to moderate EDs. This study aimed to bridge the gap between traditional clinical treatment approaches and apps with a focus on how these methods can inform app design. CBT is the "gold standard" for eating treatment (National Centre for Clinical Excellence) and frequently forms the basis for app design, with some apps hardly differing from digitized manual CBT. This study challenges the assumption that digitized traditional CBT is appropriate for app development by exploring how therapists deviate from standard CBT and understanding the justifications for this to inform the development of an effective and engaging app.

Objectives

The primary objectives of this study were to understand the following: (1) how therapists currently supported people with mild to moderate EDs in the community; (2) how therapists motivated and engaged their clients in treatment; (3) where an app might fit in the treatment pathway, who it is most suited for, and what concerns needed addressing; and (4) how therapists envisaged app content and design.

Methods

Overview

In total, 12 web-based semistructured interviews, ranging from 45 to 60 minutes, were completed with ED therapists in the United Kingdom. Participants were asked about how they supported their clients with ED and their perspectives on a proposed ED app for people with mild to moderate EDs (Multimedia Appendix 1 provides the interview guide). This qualitative study was nested within a broader research program

XSL•FO
RenderX

on the development and evaluation of an ED app module for people with mild to moderate EDs.

The COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist was followed [22]. This checklist helped to clarify the procedures, analysis, and interpretation.

Participants

Participants were therapists, counselors, and a psychiatrist working with a patient group with an ED (First Steps ED) and a mental health service provider (Thrive Mental Wellbeing) across the United Kingdom. This enabled a comprehensive understanding of treatment approaches for people with mild to moderate EDs and identified potential benefits, risks, and concerns relating to apps. They had to have been in their role for at least 3 months and completed their compulsory training. The term "therapist" is used throughout this paper as a blanket term to cover the range of professional experiences.

Recruitment

We targeted therapists who worked in a community setting, as their patient population may be more reflective of those who would be most suitable for an app and for whom self-management support may be an appropriate option. Convenience sampling was used with participants selected based on their availability and willingness to participate, given therapists' time constraints. Random sampling was not required, given the aim of this study was to develop meaningful insights and further our understanding across a range of professionals, without aiming for generalizability. Given the exploratory nature of this study, the sample was not required to be representative, so a relatively small sample size was deemed acceptable [23].

Recruitment took place via First Steps ED, a specialist ED support service, and Thrive Mental Wellbeing, an industry collaborator providing mental health support via employers. To maintain rigor and relevance, team managers at both organizations facilitated the recruitment process, ensuring that only therapists with relevant ED experience or training were invited to participate. At the time of recruitment, Thrive Mental Wellbeing had approximately 25 therapists, though not all specialized in ED treatment, and First Steps ED had 18 trained ED therapists and counselors. By engaging 2 distinct yet complementary organizations, we ensured a breadth of perspectives by capturing insights from therapists working within a dedicated ED service (First Steps ED) and those delivering broader workplace-based mental health support (Thrive Mental Wellbeing).

Interviews were online and audio and video recorded (via Microsoft Teams) between January and February 2024. Field notes were taken after the session to capture important points and initial reflections.

Materials

Semistructured interview questions were used to guide the interview discussion. The interview guide was developed by the primary researcher (PCT) in consultation with a public and patient involvement group from First Steps ED comprising the research manager, clinical service manager, and 2 specialist support officers. Questions focused on the therapists' treatment approach, how they applied this in practice, the target group for the app, engagement and motivation, and the design and functionality of the app (Multimedia Appendix 1 provides the interview guide).

Data Analysis

The audio recordings were transcribed using Microsoft Teams and then anonymized and checked for accuracy by the primary researcher (PCT). The scripts were then imported into NVivo (version 12; Lumivero) [24] for analysis. The analysis followed a structured thematic approach based on the approach by Clarke and Braun [25]. The primary researcher created initial codes across all transcripts and then reviewed and consolidated these into an initial coding framework of main codes and subcodes in an iterative process. The second coder (PB) familiarized herself with 25% of the data and independently generated initial codes and subcodes iteratively. The 2 coders refined their results in several iterations until agreeing upon a coding framework against which all the interviews were coded. The primary researcher consulted with a public and patient involvement team at First Steps ED to validate the final coding framework. A codebook was generated to support the implementation of the coding framework across both coders and ensure reliability. Given the complexity of the topic and interdependency of some of the themes, it was possible for the same quote from participants to be used against more than one theme and subtheme. We did not aim for saturation, as the approach by Clarke and Braun [25] prioritized generating meaningful insights over complete data coverage. Because data were relatively homogeneous across interviews, common themes and subthemes were easily identifiable.

Reflexivity Statement

The primary researcher was a female researcher with experience and training in qualitative research methods. She reflected on any potential biases during the coding, such as an assumption that apps have the potential to address existing service gaps and help support people with EDs who may currently not be getting help. She remained receptive to the insights and experiences shared by ED therapists, including participants who were more skeptical about the role of apps in treatment, and made sure to probe all interviewees about their potential concerns regarding apps.

Ethical Considerations

This study received ethics approval from the University College London Research Ethics Committee (23943/001). The research adhered to ethical guidelines for qualitative studies involving human participants.

Informed Consent

Participants received an information sheet outlining the aims and objectives of the study. This informed them of the study's purpose, voluntary nature, and the right to withdraw at any time without consequence. They provided consent via a digital consent form in a secure electronic data platform, REDCap (Research Electronic Data Capture; Vanderbilt University) before the participation and reconfirmed their consent at the start of the session. After providing informed consent, therapists participated in one-on-one web-based interviews from their

XSL•FO
RenderX

workplace. No participants dropped out after giving their consent. Participants did not know the researcher, although there was friendly email exchange before each session to establish a rapport, clarify session objectives, and address any participant queries. Participants were aware of the researcher's role and motivation to explore how an ED app could be developed to support treatment, aligned with the study's objectives.

Privacy and Confidentiality

Baseline characteristics of participants were collected via Microsoft Forms, hosted on a secure University College London server. All study data were securely stored, and standard anonymization protocols were followed after transcription. Identifying information was removed to maintain participant confidentiality. Only the research team had access to the data, and no personally identifiable information was included in this paper or the multimedia appendices.

Compensation

Participants were not offered financial or material compensation for their involvement in the study. Their participation was motivated by a shared commitment to improving access to treatment for individuals with EDs.

Participant Identifiability

No personally identifying details of participants are included in this paper or the multimedia appendices. In cases where quotations from transcripts are presented, they have been deidentified to prevent recognition of individual participants.

Results

In total, 12 interviews were completed and lasted an average of 50 (SD 6; range 35-60) minutes.

Participant Characteristics

In total, 8 therapists, 3 counselors, and 1 psychiatrist took part. Participants were aged between 21 and 52 (mean 28.7, SD 7.3) years. Most (11/12, 92%) therapists were trained in CBT and had a range of experience and training, including dialectical behavior therapy and acceptance and commitment therapy (ACT). All were educated till the master's level (10/12, 83%) or had higher education (2/12, 17%).

Moreover, 50% (6/12) of the participants had direct experience of working alongside apps, allowing them to speak from experience in terms of how an app could work alongside treatment, and 50% (6/12) did not. Most participants were positive (10/12, 83%) that apps had the potential to be a valuable tool in the treatment process and felt that they might or were likely to recommend a person-centered app to their clients (8/12, 67%). However, 25% (3/12) of the participants agreed with the statement that they had significant concerns about their use in treatment (refer to Multimedia Appendix 2 for the baseline characteristics).

Five themes were covered. The first 3 themes provided crucial context for informing app development by providing an in-depth exploration of how treatment was delivered to help identify elements that could be prioritized and adapted. These included therapists' treatment approach, how therapy was implemented in practice, and strategies for engaging and motivating clients. The subsequent 2 themes focused explicitly on the app itself and explored perspectives on a potential ED app and suggestions for app content and design. Therapists discussed where an ED app might fit within the treatment pathway, who it would be suitable for, and any potential concerns.

Theme 1: Therapists' Treatment Approach

Overview

The theoretical treatment approach used by all was grounded in CBT, an evidence-based path to recovery; however, in dealing with the complexity of EDs, it went beyond the traditional form. Therapists emphasized personalization in treatment, combining additional therapeutic techniques, such as ACT, compassion-focused therapy, and psychodynamic approaches.

CBT as an Evidence-Based Approach

CBT was described as the "gold standard" treatment covering a structured, goal-oriented approach that helped individuals identify and challenge negative thought patterns and behaviors, making it especially effective for addressing disordered eating behavior. Therapists appreciated the practical, skill-oriented approach of CBT and the provision of a toolkit of resources and concrete tools to change ED behaviors. A strength of CBT was the ability to address key features of EDs, such as cognitive distortions around food and body image and avoidance of feared foods:

> *A lot of people with eating disorders struggle with like perfectionism and black and white thinking. It's that very rational and everything has to be kind of logical sort of thinking in some ways and and obviously CBT helps with that, it helps change those ways of thinking.* [Therapist 8]

> *Firstly, giving them tools that will help reduce anxiety around food cause obviously that's really important. And then very gradually introducing and perhaps foods that they might deem or a food they might deem unsafe, and they avoid.* [Therapist 11]

It was suggested that an app might be useful in providing access to CBT-type resources for those who were unable to access therapy:

> *There's so many people who don't get access to therapy and don't get access to support, to have something that's in your pocket, that's just really convenient would be really useful...there's a couple of websites where you can find typical CBT style kind of interventions and worksheets and if you could just have those accessible to people.* [Therapist 8]

However, therapists noted CBT limitations for this population. The structured format, including homework and behavioral tasks, did not suit everyone, particularly those with neurodivergent conditions. For example, individuals with attention-deficit/hyperactivity disorder (ADHD) might require shorter, more creative sessions to stay engaged, and the rigidity of traditional CBT could be overwhelming. Therapists also observed that CBT assumed individuals were willing and able to engage with cognitive restructuring and behavioral change,

XSL•FO
RenderX

which might not be true for those early in their recovery or resistant to structured approaches. Finally, CBT was not seen as appropriate for deeper emotional issues, such as trauma:

> *It's not ethically OK to...start trauma work in session 8 and leave them after two sessions.* [Therapist 7]
>
> *I think there is a limit to CBT personally, because essentially there's always deeper issues that lead into the behaviors and the thoughts and the emotions.* [Therapist 5]

Person-Centered Care

While CBT principles were followed, they were adapted to fit the client's situation to a person-centered approach aligning with the client's preferences and readiness:

> *It's good to have a protocol as a guide to kind of give you some structure, but you can obviously amend it...this is where it comes into being more person-centred.* [Therapist 10]

Adaptability was key to ensuring therapy remained relevant to everyday ED challenges and individual differences:

> *We can adapt things to make it fit the service user and we follow that lead a little bit to be honest as well.* [Therapist 4]

There was recognition of the potential role of apps to provide different content adapted to individual clients:

> *Provide certain content based on certain personal characteristics...to have different sort of journeys through the things that would be the most likely to be of interest to you.* [Therapist 2]

Complementary Treatment Approaches

Complementary treatment approaches addressed emotional and psychological needs beyond the scope of CBT:

> *The integration of these complementary approaches added emotional depth to the treatment process...[they] delve into the deeper issues...that led to the eating disorder.* [Therapist 8]

The approaches most commonly mentioned were ACT [26] and compassion-focused therapy. ACT was seen as helping individuals accept their thoughts and emotions, encouraging them to commit to actions aligned with their values and not struggle against them. It was seen as helping clients link recovery to deeper, personal motivations and used to reduce self-criticism and cultivate self-compassion:

> *I integrate things like compassion-focused therapy...creating visualizations or safe spaces around compassion for self.* [Therapist 7]

Learning self-compassion was seen as particularly relevant to ED treatment, where clients often experienced shame and self-judgment, helping them develop a compassionate inner voice and offer a gentler, more empathetic approach to their issues:

> *It helps change the stories we tell ourselves.* [Therapist 8]

Theme 2: How Therapy Was Implemented in Practice

Overview

The interviews underscored the importance of structured sessions that incorporated both standard therapeutic activities and flexible content addressing the unique aspects of each client's situation and gave some insights into what could be mirrored in an adjunct app. Collaboration with other professionals was crucial, especially when symptoms were severe or complex.

Structure and Content of ED Therapy

Therapy typically included a set number of sessions, usually 10, especially when the National Health Service commissioned the patients. Key components included initial assessments, psychoeducation on nutrition and body image, and interventions, such as cognitive restructuring, behavior experiments, and relapse prevention.

The initial assessment developed trust and was designed to understand the client's history and identify treatment goals. Sessions started with broad, open-ended discussions before narrowing to specific issues, allowing a thorough exploration of the client's experiences. Some flexibility ensured treatment was kept relevant to the client's immediate needs, whether addressing restrictive behaviors or navigating challenging social situations (refer to Multimedia Appendix 3 for the treatment stages).

Tools included food and thought diaries to guide the direction of therapy:

> *Sessions begin with reviewing that food and thought diary from the week and that can guide where the support will go for that session.* [Therapist 11]

Psychoeducation was a crucial part of therapy, where clients learned about nutrition and the physical effects of ED behaviors:

> *We go through things like the real food pyramid...talking about the effects of purging on the body.* [Therapist 11]

The structure included educating clients about societal influences on body image:

> *I delve into the impact of social media and diet culture on body image...breaking down whether what's posted is a true snapshot of people's lives.* [Therapist 11]

Tailoring Sessions to Individual Needs

As clients came with different histories, experiences, and cognitive styles, therapists emphasized the importance of adapting to personal preferences, specific symptoms, and past experiences. For example, therapists sought to understand what had or had not worked in the past to avoid repetition and to build on strategies that resonated with the client:

> *So with a lot of adults, you'll often find that they've had support in the past and they will sit there and tell you that I have tried this before and that didn't work for me. I've tried this and I really enjoyed that, and I got a lot from that.* [Therapist 12]

XSL•FO
RenderX

Male participants faced specific challenges and stigma related to their ED, which needed to be addressed in treatment:

> *A lot is you will lose your periods and it's like men won't...making sure we're talking about testosterone levels...I'm working with a young gentleman at the minute that is infertile.* [Therapist 4]

The adaptability of the therapeutic process was evident in session planning:

> *We wouldn't sit there with someone and say...we need to do this in Session 5...we are able to have that flexibility and look at what they are struggling with and really just try and help that person how they need help that week essentially.* [Therapist 12]

Another key aspect was the continuous evaluation of the client's state, which allowed therapists to adjust the content and timing of the therapeutic plan:

> *If we can actively see, OK, you've you've binged quite a bit this week. Let's try and focus on that a little bit in some of our future sessions.* [Therapist 12]

If a client was struggling with body image issues, the focus would shift to body acceptance. This involved reframing how they thought about their bodies, helping them move from viewing their bodies traditionally in aesthetic terms to appreciating their functionality:

> *What does that body part do for you? Does how it look change its function?* [Therapist 11]

Therapists recognized that clients worked at different speeds, allowing clients to explore at their own pace and determine the best course of action for themselves:

> *I never forced them to do anything...if they don't feel ready, that's fine.* [Therapist 3]

Collaboration or Referral to Other Specialists

Clients were referred to specialists when they presented with severe symptoms, such as very low BMI or comorbid mental health conditions, or had needs beyond the therapist's expertise requiring more intensive medical or psychiatric support:

> *If someone is struggling with let's say suicidal ideation for example, we make sure somebody's got the general mental health support in place.* [Therapist 8]

There was a collaboration with additional professionals, and clients were provided with access to support networks and group support sessions. This included expert nutritional advice and support in meal planning:

> *Seeing a dietitian or nutritionist who was working with them, who's working very specifically with him about creating some sort of food plan.* [Therapist 5]

For clients with more severe symptoms, therapists worked within a multidisciplinary team that assessed and planned the best course of action for each client:

> *If they're a little bit more severe and it's a little bit more complicated...we will then take them to an MDT meeting and where we just got off to discuss the issue in the case with the rest of the team and kind of just decide between us what's the best course of action for them.* [Therapist 8]

For those who progressed with therapy, help was provided at the end of treatment to help clients transition to less intensive "step-down" support services, such as group sessions or befriending programs. The final sessions involved developing relapse prevention plans, where options for ongoing support were discussed:

> *We put in things like exploring triggers and their support networks...making it clear there's still support, just not one-to-one.* [Therapist 8]

This could also include referral to an ED app, which several therapists suggested could provide support at this difficult stage:

> *Really see it as as step-down option after treatment, a way to check in with themselves...because once you've left recovery, you're on your own.* [Therapist 4]

Theme 3: Strategies for Engaging and Motivating Clients

Engagement created a safe, supportive space where clients felt understood and valued, and motivation focused on helping clients recognize the need for change and overcome any resistance to change.

Therapeutic Relationship

The therapeutic relationship, defined by trust, empathy, and mutual understanding, was seen as fundamental to the success of therapy. Therapists emphasized the importance of rapport, empathy, and creating a safe space where clients felt heard and supported. Although therapists were open to the potential role of an app in treatment, they were skeptical about whether it could provide personalized support, encouragement, and empathy essential for successful treatment outcomes:

> *The relationship is everything in therapy. So I don't really know how that transcribes into an app, and so for that reason I am a little bit wary of apps in a lot of ways.* [Therapist 7]

One therapist highlighted the importance of making the first session comfortable ("The first session is very relaxed, very chilled"; therapist 11) while the majority expressed the importance of creating a ("safe space"; therapist 7). Trust was fundamental, allowing clients to open up about sensitive issues and ("do the hard things"; therapist 2):

> *You're trying to hold a safe space for them, allowing them to talk and be open...but also trying to challenge their identity and aspects of themselves.* [Therapist 8]

> *Creating a therapeutic alliance, showing congruence, showing unconditional positive regard and care for the service user that I'm working with to create a safe space for us to work on.* [Therapist 5]

Therapy was seen as a partnership, where therapists and clients worked together to set goals and navigate challenges. Clients

XSL•FO
RenderX

had to feel confident that they would not be "told off" (therapist 7) or "judged" (therapist 1).

Client Empowerment

Empowerment focused on helping clients take ownership of their recovery, such as supporting goal setting and progress tracking. This also included setting boundaries, ensuring that clients realized responsibility for change lay within themselves. Guidance extended beyond sessions with strategies, such as mindfulness and self-reflection exercises, promoting independence and self-management. One therapist mentioned using an artificial intelligence (AI) tool to encourage self-reflection outside of sessions.

Therapy helped clients increase self-awareness and take control:

> *What makes therapy successful is towards the end...the user is more self-aware, they can almost hear their therapist making their reflections and challenges.* [Therapist 5]

> *I don't take an expert role...although I'm providing techniques, what I am saying is, you are the person who will be able to affect these.* [Therapist 7]

Client Engagement

Customization of resources was important to support client engagement, with worksheets often adapted to make them more acceptable to their clients. This was particularly important for clients with specific needs, such as neurodiverse individuals with ADHD or autism who struggled with focus or traditional therapy methods:

> *I will sit there and literally create a resource...adding pictures or using bright colours.* [Therapist 7]

These types of techniques could be mirrored within an app:

> *If they have to log in, you know allowing customization of like the home page for example that you go to just then it's a bit more inviting for them.* [Therapist 11]

Providing gentle challenges, along with regular feedback on progress and positive outcomes, was seen as key to keeping participants engaged in therapy:

> *Getting them to reflect on where they are, where they've been and what kind of path they've been on is really important.* [Therapist 5]

> *Keep focusing on the wins, you know, getting to pick out a couple of positives, positive things that happened this week or positive things they learned over the last three weeks or something like that.* [Therapist 11]

Complex tasks were broken down into manageable chunks to help clients engage in activities. Setting small, regular goals was seen as valuable in giving clients a sense of accomplishment, supporting ongoing engagement:

> *The absolute best homework is set your alarm, get up in the morning and do one task and that's your homework for the week.* [Therapist 7]

These techniques, including goal setting and monitoring, were seen as easily transferable to apps:

> *Something in your phone that, for example, has your agreed goals.* [Therapist 2]

Addressing Motivation

Therapists used various techniques to enhance clients' motivation and address ambivalence or resistance. Motivational exercises such as future visualizations helped clients envision the outcomes of their behaviors and set goals for change, highlighting potential consequences of their actions and motivating change:

> *We picture ourselves and the tree has two halves...one is if I continue with these behaviors, where will I be in one year, three years, five years?* [Therapist 12]

Therapists helped clients address resistance by exploring conflicting feelings about change, providing a safe space to discuss fears, and "to help them challenge or reframe their reservations" (therapist 5):

> *If I was to keep this eating disorder, what does that mean for myself? And then also get them to look at...if I was to lose and if I was to recover, what would that look like?* [Therapist 8]

Setbacks and lapses were discussed as learning opportunities rather than failures to help maintain motivation. Therapists encouraged a growth mindset, offering support and understanding while focusing on incremental progress:

> *We're always learning...slipping backwards and forwards. The power is in those lapses; that's the learning.* [Therapist 7]

Therapists highlighted that if a client was resistant to change, there was only so much they could do and the client might not have been ready for therapy yet.

Theme 4: Perspectives on a Potential ED App

Overview

Perspectives on a potential ED app alongside therapy were generally positive. However, some therapists were skeptical, citing concerns about triggering content, safeguarding, and data privacy. There was also a critical view of the lack of moderation and poor effectiveness of existing ED apps and concerns about how well an app could personalize therapy or support a therapeutic relationship.

Value of an App

Therapists recognized the potential value of an app for ED treatment as a complementary tool to traditional therapy. It offered greater accessibility to psychoeducational content about EDs and a nonjudgmental space for users who were reluctant to engage in face-to-face treatment. It could also offer users a pretreatment, private space to explore their feelings and behaviors, potentially reducing their stigma and shame, especially for those initially hesitant to seek support:

> *You can educate yourself...without judgment because I know that a lot of people don't come to therapy because they are afraid of judgement.* [Therapist 1]

XSL•FO
RenderX

Therapists commented on the advantage for clients to have educational materials and self-help tools to enhance their sense of independence and maintain continuity between therapy sessions:

> *An app could provide continuous support, be informative and assist with self-education.* [Therapist 5]

An app could also serve as a "follow-up" (therapist 11) to in-person therapy, helping clients track their progress and reinforce the concepts discussed during sessions, as well as a helpful "step-down approach" (therapist 9) at the end of therapy.

Who Would Be Most Suitable for an App?

Therapists felt an app would be best for individuals with mild to moderate symptoms, as they were at a lower risk and more likely to engage in treatment. It could work as a form of early intervention to prevent symptoms from worsening or for those who were no longer in a severe phase. Two therapists mentioned the potential for a recovery tool for patients with more severe symptoms to help prevent relapse:

> *After people have been in the more moderate to severe category, they've had support and then they're sort of stepping away from that support, I think it will be a really useful tool to help people continue on with what they've learned.* [Therapist 9]

Where Would an App Fit in Treatment Pathway?

Regardless of severity, therapists highlighted that users would need motivation to use an app. They emphasized that insights, personal motivation, and readiness to change were key factors in determining its suitability:

> *It has to be a user who's committed to the change or committing to wanting to change.* [Therapist 5]

Therapists saw an app as beneficial at various points in the treatment journey, from before treatment, when individuals sought to understand their behaviors but were uncomfortable discussing them, to being a valuable resource in recovery. There was consensus that an app could complement therapy.

Several participants suggested that an app could be especially valuable for individuals on waiting lists for therapy, serving as a bridge to treatment by offering support and educational resources while they awaited professional help:

> *So many people are on waiting lists for such a long time...if there was even some nutritional education in an app, that would help.* [Therapist 7]

Therapists suggested it could assist with helping clients stay engaged between sessions:

> *They can use the app to follow up their homework...maybe set a goal for the week and the therapist can track their improvement.* [Therapist 3]

Many viewed an app's potential as a tool for relapse prevention and self-monitoring after formal therapy has ended:

> *Once you've left treatment...an app could help you check in with your recovery, set goals and face fear foods.* [Therapist 4]

Potential Safety and Efficacy Concerns

Concerns regarding the use of an app for ED treatment fell under safety and efficacy. First, therapists questioned whether an app could ensure a safe and secure environment for users, addressing issues such as data privacy, crisis management, and safeguarding individuals who may be at risk. Second, they were skeptical about an app's ability to support a meaningful therapeutic relationship, emphasizing the importance of emotional connection, personalized support, and the dynamic, responsive nature of face-to-face therapy for it to be effective.

Safety and Security Concerns

Therapists were concerned about how effectively the app could be moderated to prevent triggering or harmful content. In person, professionals could intervene immediately if a conversation became harmful, but an app may lack the same level of real-time oversight. For immediate safeguarding support for someone in crisis, ensuring safety could be difficult without direct therapist support:

> *One issue came up around like safeguarding, you know again it's sometimes a bit more difficult to support people particularly if there is a safeguarding issue. You know, if they're on an app and as opposed to if they're like in person coming for support.* [Therapist 11]

Therapists suggested safety concerns could be mitigated by personalizing the app so that triggering content could be hidden or by having a link to a crisis number. There were also worries about users encountering unmoderated discussions that could normalize disordered behaviors or reinforce negative self-perceptions. For discussion forums, strong moderation was seen as vital to maintain a safe and supportive space.

Strong concerns were expressed about data privacy, particularly around sharing sensitive information or ensuring that third parties handling the data were trustworthy and not using data for commercial gain:

> *I don't think I would trust any third party with any of the resources that I give out if it was capturing any of their data.* [Therapist 7]

Concerns were raised by 2 therapists about individuals becoming overly reliant on their phones, which was not perceived always to be healthy, particularly if they were engaging in social comparisons on social media or looking up dietary advice. For these people, directing them to an app could potentially reinforce negative behaviors related to phone use:

> *Quite often with my clients, they are working on moving away from their phones.* [Therapist 12]

Perceived Ability of an App to Be Effective

Some therapists expressed doubt over an app's potential to mirror the therapeutic relationship to meet an individual's needs or to engage users over time. Some questioned the potential to provide the depth and emotional connection that formed part of the therapeutic relationship, which was central to achieving successful outcomes. Another perceived challenge was how an app could deliver a treatment experience that aligned with

XSL•FO
RenderX

person-centered therapy and be adaptive to the unique needs of individuals. Some provided suggestions, although these seemed fairly limited compared to what could be achieved via face-to-face therapy:

> *Having some question prompts at the beginning to try and uncover what that person will relate to more could be helpful.* [Therapist 12]

Therapists also noted that maintaining user engagement over time may be difficult due to factors such as the lack of sustained motivation and the absence of real-time human interaction. However, they suggested that reviewing progress, gamification, personalized feedback, and celebrating small wins would be beneficial within an app:

> *If you want to look back at what you've recorded in week one compared to where you are now, how are things different and even just ask them what have you learned, you know, over the last week or two?* [Therapist 11]

> *You could have different story modes. So people finish one, they start anew. And so that will keep engagement.* [Therapist 9]

Theme 5: Suggestions for App Content and Design

App Design

Therapists recommended that an app should provide clear guidance on its purpose and benefits to users, along with personalized content ideally adapted to suit individuals' learning preferences. Design recommendations focused on ensuring the app was user-friendly and engaging, including a simple, intuitive interface with easy navigation and avoiding too much text. Reminders and notifications could be helpful to prompt users to engage with activities, such as food and thought diaries, ensuring these were timely and relevant but not overly intrusive. Rewards and gamification were suggested, with options for unlocking new content or incentives, explicitly avoiding rewards related to changes in eating behaviors. Safeguarding support was mandatory with immediate access to crisis support and coping skills.

Personalization options, such as allowing users to add their names, choose themes, and customize their home pages, were recommended to enhance user engagement. In addition, incorporating fun and light-hearted elements was suggested to make the app more inviting and enjoyable, helping to sustain user interest and participation similar to traditional therapy.

App Content

The therapists used various tools, such as CBT worksheets, food journals, and mindful eating exercises, to engage and support clients. These tools were integral to the therapeutic process, offering clients practical strategies and support as they worked toward recovery. Therapists filtered out those they thought were suitable for app inclusion and those that would be more appropriate for therapist-led activities. Although there was a good level of agreement, some therapists were more cautious about the ability of an app to support certain activities, such as exposure therapy. Table 1 presents a summary of key tools used and their perceived suitability for an app.

XSL•FO
RenderX

Table 1. Therapists' perceptions on which therapeutic tools might be transferable to an eating disorder app.

Therapeutic content and example tools	Appropriate for an app	Unsure of whether transferable
About CBT[a]		
Five areas model ("hot cross bun model")	✓	
Psychoeducational content		
Set-point theory	✓	
Food-related myths	✓	
Consequences (health, social, and emotional)	✓	
Understanding your emotions		
Emotion wheel and mood cards	✓	
Coping strategies and tools		
Safe space visualization	✓	
Breathing exercises (eg, box breathing)	✓	
Grounding exercises	✓	
Self-monitoring tools		
Food and mood diary	✓	
Thought records and cognitive restructuring		
Thought records	✓	
Challenging negative thoughts		✓
Goal setting and tracking		
SMART[b] goal framework	✓	
Action plans	✓	
Exposure therapy		
Fear ladder		✓
Behavior experiments		✓
Distraction techniques		
Urge surfing	✓	
Distress hills	✓	
Trigger toolbox	✓	
Motivational tools		
Motivational quotes	✓	
Praise and positive affirmations	✓	
Self-esteem work		
Value-based work		✓
Strengths and qualities		✓
Body image work		
Body acceptance activities	✓	
Behavior experiments		✓
Social support		
Peer stories and recovery stories	✓	
Online community		✓
Gratitude journal		
Gratitude journal	✓	

XSL•FO
RenderX

[a]CBT: cognitive behavioral therapy.

[b]SMART: specific, measurable, achievable, relevant, time-bound.

Psychoeducational support was deemed essential and valuable, offering information on nutrition, the health impacts of ED behaviors, and debunking common food-related myths. Coping strategies, such as breathing exercises, safe space visualization, and grounding techniques, were also recommended. The app could usefully introduce the fundamentals of CBT and how it worked and incorporated practical tools, such as food and mood diaries, goal setting, and recording negative or automatic thoughts. Exposure techniques, such as fear ladders were mentioned, though there were concerns about suitability as a self-guided activity. Distraction methods, including urge surfing and distress management tools, were considered important and highly suitable for app inclusion. Therapists also emphasized the need to address body image and self-esteem, suggesting that selected content be integrated throughout the app, though therapist support may be required for more challenging work.

Discussion

Principal Findings

This study advanced our understanding of how to design and develop clinically safe, evidence-based ED apps that can complement therapy by extending continuity of care and the self-management and psychoeducation of clients.

Therapists emphasized an integrative, adaptive CBT that incorporated elements of other therapeutic approaches based on individuals' needs and the requirement for strategies to support engagement and motivation within an app [27]. Digital ED interventions based on CBT have previously been developed in combination with other theoretical approaches, such as ACT [28], suggesting an integrated approach to treatment can be effective [10] and could be further extended.

Therapists' resistance to incorporating digital interventions into their treatment could be softened with content that mirrors actual practice and mitigates key concerns, such as the lack of trust, low credibility, and the lack of evidence base perceived in existing apps. Support, but not replacement, of the existing therapeutic relationship was one of the major issues. Some elements of a supportive design could be achieved through personalization, interactive features, and a language style that aims to replicate elements of the therapeutic relationship, such as trust, empathy, and developing bonds [29]. The term "digital therapeutic alliance" defines the alliance a user might perceive with a digital intervention, transferring concepts from a traditional face-to-face therapeutic alliance [30], while also considering the strengths afforded by this medium. For example, just-in-time interventions, which predict and intervene at the point of need, may be one way in which a digital therapeutic alliance could be facilitated [21]. Another way includes building empathy with the user, such as sharing relevant case studies, peer support [31], and the use of affirmative, supportive language [32]. Providing regular, person-centered communication through reminders, check-ins, and motivational messages [33] can also make users feel more involved and connected to an intervention [34].

However, the limitations and challenges of translating therapeutic elements into an app context are significant. Unlike therapists, an app is unlikely to build the deep empathy and trust needed for a therapeutic alliance. Some progress has been made through the application of AI tools, which have demonstrated an ability to encourage self-reflection [35] and some potential for recognizing and validating user emotions [32]. However, apps are currently limited in their ability to provide real-time, nuanced emotional support, which is a core component of the therapeutic alliance. Concerns regarding privacy, safety, and security mean any integration of AI tools within apps requires user involvement, ethical discussions, and extensive testing in advance [36].

Customizing the app's resources, such as providing a range of multimedia formats to accommodate different user needs (eg, ADHD or autism) and catering for the specific needs of male individuals and ethnic groups, is key to maintaining engagement [37] and inclusivity. The use of reminders and prompts could support engagement so long as they are relevant, supportive, and could be turned off, consistent with a previous study on users' app preferences [7]. Offering tailored visual elements or customizable content, such as a personal dashboard, can make users feel more connected to an app, similar to how therapists personalize their approaches to individual clients.

The collaborative aspect of therapy, where clients work with therapists to set goals and navigate challenges, suggests that an app should emphasize user agency [38]. It could achieve this by providing users with the ability to set personalized goals and offer feedback on their progress, encouraging users to take ownership of their recovery [39]. The app might include features, such as progress tracking, feedback loops, and reflective exercises, to help users remain aware of their journey, fostering motivation through visible achievements no matter how small. These small achievements should be celebrated on the way, providing further motivation for the user. Importantly, setbacks should also be managed within the app so that these are not seen as a failure but as a learning opportunity, as in therapy. Rewards could be used, though only in connection with activities that help them progress, such as completing journal entries, and not in relation to food or changes in eating behavior, which could be counterproductive. Working closely with therapists during development helps ensure the app is designed in alignment with therapeutic goals and can provide digital support that complements face-to-face treatment.

Although therapists provided suggestions for app content and design based on their therapeutic practice, apps have the potential to extend beyond therapists' expectations. For example, an app could enhance therapy and improve user understanding of key concepts, such as the five areas model [40] ("hot cross bun model"), through an interactive, step-by-step format, making the content more engaging and accessible for users. Apps could also support more challenging activities such as exposure therapy by providing a structured, supportive, and step-by-step approach to help users confront and overcome their fears. This process could be integrated with therapist support, allowing

XSL•FO
RenderX

clinicians to monitor progress, provide guidance, and adjust exposure tasks through the app interface. There is the potential for a paradigm shift in how therapists perceive apps and their potential value if they are developed using evidence-based methods and in consultation with therapists and users.

It was agreed that an app could complement therapy across various stages of the treatment pathway, including before the beginning of the therapy, during the therapy as a supplementary tool, and after the therapy to support recovery, broadening their potential scope. A recent meta-analysis supported this "blended approach," with patients who used a mobile app alongside therapy having better outcomes than those who did not [41]. Assessing readiness for [42] change before or early within the intervention may be important, as it determines whether a person is prepared to engage or requires an alternative form of support. It is imperative that the safety and security concerns of therapists are addressed within app development.

The key recommendations are mentioned subsequently.

The first recommendation is to comply with data privacy regulations (eg, General Data Protection Regulation and Health Insurance Portability and Accountability Act). Transparent communication about how user data will be used, stored, and protected is fundamental to building trust during early interactions with the app, providing an environment that is safe and secure for users.

The second recommendation is to communicate who the app is for. Apps should clearly specify that they are aimed at individuals with mild to moderate EDs, those in early stages, early recovery, or for relapse prevention, which could include step-down support for more severe patients who are in recovery.

The third recommendation is the immediate access to safeguarding support, that is, providing immediate access to crisis support either to a therapist or to an emergency service. Provide a "panic button" for immediate access to personalized coping skills.

The fourth recommendation is to manage triggering content. Users will need to be able to customize their content, block triggers, and receive warnings before sensitive material content is shown, allowing them to choose whether to proceed or skip. Therefore, understanding potential triggers in advance is crucial to ensure sensitivity around content, requiring thorough testing with users and clinicians before implementation.

The fifth recommendation is a hybrid approach where the app should complement therapy, providing summary reports or personalized feedback to therapists to enhance the therapeutic process. Therapists can address immediate concerns and integrate the app as appropriate to maximize their efficiency and provide between-session support. Therapists should handle complex tasks requiring a therapeutic relationship, such as deep emotional work or anxiety-inducing activities, while managing the overall treatment plan.

The sixth recommendation is app co-design. Collaborative co-design with therapists and users is essential to ensure a person-centered approach to app development, resulting in an app that is clinically suitable, usable, and has the potential to be effective.

Fit With Existing Research

The concerns about safety and security echo the findings of previous research on barriers to using digital interventions within routine clinical care [13], highlighting the need for data security and crisis support within apps as well as the need for evidence-based psychoeducational content [43]. Skepticism about how well digital interventions could replicate the therapeutic relationship or be able to personalize the users' experience were also consistent with previous research [23]. However, this study builds on and complements previous research by providing a deeper exploration of the clinical practice of therapists, identifying techniques and strategies that could be implemented in apps to help overcome these challenges. It paves the way for the design and development of more engaging and effective apps that can complement therapy across the treatment pathway.

Strengths and Limitations

Clinician-centered design enables interventions to be user-centered, clinically relevant, and more likely to be accepted by practitioners. Our study directly engaged with therapists' concerns, contributing to the development of a safer, more effective, and more engaging app. We have provided detailed and practical recommendations for app content and ensured alignment with current therapeutic practices. This study has highlighted the potential of an app to complement existing treatment for mild to moderate EDs and improve therapists' efficiency and patient outcomes. This level of specificity increases the likelihood that future digital interventions will include features that benefit both therapists and their clients. Although the study included only 12 participants through a convenience sampling method, we found, despite differences in experience, training, and qualifications, significant overlap during the coding process, which contributed to the creation of consistent and coherent findings.

Further Work

App co-design work with users and therapists can advance these ideas through a process of iterative development and user testing. Focus groups of potential users ensure the app's personalization features and motivational tools are usable and have the potential to be effective. Collaborating with therapists will ensure the app aligns with therapeutic best practices, including the co-design of specific features (eg, fear ladders) to make sure they can be safely implemented for self-management and to identify when therapist involvement is required. An exploration of how a blended approach would work in practice is also required in terms of how to integrate an app with therapist sessions.

Integrating and evaluating the suggested components using mixed methods will allow for measurement of their impact on user experience, adherence, and clinical outcomes. This should be completed across a diverse range of participants in terms of sociodemographics and ED experiences. Pilot studies are needed to test app efficacy and safety before real-world implementation, along with postimplementation evaluation to assess effectiveness

XSL•FO
RenderX

in practice. This work has the potential to develop usable and effective ED apps that are developed to support the treatment of EDs alongside therapy. It is important that it should capture the perspectives of diverse populations, including those with different ED symptoms, life circumstances, cultural backgrounds, and neurodiverse conditions, to ensure that apps are accessible and relevant across a broad range of users.

The therapeutic alliance was confirmed as central to engagement and a predictor of clinical outcomes and is also a key area for research in digital interventions [29]. More work is needed to understand how apps can better support or possibly replicate the emotional depth, trust, and empathy that are central to therapist-client relationships. Further work is needed around digital technology [7], including developing just-in-time adaptive interventions and using AI tools while addressing challenges around patient safety, data security, and AI or chatbot limitations regarding understanding and responding to human emotions [32]. Emerging technological innovations are capable of delivering in-the-moment interventions tailored to individual needs [15], which could enhance the user experience and improve treatment outcomes of ED apps if privacy and safety are paramount.

Conclusions

This therapist-centered approach paves the way for safer, more effective, and user-centered apps, which have the potential to play a valuable role in ED treatment alongside therapy.

Acknowledgments

The authors would like to acknowledge and express their gratitude to those who were involved in this project at First Steps ED and Thrive Mental Wellbeing and to both these organizations for their ongoing support for this program of work. This project was supported by a patient public involvement grant from University College London Hospitals (UCLH; BRC1100/PPI/PT/104990).

Data Availability

The data supporting this study's findings consist of qualitative interview transcripts and researcher field notes. Due to the sensitive nature of this information and to protect participant confidentiality, these data are not publicly available. However, deidentified excerpts pertinent to the study's conclusions can be obtained from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview guide.
[DOCX File , 42 KB - formative_v9i1e68846_app1.docx]

Multimedia Appendix 2
Baseline characteristics.
[DOCX File , 42 KB - formative_v9i1e68846_app2.docx]

Multimedia Appendix 3
Treatment stages.
[DOCX File , 46 KB - formative_v9i1e68846_app3.docx]

References

1. Ali K, Farrer L, Fassnacht DB, Gulliver A, Bauer S, Griffiths KM. Perceived barriers and facilitators towards help-seeking for eating disorders: a systematic review. Int J Eat Disord 2017 Jan;50(1):9-21. [doi: 10.1002/eat.22598] [Medline: 27526643]
2. Steinhausen HC. Outcome of eating disorders. Child Adolesc Psychiatr Clin N Am 2009 Jan;18(1):225-242. [doi: 10.1016/j.chc.2008.07.013] [Medline: 19014869]
3. Katzman DK. The COVID-19 pandemic and eating disorders: a wake-up call for the future of eating disorders among adolescents and young adults. J Adolesc Health 2021 Oct;69(4):535-537 [FREE Full text] [doi: 10.1016/j.jadohealth.2021.07.014] [Medline: 34489056]
4. Treasure J, Stein D, Maguire S. Has the time come for a staging model to map the course of eating disorders from high risk to severe enduring illness? An examination of the evidence. Early Interv Psychiatry 2015 Jun 29;9(3):173-184. [doi: 10.1111/eip.12170] [Medline: 25263388]
5. Linardon J, Messer M, Lee S, Rosato J. Perspectives of e-health interventions for treating and preventing eating disorders: descriptive study of perceived advantages and barriers, help-seeking intentions, and preferred functionality. Eat Weight Disord 2021 May;26(4):1097-1109. [doi: 10.1007/s40519-020-01005-3] [Medline: 32959274]

XSL•FO
RenderX

6. Wykes T, Lipshitz J, Schueller SM. Towards the design of ethical standards related to digital mental health and all its applications. Curr Treat Options Psych 2019 Jul 5;6(3):232-242. [doi: 10.1007/s40501-019-00180-0]
7. Juarascio AS, Manasse SM, Goldstein SP, Forman EM, Butryn ML. Review of smartphone applications for the treatment of eating disorders. Eur Eat Disord Rev 2015 Jan;23(1):1-11 [FREE Full text] [doi: 10.1002/erv.2327] [Medline: 25303148]
8. Fairburn CG, Rothwell ER. Apps and eating disorders: a systematic clinical appraisal. Int J Eat Disord 2015 Nov;48(7):1038-1046 [FREE Full text] [doi: 10.1002/eat.22398] [Medline: 25728705]
9. Fairburn CG, Murphy R. Treating eating disorders using the internet. Curr Opin Psychiatry 2015 Nov;28(6):461-467. [doi: 10.1097/YCO.0000000000000195] [Medline: 26382155]
10. Thomas PC, Curtis K, Potts HW, Bark P, Perowne R, Rookes T, et al. Behavior change techniques within digital interventions for the treatment of eating disorders: systematic review and meta-analysis. JMIR Ment Health 2024 Aug 01;11:e57577 [FREE Full text] [doi: 10.2196/57577] [Medline: 39088817]
11. Linardon J, Shatte A, Messer M, Firth J, Fuller-Tyszkiewicz M. E-mental health interventions for the treatment and prevention of eating disorders: an updated systematic review and meta-analysis. J Consult Clin Psychol 2020 Nov;88(11):994-1007. [doi: 10.1037/ccp0000575] [Medline: 32852971]
12. Linardon J, Hindle A, Brennan L. Dropout from cognitive-behavioral therapy for eating disorders: a meta-analysis of randomized, controlled trials. Int J Eat Disord 2018 May;51(5):381-391. [doi: 10.1002/eat.22850] [Medline: 29493805]
13. Mayer G, Lemmer D, Michelsen I, Schrader P, Friederich HC, Bauer S. Views of German mental health professionals on the use of digital mental health interventions for eating disorders: a qualitative interview study. J Eat Disord 2024 Feb 23;12(1):32 [FREE Full text] [doi: 10.1186/s40337-024-00978-1] [Medline: 38395950]
14. Loucas CE, Fairburn CG, Whittington C, Pennant ME, Stockton S, Kendall T. E-therapy in the treatment and prevention of eating disorders: a systematic review and meta-analysis. Behav Res Ther 2014 Dec;63:122-131 [FREE Full text] [doi: 10.1016/j.brat.2014.09.011] [Medline: 25461787]
15. Carrard I, Crépin C, Rouget P, Lam T, Golay A, Van der Linden M. Randomised controlled trial of a guided self-help treatment on the internet for binge eating disorder. Behav Res Ther 2011 Aug;49(8):482-491. [doi: 10.1016/j.brat.2011.05.004] [Medline: 21641580]
16. Yardley L, Morrison L, Bradbury K, Muller I. The person-based approach to intervention development: application to digital health-related behavior change interventions. J Med Internet Res 2015 Jan 30;17(1):e30 [FREE Full text] [doi: 10.2196/jmir.4055] [Medline: 25639757]
17. Torous J, Nicholas J, Larsen ME, Firth J, Christensen H. Clinical review of user engagement with mental health smartphone apps: evidence, theory and improvements. Evid Based Ment Health 2018 Aug;21(3):116-119 [FREE Full text] [doi: 10.1136/eb-2018-102891] [Medline: 29871870]
18. Graham AK, Kosmas JA, Massion TA. Designing digital interventions for eating disorders. Curr Psychiatry Rep 2023 Apr;25(4):125-138. [doi: 10.1007/s11920-023-01415-x] [Medline: 36928767]
19. Liu C, Linardon J. Mental health clinicians' practices and perspectives of eating disorder apps. Psychiatry Res 2025 May;347:116412 [FREE Full text] [doi: 10.1016/j.psychres.2025.116412] [Medline: 39987587]
20. Juarascio AS, Goldstein SP, Manasse SM, Forman EM, Butryn ML. Perceptions of the feasibility and acceptability of a smartphone application for the treatment of binge eating disorders: qualitative feedback from a user population and clinicians. Int J Med Inform 2015 Oct;84(10):808-816 [FREE Full text] [doi: 10.1016/j.ijmedinf.2015.06.004] [Medline: 26113461]
21. Juarascio A, Srivastava P, Presseller E, Clark K, Manasse S, Forman E. A clinician-controlled just-in-time adaptive intervention system (CBT+) designed to promote acquisition and utilization of cognitive behavioral therapy skills in bulimia nervosa: development and preliminary evaluation study. JMIR Form Res 2021 May 31;5(5):e18261 [FREE Full text] [doi: 10.2196/18261] [Medline: 34057416]
22. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
23. Boddy CR. Sample size for qualitative research. Qual Market Res 2016 Sep 12;19(4):426-432. [doi: 10.1108/qmr-06-2016-0053]
24. NVivo. Lumivero. URL: https://lumivero.com/products/nvivo/ [accessed 2025-04-18]
25. Clarke V, Braun V. Thematic Analysis: A Practical Guide. London, UK: SAGE Publications; 2021.
26. Hayes SC, Strosahl KD, Strosahl K, Wilson KG. Acceptance and Commitment Therapy: The Process and Practice of Mindful Change. New York, NY: Guilford Press; 2012.
27. National Collaborating Centre for Mental Health (UK). Eating Disorders: Core Interventions in the Treatment and Management of Anorexia Nervosa, Bulimia Nervosa and Related Eating Disorders. Leicester, UK: British Psychological Society; 2004.
28. Wyssen A, Meyer AH, Messerli-Bürgy N, Forrer F, Vanhulst P, Lalanne D, et al. BED-online: acceptance and efficacy of an internet-based treatment for binge-eating disorder: a randomized clinical trial including waitlist conditions. Eur Eat Disord Rev 2021 Nov;29(6):937-954 [FREE Full text] [doi: 10.1002/erv.2856] [Medline: 34418221]
29. Bordin ES. The generalizability of the psychoanalytic concept of the working alliance. Psych Psychother Theory Res Pract 1979;16(3):252-260. [doi: 10.1037/h0085885]

XSL•FO
RenderX

30. Kaveladze B, Schueller SM. Chapter 8 - A digital therapeutic alliance in digital mental health. In: Jacobson N, Kowatsch T, Marsch L, editors. Digital Therapeutics for Mental Health and Addiction: The State of the Science and Vision for the Future. Cambridge, MA: Academic Press; 2023:87-98.
31. Rennick-Egglestone S, Morgan K, Llewellyn-Beardsley J, Ramsay A, McGranahan R, Gillard S, et al. Mental health recovery narratives and their impact on recipients: systematic review and narrative synthesis. Can J Psychiatry 2019 Oct 02;64(10):669-679 [FREE Full text] [doi: 10.1177/0706743719846108] [Medline: 31046432]
32. Inkster B, Sarda S, Subramanian V. An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: real-world data evaluation mixed-methods study. JMIR Mhealth Uhealth 2018 Nov 23;6(11):e12106 [FREE Full text] [doi: 10.2196/12106] [Medline: 30470676]
33. Proudfoot J, Clarke J, Birch MR, Whitton AE, Parker G, Manicavasagar V, et al. Impact of a mobile phone and web program on symptom and functional outcomes for people with mild-to-moderate depression, anxiety and stress: a randomised controlled trial. BMC Psychiatry 2013 Nov 18;13:312 [FREE Full text] [doi: 10.1186/1471-244X-13-312] [Medline: 24237617]
34. Kuhn E, Greene C, Hoffman J, Nguyen T, Wald L, Schmidt J, et al. Preliminary evaluation of PTSD Coach, a smartphone app for post-traumatic stress symptoms. Mil Med 2014 Jan;179(1):12-18. [doi: 10.7205/MILMED-D-13-00271] [Medline: 24402979]
35. Nepal S, Pillai A, Campbell W, Massachi T, Heinz MV, Kunwar A, et al. MindScape study: integrating LLM and behavioral sensing for personalized AI-driven journaling experiences. Proc ACM Interact Mob Wearable Ubiquitous Technol 2024 Nov 21;8(4):1-44 [FREE Full text] [doi: 10.1145/3699761] [Medline: 39664112]
36. Sharp G, Torous J, West ML. Ethical challenges in AI approaches to eating disorders. J Med Internet Res 2023 Aug 14;25:e50696 [FREE Full text] [doi: 10.2196/50696] [Medline: 37578836]
37. Tregarthen J, Paik Kim J, Sadeh-Sharvit S, Neri E, Welch H, Lock J. Comparing a tailored self-help mobile app with a standard self-monitoring app for the treatment of eating disorder symptoms: randomized controlled trial. JMIR Ment Health 2019 Nov 21;6(11):e14972 [FREE Full text] [doi: 10.2196/14972] [Medline: 31750837]
38. Palacios JE, Erickson-Ridout KK, Paik Kim J, Buttlaire S, Ridout S, Argue S, et al. Effects of a digital therapeutic adjunct to eating disorder treatment on health care service utilization and clinical outcomes: retrospective observational study using electronic health records. JMIR Ment Health 2024 Nov 27;11:e59145 [FREE Full text] [doi: 10.2196/59145] [Medline: 39602804]
39. Tregarthen JP, Lock J, Darcy AM. Development of a smartphone application for eating disorder self-monitoring. Int J Eat Disord 2015 Nov;48(7):972-982. [doi: 10.1002/eat.22386] [Medline: 26213130]
40. Padesky CA. Collaborative case conceptualization: client knows best. Cogn Behav Pract 2020 Nov;27(4):392-404. [doi: 10.1016/j.cbpra.2020.06.003]
41. Lindhiem O, Bennett CB, Rosen D, Silk J. Mobile technology boosts the effectiveness of psychotherapy and behavioral interventions: a meta-analysis. Behav Modif 2015 Nov;39(6):785-804 [FREE Full text] [doi: 10.1177/0145445515595198] [Medline: 26187164]
42. Prochaska JO, Norcross JC. Stages of change. Psychother Theory Res Pract Train 2001;38(4):443-448. [doi: 10.1037/0033-3204.38.4.443]
43. Lemmer D, Mayer G, Schrader P, Michelsen I, Friederich HC, Bauer S. Experts' views on the implementation of digital interventions for eating disorders: a Delphi study. BMC Public Health 2024 Sep 12;24(1):2486 [FREE Full text] [doi: 10.1186/s12889-024-19989-3] [Medline: 39267038]

Abbreviations

ACT: acceptance and commitment therapy
ADHD: attention-deficit/hyperactivity disorder
AI: artificial intelligence
CBT: cognitive behavioral therapy
COREQ: Consolidated Criteria for Reporting Qualitative Research
ED: eating disorder
REDCap: Research Electronic Data Capture

XSL•FO
RenderX

Edited by A Mavragani; submitted 15.11.24; peer-reviewed by MC Daugelat, S Hughes; comments to author 21.02.25; revised version received 03.04.25; accepted 08.04.25; published 06.05.25.

Please cite as:
Thomas PC, Bark P, Rowe S
Exploring Therapists' Approaches to Treating Eating Disorders to Inform User-Centric App Design: Web-Based Interview Study
JMIR Form Res 2025;9:e68846
URL: https://formative.jmir.org/2025/1/e68846
doi:10.2196/68846
PMID:

©Pamela Carien Thomas, Pippa Bark, Sarah Rowe. Originally published in JMIR Formative Research (https://formative.jmir.org), 06.05.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Cocreating the Visualization of Digital Mobility Outcomes: Delphi-Type Process With Patients

Jack Lumsdon[1], MSc; Cameron Wilson[2], MPH; Lisa Alcock[3,4], PhD; Clemens Becker[5,6], MD; Francesco Benvenuti[7], MD; Tecla Bonci[8,9], PhD; Koen van den Brande[7], BSc; Gavin Brittain[10,11], MD; Philip Brown[12], MSc; Ellen Buckley[9,13], PhD; Marco Caruso[14], PhD; Brian Caulfield[15,16], PhD; Andrea Cereatti[14], PhD; Laura Delgado-Ortiz[17,18,19], PhD; Silvia Del Din[3,4], PhD; Jordi Evers[20], MSc; Judith Garcia-Aymerich[17,18,19], PhD; Heiko Gaßner[21,22], PhD; Tova Gur Arieh[7], BA; Clint Hansen[23], PhD; Jeffrey M Hausdorff[24,25,26], PhD; Hugo Hiden[27], PhD; Emily Hume[28], PhD; Cameron Kirk[3], PhD; Walter Maetzler[23], PhD; Dimitrios Megaritis[3,28], PhD; Lynn Rochester[3,4,12], PhD; Kirsty Scott[3], PhD; Basil Sharrack[10,11], PhD; Norman Sutton[7]; Beatrix Vereijken[29], PhD; Ioannis Vogiatzis[28], PhD; Alison Yarnall[3,4,12], PhD; Alison Keogh[15,30], PhD; Alma Cantu[27], PhD

[1]Population Health Sciences Institute, Faculty of Medical Sciences, Newcastle University, Newcastle Upon Tyne, United Kingdom
[2]School of Clinical Medicine, Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom
[3]Translational and Clinical Research Institute, Faculty of Medical Sciences, Newcastle University, Newcastle Upon Tyne, United Kingdom
[4]NIHR Newcastle Biomedical Research Centre, Newcastle University and The Newcastle Upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, United Kingdom
[5]Robert Bosch Society for Medical Research, Stuttgart, Germany
[6]Digital Geriatrics Unit, Medical Centre, University of Heidelberg, Heidelberg, Germany
[7]Mobilise-D Patient and Public Advisory Group, Newcastle Upon Tyne, United Kingdom
[8]School of Mechanical, Aerospace and Civil Engineering, University of Sheffield, Sheffield, United Kingdom
[9]Insigneo Institute for Silico Medicine, University of Sheffield, Sheffield, United Kingdom
[10]Department of Neuroscience and Sheffield NIHR Translational Neuroscience BRC, Sheffield Teaching Hospitals NHS Foundation Trust, Sheffield, United Kingdom
[11]Sheffield Institute for Translational Neuroscience, University of Sheffield, Sheffield, United Kingdom
[12]The Newcastle Upon Tyne Hospitals NHS Foundation Trust, Newcastle Upon Tyne, United Kingdom
[13]Division of Clinical Medicine, University of Sheffield, Sheffield, United Kingdom
[14]Department of Electronics and Telecommunications, Politecnico di Torino, Turin, Italy
[15]Insights Centre Data Analytics, University College Dublin, Dublin, Ireland
[16]School of Public Health, Physiotherapy and Sports Science, University of College Dublin, Dublin, Ireland
[17]Barcelona Institute for Global Health, Barcelona, Spain
[18]Department of Medicine and Life Sciences, Pompeu Fabra University, Barcelona, Spain
[19]Network in Epidemiology and Public Health, Center for Biomedical Research, Madrid, Spain
[20]McRoberts BV, The Hague, The Netherlands
[21]Department of Molecular Neurology, University Hospital Erlangen, Erlangen, Germany
[22]Fraunhofer Institute for Integrated Circuits IIS, Erlangen, Germany
[23]Department of Neurology, University Medical Center Schleswig-Holstein Campus, Kiel, Germany
[24]Sagol School of Neuroscience and Department of Physical Therapy, Faculty of Medical and Health Sciences, Tel Aviv University, Tel Aviv, Israel
[25]Rush Alzheimer's Disease Center and Department of Orthopaedic Surgery, Rush University Medical Center, Chicago, IL, United States
[26]Center for the Study of Movement, Cognition, and Mobility, Neurological Institute, Tel Aviv Sourasky Medical Center, Tel Aviv, Israel
[27]School of Computer Science, Newcastle University, Newcastle Upon Tyne, United Kingdom
[28]Department of Sport, Exercise and Rehabilitation, Northumbria University, Newcastle Upon Tyne, United Kingdom
[29]Department of Neuromedicine and Movement Science, Norwegian University of Science and Technology, Trondheim, Norway
[30]School of Medicine, Trinity College Dublin, Dublin, Ireland

Corresponding Author:
Jack Lumsdon, MSc
Population Health Sciences Institute
Faculty of Medical Sciences
Newcastle University

XSL•FO
RenderX

Floor 2, Newcastle Biomedical Research Building
Campus of Ageing and Vitality
Newcastle Upon Tyne, NE4 5PL
United Kingdom
Phone: 44 01912081103
Email: jack.lumsdon@newcastle.ac.uk

Abstract

Background: Recent technological advances in wearable devices offer new potential for measuring mobility in real-world contexts. Mobilise-D has validated digital mobility outcomes to provide novel outcomes and end points in clinical research of 4 different long-term health conditions (Parkinson disease, multiple sclerosis, chronic obstructive pulmonary disease, and proximal femoral fracture). These outcomes also provide unique information that is important to patients; however, there is limited literature that explores the optimal methods to achieve this, such as the best way to visualize patients' data.

Objective: This study aimed to identify meaningful outcomes for each condition and how to best visualize them from the perspective of end users.

Methods: Using a Delphi-type protocol with patients as subject matter experts, we gathered iterative feedback on the cocreation of visualizations through 3 rounds of questionnaires. An open-ended questionnaire was used in round 1 to understand what aspects of mobility were most influenced by their health condition. These responses were mapped onto relevant digital mobility outcomes and walking experiences and then prioritized for visualization. Using patient responses, we worked alongside researchers, clinicians, and a patient advisory group to develop visualizations that depicted a week of mobility data. During rounds 2 and 3, participants rated usefulness and ease of understanding on a 5-point Likert scale and provided unstructured feedback in comment boxes for each visualization. Visualizations were refined using the feedback from round 2 before receiving further feedback in round 3.

Results: Participation varied across rounds 1 to 3 (n=48, n=79, and n=78, respectively). Round 1 identified important outcomes and contexts for each health condition, such as walking speed and stride length for people with Parkinson disease or multiple sclerosis and number of steps for people with chronic obstructive pulmonary disease or proximal femoral fracture. The consensus was not reached for any visualization reviewed in round 2 or 3. Feedback was generally positive, and some participants reported that they were able to understand the visualization and interpret what the visualization represented.

Conclusions: Through the feedback provided and existing data visualization principles, we developed recommendations for future visualizations of mobility- and health-related data. Visualizations should be readable by ensuring that large and clear fonts are used and should be friendly for people with vision impairments, such as color blindness. Patients have a strong understanding of their own condition and its variability; hence, adding additional factors into visualizations is recommended to better reflect the nuances of a condition. Ensuring that outcomes and visualizations are meaningful requires close collaboration with patients throughout the development process.

(JMIR Form Res 2025;9:e68782) doi:10.2196/68782

KEYWORDS

mobility; data visualization; wearable devices; digital mobility outcomes; cocreation

Introduction

Background

Mobility refers to the ability to move freely and easily to carry out activities of daily living and refers to movement in all forms, from moving out of a chair to walking [1]. It is necessary for daily tasks, participating in social activities, and maintaining independence [2]. Mobility is a meaningful aspect of health for individuals across multiple health conditions and plays a large role in their physical, social, and psychological experience [3]. In conditions where motor symptoms cause functional impairment, understanding an individual's walking, posture, and how they move is important for measuring and monitoring condition severity [4]. Specifically, walking is an important clinical measure whereby factors such as walking speed and stride length can predict functional decline and increase risk of frailty and mortality [5,6]. Hence, it is essential to be able to measure walking accurately and meaningfully convey these data to patients to provide insights into progress and changes in their condition over time.

Recent technological advances in wearable devices (such as body-worn sensors) offer new potential for measuring mobility continuously in real-world contexts (collectively termed digital mobility outcomes [DMOs]). This in turn presents the opportunity for novel insights into aspects of health that are important to patients, such as their real-world walking speed and daily step count [7]. Crucially, providing individuals with access to, and insights from, their own data from such devices is linked to patient autonomy, enhanced health professional-patient communication, and improved compliance with physician recommendations through the enhanced trust between patients and those involved in their care [8-12].

XSL•FO
RenderX

However, the effectiveness of this feedback hinges upon the meaningfulness of the outcome to patients, particularly whether their symptoms and symptom burden can be reflected by the outcome and be interpreted by patients [13-15].

Integration of the wearable devices into health care systems presents significant challenges in the visualization of the complex data and the insights these generate [16]. Mobility can be affected by multiple factors, such as changes in symptoms, environmental factors, and medication, which creates challenges and added complexity when communicating insights from the data [17-19]. Hence, it is essential to ensure that data visualizations are both accessible and understandable to both patients and clinicians and reflect the nuances of patients' mobility while also representing meaningful aspects of their health [14,20,21]. Creating visualizations that enable an individual to come to a full understanding of a person's overall health condition can be challenging; hence, several approaches have been attempted without one prevailing approach [19]. In previous visualizations from wearable devices, a range of traditional visualization approaches, such as radar plots, bar charts, line plots, and box plots, have been used, as well as more novel approaches, such as a calendar depicting 6 months of data in 1 visual [21,22]. Consequently, exploring data visualization preferences is important yet nuanced, and the creation of the visualizations requires an understanding of the patient experience.

Objectives

This work builds upon the efforts of the Mobilise-D consortium to validate DMOs and provide novel insights into several long-term health conditions, including Parkinson disease (PD), multiple sclerosis (MS), chronic obstructive pulmonary disease (COPD), and proximal femoral fracture (PFF) [17,23-25]. Research has been carried out to establish the technical, clinical validity, and acceptability of wearable devices and the analytic pipelines applied to quantify DMOs [17,23,26]. An additional part of ensuring implementation of the methods and outcomes is to determine how best to visualize DMO data for patients. Consequently, this study aimed to explore patient visualization preferences by (1) determining which DMOs are meaningful to individuals living with diverse long-term health conditions and (2) identifying how to best visualize them from the perspective of end users [12].

Previous research on wearable device data visualization has highlighted the lack of consistent preferences regarding how best to visualize the data. Several approaches have explored the use of varying levels of granularity, with some studies attempting to visualize a day and others 6 months [21,22]. While this was highly variable depending upon the outcome, for mobility it is common to use a week of data [18,19,21]. A major point of interest in the literature was embedding contextual data within visualizations, particularly in relation to clinical use to enhance the inferences made [18,19]. Condition-related information, such as symptoms, emotions, and medications, can influence the lived experience, and the data from wearable devices should be able to reflect this, as well as the time and location of activities [27,28]. While we know specific design approaches and elements are key to the creation of comprehensive visualizations, it is also important to consider the individualized nature of preferences when developing meaningful visualizations [12]. Hence, this study aimed to map DMOs into the lived experiences of mobility and symptom burden and explore the visualization preferences of patients using a week of DMO data.

Methods

Study Design

This study design was based on a Delphi methodology in which patients acted as subject matter experts to iteratively provide feedback on the cocreation of data visualizations through 3 rounds of questionnaires. It was designed in collaboration with academic researchers of Mobilise-D and the members of the Mobilise-D Public and Patient Advisory Group (PPAG) [29].

In total, 3 rounds were developed to explore the impact of mobility and symptoms and how DMOs can be visualized. A Delphi methodology was decided upon as it allows us to reach a consensus of patient preferences with an iterative, anonymous, multistage approach with controlled feedback of comments and scores on a 5-point Likert scale [30,31]. The Delphi methodology recommends 3 rounds of feedback; the first round is to generate qualitative data on a topic and the remaining rounds are to gain a consensus through Likert scales [31]. We followed the recommended 3 round; however, the methods used here will be described as a "Delphi-type process" due to deviations from Delphi-specific design elements. A key aspect of a Delphi methodology is ensuring participants completed previous rounds; however, due to ethical requests, this was not possible within this study [31,32]. This process involved a collaboration between patient feedback and researcher refinement, as described in Figure 1.

XSL•FO
RenderX

Figure 1. Outline of the study design, showing how the results from each round built upon each other. DMO: digital mobility outcome.

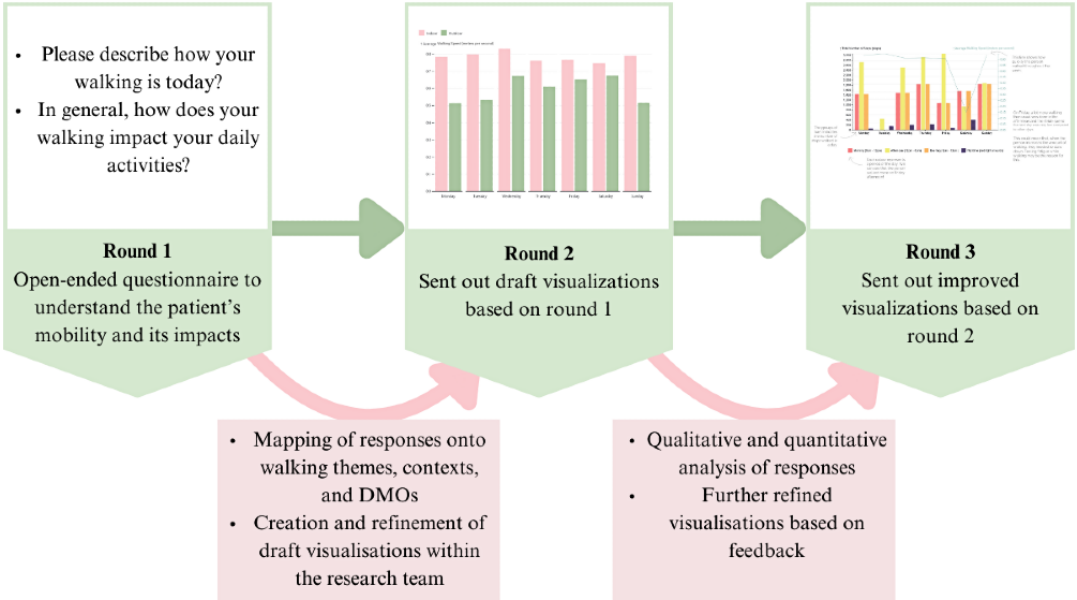

Visualization Data

The data visualized in this study were collected as part of the Mobilise-D technical validation study (TVS) [23]. Participants with long-term health conditions wore a lower back wearable device (either a McRoberts MoveMonitor+ or the Axivity AX6) to provide real-world monitoring data for 7 consecutive days. Accelerometry data were collected and processed using the algorithms validated by Mobilise-D to produce DMOs [33]. In addition, GPS data were collected for participants, which allowed the data to be categorized into indoor versus outdoor. No activity or medication diaries were collected, which limited the contextual information available when visualizing the data.

Public and Patient Involvement

This research was designed with members of the Mobilise-D PPAG [29]. In total, 4 international representatives (Belgium, Israel, Italy, and the United Kingdom) from the PPAG (2 with PD, 1 with a previous hip fracture, and 1 with MS) worked with academic researchers to develop each round of questions and visualizations to ensure they were appropriately worded and constructed and that the conclusions drawn from the data in each round were meaningful. Due to the international and remote work involved in the PPAG, perspectives of the digitally excluded unfortunately could not be included with the team.

Recruitment

People with PD, MS, COPD, or a previous hip fracture primarily from Europe (British, Irish, Norwegian, and German) were invited to participate in this web-based survey Delphi-type study. Patients were purposively recruited from newsletters sent out to various patient charities and organizations (including Voice, MS Society, and Parkinson's UK), social media, and through existing Mobilise-D clinical sites. A sample size of between 10 and 15 people per health condition was proposed in line with guidance for Delphi studies [31].

There were 239 responses to this initial form, which was then used as the distribution list for all rounds of the study. However, because of this method of recruitment, participants were unable to be individually tracked or compared per round, and this meant the participation of individuals in previous rounds could not be confirmed, which is not typical for a Delphi methodology [32]. Data were collected between June 2023 and June 2024. As this study aimed to iteratively build upon previous findings, the lack of participation in previous rounds may have impacted the participant's understanding of the context surrounding each round. However, each round started with a summary of the project and the findings from previous rounds to mitigate some of this confusion.

Data Collection and Analysis

Round 1

An open-ended questionnaire was used in round 1 to understand what aspects of mobility were most influenced by their long-term health condition (Multimedia Appendix 1). Questions were based on existing literature relating to how people with long-term health conditions experience mobility, specifically how it impacts them physically, socially, and psychologically [3]. This questionnaire was piloted with the PPAG before distribution, and the feedback led to questions being accompanied by a sample text written by the PPAG, exemplifying their mobility experiences and prompting participants (Multimedia Appendix 1).

The results of round 1 were aggregated and then analyzed using a deductive content analysis by 6 researchers with experience in digital mobility and qualitative research (AK, LD-O, AC, HG, CW, and KS). Deductive analysis was conducted by mapping responses onto walking themes, symptoms, and the most relevant DMOs (Table 1 and Multimedia Appendix 2). Responses from the questionnaire were discussed within the team regarding how elements of walking may be impacted by symptoms and how this may manifest within a DMO. This mapping involved the research team (including members of the PPAG and clinical experts—LR, GB, TGA, KvdB, and NS) discussing how elements of walking may be impacted by symptoms and how this might manifest within a DMO.

XSL•FO
RenderX

Table 1. Digital mobility outcomes used to map the patient experience during round 1, adapted from Mazzà et al [23] and Kluge et al [34].

Digital mobility outcomes	Description
Step count	Number of steps taken per day.
Number of WBs[a]	Number of WBs taken per day. A WB is a walking sequence containing at least 2 consecutive strides of both feet. The start and end of a WB are determined by a resting period or any other activity (nonwalking period).
WB duration	Duration between the start and the end of a WB. Defined as the daily average measured across each WB.
Stride length	The interval between 2 successive initial contacts of the same foot. Thereby, every stride contains 2 steps. Defined as the daily average measured across each WB.
Walking speed	Distance covered by the whole body within a certain time interval per unit time of walking. It is measured in meters per second and is the magnitude of the velocity vector. Defined as the daily average measured across each WB.
Step duration	Duration between an initial contact and the next initial contact of the opposite foot. Defined as the daily average measured across each WB.
Cadence	The number of steps taken per minute. Defined as the daily average measured across each WB.

[a]WB: walking bout.

In total, 3 rounds of cross-checking were conducted to clarify differences and determine specific coding rules before finalizing an agreed set of walking experiences deemed most important and how they map onto specific DMOs. Specific elements of walking most frequently mentioned were subsequently identified for each health condition to prioritize the aspects that were most important to patients. Following this, DMOs were identified that required the development of visualizations that addressed the most frequent walking experiences (Multimedia Appendix 2).

Round 2

The results of round 1 were used to develop draft visualizations that depicted a week of mobility data. Using the data from the Mobilise-D TVS, visualizations were drafted (Multimedia Appendix 3) [23]. Specifically, visualizations were prioritized based on the commonly listed experiences of mobility for each long-term health condition from round 1. The order of visualizations that were presented to participants was randomized to prevent bias.

On the basis of a list of relevant options provided by a visualization expert (AC), the research team and PPAG (AK, HG, CK, LD-O, JL, AC, NS, and TGA) discussed and voted for visualizations and contexts that they thought were most appropriate and useful for each element of walking. These decisions were reviewed, and the final visualizations for each element of walking were selected. The draft visualizations were then matched to a sample of text to explain why they were chosen and how the researchers felt that they related to people's experiences (Multimedia Appendix 1).

Following the guidance from Trevelyan and Robinson [31], patients were asked to rate the usefulness and ease of understanding of each individual visualization using a 5-point Likert scale (0=not useful at all or very difficult and 5=very useful or very easy), as well as to suggest potential changes to the draft visualization. We considered consensus was reached when at least 75% of participants rated a visualization as ≥4 on the Likert scale [30]. In addition, qualitative responses were reviewed for relevant feedback, and a content analysis of responses was conducted by 2 researchers (JL and AK) to suggest changes for the visualizations.

Round 3

Using the results of round 2, the draft visualizations were revised with the aim of improving them in line with patient feedback. Revised visualizations were discussed between the academic research team and PPAG (CK, HG, LD-O, AK, JL, AC, NS, and TGA) and were then presented to patients alongside the justification for their creation (Multimedia Appendix 1). The order of visualizations that were presented to participants was randomized to prevent bias. Participants rated the revised visualizations on a 5-point Likert scale and provided comments, which were analyzed using a content analysis.

Ethical Considerations

Ethics approval was granted by Newcastle University (2514/30280). To ensure anonymity in line with ethical requests, participants were asked to complete an anonymous form whereby their email address was collected and stored securely separate from demographic and response data from the study. These email addresses were stored securely on a password-protected computer and only accessed by researchers involved in questionnaire distribution. This was then used to send participants a link to take part in the study anonymously. Participation was voluntary, and responses were anonymous, with no personally identifiable data collected. All study participants received an information sheet about the study and were provided with the researcher's emails in case of any questions. Participants were informed that the study should not take any longer than 10 minutes and were assured that they could withdraw without consequence at any time during the study. All participants gave informed consent. There was no compensation for participation in this study.

Results

Participants

Participation was dependent upon a round being fully completed; the demographics of participants across the 3 rounds are presented in Table 2. Participation varied across rounds 1 to 3

XSL•FO
RenderX

(n=48, n=79, and n=78, respectively). Across all rounds and participants, 63.4% (n=130) were female participants, 29.8% (n=71) of participants were below degree level or preferred not to say, 94.1% (n=193) were White and 37.6% (n=77) wore a wearable device daily. Considering that the sample size for COPD and PFF was lower than expected, their qualitative responses were included in the analysis but were not considered to have reached consensus [31].

XSL•FO
RenderX

Table 2. Participant demographics across all 3 rounds separated by each condition.

Demographic details	Round 1 (n=48), n (%)				Round 2 (n=79), n (%)				Round 3 (n=78), n (%)			
	PD[a] (n=16)	MS[b] (n=25)	COPD[c] (n=1)	PFF[d] (n=6)	PD (n=52)	MS (n=24)	COPD (n=1)	PFF (n=2)	PD (n=54)	MS (n=21)	COPD (n=2)	PFF (n=1)
Sex												
Male	7 (44)	20 (80)	2 (100)	3 (50)	25 (48)	3 (13)	0 (0)	2 (100)	25 (46)	2 (10)	0 (0)	1 (100)
Female	9 (56)	5 (20)	0 (0)	3 (50)	27 (52)	20 (83)	1 (100)	0 (0)	29 (54)	19 (91)	1 (50)	0 (0)
Prefer not to say	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (4)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Age (y)												
18-30	0 (0)	1 (4)	0 (0)	0 (0)	1 (2)	2 (8)	0 (0)	0 (0)	0 (0)	1 (5)	0 (0)	0 (0)
31-40	0 (0)	7 (28)	0 (0)	0 (0)	0 (0)	6 (25)	0 (0)	0 (0)	0 (0)	6 (29)	0 (0)	0 (0)
41-50	1 (6)	11 (44)	0 (0)	0 (0)	1 (2)	6 (25)	0 (0)	0 (0)	2 (4)	9 (43)	0 (0)	0 (0)
51-60	1 (6)	6 (24)	0 (0)	1 (17)	8 (15)	7 (29)	0 (0)	0 (0)	14 (26)	3 (14)	0 (0)	0 (0)
61-70	7 (44)	0 (0)	2 (100)	0 (0)	21 (40)	3 (13)	1 (100)	0 (0)	22 (41)	2 (10)	2 (100)	0 (0)
71-80	6 (38)	0 (0)	0 (0)	5 (83)	15 (29)	0 (0)	0 (0)	2 (100)	14 (26)	0 (0)	0 (0)	1 (100)
>81	1 (6)	0 (0)	0 (0)	0 (0)	6 (12)	0 (0)	0 (0)	0 (0)	2 (4)	0 (0)	0 (0)	0 (0)
Race and ethnicity												
Asian or Pacific Islander	0 (0)	1 (4)	0 (0)	0 (0)	2 (4)	0 (0)	0 (0)	0 (0)	1 (2)	1 (5)	0 (0)	0 (0)
Hispanic or Latino	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (4)	0 (0)	0 (0)	0 (0)	1 (5)	0 (0)	0 (0)
White	16 (100)	22 (88)	2 (100)	6 (100)	50 (96)	21 (87)	1 (100)	2 (100)	52 (96)	19 (91)	2 (100)	1 (100)
Other[e]	0 (0)	2 (8)	0 (0)	0 (0)	0 (0)	1 (4)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Prefer not to say	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (4)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Education												
Doctorate	1 (6)	1 (4)	0 (0)	2 (33)	5 (10)	1 (4)	0 (0)	2 (100)	5 (9)	1 (5)	0 (0)	1 (100)
Masters degree	5 (31)	1 (4)	0 (0)	2 (33)	14 (27)	2 (8)	0 (0)	0 (0)	16 (30)	2 (10)	0 (0)	0 (0)
Bachelors degree	5 (31)	16 (64)	0 (0)	0 (0)	12 (23)	11 (46)	0 (0)	0 (0)	19 (35)	11 (52)	0 (0)	0 (0)
High school degree or equivalent	1 (6)	6 (24)	0 (0)	2 (33)	10 (19)	6 (25)	1 (100)	0 (0)	6 (11)	6 (29)	1 (50)	0 (0)
Less than a high school diploma	1 (6)	0 (0)	2 (100)	0 (0)	4 (8)	1 (4)	0 (0)	0 (0)	2 (4)	0 (0)	1 (50)	0 (0)
Prefer not to say or other	3 (19)	1 (4)	0 (0)	0 (0)	7 (14)	3 (13)	0 (0)	0 (0)	6 (11)	1 (5)	0 (0)	0 (0)
Do you wear a wearable?												
Yes	10 (63)	10 (40)	1 (50)	0 (0)	18 (35)	6 (25)	0 (0)	0 (0)	22 (41)	11 (52)	0 (0)	0 (0)

[a]PD: Parkinson disease.

[b]MS: multiple sclerosis.

[c]COPD: chronic obstructive pulmonary disease.

[d]PFF: proximal femoral fracture.

[e]Additional race and ethnicity demographics include round 1 (2 mixed race participants, 1 White and Latino, and 1 Turkish or Northern Irish), round 2 (1 mixed race participant, Turkish or Northern Irish), and round 3 (1 Jewish participant).

XSL•FO
RenderX

Round 1

Examples of the mapping process per long-term health condition are provided in Table 3. The experiences described by patients for each health condition are briefly summarized here.

For people with PD, scenarios and examples that related to the themes of *physical experience of walking*, *the context of the walking experience*, and *becoming aware of the personal walking experience* were frequently reported. They described how their symptoms affected their balance and noted that they felt they were not able to do as much because of their symptoms, which forced them to slow down. Participants also described how their walking changed depending on the time of day and location. This impacts where and how far they might be willing to go. They also described feeling more conscious of how they walk in terms of speed and quality to counter the changes to their balance. These experiences could be directly related to DMOs describing stride length, step duration, walking speed, and step count (Table 3), which reflect altered balance, speed of movement, and amount of activity.

People with MS frequently described scenarios and examples that related to the themes of *physical experience of walking*, *the context of the walking experience*, and *the walking experience as a link between individual's activities and sense of self*. MS symptoms led people to describe how fatigue impacts their lives, influencing their ability to walk distances and to plan activities. They also spoke about how the symptoms of MS influenced their capacity to undertake activities, depending on how they felt on a daily basis. Finally, they described changes to their balance and perception of distance, which impacted their ability to walk smoothly. The DMOs that most frequently mapped to these themes were step count, stride length, walking speed, and step duration.

Limited data were collected from people with COPD or PFF. However, both groups focused on their physical symptoms, which appeared to link most closely to the amount of activity that they could do. People with COPD described breathlessness as an important symptom that led to struggling to walk and use stairs, among others, thus limiting the amount or duration of their physical activity. For people with COPD, these experiences mapped onto step count and number of walking bouts. People recovering from PFF described changes and adaptations since their fracture, including the need to walk slower, increased fatigue, and having to rest because of this, which mapped onto DMOs, such as walking speed and step count.

Following this step, the issues and concepts that were most often mentioned were listed by the health condition and used to identify the data required to develop visualizations in line with these needs. The most relevant DMOs and contexts were extracted from the Mobilise-D TVS and were drafted into visualizations to represent varying mobility impairments across the 4 health conditions; this process is presented in Figure 2 and Multimedia Appendix 3 [23].

XSL•FO
RenderX

Table 3. Examples of how the patient experiences were mapped to mobility themes and digital mobility outcomes (DMOs).

Example quotation	Mobility themes based on Delgado-Ortiz et al [3]	Mapped DMOs
Patients with PD[a]		
"I walk slowly now so have to allow more time to get anywhere. My family and friends adapt their speed to me so no issues with interactions. I have had some issues in crowded football grounds or trains where I feel unsteady on my feet." [White male participant with PD, aged 61-70 y]	• The physical walking experience • The social experience of walking • Behavioral and attitudinal adaptations resulting from the walking experience • Becoming aware of the personal walking experience	• Stride length • Walking speed • Step duration
"Any way that I can improve my walking which is ok when outside when I concentrate hard on what I am doing, taking long steps and counting out loud. In the house I tend to stand still for ages before I can get my legs to move." [White female participant with PD, aged 71-80 y]	• The physical walking experience • Behavioral and attitudinal adaptations resulting from the walking experience • The context of the walking experience	• Step count; WB[b] duration • Stride length • Step duration
"Walking to the shops used to be 5 minutes up a hill but has become 10 to 15, my family are not impressed by the shuffle." [White male participant with PD, aged 71-80 y]	• The physical walking experience • The social experience of walking • The context of the walking experience • Becoming aware of the personal walking experience	• WB duration • Walking speed
Patients with MS[c]		
"It tends to change depending on your energy, what you've done the day before and the weather. Today has been ok but my body feels tired." [White female participant with MS, aged 31-40 y]	• The physical walking experience • The context of the walking experience	• Walking speed • Step count • Number of WBs
"There is a goldilocks zone. Below minus 3 degrees or above 26 degrees then I walk like I'm drunk. I find this frustrating and left a walking group as a man teased me about it and kept trying to get others in the group to comment on my walking or smell my breath for alcohol. I only walk with my dog alone now." [White female participant with MS, aged 51-60 y]	• The physical walking experience • The mental and emotional walking experience • The social experience of walking • The context of the walking experience • Behavioral and attitudinal adaptations resulting from the walking experience	• Stride length • Walking speed • Step duration
"Walking is normal today. depending on sleep and level of activity walking can become very slow and laboured. Also drop foot can occur and more balance/shoe scuffing issues." [White male participant with MS, aged 41-50 y]	• The physical walking experience • The context of the walking experience.	• Stride length • Walking speed • Step duration
Patients with COPD[d]		
"Yes my walking and climbing stairs impact my breathing on most days I struggle going up stairs I try to keep myself downstairs once I get up." [White female participant with COPD, aged 61-70 y]	• The physical walking experience • The context of the walking experience • Behavioral and attitudinal adaptations resulting from the walking experience	• Step count
"If I do too much I get very breathless and have to rest I also plan what I'm going to do to help my breathing." [White female participant with COPD, aged 61-70 y]	• The physical walking experience • Behavioral and attitudinal adaptations resulting from the walking experience	• Step count • Number of WBs
Patients with PFF[e]		
"My walking has a great influence for my daily activities—I have only few social interactions with friends and no hobbies." [White female participant with PFF, aged 71-80 y]	• The social experience of walking • Behavioral and attitudinal adaptations resulting from the walking experience • The walking experience as a link between an individual's activities and sense of self	• Step count
"I don't walk as fast anymore, I don't jog or cycle anymore. I go to physiotherapy regularly. My relatives rarely talk about my fracture." [White male participant with PFF, aged 71-80 y]	• The physical walking experience • The social experience of walking • Behavioral and attitudinal adaptations resulting from the walking experience	• Step count • Walking speed

[a]PD: Parkinson disease.

[b]WB: walking bout.

XSL•FO
RenderX

[c]MS: multiple sclerosis.
[d]COPD: chronic obstructive pulmonary disease.
[e]PFF: proximal femoral fracture.

Figure 2. Representation of developing the visualizations of round 2 through the feedback from round 1. DMO: digital mobility outcome.

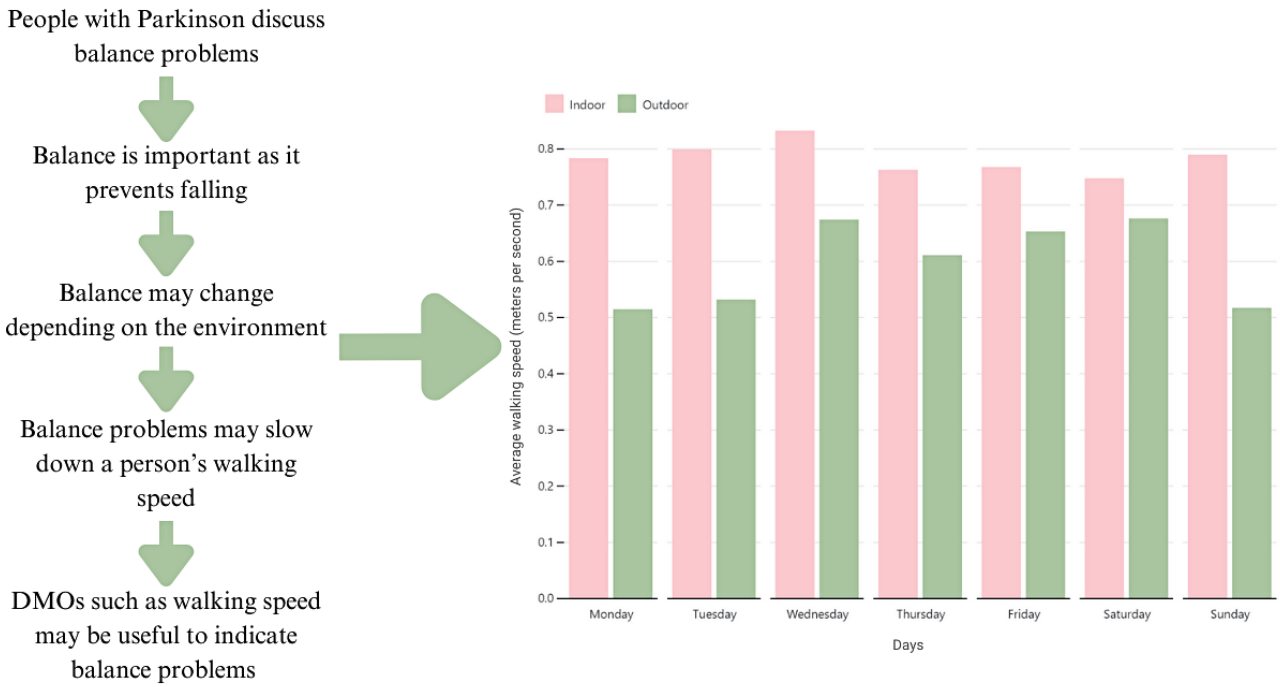

Round 2

Researchers developed and presented 15 visualizations and justifications to participants based on the feedback from round 1. Every long-term health condition was presented with 2 visualizations of fatigue, PD was presented with 2 visualizations on balance, MS was presented with 3 visualizations on doing less, and PFF was presented with 2 visualizations on slowing down (Figure 3 gives the examples). Participants then rated the usefulness and ease of understanding of each visualization. The consensus was reached for the ease of understanding in people with PD for 2 graphs, namely the bar chart showing the indoor versus outdoor data for both balance and fatigue.

The qualitative responses to the draft visualizations were generally positive (summary in Table 4). Participants reported that they were able to understand the visualizations, and some were able to interpret what these visualizations represented. For some, particularly those with PD, more data and variables were deemed necessary to allow a fuller picture of the data being presented. Specifically, people with PD highlighted a desire for additional, contextual information and patient-reported outcome measures to accompany the visualizations. However, there were individuals in each long-term health condition that found the visualizations difficult to understand.

A range of feedback was provided that touched on different aspects of the data visualizations and their method of presentation (Table 5 gives the summary). This was used to improve the visualizations for round 3 (Multimedia Appendix 1). Feedback such as superimposing variables and improving the clarity, accessibility, and visuals of the visualizations was implemented, as seen in Figure 4. Some feedback could not be implemented due to software limitations, such as adding an interactive element to switch between meters and feet.

XSL•FO
RenderX

Figure 3. Examples of draft visualizations presented to participants in round 2. A: Horizontal stacked bar chart showing number of steps across a week and through each day to indicate fatigue B: Bar chart showing average walking speed across a week both indoors and outdoors to indicate fatigue. C: Line chart showing average walking speed across a week both indoors and outdoors to indicate balance.

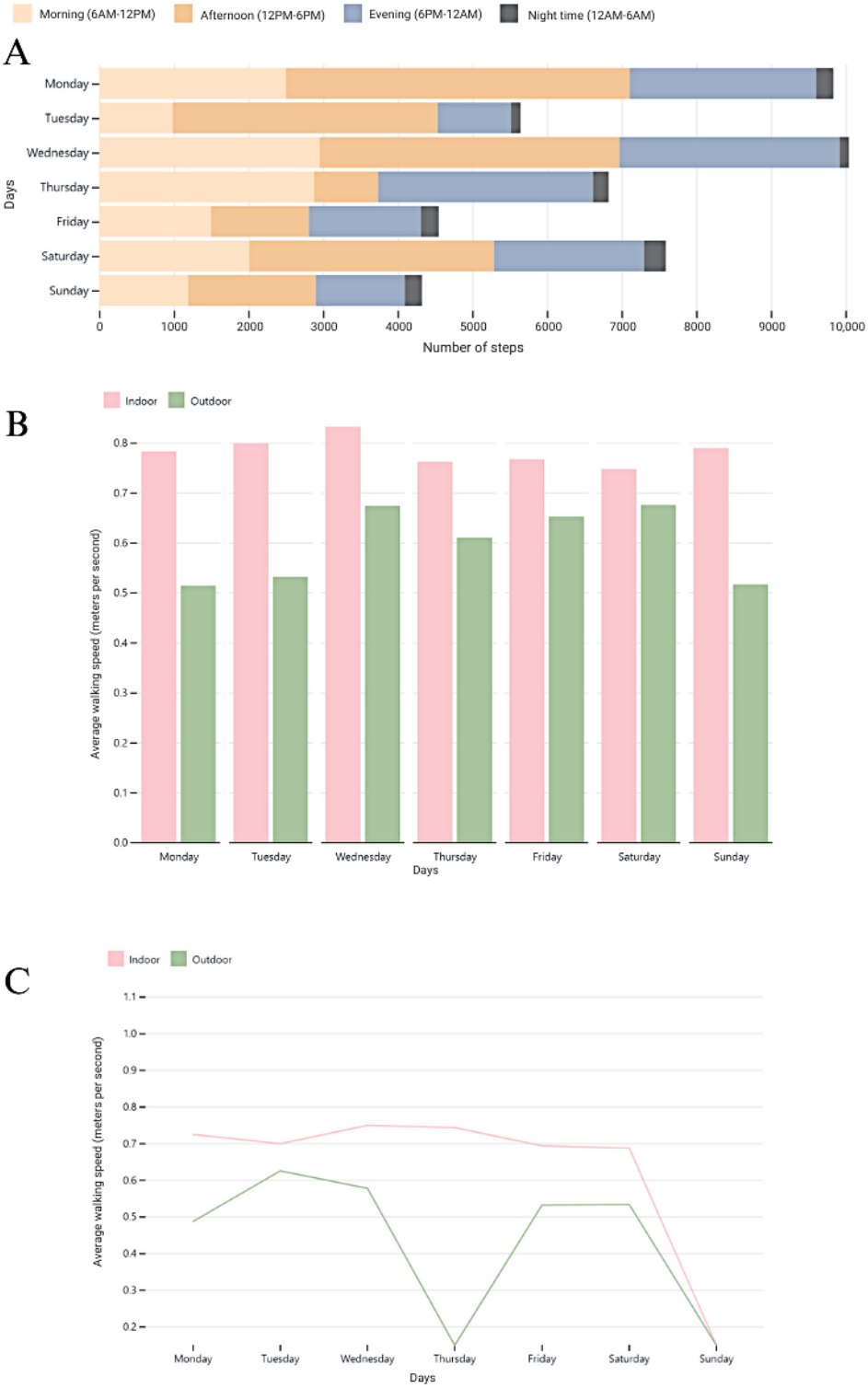

XSL•FO
RenderX

Table 4. A summary of the insights gained from participants during round 2.

Themes	Examples of accompanying quotes
Ability to understand the visualization	• "Easy to understand, maybe brighter colours would make it easier to read. I like that it spans across a week as MS can change day to day." [White female participant with MS[a] aged 18-30 y] • "Easy to read the information and digest it" [White female participant with MS aged 41-50 y]
Ability to interpret the data	• "Suggests that the subject moves around more confidently indoors than out." [White male participant with PD[b] aged 71-80 y] • "It looks like after a few busy days fatigue sets in." [Northern Irish and Turkish female participant with MS aged 41-50 y]
Desire for additional information	• "I think the information could be very useful if there was a way to link to notes on how you were feeling on a given day to see what may have impacted your stats." [White female participant with PD aged 51-60 y] • "OK, this graph also shows indoor walking speeds are consistently higher than outdoor, but it also shows a varying degree of difference on different days, implying another factor at work—maybe the distance walked outside on different days?" [White male participant with PD aged 61-70 y] • "It is useful to be able to see the difference between outdoor and indoor walking. It might be useful to record whether or not a mobility aid is being used and the type of surface being walked on." [White female participant with PD aged 61-70 y]

[a]MS: multiple sclerosis.

[b]PD: Parkinson disease.

XSL•FO
RenderX

Table 5. A summary of suggested changes from participants during round 2 and the changes that were made in round 3 following these suggestions.

Issue raised	Examples of accompanying quotes	Change made
Superimposed variables	• "I think that there has to be a better way to show stride cadence. Speed of step is one important factor, but length of stride is another. So, stepping out to say marching music with a quick march tempo helps to increase pace but what slows people with PD down is often that they take small steps to overcome uneven ground, and this also slows them down." [White female participant with PD[a] aged 61-70 y] • "I would like to see the step length line chart superimposed over this bar chart. Small steps mean more steps over the same distance. As Alexander says 'simples.'" [White female participant with PD aged 71-80 y]	Added more variables per graph to allow for deeper insight into the data shown.
Difficulty with the visualizations	• "Too complicated, too confusing." [White female participant with COPD[b] aged 61-70 y] • "Boring, too academic, too SPSS-like [data analysis software]." [White male participant with PFF[c] aged 71-80 y] • "My response was, oh, boring. More of the same! For the ordinary everyday person having to study these graphs." [White female participant with PD aged >81y]	Made the graphs more visually appealing and added graphics to enhance the visualization.
Interactivity of graphs	• "Assuming that it's plotting the same data, then providing the user with the choice of how the data is displayed would be the most user-friendly approach." [White female participant with PD aged 61-70 y]	Unable to integrate this feedback into round 3.
Improve clarity for the justifications of each graph	• "So many other issues slow walking, all would need to be taken into account, such as gait, balance issues, foot drop, pain, as well as fatigue." [White female participant with MS[d] aged 51-60 y]	Made the link between participant experience and the mapping onto the DMO[e] and context clearer.
Improve descriptions	• "There is no info on how this data has been collected. Does the graph represent one person or is it the average of various persons?" [White female participant with PD aged 71-80 y] • "This graph I found more informative. Is it possible to break it down to age, how long diagnose?" [White female participant with PD aged 51-60 y]	Made it clearer that these are single person's data and are drafted as an example.
Improve accessibility	• "The beige colours seem too close in hue." [White female participant with MS aged 51-60 y] • "Easy to understand perhaps colours could be brighter." [White female participant with PD aged 71-80 y] • "Unable to see colours as colour-blind to parts of spectrum, which is common in MS optic neuritis." [White female participant with MS aged 51-60 y] • "Very Faint" and "Too Pale" [White male participant with PD aged 71-80 y]	Increased font size and made lines on the line chart bolder. Changed to brighter colors that are more accessible for people with color blindness.
Change time distributions	• "The information is good, I like the idea of being able to compare times of day, however I think the nighttime slot is too short and should be at least 10 hours. The evening slot is too long and shouldn't go past 10pm. A better distribution would be Morning 7am-12pm; afternoon 12pm-5pm; evening 5pm-10pm, night 10pm to 7am." [White female participant with PD aged 51-60 y]	Updated the times on event visualizations.

[a]PD: Parkinson disease.

[b]COPD: chronic obstructive pulmonary disease.

[c]PFF: proximal femoral fracture.

[d]MS: multiple sclerosis.

[e]DMO: digital mobility outcome.

XSL•FO
RenderX

Figure 4. The process of refining the draft visualizations for round 3 based on the feedback from round 2. A: Bar chart showing average walking speed across a week both indoors and outdoors to indicate balance. B: Bar chart showing number of steps across a week and throughout each day superimposed by a line chart of average walking speed to indicate fatigue. C: Bar chart showing number of steps across a week to indicate fatigue.

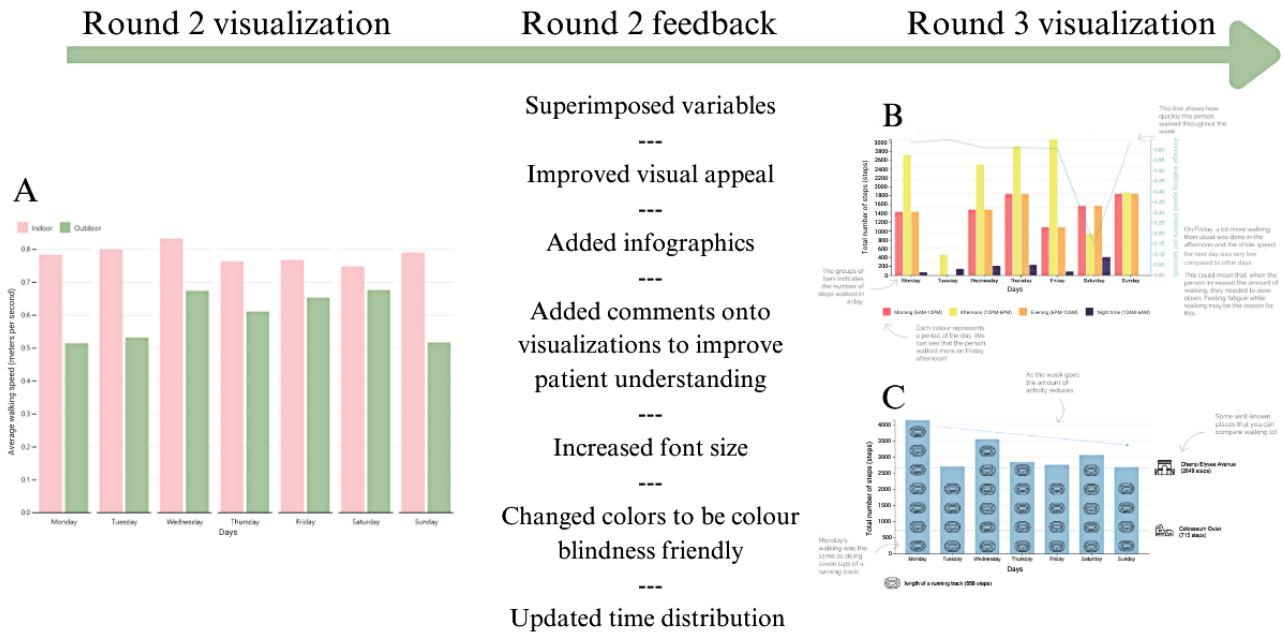

Round 3

Researchers developed and presented 8 visualizations and justifications to participants based on feedback from round 2. Every long-term health condition was presented with 2 visualizations of fatigue, PD was presented with 2 visualizations on balance, MS was presented with 3 visualizations on doing less, and PFF was presented with 2 visualizations on slowing down (Figure 5 gives the examples). Patients rated the usefulness and ease of understanding of each visualization. No consensus was reached during round 3 for any visualization.

Qualitative responses to the draft visualizations were generally positive (summary provided in Table 6). Participants reported that they generally understood the visualizations and appreciated the comments surrounding the visualizations that explained the graph and how to interpret the outcomes. Participants found the differentiation between environment and time of day included in the visualizations to be important, as this influences the presentation of their symptoms.

A range of feedback was provided throughout the comments touching on different aspects of the data visualizations and how they were presented (aggregated summary in Table 7). This feedback could not be used within this study but serves as a recommendation for future studies wishing to visualize mobility data. The feedback mainly surrounded accessibility and explanations, as well as the need for additional data to fully interpret the graphs. Interestingly, this round provided contradictory feedback between participants, with some appreciating the infographic visuals while others found them irritating. This indicates the need for personalization of data visualizations.

XSL•FO
RenderX

Figure 5. Examples of draft visualizations presented to participants in round 3. A: Line chart showing average stride length and average step duration across a week to indicate balance. B: Bar chart showing number of steps across a week and throughout each day superimposed by a line chart of average walking speed to indicate fatigue. C: Bar chart showing number of steps across a week to indicate fatigue.

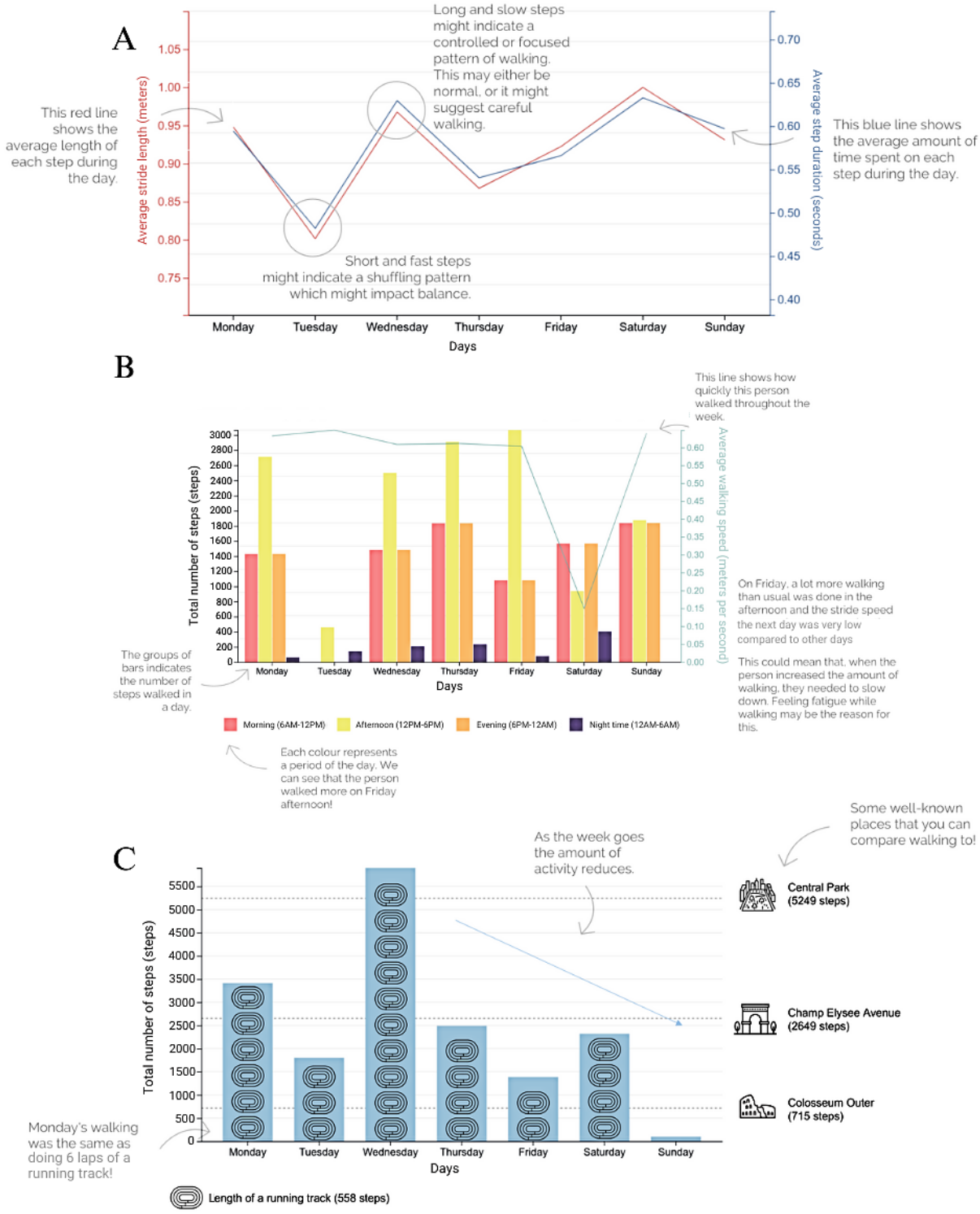

XSL•FO
RenderX

Table 6. A summary of the insights from participants gained during round 3.

Themes	Examples of accompanying quotes
Appreciated the comments to explain the graphs	• "Looks good and explained well" [White female participant with COPD[a] aged 61-70 y] • "I find this graph harder to follow than the previous one. Without the explanatory notes, not sure I'd have grasped the logic behind the walking speed trend line." [White male participant with MS[b] aged 41-50 y]
Easily understood the graphs	• "Simple and clear graphics made graph easy to understand. The purpose of each data item shown was clearly described. I would find this graph very helpful." [White male participant with PD[c] aged 51-60 y] • "Easy to understand, colours differentiate it quite well." [Hispanic female participant with MS aged 18-30 y]
Time of day and location are important to understand symptoms	• "Personally, my stride length varies not just with location (indoors/outdoors) but also where I am in the drug cycle...So it would be useful to not only plot the steps per day but throughout the day." [White female participant with PD aged 61-70 y] • "Round the house I take short quick steps, sometimes with festination and stumbles. At Sainsburys [a supermarket] with a trolley I am much more normal in my stride length and speed" [White female participant with PD aged 71-80 y]

[a]COPD: chronic obstructive pulmonary disease.

[b]MS: multiple sclerosis.

[c]PD: Parkinson disease.

XSL•FO
RenderX

Table 7. A summary of the suggested changes from participants during round 3.

Issue raised	Examples of accompanying quotes
Graphs are small	• "The graph is small and difficult to read therefore difficult to interpret." [White female participant with PD[a] aged 61-70 y]
Explanations are too faint	• "I've had to zoom in to read the comments on all the graphs—could they not be black, rather than grey?" [White female participant with PD aged 51-60 y] • "The explanations of the key features of the graph would be easier to read if in a larger font size and bolder type." [White female participant with PD aged 51-60 y]
More contextual information is needed	• "It would be useful to know if a walking aid was used." [White female participant with MS[b] aged 41-50 y] • "I don't see the purpose of monitoring strides etc unless you show a graph as to how and why the people involved changed their stride length. Parkinson's varies daily, illness affects symptoms, etc." [White female participant with PD aged 61-70 y] • "There doesn't seem to be any way of showing e.g. gradient or quality of surface that you are walking on which limits the usefulness of this graph." [White female participant with PD aged 61-70 y]
Debate surrounding the visuals used in some graphs	• "I think this is very easy to understand. I like visual graphs rather than data." [White female participant with MS aged 41-50 y] • "I am not sure I like the running tracks." [White female participant with MS aged 61-70 y] • "This graph seems to have little point beyond the step count. the graphics (colosseum etc) are unnecessary and, for me, irritating...adding CeeBeeBee [a children's TV channel in the UK] style graphics doesn't add anything useful." [White male participant with PD aged 71-80 y] • "I like this graph. Easy to understand and I really like the Running Track and famous places comparisons for distances." [White male participant with PD aged 51-60 y] • "I love the idea of the equivalent places, but could this be more localised/personalised?" [White female participant with PD aged 51-60 y]
The link between DMO[c] and symptoms is unclear	• "My walking speed can be quite fast even though I'm experiencing fatigue, because of foot drop." [Hispanic female participant with MS aged 18-30 y] • "I think that this doesn't help express fatigue." [White female participant with MS aged 41-50 y] • "People with Parkinson's sometimes shuffle and take small steps, so this will help to understand if or how often they occur." [Asian or Pacific Islander female participant with PD aged 61-70 y]
Questions on the usefulness of the graphs	• "It describes the situation but I'm not sure how I would utilise the information." [White female participant with PD aged 61-70 y] • "Not clear what use this graph for" [White male participant with PD aged 61-70 y] • "Although the graph is easy to read and gives a clear visual representation of the number of steps taken, is this information really useful? If you are tired clearly you will want to walk less." [White female participant with PD aged 51-60 y]
The need for a healthy control's average	• "Might a 'normal' comparator be useful? adjusted for gender and age. i.e. are my steps tiny compared to other women of my age??" [White female participant with PD aged 71-80 y]

[a]PD: Parkinson disease.

[b]MS: multiple sclerosis.

[c]DMO: digital mobility outcome.

Discussion

Principal Findings

This study explored patient preferences on how mobility data derived from wearable devices is visualized. Through the multistage cocreation of visualizations with patients as experts, we embedded the voice of patients throughout the research cycle so that the visualization recommendations would be meaningful to them. While these early insights cannot claim a consensus of patient preferences on mobility data visualizations, the study provides important learnings and recommendations for future mobility visualizations.

Participants were generally able to understand the visualizations and interpret what the data may indicate. The next step would be advancing and personalizing the data visualizations to empower patients to gain actionable insights from their inferences. Feedback consistently requested additional data to be presented to enhance the visualizations; this was achieved through embedding indoor versus outdoor categories in certain visualizations as well as superimposing DMOs on some visualizations in round 3. Participants appreciated this extra detail during round 3; however, this was not sufficient in terms of the information they felt was important to help them apply learnings or actions into their daily lives.

Contextual information, such as how the person felt, when medication was taken, and whether they used a walking aid, would be important when interpreting the visualizations. Previous studies have explored the use of calendars to present medication data over time using a similar approach; it would

XSL•FO
RenderX

be possible to note when medication changes or interventional changes occur by indicating them with icons [19,21]. Alternatively, exploring the granularity of data may be helpful in addressing these comments; for example, exploring when medication was taken throughout the day may lend itself better to visualizing a single day's data with icons added when medication was taken to identify any trends related to medication [18,35]. Previous patient-focused visualization studies within literature also discuss the importance of context and additional factors in determining the meaningfulness of a visualization; hence, it is important for these contexts to be embedded into visualizations [12,15,36]. Typical consumer wearable devices display a myriad of health data, such as number of steps and calories burned; however, these visualizations are often oversimplified and, consequently, do not demonstrate the relationship between multiple factors or acknowledge the presence of external, contextual factors [37]. It is important that researchers do not default to the basic bar charts of consumer wearable devices and instead recognize and highlight the intertwined nature of health, environment, and activity, particularly in those with health conditions whereby this additional information is vital.

Some participants found it difficult to interpret the information from the visualizations and apply it to themselves. This is potentially due to insufficient justifications, descriptions, and accessibility of the visualizations [12]. This indicates the importance of working alongside patient advisors to understand the lived experience of long-term health conditions and develop justifications for what and why the data are displayed [32,38]. A major barrier to understanding for participants was the rationale for why certain DMOs had been chosen to represent a symptom. This indicates the need for future qualitative studies to be conducted to map the patient experience to each DMO. Research to understand the link between patient experience and outcome measures is limited but will be essential for understanding the importance of where DMOs fit into the patient experience. Being able to use visualizations to make sense of the "chaos of lived experience" will be key to encouraging the use of wearable devices within health care [28]. While this mapping process was piloted within round 1 of this study, a larger and more in-depth mapping process is needed to bridge the gap of DMO meaningfulness to patients. Already, the benefits of this mapping process have been seen in other areas of digital health technologies, with additional studies currently ongoing [39-41].

An important barrier to interpreting the visualization was a lack of personal connection with the data. As the data presented to participants was not their own, they found it difficult to understand what was being visualized, as there was no context for what the data implied. This was a recurring theme throughout the qualitative analysis and something that should be addressed in future studies by personalizing visualizations to patients' own data to address these barriers. Visualizing personal data may empower patients to make actionable insights that reflect their own lived experiences, which makes strides toward the potential for personalized health care [42]. This may also address the lack of consensus found in this study, as participants had no familiarity with the data presented, and this may have contributed to poorer understanding and usefulness of the data.

Patients were able to read and interpret bar charts, but their ability to interpret DMOs, such as stride length and walking speed, was hindered due to a lack of understanding of a healthy control's "normal." This is a common theme within clinical data visualizations. Providing a "normal" value may be helpful when goal setting using visualizations; however, issues arise regarding how that "normal" value is set and ensuring it reflects the age, gender, and other characteristics of the patient [21]. Other studies demonstrate that the use of visual cues, such as emojis, and user-friendly visual analogies can improve the understanding of the visualization [36,43]. The use of visual cues throughout data visualization should be of particular importance in future work with difficult-to-visualize outcomes. During round 3 we included comments and additional visual cues to better explain the visualization. The feedback implied that this was well received, yet there is scope for further enhancing these visual cues and comments. For example, the ability to relate to famous monuments as a visual cue was very individualized and should be personalized based upon participant location for future visualizations. Moreover, the accessibility of the graph is important. Due to software limitations, the graphs presented were small and difficult to read for some participants; this is something to be mindful of in future studies.

Future studies should explore alternative visualization strategies. Here, data were presented using graphs to display changes during a week and to represent the dynamic nature of patient experiences. However, certain DMOs may be better conveyed through other visual methods. For example, some studies have used motion capture and videos to visualize gait, which may be an interesting avenue for future mobility-related visualization studies [44,45]. Furthermore, for data to be relevant to a person's immediate context, future work should focus on designing a tool that can provide real-time feedback that supports people's understanding of their condition [27].

Limitations

The COPD and PFF health conditions had a lower representation and did not meet the recommended number of participants for Delphi protocols (10 to 15 participants); this limits the understanding we have of how these groups prefer their data to be visualized [31]. This was due to our recruitment methods not being successful for these health conditions. While results indicate that DMOs and justifications should be personalized based on the symptoms and experience of a long-term health condition, the sample size was not adequate to explore any differences in visualization preferences between groups. Further research is necessary to fully understand any nuances that may arise from these long-term health conditions.

The sample of this study was not representative; participants were largely educated from Western countries (primarily the United Kingdom) and around a third of participants used a wearable device in their daily lives. Due to ethical requests, participant demographics were only collected after participants were recruited; it was not possible to retrospectively recruit and address this disparity. In addition, by nature of a web-based survey, it is likely that the digitally excluded were not well

XSL•FO
RenderX

represented in this study. It is pivotal that research includes and represents the experiences of underserved communities to improve the generalizability of findings [46,47]. Previous data visualization studies have looked at preferences in people with lower educational attainment; this highlights the importance of diversity within data visualization samples, as preferences may vary between groups [48]. For the integration of wearable devices into health care, it will be paramount to understand how best we can visualize DMO data to ensure a wide range of demographics can access and understand the data provided. To address this, future research should aim to provide a more representative sample [49].

It is essential that future studies are conducted with clinicians and those involved in the care of patients. With the rise of commercial wearable devices, data are being presented to clinicians by patients and being dismissed, as there is a lack of understanding on the clinician's side for how to interpret and act on the data [21]. It is imperative that the people involved in a person's care (both the patient and care team) can extract meaningful data from wearable devices. It may be that different forms of visualization may be helpful for clinicians compared to patients; potentially clinicians will be more interested in longitudinal changes, such as baseline to 6 months, than patients. Already studies have explored this for clinic-based sensor data, and this should be explored more widely using real-world data to capture the nuances of mobility [50].

Conclusions

The use of remote mobility monitoring devices is predicted to increase. Hence, how to visualize this data in an understandable and meaningful way to patient populations is of key importance. This study indicates that visualizing these data is possible and can be done in a way that patients understand. While this study did not draft any visualizations that had a strong consensus of understanding and ease of use, we were able to provide several points of recommendation for future visualizations and studies. Particularly, we recommend that future studies explore individualized visualizations, whereby the participant's own data are visualized both individually and in the context of their condition to better enhance understanding.

Recommendations

Despite the exploratory nature of this work, we have been able to develop several initial recommendations for future visualizations in tandem with existing principles that are applicable both for the visualization of mobility data as well as the visualization of health data in general, which include the following:

- Ensure readability—visualizations should be large, easy to read, and contain sufficient color contrasts [51,52].
- Be aware of potential visual impairments within individuals and conditions, for example, contrast sensitivity in MS due to optic neuritis, as well as individual additional needs that must be catered to [53,54].
- Be aware of the need for additional information. Patients understand the variability of their condition best; hence, providing the ability to dynamically interact and insert additional information to explain this variation would be greatly beneficial [12,15,36].
- Think beyond the basics of consumer device visualizations. Experiment with different visualization techniques to enhance patient understanding [19,21,43].
- Ensure a strong rationale. Engage with patients and embed public and patient involvement within the project to ensure meaningful outcomes for patients [38,55].
- Use individualized visualizations. Personalize visualizations to the data of each participant; while not done in this study, participants indicate this as the next step [42].

Acknowledgments

This work was supported by the Mobilise-D project that has received funding from the Innovative Medicines Initiative (IMI) 2 Joint Undertaking (JU; grant agreement 820820). This JU receives support from the European Union's Horizon 2020 research and innovation program and the European Federation of Pharmaceutical Industries and Associations (EFPIA). The content of the current publication reflects the authors' view, and neither IMI nor the European Union, EFPIA, nor any associated partners are responsible for any use that may be made of the information contained herein.

HG was supported by the Fraunhofer Internal Programs (grants Attract 044–602140 and 044–602150). HG also received institutional funding from the Deutsche Forschungsgemeinschaft (German Research Foundation)—SFB 1483—Project-ID 442419336, EmpkinS.

LR, AY, and SDD were also supported by the IMI2 JU project IDEA-FAST (grant agreement 853981). LR, AY, and SDD were supported by the National Institute for Health and Care Research (NIHR) Newcastle Biomedical Research Centre based at The Newcastle upon Tyne Hospital National Health Service (NHS) Foundation Trust, Newcastle University, and the Cumbria, Northumberland, and Tyne and Wear NHS Foundation Trust and the NIHR/Wellcome Trust Clinical Research Facility infrastructure at Newcastle upon Tyne Hospitals NHS Foundation Trust. LR received funding from the NIHR Senior Investigator Awards (2020-2024 and 2024-2028). EB was supported by the NIHR Sheffield Biomedical-Research Centre and Clinical Research Facility. The views expressed are those of the authors and not necessarily those of the NHS, NIHR, or the Department of Health and Social Care.

AY was supported by the Michael J Fox Foundation (grant references MJFF-022735 and MJFF-022737), the UK Engineering and Physical Sciences Research Council (grant EP/X036146/1), and Intercept Pharmaceuticals, NIHR Health Technology

XSL•FO
RenderX

Assessment (grant reference NIHR129175). She received grants from Parkinson's UK, Weston Brain Institute, Dunhill Medical Trust, and Cure Parkinson's Trust. She had received honoraria from GE Healthcare as speaker and chair outside the study.

LD-O and JG-A are supported by ISGlobal which acknowledge support from the grant CEX2023-0001290-S funded by MCIN/AEI/10.13039/501100011033, and support from the Generalitat de Catalunya through the Centres de Recerca de Catalunya Program.

Data Availability

The data collected and analyzed within this study are available upon request from the corresponding author.

Conflicts of Interest

JE is employed by McRoberts and holds shares. McRoberts is the manufacturer of the DynaPort sensor used to generate the data presented. SDD reports on consultancy activity with Hoffmann-La Roche Ltd outside of this study. HG received personal fees from Zambon GmbH outside of the submitted work.

Multimedia Appendix 1
Examples of text and visualizations presented to participants.
[DOCX File , 456 KB - formative_v9i1e68782_app1.docx]

Multimedia Appendix 2
Coding framework, based on Delgado-Ortiz et al [3] and coding framework of health condition–specific walking elements (foot drop and freezing, among others).
[DOCX File , 16 KB - formative_v9i1e68782_app2.docx]

Multimedia Appendix 3
Technical validation study data requirements.
[DOCX File , 16 KB - formative_v9i1e68782_app3.docx]

References

1. Viceconti M, Tome M, Dartee W, Knezevic I, Hernandez Penna S, Mazzà C, Mobilise-D Consortium. On the use of wearable sensors as mobility biomarkers in the marketing authorization of new drugs: a regulatory perspective. Front Med (Lausanne) 2022;9:996903 [FREE Full text] [doi: 10.3389/fmed.2022.996903] [Medline: 36213641]
2. Decade of healthy ageing: baseline report. World Health Organization. 2020. URL: https://www.who.int/publications/i/item/9789240017900 [accessed 2025-04-15]
3. Delgado-Ortiz L, Polhemus A, Keogh A, Sutton N, Remmele W, Hansen C, et al. Listening to the patients' voice: a conceptual framework of the walking experience. Age Ageing 2023 Jan 08;52(1):afac233 [FREE Full text] [doi: 10.1093/ageing/afac233] [Medline: 36729471]
4. Lusardi MM. Is walking speed a vital sign? Absolutely!. Geriatr Rehabil 2012;28(2):37. [doi: 10.1097/TGR.0b013e31824385a4]
5. Buttery SC, Williams PJ, Alghamdi SM, Philip KE, Perkins A, Kallis C, et al. Investigating the prognostic value of digital mobility outcomes in patients with chronic obstructive pulmonary disease: a systematic literature review and meta-analysis. Eur Respir Rev 2023 Dec 31;32(170):230134 [FREE Full text] [doi: 10.1183/16000617.0134-2023] [Medline: 37993126]
6. Buckley C, Alcock L, McArdle R, Rehman R, Del Din S, Mazzà C, et al. The role of movement analysis in diagnosing and monitoring neurodegenerative conditions: insights from gait and postural control. Brain Sci 2019 Feb 06;9(2):34 [FREE Full text] [doi: 10.3390/brainsci9020034] [Medline: 30736374]
7. Polhemus A, Delgado-Ortiz L, Brittain G, Chynkiamis N, Salis F, Gaßner H, Mobilise-D. Walking on common ground: a cross-disciplinary scoping review on the clinical utility of digital mobility outcomes. NPJ Digit Med 2021 Oct 14;4(1):149 [FREE Full text] [doi: 10.1038/s41746-021-00513-5] [Medline: 34650191]
8. Davis Giardina T, Menon S, Parrish DE, Sittig DF, Singh H. Patient access to medical records and healthcare outcomes: a systematic review. J Am Med Inform Assoc 2014;21(4):737-741 [FREE Full text] [doi: 10.1136/amiajnl-2013-002239] [Medline: 24154835]
9. Delbanco T, Walker J, Bell SK, Darer JD, Elmore JG, Farag N, et al. Inviting patients to read their doctors' notes: a quasi-experimental study and a look ahead. Ann Intern Med 2012 Oct 02;157(7):461-470 [FREE Full text] [doi: 10.7326/0003-4819-157-7-201210020-00002] [Medline: 23027317]
10. Earnest MA, Ross SE, Wittevrongel L, Moore LA, Lin C. Use of a patient-accessible electronic medical record in a practice for congestive heart failure: patient and physician experiences. J Am Med Inform Assoc 2004;11(5):410-417 [FREE Full text] [doi: 10.1197/jamia.M1479] [Medline: 15187074]

XSL•FO
RenderX

11. Mold F, de Lusignan S, Sheikh A, Majeed A, Wyatt JC, Quinn T, et al. Patients' online access to their electronic health records and linked online services: a systematic review in primary care. Br J Gen Pract 2015 Mar;65(632):e141-e151 [FREE Full text] [doi: 10.3399/bjgp15X683941] [Medline: 25733435]
12. Polhemus A, Novak J, Majid S, Simblett S, Morris D, Bruce S, et al. Data visualization for chronic neurological and mental health condition self-management: systematic review of user perspectives. JMIR Ment Health 2022 Apr 28;9(4):e25249 [FREE Full text] [doi: 10.2196/25249] [Medline: 35482368]
13. Manta C, Patrick-Lake B, Goldsack J. Digital measures that matter to patients: a framework to guide the selection and development of digital measures of health. Digit Biomark 2020 Sep 15;4(3):69-77 [FREE Full text] [doi: 10.1159/000509725] [Medline: 33083687]
14. Patient-focused drug development: collecting comprehensive and representative input. Food and Drug Administration. 2020. URL: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/patient-focused-drug-development-collecting-comprehensive-and-representative-input [accessed 2025-04-15]
15. Koopman RJ, Canfield SM, Belden JL, Wegier P, Shaffer VA, Valentine KD, et al. Home blood pressure data visualization for the management of hypertension: designing for patient and physician information needs. BMC Med Inform Decis Mak 2020 Aug 18;20(1):195 [FREE Full text] [doi: 10.1186/s12911-020-01194-y] [Medline: 32811489]
16. Keim DA, Mansmann F, Schneidewind J, Thomas J, Ziegler H. Visual analytics: scope and challenges. In: Simoff SJ, Böhlen MH, Mazeika A, editors. Visual Data Mining: Theory, Techniques and Tools for Visual Analytics. Cham, Switzerland: Springer; 2008:76-90.
17. Mikolaizak AS, Rochester L, Maetzler W, Sharrack B, Demeyer H, Mazzà C, clinical validation study (WP4) on behalf of Mobilise-D consortium. Connecting real-world digital mobility assessment to clinical outcomes for regulatory and clinical endorsement-the Mobilise-D study protocol. PLoS One 2022;17(10):e0269615 [FREE Full text] [doi: 10.1371/journal.pone.0269615] [Medline: 36201476]
18. Keogh A, Taraldsen K, Caulfield B, Vereijken B. It's not about the capture, it's about what we can learn": a qualitative study of experts' opinions and experiences regarding the use of wearable sensors to measure gait and physical activity. J Neuroeng Rehabil 2021 May 11;18(1):78 [FREE Full text] [doi: 10.1186/s12984-021-00874-8] [Medline: 33975600]
19. Wanderer JP, Nelson SE, Ehrenfeld JM, Monahan S, Park S. Clinical data visualization: the current state and future needs. J Med Syst 2016 Dec;40(12):275. [doi: 10.1007/s10916-016-0643-x] [Medline: 27787779]
20. Qin X, Luo Y, Tang N, Li G. Making data visualization more efficient and effective: a survey. VLDB J 2019 Nov 19;29(1):93-117. [doi: 10.1007/S00778-019-00588-3]
21. Seals A, Pilloni G, Kim J, Sanchez R, Rizzo JR, Charvet L, et al. ‘Are they doing better in the clinic or at home?’: understanding clinicians’ needs when visualizing wearable sensor data used in remote gait assessments for people with multiple sclerosis. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. 2022 Presented at: CHI '22; April 29-May 5, 2022; New Orleans, LA p. 1-16 URL: https://dl.acm.org/doi/10.1145/3491102.3501989 [doi: 10.1145/3491102.3501989]
22. Sadhu S, Solanki D, Brick LA, Nugent NR, Mankodiya K. Designing a clinician-centered wearable data dashboard (CarePortal): participatory design study. JMIR Form Res 2023 Dec 05;7:e46866 [FREE Full text] [doi: 10.2196/46866] [Medline: 38051573]
23. Mazzà C, Alcock L, Aminian K, Becker C, Bertuletti S, Bonci T, et al. Technical validation of real-world monitoring of gait: a multicentric observational study. BMJ Open 2021 Dec 02;11(12):e050785 [FREE Full text] [doi: 10.1136/bmjopen-2021-050785] [Medline: 34857567]
24. Kirk C, Küderle A, Micó-Amigo ME, Bonci T, Paraschiv-Ionescu A, Ullrich M, Mobilise-D consortium. Mobilise-D insights to estimate real-world walking speed in multiple conditions with a wearable device. Sci Rep 2024 Jan 19;14(1):1754 [FREE Full text] [doi: 10.1038/s41598-024-51766-5] [Medline: 38243008]
25. Rochester L, Mazzà C, Mueller A, Caulfield B, McCarthy M, Becker C, et al. A roadmap to inform development, validation and approval of digital mobility outcomes: the mobilise-D approach. Digit Biomark 2020 Nov 26;4(Suppl 1):13-27 [FREE Full text] [doi: 10.1159/000512513] [Medline: 33442578]
26. Keogh A, Alcock L, Brown P, Buckley E, Brozgol M, Gazit E, et al. Acceptability of wearable devices for measuring mobility remotely: observations from the mobilise-D technical validation study. Digit Health 2023 Feb 01;9:20552076221150745 [FREE Full text] [doi: 10.1177/20552076221150745] [Medline: 36756644]
27. Dove G, Guthmann MR, Charvet L, Nov O, Pilloni G. "Data is one thing, but I want to know the story behind": designing for self-tracking and remote patient monitoring in the context of multiple sclerosis care. In: Proceedings of the 2024 ACM Designing Interactive Systems Conference. 2024 Presented at: DIS '24; July 1-5, 2024; Copenhagen, Denmark p. 597-618 URL: https://doi.org/10.1145/3643834.3661499 [doi: 10.1145/3643834.3661499]
28. Vafeiadou J, Vasalou A, Roussos G. Self-tracking in Parkinson's the lived efforts of self-management. Proc ACM Hum Comput Interact 2021 Apr 22;5(CSCW1):1-25. [doi: 10.1145/3449158]
29. Keogh A, Mc Ardle R, Diaconu MG, Ammour N, Arnera V, Balzani F, Mobilise-D consortium. Mobilizing patient and public involvement in the development of real-world digital technology solutions: tutorial. J Med Internet Res 2023 Oct 27;25:e44206 [FREE Full text] [doi: 10.2196/44206] [Medline: 37889531]

XSL•FO
RenderX

30. Santaguida P, Dolovich L, Oliver D, Lamarche L, Gilsing A, Griffith LE, et al. Protocol for a Delphi consensus exercise to identify a core set of criteria for selecting health related outcome measures (HROM) to be used in primary health care. BMC Fam Pract 2018 Sep 04;19(1):152 [FREE Full text] [doi: 10.1186/s12875-018-0831-5] [Medline: 30185172]
31. Trevelyan EG, Robinson PN. Delphi methodology in health research: how to do it? Eur J Integr Med 2015 Aug;7(4):423-428 [FREE Full text] [doi: 10.1016/j.eujim.2015.07.002]
32. Jones J, Hunter D. Consensus methods for medical and health services research. BMJ 1995 Aug 05;311(7001):376-380 [FREE Full text] [doi: 10.1136/bmj.311.7001.376] [Medline: 7640549]
33. Micó-Amigo ME, Bonci T, Paraschiv-Ionescu A, Ullrich M, Kirk C, Soltani A, Mobilise-D consortium. Assessing real-world gait with digital technology? Validation, insights and recommendations from the Mobilise-D consortium. J Neuroeng Rehabil 2023 Jun 14;20(1):78 [FREE Full text] [doi: 10.1186/s12984-023-01198-5] [Medline: 37316858]
34. Kluge F, Del Din S, Cereatti A, Gaßner H, Hansen C, Helbostad JL, Mobilise-D consortium. Consensus based framework for digital mobility monitoring. PLoS One 2021 Aug 20;16(8):e0256541 [FREE Full text] [doi: 10.1371/journal.pone.0256541] [Medline: 34415959]
35. Packer E, Debelle H, Bailey HG, Ciravegna F, Ireson N, Evers J, et al. Translating digital healthcare to enhance clinical management: a protocol for an observational study using a digital health technology system to monitor medication adherence and its effect on mobility in people with Parkinson's. BMJ Open 2023 Sep 04;13(9):e073388 [FREE Full text] [doi: 10.1136/bmjopen-2023-073388] [Medline: 37666560]
36. Reading Turchioe M, Grossman LV, Myers AC, Baik D, Goyal P, Masterson Creber RM. Visual analogies, not graphs, increase patients' comprehension of changes in their health status. J Am Med Inform Assoc 2020 May 01;27(5):677-689 [FREE Full text] [doi: 10.1093/jamia/ocz217] [Medline: 31999316]
37. Fitbit: health and fitness. Apple App Store. URL: https://apps.apple.com/us/app/fitbit-health-fitness/id462638897 [accessed 2024-04-29]
38. Small N, Ong BN, Lewis A, Allen D, Bagshaw N, Nahar P, DEPEND team. Co-designing new tools for collecting, analysing and presenting patient experience data in NHS services: working in partnership with patients and carers. Res Involv Engagem 2021 Nov 27;7(1):85 [FREE Full text] [doi: 10.1186/s40900-021-00329-3] [Medline: 34838128]
39. Aryal S, Blankenship JM, Bachman SL, Hwang S, Zhai Y, Richards JC, et al. Patient-centricity in digital measure development: co-evolution of best practice and regulatory guidance. NPJ Digit Med 2024 May 16;7(1):128 [FREE Full text] [doi: 10.1038/s41746-024-01110-y] [Medline: 38755349]
40. Cesnakova L, Meadows K, Avey S, Barrett J, Calimlim B, Chatterjee M, et al. A patient-centred conceptual model of nocturnal scratch and its impact in atopic dermatitis: a mixed-methods study supporting the development of novel digital measurements. Skin Health Dis 2023 Oct;3(5):e262 [FREE Full text] [doi: 10.1002/ski2.262] [Medline: 37799371]
41. Measuring what matters - in whom it matters - ensuring diversity and equity in Parkinson's disease research and care. Brain and Movement Newcastle University. URL: https://www.bam-ncl.co.uk/mjff-project [accessed 2024-09-13]
42. Zhao Y, Parvinzamir F, Wilson S, Wei H, Deng Z, Portokallidis N, et al. Integrated visualisation of wearable sensor data and risk models for individualised health monitoring and risk assessment to promote patient empowerment. J Vis 2016 Oct 27;20(2):405-413. [doi: 10.1007/S12650-016-0402-6]
43. Stonbraker S, Porras T, Schnall R. Patient preferences for visualization of longitudinal patient-reported outcomes data. J Am Med Inform Assoc 2020 Feb 01;27(2):212-224 [FREE Full text] [doi: 10.1093/jamia/ocz189] [Medline: 31670816]
44. Bicer M, Phillips AT, Melis A, McGregor A, Modenese L. Generative deep learning applied to biomechanics: a new augmentation technique for motion capture datasets. J Biomech 2022 Nov;144:111301 [FREE Full text] [doi: 10.1016/j.jbiomech.2022.111301] [Medline: 36201910]
45. Guayacán LC, Martínez F. Visualising and quantifying relevant parkinsonian gait patterns using 3D convolutional network. J Biomed Inform 2021 Nov;123:103935 [FREE Full text] [doi: 10.1016/j.jbi.2021.103935] [Medline: 34699990]
46. Rochester L, Carroll C. Implications of research that excludes under-served populations. Nat Rev Neurol 2022 Aug 29;18(8):449-450. [doi: 10.1038/s41582-022-00688-9] [Medline: 35768654]
47. Goodwin VA, Low MS, Quinn TJ, Cockcroft EJ, Shepherd V, Evans PH, et al. Including older people in health and social care research: best practice recommendations based on the INCLUDE framework. Age Ageing 2023 Jun 01;52(6):afad082 [FREE Full text] [doi: 10.1093/ageing/afad082] [Medline: 37261448]
48. McCaffery KJ, Dixon A, Hayen A, Jansen J, Smith S, Simpson JM. The influence of graphic display format on the interpretations of quantitative risk information among adults with lower education and literacy: a randomized experimental study. Med Decis Making 2012;32(4):532-544. [doi: 10.1177/0272989X11424926] [Medline: 22074912]
49. Koopman RJ, Wakefield BJ, Johanning JL, Keplinger LE, Kruse RL, Bomar M, et al. Implementing home blood glucose and blood pressure telemonitoring in primary care practices for patients with diabetes: lessons learned. Telemed J E Health 2014 Mar;20(3):253-260 [FREE Full text] [doi: 10.1089/tmj.2013.0188] [Medline: 24350806]
50. Silva-Batista C, Harker G, Vitorio R, Studer M, Whetten B, Lapidus J, et al. Mobility Rehab visual feedback system for gait rehabilitation in older adults. J Neuroeng Rehabil 2023 Oct 24;20(1):144 [FREE Full text] [doi: 10.1186/s12984-023-01260-2] [Medline: 37875971]
51. Munzner T. Visualization Analysis and Design: Opportunities for Domestic Investments in Water and Sanitation for the Poor. Boca Raton, FL: CRC Press; 2014.

XSL•FO
RenderX

52. Web Content Accessibility Guidelines (WCAG) overview. W3C Web Accessibility Initiative (WAI). URL: https://www.w3.org/WAI/standardsguidelines/wcag [accessed 2020-10-02]
53. Goldstein EB, Brockmole JR. Sensation and Perception. 10th edition. New York, NY: Wadsworth Publishing; 2017.
54. Toosy AT, Mason DF, Miller DH. Optic neuritis. Lancet Neurol 2014 Jan;13(1):83-99 [FREE Full text] [doi: 10.1016/s1474-4422(13)70259-x]
55. Chang S, Gray L, Alon N, Torous J. Patient and clinician experiences with sharing data visualizations integrated into mental health treatment. Soc Sci 2023 Nov 22;12(12):648 [FREE Full text] [doi: 10.3390/socsci12120648]

Abbreviations

COPD: chronic obstructive pulmonary disease
DMO: digital mobility outcome
MS: multiple sclerosis
PD: Parkinson disease
PFF: proximal femoral fracture
PPAG: Public and Patient Advisory Group
TVS: technical validation study

Edited by A Mavragani; submitted 13.11.24; peer-reviewed by A Adonis, J Bowen; comments to author 04.02.25; revised version received 19.02.25; accepted 11.03.25; published 09.05.25.

Please cite as:
Lumsdon J, Wilson C, Alcock L, Becker C, Benvenuti F, Bonci T, van den Brande K, Brittain G, Brown P, Buckley E, Caruso M, Caulfield B, Cereatti A, Delgado-Ortiz L, Del Din S, Evers J, Garcia-Aymerich J, Gaßner H, Gur Arieh T, Hansen C, Hausdorff JM, Hiden H, Hume E, Kirk C, Maetzler W, Megaritis D, Rochester L, Scott K, Sharrack B, Sutton N, Vereijken B, Vogiatzis I, Yarnall A, Keogh A, Cantu A
Cocreating the Visualization of Digital Mobility Outcomes: Delphi-Type Process With Patients
JMIR Form Res 2025;9:e68782
URL: https://formative.jmir.org/2025/1/e68782
doi:10.2196/68782
PMID:

©Jack Lumsdon, Cameron Wilson, Lisa Alcock, Clemens Becker, Francesco Benvenuti, Tecla Bonci, Koen van den Brande, Gavin Brittain, Philip Brown, Ellen Buckley, Marco Caruso, Brian Caulfield, Andrea Cereatti, Laura Delgado-Ortiz, Silvia Del Din, Jordi Evers, Judith Garcia-Aymerich, Heiko Gaßner, Tova Gur Arieh, Clint Hansen, Jeffrey M Hausdorff, Hugo Hiden, Emily Hume, Cameron Kirk, Walter Maetzler, Dimitrios Megaritis, Lynn Rochester, Kirsty Scott, Basil Sharrack, Norman Sutton, Beatrix Vereijken, Ioannis Vogiatzis, Alison Yarnall, Alison Keogh, Alma Cantu. Originally published in JMIR Formative Research (https://formative.jmir.org), 09.05.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Developing the MyCancerGene Digital Health Portal to Improve Patients' Understanding of Germline Cancer Genetic Test Results: Development, User, and Usability Testing Study

Phillip Trieu[1]; Dominique Fetzer[1], BA; Briana McLeod[1], MPH; Kathryn Schweickert[1]; Lauren Gutstein[1], MS; Brian Egleston[2], PhD; Susan Domchek[1], MD; Linda Fleisher[2], MPH, PhD; Lynne Wagner[3], PhD; Kuang-Yi Wen[4], PhD; Cara Cacioppo[1], MS, LCGC; Jessica E Ebrahimzadeh[1], MS, LCGC; Dana Falcone[1], MS, LCGC; Claire Langer[1], MS, LSGC; Elisabeth Wood[1], MS, LCGC; Kelsey Karpink[1], BS; Shelby Posen[1], BS; Enida Selmani[1], BS; Angela R Bradbury[5], MD

[1]Division of Hematology-Oncology, Abramson Cancer Center, The University of Pennsylvania, Philadelphia, PA, United States

[2]Fox Chase Cancer Center, Temple University, Philadelphia, PA, United States

[3]Grillings School of Global Public Health, University of North Carolina, Chapel Hill, NC, United States

[4]Thomas Jefferson University, Philadelphia, PA, United States

[5]University of Pennsylvania, Philadelphia, PA, United States

Corresponding Author:
Angela R Bradbury, MD
University of Pennsylvania
3400 Civic Center Blvd
Philadelphia, PA, 19104
United States
Phone: 1 2156156899
Email: Angela.Bradbury@pennmedicine.upenn.edu

Abstract

Background: The use of multigene panels has significantly increased the likelihood that genetic testing will leave patients with uncertainties regarding test interpretation, implications, and recommendations, which will change over time. Effective longitudinal care models are needed to provide patients with updated information and to obtain patient and family history updates.

Objective: To bridge this gap, we aimed to develop a patient- and genetic provider–informed digital genetic health portal (GHP), MyCancerGene, to improve longitudinal patient understanding of and responses to genetic testing.

Methods: We used a 5-step process to develop MyCancerGene. To better understand their interest in and willingness to use a digital GHP, we surveyed 307 patients who completed genetic testing (step 1). We completed qualitative interviews with 10 patients and a focus group with 17 genetic providers to inform the content and function of MyCancerGene (step 2). Next, we developed initial intervention content (step 3) and completed user testing of intervention content with 25 providers and 28 patients (step 4). After developing the prototype intervention, we completed usability testing with 8 patients for their feedback on the final content, functions, and ease of use (step 5).

Results: In surveys conducted in step 1, 90% of patients with positive results reported interest in a digital GHP, and over 75% of participants with variants of uncertain significance or uninformative negative results reported similar interest. The most frequently reported advantages among patients were increasing accessibility, convenience, and efficiency (103/224, 46%); keeping genetic information organized (54/224, 24.1%); and increasing or maintaining patient understanding of the information (38/224, 17%). In qualitative interviews (step 2), both patients and genetic providers endorsed the benefit of the tool for updating personal and family history and for providers to share new risk information, test interpretation, or other medical changes. Patient and provider input informed eight key components of the tool: (1) Landing Page, (2) Summary of Care page, (3) My Genetic Test Results page, (4) My Family History page, (5) Provide an Update page, (6) Review an Update page, (7) Resources page, and (8) the Screenings Tracker. They also recommended key functions, including the ability to download and print materials and the inclusion of reminders and engagement functions. Potential challenges identified by patients included privacy and security concerns (67/206, 32.5%) and the potential for electronic information to generate distress (20/206, 9.7%). While patients were

XSL•FO
RenderX

comfortable with updates (ie, even variant reclassification upgrades or clinically significant results), 44% (11/25) of genetic providers were uncomfortable sharing variant reclassification upgrades through MyCancerGene.

Conclusions: MyCancerGene, a patient-centered digital GHP, was developed with extensive patient and genetic provider feedback and designed to enhance longitudinal patient understanding of and affective and behavioral responses to genetic testing, particularly in the era of evolving evidence and risk information.

(JMIR Form Res 2025;9:e56282) doi:10.2196/56282

KEYWORDS

genetic health portal; digital intervention development; health portal; usability; digital health; germline cancer; multigene panels; clinical practice; likelihood; effectiveness; medical history; genetic medicine; risk information; digital tool; intervention; longitudinal care; patient-centered content; electronic information

Introduction

Background

Advances in basic science in genetics have shown great promise in improving human health and reducing the burden of cancer [1,2]. The promise of precision medicine is the ability to tailor treatment or screening of individual patients based on their genotype. With these advances, there is an increasing need for multidisciplinary translational research that focuses on how to advance precision medicine discoveries into clinical practice and to capitalize on connectivity (eg, digital health solutions) in ways that benefit the health of entire populations [3-5].

One of the challenges with recent advances in clinical sequencing, including hereditary cancer multigene panel testing (MGPT), is the increasing likelihood that genetic testing will leave many patients with uncertainties regarding interpretation, implications, and recommendations, all of which will change over time. Many of the genes included in multigene panels are moderate-penetrance genes, increasing the risk of cancer by only 2- to 4-fold, and in many cases, risk estimates are based on limited data and continue to evolve over time [6-8]. There is also uncertainty regarding optimal screening, given medical management recommendations, such as those published by the National Comprehensive Cancer Network and others, are continually evolving as new data emerge. Thus, even positive results leave many unknowns regarding cancer risks, optimal management, and the value of testing unaffected relatives [6,9,10]. In addition, multigene panels have been associated with higher rates of variants of uncertain significance (VUS), which are difficult for patients and genetic providers to understand, and can be reclassified over time [11-15].

This clinical transition from discrete (ie, single gene) to broad (ie, multi, whole genome) sequencing in clinical genetic testing presents challenges that will only increase as precision medicine applications expand. With broader testing, an increasing number of patients will be left at a risk of misunderstanding, uncertainty, and evolving interpretations and recommendations [8,16]. While these uncertainties will be clarified over time, we lack clinical care models to maintain longitudinal communication with patients to provide updated information regarding their genetic test result or medical recommendations and to obtain personal and medical history updates [16-19]. For example, in the current standard of care, if a genetic variant interpretation is updated (eg, VUS to benign or VUS to likely pathogenic result), the commercial laboratory contacts the ordering genetic provider, who in turn contacts the patient [20,21]. This model is not sustainable and may not be feasible due to outdated patient contact information or change in employment of providers. Alternatively, patients can contact their provider periodically, but this places a burden on patients and has not been proven to be successful.

Equally important, while many studies have shown limited psychosocial distress with genetic testing, robust longitudinal data on cognitive, affective, and behavioral outcomes in diverse and representative patient populations and with MGPT are limited [22,23]. Recent studies suggest some subgroups are at risk for greater posttest distress with MGPT, including patients with positive results, a history of cancer, and lower formal education [24-26]. Patients with positive and VUS results have also demonstrated greater distress and uncertainty [15,27,28]. Furthermore, there is an increased use of remote counseling (eg, phone or videoconference), digital alternatives, and streamlined counseling models, which may introduce additional short-term or longitudinal knowledge gaps or testing-related distress [29-33]. Thus, effective, evidence-based alternative strategies for longitudinal communication and care for genetic patients are critical to realizing the promise of precision medicine.

The Institute of Medicine has highlighted the importance of patient-centered care and improved transparency to enhance the delivery and outcomes of medical care [34-36]. Data from the Pew Research Center in 2021 report that 93% of Americans use the internet, 80% have searched the web for health information, and patients frequently have high interest in communicating electronically with genetic providers [37-41]. Thus, interactive health communication applications, also known as digital health solutions or tools (eg, apps, health care portals, and educational or decision aid interventions), have been implemented to facilitate this goal of enhanced transparency and communication [36,42-44]. Digital health tools may be particularly effective in chronic disease management, where longitudinal care is crucial, when informational needs change over time or vary among patients, when coping and adjustment are part of the 'journey,' and when improved medical outcomes require changing health behaviors [43]. Studies from a variety of chronic disease settings, excluding genetics, have shown that digital health tools can improve knowledge [43], self-efficacy [43,45,46], satisfaction [45,47], clinical outcomes [43,47-50], and unmet communication needs [45,48]. Recent randomized studies have

XSL•FO
RenderX

shown reductions in distress with digital health tools [46,49,51]. Furthermore, digital health tools may provide education in a way that is more simple, accessible, and salient, resulting in better knowledge retention [43].

Yet, many studies have identified limitations and knowledge gaps related to digital health tools [36,42,44,45,52,53]. Various types of digital health tools have been used, and electronic patient portals with limited functions (ie, providing access to records only) may have less impact on outcomes [36]. Most randomized studies involving health portals have been relatively small or report limited outcomes or benefits [37,45,52]. Furthermore, in some studies, participants rarely or never log on to the portal [50], which may impact outcomes and diminish power and has been associated with factors such as race, ethnicity, and level of education [50,54-58]. In addition, studies including socioeconomically diverse patient populations are needed, as minority patients are significantly less likely to use electronic patient portals, which could exacerbate health care disparities [37,45,54,59-62]. Importantly, at the time we developed MyCancerGene, there were no published studies evaluating longitudinal digital health tools in clinical cancer genetics, which shares many of the characteristics of chronic illness (eg, evolving information over time, adjustment and need for behavior change, and communication with relatives and other health care providers). Thus, we propose that cancer genetic testing is an ideal clinical context to study the benefits and limitations of a theoretically (ie, diffusion of innovation theory) and patient- and genetic provider–informed patient-centered longitudinal digital health tool to optimize patients' outcomes following cancer genetic testing and the clinical implementation of precision medicine.

Objective

To address this clinically substantial gap in the delivery of genetic medicine, we sought to obtain patient and genetic provider input to develop a patient- and provider-informed longitudinal digital genetic health portal (GHP) called MyCancerGene to enhance longitudinal patient understanding of and affective and behavioral responses to genetic testing, particularly in the era of evolving evidence and risk information.

Methods

Overview

We used a 5-step process to develop the MyCancerGene digital tool (Table 1). First, we describe initial patient and genetic provider feedback and recommendations regarding the concept of MyCancerGene. Next, we describe user and usability testing, which informed the final content and functionality of MyCancerGene.

Table 1. Summary of the 5 steps, methods, and outputs involved in the development of MyCancerGene. Each step included feedback and input from both patients and genetic providers (N=395).

Steps	Methods	Outputs
1. Inquire	Surveyed patients (n=307) who have completed genetic testing in the multicenter COGENT[a] study	• Evaluated participants' interest in and barriers to a digital health portal • Explored GHP[b] advantages, disadvantages, and usefulness and potential content
2. Determine	Individual qualitative interviews with patients (n=10) and a genetic provider focus group (n=17) to better understand key intervention components and functions	• Informed by the diffusion of innovation theory, evaluating key attributes (eg, relative advantage, risk compatibility, and complexity) • Evaluated patient and provider preferences for content and functionality, including comfort with updates in test results
3. Develop	Incorporate feedback into MyCancerGene	• GHP (MyCancerGene) informed by formative interviews • Key components from patient and provider input • Initial screenshots were developed for user testing
4. User testing and refinement of the intervention	Provider (n=25) and patient (n=28) feedback on the specific drafted content and functionality	• Feedback on the purpose, content, and comfort with specific functions • Recommendations for changes
5. Usability testing and final modifications to the intervention	Patient feedback (n=8) on the initial digital version of the intervention	• Feedback on content, presentation, and functionality of the initial digital intervention

[a]COGENT: Communication of Genetic Test Results by Telephone.

[b]GHP: genetic health portal.

Step 1: Evaluating Patient Interest in a Digital GHP

To better understand interest in and willingness to use a digital GHP among patients who had undergone clinical cancer genetic testing, we surveyed patients in the National Institutes of Health–funded Communication of Genetic Test Results by Telephone (COGENT) study. The COGENT study was a multicenter, noninferiority randomized study of telephone disclosure compared to in-person disclosure of cancer genetic test results, including MGPT [29,63]. All participants in the COGENT study were English-speaking adults who had completed in-person pretest counseling with a genetic counselor and were proceeding with clinical cancer genetic testing. As participants in the study, they completed surveys before and after the disclosure of results. This was an ideal clinically diverse population to obtain patient feedback on the value, advantages,

XSL•FO
RenderX

disadvantages, and content for the development of a future digital GHP (eg, MyCancerGene).

To obtain patient feedback, we develop closed- and open-ended items that were added to the COGENT postdisclosure surveys in 2014. These included 7 close-ended questions about the interest in and barriers to a GHP, as well as 9 open-ended items asking how a GHP could be helpful, its advantages and disadvantages, and what types of information and functions would be most useful. Framework analysis was used to examine open-ended responses [64-66]. Investigators reviewed responses for a subsample (58/307, 18.9%) of participants and developed a thematic framework of primary and secondary themes for each open-ended item. Next, 2 investigators (PT and AB) independently assigned thematic codes to the open-ended responses and discussed differences to refine the thematic framework. The thematic framework was then applied to the remaining samples' open-ended responses, and themes were refined to include new ones as they emerged. Differences in code assignments were resolved through discussion and establishing agreement for all responses.

Step 2: Qualitative Patient and Genetic Provider Inquiry to Inform the MyCancerGene Intervention

To better inform specific content and functions of our initial GHP, we conducted additional qualitative interviews with patients and genetic providers. Interviews were guided by the Diffusion of Innovation Theory and key attributes of successful innovations (eg, advantage, risk, compatibility, and complexity) were evaluated [67-70]. Genetic providers included 13 genetic counselors and 4 physicians with expertise in cancer genetics (n=17) at the 5 COGENT sites, who participated in a provider focus group in 2014. All providers were female and had practiced in cancer genetics for many years. Providers were asked open-ended items evaluating the perceived usefulness of a GHP for patients and providers, perceived provider challenges, and comfort with different types of updates (eg, reminders, general testing updates, upgraded and downgraded VUS results, and patient family history updates). Open-ended responses were independently coded as described earlier by 2 authors (PT and KS) and resolved through discussion with a third coder (BM).

Patients included 10 purposefully selected COGENT participants across sites to represent the view of sociodemographically diverse patients who had undergone genetic testing. Patients completed individual interviews, including open-ended questions in 2014. These included (1) evaluating internet and health portal use in general, (2) advantages and disadvantages to health portals, (3) advantage and disadvantages to a GHP, (4) content that would be useful and not useful, (5) perceived usefulness of updates in genetic information and updates in family history, and (6) other suggestions for the GHP content and function. Similarly, the open-ended responses as described earlier were independently coded by 2 authors (PT and KS) and resolved through discussion with a third coder (BM).

Step 3: Developing the MyCancerGene Intervention

Key components and functions of our GHP (MyCancerGene) were informed by our formative interviews of the patients and genetic providers. Initial screenshots were developed for patient and provider feedback. Genetic counselors involved in content development reflected on their clinical experiences, providing clinical counseling to patients of diverse backgrounds during intervention content development. Similarly, researchers with expertise in health disparities and behavioral science were attentive to development for a clinically diverse population. Strategies included using plain language, definitions, short sentences, and more inclusive images.

Step 4: User Testing and Refinement of the Intervention

Initial screenshots for MyCancerGene content were developed based on the formative interviews of the patients and genetic providers, and individual user–testing feedback on these screenshots was obtained from both patients and providers in this step.

We obtained feedback from patients at 2 time points. The first group of patients providing feedback on user testing included 8 participants in 2014, and the second included 20 participants in 2019. In both sets of interviews, we purposefully selected for sociodemographically diverse participants. They viewed current versions of screenshots and were asked if the content was useful, what they expected to see, and why or why not for each question. The screenshots were updated in 2019 based on the feedback from 2014, although the questions were the same. In addition, they were asked for additional recommendations to improve the content. Participant responses were transcribed and summarized to inform modifications for subsequent interviews and development.

We conducted individual user–testing interviews with 25 genetic counselors from 2018 to 2019. Interviews included questions about the use of patient portals in general (ie, 7 items); feedback on the purpose, content, and comfort with specific functions of MyCancerGene (ie, 7 items); and feedback on individual screenshots and draft messaging for updates (ie, 14 items). Example questions are included in the associated tables. Participant responses were transcribed and summarized independently by 2 research staff to identify potential modifications. Feedback from the patients and providers were incorporated into the initial digital version of the MyCancerGene intervention.

Step 5: Usability Testing and Final Modifications to the Intervention

Using the initial digital version of MyCancerGene and cobrowsing software, we conducted individual usability interviews with 8 purposefully selected patients to obtain feedback on the functionality, navigation, experience, and any additional comments on the content from 2020 to 2021. Participants were asked what they liked and disliked on each page, if the information was understandable, and how it could be clarified if not. As applicable, they were asked about fonts, colors and images, how easy it was to navigate the functions, and how the content or functions could be improved. Participant responses were transcribed and summarized independently by 2 research staff to identify potential modifications.

XSL•FO
RenderX

Ethical Considerations

This study protocol was approved by the institutional review board (832,628), as well as the Cancer Center's Clinical Trials Scientific Review and Monitory Committee at the University of Pennsylvania and was determined to pose minimal risks to participants All participants provided a signed or verbal (ie, depending on step) informed consent. All data were deidentified before analysis. Participants were compensated in a range from US $0 to US $25, depending upon stakeholder group and step. None of the participants are identifiable in any of the findings reported in this manuscript.

Results

Step 1: Evaluating Patient Interest in a Digital GHP

In total, 307 COGENT participants completed the self-administered GHP patient items as presented in Table 2. Most patients reported a GHP would be useful and that they would use it as presented in Table 3. Those with a positive result were significantly more likely to report that a GHP would be helpful (odds ratio [OR] 10.9, 95% CI 2.2-54.0, *P*=.003) and that they would use a GHP (compared to true negatives: OR 7.2, 95% CI 2.1-25.0, *P*=.002). Nonetheless, over 75% (25/33) of those with a VUS or uninformative negative results reported a GHP would be helpful. Another factor associated with likely GHP use was history of prior health care portal use (OR 2.5, 95% CI 1.4-4.4, *P*=.001). Age, gender, education, cancer history, and baseline cognitive and affective measures (eg, knowledge and anxiety) were not substantially associated with interest in a GHP or the likelihood of using it.

XSL•FO
RenderX

Table 2. Characteristics of patients from the COGENT[a] study who completed genetic testing and were asked about their interest in a potential genetic health portal (N=307).

Characteristics	Values
Age (y), mean (SD)	47.60 (12.86)
Sex, n (%)	
Female	286 (93.2)
Male	21 (6.8)
Education, n (%)	
Some high school or less	1 (0.3)
High school	15 (4.9)
Associate degree, some college, or trade school	79 (25.7)
College	107 (34.9)
Post college	104 (33.9)
Refused to answer	1 (0.3)
Race, n (%)	
Asian	5 (1.6)
Black or African American	16 (5.2)
Mixed	4 (1.3)
White	282 (91.9)
Ethnicity, n (%)	
Hispanic or Latino	4 (1.3)
Non-Hispanic or Latino	303 (98.7)
Married, n (%)	226 (73.6)
Disclosure method, n (%)	
In person[b]	173 (56.4)
Telephone	134 (43.6)
Site, n (%)	
University of Pennsylvania	105 (34.2)
Fox Chase Cancer Center	110 (35.8)
University of Chicago	50 (16.3)
Stroger Hospital at Cook County	7 (2.3)
MD Anderson Cancer Center at Cooper	35 (11.4)
Known mutation in the family, n (%)	74 (24.1)
History of cancer, n (%)	153 (49.8)
Result, n (%)	
Positive	52 (16.9)
True negative	32 (10.4)
Uninformative negative	189 (61.6)
Variant of uncertain significance	34 (11.1)

[a]COGENT: Communication of Genetic Test Results by Telephone.

[b]In total, 53 patients were in the self-selected in-person group.

XSL•FO
RenderX

Table 3. Patient interest in a genetic health portal by genetic test results among patients who had completed genetic testing in the COGENT[a] study (N=307).

	All, n (%)	Positive (n=50-52)[b], n (%)	Uninformative negative (n=179-188)[b], n (%)	True negative (n=30-32)[b], n (%)	VUS[c] (n=32-34)[b], n (%)
Would a genetic health portal be helpful?[d] (n=303)	249 (82.2)	48 (96)	154 (81.9)	22 (68.8)	25 (75.8)
Any barriers to using a genetic health portal?[e] (n=291)	60 (20.6)	13 (26)	36 (20.1)	6 (20)	5 (15.6)
Would you use a genetic health portal?[f] (n=305)	227[g] (74.4)	48[g] (92.3)	135[g] (72.2)	20[g] (62.5)	24[g] (70.6)

[a]COGENT: Communication of Genetic Test Results by Telephone.

[b]A range is included as some participants did not answer all items.

[c]VUS: variant of uncertain significance.

[d]"Now that you have received your results, do you think it would be helpful to have access to a secure, password protected electronic Genetic Health Portal (similar to other health portals you might have used in the past) with the information you received from your genetic provider (genetic counselor, physician, nurse practitioner, physician assistant)? (Yes/No)."

[e]"Do you think there would be any challenges with and/or barriers to using a Genetic Health Portal?"

[f]"How likely is it that you would use a Genetic Health Portal?"

[g]Indicates somewhat or very likely on a 5-point Likert scale.

Table 4 presents patient feedback on why a GHP might be considered helpful, based on responses from 224 participants. The reasons were categorized into themes, highlighting aspects such as accessibility, organization, and the emotional benefits of having secure, easy access to genetic information. Conversely, some patients expressed concerns about the need for more human interaction, technology challenges, and potential privacy issues, which could make the portal less helpful for them. In response to open-ended questions presented in Table 4, the most frequently reported advantages of the GHP were increasing accessibility, convenience, and efficiency (103/224, 46%), keeping genetic information organized (54/224, 24.1%), and to increase or maintain patient understanding of the information (38/224, 17%). Patients also reported it could be helpful for sharing genetic information with others, downloading and printing documents, security, and reducing anxiety.

Table 5 summarizes the perceived disadvantages of using a GHP, based on feedback from 206 participants. The most common concerns include issues related to privacy and data security, potential emotional distress, and difficulties with technology, such as recalling log-in information or dealing with technical glitches. Some participants also noted that they might not use the portal frequently or prefer human interaction over digital communication. A subset of patients (35/224, 15.6%) reported that a GHP would not be helpful (Table 4), citing the lack of human interaction, potential access challenges and concerns regarding privacy, and the potential to cause anxiety. When asked specifically about potential disadvantages to a GHP, 36.4% (75/206) of participants reported no disadvantages and 32.5% (67/206) reported concerns about privacy. Other disadvantages reported at lower frequencies included that information could be distressing, trouble remembering log-in information, technical challenges, and the lack of human interaction as shown in Table 5.

Table 6 outlines the types of information that 227 participants indicated would be most useful in a GHP. The most frequently reported components or functions included access to results (112/227, 49.3%), medical recommendations (47/227, 20.7%), and an easy-to-understand review of results (29/227, 12.8%). Other suggestions included access to genetic providers and additional resources, appointment details, screening results, and family history (Table 6). Some participants also endorsed the value of a summary page, updates in information, and the ability to print and share materials with relatives or providers.

XSL•FO
RenderX

Table 4. Reasons a genetic health portal would be helpful[a] (N=224).

Themes	Examples	Participants, n (%)
Reasons a genetic health portal would be helpful		
Accessibility (85 participants mentioned convenience or accessibility, including access at any time; 18 participants mentioned efficiency, ie, being fast or quick)	• "So that the information is readily accessible." • "Again, it's just more convenient to have access to information whenever I choose." • "Because I would be able to refer back to the results. It would also be good to be able to see anything that I might have missed."	103 (46)
Organization	• "Paper records can be misplaced and this is a way to access the information." • "I would like to be able to electronically see my results so I don't have to keep track of the printout." • "Having access to electronic records alleviates having to keep track of paper records."	54 (24.1)
Maintaining understanding	• "A Genetic Health Portal would be a central information resource for information related to my specific results and general genetic information. A portal could provide a central source for information and resources." • "Access to information is always helpful, especially when details are important and the information is sensitive."	38 (17)
Ability to share results with others (6 participants specifically referred to the ability to download and print results)	• No examples	26 (11.6)
Helpful (not otherwise specified)	• No examples	13 (5.8)
For security (benefit of keeping results secure)	• "No one else should have access to my genetic information." • "[This] would be a safe way to communicate with genetic provider."	7 (3.1)
Emotional benefit	• "I feel like it would empower me more to be in control of my own health information for my future."	7 (3.1)
Ability to update information	• No examples	1 (0.4)
Reasons a genetic health portal may not be helpful		
Lacks human interaction	• "I think it would be worrisome to see this information without the assistance of someone knowledgeable explaining it to me... The genetic information is so sensitive and in some cases, unclear as to its significance regarding the individual and family members and really requires assistance to understand it."	14 (6.3)
Not helpful if you lack access to or comfort with technology	• "However, those that are not internet savvy will have difficulties." • "I haven't used (or care to use) health portals because, quite frankly, creating accounts/passwords for so many varied things has become too frustrating for me personally)."	9 (4)
No benefit or advantage	• "I prefer meeting in person so I can receive a thorough explanation and have the opportunity to ask questions."	7 (3.1)
Concern for security breach or privacy	• "Might be nice. Concerned about security of personal information."	7 (3.1)
Could cause anxiety	• "But I always worry about finding bad news this way [through GHP] and then having to wait to see doctor."	5 (2.2)
Not helpful because not combined with other portals	• No examples	1 (0.4)

[a]"Now that you have received your results, do you think it would be helpful to have access to a secure, password protected electronic Genetic Health Portal (similar to other health portals you might have used in the past) with the information you received from your genetic provider (genetic counselor, physician, nurse practitioner, physician assistant)? Why or Why not [coded responses]."

XSL•FO
RenderX

Table 5. Disadvantages of a genetic health portal[a] (N=206).

Theme	Examples	Participants, n (%)
None	• No examples	75 (36.4)
Not secure enough and concerns regarding privacy (risk of breach)	• "Data could be hacked and used against for insurance or employment purposes." • "The only disadvantage would be if the site was breached." • "I suppose digital security is always a concern."	67 (32.5)
Upsetting or distressing	• "Genetic testing that could prove to be more anxiety-provoking than beneficial." • "Finding out some life changing news without the comfort or clarification from a person." • "For some, upsetting way to get bad news, perhaps misunderstanding information."	20 (9.7)
Difficulty recalling log-in information or too many health care sites	• "One more password to remember/forget."	12 (5.8)
Technical challenges (6 participants mentioned that it would be hard for those who lack technical skills, 2 mentioned that some may lack access to technology, and 2 mentioned glitches, website malfunctioning, and maintenance)	• "It would not be so beneficial for people who do not have access to online portals or do not know how to use the system." • "If I didn't have a computer and had no access to a computer." • "Any possible glitches with the system."	10 (4.9)
May not use (too much time or no benefit)	• "Not referred to often enough." • "The time to sit down and look up/use the portal."	7 (3.4)
Lack of human interaction	• "Less personal than talking to someone."	5 (2.4)

[a]"What do you feel would be the disadvantages to having access to the Genetic Health Portal?"

Table 6. Types of information desired for a genetic health portal[a] (N=227).

Theme	Participants, n (%)
Genetic test results	112 (49.3)
Educational or informational resources (26 participants mentioned resources and education; 20 statistics, cancer risks, or other numerical information; 19 current research or updates in research; and 2 the results of this research)	62 (27.3)
Medical recommendations	47 (20.7)
Easy-to-understand explanation of results	29 (12.8)
Genetic provider information and notes (14 participants mentioned access to clinicians or contact information, 7 provider notes, and 3 medication details)	24 (10.6)
Appointment information (18 participants mentioned appointment details, such as dates seen, who did I see, or location of visit; 6 upcoming appointments or reminders for upcoming appointments)	22 (9.7)
Everything	18 (7.9)
Results of screening or procedures	16 (7.1)
Family history	14 (6.2)
Genetic test description	14 (6.2)
Summary (quick overview or snapshot)	10 (4.4)
Updates to recommendations, new information, or research	9 (4)
Print and share function (to share with relatives or providers)	6 (2.6)
Ways to share or engage with others (other patient experience or patient forum)	3 (1.3)
Recommendations for relatives	1 (0.4)
Billing or insurance	1 (0.4)

[a]"What type of information or documents do you think would be most useful to include in the Genetic Health Portal?"

XSL•FO
RenderX

Step 2: Qualitative Patient and Genetic Provider Inquiry to Inform the MyCancerGene Intervention

Qualitative Genetic Provider Focus Group

The primary potential benefits of a GHP, as reported by genetic providers, included mechanisms to improve patients' sharing of accurate information with relatives and providers and for providers to update patients with new risk or test information. Most providers felt a GHP would be useful to patients as they already expect providers to update them (ie, "call them") with new information, although providers admitted this is not always feasible in practice and over time. Many endorsed a GHP as a way for patients to maintain their genetic records electronically in a specific location instead of in paper files. At the same time, providers indicated that many patients already have electronic access to these records through the electronic health record (EHR) and asked how the GHP would be different from the access already provided.

Providers acknowledged that they were not sure that a GHP would alleviate their day-to-day challenges as they already communicate with some patients via the EHR. Other concerns with a GHP included ensuring that patients receive and understand updates shared through the GHP, and that many patients do not log on or use ancillary platforms. Genetic providers asked if there could be an alert to providers if their patient logged on and reminders to patients when new information became available. Other concerns were regarding the burden on providers to update a GHP, maintaining current information, and the potential to create additional work for providers.

Genetic providers had different levels of comfort with updates provided through a GHP. They all felt that a GHP could be a good method to alert patients to update their family history (eg, suggested once or twice a year). They were less comfortable with updates to screening recommendations as these could change overtime and could conflict with other information or recommendations patients' were receiving. In general, most providers were comfortable with new general information about testing (eg, new testing available) and downgraded VUS reclassifications. At the time, VUS downgrades were communicated via letter and a GHP could provide a good alternative to mail communication, although providers raised concerns that some patients might not log in. Most providers were less comfortable with upgraded VUS reclassifications being shared through a GHP. They felt these needed to be communicated by a genetic counselor and felt a message (eg, "please call your genetic provider") was a more appropriate method for notification through a GHP.

Qualitative Patient Interviews

Patients (n=10) were aged 29 to 69 years and included 2 men, 2 patients with less than a college degree, 1 Black patient, and patients with a range of test results (ie, positive, VUS, negative, and results pending). Most participants (9/10, 90%) thought that a GHP could be useful in referencing immediate results, concise information, and accessing reports and documents. Suggestions from participants included the addition of content and documentation (ie, results, recommendation letter, and family history); information about individual risk compared with the general population; ability to track screening and medical management; and updates in the field. Most participants (8/10, 80%) supported receiving updated test results through a GHP. Participants also felt that a GHP could help support communication with relatives. They also suggested that a GHP could be more useful if it provided tailored educational resources, more billing and security information, a medical history summary, opportunities to connect with other patients, and if it were designed to be user friendly.

Step 3: Developing the MyCancerGene Intervention

Patient-Facing Content

Components of our GHP (ie MyCancerGene) were informed by our patient and genetic provider interviews as detailed in Table 7. Components included (1) the Log-in and Landing Page (Figures 1 and 2), (2) Summary of Care page (Figure 3), (3) My Genetic Test Results page (Figures 4 and 5), (4) My Family History page (Figure 6), (5) Provide an Update page, (6) Review Updates page, (7) Resources page, and (8) Screenings Tracker.

XSL•FO
RenderX

Table 7. Components of the MyCancerGene genetic health portal[a].

Component	Description	Patient endorsement	Genetic provider endorsement
Landing Page	Includes 8 icons for intervention components and a list of learning links in the left side bar. There is a summary statement regarding any new updates and a reminder to provide an update if there is new family information.	• "First page shouldn't include any medical information. Users should be able to click and navigate to what they want to see and when."	__[b]
Summary of Care	Includes location and date of service, testing laboratory, date and provider, test result statement (eg, positive for a BRCA1 mutation) and a link to the result page	• "Would make things easier to recall when it is organized like this."	—
My Genetic Test Results	Includes type of test, test result statement, and a PDF version to the test report. There is a patient-centered simple explanation or summary of the results, a table of lifetime risks associated, and general medical recommendations for positive results	• "Would be likely to refer to this when sharing results, especially by phone." • "It's like having an electronic file cabinet!"	"Could be an easier way for patients to share results with family and other providers."
My Family History	Includes family history obtained at the medical visit and updated by the patient in the portal. There is a link to the PDF document of the last provider-generated pedigree and a link to update family history	• "Super useful section. New doctor always asks for this." • "Big improvement than when [patient] had to fill it out by hand."	"Some patients might not recall which relatives' history they have or haven't shared. Being able to view to verify before they reach out would be helpful."
Provide an Update	Includes icons to provide a family update (family history or genetic testing in relatives), personal update (medical history of genetic testing), and other updates (contact information and other information)	• "Feels there is a higher chance of 'success' in conveying updates or contact this way over a "standard" telephone approach. With a portal, there is a way to track who has been told what and when. With the phone/mail, who can say?"	"Many patients are under the impression we will call them when new testing or other information becomes available. This is often not feasible."
Review Updates	Provides a chronological list of updates with dates, provider involved, and summary of the update	• "[I] like this, but would probably be more inclined to update as needed when reminded."	"Currently, certain updates are handled by letter. If a verification that update has been received is available, some updates may be easier this way."
Resources	A list of 8 links for organizations that provide information regarding emotional and additional education resources. There is also a tool bar on the home screen with "Learning Links" including 3 videos and 3 text screens.	• "[Would want] a section for recommendation links and resources"	—
Screenings Tracker	Includes the ability to enter the details of upcoming screening appointments including date, description, and comments. These are self-entered and self-monitored.	• "Medical management recommendations for the patient specifically would be ideal...perhaps a way to track what screening/med. management had been done –a timeline."	—

[a]There is a home icon on all component screens to return to the home screen. There is a "Have a question?" icon on every page.

[b]Not applicable.

XSL•FO
RenderX

Figure 1. Log-in page.

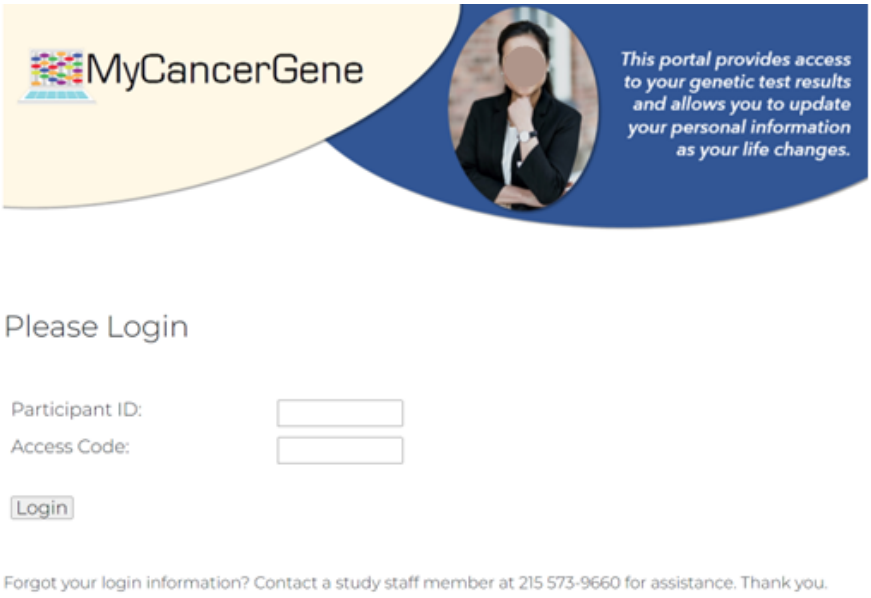

Figure 2. Landing page.

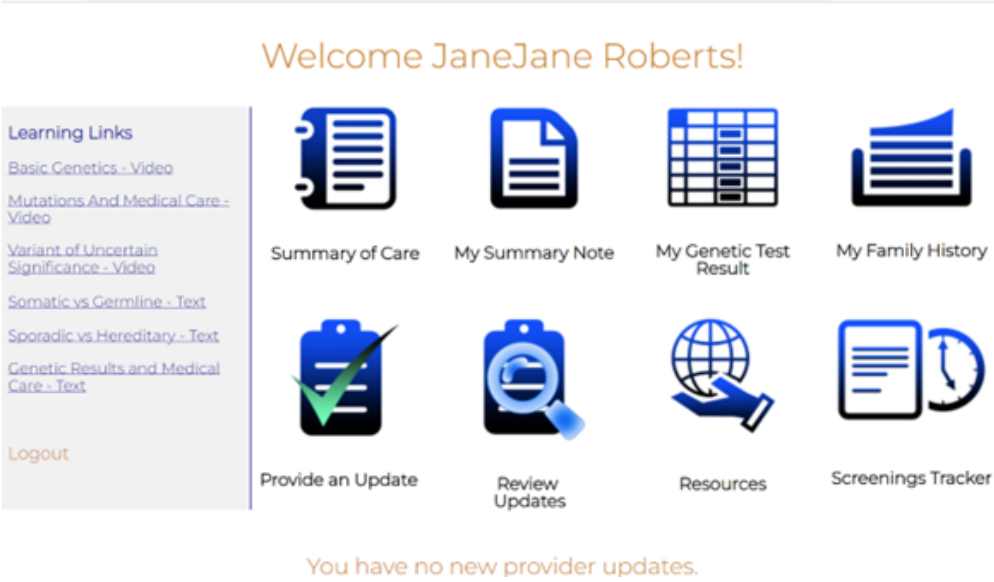

XSL•FO
RenderX

Figure 3. Summary of Care page.

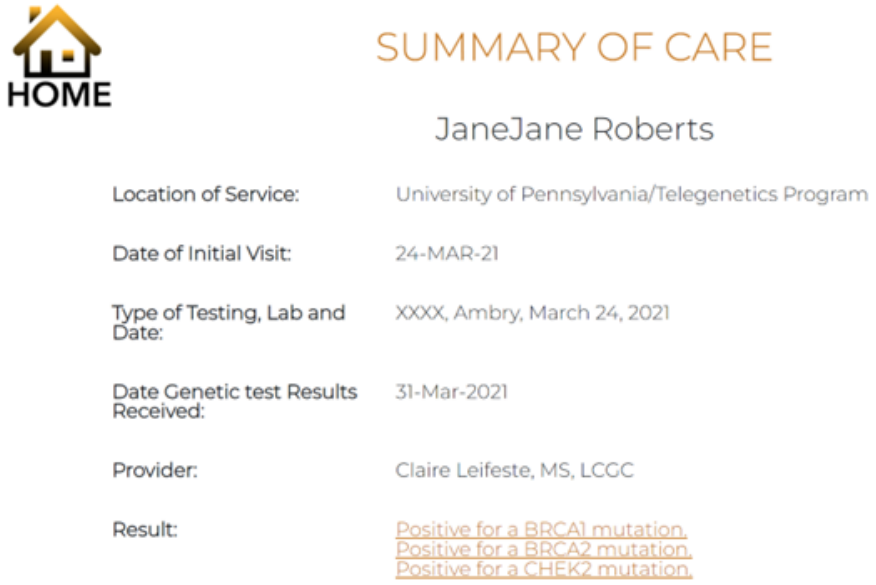

Figure 4. My Genetic Test Results page.

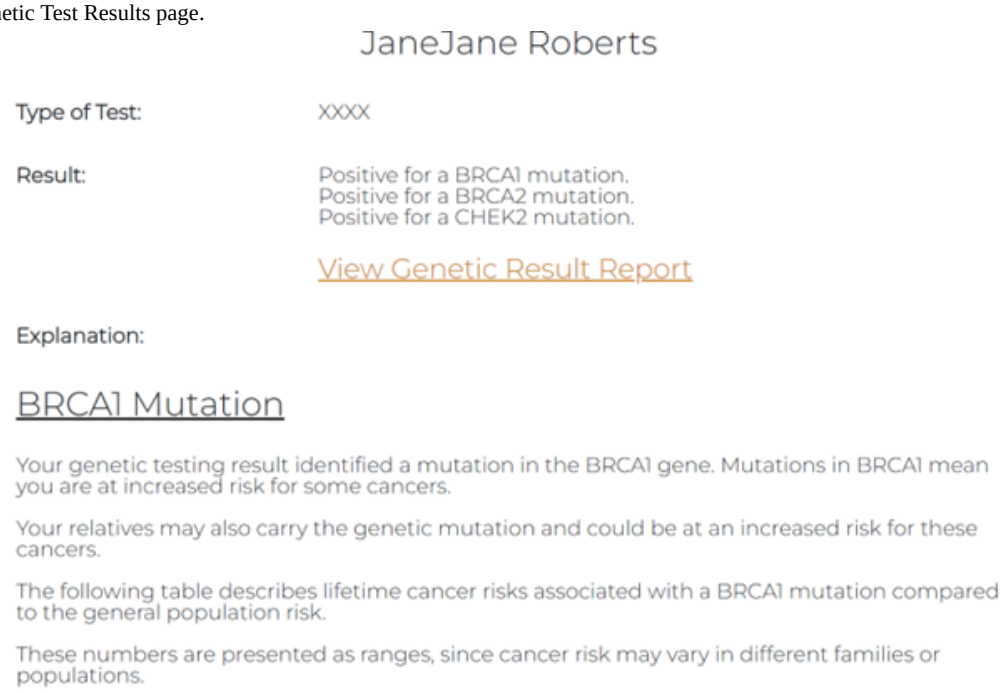

XSL•FO
RenderX

Figure 5. My Genetic Test Results page (continued).

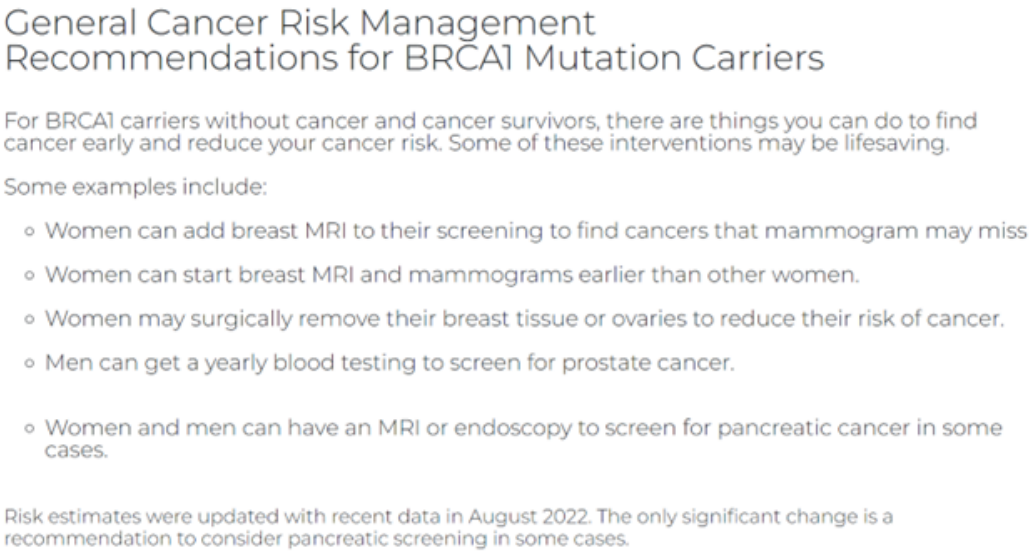

Figure 6. My Family History page.

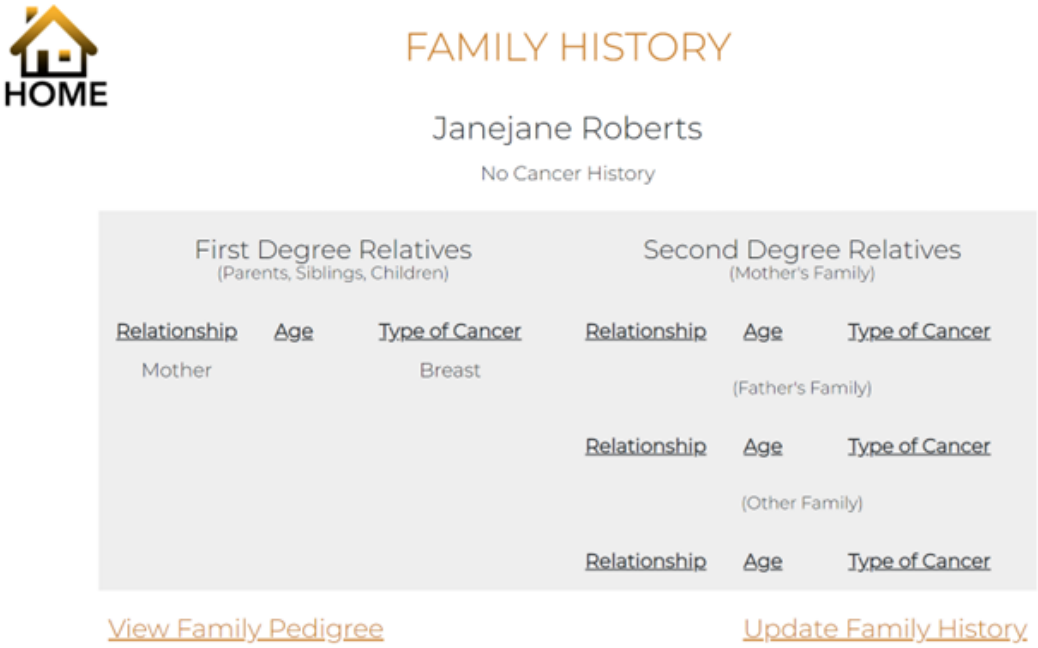

Engagement Strategies

Overview

As endorsed by patients, reminders and engagement functions were designed to remind patients about the portal (Table 8). These include initial activation reminders to encourage patients to activate their MyCancerGene portal. We also designed general reminders sent every 6 months to remind patients to update their family or personal history and any new testing in the family. We also designed educational engagement reminders, which are short educational messages regarding cancer genetics or family risk and are sent every 6 months, alternating with general reminders (Table 8).

XSL•FO
RenderX

Table 8. MyCancerGene reminders.

Type	Frequency	Timing	Examples
Initiation reminder	4 reminders	On days 5, 9, 15, and 30	__[a]
General reminders (to update personal and family history and new testing in the family)	No end date; every 6 months	At 6, 12, and 18 months	• Have you updated your family history: cancer status? • Have you updated your family history: testing status?
Engagement reminders (including educational messages and reminders about genetic and familial risk)	No end date; every 6 months	At 3, 9, and 15 months	• Have you shared your results with your relatives? • Have you had your recommended screening?

[a]Not applicable.

Patient and Genetic Provider Updates

As outlined earlier, one of the main purposes of the MyCancerGene intervention is to provide a mechanism for patients to provide updates of their personal and family history to the genetic providers and for providers to send updates to the patients.

Patient-Driven Updates

MyCancerGene was designed so that patients can provide updates to their personal and family history or new testing in the family at any time on the Provide an Update page. They may also provide an update in response to general engagement messages sent every 6 months. All updates were designed to be reviewed by their genetic provider to determine if their update prompts a change in medical management or new genetic testing in them or their relatives. Given the integration challenges and concerns about burden on genetic providers, the research staff reviewed updates, drafted responses when appropriate, and obtained a final approved response from the genetic provider. Genetic provider responses to updates were then sent to patients through the MyCancerGene portal. It is intended that this could be better automated in the future after efficacy is established.

Genetic Provider–Driven Updates

MyCancerGene was designed for 2 specific genetic provider updates. First, to provide updates in VUS classifications (eg, VUS reclassification to benign or VUS reclassification to pathogenic or likely pathogenic). When genetic providers received a laboratory update on a patient enrolled in the MyCancerGene portal, they contacted the research team with the update. The genetic provider and research team created updated content for the Summary of Care, My Summary Note, and My Genetic Test Results pages. These changes were then programmed by the Clinical Research Computing Unit and, when completed, would generate a message to be sent to the patient. For example, “Your Variant of Uncertain Significance Result has been reclassified. To view the details of your reclassification now, please log onto your MyCancerGene portal.” This message provided log-in information and the option to speak with a provider to receive the updated information. For example, “If you would rather speak with your Genetic Counselor to discuss this update, please contact a member of the research study team who will help you set up a telephone call.”

In addition to patient-specific updates to results, the MyCancerGene intervention was designed to permit other updates based on changes in the field that might apply to a group of individuals. Examples include new screening recommendations for patients with a specific genetic mutation (eg, new screening recommendations for *PALB2* carriers), changes to risk estimates for particular genes, or new genetic testing options for patients with specific personal or family history. Thus, MyCancerGene was designed to allow edits to all patient-facing content to allow for these updates and a mechanism to message patients that updates had been made. In practice, updates were classified as “informational,” with a message notifying patients that there were changes made to their genetic information but that it likely did not change their current care. In contrast, other updates were classified as “potentially actionable,” and messages suggested that they access MyCancerGene to review the updates and contact their genetic provider if they had questions.

Genetic Provider Interface and Functionality

At the time of development, there was not an easy pathway for this intervention (ie, MyCancerGene) to integrate with the existing EHR. Furthermore, at the time of development, the use of MyPennMedicine, the institutional patient portal, was relatively limited and for messaging only. Given that the intervention could not be integrated with the EHR at the time of development, and genetic provider concerns about the burden of operating in 2 separate systems, we elected to not develop a provider-facing functionality. Rather, case report forms for research staff to enter messages and submit content to programmers were used. We planned that if the intervention had proven efficacy (eg, after a randomized trial), an integrated provider-facing interface would be developed. For the randomized trial, we planned that research staff would act as facilitators to populate the genetic information in MyCancerGene (eg, testing performed, copies of reports, and letters). In addition, to reduce human error, a second verification step was included to ensure accuracy (eg, a study genetic counselor or second research staff member verifies that there is accurate information in MyCancerGene before participant access). It was expected that these steps would be reduced over time as systems permitted greater integration or at the time of wider dissemination, after efficacy of the intervention had been proven.

XSL•FO
RenderX

Programming and Technology Specifications

The MyCancerGene participant portal is based on Oracle database tools that reference tables, functions, and procedures in an Oracle database. The participant is provided with a URL link that they enter into a standard web browser. The link invokes an Oracle REST Services call that connects to the appropriate database and schema to call a stored procedure that implements the log-in process. The participant is provided credentials by the MyCancerGene staff to complete the log-in. The log-in procedure passes the credentials to a registration procedure that checks the credentials and creates an active session if the log-in is valid. The active session includes an expiration time for security purposes. Once a session is established, a page-rendering procedure is invoked to read content from database tables and dynamically generate HTML code for display on the browser. All content for the portal is maintained in tables of the database. Study research staff use a separate application that was built using Oracle ApEX to enter participant specific information into Oracle tables that is used by the portal procedures to build the web screens. The page-rendering procedure handles page navigation as well as screen display. A log-out option on the main portal screen is used to terminate the session.

Step 4: User Testing and Refinement of the Intervention

Patient User Testing

Participants for the first 8 user-testing interviews were from 4 of the COGENT sites. They were all female, were aged 39 to 57 years, included 1 Hispanic and 1 Black patient, and included 3 participants with less than a college degree and patients with a range of results (ie, 5 negative, 1 positive, and 2 VUS results). Most participants reported the content to be both what they expected and useful. Key feedback from initial interviews included adding a landing page to be seen before the summary of care, reducing text, and suggestions for descriptive text, colors, and fonts. Participants were evenly divided on presenting family history information as a list versus as a pedigree. They also reported that general reminders and notifications when MyCancerGene information was updated would be important to increasing use of the intervention. Other recommendations included the addition of educational resources, details on care providers, and dates of visits and other events.

Participants for the second set of patient user–testing interviews (n=20) were identified from recent clinical encounters. They were aged 29 to 72 years, including 5 men and 1 participant with less than a college degree and with a range of results (ie, 3 negative, 10 positive, and 7 VUS results). Participants had limited feedback for most content, with most suggestions for the Landing Page, My Family History page, messages about updates, and general suggestions to increase use. Suggestions for the Landing Page included clarifying the organizational affiliation and changes to design, fonts, and images. Suggestions on the Family History page included the option to enter relatives and their test results, inclusion of third and extended generations, and the option to view a pedigree. Other general suggestions included replace pie charts with numerical risks, include a glossary and define terms, clarify the type of update and a link to the portal in messages, add educational resources, make MyCancerGene accessible on mobile devices, integrate MyCancerGene into the existing medical record and patient portal, and provide reminders about availability of MyCancerGene. Some participants expressed concerns about privacy and security and provided recommendations to reassure them (eg, use a secure site and authentication strategies).

Participants in the second set of interviews also reviewed several options for messaging for updates to their test results. In contrast to genetic provider opinions as described subsequently, they strongly preferred the option to receive a message and directly access their updates in MyCancerGene, with the option to speak with a genetic provider of their choice, as opposed to options which alerted them of an update but then required that they speak with a genetic provider first before having access to their update (ie, provider disclosed).

Genetic Provider User Testing

Of the 25 genetic counselors who completed user-testing interviews, 88% (22/25) were female, 32% (8/25) were affiliated with the University of Pennsylvania, and 68% (17/25) were from external practices from 5 states (Pennsylvania, New Jersey, Texas, California, and Florida). Genetic providers reported that the benefits to patients included providing a centralized area for information and a tool to communicate updates (Table 9). Providers also identified benefits for themselves, including optimizing communication and saving their time. They also reported potential challenges, including creating extra work and potentially leading to inefficiencies (Table 9).

Genetic providers also had recommendations for changes to the Summary of Care page, Landing Page, My Family History page, and VUS reclassification page, which included simplifying text, changes to graphics and pictures, adding screening recommendations or care plans, changes to identify relatives on the family history page, and accounting for multiple VUS results. Providers reported variable comfort with different types of updates being provided through MyCancerGene (Table 10). They expressed high comfort with updates to general testing information, reminders, and downgrades to VUS results. However providers varied in their comfort with upgraded VUS results being provided through MyCancerGene. Some were comfortable providing patients with the option to directly access their updates in the portal, with the option to speak with a provider instead if they preferred. Others felt that changes in VUS classification (ie, particularly upgrades) could be sent as a message, which would alert patients of the update, but they felt that patients should be required to speak with a provider first before accessing their update in MyCancerGene.

XSL•FO
RenderX

Table 9. Key findings from genetic provider stakeholder interviews and user testing (n=25).

Questions	Most common themes: n (%)	Examples
When thinking about your patients, how do you think this patient portal (MyCancerGene) would benefit them?	• Centralized information: 13 (59) • Clearer updates: 8 (36)	• "All documentation would be easy to locate. Family history and VUSa updates which are typically cumbersome would be streamlined."
What types of challenges could MyCancerGene alleviate for genetic counselors?	• Centralized information: 8 (33) • More efficient communication: 8 (33) • Saves provider's time: 7 (29)	• "Could increase efficiency (less back and forth phone calls/sending letters that may or may not get to where they need to, etc...)" • "Could end up saving as much as 30 minutes with each interaction since, instead of having to scavenge through chart notes, all the pertinent points are highlighted for each patient."
What types of challenges could MyCancerGene create for genetic counselors?	• Extra work: 15 (63) • Inefficient communication: 5 (20.8)	• "More work (double the work if not integrated with the EMRb). Also potential for more work if patients have an easier means of communication."

[a]VUS: variants of uncertain significance.

[b]EMR: electronic medical record.

Table 10. Genetic provider comfort with different types of provider updates.

Comfort with different types of updates	Most common themes: n (%)	Examples
General testing information	• Comfortable: 23 (92)	• "Comfortable. Thinks this is so much better than the 'call us to follow up to see if there are any updates.'"
Reminders	• Comfortable: 24 (96)	• "Comfortable - and patients would appreciate this."
Downgraded VUS[a]	• Comfortable: 22 (88)	• "Comfortable with confirmation."
Upgraded VUS	• Comfortable: 11 (44) • Not comfortable: 11 (44)	• "Comfortable as long as follow-up with GCb is not just suggested but strongly encouraged if not mandatory." • "Not comfortable; prefers to call patients directly."

[a]VUS: variants of uncertain significance.

[b]GC: genetic counselor.

Step 5: Usability Testing and Final Modifications to the Intervention

Participants for usability testing (n=8) were White women who were aged 24 to 70 years, with a range of education levels (ie, high school only to graduate level). Participants had a range of genetic test results (ie, 6 positive, 1 VUS, and 1 negative). Overall, many participants reported that icons and features were easy to understand, useful, and what they expected. Participants reported greatest interest in the Genetic Test Results and Family History pages and additional resources. While they reported the summary letter was useful to have, several mentioned that they may not use it. Some were not clear what it was, even though they should have received a copy after their genetic counseling disclosure visit. Most found the Summary of Care and Summary of Updates pages useful, and most felt the Review Updates page was easy to understand. There were several typos or spacing recommendations to make text easier to understand, and these were generally adopted. Recommendations for changes to fonts, pictures, or colors were made when recommendations were not conflicting or were mentioned by more than 1 participant. Key recommendations included making pages available in a printable format, summaries under the videos, a note that if something does not appear accurate (eg, personal or family history) to go to share an update, and the ability to provide more notes on the family history page (eg, dates of diagnosis). Another recommendation was to add the My Screenings page, which had been a previous recommendation that was not included due to the challenge of keeping it updated with the electronic medical record when MyCancerGene was not yet integrated. However, given the repeated patient feedback, this page was created in the final version so that participants could self-track screening and data. Other recommendations that were considered but were not possible at the time included providing access to relatives, providing access to genetic providers, synchronizing a screenings tracker with their calendar, making the intervention available within their existing portal, and creating a repository where patients could store articles and websites they found

XSL•FO
RenderX

personally useful. These were not included because they were either suggested by a single participant or were technically challenging at the time, but they were recorded as potential future considerations.

Discussion

Principal Findings

With the advent of MGPT in clinical genetic testing, an increasing number of patients are left at a risk of misunderstanding, uncertainty, and evolving interpretations and recommendations after receipt of cancer genetic test results [6,8,71]. Thus, longitudinal follow-up to update risk estimates and recommendations for positive results, update VUS results, and update family history based on changing circumstances are needed. Yet, cancer genetic testing is often a one-time encounter with a genetic counselor, as opposed to an ongoing relationship with continuity of care. This leaves the responsibility for follow-up, updates, and longitudinal care to patients, who can contact their genetic provider as needed, or to their other health care providers, who may not be equipped to provide genetic-specific updates. To address this clinically significant gap in cancer genetic care, we developed a patient-centered longitudinal digital GHP to support patients with longitudinal care after receipt of their cancer genetic test results. MyCancerGene was directly informed by the feedback from patients and genetic providers directing both the content and functionality of the intervention. The initial prototype was then refined through extensive user and usability testing with patients who had received a range of genetic test results.

Patient and genetic provider interviews identified high interest in a patient-centered longitudinal digital GHP to support longitudinal care. Over 90% of patients with positive results and 75% of participants with a VUS or uninformative negative results reported an interest in a patient-centered digital GHP. The primary advantages of this type of tool, according to patients, included increased accessibility, convenience, and efficiency of accessing their genetic test reports and other documentation from the genetic visits, keeping genetic information organized, and increasing and maintaining patient understanding through easy-to-understand materials and educational resources. Patients also felt that such a digital health tool could help with communication and sharing of materials with relatives or other health care providers. Patients were also in favor of receiving electronic updates through a patient-centered digital GHP and highlighted that the intervention would need to address privacy concerns and be easy to use. Similar to patients, genetic providers reported that a patient-centered digital GHP could help patients share information with relatives and other health care providers and help patients to update the genetics team about new health information (ie, personal or family history or new test results in the family). Patients and genetic providers also felt the portal may help provide a mechanism to update patients with new information about genetic results. While many providers endorsed this tool as a place to electronically store documents for patients, some noted that this could be redundant to what is already provided in the electronic health portal, and the added value was unclear.

A key component of this formative work was to determine, from the patient and genetic provider standpoint, what content and functions this patient-centered digital GHP (ie, MyCancerGene) should include. Patient and provider input identified 8 key components of the tool and most were endorsed by both patients and providers as useful for longitudinal care. These included the following pages: Landing Page, Summary of Care, My Genetic Test Results, My Family History, Provide an Update, Review an Update, Resources, and a Screenings Tracker. Patients and providers also addressed several key functions, including the ability to download and print materials and the inclusion of reminders and engagement functions. The iterative user and usability testing helped inform changes to increase ease of access by making the layout and design more intuitive, changes to content to increase understanding, the inclusion of a glossary with defined terms, the addition of educational resources, and changes to pictures and colors to help with patient understanding and overall experience. Integrating MyCancerGene into other health portals, providing access to relatives, and synchronizing screening trackers with personal calendars were unable to be incorporated at the time but were identified as future content or functions that could be helpful.

While most patients and genetic providers endorsed advantages and benefits of a patient-centered digital GHP, some identified potential challenges. Some participants had privacy or security concerns, and a few anticipated that reviewing past or updated information regarding genetic risk electronically could be upsetting or distressing. A few also commented that MyCancerGene would be yet another health portal that could create additional log-in credentials, which can be hard to recall. Patient and provider comfort levels also varied with the return of updated VUS results in MyCancerGene. While patients were comfortable with updates (ie, even VUS reclassification upgrades or clinically significant results), genetic providers had mixed feelings on the appropriateness of sharing upgraded VUS results through MyCancerGene. After sharing the overwhelming support among patients to make a choice for themselves if they are ready for updated information about their results, genetic providers were more open to this option, although some still had reservations. This will be an important outcome to evaluate in future longitudinal studies given the variable opinions from patients and genetic providers.

Another primary challenge identified by both patients and genetic providers is the strong desire to have such a genetic portal integrated with the existing EHR. Technological challenges in integrating the Oracle based system with the existing EHR and long queues and prioritization for already existing EHR modifications were barriers to developing MyCancerGene initially as an integrated component of the EHR. Furthermore, it was felt that establishing efficacy would be important to determining prioritization, value, and investment in future integration. Another concern was the availability of a variety of different EHR platforms available and institutions might differ in their choice of platform used, which could limit future implementation. For these reasons, the intervention was built as a proof of concept for a randomized trial with plans to

XSL•FO
RenderX

engage an IT integration committee throughout the trial to consider how MyCancerGene could be integrated with the EHR in the future. Of note, the genetic provider–facing component of MyCancerGene would need to be developed and tested at the time of future integration.

Limitations

Some limitations to our formative research and development are acknowledged. The hypothetical query may not have aligned with future use and benefits. Thus, a randomized trial of the intervention compared to usual care is planned and may better define real-world benefits and outcomes. As noted earlier, some desired components and functions could not be included, and this may limit the benefits, although these could be developed for future versions of MyCancerGene. In addition, while we attempted to purposefully identify sociodemographically diverse patients, some declined participation, and given funding timelines, the user-testing evaluations were more limited in diversity and could have benefited from a more diverse patient population. Recruiting from both academic and community settings in the future randomized controlled trial is planned to assess the real-world benefits of a GHP and to reduce the possibility of contributing or exacerbating existing digital disparities [5,44,72].

Randomized Trial

The efficacy of MyCancerGene is being evaluated in a randomized study (NCT04774445) and compared to usual care following the receipt of genetic testing results in real-world clinical patients. In this ongoing study, we hypothesize that the intervention will be associated with short-term and longitudinal increases in knowledge, decreases in distress, increases in communication with relatives and health care providers, and increases in adoption of risk-reducing health behaviors.

Conclusions

The MyCancerGene digital GHP was developed with extensive feedback from patients and genetic providers and may be a useful digital health tool to enhance longitudinal patient understanding of and affective and behavioral responses to genetic testing, particularly in the era of evolving evidence and risk information. Evaluation of MyCancerGene in a randomized trial in real-world clinical settings will determine the real-world uptake and clinical risks and benefits of this portal, which will ultimately contribute to the integration and promise of personalized genetic medicine.

Acknowledgments

Primary financial support for this work was supported by a Research Scholar grant (RSG-18-192-CPHPS) from the American Cancer Society (AB).

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

BE, SD, LF, LIW, KW, and AB were responsible for conception and design. BE, SD, LF, LIW, KW, AB were responsible for conception and design. PT, D Fetzer, BM, KS, LG, CC, JEE, D Falcone, CL, EMW, KK, SP, ES contributed to collection and assembly of data. BE, AB, and PT were responsible for data analysis and interpretation. PT, D Fetzer, BM, KS, LG, BE, SD, LF, LIW, KW, CC, JEE, D Falcone, CL, EMW, KK, SP, ES, and AB contributed to manuscript writing.

Conflicts of Interest

None declared.

References

1. Khoury MJ, Gwinn M, Yoon PW, Dowling N, Moore CA, Bradley L. The continuum of translation research in genomic medicine: how can we accelerate the appropriate integration of human genome discoveries into health care and disease prevention? Genet Med 2007 Oct;9(10):665-674. [doi: 10.1097/gim.0b013e31815699d0]
2. Collins FS, Varmus H. A new initiative on precision medicine. N Engl J Med 2015 Feb 26;372(9):793-795. [doi: 10.1056/nejmp1500523]
3. Burke W, Psaty BM. Personalized medicine in the era of genomics. JAMA 2007 Oct 10;298(14):1682-1684. [doi: 10.1001/jama.298.14.1682] [Medline: 17925520]
4. Bombard Y, Bach PB, Offit K. Translating genomics in cancer care. J Natl Compr Canc Netw 2013 Nov 13;11(11):1343-1353. [doi: 10.6004/jnccn.2013.0158] [Medline: 24225968]
5. Bombard Y, Ginsburg GS, Sturm AC, Zhou AY, Lemke AA. Digital health-enabled genomics: opportunities and challenges. Am J Hum Genet 2022 Jul 07;109(7):1190-1198 [FREE Full text] [doi: 10.1016/j.ajhg.2022.05.001] [Medline: 35803232]
6. Tung N, Domchek SM, Stadler Z, Nathanson KL, Couch F, Garber JE, et al. Counselling framework for moderate-penetrance cancer-susceptibility mutations. Nat Rev Clin Oncol 2016 Sep 14;13(9):581-588 [FREE Full text] [doi: 10.1038/nrclinonc.2016.90] [Medline: 27296296]

XSL•FO
RenderX

7. Domchek SM, Bradbury A, Garber JE, Offit K, Robson ME. Multiplex genetic testing for cancer susceptibility: out on the high wire without a net? J Clin Oncol 2013 Apr 01;31(10):1267-1270. [doi: 10.1200/jco.2012.46.9403]
8. Bradbury AR, Patrick-Miller L, Domchek S. Multiplex genetic testing: reconsidering utility and informed consent in the era of next-generation sequencing. Genet Med 2015 Feb;17(2):97-98 [FREE Full text] [doi: 10.1038/gim.2014.85] [Medline: 25032987]
9. Pilarski R. How have multigene panels changed the clinical practice of genetic counseling and testing. J Natl Compr Canc Netw 2021 Jan 06;19(1):103-108. [doi: 10.6004/jnccn.2020.7674] [Medline: 33406496]
10. Hamilton JG, Robson ME. Psychosocial effects of multigene panel testing in the context of cancer genomics. Hastings Cent Rep 2019 May 03;49 Suppl 1(Suppl 1):S44-S52 [FREE Full text] [doi: 10.1002/hast.1016] [Medline: 31268573]
11. Burke W, Parens E, Chung WK, Berger SM, Appelbaum PS. The challenge of genetic variants of uncertain clinical significance. Ann Intern Med 2022 Jul;175(7):994-1000. [doi: 10.7326/m21-4109]
12. Balmaña J, Digiovanni L, Gaddam P, Walsh MF, Joseph V, Stadler ZK, et al. Conflicting interpretation of genetic variants and cancer risk by commercial laboratories as assessed by the prospective registry of multiplex testing. J Clin Oncol 2016 Dec 01;34(34):4071-4078. [doi: 10.1200/jco.2016.68.4316]
13. Susswein LR, Marshall ML, Nusbaum R, Vogel Postula KJ, Weissman SM, Yackowski L, et al. Pathogenic and likely pathogenic variant prevalence among the first 10,000 patients referred for next-generation cancer panel testing. Genet Med 2016 Aug;18(8):823-832 [FREE Full text] [doi: 10.1038/gim.2015.166] [Medline: 26681312]
14. Adam F, Fluri M, Scherz A, Rabaglio M. Occurrence of variants of unknown clinical significance in genetic testing for hereditary breast and ovarian cancer syndrome and Lynch syndrome: a literature review and analytical observational retrospective cohort study. BMC Med Genomics 2023 Jan 16;16(1):7 [FREE Full text] [doi: 10.1186/s12920-023-01437-7] [Medline: 36647026]
15. Gould D, Walker R, Makari-Judson G, Seven M. Experiences of individuals with a variant of uncertain significance on genetic testing for hereditary cancer risks: a mixed method systematic review. J Community Genet 2022 Aug 12;13(4):371-379 [FREE Full text] [doi: 10.1007/s12687-022-00600-4] [Medline: 35819584]
16. Pyeritz RE. The coming explosion in genetic testing — is there a duty to recontact? N Engl J Med 2011 Oct 13;365(15):1367-1369. [doi: 10.1056/nejmp1107564]
17. Makhnoon S, Levin B, Ensinger M, Mattie K, Volk RJ, Zhao Z, et al. A multicenter study of clinical impact of variant of uncertain significance reclassification in breast, ovarian and colorectal cancer susceptibility genes. Cancer Med 2023 Feb 24;12(3):2875-2884 [FREE Full text] [doi: 10.1002/cam4.5202] [Medline: 36426404]
18. Kwong A, Ho CY, Shin VY, Au CH, Chan TL, Ma ES. How does re-classification of variants of unknown significance (VUS) impact the management of patients at risk for hereditary breast cancer? BMC Med Genomics 2022 May 31;15(1):122 [FREE Full text] [doi: 10.1186/s12920-022-01270-4] [Medline: 35641994]
19. Cheon JY, Mozersky J, Cook-Deegan R. Variants of uncertain significance in BRCA: a harbinger of ethical and policy issues to come? Genome Med 2014 Dec 19;6(12):121 [FREE Full text] [doi: 10.1186/s13073-014-0121-3] [Medline: 25593598]
20. Murray ML, Cerrato F, Bennett RL, Jarvik GP. Follow-up of carriers of BRCA1 and BRCA2 variants of unknown significance: variant reclassification and surgical decisions. Genet Med 2011 Dec;13(12):998-1005. [doi: 10.1097/gim.0b013e318226fc15]
21. Hoffman-Andrews L. The known unknown: the challenges of genetic variants of uncertain significance in clinical practice. J Law Biosci 2017 Dec;4(3):648-657 [FREE Full text] [doi: 10.1093/jlb/lsx038] [Medline: 29868193]
22. Peterson JF, Roden DM, Orlando LA, Ramirez AH, Mensah GA, Williams MS. Building evidence and measuring clinical outcomes for genomic medicine. The Lancet 2019 Aug;394(10198):604-610. [doi: 10.1016/s0140-6736(19)31278-4]
23. Yanes T, Willis AM, Meiser B, Tucker KM, Best M. Psychosocial and behavioral outcomes of genomic testing in cancer: a systematic review. Eur J Hum Genet 2019 Jan 11;27(1):28-35 [FREE Full text] [doi: 10.1038/s41431-018-0257-5] [Medline: 30206354]
24. Culver JO, Ricker CN, Bonner J, Kidd J, Sturgeon D, Hodan R, et al. Psychosocial outcomes following germline multigene panel testing in an ethnically and economically diverse cohort of patients. Cancer 2021 Apr 15;127(8):1275-1285 [FREE Full text] [doi: 10.1002/cncr.33357] [Medline: 33320347]
25. Carlsson L, Thain E, Gillies B, Metcalfe K. Psychological and health behaviour outcomes following multi-gene panel testing for hereditary breast and ovarian cancer risk: a mini-review of the literature. Hered Cancer Clin Pract 2022 Jun 22;20(1):25 [FREE Full text] [doi: 10.1186/s13053-022-00229-x] [Medline: 35733200]
26. Jacobs C, Patch C, Michie S. Communication about genetic testing with breast and ovarian cancer patients: a scoping review. Eur J Hum Genet 2019 Apr 20;27(4):511-524 [FREE Full text] [doi: 10.1038/s41431-018-0310-4] [Medline: 30573802]
27. Lumish HS, Steinfeld H, Koval C, Russo D, Levinson E, Wynn J, et al. Impact of panel gene testing for hereditary breast and ovarian cancer on patients. J Genet Couns 2017 Oct 29;26(5):1116-1129 [FREE Full text] [doi: 10.1007/s10897-017-0090-y] [Medline: 28357778]

XSL•FO
RenderX

28. Mighton C, Shickh S, Uleryk E, Pechlivanoglou P, Bombard Y. Clinical and psychological outcomes of receiving a variant of uncertain significance from multigene panel testing or genomic sequencing: a systematic review and meta-analysis. Genet Med 2021 Jan;23(1):22-33 [FREE Full text] [doi: 10.1038/s41436-020-00957-2] [Medline: 32921787]
29. Kilbride MK, Egleston BL, Hall MJ, Patrick-Miller LJ, Daly MB, Ganschow P, et al. Longitudinal follow-up after telephone disclosure in the randomized COGENT study. Genet Med 2020 Aug;22(8):1401-1406 [FREE Full text] [doi: 10.1038/s41436-020-0808-3] [Medline: 32376981]
30. Swisher EM, Rayes N, Bowen D, Peterson CB, Norquist BM, Coffin T, et al. Remotely delivered cancer genetic testing in the making genetic testing accessible (MAGENTA) trial: a randomized clinical trial. JAMA Oncol 2023 Nov 01;9(11):1547-1555. [doi: 10.1001/jamaoncol.2023.3748] [Medline: 37707822]
31. Rana HQ, Rainville IR, Vaidya A. Genetic testing in the clinical care of patients with pheochromocytoma and paraganglioma. Curr Opin Endocrinol Diabetes Obes 2014 Jun;21(3):166-176. [doi: 10.1097/MED.0000000000000059] [Medline: 24739310]
32. Rashkin MD, Bowes J, Dunaway K, Dhaliwal J, Loomis E, Riffle S, et al. Genetic counseling, 2030: an on-demand service tailored to the needs of a price conscious, genetically literate, and busy world. J Genet Couns 2019 Apr 09;28(2):456-465. [doi: 10.1002/jgc4.1123] [Medline: 30964579]
33. Terry AB, Wylie A, Raspa M, Vogel B, Sanghavi K, Djurdjinovic L, et al. Clinical models of telehealth in genetics: a regional telegenetics landscape. J Genet Couns 2019 Jun 02;28(3):673-691. [doi: 10.1002/jgc4.1088] [Medline: 30825358]
34. Institute of Medicine, Committee on Quality of Health Care in America. Crossing the Quality Chasm: A New Health System for the 21st Century. Washington, DC: National Academies Press; 2001.
35. Epstein RM, Fiscella K, Lesser CS, Stange KC. Why the nation needs a policy push on patient-centered health care. Health Aff (Millwood) 2010 Aug;29(8):1489-1495. [doi: 10.1377/hlthaff.2009.0888] [Medline: 20679652]
36. Ammenwerth E, Schnell-Inderst P, Hoerbst A. The impact of electronic patient portals on patient care: a systematic review of controlled trials. J Med Internet Res 2012 Nov 26;14(6):e162 [FREE Full text] [doi: 10.2196/jmir.2238] [Medline: 23183044]
37. Lin CT, Wittevrongel L, Moore L, Beaty BL, Ross SE. An internet-based patient-provider communication system: randomized controlled trial. J Med Internet Res 2005 Aug 05;7(4):e47 [FREE Full text] [doi: 10.2196/jmir.7.4.e47] [Medline: 16236699]
38. Delbanco T, Sands DZ. Electrons in flight — e-mail between doctors and patients. N Engl J Med 2004 Apr 22;350(17):1705-1707. [doi: 10.1056/nejmp038209]
39. Calixte R, Rivera A, Oridota O, Beauchamp W, Camacho-Rivera M. Social and demographic patterns of health-related internet use among adults in the United States: a secondary data analysis of the health information national trends survey. Int J Environ Res Public Health 2020 Sep 19;17(18):6856 [FREE Full text] [doi: 10.3390/ijerph17186856] [Medline: 32961766]
40. Perrin A, Atske S. 7% of Americans don't use the internet. Who are they? Pew Research Center. 2021 Apr 2. URL: https://www.pewresearch.org/short-reads/2021/04/02/7-of-americans-dont-use-the-internet-who-are-they/ [accessed 2025-02-06]
41. Liederman EM, Morefield CS. Web messaging: a new tool for patient-physician communication. J Am Med Inform Assoc 2003;10(3):260-270 [FREE Full text] [doi: 10.1197/jamia.M1259] [Medline: 12626378]
42. Dendere R, Slade C, Burton-Jones A, Sullivan C, Staib A, Janda M. Patient portals facilitating engagement with inpatient electronic medical records: a systematic review. J Med Internet Res 2019 Apr 11;21(4):e12779 [FREE Full text] [doi: 10.2196/12779] [Medline: 30973347]
43. Murray E, Burns J, See TS, Lai R, Nazareth I. Interactive health communication applications for people with chronic disease. Cochrane Database Syst Rev 2005 Oct 19(4):CD004274. [doi: 10.1002/14651858.CD004274.pub4] [Medline: 16235356]
44. Bombard Y, Hayeems RZ. How digital tools can advance quality and equity in genomic medicine. Nat Rev Genet 2020 Sep 29;21(9):505-506 [FREE Full text] [doi: 10.1038/s41576-020-0260-x] [Medline: 32601319]
45. Kruse CS, Bolton K, Freriks G. The effect of patient portals on quality outcomes and its implications to meaningful use: a systematic review. J Med Internet Res 2015 Feb 10;17(2):e44 [FREE Full text] [doi: 10.2196/jmir.3171] [Medline: 25669240]
46. Børøsund E, Cvancarova M, Moore SM, Ekstedt M, Ruland CM. Comparing effects in regular practice of e-communication and web-based self-management support among breast cancer patients: preliminary results from a randomized controlled trial. J Med Internet Res 2014 Dec 18;16(12):e295 [FREE Full text] [doi: 10.2196/jmir.3348] [Medline: 25525672]
47. Goldzweig CL, Orshansky G, Paige NM, Towfigh AA, Haggstrom DA, Miake-Lye I, et al. Electronic patient portals: evidence on health outcomes, satisfaction, efficiency, and attitudes: a systematic review. Ann Intern Med 2013 Nov 19;159(10):677-687. [doi: 10.7326/0003-4819-159-10-201311190-00006] [Medline: 24247673]
48. Kruse CS, Argueta DA, Lopez L, Nair A. Patient and provider attitudes toward the use of patient portals for the management of chronic disease: a systematic review. J Med Internet Res 2015 Feb 20;17(2):e40 [FREE Full text] [doi: 10.2196/jmir.3703] [Medline: 25707035]
49. Yun YH, Lee KS, Kim YW, Park SY, Lee ES, Noh DY, et al. Web-based tailored education program for disease-free cancer survivors with cancer-related fatigue: a randomized controlled trial. J Clin Oncol 2012 Apr 20;30(12):1296-1303. [doi: 10.1200/jco.2011.37.2979]

XSL•FO
RenderX

50. Krist AH, Woolf SH, Rothemich SF, Johnson RE, Peele JE, Cunningham TD, et al. Interactive preventive health record to enhance delivery of recommended care: a randomized trial. Ann Fam Med 2012 Jul 09;10(4):312-319 [FREE Full text] [doi: 10.1370/afm.1383] [Medline: 22778119]
51. Ruland CM, Andersen T, Jeneson A, Moore S, Grimsbø GH, Børøsund E, et al. Effects of an internet support system to assist cancer patients in reducing symptom distress: a randomized controlled trial. Cancer Nurs 2013;36(1):6-17. [doi: 10.1097/NCC.0b013e31824d90d4] [Medline: 22495503]
52. Tuil WS, Verhaak CM, Braat DD, de Vries Robbé PF, Kremer JA. Empowering patients undergoing in vitro fertilization by providing internet access to medical data. Fertil Steril 2007 Aug;88(2):361-368. [doi: 10.1016/j.fertnstert.2006.11.197] [Medline: 17416366]
53. Chou WY, Prestin A, Lyons C, Wen KY. Web 2.0 for health promotion: reviewing the current evidence. Am J Public Health 2013 Jan;103(1):e9-18. [doi: 10.2105/ajph.2012.301071]
54. Goel MS, Brown TL, Williams A, Cooper AJ, Hasnain-Wynia R, Baker DW. Patient reported barriers to enrolling in a patient portal. J Am Med Inform Assoc 2011 Dec 01;18 Suppl 1(Suppl 1):i8-12 [FREE Full text] [doi: 10.1136/amiajnl-2011-000473] [Medline: 22071530]
55. Amante DJ, Hogan TP, Pagoto SL, English TM. A systematic review of electronic portal usage among patients with diabetes. Diabetes Technol Ther 2014 Nov;16(11):784-793. [doi: 10.1089/dia.2014.0078] [Medline: 24999599]
56. Wagner PJ, Dias J, Howard S, Kintziger KW, Hudson MF, Seol YH, et al. Personal health records and hypertension control: a randomized trial. J Am Med Inform Assoc 2012 Jul 01;19(4):626-634 [FREE Full text] [doi: 10.1136/amiajnl-2011-000349] [Medline: 22234404]
57. Ronda MC, Dijkhorst-Oei LT, Rutten GE. Reasons and barriers for using a patient portal: survey among patients with diabetes mellitus. J Med Internet Res 2014 Nov 25;16(11):e263 [FREE Full text] [doi: 10.2196/jmir.3457] [Medline: 25424228]
58. Coughlin SS, Prochaska JJ, Williams LB, Besenyi GM, Heboyan V, Goggans DS, et al. Patient web portals, disease management, and primary prevention. Risk Manag Healthc Policy 2017 Apr;Volume 10:33-40. [doi: 10.2147/rmhp.s130431]
59. Irizarry T, DeVito Dabbs A, Curran CR. Patient portals and patient engagement: a state of the science review. J Med Internet Res 2015 Jun 23;17(6):e148 [FREE Full text] [doi: 10.2196/jmir.4255] [Medline: 26104044]
60. Lyles CR, Nelson EC, Frampton S, Dykes PC, Cemballi AG, Sarkar U. Using electronic health record portals to improve patient engagement: research priorities and best practices. Ann Intern Med 2020 Jun 02;172(11_Supplement):S123-S129. [doi: 10.7326/m19-0876]
61. Kim T, White K, DuGoff E. Racial/ethnic variations in social determinants of mental health among medicare advantage beneficiaries. J Appl Gerontol 2022 Mar 17;41(3):690-698. [doi: 10.1177/07334648211039311] [Medline: 34404243]
62. Tuan WJ, Mellott M, Arndt BG, Jones J, Simpson AN. Disparities in use of patient portals among adults in family medicine. J Am Board Fam Med 2022 May 31;35(3):559-569 [FREE Full text] [doi: 10.3122/jabfm.2022.03.210486] [Medline: 35641056]
63. Bradbury A, Patrick-Miller L, Egleston B, Hall MJ, Domchek SM, Daly MB, et al. Randomized noninferiority trial of telephone vs in-person disclosure of germline cancer genetic test results. J Natl Cancer Inst 2018 Sep 01;110(9):985-993 [FREE Full text] [doi: 10.1093/jnci/djy015] [Medline: 29490071]
64. Pope C, van Royen P, Baker R. Qualitative methods in research on healthcare quality. Qual Saf Health Care 2002 Jun;11(2):148-152 [FREE Full text] [doi: 10.1136/qhc.11.2.148] [Medline: 12448807]
65. Ritchie J, Spencer L. Qualitative data analysis for applied policy research. In: Bryman A, Burgess B, editors. Analyzing Qualitative Data. London, UK: Routledge; 1994.
66. Gale NK, Heath G, Cameron E, Rashid S, Redwood S. Using the framework method for the analysis of qualitative data in multi-disciplinary health research. BMC Med Res Methodol 2013 Sep 18;13(1):117 [FREE Full text] [doi: 10.1186/1471-2288-13-117] [Medline: 24047204]
67. Rogers EM. Diffusion of Innovations. New York, NY: Free Press; 1983.
68. Berwick DM. Disseminating innovations in health care. JAMA 2003 Apr 16;289(15):1969-1975. [doi: 10.1001/jama.289.15.1969] [Medline: 12697800]
69. Greenhalgh T, Stramer K, Bratan T, Byrne E, Mohammad Y, Russell J. Introduction of shared electronic records: multi-site case study using diffusion of innovation theory. BMJ 2008 Oct 23;337(oct23 1):a1786 [FREE Full text] [doi: 10.1136/bmj.a1786] [Medline: 18948344]
70. Pelletier AC, Jethwani K, Bello H, Kvedar J, Grant RW. Implementing a web-based home monitoring system within an academic health care network: barriers and facilitators to innovation diffusion. J Diabetes Sci Technol 2011 Jan 01;5(1):32-38 [FREE Full text] [doi: 10.1177/193229681100500105] [Medline: 21303622]
71. Bradbury AR, Patrick-Miller L, Long J, Powers J, Stopfer J, Forman A, et al. Development of a tiered and binned genetic counseling model for informed consent in the era of multiplex testing for cancer susceptibility. Genet Med 2015 Jun;17(6):485-492 [FREE Full text] [doi: 10.1038/gim.2014.134] [Medline: 25297947]
72. Kilbride M, Egleston BL, Chung WK, Olopade O, Maxwell KN, Shah P, et al. Uptake of genetic research results and patient-reported outcomes with return of results incorporating web-based predisclosure education. J Clin Oncol 2023 Nov 01;41(31):4905-4915. [doi: 10.1200/jco.22.00516]

XSL•FO
RenderX

Abbreviations

COGENT: Communication of Genetic Test Results by Telephone
EHR: electronic health record
GHP: genetic health portal
MGPT: multigene panel testing
OR: odds ratio
VUS: variants of uncertain significance

Edited by A Mavragani; submitted 29.02.24; peer-reviewed by A Hassan, R Bidkar, J Rivera Rivera; comments to author 20.05.24; revised version received 08.07.24; accepted 06.09.24; published 14.05.25.

Please cite as:
Trieu P, Fetzer D, McLeod B, Schweickert K, Gutstein L, Egleston B, Domchek S, Fleisher L, Wagner L, Wen KY, Cacioppo C, Ebrahimzadeh JE, Falcone D, Langer C, Wood E, Karpink K, Posen S, Selmani E, Bradbury AR
Developing the MyCancerGene Digital Health Portal to Improve Patients' Understanding of Germline Cancer Genetic Test Results: Development, User, and Usability Testing Study
JMIR Form Res 2025;9:e56282
URL: https://formative.jmir.org/2025/1/e56282
doi:10.2196/56282
PMID:

©Phillip Trieu, Dominique Fetzer, Briana McLeod, Kathryn Schweickert, Lauren Gutstein, Brian Egleston, Susan Domchek, Linda Fleisher, Lynne Wagner, Kuang-Yi Wen, Cara Cacioppo, Jessica E Ebrahimzadeh, Dana Falcone, Claire Langer, Elisabeth Wood, Kelsey Karpink, Shelby Posen, Enida Selmani, Angela R Bradbury. Originally published in JMIR Formative Research (https://formative.jmir.org), 14.05.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Evaluating the Accuracy and Reliability of Real-World Digital Mobility Outcomes in Older Adults After Hip Fracture: Cross-Sectional Observational Study

Martin A Berge[1], MSc; Anisoara Paraschiv-Ionescu[2], PhD; Cameron Kirk[3], PhD; Arne Küderle[4], MSc; Encarna Micó-Amigo[3], PhD; Clemens Becker[5,6], PhD; Andrea Cereatti[7], PhD; Silvia Del Din[3,8], PhD; Monika Engdal[1], MSc; Judith Garcia-Aymerich[9,10,11], PhD; Karoline B Grønvik[1], MSc; Clint Hansen[12], PhD; Jeffrey M Hausdorff[13,14,15,16,17], PhD; Jorunn L Helbostad[1], PhD; Carl-Philipp Jansen[5,6], PhD; Lars Gunnar Johnsen[1,18], PhD; Jochen Klenk[6,19,20], PhD; Sarah Koch[9,10,11,21], PhD; Walter Maetzler[12], PhD; Dimitrios Megaritis[3,22], PhD; Arne Müller[23], PhD; Lynn Rochester[3,8], PhD; Lars Schwickert[6], PhD; Kristin Taraldsen[24], PhD; Beatrix Vereijken[1], PhD

[1]Department of Neuromedicine and Movement Science, Norwegian University of Science and Technology, Trondheim, Norway

[2]Laboratory of Movement Analysis and Measurement, Ecole Polytechnique Federale de Lausanne, Lausanne, Switzerland

[3]Translational and Clinical Research Institute, Faculty of Medical Sciences, Newcastle University, Newcastle Upon Tyne, United Kingdom

[4]Machine Learning and Data Analytics Lab, Department Artificial Intelligence in Biomedical Engineering, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

[5]Geriatric Center, Medical Faculty Heidelberg, Heidelberg University, Heidelberg, Germany

[6]Department of Geriatrics and Rehabilitation, Robert Bosch Hospital, Stuttgart, Germany

[7]Department of Electronics and Telecommunications, Politecnico di Torino, Turin, Italy

[8]National Institute for Health and Care Research Newcastle Biomedical Research Centre, Newcastle University and the Newcastle Upon Tyne Hospitals NHS Foundation Trust, Newcastle Upon Tyne, United Kingdom

[9]Barcelona Institute for Global Health, Barcelona, Spain

[10]Department of Medicine and Life Sciences, Universitat Pompeu Fabra, Barcelona, Catalonia, Spain

[11]CIBER Epidemiología y Salud Pública, Madrid, Spain

[12]Department of Neurology, University Hospital Schleswig-Holstein and Kiel University, Kiel, Germany

[13]Center for the Study of Movement, Cognition, and Mobility, Neurological Institute, Tel Aviv Medical Center, Tel Aviv, Israel

[14]Department of Physical Therapy, Faculty of Medical & Health Sciences, Tel Aviv University, Tel Aviv, Israel

[15]Sagol School of Neuroscience, Tel Aviv University, Tel Aviv, Israel

[16]Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago, IL, United States

[17]Department of Orthopedic Surgery, Rush Medical College, Rush University, Chicago, IL, United States

[18]Department of Orthopaedic Surgery, St. Olav's Hospital, Trondheim, Norway

[19]Institute of Epidemiology and Medical Biometry, Ulm University, Ulm, Germany

[20]IB University of Health and Social Sciences, Study Centre Stuttgart, Stuttgart, Germany

[21]Department of Sport, Exercise, and Health, University Basel, Basel, Switzerland

[22]Department of Sport, Exercise and Rehabilitation, Northumbria University Newcastle, Newcastle Upon Tyne, United Kingdom

[23]Novartis Biomedical Research, Novartis Pharma AG, Basel, Switzerland

[24]Department of Rehabilitation Science and Health Technology, OsloMet, Oslo, Norway

Corresponding Author:
Martin A Berge, MSc
Department of Neuromedicine and Movement Science
Norwegian University of Science and Technology
Helgasetr, 4th Fl.
Vanglunds gate 2
Trondheim, 7030
Norway
Phone: 47 92885664
Email: martin.a.berge@ntnu.no

XSL•FO
RenderX

Abstract

Background: Algorithms estimating real-world digital mobility outcomes (DMOs) are increasingly validated in healthy adults and various disease cohorts. However, their accuracy and reliability in older adults after hip fracture, who often walk slowly for short durations, is underexplored.

Objective: This study examined DMO accuracy and reliability in a hip fracture cohort considering walking bout (WB) duration, physical function, days since surgery, and walking aid use.

Methods: In total, 19 community-dwelling participants were real-world monitored for 2.5 hours using a lower back wearable device and a reference system combining inertial modules, distance sensors, and pressure insoles. A total of 6 DMO estimates from 164 WBs from 58% (11/19) of the participants (aged 71-90 years; assessed 32-390 days after surgery; Short Physical Performance Battery [SPPB] scores of 3-12; gait speed range 0.39-1.34 m/s) were assessed against the reference system at the WB and participant level. We stratified by WB duration (all WBs, WBs of >10 seconds, WBs of 10-30 seconds, and WBs of >30 seconds) and lower versus higher SPPB scores and observed whether days since surgery and walking aid use affected DMO accuracy and reliability.

Results: Across WBs, walking speed and distance ranged from 0.25 to 1.29 m/s and from 1.7 to 436.5 m, respectively. Estimation of walking speed, cadence, stride duration, number of steps, and distance stratified by WB duration showed intraclass correlation coefficients (ICCs) ranging from 0.50 to 0.99 and mean relative errors (MREs) from –6.9% to 12.8%. Stride length estimation showed poor reliability, with ICCs ranging from 0.30 to 0.49 and MREs from 6.1% to 13.2%. Walking speed and distance ICCs in the higher–SPPB score group ranged from 0.85 to 0.99, and MREs ranged from –10.1% to –1.7%. In the lower–SPPB score group, walking speed and distance ICCs ranged from 0.17 to 0.99, and MREs ranged from 13.5% to 32.6%. There was no discernible effect of time since surgery or walking aid use.

Conclusions: In total, 5 accurate and reliable real-world DMOs were identified in older adults after hip fracture: walking speed, cadence, stride duration, number of steps, and distance. Accuracy and reliability of most DMOs improved when excluding WBs of <10 seconds and were higher for WBs of >30 seconds than for WBs of 10 to 30 seconds and for participants with higher physical function. DMOs capture daily gait as early as 1 month after surgery also in people using walking aids. However, as most WBs in this cohort were short, there was a trade-off between improving accuracy and reliability by excluding short WBs and losing a substantial amount of data. These results have important implications for establishing the clinical validity of DMOs and evaluating the effects of interventions on daily-life gait, thereby facilitating the design of optimal care pathways.

(JMIR Form Res 2025;9:e67792) doi:10.2196/67792

KEYWORDS

gait; free living; digital mobility assessment; wearable device; walking speed; distance; cadence; active aging

Introduction

Background

Hip fracture after a fall is one of the most serious injuries in older adults, with a high rate of morbidity and a 1-year mortality of 22% [1,2]. Survivors of hip fracture experience a substantial decline in quality of life and mobility [3]. Mobility, the ability to move or walk around freely and easily, is essential for functioning well, maintaining independence, and ensuring social and emotional well-being [4]. Between 40% and 60% of people after hip fracture do not recover their prefracture level of mobility and ability to perform instrumental activities of daily living [3], and the incidence of nursing home admissions is high [5]. Hip fractures require surgery to restore mobility, but there is large variation in the average length of stay at the hospital across countries, ranging from 4 to 40 days after surgery [6,7]. The main goals of care in the first days are typically pain management and early mobilization, followed by rehabilitation targeting physical function and independent walking to prepare patients for returning to their daily life [8,9]. However, the different recovery trajectories and the factors determining mobility outcomes for various individuals are poorly understood. The ability to walk and regain mobility is an important indicator of recovery, making it a critical focus in the rehabilitation of patients after a hip fracture. However, getting patients back on their feet requires more than just a step in the right direction. Comprehensive mobility assessment is key to supporting effective recovery, providing valuable information about patients' progress and informing rehabilitation strategies and optimal care pathways.

Assessing mobility requires objective, quantifiable metrics that accurately reflect a patient's ability to move around efficiently and independently. While mobility can be described using broader measures such as upright time and number of sit-to-stand transitions [10,11], these primarily provide information about activity amount. In contrast, gait-related metrics provide information about the quality and efficiency of movement, reflecting musculoskeletal function, coordination, and stability. In addition, gait impairments are strong predictors of falls, disability, and overall health and inform about the status and progression of different health challenges [12-14], making walking a key aspect of mobility.

Gait after a hip fracture has mostly been assessed using patient-reported outcome measures that are prone to response bias or via standardized walking tests in clinical settings and laboratory assessments [15]. Traditional supervised gait

XSL•FO
RenderX

assessments using sophisticated technology in laboratory settings, such as gold standard instrumented gait analysis, allow for accurate spatiotemporal measurements of gait in a controlled laboratory environment [14,16]. However, gait assessments under such conditions lack ecological validity for several reasons [17]. Laboratory assessments provide a snapshot in time, are limited by space and infrastructure, are prone to white coat effects, and often consist of isolated and structured tasks that do not necessarily reflect what people with and without mobility impairment do in their daily lives [17-19]. Importantly, they may also limit the inclusion of people from rural areas, potentially overlooking a key demographic [17]. In addition, in-laboratory assessments typically evaluate gait on even, uncluttered ground over short, straight distances only [15]. Therefore, we need more knowledge regarding gait in people's real-world environments. In addition, there is a critical need to complement current short-term follow-up and infrequent clinical tests or assessments with more comprehensive and detailed knowledge about daily-life gait recovery throughout the first years after a hip fracture [15]. Hence, extended monitoring of gait is vital for understanding the long-term recovery process and optimizing intervention strategies [20].

However, assessing accurate gait characteristics in the real world is challenging due to internal and external confounding factors. Real-world gait is highly variable and complex as it entails navigating different walking surfaces and uneven terrain and obstacles and consists of a wide variety of related activities such as turning, stopping, and starting [21]. Methods based on the use of a single inertial measurement unit (IMU) on the lower back are available to assess gait in the real world through the quantification of digital mobility outcomes (DMOs) [20]. However, until recently, algorithms to estimate DMOs lacked comprehensive and systematic validation [22].

The Mobilise-D consortium [23] has made significant strides to change this. Mobilise-D has used a rigorous and comprehensive validation process, including a multisensor wearable setup for real-world analysis [24] and involving 1 healthy and 5 patient cohorts performing an extensive technical validation protocol [22,25]. This resulted in the identification and refinement of algorithms that enable robust gait sequence detection and subsequent estimation of key DMOs from a single inertial device worn on the lower back, such as initial foot contact, cadence, stride length, and walking speed. Previous Mobilise-D work has shown that these DMOs can be estimated accurately and reliably in the real world across a range of cohorts (healthy older adults and adults with Parkinson disease, multiple sclerosis, congestive heart failure, chronic obstructive pulmonary disease, and hip fracture), tasks, and contextual factors [26,27]. Analyses of all 6 Mobilise-D cohorts showed that the real-world algorithm performances were valid and accurate but more challenging in the case of short walking bouts (WBs; <10 seconds) and slower gait speeds (<0.5 m/s) [26,27], typical hallmarks of gait after a hip fracture. Analyses stratified by cohort indicated that the estimated walking speed and cadence in the hip fracture cohort showed moderate reliability and mean relative error (MRE) of 10% and –0.5%, respectively, and estimated stride length with poor reliability and an MRE of 11% [26]. However, these analyses did not consider hip fracture–specific factors that may affect the reliability and accuracy of the DMO estimates, which may be of critical relevance for the hip fracture cohort.

Objectives

Building on the previous Mobilise-D publications [26,27], this study investigated additional DMOs and delved deeper into potential factors that may affect the reliability and accuracy of the real-world DMO estimates in the hip fracture cohort. As short WBs are frequent after a hip fracture [15] and walking distance can influence spatiotemporal parameters [28], we investigated whether different WB duration categories affected the estimates. Furthermore, physical functioning is associated with variations in gait parameters and is highly compromised after a hip fracture [13,29], particularly in the early phases of recovery [30], often necessitating the use of a walking aid [31].

Hence, this study investigated the following research question: can the validated Mobilise-D real-world DMOs be accurately and reliably estimated in older adults after hip fracture considering WB duration, time since surgery, level of physical function, and potential walking aid use? To answer this question, we compared 6 DMOs estimated from a single wearable device against a reference system [24,32] and investigated whether WB duration, physical function, walking aid use, and days since surgery affected the accuracy and reliability of the DMO estimates.

Given earlier results, we expected that most real-world DMOs could be estimated accurately and reliably in the hip fracture cohort and that results would improve further when excluding WBs of short duration. In addition, as shorter WBs tend to have higher errors [26,27], we expected that DMO accuracy and reliability might be lower in older adults with low physical function than in those with better physical functioning. Moreover, as gait function is especially impaired early after a hip fracture, the accuracy and reliability of DMO estimates may be lower in the earlier stages of recovery than in later stages.

Methods

Study Design and Participants

This multicenter observational study used data from the technical validation study (TVS) of the Innovative Medicines Initiative 2 Joint Undertaking–funded Mobilise-D project [22,23,25]. Data were collected between July 2020 and March 2022. The participants after hip fracture were recruited from the Robert Bosch Foundation for Medical Research (Germany) and Kiel University (Germany).

The participants were recruited within 13 months of surgical treatment (fixation or arthroplasty) for a low-energy fracture of the proximal femur (*International Classification of Diseases, 10th Revision*, diagnosis codes S72.0, S72.1, and S72.2), as diagnosed through x-rays of the hip and pelvis.

Participants had to be aged ≥65 years to be included. Participants were excluded if they were unable to walk 4 m independently with or without a walking aid; had a shoe size of <36 European size; had a Montreal Cognitive Assessment [33] score of ≤15; or had an occurrence of any of the following within 3 months

XSL•FO
RenderX

before inclusion: myocardial infarction, hospitalization for unstable angina, stroke, coronary artery bypass graft, percutaneous coronary intervention, or implantation of a cardiac resynchronization therapy device [22]. We aimed to recruit 20 participants. Due to challenges posed by the COVID-19 pandemic, recruitment was much slower than anticipated and further hampered by us not being allowed to recruit participants during their hospital stay. As a result, we needed to almost double the recruitment period to be able to enroll 19 participants. Of these 19 participants, 6 (32%) had to be excluded because of technological issues with the pressure insoles of the reference system. Data from the reference system from another 11% (2/19) of the participants had insufficient signal quality. These 2 participants were relatively young (aged 70 and 66 years), had a very low Short Physical Performance Battery (SPPB) score of 3, and were assessed 28 and 64 days after surgery. Data from the remaining 58% (11/19) of the participants were included in the final analysis.

Protocol and Equipment

Participants' activities were monitored for 2.5 hours in a real-world setting within their preferred habitual environment (home, work, or community). The activities were unstructured, but to ensure sufficient variability in the collected data, participants were encouraged to complete several activities, such as rising from a chair and walking to another room; walking up and down a flight of stairs; walking to the kitchen and preparing a drink; walking outdoors (if possible for a minimum of 2 minutes); and, if walking outside, walking up and down an inclined path [22].

Gait data were collected using a single wearable device (McRoberts DynaPort MM+; 100-Hz sampling frequency, –8 g to +8 g triaxial acceleration range and 1-mg resolution, and –2000 to +2000 degrees per second triaxial gyroscope range with a 70–millidegrees per second resolution) worn at the lower back with a Velcro belt. In addition, participants were equipped with a multisensor reference system consisting of IMUs, distance sensors, and pressure insoles named Inertial Module With Distance Sensors and Pressure Insoles (sampling frequency of 100 Hz) [22,24,32,34,35] previously validated with excellent reliability in different cohorts, including hip fracture, across a complex set of different motor tests, including simulated daily activities [24]. Specifically, 2 magneto-IMUs were fixed to the instep using clips. A third IMU was attached to the lower back using Velcro. Asymmetrically positioned distance sensors were fixed above the ankles using Velcro, and 2 pressure insoles were inserted into the shoes. To ensure synchronization, time stamps of the reference system and the single wearable device were aligned (–10 to +10 ms).

Evaluation of DMOs

The evaluation of the DMOs was based on previous Mobilise-D work that selected the top-ranked algorithms to detect gait sequences and estimate initial contact events, stride length, and cadence within identified gait sequences [27]. The best-performing cadence and stride length algorithms for hip fracture were then used to estimate walking speed [26]. The validation of this processing pipeline was based on a minimum 95% CI intraclass correlation coefficient (ICC) threshold for performance metrics (ie, sensitivity, positive predictive value, and accuracy) of at least 0.7 and a relative error of <20%, as described in the studies by Micó-Amigo et al [27] and Kirk et al [26]. All algorithms are available in the MobGap Python library (Python Software Foundation) [36].

DMOs were evaluated at a WB level. A WB was defined as a continuous walking sequence comprising a minimum of 2 consecutive strides of both feet [37]. WBs were separated by breaks of >3 seconds, and for a stride to be included, it required a duration of 0.2 to 3 seconds and a minimum length of 0.15 m [27]. These criteria were applied to generate WBs for both the single wearable device and the reference system by initially filtering the identified strides according to the stride level definition and then assembling them into final WBs by identifying breaks in the stride sequence. Final DMOs for both systems were calculated as the average value over all strides within a WB. For a rigorous comparison of DMOs at a WB level, it was essential to focus on WBs concurrently detected by both systems using a true-positive analysis approach. Accordingly, we considered WBs with a time overlap exceeding 80% of their duration as true positives, as detailed in the work by Kirk et al [26].

For each WB, 6 gait characteristics were obtained from the single wearable device and the reference system: cadence (steps per minute; the number of steps taken per minute), stride length (meters; the length of 2 consecutive steps), number of steps, stride duration (seconds) [27], walking speed (m/s) [26], and distance (meters). The walked distance was calculated by multiplying 2 validated DMOs [27]: average walking speed × WB duration.

Variables

Participant Characteristics

Age, height, weight, sex, cognitive function (Montreal Cognitive Assessment), fracture type, and surgical implant were collected for all participants. Pain while walking was assessed using a visual analogue scale (from 0 to 100, where 0 is no pain and 100 is worst pain imaginable).

WB Duration

Given the larger single wearable device DMO errors observed in shorter WBs [26,27], we assessed whether the accuracy and reliability of DMO estimates differed when excluding WB durations of <10 seconds [38,39]. In addition, we divided WBs of >10 seconds into 2 subcategories: WBs of 10 to 30 seconds and WBs of >30 seconds. Thus, we analyzed the DMOs in 4 categories of WB duration: all WBs combined, WBs of >10 seconds, WBs of 10 to 30 seconds, and WBs of >30 seconds. This was performed at both a WB level and a participant level.

Physical Function

The participants' physical function was assessed using the SPPB 7 days before or after the 2.5-hour activity monitoring. The SPPB consists of a static balance test, a 5-time chair rise test, and a 4-m walk test at a comfortable gait speed. The total SPPB score ranges from 0 to 12, where higher scores indicate better mobility capacity and a score of <8 points indicates impaired activity of daily living functions [40]. Due to the larger DMO

XSL•FO
RenderX

errors in cohorts with more impaired gait and slower walking speeds [26,27], we divided participants into 2 physical function groups based on the 8-point SPPB threshold: a lower–SPPB score group consisting of participants with a total score between 0 and 7 and a higher–SPPB score group consisting of participants with a total score between 8 and 12. This threshold resulted in 2 groups of 45% (5/11) and 55% (6/11) of the participants, respectively.

Time Since Surgery and Walking Aid Use

Physical function and mobility typically improve gradually in the first year after a hip fracture [15,41,42]. We collected the number of days between the surgery and the 2.5-hour free-living testing date, aggregated participant DMOs, and sorted them by number of days since surgery. We also collected information on walking aid use (yes or no and type of walking aid).

Statistical Analyses

DMOs were evaluated at a WB level to quantify DMO estimation errors (accuracy) and reliability using the validation metrics detailed in previous Mobilise-D work [26,27]. For accuracy, absolute agreement was assessed by calculating the mean error, mean absolute error, and precision (limits of agreements) [43] between DMO estimates from the single wearable device and the reference system. MREs and mean absolute relative errors (MAREs) were calculated by dividing the (absolute) errors per WB by the corresponding estimates from the reference system, expressed in percentage, as shown in the following formulas (*SWD* refers to the single wearable device, and *RS* refers to the reference system):

1. Mean error = $(1 / n) \times \Sigma$ from i=1 to n $([DMO_SWD_i - DMO_RS_i])$
2. MRE = $(1 / n) \times \Sigma$ from i=1 to n $([(DMO_SWD_i - DMO_RS_i)/DMO_RS_i] \times 100)$
3. Mean absolute error = $(1 / n) \times \Sigma$ from i=1 to n $(|DMO_SWD_i - DMO_RS_i|)$
4. MARE = $(1 / n) \times \Sigma$ from i=1 to n $([|DMO_SWD_i - DMO_RS_i|/DMO_RS_i] \times 100)$

For reliability, the ICC(2,1) [44] was computed to evaluate how closely each of the DMOs of the 2 systems were related. ICC values of <0.5 were considered poor, ICC values between 0.5 and 0.75 were considered moderate, ICC values between 0.75 and 0.9 were considered good, and ICC values of >0.90 were considered excellent [45].

For analyses stratified by WB duration, all 6 DMOs were included. In analyses stratified by the 2 SPPB score groups, we included the 2 most relevant DMOs for clinicians and patients: walking speed [15] and distance [46].

Walking speed and distance were also evaluated at a participant level. Histograms, *Q-Q* plots, and the Anderson-Darling tests were used for both variables to assess normality. Given that the DMOs were not normally distributed and the relatively low number of participants, median DMO values were computed across all WBs for each participant for the single wearable device and the reference system. The median DMO values from both systems were visualized in bar graphs with IQRs.

Data preparation and visualization were conducted in MATLAB (R2022a; MathWorks), whereas statistical analyses were conducted in Stata (version 18.0; StataCorp).

Ethical Considerations

The protocol was approved by the Faculty of Medicine of the University of Tübingen (647/2019BO2) and the Faculty of Medicine of Kiel University (D540/19). All participants gave their written informed consent. The analysis was conducted using pseudonymized data. No financial compensation was provided to participants in this study.

Results

Participant Characteristics

The 11 participants included in the analyses (n=6, 55% men and n=5, 45% women) were community dwelling, and their characteristics are outlined in Table 1. Notably, the study sample included older adults after hip fracture assessed between 32 and 390 days after surgery. Across the 164 WBs, walking speed and distance ranged from 0.25 to 1.29 m/s and from 1.7 to 436.5 m, respectively, as measured using the reference system, and 22% (36/164) of these WBs were of <0.5 m/s. Of the 11 included participants, 9 (82%) had ≥2 WBs of <0.5 m/s, and 2 (18%) used a walking aid during the observation period, both using a single cane or crutch. Of the 11 participants, 5 (45%) had trochanteric fractures and were treated with intramedullary nails, whereas 6 (55%) had cervical fractures, of whom 4 (67%) received a hemiprosthesis and 2 (33%) received a total prosthesis.

XSL•FO
RenderX

Table 1. Characteristics of individual participants and the overall sample.

	Participant											Median (IQR)
	1	2	3	4	5	6	7	8	9	10	11	
Days since surgery	32	39	60	114	141	179	193	200	244	369	390	179 (60-244)
Age (y)	71	76	80	84	79	83	87	72	71	83	90	80 (72-84)
Height (cm)	180	182	169	180	158	159	174	174	158	165	174	173.5 (159-180)
Weight (kg)	78	85	66	69	54	44	71	95	53	66	96	68.5 (54-85)
MoCA[a] score	26	30	27	28	30	18	25	30	22	21	19	26 (21-30)
Walking pain (0-100)	5	3	2	4	0	30	19	4	52	2	1	4 (2-19)
SPPB[b] score	9	4	3	9	10	7	11	10	6	12	4	9 (4-10)
Gait speed (4MWT[c]; m/s)	0.79	0.39	0.41	0.89	0.80	0.83	1.34	1.05	0.75	1.08	0.62	0.80 (0.69-0.97)
WBs[d] of <0.5 m/s, n (%)[e]	2 (14)[f]	0 (0)	10 (83)[g]	3 (14)[h]	2 (7)[i]	3 (25)[g]	2 (14)[f]	3 (18)[j]	8 (23)[k]	0 (0)	3 (30)[l]	—[m]
Sex	Male	Male	Female	Male	Female	Female	Male	Male	Female	Female	Male	—
Fracture type	C[n]	C	C	C	T[o]	T	T	C	T	T	C	—
Implant	H[p]	TP[q]	TP	H	Nail	Nail	Nail	H	Nail	Nail	H	—

[a]MoCA: Montreal Cognitive Assessment.
[b]SPPB: Short Physical Performance Battery.
[c]4MWT: 4-m walk test.
[d]WB: walking bout.
[e]Number of matched WBs with a speed of <0.5 m/s as measured using the reference system.
[f]n=14.
[g]n=12.
[h]n=21.
[i]n=27.
[j]n=17.
[k]n=35.
[l]n=10.
[m]Not applicable.
[n]C: cervical.
[o]T: trochanteric.
[p]H: hemiprosthesis.
[q]TP: total prosthesis.

WB Duration

In total, 164 WBs met our true-positive approach [26], of which 65 (39.6%) were of <10 seconds, 60 (36.6%) were between 10 and 30 seconds, and 39 (23.8%) were of >30 seconds (Table 2). Estimations of the number of steps and distance demonstrated good to excellent reliability in all WB duration categories, with ICCs exceeding 0.83, MREs ranging from –6.87% to 5.34%, and MAREs ranging from 4.96% to 19.79% (Table 3). Walking speed and stride duration showed moderate reliability for all WBs combined and for WBs of >10 seconds, with ICCs ranging from 0.64 to 0.71, MREs ranging from –5.68% to 10.35%, and MAREs ranging from 7.51% to 19.75%. Walking speed showed good reliability and less error in longer WBs of >30 seconds (ICC=0.80; MRE=5.09%; MARE=14.5%) compared to shorter WBs of 10 to 30 seconds (ICC=0.50; MRE=12.78%; MARE=20.81%). Similarly, stride duration showed good reliability and less error in WBs of >30 seconds (ICC=0.77; MRE=–4.98%; MARE=5.3%) compared to WBs of 10 to 30 seconds (ICC=0.54; MRE=–6.13%; MARE=8.94%). Cadence

XSL•FO
RenderX

showed moderate to excellent reliability in all WB categories, with ICCs ranging from 0.74 to 0.98, low MREs ranging from –0.56% to –0.34%, and MAREs ranging from 1.75% to 5.62%. In contrast, stride length showed poor reliability, with ICCs ranging from 0.30 to 0.49, an overestimation in MRE varying from 6.05% to 13.23%, and MAREs ranging from 15.46% to 20.61%.

Table 2. The number of true-positive walking bouts (WBs) in each WB duration category for all participants in total, per Short Physical Performance Battery (SPPB) score group, and per participant sorted by days since surgery (N=164).

WB duration[a]	Number of WBs, n (%)	True-positive WBs per SPPB score group, n (%)		True-positive WBs by days since surgery, n (%)										
		Lower	Higher	32 days	39 days	60 days	114 days	141 days	179 days	193 days	200 days	244 days	369 days	390 days
All	164 (100)	70 (42.7)	94 (57.3)	14 (8.5)	1 (0.6)	12 (7.3)	21 (12.8)	27 (16.5)	12 (7.3)	14 (8.5)	17 (10.4)	35 (21.3)	1 (0.6)	10 (6.1)
>10 s	99 (60.4)	42 (25.6)	57 (34.8)	10 (6.1)	1 (0.6)	8 (4.9)	15 (9.1)	15 (9.1)	9 (5.5)	8 (4.9)	8 (4.9)	14 (8.5)	1 (0.6)	10 (6.1)
10-30 s	60 (36.6)	27 (16.5)	33 (20.1)	5 (3)	0 (0)	4 (2.4)	8 (4.9)	13 (7.9)	7 (4.3)	5 (3)	1 (0.6)	11 (6.7)	1 (0.6)	5 (3)
>30 s	39 (23.8)	15 (9.1)	24 (14.6)	5 (3)	1 (0.6)	4 (2.4)	7 (4.3)	2 (1.2)	2 (1.2)	3 (1.8)	7 (4.3)	3 (1.8)	0 (0)	5 (3)

[a]As described in the Methods section, WBs of <10 seconds (n=65) were included in the *all WBs* category and were not reported separately.

XSL•FO
RenderX

Table 3. Digital mobility outcome estimates from the reference system (RS) and single wearable device (SWD), mean errors and mean relative errors (MREs) with limits of agreement (LoA), mean absolute errors (MAEs) and mean absolute relative errors (MAREs), and intraclass correlation coefficients (ICC) for all walking bouts (WBs) and WBs of >10 seconds (the latter also split into WBs of 10 to 30 seconds and WBs of >30 seconds).

WB duration	RS, mean (5% quantile, 95% quantile)	SWD, mean (5% quantile, 95% quantile)	Error, mean (LoA)	MRE (%), mean (LoA)	MAE, mean (5% quantile, 95% quantile)	MARE (%), mean (5% quantile, 95% quantile)	ICC(2,1), mean (5% quantile, 95% quantile)
Walking speed (m/s)							
All[a]	0.65 (0.37, 1.03)	0.69 (0.45, 0.96)	0.03 (–0.25 to 0.31)	10.35 (–50.12 to 70.82)	0.11 (0.01 to 0.33)	19.75 (1.80, 66.32)	0.67 (0.58, 0.73)[b]
>10 s	0.68 (0.39, 1.05)	0.71 (0.44, 1.02)	0.03 (–0.24 to 0.30)	9.75 (–53.53 to 73.03)	0.10 (0.01, 0.33)	18.32 (1.27, 66.32)	0.71 (0.62, 0.79)[b]
10-30 s	0.61 (0.38, 0.98)	0.66 (0.43, 0.89)	0.05 (–0.23 to 0.32)	12.78 (–56.61 to 82.16)	0.11 (0.01, 0.32)	20.81 (1.74, 65.57)	0.50 (0.32, 0.64)[b]
>30 s	0.77 (0.46, 1.19)	0.78 (0.45, 1.10)	0.01 (–0.25 to 0.26)	5.09 (–46.99 to 57.16)	0.09 (0.01, 0.39)	14.50 (1.11, 79.76)	0.80 (0.69, 0.88)[c]
Cadence (steps per min)							
All[a]	89.81 (64.25, 114.21)	88.52 (65.76, 110.95)	–1.29 (–21.79 to 19.21)	–0.56 (–19.50 to 18.39)	5.47 (0.15, 17.50)	5.62 (0.16, 19.43)	0.74 (0.68, 0.80)[b]
>10 s	88.28 (62.32, 111.14)	87.42 (62.90, 104.36)	–0.86 (–14.73 to 13.02)	–0.39 (–13.81 to 13.04)	3.58 (0.08, 12.36)	3.84 (0.08, 13.61)	0.86 (0.80, 0.89)[c]
10-30 s	88.60 (64.83, 110.10)	87.52 (64.83, 102.66)	–1.08 (–18.31 to 16.15)	–0.42 (–17.01 to 16.17)	4.89 (0.16, 15.30)	5.20 (0.20, 15.89)	0.75 (0.64, 0.83)[c]
>30 s	87.77 (58.29, 111.48)	87.26 (59.51, 108.71)	–0.52 (–6.42 to 5.39)	–0.34 (–6.45 to 5.77)	1.57 (0.02, 10.59)	1.75 (0.03, 10.36)	0.98 (0.96, 0.99)[d]
Stride length (m)							
All[a]	0.88 (0.54, 1.24)	0.93 (0.72, 1.19)	0.05 (–0.33 to 0.43)	11.42 (–49.51 to 72.35)	0.15 (0.01, 0.41)	20.40 (1.38, 64.86)	0.45 (0.34, 0.55)[e]
>10 s	0.91 (0.57, 1.29)	0.96 (0.72, 1.26)	0.05 (–0.33 to 0.42)	10.40 (–49.91 to 70.71)	0.14 (0.01, 0.43)	18.58 (1.03, 64.86)	0.49 (0.36, 0.61)[e]
10-30 s	0.83 (0.56, 1.23)	0.90 (0.71, 1.11)	0.07 (–0.30 to 0.43)	13.23 (–47.34 to 73.80)	0.14 (0.01, 0.42)	20.61 (1.28, 63.57)	0.41 (0.21, 0.57)[e]
>30 s	1.04 (0.60, 1.38)	1.06 (0.89, 1.31)	0.02 (–0.38 to 0.42)	6.05 (–53.63 to 65.73)	0.13 (0.01, 0.72)	15.46 (0.80, 119.91)	0.30 (0.04, 0.53)[e]
Number of steps							
All	49.77 (6.00, 266.00)	47.00 (5.00, 233.00)	–2.77 (–18.78 to 13.23)	–5.62 (–31.00 to 19.76)	3.41 (0.00, 10.00)	10.39 (0.00, 30.00)	0.99 (0.99, 1.00)[d]
>10 s	76.32 (13.00, 317.00)	72.10 (13.00, 306.00)	–4.22 (–24.21 to 15.77)	–5.80 (–25.88 to 14.28)	4.87 (0.00, 21.00)	8.61 (0.00, 25.81)	0.99 (0.98, 1.01)[d]
10-30 s	22.42 (13.00, 38.50)	20.62 (12.50, 37.00)	–1.80 (–7.52 to 3.92)	–6.87 (–30.79 to 17.06)	2.50 (0.00, 8.00)	10.98 (0.00, 29.89)	0.90 (0.79, 0.95)[c]
>30 s	159.26 (40.00, 479.00)	151.31 (36.00, 423.00)	–7.95 (–37.77 to 21.87)	–4.17 (–15.72 to 7.39)	8.51 (0.00, 56.00)	4.96 (0.00, 16.86)	0.99 (0.98, 1.00)[d]
Stride duration (s)							
All	1.41 (1.06, 1.88)	1.33 (1.03, 1.65)	–0.08 (–0.40 to 0.24)	–4.65 (–25.80 to 16.50)	0.12 (0.00, 0.41)	8.15 (0.15, 26.82)	0.70 (0.55, 0.79)[b]
>10 s	1.43 (1.09, 1.95)	1.33 (1.08, 1.65)	–0.10 (–0.41 to 0.22)	–5.68 (–24.00 to 12.64)	0.12 (0.00, 0.42)	7.51 (0.16, 24.25)	0.64 (0.41, 0.77)[b]

XSL•FO
RenderX

WB duration	RS, mean (5% quantile, 95% quantile)	SWD, mean (5% quantile, 95% quantile)	Error, mean (LoA)	MRE (%), mean (LoA)	MAE, mean (5% quantile, 95% quantile)	MARE (%), mean (5% quantile, 95% quantile)	ICC(2,1), mean (5% quantile, 95% quantile)
10-30 s	1.42 (1.10, 1.91)	1.32 (1.09, 1.63)	–0.10 (–0.45 to 0.24)	–6.13 (–27.08 to 14.82)	0.14 (0.01, 0.42)	8.94 (0.52, 25.02)	0.54 (0.29, 0.70)[b]
>30 s	1.43 (1.09, 2.07)	1.34 (1.08, 1.67)	–0.09 (–0.35 to 0.18)	–4.98 (–18.43 to 8.47)	0.09 (0.00, 0.42)	5.30 (0.04, 20.19)	0.77 (0.54, 0.88)[c]
Distance (m)							
All	25.75 (2.36, 125.62)	24.22 (2.44, 119.63)	–1.53 (–22.16 to 19.11)	4.91 (–53.08 to 62.90)	3.75 (0.06, 15.48)	19.79 (0.92, 65.29)	0.98 (0.97, 0.98)[d]
>10 s	40.08 (5.40, 203.72)	37.50 (5.04, 168.91)	–2.58 (–28.93 to 23.77)	3.14 (–54.18 to 60.46)	5.66 (0.07, 26.48)	17.72 (0.86, 58.11)	0.98 (0.97, 0.98)[d]
10-30 s	9.62 (4.84, 19.38)	9.87 (4.69, 19.41)	0.25 (–4.61 to 5.12)	5.34 (–56.44 to 67.13)	1.71 (0.05, 5.55)	19.53 (0.86, 57.10)	0.83 (0.74, 0.89)[c]
>30 s	86.93 (15.61, 253.43)	80.00 (20.80, 233.73)	–6.93 (–47.31 to 33.44)	–0.25 (–49.97 to 49.47)	11.74 (0.32, 59.36)	14.93 (0.78, 79.93)	0.96 (0.94, 0.98)[d]

[a]As reported in the work by Kirk et al [26].

[b]Moderate ICC level (0.50-0.75).

[c]Good ICC level (0.75-0.90).

[d]Excellent ICC level (>0.90).

[e]Poor ICC level (<0.50).

Physical Function

The reliability of the walking distance estimates was excellent for both SPPB groups and all WB duration categories (ICCs≥0.94) except for WBs of 10 to 30 seconds in the lower–SPPB score group (ICC=0.63), which showed moderate reliability (Table 4). Distance was overestimated in the lower SPPB–score group for all WBs (MRE=20.05%; MARE=26.83%) and WBs of >10 seconds (MRE=20.34%; MARE=26.3%), and the error was lower for longer WBs of >30 seconds (MRE=13.52%; MARE=20.43%) than for shorter WBs of 10 to 30 seconds (MRE=24.13%; MARE=29.56%). Conversely, the distance was underestimated in the higher–SPPB score group for all WBs (MRE=–6.37%; MARE=14.55%) and WBs of >10 seconds (MRE=–9.53%; MARE=11.39%), and the error was lower for WBs of >30 seconds (MRE=–8.85%; MARE=11.49%) than for WBs of 10 to 30 seconds (MRE=–10.03%; MARE=11.32%).

In the higher–SPPB score group, walking speed reliability was good for all WB duration categories (ICC≥0.85), with a slightly underestimated walking speed and MRE ranging from –1.73% to –3.59% (MARE reaching 8.62% to 11.15%). Conversely, the lower–SPPB score group showed poor walking speed reliability and higher errors for all WBs combined, WBs of >10 seconds, and WBs of 10 to 30 seconds, with ICCs ranging from 0.17 to 0.44, MREs reaching 26.58% to 32.56%, and MAREs reaching 31.22% to 35.71%. Notably, the lower–SPPB score group had a moderate walking speed reliability for WBs of >30 seconds (ICC=0.66), with an MRE of 18.96% and MARE of 23.14%.

XSL•FO
RenderX

Table 4. Digital mobility outcome estimates for the higher–Short Physical Performance Battery (SPPB) score (n=6) and lower–SPPB score (n=5) groups from the reference system (RS) and single wearable device (SWD), mean errors and mean relative errors (MREs) with limits of agreement (LoA), mean absolute errors (MAEs) and mean absolute relative errors (MAREs), and intraclass correlation coefficients (ICC).

WB[a] duration and SPPB group (score)	RS, mean (5% quantile, 95% quantile)	SWD, mean (5% quantile, 95% quantile)	Error, mean (LoA)	MRE (%), mean (LoA)	MAE, mean (5% quantile, 95% quantile)	MARE (%), mean (5% quantile, 95% quantile)	ICC(2,1), mean (5% quantile, 95% quantile)
Walking speed (m/s)							
Lower (0-7)							
All	0.59 (0.34, 0.91)	0.70 (0.43, 0.94)	0.12 (–0.19 to 0.42)	26.58 (–49.26 to 102.41)	0.15 (0.02, 0.39)	31.30 (2.76, 95.98)	0.44 (0.14, 0.63)[b]
>10 s	0.56 (0.34, 0.91)	0.68 (0.43, 0.92)	0.12 (–0.18 to 0.41)	27.70 (–54.84 to 110.25)	0.14 (0.02, 0.39)	31.22 (2.76, 95.98)	0.42 (0.01, 0.64)[b]
10-30 s	0.53 (0.32, 0.66)	0.67 (0.42, 0.87)	0.14 (–0.16 to 0.43)	32.56 (–54.61 to 119.72)	0.16 (0.02, 0.35)	35.71 (4.02, 95.98)	0.17 (0.00, 0.42)[b]
>30 s	0.63 (0.34, 1.04)	0.71 (0.44, 1.05)	0.08 (–0.21 to 0.38)	18.96 (–54.21 to 92.13)	0.11 (0.01, 0.44)	23.14 (0.77, 127.33)	0.66 (0.33, 0.85)[c]
Higher (8-12)							
All	0.70 (0.39, 1.05)	0.68 (0.48, 0.97)	–0.03 (–0.21 to 0.15)	–1.73 (–30.26 to 26.80)	0.08 (0.01, 0.19)	11.15 (1.24, 29.51)	0.86 (0.80, 0.90)[d]
>10 s	0.76 (0.49, 1.19)	0.72 (0.47, 1.08)	–0.04 (–0.20 to 0.12)	–3.49 (–23.62 to 16.65)	0.07 (0.01, 0.19)	8.82 (1.24, 22.25)	0.89 (0.80, 0.93)[d]
10-30 s	0.68 (0.44, 1.00)	0.65 (0.46, 0.89)	–0.03 (–0.18 to 0.12)	–3.41 (–23.29 to 16.47)	0.06 (0.01, 0.16)	8.62 (1.27, 22.25)	0.85 (0.73, 0.91)[d]
>30 s	0.86 (0.63, 1.19)	0.81 (0.65, 1.10)	–0.05 (–0.22 to 0.13)	–3.59 (–24.49 to 17.32)	0.08 (0.01, 0.21)	9.09 (1.24, 18.61)	0.87 (0.72, 0.93)[d]
Distance (m)							
Lower (0-7)							
All	22.80 (1.98, 125.62)	24.31 (2.30, 130.75)	1.50 (–9.72 to 12.72)	20.05 (–49.57 to 89.67)	3.10 (0.07, 15.48)	26.83 (0.93, 93.85)	0.99 (0.99, 0.99)[e]
>10 s	35.68 (5.55, 125.63)	37.85 (5.75, 145.84)	2.16 (–12.14 to 16.46)	20.34 (–52.06 to 92.75)	4.57 (0.15, 20.23)	26.30 (0.86, 90.93)	0.99 (0.98, 0.99)[e]
10-30 s	8.79 (4.31, 16.80)	10.66 (5.04, 16.80)	1.87 (–3.29 to 7.02)	24.13 (–50.53 to 98.80)	2.31 (0.07, 6.03)	29.56 (0.86, 90.93)	0.63 (0.29, 0.81)[c]
>30 s	84.09 (14.35, 237.84)	86.79 (27.33, 233.73)	2.70 (–20.70 to 26.10)	13.52 (–55.05 to 82.09)	8.65 (0.32, 26.48)	20.43 (0.33, 107.91)	0.99 (0.96, 0.99)[e]
Higher (8-12)							
All	27.94 (2.56, 139.68)	24.16 (2.49, 119.63)	–3.78 (–28.41 to 20.85)	–6.37 (–40.14 to 27.41)	4.24 (0.04, 20.05)	14.55 (0.87, 36.45)	0.97 (0.96, 0.98)[e]
>10 s	43.32 (5.40, 203.72)	37.25 (4.92, 168.91)	–6.07 (–36.94 to 24.80)	–9.53 (–30.31 to 11.24)	6.46 (0.06, 34.81)	11.39 (0.87, 25.97)	0.97 (0.95, 0.98)[e]
10-30 s	10.30 (5.29, 21.23)	9.23 (4.46, 22.02)	–1.07 (–3.66 to 1.53)	–10.03 (–30.75 to 10.69)	1.21 (0.04, 4.03)	11.32 (0.87, 28.68)	0.94 (0.80, 0.97)[e]
>30 s	88.71 (21.79, 253.43)	75.76 (20.80, 194.07)	–12.95 (–57.48 to 31.58)	–8.85 (–30.07 to 12.37)	13.67 (0.56, 59.36)	11.49 (0.92, 23.42)	0.96 (0.89, 0.98)[e]

[a]WB: walking bout.

[b]Poor ICC level (<0.50).

[c]Moderate ICC level (0.50-0.75).

[d]Good ICC level (0.75-0.90).

[e]Excellent ICC level (>0.90).

XSL•FO
RenderX

Time Since Surgery and Walking Aid Use

To explore whether time since surgery influenced DMO accuracy, we plotted the participants' walking speed and distance estimates from both systems sorted by the number of days since surgery, which ranged from 32 to 390. As shown in Figures 1 and 2, there was no evident pattern in DMO accuracy across the number of days since surgery for any of the WB duration categories. Similarly, no clear deviation was found for the 18% (2/11) of participants who used a walking aid.

Figure 1. Walking speed estimates from the reference system (RS) and single wearable device (SWD) for each participant for all walking bouts (WBs) combined (red bars), WBs of >10 seconds (blue bars), and WBs of >30 seconds (green bars). The bars indicate the median, and the error bars indicate the IQR. Participants are sorted by days since surgery (DSS). The total Short Physical Performance Battery (SPPB) scores for each participant are listed below the x-axis; scores in green indicate medium or high functioning (8-12), and scores in red indicate impaired activity of daily living functions (0-7). The 2 walking aid users are marked with an icon.

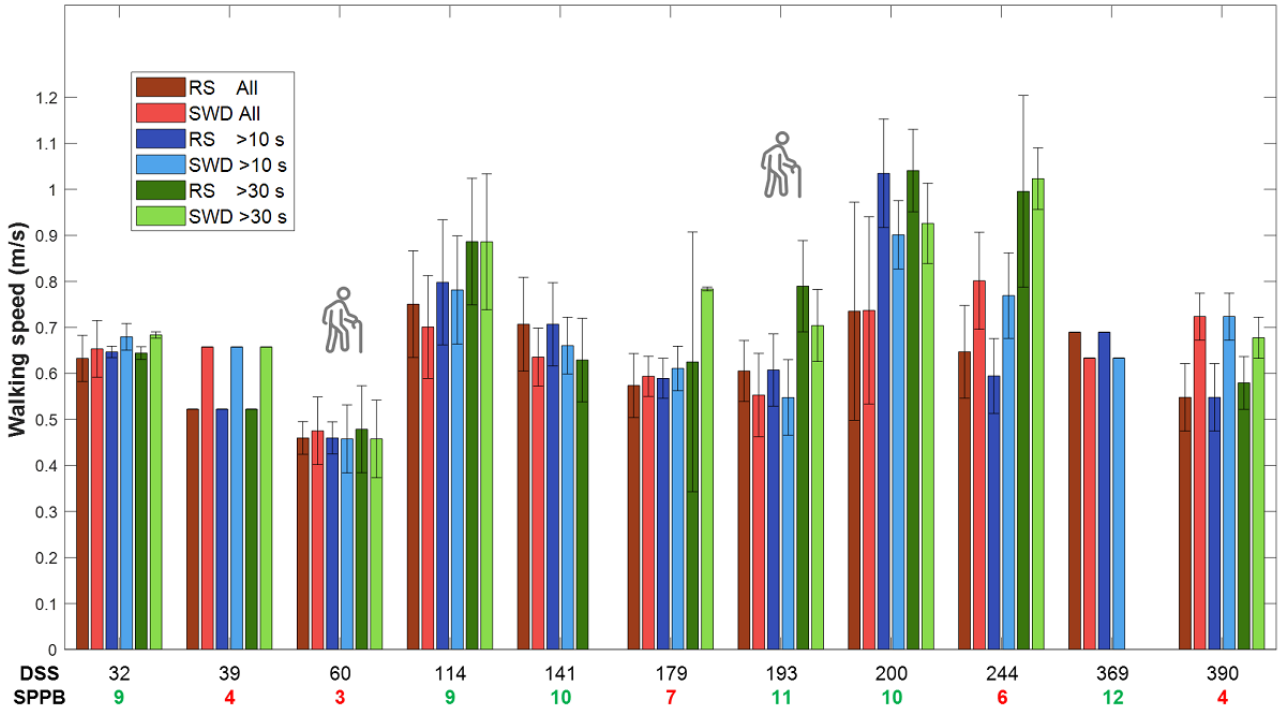

XSL•FO
RenderX

Figure 2. Mean relative error (MRE; A) and mean error (B) of the distance estimates from the single wearable device, and distance and median walking speed from the reference system (RS; C) for each participant for all walking bouts (WBs; red), WBs of >10 seconds (blue) and >30 seconds (green). Box plots (shown if ≥5 data points) display median, IQR, and 1.5×IQR whiskers. Participants are sorted by days since surgery (DSS). Total SPPB scores are shown below the x-axis (red: 0–7, green: 8–12). Walking aid users are marked with an icon.

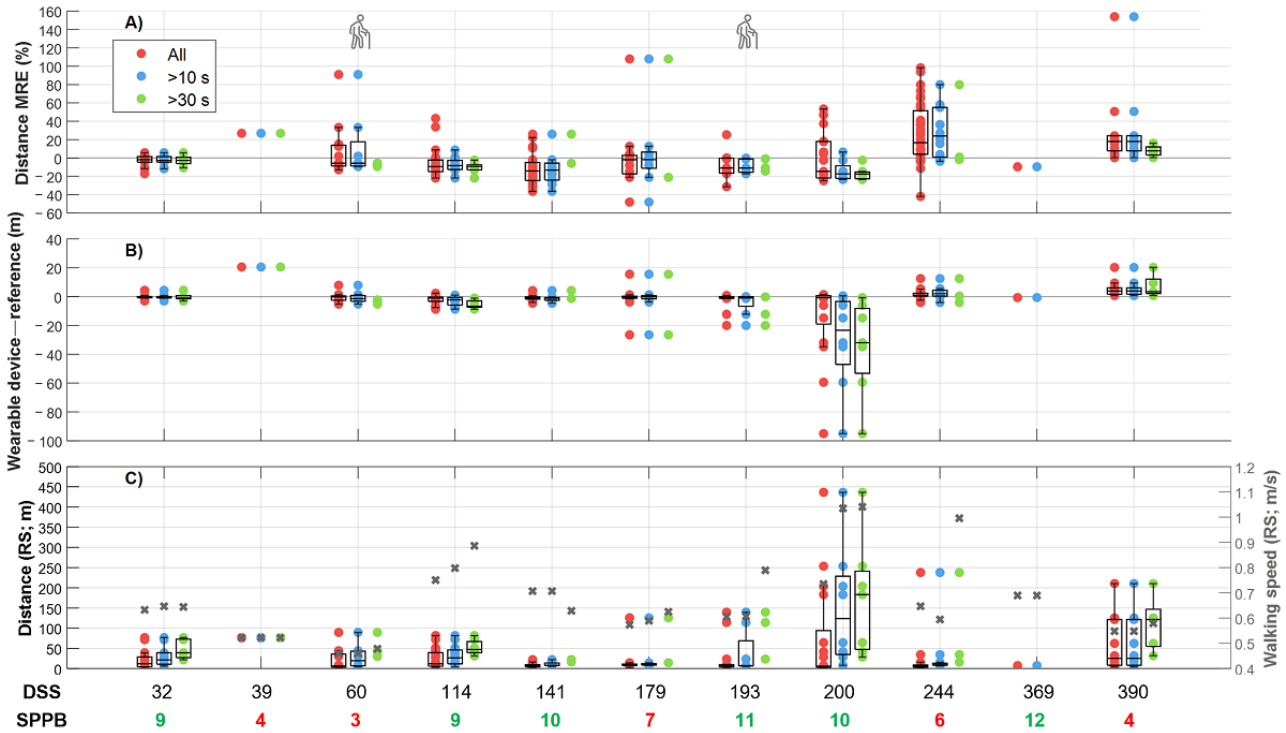

Discussion

Principal Findings

This study is the first to investigate critical hip fracture–specific factors that potentially influence the accuracy and reliability of validated DMOs, building on the rigorous methods and data from the Mobilise-D TVS [22,26,27]. We compared 6 real-world DMO estimates from a single wearable device against a validated multisensor reference system [24] and investigated factors that may affect the accuracy and reliability of validated DMOs in older adults after a hip fracture: WB duration, physical function, time since surgery, and walking aid use. Overall, 5 of the DMOs (walking speed, cadence, stride duration, number of steps, and distance) showed moderate to excellent accuracy and reliability across WB durations. The sixth DMO (stride length) showed poor accuracy and reliability in this hip fracture cohort. Furthermore, walking speed and distance were more accurate and reliable in participants with better physical function. We did not observe a discernible effect of time since surgery or walking aid use on the walking speed and distance estimates. These results indicate that real-world gait assessment is feasible in a hip fracture cohort and that gait characteristics can be estimated accurately and reliably from as early as 1 month after hip fracture surgery.

Impact of WB Duration

Previous Mobilise-D studies that validated DMOs across 6 cohorts showed generally lower algorithm performances when including very short WBs (<10 seconds) and for cohorts walking slowly and relying more on walking aids, such as many older adults after hip fracture [26,27]. As these previous studies did not perform cohort-specific analyses stratified by WB duration, this study further examined the hip fracture cohort following the true-positive evaluation of DMO accuracy and reliability at a WB level as outlined in the work by Kirk et al [26] and Micó-Amigo et al [27]. As expected, excluding shorter WBs increased the accuracy and reliability of walking speed, cadence, number of steps, stride duration, and distance, with WBs of >30 seconds having the highest accuracy and reliability. These lower errors are likely due to longer WBs capturing more continuous, consistent gait [38]. In addition, longer WB durations likely correspond to outdoor walks, which are generally not as slow and intermittent as indoor walking [38,39]. For short-duration WBs, contextual factors and turns might be more prominent [21], likely leading to higher DMO errors. However, as patients after a hip fracture typically walk in short bouts, excluding these would result in the loss of a potentially large amount of data, jeopardizing the representativeness of the data.

In addition to slow gait in short bouts, the complexity and variability of real-world environments further challenge the performance of algorithms. Factors such as deviations from a straight path, turns and obstacles, limited visibility, crowded areas, and other mobility tasks such as navigating stairs or slopes may have influenced the DMO estimates. These conditions make stride length estimation particularly challenging as this DMO is sensitive to intermittent gait [27]. Stride length exhibited larger errors and poor reliability for all WB durations, contributing to the errors in walking speed and distance estimates. This lower performance in stride length estimation could be related to the algorithms assuming an inverted pendulum, which might not properly fit the hip fracture cohort due to their asymmetrical gait [13]. Further work is necessary

XSL•FO
RenderX

to improve stride length estimation in older adults after hip fracture.

Impact of Physical Function

Our results show that walking speed can be estimated with good reliability in older adults with higher SPPB scores in all WB duration categories, with errors ranging from –3.59% to –1.73%. Conversely, the lower–SPPB score group exhibited larger errors, ranging from 26.58% to 32.56%, and demonstrated poor reliability for all WB duration categories except for WBs lasting >30 seconds, which showed moderate reliability and lower errors (18.96%). These group differences are likely due to the differences in participants' real-world walking speed. As reported in our previous work, the performance of all algorithms decreased for speeds of <0.5 m/s [27], a threshold distinguishing between slow and medium-speed walkers [47]. Those with lower SPPB scores walked slower on average than those with higher SPPB scores (0.59, SD 0.17 m/s vs 0.70, SD 0.20 m/s, respectively), as measured using the reference system. The lower–SPPB score group walked slower and had higher errors in 10- to 30-second WBs. Although 22% (36/164) of all WBs across participants were of <0.5 m/s, only 9% (1/11) of the participants had a median speed slower than this (Figure 1 and Table 1; participant 3). Notably, this participant was among those with the lowest errors in walking speed and distance estimates despite having an SPPB score of only 3. In general, errors in slow walkers may be due to lower amplitude in the acceleration signals, inconsistent gait cycles [48,49], irregular gait patterns [50,51], short step length, shuffling gait [8], and less symmetrical gait [13]. Nevertheless, DMOs in the lower–SPPB score group demonstrated better accuracy and moderate reliability in WBs of >30 seconds, indicating that accuracy improves during longer WBs irrespective of physical function.

Patients have individual preferences regarding whether walking faster or further is more important [46], and this often depends on the context. For instance, walking speed is essential for tasks such as crossing the road, whereas walking distance is more relevant for engaging in activities such as going to the grocery store or participating in the community. As most patients report walking further as more important [46], a reliable distance estimate would provide clinicians with valuable information that is meaningful to their patients. Overall, distance estimates showed good to excellent reliability and low errors, especially for the longest WBs. When stratified by SPPB score, distance exhibited the same error pattern as walking speed, with errors being larger in shorter WBs. Although longer WBs could be more informative as these likely represent more consistent outdoor walking, for many patients, life space after a hip fracture may be confined primarily to their home environment, where furniture, walls, and doors are likely to result in significantly shorter WBs.

Distance was overestimated in older adults with lower SPPB scores and underestimated in those with higher SPPB scores, showing large variations in errors despite excellent reliability. This could be explained by the combination of overestimated average walking speed and underestimated WB duration [27], the multiplication of which yielded the distance DMO. Furthermore, high errors in the lower–SPPB score group could be explained by an inclusive WB definition that allowed for brief pauses of up to 3 seconds [37]. Such brief pauses are included in the total duration of a WB but are not taken into account in the calculation of mean walking speed, leading to a potential overestimation of distance. Older adults with lower physical function after a hip fracture may be more likely to take brief pauses and make slower turns within a WB than those with higher physical function. Therefore, when interpreting walked distance, researchers and clinicians should focus on longer and more continuous WBs if possible.

Impact of Time Since Surgery and Walking Aids

Promisingly, our results showed no discernible change in the accuracy of walking speed and distance estimates across 32 to 390 days after hip fracture surgery, suggesting that these DMOs can be used to capture daily-life gait from as early as 1 month after surgery. As this study assessed DMO errors in the real world, we only included participants who were mobile in their home environment, which may have contributed to not having included patients within the first month after surgery. Therefore, we cannot conclude on the accuracy and reliability of DMOs in the acute phase after hip fracture. However, in the initial days and weeks following hip fracture surgery, when many patients are still in the hospital or rehabilitation center [6,7], clinical assessments may be more appropriate and informative than real-world gait monitoring. The primary focus during this period is typically first on pain management and early mobilization, subsequently followed by rehabilitation aimed at improving physical function and enabling independent walking to prepare patients for discharge to home environments [8,9]. Importantly, physical function usually improves the most during the first 3 months after hip fracture surgery [41], with DMO accuracy for participants assessed within this period showing no discernible deviation from that of participants assessed after the first 3 months. Similarly, DMO accuracy for the 18% (2/11) of participants who used walking aids did not deviate visibly from that of participants who did not. Although we found no evidence that days since surgery or the use of walking aids systematically affect DMO accuracy and reliability, these results in our small sample size should be confirmed in a larger sample.

General Discussion

In patients after hip fracture with reduced walking ability, shorter walking duration, or slower walking speed, additional mobility outcomes may provide relevant information about their movement characteristics. The core DMOs from the Mobilise-D TVS allow for a broader assessment of additional DMOs, such as walking duration and number of WBs exceeding different duration thresholds. Other relevant metrics in this population may include the number of turns, time spent in an upright position, and the frequency of posture transitions (eg, sitting to standing). In particular, the number of sit-to-stand transitions may be an important mobility outcome for patients after hip fracture, especially in the acute phase, and ongoing research is exploring this further.

While the lower back remains a viable location for gait monitoring in the daily lives of patients after hip fracture, it may not be optimal for long-term recordings. Extended

XSL•FO
RenderX

monitoring periods require sufficient battery life and may cause discomfort, particularly when sitting or lying down. In addition, adhesive attachments may irritate the skin, whereas belts used for fixation can introduce movement artifacts that compromise data quality. Other wearable device locations may help overcome some of these challenges. Although wrist-worn devices could be an option, wrist-based data are currently less accurate than lower back data [52] and have shown lower accuracy in patients recovering from hip fracture [53]. In patients with frailty and slow and cautious gait, a more suitable location for a single wearable device may be the foot or shank, which is better suited for detecting initial contact events than the lower back [54]. Furthermore, promising alternative approaches for gait monitoring in patients after hip fracture include instrumented walking aids [55] and even the prostheses themselves [56], a recent area of innovative development.

Recent comprehensive validation studies represent a significant step toward advancing real-world gait monitoring in multiple cohorts by providing robust reliability and accuracy data for DMOs [26,27]. Currently, clinically relevant accuracy thresholds have not yet been established for these DMOs in the hip fracture cohort, and they may also differ for different DMOs, making it challenging to define their appropriateness for clinical use. Our comprehensive and transparent reporting of accuracy and reliability measures allows readers to interpret the results in the context of their own research or clinical practice. Furthermore, based on the DMOs in this study, ongoing research is investigating the minimal important difference using the Mobilise-D clinical validation study dataset [57]. Further research is needed to establish clinically relevant thresholds to strengthen clinical applicability.

Clinical Implications

As most walking activity happens in daily life, outside a laboratory or clinic, understanding gait recovery after hip fracture surgery should not only rely on snapshot tests in the clinic or assessments in the laboratory but be complemented with accurate and reliable real-world monitoring. This study shows that real-world gait monitoring can provide much-needed information about a patient's functional recovery. Currently, the Mobilise-D consortium is undertaking a comprehensive clinical validation of all validated DMOs [57], including construct validity, predictive capacity, and the ability to detect change. Together with the encouraging results of this study, this opens up the study of real-world trajectories of gait recovery after a hip fracture and investigation of what characteristics of the patient, treatment, or rehabilitation may predict these trajectories.

This study showed that, for the hip fracture cohort, the number of steps and cadence DMOs are the most accurate and reliable, as well as suitable for both short and longer walking periods. These DMOs can also account for participants with functional constraints preventing them from performing outdoor walking. Furthermore, we found that walking speed, stride duration, and walking distance can be used to provide additional information beyond the most frequently studied DMOs, such as number of steps, cadence, and walking time [58-60]. In terms of walking speed, we recommend excluding the shortest WBs of <10 seconds to ensure accurate and reliable estimates without disregarding a disproportionate amount of data. Furthermore, clinicians should consider the patients' physical function when interpreting walking speed due to the overestimation in people with lower SPPB scores and underestimation in people with higher SPPB scores. Distance and stride duration showed promising results across all WB durations. However, in subgroups with lower SPPB scores (<8), WBs of >30 seconds are required for obtaining more accurate distance estimates, and this may be challenging for patients with the most frailty. We showed that the distance and walking speed DMOs can be used as early as 1 month after surgery also in people using a single cane or crutch, but more data on walking aid use are needed from a larger sample.

Strengths and Limitations

To the best of our knowledge, this is the first study to rigorously investigate the accuracy and reliability of real-world DMOs while considering critical hip fracture–related factors. A key strength is the sufficient number of 164 WBs [25], which improves the statistical power. However, the number of available WBs decreased substantially when excluding short WBs. Although this improved the accuracy and reliability of several DMOs, it also excluded a large part of participants' data, potentially jeopardizing their representativeness of daily life activities. Moreover, some participants, especially those with better physical function, contributed more WBs than others. This imbalance may have biased the analyses by overrepresenting the walking patterns of more mobile participants. Future studies could address this by accounting for within-subject variability and ensuring more balanced data collection across participants.

While our study sample included participants with a wide range of physical function and walking speeds, with SPPB scores ranging from 3 to 12 and 21.9% (36/164) of the WBs being of <0.5 m/s, our hip fracture sample was relatively well functioning compared to the general population after hip fracture, who may have SPPB scores ranging from 1 to 5 in the first year after surgery [30]. In addition, while we had enough WBs to compare sample by sample in the true-positive analysis [25], our number of participants included in the analyses was small. As a result, formal analyses of the potential effect of days since surgery and walking aid use on DMO accuracy and reliability were not feasible.

Several factors might explain the specifics of our study sample. First, we only included participants who were mobile in their home environment. Second, as we recruited during the pandemic, we were prohibited from recruiting participants during their hospital stay, which consequently excluded those in the acute phase. Third, the complexity and thoroughness of the protocol likely also limited the recruitment of patients with the most frailty. Fourth, while we initially recruited 19 participants from a target sample size of 20, the sample size was reduced to 11 (58%) in the final analysis, primarily due to technical errors with the reference system. However, 11% (2/19) of the participants were excluded due to insufficient signal quality, which may have been related to their low physical function (SPPB scores of 3), potentially accompanied by

XSL•FO
RenderX

shuffling or a very cautious gait, making it challenging to detect gait events. As a result, our findings may not fully generalize to patients with more severe impairments, and a larger sample is needed to provide stronger evidence. These limitations constrain our findings but offer a promising foundation for future research to improve the robustness and generalizability of DMOs in frail populations.

Conclusions

Considering specific factors critical for older adults after a hip fracture, our study identified 5 accurate and reliable real-world DMO estimates from a single wearable device worn on the lower back: walking speed, cadence, stride duration, number of steps, and distance. The accuracy and reliability of most DMOs improved when excluding WBs of <10 seconds and were higher for WBs of >30 seconds than for WBs of 10 to 30 seconds and for participants with higher physical function. DMOs can capture daily gait as early as 1 month after surgery also in people using walking aids. However, as most patients after hip fracture perform WBs of short duration, there is a trade-off between more accurate and reliable walking speed and distance estimates and the disregard of substantial amounts of data. Our results add more granularity to real-world gait assessments in populations with severe gait impairments. They have important implications for future research as they can provide a significant contribution to clinical validation studies, randomized controlled trials, and descriptive studies on gait recovery after hip fracture that is meaningful for clinicians and patients alike. Finally, these results support the use of these DMOs to assess intervention effects on real-world gait in detail, thereby aiding to the design of optimal care pathways.

Acknowledgments

This work was supported by the Mobilise-D project that has received funding from the Innovative Medicines Initiative (IMI) 2 Joint Undertaking under grant 820820. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation program and the European Federation of Pharmaceutical Industries and Associations. The authors thank the Mobilise-D technical validation study team of work package 2, all participants who devoted their time and energy to this taxing protocol in times of COVID-19, and Miriam Dinger at the Robert Bosch Foundation for Medical Research and Dr Tim Klüter at University Hospital Schleswig-Holstein for their indispensable help in recruiting patients. MAB acknowledges ChatGPT, an artificial intelligence model by OpenAI, for suggestions that helped improve the clarity of some sentences in this paper. SDD and LR were also supported by the Identifying Digital Endpoints to Assess Fatigue, Sleep, and Activities in Daily Living in Neurodegenerative Disorders and Immune-Mediated Inflammatory Diseases project that has received funding from the IMI 2 Joint Undertaking under grant agreement 853981. LR is a National Institute for Health and Care Research (NIHR) senior investigator. Both SDD and LR were supported by the NIHR Newcastle Biomedical Research Centre based at the Newcastle upon Tyne Hospitals National Health Service (NHS) Foundation Trust; Newcastle University; and the Cumbria, Northumberland, Tyne, and Wear NHS Foundation Trust. SDD and LR were also supported by the NIHR and Wellcome Trust Clinical Research Facility infrastructure at Newcastle upon Tyne Hospitals NHS Foundation Trust. SDD was supported by the UK Research and Innovation Engineering and Physical Sciences Research Council (grant reference EP/X031012/1 and EP/X036146/1). The Barcelona Institute for Global Health acknowledges support from grant CEX2018-000806-S funded by the Ministry of Science, Innovation, and Universities and the State Research Agency (MCIN/AEI/10.13039/501100011033) and support from the Government of Catalonia through the Research Centers of Catalonia program. The views expressed in this publication are those of the authors and do not necessarily reflect those of the IMI, the European Union, the European Federation of Pharmaceutical Industries and Associations, any associated partners, the NIHR, or the Department of Health and Social Care, who are not responsible for any use that may be made of the information contained herein.

Data Availability

The datasets generated or analyzed during the Mobilise-D technical validation study are available in the Zenodo repository [61].

Conflicts of Interest

SDD reports consultancy activity with Hoffmann-La Roche Ltd outside of this study. All other authors declare no other conflicts of interest.

References

1. Emmerson BR, Varacallo M, Inman D. Hip fracture overview. StatPearls. URL: https://www.ncbi.nlm.nih.gov/books/NBK557514/ [accessed 2024-04-29]
2. Downey C, Kelly M, Quinlan JF. Changing trends in the mortality rate at 1-year post hip fracture - a systematic review. World J Orthop 2019 Mar 18;10(3):166-175 [FREE Full text] [doi: 10.5312/wjo.v10.i3.166] [Medline: 30918799]
3. Dyer SM, Crotty M, Fairhall N, Magaziner J, Beaupre LA, Cameron ID, Fragility Fracture Network (FFN) Rehabilitation Research Special Interest Group. A critical review of the long-term disability outcomes following hip fracture. BMC Geriatr 2016 Sep 02;16(1):158 [FREE Full text] [doi: 10.1186/s12877-016-0332-0] [Medline: 27590604]

XSL•FO
RenderX

4. Delgado-Ortiz L, Polhemus A, Keogh A, Sutton N, Remmele W, Hansen C, et al. Listening to the patients' voice: a conceptual framework of the walking experience. Age Ageing 2023 Jan 08;52(1):afac233 [FREE Full text] [doi: 10.1093/ageing/afac233] [Medline: 36729471]
5. Wahlsten LR, Smedegaard L, Brorson S, Gislason G, Palm H. Living settings and cognitive impairment are stronger predictors of nursing home admission after hip fracture surgery than physical comorbidities: a nationwide Danish cohort study. Injury 2020 Oct;51(10):2289-2294. [doi: 10.1016/j.injury.2020.06.041] [Medline: 32622625]
6. Grigoryan KV, Javedan H, Rudolph JL. Orthogeriatric care models and outcomes in hip fracture patients: a systematic review and meta-analysis. J Orthop Trauma 2014 Mar;28(3):e49-e55 [FREE Full text] [doi: 10.1097/BOT.0b013e3182a5a045] [Medline: 23912859]
7. Werner M, Macke C, Gogol M, Krettek C, Liodakis E. Differences in hip fracture care in Europe: a systematic review of recent annual reports of hip fracture registries. Eur J Trauma Emerg Surg 2022 Jun 08;48(3):1625-1638 [FREE Full text] [doi: 10.1007/s00068-021-01797-8] [Medline: 34623474]
8. Nijdam TM, Laane DW, Schiepers TE, Smeeing DP, Kempen DH, Willems HC, et al. The goals of care in acute setting for geriatric patients in case of a hip fracture. Eur J Trauma Emerg Surg 2023 Aug 18;49(4):1835-1844 [FREE Full text] [doi: 10.1007/s00068-023-02258-0] [Medline: 36933048]
9. Reyes BJ, Mendelson DA, Mujahid N, Mears SC, Gleason L, Mangione KK, et al. Postacute management of older adults suffering an osteoporotic hip fracture: a consensus statement from the international geriatric fracture society. Geriatr Orthop Surg Rehabil 2020 Jul 16;11:2151459320935100 [FREE Full text] [doi: 10.1177/2151459320935100] [Medline: 32728485]
10. Taraldsen K, Thingstad P, Sletvold O, Saltvedt I, Lydersen S, Granat MH, et al. The long-term effect of being treated in a geriatric ward compared to an orthopaedic ward on six measures of free-living physical behavior 4 and 12 months after a hip fracture - a randomised controlled trial. BMC Geriatr 2015 Dec 04;15:160 [FREE Full text] [doi: 10.1186/s12877-015-0153-6] [Medline: 26637222]
11. Taylor NF, Peiris CL, Thompson AL, Prendergast LA, Harding KE, Hau R, et al. Association between physical activity and short-term physical function changes after hip fracture: an observational study. Physiother Res Int 2021 Jan;26(1):e1876. [doi: 10.1002/pri.1876] [Medline: 32918389]
12. Dargent-Molina P, Favier F, Grandjean H, Baudoin C, Schott AM, Hausherr E, et al. Fall-related factors and risk of hip fracture: the EPIDOS prospective study. Lancet 1996 Jul 20;348(9021):145-149. [doi: 10.1016/s0140-6736(96)01440-7] [Medline: 8684153]
13. Thingstad P, Egerton T, Ihlen EF, Taraldsen K, Moe-Nilssen R, Helbostad JL. Identification of gait domains and key gait variables following hip fracture. BMC Geriatr 2015 Nov 18;15(1):150 [FREE Full text] [doi: 10.1186/s12877-015-0147-4] [Medline: 26581616]
14. Hulleck AA, Menoth Mohan D, Abdallah N, El Rich M, Khalaf K. Present and future of gait assessment in clinical practice: towards the application of novel trends and technologies. Front Med Technol 2022;4:901331 [FREE Full text] [doi: 10.3389/fmedt.2022.901331] [Medline: 36590154]
15. Taraldsen K, Polhemus A, Engdal M, Jansen CP, Becker C, Brenner N, et al. Evaluation of mobility recovery after hip fracture: a scoping review of randomized controlled studies. Osteoporos Int 2024 Mar 06;35(2):203-215 [FREE Full text] [doi: 10.1007/s00198-023-06922-4] [Medline: 37801082]
16. Cappozzo A. Gait analysis methodology. Hum Mov Sci 1984 Mar;3(1-2):27-50. [doi: 10.1016/0167-9457(84)90004-6]
17. Warmerdam E, Hausdorff JM, Atrsaei A, Zhou Y, Mirelman A, Aminian K, et al. Long-term unsupervised mobility assessment in movement disorders. Lancet Neurol 2020 May;19(5):462-470. [doi: 10.1016/S1474-4422(19)30397-7] [Medline: 32059811]
18. Van Ancum JM, van Schooten KS, Jonkman NH, Huijben B, van Lummel RC, Meskers CG, et al. Gait speed assessed by a 4-m walk test is not representative of daily-life gait speed in community-dwelling adults. Maturitas 2019 Mar;121:28-34. [doi: 10.1016/j.maturitas.2018.12.008] [Medline: 30704562]
19. Hillel I, Gazit E, Nieuwboer A, Avanzino L, Rochester L, Cereatti A, et al. Is every-day walking in older adults more analogous to dual-task walking or to usual walking? Elucidating the gaps between gait performance in the lab and during 24/7 monitoring. Eur Rev Aging Phys Act 2019 May 3;16(1):6 [FREE Full text] [doi: 10.1186/s11556-019-0214-5] [Medline: 31073340]
20. Rochester L, Mazzà C, Mueller A, Caulfield B, McCarthy M, Becker C, et al. A roadmap to inform development, validation and approval of digital mobility outcomes: the Mobilise-D approach. Digit Biomark 2020 Nov 26;4(Suppl 1):13-27 [FREE Full text] [doi: 10.1159/000512513] [Medline: 33442578]
21. Rehman RZ, Klocke P, Hryniv S, Galna B, Rochester L, Del Din S, et al. Turning detection during gait: algorithm validation and influence of sensor location and turning characteristics in the classification of Parkinson's disease. Sensors (Basel) 2020 Sep 19;20(18):5377 [FREE Full text] [doi: 10.3390/s20185377] [Medline: 32961799]
22. Mazzà C, Alcock L, Aminian K, Becker C, Bertuletti S, Bonci T, et al. Technical validation of real-world monitoring of gait: a multicentric observational study. BMJ Open 2021 Dec 02;11(12):e050785 [FREE Full text] [doi: 10.1136/bmjopen-2021-050785] [Medline: 34857567]
23. Connecting digital mobility assessment to clinical outcomes for regulatory and clinical endorsement. Mobilise-D. URL: https://mobilise-d.eu/ [accessed 2024-08-23]

XSL•FO
RenderX

24. Salis F, Bertuletti S, Bonci T, Caruso M, Scott K, Alcock L, et al. A multi-sensor wearable system for the assessment of diseased gait in real-world conditions. Front Bioeng Biotechnol 2023 Apr 21;11:1143248 [FREE Full text] [doi: 10.3389/fbioe.2023.1143248] [Medline: 37214281]
25. Scott K, Bonci T, Salis F, Alcock L, Buckley E, Gazit E, Mobilise-D consortium. Design and validation of a multi-task, multi-context protocol for real-world gait simulation. J Neuroeng Rehabil 2022 Dec 16;19(1):141 [FREE Full text] [doi: 10.1186/s12984-022-01116-1] [Medline: 36522646]
26. Kirk C, Küderle A, Micó-Amigo ME, Bonci T, Paraschiv-Ionescu A, Ullrich M, Mobilise-D consortium. Mobilise-D insights to estimate real-world walking speed in multiple conditions with a wearable device. Sci Rep 2024 Jan 19;14(1):1754 [FREE Full text] [doi: 10.1038/s41598-024-51766-5] [Medline: 38243008]
27. Micó-Amigo ME, Bonci T, Paraschiv-Ionescu A, Ullrich M, Kirk C, Soltani A, Mobilise-D consortium. Assessing real-world gait with digital technology? Validation, insights and recommendations from the Mobilise-D consortium. J Neuroeng Rehabil 2023 Jun 14;20(1):78 [FREE Full text] [doi: 10.1186/s12984-023-01198-5] [Medline: 37316858]
28. Najafi B, Helbostad JL, Moe-Nilssen R, Zijlstra W, Aminian K. Does walking strategy in older people change as a function of walking distance? Gait Posture 2009 Mar;29(2):261-266. [doi: 10.1016/j.gaitpost.2008.09.002] [Medline: 18952435]
29. Kline Mangione K, Craik RL, Lopopolo R, Tomlinson JD, Brenneman SK. Predictors of gait speed in patients after hip fracture. Physiother Can 2008;60(1):10-18 [FREE Full text] [doi: 10.3138/physio/60/1/10] [Medline: 20145738]
30. Prestmo A, Hagen G, Sletvold O, Helbostad JL, Thingstad P, Taraldsen K, et al. Comprehensive geriatric care for patients with hip fractures: a prospective, randomised, controlled trial. Lancet 2015 Apr 25;385(9978):1623-1633. [doi: 10.1016/S0140-6736(14)62409-0] [Medline: 25662415]
31. Thomas S, Halbert J, Mackintosh S, Cameron ID, Kurrle S, Whitehead C, et al. Walking aid use after discharge following hip fracture is rarely reviewed and often inappropriate: an observational study. J Physiother 2010;56(4):267-272 [FREE Full text] [doi: 10.1016/s1836-9553(10)70010-2] [Medline: 21091417]
32. Bertuletti S, Della Croce U, Cereatti A. A wearable solution for accurate step detection based on the direct measurement of the inter-foot distance. J Biomech 2019 Mar 14;84:274-277. [doi: 10.1016/j.jbiomech.2018.12.039] [Medline: 30630626]
33. Nasreddine ZS, Phillips NA, Bédirian V, Charbonneau S, Whitehead V, Collin I, et al. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. J Am Geriatr Soc 2005 Apr;53(4):695-699. [doi: 10.1111/j.1532-5415.2005.53221.x] [Medline: 15817019]
34. Salis F, Bertuletti S, Bonci T, Della Croce U, Mazzà C, Cereatti A. A method for gait events detection based on low spatial resolution pressure insoles data. J Biomech 2021 Oct 11;127:110687. [doi: 10.1016/j.jbiomech.2021.110687] [Medline: 34455233]
35. Caruso M, Sabatini AM, Knaflitz M, Gazzoni M, Croce UD, Cereatti A. Orientation estimation through magneto-inertial sensor fusion: a heuristic approach for suboptimal parameters tuning. IEEE Sensors J 2021 Feb 1;21(3):3408-3419. [doi: 10.1109/jsen.2020.3024806]
36. mobilise-d / mobgap. GitHub. URL: https://github.com/mobilise-d/mobgap [accessed 2024-08-28]
37. Kluge F, Del Din S, Cereatti A, Gaßner H, Hansen C, Helbostad JL, Mobilise-D consortium. Consensus based framework for digital mobility monitoring. PLoS One 2021 Aug 20;16(8):e0256541 [FREE Full text] [doi: 10.1371/journal.pone.0256541] [Medline: 34415959]
38. Baroudi L, Barton K, Cain SM, Shorter KA. Classification of human walking context using a single-point accelerometer. Sci Rep 2024 Mar 06;14(1):3039 [FREE Full text] [doi: 10.1038/s41598-024-53143-8] [Medline: 38321039]
39. Schmitt AC, Baudendistel ST, Lipat AL, White TA, Raffegeau TE, Hass CJ. Walking indoors, outdoors, and on a treadmill: gait differences in healthy young and older adults. Gait Posture 2021 Oct;90:468-474. [doi: 10.1016/j.gaitpost.2021.09.197] [Medline: 34619613]
40. Guralnik JM, Simonsick EM, Ferrucci L, Glynn RJ, Berkman LF, Blazer DG, et al. A short physical performance battery assessing lower extremity function: association with self-reported disability and prediction of mortality and nursing home admission. J Gerontol 1994 Mar;49(2):M85-M94. [doi: 10.1093/geronj/49.2.m85] [Medline: 8126356]
41. Beckmann M, Bruun-Olsen V, Pripp AH, Bergland A, Smith T, Heiberg KE. Recovery and prediction of physical function 1 year following hip fracture. Physiother Res Int 2022 Jul 24;27(3):e1947 [FREE Full text] [doi: 10.1002/pri.1947] [Medline: 35332627]
42. Keppler AM, Holzschuh J, Pfeufer D, Gleich J, Neuerburg C, Kammerlander C, et al. Mobility improvement in the first 6 postoperative weeks in orthogeriatric fracture patients. Eur J Trauma Emerg Surg 2022 Aug;48(4):2867-2872 [FREE Full text] [doi: 10.1007/s00068-021-01856-0] [Medline: 34932124]
43. Giavarina D. Understanding Bland Altman analysis. Biochem Med (Zagreb) 2015;25(2):141-151 [FREE Full text] [doi: 10.11613/BM.2015.015] [Medline: 26110027]
44. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychol Methods 1996 Mar;1(1):30-46. [doi: 10.1037/1082-989X.1.1.30]
45. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016 Jun;15(2):155-163 [FREE Full text] [doi: 10.1016/j.jcm.2016.02.012] [Medline: 27330520]

XSL•FO
RenderX

46. Combs SA, Van Puymbroeck M, Altenburger PA, Miller KK, Dierks TA, Schmid AA. Is walking faster or walking farther more important to persons with chronic stroke? Disabil Rehabil 2013 May;35(10):860-867. [doi: 10.3109/09638288.2012.717575] [Medline: 23035811]
47. Studenski S, Perera S, Patel K, Rosano C, Faulkner K, Inzitari M, et al. Gait speed and survival in older adults. JAMA 2011 Jan 05;305(1):50-58 [FREE Full text] [doi: 10.1001/jama.2010.1923] [Medline: 21205966]
48. Hebenstreit F, Leibold A, Krinner S, Welsch G, Lochmann M, Eskofier BM. Effect of walking speed on gait sub phase durations. Hum Mov Sci 2015 Oct;43:118-124. [doi: 10.1016/j.humov.2015.07.009] [Medline: 26256534]
49. Quintero D, Lambert DJ, Villarreal DJ, Gregg RD. Real-time continuous gait phase and speed estimation from a single sensor. Control Technol Appl 2017 Aug;2017:847-852 [FREE Full text] [doi: 10.1109/CCTA.2017.8062565] [Medline: 30148285]
50. Soltani A, Aminian K, Mazza C, Cereatti A, Palmerini L, Bonci T, et al. Algorithms for walking speed estimation using a lower-back-worn inertial sensor: a cross-validation on speed ranges. IEEE Trans Neural Syst Rehabil Eng 2021;29:1955-1964. [doi: 10.1109/tnsre.2021.3111681]
51. Taraldsen K, Thingstad P, Døhl Ø, Follestad T, Helbostad JL, Lamb SE, et al. Short and long-term clinical effectiveness and cost-effectiveness of a late-phase community-based balance and gait exercise program following hip fracture. The EVA-Hip randomised controlled trial. PLoS One 2019;14(11):e0224971 [FREE Full text] [doi: 10.1371/journal.pone.0224971] [Medline: 31738792]
52. Kim DW, Hassett LM, Nguy V, Allen NE. A comparison of activity monitor data from devices worn on the wrist and the waist in people with Parkinson's disease. Mov Disord Clin Pract 2019 Nov;6(8):693-699 [FREE Full text] [doi: 10.1002/mdc3.12850] [Medline: 31745480]
53. Kluge F, Brand YE, Micó-Amigo ME, Bertuletti S, D'Ascanio I, Gazit E, et al. Real-world gait detection using a wrist-worn inertial sensor: validation study. JMIR Form Res 2024 May 01;8:e50035 [FREE Full text] [doi: 10.2196/50035] [Medline: 38691395]
54. Pacini Panebianco G, Bisi MC, Stagni R, Fantozzi S. Analysis of the performance of 17 algorithms from a systematic review: influence of sensor position, analysed variable and computational approach in gait timing estimation from IMU measurements. Gait Posture 2018 Oct;66:76-82. [doi: 10.1016/j.gaitpost.2018.08.025] [Medline: 30170137]
55. Fernandez-Carmona M, Ballesteros J, Díaz-Boladeras M, Parra-Llanas X, Urdiales C, Gómez-de-Gabriel JM. Walk-IT: an open-source modular low-cost smart rollator. Sensors (Basel) 2022 Mar 08;22(6):2086 [FREE Full text] [doi: 10.3390/s22062086] [Medline: 35336255]
56. Merle G, Parent-Harvey A, Harvey EJ. Sensors and digital medicine in orthopaedic surgery. OTA Int 2022 Apr;5(2 Suppl):e189 [FREE Full text] [doi: 10.1097/OI9.0000000000000189] [Medline: 35949264]
57. Mikolaizak AS, Rochester L, Maetzler W, Sharrack B, Demeyer H, Mazzà C, clinical validation study (WP4) on behalf of Mobilise-D consortium. Connecting real-world digital mobility assessment to clinical outcomes for regulatory and clinical endorsement-the Mobilise-D study protocol. PLoS One 2022;17(10):e0269615 [FREE Full text] [doi: 10.1371/journal.pone.0269615] [Medline: 36201476]
58. O'Halloran PD, Shields N, Blackstock F, Wintle E, Taylor NF. Motivational interviewing increases physical activity and self-efficacy in people living in the community after hip fracture: a randomized controlled trial. Clin Rehabil 2016 Nov;30(11):1108-1119. [doi: 10.1177/0269215515617814] [Medline: 26603892]
59. Senserrick C, Lawler K, Scroggie GD, Williams K, Taylor NF. Three short sessions of physiotherapy during rehabilitation after hip fracture were no more effective in improving mobility than a single longer session: a randomised controlled trial. Physiotherapy 2021 Sep;112:87-95. [doi: 10.1016/j.physio.2020.12.002] [Medline: 34052569]
60. Zusman EZ, Dawes M, Fleig L, McAllister MM, Cook WL, Guy P, et al. Older adults' sedentary behavior and physical activity after hip fracture: results from an outpatient rehabilitation randomized controlled trial. J Geriatr Phys Ther 2019;42(2):E32-E38. [doi: 10.1519/JPT.0000000000000193] [Medline: 30028352]
61. Küderle A. Mobilise-D Technical Validation Study (TVS) dataset. Zenodo. URL: https://zenodo.org/records/13899386 [accessed 2024-10-21]

Abbreviations

DMO: digital mobility outcome
ICC: intraclass correlation coefficient
IMU: inertial measurement unit
MARE: mean absolute relative error
MRE: mean relative error
SPPB: Short Physical Performance Battery
TVS: technical validation study
WB: walking bout

XSL•FO
RenderX

Edited by A Mavragani; submitted 23.10.24; peer-reviewed by B Cornish, S Lord; comments to author 03.03.25; revised version received 31.03.25; accepted 02.04.25; published 20.05.25.

Please cite as:
Berge MA, Paraschiv-Ionescu A, Kirk C, Küderle A, Micó-Amigo E, Becker C, Cereatti A, Del Din S, Engdal M, Garcia-Aymerich J, Grønvik KB, Hansen C, Hausdorff JM, Helbostad JL, Jansen CP, Johnsen LG, Klenk J, Koch S, Maetzler W, Megaritis D, Müller A, Rochester L, Schwickert L, Taraldsen K, Vereijken B
Evaluating the Accuracy and Reliability of Real-World Digital Mobility Outcomes in Older Adults After Hip Fracture: Cross-Sectional Observational Study
JMIR Form Res 2025;9:e67792
URL: https://formative.jmir.org/2025/1/e67792
doi:10.2196/67792
PMID:

©Martin A Berge, Anisoara Paraschiv-Ionescu, Cameron Kirk, Arne Küderle, Encarna Micó-Amigo, Clemens Becker, Andrea Cereatti, Silvia Del Din, Monika Engdal, Judith Garcia-Aymerich, Karoline B Grønvik, Clint Hansen, Jeffrey M Hausdorff, Jorunn L Helbostad, Carl-Philipp Jansen, Lars Gunnar Johnsen, Jochen Klenk, Sarah Koch, Walter Maetzler, Dimitrios Megaritis, Arne Müller, Lynn Rochester, Lars Schwickert, Kristin Taraldsen, Beatrix Vereijken. Originally published in JMIR Formative Research (https://formative.jmir.org), 20.05.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Public Health Messaging About Dengue on Facebook in Singapore During the COVID-19 Pandemic: Content Analysis

Shirley S Ho[1], PhD; Mengxue Ou[1], MA; Nova Mengxia Huang[1], MA; Agnes SF Chuah[1], PhD; Vanessa S Ho[1], BA; Sonny Rosenthal[2], PhD; Hye Kyung Kim[1], PhD

[1]Wee Kim Wee School of Communication and Information, Nanyang Technological University, Singapore, Singapore

[2]College of Integrative Studies, Singapore Management University, Singapore, Singapore

Corresponding Author:
Shirley S Ho, PhD
Wee Kim Wee School of Communication and Information
Nanyang Technological University
WKWSCI Building, 31 Nanyang Link
Singapore, 637718
Singapore
Phone: 65 65138667
Email: tsyho@ntu.edu.sg

Abstract

Background: Dengue, a mosquito-borne disease, has been a health challenge in Singapore for decades. In 2020, during the COVID-19 pandemic, Singapore encountered a serious dengue outbreak and deployed various communication strategies to raise public awareness and mitigate dengue transmission.

Objective: Drawing on the Crisis and Emergency Risk Communication (CERC) framework, this study examines how dengue-related messages communicated on Facebook (Meta) during the COVID-19 pandemic fall into the CERC themes. This study also seeks to understand how these themes differ between dengue outbreak (eg, 2020) and nonoutbreak years (eg, 2021). In addition, we explore how message themes on dengue changed across different CERC phases within the dengue outbreak year.

Methods: We conducted a content analysis on 314 Facebook posts published by public health authorities in Singapore between January 1, 2020, and September 30, 2022. We conducted chi-square tests to examine the differences in message themes between the dengue outbreak and nonoutbreak years. We also conducted chi-square tests to examine how these message themes varied across 3 CERC phases during the dengue outbreak year.

Results: Our findings suggest that during the dual epidemics of dengue and COVID-19, Singapore's public health communication on dengue largely adhered to CERC principles. Dengue-related messaging, particularly regarding intelligence and requests for contributions, significantly varied between outbreak and nonoutbreak years. In addition, messages on general advisories and vigilance, as well as those on social and common responsibility, significantly differed across the CERC phases during the dengue outbreak year.

Conclusions: Singapore's public health authorities flexibly adjusted their messaging strategies on social media platforms in response to the evolving dengue situation during the COVID-19 pandemic, demonstrating the high adaptability of the government's health communication amid the dual epidemics. However, several areas for improvement should also be noted for future public health communication to mitigate dengue transmission.

(JMIR Form Res 2025;9:e66954) doi:10.2196/66954

KEYWORDS

dengue; environment and public health; social media; Singapore; crisis and emergency risk communication; media; mosquito-borne; viruses; dengue outbreaks; epidemics; dengue transmission

XSL•FO
RenderX

Introduction

Background

Dengue, a mosquito-borne virus, is prevalent in tropical regions, with Asia accounting for 70% of cases [1]. Singapore, located in Southeast Asia, is particularly vulnerable, experiencing cyclical dengue outbreaks and remaining hyperendemic for the past 3 decades [2]. Research suggests a link between dengue and COVID-19 incidence in Asia [3], raising concerns about a dual epidemic [4,5]. Singapore experienced its most significant dengue outbreak in 2020, peaking during the COVID-19 circuit breaker period when restrictions were placed on nonessential activities [4]. In response, the government implemented various communication strategies to combat dengue during the pandemic. This study examines Singapore's public health messaging on dengue amid the dual epidemic.

Despite the prevalence of dengue in Singapore due to reasons including environmental factors and reduced immunity levels [6], the country has managed to keep massive outbreaks under control through continuous enhancements of dengue management strategies. This is evident from the low dengue seroprevalence and consistently low force of infection rate (ie, the rate at which susceptible individuals get infected) [7] which has seen a 10-fold decline since the 1960s when strategies to manage dengue were first introduced [6].

Singapore has implemented various strategies to combat dengue, including Project Wolbachia-Singapore, which releases Wolbachia-carrying male mosquitoes to reduce mosquito populations [8]. While government efforts are crucial, public involvement is essential. To engage the community, institutions like the Ministry of Health (MOH) and National Environment Agency (NEA) tend to use Facebook, Singapore's second most popular social media platform, to disseminate health crisis–related information [9]. However, the specific strategies used remain unclear. Understanding how governments use social media for health communication can improve issue management and enhance public adherence [10,11]. Given Singapore's success in reducing dengue [6], studying its use of Facebook during the dual epidemic can provide valuable insights for other nations developing health communication strategies.

Drawing on the Crisis and Emergency Risk Communication (CERC) framework [12], this study has 3 objectives. The CERC framework proposes that communication during a health crisis can be divided into different phases and the communication messages across phases can be categorized into several main themes and subthemes. Specifically, this study first examines how dengue-related messages communicated on Facebook during the COVID-19 pandemic fall into the CERC themes. Second, it examines how these themes differ between the dengue outbreak (ie, 2020) and nonoutbreak years (ie, 2021) during the pandemic. Third, it also explores how message themes on dengue changed across different CERC phases within the dengue outbreak year (ie, 2020). This study contributes to the CERC literature by applying the CERC framework in a unique context—an ongoing hyperendemic disease during a pandemic. Practically, this study can help inform relevant authorities in developing effective public health messaging strategies to navigate the challenges posed by concurrent epidemics.

The CERC Framework

This study applies the CERC framework to analyze how the Singapore government communicated dengue risks during the COVID-19 pandemic. The CERC framework [12] is well suited for this purpose, as it structures crisis communication into distinct phases, allowing strategies to adapt to evolving health crises. This framework operates on the fundamental assumption that most crises follow a largely predictable and replicable trajectory—starting with potential risks, escalating into an active crisis, and progressing toward recovery and eventual reflection or evaluation [12]. By organizing communication into clearly defined phases, the CERC framework enables risk communicators to adopt a proactive rather than reactive approach to crises, ensuring that messaging strategies are tailored to each phase for more effective communication and public response. Specifically, the CERC framework outlines five stages for tailoring communication during a health crisis: (1) precrisis stage, focusing on preparation and message development; (2) initial stage, providing immediate information to minimize risk [4]; (3) maintenance stage, offering updates and clarifications [13]; (4) resolution stage, emphasizing recovery and protective behaviors [4]; and (5) evaluation stage, reviewing postcrisis communication efforts [4,12].

The CERC framework has guided content analyses in health crises like COVID-19 [13], Zika [9], and Middle East Respiratory Syndrome [14]. Malik et al [13] expanded CERC during the COVID-19 pandemic by categorizing public communication into four themes: (1) Risk and Crisis Information, (2) Self-Efficacy and Sense-Making, (3) Preparation and Uncertainty Reduction, and (4) Advisories and Alerts. These themes include subthemes like epidemic intelligence, personal preventive measures, social responsibility, clarification, and general advisories. Building on these classifications, Malik et al [13] organize Singapore's dengue-related Facebook content into 4 main themes and 11 subthemes, as outlined in Table 1.

XSL•FO
RenderX

Table 1. Definitions of Crisis and Emergency Risk Communication (CERC) message themes and examples.

CERC themes and subthemes	Definitions
Risk and crisis information	
Pandemic intelligence	Messages disseminating generic, basic-level information (such as statements or numbers) about the disease, to highlight or raise awareness of the disease or the current situation [13]. In the case of dengue, this could be before, during, or after the dengue outbreak.
Self-efficacy and sense-making	
Personal preventive measures and mitigation	Messages that contain measures or precautions that can be taken by an individual to protect themselves from infection, or mitigation of disease-related issues [13]. This includes mental and physical health and wellness and includes measures taken after testing or contracting the disease to prevent further transmission.
Social and common responsibility	Messages highlighting the measures or precautions that can be taken by an individual or community to prevent the spread of the disease, for the greater good of society [13].
Inquisitive messaging	Messages that address public queries about issues related to the disease [13].
Preparations and uncertainty reduction	
Clarification	Messages intended to alert about or dispel myths, fake news, or misinformation about the disease [13].
Events, campaigns, and activities	Messages promoting communication campaigns, events or activities for awareness, relief, or treatment of the disease [13].
Request for contributions	Messages with a call-to-action seeking financial and voluntary contributions for tackling the disease [13].
Showing gratitude	Messages expressing thanks, approval, regards, reassurance, and paying tribute to the frontline workers (eg, doctors, nurses, cleaners, volunteers, etc) [13].
Reassurance	Messages that calm the public and remove their fears of the disease [13].
Advisories and alerts	
Risk groups	Messages regarding people who are at greater risk of contracting the disease or experiencing the negative consequences of contracting the disease. Such messaging involves all warnings or advice pertaining to risk groups, whether it is directly to them or to the people around them [13].
General advisories and vigilance	Messages in the form of alerts, tips, or cautions to help the public and entities in responding to the disease in certain situations, such as travel and workplace [13]. These can be presented in the form of announcements, such as the implementation of or changes in rules and regulations.

Public Messaging About Dengue

During the COVID-19 pandemic, dengue re-emerged as a significant health threat, creating a dual epidemic in countries like Singapore, Peru, and Bangladesh [4,5]. Despite evidence linking COVID-19 and dengue in Asia [15], little research has examined how health authorities communicated dengue risks during this dual epidemic. In 2020, Singapore experienced a major dengue outbreak, with a total of 35,315 dengue cases reported [16]. Yet, studies have not explored how Singaporean authorities communicated dengue risks and preventive strategies to the public during the COVID-19 pandemic, or how these messages aligned with the CERC framework. This study aims to fill that gap by analyzing how public messaging on dengue from Singapore's health authorities, particularly on Facebook, aligns with the themes proposed by the CERC framework.

The CERC framework has been widely applied to single emergencies, including the California measles outbreak [17], the COVID-19 pandemic [18], and natural disasters such as hurricanes [19]. Research on these cases has shown that health authorities' communication strategies typically evolve across the phases outlined in the CERC framework, from the initial phase to resolution. Since these emergencies (eg, COVID-19 or hurricanes) have a clear and definitive endpoint, prior studies have primarily focused on examining CERC communication strategies throughout the full crisis progression, from the initial phase through to resolution.

Beyond acute emergencies, researchers have also applied the CERC framework to hyperendemic diseases, such as HIV outbreaks in Africa [20] and tuberculosis outbreaks [21]. Unlike single-crisis events, these diseases recur cyclically and lack a clear resolution, presenting unique challenges for risk communication. Previous studies have primarily focused on how CERC-based crisis communication influences behavior change [20] and its effects on public outrage during crises [21], rather than systematically comparing risk communication strategies across different CERC phases. To the best of our knowledge, existing research has yet to examine the alignment between health communication strategies and the communication principles outlined in the CERC model across different phases of hyperendemic health crises like dengue. Understanding these communication strategies would provide critical insights for improving risk messaging during hyperendemic outbreaks and contribute to refining the CERC model to better address the complexities of such recurring public health challenges.

In tropical regions like Singapore, dengue is hyperendemic, characterized by continuous transmission with intermittent significant outbreaks [6]. While dengue cases decline during nonoutbreak periods, the disease persists year-round, lacking a

XSL•FO
RenderX

clear resolution or endpoint. Consequently, risk communication for hyperendemic diseases extends beyond the initial outbreak, continuing through the Maintenance to Evaluation phases [6]. Given the recurring nature of such diseases, communication strategies primarily vary across the precrisis, initial, and postcrisis phases, with the postcrisis phase incorporating elements of the maintenance, resolution, and evaluation phases outlined in the CERC framework [9]. This study, therefore, examines Singapore's public health communication strategies during these 3 critical phases—precrisis, initial, and postcrisis—and analyzes how dengue-related messaging was conveyed across these phases during outbreaks.

In addition, the application of the CERC model to analyzing communication strategies during a dual epidemic, such as the concurrent outbreaks of COVID-19 [5], is lacking. Given dengue's recurring nature and the unique challenges posed by dual epidemics [22], public health preparedness is crucial for early detection and timely response in such situations [2]. While intensified risk communication is essential during outbreaks, understanding communication strategies during nonoutbreak periods is equally important for establishing continuous, preventive approaches tailored to hyperendemic diseases like dengue [2,23]. These strategies include developing preemptive plans to enhance outbreak preparedness and facilitating swift, adaptive transitions when outbreaks occur [2,23].

By analyzing Singapore's public health messaging on dengue across both outbreak and nonoutbreak periods, this study seeks to uncover how CERC strategies can be adapted to varying dengue scenarios. The findings from this study will provide valuable insights into optimizing health communication frameworks for managing both persistent and emergent public health threats, ultimately enhancing long-term disease control and crisis preparedness [2,23].

Taken together, we propose the following research questions (RQ):

- RQ1: How does Singapore's public messaging about dengue align with the communication themes proposed by the CERC framework?
- RQ2: How do the CERC-themed messages in Singapore's public health communication about dengue vary across dengue outbreak (ie, 2020) and nonoutbreak (ie, 2021) years?
- RQ3: How do the messages in Singapore's public health communication about dengue vary across different CERC phases (precrisis, initial, and postcrisis) during the dengue outbreak years (ie, 2020)?

Methods

Overview

To achieve our objectives, we used a deductive approach in our analysis. We conducted a quantitative content analysis of Facebook messages posted by different Singapore government institutions that were involved in public health communication on dengue. These institutions include Gov.sg, MOH, Ministry of Sustainability and the Environment, NEA, and Health Promotion Board. This study focused on messages posted on Facebook, a social media platform commonly used by Singapore government institutions to communicate with the general public.

Ethical Considerations

Before collecting data, we received ethical approval from the institutional review board (IRB) at Nanyang Technological University (IRB-2022-725) in exempt category 4. Our study involved analyzing secondary data from publicly available social media accounts, which met the exemption criteria specified under the exempt category 4. Obtaining IRB approval ensured that this study followed ethical guidelines, safeguarding the privacy and rights of individuals whose data were being analyzed.

Data Collection and Sampling

Using the keywords "Dengue," "dengue," "Mozzie," "mozzie," "mosquito," "breed," "wolbachia," "Wolbachia," "aedes," "BLOCK," "B-L-O-C-K," "SAW," "S-A-W," "#mozziewipeout," "#Mozziewipeout," "repellent," "insecticide," "vase," "pails," "pot," and "soil," we crawled Facebook posts of the 5 Singapore governmental institutions. We used the Python Web Crawler to crawl the Facebook posts, starting from January 1, 2020. The data collection concluded on September 30, 2022, as we needed to halt data gathering and proceed to the next phase. A total of 560 Facebook posts were collated during this period. We cleaned the data manually by removing Facebook posts that were unrelated to public health communication about dengue and the duplicates, leaving out 378 Facebook posts. Facebook posts that were excluded are posts that solely focus on the call-out to subscribe for updates, mentions of dengue as a time frame where other activities or programs are the main topic, not dengue-focused, speeches by public figures, and press releases. The data were then randomly sampled with a confidence level of 99% and a 3% margin of error, resulting in a sample of 314 Facebook posts for final analysis.

Units of Analysis

Each Facebook post, including all text and visual elements, and everything visible on the web pages, were coded as one unit of analysis, respectively.

Developing the Codebook

Prior to data coding, we created a codebook (see Codebook for communication strategies from the government in Multimedia Appendix 1) based on the CERC message themes adapted from existing literature [13]. The message themes included are (1) pandemic intelligence, (2) personal preventive measures and mitigation, (3) social and common responsibility, (4) inquisitive messaging, (5) clarification, (6) events, campaigns, and activities, (7) request for contributions, (8) showing gratitude, (9) reassurance, (10) risk groups, and (11) general advisories and vigilance. The definitions and examples of each code were included in the full codebook in Multimedia Appendix 1.

Intercoder Reliability

We recruited 3 coders to code the data. The research team trained the coders before the actual coding. During the training sessions, coders were given the same units of data to code. After coding the data, they discussed their coding process to ensure

XSL•FO
RenderX

a common understanding of the codebook. Once consensus was achieved, the coders received 10% of the data to code, which we used to estimate the intercoder reliability. This process was repeated twice until we achieved an average Krippendorff α value of 0.72, which is an acceptable intercoder reliability that is greater than 0.70 [24]. Subsequently, the data were split equally and coded by the coders.

Results

Our analysis (Table 2) showed that most of the messages about dengue posted on Facebook were communicated by NEA (288/314, 91.7%), followed by the Ministry of Sustainability and the Environment (15/314, 4.8%), Gov.sg (8/314, 2.6%), Health Promotion Board (2/314, 0.6%), and MOH (1/314, 0.3%).

For RQ1, our results (Table 3) showed that most of the message themes communicated by the Singapore government were aligned with the communicated themes proposed by the CERC framework, except for inquisitive messaging (n=0). Events, campaigns, and activities were the most frequently communicated message theme (n=267), followed by personal preventive measures and mitigation (n=261), pandemic intelligence (n=231), general advisories and vigilance (n=59), social and common responsibility (n=50), and risk groups (n=16). Message themes that were less frequently communicated were clarification (n=8), showing gratitude (n=5), reassurance (n=4), and request for contributions (n=2).

Table 2. Distribution of Facebook posts by Singapore government institutions (N=314).

Institutions	Facebook posts, n (%)			
	2020	2021	2022	Total
Gov.sg	1 (0.6)	2 (2.7)	5 (6.1)	8 (2.6)
Ministry of Health	0 (0.0)	0 (0.0)	1 (1.2)	1 (0.3)
Ministry of Sustainability and Environment	11 (6.9)	1 (1.4)	3 (3.7)	15 (4.8)
National Environmental Agency	147 (92.5)	69 (94.5)	72 (87.8)	288 (91.7)
Health Promotion Board	0 (0.0)	1 (1.4)	1 (1.2)	2 (0.6)
Total	159 (100)	73 (100)	82 (100)	314 (100)

Table 3. Distribution of Facebook posts based on Crisis and Emergency Risk Communication themes and Singapore government institutions (N=314).

CERC[a] Message themes	Gov.sg	MOH[b]	MSE[c]	NEA[d]	HPB[e]	Total
Risk and crisis information, n						
Pandemic intelligence	5	1	14	210	1	231
Self-efficacy and sense-making, n						
Personal preventive measures and mitigation	5	1	13	240	2	261
Social and common responsibility	1	0	2	46	1	50
Inquisitive messaging	0	0	0	0	0	0
Preparations and uncertainty reduction, n						
Clarification	0	0	1	7	0	8
Events, campaigns, and activities	5	0	14	246	2	267
Request for contributions	0	0	0	2	0	2
Showing gratitude	0	0	0	5	0	5
Reassurance	0	0	1	3	0	4
Advisories and alerts, n						
Risk groups	0	0	0	16	0	16
General advisories and vigilance	2	1	2	54	0	59

[a]CERC: Crisis and Emergency Risk Communication.

[b]MOH: Ministry of Health.

[c]MSE: Ministry of Sustainability and the Environment.

[d]NEA, National Environment Agency.

[e]HPB: Health Promotion Board.

XSL•FO
RenderX

To answer RQ2, we conducted chi-square tests to examine the differences in message themes between dengue outbreak (ie, 2020) and nonoutbreak years (ie, 2021). We limited our comparison to 2020 and 2021, excluding 2022, as we only had 9 months of data for that year, making it less reliable for comparison.

The results of chi-square tests revealed that the number of messages on general advisories and vigilance was significantly different between 2020 (35/159, 22%) and 2021 (4/73, 5.5%; χ^2_1=9.8; *P*=.002). General advisories and vigilance messages were more frequently posted on Facebook during the dengue outbreak year. The number of messages on request for contributions also significantly varied between 2020 (0/159, 0%) and 2021 (2/73, 2.7%; χ^2_1=4.4; *P*=.04; Table 4). Furthermore, the differences in the number of messages on personal preventive measures and mitigation between 2020 and 2021 were almost significant (χ^2_1=3.7; *P*=.05), with the number of messages related to this theme being 138/159 (86.8%) in 2020. This was twice the number of messages (56/73, 76.7%) focusing on this theme in 2021.

For RQ3, we conducted chi-square tests to examine how CERC message themes vary across three CERC phases—precrisis, initial, and postcrisis—in 2020 (ie, the dengue outbreak year). As shown in Table 5, message themes on pandemic intelligence (χ^2_2=6.1; *P*=.047) and social and common responsibility (χ^2_2=6.1; *P*=.047) change across the CERC phases. Specifically, messages on pandemic intelligence were frequently posted on Facebook during the initial (47/69, 68.1%) and postcrisis phases (45/52, 86.5%), while messages on social and common responsibility were often posted during the initial phase (17/69, 24.6%).

Table 4. Comparison of Crisis and Emergency Risk Communication (CERC) message themes posted between 2020 (outbreak year) and 2021 (nonoutbreak year) (N=314).

CERC Message themes	2020, n (%)		2021, n (%)		Chi-square (*df*)	*P* value
	Yes	No	Yes	No		
Risk and crisis information						
Pandemic intelligence	118 (74.2)	41 (25.8)	47 (64.4)	26 (35.6)	2.4 (1)	.13
Self-efficacy and sense-making						
Personal preventive measures and mitigation	138 (86.8)	21 (13.2)	56 (76.7)	17 (23.3)	3.7 (1)	.05
Social and common responsibility	26 (16.4)	133 (83.6)	6 (8.2)	67 (91.8)	2.8 (1)	.10
Inquisitive messaging	0 (0)	159 (100)	0 (0)	73 (100)	—[a]	—
Preparations and uncertainty reduction						
Clarification	6 (3.8)	153 (96.2)	2 (2.7)	71 (97.3)	0.2 (1)	.69
Events, campaigns, and activities	133 (83.6)	26 (16.4)	63 (86.3)	10 (13.7)	0.3 (1)	.60
Request for contributions	0 (0)	159 (100)	2 (2.7)	71 (97.3)	4.4 (1)	.04
Showing gratitude	1 (0.6)	158 (99.4)	3 (4.1)	70 (95.9)	3.6 (1)	.06
Reassurance	1 (0.6)	158 (99.4)	3 (4.1)	70 (95.9)	3.6 (1)	.06
Advisories and alerts						
Risk groups	11 (6.9)	148 (93.1)	1 (1.4)	72 (98.6)	3.1 (1)	.08
General advisories and vigilance	35 (22.0)	124 (78.0)	4 (5.5)	69 (94.5)	9.8 (1)	<.001

[a]Not available.

XSL•FO
RenderX

Table 5. Message themes across different Crisis and Emergency Risk Communication (CERC) phases in 2020 (N=314).

CERC Message themes	Precrisis, n (%)		Initial, n (%)		Postcrisis, n (%)		Chi-square (*df*)	*P* value
	Yes	No	Yes	No	Yes	No		
Risk and crisis information								
Pandemic intelligence	26 (68.4)	12 (31.6)	47 (68.1)	22 (31.9)	45 (86.5)	7 (13.5)	6.1 (2)	<.05
Self-efficacy and sense-making								
Personal preventive measures and mitigation	35 (92.1)	3 (7.9)	56 (81.2)	13 (18.8)	47 (90.4)	5 (9.6)	3.3 (2)	.18
Social and common responsibility	4 (10.5)	34 (89.5)	17 (24.6)	52 (75.4)	5 (9.6)	47 (90.4)	6.1 (2)	<.05
Inquisitive messaging	0 (0)	38 (100)	0 (0)	69 (100)	0 (0)	52 (100)	__[a]	—
Preparations and uncertainty reduction								
Clarification	0 (0)	38 (100)	4 (5.8)	65 (94.2)	2 (3.8)	50 (96.2)	2.3 (2)	.32
Events, campaigns, and activities	34 (89.5)	4 (10.5)	55 (79.7)	14 (20.3)	44 (84.6)	8 (15.4)	1.8 (2)	.42
Request for contributions	0 (0)	38 (100)	0 (0)	69 (100)	0 (0)	52 (100)	—	—
Showing gratitude	0 (0)	38 (100)	1 (1.4)	68 (98.6)	0 (0)	52 (100)	1.3 (2)	.52
Reassurance	0 (0)	38 (100)	1 (1.4)	68 (98.6)	0 (0)	52 (100)	1.3 (2)	.52
Advisories and alerts								
Risk groups	4 (10.5)	34 (89.5)	3 (4.3)	66 (95.7)	4 (7.7)	48 (92.3)	1.5 (2)	.47
General advisories and vigilance	8 (21.1)	30 (78.9)	16 (23.2)	53 (76.8)	11 (21.2)	41 (78.8)	0.1 (2)	.95

[a]Not available.

Discussion

Principal Findings

Drawing on the CERC model, this study analyzed how Singapore's public health authorities communicated dengue risks to the public on Facebook. Specifically, we examined how public health communication regarding dengue aligns with the message themes outlined in the CERC model. We explored how the CERC strategies used by the Singapore public health authorities vary between outbreak and nonoutbreak years, as well as the differences in the CERC strategies used across various phases within the dengue outbreak year. The findings revealed that during the dual epidemic of dengue and COVID-19, Singapore's public health communication about dengue adhered to CERC principles. These communication strategies differed between dengue outbreak and nonoutbreak years, with a greater emphasis on requests for contributions, as well as prevention and vigilance messages, during the outbreak year as compared to the nonoutbreak year. During the dengue outbreak year, messages related to dengue cases and collective efforts were more prominent in the initial phase as compared to other phases (ie, precrisis and postcrisis).

Public health communication on dengue during the COVID-19 pandemic closely followed the message themes outlined in the CERC model, with a greater emphasis on pandemic intelligence, personal preventive measures and mitigation, and disease-related events and campaigns. These findings were consistent with previous research on the Singaporean public health communication on COVID-19 [18]. This shows that public health authorities in Singapore have put in much effort to raise public awareness about the severity of dengue and the prevention measures that the public should adopt. However, there was an absence of inquisitive messaging on Facebook, which could have been used to address the public's inquiries about dengue-related issues. The communication of vigilant information by the public health authorities, such as general advisories and vigilance and pandemic intelligence, may contribute to increased uncertainty among the public [18,25]. A useful strategy to reduce the public's uncertainty about crises and prevent confusion is to enhance messaging on inquisitive, especially during the initial phase. Therefore, given the lack of inquisitive messaging on dengue, our results suggest that Singapore health authorities should consider addressing the public's questions about dengue on social media during times of uncertainty [13,25]. However, as this study only examined Singapore's dengue-related public health messaging on Facebook, it remains unclear whether the public health authorities incorporated any inquisitive messaging in its dengue-related risk communication on other media platforms. Future research should investigate health authorities' efforts in addressing public queries by scrutinizing risk communication messages across different platforms.

We found that dengue-related message themes varied between the dengue outbreak and nonoutbreak years. Notably, a higher proportion of alert-related messages, such as "general advisories and vigilance", were communicated more during the dengue outbreak year as compared to the nonoutbreak year. This finding highlights the health authorities' adaptive communication strategies in implementing alerts and regulations in response to the severe dengue outbreak. The emphasis on general advisories

XSL•FO
RenderX

and vigilance aligns with the principles recommended by the World Health Organization for global vector control, such as risk-based prevention and intervention through data analytics [6]. In line with the World Health Organization principles, relevant health authorities in Singapore promptly alert stakeholders and initiate national dengue campaigns upon detecting outbreak signals, ensuring collective preparedness for outbreaks [4]. These alerting messages have also been commonly posted on social media by public health authorities in response to the outbreaks of other infectious diseases, such as COVID-19 and Ebola [4,25]. These messages can help increase public awareness about disease outbreaks and prompt immediate action, thereby helping to prevent a surge in dengue cases.

Our results also showed that the number of preventive messages conveyed by Singapore's public health authorities on social media was significantly higher during the outbreak year (2020) as compared to the nonoutbreak year (2021), indicating a potential relaxation of dengue communication campaigns during nonoutbreak periods. However, the recurrent nature of dengue hyperendemicity requires public vigilance and preparedness, even during nonoutbreak years [26]. It would be beneficial for health authorities to also deliver dengue preventive messages during nonoutbreak periods. Maintaining consistent risk communication across outbreak and nonoutbreak years would help sustain public vigilance and better equip individuals for the potential future dengue outbreaks.

Furthermore, our study revealed variations in the CERC message themes communicated across different phases during the dengue outbreak year. The findings align with the principles of the CERC framework, suggesting that health authorities should emphasize different themes at various stages of a crisis [12]. For example, public health messaging on "pandemic intelligence" was disseminated more during the initial and postcrisis phases compared to the precrisis phase in 2020. Reports indicated a significant rise in dengue cases during the initial stage, and although the situation was brought under control in the postcrisis phase, the issue remained a key focus [16]. Therefore, our results demonstrated that the public health authorities in Singapore transparently reported the reality of dengue transmission during the initial outbreak to raise public awareness. This adherence to transparency in dengue-related communication echoes Singapore's past risk communication strategies on Facebook during the Zika outbreak, as noted by Lwin et al [9], where authorities addressed the public's need for information by sharing pandemic intelligence during the initial phase.

Messages about social and common responsibilities were also communicated more frequently during the initial stage, reflecting the health authorities' efforts to emphasize the importance of community-level measures in disease prevention. The CERC model suggests that public health authorities should promote self-efficacy and encourage personal response actions [12]. One effective communication strategy is to emphasize shared responsibility among the public and stakeholders through social media [27]. By highlighting social responsibilities, individuals are more likely to actively comply with preventive measures [9,28]. This is particularly relevant in collectivist cultures, such as those in Asian societies, where social norms and community well-being are strongly emphasized [29]. Therefore, our results highlighted Singapore's proactive adoption of CERC principles in responding to various stages of dengue outbreaks, while also considering social contexts in crisis communication.

We also found that the CERC theme—preparations and uncertainty reduction—did not differ significantly between outbreak and nonoutbreak years or across different phases. This suggests that Singapore's public health authorities consistently engaged in risk communication on social media to prepare the public for potential dengue outbreaks, regardless of the phase or outbreak status. This finding aligns with the hyperendemic nature of dengue in Singapore, where cyclical outbreaks remain a persistent threat. Hyperendemic diseases necessitate continuous and preventive communication strategies to maintain public awareness and preparedness while enabling adaptive transitions during outbreaks [23]. The findings of this study highlight how Singapore's public health authorities have adopted the CERC principles by consistently implementing preemptive communication plans. These efforts are critical for fostering sustained public preparedness and are essential for effectively managing hyperendemic diseases like dengue.

Implications and Limitations

This study contributes to the literature on crisis communication during infectious disease outbreaks in 3 notable aspects. First, it adds to the body of work on CERC by being one of the first to analyze and evaluate public health communication in Singapore, a country renowned for its effective health management during both the COVID-19 pandemic and dengue outbreaks. In particular, we examined how Singapore's health authorities applied the communication themes suggested by the CERC framework during both outbreak and nonoutbreak periods of a hyperendemic. Hence, this study provides a nuanced understanding of crisis communication principles across different epidemiological scenarios, thereby expanding the application of CERC model in the context of a hyperendemic disease. Second, this study contributes to the extant literature on public health messaging about infectious diseases by examining the communication strategies used by health authorities to address dengue within the unique context of dual epidemics.

Third, by identifying risk communication strategies within the context of a hyperendemic disease, this study highlights the need to refine the CERC framework to address the persistent and cyclical nature of hyperendemic diseases. Specifically, the absence of clear resolution and evaluation phases calls for reconsidering how to integrate continuous assessment into the framework. This adaptation could enhance the relevance of the framework for hyperendemic diseases. Our findings provide a foundation for future research to refine the CERC framework by emphasizing the importance of analyzing and comparing health authorities' risk communication strategies across different phases of hyperendemic health crises. This includes considering both the phases outlined in the CERC model and the actual trends in disease progression. Furthermore, future research could focus on developing a modified version of the CERC framework that incorporates effective communication principles tailored

XSL•FO
RenderX

to the unique characteristics of hyperendemic diseases. For example, this refined framework could highlight the importance of continuous preparation, uncertainty reduction, and long-term prevention efforts across various phases while allowing for the flexible adjustment of messaging strategies in response to disease outbreaks.

This study offers practical insights for health authorities and policymakers on how to manage a hyperendemic disease (eg, dengue), especially when faced with another epidemic (eg, COVID-19). Understanding how CERC themes were used in dengue-related communication helps identify the strengths and areas for improvement in Singapore's risk communication strategies. For instance, our findings indicate a lack of inquisitive messaging on social media. Addressing the public questions about dengue during the initial stages of dengue outbreaks could reduce uncertainty in the future. In addition, establishing health emergency preparedness is crucial given the recurring nature of a hyperendemic (ie, dengue) [30]. During nonoutbreak periods, a proactive communication strategy focused on constant preventive messaging on social media can keep the public vigilant and prepared for potential health outbreaks [26]. Conversely, during outbreak years, policymakers could use reactive crisis response strategies, such as disseminating dengue-related advisories. As a result, health authorities and policymakers are strongly recommended to adapt these multifaceted communication strategies to different hyperendemic contexts.

While this study presented the CERC strategies used in Singapore's public health communication about dengue during COVID-19, several limitations should be acknowledged. First, this study focused solely on dengue-related messages posted on Facebook. Dengue-related messages posted on other social media platforms, such as X and Instagram, as well as online and offline news reports and printed communication materials such as banners and brochures (which lack publication dates), were excluded. As a result, our findings may not provide a comprehensive representation of Singapore's health messaging on dengue. The sole focus on Facebook messages may partly explain the limited representation of health authorities like the MOH in our data. Given that the MOH and other health authorities may also communicate about dengue on other platforms, such as their official websites, our data may underrepresent these institutions. Future research should consider incorporating data from a wider range of platforms and health authorities.

Second, it is important to note that this study encompassed only a portion of the Facebook messages posted in 2022, as data collection concluded on September 30, 2022. Hence, we only analyzed messages posted before that date, which may affect the interpretation of Singapore's dengue-related communication for the full year in 2022. In addition, for statistical comparison of communication strategies between outbreak and nonoutbreak years, we only used data from 2020 and 2021, excluding the 2022 data since it only covers 9 months. To gain a more comprehensive understanding of the communication strategies adopted by health authorities and their variation across different years, future research could include Facebook posts spanning a longer time frame.

Third, this study did not evaluate the effectiveness of communication strategies by analyzing public engagement or responses to different message themes. Given that public engagement can provide valuable insights into which messages are most effective and how the public perceives government communication, future research should incorporate sentiment analysis or engagement metrics (eg, number of shares, comments, and reactions) to assess the public response to different message themes.

Conclusions

Drawing on the CERC model, this study examined Singapore's public health messaging about dengue during the COVID-19 pandemic. Overall, Singapore's public health messaging about dengue was well aligned with those suggested by the CERC model. Compared to the nonoutbreak year of dengue, public health messaging during the outbreak year emphasized raising public awareness by prioritizing the dissemination of advisories and personal prevention messages related to dengue. Furthermore, public health messaging also actively provided updates about dengue cases to fulfill the public's information needs and to encourage collective action for dengue prevention, particularly during the initial stage of the outbreak. By analyzing Singapore's public health communication efforts about dengue, this study could provide suggestions for health authorities to effectively communicate in the hyperendemic context.

Acknowledgments

This research was supported by the National Research Foundation, Prime Minister's Office, Singapore, under its Campus for Research Excellence and Technological Enterprise program (grant NRF2020-ITC003-0001).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Codebook for communication strategies from public health authorities in Singapore.
[DOCX File , 62 KB - formative_v9i1e66954_app1.docx]

References

XSL•FO
RenderX

1. Bhatt S, Gething PW, Brady OJ, Messina JP, Farlow AW, Moyes CL, et al. The global distribution and burden of dengue. Nature 2013;496(7446):504-507 [FREE Full text] [doi: 10.1038/nature12060] [Medline: 23563266]
2. Lim JT, Dickens BS, Tan KW, Koo JR, Seah A, Ho SH, et al. Hyperendemicity associated with increased dengue burden. J R Soc Interface 2021;18(182):20210565 [FREE Full text] [doi: 10.1098/rsif.2021.0565] [Medline: 34520691]
3. Lim JT, Chew LZX, Choo ELW, Dickens BSL, Ong J, Aik J, et al. Increased dengue transmissions in singapore attributable to SARS-CoV-2 social distancing measures. J Infect Dis 2021;223(3):399-402 [FREE Full text] [doi: 10.1093/infdis/jiaa619] [Medline: 33000172]
4. Cheema HA, Mujtaba RS, Siddiqui A, Vohra LI, Shahid A, Shah J, et al. Singapore's dengue outbreak amidst the COVID-19 pandemic: Challenges, responses, and lessons. IDR 2023;Volume 16:1081-1085. [doi: 10.2147/idr.s397407]
5. Huang L, Xiao G, Chen H, Niu X, Fu X, Xu H, et al. Geographical clusters of dengue outbreak in Singapore during the Covid-19 nationwide lockdown of 2020. Sci Data 2022;9(1):547 [FREE Full text] [doi: 10.1038/s41597-022-01666-y] [Medline: 36071062]
6. Ho SH, Lim JT, Ong J, Hapuarachchi HC, Sim S, Ng LC. Singapore's 5 decades of dengue prevention and control-Implications for global dengue control. PLoS Negl Trop Dis 2023;17(6):e0011400 [FREE Full text] [doi: 10.1371/journal.pntd.0011400] [Medline: 37347767]
7. Egger J, Ooi EE, Kelly DW, Woolhouse ME, Davies CR, Coleman PG. Reconstructing historical changes in the force of infection of dengue fever in Singapore: implications for surveillance and control. Bull World Health Organ 2008;86(3):187-196 [FREE Full text] [doi: 10.2471/blt.07.040170] [Medline: 18368205]
8. National Environment Agency. URL: https://www.nea.gov.sg/corporate-functions/resources/research/environmental_health_institute/wolbachia-aedes-mosquito-suppression-strategy [accessed 2024-04-18]
9. Lwin M, Lu J, Sheldenkar A, Schulz P. Strategic uses of Facebook in zika outbreak communication: Implications for the crisis and emergency risk communication model. Int J Environ Res Public Health 2018;15(9):1974 [FREE Full text] [doi: 10.3390/ijerph15091974] [Medline: 30201929]
10. Bao H, Cao B, Xiong Y, Tang W. Digital media's role in the COVID-19 pandemic. JMIR Mhealth Uhealth 2020;8(9):e20156 [FREE Full text] [doi: 10.2196/20156] [Medline: 32530817]
11. Mheidly N, Fares J. Leveraging media and health communication strategies to overcome the COVID-19 infodemic. J Public Health Policy 2020;41(4):410-420 [FREE Full text] [doi: 10.1057/s41271-020-00247-w] [Medline: 32826935]
12. Reynolds B, W Seeger M. Crisis and emergency risk communication as an integrative model. J Health Commun 2005;10(1):43-55. [doi: 10.1080/10810730590904571] [Medline: 15764443]
13. Malik A, Khan ML, Quan-Haase A. Public health agencies outreach through Instagram during the COVID-19 pandemic: Crisis and emergency risk communication perspective. Int J Disaster Risk Reduct 2021;61:102346 [FREE Full text] [doi: 10.1016/j.ijdrr.2021.102346] [Medline: 36337987]
14. Yoon HY. Is crisis and emergency risk communication as effective as vaccination for preventing virus diffusion? Measuring the impacts of failure in CERC with MERS-CoV outbreak in South Korea. Risk Anal 2022;42(7):1504-1523. [doi: 10.1111/risa.13842] [Medline: 34655090]
15. Hasan MM, Sahito AM, Muzzamil M, Mohanan P, Islam Z, Billah MM, et al. Devastating dengue outbreak amidst COVID-19 pandemic in Bangladesh: an alarming situation. Trop Med Health 2022;50(1):11 [FREE Full text] [doi: 10.1186/s41182-022-00401-y] [Medline: 35078540]
16. Accessed April 18, 2024. NEA Vox. URL: https://www.nea.gov.sg/media/nea-vox [accessed 2024-04-18]
17. Meadows CW, Meadows CZ, Tang L, Liu W. Unraveling public health crises across stages: Understanding Twitter emotions and message types during the california measles outbreak. Communication Studies 2019;70(4):453-469. [doi: 10.1080/10510974.2019.1582546]
18. Ho SS, Chuah ASF, Ho VS, Rosenthal S, Kim HK, Soh SSH. Crisis and emergency risk communication and emotional appeals in COVID-19 public health messaging: Quantitative content analysis. J Med Internet Res 2024;26:e56854 [FREE Full text] [doi: 10.2196/56854] [Medline: 39288399]
19. Jin X, Spence PR. Understanding crisis communication on social media with CERC: topic model analysis of tweets about hurricane maria. Journal of Risk Research 2020;24(10):1266-1287. [doi: 10.1080/13669877.2020.1848901]
20. Gondwe G, Adae EK. Media campaign strategies in communicating HIV/AIDS in zambia: Comparing risk and crisis communication strategies in mitigating behavior change among sex workers. International Journal of Communication 2021;15(2021):5252-5271.
21. Gentili D, Bardin A, Ros E, Piovesan C, Ramigni M, Dalmanzio M, et al. Impact of communication measures implemented during a school tuberculosis outbreak on risk perception among parents and school staff, Italy, 2019. Int J Environ Res Public Health 2020;17(3):911 [FREE Full text] [doi: 10.3390/ijerph17030911] [Medline: 32024183]
22. Khan S, Akbar SMF, Yahiro T, Mahtab MA, Kimitsuki K, Hashimoto T, et al. Dengue infections during COVID-19 period: Reflection of reality or elusive data due to effect of pandemic. Int J Environ Res Public Health 2022;19(17):10768 [FREE Full text] [doi: 10.3390/ijerph191710768] [Medline: 36078486]
23. Cairns G, de Andrade M, MacDonald L. Reputation, relationships, risk communication, and the role of trust in the prevention and control of communicable disease: a review. J Health Commun 2013;18(12):1550-1565. [doi: 10.1080/10810730.2013.840696] [Medline: 24298887]

XSL•FO
RenderX

24. Hayes AF, Krippendorff K. Answering the call for a standard reliability measure for coding data. Communication Methods and Measures 2007;1(1):77-89. [doi: 10.1080/19312450709336664]
25. Dalrymple KE, Young R, Tully M. “Facts, Not Fear”. Science Communication 2016;38(4):442-467. [doi: 10.1177/1075547016655546]
26. Adalja AA, Sell TK, Bouri N, Franco C. Lessons learned during dengue outbreaks in the United States, 2001-2011. Emerg Infect Dis 2012;18(4):608-614 [FREE Full text] [doi: 10.3201/eid1804.110968] [Medline: 22469195]
27. Low LL, Tong SF, Ang JY, Abdullah Z, Hamid MA, Risman MS, et al. Social responsibility perspective in public response to the COVID-19 pandemic: a grounded theory approach. BMC Public Health 2022;22(1):469 [FREE Full text] [doi: 10.1186/s12889-022-12819-4] [Medline: 35264136]
28. Finset A, Bosworth H, Butow P, Gulbrandsen P, Hulsman RL, Pieterse AH, et al. Effective health communication - a key factor in fighting the COVID-19 pandemic. Patient Educ Couns 2020;103(5):873-876 [FREE Full text] [doi: 10.1016/j.pec.2020.03.027] [Medline: 32336348]
29. Hofstede G. Culture? Consequences: Comparing Values, Behaviors, Institutions, and Organizations across Nations. 2. ed. Thousand Oaks, CA: SAGE; 2013.
30. Savoia E, Lin L, Viswanath K. Communications in public health emergency preparedness: a systematic review of the literature. Biosecur Bioterror 2013;11(3):170-184 [FREE Full text] [doi: 10.1089/bsp.2013.0038] [Medline: 24041193]

Abbreviations

CERC: Crisis and Emergency Risk Communication
IRB: institutional review board
MOH: Ministry of Health
NEA: National Environment Agency
RQ: research question

Edited by A Mavragani; submitted 27.09.24; peer-reviewed by H Zheng, CX Yang; comments to author 20.02.25; revised version received 12.03.25; accepted 07.04.25; published 22.05.25.

Please cite as:
Ho SS, Ou M, Huang NM, Chuah ASF, Ho VS, Rosenthal S, Kim HK
Public Health Messaging About Dengue on Facebook in Singapore During the COVID-19 Pandemic: Content Analysis
JMIR Form Res 2025;9:e66954
URL: https://formative.jmir.org/2025/1/e66954
doi:10.2196/66954
PMID:

©Shirley S Ho, Mengxue Ou, Nova Mengxia Huang, Agnes SF Chuah, Vanessa S Ho, Sonny Rosenthal, Hye Kyung Kim. Originally published in JMIR Formative Research (https://formative.jmir.org), 22.05.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Needs and Preferences of Swedish Young Adults for a Digital App Promoting Mental Health Literacy, Occupational Balance, and Peer Support: Qualitative Interview Study

Martin Karaba Bäckström[1*], OTR, MSc; Sonya Girdler[1,2,3,4,5,6*], PhD, OTR; Ben Milbourn[2,6*], PhD, OTR; Annika Lexén[1,2*], PhD, OTR

[1]Mental Health, Activity and Participation, Faculty of Medicine, Lund University, Lund, Sweden
[2]Curtin Autism Research Group (CARG), Curtin University, Perth, Australia
[3]Center of Neurodevelopmental Disorders (KIND), Karolinska Institutet, Stockholm, Sweden
[4]Stockholm Health Care Services, Stockholm County Council, Stockholm, Sweden
[5]School of Allied Health, University of Western Australia, Perth, Australia
[6]Curtin School of Allied Health, Perth, Australia
[*]all authors contributed equally

Corresponding Author:
Martin Karaba Bäckström, OTR, MSc
Mental Health, Activity and Participation
Faculty of Medicine
Lund University
Sölvegatan 19
Lund, 223 62
Sweden
Phone: 46 762205058
Email: martin.karaba_backstrom@med.lu.se

Abstract

Background: Young adults experience stressors in their transition to adulthood and are at increased risk of mental ill-health. This risk is compounded by young adults' low levels of mental health literacy and limited competencies in implementing strategies promoting mental health and well-being in their daily lives. Previous research suggests that digital mental health apps may be particularly effective in increasing the mental health literacy of young adults. In Sweden, there is a lack of research on young adults' unique perspectives on what constitutes mental health, well-being, and ill-health—perspectives that could inform the coproduction of evidence-based interventions targeting these issues.

Objective: The overarching aim of this study was to conduct a needs assessment as part of coproducing a digital mental health app for Swedish young adults. More specifically, the study addressed two research questions: (1) What do Swedish young adults perceive as contributing to the mental health, well-being, and ill-health of themselves and their peers? (2) What are Swedish young adults' preferences and ideas on how a digital mental health app can support their mental health during young adulthood, including their perspectives on the app's usability?

Methods: We conducted semistructured interviews with 16 young adults and analyzed the data using reflexive thematic analysis.

Results: Of the 16 study participants, 9 (56%) identified as women and 7 (44%) as men. Their mean age was 23.6 (SD 4.22; range 18-29) years. Furthermore, 56% (9/16) were pursuing or had obtained a higher education degree, while 44% (7/16) had completed or were in the process of completing a high school diploma. The interviews and subsequent analysis revealed three main themes: (1) "To feel that life is worth living"—pathways through pressures and pursuit of mental well-being during young adulthood, (2) "A personal space for working on one's own mental well-being"—digital companionship with others, and (3) "Something that is designed for me"—customizing one's digital mental health journey.

Conclusions: In line with the preferences of Swedish young adults, the promotion of mental health and well-being through digital technology and eHealth should focus on a customizable app that supports balance in daily life while strengthening mental health competencies. The content should center on fostering and maintaining meaningful relationships and activities, addressing challenges such as negative social media use and stress recovery, and enhancing mental health knowledge and peer support.

XSL•FO
RenderX

Future efforts should focus on researching young adults' experiences of the life phase of emerging adulthood and its implications for mental health. In addition, future technical development and research on digital mental health apps should include the perspectives of stakeholders, such as mental health professionals, and involve prototype testing with diverse groups to ensure the app's relevance, user engagement, and effectiveness.

(JMIR Form Res 2025;9:e71563) doi:10.2196/71563

KEYWORDS

emerging adulthood; design thinking; mental health promotion; mental ill-health; needs assessment; artificial intelligence; AI

Introduction

Background

Globally, young adults aged between 18 and 29 years disproportionally bear the burden of mental ill-health [1]. Young adults are at greater risk of being exposed to anxiety, stress, and depression compared to the general population [2]. As a result, 75% of individuals experiencing long-term mental ill-health report onset before the age of 24 years, affecting their transition to adulthood and limiting their participation in education, employment, and the community [3]. Furthermore, the World Health Organization [4,5] reports that suicide is the second leading cause of death among young adults. In Sweden, mental ill-health costs the Swedish economy approximately 5% of the gross domestic product [6], with young adults carrying a significant proportion of these costs. Research points to a mental health crisis among young adults, which is impacting national and global economies as well as welfare systems [3,7].

In this study, mental health is conceptualized as existing on a continuum ranging between mental ill-health and mental well-being [8]. This approach is favored over the more commonly used terms and discourse around mental illness, mental health conditions, or mental disorders [4]. This choice was made for 2 reasons: first, to align with the terminology guidelines set by the Swedish National Board of Health and Welfare; and second, because these terms are commonly used by Swedes to describe mental health on individual, group, and societal levels [8,9].

Swedish young adults report high levels of mental ill-health, driven by worries about their future, daily stressors, and challenges in balancing educational and employment demands [9]. This is not surprising, given the common difficulties associated with the phase of "emerging adulthood" [10]. Emerging adulthood as a developmental life stage encompasses young adults aged 18 to 29 years [9]. It is characterized by the challenges and opportunities in navigating new environments, establishing major life roles, and meeting new demands—often in the context of diminished social support and resources [9]. During this life stage, young adults describe experiencing feelings of uncertainty and being "in between," struggling to choose between multiple life choices and possibilities [9]. Recent research [3,8] also points to the negative impact of new mega trends on the mental health of young adults, including concerns regarding climate change and the loss of biodiversity, increasingly toxic social media, and rising intergenerational inequalities. Increasingly, there is recognition [3] that young adults are experiencing an unprecedented mental health crisis and that there is a need for new insights and approaches.

In addition, negative community attitudes toward mental ill-health contribute to stigma and act as barriers to accessing support and mental health care [11]. Furthermore, limited community understanding of the signs signaling deteriorating mental ill-health and approaches to promoting well-being negatively impact the community's ability to support young adults [12]. While there is a heightened awareness and normalization of mental ill-health among youth, driving an increasing demand for mental health services, this has not been paralleled by investment in mental health services [2,13]. Therefore, while there is increasing recognition of the stressors associated with young adulthood and the importance of this period in preventing long-term ill-health, young adults still infrequently receive or seek targeted support from mental health services [8,14].

Mental health promotion—supporting protective factors and healthy behaviors—has been proposed as a means of enhancing the mental health and well-being of young adults [13,15] and reducing the associated economic costs [7]. Advances in technology have enabled the development of digital health technologies and apps, which have emerged as key tools in promoting mental health across populations [16]. Digital tools have demonstrated particular efficacy in improving mental health literacy and supporting individuals with mental ill-health [17-19]. Mental health literacy, that is, the ability to recognize and manage challenges to one's mental health [12], empowers individuals to adopt positive behaviors in daily life and support those around them [12]. At the community level, mental health literacy plays a critical role in preventing long-term mental ill-health, supporting suicide prevention efforts, reducing stigma, and promoting thriving communities [4]. Given their limited life experiences, many young adults commonly have particularly low levels of mental health literacy [8]. Globally, several digital mental health literacy interventions have targeted adolescents and young adults [18,19], demonstrating improvements in mental health outcomes through strengths-based approaches, peer support, life skills training, and increased resilience [19-21]. However, there is a lack of evidence-based digital mental health apps specifically tailored to the needs of Swedish young adults [22].

In summary, the stressors that young adults face in transitioning to adulthood can threaten their long-term mental health [3,9]. Young adults' risk of mental ill-health is compounded by their generally low levels of mental health literacy, including their limited understanding of strategies promoting mental well-being in their daily lives [8]. Digital technologies, such as a mental health app, may provide a useful resource in delivering a strengths-based approach to mitigating mental health threats.

XSL•FO
RenderX

Prior research highlights the importance of engaging stakeholders and coproduction in developing successful and culturally appropriate digital health interventions targeted toward the needs of a specific group [23-25]. However, less attention has been given to how young adults themselves perceive their everyday mental health challenges and specifically what kinds of digital support they find both relevant and accessible. Without this user-centered perspective, digital interventions risk missing the mark in addressing actual needs. Thus, there is a need for in-depth understanding of young adults' experiences and preferences to inform the cocreation of more meaningful and effective digital mental health tools. In line with this thinking, this study aimed to explore Swedish young adults' mental health needs during emerging adulthood, examining their perspectives on how digital technology could meet these needs, laying the foundation for the development of a culturally and contextually relevant digital mental health app. This approach is likely to enhance the efficacy, accessibility, and utility of a future digital mental health solution [24,26].

Study Aim

The overarching aim of this study was to conduct a needs assessment as part of coproducing a digital mental health app for Swedish young adults. More specifically, the study addressed two research questions:

1. What do Swedish young adults perceive as contributing to the mental health, well-being, and ill-health of themselves and their peers?
2. What are Swedish young adults' preferences and ideas on how a digital mental health app can support their mental health during young adulthood, including their perspectives on the app's usability?

Methods

Study Context

The ultimate goal of this line of research is to coproduce an eHealth app supporting the mental health and well-being of Swedish young adults. The Medical Research Council's framework for developing and evaluating complex interventions [27] and the design thinking approach [28] provide the theoretical foundation for this work, outlining the necessary steps in innovating new health care solutions—from inception to implementation. The initial phase, referred to as a "needs assessment" [27,28], involves understanding and defining the needs of the target audience of a proposed intervention as well as identifying the challenges and opportunities related to its successful implementation and use. Together, these approaches have demonstrated utility in coproducing innovative health promotion initiatives [29-31].

Study Design

A qualitative descriptive interview study [32], guided by a formative research approach [33], was considered the most appropriate method to address this study's aim and research questions. The formative research approach focuses on exploring a target group's needs and preferences to ensure that a planned intervention is relevant, user-centered, and effective [33]. This research design has proven particularly useful in conducting needs assessments and informing eHealth interventions tailored to the specific needs of a defined target group [29,30]. In addition, this study is designed to align with the Standards for Reporting Qualitative Research [34], strengthening the study's dependability and confirmability. The completed Standards for Reporting Qualitative Research checklist is available in Multimedia Appendix 1.

Recruitment and Participants

A total of 16 participants were recruited via convenience and snowballing sampling [32] (Table 1). The number of participants considered adequate for a qualitative descriptive study is estimated to range from 15 to 20 [35]. Furthermore, previous studies with similar aims support the sufficiency of this number of participants [36,37]. The mean age of the participants was 23.6 (SD 4.22; range 18-29) years. They were literate in Swedish and actively using digital devices and apps.

XSL•FO
RenderX

Table 1. Sociodemographic characteristics of Swedish young adults participating in a qualitative study (March-September 2024) exploring mental health and preferences for digital mental health support.

Participant IDs	Gender	Age (years)	Educational level	Employment status or role
1	Man	28	High school diploma	Teaching assistant
2	Man	18	Ongoing high school diploma	Student
3	Man	28	University degree	Nurse in a psychiatric acute clinic
4	Man	25	University degree	Industrial technician
5	Woman	29	University degree	Occupational therapist at psychiatric rehabilitation unit
6	Woman	29	High school diploma	Care assistant
7	Woman	28	University degree	Employed at a municipality
8	Woman	18	Ongoing high school diploma	Student
9	Woman	20	High school diploma	Unemployed
10	Woman	18	Ongoing high school diploma	Student
11	Woman	26	Higher vocational education diploma	Video editor
12	Woman	22	Ongoing university degree	Student
13	Woman	25	Ongoing university degree	Student
14	Woman	24	University degree	Sustainability strategist
15	Man	18	High school diploma	Unemployed
16	Man	23	Ongoing university degree	Student

Convenience and snowballing sampling took place between March and September 2024. The first author (MKB) contacted 28 young adults who were distant acquaintances from past or current educational settings; workplaces; or sports clubs in southern, central, and northern Sweden. From a perspective of positionality, the author (MKB) had no ongoing contact or relationship with the young adults approached. Initial contact was made via SMS text messaging or a messaging app. The young adults were asked whether they were interested in participating or whether they could forward the invitation to someone else who met the inclusion criteria. If the young adult was interested, they were provided with an information letter outlining the key aspects of their participation. At the end of each interview, participants were asked whether they knew any peers who might be interested in taking part in the study.

Data Collection

A semistructured interview guide (Multimedia Appendix 2) was developed in line with the study aims and research questions to ensure consistency across interviews [32]. The guide consisted of 2 focus areas: one eliciting the perspectives and experiences of the young adults related to mental health (eg, “Can you give examples of what mental health looks like among young adults in your community?”) and the other focusing on gathering the young adults’ preferences for the development of a digital mental health app (eg, “Can you give examples of important functionalities to make the digital mental health application usable?”). Interviews loosely adhered to the interview guide, with broad, open-ended questions establishing the context, followed up with more specific, probing questions, ensuring a natural flow and an in-depth understanding of the topic. However, several interviews followed a more iterative style, with conversations naturally following participants’ views on the topic.

A pilot test of the interview guide was undertaken with 2 young adults who met the inclusion criteria, resulting in a simplification of the language, with the questions made more open-ended. Given that the pilot interviews provided valuable information relevant to the study’s aim, they were included in the data analysis. In line with participants’ preferences, interviews were conducted either digitally or in person by the first author (MKB) between March 17 and September 11, 2024. The interviews lasted an average of 48 (SD 7.09) minutes.

Data Analysis

The interview transcripts were analyzed inductively, using reflexive thematic analysis, with themes emerging from the data [38]. This analysis method was chosen because of its emphasis on creating a narrative and story reflecting the study context and the voices of the participants [38]. Reflexive thematic analysis acknowledges researchers’ biases and accepts that the interpretation of the data is influenced by prior knowledge and experiences [38].

The ontological and epistemological position of the authors and this research is based on constructivism [39]. This approach assumes that knowledge is created relationally through humans interacting with each other in daily life [39]. The analysts (MKB, SG, BM, AL, and Moa Yngve, an external researcher [ER]) involved in this study have research roles and previous experience in conducting research within fields related to young adults’ mental health and well-being. Their broader research experience includes promoting community participation and functioning in daily life through a wide array of health promotion interventions. The first author (MKB) has personal

XSL•FO
RenderX

experiences of mental ill-health during young adulthood. Both MKB and AL also have experience as close relatives of family members with long-term mental ill-health.

After transcribing the interviews verbatim, the first author (MKB) read the transcripts, familiarizing himself with the content. Although the interview guide questions were structured around 2 focus areas, the young adults' answers were analyzed as a whole to capture overarching themes and patterns across the data. The initial 6 interviews were coded semantically, and a code chart reflecting the codes and initial themes from these interviews was created and developed iteratively and latently. Thereafter, using this chart, MKB coded the remaining interview transcripts. During the coding process, the code chart was revised and adapted several times as the understanding of overall themes emerged. The ER coded 2 interviews, providing feedback and suggestions for additional codes (MKB and the ER). Subsequently, the codes were mapped into clusters, which underpinned the themes (MKB, the ER, and AL). Across the data analysis process, the research team engaged in an iterative discussion of the coding and thematization until reaching a level of abstraction that was deemed logical and reflective of the material (MKB, SG, BM, AL, and the ER).

Ethical Considerations

This research was approved by the Swedish Ethical Review Authority (2024-00582-01) and adhered to ethical principles outlined by the World Medical Association [40]. All participants provided written informed consent before the interviews and were informed of their right to withdraw from the study at any time. The collected data were deidentified, and only members of the research team involved in the analysis had access to the interview transcripts. No financial or other compensation was offered for participating.

As sharing one's views on mental health can be sensitive and potentially evoke stress, anxiety, or worry, available supports were detailed in the information letter. In the event that participants experienced uncomfortable feelings or thoughts after engaging in the interviews, they were encouraged to contact the first (MKB) or last (AL) author for support—both have clinical experience in working with young people with mental health conditions. If this initial support was deemed insufficient by the young adult, given their consent, they would be referred to collaborating partners in mental health services. Sufficient time was allowed both before and after each interview as a strategy to promote rapport and create a safe space for sharing [41]. The sharing of an emotionally charged experience can bring meaning when participants deem the aims of a study worthwhile personally and to others [41]. Therefore, the overarching aims of the research project and its future implications for young adults' mental well-being was emphasized in the information letter and before each interview started.

Results

Overview

The data analysis process revealed three themes: (1) "To feel that life is worth living"—pathways through pressures and pursuit of mental well-being during young adulthood, (2) "A personal space for working on one's own mental well-being"—digital companionship with others, and (3) "Something that is designed for me"—customizing one's digital mental health journey. The themes and subthemes are presented in Figure 1. Similarly, in Table 2, the themes and subthemes are illustrated with exemplar quotes, referred to as "illustrative examples," along with condensed implications for designing the app.

Figure 1. Themes and subthemes identified through thematic analysis of interviews with Swedish young adults (aged 18-29 years) between March and September 2024, reflecting their views on mental health and digital mental health support preferences.

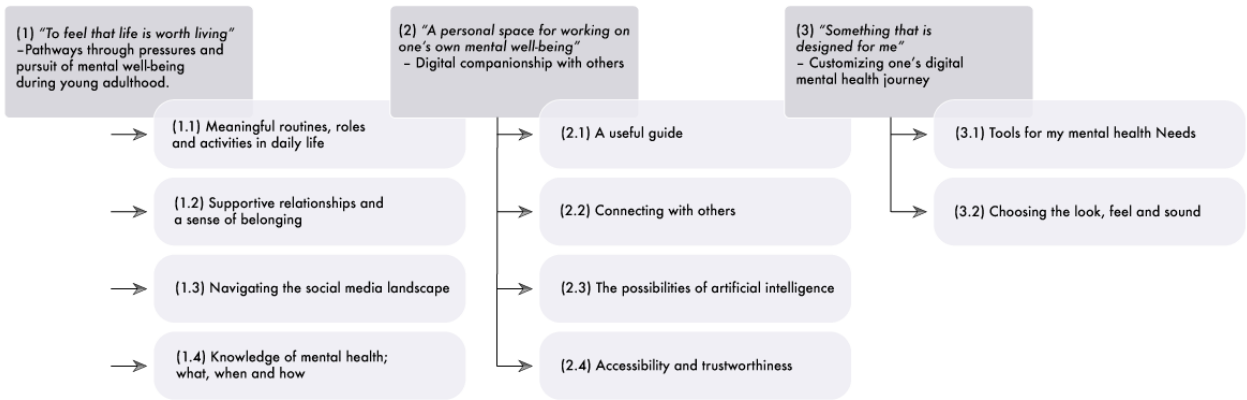

XSL•FO
RenderX

Table 2. Overview of themes, subthemes, and supporting quotes derived from interviews with Swedish young adults, showcasing their implications for the design of a digital mental health app.

Themes and subthemes	Illustrative examples	Implications for the design of a digital mental health app
1. "To feel that life is worth living"—pathways through pressures and pursuit of mental well-being during young adulthood		
1a. Meaningful routines, roles, and activities in daily life	"I hope it [the app] will allow people to reallocate their time so that, in some way, they can spend more time on what makes them feel good and on what is important to them." [Study participant 14]	• Include information about the importance of meaningful activities for mental well-being • Support balancing activities that young adults need to perform with those that they want to perform
1b. Supportive relationships and a sense of belonging	"I think that today a good safety net is probably essential for young adults to maintain mental well-being. We have so many factors, both societal and social, that easily lead to mental health issues. And that's exactly why a very strong safety net, both socially and societally, is needed to support good mental health." [Study participant 16]	• Increase competencies in young adults collectively taking care of each other's mental health
1c. Navigating the social media landscape	"Just based on myself and others, I feel that on social media and the internet today, we are bombarded with so much information about these, what should we call them, elite people who might be just one percent of the population, and they are the only ones we see." [Study participant 14]	• Support young adults in navigating a balanced and healthy relationship with social media
1d. Knowledge of mental health: what, when, and how	"My hypothesis is that we might have become a bit unaccustomed to talking about difficult feelings or tough moments. And that we could get better at showing our bad moments in public so people understand that life isn't always good—you have your good times and your bad times. And by sharing, we can learn how to handle those tough moments and learn from each other." [Study participant 16]	• Provide young adults with tools to proactively increase knowledge about their mental well-being • Teach young adults self-care and peer support strategies in managing mental health challenges • Offer young adults direct access to additional mental health resources and support
2. "A personal space for working on one's own mental well-being"—digital companionship with others		
2a. A useful guide	"I think an app like this could be very helpful and supportive for those who are in this transition phase." [Study participant 7]	• The digital mental health app should be a guide to young adults' mental well-being
2b. Connecting with others	"A way to get to know other people and their experiences, to learn from them, and perhaps gain insights. Their experiences can provide support for our own lives and the situations we face." [Study participant 15]	• Provide safe ways for young adults to interact with others for mutual support and learning • Facilitate the sharing of experiences to collectively address mental health needs
2c. The possibilities of artificial intelligence	"I've also noticed, or heard, and even tested and experienced a pretty good effect from talking with AI [artificial intelligence], like ChatGPT." [Study participant 3]	• Cocreate an iterative, empathetic conversational AI agent to support mental health
2d. Accessibility and trustworthiness	"If you turn to a tool that focuses specifically on mental health, you can expect it to meet that confidentiality standard that is essential for such conversations." [Study participant 12]	• Ensure that the app is available for free or offered at a minimal cost • Maintain transparency in how personal data and their use are kept safe and secure
3. "Something that is designed for me"—customizing one's digital mental health journey		
3a. Tools for my mental health needs	"If someone wants a motivating app to mutually support themselves and others, they should be able to have that. And if someone just wants a listening, accepting app that doesn't promote positive feelings, but is simply okay with everything, then maybe they should have that too." [Study participant 11]	• Offer customizable features to address diverse situations, challenges, and goals

XSL•FO
RenderX

Themes and subthemes	Illustrative examples	Implications for the design of a digital mental health app
3b. Choosing the look, feel, and sound	"That you can choose what color you want on the outside of the app, or that the inside of the app allows you to select whether you want it to be pink or blue, and little things like that. I think it makes it easier to feel like, 'This is my app.'" [Study participant 8]	• Allow customization of the app's appearance and interface to enhance use

Theme 1: "To Feel That Life Is Worth Living"—Pathways Through Pressures and Pursuit of Mental Well-Being During Young Adulthood

Overview

This theme revolves around the needs identified as essential to nurturing the mental health and well-being of Swedish young adults. The participants in the study described mental health as a deeply personal experience, with an individual's unique needs and preferences playing a crucial role in shaping their overall well-being. The importance of daily routines is captured in the first subtheme, (1a) meaningful routines, roles, and activities in daily life, highlighting the importance of nurturing mental well-being and growth as a young adult by actively doing what is important. In addition, the critical role of family, friends, and peers in supporting mental health was reflected in the second subtheme, (1b) supportive relationships and a sense of belonging, pointing to the importance of connectedness in everyday life. Furthermore, the study participants described contemporary societal phenomena as negatively impacting their mental well-being. This was mainly reflected in the third subtheme, (1c) navigating the social media landscape, with most of the study participants stressing the disempowering effect that social media has on their mental well-being and the experience of transitioning to adulthood. The fourth subtheme, (1d) knowledge of mental health: what, when, and how, captures participants' experiences of inadequate mental health support during emerging adulthood, as well as their perception that they were not competent in actively promoting their own mental well-being.

Subtheme 1a: Meaningful Routines, Roles, and Activities in Daily Life

The young adults involved in this study fundamentally described "mental health" as a state of emotional well-being, enabling them to engage in personally meaningful activities and fulfill roles within their communities. Engaging in these activities and roles was seen as key in fostering a positive sense of self, essential in supporting participants' transition to adulthood:

> *My personal experience...is that there have been things that...have been a bit tough. For example, not getting a job, or if I didn't get a job, or things like getting a driver's license.... There's a pressure there because it feels like those are adult things one should be achieving in life. But by having things like training, going to the gym, playing floorball, and people to talk to, I've gained so much energy that even the things that have been heavy, difficult and full of anxiety to some extent have become easier.* [Study participant 14]

Being "responsible" and managing their activities of daily life, including getting sufficient sleep, exercising regularly, eating well, and adhering to healthy routines, were important to "staying on track" during challenging periods. Engaging in leisure activities "protected" against mental ill-health, while engaging in meaningful activities provided opportunities for exploring and embracing alternative roles, relieving their focus on careers and "succeeding." Spending time on leisure activities enabled participants to develop and strengthen roles important for their relaxation, mental well-being, and rest, as well as support the growth of others.

However, several of the young adults described experiencing conflict in prioritizing leisure activities over work, study, or future-oriented obligations. Participants described themselves and their peers as focusing on meeting the necessary demands in achieving their career goals or "succeeding," seeing this as negatively impacting their mental health.

Subtheme 1b: Supportive Relationships and a Sense of Belonging

Social support was key in sustaining participants' mental health and aiding their recovery from the challenges stemming from their transition experiences. The availability of support from friends, family, and peers was seen as essential for both "surviving" and "living a good life":

> *Relationships with people. Deep relationships with people. Being able to have people you can talk to and feel secure with. I would say that's something important. It's kind of one of those basic things, like food and water.* [Study participant 14]

By contrast, participants recognized feeling "isolated" and "withdrawing" from relationships, preferred activities, and their community as signs of mental ill-health:

> *I feel like if a friend, for example, is having a bad day, they can call or text and talk about it. But when things are genuinely tough in life right now, you don't hear about it until afterward, when things are better again.* [Study participant 15]

Subtheme 1c: Navigating the Social Media Landscape

Social media use emerged as a complex factor influencing participants' mental health. While it enabled connection with friends as well as access to inspiring influencers, an online community, and content that resonated with participants' interests and sense of humor, social media also contributed to the feeling of being "locked in," particularly in response to compulsive scrolling or "doom scrolling." Excessive social media use resulted in participants feeling stressed, anxious, or "worthless," as though they had "wasted" the day. A participant described the experience as follows:

XSL•FO
RenderX

> *If you feel like you haven't done much in a day [and then spending time on social media], you end up feeling worse. You want to feel like you did something fun or something important with your day.* [Study participant 12]

Social media negatively impacted participants' self-esteem, promoting narrow, "one-size-fits-all" ideals related to wealth, social status, and "successful" career paths. Instances of online harassment and exposure to unrealistic beauty ideals were described as further contributing to participants' distress. While participants were aware of the potential harmful impact of excessive social media use, they reported struggling to reduce their engagement due to "fear of missing out." This fear made disengaging from social media platforms particularly challenging. Several of the young adults expressed a specific desire for support and guidance on how to develop a healthier, more balanced relationship with social media. They emphasized the need for tools and strategies to mitigate its negative impacts while retaining its positive aspects:

> *The reason why many people feel bad today is probably also because of social media, I would say....There are influencers or other famous people, or maybe people who live in the same city as you, who you think look good. And then you want to be like them. And then when you don't look like them maybe, or don't wear makeup like they do, you feel like you don't fit in. And then you can feel bad about it.* [Study participant 8]

Subtheme 1d: Knowledge of Mental Health—What, When, and How

Many participants described existing mental health care services as inadequate in meeting the diverse and evolving needs of young adults, leaving them disempowered in both managing their own mental health and supporting their peers. Participants further perceived that they lacked the knowledge to maintain their own mental well-being, which they felt contributed to their risk of mental ill-health:

> *If I take my own experiences, considering that I've dealt with mental health issues and still struggle with them from time to time, it took me four years to recognize the warning signs before I asked for help. When I first began, at the start of it all, I asked for help, but didn't receive what I needed, and that made me feel like this was something to stay quiet about.* [Study participant 10]

During the interviews, the majority of the young adults provided more comprehensive responses about aspects of mental ill-health than about the positive, health-promoting factors associated with mental well-being:

> *Mental health...for some reason, sadly, one immediately thinks of mental ill-health.* [Study participant 5]

Most participants' knowledge of mental health centered on noticing and responding to alarming symptoms in themselves and others. However, they expressed a strong desire to better understand mental health during emerging adulthood. Participants felt that collectively improving their understanding of approaches to fostering mental well-being in their daily lives could empower their generation.

Theme 2: "A Personal Space for Working on One's Own Mental Well-Being"—Digital Companionship With Others

Overview

This theme highlights young adults' largely positive view of digital technology as a mental health support tool, as seen in the first subtheme, (2a) a useful guide. A digital mental health app held promise as a way to "actively put effort" into their well-being. The second subtheme, (2b) connecting with others—relates to the participants' perception of a digital mental health app as a potentially valuable tool for fostering connections and increasing mental health knowledge. The third subtheme, (2c) the possibilities of artificial intelligence, explores the young adults' views on artificial intelligence, while the fourth, (2d) accessibility and trustworthiness, expands on data security and access to a digital mental health app.

Subtheme 2a: A Useful Guide

Many young adults expressed a strong desire to learn more about maintaining mental health and developing knowledge of how mental ill-health is manifested, coping mechanisms, and routines promoting well-being. However, they described feeling frequently overwhelmed by the vast amount of information available online and struggling to navigate and curate it. They envisioned a mental health app as offering clear guidance through an intuitive interface, making it easier to browse relevant topics, categories, and information relating to mental health:

> *Perhaps guides on how to proceed in order to somehow succeed in getting out of a dip. Even though it's strange maybe to say "just do it"...But if you don't have it on paper, it's often very difficult to see that you can get there. I think the opportunity to get a clear picture of how things can be handled and how one can get help, simply.* [Study participant 7]

Participants stressed the need for an app that offers practical advice for "handling daily life," helping to bridge the gap between their desire for knowledge and their difficulty in finding and prioritizing the right information online. A digital guide tailored to young adults' mental health challenges was seen as a promising tool for navigating the life phase of emerging adulthood.

Despite the aforementioned features and optimism, some participants questioned whether a digital app could effectively support the mental well-being of young adults; for instance, some viewed the current digitalization of society and health care services as "ineffective" and disturbing due to their belief that therapy provided in person is "better" and more "humane."

Subtheme 2b: Connecting With Others

A digital solution focusing on enabling interactions between young adults and providing a forum for them to discuss their mental health needs was viewed positively:

XSL•FO
RenderX

> *I also think it's important for an application like this to have some kind of forum that promotes interaction, whether it's between users or between a moderator and users. I think that would be helpful. We need more platforms to talk about this in safe spaces.* [Study participant 16]

Existing forums were frequently mentioned and described as providing a daily space to discuss worries or questions, as well as to gain insights and competencies. However, participants noted the importance of ensuring a safe space for discussing mental health and ill-health, while maintaining openness, but they cautioned that an open forum could be misused or abused. Participants noted that peer support could take the form of inspirational quotes as well as stories and narratives showcasing elements of "recovery" and "meaningfulness" in daily life despite crises, injuries, or chronic disease:

> *For me it has meant a lot to see people who may have experienced crises or changes in life but who have found other paths.* [Study participant 9]

Subtheme 2c: The Possibilities of Artificial Intelligence

Participants frequently identified opportunities for integrating artificial intelligence into the app, imagining an empathetic conversational chatbot capable of helping users process difficult thoughts and feelings through both voice and text. The young adults shared their expectations and hopes for a cutting-edge technological solution (eg, ChatGPT) that would provide a seamless and fully responsive user experience:

> *It should be easy speaking with "it," and that you don't have to correct your phrases all the time. That it understands you while you speak. That it just doesn't straight up read a website from "1177" [Sweden's national health care guide, offering medical information and digital services]. The information you receive should be adjusted and align specifically with the challenge described, and only information related to what you've already told.* [Study participant 3]

Subtheme 2d: Accessibility and Trustworthiness

Another potential concern voiced by all participants was whether it would cost money to use the app. They stressed the importance of making it free of charge, or at most, available for a small fee. Data security was expressed as a key concern, with participants questioning where and by whom confidential content could be accessed or potentially misused. Transparency regarding the measures taken to ensure data security and confidentiality was seen as of high importance for building trust from future users:

> *I think being honest with the user is the most important thing. So that people feel like individuals and not just one in the crowd.* [Study participant 12]

Some of the young adults described their hesitation in using existing digital mental health apps, citing uncertainty or lack of clarity about the companies behind them. They were instinctively more positive and trusting toward digital mental health apps created and offered by public health care agencies or universities.

Theme 3: "Something That Is Designed for Me"—Customizing One's Digital Mental Health Journey

Overview

This theme emphasizes the critical importance of enabling users to tailor the app to their personal needs, preferences, and wishes. The first subtheme, (3a) tools for my mental health needs, pertains to customizable functionality and features that support each individual's unique mental health needs and aspirations. The second subtheme, (3b) choosing the look, feel, and sound, reflects the young adults' desire to control how the digital mental health app interacts with them, including customization of its appearance, tone, and user experience.

Subtheme 3a: Tools for My Mental Health Needs

Ultimately, the study participants envisioned a digital mental health app as not only providing mental health support but also empowering them to navigate the complexities of transitioning to adulthood:

> *The right way to handle one's problem—based on one's own circumstances and issues. And what feels best, what fits best at the moment.* [Study participant 3]

By incorporating peer support, personalized recommendations, and artificial intelligence–driven technology, the app could potentially serve as a hub for addressing the mental health needs of young adults:

> *I have a lot of anxiety when I have to go out, but last time I did it, I called and talked to someone, and it worked. I can try that again.* [Study participant 4]

Some of the young adults highlighted the potential benefits of using digital technology to structure daily activities, manage time, and balance priorities. Examples included checklists; tools for tracking health and activity patterns; and common calendar functions such as reminders and visualizing schedules over days, weeks, or months. Similarly, many of the young adults expressed a desire to ideate, display, track, and follow up on personally meaningful mental well-being goals over time. Gamification elements, such as leveling up or receiving rewards, were also suggested as effective mechanisms to encourage continued use and motivate young adults to pursue their mental health goals.

Subtheme 3b: Choosing the Look, Feel, and Sound

Customization emerged as a key design feature, with participants wanting the app to feel uniquely tailored to them. They requested customizable features such as color schemes, interface layouts, and content preferences; for example, they wanted options for receiving information in various formats—text, audio, or video—based on individual learning styles. Some even suggested options to choose the voice reading information:

> *I remember that when I was feeling very low when I was younger, at my worst moments of anxiety, I couldn't even read. There was just too much swirling around in my head, so I would read the words, but nothing would actually sink in. In those situations, it*

XSL•FO
RenderX

was always easier for me to listen to something. [Study participant 9]

Discussion

Principal Findings

Overview

To the best of our knowledge, this is the first study in Sweden to conduct a needs assessment with young adults as part of coproducing a digital mental health app. In summary, young adults emphasized the importance of nurturing meaningful relationships and activities to support their mental health and well-being. Participants identified several key contributors to mental ill-health during emerging adulthood, including negative social media use, societal pressures, occupational imbalance, and difficulties in managing and recovering from stress and anxiety. They also expressed a positive outlook toward a tailored app addressing these challenges, highlighting its potential to enhance mental health literacy, foster peer support, and serve as a tool for achieving a better balance between what they "wanted to do" and "needed to do." Customization, both in terms of visual design and functional features, was considered essential for fostering user engagement and achieving effective outcomes. The study's 2 guiding research questions structure the discussion in the following subsections.

Swedish Young Adults' Views on the Causes of Mental Health, Well-Being, and Ill-Health

The findings of this study highlight young adults' perspectives on mental health and ill-health, which result from the complex interplay between individuals, groups, and sociocultural phenomena. First, the participants essentially viewed mental health and well-being as a state of being—characterized by the ability to engage in personally meaningful activities and to balance what they wanted to do with what they needed to do in daily life. Furthermore, the findings show that young adults associate mental well-being with creating room for leisure activities and engaging in activities that bring them joy amid stressful career- and success-dominated lives. However, the young adults often struggled with the skills and competencies needed to prioritize these types of activities over work or study obligations.

Research examining young adults' mental health and well-being points to the importance of meaningful routines, habits, and activities as protective factors against mental ill-health [16,42,43] and essential in developing and cultivating a sense of identity as an adult [8,9]. The findings of this study align with the concept of "occupational balance" [44] from the fields of occupational science and therapy, which refers to an individual's experience of maintaining a sufficiently varied and meaningful range of occupations and activities in daily life [45]. Research in this field has shown positive relationships between occupational balance, quality of life, and mental well-being [45,46]. Despite these insights, there is a notable lack of research specifically examining the components of occupational balance among young adults, particularly during emerging adulthood. More qualitative and quantitative research is needed to better understand their experiences and needs at a population level. Given the challenge that emerging adulthood poses to established routines and roles [9], promoting occupational balance through digital tools may offer an untapped opportunity to support young adults in managing their stress, enhancing their life satisfaction, and fostering their identity development.

The findings reveal 2 primary contributors to mental ill-health among young adults. The first is the adverse effects of social media use, which fosters unhealthy comparisons, negative self-image, and self-doubt regarding life choices. The adverse effects of social media use on the study participants' mental ill-health were mainly exacerbated by "doom scrolling" [47], a behavior characterized by compulsive engagement with distressing content. This habit often leads to feelings of entrapment, frustration, and involuntary time loss. Participants often described a cycle of overuse of social media despite recognizing its negative consequences, underscoring the addictive nature of social media and its impact on mental well-being. These patterns align with existing research, which consistently links excessive social media use to heightened psychological distress and a diminished sense of self-worth [48-50]. Of note was the finding that young adults also experienced social media positively. For some, it provided inspiration, connections, and a source of humor in their daily life. Consistent with previous research, our findings highlight the complex meaning and effects of engaging with social media and its impact on the mental health of young adults [51,52]. Collectively, these findings suggest that any future digital mental health app aiming to support the mental health of young adults should focus on helping them to develop a more healthy, conscious relationship with social media through knowledge, competencies, and peer support.

The second contributor to the mental ill-health of the young adults in this study was the challenge of navigating emerging adulthood amid increasing societal uncertainty. The young adults highlighted the connection of societal and sociocultural uncertainty with increased mental ill-health, a phenomenon that can be understood through the concept of "liquid modernity" proposed by Bauman [53]. In this fluid and ever-changing societal landscape, traditional markers of adulthood—such as financial independence, home ownership, and career stability—have become increasingly elusive, exacerbating anxiety, insecurity, and mental ill-health among young adults [3,9,53]. This is driven by economic inequalities, a challenging labor market, digitalization, global crises, and shifting social norms, leaving young adults uncertain about their futures [13]. Young adults report increasing experiences of inadequacy, isolation, and self-doubt among themselves and their peers [3,9]. The expectation to succeed in a world of constant change heightens the risks of burnout from stress, anxiety disorders, and depression. The study participants expressed feeling unprepared for these challenges, grappling with a lack of knowledge and competencies as well as insufficient mental health support at institutional or societal levels that is tailored to their generation's needs. To counter these rapid and often pathogenic societal developments, increasing the individual and group resilience of young adults through efforts to bolster their mental health and well-being may be a viable and impactful solution.

XSL•FO
RenderX

Swedish Young Adults' Preferences and Thoughts on Digital Mental Health Support and Usability

Conducting a needs assessment to research and incorporate the preferences, needs, and suggestions of future users is recommended as a key step in developing eHealth innovations [27]. This approach not only enhances user engagement, acceptability, and use but also helps to identify critical factors for successful implementation [26].

A key finding of this study was that the young adults were positive toward using digital technology as a self-support tool for mental well-being and the prevention of mental ill-health. Similar attitudes and beliefs have been reported in recent studies [54,55]. Previous research [56-58] has demonstrated the efficacy of online tools in enhancing young adults' awareness and competencies in proactively promoting mental health and responding to early signs of worsening mental ill-health. However, most eHealth-based mental health literacy interventions have primarily targeted individuals already experiencing mental ill-health, with an underrepresentation of initiatives developed to strengthening and improving mental well-being and resilience [56]. Thus, there are strong incentives for supporting young adults to handle the stressful life phase of emerging adulthood. Health-promoting efforts can play a vital role in improving the daily lives of young adults in Sweden, thereby helping to prevent suicides and distress as consequences of mental ill-health.

Another key finding of this study points to the importance of enabling personal customization in the functionality and appearance of the digital mental health app. Inviting users to choose from a variety of options – for instance, color schemes, media formats, screen content, or areas of mental health – can significantly increase the likelihood of sustained engagement among young adults. Prior studies support the interactions between customization, user empowerment, and use [59-61]. These findings underscore the need for the digital mental health app to incorporate customization features wherever technologically feasible.

Another mode of customization highlighted by the participating young adults was the use of intuitive and flexible generative artificial intelligence—particularly an empathetic chatbot—to meet their specific mental health literacy needs. Research [62,63] has shown that chatbots can be acceptable and efficacious in supporting mental health, provided they make accurate assumptions and personalize the interactions with each user. However, significant limitations of this technology remain, particularly in adequately understanding and addressing mental health crises, as well as in navigating ethical and privacy concerns [62,63]. Apps that use artificial intelligence should be cautiously coproduced with young adults to harness its capabilities to deliver a personalized user experience, simultaneously accounting for its limitations and associated risks.

Access to peer support through videos, meetings, or forums was identified as a central feature to be built in for inspiration, destigmatization, learning, and a holistic understanding of mental health. Previous research on implemented digital mental health apps has shown them to be effective in providing empowering peer support [64]. One important factor for successful implementation is balancing a moderated safe space with openness to discuss diverse topics [65]. Therefore, the incorporation of feasible and secure software features to enable meaningful and safe peer support should be prioritized in the development of the digital mental health app.

Another key finding of this study is the appreciation young adults expressed when seeing peers share their paths to recovery on social media. Such narratives had a positive impact on their own sense of hopefulness and connectedness. A previous study [66] similarly demonstrated the positive effects of using others' recorded recoveries from mental ill-health to strengthen one's sense of hopefulness and connectedness. In the article "The Lancet Psychiatry Commission on Youth Mental Health," McGorry et al [3] emphasize the importance of viewing emerging adulthood through the lens of the "hero's journey." This perspective frames this life phase as a period of normal struggle and adventure with the goal of gaining maturity and exploring the world, coupled with a "return" or "homecoming" where individuals support and guide others on similar paths. Applying this perspective to mental health challenges and periods of ill-health can, according to the authors, serve to normalize experiences, foster empowerment, and reduce stigma [3]. Other research further supports this approach, finding that the "hero's journey" perspective can increase an individual's sense of agency and provide deeper meaning in life [67]. Therefore, providing stories of other young adults' recovery journeys in the digital mental health app could be valuable for enhancing the mental well-being of young adults.

While the young adults generally perceived a digital mental health app as accessible and promising in meeting their needs, notable caveats exist. Study participants identified risks related to online security and the potential leakage of private information as potential barriers to active use. Transparency regarding how users' data are stored and protected from misuse was emphasized as essential for building trust among young adults.

These findings align with previous research [54,55]. Therefore, to address these concerns, it is crucial to carefully choose a secure storage solution for users' confidential information and to actively communicate how their data are managed and safeguarded.

Strengths and Limitations

In this study, collecting data via semistructured interviews enabled participants' experiences and perspectives to be explored in depth [32]. Furthermore, a clear description of the sociodemographic backgrounds of participants, including age, educational attainment, and employment status, informs judgments about the transferability of the findings [68] and exemplifies the contemporary and diverse life circumstances of Swedish young adults. We believe that the transferability of the findings is further supported by the broad inclusion criteria adopted in this study, which resulted in a diverse sample of Swedish-speaking young adults aged 18 to 29 years who used digital devices daily.

XSL•FO
RenderX

The inclusion of quotes and exemplars representing the majority of study participants further strengthens the trustworthiness of the study, as recommended by Korstjens and Moser [68]. In addition, the credibility of our analysis and results is reinforced by the reflexivity statement presented in the Methods section [35,38], which details the theoretical assumptions underpinning the data analysis and acknowledges the authors' positionality. In regard to this study, the coding and generation of themes were influenced by both the analysts' prior experiences of conducting research in the field of young adults' mental health and their personal or familial experiences of mental ill-health. This may in turn bolster the credibility [68] of the study findings.

The confirmability of the study was further strengthened by the involvement of 3 analysts (MKB, AL, and the ER) in the data analysis process, including reading and discussing the interview transcripts, and with all analysts (MKB, AL, SG, the ER, and BM) actively engaged in developing the codes and themes and subthemes, as well as participating in debriefing sessions. Overall, these processes of intersubjective interpretation during data analysis [38,68] contributed to enhancing the study's credibility.

The findings of the study are limited by the absence of member checking of the transcripts or final analysis with participants, as recommended by Korstjens and Moser [68]. While the first author (MKB) reflected deeply during data collection and analysis, these thoughts were not captured in a reflexive journal [38,68] during the research process, detracting from the confirmability of the findings. Finally, the conformability and neutral framing [68] of the research data may have been weakened by the analysts' prior subject experience and knowledge.

Conclusions

This needs assessment provides valuable insights for coproducing digital mental health apps tailored to young adults. In line with the preferences of Swedish young adults, the promotion of mental health and well-being through digital technology and eHealth should focus on a customizable app that supports balance in daily life while strengthening mental health competencies. The content should center on fostering and maintaining meaningful relationships and activities, addressing challenges such as negative social media use and stress recovery, and enhancing mental health knowledge and peer support. Future efforts should focus on researching young adults' experiences of the life phase of emerging adulthood and its implications for mental health. In addition, future technical development and research on digital mental health apps should include the perspectives of stakeholders, such as mental health professionals, and involve prototype testing with diverse groups to ensure the app's relevance, user engagement, and effectiveness.

Acknowledgments

The authors want to extend their heartfelt gratitude to the participating young adults for generously sharing their time and perspectives during the interviews. Special thanks to Associate Professor Moa Yngve at Linköping University for her excellent collaboration. This research is funded by FORTE (2023-00254), the Swedish research council for health, working life, and welfare. The generative artificial intelligence tool ChatGPT (OpenAI) was used to revise the language and style of early versions of the manuscript, as well as to condense certain paragraphs for clarity and readability.

Data Availability

Due to the sensitivity of research on young adults' mental health, deidentified interview transcripts are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Standards for Reporting Qualitative Research (SRQR) checklist.
[PDF File (Adobe PDF File), 160 KB - formative_v9i1e71563_app1.pdf]

Multimedia Appendix 2
Interview guide for data collection between March and September 2024.
[DOCX File , 43 KB - formative_v9i1e71563_app2.docx]

References

1. Solmi M, Radua J, Olivola M, Croce E, Soardo L, Salazar de Pablo G, et al. Age at onset of mental disorders worldwide: large-scale meta-analysis of 192 epidemiological studies. Mol Psychiatry 2022 Jan;27(1):281-295 [FREE Full text] [doi: 10.1038/s41380-021-01161-7] [Medline: 34079068]

XSL•FO
RenderX

2. Auerbach RP, Mortier P, Bruffaerts R, Alonso J, Benjet C, Cuijpers P, WHO WMH-ICS Collaborators. WHO world mental health surveys international college student project: prevalence and distribution of mental disorders. J Abnorm Psychol 2018 Oct;127(7):623-638 [FREE Full text] [doi: 10.1037/abn0000362] [Medline: 30211576]
3. McGorry PD, Mei C, Dalal N, Alvarez-Jimenez M, Blakemore S, Browne V, et al. The Lancet Psychiatry commission on youth mental health. Lancet Psychiatry 2024 Sep;11(9):731-774. [doi: 10.1016/S2215-0366(24)00163-9] [Medline: 39147461]
4. World mental health report: transforming mental health for all. World Health Organization. URL: https://www.who.int/publications/i/item/9789240049338 [accessed 2024-05-29]
5. Preventing suicide: a global imperative. World Health Organization. URL: https://www.who.int/publications/i/item/9789241564779 [accessed 2024-12-05]
6. Health at a glance: Europe 2018. Organisation for Economic Cooperation and Development. 2018. URL: https://www.oecd.org/en/publications/health-at-a-glance-europe-2018_health_glance_eur-2018-en.html [accessed 2024-12-05]
7. Arias D, Saxena S, Verguet S. Quantifying the global burden of mental disorders and their economic value. EClinicalMedicine 2022 Dec;54:101675. [doi: 10.1016/j.eclinm.2022.101675] [Medline: 36193171]
8. Begrepp inom området psykisk hälsa Version 1.1. Swedish National Board of Health and Welfare. 2024. URL: https://www.socialstyrelsen.se/globalassets/sharepoint-dokument/artikelkatalog/ovrigt/2024-6-9186.pdf [accessed 2024-12-05]
9. Att inte bara överleva utan att faktiskt också leva: En kartläggning om ungas psykiska hälsa – KORTVERSION. The Public Health Agency of Sweden. 2023. URL: https://www.folkhalsomyndigheten.se/contentassets/66fd62dfd17141f28b8e03316d295952/att-inte-bara-overleva-utan-att-faktiskt-ocksa-leva.pdf
10. Arnett JJ, Žukauskienė R, Sugimura K. The new life stage of emerging adulthood at ages 18-29 years: implications for mental health. Lancet Psychiatry 2014 Dec;1(7):569-576. [doi: 10.1016/S2215-0366(14)00080-7] [Medline: 26361316]
11. Thornicroft G, Sunkel C, Alikhon Aliev A, Baker S, Brohan E, El Chammay R, et al. The Lancet Commission on ending stigma and discrimination in mental health. Lancet 2022 Oct 22;400(10361):1438-1480. [doi: 10.1016/S0140-6736(22)01470-2] [Medline: 36223799]
12. Jorm AF. Why we need the concept of "mental health literacy". Health Commun 2015 Sep 15;30(12):1166-1168. [doi: 10.1080/10410236.2015.1037423] [Medline: 26372027]
13. Kirkbride JB, Anglin DM, Colman I, Dykxhoorn J, Jones PB, Patalay P, et al. The social determinants of mental health and disorder: evidence, prevention and recommendations. World Psychiatry 2024 Feb 12;23(1):58-90 [FREE Full text] [doi: 10.1002/wps.21160] [Medline: 38214615]
14. Blom S, Lindh F, Lundin A, Burström B, Hensing G, Löve J. How gender and low mental health literacy are related to unmet need for mental healthcare: a cross-sectional population-based study in Sweden. Arch Public Health 2024 Jan 25;82(1):12 [FREE Full text] [doi: 10.1186/s13690-023-01228-7] [Medline: 38273389]
15. Brunette MF, Erlich MD, Edwards ML, Adler DA, Berlant J, Dixon L, et al. Addressing the increasing mental health distress and mental illness among young adults in the United States. J Nerv Ment Dis 2023 Dec 01;211(12):961-967. [doi: 10.1097/NMD.0000000000001734] [Medline: 38015186]
16. Thomas SA, Browning CJ, Charchar FJ, Klein B, Ory MG, Bowden-Jones H, et al. Transforming global approaches to chronic disease prevention and management across the lifespan: integrating genomics, behavior change, and digital health solutions. Front Public Health 2023 Oct 13;11:1248254 [FREE Full text] [doi: 10.3389/fpubh.2023.1248254] [Medline: 37905238]
17. Kählke F, Buntrock C, Smit F, Ebert DD. Systematic review of economic evaluations for internet- and mobile-based interventions for mental health problems. NPJ Digit Med 2022 Nov 23;5(1):175 [FREE Full text] [doi: 10.1038/s41746-022-00702-w] [Medline: 36424463]
18. Afsharnejad B, Milbourn B, Hayden-Evans M, Baker-Young E, Black MH, Thompson C, et al. The efficacy of the "Talk-to-Me" suicide prevention and mental health education program for tertiary students: a crossover randomised control trial. Eur Child Adolesc Psychiatry 2023 Dec 04;32(12):2477-2489 [FREE Full text] [doi: 10.1007/s00787-022-02094-4] [Medline: 36194311]
19. Yeo G, Reich SM, Liaw NA, Chia EY. The effect of digital mental health literacy interventions on mental health: systematic review and meta-analysis. J Med Internet Res 2024 Feb 29;26:e51268 [FREE Full text] [doi: 10.2196/51268] [Medline: 38421687]
20. Wang Q, Zhang W, An S. A systematic review and meta-analysis of Internet-based self-help interventions for mental health among adolescents and college students. Internet Interv 2023 Dec;34:100690 [FREE Full text] [doi: 10.1016/j.invent.2023.100690] [Medline: 38023965]
21. Serlachius A, Boggiss A, Lim D, Schache K, Wallace-Boyd K, Brenton-Peters J, et al. Pilot study of a well-being app to support New Zealand young people during the COVID-19 pandemic. Internet Interv 2021 Dec;26:100464 [FREE Full text] [doi: 10.1016/j.invent.2021.100464] [Medline: 34631433]
22. Digitalt stöd för psykisk hälsa: Möjligheter och hinder, Fokusrapport. Swedish eHealth Agency. 2021. URL: https://www.ehalsomyndigheten.se/globalassets/ehm/3_om-oss/rapporter/fokusrapport---digitalt-stod-for-psykisk-halsa.pdf [accessed 2024-12-05]

XSL•FO
RenderX

23. Bell J, Lim A, Williams R, Girdler S, Milbourn B, Black M. ‘Nothing about us without us': co-production ingredients for working alongside stakeholders to develop mental health interventions. Adv Ment Health 2023;21(1):4-16 [FREE Full text] [doi: 10.1080/18387357.2021.2020143]
24. Bakker D, Kazantzis N, Rickwood D, Rickard N. Mental health smartphone apps: review and evidence-based recommendations for future developments. JMIR Ment Health 2016 Mar 01;3(1):e7 [FREE Full text] [doi: 10.2196/mental.4984] [Medline: 26932350]
25. Working well together: evidence and tools to enable co-production in mental health commissioning. National Collaborating Centre for Mental Health. 2019. URL: https://www.rcpsych.ac.uk/docs/default-source/improving-care/nccmh/working-well-together/working-well-together---evidence-and-tools-to-enable-co-production-in-mental-health-commissioning.pdf?sfvrsn=4e2924c1_6 [accessed 2024-12-05]
26. Schweitzer R, Schlögl S, Schweitzer M. Technology-supported behavior change-applying design thinking to mHealth application development. Eur J Investig Health Psychol Educ 2024 Mar 07;14(3):584-608 [FREE Full text] [doi: 10.3390/ejihpe14030039] [Medline: 38534900]
27. Skivington K, Matthews L, Simpson SA, Craig P, Baird J, Blazeby JM, et al. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. Int J Nurs Stud 2024 Jun;154:104705. [doi: 10.1016/j.ijnurstu.2024.104705] [Medline: 38564982]
28. Altman M, Huang TT, Breland JY. Design thinking in health care. Prev Chronic Dis 2018 Sep 27;15:E117 [FREE Full text] [doi: 10.5888/pcd15.180128] [Medline: 30264690]
29. Babbage C, Jackson GM, Nixon E. Desired features of a digital technology tool for self-management of well-being in a nonclinical sample of young people: qualitative study. JMIR Ment Health 2018 Dec 18;5(4):e10067 [FREE Full text] [doi: 10.2196/10067] [Medline: 30563820]
30. Park SY, Yourell J, McAlister KL, Huberty J. Exploring generation Z and young millennials' perspectives of a spiritual self-care app and their spiritual identity (Skylight): qualitative semistructured interview study. JMIR Form Res 2023 Dec 28;7:e54284 [FREE Full text] [doi: 10.2196/54284] [Medline: 38064199]
31. Stoyanov SR, Zelenko O, Staneva A, Kavanagh DJ, Smith C, Sade G, et al. Development of the niggle app for supporting young people on their dynamic journey to well-being: co-design and qualitative research study. JMIR Mhealth Uhealth 2021 Apr 20;9(4):e21085 [FREE Full text] [doi: 10.2196/21085] [Medline: 33877050]
32. Colorafi KJ, Evans B. Qualitative descriptive methods in health science research. HERD 2016 Jul;9(4):16-25 [FREE Full text] [doi: 10.1177/1937586715614171] [Medline: 26791375]
33. Meyerhoff J, Kornfield R, Lattie EG, Knapp AA, Kruzan KP, Jacobs M, et al. From formative design to service-ready therapeutic: a pragmatic approach to designing digital mental health interventions across domains. Internet Interv 2023 Dec;34:100677 [FREE Full text] [doi: 10.1016/j.invent.2023.100677] [Medline: 37808416]
34. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014 Sep;89(9):1245-1251 [FREE Full text] [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]
35. Moser A, Korstjens I. Series: practical guidance to qualitative research. Part 3: sampling, data collection and analysis. Eur J Gen Pract 2018 Dec 04;24(1):9-18 [FREE Full text] [doi: 10.1080/13814788.2017.1375091] [Medline: 29199486]
36. Posselt J, Baumann E, Dierks ML. A qualitative interview study of patients' attitudes towards and intention to use digital interventions for depressive disorders on prescription. Front Digit Health 2024 Feb 5;6:1275569 [FREE Full text] [doi: 10.3389/fdgth.2024.1275569] [Medline: 38375490]
37. Taher R, Stahl D, Shergill S, Yiend J. The safety of digital mental health interventions: findings and recommendations from a qualitative study exploring users' experiences, concerns, and suggestions. JMIR Hum Factors 2025 Feb 07;12:e62974 [FREE Full text] [doi: 10.2196/62974] [Medline: 39919292]
38. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101 [FREE Full text] [doi: 10.1191/1478088706qp063oa]
39. Moon K, Blackman D. A guide to understanding social science research for natural scientists. Conserv Biol 2014 Oct 24;28(5):1167-1177. [doi: 10.1111/cobi.12326] [Medline: 24962114]
40. World Medical Association. World medical association declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA 2013 Nov 27;310(20):2191-2194. [doi: 10.1001/jama.2013.281053] [Medline: 24141714]
41. Thompson AR, Chambers E. Ethical issues in qualitative mental health research. In: Harper D, Thompson AR, editors. Qualitative research methods in Mental Health and Psychotherapy. Chichester, UK: Wiley-Blackwell; 2011:23-37.
42. Yung JY, Wong V, Ho GW, Molassiotis A. Understanding the experiences of hikikomori through the lens of the CHIME framework: connectedness, hope and optimism, identity, meaning in life, and empowerment; systematic review. BMC Psychol 2021 Jul 10;9(1):104 [FREE Full text] [doi: 10.1186/s40359-021-00605-7] [Medline: 34246318]
43. Gunnarsson AB, Hedberg AK, Håkansson C, Hedin K, Wagman P. Occupational performance problems in people with depression and anxiety. Scand J Occup Ther 2023 Feb 11;30(2):148-158 [FREE Full text] [doi: 10.1080/11038128.2021.1882562] [Medline: 33569993]

XSL•FO
RenderX

44. Eklund M, Parsonage-Harrison J, Argentzell E. Occupation- and lifestyle-based mental health interventions - a hallmark for the occupational therapy profession? Br J Occup Ther 2024 Jul;87(7):395-397 [FREE Full text] [doi: 10.1177/03080226241228750] [Medline: 40336722]
45. Håkansson C, Ahlborg Jr G. Occupational imbalance and the role of perceived stress in predicting stress-related disorders. Scand J Occup Ther 2018 Jul 02;25(4):278-287 [FREE Full text] [doi: 10.1080/11038128.2017.1298666] [Medline: 28276962]
46. Lillo-Navarro C, Fernández-Pires P, Benavides Gil G, Martínez-Zaragoza F, Chaves C, Roca P, et al. Effects of a mindfulness-based program on the occupational balance and mental health of university students. Protocol for a randomized controlled trial. PLoS One 2024 May 2;19(5):e0302018 [FREE Full text] [doi: 10.1371/journal.pone.0302018] [Medline: 38696406]
47. Satici SA, Gocet Tekin E, Deniz ME, Satici B. Doomscrolling scale: its association with personality traits, psychological distress, social media use, and wellbeing. Appl Res Qual Life 2023 Oct 19;18(2):833-847 [FREE Full text] [doi: 10.1007/s11482-022-10110-7] [Medline: 36275044]
48. Zubair U, Khan MK, Albashari M. Link between excessive social media use and psychiatric disorders. Ann Med Surg (Lond) 2023 Apr;85(4):875-878 [FREE Full text] [doi: 10.1097/MS9.0000000000000112] [Medline: 37113864]
49. Brand C, Fochesatto CF, Gaya AR, Schuch FB, López-Gil JF. Scrolling through adolescence: unveiling the relationship of the use of social networks and its addictive behavior with psychosocial health. Child Adolesc Psychiatry Ment Health 2024 Aug 31;18(1):107 [FREE Full text] [doi: 10.1186/s13034-024-00805-0] [Medline: 39217325]
50. Seabrook EM, Kern ML, Rickard NS. Social networking sites, depression, and anxiety: a systematic review. JMIR Ment Health 2016 Nov 23;3(4):e50 [FREE Full text] [doi: 10.2196/mental.5842] [Medline: 27881357]
51. Zsila Á, Reyes ME. Pros and cons: impacts of social media on mental health. BMC Psychol 2023 Jul 06;11(1):201 [FREE Full text] [doi: 10.1186/s40359-023-01243-x] [Medline: 37415227]
52. Plackett R, Sheringham J, Dykxhoorn J. The longitudinal impact of social media use on UK adolescents' mental health: longitudinal observational study. J Med Internet Res 2023 Mar 24;25:e43213 [FREE Full text] [doi: 10.2196/43213] [Medline: 36961482]
53. Bauman Z. Liquid Modernity. Oxfordshire, UK: Wiley-Blackwell; 2000.
54. Ito-Jaeger S, Perez Vallejos E, Logathasan S, Curran T, Crawford P. Young people's trust in cocreated web-based resources to promote mental health literacy: focus group study. JMIR Ment Health 2023 Jan 09;10:e38346 [FREE Full text] [doi: 10.2196/38346] [Medline: 36622752]
55. Riboldi I, Calabrese A, Piacenti S, Capogrosso CA, Paioni SL, Bartoli F, et al. Understanding university students' perspectives towards digital tools for mental health support: a cross-country study. Clin Pract Epidemiol Ment Health 2024 Feb 21;20(1):e17450179271467 [FREE Full text] [doi: 10.2174/0117450179271467231231060255] [Medline: 38660572]
56. Tian L, Wong EL, Dong D, Cheung AW, Chan SK, Cao Y, et al. Improving mental health literacy using web- or app-based interventions: a scoping review. Digit Health 2024 Apr 05;10:20552076241243133 [FREE Full text] [doi: 10.1177/20552076241243133] [Medline: 38585191]
57. Kerber A, Beintner I, Burchert S, Knaevelsrud C. Effects of a self-guided transdiagnostic smartphone app on patient empowerment and mental health: randomized controlled trial. JMIR Ment Health 2023 Nov 06;10:e45068 [FREE Full text] [doi: 10.2196/45068] [Medline: 37930749]
58. Saboor S, Medina A, Marciano L. Application of positive psychology in digital interventions for children, adolescents, and young adults: systematic review and meta-analysis of controlled trials. JMIR Ment Health 2024 Aug 14;11:e56045 [FREE Full text] [doi: 10.2196/56045] [Medline: 39141906]
59. Kozelka EE, Acquilano SC, Al-Abdulmunem M, Guarino S, Elwyn G, Drake RE, et al. Digital mental health and its discontents: assumptions about technology that create barriers to equitable access. Psychiatr Serv 2024 Mar 01;75(3):299-302. [doi: 10.1176/appi.ps.20230238] [Medline: 38050440]
60. Ho TQ, Le LK, Engel L, Le N, Melvin G, Le HN, et al. Barriers to and facilitators of user engagement with web-based mental health interventions in young people: a systematic review. Eur Child Adolesc Psychiatry 2025 Jan 14;34(1):83-100. [doi: 10.1007/s00787-024-02386-x] [Medline: 38356043]
61. Saleem M, Kühne L, De Santis KK, Christianson L, Brand T, Busse H. Understanding engagement strategies in digital interventions for mental health promotion: scoping review. JMIR Ment Health 2021 Dec 20;8(12):e30000 [FREE Full text] [doi: 10.2196/30000] [Medline: 34931995]
62. Balan R, Dobrean A, Poetar CR. Use of automated conversational agents in improving young population mental health: a scoping review. NPJ Digit Med 2024 Mar 19;7(1):75 [FREE Full text] [doi: 10.1038/s41746-024-01072-1] [Medline: 38503909]
63. Xue J, Zhang B, Zhao Y, Zhang Q, Zheng C, Jiang J, et al. Evaluation of the current state of chatbots for digital health: scoping review. J Med Internet Res 2023 Dec 19;25:e47217 [FREE Full text] [doi: 10.2196/47217] [Medline: 38113097]
64. Gammon D, Strand M, Eng LS, Børøsund E, Varsi C, Ruland C. Shifting practices toward recovery-oriented care through an e-recovery portal in community mental health care: a mixed-methods exploratory study. J Med Internet Res 2017 May 02;19(5):e145 [FREE Full text] [doi: 10.2196/jmir.7524] [Medline: 28465277]

XSL•FO
RenderX

65. Marshall P, Booth M, Coole M, Fothergill L, Glossop Z, Haines J, et al. Understanding the impacts of online mental health peer support forums: realist synthesis. JMIR Ment Health 2024 May 09;11:e55750 [FREE Full text] [doi: 10.2196/55750] [Medline: 38722680]
66. Ng F, Newby C, Robinson C, Llewellyn-Beardsley J, Yeo C, Roe J, et al. How do recorded mental health recovery narratives create connection and improve hopefulness? J Ment Health 2022 Apr 05;31(2):273-280 [FREE Full text] [doi: 10.1080/09638237.2021.2022627] [Medline: 34983300]
67. Rogers BA, Chicas H, Kelly JM, Kubin E, Christian MS, Kachanoff FJ, et al. Seeing your life story as a Hero's Journey increases meaning in life. J Pers Soc Psychol 2023 Oct;125(4):752-778. [doi: 10.1037/pspa0000341] [Medline: 36972106]
68. Korstjens I, Moser A. Series: practical guidance to qualitative research. Part 4: trustworthiness and publishing. Eur J Gen Pract 2018 Dec;24(1):120-124 [FREE Full text] [doi: 10.1080/13814788.2017.1375092] [Medline: 29202616]

Abbreviations

ER: external researcher

Edited by A Mavragani; submitted 22.01.25; peer-reviewed by E Tengelin; comments to author 21.03.25; revised version received 11.04.25; accepted 05.05.25; published 23.05.25.

Please cite as:
Karaba Bäckström M, Girdler S, Milbourn B, Lexén A
Needs and Preferences of Swedish Young Adults for a Digital App Promoting Mental Health Literacy, Occupational Balance, and Peer Support: Qualitative Interview Study
JMIR Form Res 2025;9:e71563
URL: https://formative.jmir.org/2025/1/e71563
doi:10.2196/71563
PMID:40406829

©Martin Karaba Bäckström, Sonya Girdler, Ben Milbourn, Annika Lexén. Originally published in JMIR Formative Research (https://formative.jmir.org), 23.05.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Unlocking the Potential of Wear Time of a Wearable Device to Enhance Postpartum Depression Screening and Detection: Cross-Sectional Study

Eric Hurwitz[1], PhD; Samantha Meltzer-Brody[2], MPH, MD; Zachary Butzin-Dozier[3], PhD; Rena C Patel[4], MPhil, MPH, MD; Noémie Elhadad[5], PhD; Melissa A Haendel[1], PhD

[1]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States
[2]Department of Psychiatry, University of North Carolina School of Medicine, Chapel Hill, NC, United States
[3]Division of Biostatistics, School of Public Health, University of California, Berkeley, Berkeley, CA, United States
[4]Department of Infectious Disease, University of Alabama at Birmingham, Birmingham, AL, United States
[5]Department of Biomedical Informatics, Columbia University Irving Medical Center, New York, NY, United States

Corresponding Author:
Eric Hurwitz, PhD
Department of Genetics
University of North Carolina at Chapel Hill
120 Mason Farm Road
Chapel Hill, NC, 27599
United States
Phone: 1 919 843 8724
Email: eric_hurwitz@med.unc.edu

Abstract

Background: Postpartum depression (PPD) is a mood disorder affecting 1 in 7 women after childbirth that is often underscreened and underdetected. If not diagnosed and treated, PPD is associated with long-term developmental challenges in the child and maternal morbidity. Wearable technologies, such as smartwatches and fitness trackers (eg, Fitbit), offer continuous and longitudinal digital phenotyping for mood disorder diagnosis and monitoring, with device wear time being an important yet understudied aspect.

Objective: We aimed to suggest that wear time of a wearable device may provide additional information about perinatal mental health to facilitate screening and early detection of PPD. We proposed that wear time of a wearable device may also be valuable for managing other mental health disorders.

Methods: Using the *All of Us* Research Program dataset, we identified females who experienced childbirth with and without PPD using computational phenotyping. We compared the percentage of days and number of hours per day females with and without PPD wore Fitbit devices during prepregnancy, pregnancy, postpartum, and PPD periods, determined by electronic health records. Comparisons between females with and without PPD were conducted using linear regression models. We also assessed the correlation between Fitbit wear time consistency (measured as the maximum number of consecutive days the Fitbit was worn) during prepregnancy and PPD periods in females with and without PPD using the Pearson correlation. All analyses were run with Bonferroni correction.

Results: Our findings showed a strong trend, although nonsignificant after multiple testing correction, that females in the PPD cohort wore their Fitbits more than those in non-PPD cohort during the postpartum (PPD cohort: mean 69.9%, 95% CI 42.7%-97%; non-PPD cohort: mean 50%, 95% CI 25.5%-74.4%; P=.02) and PPD periods (PPD cohort: mean 66.6%, 95% CI 37.9%-95.3%; non-PPD cohort: mean 46.4%, 95% CI 20.5%-72.2%; P=.02). We found no difference in the number of hours per day females in the PPD and non-PPD cohorts wore their Fitbit during any period of pregnancy. Finally, there was no relationship between the consistency of Fitbit wear time during prepregnancy and PPD periods (r=–0.05, 95% CI –0.46 to 0.38; P=.84); however, there was a trend, though nonsignificant, in Fitbit wear time consistency among females without PPD (r=0.25, 95% CI –0.02 to 0.49; P=.07).

Conclusions: We hypothesize that increased Fitbit wear time among females with PPD may be attributed to hypervigilance, given the common co-occurrence of anxiety symptoms. Future studies should assess the link between PPD, hypervigilance, and

XSL•FO
RenderX

wear time patterns. We envision that wear time patterns of a wearable device combined with digital biomarkers such as sleep and physical activity could enhance early PPD detection using machine learning by alerting clinicians to potential concerns and facilitating timely screenings, which may have implications for other mental health disorders.

(JMIR Form Res 2025;9:e67585) doi:10.2196/67585

KEYWORDS

wearable device; All of Us; postpartum depression; Fitbit; wear time; screening

Introduction

Background

The rise of wearable device ownership, such as Fitbits, has led to significant advancements in the realm of digital phenotyping [1]. Because wearables can be used to monitor the same individual in a continuous and longitudinal manner, their use for personalized medicine is exciting, especially for mental health where individualized tools for diagnosis and treatment monitoring are lacking. Digital biomarkers from wearables are collected in a passive manner in nonclinical settings, thus enabling these devices to offer potential enhancements to several clinical aspects of the mental health care continuum [2].

Postpartum depression (PPD) is a mood disorder that is one of the most common complications of childbirth [3]. PPD has significant implications for maternal morbidity, associations with developmental delays for the child, and incurs significant costs to society [4-7]. Because PPD is a highly heterogeneous condition and often stigmatized, many patients go undiagnosed [8]. One significant issue with PPD is that most women do not receive sufficient screening, as only about 31% of women with PPD receive a diagnosis [4]. As noted by Cox et al [4], there are reliable screening instruments for PPD (eg, the Edinburgh-Postnatal Depression Scale [EPDS]) [9] and specific treatments for PPD (eg, brexanalone and zuranolone) [10,11]; yet screening and diagnosis of PPD lag behind and novel approaches for PPD detection are direly needed.

Wearables have provided an opportunistic route for enhanced behavioral phenotyping during pregnancy and in the postpartum period, including for PPD [2,12,13]. Given the underdiagnostic rate of PPD, readily available consumer wearables may aid in its early detection due to their rise in ownership and passive data collection, thereby improving patient outcomes. For example, our recent work demonstrated that individualized machine learning (ML) models using digital biomarkers (heart rate, physical activity, and energy expenditure) from a Fitbit were able to distinguish between 4 phases of pregnancy, including prepregnancy, pregnancy, postpartum, and during PPD [12]. Wearable devices have also been shown to predict whether a woman will experience preterm birth using only 1 week of activity and sleep data [13]. In addition, studies have demonstrated that activity intensity distribution during the day, resting heart rate, and heart rate variability captured from a wearable device were predictive of maternal loneliness, which is associated with PPD [14]. Collectively, these studies highlight a relationship between digital biomarkers and perinatal mental health, suggesting that wearables may enhance longitudinal monitoring.

While it has been shown that digital biomarkers from wearables, such as Fitbit, combined with ML can provide insight into mental health conditions, patterns of wear time remain relatively unexplored. Previous studies exploring wear time of a wearable device have mainly taken place in the human-computer interaction field in a general population and disease-agnostic setting [15-19]. A few studies have looked at wear time behavior in the context of biomedical research, but only in a limited capacity. For instance, analyses from the Framingham Heart Study suggest that higher depressive symptoms are associated with lower smartwatch use, defined as wearing the device for more than 5 hours at least one day of the week. The authors suggest this observation is due to the link between motivation and depressive symptoms, where individuals are less likely to engage with a smartwatch for health-related activities such as tracking daily steps or promoting healthy lifestyle behaviors [20]. While this result posits a relationship between device wear time and mental health, there is a need to explore wear time of a wearable device in pregnancy cohorts, which are heterogeneous and constantly changing, making it difficult to identify potential screening tools and biomarkers.

Objectives

In this study, we sought to demonstrate the value of wear time of a wearable device as an additional insightful digital biomarker for mental health. Our previous work demonstrated that several digital biomarkers were altered during PPD relative to other periods [12]; therefore, we next wanted to assess differences in wear time of a wearable device across multiple periods of pregnancy. We leveraged the *All of Us* Research Program (AoURP) dataset, a longitudinal, observational dataset with several health-related data types, including electronic health records (EHRs), surveys, physical measurements, and Fitbit data [21]. To highlight the potential value of Fitbit wear time in facilitating early detection of PPD, we characterized differences in wear time between females with and without PPD. We proposed that wear time of a wearable device may contribute to serving as a clinically informative biomarker to help facilitate early detection of mental health disorders in a continuous, passive, and nonclinical setting. For PPD specifically, gaining insight into wearable behavior patterns could offer valuable understanding of perinatal mental health, potentially enhancing screening and diagnosis in real-world settings.

Methods

Data Sources and Platforms

Data in this study leveraged the AoURP Controlled Tier version 7 dataset. Analysis was conducted using the AoURP Researcher

XSL•FO
RenderX

Workbench cloud platform. All phenotyping and data analysis were conducted using R software (R Foundation for Statistical Computing). Fitbit data in the AoURP operates under a bring-your-own-device model, where participants who consent to participate in the study share data from their device that they already own [22].

Study Setting, Clinical Setting, and Recruitment Procedures by AoURP

AoURP is a longitudinal, observational study [21]. The data used in this study were obtained from the AoURP, which was responsible for all study and recruitment procedures described in the AoURP Operational Protocol [23]. To briefly summarize study and recruitment procedures from the protocol, anyone who lives in the United States (or territory of the United States) is eligible to enroll in the AoURP. The program emphasizes recruitment of minority populations that have been underrepresented in biomedical research historically. Specific inclusion criteria consist of (1) adults ≥18 years with the ability to provide consent and (2) individuals who currently reside in the United States. The only exclusion criteria include individuals who are incarcerated at the time of enrollment. Recruitment is conducted through targeted advertising (ie, print flyers, brochures, posters, Television, radio, web, mobile, billboards, bus advertisements, email, and snail mail); personal interest groups (ie, social media, community events, and press coverage); and directly at health care provider organizations (HPOs) or direct volunteer (DV) partner sites (ie, waiting areas, regular course of clinical care at HPOs, local informational events, and regional informational events organized by research program awardees, HPOs, or DV partners). Participants enroll to participate in the program through the *All of Us* website or a smartphone app, go through electronic consent modules, and watch explanatory videos with text, icons, and formative questions. Once consented, participants are given baseline health surveys that each take about 15 minutes to complete. In addition, participants then provide authorization to share EHR data and are provided the opportunity to share additional physical measurements and biospecimens at an HPO or DV [21]. The reimbursement procedures are described in the following section.

Ethical Considerations

The protocol for the AoURP study was reviewed by the institutional review board of the AoURP (protocol 2021-02-TN-001). The institutional review board follows the regulations and guidance of the National Institutes of Health Office for Human Research Protections for all studies, ensuring that the rights and welfare of research participants are overseen and protected uniformly. The informed consent process states that participants have the option to withdraw at any time. Privacy of participant data is maintained in the following three ways: (1) storing data on protected computers, (2) preventing researchers from seeing identifiable patient information, such as name or social security number, and (3) having researchers sign a contract declaring that they will not try to identify participants. Furthermore, access to the AoURP dataset is only available through the Researcher Workbench, which is only accessible to researchers who have completed the requisite training at institutions with a signed data use agreement. For compensation, participants are offered US $25 one-time in the form of cash, gift card, or an electronic voucher if they are asked and decide to go to an *All of Us* partner center for physical measurements to give blood, saliva, or urine samples. Notably, other racial or ethnic groups—including Asian non-Hispanic, Black non-Hispanic, Hispanic or Latinx of any race, more than one population, none of these, and those who skipped the reporting—were not reported in this study because the sample sizes for several of them were less than 20 and could risk patient reidentification, which violates the AoURP Dissemination policy [24].

Computational Phenotyping of PPD and Non-PPD Cohorts

Females were assigned to the PPD cohort using the same method that we described previously [12,25]. Briefly, identifying females with PPD consists of a three-fold approach: (1) a PPD diagnosis, (2) a diagnosis of depression during the postpartum period, or (3) antidepressant drug exposure during the postpartum period, which is consistent with other computational phenotyping approaches for PPD [26]. Females were assigned to the non-PPD cohort by identifying those with available pregnancy or delivery EHR data in a similar manner to the PPD cohort and then excluding those who were in the PPD cohort.

To assess wear time behavior in a longitudinal manner, Fitbit wear time data for each female in the PPD cohort were assigned to one of the four periods: (1) prepregnancy (starting from 2 years before the PPD index date), (2) pregnancy, (3) postpartum without depression (after the delivery date and before the PPD diagnosis date), or (4) PPD (a diagnosis up to 24 months from the date of delivery, which has been done in previous work and in this study also represents a period) [27,28]. The PPD period ranged from 14 days before the index date to 30 days after the index date, which was selected because (1) the diagnostic criteria for PPD requires displaying 5 depressive symptoms lasting 2 weeks and (2) some individuals received antidepressant medication on the same date as their index date, which can take effect after 4 weeks of use [29,30].

Because females in the non-PPD group did not undergo a *fourth* phase of PPD as seen in the PPD group, we introduced a pseudotime period called the *PPD-equivalent* period as a time frame for the non-PPD group to align with the PPD period. The index date for the PPD-equivalent period was set at 58 days following delivery, corresponding to the median number of days after delivery of PPD diagnosis among females in the PPD group, following the same strategy we used in our previous work. Similarly, 14 days before the index date was not used as these females did not actually experience PPD [12]. Females were only included in the PPD or non-PPD cohorts if they had any Fitbit data during any of the 4 periods.

Covariates

Models were run with covariates of age at PPD diagnosis (or age at the index date for the non-PPD cohort), race/ethnicity, and annual income, which was determined from self-reported responses in the Basics survey as a measure of socioeconomic status [31]. Potential responses for annual income consisted of

XSL•FO
RenderX

the following: (1) <US $10,000, (2) US $10,000-US $24,999, (3) US $25,000-US $34,999, (4) US $35,000-US $49,999, (5) US $50,000-US $74,999, (6) US $75,000-US $99,999, (7) US $100,000-US $149,999, (8) US $150,000-US $199,999, (9) >US $200,000, or (10) prefer not to answer. Notably, previous history of mental health disorders was intentionally excluded as a covariate because its prevalence was significantly higher in the PPD cohort than in the non-PPD cohort. Including it in the model could result in unreliable estimates and therefore we intentionally chose not to adjust for it.

Outcome of Measuring Fitbit Wear Time in PPD and Non-PPD Cohorts

Fitbit wear time was measured by first determining the number of hours the device was worn in a day using methods described previously [32]. Previous studies have indicated that a *valid* day of smartwatch data requires 10 hours of wear time and between 100 and 45,000 steps [32]. In this study, rather than analyzing days of *valid* data, we wanted to understand patterns of Fitbit wear time behavior among females with PPD. Hence, we established a binary variable for each day to indicate whether the device was worn or not based on the presence of at least 1 hour of wear time, where hours of wear time were measured based on the presence of step data, similar to prior studies [32]. We then determined the percentage of days the Fitbit was worn during each of the 4 periods (ie, prepregnancy, pregnancy, postpartum, and PPD [or PPD-equivalent for the non-PPD cohort]) by counting the number of days the device was worn divided by the total number of days during that period for each female. The total number of days was determined for each person by filtering data after the first recorded date of any Fitbit data to ensure that we were not labeling someone as not wearing their Fitbit if they did not own one at the time.

Outcome of Measuring the Number of Hours Fitbit Devices Were Worn Per Day in PPD and Non-PPD Cohorts

We determined the number of hours per day Fitbit was worn using the same logic as described in aforementioned section. The dataset was filtered on individuals who had at least 1 hour of wear time, as we wanted to ensure that we were assessing whether there was a difference in the number of hours the device was worn per day among days that the device was actually worn.

Outcome of Measuring the Percentage of Days Fitbits Were Worn to Sleep in PPD and Non-PPD Cohorts

To assess how often females with PPD wore their device to sleep, we focused on whether females had any record of *main sleep* for each date as a binary variable for yes or no. We then determined the percentage of days the Fitbit was worn to sleep during each of the 4 periods (ie, prepregnancy, pregnancy, postpartum, and PPD [or PPD-equivalent for the non-PPD cohort]) by counting the number of days the device was worn to sleep divided by the total number of days during that period for each female.

Statistical Analysis Comparing the Percentage of Wear Time, Hours of Wear Time Per Day, and Percentage of Days Fitbits Were Worn to Sleep

The percentage of Fitbit wear time was compared between PPD and non-PPD cohorts using linear regression, where 4 total models were run (one for each period). Models were run with covariates of age at PPD diagnosis (or age at the index date for the non-PPD cohort), race/ethnicity, and annual income. Each model filtered data during one period and the means were calculated using the emmeans() function [33]. Similar methodology was used for comparing the percentage of days Fitbits were worn to sleep among PPD and non-PPD cohorts. For comparing the hours per day of Fitbit wear time between PPD and non-PPD cohorts, we ran a linear mixed-effects model using the lme4 package in R because there were multiple days of data per person (ie, person ID was included as the random effect) [34,35]. All models were run at a significance level of P=.05; however, because 4 statistical tests were performed in each analysis (ie, one for each period of pregnancy), we performed Bonferroni correction [36] creating a corrected significance level of P=.013.

Assessing the Correlation Between Device Wear Time Before and During PPD

The correlation between device wear time before and during PPD was performed by filtering on 2 periods of interest (eg, prepregnancy and PPD) and assessing the correlation among all females in the PPD cohort. The *ggpubr* package was used to determine the correlation and was evaluated at a significance level of P=.05, which was adjusted to a threshold of P=.013 using Bonferroni correction. The same analysis was conducted for comparing device use during pregnancy and PPD periods as well. Both analyses were repeated in females without PPD for comparison.

Assessing the Correlation Between Device Wear Time Consistency Before and During PPD

Device wear time consistency was measured by determining the maximum number of consecutive days the Fitbit was worn for each period for each unique person. Device wear time for one day was defined using the same definition as before, where we considered an individual wore the device if they had at least 1 hour of wear time for each date based on the presence of step data [32]. We then determined the relationship between device wear time consistency during prepregnancy and PPD by calculating the Pearson correlation coefficient between the maximum number of days the device was worn during the PPD period versus the prepregnancy period at a significance level of P=.05, which was adjusted to a threshold of P=.013 using Bonferroni correction, using the *ggpubr* package in R [37,38]. We also performed the same analysis replacing the prepregnancy period with the pregnancy period to assess the correlation between device wear time consistency during pregnancy and PPD. This process was repeated in the non-PPD cohort for comparison.

XSL•FO
RenderX

Sensitivity Analysis Comparing the Percentage of Days Fitbits Were Worn in PPD and Non-PPD Cohorts Across Periods

To validate our results, we conducted a sensitivity analysis using a stricter definition of days worn (≥10 hours, >100 steps, <45,000 steps, consistent with other studies [12,32]) to avoid overestimation of wear time. Similar to methods described in previous sections, for days that met these criteria, we created a binary variable for each day to indicate whether the device was worn or not. We then determined the percentage of days the Fitbit was worn during each of the 4 periods (ie, prepregnancy, pregnancy, postpartum, and PPD [or PPD-equivalent for the non-PPD cohort]) by counting the number of days the device was worn divided by the total number of days during that period for each female. The total number of days was determined for each person by filtering data after the first recorded date of any Fitbit data to ensure that we were not labeling someone as not wearing their Fitbit if they did not own one at the time. A sensitivity analysis was not conducted to compare hours of wear time, as it would not be appropriate to perform a comparison between groups for the number of hours the device was worn only in days it was worn more than 10 hours. We also did not pursue a sensitivity analysis to compare the percentage of days of Fitbit sleep data, as determining whether a device was worn during sleep was based on the presence of *main sleep* Fitbit data rather than step data and hours of wear time.

Results

Descriptive Statistics

Females were assigned to the PPD cohort using methods as previously described (Methods section; Figure 1) [12,25]. To summarize, females were assigned to the PPD or non-PPD cohorts using a combination of EHR and Fitbit data. Starting with the entire AoURP longitudinal observational cohort with EHR data, we filtered on (1) individuals assigned female sex at birth, (2) female participants with any Fitbit data (ie, those who already owned a Fitbit and consented to share data), (3) female participants with any pregnancy data, and (4) female participants with Fitbit and pregnancy data at the same time. Females were then assigned to the PPD cohort based on the presence of a PPD diagnostic code, depression diagnostic code during the postpartum period, or antidepressant prescription during the postpartum window using EHR data. The remaining cohort of females not assigned to the PPD cohort were labeled as females without PPD.

Our AoURP cohort consisted of 142 females who experienced pregnancy and had available Fitbit data with a total of 108,062 days of data, where 41 (28.9%) females experienced PPD (31,201 days of data) and 101 (71.1%) females (76,861 days of data) did not. To achieve an accurate comparison in females without PPD, we created a pseudotime period labeled *PPD-equivalent* starting 58 days following delivery, which was the median number of days after delivery for PPD diagnosis (more details about pregnancy periods are provided in the Methods section). In the PPD and non-PPD cohorts, there were 13,225 days and 40,212 days of data during prepregnancy, 11,055 days and 27,559 days of data during pregnancy, 5089 days and 6060 days of data during postpartum, and 1832 days and 3030 days of data during the PPD or PPD-equivalent periods, respectively. The median age of the PPD cohort was 33.1 years (IQR 29.1-35.7 years) compared to 33.9 years (IQR 30.9-37.1 years) for females in the non-PPD cohort. Both the PPD (36/41, 88%) and non-PPD cohorts (76/101, 75.2%) were predominantly White non-Hispanic. Approximately 50% of the PPD cohort and non-PPD cohorts had an annual income of more than US $100,000 (Table 1).

The count and percentage (or median and IQR) of the number of days of Fitbit data during each period of pregnancy; demographics, including age and race/ethnicity; and annual income are provided in Table 1. Periods of pregnancy were defined using EHR data of (1) the PPD index date (14 days before the EHR date of PPD diagnosis to align with the diagnostic requirements of having symptoms for at least 2 weeks), (2) the date of delivery (separating the pregnancy and postpartum periods), and (3) the date of pregnancy (either the first EHR recorded date of pregnancy or 9 months [standard length of pregnancy] before the date of delivery) creating 4 distinct periods of pregnancy. Because females in the non-PPD cohort did not experience PPD, a fourth comparable pseudotime period labeled PPD-equivalent (starting 58 days postdelivery—the median time to PPD diagnosis after delivery in the PPD cohort) was created.

XSL•FO
RenderX

Figure 1. Flow diagram describing the inclusion and exclusion criteria of postpartum depression (PPD) and non-PPD cohorts. AoURP: *All of Us* Research Program.

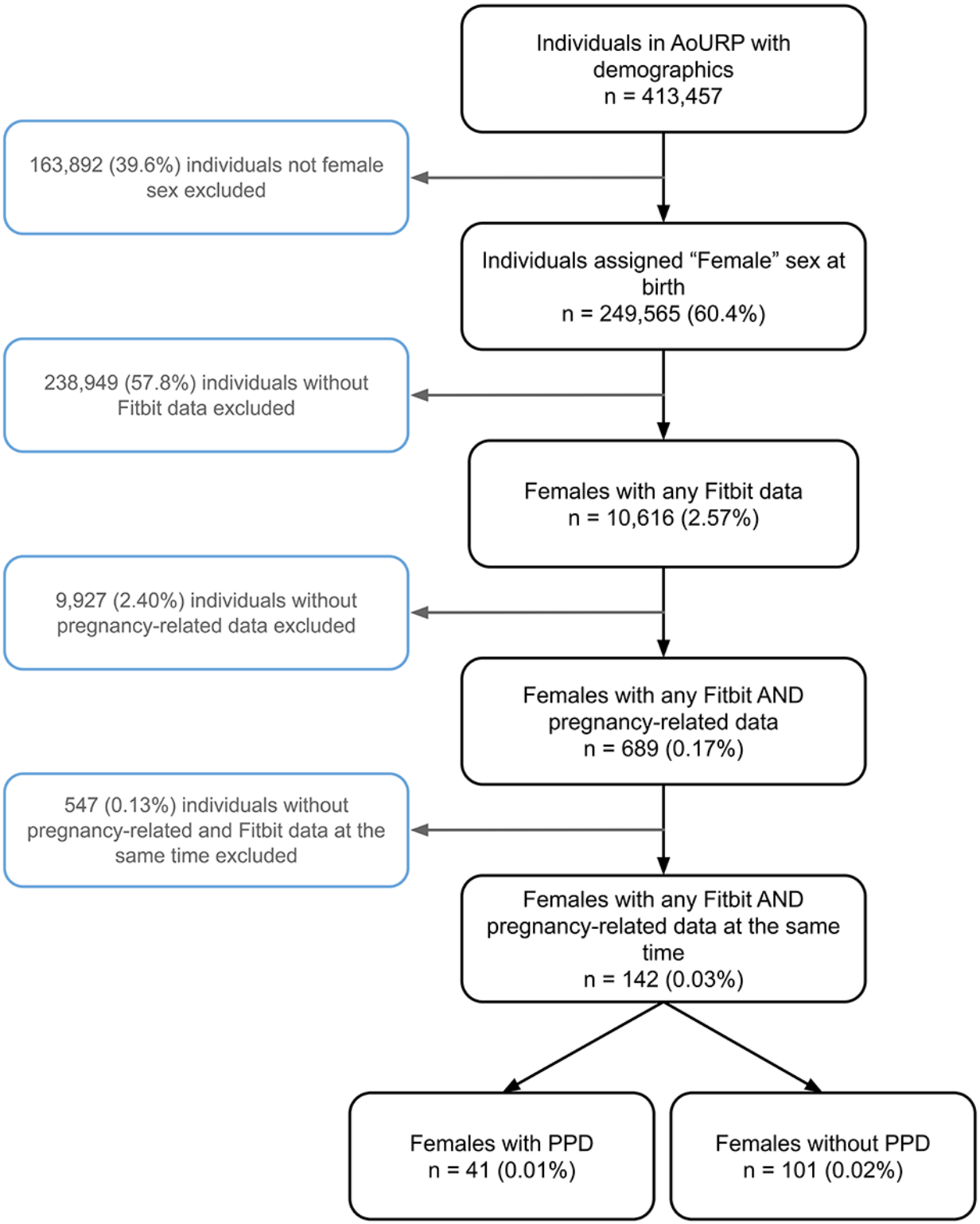

XSL•FO
RenderX

Table 1. Descriptive statistics of postpartum depression (PPD) and non-PPD cohorts.

Statistics	Cohort		
	Total	PPD	Non-PPD
Number of females, n (%)	142 (100)	41 (28.9)	101 (71.1)
Number of days of data, n (%)	108,062 (100)	31,201 (28.9)	76,861 (71.1)
Number of days of data by period, n (%)			
Prepregnancy	53,437 (49.5)	13,225 (42.4)	40,212 (52.3)
Pregnancy	38,614 (35.7)	11,055 (35.4)	27,559 (35.9)
Postpartum	11,149 (10.3)	5089 (16.3)	6060 (7.9)
PPD (or PPD-equivalent)	4862 (4.5)	1832 (5.9)	3030 (3.9)
Age (y), median (IQR)	33.7 (30.6-36.5)	33.1 (29.1-35.7)	33.9 (30.9-37.1)
White non-Hispanic females, n (%)	112 (78.9)	36 (87.8)	76 (75.2)
Females with annual income >US $100,000, n (%)	72 (51.1)[a]	<20 (48.8)[a]	52 (51.5)

[a]Note: to comply with *All of Us* Research Program guidelines for counts <20 and mitigate the risk of participant reidentification, the number of females in the PPD cohort and the combined PPD and non-PPD groups with an income >US $100,000 is reported as an estimate.

Wear Time Patterns in Females With and Without PPD

We first sought to evaluate whether females with PPD displayed Fitbit wear time behavior that differed to those without PPD. We calculated the percentage of days that each female wore their device during the PPD and PPD-equivalent periods and built a linear regression model adjusted for age at PPD diagnosis, race/ethnicity, and annual income. The results revealed a trend (although nonsignificant after using Bonferroni adjusted *P*=.013) that the percentage of days the device was worn in the PPD cohort (mean 66.6%, 95% CI 37.9%-95.3%) was greater than the non-PPD cohort (mean 46.4%, 95% CI 20.5%-72.2%; *P*=.02; Figure 2).

Observing this pattern during the PPD or PPD-equivalent periods, we proceeded to explore potential disparities in wear time between PPD and non-PPD cohorts across other pregnancy stages, including prepregnancy, pregnancy, and postpartum periods. Such analysis aimed to discern potential associations between Fitbit wear time behavior and future PPD onset. Models were run in a similar fashion for the prepregnancy, pregnancy, and postpartum periods, where we also detected a trend, though nonsignificant after multiple testing correction, of increased wear time during the postpartum period, with a mean of 69.9% (95% CI 42.7%-97%) in the PPD cohort compared to 50% (95% CI 25.5%-74.4%) in the non-PPD cohort (*P*=.02; Figure 2). These results suggest that females who go on to develop PPD may wear their device more than those who do not in the postpartum period. Alternatively, there was no significant difference in the percentage of days the device was worn during prepregnancy (PPD cohort: mean 47.5%, 95% CI 27.3%-67.7%; non-PPD cohort: mean 48.6%, 95% CI 30.5%-66.7%; *P*=.87) or pregnancy (PPD cohort: mean 46.8%, 95% CI 22.4%-71.3%; non-PPD cohort: mean 40.5%, 95% CI 18.4%-62.6%; *P*=.41) periods between the 2 cohorts (Figure 2). Sensitivity analysis results showed the same patterns of (1) no difference in Fitbit wear time during prepregnancy (PPD cohort: mean 45.5%, 95% CI 24.1%-66.9%; non-PPD cohort: mean 45.2%, 95% CI 26%-64.4%; *P*=.97) and pregnancy (PPD cohort: mean 44.1%, 95% CI 19.6%-68.5%; non-PPD cohort: mean 38%, 95% CI 15.9%-60.1%; *P*=.44) time periods between PPD and non-PPD cohorts and (2) a trend, though nonsignificant, of increased wear time during the PPD (PPD cohort: mean 64.5%, 95% CI 37%-92%; non-PPD cohort: mean 46.1%, 95% CI 21.3%-70.8%; *P*=.03) and postpartum (PPD cohort: mean 62.6%, 95% CI 36.5%-88.7%; non-PPD cohort: mean 47.8%, 95% CI 24.4%-71.3%; *P*=.06) time periods among females in the PPD cohort compared to females without PPD, with trends slightly less strong than the main analysis (Multimedia Appendix 1).

XSL•FO
RenderX

Figure 2. The percentage of days females in the postpartum depression (PPD) and non-PPD *All of Us* Research Program (AoURP) cohorts wore their Fitbit across the prepregnancy (top left), pregnancy (top right), postpartum (bottom left), and PPD (or PPD-equivalent; bottom right) periods. Data in the PPD and non-PPD cohorts were compared using linear regression adjusted for age at PPD diagnosis, race/ethnicity, and annual income and are expressed as mean and 95% CI. PPD-eq: postpartum depression–equivalent.

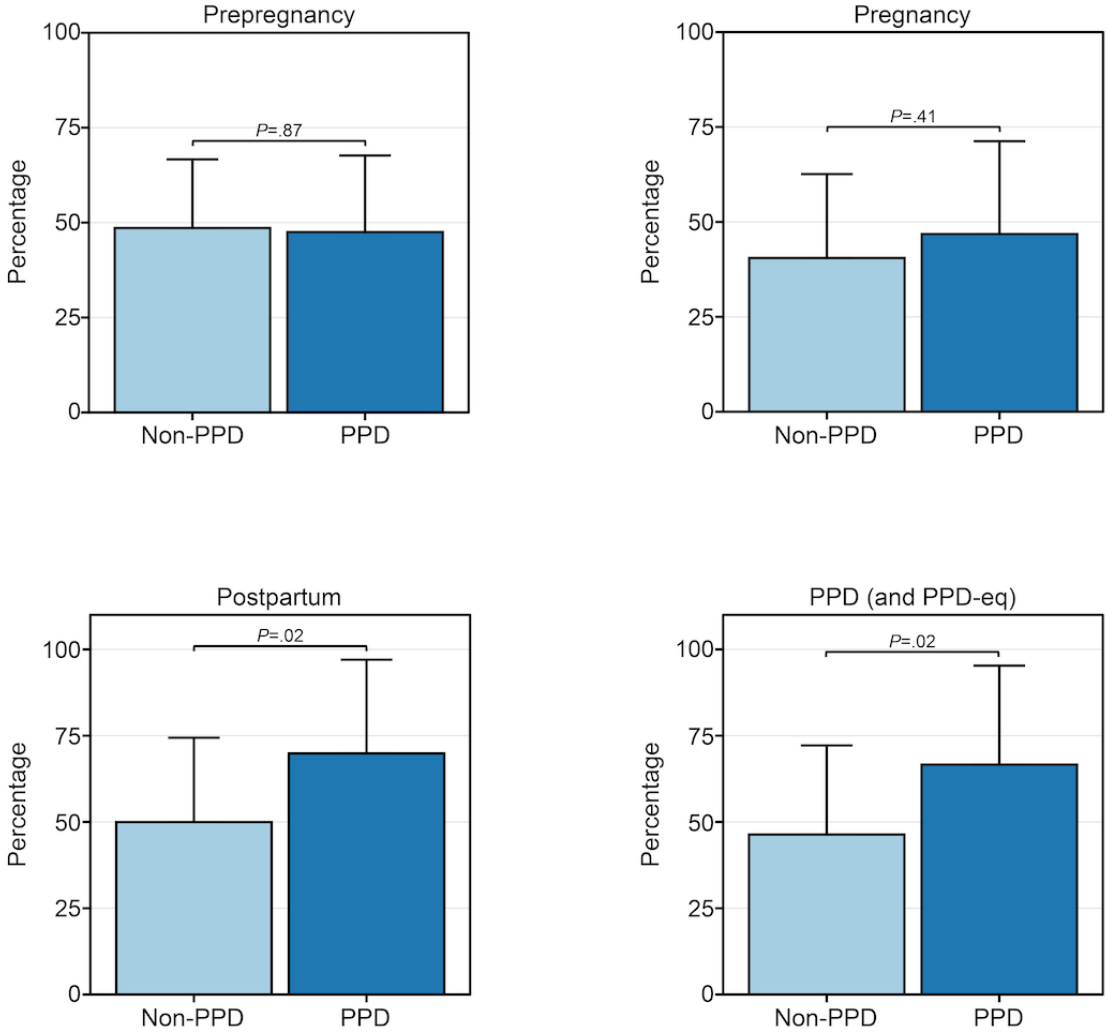

Hours Per Day of Wear Time in Females With and Without PPD

Observing variation in the percentage of days wearable devices were worn between PPD and non-PPD cohorts, we subsequently evaluated if there were any differences in the daily duration of device wear time adjusted for age at PPD diagnosis, race/ethnicity, and annual income. Surprisingly, our findings revealed no trends or significant differences between PPD and non-PPD cohorts during prepregnancy (PPD cohort: mean 14.5, 95% CI 12.5-16.5; non-PPD cohort: mean 15.1, 95% CI 13.3-16.9; *P*=.35), pregnancy (PPD cohort: mean 16.1, 95% CI 14-18.1; non-PPD cohort: mean 16.8, 95% CI 14.9-18.6; *P*=.30), postpartum (PPD cohort: mean 16.6, 95% CI 14.6-18.6; non-PPD cohort: mean 17.4, 95% CI 15.6-19.1; *P*=.27), or the PPD or PPD-equivalent periods (PPD cohort: mean 17.2, 95% CI 15-19.3; non-PPD cohort: mean 17.8, 95% CI 16-19.6; *P*=.38; Figure 3).

XSL•FO
RenderX

Figure 3. The number of hours per day females in the postpartum depression (PPD) and non-PPD *All of Us* Research Program (AoURP) cohorts wore their Fitbit across the prepregnancy (top left), pregnancy (top right), postpartum (bottom left), and PPD (or PPD-equivalent; bottom right) time periods. Data in the PPD and non-PPD cohorts were compared using linear mixed-effects models with person ID as the random effect adjusted for age at PPD diagnosis, race/ethnicity, and annual income and are expressed as mean and 95% CI. PPD-eq: postpartum depression–equivalent.

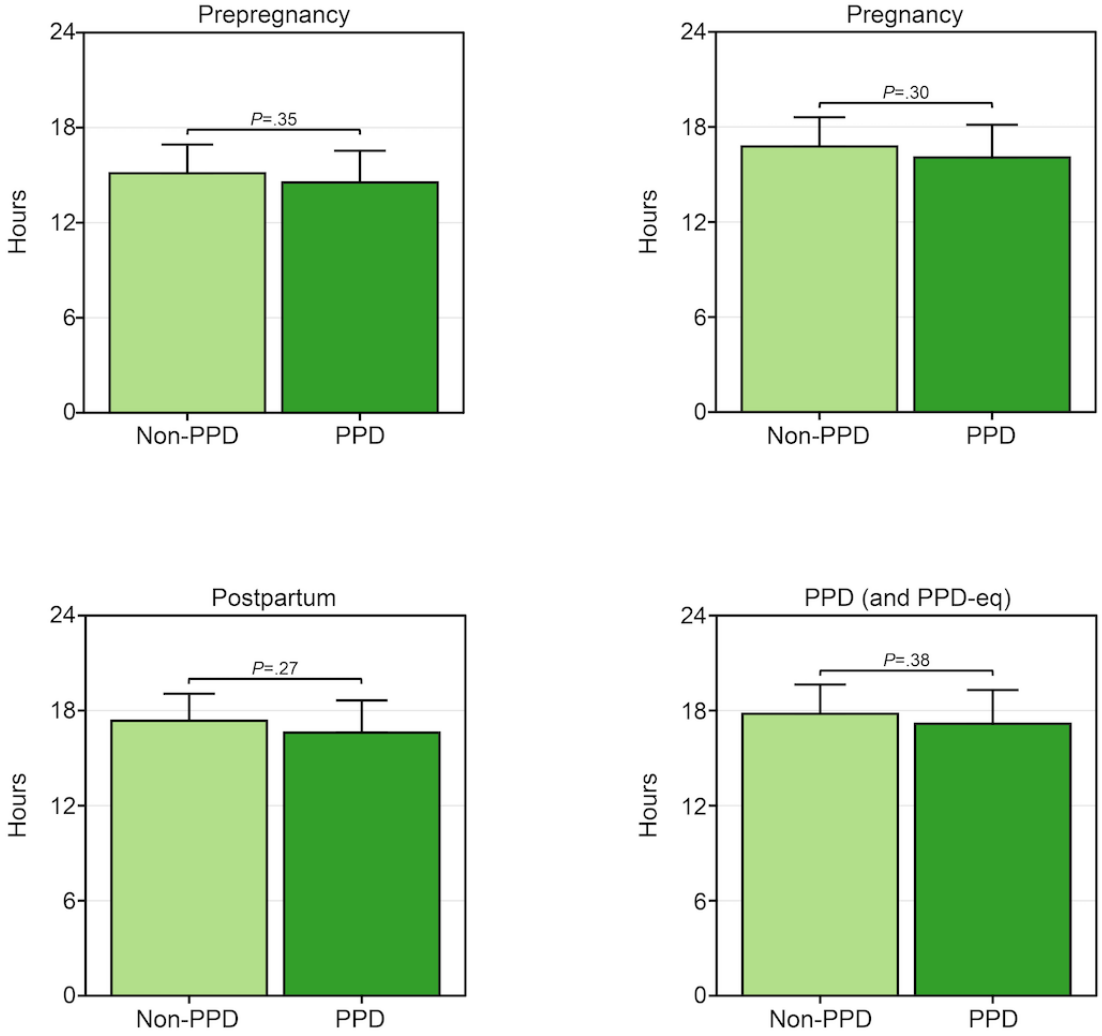

Wear Time Patterns to Sleep in Females With and Without PPD

Given the extensive connection between sleep and PPD, we sought to describe and compare the percentage of days females in the PPD cohort wore their Fitbit to sleep compared to those without PPD during each phase of pregnancy [39-45]. When comparing the percentage of days females in each cohort wore their Fitbit to sleep, the results showed similar observations to the percentage of wear time findings (Figure 2), where we noticed a trend, though nonsignificant after multiple testing correction, of females wearing the device to sleep more during the postpartum period in the PPD cohort (mean 55.4%, 95% CI 30.8%-79.9%) compared to the non-PPD cohort (mean 37.3%, 95% CI 15.2%-59.4%; *P*=.02) adjusted for age at PPD diagnosis, race/ethnicity, and annual income (Figure 4). There was a similar pattern during the time period females experienced PPD (mean 60.3%, 95% CI 33.4%-87.2%) compared to those who did not (mean 41.8%, 95% CI 17.6%-66%; *P*=.02; Figure 4). No differences were detected in the percentage of days Fitbits were worn to sleep between PPD and non-PPD cohorts during the prepregnancy (PPD cohort: mean 46.4%, 95% CI 20.6%-72.1%; non-PPD cohort: mean 44.7%, 95% CI 21.7%-67.7%; *P*=.84) or pregnancy (PPD cohort: mean 60.7%, 95% CI 35.1%-86.3%; non-PPD cohort: mean 49.4%, 95% CI 26.2%-72.5%; *P*=.16) periods (Figure 4).

XSL•FO
RenderX

Figure 4. The percentage of days females in the postpartum depression (PPD) and non-PPD *All of Us* Research Program (AoURP) cohorts wore their Fitbit to sleep across the prepregnancy (top left), pregnancy (top right), postpartum (bottom left), and PPD (or PPD-equivalent; bottom right) periods. Data in the PPD and non-PPD cohorts were compared using linear regression adjusted for age at PPD diagnosis, race/ethnicity, and annual income and are expressed as mean and 95% CI. PPD-eq: postpartum depression–equivalent.

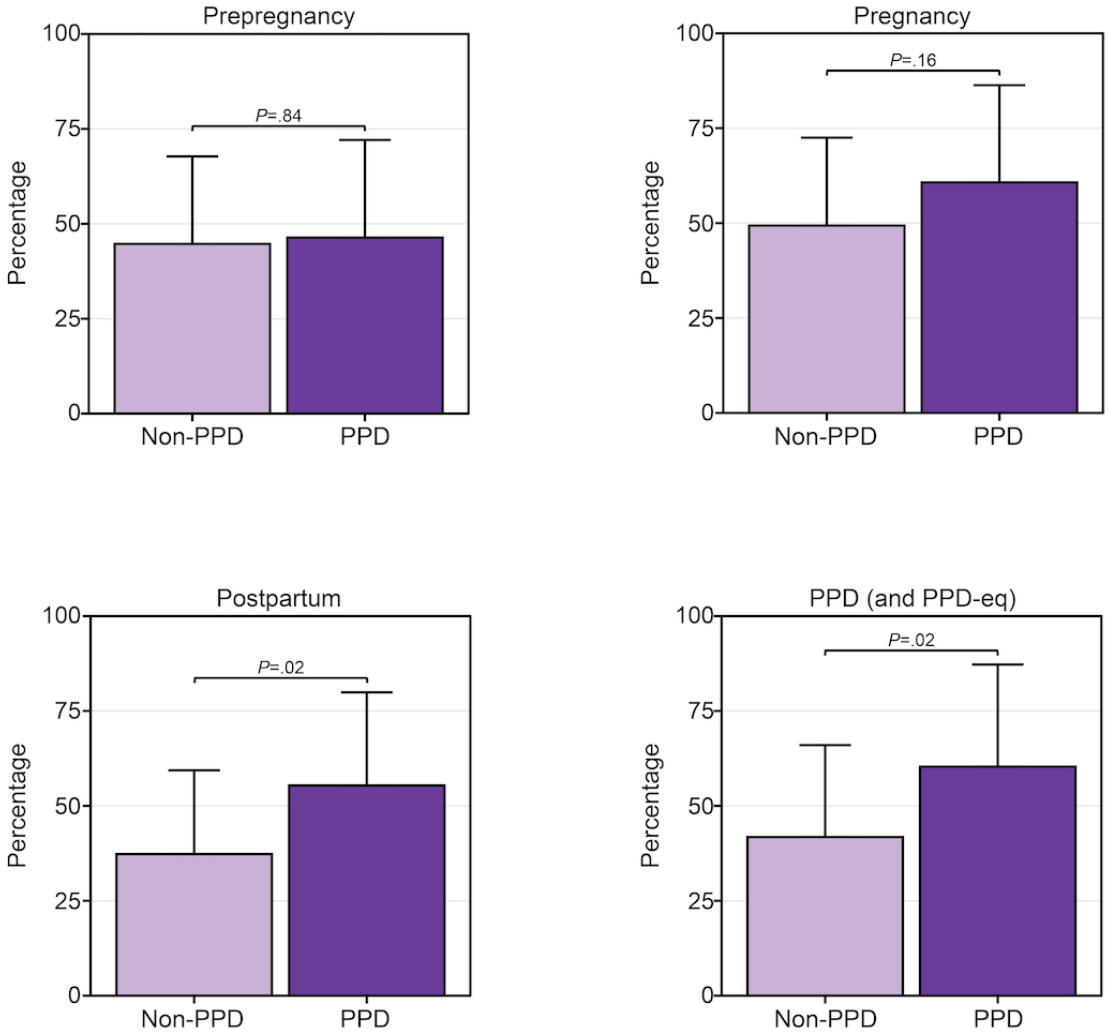

Fitbit Wear Time Consistency During PPD and Other Time Periods of Pregnancy

Finally, we wanted to explore individual-level device wear time patterns before and during PPD. Therefore, we examined the correlation between the percentage of days within females who wore their Fitbit during periods before PPD (ie, prepregnancy and pregnancy) with the PPD period. For instance, a positive correlation would suggest that those who wore their Fitbit more frequently during the prepregnancy period tended to do so during the PPD period. We conducted this analysis in parallel with the non-PPD cohort for comparison. In females with PPD, the results displayed a significant positive correlation between the percentage of days the Fitbit was worn during prepregnancy and PPD periods (r=0.48, 95% CI 0.16-0.71; P=.005; Figure 5). A positive correlation was also detected during prepregnancy and PPD-equivalent periods among females without PPD, but it did not reach statistical significance after Bonferroni correction (r=0.24, 95% CI 0.04-0.42; P=.02; Figure 5). There also existed a strong positive correlation between the percentage of wear time during the pregnancy and PPD periods among females in the PPD (r=0.77, 95% CI 0.59-0.88; P<.001) and non-PPD cohorts (r=0.66, 95% CI 0.53-0.76; P<.001; Figure 5).

XSL•FO
RenderX

Figure 5. The Pearson correlation coefficient and associated *P* values demonstrate the relationship between the percentage of days females wore their Fitbit across different periods separated by postpartum depression (PPD) status—without PPD (left) or with PPD (right). Top panels display the correlation of females wearing their Fitbit during prepregnancy and PPD periods (or PPD-equivalent for the non-PPD cohort), while bottom panels show the correlation during pregnancy and PPD periods (or PPD-equivalent for those in the non-PPD cohort). The blue line represents the line of best fit and gray shading shows the 95% CI. PPD-eq: postpartum depression–equivalent.

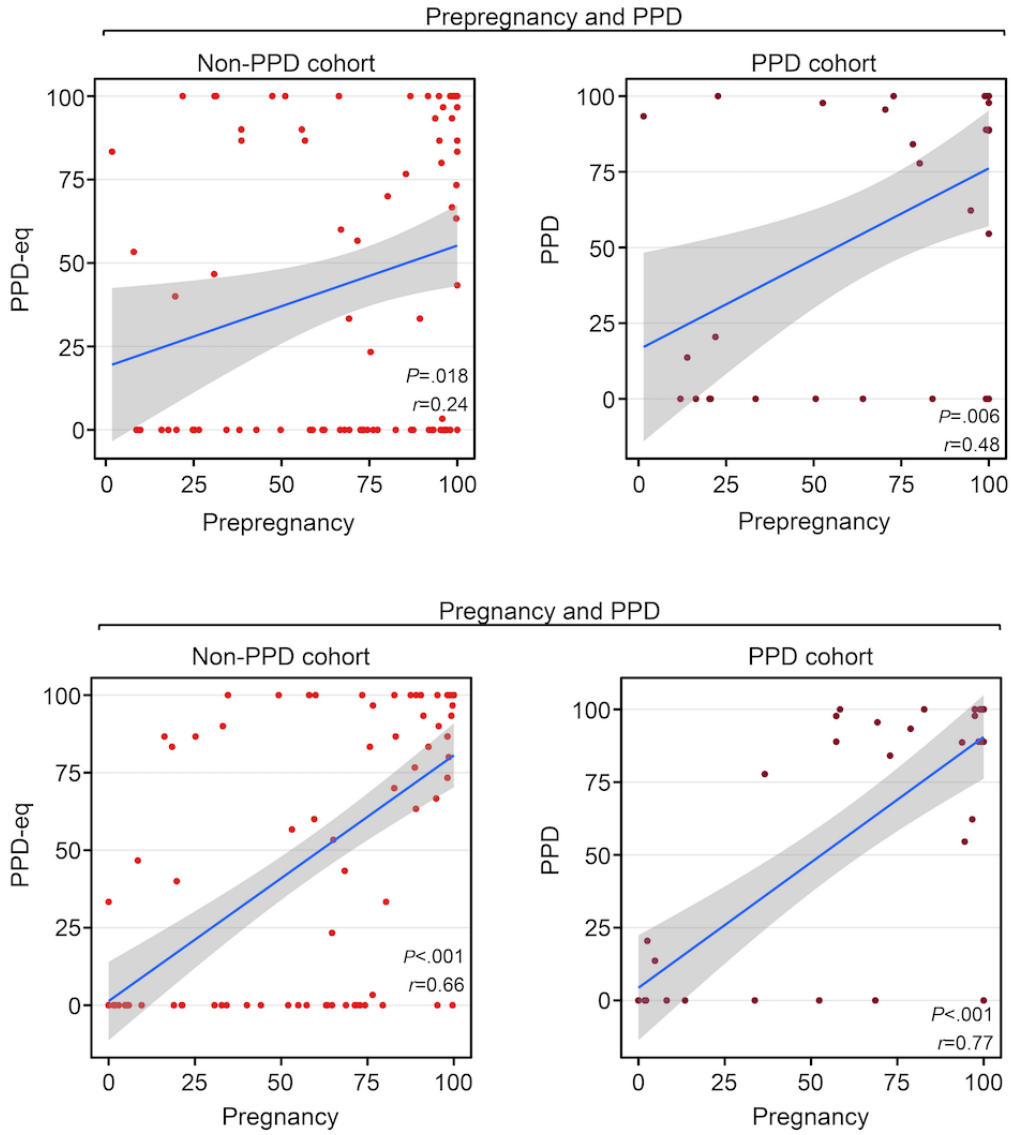

To further understand the Fitbit wear time behavior of females during each period, we also sought to analyze the consistency with which Fitbit devices were worn. Our earlier analyses focused on comparing the percentage of days the device was worn across different periods; however, we acknowledged that wear patterns could vary. For example, if there were 50 days in total to potentially wear the device during one of the periods, wearing it consistently for 25 consecutive days followed by 25 days of nonuse is different from alternating between wearing and not wearing the device every other day, even though both scenarios indicate 50% wear time. Therefore, to assess individual-level consistency during each period, we determined the maximum consecutive number of days the device was worn during each period and examined the correlation across periods (ie, during prepregnancy and PPD [or PPD-equivalent] and during pregnancy and PPD [or PPD-equivalent]). The results displayed a trend, though nonsignificant, in wear time consistency between prepregnancy and PPD-equivalent periods among females without PPD (r=0.25; P=.07), while those with PPD did not exhibit any correlation (r=–0.05, 95% CI –0.46 to 0.38; P=.84; Figure 6). Alternatively, a significant correlation was present in the non-PPD (r=0.54, 95% CI 0.31-0.7; P<.001) cohort between the pregnancy and PPD (or PPD-equivalent) periods, but only a strong trend, which was nonsignificant after multiple testing correction, was observed in the PPD cohort between the 2 periods (r=0.48, 95% CI 0.08-0.74; P=.02; Figure 6). These data suggest a relationship between the consistency of Fitbit use during pregnancy and the PPD (or PPD-equivalent) periods in both cohorts. Notably, when analyzing the consistency of Fitbit wear time during prepregnancy, the relationship only was present among females in the non-PPD cohort (Figure 6).

XSL•FO
RenderX

Figure 6. The Pearson correlation coefficient and associated *P* values demonstrate the relationship between the maximum number of days in a row (ie, consistency) females wore their Fitbit across different periods separated by postpartum depression (PPD) status—without PPD (left) or with PPD (right). Top panels display the correlation of females wearing their Fitbit during prepregnancy and PPD periods (or PPD-equivalent for the non-PPD cohort), while bottom panels show the correlation during pregnancy and PPD periods (or PPD-equivalent for those in the non-PPD cohort). The blue line represents the line of best fit and gray shading shows the 95% CI. PPD-eq: postpartum depression–equivalent.

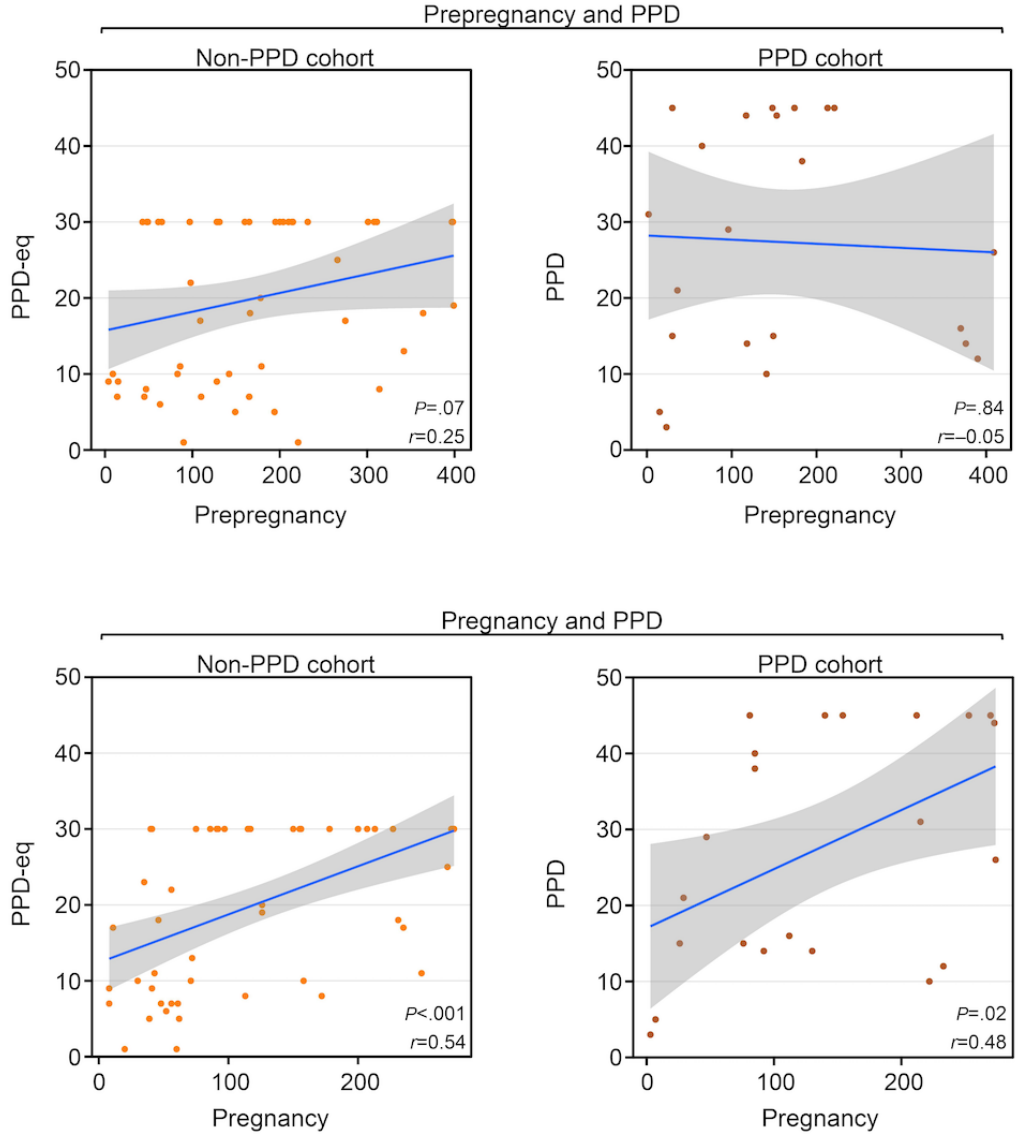

Discussion

Principal Findings

Our study elucidated numerous insights related to Fitbit wear time and PPD across periods of pregnancy. First, we observed a trend, though nonsignificant after multiple testing correction, that females with PPD wear their device a higher percentage of days than those without PPD during the postpartum and PPD periods (Figure 2). However, among the days that females with and without PPD wore their device, there was not a significantly different number of daily hours the device was worn (Figure 3). Regarding Fitbit wear time during sleep, a similar pattern was observed as the percentage of daily wear time. Females in the PPD cohort wore their Fitbit to sleep on a higher percentage of days during the postpartum and PPD periods compared to those in the non-PPD cohort (Figure 4). It was also observed that females in the PPD and non-PPD cohorts displayed the same correlation patterns between wear time during earlier and later pregnancy periods (Figure 5). Finally, we found that females in both cohorts who wore their devices more consistently during pregnancy also maintained higher levels of device wear consistency during the PPD (or PPD-equivalent) periods. However, there was no correlation in the consistency of Fitbit wear time during the prepregnancy and PPD periods among females in the PPD cohort (Figure 6).

Our study's first key finding showed that females with PPD wore their Fitbit a higher percentage of days compared to females without PPD. A similar pattern was detected during the postpartum periods among females from PPD and non-PPD cohorts (Figure 2). The trend was not as strong in our sensitivity analysis using a stricter definition of Fitbit wear time (Multimedia Appendix 1). One reason we hypothesize that females with PPD wore their devices more frequently than those without PPD is due to anxiety and hypervigilance, which commonly occurs in females with PPD, and may drive increased personal tracking behavior [46]. Because PPD often goes

XSL•FO
RenderX

undetected, females in the PPD cohort during the postpartum period (before EHR diagnosis) may have already been experiencing PPD symptoms, which could explain the similarities in patterns observed between the postpartum and PPD periods [4]. Unfortunately, AoURP does not have symptom-related data so we cannot know for sure when symptoms began, which could also explain the differences in our analyses during the postpartum period where some females may have been experiencing symptoms while others were not. While it was originally suggested that PPD symptoms peak between 4 and 6 weeks in the postpartum period [47,48], recent work suggests subgroups of females display unique symptom trajectories [49]. Furthermore, although females in the PPD cohort tended to wear their Fitbit a higher percentage of days during the postpartum and PPD periods, we did not detect any difference in the number of hours per day the device was worn compared to females without PPD (Figure 3). Our findings showed that both the PPD and non-PPD cohorts wore their devices approximately 15 to 17 hours of the day, which is consistent with other studies involving Fitbits [50,51]. Our findings also align with previous research in perinatal populations, where (1) Sarhaddi et al [52] reported that women wore their wearable devices for an average of 17 hours per day during pregnancy and 13.7 hours per day postpartum (up to 12 weeks after delivery) and (2) Grym et al [53] reported device wear time for an average of 17.3 hours per day during pregnancy and 14.4 hours per day postpartum (up to 4 weeks after delivery). These findings are also similar to individuals with depression who wore their device for an average of 15 to 17 hours per day [54].

The next component of our study was to investigate the percentage of days females with PPD wear their device to sleep across each phase of pregnancy compared to those without PPD given the extensive relationship between sleep and PPD [39-45]. Our findings revealed that females in the PPD cohort tended to wear their device more to sleep during the postpartum and PPD periods compared to those without PPD (Figure 4). Considering the similar pattern observed in Fitbit wear time frequency (Figure 2), it was not surprising to find the same result in sleep data. Unfortunately, it is not possible to determine whether the device was intentionally worn for sleep tracking or simply due to continued use, but it could be interesting to investigate this in future studies. The fact that females on average wear their device between 15 and 17 hours per day suggests that when females in these cohorts do wear their device, they also wear it to sleep [50,51].

Finally, our study sought to assess whether Fitbit wear time behavior during periods before PPD may correlate with behavior during PPD, with the potential that wear time behavior during prepregnancy or pregnancy periods may be able to help predict PPD onset. Our findings displayed that females who wore their device more during prepregnancy also wore their device more during PPD (Figure 5). A similar observation was detected when comparing pregnancy and PPD periods; however, this relationship also persisted among females in the non-PPD cohort (Figure 5). When assessing the consistency of Fitbit wear time, we noticed a trend, though nonsignificant, only in females without PPD that greater wear time consistency in prepregnancy correlated with greater consistency during the PPD-equivalent period (Figure 6). This may be attributed to females with PPD experiencing cooccurring mood and anxiety symptoms, leading them to wear their devices more frequently regardless of prepregnancy consistency [46]. In addition, wear time patterns of wearable device during prepregnancy and PPD periods findings may also be due to behavioral or lifestyle factors, such as increased motivation for sleep tracking, as wearables increase perception of sleep quality and are increasingly used for sleep assessment, which is important for new mothers [55-57].

Limitations

While this study is the first to evaluate wear time behavior of wearable device among females with PPD across phases of pregnancy, it is not without limitations. First, the number of hours the device was worn was estimated based on recorded Fitbit steps data using previously established methods; however, it is not *ground truth* data and therefore may contain some level of inaccuracy [32]. We estimated the hours per day (and percentage of days) females in the PPD and non-PPD cohort wore their Fitbit based on the presence of step data, which is consistent with our previous work and that of others [12,32]. Unfortunately, it is not possible to know exactly how long someone wore their device from the AoURP retrospective data. Second, we do not have access to study participants in AoURP to perform any qualitative analysis to further understand causal relationships about individual-level Fitbit wear time patterns and disease symptoms or severity. Wear time could be impacted by factors not controlled for in the study, such as skin issues from using a wristband, breastfeeding status (which was not available in AoURP), and mental health history (which was intentionally excluded due to its significantly higher prevalence in the PPD cohort compared to the non-PPD cohort, and including it in the model could result in unreliable estimates) [47,58,59]. Future studies should include user-experience-related questionnaires and qualitative methods tailored toward women with and without PPD during the postpartum period to better assess the causality between PPD, hypervigilance, other potential confounding factors, and device wear time [15-19]. Third, this study only investigated wearable device behavior for Fitbit. While Fitbit is the most commonly wearable device used for research purposes, it would be valuable to incorporate other devices, such as the Apple Watch, Google Fit, Garmin smartwatch, or Oura ring, which have shown high levels of adherence, and the type of device could be adjusted as a covariate [60-63]. Fourth, the PPD and non-PPD cohorts were relatively small, and we posit we may have observed statistical significance with larger sample sizes. Furthermore, future studies with larger sample sizes should include a sensitivity analysis in females with a confirmed PPD diagnosis to robustly validate our findings to address potential overclassification of PPD diagnosis, which we could not do in this study as AoURP Data and Statistics Dissemination policy prohibits analyses on samples sizes less than 20 individuals to protect patient privacy [24]. Finally, there was likely a selection bias, because our cohorts consisted primarily of females who were White and non-Hispanic. The lack of racial or ethnic diversity may influence study results and limit their generalizability to other patient populations. Previous studies have shown racial or ethnic

XSL•FO
RenderX

disparities in PPD prevalence, such as the substantial increase in PPD among Asian and Pacific Islanders and Black non-Hispanic women [64]. Future work should replicate our study findings in diverse populations to evaluate differences in wear time behavior of a wearable device between PPD and non-PPD cohorts across different racial or ethnic groups with additional confounding variables (eg, mental health history and breastfeeding status), if available. We want to clarify that all Fitbit data used in this study was collected by AoURP, where the program operates under a bring-your-own-device model and participants who consent to share Fitbit data already own their device [22]. Notably, one strength of this study is that AoURP does not send any type of notification or reminders for continued use, thus our work provides a great foundation for the first study to assess real-world wear time behavior of a wearable device in females with PPD.

Interpretation

PPD remains underscreened and consequently underdiagnosed for several reasons, such as its heterogeneity and stigmatization [8]. Wearable devices offer a promising avenue for continuous mental health monitoring and early detection through their ability to capture high-density longitudinal data, including physical activity and sleep patterns—factors known to influence PPD [45,55]. Our findings showed that females with PPD exhibit higher wear time of a wearable device compared to those without PPD, suggesting that wear time patterns may serve as an additional digital biomarker beyond traditional sensor data. While wear time alone may not be specific enough to detect PPD, its potential clinical value lies in its integration with other wearable digital biomarkers, such as longitudinal patterns of physical activity and sleep. Building on our previous work that demonstrated individualized ML models using wearable sensor data (eg, day-level average heart rate, sum of steps, and activity calories) can differentiate PPD from other pregnancy periods, the results of this study suggest that combining sensor data with wear time patterns could enhance PPD detection capabilities [12]. We propose that future clinical implementation of PPD detection ML algorithms using wearables can include wear time of a wearable device in addition to sensor-derived digital biomarkers to prompt PPD screening using the EPDS [9]. However, challenges such as ensuring data privacy, usability, providing adequate clinician training, acceptability to clinicians, improving patient accessibility, and compatibility with existing screening tools (eg, EPDS) must be addressed, along with fostering collaboration between maternal and mental health services to create dedicated care pathways for comprehensive perinatal support [9,59,60,65]. To eventually achieve the goal of clinical implementation, additional work is needed using advanced ML approaches to quantify the relative importance of wear time of a wearable device compared to other features in PPD prediction models, ultimately working toward more effective early screening and diagnosis protocols.

Conclusions

Understanding wear time behavior of a wearable device can provide insightful clinical information related to women with PPD. Considering that screening and diagnosis of PPD pose significant challenges, wearables, including features of wear time behavior, could potentially offer a viable solution. We envision a future using wearables combined with an ML algorithm that incorporates wear time of a wearable device with other digital biomarkers, such as sleep and physical activity, to facilitate early detection of PPD by notifying the clinician with potential concern to prompt timely screening. Wear time behavior presents a passive and relatively straightforward feature to aid in evaluating PPD in nonclinical environments, and its application could potentially extend to other perinatal and general mental health disorders.

Acknowledgments

First, the *All of Us* Research Program would not be possible without the partnership of its participants. The *All of Us* Research Program is supported by the National Institutes of Health, Office of the Director: Regional Medical Centers: 1 OT2 OD026549; 1 OT2 OD026554; 1 OT2 OD026557; 1 OT2 OD026556; 1 OT2 OD026550; 1 OT2 OD 026552; 1 OT2 OD026553; 1 OT2 OD026548; 1 OT2 OD026551; 1 OT2 OD026555; IAA #: AOD 16037; Federally Qualified Health Centers: HHSN 263201600085U; Data and Research Center: 5 U2C OD023196; Biobank: 1 U24 OD023121; The Participant Center: U24 OD023176; Participant Technology Systems Center: 1 U24 OD023163; Communications and Engagement: 3 OT2 OD023205; 3 OT2 OD023206; and Community Partners: 1 OT2 OD025277; 3 OT2 OD025315; 1 OT2 OD025337; 1 OT2 OD025276. The authors would also like to acknowledge that they greatly appreciate support from the University of North Carolina Department of Genetics and this work would not have been possible without them.

Individual authors were supported by NIMH R01131542 (RCP), NHGRI U24HG011449 (EH and MAH), and NHGRI RM1HG010860 (EH and MAH).

ChatGPT (GPT-3.5; OpenAI) [66] and Claude (Anthropic) [67], were used to edit some portions of the manuscript, including grammar, language, and synonyms. All recommendations from ChatGPT and Claude were reviewed by the authors and were not used for generating ideas or content.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to data security and privacy guidelines defined by *All of Us* Research Program (AoURP). However, access to the entire AoURP cohort is available to those who (1) complete the required training activities requested by AoURP and (2) create an account on the Researcher Workbench. To individuals who

XSL•FO
RenderX

complete the necessary requirements in AoURP, data and code for this study are available upon reasonable request to the corresponding author.

Authors' Contributions

EH conceived the study. EH and MAH designed the study. EH analyzed the data. EH, SM-B, ZB-D, RCP, NE, and MAH interpreted the results. EH wrote the draft manuscript. MAH provided supervision and acquired funding. All authors reviewed and edited the final manuscript for publication.

Conflicts of Interest

MAH is a founder of Alamya Health. SMB has received sponsored research grant funding from Sage Therapeutics, ElectroMedical Products International, and Sirtsei Pharmaceuticals for the conducts of clinical trials. SMB has also served as a Scientific Advisor for EmbarkNeuro and Seaport Therapeutics, and serves as a professional corporation owner and clinical advisor for Modern Health. All other authors declare no other conflicts of interest.

Multimedia Appendix 1
The percentage of days females in the PPD and non-PPD AoURP cohorts wore their Fitbit using a stricter definition of wear time. AoURP: *All of Us* Research Program; PPD: postpartum depression.
[DOCX File , 184 KB - formative_v9i1e67585_app1.docx]

References

1. Insel T. Digital phenotyping: technology for a new science of behavior. JAMA 2017 Oct 03;318(13):1215-1216. [doi: 10.1001/jama.2017.11295] [Medline: 28973224]
2. Dunn J, Runge R, Snyder M. Wearables and the medical revolution. Per Med 2018 Sep 27;15(5):429-448 [FREE Full text] [doi: 10.2217/pme-2018-0044] [Medline: 30259801]
3. Gaynes BN, Gavin N, Meltzer-Brody S, Lohr KN, Swinson T, Gartlehner G, et al. Perinatal depression: prevalence, screening accuracy, and screening outcomes. Evid Rep Technol Assess (Summ) 2005 Feb(119):1-8. [doi: 10.1037/e439372005-001] [Medline: 15760246]
4. Cox EQ, Sowa NA, Meltzer-Brody SE, Gaynes BN. The perinatal depression treatment cascade: baby steps toward improving outcomes. J Clin Psychiatry 2016 Sep 28;77(09):1189-1200. [doi: 10.4088/jcp.15r10174]
5. Kingston D, Tough S, Whitfield H. Prenatal and postpartum maternal psychological distress and infant development: a systematic review. Child Psychiatry Hum Dev 2012 Oct 10;43(5):683-714. [doi: 10.1007/s10578-012-0291-4] [Medline: 22407278]
6. Luca DL, Margiotta C, Staatz C, Garlow E, Christensen A, Zivin K. Financial toll of untreated perinatal mood and anxiety disorders among 2017 births in the United States. Am J Public Health 2020 Jun;110(6):888-896. [doi: 10.2105/ajph.2020.305619]
7. Reid HE, Pratt D, Edge D, Wittkowski A. Maternal suicide ideation and behaviour during pregnancy and the first postpartum year: a systematic review of psychological and psychosocial risk factors. Front Psychiatry 2022 Mar 24;13:765118 [FREE Full text] [doi: 10.3389/fpsyt.2022.765118] [Medline: 35401283]
8. Postpartum Depression: Action Towards Causes and Treatment (PACT) Consortium. Heterogeneity of postpartum depression: a latent class analysis. Lancet Psychiatry 2015 Jan;2(1):59-67. [doi: 10.1016/s2215-0366(14)00055-8]
9. Guedeney N, Fermanian J, Guelfi JD, Kumar RC. The Edinburgh Postnatal Depression Scale (EPDS) and the detection of major depressive disorders in early postpartum: some concerns about false negatives. J Affect Disord 2000 Dec;61(1-2):107-112. [doi: 10.1016/s0165-0327(99)00186-x] [Medline: 11099748]
10. Deligiannidis KM, Meltzer-Brody S, Maximos B, Peeper EQ, Freeman M, Lasser R, et al. Zuranolone for the treatment of postpartum depression. Am J Psychiatry 2023 Sep 01;180(9):668-675. [doi: 10.1176/appi.ajp.20220785] [Medline: 37491938]
11. Meltzer-Brody S, Colquhoun H, Riesenberg R, Epperson C, Deligiannidis KM, Rubinow DR, et al. Brexanolone injection in post-partum depression: two multicentre, double-blind, randomised, placebo-controlled, phase 3 trials. The Lancet 2018 Sep;392(10152):1058-1070. [doi: 10.1016/s0140-6736(18)31551-4]
12. Hurwitz E, Butzin-Dozier Z, Master H, O'Neil ST, Walden A, Holko M, et al. Harnessing consumer wearable digital biomarkers for individualized recognition of postpartum depression using the All of Us research program data set: cross-sectional study. JMIR Mhealth Uhealth 2024 May 02;12:e54622 [FREE Full text] [doi: 10.2196/54622] [Medline: 38696234]
13. Ravindra NG, Espinosa C, Berson E, Phongpreecha T, Zhao P, Becker M, et al. Deep representation learning identifies associations between physical activity and sleep patterns during pregnancy and prematurity. NPJ Digit Med 2023 Sep 28;6(1):171 [FREE Full text] [doi: 10.1038/s41746-023-00911-x] [Medline: 37770643]
14. Sarhaddi F, Azimi I, Niela-Vilen H, Axelin A, Liljeberg P, Rahmani AM. Maternal social loneliness detection using passive sensing through continuous monitoring in everyday settings: longitudinal study. JMIR Form Res 2023 Aug 09;7:e47950 [FREE Full text] [doi: 10.2196/47950] [Medline: 37556183]

XSL•FO
RenderX

15. Jeong H, Kim H, Kim R, Lee U, Jeong Y. Smartwatch wearing behavior analysis: a longitudinal study. Proc ACM Interact Mob Wearable Ubiquitous Technol 2017 Sep 11;1(3):1-31. [doi: 10.1145/3131892]
16. Ogbanufe O, Gerhart N. Watch it! Factors driving continued feature use of the smartwatch. Int J Hum Comput Interact 2017 Nov 30;34(11):999-1014. [doi: 10.1080/10447318.2017.1404779]
17. Saheb T, Cabanillas FJ, Higueras E. The risks and benefits of Internet of Things (IoT) and their influence on smartwatch use. Span J Mark ESIC 2022 Feb 16;26(3):309-324. [doi: 10.1108/sjme-07-2021-0129]
18. Siepmann C, Kowalczuk P. Understanding continued smartwatch usage: the role of emotional as well as health and fitness factors. Electron Markets 2021 Feb 18;31(4):795-809. [doi: 10.1007/s12525-021-00458-3]
19. Visuri A, van Berkel N, Goncalves J, Rawassizadeh R, Ferreira D, Kostakos V. Understanding usage style transformation during long-term smartwatch use. Pers Ubiquit Comput 2021 Jan 19;25(3):535-549. [doi: 10.1007/s00779-020-01511-2]
20. Pathiravasan CH, Zhang Y, Wang X, Trinquart L, Benjamin EJ, Borrelli B, et al. Factors associated with long-term use of digital devices in the electronic Framingham Heart Study. NPJ Digit Med 2022 Dec 27;5(1):195 [FREE Full text] [doi: 10.1038/s41746-022-00735-1] [Medline: 36572707]
21. The All of Us Research Program Investigators. The “All of Us” research program. N Engl J Med 2019 Aug 15;381(7):668-676. [doi: 10.1056/nejmsr1809937]
22. Holko M, Litwin TR, Munoz F, Theisz KI, Salgin L, Jenks NP, et al. Wearable fitness tracker use in federally qualified health center patients: strategies to improve the health of all of us using digital health devices. NPJ Digit Med 2022 Apr 25;5(1):53 [FREE Full text] [doi: 10.1038/s41746-022-00593-x] [Medline: 35469045]
23. All of Us research program protocol. National Institutes of Health. URL: https://allofus.nih.gov/about/all-us-research-program-protocol [accessed 2025-01-28]
24. Data and statistics dissemination policy. All of Us Research Hub. URL: https://www.researchallofus.org/faq/data-and-statistics-dissemination-policy/ [accessed 2025-05-14]
25. Jones S, Bradwell KR, Chan LE, Olson-Chen C, Tarleton J, Wilkins KJ, et al. Who is pregnant? Defining real-world data-based pregnancy episodes in the National COVID Cohort Collaborative (N3C). medRxiv Preprint posted online on August 6, 2022 [FREE Full text] [doi: 10.1101/2022.08.04.22278439] [Medline: 35982668]
26. Amit G, Girshovitz I, Marcus K, Zhang Y, Pathak J, Bar V, et al. Estimation of postpartum depression risk from electronic health records using machine learning. BMC Pregnancy Childbirth 2021 Sep 17;21(1):630 [FREE Full text] [doi: 10.1186/s12884-021-04087-8] [Medline: 34535116]
27. Goodman JH. Postpartum depression beyond the early postpartum period. J Obstet Gynecol Neonatal Nurs 2004 Jul;33(4):410-420. [doi: 10.1177/0884217504266915] [Medline: 15346666]
28. Horowitz JA, Goodman J. A longitudinal study of maternal postpartum depression symptoms. Res Theory Nurs Pract 2004 Jun 01;18(2-3):149-163. [doi: 10.1891/rtnp.18.2.149.61285] [Medline: 15553344]
29. Machado-Vieira R, Baumann J, Wheeler-Castillo C, Latov D, Henter ID, Salvadore G, et al. The timing of antidepressant effects: a comparison of diverse pharmacological and somatic treatments. Pharmaceuticals (Basel) 2010 Jan 06;3(1):19-41 [FREE Full text] [doi: 10.3390/ph3010019] [Medline: 27713241]
30. Carlson K, Mughal S, Azhar Y, Siddiqui W. Perinatal depression. In: StatPearls [Internet]. Treasure Island, FL: StatPearls Publishing; 2025.
31. The basics. All of Us Research Hub. URL: https://tinyurl.com/2hb323a5 [accessed 2024-12-02]
32. Master H, Annis J, Huang S, Beckman JA, Ratsimbazafy F, Marginean K, et al. Association of step counts over time with the risk of chronic disease in the All of Us Research Program. Nat Med 2022 Nov 10;28(11):2301-2308 [FREE Full text] [doi: 10.1038/s41591-022-02012-w] [Medline: 36216933]
33. Lenth RV. R package emmeans: estimated marginal means. GitHub. URL: https://rvlenth.github.io/emmeans/ [accessed 2025-05-12]
34. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. J Stat Softw 2015;67(1):1-48. [doi: 10.18637/jss.v067.i01]
35. Kuznetsova A, Brockhoff PB, Christensen RH. lmerTest package: tests in linear mixed effects models. J Stat Softw 2017;82(13):1-26. [doi: 10.18637/jss.v082.i13]
36. Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. BMJ 1995 Jan 21;310(6973):170 [FREE Full text] [doi: 10.1136/bmj.310.6973.170] [Medline: 7833759]
37. Benesty J, Chen J, Huang Y, Cohen I. Pearson correlation coefficient. In: Noise Reduction in Speech Processing. Berlin, Heidelberg: Springer; 2009.
38. Kassambara A. ggpubr: 'ggplot2' based publication ready plots. The Comprehensive R Archive Network. 2023 Feb 10. URL: https://cran.r-project.org/web/packages/ggpubr/index.html [accessed 2024-05-24]
39. Iranpour S, Kheirabadi GR, Esmaillzadeh A, Heidari-Beni M, Maracy MR. Association between sleep quality and postpartum depression. J Res Med Sci 2016;21(1):110 [FREE Full text] [doi: 10.4103/1735-1995.193500] [Medline: 28250787]
40. Lawson A, Murphy KE, Sloan E, Uleryk E, Dalfen A. The relationship between sleep and postpartum mental disorders: a systematic review. J Affect Disord 2015 May 01;176:65-77. [doi: 10.1016/j.jad.2015.01.017] [Medline: 25702602]

XSL•FO
RenderX

41. Lewis BA, Gjerdingen D, Schuver K, Avery M, Marcus BH. The effect of sleep pattern changes on postpartum depressive symptoms. BMC Womens Health 2018 Jan 09;18(1):12 [FREE Full text] [doi: 10.1186/s12905-017-0496-6] [Medline: 29316912]
42. McEvoy KM, Rayapati D, Washington Cole KO, Erdly C, Payne JL, Osborne LM. Poor postpartum sleep quality predicts subsequent postpartum depressive symptoms in a high-risk sample. J Clin Sleep Med 2019 Sep 15;15(9):1303-1310 [FREE Full text] [doi: 10.5664/jcsm.7924] [Medline: 31538601]
43. Okun ML. Disturbed sleep and postpartum depression. Curr Psychiatry Rep 2016 Jul 25;18(7):66. [doi: 10.1007/s11920-016-0705-2] [Medline: 27222140]
44. Park EM, Meltzer-Brody S, Stickgold R. Poor sleep maintenance and subjective sleep quality are associated with postpartum maternal depression symptom severity. Arch Womens Ment Health 2013 Dec 4;16(6):539-547 [FREE Full text] [doi: 10.1007/s00737-013-0356-9] [Medline: 23733081]
45. Posmontier B. Sleep quality in women with and without postpartum depression. J Obstet Gynecol Neonatal Nurs 2008 Nov;37(6):722-35; quiz 735 [FREE Full text] [doi: 10.1111/j.1552-6909.2008.00298.x] [Medline: 19012723]
46. Thurgood S, Avery DM, Williamson L. Postpartum Depression (PPD). Am J Clin Med 2009;6(2):17-22 [FREE Full text]
47. Haga SM, Ulleberg P, Slinning K, Kraft P, Steen TB, Staff A. A longitudinal study of postpartum depressive symptoms: multilevel growth curve analyses of emotion regulation strategies, breastfeeding self-efficacy, and social support. Arch Womens Ment Health 2012 Jun 27;15(3):175-184. [doi: 10.1007/s00737-012-0274-2] [Medline: 22451329]
48. Vliegen N, Casalin S, Luyten P. The course of postpartum depression: a review of longitudinal studies. Harv Rev Psychiatry 2014;22(1):1-22. [doi: 10.1097/HRP.0000000000000013] [Medline: 24394219]
49. Drozd F, Haga SM, Valla L, Slinning K. Latent trajectory classes of postpartum depressive symptoms: a regional population-based longitudinal study. J Affect Disord 2018 Dec 01;241:29-36. [doi: 10.1016/j.jad.2018.07.081] [Medline: 30096589]
50. Liao Y, Robertson MC, Winne A, Wu IH, Le TA, Balachandran DD, et al. Investigating the within-person relationships between activity levels and sleep duration using Fitbit data. Transl Behav Med 2021 Mar 16;11(2):619-624 [FREE Full text] [doi: 10.1093/tbm/ibaa071] [Medline: 32667039]
51. Singh B, Spence RR, Sandler CX, Tanner J, Hayes SC. Feasibility and effect of a physical activity counselling session with or without provision of an activity tracker on maintenance of physical activity in women with breast cancer - a randomised controlled trial. J Sci Med Sport 2020 Mar;23(3):283-290. [doi: 10.1016/j.jsams.2019.09.019] [Medline: 31640924]
52. Sarhaddi F, Azimi I, Labbaf S, Niela-Vilén H, Dutt N, Axelin A, et al. Long-term IoT-based maternal monitoring: system design and evaluation. Sensors (Basel) 2021 Mar 24;21(7):2281 [FREE Full text] [doi: 10.3390/s21072281] [Medline: 33805217]
53. Grym K, Niela-Vilén H, Ekholm E, Hamari L, Azimi I, Rahmani A, et al. Feasibility of smart wristbands for continuous monitoring during pregnancy and one month after birth. BMC Pregnancy Childbirth 2019 Jan 17;19(1):34 [FREE Full text] [doi: 10.1186/s12884-019-2187-9] [Medline: 30654747]
54. Pedrelli P, Fedor S, Ghandeharioun A, Howe E, Ionescu DF, Bhathena D, et al. Monitoring changes in depression severity using wearable and mobile sensors. Front Psychiatry 2020 Dec 18;11:584711 [FREE Full text] [doi: 10.3389/fpsyt.2020.584711] [Medline: 33391050]
55. Chandrasekaran R, Katthula V, Moustakas E. Patterns of use and key predictors for the use of wearable health care devices by US adults: insights from a national survey. J Med Internet Res 2020 Oct 16;22(10):e22443 [FREE Full text] [doi: 10.2196/22443] [Medline: 33064083]
56. Berryhill S, Morton CJ, Dean A, Berryhill A, Provencio-Dean N, Patel SI, et al. Effect of wearables on sleep in healthy individuals: a randomized crossover trial and validation study. J Clin Sleep Med 2020 May 15;16(5):775-783 [FREE Full text] [doi: 10.5664/jcsm.8356] [Medline: 32043961]
57. Guillodo E, Lemey C, Simonnet M, Walter M, Baca-García E, Masetti V, HUGOPSY Network, et al. Clinical applications of mobile health wearable-based sleep monitoring: systematic review. JMIR Mhealth Uhealth 2020 Apr 01;8(4):e10733 [FREE Full text] [doi: 10.2196/10733] [Medline: 32234707]
58. Wei JJ, Hayward EM, Leung TH, Rosenbach M. Unbearable wearables. Dermatol Online J 2019 Dec 15;25(12):13030/qt9rv5r4x6 [FREE Full text] [Medline: 32045171]
59. Nagappan A, Krasniansky A, Knowles M. Patterns of ownership and usage of wearable devices in the United States, 2020-2022: survey study. J Med Internet Res 2024 Jul 26;26:e56504 [FREE Full text] [doi: 10.2196/56504] [Medline: 39058548]
60. Browne JD, Boland DM, Baum JT, Ikemiya K, Harris Q, Phillips M, et al. Lifestyle modification using a wearable biometric ring and guided feedback improve sleep and exercise behaviors: a 12-month randomized, placebo-controlled study. Front Physiol 2021;12:777874 [FREE Full text] [doi: 10.3389/fphys.2021.777874] [Medline: 34899398]
61. Wearable devices. Care Evolution. URL: https://careevolution.com/white-papers/choosing-a-commercially-available-wearable-device-for-your-research-study/ [accessed 2024-05-23]
62. Nolasco HR, Vargo A, Bohley N, Brinkhaus C, Kise K. Examining participant adherence with wearables in an in-the-wild setting. Sensors (Basel) 2023 Jul 18;23(14):6479 [FREE Full text] [doi: 10.3390/s23146479] [Medline: 37514773]

XSL•FO
RenderX

63. Shiba SK, Temple CA, Krasnoff J, Dilchert S, Smarr BL, Robishaw J, et al. Assessing adherence to multi-modal oura ring wearables from COVID-19 detection among healthcare workers. Cureus 2023 Sep;15(9):e45362 [FREE Full text] [doi: 10.7759/cureus.45362] [Medline: 37849583]
64. Khadka N, Fassett MJ, Oyelese Y, Mensah NA, Chiu VY, Yeh M, et al. Trends in postpartum depression by race, ethnicity, and prepregnancy body mass index. JAMA Netw Open 2024 Nov 04;7(11):e2446486 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.46486] [Medline: 39565621]
65. Luciano M, Di Vincenzo M, Brandi C, Tretola L, Toricco R, Perris F, et al. Does antenatal depression predict post-partum depression and obstetric complications? Results from a longitudinal, long-term, real-world study. Front Psychiatry 2022;13:1082762 [FREE Full text] [doi: 10.3389/fpsyt.2022.1082762] [Medline: 36590632]
66. ChatGPT homepage. ChatGPT. URL: https://chatgpt.com/ [accessed 2025-01-29]
67. Claude homepage. Claude. URL: https://claude.ai/new [accessed 2025-01-29]

Abbreviations

AoURP: *All of Us* Research Program
DV: direct volunteer
EHR: electronic health record
EPDS: Edinburgh-Postnatal Depression Scale
HPO: health care provider organization
ML: machine learning
PPD: postpartum depression

Edited by A Mavragani; submitted 15.10.24; peer-reviewed by F Sarhaddi, M La Verde; comments to author 23.01.25; revised version received 10.02.25; accepted 02.05.25; published 23.05.25.

Please cite as:
Hurwitz E, Meltzer-Brody S, Butzin-Dozier Z, Patel RC, Elhadad N, Haendel MA
Unlocking the Potential of Wear Time of a Wearable Device to Enhance Postpartum Depression Screening and Detection: Cross-Sectional Study
JMIR Form Res 2025;9:e67585
URL: https://formative.jmir.org/2025/1/e67585
doi:10.2196/67585
PMID:

©Eric Hurwitz, Samantha Meltzer-Brody, Zachary Butzin-Dozier, Rena C Patel, Noémie Elhadad, Melissa A Haendel. Originally published in JMIR Formative Research (https://formative.jmir.org), 23.05.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Data-Driven Approach to Assessing Hepatitis B Mother-to-Child Transmission Risk Prediction Model: Machine Learning Perspective

Dung Nguyen Tien[1], MD, PhD; Huong Thi Thu Bui[1,2], MD, PhD; Tram Hoang Thi Ngoc[1], MD, PhD; Thuy Thi Pham[1], MD, PhD; Dac Trung Nguyen[1], MD, PhD; Huyen Nguyen Thi Thu[1], MD, PhD; Thi Thu Hang Vu[1], MD, PhD; Thi Lan Anh Luong[3,4], MD, PhD; Lan Thu Hoang[3,4], MSc; Ho Cam Tu[3,5], MSc; Nina Körber[6], PhD; Tanja Bauer[6], PhD; Lam Khanh Ho[7], PhD

[1]Department of Microbiology, Thai Nguyen University of Medicine and Pharmacy, Thái Nguyên, Vietnam
[2]Department of Immunology - Molecular Genetics, Thai Nguyen National General Hospital, Thái Nguyên, Vietnam
[3]Department of MBG, Hanoi Medical University, Hanoi, Vietnam
[4]Center of Clinical Genetics and Genomics, Hanoi Medical University Hospital, Hanoi, Vietnam
[5]Technical University of Munich, Munich, Germany
[6]Institute of Virology (VIRO), Molecular Targets and Therapeutics Center, Helmholtz Zentrum München, Munich, Germany
[7]Faculty of Information Technology, Hung Yen University of Technology and Education, H ng Yên, Vietnam

Corresponding Author:
Huong Thi Thu Bui, MD, PhD
Department of Microbiology
Thai Nguyen University of Medicine and Pharmacy
284 L ng Ng c Quy n
Thái Nguyên, 250000
Vietnam
Phone: 84 912916863
Email: huongbuithithu@tnmc.edu.vn

Abstract

Background: Hepatitis B virus (HBV) can be transmitted from mother to child either through transplacental infection or via blood-to-blood contact during or immediately after delivery. Early and accurate risk assessments are essential for guiding clinical decisions and implementing effective preventive measures. Data mining techniques are powerful tools for identifying key predictors in medical diagnostics.

Objective: This study aims to develop a robust predictive model for mother-to-child transmission (MTCT) of HBV using decision tree algorithms, specifically Iterative Dichotomiser 3 (ID3) and classification and regression trees (CART). The study identifies clinically and paraclinically relevant predictors, particularly hepatitis B e antigen (HBeAg) status and peripheral blood mononuclear cell (PBMC) concentration, for effective risk stratification and prevention. Additionally, we will assess the model's reliability and generalizability through cross-validation with various training-test split ratios, aiming to enhance its applicability in clinical settings and inform improved preventive strategies against HBV MTCT.

Methods: This study used decision tree algorithms—ID3 and CART—on a data set of 60 hepatitis B surface antigen (HBsAg)–positive pregnant women. Samples were collected either before or at the time of delivery, enabling the inclusion of patients who were undiagnosed or had limited access to treatment. We analyzed both clinical and paraclinical parameters, with a particular focus on HBeAg status and PBMC concentration. Additional biochemical markers were evaluated for their potential contributory or inhibitory effects on MTCT risk. The predictive models were validated using multiple training-test split ratios to ensure robustness and generalizability.

Results: Our analysis showed that 20 out of 48 (based on a split ratio of 0.8 from a total of 60 cases, 42%) to 27 out of 57 (based on a split ratio of 0.95 from a total of 60 cases, 47%) training cases with HBeAg-positive status were associated with a significant risk of MTCT of HBV (χ^2_8=21.16, *P*=.007, *df*=8). Among HBeAg-negative women, those with PBMC concentrations ≥8 × 10^6 cells/mL exhibited a low risk of MTCT, whereas individuals with PBMC concentrations <8 × 10^6 cells/mL demonstrated a

XSL•FO
RenderX

negligible risk. Across all training-test split ratios, the decision tree models consistently identified HBeAg status and PBMC concentration as the most influential predictors, underscoring their robustness and critical role in MTCT risk stratification.

Conclusions: This study demonstrates that decision tree models are effective tools for stratifying the risk of MTCT of HBV by integrating key clinical and paraclinical markers. Among these, HBeAg status and PBMC concentration emerged as the most critical predictors. While the analysis focused on untreated patients, it provides a strong foundation for future investigations involving treated populations. These findings offer actionable insights to support the development of more targeted and effective HBV MTCT prevention strategies.

(JMIR Form Res 2025;9:e69838) doi:10.2196/69838

KEYWORDS

chronic hepatitis B virus infection; liver; pregnant women; cord blood; PBMCs (peripheral blood mononuclear cells); ID3 (Iterative Dichotomiser 3); CART (classification and regression trees)

Introduction

Hepatitis B virus (HBV) mother-to-child transmission (MTCT) can occur through transplacental infection or blood-to-blood contact during or after delivery and accounts for a significant proportion of chronic HBV infections worldwide [1]. In high-prevalence countries such as Vietnam, MTCT remains the most common mode of transmission. Children who acquire chronic infection have a 40% lifetime risk of dying from HBV-related complications. Antiviral treatment to reduce high viral loads, along with immunoprophylaxis using anti-HBV immunoglobulin shortly after delivery and active HBV immunization of neonates, can significantly reduce the incidence of MTCT. However, these measures are rarely implemented in most countries [2]. Studies examining high-risk factors associated with MTCT in patients with HBV lack depth, and the relationship between HBV DNA in maternal serum and cord blood remains unclear. We conducted a clustering study to explore the potential link between HBV infection in pregnant women and cord blood, aiming to identify a clinical reference marker for prenatal surveillance and postnatal management, and to strengthen the prevention of HBV MTCT.

Data mining techniques play a crucial role in clinical decision-making by providing physicians with accurate, reliable, and timely predictions through various models. Machine learning is broadly categorized into 3 types: supervised learning, unsupervised learning, and reinforcement learning. The decision tree is a supervised learning algorithm capable of handling both regression and classification tasks. A typical machine learning algorithm involves 2 main steps: training (where the algorithm learns a model from data) and prediction (where the learned model is used to predict new values). The training step in the decision tree algorithm constructs a decision tree. A decision tree is an effective support tool for engineers' decision-making [3], using a tree model that illustrates decisions and their possible outcomes, including random outcomes, resource costs, and benefits. Based on published premise cluster analysis research, we set up a machine learning experiment. This research may help identify potential risk factors for MTCT.

This study was conducted at a single center in Thai Nguyen, Vietnam. It is valuable to explore how variations in HBV circulation rates, genotypes, and health care practices across different regions of the country might influence both the detection and relevance of the study. Thai Nguyen, where the research took place, serves as the economic, political, and social hub of the northeastern region, the Central Highlands, and the northern mountainous areas of Vietnam. As a result, although the patients come from various hometowns, they share characteristics typical of the northern mountainous region, where many ethnic minorities reside. Viral hepatitis—particularly HBV, hepatitis D virus, and hepatitis E virus—remains a significant public health concern in Vietnam, especially among ethnic minority communities. Higher infection rates in these groups are often linked to limited access to health care, low socioeconomic status, high-risk living conditions, and a lack of awareness about the disease.

In this study, we enrolled 60 pregnant women who tested positive for hepatitis B surface antigen (HBsAg) from a clinical setting, applying strict inclusion criteria to ensure that only those with chronic HBV infection—defined as being HBsAg-positive for more than 6 months—and who had not received any HBV treatment were included.

In our study, we focused exclusively on HBsAg-positive pregnant women who were not receiving antiviral treatment. This approach was intentionally chosen to better understand the natural history and intrinsic risk factors for MTCT in untreated patients. According to World Health Organization (WHO) guidelines and current Vietnamese recommendations, pregnant women identified as being at high risk of MTCT are typically advised to begin treatment early. However, because sampling in our study took place either before delivery or immediately at the time of delivery, some participants had not yet been diagnosed or had not had the opportunity to access treatment.

After delivery, we strongly recommended that all patients initiate antiviral therapy promptly, emphasizing that postnatal treatment does not negatively impact milk production or the quality of breast milk for the infant. Although excluding women who were already receiving treatment may limit the generalizability of our findings, our study offers valuable baseline data on MTCT risk in untreated patients. This baseline can serve as a crucial reference for future studies evaluating the effectiveness of early treatment interventions.

After obtaining informed consent, we collected maternal and cord blood samples to measure various biomarkers, including viral markers (HBsAg, hepatitis B e antigen [HBeAg], HBV DNA) and biochemical parameters (eg, alanine aminotransferase [ALT], aspartate aminotransferase [AST], and peripheral blood

XSL•FO
RenderX

mononuclear cells [PBMCs]). The raw data were preprocessed using R (R Foundation), which involved cleaning and organizing the data for further analysis. We conducted descriptive and univariate analyses to summarize the characteristics of the study population, followed by correlation and clustering analyses to explore relationships among the measured variables. Based on logistic regression models and initial analyses, we identified specific cutoff values (eg, PBMCs/mL and HBV DNA copies/mL). Feature selection was performed using information gain techniques based on Iterative Dichotomiser 3 (ID3) theory, identifying key predictors such as HBeAg status and PBMC levels. The selected features were then used to construct decision tree models with both the ID3 and classification and regression trees (CART) algorithms to classify participants into distinct MTCT risk categories.

We validated the models through 1000 simulation runs using various training-test split ratios, calculating performance metrics—including accuracy, sensitivity, specificity, and AUC—to assess model performance. The validated models enabled risk stratification based on Cohen effect size classifications (trivial, small, medium, and large), and the results were interpreted to provide actionable clinical insights, highlighting the critical roles of HBeAg and PBMCs in predicting the risk of HBV MTCT (Figure 1).

The goal of this study is to develop and validate a machine learning–based decision tree model to effectively predict the risk of HBV transmission from mother to child. This will be accomplished by incorporating key clinical and paraclinical markers—particularly HBeAg status and PBMC concentration—to inform targeted prevention strategies.

Figure 1. Study flowchart. CART: classification and regression trees; HBsAg: hepatitis B surface antigen; ID3: Iterative Dichotomiser 3.

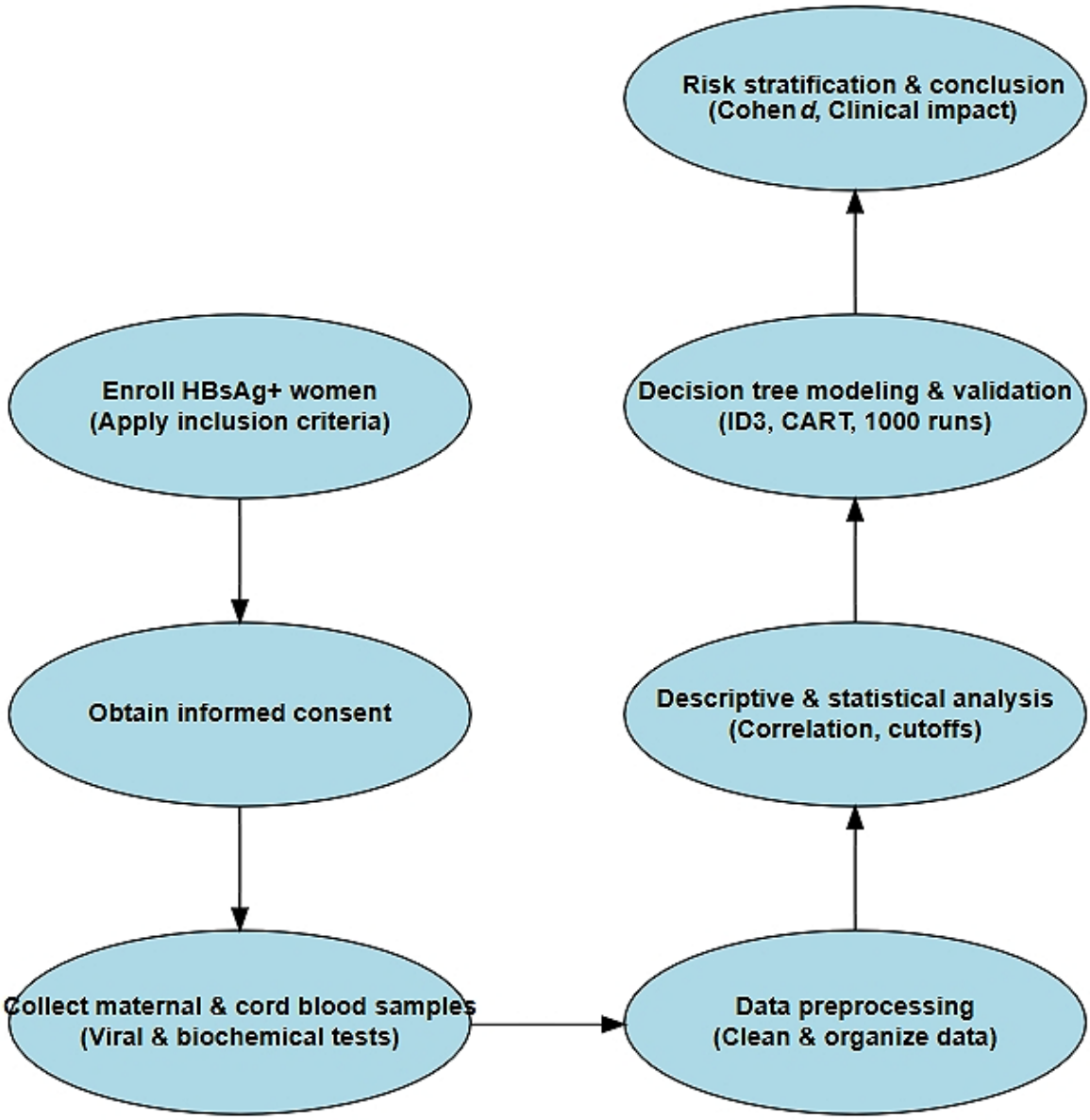

XSL•FO
RenderX

Methods

Study Population

Between 2020 and 2021, we conducted a pilot study at Thai Nguyen National Hospital in Vietnam, focusing on pregnant women who received regular check-ups and delivered at the facility. A total of 60 pregnant women who tested positive for HBsAg were enrolled.

Participants were provided with detailed information about the risks of hepatitis B transmission and were counseled on available treatment options. They could choose to begin treatment immediately or defer it until after delivery. Following birth, both mothers and their newborns were monitored for 3-6 months. Newborns received routine prophylactic care, and mothers with HBV DNA levels exceeding 200,000 IU/mL were offered treatment in accordance with WHO guidelines (2021) [4].

Only pregnant women with chronic hepatitis B (CHB) infection (HBsAg positive for more than 6 months) who opted to delay antiviral treatment were included in the study. We collected various biological samples and conducted clinical surveillance to establish a comprehensive hepatitis B patient cohort, aimed at exploring the multifactorial risk factors associated with HBV MTCT.

Ethical Considerations

This study was approved by the Institutional Review Board for Ethics in Biomedical Research of Hanoi Medical University (approval number NCS22/HMU-IRD), ensuring the protection of participants' rights. Informed consent for future data use was obtained during the original data collection, and this secondary analysis was conducted in accordance with that provision. All data used in this study were deidentified and securely stored. No personally identifiable information was collected or retained, and no identifiable images of individuals are included in the manuscript or multimedia appendices.

Diagnostic Criteria

We collected clinical data from pregnant women who tested positive for HBsAg and were not receiving HBV treatment. The information gathered included general demographic and obstetric details such as age at delivery, gestational age, and pregnancy history (including the number of pregnancies and births, cesarean sections, and any preexisting conditions). Laboratory test results—such as ALT, AST, HBV DNA levels, and records of any antiviral therapy—were also included. Additionally, we assessed various maternal conditions, including preeclampsia, chronic hypertension, history of abortion, placental abruption, hyperthyroidism, gestational diabetes mellitus, pregnancy-induced hypertension, intrahepatic cholestasis of pregnancy, and other related complications. A retrospective analysis was conducted to examine the association between HBV infection status and preclinical factors.

Isolation of Peripheral Blood Mononuclear Cells

We collected maternal blood in EDTA (ethylenediaminetetraacetic acid) tubes and also collected umbilical cord blood. The cord blood was obtained from the umbilical vein of the umbilical cord and placed in a 20-mL cylinder with anticoagulant immediately after birth. The process of collecting cord blood takes only 2-3 minutes and involves the following steps: (1) Immediately after the mother gives birth, the medical staff clamps a section of the umbilical cord that is at least 10 cm long. (2) This section can be cut immediately to obtain a blood sample or left intact until after the placenta is delivered. (3) The surface of the umbilical cord is disinfected with povidone-iodine solution. (4) The needle of the collection cylinder is then inserted into the umbilical vein to draw the blood. (5) The collection cylinder is clamped, and the needle is withdrawn. (6) Finally, the collection cylinder is gently shaken to mix the blood with the anticoagulant.

Serum and plasma samples were analyzed for viral markers (HBsAg, HBeAg, and HBV DNA copies/mL) and other preclinical factors, including platelet count ($\times 10^3$ cells/mL), prothrombin time (seconds), prothrombin ratio (%), hemoglobin (g/L), red blood cell count ($\times 10^6$ cells/mL), creatinine (µmol/L), AST (U/L), and ALT (U/L) [5].

Statistical and Decision Tree Analysis

Data collection, storage, and analysis in this study were conducted using the R 4.1.0 package tools. The correlation R value measures the strength of the linear relationship between 2 quantitative variables. The Pearson R formula is as follows:

where R is the Pearson correlation coefficient; and x and y are 2 vectors of length i and j, respectively [6]. The value of R ranges from –1 to 1, with $R>0$ indicating a positive association and $R<0$ indicating a negative association [6].

We conducted the clustering analysis again using the 5 most significant factors identified in our earlier study [7-9]. To determine the effect size for each factor, we transformed the natural logarithm of the odds ratio, that is, ln(odds ratio), by dividing it by 1.81, based on their respective odds ratios [10].

Additionally, we utilized the ID3 algorithm to generate general rules and predictions for new cases. This algorithm requires specifying the order in which attributes are evaluated at each step. As finding the optimal solution can be challenging when dealing with numerous attributes (such as varied patient test results), we opted for a simpler approach: at each step, we selected the attribute that best satisfied a chosen criterion. After selecting an attribute, the data are divided into child nodes based on their values, and this process continues recursively for each child node. Although this greedy selection method may not always yield the optimal solution, it is intuitively close to the best outcome and significantly simplifies the problem [11].

A crucial element of this approach is assessing the quality of each partition. Ideally, a good partition is one in which each child node predominantly contains data from a single class, allowing it to be treated as a leaf with no further division. Conversely, a partition that produces child nodes with mixed classes is less desirable. To evaluate this, we require a function that measures the purity or impurity of a partition. This function should yield the lowest value when each node contains data

XSL•FO
RenderX

from only 1 class (indicating high purity), and a higher value when nodes include a diverse mix of classes. The entropy function, commonly used in information theory, serves this purpose (see equation 2).

In the ID3 algorithm, the loss function for a decision tree is defined as the weighted sum of the entropies at its leaf nodes, with the weights corresponding to the number of data points in each node. The goal of ID3 is to determine the order of attribute splits in a way that minimizes this total loss. This is achieved by selecting the attribute that leads to the greatest reduction in entropy at each step. Essentially, constructing the decision tree using ID3 can be viewed as a series of smaller tasks, where at each nonleaf node, we choose the attribute that most effectively improves the split. We then develop a calculation method for each of these nodes (see equation 3). Thus, the entropy at this node is given by:

Next, suppose the selected attribute is x. We define it as the weighted sum of the entropy of each *child node*, computed similarly to equation 3. This weighting is important because *nodes often* contain different numbers of points (see equation 4).

Let the selected attribute be x. Based on x, the data points in S are divided into *Kchildnode* S_1, S_2, ..., $S_{K,}$ with the number of points in each *child node* being m_1, m_2, ..., m_K, respectively.

Next, we define the *information gain* based on the attribute x, as given by equation 5.

$G(x, S) = H(S) - H(x,S)$ **(5)**

Let the selected attribute be x. Based on x, the data points in S are divided into *Kchildnode* S_1, S_2, ..., S_K, with the number of points in each *child node* being m_1, m_2, ..., m_K, respectively. In equation 5, $H(s)$ is the *root node entropy.* In ID3, at each node, the selected attribute is determined based on equation 6, which identifies the property that maximizes the information gain.

Identifying important variables helps eliminate less important ones, simplifying the model and reducing noise. Thus, the final step determines the importance of each variable using a different splitting ratio. These variables can be quantified by the reduction in impurity (such as the Gini index) achieved when they are used for splitting. This imputation follows the CART formula (see equations 7 and 8).

Let us suppose an object is selected at random from one of the C classes according to the probabilities (p_1, p_2, ..., p_C) and is randomly assigned to a class using the same distribution. In this scenario, we get the following:

In equation 8, let $L(i, j)$ be the loss of assigning class j to an object which actually belongs to class i. The expected cost of misclassification is

In our study, we simulate 1000 runs for each splitting ratio. The best variable is the one that shows the highest Gini index score. The Gini index has been adapted to assess health inequality across populations by providing estimates that capture the distribution of risk, or lack of risk, among the entire population or within specific groups [12].

Results

Power Estimation in a Multivariate Regression

We generate a power estimation curve for a multivariate regression (using f^2 as the effect size measure) with a total sample size of 60 and 5 predictors. This plot helps illustrate how adequate—or inadequate—the sample size is for detecting effects in a multivariate context. In multiple regression, Cohen [13] suggested the following guidelines for f^2 effect sizes: small (f^2=0.002), medium (f^2=0.15), and large (f^2=0.35). We calculate the statistical power for a range of f^2 values given the following: total sample size N=60, number of predictors p=5, degrees of freedom for the error $v=N-p-1=54$, and significance level α=.05. Figure 2 illustrates how statistical power changes in a multiple regression model (with 5 predictors and a total sample size of 60) as the effect size (f^2) ranges from small (0.02) to large (0.35). The vertical axis shows the probability (power) of detecting an actual effect at the 5% significance level, while the horizontal axis shows the size of that effect. The red dashed line at 0.80 marks the conventional threshold for sufficient power (80%). When (f^2) is small (0.02), the power is around 0.2, indicating only a 20% chance of detecting such a minor effect with the given sample size and number of predictors. As (f^2) approaches medium (0.15), the power increases but remains below the 0.80 line, suggesting that moderate effects are not reliably detected. Only when (f^2) nears the large range (around 0.35) does the power surpass or approach the 80% mark, implying that the study can reliably detect larger effects but may struggle to identify smaller or moderate ones (Figure 2).

XSL•FO
RenderX

Figure 2. Power estimation in a multivariate regression with 60 participants and 5 predictors. We set n=60 (sample size) and p=5 (number of predictors). The error degrees of freedom are calculated as v=60 − 5 − 1 = 54. A sequence of f^2 values ranging from 0.02 (small) to 0.35 (large) is generated. For each f^2 value, we use pwr.f2.test() to estimate the power of the overall F-test in a multiple regression. The y-axis shows statistical power, and the x-axis shows the effect size (f^2). A horizontal red dashed line marks the 80% power threshold. Conventional effect sizes (small, medium, and large) are highlighted with dark green points and labeled accordingly.

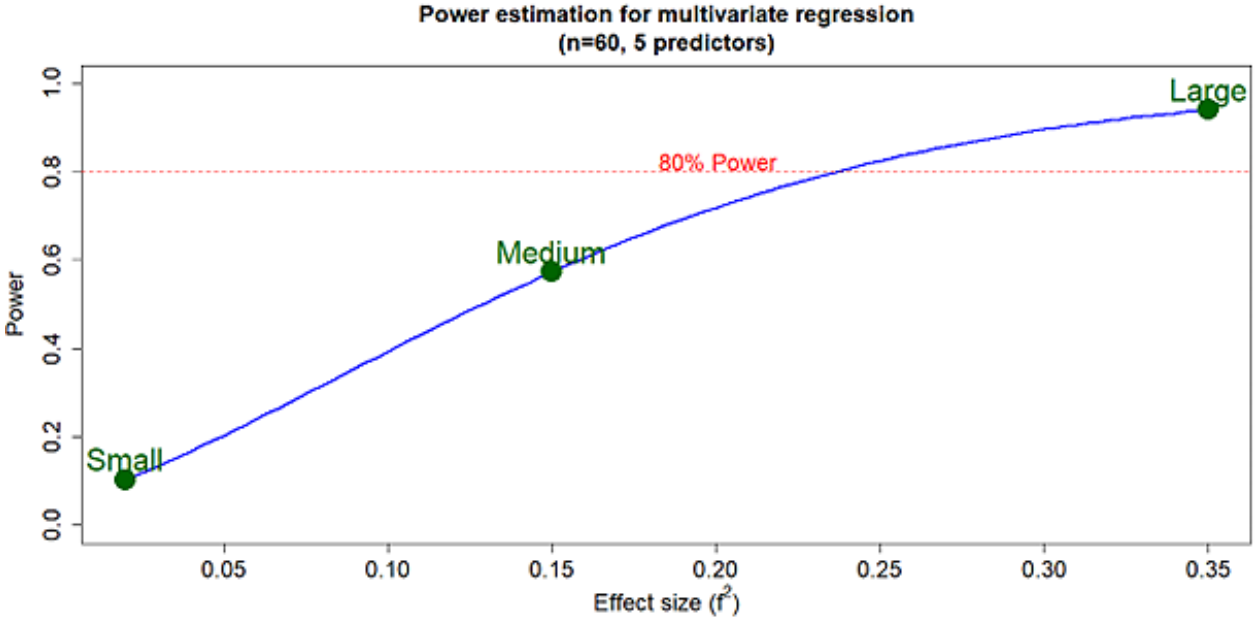

The Five Most Important Factors of MTCT Risk: HBeAg, ALT, AST, HBV DNA, and PBMCs

The mean age is 27.6 (SD 4.2) years. Pregnancy for the second time or more accounted for 43 out of 60 (72%) cases. Clinical symptoms were edema, fatigue, and loss of appetite, similar to those experienced during pregnancy; 40 (67%) women had cesarean section. Of the 60 women, 27 (45%) were HBeAg positive, and HBV DNA ≥ 10^7 copies/mL was reported in 20 (33%). The mean gestational age is 38.9 (SD 1.2) weeks. Among the babies, 32 (53%) were HBsAg positive, and 23 (38%) were HBeAg positive (Tables 1-4).

In 60 Vietnamese CHB pregnant women, 32 (53%) cord-blood samples were HBsAg positive, and 28 (47%) were HBsAg negative. We fit a logistic regression model to predict the subclinical values (cutoff value), which correspond to 50:50 probabilities that HBsAg in cord blood is positive. Two variables have a positive association when above-average values of one tend to accompany above-average values of the other, and below-average values tend to occur together as well. Two variables have a negative association when above-average values of one tend to accompany below-average values of the other [14] (see Table 5 and Figure S1 in Multimedia Appendix 1).

Table 1. General characteristics of 60 Vietnamese pregnant women with chronic hepatitis B[a].

General characteristics of the study participants	Values, n (%)
Mother's age (years)	
18-35	58 (97)
>35	2 (3)
Mean (SD)	27.6 (4.2)
Number of pregnancies	
First time	17 (28)
From the second time	43 (72)
Time of detection of hepatitis B virus infection	
Before getting pregnant	25 (42)
This time	35 (58)

[a]32 (53%) cord blood samples were hepatitis B surface antigen positive, and 28 (47%) hepatitis B surface antigen negative.

XSL•FO
RenderX

Table 2. Clinical and preclinical characteristics of the 60 Vietnamese pregnant women with chronic hepatitis B.

Characteristics	Values, n (%)
Clinical characteristics	
Edema	7 (12)
Tired	8 (13)
Anorexia	1 (2)
Nausea/vomiting	4 (7)
Insomnia	4 (7)
Joint pain	3 (5)
Right lower quadrant pain	2 (3)
Birth method	
Birth without episiotomy	4 (7)
Birth with episiotomy	16 (27)
Caesarean section	40 (67)
Preclinical characteristics	
An increase of aspartate aminotransferase	14 (23)
An increase of alanine aminotransferase	11 (18)
Hepatitis B e antigen positive	27 (45)
Hepatitis B e antigen negative	33 (55)
Hepatitis B virus DNA≥10^7	20 (33)
Hepatitis B virus DNA<10^7	40 (67)

Table 3. Clinical and preclinical characteristics of babies (N=60).

Characteristics	Values
Clinical	
Apgar score≥8, n (%)	58 (97)
Apgar score<7, n (%)	2 (3)
Gestational age, mean (SD)	38.9 (1.2)
Weight (g)	
<3500, n (%)	47 (78)
≥3500, n (%)	13 (22)
Mean (SD)	3198.3 (362.9)
Preclinical (cord blood), n (%)	
Hepatitis B surface antigen positive	32 (53)
Hepatitis B surface antigen negative	28 (47)
Hepatitis B e antigen positive	23 (38)
Hepatitis B e antigen negative	37 (62)

XSL•FO
RenderX

Table 4. Preclinical measures.

Details of preclinical measures results	Range	Median (95% CI)	Mean (SD)
Maternal prothrombin time (seconds) (N=60)	10.5-47.8	12.1 (11.6-12.8)	12.9 (4.76)
Maternal prothrombin ratio (%) (N=60)	10.0-133.0	108.0 (99.6-117.0)	106 (21.6)
Maternal red blood cells ($\times 10^6$ cells/mL) (N=60)	3.25-6.25	4.33 (4.06-4.69)	4.38 (0.539)
Maternal platelet count ($\times 10^3$ cells/mL) (N=60)	109.0-344.0	214.0 (180.0-260.0)	219 (59.6)
Maternal creatinine (µmol/L) (N=60)	41.0-99.0	59.6 (54.4-68.8)	62.1 (11.1)
Maternal aspartate transaminase (U/L) (N=60)	13.0-285.0	22.1 (18.6-29.2)	33.2 (41.3)
Maternal alanine transaminase (U/L) (N=60)	6.40-217.0	15.8 (12.8-24.3)	27.7 (38.0)
Maternal protein in blood (g/L)	12.3-80.2	67.9 (65.8-71.1)	66.7 (9.82)
Maternal albumin in blood (g/L)	25.6-44.9	34.5 (33.0-35.7)	34.4 (3.54)
Maternal hepatitis B virus DNA (copies/mL) (N=60)	35.0-1,350,000,000	49,600 (771.0-129,000,000)	152,000,000 (310,000,000)
Maternal peripheral blood mononuclear cells (cells/mL) (N=60)	1,300,000-12,300,000	5,500,000 (3,000,000-7,510,000)	5,620,000 (3,050,000)
Cord blood mononuclear cell concentration (cells/mL) (N=60)	3,640,000-51,000,000	12,200,000 (6,500,000-15,000,000)	12,600,000 (7,630,000)

Table 5. Cutoff values.

Variables	$Cutoff_{50}$[a] for hepatitis B surface antigen–positive probability in cord blood	The direction of the relationship between the 2 variables	Cross-reference
Maternal peripheral blood mononuclear cell concentration	8.03×10^6 cells/mL	Negative	Figure S1A in Multimedia Appendix 1
Maternal hepatitis B virus DNA	5.40×10^7 copies/mL	Positive	Figure S1B in Multimedia Appendix 1
Maternal platelet count	317.89×10^3 cells/mL	Negative	Figure S1C in Multimedia Appendix 1
Maternal prothrombin time	11.00 seconds	Positive	Figure S1D in Multimedia Appendix 1
Cord blood mononuclear cell concentration	6.64×10^6 cells/mL	Positive	Figure S1E in Multimedia Appendix 1
Maternal hemoglobulin	128.53 g/L	Negative	Figure S1F in Multimedia Appendix 1
Maternal red blood cells	5×10^6 cells/mL	Negative	Figure S1G in Multimedia Appendix 1
Maternal creatinine	37.46 µmol/L	Positive	Figure S1H in Multimedia Appendix 1
Maternal aspartate aminotransferase	14.15 U/L	Positive	Figure S1I in Multimedia Appendix 1
Maternal alanine aminotransferase	43.34 U/L	Negative	Figure S1K in Multimedia Appendix 1
Maternal prothrombin ratio	76.34%	Positive	Figure S1L in Multimedia Appendix 1

[a]Predicted values linked with a 50:50 probability that hepatitis B surface antigen is detectable in cord blood.

We study the Pearson correlation between each factor in the matrix model. The *R* score with a significant *P* value will be considered for further steps (see Figure 3 and Multimedia Appendices 2-5). When maternal viral load exceeds 5×10^7 copies/mL, the risk of being HBsAg positive in cord blood increases by 123% (risk ratio 2.23, 95% CI 1.48-3.36); when the viral load is lower than this baseline, the risk decreases by 55% (risk ratio 0.45, 95% CI 0.30-0.67; $P<.001$; see Table 6 and Multimedia Appendices 6 and 7). We calculate the risk ratio and odds ratio based on the new value indications (see Table 6) for 2 groups: HBsAg cord blood positive and negative, using the results of the HCA analysis. A dendrogram and principal component analysis plot were constructed based on the correlation between each factor (see Multimedia Appendices 8-10).

XSL•FO
RenderX

Figure 3. Correlation matrix between biomarkers depicted as a heat map. The heat map compares 2 groups: HBV DNA ≥ 5 × 10 copies/mL (left panel) and HBV DNA < 5 × 10 copies/mL (right panel). It illustrates the color-coded Pearson correlation coefficients between subclinical indices, including prothrombin time, AST, ALT, RBC, and haemoglobin in maternal blood; the concentration and density of PBMCs; and the status of HBeAg and anti-HBs in both cord and maternal blood. Cell colors represent the strength and direction of correlations, ranging from red (negative correlation) to blue (positive correlation). The intensity of the color reflects the magnitude of the correlation, as indicated by the color scale shown to the right of the panel. Pairwise Pearson correlation coefficients are detailed in Multimedia Appendices 2-5. ALT: alanine aminotransferase; AST: aspartate aminotransferase; CB: cord blood; CBMC: cord blood mononuclear cell; Hb: hemoglobin; HBeAg: hepatitis B e antigen; HBsAg: hepatitis B surface antigen; RBC: red blood cell.

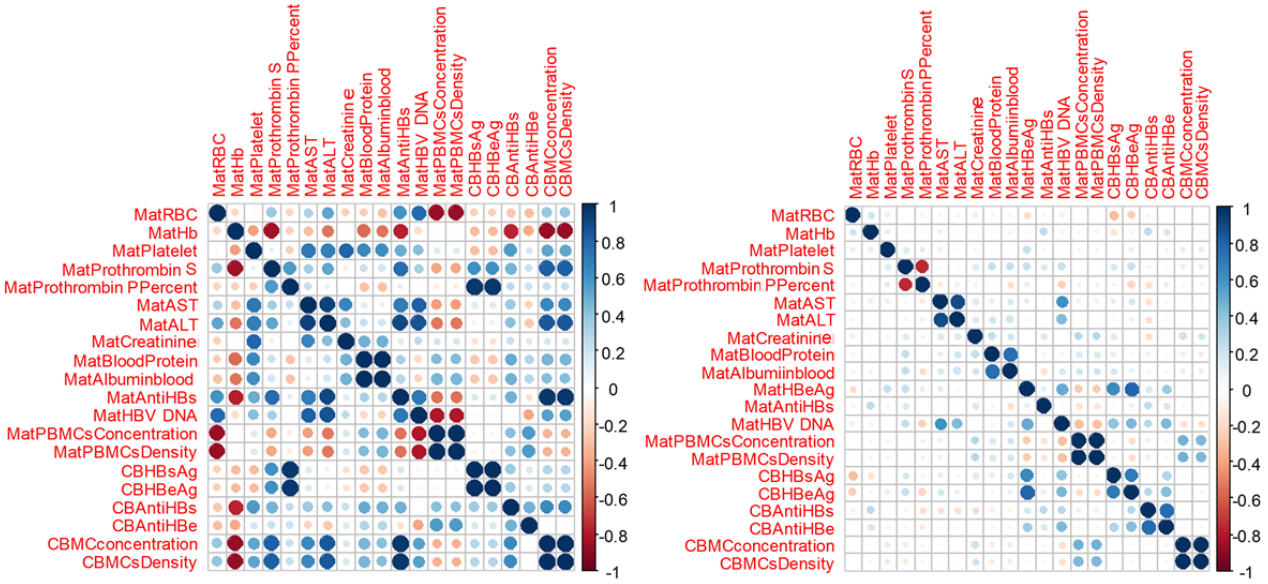

XSL•FO
RenderX

Table 6. Risk ratios for hepatitis B surface antigen–positive cord blood by factor, with corresponding threshold values.

Variables	Cord blood hepatitis B surface antigen positive (n=32), n	Cord blood hepatitis B surface antigen negative (n=28), n	Risk ratio (95% CI)	OR (95% CI)	Chi-square (*df*)	*P* value>chisq
Maternal peripheral blood mononuclear cell concentration (cells/mL)					0.34 (1)	.56
≥8.03 × 10^6	6	7	0.83 (0.44-1.58)	0.69 (0.20-2.38)		
<8.06 × 10^6	26	21	1.2 (0.63-2.28)	1.44 (0.42-4.96)		
Maternal hepatitis B virus DNA (copies/mL)					11.61 (1)	<.001
≥5 × 10^7	15	2	2.23 (1.48-3.36)	11.47 (2.32-56.65)		
<5 × 10^7	17	26	0.45 (0.30-0.67)	0.09 (0.02-0.43)		
Maternal platelets (×10^3cells/mL)					0 (1)	>.99
≥317.89	2	2	0.93 (0.34-2.56)	0.87 (0.11-6.59)		
<317.89	30	26	1.07 (0.39-2.94)	1.15 (0.15-8.78)		
Maternal prothrombin (seconds)					0 (1)	>.99
≥11	29	25	1.07 (0.46-2.48)	1.15 (0.21-6.27)		
<11	3	3	0.93 (0.40-2.15)	0.86 (0.16-4.66)		
Cord blood mononuclear cell concentration (cells/mL)					0.73 (1)	.39
≥6.64 × 10^6	22	22	0.8 (0.49-1.29)	0.6 (0.19-1.94)		
<6.64 × 10^6	10	6	1.25 (0.77-2.02)	1.67 (0.52-5.38)		
Maternal hemoglobin (g/L)					0.46 (1)	.52
≥128.53	11	12	0.84 (0.51-1.40)	0.7 (0.25-1.99)		
<128.53	21	16	1.19 (0.71-1.98)	1.43 (0.50-4.07)		
Maternal red blood cells (cells/mL)					0.024 (1)	.88
≥5 × 10^6	2	3	0.73 (0.24-2.20)	0.56 (0.09-3.59)		
<5 × 10^6	30	25	1.36 (0.45-4.10)	1.8 (0.28-11.64)		
Maternal creatinine (μmol/L)					0.047 (1)	.83
≥37.46	28	23	1.24 (0.57-2.67)	1.52 (0.37-6.33)		
<37.46	4	5	0.81 (0.37-1.75)	0.66 (0.16-2.73)		
Maternal aspartate transaminase (U/L)					0.34 (1)	.56
≥14.15	29	23	1.49 (0.59-3.76)	2.1 (0.45-9.73)		
<14.15	3	5	0.67 (0.27-1.70)	0.48 (0.10-2.20)		
Maternal alanine transaminase (U/L)					0.61 (1)	.44
≥43.34	4	1	1.57 (0.94-2.62)	3.86 (0.40-36.75)		
<43.34	28	27	0.64 (0.38-1.06)	0.26 (0.03-2.47)		
Maternal prothrombin ratio (%)					0 (1)	>.99
≥76.34	31	27	1.07 (0.26-4.36)	1.15 (0.07-19.25)		
<76.34	1	1	0.94 (0.23-3.82)	0.87 (0.05-14.60)		

PBMCs Gain the Most Information Following ID3 Theory Calculation

We next calculated the information gain for each factor in our actual data based on the correlation and clustering study results, following the ID3 theory. There are 5 attributes of pregnant women that may increase the risk of infection in infants. Each factor has 2 types of variants: HBeAg (positive and negative); ALT (<43.34 U/L and ≥43.34 U/L); AST (≥14.15 U/L and

XSL•FO
RenderX

<14.15 U/L); HBV DNA (≥5 × 10^7 and <5 × 10^7 copies/mL); and PBMCs (≥8 × 10^6 and <8 × 10^6 cells/mL).

The analysis of Cohen *h* across the groups reveals varying degrees of effect size between the proportions of cord blood HBsAg-positive and -negative cases. Group 14 demonstrates a very large effect (*h*=1.46), characterized by a notably low proportion of positive cases (3/18, 17%) and a high proportion of negative cases (15/18, 83%). Group 12 shows a medium effect (*h*=0.51), indicating a moderate difference between the 2 proportions. Groups 10 (*h*=1.70) and 7 (*h*=1.85) both reflect very large effects, with high positive rates of 7 out of 8 (88%) and 9 out of 10 (90%), respectively, contrasted with much lower negative rates. Group 13 exhibits the maximum possible difference (*h*=3.14), with a proportion of 3 out of 3 (100%) positives and 0 out of 3 (0%) negatives, corresponding to the theoretical limit of Cohen *h*, π. By contrast, other_group shows a small effect (*h*=0.15), with nearly equal positive and negative proportions, suggesting minimal difference between the 2 groups (Multimedia Appendix 11). Using the pooled SD method, we calculated Cohen *d* to measure the standardized difference between groups 10 and 7 (Multimedia Appendix 11). For instance, if our calculated Cohen *d* is approximately 0.8 or higher, it suggests a large effect size, indicating that the difference in risk or related measures between these 2 groups is not only statistically significant but also clinically meaningful (*P*=.01 for group 10 and *P*=.001 for group 7).

In the 17 output values in Tables 1-4 (N=17 scores), there are 12 values showing effect increases: 1 value is small, 2 values are large (*P*=.01 for group 10 and *P*=.001 for group 7; also see Multimedia Appendix 11), and 2 are trivial. The probability that each data point falls into class c=*medium* is determined by *Nc/N*=12/17. The probability that each data point falls into class c=*small* is determined by *Nc/N*=1/17. The probability that each data point falls into class c=*large* is determined by *Nc/N*=2/17, and the probability that each data point falls into class c=*trivial* is determined by *Nc/N*=2/17. Therefore, the entropy at *the root node* is calculated according to equations 1 and 2 of this formula, with 2 classes "no" and "yes" (C=4), which is given as follows:

If one of the attributes—HBeAg, ALT, AST, HBV DNA, or PBMCs—is selected to divide the data, we calculate the weighted sum of the entropy of the child nodes. The result is shown in Table 5. PBMC concentrations were chosen because they have the highest information gain of 0.247. We could construct a decision tree by selecting PBMCs as the root node, as they provide the greatest information gain. From the root node (PBMCs), we branch out to other nodes, each named after the corresponding attribute (Multimedia Appendix 12).

The Five Most Important Factors of MTCT Risk: HBeAg, ALT, AST, Serum HBV DNA, and PBMCs

We verified our calculation in R with 1000 runs (observations) for 5 split ratios: 0.50, 0.75, 0.80, 0.85, 0.90, and 0.95. The value in each node represents the number of observations in the data set that fall into that particular node. In Figure 4, we observed that the strongest information gain scores are found in HBeAg and PBMC concentration. From this dot plot, we can observe multiple runs of the ID3-based decision tree analysis at various training-test split ratios (shown in each column). In each run, the algorithm identifies the factor (AST, HBeAg, HBV DNA, or PBMCs) that provides the highest information gain for classification. The y-axis represents the extent of predictive "value" that each factor contributes. The results indicate that HBeAg (green dots) and PBMCs (purple dots) frequently exhibit high information gain, suggesting that these 2 factors play a prominent role in most runs, regardless of the split ratio. Meanwhile, AST (red dots) occasionally shows considerable information gain but appears less frequently, and HBV DNA (blue dot) rarely emerges as the most important factor. Overall, these findings highlight that HBeAg and PBMCs are the dominant factors in the model, maintaining stability across different training-test splits (see Figure 4).

Comparing the medians of the score values for HBeAg and PBMCs, we see that they are higher than those of the other factors. These dot plots show the information gain values for 4 factors (AST, HBeAg, HBV DNA, and PBMCs) across various runs of the model at different data split ratios (0.5, 0.75, 0.8, 0.85, 0.9, and 0.95). Each panel represents a specific split ratio, with the y-axis displaying the distribution of information gain scores for the factor deemed most important in that particular run. In general, HBeAg (green box) and PBMCs (purple box) consistently show higher median values or broader ranges of information gain, indicating that these 2 factors are crucial in the classification model, regardless of the training-test split used. In some panels (for instance, at split ratios of 0.5 or 0.75), AST (red box) occasionally records a high information gain score, but it does not match the stability or frequency of HBeAg or PBMCs. By contrast, HBV DNA (blue) seldom ranks with the highest information gain score (and hence not visible), suggesting it plays a lesser role in the classification decisions during these ID3 runs. Therefore, HBeAg and PBMCs are the primary contributors to the model's predictive performance, while AST shows occasional dominance, and HBV DNA rarely serves as the top predictor (see Figure 5).

XSL•FO
RenderX

Figure 4. The dot plot of the most important factor in each run following different split ratios. AST: aspartate aminotransferase; HBeAg: hepatitis B e antigen; HBV: hepatitis B virus; PBMC: peripheral blood mononuclear cell.

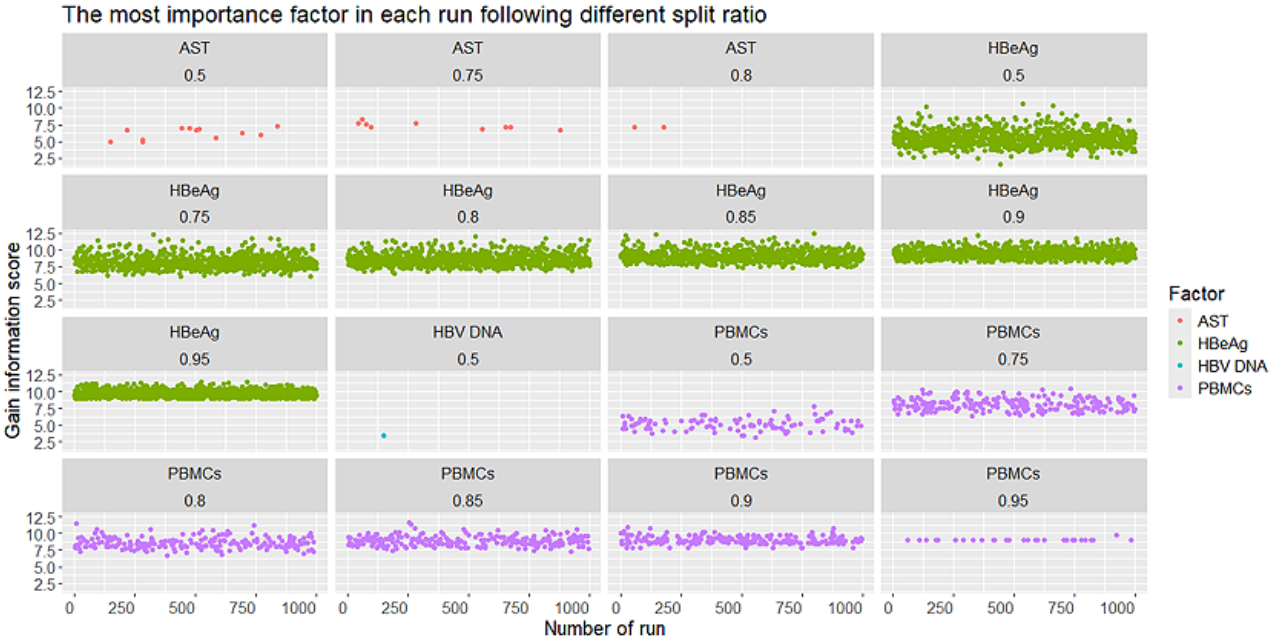

Figure 5. The bar plot of the most important factor in each run following different split ratios. AST: aspartate aminotransferase; HBeAg: hepatitis B e antigen; HBV: hepatitis B virus; PBMC: peripheral blood mononuclear cell.

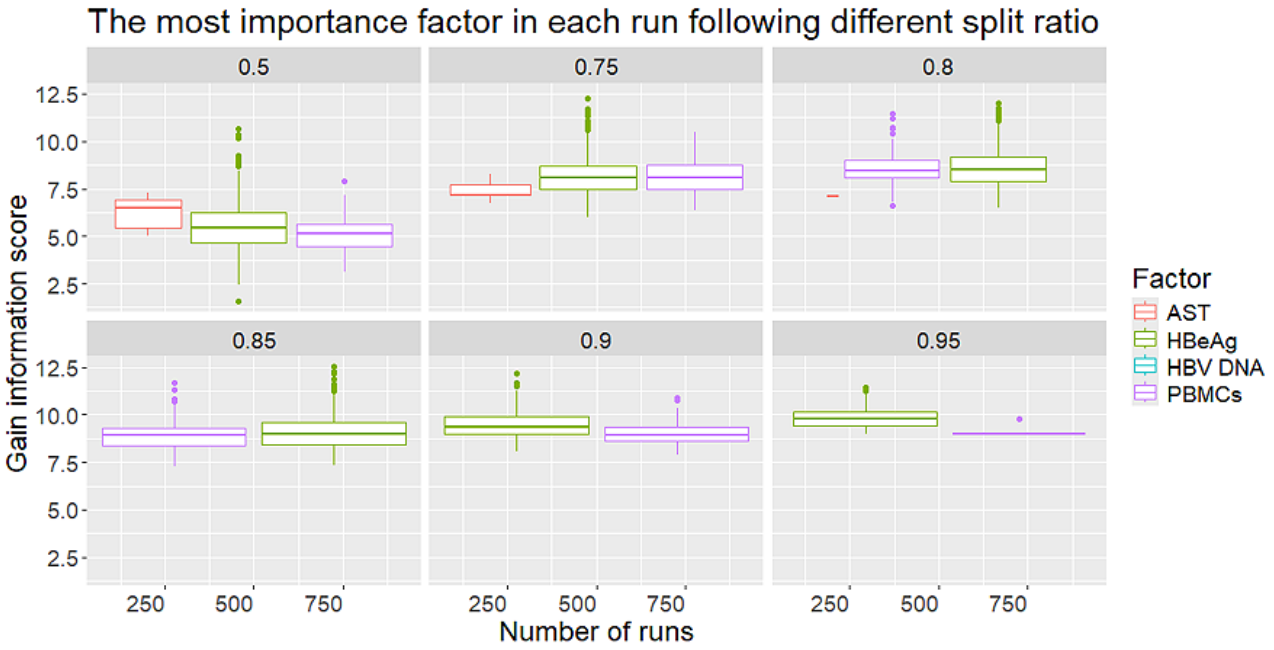

In the decision tree representation, each node's split is based on the predictor variables. The result provides several pieces of information that help us interpret the tree. Figure 5 illustrates how the decision trees (ID3/CART) divide the data based on various training-test split ratios (0.95, 0.90, 0.85, 0.80, 0.75, and 0.50), along with the associated MTCT risk classifications (trivial, small, medium, and large) and gain score results. In all splits, HBeAg consistently emerges as the primary splitting factor. The HBeAg-positive branch generally leads to the large risk category when PBMCs are ≥8 × 10^6 cells/mL, and to medium or small risk when PBMCs are lower. By contrast, HBeAg-negative cases typically split into small risk (when PBMCs are high) or trivial risk (when PBMCs are low). The gain score tables further validate that HBeAg and PBMCs are dominant. Although the precise distributions of risk categories (trivial, small, medium, and large) vary slightly with different splits, the model consistently highlights HBeAg and PBMCs as key factors, reinforcing earlier findings that these 2 variables are essential predictors of mother-to-child HBV transmission risk (Figure 5). According to Figure 6, the risk of MTCT of HBV is stratified based on HBeAg status and PBMC concentration as follows: among HBeAg-positive women, 20 out of 48 training cases with a split ratio of 0.80 (42%) to 27 out of 57 training cases with a split ratio of 0.95 (47%), or 16 out of 30 training cases with a split ratio of 0.50 (53%), were

XSL•FO
RenderX

classified as having a high risk of MTCT. Among HBeAg-negative individuals with PBMC concentrations ≥8 × 10^6 cells/mL, 7 out of 51 training cases with a split ratio of 0.85 (14%) to 8 out of 45 training cases with a split ratio of 0.75 (18%) were categorized as having a small risk. The remaining 21 out of 57 training cases with a split ratio of 0.95 (37%), or 20 out of 53 training cases with a split ratio of 0.90 (38%) to 16 out of 30 training cases with a split ratio of 0.50 (53%) of HBeAg-negative cases with PBMC concentrations <8 × 10^6 cells/mL were classified as having negligible risk (see Figure 6).

From the dot plot (Figure 4), we can observe several runs of the ID3-based decision tree analysis at different training-test split ratios, indicated in each column. In each run, the algorithm identifies the factor (AST, HBeAg, HBV DNA, or PBMCs) that provides the highest information gain for classification. The y-axis shows the predictive "value" contributed by each factor. The results reveal that HBeAg (green) and PBMCs (purple) frequently demonstrate high information gain, indicating that these 2 factors are dominant in most runs, regardless of the split ratio. By contrast, AST (red) occasionally shows a notable information gain but is less frequent, while HBV DNA (blue) rarely appears as the most critical factor. Overall, these results emphasize that HBeAg and PBMCs are the key factors in the model, remaining consistent across various training-test splits.

In summary, these plots show that HBeAg and PBMCs are typically the most influential factors across different training-test splits, while AST plays a more inconsistent role, and HBV DNA contributes less frequently to classification decisions in these ID3 runs.

Figure 6. Decision tree diagram based on different split ratios: (A) 0.95, (B) 0.90, (C) 0.85, (D) 0.80, (E) 0.75, and (F) 0.50. Cohen classified effect sizes on MTCT risk as trivial (d<0.2), small (0.2≤d<0.5), medium (0.5≤d<0.8), and large (d≥0.8). ALT: alanine aminotransferase; AST: aspartate aminotransferase; HBeAg: hepatitis B e antigen; HBV: hepatitis B virus; MTCT: mother-to-child transmission; PBMC: peripheral blood mononuclear cell.

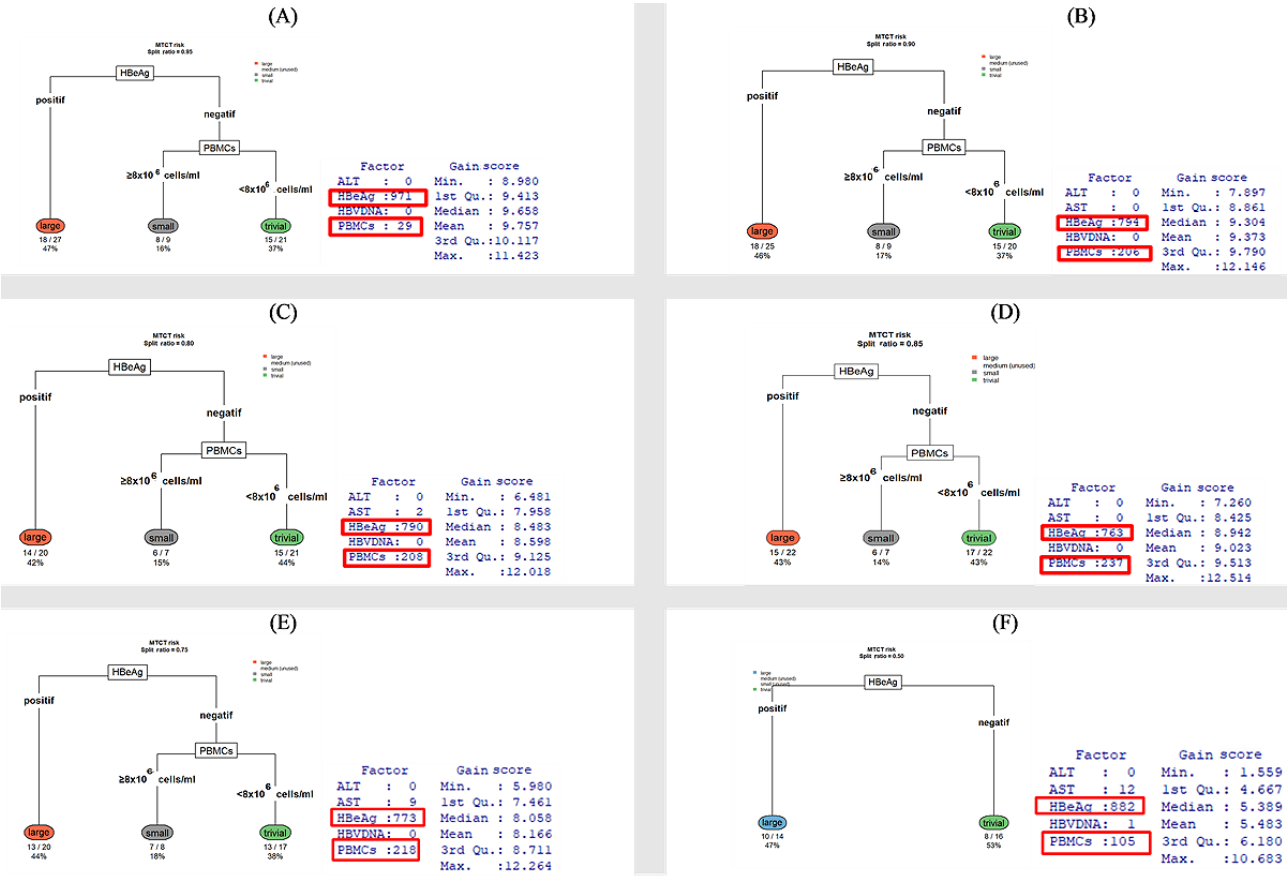

Boosting the Assembly of the Five Factors Shows the Important Groups to Predict the Risk of MTCT

We have 14 cases in which we could enhance the assembly simulation (Multimedia Appendix 11). The split ratio and number of runs influence our predictions for all cases in the group. Figure 6 shows the raw distribution, reflecting the accuracy of each observation in relation to the split ratio. The results of the Pearson chi-square test for the contingency table show that the distribution of risk groups (large, small, and trivial) across different training data split ratios (ranging from 0.75 to 0.95) differs significantly (χ^2_8=21.16, *P*=.007). When testing each group individually using a 1-way chi-square test, the large group has a *P* value of .002, indicating a clear change in distribution across different training ratios. The small group has a *P* value of .07, which is close to the significance threshold, suggesting a potential trend in distribution change, while the trivial group has a *P* value of .84, indicating a stable distribution that is not significantly affected by the data split ratio. Thus, changes in the train/test ratio may influence how decision trees

XSL•FO
RenderX

learn, particularly for groups with distinct characteristics, such as the "large" group. An accuracy measure for classification tasks, using the confusion matrix, provides a better evaluation of classification performance. The general idea is to count how often true instances (true positive and true negative) are misclassified as false (false positive and false negative). We compute the accuracy of the test from the confusion matrix using the following formula: accuracy = (true positive + true negative)/(true positive + true negative + false positive + false negative). The accuracy score reflects the probability that the test data produces the same result as the training data set. Figure 7 shows the results of the correlation test, confirming that the number of observations and the split ratio strongly correlate with accuracy. We repeat this random selection 1000 times for each division ratio, aiming to achieve the highest accuracy. Figure 8 provides an overview of the MTCT risk. Groups 7 and 10 have the highest MTCT risk, with a prevalence ranging from 13 out of 48 (a split ratio of 0.8 from a total of 60 cases, 27%) to 18 out of 54 (a split ratio of 0.9 from a total of 60 cases, 33%). Groups 13 and 14 have the lowest risk, with 9 out of 30 (a split ratio of 0.5 from a total of 60 cases, 30%) to 18 out of 48 (a split ratio of 0.8 from a total of 60 cases, 38%) cases falling into the trivial group, indicating negligible MTCT risk. The other groups show an accuracy ranging from 8 out of 57 (a split ratio of 0.95 from a total of 60 cases, 14%) to 9 out of 30, a split ratio of 0.5 from a total of 60 cases (30%), with cases categorized into the medium- or small-risk groups.

Figure 7. Distribution of accuracy score. (A) Raw distribution. (B) Violin plot shows the median of accuracy.

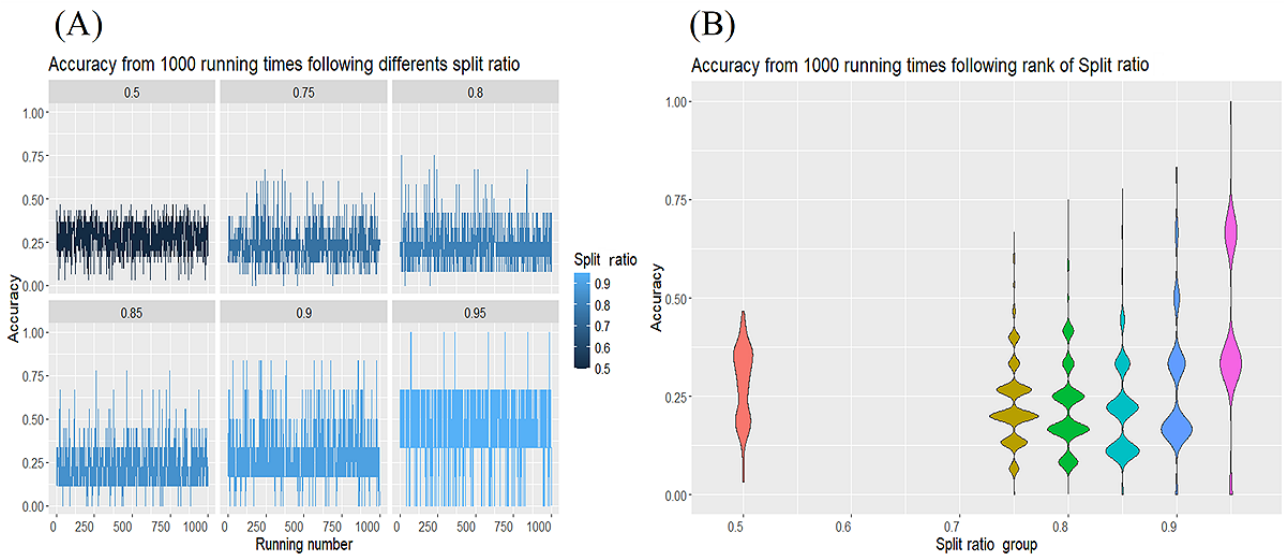

From these plots, we can observe the classification accuracy achieved by the ID3 model over 1000 runs at various training-test split ratios (0.5, 0.75, 0.8, 0.85, 0.9, and 0.95). Each panel corresponds to a specific split ratio, with the y-axis displaying accuracy scores and the x-axis representing the number of runs. Although there is considerable fluctuation within each split ratio, some general patterns emerge. At lower split ratios (eg, 0.5), the model's accuracy tends to cluster in the lower to mid range (approximately 0.25-0.50). As the split ratio increases (eg, 0.8 or 0.85), accuracy occasionally reaches higher peaks, with some runs exceeding 0.60 or 0.70, though variability remains evident. With even larger training proportions (0.9 and 0.95), the accuracy range broadens further, with some runs achieving relatively high performance while others drop close to 0. Overall, these results indicate that the model's accuracy is highly sensitive to the specific partitioning of the data set, showing moderate gains and substantial variation as the training-test split ratio changes (Figure 7A).

The violin plots illustrate the distribution of classification accuracy across 1000 runs for each training-test split ratio. On the left, with a split ratio of 0.5, the distribution is relatively narrow, centering around the 0.25-0.40 range. This suggests that using half of the data for training typically results in modest accuracy. As the training proportion increases to 0.75 or 0.8, the distribution shifts upward, revealing higher accuracy values. At split ratios of 0.9 and 0.95, the distribution widens considerably, with some runs achieving very high accuracy—approaching or exceeding 0.75—while others fall closer to 0.20. This suggests that while more training data can improve the model's performance in some cases, a smaller test set may lead to greater variance, resulting in a wider range of accuracy outcomes (Figure 7B).

Figure 8A shows a strong positive correlation (R=0.91, P<2.2 × 10^{-16}) between the training-test split ratio (on the x-axis) and classification accuracy (on the y-axis). As the amount of data allocated for training increases, the model's accuracy tends to improve, resulting in a nearly linear upward trend. In Figure 8B, the correlation between the number of runs (indicated on the x-axis) and accuracy is considerably weaker (R=0.17, P<2.2 × 10^{-16}). This suggests that while accuracy shows a slight upward trend over multiple runs, the majority of the variation is better explained by other factors, particularly the split ratio, rather than the order or total number of runs.

XSL•FO
RenderX

Figure 8. Correlation line between accuracy and split ratio or running number (number of observations). (A) Correlation plot between accuracy and split ratio. (B) Correlation plot between accuracy and running number.

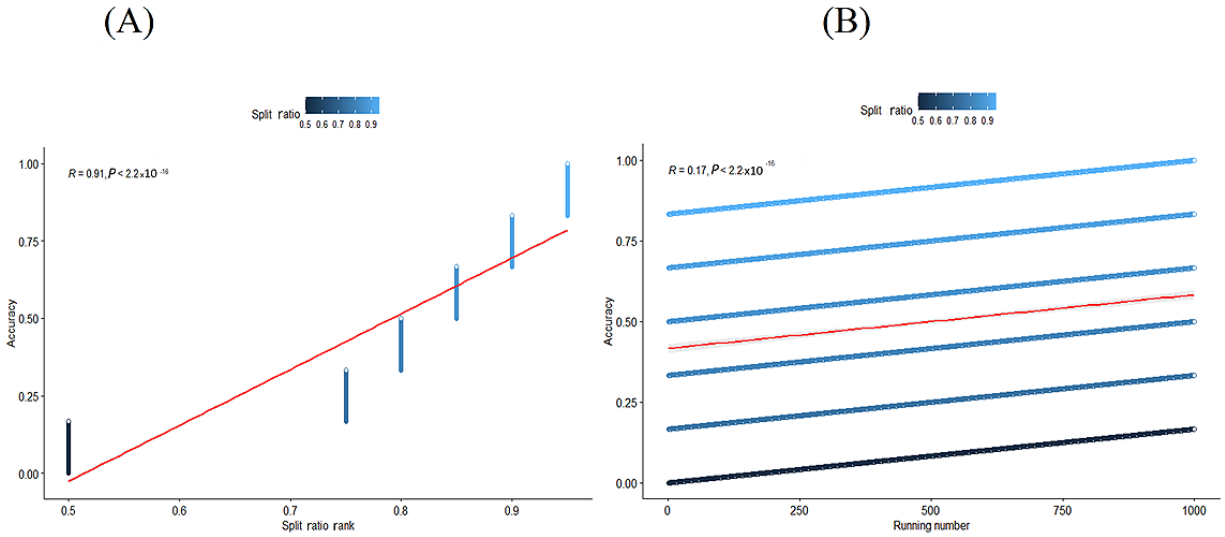

In these CART diagrams (Figure 9 and Multimedia Appendix 13), we observe how the data set is consistently divided by key factors such as PBMCs, AST, ALT, HBeAg, and HBV DNA across various training-test split ratios, ranging from 0.50 to 0.95. Each final branch represents a category of MTCT risk—classified as "large," "medium," "small," or "trivial"—according to the Cohen *d* index. Interestingly, some groups—such as group 7 (with PBMCs < 8×10^6 cells/mL, AST ≥ 14.15 U/L, ALT < 43.34 U/L, HBeAg positive, HBV DNA ≥ 5×10^7 copies/mL) and group 10 (with PBMCs < 8×10^6 cells/mL, AST ≥ 14.15 U/L, ALT < 43.34 U/L, HBeAg positive, HBV DNA < 5×10^7 copies/mL)—are frequently associated with "very large" effect sizes. By contrast, others—such as group 13 (with PBMCs < 8×10^6 cells/mL, AST ≥ 14.15 U/L, ALT ≥ 43.34 U/L, HBeAg positive, HBV DNA ≥ 5×10^7 copies/mL) and group 14 (with PBMCs < 8×10^6 cells/mL, AST ≥ 14.15 U/L, ALT < 43.34 U/L, HBeAg negative, HBV DNA < 5×10^7 copies/mL)—are categorized as "trivial" in terms of increasing the risk of MTCT. The proportion and number of runs leading to each category are displayed beneath each node. As the training ratio increases, the model becomes more effective at assigning specific groups to their corresponding risk categories, although the overall distribution of runs across large, medium, and trivial categories still fluctuates. This suggests that while certain subgroup characteristics, such as PBMC levels or HBeAg status, influence classification toward high or negligible MTCT risk, the model's consistency and accuracy are also impacted by how the data are divided for training and testing.

In this study, a "large effect" is defined as a Cohen *d* value of 0.8 or higher. Our decision tree analysis (utilizing ID3 and CART) reveals that the groups identified as 7 and 8 typically show a large effect size, indicating a high risk of MTCT. In these groups, factors such as positive HBeAg status, elevated PBMC concentration (≥8×10^6 cells/mL), and other relevant biochemical markers contribute to an effect size (Cohen *d*) that meets or exceeds the 0.8 threshold. This indicates an increased risk of transmission in these groups.

XSL•FO
RenderX

Figure 9. Classification and regression trees with different split ratios. (A) Split ratio 0.95. (B) Split ratio 0.90. Group numbers correspond to those listed in Table 3. The range of mother-to-child transmission (MTCT) risk is based on the Cohen d index. According to Cohen classification, effect sizes on MTCT risk are defined as follows: trivial (d<0.2), small (0.2≤d<0.5), medium (0.5≤d<0.8), and large (d≥0.8). ALT: alanine aminotransferase; AST: aspartate aminotransferase; HBeAg: hepatitis B e antigen; HBV: hepatitis B virus; PBMC: peripheral blood mononuclear cell.

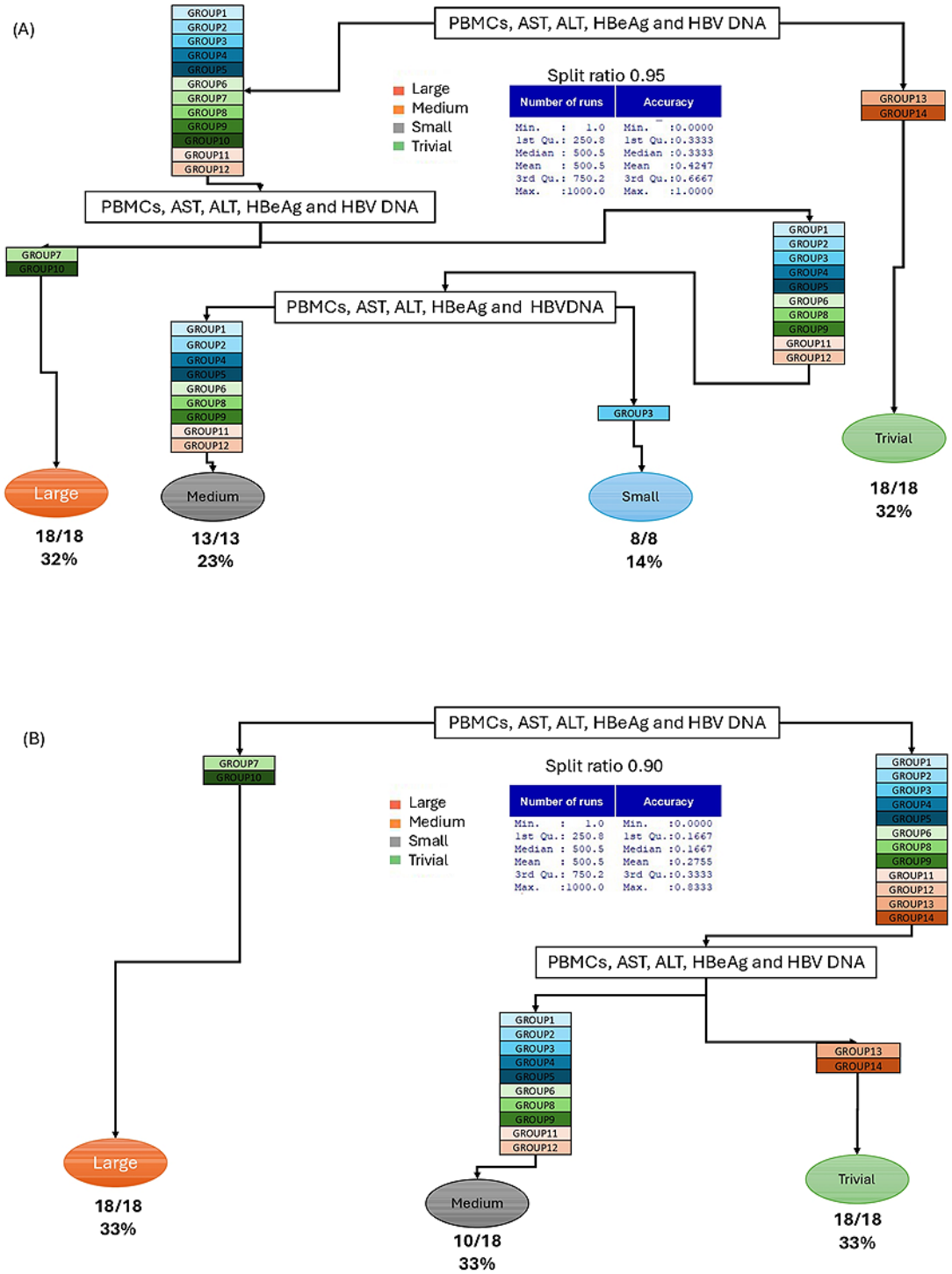

Discussion

Principal Findings

This study aimed to identify key predictors of MTCT of HBV, with a particular focus on maternal HBeAg status and PBMC concentration. We developed and applied an ID3-based decision tree model to analyze clinical data from pregnant Vietnamese women. Our primary finding is that HBeAg positivity, combined with elevated PBMC levels (≥8 × 10^6 cells/mL), is strongly associated with an increased risk of MTCT. The decision tree model effectively stratified risk based on a combination of virological and biochemical indicators, confirming our hypothesis regarding the predictive value of both PBMCs and HBeAg.

HBV is known to cross the placental barrier, and the presence of maternal HBeAg in newborns at birth supports the occurrence of vertical transmission. Our findings add to the growing body of evidence suggesting that PBMCs may play a critical role in

XSL•FO
RenderX

facilitating intrauterine HBV infection [15]. Prior research has shown a 9.5-fold increased risk of HBV infection among neonates who are HBV DNA positive in PBMCs [16]. Similarly, our earlier study involving a cohort of 60 Vietnamese mothers and their infants found a correlation between PBMC concentrations and maternal viral load, particularly in cases with HBV DNA levels below 5×10^7 copies/mL. HBV can be transmitted from mother to child due to its ability to cross the placental barrier. Notably, HBeAg detected in newborns at birth is of maternal origin [17]. In this study, we aim to utilize real-world clinical data to develop a predictive model to determine the risk of MTCT. Our approach uses a machine learning model based on the ID3 decision tree algorithm. The critical decision nodes are determined through a series of calculations, starting from individual (single) entropy to total entropy. Information gain, which quantifies the reduction in entropy, is used to evaluate how effectively a given feature separates or classifies the target outcomes. The feature with the highest information gain is selected as the most informative. This classical entropy-based measure plays a fundamental role in various machine learning algorithms, including decision tree models. In the ID3 algorithm, the attribute selected as the splitting criterion is the one with the highest information gain. This method has proven effective in various machine learning and signal processing applications. Beyond facilitating feature selection, entropy-based measures serve as reliable indicators of data complexity and classification difficulty in real-world scenarios, as demonstrated by Juszczuk et al [18]. Their research underscores the critical importance of using entropy in model construction, especially when dealing with complex biomedical data sets where noise and variability can impact model performance. This supports our application of information gain in identifying PBMC concentration as the most informative variable, aligning both with clinical relevance and theoretical expectations in predicting the risk of MTCT [18]. In machine learning theory, information gain is synonymous with Kullback-Leibler divergence; this index quantifies the amount of information obtained about a random variable by observing another. In the first step, the ID3 algorithm was applied to our data subset without splitting. PBMCs achieved the highest information gain score, which was 0.247 (Multimedia Appendix 13). We could draw a decision tree by choosing PBMCs as the root node, as it has the most significant information gain. From the root node PBMCs, branches extend to other attribute nodes listed in Multimedia Appendix 13. We found that groups 7 and 10 (Multimedia Appendix 11) are associated with a high risk of MTCT, characterized by positive HBeAg status, PBMC serum concentrations above 8×10^6 cells/mL, elevated AST levels (≥14.15 U/L), and low ALT levels (<43.34 U/L). These groups have Cohen *d* values of 1.855 and 1.696, respectively (Multimedia Appendix 11), with associated probabilities ranging from 13 out of 48 (a split ratio of 0.8 from a total of 60 cases, 27%) to 18 out of 54 (a split ratio of 0.9 from a total of 60 cases, 33%), respectively (Figure 9 and Multimedia Appendix 12). The dual roles of HBeAg as both a tolerogen and an immunogen, along with its ability to either suppress or activate the immune response, highlight the complexity of its interactions with the host. Numerous studies have demonstrated that HBeAg can influence both innate and adaptive immune responses, contributing to the persistence of HBV. HBeAg can bind to PBMCs, neutrophils, and B lymphocytes, but not to T lymphocytes. The interaction between HBeAg and monocytes or neutrophils has been shown to be dose dependent, resulting in the inhibition of both cell types. Monocytic myeloid-derived suppressor cells (mMDSCs) are derived from myeloid progenitor cells and account for approximately 0.5% of PBMCs in healthy individuals. The mMDSC population expands during infection, inflammation, and cancer. HBeAg plays a crucial role in the expansion of the mMDSC population and the induction of immune tolerance. Compared with HBeAg-negative patients, HBeAg-positive patients were shown to have significantly higher levels of mMDSCs. When PBMCs from healthy individuals were exposed to HBeAg, there was an increase in mMDSCs and the expression of IL-6 and IL-1β. Additionally, mMDSCs from HBeAg-positive patients suppressed the proliferation of CD4+ and CD8+ T cells. This may represent a potential mechanism by which HBeAg modulates the host immune response during CHB by physically depleting or weakening virus-specific CD4+ and CD8+ T cells. As a result, these cells are unable to proliferate in response to viral antigens or produce essential antiviral and immunostimulatory cytokines, which are crucial for controlling the virus in patients with CHB [14]. Additionally, new research by Padarath et al [14] provided further insights into the various functions of HBeAg and its precursors in the development of chronic HBV infection. According to their review, HBeAg may influence hepatocarcinogenesis through long-term immune modulation and chronic inflammation, in addition to promoting immune evasion and tolerance. This demonstrates that HBeAg serves various functions, including promoting vertical transmission, inhibiting host defenses, and potentially contributing to chronic conditions such as liver cancer. These findings confirm our results, which indicate that HBeAg plays a major role in MTCT risk, and emphasize the need for close monitoring of pregnancies involving HBeAg-positive women. It is well known that both HBeAg and AST are independent risk factors for predicting nonminimal liver inflammation in untreated patients with CHB. In HBeAg-positive untreated patients with CHB, the liver inflammation associated with CHB is linked to the balance between the immune system and HBV infection. Quantitative changes in indicators such as HBsAg and HBeAg can signal the breakdown of immune tolerance and the onset of immune clearance in CHB infection. Including immune-related indicators in the inflammatory prediction model for CHB infection is essential. In general, ALT or AST is included in most models for liver inflammation or fibrosis. It has been found that AST is a better predictor than ALT. HBV DNA, HBsAg, and HBeAg reflect the replication capacity of HBV in untreated patients with CHB. Previous studies have shown that HBsAg and HBeAg are negatively correlated with liver inflammation. In this study, HBsAg, HBeAg, and HBV DNA were all included, with HBeAg demonstrating the best predictive ability. Compared with other models, the significance of this nonminimal liver inflammation model lies in its confirmation of the importance of HBeAg in identifying liver inflammation. On the one hand, HBeAg reflects the replication level of HBV in HBeAg-positive patients with CHB; on the other hand, a decline in HBeAg is often an early sign of the breakdown of immune tolerance [19]. Based on the

XSL•FO
RenderX

above calculations, it is evident that further experiments with various random split ratios of the data file are necessary. Different tests will improve prediction performance. The positive correlation between accuracy, division rate, and the number of runs supports this. A clear result is that, after 1000 random runs for each division type, HBeAg and PBMC concentrations consistently show the highest information gain scores. When testing all 5 factors simultaneously, fixed groups emerged in each risk category, highlighting the additive or inhibitory influence of other biochemical indicators on the impact of HBeAg and PBMCs on the risk of MTCT.

Nonetheless, this study has several limitations that should be noted. First, the sample size was relatively small and geographically limited to a single region in Vietnam, which may affect the generalizability of the findings to other populations or ethnicities. Second, while the ID3 decision tree provided valuable insights into variable importance and risk classification, predictive performance could potentially be improved with larger, more heterogeneous data sets, and by comparing it with other, yet untested, machine learning approaches (eg, random forest, XGBoost). Third, additional maternal or fetal factors (eg, nutritional status, coinfections, genetic predisposition) may have been overlooked and could be important, given the constraints of the available data. Finally, the cross-sectional design and retrospective data collection may have introduced bias, limiting the ability to establish a definitive causal relationship. Nevertheless, the findings of the study are still valuable and provide a strong foundation for future research. By identifying key predictors of risk, particularly HBeAg status and PBMC concentration, it was possible to develop a simple yet practical model for predicting risk. These findings represent an essential first step toward developing more comprehensive tools, which can later be refined and validated through larger prospective studies aimed at enhancing clinical decision-making in HBV-infected pregnancies.

Our findings underscore the predictive importance of maternal HBeAg status and PBMC concentration in assessing the risk of HBV MTCT. Incorporating these indicators into clinical screening procedures could enable more targeted interventions during pregnancy. Additionally, our results support the use of machine learning tools in epidemiological risk assessment, particularly in resource-limited settings. Future studies should aim to validate these findings in larger, more diverse cohorts and explore the potential pathways through which PBMCs mediate HBV transmission. These insights may influence decisions regarding the timing and choice of antiviral treatments in high-risk pregnancies, as well as the design of other preventive strategies.

Limitation

Decision trees are robust and flexible machine learning algorithms. They are interpretable, capable of handling nonlinear relationships, and efficient in computation. Additionally, decision trees can accommodate mixed data types. However, if a small sample is tested, there is a risk of overfitting or overclassification. It is also important to note that only 1 attribute is tested at a time when making a decision.

Conclusions

This study demonstrated that by combining the ID3 and CART algorithms, data can be interpreted as a decision tree to assist clinicians in their understanding. Additionally, the proposed system provides improved performance by category. The resulting prediction rules, derived from the training data, construct the fastest and most efficient tree. This approach only requires testing enough attributes to classify all the data. By identifying leaf nodes, the test data can be pruned, reducing the number of tests required. The entire data set is explored to construct the tree. This strategy provides clear, structured choices with potential outcomes, making it especially useful in complex diagnostics. It allows health care professionals to review symptoms and test results systematically, supporting more informed decision-making.

Acknowledgments

This study is part of a collaborative research effort between Germany and Vietnam. It is funded by the BMBF (German Federal Ministry of Education and Research) under reference number 01DP19001 for its participation in the project “Identification of Viral and Host Factors in Vertical Mother-to-Child Transmission of Hepatitis B Virus,” and by MOST (Ministry of Science and Technology of Vietnam) under reference number 232/QD-BKHCN for its involvement in the project “Application of DNA Sequencing and ELISPOT Technologies in Evaluation of HBV Transmission Risk from Mother to Child.” We express our gratitude to the BMBF of Germany and MOST of Vietnam for supporting bilateral scientific cooperation between the two countries. We also thank the patients, health care workers, midwives, and technicians who participated in and assisted with the completion of this research. Our heartfelt thanks go to Professor Ulrike Protzer’s research group at the Institute of Virology, School of Medicine, Technical University of Munich, Munich, Germany, and Helmholtz Zentrum München, Munich, Germany; the German Center for Infection Research (DZIF), Munich and Hamburg Partner sites, Germany, for their invaluable collaboration and support during this study. Their expertise, insightful contributions, and dedicated efforts have been crucial to the success of this project.

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Authors' Contributions

DNT contributed to the study design, study supervision, and manuscript review as the project leader. HBTT was involved in the study design, data compilation and analysis, as well as manuscript writing and review. THTN contributed to sample collection, data compilation and analysis, and manuscript writing and review. TPT participated in sample collection and data compilation. TND, HNTT, and HVTT provided technical supervision. LATL was responsible for the technical supervision of LTH. HCT contributed to the study design, proposal writing, bioinformatic analysis, and manuscript writing. NK and TB were involved in the study design, study supervision, and manuscript review. LKH contributed to bioinformatic analysis planning, analysis supervision, and manuscript review. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Relative risk ratio plot. Probability: 1; CBHBsAg positif, 0; and CBHBsAg negatif: 0. CB: cord blood; HBsAg: hepatitis B surface antigen.

[PDF File (Adobe PDF File), 422 KB - formative_v9i1e69838_app1.pdf]

Multimedia Appendix 2

Pearson correlation coefficients (R) in the HBV-DNA $\geq 5\times10^7$ copies/mL group. The color-coded Pearson correlation coefficients between subclinical indices, including levels of Prothrombin, AST, ALT, RBC, and hemoglobin in maternal blood; the concentration and density of PBMCs; and the status of HBeAg and anti-HBs in both cord and maternal blood. The color of each cell reflects the strength and direction of the correlation, ranging from red (negative correlation) to blue (positive correlation), with intensity corresponding to the magnitude of the association. The correlation strength is indicated by the accompanying color scale. ALT: alanine aminotransferase; AST: aspartate aminotransferase; HBV: hepatitis B virus; HBeAg: hepatitis B e antigen; RBC, red blood cell; PBMC: peripheral blood mononuclear cell.

[PDF File (Adobe PDF File), 490 KB - formative_v9i1e69838_app2.pdf]

Multimedia Appendix 3

Pearson correlation coefficients (R values) in the HBV DNA $< 5 \times 10^7$ copies/mL group are presented in this table. It displays color-coded Pearson correlation coefficients between various subclinical indices, including prothrombin time, AST, ALT, RBC, and hemoglobin levels in maternal blood; the concentration and density of PBMCs; and the status of HBeAg and anti-HBs in both cord and maternal blood. Cell colors indicate the strength and direction of the correlations, ranging from red (negative correlations) to blue (positive correlations), with the intensity corresponding to the magnitude, as shown by the color scale. ALT: alanine aminotransferase; AST: aspartate aminotransferase; HBeAg: hepatitis B e antigen; HBV: hepatitis B virus; RBC: red blood cell; PBMC: peripheral blood mononuclear cell.

[PDF File (Adobe PDF File), 324 KB - formative_v9i1e69838_app3.pdf]

Multimedia Appendix 4

P values from the Pearson correlation tests in the HBV DNA $\geq 5 \times 10^7$ copies/mL group. HBV: hepatitis B virus.

[PDF File (Adobe PDF File), 486 KB - formative_v9i1e69838_app4.pdf]

Multimedia Appendix 5

P values from the Pearson correlation tests in the HBV DNA $< 5 \times 10^7$ copies/mL group. HBV: hepatitis B virus.

[PDF File (Adobe PDF File), 315 KB - formative_v9i1e69838_app5.pdf]

Multimedia Appendix 6

P values from the Fisher exact test comparing the 2 groups: HBV DNA $< 5 \times 10^7$ copies/mL and HBV DNA $\geq 5 \times 10^7$ copies/mL. HBV: hepatitis B virus.

[PDF File (Adobe PDF File), 651 KB - formative_v9i1e69838_app6.pdf]

Multimedia Appendix 7

P values from the Fisher exact test and Pearson correlation test comparing the correlation coefficients (R) between the 2 groups: HBV DNA $< 5 \times 10$ copies/mL and HBV DNA $\geq 5 \times 10^7$ copies/mL. HBV: hepatitis B virus.

[PDF File (Adobe PDF File), 664 KB - formative_v9i1e69838_app7.pdf]

XSL•FO
RenderX

Multimedia Appendix 8

Cluster dendrogram of the 2 groups. Left: group with higher maternal viral load (HBV DNA ≥ 5×10^7 copies/mL); right: group with lower maternal viral load (HBV DNA < 5×10^7 copies/mL). The vertical axis represents the Euclidean distance between observations or clusters. Horizontal bars indicate the linkage points at which clusters or individual observations are merged. HBV: hepatitis B virus.

[PDF File (Adobe PDF File), 349 KB - formative_v9i1e69838_app8.pdf]

Multimedia Appendix 9

Top: Group with higher viral load (HBV DNA ≥ 5×10^7 copies/mL); bottom: group with lower viral load (HBV DNA < 5×10^7 copies/mL). K-means clustering was optimized using 3 methods: average silhouette, elbow, and gap statistic. The cluster analysis was performed without any missing values. HBV: hepatitis B virus.

[PDF File (Adobe PDF File), 290 KB - formative_v9i1e69838_app9.pdf]

Multimedia Appendix 10

Cluster plot. Left: group with higher viral load (HBV DNA ≥ 5×10^7 copies/mL); right: group with lower viral load (HBV DNA < 5×10^7 copies/mL). Clustering was performed with the number of clusters set to 5 and 8, respectively, to visualize the parameter groupings within each viral load group. HBV: hepatitis B virus.

[PDF File (Adobe PDF File), 457 KB - formative_v9i1e69838_app10.pdf]

Multimedia Appendix 11

Cohen classified effect sizes on MTCT risk as follows: trivial (d<0.2), small (0.2≤d<0.5), medium (0.5≤d<0.8), and large (d≥0.8). To determine the effect size for each factor, we transformed the natural logarithm of the odds ratio (ln[odds ratio]) by dividing it by 1.81. Additionally, Cohen h was calculated using the formula $h = 2 \times |\arcsin(\sqrt{p1}) - \arcsin(\sqrt{p2})|$, where p1 represents the proportion of CBHBsAg-positive cases and p2 the proportion of CBHBsAg-negative cases [9]. The analysis of Cohen h across the groups revealed varying degrees of effect size between CBHBsAg-positive and CBHBsAg-negative proportions. Group 14 demonstrated a very large effect (h=1.46), with a low positive rate (17%) and high negative rate (83%). Group 12 showed a medium effect (h=0.51), indicating a moderate difference. Groups 10 (h=1.70) and 7 (h=1.85) both exhibited very large effects, with positive rates of 88% and 90%, respectively, contrasted with much lower negative rates. Group 13 showed the maximum possible difference (h=3.14), with a 100% positive and 0% negative rate, corresponding to the theoretical upper limit of Cohen h, Π. By contrast, the "other_group" displayed a small effect (h=0.15), with nearly equal proportions of positive and negative cases, suggesting minimal difference between the 2. HBsAg: hepatitis B surface antigen; MTCT: mother-to-child transmission.

[DOCX File , 23 KB - formative_v9i1e69838_app11.docx]

Multimedia Appendix 12

Classification and regression trees (CART) with varying split ratios. Panels show CART results with different split ratios: (A) 0.85, (B) 0.80, (C) 0.75, and (D) 0.50. Group numbers correspond to those listed in Table 3. The range of mother-to-child transmission (MTCT) risk is based on Cohen d index. According to the Cohen classification, effect sizes for MTCT risk are defined as follows: trivial (d<0.2), small (0.2≤d<0.5), medium (0.5≤d<0.8), and large (d≥0.8).

[PDF File (Adobe PDF File), 492 KB - formative_v9i1e69838_app12.pdf]

Multimedia Appendix 13

The weighted sum of the entropy of the child nodes and the corresponding information gain reaches its maximum value. Cohen classified effect sizes related to mother-to-child transmission risk as follows: trivial (d<0.2), small (0.2≤d<0.5), medium (0.5≤d<0.8), and large (d≥0.8).

[DOCX File , 33 KB - formative_v9i1e69838_app13.docx]

References

1. Nelson NP, Jamieson DJ, Murphy TV. Prevention of perinatal hepatitis B virus transmission. J Pediatric Infect Dis Soc 2014 Sep;3 Suppl 1(Suppl 1):S7-S12 [FREE Full text] [doi: 10.1093/jpids/piu064] [Medline: 25232477]
2. Dunford L, Carr MJ, Dean J, Nguyen LT, Ta Thi TH, Nguyen BT, et al. A multicentre molecular analysis of hepatitis B and blood-borne virus coinfections in Viet Nam. PLoS One 2012 Jun 13;7(6):e39027 [FREE Full text] [doi: 10.1371/journal.pone.0039027] [Medline: 22720022]
3. Sarker IH. Machine learning: algorithms, real-world applications and research directions. SN Comput Sci 2021 Mar 22;2(3):160 [FREE Full text] [doi: 10.1007/s42979-021-00592-x] [Medline: 33778771]
4. World Health Organization (WHO). Prevention of Mother-to-Child Transmission of Hepatitis B Virus: Guidelines on Antiviral Prophylaxis in Pregnancy. Geneva, Switzerland: WHO; 2021.

XSL•FO
RenderX

5. Körber N, Pohl L, Weinberger B, Grubeck-Loebenstein B, Wawer A, Knolle PA, et al. Hepatitis B vaccine non-responders show higher frequencies of CD24CD38 regulatory B cells and lower levels of IL-10 expression compared to responders. Front Immunol 2021 Sep 10;12:713351 [FREE Full text] [doi: 10.3389/fimmu.2021.713351] [Medline: 34566969]
6. Starnes DS, Tabor J. The Basic Practice of Statistics (6th Edition). New York, NY: BFW High School Publishers; 2018.
7. Thu Nguyen T, Van Tran K, Cam Ho T, Xuan Nguyen H, Trong Nguyen T. A systematic analysis with the hierarchical cluster analysis strategy on the complex interaction of TERT and CTNNB1 somatic mutations in Vietnamese hepatocellular carcinoma patients. Gene 2024 Nov 15;927:148646. [doi: 10.1016/j.gene.2024.148646] [Medline: 38851365]
8. Nguyen TT, Ho TC, Bui HTT, Tran V, Nguyen TT. Multi-clustering study on the association between human leukocyte antigen and hepatitis B virus-related hepatocellular carcinoma and cirrhosis in Viet Nam. World J Gastroenterol 2024 Dec 14;30(46):4880-4903 [FREE Full text] [doi: 10.3748/wjg.v30.i46.4880] [Medline: 39679310]
9. Nguyen TT, Ho CT, Bui HTT, Ho LK, Ta VT. Multidimensional machine learning for assessing parameters associated with COVID-19 in Vietnam: validation study. JMIR Form Res 2023 Feb 16;7:e42895. [doi: 10.2196/42895] [Medline: 36668902]
10. Chinn S. A simple method for converting an odds ratio to effect size for use in meta‐analysis. Statist Med 2000 Nov;19(22):3127-3131. [doi: 10.1002/1097-0258(20001130)19:22<3127::aid-sim784>3.3.co;2-d]
11. Blockeel H, Devos L, Frénay B, Nanfack G, Nijssen S. Decision trees: from efficient prediction to responsible AI. Front Artif Intell 2023 Jul 26;6:1124553 [FREE Full text] [doi: 10.3389/frai.2023.1124553] [Medline: 37565044]
12. Sánchez-Hechavarría ME, Ghiya S, Carrazana-Escalona R, Cortina-Reyna S, Andreu-Heredia A, Acosta-Batista C, et al. Introduction of application of Gini coefficient to heart rate variability spectrum for mental stress evaluation. Arq Bras Cardiol 2019;113(4):725-733. [doi: 10.5935/abc.20190185] [Medline: 31508693]
13. Cohen J. Statistical Power Analysis for the Behavioral Sciences (2nd ed). Hillsdale, NJ: Lawrence Erlbaum Associates; 1998.
14. Padarath K, Deroubaix A, Kramvis A. The complex role of HBeAg and its precursors in the pathway to hepatocellular carcinoma. Viruses 2023 Mar 27;15(4):857. [doi: 10.3390/v15040857] [Medline: 37112837]
15. Jiang H, Chen C, Yuan D, Ye X, Chen Y, Han G, et al. The relationship of maternal hepatitis B e antigen and response to vaccination of infants born to women with chronic infection. BMC Pregnancy Childbirth 2023 Jul 15;23(1):518 [FREE Full text] [doi: 10.1186/s12884-023-05815-y] [Medline: 37454068]
16. Xu Y, Liu H, Zhong Y, Liu C, Wang Y, Jia L, et al. Peripheral blood mononuclear cell traffic plays a crucial role in mother-to-infant transmission of hepatitis B virus. Int J Biol Sci 2015;11(3):266-273 [FREE Full text] [doi: 10.7150/ijbs.10813] [Medline: 25678845]
17. Ngoc Tram HT, Thu Huong BT, Duc Hinh N, Thuy PT, Lan Anh LT, Kim Phuon DT, et al. Multidimensional analysis of the mother-to-child transmission risk factors in chronic hepatitis B virus infection in pregnant women in Vietnam. Clin Ter 2023;174(3):266-274 [FREE Full text] [doi: 10.7417/CT.2023.2533] [Medline: 37199363]
18. Juszczuk P, Kozak J, Dziczkowski G, Głowania S, Jach T, Probierz B. Real-world data difficulty estimation with the use of entropy. Entropy 2021 Dec 01;23(12):1621 [FREE Full text] [doi: 10.3390/e23121621] [Medline: 34945927]
19. Zhang L, Bi X, Chen X, Zhang L, Xiong Q, Cao W, et al. Front Immunol 2022 Jan 19;13:1119124. [doi: 10.3389/fimmu.2022.1119124] [Medline: 36741383]

Abbreviations

ALT: alanine aminotransferase
AST: aspartate aminotransferase
CART: classification and regression trees
CHB: chronic hepatitis B
EDTA: ethylenediaminetetraacetic acid
HBeAg: hepatitis B e antigen
HBsAg: hepatitis B surface antigen
HBV: hepatitis B virus
ID3: Iterative Dichotomiser 3
mMDSC: monocytic myeloid-derived suppressor cel
MTCT: mother-to-child transmission
PBMC: peripheral blood mononuclear cell
WHO: World Health Organization

XSL•FO
RenderX

Edited by A Mavragani; submitted 10.12.24; peer-reviewed by SM Atah, K Ko; comments to author 02.03.25; revised version received 19.03.25; accepted 30.04.25; published 23.05.25.

Please cite as:
Nguyen Tien D, Thi Thu Bui H, Hoang Thi Ngoc T, Thi Pham T, Trung Nguyen D, Nguyen Thi Thu H, Thu Hang Vu T, Lan Anh Luong T, Thu Hoang L, Cam Tu H, Körber N, Bauer T, Khanh Ho L
A Data-Driven Approach to Assessing Hepatitis B Mother-to-Child Transmission Risk Prediction Model: Machine Learning Perspective
JMIR Form Res 2025;9:e69838
URL: https://formative.jmir.org/2025/1/e69838
doi:10.2196/69838
PMID:

©Dung Nguyen Tien, Huong Thi Thu Bui, Tram Hoang Thi Ngoc, Thuy Thi Pham, Dac Trung Nguyen, Huyen Nguyen Thi Thu, Thi Thu Hang Vu, Thi Lan Anh Luong, Lan Thu Hoang, Ho Cam Tu, Nina Körber, Tanja Bauer, Lam Khanh Ho. Originally published in JMIR Formative Research (https://formative.jmir.org), 23.05.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Ambivalent User Needs as a Challenge and Chance for the Design of a Web-Based Intervention for Gaming Disorder: Qualitative Interview Study With Adolescents and Young Adults

Birte Linny Geisler[1,2], MA, Dr sc hum; Kay Uwe Petersen[2,3], Dipl Psych, PhD; Sara Hanke[2,3], Dipl Psych; Simon Schurer[1,2], MA; Anne Schreiber[1,2], MD; Christine Lämmle[1,2], MA, Dr biol hum; Anil Batra[2,3], Prof Dr Med; Tobias Renner[1,2], Prof Dr Med; Isabel Brandhorst[1,2], Dipl Psych, Dr rer nat

[1]Department of Child and Adolescent Psychiatry, Psychosomatics and Psychotherapy, University Hospital Tübingen, Tübingen, Germany

[2]German Center for Mental Health (DZPG), Partner Site Tübingen, Tübingen, Germany

[3]Department of Psychiatry and Psychotherapy, Section of Addiction Medicine and Addiction Research, University Hospital Tübingen, Tübingen, Germany

Corresponding Author:
Isabel Brandhorst, Dipl Psych, Dr rer nat
Department of Child and Adolescent Psychiatry, Psychosomatics and Psychotherapy
University Hospital Tübingen
Osianderstr. 14
Tübingen, 72076
Germany
Phone: 49 07071 29 ext 82338
Email: isabel.brandhorst@med.uni-tuebingen.de

Abstract

Background: In Germany, there are still many young people with gaming disorder (GD) who do not use or cannot access existing treatment services. Given the increasing prevalence of internet use disorders and GD, especially among young people in Germany, there is a need to provide additional low-threshold treatment options that are easily accessible anywhere. Web-based interventions (WBIs) can be used to achieve this goal.

Objective: The aim of this study was to explore the treatment needs of young people with GD in Germany and derive implications for the development of a self-guided WBI for GD.

Methods: Using a qualitative study design, we conducted a focus group with 3 young male adults and semistructured individual interviews with 3 male adolescents. Data were analyzed using qualitative content analysis. The reporting of this study followed the COREQ (Consolidated Criteria for Reporting Qualitative Research) guidelines.

Results: Participants' expectations of web-based help in general and of a self-guided WBI for GD revealed a wide variety of sometimes conflicting user needs. For example, by analyzing participants' experiences with successful strategies, we found that external stabilizers (eg, parental control and support group meetings) were helpful in managing GD. However, with regard to a WBI, participants described it as a barrier if the WBI created "too much pressure." On the other hand, "not enough pressure" (ie, not enough external control) was also mentioned as a barrier. The belief that gaming is rewarding and that only equally rewarding activities are successful alternatives to gaming is in tension with the fact that changing problematic gaming behavior can be stressful and not feel rewarding at all. The data also showed that, on the one hand, a WBI should be designed to be attractive (eg, by incorporating gaming elements) but that it should not be too attractive as this, in turn, could trigger GD.

Conclusions: A self-guided WBI for GD should consider and address conflicting user needs. Ambivalence of needs in the face of coping with GD should not be seen as a problem but as a normal part of a change process and, therefore, actively integrated into the WBI concept and storyline.

Trial Registration: German Clinical Trials Register DRKS00032334; https://drks.de/search/en/trial/DRKS00032334

(JMIR Form Res 2025;9:e63258) doi:10.2196/63258

XSL•FO
RenderX

KEYWORDS

gaming disorder; user-centered design; self-guided web-based intervention; adolescents; young adults; youth; treatment needs; digital intervention; artificial intelligence; AI

Introduction

Background

According to the *International Classification of Diseases, 11th Revision* (code 6C51), gaming disorder (GD) can be defined as an online or offline gaming behavior that involves reduced control over gaming, an increase in the prioritization of gaming over other life interests and activities, and the continuation or escalation of gaming despite its negative consequences. Gaming behavior significantly interferes with important areas of life, such as family, education, and work. To be diagnosed with GD, the behavior must typically occur over a period of at least 12 months [1]. In addition to GD, there are other problematic behaviors that can occur in the context of internet use, including those related to social media, online shopping, online pornography, and online gambling. The term internet use disorder (IUD) is sometimes used as an umbrella term for these problem areas. In some cases, IUD is also used as a separate disorder category because users often problematically use multiple internet applications at the same time. The relevance and prevalence of IUD and GD is increasing, especially since the COVID-19 pandemic [2-4].

In Germany, it can be assumed that young people with IUD or GD are still not adequately cared for due to the lack of specialized care centers [5]. During the COVID-19 pandemic, the prevalence of GD increased rapidly among young people [6]. In 2023, there was an estimated 4.3% of individuals aged 10 to 17 years in Germany who had GD. A total of 11.1% met the *International Classification of Diseases, 11th Revision*, criteria for risky gaming behavior, which corresponds to approximately 680,000 affected children and adolescents [7]. Therefore, there is a need to provide low-threshold treatment options that are easily accessible anywhere. Web-based treatment options are ideal for achieving this goal.

With the Online Bridge project (in German: Onlinebrücke), the authors of this paper created Breaking the Game, an innovative treatment offer in the form of a self-guided web-based intervention (WBI; Textbox 1). WBIs can be defined as digital programs on the internet aimed at treating or counseling people with (mental) health problems [8]. Breaking the Game focuses on adolescents and young adults aged ≥12 years with symptoms of GD. The first part of the project, focusing on the development of the WBI, ran from November 2021 to December 2023 and included the qualitative study presented in this paper. This study was registered in the German Clinical Trials Register (DRKS00032334). The project was funded by the Ministry of Social Affairs, Health, and Integration of the state of Baden-Württemberg in Germany.

Textbox 1. Types of web-based interventions (WBIs).

- Guided WBIs include human therapists providing treatment via webcam, telephone, email, or messages [9,10], such as for internet use disorder (IUD) treatment in the form of a manualized online counseling service using motivational interviewing techniques [2,11,12].
- Self-guided WBIs do not involve a human therapist but provide a predesigned program [9,10], such as an app-based intervention for IUD treatment [2,13].
- Blended WBIs combine digital and analogue treatment or self-guided and guided elements [9,10], such as for IUD treatment as a stepped-care approach combining an app-based self-guided program with webcam counseling [2,14].

However, in conceptualizing the WBI, the research group was faced with questions that could not be answered by the research literature at the time. In WBIs for GD, treatment is offered over the internet, where gaming itself takes place. This distinguishes GD from other (mental) health problems addressed in WBIs, such as diabetes or depression. From the experience of the German WBI OMPRIS, which offered webcam-based counseling, we know that guided WBIs for IUD and GD can work successfully. However, due to the potentially noncommittal nature of the internet environment, aspects of motivation should be considered when developing a successful web-based treatment program [11]. The question was how this could be applied to a *self-guided* WBI for GD. In a self-guided WBI, there is no contact with a counselor or therapist, which could increase the risk of leaving the WBI with a single click. In addition, the question was raised of how to design an attractive WBI for treating GD so that young people affected by GD who are familiar with well-designed games would find it attractive enough to return to it regularly.

In addition, a high variability of characteristics within the target population of young people with GD had to be considered, including age, personality traits, comorbidities, educational background, mental health literacy, and gender [15-17]. Research shows that the development of GD is influenced not only by the amount of time spent online [4] but also by the interaction of several risk and influencing factors [18]. Biographical influences, such as family factors, play a role [19]. For example, the parent-child relationship is an important factor in both the development and maintenance of IUDs [20]. Gender differences in the development and maintenance of GD should be considered, such as the level of impulsivity and impaired inhibitory control or the role of aggressive feelings and craving [21]. In addition, many people do not yet play games pathologically but already do so in an uncontrolled manner [22]. Problem gamers should also be reached with help offers to prevent them from developing GD.

Research into the treatment of GD has examined not only analogue but also digital treatment options. Studies show that

XSL•FO
RenderX

both adults and young people with mental health problems can be effectively reached through WBIs, possibly even more effectively than through analogue interventions. For example, an Australian study of adolescents' attitudes toward online therapy found that 72% of the adolescents surveyed would use online therapy if they had mental health problems. A third would even prefer online therapy to traditional face-to-face options, citing reduced stigma and increased accessibility [23]. Adolescents and young adults already spend a significant amount of time online. When seeking help, it is plausible that young people will (as a first step) turn to the internet as a familiar environment rather than to a local therapist. Therefore, WBIs can fill a gap in care, especially when it comes to reaching young people with mental health problems [24,25].

Studies suggest that WBIs for IUD and GD may also be helpful and effective [2,26], for example, in preventing GD by improving self-regulation [27]. As people with IUD or GD spend even more time online than average, it is likely that treatment services offered online are accessible at a lower threshold than analogue treatment services [11]. In addition, IUD and GD are associated with comorbidities such as depression, attention-deficit/hyperactivity disorder, and social phobia [28]. It is reasonable to assume that people with IUD or GD combined with these comorbidities may have additional problems seeking analogue therapy. Therefore, easily accessible WBIs may be an important treatment option.

The development of a self-guided WBI differs in many ways from the development of a guided WBI (eg, webcam-based individual counseling; Textbox 1) [8]. Self-guided WBI content must be predesigned and standardized, with the risk that the final version will not meet all the needs of a heterogeneous target group. This may be the reason poor adherence to self-guided WBIs remains a problem [29]. However, self-guided WBIs have a number of advantages that distinguish them from other intervention types. They are cost-effective because they do not require human therapists. They can be used at the user's own pace and needs, with time flexibility and little effort and without long waiting times [9,30]. Users of self-guided WBIs can manage their treatment processes independently, which can increase users' autonomy, self-determination, and self-efficacy. In addition, self-guided WBIs can be used anonymously and without commitment, which is likely to increase willingness to register. Studies suggest that guided WBIs involving human therapists are more effective than self-guided WBIs (ie, they achieve higher rates of adherence) [9,10,31-35]. However, other research shows that the design of the intervention determines its ultimate effectiveness [9,36]. In the area of disordered gambling, a study showed that intensively designed self-guided WBIs with more than 6 thematic modules were more effective than short and less intensively designed face-to-face options [37]. This illustrates the importance of specific design choices for the subsequent effectiveness of a WBI. Therefore, although research suggests that WBIs are generally effective, this proved to be insufficient information for the development of a new self-guided WBI for a young German target group because of the lack of detailed information from the users' perspective about the assumed effectiveness based on the WBI design.

Participatory and collaborative approaches are playing an increasingly important role in the development and research of treatment options in psychiatry and psychotherapy [38-40]. However, the involvement of future users in the development of treatment services for GD and IUD from the outset has long been lacking in German studies. This applies to both analogue and digital services. To our knowledge, no study has yet investigated the treatment needs of young Germans with GD symptoms regarding a self-guided WBI using a participatory approach before and during the WBI development process. In general, participatory approaches (eg, by using qualitative methods that invite people with IUDs to share their personal experiences with IUDs and their treatment needs) are still rarely used in German web-based IUD treatment research. Dreier and Wölfling [2] conducted a qualitative study on patients' perceptions of online therapy. However, this analysis showed how patients perceived treatment services after they had already been developed. Chen et al [41], who developed a self-guided WBI for adolescents with GD, involved end users by observing participants using the first version of the WBI and conducting a usability survey. Hanke et al [42] conducted a survey to investigate the needs of parents of adolescents with IUD regarding a WBI. However, there are still WBIs for GD that have not been developed based on a user-centered or human-centered design approach [43,44], which involves the cocreation of an intervention with end-user representatives [45,46]. Therefore, to develop a self-guided WBI for young Germans with GD, the research group lacked information that can only be provided by people with GD and potential users themselves.

Due to the presented research gaps, especially regarding the young German target population with GD, a qualitative interview study with German adolescents and young adults with GD symptoms was conducted within the Onlinebrücke project to collect information from the perspective of potential users to be included in the development of the self-guided WBI Breaking the Game. The interim and final results of the qualitative study were incorporated into the parallel development of the WBI. In addition to integrating the results of this study, we developed the WBI in close collaboration with addiction counseling centers.

Study Objectives

The aim of this qualitative study was to understand the treatment needs of German adolescents and young adults affected by GD to inform the development of a self-guided WBI for a German target group. The overall research question was as follows: What user needs must be considered when developing a target group–oriented self-guided WBI for adolescents and young adults with GD symptoms in Germany?

This led to the following subquestions: (1) What are the treatment needs of adolescents and young adults with GD symptoms in Germany regarding a self-guided WBI for GD? (2) What are the implications for the development and design of a self-guided WBI for GD for a young German target group?

XSL•FO
RenderX

Methods

Study Design

A qualitative interview study was conducted. One focus group and 3 semistructured interviews were used for data collection as these methods are suitable for capturing people's perspectives and experiences on a predefined topic. Qualitative content analysis according to Kuckartz and Rädiker [47] was used to analyze the data as this method allows for deductive and inductive development of coding categories. In this way, it was possible to find unexpected results as well as follow predefined categories based on the research questions. The ability to analyze the data in a more interpretive way when needed helped gain insights into user needs that are often hidden in implicit statements [47]. The COREQ (Consolidated Criteria for Reporting Qualitative Research) guidelines were followed in reporting this study (Multimedia Appendix 1) [48].

Ethical Considerations

Approval was granted by the ethics committee at the Medical Faculty of the Eberhard Karls University of Tübingen and University Hospital Tübingen on February 15, 2023, under the reference number 843/2022BO2. This study was conducted in accordance with the Declaration of Helsinki. Participants gave their informed consent to participate before data collection. In the case of minors, the parents also signed the consent form. Participants received no monetary compensation for taking part in this study. The interviews were transcribed pseudonymously by a certified transcription service (Amanu). The transcripts were stored on protected servers at University Hospital Tübingen.

Study Population

The aim was to interview adolescents and young adults with GD symptoms as this was the target group for the Breaking the Game WBI. The young adults in the study population were recruited through a self-help network in Germany by author IB as the research group leader. Adolescents were recruited at a German child and adolescent psychiatric outpatient clinic by author SS, who was a master's student in psychology. Finding interviewees proved to be a challenge. Due to limited time and personnel resources, the search focused on institutions within the health care system and did not extend to settings of everyday life (eg, schools, universities, and youth centers). Finally, participants were selected through a combination of purposive sampling (ie, directly contacting an IUD support group where people who met the inclusion criteria could be found) and convenience sampling (ie, spreading the word among clinic staff to look for young patients with GD symptoms). During recruitment, the inclusion criteria were adjusted to include current or past experience with GD symptoms and an age range that approximated the age range of future WBI users. For example, although the former age range of the WBI was 12 to 21 years (later changed to ≥12 years), participants aged >21 years were recruited because of their experience with GD at a younger age and their interest in speaking with researchers. A diagnosis of GD was not required. One potential participant dropped out of the study. He did not attend the agreed appointment, and no alternative date could be arranged. The time frame for data collection was 6 months. Data collection was completed when 3 male young adults and 3 male adolescents had been interviewed. The heterogeneity of the study population in terms of GD experience was acceptable. However, it can be assumed that, if recruitment had also taken place outside the health care system, female participants would most likely have been reached. In Germany, there are still more men with GD than women with GD in the health care system.

The sample in this study consisted of 3 male young adults in their 20s (P1, P2, and P3) and 3 male adolescents between the ages of 12 and 16 years (P4, P5, and P6). Textbox 2 provides more details.

Textbox 2. Sample description.

- P1 (male; in his 20s; university student) first noticed symptoms of gaming disorder (GD) at the beginning of his studies. During exam periods, he used games to procrastinate. During this time, he felt the need to change. Therefore, several years ago, he got rid of his notebook. Since then, his gaming problems have gradually improved. However, in retrospect, there were already symptoms of GD during his school years.
- P2 (male; in his 20s; university student) experienced symptoms of GD as a teenager. However, when he lived at home with his parents, gaming was not a problem because his parents had some control over his gaming. The peak of the problem occurred at the beginning of his university years, after he had moved out of his parents' home. During the first semester, he procrastinated a lot and experienced a relationship breakup. At that time, he was playing games for up to 16 hours a day. At the time of data collection, GD symptoms remained a challenge.
- P3 (male; in his 20s; university student) dropped out of university because of GD symptoms. He procrastinated a lot by playing games, which negatively affected his academic performance. This was the first time that he realized that he had no control over his gaming behavior. At the time of data collection, he was just beginning to change his gaming behavior.
- P4 (male; aged 12-16 years; high school student) played on PC and PlayStation mostly with friends he knew from offline life. In addition to gaming, he mentioned several analogue hobbies (eg, sports and music). He mentioned symptoms of GD when he was in elementary school. When his parents took away his cell phone, he played games on Nintendo Switch and hid it from his parents. At the time of data collection, he did not consider his gaming behavior problematic, although his mother did.
- P5 (male; aged 12-16 years; high school student) reported negative effects of gaming, such as not being on time for dinner or not studying for school. However, he himself did not seem to find his gaming problematic, seeing it primarily as an opportunity to have fun and be in contact with his friends. He would stop gaming if his friends asked him to play soccer with them and also mentioned various sports as analogue hobbies.
- P6 (male; aged 12-16 years; high school student) played games after stressful days because it relaxed him. He also played games when he was bored and when he needed to feel successful. At the time of data collection, he did not seem to consider his gaming as problematic, but there were symptoms of GD in the past. He had already reduced his gaming.

XSL•FO
RenderX

Data Collection

Data collection was conducted between January 2023 and July 2023 by author SS, who was new to qualitative methods and IUD research at that time. Therefore, authors IB and BLG and other members of the project group, who have experience in qualitative research, IUD research, or IUD treatment, were involved in the design of the interview guides (Textbox 3). The project group consisted of male and female researchers between the ages of 20 and 60 years with backgrounds in psychology, medicine, health sciences, and communication sciences and with many years of experience in IUD research and therapy.

Textbox 3. Interview guide development.

The interview guide was developed by translating the research interests into interview questions based on the "SPSS principle for creating interview guidelines" according to Helfferich [49]. In step 1, author SS collected as many questions as possible that could be of interest in relation to the research topic. In step 2, under the supervision of author BLG, SS gradually reduced the questions to those that were actually useful for answering the research questions. In a next round, and in exchange with BLG and the rest of the project group, SS performed steps 3 and 4 (sorting and subsuming). Finally, SS presented the interview guide to the Qualitative Methods Research Workshop at University Hospital Tübingen. After incorporating feedback from the workshop members and further consultation with the research team, SS finalized the interview guide.

The final interview guides, one for the focus group and one for the semistructured interviews, included the following topics: (1) participants' self-introduction and their reasons for playing computer games, (2) experiences with (problematic) use of computer games and motivation to seek help, (3) experiences with strategies and offers of help that ameliorated problematic gaming behavior, (4) experience with online services in general, (5) user needs for a self-guided web-based intervention for gaming disorder, and (6) an open final question—"Is there anything else you would like to add at the end?"

The focus group lasted 98 minutes. It was moderated by author SS as a young male individual whom the participants might not associate with an authority figure, such as a parent or teacher. For this reason, author KUP, as the senior researcher who assisted and took field notes, deliberately remained in the background. Author SS introduced himself not only as a psychologist but also as a recreational gamer himself to establish a trusting relationship and invite participants to respond in detail. The semistructured interviews lasted 60 minutes each and were conducted by author SS alone. All interviews were audiotaped and transcribed. Participants in the semistructured interviews were familiar with author SS before the interviews. SS had no previous relationship with the focus group participants.

Analysis

Data analysis followed the variant of qualitative content analysis by Kuckartz and Rädiker [47]. The pseudonymized transcripts were coded and analyzed by an experienced qualitative researcher (author BLG) combined with peer debriefings. The coding process was based on the principle of consensual coding [47] (ie, in this case, joint coding and joint discussions of codings and themes). It was decided against an approach that uses the principle of intercoder reliability as a validation strategy [50] because it is based on a quantitative paradigm [51] or postpositivist paradigm [52]. Rather, we followed a "Big Q" approach based on a qualitative paradigm [51] or interpretivist paradigm [52]. In this case, the validity of the analysis is not shown by the fact that different coders were able to agree on one "truth" but by allowing and discussing different perspectives on the data in exchange with other researchers. During the coding process, BLG repeatedly selected excerpts from all interview transcripts for joint coding. Data sessions took place with members of the research team (5 additional researchers participated: AS, KUP, IB, SH, and CL) and at the Qualitative Methods Research Workshop at University Hospital Tübingen, which is attended by an average of 5 to 8 researchers. In this way, a joint analysis took place to ensure the quality of the analysis process. This included the structuring and revision of the category system as well as the joint coding and discussing of rich and complex text passages. BLG also used the exchange to continually reflect and critically question her own perspective on the material. BLG wrote protocols for the analysis sessions and integrated the findings into the subsequent analysis process.

The coding process (see Textbox 4 [47] for details) was supported by the MAXQDA software (VERBI GmbH). Further in-depth and cross-case analyses were conducted using the coding tree (Multimedia Appendix 2) and theme matrices (see an example in Multimedia Appendix 3 [47]).

XSL•FO
RenderX

Textbox 4. Data analysis process. The data analysis in this study followed the qualitative content analysis by Kuckartz and Rädiker. It was slightly modified.

- Familiarization with the material through repeated readings and initial comments
- Deductive development of major themes in advance based on research questions and interview guide
- Coding one transcript at a time guided by major themes and, during the coding process, deriving inductive codes (called subcategories in the coding tree) from the data while simultaneously coding the transcript
- Writing case summaries and memos
- Reorganizing the coding tree based on 3 levels of categories: superordinate categories, main categories, and subcategories; inductively generating the subcategories that contain detailed facets of the main category
- After coding all transcripts, creating theme matrices "cases times categories" for each superordinate category
- Coding the material again
- Comparisons within and across cases and further in-depth analysis using the theme matrices
- Identifying patterns and ambivalences in the data using the coding tree and theme matrices with repeated returns to the original data

Results

Overview

This study aimed to understand the treatment needs of young people with GD in relation to the development of a self-guided WBI for GD. However, during the analysis, it became clear that focusing solely on explicitly stated user needs would not be sufficient. In analyzing the data, it became necessary to differentiate the treatment needs, which resulted in the following overarching themes: experiences and attitudes as contextual factors, as well as specific expectations of a WBI, assuming that contextual factors, in turn, influence expectations of a GD intervention (Figure 1). This is in line with the International Organization for Standardization 9241-210:2010 standard, which states that human-centered design activities should also include understanding and specifying the context of use [53]. In the following sections, we first present major themes that involve contextual factors: (1) experience with GD symptoms, (2) experience with successful strategies and barriers, (3) experience with web-based support in general, (4) beliefs about gaming and GD, and (5) motivations for change. We then present the major themes: (6) expectations of a self-guided WBI in general, (7) expectations of a self-guided WBI for GD, and (8) expected barriers to its use. Direct quotes were translated from German and edited for clarity (ie, condensed to the most relevant parts or edited if jargon was used or if the original sentence was not grammatically correct).

XSL•FO
RenderX

Figure 1. Operationalization of treatment needs in this study as major themes. GD: gaming disorder; WBI: web-based intervention.

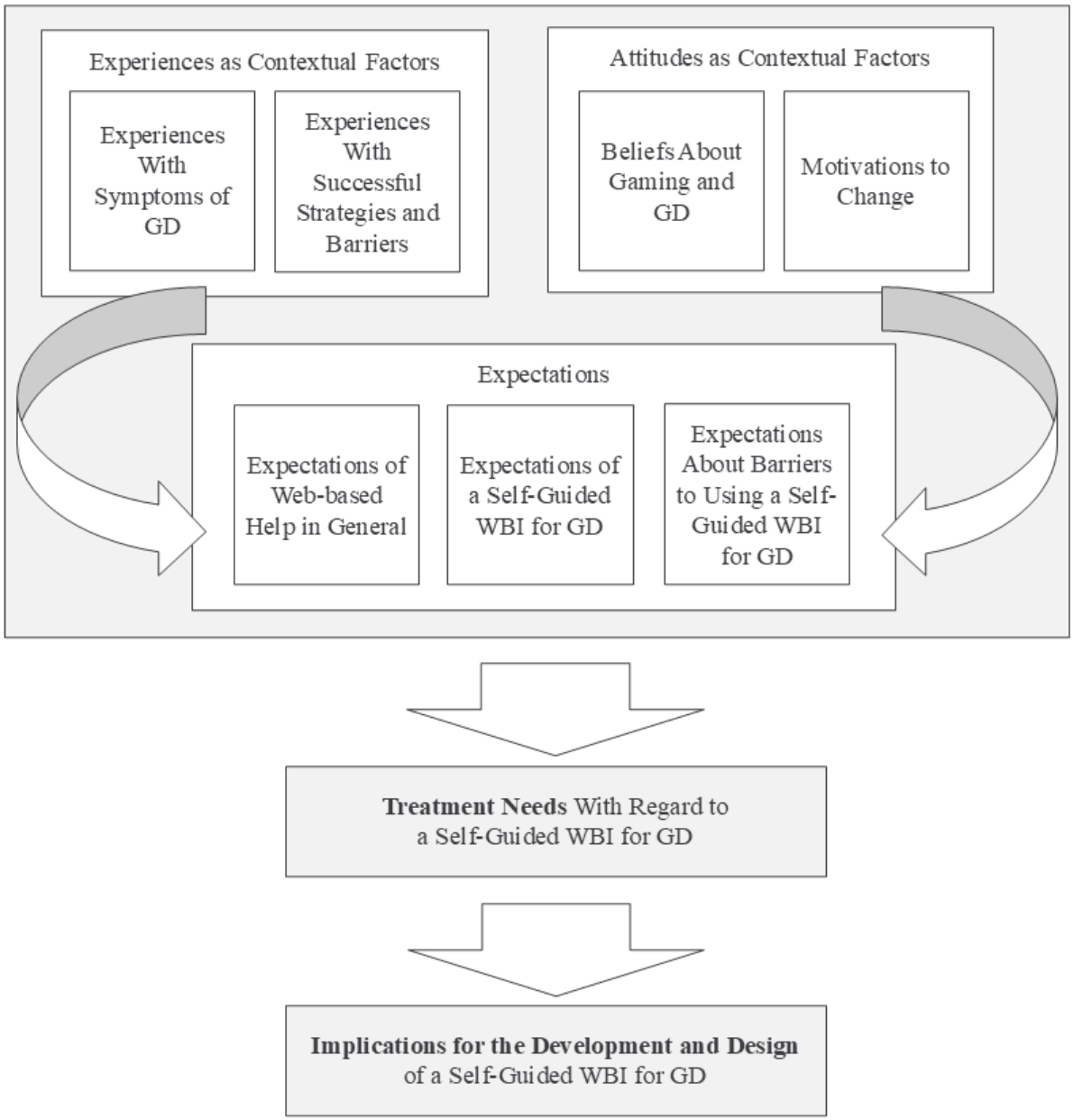

Experiences as Contextual Factors

Experiences With Symptoms of GD

All participants had been gaming for many years, had experienced (temporary) symptoms of GD, and had experienced negative consequences of gaming. The participants identified needs that are met through gaming, as well as risk situations.

Several participants reported negative social consequences, such as a negative impact on school, studies, and family life (P1, P3, P5, and P6). Some participants reported unpleasant emotional consequences of gaming (P1, P3, P5, and P6), such as loss of control, anger after failing in the game, and feeling empty after finishing the game. Physical consequences were also reported, such as experiencing a “hot head” from playing for too long:

> *But if you play too much, you can get stupid...Because then I’ve noticed that my head gets really hot.* [P6; aged 12-16 years]

Nevertheless, regardless of the negative effects, gaming and even symptoms of GD seemed to meet certain needs. All participants used gaming to regulate unpleasant emotional states, such as dissatisfaction, boredom, or stress. In addition, several participants used gaming to create positive emotional states, such as feeling successful, feeling stimulated, having fun, experiencing adventure, and as a balance after a day full of obligations (P3, P4, P5, and P6). Another important aspect was the regulation of social needs through gaming—for P4 and P5, gaming provided an opportunity to meet friends from school. On the other hand, for P3, one of the older participants, gaming

XSL•FO
RenderX

was about "finally" having time for himself (ie, distancing oneself from social contacts).

In addition, the data indicated various risk situations in which participants showed symptoms of GD. For older participants, the importance of transitional periods as risk situations became clear. For P1, P2, and P3, GD symptoms appeared at the beginning of university combined with moving out of the parental home (which led to the loss of parental control). Other factors contributing to the development of GD symptoms included study-related stress situations, such as exams, a general lack of structure in daily life, and personal crises, such as separation from a romantic partner. In addition, P1, P2, and P3 reflected that their gaming had long been an invisible problem that developed unnoticed until they moved out:

> *P1: I think it was in 9th or 10th grade when I first started thinking about it. StarCraft II was a problem at that time. But it never had any direct consequences because I was doing well in school and getting good grades. So my parents were not worried...And up until then, I didn't really have much of a social environment that I would have lost touch with. So, it...lay dormant for quite a while, I'd say...without it hurting.*
>
> *P3: Yes...It lies dormant for a couple of years. Until you get out of school. And then it really hits you...during your studies. I also feel like it starts earlier and then it boils over.* [Focus group; participants in their 20s]
>
> *This is consistent with the experiences of the younger participants (P4, P5, and P6), who were still living with their parents at the time of the study. They also showed uncontrolled gaming, especially when parental control was lacking, such as when parents were absent. Other risk factors affecting the adolescents were sleep problems (ie, playing because of sleeplessness) and peer groups that also played a lot.*

Experiences With Successful Strategies and Barriers

All participants reported successful use of strategies to cope with GD symptoms outside of the internet, as well as barriers to using these strategies.

All participants seemed to benefit from external stabilizers, such as parental control, daily routines, and support group meetings. Strategies also included reducing gaming or taking a break from gaming (P1, P2, P3, and P6), such as uninstalling the game, deleting the account, or using an app blocker. The older participants (P1, P2, and P3) mentioned different variations of self-reflection as a strategy, such as reflecting on general life satisfaction. The strategy of taking personal responsibility was also important (P2, P4, P5, and P6; eg, making rules for oneself). On the other hand, not expecting too much of oneself and avoiding extremes (P4, P5, and P6) was also mentioned as a strategy, for example, not imposing excessive restrictions on oneself and allowing oneself a defined daily dose of gaming. For the older participants and P5, alternative activities were important, especially sports and activities that were fun or provided a "high" (see the Beliefs About Gaming and GD section).

However, participants also experienced barriers when using strategies. In particular, the barrier of too much pressure should be noted (P1, P2, P3, and P6) as opposed to the barrier of not enough pressure (P1, P2, P3, P4, and P5). On the one hand, the barrier of too much pressure (see the following quotation) referred to strategies that were too strict or overwhelming and, therefore, actually prevented addressing GD symptoms, such as stopping gaming overnight without building alternatives in parallel, trying to cope with GD alone, using treatment tools that were not applicable, and making agreements with others to reduce or stop gaming (P1, P2, P3, and P6). On the other hand, the barrier of not enough pressure also prevented them from addressing GD symptoms, including not being sufficiently aware of the GD symptoms and, therefore, not feeling the need for help or negative consequences not being "painful" enough:

> *P2: If you promise your parents that you won't play this week and then you don't keep the promise because the pressure of addiction is too great, then it's rather counterproductive [to make agreements at all]. Then I feel like I've disappointed my parents...I'm so ashamed that I feel...even worse...And then I actually play even more.*
>
> *P1: Right, then you might want to drown this feeling of shame, and then it gets worse.*
>
> *P2: Yes, yes, exactly.*
>
> *P3: Agreements in general never helped me either...My mother once wanted to make a plan with me. I stuck to the plan for one day at most, and then I threw it out again.*
>
> *P2: It's the same with making plans with oneself.*
>
> *P3: Making plans with oneself is also terrible. It doesn't work either.* [Focus group; participants in their 20s]

Attitudes as Contextual Factors

Beliefs About Gaming and GD

Participants held certain beliefs about gaming and GD. Of particular relevance were the beliefs of lack of agency versus agency and the dopamine hypothesis.

Regarding the lack of agency belief, the older participants seemed to be partially convinced that they had no agency in their GD (P1, P2, P3, and P6). This included the belief that there was a "gamer gene" that could be activated repeatedly. In this case, the participants seemed to experience their game as an overwhelming force rather than as a behavior of their own. The game had agency, not the player. Therefore, when using a self-guided WBI for GD, technological barriers to the game would be necessary:

> *I need some kind of [separate] electronic device on which I can do this online training...A blocker while I'm using this training, so that I don't just drop out and end up in a game.* [P1; focus group; participants in their 20s]

XSL•FO
RenderX

In addition, P2 and P3 believed that proper parental behavior regarding their children's media use was crucial in preventing the child from later developing GD. This can be interpreted as parents being responsible for later GD, not the gamers themselves. P6 believed that too much parental control could lead to the development of addiction:

> *Parents sometimes create the addiction themselves by letting you play so little.* [P6; aged 12-16 years]

In this case, the agency lay with the parents and not with the gamers. However, at the same time, many participants emphasized the importance of personal responsibility, which in turn implies having agency (P2, P3, P4, and P6).

This belief was evident in both older and younger participants, for example, the belief that a person has to really want something; otherwise, nothing will change. P2 said that it was not the games themselves that caused the problem but people's thoughtless use of them. These beliefs implied that it was not the game that had the power but the gamer. It is important to note that this is not an "either/or" situation; rather, participants seemed to feel both a lack of agency on the one hand and full responsibility for their behavior on the other.

Another relevant belief was the so-called dopamine hypothesis (P1, P2, P3, and P6). This belief implied that gaming was associated with rewarding effects and that the "highs" experienced in gaming were difficult to find elsewhere. Therefore, if the alternative activity to gaming was not sufficiently attractive, one would inevitably play again. Only those activities that provided a "kick" were worthwhile:

> *I realized that doing sports helps me very little. Especially weight training doesn't help me because I don't get anything out of it. It's all about dopamine, and I just don't get any dopamine from it.* [P3; participant in his 20s]

Belief in the dopamine hypothesis seemed to influence expectations regarding the design of a WBI for GD—if the WBI did not provide the necessary rewarding experience, it might even be a risk factor as switching to the game was only a click away:

> *And when the game is only 2 clicks away, and your dopamine is only 2 clicks away, you end up in the game very quickly.* [P3; participant in his 20s]

In this case, the dopamine hypothesis theme can be seen in connection with the lack of agency theme—the rewarding effect had ultimate power over the player.

Motivations to Change

Participants mentioned several motivational factors as to why they had accepted help regarding their GD symptoms in the past: the opportunity to connect with others; autonomy and making one's own decisions; positive, solution-oriented attitudes; and emotional distress.

The ability to connect with others was an important motivational factor for most participants (P1, P2, P3, P5, and P6). For example, they felt motivated when another person gently approached them about their GD, when they could share general life experiences, and when they received appreciative feedback from others about their progress. Connecting with others seemed to be particularly motivating when a participant was overly self-critical:

> *If you are very, very negative in your self-evaluation, sometimes an outside influence is very good...You think: I've just been gaming here all week. And then somebody else tells you: Yeah, but...first of all, you came back [to the support group]. And secondly, you did this and that differently that day.* [P1; participant in his 20s]

Another motivating factor was autonomy and making one's own decisions (P2, P5, and P6). This included, for example, being able to decide whether an offer of help was right or develop one's own solutions. P6 (aged 12-16 years) in particular seemed to be someone who was motivated by a sense of autonomy. Gaining autonomy and developing one's own identity is important during adolescence. Therefore, it seemed to be very important for P6 to be actively involved in decisions—arrangements about gaming between him and his parents had to be workable and understandable, and he wanted to be in agreement with them. In addition, he did not like to be controlled when following the rules; instead, agreements should be based on trust.

This was consistent with another motivating factor: offering help in a positive, solution-oriented attitude rather than a narrow focus on the problem (P1, P2, P3, P5, and P6). Almost all participants felt motivated when an offer of help included the development of attractive alternatives and goals and encouragement to change by focusing on positive prospects and promises of success. Interestingly, P6 needed to be first shaken and startled and then presented with a solution:

> *In my case, it...should scare me a little bit. That is how advertising works: If you want to get money from customers, you have to scare them a little [laughs]. Then you have to offer a solution, and then [the customer] clicks on it.* [P6; aged 12-16 years]

In this case, P6 combined the solution-oriented motivating factor with emotional distress as a motivating factor. Emotional distress was especially reported by older participants as the initial motivation for seeking help (P1, P2, and P3), for example, caused by the negative impact of gaming on academic performance. P1 reported that the experience of losing control over gaming led to feelings of shame, which motivated him to seek help. However, at another point in the focus group, shame was mentioned as a barrier. Emotional distress seemed to work in both ways: it could motivate change, but it could also lead to the avoidance of change.

Expectations

Expectations of Web-Based Help in General

Interestingly, only 1 participant, P5, had experience with web-based help in the form of a self-guided online media literacy training. The following characteristics of a self-guided WBI were derived from P5's positive and negative experiences with this type of web-based help:

XSL•FO
RenderX

- Quality design—the WBI should be attractively designed. Content should be presented in a variety of ways, for example, not only textual content but also playful elements.
- Comprehensibility—the WBI should be easy to understand, for example, by providing a guide with explanatory elements on how to use the WBI and providing age-appropriate content while avoiding unfamiliar terms.
- Challenging—the WBI should encourage self-reflection, for example, by offering relevant new information or providing a final quiz at the end of a unit to check the user's learning status.
- Autonomy enablement—the user should be able to actively search for relevant content. The WBI should be accessible through a variety of devices.

These ideal characteristics of a self-guided WBI in general came from only 1 participant. However, they were formulated in such a nuanced way and related so well to the experiences described in the following section that P5's evaluations are presented here anyway.

Expectations of a Self-Guided WBI for GD

The participants expressed a variety of expectations for a potential self-guided WBI for GD.

In terms of content, all participants expected the self-guided WBI to connect with their everyday life, for example, by referring to the lived experiences of real people rather than to scientific theory:

> *Interviewer: For example, what would an online training absolutely have to do or include for you to say I'm going to take a look at it?*
>
> *P6: It would have to appeal to the person; it would have to come from experience; and it would definitely have to be logical. And it would have to be genuine...You know, a study can mean a lot of things, but...what worked in the end is what worked and not what the study says.* [Aged 12-16 years]

In addition, a self-guided WBI should provide content that invites self-reflection (P1, P2, P3, P4, and P5), for example, by asking questions about gaming behavior and tracking it over time or by inviting nonjudgmental reflection on negative consequences. All adolescent participants wanted the WBI to offer personal counseling (P4, P5, and P6), whereas 2 of the older participants (P1 and P2) saw the self-guided WBI as a bridge to local counseling. Again, it was P5 who had detailed ideas when asked about the content design of a self-guided WBI for GD:

> *I can imagine that you can...choose how old you are...You can put in different hobbies or what you do. Then you can put in what your schedule or a typical week looks like, when you play, and how much you play. Every 3 or 4 days you can go back to the website and put in how long you've been playing, what you've been playing, and what you've been doing. At the end of the week, you can discuss with someone how long you've been playing. And then they can have a look at it, and then you can make a plan for what you want to do.* [P5; aged 12-16 years]

P4 also provided detailed information. The WBI should (1) discuss the initial gaming behavior and develop a plan for change, (2) facilitate the solution-oriented development of a vision or goal, (3) keep in touch with the user and ask about the status of the change, (4) allow for tracking of change implementation, (5) provide a reminder function to limit play time, and (6) suggest motivating and rewarding alternatives to make it easier to stop playing.

What can be seen here is the users' need for a self-guided WBI that repeatedly communicates with them, similar to a human counselor or therapist.

In terms of structure, participants wanted a self-guided WBI to be flexibly accessible at a low threshold (P2, P4, P5, and P6). For example, it should be usable on all devices, browser based, and spontaneously accessible in critical situations without a time-consuming registration process. Participants' statements about the desired duration and frequency of use varied—P1, P2, P3, and P5 expected to use a WBI for 15 to 30 minutes per day, whereas P6 expected to use a WBI only once for 30 minutes and the other participants would use it several times. P4, on the other hand, only wanted short videos of 3 to 5 minutes:

> *Interviewer: What do you think? Fifteen minutes, half an hour, an hour? How long would it take to watch videos to get back on track?*
>
> *P4: I think the TikTok video I watched was 2 or 3 minutes long. But it was not unnecessary talk; it was real words. Nice and clear, all well-spoken. Right. And I really got it. It's true. And it doesn't take a long time. But the longer you think about it, the better it is. If you get it right, then you need...5 minutes to look at it. And then you sit down somewhere and...think about it a little bit. And in this case, I realized: Yeah, that's right. And then I got a better handle on this and that.* [Aged 12-16 years]

Regarding optimal promotion, almost all participants (P2, P3, P4, P5, and P6) expected video-based advertising to be easy to find on social networks, especially YouTube and TikTok. One expectation was that the WBI would be promoted by social media influencers (P1, P2, P3, and P6) or by a therapist (P6).

Expectations About Barriers to Using a Self-Guided WBI for GD

Participants mentioned several possible barriers if they were to actually use a self-guided WBI for GD: negative framing, lack of differentiation, no reflection of the reality of the user's life, use of gamification elements, and focus only on digital help.

Negative framing was seen as a barrier by all participants. For example, some participants found it critical to confront users with a negative assessment of their gaming behavior. The use of the term *addiction* was also viewed critically by all participants. Some felt that being labeled as addicts could be perceived as an attack or devaluation. P1 found it problematic to title the WBI using the term *addiction* as some of those affected were not yet aware that they had a problem and, therefore, would not feel addressed by it. Therefore, a neutral or solution-oriented title such as "media training" or "screen time reduction training" was suggested. The negative framing

XSL•FO
RenderX

barrier shows that, from the perspective of potential users, a self-guided WBI should avoid echoing the familiar restrictions and prohibitions that young users already hear all the time, for example, from their parents:

> *When you google something like that, the first thing you find is always the parents, who...say: listen to your father, and this and that...It's the same thing you hear from your parents. It's just annoying...Let people talk who have already been through it, let them talk, that is much more pleasant and refreshing.* [P6; aged 12-16 years]

A lack of differentiation was also seen as a barrier (P2, P3, P4, P5, and P6). This included, for example, too much standardization of the WBI content, such as presenting a one-size-fits-all solution plan instead of responding individually to users, or focusing solely on the goal of stopping gaming and demonizing gaming across the board. A boring design that repeatedly presented content in the same manner was also considered a barrier.

Another barrier mentioned was the WBI not reflecting the reality of the user's life (P1, P2, P3, P5, and P6), for example, by mainly reflecting the perspective of nongamers, parents, or scientists. It would also be considered problematic if the intervention was mainly based on textual content and if it could only be used on nonmobile devices.

In contrast to the barriers described previously, the barrier of using gamification elements stood out. One might think that gamification elements would be seen as helpful to connect with the users' world. However, some participants found them problematic (P1, P3, and P6). At the same time (eg, regarding the dopamine hypothesis), gamification elements are a good way to make a WBI attractive. This shows that gamification elements were seen as both beneficial and risky. P6 felt that the use of educational games in a WBI could again lead to GD:

> *That sounds stupid. I don't know, I can't really imagine a game like that...So I play this [educational] game. But I would end up playing again. I would tell myself that I'm playing less, but in reality, I'm playing again. So I'd have invested 2 hours in this [educational] game, whereas I used to invest 6 hours in the real game. So, 4 hours in this game [the real game] and 2 hours in that game [the educational game]. And then what? Well, it's not much use.* [P6; aged 12-16 years]

Finally, focusing only on a self-guided WBI was also seen as a barrier to addressing GD symptoms (P2 and P4). For example, instead of offering only an online chat with other users affected by GD, the user should be guided into conversations with people outside the internet.

Discussion

Principal Findings

Overview

This qualitative study took initial exploratory steps to identify user needs regarding the characteristics of a self-guided WBI for German adolescents and young adults with GD. Changing a GD or any other harmful behavior is a major challenge. Therefore, ambivalence toward change is a key issue that any intervention should address, especially a self-guided WBI that can only provide standardized, "fixed" content. The following ambivalent treatment needs represent, at a further level of abstraction, a condensed version of the ambivalences that were presented in the Results section. We then discuss the implications of ambivalence for the design of a WBI for GD. The results indicate that, although a self-guided WBI for GD can be considered a low-threshold treatment option, its development is challenging. A standardized, self-guided WBI must find ways to address the complexity of young people's experiences of GD and their internal ambivalences during a change process. The content of the WBI, as well as its structure, storyline, protagonists, tasks, and language, must be adapted to the fact that facing and changing a possible GD can be associated with internal struggles for the user.

Ambivalent Treatment Needs

"Be Entertaining but Not Triggering"

All participants expressed the need for a WBI to be attractively designed. This finding is consistent with the observations by Bosworth et al [54], who used the System Usability Scale [55] and qualitative interviews to explore user requirements and design preferences during the development of a mobile health app for adolescents. For example, they found that "adolescents prefer vibrant colors, modern, easy-to-use interface, gamification and rewards, customizable and personalized, simple and mature graphics. Adolescents were especially motivated by gamification techniques in maintaining their interest in the application and their health behavior goals" [54]. Findings such as these are the reason gamification elements were integrated into the design of the Breaking the Game WBI. However, the results of our study showed that some participants were concerned that gamification elements in the WBI would trigger GD symptoms. In this respect, the results of our study contrast with those of the study by Bosworth et al [54], who generalize their findings in a way that does not reflect adolescents with GD. The results of the study presented in this paper draw attention to the risk of a "too attractive" design in the development of self-guided WBIs for GD. This has not been thoroughly reflected in previous research on WBI development as the WBI being too attractive is a specific problem for people with GD or other IUDs, and there is still a lack of research literature reflecting on the development of a self-guided WBI for GD, especially with regard to design. Studies of web-based treatment options for people with GD or other IUDs still focus too much on the therapeutic concept and technological advantages (eg, accessibility and anonymity) rather than on the way in which the treatment option is presented [26].

In this context, it is also necessary to consider users' beliefs about their problem and possible solutions that might influence the use of a WBI. For example, the idea that alternatives to a game should be as rewarding as the game itself was mentioned by several participants in this study. From the research team's perspective, a self-guided WBI should, while nicely designed as a virtual place to return to, also address the question of

XSL•FO
RenderX

whether it is helpful to constantly seek reward and comfort. The dopamine hypothesis should be addressed in the WBI, for example, by informing users that the brain needs time to adjust to lower levels of dopamine and that it may be necessary to initially forgo the "high" in alternative activities.

"I Want to Change, but Change Means Stress, and I Cannot Handle Stress"

The data from this study showed that gaming, even symptoms of a GD, could meet important emotional self-regulation needs, such as stress management or generating positive emotional states. This is consistent with the findings of Rollo et al [56] that internet use can meet important needs of young people that are not met in the offline world. It is also consistent with the findings of other studies linking the occurrence and persistence of IUD or GD symptoms to deficits in emotion regulation [57,58]. This means that using a WBI for GD may feel risky. Young people with GD may be used to suppressing feelings, avoiding direct confrontation with problems [59], and regulating unpleasant feelings through gaming [60]. Therefore, a self-guided WBI for GD must be designed to encourage users to reflect on the needs they are satisfying through gaming. In addition, users should be provided with alternative emotion regulation strategies and shown new ways of satisfying the needs that are currently mainly met through computer games. This is in line with the work by Gurdal et al [61], who concluded that adolescents with GD find it helpful to understand the reasons for their problematic gaming. This also means that IUD or GD interventions should not focus solely on controlling or reducing internet use [56].

However, there is a gap in knowledge on how to design a WBI for GD to help users who are stressed with their change process. Guided WBIs for GD and IUD already include change-sensitive techniques, such as motivational interviewing in their concept [11], but there is still a lot to learn regarding self-guided WBIs for GD because there is no counselor or therapist to apply motivational interviewing techniques.

"Help Me With My Gaming Problem but Don't Call Me an Addict"

The comments of some study participants showed that too much pressure was a significant barrier to changing problematic gaming behavior. From their perspective, a self-guided WBI should not negatively frame the user's gaming behavior, for example, by calling it an "addiction." It should also avoid creating too much pressure by repeating parental restrictions. At the same time, the lack of pressure seemed to be an equally important barrier in addressing GD symptoms.

For the development of a self-guided WBI, this means finding a good balance between demonstrating the negative consequences of GD symptoms and simultaneously avoiding high demands and rigidly formulated strategies. Instead, the WBI could make it transparent to the user that a change in gaming behavior may cause temporary stress and provide coping strategies. The WBI should not say that people *are* addicts but that they *have* an addiction—and they can cope with it, even if it is hard to imagine at first. Stigmatization or self-stigmatization can have a negative impact on mental health patients' self-efficacy and adherence to treatment [62,63]. Therefore, the importance of wording is significant in the development of a self-guided WBI. As logging out is only a click away, balanced language is critical.

"The Game Is Not the Real Problem, My Life Is"

The data from this study showed that life transitions, such as leaving home, were experienced as challenging, which, in turn, increased the risk of developing a GD. A self-guided WBI for GD should take this into account—in addition to their GD, users may be experiencing a general life crisis. Therefore, it is not enough to focus on GD alone. The WBI should also address other problems in the user's life, such as stress at school, university, and work, as well as social problems. This is in line with the study by Rollo et al [56], which explored the needs of adolescents in relation to the prevention of problematic internet use. From their perspective, interventions should not be limited to the level of the individual but should take a broader perspective by reflecting on the social environment [56].

"I Resent External Control, but I Also Need It"

The results of this study suggest that an intervention that lacks freedom of choice and autonomy may discourage coping with GD symptoms. However, at the same time, external control, for example, from parents, worked as a helpful strategy for participants, preventing escalation of gaming behavior for many years. Technology also seemed to provide external control. For example, app blockers were mentioned as helpful for several participants. Does this mean that technology is allowed to control gaming behavior, whereas people are not? The key point seems to be that it was the participants' choice to allow external control in the form of technology into their lives, whereas parental control was not a choice but was imposed on them. This suggests that the need for external control does not necessarily contradict the need for autonomy as long as it is based on free choice. However, age-related differences should be taken into account. From a developmental perspective, it can be assumed that adolescents are less able to control their gaming than young adults [64]. At the same time, adolescents may be driven by a strong need for autonomy.

For the development of a self-guided WBI, this means finding the right amount of guidance while considering different age-related autonomy needs. The WBI should allow for autonomy in its use while providing a clear structure and social support, such as a forum. Language and wording should address the target group in a way that does not patronize them but invites them to participate. Control tools, such as tracking gaming behavior, should be offered in a positive and inviting way.

Integrating Ambivalence Into WBI Design

Ambivalence about changing a GD (or any other problem behavior) is very normal. It is also a significant challenge in offline face-to-face therapy. However, as a self-guided WBI cannot respond to the user's ambivalence as flexibly as a human therapist can, it must find other ways to accommodate the user's ambivalence.

One basic attitude that a self-guided WBI can adopt is to communicate it openly to the user, such as in the following way: "We recognize that certain aspects may make it difficult for you to change. You may sometimes feel ambivalent and ask yourself:

XSL•FO
RenderX

Do I really want to change my gaming? While another part of you may actually want to change. We believe that change is possible, and we want to help you get there." To be successful, a self-guided WBI, just like a face-to-face therapy, must be able to oscillate between the ambivalent poles of the target group.

The technique of motivational interviewing can be considered the gold standard in addiction treatment and has been used in IUD WBIs with human therapists with the specific goal of resolving ambivalence and building motivation to change [11,12]. A self-guided WBI may not be able to replace face-to-face motivational interviewing. However, it can incorporate design elements based on motivational interviewing principles to promote user compliance.

Accepting and Expanding the Boundaries of Self-Guided WBIs

As mentioned previously, for some people with GD symptoms, a self-guided WBI may not be sufficient. In some cases, it may be better to seek more individualized help to address the underlying problem for which gaming may be a coping strategy. Some of the study participants expressed a clear need for a more personalized approach, suggesting that the WBI should include counseling by a human therapist.

However, studies show that a self-guided WBI can be designed to strengthen the bond between the user and the treatment program. In a scoping review, Scholten et al [33] investigated how embodied conversational agents can support users of self-guided WBIs and contribute to adherence. Therefore, the effectiveness of artificial intelligence, such as conversational chatbots and dialogue systems, is currently the subject of intense research [65]. Holter et al [66] investigated the person-to-program alliance between users and self-guided WBIs. The authors found 3 typical relational patterns. They showed that some of the interviewed participants developed semisocial interactions and even semisocial relationships with the digital program [66].

Limitations and Perspectives

This study has limitations that must be noted.

The results are based on a small sample of 6 young gamers. As this study revealed heterogeneous and sometimes contradictory user needs, it was sufficient to provide guidance for the development of the WBI Breaking the Game for a German target population. It revealed a variety of experiences and perspectives of young people. The research group was successful in integrating different age groups and educational backgrounds. The triangulation of different data collection methods depending on the age group (a focus group with the older participants and individual interviews with the younger participants) contributed to the quality of the data. However, the sample of this study could only provide an initial basis for exploring and identifying relevant aspects that could be built upon in a larger study (eg, to derive quantitative items). Follow-up studies are needed to deepen and generalize the initial findings of our study.

In addition to the limited sample size, it is critical to note that the sample consisted entirely of male participants, although studies have shown that young women can also be affected by GD [67,68]. A German prevalence study showed that boys were twice as likely as girls to be affected by GD in 2023 [7]. In our view, this reflects the limited perspective of the research field and the continuing gap in care in Germany, which is reflected in the fact that women with GD continue to be significantly less represented in the health care system than men. It is possible that only male individuals were recruited because of the focus on the health care system. The support group consisted of male participants only. Therefore, the sample of this study should be viewed critically as it reproduces the research gap on GD in women. It may also have been an obstacle that SS, as a male researcher, recruited participants in the outpatient clinic. This may have made it less likely for female participants to take part. This limitation of our study demonstrates the importance of sensitivity to diversity but also diversity among researchers themselves to generate findings that reflect the diversity of reality. It would be important for researchers to be more aware of their field affiliation and actively seek participants outside their field.

Another limitation relates to the research design and the level of participation. During the analysis, it was interesting to observe that some of the participants' statements were particularly irritating to the members of the research team responsible for developing the WBI. "What did he mean by that?" was a common question. This suggests that additional participatory elements would have been useful. For example, an additional group discussion or member-checking workshop could have been conducted with the participants after the analysis, with questions played back to the participants to validate or clarify their meaning. We believe that a strength of this study is that, by including the end-user perspective in the development of the Breaking the Game WBI, the project was in line with the World Health Organization's advocacy for collaborative and responsible development of youth-centered digital health interventions. The World Health Organization considers it as a no-go to involve young people "only at the end of the process or only at one stage of the process" [69]. However, one might ask what other collaborative formats might be possible besides qualitative focus groups and interviews. This is, of course, also a question of resources, which would need to be planned for at the outset. The World Health Organization framework states that a part of the process should involve young people as cocreators. "They are the experts on what health information young people need and what technology young people are using" [69]. Their involvement "can range from having several young people engaged in every aspect of planning, development and implementation to having a youth advisory board to consult throughout the process" [69].

The financial and time constraints of research projects may make this difficult to achieve. Nevertheless, a more intensive involvement of end users in the development of the self-guided WBI would have been instructive and should be considered in future research projects.

Conclusions

Given the diversity of expectations expressed by a group of 6 young people, it can be assumed that it is impossible to design a self-guided WBI that meets all user needs. However, it must

XSL•FO
RenderX

not be the goal to map user needs 1:1 in a WBI. On the contrary, ambivalence of needs in the face of a possible GD should be seen as a normal part of a change process and, therefore, actively integrated into the WBI concept and storyline. A self-guided WBI for GD should make the consideration and management of conflicting user needs its main concept—this is what the reality of people with GD is all about. In the development of the Breaking the Game WBI, for example, this was taken into account by having the protagonist himself oscillate between the motivation to change and the urge to return to the game in stressful situations. This is not presented as a problem but as his personal way of getting to know himself better, which turns out to be the basis for a sustainable change toward healthier internet use.

Acknowledgments

This study is part of the Onlinebrücke project, which was funded by the Ministry of Social Affairs, Health, and Integration of the State of Baden-Württemberg, Germany. The authors also acknowledge the support of the Open Access Publication Fund of the Eberhard Karls University of Tübingen and the support of the German Center for Mental Health (01EE2306B). They would also like to thank Dr Christine Preiser and the participants of the Qualitative Methods Research Workshop at University Hospital Tübingen for their support in the analysis of the interview data and their repeated feedback on the study design. Finally, the authors would like to thank the study participants for providing rich insights into their personal experiences.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

IB, BLG, SS, KUP, AS, CL, and SH were involved in study design and interview guide development; IB and SS were involved in participant recruitment; and SS and KUP were involved in data collection. BLG, IB, KUP, AS, CL, and SH were involved in data analysis. BLG wrote the initial draft of this paper. SH was involved in writing the original draft, and IB and KUP were involved in the initial review and editing. All other authors were involved in subsequent review and editing. All authors read and approved the published version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
COREQ checklist.
[PDF File (Adobe PDF File), 187 KB - formative_v9i1e63258_app1.pdf]

Multimedia Appendix 2
Coding tree.
[DOCX File , 15 KB - formative_v9i1e63258_app2.docx]

Multimedia Appendix 3
Theme matrix example.
[DOCX File , 21 KB - formative_v9i1e63258_app3.docx]

References

1. 6C51 Gaming disorder: ICD-11 for mortality and morbidity statistics. World Health Organization. 2024 Mar. URL: https://icd.who.int/browse/2024-01/mms/en#1448597234 [accessed 2024-05-08]
2. Dreier M, Wölfling K. The treatment of computer gaming and internet use disorder in times of digitalized addiction treatment. In: Beiglböck W, Gottwald-Nathaniel G, Preinsperger W, Scheibenbogen O, editors. Addiction Treatment and Digitalization: Addiction Prevention and Addiction Therapy Between Human Encounters and Virtual Reality. Berlin, Heidelberg: Springer; 2023:159-176.
3. Han TS, Cho H, Sung D, Park M. A systematic review of the impact of COVID-19 on the game addiction of children and adolescents. Front Psychiatry 2022 Aug 18;13:976601 [FREE Full text] [doi: 10.3389/fpsyt.2022.976601] [Medline: 36061296]
4. Montag C, Pontes HM, Kannen C, Rozgonjuk D, Brandt D, Bischof A, et al. Examining the interplay between internet use disorder tendencies and well-being in relation to sofalizing during the COVID-19 pandemic. Compr Psychiatry 2024 Apr;130:152452 [FREE Full text] [doi: 10.1016/j.comppsych.2024.152452] [Medline: 38290293]

XSL•FO
RenderX

5. Pape M, Geisler BL, Cornelsen L, Bottel L, Te Wildt BT, Dreier M, et al. A short-term manual for webcam-based telemedicine treatment of internet use disorders. Front Psychiatry 2023 Feb 23;14:1053930 [FREE Full text] [doi: 10.3389/fpsyt.2023.1053930] [Medline: 36911137]
6. Mediensucht 2020 - Gaming und Social Media in Zeiten von Corona. DAK-Längsschnittstudie: Befragung von Kindern, Jugendlichen (12-17 Jahre) und deren Eltern. DAK-Gesundheit. 2020. URL: https://caas.content.dak.de/caas/v1/media/12632/data/e364341b499ec01105a44cdd5eed6f97/dak-studie-mediensucht-zwischenergebnisse.pdf [accessed 2025-03-02]
7. Problematische Mediennutzung im Kindes- und Jugendalter in der post-pandemischen phase: Ergebnisbericht 2023. DAK-Gesundheit. 2024. URL: https://caas.content.dak.de/caas/v1/media/57130/data/1e99c74f76c82194594692bcf2e2337d/dzskj-dak-mediensuchtstudie-2023-24-ergebnisbericht.pdf [accessed 2024-05-08]
8. Horvath KJ, Ecklund AM, Hunt SL, Nelson TF, Toomey TL. Developing Internet-based health interventions: a guide for public health researchers and practitioners. J Med Internet Res 2015 Jan 23;17(1):e28 [FREE Full text] [doi: 10.2196/jmir.3770] [Medline: 25650702]
9. Bücker L, Gehlenborg J, Moritz S, Westermann S. A randomized controlled trial on a self-guided internet-based intervention for gambling problems. Sci Rep 2021 Jun 22;11(1):13033 [FREE Full text] [doi: 10.1038/s41598-021-92242-8] [Medline: 34158553]
10. Baumeister H, Reichler L, Munzinger M, Lin J. The impact of guidance on Internet-based mental health interventions — a systematic review. Internet Interv 2014 Oct;1(4):205-215. [doi: 10.1016/j.invent.2014.08.003]
11. Dieris-Hirche J, Bottel L, Basten J, Pape M, Timmesfeld N, Te Wildt BT, OMPRIS Study Group. Efficacy of a short-term webcam-based telemedicine treatment of internet use disorders (OMPRIS): a multicentre, prospective, single-blind, randomised, clinical trial. EClinicalMedicine 2023 Oct;64:102216 [FREE Full text] [doi: 10.1016/j.eclinm.2023.102216] [Medline: 37745023]
12. Bottel L, Brand M, Dieris-Hirche J, Herpertz S, Timmesfeld N, Te Wildt BT. Efficacy of short-term telemedicine motivation-based intervention for individuals with internet use disorder - a pilot-study. J Behav Addict 2021 Nov 17;10(4):1005-1014 [FREE Full text] [doi: 10.1556/2006.2021.00071] [Medline: 34797218]
13. Grüning DJ, Riedel F, Lorenz-Spreen P. Directing smartphone use through the self-nudge app one sec. Proc Natl Acad Sci U S A 2023 Feb 21;120(8):e2213114120 [FREE Full text] [doi: 10.1073/pnas.2213114120] [Medline: 36795756]
14. Bischof A, Brandt D, Schlossarek S, Vens M, Rozgonjuk D, Wernicke J, et al. Study protocol for a randomised controlled trial of an e-health stepped care approach for the treatment of internet use disorders versus a placebo condition: the SCAPIT study. BMJ Open 2022 Nov 02;12(11):e061453 [FREE Full text] [doi: 10.1136/bmjopen-2022-061453] [Medline: 36323482]
15. Jeon M, Lee MS, Yoon JY, Bhang SY. Mental health literacy of internet gaming disorder and problematic smartphone use among Korean teenagers. PLoS One 2022 Jul 21;17(7):e0270988 [FREE Full text] [doi: 10.1371/journal.pone.0270988] [Medline: 35862446]
16. Granero R, Fernández-Aranda F, Castro-Calvo J, Billieux J, Valero-Solís S, Mora-Maltas B, et al. Subtyping treatment-seeking gaming disorder patients. Addict Behav 2021 Dec;123:107086 [FREE Full text] [doi: 10.1016/j.addbeh.2021.107086] [Medline: 34450349]
17. Torres-Rodríguez A, Griffiths MD, Carbonell X, Oberst U. Internet gaming disorder in adolescence: psychological characteristics of a clinical sample. J Behav Addict 2018 Sep 01;7(3):707-718 [FREE Full text] [doi: 10.1556/2006.7.2018.75] [Medline: 30264606]
18. Gao YX, Wang JY, Dong GH. The prevalence and possible risk factors of internet gaming disorder among adolescents and young adults: systematic reviews and meta-analyses. J Psychiatr Res 2022 Oct;154:35-43. [doi: 10.1016/j.jpsychires.2022.06.049] [Medline: 35926424]
19. Schneider LA, King DL, Delfabbro PH. Family factors in adolescent problematic internet gaming: a systematic review. J Behav Addict 2017 Sep 01;6(3):321-333 [FREE Full text] [doi: 10.1556/2006.6.2017.035] [Medline: 28762279]
20. Brandhorst I, Renner T, Barth GM. Parental factors in internet and computer game addiction in adolescence: an overview. Z Kinder Jugendpsychiatr Psychother 2021 Jan 01;50(1):37-50 [FREE Full text] [doi: 10.1024/1422-4917/a000817] [Medline: 34238056]
21. Dong GH, Potenza MN. Considering gender differences in the study and treatment of internet gaming disorder. J Psychiatr Res 2022 Sep;153:25-29 [FREE Full text] [doi: 10.1016/j.jpsychires.2022.06.057] [Medline: 35793576]
22. André F, Broman N, Håkansson A, Claesdotter-Knutsson E. Gaming addiction, problematic gaming and engaged gaming - prevalence and associated characteristics. Addict Behav Rep 2020 Dec;12:100324 [FREE Full text] [doi: 10.1016/j.abrep.2020.100324] [Medline: 33354616]
23. Sweeney GM, Donovan CL, March S, Forbes Y. Logging into therapy: adolescent perceptions of online therapies for mental health problems. Internet Interv 2019 Mar;15:93-99 [FREE Full text] [doi: 10.1016/j.invent.2016.12.001] [Medline: 30792959]
24. Howes ST, Gorey KM, Charron CM. Relative effectiveness of online cognitive behavioural therapy with anxious or depressed young people: rapid review and meta-analysis. Aust Soc Work 2021 Nov 29;76(4):547-561 [FREE Full text] [doi: 10.1080/0312407X.2021.2001832]

XSL•FO
RenderX

25. Domhardt M, Ebert D, Baumeister H. Internet and mobile-based interventions in children and adolescents. In: Fegert J, Eggers C, Resch F, editors. Psychiatry and Psychotherapy of Children and Adolescents. Berlin, Germany: Springer; 2021:1-12.
26. Gorowska M, Tokarska K, Zhou X, Gola MK, Li Y. Novel approaches for treating internet gaming disorder: a review of technology-based interventions. Compr Psychiatry 2022 May;115:152312 [FREE Full text] [doi: 10.1016/j.comppsych.2022.152312] [Medline: 35405548]
27. Balhara YP, Sarkar S, Laspal N, Bhargava R, Yadav Z. A randomized controlled trial to assess effectiveness of GamE- an e-Health intervention to self-manage gaming with an aim to prevent gaming disorder. Asian J Psychiatr 2023 Feb;80:103389 [FREE Full text] [doi: 10.1016/j.ajp.2022.103389] [Medline: 36516699]
28. Sioni SR, Burleson MH, Bekerian DA. Internet gaming disorder: social phobia and identifying with your virtual self. Comput Human Behav 2017 Jun;71:11-15. [doi: 10.1016/j.chb.2017.01.044]
29. Gulliver A, Calear AL, Sunderland M, Kay-Lambkin F, Farrer LM, Banfield M, et al. Consumer-guided development of an engagement-facilitation intervention for increasing uptake and adherence for self-guided web-based mental health programs: focus groups and online evaluation survey. JMIR Form Res 2020 Oct 29;4(10):e22528 [FREE Full text] [doi: 10.2196/22528] [Medline: 33118939]
30. Karyotaki E, Riper H, Twisk J, Hoogendoorn A, Kleiboer A, Mira A, et al. Efficacy of self-guided internet-based cognitive behavioral therapy in the treatment of depressive symptoms: a meta-analysis of individual participant data. JAMA Psychiatry 2017 Apr 01;74(4):351-359 [FREE Full text] [doi: 10.1001/jamapsychiatry.2017.0044] [Medline: 28241179]
31. Khan K, Hall CL, Davies EB, Hollis C, Glazebrook C. The effectiveness of web-based interventions delivered to children and young people with neurodevelopmental disorders: systematic review and meta-analysis. J Med Internet Res 2019 Nov 01;21(11):e13478 [FREE Full text] [doi: 10.2196/13478] [Medline: 31682573]
32. Rogers MA, Lemmen K, Kramer R, Mann J, Chopra V. Internet-delivered health interventions that work: systematic review of meta-analyses and evaluation of website availability. J Med Internet Res 2017 Mar 24;19(3):e90 [FREE Full text] [doi: 10.2196/jmir.7111] [Medline: 28341617]
33. Scholten MR, Kelders SM, Van Gemert-Pijnen JE. Self-guided web-based interventions: scoping review on user needs and the potential of embodied conversational agents to address them. J Med Internet Res 2017 Nov 16;19(11):e383 [FREE Full text] [doi: 10.2196/jmir.7351] [Medline: 29146567]
34. Andersson G, Cuijpers P, Carlbring P, Riper H, Hedman E. Guided Internet-based vs. face-to-face cognitive behavior therapy for psychiatric and somatic disorders: a systematic review and meta-analysis. World Psychiatry 2014 Oct 01;13(3):288-295 [FREE Full text] [doi: 10.1002/wps.20151] [Medline: 25273302]
35. Carlbring P, Andersson G, Cuijpers P, Riper H, Hedman-Lagerlöf E. Internet-based vs. face-to-face cognitive behavior therapy for psychiatric and somatic disorders: an updated systematic review and meta-analysis. Cogn Behav Ther 2018 Jan;47(1):1-18 [FREE Full text] [doi: 10.1080/16506073.2017.1401115] [Medline: 29215315]
36. Talbot F. Client contact in self-help therapy for anxiety and depression: necessary but can take a variety of forms beside therapist contact. Behav Change 2012 Jul 06;29(2):63-76. [doi: 10.1017/bec.2012.4]
37. Goslar M, Leibetseder M, Muench HM, Hofmann SG, Laireiter A. Efficacy of face-to-face versus self-guided treatments for disordered gambling: a meta-analysis. J Behav Addict 2017 Jun 01;6(2):142-162 [FREE Full text] [doi: 10.1556/2006.6.2017.034] [Medline: 28662618]
38. Dziobek I, Lipinski S. Participatory research in clinical psychology and psychiatry in Germany – achievements, implementation, and a look to the future. Psychiatr Prax 2021 Oct 05;48(7):337-340 [FREE Full text] [doi: 10.1055/a-1614-0719] [Medline: 34610643]
39. Levitan B, Getz K, Eisenstein EL, Goldberg M, Harker M, Hesterlee S, et al. Assessing the financial value of patient engagement: a quantitative approach from CTTI's patient groups and clinical trials project. Ther Innov Regul Sci 2018 Mar 30;52(2):220-229 [FREE Full text] [doi: 10.1177/2168479017716715] [Medline: 29714515]
40. von Peter S. Participatory and collaborative strategies in psychiatric research. Psychiatr Prax 2017 Nov 06;44(8):431-433. [doi: 10.1055/s-0043-120241] [Medline: 29108084]
41. Chen Y, Goh K, Razak M. Development of a web-based tailored intervention for excessive gaming. In: Proceedings of the 2012 World Congress Conference on Engineering and Computer Science. 2012 Presented at: WCECS '12; October 24-26, 2012; San Francisco, CA p. 24-26 URL: https://www.iaeng.org/publication/WCECS2012/WCECS2012_pp134-139.pdf
42. Hanke S, Brecht L, Petersen K, Barth GM, Renner T, Batra A, et al. Preparation of an online training programme for the parents of adolescents and young adults with internet use disorders. Kindh Entwickl 2022 Oct;31(4):220-228. [doi: 10.1026/0942-5403/a000393]
43. Bernstein K, Zarski A, Pekarek E, Schaub MP, Berking M, Baumeister H, et al. Case report for an internet- and mobile-based intervention for internet use disorder. Front Psychiatry 2023 Jul 25;14:700520 [FREE Full text] [doi: 10.3389/fpsyt.2023.700520] [Medline: 37559920]
44. Chau C, Tsui YY, Cheng C. Gamification for internet gaming disorder prevention: evaluation of a wise IT-use (WIT) program for Hong Kong primary students. Front Psychol 2019 Nov 1;10:2468 [FREE Full text] [doi: 10.3389/fpsyg.2019.02468] [Medline: 31736842]

XSL•FO
RenderX

45. Dominguez-Rodriguez A, De La Rosa-Gómez A. A perspective on how user-centered design could improve the impact of self-applied psychological interventions in low- or middle-income countries in Latin America. Front Digit Health 2022 Jun 2;4:866155 [FREE Full text] [doi: 10.3389/fdgth.2022.866155] [Medline: 35721795]
46. Larsson Lund M, Månsson Lexell E, Nyman A. Optimising the development of sustainable internet-based occupational therapy interventions: important key actions and perspectives to consider. Scand J Occup Ther 2022 May 19;29(4):259-269 [FREE Full text] [doi: 10.1080/11038128.2021.1950206] [Medline: 34280329]
47. Kuckartz U, Rädiker S. Qualitative Content Analysis: Methods, Practice and Software. Weinheim, Germany: Juventa Verlag; 2022.
48. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
49. Helfferich C. The Quality of Qualitative Data. A Manual for Conducting Qualitative Interviews. Wiesbaden, Germany: Springer; 2011.
50. Mayring P. Qualitative Inhaltsanalyse: Grundlagen und Techniken. Weinheim, Germany: Springer; 2022.
51. Clarke V, Braun V, Hayfield N. Thematic analysis. In: Smith JA, editor. Qualitative Psychology: A Practical Guide to Research Methods. 3rd edition. Thousand Oaks, CA: Sage Publications; 2015:222-248.
52. Sabnis SV, Wolgemuth JR. Validity practices in qualitative research in school psychology. Sch Psychol Int 2023 Aug 17;45(2):87-114 [FREE Full text] [doi: 10.1177/01430343231194731]
53. ISO 9241-210:2019: ergonomics of human-system interaction: part 210: human-centred design for interactive systems. International Organization for Standardization (ISO). URL: https://www.iso.org/standard/77520.html [accessed 2024-03-03]
54. Bosworth T, Williams A, Wilson G, Flowers L. Development and design needs of mobile health (mHealth) apps for adolescents. Ann Fam Med 2023;21(Supplement 1):4225 [FREE Full text] [doi: 10.1370/afm.21.s1.4225]
55. Brooke J. SUS: a quick and dirty usability scale. ResearchGate. 1995 Nov. URL: https://www.researchgate.net/publication/228593520_SUS_A_quick_and_dirty_usability_scale [accessed 2024-06-13]
56. Rollo S, Venuleo C, Ferrante L, De Luca Picione R. What adolescents have to say about problematic internet use: a qualitative study based on focus groups. Int J Environ Res Public Health 2023 Nov 02;20(21):7013 [FREE Full text] [doi: 10.3390/ijerph20217013] [Medline: 37947570]
57. Wartberg L, Lindenberg K. Predictors of spontaneous remission of problematic internet use in adolescence: a one-year follow-up study. Int J Environ Res Public Health 2020 Jan 09;17(2):448 [FREE Full text] [doi: 10.3390/ijerph17020448] [Medline: 31936677]
58. Lin PY, Lin HC, Lin PC, Yen JY, Ko CH. The association between emotional regulation and internet gaming disorder. Psychiatry Res 2020 Jul;289:113060. [doi: 10.1016/j.psychres.2020.113060] [Medline: 32450453]
59. Yen JY, Yeh YC, Wang PW, Liu TL, Chen YY, Ko CH. Emotional regulation in young adults with internet gaming disorder. Int J Environ Res Public Health 2017 Dec 25;15(1):30 [FREE Full text] [doi: 10.3390/ijerph15010030] [Medline: 29295597]
60. Villani D, Carissoli C, Triberti S, Marchetti A, Gilli G, Riva G. Videogames for emotion regulation: a systematic review. Games Health J 2018 Apr;7(2):85-99 [FREE Full text] [doi: 10.1089/g4h.2017.0108] [Medline: 29424555]
61. Gurdal S, Kapetanovic S, Einarsson I, Boson K, Claesdotter-Knutsson E. Adolescents' perceptions of a relapse prevention treatment for problematic gaming-a qualitative study. Healthcare (Basel) 2023 Aug 22;11(17):2366 [FREE Full text] [doi: 10.3390/healthcare11172366] [Medline: 37685400]
62. Surmann M, Gruchalla LV, Falke S, Maisch B, Uhlmann C, Bock E, et al. The importance of strengthening competence and control beliefs in patients with psychosis to reduce treatment hindering self-stigmatization. Psychiatry Res 2017 Sep;255:314-320. [doi: 10.1016/j.psychres.2017.05.053] [Medline: 28601714]
63. Matthews S, Dwyer R, Snoek A. Stigma and self-stigma in addiction. J Bioeth Inq 2017 Jun 3;14(2):275-286 [FREE Full text] [doi: 10.1007/s11673-017-9784-y] [Medline: 28470503]
64. Sugaya N, Shirasaka T, Takahashi K, Kanda H. Bio-psychosocial factors of children and adolescents with internet gaming disorder: a systematic review. Biopsychosoc Med 2019 Feb 14;13(1):3 [FREE Full text] [doi: 10.1186/s13030-019-0144-5] [Medline: 30809270]
65. Bendotti H, Lawler S, Chan GC, Gartner C, Ireland D, Marshall HM. Conversational artificial intelligence interventions to support smoking cessation: a systematic review and meta-analysis. Digit Health 2023 Nov 03;9:20552076231211634 [FREE Full text] [doi: 10.1177/20552076231211634] [Medline: 37928336]
66. Holter MT, Ness O, Johansen AB, Brendryen H. Making come-alive and keeping un-alive: how people relate to self-guided web-based health interventions. Qual Health Res 2020 May 11;30(6):927-941 [FREE Full text] [doi: 10.1177/1049732320902456] [Medline: 32046613]
67. Geisler BL. Lebensbewältigung im Spiegel der Internetnutzung: Eine qualitative Biografiestudie zur problematischen Internetnutzung bei Frauen. Tübingen, Germany: Eberhard Karls Universität Tübingen: Tübingen Library Publishing; 2023.
68. Kim HS, Son G, Roh EB, Ahn WY, Kim J, Shin S, et al. Prevalence of gaming disorder: a meta-analysis. Addict Behav 2022 Mar;126:107183. [doi: 10.1016/j.addbeh.2021.107183] [Medline: 34864436]

XSL•FO
RenderX

69. Youth-centred digital health interventions: a framework for planning, developing and implementing solutions with and for young people. World Health Organization. 2020. URL: https://iris.who.int/bitstream/handle/10665/336223/9789240011717-eng.pdf [accessed 2024-05-08]

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Research
GD: gaming disorder
IUD: internet use disorder
WBI: web-based intervention

Edited by A Mavragani; submitted 14.06.24; peer-reviewed by S Meister; comments to author 20.11.24; revised version received 02.03.25; accepted 05.03.25; published 26.05.25.

Please cite as:
Geisler BL, Petersen KU, Hanke S, Schurer S, Schreiber A, Lämmle C, Batra A, Renner T, Brandhorst I
Ambivalent User Needs as a Challenge and Chance for the Design of a Web-Based Intervention for Gaming Disorder: Qualitative Interview Study With Adolescents and Young Adults
JMIR Form Res 2025;9:e63258
URL: https://formative.jmir.org/2025/1/e63258
doi:10.2196/63258
PMID:

©Birte Linny Geisler, Kay Uwe Petersen, Sara Hanke, Simon Schurer, Anne Schreiber, Christine Lämmle, Anil Batra, Tobias Renner, Isabel Brandhorst. Originally published in JMIR Formative Research (https://formative.jmir.org), 26.05.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Insights From the Nihon Housou Kyoukai's Virtual Reality–Based Social Interaction Television Program "Project Aliens" for Adolescents With Psychiatric Disorders: Single-Center Case Series Study

Junichi Fujita[1], MD, PhD; Mizuho Takayama[1], BSc, MHSW; Emi Kamono[2], MMS; Satoru Shinoda[2], PhD; Hiroyuki Yamaguchi[3], MD, PhD; Tomoko Moroga[1], BSc, RN; Mio Ishii[1,3], MD, PhD; Tomoyuki Miyazaki[4], MD, PhD

[1]Department of Child Psychiatry, Yokohama City University Hospital, Yokohama, Japan

[2]Department of Biostatistics, Graduate School of Medicine, Yokohama City University, Yokohama, Japan

[3]Department of Psychiatry, Yokohama City University School of Medicine, Yokohama, Japan

[4]Center for Research and Industry-Academia-Government Collaboration, Yokohama City University, Yokohama, Japan

Corresponding Author:
Junichi Fujita, MD, PhD
Department of Child Psychiatry
Yokohama City University Hospital
3-9, Fukuura, Kanazawa-ku
Yokohama,
Japan
Phone: 81 45 787 2800
Email: jun1182@yokohama-cu.ac.jp

Abstract

Background: Virtual reality (VR) technology is emerging as a tool in mental health care, providing a safe space for social interaction and therapeutic engagement. A social VR-based television program broadcast on Japanese public television offers a virtual environment where adolescents with mental health challenges can engage in peer support using alien avatars, reducing barriers to communication and encouraging emotional expression.

Objective: This case series aimed to document the psychological trajectories of adolescents with psychiatric disorders participating in a social VR-based television program.

Methods: A single-center case series was conducted with 3 adolescents with psychiatric disorders (aged 15, 18, and 19 years) who participated in the social VR-based television program. The study focused on examining patient-reported outcomes (PROs), including psychological measures and qualitative experiences, and clinical observations across program participation and broadcast viewing. Psychological measures, including the Japanese versions of the 3-item Short-Form University of California, Los Angeles Loneliness Scale (UCLA-LS3-J SF-3), the 14-item Resilience Scale, short form (RS-14), and the 9-item Patient Health Questionnaire (PHQ-9), were assessed at 3 time points: baseline, prebroadcast, and postbroadcast. Qualitative analysis of participant dialogue explored themes of self-disclosure, emotional expression, and social dynamics.

Results: Participants showed improvements in loneliness, resilience, and depressive symptoms after participating in the social VR-based program, as indicated by psychological measures and PROs. Qualitative analysis suggested that the structured facilitation embedded in the program enabled participants to express positive and negative emotions, promoting self-reflection and mutual support.

Conclusions: This case series suggests that structured social VR programs can provide a supportive platform for emotional exploration and psychological growth among adolescents with psychiatric disorders. The combination of avatar-based interaThis case series suggests that structured social VR-based programs can provide a supportive platform for emotional exploration and psychological growth among adolescents with psychiatric disorders. The combination of avatar-based interaction and therapeutic facilitation may offer a novel approach to engaging young people in mental health care, particularly during waiting periods for traditional psychiatric services.ction and therapeutic facilitation may offer a novel approach to engaging young people in mental health care, particularly during waiting periods for traditional psychiatric services.

XSL•FO
RenderX

(JMIR Form Res 2025;9:e74401) doi:10.2196/74401

KEYWORDS

virtual reality; social interaction; adolescents; mental health; case series; peer support; therapeutic facilitation

Introduction

Virtual reality (VR) technology has emerged as an innovative therapeutic tool in mental health care, particularly for individuals with social anxiety, depression, and related conditions [1]. Although VR-based exposure therapy and cognitive behavioral interventions have been well documented [2], VR environments are increasingly being used as therapeutic spaces providing safe gathering places for group therapy settings or social skill training for individuals with mental health challenges. However, the therapeutic potential of social VR environments—spaces where users can interact through avatars—remains largely unexplored [3,4]. In Japan, various VR-based social participation programs have been implemented to address school refusal—a significant educational challenge in which students do not or cannot attend school due to psychological, emotional, physical, or social reasons, excluding illness, financial difficulties, or COVID-19–related concerns—affecting over 240,000 students in 2021 [5].

Building on these developments, Japan's public broadcaster (Nihon Housou Kyoukai [NHK]) developed *Project Aliens*, an innovative television program that provides a VR environment where adolescents interact through alien avatars [6]. Participants communicate in a virtual space using alien avatars, which provides a psychologically safe medium for self-disclosure and fostering connections based on shared experience. Therefore, the *Project Aliens* series offers a compelling model for VR-based social experiments with clinical relevance. By lowering barriers to self-expression and creating a supportive VR-based environment, the series aims to help participants confront personal and interpersonal challenges in ways that may be less achievable in conventional therapeutic settings. This social VR-based approach aligns with broader explorations into the metaverse's potential for mental health care, where users can interact within a customizable, immersive environment that offers emotional safety and flexibility, critical for therapeutic engagement among young people. Each episode of *Project Aliens* explores a unique theme, including social stigma, family dynamics, and self-identity. Thus, it can be recommended as a tool for fostering emotional growth and social connection.

Among children and adolescents, barriers to seeking and accessing mental health services remain significant challenges, with the most prominent barriers being social stigma and individual factors, such as limited mental health knowledge [7]. Although peer support has shown promise in promoting recovery among young people with mental health conditions [8-10], traditional face-to-face interventions often face barriers related to social anxiety and stigma [11]. Social VR-based platforms may address these challenges by enabling anonymous, avatar-based interactions [12,13]. However, the psychological impact of structured, VR-based peer support programs for adolescent psychiatric patients remains poorly understood. To address this gap, this case series investigated the impact of participating in *Project Aliens* on adolescents' psychological well-being, specifically, by examining changes in loneliness, depression, resilience, and patterns of emotional expression during VR-based interactions.

Methods

Study Design

This single-center, observational case series evaluated the psychological effects of participation in *Project Aliens*, a Japanese public television program using social VR-based peer dialogue among adolescents with psychiatric disorders.

Participants meeting the inclusion and exclusion criteria were enrolled after providing consent for both program participation and study involvement. Changes in psychological indicators were assessed through data collection before recording and after broadcast.

Participants

Eligible participants were child and adolescent patients with psychiatric disorders receiving care at Yokohama City University Hospital who met the following criteria inclusion criteria:

- Participants must be aged 11-19 years at the time of consent.
- Participants must be in a stable mental health condition, as assessed by their primary care physician.
- Both the participant and their family must fully understand the purpose of the NHK program
  Project Aliens
  and agree to appear on the show.
- Written consent for participation in the study must be obtained.

The exclusion criteria were as follows:

- The presence of intellectual disabilities that may interfere with understanding the program or study
- Imminent risk of self-harm or suicide
- Suspected cases of abuse by family members
- Severe depression, as indicated by a score of ≥21 in the 9-item Patient Health Questionnaire (PHQ-9)
- Deemed unsuitable for participation based on clinical judgment by the primary care physician

Recruitment

Participants were recruited from the Department of Child Psychiatry at Yokohama City University Hospital. Between March 14 and April 12, 2024, a total of 15 adolescent patients receiving outpatient care were invited to participate. After follow-up reminders, 10 (67%) patients expressed interest and completed a program-specific questionnaire designed to assess their motivation for participation and potential for engagement in the virtual environment.

XSL•FO
RenderX

A sample size of 3 participants was prespecified based on both production requirements and methodological considerations. Although some previous episodes of *Project Aliens* had featured 4 participants, for this specific episode, the production team determined that 3 participants would be optimal to balance participant interaction with the facilitation role of the Moon Rabbit character. This decision ensured sufficient dialogue depth and narrative clarity for viewers, while maintaining appropriate pacing for the therapeutic journey.

The final selection of participants was conducted collaboratively between clinical staff and NHK producers, with priority given to 3 primary criteria, as documented in preproduction meetings: (1) participant motivation for virtual interaction rather than specific demographic attributes, (2) communication capabilities that would enable meaningful dialogue within the virtual environment, and (3) complementary personality characteristics or "interpersonal variability" that would facilitate dynamic group interaction. Although clinical considerations ensured participant safety and appropriateness for the program, the production team prioritized selecting participants who demonstrated genuine interest in connecting with others in the virtual environment.

From a methodological standpoint, this case series was conceptualized as an in-depth exploration of individual psychological trajectories rather than a study aimed at statistical generalization, making the small sample size appropriate for the detailed qualitative analysis planned. The selected participants included 1 (33%) male and 2 (67%) females, aged 15, 18, and 19 years, respectively, at the time of filming.

Program Description

This study documented the participants' experiences during their involvement in *Project Aliens*, a 45-minute late-night special program on Japanese public television. Episode 8 of the series, titled "A Healing Field Trip," was recorded on July 13-14, 2024, and broadcast on October 27, 2024. Participants joined the *Project Aliens* virtual world remotely from their homes, connecting through personal computers rather than VR headsets.

The program offered a VR-based journey through 3 distinct stages: (1) Cityscape, (2) Spaceship and Spaceport, and (3) Lunar Base, as shown in Figure 1A-C. This staged progression—from an open urban setting to an enclosed spaceship and, finally, to an expansive lunar landscape—was designed to gradually foster trust, emotional connection, and self-disclosure. In particular, the enclosed spaceship setting encouraged focused conversations and deeper engagement, whereas the openness of the moon symbolized emotional release and perspective expansion, reinforcing participants' psychological progression.

Figure 1. Sequential stages of the VR-based therapeutic journey. The figure depicts 3 sequential therapeutic environments: (A) Cityscape: designed to facilitate initial collaboration and trust formation; (B) Spaceship and Spaceport: structured for deep reflection through narrative sharing; and (C) Lunar Base: focused on personal growth and future orientation. Moon Rabbit (white figure) functions as the therapeutic facilitator, guiding structured activities and maintaining psychological safety throughout the program. This facilitation is crucial in creating a supportive environment that enables meaningful participant interaction and emotional expression. Alien 1 is named Ramune, colored green; alien 2 is named Puchi, colored purple; and alien 3 is named Maruhachi, colored yellow. VR: virtual reality.

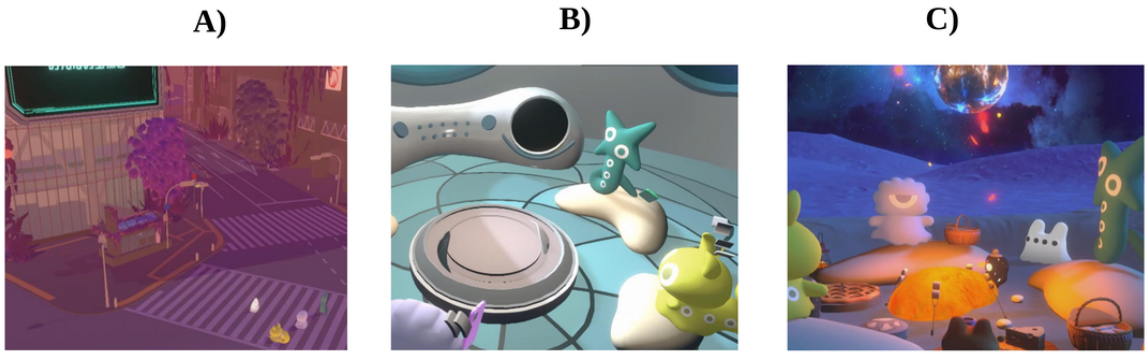

A) Cityscape, B) Spaceship, and spaceport, C) Lunar Base
Alien1) Ramune (green), Alien2) Puchi (purple), Alien3) Maruhachi (Yellow)

This approach was rooted in the belief that meaningful dialogue and the breakdown of preconceptions require a carefully structured process of interaction. Each stage was designed to not only support psychological transition but also create an environment where participants could engage in progressively deeper conversations, while feeling psychologically safe.

The virtual environment was created by an artist renowned for distinctive illustrations, 3D artwork, and youth-oriented video projects that have received national and international recognition [14]. To further support participant comfort and engagement, the program incorporated environmental audio elements, such as ambient nature sounds and underwater acoustics, aimed at fostering an immersive, calming atmosphere. These auditory cues were included to reinforce a shared sense of presence among participants and elevate their perception of a unique, transformative experience.

Throughout the program, the participants' journey was facilitated by a virtual navigator named Moon Rabbit, portrayed in real time by a professional actress. Moon Rabbit guided the 3 participants on their virtual trip to the moon. Unlike a prescripted artificial intelligence (AI) system, Moon Rabbit engaged dynamically with the participants, responding to their interactions and guiding them through each stage of the journey. Its role was carefully designed to foster psychological safety, manage psychological transitions, and encourage peer support and self-reflection. This approach was developed in

XSL•FO
RenderX

collaboration with mental health professionals to ensure that it aligned with therapeutic principles. Moon Rabbit is the white character shown in Figure 1.

Cityscape: Discovering Life Through Each Other's Treasures

The journey began in a virtual recreation of Shibuya City in Tokyo, a lively urban environment popular among young people. The participants collaborated in a treasure hunt to locate hidden "treasures" within the city. This activity encouraged teamwork and provided an initial opportunity for participants to share personal insights, building trust and familiarity within the group.

The urban setting was intentionally chosen as a familiar yet dynamic social space where participants could ease into interactions. For those from rural areas, it introduced them to a different social environment. The open nature of the cityscape provided a sense of exploration and informal engagement, helping participants gradually become comfortable in the setting before transitioning to deeper discussions.

Spaceship and Spaceport: Discovering We Are Not Alone

In this stage, the participants boarded a spaceship heading to the moon. Here, they encountered narratives from former participants, or "senior aliens," who shared their transformative journeys. The participants also shared personal photos and discussed their daily struggles, fostering deeper self-reflection and mutual understanding.

The spaceship, a deliberately enclosed environment, was designed to foster focused, structured conversations. The transition from an open cityscape to a confined space created a sense of intimacy, encouraging self-disclosure and deeper emotional engagement. At the spaceport, participants undertook the final task of the journey: writing letters to themselves. These self-reflective letters explored past challenges, present struggles, and aspirations for the future. Reading these letters aloud allowed the participants to express self-compassion and find closure. This transitional stage, just before reaching the moon, offered the participants an opportunity to consolidate their thoughts and prepare for the final phase of the journey.

Lunar Base: Sharing Painful Pasts, Stepping Toward the Future

The final stage took place on the moon, representing a safe space for introspection and transformation. The participants reflected on their experiences, shifting the focus from their past to envisioning their future. This phase highlighted the progress they had made throughout the journey.

The moon was chosen for its vast, open landscape, symbolizing emotional release and perspective expansion. In stark contrast to the confined spaceship, the moon represented liberation and psychological progression. The openness of the moon encouraged the participants to engage in forward-looking discussions, marking a symbolic transition from structured reflection to envisioning new possibilities in their lives.

Case Presentations

The following cases describe the 3 participants who engaged in the VR-based journey described before. Each participant's avatar appearance is indicated in Figure 1.

Case 1: Alien 1 (Ramune)

A 19-year-old male (Ramune) grew up in a socioeconomically disadvantaged family with significant mental health comorbidities. His father had bipolar disorder, his mother experienced depression and asthma, and his sister was diagnosed with autism spectrum disorder (ASD) and bipolar disorder. Diagnosed with ASD and a learning disability at age 8 years, the patient demonstrated communication difficulties, experienced persistent bullying, and subsequently developed severe social anxiety disorder, leading to extended periods of social withdrawal. Prolonged periods of housebound behavior further isolated him. Telemedicine during the COVID-19 pandemic enabled him to reconnect with psychiatric care, leading to improvements in his emotional stability. By the time of his participation, Ramune was attending a vocational training facility and a youth support center but remained hesitant about interpersonal interactions. Ramune is the green character in Figure 1.

Case 2: Alien 2 (Puchi)

A 15-year-old female (Puchi) presented with a history of exposure to high academic pressure. The patient developed major depressive disorder with psychotic features at age 11 years, coinciding with preparation for private school entrance examinations. Subsequently, she developed body dysmorphic disorder, exacerbated by social media exposure, which led to familial discord and suicidal behavior. Despite progress, Puchi continued to experience fatigue and emotional fragility. Puchi is the purple character in Figure 1.

Case 3: Alien 3 (Maruhachi)

An 18-year-old female (Maruhachi) presented with a history of severe academic and familial stressors. By age 15 years, her rigorous schedule led to severe weight loss, panic attacks, and eventually hospitalization for generalized anxiety disorder and an eating disorder. After 1 month of inpatient nutritional rehabilitation and antidepressant treatment, she regained physical health but continued to struggle with social isolation and reintegration into school life. Maruhachi is the yellow character in Figure 1.

Outcome Measures

Primary Outcome Measure

The primary outcome measure in this study was the longitudinal progression of each participant, assessed from baseline (study registration) through prebroadcast and postbroadcast evaluations. This case series aimed to document individual trajectories by examining changes in self-reported experiences (patient-reported outcomes [PROs]) and clinical observations.

To capture subjective experiences, PROs included self-reported loneliness, resilience, and depressive symptoms, as well as open-ended reflections on interpersonal relationships, emotional states, and coping strategies. These qualitative data were

XSL•FO
RenderX

supplemented with psychiatrist assessments, based on routine consultations and medical records, to provide a clinical perspective on participants' psychological status.

Secondary Outcome Measures

Psychological Measures

The Japanese version of the 3-item Short-Form University of California, Los Angeles Loneliness Scale (UCLA-LS3-J SF-3) was used. The UCLA-LS3-J SF-3 is a concise tool developed for the rapid assessment of loneliness. This scale has been validated in various contexts, including among mothers with infants and toddlers, demonstrating its versatility and robustness in capturing loneliness across different populations. The 3-item version is derived from the 10-item version, with a high correlation to the original UCLA Loneliness Scale version 3. It has shown adequate reliability and validity in Japanese studies, making it particularly suitable for brief interventions [15]. The total score on the UCLA-LS3-J SF-3 ranges from 3 to 9.

The 14-item Resilience Scale, short form (RS-14) is a validated tool for assessing resilience, which is conceptualized as the capacity to recover from adversity. Its reliability and validity have been established in Japanese populations, including its psychometric equivalence to the original RS. The short form reduces respondent burden, while maintaining robust internal consistency and test-retest reliability [16]. The total RS-14 score ranges from 14 to 98.

The PHQ-9 was selected because assessing treatment efficacy through PHQ-9 score changes is commonly recommended and widely accepted in clinical research, including in Japan [17]. The PHQ-9, a self-administered questionnaire comprising 9 items, evaluates the presence and severity of depressive symptoms within the past 2 weeks, based on *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition* (DSM-IV) criteria for major depressive disorder. The total PHQ-9 score ranges from 0 to 27, with higher scores indicating more severe depressive symptoms.

Assessments were conducted at 3 time points: baseline (study enrollment), interim (4-10 days pre-recording), and final (4-10 days postbroadcast). For all evaluations, participants completed the questionnaires at home on paper. The completed questionnaires were then either returned during their next outpatient visit or sent by mail. This schedule was consistently applied for all 3 measures to ensure data reliability and consistency.

Speech Data

Speech data were collected from participants during the program's metaverse-based interactions.

Data Analysis

To analyze the primary outcome, we created a structured summary table outlining the longitudinal progression of each participant from baseline (study registration) to postbroadcast. This table summarized participants' qualitative self-reports at 3 time points (baseline, prebroadcast, and postbroadcast), capturing their experiences of loneliness, emotional responses, and coping strategies. As an ad hoc analysis, a case-by-case qualitative analysis was conducted to further explore these individual trajectories. This analysis focused on participants' self-reported reflections on loneliness, emotional states, and coping strategies, based on responses to the following open-ended questions:

- During the past 10 days, have you felt lonely in your relationships with friends, family, or society?
- How did you feel at that moment?
- How did you cope with this feeling of loneliness?

Thematic patterns were categorized based on predefined domains aligned with the psychological scales UCLA-LS3-J SF-3, RS-14, and PHQ-9. The identified themes included (1) changes in perceived loneliness (eg, descriptions of social isolation or re-engagement with peers), (2) changes in resilience and shifts in coping strategies (eg, engagement in self-reflection, use of peer support, and (3) emotional responses to program participation (eg, depressive state, expressions of hope, anxiety, or relief). These thematic categories were aligned with predefined psychological constructs measured by the (1) UCLA-LS3-J SF-3 (loneliness), (2) RS-14 (resilience), and (3) PHQ-9 (depressive symptoms). This alignment was achieved by examining the semantic content of participant narratives in relation to established dimensions of the respective psychological scales. Two independent researchers performed coding, and discrepancies were resolved through discussion. In addition, clinical observations from psychiatrists' records were reviewed to assess changes in participants' social engagement and emotional regulation. The goal was to identify patterns of psychological adaptation throughout the program.

To analyze secondary outcome measures, changes in pre- and postintervention scores were calculated for the psychological measures, including the UCLA-LS3-J SF-3 (loneliness), the RS-14 (resilience), and the PHQ-9 (depressive symptoms). Descriptive statistics were used to summarize individual trajectories and group-level trends across the 3 assessment points: baseline, prebroadcast, and postbroadcast. Given the small sample size, statistical significance testing was not performed, and results were interpreted based on individual-level changes rather than inferential statistics.

Exploratory sentiment analysis of participants' utterances was conducted to capture patterns of emotional expression. Words were categorized into positive and negative terms, and their frequencies were counted to assess emotional dynamics across different stages of the program (Cityscape, Spaceship and Spaceport, Lunar Base). Initially, text segmentation and morphological analysis were attempted using MeCab, a widely used Japanese text analysis tool. However, existing sentiment dictionaries did not adequately capture the context-specific emotional expressions in this study. Therefore, a custom lexicon was developed to classify words into sentiment categories. The distribution of emotional expressions was analyzed across different program stages (Cityscape, Spaceship and Spaceport, and Lunar Base) to identify shifts in emotional dynamics. The frequency of positive and negative words was calculated, and sentiment distributions were aggregated for each scene to illustrate emotional transitions across the program's stages. The analysis was performed using R version 4.4.2 (R Foundation for Statistical Computing), with custom scripts using the *dplyr*,

XSL•FO
RenderX

ggplot2, and *stringr* packages for text processing and visualization. Sentiment classification was performed using a stepwise approach to systematically categorize words and phrases based on their emotional valence. The process involved 3 steps: removal of neutral elements, sentiment categorization, and validation and refinement.

Sentiment classification was performed using a stepwise approach to systematically categorize words and phrases based on their emotional valence. The process involved 3 steps: removal of neutral elements, sentiment categorization, and validation and refinement.

First, neutral elements, such as proper nouns (eg, place or object names), conjunctions, auxiliary verbs, and particles were identified and excluded from the analysis to focus on emotionally meaningful content. Second, the remaining words and phrases were categorized into 3 sentiment categories: positive, negative, or neutral. Positive expressions included words indicating happiness, encouragement, or social connection (eg, fun, excited, relieved), whereas negative expressions encompassed terms associated with distress, anxiety, or isolation (eg, worried, lonely, frustrated). Ambiguous words (eg, different, interesting) were assessed in context before classification. Additionally, commonly used expressions that do not convey strong emotional valence, such as acknowledgments and greetings, were also categorized as neutral. For instance, frequently used words, such as hello and indeed, appeared frequently in the dataset but lacked a clear positive or negative emotional connotation. Finally, to ensure consistency and accuracy, the primary author and co-author (MT) carefully reviewed the classification together. When ambiguous terms were encountered, they analyzed the recorded broadcast content to determine the appropriate classification. This collaborative analysis considered contextual factors, including (1) the surrounding linguistic context within the utterance; (2) the participant's tone and delivery, as observed in the broadcast; (3) the narrative arc of the participant's journey within the program; and (4) the clinical understanding of adolescent emotional expression in psychiatric contexts. For example, terms such as "challenge" and "different" were particularly context dependent, as they could indicate either positive growth opportunities or negative stressors. In these cases, the authors reviewed the complete recorded dialogue and surrounding context to determine the emotional valence. Cultural nuances and adolescent-specific expressions were given special consideration, with disagreements resolved through detailed discussion until consensus was reached.

This systematic approach to sentiment classification, though qualitative in nature, ensured that ambiguous terms were consistently categorized, while respecting the complex emotional landscape of adolescents with psychiatric disorders participating in the virtual environment.

Any discrepancies were resolved through discussion and consensus. This rigorous process ensured a standardized, reproducible approach to sentiment categorization.

All available data were analyzed without imputation. No missing responses were reported for psychological measures, as participants completed and returned all self-reported questionnaires, as instructed. For qualitative responses, minor transcription gaps were resolved through researcher consensus, ensuring accurate thematic analysis. Due to the small sample size, inferential statistical analysis was not conducted.

Ethics Approval

This study complied with the Declaration of Helsinki. Ethical approval was obtained from the Yokohama City University Ethics Committee (approval number F240600003). Detailed explanations of the study's purpose, methods, potential risks, and benefits were provided to the participants and their legal guardians. Written informed consent was obtained from all participants, and the assent of minors aged 16 years or older with sufficient capacity for independent decision-making was also obtained.

The participants and their families provided explicit consent for appearing on the NHK television program *Project Aliens* through an agreement with the broadcaster. The research team only had access to transcriptions of the publicly broadcasted content from the program and did not have access to unedited interactions within the VR space. Additionally, the study used only predefined outcome measures, including primary and secondary outcomes. Personal data were anonymized using unique identification codes, ensuring confidentiality. The correspondence table linking codes to individual identities was securely stored at the research institution and was inaccessible to unauthorized personnel.

Participants were informed of their right to withdraw from the study at any time without consequences, and that their data would be excluded unless anonymized and aggregated. These measures ensured the protection of participant privacy, confidentiality of data, and compliance with ethical standards throughout the study.

In accordance with NHK's standard practices for television production, participants received a modest honorarium as a token of appreciation for their cooperation in the program. This honorarium was not tied to their participation in the research study and was unrelated to the data collection or analysis. The amount was within the customary range for adolescent contributors in similar media contexts.

Results

Primary Outcome Measure

A case-by-case analysis of patient trajectories, based on structured evaluation, as shown in Table 1, revealed key psychological and social changes. Longitudinal changes in psychological status for the 3 adolescent participants. Data were collected at 3 time points: baseline (study registration), prebroadcast, and postbroadcast. Each participant's subjective experiences and clinical observations were categorized into 4 domains: loneliness, resilience, depression (PRO), and psychiatrist observation. Psychiatrist observations were documented during routine clinical consultations at each time point.

XSL•FO
RenderX

Table 1. Longitudinal psychological changes in *Project Aliens* participants (N=3).

Participant and time points	Loneliness (PRO[a])	Resilience (PRO)	Depression (PRO)	Psychiatrist observation
Alien 1				
Baseline	Feeling disconnected when unable to talk to online friends	Distracted self with videos and games	No specific episode	Limited social engagement
Prebroadcase	Feeling excluded when unable to join conversations	Watched videos to change their mood	No specific episode	Limited social interaction skills
Postbroadcase	No specific episode	Found interacting with various people was meaningful and enriching	No specific episode	Developed better coping strategies and became more socially engaged
Alien 2				
Baseline	Feeling abandoned and lonely when a classmate stopped interacting after a seating change	No specific episode	Feeling overwhelmed by loneliness and sadness	Emotional distress, avoidance tendencies, and negative thought patterns
Prebroadcase	No specific episode	No specific episode	Feeling anxious about participating in the program	Anxiety about social interaction
Postbroadcase	"Talking with other participants who shared similar struggles helped me realize that I was not alone."	"Expressing sadness and loneliness to others brought me relief."	Feeling warmth from kind program staff, reducing isolation	Reduction in loneliness, improvement in coping skills, and alleviation of interpersonal distrust
Alien 3				
Baseline	"After missing school, pressure from my family to return caused tension, and I also became isolated from my friends."	"Retreating to my room provided me with a sense of comfort."	"However, I still felt distressed, as if I had no place to belong."	Social isolation, depressive state, and avoidance behavior
Prebroadcase	"Continued school absence has resulted in prolonged feelings of isolation."	"I tried going outside."	"Feelings of depression and hopelessness persisted."	Despite improvements in proactive coping strategies, the depressive symptoms remained.
Postbroadcase	"I started spending time and chatting with my friends after school."	"I stopped avoiding tasks due to perfectionism and realized that taking action matters. Now, l have learned to accept my imperfections."	"At first, I struggled to connect with my new classmates. Now I feel sad to part ways with them."	Enhanced coping strategies, relief from depressive symptoms, and reduced feelings of isolation

[a]PRO: patient reported outcome.

Patient-Reported Outcomes

The participants described changes in their experiences of loneliness, emotional responses, and coping mechanisms across the 3 time points (baseline, prebroadcast, postbroadcast).

Alien 1 initially reported feeling left out in social communication. Postbroadcast, he found the experience of talking to various people meaningful and enriching. From a psychiatrist's observation, his social engagement had increased.

Alien 2 initially felt isolated and experienced sadness and loneliness. Postbroadcast, she was able to express her feelings directly to a friend, which provided relief. From a psychiatrist's observation, postbroadcast, her loneliness decreased, coping skills improved, and interpersonal distrust was alleviated.

Alien 3 initially described feeling severe loneliness due to being unable to attend school and lacking communication with family. Postbroadcast, she reported positive social interactions, noting that she had been spending more time with friends. From a psychiatrist's observation, her coping strategies improved, depressive symptoms were alleviated, and her feelings of isolation decreased.

Secondary Outcome Measures

Figure 2 demonstrates changes in individual scores and group means across 3 psychological domains:

- Loneliness (UCLA-LS3-J SF-3): Scores increased slightly for alien 1 (+2) and alien 2 (+1), whereas alien 3 remained stable. This subtle increase in perceived loneliness may reflect heightened awareness of social needs following the program's interpersonal engagement.
- Resilience (RS-14): Resilience improved across participants, with the largest increase observed in alien 3 (+24). This substantial improvement suggests enhanced psychological capacity to cope with adversity, particularly notable in alien

XSL•FO
RenderX

3, whose resilience score moved from below average to above average.

- Depression (PHQ-9): Depression scores decreased for all participants, with alien 3 showing the most significant reduction (−7). When interpreted according to established clinical severity categories (0-4: minimal; 5-9: mild; 10-14: moderate; 15-19: moderately severe; 20-27: severe), all participants showed clinically meaningful improvements. Alien 1 showed gradual improvement, while remaining in the mild category (7→6→5); alien 2 initially improved then slightly increased but stayed within the mild range (8→5→6); and alien 3 achieved the most significant change, transitioning from mild depression to minimal/none (8→5→1).

Each psychological measure exhibited different patterns of change across the 3 time points, as shown in Figure 2:

- Loneliness (UCLA-LS3-J SF-3): The group mean increased from 6.3 (SD 1.5) at registration to 7.3 (SD 1.2) prebroadcast and then remained stable at 7.3 (SD 1.5) postbroadcast.
- Resilience (RS-14): The group mean increased from 56.0 (SD 2.6) at registration to 56.7 (SD 6.0) prebroadcast, followed by a more substantial rise to 70.7 (SD 6.0) postbroadcast.
- Depression (PHQ-9): The group mean decreased from 7.7 (SD 0.6) at registration to 5.3 (SD 0.6) prebroadcast, and then further dropped to 4.0 (SD 2.6) postbroadcast.

Figure 2. Longitudinal changes in psychological outcome measures. Visualization of individual trajectories and group means across 3 time points for (A) loneliness (UCLA-LS3-J SF-3, range 3-9), (B) resilience (RS-14, range 14-98), and (C) depression (PHQ-9, range 0-27). Each participant's scores were measured at 3 time points: registration, prebroadcasting, and postbroadcasting. Alien 1 (diamond), alien 2 (square), and alien 3 (triangle) are represented by different markers. Group mean trends, indicated by × markers and dashed lines, illustrate the overall trajectory of changes in loneliness, resilience, and depressive symptoms among participants. PHQ-9: 9-item Patient Health Questionnaire; RS-14: 14-item Resilience Scale, short form; UCLA-LS3-J SF-3: 3-item Short-Form University of California, Los Angeles Loneliness Scale.

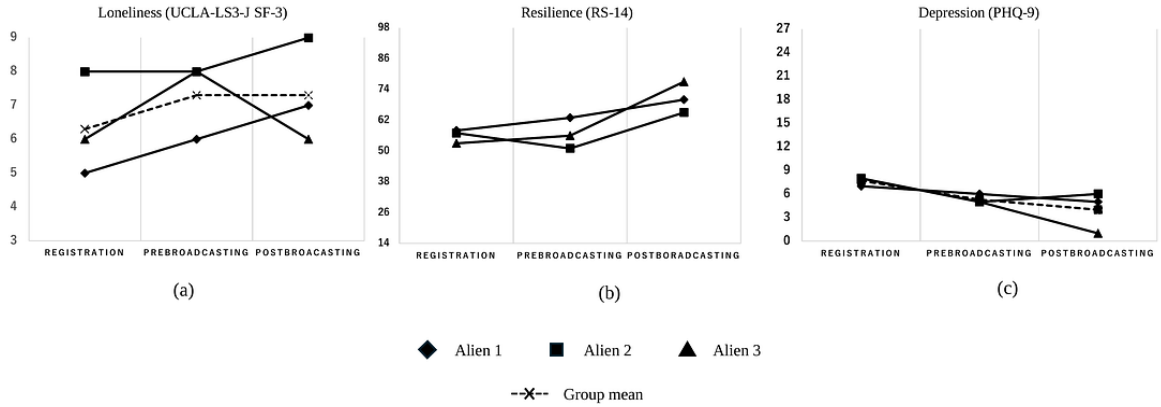

Sentiment Analysis of Speech

The sentiment analysis, conducted as a secondary outcome measure, revealed distinct emotional dynamics across the program's stages. Positive expressions dominated scene A, whereas scenes B and C reflected more balanced emotional expressions, indicating deeper emotional engagement.

Figure 3 illustrates the sentiment distribution across different scenes, providing a visual representation of how the participants' emotions varied during each scene of *Project Aliens*. Scene A contained a total of 48 emotional expressions: 18 (37.5%) negative and 30 (62.5%) positive. Scene B contained a total of 115 emotional expressions: 64 (55.7%) negative and 51 (44.3%) positive. Scene C contained a total of 68 emotional expressions: 29 (42.6%) negative and 39 (57.4%) positive.

In scene A, the participants frequently expressed positive emotions, such as excitement and curiosity, using terms such as "cute" and "fun." These findings suggest that the design of the initial VR environment successfully engaged participants and reduced their apprehension. However, as the journey progressed to scene B, more complex emotional responses emerged. Participants used words such as "painful" and "anxious" while mentioning "courage" and "challenge," indicating the difficulties they faced and their ability to confront them. By scene C, the focus shifted to closure, as reflected in expressions such as "finish" and "miss you," and optimism, as reflected in expressions such as "hope" and "wish."

XSL•FO
RenderX

Figure 3. Quantitative analysis of emotional expression patterns. Temporal distribution of emotional expressions categorized by valence (positive/negative) across program stages. Frequencies represent word-level sentiment analysis of participant utterances. Scene (A) Cityscape shows predominantly positive expressions during initial interactions. Scene (B) Spaceship/Spaceport demonstrates increased emotional complexity with both positive and negative expressions. Scene (C) Moon reflects a shift toward resolution and hope. *Word frequency counts.

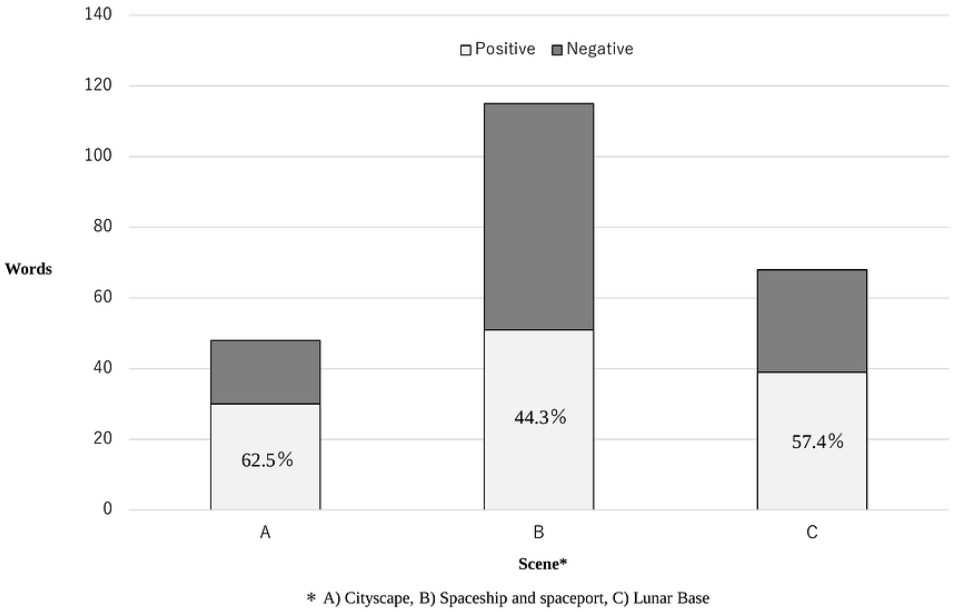

Participant Dialogue Across Scenes

To illustrate the emotional progression observed in the sentiment analysis, representative participant dialogues from each scene are provided next.

In scene A (Cityscape), the participants introduced themselves through their avatars, using them as tools for self-expression and connection.

> *Hi, I'm Ramune. I chose a light-blue color that feels like "me."* [Alien 1]
>
> *I'm Pichi. I wanted to get better at thinking before I speak, so I made my head look like a brain.* [Alien 2]
>
> *I'm Maruhachi. I put eyes in different places so I can see things from multiple perspectives.* [Alien 3]

In scene B (Spaceship and Spaceport), the participants began sharing deeper emotions, revealing vulnerabilities and finding validation in each other's experiences.

> *Hearing others talk about their struggles makes me realize I'm not the only one who feels trapped sometimes.* [Alien 3]
>
> *When I have a panic attack, I can't move or do anything. Having something comforting nearby really helps in those moments.* [Alien 1]
>
> *There's a song lyric that really speaks to me: "All the things I can't express just race past me with a face of sorrow." I tend to bottle up my emotions, but I don't have much capacity, so when it overflows, I just break down.* [Alien 2]

The final scene C (Lunar Base) reflected a shift toward self-acceptance and motivation for personal growth, with the participants expressing both lingering concerns and new resolutions.

> *I've been through a lot, but I feel like I'm finally looking forward. Maybe that's the most important thing.* [Alien 2]
>
> *I think I need to trust myself a little more.* [Alien 3]
>
> *I still worry about going back to work or school. What if my panic attacks come back?* [Alien 1]
>
> *There's no need to force change, but we should keep challenging ourselves. That's how we take the first step.* [Alien 3]

Discussion

Principal Findings

This case series examined the psychological impact of *Project Aliens*, an innovative VR-based television program on Japanese public broadcasting, demonstrating its therapeutic potential for adolescents with mental health challenges. The program provided a unique social VR-based experience, offering a safe space for emotional expression and peer support through structured facilitation and immersive virtual environments. The findings indicate potential improvements in psychological well-being, including reduced loneliness and depressive symptoms and enhanced resilience. These improvements were observed in psychological measures, as well as PROs throughout the duration of the program, from recording through broadcast viewing. Social VR-based programs, such as *Project Aliens*, which create immersive spaces for dialogue, peer support, and emotional processing, may pave the way for new mental health interventions. Within this psychologically safe VR environment, the participants could openly express positive and negative emotions while sharing their recovery journey and adverse experiences through mutual empathy [12,13].

A key factor in the program's efficacy was the structured facilitation of participant engagement. The program

XSL•FO
RenderX

design—featuring warm colors, relatable visual elements, and playful alien avatars—fostered a sense of safety and accessibility, reducing communication barriers. Although realistic avatars can enhance trust in virtual environments [18], research suggests that nonhumanoid avatars, such as animal avatars, may promote self-disclosure and intimacy by reducing self-presentation anxiety [19]. This aligns with *Project Aliens*' approach, where the alien avatars may have encouraged freer self-expression and deeper emotional dialogue. Online self-disclosure, although different from face-to-face interactions, has been recognized as a valuable preliminary step for highly anxious adolescents transitioning to offline communication [20].

This social VR-based program actively encouraged participant-led interactions, while ensuring their psychological safety. Previous studies have shown that online peer support groups enhance emotional regulation and coping strategies [21]. Additionally, structured VR-based learning environments have been found to boost engagement, self-efficacy, and emotional expression [22]. Likewise, virtual therapeutic spaces require structured facilitation to sustain meaningful interactions and maximize psychological benefits [23,24]. In *Project Aliens*, facilitation has been embedded into the program's design, helping participants build trust and develop communication skills.

The quantitative findings indicate improvements in resilience and reductions in depressive symptoms among participants. Along with these measurable changes, the participants' reflections provided valuable qualitative insights. Many reported feeling less isolated and more encouraged to face challenges positively. Although these subjective accounts were not part of formal outcome measures, they complemented the quantitative data, offering a broader understanding of the program's impact. The structured VR environment facilitated collaborative engagement, mutual support, and sharing of emotional experiences, fostering a strong sense of solidarity among participants. Although this study did not directly measure changes in social attitudes, previous studies suggest that VR experiences can enhance interpersonal engagement and emotional connections [25,26].

Notably, improvements in resilience and other PROs following the program indicate that these transformations may extend beyond the virtual environment. This process aligns with the *Proteus effect* [27], a psychological phenomenon where individuals adopt behaviors and attitudes consistent with the characteristics and expectations of their digital avatars, often unconsciously altering their self-perception based on how they believe others perceive their virtual representation. As participants engaged with their avatars in *Project Aliens*, they likely internalized aspects of their virtual identities, reinforcing self-efficacy, emotional expression, and adaptive coping strategies. These factors may have contributed to the psychological improvements observed postbroadcast.

Limitations

This study has several limitations. First, the small sample size and brief follow-up period limit the generalizability of findings and the ability to assess long-term outcomes. Additionally, the participants' subjective reflections may introduce recall bias, affecting the accuracy of reported psychological changes. Furthermore, the unique nature of this intervention—a VR-based television program—may limit its applicability to broader adolescent psychiatric populations. Additionally, the participants' responses may have been influenced by factors, such as feeling excited for participating in media, receiving attention from production staff, and feeling aware of being filmed for public broadcast. These media-related factors may have impacted psychological measures independently of the VR intervention.

All participants were VR novices, which helped minimize the influence of prior VR experience; however, individual differences in VR adaptability (eg, motion sickness, interface familiarity) may have affected engagement and the overall effectiveness of the intervention. Future studies should account for these variations by implementing standardized conditions, such as screening for VR tolerance and stratifying participants accordingly. Additionally, the participant selection process may have introduced bias. As the participants were selected by NHK producers based on thematic and logistical considerations rather than through random selection, this method could have influenced the study's findings, limiting the generalizability of results.

Future Direction

Despite these limitations, this study highlights the potential of VR-based group psychotherapy and day-care programs for young psychiatric patients. The structured facilitation and avatar-based interaction elements used in *Project Aliens* could address unresolved clinical challenges in various clinical settings.

The therapeutic framework of *Project Aliens* could potentially be adapted for conventional clinical practice through simplified implementations without the broadcasting component. Particularly valuable is the inherent structure of *Project Aliens* that allows viewers to observe social interactions between participants who maintain their anonymity. This format enables individuals with psychiatric disorders to interact through avatars while allowing observation by others. Similar to how *Project Aliens* viewers observed authentic interactions without compromising participant identities, this model could enable medical students, residents, mental health professionals, or even patient and family organizations to observe therapeutic exchanges, deepening understanding and reducing stigma toward psychiatric disorders.

Future research should use methodologically rigorous designs to isolate the specific therapeutic effects of VR elements. During our recording process, the involvement of television staff prior to filming undoubtedly provided positive stimulation for the young participants. By controlling conditions and establishing verification scenarios different from television program settings, while only changing the delivery medium, it would be possible to accurately evaluate the specific effects of the VR component. Additionally, if we could quantify neurological synchronization patterns related to empathy development and emotional regulation during avatar-mediated communication by collecting physiological data, such as electroencephalograms, heart rate variability, and oxytocin sampling, we might gain valuable

XSL•FO
RenderX

insights into the mechanisms through which VR-based social interactions facilitate therapeutic change.

Through such validation processes, social VR could be refined as a structured therapeutic tool in psychiatric care, potentially creating new therapeutic support methods to expand mental health services for young people.

Acknowledgments

We extend our gratitude to those who contributed to this study. Minori Ito assisted with participant recruitment and study progress management as a research assistant. Mikiko Yuzawa from Nouvelle Plus Inc. prepared the ethics application documents. Ryo Urabe from the Nihon Housou Kyoukai (NHK) provided critical support during the program's production process, facilitating coordination between the research team and the broadcasting staff. We also acknowledge the assistance of OpenAI's ChatGPT and Anthropic's Claude in supporting the drafting and English editing process of this manuscript. This study was funded by the Japan Science and Technology Agency (JST) under the Co-creation Opportunity Formation Support Program (COI-NEXT; FY2022–2031).

Data Availability

The datasets generated and analyzed during the study are not publicly available due to privacy protection and the scope of the research plan but are available from the corresponding author upon reasonable request.

Authors' Contributions

JF, EK, and SS conceived the study design, developed the analytical framework, designed the analysis, conducted data analysis, and interpreted the results. TM and MT collected and curated the data used in the study. JF provided essential data and developed analysis tools for the research. JF, EK, SS, MI, HY, and MT drafted and revised the manuscript.

Conflicts of Interest

TM, MT, MI, MT, HY, YM, SS, EK, and MT have no conflicts of interest. JF received research grants from KAKENHI (21K01994). JF also served as a member of an advisory board for the Seisa Yokohama Educational Counseling Center. The funding agency had no role in study design, data collection, analysis, interpretation, or manuscript submission decisions.

References

1. Fodor LA, Cote CD, Cuijpers P, Szamoskozi S, David D, Cristea IA. The effectiveness of virtual reality based interventions for symptoms of anxiety and depression: a meta-analysis. Sci Rep 2018 Jul 09;8(1):10323 [FREE Full text] [doi: 10.1038/s41598-018-28113-6] [Medline: 29985400]
2. Park MJ, Kim DJ, Lee U, Na EJ, Jeon HJ. A literature overview of virtual reality (VR) in treatment of psychiatric disorders: recent advances and limitations. Front Psychiatry 2019 Jul 19;10:505 [FREE Full text] [doi: 10.3389/fpsyt.2019.00505] [Medline: 31379623]
3. Geraets CN, van der Stouwe EC, Pot-Kolder R, Veling W. Advances in immersive virtual reality interventions for mental disorders: a new reality? Curr Opin Psychol 2021 Oct;41:40-45 [FREE Full text] [doi: 10.1016/j.copsyc.2021.02.004] [Medline: 33714892]
4. Mosher MA, Carreon AC, Craig SL, Ruhter LC. Immersive technology to teach social skills to students with autism spectrum disorder: a literature review. Rev J Autism Dev Disord 2021 Apr 30;9(3):334-350. [doi: 10.1007/s40489-021-00259-6]
5. Survey on various issues in student guidance, such as problem behavior and school absenteeism among children and students. Ministry of Education, Culture, Sports, Science and Technology. 2021. URL: https://www.mext.go.jp/a_menu/shotou/seitoshidou/1302902.htm [accessed 2025-05-19]
6. NHK launches new VR documentary series Project Aliens. ITmedia. 2022 Sep 13. URL: https://nlab.itmedia.co.jp/nl/articles/2209/13/news183.html [accessed 2025-02-09]
7. Radez J, Reardon T, Creswell C, Lawrence PJ, Evdoka-Burton G, Waite P. Why do children and adolescents (not) seek and access professional help for their mental health problems? A systematic review of quantitative and qualitative studies. Eur Child Adolesc Psychiatry 2021 Feb 21;30(2):183-211 [FREE Full text] [doi: 10.1007/s00787-019-01469-4] [Medline: 31965309]
8. Chmielowska M, Mannocci N, Tansel A, Zisman-Ilani Y. Peer support and shared decision making in open dialogue: opportunities and recommendations. Front Psychol 2022 Dec 8;13:1059412 [FREE Full text] [doi: 10.3389/fpsyg.2022.1059412] [Medline: 36571005]
9. Murphy R, Huggard L, Fitzgerald A, Hennessy E, Booth A. A systematic scoping review of peer support interventions in integrated primary youth mental health care. J Community Psychol 2024 Jan 23;52(1):154-180. [doi: 10.1002/jcop.23090] [Medline: 37740958]

XSL•FO
RenderX

10. Scanlon CL, Del Toro J, Wang M. Socially anxious science achievers: the roles of peer social support and social engagement in the relation between adolescents' social anxiety and science achievement. J Youth Adolesc 2020 May 23;49(5):1005-1016. [doi: 10.1007/s10964-020-01224-y] [Medline: 32206958]
11. Lynch H, McDonagh C, Hennessy E. Social anxiety and depression stigma among adolescents. J Affect Disord 2021 Feb 15;281:744-750 [FREE Full text] [doi: 10.1016/j.jad.2020.11.073] [Medline: 33257039]
12. Kenyon K, Kinakh V, Harrison J. Social virtual reality helps to reduce feelings of loneliness and social anxiety during the COVID-19 pandemic. Sci Rep 2023 Nov 07;13(1):19282 [FREE Full text] [doi: 10.1038/s41598-023-46494-1] [Medline: 37935718]
13. Karami B, Koushki R, Arabgol F, Rahmani M, Vahabie A. Effectiveness of virtual/augmented reality–based therapeutic interventions on individuals with autism spectrum disorder: a comprehensive meta-analysis. Front Psychiatry 2021 Jun 23;12:665326 [FREE Full text] [doi: 10.3389/fpsyt.2021.665326] [Medline: 34248702]
14. Cho H. Hikaru Cho's official website. Hikaru Cho. URL: https://www.hikarucho.com/ [accessed 2025-03-15]
15. Arimoto A, Tadaka E. Reliability and validity of Japanese versions of the UCLA loneliness scale version 3 for use among mothers with infants and toddlers: a cross-sectional study. BMC Womens Health 2019 Jul 26;19(1):105 [FREE Full text] [doi: 10.1186/s12905-019-0792-4] [Medline: 31349835]
16. Nishi D, Uehara R, Kondo M, Matsuoka Y. Reliability and validity of the Japanese version of the Resilience Scale and its short version. BMC Res Notes 2010 Nov 17;3(1):310 [FREE Full text] [doi: 10.1186/1756-0500-3-310] [Medline: 21083895]
17. Muramatsu K, Miyaoka H, Kamijima K, Muramatsu Y, Tanaka Y, Hosaka M, et al. Performance of the Japanese version of the Patient Health Questionnaire-9 (J-PHQ-9) for depression in primary care. Gen Hosp Psychiatry 2018;52:64-69. [doi: 10.1016/j.genhosppsych.2018.03.007] [Medline: 29698880]
18. Aseeri S, Interrante V. The influence of avatar representation on interpersonal communication in virtual social environments. IEEE Trans Visual Comput Graphics 2021 May;27(5):2608-2617. [doi: 10.1109/tvcg.2021.3067783]
19. Ichikawa A, Ihara K, Kawaguchi I. Investigation of how animal avatar affects users' self-disclosure and subjective responses in one-on-one interactions in VR space. 2023 Presented at: Asian HCI Symposium 2023; April 28, 2023; Indonesia p. 70-75. [doi: 10.1145/3604571.3604582]
20. Towner E, Grint J, Levy T, Blakemore S, Tomova L. Revealing the self in a digital world: a systematic review of adolescent online and offline self-disclosure. Curr Opin Psychol 2022 Jun;45:PMC10561581 [FREE Full text] [doi: 10.1016/j.copsyc.2022.101309] [Medline: 37941912]
21. McGill BC, Sansom-Daly UM, Wakefield CE, Ellis SJ, Robertson EG, Cohn RJ. Therapeutic alliance and group cohesion in an online support program for adolescent and young adult cancer survivors: lessons from "Recapture Life". J Adolesc Young Adult Oncol 2017 Dec;6(4):568-572. [doi: 10.1089/jayao.2017.0001] [Medline: 28581346]
22. Chen C, Chang S, Hwang G, Zou D. Facilitating EFL learners' active behaviors in speaking: a progressive question prompt-based peer-tutoring approach with VR contexts. Interact Learn Environ 2021 Feb 10;31(4):2268-2287. [doi: 10.1080/10494820.2021.1878232]
23. Redburn J, Hayes B. Facilitators and barriers to "positive outcomes" from cognitive–behavioral therapy, according to young people: a thematic synthesis. J Clin Psychol 2024 May 08;80(5):968-1002. [doi: 10.1002/jclp.23653] [Medline: 38328892]
24. Aebersold M, Villarruel A, Tschannen D, Valladares A, Yaksich J, Yeagley E, et al. Using a virtual environment to deliver evidence-based interventions: the facilitator's experience. JMIR Serious Games 2015 Jul 21;3(2):e5 [FREE Full text] [doi: 10.2196/games.4293] [Medline: 26199045]
25. Ichino J, Ide M, Yokoyama H, Asano H, Miyachi H, Okabe D. "I've talked without intending to": self-disclosure and reciprocity via embodied avatars. Proc ACM Hum-Comput Interact 2022 Nov 11;6(CSCW2):482. [doi: 10.1145/3555583]
26. Nikolaou A, Schwabe A, Boomgaarden H. Changing social attitudes with virtual reality: a systematic review and meta-analysis. Ann Int Commun Assoc 2022 May 14;46(1):30-61. [doi: 10.1080/23808985.2022.2064324]
27. Yee N, Bailenson J. The Proteus effect: the effect of transformed self-representation on behavior. Human Commun Res 2007 Jul;33(3):271-290. [doi: 10.1111/j.1468-2958.2007.00299.x]

Abbreviations

ASD: autism spectrum disorder
NHK: Nihon Housou Kyoukai
PHQ-9: 9-item Patient Health Questionnaire
PRO: patient-reported outcome
RS-14: 14-item Resilience Scale, short form
UCLA: University of California, Los Angeles
UCLA-LS3-J SF-3: 3-item Short-Form UCLA Loneliness Scale
VR: virtual reality

XSL•FO
RenderX

Edited by A Mavragani; submitted 07.04.25; peer-reviewed by M Adnan; comments to author 07.05.25; revised version received 12.05.25; accepted 13.05.25; published 30.05.25.

Please cite as:
Fujita J, Takayama M, Kamono E, Shinoda S, Yamaguchi H, Moroga T, Ishii M, Miyazaki T
Insights From the Nihon Housou Kyoukai's Virtual Reality–Based Social Interaction Television Program "Project Aliens" for Adolescents With Psychiatric Disorders: Single-Center Case Series Study
JMIR Form Res 2025;9:e74401
URL: https://formative.jmir.org/2025/1/e74401
doi:10.2196/74401
PMID:

©Junichi Fujita, Mizuho Takayama, Emi Kamono, Satoru Shinoda, Hiroyuki Yamaguchi, Tomoko Moroga, Mio Ishii, Tomoyuki Miyazaki. Originally published in JMIR Formative Research (https://formative.jmir.org), 30.05.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Self-Harm Awareness Training Module for School Staff: Co-Design and User Testing Study

Anne-Marie Burn[1], PhD; Hayley Gains[1], MSc; Joanna K Anderson[1], PhD

Department of Psychiatry, University of Cambridge, Cambridge, United Kingdom

Corresponding Author:
Anne-Marie Burn, PhD
Department of Psychiatry
University of Cambridge
Herchel Smith Building for Brain and Mind Sciences,
Forvie Site, Robinson Way,
Cambridge, CB2 0SZ
United Kingdom
Phone: 44 01223 336961
Email: amb278@cam.ac.uk

Abstract

Background: The increasing prevalence of self-harm among adolescents is a significant public health concern. School staff are often the first professionals to notice when a young person is self-harming and are in a unique position to intervene and offer support. However, research indicates that many school staff members feel ill-equipped and lack confidence in how to respond. Negative or dismissive responses may discourage young people from seeking further help. There is an urgent need for targeted training interventions to equip school staff with the skills and knowledge necessary to support students who self-harm.

Objective: This study aimed to co-design a self-harm awareness e-learning module for school staff in the United Kingdom.

Methods: The e-learning module design and development was guided by a person-based approach over three participatory design cycles: (1) co-design sessions with experts in mental health, self-harm, and school-based training; (2) workshops with school staff to co-design the e-learning module components and explore their views on supporting students who self-harm; and (3) user testing of the prototype and focus groups with school staff to explore acceptability and feasibility. Data were thematically analyzed using the framework method.

Results: Training content, videos, and quizzes were developed in collaboration with a panel of experts. Co-design workshops with school staff (n=11) informed the prototype module design, structure, and scripts for the training content and filmed scenarios, as well as highlighting potential barriers to and facilitators of implementation. User testing of the prototype with staff (n=20) yielded high usability ratings, demonstrating high levels of acceptability. Analysis of the qualitative user testing data generated four themes: (1) usability, (2) content and design, (3) feasibility, and (4) views on how the training improved knowledge and confidence.

Conclusions: The Supportive Response to Self-Harm e-learning module was developed to enhance school staff's knowledge and confidence in responding to self-harm. It was created with a user-centered design and a person-based approach and underpinned by psychological theory. Initial findings indicate that the training is acceptable and feasible. Further research will involve a mixed methods pilot feasibility study to assess the effectiveness of the program. This will provide the necessary evidence for a large-scale rollout in schools.

(JMIR Form Res 2025;9:e69309) doi:10.2196/69309

KEYWORDS

self-harm; schools; young people; youth; school staff; training; co-design; qualitative

XSL•FO
RenderX

Introduction

Self-Harm Among Adolescents

In the United Kingdom, self-harm is defined as "all acts of intentional self-poisoning or self-injury, irrespective of the apparent purpose" [1]. Self-harm has been widely researched as a feature of clinical populations, including those with a diagnosis of borderline personality disorder, bipolar disorder, and posttraumatic stress disorder. However, self-harm is increasingly prevalent among adolescents in the general population [2]. A survey of individuals aged 15 years in England found that 34% had deliberately hurt themselves at some point in their lifetime, an increase from 22% in 2014 and 25% in 2018 [3]. Between 10% and 20% of young people who self-harm present to hospital, whereas many other individuals sustain injuries that are classified as low risk and, therefore, do not receive hospital care [4]. Despite this, low-level self-harm behaviors are indicators of psychological distress and may require professional intervention. Young people who self-harm are at a high risk of adverse outcomes, including poor mental health in adulthood and an increased risk of suicide [5,6]. Research indicates that self-harm can be a persistent behavior. In one study, 19% of those who self-harmed in adolescence were still self-harming 10 years later [7]. This underscores the need for early detection and timely interventions to prevent long-term negative outcomes in young people who self-harm.

Self-Harm in the School Setting

The UK government's green paper highlights the essential role of schools in supporting young people's mental health and providing earlier access to support [8]. Self-harm in school settings often stems from various factors, including bullying, anxiety, and the stress associated with academic pressures, and, in some cases, might serve as an attempt to seek acceptance within a peer group [9]. A recent study identified bullying, adults not listening, being picked on by a teacher, and racism as risk factors for self-harm in school settings [10]. Despite high prevalence, there is a notable absence of policies in the United Kingdom that outline how schools should address and support students who self-harm. Consequently, it becomes the responsibility of each school to navigate this challenging issue and establish their own protocols and practices for intervention and support.

School staff are often the first adults to notice when a young person is self-harming and are in a key position to intervene and provide timely support. However, research has shown that many staff members feel ill-equipped, lacking the knowledge, confidence, and resources to effectively support students [11,12]. While many staff members are willing to help, they report a need for training to improve their knowledge and confidence. A survey of UK teachers found that awareness of self-harm was inconsistent and their responses often included shock, panic, and anxiety [13]. Such reactions may discourage young people from seeking help and urge them to conceal their behavior [14]. There is evidence suggesting that most young people who self-harm do not seek support due to concerns about receiving a negative response from those they confide in [14-16]. Furthermore, the prevalence and the impact of self-harm is often underestimated in schools and not prioritized in the curriculum despite students expressing a need for information on the topic [9].

Self-Harm Training for School Staff

Training interventions could be an effective way to increase school staff's knowledge and confidence in responding to self-harm. However, there is a lack of rigorously evaluated training programs for school staff to support students who self-harm [17]. A UK survey found that 52% of staff members recruited from 222 schools reported having received some training about self-harm, but only 22% rated the training as adequate [18]. In a smaller study of 30 teachers, 5 had encountered a student who self-harmed, but none had received any self-harm training, and all expressed a need for supervision and increasing their knowledge [19]. Self-harm training for pastoral staff has been shown to provide a safe, reflective space where staff can better understand self-harm, address their own anxieties about the topic, and promote a shift in their attitudes and beliefs [20].

Consequently, there is a significant gap in accessible and feasible training programs for schools to effectively address student self-harm. Researchers in the United Kingdom developed a 30-minute e-learning module on self-harm for teachers [21]. The module was created with input from a head teacher and 2 mental health professionals, but limited information is available on the content or the development process. Teachers' actual knowledge of self-harm measured using the Self-Injury Knowledge Questionnaire [22], significantly increased immediately after the intervention (z=9.62; P<.001; r=0.73; N=167 teachers). Perceived knowledge and confidence in addressing self-harm and speaking to young people about self-harm also showed significant improvement. However, the sustainability of these outcomes is not known. Acceptability ratings were high, with 90.7% of teachers reporting that e-learning was a good method for receiving training and that they found the training both accessible and engaging. However, as the training was developed specifically for teachers, the study does not provide evidence on the feasibility and acceptability of self-harm e-learning among nonteaching school staff, whose involvement is essential for a whole-school approach to prevention and intervention.

This Study

The aim of this study was to co-design a stand-alone e-learning training module to equip school staff with the skills to communicate supportively with students about self-harm, respond appropriately to disclosures or incidents, and encourage students to seek help. This training intervention builds on the existing Supportive Response to Self-Harm (SORTS) web-based training program and resource toolkit as part of a whole-school approach [23-25]. This previous work has highlighted the need for an e-learning module incorporating scenarios and conversation starters, which staff members could complete as part of their induction or continuing professional development (CPD) training.

This paper describes the design and development of the SORTS e-learning module using a person-based approach as a

XSL•FO
RenderX

framework [26]. The training intervention was theoretically informed and guided by three principles from existing research evidence to (1) raise awareness and understanding of self-harm among school staff, (2) improve staff confidence when responding to self-harm, and (3) encourage a whole-school approach to self-harm. Using a participatory development cycle, we co-designed the e-learning module components with an expert panel and potential end users (school staff members) to inform a rapid prototype, followed by user testing to assess its usability, acceptability, and feasibility.

Methods

Ethical Considerations

Ethics approval was granted by the University of Cambridge Department of Psychology Research Ethics Committee (reference PRE.2024.006). All participants were fully informed about the project, including the limits on confidentiality and their right to withdraw from the study at any time without giving a reason. They were informed that personal data would be kept confidential and managed in accordance with the Data Protection Act, the Research Governance Framework for Health and Social Care, and the conditions of Research Ethics Committee review. Electronic consent was obtained from each participant before taking part. All audio recordings of the focus groups were taken on an encrypted voice recorder and transcribed by a university-approved transcription company. Transcripts were deidentified and stored electronically in separate, password-protected folders that only the research team had access to. All participants were offered a £40 (US $51.93) online shopping voucher for their time. Reporting followed the COREQ (Consolidated Criteria for Reporting Qualitative Research) guidelines.

Underlying Theory and Logic Model

The theoretical underpinning of the SORTS program is the self-efficacy theory by Bandura and Adams [27], which describes self-efficacy as one's beliefs about their capabilities to plan and execute actions that are required to produce a desired outcome. Studies show that teachers with high self-efficacy are more confident and effective in their teaching and classroom management. Students also view these teachers as more competent and trustworthy [28]. Teachers' attitudes and confidence to support students' mental health may be determined by their knowledge, interests, experience, and self-efficacy [29]. The theory of change hypothesis is that completing the SORTS training program will result in school staff's improved knowledge about self-harm and increased confidence when responding to students who self-harm, which in turn will facilitate student help seeking and earlier access to support. The theory of change model (Multimedia Appendix 1) will be revised iteratively as new findings emerge.

Research Plan

We were able to draw on a previous systematic review and focus groups with young people and school staff to create the e-learning content [17,23,25]. Following Medical Research Council guidance, the e-learning module development involved appropriate users at all stages of the process [30]. In addition, the study used a person-based approach, which actively involves potential end users in intervention planning and leads to the development of guiding principles [26].

The research and development cycle included three participatory design phases: (1) consultation with an expert panel, (2) co-design workshops, and (3) user testing (the process is outlined in Figure 1). This iterative approach enabled us to assess the utility, acceptability, and feasibility of the training intervention and incorporate feedback into the design and development. At the start of the study, we invited public collaborators with a range of perspectives and roles to join an "expert" public involvement panel (subject matter experts). Qualitative co-design workshops were conducted with school staff to refine and finalize the training scenarios, gather feedback on the user interface design, and capture potential users' views on prospective acceptability and feasibility [31]. Findings from the co-design workshops informed the design and development of the prototype e-learning module. We commissioned a technical developer with an established record of creating web-based training for the English National Health Service on mental health topics. In addition, we commissioned a filmmaker who specializes in creating films about children and young people's mental health to create two filmed scenarios (1) depicting a staff member discovering a student's self-harm and (2) showing a poor response to a student's disclosure compared to a more supportive response. In the user testing phase, the prototype e-learning module was given to a sample audience to identify any usability issues, acceptability, and feasibility [32] and measure changes in knowledge and confidence.

XSL•FO
RenderX

Figure 1. Research process for the design and content of the e-learning module.

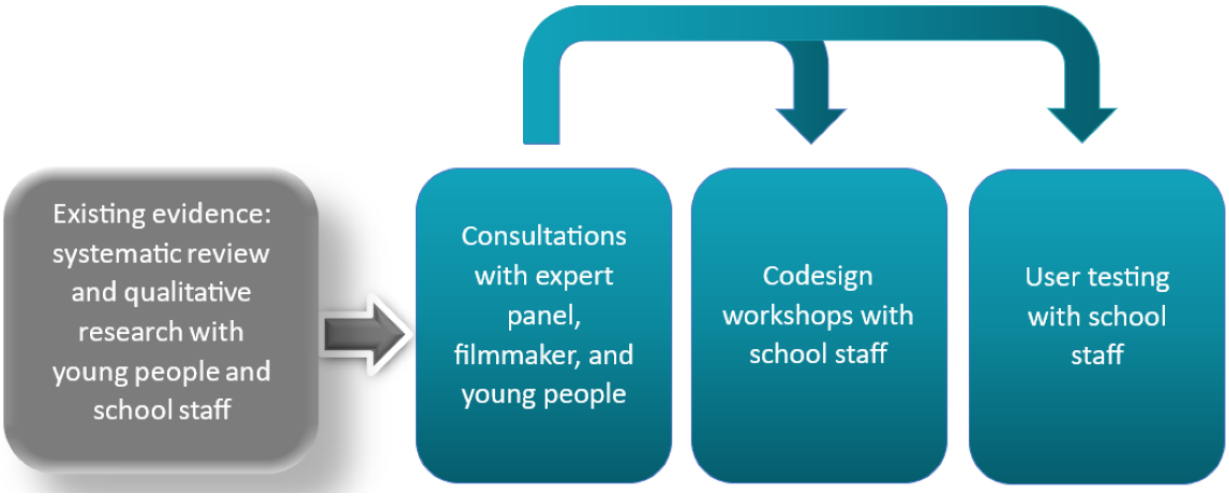

Procedures

Initial Design and Development of the e-Learning Module With Subject Matter Experts

We held a series of co-design meetings with a panel of subject matter experts to advise and develop ideas for the e-learning module training content and web elements. Panel members had expertise in young people's mental health and experience working in school settings. They included 2 senior mental health school staff members, 2 well-being leads, a teaching assistant, an intervention teacher, and a behavioral support officer. In addition, the panel included 2 self-harm trainers from a UK mental health charity that delivers self-harm awareness training for schools. Experts provided valuable insight on the training needs of school staff, the appropriate level of information about self-harm, and the challenges of managing self-harm in the school context. Co-design sessions focused on designs of mock-ups of the user interface, a wireframe structure of the module, types of questions for the interactive quizzes, draft video scripts, and storyboards depicting school-based scenarios of student-staff interactions regarding self-harm. Scenarios enable the learner to think about real-life situations as well as creating a more immersive and engaging learning environment. We worked with a filmmaker with expertise in producing mental health–related content to further refine the scripts for the filmed scenarios. We then met with a group of 3 young people (aged 18 years) to obtain their perspectives to ensure that the scripts were realistic depictions of a pupil-teacher interaction.

Participant Recruitment for Staff Workshops and User Testing

For both data collection phases, we aimed to recruit a range of school staff members working in different roles (eg, student support, administration, teaching assistants, and teachers). As far as possible, we purposefully sampled staff to ensure that a diversity of backgrounds was represented. In total, 4 state-maintained secondary schools and colleges took part in the study. We recruited schools via the team's local networks in the South and East of England. One school was situated in an urban area (London), and 3 schools were within areas that have been identified as "populations in focus" by the National Institute for Health and Care Research Applied Research Collaboration East of England due to having high health needs [33]. An email detailing information about the study, research activities, and potential involvement was sent to a key contact person (eg, mental health lead) within each school. The school contact person helped identify and approach potential participants and raise awareness of the study by distributing study information to school staff. School staff members who expressed an interest in participating were asked to contact a study researcher for further details, who then sent a participant information sheet and a consent form. Informed consent was obtained from each participant before taking part. Participants also completed a demographics form and received thank-you vouchers for their time.

Co-design Workshops With School Staff

A semistructured topic guide was developed for the staff workshops (Multimedia Appendix 2). Participants were asked to complete a preworkshop task to familiarize themselves with the e-learning training content and designs and note any other knowledge areas, skills, and techniques that they would like to incorporate into the training. We then ran in-person workshops at 2 schools to obtain their initial feedback on the proposed designs and generate new ideas on the content and training scenarios. Sessions lasted up to 2 hours and were facilitated by an experienced qualitative researcher (AMB) with support from another member of the research team (HG). Materials were presented to participants, including the initial wireframe structure, visual design concepts, mock-ups of the module, interactive quizzes, and scenarios that had been developed with the expert panel (Figure 2).

XSL•FO
RenderX

Figure 2. Example materials used in the co-design workshops with school staff. Note: images with visible people are stock images sourced from iStock.

User Testing of the e-Learning Module

Once the e-learning prototype was finalized, we conducted a user testing phase. Participants were given 2 weeks to complete the 30-minute e-learning module and a short survey to capture metrics on usability (Multimedia Appendix 3). A topic guide was developed to explore usability and acceptability, as well their perspectives on changes in their knowledge and confidence (Multimedia Appendix 4). Follow-up focus groups were conducted at 3 schools; one of these schools had taken part in the previous workshops, but new staff members were recruited for the user testing. Focus group sessions lasted approximately 60 to 70 minutes. During the sessions, we explored staff's views on the design, usability, and navigability of the module, and similarly to the previous design and development phase, participants were asked about their views on the acceptability, feasibility, and barriers to implementation. The finalized module structure is shown in Figure 3. Visuals of the e-learning module prototype are shown in Figures 4 and 5, and screenshots from the filmed scenarios are shown in Figure 6.

XSL•FO
RenderX

Figure 3. Final structure of the e-learning module for user testing. NICE: National Institute for Health and Care Excellence; SORTS: Supportive Response to Self-Harm.

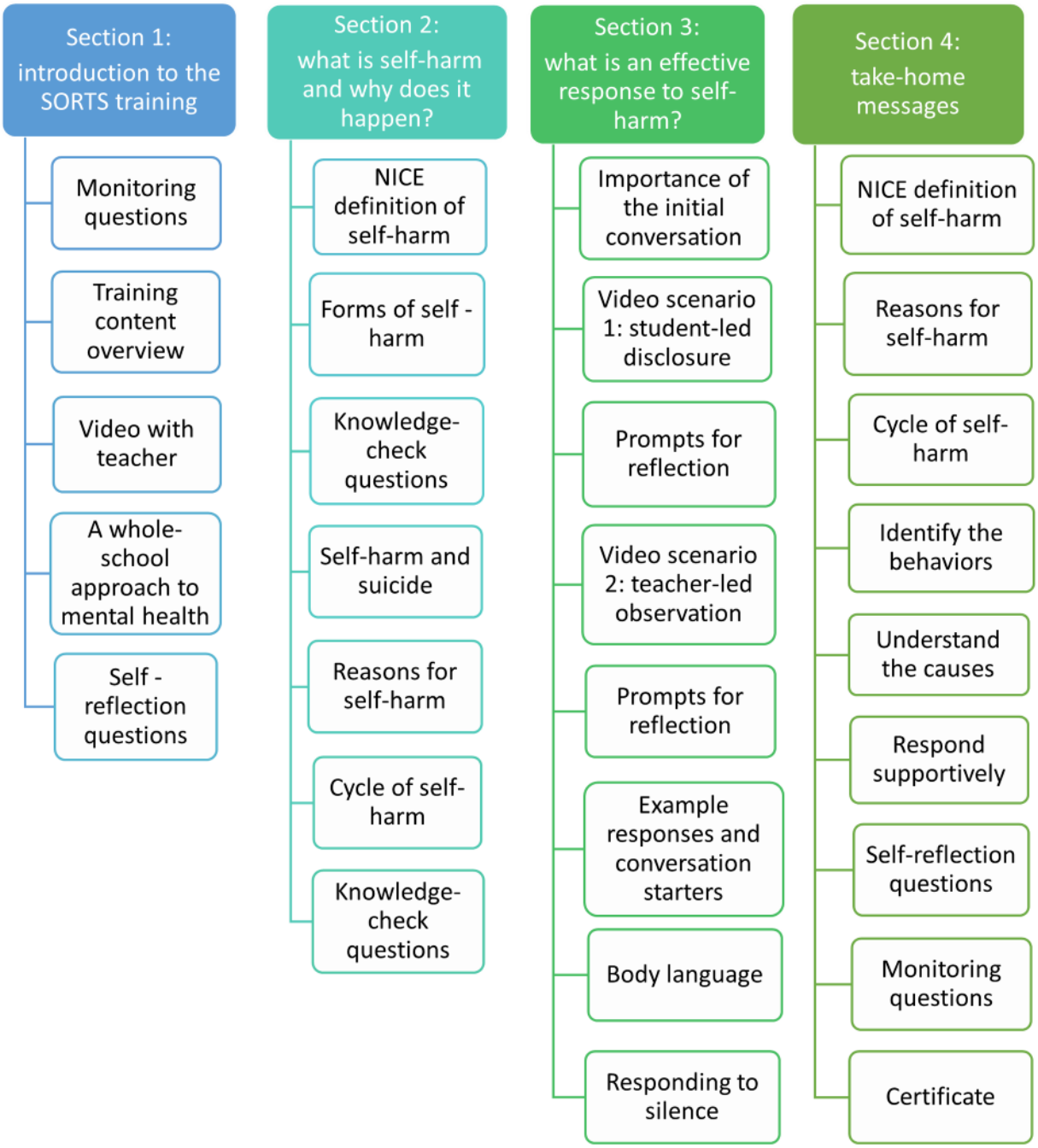

XSL•FO
RenderX

Figure 4. Supportive Response to Self-Harm (SORTS) e-learning module—welcome screen (A) and menu screen (B) for user testing.

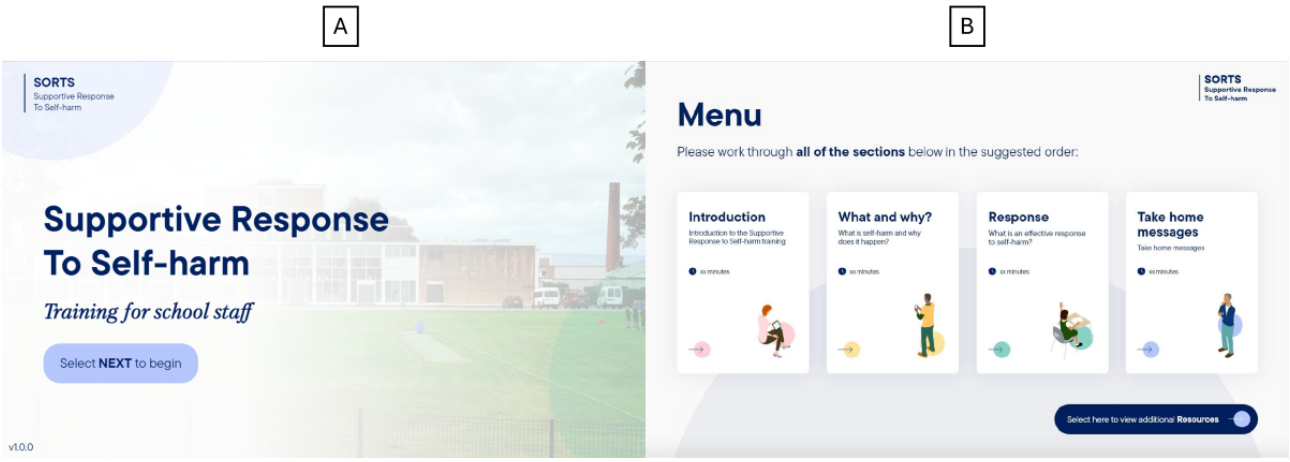

Figure 5. Examples of quiz questions in the e-learning module.

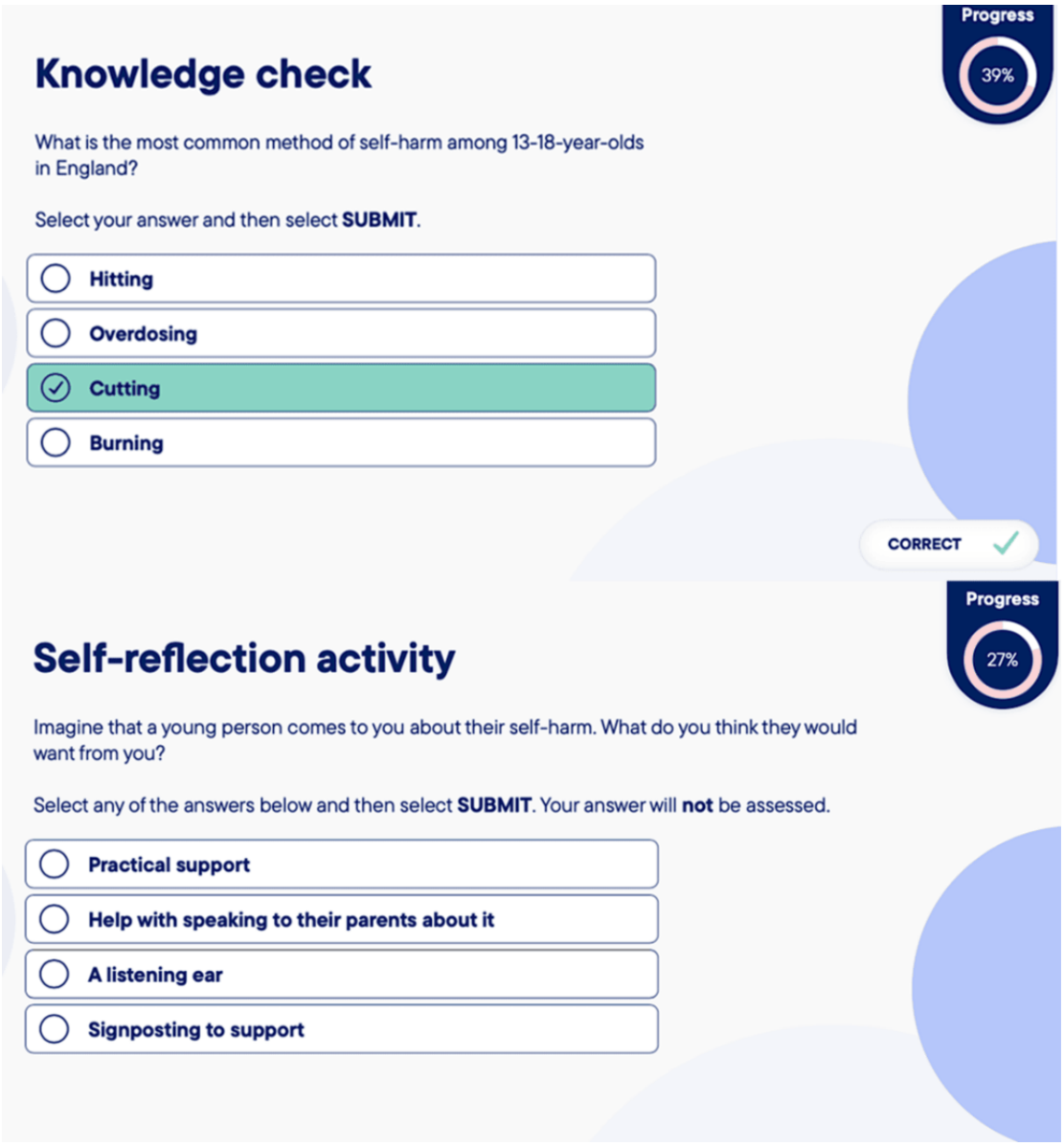

XSL•FO
RenderX

Figure 6. Screenshots of filmed scenarios—Chloe's story (A) and Jack's story (B). Note: images of actors from the film are intellectual property of the authors.

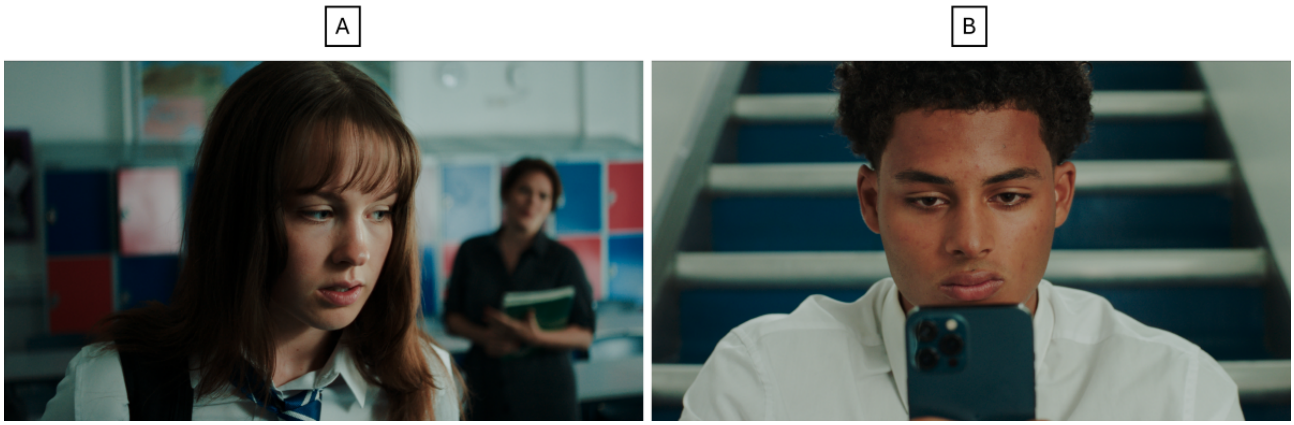

Analysis

All focus groups and workshops were recorded, transcribed professionally, thematically analyzed using the framework method [34], and triangulated with usability data. The transcripts were checked for accuracy, and all identifying information was removed before being entered into NVivo (version 12; QSR International) for data management. A detailed analysis plan was drawn up to guide members of the research team through the analysis. The process comprised the following stages: transcription; familiarization with the transcripts; coding; and development of a working analytical framework, application of the framework, charting of data into the framework matrix, and interpretation of the data. The team (AMB, HG, and JKA) began by familiarizing themselves with a subset of transcripts from each participant group and using a combination of deductive and inductive coding. Team members met to discuss how the codes could be grouped into categories, and these formed 2 initial working analytical frameworks (one for student data and one for school staff data). The working analytical frameworks were entered into NVivo and applied to all transcripts. Once all the transcripts had been coded, Microsoft Excel framework matrices were generated, which enabled the data to be charted by AMB and HG This involved creating detailed summaries in each cell of the matrix while retaining key verbatim quotes. The team met regularly to discuss potential themes and interpret the data. Monitoring data were exported from the SORTS e-learning module and analyzed using R Statistical Software (version 4.3.1; R Foundation for Statistical Computing). Two-tailed, paired samples *t* tests were conducted to assess whether there had been a significant change in participants' self-reported understanding and confidence rating at the beginning and end of the module.

Results

This section provides the results from the co-design workshops and user testing phase. Themes from the qualitative analysis are illustrated with example quotes from school staff.

Participants in the Workshops and User Testing

In total, 31 school staff members (n=23, 74% female; n=8, 26% male) participated in the study from 4 schools (n=3, 75% mixed sex and n=1, 25% single-sex boys' school). The single-sex boys' school took part in both phases, but different staff members attended the sessions. Participants' years of experience were 0 to 5 (21/31, 68%), 5 to 10 (4/31, 14%), 15 to 20 (3/31, 9%), and ≥20 (3/31, 9%), and they had a range of roles (Table 1).

XSL•FO
RenderX

Table 1. Roles of school staff members who participated in the co-design workshops and user testing (N=31).

Role	Participants, n (%)	Previous mental health training
Co-design workshop 1 (n=5)		
Associate assistant head (SLT[a])	1 (20)	No
Deputy SENCO[b]	1 (20)	Yes
Head of department for art and design	1 (20)	No
TA[c] and food technician	1 (20)	No
Safeguarding officer	1 (20)	Yes
Co-design workshop 2 (n=6)		
Deputy head (SLT)	1 (17)	No
Head of humanities (SLT)	1 (17)	Yes
School chaplain	1 (17)	No
Pastoral lead	1 (17)	Yes
Assistant principal and DSL[d] (SLT)	1 (17)	Yes
Administrator (secretary)	1 (17)	No
User testing in school A (n=9)		
Assistant head teacher (SLT)	1 (11)	Yes
Higher-level TA	1 (11)	Yes
Head of drama	1 (11)	No
Teachers (English and social sciences)	2 (22)	No
Career and marketing coordinator	1 (11)	No
Head of geography	1 (11)	No
Post 16 support officer	1 (11)	Yes
Librarian	1 (11)	No
User testing in school B (n=6)		
Senior mental health lead	1 (17)	Yes
Head of humanities	1 (17)	No
Cover supervisor (previously TA)	1 (17)	Yes
Head of maths	1 (17)	No
Teacher of religious education	1 (17)	No
Unit manager of the kitchen	1 (17)	No
User testing in school C (n=5)		
Safeguarding officer	1 (20)	Yes
Assistant head teacher and DSL	1 (20)	Yes
Designated mental health lead	1 (20)	Yes
Administrator (receptionist)	1 (20)	No
Deputy SENCO	1 (20)	Yes

[a]SLT: senior leadership team.
[b]SENCO: special educational needs coordinator.
[c]TA: teaching assistant.
[d]DSL: designated safeguarding lead.

XSL•FO
RenderX

Findings From Staff Co-Design Workshops

The qualitative analysis from the co-design workshops generated three main themes: (1) prospective acceptability of a self-harm awareness training module; (2) module design, structure, and content; and (3) potential barriers to and facilitators of implementation. This feedback was used to inform the design, development, and implementation of the training module.

Theme 1: Prospective Acceptability of a Self-Harm Awareness Training Module

Participants highlighted the need for guidance and training on how to manage self-harm, particularly because they worried about saying the wrong thing and inadvertently causing further distress for a student:

> *It's [having] that confidence of going, "Right, what you're going to say is then not going to make a negative impact to that child five hours later at home." And not having that fear of going—if you say the wrong thing, this might then make them progress even worse.* [Head of drama]

Some staff members noted that there were some practical subjects, such as food technology and art, where they had seen students use objects to self-harm and they had not received any training in how to manage that:

> *It just needs to be addressed. Because I've taught for a really long time and, as an art teacher and probably in food, there's a lot of objects where students can self-harm. And it's a practical subject, so I have seen students self-harming in lessons, I have seen students going to self-harm, or I notice things. I have never had any formal training on that.* [Associate assistant head]

Staff welcomed the idea of a training module but expressed concerns about being overwhelmed by the content and the potential for it to be distressing for some staff members. They felt that it was important for the training module to be sensitive to the lived experiences of school staff as they may have encountered self-harm both personally and professionally:

> *Depends on your own personal circumstances, as well, as to how you feel you can answer the questions. Because some people are more personally affected by it than others.* [Administrator]

Some staff members shared that they had personal experience with self-harm, which made them more invested in completing the training to better support students. While school therapists have supervision, staff members may complete the training and then return to their lessons without fully realizing how the content has affected them. As such, the module should also prioritize staff well-being:

> *I just want to say, make the point that, well, I'm not only speaking for myself, but we're here, obviously, as professionals, but as parents ourselves, I've personally had experience of this, so I just wanted to say that we'll be speaking from both a professional and a personal perspective.* [Assistant principal]

There was a strong consensus that all staff members, including those in nonteaching roles, would benefit from the e-learning module. Participants emphasized the importance of equipping all staff members with the skills to respond to potential discoveries or disclosures of self-harm. It was noted that some staff members who are often on the periphery of school training would benefit from being included:

> *...the canteen staff, for example, do have quite a lot of interactions with the kids.* [Associate assistant head]

> *I was a catering assistant for 10 years, and you're part of a school, but you're not part of a school. You know nothing. You're not really included. So, I think catering staff would probably benefit a lot, because they will know absolutely nothing. Because you are secluded.* [Food technician and teaching assistant]

Theme 2: Module Design, Structure, and Content

Staff members suggested that the training module include examples of good practice and opportunities for self-reflection through example scenarios of supportive communication with students. Incorporating scenarios would enhance engagement and help staff understand how to respond more confidently to potential real-life situations. Staff members emphasized the need for training videos to include empathetic and realistic safeguarding messages, clearly informing students that any disclosures may need to be shared with others. Videos should address the common hesitation that staff members can feel in challenging situations and encourage them to seek guidance if unsure. In addition, they recommended including sentence starters and response prompts to help those who are uncertain about how to respond to a discovery or disclosure:

> *Sentence-starters for a staff member who would have no idea what to say and would panic. If they read those and could see them, I think that's really helpful.* [Deputy special educational needs coordinator]

They highlighted that visuals and animations are crucial to keep staff engaged, although some pointed out that animations could be distracting, especially when trying to read the on-screen text simultaneously. For imagery, participants advised against stereotypical depictions of self-harm, recommending instead the use of subtle visuals that reflect the hidden or isolating aspects of self-harm. Importantly, they emphasized that images should be "realistic" and "reflect the diversity" of school communities.

Staff members recommended building in self-reflection activities at the beginning and end of the module to help learners gauge the impact of their learning:

> *Do you know what would be really powerful, is to do it [self-reflection] at the start and the end. In terms of realizing the students' perspective, in terms of what's really going on for them...I just think that that's, better than a certificate, is for me to know that that was worthwhile. That I've got it now, or I feel better equipped if I walk into a scenario like that.* [Assistant principal]

XSL•FO
RenderX

> *That depends on your own personal circumstances, as well, as to how you feel you can answer the questions. Because some people are more personally affected by it than others. And like you say—sometimes you get detached. You don't mean to, but you've seen a lot, and—No, but you know, it's true that in your role, if you see it lot—it's—it does put a different perspective on it.* [Administrator]

To meet diverse needs, participants suggested giving learners control over presentation formats, including options for subtitles, voice-overs, and playback features. Videos should ideally be approximately 2 minutes long and interspersed with quiz questions to maintain user engagement. Participants recommended allowing for flexible navigation within the module so that staff members could revisit or skip sections as needed:

> *Is there an option for you to make a video with no sound and then make a video with sound, and people could choose which one they want. Because I'm dyslexic myself, it's probably easier for people with learning disabilities to be able to access what you're talking about if there is sound. But then some people might find that too distracting.* [Head of department for art and design]

Theme 3: Potential Barriers to and Facilitators of Implementation

Staff members supported a whole-school approach that fosters a culture of openness around self-harm. Participants agreed that the training should be provided to all staff members as it would emphasize that safeguarding is a shared responsibility and help create a more supportive school environment:

> *Really, we all should have knowledge. Because actually, we all have a duty to safeguard. That's every adult in the college has a duty to safeguard. So, we could provide more benefit and help to students when they're struggling with other emotional issues...And so, I think every member of staff being trained would help everybody respond...it might just be that we get a better relationship to talking about it, so it's not such a taboo subject. And then there might be more understanding to help.* [Administrator]

Participants felt that making the training mandatory and allocating time for its completion would be key to successful implementation. Given busy school schedules, inset training days would be ideal for staff to complete the e-learning module. They recommended that a concise 20- to 30-minute module with the flexibility to start and stop completion would make it more manageable for all staff members:

> *Time [laughs]. It's not even that, I don't know, we could—the whole school, but it would have to be in an inset day, and with that we'd have to find out who the person that runs CPD, if we can have an allocated slot. Because no one's going to do it in their own time.* [Safeguarding officer]

While most staff members said that they would expect to complete the training on a desktop, they thought that the training module should be compatible with both desktop and mobile platforms to offer greater choice and flexibility:

> *It would be good to have the option, like...if you wanted to do it at home, and people had the option to do it on a mobile, it would be useful if it is kind of transferable—I'm sure I looked at it on my phone when I looked at it.* [Head of humanities]

Ensuring that the training is suitable for catering, administrative, and cleaning staff was seen as important as students may feel more comfortable approaching these individuals. Providing learners with CPD certificates on completion could boost participation and enable senior leaders in the school to track completions and follow up with staff members who have not finished. In terms of scaling up the training, they recommended partnering with reputable educational and mental health organizations and integrating it into teacher education programs:

> *I did my recent level 3 safeguarding training on a similar module, like this, and it was brilliant, because I could go away, come back, do it, and it worked really, really well.* [Assistant principal and teacher]

Concerns were raised about the possibility of staff discussions on self-harm with students conflicting with safeguarding protocols, which could potentially hinder schools from adopting the training intervention. To address this, the training should include guidance for staff on responding thoughtfully and nonjudgmentally to disclosures while clearly explaining to students the necessity to share information for their safety.

Findings From the User Testing

Overview

Monitoring questions were embedded into the SORTS module and stored anonymously in a learner record system. First, learners were asked for their "name of school or academy" and could select from a drop-down list of schools drawn from a UK database or "other organization." Learners were also asked "how would you classify your role?" We generated 3 categories to capture similar roles, responsibilities, and previous training in UK school settings. Participants from the user testing schools self-reported their roles as "teacher, senior leadership or pastoral" (12/20, 60%), "teaching assistant, administrative staff, librarian" (7/20, 35%), and "school facilities (catering, cleaning, technician)" (1/20, 5%). Learners were asked to rate their confidence and understanding of self-harm at the beginning and end of the training. Finally, learners were asked to rate the training. Mean scores are shown in Table 2.

XSL•FO
RenderX

Table 2. Participants' ratings for the monitoring questions embedded into the Supportive Response to Self-Harm e-learning module (N=19)[a].

Monitoring question	Response scale	Score, mean (SD)	
		Start of module	End of module
"I feel confident that I can effectively handle a conversation with a student about self-harm."	0—"not at all confident" to 9—"very confident"	5.32 (1.86)	7.47 (1.07)
"I feel confident that I have a good understanding of what causes young people to self-harm."	0—"not at all confident" to 9—"very confident"	6.05 (1.51)	7.84 (1.01)
"How would you rate this training?"	1—"poor" to 5—"excellent"	__[b]	4.42 (0.77)

[a]Embedded monitoring data did not save for 5% (1/20) of the participants. Therefore, the mean scores presented in this table are based on 95% (19/20) of the user testing participants.

[b]Participants only provided a response for "how would you rate this training?" at the end of the module.

Paired-sample *t* tests (2-tailed) were conducted to compare responses from the start and end of the module. There were significant differences in the scores for the following statements—"I feel confident that I can effectively handle a conversation with a student about self-harm" (t_{18}=−7.4717; P<.001; equal variances assumed) and "I feel confident that I have a good understanding of what causes young people to self-harm" (t_{18}=−7.5607; P<.001; equal variances not assumed)—at the start and end of the module.

Tables 3 and 4 show the findings from the usability survey. Results show that there were high ratings for usability metrics, including ease of use and the acceptability of the design and content. Recommendations for improvements were provided in the open-ended survey responses (Table 4) and are discussed in the qualitative themes in the following sections. The key themes from the user testing with staff were (1) usability, (2) content and design, (3) feasibility, and (4) views on how the training improved knowledge and confidence.

Table 3. Participant responses to the user testing survey (N=20)[a].

Survey item	Score, mean (SD)
Usability	
"It was easy for me to access the training module."	4.76 (0.43)
"The objectives of the training module were clear to me."	4.81 (0.93)
"I found the training module easy to use."	4.71 (0.70)
"I needed help to use the SORTS module."	1.33 (0.89)
"I found the videos engaging."	4.33 (0.78)
"I found the quizzes engaging."	4.05 (0.58)
"I liked the design of the module."	4.29 (0.63)
"The images used were relevant to schools."	4.48 (0.50)
"The training content is relevant to my job."	4.62 (0.58)
"I understood the training content."	4.81 (0.39)
"I could easily download and save my CPD certificate."	4.67 (0.71)
Overall satisfaction	
"The training was engaging."	4.38 (0.49)
"After completing the training, I feel more confident in how to respond to self-harm."	4.38 (0.79)
"How likely are you to recommend this product to a colleague?"	4.43 (0.58)
"The training is important for school staff to complete."	4.62 (0.49)
"I could fit this training into my work schedule."	4.38 (0.84)
"Overall, I was satisfied with the quality of this training module."	4.62 (0.49)

[a]All survey items were scored from 1 to 5, with 1=strongly disagree, 2=disagree, 3=neither agree or disagree, 4=agree, and 5=strongly agree.

XSL•FO
RenderX

Table 4. Participant feedback on the open-ended usability survey questions (N=20).

Open-ended question	Positive feedback	Suggestions for improvement
"What, if anything, did you not like about the training?"	A total of 14 participants said that there was nothing they disliked.	In total, 7 participants commented on features they did not like: the voice-over was too slow, repeated instructions, and difficult to see on a phone.
"Did you experience anything in the module that would prevent you from completing it?"	All participants said that there was nothing in the module that would prevent them from completing it.	One participant reiterated that they expected the module to be more in depth and show images to help identify self-harm injuries.
"Is there anything that would make you more likely to complete this training?"	A total of 16 participants said that there was nothing that would make them more likely to complete the training and left positive comments, such as describing the module as very informative and succinct.	In total, 5 participants said that they would be more likely to complete the module if there were more videos representing a wider range of self-harm methods, more statistics and information on the causes of self-harm, and allocated CPD[a] time to complete the module.
"Were there any difficulties for you accessing/beginning the module?"	In total, 20 participants reported no difficulties accessing or beginning the module and described it as very easy, with clear explanations of how the module would be laid out.	One participant thought that their progress would be saved if they exited the module and, therefore, reported having to start it from the beginning after exiting.
"Overall, how would you describe the module?"	A total of 19 participants described the module in very positive terms, including (1) easy-to-digest and accessible entry-level training for all school staff members; (2) very practical, useful, and clear advice; (3) engaging and user-friendly format; (4) soft and relaxing design and colors; (5) information that staff had not encountered in other training; and (6) feeling more confident after completing the module.	__[b]
"If you could change anything about it, what would you change?"	In total, 10 participants said that there was nothing they would change and left positive comments regarding the diversity represented in the visuals and the mixture of interactive features.	Suggestions for changes included myth busting; more challenging quiz questions; further explanation of the scenarios, what went well, and what could be improved; links to additional learning resources throughout the module for those seeking a higher training level; more content representing the student voice; and strategies to put in place beyond the initial conversation.

[a]CPD: continuing professional development.

[b]Not reported.

Theme 1: Usability (Ease of Use and Navigability)

Staff generally found the training module easy to start and navigate. They found the structure intuitive and valued the ability to easily return to previous sections:

> *Really straightforward, and you could just navigate through really easily. I liked how—I think I did something wrong at one point, and I could just really easily navigate back as well. And I liked that because sometimes when you do these trainings, once you've skipped forward you can't go back. And then you think, "Oh no, I didn't grasp the right thing," or, "Actually, I want to re-read that." So, I felt that that was nice, to be able to do that.* [Post 16 support officer]

However, participants recommended some minor usability improvements to streamline the experience. For example, the "Continue" button was not always visible on certain devices, possibly due to the screen zoom settings, which caused confusion. A couple of participants found it difficult to determine where they were within the module, especially when they returned to the home page without a section being marked as completed. They suggested that a clearer overview at the start could help clarify progress:

> *I did have a bit where it kind of takes you back out. And it hadn't said that I'd completed the first module, so I just went back into it. And then I was like, hang on, I've done this before. I ended up flicking through the first module again just to check that I had done it before I came, went into the second one. So maybe just like a tick screen or something [showing] that that's completed.* [Librarian]

Theme 2: Content and Design

Participants generally felt that the training module provided an appropriate level of information, with content that was easy to understand and well balanced for both teaching and nonteaching staff. However, some information was repeated multiple times, which some users found frustrating. Overall, they found the module easy to follow as information was presented in manageable "bite-sized pieces" that were not overwhelming:

> *I felt like it was really digestible chunks of information. I never felt overwhelmed by it. And I*

XSL•FO
RenderX

> *think sometimes, especially with that content—and I think that's one of the reasons I wanted to do it, so I didn't feel that overwhelm. But I never did. I felt like I could follow it really nicely through.* [Head of geography]

Staff members with previous mental health training said that the training content reinforced their existing knowledge and validated their instincts on how to approach sensitive situations and handle difficult conversations with students:

> *The thing I found the most useful from it was there were some bits where I was going, "OK, I thought that would have been the approach I would have taken" on a scenario...So it wasn't for me going, "Oh, I know this." It was actually gaining that confidence of going, "Actually, what my instincts were saying were down the right lines." I think that would be helpful for everyone in that way.* [Head of drama]

All staff members valued the "knowledge check" and "self-reflection" questions, which helped reinforce understanding of key facts and gauge their understanding and progress throughout the module. They found that the self-reflection prompts specifically encouraged them to consider their own feelings and responses in real-life scenarios:

> *I really liked the one with the timer...because it wasn't too long, and it just made me sit there for 20 seconds and go, "OK, how would I actually feel?" and actually think about it. Not for so long that it felt like a drag. But for just a little bit of time to have something slightly more than a surface-level thought.* [Assistant principal]

> *Yeah, I quite like the fact that there were two different types. And I like that it's kind of integrated throughout the modules as well, because I find that, well, it's all saved up towards the end. If I get something wrong, I don't always go back and double check what I was supposed to have. And it's nice that it's kind of right there, so that it's you've just read it through, and then you kind of have the reconfirmation of the facts as you've completed the questions as well. And then the more open-ended ones I did. I quite like those, because it was like, well, it made you really, actually consider what would happen in a real-life scenario. So that was good too.* [Librarian]

Participants found the module visually engaging, particularly the filmed scenarios, which they said were realistic and inclusive. While some initially worried that one of the videos depicted an incorrect approach, they soon realized that this was intentional and illustrated a progression to best practice. This highlighted the need to provide additional information before each video scenario.

In addition, they appreciated the inclusion of a male student in one of the videos as it helped challenge the stereotype that self-harm is primarily a female issue. This was seen as a constructive way to break down gender stereotypes around mental health, making the module more inclusive and relatable to a broader audience:

> *I really liked that the video was a boy. I think that made it so that—I feel like sometimes the stereotype is that it's females who will cut themselves, not necessarily different types of self-harm. And I thought it was really good that it was a man speaking to a boy. It just jumped over that hurdle, that barrier or the stereotype that we have.* [Social sciences teacher]

> *Quite a few of us are support staff. I would have quite liked something that was not classroom teacher-based, because the situations in which we interact with students are not the same.* [Assistant principal]

Theme 3: Feasibility (Implementation and Adoption)

There was a consensus that the module's flexible format would make it easy to implement in schools. The 30-minute duration was considered ideal as it would allow staff members to gain essential insights without a major time commitment. Senior staff members said that they would incorporate the module into allocated training time, such as staff inset days and CPD sessions or as part of new staff induction programs:

> *I think I prefer doing it as a whole-staff CPD, not in one go. Because poor people, when they start at schools, it's information...it can't go on forever, because you won't get the buy-in from schools. But I think to go, "Right, this is going to take you 20 minutes guys, and it could be really helpful," would be how I see it being done.* [Assistant principal]

One senior staff member suggested that staff members could complete the training as a group rather than individually. Participants highlighted the benefits of allowing time for discussions during or after the training to reinforce learning and facilitate self-reflection. They acknowledged that the topic could be distressing for some staff members and support should be available immediately after training to address any issues that arise. This highlighted the need for clear implementation guidelines for schools using the training.

Theme 4: Views on How the Training Improved Knowledge and Confidence

The training was valued for fostering a shared understanding among staff members and establishing a consistent "baseline" for approaching self-harm within the school. Staff members appreciated how the module explained the complex and personal nature of self-harm, highlighting that it may not always have a clear cause. The training broadened their awareness of the different forms that self-harm can take beyond visible injuries, helping them recognize that less obvious behaviors, such as ingesting harmful substances, can also be forms of self-harm. This increased their understanding and awareness of different warning signs:

> *I think it was really beneficial having the different types of self-harm, as we automatically just think of cutting, and obviously there's so many. So I think that was really good to have a dedicated section on that.* [Special educational needs coordinator]

XSL•FO
RenderX

> *Things like hair-pulling and all those things that are maybe less easy to pick up on. And in general, it's an incredibly secretive thing anyway, so people usually go to great pains to keep it to themselves.* [Designated mental health lead]

Some participants already felt knowledgeable about self-harm due to their personal experiences or from interactions with students who were struggling with self-harm. For these individuals, the training provided reassurance that they were handling disclosures appropriately and offering a supportive response. Others noted that this training would be particularly beneficial for nonteaching staff members who do not work as closely with students, as it would provide them with essential insights and help them notice the warning signs:

> *I would feel really good about other people having done the same knowledge, and knowing that this is the level that other people—that everyone in the school has had...So that if I had to refer it on to someone—I know that they'd be getting this kind of positive approach.* [Assistant head]

Overall, participants said that the training improved their confidence in recognizing and responding to self-harm, and it emphasized the importance of a compassionate approach. Staff with previous mental health training felt that the e-learning module reinforced their confidence in addressing self-harm and affirmed their ability to approach situations appropriately. They particularly liked the training's emphasis on maintaining a supportive, conversational approach rather than focusing solely on procedural aspects. However, some suggested including guidance on handling situations in which students are unwilling to talk:

> *For me, that was my first training I've done on the mental health. So I was going into it not really knowing too much. But after, I felt a lot more confident of signs, what to do. So yeah, it made me feel like, OK, if someone comes to me, I knew how to approach it and go about it.* [Administrator]

Several staff members found it helpful that the training highlighted the importance of positive body language and eye contact, which they felt would build trust and make students feel genuinely heard. One participant suggested that body language tips may need to be adapted for students with autism or attention-deficit/hyperactivity disorder, and this led to a discussion on the need for an additional training module specifically addressing self-harm among neurodiverse students:

> *I really, really, really agree with the whole you have to communicate non-verbally. But I will say on the body language and eye contact thing, as someone who works almost exclusively with neurodiverse children, sometimes the things that are good advice if you're talking to a neurotypical child, would come across very intimidating and demanding to a neurodiverse child.* [Higher-level teaching assistant]

Discussion

Summary of Findings

This manuscript outlines the collaborative design and development of a self-harm awareness e-learning module for school staff. We also present the qualitative and quantitative evidence of the module's usability and acceptability from a diverse sample of school staff members, including those in senior leadership, teaching, administrative, and school catering roles. The training module builds on previous work [23,25] and addresses the lack of effective and acceptable training to support whole-school awareness of self-harm [17,18]. We identified one recent study that developed an e-learning module on self-harm for schools; however, it was specifically designed for teachers rather than a wider range of teaching or nonteaching staff, which does not fit with a whole-school approach. In addition, the study provided limited details on the training content and development process and did not appear to use a user-centered or person-based approach. Moreover, the absence of follow-up data makes it challenging to assess the long-term impacts of the training [21].

In contrast, the SORTS e-learning module follows a whole-school approach in alignment with the UK government's national policy to improve mental health prevention and intervention for children and young people in educational settings [8]. It recognizes that all school staff members are uniquely positioned to intervene when they identify a young person who has self-harmed, helping prevent further escalation and guiding students toward appropriate support. Research has shown that school staff are often the first point of contact for parents concerned about their children's well-being, with schools being described as a "one-stop-shop" for advice for parents before seeking help from health professionals [25,35,36]. To support a whole-school approach, it is essential that all school staff members receive the necessary training and guidance to effectively respond to students who self-harm [37].

In this study, feedback from the 2 participatory design cycles was instrumental in shaping the design and development of the SORTS e-learning module. We involved relevant users throughout the development process, a key step in improving the acceptability and uptake of interventions [38,39]. Consistent with previous research, most school staff members reported feeling ill-equipped and stated that they had not received self-harm awareness training, highlighting an unmet need for such training [11,17,18,40]. Some participants noted that their awareness and understanding of self-harm were shaped by their previous personal and professional experiences, resulting in considerable variation in staff knowledge within schools. The co-design workshops gathered valuable staff input on the content, design, and structure of the module before prototype development. Several user requirements were identified, including a simple, user-friendly interface; accessibility features; interactive elements to engage users; and concise information to avoid overwhelming staff. Participants also requested model responses in video format and suggested including sentence starters for situations in which they felt unsure of what to say. In addition, they highlighted the importance of opportunities

XSL•FO
RenderX

for reflection to explore their own beliefs and experiences related to self-harm. This aligns with evidence that reflection is a key mechanism supporting effective implementation of interventions in school settings [41]. In response, we created 2 video scenarios demonstrating best practices with suggested sentence starters to support staff in navigating difficult conversations regarding discovery or disclosure. We included references to the school's safeguarding protocols to reinforce safe, effective procedures. For accessibility, a voice-over option was added, and the module was designed to function seamlessly across different devices. To enhance usability, we ensured that the development was based on the general principles of interaction design [42].

User testing showed that the training was easy to use and navigate, with both the content and design being highly acceptable to staff. Results showed an increase in staff's knowledge and confidence, and user testing focus groups indicated that this was particularly true for those with no previous mental health training. This is important because previous research has found that a lack of knowledge and confidence can lead to negative responses [13,40]. For those with previous mental health training, the SORTS module reinforced their existing knowledge and provided reassurance that they were responding supportively to students. For many participants, the module helped them recognize a wider range of self-harm behaviors, suggesting that, without training, school staff may overlook certain behaviors or fail to respond effectively due to a lack of awareness.

Stigma is a major barrier to help seeking. Research indicates that young people are often reluctant to disclose their self-harm due to concerns about receiving a negative response [14,16]. However, supportive student-teacher relationships, characterized by warmth and understanding, play a crucial role in promoting student well-being [25]. Therefore, training must be integrated into a broader cultural shift that fosters a whole-school approach, which includes implementing a school policy on self-harm; incorporating relevant education into the curriculum; and providing support for parents, siblings, and peers of students who self-harm [23,37].

Strengths and Limitations

A key strength of this study is the rigorous co-design methodology. It was conducted in line with Medical Research Council guidelines for complex interventions, involving regular input from "appropriate users" at all stages of intervention development to ensure that the training intervention is acceptable, feasible, and engaging [43]. It applied a person-based approach to guide the development [26]. Staff workshops ensured that school staff members' perspectives were incorporated throughout the iterative design process. Another strength of this study is the mix of urban and rural schools with diverse student populations that participated in the study. Schools were in regions where the local population has limited opportunity to take part in health research. We recruited school staff members in a range of roles to ensure a range of views from teaching and nonteaching staff.

Despite the key strengths outlined previously, some limitations should be acknowledged. Those who volunteered to participate in this study are more likely to have an interest in student mental health, and therefore, we may not have captured the views of staff members who are more concerned about responding to a student who is self-harming. Some decisions about the e-learning module content and design were made before the staff workshops as it was necessary to have content for staff to consult on. However, this previous work was developed in consultation with experts in mental health, education, and specifically self-harm. A limitation of this study could have been that young people were not involved in the development work. However, we have previously conducted 2 separate studies with young people about training school staff about self-harm, which directed the research in this study [23,25].

Implications and Future Work

The SORTS e-learning module was designed for all school staff members to support a whole-school approach to self-harm prevention and intervention. We anticipate that implementing this approach in schools will encourage young people to seek help from staff members, improve signposting, and facilitate access to community resources and support. While it is targeted at school staff, it may also be relevant for other professionals working with young people, for example, those in youth organizations, social care, or juvenile justice.

The module is already being used by schools, and preliminary data collected before and after training suggest that staff knowledge and confidence increase after completing the training-learning module. We are now planning a feasibility study to evaluate the module's effectiveness across several schools, assessing both staff and student outcomes.

In addition, staff members in this study highlighted gaps in their training related to other areas of mental health. Future work could involve co-designing similar modules on topics such as absenteeism, eating disorders, and supporting neurodiverse students.

Acknowledgments

The authors would like to thank all the participating schools, in particular the senior school staff members who helped with recruiting school staff for the workshops and user testing phases. They would also like to thank the expert panel members and members of the Charlie Waller Trust [44] for their support and contributions to the development of the Supportive Response to Self-Harm program. The authors thank Lauren Hitchcock, who helped create Figure 3. This work is funded by the UK Research and Innovation Medical Research Council (Public Health Intervention Development; MR/Y503277/1). All research at the Department of Psychiatry at the University of Cambridge is supported by the National Institute for Health and Care Research (NIHR) Cambridge Biomedical Research Centre (NIHR203312) and the NIHR Applied Research Collaboration East of England.

XSL•FO
RenderX

The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care. AMB is supported by the NIHR Applied Research Collaboration East of England.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

AMB and JKA conceived the study design and wrote the protocol. AMB, JKA, and HG were responsible for the development of study materials, recruitment, and data collection and analysis. AMB, JKA, and HG developed the e-learning module content, videos, and prototype designs. AMB, JKA, and HG wrote the first draft of this paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supportive Response to Self-Harm theory of change model.
[PDF File (Adobe PDF File), 156 KB - formative_v9i1e69309_app1.pdf]

Multimedia Appendix 2
Co-design workshop topic guide.
[DOCX File , 22 KB - formative_v9i1e69309_app2.docx]

Multimedia Appendix 3
User testing survey.
[DOCX File , 67 KB - formative_v9i1e69309_app3.docx]

Multimedia Appendix 4
User testing topic guide.
[DOCX File , 24 KB - formative_v9i1e69309_app4.docx]

References

1. Self-Harm: the short-term physical and psychological management and secondary prevention of self-harm in primary and secondary care. National Institute for Health and Clinical Excellence (NICE). 2004. URL: https://www.ncbi.nlm.nih.gov/books/NBK56385/pdf/Bookshelf_NBK56385.pdf [accessed 2025-05-29]
2. Hawton K, Bale L, Brand F, Townsend E, Ness J, Waters K, et al. Mortality in children and adolescents following presentation to hospital after non-fatal self-harm in the Multicentre Study of Self-harm: a prospective observational cohort study. Lancet Child Adolesc Health 2020 Feb;4(2):111-120. [doi: 10.1016/S2352-4642(19)30373-6] [Medline: 31926769]
3. Hulbert S, Eida T, Ferris E, Hrytsenko V, Kendall S. HBSC England National Report Findings from the 2021-2022 HBSC study for England Health Behaviour in School-aged children (HBSC): World Health Organisation collaborative cross-national study. Health Behaviour in School-Aged Children (HBSC). 2022. URL: https://hbscengland.org/wp-content/uploads/2023/11/2022_FULL_REPORT_final_21.11.23-1.pdf [accessed 2025-05-29]
4. Hawton K, Rodham K, Evans E, Weatherall R. Deliberate self harm in adolescents: self report survey in schools in England. BMJ 2002 Nov 23;325(7374):1207-1211 [FREE Full text] [doi: 10.1136/bmj.325.7374.1207] [Medline: 12446536]
5. Borschmann R, Becker D, Coffey C, Spry E, Moreno-Betancur M, Moran P, et al. 20-year outcomes in adolescents who self-harm: a population-based cohort study. Lancet Child Adolesc Health 2017 Nov;1(3):195-202. [doi: 10.1016/S2352-4642(17)30007-X] [Medline: 30169168]
6. Hawton K, Saunders KE, O'Connor RC. Self-harm and suicide in adolescents. Lancet 2012 Jun 23;379(9834):2373-2382. [doi: 10.1016/S0140-6736(12)60322-5] [Medline: 22726518]
7. Daukantaitė D, Lundh LG, Wångby-Lundh M, Claréus B, Bjärehed J, Zhou Y, et al. What happens to young adults who have engaged in self-injurious behavior as adolescents? A 10-year follow-up. Eur Child Adolesc Psychiatry 2021 Mar;30(3):475-492 [FREE Full text] [doi: 10.1007/s00787-020-01533-4] [Medline: 32318877]
8. Transforming children and young people's mental health provision. UK Government. URL: https://www.gov.uk/government/publications/transforming-children-and-young-peoples-mental-health-provision [accessed 2025-05-29]
9. Evans R, Hurrell C. The role of schools in children and young people's self-harm and suicide: systematic review and meta-ethnography of qualitative research. BMC Public Health 2016 May 14;16:401 [FREE Full text] [doi: 10.1186/s12889-016-3065-2] [Medline: 27179765]

XSL•FO
RenderX

10. Nawaz RF, Anderson JK, Colville L, Fraser-Andrews C, Ford TJ. Review: interventions to prevent or manage self-harm among students in educational settings - a systematic review. Child Adolesc Ment Health 2024 Feb;29(1):56-69. [doi: 10.1111/camh.12634] [Medline: 36625166]
11. Berger E, Hasking P, Reupert A. “We’re working in the dark here”: education needs of teachers and school staff regarding student self-injury. School Mental Health 2013 Nov 27;6(3):201-212. [doi: 10.1007/S12310-013-9114-4]
12. Heath NL, Toste JR, Sornberger MJ, Wagner C. Teachers’ perceptions of non-suicidal self-injury in the schools. School Mental Health 2010 Oct 5;3(1):35-43. [doi: 10.1007/S12310-010-9043-4]
13. Best R. Deliberate self-harm in adolescence: a challenge for schools. Br J Guid Couns 2006 May;34(2):161-175. [doi: 10.1080/03069880600583196]
14. Rowe SL, French RS, Henderson C, Ougrin D, Slade M, Moran P. Help-seeking behaviour and adolescent self-harm: a systematic review. Aust N Z J Psychiatry 2014 Dec;48(12):1083-1095. [doi: 10.1177/0004867414555718] [Medline: 25335872]
15. Nada-Raja S, Morrison D, Skegg K. A population-based study of help-seeking for self-harm in young adults. Aust N Z J Psychiatry 2003 Oct;37(5):600-605. [doi: 10.1046/j.1440-1614.2003.01252.x] [Medline: 14511089]
16. Curtis S, Thorn P, McRoberts A, Hetrick S, Rice S, Robinson J. Caring for young people who self-harm: a review of perspectives from families and young people. Int J Environ Res Public Health 2018 May 10;15(5):950 [FREE Full text] [doi: 10.3390/ijerph15050950] [Medline: 29747476]
17. Pierret AC, Anderson JK, Ford TJ, Burn AM. Review: education and training interventions, and support tools for school staff to adequately respond to young people who disclose self-harm - a systematic literature review of effectiveness, feasibility and acceptability. Child Adolesc Ment Health 2022 May;27(2):161-172. [doi: 10.1111/camh.12436] [Medline: 33277965]
18. Evans R, Parker R, Russell AE, Mathews F, Ford T, Hewitt G, et al. Adolescent self-harm prevention and intervention in secondary schools: a survey of staff in England and Wales. Child Adolesc Ment Health 2019 Sep;24(3):230-238 [FREE Full text] [doi: 10.1111/camh.12308] [Medline: 31588199]
19. Timson D, Priest H, Clark-Carter D. Adolescents who self-harm: professional staff knowledge, attitudes and training needs. J Adolesc 2012 Oct;35(5):1307-1314. [doi: 10.1016/j.adolescence.2012.05.001] [Medline: 22705150]
20. Lee F. Self-harm training in secondary schools: an educational psychology intervention using interpretative phenomenological analysis. Educ Child Psychol 2016 Jun 01;33(2):105-116. [doi: 10.53841/bpsecp.2016.33.2.105]
21. Price C, Satherley RM, Jones CJ, John M. Development and evaluation of an eLearning training module to improve United Kingdom secondary school teachers’ knowledge and confidence in supporting young people who self-harm. Front Educ 2022 Jun 20;7:52. [doi: 10.3389/feduc.2022.889659]
22. Jeffery D, Warm A. A study of service providers' understanding of self-harm. J Ment Health 2009 Jul 06;11(3):295-303. [doi: 10.1080/09638230020023679]
23. Burn AM, Hall P, Anderson J. A web-based training program for school staff to respond to self-harm: design and development of the supportive response to self-harm program. JMIR Form Res 2024 Jun 04;8:e50024 [FREE Full text] [doi: 10.2196/50024] [Medline: 38833286]
24. Recognising and responding to self-harm in schools. SORTS. URL: https://sorts4schools.org.uk/ [accessed 2024-10-31]
25. Colville L, Anderson JK, Burn AM. How schools can respond to pupils who self-harm: a qualitative study with young people and school staff. Emot Behav Diffic 2024 Oct 14;29(3-4):123-137. [doi: 10.1080/13632752.2024.2405806]
26. Yardley L, Morrison L, Bradbury K, Muller I. The person-based approach to intervention development: application to digital health-related behavior change interventions. J Med Internet Res 2015 Jan 30;17(1):e30 [FREE Full text] [doi: 10.2196/jmir.4055] [Medline: 25639757]
27. Bandura A, Adams NE. Analysis of self-efficacy theory of behavioral change. Cogn Ther Res 1977 Dec;1(4):287-310. [doi: 10.1007/bf01663995]
28. Gunning AM, Mensah FM. Preservice elementary teachers’ development of self-efficacy and confidence to teach science: a case study. J Sci Teacher Educ 2017 Feb 23;22(2):171-185. [doi: 10.1007/S10972-010-9198-8]
29. Finney D. The road to self‐efficacy: a discussion of generic training in mental health competencies for educational professionals. Pastoral Care Educ 2009 Mar;27(1):21-28. [doi: 10.1080/02643940902731831]
30. Moore GF, Evans RE. What theory, for whom and in which context? Reflections on the application of theory in the development and evaluation of complex population health interventions. SSM Popul Health 2017 Dec;3:132-135 [FREE Full text] [doi: 10.1016/j.ssmph.2016.12.005] [Medline: 29302610]
31. Sekhon M, Cartwright M, Francis JJ. Acceptability of healthcare interventions: an overview of reviews and development of a theoretical framework. BMC Health Serv Res 2017 Jan 26;17(1):88 [FREE Full text] [doi: 10.1186/s12913-017-2031-8] [Medline: 28126032]
32. Boren MT, Ramey J. Thinking aloud: reconciling theory and practice. IEEE Trans Profess Commun 2000;43(3):261-278. [doi: 10.1109/47.867942]
33. Porter B, Wills W. Understanding the barriers and enablers of the applied research collaboration (East of England) population-in-focus approach a qualitative review. Applied Research Collaboration, East of England. URL: https://arc-eoe.nihr.ac.uk/sites/default/files/uploads/files/Full%20PiF%20Review%20Report%202023.pdf [accessed 2024-11-01]

XSL•FO
RenderX

34. Gale NK, Heath G, Cameron E, Rashid S, Redwood S. Using the framework method for the analysis of qualitative data in multi-disciplinary health research. BMC Med Res Methodol 2013 Sep 18;13:117 [FREE Full text] [doi: 10.1186/1471-2288-13-117] [Medline: 24047204]
35. Newlove-Delgado T, Moore D, Ukoumunne O, Stein K, Ford T. Mental health related contact with education professionals in the British Child and Adolescent Mental Health Survey 2004. J Ment Health Train Educ Pract 2015;10(3):169 [FREE Full text] [doi: 10.1108/JMHTEP-02-2015-0007]
36. Mental health of children and young people in England. National Health Service England. 2018. URL: https://digital.nhs.uk/data-and-information/publications/statistical/mental-health-of-children-and-young-people-in-england/2017/2017 [accessed 2024-05-20]
37. Meinhardt I, Cuthbert S, Gibson K, Fortune S, Hetrick SE. Young people and adult stakeholders' reflections on how school staff should support students who self-harm: a qualitative study. J Adolesc 2022 Oct;94(7):969-980. [doi: 10.1002/jad.12078] [Medline: 35880459]
38. Skivington K, Matthews L, Simpson SA, Craig P, Baird J, Blazeby JM, et al. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. BMJ 2021 Sep 30;374:n2061 [FREE Full text] [doi: 10.1136/bmj.n2061] [Medline: 34593508]
39. Hawkins J, Madden K, Fletcher A, Midgley L, Grant A, Cox G, et al. Development of a framework for the co-production and prototyping of public health interventions. BMC Public Health 2017 Sep 04;17(1):689 [FREE Full text] [doi: 10.1186/s12889-017-4695-8] [Medline: 28870192]
40. Berger E, Hasking P, Reupert A. Response and training needs of school staff towards student self-injury. Teach Teach Educ 2014 Nov;44:25-34. [doi: 10.1016/j.tate.2014.07.013]
41. Moore D, Proctor R, Benham-Clarke S, Gains H, Melendez-Torres G, Axford N, et al. Review of evidence on implementation in education. The Education Endowment Foundation (EEF). URL: https://educationendowmentfoundation.org.uk/education-evidence/evidence-reviews/implementation-in-education [accessed 2025-05-29]
42. 10 usability heuristics for user interface design. Nielsen Norman Group. 2024. URL: https://www.nngroup.com/articles/ten-usability-heuristics/ [accessed 2025-05-29]
43. Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M, Medical Research Council Guidance. Developing and evaluating complex interventions: the new Medical Research Council guidance. BMJ 2008 Sep 29;337:a1655 [FREE Full text] [doi: 10.1136/bmj.a1655] [Medline: 18824488]
44. Charlie Waller Trust. URL: https://charliewaller.org/ [accessed 2024-05-20]

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Research
CPD: continuing professional development
SORTS: Supportive Response to Self-Harm

Edited by A Mavragani; submitted 03.12.24; peer-reviewed by N Patel; comments to author 13.04.25; revised version received 17.04.25; accepted 09.05.25; published 02.06.25.

Please cite as:
Burn AM, Gains H, Anderson JK
A Self-Harm Awareness Training Module for School Staff: Co-Design and User Testing Study
JMIR Form Res 2025;9:e69309
URL: https://formative.jmir.org/2025/1/e69309
doi:10.2196/69309
PMID:

©Anne-Marie Burn, Hayley Gains, Joanna K Anderson. Originally published in JMIR Formative Research (https://formative.jmir.org), 02.06.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Leveraging Cognitive and Speech Ecological Momentary Assessment in Individuals With Phenylketonuria: Development and Usability Study of Cognitive Fluctuations in a Rare Disease Population

Shifali Singh[1], PhD; Lisa Kluen[1], PhD; Katelin Curtis[1], BA; Raquel Norel[2], PhD; Carla Agurto[2], PhD; Elizabeth Grinspoon[1], PhD; Zoe Hawks[1], PhD; Shawn Christ[3], PhD; Susan Waisbren[4], PhD; Guillermo Cecchi[2], PhD; Laura Germine[1], PhD

[1]McLean Hospital, Harvard Medical School, Belmont, MA, United States

[2]Digital Health, IBM Research, Yorktown Heights, NY, United States

[3]Department of Psychological Sciences, University of Missouri, Columbia, MO, United States

[4]Genetics and Metabolism Program, Boston Children's Hospital, Department of Pediatrics, Harvard Medical School, Boston, MA, United States

Corresponding Author:
Shifali Singh, PhD
McLean Hospital
Harvard Medical School
115 Mill Street, South Belknap
Belmont, MA, 02478
United States
Phone: 1 6178552675
Email: ssingh@mclean.harvard.edu

Abstract

Background: Phenylketonuria (PKU) is a rare, hereditary disease that causes disruption in phenylalanine (Phe) metabolism. Despite early intervention, individuals with PKU may have difficulty in several different cognitive domains, including verbal fluency, processing speed, and executive functioning.

Objective: The overarching goal of this study is to characterize the relationships among cognition, speech, mood, and blood-based biomarkers (Phe, tyrosine) in individuals with early treated PKU. We describe our initial optimization pilot results that are guiding this study while establishing the feasibility and reliability of using ecological momentary assessment (EMA) in this clinical population.

Methods: In total, 20 adults with PKU were enrolled in this study between December 2022 and March 2023 through the National PKU Alliance. Of the total, 18 participants completed an extended baseline assessment followed by 6 EMAs over 1 month. The EMAs included digital cognitive tests measuring processing speed, sustained attention, and executive functioning, as well as speech (semantic fluency) and mood measures. Participants had 60 minutes to complete the assessment.

Results: Completion rates of EMAs were above 70% (on average 4.78 out of 6 EMAs), with stable performances across baseline measures and EMAs. Between-person reliability (BPR) of the EMAs, representing the variance due to differences between individuals versus within individuals, is satisfactory with values close to (semantic fluency BPR: 0.7, sustained attention BPR: 0.72) or exceeding (processing speed: 0.93, executive functioning: 0.88) data collected from a large normative database (n=5039-10,703), as well as slightly below or matching a previous study using a clinical group (n=18). As applicable, within-person reliability was also computed; we demonstrated strong reliability for processing speed (0.87). A control analysis ensured that time of day (ie, morning, afternoon, and evening) did not impact performance; performance on tasks did not decrease if tested earlier versus later in the day (all P values >.09). Similarly, to assess variability in task performance over the course of all EMAs, the coefficient of variability was computed; 28% for the task measuring sustained attention, 37% for semantic fluency, 15.8% for the task measuring executive functioning, and 17.6% for processing speed. Performance appears more stable in tasks measuring processing speed and executive functioning than on tasks of sustained attention and semantic fluency.

XSL•FO
RenderX

Conclusions: Preliminary results of this study demonstrate strong reliability of cognitive EMA, indicating that EMA is a promising tool for evaluating fluctuations in cognitive status in this population. Future work should refine and expand the utility of these digital tools, determine how variable EMA frequencies might better characterize changes in functioning as they relate to blood-based biomarkers, and validate a singular battery that could be rapidly administered at scale and in clinical trials to determine the progression of disease.

(JMIR Form Res 2025;9:e63644) doi:10.2196/63644

KEYWORDS

neuropsychology; ecological momentary assessment; rare diseases; metabolism; cognition; phenylketonuria; PKU; hereditary; phenylalanine hydroxylase deficiency; phenylalanine

Introduction

Phenylketonuria and Cognition

Each year, 6 out of 100,000 newborns are diagnosed with phenylketonuria (PKU) [1]. PKU is a rare, hereditary disease [2,3] that, if left untreated, can lead to severe brain damage, intellectual disabilities, and behavioral issues [1,3,4]. PKU is characterized by a deficiency in the phenylalanine hydroxylase enzyme, necessary for the metabolism of the amino acid or phenylalanine (Phe) [3,5-7]. This causes disruptions in Phe metabolism [2] and deficiencies in tyrosine, or Tyr, with significant downstream effects on serotonin and dopamine [5,7]. Individuals with PKU tend to experience difficulties on tests measuring verbal fluency [2,8], processing speed [3,9], and executive functioning [5,6]. In addition, despite early intervention, patients with PKU also typically exhibit lower IQ scores, but to a lesser degree than executive functioning and processing speed [1,3,4,9].

PKU Biomarkers and Cognitive Functioning

Cognitive functioning in patients with PKU is greatly affected by fluctuations in blood Phe and blood Tyr levels. Although the relationship between Phe and cognitive functioning is more like a continuum, such that higher Phe is associated with greater impairment in cognitive functioning, in general, patients with PKU with Phe levels above 600 mmol/L demonstrate impaired cognition [3,4] and poor frontal lobe function [10]. Participants with even higher Phe levels, above 1000 mmol/L, scored lower on a greater number of cognitive tests assessing cognitive domains; this includes attention, verbal fluency, reaction time, verbal recognition memory, visual memory, and naming, compared with those with Phe levels lower than 1000 mmol/L [11,12]. Greater variability in Phe levels also appears to contribute to the severity of neurocognitive sequelae [3]. Indeed, previous work points to variability in Phe control as the strongest predictor of executive function and general cognitive outcomes, wherein Phe variability may be a better indicator of cognitive functioning than both metabolic control and age [3].

Assessing Fluctuations in Cognitive Status

Most previous studies of neurocognitive functioning in individuals with PKU have relied on traditional neuropsychological evaluations, which only capture a single time point of assessment, or a “snapshot” of cognitive functioning. This is not necessarily problematic, as this is consistent with standard clinical practice; however, these assessments are often limited in that individuals with PKU may experience fluctuations in cognitive status, wherein their cognition may shift within hours, days, weeks, and months. Thus, a typical evaluation is unlikely to capture this variance adequately [13]. Thus, we use a methodological approach known as ecological momentary assessment (EMA) in this study; this facilitates repeated assessments over time to capture intraindividual and interindividual variability [14-16]. EMA has been particularly effective with the widespread use of smartphones, allowing real-time capture of behavioral, psychological, and cognitive processes. Furthermore, it is a particularly [17] robust method of data capture given that it enables investigators to account for contextual and environmental factors, which are typical limitations of traditional research methodologies that most frequently capture functioning in a lab at a single time point, without any external distractions [18,19].

Previous work has demonstrated that EMA can be used to evaluate everyday fluctuations in *cognitive status*, enabling assessment of cognitive functioning in naturalistic contexts [20,21]. Yet this approach is just recently being incorporated into studies with patients diagnosed with PKU, despite EMA being an effective tool to measure everyday variability, or fluctuations, in cognitive status [3,17,22]. EMA is an ideal methodology for the study of rare disease populations like PKU, because it allows for the collection of multiple data points from each individual by increasing reliability and improves statistical power to detect clinically meaningful findings with small sample sizes—a key factor in rare disease populations. In addition, EMA study designs enable detailed analyses of multiple dynamic biological, cognitive, and psychological processes, such that one can effectively characterize specific clinical populations [23-25].

This Study

The primary overarching goal of this study is to characterize the relationships among cognition, speech, mood, and blood-based biomarkers (Phe, Tyr) in individuals with early treated PKU. This study, led by principal investigator SS of McLean Hospital of Harvard Medical School and funded by the Phenylalanine Families and Researchers Exploring Evidence (PHEFREE) Consortium, leverages EMA to evaluate real-time cognitive status, speech or voice biomarkers, and psychological functioning. Blood-based biomarkers, including Phe and Tyr, are also useful in contextualizing these relationships and were assessed on the days of EMA administration; further work on this project will aim to gather more Phe and Tyr data to better

XSL•FO
RenderX

establish psychometric robustness. Participants are continuing to be recruited for this study.

This paper reports results from an initial EMA optimization pilot of 20 participants diagnosed with PKU. This study was conducted before the completion of a longer protocol to determine the most optimal battery for individuals with PKU, including the number, length, and frequency of EMAs. We describe our initial optimization pilot results that are guiding the study design. We collected EMAs 6 times within a month; finger prick tests were completed on the day of each EMA to determine how variations in amino acid metabolism might relate to or predict fluctuations in certain aspects of functioning. This frequency was chosen to limit participant burden while still accounting for a wide range of data capture, varying between time of day and day of week over 1 month. Our goal is to determine whether the current method of evaluation, using EMA, is appropriate for individuals diagnosed with PKU, as well as how performance on cognitive tests might vary over time. This has implications for widespread dissemination of rapid, repeatable, scalable batteries that can be administered completely remotely, offering greater equity and accessibility in evaluating individuals with PKU on an international scale.

Methods

Participants

All participants were recruited through flyers distributed by the National PKU Alliance (NPKUA) through email with individuals in their patient registry and by postings on social media platforms associated with the organization. Recruitment specifically targeted individuals already enrolled in the NPKUA registry, a database that connects individuals diagnosed with PKU because it aligns with the study's objective of examining the cognitive and behavioral impacts of the rare disease. Recruitment materials provided detailed information about the study and instructions for those who expressed interest in participation. To be eligible, participants had to meet the following inclusion criteria: current US resident, normal or corrected-to-normal vision, ability to provide consent, and a diagnosis of PKU. Participants were excluded based on the following: significant physical disabilities affecting their ability to perform digital assessments (eg, due to visual, motor, or hearing impairments) or their inability to complete EMAs during the study period (eg, due to planned travel, night shift work, or occupation that does not allow time to complete assessments within 60 minutes). Potential participants were screened during an initial virtual visit to confirm eligibility. The study's purpose, procedures, risks, and expectations were thoroughly explained. Informed consent was collected electronically via a secure REDCap (Research Electronic Data Capture; Vanderbilt University) form before enrollment.

Materials

Baseline Assessment

Baseline cognitive tests, speech assessments, and psychological questionnaires were completed by all participants via their smartphones upon enrollment in the study. We used two platforms, including (1) TestMyBrain (TMB), which is an open-source, digital cognitive test platform with data collected from approximately 3 million participants worldwide [20,22,26-28], and (2) SurveyLex, developed by Sonde Health, which is a speech acquisition platform that enables voice recordings in response to predefined prompts or free response questions. This tool has been used to effectively validate voice biomarkers in other work [29]. Tasks were selected based on traditional neuropsychological evaluations, as well as speech measures typically used by collaborators on this project from the IBM Thomas J. Watson Research Center. In total, the baseline battery took approximately 90 minutes to complete and included "full versions" of all cognitive EMAs (refer to Table 1 for a complete list); of note, this is significantly longer than the ultrabrief, EMA versions of cognitive tests used for repeated assessments throughout the study. Please refer to Table 1 for a complete list of baseline assessments and their descriptions.

XSL•FO
RenderX

Table 1. Baseline assessments and constructs measured.

Questionnaire or assessment	Description
Baseline questionnaires (approximately 20 min)	
General questionnaire	Demographic characteristics, sleep and wake times in a typical work week, and employment.
PROMIS[a] Scale [30]	Questionnaire assessing self-reported anxiety and depression using the PROMIS Short Form v1.0 Anxiety 7a and PROMIS Short Form v1.0 Depression 8b scales. Together, they form a combined 15-item measure.
Global perceived stress scale [31]	Questionnaire assessing chronic experiences of stress. It is a 10-item scale measuring the degree to which situations are appraised as stressful. It takes approximately 5 min.
Quality of Life in Neurological Disorders (Neuro-QoL)—Cognitive Function Short Form [32]	Questionnaire assessing self-reported cognitive problems in daily life. It is an 8-item questionnaire and takes approximately 5 min.
Mental Health Questionnaire [33,34]	Questionnaire assessing cross-cutting symptoms for psychopathology based on the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition*. It is a 6-item questionnaire assessing possible broad psychopathology and takes approximately 2 min.
World Health Organization Alcohol, Smoking and Substance Involvement Screening Test (ASSIST) [35]	Screening for alcohol consumption, smoking, and other substance use over lifetime and the last 3 mo before the assessment. It takes approximately 3 min.
Snoring, tiredness, observed apnea, high blood pressure, BMI, age, neck circumference, and male gender (STOP-Bang) Questionnaire [36]	Questionnaire assessing obstructive sleep apnea risk. It is an 8-question measure and takes approximately 2 min.
Baseline cognitive assessment (approximately 60 min)	
TMB[b] website simple reaction time	Cognitive test assessing basic psychomotor speed. Participants press a button every time a green square appears on screen.
TMB vocabulary	Cognitive test assessing verbal reasoning. Participants indicate which of the 5 words is the closest in meaning to a target word.
TMB digit symbol matching (DSM)	Cognitive test assessing psychomotor processing speed. Participants match a set of symbols to the numbers 1, 2, or 3 based on a key presented on screen.
TMB gradual onset continuous performance test (gradCPT)	Cognitive test assessing sustained attention. Participants see a series of city or mountain scenes and are asked to press a button whenever they see a city scene and withhold a response whenever they see a mountain scene.
TMB choice reaction time (Choice RT)	Cognitive test assessing psychomotor processing speed. Participants indicate the direction of the one arrow that is a different color from the rest of the arrows.
TMB matrix reasoning	Cognitive test assessing general cognitive ability and nonverbal reasoning. Participants solve a series of visual puzzles.
TMB paced serial addition task (PSAT)	Cognitive test assessing sustained attention and working memory. Participants add pairs of numbers that appear one after another and determine whether the sum is >10 or <10.
TMB flicker change detection (Flicker)	Cognitive test assessing visual working memory. Participants view a series of visual scenes with blue and yellow dots. One of the dots is changing color from blue to yellow. Participants are asked to indicate the dot that is changing color.
TMB adaptive delay discounting	Cognitive test assessing decision-making. Participants indicate whether they would prefer differencing amounts of hypothetical money now vs. in the future.
TMB visual paired associates memory – learn	Cognitive test assessing visual memory. Participants learn a set of picture pairs.
TMB multiple object tracking (MOT)	Cognitive test assessing sustained visual attention. Participants remember and track a set of target circles as they move around the screen among a larger set of identical distractor circles.
TMB visual paired associates memory – test	Cognitive test assessing episodic memory. Participants indicate which pictures go together based on the set they learned.
TMB letter-number switching	Cognitive test assessing cognitive flexibility and task switching. Participants indicate which response fits the instruction cue shown on screen.
Baseline (and EMA[c]) voice survey (approximately 5 min)	
Short sentence	Speech test assessing speech abnormalities. Participants repeat the following sentence: “The quick brown fox jumps over the lazy dog.”
Sustained phonation	Speech test assessing vocal instability. Participants take a deep breath in and then say the vowel “aaa” for 30 s, taking breaks as needed.

XSL•FO
RenderX

Questionnaire or assessment	Description
Diadochokinetic task	Speech test assessing articulation. Repeat “pah-tah-kah“ as many times as you can in 10 s.
Paragraph (amusement park)	Speech test assessing speech patterns and abnormalities. Participants are instructed to read the following paragraph about a day at the amusement park.
Feeling question	Question asking the participant to share how they are feeling and why for 1 min.
Semantic fluency	Speech test assessing verbal fluency. Participants are given 1 minute to come up with as many words as they can that fit the given category.

[a]PROMIS: Patient-Reported Outcomes Measurement Information System.

[b]TMB: TestMyBrain.

[c]EMA: ecological momentary assessment.

Cognitive EMA, Speech EMA, and Blood-Based Biomarkers

Overview

Ultrabrief versions of the TMB full length cognitive tests were selected for the optimization pilot; these versions of the full-length tests demonstrate robust reliability and good sensitivity [37]. These specific cognitive tests were chosen based on their domain of assessment; that is, given processing speed and executive functioning are implicated in PKU, brief versions of cognitive tests measuring these domains were chosen for the EMA study portion. All tests were developed using a combination of JavaScript and HTML, delivered through web apps that were downloaded to the participants local device, ran in the browser, and then delivered data back to a central server. Analyses required participants to complete at least 4 assessments, 1 of which was the baseline assessment. A brief description of the tests used are below, with further detailed information and psychometric characteristics described in Germine et al [38] and Singh et al [28].

TMB Digit Symbol Matching

In digit symbol matching (DSM) [22,26,39], participants are presented with 6 symbols, each of which are paired with a single digit between 1 and 3 (ie, 2 symbols were paired with each digit). These pairings remain visible throughout the duration of the test. Individual probe symbols are sequentially presented above these pairings, to which patients respond by selecting the corresponding digit as quickly as possible. Each probe symbol remains visible until the patient makes a response. Scores are recorded as the total number of correct responses in 90 seconds. The primary test score of interest used for psychometric analyses is number of correctly completed matches (DSM.score).

TMB Gradual Onset Continuous Performance Test

In gradual onset continuous performance test (gradCPT) [40], the participant presses a key when a city image appears and does not press it when a mountain image appears. Images rapidly transition from one to the next, with mountains appearing only 10%-20% of the time. Scores are recorded as a measure of response bias where a larger value indicates greater response impulsivity, or tendency to press a key regardless of the picture type. The primary test score of interest and used for psychometric analyses is response bias (CPT.dprime).

TMB Multiple Object Tracking Test

In multiple object tracking (MOT) test [41], the participant remembers and tracks a set of target circles as they move around the screen, among a larger set of identical distractor circles. The primary test score of interest and used for psychometric analyses is percent correct (MOT.score).

Speech EMA includes the same speech tests used in the baseline voice survey (refer to Table 1 for a comprehensive list, tests’ relative duration, and response style, ie, repetition or free response).

To determine Phe and Tyr levels, participants were supplied with test kits from PerkinElmer that were mailed to their homes. Participants provided blood samples that were mailed directly to PerkinElmer after obtaining a fasting sample via finger prick on the morning of their scheduled EMA.

In addition to cognitive EMA, speech EMA, and blood biomarkers, we also collected passive measures, including metadata about browser, screen size, and operating system. This information was used in data analysis to ensure consistent data quality throughout participant responses.

Procedure

This study was compliant with ethical principles and approved by the Mass General Brigham (MGB) Institutional Review Board (IRB) and the NPKUA Ethics Committee. All participants completed an orientation and signed the informed consent form through the secure REDCap platform. In total, 23 participants were recruited through flyers distributed by the NPKUA. Participants were asked to complete 6 EMAs over 1 month, varying by day of the week (weekday vs weekend) and time of day (morning, afternoon, or evening). Ultrabrief versions of selected baseline tasks, based on areas of functioning typically impaired in individuals with PKU, were administered on a mobile device at varying times throughout the day to minimize participant burden. More specifically, EMAs were sent on predetermined days throughout the month (week 1: Wednesday; week 2: Tuesday and Friday; week 3: Monday and Sunday; and week 4: Thursday). All participants followed the same EMA schedule in Eastern Standard Time: EMA 1 at 10:13 AM, EMA 2 at 10:13 AM, EMA 3 at 7:45 PM, EMA 4 at 12:05 PM, EMA 5 at 1:46 PM, and EMA 6 at 5:57 PM. There was a minimum of 3 days and a maximum of 6 days between EMAs. All EMAs were delivered between 9 AM and 9 PM local time to reduce interference with participants’ daily schedules and sleep routines. All participants completed EMAs on their personal

XSL•FO
RenderX

devices (smartphones). They were routinely sent push notifications that enabled them to access the assessments via a web link; they had 60 minutes to complete the assessment. Participants were compensated US $300 (per EMA and blood collection). The between-person reliability of each EMA task was evaluated, which is especially relevant given that each task was administered in real-life settings with a clinical population. Refer to Figure 1 for the overall study design.

Figure 1. Study design. Overview of the study design for participants with phenylketonuria. EMA: ecological momentary assessment; PKU: phenylketonuria.

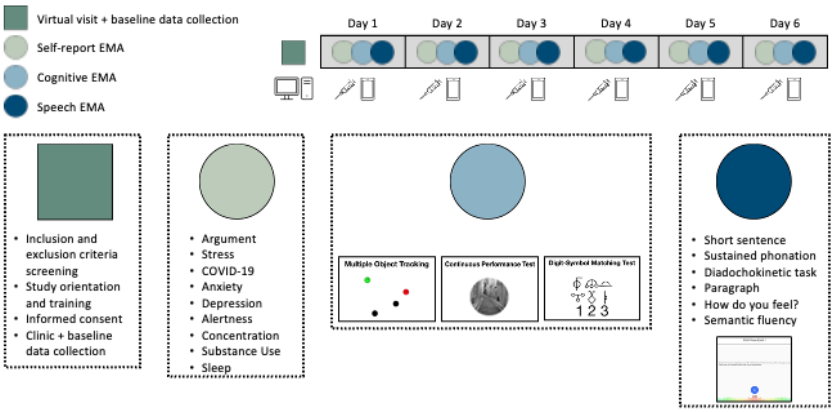

The overall study design is the same as the optimization pilot discussed in this paper. Conducted over 1 month, the study includes a virtual prestudy visit conducted over Zoom (screening for inclusion and exclusion criteria, orientation, informed consent, and baseline assessment). From Days 1-6, participants complete repeated EMAs across three domains: (1) self-report EMAs that track variables such as argument occurrence, stress, COVID-19 impact, anxiety, depression, alertness, concentration, substance use, and sleep, (2) cognitive EMAs that include tasks like MOT, gradCPT, and DSM to measure cognitive function, and (3) speech EMAs include tasks such as short sentence repetition, sustained phonation, diadochokinetic exercises, paragraph reading, responses to “How do you feel?” prompts, and semantic fluency tasks.

Ethical Considerations

This study adhered to ethical guidelines for human participants research as outlined by McLean Hospital, Harvard Medical School, and MGB. Before study initiation, a formal review of the research protocol was conducted by the MGB IRB. Based on the nature of the study, the IRB determined that no exemptions were applicable. Once the study documents were approved (IRB approval number 2024P001922), all research activities were conducted in full compliance with institutional policies and federal regulations to uphold ethical standards.

Before providing consent, participants received a detailed informed consent form that outlined key components of the study, such as the purpose, procedures, risks, benefits, confidentiality, and privacy rights. Participants were given adequate time to review the consent form, ask questions, and address concerns with study staff before providing consent. The consent process was conducted virtually through the secure online platform REDCap.

To safeguard participant privacy and confidentiality, each participant was assigned a unique participant ID (eg, PKU1) generated in REDCap during the virtual consent process. Identifiable information was securely stored in password-protected files, accessible only to authorized study staff, and all data were deidentified. All TMB data were automatically backed up nightly, with access restricted to authorized users. Data collected via SurveyLex, a Health Insurance Portability and Accountability Act (HIPAA)–compliant software administered by Sonde Health, has been approved for use in other MGB IRB protocols (eg, 2019P003458 and 2019P002752). Any data that our team retrieved from the SurveyLex website were encrypted in transit and stored securely in a designated Partners Dropbox folder.

Participants were compensated up to US $300 for their participation in the study. Compensation included US $10 for each brief daily assessment (6 assessments × US $10 = US $60) and US $40 for each blood collection (6 collections × US $40 = US $240). Payments were issued via checks mailed to the participants’ homes.

Statistical Analyses

EMA and baseline tasks were parsed using Python 3.11 via Spyder IDE 5.4.3. Analysis was done using R 4.3.1 [42] in RStudio 2023.06.1+524 [43], with the packages tidyverse [44], plyr [45], and psych [46]. Plots were made with ggplot, part of tidyverse, as well as patchwork [47]. The between-person reliability of each EMA task was evaluated, which is especially relevant given that each task was administered in real-life settings with a clinical population. The between-person reliability, which represents the variance that is due to differences between individuals versus within individuals, was assessed using 2 different approaches: the mlr (multilevel reliability) function from the psych package in R (for gradCPT, MOT, and DSM) and a regression-based approach (for semantic fluency, due to the missing trial structure), modeled after Mascarenhas Fonseca et al [37]. The mlr function computes reliability as well as generalizability, using unconditional multilevel mixed models to predict performance for each EMA, with nested within-person random effects (between-person reliability: mlr ‘RkRn’; within-person reliability mlr ‘Rcn’ [24]). For gradCPT EMAs, response bias (dprime) was calculated for odd and even trials, and performance was predicted for the other half of the EMA. For DSM, scores were also calculated based

XSL•FO
RenderX

on odd and even trials [27]. For MOT, scores were calculated on a trial-by-trial level [21] based on previous works [21,48]. EMA number and trial number (in case of MOT) or for both halves of the task (gradCPT and DSM) were coded from –2.5 to 2.5.

For the regression-based approach, using the lmer function in r, the calculated test scores were used and entered into an unconditional multilevel mixed model (equation 1) to predict scores on each EMA. Fitting the model allowed partitioning of variance, which could then be entered into equation 2 [37]:

semanticFluency<–lmer(score~1 + (1 + EMA_num|participant),data = df) **(1)**

EMA_num was a vector coded from –2.5 to 2.5 for EMAs 1 to 6.

Where Var (BP) is the total variance in scores between participants, Var (WP) is the variance in scores within participants (ie, variance between EMA sessions and residual variance), and n is the total number of measurements (in this case, the mean number of EMAs completed across all participants) [48]. Between-person reliability was reported for each EMA task, together with average completion rate and average EMA performance.

To assess variability in task performance (in percent) over the course of the EMAs, the coefficient of variability (CV) was calculated (equation 3) for each of the 3 tasks.

To make sure that we understand if performance was impacted by time of day, we used *t* tests to compute whether there was a significant difference in performance between morning EMAs (EMAs 1 and 2), evening (EMA 3 and 6), and morning and midday EMAs (EMA 4 and 5).

Finally, for baseline tasks, average performance was reported. Notably, although we collected self-reported mood and psychological data, given the small sample size, further data will need to be collected to sufficiently power additional analyses regarding the relationship among EMA and mood or self-report.

Results

Participants

In total, 21 adults with PKU were enrolled in this study between December 2022 and March 2023 (Table 2 illustrates descriptive statistics); however, only 20 participants were included in the EMA analyses due to 1 participant's participation in a very similar study that was not disclosed before enrollment. Following the exclusion of that participant, 2 participants had missing baseline data, so 18 total participants were included in analyses. All 20 participants were diagnosed with PKU at or around the time of birth, and diagnosis details were self-reported. Among the participants, 19 had classic PKU (Phe >1200 mmol/L), 1 had hyperphenylalanemia (Phe 120-600 mmol/L), and 3 were unsure of their specific diagnosis type.

XSL•FO
RenderX

Table 2. Description of study sample (N=20).

Characteristic	Values
Age (years), mean (SD)	37.24 (11.33)
Sex, n (%)	
Female	12 (60)
Male	8 (40)
Race or ethnicity, n (%)	
American Indian or Alaskan Native	0 (0)
Asian	0 (0)
African or Black	0 (0)
Native Hawaiian or Pacific Islander	0 (0)
European or White	20 (100)
Hispanic or Latino	0 (0)
Education, n (%)	
Primary school (less than 7 years)	0 (0)
Middle or junior high school (7-10 years)	0 (0)
Secondary school (high school diploma or GED[a])	2 (10)
Some college or University	3 (15)
Technical training or associate degree	2 (10)
Bachelor's degree	5 (25)
Master's degree	5 (25)
Graduate or professional degree (eg, PhD, MD, and JD)	3 (15)

[a]GED: General Educational Development.

Baseline Results

Participants performed several baseline tasks that will be used for further analysis in subsequent publications as the study continues. Descriptive results are reported in Table 3 and presented alongside normative data from the TMB database for each test, respectively. Data in the table are reported as mean (SD) across the entire participant sample (for this study data as well as the normative sample). This is true for all score and accuracy data but also for the mean and median reaction times data.

XSL•FO
RenderX

Table 3. Baseline descriptive results.

Cognitive test and task measure	PKU[a] pilot sample		TMB[b] normative data	
	Participants, n	Result, mean (SD)	Participants, n	Result, mean (SD)
TMB MOT[c]	18		1931	
Accuracy		78.55 (9.32)		79.86 (10.22)
Score		56.56 (6.71)		57.5 (7.36)
TMB letter or number switching[d]	18		3421	
Accuracy		98.55 (1.96)		96.46 (0.66)
Mean RTc[e]		1169.44 (270.50)		1252.61 (364.43)
Median RTc		1060.74 (249.63)		1134.07 (349.32)
TMB delay discounting[f]	18		24,081	
ln(k)[g]		–6.22 (1.68)		–4.73 (2.18)
TMB choice RTc	18		34,641	
Score		15.32 (21.22)		11.22 (3.51)
Accuracy		87.41 (27.21)		93.44 (13.48)
Mean RTc		1009.82 (252.23)		969.07 (317.23)
Median RTc		964.28 (222.79)		912.02 (290.51)
TMB flicker[h]	18		13,363	
Score		11.61 (3.72)		10.79 (6.11)
Accuracy		96.58 (9.19)		94.31 (0.93)
RTc Mean		6927.79 (1961.00)		7271.13 (2347.85)
RTc Median		5859.93 (1973.97)		6028.51 (2314.93)
flipscMean[i]		9.42 (2.85)		9.90 (3.37)
flipscMedian		7.91 (2.78)		8.13 (3.32)
TMB paced serial addition task[j]	18		3835	
Score		46.44 (7.72)		44.59 (11.03)
Accuracy		77.41 (12.86)		74.32 (11.03)
Mean RTc		881.82 (146.04)		825.09 (181.44)
Median RTc		852.79 (161.77)		797.57 (196.34)
TMB SRT[k]	18		60,874	
Mean RTc		341.40 (111.7)		317.67 (72.05)
Median RTc		323.08 (94.72)		304.35 (67.70)
TMB visual paired associates memory task[l]	18		9758	
Score		15.5 (5.73)		15.86 (4.57)
Accuracy		63.23 (23.41)		66.1 (19.04)
Mean RTc		3941.48 (819.75)		3850.91 (840.13)
Median RTc		3639.84 (843.8)		3561.11 (972.49)
TMB matrix reasoning[m]	18		25,829	
Score		26.06 (4.77)		27.26 (5.91)
Accuracy		74.44 (13.63)		82.80 (0.9)
Mean RTc		7275.12 (2209.59)		10,098.3 (5905.12)
Median RTc		5039.36 (1366.61)		6056.32 (2829.13)

XSL•FO
RenderX

Cognitive test and task measure	PKU[a] pilot sample		TMB[b] normative data	
	Participants, n	Result, mean (SD)	Participants, n	Result, mean (SD)
TMB vocabulary[n]	18		36,230	
Score		26.78 (1.59)		23.22 (5.52)
Accuracy		86.7 (5.96)		77.4 (18.41)
Mean RTc		3544.74 (778.43)		7279.86 (3607.23)
Median RTc		3164.19 (856.14)		5718.68 (2520.94)
TMB gradCPT[o]	18		4669	2.70 (0.76)
dprime[p]		2.64 (0.74)		

[a]PKU: Phenylketonuria.

[b]TMB: TestMyBrain website [49].

[c]MOT: multiple object tracking (visuospatial attention and visual working memory). A higher score or accuracy indicates better performance.

[d]Letter or number switching: switching between 2 tasks, testing response selection or inhibition. Higher accuracy indicates better performance.

[e]RTc: reaction time (processing speed and response selection or inhibition). A higher score indicates higher processing speed. A higher accuracy indicates better performance. The measures are shown in ms.

[f]Delay discounting: adaptive delay discounting, choosing between smaller immediate and larger delayed rewards (temporal discounting and impulsivity). Delay discounting is measured using the natural logarithm of the discounting factor k (ln(k)). ln(k) is negative for small values, indicating more future-oriented individuals, and positive for larger k values, indicating more impulsive, immediate-reward focused individuals.

[g]ln(k): natural logarithm of the discounting factor k.

[h]Flicker: flicker change detection (visual search, change detection, and visual working memory). A higher accuracy indicates better performance. The same is true for the score.

[i]number of image flips for correct responses to “test” trials (ms).

[j]Paced serial addition: adding pairs of numbers, sustained attention, working memory. A higher score (number of correct trials) and accuracy (proportion of correct trials) indicate better performance.

[k]SRT: simple reaction time (psychomotor response speed). The mean and median correct response times are indicated, and shorter response times indicate faster processing speed.

[l]Visual paired associates memory: visual memory, episodic memory (remembering pictures). A higher score (number of correct trials) and accuracy (proportion of correct trials) indicate better performance.

[m]Matrix reasoning: fluid cognitive ability and nonverbal reasoning. A higher score (number of correct trials) and accuracy (proportion of correct trials) indicate better performance.

[n]Vocabulary: identifying synonyms (crystallized cognitive ability and verbal reasoning). A higher score (number of correct trials) and accuracy (proportion of correct trials) indicate better performance.

[o]gradCPT: gradual onset continuous performance test. Performance is measured using dprime. Dprime can be interpreted as the discrimination sensitivity in the task, with higher values indicating a better ability to perform the task.

[p]dprime: response bias.

Biomarkers

The averages for Phe, Tyr, and the Phe:Tyr ratio were recorded daily for each scheduled EMA. These blood-based biomarkers were measured in micromoles per liter (mmol/L), as illustrated in Table 4.

XSL•FO
RenderX

Table 4. Phenylalanine, tyrosine, and phenylalanine:tyrosine ratio averages by day of ecological momentary assessment administration.

Day	Values, mean (SD)
Day 1 (n=17)	
Phe[a]	425.18 (429.43)
Tyr[b]	43.92 (19.13)
Phe:Tyr ratio	10.66 (11.13)
Day 2 (n=18)	
Phe	480.61 (504.59)
Tyr	42.41 (11.44)
Phe:Tyr ratio	11.76 (12.98)
Day 3 (n=17)	
Phe	482.98 (448.82)
Tyr	42.41 (14.61)
Phe:Tyr ratio	11.52 (10.53)
Day 4 (n=17)	
Phe	478.17 (456.04)
Tyr	39.47 (13.19)
Phe:Tyr ratio	13.02 (10.02)
Day 5 (n=17)	
Phe	515.36 (432.93)
Tyr	38.36 (11.83)
Phe:Tyr ratio	12.61 (10.20)
Day 6 (n=14)	
Phe	524.23 (360.14)
Tyr	43.91 (12.33)
Phe:Tyr ratio	11.85 (6.85)

[a]Phe: phenylalanine.

[b]Tyr: tyrosine.

EMA Results

Participants were prompted to complete 4 tasks as part of each of the 6 EMAs (“measurement time points”) in the study. On average, each task was completed between 4.78 times out of the 6 measurement time points. Between-person reliability (ie, the consistency of the differences in scores between individuals [27]) was slightly lower or comparable to previous EMA studies [27,40] specifically for the TMB MOT and TMB DSM, and comparable with or exceeding a representative sample from TestMyBrain [40,49]. However, while reliability was slightly lower for the TMB gradCPT (0.72) and semantic fluency (0.70), they still fall within a good range (reliability between 0.4 and 0.59 is considered fair, 0.60 and 0.74 good, and reliability above 0.75 excellent). Results are listed in Table 5.

XSL•FO
RenderX

Table 5. Initial reliability data for ecological momentary assessments, based on the data collected from the phenylketonuria pilot sample.

Outcome	PKU[a] pilot sample			
	Participants, n	Mean (SD)	Between-person reliability of the EMA[b]	Within-person reliability of the EMA
Brief TMB[c] gradCPT[d]—dprime[e]	20	2.66 (0.75)	0.72	0
Brief TMB DSM[f]—score (#correct)	20	23.18 (4.07)	0.93	0.87
Brief TMB MOT[g]—accuracy	20	71.37 (22.13)	0.88	0
Brief semantic fluency—score	20	20.97 (7.91)	0.70	__[h]

[a]PKU: phenylketonuria.
[b]EMA: ecological momentary assessment.
[c]TMB: TestMyBrain website [49].
[d]gradCPT: gradual onset continuous performance test.
[e]dprime: response bias.
[f]DSM: digit symbol matching.
[g]MOT: multiple object tracking.
[h]Not available (due to missing trial structure, this measure cannot computed for this task, because there is no variance within each individual ecological momentary assessment).

Figure 2 provides an overview of participant performance in all 4 tasks that were administered as part of an EMA, over the course of the 6 EMAs. Boxplots depict the data distribution, with the median shown as a horizontal line, the lower and upper quartile values as the edges of the box, while the whiskers represent the minimum and maximum data values. The blue points represent outliers (defined as the standard to be above or below the upper or lower quartile value minus 1.5 times the IQR) and the red points show individual participant values.

Figure 2. Ecological momentary assessment performance across cognitive tasks. dprime: response bias; DSM: digit symbol matching; EMA: ecological momentary assessment; gradCPT: gradual onset continuous performance test; MOT: multiple object tracking; TMB: TestMyBrain website.

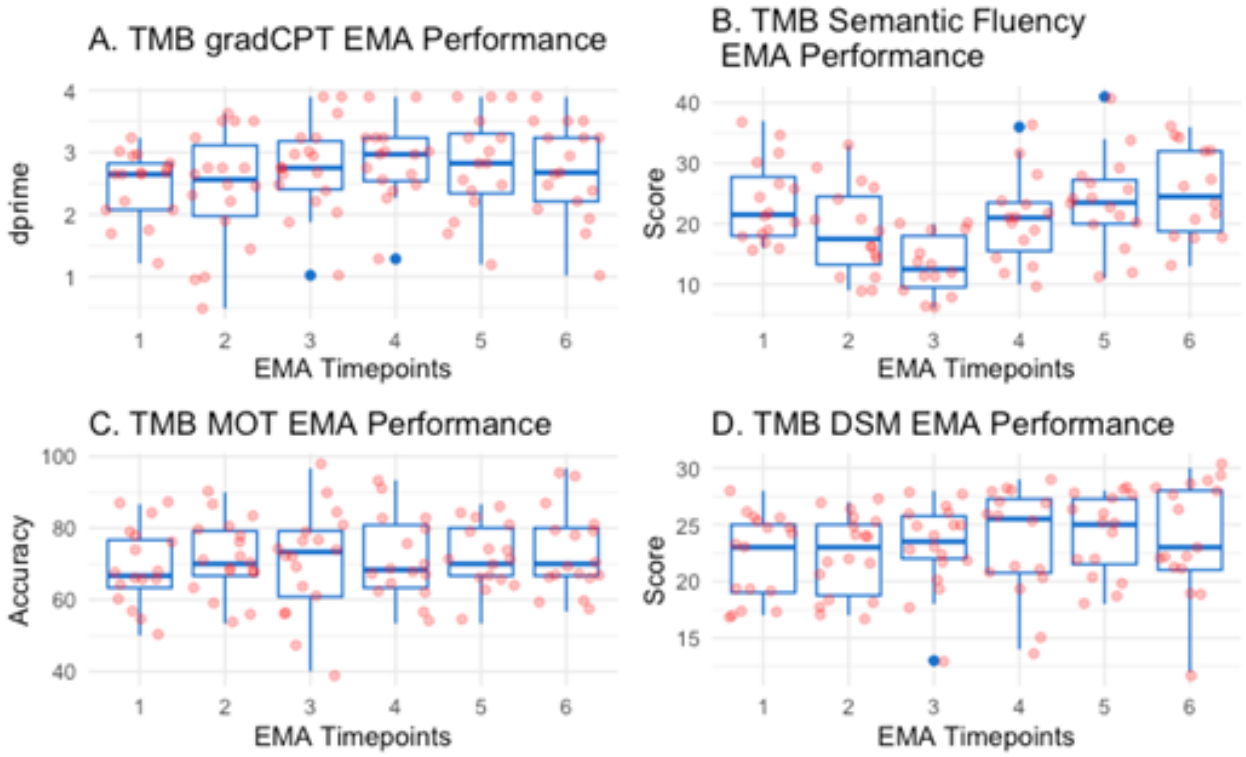

Overview of participants' performance on individual cognitive tasks administered as part of each EMA conducted over 1 month. The x-axes represent EMA time points from day 1 to day 6, and the y-axes represent participant performance for each task. Panel A shows performance on the gradCPT (n=20), with dprime values shown on the y-axis. The gradCPT assesses sustained attention, and participants are required to distinguish between target and nontarget stimuli. Dprime can be interpreted as the

XSL•FO
RenderX

discrimination sensitivity in the task, with higher values indicating a better ability to perform the task. Panel B shows performance on the semantic fluency task (n=20), where the y-axis represents participants' ability to produce words within a specific category (eg, animals) during a set time frame. A higher score indicates better semantic fluency. Panel C shows performance on the MOT (n=20), with the y-axis reflecting participants' accuracy. The MOT assesses speeded visual attention, and participants are asked to track a set of target circles. A higher number indicates better accuracy, meaning how well participants tracked objects. Finally, panel D shows performance on the Digit Symbol Matching (DSM; n=20), with the y-axis representing the number of correctly completed matches. The DSM measures processing speed, where participants match symbols to corresponding digits as quickly as possible. The EMAs were sent on predetermined days throughout the month (week 1, EMA 1: Wednesday; week 2, EMA 2 and 3: Tuesday and Friday; week 3, EMA 4 and 5: Monday and Sunday; and week 4, EMA 6: Thursday). All participants followed the same EMA schedule in Eastern Standard Time: EMA 1 at 10:13 AM (gradCPT, MOT, and DSM, 17 out of 20 participants completed, verbal fluency: 16 out of 20 participants completed EMA), EMA 2 at 10:13 AM (gradCPT, MOT, and DSM, 18 out of 20 participants completed, verbal fluency: 16 out of 20 participants completed EMA), EMA 3 at 7:45 PM (gradCPT, MOT, and DSM, 18 out of 20 participants completed, verbal fluency: 14 out of 20 participants completed EMA), EMA 4 at 12:05 PM (gradCPT, MOT, and DSM, 16 out of 20 participants completed, verbal fluency: 15 out of 20 participants completed EMA), EMA 5 at 1:46 PM (gradCPT, MOT, and DSM, 16 out of 20 participants completed, verbal fluency: 16 out of 20 participants completed EMA), and EMA 6 at 5:57 PM (gradCPT, MOT, and DSM, 17 out of 20 participants completed, verbal fluency: 14 out of 20 participants completed EMA). There was a minimum of 3 days and a maximum of 6 days between EMAs. A total of 12 participants resided in the Eastern Time Zone, 4 in the Central Time Zone, 1 in the Mountain Time Zone, and 3 in the Pacific Time Zone.

Quantitative assessment of variability indicates a coefficient of variability of 28% for the gradCPT, 37% for semantic fluency, 15.8% for the MOT, and 17.6% for the DSM. Hence, performance is more stable in those tasks measuring processing speed, visual short-term memory, and visuospatial attention (DSM and MOT), while performance on the gradCPT (measuring sustained attention, response inhibition, and cognitive control) as well as semantic fluency (also measuring executive control) is much more variable over time. The semantic fluency prompt (ie, "name everything you can think of in this category") was changed for each EMA, and therefore greater variability was expected for this speech-based task. This task also represents domains (eg, verbal fluency and aspects of executive functioning) known to be difficult for patients with PKU [50]. Performance in the TMB gradCPT is measured using dprime, which indicates the discrimination sensitivity in the task. Higher values reflect a better ability to perform the task. Performance in the TMB MOT is measured using accuracy, which assesses the ability to correctly track objects. Median performance across all EMAs is at 70%, indicating participants' general ability to do well in the task. Furthermore, TMB DSM performance is measured using the DSM score, which indicates the number of correctly identified matches in the task. Higher values indicate a higher processing speed and better visual short-term memory. Median performance appears stable across all EMA measurements, with little fluctuation.

Finally, we assessed whether performance was significantly different between the morning (EMA 1 and 2) and evening (EMA 3 and 6) and morning and midday (EMA 4 and 5) EMAs. Results showed that for the DSM, there was no significant difference in performance between the morning and evening ($t_{13.72}$=1.46, P=.17) or morning and midday EMAs ($t_{13.90}$=–0.98, P=.34; average performance morning: mean 23, SD 3.76; midday: mean 24.94, SD 4.11; evening: mean 25.56, SD 3.26). For semantic fluency, we again did not see a difference in mean performance between morning (mean 19.2, SD 4.73) and evening (mean 19.3, SD 4.75) EMAs (t_8=–0.033, P=.97) and morning and midday (mean 20.4, SD 5.18) EMAs ($t_{7.93}$=–0.38, P=.71). A similar pattern emerges for the MOT, with no significant difference between morning (percent correct: mean 71.46, SD 19.35) and evening (mean 76.04, SD 19.54) EMAs (t_{94}=–1.16, P=.25) or morning and midday (mean 74.79, SD 19.46) EMAs (t_{94}=–0.84, P=.40). Finally, the same is true for the comparison between morning (dprime: mean 2.83, SD 0.32) and evening (mean 3.07, SD 0.51) EMAs in the gradCPT ($t_{11.68}$=–1.13, P=.29), and between the morning and midday (mean 3.17, SD 0.40) performance ($t_{13.25}$=–1.86, P=.09). Hence, performance does not decrease when EMAs are administered later in the day compared with earlier (all reported t tests are 2-tailed).

Attrition

In total, 21 participants were recruited through the NPKUA. Furthermore, 1 participant was removed because of a conflict of interest (they were a part of another study with a laboratory we are collaborating with at the University of Missouri-Columbia). All participants completed the study in its entirety. Regarding EMAs, completion rates were around 70% (between 4.6 and 5 measurements out of 6); 1 participant only completed 1 EMA and 2 blood samples.

Discussion

Principal Findings

Individuals with PKU face a large health care burden given that they endure multiple hospital visits from childhood, and must consistently monitor diet and blood levels, without the advent of continuous monitoring digital tools [51,52]. These individuals are more likely to experience chronic conditions of organ systems, further increasing their health care burden [52,53]. In addition, given that PKU is inherently a rare disease, there is often a dearth of specialty providers who can routinely monitor patients with the disease; thus, the time burden on caregivers and patients alike is further exacerbated [51]. Studies on this population need to focus on scalable, accessible ways of remotely monitoring individuals with PKU, so they can achieve a better quality of life with less health care burden. Thus, EMA studies in this population fill a unique gap, wherein providers and researchers alike can continuously monitor patients at scale,

XSL•FO
RenderX

enabling greater access to individuals with PKU both nationally and internationally. Smartphones and the widespread availability of personal digital devices further facilitate larger-scale studies that can readily incorporate novel or experimental measures as they are developed.

The goal of this pilot study was to (1) demonstrate that EMA is a valid and reliable methodology for evaluating fluctuations in cognitive status in individuals diagnosed with PKU and (2) optimize a test battery that can be iterated on in the larger protocol based on these results. Given that this is the first study examining cognitive and speech EMA in individuals diagnosed with PKU, there were limited previous studies on which to develop appropriate frequency or timepoint decisions. Therefore, we turned to other clinical studies that use EMA in clinical and community samples [24,37]; based on the results demonstrating robust reliability (between-person for all tests and within-person for processing speed), it appears that this frequency can be deemed appropriate for this population.

Results suggest that EMAs were completed adequately well in this clinical sample, with completion rates above 70% (4.78 measurements out of 6). Furthermore, performance in both EMA measures and baseline tests appears in line with expectations based on expansive normative data, both for community and clinical samples, which demonstrate participants can complete these types of tasks with relatively low attrition or participant burden. Reliability is also close to what has been recorded in other samples, though slightly lower for the gradCPT and semantic fluency (0.7 and above), with the latter, however, representing a domain that is known to be specifically difficult for patients with PKU [53]. The TMB DSM and TMB MOT tests demonstrate the strongest between-person reliability in this sample, consistent with previous studies and the TMB representative sample [37].

Notably, there appeared to be small “dips” in semantic fluency on EMA 3; however, this qualitative decline cannot only be accounted for by time of assessment (ie, evening), given that EMA 6 was also in the evening, and time of day did not significantly impact performance. Furthermore, we did not see a significant difference between EMA 3 (mean 15.6, SD 3.85) and 6 (mean 23, SD 7.12; $t_{6.16}$=–2.05, P=.09). Similarly, TMB MOT appears to be somewhat more variable during EMA 3 as well, so we compared variability in performance on EMA 3 versus EMA 6, given that they are both administered in the evening, and there was no significant difference in performance ($t_{93.98}$=–0.28, P=.78) between EMA 3 (mean 75.41, SD 21.92) and EMA 6 (mean 76.67, SD 21.57). In addition, performance on the MOT during EMA 3 was in no way significantly different (Fligner-Killeen test of homogeneity of variances: P=.93). Therefore, it appears that there may be natural peaks and troughs in performance that, although not statistically significant, strengthen the argument for this methodology, which enables one to assess fluctuations in performance over one full month.

Limitations

Future work might examine whether higher-frequency EMAs could potentially offer greater utility in capturing nuanced changes in everyday cognition. Previous work by Mascarenhas Fonseca et al [37] has demonstrated that 3 EMAs per day for 10 days was particularly effective in capturing changes in glucose levels in individuals diagnosed with type 1 diabetes; this frequency should be examined within the PKU population, with finger prick tests completed at the time of EMA, rather than first thing in the morning, as done in this study, but no conclusions were drawn based on blood-based biomarkers given the limited range of Phe or Tyr. Ongoing analyses are determining whether any specific changes in blood levels within individuals may offer some interesting insights in relation to fluctuations in cognitive status. In addition, given the small sample size and relatively (racially or ethnically) homogeneous population, it is difficult to determine the generalizability of the results. All participants were recruited through the NPKUA, and therefore special efforts aimed at encouraging diversity were limited. Future work might attempt to recruit individuals with PKU from the broader community to better capture diversity. As this study continues, obtaining a larger sample size, though difficult in a rare disease population, will be critical in determining how cognition, other speech characteristics (besides semantic fluency), and blood-based biomarkers (Phe, Tyr) interact. Collecting this information will facilitate a data-driven approach in streamlining and refining the battery used in this study.

Conclusion

The EMA pilot study described in this manuscript demonstrated the psychometric reliability and feasibility of EMA studies in individuals with PKU. By leveraging digital tools, EMA offers the ability to remotely capture everyday cognitive functioning, outside of a single time point of assessment. The digital nature of EMA batteries facilitates entirely remote test administration, enabling more rapid and scalable patient monitoring while improving equity and accessibility, particularly in clinical trials interested in outcome-based research. Future projects may focus on validating a singular battery that could be rapidly administered at scale and in clinical trials to determine the progression of disease, with or without pharmacological intervention.

Acknowledgments

This research was funded through a research fellowship administered by the Phenylalanine Families and Researchers Exploring Evidence (PHEFREE) Consortium. The PHEFREE Consortium (U54HD100982) is a part of the National Institutes of Health Rare Disease Clinical Research Network, supported through collaboration between the National Center for Advancing Translational Science, the Eunice Kennedy Shriver National Institute of Child Health and Human Development, the National Institute of Neurological Disorders and Stroke, and the Office of Dietary Supplements. The authors confirm independence from the funders; the content of the article has not been influenced by the funders.

XSL•FO
RenderX

Data Availability

The datasets generated during or analyzed during this study are not publicly available due to participant confidentiality and privacy concerns but are available from the corresponding author on reasonable request.

Conflicts of Interest

SS is a Venture Associate with a venture capital fund that invests in medical technology companies. SW consults for BioMarin Pharmaceutical and Sanofi regarding functional end points for clinical trials in phenylketonuria. SC has received consulting fees and honoraria from BioMarin Pharmaceutical, Synlogic Therapeutics, and Jnana Therapeutics. Biomarin has also funded current and past research in SC's laboratory. LG is the president of the Many Brains Project, a nonprofit organization that supports open-source digital cognitive assessment. ZH has received consulting fees from Blueprint Health and Kintsugi Health. No authors reported biomedical financial interests or potential conflicts of interest.

References

1. Romani C, Olson A, Aitkenhead L, Baker L, Patel D, Spronsen FV, et al. Meta-analyses of cognitive functions in early-treated adults with phenylketonuria. Neurosci Biobehav Rev 2022;143:104925 [FREE Full text] [doi: 10.1016/j.neubiorev.2022.104925] [Medline: 36283539]
2. Hawks ZW, Strube MJ, Johnson NX, Grange DK, White DA. Developmental trajectories of executive and verbal processes in children with phenylketonuria. Dev Neuropsychol 2018;43(3):207-218 [FREE Full text] [doi: 10.1080/87565641.2018.1438439] [Medline: 29432026]
3. Hood A, Grange DK, Christ SE, Steiner R, White DA. Variability in phenylalanine control predicts IQ and executive abilities in children with phenylketonuria. Mol Genet Metab 2014;111(4):445-451 [FREE Full text] [doi: 10.1016/j.ymgme.2014.01.012] [Medline: 24568837]
4. Jahja R, van Spronsen FJ, de Sonneville LMJ, van der Meere JJ, Bosch AM, Hollak CEM, et al. Long-term follow-up of cognition and mental health in adult phenylketonuria: A PKU-COBESO Study. Behav Genet 2017;47(5):486-497 [FREE Full text] [doi: 10.1007/s10519-017-9863-1] [Medline: 28776207]
5. Christ SE, Huijbregts SC, de Sonneville LM, White DA. Executive function in early-treated phenylketonuria: profile and underlying mechanisms. Mol Genet Metab 2010;99 Suppl 1:S22-S32. [doi: 10.1016/j.ymgme.2009.10.007] [Medline: 20123466]
6. Janzen D, Nguyen M. Beyond executive function: non-executive cognitive abilities in individuals with PKU. Mol Genet Metab 2010;99 Suppl 1:S47-S51. [doi: 10.1016/j.ymgme.2009.10.009] [Medline: 20123470]
7. Hofman DL, Champ CL, Lawton CL, Henderson M, Dye L. A systematic review of cognitive functioning in early treated adults with phenylketonuria. Orphanet J Rare Dis 2018;13(1):150 [FREE Full text] [doi: 10.1186/s13023-018-0893-4] [Medline: 30165883]
8. Palermo L, Geberhiwot T, MacDonald A, Limback E, Hall SK, Romani C. Cognitive outcomes in early-treated adults with phenylketonuria (PKU): A comprehensive picture across domains. Neuropsychology 2017;31(3):255-267 [FREE Full text] [doi: 10.1037/neu0000337] [Medline: 28080075]
9. Anderson PJ, Wood SJ, Francis DE, Coleman L, Anderson V, Boneh A. Are neuropsychological impairments in children with early-treated phenylketonuria (PKU) related to white matter abnormalities or elevated phenylalanine levels? Dev Neuropsychol 2007;32(2):645-668. [doi: 10.1080/87565640701375963] [Medline: 17931123]
10. Leuzzi V, Pansini M, Sechi E, Chiarotti F, Carducci C, Levi G, et al. Executive function impairment in early-treated PKU subjects with normal mental development. J Inherit Metab Dis 2004;27(2):115-125. [doi: 10.1023/B:BOLI.0000028781.94251.1f] [Medline: 15159642]
11. Brumm VL, Azen C, Moats RA, Stern AM, Broomand C, Nelson MD, et al. Neuropsychological outcome of subjects participating in the PKU adult collaborative study: a preliminary review. J Inherit Metab Dis 2004;27(5):549-566. [doi: 10.1023/b:boli.0000042985.02049.ff] [Medline: 15669671]
12. Clocksin HE, Abbene EE, Christ S. A comprehensive assessment of neurocognitive and psychological functioning in adults with early-treated phenylketonuria. J Int Neuropsychol Soc 2023;29(7):641-650. [doi: 10.1017/S1355617722000686] [Medline: 36134437]
13. Singh S, Germine L. Technology meets tradition: a hybrid model for implementing digital tools in neuropsychology. Int Rev Psychiatry 2021;33(4):382-393. [doi: 10.1080/09540261.2020.1835839] [Medline: 33236657]
14. Ebner-Priemer UW, Trull TJ. Ecological momentary assessment of mood disorders and mood dysregulation. Psychol Assess 2009;21(4):463-475. [doi: 10.1037/a0017075] [Medline: 19947781]
15. Cartwright J. Technology: smartphone science. Nature 2016;531(7596):669-671. [doi: 10.1038/nj7596-669a] [Medline: 27035011]
16. Germine L, Reinecke K, Chaytor NS. Digital neuropsychology: challenges and opportunities at the intersection of science and software. Clin Neuropsychol 2019;33(2):271-286. [doi: 10.1080/13854046.2018.1535662] [Medline: 30614374]

XSL•FO
RenderX

17. Germine L, Strong RW, Singh S, Sliwinski MJ. Toward dynamic phenotypes and the scalable measurement of human behavior. Neuropsychopharmacology 2021;46(1):209-216 [FREE Full text] [doi: 10.1038/s41386-020-0757-1] [Medline: 32629456]
18. Trull TJ, Ebner-Priemer U. Ambulatory assessment. Annu Rev Clin Psychol 2013;9:151-176 [FREE Full text] [doi: 10.1146/annurev-clinpsy-050212-185510] [Medline: 23157450]
19. Liao Y, Skelton K, Dunton G, Bruening M. A systematic review of methods and procedures used in ecological momentary assessments of diet and physical activity research in youth: an adapted STROBE checklist for reporting EMA studies (CREMAS). J Med Internet Res 2016;18(6):e151 [FREE Full text] [doi: 10.2196/jmir.4954] [Medline: 27328833]
20. Germine L, Nakayama K, Duchaine BC, Chabris CF, Chatterjee G, Wilmer JB. Is the web as good as the lab? Comparable performance from web and lab in cognitive/perceptual experiments. Psychon Bull Rev 2012;19(5):847-857. [doi: 10.3758/s13423-012-0296-9] [Medline: 22829343]
21. Sliwinski MJ, Mogle JA, Hyun J, Munoz E, Smyth JM, Lipton RB. Reliability and validity of ambulatory cognitive assessments. Assessment 2018;25(1):14-30 [FREE Full text] [doi: 10.1177/1073191116643164] [Medline: 27084835]
22. Chaytor NS, Barbosa-Leiker C, Germine LT, Fonseca LM, McPherson SM, Tuttle KR. Construct validity, ecological validity and acceptance of self-administered online neuropsychological assessment in adults. Clin Neuropsychol 2021;35(1):148-164 [FREE Full text] [doi: 10.1080/13854046.2020.1811893] [Medline: 32883156]
23. Fonseca L, Kanapka L, Miller K, Pratley R, Rickels M, Chaytor N. Cognitive impairment in older adults with type 1 diabetes: longitudinal data from the wireless innovation for seniors with diabetes mellitus (WISDM) Study. Alzheimer's & Dementia 2022;18(S7) [FREE Full text] [doi: 10.1002/alz.066626]
24. Singh S, Strong R, Xu I, Fonseca LM, Hawks Z, Grinspoon E, et al. Ecological momentary assessment of cognition in clinical and community samples: reliability and validity study. J Med Internet Res 2023;25:e45028 [FREE Full text] [doi: 10.2196/45028] [Medline: 37266996]
25. Hawks Z, Beck E, Jung L. Dynamic associations between glucose and ecological momentary cognition in Type 1 Diabetes. npj Digit Med 2024;59(7) [FREE Full text] [doi: 10.1093/geroni/igae098.2212]
26. Hartshorne JK, Germine LT. When does cognitive functioning peak? The asynchronous rise and fall of different cognitive abilities across the life span. Psychol Sci 2015;26(4):433-443 [FREE Full text] [doi: 10.1177/0956797614567339] [Medline: 25770099]
27. Passell E, Dillon DG, Baker JT, Vogel SC, Scheuer LS, Mirin NL. Digital Cognitive Assessment: Results from the TestMyBrain NIMH Research Domain Criteria (RDoC) Field Test Battery Report. 2019. URL: https://osf.io/preprints/psyarxiv/dcszr_v1 [accessed 2025-03-21]
28. Singh S, Strong RW, Jung L, Li FH, Grinspoon L, Scheuer LS, et al. The testMyBrain digital neuropsychology toolkit: development and psychometric characteristics. J Clin Exp Neuropsychol 2021;43(8):786-795 [FREE Full text] [doi: 10.1080/13803395.2021.2002269] [Medline: 34907842]
29. Larsen E, Murton O, Song X, Joachim D, Watts D, Kapczinski F, et al. Validating the efficacy and value proposition of mental fitness vocal biomarkers in a psychiatric population: prospective cohort study. Front Psychiatry 2024;15:1342835 [FREE Full text] [doi: 10.3389/fpsyt.2024.1342835] [Medline: 38505797]
30. Pilkonis PA, Choi SW, Reise SP, Stover AM, Riley WT, Cella D, PROMIS Cooperative Group. Item banks for measuring emotional distress from the patient-reported outcomes measurement information system (PROMIS®): depression, anxiety, and anger. Assessment 2011;18(3):263-283 [FREE Full text] [doi: 10.1177/1073191111411667] [Medline: 21697139]
31. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. Journal of Health and Social Behavior 1983;24(4):385. [doi: 10.2307/2136404]
32. Cella D, Lai JS, Nowinski CJ, Victorson D, Peterman A, Miller D, et al. Neuro-QOL: brief measures of health-related quality of life for clinical research in neurology. Neurology 2012;78(23):1860-1867 [FREE Full text] [doi: 10.1212/WNL.0b013e318258f744] [Medline: 22573626]
33. NA. Diagnostic and Statistical Manual of Mental Disorders. Virginia, United States: American Psychiatric Association; 2013.
34. Narrow WE, Clarke DE, Kuramoto SJ, Kraemer HC, Kupfer DJ, Greiner L, et al. DSM-5 field trials in the United States and Canada, part III: development and reliability testing of a cross-cutting symptom assessment for DSM-5. Am J Psychiatry 2013;170(1):71-82. [doi: 10.1176/appi.ajp.2012.12071000] [Medline: 23111499]
35. WHO ASSIST Working Group. The alcohol, smoking and substance involvement screening test (ASSIST): development, reliability and feasibility. Addiction 2002;97(9):1183-1194 [FREE Full text] [doi: 10.1046/j.1360-0443.2002.00185.x] [Medline: 12199834]
36. Chung F, Abdullah HR, Liao P. STOP-bang questionnaire: A practical approach to screen for obstructive sleep apnea. Chest 2016;149(3):631-638 [FREE Full text] [doi: 10.1378/chest.15-0903] [Medline: 26378880]
37. Mascarenhas Fonseca L, Strong RW, Singh S, Bulger JD, Cleveland M, Grinspoon E, et al. Glycemic variability and fluctuations in cognitive status in adults with type 1 diabetes (GluCog): observational study using ecological momentary assessment of cognition. JMIR Diabetes 2023;8:e39750 [FREE Full text] [doi: 10.2196/39750] [Medline: 36602848]

XSL•FO
RenderX

38. Germine LT, Joormann J, Passell E, Rutter LA, Scheuer L, Martini P, et al. Neurocognition after motor vehicle collision and adverse post-traumatic neuropsychiatric sequelae within 8 weeks: initial findings from the AURORA study. J Affect Disord 2022;298(Pt B):57-67 [FREE Full text] [doi: 10.1016/j.jad.2021.10.104] [Medline: 34800569]
39. D'Ardenne K, Savage CR, Small D, Vainik U, Stoeckel LE. Core neuropsychological measures for obesity and diabetes trials: initial report. Front Psychol 2020;11:554127 [FREE Full text] [doi: 10.3389/fpsyg.2020.554127] [Medline: 33117225]
40. Fortenbaugh FC, DeGutis J, Germine L, Wilmer JB, Grosso M, Russo K, et al. Sustained attention across the life span in a sample of 10,000: dissociating ability and strategy. Psychol Sci 2015;26(9):1497-1510 [FREE Full text] [doi: 10.1177/0956797615594896] [Medline: 26253551]
41. Pylyshyn ZW, Storm RW. Tracking multiple independent targets: evidence for a parallel tracking mechanism. Spat Vis 1988;3(3):179-197. [doi: 10.1163/156856888x00122] [Medline: 3153671]
42. The R foundation for statistical computing. R version 4.3.1. 2023. URL: https://www.r-project.org/ [accessed 2025-03-21]
43. RStudio, PBC. 2022. URL: https://posit.co/ [accessed 2025-03-21]
44. Wickham H, Averick A, Bryan J, Chang W, McGowan L, François R, et al. Welcome to the Tidyverse. JOSS 2019;4(43):1686 [FREE Full text] [doi: 10.21105/joss.01686]
45. Wickham H. The split-apply-combine strategy for data analysis. J Statistical Software 2011;40:1-29 [FREE Full text] [doi: 10.18637/jss.v040.i01]
46. Revelle W. psych: procedures for psychological, psychometric, and personality research. R package version 2.4.3. 2024. URL: https://CRAN.R-project.org/package=psych [accessed 2025-03-21]
47. Pedersen T. patchwork: the composer of plots. R package version 1.2.0. 2024. URL: https://github.com/thomasp85/patchwork [accessed 2025-03-21]
48. Cranford JA, Shrout PE, Iida M, Rafaeli E, Yip T, Bolger N. A procedure for evaluating sensitivity to within-person change: can mood measures in diary studies detect change reliably? Pers Soc Psychol Bull 2006;32(7):917-929 [FREE Full text] [doi: 10.1177/0146167206287721] [Medline: 16738025]
49. Featured brain tests. Test My Brain. URL: https://testmybrain.org [accessed 2025-03-21]
50. Banerjee P, Grange D, Steiner R, White DA. Executive strategic processing during verbal fluency performance in children with phenylketonuria. Child Neuropsychol 2011;17(2):105-117 [FREE Full text] [doi: 10.1080/09297049.2010.525502] [Medline: 21140312]
51. Eijgelshoven I, Demirdas S, Smith TA, van Loon JM, Latour S, Bosch AM. The time consuming nature of phenylketonuria: a cross-sectional study investigating time burden and costs of phenylketonuria in the Netherlands. Mol Genet Metab 2013;109(3):237-242 [FREE Full text] [doi: 10.1016/j.ymgme.2013.05.003] [Medline: 23731533]
52. Trefz KF, Muntau AC, Kohlscheen KM, Altevers J, Jacob C, Braun S, et al. Clinical burden of illness in patients with phenylketonuria (PKU) and associated comorbidities - a retrospective study of German health insurance claims data. Orphanet J Rare Dis 2019;14(1):181 [FREE Full text] [doi: 10.1186/s13023-019-1153-y] [Medline: 31331350]
53. Darbà J. Characteristics, comorbidities, and use of healthcare resources of patients with phenylketonuria: a population-based study. J Med Econ 2019;22(10):1025-1029 [FREE Full text] [doi: 10.1080/13696998.2019.1636381] [Medline: 31237171]

Abbreviations

BPR: between-person reliability
dprime: response bias
DSM: digit symbol matching
EMA: ecological momentary assessment
gradCPT: gradual onset continuous performance test
HIPAA: Health Insurance Portability and Accountability Act
IRB: Institutional Review Board
MGB: Mass General Brigham
mlr: multilevel reliability
MOT: multiple object tracking
NPKUA: National PKU Alliance
Phe: phenylalanine
PHEFREE: Phenylalanine Families and Researchers Exploring Evidence
PKU: phenylketonuria
REDCap: Research Electronic Data Capture
TMB: TestMyBrain website

XSL•FO
RenderX

Edited by A Mavragani; submitted 25.06.24; peer-reviewed by MS Schoen, J Curcic; comments to author 07.10.24; revised version received 18.11.24; accepted 04.03.25; published 03.06.25.

Please cite as:
Singh S, Kluen L, Curtis K, Norel R, Agurto C, Grinspoon E, Hawks Z, Christ S, Waisbren S, Cecchi G, Germine L
Leveraging Cognitive and Speech Ecological Momentary Assessment in Individuals With Phenylketonuria: Development and Usability Study of Cognitive Fluctuations in a Rare Disease Population
JMIR Form Res 2025;9:e63644
URL: https://formative.jmir.org/2025/1/e63644
doi:10.2196/63644
PMID:40072884

©Shifali Singh, Lisa Kluen, Katelin Curtis, Raquel Norel, Carla Agurto, Elizabeth Grinspoon, Zoe Hawks, Shawn Christ, Susan Waisbren, Guillermo Cecchi, Laura Germine. Originally published in JMIR Formative Research (https://formative.jmir.org), 03.06.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Caregivers' Perceptions of Clinical Symptoms, Disease Management, and Quality of Life Impact in Cases of Cyclin-Dependent Kinase-Like 5 Deficiency Disorder: Cross-Sectional Online Survey

Sam Amin[1], MBChB; Carol-Anne Partridge[2], MSc; Helen Leonard[3], MBChB; Jenny Downs[3], PhD; Helen Allvin[4], MSc; Valentine Ficara[5], MSc; Emilie Pain[5], MSc; Minna A Korolainen[4], PhD

[1]Department of Paediatric Neurology, Bristol Royal Hospital for Children, Bristol, United Kingdom
[2]CDKL5 UK, Bristol, United Kingdom
[3]The Kids Research Institute Australia, The Centre for Child Health Research, University of Western Australia, Perth, Australia
[4]Orion Corporation, Orion Pharma, Espoo, Finland
[5]Carenity, Paris, France

Corresponding Author:
Sam Amin, MBChB
Department of Paediatric Neurology
Bristol Royal Hospital for Children
Paediatric Neurology, Level 6, Education Centre
Upper Maudlin Street
Bristol, BS 28AE
United Kingdom
Phone: 44 7886660762
Email: sam.amin@bristol.ac.uk

Abstract

Background: Cyclin-dependent kinase-like 5 (CDKL5) deficiency disorder (CDD) is an ultrarare genetic condition causing developmental epileptic encephalopathy characterized by seizures and motor and intellectual disabilities. No disease-modifying therapies are available, and treatments focus mainly on symptom management to improve quality of life.

Objective: The aim of this study was to better understand the burden of CDD based on family caregivers' perceptions.

Methods: The study was a cross-sectional, web-based survey comprising 40 questions for caregivers of patients with CDD and focusing on sociodemographic and medical characteristics, disease burden, unmet needs, treatments, and support. An adapted version of the EQ-5D-5L instrument was included to measure patients' health-related quality of life as perceived by their caregivers.

Results: A total of 132 caregivers, mostly from western parts of Europe, responded. The median patient age was 7.6 (IQR 2.9-12.2) years. Seizure onset occurred early, with the median onset at 2.0 (IQR 1.0-3.0) months of age. The median age at diagnosis was 1.2 (IQR 0.6-4.0) years. Epilepsy (123/132, 93.2%) and limited communication skills (111/132, 84.1%) were the most commonly reported symptoms. The highest number of different types of symptoms was reported for patients aged 5-9 years, with a median of 9.0 (IQR 7.5-10.0) symptoms. Most patients with epilepsy experienced daily seizures (81/123, 65.9%), and nearly all (119/123, 96.7%) were on antiseizure medications. A minority was on a ketogenic diet (21/123, 17.1%) or underwent vagus nerve stimulation (14/123, 11.4%). The care received was multidisciplinary. Compared to younger patients, adults had fewer medical appointments and a smaller variety of health care professionals in their care team. The EQ-5D-5L, adapted for caregivers, indicated low health-related quality of life for patients, with a median global index value of 0.18 (IQR 0.11-0.32). The most severe consequences of CDD on patients' daily lives were reported for mobility (88/132, 66.7%), self-care (120/132, 90.9%), and everyday activities (103/132, 78.0%). Caregiver burden was also substantial, with all life aspects reportedly impacted by CDD, including professional life and financial resources (median impact ratings of 9.0/10 and 7.0/10, respectively). Access to support and care varied depending on location. Caregivers outside Europe reported a longer time between the first seizure and diagnosis (26.5, IQR 3.2-47.0 months) compared to European caregivers (11.0, IQR 5.0-45.0 months). They also reported a higher impact of CDD on their financial resources (rating of 10/10) compared to European caregivers (rating of 6/10) and greater challenges in covering costs.

XSL•FO
RenderX

Conclusions: The study findings provide valuable insights on symptoms and disease burden related to CDD. This burden was quantitatively characterized with the EQ-5D-5L for the first time and was perceived as substantial by family caregivers. Discrepancies between geographic regions and age groups were highlighted, especially regarding available support and access to resources and care.

(JMIR Form Res 2025;9:e72489) doi:10.2196/72489

KEYWORDS

CDKL5 deficiency disorder; cyclin-dependent kinase-like 5; developmental epileptic encephalopathy; caregiver perception; burden; cross-sectional survey; online survey; health-related quality of life; epilepsy; rare disorder

Introduction

Background

Cyclin-dependent kinase-like 5 (CDKL5) deficiency disorder (CDD) is an ultrarare genetic condition, with an estimated incidence of 2.36 per 100,000 live births [1]. It is a developmental epileptic encephalopathy condition and was first described in 2004 [2]. Initially, CDD was believed to be an early-onset variant of Rett syndrome, another rare neurological disorder, due to the similarity in symptoms (seizures, developmental delay, gastrointestinal dysfunction, scoliosis, etc). However, CDD has a distinct genetic cause and symptom profile, as severe developmental delay is exhibited from birth and seizure onset occurs earlier, and it was thus established as a separate clinical condition in 2013 [3-5]. CDD is caused by pathogenic variants in the X-linked gene *CDKL5*, which plays a key role in normal brain development and function [6]. The disorder is estimated to be 4 times more prevalent in females than in males, but male patients often experience more severe symptoms than their female counterparts [7].

CDD is characterized by a broad range of clinical features, including early-onset refractory epilepsy; hypotonia; developmental, motor, and intellectual disabilities; cerebral (cortical) visual impairment; and microcephaly [3]. Epileptic seizures are usually the first symptom to occur, with a median age of onset of 6 weeks (by the age of 3 months in 90% of cases) [3,8,9]. Psychomotor impairments, hypotonia, and significant intellectual disability in children with CDD result in difficulties with gross motor skills, such as sitting, standing, or walking, and communication skills, which are often limited to body language, facial expressions, and vocalizations [5,10-12]. Other common comorbidities of CDD include sleep disturbances, movement disorders, respiratory issues, and gastrointestinal disorders [13,14].

CDD can be suspected through clinical examination, but genetic testing is required to confirm the diagnosis [3]. Phenotypic variability exists between individuals in terms of symptom range and severity [3,7,15]. Management of CDD typically starts with pharmacological interventions to alleviate seizures and other comorbid symptoms, as no disease-modifying therapies are currently available [3,16,17]. Antiseizure medications (ASMs) are prescribed based on seizure types and patterns, with polypharmacy often necessary due to seizure refractoriness [3,18]. The most commonly used ASMs for CDD include levetiracetam, topiramate, sodium valproate, vigabatrin, phenobarbital, and clobazam [18]. Ganaxolone, a neuroactive steroid, demonstrated long-term reduction of CDD-associated seizure frequency in clinical trials and became the first ASM approved by US and European regulatory agencies for the treatment of seizures associated with CDD in patients aged 2 years or older [19-22]. In addition to medications, a ketogenic diet and vagus nerve stimulation (VNS; for patients over 4 years old) may also help to reduce the frequency and intensity of seizures [17,23,24]. For developmental and social issues, a multidisciplinary approach is considered essential, including speech therapy, physiotherapy, psychosocial support, and occupational therapy, to address the broader needs of patients [17]. Given the symptom load and the necessary care, the burden of CDD is considered substantial for patients and their family members. Families caring for children with CDD experience significant emotional and financial burdens, with primary caregivers, particularly mothers, reporting lower mental well-being. The severity of the child's sleep disturbances and the financial hardships further exacerbate stress levels, while the significant care needs likely contribute to a higher caregiving burden and reduced family quality of life [25]. Emotional well-being appears to be more impacted in the parents of children with CDD than in those caring for children with other genetic disorders associated with intellectual disability, such as Down syndrome or Rett syndrome [26]. Although recent studies have provided major insights on the clinical features and disease management of CDD, research quantifying and qualifying this burden remains limited, given the scarcity of the disorder and its recent identification [14,25-29].

Goal

The aim was to better understand the burden of CDD on both patients and their family caregivers and identify unmet needs by analyzing the perspectives of family caregivers in an international cross-sectional survey.

Methods

Study Design

This study included a descriptive, cross-sectional, web-based, self-administered survey completed by the caregivers of patients with CDD. The survey was designed and approved by a multidisciplinary steering committee of CDD specialists and caregivers.

Study Population

Participants were recruited through CDD patient advocacy organizations, including CDKL5 UK (United Kingdom), CDKL5 Alliance Francophone (France), CDKL5 Deutschland (Germany), Asociacion de Afectados CDKL5 (Spain), CDKL5

XSL•FO
RenderX

Ireland, CDKL5 Italy, CDKL5 Alliance (World), Finnish Epilepsy Association (Finland), Svenska Epilepsiförbundet - Swedish association from the International Bureau for Epilepsy (IBE), and EpiCare-network (Europe), as well as through Carenity, a web-based community in which both patients and caregivers can share their experiences, provide support, and receive information. CDD caregivers were recruited via digital communication and invited by patient organizations via email to voluntarily complete an online survey about the burden of this disease and its impact on the quality of life of patients with CDD and their caregivers. Eligible participants included adult legal representatives or guardians (18 years or older at the time of inclusion) of individuals with genetically confirmed CDD. While the study initially targeted caregivers living in western parts of Europe, no respondents were excluded based on their country of residence. All participants provided electronic informed consent prior to participating in the survey.

Survey and Data Collection

The questionnaire consisted of 3 sections with a total of 40 questions: sociodemographic and medical characteristics (25 questions); quality of life, disease burden, and unmet needs (10 questions); and treatment and support received (5 questions) (Multimedia Appendix 1). There were various question formats, such as single- and multiple-choice questions, numerical questions, slider-scale questions, and open-ended questions, and there was an adapted version of the self-administered EQ-5D-5L instrument, developed by the EuroQol Research Foundation to measure patients' health-related quality of life [30,31]. The estimated completion time for the survey was 25 minutes, and the average completion time was 24.5 minutes.

The EQ-5D-5L proxy version 1 questionnaire was modified for caregivers to describe patients' health state. The adapted version was validated by EuroQol [31]. Using a 5-level scale (no problems, slight problems, moderate problems, severe problems, and extreme problems), caregivers reported the severity of issues patients faced across 5 dimensions: mobility, self-care, everyday activities, pain/discomfort, and anxiety/depression. These levels are combined to obtain a single index value, reflecting the preferences of the general population in the caregiver's country, with each dimension weighted accordingly (Multimedia Appendix 2) [32,33]. Additionally, using a visual analog scale (VAS), caregivers rated the patient's health on the day the questionnaire was completed, ranging from 0 (worst health imaginable) to 100 (best health imaginable).

The impacts of CDD on the caregiver's family, social, and work life; financial resources; quality of sleep; level of stress; and general quality of life were self-rated using a slider scale, with responses ranging from "no impact at all" (value=0) to "very high impact" (value=10).

The data collection occurred from May 25, 2023, to December 31, 2023.

Statistical Analysis

Data management and analysis were performed using R (version 4.1.2; R Core Team).

A comprehensive quality check was performed prior to analysis on fully completed questionnaires in order to exclude respondents with inconsistent answers (eg, conflicting responses between similar questions, unrealistic values in numerical fields, or incoherent open-ended responses), response times shorter than the calculated minimum for the shortest completion path (2.6 minutes or less than 5 seconds per question), and suspected duplicate entries.

We conducted descriptive analyses to characterize our study population and summarize responses to the survey. The data have been presented as median and IQR or frequency and percentage values as appropriate. Responses to open-ended questions were translated into English when necessary and manually analyzed to identify categories and subcategories. Where relevant, subgroup analyses were conducted based on caregivers' sociodemographic characteristics and patients' medical characteristics.

Ethical Considerations

This study was conducted in accordance with the current version of the Declaration of Helsinki, the ICH Good Epidemiological Practice guidelines, any local regulations, and the General Data Protection Regulation (GDPR). The study did not undergo an ethical review as it was considered as "a satisfaction survey" with a focus to better understand the impact of a disease (CDD) on patients and caregivers [34].

All patients provided electronic informed consent (translated to the language of the questionnaire) before any patient data collection. Participation was voluntary, and participants received no incentives to participate in the study.

Results

Participant Characteristics

Of 245 participants who started the survey, 132 (53.9% response rate) completed a valid questionnaire (Figure 1). Most participants were from Europe (106/132, 80.3%), more specifically western Europe (France and Benelux; 32/132, 24.2%), southern Europe (Spain, Italy, and Portugal; 30/132, 22.7%), central Europe (Germany, Austria, and Switzerland; 21/132, 15.9%), and northern Europe (United Kingdom, Ireland, and Sweden; 20/132, 15.2%). A few responses were gathered from eastern Europe (Poland, Slovakia, and Montenegro; 4/132, 3.0%). Respondents outside Europe (26/132, 19.7%) were from Latin America (Argentina, Bolivia, Chile, Colombia, Peru, Mexico, and Uruguay; 15/132, 11.4%), Turkey (8/132, 6.1%), Australia (2/132, 1.5%), and the United States (1/132, 0.8%).

XSL•FO
RenderX

Figure 1. Flow diagram outlining the survey population of caregivers of patients with cyclin-dependent kinase-like 5 deficiency disorder (CDD): number of individuals who started the survey, number of individuals who were excluded during the questionnaire completion or before analysis, and final number of questionnaires analyzed. *Reasons for not finishing the survey cannot be precisely determined, and they may include individuals who started the survey out of curiosity; **Not a caregiver of a patient with CDD (n=38), automatically screened out after entering the child's date of birth instead of the caregiver's date of birth (n=11), not the legal guardian of the patient they cared for (n=10), and no genetic testing performed (n=4).

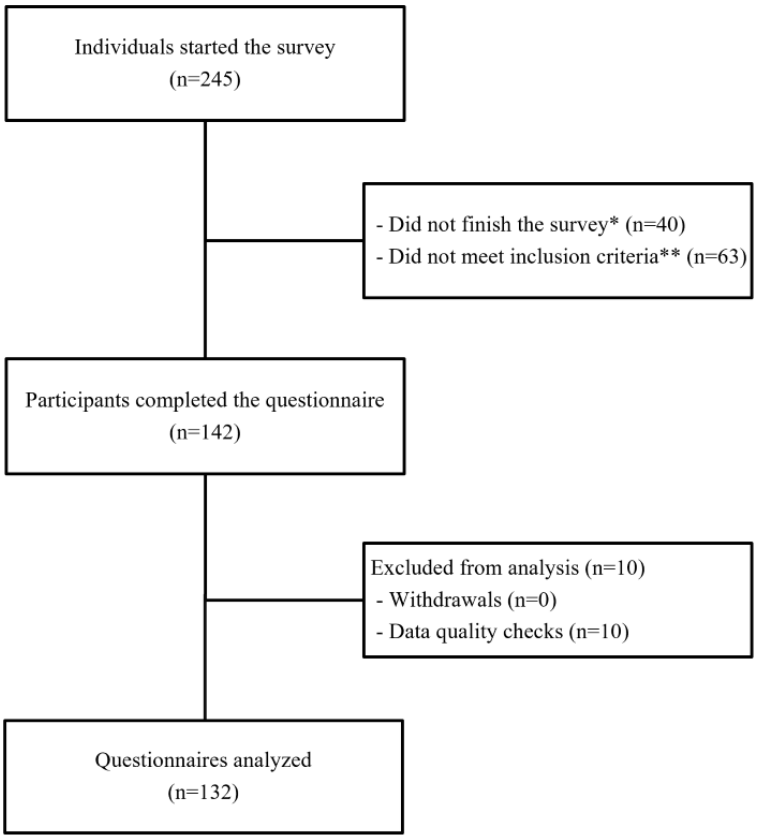

Sociodemographic Characteristics of Family Caregivers

Caregivers' demographics are summarized in Table 1. All 132 respondents were parents of patients with CDD, with a median age of 41 years (IQR 35.0-46.0 years), and most were women (112/132, 84.8%). Most caregivers (85/132, 64.4%) did not work or worked part-time because of their children's CDD. Caregivers outside Europe were slightly more likely to have their employment status impacted by CDD (19/26, 73%; either working part-time or not working because of CDD) compared to European caregivers (66/106, 62.3%).

XSL•FO
RenderX

Table 1. Sociodemographic characteristics of caregivers caring for patients with cyclin-dependent kinase-like 5 deficiency disorder.

Characteristic	Value (N=132)
Gender, n (%)	
Female	112 (84.8)
Male	19 (14.4)
Did not answer	1 (0.8)
Age (years)[a], median (IQR)	41.0 (35.0-46.0)
Age categories[a], n (%)	
<40 years	57 (43.8)
≥40 years	73 (56.2)
Relationship with the patient, n (%)	
Parent	132 (100)
Caregiver's role, n (%)	
Main caregiver	131 (99.2)
Not the main caregiver	1 (0.8)
Employment status, n (%)	
Working full time	32 (24.2)
Working part-time for reasons unrelated to CDD[b]	9 (6.8)
Working part-time because of CDD	39 (29.5)
Not working for reasons unrelated to CDD	4 (3.0)
Not working because of CDD	46 (34.8)
Other[c]	2 (1.5)

[a]N=130; 2 answers were removed after data quality check.

[b]CDD: cyclin-dependent kinase-like 5 deficiency disorder.

[c]Other (n=2): working from home (n=1) and 99% reduction in working hours (n=1).

Sociodemographic Characteristics of Patients

Most patients with CDD were female (117/132, 88.6%), and two-thirds (88/131, 67.2%) were under 10 years old, with age ranging from 4 months to 38 years (Table 2). Patients in and outside Europe had similar demographic characteristics, although European patients were slightly older, with a median age of 7.7 years (IQR 3.0-14.4 years) compared to a median age of 5.2 years (IQR 1.9-8.2 years) for those living outside Europe. Most patients (127/132, 96.2%) lived in their family home.

XSL•FO
RenderX

Table 2. Sociodemographic characteristics of patients with cyclin-dependent kinase-like 5 deficiency disorder.

Characteristic	Value (N=132)
Gender, n (%)	
Female	117 (88.6)
Male	15 (11.4)
Age (years)[a], median (IQR)	7.6 (2.9-12.2)
Age categories[a], n (%)	
<5 years	49 (37.4)
≥5 and <10 years	39 (29.8)
≥10 and <18 years	23 (17.6)
≥18 years	20 (15.3)
Living situation, n (%)	
Family home	127 (96.2)
Specialized institution	3 (2.3)
Community home	2 (1.5)
Location, n (%)	
Central Europe (Germany, Austria, and Switzerland)	21 (15.9)
Eastern Europe (Poland, Slovakia, and Montenegro)	4 (3.0)
Northern Europe (United Kingdom, Ireland, and Sweden)	20 (15.6)
Southern Europe (Spain, Italy, and Portugal)	30 (22.7)
Western Europe (France and Benelux)	32 (24.2)
Latin America (Argentina, Bolivia, Chile, Colombia, Peru, Mexico, and Uruguay)	15 (11.4)
Turkey	8 (6.1)
Other (Australia and United States)	3 (2.3)

[a]N=131; 1 answer was removed after data quality check.

CDD Journey for Patients and Caregivers

Patients' Medical Characteristics and Diagnoses

At inclusion, the median disease duration (time since the appearance of the first symptoms) was 7 years (IQR 2.7-12.3 years) and was longer for European patients (7.5 years, IQR 2.8-13.9 years) than for patients outside Europe (4.8 years, IQR 1.7-7.7 years). Nearly half (59/132, 44.7%) of the patients had access to a medical center specialized in CDD care, while the other half (61/132, 46.2%) did not. Half of the patients in Europe (54/106, 50.9%) had access to specialized centers compared to only 19% (5/26) of those outside Europe. The median time since diagnosis was 4.4 years (IQR 1.8-8.2 years).

All included patients had a genetically confirmed diagnosis, and the median age at diagnosis was 1.2 years (IQR 0.6-4.0 years), while the median age at the first seizure or symptoms was 2.0 months (IQR 1.0-3.0 months). Most patients (127/132, 96.2%) experienced their first CDD symptoms (seizures) before 12 months of age. No difference was observed based on geographic location, except for the median age at diagnosis, which was higher for patients outside Europe (2.7 years, IQR 0.5-4.0 years) than for European patients (1.1 years, IQR 0.6-4.0 years).

Medical attention for CDD symptoms was sought for a majority of patients (123/132, 93.2%) after the first seizure, with little to no delay (median 0.0 months). However, the median time between the first seizure or symptoms and diagnosis was 11 months (IQR 4.0-45.5 months). This time was longer for patients outside Europe (26.5 months, IQR 3.2-47.0 months) than for European patients (11 months, IQR 5.0-45.0 months) (Figure S1 in Multimedia Appendix 3).

The median age at diagnosis increased with the patient age at the time of the study (Figure 2).

XSL•FO
RenderX

Figure 2. Mean and median age at diagnosis (in years) by age group at inclusion among patients with cyclin-dependent kinase-like 5 deficiency disorder (N=132). Patients are categorized into age subgroups: <5 years (n=48), ≥5 and <10 years (n=39), ≥10 and <18 years (n=23), and ≥18 years (n=20). Q1: first quartile; Q3: third quartile; y/o: years old.

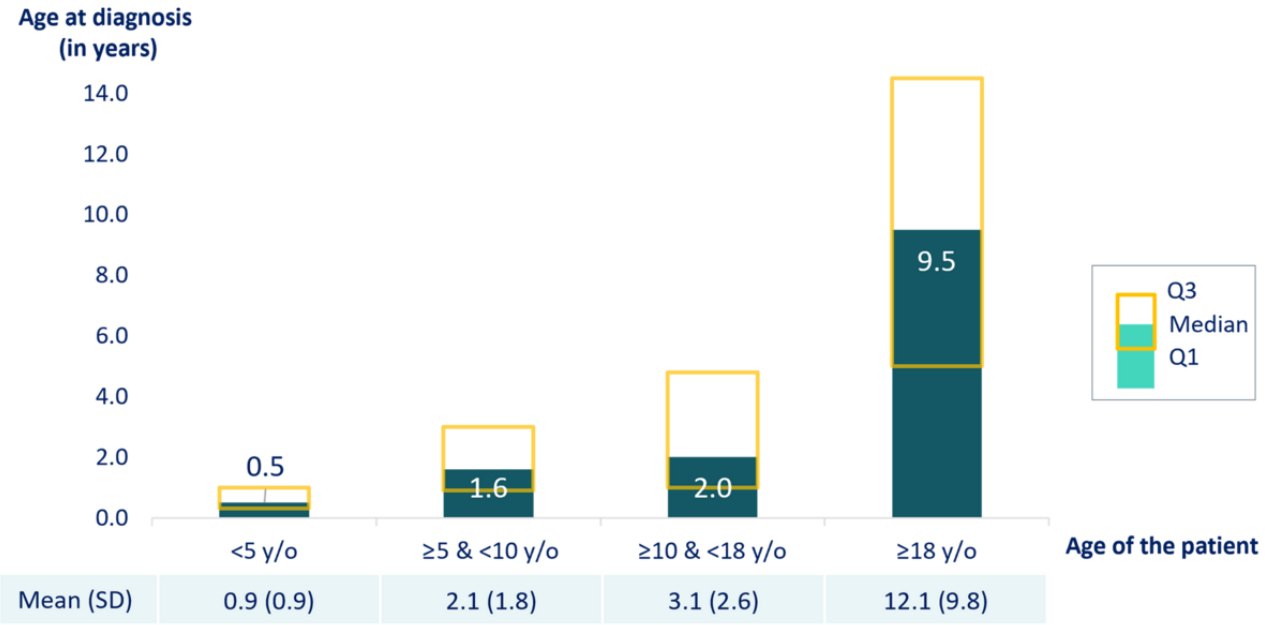

CDD Symptoms

All patients experienced at least one symptom over the past year, and 56.8% (75/132) experienced 8 or more symptoms (Figure 3). Epilepsy or seizures were reported in 93.2% (123/132) of patients, functional disabilities (eg, limited communication skills, difficulty with walking, stereotypies, limited hand function, movement disorders, visual impairment, and behavior disturbances) in 97.7% (129/132), and other medical problems (sleep problems, gastrointestinal and feeding problems, scoliosis or kyphosis, respiratory problems, and cardiac issues) in 89.4% (118/132). Patients aged 5-9 years experienced more symptoms, with a median of 9.0 (IQR 7.5-10.0) compared to older and younger age groups. Patients aged 10-17 years had a median of 8.0 symptoms (IQR 6.5-10.0), similar to adult patients (IQR 5.8-9.0). Patients aged <5 years had a median of 7.0 symptoms (IQR 7.0-9.0). Notably, 74% (29/39) of patients aged 5-9 years experienced 8 or more symptoms compared to only 39% (19/48) of patients aged <5 years.

XSL•FO
RenderX

Figure 3. Cyclin-dependent kinase-like 5 deficiency disorder symptoms experienced over the past year as reported by caregivers. Percentage of patients who experienced the symptoms in the last 12 months (N=132). *Other (n=9): developmental disorder (n=1), scratching (n=1), hyperactivity (n=1), immune system issues (n=1), precocious puberty (n=1), kidney stones (n=1), hypotonia (n=1), osteoporosis (n=1), and very intense outbreaks of aggression (n=1).

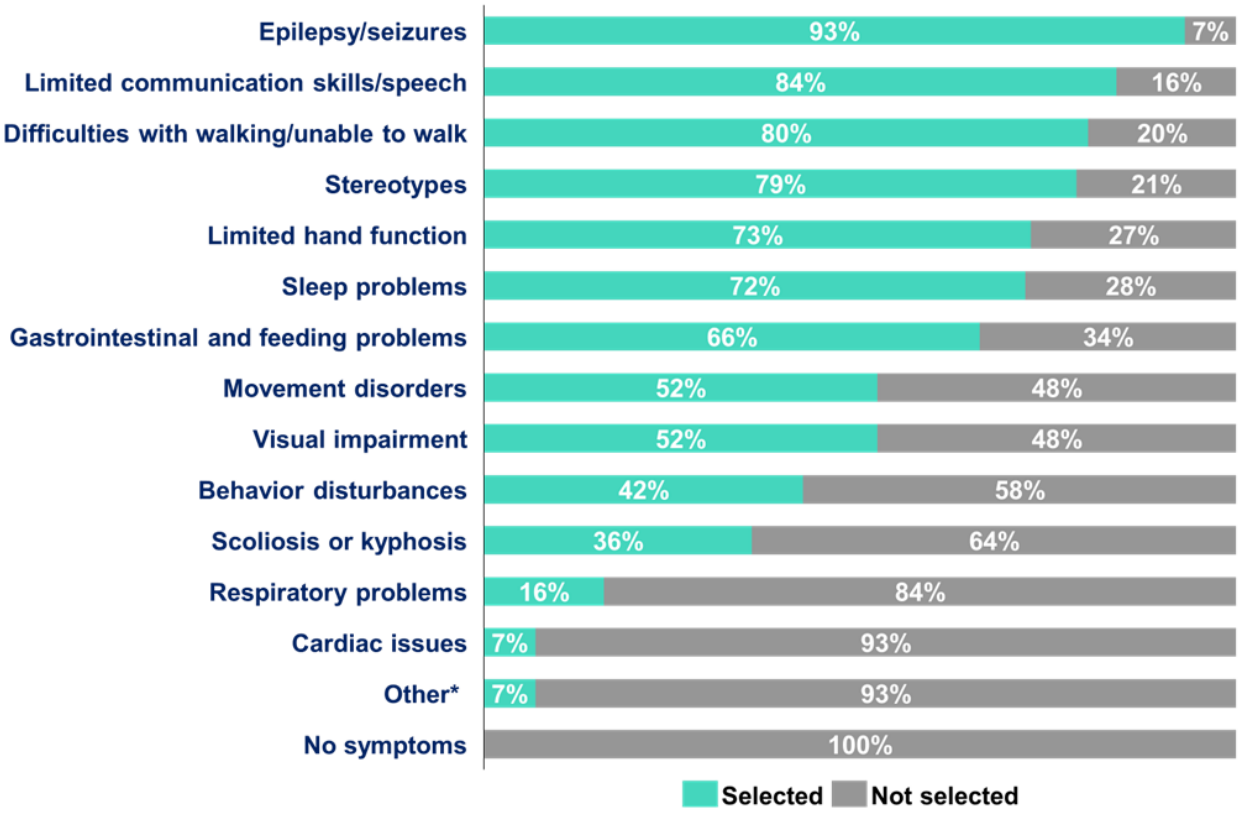

Of 123 respondents who reported epilepsy or seizures, 120 (97.6%) were able to identify at least one type of seizure over the past year. The majority (99/120, 82.5%) reported multiple types, with a median of 3.0 different types (IQR 2.0-4.0). Patients aged 5-9 years had the most varied seizure types, with 23% (8/35) experiencing 5 or more types compared to 10% (2/21) of those aged 10-17 years, 11% (5/45) of those aged <5 years, and 11% (2/18) of those aged ≥18 years.

The most commonly reported type of seizure over the past year was major motor seizure (91/123, 74.0%), followed by epileptic spasms (81/123, 65.9%), absence seizure (71/123, 57.7%), and myoclonic seizure (59/123, 48.0%). Hypermotor-tonic-spasm sequence was the least reported type of seizure (43/123, 35.0%).

Two-thirds (81/123, 65.9%) of patients with epilepsy or seizures experienced daily seizures over the past month, and 27 (22.0%) had >5 seizures per day. For only 8 patients (6.5%), no seizures were reported in the month prior to completing the survey.

Patients aged 5-18 years were identified as having more frequent seizures compared to younger and older patients. We found that 83% (29/35) of patients aged 5-9 years and 77% (16/21) of those aged 10-17 years experienced daily seizures, whereas only 42% (8/19) of patients aged ≥18 years and 60% (28/47) of patients aged <5 years experienced daily seizures.

Treatments

Almost all patients with seizures (119/123, 96.7%) were receiving ASMs on the day caregivers completed the questionnaire, with a median of 2 medications per day (IQR 2.0-3.0). Patients aged 5-9 years were taking more daily ASMs, with a median of 3.0 (IQR 2.0-4.0) compared to other age groups (2.0, IQR 2.0-3.0).

The number of daily ASMs taken tended to correlate positively with the frequency of seizures (Figure S2 in Multimedia Appendix 3). Patients experiencing >5 seizures per day took a median of 2.5 ASMs daily, and those with 1-5 seizures per day had a median of 3.0 ASMs. In contrast, patients who had experienced weekly, monthly, or no seizures over the past month had a lower median of 2.0 daily medications. Notably, 73% (24/33) of patients aged 5-9 years were on three or more daily ASMs compared to approximately 40% of patients in all other age groups.

Most respondents (108/132, 81.8%) reported discontinuing at least one ASM since diagnosis, with a median of 4.0 (IQR 2.0-7.0) ASMs stopped. The number was higher for older patients and for those with higher seizure frequency. A median of 5 ASMs had been discontinued for patients experiencing daily seizures (either 1-5 or >5), while those with weekly or monthly seizures discontinued only 2 ASMs.

At the time of the survey, only 17.1% (21/123) of patients were on a ketogenic diet and 11.4% (14/123) were on VNS. The

XSL•FO
RenderX

percentage of patients on a ketogenic diet was consistent across regions, with 17% (16/97) in Europe and 19% (5/26) outside Europe. However, VNS was reported for 14% (14/97) of European patients, while none (0/26) of the patients outside Europe had undergone this treatment (Figure S3 in Multimedia Appendix 3). Additionally, differences were seen within Europe as well regarding VNS treatment, as it was not reported for any patient living in southern Europe (0/27; Italy, Spain, and Portugal). Patient age had a clear impact on the types of treatments received. Among patients aged <5 years, 34% (16/47) were on a ketogenic diet compared to only 6% (2/35) among those aged 5-9 years, 14% (3/21) among those aged 10-17 years, and none among those aged ≥18 years. VNS is not recommended for children aged <4 years, and it was not reported for any patient aged <5 years [23]. However, VNS was used in 11% (4/35) of patients aged 5-9 years, 29% (6/21) of those aged 10-17 years, and 21% (4/19) of those aged ≥18 years.

Seizure frequency did not influence the number of patients on a ketogenic diet, but the use of VNS tended to correlate positively with seizure frequency as follows: 8% (1/12) of patients with a few seizures per month, 9% (2/22) of those with weekly seizures, 11% (6/54) of those with 1-5 seizures a day, and 19% (5/27) of those with >5 seizures a day.

More than half (75/123, 61.0%) of the caregivers of patients with seizures reported not using any rescue medications during the past year. This percentage was higher among the caregivers of younger patients (33/47, 70% for those aged <5 years) compared to other age groups (Figure S4 in Multimedia Appendix 3).

The use of rescue medications tended to be higher in patients with higher seizure frequency (Figure 4). Among the caregivers of patients with >5 seizures a day, 67% (18/27) used rescue medications at least once in the past year compared to 37% (20/54) of those with 1-5 seizures a day, 18% (4/22) of those with weekly seizures, and 17% (2/12) of those experiencing only a few seizures per month. Rescue medications were also used more frequently in groups with higher seizure frequency, with 44% (12/27) of caregivers of patients with >5 seizures a day using them several times a month or more often. This number did not exceed 11% in any other group.

Figure 4. Frequency of rescue medication use in the past year by frequency of seizures experienced over the past month by patients with cyclin-dependent kinase-like 5 deficiency disorder. Patients who experienced seizures (n=123) were categorized into subgroups according to seizure frequency over the past year: no seizure (n=8), a few seizures a month (n=12), weekly seizures (n=22), 1-5 seizures a day (n=54), and >5 seizures a day (n=27). The frequency of rescue medication is represented in percentage within each category of seizure frequency.

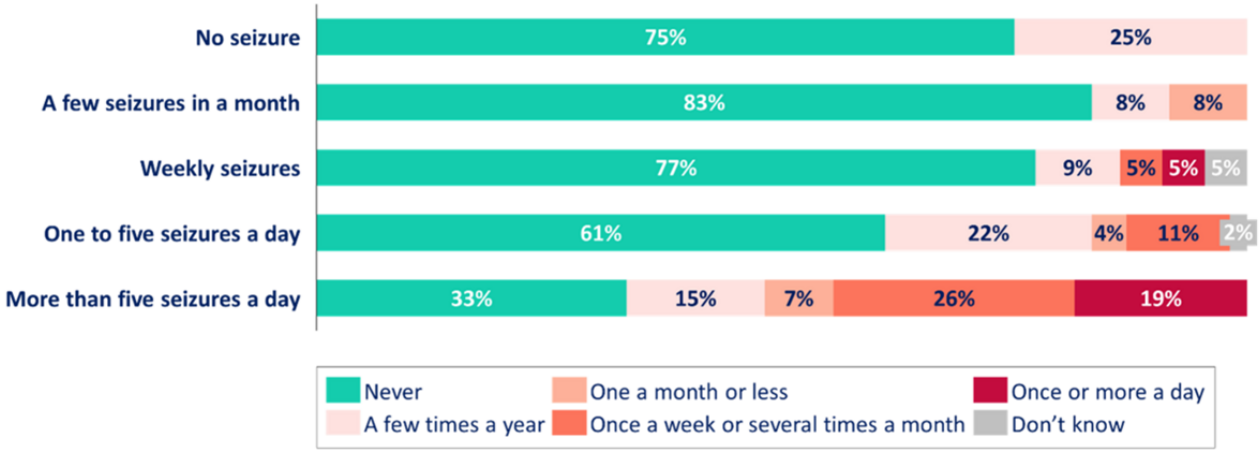

Patients who had discontinued fewer ASMs received rescue medications less frequently in the past year. Only 25% (3/12) of those who had not stopped any medications and 29% (13/45) of those who had stopped up to four medications reported using rescue medications. In contrast, the frequency of rescue medication use was higher among patients who had discontinued more medications. We found that 49% (19/39) of those who had stopped 5-10 medications and 44% (8/18) of those who had stopped ≥10 medications had used rescue medications in the past year.

Family caregivers of patients treated for epileptic seizures (n=119) reported dissatisfaction with both the effectiveness of ASMs in reducing seizures, rating it a median of 4 out of 10 (IQR 1.8-7.0), and the side effects (4/10, IQR 2.0-7.0). However, satisfaction regarding the burden of administration was slightly higher, with a median rating of 5.5 out of 10 (IQR 3.0-8.2).

Satisfaction with the effectiveness of seizure-reducing treatments and side effects decreased with time since the onset of the first seizure or symptoms (or time since disease onset) in contrast to satisfaction with the method and administration of treatment, which increased with time (Figure 5).

XSL•FO
RenderX

Figure 5. Mean and median caregiver satisfaction with the effectiveness, side effects, treatment method, and burden of administration of seizure-reducing treatments by time since the first seizure or symptoms, reported on a scale from 0 (not satisfied at all) to 10 (very satisfied) (n=118). Patients with cyclin-dependent kinase-like 5 deficiency disorder were categorized into subgroups according to the time since their first seizure or symptoms: <5 years (n=47), ≥5 and <10 years (n=33), and ≥10 years (n=38). Q1: first quartile; Q3: third quartile.

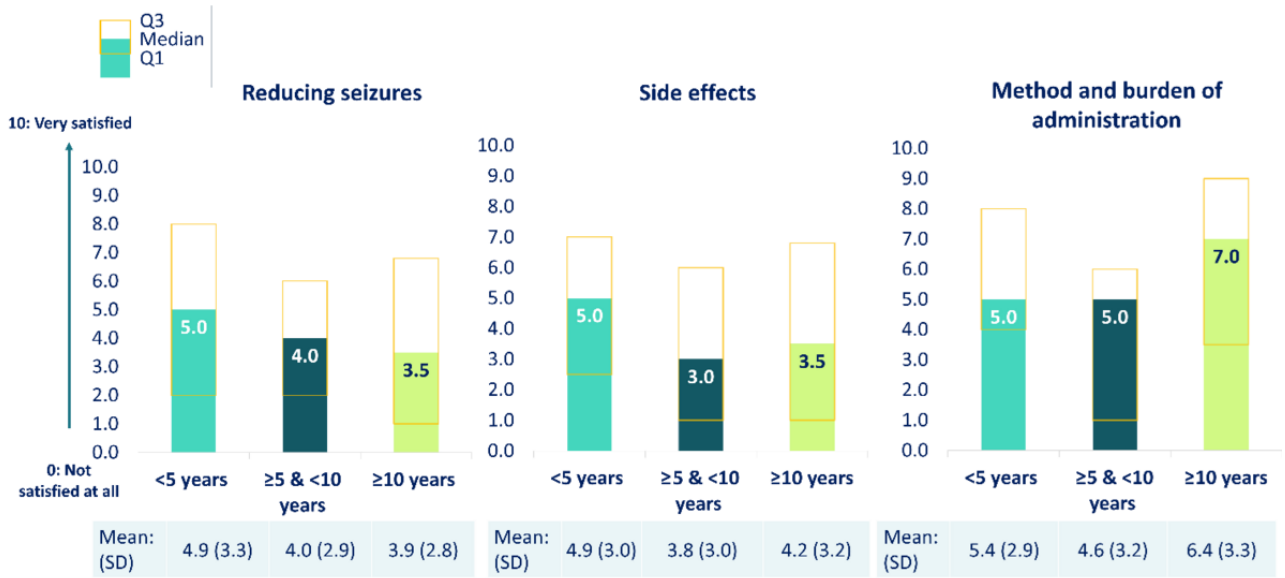

Satisfaction with treatment effectiveness decreased with the frequency of seizures experienced over the past month. Caregivers gave a median rating of 9 out of 10 (IQR 8.0-10.0) for patients with no seizures, 5 out of 10 for patients with a few seizures per month (IQR 2.5-8.0) and those with weekly seizures (IQR 4.0-7.0), and 3 out of 10 (IQR 0.2-4.8) for patients with daily seizures.

Health Care Appointments

Most patients (81/132, 61.4%) had appointments with 6 or more different types of health care specialists over the past year, with a median of 7.0 (IQR 5.0-8.0) different health care professionals involved (Figure S5 in Multimedia Appendix 3). The overall median number of total appointments per year was 57.5 (IQR 19.0-115.5), equating to over one appointment per week. Adult patients had fewer total appointments (median 13.0, IQR 4.8-60.8) and consulted a smaller variety of health care professionals, with a median number of 4.5 types.

The median number of appointments over the past year was also higher (62.0, IQR 26.0-127.0) among caregivers impacted in their professional life compared to caregivers who experienced no such impact (44.0, IQR 18.0-102.0).

Overall, the most frequently consulted health care professionals were physiotherapists and pediatric neurologists or epileptologists (97/132, 73.5% and 96/132, 72.7%, respectively), followed by pediatricians (81/132, 61.4%). Half of the patients (69/132, 52.3%) consulted a speech therapist. Over the past year, patients with more frequent seizures consulted more frequently with pediatric neurologists or epileptologists. The consultation rates were as follows: 80% (43/54) for patients with 1-5 seizures per day, 78% (21/27) for patients with >5 seizures per day, 25% (3/12) for patients with a few seizures in a month, and 50% (4/8) for patients without seizures in the past month.

General practitioners were consulted more often in Europe (37/106, 34.9%) than elsewhere (3/26, 11.5%). Conversely, nutritionists, gastroenterologists, and geneticists were consulted 1.5 to 4 times more often outside Europe.

Patient age influenced the type of health care professionals consulted. Visits to the dentist in the past year were more frequently reported for older patients (16/20, 80% for patients aged ≥18 years and 20/23, 87% for patients aged 10-17 years) compared to younger patients (16/49, 33% for patients aged <5 years and 25/39, 64% for patients aged 5-9 years). Inversely, older patients had fewer appointments with ophthalmologists. They were visited by 80% (39/49) of those aged <5 years in the past year and only 10% (2/20) of those aged ≥18 years.

Of the 132 patients, 102 (77.3%) had been hospitalized and 89 (67.4%) visited the emergency department for CDD-related symptoms at least once in their lifetime (Figure S6 in Multimedia Appendix 3). The hospitalization rate was higher in Europe (86/106, 81.1%) than outside Europe (16/26, 62%), but the median length of stay during the past year was shorter in Europe (2.0 nights, IQR 0.0-11.5) than outside Europe (4.0 nights, IQR 0.8-10.8).

The percentage of hospitalized patients was consistent across age groups, but younger patients spent more nights at the hospital during the past year. Patients aged <5 years spent a median of 8.0 nights (IQR 2.0-30.0), while those aged 5-9 years spent a median of 3.5 nights (IQR 1.0-10.0). Older patients had fewer hospital stays, with a median of 0.0 for both age groups (those aged 10-17 years: IQR 0.0-2.2; those aged ≥18 years: IQR 0.0-2.5).

Both the number of hospital stays and the length of these stays tended to correlate positively with the frequency of seizures. Altogether 89% (24/27) of patients with >5 seizures a day were hospitalized at least once in their lifetime compared to 42% (5/12) of patients with only a few seizures per month. Patients

XSL•FO
RenderX

experiencing >5 seizures per day had a median stay of 4.0 nights (IQR 0.8-13.2) at the hospital in the past year, while those with 1-5 seizures a day had a median stay of 6.0 nights (IQR 1.0-23.8). In contrast, patients with weekly seizures and those with monthly seizures spent a median of 0.0 nights (IQR 0.0-4.0 and IQR 0.0-4.5, respectively) at the hospital.

Disease Burden for Patients With CDD

When completing the adapted version of the EQ-5D-5L, most of the caregivers declared that patients' mobility (88/132, 66.7%), self-care (120/132, 90.9%), and everyday activities (103/132, 78.0%) were severely affected by CDD (level 5 "extreme problems/unable to do") on the day of completion (Figure 6). For most of the patients, there was little to no pain or discomfort (77/132, 58.3%) and depression or anxiety (97/132, 73.5%) reported that day.

Figure 6. Caregiver perception of the severity of cyclin-dependent kinase-like 5 deficiency disorder on patients' quality of life across the 5 EQ-5D-5L dimensions (N=132). For each dimension (mobility, self-care, usual activities, pain/discomfort, and anxiety/depression), the percentages of respondents are reported by the level of severity selected (1, "no problem" to 5, "extreme problems/unable to do").

100% / 0%	Mobility	Self-care	Usual activities	Pain/discomfort	Anxiety/depression
Level 1 (no problems)	5%	1%	1%	24%	42%
Level 2 (slight problems)	7%	2%	2%	34%	31%
Level 3 (moderate problems)	9%	2%	5%	27%	20%
Level 4 (severe problems)	12%	4%	14%	11%	5%
Level 5 (extreme problems/unable to do)	67%	91%	78%	5%	2%

Caregivers mentioned consistent levels of severity for mobility, pain or discomfort, and depression or anxiety across all countries, but more European caregivers reported that patients with CDD were not able to wash or dress themselves (99/106, 93.4%) or perform usual activities (86/106, 81.1%) that day compared to non-European caregivers (21/26, 81% and 17/26, 65%, respectively) (Figure S7 in Multimedia Appendix 3).

Overall, severity appeared to be similar across most age groups, with the exception of mobility, which seemed to be less affected among older patients in this study. Among patients aged ≥18 years, only 35% (7/20) had a rating of level 5, indicating "extreme problems/unable to do" compared to 65% (15/23) of patients aged 10-17 years, 74% (29/39) of patients aged 5-9 years, and 76% (37/49) of patients aged <5 years.

The survey population showed a notably low health-related quality of life for patients with CDD, with an overall median EQ-5D-5L index value of 0.18 (IQR 0.11-0.32) on a scale ranging from 0 (a state comparable to being dead) to 1 (representing full health). The VAS results were different, with an overall median score of 56 out of 100 (IQR 39.0-71.0), suggesting a neutral stance on the health status of patients.

Median index values were slightly higher for patients aged <5 years (0.21, IQR 0.13-0.31) and patients aged ≥18 years (0.25, IQR 0.16-0.36) that for patients aged 5-17 years (patients aged 5-9 years: 0.13, IQR –0.09 to 0.27; patients aged 10-17 years: 0.17, IQR 0.08-0.35). Patients aged <10 years had lower VAS scores (patients aged <5 years: 52/100, IQR 35.0-75.0; patients aged 5-9 years: 50/100, IQR 39.0-62.5) compared to older patients (patients aged 10-17 years: 59/100, IQR 41.5-70.5), and patients aged ≥18 years had the highest median score (70/100, IQR 50.5-80.25), indicating a better health status as perceived by family caregivers.

Caregiver Burden

Impact of CDD on Caregivers

In our sample, caring for a person with CDD had a substantial negative impact on all aspects of the caregiver's life, with median ratings consistently exceeding 7 out of 10 across different domains (Figure S8 in Multimedia Appendix 3). The burdens on family, professional, and social life, and the levels of stress were mentioned particularly often. Family life and stress levels were rated with a median of 9 out of 10 (IQR 7.0-10.0 for both), and the social and professional spheres had a median rating of 9 out of 10 (IQR 6.0-10.0 for both).

Caregivers residing outside Europe reported a higher overall impact of CDD on their quality of life compared to their European counterparts (Figure S8 in Multimedia Appendix 3). The most striking difference was observed in the financial field, where European caregivers reported a median rating of 6 out of 10 (IQR 3.0-9.0), while caregivers from other regions of the world rated the financial impact at a median of 10 out of 10 (IQR 8.5-10.0), indicating a more severe financial strain outside Europe.

Three-quarters of caregivers (97/132, 73.5%) reported epilepsy or seizures as one of the most difficult symptoms to manage. This was followed by difficulty interpreting communication (51/132, 38.6%) and difficulty with walking or the inability to walk (45/132, 34.1%).

Caregivers who reported epilepsy as one of the most difficult symptoms to manage (n=97) also reported higher burdens on family, social, and professional life; quality of sleep; and levels of stress compared to those who did not mention this symptom (n=35; Table 3).

XSL•FO
RenderX

Table 3. Self-reported impact of cyclin-dependent kinase-like 5 deficiency disorder on caregivers' family, social, and professional life; financial resources; sleep; stress; and overall quality of life, according to the level of epilepsy-related burden.

Aspects of life impacted by CDD[a,b]	Selected epilepsy/seizures as one of the hardest symptoms to manage (n=97)	Did not select epilepsy/seizures as one of the hardest symptoms to manage (n=35)
Family life		
Mean (SD)	8.3 (2.5)	7.7 (2.5)
Median (IQR)	10.0 (7.0-10.0)	8.0 (6.0-10.0)
Social life		
Mean (SD)	8.0 (2.5)	7.5 (2.1)
Median (IQR)	9.0 (7.0-10.0)	8.0 (6.0-10.0)
Professional life		
Mean (SD)	8.1 (2.4)	6.8 (3.2)
Median (IQR)	9.0 (7.0-10.0)	7.0 (5.0-10.0)
Financial resources		
Mean (SD)	6.6 (3.2)	6.6 (3.5)
Median (IQR)	7.0 (5.0-10.0)	7.0 (3.5-10.0)
Quality of sleep		
Mean (SD)	7.6 (2.8)	7.2 (2.5)
Median (IQR)	9.0 (6.0-10.0)	8.0 (6.0-10.0)
Level of stress		
Mean (SD)	8.2 (2.3)	7.6 (2.2)
Median (IQR)	9.0 (7.0-10.0)	7.0 (6.0-10.0)
General quality of life		
Mean (SD)	7.9 (2.2)	7.2 (2.0)
Median (IQR)	8.0 (7.0-10.0)	8.0 (5.0-9.0)

[a]CDD: cyclin-dependent kinase-like 5 deficiency disorder.

[b]Rated on a scale from 0 (no impact at all) to 10 (very high impact).

Out-of-Pocket Expenses

The majority of respondents (110/132, 83.3%) reported having out-of-pocket costs related to managing patients with CDD, with 44.7% (59/132) covering these expenses through household income (Figure S9 in Multimedia Appendix 3). However, 38.6% (51/132) found it challenging to meet these costs. While the number of caregivers reporting out-of-pocket expenses was consistent across all countries, more respondents from Europe were able to cover the costs (54/106, 50.9%) compared to non-Europeans (5/26, 19%).

Half of the respondents (55/110, 50.0%) reported incurring four or more different types of out-of-pocket expenses. The most common reason for expenses was alternative therapies, such as food supplements (66/110, 60.0%), followed by indirect costs (65/110, 59.1%) and nonpharmaceutical treatments (63/110, 57.3%).

The types of out-of-pocket costs varied greatly depending on the respondents' region of residence (Figure 7). Across all regions, alternative therapy fees were the most frequently reported. However, outside Europe, fees for medical appointments or work-up evaluations (18/23, 78%) and medical treatments (17/23, 74%) were more frequently reported compared to Europe (16/87, 18% for consultations and 22/87, 25% for medical treatments).

XSL•FO
RenderX

Figure 7. For caregivers who reported having out-of-pocket costs related to cyclin-dependent kinase-like 5 deficiency disorder management (n=110), the types of costs covered by themselves by region (Europe, n=87; outside Europe, n=23) are presented in terms of the percentage of respondents who selected each expense type.

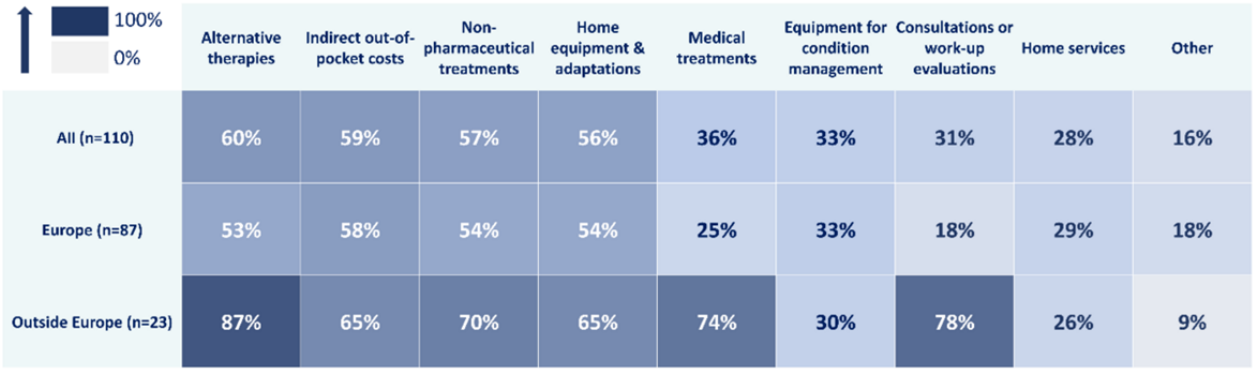

Caregivers who reported professional life disruptions due to the disease faced greater difficulties in covering these costs (41/85, 48%) compared to those whose employment was not impacted (9/45, 20%).

Unmet Needs

When asked what kind of additional help they would like to have received, family caregivers mostly reported wishing for help with daily care and everyday activities (60/132, 45.5%) and improved medical care (57/132, 43.2%). Caregivers less frequently mentioned administrative and financial help (22/132, 16.7%), social and emotional support (20/132, 15.2%), and information on CDD (16/132, 12.1%).

Discussion

Main Results

In this study, we found that the perceived burden of CDD was substantial for patients and their caregivers. Using a tailored version of the EQ-5D-5L for this condition for the first time, we were able to better assess patients' health-related quality of life. Caregiver burden was also reported to be significant, with CDD affecting multiple life aspects. Unmet needs were identified and appeared to be population-dependent, as variations between age groups and world regions were observed.

As previously reported for CDD, patients in this study had an early onset of seizures (median of 2.0 months), and most experienced daily seizures (66%), often of multiple types [4,9,14,27]. Variations regarding the proportion of patients experiencing different types of seizures, especially absence seizures, can be observed between this work and previous ones [14]. Identification of seizure types can be challenging for caregivers, as CDD seizures can present mixed motor and sometimes nonmotor features [14]. It is notably possible for caregivers to mistake epileptic spasms for hypermotor-tonic-spasm sequence seizures. Symptom load was high, with 57% of caregivers reporting eight or more symptoms. Over two-thirds of patients were hospitalized for CDD-related symptoms over the past year, and the proportion of hospitalized patients increased with the frequency of seizures. This aligns with previous findings that most hospitalizations of patients with CDD are related to seizures [29]. Almost all patients with seizures were on one or more ASMs (97%), and 82% of them discontinued at least one ASM. In previous work, the number of seizure-related hospital admissions increased with the number of ASMs used during the lifetime of patients [29]. All these findings may be an indication of the refractoriness of seizures, which is characteristic of CDD epilepsy [6,9,18,23,35].

The steady decline in caregiver satisfaction with the effectiveness of treatments in reducing seizures as time passed could be a further indication of potential treatment fatigue or habituation. Dissatisfaction with side effects, which usually consist of increased fatigue, irritability, and behavior disturbance, was also reported [18]. Polypharmacy, which was common in our study population, was previously associated with a higher likelihood of experiencing side effects and a lower quality of life for patients [18,28]. A ketogenic diet previously showed promising results for a short-term decrease in seizure frequency but poor long-term efficacy, which might explain the decrease in patients treated with this approach as they age [24].

The EQ-5D-5L is a standardized instrument used to quantitatively measure health-related quality of life, and it involves questions on health status [30,36]. In this study, most caregivers reported the highest severity in 3 out of the 5 dimensions measured: mobility, self-care, and everyday activities. Moreover, index values were low, with a median of 0.18, and were even negative for 17 respondents, indicating a perceived health state worse than death. This is consistent with previous research indicating that children with intellectual disabilities had lower EQ-5D scores than other populations [37]. Overall, these findings highlight the poor health-related quality of life of patients with CDD as perceived by their caregivers. The findings align with previous work using the Quality of Life Inventory Disability (QI-Disability), an instrument developed for children and adolescents with intellectual disability, which showed an association between more limited functional abilities and a poorer quality of life in patients with CDD [28,38]. These results provide some baseline data for comparison when using EQ-5D-5L as a standardized method for quantitatively assessing health-related quality of life in patients of all ages, enabling the analysis of various conditions and contributing factors. Furthermore, it may serve as a valuable complement to the QI-Disability scores in studies involving children and adolescents with CDD.

XSL•FO
RenderX

Family caregiver burden was reported as substantial in several aspects, both in this study and in previous work [25,26]. Most caregivers (64%) reported that their employment status was impacted by CDD. CDD care involved a wide range of health care professionals and was demanding and time-consuming, with the majority of patients requiring at least one appointment per week. Caregivers whose employment was impacted by CDD reported attending more appointments compared to those whose employment was not affected, highlighting the time-intensive nature of multidisciplinary care. Additionally, these caregivers faced greater financial challenges in covering costs related to the disease. This is a crucial issue, as a previous study has shown that financial strain is closely linked to a decline in parental well-being for CDD caregivers, emphasizing the importance of financial support for families [25].

Overall, caregivers reported substantial impacts of CDD on their family, social, and professional lives, as well as on their quality of sleep and stress levels. The burden of seizures was particularly significant for caregivers, with 73% identifying seizures as one of the most challenging symptoms to manage and reporting a greater impact on their own quality of life. This is consistent with findings that epilepsy, even in the absence of CDD or developmental delay, is associated with poorer family functioning [39].

These burdens and resulting unmet needs should be considered in the context of the profiles of both caregivers and patients, as noticeable differences were highlighted between world regions and age groups.

Caregivers outside Europe seemed to have more difficulties with access to resources and support, with only 19% of them reporting having access to a specialized CDD center compared to 51% of European respondents. Delay between the first seizure and diagnosis was also longer outside Europe than in Europe (26.5 months vs 11 months). This difference may be attributed to several factors, including varying levels of awareness of CDD, access to specialized medical centers, and the availability or reimbursement of genetic testing across different countries. CDD care varied by geographic location, with differences in the types of health care professionals consulted between European countries and countries outside Europe. The proportion of patients on ASMs and a ketogenic diet were similar across world regions. VNS was however reported exclusively among patients living in central, western, northern, and eastern Europe and was not reported among patients living outside Europe or in southern Europe. Major disparities in the availability of resources for neurological disorders exist across countries, with low-income countries being most severely disadvantaged [40]. These disparities could impact the level of care available for patients with CDD, potentially adversely impacting their quality of life [28].

Financial strain was also more severe for caregivers outside Europe, with a median rating of CDD impact on financial resources of 10 out of 10 compared to 6 out of 10 for European caregivers. Non-European caregivers reported greater difficulties in covering out-of-pocket fees related to medical treatments, medical appointments, or work-up evaluations, which were less frequently reported by Europeans. This may further highlight differences in health care systems and may suggest that European families benefit from better coverage and financial support, reducing the financial burden of CDD compared to families in other countries. Caregivers outside Europe were also more impacted in their professional life. All these difficulties had an impact on their quality of life, as caregivers outside Europe had higher median ratings of CDD impact on stress levels, social life, and general quality of life. All these findings are aligned with previous ones regarding the impacts of a family's residential location and especially the presence of universal health care on financial well-being, which is often closely linked to overall well-being [25,41].

Many CDD symptoms are age-dependent [7,10,14]. In our survey, patients aged 5-9 years exhibited a higher number of symptoms, had more varied and frequent seizures, and were on more medications. This may be due to the course of the disease or individual variability, or because symptoms become more detectable by parents from the age of 5 years [10]. These patients also had poorer functional ability, as reflected by their EQ-5D-5L index values and VAS results. Adult patients in this study were reported to be notably less severely affected than children, exhibiting less frequent seizures, reduced visual impairment, and fewer walking difficulties. Moreover, they were reported to receive less comprehensive care, with fewer appointments and fewer different types of health care professionals involved. This suggests that the transition to adult care may pose challenges for patients with CDD, resulting in poorer care. Further evidence would be needed to better understand the biological and nonbiological factors that could impact the life and care of adult patients, as well as the rationale behind them being reported to be less affected than younger patients.

This highlights the need for an evolving, age-specific multidisciplinary care approach, as supported by previous work by Amin et al [3]. Finally, the age at diagnosis was reported to be higher for older patients, which may be explained by the relatively recent discovery of the disease in 2004 and the growing accessibility of genetic testing [2,14]. At present, patients are diagnosed at a younger age, and this trend is expected to continue, with the age at diagnosis anticipated to decrease and stabilize in the coming years.

Strengths and Limitations

This study is one of the few studies illustrating the burden and impact of CDD on patients and caregivers with 132 respondents, mainly from western parts of Europe, Latin America, and Turkey. Clinical features of patients reported by caregivers were consistent with the epidemiological and clinical characteristics reported in the literature, including female predominance, early onset of seizures, impaired development, limited communication and gross motor skills, high seizure frequency, and multiple seizure types [5,10,14,42,43]. Treatments reported aligned with those commonly described in the literature [18,23,24]. Despite self-reporting by caregivers, the consistency of reported medical data with the literature verifies the representativeness of our study's sample.

To measure health-related quality of life, we used an adapted and validated version of a standardized tool, the EQ-5D-5L, as

XSL•FO
RenderX

described in the methods. This tool has been used in several studies and demonstrated excellent psychometric properties across a broad range of populations [44]. However, dimension scores for conditions associated with intellectual disability are usually much lower than for other chronic health conditions but map to other measures of the dimension (eg, mobility and self-care), highlighting that scores reflect relevant impairments and functions [28,36,37]. Different value sets were used to obtain the EQ-5D-5L index values of countries, complicating cross-country comparisons, as differences may result from the value sets rather than from the respondents' answers [45] (Multimedia Appendix 2).

In this study, comparisons between European respondents (n=106) and those from outside Europe (n=26; mostly from Southern America and Turkey) were unbalanced, with a higher number of participants from Europe and especially from Western Europe. Patients' medical characteristics varied between countries, and these disparities may have influenced the cross-country analysis, limiting the generalizability of the results. Participation in the survey was voluntary, with most participants recruited via support organizations and the Carenity platform. This may have led to an overrepresentation of caregivers for patients with severe symptoms, while isolated caregivers or those without internet access were underrepresented. This also may have impacted the geographic distribution of the respondents, as they were recruited from locations where active support organizations exist. Additionally, women and younger adults are often overrepresented in online communities, potentially influencing the sample composition [46].

If a respondent stopped and closed the questionnaire during completion, it was technically not possible for them to resume where they left off and they had to start over. This limitation may have contributed to a reduction in the number of fully completed and analyzable surveys.

The same team member conducted both the quality check and analysis, which may have introduced slight bias. However, another analyst, independent of the project, reviewed both the quality check and analysis results.

Future Directions and Implications

Our findings highlight the critical need for new effective ASMs as part of CDD treatment, benefiting not only patients but also their families, whose well-being is closely tied to the management of the condition. The use of a tailored version of the EQ-5D-5L to quantify health status in this survey generated values that could potentially serve as reference points for future studies to evaluate the impact of treatments, other therapies, practices, or other factors. As the EQ-5D-5L is a measure of health-related quality of life and is therefore influenced by impairments, it could benefit from being complemented by a measure of quality of life such as the QI-Disability score.

Solutions regarding support and access to resources must be adjusted depending on the family location, as considerable geographical discrepancies exist. The care provided for patients with CDD should evolve as they age, be multidisciplinary, and be adapted to the clinical features exhibited by the patient, as significant individual variations exist for this disorder [10].

Conclusion

Our results reveal the significant burden of CDD on both patients and caregivers, affecting caregivers' quality of life, including financial resources, sleep, stress levels, and professional and social aspects. Seizures in particular have a profound impact on caregivers' quality of life, underscoring the need to develop efficient and long-lasting ASMs. Furthermore, the study highlights geographical discrepancies regarding access to specialized medical centers, diagnosis delays, health-related cost coverage, and financial resources. These findings suggest that disease management and support should be tailored to the specific needs and health care systems of each region.

For the first time, the health-related quality of life of patients with CDD was measured by caregivers using a tailored version of the standardized and validated EQ-5D-5L tool, which provides a valuable baseline for future research.

The observed differences between age groups emphasize the need for multidisciplinary care for all patients, including adult patients with CDD, who appear to receive less comprehensive care. Further research is needed on caregivers outside Europe, children aged 5-9 years, and adult patients.

Acknowledgments

The authors would like to thank all the participants involved in this survey and the patient advocacy groups involved in the recruitment. The authors thank Angel Aledo-Serrano, MD, PhD, from the Epilepsy Program, Department of Neurology, Ruber Internacional Hospital, as well as Sandra Lopez Cabeza from Asociación de Afectados CDKL5 in Spain for being part of the steering committee and for participating in designing this study. The authors thank Toni Sarapohja, MSc, Head of Biostatistics and Medical Writing at Orion Corporation, who reviewed the statistical analysis plan and this manuscript. The authors also thank Cynthia Lesbros, PhD, project manager at Carenity, who participated in the supervision of this project and the analysis and interpretation of the results, and reviewed this manuscript.

This study was funded by Orion Corporation. Alongside Carenity, this sponsor was involved in the study design, data interpretation, and writing of the manuscript, and was kept informed during data collection and data analysis. All authors participated in the drafting and reviewing of the manuscript and have read and approved the final version.

XSL•FO
RenderX

Data Availability

The datasets generated or analyzed during this study are not publicly available due to personal data protection policies but are available from the corresponding author upon reasonable request.

Conflicts of Interest

HA and MAK were employees of Orion Corporation at the time the research was conducted. EP and VF were employees of Carenity at the time the research was conducted. CAP provided consultancy for Orion, and received an honorarium from Ultragenyx. HL provided consulting for Marinus, Acadia, Avexis, Orion, Neurogene and Taysha and conducted clinical trials with Anavex and Newron. All remuneration has been made to her department. JD provided consulting for Marinus, Ultragenyx, Acadia, Avexis, Orion, Takeda, Neurogene, and Taysha; conducted clinical trials with Anavex and Newron; and was a consulting/advisory board member for SCN2A Australia. All remuneration has been made to her department. SA has received funding from GW Pharmaceuticals, Novartis, PTC Therapeutics, Boston Scientific, Nutricia, UCB, BioMarin, LivaNova, Medtronic, Desitin, Ipsen, Orion, CDKL5 UK, TSA, University of Bristol, CASK Foundation, Elizabeth Blackwell Institute, and the National Institute for Health Research.

Multimedia Appendix 1
Survey questionnaire.
[DOCX File , 72 KB - formative_v9i1e72489_app1.docx]

Multimedia Appendix 2
Value sets used for index value calculation by country.
[DOCX File , 16 KB - formative_v9i1e72489_app2.docx]

Multimedia Appendix 3
Additional figures related to geographic location, age, and seizure frequency.
[DOCX File , 532 KB - formative_v9i1e72489_app3.docx]

Multimedia Appendix 4
CHERRIES checklist.
[PDF File (Adobe PDF File), 235 KB - formative_v9i1e72489_app4.pdf]

References

1. Symonds J, Zuberi S, Stewart K, McLellan A, O'Regan M, MacLeod S, et al. Incidence and phenotypes of childhood-onset genetic epilepsies: a prospective population-based national cohort. Brain 2019 Aug 01;142(8):2303-2318 [FREE Full text] [doi: 10.1093/brain/awz195] [Medline: 31302675]
2. Weaving LS, Christodoulou J, Williamson SL, Friend KL, McKenzie OL, Archer H, et al. Mutations of CDKL5 cause a severe neurodevelopmental disorder with infantile spasms and mental retardation. Am J Hum Genet 2004 Dec;75(6):1079-1093 [FREE Full text] [doi: 10.1086/426462] [Medline: 15492925]
3. Amin S, Møller RS, Aledo-Serrano A, Arzimanoglou A, Bager P, Jóźwiak S, et al. Providing quality care for people with CDKL5 deficiency disorder: A European expert panel opinion on the patient journey. Epilepsia Open 2024 Jun;9(3):832-849 [FREE Full text] [doi: 10.1002/epi4.12914] [Medline: 38450883]
4. Kadam S, Sullivan B, Goyal A, Blue M, Smith-Hicks C. Rett syndrome and CDKL5 deficiency disorder: from bench to clinic. Int J Mol Sci 2019 Oct 15;20(20):5098 [FREE Full text] [doi: 10.3390/ijms20205098] [Medline: 31618813]
5. Fehr S, Wilson M, Downs J, Williams S, Murgia A, Sartori S, et al. The CDKL5 disorder is an independent clinical entity associated with early-onset encephalopathy. Eur J Hum Genet 2013 Mar 8;21(3):266-273 [FREE Full text] [doi: 10.1038/ejhg.2012.156] [Medline: 22872100]
6. Hector R, Kalscheuer V, Hennig F, Leonard H, Downs J, Clarke A, et al. variants: Improving our understanding of a rare neurologic disorder. Neurol Genet 2017 Dec;3(6):e200 [FREE Full text] [doi: 10.1212/NXG.0000000000000200] [Medline: 29264392]
7. Jakimiec M, Paprocka J, Śmigiel R. CDKL5 deficiency disorder-a complex epileptic encephalopathy. Brain Sci 2020 Feb 17;10(2):107 [FREE Full text] [doi: 10.3390/brainsci10020107] [Medline: 32079229]
8. Wong K, Davies G, Leonard H, Downs J, Junaid M, Amin S. Growth patterns in individuals with CDKL5 deficiency disorder. Dev Med Child Neurol 2024 Apr 07;66(4):469-482 [FREE Full text] [doi: 10.1111/dmcn.15777] [Medline: 37804112]
9. Fehr S, Wong K, Chin R, Williams S, de Klerk N, Forbes D, et al. Seizure variables and their relationship to genotype and functional abilities in the CDKL5 disorder. Neurology 2016 Nov 22;87(21):2206-2213. [doi: 10.1212/WNL.0000000000003352] [Medline: 27770071]

XSL•FO
RenderX

10. Fehr S, Downs J, Ho G, de Klerk N, Forbes D, Christodoulou J, et al. Functional abilities in children and adults with the CDKL5 disorder. Am J Med Genet A 2016 Nov 16;170(11):2860-2869. [doi: 10.1002/ajmg.a.37851] [Medline: 27528505]
11. Fehr S, Leonard H, Ho G, Williams S, de Klerk N, Forbes D, et al. There is variability in the attainment of developmental milestones in the CDKL5 disorder. J Neurodev Disord 2015 Jan 05;7(1):2 [FREE Full text] [doi: 10.1186/1866-1955-7-2] [Medline: 25657822]
12. Keeley J, Benson-Goldberg S, Saldaris J, Lariviere J, Leonard H, Marsh ED, et al. Communication of individuals with CDKL5 deficiency disorder as observed by caregivers: A descriptive qualitative study. Am J Med Genet A 2024 Jul;194(7):e63570. [doi: 10.1002/ajmg.a.63570] [Medline: 38425131]
13. Mangatt M, Wong K, Anderson B, Epstein A, Hodgetts S, Leonard H, et al. Prevalence and onset of comorbidities in the CDKL5 disorder differ from Rett syndrome. Orphanet J Rare Dis 2016 Apr 14;11(1):39 [FREE Full text] [doi: 10.1186/s13023-016-0418-y] [Medline: 27080038]
14. Leonard H, Downs J, Benke T, Swanson L, Olson H, Demarest S. CDKL5 deficiency disorder: clinical features, diagnosis, and management. The Lancet Neurology 2022 Jun;21(6):563-576. [doi: 10.1016/s1474-4422(22)00035-7]
15. MacKay CI, Wong K, Demarest ST, Benke TA, Downs J, Leonard H. Exploring genotype-phenotype relationships in the CDKL5 deficiency disorder using an international dataset. Clin Genet 2021 Jan 20;99(1):157-165. [doi: 10.1111/cge.13862] [Medline: 33047306]
16. Amin S, Majumdar A, Mallick AA, Patel J, Scatchard R, Partridge CA, et al. Caregiver's perception of epilepsy treatment, quality of life and comorbidities in an international cohort of CDKL5 patients. Hippokratia 2017;21(3):130-135 [FREE Full text] [Medline: 30479474]
17. Amin S, Monaghan M, Aledo-Serrano A, Bahi-Buisson N, Chin RF, Clarke AJ, et al. International consensus recommendations for the assessment and management of individuals with CDKL5 deficiency disorder. Front Neurol 2022 Jun 20;13:874695 [FREE Full text] [doi: 10.3389/fneur.2022.874695] [Medline: 35795799]
18. Wong K, Junaid M, Alexander S, Olson HE, Pestana-Knight EM, Rajaraman RR, et al. Caregiver perspective of benefits and side effects of anti-seizure medications in CDKL5 deficiency disorder from an international database. CNS Drugs 2024 Sep 26;38(9):719-732 [FREE Full text] [doi: 10.1007/s40263-024-01105-z] [Medline: 39060900]
19. Gould A, Amin S. An overview of ganaxolone as a treatment for seizures associated with cyclin-dependent kinase-like 5 deficiency disorder. Expert Rev Neurother 2024 Oct 31;24(10):945-951. [doi: 10.1080/14737175.2024.2385937] [Medline: 39082513]
20. Knight E, Amin S, Bahi-Buisson N, Benke T, Cross J, Demarest S, Marigold Trial Group. Safety and efficacy of ganaxolone in patients with CDKL5 deficiency disorder: results from the double-blind phase of a randomised, placebo-controlled, phase 3 trial. Lancet Neurol 2022 May;21(5):417-427. [doi: 10.1016/S1474-4422(22)00077-1] [Medline: 35429480]
21. Olson H, Amin S, Bahi-Buisson N, Devinsky O, Marsh ED, Pestana-Knight E, et al. Long-term treatment with ganaxolone for seizures associated with cyclin-dependent kinase-like 5 deficiency disorder: Two-year open-label extension follow-up. Epilepsia 2024 Jan;65(1):37-45. [doi: 10.1111/epi.17826] [Medline: 37950390]
22. De S. Ganaxolone: first FDA-approved medicine for the treatment of seizures associated with cyclin-dependent kinase-like 5 deficiency disorder. Curr Med Chem 2024;31(4):388-392. [doi: 10.2174/0929867330666230320123952] [Medline: 36959132]
23. Lim Z, Wong K, Downs J, Bebbington K, Demarest S, Leonard H. Vagus nerve stimulation for the treatment of refractory epilepsy in the CDKL5 Deficiency Disorder. Epilepsy Res 2018 Oct;146:36-40. [doi: 10.1016/j.eplepsyres.2018.07.013] [Medline: 30071384]
24. Lim Z, Wong K, Olson HE, Bergin AM, Downs J, Leonard H. Use of the ketogenic diet to manage refractory epilepsy in CDKL5 disorder: Experience of >100 patients. Epilepsia 2017 Aug 12;58(8):1415-1422. [doi: 10.1111/epi.13813] [Medline: 28605011]
25. Mori Y, Downs J, Wong K, Anderson B, Epstein A, Leonard H. Impacts of caring for a child with the CDKL5 disorder on parental wellbeing and family quality of life. Orphanet J Rare Dis 2017 Jan 19;12(1):16 [FREE Full text] [doi: 10.1186/s13023-016-0563-3] [Medline: 28103894]
26. Mori Y, Downs J, Wong K, Heyworth J, Leonard H. Comparing parental well-being and its determinants across three different genetic disorders causing intellectual disability. J Autism Dev Disord 2018 May 30;48(5):1651-1665. [doi: 10.1007/s10803-017-3420-x] [Medline: 29192378]
27. Olson H, Demarest S, Pestana-Knight E, Swanson LC, Iqbal S, Lal D, et al. Cyclin-dependent kinase-like 5 deficiency disorder: clinical review. Pediatr Neurol 2019 Aug;97:18-25 [FREE Full text] [doi: 10.1016/j.pediatrneurol.2019.02.015] [Medline: 30928302]
28. Leonard H, Junaid M, Wong K, Demarest S, Downs J. Exploring quality of life in individuals with a severe developmental and epileptic encephalopathy, CDKL5 Deficiency Disorder. Epilepsy Res 2021 Jan;169:106521. [doi: 10.1016/j.eplepsyres.2020.106521] [Medline: 33341033]
29. Junaid M, Wong K, Korolainen MA, Amin S, Downs J, Leonard H. Measuring the burden of epilepsy hospitalizations in CDKL5 deficiency disorder. Pediatr Neurol 2025 Feb;163:68-75. [doi: 10.1016/j.pediatrneurol.2024.11.010] [Medline: 39705835]
30. EQ-5D-5L User Guide. EuroQol. URL: https://euroqol.org/publications/user-guides [accessed 2024-07-18]

XSL•FO
RenderX

31. EQ-5D-5L. EuroQol. URL: https://euroqol.org/information-and-support/euroqol-instruments/eq-5d-5l/ [accessed 2024-07-15]
32. Gerlinger C, Bamber L, Leverkus F, Schwenke C, Haberland C, Schmidt G, et al. Comparing the EQ-5D-5L utility index based on value sets of different countries: impact on the interpretation of clinical study results. BMC Res Notes 2019 Jan 14;12(1):18 [FREE Full text] [doi: 10.1186/s13104-019-4067-9] [Medline: 30642397]
33. Roudijk B, Janssen B, Olsen JA. How Do EQ-5D-5L Value Sets Differ? In: Devlin N, Roudijk B, Ludwig K, editors. Value Sets for EQ-5D-5L. Cham: Springer; 2022:235-258.
34. Section 1 : Définitions. (Articles R1121-1 à R1121-2). Legifrance. URL: https://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000043723460 [accessed 2025-04-28]
35. Olson HE, Daniels CI, Haviland I, Swanson LC, Greene CA, Denny AMM, et al. Current neurologic treatment and emerging therapies in CDKL5 deficiency disorder. J Neurodev Disord 2021 Sep 16;13(1):40 [FREE Full text] [doi: 10.1186/s11689-021-09384-z] [Medline: 34530725]
36. Karimi M, Brazier J. Health, health-related quality of life, and quality of life: what is the difference? Pharmacoeconomics 2016 Jul 18;34(7):645-649 [FREE Full text] [doi: 10.1007/s40273-016-0389-9] [Medline: 26892973]
37. Downs J, Norman R, Mulhern B, Jacoby P, Reddihough D, Choong CS, et al. Psychometric properties of the EQ-5D-Y-5L for children with intellectual disability. Value Health 2024 Jun;27(6):776-783. [doi: 10.1016/j.jval.2024.02.016] [Medline: 38467188]
38. Downs J, Jacoby P, Leonard H, Epstein A, Murphy N, Davis E, et al. Psychometric properties of the Quality of Life Inventory-Disability (QI-Disability) measure. Qual Life Res 2019 Mar 20;28(3):783-794. [doi: 10.1007/s11136-018-2057-3] [Medline: 30460513]
39. Rozensztrauch A, Kołtuniuk A. The quality of life of children with epilepsy and the impact of the disease on the family functioning. Int J Environ Res Public Health 2022 Feb 17;19(4):2277 [FREE Full text] [doi: 10.3390/ijerph19042277] [Medline: 35206465]
40. Atlas: country resources for neurological disorders, 2nd ed. World Health Organization. 2017. URL: https://iris.who.int/handle/10665/258947 [accessed 2025-01-27]
41. John Cherian D, Ta D, Smith J, Downs J, Leonard H. How families manage the complex medical needs of their children with duplication syndrome. Children (Basel) 2023 Jul 11;10(7):1202 [FREE Full text] [doi: 10.3390/children10071202] [Medline: 37508699]
42. Leonard H, Junaid M, Wong K, Aimetti AA, Pestana Knight E, Downs J. Influences on the trajectory and subsequent outcomes in CDKL5 deficiency disorder. Epilepsia 2022 Feb 27;63(2):352-363. [doi: 10.1111/epi.17125] [Medline: 34837650]
43. Benke TA, Demarest S, Angione K, Downs J, Leonard H, Saldaris J, et al. CDKL5 Deficiency Disorder. In: Adam MP, Feldman J, Mirzaa GM, editors. GeneReviews. Seattle, WA: University of Washington, Seattle; 2024.
44. Feng Y, Kohlmann T, Janssen MF, Buchholz I. Psychometric properties of the EQ-5D-5L: a systematic review of the literature. Qual Life Res 2021 Mar 07;30(3):647-673 [FREE Full text] [doi: 10.1007/s11136-020-02688-y] [Medline: 33284428]
45. Joelson A, Wildeman P, Sigmundsson FG, Rolfson O, Karlsson J. Properties of the EQ-5D-5L when prospective longitudinal data from 28,902 total hip arthroplasty procedures are applied to different European EQ-5D-5L value sets. Lancet Reg Health Eur 2021 Sep;8:100165 [FREE Full text] [doi: 10.1016/j.lanepe.2021.100165] [Medline: 34557854]
46. Raïs S, Radoszycki L, Dourgnon P, Rochaix L, Chekroun M. Accurate representation of patients' opinions for decision making: are online health communities good candidates? Carenity. URL: https://www.carenity.co.uk/static/doc/corporate/poster-ISPOR2017-Accurate-Representation-of-patients-opinions-for-decision-making.pdf [accessed 2025-05-12]

Abbreviations

ASM: antiseizure medication
CDD: cyclin-dependent kinase-like 5 deficiency disorder
CDKL5: cyclin-dependent kinase-like 5
QI-Disability: Quality of Life Inventory-Disability
VAS: visual analog scale
VNS: vagus nerve stimulation

XSL•FO
RenderX

Edited by A Mavragani; submitted 12.02.25; peer-reviewed by A Verrotti, F Pabon-Rodriguez; comments to author 31.03.25; revised version received 07.04.25; accepted 05.05.25; published 10.06.25.

Please cite as:
Amin S, Partridge CA, Leonard H, Downs J, Allvin H, Ficara V, Pain E, Korolainen MA
Caregivers' Perceptions of Clinical Symptoms, Disease Management, and Quality of Life Impact in Cases of Cyclin-Dependent Kinase-Like 5 Deficiency Disorder: Cross-Sectional Online Survey
JMIR Form Res 2025;9:e72489
URL: https://formative.jmir.org/2025/1/e72489
doi:10.2196/72489
PMID:

©Sam Amin, Carol-Anne Partridge, Helen Leonard, Jenny Downs, Helen Allvin, Valentine Ficara, Emilie Pain, Minna A Korolainen. Originally published in JMIR Formative Research (https://formative.jmir.org), 10.06.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Experiences of Individually Tailored Internet-Based Cognitive Behavioral Therapy During the COVID-19 Pandemic: Qualitative Interview Study

Victoria Aminoff[1], MSc; Matilda Baltius[1], MSc; Emelie Lundström[1], MSc; Matilda Berg[1], PhD; Gerhard Andersson[1,2,3], Prof Dr; Mikael Ludvigsson[2,4,5], PhD

[1]Department of Behavioural Sciences and Learning, Linköping University, Linköping, Sweden
[2]Department of Biomedical and Clinical Sciences, Linköping University, Linköping, Sweden
[3]Department of Clinical Neuroscience, Karolinska Institute, Stockholm, Sweden
[4]Department of Psychiatry, Linköping University, Linköping, Sweden
[5]Department of Geriatrics and Palliative Medicine, and Department of Health, Medicine and Caring Sciences, Linköping University, Linköping, Sweden

Corresponding Author:
Victoria Aminoff, MSc
Department of Behavioural Sciences and Learning
Linköping University
Olaus Magnus väg 42-44
Linköping, 581 83
Sweden
Phone: 46 761339009
Fax: 46 13288909
Email: victoria.aminoff@liu.se

Abstract

Background: During the COVID-19 pandemic, both physical and psychological health were at risk. Internet-based cognitive behavioral therapy (ICBT) is a psychological treatment alternative that does not inherently increase the risk of virus transmission because face-to-face interactions are not required. ICBT has been found to be effective for a variety of mental health problems, both before and during the COVID-19 pandemic. Although the experiences of undergoing ICBT have been investigated in previous studies, the specific experiences of participating in ICBT during the COVID-19 pandemic have been less examined.

Objective: This qualitative study aimed to investigate the experiences of participants undergoing individually tailored ICBT with weekly therapist support during the COVID-19 pandemic.

Methods: We approached trial participants who had received ICBT for psychological symptoms related to the COVID-19 pandemic during the summer of 2020. A strategic sample, based on the number of log-ins to the treatment platform, among other factors, was selected in an effort to achieve the highest possible variation. Semistructured telephone interviews were conducted 4 to 6 months after treatment completion, depending on whether the participant was initially assigned to the treatment or control group. Data were transcribed and then analyzed based on thematic analysis.

Results: A total of 16 participants aged between 23 and 78 years were interviewed. Four main themes and 10 subthemes were derived from the thematic analysis: (1) functions of the treatment (initiating and motivating, perspective widening), (2) treatment equals work (experience of the treatment as demanding, going from text to action, posttreatment engagement, participant agency), (3) changes experienced (changes in relation to the COVID-19 pandemic, other changes not related to the COVID-19 pandemic), and (4) wishing for something else (individually tailored, contact with the therapist).

Conclusions: The results closely align with those of previous qualitative studies on experiences of ICBT. Participants expressed appreciation of the treatment's content and format. Suggestions and wishes for changes were also expressed in the interviews. However, a unique finding was that participants described experiencing changes in well-being related to the COVID-19 pandemic. At the same time, there were also reports of changes in other symptoms not related to the pandemic. Further studies are needed on the experiences of participants who drop out of ICBT and the type of therapist contact they prefer.

(JMIR Form Res 2025;9:e66908) doi:10.2196/66908

XSL•FO
RenderX

KEYWORDS

internet-based cognitive behavioral therapy; COVID-19; internet; thematic analysis; qualitative methods; interview

Introduction

Background

Throughout history, there have been repeated occurrences of infectious diseases [1]. Increased globalization and human mobility, despite advancements in diagnostics and vaccines, have led to more complex consequences of disease spread [1]. The impact extends beyond physical health to include psychological [2], social, political, and economic dimensions [3]. This was salient during previous pandemics [4], as well as during the COVID-19 pandemic. Research investigating psychological health during the COVID-19 pandemic identified a coemerging psychiatric epidemic [5]. In addition to the fear of death it instilled, the pandemic also had an impact on family organization and work routines and led to the closure of public places, companies, and schools [5]. Furthermore, socioeconomic consequences such as unemployment [6] and reduced access to social support because of isolation and quarantine were reported [7]. An increasing number of studies investigating individuals infected by the SARS-CoV-2 virus compared to the noninfected population have found that those who were infected had higher levels of anxiety, depression, and posttraumatic stress symptoms [8,9]. In addition, strategies to prevent the spread of the virus, such as quarantine, were also associated with psychological distress, including symptoms of depression, posttraumatic stress, and insomnia [10,11]. Thus, the COVID-19 pandemic prompted action not only in response to physical health concerns but also in relation to mental health problems [5].

Internet-based cognitive behavioral therapy (ICBT) has been shown in a growing body of research to be effective for depression and anxiety disorders [12] as well as symptoms related to somatic conditions such as tinnitus and chronic pain [13]. ICBT is commonly delivered through a digital platform [14], and although the treatment can be provided with minimal or no therapist support [15], unguided treatments tend to be less effective than ICBT with scheduled therapist support [16]. Therapist-supported ICBT has shown effects comparable to those of face-to-face cognitive behavioral therapy (CBT) [17], with additional evidence supporting its effectiveness and acceptability in routine care settings [18].

An increased number of monthly course registrations for ICBT was observed during the early phase of the COVID-19 pandemic [19]. In line with this, there was a call in the literature for digital psychological health interventions during the pandemic [5,19,20], along with an emphasis on evaluating their effectiveness, quality, and safety [21]. Thereafter, an increasing number of studies have been conducted on ICBT and its effects on psychological symptoms during the pandemic. Our research team conducted a pilot study in the summer of 2020 to explore the effects of individually tailored ICBT during the early phase of the pandemic [22]. The results indicated that the treatment was effective for symptoms of depression, anxiety, and stress in participants who underwent ICBT compared to a waitlist control condition. The findings align with those of Wahlund et al [23], who examined a brief ICBT intervention for dysfunctional worry related to the COVID-19 pandemic. In addition to improvements in worry, the study found reductions in other symptoms, such as depressive symptoms [23]. A few months after the pilot study was conducted by our research team [22], a randomized controlled trial (RCT) of the same treatment demonstrated effects on symptoms of depression and insomnia as well as other reported symptoms [24]. In a meta-analysis, Komariah et al [25] summarized findings from 12 studies conducted during the COVID-19 pandemic and showed that ICBT could decrease symptoms of depression and anxiety. Similar effects of ICBT on symptoms of depression and anxiety have been reported during the COVID-19 pandemic as before the pandemic outbreak [19]. In line with previous research, ICBT with therapist support had superior outcomes compared to self-guided interventions [25], although both therapist-supported and self-guided ICBT were found to have effects on depressive and anxiety symptoms [26]. Most of the ICBT studies examined in the meta-analysis by Komariah et al [25] included participants regardless of whether they had contracted COVID-19. However, Liu et al [27] found improvements in depressive symptoms, anxiety, and insomnia among individuals specifically diagnosed with COVID-19. At the same time, not all studies have demonstrated ICBT to be effective for all aforementioned symptoms; for instance, Brog et al [28] found no effects of ICBT on depressive or anxiety symptoms, although an improvement in emotion regulation skills was observed. Overall, however, ICBT seems to have been a promising treatment option for psychological symptoms during the COVID-19 pandemic, as evaluated quantitatively through self-assessment questionnaires.

Several sociodemographic and psychosocial factors were proposed to be associated with declining mental health during the COVID-19 pandemic. This included, for example, gender—women had a higher risk of developing mental health problems than men [29,30]—and preexisting mental health problems, with comorbid physical and mental health conditions before the pandemic being predictive of worse outcomes [31]. Interviewing participants who underwent ICBT for psychological symptoms during the COVID-19 pandemic could provide a better understanding of how the treatment works for each individual and highlight problems experienced during treatment.

Qualitative methods complement quantitative data by capturing participants' views, thus providing insights into how interventions are experienced. Several studies have examined ICBT from various research perspectives before the COVID-19 pandemic, investigating participants' experiences with this form of treatment. These studies have focused on individuals with different mental health conditions who have undergone distinct ICBT programs for, for example, depression [32], anxiety [33], and procrastination [34]. In a systematic review, Patel et al [35] synthesized qualitative research and literature on participants' experiences of internet interventions aimed at targeting depression, anxiety, and somatoform disorders, including their

XSL•FO
RenderX

acceptability and usability. The authors identified 3 main themes (with subordinate subthemes): initial motivations and approaches, personalization of treatment, and the value of receiving personal support [35]. The first main theme concerned the important role played by the participants' initial motivations and attitudes for engaging with internet interventions; for example, the participants' approach toward the interventions at the beginning was found to have an impact on both recruitment and the therapeutic process. The second theme emphasized availability and flexibility, which could contribute to an increased sense of autonomy. With regard to the third main theme, Patel et al [35] reported that almost all studies they included in their review had noted the value of human support. The support was acknowledged as helpful in understanding the interventions better and enhancing commitment and motivation [35].

However, fewer studies have examined the experiences of participants undergoing ICBT during the COVID-19 pandemic. Among these, the study by Thapar et al [36] examined both patients' and therapists' perceptions of ICBT. The themes identified in the interviews they conducted included facilitators, such as the accessibility of registering for the program and the feasibility of completing the psychotherapeutic content of the program, as well as barriers, such as challenges with the web-based modality for developing a therapeutic alliance and limited opportunities for customization [36]. In another study, Ncheka et al [37] interviewed 50 individuals who had participated in a CBT program called Moodgym [38], identifying themes such as Moodgym improved understanding of how thoughts influence feelings and students experienced some challenges with Moodgym. Within the same field, adolescents' experiences of dialectical behavior therapy during the COVID-19 pandemic were also investigated, although the treatment was provided via telephone or video calls rather than through text-based modules and asynchronous communication with the therapist [39]. The themes identified included benefits of virtual dialectical behavior therapy and sense of loss [39]. In addition, experiences of CBT via video calls for depression during the pandemic have been explored, focusing on patients' and therapists' perspectives [40]. Both similarities, such as clinical efficacy, and differences, such as the structure of therapy, were identified in the experiences of both patients and therapists [40]. Thus, while some studies have focused on participants' experiences of ICBT during the COVID-19 pandemic, to the best of our knowledge, no study has specifically examined participants' experiences of individually tailored ICBT during the COVID-19 pandemic. The individually tailored approach could help address the previously noted need for greater ICBT customization [36].

Objectives

ICBT has been associated with positive experiences both before and during the COVID-19 pandemic. However, some participants have also reported negative experiences or unmet needs within the ICBT programs examined in previous studies. Therefore, further research on ICBT experiences remains necessary, particularly in the context of the COVID-19 pandemic because limited studies have been conducted under these conditions. Thus, this study aimed to investigate and develop a deeper understanding of participants' experiences of a 7-week individually tailored ICBT program with weekly therapist support during the COVID-19 pandemic, more specifically during the summer of 2020, as well as to investigate the experience of the treatment's usefulness. The program's effects were examined separately in a pilot RCT [22].

Methods

Study Design

This study was part of a larger RCT project on ICBT [22] (ClinicalTrials.gov NCT04424212). In that RCT project, adult individuals in Sweden with psychological symptoms related to the COVID-19 pandemic were recruited and randomly assigned to either a 7-week ICBT program with weekly therapist support or a waitlist control group. The control group received the same treatment 2 months later, after posttreatment measurements had been collected. The interviews reported in this study were conducted by telephone 4 to 6 months after the treatment ended, depending on whether the participant had been allocated to the immediate treatment or waitlist control group. This manuscript was prepared in accordance with the COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist [41].

Participants and Recruitment

Participants were recruited for this study from the aforementioned pilot RCT [22]. Briefly, participants were recruited for the RCT through advertisements on social media and in a national newspaper. Participants completed a web-based informed consent form and a pretreatment assessment, followed by a brief clinical telephone interview. As the intervention was provided through a digital platform, and both recruitment and eligibility assessment were conducted via the internet and telephone, individuals from across Sweden could apply to participate. A complete list of the inclusion and exclusion criteria for the pilot RCT is provided in the study by Aminoff et al [22]. The exclusion criteria were, among others, severe mental or somatic illness or acute suicidality that could hinder participation. Individuals meeting these criteria were referred to other health care services and encouraged to seek help through them (health care in Sweden is tax funded and does not require private insurance).

For this study, a strategic sampling approach aimed at maximum variation was used: we included participants from both the immediate and waitlist treatment groups in the pilot RCT [22]. Recruitment began with an email to 23 participants outlining the study's purpose and informing them that they had the right to decline or withdraw participation at any time without providing a reason and that a research team member would contact them by telephone to schedule an interview. Individuals who had not engaged with the treatment at all and those who had dropped out were excluded because the aim of the study was to explore participants' experiences of undergoing ICBT during the early phase of the COVID-19 pandemic. During the brief clinical telephone interview conducted after the treatment in the pilot RCT, participants had rated their improvement using the Clinical Global Impressions–Improvement scale [42], which measures perceived change in well-being, comparing the state just after treatment to the state before treatment. Participants

XSL•FO
RenderX

rate this change on a 7-point Likert scale, ranging from 1 (very much improved) to 7 (very much worse). We used these ratings to gather a broad range of experiences. Participants with varying numbers of log-ins to the treatment platform were also considered [14] because log-in frequency may reflect different levels of engagement with the treatment. Other aspects considered when striving for maximum variation in the sample were age, gender, educational level, and employment status.

The Pilot RCT Intervention

The pilot RCT treatment was a 7-week individually tailored ICBT program that included weekly therapist support. All participants received the same first and last module, namely, "Introduction" and "Conclusion and maintenance plan." For the remaining 5 modules, the participants could indicate their preferences. The final selection was made by the research team based on pretreatment assessments, the clinical telephone interview, and the participants' stated preferences. The therapists involved in the treatment were 1 licensed clinical psychologist and 3 recently graduated clinical psychologists with experience in ICBT who were supervised by the licensed clinical psychologist. The participants received 1 module per week, chosen from 14 possible modules. These optional modules focused on cognitive techniques, behavioral activation, emotion regulation, acceptance, stress management, perfectionism, relaxation, sleep strategies, problem-solving, anxiety and exposure, anxiety and worry, anxiety and panic, social anxiety, and imaginal exposure for difficult memories. In line with the purpose of the pilot study [22], the treatment modules aimed to address psychological symptoms related to the COVID-19 pandemic. The modules were adapted to the recommendations regarding the COVID-19 pandemic issued by Sweden's Public Health Agency. A detailed description of the modules is provided in the study by Aminoff et al [22]. All communication between the participants and the therapists took place on the same platform where the modules were delivered [14].

Data Collection

The interviews were conducted in February 2021 by telephone, allowing both interviewers and participants flexibility regarding location. The interviewers (M Baltius and EL) were either at home or at Linköping University, while the participants were at home. No other person was present with the interviewers, and to the best of our knowledge, no one else was present with the participants, except for one who mentioned holding her infant during the interview.

The interviewers had no contact with the participants before the study. The interviews began with the interviewers introducing themselves and reiterating the purpose and procedures of the study. Informed consent was also obtained at this stage. Pseudonyms have been used in the Results section to preserve the participants' anonymity. The semistructured interviews were conducted using an interview guide (Multimedia Appendix 1), adapted from Beukes et al [43], which was not pilot-tested because its structure and questions had previously proven adequate. The questions were mostly open ended and exploratory. Participants were asked about their expectations of the treatment as well as their experiences of the treatment and the contact with the therapists with questions such as "How did you experience the treatment?" "How has the treatment affected you and your situation?" Follow-up questions were asked to obtain richer and more detailed data. All interviews were conducted in Swedish, audio recorded, and then transcribed verbatim. The transcripts were not returned to the participants for validation or correction. The recordings were made using the Voice Memos app on an Apple device, transferred to a USB drive, and then deleted from the app. All transcriptions were saved on the same USB drive. The mean interview length was 19 minutes and 11 seconds (SD 7 min and 27 s), with the shortest interview being 8 minutes and 11 seconds and the longest being 35 minutes and 22 seconds.

Data Analysis

Overview

Data were analyzed using thematic analysis (TA) following the guidelines proposed by Braun and Clarke [44]. TA is used to identify and analyze patterns or themes that are meaningful to the research questions [44]. We chose TA with an inductive, bottom-up approach to openly explore and encapsulate data, rather than generate new theory or compare findings to previous research. In contrast to a deductive approach, this method is considered to be more data driven, not aiming to fit the data into already existing frameworks. However, it remains important to acknowledge that researchers' theoretical and epistemological orientations influence the analysis [44]. To explore the participants' experiences of the ICBT program, an essentialist and realist approach was adopted.

In line with the analysis procedure described by Braun and Clarke [44], we conducted the qualitative analysis according to the 6 phases outlined in the following subsections. Although Braun and Clarke [44] describe the analysis in 6 phases, they emphasize its recursive nature. Accordingly, we moved back and forth between phases rather than following a strictly linear process.

Phase 1: Familiarization With the Data

Initially, we transcribed the interviews verbatim and familiarized ourselves with the data by thoroughly reading the transcripts. An overview of the data material was established by actively and repeatedly reading all transcripts. This process also involved generating keywords and noting preliminary ideas.

Phase 2: Generating Initial Codes

We then systematically analyzed the entire dataset and generated initial codes. Codes represent the most fundamental element that can be seen as meaningful regarding the phenomena under study [44]. Thus, the data were organized into relevant categories upon which further analysis could be built.

Phase 3: Searching for Themes

Codes are narrower than themes, which we sought in this stage. Themes, defined as elements capturing something relevant to the phenomenon and research question being investigated [44,45], were identified by grouping codes that were assessed as sharing common characteristics. Themes reflect patterns in the data rather than their frequency [44] and were developed based on their capacity to meaningfully represent aspects of participants' experiences relevant to the research aim.

XSL•FO
RenderX

Phase 4: Reviewing Themes

In an iterative process, we sought codes, subthemes, and themes and compared them to each other. This iterative process aimed to digest and synthesize the findings continuously, as well as to enhance the confirmability of the analysis. Following the recommendations by Braun and Clarke [44], this process involved 2 stages: first, reviewing and refining the themes; and second, comparing them to the entire dataset. We assessed whether the interview extracts formed coherent patterns and whether the themes accurately reflected the dataset as a whole.

Phase 5: Defining and Naming Themes

Themes were defined and labeled. Through this process, the themes were refined by clarifying their core meaning. A detailed analysis of each theme and the relationships between the themes was conducted to identify potential overlaps and determine whether they were distinct from each other. The same process was conducted within and between the subthemes.

Phase 6: Producing the Report

Although writing and presenting the results is a continuous part of the process rather than something reserved for the final phase of analysis [44], the later stage of the work involved describing and reporting the themes, subthemes, and codes, along with selected excerpts that reflect the essence of each particular theme or subtheme. Subsequently, the themes, subthemes, codes, and chosen excerpts were translated from Swedish, the original interview language, into English.

Research Trustworthiness

As several members of the research team have prior experience in and knowledge of ICBT, qualitative research, and qualitative research on ICBT, we reflected on any preexisting assumptions. The research team included researchers of varying ages, genders, and levels of research experience. At the time the interviews were conducted, M Baltius and EL were MSc psychology students in their final (tenth) semester. VA was a PhD student and M Berg, GA, and ML held PhDs. VA, M Berg, GA, and ML were working at Linköping University in research and teaching roles. VA, M Baltius, EL, and M Berg are women, whereas GA and ML are men. M Berg, GA, and ML have prior experience with qualitative research, both within and outside the field of clinical psychology. VA had no prior experience conducting qualitative research but had completed relevant coursework at the master's and doctoral levels. M Baltius and EL had taken master's-level courses in qualitative methods and were supervised by the other authors.

Before the interviews, conducted by M Baltius and EL, who were not involved in the pilot RCT, the research team discussed the importance of open-ended questions and possible follow-up questions in the semistructured interview guide, as well as the importance of remaining receptive to information that might not align with previous findings or expectations. The data analysis was conducted independently by M Baltius, EL, and VA, with additional contributions from ML. GA and M Berg oversaw and supported the entire study, from the development of the interview guide to the qualitative data analysis. All authors reviewed and provided feedback on the analysis.

Ethical Considerations

The study involved human participants and focused on their experiences; thus, it was submitted for ethical review and approved by the Swedish National Ethics Committee (Dnr 2020-02313). Participants were sent an email containing information about the study and a description of the purpose of the interview. The email also informed participants that participation was voluntary, that the interviews would be recorded, and that their information and answers would be presented in an anonymous format. The participants were also informed that they could withdraw from the study at any time without providing a reason. The research team subsequently contacted the participants by telephone and provided the same information verbally, after which informed consent was obtained. No compensation was offered to the participants.

Results

Overview

Of the 23 participants who were invited to take part in the study, 16 (70%) completed the telephone interview, while 7 (30%) did not take part in the study. Of these 7 invitees, 3 (43%) could not be reached by email or telephone, while the remaining 4 (57%) declined to participate, citing time constraints or lack of interest. The participants' mean age was 51.87 (SD 18.64; range 23-78) years. Of the 16 participants, 11 (69%) were female, and 5 (31%) were male; moreover, 9 (56%) were from the initial treatment group and 7 (44%) were from the initial control group. The participants logged in to the treatment platform 34.44 (SD 18.82; range 11-85) times on average and completed an average of 6.19 (SD 1.05; range 4-7) modules. On the basis of these interviews, data saturation was considered to have been achieved.

Four main themes were identified regarding the participants' experiences—functions of the treatment, treatment equals work, changes experienced, and wishing for something else—with 10 related subthemes. A summary of the themes and subthemes, along with examples of codes and illustrative quotes, is presented in Table 1.

XSL•FO
RenderX

Table 1. Overview of the main themes and subthemes, along with examples of codes and illustrative quotes.

Main themes and subthemes	Examples of codes	Illustrative quotes
Functions of the treatment		
Initiating and motivating	Treatment as motivating, therapist helped to get activated, and became activated	"I found a lot of things that...like because the weather outside; it was boring; and I was tired and so on. I detest going by myself, still I need to exercise and come outdoors and so. And then, then I used this strategy to do the opposite of what one feels like doing, yes then I do an exercise. 'Do the opposite.' I go out, then. So, I got dressed and went out for a walk. And that was, well I...You know! Then I laughed out loud to myself. Was it this easy?" [P6; female; 67 y]
Perspective widening	Treatment gives hope, helps identify why feeling in a certain way, and identifies other problems	"[I]f I am stressed about something, then it is easier [now after the treatment] for me to take a step back and think about why. And see if you, for example, can solve the stress in a different way than just accept that you are stressed and do not feel good because of it...If I am having a bad day nowadays, then I can have a bad day for a while and then perhaps think that "well, then it will be a better afternoon and that yes, I let this go for a second and try again later." While before, I hung around longer with [the thought of] the bad day, so to speak." [P11; male; 40 y]
Treatment equals work		
Experience of the treatment as demanding	Demanding aspects of the treatment, obstacles to engaging with the treatment, and treatment perceived as demanding when feeling unwell	"[The treatment was] useful but tough. It is this I mean...the hardest [parts of the treatment] was probably also the best for me." [P6; female; 67 y]
Going from text to action	Practiced the treatment content, using strategy from the treatment, and concrete work with strategy	"[J]ust that I can go back and read, well, this was probably a thought trap, or that it happens that I automatically can think now and wonder if this is a thought trap that tries to trick me now." [P8; female; 54 y]
Posttreatment engagement	Experiencing use of the treatment afterward, continuous engagement with treatment strategy, and need for repetition	"[I]f one wants to get an effect, then one has to work with this all the time, yes but actually all the time, so to say." [P7; male; 42 y]
Participant agency	Experience of having to be active, adapting the treatment to self, and oneself has the responsibility	"My own contribution was the largest contribution. It was completely, completely decisive for how I feel today." [P6; female; 67 y]
Changes experienced		
Changes in relation to the COVID-19 pandemic	Treatment helped with managing the COVID-19 pandemic, treatment has helped managing pandemic consequences, and other circumstances helped to handle the pandemic	"[T]here are several such, eh, things, factors that should make me feel, even worse in the summer compared to the spring, but, but it did not turn out that way...and it is surely to some extent, eh, [thanks to] this [treatment] from which I have these tools and that I have received this." [P7; male; 42 y]
Other changes not related to the COVID-19 pandemic	Treatment has helped not only with the pandemic but also other things, experienced the treatment as helpful, and treatment helped with issues other than the pandemic	"I usually think November, like around that time, is pretty tough normally, but I felt better in November than I usually do and maybe it also has to do with the fact that I had precisely boosted on a little, like, received some tools [from the treatment] and could hold onto these tools." [P5; female; 45 y]
Wishing for something else		
Individually tailored	Treatment too general, treatment not individually tailored enough, and wish for more variation in the treatment content	"I did not really feel that it was directly related to my specific problems. And this is what happens with such a general treatment method...There were areas around the anxiety problem, they were so general because it [the treatment] should fit many." [P4; female; 46 y]
Contact with the therapist	Wish for verbal contact with therapist, wish for face-to-face contact with therapist, and request for other kind of contact with therapist	"One got great response, but it would also have been nice to have had someone to talk to beyond that...at least at some time. Some kind of telephone contact." [P3; female; 28 y]

Main Theme: Functions of the Treatment

Two different essential functions of the treatment were identified in the data corresponding to the subthemes initiating and motivating, which reflects a shift in the participants' level of activation regarding their mental health; and perspective widening, which highlights participants' new reflections and insights during the treatment.

Subtheme: Initiating and Motivating

During the treatment, participants received modules containing reading material and practical exercises, which encouraged them

XSL•FO
RenderX

to take an active role in the treatment. Allowing participants to request specific modules was an aspect of the treatment that was described as motivating and prompted active engagement in and reflection about the treatment. Even when not feeling well, just doing something about it can be a good thing. Taking control by being the one creating or structuring the day's activity was also mentioned as motivating. As another example, a participant described structure as a helpful way of getting started with positive activities and also noted that she did not need to think much about a positive activity (so-called plus activity) before engaging with it:

> *One thing that was more powerful than I had understood before was this about creating simple, good...Or creating structure at least, and to create simple, good "plus activities." [I realized] that it does not have to be a big effort, but instead that you make sure to not think so much before doing it...And just think to yourself: "Mm, today I have planned to do this when I quit work. Yes, then I will do it now," or something like that.* [P5; female; 45 y]

This participant stated that it did not require a big step to initiate an activity that had been on her mind. When she planned to engage in an activity, she acted on it without further deliberation. Through different aspects of the treatment, participants began, or became motivated to perform, certain everyday activities. This included activities they had enjoyed before the pandemic but did not engage in anymore. The treatment itself could also be viewed as a substitute for activities they missed because of the pandemic:

> *[I]t was actually fun to get to do something with the brain and to write a little and formulate as well, I thought. Because it was also...the problem I thought was that I, that I...I have been working for a long time and been active...It hit me a little hard that I...had to quit my job.* [P14; female; 75 y]

This participant described the treatment as stimulating. The treatment became something else in which she could become active and involved when she had to quit her job. Overall, the treatment was described as helping the participants engage with something actively, both within the treatment itself and with other activities. However, this was not the case for all. There were also descriptions of difficulty in finding motivation or energy to engage with the treatment.

Subtheme: Perspective Widening

Participants reported gaining new insights and broader perspectives on mental health and emotions as well as their personal situations during the treatment. The treatment helped participants identify feelings that bothered them (eg, if it was mainly fear or worry) and provide explanations for why these feelings had emerged:

> *I have received help to, like, [understand] what it was with this with corona...And that my freedom to make own decisions was taken away from me. Because it was this that was my problem, that, that it, independence was taken away from me. I was not allowed to go to work, I was forbidden to do things.* [P4; female; 46 y]

The perspective-widening function of the treatment also involved descriptions of feeling less lonely in one's life situation during the pandemic and in handling the COVID-19 situation:

> *[I]t was that I felt that I got...it [the treatment] gave me power and new thoughts and to see myself from the outside. And that, and that it was not that strange what I experienced, but it was many other [people who felt the same]...From what I understood from questions and such in the treatment, I understood that it was quite common to feel this...With the stress related to COVID.* [P14; female; 75 y]

The participant described gaining the perspective of observing herself from the outside and realizing that she was not alone in her experiences. To summarize, this function of the treatment involved gaining wider perspectives on which feelings were present and why they were present, as well as the realization that others also struggled with similar emotions.

Main Theme: Treatment Equals Work

Four subthemes were derived from the interviews regarding the treatment and its effects: experience of the treatment as demanding, going from text to action, posttreatment engagement, and participant agency. These subthemes corresponded to the participants' impressions of working with the treatment, how they engaged with the text-based material in practice, their continued efforts after the 7-week treatment period, and their own role in the treatment.

Subtheme: Experience of the Treatment as Demanding

Participants described the treatment as demanding in terms of both time and effort. One participant reflected on the timing of the treatment and emphasized the advantage of having more free time when engaging with it:

> *I did this [the treatment] when I was on vacation, so it was very good. Because, basically, it takes a lot of time and effort. Like to do these things.* [P8; female; 54 y]

The experience of the effort required was described as both tough and nice—sometimes simultaneously:

> *It can...Things that are a bit nice-pain. It is nice to work with it and eh work with it structured...And still, it is tough and hurts and I had to fight very much with certain things [components of the treatment].* [P6; female; 67 y]

The participant experienced the treatment work as both rewarding and painful. The interviews revealed that the workload was experienced as both moderate and overwhelming. As the treatment involved many components, it could be challenging to keep up: each week introduced a new module, making it difficult to engage in other activities.

Subtheme: Going From Text to Action

Descriptions of participants' engagement with the treatment and CBT strategies were provided in the interviews; for example, this could involve using a smartphone app to register

XSL•FO
RenderX

thoughts of worry that occurred during the day and later dealt with during scheduled worry time (ie, when working with the worry module). Other participants mentioned keeping the CBT strategies in mind:

> *[S]o instead of having a new nuisance every time, you have a name for it, and then it does not become as scary...yes, now that feeling comes again, and that I, and so you call it something...then it, like, this went good last time so it will also go well this time.* [P7; male; 42 y]

By noticing and recognizing a feeling, the participant practically used a common first step in emotion regulation. This involved observing the upcoming emotion in that very moment. The practical implications described in the interviews could also be on a more abstract level. A participant described how she had worked with behavioral activation:

> *I do not use this activity plan within the daily routine as strictly as I did then [during the treatment weeks]. But it exists in the sense that I think what I will do, eh. I engage in some of these plus activities each day.* [P6; female; 67 y]

Although this participant was not fully using behavioral activation as described in the treatment, the thought of "plus activities" was on her mind. In sum, the subtheme treatment equals work involved being influenced by the treatment and using it in both concrete and abstract ways.

Subtheme: Posttreatment Engagement

This subtheme involved statements regarding the continued use of the treatment strategies after completing the 7-week treatment schedule. The treatment could be viewed as continuously ongoing rather than something to finish:

> *I think that CBT and these things, they are actually, yes, it is something ongoing. It is nothing, it is not a wound that heals and then you are done, but it rather leads to new questions. You become equipped with a toolbox.* [P9; female, 70 y]

In the interviews, there were different ways in how engagement with the treatment was described; for instance, a participant described how he tried to handle feelings by viewing difficult feelings as a beach ball, metaphorically speaking:

> *[T]here are certain keywords...that I have taken with me, yes, that I still work with, so, yes, like this beach ball is one example. That if you notice that, "now I am trying to smooth things over here," and pretend like that it does not exist and so on, eh, and I should not do that then, but I should try, you are supposed to be with your feelings, I should learn to live with this, so to speak.* [P7; male; 42 y]

Participants shared that repeating a specific strategy was also considered important and helpful, especially in stressful situations, when strategies were not routinely applied. It was clear that the treatment had not fully prevented the occurrence of difficult situations, feelings, and thoughts but had rather provided strategies and ways to handle problems.

Subtheme: Participant Agency

Participants described their own role in the treatment, including control and responsibility; for instance, engagement was described as crucial for the treatment to work: although it was individually tailored in terms of the choice of modules, it was standardized because the specific modules were the same for all who received them. This is exemplified by a participant's description of her experience of the support she received from the therapist:

> *[I]t was the support I felt I needed there and then that I got because the other parts of the treatment are such things that lie with me, it is I who must like do the things, it is no one else but I who must do this and challenge.* [P8; female; 54 y]

The participant emphasized her own responsibility in the process, explaining that no one else—not even the therapist—could take over. Another participant gave an example of the demands she experienced:

> *Just this that one must go into the deep to be able to understand and absorb something. Yes. And making it to be my own [treatment], how can I use this in a good way?* [P15; female; 75 y]

This participant underscored the necessity of engaging in the treatment to make it relevant and adapt it to her situation. Various descriptions of the participants' relationship with their therapist emerged in the interviews, with the idea of being a team serving as 1 example. The relationship was described as collaborative work, rather than a one-way relationship in which the therapist delivered the treatment. Along these lines, a participant emphasized being in the driver's seat with responsibility:

> *[I]t is clear that it [the treatment] makes huge demands that you are the active one, you should be the proactive, you should drive it.* [P10; male; 48 y]

This interviewee described an experience characterized by responsibility and the obligations placed on him as a participant. Overall, the participant described his role in the treatment as vital but not as standing alone. In the interviews, it became clear that this role can only be taken on by the participants themselves to achieve the desired treatment effects.

Main Theme: Changes Experienced

As described by participants, the changes that emerged during and after the treatment differed in both cause and content. Accordingly, 2 subthemes were identified: changes in relation to the COVID-19 pandemic and other changes not related to the COVID-19 pandemic.

Subtheme: Changes in Relation to the COVID-19 Pandemic

Changes regarding symptoms related to the COVID-19 pandemic were described. As explained by a participant, the treatment could be used to manage the circumstances brought about by the pandemic:

> *I barely meet anyone. But...eh...and then there are many who ask, how do you cope? Yes, but I think I*

XSL•FO
RenderX

> *cope because I have...all this other stuff [the treatment] that I work with, and in that way, I have been helped by these tools.* [P9; female; 70 y]

Another way the treatment was described as helpful was by changing how the pandemic was viewed, for instance, how much worry it caused and how it was managed. These psychological changes were not necessarily linked to how the pandemic situation changed; for example, a participant described her mental health improving despite the pandemic worsening:

> *It was therefore it became like...almost another life. Before CBT, I did not know exactly how I felt, I can remember that. And now after this, I mean the pandemic, this has not changed. It [things] has even become much worse now for me. But I feel much better, and I can handle it.* [P6; female; 67 y]

The interviews also included accounts in which participants attributed their improved mental health to factors beyond the treatment. Other changes in their everyday life were also described, such as getting a new job.

Subtheme: Other Changes Not Related to the COVID-19 Pandemic

The treatment was also described as having an impact on issues not directly related to the pandemic. One participant realized that the treatment material would not only be useful during the pandemic but also in other areas of life:

> *[A]fter I got started, I felt that this is good and will be good for me like, it is not only now during the pandemic, but also like, yes, but you know, in working life and so on.* [P5; female; 45 y]

The treatment was also described as addressing psychological symptoms that existed before the COVID-19 pandemic began:

> *I was quite stuck before I started with the treatment, so, it was nice to get some keys to how I unlock my problems that I have actually had for a long time.* [P3; female; 28 y]

Main Theme: Wishing for Something Else

This main theme consisted of 2 subthemes: individually tailored and contact with the therapist. The main theme and subthemes reflect treatment components that participants wished had been included or implemented differently. Although not included in this theme or highlighted as a separate theme, participants also described the treatment as helpful and useful and the treatment format as applicable. The treatment was described as individually tailored, with the possibility to request modules relevant to one's situation and symptoms. Therapist contact was appreciated and at times described as essential. Nonetheless, as illustrated by this theme, participants expressed wishes for additional treatment components.

Subtheme: Individually Tailored

The treatment was individually tailored through the selection of modules, while the content of each module remained unchanged. As a result, the material was not always fully applicable to each participant. Even if the focus of the module had been judged as potentially helpful, the treatment was also described as too general to be useful:

> *[T]he treatment did not give me so much as it was too general. It was not as specific as I wanted it to be. Sure I could choose, eh, which areas I wanted to look into further or wanted to work with and so. But it was still that every area was so very general and large, that it was difficult to apply specifically for what I felt.* [P4; female; 46 y]

This participant wanted more specific treatment material. To make the treatment more user-friendly, accessible, and helpful, participants offered some concrete suggestions; for instance, a participant mentioned both her own preferences and what she thought could help other people:

> *[W]hat I missed very much was that there should have been small, short video recordings...I would really like to have visual, more that...and then it can be good to listen to someone who reads the text...I think that could make it easier to absorb the text for some people.* [P9; female; 70 y]

Subtheme: Contact With the Therapist

There were some suggestions or requests for different types of contact with the therapist within the treatment; for instance, a real-time chat function as a complement to the asynchronous text message function was suggested. Alternatives to internet-based contact with the therapist were also mentioned, such as telephone contact:

> *[I]t would have been good if one had both, at least for me. Like some human contact too and not just email...that there could have been a mix of it email and telephone contact then I think it would have been better.* [P10; male; 48 y]

Another example of a request was for face-to-face meetings with the therapist. One participant expressed a wish to meet the therapist in person to discuss the problems she was experiencing:

> *I would personally have needed a little more direct personal contact...I felt I did not have anyone to bounce ideas directly with. Well, sure, I could write to my therapist, and I received some good answers, but I would have needed to meet someone. Where one can sit down and discuss things...I have nothing against CBT, well, I believe in the method...but I would have preferred a face-to-face CBT and not an internet-based.* [P4; female; 46 y]

Discussion

Principal Findings

This study aimed to explore participants' experiences of undergoing individually tailored ICBT with weekly therapist support during the COVID-19 pandemic. The intention was also to investigate the experienced usefulness and possible limitations of the treatment. After semistructured telephone interviews, data analysis, and interpretation, 4 main themes and 10 underlying subthemes were identified. The 4 main themes

XSL•FO
RenderX

generated were functions of the treatment, treatment equals work, changes experienced, and wishing for something else.

Limitations

There are several limitations to this study that should be addressed. First, as in any study involving participants who have received treatment, there is a risk that social desirability influenced the responses. The interviewers had not had any contact with the participants before the interviews, that is, they were not the participants' therapists. However, participants may still have withheld honest responses to avoid disappointing the interviewers, who are members of the research team. Hopefully, participants understood that they had the chance to express what they did not like about the treatment and suggest changes. Questions about this aspect were included in the interview guide; moreover, the people who administered the interviews were not involved in the project when the pilot RCT [22] was conducted. Second, the shortest interview lasted only 8 minutes, while the mean interview length was 19 minutes, with some interviews being longer. There is a risk in all interviews, but especially in the shortest ones, that important aspects may have been missed, and the shorter interviews could be due to the interviews being conducted over the telephone. However, it could also be that the participants experienced that it was more convenient to give feedback when not having eye contact with the interviewer. In addition, in line with the discussion by Biliunaite et al [46], participants who agreed to take part in the study could be more positive about the treatment compared to those who declined to participate. The results might have differed if the 7 invited individuals (n=3, 43% were unreachable; n=4, 57% declined) who did not participate had been included. We recognize the need for broader recruitment to further explore the experiences of individuals who felt that the ICBT was not suitable for them. Moreover, the participants were interviewed approximately 4 to 6 months after completing the treatment. There is a risk of recall bias because people in the control group, receiving the treatment later and completing the interviews closer to the treatment, may have had the treatment more readily available in their memory. However, this was not observed as a systematic problem by the interviewers or by the other authors when analyzing the data. The relatively long period of time between the treatment and the interviews could have affected what the participants remembered from the treatment, and answers could have been different if the interviews had been conducted directly after the treatment concluded. At the same time, the delay between treatment and interview could also have allowed participants time to reflect and assess what, if anything, they found useful regarding the treatment, which was part of our aim. Thus, there was an opportunity to investigate whether participants continued to use the treatment material after treatment conclusion and whether this had been helpful to them, as illustrated by the identified subtheme posttreatment engagement.

It is possible that our prior understanding and knowledge of ICBT impacted the analysis of the results. Several members of the research team have extensive experience with ICBT and with qualitative methodology. Furthermore, all authors, except for the 2 interviewers, were involved in the pilot RCT [22]. However, this is not necessarily a problem because researcher subjectivity can be viewed as an advantage [47]; for example, when participants sometimes referred to a certain module or assignment within the treatment, the interviewers and authors could follow up with clarifying questions because they were familiar with the treatment approach and the modules. It is also important to acknowledge that there could be other possible ways to code and interpret the participants' statements. Furthermore, it is possible that we would have ended up with different results if we had used an alternative qualitative analysis method. The process of thematization or pattern identification can be found in other qualitative approaches, such as interpretative phenomenological analysis or grounded theory [45]. However, both interpretative phenomenological analysis and grounded theory are theoretically grounded, in contrast to TA [44]. TA is not inherently linked to any particular theory or epistemological perspective, which we deemed most suitable for our purpose. Moreover, the results might have differed if a different semistructured interview guide had been used, for instance, one with a greater focus on core CBT principles. Our intention with the broad questions was to provide participants with as much space as possible to freely reflect on their experiences, particularly because the ICBT program was individually tailored, and not all participants engaged with the same materials and CBT strategies. However, more focused questions could have served as a reminder of the components to which the participants had access during the treatment and potentially elicited more detailed descriptions of their experiences.

It is also worth mentioning that, as with other qualitative research, the possibility of generalizing the results is limited and is not even a goal of the research [48]. However, the themes identified do provide us with information on how the participants in this study experienced undergoing ICBT during the COVID-19 pandemic. The results are to a great extent similar to those of previous qualitative studies in the ICBT research field [49,50]. This supports the notion of transferability, which, rather than generalizability, is commonly used as a quality criterion in qualitative research [51]. As such, the findings may also be informative for other ICBT programs tested under more typical conditions, when the COVID-19 pandemic is no longer a major societal concern.

Comparison to Previous Work

Considering the first main theme—functions of the treatment—participants described in different ways how the treatment and their work with the treatment affected them. The motivating and activating function that was described aligns with previous findings regarding increased self-efficacy as a positive aspect of ICBT [34]. The function of the treatment as perspective widening also mirrors previous qualitative studies in the ICBT field [49,50,52,53]. Participants described gaining new insights and a broader understanding of their mental health, contributing to an increased sense of control. Similar experiences have been reported by other individuals who underwent ICBT during the COVID-19 pandemic [37]. As important as these findings are, it is not surprising that the treatment can help one understand oneself and the problems experienced, in particular given that psychoeducation is part of the treatment. Psychoeducation alone has been shown to be effective for

XSL•FO
RenderX

symptoms of depression and psychological distress [54]. A possible extension of this finding is that identifying and understanding one's feelings might be therapeutic in itself. Affect labeling—putting feelings into words—has been observed to have an emotion-regulating function [55].

Regarding the second main theme, related to the treatment effects being dependent on work and effort, participants described the treatment as demanding in terms of time and energy. This has been reported before in a previous qualitative study on ICBT [50]. In this study, descriptions about this experience were salient both during the treatment period and afterward, which illustrates the importance of motivation. Moreover, participants described their role and agency in the treatment, which can be seen as reflecting not only self-efficacy but also a recognition of their responsibility for their own mental health. This active approach to the treatment, both during the treatment period and afterward, is in line with previous findings that emphasize the importance of processing and applying the treatment material [49], as well as the value of autonomy and self-sufficiency [52].

The original purpose of the pilot RCT [22] was to investigate the effects of ICBT on psychological symptoms related to the COVID-19 pandemic. It became evident in this qualitative investigation, as illustrated by the third main theme about the changes experienced, that working with the treatment was not isolated to pandemic-related symptoms. The modules used in the treatment were derived from previous ICBT studies, conducted before the pandemic, and dealt with problems that very well could have been present before the pandemic. However, the modules were adapted and focused on the COVID-19 pandemic context. In addition, this is in line with the notion that comorbidity is common [56] and the conception that common underlying mechanisms, such as anxiety sensitivity [57], contribute to a range of psychological symptoms [58,59].

The fourth and final theme concerned the participants' wishes and suggestions regarding how the treatment format and the therapist contact could have been designed differently to better suit their needs. Some of the suggested changes have been raised in previous studies [46,50,60]; for example, the experience of ICBT not being sufficiently individually tailored has been reported both before [61] and during the COVID-19 pandemic [36]. It was mentioned that the existing form of tailoring was not sufficient to make the treatment useful for everyone. People in general and the participants in this study were impacted by the COVID-19 pandemic in very different ways, with some, for example, experiencing worry as their main problem, while others reported feeling more depressed. Creating different combinations of modules in this study was an attempt to address this variety of needs, but the way the tailoring was implemented could be improved further in future studies [62].

Similar to the findings by Asplund et al [50], participants in this study expressed a need for extended or alternative forms of support and contact with their therapist, such as through telephone or video. At the same time, other participants indicated that they did not feel a significant need for extended contact with their therapist or appreciated that the contact was only through asynchronous text messages. No participants reported problems with the technology, such as logging into the platform. However, as discussed by Patel et al [35], this may be due to the likelihood that the participants interviewed, who had chosen to take part in an ICBT study, were computer literate.

The findings of this study are valuable in enhancing the understanding of how participants may experience undergoing ICBT, particularly in the context of the COVID-19 pandemic. The results largely align with previous research on experiences of internet-based interventions, highlighting themes such as the development of understanding and perspectives on thoughts and feelings [37], the importance of the participant's active involvement in the treatment process, and the significance of both treatment personalization and individual feedback [35]. These insights further expand the current perspectives on how ICBT may be perceived, elucidating the functions the treatment is seen to serve, how participants' engagement with the treatment is described and experienced, and what aspects would need to be altered or improved. A distinctive contribution of this study, in comparison to previous research conducted both before [35] and during the COVID-19 pandemic [36,37], is the exploration of participants' perceptions of the underlying causes and specific content of the changes attributed to ICBT. This includes how participants described changes in their mental well-being, both in relation to the COVID-19 pandemic and in more general terms.

Future Research

To gain further insights into how people experience undergoing ICBT focused on psychological symptoms related to the COVID-19 pandemic and to explore the experiences of undergoing ICBT in general, it would be valuable to interview an even broader range of participants. This includes all participants, from those completing a single module to those who completed all modules and remained active throughout the treatment. Demonstrating the effectiveness of a therapy does not necessarily yield evidence regarding why the therapy works [63]. Qualitatively investigating what participants perceive as helpful or less helpful could be a way to gain more insight into their experiences, including reasons for ending treatment or dropping out prematurely. In addition to exploring the experiences of a broader sample of participants, it would be necessary to investigate participants' experiences throughout treatment. This would allow adjustments to be made during the treatment process, rather than insights emerging only after treatment ends. Furthermore, comparing experiences during and after treatment could help evaluate whether any adjustments could improve treatment outcomes and adherence.

It would be relevant to implement changes and further develop the material of this ICBT program based on the findings of this study. This would enable a comparison of treatment outcomes between the revised and original program versions. At the same time, several aspects of the program were described as suitable and helpful. Other ICBT programs, not necessarily focused on psychological symptoms during the COVID-19 pandemic, could also be developed based on these results because a considerable portion of what participants described in the interviews pertains to generic aspects of ICBT. Furthermore, it could be valuable to mix quantitative and qualitative methods to evaluate ICBT.

XSL•FO
RenderX

As ICBT relies on technology, including treatment platforms that feature exercises and communication with a therapist, it is essential to examine participants' experiences with the technology itself. This should include an exploration of how comfortable participants feel with the technology based on their technical competence and sense of security, as well as how the technology may influence their overall experience of the treatment.

Finally, based on the subtheme contact with the therapist, it would be of interest to investigate experiences of ICBT in which participants can choose their preferred mode of therapist contact. Gaining further understanding of participants' preferences for both treatment content and therapist support could inform further development of ICBT.

Overall, this study should be followed by other similar investigations into how ICBT programs are experienced. Each new treatment program and controlled trial could be complemented by a qualitative study because such studies provide additional information and can be useful when refining an intervention [64].

Conclusions

This study investigated participants' experiences of undergoing ICBT during the COVID-19 pandemic in the summer of 2020, focusing on psychological symptoms related to the pandemic. The results largely align with previous research in the field, but pandemic-specific findings were also identified. The treatment seemed to serve various functions, such as helping to widen participants' understanding of their own feelings; it required effort both during and after the treatment period; and its format, in terms of the level of individual tailoring and therapist contact, did not suit everyone's needs. Moreover, the treatment content seemed to be relevant not only for psychological symptoms related to the COVID-19 pandemic but also for other symptoms. In other words, participants described changes in their well-being as a result of the treatment in ways that extended beyond the pandemic context, highlighting the potential of ICBT to facilitate psychological improvements independent of situational factors.

Acknowledgments

The authors thank all participants for taking part in the interviews and giving their time and effort to describe their experiences of the internet-based cognitive behavioral therapy program. The authors would also like to thank their webmaster, George Vlaescu, for managing the treatment platform.

Data Availability

The data set generated or analyzed during this study are not publicly available due to the inclusion of sensitive information but additional examples of quotes for each theme are available from the corresponding author on reasonable request.

Authors' Contributions

M Baltius and EL conducted the telephone interviews. The transcriptions were carried out by M Baltius and EL, as well as VA. Data analysis was conducted by M Baltius, EL, and VA independently, with ML also contributing to the analysis. GA and M Berg monitored and supported the entire study, from developing the interview guide to processing the qualitative analysis. All authors have provided feedback on the analysis as well as the manuscript in its entirety. Furthermore, all authors have read, commented on, and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview guide.
[DOCX File , 18 KB - formative_v9i1e66908_app1.docx]

References

1. Morens DM, Fauci AS. Emerging infectious diseases: threats to human health and global stability. PLoS Pathog 2013;9(7):e1003467 [FREE Full text] [doi: 10.1371/journal.ppat.1003467] [Medline: 23853589]
2. Taylor S. The Psychology of Pandemics: Preparing for the Next Global Outbreak of Infectious Disease. Cambridge, MA: Cambridge Scholars Publishing; 2020.
3. Bloom DE, Cadarette D. Infectious disease threats in the twenty-first century: strengthening the global response. Front Immunol 2019;10:549 [FREE Full text] [doi: 10.3389/fimmu.2019.00549] [Medline: 30984169]
4. Tucci V, Moukaddam N, Meadows J, Shah S, Galwankar S, Kapur G. The forgotten plague: psychiatric manifestations of Ebola, Zika, and emerging infectious diseases. J Glob Infect Dis 2017;9(4):151-156 [FREE Full text] [doi: 10.4103/jgid.jgid_66_17] [Medline: 29302150]

XSL•FO
RenderX

5. Hossain MM, Tasnim S, Sultana A, Faizah F, Mazumder H, Zou L, et al. Epidemiology of mental health problems in COVID-19: a review. F1000Res 2020;9:636 [FREE Full text] [doi: 10.12688/f1000research.24457.1] [Medline: 33093946]
6. The employment situation. Bureau of Labor Statistics (BLS). 2020. URL: https://www.bls.gov/news.release/archives/empsit_05082020.pdf [accessed 2023-05-24]
7. Horesh D, Brown AD. Traumatic stress in the age of COVID-19: a call to close critical gaps and adapt to new realities. Psychol Trauma 2020 May;12(4):331-335. [doi: 10.1037/tra0000592] [Medline: 32271070]
8. Guo Q, Zheng Y, Shi J, Wang J, Li G, Li C, et al. Immediate psychological distress in quarantined patients with COVID-19 and its association with peripheral inflammation: a mixed-method study. Brain Behav Immun 2020 Aug;88:17-27 [FREE Full text] [doi: 10.1016/j.bbi.2020.05.038] [Medline: 32416290]
9. Xie Y, Xu E, Al-Aly Z. Risks of mental health outcomes in people with COVID-19: cohort study. BMJ 2022 Mar 16;376:e068993 [FREE Full text] [doi: 10.1136/bmj-2021-068993] [Medline: 35172971]
10. Brooks SK, Webster RK, Smith LE, Woodland L, Wessely S, Greenberg N, et al. The psychological impact of quarantine and how to reduce it: rapid review of the evidence. Lancet 2020 Mar 14;395(10227):912-920 [FREE Full text] [doi: 10.1016/S0140-6736(20)30460-8] [Medline: 32112714]
11. Tang W, Hu T, Hu B, Jin C, Wang G, Xie C, et al. Prevalence and correlates of PTSD and depressive symptoms one month after the outbreak of the COVID-19 epidemic in a sample of home-quarantined Chinese university students. J Affect Disord 2020 Sep 01;274:1-7 [FREE Full text] [doi: 10.1016/j.jad.2020.05.009] [Medline: 32405111]
12. Andersson G, Carlbring P, Titov N, Lindefors N. Internet interventions for adults with anxiety and mood disorders: a narrative umbrella review of recent meta-analyses. Can J Psychiatry 2019 Jul;64(7):465-470 [FREE Full text] [doi: 10.1177/0706743719839381] [Medline: 31096757]
13. Mehta S, Peynenburg VA, Hadjistavropoulos HD. Internet-delivered cognitive behaviour therapy for chronic health conditions: a systematic review and meta-analysis. J Behav Med 2019 Apr;42(2):169-187. [doi: 10.1007/s10865-018-9984-x] [Medline: 30387008]
14. Vlaescu G, Alasjö A, Miloff A, Carlbring P, Andersson G. Features and functionality of the Iterapi platform for internet-based psychological treatment. Internet Interv 2016 Nov;6:107-114 [FREE Full text] [doi: 10.1016/j.invent.2016.09.006] [Medline: 30135819]
15. Dear BF, Staples LG, Terides MD, Karin E, Zou J, Johnston L, et al. Transdiagnostic versus disorder-specific and clinician-guided versus self-guided internet-delivered treatment for generalized anxiety disorder and comorbid disorders: a randomized controlled trial. J Anxiety Disord 2015 Dec;36:63-77 [FREE Full text] [doi: 10.1016/j.janxdis.2015.09.003] [Medline: 26460536]
16. Baumeister H, Reichler L, Munzinger M, Lin J. The impact of guidance on internet-based mental health interventions — a systematic review. Internet Interventions 2014 Oct;1(4):205-215. [doi: 10.1016/j.invent.2014.08.003]
17. Hedman-Lagerlöf E, Carlbring P, Svärdman F, Riper H, Cuijpers P, Andersson G. Therapist-supported internet-based cognitive behaviour therapy yields similar effects as face-to-face therapy for psychiatric and somatic disorders: an updated systematic review and meta-analysis. World Psychiatry 2023 Jun;22(2):305-314 [FREE Full text] [doi: 10.1002/wps.21088] [Medline: 37159350]
18. Etzelmueller A, Vis C, Karyotaki E, Baumeister H, Titov N, Berking M, et al. Effects of internet-based cognitive behavioral therapy in routine care for adults in treatment for depression and anxiety: systematic review and meta-analysis. J Med Internet Res 2020 Aug 31;22(8):e18100 [FREE Full text] [doi: 10.2196/18100] [Medline: 32865497]
19. Mahoney A, Li I, Haskelberg H, Millard M, Newby JM. The uptake and effectiveness of online cognitive behaviour therapy for symptoms of anxiety and depression during COVID-19. J Affect Disord 2021 Sep 01;292:197-203 [FREE Full text] [doi: 10.1016/j.jad.2021.05.116] [Medline: 34126311]
20. Ho CS, Chee CY, Ho RC. Mental health strategies to combat the psychological impact of coronavirus disease 2019 (COVID-19) beyond paranoia and panic. Ann Acad Med Singap 2020 Mar 16;49(3):155-160 [FREE Full text] [Medline: 32200399]
21. Yao H, Chen JH, Xu YF. Rethinking online mental health services in China during the COVID-19 epidemic. Asian J Psychiatr 2020 Apr;50:102015 [FREE Full text] [doi: 10.1016/j.ajp.2020.102015] [Medline: 32247261]
22. Aminoff V, Sellén M, Sörliden E, Ludvigsson M, Berg M, Andersson G. Internet-based cognitive behavioral therapy for psychological distress associated with the COVID-19 pandemic: a pilot randomized controlled trial. Front Psychol 2021;12:684540 [FREE Full text] [doi: 10.3389/fpsyg.2021.684540] [Medline: 34194374]
23. Wahlund T, Mataix-Cols D, Olofsdotter Lauri K, de Schipper E, Ljótsson B, Aspvall K, et al. Brief online cognitive behavioural intervention for dysfunctional worry related to the COVID-19 pandemic: a randomised controlled trial. Psychother Psychosom 2021;90(3):191-199 [FREE Full text] [doi: 10.1159/000512843] [Medline: 33212440]
24. Aminoff V, Bobeck J, Hjort S, Sörliden E, Ludvigsson M, Berg M, et al. Tailored internet-based psychological treatment for psychological problems during the COVID-19 pandemic: a randomized controlled trial. Internet Interv 2023 Dec;34:100662 [FREE Full text] [doi: 10.1016/j.invent.2023.100662] [Medline: 37671335]
25. Komariah M, Amirah S, Faisal EG, Prayogo SA, Maulana S, Platini H, et al. Efficacy of internet-based cognitive behavioral therapy for depression and anxiety among global population during the COVID-19 pandemic: a systematic review and

XSL•FO
RenderX

meta-analysis of a randomized controlled trial study. Healthcare (Basel) 2022 Jun 30;10(7):1224 [FREE Full text] [doi: 10.3390/healthcare10071224] [Medline: 35885751]
26. Al-Alawi M, McCall RK, Sultan A, Al Balushi N, Al-Mahrouqi T, Al Ghailani A, et al. Efficacy of a six-week-long therapist-guided online therapy versus self-help internet-based therapy for COVID-19-induced anxiety and depression: open-label, pragmatic, randomized controlled trial. JMIR Ment Health 2021 Mar 12;8(2):e26683 [FREE Full text] [doi: 10.2196/26683] [Medline: 33512323]
27. Liu Z, Qiao D, Xu Y, Zhao W, Yang Y, Wen D, et al. The efficacy of computerized cognitive behavioral therapy for depressive and anxiety symptoms in patients with COVID-19: randomized controlled trial. J Med Internet Res 2021 May 14;23(5):e26883 [FREE Full text] [doi: 10.2196/26883] [Medline: 33900931]
28. Brog NA, Hegy JK, Berger T, Znoj H. Effects of an internet-based self-help intervention for psychological distress due to COVID-19: results of a randomized controlled trial. Internet Interv 2022 Mar;27:100492 [FREE Full text] [doi: 10.1016/j.invent.2021.100492] [Medline: 34956841]
29. Özdin S, Bayrak Özdin Ş. Levels and predictors of anxiety, depression and health anxiety during COVID-19 pandemic in Turkish society: the importance of gender. Int J Soc Psychiatry 2020 Aug;66(5):504-511 [FREE Full text] [doi: 10.1177/0020764020927051] [Medline: 32380879]
30. Wang Y, Di Y, Ye J, Wei W. Study on the public psychological states and its related factors during the outbreak of coronavirus disease 2019 (COVID-19) in some regions of China. Psychol Health Med 2021 Jan;26(1):13-22. [doi: 10.1080/13548506.2020.1746817] [Medline: 32223317]
31. Hao F, Tan W, Jiang L, Zhang L, Zhao X, Zou Y, et al. Do psychiatric patients experience more psychiatric symptoms during COVID-19 pandemic and lockdown? A case-control study with service and research implications for immunopsychiatry. Brain Behav Immun 2020 Jul;87:100-106 [FREE Full text] [doi: 10.1016/j.bbi.2020.04.069] [Medline: 32353518]
32. Holst A, Nejati S, Björkelund C, Eriksson MC, Hange D, Kivi M, et al. Patients' experiences of a computerised self-help program for treating depression - a qualitative study of internet mediated cognitive behavioural therapy in primary care. Scand J Prim Health Care 2017 Mar;35(1):46-53 [FREE Full text] [doi: 10.1080/02813432.2017.1288813] [Medline: 28277055]
33. Pedersen MK, Mohammadi R, Mathiasen K, Elmose M. Internet-based cognitive behavioral therapy for anxiety in an outpatient specialized care setting: a qualitative study of the patients' experience of the therapy. Scand J Psychol 2020 Dec 23;61(6):846-854. [doi: 10.1111/sjop.12665] [Medline: 32705703]
34. Rozental A, Forsström D, Tangen JA, Carlbring P. Experiences of undergoing internet-based cognitive behavior therapy for procrastination: a qualitative study. Internet Interv 2015 Sep;2(3):314-322. [doi: 10.1016/j.invent.2015.05.001]
35. Patel S, Akhtar A, Malins S, Wright N, Rowley E, Young E, et al. The acceptability and usability of digital health interventions for adults with depression, anxiety, and somatoform disorders: qualitative systematic review and meta-synthesis. J Med Internet Res 2020 Jul 06;22(7):e16228 [FREE Full text] [doi: 10.2196/16228] [Medline: 32628116]
36. Thapar S, Nguyen M, Khan BN, Fanaieyan R, Kishimoto V, Liu R, et al. Patient and therapist perceptions of a publicly funded internet-based cognitive behavioral therapy (iCBT) program for Ontario adults during the COVID-19 pandemic: qualitative study. JMIR Form Res 2024 Mar 19;8:e50113 [FREE Full text] [doi: 10.2196/50113] [Medline: 38373027]
37. Ncheka JM, Menon JA, Davies EB, Paul R, Mwaba SO, Mudenda J, et al. Implementing internet-based cognitive behavioural therapy (moodgym) for African students with symptoms of low mood during the COVID-19 pandemic: a qualitative feasibilty study. BMC Psychiatry 2024 Mar 01;24(1):92 [FREE Full text] [doi: 10.1186/s12888-024-05542-4] [Medline: 38302998]
38. Twomey C, O'Reilly G, Byrne M, Bury M, White A, Kissane S, et al. A randomized controlled trial of the computerized CBT programme, MoodGYM, for public mental health service users waiting for interventions. Br J Clin Psychol 2014 Nov;53(4):433-450. [doi: 10.1111/bjc.12055] [Medline: 24831119]
39. Ramzan N, Dixey R, Morris A. A qualitative exploration of adolescents' experiences of digital Dialectical Behaviour Therapy during the COVID-19 pandemic. Cogn Behav Ther 2022 Nov 18;15:e48. [doi: 10.1017/s1754470x22000460]
40. Faisal K, Masood A. Efficacy of therapist guided internet based cognitive behavioural therapy for depression: a qualitative exploration of therapists and clients experiences. J Prof Appl Psychol 2022 Dec 31;3(4):453-468. [doi: 10.52053/jpap.v3i4.130]
41. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
42. Busner J, Targum SD. The clinical global impressions scale: applying a research tool in clinical practice. Psychiatry (Edgmont) 2007 Jul;4(7):28-37. [Medline: 20526405]
43. Beukes EW, Manchaiah V, Davies AS, Allen PM, Baguley DM, Andersson G. Participants' experiences of an internet-based cognitive behavioural therapy intervention for tinnitus. Int J Audiol 2018 Dec;57(12):947-954. [doi: 10.1080/14992027.2018.1514538] [Medline: 30295113]
44. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]

XSL•FO
RenderX

45. Willig C. Introducing Qualitative Research in Psychology. 3rd edition. New York, NY: Open University Press; 2013.
46. Biliunaite I, Dumarkaite A, Kazlauskas E, Sanderman R, Andersson G. ICBT program for improving informal caregiver well-being: a qualitative study. Internet Interv 2021 Mar;23:100361 [FREE Full text] [doi: 10.1016/j.invent.2021.100361] [Medline: 33489781]
47. Braun V, Clarke V. Reflecting on reflexive thematic analysis. Qual Res Sport Exerc Health 2019 Jun 13;11(4):589-597. [doi: 10.1080/2159676X.2019.1628806]
48. Larsson S. A pluralist view of generalization in qualitative research. Int J Res Method Educ 2009 Apr;32(1):25-38. [doi: 10.1080/17437270902759931]
49. Berg M, Malmquist A, Rozental A, Topooco N, Andersson G. Knowledge gain and usage of knowledge learned during internet-based CBT treatment for adolescent depression - a qualitative study. BMC Psychiatry 2020 Sep 10;20(1):441 [FREE Full text] [doi: 10.1186/s12888-020-02833-4] [Medline: 32912154]
50. Person Asplund R, Jäderlind A, Björk IH, Ljótsson B, Carlbring P, Andersson G. Experiences of internet-delivered and work-focused cognitive behavioral therapy for stress: a qualitative study. Internet Interv 2019 Dec;18:100282 [FREE Full text] [doi: 10.1016/j.invent.2019.100282] [Medline: 31737490]
51. Stalmeijer RE, Brown ME, O'Brien BC. How to discuss transferability of qualitative research in health professions education. Clin Teach 2024 Dec 18;21(6):e13762. [doi: 10.1111/tct.13762] [Medline: 38497107]
52. Bendelin N, Hesser H, Dahl J, Carlbring P, Nelson KZ, Andersson G. Experiences of guided internet-based cognitive-behavioural treatment for depression: a qualitative study. BMC Psychiatry 2011 Jun 30;11:107 [FREE Full text] [doi: 10.1186/1471-244X-11-107] [Medline: 21718523]
53. Lindegaard T, Kashoush F, Holm S, Halaj A, Berg M, Andersson G. Experiences of internet-based cognitive behavioural therapy for depression and anxiety among Arabic-speaking individuals in Sweden: a qualitative study. BMC Psychiatry 2021 Jun 03;21(1):288 [FREE Full text] [doi: 10.1186/s12888-021-03297-w] [Medline: 34082745]
54. Donker T, Griffiths KM, Cuijpers P, Christensen H. Psychoeducation for depression, anxiety and psychological distress: a meta-analysis. BMC Med 2009 Dec 16;7:79 [FREE Full text] [doi: 10.1186/1741-7015-7-79] [Medline: 20015347]
55. Torre JB, Lieberman MD. Putting feelings into words: affect labeling as implicit emotion regulation. Emotion Review 2018 Mar 20;10(2):116-124. [doi: 10.1177/1754073917742706]
56. Kessler RC, Sampson NA, Berglund P, Gruber MJ, Al-Hamzawi A, Andrade L, et al. Anxious and non-anxious major depressive disorder in the World Health Organization World Mental Health surveys. Epidemiol Psychiatr Sci 2015 Jun;24(3):210-226 [FREE Full text] [doi: 10.1017/S2045796015000189] [Medline: 25720357]
57. Dalgleish T, Black M, Johnston D, Bevan A. Transdiagnostic approaches to mental health problems: current status and future directions. J Consult Clin Psychol 2020 Mar;88(3):179-195 [FREE Full text] [doi: 10.1037/ccp0000482] [Medline: 32068421]
58. Lynch SJ, Sunderland M, Newton NC, Chapman C. A systematic review of transdiagnostic risk and protective factors for general and specific psychopathology in young people. Clin Psychol Rev 2021 Jul;87:102036 [FREE Full text] [doi: 10.1016/j.cpr.2021.102036] [Medline: 33992846]
59. Olatunji BO, Wolitzky-Taylor KB. Anxiety sensitivity and the anxiety disorders: a meta-analytic review and synthesis. Psychol Bull 2009 Nov;135(6):974-999. [doi: 10.1037/a0017428] [Medline: 19883144]
60. Johnsen HM, Haddeland K. User evaluation of a therapist-guided internet-delivered treatment program for anxiety disorders: a qualitative study. Internet Interv 2021 Sep;25:100389 [FREE Full text] [doi: 10.1016/j.invent.2021.100389] [Medline: 33996506]
61. Earley C, Joyce C, McElvaney J, Richards D, Timulak L. Preventing depression: qualitatively examining the benefits of depression-focused iCBT for participants who do not meet clinical thresholds. Internet Interv 2017 Sep;9:82-87 [FREE Full text] [doi: 10.1016/j.invent.2017.07.003] [Medline: 30135841]
62. Fernández-Álvarez J, Díaz-García A, González-Robles A, Baños R, García-Palacios A, Botella C. Dropping out of a transdiagnostic online intervention: a qualitative analysis of client's experiences. Internet Interv 2017 Dec;10:29-38. [doi: 10.1016/j.invent.2017.09.001] [Medline: 30135750]
63. Kazdin AE. Mediators and mechanisms of change in psychotherapy research. Annu Rev Clin Psychol 2007;3:1-27. [doi: 10.1146/annurev.clinpsy.3.022806.091432] [Medline: 17716046]
64. Skivington K, Matthews L, Simpson SA, Craig P, Baird J, Blazeby JM, et al. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. BMJ 2021 Sep 30;374:n2061 [FREE Full text] [doi: 10.1136/bmj.n2061] [Medline: 34593508]

Abbreviations

CBT: cognitive behavioral therapy
COREQ: Consolidated Criteria for Reporting Qualitative Research
ICBT: internet-based cognitive behavioral therapy
RCT: randomized controlled trial
TA: thematic analysis

XSL•FO
RenderX

Edited by A Mavragani; submitted 26.09.24; peer-reviewed by A Eisingerich, C Webb; comments to author 03.03.25; revised version received 14.03.25; accepted 21.05.25; published 11.06.25.

Please cite as:
Aminoff V, Baltius M, Lundström E, Berg M, Andersson G, Ludvigsson M
Experiences of Individually Tailored Internet-Based Cognitive Behavioral Therapy During the COVID-19 Pandemic: Qualitative Interview Study
JMIR Form Res 2025;9:e66908
URL: https://formative.jmir.org/2025/1/e66908
doi:10.2196/66908
PMID:

©Victoria Aminoff, Matilda Baltius, Emelie Lundström, Matilda Berg, Gerhard Andersson, Mikael Ludvigsson. Originally published in JMIR Formative Research (https://formative.jmir.org), 11.06.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Development of a Serious Game to Simulate Neonatal Intensive Care Unit Experiences: Collaborative Quasi-Experimental Study

Yukihide Miyosawa[1], MD; Koichi Hirabayashi[1], MD; Kodai Yamada[2]; Fumiya Kobayashi[2]; Nanami Ogihara[2]; Noa Takeda[3]; Eri Okamura[2]; Shogo Matsumura[4]

[1]Department of Pediatrics, Shinshu University School of Medicine, Matsumoto, Japan

[2]Department of Medical Education, Shinshu University School of Medicine, Matsumoto, Japan

[3]Department of Illustration, Tokyo Designer Gakuin College, Tokyo, Japan

[4]Department of Medical Education, Keio University School of Medicine, Keio University, Tokyo, Japan

Corresponding Author:
Yukihide Miyosawa, MD
Department of Pediatrics
Shinshu University School of Medicine
3-1-1, Asahi
Matsumoto, 390-8621
Japan
Phone: 81 0263372642
Fax: 81 0263373089
Email: miyosawa@shinshu-u.ac.jp

Abstract

Background: Opportunities for neonatal intensive care unit (NICU) training are limited for medical and nursing students due to patient safety concerns and the complexities of neonatal care. In addition, the COVID-19 pandemic significantly disrupted clinical training opportunities, further underscoring the need for alternative educational tools that can provide immersive and practical learning experiences. Serious games have garnered attention as potential tools for medical education; however, few are designed to simulate the complete NICU environment and its unique challenges.

Objective: To address the educational gaps in neonatal care training, we aimed to develop and evaluate a serious game that provides a comprehensive NICU simulation experience for students and the general public.

Methods: The game was developed over 14 months by a collaborative team that included a neonatologist, 4 medical students, and 1 art student, with a total cost of US $10,000. Initially created in TyranoBuilder (STRIKEWORKS), the game was later redeveloped in Unity with Naninovel to support multilingual functionality. Structured as a 6-chapter visual novel, the game follows a high school student observing the NICU during a hospital internship. Scenario-based decision-making and interactive dialogues guide the player through both the clinical and emotional aspects of neonatal care. After completing the game, players were invited to participate in an optional web-based survey that assessed demographic information, gameplay quality, and educational value using Likert scales. Descriptive and inferential statistics were used for data analysis.

Results: The game, titled First Steps in the NICU, was released for iOS, Android, and Steam. As of May 2025, it has been downloaded 2799 times (2260 on iOS and 539 on Android). A total of 160 survey responses were collected, with 46.3% of respondents identifying as health care professionals or students. The majority of participants were female (114/160, 71.3%) and aged 20-29 years (59/160, 36.9%). Mean scores for length, difficulty, and gameplay were 3.05 (SD 0.62), 2.49 (SD 0.76), and 3.65 (SD 0.77), respectively, indicating a well-balanced design. The educational usefulness of the game received high ratings: empathy with the story (4.24), usefulness for knowledge acquisition (4.16), and effectiveness of serious games as a learning tool (4.37). No significant differences in evaluations were found between health care professionals and students and the general public, suggesting broad accessibility and appeal.

Conclusions: We developed a low-cost serious game that simulates NICU experiences through collaboration between a neonatologist and students. The game received positive feedback and demonstrated educational value for a diverse audience. Positioned as formative research, this study highlights the potential of serious games to supplement neonatal care education. Future updates will incorporate user feedback, leading to improvements in gameplay and expanded content.

XSL•FO
RenderX

(JMIR Form Res 2025;9:e73009) doi:10.2196/73009

KEYWORDS

collaborative development; medical education; neonatal care; neonatal intensive care unit; physicians and students; serious game; simulation

Introduction

Serious games are designed to address social issues and are used across various fields, including education, government, and health care [1,2]. These games offer several benefits, such as increasing engagement, providing a safe environment to practice skills, and enabling personalized learning experiences [3,4].

In health care, serious games have been widely adopted as educational tools [5], with numerous reports highlighting their effectiveness in aiding the acquisition of diagnostic skills and surgical techniques [6]. However, their application in simulating immersive medical experiences across broader health care fields remains limited [7].

In neonatal care, serious games have primarily been implemented in neonatal resuscitation training, with many studies demonstrating their effectiveness in this area [8,9]. However, there are few reports on serious games that allow students to experience the full scope of neonatal intensive care unit (NICU) settings and the complexities of neonatal care. Premature and critically ill neonates in the NICU are at high risk of sudden deterioration, making it challenging for students to conduct examinations or perform medical procedures. For medical and nursing students, these high-stakes situations create substantial psychological barriers to engaging with neonatal care [10]. Furthermore, the COVID-19 pandemic significantly restricted clinical training opportunities, highlighting the potential for future pandemics to impose similar limitations, resulting in the loss of valuable learning experiences [11].

Based on the above considerations, we hypothesized that a serious game simulating a comprehensive NICU experience could serve as an effective tool to enhance medical and nursing students' understanding of neonatal care. In 2021, we developed a serious game to simulate clinical training in the NICU. This project was undertaken solely by a team of a neonatologist, medical students, and a design student without involvement from professional game developers. This study outlines the development process of the serious game and evaluates its educational effectiveness based on player feedback. Therefore, the objective of this study was to investigate whether a collaboratively developed serious game could enhance users' understanding of neonatal intensive care, particularly among medical and nursing students. We hypothesized that this game would provide educational value and promote engagement among a diverse audience, including health care professionals and the general public.

Methods

Preparation and Design

The game development spanned 14 months, from July 2021 to August 2022. The Serious Game Development Committee consisted of 6 members: 1 neonatologist who served as the project leader, 4 medical students, and 1 art student specializing in illustration. The game's storyline was created by the neonatologist to accurately represent the NICU experience. Nurses, midwives, and obstetricians from Shinshu University Hospital contributed to the medical accuracy of the content through interviews. After finalizing the storyline, the team programmed the game. Programming was conducted by 4 medical students in collaboration with the neonatologist, while 1 student processed background images by converting photographs into illustrations to create a cohesive visual style. Character design was performed by the art student specializing in illustration. All students involved in this project were used as part-time assistants and met with the neonatologist in the laboratory after classes to work on the development of the game. After the prototype was completed, perinatal health care professionals tested the game. Their feedback on the accuracy of medical information and overall functionality was incorporated into the final version to ensure that it met educational and practical standards.

Story

The game is structured as a visual novel, allowing players to follow a narrative-driven experience with interactive choices. Players assume the role of a high school student observing the NICU as part of a 1-day hospital internship with the goal of compiling a report based on their experience.

The game is divided into 6 chapters and offers a comprehensive NICU experience, gradually introducing players to the NICU environment and its unique challenges (Figure 1).

Throughout the game, players interact with doctors, nurses, and families, and their choices influence the storyline, leading to multiple possible outcomes. In the exploration phase, players navigate the NICU freely, examining various aspects of neonatal care. This includes interacting with medical equipment such as ventilators and ultrasound machines, learning emergency response protocols, and understanding family support systems. The estimated total playtime for the game is approximately 1 hour.

XSL•FO
RenderX

Figure 1. Overview of the neonatal intensive care unit simulation game stages. (A) High school scene: the protagonist, a high school student, expresses their interest in the neonatal intensive care unit to classmates and prepares a previsit report, setting the foundation for their upcoming hospital experience. (B) Neonatal intensive care unit exploration: the player freely explores the neonatal intensive care unit environment, learning about medical equipment such as ventilators and ultrasound machines, neonatal care practices, and the emotional perspectives of health care providers and families involved in neonatal care. (C) Lunch with staff: during a meal break with doctors and nurses, the player can engage in open conversations, gaining insights into the motivations and experiences of neonatal intensive care unit health care professionals. (D) Emergency scenario: the player participates in a simulated emergency response, providing an opportunity to learn about critical actions during neonatal resuscitation and acute care. (E) General care unit and discharge support: the player transitions to the general care unit to learn about discharge planning and the support necessary for families preparing to take their newborns home. (F) Report compilation: after completing the neonatal intensive care unit experience, the player consolidates their knowledge and reflections into a report, reinforcing the learning objectives from each stage of the simulation.

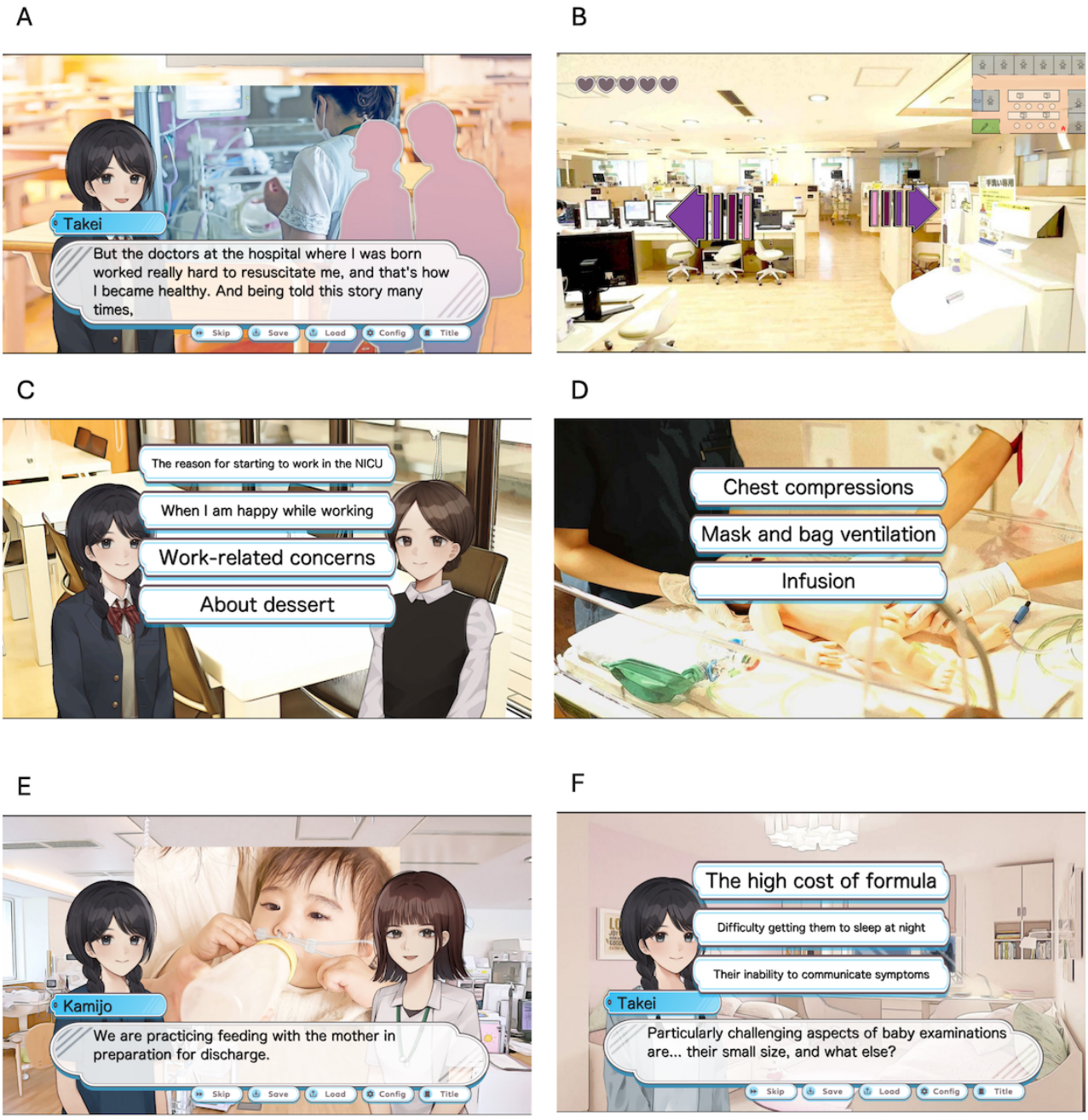

Technology

The game was initially developed using TyranoBuilder (STRIKEWORKS), a game engine specialized for visual novels. TyranoBuilder's visual scripting interface facilitates learning and development even for novice developers. This tool enables the integration of branching storylines, achievement tracking, and multiple endings. The game was first released as an iOS and Android application in September 2022. However, as of 2023, TyranoBuilder did not support multilingual functionality (a feature introduced in 2024). To provide multilingual versions in Japanese and English, the game was redeveloped in September 2023 using Naninovel, a visual novel engine for Unity (Naninovel, Elringus). Unity is a widely used game

XSL•FO
RenderX

development platform and Naninovel simplifies the creation of interactive narratives within Unity. The multilingual version was successfully released in November 2024. The game, titled "First Steps in the NICU" in English, was made freely available for download on the Apple App and Google Play stores without restrictions. In addition to releasing the game on the App Store and Google Play, the game was also made available on Steam in February 2025, expanding its accessibility to a wider audience on PC platforms. This multiplatform approach was intended to reach a diverse demographic and enhance user engagement.

Game Evaluation

After completing the game, players were invited to participate in an optional survey conducted via Google Forms. Textbox 1 outlines the survey questions.

The survey gathered demographic information such as occupation, age, gender, device used for the game, and playtime. To assess the game's quality and educational value, the survey used a series of 5-point Likert scales listed in Textbox 2.

The scenes that left the strongest impressions and requests regarding topics for future games were also collected. Finally, participants were given the opportunity to provide additional insights through open-ended comments. We also retrieved data on user engagement from App Store Connect and Google Play Console. These include download trends over time and device usage distributions.

XSL•FO
RenderX

Textbox 1. The list of questions in the survey.

Questions and their choices

1. What is your occupation?

- Pediatrician
- Physician other than pediatrician
- Nonphysician health care professional
- Medical student
- Nursing student
- Non–health care professional
- Non–health care university student
- Elementary, junior high, and high school student
- Other

2. What is your age? (years)

- 10-19
- 20-29
- 30-39
- 40-49
- 50-59
- 60+

3. What is your gender?

- Men
- Women
- Other

4. What device did you use?

- Smartphone
- Tablet
- Personal computer
- Other

5. How long did it take you to complete the game? (hours)

- <1
- 1-2
- 2-4
- >4

6. How did you find the game length?

- 1 (too short), 2, 3, 4, 5 (too long)

7. How did you find the game difficulty?

- 1 (too easy), 2, 3, 4, 5 (too hard)

8. How much gameplay should the game include?

- 1 (less gameplay), 2, 3, 4, 5 (more gameplay)

9. Could you empathize with the story?

- 1 (slightly empathetic), 2, 3, 4, 5 (very empathetic)

XSL•FO
RenderX

10. Was the game useful for knowledge acquisition?

- 1 (slightly useful), 2, 3, 4, 5 (very useful)

11. Do you think serious games are effective as a learning tool?

- 1 (slightly effective), 1, 2, 3, 4, 5 (very effective)

12. What scenes left the strongest impression on you? (Select all that apply.)

- From high school to neonatal intensive care unit (NICU) admission
- NICU exploration
- Learning how to care for the baby
- Chatting over lunch
- Baby's hospitalization
- Discharge support at the general care unit
- Report writing to the ending

13. What topics would you like future serious games to cover? (select all that apply)

- Discharge support from the NICU
- Infant care institutions
- Disaster preparedness for children requiring medical care
- Home support for children requiring medical care
- Child development support facilities
- School attendance for children requiring medical care
- Employment support for children requiring medical care
- Using administrative services for children requiring medical care
- Other

14. Please feel free to share any additional comments or feedback.

Textbox 2. Likert scales for game quality and educational usefulness matrices.

Game quality metrics

- Length: 1 (too short) to 5 (too long).
- Difficulty: 1 (too easy) to 5 (too hard).
- Gameplay: 1 (needs less gameplay) to 5 (needs more gameplay).

Educational usefulness metrics

- Empathy with the story: 1 (slightly empathetic) to 5 (very empathetic).
- Usefulness for knowledge acquisition: 1 (slightly useful) to 5 (very useful).
- Effectiveness of serious games for learning: 1 (slightly effective) to 5 (very effective).

Statistical Analyses

Participants' characteristics were summarized using descriptive statistics. A comparison of the results between the 2 groups was performed using unpaired *t* tests for continuous variables and Fisher exact tests for categorical variables. Graph generation and statistical analyses were performed using GraphPad Prism version 9 (GraphPad Software) or EZR (Saitama Medical Center, Jichi Medical University) software [12]. Statistical significance was set at $P<.05$ (2-sided).

Ethical Considerations

This study was approved by the Ethics Committee of Shinshu University (approval number 5794) on March 14, 2023. The research protocol was titled "Educational Effects of Experiencing Neonatal Medical Care Through a Serious Game." Participation in the postgame survey was voluntary, and informed consent was obtained through a statement provided at the beginning of the questionnaire. The survey was anonymous, and no personally identifiable information was collected. Participants did not receive any compensation for their participation.

XSL•FO
RenderX

Results

Overview of Downloads by Device Type, Operating System, and Country

The game has been downloaded 2260 times on iOS and 539 times on Android as of May 2025. In terms of device type, the majority of downloads were from smartphones (2540/2796, 90.8%), followed by tablets (230/2796, 8.2%) and other devices (29/2796, 1%; Table S1 in Multimedia Appendix 1). Geographically, Japan accounted for the overwhelming majority of downloads (2703/2796, 96.6%), reflecting the fact that the Japanese version was the primary distribution channel at the time of data collection. Notably, the application was also accessed by users in North America, South and Southeast Asia, and the Middle East, demonstrating its potential international reach (Table S2 in Multimedia Appendix 1).

Overall Respondent Characteristics and Game Evaluations

Responses were collected from 160 participants. Table 1 presents the respondents' characteristics. Health care professionals and students comprised 74 out of 160 (46.3%) participants, while the general public accounted for 86 out of 160 (53.8%) participants. The highest proportion of respondents was in the 20- to 29-year age range, accounting for 59 out of 160 (36.9%) participants, with 114 out of 160 (71.3%) participants identifying as female.

Table 1. Characteristics of survey respondents (N=160).

Characteristics	Values, n (%)
Occupation	
Pediatricians	10 (6.3)
Physicians other than pediatricians	3 (1.9)
Nonphysician health care professionals	31 (19.4)
Medical students	18 (11.3)
Nursing students	12 (7.5)
Non–health care professionals	44 (27.5)
Non–health care university students	6 (3.7)
Elementary, junior high, and high school students	24 (15)
Other	12 (7.5)
Age group (years)	
10-19	25 (15.6)
20-29	59 (36.9)
30-39	38 (23.7)
40-49	26 (16.2)
50-59	10 (6.3)
60+	2 (1.2)
Gender	
Men	46 (28.7)
Women	114 (71.3)
Other	0 (0)
Playtime (hours)	
<1	99 (61.9)
1-2	53 (33.1)
2-4	3 (1.9)
>4	5 (3.1)

The mean scores for length, difficulty, gameplay, story empathy, usefulness for knowledge acquisition, and effectiveness of serious games for learning were 3.05 (SD 0.62), 2.49 (SD 0.76), 3.65 (SD 0.77), 4.24 (SD 0.84), 4.16 (SD 0.96), and 4.37 (SD 0.78), respectively (Figure 2).

XSL•FO
RenderX

Figure 2. Game evaluations in overall respondents The respondents rated six aspects of the game: length, difficulty, gameplay, empathy with the story, usefulness for knowledge acquisition, and effectiveness of serious games for learning. The black spots and error bars indicate the mean and standard error, respectively.

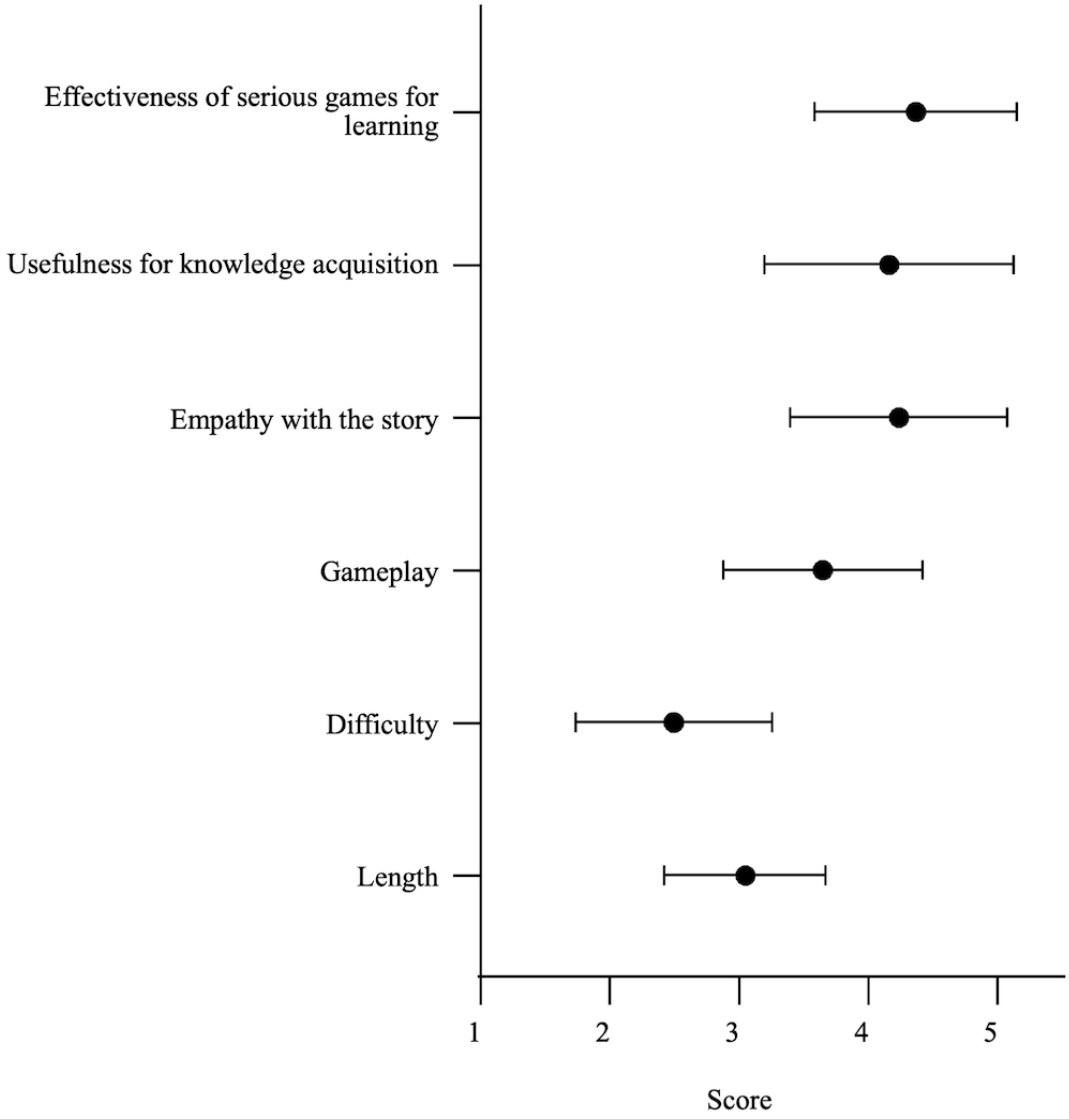

Comparison Between the General Public and Health Care Professionals and Students

To assess the differences in ratings for the serious game between the general public and health care professionals and students, we compared the questionnaire responses. While there were no differences in age or gender, playtime was significantly shorter among health care professionals and students than among the general public (Table 2).

There were no differences in the ratings for the 6 questions between the 2 groups (Figure 3).

XSL•FO
RenderX

Table 2. Comparison of the characteristics of health care professionals and students and the general public.

Characteristics	Health care professionals and students (n=74), n	General public (n=86), n	*P* value
Age group (years)			.75
10-19	1	24	
20-29	41	18	
30-39	15	23	
40-49	15	11	
50-59	2	8	
60+	0	2	
Gender			.49
Male	19	27	
Female	55	59	
Playtime			.03
Less than 1 h	53	46	
1-2 h	19	34	
2-4 h	0	3	
More than 4 h	2	3	

XSL•FO
RenderX

Figure 3. Comparison of game evaluations between health care professionals and students and the general public. The mean ratings of the game evaluations by the general public and health care professionals and students are shown as open squares and circles, respectively. The error bars indicate the SE.

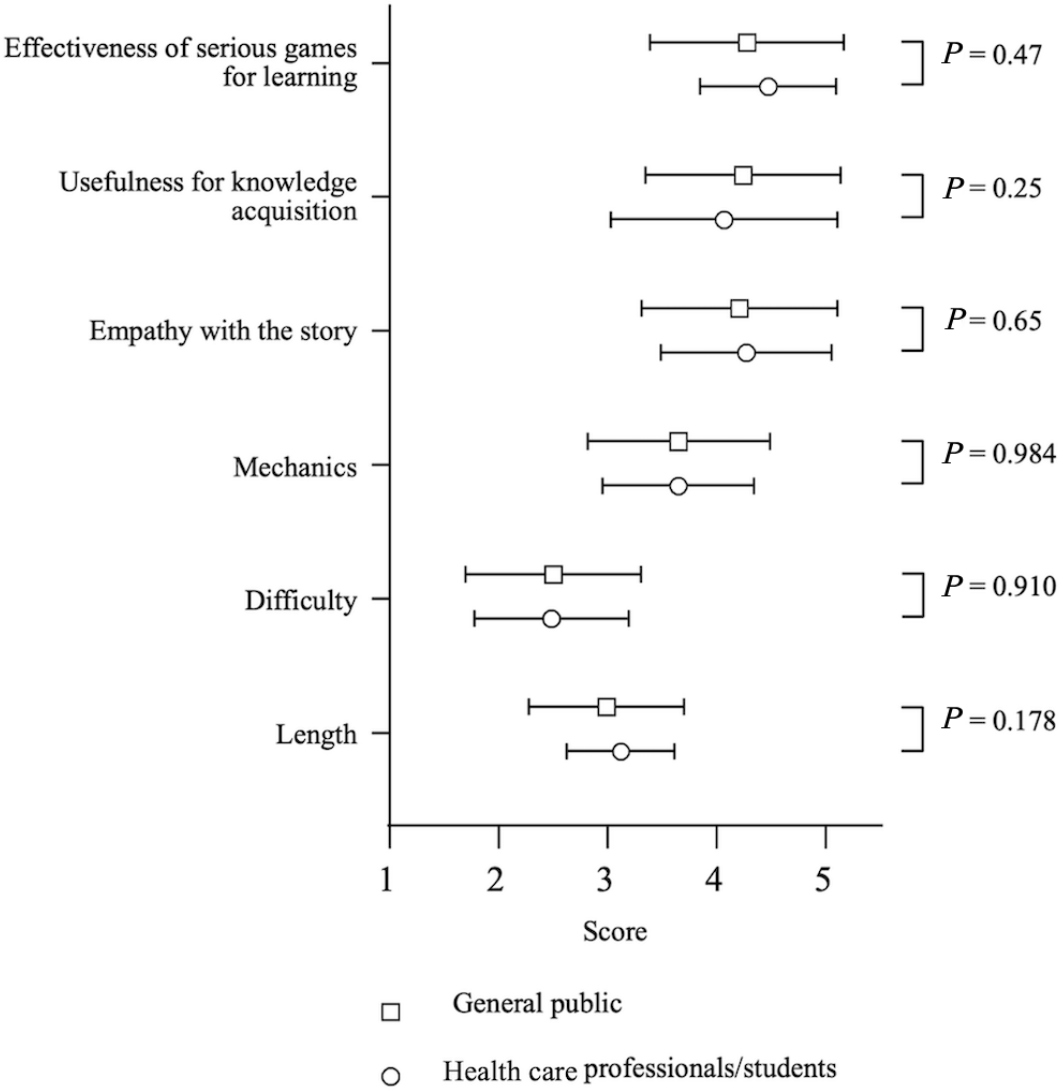

Discussion

Main Findings

This study demonstrated that a serious game simulating the NICU experience, developed collaboratively by a neonatologist and university students, was positively evaluated by both health care professionals and the general public. The game was found to offer high educational value and accessibility across diverse user groups.

Development Process and Cost Efficiency

A serious game designed to simulate clinical training in the NICU for medical and nursing students was created by a team comprising a neonatologist, medical students, and a design student. Unlike existing tools that focus on specific skills, such as neonatal resuscitation [8,9], this game offers a broader educational experience covering various aspects of neonatal medicine, making it a unique resource in medical education.

A major concern regarding the development of serious games among health care professionals is the lack of technical expertise in programming and limited funding [13,14]. In this study, the game was developed without the involvement of professional game developers. Despite a limited budget of approximately 1.5 million yen (US $10,000) for this project, a serious game was successfully created using TyranoBuilder, a low-cost visual novel development tool with an easy learning curve. This tool enabled medical students with no prior programming experience to contribute directly to the creation of the game. In addition to reducing costs, this approach likely helped students gain a deeper understanding of neonatal care. The art student's participation in a socially impactful project may have also enhanced their confidence, emphasizing the interdisciplinary value of educational game development. Our serious game development process demonstrates that health care professionals without technical programming expertise can create serious games, suggesting that these games can be developed at low costs.

XSL•FO
RenderX

Game Design and Educational Value

Balancing game design is crucial for maximizing both game engagement and educational value. In this study, we assessed the balance of the serious game using 3 parameters: length, difficulty, and gameplay. On the 5-point Likert scale, a score of 3 indicates an ideal rating. The mean scores for length, difficulty, and gameplay were 3.05, 2.49, and 3.65, respectively, indicating a balanced design. However, to improve the experience, increasing both the difficulty and gameplay could be ideal for future versions.

The main purpose of the serious game was to educate medical and nursing students. We assessed the educational usefulness of our serious game using 3 parameters: empathy with the story, usefulness for knowledge acquisition, and effectiveness of serious games for learning. The mean scores for all 3 parameters exceeded 4 on a 5-point Likert scale (Figure 3), suggesting that our serious game offers substantial educational value. Although previous studies have criticized serious games for offering lower levels of enjoyment compared to commercial video games [15], our findings suggest that this game may overcome this problem. These findings suggest that serious games have the potential to serve as educational tools and models for future game development in medical education. These findings can also be understood through educational theory. Kolb experiential learning model emphasizes learning through active experience and reflection, which are incorporated into this game through scenario-based choices and integrated quizzes [16]. In particular, the 4 stages of the Kolb model (concrete experience, reflective observation, abstract conceptualization, and active experimentation) are represented, respectively, through the immersive NICU simulation, narrative reflection points, in-game medical content, and decision-making paths. This structure helps players internalize neonatal care concepts, potentially leading to real-world applications. Similarly, the game aligns with Bloom taxonomy by promoting not only memorization but also higher-order thinking skills such as application and analysis, suggesting it may be more effective than traditional knowledge-based education [17]. By integrating these learning theories into game design, the educational impact of serious games can be theoretically grounded and systematically enhanced, supporting their use in structured medical curricula.

Comparison Between Health Care Professionals and the Public

We also compared the game's ratings between health care professionals and students and the general public. Because the NICU is an unfamiliar and highly specialized environment, we anticipated lower ratings from the general public regarding balance and educational value. Contrary to our expectations, no significant differences were observed between the 2 groups. This indicates that our serious game is suitable for both health care professionals and students and the general public, providing NICU learning and immersive experiences.

Although the primary target of this serious game was medical and nursing students, a wide variety of people played the game and participated in the survey. Teenagers accounted for 25 out of 160 (15.6%) respondents. This is particularly important, as early exposure to medical concepts may inspire future interest in health care careers. The game's accessible design makes it appealing to the general public, suggesting that increased outreach to younger demographics could be beneficial.

Formative Nature and Future Development

This study can be positioned as formative research, as it focuses on the early-stage development and evaluation of an innovative educational tool. The postgame survey revealed that many participants provided positive feedback regarding the educational content and emotional impact of the game. However, some respondents expressed a desire for enhanced gameplay features and increased entertainment value. Based on these insights, future iterations of the game will explore ways to integrate more engaging mechanics while preserving educational depth. The authors are also involved in the ongoing development of serious games related to perinatal care and pediatric home medical support. In future projects, they aim to create tools that are both informative and enjoyable, appealing to a broader audience beyond health care professionals.

Limitations and Future Directions

This study has several limitations. First, participation in the survey was voluntary and limited to players of the Japanese version of the game, potentially limiting the generalizability of the findings. In addition, the absence of a defined sampling frame and randomization process limits our ability to assess how representative the respondent population is. Second, the lack of an advertising budget may have reduced the diversity of respondents, potentially overrepresenting individuals connected to the Serious Game Development Committee. Moreover, the absence of a pretest and posttest design limits our ability to assess learning gains objectively. Consequently, the findings should be interpreted as preliminary and exploratory. The voluntary nature of the survey may also have introduced self-selection bias, potentially influencing the generalizability of the findings. Furthermore, the survey instrument was not formally validated or psychometrically tested and its reliability remains unverified. This lack of validation may affect the accuracy and consistency of the responses. Future studies should evaluate the multilingual version of the game and explore its impact across broader demographics and cultural contexts.

Conclusions

We developed a serious game designed to enhance neonatal care education at a low cost, relying solely on the collaboration between a neonatologist and students. This serious game offers a comprehensive educational experience accessible to both health care professionals and students and the general public.

XSL•FO
RenderX

Acknowledgments

We sincerely thank Yuka Takinami, Sachiha Kamijo, Mayuko Yagi, and Haruna Ariizumi for their invaluable contributions to this study. Their cooperation during the interviews and their participation in the game design process greatly enhanced the accuracy and quality of the content. We also wish to thank Editage for the English language editing. The development was supported by the Yuumi Memorial Foundation for Home Health Care. No generative artificial intelligence tools were used in the preparation of this manuscript.

Authors' Contributions

YM contributed to conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, supervision, visualization, writing—original draft, writing—review & editing. KY, FK, and NO assisted with software. NT, EO, and SM contributed to resources. KH contributed to formal analysis, investigation, methodology, visualization, writing—original draft, and writing—review & editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Overview of download by device type, operating system, and country.
[DOCX File , 37 KB - formative_v9i1e73009_app1.docx]

References

1. Graafland M, Schraagen JM, Schijven MP. Systematic review of serious games for medical education and surgical skills training. Br J Surg 2012;99(10):1322-1330. [doi: 10.1002/bjs.8819] [Medline: 22961509]
2. Boeker M, Andel P, Vach W, Frankenschmidt A. Game-based e-learning is more effective than a conventional instructional method: a randomized controlled trial with third-year medical students. PLoS One 2013;8(12):e82328 [FREE Full text] [doi: 10.1371/journal.pone.0082328] [Medline: 24349257]
3. Gorbanev I, Agudelo-Londoño S, González RA, Cortes A, Pomares A, Delgadillo V, et al. A systematic review of serious games in medical education: quality of evidence and pedagogical strategy. Med Educ Online 2018;23(1):1438718 [FREE Full text] [doi: 10.1080/10872981.2018.1438718] [Medline: 29457760]
4. Gentry SV, Gauthier A, L'Estrade Ehrstrom B, Wortley D, Lilienthal A, Tudor Car L, et al. Serious gaming and gamification education in health professions: systematic review. J Med Internet Res 2019;21(3):e12994 [FREE Full text] [doi: 10.2196/12994] [Medline: 30920375]
5. Maheu-Cadotte M, Cossette S, Dubé V, Fontaine G, Lavallée A, Lavoie P, et al. Efficacy of serious games in healthcare professions education: a systematic review and meta-analysis. Simul Healthc 2021;16(3):199-212. [doi: 10.1097/SIH.0000000000000512] [Medline: 33196609]
6. Dankbaar MEW, Alsma J, Jansen EEH, van Merrienboer JJG, van Saase JLCM, Schuit SCE. An experimental study on the effects of a simulation game on students' clinical cognitive skills and motivation. Adv Health Sci Educ Theory Pract 2016;21(3):505-521 [FREE Full text] [doi: 10.1007/s10459-015-9641-x] [Medline: 26433730]
7. Sardi L, Idri A, Fernández-Alemán JL. A systematic review of gamification in e-Health. J Biomed Inform 2017;71:31-48 [FREE Full text] [doi: 10.1016/j.jbi.2017.05.011] [Medline: 28536062]
8. Sawyer T, Umoren RA, Gray MM. Neonatal resuscitation: advances in training and practice. Adv Med Educ Pract 2017;8:11-19 [FREE Full text] [doi: 10.2147/AMEP.S109099] [Medline: 28096704]
9. Sarvan S, Efe E. The effect of neonatal resuscitation training based on a serious game simulation method on nursing students' knowledge, skills, satisfaction and self-confidence levels: a randomized controlled trial. Nurse Educ Today 2022;111:105298. [doi: 10.1016/j.nedt.2022.105298] [Medline: 35158135]
10. Waggoner JK, Waskosky A. Benefits of an integrated nurse residency program in the neonatal intensive care unit: a review of the literature. J Perinat Neonatal Nurs 2023;37(2):148-152. [doi: 10.1097/JPN.0000000000000681] [Medline: 37102562]
11. Rose S. Medical student education in the time of COVID-19. JAMA 2020;323(21):2131-2132. [doi: 10.1001/jama.2020.5227] [Medline: 32232420]
12. Kanda Y. Investigation of the freely available easy-to-use software 'EZR' for medical statistics. Bone Marrow Transplant 2013;48(3):452-458 [FREE Full text] [doi: 10.1038/bmt.2012.244] [Medline: 23208313]
13. McCallum S. Gamification and serious games for personalized health. Stud Health Technol Inform 2012;177:85-96. [Medline: 22942036]
14. Akl EA, Kairouz VF, Sackett KM, Erdley WS, Mustafa RA, Fiander M, et al. Educational games for health professionals. Cochrane Database Syst Rev 2013;2013(3):CD006411 [FREE Full text] [doi: 10.1002/14651858.CD006411.pub4] [Medline: 23543543]

XSL•FO
RenderX

15. Connolly TM, Boyle EA, MacArthur E, Hainey T, Boyle JM. A systematic literature review of empirical evidence on computer games and serious games. Comput Educ 2012;59(2):661-686. [doi: 10.1016/j.compedu.2012.03.004]
16. Kayes D. Experiential learning and its critics: preserving the role of experience in management learning and education. AMLE 2002;1(2):137-149. [doi: 10.5465/amle.2002.8509336]
17. Krathwohl DR. A Revision of Bloom's Taxonomy: An Overview. Theory Into Practice 2010 Jun 24;41(4):212-218. [doi: 10.1207/s15430421tip4104_2]

Abbreviations

NICU: neonatal intensive care unit

Edited by J Sarvestan; submitted 25.02.25; peer-reviewed by S Thananithichot; comments to author 22.04.25; revised version received 12.05.25; accepted 22.05.25; published 18.06.25.

Please cite as:
Miyosawa Y, Hirabayashi K, Yamada K, Kobayashi F, Ogihara N, Takeda N, Okamura E, Matsumura S
Development of a Serious Game to Simulate Neonatal Intensive Care Unit Experiences: Collaborative Quasi-Experimental Study
JMIR Form Res 2025;9:e73009
URL: https://formative.jmir.org/2025/1/e73009
doi:10.2196/73009
PMID:

©Yukihide Miyosawa, Koichi Hirabayashi, Kodai Yamada, Fumiya Kobayashi, Nanami Ogihara, Noa Takeda, Eri Okamura, Shogo Matsumura. Originally published in JMIR Formative Research (https://formative.jmir.org), 18.06.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Health Services Usage in Patients Receiving Buprenorphine for Opioid Use Disorder or Long-Term Opioid Therapy for Chronic Pain: Retrospective Cohort Study

Samuel T Savitz[1], PhD; Maria A Stevens[1], MA, MPH, PhD; Bidisha Nath[2], MPH, MBBS; Gail D'Onofrio[2,3,4], MS, MD; Edward R Melnick[2,5], MHS, MD; Molly M Jeffery[1,6], PhD

[1]Division of Health Care Delivery Research, Kern Center for the Science of Health Care Delivery, Mayo Clinic, Rochester, MN, United States
[2]Department of Emergency Medicine, Yale School of Medicine, Yale University, New Haven, CT, United States
[3]Department of Chronic Disease Epidemiology, Yale School of Public Health, Yale University, New Haven, CT, United States
[4]Department of Internal Medicine, Yale School of Medicine, Yale University, New Haven, CT, United States
[5]Department of Biostatistics, Yale School of Public Health, Yale University, New Haven, CT, United States
[6]Department of Emergency Medicine, Mayo Clinic, Rochester, MN, United States

Corresponding Author:
Samuel T Savitz, PhD
Division of Health Care Delivery Research
Kern Center for the Science of Health Care Delivery
Mayo Clinic
200 1st Street SW, CSHCD – Harwick 2nd Floor
Rochester, MN, 55905
United States
Phone: 1 919 259 0843
Email: Savitz.Samuel@mayo.edu

Abstract

Background: Patients using buprenorphine for opioid use disorder (OUD) or long-term opioid therapy for chronic pain are at risk for poor outcomes if care is interrupted. Both treatments are highly regulated, with prepandemic requirements for in-person care. COVID-19 may have resulted in barriers to accessing in-person care through disruptions in care delivery. However, there were also opportunities for improved access to telemedicine visits through policy changes.

Objective: This study aims to evaluate changes in health care and telemedicine use during the COVID-19 pandemic among patients using buprenorphine for OUD and long-term opioid therapy for chronic pain.

Methods: We used administrative claims data for commercially insured and Medicare Advantage patients from the OptumLabs Data Warehouse. We included patients using buprenorphine for OUD or long-term opioid therapy for chronic pain compared to patients with another chronic condition without similar prescribing restrictions: serious mental illness. We evaluated changes in in-person and telemedicine care by comparing rates of services by physician specialty, type of service, and the percentage of visits through telemedicine. Changes in usage were measured using a difference-in-differences approach with Poisson regression. The results are presented as incident rate ratios (IRR).

Results: We found declines in in-person visits in April 2020 across the buprenorphine, chronic opioids, and serious mental illness cohorts. The largest declines were for specialties that rely on in-person treatment, such as emergency medicine (IRR range 0.60-0.62), orthopedics (IRR 0.48-0.52), cardiology (IRR 0.64-0.78), and oncology (IRR 0.77-0.81). In contrast, there were smaller declines for specialties that could more easily transition to telemedicine, namely family practice (IRR 0.80-0.92), mental health (IRR 0.92-1.01), and pain medicine (IRR 0.87-1.08). The percentage of telemedicine visits for these specialties ranged from 30% to 51% in the period. There were also large declines for specific services, including emergency medicine (IRR 0.53-0.89), physical therapy (IRR 0.24-0.72), and new office visits (IRR 0.38-0.64). By January 2022, usage was similar to prepandemic levels, but the percentage of telemedicine visits remained elevated for family practice (10%-14%), mental health (34%-43%), and pain medicine (11%-15%) through January 2022. The results were similar across the cohorts, although in April 2020 there was a modest decrease (IRR 0.87) for pain medicine in the serious mental illness cohort, but the differences were not significant for the buprenorphine (IRR 1.08) and chronic opioids (IRR: 0.99) cohorts.

XSL•FO
RenderX

Conclusions: These findings highlight the value of telemedicine to maintain access among people at risk for poor outcomes if care is interrupted. While flexibilities in the regulation of telemedicine services that arose during the pandemic have been temporarily extended multiple times, they are set to expire in 2025 without further action. Making these changes to telemedicine regulation permanent may benefit vulnerable patient populations who face access to care challenges.

(JMIR Form Res 2025;9:e66596) doi:10.2196/66596

KEYWORDS

telemedicine; opioids; buprenorphine; health care utilization; telehealth; utilization; usage; claims data; services; OUD; opioid use disorder; retrospective; cohort; pain; chronic

Introduction

Background

The COVID-19 pandemic caused disruptions to the delivery of health services [1-3]. During the early pandemic period, there were large decreases in care [2]. The decreases were most pronounced for services that could not easily transition to telemedicine such as emergency medicine, which relies heavily on physical examinations, laboratory tests, and imaging that must be done in person [3,4]. In contrast, declines tended to be smaller for services like mental health care that are less dependent on physical examinations and tests [3,5].

In response to the disruption in the delivery of health services, both public and private insurers loosened restrictions on telemedicine visits. Medicare expanded telemedicine coverage nationwide starting March 6, 2020, under an 1135 waiver [6]. The Centers for Medicare and Medicaid Services also allowed insurers to modify plan coverage midyear to provide greater coverage of telemedicine services [7]. The Coronavirus Aid, Relief, and Economic Security (CARES) Act in May 2020 provided additional flexibility for insurers with respect to telemedicine coverage [8]. In response, most private plans began offering telemedicine services with predeductible coverage [9].

There were also changes to the regulation of prescribing controlled substances. Starting on March 31, 2020, the Drug Enforcement Administration (DEA) allowed clinicians to prescribe controlled substances such as chronic opioid analgesic therapy and buprenorphine through telemedicine without first having an in-person visit [10]. In addition, the DEA began to relax the requirements for obtaining a Drug Addiction Treatment Act (2000) waiver, also known as the X-Waiver [11]. On April 27, 2021, the Department of Health and Human Services relaxed the training requirement for clinicians who were treating up to 30 patients with buprenorphine at a given time [11]. Then, the X-Waiver was eliminated on December 29, 2022, as part of the Consolidated Appropriations Act of 2023 [12].

Buprenorphine for Opioid Use Disorder and Chronic Opioid Analgesic Therapy

Disruptions in access to care during the pandemic presented health risks to all patients [13]. In this context, 2 populations of particular concern are patients who use buprenorphine for opioid use disorder (OUD) and patients who use chronic opioid analgesic therapy. These patients need ongoing treatment, including monthly prescriptions, to manage chronic health problems and already faced access-to-care barriers before the pandemic [14]. These patients are at high risk for fatal and nonfatal opioid overdoses [15-19]. In the early months of the pandemic, there were increases in emergency usage related to opioids including emergency department (ED) visits for nonfatal opioid overdoses [20] and emergency medical service runs for fatal and nonfatal opioid-related concerns [21]. Disruption in access for patients who use buprenorphine for OUD and patients who use chronic opioid analgesic therapy may have contributed to these issues.

Patients receiving opioids for chronic pain or buprenorphine for OUD may be more likely to experience disruptions in care since buprenorphine and chronic opioid analgesic therapy are controlled substances with restrictions on prescribing [22-26]. Further, clinicians prescribing opioid analgesics face restrictions that vary by state and by payer, including the use of prescription drug monitoring programs, quantity limits, and duration limits [24,25,27]. Therefore, disruptions in care may affect continued access to buprenorphine or opioid therapy.

Gap in the Literature

It is important to understand how pandemic-related care disruptions have affected these populations to inform future resource allocation and policy decisions related to ensuring ongoing access. Understanding the impact of the increase in access to telemedicine is also important, given that its continuation past the pandemic period is in question. While general population use of health care services and telemedicine has been reported [3], there is limited evidence on the experience among patients particularly vulnerable to access disruption.

Study Overview

Our objective was to assess the change in the use of health services and the impact of telemedicine on access to care for 2 patient cohorts facing barriers to access. We used a large claims database to analyze trends in health services usage for patients who have fills for buprenorphine for OUD or chronic opioid analgesic therapy and compare them to patients with another chronic condition that does not generally have prescribing restrictions: serious mental illness (SMI). We selected SMI for comparison since SMI includes chronic conditions, but unlike buprenorphine and most opioid analgesics, most treatments for SMI can be written with refills and there are few or no requirements for in-person care. SMI is also a common comorbidity for patients with OUD or chronic pain [28-30]. The purpose of including SMI as a comparison group was to descriptively assess whether the observed trends were specific to patients with buprenorphine for OUD or chronic opioid analgesic therapy.

XSL•FO
RenderX

Methods

Data Source

We evaluated the use of services and access to care from January 2018 to February 2022, using data from the OptumLabs Data Warehouse (OLDW), a deidentified, national, medical claims database [31]. OLDW includes claims data for individuals with commercial and Medicare Advantage (MA) coverage. We report the study in accordance with the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) checklist for cohort studies (Multimedia Appendix 1) [32].

Patients

We compared 3 cohorts of patients with chronic conditions that require ongoing treatment and may be affected by changes in access to care: patients with fills for buprenorphine formulations to treat OUD, patients with fills for chronic opioid analgesic therapy, and patients with SMI. We also evaluated heart failure as an additional nonmental health comparison condition and included results in Multimedia Appendix 2 as an additional sensitivity analysis. We separately identified patients belonging to each cohort. First, we identified patients with fills for buprenorphine to treat OUD. The definition for a buprenorphine episode is a period in which the patient never went more than 60 days without an active buprenorphine prescription. We only included buprenorphine formulations that were for treating OUD, excluding those used only to treat pain. We did not require patients to have a diagnosis for OUD to be included in the cohort. Second, we identified patients with fills for chronic opioid analgesic therapy. The definition of a chronic opioid episode is a period in which the patient never went more than 30 days without an active opioid prescription and had at least 10 fills or 120 days' supply. We only included buprenorphine formulations that were for treating pain. Third, we identified patients who had at least one inpatient diagnosis or two outpatient diagnoses at least 1 day apart for SMI in the previous 6 months. SMI codes came from an Arizona block grant for serious emotional disturbance patients. Patients in the SMI cohort did not have to have received medication to be in the cohort. The cohorts were not mutually exclusive. Patients in all cohorts had to have at least 30 days of medical and pharmacy insurance enrollment in January 2018-February 2022 and 183 days of continuous medical and pharmacy coverage (see "Sample Flow Diagram" in Multimedia Appendix 3). To make the cohorts more comparable, we made the start date for the evaluation of the use of at least 60 days after the beginning of continuous enrollment. Censoring occurred upon the end of continuous enrollment, February 28, 2022, at the end of buprenorphine or chronic opioid episode, and when the patient no longer had at least one inpatient diagnosis or two outpatient diagnoses for SMI at least 1 day apart for the SMI cohort. During the study period, patients could enter and be removed from the analysis depending on whether they meet the inclusion and exclusion criteria on a given date.

Measures

We first characterized the cohorts using descriptive statistics, including diagnoses for SMI [33], and selected Elixhauser comorbidities [34,35]. The key outcome was the use of health care services. We measured in-person and telemedicine visits as the specialty of the provider on the claim and the Berenson-Eggers Types of Service Codes (BETOS) 2.0, which categorizes procedure codes into meaningful groups for the management of clinical conditions [36]. Previous work has analyzed usage patterns using the provider specialty field in OLDW [3]. We defined telemedicine usage as claims with either a procedure modifier of GT, GQ, or 95 or a Current Procedural Terminology (CPT) code of 99441, 99442, 99443, G0425, G0426, G0427, G0406, G0407, G0408, or G045 [6]. We measured usage both overall among the 3 cohorts and broken out by insurance categories (commercial insurance [COM], MA younger than 65 y [MA <65 y], and MA aged 65 y or older [MA 65+]). People younger than 65 years qualify for MA due to long-term disability or certain serious medical conditions, including end-stage renal disease and experience different patterns in health care usage [37]. The numerator was the number of person-days of service, and the denominator was the number of enrollee months in the period.

We initially focused on commonly accessed clinician specialties that are important for these patient populations, selecting a range of specialties that differed in how readily they could be performed through telemedicine. The specialties initially considered were cardiology, emergency medicine, family practice, mental health, oncology, orthopedics, and pain medicine. The mental health category included visits with clinicians who have a specialty in psychiatry, psychology, or social work. The pain medicine category included visits with a specialty of anesthesiology for which the claim had an evaluation and management code. After exploring these specialties, we narrowed the focus to family practice, mental health, and pain medicine. We selected these specialties as particularly conducive to telemedicine visits since they are generally not procedure-based and often focus on the management of chronic conditions. Pharmacotherapy can be prescribed through telemedicine, and it is the most common treatment modality for mental health conditions like depression [5] and is one of the most common treatment modalities for pain medicine [38]. Many of the most common reasons for primary care visits include conditions that could be managed, in part, through telemedicine, including hypertension, diabetes, depression, and anxiety [39]. Further, these specialties are especially important for the study populations since clinicians in these specialties are among the most likely to prescribe buprenorphine to treat opioid use disorder [40]. Results for the other specialties are available in Multimedia Appendices 4 and 5.

We also evaluated 3 BETOS procedure categories: emergency medicine, physical therapy, and new office visits. These categories represent services that may be important for the included patient cohorts and for which there may have been reduced access because of their reliance on in-person visits. Emergency medicine often requires physical examinations and tests [4], as well as time-sensitive physical interventions, including, for example, administering naloxone to patients presenting with an overdose [20]. Similarly, physical therapy relies on physical examinations and treatments [41] and is an important nonpharmacological approach to pain management

XSL•FO
RenderX

[42]. New office visits may be particularly important for patients who no longer had an established provider for managing prescriptions or initiated buprenorphine in the emergency department after an overdose [43].

We also assessed how the pandemic-associated changes in usage have affected access to care for mental health. We defined mental health access in each month as the percentage of patients with a psychiatry or behavioral health visit in the 6 months before that month. We measured access among the subset of patients in the buprenorphine and chronic opioid cohorts who also had SMI. The purpose of this measure was to assess whether patients with chronic mental health condition were receiving regular follow-up care.

Design

For usage outcomes, we calculated the rate of person-days of service per 100,000 enrollee-months. To assess whether there were significant differences in usage, we performed an analysis using a difference-in-differences (DiD) approach. This approach evaluates the difference before and after COVID-19 and accounts for temporal patterns in usage by month. While DiD is often used for causal inference [44,45], in this study, we use the DiD approach to descriptively assess whether usage trends changed over time differentially for one group versus another. We focused on four comparisons to January 2020 as the reference: April 2020, July 2020, July 2021, and January 2022. We selected the 4 time points to represent key periods during the pandemic: early pandemic, April 2020; second wave, July 2020; Delta wave, July 2021; and Omicron wave, January 2022. For example, usage changes in April 2020 were calculated by evaluating whether the change in usage between January and April 2020 was different from that between January and April 2019. We chose the DiD approach for 2 reasons: first, there are no clear-cut points to define the major waves of the pandemic, which shifted somewhat gradually over time and hit at different times across the country. Second, we wanted to compare the changes with respect to a prepandemic baseline. Using our approach, we were able to compare the 4 time points to the reference time point of January 2020 to see if usage was lower or higher than that before the pandemic. We implemented the analysis using Poisson regression with the use rate as the outcome. The models also included an offset term for the natural log of the number of eligible enrollees in each cohort to model the number of events as a rate of events per enrollee [46]. We present the results as incident rate ratios (IRRs) with 95%. All analyses were performed using Stata (version 16; StataCorp, 2019).

Ethical Considerations

The study was submitted to the Mayo Clinic institutional review board (application ID 25-003376) and deemed exempt from institutional board review under Exemption Category 4 (secondary research of existing data). The data for this study were deidentified to ensure privacy for sensitive health care information.

Results

Descriptive Statistics

We identified 69,032 patients in the buprenorphine cohort, 817,381 in the chronic opioid use cohort, and 4,247,090 in the SMI cohort. The cohorts ranged (Table 1) from 32% (4426/13,717) female (COM buprenorphine cohort) to 72% (223,546/312,291) female (MA 65+ SMI cohort). The buprenorphine cohort was more likely to meet the Elixhauser comorbidity drug use disorder definition [53% (7217/13,717) COM compared to <10% for other cohort] and less likely to be female.

XSL•FO
RenderX

Table 1. Descriptive statistics for buprenorphine, chronic opioids, and serious mental illness cohorts as of January 2020.

Variable	Commercial			Medicare Advantage <65 years			Medicare Advantage 65+ years		
	Buprenorphine	CO[b]	SMI[c]	Buprenorphine	CO	SMI	Buprenorphine	CO	SMI
Total (N)	13,717	69,979	534,318	5059	92,432	118,077	2506	180,900	312,291
Age range (years), % (n)									
0-25	3.2 (440)	0.5 (344)	24.6 (131,297)	8.0 (404)	2.3 (2083)	6.9 (8133)	—[d]	—	—
0-39									
26-39	43.5 (5967)	10.2 (7149)	30.3 (161,984)	92.0 (4655)	97.8 (90,349)	93.1 (109,944)	—	—	—
40-64	51.7 (7091)	79.7 (55,764)	41.4 (221,202)	—	—	—	87.3 (2188)	58.0 (104,904)	50.2 (156,638)
65+	1.6 (219)	9.6 (6722)	3.7 (19,835)	—	—	—	12.7 (318)	42.0 (75,996)	49.8 (155,653)
65-74	—	—	—	—	—	—	87.3 (2188)	58.0 (104,904)	50.2 (156,638)
75+	—	—	—	—	—	—	12.7 (318)	42.0 (75,996)	49.8 (155,653)
Gender, % (n)									
Women	32.3 (4426)	57.0 (39,904)	65.0 (346,965)	49.4 (2500)	59.4 (54,939)	63.2 (74,578)	43.4 (1088)	64.1 (115,889)	71.6 (223,546)
Race and ethnicity, % (n)									
Asian or PI[e]	0.9 (129)	1.1 (764)	2.2 (11,991)	0.6 (30)	0.7 (648)	1.0 (1145)	0.5[f] (12)	0.9 (1557)	1.3 (4064)
Black	5.7 (781)	8.4 (5872)	6.4 (33,959)	12.0 (607)	21.7 (20,046)	18.5 (21,840)	13.7 (343)	17.3 (31,373)	11.5 (35,813)
Hispanic	5.1 (705)	6.7 (4713)	7.7 (41,075)	5.5 (276)	6.4 (5865)	8.2 (9708)	5.3 (132)	6.0 (10,889)	9.3 (28,990)
White	70.3 (9644)	71.5 (50,066)	66.7 (356,143)	69.1 (3494)	59.8 (55,299)	59.1 (69,760)	69.4 (117,989)	65.2 (117,989)	65.7 (205,273)
Other	17.9 (2458)	12.2 (8564)	17.1 (91,150)	12.9 (652)	11.4 (10,574)	13.2 (15,624)	11.2 (281)	10.6 (19,092)	12.2 (38,151)
Elixhauser comorbidities, % (n)									
Alcohol use disorder	3.8 (524)	0.9 (659)	3.0 (16,247)	6.8 (345)	1.9 (1780)	5.1 (6072)	5.5 (137)	1.1 (1905)	2.4 (7351)
Drug use disorder	52.6 (7217)	4.3 (2998)	3.3 (17,655)	57.4 (2904)	7.5 (6973)	9.6 (11,288)	50.2 (1259)	4.0 (7287)	2.5 (7772)
Diabetes	2.8 (384)	11.0 (7713)	4.8 (25,686)	14.8 (746)	29.6 (27,327)	27.6 (32,540)	18.4 (462)	30.7 (55,500)	28.6 (89,165)
Other comorbidities, % (n)									
SMI[g]	25.7 (3526)	19.3 (13,528)	100 (534,318)	50.2 (2537)	34.5 (31,907)	100 (118,077)	38.9 (974)	23.5 (42,452)	100 (312,291)
Anxiety disorders	17.1 (2348)	12.8 (8976)	62.4 (333,431)	30.7 (1552)	20.5 (18,976)	42.9 (50,604)	22.8 (571)	13.6 (24,650)	44.2 (138,104)
Bipolar disorders	2.7 (376)	1.5 (1055)	6.7 (35,969)	13.4 (678)	5.7 (5243)	20.9 (24,676)	3.4 (84)	1.5 (2780)	5.9 (18,289)
Depressive disorders	14.4 (1977)	10.8 (7543)	50.1 (267,522)	30.0 (1518)	21.3 (19,961)	50.0 (59,066)	25.8 (27,036)	15.0 (27,036)	52.6 (164,331)
Dissociative disorders	0.0 (0)[f]	0.0 (0)[f]	0.0 (188)	0.1 (0)[f]	0.0 (18)	0.1 (123)	0.0 (0)[f]	0.0 (0)[f]	0.0 (70)

XSL•FO
RenderX

Variable	Commercial			Medicare Advantage <65 years			Medicare Advantage 65+ years		
	Buprenorphine	CO[b]	SMI[c]	Buprenorphine	CO	SMI	Buprenorphine	CO	SMI
Obsessive-compulsive disorders	0.4 (53)	0.2 (117)	2.6 (11,653)	0.6 (28)	0.3 (241)	1.3 (1529)	0.2 (0)[f]	0.1 (180)	0.6 (1898)
Other mood disorders	0.8 (104)	0.3 (173)	1.7 (9321)	1.3 (67)	0.5 (442)	1.7 (1966)	1.0 (26)	0.3 (536)	1.2 (3841)
Personality disorders	0.3 (47)	0.1 (100)	0.9 (5049)	1.7 (88)	0.6 (571)	2.7 (3126)	0.4 (0)[f]	0.2 (275)	0.5 (1616)
Psychotic disorders	0.5 (65)	0.2 (121)	1.4 (7640)	4.2 (212)	2.2 (2054)	16.1 (18,977)	1.5 (38)	1.0 (1774)	5.7 (17,715)
Trauma-related disorders	2.4 (334)	1.3 (909)	5.9 (31,661)	8.6 (437)	3.4 (3150)	10.5 (12,345)	2.6 (66)	0.7 (1246)	2.2 (6699)

[b]CO: chronic opioid.

[c]SMI: serious mental illness.

[d]Not applicable.

[e]PI: Pacific Islander.

[f] Number suppressed due to a cell size <11

[g]The number of Asian or PI patients was <11 for the Medicare Advantage 65+ buprenorphine cohort, so we combined this group with the Other/Unknown group.

Specialty Usage

In April 2020, the IRR for changes in use and the percentage of telemedicine visits, compared to those before COVID-19, were similar for patients in the buprenorphine and chronic opioid cohorts relative to the SMI cohort across 7 key specialties (Figure 1; Multimedia Appendix 4). Usage declined relative to prepandemic for emergency medicine (IRR range 0.60-0.62), orthopedics (IRR 0.48-0.52), cardiology (IRR 0.64-0.78), and oncology (IRR 0.77-0.81), while the corresponding percentage of telemedicine visits for these services remained below 20%. In contrast, usage was similar during COVID-19 compared to that before COVID-19 for family practice (IRR 0.80-0.92), mental health (IRR 0.92-1.01), and pain medicine (IRR 0.87-1.08). The use of telemedicine visits was also greater than before the pandemic and ranged from 30% to 51%.

Figure 1. Difference-in-differences incidence rate ratio for change in utilization and telemedicine use across seven specialty categories in the buprenorphine, chronic opioids, and serious mental illness cohorts pre versus post COVID-19 April 2020.

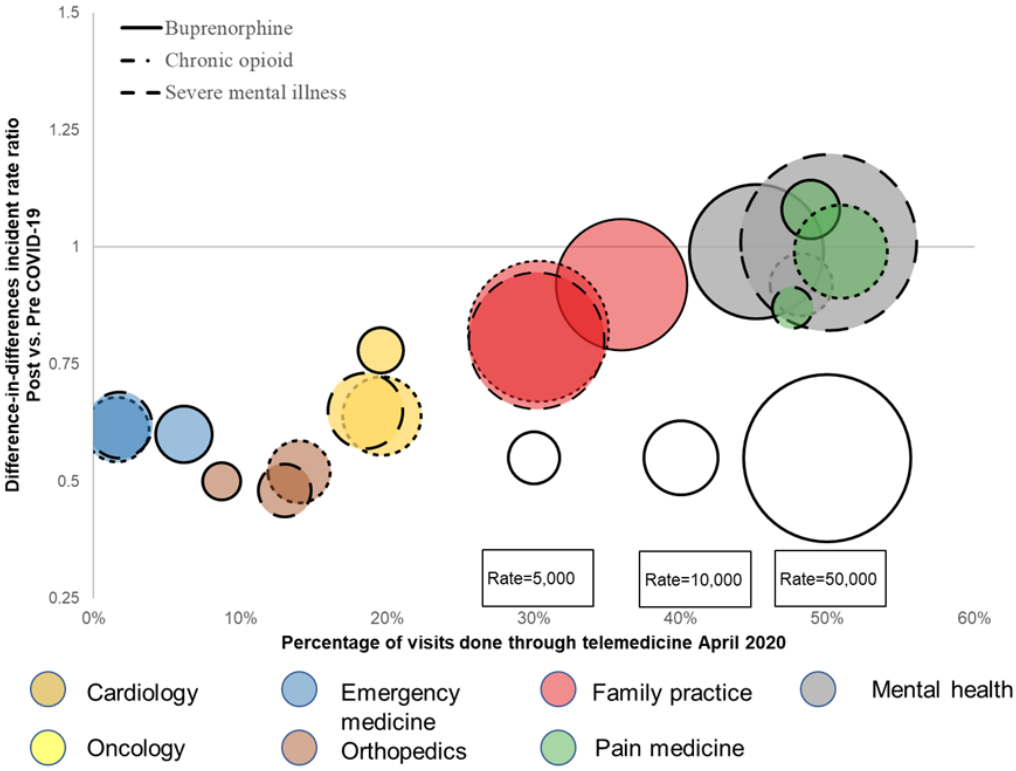

XSL•FO
RenderX

While the results were generally similar for the buprenorphine and chronic opioid cohorts compared to the SMI cohort, some differences were present. For pain medicine, the buprenorphine (IRR 1.08) and chronic opioid (IRR 0.99) cohorts experienced no significant difference in usage, whereas the SMI (IRR 0.87) cohort experienced a modest decrease compared to each cohort's usage before the pandemic.

In Figure 1, bubble size is the rate of person-days of service per 100,000 patient-months in April 2020. The incidence rate ratio for oncology in the buprenorphine cohort is not displayed because the rate is below 1000 person-days of service per 100,000 person-months.

In January 2022 during the Omicron wave, usage was similar to prepandemic levels (Figure 2; Multimedia Appendix 4). Telemedicine use was also much lower than in April 2020. The percentage of visits that were telemedicine was 5% or lower for all cohorts of emergency medicine, oncology, orthopedics, and cardiology. In contrast, telemedicine use remained moderate for family practice (10%-14% of visits) and pain medicine (11%-15%) and high for mental health (34%-43%).

Figure 2. Difference-in-differences incidence rate ratio for change in utilization and telemedicine use across seven specialty categories in the buprenorphine, chronic opioids, and serious mental illness cohorts pre versus post COVID-19 January 2022.

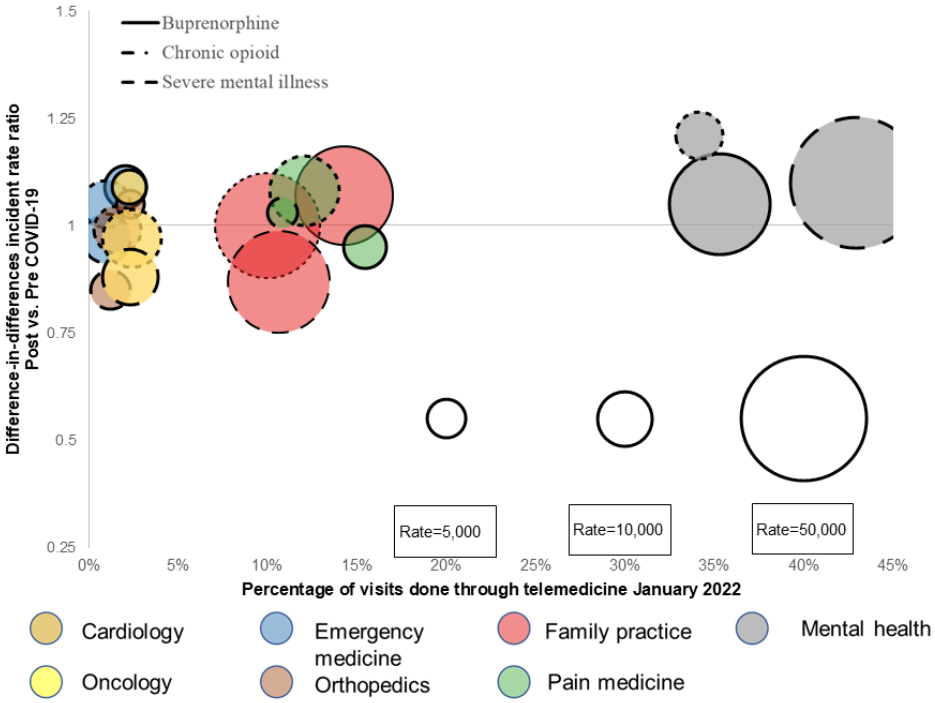

Separating the cohorts by insurance categories revealed similar patterns (Table 2, Multimedia Appendix 4). Usage was lower for family practice (IRR 0.74-0.95 for COM, 0.84-0.87 for MA <65 y, and 0.79-0.93 for MA 65+) in April 2020 than in April 2019 (Table 2). However, there was a smaller reduction or no decrease for pain medicine (IRR 0.84-0.96 for COM, 0.97-1.13 for MA <65 y, and 0.79-1.12 for MA 65+) and mental health (IRR 0.92-1.03 for COM, 1.00-1.08 for MA <65 y, and 0.80-1.05 for MA 65+). The percentage of telemedicine visits for April 2020 was high for all specialties across all cohorts and insurance groups. Telemedicine ranged from 23% to 47% of visits for family practice, 45%-55% for pain medicine, and 40%-56% for mental health. In contrast, the percentage of telemedicine visits was below 2% for all cohorts in April 2019.

XSL•FO
RenderX

Table 2. Rate per 100,000 person-months, difference-in-differences incidence rate ratios, and telemedicine use for select specialties in the buprenorphine, chronic opioids, and serious mental illness cohorts April 2020-January 2022. The rate is in terms of person-days of service per 100,000 patient-months.

Insurance (cohort)	January 2020, rate per 100,000 PMs[a]	April 2020		July 2020		July 2021		January 2022	
		IRR[b]	Tel (%)[c]	IRR	Tel (%)	IRR	Tel (%)	IRR	Tel (%)
Family practice									
Commercial									
Bup[d]	23,467.1	0.95	35	1.03	20	0.90[e]	11	0.90[e]	15
CO[f]	21,461.9	0.84[g]	42	0.96[h]	19	0.93[i]	8	1.00	12
SMI[i]	16,491.2	0.74[g]	47	1.00	23	0.89[g]	10	0.98[e]	15
MA[j] <65 y									
Bup	44,592.4	0.84[g]	37	0.93	20	0.94	11	1.01	15
CO	38,879.2	0.87[g]	31	0.99	14	0.98	6	1.00	12
SMI	47,487.2	0.85[g]	30	0.99	15	0.87[g]	8	0.87[g]	12
MA 65+									
Bup	42,316.4	0.93	36	0.86[h]	18	1.01	9	1.19[h]	12
CO	38,328.8	0.79[g]	27	0.95[g]	12	0.93[g]	5	0.97[g]	9
SMI	53,511.9	0.79[g]	23	0.95[g]	12	0.79[g]	5	0.80[g]	8
Pain medicine									
Commercial									
Bup	3886.0	0.95	45	0.98	34	0.91	10	0.71[g]	13
CO	16,513.7	0.96	55	0.99	31	0.97	11	0.92[g]	14
SMI	1310.1	0.84[g]	51	1.02	25	0.91[g]	9	0.88[g]	13
MA <65 y									
Bup	10,583.7	1.13	51	1.03	35	0.96	13	0.78[h]	16
CO	20,565.7	1.05[e]	47	1.08[g]	23	1.05[g]	9	0.97[h]	12
SMI	8394.9	0.97	48	1.11[g]	22	1.03	9	0.91[g]	12
MA 65+									
Bup	10,655.2	1.12	50	1.18	24	1.22	12	1.44[h]	16
CO	13,132.7	0.95[g]	52	1.03[h]	26	1.15[g]	9	1.17[g]	12
SMI	4057.6	0.79[g]	45	1.04	20	1.12[g]	7	1.12[g]	9
Mental health									
Commercial									
Bup	31,061.2	0.97	41	1.11[e]	38	1.01	31	0.89[e]	33
CO	7047.7	0.92[e]	56	1.05	50	1.02	44	1.08[h]	46
SMI	72,180.4	1.03[g]	52	1.18[g]	51	1.08[g]	45	1.08[g]	48
MA <65 y									
Bup	39,397.7	1.00	50	1.06	48	1.15[e]	33	1.12[f]	36
CO	9484.2	1.06[e]	52	1.12[g]	47	1.18[g]	37	1.32[g]	40

XSL•FO
RenderX

Insurance (cohort)	January 2020, rate per 100,000 PMs[a]	April 2020		July 2020		July 2021		January 2022	
		IRR[b]	Tel (%)[c]	IRR	Tel (%)	IRR	Tel (%)	IRR	Tel (%)
SMI	51,595.6	1.08[g]	48	1.14[g]	43	1.09[g]	34	1.19[g]	37
MA 65+									
Bup	26,536.6	1.05	54	1.22[h]	48	1.38[g]	38	1.04	41
CO	6033.9	0.80[g]	40	0.89[g]	37	0.98	22	1.11[g]	25
SMI	30,906.6	0.91[g]	42	1.14[g]	35	1.01	28	1.10[g]	31

[a]PM: person-month.
[b]IRR: incident rate ratio.
[c]Tel: telemedicine.
[d]Bup: Buprenorphine.
[e]$P<.01$.
[f]CO: chronic opioid.
[g]$P<.001$.
[h]$P<.05$.
[i]SMI: serious mental illness.
[j]MA: Medicare Advantage.

In Figure 2, bubble size is the rate of person-days of service per 100,000 patient-months in January 2022. The IRR for oncology in the buprenorphine cohort is not displayed because the rate is below 1000 person-days of service per 100,000 person-months.

Total usage for the same specialties largely recovered in July 2020 compared to July 2019 and remained at similar levels in July 2021 and January 2022. The percentage of telemedicine visits was lower for family practice and pain medicine, but not mental health (July 2021: 35%-51%; July 2021: 22%-45%; January 2022: 25%-48%).

BETOS Usage

Using the BETOS classification, there were large declines in total usage in emergency medicine (IRR 0.53-0.89), physical therapy (IRR 0.24-0.72), and new office visits (IRR 0.38-0.64) in April 2020 compared to April 2019 (Table 3). Except for the buprenorphine cohort in the MA 65+ group for emergency medicine and physical therapy, usage significantly decreased for all cohorts and insurance groups. The declines were consistent across insurance groups. Usage mostly recovered by July 2020 for physical therapy and new office visits. However, usage was still moderately lower for emergency medicine for 10 out of 12 cohorts and insurance groups in July 2020. By July 2021, use of emergency medicine had further recovered. In January 2022, there was a mix of increases and decreases with no consistent pattern.

XSL•FO
RenderX

Table 3. Rate and difference-in-differences incidence rate ratios for emergency medicine, physical therapy, and new office visits Berenson-Eggers Types of Service Codes categories in the buprenorphine, chronic opioids, and serious mental illness cohorts April 2020-January 2022.

Insurance (cohort)	January 2020, rate	April 2020, DiD[a]	July 2020, DiD	July 2021, DiD	January 2022, DiD
Emergency medicine					
COM[b]					
Bup[c]	3228	0.53[d]	0.91	1.10	1.33[e]
CO[f]	5108	0.57[d]	0.79[d]	0.95	1.14[d]
SMI[g]	3994	0.54[d]	0.93[d]	0.93[d]	0.89[d]
MA[h] <65 y					
Bup	8545	0.60[d]	0.79[e]	0.86	1.25
CO	9614	0.58[d]	0.80[d]	1.08[d]	1.06[e]
SMI	15,770	0.63[d]	0.85[d]	0.92[d]	0.88[d]
MA 65+					
Bup	6495	0.89	1.39[e]	1.67[i]	0.83
CO	6637	0.57[d]	0.87[d]	1.25[d]	1.16[d]
SMI	9567	0.60[d]	0.84[d]	1.01	1.02
Physical therapy					
COM					
Bup	3228	0.63[d]	1.04	1.05	0.67[d]
CO	14,109	0.49[d]	0.86[d]	1.05[e]	1.02
SMI	13,228	0.44[d]	0.87[d]	0.92[d]	0.79[d]
MA<65 y					
Bup	5243	0.24[d]	0.98	0.96	1.53[e]
CO	13,176	0.50[d]	0.99	1.14[d]	0.81[d]
SMI	22,090	0.62[d]	0.98	0.96[i]	0.72[d]
MA 65+					
Bup	11,771	0.72	1.18	0.95	0.98
CO	21,934	0.64[d]	0.95[d]	1.16[d]	1.02
SMI	46,280	0.71[d]	1.02[i]	0.98[d]	0.84[d]
New office visits					
COM					
Bup	6349	0.64[d]	1.05	1.04	1.27[i]
CO	8779	0.55[d]	0.90[d]	1.02	1.07[e]
SMI	9345	0.51[d]	1.02	0.97[i]	1.00
MA<65 y					
Bup	8812	0.60[d]	1.08	0.98	0.75[e]
CO	8689	0.51[d]	0.95[e]	1.05	0.97
SMI	10,857	0.49[d]	0.97	0.96[e]	0.85[d]
MA 65+					

XSL•FO
RenderX

Insurance (cohort)	January 2020, rate	April 2020, DiD[a]	July 2020, DiD	July 2021, DiD	January 2022, DiD
Bup	9590	0.51[d]	0.82	1.08	1.03
CO	7444	0.41[d]	0.90[d]	1.16[d]	1.14[d]
SMI	9744	0.38[d]	0.92[d]	1.04[i]	0.94[d]

[a]DiD: difference-in-differences.
[b]COM: Commercial insurance.
[c]Bup: Buprenorphine.
[d]$P<.001$.
[e]$P<.05$.
[f]CO: chronic opioid.
[g]SMI: serious mental illness.
[h]MA: Medicare Advantage.
[i]$P<.01$.

Mental Health Access

Among patients in the buprenorphine and chronic opioid cohorts who also had a diagnosis for SMI, there were mixed results for access to mental health services. The percentage of patients with behavioral health visits in the previous 6 months fell slightly or stayed flat during the early COVID-19 period (Figure 3). The change from January 2020 to January 2022 ranged from a decrease of 4% points to an increase of 5% points. Patients in the buprenorphine cohort had decreases (–1% to –3%) whereas patients in the chronic opioid analgesic therapy cohort (2%-5%) had increases. The results were largely similar when evaluating psychiatry visits (Multimedia Appendix 5).

In Figure 3, The line indicates the beginning of the COVID-19 pandemic in April 2020. The outcome measures the percentage of patients with a visit in the usage category within the prior 6 months conditional on having coverage for at least 6 months.

Figure 3. Percent of patients with behavioral health services in previous six months in the buprenorphine and chronic opioids cohorts January 2019-February 2022. The line indicates the beginning of the COVID-19 pandemic in April 2020. MA: Medicare Advantage.

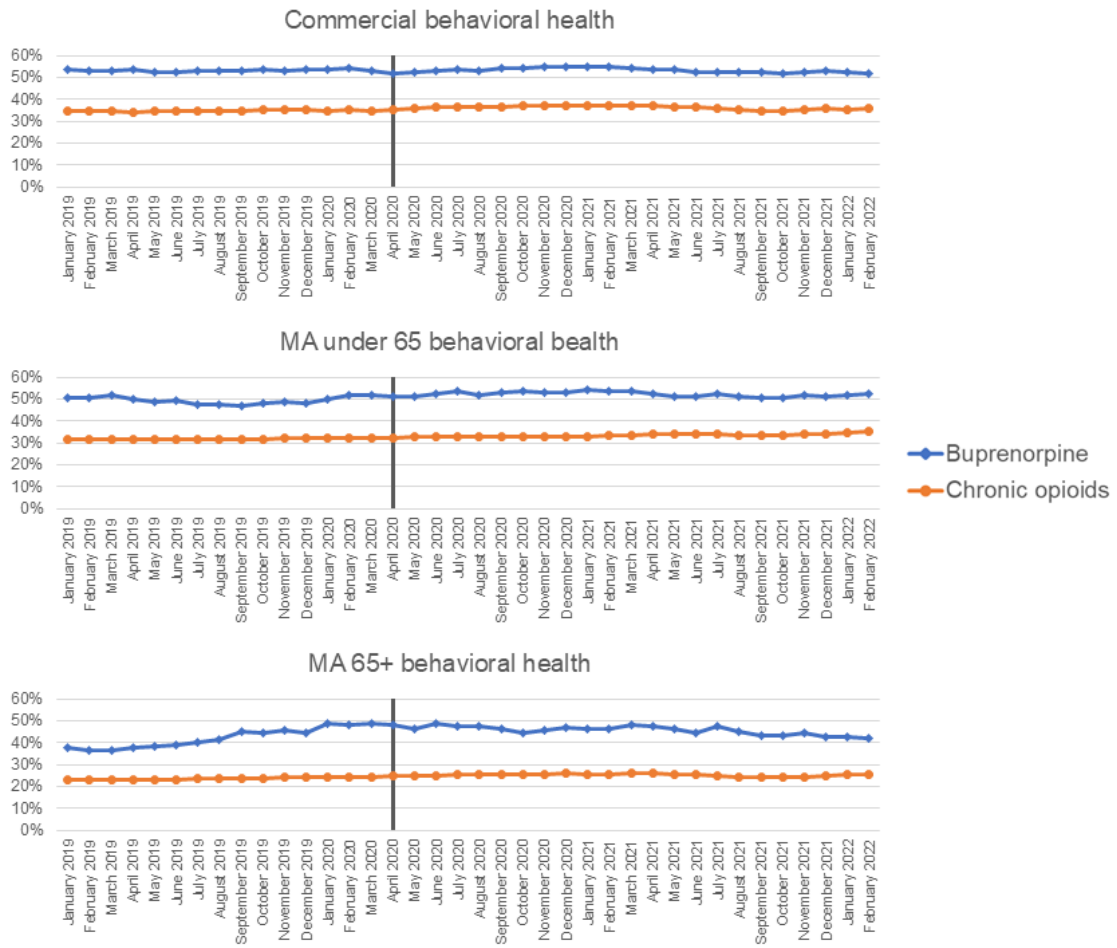

XSL•FO
RenderX

Discussion

Summary

We found that among patients with fills for buprenorphine for OUD or chronic opioid analgesic therapy during the early period of COVID-19, the total use rate for many key specialties was similar or only slightly lower than prepandemic. This pattern appeared to be supported by the substitution of in-person visits with telemedicine. In addition, the relative usage and access for patients who use buprenorphine or chronic opioid analgesic therapy were similar or higher than another comparative chronic condition cohort, suggesting that these cohorts were able to maintain access to care during the pandemic.

We also found an expected substantial decrease in usage for specialties and services that could not be readily performed via telemedicine. While usage had largely recovered to prepandemic levels by the end of the study period, telemedicine use remained elevated, especially for mental health services. However, despite the continued use of telemedicine in mental health care, access did not improve during the study period—the availability of telemedicine appears primarily to have substituted for in-person visits rather than increasing overall access to mental health care. This may reflect the shortages in the mental health care workforce prevalent across much of the country [47].

Services That Could Not Easily Shift to Telemedicine

We found decreased usage in April 2020 for service categories that could not easily transition to telemedicine, including emergency medicine, orthopedics, and physical therapy. These services are critical for many patients among the buprenorphine and chronic opioid analgesic cohorts, and the reduced access may have contributed to worse outcomes. Patients with overdoses may be seen in the ED, with about 21% of patients who have an opioid-related condition or opioid overdose as the primary diagnosis for the ED visit being admitted to the hospital [48]. One concern is that the decreased use of emergency medicine may include delayed care for opioid overdoses similar to what was found for myocardial infarction [49]. Instead, the rate of opioid overdose ED visits actually increased despite a decrease in all-cause ED visits [20]. However, the number of excess deaths from drug overdoses also increased during the early months of COVID-19 [50], and there may have also been more patients experiencing overdoses that did not reach the hospital. In addition, some patients who receive treatment for OUD have it initiated at the ED, so it is possible that the decline in access to the ED may have reduced opportunities to initiate treatment. This pattern may be related to the lower initiation of medications for OUD during the early months of the pandemic [51].

The decrease in usage for physical therapy is also concerning because chronic opioid analgesic patients who receive early physical therapy tend to have shorter durations and lower intensity of opioid use [52]. More research is needed to understand how the declines in access to emergency medicine, orthopedics, and physical therapy potentially affected buprenorphine and chronic opioid analgesic use during this period and whether there is an association with the increase in fatal opioid overdoses [53].

Role of Telemedicine for Maintaining Access

These findings also suggest that the expanded use of telemedicine was important for maintaining access to certain specialties. Our finding that greater use of telemedicine was associated with a lower decline in usage was similar to another study focused on all patients [3]. Our study confirms that vulnerable patient populations, including those who receive buprenorphine or chronic opioid analgesic therapy as well as patients with SMI, depended on telemedicine for continued access to care. We also confirmed that usage largely recovered after the initial months of the pandemic. Telemedicine use declined for all specialties but remained higher than prepandemic for family practice and pain medicine and much higher for mental health. Despite the increased use of telemedicine for mental health, access to mental health services was largely unchanged.

Policy makers should consider our findings with respect to the role of telemedicine in maintaining access to key specialties when deciding regulations, including recent proposed changes from the DEA [54,55]. Continued access to telemedicine may be important for maintaining access to care during other disruptions in care delivery such as ongoing provider shortages and the need for OUD treatment given the high numbers of fatal opioid overdoses [53]. Telemedicine also has the potential to reduce disparities for patients in rural areas with limited access to mental health providers [56-59]. However, there are also inequities in broadband availability [60]. and digital literacy [61] that may affect access to telemedicine. Individuals needing OUD treatment tend to have lower income, with 7.5% of those with income <100% of the federal poverty line needing OUD treatment compared to only 2.5% of those with income ≥200% of the federal poverty line. Further, individuals needing OUD treatment are also more likely to live in counties outside of metropolitan areas (4.7% vs 3.5%) [62]. As such, difficulty accessing broadband may be particularly salient for many individuals needing OUD treatment. Other access barriers, such as shortages of mental health providers [63], and inequities in who uses telemedicine [64], may also limit improvements in access to telemedicine. Improvements in access may depend on telemedicine being available with minimal restrictions by encounter type and for both new and established patients and on policies being designed with consideration of equity principles [65].

However, access to telemedicine may worsen if temporary waivers of telemedicine regulation made during the pandemic are allowed to lapse. The DEA has continued to grant temporary extensions for permitting the prescribing of controlled medications for OUD such as buprenorphine through telemedicine. Most recently, the DEA extended the flexibility for the third time through December 31, 2025 [66]. However, since the extensions are temporary, the potential remains for this flexibility to no longer be extended in the future. Similarly, the CARES Act included telemedicine waivers that (1) allowed patients to reside in any part of the country and to be in any setting, (2) eliminated the need for an in-person visit before receiving behavioral health visits through telemedicine, (3) extended which practitioners were eligible to bill for telemedicine services, and (4) allowed audio-only telemedicine

XSL•FO
RenderX

for some services [8,67]. These waivers were extended in the Consolidated Appropriations Act of 2023 and were set to expire on December 31, 2024 [12]. Then, the American Relief Act was passed in December 2024, and it temporarily extended the waivers until March 31, 2025 [68]. Instead of additional temporary extensions for the prescribing of buprenorphine through telemedicine and the general telemedicine waivers, a permanent revision to the regulations would help ensure continued access.

Limitations

A limitation of this analysis is that the OLDW does not include uninsured individuals or those with Medicaid coverage, who are more likely to have challenges in access to care such as a lower rate of physicians accepting new patients with Medicaid compared to new patients with commercial or Medicare coverage [69]. Medicaid accounts for approximately 25% of spending for mental health treatment and 21% of spending for substance use disorder treatment [70]. As such, it is an important population, and more research is needed to understand how access and usage changed during the period for individuals with Medicaid coverage as well as individuals who were uninsured. While Medicaid is a major payer for substance use disorders, commercial insurance is also an important payer, covering 58% of nonolder adults with substance use disorder [71]. In addition, it is important to note that this study's focus on people with health insurance allows us to assess the impact of the pandemic and increased coverage of telemedicine on people with important chronic healthcare needs, while abstracting away from access challenges related to insurance status.

Conclusions

Among commercially insured and MA patients, the availability of telemedicine in the early period of the pandemic may have helped patients, including those who use buprenorphine and chronic opioid analgesic therapy, retain access to needed services. However, the continued use of telemedicine in mental health care did not improve access to care. For services among which access decreased, the decline was no greater for patients who use buprenorphine or chronic opioid analgesic therapy compared to patients with SMI. It will remain important to maintain access for these vulnerable populations in future disruptions to health care delivery.

Acknowledgments

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: this work is supported within the National Institutes of Health (NIH) Health Care Systems Research Collaboratory by the NIH Common Fund (cooperative agreement U24AT009676 from the Office of Strategic Coordination within the Office of the NIH Director and cooperative agreement UH3DA047003 from the National Institute on Drug Abuse of the National Institutes of Health). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. Generative artificial intelligence was not used in the generation of this manuscript.

Data Availability

The datasets generated or analyzed during this study are not publicly available because they were used under license for the current study in an agreement with OLDW. Data are, however, available from the authors upon reasonable request and with permission of OLDW.

Conflicts of Interest

None declared.

Multimedia Appendix 1
STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) checklist for cohort studies.
[DOCX File , 34 KB - formative_v9i1e66596_app1.docx]

Multimedia Appendix 2
Heart failure cohort sensitivity analysis.
[DOCX File , 28 KB - formative_v9i1e66596_app2.docx]

Multimedia Appendix 3
Sample flow diagram.
[PNG File , 178 KB - formative_v9i1e66596_app3.png]

Multimedia Appendix 4
Additional results for specialty usage in the buprenorphine and chronic opioids cohorts, April 2020-January 2022.
[DOCX File , 33 KB - formative_v9i1e66596_app4.docx]

XSL•FO
RenderX

Multimedia Appendix 5
Access to psychiatry visits in the buprenorphine and chronic opioids, cohorts, April 2020-January 2022.
[PNG File , 39 KB - formative_v9i1e66596_app5.png]

References

1. Alexander GC, Tajanlangit M, Heyward J, Mansour O, Qato DM, Stafford RS. Use and content of primary care office-based vs telemedicine care visits during the COVID-19 pandemic in the US. JAMA Netw Open 2020;3(10):e2021476 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.21476] [Medline: 33006622]
2. Whaley CM, Pera MF, Cantor J, Chang J, Velasco J, Hagg HK, et al. Changes in health services use among commercially insured US populations during the COVID-19 pandemic. JAMA Netw Open 2020;3(11):e2024984 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.24984] [Medline: 33151319]
3. Patel SY, Mehrotra A, Huskamp HA, Uscher-Pines L, Ganguli I, Barnett ML. Variation in telemedicine use and outpatient care during the COVID-19 pandemic In the United States. Health Aff (Millwood) 2021;40(2):349-358 [FREE Full text] [doi: 10.1377/hlthaff.2020.01786] [Medline: 33523745]
4. Cameron P, Jelinek G, Kelly AM, Murray L, Brown A. Textbook of Adult Emergency Medicine. New York, NY: Churchill Livingstone; 2015.
5. Olfson M, Blanco C, Marcus SC. Treatment of adult depression in the United States. JAMA Intern Med 2016;176(10):1482-1491. [doi: 10.1001/jamainternmed.2016.5057] [Medline: 27571438]
6. Medicare Telemedicine Health Care Provider Fact Sheet. CMS. 2020. URL: https://www.cms.gov/newsroom/fact-sheets/medicare-telemedicine-health-care-provider-fact-sheet [accessed 2020-03-07]
7. FAQs on availability and usage of telehealth services through private health insurance coverage in response to Coronavirus disease. CMS. 2020. URL: https://www.cms.gov/files/document/faqs-telehealth-covid-19.pdf [accessed 2020-03-24]
8. S. 3548 - Coronavirus Aid, Relief, and Economic Security (CARES) Act. Taxnotes. 2020. URL: https://www.taxnotes.com/research/federal/legislative-documents/proposed-legislation/s-3548-coronavirus-aid-relief-and-economic-security-cares-act/2c9nz [accessed 2020-03-19]
9. America's Health Insurance Plans. Telehealth Coverage During the COVID-19 Pandemic. 2020. URL: https://www.ahip.org/resources/telehealth-coverage-during-the-covid-19-pandemic [accessed 2022-03-17]
10. Drug Enforcement Administration (DEA). How to prescribe controlled substances to patients during the COVID-19 public health emergency. URL: https://www.deadiversion.usdoj.gov/GDP/RESCINDED_(DEA-DC-023)(DEA075)Decision_Tree_(Final)_33120_2007.pdf [accessed 2024-11-19]
11. SAMHSA - Become a Buprenorphine Waivered Practitioner. AHRQ. 2022. URL: https://integrationacademy.ahrq.gov/resources/17556 [accessed 2025-05-07]
12. Consolidated Appropriations Act, 2023, US Congress, 117th Congress (2021-2022) Sess. 2022. URL: https://www.govtrack.us/congress/bills/117/hr2617 [accessed 2022-12-28]
13. Zhang X, Elsaid MI, DeGraffinreid C, Champion VL, Paskett ED, Impact of COVID-19 on Behaviors across the Cancer Control Continuum in Ohio group. Impact of the COVID-19 pandemic on cancer screening delays. J Clin Oncol 2023;41(17):3194-3202 [FREE Full text] [doi: 10.1200/JCO.22.01704] [Medline: 36735899]
14. Alexander GC, Stoller KB, Haffajee RL, Saloner B. An epidemic in the midst of a pandemic: Opioid use disorder and COVID-19. Ann Intern Med 2020;173(1):57-58. [doi: 10.7326/m20-1141]
15. Olfson M, Rossen LM, Wall MM, Houry D, Blanco C. Trends in intentional and unintentional opioid overdose deaths in the United States, 2000-2017. JAMA 2019;322(23):2340-2342 [FREE Full text] [doi: 10.1001/jama.2019.16566] [Medline: 31846008]
16. Bohnert ASB, Valenstein M, Bair MJ, Ganoczy D, McCarthy JF, Ilgen MA, et al. Association between opioid prescribing patterns and opioid overdose-related deaths. JAMA 2011;305(13):1315-1321. [doi: 10.1001/jama.2011.370] [Medline: 21467284]
17. Dunn KM, Saunders KW, Rutter CM, Banta-Green CJ, Merrill JO, Sullivan MD, et al. Opioid prescriptions for chronic pain and overdose: a cohort study. Ann Intern Med 2010;152(2):85-92 [FREE Full text] [doi: 10.7326/0003-4819-152-2-201001190-00006] [Medline: 20083827]
18. Morgan JR, Schackman BR, Weinstein ZM, Walley AY, Linas BP. Overdose following initiation of naltrexone and buprenorphine medication treatment for opioid use disorder in a United States commercially insured cohort. Drug Alcohol Depend 2019;200:34-39 [FREE Full text] [doi: 10.1016/j.drugalcdep.2019.02.031] [Medline: 31082666]
19. Williams AR, Samples H, Crystal S, Olfson M. Acute care, prescription opioid use, and overdose following discontinuation of long-term buprenorphine treatment for opioid use disorder. Am J Psychiatry 2020;177(2):117-124. [doi: 10.1176/appi.ajp.2019.19060612] [Medline: 31786933]
20. Soares WE, Melnick ER, Nath B, D'Onofrio G, Paek H, Skains RM, et al. Emergency department visits for nonfatal opioid overdose during the COVID-19 pandemic across six US health care systems. Ann Emerg Med 2022;79(2):158-167 [FREE Full text] [doi: 10.1016/j.annemergmed.2021.03.013] [Medline: 34119326]

XSL•FO
RenderX

21. Slavova S, Rock P, Bush HM, Quesinberry D, Walsh SL. Signal of increased opioid overdose during COVID-19 from emergency medical services data. Drug Alcohol Depend 2020;214:108176 [FREE Full text] [doi: 10.1016/j.drugalcdep.2020.108176] [Medline: 32717504]
22. Fiscella K, Wakeman SE, Beletsky L. Buprenorphine deregulation and mainstreaming treatment for opioid use disorder: X the X waiver. JAMA Psychiatry 2019;76(3):229-230. [doi: 10.1001/jamapsychiatry.2018.3685] [Medline: 30586140]
23. Hartnett KP, Kite-Powell A, DeVies J, Coletta MA, Boehmer TK, Adjemian J, National Syndromic Surveillance Program Community of Practice. Impact of the COVID-19 pandemic on emergency department visits - United States, January 1, 2019-May 30, 2020. MMWR Morb Mortal Wkly Rep 2020;69(23):699-704 [FREE Full text] [doi: 10.15585/mmwr.mm6923e1] [Medline: 32525856]
24. Meara E, Horwitz JR, Powell W, McClelland L, Zhou W, O'Malley AJ, et al. State legal restrictions and prescription-opioid use among disabled adults. N Engl J Med 2016;375(1):44-53. [doi: 10.1056/nejmsa1514387]
25. Prescribing policies: states confront opioid overdose epidemic. NCSL. 2017. URL: https://www.legis.iowa.gov/docs/publications/SD/864399.pdf [accessed 2022-07-08]
26. Haffajee RL, Bohnert ASB, Lagisetty PA. Policy pathways to address provider workforce barriers to buprenorphine treatment. Am J Prev Med 2018;54(6 Suppl 3):S230-S242 [FREE Full text] [doi: 10.1016/j.amepre.2017.12.022] [Medline: 29779547]
27. Lin DH, Jones CM, Compton WM, Heyward J, Losby JL, Murimi IB, et al. Prescription drug coverage for treatment of low back pain among US medicaid, medicare advantage, and commercial insurers. JAMA Netw Open 2018;1(2):e180235 [FREE Full text] [doi: 10.1001/jamanetworkopen.2018.0235] [Medline: 30646077]
28. Hooten WM. Chronic pain and mental health disorders: shared neural mechanisms, epidemiology, and treatment. Mayo Clin Proc 2016;91(7):955-970. [doi: 10.1016/j.mayocp.2016.04.029] [Medline: 27344405]
29. Jones CM, McCance-Katz EF. Co-occurring substance use and mental disorders among adults with opioid use disorder. Drug Alcohol Depend 2019;197:78-82. [doi: 10.1016/j.drugalcdep.2018.12.030] [Medline: 30784952]
30. Freda PJ, Moore JH, Kranzler HR. The phenomics and genetics of addictive and affective comorbidity in opioid use disorder. Drug Alcohol Depend 2021;221:108602 [FREE Full text] [doi: 10.1016/j.drugalcdep.2021.108602] [Medline: 33652377]
31. Wallace PJ, Shah ND, Dennen T, Bleicher PA, Crown WH. Optum Labs: building a novel node in the learning health care system. Health Aff (Millwood) 2014;33(7):1187-1194. [doi: 10.1377/hlthaff.2014.0038] [Medline: 25006145]
32. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. Int J Surg 2014;12(12):1495-1469 [FREE Full text] [doi: 10.1016/j.ijsu.2014.07.013] [Medline: 25046131]
33. Arizona Health Care Cost Containment System (AHCCS). 320, Services with Special Circumstances. 2019. URL: https://www.azahcccs.gov/shared/MedicalPolicyManual/ [accessed 2021-07-25]
34. Elixhauser A, Steiner C, Harris DR, Coffey RM. Comorbidity measures for use with administrative data. Med Care 1998;36(1):8-27. [doi: 10.1097/00005650-199801000-00004] [Medline: 9431328]
35. Quan H, Sundararajan V, Halfon P, Fong A, Burnand B, Luthi JC, et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. Med Care 2005;43(11):1130-1139. [doi: 10.1097/01.mlr.0000182534.19832.83] [Medline: 16224307]
36. Berenson RA, Braid-Forbes MJ. Development and structure of BETOS 2.0 with illustrative data. 2020. URL: https://www.urban.org/research/publication/development-and-structure-betos-20-illustrative-data [accessed 2025-05-10]
37. Riley GF, Lubitz JD, Zhang N. Patterns of health care and disability for medicare beneficiaries under 65. Inquiry 2003;40(1):71-83 [FREE Full text] [doi: 10.5034/inquiryjrnl_40.1.71] [Medline: 12836909]
38. Alorfi NM. Pharmacological methods of pain management: narrative review of medication used. Int J Gen Med 2023;16:3247-3256 [FREE Full text] [doi: 10.2147/IJGM.S419239] [Medline: 37546242]
39. Finley CR, Chan DS, Garrison S, Korownyk C, Kolber MR, Campbell S, et al. What are the most common conditions in primary care? Systematic review. Can Fam Physician 2018;64(11):832-840 [FREE Full text] [Medline: 30429181]
40. Savitz ST, Stevens MA, Nath B, D'Onofrio G, Melnick ER, Jeffery MM. Trends in the prescribing of buprenorphine for opioid use disorder, 2019-2023. Mayo Clin Proc Innov Qual Outcomes 2024;8(3):308-320 [FREE Full text] [doi: 10.1016/j.mayocpiqo.2024.04.004] [Medline: 38841599]
41. Chapman BL, Liebert RB, Lininger MR, Groth JJ. An introduction to physical therapy modalities. Adolesc Med State Art Rev 2007;18(1):11-23, vii. [Medline: 18605388]
42. Groenewald CB, Murray CB, Battaglia M, Scaini S, Quinn PD. Prevalence of pain management techniques among adults with chronic pain in the United States, 2019. JAMA Netw Open 2022;5(2):e2146697 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.46697] [Medline: 35129599]
43. Herring AA, Perrone J, Nelson LS. Managing opioid withdrawal in the emergency department with buprenorphine. Ann Emerg Med 2019;73(5):481-487. [doi: 10.1016/j.annemergmed.2018.11.032] [Medline: 30616926]
44. Dimick JB, Ryan AM. Methods for evaluating changes in health care policy: the difference-in-differences approach. JAMA 2014;312(22):2401-2402. [doi: 10.1001/jama.2014.16153] [Medline: 25490331]
45. Wooldridge JM. Econometric analysis of cross section and panel data. Cambridge, MA: MIT press; 2010.

XSL•FO
RenderX

46. Hutchinson MK, Holtman MC. Analysis of count data using poisson regression. Res Nurs Health 2005;28(5):408-418. [doi: 10.1002/nur.20093] [Medline: 16163676]
47. Counts N. Understanding the US behavioral health workforce shortage. The Commonwealth Fund. 2023. URL: https://www.commonwealthfund.org/publications/explainer/2023/may/understanding-us-behavioral-health-workforce-shortage [accessed 2023-05-18]
48. Langabeer JR, Stotts AL, Bobrow BJ, Wang HE, Chambers KA, Yatsco AJ, et al. Prevalence and charges of opioid-related visits to U.S. emergency departments. Drug Alcohol Depend 2021;221:108568. [doi: 10.1016/j.drugalcdep.2021.108568] [Medline: 33578297]
49. Solomon MD, McNulty EJ, Rana JS, Leong TK, Lee C, Sung SH, et al. The Covid-19 pandemic and the incidence of acute myocardial infarction. N Engl J Med 2020;383(7):691-693. [doi: 10.1056/nejmc2015630]
50. Faust JS, Du C, Mayes KD, Li SX, Lin Z, Barnett ML, et al. Mortality from drug overdoses, homicides, unintentional injuries, motor vehicle crashes, and suicides during the pandemic, march-august 2020. JAMA 2021;326(1):84-86 [FREE Full text] [doi: 10.1001/jama.2021.8012] [Medline: 34019096]
51. Currie JM, Schnell MK, Schwandt H, Zhang J. Prescribing of opioid analgesics and buprenorphine for opioid use disorder during the COVID-19 pandemic. JAMA Netw Open 2021;4(4):e216147 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.6147] [Medline: 33856474]
52. Sun E, Moshfegh J, Rishel CA, Cook CE, Goode AP, George SZ. Association of early physical therapy with long-term opioid use among opioid-naive patients with musculoskeletal pain. JAMA Netw Open 2018;1(8):e185909 [FREE Full text] [doi: 10.1001/jamanetworkopen.2018.5909] [Medline: 30646297]
53. Ahmad FB, Rossen LM, Sutton P. Provisional drug overdose death counts. CDC. 2022. URL: https://www.cdc.gov/nchs/nvss/vsrr/drug-overdose-data.htm [accessed 2025-05-07]
54. DEA, SAMHSA extend COVID-19 telemedicine flexibilities for prescribing controlled medications for six months while considering comments from the public. American Academy of Child and Adolescent Psychiatry. 2023. URL: https://www.aacap.org/AACAP/zLatest_News/DEA_SAMHSA_Extend_COVID-19_Telemedicine_Flexibilities.aspx [accessed 2023-05-09]
55. Second temporary extension of COVID-19 telemedicine flexibilities for prescription of controlled medications. Federal Register. 2023. URL: https://www.federalregister.gov/documents/2023/10/10/2023-22406/second-temporary-extension-of-covid-19-telemedicine-flexibilities-for-prescription-of-controlled [accessed 2023-10-10]
56. Hubley S, Lynch SB, Schneck C, Thomas M, Shore J. Review of key telepsychiatry outcomes. World J Psychiatry 2016;6(2):269-282 [FREE Full text] [doi: 10.5498/wjp.v6.i2.269] [Medline: 27354970]
57. Pierce RP, Stevermer JJ. Disparities in the use of telehealth at the onset of the COVID-19 public health emergency. J Telemed Telecare 2020;29(1):3-9. [doi: 10.1177/1357633x20963893]
58. Mehrotra A, Huskamp HA, Souza J, Uscher-Pines L, Rose S, Landon BE, et al. Rapid growth in mental health telemedicine use among rural medicare beneficiaries, wide variation across states. Health Aff (Millwood) 2017;36(5):909-917. [doi: 10.1377/hlthaff.2016.1461] [Medline: 28461359]
59. Bashshur RL, Shannon GW, Bashshur N, Yellowlees PM. The empirical evidence for telemedicine interventions in mental disorders. Telemed J E Health 2016;22(2):87-113 [FREE Full text] [doi: 10.1089/tmj.2015.0206] [Medline: 26624248]
60. Zahnd WE, Bell N, Larson AE. Geographic, racial/ethnic, and socioeconomic inequities in broadband access. J Rural Health 2022;38(3):519-526. [doi: 10.1111/jrh.12635] [Medline: 34792815]
61. Neter E, Brainin E. eHealth literacy: extending the digital divide to the realm of health information. J Med Internet Res 2012;14(1):e19 [FREE Full text] [doi: 10.2196/jmir.1619] [Medline: 22357448]
62. Dowell D, Brown S, Gyawali S, Hoenig J, Ko J, Mikosz C, et al. Treatment for opioid use disorder: population estimates - United States, 2022. MMWR Morb Mortal Wkly Rep 2024;73(25):567-574 [FREE Full text] [doi: 10.15585/mmwr.mm7325a1] [Medline: 38935567]
63. Thomas KC, Ellis AR, Konrad TR, Holzer CE, Morrissey JP. County-level estimates of mental health professional shortage in the United States. Psychiatr Serv 2009;60(10):1323-1328. [doi: 10.1176/ps.2009.60.10.1323] [Medline: 19797371]
64. Ye S, Kronish I, Fleck E, Fleischut P, Homma S, Masini D, et al. Telemedicine expansion during the COVID-19 pandemic and the potential for technology-driven disparities. J Gen Intern Med 2021;36(1):256-258 [FREE Full text] [doi: 10.1007/s11606-020-06322-y] [Medline: 33105000]
65. Khoong EC. Policy considerations to ensure telemedicine equity. Health Aff (Millwood) 2022;41(5):643-646. [doi: 10.1377/hlthaff.2022.00300] [Medline: 35500190]
66. Third Temporary Extension of COVID-19 Telemedicine Flexibilities for Prescription of Controlled Medication,. Federal Register. URL: https://tinyurl.com/4pchsnej [accessed 2024-11-19]
67. Advocacy issue: telehealth waivers. American Hospital Association. 2024. URL: https://www.aha.org/advocacy/advocacy-issues/2024-10-31-advocacy-issue-telehealth-waivers [accessed 2024-10-31]
68. H.R.10545 - American Relief Act, 2025. CONGRESS. URL: https://www.congress.gov/bill/118th-congress/house-bill/10545 [accessed 2025-05-07]
69. Physician acceptance of new medicaid patients: findings from the National Electronic Health Records Survey. MACPAC. 2021. URL: https://www.macpac.gov/wp-content/uploads/2021/06/

XSL•FO
RenderX

Physician-Acceptance-of-New-Medicaid-Patients-Findings-from-the-National-Electronic-Health-Records-Survey.pdf [accessed 2025-05-07]
70. Mark TL, Yee T, Levit KR, Camacho-Cook J, Cutler E, Carroll CD. Insurance financing increased for mental health conditions but not for substance use disorders, 1986-2014. Health Aff (Millwood) 2016;35(6):958-965. [doi: 10.1377/hlthaff.2016.0002] [Medline: 27269010]
71. Saunders H, Rudowitz R. Demographics and health insurance coverage of nonelderly adults with mental illness and substance use disorders in 2020. KFF. 2022. URL: https://www.kff.org/mental-health/issue-brief/demographics-and-health-insurance-coverage-of-nonelderly-adults-with-mental-illness-and-substance-use-disorders-in-2020/ [accessed 2022-06-06]

Abbreviations

BETOS: Berenson-Eggers Types of Service Codes
CARES: Coronavirus Aid, Relief, and Economic Security
CPT: Current Procedural Terminology
DEA: Drug Enforcement Agency
DiD: difference-in-differences
ED: emergency department
IRR: incident rate ratio
MA: Medicare Advantage
OLDW: OptumLabs Data Warehouse
OUD: opioid use disorder
SMI: serious mental illness
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by A Mavragani; submitted 17.09.24; peer-reviewed by P Freda, J Cantor; comments to author 22.11.24; revised version received 13.01.25; accepted 30.04.25; published 19.06.25.

Please cite as:
Savitz ST, Stevens MA, Nath B, D'Onofrio G, Melnick ER, Jeffery MM
Health Services Usage in Patients Receiving Buprenorphine for Opioid Use Disorder or Long-Term Opioid Therapy for Chronic Pain: Retrospective Cohort Study
JMIR Form Res 2025;9:e66596
URL: https://formative.jmir.org/2025/1/e66596
doi:10.2196/66596
PMID:

©Samuel T Savitz, Maria A Stevens, Bidisha Nath, Gail D'Onofrio, Edward R Melnick, Molly M Jeffery. Originally published in JMIR Formative Research (https://formative.jmir.org), 19.06.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

User Experiences With Digital Future-Self Interventions in the Contexts of Smoking and Physical Inactivity: Mixed Methods Multistudy Exploration

Kristell M Penfornis[1], MSc; Nele Albers[2], PhD; Willem-Paul Brinkman[2], PhD; Mark A Neerincx[2,3], PhD; Andrea WM Evers[1], PhD; Winifred A Gebhardt[1], PhD; Eline Meijer[4], PhD

[1]Unit Health-, Medical and Neuropsychology, Institute of Psychology, Leiden University, Leiden, The Netherlands

[2]Department of Intelligent Systems, Delft University of Technology, Delft, The Netherlands

[3]Department of Perceptual and Cognitive Systems, Nederlandse Organisatie voor Toegepast Natuurwetenschappelijk Onderzoek, Soesterberg, The Netherlands

[4]Department of Public Health and Primary Care, Leiden University Medical Center, Leiden, The Netherlands

Corresponding Author:
Kristell M Penfornis, MSc
Unit Health-, Medical and Neuropsychology
Institute of Psychology
Leiden University
Wassenaarseweg 52
Leiden, 2333 AK
The Netherlands
Phone: 31 71 527 5081
Email: k.m.penfornis@fsw.leidenuniv.nl

Abstract

Background: Smoking and physical inactivity compromise health, especially in combination. Interventions to promote smoking cessation and increased physical activity (PA) often lack impact, especially in the long term. Digital future-self interventions (FSIs), which prompt individuals to imagine who they do and do not want to become (ie, their desired and undesired future selves), show promise in encouraging sustainable changes in both behaviors. However, knowledge of user experiences with digital FSIs is limited. A deeper understanding of these experiences could help optimize FSIs, enhancing their efficacy in supporting smoking cessation and increased PA sustainably.

Objective: This study examined behavioral, cognitive, and affective experiences with digital FSIs focused on smoking, PA, or both. Potential differences in user experiences based on behavior (smoking vs PA), polarity (desired vs undesired future self), and modality (verbal vs visual description of future selves) were explored.

Methods: Secondary analyses of quantitative and qualitative survey data from 3 studies using digital FSIs as a means to encourage smoking cessation or increase PA were conducted. In study 1, participants (N=144) thought about how it would be to complete the FSI. In studies 2 (N=447) and 3 (N=87), they completed an FSI. Each study highlighted different aspects of user experiences with FSIs, namely, behavioral (eg, time spent), cognitive (eg, mental effort exerted), or affective (eg, emotions) experiences. Quantitative and qualitative findings were integrated for a comprehensive interpretation.

Results: Regarding behavioral experiences, participants completed future-self tasks promptly (mean 6.64, SD 8.30 minutes), spent less time completing the desired- versus undesired-future-self ($P<.001$; $\eta_p^2=0.227$) and verbal versus visual ($P=.03$; $\eta_p^2=0.060$; quantitative) tasks, and integrated the tasks into their lives (qualitative). Despite tasks being preparatory and not actively encouraging behavior change, multiple participants reported implementing changes in their smoking or PA (qualitative). Regarding cognitive experiences, moderate effort (mean 5.85/10, SD 2.56) was exerted on the tasks regardless of behavior ($P=.69$; $\eta_p^2=0.002$), modality ($P=.45$; $\eta_p^2=0.004$), or polarity ($P=.69$; $\eta_p^2=0.002$; quantitative). Experiences of task difficulty were inconsistent across studies, individuals, and tasks, although mental visualization and describing one's future self using images were consistently reported as challenging (quantitative and qualitative). Future-self tasks were reported to prompt cognitive processes such as contemplating consequences of smoking and PA behavior (qualitative). Regarding affective experiences, desired- and undesired-future-self tasks elicited different emotions ($P<.001$; $\eta_p^2=0.630$; quantitative). Desired-future-self tasks

XSL•FO
RenderX

were perceived as enjoyable and happiness inducing, whereas undesired-future-self tasks were perceived as confronting and unpleasant, evoking feelings of sadness, fear, and anger (quantitative and qualitative).

Conclusions: Digital FSIs appeared to be a time-efficient, feasible, and acceptable way of strengthening identities as a means to encourage smoking cessation and PA. Findings support continued implementation of digital FSIs, although further research is required to optimize their operationalization. Avenues in that regard are proposed and discussed.

(JMIR Form Res 2025;9:e63893) doi:10.2196/63893

KEYWORDS

smoking; physical activity; future-self intervention; user experiences; multistudy; mixed methods

Introduction

Background

Smoking is the leading behavioral risk factor for disease and premature death worldwide, and physical inactivity is the tenth [1]. Annually, tobacco is responsible for approximately 8 million deaths worldwide, and physical inactivity is responsible for approximately 1 million [2,3]. Combined, smoking and physical inactivity more than double the likelihood of premature death [4] and decrease disease-free life years by 6 years [5]. Consequently, interventions targeting both smoking cessation and physical activity (PA) have great potential to reduce worldwide mortality and morbidity, with multi-behavior interventions potentially enhancing effectiveness through synergistic effects [6,7]. Digital platforms such as eHealth and mobile health further enable scalable, accessible support [8,9].

Quitting smoking and increasing PA immediately mitigate health risks even among older individuals and those with long-standing histories of smoking or sedentary behavior [10,11]. However, quitting smoking and increasing PA are often challenging. To illustrate, long-term smoking abstinence may require 20 to 30 attempts [12]. While numerous smoking cessation and PA promotion interventions exist, their effects tend to be small to moderate and are rarely sustained beyond a year [13-15]. This leaves individuals vulnerable to relapses and renewed risks of negative health outcomes.

Identity-focused health behavior change interventions, which aim to connect healthy behavior to central components of who one is (ie, one's self-identity), present a promising avenue to enhance the effectiveness of smoking cessation and PA promotion programs. Such interventions are rooted in identity theories that propose that, as people prefer to act in line with their self-identity [16,17], one is more likely to abstain from smoking or engage in regular PA when such healthy behavior is integrated into one's self-identity. Empirical research supports this—individuals identifying with smoking show weaker intentions to quit [18,19] and are less likely to attempt cessation [20-22], whereas those identifying as quitters or nonsmokers are more likely to attempt [20,22] and achieve cessation [18]. Similarly, identifying with PA is linked to more intensive and frequent PA [23-26].

Among identity-focused interventions, future-self interventions (FSIs) appear particularly promising. FSIs are rooted in possible self–related theories, one branch of identity theory, which posit that possible future identities and not only current identity influence current behavior [27,28]. A clear vision of one's desired or undesired future self can motivate self-regulatory processes to become the desired future self [28,29]. This is why FSIs generally prompt people to imagine who they want to become (ie, desired future self) or who they want to avoid becoming (ie, undesired future self). Research recommends envisioning both desired and undesired future selves for greater effect on identity and, subsequently, behavior [29,30]. Common tasks include writing about the envisioned future selves (ie, verbal future-self tasks), and searching for images that describe them (ie, visual future-self tasks) is also common [20,31]. Such tasks help create a more vivid image of the future self, ultimately increasing the likelihood of behaving in ways confirming or disconfirming that possible future identity [31]. FSIs are increasingly being delivered digitally [20,32].

Despite a solid theoretical foundation, experimental studies using FSIs in the context of smoking and PA have shown mixed results. While FSIs have repeatedly promoted PA [31,33-35] and impacted smoking cessation–related outcomes (eg, lower craving intensity, greater quit intention, and smoking reduction and abstinence [36-38]), some studies have found no effect on smoking [20,32] or PA [39]. The effectiveness of digital health behavior change interventions appears to hinge on user engagement [40]. Understanding user experiences is essential to determine the ideal type and level of engagement [41]. However, beyond one study reporting FSIs as useful to reduce smoking [37] and 2 reporting that future-self imagery can influence affect [42,43], little is known about user experiences with FSIs. Therefore, understanding user experiences is crucial for optimizing FSIs to facilitate smoking cessation and PA promotion.

Objectives

Examination of user experiences is most complete when considering behavioral (eg, use and time spent), cognitive (eg, [mental] effort, interest, and attention), and affective (eg, emotional response) experiences [44-47]. Relying on a single study risks offering a narrow or skewed viewpoint influenced by the specific FSI examined. Therefore, this multistudy paper triangulates behavioral, cognitive, and affective user experiences across several digital FSIs aimed at facilitating smoking cessation or PA promotion to offer a more comprehensive overview.

The 3 digital FSIs in this study asked people to envision smoking- or PA-related desired and undesired future selves and complete verbal and/or visual tasks about these self-views. Given that smoking is a health-compromising behavior that is preferably quit and PA is a health-promoting behavior that is

XSL•FO
RenderX

preferably increased, user experiences with FSIs may differ by behavior. Therefore, we explored potential differences in user experiences by behavior (smoking vs PA). As examining desired and undesired future selves jointly may mask varying experiences with one versus the other, we also explored differences in user experiences by task polarity (desired- vs undesired-future-self task). Third, because individuals may experience verbal and visual tasks differently, we explored whether user experiences vary by modality (visual vs verbal future-self task). Finally, to gain full insight into user experiences with FSIs, we explored interactions among behavior, polarity, and/or modality and also explored free-text responses.

Methods

Overview

This multistudy report is an observational, exploratory analysis of secondary quantitative and qualitative survey data from 3 empirical studies that used distinct digital FSIs. Combining quantitative and qualitative methods from multiple studies enabled data triangulation across contexts while providing deeper, more nuanced insights into individuals' experiences [47-49]. The multistudy approach enhanced comprehensive understanding of experiences with FSIs and reduced the likelihood of chance effects. This study followed the Mixed Methods Article Reporting Standards. The 3 studies are, from this point on, referred to as study 1, study 2, and study 3. Personal characteristics of the participants included in the 3 studies are presented in Table 1.

The characteristics of the 3 original studies and outcomes used in this multistudy report to explore user experiences are presented in Table 2. Each study provides complementary insights into user experiences compared to the previous ones. Study 1 primarily explored cognitive experiences with a smoking- and PA-related desired- and undesired-FSI. Study 2 primarily explored cognitive and behavioral experiences with a smoking- and PA-related desired- and undesired-FSI and provided some insights into affective experiences as well. Study 3 explored behavioral, cognitive, and affective experiences with a smoking-related desired- and undesired-FSI. Collectively, the 3 studies offer a comprehensive understanding of experiences with FSIs.

Table 1. Personal characteristics of the participants included in studies 1, 2, and 3.

Characteristic	Study 1 (N=144)	Study 2 (N=447)	Study 3 (N=87)
Age (y), mean (SD; range)	37.55 (11.76; 20-69)	36.25 (11.44; 19-71)	37.84 (19.22; 18-82)
Gender[a], n (%)			
Men	72 (50)	211 (47.2)	27 (31)
Women	72 (50)	222 (49.7)	60 (69)
Nonbinary or other	0 (0)	14 (3.1)	0 (0)
Socioeconomic position[b], n (%)			
Do not know	__[c]	3 (0.7)	—
Lower	1 (0.7)	2 (0.4)	13 (14.9)
Middle	32 (22.2)	126 (28.2)	45 (51.7)
Higher	111 (77.1)	316 (70.7)	29 (33.3)
Cigarette or e-cigarette consumption[d], n (%)			
Once per day	5 (3.5)	25 (5.6)	1 (1.1)
2-5 times per day	97 (67.4)	73 (16.3)	28 (32.2)
6-10 times per day	48 (33.3)	93 (20.8)	21 (24.1)
11-19 times per day	45 (31.3)	104 (23.3)	13 (14.9)
≥20 times per day	25 (17.4)	152 (34)	24 (27.6)
Weekly exercise[e], n (%)			
Never to little	30 (20.8)	124 (27.7)	—
Sometimes	80 (55.6)	193 (43.2)	—
Often	34 (23.6)	129 (28.9)	—

[a]For studies 1 and 2, the "men" and "women" categories include transgender individuals.

[b]Lower: no formal education, primary education, and high school or equivalent; middle: tertiary education; higher: technical or community college or undergraduate, graduate, or doctoral degree as per the International Standard Classification of Education [50].

[c]This answer option was not provided in the study.

[d]For study 2, the numbers reflect smoking and vaping together.

[e]Never to little: 0 to 60 minutes per week; sometimes: 60 to 150 minutes per week; often: >150 minutes per week.

XSL•FO
RenderX

Table 2. Characteristics of the 3 original studies and outcome variables used to explore behavioral, cognitive, or affective user experiences with the corresponding future-self interventions.

	Study 1: cognitive experiences with a future-self intervention	Study 2: behavioral and cognitive experiences with a future-self intervention	Study 3: behavioral, cognitive, and affective experiences with a future-self intervention
Design, measurements, and future-self intervention			
Study design	• Online observational cross-sectional study	• Online longitudinal observational study	• Online longitudinal experimental study
Data collection mode	• Online survey	• Online survey	• Online survey
Type of data	• Quantitative and qualitative	• Quantitative and qualitative	• Quantitative and qualitative
Measurement number and moments	• 1, during the cross-sectional study survey	• 5, once per week for the duration of the intervention	• 1, during the postintervention questionnaire completed directly after the intervention
Survey language	• English	• English	• Dutch
Studied health behaviors	• Smoking and PA[a]	• Smoking and PA	• Smoking
Future-self intervention	• Participants (N=144) thought about completing digital verbal and visual future-self tasks regarding their desired and undesired smoking- and PA-related future selves	• Participants (N=447) completed 1 to 3 digital verbal or visual future-self tasks regarding their desired and undesired smoking- and PA-related future selves	• Participants randomized to the intervention condition (n=87) completed both a digital verbal and visual future-self task regarding their desired and undesired smoking-related future selves
Outcomes reflecting behavioral, cognitive, and affective user experiences			
Behavioral experiences	—[b]	• Quantitative: — • Qualitative: how future-self tasks were approached, done, or experienced[c]	• Quantitative: time spent on future-self tasks • Qualitative: what it was like to perform the future-self tasks[d]
Cognitive experiences	• Quantitative: anticipated difficulty of future-self tasks and anticipated completion time of future-self tasks • Qualitative: what was anticipated to make tasks more difficult or easier than others	• Quantitative: mental effort deployed on future-self tasks • Qualitative: how future-self tasks were approached, done, or experienced[c]	• Quantitative: experienced difficulty of future-self tasks • Qualitative: what it was like to perform the future-self tasks[d]
Affective experiences	—	• Quantitative: — • Qualitative: how future-self tasks were approached, done, or experienced[c]	• Quantitative: emotional response to future-self tasks • Qualitative: what it was like to perform the future-self tasks[d]

[a]PA: physical activity.

[b]Not applicable.

[c]The "do" part of the question was expected to probe participants to share their behavioral experiences with the future-self tasks, the "approach" part of the question was expected to probe participants to share their behavioral and cognitive experiences, and the "experience" part of the question was expected to probe participants to share their affective experiences.

[d]This question was expected to probe participants to share their behavioral, cognitive, and affective experiences with the future-self tasks. Deductive coding during data analysis allowed for the determination of which answer informed which experiential dimension.

Data Analysis

For studies 1, 2, and 3, quantitative analyses were conducted in SPSS Statistics for Windows (version 29; IBM Corp), with statistical significance set at $P<.05$ (2-tailed). Bonferroni corrections were applied for multiple testing [51]. Effect sizes were interpreted according to the Cohen d (small≥0.2, medium≥0.5, and large≥0.8) for t tests and the Cohen f (small≥0.1, medium≥0.25, and large≥0.4) for ANOVAs [52]. The syntaxes are publicly available [53].

Qualitative survey data were analyzed in Microsoft Excel using qualitative content analysis principles [54]. Two researchers (KMP and a trained psychology master student) independently familiarized themselves with the responses, developed initial

XSL•FO
RenderX

inductive codes, compared these, and created a coding tree. Participant responses were then coded jointly, and the coding trees were refined as needed. One researcher (KMP) subsequently created a matrix indexing all codes and counted their frequency for a weighted data representation. Finally, the findings were categorized into behavioral, cognitive, or affective experiences with future-self tasks. Specifications are detailed in the respective Data Analysis sections. Reported coding frequencies indicate how often the code appeared, not the number of participants to whom it applied.

Study 1: Cognitive Experiences With a PA- and Smoking-Related FSI

In study 1, quantitative and qualitative survey data from the original online cross-sectional observational study [55] were used to explore cognitive experiences with a smoking- and PA-related FSI. Additional information regarding the original study can be found in the preregistration on the Open Science Framework [56], and the data related to this study are openly accessible elsewhere [57].

Methods

Participants

A total of 144 participants were recruited via the online recruitment platform Prolific [58] between September 2022 and November 2022. Participants were from countries that belonged to the Organisation for Economic Co-operation and Development, excluding Turkey, Lithuania, Colombia, and Costa Rica but including South Africa [59]. Inclusion criteria required being aged 18 years, fluency in English, daily smoking, and an intention to quit within 30 days to 6 months. The original study was primarily focused on smoking cessation, which is why an inclusion criterion was formulated regarding smoking but not PA. Those familiar with similar tasks from a previous study [60] were excluded. Table 1 provides the participant characteristics.

Procedure

In total, 2 online surveys were administered using Qualtrics XM (Qualtrics International Inc) [61]. Interested Prolific members provided digital informed consent and completed a screening survey. Upon meeting the inclusion criteria, participants proceeded to the study survey, which included instructions for 44 different tasks that either aimed to aid smoking cessation or PA enhancement or could be beneficial to either behavior. This multistudy report focuses only on the 8 tasks targeted at one's smoking- or PA-related future self, involving writing about or searching images describing the desired or undesired future self. Detailed instructions are provided in Multimedia Appendix 1 [55,60]. Participants were presented with all 44 preparatory tasks in a random order and provided feedback, which required approximately 40 minutes.

Ethical Considerations

Ethics approval for this study was granted by the Human Research Ethics Committee of Delft University of Technology (letter of approval 2338). Digital consent from participants in the original study enabled secondary analyses without additional approval. Participant data were deidentified for analysis to ensure confidentiality and privacy. Participants were compensated with £6.00 (US $7.69) per hour following Prolific regulations [62].

Measures

Personal Characteristics

These were collected from Prolific participant profiles and the screening survey and used to describe the participants. Variables included *age* (derived from year of birth), *gender* (male, including transgender male; female, including transgender female; or other), and highest completed educational level. The latter was used as an indicator of *socioeconomic position* [63] and recoded according to the International Standard Classification of Education [50]. *Smoking frequency* was assessed using the following question: "How often do you smoke tobacco products?" Participants indicating not smoking daily were informed that they could not continue taking part in the study. One question ascertained participants' *weekly exercise* (ie, "How often do you engage in physical exercise per week?").

Cognitive Experiences

In total, 3 items from the study survey were used to explore cognitive experiences with the FSI. Participants rated the *anticipateddifficulty* of the 8 future-self tasks on a scale in which −5=*very difficult*, 0=*neutral*, and 5=*very easy*. Second, they provided an *estimatedcompletion time* ranging from 0 to 30 minutes, indicating exact minutes if the anticipated time exceeded 30 minutes. Third, an open question, "Think of the preparatory activities you have just seen. What makes an activity more difficult than others?" was expected to probe participants to share their cognitive experiences with future-self tasks.

Data Analysis

After computing descriptive statistics, a 1-sample *t* test examined whether anticipated difficulty scores differed from the neutral 0 point. One- and 3-way repeated-measure (RM) ANOVAs explored differences in anticipated difficulty and anticipated completion time based on behavior (smoking vs PA), polarity (desired vs undesired future self), and modality (verbal vs visual description) of the future-self tasks. Assumptions were verified (Multimedia Appendix 2).

Qualitative responses ranged from 1 word to 3 sentences. All participant responses (N=144) were coded. The full coding tree can be found in Multimedia Appendix 3.

Results

Quantitative Results

Overview

The main findings of study 1 are summarized in Table 3.

XSL•FO
RenderX

Table 3. Summary of quantitative and qualitative findings regarding behavioral, cognitive, and affective experiences with smoking- or physical activity (PA)–related future-self tasks from studies 1, 2, and 3.

	Study 1	Study 2	Study 3
Behavioral experiences			
Quantitative findings	__[a]	—	• Future-self tasks were completed in <7 minutes. • Participants spent significantly more time on the verbal than on the visual task and on the desired- than on the undesired-future-self task.
Qualitative findings	—	• Participants successfully completed the future-self tasks and integrated them into daily life. • Frequency, duration, time of day, and location were adapted to preferences and lifestyle. • While the tasks were aimed at preparing for behavior change, multiple participants reported having changed their PA and smoking behavior or having formulated goals and action plans to do so.	—
Cognitive experiences			
Quantitative findings	• Future-self tasks were anticipated to be completed relatively easily and in <14 minutes. • Verbal description of future selves was anticipated to take more time than visual description. • It was anticipated that desired tasks, smoking-related tasks, and verbal descriptions would be easier than undesired tasks, PA-related tasks, and visual descriptions, respectively	• Participants deployed slightly above-average effort on the future-self tasks. • There were no differences in effort deployed on the future-self tasks depending on behavior, modality, or polarity.	• Future-self tasks were experienced as relatively difficult. • Visual description of future selves was experienced as more difficult than simply envisioning them.
Qualitative findings	• Tasks consisting of multiple components (eg, visualizing and writing) and requiring more time, effort, and mental visualization, particularly of one's undesired future self, were anticipated to be more difficult. • Opinions differed on whether desired- or undesired-future-self tasks were considered more difficult. • Greater difficulty was anticipated to be linked to task unfamiliarity, motivation to perform the tasks, degree of self-honesty about the consequences of current behavior, attainability of the future selves, negative emotions, and finding it easier to take concrete steps to change behavior.	• Some found the tasks difficult because of the mental visualization. Some found them easy because they were familiar with the tasks. • The tasks appeared to trigger numerous cognitive processes, including contemplating the consequences of current and changed behavior; identifying role models; altering motivation to change; and mentally contrasting past, present, and future selves.	• The desired-future-self task was seen as almost equally easy and difficult, and the undesired-future-self task was seen as (very) difficult, with individual variations. • Unfamiliarity with envisioning one's future self was reported to hinder task completion, but familiarity did not always ease it. • The desired-future-self task generally gave something to look forward to and sometimes prompted reflection about reasons for quitting smoking. • The undesired-future-self task illustrated who participants aimed not to become, confronted them with its probability of happening or addiction, and sometimes decreased motivation to quit or prompted negative self-reflection. • Future-self tasks seemed to prompt cognitive processes, including reflection about the consequences of smoking or being a smoker; increased motivation to quit smoking; cognitive dissonance; resistance to future-self thinking; or comparisons between future, current, or past selves.

XSL•FO
RenderX

	Study 1	Study 2	Study 3
Affective experiences			
Quantitative findings	—	—	• Participants generally felt happy during the desired-future-self task and sad, angry, or anxious during the undesired-future-self task.
Qualitative findings	—	• Desired-future-self tasks were reported to elicit positive emotions, whereas undesired-future-self tasks were reported to elicit negative emotions.	• The desired-future-self task was experienced as enjoyable, and the undesired-future-self task was experienced as confronting and unpleasant, although experiences varied per individual.

[a]Not applicable for this study.

Anticipated Difficulty of Future-Self Tasks

Figure 1 shows mean anticipated ease by behavior, modality, and polarity. A 1-sample *t* test showed that anticipated difficulty scores significantly differed from the neutral 0 point, meaning that future-self tasks were anticipated to be relatively easy to complete (mean 1.19, SD 1.75; t_{143}=8.20; P<.001; Cohen d=0.68). RM ANOVAs assessed differences in anticipated difficulty based on behavior, polarity, and modality. There was a significant 3-way interaction among behavior, polarity, and modality on anticipated difficulty of the future-self tasks ($F_{1,143}$=4.26; P=.04; η_p^2=0.029). Follow-up ANOVAs revealed no significant 2-way interactions or main effects of behavior and modality (lowest P>.01). However, they revealed significant main effects of polarity. Specifically, desired-future-self tasks were anticipated to be easier than undesired-future-self tasks for both smoking ($F_{1,143}$=24.16; P<.001; η_p^2=0.145) and PA ($F_{1,143}$=33.90; P<.001; η_p^2=0.192) and when asked to describe the future self verbally ($F_{1,143}$=23.93; P<.001; η_p^2=0.143) and visually ($F_{1,143}$=30.85; P<.001; η_p^2=0.177).

XSL•FO
RenderX

Figure 1. Mean anticipated difficulty scores and anticipated completion time of physical activity (PA)– and smoking-related future-self tasks by behavior (smoking or PA), polarity (desired or undesired future self), and modality (verbal or visual) in study 1 (N=144).

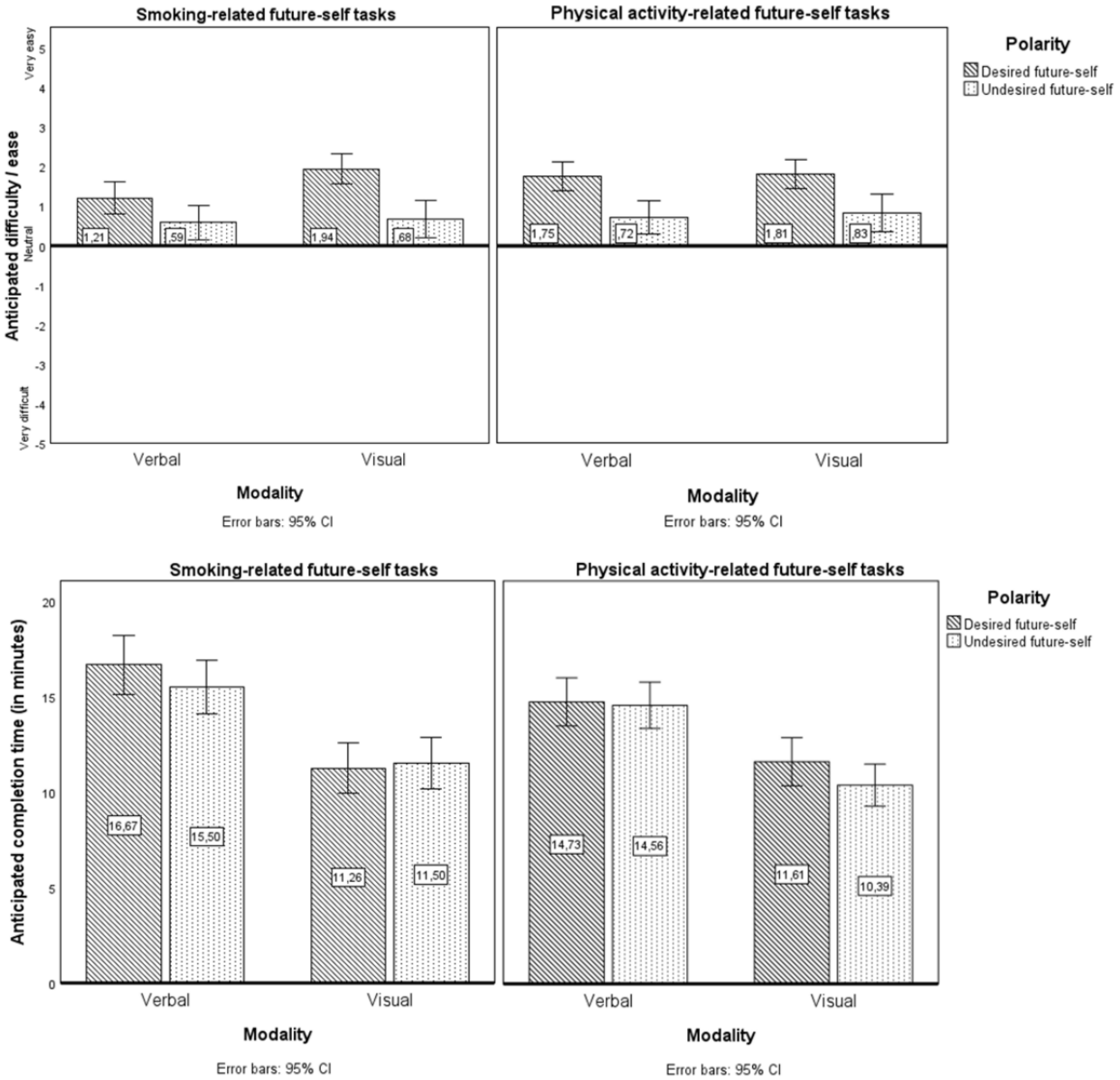

Anticipated Completion Time for Future-Self Tasks

Figure 1 shows mean anticipated completion time by behavior, modality, and polarity. Participants estimated that it would take, on average, 13.28 (SD 5.79) minutes to complete future-self tasks. RM ANOVAs assessed differences in anticipated completion time based on behavior, polarity, and modality. There was no statistically significant 3-way ($F_{1,130}$=3.54; P=.06; η_p^2=0.027) or 2-way interaction for anticipated completion time (P>.05 in all cases), but there were significant main effects of behavior ($F_{1,130}$=11.30; P=.001; η_p^2=0.080) and modality ($F_{1,130}$=78.66; P<.001; η_p^2=0.377). Specifically, the smoking-related future-self task was anticipated to take more time to complete than the PA-related one (Δ=0.91, 95% CI 0.38-1.45). In addition, verbal description of future selves was anticipated to take more time than visual description (Δ=4.17, 95% CI 3.24-5.11).

Qualitative Results

Cognitive Experiences With Future-Self Tasks

Qualitative analysis indicated that tasks were anticipated to be more difficult if they involved mental visualization (25/144, 17.4%), cognitive effort (24/144, 16.7%), more time (19/144, 13.2%), physical effort (14/144, 9.7%), or multiple components (eg, both visualizing and writing about one's future self; 4/144, 2.8%). Six reasons emerged for why visualization tasks were anticipated to be difficult: (1) unfamiliarity with mental visualization (3/144, 2.1%; "Visualization is a concept that is foreign to some people. A course is needed" [female participant; aged 64 years]), (2) motivation to perform the tasks or change behavior (3/144, 2.1%), (3) willingness to be honest about (the consequences of) current behavior (3/144, 2.1%), (4) perceived

XSL•FO
RenderX

attainability of the future self (2/144, 2.1%), (5) expectation of a negative emotional response (2/144, 2.1%), and (6) the challenge of visualizing compared to taking concrete actions to change the behavior (2/144, 2.1%). In total, 0.7% (1/144) of the participants anticipated envisioning a positive future to be easier, whereas 2.1% (3/144) anticipated the undesired-future-self task to be easier:

> *For me, thinking about...positives would be difficult, as I don't often praise myself or congratulate myself or anything like that.* [Female participant; aged 50 years]

Discussion

Regarding cognitive experiences, quantitative results from study 1 showed that future-self tasks were anticipated to be completed relatively easily and in less than 14 minutes, suggesting FSIs can be anticipated to be time efficient and feasible. Desired-future-self tasks were perceived as easier than undesired-future-self tasks. Furthermore, smoking-related future-self tasks and verbal descriptions of future selves were estimated to take more time than PA-related tasks and visual descriptions. These results imply differences in experiences between specific future-self tasks.

Qualitative results indicated that tasks requiring mental visualization, cognitive or physical effort, or more time or tasks having multiple components (eg, visualizing and writing) were anticipated to be difficult. Visualization tasks were seen as challenging due to factors such as unfamiliarity, motivation, and anticipated negative emotional responses. Views varied on whether desired- or undesired-future-self tasks were more challenging in these respects. These findings suggest that, while FSIs are generally considered fairly easy, the mental visualization aspect is viewed as challenging.

Study 2: Behavioral and Cognitive Experiences With PA- or Smoking-Related FSIs

In study 2, quantitative and qualitative survey data from the original online longitudinal observational study [55] were used to explore behavioral; cognitive; and, where possible, affective experiences with a smoking- or PA-related FSI. Additional information regarding the original study can be found in the preregistration on the Open Science Framework [64], and the data underlying this study have been published elsewhere [53].

Methods

Participants

Participants were 52.5% (447/852) of the included individuals, who completed future-self tasks in the original study between February 2024 and March 2024. Inclusion criteria were being aged 18 years, fluency in English, daily smoking or vaping, intention to quit within 30 days to 6 months, not taking part in a smoking or vaping cessation intervention, and not being familiar with similar tasks from previous studies [65,66]. Participants failing attention checks integrated into the survey or not completing tasks within 2 days of having received the invitation were excluded. Dropouts were replaced until the budget of approximately €5000 (US $5475.65) was spent. The personal characteristics of study 2 participants are presented in Table 1.

Procedure

Prolific members who provided digital informed consent and met the inclusion criteria completed a baseline survey via Qualtrics XM, including questions about demographic characteristics. The intervention consisted of up to 5 conversational sessions of 6 to 8 minutes spaced 3 to 5 days apart with coach Kai, a text-based chatbot [67]. During these conversation sessions, participants received instructions for 1 task randomly selected from 37 tasks meant to aid smoking cessation and PA enhancement, of which 8 (22%) were future-self tasks. The tasks were a subset of those in study 1, with more detailed instructions based on the work by Albers et al [60] (Multimedia Appendix 1). In sessions 2 to 5, coach Kai asked questions about the most recent task [68].

Ethical Considerations

Ethics approval was granted by the Human Research Ethics Committee of Delft University of Technology (letter of approval 3683). Digital consent from participants in the original study enabled secondary analyses without additional approval. Participant data were deidentified for analysis to ensure confidentiality and privacy. Participants were compensated with £6.00 (US $7.69) per hour following Prolific regulations [62].

Measures

Personal Characteristics

The same variables as in study 1 were collected from Prolific profiles and the baseline survey with the exception of *smoking or vaping frequency*, which were combined into 1 variable. For vaping, the question was adapted to "How often do you vape?"

Cognitive Experiences

In total, 2 survey items administered at the start of each conversational session with coach Kai were used to explore cognitive experiences with the future-self tasks. In conversational sessions 2 to 5, coach Kai asked participants to rate their *effort* on the assigned future-self task from 0=*nothing* to 10=*extremely strong*. This served as an indication of mental effort. Participants then answered an open question, "How did you approach, do, or experience your assigned activity?" requiring a response of at least 20 characters. After providing a response, participants were asked by coach Kai whether they wanted to add something to or modify their response. The *approach* part of the question was expected to probe participants to share their cognitive experiences with future-self tasks.

Behavioral and Affective Experiences

The *do* and *approach* parts of a question "How did you approach, do, or experience your assigned task?" were expected to probe participants to share their behavioral experiences with future-self tasks, whereas the *experience* part was expected to probe them to share affective experiences.

XSL•FO
RenderX

Data Analysis

Descriptive statistics were computed. Two 1-sample *t* tests examined whether mean effort on the 8 future-self tasks differed from (1) the midpoint of the scale (ie, 5) and (2) mean effort on the other 29 preparatory tasks. The effort score of 0.2% (1/447) of the participants was adjusted from 0 to 5 following self-report in the free-text response field that they had erroneously selected 0. Three-way and follow-up 1-way ANOVAs explored differences in effort based on behavior (smoking, PA, or both), polarity (desired or undesired future self or both), and modality (verbal or visual description or both) of the future-self tasks. Assumptions were verified (Multimedia Appendix 2).

Free-text survey answers were coded until no new codes emerged. This resulted in 17.6% (103/586) of free-text responses from 21.3% (95/447) of the participants. Responses ranged from 1 word to 13 sentences. The full coding tree can be found in Multimedia Appendix 3.

Quantitative Results: Effort Deployed on Future-Self Tasks

Participants completed between 1 and 3 future-self tasks. One-sample *t* tests showed that participants deployed significantly above-average amounts of effort on future-self tasks (mean 5.85, SD 2.56; t_{446}=7.02; P<.001; Cohen d=0.33), akin to the effort deployed on the other preparatory tasks from the same study (t_{446}=0.92; P=.36; Cohen d=0.04). Figure 2 shows mean effort by behavior, polarity, and modality. A 3-way ANOVA assessed possible differences in effort based on behavior, polarity, and modality. There was no significant interaction between behavior, polarity, or modality and effort ($F_{1,424}$=1.04; P=.39; η_p^2=0.010) and no significant main effects of behavior ($F_{1,444}$=0.38; P=.69; η_p^2=0.002), modality ($F_{1,444}$=0.80; P=.45; η_p^2=0.004), or polarity ($F_{1,444}$=0.37; P=.69; η_p^2=0.002).

Results

The main findings of study 2 are summarized in Table 3.

Figure 2. Mean effort deployed on physical activity (PA)– or smoking-related future-self tasks by behavior (smoking, PA, or both), polarity (desired or undesired future self or both), and modality (verbal or visual or both) in study 2 (N=447).

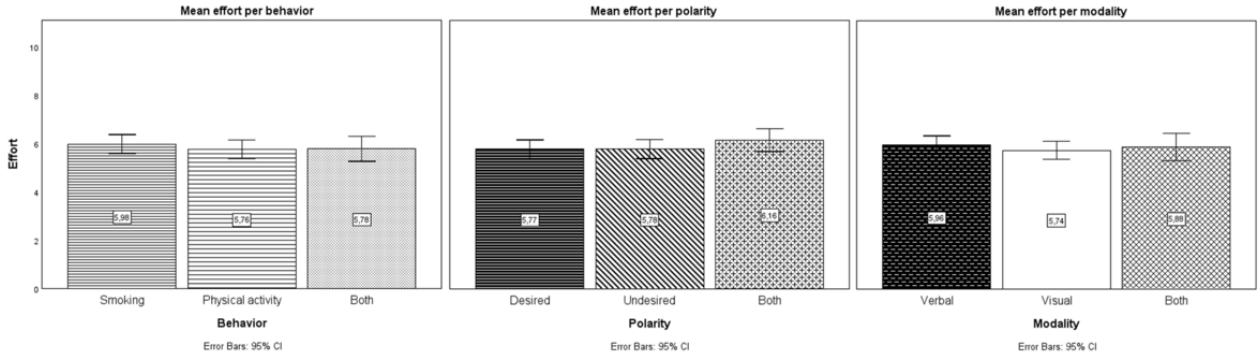

Qualitative Results

Behavioral Experiences With Future-Self Tasks

Participants reported success in writing about their future selves (2/95, 2%) or finding an image representing them (13/95, 14%):

> *I found a picture of a graveyard, and imagined being laid to rest there earlier than I might be.* [Male participant; aged 59 years]

Several participants reported on the frequency (eg, 1-2 times a day or when craving a cigarette), duration (eg, needing several days to form a clear mental image [male participant; aged 31 years]), time of day (eg, before sleep [male participant; aged 23 years]), or conditions (eg, in a relaxed, quiet, distraction-free environment, 3/95, 3%) required for completing their future-self tasks. In total, 3% (3/95) of the participants were unable to complete the tasks due to forgetting, illness, or parental duties. Overall, participants adapted the future-self tasks to their preferences and lifestyles and integrated them into their lives.

The future-self tasks were reported to have various behavioral outcomes. A total of 15% (14/95) of the participants saw instructions as encouraging them to quit smoking or become more physically active:

> *I just made an effort to go for a short walk every day after work to destress.* [Female participant; aged 33 years]

In total, 2% (2/95) of the participants searched for information on smoking's health effects, and others felt invited to formulate goals (11/95, 12%) or an action plan (5/95, 5%) regarding smoking or PA. A total of 8% (8/95) reported the task to be ineffective in changing their PA or smoking behavior:

> *I found that this task to help quit vaping was ineffective for me as I found it difficult to visualize myself in the future.* [Male participant; aged 59 years]

Cognitive Experiences With Future-Self Tasks

There was quite some variation in how participants experienced future-self tasks. In total, 2% (2/95) found them easy, whereas 5% (5/95) found them difficult. Reported reasons for seeing them as difficult included the inability to produce a clear mental image (5/95, 5%), refusal to think about (1/95, 1%) or identify with (1/95, 1%) an undesired future self, not perceiving the need to quit vaping (1/95, 1%), or addiction standing in the way of visualizing a future without smoking (1/95, 1%).

The future-self tasks were reported to trigger various cognitive processes. The most common was thinking about the

XSL•FO
RenderX

consequences of current (14/95, 15%) or changed behavior (9/95, 9%):

> *I took my time to think about the negative things about vaping and be clear about the type of person I don't want to become.* [Male participant; aged 42 years]

The second most common cognitive process was facing one's fears (eg, of disease, premature death, or being a bad romantic partner and example to children, 6/95, 6%). The third most common was identifying positive (3/95, 3%) or negative (3/95, 3%) role models representative of desired or undesired future selves. The fourth most common was comparing desired and undesired future selves (1/95, 1%), desired future or past selves with the undesired current self (1/95, 1%), or past with desired future selves (5/95, 5%; eg, "I looked at pictures of myself when I was at my heaviest weight. I was close to 200 lbs, smoking, feeling like garbage. I don't EVER want to become that person again" [female participant; aged 57 years]). The fifth most common was increased or decreased motivation for changing behavior following the tasks (1/95, 1%).

Affective Experiences With Future-Self Tasks

Tasks related to the undesired future self were experienced as confronting (6/95, 6%), anxiety inducing (2/95, 2%), worrisome (1/95, 1%), or saddening (1/95, 1%). In contrast, desired-future-self tasks felt inspiring (1/95, 1%) and provided something to look forward to (3/95, 3%).

Discussion

Regarding behavioral experiences, qualitative findings from study 2 indicated that participants successfully integrated future-self tasks into their lives by adjusting frequency, duration, timing, and location to suit their lifestyles and preferences. This aligns with study 1, suggesting that FSIs are feasible for smokers intending to quit and for varying levels of PA. Although the tasks aimed to help prepare for behavior change, numerous participants actually changed their PA and smoking behaviors or formulated goals and action plans to do so, suggesting that the tasks were stimulating enough to prompt action.

Regarding cognitive experiences, quantitative findings showed that participants exerted above-average effort on future-self tasks regardless of behavior, polarity, or modality. This implies that tasks were appropriately challenging. Similar to study 1, mental visualization was seen as complicating the tasks, a result formally assessed in study 3. Future-self tasks were perceived as easier with increased familiarity. They also appeared to stimulate various cognitive processes, suggesting that repeated practice may further enhance engagement.

Regarding affective experiences, future-self tasks elicited contrasting emotional responses—negative for undesired-future-self tasks and positive for desired-future-self tasks—highlighting the importance of considering emotional engagement with FSIs.

Study 3: Behavioral, Cognitive, and Affective Experiences With a Smoking-Related FSI

In study 3, quantitative and qualitative survey data from the original online longitudinal experimental study were used to explore behavioral, cognitive, and affective experiences with a smoking-related FSI. Additional information regarding the methods and results of the original study is reported elsewhere [20,69], and the underlying data are openly accessible [70].

Methods

Participants

Participants were 43.3% (87/201) of the individuals included in the original study (ie, those who completed the FSI). Recruited in the Netherlands and Dutch-speaking part of Belgium between July 2017 and October 2018, they were sourced through various media channels (eg, participation in previous research and social media—the full list of recruitment methods is available elsewhere [20]). Inclusion criteria were being aged ≥18 years, smoking daily, and intending to quit smoking sometime in the future. Participant characteristics are presented in Table 1.

Procedure

Data were collected using Qualtrics XM. After providing digital informed consent, participants completed the baseline survey and were randomized 1:1 to the intervention or waitlist control condition. Intervention participants imagined a future in which they successfully quit smoking, wrote about this future self (verbal task), and uploaded images describing it (visual task). Afterward, they completed a postintervention survey, which took approximately 20 minutes. Control participants completed similar tasks about washing their hands more often.

Ethical Considerations

Ethics approval was granted by the Research Ethics Committee of Leiden University's Institute of Psychology (CEP17-0505/192). Digital consent from participants in the original study enabled secondary analyses without additional approval. Participant data were deidentified for analysis to ensure confidentiality and privacy. In total, 2 gift coupons of €100 (US $109.51) and 6 coupons of €50 (US $54.76) were randomly distributed among participants who completed the study.

Measures

Personal Characteristics

Participants indicated their *gender*, *birth year*, and *cigarette consumption*, which was recoded to match smoking frequency in studies 1 and 2. Educational level was used as an indicator of *socioeconomic position* (as in the studies by Penfornis et al [20] and Meijer et al [71]), measured using answer options from *no education* to *university* and recoded according to the International Standard Classification of Education [50].

XSL•FO
RenderX

Behavioral Experiences

This was measured through *time spent* on each future-self task (in minutes), extracted from Qualtrics, and 2 open questions about *experiences with the desired- and undesired-future-self tasks* (ie, "What was it like for you to perform the tasks about yourself in a future when you have [successfully quit/continued] smoking?"). The 2 open questions were expected to probe participants to share their behavioral experiences with future-self tasks. There was no minimum or maximum length for answers.

Cognitive Experiences

A total of 8 items assessed the *experienced difficulty* of different elements of future-self tasks (ie, "How difficult did you find it to do the following things: [visualize/search for images/write a text/write keywords] describing yourself [as successfully quit smoker/continued smoker]?"), with answer options 1=*very difficult* to 7=*very easy*. The variable was recoded so that a higher score indicated greater difficulty. Scores for writing a short text and writing keywords were averaged to create 1 verbal difficulty score. The 2 open questions about *experiences with the desired- and undesired-future-self tasks* were expected to probe participants to share their cognitive experiences with future-self tasks.

Affective Experiences

In total, 8 items ascertained *emotional responses* to future-self tasks, asking participants to what extent they felt happy, scared, sad, or angry when imagining a future in which they successfully quit or continued smoking, rated from 1=*totally disagree* to 5=*totally agree*. An average score was computed for each emotion. The 2 open questions about *experiences with the desired- and undesired-future-self tasks* were expected to probe participants to share their affective experiences with future-self tasks.

Data Analysis

Descriptive statistics were computed, and a 1-sample *t* test examined differences from 3.5, the midpoint of the scale, in the average experienced difficulty across the 2 future-self tasks. RM ANOVAs explored differences in (1) time spent based on polarity (desired vs undesired future self) and modality (verbal vs visual description of the future self) and (2) experienced difficulty based on polarity and future-self task element (ie, mental visualization vs verbal vs visual description of the future self). A multivariate ANOVA ascertained differences in the strength of the 4 emotions (happy, scared, sad, and angry) based on polarity. Assumptions were verified (Multimedia Appendix 2).

Answers to both open questions—which ranged from 1 word to 3 sentences—were coded for all participants who provided one (83/87, 95%). In total, 6% (5/83) of the participants provided identical answers for both future-self tasks, which were counted and interpreted only on first occurrence. A total of 6 answers of only a few characters in length or irrelevant for describing the experience (eg, "looking in the future") were excluded from the analysis. The full coding tree can be found in Multimedia Appendix 3.

Results

The main findings are summarized in Table 3.

Quantitative Results

Time Spent on Future-Self Tasks

Future-self tasks were completed in an average of 6.64 (SD 8.30) minutes. Figure 3 shows mean time spent by polarity and modality. An RM ANOVA assessed differences in time spent based on polarity and modality. There was no significant interaction effect ($F_{1,81}$=0.74; P=.39; η_p^2=0.009), but there were significant main effects of polarity ($F_{1,81}$=23.81; P<.001; η_p^2=0.227) and modality ($F_{1,81}$=5.18; P=.03; η_p^2=0.060) on time spent. Specifically, participants spent significantly more time on the verbal than on the visual task (Δ=1.26, 95% CI 0.16-2.36) and on the desired- than on the undesired-future-self task (Δ=2.16, 95% CI 1.28-3.04).

Figure 3. Mean experienced difficulty with, time spent on, and emotional response to smoking-related future-self tasks by polarity (desired or undesired future self) and modality (visualizing the future self, describing it verbally, or describing it visually) in study 3 (N=87).

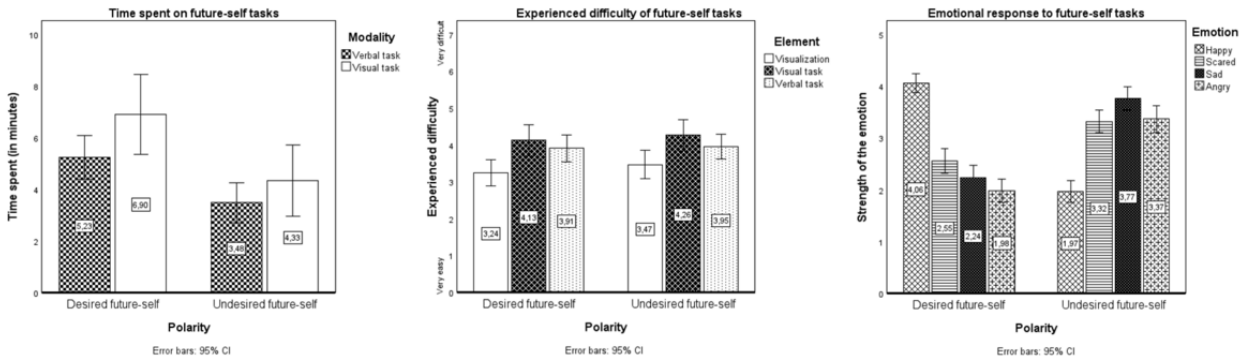

Experienced Difficulty of Future-Self Tasks

Future-self tasks were experienced as relatively difficult (mean 3.83, SD 1.02; t_{86}=3.00; P=.002; Cohen d=0.32). Figure 3 shows mean experienced difficulty by polarity and task element. An RM ANOVA assessed differences in experienced difficulty based on polarity and task element. There was no significant interaction between polarity and element regarding experienced difficulty ($F_{2,85}$=0.23; P=.80; η_p^2=0.005). There was a significant main effect of element ($F_{2,85}$=8.32; P<.001; η_p^2=0.134). Specifically, searching for images summarizing the

XSL•FO
RenderX

future selves was experienced as significantly more difficult than merely envisioning the future self (Δ=−0.84, 95% CI −1.30 to −0.30).

Emotional Response to Future-Self Tasks

Figure 3 shows mean strength of each emotion by polarity. A multivariate ANOVA ascertained differences in the strength of the 4 emotions based on polarity. There was a significant main effect of polarity ($F_{4,83}$=35.38; P<.001; η_p^2=0.630). Specifically, the desired-future-self task elicited significantly more happiness than the undesired-future-self task ($F_{1,86}$=141.58; P<.001; η_p^2=0.622). The undesired-future-self task elicited significantly more anger ($F_{1,86}$=61.14; P<.001; η_p^2=0.416), sadness ($F_{1,86}$=66.31; P<.001; η_p^2=0.435), and anxiety ($F_{1,86}$=21.60; P<.001; η_p^2=0.201).

Qualitative Results

Behavioral Experiences With Future-Self Tasks

Only 2% (2/83) of the responses informed about behavioral experiences with the tasks. Specifically, one participant printed and kept the images describing his future selves, and another mentioned finding the task to be difficult to translate into action.

Cognitive Experiences With Future-Self Tasks

Nearly equal numbers of participants found the desired-future-self task easy (11/83, 13%) or difficult (8/83, 10%), with some describing it as inspiring (3/83, 4%), interesting (3/83, 4%), or important (2/83, 2%). Individual cases found it empowering, insightful, and peculiar. The undesired-future-self task was mostly considered (very) difficult (17/83, 20%). A few participants (2/83, 2%) found it easy, and individual cases found it peculiar, important, interesting, useful, and inspiring. In total, 1% (1/83) of the participants found both tasks equally easy, 2% (2/83) found the undesired-future-self task easier, and 5% (4/83) found the desired-future-self task easier due to difficulties imagining a negative, unwanted outcome.

Familiarity with envisioning one's future self seemed impactful for task completion. Some participants (2/83, 2%) mentioned that it was their first time considering their future self, whereas others had done so once (1/83, 1%) or repeatedly (3/83, 4%) before. In total, 2% (2/83) of the participants, who were first timers, found the task difficult or "peculiar. I suddenly had to think about what a future without smoking actually means" (female participant; aged 28 years). Among those with previous experience, one participant reported no effect, another participant visualized a clearer future self this time around, and 2 others found the task easy or still difficult.

The desired future self provided something to look forward to (7/83, 8%), although some (3/83, 4%) found it unattainable. It fostered positive perceptions of the future or future self (2/83, 2%) and, in individual cases, prompted reflection on reasons to quit smoking and attitudes toward smoking or had little effect. The undesired-future-self task made clear who participants did not want to become (10/83, 12%):

> *I don't want to become the person I outlined myself—surely a right-thinking person doesn't poison herself!* [Female participant; aged 35 years]

It confronted some with their addiction (2/83, 2%) or its likelihood (2/83, 2%), decreased motivation to quit smoking (1/83, 1%), or felt like "you have to tell bad things about yourself" (male participant; aged 75 years).

Participants' responses suggest that future-self tasks triggered various cognitive processes, including reflection on smoking consequences (10/83, 12%), cognitive dissonance (8/83, 10%), increased motivation to quit smoking (7/83, 8%), resistance to future-self thinking (3/83, 4%), comparing current and future selves (5/83, 6%) or future and past selves (2/83, 2%), and considering the meaning of being a smoker (2/83, 2%).

Preferences regarding the order of the future-self tasks were mixed. One participant preferred the desired-future-self task first for its pleasantness, whereas another favored the undesired-future-self task first to end positively:

> *Now the smoker is in my head instead of the nonsmoker.* [Male participant; aged 53 years]

Affective Experiences With Future-Self Tasks

A total of 17 adjectives were used to describe affective experiences with future-self tasks, both positive and negative. The desired-future-self task was most often called *enjoyable* (21/83, 25%). It was also described as positive (6/83, 7%), confronting (3/83, 4%), unpleasant (1/83, 1%), and depressing (1/83, 1%).

The undesired-future-self task was primarily described as confronting (21/83, 25%), unpleasant (10/83, 12%), depressing (5/83, 6%), negative (6/83, 7%), enlightening (3/83, 4%), anxiety inducing (3/83, 4%), and uncomfortable (2/83, 2%). Some described it as frustrating, humiliating, neutral, positive, and triggering insecurity. Notably, 2% (2/83) of the participants found the task enjoyable.

Discussion

Regarding behavioral experiences, quantitative findings from study 3 showed that future-self tasks were completed in less than 7 minutes, with more time devoted to the desired-future-self and verbal tasks. This indicates that FSIs can be time efficient.

Concerning cognitive experiences, quantitative results indicated that future-self tasks were considered relatively difficult, with visual description of future selves being harder than visualization. Qualitative results revealed variations in experienced difficulty—the desired-future-self task was viewed as both easy and difficult, whereas the undesired-future-self task was mostly seen as difficult. Furthermore, as observed in studies 1 and 2, unfamiliarity with envisioning one's future self seemed to hinder task performance, although familiarity did not always ease it. Finally, task order preferences varied. These findings highlight the importance of considering differences in individual experiences with FSIs.

In line with the findings of studies 1 and 2, future-self tasks appeared to trigger numerous cognitive processes, further hinting at FSIs having the capacity to prompt action. Coherent with its

XSL•FO
RenderX

concept, the desired future self generally provided something to look forward to and sometimes led to reflections on quitting smoking. Conversely, the undesired future self served as something to avoid, was seen as confronting, and sometimes decreased motivation to quit or prompted negative self-reflection.

Regarding affective experiences, quantitative results showed that participants mostly felt happy during the desired-future-self task and sad, angry, or anxious during the undesired-future-self task. Qualitative results supported this, with the desired-future-self task described as enjoyable and the undesired-future-self task described as confronting and unpleasant, although with individual variations. These experiences are understandable given the concept of the future-self tasks and suggest that they had the intended effect.

General Discussion

Principal Findings

Overview

This multistudy report involved various smoking- and PA-related FSIs from multiple empirical studies comprising a large number of participants. It is the first to explore user experiences with smoking- and PA-related FSIs, triangulating behavioral, cognitive, and affective dimensions. Regarding behavioral experiences, participants generally reported successful and timely integration of FSIs into daily life, with some noting changes in PA or smoking behavior despite the tasks being preparatory. Regarding cognitive experiences, a moderate effort was exerted, with task difficulty varying by individual and task, yet future-self tasks consistently prompted cognitive processes supporting behavior change. Regarding affective experiences, desired-future-self tasks were generally seen as enjoyable and happiness inducing, whereas undesired-future-self tasks were perceived as confronting, unpleasant, and evoking negative emotions. The findings of all 3 studies are summarized in Table 3.

Behavioral Experiences With FSIs

Study 3 reflected that participants successfully and quickly completed digital FSIs and integrated them into their lives. With a completion time of less than 7 minutes—much less than the 15 minutes anticipated in study 1 and 14 to 20 minutes reported in previous studies [37,38]—these interventions seem time efficient. Differences in completion time in this study compared to previous ones may be due to the digital format allowing for more freedom over exercise length. The discrepancy with study 1's estimate may stem from difficulties estimating time requirements without completing the intervention.

The successful integration of FSIs into daily life suggests feasibility and acceptability. Participants did not provide negative comments about the digital format, which further supports the feasibility and acceptability of digital implementation. Digital implementation, in turn, optimizes human and monetary resources for smoking cessation and PA promotion and allows for delivering support anywhere, anytime [8].

Studies 1 and 3 provided additional insights into future-self task completion time. Study 3 revealed that verbal descriptions took longer than visual ones, consistent with study 1's expectations. This may be because visually describing the future self extends the process of producing mental images, partly relying on the same visual cognitive processes [69,72], whereas verbal future-self tasks, which rely on different cognitive processes, would understandably take longer. Alternatively, participants generally used existing images for visual descriptions, which is quicker than creating written materials. Future research could validate and shed light on the origin of this difference. Study 3 also showed that desired-future-self tasks took longer than undesired-future-self ones, possibly because participants were more inclined to invest time in considering and describing a positive future self. Therefore, future research and FSI designs could explore how to optimize task formats to balance cognitive engagement, practical feasibility, and impact.

One study 2 participant attempted to quit smoking, and multiple participants increased their PA after the future-self tasks, suggesting that FSIs can be effective in influencing behavior. These findings echo previous successes of FSIs in influencing smoking and PA behaviors [31,33-38] and support the continued use of these interventions to promote change in these health behaviors. In contrast, no behavior changes were observed in study 3, possibly because study 2 participants were explicitly told that FSIs might help quit smoking—a hypothesis for future research. Reported changes mainly involved increased PA, possibly because the PA-related tasks in study 2 presented PA as a potential facilitator of smoking cessation, empowering participants to increase their PA as a first step. Alternatively, as the participants were daily smokers and, in most cases, regular exercisers, smoking and PA were likely anchored in their identity. We know that people prefer to act in line with their identity and will try to avoid behaviors that do not align with or threaten it [16,73]. Thus, tasks aiming to connect nonsmoking with their identity may have felt less self-relevant or threatening, discouraging changes, whereas PA-related tasks possibly reinforced PA identity, encouraging the behavior. These findings suggest that interventions are more effective when they build on established identity or carefully frame new behaviors, such as nonsmoking, to align with or positively reinforce the individual's current identity.

Cognitive Experiences With FSIs

Slightly above-average effort was exerted to complete the tasks in study 2, akin to effort deployed in comparable future-self tasks [74] and the other preparatory tasks from the study. This effort level suggests that the tasks either included fitting elements to keep individuals engaged [71] or lacked such elements, resulting in individuals not being sufficiently engaged to exert more effort. This effort level further implies that participants took the tasks seriously, possibly because they found them useful and meaningful [37,60,76,77], and may indicate tasks of appropriate difficulty and emotional engagement (see the motivational intensity theory [74,75]). No significant differences in effort were observed based on behavior, polarity, or modality, suggesting equal willingness to describe the desired or undesired future self verbally or visually regardless of whether the behavior is health compromising or

XSL•FO
RenderX

health promoting. While the aforementioned findings are insightful, more research (eg, conducting in-depth interviews with users to understand their perceptions of effort) would be beneficial for unequivocal conclusions.

Across all studies, task difficulty varied depending on individuals, tasks, and quantitative versus qualitative data. More specifically, in study 1, future-self tasks were anticipated to be relatively easy, whereas in study 3, they were experienced as fairly difficult. In study 1, undesired-future-self, PA-related, and visual future-self tasks were anticipated to be harder, but in study 3, only visual tasks were perceived as more challenging. In addition, while in studies 1 and 2 participants often attributed difficulty to visualization, study 3 found it to be the easiest task component, with visual description being the most challenging. Given that study 3 actually measured difficulty directly—unlike study 1, which measured anticipated completion time—its findings could be considered more reliable. However, its qualitative results partly contradicted its quantitative findings as envisioning future selves was generally considered difficult. In summary, conclusions about future-self task difficulty are inconsistent, warranting further investigation. In total, 2 exceptions emerged. First, visual future-self tasks were generally seen as difficult, suggesting that verbal tasks may be preferable for future interventions, although more time-consuming. Second, all 3 studies found that unfamiliarity with mental visualization complicated future-self tasks, supporting the teaching [38,41] and practice of mental visualization.

Future-self tasks in studies 2 and 3 were reported to trigger cognitive processes (eg, considering the consequences of PA or smoking behavior or behavior change motivation fluctuations, which typically precede behavior change [78-81]). Sometimes, the tasks seemed to trigger mental contrasting among past, present, or future selves, a process beneficial for achieving the desired identity and avoiding the undesired one [29,82]. Coupled with behavior change reports, it seems that future-self tasks can effectively prepare participants for behavior change by connecting healthier behaviors with their identity. Future research could assess whether FSIs are more suited to change certain health behaviors and whether they have the capacity to change multiple health behaviors simultaneously [7,83].

Affective Experiences With FSIs

All studies showed that desired-future-self tasks were generally experienced as positive and enjoyable and associated with happiness, aligning with previous research [42,43]. Conversely, undesired-future-self tasks were reported as negative, confronting, and unpleasant and linked to feeling scared, sad, or angry, also supporting previous findings [43]. By allowing individuals to anticipate feeling what it is like to (not) be someone who engages in a certain behavior, future-self imagery can drive intentional behavior change [27,28] and sustain desire for change until the behavior becomes habitual [84]. Therefore, despite causing some measure of negative arousal, the undesired-future-self task appears crucial for motivating behavior change and should remain part of FSIs. To minimize psychological distress, strategies such as ending with desired-future-self tasks, allowing for task order choice, rehearsing steps to avoid or achieve the undesired or desired future self to boost self-efficacy [33,37,84], or arranging access to (professional) human support could be used.

Strengths and Limitations

This multistudy report has several limitations. First, conducting secondary analyses on data from different studies with diverse original aims, interventions, and methodologies complicated the triangulation and interpretation of findings. However, combining mixed methods data across interventions and research fields provided richer insights into user experiences, which can guide hypothesis formulation in future research. Second, post hoc power analyses showed that some effects, especially interaction effects in studies 1 and 3, fell below the conventional 0.80 threshold. Hence, certain effects may have been missed, which calls for research using larger, adequately powered samples to validate the findings. Nevertheless, we systematically tested main and simple main effects across all factors, reducing the likelihood of having missed essential patterns in the data. Third, as the participants were predominantly highly educated—and, therefore, likely digitally literate and used to performing tasks requiring cognitive load—some sampling bias may have been introduced. The findings of this multistudy report may not transfer to populations with middle and lower educational levels [85], in particular people who smoke [86-88], and have to be interpreted carefully. Fourth, while all studies mostly relied on self-report measures, which are less objective than use or performance measures, such measures are commonly used in research and valued for their convenience, efficiency [89], and capacity to capture complex experiences. In addition, combining qualitative and quantitative data and exploring behavioral, cognitive, and affective aspects likely mitigated the limitations of individual studies, offering a more thorough perspective on user experiences with FSIs.

Conclusions

The findings of this multistudy report support the future use of FSIs as a time-efficient, feasible, and acceptable method for promoting smoking cessation and PA by connecting healthier behaviors with self-identity. Experiences with FSIs did not appear to significantly differ between smoking, a health-compromising behavior, and PA, a health-promoting behavior. Digital administration seemed acceptable and is beneficial for efficient use of human and monetary resources. We encourage FSIs to include both a desired- and undesired-future-self task and train users in mental visualization. The intervention may be improved by tailoring it to individual preferences, such as choosing to start with the desired- or undesired-future-self task. The results enhance the scientific understanding of FSIs and offer guidance for designing tailored, effective, scalable, user-centered digital interventions promoting smoking cessation and PA. While the results encourage further use, inconsistencies in them highlight the need for more research to optimize FSIs, such as further assessing user experiences; the effectiveness of FSIs in influencing multiple health behaviors; and the impact of factors such as socioeconomic position, self-identity strength, or behavioral history.

XSL•FO
RenderX

Acknowledgments

The authors would like to thank Emma von Haeseler, BS, for her invaluable contribution to the qualitative analyses in this multistudy report. They also give special thanks to Professor Dr Ir Geke Ludden for her input on the manuscript's structure and narrative and valuable literature references in that regard. This work is part of the multidisciplinary research project Perfect Fit. This research received funding from the Netherlands Organisation for Scientific Research Commit2Data – Big Data & Health program (project 628.011.211). The program was funded by the following parties: the Netherlands Organisation for Scientific Research; the Netherlands Organisation for Health Research and Development; the Dutch Heart Foundation; the Ministry of Health, Welfare, and Sport; Health~Holland; and the Netherlands eScience Center. The publication reflects only the authors' views, and the funders are not liable for any use that may be made of the information contained herein. ChatGPT (OpenAI) was used to correct spelling and grammatical errors and to shorten the manuscript. All corrections were verified by the first author.

Data Availability

The datasets generated or analyzed during study 1 [57], study 2 [53], and study 3 [20], as well as the analysis codes for all 3 studies are available online [53].

Conflicts of Interest

None declared.

Multimedia Appendix 1
Instructions for the future-self interventions used in studies 1, 2, and 3.
[DOCX File , 19 KB - formative_v9i1e63893_app1.docx]

Multimedia Appendix 2
Results of assumption testing and robustness analyses.
[DOCX File , 18 KB - formative_v9i1e63893_app2.docx]

Multimedia Appendix 3
Coding trees for qualitative analyses.
[DOCX File , 21 KB - formative_v9i1e63893_app3.docx]

References

1. Global burden of disease (GBD). Institute for Health Metrics and Evaluation. URL: https://www.healthdata.org/research-analysis/gbd [accessed 2024-02-09]
2. Child mortality rate, 2022. Our World in Data. URL: https://ourworldindata.org/explorers/global-health [accessed 2025-04-23]
3. Ritchie H, Roser M. Smoking. Our World in Data. URL: https://ourworldindata.org/smoking [accessed 2024-02-09]
4. Yang Y, Xu H, Liu X, Li J, Liew Z, Liu X, et al. Joint association of smoking and physical activity with mortality in elderly hypertensive patients: a Chinese population-based cohort study in 2007-2018. Front Public Health 2022 Sep 29;10:1005260 [FREE Full text] [doi: 10.3389/fpubh.2022.1005260] [Medline: 36249230]
5. Stenholm S, Head J, Kivimäki M, Kawachi I, Aalto V, Zins M, et al. Smoking, physical inactivity and obesity as predictors of healthy and disease-free life expectancy between ages 50 and 75: a multicohort study. Int J Epidemiol 2016 Aug;45(4):1260-1270 [FREE Full text] [doi: 10.1093/ije/dyw126] [Medline: 27488415]
6. Duan Y, Shang B, Liang W, Du G, Yang M, Rhodes RE. Effects of eHealth-based multiple health behavior change interventions on physical activity, healthy diet, and weight in people with noncommunicable diseases: systematic review and meta-analysis. J Med Internet Res 2021 Feb 22;23(2):e23786 [FREE Full text] [doi: 10.2196/23786] [Medline: 33616534]
7. Prochaska JJ, Prochaska JO. A review of multiple health behavior change interventions for primary prevention. Am J Lifestyle Med 2011 May;5(3):208-221 [FREE Full text] [doi: 10.1177/1559827610391883] [Medline: 24358034]
8. Marcolino MS, Oliveira JA, D'Agostino M, Ribeiro AL, Alkmim MB, Novillo-Ortiz D. The impact of mHealth interventions: systematic review of systematic reviews. JMIR Mhealth Uhealth 2018 Jan 17;6(1):e23 [FREE Full text] [doi: 10.2196/mhealth.8873] [Medline: 29343463]
9. Fang YE, Zhang Z, Wang R, Yang B, Chen C, Nisa C, et al. Effectiveness of eHealth smoking cessation interventions: systematic review and meta-analysis. J Med Internet Res 2023 Jul 28;25:e45111 [FREE Full text] [doi: 10.2196/45111] [Medline: 37505802]
10. Smoking cessation: a report of the surgeon general. U.S. Department of Health and Human Services. 2020. URL: https://www.hhs.gov/sites/default/files/2020-cessation-sgr-full-report.pdf [accessed 2024-09-02]

XSL•FO
RenderX

11. Europe physical activity factsheet 2021. European Commission. URL: https://sport.ec.europa.eu/document/europe-physical-activity-factsheet-2021 [accessed 2025-05-15]
12. Chaiton M, Diemert L, Cohen JE, Bondy SJ, Selby P, Philipneri A, et al. Estimating the number of quit attempts it takes to quit smoking successfully in a longitudinal cohort of smokers. BMJ Open 2016 Jun 09;6(6):e011045 [FREE Full text] [doi: 10.1136/bmjopen-2016-011045] [Medline: 27288378]
13. Hartmann-Boyce J, Livingstone-Banks J, Ordóñez-Mena JM, Fanshawe TR, Lindson N, Freeman SC, et al. Behavioural interventions for smoking cessation: an overview and network meta-analysis. Cochrane Database Syst Rev 2021 Jan 04;1(1):CD013229. [doi: 10.1002/14651858.CD013229.pub2] [Medline: 33411338]
14. Jepson RG, Harris FM, Platt S, Tannahill C. The effectiveness of interventions to change six health behaviours: a review of reviews. BMC Public Health 2010 Sep 08;10:538 [FREE Full text] [doi: 10.1186/1471-2458-10-538] [Medline: 20825660]
15. Murray JM, Brennan SF, French DP, Patterson CC, Kee F, Hunter RF. Effectiveness of physical activity interventions in achieving behaviour change maintenance in young and middle aged adults: a systematic review and meta-analysis. Soc Sci Med 2017 Nov;192:125-133. [doi: 10.1016/j.socscimed.2017.09.021] [Medline: 28965003]
16. Burke PJ, Stets JE. Identity Theory. Oxford, UK: Oxford University Press; 2009.
17. Oyserman D. Identity‐based motivation: implications for action‐readiness, procedural‐readiness, and consumer behavior. J Consum Psychol 2009 Jun 25;19(3):250-260. [doi: 10.1016/j.jcps.2009.05.008]
18. Meijer E, Van den Putte B, Gebhardt WA, Van Laar C, Bakk Z, Dijkstra A, et al. A longitudinal study into the reciprocal effects of identities and smoking behaviour: findings from the ITC Netherlands Survey. Soc Sci Med 2018 Mar;200:249-257 [FREE Full text] [doi: 10.1016/j.socscimed.2017.12.006] [Medline: 29321102]
19. Smit ES, Hoving C, Schelleman-Offermans K, West R, de Vries H. Predictors of successful and unsuccessful quit attempts among smokers motivated to quit. Addict Behav 2014 Sep;39(9):1318-1324 [FREE Full text] [doi: 10.1016/j.addbeh.2014.04.017] [Medline: 24837754]
20. Penfornis KM, Gebhardt WA, Rippe RC, Van Laar C, van den Putte B, Meijer E. My future-self has (not) quit smoking: an experimental study into the effect of a future-self intervention on smoking-related self-identity constructs. Soc Sci Med 2023 Mar;320:115667 [FREE Full text] [doi: 10.1016/j.socscimed.2023.115667] [Medline: 36641885]
21. Tombor I, Shahab L, Brown J, West R. Positive smoker identity as a barrier to quitting smoking: findings from a national survey of smokers in England. Drug Alcohol Depend 2013 Dec 01;133(2):740-745. [doi: 10.1016/j.drugalcdep.2013.09.001] [Medline: 24075070]
22. van den Putte B, Yzer M, Willemsen MC, de Bruijn GJ. The effects of smoking self-identity and quitting self-identity on attempts to quit smoking. Health Psychol 2009 Sep;28(5):535-544. [doi: 10.1037/a0015199] [Medline: 19751079]
23. de Bruijn GJ, van den Putte B. Exercise promotion: an integration of exercise self‐identity, beliefs, intention, and behaviour. Eur J Sport Sci 2012 Feb 29;12(4):354-366. [doi: 10.1080/17461391.2011.568631]
24. Havitz ME, Kaczynski AT, Mannell RC. Exploring relationships between physical activity, leisure involvement, self-efficacy, and motivation via participant segmentation. Leis Sci 2013;35(1):45-62. [doi: 10.1080/01490400.2013.739890]
25. Huffman MK, Christ SL, Ferraro KF, Klenosky DB, Marceau K, Amireault S. Measuring physical activity regulatory styles and identity among adults 55 years or older. J Aging Health 2022 Oct;34(6-8):775-785. [doi: 10.1177/08982643211063349] [Medline: 35100880]
26. Strachan SM, Perras MG, Forneris T, Stadig GS. I'm an exerciser: common conceptualisations of and variation in exercise identity meanings. Int J Sport Exerc Psychol 2015 Oct 19;15(3):321-336. [doi: 10.1080/1612197x.2015.1096292]
27. Markus H, Nurius P. Possible selves. Am Psychol 1986;41(9):954-969. [doi: 10.1037/0003-066X.41.9.954]
28. Oyserman D, Bybee D, Terry K, Hart-Johnson T. Possible selves as roadmaps. J Res Pers 2004 Apr;38(2):130-149. [doi: 10.1016/s0092-6566(03)00057-6]
29. Oettingen G. Future thought and behaviour change. Eur Rev Soc Psychol 2012 Mar 13;23(1):1-63. [doi: 10.1080/10463283.2011.643698]
30. Oyserman D, Destin M, Novin S. The context-sensitive future self: possible selves motivate in context, not otherwise. Self Identity 2014 Oct 08;14(2):173-188. [doi: 10.1080/15298868.2014.965733]
31. Rutchick AM, Slepian ML, Reyes MO, Pleskus LN, Hershfield HE. Future self-continuity is associated with improved health and increases exercise behavior. J Exp Psychol Appl 2018 Mar;24(1):72-80. [doi: 10.1037/xap0000153] [Medline: 29595304]
32. Meijer E, Gebhardt WA, van Laar C, van den Putte B, Evers AW. Strengthening quitter self-identity: an experimental study. Psychol Health 2018 Oct;33(10):1229-1250 [FREE Full text] [doi: 10.1080/08870446.2018.1478976] [Medline: 29886765]
33. Chan CK, Cameron LD. Promoting physical activity with goal-oriented mental imagery: a randomized controlled trial. J Behav Med 2012 Jun;35(3):347-363. [doi: 10.1007/s10865-011-9360-6] [Medline: 21695405]
34. Murru EC, Martin Ginis KA. Imagining the possibilities: the effects of a possible selves intervention on self-regulatory efficacy and exercise behavior. J Sport Exerc Psychol 2010 Aug;32(4):537-554. [doi: 10.1123/jsep.32.4.537] [Medline: 20733212]
35. Ouellette JA, Hessling R, Gibbons FX, Reis-Bergan M, Gerrard M. Using images to increase exercise behavior: prototypes versus possible selves. Pers Soc Psychol Bull 2005 May;31(5):610-620. [doi: 10.1177/0146167204271589] [Medline: 15802656]

XSL•FO
RenderX

36. Shadel WG, Cervone D. Evaluating social-cognitive mechanisms that regulate self-efficacy in response to provocative smoking cues: an experimental investigation. Psychol Addict Behav 2006 Mar;20(1):91-96. [doi: 10.1037/0893-164X.20.1.91] [Medline: 16536671]
37. Tindle HA, Barbeau EM, Davis RB, Eisenberg DM, Park ER, Phillips RS, et al. Guided imagery for smoking cessation in adults: a randomized pilot trial. Complement Health Pract Rev 2006 Jan;11(3):166-175. [doi: 10.1177/1533210106296773]
38. Wynd C. Guided health imagery for smoking cessation and long-term abstinence. J Nurs Scholarsh 2005;37(3):245-250 [FREE Full text] [doi: 10.1111/j.1547-5069.2005.00042.x] [Medline: 16235865]
39. Meslot C, Gauchet A, Allenet B, François O, Hagger MS. Theory-based interventions combining mental simulation and planning techniques to improve physical activity: null results from two randomized controlled trials. Front Psychol 2016 Nov 16;7:1789 [FREE Full text] [doi: 10.3389/fpsyg.2016.01789] [Medline: 27899904]
40. Yardley L, Spring BJ, Riper H, Morrison LG, Crane DH, Curtis K, et al. Understanding and promoting effective engagement with digital behavior change interventions. Am J Prev Med 2016 Nov;51(5):833-842. [doi: 10.1016/j.amepre.2016.06.015] [Medline: 27745683]
41. Scarpa MP, Prilletensky I, McMahon A, Myers ND, Prilleltensky O, Lee S, et al. Is fun for wellness engaging? Evaluation of user experience of an online intervention to promote well-being and physical activity. Front Comput Sci 2021 Jun 18;3. [doi: 10.3389/fcomp.2021.690389]
42. Boyatzis RE, Akrivou K. The ideal self as the driver of intentional change. J Manag Dev 2006 Aug 1;25(7):624-642. [doi: 10.1108/02621710610678454]
43. Schubert T, Eloo R, Scharfen J, Morina N. How imagining personal future scenarios influences affect: systematic review and meta-analysis. Clin Psychol Rev 2020 Feb;75:101811. [doi: 10.1016/j.cpr.2019.101811] [Medline: 31884148]
44. Kelders SM, van Zyl LE, Ludden GD. The concept and components of engagement in different domains applied to eHealth: a systematic scoping review. Front Psychol 2020 May 27;11:926 [FREE Full text] [doi: 10.3389/fpsyg.2020.00926] [Medline: 32536888]
45. O'Brien HL, Toms EG. What is user engagement? A conceptual framework for defining user engagement with technology. J Am Soc Inf Sci Technol 2008 Feb 28;59(6):938-955. [doi: 10.1002/asi.20801]
46. Perski O, Blandford A, West R, Michie S. Conceptualising engagement with digital behaviour change interventions: a systematic review using principles from critical interpretive synthesis. Transl Behav Med 2017 Jun;7(2):254-267 [FREE Full text] [doi: 10.1007/s13142-016-0453-1] [Medline: 27966189]
47. Short CE, DeSmet A, Woods C, Williams SL, Maher C, Middelweerd A, et al. Measuring engagement in eHealth and mHealth behavior change interventions: viewpoint of methodologies. J Med Internet Res 2018 Nov 16;20(11):e292 [FREE Full text] [doi: 10.2196/jmir.9397] [Medline: 30446482]
48. Dawadi S, Shrestha S, Giri RA. Mixed-methods research: a discussion on its types, challenges, and criticisms. J Pract Stud Educ 2021 Feb 24;2(2):25-36. [doi: 10.46809/jpse.v2i2.20]
49. Lynn M. More multi-study articles wanted. Cornell Hosp Q 2017 Jun 23;58(3):228. [doi: 10.1177/1938965517714372]
50. International standard classification of education (ISCED). International Labor Organization. URL: https://ilostat.ilo.org/methods/concepts-and-definitions/classification-education/ [accessed 2024-02-20]
51. Tabachnick BG, Fidell LS. Using Multivariate Statistics. London, UK: Pearson; 2018.
52. Cohen J. Statistical Power Analysis for the Behavioral Sciences Second Edition. Milton Park, UK: Routledge; 1988.
53. Penfornis KM, Albers N. User experiences with digital future-self interventions in the contexts of smoking and physical activity: a mixed methods multi-study exploration - data and analysis code. 4TU.ResearchData. 2025. URL: https://data.4tu.nl/datasets/951b2dc4-a59e-48ed-9856-af484b125393 [accessed 2025-04-23]
54. White MD, Marsh EE. Content analysis: a flexible methodology. Libr Trends 2006;55(1):22-45. [doi: 10.1353/lib.2006.0053]
55. Albers N. Psychology-informed reinforcement learning for situated virtual coaching in smoking cessation. Delft University of Technology. 2025. URL: https://doi.org/10.4233/uuid:5e1d1500-1c45-4869-b995-6ed7db8f8b3b [accessed 2025-04-22]
56. Albers N, Brinkman WP. Perfect Fit - beliefs about and competencies built by preparatory activities for quitting smoking and becoming more physically active. OSF Registries. 2022. URL: https://osf.io/cax6f [accessed 2024-01-16]
57. Albers N, Neerincx MA, Brinkman WP. Difficulty and time perceptions of preparatory activities for quitting smoking: Dataset. 4TU.ResearchData. 2023 Nov 6. URL: https://data.4tu.nl/datasets/5198f299-9c7a-40f8-8206-c18df93ee2a0/1 [accessed 2025-04-23]
58. Prolific homepage. Prolific. URL: https://www.prolific.com [accessed 2024-10-23]
59. Researcher help center. Prolific. URL: https://researcher-help.prolific.com/hc/en-gb/articles/360009220833-Who-are-the-participants-on-Prolific [accessed 2024-01-16]
60. Albers N, Neerincx MA, Penfornis KM, Brinkman WP. Users' needs for a digital smoking cessation application and how to address them: a mixed-methods study. PeerJ 2022 Aug 19;10:e13824 [FREE Full text] [doi: 10.7717/peerj.13824] [Medline: 36003307]
61. Qualtrics XM homepage. Qualtrics. URL: https://www.qualtrics.com/nl/ [accessed 2024-10-23]
62. What is your pricing? Prolific. URL: https://researcher-help.prolific.com/hc/en-gb/articles/360009223533-What-is-your-pricing [accessed 2024-01-17]

XSL•FO
RenderX

63. Schaap MM, Kunst AE. Monitoring of socio-economic inequalities in smoking: learning from the experiences of recent scientific studies. Public Health 2009 Feb;123(2):103-109. [doi: 10.1016/j.puhe.2008.10.015] [Medline: 19147163]
64. Perfect Fit - learning when to involve a human coach in an eHealth application for preparing for quitting smoking or vaping. OSF Registries. URL: https://osf.io/78cnr [accessed 2025-05-30]
65. Perfect Fit - experiment to gather data for and test a reinforcement learning-approach for motivating people. OSF Registries. URL: https://osf.io/k2uac [accessed 2025-05-30]
66. Albers N, Brinkman WP. Perfect Fit - learning to propose useful preparatory activities for quitting smoking and becoming more physically active. OSF Registries. 2023 Jul 21. URL: https://osf.io/nuy4w [accessed 2025-04-09]
67. Albers N. PerfectFit-project/virtual_coach_human_inv: virtual coach kai for preparing for quitting smoking with human support. Zenodo. 2024 May 2. URL: https://zenodo.org/records/11102861 [accessed 2025-04-22]
68. Example dialog with the virtual coach kai for preparing to quit smoking with human support. YouTube. URL: https://www.youtube.com/watch?v=gLiy5AnolYE [accessed 2024-11-08]
69. Meijer E, Gebhardt WA, van Laar C, Chavannes NH, van den Putte B. Identified or conflicted: a latent class and regression tree analysis explaining how identity constructs cluster within smokers. BMC Psychol 2022 Oct 07;10(1):231 [FREE Full text] [doi: 10.1186/s40359-022-00937-y] [Medline: 36207742]
70. Penfornis KM. My future-self has (not) quit smoking: an experimental study into the effect of a future-self intervention on smoking-related self-identity constructs. DataverseNL 2022. [doi: 10.34894/9GZFTX]
71. Meijer E, Gebhardt WA, Dijkstra A, Willemsen MC, Van Laar C. Quitting smoking: the importance of non-smoker identity in predicting smoking behaviour and responses to a smoking ban. Psychol Health 2015;30(12):1387-1409. [doi: 10.1080/08870446.2015.1049603] [Medline: 25959600]
72. Kavanagh DJ, Andrade J, May J. Imaginary relish and exquisite torture: the elaborated intrusion theory of desire. Psychol Rev 2005 Apr;112(2):446-467 [FREE Full text] [doi: 10.1037/0033-295X.112.2.446] [Medline: 15783293]
73. Oyserman D, Elmore K, Smith G. Self, self-concept, and identity. In: Leary M, Tangney J, editors. Handbook of Self and Identity Second Edition. New York City, NY: Guilford Publications; Jan 2012.
74. Albers N, Neerincx MA, Brinkman WP. Addressing people's current and future states in a reinforcement learning algorithm for persuading to quit smoking and to be physically active. PLoS One 2022 Dec 1;17(12):e0277295 [FREE Full text] [doi: 10.1371/journal.pone.0277295] [Medline: 36454782]
75. Szinay D, Jones A, Chadborn T, Brown J, Naughton F. Influences on the uptake of and engagement with health and well-being smartphone apps: systematic review. J Med Internet Res 2020 May 29;22(5):e17572 [FREE Full text] [doi: 10.2196/17572] [Medline: 32348255]
76. Baumel A, Muench FJ. Effort-optimized intervention model: framework for building and analyzing digital interventions that require minimal effort for health-related gains. J Med Internet Res 2021 Mar 12;23(3):e24905 [FREE Full text] [doi: 10.2196/24905] [Medline: 33709943]
77. Venkatesh V, Thong JY, Xu X. Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology. MIS Q 2012 Mar;36(1):157-178. [doi: 10.2307/41410412]
78. Barnett G, Boduszek D, Willmott D. What works to change identity? A rapid evidence assessment of interventions. J Appl Soc Psychol 2021 May 03;51(7):698-719. [doi: 10.1111/jasp.12776]
79. Deci EL, Ryan RM. Self-determination theory: a macrotheory of human motivation, development, and health. Can Psychol 2008;49(3):182-185. [doi: 10.1037/a0012801]
80. Michie S, Ashford S, Sniehotta FF, Dombrowski SU, Bishop A, French DP. A refined taxonomy of behaviour change techniques to help people change their physical activity and healthy eating behaviours: the CALO-RE taxonomy. Psychol Health 2011 Nov;26(11):1479-1498. [doi: 10.1080/08870446.2010.540664] [Medline: 21678185]
81. Michie S, Hyder N, Walia A, West R. Development of a taxonomy of behaviour change techniques used in individual behavioural support for smoking cessation. Addict Behav 2011 Apr;36(4):315-319. [doi: 10.1016/j.addbeh.2010.11.016] [Medline: 21215528]
82. Oyserman D. Pathways to Success Through Identity-Based Motivation. Oxford, UK: Oxford University Press; 2015.
83. James E, Freund M, Booth A, Duncan MJ, Johnson N, Short CE, et al. Comparative efficacy of simultaneous versus sequential multiple health behavior change interventions among adults: a systematic review of randomised trials. Prev Med 2016 Aug;89:211-223. [doi: 10.1016/j.ypmed.2016.06.012] [Medline: 27311332]
84. Solbrig L, Whalley B, Kavanagh DJ, May J, Parkin T, Jones R, et al. Functional imagery training versus motivational interviewing for weight loss: a randomised controlled trial of brief individual interventions for overweight and obesity. Int J Obes (Lond) 2019 Apr;43(4):883-894. [doi: 10.1038/s41366-018-0122-1] [Medline: 30185920]
85. Ritchie H, Samborska V, Ahuja N, Ortiz-Ospina E, Roser M. Global education. Our World in Data. URL: https://ourworldindata.org/global-education [accessed 2024-06-28]
86. Cao P, Jeon J, Tam J, Fleischer NL, Levy DT, Holford TR, et al. Smoking disparities by level of educational attainment and birth cohort in the U.S. Am J Prev Med 2023 Apr;64(4 Suppl 1):S22-S31 [FREE Full text] [doi: 10.1016/j.amepre.2022.06.021] [Medline: 36935129]

XSL•FO
RenderX

87. Ruokolainen O, Härkänen T, Lahti J, Haukkala A, Heliövaara M, Rahkonen O. Association between educational level and smoking cessation in an 11-year follow-up study of a national health survey. Scand J Public Health 2021 Dec;49(8):951-960 [FREE Full text] [doi: 10.1177/1403494821993721] [Medline: 33648397]
88. Tomioka K, Kurumatani N, Saeki K. The association between education and smoking prevalence, independent of occupation: a nationally representative survey in Japan. J Epidemiol 2020 Mar 05;30(3):136-142 [FREE Full text] [doi: 10.2188/jea.JE20180195] [Medline: 30828035]
89. O'Brien HL, Toms EG. The development and evaluation of a survey to measure user engagement. J Am Soc Inf Sci Technol 2009 Dec 09;61(1):50-69. [doi: 10.1002/asi.21229]

Abbreviations

FSI: future-self intervention
PA: physical activity
RM: repeated-measure

Edited by A Mavragani; submitted 29.07.24; peer-reviewed by H Bhurawala, A Eisingerich; comments to author 10.10.24; revised version received 08.11.24; accepted 07.02.25; published 20.06.25.

Please cite as:
Penfornis KM, Albers N, Brinkman WP, Neerincx MA, Evers AWM, Gebhardt WA, Meijer E
User Experiences With Digital Future-Self Interventions in the Contexts of Smoking and Physical Inactivity: Mixed Methods Multistudy Exploration
JMIR Form Res 2025;9:e63893
URL: https://formative.jmir.org/2025/1/e63893
doi:10.2196/63893
PMID:40540315

©Kristell M Penfornis, Nele Albers, Willem-Paul Brinkman, Mark A Neerincx, Andrea WM Evers, Winifred A Gebhardt, Eline Meijer. Originally published in JMIR Formative Research (https://formative.jmir.org), 20.06.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Research Dissemination Strategies in Pediatric Emergency Care Using a Professional Twitter (X) Account: A Mixed Methods Developmental Study of a Logic Model Framework

Gwendolyn C Hooley[1,2*], MD; Julia N Magana[3], MD; Jason M Woods[4,5], MD; Shyam Sivasankar[6,7], MD; Lauren VonHoltz[8], MD; Anita R Schmidt[1], MPH; Todd P Chang[1], MD, MAcM; Michelle Lin[9*], MD

[1]Children's Hospital Los Angeles, Los Angeles, CA, United States
[2]University of California, Los Angeles, Los Angeles, CA, United States
[3]University of California Irvine, Irvine, CA, United States
[4]University of Colorado School of Medicine, Denver, CO, United States
[5]Children's Hospital Colorado, Denver, CO, United States
[6]Baylor College of Medicine, Houston, TX, United States
[7]Stanford University School of Medicine, Stanford, CA, United States
[8]Children's Hospital of Philadelphia, Philadelphia, PA, United States
[9]University of California, San Francisco, San Francisco, CA, United States
[*]these authors contributed equally

Corresponding Author:
Michelle Lin, MD
University of California, San Francisco
533 Parnassus Ave
San Francisco, CA, 94143
United States
Phone: 1 (415) 206 5753
Email: mlin@aliem.com

Abstract

Background: Research dissemination is a vital step in bridging the gap between the publication of cutting-edge research and its adoption into clinical practice. Social media platforms like Twitter (rebranded as X) offer promising channels for dissemination, yet research organizations lack clear guidance on establishing a professional social media presence. We present a structured framework based on our research network's multiyear experience developing a Twitter account for research dissemination.

Objective: This study aimed to provide a roadmap for organizations aiming to create a professional Twitter account for research dissemination.

Methods: This was a mixed methods study analyzing the Pediatric Emergency Care Applied Research Network (PECARN) Twitter team's 4-year experience (2020-2023) with building a social media account. Using the nominal group technique qualitative approach, we recorded insights from the 6 team members' experiences in a round-robin fashion until response saturation. In addition, we analyzed internal Slack (Slack Technologies) communications to identify key developmental events. Together, these were then prioritized by consensus to elucidate key developmental events that enhanced both social media and scientific engagement. This process was informed by quantitative data from Twitter performance metrics and Altmetric Attention Scores for journal publications collected over a 39-month period. Together, these elements informed the design of a logic model framework.

Results: The nominal group technique generated 63 thematic statements which included issues such as organizational structure, content strategy, technologies, analytics, organizational priorities, and challenges. These statements coalesced into the 7 domains (priorities, assumptions, inputs, outputs, outcomes, and external factors) that comprise the logic model. Inputs included organizational support (eg, executive-level champion and funding), specialized personnel (eg, content writer and analytics manager), and operational technologies (eg, communications and data analytics tools). Outputs encompassed targeted activities, such as engaging with other Twitter accounts, publishing high-quality tweets highlighting scholarly work, and developing a dynamic operations manual for the Twitter team. Outcomes were measured through tweet metrics, account analytics, and article-level impact scores.

XSL•FO
RenderX

Conclusions: Our logic model roadmap, based on our practical multiyear experience and data-driven strategies, can serve as a guide for research organizations or medical institutions aiming to incorporate Twitter or other social media platforms for research dissemination.

(JMIR Form Res 2025;9:e59481) doi:10.2196/59481

KEYWORDS

social media; research dissemination; knowledge translation; twitter; professional social media account; medical social media; professional twitter; medical education; clinical practice; academic medicine; mixed methods study; qualitative approach; quantitative data; logic model; medical institution; data-driven

Introduction

Despite technological advances dramatically increasing the rapidity with which new information becomes available, studies demonstrate that there is a significant lag between the publication of new research and its adoption into clinical practice [1,2]. Social media holds the potential to reduce this lag by promptly informing clinicians and policy makers about the latest practice-changing literature [3,4]. In particular, Twitter (now rebranded as X, though "Twitter" is used in this article given its widely recognized name) has over 600 million monthly active users [5] and has gained traction across clinical subspecialties, journals, and academia for research dissemination [6-8]. Having a Twitter presence is becoming increasingly important as a means to reach target audiences and affect bibliometric markers of research impact. As academic Twitter usage has increased, studies have identified characteristics that make tweets more effective from a research dissemination perspective, such as the incorporation of graphics and campaigns coinciding with target events [9-11]. While numerous studies have outlined content strategies to boost engagements and impressions in research dissemination, there is limited guidance for research organizations and medical institutions on the key steps to building and maintaining a professional Twitter account for research dissemination [12-14], such as the personnel needed to manage a social media account and technical resources required in addition to the platform itself.

The Pediatric Emergency Care Applied Research Network (PECARN) is a US-based, federally funded, multi-institutional consortium specializing in pediatric emergency research, with over 200 peer-reviewed publications on improving pediatric emergency care. PECARN is funded by the Health Services and Resources Administration, a governmental agency dedicated to improving access to health care, particularly in vulnerable communities. In 2019, PECARN renewed federal funding with an added focus on information dissemination. In partnership with Academic Life in Emergency Medicine (ALiEM), an education organization with expertise in social media–based scholarship and dissemination, a joint decision was reached to launch a professional Twitter account (@PECARNteam) for showcasing PECARN's publications. This led to the formation of the PECARN Twitter team, which began posting tweets in 2020. The team was ultimately composed of 8 members including 2 content writers, 2 peer reviewers, 2 account monitors, an analytics manager, and a graphic designer. Drawing from the team's experiences along with the best practices from the business, marketing, and publishing sectors, we launched and grew the professional account through iterative design.

In this report, we share the PECARN Twitter team's experiences, reflections, and analytics over a 4-year period (2020-2023) to establish a professional social media presence for a national research organization. We identify key components for the creation and maintenance of an organizational social media account, supported by quantitative data from the study period. Our objective is to offer a structured roadmap in logic model format [15] for research organizations, national societies, and medical institutions aiming to accelerate research dissemination using a professional Twitter account.

Methods

Overview

We used a convergent mixed methods design to encapsulate the Twitter team's experience during 2020-2023 [16]. For the qualitative strand, we used the nominal group technique (NGT) to elicit as many opinions and recollections of key events, developments, and challenges, from each Twitter team member; our goal was to build consensus among the Twitter team members about which essential steps were undertaken to build the PECARN Twitter account [17]. In addition, we analyzed our team's 4-year Slack (Slack Technologies) communication history to extract significant events and decisions, ensuring no major events were overlooked. For the quantitative strand, we performed an analysis of Twitter metrics and the Altmetric Attention Score. We combined these data sources to create a logic model, which outlines a roadmap for institutions and research organizations aiming to build a professional Twitter account for research dissemination.

Ethical Considerations

This study received exempted approval from the Institutional Review Board at Children's Hospital Los Angeles. An information sheet described the research purpose and NGT. Participants understood that their identities, which were publicly available on PECARN and ALiEM websites, could not be fully anonymized. They could email questions to the principal investigator before participating or opting out. Participation implied consent to using responses for the study. No participants were compensated for participating in the study. The NGT occurred on a secure internet platform, with transcripts stored on a password-protected laptop.

XSL•FO
RenderX

Qualitative Methods

NGT

The NGT is a consensus-building technique led by a facilitator, whereby a group of 5-9 participants answer a discussion prompt and then participate in a structured brainstorming session [17,18]. The purpose of our NGT was to compile a comprehensive list of the essential steps taken to build the PECARN Twitter account. In total, 6 core members of the PECARN Twitter team gathered on an online videoconference platform over the course of 2 hours with 2 external facilitators (ARS and TPC) with NGT expertise. The session omitted 2 recently added team members due to their limited involvement over the study period.

Our prompt was as follows: "Please describe, in as much detail as possible, the specific activities, events, and milestones that happened while creating the current PECARN Twitter account, including barriers encountered along the way." The language of the prompt was crafted through iterative feedback from 3 external qualitative researchers to ensure clarity, appropriateness, and scope. After being presented with the prompt, participants were first given 5 minutes to independently record their ideas. Subsequently, participants participated in a round-robin fashion, each sharing ideas in turn until all new ideas were exhausted. The final phase of the NGT entailed a group discussion allowing individuals to clarify any concepts or ideas. In contrast to typical NGT brainstorms, participants did not rank the ideas. Instead, we grouped related concepts together and condensed similar ideas into representative statements. The overarching themes laid the groundwork for what would eventually become the logic model.

Slack Analysis

Slack is a real time private messaging and collaboration platform employed by the Twitter team to assist with running the PECARN Twitter account. It served as the primary means of communication during our study. All conversations are saved to the platform in chronological order with dates and timestamps, providing an objective record of major developments. To avoid overlooking key events, 2 of the researchers who were part of the Twitter team reviewed the conversations and subthreads over the course of our study period to generate a timeline of events. The timeline was reviewed and revised several times with input from the Twitter team members. These events were then compared with the events recalled in the NGT. Events not mentioned in the NGT but felt to be crucial to the development of a social media account for dissemination, such as the incorporation of content creators and analytics managers, were then included in the development of our logic model. Using author consensus, events were deemed crucial if they led to a key development in the account, such as creating a new role fundamental to the execution of the account, initiation of a new activity that became routinely included in our content, or led to new guidelines surrounding the management of the account.

Quantitative Methods

Tweet-Level Analysis

We manually collected descriptive data on individual tweet characteristics during a 39-month period (May 2020-August 2023). Tweets are posts on the Twitter platform limited to 280 characters and can include graphics, videos, and links. We performed a multiple linear regression analysis (Microsoft Excel v16.81) using data from December 2020-August 2023 to determine whether 6 tweet characteristics influenced impressions (view count of a tweet) and engagement (user interactions with a tweet including retweets, likes, replies, poll votes, media views, mentions, and clicks on URLs or hashtags).

We excluded the May-November 2020 period because of the "honeymoon period" of starting a Twitter account, where the promotion and novelty surrounding a Twitter account generate a spike and wide variability in engagement [19]. In addition, this was the beginning of the COVID-19 pandemic, which may have altered the online presence of health care professionals. Based on existing marketing literature [12,13,20], 6 independent variables—account tagging, emojis, graphics (image or video), hashtags, polls, and URL links—were selected. Each variable was binary and coded with a 0 (absent) or 1 (present) for each tweet. All 6 independent variables were entered simultaneously in the multiple linear regression model as binary dummy variables. We then calculated standardized beta coefficients (β) post hoc using the formula: $\beta = b \times (SD_{predictor} \div SD_{outcome})$ where b is the unstandardized coefficient, $SD_{predictor}$ is the standard deviation of the tweet characteristic (binary variable), and $SD_{outcome}$ is the standard deviation of impressions or engagements. The $SD_{predictor}$ was derived from tweet characteristic frequencies using for binomial distributions. Using a standardized beta coefficient–enabled direct comparison of effect sizes across variables measured on different scales.

Before this analysis, we conducted a multicollinearity check using a correlation matrix (Multimedia Appendix 1). We also used Seaborn to assess linearity and homoscedasticity [21]. These analyses confirmed that our independent variables met the necessary assumptions required for linear regression.

Twitter Account Analysis

We monitored monthly account-level data for follower counts from May 2020 to August 2023, using the Twitter analytics services and the Fedica software. We also assessed follower demographics by manually categorizing the Twitter followers by profession (any physician, emergency physician, emergency medical service [EMS] provider, other health care worker, non–health care worker, or medical organization) as well as domestic or international. The follower list was downloaded in April 2021 and October 2022 by 5 PECARN research associates, with codification ambiguities adjudicated by consensus among the team.

Altmetric Attention Score Analysis

The Altmetric Attention Score, a newer bibliometric measure in academia to complement traditional citation counts, gauges a journal article's digital reach and attention [22]. Although

XSL•FO
RenderX

tweet mentions of the journal article correlate with increased citation counts [13,22-31], citation counts experience a publication lag. The Altmetric score instead is a weighted, composite, real-time measure of the amount of attention a publication has received through various outlets, including Twitter, blogs, news media, and policy documents [22,32]. The Altmetric website asserts that tweets mentioning a journal article's Digital Object Identifier (DOI) or PubMed URL will positively raise the Altmetric score, although the exact calculation is not publicly available. To assess this association between our Twitter activity and a journal article's attention, we conducted an exploratory analysis of the Altmetric score over a 9-month period (August 2020-April 2021). Specifically, for each PECARN publication featured on our Twitter account, we manually recorded the Altmetric score directly from the Altmetric website at 2 specific time points, that is, Sunday immediately before posting the Monday tweet about the article and again the next Sunday (1 week later). This before-and-after approach allowed us to capture any changes in the Altmetric score potentially attributable to our tweets. These scores were publicly available and free on the Altmetric website only for the current day.

Logic Model

A logic model is a visual way to present the relationships and resources required to execute a program [33]. We derived a logic model using abductive reasoning based on the qualitative NGT and quantitative analytic data to illustrate the relationships between inputs (essential resources), outputs (activities and audience), and outcomes (direct result of activities) in managing a professional Twitter account for research dissemination [15,34]. We cross-referenced our final logic model with the W.K. Kellogg Foundation Logic Model Development Guide's checklists to ensure comprehensive coverage and internal consistency of all essential elements in the framework [33]. To strengthen the model, we examined our underlying assumptions and unforeseen external factors. Because of the complex, intersecting relationships of the NGT data, any concepts that could be categorized into more than 1 logic model domain were adjudicated by the authorship team.

To enhance generalizability, the final proposed logic model maintains the core elements of the original framework but uses broader terminology. For logic model elements that were not directly addressed in the NGT, such as personnel or audience, we referenced our internal Slack communications and the PECARN website.

Results

Overview

Overall, the NGT generated 63 unique thematic statements, which addressed issues such as organizational structure and management, content strategy and outreach, technologies, analytics, organizational priorities, and challenges. These concepts were grouped based on the logic model domains (Textbox 1). Of note, some statements were slightly reworded in Textbox 1 to protect proprietary and private content.

We then transformed our project-specific findings into a more generalizable logic model to provide professional organizations with a roadmap for research dissemination using Twitter (Figure 1). For instance, external factors beyond our organizational control that impacted the Twitter campaign included the COVID-19 pandemic and Elon Musk's acquisition and rebranding of the platform in 2022. These 2 specific scenarios were generalized into "world events" and "changing social media landscape and platform dynamics." The following paragraphs provide a deeper exploration of the assumptions, inputs, outputs, and outcomes with each section outlining our general recommendations and our team's experiences.

XSL•FO
RenderX

Textbox 1. Nominal group technique concepts categorized into the logic model domains with each unique concept designated by a unique letter code.

Situation

- Pediatric Emergency Care Applied Research Network (PECARN) contracts Academic Life in Emergency Medicine (ALiEM) for social media management (M)
- Integration of interdisciplinary and intergenerational members (BD)
- PECARN Champion (BE)
- Team lead with expertise and connections (AI)
- Creation of a dissemination working group (F)
- Securing funding for dissemination (AO)
- Recruitment of Twitter-savvy users to the team (A)
- P3 Project (1)

Priorities

- Preemptive education on responding to negative feedback (O1)
- Navigating nodes with equity or balance (AR)
- Getting the word out within PECARN equitably (BF)
- Presentations to the PECARN Steering Committee including updates to increase awareness (BG)
- Encouraging researcher engagement (AT, V1)
- Becoming aware of new publications (AS)
- Anticipating events or bursts of demand (AE)
- Establishing credibility on Twitter (AP)

Assumptions

- Peer review of tweets for credibility or trustworthiness (H)
- No gold standard for analytics makes measuring impact or effectiveness difficult (N)

Inputs

- Incorporation of best Twitter practices from the marketing world for visibility and engagement (AQ)
- Working with and managing expectations from stakeholders (W)
- Balancing researcher involvement (B, BH)
- Organized structure and systematic review process for creating and editing tweets before publication (E)
- Survey to authors for impending publications (AY)
- Having multiple people making content to cover for each other (AL)
- Reach out to authors for live tweeting (BC)
- Regular meetings, feedback incorporation, and strategic development (R)
- Acquisition of external graphics designer (G)
- Selection of software and metrics (J)
- Learning to work with existing technologies for scheduling and publishing (AF)
- Sudden unexpected team member change (U)

Outputs

- Extracting key teaching points from top PECARN articles into tweet (X)
- Tweet frequency established at the same time each week (3 times per week) (AH)
- Tweet visual structure: Intentional or strategic polls, emojis, bullet points, and multimedia such as images and videos (AA, AB, AC, C, D)
- Linking to open-access and other websites, blogs, and source material (AZ)
- Threading continuous tweets (BB)
- Use of hashtags including tagging other existing conferences for more visibility (AN)

XSL•FO
RenderX

- Researcher profile highlights including narrative story to provide sufficient content (BA)
- Priority list of timely content for tweets such as newly published PECARN studies with icons to assist readers (AG)
- Iterative operation manual with protocols and policies to align with stakeholders and unexpected obstacles (S, T, Y)
- Developing approach to unexpected events and outside-of-scope requests (O, AD, AM, AU)

Outcomes

- Development of analytic metrics with dashboard monitoring (Q)
- Increasing and broadening followership for Twitter account (AK, AX, K, P)
- Tracking pre- and post-Altmetric scores following tweets (AJ)
- Understanding how PECARN account interacted with other accounts (AV)
- Researching and categorizing follower characteristics to assess who we were reaching (AW)

External factors

- COVID-19 pandemic at launch led to concerns over delay (L)
- Flexibility and responding to events occurring on Twitter (Y1, AE1)

Figure 1. Proposed logic model for implementing a professional organizational Twitter account for research dissemination.

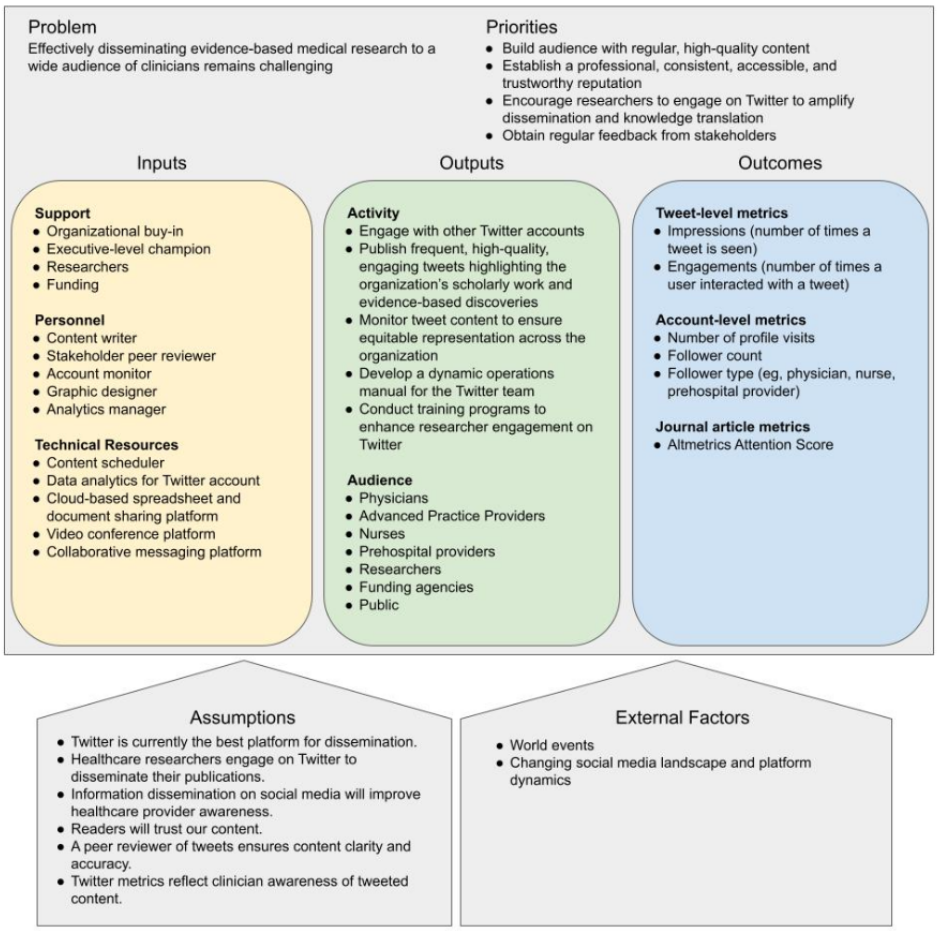

Assumptions

Recommendation

Organizations creating a social media account for research dissemination must assume their chosen platform is widely used by the target audience and that sharing content will effectively disseminate research. They must also presume that selected metrics serve as the best surrogate measures reflecting dissemination and clinician awareness, despite the inherent challenges in measuring social media impact.

PECARN Experience

In building our logic model, we made several assumptions. We prioritized Twitter over other popular social media platforms

XSL•FO
RenderX

because Twitter has been a dominant platform in academic medicine, used by health care professionals, researchers, institutions, organizations, and academic journals [6-8,12]. The platform has many advantages for dissemination including accessibility, reach, conversational tone, and communal nature which allows for a dynamic exchange of ideas within the medical community. Twitter's format, with its 280-character limit, makes it easily digestible and well-suited for busy health care professionals.

Another assumption we made was that a pre-tweet peer review process ensured the dissemination of trustworthy, unambiguous, and accurate content. In the era of misinformation, this vetting process ensured our tweets could better withstand academic scrutiny and aligned with the values of PECARN. We also assumed our use of a formal Twitter handle which included our organization's name (@PECARNteam), prominent academic publications, and professional tone would engender trust.

Finally, we assumed the Twitter metrics of impressions and engagement were the best surrogate measures to reflect clinician awareness of the PECARN articles and their application toward patient care. We also explored the impact our tweets may have had on each featured article's Altmetric Attention Score as another indirect measure of impact that organizations may use.

Input: Support

Recommendation

Securing endorsement from leaders and stakeholders is a fundamental starting point. Stakeholders may include researchers, an internal communications department, funding organizations, and institutions that have a vested interest in the research being disseminated. Ideally, this includes an executive-level champion who can advocate for the social media team.

PECARN Experience

In 2017, a senior PECARN researcher met with the founder of an external organization, (ALiEM), who had over 10 years of Twitter experience in the health professions education domain. Their interests aligned when PECARN's federal funding increasingly valued research dissemination since ALiEM's mission was to accelerate knowledge translation using social media. Continued discussions ultimately led to a funded partnership whereby ALiEM would serve as PECARN's Twitter account manager, with peer review oversight by designated PECARN members.

Input: Personnel

Recommendation

In total, 5 integral roles form the foundation of a Twitter research dissemination team. A content writer crafts tweets and schedules content for publication. This person also serves as the team lead, collaborating with the peer reviewer (verifies tweet accuracy and ensures alignment with organizational mission), graphic designer (optimizes visual elements), account monitor (engages with other Twitter accounts), and analytics manager (monitors and reports account metrics). All roles should ideally have previous Twitter experience.

PECARN Experience

In building our 8-member Twitter team, deliberate consideration was given to the group composition, reflecting a strategic blend of expertise and flexibility. Of the total, 2 were content writers, who were physicians with an established Twitter presence focusing on emergency medicine and pediatric education. The decision to enlist 2 content writers was intentional, as they had full-time physician jobs competing for their time. Having duplicate writers ensured consistency despite scheduling constraints and individual availability. The 2-writer model served as a safeguard against disruptions, providing a safety net for when a content writer departed the project, requiring the rapid onboarding of a new writer. This redundancy ensured project momentum, content quality, and institutional memory.

For similar reasons, duplicate roles were also created for the peer reviewer and account monitor roles. Peer reviewers, selected from the PECARN organization's Dissemination Working Group, ensured tweet quality and accuracy. The reviewers also ensured that the studies highlighted by the account equitably represented all of the researchers within the organization and that tweets aligned with PECARN's mission and values. Account monitors were recruited based on Twitter and research dissemination expertise. They surveilled the Twitter account daily for relevant mentions from other Twitter accounts and replying to, liking, or retweeting their content. A dedicated graphic designer was added 2 years after the launch of the Twitter campaign. This person played a pivotal role in enhancing the visual appeal of the tweets, especially given evidence showing the importance of graphics in promoting engagement [35]. Finally, the analytics manager reported data-driven insights on a monthly basis that informed changes and decisions.

Input: Technical Features

Recommendation

Digital tools are necessary to streamline team communication, documentation, analytics tracking, tweet scheduling, and Twitter account monitoring. These tools should be cloud-based to ensure version control and facilitate asynchronous collaboration among team members, who may be geographically distributed with different work schedules.

PECARN Experience

Our Twitter team used digital tools that optimized workflow, efficiency, collaboration, and data-driven decision-making. Slack served as our communication platform, proving more effective than traditional email channels by facilitating real-time collaboration, quick decision-making, and seamless information exchange among team members. Google Sheets (Alphabet Inc) served as a repository for both past and upcoming tweets. This spreadsheet included details such as the responsible team member for each tweet and its current stage (awaiting peer review, graphic design, or publication). The accessibility of this document to the entire team ensured transparency and collaborative tracking of tweet progress. Google Docs (Alphabet Inc) allowed a centralized record of monthly meeting minutes and action plans. In addition, this platform was used to craft an operations manual and style guide. Fedica, a social media management platform, provided data analytics and allowed

XSL•FO
RenderX

tweet scheduling. Zoom (Zoom Communications Inc) was the platform for hosting monthly videoconference meetings, allowing for real-time discussions while fostering a sense of community and accountability among team members across locations. Finally, the Twitter mobile app (X Holdings Corp) offered account managers the flexibility to promptly respond to engagements from other Twitter accounts.

Output: Activity

Recommendation

Establishing goals for social media accounts allows for a focused approach in content creation. Growing a following on a professional Twitter account requires a trustworthy, consistent, professional, and engaging presence. Tweets should adhere to the organization's established content focus and writing style, as well as follow Twitter marketing best practices. Content should be monitored to ensure equitable representation of all individuals and subgroups within the organization. A regularly updated, cloud-based operations manual facilitates easy guideline modifications and streamlines the onboarding process for new team members. To grow an online community and facilitate discussions with intended audiences, conduct training sessions to encourage researchers within the organization to engage directly on Twitter. Iterative evaluation of outcomes data and stakeholder feedback should regularly guide the reassessment of these activities.

PECARN Experience

Our overarching goals were to use Twitter to increase awareness of new PECARN publications, provide ongoing education about established PECARN research, introduce evidence-based decision aids for providers to incorporate into their clinical practice, and highlight the PECARN organization as well as its researchers. In the early stages of the PECARN Twitter account, we concentrated on growing our follower base to expand our visibility and reach. We tagged relevant entities, including journals, universities, funding agencies, educational blogs, and prominent physicians in the field.

In conjunction with growing an expanded audience, the Twitter team focused on creating a consistent tweet publication schedule focusing on highlighting both previous and recent PECARN studies. Our group decided to only tweet about full articles, rather than abstracts or preprints, after they were published to ensure the highest quality evidence. To identify new studies, we requested PECARN authors notify us of impending publications. We also configured a PubMed email as well as a Google Alert to notify us of PECARN-related articles. Acknowledging that our Twitter audience might not be aware of previous PECARN research, highlighting older studies was seen as valuable for research dissemination. This approach counters the common trend among journals and news media to concentrate on newer publications. It also allowed us to have a larger pool of content to choose from, allowing us to tweet consistently, even during periods when there were no new publications.

Informed by the medical literature, business insights, influencer marketing practices, and our data analytics, our tweet structure aimed to incorporate bullet points, emojis, graphics (image or video), URL links, and hashtags to maximize Twitter engagement and reach [12,13,35-38]. Using Twitter's threading feature (allowing for a series of connected tweets), each week focused on a single publication. Monday's tweet introduced the journal article, followed by threaded tweets over the next 5 days, highlighting key results and insights (Figure 2). Tweets were published at the same time of day for consistency.

As the account grew, new strategies were implemented to bolster engagement. In the eighth month after launching the Twitter account (December 2020), polls were incorporated, when possible, to introduce the featured research article of the week. Polls allowed readers to vote on a clinical scenario and see real time polling results. Subsequent threaded tweets reviewed salient teaching points and the significance of the featured journal article. In 2022, a PECARN graphic designer joined the team, incorporating professional graphics into our tweets, allowing for a more detailed description of the article, along with a layout, icons, color scheme, and typeface that better aligned with the PECARN brand. Notably, because the text was embedded in the graphic, we were not constrained to the 280-character tweet limit. During periods of fewer new PECARN publications, selected older studies that we featured on Twitter before 2022 were rehighlighted, incorporating new layout strategies such as professionally designed graphics.

Another goal of the account was the promotion of the PECARN organization and its researchers. To this end, we also monitored content for equitable representation across the organization. Each subgroup and its members had their own unique achievements and perspectives, meriting representation on the organization's official Twitter account. These tweets were not limited to publications and included national conference presentations, awards, and study achievements. We also occasionally created a series of tweets highlighting specific researchers and their interests.

As we developed our unique style and content strategy, we created an operations manual to help ensure continuity and accountability over time as team members changed and new members joined. This document outlined our tone, structure, content creation workflow, key resources, and general guidelines. It also included a conflict resolution plan, created after a negative reply to one of our tweets regarding concerns about distilling large studies into 280 characters. Furthermore, guidelines on boundary setting outlined which accounts to follow and retweet, as account activity is a direct reflection of the organizational brand. This continuously updated, cloud-based document maintained a shared understanding across time, team members, and stakeholders.

Because of the open community nature of the Twitter platform, researchers can also engage directly with the audience in addressing any questions or amplifying the PECARN tweet messages. Although many PECARN researchers were not active on Twitter at the time of the study, researchers who were active played key roles in retweeting or replying to conversations about their publication. Encouraging researchers to join Twitter through onboarding or training programs may further enhance the reach and visibility of their publications on the platform.

XSL•FO
RenderX

Figure 2. Example of a typical week's tweet thread discussing a PECARN publication. The first tweet in the series included an image, hashtags, emojis, and a URL link. Subsequent threaded tweets also included bullet points.

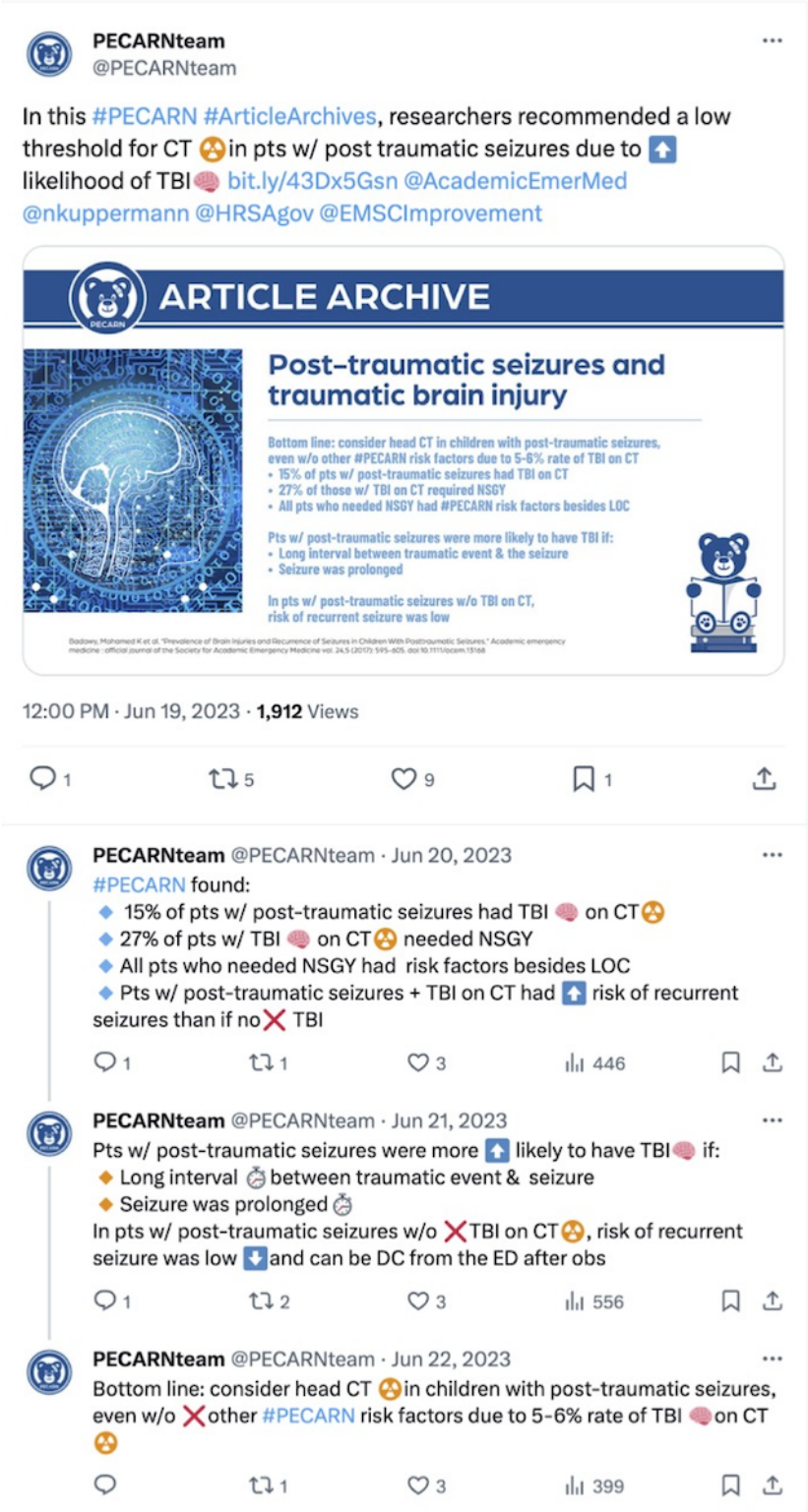

Output: Audience

Recommendation

Every tweet should be tailored with the target audience in mind. This may include health care providers, trainees, researchers, funding agencies, and the lay public.

PECARN Experience

The PECARN organization generally aims to reach physicians, nurses, researchers, prehospital providers, and funding agencies with interest or expertise in emergency medicine, pediatrics, and pediatric emergency medicine. Their priority focuses on the US emergency physicians, which includes adult and pediatric emergency physicians as well as their trainees, and EMS providers. Although some tweet content may have been relevant

XSL•FO
RenderX

to the lay public, the primary goal was to educate other health care professionals. Thus, tweets were crafted assuming some foundational knowledge of medicine and critical appraisal of clinical research. We distilled each article into relevant key points that might impact practice.

Outcomes: Tweet-Level Metrics

Recommendation

Impressions and engagements are commonly used to measure a tweet's performance. Additional metrics may be helpful to monitor progress in reaching an organization's objectives. Native Twitter analytics and third-party data services can provide valuable real-time feedback on tweet performance and audience engagement.

PECARN Experience

During May 2020-August 2023, the @PECARNteam Twitter account published 569 tweets, highlighting 99 PECARN publications. We piloted different tweet formats and made data-driven decisions to incorporate at least 1 of the following in each tweet: account tags, emojis, hashtags, graphics, polls, and URL links. A descriptive summary of tweet characteristics and their resultant median number of likes, retweets, replies, impressions, and engagements are reported in Table 1. Notably, polls demonstrated the highest median number of impressions and engagements. Each month, we monitored the best-performing tweets for impressions and engagements, because they provided insights on topics that resonated most with the audience. The median number of clicks for URL links of our featured journal article was 7 (SD 14.9).

We also evaluated whether the above 6 tweet characteristics had an impact on impression and engagement metrics using multiple linear regression analyses. Our model revealed an adjusted R^2=0.22 (impressions) and 0.26 (engagements). Table 2 reports the standardized regression coefficient scores (β), revealing 3 tweet characteristics associated with a significant increase on both impressions and engagements—polls, graphics, and URL links. Notably the inclusion of emojis was associated with a decrease in impressions and engagements.

Table 1. Tweet-level analytics associated with the presence of a tweet characteristic (May 2020-August 2023).

Tweet characteristic (number of tweets)	Likes, median (SD)	Retweets, median (SD)	Replies, median (SD)	Impressions, median (SD)	Engagements, median (SD)
Account tagging (n=295)	6 (7.3)	3 (4.9)	1 (0.7)	1008 (2043.3)	29 (58.6)
Emoji (n=347)	3 (6.1)	1(3.8)	1 (0.6)	718 (1697.6)	17 (53.8)
Graphic (n=158)	9 (8.0)	4 (5.3)	1 (0.7)	1608 (2300.1)	47 (74.0)
Hashtag (n=275)	7 (7.8)	3 (5.4)	1 (1.2)	1011 (2155.8)	31 (94.0)
Poll (n=50)	3 (2.4)	2 (1.7)	0 (1.0)	1793 (1733.0)	55 (82.6)
URL link (n=188)	7 (8.1)	3 (5.0)	1 (0.6)	1338 (2272.4)	38.5 (73.2)

Table 2. Standardized regression coefficient scores (β) from multiple linear regression analysis for engagement and impressions as influenced by various tweet characteristics for the 461 tweets during December 2020-August 2023.

Tweet characteristic	Impressions (n=461)		Engagements (n=461)	
	Standardized regression coefficient scores (β)	*P* value	Standardized regression coefficient scores (β)	*P* value
Poll	0.243[a]	<.001	0.278[a]	<.001
Graphic	0.225[a]	<.001	0.195[a]	<.001
URL Link	0.198[a]	<.001	0.173[a]	<.001
Hashtag	0.102[a]	.05	0.053	.18
Account Tagging	0.034	.51	–0.018	.64
Emoji	–0.127[a]	.01[a]	–0.073	.06

[a]Coefficient scores with *P*<.05.

Outcomes: Twitter Account-Level Metrics

Recommendation

Tracking longitudinal quantitative data for an organization's Twitter account can help to identify strategies for growth. Common account-level metrics include follower insights (eg, follower count and demographics), monthly and overall impression and engagement counts, and profile visits (number of times the Twitter profile page is viewed).

PECARN Experience

We focused primarily on trending the follower count, which grew consistently on a monthly basis (Figure 3). In April 2021 and later October 2022, we analyzed follower demographics to assess if we were reaching our target audience (Table 3).

XSL•FO
RenderX

Emergency physicians comprised 47.7% (636/1332; April 2021) and 48.6% (927/1908; October 2022) of the followers. EMS providers consistently represented a small fraction of followers, accounting for only 1.7% (22/1332; April 2021) and 1.2% (23/1908; October 2022).

Figure 3. Total number of @PECARNteam Twitter account followers by month (May 2020-August 2023).

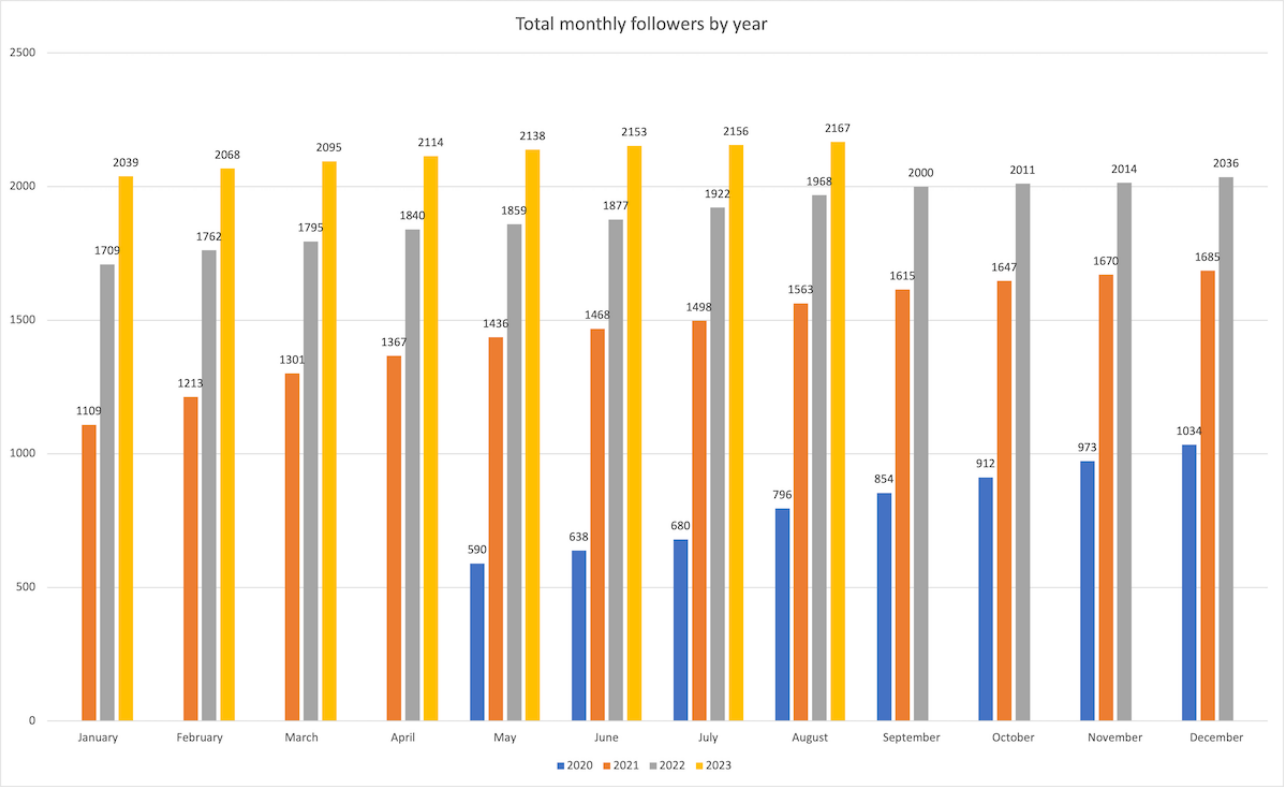

Table 3. Demographic distribution of the @PECARNteam Twitter account followers calculated in April 2021 and October 2022.

Twitter follower demographic	Year 2021 (n=1332), n (%)	Year 2022 (n=1908), n (%)
Health care worker		
All physicians	933 (70.0)	1351 (70.8)
Emergency physicians[a]	636 (47.7)	927 (48.6)
Emergency medical services providers	22 (1.7)	23 (1.2)
Other health care workers[b]	83 (6.2)	137 (7.2)
Non–health care workers	62 (4.7)	30 (1.6)
Medical organizations	76 (5.7)	116 (6.1)
Unidentified	156 (11.7)	251 (13.2)
Based in the United States	793 (59.5)	1158 (60.7)

[a]Emergency physicians included pediatric emergency physicians, general emergency physicians, and emergency medicine residents.

[b]Other health care workers included nurses, advanced practice providers, and medical students.

Outcomes: Journal Article Metrics

Recommendation

Bibliometric data, such as the Altmetric Attention Score, should also be considered for tweets highlighting journal publications. In conjunction with traditional metrics such as journal publication citation counts, this score may serve as a complementary bibliometric and more real time, multisource reflection of online reach and attention.

PECARN Experience

Our 9-month exploratory analysis of the pre-post Altmetric Attention Scores demonstrated a median positive increase of 4.0 (range 0-24, SD 5.7) for the 41 consecutively featured PECARN journal publications (Multimedia Appendix 2). All but one of the tweets mentioning a journal article's DOI or PubMed URL were associated with a nonzero increase in score. The single exception case resulted in the same pretweet and posttweet Altmetric score of 8.

XSL•FO
RenderX

Discussion

Principal Findings

Recent trends in academia and funding priorities emphasize the need for research dissemination extending beyond traditional journal publications, with a focus on reaching target audiences directly. Twitter and other social media platforms have emerged as potential tools to raise clinician awareness of the latest research. Unfortunately, there is limited literature to provide guidance for research organizations and medical institutions wishing to create an effective research dissemination campaign using social media [7,8,12,13].

This study presents a comprehensive roadmap for research organizations and medical institutions aiming to establish an effective Twitter presence for research dissemination. This roadmap, developed through a mixed methods approach, combining qualitative insights from our 4-year experience and quantitative data analysis, offers a structured framework in the form of a logic model. Our approach was particularly well-suited to the PECARN organization, which releases numerous practice-changing publications each year, such as the PECARN pediatric head trauma clinical decision tool [39]. This prolific output of high-quality studies allowed us to feature 99 unique publications over the study period, providing a rich dataset for our analysis and framework development.

The logic model we propose illustrates a complex, interconnected structure for organizations hoping to launch and maintain a social media account. Key inputs include organizational support, dedicated personnel, and technical resources. Our experience underscores the importance of a diverse team—including content writers, peer reviewers, account monitors, and graphic designers—to create and maintain a professional social media presence. The model also highlights the significance of consistent, high-quality content creation and strategic audience engagement as essential outputs. Importantly, our framework recognizes the intricate relationships between these elements, where inputs influence outputs, which in turn affect outcomes. These outcomes, such as engagement metrics and follower growth, provide valuable feedback that informs iterative refinement of our strategies. This continuous loop of implementation, assessment, and adjustment helps ensure adaptability in the dynamic social media landscape.

Our quantitative analysis yielded several key insights supporting the logic model framework, with impressions and engagement serving as our primary tweet-level outcome measures. Multiple regression analysis examining 6 tweet characteristics demonstrated meaningful explanatory power for both metrics.

Furthermore, 3 tweet elements—polls, graphics, and URL links—showed a stronger and statistically significant relationship with both impressions and engagements. The inclusion of polls demonstrated the strongest positive relationship, suggesting their effectiveness in stimulating audience attention and participation. This corroborates marketing strategies and recommendations [40,41]. This could be attributed to the ease of anonymous participation, the opportunity to engage with potentially controversial topics, and the instant gratification of immediate results. Graphics also showed a significant positive impact, likely due to their visually appealing nature and alignment with Mayer's [42] multimedia principle, which states that people learn better from words and visuals juxtaposed together. This particularly reinforces the importance of including a graphic designer on the social media team, as outlined in our logic model. The inclusion of URL links, usually to journal articles, was also impactful, likely because it provided a primary reference for critical appraisal.

Interestingly, our analysis revealed that emojis had a negative association with impressions, contradicting some previous findings in social media research [38]. This may reflect the relatively small sample size of the regression analysis or possibly even the preferences of our specific audience of health care professionals, who might favor a more formal tone in academic communications. This suggests that strategies applicable in general social media marketing may not always apply to professional and academic contexts.

Thus, our current recommended tweet structure is to prioritize the inclusion of at least 3 tweet characteristics (polls, graphics, and URL links) to full articles or resources. Despite hashtags and account tagging being weakly significant in our study, current marketing recommendations support their use [43-45]. These 2 characteristics should be considered for inclusion depending on one's audience, content, and organizational priorities. Hashtags and account tagging specifically may garner the attention of a community, field expert, or influencer who can amplify tweets to new audiences. Emojis could also be incorporated cautiously after careful consideration of the target audience given previous research supports their use. The evolution of the tweet structure should be informed by continued monitoring of Twitter analytics.

Our regression model accounted for 22%-26% of the outcome variance. Although this may seem a very modest association, it actually represents substantial explanatory strength within social science research contexts. Digital engagement patterns on platforms like Twitter are inherently influenced by numerous factors outside researcher control—algorithmic content prioritization, fluctuating user attention spans, competing real-world events, and platform-specific interaction dynamics. For context, established health communication and social science research typically considers R^2 values of ≥0.1 as meaningful, recognizing that no statistical model can fully capture the complex interplay of individualized preferences, temporal factors, and contextual variables driving social media engagement behaviors [46].

Our R^2 values indicate that while specific tweet characteristics contribute significantly to performance, additional factors beyond our model—such as content relevance, subject matter alignment with audience interests, and timing—may also influence tweet success. This finding reinforces the cross-disciplinary principle that "content is king" in digital communication environments. Our steady follower growth throughout the 4-year study period validates this content-focused strategy. Importantly, our follower demographics reveal that emergency physicians constitute nearly half of our audience, confirming an effective reach of our targeted professional

XSL•FO
RenderX

community and suggesting a successful implementation of our dissemination approach.

We performed a 9-month, nonrandomized, exploratory, descriptive analysis of the Altmetric Attention Score for 41 consecutively featured journal articles, to confirm the Altmetric website's assertion that a Twitter mention is part of their composite formula [43]. We reasoned that if any article showed no change in Altmetric score following our tweet activity, it would cast doubt on the reliability of Altmetric scores as a long-term outcome measure for our Twitter dissemination efforts. The observed increases in scores across all but one of the articles reinforced our confidence in using Altmetric data as one component of our evaluation strategy. Although the analysis was not designed to demonstrate a direct causal link, this analysis of both new and old PECARN publications supports Altmetric's attestation. However, more research needs to be done to determine the impact of tweets on Altmetric scores.

Our proposed logic model for research dissemination has significant implications for knowledge translation in the digital age. Knowledge translation, often described as bridging the "know-do gap," is the process of moving research findings into practical application and remains a significant challenge in health care [47,48]. Our findings align well with established frameworks in this field, particularly the knowledge to action (KTA) model and Rogers' diffusion of innovations theory [49,50].

The KTA model comprises 2 main components—knowledge creation and action cycle. Knowledge creation involves knowledge inquiry, synthesis, and the development of knowledge tools [49]. The action cycle includes identifying problems, adapting knowledge to a local context, assessing barriers, implementing interventions, monitoring knowledge use, evaluating outcomes, and sustaining knowledge use. Our Twitter-based approach addresses several of these components. It facilitates knowledge synthesis and tool development by presenting research findings in concise, visually appealing formats easily accessible to clinicians. The consistent growth in followers and engagement metrics we observed suggests that this method effectively supports the knowledge-creation phase. In addition, our approach contributes to multiple steps in the action cycle. The elements we used, such as polls, further support the implementation and monitoring stages of the action cycle by actively engaging health care professionals.

Our study also aligns with the diffusion of innovations theory, which describes how new ideas spread through social systems [50]. Twitter, as a social media platform, acts as a channel for diffusion, allowing for rapid dissemination of research findings. The elements we used, such as polls, align with the theory's emphasis on communication channels and social systems in the adoption of innovations. These elements potentially accelerate the diffusion process by engaging health care professionals and increasing their exposure to new research findings, thus facilitating the "knowledge" and "persuasion" stages of the innovation-decision process.

The effectiveness of visual content in our tweets underscores the crucial role of graphics in capturing attention, conveying complex information accurately, and summarizing content succinctly. This addresses a core principle in both the KTA model and diffusion of innovations theory—the need to present information in a clear, easily understandable format to promote its adoption. By translating complex research findings into digestible visual formats, we can potentially lower the barriers to adoption and increase the likelihood of research influencing clinical practice.

Several areas warrant further investigation. Longitudinal studies are needed to assess the long-term impact of social media dissemination on clinical practice changes. Exploration of other social media platforms for research dissemination could provide valuable comparisons to Twitter's effectiveness. In addition, investigating the optimal frequency and timing of posts for maximum engagement among health care professional audiences could further refine social media strategies.

In summary, our logic model and mixed method findings provide a foundational framework for organizations seeking to incorporate Twitter and potentially other social media platforms for research dissemination. While our study demonstrates the potential of this approach, it also highlights the complexities involved in effective social media engagement for academic purposes. As the digital landscape continues to evolve, ongoing research and adaptation will be needed to optimize the dissemination of research findings to health care practitioners.

Limitations

The methodology of our study faced certain limitations. While our logic model was crafted using qualitative and quantitative data with the PECARN Twitter account, there remains a possibility that critical elements for launching a professional Twitter account might have been overlooked. However, during our multiyear experience, we regularly received multisource feedback, likely identifying major challenges and uncovering hidden weaknesses.

Our experience was significantly influenced by 2 major external factors, potentially limiting the generalizability of our findings and recommendations. First, we launched our campaign just as the COVID-19 pandemic started in 2020. This period potentially enhanced our account's visibility because of the shift toward social distancing and more digital connectivity. Second, since the time of our study, Twitter has since been sold and rebranded as X. It is unclear how the new leadership of Twitter will impact academic social media going forward and if the platform will continue in the same manner as during the study period. If Twitter were to be replaced with a new platform for professional discourse among health care professionals, our logic model principles can likely be adapted for other platforms.

Traditionally, the NGT is conducted in person, but due to the locations of our participants, we had to adapt to an online approach. Although a precedent exists in the literature, this shift to a remote format may have influenced the dynamics and outcomes of the discussions [51,52]. In our study, we chose to exclude the rank-order component of the NGT, opting instead to classify the events and insights based on their relationships to one another. This decision was because different elements carry different weights at different points in time when launching

XSL•FO
RenderX

a professional social media account, thereby making ranking in order of importance irrelevant for our purposes.

Finally, 1 author (ML), the founder of the ALiEM organization (which was funded by the PECARN organization to manage the PECARN Twitter account), disclosed a potential conflict of interest. However, we mitigated this by having 2 external facilitators lead the entire NGT brainstorming session and identify the themes. Also, the collection and reporting of Twitter data was performed by the data analytics leader.

Conclusion

Social media plays an increasingly vital role in research dissemination. Our initiative highlights the value of a deliberately designed Twitter presence in expanding the reach and awareness of a research organization's scientific discoveries. Through practical experience, strategic planning, and data-driven iterative change, we propose a logic model roadmap designed to help other professional organizations in launching a social media campaign for research dissemination purposes.

Acknowledgments

We wish to thank Dr Nathan Kuppermann for his continued support, mentorship, and vision regarding implementation science. We also value the trust and support from the Pediatric Emergency Care Applied Research Network (PECARN) organization and specifically the Dissemination Working Group, allowing us to represent the national research network.

This work was performed by the PECARN Dissemination Working Group. PECARN is supported by the Health Resources and Services Administration (HRSA) of the US Department of Health and Human Services (HHS), in the Maternal and Child Health Bureau (MCHB), under the Emergency Medical Services for Children (EMSC) program through the following cooperative agreements: EMSC Data Center (EDC), University of Utah; CHaMP E-RNC; GLACiER; HOMERUN; PEM-NEWS; PRIME; SPARC; and STELAR. This information or content and conclusions are those of the authors and should not be construed as the official position or policy of, nor should any endorsements be inferred by HRSA, HHS, PECARN or the US government.

Data Availability

All data analyzed during this study are included in the main manuscript and supplementary appendices of this publication, except for Twitter analytic data (available from the corresponding author upon request).

Authors' Contributions

Conceptualization: GCH (lead), ML (supporting), ARS, TPC (qualitative methodology), JNM, JMW, SS, LV
Data curation: ARS, TPC (qualitative data); JNM, SS, LV, ML (quantitative data)
Formal analysis: ARS, TPC (qualitative analysis); ML, GCH (quantitative analysis)
Investigation: ARS, TPC (nominal group technique); LV (Twitter analytics); SS (Altmetric scores); ML, GCH (Slack communications)
Methodology: GCH (lead), ML (supporting), ARS (supporting), TPC (supporting)
Project administration: GCH (lead), ML (supporting)
Supervision: ML (lead), ARS (supporting), TPC (supporting)
Validation: ARS, TPC (qualitative results interpretation)
Visualization: GCH (lead), ML (supporting)
Writing – original draft: GCH (lead), JNM (supporting), JMW (supporting), SS (supporting), LV (supporting), ML (supporting)
Writing – review and editing: GCH, ML (equal); JNM, JMW, SS, LV (supporting); ARS, TPC

Conflicts of Interest

ML is the founder of Academic Life in Emergency Medicine (ALiEM), which was the organization funded by the PECARN organization to manage their Twitter account.

Multimedia Appendix 1
Correlation coefficient matrix for tweet characteristics, impressions, and engagement.
[DOCX File , 15 KB - formative_v9i1e59481_app1.docx]

Multimedia Appendix 2
Altmetric Attention Scores of 41 Pediatric Emergency Care Applied Research Network (PECARN) articles before and after tweet dissemination (August 2020-April 2021).
[DOCX File , 76 KB - formative_v9i1e59481_app2.docx]

References

XSL•FO
RenderX

1. de Wit K, Curran J, Thoma B, Dowling S, Lang E, Kuljic N, et al. Review of implementation strategies to change healthcare provider behaviour in the emergency department. CJEM 2018;20(3):453-460. [doi: 10.1017/cem.2017.432] [Medline: 29429430]
2. Densen P. Challenges and opportunities facing medical education. Trans Am Clin Climatol Assoc 2011;122:48-58 [FREE Full text] [Medline: 21686208]
3. Chan TM, Dzara K, Dimeo SP, Bhalerao A, Maggio LA. Social media in knowledge translation and education for physicians and trainees: a scoping review. Perspect Med Educ 2020;9(1):20-30 [FREE Full text] [doi: 10.1007/s40037-019-00542-7] [Medline: 31834598]
4. Gray JM, Schnadower D, LaFollette R, Shah AS, Sobolewski B. Going viral: A scoping review of the current state and impact of online research dissemination in emergency medicine. AEM Educ Train 2022;6(1):e10725 [FREE Full text] [doi: 10.1002/aet2.10725] [Medline: 35224409]
5. Singh S. How many people use Twitter (X) – active users [2025 stats]. Demand Sage. 2025. URL: https://www.demandsage.com/twitter-statistics/ [accessed 2025-05-01]
6. Cardona-Grau D, Sorokin I, Leinwand G, Welliver C. Introducing the Twitter impact factor: an objective measure of urology's academic impact on Twitter. Eur Urol Focus 2016;2(4):412-417. [doi: 10.1016/j.euf.2016.03.006] [Medline: 28723474]
7. Sterling M, Leung P, Wright D, Bishop TF. The use of social media in graduate medical education: A systematic review. Acad Med 2017;92(7):1043-1056 [FREE Full text] [doi: 10.1097/ACM.0000000000001617] [Medline: 28225466]
8. Jalali A, Sherbino J, Frank J, Sutherland S. Social media and medical education: exploring the potential of Twitter as a learning tool. Int Rev Psychiatry 2015;27(2):140-146. [doi: 10.3109/09540261.2015.1015502] [Medline: 25768325]
9. Roberts-Lewis S, Baxter H, Mein G, Quirke-McFarlane S, Leggat FJ, Garner H, et al. Examining the effectiveness of social media for the dissemination of research evidence for health and social care practitioners: systematic review and meta-analysis. J Med Internet Res 2024;26:e51418 [FREE Full text] [doi: 10.2196/51418] [Medline: 38838330]
10. Buettmann EG, Chlebek C, Lockard CA, Clayton SW, Lewis KJ, Collins KH. Post or perish? Social media strategies for disseminating orthopedic research. J Orthop Res 2023;41(8):1643-1652 [FREE Full text] [doi: 10.1002/jor.25588] [Medline: 37163368]
11. Yamamoto T, Goto K, Kitano S, Maeshima Y, Yamada T, Azuma Y, et al. Current insights on social media as a tool for the dissemination of research and education in surgery: a narrative review. Surg Today 2024;54(10):1113-1123. [doi: 10.1007/s00595-024-02891-1] [Medline: 38980332]
12. Bhatt N, Czarniecki S, Borgmann H, van Oort IM, Esperto F, Pradere B, EAU Guidelines Office Dissemination Committee. A systematic review of the use of social media for dissemination of clinical practice guidelines. Eur Urol Focus 2021;7(5):1195-1204. [doi: 10.1016/j.euf.2020.10.008] [Medline: 33172773]
13. Erskine N, Hendricks S. The use of Twitter by medical journals: systematic review of the literature. J Med Internet Res 2021;23(7):e26378 [FREE Full text] [doi: 10.2196/26378] [Medline: 34319238]
14. Elliott SA, Dyson MP, Wilkes GV, Zimmermann GL, Chambers CT, Wittmeier KD, et al. Considerations for health researchers using social media for knowledge translation: multiple case study. J Med Internet Res 2020;22(7):e15121 [FREE Full text] [doi: 10.2196/15121] [Medline: 32706653]
15. Bickman L. The functions of program theory. New Directions for Program Evaluation 2004;1987(33):5-18 [FREE Full text] [doi: 10.1002/ev.1443]
16. Fetters MD, Curry LA, Creswell JW. Achieving integration in mixed methods designs-principles and practices. Health Serv Res 2013;48(6 Pt 2):2134-2156 [FREE Full text] [doi: 10.1111/1475-6773.12117] [Medline: 24279835]
17. McMillan SS, King M, Tully MP. How to use the nominal group and Delphi techniques. Int J Clin Pharm 2016;38(3):655-662 [FREE Full text] [doi: 10.1007/s11096-016-0257-x] [Medline: 26846316]
18. Delbecq A, Van DVA, Gustafson DH. Group Techniques for Program Planning: A Guide to Nominal Group and Delphi Processes. Glenview, IL: Scott Foresman; 1975.
19. How we used Twitter (now X) threads to increase impressions by 2417%. Fedica. 2022. URL: https://fedica.com/blog/how-we-used-twitter-threads-to-increase-impressions-by-2417/ [accessed 2024-01-31]
20. Nielsen J. F-shaped pattern for reading web content (original study). Norman Nielsen Group. 2006. URL: https://www.nngroup.com/articles/f-shaped-pattern-reading-web-content-discovered/ [accessed 2024-01-14]
21. Waskom ML. seaborn: statistical data visualization. JOSS 2021;6(60):3021. [doi: 10.21105/joss.03021]
22. Araujo AC, Vanin AA, Nascimento DP, Gonzalez GZ, Costa LOP. What are the variables associated with altmetric scores? Syst Rev 2021;10(1):193 [FREE Full text] [doi: 10.1186/s13643-021-01735-0] [Medline: 34187573]
23. Richardson M, Park W, Echternacht S, Bell D. Altmetric Attention Score: evaluating the social media impact of burn research. J Burn Care Res 2021;42(6):1181-1185. [doi: 10.1093/jbcr/irab026] [Medline: 33528573]
24. Floyd AR, Wiley ZC, Boyd CJ, Roth CG. Examining the relationship between Altmetric score and traditional bibliometrics in the pathology literature. J Pathol Inform 2021;12:8 [FREE Full text] [doi: 10.4103/jpi.jpi_81_20] [Medline: 34012712]
25. Smartz TM, Jabori S, Djulbegovic M, Watane A, Sayegh Y, Lyons N, et al. Correlation between Altmetric scores and citation count in 4 high-impact plastic surgery journals. Aesthet Surg J 2023;43(11):NP943-NP948. [doi: 10.1093/asj/sjad239] [Medline: 37474313]

XSL•FO
RenderX

26. Chi AJ, Lopes AJ, Rong LQ, Charlson ME, Alvarez RD, Boerner T. Examining the correlation between Altmetric Attention Score and citation count in the gynecologic oncology literature: does it have an impact? Gynecol Oncol Rep 2021;37:100778 [FREE Full text] [doi: 10.1016/j.gore.2021.100778] [Medline: 34036141]
27. Nocera AP, Boyd CJ, Boudreau H, Hakim O, Rais-Bahrami S. Examining the correlation between Altmetric score and citations in the urology literature. Urology 2019;134:45-50. [doi: 10.1016/j.urology.2019.09.014] [Medline: 31560915]
28. Demir HA, Dogan S. Correlation between academic citations in emergency medicine journals and Twitter mentions. Am J Emerg Med 2022;58:33-38. [doi: 10.1016/j.ajem.2022.05.027] [Medline: 35623181]
29. Luc JGY, Archer MA, Arora RC, Bender EM, Blitz A, Cooke DT, et al. Social media improves cardiothoracic surgery literature dissemination: results of a randomized trial. Ann Thorac Surg 2020;109(2):589-595. [doi: 10.1016/j.athoracsur.2019.06.062] [Medline: 31404547]
30. Klar S, Krupnikov Y, Ryan JB, Searles K, Shmargad Y. Using social media to promote academic research: identifying the benefits of Twitter for sharing academic work. PLoS One 2020;15(4):e0229446 [FREE Full text] [doi: 10.1371/journal.pone.0229446] [Medline: 32251463]
31. Altmetric Attention Score modifiers. Altmetric. 2023. URL: https://help.altmetric.com/support/solutions/articles/6000234288-altmetric-attention-score-modifiers [accessed 2024-01-14]
32. How is the Altmetric Attention Score calculated? Altmetric. URL: https://help.altmetric.com/support/solutions/articles/6000233311-how-is-the-altmetric-attention-score-calculated- [accessed 2024-01-13]
33. Logic Model Development Guide. W.K. Kellogg Foundation. 2004. URL: https://wkkf.issuelab.org/resource/logic-model-development-guide.html [accessed 2024-01-31]
34. Veen M. Creative leaps in theory: the might of abduction. Adv Health Sci Educ Theory Pract 2021;26(3):1173-1183 [FREE Full text] [doi: 10.1007/s10459-021-10057-8] [Medline: 34142300]
35. What fuels a tweet's engagement. Twitter blog. 2014. URL: https://blog.twitter.com/en_us/a/2014/what-fuels-a-tweets-engagement [accessed 2024-01-14]
36. McLaughlin J. The Anatomy of a Tweet - Twitter Marketing Tips. 2021. URL: https://www.brandcredential.com/post/the-anatomy-of-a-tweet-twitter-marketing-tips [accessed 2024-01-14]
37. Lua A. Can Twitter threads increase reach, engagement, and referral traffic' an experiment. Buffer. URL: https://buffer.com/resources/twitter-thread-experiment/ [accessed 2024-01-14]
38. McShane L, Pancer E, Poole M, Deng Q. Emoji, playfulness, and brand engagement on Twitter. Journal of Interactive Marketing 2022;53(1):96-110 [FREE Full text] [doi: 10.1016/j.intmar.2020.06.002]
39. Kuppermann N, Holmes JF, Dayan PS, Hoyle JD, Atabaki SM, Holubkov R, Pediatric Emergency Care Applied Research Network (PECARN). Identification of children at very low risk of clinically-important brain injuries after head trauma: a prospective cohort study. Lancet 2009;374(9696):1160-1170. [doi: 10.1016/S0140-6736(09)61558-0] [Medline: 19758692]
40. Berry R. 9 interesting examples of Twitter polls to boost engagement. SocialPilot. URL: https://www.socialpilot.co/blog/twitter-poll [accessed 2024-10-01]
41. Lake C. How one Twitter poll helped us double impressions, increase traffic, and land news coverage. Foundr. URL: https://foundr.com/articles/social-media/twitter-polls [accessed 2024-10-01]
42. Mayer RE. Multimedia Learning. Cambridge, England: Cambridge University Press; 2020.
43. Zote J. Twitter hashtags: How to find and use the right hashtags. Sprout Social. URL: https://sproutsocial.com/insights/twitter-hashtags/ [accessed 2021-06-10]
44. Mahapatra C. Using Twitter hashtags for growth in 2024: do they work? Blaze. 2023. URL: https://www.withblaze.app/blog/using-twitter-hashtags-for-growth [accessed 2024-10-01]
45. Radu A. SocialBee. 2024. URL: https://socialbee.com/blog/twitter-hashtags/ [accessed 2024-10-01]
46. Ozili, Peterson K. The acceptable R-square in empirical modelling for social science research. Social Research Methodology and Publishing Results. 2022. URL: https://ssrn.com/abstract=4128165 [accessed 2025-04-09]
47. About us. Canadian Institutes of Health Research. URL: https://www.cihr-irsc.gc.ca/e/29418.html [accessed 2024-10-01]
48. Bridging the 'Know-Do' gap: meeting on knowledge translation in global health. MEASURE Evaluation. 2005. URL: https://www.measureevaluation.org/resources/training/capacity-building-resources/high-impact-research-training-curricula/bridging-the-know-do-gap.pdf [accessed 2024-10-01]
49. Graham ID, Logan J, Harrison MB, Straus SE, Tetroe J, Caswell W, et al. Lost in knowledge translation: time for a map? J Contin Educ Health Prof 2006;26(1):13-24. [doi: 10.1002/chp.47] [Medline: 16557505]
50. Rogers E. Diffusion of Innovations. New York City: Simon and Schuster; 2003.
51. Smith D, Cartwright M, Dyson J, Aitken LM. Use of nominal group technique methods in the virtual setting: A reflective account and recommendations for practice. Aust Crit Care 2024;37(1):158-165. [doi: 10.1016/j.aucc.2023.09.004] [Medline: 37880060]
52. Khurshid F, O'Connor E, Thompson R, Hegazi I. Twelve tips for adopting the virtual nominal group technique (vNGT) in medical education research. MedEdPublish (2016) 2023;13:18 [FREE Full text] [doi: 10.12688/mep.19603.1] [Medline: 37484833]

XSL•FO
RenderX

Abbreviations

ALiEM: Academic Life in Emergency Medicine
DOI: Digital Object Identifier
EMS: emergency medical service
KTA: knowledge to action
NGT: nominal group technique
PECARN: Pediatric Emergency Care Applied Research Network

Edited by A Mavragani; submitted 12.04.24; peer-reviewed by M Barwick, L Li, S Zörgő, S Desroches; comments to author 19.08.24; revised version received 18.12.24; accepted 01.04.25; published 24.06.25.

Please cite as:
Hooley GC, Magana JN, Woods JM, Sivasankar S, VonHoltz L, Schmidt AR, Chang TP, Lin M
Research Dissemination Strategies in Pediatric Emergency Care Using a Professional Twitter (X) Account: A Mixed Methods Developmental Study of a Logic Model Framework
JMIR Form Res 2025;9:e59481
URL: https://formative.jmir.org/2025/1/e59481
doi:10.2196/59481
PMID:40554778

©Gwendolyn C Hooley, Julia N Magana, Jason M Woods, Shyam Sivasankar, Lauren VonHoltz, Anita R Schmidt, Todd P Chang, Michelle Lin. Originally published in JMIR Formative Research (https://formative.jmir.org), 24.06.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Video- Versus Text-Based Psychoeducation in Web-Based E-Mental Health Programs: Randomized Controlled Trial

Swantje Borsutzky[1], PhD; Josefine Gehlenborg[1], PhD; Lara Rolvien[1], PhD; Steffen Moritz[1], Prof Dr

Department of Psychosocial Medicine, University Hospital Hamburg-Eppendorf, Clinic for Psychiatry and Psychotherapy, Hamburg, Germany

Corresponding Author:
Swantje Borsutzky, PhD
Department of Psychosocial Medicine
University Hospital Hamburg-Eppendorf
Clinic for Psychiatry and Psychotherapy
Martinistraße 52
Hamburg, 20246
Germany
Phone: 49 040741056565
Email: s.borsutzky@uke.de

Abstract

Background: Mental health disorders affect 1 in 8 people worldwide, yet many face barriers to accessing care. E-mental health interventions, including self-guided internet-based programs, offer promising solutions. However, the mechanisms driving knowledge gain in such programs remain poorly understood. The role of medium, topic, sequence, and confidence and their interaction in learning outcomes need further investigation. Additionally, the influence of knowledge gaps on the outcome of psychoeducational intervention is not well understood (eg, whether psychoeducation requires an existing knowledge gap to be effective).

Objective: This randomized controlled trial investigated the role of medium, topic, sequence, and participants' initial knowledge levels on knowledge gain and confidence in fully automated self-guided e-mental health psychoeducation.

Methods: A total of 158 adults (mean age 34, SD 12.4 years; n=118, 74.7% female) were randomized to 8 experimental conditions (receiving video, texts, or both containing psychoeducational content on sleep or social competence; n=142) or a control group (neutral video; n=16). The fully automated interventions (videos) were developed for use in web-based e-mental health interventions. They address transdiagnostic symptoms and hence are relevant across various disorders. To assess the added value of video production for knowledge gain, text-based scripts corresponding to the video content were created and compared. All interventions and outcome assessments were delivered on the web via Qualtrics without face-to-face components. Pre- and postintervention knowledge was assessed using a validated 30-item knowledge test (true/false). Confidence in responses was rated on a 0% to 100% scale. Statistical analyses included 3-way ANOVA and multivariate ANOVA.

Results: Knowledge significantly increased across experimental groups ($F_{1,156}$=17.272; P<.001; η_p^2=0.10). Participants with social competence deficits had significantly lower baseline knowledge (P=.04; d=0.41). For sleep deficits, a nonsignificant trend emerged (P=.09; d=0.28). Participants with social competence deficits demonstrated greater knowledge improvement (t_{141}=7.12; P<.001; d=0.60). Participants with sleep deficits showed smaller but significant gains (t_{141}=2.43; P=.02; d=0.20). No significant differences in knowledge gain were found between video and text formats. Confidence in correct answers increased significantly in the experimental group (mean 42.82, 95% CI 41.15-44.50 to mean 51.67, 95% CI 49.28-54.04), with larger gains for social competence than sleep. Confidence in the control group remained unchanged.

Conclusions: Both video and text formats effectively facilitated knowledge gain in e-mental health interventions, with no clear advantage of one medium over the other. Participants with prior deficits learned more in areas where they initially lacked knowledge. Confidence in correct answers increased alongside knowledge, highlighting psychoeducation's role in promoting self-efficacy. Future research should explore multimedia integration to enhance adherence and symptom improvement.

Trial Registration: German Clinical Trials Register DRKS00026722; https://drks.de/search/en/trial/DRKS00026722

(JMIR Form Res 2025;9:e65478) doi:10.2196/65478

XSL•FO
RenderX

KEYWORDS

e-mental health; psychoeducation; knowledge gain; unguided interventions; sleep disorders; social competence deficits; confidence in learning; experimental conditions; transdiagnostic interventions

Introduction

Mental health disorders affect 1 in 8 people worldwide, with depression and anxiety being the most common conditions [1]. Notwithstanding effective prevention and treatment options (eg, cognitive behavioral therapy or medication), one-third of people with depression worldwide lack access to care, according to the World Health Organization [2]. Reasons for this treatment gap range from structural barriers (eg, long waiting lists to receive therapy due to a lack of psychotherapists, being in a remote location, or facing financial barriers) to personal factors (eg, fear of stigmatization, negative health beliefs, or a lack of awareness of treatment options) [3]. Failing to provide adequate treatment for mental disorders not only worsens the condition but also increases societal and economic costs [3], highlighting the urgent need for access to support. In response to this need, alternatives to traditional face-to-face psychotherapy, such as self-guided e-mental health programs, are gaining attention in mental health care. These programs offer flexibility and anonymity and therefore appeal to a broad audience [4,5]. Many self-guided e-mental health programs are based on psychoeducation, which involves providing structured information about a patient's disorder and treatment options. The term *psychoeducation* was first used by Anderson and colleagues [6] in a manual on schizophrenia that aimed to empower patients and their families by enhancing their understanding of the illness and improving their coping strategies [7]. Psychoeducation today covers various topics, from improving sleep to social competence skills, and it has shown promising results across different mental disorders, according to meta-analyses and literature reviews [8-13]. Transdiagnostic problems like sleep problems and poor social competence are particularly relevant in e-mental health programs, as they address symptoms that are common across multiple mental health disorders [14].

Knowledge gain as part of psychoeducation is essential in successful treatment because well-informed patients tend to benefit more from therapy [15]. In self-guided e-mental health interventions, which lack the direct interaction with therapists of face-to-face psychotherapy, adequate knowledge gain may not occur, thus potentially compromising treatment outcomes. So far, only a few studies have examined the mechanisms of knowledge gain that may lead to better therapy outcomes in face-to-face psychotherapeutic settings. Previous research [14] suggests enhancing the recall of therapy content to increase the effectiveness of psychoeducational interventions. Abramowitz and colleagues [16] found that patients with obsessive-compulsive disorder who received psychoeducation prior to psychotherapy better understood the rationale and were more compliant, suggesting that knowledge gain is a powerful mechanism of change in psychotherapy. Abramowitz and colleagues [16] suggested using quizzes to test knowledge and assumed that quantity (eg, attending all therapy sessions) would be less important than gaining knowledge about the treatment rationale. However, the extent of patients' knowledge about their symptoms generally and prior to therapy remains unclear and requires further investigation.

Research on methods of knowledge gain in therapeutic settings is minimal. In educational settings, effective ways to convey knowledge have been studied for decades. A meta-analysis of 51 studies found that students who took online classes outperformed students receiving face-to-face instruction, with an average effect size that was small (d=0.24) [17]. Moreover, online learning has no time or location limitations and thus increases accessibility. Recent meta-analytic evidence further supports these findings. For example, a meta-analysis by Ebner and Gegenfurtner [18] demonstrated that online learning formats, including webinars, can achieve comparable or even superior outcomes in terms of learning effectiveness when compared to traditional face-to-face formats. Another recent meta-analysis found that instructional in-class videos in educational contexts achieved stronger learning effects (d=0.69) than other formats such as outside-class videos, texts, or video games [19].

Two prevalent media for conveying knowledge in e-mental health interventions are videos and texts. In the framework of e-learning for clinical procedures, videos were more effective for learning outcomes than illustrated, text-based online instructions in a sample of medical students [20]. However, the literature on optimal media for knowledge gain in e-mental health programs is limited. Addressing this gap, Schlarb and colleagues [21] explored in their pilot study the acceptance and feasibility of online prevention and early intervention training for university staff with self-reported sleep issues. Participants found the videos helpful and entertaining and had improved insomnia symptoms with medium effect sizes. Further research suggests that, compared to text, information provided via video is more engaging and enhances content retention [22].

Researchers have also investigated judgment certainty, which is another moderator of knowledge gain. As early as 1977, Fischhoff and colleagues [23] examined people's tendency to be overconfident concerning their knowledge. Subsequent studies, such as one by Martí [24], have linked certainty to learners' performance and feedback. However, in educational settings, Dory and colleagues [25] found that medical students' confidence levels were not necessarily correlated with their test performance. To date, the role of judgment certainty on knowledge gain in clinical settings remains largely unexplored.

Objectives

One aim of this research project was to understand moderators of success in knowledge gain. Specifically, this study explored the role of the medium (text or video), topic (sleep or social competence), and sequence of presentation (ie, whether the medium changed or remained the same, that is, video to text vs text to text) in knowledge gain in low-threshold online experimental conditions targeting the topics of sleep and social competence. We also aimed to examine whether participants

XSL•FO
RenderX

with deficits related to sleep and/or social competence had knowledge gaps in these areas and whether this could be compensated for through targeted psychoeducation. Finally, we investigated whether knowledge gain would lead to an increase in high-confidence correct judgments and/or a decrease in overconfidence in incorrect judgments.

Methods

Participants

A total of 158 adults with a mean age of 34 (SD 12.4) years participated in the study. Most participants were female (n=118, 74.7%); the majority (n=88, 55.5%) had an education level above a high school degree.

To analyze knowledge gaps and improvements, participants were divided into 2 groups based on reported deficits: those with subjective social competence deficits (n=21) and those with subjective sleep deficits (n=48). Social competence deficits were defined as scoring below 60% on a self-report scale (0% to 100%); sleep deficits were based on a binary "yes" or "no" self-report. "Yes" responses assigned participants to the sleep deficit group. Participants were randomly assigned to 1 of 8 experimental conditions or a control group (neutral videos: n=16), as described in the Procedure section.

Sample Size and Power Calculation

A priori power analysis was conducted using G*Power (version 3.1; Heinrich-Heine-Universität) to determine the minimum sample size required for the study. Assuming a small to medium effect size (f=0.25), an α level of .05, and a power of 0.80, the estimated sample size required for a 3-way ANOVA with repeated measures (eg, time, topic, and condition) was 156 participants. This calculation ensured sufficient statistical power to detect meaningful differences between the experimental and control conditions across time and topics. The recruited sample size of 158 participants met this requirement.

Ethical Considerations

Ethics Approval and Trial Registration

The study was approved by the ethics committee of the University Medical Center Hamburg-Eppendorf, Germany (LPEK-0399) and was conducted in accordance with the Declaration of Helsinki. This study is part of a larger research project aimed at developing and evaluating an e-mental health intervention. Sleep and social competence represent two key modules within the program. The interventions examined in this trial were specifically designed as integral components of this broader program. The trial is registered in the German Clinical Trials Registration under the identifier DRKS00026722. This is a pilot trial informing a larger and independent randomized controlled trial that has not yet been published.

Informed Consent

All participants provided electronic informed consent before beginning the study. They were informed about the purpose and procedures, as well as their right to withdraw at any time without consequences. No waiver of consent was applied.

Privacy and Confidentiality

The data collected in this study were anonymized and stored in compliance with the General Data Protection Regulation. No personally identifiable information was collected or retained, and all data were analyzed in aggregated form to ensure confidentiality.

Compensation

Participants received access to the COGITO (www.uke.de/cogito) self-help smartphone app as an incentive for their participation.

Procedure

Participants were recruited online through Facebook groups and Facebook pages focused on mental health, well-being, and psychology (eg, forums for self-help and personal development). Recruitment posts included a brief explanation of the study purpose, eligibility criteria (eg, being aged 18 to 75 years, having internet access, and being proficient in German), and a link to the survey on the Qualtrics platform. The recruitment messages emphasized the voluntary nature of participation and informed potential participants about the incentive for completing the study—access to the COGITO self-help app.

Individuals reporting suicidal ideation, assessed using item 15 of the Web Screening Questionnaire [25], were excluded to ensure participant safety. The recruitment process aimed to reach a broad audience. The study was conducted online without face-to-face assessment. After providing informed consent, participants completed questions on demographic characteristics as well as questions on sleep and social competence. A baseline knowledge test followed. The knowledge test was designed specifically for this study and was reviewed by experts to ensure content validity. Participants were then randomly assigned to 1 of 9 experimental or control conditions using Qualtrics software, ensuring balanced group sizes. The study used a parallel-group randomized controlled trial design. There were no participant dropouts during the study, and all randomized participants completed the interventions. No changes to the study methods, intervention materials, or randomization procedures were made after the trial commenced.

Experimental groups received videos, texts, or both containing content on both topics (sleep and social competence), while the control group watched a video on a neutral topic unrelated to sleep or social competence (a tutorial on using a self-help app). To assess whether the order of the media might be a mechanism in knowledge gain, the order of text versus videos varied between experimental conditions. To prevent participants from skipping information by continuing without watching a video or reading a text, a timer was added. After the timer had expired (after 2.5 minutes), participants could continue the program. After presentation of the learning materials or of neutral stimuli in the control group, the postintervention knowledge test was administered. Blinding was not possible in this study due to the nature of the interventions, as participants were aware of whether they were receiving video or text materials. However, assessments were conducted as planned after randomization, and participants were not fully informed about the specific hypotheses of the study.

XSL•FO
RenderX

Outcomes

The primary outcome measure was knowledge gain. Knowledge was measured before and after the experimental condition for the topics of sleep and social competence. The primary outcome measure was an a priori designed and validated knowledge test consisting of 30 statements (15 per topic) that participants had to answer on a 2-point scale (true or false). Participants received 1 point for each correct answer. Thus, a maximum of 30 points could be achieved in the knowledge test. The internal consistency of the knowledge test was assessed using Cronbach α, which yielded a value of α=0.83, indicating good reliability of the test items. Subsequently, for each statement participants had to report their level of confidence in each response on a scale ranging from 0% to 100%. The outcome score for knowledge gain was the difference in the knowledge test scores before and after the intervention. The experimental condition was either a video or a text (a transcript of the video) or a combination of video and text conveying knowledge about the topics of sleep and social competence (see Experimental Conditions).

Experimental Conditions

Overview

Within the framework of this research project, we developed transdiagnostic videos on sleep and social competence deficits that might serve as psychoeducation material for many disorders (eg, chronic pain or pathological gambling) within the framework of e-mental health programs. The videos were created with the software Videoscribe (Sparkol Limited) and Final Cut Pro X2 (Apple Inc). Participants in the control condition received a 4-min video tutorial of a self-help app with content that was unrelated to sleep or social competence.

Video on Sleep

The video on sleep (VS) explained the theoretical framework for the formation and maintenance of sleep disorders, highlighting environmental (eg, noise) and internal (eg, ruminating) maintenance factors. The video explained how disorders are maintained by a learned connection of tension, sleep environment, and stressful thoughts. It also highlighted the negative consequences of sleep deprivation. The total length of the video was 2 minutes, 14 seconds, with a vertical resolution of 1080 pixels and framerate of 15 fps.

Text on Sleep

The text on sleep (TS) was a transcript of the video. Thus, the wording was identical in the video and the text.

Video on Social Competence

The video on social competence (VSC) explained that social competence is the ability to find a balance between one's own needs on the one hand and social adaptation on the other. Collectively, 3 different skills are crucial for social competence, all of which can be learned and practiced. The first skill is asserting one's own interests. The second skill is positive relationship building, such as in a partnership, to protect both one's own interests and at the same time the interests of the other person. The third skill is the ability to solicit sympathy from other people, such as when contacting strangers or in situations where one wants something from one's counterpart. The total length of the video was 3 minutes, 57 seconds, with a resolution of 1080 pixels and framerate of 24 fps.

Text on Social Competence

The text on social competence (TSC) was a transcript of the video. Thus, the wording was identical in the video and the text.

Conditions

The conditions were categorized by sequence (ie, whether the medium alternated or remained the same, that is, video to text vs text to text), topic (sleep vs social competence), and media type (video vs text), resulting in 8 experimental conditions. A control condition was also included in the study. The 8 experimental conditions were as follows:

1. VS, then TSC (n=19)
2. TS, then VSC (n=17)
3. VS, then VSC (n=20)
4. TS, then TSC (n=16)
5. VSC, then TS (n=18)
6. TSC, then VS (n=16)
7. VSC, then VS (n=18)
8. TSC, then TS (n=18)

Condition 9 served as the control condition, consisting of a neutral video stimulus (n=16).

Statistical Analysis

Data were analyzed using SPSS Statistics (version 29; IBM Corp). To gain a better understanding of knowledge gain in e-mental health programs, multiple statistical analyses were conducted. The analyses aimed to compare the experimental conditions to the control condition and to assess the impact of knowledge deficits on knowledge gain as well as the potential effects of medium, topic, and sequence on knowledge gain. Specifically, to compare knowledge gain between the intervention and control conditions, a 3-way ANOVA considering topic (sleep and social competence) and time (pre- and postintervention knowledge tests) as within-subject factors and condition (experimental vs control) as the between-subject factor was calculated. To assess knowledge gain in individuals with initial deficits in sleep or social competence, a 2-way ANOVA with deficit groups (participants with deficits in sleep, social competence, both, or none) as the between-subject factor and topic as the within-subject factor was performed. To examine the effects of medium, topic, and sequence on learning outcomes, a multivariate ANOVA (MANOVA) with knowledge gain differences as the dependent variables and various independent factors, including medium and deficit group, was applied. Subsequently, to evaluate the impact of knowledge gain on confidence, a 4-way ANOVA using, time, accuracy, and topic as the within-subject factors and condition as the between-subject factor was conducted. These analyses provided insights into factors influencing knowledge gain in the experimental conditions used in this study. An α error of 5% (2-sided) was used for all analyses. As the primary aim of this study was exploratory, parametric tests such as *t* tests, ANOVA, and MANOVA were used. A visual inspection of assumptions

XSL•FO
RenderX

did not reveal serious violations that would impact the validity of the analyses.

Results

Table 1 presents the knowledge score (primary outcome) for all groups, showing the changes in the pre- and postintervention assessments. The flowchart in Figure 1 presents a detailed description of the participant flow through the study.

Table 1. Mean knowledge scores before and after interventions for sleep and social competence across 9 experimental conditions in a web-based e-mental health study. Participants were exposed to combinations of video- and text-based psychoeducation or served as controls. The study was conducted online.

	Participants, n	Preintervention score for sleep, mean (SD)	Postintervention score for sleep, mean (SD)	Preintervention score for social competence, mean (SD)	Postintervention score for social competence, mean (SD)
Condition 1: video on sleep, then text on social competence	19	11.47 (1.95)	11.84 (1.61)	8.68 (1.38)	9.32 (1.25)
Condition 2: text on sleep, then video on social competence	17	9.94 (1.71)	10.71 (1.96)	8.59 (1.77)	8.65 (1.87)
Condition 3: video on sleep, then video on social competence	20	9.45 (2.52)	10.10 (2.77)	8.70 (1.17)	8.90 (1.12)
Condition 4: text on sleep, then text on social competence	16	9.81 (2.32)	9.38 (2.45)	8.31 (1.66)	8.81 (1.64)
Condition 5: video on social competence, then text on sleep	18	9.50 (1.86)	10.11 (2.22)	9.22 (1.17)	9.89 (1.23)
Condition 6: text on social competence, then video on sleep	16	11.06 (1.53)	10.94 (2.21)	8.25 (1.53)	8.94 (1.57)
Condition 7: video on social competence, then video on sleep	18	9.78 (2.67)	10.33 (2.83)	8.50 (1.62)	9.17 (1.54)
Condition 8: text on social competence, then text on sleep	18	10.17 (2.26)	10.50 (2.26)	8.50 (1.42)	8.83 (1.29)
Condition 9: control	16	10.00 (2.31)	9.25 (2.32)	8.12 (1.20)	8.88 (1.20)

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) flow diagram. TS: text on sleep; TSC: text on social competence; VS: video on sleep; VSC: video on social competence.

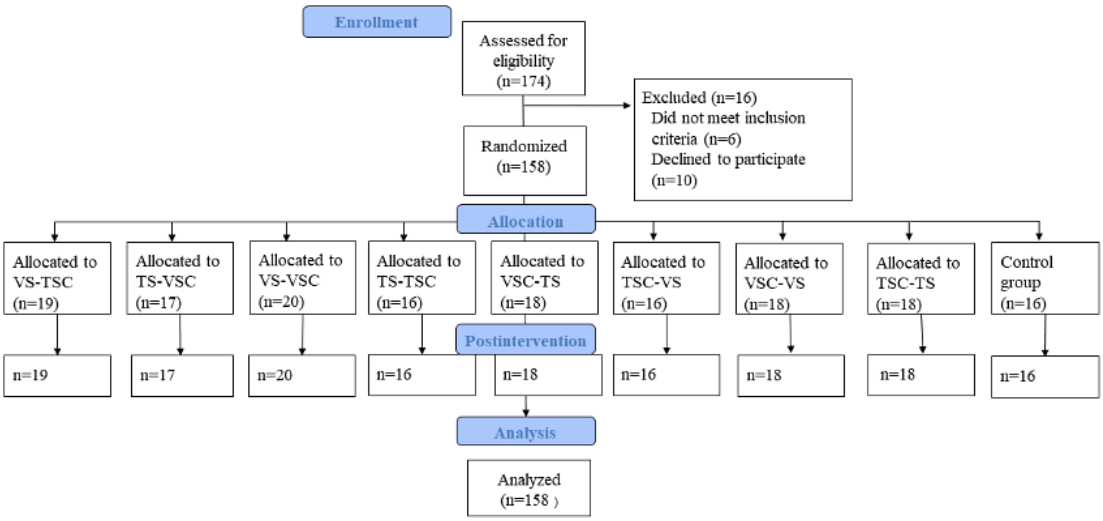

Effectiveness of the Experimental Conditions for Knowledge Gain in Comparison to the Control Condition

To assess knowledge gain from the first to the second (pre- and postexperimental) knowledge tests, the 8 experimental conditions were pooled into 1 group and compared with the control group (Figure 2). A 3-way ANOVA with domain (sleep and social competence) and time (pre- and postintervention) as the within-subject factors and group (pooled experimental group vs control group) as the between-subject factor was performed. The analysis yielded a significant main effect for time (pre- to postexperimental condition: $F_{1,156}$=17.272; P<.001; ηp^2=0.10), indicating that knowledge was successfully increased from

XSL•FO
RenderX

before to after the intervention. Furthermore, a significant interaction of topic, time, and condition was observed ($F_{1,156}$=6.737; P=.01; ηp^2=0.04), reflecting that the pooled experimental condition improved scores on the knowledge test for both topics (sleep: preintervention mean 10.14, 95% CI 9.78-10.50 and postintervention mean 10.50, 95% CI 10.11-10.89; social competence: preintervention mean 8.61, 95% CI 8.37-8.85 and postintervention mean 9.07, 95% CI 8.83-9.31) across time. In contrast, the control condition did not lead to improvement on the knowledge test on sleep over time (preintervention mean 10.00, 95% CI 8.72-11.28 and postintervention mean 9.25, 95% CI 7.97-10.53). Unexpectedly, the control group showed some improvement on the knowledge test on social competence (preintervention mean 8.13, 95% CI 7.47-8.80 and postintervention mean 8.88, 95% CI 8.22-9.54).

Figure 2. Pre- and postintervention knowledge scores for sleep and social competence in the intervention and control groups. The results demonstrate changes in knowledge levels following the intervention, highlighting differences between the two groups for both topics.

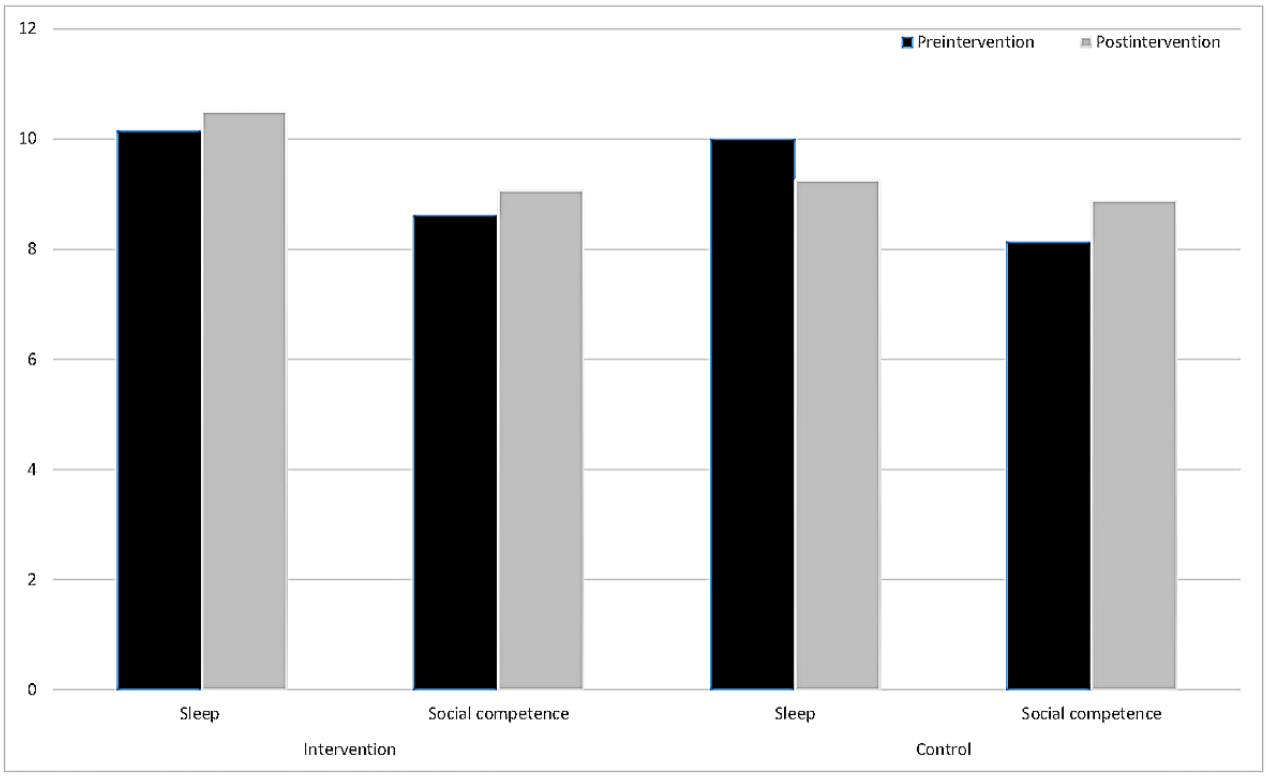

Knowledge Gain in Participants With Initial Deficits in Sleep and Social Competence

Individuals who reported social competence deficits showed less knowledge about social competence as assessed by the knowledge test at baseline compared to participants without reported deficits in social competence (t_{156}=2.12; P=.04; d=0.41, 95% CI 0.04-1.14). At a statistical trend level, this was also true for individuals with reported sleep deficits (t_{156}=1.71-.09; d=0.28, 95% CI −1.04 to 0.67).

To further assess knowledge gain in individuals with sleep deficits or social competence deficits, a 2-way ANOVA with the deficit groups (participants with deficits in social competence, deficits in sleep, none, or both) as the between-subject factor and topic (sleep and social competence) as the within-subject factor was performed. Gain in knowledge (from the preintervention to the postintervention knowledge test) served as the dependent variable. A significant main effect for topic was present ($F_{1,154}$=4.883-.03; η_p^2=0.03), reflecting greater knowledge gain in the social competence knowledge test compared to the sleep knowledge test for all participants. Furthermore, an interaction effect, which bordered on significance, was found for topic and sample ($F_{3,154}$=2.210-.09; η_p^2=0.04), showing a greater knowledge gain for social competence than for sleep in individuals with prior social competence deficits and a greater knowledge gain for sleep than for social competence in individuals with sleep deficits. Individuals who reported both sleep deficits and social competence deficits or who did not report deficits in either sleep or social competence showed similar knowledge gain for both topics. Post hoc tests showed that those with social competence deficits had significant knowledge gain (t_{141}=7.12; P<.001; d=0.60, 95% CI –0.76 to −0.42) as did those with sleep deficits (t_{141}=2.43-.016; d=0.20; 95% CI −0.37 to −0.04) from the preintervention to postintervention knowledge tests.

Examining the Impact of Medium, Sequence, and Topic

To investigate the potential effects of medium, sequence, and topic on learning outcomes, a MANOVA was conducted. Two indices of knowledge gain (ie, differences in knowledge from baseline to postassessment for the topics of sleep and social competence) served as dependent variables. Independent factors were the different sequences of the two media (ie, video to text or text to video), the medium shown first (video or text), topic

XSL•FO
RenderX

displayed first (sleep or social competence), use of only 1 medium (ie, video and video, text and text), and deficit group (participants with deficits in sleep, social competence, none, or both). For the omnibus analyses for sleep, none of the main effects or interaction effects achieved significance (P>.05). For social competence, subsidiary analyses showed only 1 significant main effect, which was for the medium shown first ($F_{1,157}$=4.058-.05; η_p^2=0.04), indicating greater knowledge gain when the video was shown first. This result would likely not remain significant after correction for multiple comparisons.

Effects of Knowledge Gain on Confidence in Correct and Erroneous Items

To assess the influence of knowledge gain through video or text on confidence in a knowledge test on the topics of sleep and social competence, a 4-way ANOVA with time (pre- and postintervention knowledge tests), accuracy (correct and incorrect), and topic (sleep and social competence) as the within-subject factors and condition (pooled experimental condition and control condition) as the between-group factor was conducted (Figure 3). Confidence served as the dependent variable. A significant main effect for time was found ($F_{1,157}$=9.354; P<.001; η_p^2=.06), indicating an overall rise in confidence from pre- to postintervention. This was qualified by a significant interaction for time and accuracy ($F_{1,157}$=9.354; P<.001; η_p^2=0.06), suggesting a confidence gain from preintervention (mean 43.95, 95% CI 41.32-46.59) to postintervention (mean 48.32, 95% CI 44.58-52.07) for correct items only; confidence for incorrect items remained unchanged over time (at both times: mean 27.18, 95% CI 24.79-29.58). Additionally, analyses revealed a significant main effect for accuracy ($F_{1,157}$=98.735; P<.001; η_p^2=0.39), indicating higher confidence for correct than incorrect items overall. Another significant interaction emerged for time, accuracy, and group ($F_{1,157}$=9.797-.002; η_p^2=0.06), indicating a gain in confidence for correct items from preintervention (mean 42.82, 95% CI 41.15-44.50) to postintervention (mean 51.67, 95% CI 49.28-54.04) in the pooled experimental condition, whereas the control condition showed essentially the same confidence from preintervention (correct: mean 45.09, 95% CI 40.09-50.08; incorrect: mean 29.19, 95% CI 24.65-33.73) to postintervention (correct: mean 44.98, 95% CI 37.89-52.08; incorrect: mean 29.19, 95% CI 24.65-33.73) for both correct and incorrect items. Moreover, a significant interaction for time, accuracy, and topic was found ($F_{1,157}$=10.213; P<.001; η_p^2=0.61, suggesting a higher confidence gain for social competence from preintervention (correct: mean 42.86, 95% CI 41.15-44.50) to postintervention (correct: mean 50.27, 95% CI 46.52-54.03), whereas there was essentially no change in scores for sleep from preintervention (correct: mean 45.05, 95% CI 41.57-48.53) to postintervention (correct: mean 46.37, 95% CI 41.81-50.93) for correct items.

Figure 3. Percentages of correct and incorrect answers before and after the interventions across the control (gray bars) and intervention (black bars) groups. The data highlight changes in confidence levels for correct and incorrect responses, showing the effects of the intervention on participant confidence.

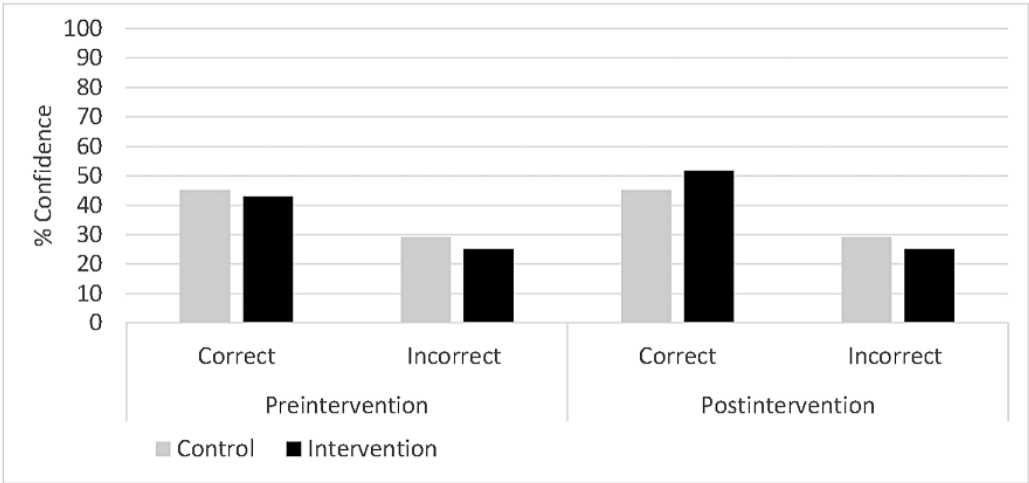

Discussion

Overview

One of the aims of this study was to examine which medium (video vs text) is most effective in conveying knowledge in e-mental health programs for individuals with and without psychological deficits, using sleep and social competence as example topics. Our primary outcome was knowledge gain, with a particular focus on the interactions between type of medium (video or text), topic (sleep or social competence), and sequence of presentation. We hypothesized that individuals with sleep deficits or social competence deficits would initially have less knowledge on these topics and that the experimental conditions (information provided by video and/or text) would compensate for these knowledge gaps. Additionally, we investigated how increased knowledge affected confidence in judgment.

Efficacy of the Interventions for Knowledge Gain and Moderators

The results of this study indicate that the experimental conditions (psychoeducational videos and texts) significantly increased participants' knowledge from before to after the intervention,

XSL•FO
RenderX

aligning with our expectations and corroborating previous research that highlights the effectiveness of psychoeducation across various mental disorders, as confirmed by meta-analyses and literature reviews [8-13]. The two media (video and text) led to roughly the same increase in knowledge. In addition, topic, medium, and sequence generally did not impact knowledge gain, with the exception that for the topic of social competence, the gain in knowledge was greater if videos were shown before texts. However, this result would not have withstood correction for multiple comparisons. To the best of our knowledge, no comparable studies have been conducted in an e-mental health setting to date. Preliminary studies, such as those in the field of education involving medical students, have demonstrated that video can have advantages over text, as learning outcomes are better. However, our results do not confirm this finding.

As expected, participants with subjective deficits did have less knowledge on the respective topics at baseline (for sleep, this was only at the statistical trend level). The results thus lend support to the implicit claim of psychoeducation that knowledge deficits are associated with symptoms. Importantly, those who reported deficits in a topic learned more about that topic compared to participants without reported deficits in that topic as a result of the intervention.

Impact of Confidence on Knowledge Gain

Our results show that confidence in correct answers increased with increasing knowledge. Thus, our low-threshold approach did more than convey knowledge; the participants also became more confident in their correct answers. In line with these results, overall, psychoeducation interventions have been shown to positively impact confidence in various areas, such as childbirth, newborn care, parenting children with autism spectrum disorder, communication skills in health care settings, disaster preparedness, and managing mental disorders [26-28]. Based on previous research, however, we know that the effect can go both ways: increased confidence can also lead to improved learning outcomes [29,30].

Strengths and Limitations

This study is among the first to explore knowledge gain mechanisms in e-mental health programs and the relationship between knowledge gain and psychological deficits. To ensure the robustness of our findings, we used two different topics to control for the potential confounding variables of knowledge content and topic. Additionally, we measured not only knowledge gain but also confidence levels. Random group allocation and a control condition enhanced the validity of the results. However, our study also had several limitations. First, the recruitment strategy relied on social media, raising concerns about the representativeness of the sample. Moreover, many university students were included in the sample; as a result, the high educational level of our sample compromises the generalizability of our findings to the general population and especially to patient samples, which often have a lower educational level. The high educational level of the sample may have facilitated knowledge acquisition, particularly through text-based interventions, as such participants are likely more familiar with learning in text-dominant formats. In contrast, individuals with lower educational levels may have difficulty understanding text-heavy content, which could reduce the effectiveness of such interventions. Video-based formats might be particularly beneficial for such populations, given videos' ability to simplify complex concepts and enhance engagement.

Future Directions

Our findings suggest that the question might not be about determining whether video or text is superior in psychoeducation. Instead, it might be more beneficial to focus on multimedia integration. Combining various media formats could be key to maintaining learner engagement and motivation. The synergy of text, video, and other multimedia elements might enhance the overall knowledge gain more effectively than any single medium alone. This multimodal approach could cater to diverse learning preferences and aid in sustaining attention and interest over prolonged periods. Future studies should explore the impact of multimedia combinations on knowledge gain, especially in the context of psychoeducational content in e-learning platforms. Based on 14 trials, Emmerson and colleagues [31] found that multimedia approaches (including videos) led to increased adherence compared to written or verbal approaches. However, previous studies indicate that to increase adherence, which could not be tested in our single-session design, videos alone might be better [20,31]. Long-term designs are necessary to test the effects of different media on adherence, ensuring sustained engagement and benefit. Additionally, our study's limited scope compared to comprehensive e-learning programs highlights the need to investigate how media effectiveness varies with content complexity.

E-mental health programs save resources in comparison to conventional therapies by automating interventions (eg, video vs text), without therapist support. Acting as a versatile toolbox, e-mental health programs offer different media formats. Personalizing these programs to individual preferences—such as allowing the user to choose between videos and text—might enhance their effectiveness. For instance, individuals with cognitive impairments might benefit more from videos if they find text-heavy content challenging, while others might prefer reading to avoid the fast pace of videos. Future studies should explore how such preferences influence knowledge gain and symptom improvement to ensure that e-mental health programs are both individualized and effective.

Future research should also examine the impact of various media (video, text, and audio) on symptom improvement in e-mental health interventions. In other words, it still needs to be demonstrated that an increase in knowledge leads to a reduction in symptoms. If true, this would highlight the crucial role of psychoeducation in psychotherapy [7,20].

Considering the strong tie between confidence and knowledge gain, mental health programs could be improved by including strategies to increase confidence, as increased confidence may lead to a decrease in symptoms. Future studies would be helpful to better understand this relationship.

Although the psychoeducational interventions in this study showed promising results in terms of knowledge gain and improved confidence, their implementation in real-world settings

XSL•FO
RenderX

may face challenges. For instance, the high educational level of the sample could limit the transferability of the findings to populations with less education, who might require more tailored or simplified formats, such as videos or interactive content. Furthermore, the interventions were tested in a controlled, single-session setting, which does not account for the sustained engagement required in real-world applications. To address these challenges, future implementations could incorporate adaptive learning pathways that adjust content complexity and format based on the user's educational background or cognitive abilities. Additionally, the interventions could be embedded in existing e-mental health programs. Future research could also benefit from the inclusion of participants with diagnosed mental disorders to improve external validity and the transferability of results to clinical populations.

Conclusion

E-mental health is essential in modern mental health care. To maximize its potential, research should focus on optimizing e-mental health programs and understanding the mechanisms of success. This study contributes to this goal by investigating the roles of topic, sequence, and medium in knowledge gain. We found no differences in knowledge gain between video and text formats. Interestingly, individuals with self-reported deficits in 1 of the 2 topics did have less knowledge on the topic but also gained more new knowledge on that topic than those without self-reported deficits. Future studies should explore multimedia effects (video, text, and audio) on symptom improvement in clinical intervention studies. Future research should also examine whether individuals' preferences for different media (eg, video vs text) influence knowledge gain and confidence.

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability

The anonymized datasets generated and analyzed during this study are available from the corresponding author upon reasonable request. Requests must include a clear statement of purpose and comply with institutional guidelines.

Conflicts of Interest

None declared.

Multimedia Appendix 1
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 2743 KB - formative_v9i1e65478_app1.pdf]

References

1. World mental health report: transforming mental health for all. World Health Organization. 2022. URL: https://iris.who.int/bitstream/handle/10665/356119/9789240049338-eng.pdf?sequence=1 [accessed 2025-05-22]
2. Andrade LH, Alonso J, Mneimneh Z, Wells JE, Al-Hamzawi A, Borges G, et al. Barriers to mental health treatment: results from the WHO World Mental Health surveys. Psychol Med 2014 Apr;44(6):1303-1317 [FREE Full text] [doi: 10.1017/S0033291713001943] [Medline: 23931656]
3. Knapp M. Hidden costs of mental illness. Br J Psychiatry 2003 Dec;183(6):477-478. [doi: 10.1192/bjp.183.6.477] [Medline: 14645015]
4. Karyotaki E, Riper H, Twisk J, Hoogendoorn A, Kleiboer A, Mira A, et al. Efficacy of self-guided internet-based cognitive behavioral therapy in the treatment of depressive symptoms: a meta-analysis of individual participant data. JAMA Psychiatry 2017 Apr 01;74(4):351-359 [FREE Full text] [doi: 10.1001/jamapsychiatry.2017.0044] [Medline: 28241179]
5. Karyotaki E, Efthimiou O, Miguel C, Bermpohl FMG, Furukawa TA, Cuijpers P, Individual Patient Data Meta-Analyses for Depression (IPDMA-DE) Collaboration, et al. Internet-based cognitive behavioral therapy for depression: a systematic review and individual patient data network meta-analysis. JAMA Psychiatry 2021 Apr 01;78(4):361-371 [FREE Full text] [doi: 10.1001/jamapsychiatry.2020.4364] [Medline: 33471111]
6. Anderson CM, Hogarty GE, Reiss DJ. Family treatment of adult schizophrenic patients: a psycho-educational approach. Schizophr Bull 1980;6(3):490-505. [doi: 10.1093/schbul/6.3.490] [Medline: 7403810]
7. Srivastava P, Panday R. Psychoeducation an effective tool as treatment modality in mental health. Int J Indian Psychol 2016 Dec 25;4(1):82 [FREE Full text] [doi: 10.25215/0401.153]
8. Okafor AJ, Monahan M. Effectiveness of psychoeducation on burden among family caregivers of adults with schizophrenia: a systematic review and meta-analysis. Nurs Res Pract 2023;2023:2167096 [FREE Full text] [doi: 10.1155/2023/2167096] [Medline: 37767330]

XSL•FO
RenderX

9. Powell LA, Parker J, Weighall A, Harpin V. Psychoeducation intervention effectiveness to improve social skills in young people with ADHD: a meta-analysis. J Atten Disord 2022 Feb;26(3):340-357 [FREE Full text] [doi: 10.1177/1087054721997553] [Medline: 33666104]
10. Quintiliani MI, Imperatori C, Testani E, Losurdo A, Tamburello S, Contardi A, et al. Usefulness of psychoeducational intervention in chronic insomnia: an actigraphic study. J Ment Health 2020 Feb;29(1):20-26. [doi: 10.1080/09638237.2017.1417563] [Medline: 29271275]
11. Oliveira CTD, Dias ACG. How can psychoeducation help in the treatment of mental disorders? Estud.Psicol (Campinas) 2023;40:e190183. [doi: 10.1590/1982-0275202340e190183]
12. Brouzos A, Vatkali E, Mavridis D, Vassilopoulos SP, Baourda VC. Psychoeducation for adults with post-traumatic stress symptomatology: a systematic review and meta-analysis. J Contemp Psychother 2021 Oct 25;52(2):155-164. [doi: 10.1007/s10879-021-09526-3]
13. Sin J, Gillard S, Spain D, Cornelius V, Chen T, Henderson C. Effectiveness of psychoeducational interventions for family carers of people with psychosis: a systematic review and meta-analysis. Clin Psychol Rev 2017 Aug;56:13-24 [FREE Full text] [doi: 10.1016/j.cpr.2017.05.002] [Medline: 28578249]
14. Harvey AG, Lee J, Williams J, Hollon SD, Walker MP, Thompson MA, et al. Improving outcome of psychosocial treatments by enhancing memory and learning. Perspect Psychol Sci 2014 Mar;9(2):161-179 [FREE Full text] [doi: 10.1177/1745691614521781] [Medline: 25544856]
15. Lukens E, McFarlane W. Psychoeducation as evidence-based practice: considerations for practice, research, and policy. Brief Treat Crisis Interv 2004;4(3):205-225. [doi: 10.1093/brief-treatment/mhh019]
16. Abramowitz JS, Franklin ME, Zoellner LA, DiBernardo CL. Treatment compliance and outcome in obsessive-compulsive disorder. Behav Modif 2002 Sep;26(4):447-463. [doi: 10.1177/0145445502026004001] [Medline: 12205821]
17. Means B, Toyama Y, Murphy R, Bakia M, Jones K. Evaluation of evidence-based practices in online learning: a meta-analysis and review of online learning studies. US Department of Education. 2009. URL: https://www2.ed.gov/rschstat/eval/tech/evidence-based-practices/finalreport.pdf [accessed 2025-05-22]
18. Ebner C, Gegenfurtner A. Learning and satisfaction in webinar, online, and face-to-face instruction: a meta-analysis. Front Educ 2019 Sep 3;4:92. [doi: 10.3389/feduc.2019.00092]
19. Lin Y, Yu Z. A meta-analysis evaluating the effectiveness of instructional video technologies. Tech Know Learn 2023 Jul 12;29(4):2081-2115. [doi: 10.1007/s10758-023-09669-3]
20. Buch SV, Treschow FP, Svendsen JB, Worm BS. Video- or text-based e-learning when teaching clinical procedures? A randomized controlled trial. Adv Med Educ Pract 2014;5:257-262 [FREE Full text] [doi: 10.2147/AMEP.S62473] [Medline: 25152638]
21. Schlarb AA, Fründ JP, Kovacevic T, Faber J. Modularized iCBT-I self-learn training for university staff-prevention and early intervention in the SARS-CoV-2 crisis: A pilot study. Somnologie (Berl) 2021;25(1):29-37 [FREE Full text] [doi: 10.1007/s11818-021-00301-z] [Medline: 33649702]
22. Lim SL, Yang J, Ehrisman J, Havrilesky LJ, Reed SD. Are videos or text better for describing attributes in stated-preference surveys? Patient 2020 Aug;13(4):401-408. [doi: 10.1007/s40271-020-00416-9] [Medline: 32239442]
23. Fischhoff B, Slovic P, Lichtenstein S. Knowing with certainty: the appropriateness of extreme confidence. J Exp Psychol Hum Percept Perform 1977;3(4):552-564. [doi: 10.1037//0096-1523.3.4.552]
24. Martí L. Certainty and the source of misinformed beliefs [dissertation]. University of California Berkeley. 2021. URL: https://escholarship.org/uc/item/46n2b569 [accessed 2025-05-22]
25. Dory V, Degryse J, Roex A, Vanpee D. Usable knowledge, hazardous ignorance - beyond the percentage correct score. Med Teach 2010;32(5):375-380. [doi: 10.3109/01421590903197027] [Medline: 20423255]
26. Donker T, van Straten A, Marks I, Cuijpers P. A brief web-based screening questionnaire for common mental disorders: development and validation. J Med Internet Res 2009 Jul 24;11(3):e19 [FREE Full text] [doi: 10.2196/jmir.1134] [Medline: 19632977]
27. Fenwick J, Toohill J, Slavin V, Creedy DK, Gamble J. Improving psychoeducation for women fearful of childbirth: evaluation of a research translation project. Women Birth 2018 Feb;31(1):1-9. [doi: 10.1016/j.wombi.2017.06.004] [Medline: 28684046]
28. Shorey S, Chan SW, Chong YS, He H. Perceptions of primiparas on a postnatal psychoeducation programme: the process evaluation. Midwifery 2015 Jan;31(1):155-163. [doi: 10.1016/j.midw.2014.08.001] [Medline: 25174540]
29. Magaña S, Lopez K, Salkas K, Iland E, Morales MA, Garcia Torres M, et al. A randomized waitlist-control group study of a culturally tailored parent education intervention for Latino parents of children with ASD. J Autism Dev Disord 2020 Jan;50(1):250-262. [doi: 10.1007/s10803-019-04252-1] [Medline: 31606887]
30. Cummings CL, Connelly LK. Can nursing students' confidence levels increase with repeated simulation activities? Nurse Educ Today 2016 Jan;36:419-421. [doi: 10.1016/j.nedt.2015.11.004] [Medline: 26599594]
31. Emmerson KB, Harding KE, Taylor NF. Providing exercise instructions using multimedia may improve adherence but not patient outcomes: a systematic review and meta-analysis. Clin Rehabil 2019 Apr 24;33(4):607-618. [doi: 10.1177/0269215518819706] [Medline: 30582358]

XSL•FO
RenderX

Abbreviations

MANOVA: multivariate ANOVA
TS: text on sleep
TSC: text on social competence
VS: video on sleep
VSC: video on social competence

Edited by A Mavragani; submitted 16.08.24; peer-reviewed by OE Santangelo, K Muroi; comments to author 15.01.25; revised version received 04.02.25; accepted 07.02.25; published 24.06.25.

Please cite as:
Borsutzky S, Gehlenborg J, Rolvien L, Moritz S
Video- Versus Text-Based Psychoeducation in Web-Based E-Mental Health Programs: Randomized Controlled Trial
JMIR Form Res 2025;9:e65478
URL: https://formative.jmir.org/2025/1/e65478
doi:10.2196/65478
PMID:40554780

©Swantje Borsutzky, Josefine Gehlenborg, Lara Rolvien, Steffen Moritz. Originally published in JMIR Formative Research (https://formative.jmir.org), 24.06.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Patient–Health Care Professional Communication via a Secure Web-Based Portal in Severe Mental Health Conditions: Qualitative Analysis of Secure Messages

Eva Meier-Diedrich[1,2], MSc; Carolyn Turvey[3,4,5], PhD; Jonas Maximilian Wördemann[1,2], MD; Justin Speck[1,2], BSc; Mareike Weibezahl[1,2], MD; Julian Schwarz[1,2,6], MD

[1]Department of Psychiatry and Psychotherapy, Center for Mental Health, Immanuel Hospital Rüdersdorf, Brandenburg Medical School Theodor Fontane, Rüdersdorf, Germany

[2]Faculty of Health Sciences Brandenburg, Brandenburg Medical School Theodor Fontane, Neuruppin, Germany

[3]Center for Access & Delivery Research and Evaluation, Iowa City Veterans Affairs Health Care System, Iowa City, IA, United States

[4]Office of Rural Health, Veterans Rural Health Resource Center – Iowa City, Iowa City Veterans Affairs Health Care System, Iowa City, IA, United States

[5]Department of Psychiatry, University of Iowa, Iowa City, IA, United States

[6]Center for Health Service Research Brandenburg, Brandenburg Medical School Theodor Fontane, Rüdersdorf, Germany

Corresponding Author:
Eva Meier-Diedrich, MSc
Department of Psychiatry and Psychotherapy, Center for Mental Health
Immanuel Hospital Rüdersdorf
Brandenburg Medical School Theodor Fontane
Seebad 82/83
Rüdersdorf, 15562
Germany
Phone: 49 33638 83 50
Email: Eva.Meier-Diedrich@mhb-fontane.de

Abstract

Background: Patients' web-based access to their medical records and secure messaging (SM) via patient portals is becoming increasingly prevalent worldwide. SM offers several potential benefits, including improved health outcomes and increased patient engagement. However, SM also raises concerns about effects on the therapeutic relationship and may be constrained by factors such as limited digital literacy and access to digital devices. Evidence on the use of SM in mental health is limited, and results are inconclusive.

Objective: This study aimed to examine (1) the purposes for which health care professionals (HCPs) and patients with psychiatric disorders use SM to communicate and (2) the specific use patterns associated with both patients and HCPs.

Methods: The secure messages (n=274) of 38 patients with psychiatric disorders and 4 HCPs (psychiatrists) from 3 psychiatric outpatient clinics in Brandenburg, Germany, was analyzed using thematic analysis. The data selected for this study represent a subsample from a larger study comprising a total of 116 patients. The subsample consists of the patients and HCPs who used SM.

Results: A total of 274 messages were analyzed: 22.3% (61/274) were initial notes from HCPs, 44.5% (122/274) were patient responses, and 33.2% (91/274) were HCP replies. Patients sent between 1 and 15 messages (mean 4.16, SD 3.42) and logged in 1 to 42 times (mean 10.78, SD 9.38). Most messages were sent during the day, although some were also sent at night and in the early morning. Regarding the purposes of SM, 4 core functions of SM were identified: reporting and feedback, interpersonal uses, intrapersonal uses, and organizational uses. Both patients and HCPs used SM to share treatment-relevant information and elicited feedback on treatment and medication. Furthermore, secure messages included expressions of gratitude by the patients, in addition to well-wishes and emotional support from the HCPs. SM allowed patients to reflect on their treatment and provide self-encouragement. Finally, secure messages were used to address organizational aspects such as scheduling, appointments, and administrative tasks.

Conclusions: SM in outpatient mental health care is multifaceted and holds the potential to enhance therapeutic contact and improve access to care by enabling quick, low-threshold communication between patients and HCPs, allowing treatment-related

XSL•FO
RenderX

concerns to be addressed promptly and effectively. However, the asynchronous nature of SM also poses new challenges, particularly in managing acute mental health crises and in setting boundaries to prevent HCPs from being perceived as constantly available. Therefore, specific training for HCPs—both during medical education and in clinical practice—is essential, along with clear guidelines on handling crises and managing sensitive information.

(JMIR Form Res 2025;9:e63713) doi:10.2196/63713

KEYWORDS

secure messaging; open notes; online record access; ORA; psychiatry; eHealth; mobile health; mHealth

Introduction

Background

With a growing number of countries worldwide allowing web-based medical record access to patients, the integration of secure messaging (SM) systems into patient portals is also increasing [1]. SM is recognized as a valuable tool with multiple benefits in health care. It is one of the most popular features of patient portals among both patients and health care professionals (HCPs) [2] and has been rapidly adopted in recent years [3,4]. SM enables patients and HCPs to communicate effectively, provide treatment feedback, and address urgent questions without the need for them to be postponed until the next in-office visit [5]. Therefore, SM serves as a preventive measure against treatment delays caused by long wait times [6], which can be particularly prominent in the field of mental health [7].

Providing patients with access to a patient portal is already well established internationally, particularly in the United States and the northern European countries [8]. However, the introduction of an SM feature is a relatively recent development in the field. In Sweden, patients were previously able to comment on their electronic health records (EHRs) [9], and while this feature was well accepted and frequently used by patients, it was discontinued in 2022 due to technical issues [10]. In Finland, patients can save data to their personal health record through certain (limited) well-being applications; however, HCPs currently do not have access to the data [11]. Similarly, in Germany, the need for an effective and secure digital communication tool in health care has been recognized, leading to the enactment of various laws in recent years, such as the "Act for Secure Digital Communication and Applications in Healthcare" (E-Health Act). This legislation paved the way for the mandatory implementation of an EHR for all patients by 2025 (so-called opt-out). In addition, the Hospital Future Act, which came into force in Germany in 2020, mandates the implementation of patient portals, including (asynchronous) communication between patients and HCPs [12].

Studies indicate that SM via patient portals can facilitate access to care [13], enhance patient-clinician communication, and enable efficient and location-independent information exchange between patients and HCPs [5,14]. This appears to be of particular importance in psychiatric systems, where therapeutic contacts are often brief and infrequent—typically 1 appointment every 3 months lasting 7.5 to 30 minutes for patients affected by mental health conditions. Furthermore, SM has been associated with improved health outcomes [15], greater patient engagement (eg, better treatment adherence [16], higher clinic attendance [17], and improved disease management and disease awareness [18]), and increased patient satisfaction [13]. Reported effects of SM on the frequency of in-office visits remain inconclusive [3-5,19]. SM is commonly used by patients for self-management of chronic diseases, accessing preventive services, and communicating with HCPs (eg, to inquire about symptoms or laboratory results, request prescription refills, or schedule appointments) [15,20]. Both patients and HCPs are generally in favor of using SM [14]. Patients particularly value the speed, ease, and directness of communicating through SM, perceiving it as an extension—and sometimes even replacement—of the in-office treatment [5,20]. Similarly, HCPs generally have a positive view of SM, appreciating its usefulness for tasks such as prescription refills and patient monitoring [14,20].

Despite its many advantages, both patients and HCPs have some concerns. Patients worry about the loss of interpersonal contact [21], the appropriate manner of engagement [14], and the possibility of taking up too much of their clinicians' (uncompensated) time [5,14]. Additional patient use barriers include time constraints, limited usability and patient portal accessibility, and lack of digital literacy [5]. The privacy, security, and governance of sensitive health data have been a major concern for both parties [22]. HCPs encounter new demands in the practical implementation, such as potential institutional and workflow barriers (uncompensated hours and poor design), which could result in an additional workload [20,23] and the pressure to respond immediately [14,20]. HCPs express concerns that patients may misinterpret the proper use of the messaging feature [14] and already face difficulties in managing certain "high-demand" patient messages due to their length, frequency, urgency, complexity, or lack of clarity [14,20].

Objectives

There is only limited evidence on patient portals in mental health [24-28]. Studies on SM through patient portals in mental health are even fewer and vary depending on the outcome measures. While there is some indication of increased engagement through SM [29], the effects on medication adherence remain limited [30]. Nevertheless, both patients and HCPs in the mental health care field seem to welcome the introduction of SM [20,31]. Evidently, the implementation of SM via patient portals in mental health needs further exploration [20,29]. Therefore, this study aims to thoroughly explore the use of the SM function in patient portals by patients affected by a mental health condition. The following research questions will be addressed: (1) for what purposes do HCPs and patients affected by a severe mental health condition use SM to communicate in the outpatient

XSL•FO
RenderX

treatment setting? and (2) what characterizes the use patterns of HCPs and patients?

Methods

Design

This study is part of the PEPPPSY (Piloting and Evaluation of Participatory Patient-Accessible Electronic Health Record in Psychiatry and Somatics) project (since 2021), which centers on piloting and evaluating a participatory patient record in psychiatry and somatic medicine [32,33]. The main objective is to investigate the development, implementation, processes, and outcomes of the corresponding patient portal, referred to as PEPPPSY, from the perspectives of patients and HCPs. Following the exploratory nature of this study, a qualitative method was chosen to comprehensively analyze the use of SM through patient portals.

PEPPPSY App

The pilot patient portal PEPPPSY was created through a research collaboration between the Norwegian University of Science and Technology and the Brandenburg Medical School. PEPPPSY is a stand-alone web-based technology designed to facilitate access to specific elements of a (digital) medical health record for research purposes. PEPPPSY is a patient portal that accesses data from the EHR and extends it with additional functions, as illustrated in Figure 1.

Figure 1. Relationship between the electronic health record (EHR) and the web-based patient portal PEPPPSY (Piloting and Evaluation of Participatory Patient-Accessible Electronic Health Record in Psychiatry and Somatics) and its functions.

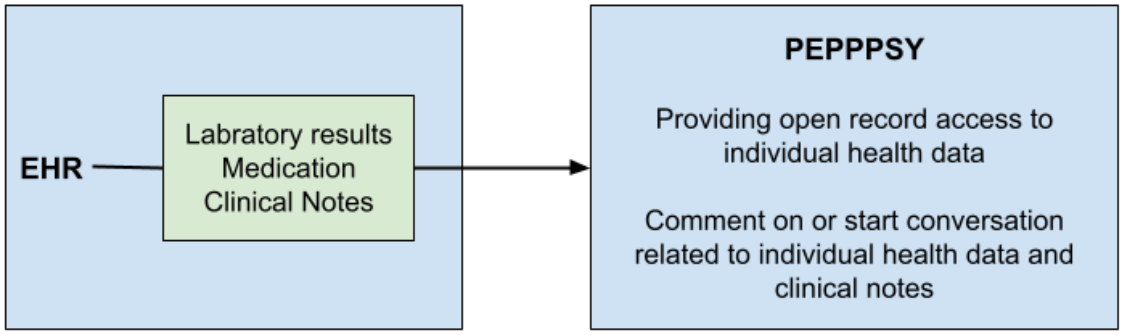

PEPPPSY is the result of an (ongoing) iterative process of participatory design, development, application, and evaluation [32]. PEPPPSY's primary function is to provide patients with secure access to view their clinicians' clinical notes. The practice of sharing clinical notes with patients is usually described with the term Open Notes. The pilot does not yet have a messaging function separate from the clinical notes. Nevertheless, patients have the possibility to respond to each clinical note and initiate SM with their HCPs. It is not uncommon for patients to use this SM option, independently of the initial clinical notes, both in terms of time and content, to contact their HCPs. In the patient portal settings, patients and HCPs could configure email notifications for new messages in the patient portal. HCPs had no specific requirements regarding how quickly they needed to respond or the time of day responses should be sent. Table 1 illustrates a fictional example of a messaging interaction via PEPPPSY that closely resembles actual practice. In the results, each quote is linked to a specific HCP (HCPXX) or patient (PTXX) using an abbreviation followed by a number. In the current second phase of the pilot, PEPPPSY is being expanded to serve a broader patient population, offering additional features such as the SM function or a proxy access for care partners.

Table 1. Example of messaging via the patient portal, including an initial clinical note by a health care professional (HCP), the patient's response, and the HCP's reply.

Sequence of the interaction	Excerpt from fictional secure message
Initial clinical note by an HCP	"The dosage of venlafaxine was decreased gradually from 150 mg to 75 mg due to the patient's report of excessive perspiration." [HCPXX]
Patient's response	"Unfortunately, the medication hasn't helped. I've been trying 75mg for two weeks now, but I'm still sweating. What can we do?" [PTXX]
HCP's response	"Thank you for your message. Please come to my office. We should reduce or discontinue the current medication and then switch to a different one." [HCPXX]

Ethical Considerations

Ethics approval was obtained from the Ethics Committee of the Brandenburg Medical School (E-01-20210727), and the study was registered with the German Clinical Trial Register (DRKS00030188). All data collected were deidentified. Within the SM, names, age, family relationships, location, gender, and times were randomly pseudonymized. The participating HCPs received a single compensation payment of €50 (US $57) for their participation in the study.

Study Setting

The study was conducted at 3 psychiatric outpatient university clinics of Immanuel Hospital Rüdersdorf, part of Brandenburg Medical School. Located in the Berlin/Brandenburg metropolitan region, these clinics provide mental health care services to

XSL•FO
RenderX

approximately 255,000 residents in the catchment area. These outpatient clinics are dedicated to offer specialized mental health care for patients who need a comprehensive and multidisciplinary approach, considering the nature, severity, or duration of their condition. There are approximately 500 psychiatric outpatient clinics (POCs) in Germany, treating approximately 2 million patients a year [34]. Treatment eligibility in respective clinics is based on specific diagnoses—including severe mental illness and other approved conditions—as determined by insurance providers and the hospital association [35]. The implemented multidisciplinary team approach includes psychiatrists, psychotherapists, nurses, and social workers.

Recruitment and Sampling

The inclusion period spanned from October 2022 to February 2024. For this branch of the study, all patients who used the SM function between December 2022 and June 2023 were included. Consequently, this study exclusively comprises HCPs whose patients had sent at least 1 secure message to them.

For the PEPPPSY study, patients were recruited using a purposive sampling strategy, targeting individuals who met the following inclusion criteria: a minimum age of 18 years; a diagnosis of severe mental illness based on the German version of the *International Classification of Diseases, Tenth Revision* (*ICD-10*) criteria; ongoing outpatient treatment in one of the study centers within the last 6 months; and sufficient capacity to provide informed consent for the study. Qualified HCPs (eg, treating psychiatrists or psychotherapists) assessed suitability for participation. Patients with suicidal tendencies, patients who posed a danger to others, as well as patients affected by psychosis or severe cognitive impairment regardless of their diagnosis were excluded from participation in the study. Eligible patients were approached face-to-face by their HCPs and, if interested, were provided with information about the study's objectives and procedures. Those who agreed to participate provided written informed consent. Patients were then introduced to the patient portal by their HCP or a study team member and granted access. They also received a written user manual, which included detailed step-by-step instructions and screenshots. Study team members were available to assist both patients and HCPs with any technical or other questions. Typically, patients accessed the patient portal via their personal digital devices (eg, smartphones or laptops). If patients lacked such devices, they could use a study tablet at the study center to access the portal. However, all participating patients were required to have a device capable of receiving SMS text messages, as this was essential for secure 2-factor authentication to log into the patient portal.

HCPs at the study centers were contacted by the study team and informed about the PEPPPSY study's objective, procedures, and compensation (€50 per HCP). Participating HCPs provided informed consent and were introduced to the patient portal by a member of the study team. HCPs typically accessed the patient portal using their professional computers and professional mobile devices. To be eligible, they had to be older than 18 years, capable of providing informed consent, and employed at one of the study centers.

Data Collection

All participants were informed about the study both verbally and in writing and provided written informed consent. Sociodemographic data were collected. All patients and HCPs were able to contact the study team (patients could also reach out directly to their HCPs for support) at any time if they had technical problems with the patient portal.

All SM between HCPs and patients from December 2022 to July 2023 was retrieved from the PEPPPSY patient portal and stored in a tabular format. For improved clarity and readability, the term messages will consistently be used throughout the paper.

Data Analysis

Thematic analysis of the SM was conducted using the MAXQDA software (VERBI GmbH) [36]. A category system was devised to examine the use patterns of SM exchanged between patients and HCPs. Initially, one-third of the retrieved messages were analyzed deductively, and then, the analysis was extended to include inductive categories emerging from the data [37]. Two researchers performed the initial analysis and ensured consensus. When consensus could not be reached, a third researcher joined the analysis. In addition, participants' sociodemographic data were analyzed based on their messaging behavior. The COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist was used for quality assurance (Multimedia Appendix 1).

Results

Patient Characteristics

A total of 123 patients were included in the overarching PEPPPSY study, where initially 7 patients dropped out due to predefined exclusion criteria and personal reasons. All patients who used the SM function between December 2022 and June 2023 were included in this branch of the study. Thus, a total of 38 patients and 4 HCPs were included in the analysis. For 1 patient, sociodemographic data cannot be reported due to missing values.

Patients ranged in age from 18 to ≥60 years, with a mean age of 46.36 (SD 14.61; median 44.5, IQR 34.5-64.5) years. The overall study population (N=116 patients) had a similar age distribution to our sample, with a mean age of 45.39 (SD 15.76) years. As statistics show, the age distribution of the patients participating in this study also largely matches the age distribution of the general patient population in POCs in Germany (30% of patients are aged between 19 and 45 years and 40% of patients are aged between 45 and 69 years) [35]. Of the 38 patients in this sample, 25 (66%) were female and 13 (34%) were male. In the overall PEPPPSY study population, the gender distribution was comparable (male: 45/116, 38.8% and female: 67/116, 57.8%). Of the 116 participants, 2 (1.7%) did not disclose their gender identity. All patients in our sample were White and spoke German as their first language. Of the 38 patients, 4 (11%) had a family background of migration. The overall study population was slightly more diverse, with 3.5% (4/116) non-White patients. Similar to our sample, 10.3% (12/116) of the patients in the total population had a personal

XSL•FO
RenderX

or family migrant background. In addition, 4.3% (5/116) of the patients did not speak German as their first language. Regarding employment status, 32% (12/38) of the patients in the study sample were employed, while 50% (19/38) were retired. Of the 38 patients, 1 (3%) was unemployed and 5 (13%) were unable to work. One patient (1/38, 3%) did not disclose the employment status. The employment distribution in the total study population was similar, with 32.8% (38/116) employed. However, there was a slightly higher proportion of unemployed patients (10/116, 8.6%) and patients who were unable to work (21/116, 18%), while the proportion of retired patients was significantly lower (44/116, 38%).

Table 2 shows the distribution of *ICD-10* diagnoses for both the entire PEPPPSY study population and our sample. In the *ICD-10* classification system, the numerals F1 to F9 are assigned to represent distinct diagnostic categories. Each code is designated to specify a particular group of mental and behavioral disorders.

Table 2. Distribution of diagnoses according to the German version of the International Classification of Diseases, Tenth Revision (ICD-10) categories in the total population of the PEPPPSY (Piloting and Evaluation of Participatory Patient-Accessible Electronic Health Record in Psychiatry and Somatics) study (n=116) and the sample reported in this study (n=38).

ICD-10 codes	Study population (PEPPPSY study), n (%)	Study population with SM[a] use, n (%)
F1[b]: Mental and behavioral disorders due to psychoactive substance use	2 (1.72)	0 (0)
F2: Schizophrenia, schizotypal, delusional, and other nonmood psychotic disorders	12 (10.34)	5 (13.51)
F3: Mood (affective) disorders	60 (51.72)	17 (45.59)
F4: Anxiety, dissociative, stress-related, somatoform, and other nonpsychotic mental disorders	20 (17.24)	7 (18.92)
F5: Behavioral syndromes associated with physiological disturbances and physical factors	0 (0)	0 (0)
F6: Disorders of adult personality and behavior	16 (13.79)	5 (13.51)
F7: Intellectual disabilities	0 (0)	0 (0)
F8: Pervasive and specific developmental disorders	1 (0.86)	0 (0)
F9: Behavioral and emotional disorders with onset usually occurring in childhood and adolescence	2 (1.72)	1 (2.7)
Missing diagnosis	3 (2.59)	2 (5.41)

[a]SM: secure messaging.

[b]F1-F9 refer to diagnostic categories according to the *ICD-10* classification system, where each code represents a specific group of mental and behavioral disorders.

The diagnoses in our sample largely reflect those in the total study population, although the latter exhibits slightly greater diagnostic diversity. As statistics show, the diagnosis distribution of the patients participating in this study largely corresponds to the diagnosis distribution of the general patient population in POCs in Germany. Most patients in POCs are treated for psychotic disorders (*ICD-10* F2) and affective disorders (*ICD-10* F3). However, compared to the general population of German patients in POCs, substance-related disorders are significantly underrepresented in our sample [38].

HCP Characteristics

All HCPs were male medical physicians trained in psychiatry and behavioral psychotherapy and provided both medical and psychotherapeutic treatment. Of 4 HCPs, 2 (50%) were senior physicians, and 2 (50%) were specialist physicians in psychiatry. Furthermore, 3 HCPs worked part time (50% positions), while 1 HCP held a full-time (100%) position in the POC. HCPs ranged in age from 38 to 56 (mean 45.25, SD 8.14) years. Their clinical experience ranged from 6 to 23 (mean 11.5, SD 7.85) years. Of the 4 HCPs, 2 (50%) stated that they had never shared their clinical notes with patients before the study. The other 2 HCPs, however, reported that they had already occasionally provided patients with (parts of) the clinical documentation, for example, by documenting together or providing printouts. Moreover, 71% of the specialists in psychiatry and psychotherapy working in Germany are aged > 50 years. This means that the HCP study population is somewhat younger than the average [34]. In addition, it is likely that male psychiatrists are significantly overrepresented in our sample compared to the overall population. In 2023, an equal number of new specialist titles in psychiatry and psychotherapy were awarded to women and men [39].

Message Characteristics

A total of 274 messages were analyzed: 61 (22.3%) initial clinical notes written by HCPs, 122 (44.5%) responses from patients, and 91 (33.2%) responses composed by HCPs in reaction to the patients' messages. The number of messages sent by each patient ranged from 1 to 15, with an average of 4.16 and an SD of 3.42. The number of log-ins by patients to the web portal ranged from 1 to 42 (mean 10.78, SD 9.38). Most messages were sent by both patients and HCPs during the daytime. Nevertheless, there were also instances of messages being sent at night, at the earliest hours of the morning, and during the nighttime (Figure 2).

XSL•FO
RenderX

Figure 2. Use patterns among patients and health care professionals (HCPs) of the PEPPPSY (Piloting and Evaluation of Participatory Patient-Accessible Electronic Health Record in Psychiatry and Somatics) patient portal over time, showing the distribution of secure messages.

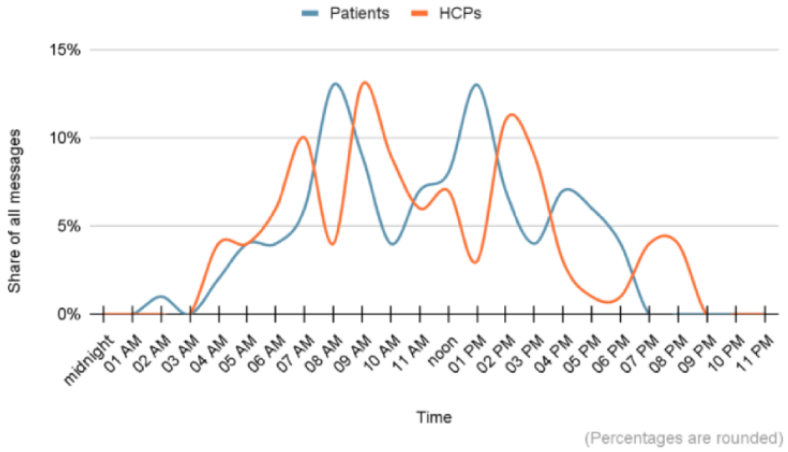

Qualitative Results

All 274 messages were coded. Overall, the following four core functions and a number of subfunctions were identified: (1) reporting or feedback, (2) interpersonal, (3) intrapersonal, and (4) organizational. The presentation follows these core functions and is illustrated by example quotes from the messages of the study participants.

Organizational Function

All 3 types of messages (initial clinical notes, patients' responses, and HCPs' responses) were used to discuss organizational matters. These included matters pertaining to appointments and the issuing of formalities such as sick notes. Example quotes can be found in Table 3. Initial clinical notes contain only very few organizational matters. In their responses to initial clinical notes, patients contact HCPs with organizational requests, queries, and information regarding appointments (including via telephone), as well as the issuance of documents such as sick notes. Many patients are polite in their entries, while others express their concerns urgently or even demandingly. In their responses to patient entries, HCPs usually address the patients' concerns in a friendly, supportive, and transparent manner. A significant proportion of organizational and scheduling requests are accommodated, but it is also clearly communicated when certain needs exceed the treatment scope or technical capabilities. HCPs frequently provided transparent information about their availability and absences, as well as alternative contact options (such as registering with the POC).

Table 3. Example quotes for organizational function, including initial clinical notes by health care professionals (HCPs), patients' responses, and HCPs' replies.

Message type	Example quotes
Initial clinical notes	• "Please bring your treatment records from 2019 with you to your next appointment and also all your sick leave records from 2021, 2022 and 2023." [HCP01]
Patients' responses	• "Please call me today as I urgently need a new prescription and need to discuss some things with you." [PT01] • "Could you please send me a scanned copy of the sick note to my email? I need it for work immediately." [PT02] • "Since my depressive symptoms have tended to worsen over the past week, I was wondering if there would be the possibility of obtaining a sick leave for the rest of the week?" [PT03]
HCPs' responses	• "This [sending scan to the employer] unfortunately doesn't work. But I can assure you that your employer doesn't need a scan directly. We will notify the employer automatically." [HCP02] • "If you need a sick leave for next week due to mental overload, feel free to drop by spontaneously. Unfortunately, I will only be available on Thursday and Friday next week—as those days are already fully booked and I'll be on vacation for 3 weeks afterwards, you can just come by spontaneously and I'll fit you in on Thursday/Friday in between." [HCP03] • "Thanks for the information, that's enough for me. Then we won't do another check-up for about 6 months. You don't need to bring anything with you, I've already looked at it." [HCP04]

Reporting and Feedback Function

Overview

Furthermore, all 3 message types were used to convey treatment-relevant information or to elicit feedback. Table 4 presents sample example quotes of the reporting and feedback functions found in each message type.

XSL•FO
RenderX

Table 4. Example quotes for reporting and feedback functions, including initial clinical notes by health care professionals (HCPs), patients' responses, and HCPs' replies.

Message type and reporting on	Example quotes
Initial clinical notes	
Current situation	• "Patient herself reports gradually increasing memory problems, especially forgetting appointments and conversations, probably said the same thing more often. The son reports that when the husband died in 2020, the things he had previously taken over no longer worked." [HCP01] • "Multiple pre-existing conditions, malfunctioning pituitary gland leading to increased urine production, requires self-administration of cortisone. His major issue is the lack of energy, along with constant sweating. Overall, there is excessive need for sleep. Additional problems include low mood and lack of drive." [HCP02]
Treatment	• "Personal appointment: Midday dose of haloperidol made quite tired in the afternoon, stopped on his own. Quetiapine increased, short-term follow-up appointment." [HCP04] • "So it was discussed with the patient that it was very important to stand up for herself and to set boundaries. This is just as important for the family, for family members who have concerns. Discussed with the patient how she could do this with her family." [HCP03]
Patients' responses	
Current situation	• "At the moment, I am feeling very well, but I haven't told you yet that I've had a high heart rate for years. My average is around 110." [PT05] • "I am under constant tension, stress, feeling unmotivated, irritable, and prone to outbursts. I just can't seem to calm down anymore. My colleagues at work, who have a strong backbone, are experiencing the same thing. We have stomach and intestinal complaints in the office due to all the frustration because nothing changes. My migraine attacks and physical joint complaints have also increased again. I am experiencing insomnia despite taking sleeping pills." [PT06] • "My self-sacrifice began in my childhood, around the age of 10. I had to do everything perfectly and neatly in the family home with my mother's changing partners. She also expected me to look after my younger siblings....I still suffer from this." [PT06]
Treatment	• "The Haloperidol works very well in the morning. The evening medication is also right and important as it is." [PT07] • "I have tolerated it well so far, I also notice a slight improvement. However, the tension is still very high several times a day. Does it take a while for the medication to build up (as with antidepressants, for example)?" [PT08]
Initial clinical notes	• "Hypohedonia? A hint on whether that requires treatment would be very helpful for me, or if you've noticed this more often." [PT09] • "Very well summarized. Therapy sessions and subsequent reading help in working towards the goal." [PT06] • "Thank you for your note. Wouldn't it also make sense to describe why I've been gradually tapering off the medication since summer, in consultation with my GP?" [PT10]
Experiences with SM[a]	• "I've also discovered the commenting feature now. You just need to click on the displayed entry—I should be more courageous." [PT11] • "I'm not sure if this [SM] is even meant for that purpose, but since I find it difficult to mention things/problems, etc., in the appointment, I just wanted to write something briefly here." [PT12]
HCPs' responses	
Patients' entries	• "Thank you very much for your report, which is very helpful for further treatment decisions. I have read it and will incorporate it into the next appointment." [HCP02] • "Oh, you're absolutely right! I fortunately noted it correctly in our medication module and corrected it here." [HCP03]
Treatment and current situation	• "It's indeed no problem to switch to Bupropion or Moclobemide. Would you like to come in on Monday for that?" [HCP02] • "Firstly, bisoprolol has fewer side effects than metoprolol and can do the same thing, which is good. Secondly, you are welcome to submit the documents, but I cannot promise that I will be able to process them before Christmas. It often makes sense to clarify open questions together. In any case, it would make sense to add half an hour to the December appointment. Please let me know when you hand in your documents at the registration desk." [HCP01]
Experiences with SM	• "Please feel free to continue documenting here if you'd like. It is helpful for our next appointment." [HCP02] • "By the way, the issue of comments appearing more than once is more of a technical problem; it's happening to others as well." [HCP01] • "I'm sorry I'm just now replying, unfortunately the programme didn't show correctly that you had written something. Do you have any questions? If so, please make an appointment with me. Kind regards" [HCP03]

[a]SM: secure messaging.

XSL•FO
RenderX

Initial Clinical Notes

In initial clinical notes of the HCPs, the reporting function predominated. HCPs described the current psychological, somatic, and social condition through the patient's medical and sociobiographical history, previous treatments, current symptoms, social and occupational situation, and acute stressors in daily life. This diverse information enables a holistic understanding of the patient's current concerns and sheds light on the reciprocal interactions between somatic health, mental health, and environmental factors. Furthermore, HCPs documented information on previous, current, and planned therapeutic and psychotherapeutic and medical treatments, usually considering the patient's wishes. HCPs seem to take great care to report information in neutral, nonjudgmental language and usually do not address patients directly (refer to the Interpersonal Function section for exceptions). Although HCPs mention their own impressions of the patient's status in the messages, they mostly focus on the patient's own perception and refer to it in direct quotations or indirect speech (using phrases such as "Patient summarizes..." and "Patient reports...").

Patients' Responses

In their responses, patients very frequently provided specific feedback on the course of treatment, particularly regarding the tolerability of the medication. While some patients conveyed their experiences in an objective manner or expressed agreement with the practitioners' approach, others voiced doubts, concerns, or hopes regarding their treatment. Specifically, when patients harbored doubts about their medical treatment or felt dissatisfied with side effects or the absence of improvement, they turned to the HCPs with direct questions. Subsequently, the patient would inquire about potential adjustments to their medication or seek clarification on its mechanism of action. Furthermore, patients supplement the information documented by HCPs in the initial clinical notes and provide feedback on changes to their current psychological and somatic condition, as well as on changes or stressors in their social or professional environment. In addition, some patients offer further insight into their personal biographical history, which seems pertinent to the development of their mental health condition and its (psychotherapeutic) treatment. For some patients, reporting appears to have a relieving effect, allowing them to express their distress instead of carrying it alone. Others use the entries as a sort of diary for self-motivation and empowerment. This possibility to relieve oneself appears to be of particular significance for patients with severe mental health conditions and is perceived as beneficial for their mental health. In addition to stressors and deteriorations, small improvements and joyful moments are also shared with the HCPs. A novel aspect is that patients provide explicit feedback on the initial treatment documentation by their HCPs as well as on the SM web application. Regarding SM, issues pertaining to usability and accessibility (eg, technical difficulties, errors, and features) are typically addressed, particularly when patients have cognitive impairments. However, patients also refer to the initial documentation in terms of content; they pose questions, express rejections and suggestions for corrections, and also offer praise and agreement. Furthermore, patients reassure themselves with their HCPs that they are allowed to use the web application for writing entries.

HCPs' Responses

In addition, HCPs reference the entries of the patients in their responses; they express gratitude for composing the comments, ascribe significance to them, and encourage further use of the function. Furthermore, they rectify any potential errors in their initial clinical notes and pose content-related questions regarding the entries of the patients and their current situation. The responses to patients' questions on further treatment are of a high standard. Patients receive timely appointment offers, psychoeducational explanations, treatment recommendations, information on dealing with emergencies, and collaborations with other HCPs. In the event of technical difficulties with the patient portal, HCPs reassure patients and clarify that the problems are not the patients' fault but rather technical errors in the system. Overall, it is noticeable that the HCPs patiently and empathetically address every concern of the patients.

Interpersonal Function

Overview

In most entries from both HCPs and patients, explicit relationship messages can be identified (Table 5). In addition, the entries also contain a multitude of implicit relationship messages. For example, the predominantly empathetic and compassionate composition of the entries already indicates a certain fundamental interpersonal character in almost all entries. At the same time, they are kept very professional and appropriately focused. Furthermore, the salutation and farewell messages (which range from formal to friendly and from distant to familiar) contain relevant relational aspects. Interpersonal aspects can thus be found in almost all other categories in one way or another. However, because interpersonal aspects are both distinct and highly important, they are discussed separately in this section.

XSL•FO
RenderX

Table 5. Example quotes for interpersonal function, including initial clinical notes by health care professionals (HCPs), patients' responses, and HCPs' replies.

Message type	Example quotes
Initial clinical notes	• "Discussed with patient that if condition worsens, sick leave is possible at any time. Should take good care of herself, set boundaries if something becomes too much, listen to herself." [PT03] • "Please observe in the next period: Sweating? Improvement in daytime drive/sleep needs? Erectile problems?" [HCP02]
Patients' responses	• "Thank you for listening and for these messages here." [PT10] • "By the way, I'm terribly afraid that I'm bothering you, that you might think I'm stupid... That's why I've been 'writing' this here for almost 2 months." [PT12]
HCPs' responses	• "I'm sorry to hear that the group therapy didn't go as you hoped—if I understand correctly. I can only say that I am sure both the therapist and the fellow patients should (and will) have the utmost understanding that not everything always goes as planned in such a group. I want to encourage you and say that you are not to blame. I believe whatever has been/is okay. I hope you can put these thoughts aside. If you feel worse, please come to the outpatient clinic." [HCP02] • "I hope you were able to regain some stability in this difficult phase of life. Feel free to schedule another appointment if needed. I wish you much strength!"[HCP03]

Initial Clinical Notes

Given that HCPs primarily focus on reporting in the initial clinical notes, the interpersonal function is only subtly present. Nevertheless, a few HCPs include direct questions to patients (regarding the course and tolerability of medications) or "homework" in the initial clinical notes, thus initiating an active exchange.

Patients' Responses

In their responses, patients convey relationship messages in multifaceted ways. A common theme in patients' responses is gratitude for the messages, responses, and support provided by HCPs. Furthermore, many HCPs also receive positive sentiments, such as well-wishes from patients. Both gratitude and good wishes can contribute to the strengthening of the therapeutic relationship. Similarly, patients use the entries to seek reassurance, unload burdens, or address sensitive topics that they may be reluctant to discuss in person. In general, patients write their entries in a friendly and appropriate manner, which is almost professional and unobtrusive. However, in other responses, their distress becomes more apparent and is (more or less) implicitly directed as an appeal to the HCPs.

HCPs' Responses

The responses of HCPs very frequently include messages and offers related to the establishment and maintenance of the therapeutic relationships. HCPs strengthen, encourage, and reassure patients through their compassionate and empathetic messages. As frequently described, HCPs always address the wishes and questions of patients in a patient and considerate manner. Similar to the patients, HCPs promote a good therapeutic relationship by expressing gratitude and good wishes to the patients and repeatedly encourage further exchange by posing specific questions. Furthermore, HCPs endeavor to foster transparency regarding their treatment and documentation practices and are not afraid to use self-disclosures. They respond authentically to patients, expressing their joy over (small) progress and demonstrating genuine understanding and compassion in the face of setbacks.

Intrapersonal Function

In the initial clinical notes, and particularly in the responses of the patients, an intrapersonal function of the SMS text messages could be identified. Both HCPs and patients used their messages as a direct or more indirect "note to themselves." Table 6 provides example quotes. HCPs occasionally record information in the initial clinical notes that serve as reminders for their next appointment, specifying what they should pay attention to. This information is typically brief and documented only with keywords. These messages often contain questions and naturally also have an interpersonal character, as previously discussed. Patients use their entries to reflect on their treatment, recap their insights, and provide themselves with words of encouragement.

Table 6. Example quotes for intrapersonal function, including initial clinical notes by health care professionals (HCPs) and patient's response.

Message type	Example quotes
Initial clinical notes	• "For the next appointment: Plan blood sample, document weight." [HCP04] • "Alternative medication: Milnacipran? Bupropion?" [HCP02]
Patient's response	• "Realization: Patience pays off—miracles are possible...Do not be intimidated, continue exercising, overcome fear of driving, and overall work on self-confidence." [PT11]

XSL•FO
RenderX

Discussion

Principal Findings

In the clinical text messages of patients affected by mental health conditions and HCPs at outpatient clinics in a rural area of Brandenburg, 4 main functions have been identified: organization, reporting and feedback, interpersonal uses, and intrapersonal uses. Despite the pervasive early fears about using SM with patients with psychiatric disorders, the interactions between HCPs and patients via SM were overwhelmingly positive in this study. This discrepancy between the anticipated and actual outcomes of SM is similarly reflected in other current studies [14]. The predominantly patient-centered, respectful, and friendly composition of the entries aligns with other empirical evidence [40,41].

Communication via SM poses challenges not only for patients but also for HCPs. The possible multitude of messages that must be managed alongside the already tightly organized treatment routine can be overwhelming [22,42]. Given that patients' responses via patient portals could be read and answered at any time, including weekends or after work hours, patients may be under the impression that HCPs are available at all times via SM and will respond promptly to any queries [43,44]. Thus, HCPs must be particularly mindful of their own boundaries [44]. In our experience, attempts to facilitate boundary setting by shutting down the patient portal on weekends caused more confusion than relief among both HCPs and patients. In our study as well as those by others, HCPs endeavor to respond to queries in a timely manner, yet their work demands often preclude them from doing so. Consequently, responses may be delayed or occur outside of their scheduled clinic hours [45,46]. Furthermore, delays may occur due to technical issues, such as messages not being displayed immediately [5,23,42]. On the basis of the use patterns of patients in this study, it can be assumed that patients primarily contact their HCPs during main working hours. Nevertheless, web portals should provide clear indications that HCPs will not read and respond to entries immediately, emphasizing that it is an asynchronous communication process. This is particularly pertinent in the context of acute crises, such as suicidal intentions [47,48]. In the event of a crisis, it is imperative that the web portal provides clear and readily accessible information on emergency numbers. No such emergency occurred in this study. However, in emerging crises, HCPs provided relevant emergency contact information to patients via SM. As the web application develops, it may be possible to implement an automated search function that identifies messages containing specific keywords (such as suicidal, suicide, and self-harm) using artificial intelligence technology [49]. This then could trigger emergency responses or display the appropriate emergency contact numbers.

As the results of this study show, both HCPs and patients use SM for organizational matters (organizational function). Communication via the patient portal can offer a simple and fast alternative to telephone or in-person coordination in the POC, which is often associated with long waiting times. However, HCPs must then reliably and consistently review and respond to incoming messages, which could lead to an additional workload. Besides the patients, only medical personnel had access to our web application. However, nonmedical practice assistants could also receive access to the patient portal to assist and support with organizational matters, thereby relieving the workload of HCPs. As is shown in our study as well as in those by others, this seems to be particularly useful for dealing with organizational issues, such as scheduling appointments via patient portals, which was frequently observed in the SM interactions [50].

SM offers patients the opportunity to provide feedback and ask questions related to their HCPs' clinical notes (reporting and feedback function). This allows patients to better understand their condition and treatment while also identifying potential errors and clarifying misunderstandings. Such processes were frequently observed in the message exchanges analyzed in this study. In line with previous research, it can be assumed that SM enhances treatment transparency and strengthen patient engagement in their own care.

Alongside the challenges it presents, communication via SM offers several benefits. In alignment with other studies, our research findings indicate that enhancing the therapeutic communication represents one of the most significant potential benefits (interpersonal factor) [44,48]. Some patients may feel more comfortable discussing sensitive topics through this mediated format than in direct contact with HCPs, often because of feelings of shame. In outpatient psychiatric care with brief and infrequent therapeutic contacts, patients often encounter difficulties in developing sufficient trust to openly discuss shameful or stigmatizing topics with their HCPs [51]. The interpersonal function of communication via SM facilitates more frequent and closer contact with HCPs, thereby possibly strengthening the therapeutic contact [52]. Moreover, HCPs sometimes seem to act as auxiliary egos for patients, offering support in important functions of the self, such as holding, containing, marking emotions, and self-reflecting [53]. However, it also becomes evident that in nearly every entry—often in just 1 sentence—many of these functions simultaneously come into play, interact, and overlap. On a communicative level, a sentence can possess diverse functions. Thus, it can be presumed that already established models of multifaceted communication, such as those by von Thun [54], also apply to communication mediated via SM [54]. Although this communication is conducted through digital and written channels, it is not merely a matter of exchanging factual and technical information. Instead, it is an intricate process of communication between HCPs and patients, with similarities to nondigital forms of communication. Conversely, the absence of nonverbal cues and the lack of immediacy inherent to digital communication via SM present HCPs with a set of novel challenges [42]. These insights must be considered when composing clinical notes and responses. It requires appropriate training and guidelines so that HCPs can have a certain level of expertise and confidence in this area [14,23,47].

Furthermore, SM offers patients the opportunity to engage with the content of their therapeutic sessions even after their appointments have ended (intrapersonal function). Some HCPs assign dedicated "homework" to patients, while other patients describe how reading the initial clinical notes and subsequent

XSL•FO
RenderX

SM has helped them continue reflecting on the session's content beyond the face-to-face therapy. This suggests that therapeutic interactions may extend into the digital space through SM.

Regarding the technical realization of SM in Germany, it can be stated that the German digital health agency "gematik" has briefly certified 7 different SM apps—so-called telematic infrastructure messengers, which all communicate using the matrix protocol via the German health data infrastructure [55]. The telematic infrastructure messenger currently supports communication only between HCPs, with patient-to-HCP communication planned for a future update. The patient portals enabled by the Hospital Future Act, which in turn should allow asynchronous, secure communication between patients treated by a hospital and their HCPs, are still used in only a few cases and have not yet met expectations. In this respect, the intervention of a patient portal with open notes and text-based asynchronous patient-HCP communication, as tested here, will only be possible under study conditions in the near future, at least in Germany [55].

Limitations

The explorative nature of the study and its qualitative approach, while providing in-depth insights, may not be generalizable to the broader population. Furthermore, the sample was drawn from 3 POCs in a rural region, which may not represent the diversity of patients in different (eg, more metropolitan) geographical locations or other health care settings, such as the somatic field. In addition, no validated instruments were used to assess the digital and technical literacy of patients and HCPs. As a result, the study could not account for varying levels of technology and digital literacy among patients, which could have led to the exclusion of patients with lower technology or digital literacy. In contrast, our experience showed that even patients with limited digital skills were willing to participate in the study, though they required increased support and guidance from the study team. To foster the engagement of patient SM in routine care, trained nonmedical professionals (so-called "digital navigators") who assist patients and HCPs in using the technology, thereby alleviating the burden on the treatment team, could be a promising option [56]. The health literacy and overall communication skills of the patients were also not specifically assessed, which may represent a limiting factor.

Furthermore, a potential gatekeeper effect should be considered, given that HCPs were responsible for patient recruitment in the PEPPPSY study [57,58]. However, all HCPs received comprehensive training on the eligibility criteria in advance and included a diverse range of patients with various psychiatric disorders and varying levels of digital literacy. Moreover, a potential self-selection bias should be considered. Patients who opt to use SM might be more engaged or proactive about their health care, whereas those who do not use the messaging feature might have different, unstudied needs or barriers. As nonusers of SM were not included, conclusions drawn from this study may primarily reflect the experiences of more engaged or technologically adept patients. In addition, by excluding patients with acute risk factors and significant cognitive impairments, the study might overlook the perspectives of those who are most vulnerable and might have unique challenges or needs related to the use of patient portals. Further research should prioritize the implementation of SM within these patient populations. In this study, SM was analyzed exclusively in the context of open medical record access. Future research could explore the use of SM in broader or alternative contexts.

Conclusions

The study demonstrated the multifaceted nature of communication through SM in psychiatric outpatient care of patients affected by severe mental health conditions, with messages serving multiple purposes simultaneously. The complex communication via SM resembles traditional forms of communication while presenting unique challenges and opportunities. In addition to the necessity for training and guidelines for HCPs and personal assistance for patients to navigate digital communication effectively, structural adjustments in workflow are essential to manage the increased communication load and maintain professional boundaries.

Despite these challenges, SM offers notable benefits, including enhanced therapeutic contact and increased accessibility to care. However, further research is needed to ensure the meaningful use and responsible implementation of SM in health care, with particular attention to ethical, legal, and practical considerations.

Acknowledgments

Funded by the Brandenburg Medical School publication fund supported by the Ministry of Science, Research and Cultural Affairs of the State of Brandenburg.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

All authors contributed to planning, analysis, and the critical interpretation of findings. J Schwarz and EM-D contributed to the study design, and J Schwarz supervised the study. J Speck and JMW contributed to data collection. EM-D conducted the data analysis. EM-D wrote the first draft of the manuscript and prepared all tables and figures. EM-D, J Schwarz, and CT modified successive drafts. J Schwarz edited and approved the final version.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist.
[PDF File (Adobe PDF File), 75 KB - formative_v9i1e63713_app1.pdf]

References

1. Lawal A, Menon D, Affleck E, Stafinski T. The impact of secure messaging in the treatment of patients with diabetes within a primary care setting: protocol for a scoping review. JMIR Res Protoc 2023 May 02;12:e42339 [FREE Full text] [doi: 10.2196/42339] [Medline: 37129935]
2. Antonio MG, Petrovskaya O, Lau F. Correction: the state of evidence in patient portals: umbrella review. J Med Internet Res 2021 Aug 16;23(8):e32421 [FREE Full text] [doi: 10.2196/32421] [Medline: 34398799]
3. Cronin RM, Davis SE, Shenson JA, Chen Q, Rosenbloom ST, Jackson GP. Growth of secure messaging through a patient portal as a form of outpatient interaction across clinical specialties. Appl Clin Inform 2015;6(2):288-304 [FREE Full text] [doi: 10.4338/ACI-2014-12-RA-0117] [Medline: 26171076]
4. Shenson JA, Cronin RM, Davis SE, Chen Q, Jackson GP. Rapid growth in surgeons' use of secure messaging in a patient portal. Surg Endosc 2016 Apr;30(4):1432-1440 [FREE Full text] [doi: 10.1007/s00464-015-4347-y] [Medline: 26123340]
5. Hefner JL, MacEwan SR, Biltz A, Sieck CJ. Patient portal messaging for care coordination: a qualitative study of perspectives of experienced users with chronic conditions. BMC Fam Pract 2019 May 03;20(1):57 [FREE Full text] [doi: 10.1186/s12875-019-0948-1] [Medline: 31053063]
6. Naiker U, FitzGerald G, Dulhunty JM, Rosemann M. Time to wait: a systematic review of strategies that affect out-patient waiting times. Aust Health Rev 2018 Jun;42(3):286-293. [doi: 10.1071/AH16275] [Medline: 28355525]
7. Rastpour A, McGregor C. Predicting patient wait times by using highly deidentified data in mental health care: enhanced machine learning approach. JMIR Ment Health 2022 Aug 09;9(8):e38428 [FREE Full text] [doi: 10.2196/38428] [Medline: 35943774]
8. Blease C, McMillan B, Salmi L, Davidge G, Delbanco T. Adapting to transparent medical records: international experience with "open notes". BMJ 2022 Nov 21;379:e069861. [doi: 10.1136/bmj-2021-069861] [Medline: 36410770]
9. Hägglund M. Electronic health records in Sweden—how can we go from transparency to collaboration? The BMJ Opinion. 2017. URL: https://blogs.bmj.com/bmj/2017/06/23/maria-hagglund-electronic-health-records-in-sweden-how-can-we-go-transparency-to-collaboration/ [accessed 2023-07-14]
10. Hägglund M, Kharko A, Hagström J, Bärkås A, Blease C, Cajander Å, et al. The NORDeHEALTH 2022 patient survey: cross-sectional study of national patient portal users in Norway, Sweden, Finland, and Estonia. J Med Internet Res 2023 Nov 13;25:e47573 [FREE Full text] [doi: 10.2196/47573] [Medline: 37955963]
11. Jormanainen V. Large-scale implementation and adoption of the Finnish national Kanta services in 2010–2017: a prospective, longitudinal, indicator-based study. FinJeHeW 2018 Dec 04;10(4):15. [doi: 10.23996/fjhw.74511]
12. Wurster F, Di Gion P, Goldberg N, Hautsch V, Hefter K, Herrmann C, et al. Roger's diffusion of innovations theory and the adoption of a patient portal's digital anamnesis collection tool: study protocol for the MAiBest project. Implement Sci Commun 2024 Jul 15;5(1):74 [FREE Full text] [doi: 10.1186/s43058-024-00614-8] [Medline: 39010236]
13. Wade-Vuturo AE, Mayberry LS, Osborn CY. Secure messaging and diabetes management: experiences and perspectives of patient portal users. J Am Med Inform Assoc 2013 May 01;20(3):519-525 [FREE Full text] [doi: 10.1136/amiajnl-2012-001253] [Medline: 23242764]
14. Sieck CJ, Hefner JL, Schnierle J, Florian H, Agarwal A, Rundell K, et al. The rules of engagement: perspectives on secure messaging from experienced ambulatory patient portal users. JMIR Med Inform 2017 Jul 04;5(3):e13 [FREE Full text] [doi: 10.2196/medinform.7516] [Medline: 28676467]
15. Han HR, Gleason KT, Sun C, Miller HN, Kang SJ, Chow S, et al. Using patient portals to improve patient outcomes: systematic review. JMIR Hum Factors 2019 Dec 19;6(4):e15038 [FREE Full text] [doi: 10.2196/15038] [Medline: 31855187]
16. Saberi P, Catz SL, Leyden WA, Stewart C, Ralston JD, Horberg MA, et al. Antiretroviral therapy adherence and use of an electronic shared medical record among people living with HIV. AIDS Behav 2015 Jun;19 Suppl 2(0 2):177-185 [FREE Full text] [doi: 10.1007/s10461-014-0982-x] [Medline: 25572829]
17. de Lusignan S, Mold F, Sheikh A, Majeed A, Wyatt JC, Quinn T, et al. Patients' online access to their electronic health records and linked online services: a systematic interpretative review. BMJ Open 2014 Sep 08;4(9):e006021 [FREE Full text] [doi: 10.1136/bmjopen-2014-006021] [Medline: 25200561]
18. Lyles CR, Sarkar U, Schillinger D, Ralston JD, Allen JY, Nguyen R, et al. Refilling medications through an online patient portal: consistent improvements in adherence across racial/ethnic groups. J Am Med Inform Assoc 2016 Apr;23(e1):e28-e33 [FREE Full text] [doi: 10.1093/jamia/ocv126] [Medline: 26335983]

XSL•FO
RenderX

19. North F, Crane SJ, Chaudhry R, Ebbert JO, Ytterberg K, Tulledge-Scheitel SM, et al. Impact of patient portal secure messages and electronic visits on adult primary care office visits. Telemed J E Health 2014 Mar;20(3):192-198 [FREE Full text] [doi: 10.1089/tmj.2013.0097] [Medline: 24350803]
20. Turvey CL, Fuhrmeister LA, Klein DM, Moeckli J, Howren MB, Chasco EE. Patient and provider experience of electronic patient portals and secure messaging in mental health treatment. Telemed J E Health 2022 Feb;28(2):189-198 [FREE Full text] [doi: 10.1089/tmj.2020.0395] [Medline: 33887164]
21. Tieu L, Sarkar U, Schillinger D, Ralston JD, Ratanawongsa N, Pasick R, et al. Barriers and facilitators to online portal use among patients and caregivers in a safety net health care system: a qualitative study. J Med Internet Res 2015 Dec 03;17(12):e275 [FREE Full text] [doi: 10.2196/jmir.4847] [Medline: 26681155]
22. Kruse CS, Argueta DA, Lopez L, Nair A. Patient and provider attitudes toward the use of patient portals for the management of chronic disease: a systematic review. J Med Internet Res 2015 Feb 20;17(2):e40 [FREE Full text] [doi: 10.2196/jmir.3703] [Medline: 25707035]
23. Laukka E, Huhtakangas M, Heponiemi T, Kujala S, Kaihlanen A, Gluschkoff K, et al. Health care professionals' experiences of patient-professional communication over patient portals: systematic review of qualitative studies. J Med Internet Res 2020 Dec 08;22(12):e21623 [FREE Full text] [doi: 10.2196/21623] [Medline: 33289674]
24. Schwarz J, Esch T. OpenNotes in mental health? Pro. Nervenarzt 2022 Nov 01;93(11):1163-1164. [doi: 10.1007/s00115-022-01357-4] [Medline: 35913587]
25. Schwarz J, Meier-Diedrich E, Neumann K, Heinze M, Eisenmann Y, Thoma S. Reasons for acceptance or rejection of online record access among patients affected by a severe mental illness: mixed methods study. JMIR Ment Health 2024 Feb 05;11:e51126 [FREE Full text] [doi: 10.2196/51126] [Medline: 38315523]
26. Schwarz J, Neumann K, Meier-Diedrich E, Mühlensiepen F, Hochwarter S, Blease C. Psychotherapists' views on open notes: an online survey from Germany. Digit Health 2024 Sep 11;10:20552076241271813 [FREE Full text] [doi: 10.1177/20552076241271813] [Medline: 39291155]
27. Meier-Diedrich E, Neumann K, Heinze M, Schwarz J. Attitudes and expectations of psychological and medical psychotherapists towards open notes: analysis of qualitative survey responses. Psychiatr Prax 2024 Oct;51(7):376-384. [doi: 10.1055/a-2320-8929] [Medline: 38810902]
28. Schwarz J, Hoetger C, Pluschke LS, Muehlensiepen F, Schneider M, Thoma S, et al. Psychiatrists' perceptions of conditions and consequences associated with the implementation of open notes: qualitative investigation. BMC Psychiatry 2024 Jun 10;24(1):430 [FREE Full text] [doi: 10.1186/s12888-024-05845-6] [Medline: 38858711]
29. Stewart MT, Hogan TP, Nicklas J, Robinson SA, Purington CM, Miller CJ, et al. The promise of patient portals for individuals living with chronic illness: qualitative study identifying pathways of patient engagement. J Med Internet Res 2020 Jul 17;22(7):e17744 [FREE Full text] [doi: 10.2196/17744] [Medline: 32706679]
30. Turvey C, Fuhrmeister L, Klein D, McCoy K, Moeckli J, Stewart Steffensmeier KR, et al. Secure messaging intervention in patients starting new antidepressant to promote adherence: pilot randomized controlled trial. JMIR Form Res 2023 Dec 08;7:e51277 [FREE Full text] [doi: 10.2196/51277] [Medline: 38064267]
31. Brearly TW, Goodman CS, Haynes C, McDermott K, Rowland JA. Improvement of postinpatient psychiatric follow-up for veterans using telehealth. Am J Health Syst Pharm 2020 Feb 07;77(4):288-294. [doi: 10.1093/ajhp/zxz314] [Medline: 32031208]
32. Hochwarter S, Fehler G, Muente C, Eisenmann Y, Heinze M, Hibsch C, et al. Design of a patient-accessible electronic health record system in mental health. Stud Health Technol Inform 2022 May 25;294:583-584. [doi: 10.3233/SHTI220534] [Medline: 35612155]
33. Hochwarter S, Schwarz J. Facilitating collaboration by opening the common information space in mental health care. In: Proceedings of the 9th International Conference on Infrastructures in Healthcare. 2023 Presented at: Infrahealth '23; September 11-12, 2023; Siegen, Germany p. 22 URL: https://www.researchgate.net/publication/373580272_Facilitating_Collaboration_by_Opening_the_Common_Information_Space_in_Mental_Health_Care [doi: 10.48340/ihc2023_p013]
34. Basisdaten zu psychischen Erkrankungen in Deutschland. DGPPN. URL: https://www.dgppn.de/_Resources/Persistent/3067cbcf50e837c89e2e9307cecea8cc901f6da8/DGPPN_Factsheet_Kennzahlen.pdf [accessed 2025-02-07]
35. Klauber J. Krankenhaus-Report 2016: Schwerpunkt: Ambulant im Krankenhaus - Mit Online-Zugang. Berlin, Germany: Schattauer; 2016.
36. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
37. Clarke V, Braun V, Hayfield N, Terry G. Thematic analysis. In: Liamputtong P, editor. Handbook of Research Methods in Health Social Sciences. Cham, Switzerland: Springer; 2015:843-860.
38. Koch-Stoecker S, Driessen M, Gouzoulis-Mayfrank E, Pollmächer T. Structure and spectrum of activities of psychiatric outpatient clinics in Germany. Psychiatr Prax 2016 Apr;43(3):129-130. [doi: 10.1055/s-0042-102862] [Medline: 27055980]
39. Ärztestatistik zum 31. Dezember 2023. Bundesärztekammer. URL: https://www.bundesaerztekammer.de/fileadmin/user_upload/BAEK/Ueber_uns/Statistik/AErztestatistik_2023_Update_Juni_2024.pdf [accessed 2025-02-07]

XSL•FO
RenderX

40. Hogan TP, Luger TM, Volkman JE, Rocheleau M, Mueller N, Barker AM, et al. Patient centeredness in electronic communication: evaluation of patient-to-health care team secure messaging. J Med Internet Res 2018 Mar 08;20(3):e82 [FREE Full text] [doi: 10.2196/jmir.8801] [Medline: 29519774]
41. Alpert JM, Wang S, Bylund CL, Markham MJ, Bjarnadottir RI, Lee J, et al. Improving secure messaging: a framework for support, partnership and information-giving communicating electronically (SPICE). Patient Educ Couns 2021 Jun;104(6):1380-1386. [doi: 10.1016/j.pec.2020.11.021] [Medline: 33280967]
42. Kelly MM, Dean SM, Carayon P, Wetterneck TB, Hoonakker PL. Healthcare team perceptions of a portal for parents of hospitalized children before and after implementation. Appl Clin Inform 2017 Mar 15;8(1):265-278 [FREE Full text] [doi: 10.4338/ACI-2016-11-RA-0194] [Medline: 28293685]
43. Shimada SL, Petrakis BA, Rothendler JA, Zirkle M, Zhao S, Feng H, et al. An analysis of patient-provider secure messaging at two veterans health administration medical centers: message content and resolution through secure messaging. J Am Med Inform Assoc 2017 Sep 01;24(5):942-949 [FREE Full text] [doi: 10.1093/jamia/ocx021] [Medline: 28371896]
44. Lieu TA, Altschuler A, Weiner JZ, East JA, Moeller MF, Prausnitz S, et al. Primary care physicians' experiences with and strategies for managing electronic messages. JAMA Netw Open 2019 Dec 02;2(12):e1918287 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.18287] [Medline: 31880798]
45. Rohrer JE, North F, Angstman KB, Oberhelman SS, Meunier MR. Timely response to secure messages from primary care patients. Qual Manag Health Care 2013;22(2):161-166. [doi: 10.1097/QMH.0b013e31828be314] [Medline: 23542371]
46. North F, Luhman KE, Mallmann EA, Mallmann TJ, Tulledge-Scheitel SM, North EJ, et al. A retrospective analysis of provider-to-patient secure messages: how much are they increasing, who is doing the work, and is the work happening after hours? JMIR Med Inform 2020 Jul 08;8(7):e16521 [FREE Full text] [doi: 10.2196/16521] [Medline: 32673238]
47. Hefner JL, Sieck CJ, Walker DM, Huerta TR, McAlearney AS. System-wide inpatient portal implementation: survey of health care team perceptions. JMIR Med Inform 2017 Sep 14;5(3):e31 [FREE Full text] [doi: 10.2196/medinform.7707] [Medline: 28912115]
48. Hoonakker PL, Carayon P, Cartmill RS. The impact of secure messaging on workflow in primary care: results of a multiple-case, multiple-method study. Int J Med Inform 2017 Apr;100:63-76 [FREE Full text] [doi: 10.1016/j.ijmedinf.2017.01.004] [Medline: 28241939]
49. Duvall M, North F, Leasure W, Pecina J. Patient portal message characteristics and reported thoughts of self-harm and suicide: a retrospective cohort study. J Telemed Telecare 2021 Sep;27(8):501-508. [doi: 10.1177/1357633X19887262] [Medline: 31726902]
50. Masterman M, Cronin RM, Davis SE, Shenson JA, Jackson GP. Adoption of secure messaging in a patient portal across pediatric specialties. AMIA Annu Symp Proc 2016;2016:1930-1939 [FREE Full text] [Medline: 28269952]
51. Haun JN, Lind JD, Shimada SL, Martin TL, Gosline RM, Antinori N, et al. Evaluating user experiences of the secure messaging tool on the Veterans Affairs' patient portal system. J Med Internet Res 2014 Mar 06;16(3):e75 [FREE Full text] [doi: 10.2196/jmir.2976] [Medline: 24610454]
52. Kruse CS, Bolton K, Freriks G. The effect of patient portals on quality outcomes and its implications to meaningful use: a systematic review. J Med Internet Res 2015 Feb 10;17(2):e44 [FREE Full text] [doi: 10.2196/jmir.3171] [Medline: 25669240]
53. Heigl-Evers A, Ott J. Die psychoanalytisch-interaktionelle Methode. Psychotherapeut 1996 Mar 1;41(2):77-83. [doi: 10.1007/S002780050020]
54. von Thun FS. Miteinander reden 1: Störungen und Klärungen: Allgemeine Psychologie der Kommunikation. Berlin, Germany: Rowohlt Verlag; 2013.
55. gematik GmbH. TI-Messenger. URL: https://www.gematik.de/anwendungen/ti-messenger [accessed 2024-05-05]
56. Perret S, Alon N, Carpenter-Song E, Myrick K, Thompson K, Li S, et al. Standardising the role of a digital navigator in behavioural health: a systematic review. Lancet Digit Health 2023 Dec;5(12):e925-e932 [FREE Full text] [doi: 10.1016/S2589-7500(23)00152-8] [Medline: 38000876]
57. Reinders H. Qualitative Interviews mit Jugendlichen führen: Ein Leitfaden. München, Germany: Oldenbourg Wissenschaftsverlag; 2012.
58. Merkens H. Stichproben bei qualitativen Studien. In: Friebertshäuser B, Prengel A, editors. Handbuch Qualitative Forschungsmethoden in der Erziehungswissenschaft. Berlin, Germany: Julius Beltz; 1997:97-106.

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Research
EHR: electronic health record
HCP: health care professional
ICD-10: International Classification of Diseases, Tenth Revision
PEPPPSY: Piloting and Evaluation of Participatory Patient-Accessible Electronic Health Record in Psychiatry and Somatics
POC: psychiatric outpatient clinic
SM: secure messaging

XSL•FO
RenderX

Edited by A Mavragani; submitted 27.06.24; peer-reviewed by M Rigby, F Wurster; comments to author 20.12.24; revised version received 09.01.25; accepted 12.03.25; published 27.06.25.

Please cite as:
Meier-Diedrich E, Turvey C, Wördemann JM, Speck J, Weibezahl M, Schwarz J
Patient–Health Care Professional Communication via a Secure Web-Based Portal in Severe Mental Health Conditions: Qualitative Analysis of Secure Messages
JMIR Form Res 2025;9:e63713
URL: https://formative.jmir.org/2025/1/e63713
doi:10.2196/63713
PMID:

©Eva Meier-Diedrich, Carolyn Turvey, Jonas Maximilian Wördemann, Justin Speck, Mareike Weibezahl, Julian Schwarz. Originally published in JMIR Formative Research (https://formative.jmir.org), 27.06.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Remotely Delivered GLP-1RA–Supported Specialist Weight Management Program in Adults Living With Obesity: Retrospective Service Evaluation

Rebecca Richards[1], PhD, CPsychol; Michael Whitman[1], BSc; Gina Wren[2], MRES; Peta Campion[1], BBiomedSc, MHSc

[1]Second Nature, London, United Kingdom

[2]Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, United Kingdom

Corresponding Author:
Rebecca Richards, PhD, CPsychol
Second Nature
483 Green Lanes
London, N13 4BS
United Kingdom
Phone: 44 020 3488 076
Email: michael.whitman@secondnature.io

Abstract

Background: Limited access to specialist weight management services restricts the implementation of novel pharmacotherapies for obesity such as glucagon-like peptide-1 receptor agonists (GLP-1RAs) in the UK National Health Service (NHS). Second Nature, a commercial digital health company, offers a remotely delivered program combining a GLP-1RA medication (semaglutide) with digital behavioral support, potentially providing a scalable solution. However, evidence for long-term effectiveness in this real-world context is limited.

Objective: This study aimed to evaluate the 12-month effectiveness, feasibility, acceptability, and potential cost-effectiveness of the remotely delivered, semaglutide-supported weight management program by Second Nature.

Methods: This retrospective service evaluation analyzed data from participants who initiated the program between September and December 2023. The primary outcome was weight change at 12 months among participants with available data (completers). Secondary outcomes included retention, program engagement (measured by views of the Home screen in the app), behavioral changes, side effects, participant experience (qualitative analysis), and a comparative cost analysis against an NHS specialist weight management service. An "active subscription" was defined as maintaining a paid subscription for the full 12-month period. Descriptive statistics and paired 2-tailed *t* tests evaluated outcomes.

Results: Data from 341 participants were included at baseline (282/341, 82.7% women; mean age 49, SD 11.1 years; mean baseline BMI 37.9, SD 6.9 kg/m^2). At 12 months, 39.6% (135/206) maintained an active subscription, while 60.4% (206/341) became inactive. Weight data at 12 months were available for 179 participants (52.5% of the baseline cohort; 100% of active and 19.4% of inactive participants). Among completers who maintained an active subscription, the mean weight loss was 20.0 kg (SD 8.7 kg; $P<.001$), representing 19.1% of starting weight. Overall, 77.7% (139/179) of completers achieved ≥10% weight loss and 61.5% (110/179) achieved ≥15%. Program engagement declined over time. Side effects also decreased, with 69.6% (81/116) of respondents reporting none by month 12. Most participants completing the 12-month survey reported positive (41/120, 34.2%) or neutral (68/120, 56.7%) experiences.

Conclusions: This evaluation suggests that remotely delivered GLP-1RA–supported weight management can achieve significant weight loss in participants remaining engaged for 12 months. However, the high rate of withdrawal limits generalizability. The program appears feasible, acceptable, and potentially cost-effective for completers. Further research, ideally in public health care settings using intent-to-treat analyses, is needed to confirm clinical outcomes, assess sustained results, and understand factors influencing retention.

(JMIR Form Res 2025;9:e72577) doi:10.2196/72577

XSL•FO
RenderX

KEYWORDS

obesity; weight management; glucagon-like peptide-1 receptor agonist; GLP-1RA; semaglutide; digital health; remote health care delivery; telemedicine; behavioral intervention; weight loss; cost-effectiveness; service evaluation; real-world evidence

Introduction

Background

Obesity remains a global health crisis, affecting 890 million adults worldwide and contributing to diseases such as type 2 diabetes, cardiovascular disease, and certain cancers [1]. In the United Kingdom, projections suggest that by 2050, the annual cost of obesity to the National Health Service (NHS) could reach £9.7 billion (US $12.1 billion), with wider societal expenses approaching £49.9 billion (US $62.4 billion) [2]. Although lifestyle interventions form the foundation of obesity treatment, their long-term effectiveness varies considerably, with more than half of lost weight typically regained within 2 years of weight loss [3].

Clinical trials of glucagon-like peptide-1 receptor agonists (GLP-1RAs), including semaglutide and tirzepatide, have demonstrated significant weight reductions, ranging from 14.9% to 20.9% over 68 to 72 weeks, respectively [4-6]. However, significant weight regain occurs following treatment discontinuation, with STEP-1 trial participants regaining approximately two-thirds of lost weight within 1 year of withdrawing from semaglutide (2.4 mg) and lifestyle intervention [7]. The National Institute for Health and Care Excellence (NICE) guidance recommending tirzepatide for obesity management alongside lifestyle modification [8] has no maximum treatment duration, suggesting potential long-term use. This necessitates the development of sustainable, cost-effective approaches to implement prolonged GLP-1RA treatment and lifestyle modifications.

To date, eligible people living with obesity in the United Kingdom have primarily accessed GLP-1RAs and lifestyle support through NHS specialist weight management services (SWMSs). SWMSs typically offer multidisciplinary team (MDT) support including dietitians, physicians, psychologists, and physiotherapists, often through a tiered system requiring general practitioner (GP) referrals [9]. However, because of underfunding, these services face significant challenges including long waiting lists, understaffing, and limited treatment flexibility, hindering timely access and potentially impacting outcomes [8,9]. In response to these issues and the anticipated demand for newer medications such as tirzepatide, NICE guidance suggests prescribing, monitoring, and multidisciplinary behavioral support could potentially be implemented in primary care [8].

However, concerns exist regarding the capacity of primary care services, particularly GPs, to absorb this additional workload. Despite manufacturer surveys indicating GP interest [8], recent data highlight increasing patient caseloads (17% more patients per GP since 2015) and appointment demands (5.46 million more appointments in February 2024 vs 2019), alongside a 3% decrease in full-time equivalent GPs [10,11]. Significant service transformation appears necessary to meet this demand within the NHS.

The NICE Health Technology Evaluation 14 (HTE14) guidance specifically highlights the potential role of remote interventions in increasing access to GLP-1RA medication and associated support [9]. Emerging evidence suggests that remotely delivered SWMSs could offer a scalable approach to address service provision gaps [9,10,12]. For example, a preliminary 12-week service evaluation of a remotely delivered program combining GLP-1RA (semaglutide) with app-based behavioral support (the same program evaluated in this study) showed that participants lost an average of 6.4% of body weight [12]. This earlier evaluation also suggested feasibility, with relatively low withdrawal rates over 12 weeks, manageable side effects, and generally positive or neutral participant experiences [12]. Furthermore, early economic modeling within the NICE HTE14 process indicated potential cost-effectiveness for remote SWMSs compared with traditional in-person services [9]. However, longer-term (≥12 months) real-world evidence for remotely delivered GLP-1RA–supported weight management interventions is crucial to substantiate their potential role in expanding access to NICE-recommended obesity treatments.

Objective

This retrospective service evaluation of a remotely delivered, semaglutide-supported weight management program by Second Nature was conducted to evaluate its effectiveness, feasibility, acceptability, and potential cost-effectiveness over a 12-month period in a real-world commercial setting.

Methods

Overview

This retrospective analysis used routinely collected service data from participants of a remotely delivered, semaglutide-supported commercial weight management program provided by Second Nature Healthy Habits Ltd, a UK-based digital health company. The study focused on participants who initiated the program between September and December 2023. As a service evaluation of existing data, the analysis primarily reflects outcomes among those who remained in the program and provided data at 12 months (ie, a completer analysis), particularly for the primary effectiveness outcome.

The core intervention, delivery methods, and data collection procedures were consistent with our previously reported 12-week evaluation [12]. Program adaptations were implemented following NICE guidance for tirzepatide, which, unlike semaglutide at the time, had no recommended treatment duration limit [8]. Consequently, the intervention was modified from a fixed duration to an open-ended format supporting individualized treatment periods. Program phases were redesignated from “transition” and “maintain” to “progress” and “momentum” to reflect observed participant behavior patterns, as significant weight loss often continued during these later phases. While this cohort used semaglutide exclusively and was followed for 12 months, these structural modifications

XSL•FO
RenderX

enabled extended support for ongoing GLP-1RA–assisted weight management where clinically appropriate. No participants in this cohort received tirzepatide.

Ethical Considerations

This evaluation study met Health Research Authority and Medical Research Council criteria for service evaluation [13]: it assessed current care using existing, anonymized data from an established intervention, did not involve any change to standard practice, and did not involve randomization or a control group. The Health Research Authority explicitly states that "because both audit and service evaluation are considered part of usual professional practice, they are exempt from oversight processes that govern research—that is, they do not require review by a research ethics committee" [14]. The study analyzed anonymized, non-NHS data that could not identify individuals, and thus did not require formal ethics approval (Multimedia Appendix 1). Participants were not compensated for taking part in the study. Participants consented to the research use of anonymized data through acceptance of the program's privacy policy at registration. The evaluation adhered to General Data Protection Regulations, maintaining participants' right to request data deletion.

Participants

Retrospective data were extracted from the Second Nature database in December 2024 for all participants (N=341) who initiated the intervention between September and December 2023. This timeframe allowed for a full 12-month potential follow-up period. All participants meeting the eligibility criteria during this period were included in the baseline analysis.

Eligible participants were adults aged 18 to 75 years living with obesity (BMI ≥30 kg/m^2) who had access to and could use a smartphone or tablet. Exclusion criteria included current or a history of eating disorders; pregnancy, breastfeeding, or actively trying to conceive; specific allergies to medication ingredients; concurrent use of certain medications (eg, other GLP-1RAs and insulin); and specific health conditions (eg, current or historical thyroid cancer, active cancer, inflammatory bowel disease, pancreatitis, severe liver or kidney disease, heart failure, multiple endocrine neoplasia type 2, gallbladder issues, or diabetic retinopathy). People living with type 2 diabetes were eligible if they were not taking any of the excluded medications (eg, metformin was allowed).

Participants were allowed to transition to a version of the digital program that focused only on behavioral support without pharmacotherapy; they were transferred to a separate intervention and were not included in this analysis.

Participants self-referred to the program, joining through digital marketing channels such as Google or Facebook, through word-of-mouth referrals from friends or family, or through guides written by Second Nature that were discovered through search engine results. The cost to the participants varied between £179 (US $224) and £299 (US $374) per month, depending on the dose of semaglutide prescribed.

Intervention

Overview

The intervention combined GLP-1RA medication (semaglutide) prescribed by independent prescribing pharmacists from Pharmalogic Chemist (a UK-based pharmacy provider) with Second Nature's digital weight management platform. Following clinical assessment and approval, participants received semaglutide, injection safety equipment, wireless weighing scales (Renpho Elis 1), and physical program materials (a FAQ leaflet, nutrition guidelines, and a recipe book) delivered to their address.

Support was provided by registered dietitians or nutritionists ("health coaches") with the Health and Care Professions Council or Association for Nutrition, respectively. Nutritionist-led coaching was supervised by a dietitian. Health coaches completed training covering medication management (including side-effect support), program delivery protocols, clinical governance, and safeguarding. The training was developed by senior MDT members and approved by the lead prescriber, with ongoing quality assurance via coaching reviews. The MDT included registered dietitians, nutritionists, clinical pharmacists, and a safeguarding lead.

App-Based Program

The digital program, delivered via Second Nature's smartphone and web applications, consisted of 5 phases: prepare (0-12 d), adapt (17 wk), grow (17 wk), progress (17 wk), and momentum (≥17 wk). This 17-week phase was designed to provide sufficient time for adapting and building new habits. The program used evidence-based behavior change techniques throughout all phases to help participants adapt to medication, adopt healthier long-term lifestyle habits (nutrition, physical activity, sleep, and stress management), and support sustainable weight management.

Educational content (nutrition, medication or side-effect management, behavioral psychology, physical activity, sleep hygiene, mental well-being, and stress management) was delivered through short daily articles. The app integrated self-monitoring tools for weight (via connected scales), activity, and sleep. The recommended frequency for self-monitoring weight was individualized for each participant based on their health coach's assessment of their progress, adherence pattern, and personal preferences. However, general guidelines were given to record weight at a consistent time each week, typically in the morning after waking and before eating or drinking. Physical activity was tracked through daily steps using the participant's smartphone or fitness tracker (no additional device was provided by the program), and participants were encouraged to self-monitor changes in sleep quality, duration, and energy levels. The program was developed by an MDT following NICE guidance for obesity management and behavior change principles [9], building upon Second Nature's established digital weight management program [12,15,16].

Participants received feedback on their progress through multiple channels. The app displayed weight trends visually through graphs showing progress toward individualized weight loss goals. Weekly automated messages summarized their weight

XSL•FO
RenderX

change to date and the percentage of their starting weight lost. Health coaches provided personalized feedback on progress during one-to-one messaging conversations, discussing weight trajectory, dietary choices, physical activity achievements, and addressing any challenges or plateaus.

Goal setting was an integral component of the program. At initiation, participants established an individualized target weight loss goal in consultation with their health coach, typically aiming for 15% to 20% of starting weight over 12 months based on GLP-1RA clinical trial outcomes. Weekly step goals were automatically suggested based on the participant's recent activity levels, gradually increasing as their activity improved. Sleep goals were more personalized, with participants encouraged to establish consistent sleep routines, aiming for 7 to 9 hours per night.

Participants received personalized support via asynchronous one-to-one messaging with their health coach during business hours, alongside moderated peer support groups within the app. Health coaches posted structured content weekly and responded to messages within specified timeframes. Interactions with the health coaches were ad hoc and based on participant needs, questions, and the educational content provided (covering topics such as the GLP-1RA medication [semaglutide], side effects, nutrition, physical activity, sleep, stress management, and mental well-being). If a participant became at risk of disengaging (defined as 4 d since their last Home screen view), health coaches were prompted to message them proactively. Program engagement was monitored through automated tracking of app interactions (see the Effectiveness Measures section). A notification system prompted the completion of mandatory monthly check-ins and weight measurements.

Medication

Semaglutide was prescribed with a 4-weekly dose titration schedule until the optimal therapeutic effect was achieved with minimal side effects. Medication management followed standardized safety protocols, with pharmacist oversight of initiation and repeat prescribing based on monthly digital monitoring surveys and consultations when clinically indicated. By month 12, all participants with active subscriptions (100%) had reached the maximum 2.4 mg dose.

Participants were guided on how to take the medication through educational resources, such as manufacturer's guides and injection demonstration videos [17]. The Second Nature web and smartphone application served as a central hub for ongoing support, featuring a dedicated medication "toolbox" and relevant daily content.

Safety monitoring comprised mandatory monthly digital surveys reviewed by pharmacists for prescription decisions. The platform enabled real-time adverse event reporting through health coach messaging, with 24-hour response times (business days). Medical concerns were escalated to prescribing teams via secure messaging, with pharmacist responses provided within 4 hours (business days). The system monitored safety indicators including BMI and rapid weight changes. Weight was tracked via Bluetooth-connected scales, and participants were encouraged to record their weight weekly for safety and to measure progress toward goals.

A tiered adverse event management protocol was implemented. Mild symptoms were initially managed through health coach–led lifestyle modifications, with a pharmacist reviewing if symptoms persisted beyond 7 days. Moderate or severe symptoms prompted immediate pharmacist assessment to determine intervention pathways. Emergency protocols directed participants to urgent care services for acute medical events. All clinical interactions were documented systematically across coaching and clinical platforms.

Data Collection and Measures

Overview

Data were analyzed from participants who initiated the intervention between September and December 2023. Participants were categorized at 12 months based on their subscription status. Active participants maintained a paid program subscription for the full 12 months. Inactive participants cancelled their self-funded subscription before the 12-month mark. Cancellation terminated medication provision and health coach support but maintained limited app access (educational resources and weight tracking).

Baseline characteristics (height, weight, age, and gender) were collected via web-based onboarding surveys and verified by customer support staff. Identity verification used LexisNexis third-party authentication, requiring a minimum of 2 database matches across Electoral Roll, Companies House, Experian, and Equifax records. Authentication failure criteria included age <18 years, deceased status, or birth data inconsistencies. Data collection comprised Bluetooth-transmitted weight measurements, monthly surveys capturing information on clinical suitability behavioral changes and participant experience, and digital platform engagement metrics. All data were stored in Second Nature's secure database.

Effectiveness Measures

The primary outcome was mean weight change (kg and % of starting weight) only at 12 months, calculated using the last recorded weight between months 11 and 13 (ie, day 330-420) compared with baseline weight. This analysis was performed on the subset of participants with available 12-month weight data (completers). The weight reading recorded for inactive participants was their last available weight reading. Secondary effectiveness outcomes included the proportion of completers achieving clinically significant weight loss thresholds: ≥10% and ≥15% of baseline weight. Weight measurements were recorded using the provided wireless scales. Participants were instructed on proper use (firm, flat surface, morning, postvoiding, and consistent day). A validation algorithm flagged potentially erroneous readings based on expected ranges, prompting user confirmation or contact with support staff.

Cost Measures

The cost-effectiveness of the intervention was evaluated through a comparative cost analysis between the commercial medication-supported intervention and NHS SWMSs. Cost data for NHS SWMSs were derived from NICE HTE14 guidance,

XSL•FO
RenderX

economic modeling, and resource impact assessments [9,18], combining expert consultation with 2022 Personal Social Services Research Unit costs. Analysis of 12 SWMS experts (11 providing detailed data) showed that standard care includes multidisciplinary support with weekly to fortnightly dietetic appointments for 2 to 3 months, followed by monthly follow-ups. Consultant reviews occur every 3 to 6 months, while psychological support comprises 6-week group sessions or bi- or triweekly individual therapy. Physiotherapy is provided at 1- to 3-month intervals. Attendance rates range from 60% to 90% (median 70%-75%), with higher rates (approximately 90%) reported for consultant and psychology appointments [9].

Additional costs were incorporated from NICE TA6179 guidance on tirzepatide for managing overweight and obesity [8], which provided broader cost ranges for implementing wraparound support alongside medication in community or primary care settings.

Using the NICE Impact Assessment Tool (IAT) [18], we analyzed costs per 100,000 population. The eligible proportion for semaglutide obesity treatment and SWMSs was set to the default value suggested in the calculator (9.50064%). On the basis of SWMS expert feedback from the NICE HTE14 supporting documentation [9], we estimated that 30% of the eligible population would prefer a remote option, with 75% ultimately starting the service based on data from Second Nature's NHS-delivered SWMS. We used base-case costs from NICE HTE14 guidance in the "Standard care (without app)" tab of the IAT and confidential costs for Second Nature's intervention [9,18].

For commercial sensitivity, percentage reductions were rounded to the nearest 10% using a relative cost-reporting approach. Costs were standardized to per-patient figures for 12 months of treatment. Percentage differences were calculated between the intervention and both primary and secondary care standard models detailed in NICE HTE14 guidance and the tirzepatide technology appraisal [8,9,18]. The analysis focused on direct intervention costs only, excluding long-term health improvement savings.

Behavioral Changes

Lifestyle-related behavioral changes, including diet, fruit and vegetable intake, and physical activity levels, were assessed through a series of mandatory web-based surveys administered at months 1, 3, 6, and 12 of the program. Participants were asked to report their daily vegetable and fruit consumption by responding to the question, "How many different vegetables or fruits do you eat a day on average?" Responses were recorded numerically to reflect the average number of different items consumed daily. To evaluate physical activity levels, participants were asked, "How many times do you exercise each week on average?" This question required participants to report the average number of exercise sessions completed per week. In addition, participants were asked about their meal preparation habits with the question, "How many times do you cook each week on average?" Responses were recorded as the number of cooking sessions per week. Finally, to assess dietary modifications, participants responded to the question, "What changes have you made to your diet over the last few weeks?" The number of dietary changes was reported numerically and analyzed at each survey time point.

Acceptability

Secondary outcomes included the acceptability of the intervention, which was assessed using the following measures.

Retention and Cancellation Reasons

Withdrawal was defined as participants cancelling their subscription to the program. The number of people who maintained an active subscription or cancelled their subscription was recorded, and reasons for cancelations were collected through an in-app cancellation form.

Program Engagement

Engagement with the app program was assessed using Home screen views, which measured the number of times the app was opened. The duration of program participation was calculated as the number of months between the participant's intervention start date and the effective cancelation date (defined as the final day in a month when the participant had full access to all intervention features).

Feasibility

Side Effects

Participants were emailed a mandatory web-based check-in survey at months 1, 3, 6, and 12 to collect information on medication side effects. Participants could select from a drop-down menu of 15 possible side effects, "none," or "other." Selecting "other" prompted a mandatory free-text response for the description. If a side effect was selected, participants were asked an additional question: "Are these side effects tolerable?" with response options of "yes" or "no." If "no" was selected, participants were asked, "Do you want advice or support with these side effects from our clinical team?" with the same response options.

Medication Administration

Responses to a question from the mandatory web-based check-in survey administered at months 1, 3, 6, and 12 were extracted to assess the acceptability of medication administration. The question, "Are you having any difficulty injecting yourself?" included 3 preset responses: "No"; "Yes, and it means I'm not injecting once a week on the same day"; and "Yes, but I'm still injecting once a week on the same day."

Participant Experience

Participants' experiences with the program were assessed by analyzing free-text responses to the question, "How have you found the last few weeks of the program?"

Statistical Analysis

Analyses were performed using R statistical software (version 4.3.3). For effectiveness, we calculated mean weight loss and SD. We evaluated statistical significance using paired 2-tailed *t* tests ($P<.05$) and also calculated the proportions of participants achieving ≥10% and ≥15% weight loss thresholds. Statistical analysis included descriptive statistics for baseline characteristics and primary weight change outcomes. We also descriptively analyzed weight change by subscription status

XSL•FO
RenderX

("active" vs "inactive" paid subscriptions). Program retention was analyzed by calculating active subscription rates and categorizing withdrawal reasons. Monthly behavioral changes, side effects, and participant experiences were analyzed as proportions of total survey respondents. Program engagement was assessed through mean monthly Home screen views. Cost comparisons used the NICE IAT to model costs per 100,000 population, assuming 9.50064% eligibility, 30% digital service uptake, and 75% initiation rate. We standardized annual per-patient costs and calculated percentage differences between the intervention and NHS standard care models [18].

Qualitative Analysis

Content analysis was conducted to analyze the free-text responses to the monthly survey question, "How have you found the last few weeks of the program?" For each question, at each time point, the responses were coded and organized into higher-level categories by MW. These frameworks were then refined based on the protocol outlined in a previous study to create 9 higher-level categories that captured key themes in participants' experiences [12]. Each response could be categorized into multiple themes to reflect the complexity of participants' experiences. For example, when a participant submitted a free-text response of "ok," "fine," "not bad," "nothing to report," or "same as before" to the "How have you found the last few weeks on the program?" question, it would be categorized as a "neutral experience."

Results

Baseline Characteristics and Retention

A total of 341 participants initiated the program between September and December 2023 and were included in the baseline analysis (Figure 1). Baseline characteristics are presented in Table 1. The majority were women (282/341, 82.7%), with a mean age of 48.2 (SD 11.7) years and a mean baseline BMI of 37.7 (SD 6.9) kg/m^2. There were no statistically significant differences in baseline characteristics between those who remained active and those who became inactive at 12 months (data not shown; $P>.05$ for all comparisons).

Figure 1. Participant flow through the 12-month semaglutide-supported weight management program, showing retention, data availability, and withdrawal reasons.

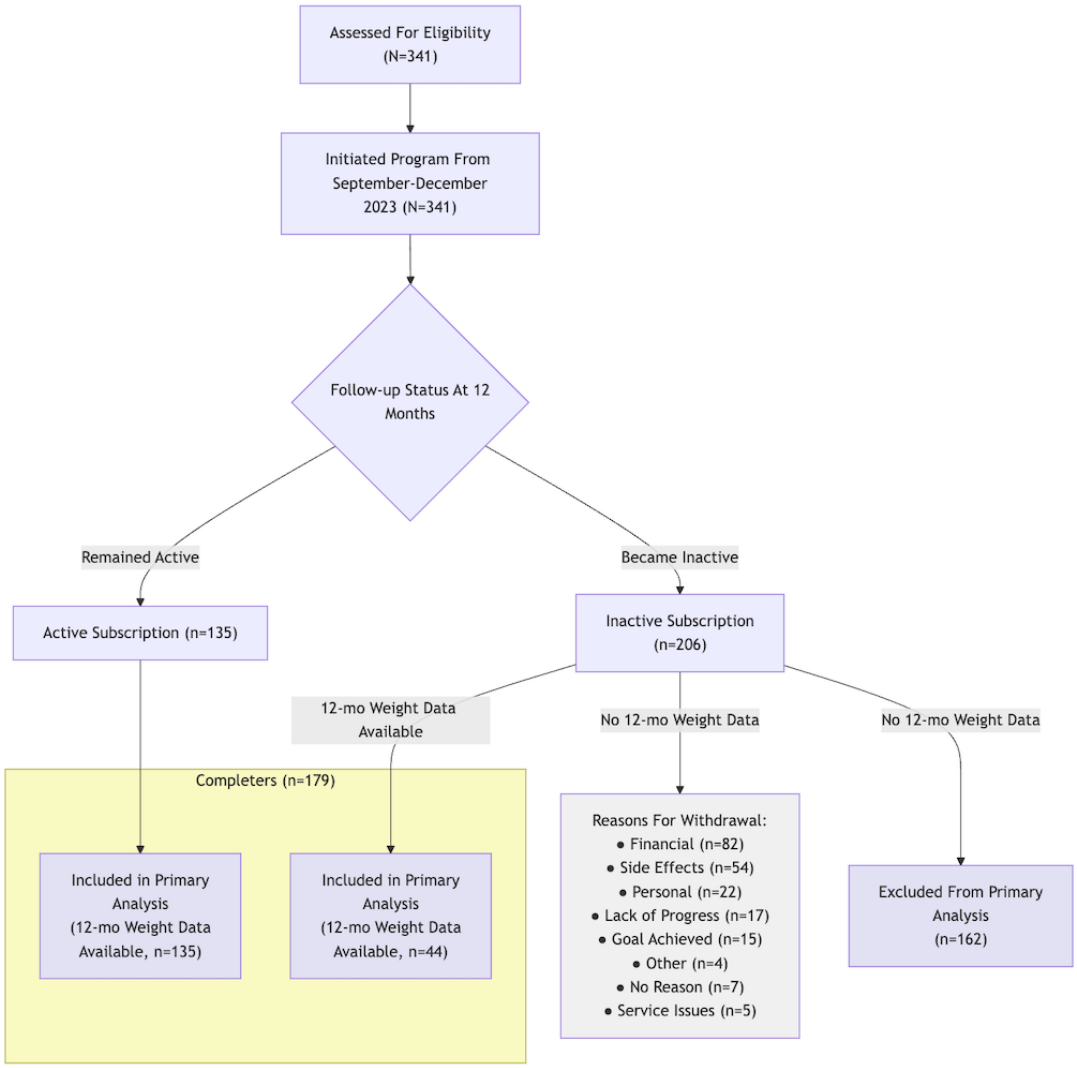

XSL•FO
RenderX

Table 1. Baseline characteristics of all participants: those with active subscriptions at 12 months and those with inactive subscriptions at 12 months.

Characteristics	All participants (N=341)	Active participants at 12 mo (n=135)	Inactive participants at 12 mo (n=206)
Age (y), mean (SD)	48.2 (11.7)	49 (11.1)	47.7 (12.1)
Female, n (%)	282 (82.7)	111 (82.2)	171 (83)
Baseline weight (kg), mean (SD)	105.7 (21.3)	106.6 (20.5)	105.1 (21.8)
Baseline BMI (kg/m²), mean (SD)	37.7 (6.9)	37.9 (6.9)	37.6 (6.9)

At the 12-month time point, of the 341 participants, 135 (39.6%) maintained an active paid subscription. The remaining 206 (60.4%) cancelled their subscriptions and became inactive. The mean duration of program participation for the entire cohort was 8.2 (SD 4.2) months. For active participants, the mean duration was 12.5 (SD 0.6) months (reflecting ongoing participation at data extraction), while for inactive participants, the mean duration before cancellation was 5.4 (SD 3.0) months. Weight data at 12 months were available for 179 (N=341, 52.5%) participants of the baseline cohort. This included all 135 (100%) active participants and 44 (n=206, 21.4%) inactive participants who continued to record weight data after cancellation.

Effectiveness

Weight Loss

Among the 179 participants with 12-month weight data (completers), the mean weight loss from baseline was 18.1 kg (SD 9.5 kg; $P<.001$), corresponding to 17.3% (SD 8.9%) of starting weight. Weight loss differed by subscription status among completers. Participants with an active subscription at 12 months (135/179, 75%) achieved a mean weight loss of 20.0 kg (SD 8.7 kg; $P<.001$), representing 19.1% (SD 8.0%) of their starting weight. Participants who were inactive but provided 12-month weight data (44/179, 24%) achieved a mean weight loss of 12.4 kg (SD 10.6 kg; $P<.001$), representing 11.9% (SD 9.8%) of their starting weight. Overall, among the 179 completers, 139 (77.7%) achieved ≥10% weight loss and 110 (61.5%) achieved ≥15% weight loss at 12 months. Among 135 active completers, 116 (85.9%) achieved ≥10% weight loss and 94 (69.6%) achieved ≥15% weight loss. Among 44 inactive completers, 23 (52.3%) achieved ≥10% weight loss and 16 (36.4%) achieved ≥15% weight loss.

Cost Compared With Usual Care

The external assessment group for the NICE HTE14 guidance estimated the cost of delivering a secondary care SWMS to be £1796 (US $2245) per patient per year [18]. Depending on the specific implementation requirements for an NHS health care setting, we estimated that the cost of Second Nature's service would be between 30% and 40% of the current standard care for NHS SWMSs. In addition, costs provided in the supporting documentation for the NICE guidance on tirzepatide for managing overweight and obesity estimated the cost of an MDT-supported weight management intervention delivered in the community or primary care settings at £868.21 to £1239.21 (US $1085.13 to $1549.01) per patient per year [8]. The intervention costs were estimated to be 30% to 90% of the costs for these hypothetical models of delivering an in-person or hybrid MDT in a community or primary care setting.

Behavioral Changes

When asked about fruit and vegetable consumption, most participants (151/280, 54%) at month 1 and (62/119, 52.1%) at month 12 reported consuming 3 to 4 different items per day across all time points (Tables 2 and 3). The proportion consuming ≥5 items remained relatively stable from 63 (n=280, 22.4%) at month 1 to 28 (n=119, 23.5%) at month 12.

XSL•FO
RenderX

Table 2. Responses to a lifestyle-related behavioral survey by month (% of total respondents; months 1-6).

Category and subcategory	Month					
	1 (n=367)	2 (n=308)	3 (n=272)	4 (n=250)	5 (n=220)	6 (n=191)
Fruit and vegetable consumption (items per d), n (%)						
0	4 (1.1)	4 (1.3)	3 (1.1)	3 (1.2)	1 (0.5)	2 (1)
1-2	81 (22.4)	71 (23)	71 (25.7)	60 (24.3)	53 (24.1)	45 (23.4)
3-4	195 (54)	166 (53.7)	141 (50.7)	132 (53.4)	115 (52.3)	102 (53.1)
≥5	81 (22.4)	68 (22)	62 (22.5)	52 (21.1)	52 (23.2)	43 (22.4)
Exercise (times/wk), n (%)						
0	78 (21.6)	48 (15.5)	28 (9.7)	20 (7.9)	14 (6.5)	13 (6.6)
1-2	152 (42.1)	139 (44.8)	120 (43.2)	115 (45.3)	101 (46.5)	78 (39.8)
3-4	91 (25.2)	91 (29.4)	96 (34.5)	86 (33.9)	77 (35)	73 (37.2)
≥5	40 (11.1)	32 (10.3)	35 (12.6)	33 (13)	26 (12)	32 (16.3)
Cooking frequency, n (%)						
Never	5 (1.4)	3 (1)	4 (1.5)	2 (0.8)	2 (0.9)	1 (0.5)
Less than once per wk	7 (1.9)	3 (1)	9 (3.3)	6 (2.4)	5 (2.3)	10 (5.2)
Once per wk	10 (2.7)	9 (2.9)	8 (3)	12 (4.8)	6 (2.7)	11 (5.8)
Few times per wk	90 (24.5)	91 (29.5)	69 (25.5)	75 (30)	71 (32.4)	57 (29.8)
Nearly every day	133 (36.2)	117 (38)	109 (40.2)	96 (38.4)	85 (38.4)	80 (41.9)
Every day	122 (33.2)	85 (27.6)	73 (26.6)	59 (23.6)	51 (23.3)	32 (16.8)

Table 3. Responses to a lifestyle-related behavioral survey by month (% of total respondents; months 7-12).

Category and subcategory	Month					
	7 (n=181)	8 (n=127)	9 (n=138)	10 (n=133)	11 (n=136)	12 (n=119)
Fruit and vegetable consumption (items/d), n (%)						
0	2 (1.1)	2 (1.5)	1 (0.7)	1 (0.8)	2 (1.4)	1 (0.8)
1-2	37 (20.2)	30 (22.9)	24 (17.5)	27 (20.3)	24 (17.4)	28 (23.5)
3-4	95 (51.9)	63 (48.1)	73 (53.3)	74 (55.6)	83 (60.1)	62 (52.1)
≥5	49 (26.8)	36 (27.5)	39 (28.5)	31 (23.3)	29 (21)	29 (23.5)
Exercise (times/wk), n (%)						
0	9 (4.9)	10 (7.6)	7 (5.1)	9 (6.7)	11 (8)	9 (7.6)
1-2	73 (39.7)	50 (38.2)	49 (35.8)	55 (41)	58 (42)	47 (39.5)
3-4	75 (40.8)	51 (38.9)	57 (41.6)	47 (35.1)	50 (36.2)	48 (39.5)
≥5	27 (14.7)	20 (15.3)	24 (17.5)	23 (17.2)	19 (13.8)	16 (13.4)
Cooking frequency, n (%)						
Never	1 (0.6)	2 (1.6)	0 (0)	1 (0.8)	1 (0.7)	1 (0.8)
Less than once per wk	5 (2.8)	2 (1.6)	4 (2.9)	4 (3)	6 (4.4)	3 (2.5)
Once per wk	6 (3.3)	5 (3.9)	8 (5.8)	8 (6)	8 (5.9)	7 (5.9)
Few times per wk	60 (33.1)	46 (36.2)	41 (29.7)	39 (29.3)	37 (27.2)	34 (28.8)
Nearly every day	73 (40.3)	38 (29.9)	58 (42)	56 (42.1)	59 (43.4)	52 (43.2)
Every day	36 (19.9)	34 (26.8)	27 (19.6)	25 (18.8)	25 (18.4)	22 (18.6)

The proportion exercising 3 to 4 times per week or every day increased, whereas the proportion exercising 1 to 2 times per week stayed relatively consistent. In addition, the number of participants who reported no exercise per week decreased from 60 (n=280, 21.5%) participants in month 1 to 9 (n=119, 7.6%) in month 12. An increase in cooking frequency was observed

XSL•FO
RenderX

over 12 months, except for those who reported cooking "every day" at baseline, which decreased from 92 (n=280, 33%) participants to 22 (n=119, 18.6%) participants, and those who "never" cooked, which decreased from 4 (n=280, 1.4%) participants to 1 (n=119, 0.8%) participant.

Acceptability

Retention and Cancellation Reasons

Of the 341 participants included in the analysis, 135 (39.6%) maintained an active paid subscription after 12 months. The remaining 206 (60.4%) cancelled their subscriptions before 12 months. However, 44 (21.4%) of these participants continued to record weight data. Cancellation reasons for those who withdrew were categorized as follows: financial concerns (70/206, 34%), medical issues and side effects (49/206, 23.8%), no reason given (29/206, 14.1%), personal circumstances (19/206, 9.2%), perceived lack of weight loss progress (16/206, 7.8%), perceived achievement of desired weight goal (15/206, 7.3%), support and service issues (5/206, 2.4%), and other (eg, where ambiguous information was provided in response to determine an accurate reason for cancellation, such as "I can not be on the injections any longer"; 3/206, 1.5%). The mean program participation was 8.2 (SD 4.2) months overall, with active participants averaging 12.5 (SD 0.6) months at the point of data collection and inactive participants averaging 5.4 (SD 3.0) months. All participants who maintained an active subscription submitted weight readings. Of the 206 participants who withdrew before the 12-month mark, 44 (21.4%) participants submitted their weight readings.

Program Engagement

The mean Home screen view count started at 113 (SD 78.4) in the first month, reduced to 73 (SD 65) in the second month, and then steadily declined each month to 25 (SD 29.06) by month 12 (Figure 2).

Figure 2. The mean number of application Home screen views by month on the program among the active participants.

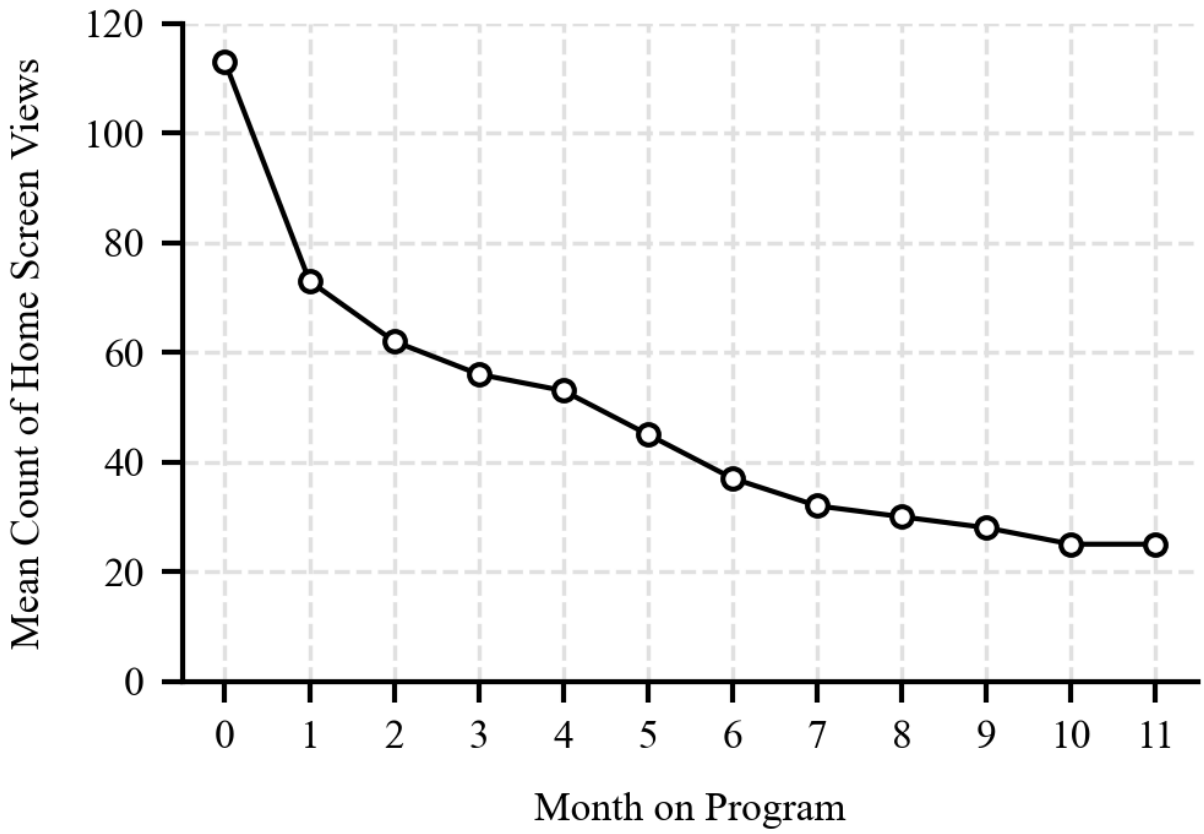

Feasibility

Side Effects

In the first month, the most reported side effects were feeling sick (90/313, 28.8%), constipation (80/313, 25.6%), and feeling more tired than usual (57/313, 18.2%); 154 (n=313, 49.2%) participants reported no side effects at this stage. After 12 months, 80 (n=116, 69%) participants reported no side effects, and the most common side effects were constipation (18/116, 15.5%), feeling sick (8/116, 7%), and hair loss (7/116, 6%; Table 4).

XSL•FO
RenderX

Table 4. Percentage of participants reporting side effects by survey month (% of total respondents).

Side effect	Month, n (%)											
	1 (n=313)	2 (n=297)	3 (n=264)	4 (n=239)	5 (n=216)	6 (n=186)	7 (n=179)	8 (n=130)	9 (n=136)	10 (n=132)	11 (n=135)	12 (n=116)
Constipation	80 (25.6)	91 (30.6)	83 (31.4)	65 (27.2)	53 (24.5)	35 (18.8)	35 (19.6)	21 (16.2)	25 (18.4)	27 (20.5)	21 (15.6)	18 (15.5)
Diarrhea	27 (8.6)	31 (10.4)	29 (11)	34 (14.2)	28 (13)	17 (9.1)	17 (9.5)	8 (6.2)	12 (8.8)	12 (9.1)	8 (5.9)	7 (6)
Dry skin and mouth	1 (0.3)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Fast, deep breathing	0 (0)	0 (0)	1 (0.4)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Feeling anxious	3 (1)	10 (3.4)	4 (1.5)	7 (2.9)	1 (0.5)	4 (2.2)	2 (1.1)	2 (1.5)	1 (0.7)	2 (1.5)	1 (0.7)	2 (1.7)
Feeling dizzy	13 (4.2)	12 (4)	12 (4.5)	9 (3.8)	8 (3.7)	4 (2.2)	4 (2.2)	2 (1.5)	5 (3.7)	3 (2.3)	2 (1.5)	6 (5.2)
Feeling more tired than usual	57 (18.2)	64 (21.5)	64 (24.2)	56 (23.4)	41 (19)	30 (16.1)	18 (10.1)	16 (12.3)	10 (7.4)	9 (6.8)	8 (5.9)	6 (5.2)
Feeling sick	90 (28.8)	99 (33.3)	94 (35.6)	67 (28)	64 (29.6)	38 (20.4)	24 (13.4)	17 (13.1)	13 (9.6)	17 (12.9)	13 (9.6)	8 (6.9)
Fever or high temperature	0 (0)	1 (0.3)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Flushed face	0 (0)	0 (0)	1 (0.4)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Hair loss	7 (2.2)	3 (1)	5 (1.9)	3 (1.3)	3 (1.4)	3 (1.6)	9 (5)	10 (7.7)	8 (5.9)	10 (7.6)	10 (7.4)	7 (6)
Headaches	32 (10.2)	25 (8.4)	25 (9.5)	17 (7.1)	20 (9.3)	8 (4.3)	5 (2.8)	6 (4.6)	2 (1.5)	5 (3.8)	2 (1.5)	4 (3.4)
Heartburn or indigestion	42 (13.4)	59 (19.9)	56 (21.2)	42 (17.6)	32 (14.8)	18 (9.7)	14 (7.8)	7 (5.4)	10 (7.4)	8 (6.1)	8 (5.9)	5 (4.3)
Muscle stiffness or aches	1 (0.3)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (0.6)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
No side effects	154 (49.2)	100 (33.7)	80 (30.3)	91 (38.1)	88 (40.7)	99 (53.2)	98 (54.7)	76 (58.5)	89 (65.4)	82 (62.1)	87 (64.4)	80 (69)
None of the above	4 (1.3)	2 (0.7)	3 (1.1)	3 (1.3)	2 (0.9)	3 (1.6)	2 (1.1)	1 (0.8)	1 (0.7)	1 (0.8)	1 (0.7)	0 (0)
Reactions in the injection site	3 (1)	2 (0.7)	4 (1.5)	3 (1.3)	4 (1.9)	0 (0)	2 (1.1)	1 (0.8)	4 (2.9)	5 (3.8)	7 (5.2)	1 (0.9)
Severe stomach pain	2 (0.6)	5 (1.7)	2 (0.8)	5 (2.1)	5 (2.3)	1 (0.5)	0 (0)	1 (0.8)	1 (0.7)	1 (0.8)	3 (2.2)	1 (0.9)
Tender or swollen stomach	6 (1.9)	7 (2.4)	11 (4.2)	4 (1.7)	2 (0.9)	2 (1.1)	1 (0.6)	1 (0.8)	1 (0.7)	1 (0.8)	1 (0.7)	0 (0)
Vomiting	8 (2.6)	14 (4.7)	20 (7.6)	20 (8.4)	11 (5.1)	9 (4.8)	11 (6.1)	3 (2.3)	6 (4.4)	6 (4.5)	3 (2.2)	2 (1.7)
Yellowing of the skin or eyes	0 (0)	0 (0)	0 (0)	0 (0)	1 (0.5)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)

Medication Administration

Most participants reported no difficulties with medication administration. At month 1, of the 315 participants, 310 (98.4%) reported no difficulties, and only 5 (1.5%) indicated they needed support with injecting. At months 2 and 3, a total of 297 (n=297, 100%) participants reported no difficulties. Small fluctuations were observed at month 4 by 1 (n=242, 0.4%) participant and month 5 by 1 (n=216, 0.5%) participant where participants reported injection difficulties. At month 7, only 1 (n=182, 0.5%) participant reported difficulties. From month 8 to month 12, all participants consistently reported no difficulties with injection administration. Throughout the program, all participants who reported difficulties still maintained their prescribed weekly injection schedule.

Participant Experience

The participant responses were categorized into 9 distinct areas: change in appetite, illness, intervention benefits, negative experience, neutral experience, other priorities, positive experience, and slow progress. Initially, most participants reported positive experiences (211/320, 65.9% at month 1), but this steadily declined to just over a third (41/120, 34.2%) by month 12. Conversely, neutral experiences increased from 15.9% (51/320) at month 1 to 56.7% (68/120) at month 12 (Table 5). Reported intervention benefits were reduced throughout the program, decreasing from 17.2% (55/320) at month 1 to 3.3% (4/120) at month 12. Similarly, reports of appetite changes declined from 8.4% (27/320) to <1%. Negative

XSL•FO
RenderX

experiences and reduced effectiveness were consistently reported by 1% (1/120) of participants.

Table 5. Percentage of participants reporting different experiences throughout by program month (% of total respondents).

Category	Month, n (%)											
	1 (n=320)	2 (n=301)	3 (n=274)	4 (n=242)	5 (n=218)	6 (n=192)	7 (n=182)	8 (n=130)	9 (n=137)	10 (n=133)	11 (n=137)	12 (n=120)
Change in appetite	27 (8.4)	32 (10.6)	25 (9.1)	13 (5.4)	9 (4.1)	6 (3.1)	6 (3.3)	2 (1.5)	3 (2.2)	4 (3)	1 (0.7)	1 (0.8)
Illness	6 (1.9)	4 (1.3)	4 (1.5)	4 (1.7)	4 (1.8)	3 (1.6)	6 (3.3)	1 (0.8)	2 (1.5)	0 (0)	0 (0)	0 (0)
Intervention benefits	55 (17.2)	32 (10.6)	41 (15)	27 (11.2)	22 (10.1)	20 (10.4)	15 (8.2)	10 (7.7)	9 (6.6)	12 (9)	7 (5.1)	4 (3.3)
Negative experience	2 (0.6)	1 (0.3)	1 (0.4)	2 (0.8)	0 (0)	2 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Neutral experience	51 (15.9)	46 (15.3)	58 (21.2)	84 (34.7)	94 (43.1)	90 (46.9)	84 (46.2)	73 (56.2)	66 (48.2)	76 (57.1)	82 (59.9)	68 (56.7)
Other	13 (4.1)	6 (2)	5 (1.8)	3 (1.2)	3 (1.4)	3 (1.6)	4 (2.2)	3 (2.3)	4 (2.9)	2 (1.5)	0 (0)	2 (1.7)
Other priorities	5 (1.6)	8 (2.7)	5 (1.8)	7 (2.9)	8 (3.7)	1 (0.5)	5 (2.7)	2 (1.5)	1 (0.7)	2 (1.5)	1 (0.7)	4 (3.3)
Positive experience	211 (65.9)	183 (60.8)	143 (52.2)	111 (45.9)	86 (39.4)	70 (36.5)	65 (35.7)	39 (30)	53 (38.7)	37 (27.8)	46 (33.6)	41 (34.2)
Reduced effectiveness or slow progress	1 (0.3)	0 (0)	0 (0)	0 (0)	2 (0.9)	0 (0)	0 (0)	1 (0.8)	0 (0)	0 (0)	1 (0.7)	0 (0)

Discussion

Principal Findings

This 12-month retrospective service evaluation provides real-world insights into a remotely delivered, semaglutide-supported commercial weight management program. The primary finding is that participants who completed 12 months of the program (179/341, 52.5%) of the initial cohort achieved substantial and clinically significant weight loss, with a mean reduction of 17.3% (18.1 kg). Those who maintained an active subscription throughout achieved even greater mean weight loss (19.1%, 20 kg). However, these findings must be interpreted cautiously because of the high withdrawal rate, with 47.5% (162/341) of participants providing no 12-month weight data, and 60.4% (206/341), cancelling their subscriptions, meaning the results primarily reflect outcomes in a self-selected group of completers.

The program demonstrated feasibility regarding medication administration and side-effect management among those who remained. Side effects were most common early in the program but decreased significantly over time, with the vast majority of participants reporting no side effects after 12 months. Self-injection was not a significant barrier. Acceptability, as judged by participant experience surveys among respondents, remained largely positive or neutral at 12 months; although positive sentiment decreased over time, neutral responses increased. Engagement with the digital platform, measured by Home screen views, showed a typical pattern of decline over the 12 months. Encouragingly, positive lifestyle changes related to exercise and cooking frequency appeared to be sustained or improved over the 12 months among survey respondents.

The intervention's estimated cost was substantially lower than traditional NHS SWMS models, suggesting the potential for cost savings if implemented within a public health care system. However, this requires confirmation through formal health economic analysis in that context.

Comparison With Prior Work

The mean weight reduction of 19.1% (SD 8.0%) for participants with an active subscription observed in this study are comparable with outcomes reported in both controlled trials and real-world evaluations of GLP-1RA medications, respectively [19-21]. For example, reviews of the semaglutide STEP randomized controlled trials showed mean weight reductions of 14% to 17% over 68 weeks [19], while tirzepatide trials demonstrated weight loss of 14.7% to 20.9% across populations with and without diabetes [20]. However, direct comparisons are limited by differences in study design (randomized controlled trials vs real-world evaluation), participant populations (trial eligibility vs commercial users), follow-up duration, the intensity of lifestyle support, and analytic approach (intent-to-treat [ITT] vs completer analysis).

Comparison with other real-world, remotely delivered, GLP-1RA–supported behavioral change programs with completer analyses are more relevant. Two real-world studies in the United States reported 12-month weight reductions with semaglutide when combined with remote support: one found a reduction of 13.8% to 15.3% and another reported a reduction of 13.9% [21,22]. This study demonstrated higher rates of substantial weight loss, with 77.7% (139/179) of participants achieving ≥10% and 61.5% (110/179) achieving ≥15% at 12 months, compared with 38.7% and 18.4%, respectively, in comparable remotely delivered program in the United States [22].

An important finding relates to the inactive participants who provided 12-month weight data (44/206, 21% of those who cancelled). These individuals, who stopped the paid program (including medication and coaching) after an average of 5.4

XSL•FO
RenderX

months, still maintained a mean weight loss of 11.9% at the 12-month mark. This observation requires cautious interpretation because of the small sample size and potential for selection bias. However, it aligns with recent evidence suggesting less weight regain after stopping semaglutide when combined with behavioral support compared with medication withdrawal alone [7,23]. The STEP 1 trial extension showed a regain of about two-thirds (approximately 11.6% absolute weight) of the weight 1 year after stopping semaglutide and intensive behavioral therapy [7], whereas another study reported only 3% regain 12 months after semaglutide cessation in participants who continued a nutritional program [23]. Our finding, while preliminary and based on limited data, suggests the potential role of the initial behavioral intervention component in mitigating weight regain after program cessation.

Program retention rates in remotely delivered weight management services vary substantially. The 12-month retention rate of 39.6% (135/341) observed in this study falls between previously reported rates ranging from 22.7% at 5 months [24] to 58% at 9 months [25] for other GLP-1RA–supported remotely delivered interventions, although structural and financial differences limit direct comparisons. The observed decline in app engagement (Home screen views) is typical for digital health interventions [22,26] and does not necessarily equate to a lack of benefit but highlights the need for strategies to maintain long-term engagement, if required. The improvement in side-effect tolerability over time is consistent with clinical trial data [19] and is reassuring for long-term use. However, the fact that 26% (49/206) of participants who withdrew cited side effects suggests that early intolerance remains a significant barrier for a subset of users, contributing to the lower prevalence of side effects observed in completers (a form of survivorship bias).

It is possible that the behavioral support provided in this study increased and supported the maintenance of exercise, dietary quality, and home cooking to support and enhance the benefits of the GLP-1RA medication. For example, exercise and maintaining dietary quality have been recommended during the use of GLP-1RAs to retain muscle mass during weight loss [27-31]. However, comparison with existing literature is constrained because of the lack of comprehensive behavioral data in GLP-1RA research. Typically, studies of GLP-1RA medication have focused on caloric intake and satiety measures rather than broader lifestyle modifications [32]. Therefore, further research on the impact of behavioral support on GLP-1RA outcomes is needed.

Cost Comparison and Implications

On the basis of NICE modeling data [9,18], this remotely delivered intervention appears to be substantially less costly than traditional secondary care SWMSs. Calculations using the NICE IAT suggest potential direct cost reductions of 40% to 50% per 100,000 population (£1.5-1.8 million; US $1.9-2.25 million) compared with the HTE14 standard care estimate [18]. While these are preliminary estimates based on modeling assumptions and confidential commercial costs, they support the hypothesis that remote delivery could offer a more resource-efficient way to provide the necessary wraparound support for GLP-1RA therapy, potentially freeing up NHS resources or enabling wider access. A formal health economic evaluation within an NHS setting is required to confirm these potential savings.

Strengths and Limitations

The strengths of this study include its relatively large initial sample size, 12-month follow-up duration, and real-world setting, providing insights beyond those from controlled trials. The use of connected scales offers objective weight data, and the collection of data on retention, engagement, side effects, behaviors, and experience provides a multifaceted evaluation.

However, the findings of this study have several limitations. The observational design without a comparator group inherently limits the ability to definitively attribute observed weight loss solely to the intervention, as secular trends or unmeasured participant factors could play a role. This design issue is compounded by the substantial participant withdrawal rate (206/341, 60.4%) by 12 months. Consequently, the primary effectiveness analysis relies on data from completers (179/341, 52.5% of the initial cohort), a group likely to differ systematically from those who dropped out (eg, in motivation, medication tolerability, or initial success). This reliance on a completer analysis significantly limits the generalizability of the effectiveness findings. An ITT analysis with appropriate imputation for missing data, while beyond the scope of this service evaluation, would be necessary to estimate effectiveness across all individuals who initiated the program.

Generalizability is also influenced by the nature of the participant cohort. As it is a commercial program requiring self-payment, participants primarily enrolled via digital marketing, search engines, or word-of-mouth referrals. This self-selection process likely yielded a group with sufficient financial resources and motivation, potentially differing from the broader NHS patient population. The high proportion citing cost as a reason for withdrawal underscores this potential selection bias and suggests that retention patterns and possibly outcomes might differ substantially in a publicly funded model. Relatedly, while initial health conditions were reviewed during onboarding (partly via NHS Summary Care Record access), detailed baseline data on comorbidities (including type 2 diabetes status), ethnicity, or concurrent medications were not accessible for this specific analysis, restricting a full understanding of the cohort's characteristics and factors influencing outcomes.

Methodological limitations also extend to data collection and measurement. Behavioral changes were assessed using self-reported data from nonvalidated questionnaires, introducing potential recall and social desirability bias. Furthermore, the absence of baseline behavioral data precludes a direct assessment of change from preintervention habits. Program engagement was measured by Home screen views, which is an unvalidated metric of engagement. More granular data on the use of specific features such as coaching interactions or self-monitoring tool entries were not included in this analysis but may offer avenues for future investigation. Although connected scales provided objective weight data, the real-world

XSL•FO
RenderX

setting meant weighing frequency and conditions were not strictly standardized.

The evaluation also lacked data on broader clinical outcomes (eg, metabolic parameters) or patient-reported outcomes, such as quality of life. In addition, participants could transition to a separate nonmedication program, and these individuals were excluded; thus, the analysis does not capture outcomes for those who may have stopped medication but continued behavioral support under a different structure. Moreover, while the last available weight was used for inactive participants with 12-month data, the specific weight trajectory immediately preceding withdrawal for the majority who became inactive earlier was not analyzed.

Finally, although the 12-month duration exceeds that of many digital health evaluations, it may be insufficient to assess the long-term sustainability of weight loss and behavioral changes, particularly concerning potential medication cessation effects. Future research should aim to address these limitations, for instance, through designs incorporating matched controls or randomization, standardized measurements using validated instruments, comprehensive baseline and outcome data collection, ITT analyses, and extended follow-up periods.

Conclusions

This real-world service evaluation suggests that a remotely delivered, GLP-1RA–supported weight management program is an effective, feasible, acceptable, and cost-effective approach for achieving weight loss and improving health behaviors in people living with obesity. These programs could address the current service provision issues with specialist in-person services, but further research is needed to evaluate their use within a public health care setting, assess their impact on other clinical and psychological measures, and evaluate sustained effects with continuation and discontinuation of the medication.

Acknowledgments

The authors thank the participants and the clinical and administrative teams at Second Nature Healthy Habits Ltd for delivering the service.

Conflicts of Interest

RR, PC, and MW are employees of Second Nature Healthy Habits Ltd. Second Nature is the industrial partner on GW's Medical Research Council Industrial Collaborative Awards in Science and Engineering studentship.

Multimedia Appendix 1
UK Health Research Authority ethics exemption: result—not research.
[PDF File (Adobe PDF File), 187 KB - formative_v9i1e72577_app1.pdf]

References

1. Obesity and overweight. World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight [accessed 2024-12-16]
2. Health matters: obesity and the food environment. Public Health England. URL: https://www.gov.uk/government/publications/health-matters-obesity-and-the-food-environment/health-matters-obesity-and-the-food-environment--2 [accessed 2024-12-16]
3. Hall KD, Kahan S. Maintenance of lost weight and long-term management of obesity. Med Clin North Am 2018 Jan;102(1):183-197 [FREE Full text] [doi: 10.1016/j.mcna.2017.08.012] [Medline: 29156185]
4. Wilding JP, Batterham RL, Calanna S, Davies M, Van Gaal LF, Lingvay I, et al. Once-weekly semaglutide in adults with overweight or obesity. N Engl J Med 2021 Mar 18;384(11):989-1002. [doi: 10.1056/nejmoa2032183]
5. Jastreboff AM, Aronne LJ, Ahmad NN, Wharton S, Connery L, Alves B, et al. Tirzepatide once weekly for the treatment of obesity. N Engl J Med 2022 Jul 21;387(3):205-216. [doi: 10.1056/nejmoa2206038]
6. Wadden TA, Bailey TS, Billings LK, Davies M, Frias JP, Koroleva A, STEP 3 Investigators. Effect of subcutaneous semaglutide vs placebo as an adjunct to intensive behavioral therapy on body weight in adults with overweight or obesity: the STEP 3 randomized clinical trial. JAMA 2021 Apr 13;325(14):1403-1413 [FREE Full text] [doi: 10.1001/jama.2021.1831] [Medline: 33625476]
7. Wilding JP, Batterham RL, Davies M, Van Gaal LF, Kandler K, Konakli K, STEP 1 Study Group. Weight regain and cardiometabolic effects after withdrawal of semaglutide: the STEP 1 trial extension. Diabetes Obes Metab 2022 Aug 19;24(8):1553-1564 [FREE Full text] [doi: 10.1111/dom.14725] [Medline: 35441470]
8. Tirzepatide for managing overweight and obesity. National Institute for Health and Care Excellence. URL: https://www.nice.org.uk/guidance/ta1026 [accessed 2025-01-09]
9. Digital technologies for delivering specialist weight-management services to manage weight-management medicinearly value assessment. National Institute for Health and Care Excellence. URL: https://www.nice.org.uk/guidance/indevelopment/gid-hte10007 [accessed 2025-01-08]
10. Pressures in general practice data analysis. British Medical Association. URL: https://www.bma.org.uk/advice-and-support/nhs-delivery-and-workforce/pressures/pressures-in-general-practice-data-analysis [accessed 2025-01-08]

XSL•FO
RenderX

11. GP appointments breaking point. Royal College of General Practitioners. URL: https://www.rcgp.org.uk/news/gp-appointments-breaking-point [accessed 2025-01-08]
12. Richards R, Wren GM, Campion P, Whitman M. A remotely delivered, semaglutide-supported specialist weight management program: preliminary findings from a retrospective service evaluation. JMIR Form Res 2023 Dec 28;7:e53619 [FREE Full text] [doi: 10.2196/53619] [Medline: 38153780]
13. Is my study research? Health Research Authority. URL: http://www.hra-decisiontools.org.uk/research/ [accessed 2025-01-17]
14. Scott AM, Kolstoe S, Ploem MC, Hammatt Z, Glasziou P. Exempting low-risk health and medical research from ethics reviews: comparing Australia, the United Kingdom, the United States and the Netherlands. Health Res Policy Syst 2020 Jan 28;18(1):11 [FREE Full text] [doi: 10.1186/s12961-019-0520-4] [Medline: 31992320]
15. Idris I, Hampton J, Moncrieff F, Whitman M. Effectiveness of a digital lifestyle change program in obese and type 2 diabetes populations: service evaluation of real-world data. JMIR Diabetes 2020 Jan 20;5(1):e15189 [FREE Full text] [doi: 10.2196/15189] [Medline: 31958064]
16. Kar P, Goward C, Whitman M, Davies M, Willner T, Shaw K. Engagement and effectiveness of digitally enabled behavioural change support for people living with type 2 diabetes. Pract Diabetes 2020 Oct 07;37(5):167. [doi: 10.1002/pdi.2295]
17. Patient resources for Wegovy. Novo Nordisk. URL: https://www.wegovy.co.uk/patientresources.html [accessed 2025-01-17]
18. National Institute for HealthCare Excellence. Resource impact template: digital technologies for delivering multidisciplinary weight-management services. Health technology evaluation HTE14. URL: https://www.nice.org.uk/guidance/hte14/resources/resource-impact-template-excel-13197081133 [accessed 2025-01-13]
19. Bergmann NC, Davies MJ, Lingvay I, Knop FK. Semaglutide for the treatment of overweight and obesity: a review. Diabetes Obes Metab 2023 Jan 18;25(1):18-35 [FREE Full text] [doi: 10.1111/dom.14863] [Medline: 36254579]
20. Qin W, Yang J, Ni Y, Deng C, Ruan Q, Ruan J, et al. Efficacy and safety of once-weekly tirzepatide for weight management compared to placebo: an updated systematic review and meta-analysis including the latest SURMOUNT-2 trial. Endocrine 2024 Oct 08;86(1):70-84. [doi: 10.1007/s12020-024-03896-z] [Medline: 38850440]
21. Ard JD, Hong YR, Foster GD, Medcalf A, Nadolsky S, Cardel MI. Twelve-month analysis of real-world evidence from a telehealth obesity-treatment provider using antiobesity medications. Obesity (Silver Spring) 2024 Dec 30;32(12):2246-2254. [doi: 10.1002/oby.24169] [Medline: 39478294]
22. Clark JM, Smith BJ, Juusola JL, Kumar RB. Long-term weight loss outcomes in a virtual weight care clinic prescribing a broad range of medications alongside behavior change. Obes Sci Pract 2025 Feb 08;11(1):e70036. [doi: 10.1002/osp4.70036] [Medline: 39790443]
23. McKenzie AL, Athinarayanan SJ. Impact of glucagon-like peptide 1 agonist deprescription in type 2 diabetes in a real-world setting: a propensity score matched cohort study. Diabetes Ther 2024 Apr 29;15(4):843-853 [FREE Full text] [doi: 10.1007/s13300-024-01547-0] [Medline: 38421559]
24. Talay LA, Vickers M, Lagesen L, Liu N. The effect of lifestyle coaching design on patient engagement and weight loss in non-diabetic patients of a semaglutide-supported digital obesity program in the UK: a comparative retrospective cohort study. Cureus 2024 Nov;16(11):e74321. [doi: 10.7759/cureus.74321] [Medline: 39583610]
25. Talay L, Vickers M. Patient adherence to a digital real-world GLP-1 RA-supported weight-loss program in the UK: a retrospective cohort study. J Community Med Public Health 2024 Mar 27;8:422. [doi: 10.29011/2577-2228.100422]
26. Talay L, Vickers M, Bell C, Galvin T. Weight loss and engagement in a Tirzepatide-supported digital obesity program: a four-arm patient-blinded retrospective cohort study. Telemed Rep 2024 Apr 01;5(1):311-321. [doi: 10.1089/tmr.2024.0058]
27. Locatelli JC, Costa JG, Haynes A, Naylor LH, Fegan PG, Yeap BB, et al. Incretin-based weight loss pharmacotherapy: can resistance exercise optimize changes in body composition? Diabetes Care 2024 Oct 01;47(10):1718-1730. [doi: 10.2337/dci23-0100] [Medline: 38687506]
28. Jensen SB, Blond MB, Sandsdal RS, Olsen LM, Juhl CR, Lundgren JR, et al. Healthy weight loss maintenance with exercise, GLP-1 receptor agonist, or both combined followed by one year without treatment: a post-treatment analysis of a randomised placebo-controlled trial. EClinicalMedicine 2024 Mar;69:102475 [FREE Full text] [doi: 10.1016/j.eclinm.2024.102475] [Medline: 38544798]
29. Jakicic JM, Rogers RJ, Church TS. Physical activity in the new era of antiobesity medications. Obesity (Silver Spring) 2024 Feb 17;32(2):234-236. [doi: 10.1002/oby.23930] [Medline: 37849057]
30. Rogers RJ. Exercise for clients taking an anti-obesity medication. ACSMs Health Fit J 2024;28(3):21-26. [doi: 10.1249/fit.0000000000000959] [Medline: 39279918]
31. Stewart J. Help patients prevent weight gain after stopping GLP-1s. Medscape. URL: https://www.medscape.com/viewarticle/help-patients-prevent-weight-gain-after-stopping-glp-1s-2024a10004z9 [accessed 2025-01-13]
32. Christensen S, Robinson K, Thomas S, Williams DR. Dietary intake by patients taking GLP-1 and dual GIP/GLP-1 receptor agonists: a narrative review and discussion of research needs. Obes Pillars 2024 Sep;11:100121 [FREE Full text] [doi: 10.1016/j.obpill.2024.100121] [Medline: 39175746]

Abbreviations

GLP-1RA: glucagon-like peptide-1 receptor agonist

XSL•FO
RenderX

GP: general practitioner
HTE14: Health Technology Evaluation 14
IAT: Impact Assessment Tool
ITT: intent-to-treat
MDT: multidisciplinary team
NICE: National Institute for Health and Care Excellence
SWMS: specialist weight management service

Edited by A Mavragani; submitted 14.02.25; peer-reviewed by J Bittman, M Patel; comments to author 01.04.25; revised version received 09.04.25; accepted 28.05.25; published 30.06.25.

Please cite as:
Richards R, Whitman M, Wren G, Campion P
A Remotely Delivered GLP-1RA–Supported Specialist Weight Management Program in Adults Living With Obesity: Retrospective Service Evaluation
JMIR Form Res 2025;9:e72577
URL: https://formative.jmir.org/2025/1/e72577
doi:10.2196/72577
PMID:40434299

©Rebecca Richards, Michael Whitman, Gina Wren, Peta Campion. Originally published in JMIR Formative Research (https://formative.jmir.org), 30.06.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Challenges and Mitigation Strategies in the Development and Feasibility Assessment of a Digital Mental Health Intervention for Depression (VMood): Mixed Methods Feasibility Study

Leena W Chau[1], MSc; Jill K Murphy[2], PhD; Vu Cong Nguyen[3], MD; Hai Nhu Tran[3], MD; Harry Minas[4], MD; Raymond W Lam[5], MD; Kanna Hayashi[1], PhD; Thi Thanh Xuan Nguyen[3], BPH; Emanuel Krebs[6], MA; John O'Neil[1], PhD

[1]Simon Fraser University, Burnaby, BC, Canada
[2]St. Francis Xavier University, Antigonish, NS, Canada
[3]Institute of Population, Health and Development, Hanoi, Vietnam
[4]University of Melbourne, Melbourne, Australia
[5]University of British Columbia, Vancouver, BC, Canada
[6]BC Cancer Agency, Vancouver, BC, Canada

Corresponding Author:
Leena W Chau, MSc
Simon Fraser University
8888 University Dr W
Burnaby, BC, V5A 1S6
Canada
Phone: 1 778 782 5870
Email: leena_chau@sfu.ca

Abstract

Background: Worldwide, the COVID-19 pandemic contributed to further gaps in mental health care, particularly in low- and middle-income countries such as Vietnam, where care is inaccessible for 90% of those who need it. There has subsequently been a considerable increase in the use of digital mental health interventions such as smartphone apps. Presently, the evidence for such interventions is limited, especially in cases in which the interventions have been adapted from evidence-based in-person formats. Implementation science aims to promote the incorporation of scientific findings into practice. A key determinant of implementation success is an intervention's usability. Hurdles to usability include an intervention being too confusing or time-intensive to use. Facilitators include incorporating a greater number of engagement features and integrating human support.

Objective: The aim of this implementation science feasibility study was to describe the challenges and mitigation strategies used in the development, usability testing, and implementation of a digital depression intervention (VMood smartphone app) developed in Vietnam. VMood was adapted from an evidence-based in-person intervention originally developed in Canada that is grounded in principles of cognitive behavioral therapy with supportive coaching by a lay health or social services worker. The research team is currently testing the effectiveness and cost-effectiveness of VMood in a randomized controlled trial across 8 provinces in Vietnam informed by the results of this feasibility assessment.

Methods: This mixed methods feasibility study was organized using an implementation outcome framework focused on acceptability, adoption, appropriateness, and feasibility. This study involved three data collection components: (1) usability testing (interviews and focus groups with app user and provider participants who tested VMood in 1 Vietnamese province), (2) app metrics (from the early phase of the randomized controlled trial in the same province but from different municipalities), and (3) discourse data (notes from various team meetings, communications, and reports on VMood's development and implementation). Qualitative data were analyzed using thematic content analysis. App use data were analyzed using basic descriptive statistics.

Results: The findings of the 3 data components showed that there were seven main challenges: (1) challenges with recruitment and uptake of the app, (2) challenges with use and engagement, (3) screening challenges, (4) digital divide, (5) limitations to digital applications for mental health, (6) technological challenges, and (7) funding and policy constraints. Various solutions to help mitigate the challenges were used by the team.

XSL•FO
RenderX

Conclusions: The findings contribute important evidence on the challenges to the development and feasibility assessment of a digital depression app adapted from an in-person intervention in Vietnam. The findings have applicability for others looking to develop and implement digital interventions in similar contexts, serving as a unique opportunity to share the lessons learned regarding the development and testing process.

(JMIR Form Res 2025;9:e68297) doi:10.2196/68297

KEYWORDS

digital mental health; depression; implementation science; usability; low- and middle-income countries; LMICs; Vietnam

Introduction

Background

Mental disorders are a leading cause of global disease burden [1], with approximately 1 in every 8 people living with a mental disorder in 2019 [2]. Improving access to mental health care as a global priority was further magnified in the context of the COVID-19 pandemic, which has contributed to deteriorations in population mental health [3]. The Global Burden of Disease 2020 study estimated that the pandemic has led to increases of 28% in major depressive disorders and 26% in anxiety disorders [3], 2 of the most common mental disorders [4]. The COVID-19 pandemic has also led to severe disruptions to already limited in-person services [5]. This has had a profound impact worldwide, including in low- and middle-income countries such as Vietnam [6-8]. Recent evidence shows the prevalence of mental disorders in Vietnam to be of 10.1 million (rate of 10.2%), with the prevalence (and rates) of depression and anxiety being 2.7 million (3%) and 2.9 million (2.7%) [9], respectively, based on the population of 98.9 million in 2021 [10]. This is especially concerning in Vietnam as mental health care is inaccessible for up to 90% of those who need it [11]. In response, there has been a considerable increase in mental health interventions delivered through digital means [12] such as smartphone apps [13] as they have emerged as a promising way to expand access to care within the additional constraints imposed by the COVID-19 pandemic. Presently, the evidence for such interventions is limited [14,15].

Implementation science aims to promote the incorporation of scientific findings and other evidence-based approaches into practice to improve health services [16]. A key determinant of implementation success is an intervention's usability [17]. Usability is the extent to which a product can be used easily, effectively, and with satisfaction to obtain specified goals [18]. Usability has been commonly referred to as a key factor in the successful implementation and sustainability of digital health apps [19], with poor usability being a primary reason for failed adoption [20]. Approximately one-third of users who download health apps stop using them after a short time [21]. An additional consideration for individuals with depression is specific symptoms such as low motivation, concentration difficulties, and behavioral avoidance that may make it more difficult for them to engage with digital health interventions [22], posing further challenges to real-world effectiveness. Facilitators of usability include incorporating a greater number of engagement features (eg, praises and reminders) [23] and integrating human support [24].

Prior Work

VMood has been adapted from an in-person supported self-management (SSM) intervention originally developed in Canada [25]. This intervention is grounded in principles of cognitive behavioral therapy (CBT) and uses an SSM approach consisting of an *Antidepressant Skills Workbook* (ASW) and supportive coaching delivered in person by a lay health or social services worker (eg, community worker or peer navigator). SSM has been shown to be effective in the Vietnamese context through a previous randomized controlled trial (RCT; funded by Grand Challenges Canada; 2016-2019) conducted by this research team across 8 Vietnamese provinces [26]. In response to the Government of Vietnam's budgetary changes and the restrictions on in-person care imposed by the COVID-19 pandemic, the team accepted the request from the Vietnamese Ministry of Labour, Invalids, and Social Affairs (MOLISA) to adapt the intervention to a digital format rather than scale up the in-person SSM intervention as originally planned. This work is funded by the Canadian Institutes of Health Research and Grand Challenges Canada (with matching funding provided by MOLISA).

Ongoing Work

VMood, delivered via a smartphone app, comprises 3 months of engagement with the VMood program, which includes the various modules (eg, *Introducing depression*, *Antidepressant skills*, and *Reactivating your life*) from the ASW being delivered in an interactive multiple-choice format. Each module includes an introductory video with a text option. The VMood program also includes support delivered through the app via messages from a social worker. Social workers, a relatively new profession in Vietnam that received official designation as a profession in 2010 [27], work under the purview of MOLISA [27]. The profession is expanding in Vietnam, reflected in the number of schools providing social worker training increasing from 2 in 2003 to 35 in 2024 [28]. Social workers were envisioned to play a role in the delivery of mental health care in community-based settings but have in general received minimal mental health training [29]. Participating social workers will be trained on basic depression, addressing stigma, and providing coaching via the app through a web-based training program developed by the Institute of Population, Health, and Development (PHAD). PHAD is a leading nongovernmental research institute in Vietnam and the research team's implementing partner.

In Vietnam, responsibility for mental health is shared between MOLISA and the Ministry of Health. MOLISA oversees 24 long-term care centers for individuals with severe mental health conditions attached to their provincial Social Protection Centers

XSL•FO
RenderX

(SPCs), whereas the Ministry of Health oversees 36 psychiatric hospitals and 25 psychiatry departments within provincial general hospitals [30]. Despite a policy shift since 1998 toward a community-based approach to mental health care with the focus on prevention and delivering low-barrier care to those with common mental disorders such as mild to moderate depression [29,31,32], progress has been slow. Vietnam's mental health system still comprises mainly large tertiary hospitals providing care only for those with severe mental health conditions using predominantly pharmacological treatment, with very little to no community-based care available for those with mild to moderate depression [33].

In this context, there is an added urgency to expand community-based mental health care for the large and growing population of those who need it. There is, at present, limited critical evidence on the unique considerations and practical guidance for the development and implementation of digital apps for depression adapted from evidence-based in-person formats. This paper will describe the challenges encountered in the development and feasibility assessment of VMood in real-world settings in 1 province of Vietnam and mitigation strategies that the team used to address them. At the time of writing, the larger RCT is being implemented across the 8 participating provinces in Vietnam informed by the lessons learned described in this paper. The findings have applicability for others looking to develop and implement digital interventions in similar contexts.

Goal of This Study

The aim of this implementation science feasibility study was to describe the challenges in the development and feasibility assessment of a digital depression intervention (VMood smartphone app) adapted from an in-person format in Vietnam, along with mitigation strategies regarding how the research team addressed them. Specific objectives were as follows:

1. To describe the challenges in the process of adapting an in-person depression intervention to a digital format (VMood app) and the challenges encountered in VMood's feasibility assessment
2. To describe the strategies that the team used to mitigate and address the challenges and the lessons learned

Methods

Overview

This feasibility assessment used mixed methods to examine the challenges influencing the development, usability testing, and implementation of the VMood digital depression intervention in Vietnam and describe the strategies used by the research team to mitigate these challenges. The team comprises researchers from diverse backgrounds (eg, global mental health, anthropology, psychiatry, statistics, and computer science) in Vietnam, Canada, and Australia. In total, 3 data components informed the findings (Table 1). The overall time frame for the feasibility study was from the beginning of app development (January 2021) to the end of the first phase of the RCT in 1 province in Vietnam (June 2024).

Table 1. Data collection components for the overall feasibility study conducted from June 2023 to March 2024.

Data type	Description	Sources
Usability testing	Early testing of the app to gather data on usability and acceptability. Conducted in Thanh Hoa province from June 2023 to September 2023	• Interviews and focus groups
App metrics	Data on rates of enrollment, log-ins, and frequency and duration of engagement with specific components of the app to supplement qualitative usability testing results. Gathered from Thanh Hoa from December 2023 to June 2024	• Use data from the VMood app • Due to technological challenges with VMood, app metrics data were not available from the usability testing and were instead gathered from the first phase of the RCT[a], also conducted in Thanh Hoa; the RCT involves different regions (ie, the participants are different from those in the usability testing)
Discourse data	Notes from various meetings and emails that describe the challenges in the development, usability testing, and implementation of VMood, along with reports from the early phase of RCT implementation in Thanh Hoa	• Meeting notes (monthly team, small-group, and internal meetings) • Personal communications (emails) • Field trip report from first phase of RCT implementation in Thanh Hoa (January 2024) • Progress reports from the Thanh Hoa RCT (January 2024-March 2024)

[a]RCT: randomized controlled trial.

Conceptual Framework

The feasibility study was informed by the implementation outcomes framework by Proctor et al [34], focusing on *acceptability*, *adoption*, *appropriateness*, and *feasibility* (Textbox 1). Additional outcomes—*costs*, *penetration*, and *sustainability*—are beyond the scope of this study, and the last outcome, *fidelity*, is captured in more detail elsewhere [35]. This framework evaluates implementation outcomes, which are distinct from clinical outcomes examining an intervention's effectiveness. However, they serve as intermediate outcomes to clinical outcomes because, if an intervention is not implemented well, it will likely not be clinically effective [34]. Implementation science frameworks are different from theories, sets of analytical statements about relationships between constructs, and models, a simplification of a specific aspect of a more complex world in the sense that frameworks provide a

XSL•FO
RenderX

series of steps summarizing how the implementation process should be planned, carried out, and evaluated [36]. The framework by Proctor et al [37] was chosen because it enabled us to selectively focus on outcomes that are important in the early stages of the implementation process.

Textbox 1. Conceptual framework—selected implementation outcomes from the implementation outcomes framework by Proctor et al. Definitions are taken directly from the work by Proctor et al.

Implementation outcome and definition

- Acceptability: "Perception among implementation stakeholders that a given treatment, service, practice, or innovation is agreeable, palatable, or satisfactory"
- Adoption: "Intention, initial decision, or action to try or employ an innovation or evidence-based practice"
- Appropriateness: "Perceived fit, relevance, or compatibility of the innovation or evidence-based practice for a given practice setting"
- Feasibility: "Extent to which a new treatment, or an innovation, can be successfully used or carried out within a given agency or setting"

Procedures

Usability Testing

The usability testing component took place across 6 communes (municipal subdivisions) in 1 district of the Thanh Hoa province in Vietnam. The names of the communes and district have been omitted for confidentiality purposes. The communes were selected in collaboration with MOLISA to represent diversity in terms of economic status (eg, urban and rural) and population composition (eg, older and younger). Usability testing involved participants testing the VMood app followed by the collection of qualitative data (interviews and focus groups). Project activities were conducted at the Thanh Hoa provincial SPC or in the community. Table 2 provides a summary of the participant categories.

Table 2. Overview of participants of the usability testing conducted between June 2023 and September 2023.

Category and description	Source (participant code)
Policy stakeholder (n=1)	
The director of the provincial SPC[a]. Responsible for providing oversight in project implementation in the province.	Interview (IPS-01)
Providers (n=15)	
2 social workers based in the SPC and responsible for project administration: 1 project coordinator responsible for all project activities and 1 provincial coordinator who served as the focal point for requests and recommendations from the province to the project team	Interviews (IPR-01 and 02)
13 community women's union staff members (ie, lay workers) based in the communes who are primarily volunteers active in their communities supporting social and health initiatives. Responsible for helping app users navigate the app.	Focus groups (FGPR-01 [n=8] and FGPR-02 [n=5])
App users (n=6)	
With depression caseness as defined by a PHQ-9[b] [38] score of ≥5	Interviews (IAU-01 to 06)
Clinical experts (n=5)	
Familiar with CBT[c] principles and depression—2 from Vietnam and 3 from Canada	Interviews (IEX-01 to 05)

[a]SPC: Social Protection Center.

[b]PHQ-9: 9-item Patient Health Questionnaire.

[c]CBT: cognitive behavioral therapy.

Our research team worked with MOLISA to have an individual appointed as the main national MOLISA coordinator who would facilitate recruitment for the overall study. The policy stakeholder, along with the 2 social workers based at the Thanh Hoa SPC, were recruited through official government channels by the MOLISA coordinator. The community women's union staff members (lay workers active in their communities; n=13) based in the communes were recruited by the Thanh Hoa project coordinator to support app deployment. PHAD team members administered the training program to the intervention providers in person. App users who self-identified as experiencing symptoms of depression were recruited from the community through the 6 local Commune Health Centers (CHCs). They were recruited through community engagement meetings held by the PHAD team and recruitment flyers posted at the CHCs. Staff at the CHCs, trained by PHAD staff on study objectives, helped facilitate recruitment. Recruitment began in May 2023 and took place over a 1-month period. A total of 642 app users were recruited and invited to download VMood and complete the registration and consent forms embedded within the app. Once app users provided consent, they were asked to self-administer the embedded 9-item Patient Health Questionnaire (PHQ-9), a 9-item depression screening measure, through the app [38]. The PHQ-9 is a widely used short

XSL•FO
RenderX

depression screening measure that was selected for this study because it has been validated for use in Vietnam [39-41] and its screening effectiveness compared to traditional (paper-based) formats has been demonstrated on mobile devices [42-45]. Only participants identified with depression caseness as determined by a PHQ-9 score of ≥5 (154/642, 24%) were invited to participate. A PHQ-9 score of ≥5 indicates depression, with scores of 5 to 9 indicating mild depression [38]. We included individuals with mild depression caseness as the VMood app is meant to help individuals with mild to moderate depression. Community women's union staff assisted app users with app download and registration. Clinical experts familiar with CBT principles were recruited from Thanh Hoa, Vietnam (2/5, 40%), and British Columbia, Canada (3/5, 60%). Experts in Vietnam were recruited through Thanh Hoa's SPC facilitated by the MOLISA coordinator, with the recruitment process coordinated by PHAD. Experts in Canada were recruited from a local community-based counseling and support services organization identified through similar research projects, along with an expert who was invited based on their familiarity with the in-person SSM intervention. All clinical experts from Canada were native Vietnamese speakers. The recruitment process for all participants in Vietnam was supported by the national MOLISA coordinator based in Hanoi, who has been involved with recruitment and project-related activities throughout the research program (2013-present).

Usability testing involved all participants downloading and testing the app. App users were notified that they had the option of connecting with a commune social worker through a messaging function built into the app for any questions regarding app navigation and basic inquiries about depression. Formative literature on sample sizes for usability testing studies indicates that a sample of 5 participants can identify 80% of a system's usability issues, additional participants will reveal fewer and less novel usability issues, and the most severe usability issues are likely to have been captured in the first few participants [46].

Exit interviews were conducted with the policy stakeholder and the 2 social workers, and focus groups were conducted with the 13 community women's union staff members to obtain their feedback on app usability. Focus groups are helpful for elucidating additional information on issues of interest due to interactions and group dynamics among multiple users [47] and can help generate a richer understanding of participants' experiences and views [47,48], in addition to being time-efficient. Interviews were also conducted with a convenience sample of the 154 app users with depression caseness involved in VMood testing (n=6, 3.9%) based on their availability to gather feedback on their experience using the app. Questions explored the implementation outcomes of acceptability, appropriateness, adoption, and feasibility. Finally, interviews were conducted with the clinical experts to obtain feedback on the conceptual components of the intervention adaptation, along with any strengths and suggestions for improvement.

All interviews and focus groups were conducted in person by the first author (LC), with an interpreter assisting in Vietnam. The interviews lasted between 30 and 60 minutes, and the focus group discussions lasted between 65 and 75 minutes. The interviews and focus groups were recorded, transcribed, and translated into English using forward and backward translation. Detailed field notes during and after the interviews and focus groups were recorded to capture thoughts and observations about the interactions that could help inform the data analysis process [47].

App Metrics

Due to technological challenges and the way in which the beta version of the app was designed (reported in the Results section), app use data could not be collected from the participants in the usability testing and, thus, are not reported in this paper. Instead, app use data from the first phase of the RCT, also conducted in the Thanh Hoa province but in a different district, are included. Recruitment in the first 2 communes from the first RCT district in Thanh Hoa began in December 2023 and was completed in March 2024. App use data from these 2 communes are reported in this paper and include data on duration of engagement with the VMood app from 20 app users, all of whom have completed the 3-month VMood program as of June 2024. Data collected using built-in analytics include metrics for evidence on the *adoption* implementation outcome.

Discourse Data

The feasibility study also involved an examination of discourse data (meeting minutes, emails, and progress reports) on the app development process that took place from January 2021 to June 2023, along with usability testing in Thanh Hoa from June 2023 to September 2023 and the implementation of VMood in the first phase of the RCT in Thanh Hoa from December 2023 to June 2024. The discussions examined the challenges encountered throughout the various phases of the project and also centered on brainstorming strategies to help ensure continued implementation and success. Certain small-group meetings focused on specific aspects of the app development (eg, technical challenges) or implementation (eg, screening concerns) processes as they arose among select members of the team with relevant expertise.

A main source of data for the app development discussion was the in-person team meetings in Hanoi, Vietnam (August 22-25, 2022). Most members of the geographically distributed research team (from Vietnam, Canada, and Australia) attended. The workshop focused on the app development process, including examining the associated challenges that the team encountered from initial development of the app to subsequent revisions and strategies to help address some of those challenges. All discourse data were either transcribed or captured through detailed meeting minutes and translated where applicable.

Monthly reports on the progress of the first phase of the RCT in Thanh Hoa (monthly reports, n=4) were prepared by PHAD, the implementing partner in Vietnam, detailing implementation updates as well as the challenges encountered and proposed solutions. Challenges included those that the team had experienced in addition to challenges discussed with the provincial implementing partners in Thanh Hoa, who shared their experiences in the community when promoting and implementing VMood. Detailed field notes were also

XSL•FO
RenderX

documented throughout the study period by LC to summarize and interpret the various discourse data.

Data Analysis

Qualitative interview, focus group, and discourse data were analyzed using thematic analysis as outlined by Braun and Clarke [49] to identify and analyze common themes across transcripts that captured important meanings and patterns in the data. A coding framework was developed deductively informed by the outcomes for implementation research framework [50] and also inductively through analysis of the qualitative data. In addition, analysis of focus group data took into account the group dynamics captured via the focus group discussions themselves and researcher field notes. Coding was a multistage process that involved open coding, where preliminary summary statements were provided for elements discussed in the transcript and formed an initial coding framework, and gathering all statements from open coding to collapse into main categories and subsequently into themes, along with validation of the coding by another research team member (senior author JON) in the form of reviewing independently a percentage of the transcripts [51].

App use data were analyzed using basic descriptive statistics.

Data triangulation and reflexivity were used to help ensure the trustworthiness of the data analysis and research findings [52]. Triangulation was used to explore congruence and divergence between data sources, thus increasing the rigor of the findings and reflexivity, aided by the detailed field notes and memos made during the qualitative coding process, to ensure the trustworthiness of the research findings [49]. These strategies provide an audit trail of the analytic process to ensure that other researchers can arrive at a comparable conclusion following the same processes [53]. The audit trail also includes detailed records of the study's methods and procedures. The NVivo software (version 14; QSR International) [54] was used for all qualitative analysis.

Ethical Considerations

All procedures were approved by the Research Ethics BC board in Canada (H21-02938) and the institutional review board at PHAD in Vietnam (PHAD-2022/VMOOD-01). Informed consent to take part was obtained from all participants. All VMood app processes for data collection, handling, and storage were encrypted. All interview data were deidentified and stored on a Simon Fraser University–approved cloud storage system. All app data were stored on PHAD's secure data center using industry-standard 256-bit encryption. The data from this study were coded only using the project name and participant identification number.

The policy stakeholder and providers were given an honorarium equivalent to CAD $10 (approximately VND 185,000 [US $7.36]) for participating in an interview or focus group. App users were provided with CAD $20 (VND 370,000 [US $14.72]) for their testing of the app, along with a phone credit of CAD $3 (VND 56,000 [US $2.21]) for data use on their mobile phones to access the app. Individuals who participated in the interviews were provided with an additional CAD $5 (VND 95,000 [US $3.68]). Finally, clinical experts in Vietnam received an honorarium of CAD $50 (VND 925,000 [US $36.80]) for their participation, including the interview. Clinical experts in Canada received CAD $150 (VND 2,775,000 [US $110.40]). All currency used a conversion rate of CAD $1=US $1.36. The difference in honoraria between experts in Vietnam and Canada was due to differences in the salaries of mental health experts and living costs between the 2 countries and because the ethics committees in Vietnam would not allow for incentives outside their range.

Results

Overview

The findings of the feasibility study demonstrated that there were 7 main challenges that fall within the implementation outcomes framework by Proctor et al [34]. Textbox 2 [34] provides a summary.

Textbox 2. Summary of key themes from the feasibility testing results informed by the implementation outcomes framework by Proctor et al.

Acceptability

- Challenges with recruitment and uptake of the app

Adoption

- Challenges with use and engagement

Appropriateness

- Screening challenges
- Digital divide
- Limitations to digital applications for mental health

Feasibility

- Technological challenges
- Funding and policy constraints

XSL•FO
RenderX

Acceptability: Challenges With Recruitment and Uptake of the App

The first challenge reported was regarding the uptake of the app as social workers and others from the SPC shared that many participants were reluctant to download the VMood app and participate. Reasons cited included resistance to more apps and new technology and security concerns. The policy stakeholder reported that "people are exposed to an abundance of apps; therefore, it is difficult to keep users engaged and accessing the app frequently if there is nothing of interest inside it" and, furthermore, "people in Thanh Hoa ... are reluctant to install unfamiliar apps on their phones out of fear of viruses and malware" (IPS-01), which was another factor impacting participants' willingness to download and engage with the app. A provider shared the following:

> *...some people are afraid and worried about their information being leaked.* [FGPR-02]

App users added the following:

> *I don't use others [software/apps] since I don't find them useful or practical.* [IAU-01]

Therefore, "the first challenge in terms of getting people to try out this app is to encourage them to download and install a completely new app" [IPS-01].

In addition, the policy stakeholder shared that many individuals interested in the usability testing were not able to participate because their mental health was good according to the PHQ-9 measure and, thus, were ineligible. Many of the individuals were recruited through community events and were quite active in their community and, thus, less likely to have depression. A solution used by the team to address this was the shift to recruiting using a list of high-risk individuals prepared by the communes. A provider suggested the following:

> *Instead, we could target specifically those who have recently experienced a traumatic event, such as the demise of a family member; then, the likelihood of them experiencing symptoms of depression will be much higher. By doing so, we would be able to reach a group in need of our services and support much more efficiently than if we were to search and look for 1,000 random individuals.* [IPR-01]

Another provider suggested the following:

> *...we must consider the commune people's committee (the administrative centre for government in the commune) and the medical centre, as they are frequently packed with people and have a high likelihood of having someone in need of our services or projects.* [IPR-02]

Solutions to address these recruitment and uptake challenges centered on using promotional efforts to increase the legitimacy of VMood and awareness of its benefits "to get more people interested in [the] app" (IAU-03). Another app user stated the following:

> *We must demonstrate the superior features of the app, particularly the application's features and functionality, so that people understand that using the app is advantageous for them and, as a result, more people are inclined to use it voluntarily.* [IAU-04]

A social worker reported the following:

> *...the communication activity to promote this model is quite limited, so sometimes only people in the Project's region have heard of it, and not many others are aware of it.* [IPR-02]

An app user mentioned the following:

> *...in order for this software to reach a wider audience, communication and propaganda processes for people must be emphasized...When they understand how to use the app and enjoy using it, they will tell others, and ultimately, everyone will start using it.* [IAU-06]

Strategies included sharing more information and education through Zalo (Vietnam's messaging platform) and Facebook group chats of the commune's various unions (eg, women's, youth, and pregnant women unions) and by combining with promotional strategies for other initiatives. The team has also created a series of promotional broadcast messages to be sent via Zalo and Facebook groups to help address concerns among the community about scams and phishing.

Participants also spoke about the importance of increasing the visibility of government support to encourage trust and uptake, especially in the context of increased scams orchestrated through digital media such as smartphones. For example, a provider indicated the following:

> *I think it [VMood] should have a QR code or an official, legal advertisement for this app; and the site where we place promotion signs or posters, or where we help users in installing this software, should be a government-related and trustworthy location, such as a medical station. [Otherwise,] people might think it's a scam.* [FGPR-01]

An app user echoed this, stating the following:

> *...people today frequently have faith in initiatives that receive support from the Authority. So if it is protected by the government, people will certainly follow and try it out.* [IAU-04]

There was a recommendation from the provincial SPC for MOLISA to dispatch notifications to the provinces to clarify the information on VMood and increase trust in the app. Another provider shared their perspectives from a different angle:

> *As a professional, I have no right to impose social issues (such as forcing them to install the app) on people. However, if we promote the use and installation of this app through and with the assistance of the government, local authorities, and the district or commune-level government, then participation will undoubtedly increase.* [FGPR-02]

Another challenge to recruitment was regarding low mental health awareness and stigma. SPC staff reported that commune health staff at times introduced the project using words that held sensitivity in the community, such as mental illness, which people still associate with schizophrenia. Consequently, some

XSL•FO
RenderX

individuals in the community were afraid of being discriminated against and, therefore, were reluctant to participate in the project. To help address this, the PHAD team has prepared an introduction sheet for the local commune health staff to refer to when recruiting participants.

Promisingly, participants in all 3 groups from the usability testing spoke about the convenience and general acceptability of VMood in their communities, indicating that it was simple, easy to use, and suitable for all ages. Once individuals install the app, "from then, people will then begin using and recognizing the app's utility and usefulness" (IPS-01).

Adoption: Challenges With Use and Engagement

Findings from the qualitative interviews and focus groups indicated that app user engagement was fairly low. The first factor influencing participant's adoption of VMood was the busy schedules of the app users. While they reported enjoying the app and certain features, including the "multiple-choice questions" and videos, one app user indicated the following:

> *I haven't watched all of the videos because I do not have sufficient time to watch them all.* [FGPR-02]

Another participant similarly reported the following:

> *I was able to use the app perhaps 2-3 times [in a month]. It's not that often because I don't have a lot of time due to my workload.*

A provider mentioned that people working in agriculture may not find the app as acceptable as perhaps a younger participant because of their busy schedule during "agricultural seasons," where "people will be preoccupied with work or study," which may result in "an overall drop in their interest in mental health" (IPR-01).

Strategies to increase participant engagement included incentives such as "monetary support" (FGPR-01) in the form of a phone card or, as another provider from the same focus group mentioned, "a form of award that could be turned into cash every time a user correctly answers a question." This provider recognized that "the project won't be able to offer monetary assistance endlessly" but perhaps "could opt to keep it simple, such as 10 points=VND 10,000 (CAD $0.50)." They concluded that "this would very certainly encourage users to open the app and use it more frequently." Another strategy was suggested by PHAD, informing participants that they would be entered into a draw for an Android smartphone.

Another strategy suggested was "a management team, or group leader" who "should be assigned to each locality. They could be in charge of contacting or reminding people [from that area] through this app" (FGPR-01). The participant clarified that "an administrator account must be created for the medical station so that they can administrate their patients" and "keep track of them." Another provider from the first focus group spoke about this need, indicating that, in their commune, there was "a large number of people who have installed this app; nevertheless, the management provided to manage and follow up with this group is insufficient" (FGPR-01).

App use data from the early phase of the Thanh Hoa RCT implementation showed that more than half (12/20, 60%) of the app user participants had high levels of engagement with the app (>120 minutes), 30% (6/20) had moderate levels of engagement, and only 10% (2/10) of the participants engaged with the app for <60 minutes since enrollment. It should be noted that use for app users was tracked beyond the 3-month intervention period for an additional 3-month follow-up period (total of 6 months). These data, which were last extracted on June 24, 2024, showed that, since enrollment, the 20 app user participants engaged with the app, on average, for 145.9 (SD 96.2) minutes, with the highest engagement being 366 minutes and the lowest being 13 minutes (median 125.5, IQR 83.5-168.25, range 13-353). These findings suggest that some of the strategies to increase participant engagement may have been effective [55]. For the purpose of this study, the engagement cutoffs were established based on preliminary interviews exploring engagement with app users in the RCT from 2 provinces, and the clinically significant treatment dose will be investigated further in the RCT.

Appropriateness

Screening Challenges

The first challenge reported that relates to VMood's appropriateness was regarding screening. Thanh Hoa SPC participants reported that some individuals responded falsely to the embedded PHQ-9 screening measure to gain access to the app. Certain app users who were invited to the usability testing initially scored low on the PHQ-9 and, per the study protocol, were asked to retake the PHQ-9 in 2 weeks. Some participants then asked for and received instructions from the commune health staff on how to obtain access to the app by scoring high on the PHQ-9. Thus, many participants in the usability testing with depression as measured by a PHQ-9 score of ≥5 (n=154) did not, in fact, have depression. Commune health staff confirmed that the 3.9% (6/154) of app users who participated in the interviews were true positives. This screening challenge was discussed at team meetings, and it was decided that the full version of the app would be provided to all app users regardless of depression caseness, so there would be no need for faking their PHQ-9 scores. Another suggestion was made to communicate clearly to the communes on the intervention benefits for those who need it.

Digital Divide

Another challenge that impacted VMood's appropriateness was the digital divide. Although most participants commented on the strengths of the VMood digital intervention in reaching a broader population, there were concerns raised regarding the digital divide—namely, how certain populations may be left behind. For example, a clinical expert shared the following:

> *I can see the app can be very applicable to my generation who are quite familiar with using apps and internet to navigate life situations. But for my parents' generation, for example, using this app can prompt a lot of anxiety for them to be logged in and where are the buttons and all that.* [IEX-05]

Similarly, an app user stated the following:

XSL•FO
RenderX

> *...senior citizens ... have a more difficult time processing than young people like me.* [IAU-03]

Another app user also commented on their parents, who "are not proficient with technology" and may not be able to log in and "use this app on their own." This was similar to the in-person RCT, where older individuals had difficulty reading the ASW and asked their family members to help read to them. Despite this, this app user indicated that they would introduce the VMood app "if any family members needed it."

Other participants also commented on structural barriers to widespread implementation and uptake of VMood. One provider reported the following:

> *As urbanization expands to the countryside, a group of young people might be affected from the use of the Internet.* [FGPR-01]

Another provider reported the same concern regarding internet connectivity issues:

> *In some locations, the connection or Wi-Fi signals are inadequate or unstable, which could have an effect on the implementation of this project.* [IPR-01]

Despite potential issues with Wi-Fi signals, it was reported by the Thanh Hoa leadership that Wi-Fi is available in most homes because television typically runs on internet connection, delivered through Wi-Fi. Similarly, 4G is available in most households except for some individuals from older generations. In addition, the project provides app users with a phone credit for their participation.

The team discovered that a larger number of individuals than anticipated did not own newer smartphones. An app user mentioned the following:

> *...while technology coverage in Vietnam is high, the proportion of the population owning smart devices is limited.* [IAU-06]

As a result, they "rarely have the opportunity to interact with (VMood)" (IEX-05). A clinical expert shared that "farmers, for example, do not know how to use smartphones and do not even own them ... if you move further away from the city, then you would see that not everyone knows how to use them." Another example of lack of smartphone ownership was given by a provider, who explained that "it can be difficult to install the app if a family has 2-3 members who are depressed, but only one of them has a smartphone" (FGPR-02). The decision made after discussing this with the SPC staff and staff at the local communes was to recruit only those with smartphones for the RCT as it would not be feasible to provide smartphones to all app user participants.

Finally, the team also learned during the usability testing that VMood did not always function properly on older smartphones that were not able to support newer apps. After discussing with the app development company, they confirmed that this limitation was due to Apple and Android providers that do not support newer software, which means that VMood will likely be available only to users of newer smartphones.

Limitations to Digital Applications for Mental Health

Overview

Another challenge influencing VMood's appropriateness was limitations to digital applications for mental health. These included the complexity of the app, such as the CBT principles VMood is based on, along with cultural and linguistic challenges. The first point participants spoke about was the complexity of the app. While one provider from the first focus group indicated that "in my opinion, the installation process is relatively easy and convenient for most people" (FGPR-01), another provider from that focus group, who indicated that they were a "collaborator [community women's union staff] who provides advice to people ... making it easier for them to access the app," stated the following:

> *The initial installation phase is easy and convenient for people to follow, but because this app is quite new, the app has not yet reached a high level of usage. Therefore, I believe that this initiative still requires collaborators to assist users in the process of using this app to evaluate their health.* [FGPR-01]

App users similarly mentioned the following:

> *It might be difficult if I let myself figure it out because it isn't written on the software about what to do or why. How can I accomplish it if I don't know what it is?* [IAU-05]

App users also spoke about the complexity of the VMood content. For example, one app user indicated that, while they enjoyed watching the videos on the app, "the explanation for the app's content is a concern. I'm not sure if it's because of my age or my academic level, but there are several things in the app that I have trouble understanding" (IAU-02). Suggestions were provided by participants to help reduce some of these concerns. A clinical expert proposed the creation of "a group of the people who are using the app and having a little talking circle ... empowering each other" (IEX-04).

Cultural and Linguistic Challenges and Suggestions

Despite the ASW being tailored to fit the local context, participants shared feedback on including more culturally appropriate examples for app components, such as components on reactivating one's life, a food guide, and creating a better sleep environment (eg, "including traditional local remedies for sleep" [IEX-03]), to improve VMood's cultural appropriateness. One clinical expert mentioned that the app is focused more on the "brain," but "Vietnamese people, they can be superstitious" (IEX-04). Another participant remarked the following:

> *...a few questions only apply to a specific group of workers such as office workers, whereas lots of Vietnamese are freelance or manual labor. ... The context of the question should be relevant and close to a normal daily life of a Vietnamese.* [FGPR-01]

Another provider from the same focus group suggested the following:

> *...the next training session at the centre should be tailored to suit the local area.* [FGPR-01]

XSL•FO
RenderX

This was noted as being especially important considering that many of the participating districts in Thanh Hoa province are "mostly in rural areas." Finally, a provider spoke about the implementation and scale-up challenges posed by cultural and linguistic issues:

> *I imagine that each region will face distinct difficulties, such as mountainous regions and remote islands. Because we are discussing a nationwide implementation of this app, each region will face particular difficulties. For instance, it might be challenging to introduce a Kinh-language app in areas where people speak ethnic languages. These are the difficulties I could imagine if we were to deploy this application nationwide.* [IPR-01]

The team reviewed the suggestions, and Vietnamese-speaking members of the team, consulting with a social worker living in a rural community, added more culturally relevant examples for the realistic thinking and problem-solving skills components to increase appropriateness. Promoting VMood acceptability and usability requires an understanding of and tailoring to local traditions and culture, which participants indicated will be a key factor to its sustainability across Vietnam's diverse communities, each with its own unique needs.

Feasibility

Technological Challenges

Technological challenges included delays experienced with the app development process and missing app use data from the usability testing. The team wanted to ensure that VMood had theoretical fidelity to the original in-person intervention and practical fidelity in terms of usability in its digital format. The study by Chau et al [35] provides the results of fidelity testing. As a result of the technological challenges, the development process was long and required frequent exchanges among the larger project team, our implementing partners in Vietnam who were working directly with the app development company, and the app development company. Any edits made to the app also required testing and fixing any bugs in addition to approval from the Apple App Store. This further delayed the editing process, sometimes causing updates to take >2 months to go live. Developing the app was a major undertaking that required significant input and resources.

Some participants from the usability testing indicated that the text size was a bit small, especially for older adults, and suggested increasing the font size. The team discussed this with the app development company and agreed to delay this change until after the RCT is launched in Thanh Hoa as making this update would require weeks and, furthermore, with an increase in font size, the text may flow out of the frame on smartphones. Another issue experienced during the usability testing was that the instructional videos lagged in some areas with slower internet speed. To help mitigate this, YouTube was chosen as the video hosting platform as it can adjust the video stream quality based on the user's internet connection.

Another technological challenge encountered was with the technical complications and oversight in the original app design despite having a sophisticated design process in place. The app did not have the functionality for capturing app use data, including the number of log-ins. We were under some pressure from the Government of Vietnam to launch VMood fairly quickly to meet their fiscal timelines, and the app design company had not yet set up the app to capture use frequency and duration. We were made aware of this after usability testing had been completed and, thus, were not able to include any app use data from the usability testing participants. The app was subsequently updated to include built-in metrics to capture use data such as frequency and duration of engagement with the various skill components.

Funding and Policy Constraints

Funding and policy constraints were the last challenge that the team encountered. Social workers participating in the larger RCT will provide support as part of their regular job responsibilities, which typically involve a heavy workload. Thus, their time will be limited for both training and project implementation. Some social workers commented on the limitations to the training program:

> *Regarding in-depth training or how to deliver counseling for a case of depression, we have not received such training. We have not received specialized training on depression; we have only received training related to the book, including how to use the book/handouts and how to assist others in using/reading this book. ... I am interested in attending other in-depth training sessions on this topic if they are offered.* [IPR-02]

Despite this, social workers indicated that they were satisfied overall and appreciative of the training and mentioned that, before receiving the training provided by project team members on how to administer the intervention and on basic depression, they had received little or no training on depression. For example, one provider stated the following:

> *I thought the training session was very sociable and enjoyable, and I learned some new things [about depression] that I had never known before.* [FGPR-01]

Another provider also reported that the training program was "quite handy. I was able to gain extra knowledge and also use what I learned in these sessions for my everyday tasks" (IPS-01). Furthermore, for VMood installation and navigation, which are the social workers' main functions on this project, a provider shared the following:

> *If we are only discussing whether the training is sufficient to allow us to assist users with accessing and using the application, then I believe the training materials are adequate.* [IPS-01]

Other social workers also shared that they had limited capacity to provide extensive support through the app. The app is not meant to be a magic bullet for everything but is intended as a widely accessible tool for those with mild to moderate depression, with minimal support on app navigation and basic depression provided by social workers.

XSL•FO
RenderX

High turnover of social workers posed a challenge for the feasibility testing, slowing down implementation in the early phase of the RCT in Thanh Hoa as there was a need to train new social workers fairly frequently, but the team has developed clear training guidelines supporting the web-based training program, and provincial project managers will be responsible for training new personnel carefully according to the guidelines. In addition, the director of Thanh Hoa's SPC changed just before the expected launch of the RCT, which also contributed to a delayed RCT launch in Thanh Hoa.

Despite the challenges described previously, ultimately, most participants indicated that the VMood app is an important tool that can assist individuals in the community experiencing depression. Strengths reported for the digital intervention included being able to use the app privately "at home" (IEX-01) in their own environment, potentially mitigating some of the effects of stigma. One clinical expert stated the following:

> *It's great because depressed patients frequently don't want to tell anyone about their problems, but they can download this app, watch it, understand it for themselves, and change their mindset.* [IEX-02]

In addition to the ability to use VMood in the comfort of participants' environment and at their convenience, another advantage is the ability of VMood to "reach a broader audience" (FGPR-01).

A provider who had been involved with the RCT for the in-person intervention compared it with the digital version and shared the following about the previous intervention:

> *Since the release of the VMOOD software, I've realized that it's much better in terms of security. Because previously we had to send social workers to people's homes to ask questions, we no longer need to do so, and users can now experience this software on their own. This will reduce the likelihood of people being shy while answering the PHQ-9 questionnaire, which could contribute to inaccuracy. ... In the past, when we went to collect people's responses, they sometimes responded differently despite their obvious illness behavior.* [IPR-02]

In summary, participants expressed unanimous support for the VMood intervention and highlighted the need for increased community-based mental health care because, currently, "there is no infrastructure in Thanh Hoa to support these people" (IPR-01). This same provider reported that "because in today's developed, modern society, life and work are fraught with pressures and difficulties. Consequently, I believe that numerous people are struggling with their mental health. I believe that many people will be interested in this topic." A clinical expert also recognized the limitations to the current mental health care system and shared that they hoped that the intervention would be scaled up to assist more people to "relieve some of the burden on [them]" (IEX-01). The Government of Vietnam recognizes the gaps in community-based mental health care and has provided ongoing support, including commitment in the form of matching funds, to support VMood implementation and scale-up.

Discussion

Principal Findings

Feasibility testing of the VMood app demonstrated a number of key challenges to its development and implementation that fell within 4 of the implementation outcomes by Proctor et al [34]: acceptability, adoption, appropriateness, and feasibility. Recruitment challenges arose that impacted VMood's *acceptability*; app use and engagement challenges impacted its *adoption* in the community; screening challenges, along with considerations regarding a digital divide and limitations to digital applications for mental health, impacted VMood's *appropriateness*; and, finally, technological challenges and funding and policy constraints needed to be considered to ensure VMood's *feasibility* within the Vietnamese context. Extensive discussions at various meetings throughout the app development process, usability testing, and launch of the RCT in Thanh Hoa among team members and with leadership at the Thanh Hoa SPC helped the team fully understand the challenges and brainstorm timely solutions to help mitigate them.

Comparison With Prior Work

Participants found VMood to be acceptable within the Vietnamese context, although there were extensive challenges reported with recruitment, such as resistance to more apps and new technology, security concerns, and persisting low mental health awareness and stigma. Similarly, other studies have shown that data privacy and security are concerns that impact acceptability [56]. Mental illness stigma is high in Vietnam, as in many other Asian countries [57], but promisingly, depression app use has been shown to increase mental health literacy and awareness and mitigate the effects of stigma as people are able to access them in their own environment and experience increased anonymity [58]. A main solution to enhance the acceptability and legitimacy of VMood was to increase promotional efforts, including sharing more information and education through social media and exhibiting government support more explicitly on promotional materials. Usability testing also highlighted the importance of a targeted recruitment approach using lists prepared by commune health staff of individuals considered to be at a higher risk of developing depression based on recent adverse life changes. This strategy appears to have been effective in the first phase of the RCT. Participants also reported on some challenges to adoption, including the busy schedule of app users, which may prevent them from fully engaging with the app, especially during busy periods such as harvesting season. Incentives in the form of money and a chance to win an Android smartphone were incorporated for the RCT to encourage participant engagement. Finally, the specific symptoms of depression, such as low motivation, concentration difficulties, and behavioral avoidance, may pose additional challenges to app engagement [22]. For example, Arean et al [55] found that participants with a higher baseline depression score (measured using the PHQ-9) accessed their 2 depression apps less frequently, suggesting that depression apps may have the greatest impact on individuals with milder levels of depression [55]. Screening through depression apps could help provide an indicator of those who require and should be prioritized for more in-person treatment.

XSL•FO
RenderX

App use metrics from the early phase of the Thanh Hoa RCT indicate that 60% (12/20) of the participants had high engagement with VMood (>120 minutes), 30% (6/20) had moderate engagement (60-120 minutes), and 10% (2/20) had low engagement (<60 minutes). A closer comparison between app user engagement and their PHQ-9 scores is warranted and will be conducted in the larger RCT.

The appropriateness of VMood was impacted by screening challenges in which a number of participants responded falsely to questions in the embedded PHQ-9 to gain access to VMood. Data integrity has been reported as an increasing concern in web-based research [59], defined as "any research involving the remote acquisition of data from or about human participants using the internet and its associated technologies" [60]. Specific concerns relevant to this study include nongenuine participants, which is when participants lie about their experiences or identity, and misrepresentation, when participants exaggerate specific details [59]. Possible motivations for participants falsely responding to the PHQ-9 include gaining access to the intervention and monetary incentive [59]. As discussed in the Results section, one strategy that the team decided to use to help address this was to provide the full app to all app users who signed up. There are also considerations that need to be taken for certain groups (eg, older adults) to ensure equitable access for all who might benefit and not further increase the digital divide, which is "a division between people who have access to and use of digital media and those who do not" [61]. The last challenge impacting appropriateness was limitations to digital applications for mental health. Lower-income countries have experienced faster mobile phone use growth rates that higher-income countries [58]. For example, Vietnam, a country of approximately 100 million people, has approximately 140 million mobile telephone subscriptions [62]. Despite this, it was found that many individuals in the communities did not, in fact, own smartphones or owned older versions that were not able to support newer apps. In addition, for a tool that is meant to be widely accessible and available, there were issues reported on the complexity of the app, including the CBT principles and cultural and linguistic concerns. These were addressed by encouraging participants to review the instructional videos thoroughly and incorporating more culturally relevant examples, respectively. The structured approach of CBT translates well to a mobile format [63], and with the accompanying videos and any necessary support provided by a social worker, community leaders were confident on VMood's appropriateness. Cultural appropriateness is a key consideration when adapting an intervention from one context to another. The study by Chau et al [64] provides an in-depth discussion on cultural fidelity. There is a growing understanding of and attention to ensuring culturally appropriate mental health care for people in diverse settings [65]. Despite some of the challenges reported on the complexity of the app, participants did not comment on some of the more common hurdles to usability found in the literature, such as the tool requiring too much time to enter data or being irrelevant to their needs [66].

As discussed, a main challenge encountered in this study was related to technological issues resulting in missing app use data from the usability testing. Despite this, through interviews and focus groups, usability testing participants were unanimous in expressing their support for the app and provided positive feedback on various components of the app, such as the instructional videos and realistic thinking skills. Interviews and focus groups are 2 of the most commonly used methods for usability testing, in addition to questionnaires and use data, and allow for the elucidation of users' experiences and the likeability of the app in addition to barriers to use and ideas for improvement [67]. App users also indicated that they would recommend VMood to friends and family based on their experiences up to the date of their interviews. App metrics from the first phase of the RCT launch in Thanh Hoa complemented the qualitative reports from the usability testing participants and showed that 60% (12/20) of the app users demonstrated high engagement with the app (>120 minutes), suggesting that they found it to be helpful for their needs.

In relation to the feasibility of VMood implementation, participants also commented on the importance of having support from social workers through the app and corresponding social worker training to ensure that they are adequately prepared to provide basic support. A systematic review examining the effects of mobile apps on stress, anxiety, and depression similarly found that integrating human support into digital health interventions has been shown to increase usability [24] and produce larger clinical effects than those of stand-alone apps [68]. Participants spoke about the importance of social worker support but also recognized the need for a training program that can facilitate rapid deployment of social workers and agreed that a balance needs to be struck between feasibility and appropriateness within current budgetary and policy constraints. It is important to recognize the concurrent responsibilities of social workers, who were participating in this project as part of their regular work duties, and avoid overburdening them, which in turn can contribute to high stress and high turnover.

Presently, evidence on the effectiveness and cost-effectiveness of smartphone-based apps is fairly limited [21,69], with at most only approximately 4% being evidence based [70]. There is also limited evidence on whether effectiveness studies are impacted by failed implementation [71]. Implementation science focuses on practical approaches to enhance the scale-up and sustainability of various evidence-based interventions [72] by helping bridge the longstanding knowledge-practice or know-do gap [73]. The various challenges reported during VMood development and implementation, along with solutions to address them, will be critical as the research team conducts the RCT across the remaining 7 provinces (this work is ongoing) to help avoid similar challenges. Suggestions from the usability testing participants, leadership in the communities, and the research team may help mitigate the high dropout rates and low engagement with technology that are increasingly recognized as challenges for digital health interventions in terms of both RCT validity and sustained engagement [66]. The findings of this feasibility study will also contribute important evidence informing other research examining the development and implementation of similar digital interventions for mental health. However, it is important to recognize that certain strategies to increase recruitment and engagement, such as providing app

XSL•FO
RenderX

users with a phone credit for their participation and entering them into a draw for an Android smartphone, were necessary and suitable for the RCT but may not be feasible for scale-up. Thus, should the RCT demonstrate VMood's effectiveness, other strategies supporting real-world implementation beyond the context of an RCT will be required for its scale-up. Ultimately, despite these challenges and limitations, participants from the usability testing spoke unanimously about the importance of a low-cost and accessible intervention such as VMood that could help fill a growing treatment gap in the community, where mental health conditions are prominent and on the rise, especially in the context of the COVID-19 pandemic [3].

Limitations

Due to challenges encountered during the development, feasibility, and usability exploration of the VMood app, we were unable to carry out the feasibility study as originally planned. However, this formative paper focuses on the immediate strategies and adjustments that the team used to address the challenges, taking advantage of a unique opportunity to share the lessons learned regarding the development and feasibility testing of a digital mental health intervention that we believe could be helpful for others similarly interested in developing and testing similar interventions. In addition, while the team used strategies to help ensure cultural and linguistic appropriateness capturing Vietnam's diversity, such as conducting beta testing with individuals familiar with different dialects, we recognize that it may be difficult for 1 app to fully capture Vietnam's diversity. Should RCT findings demonstrate VMood's effectiveness, we will adapt the app to fit different Vietnamese dialects and subgroups (eg, ethnic minority groups and individuals living in the industrial zones) and subsequently examine data across the different subgroups.

Conclusions

Adjustments based on recommendations critical to VMood's usability were made to the app, which is currently being tested in an RCT across 8 provinces of Vietnam. The RCT will examine VMood's effectiveness along with cost-effectiveness. The work by Chau et al [74] provides the RCT protocol. The Government of Vietnam has provided ongoing support for this program of research as part of a strong and sustained research-policy collaboration since 2013 (see the work by Murphy et al [75] for further details) and has committed to scale-up in the other Vietnamese provinces should results demonstrate effectiveness. The findings of this implementation science feasibility study contribute important evidence on the challenges to the development, feasibility, and usability testing of a digital depression app adapted from an in-person intervention in an RCT in Vietnam, a low- to middle-income country. The findings have applicability for others looking to develop and implement digital interventions in similar contexts, particularly given the demand for digital mental health solutions to expand mental health care, serving as a unique opportunity to share the lessons that our team learned regarding the development and testing process.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

RWL has received honoraria for ad hoc speaking or advising and consulting or received research funds from AbbVie, the Asia-Pacific Economic Cooperation, Bausch, BC Leading Edge Foundation, Brain Canada, the Canadian Institutes of Health Research, the Canadian Medical Protective Association, the Canadian Network for Mood and Anxiety Treatments, Carnot, CB Solutions, Genome BC, Grand Challenges Canada, Healthy Minds Canada, Janssen, Lundbeck, the Michael Smith Foundation for Health Research, Mitacs, Neurotorium, the Ontario Brain Institute, Otsuka, the Shanghai Mental Health Center, Unity Health, the Vancouver Coastal Health Research Institute, and the Vancouver General Hospital & University of British Columbia Hospital Foundation. The other authors declare that they have no competing interests.

References

1. GBD 2019 Mental Disorders Collaborators. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet Psychiatry 2022 Feb;9(2):137-150 [FREE Full text] [doi: 10.1016/S2215-0366(21)00395-3] [Medline: 35026139]
2. Mental disorders. World Health Organization. 2022 Jun 8. URL: https://www.who.int/news-room/fact-sheets/detail/mental-disorders [accessed 2025-06-02]
3. COVID-19 Mental Disorders Collaborators. Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. Lancet 2021 Nov 06;398(10312):1700-1712 [FREE Full text] [doi: 10.1016/S0140-6736(21)02143-7] [Medline: 34634250]
4. Patel V, Saxena S, Lund C, Thornicroft G, Baingana F, Bolton P, et al. The Lancet Commission on global mental health and sustainable development. Lancet 2018 Oct 27;392(10157):1553-1598. [doi: 10.1016/S0140-6736(18)31612-X] [Medline: 30314863]
5. COVID-19 disrupting mental health services in most countries, WHO survey. World Health Organization. 2020 Oct 5. URL: https://www.who.int/news/item/05-10-2020-covid-19-disrupting-mental-health-services-in-most-countries-who-survey [accessed 2023-03-20]

XSL•FO
RenderX

6. Duong CB, Van Tran N, Nguyen AH, Le TN, Ha BH, Do CN, et al. Impacts of COVID-19 crisis and some related factors on the mental health of 37150 Vietnamese students: a cross-sectional online study. BMC Public Health 2023 Mar 07;23(1):445 [FREE Full text] [doi: 10.1186/s12889-023-15317-3] [Medline: 36882752]
7. Hoang NA, Van Hoang N, Quach HL, Nguyen KC, Duong LH, Pham TQ, et al. Assessing the mental effects of COVID-19-related work on depression among community health workers in Vietnam. Hum Resour Health 2022 Aug 19;20(1):64 [FREE Full text] [doi: 10.1186/s12960-022-00760-x] [Medline: 35986294]
8. Aksunger N, Vernot C, Littman R, Voors M, Meriggi NF, Abajobir A, et al. COVID-19 and mental health in 8 low- and middle-income countries: a prospective cohort study. PLoS Med 2023 Apr 6;20(4):e1004081 [FREE Full text] [doi: 10.1371/journal.pmed.1004081] [Medline: 37023021]
9. GBD 2021 ASEAN Mental Disorders Collaborators. The epidemiology and burden of ten mental disorders in countries of the Association of Southeast Asian Nations (ASEAN), 1990-2021: findings from the Global Burden of Disease Study 2021. Lancet Public Health 2025 Jun;10(6):e480-e491 [FREE Full text] [doi: 10.1016/S2468-2667(25)00098-2] [Medline: 40441815]
10. Population, total - Viet Nam. World Bank Group. URL: https://data.worldbank.org/indicator/SP.POP.TOTL?locations=VN [accessed 2025-06-02]
11. Eilinghoff L, Nguyên VT, Hahn E, Nguyên VP, Lê CT, Lê TT, et al. Changes in attitudes toward persons with mental disorders after attendance of a psychiatric curriculum among medical students in Vietnam: a cross-sectional study. Asian J Psychiatr 2024 Mar;93:103949 [FREE Full text] [doi: 10.1016/j.ajp.2024.103949] [Medline: 38335892]
12. Li J. Digital technologies for mental health improvements in the COVID-19 pandemic: a scoping review. BMC Public Health 2023 Mar 01;23(1):413 [FREE Full text] [doi: 10.1186/s12889-023-15302-w] [Medline: 36859184]
13. Philippe TJ, Sikder N, Jackson A, Koblanski ME, Liow E, Pilarinos A, et al. Digital health interventions for delivery of mental health care: systematic and comprehensive meta-review. JMIR Ment Health 2022 May 12;9(5):e35159 [FREE Full text] [doi: 10.2196/35159] [Medline: 35551058]
14. Firth J, Torous J, Nicholas J, Carney R, Pratap A, Rosenbaum S, et al. The efficacy of smartphone-based mental health interventions for depressive symptoms: a meta-analysis of randomized controlled trials. World Psychiatry 2017 Oct;16(3):287-298 [FREE Full text] [doi: 10.1002/wps.20472] [Medline: 28941113]
15. Torous J, Bucci S, Bell IH, Kessing LV, Faurholt-Jepsen M, Whelan P, et al. The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. World Psychiatry 2021 Oct;20(3):318-335 [FREE Full text] [doi: 10.1002/wps.20883] [Medline: 34505369]
16. Eccles MP, Mittman BS. Welcome to implementation science. Implement Sci 2006 Feb 22;1:1. [doi: 10.1186/1748-5908-1-1]
17. Lyon AR, Pullmann MD, Jacobson J, Osterhage K, Achkar MA, Renn BN, et al. Assessing the usability of complex psychosocial interventions: the intervention usability scale. Implement Res Pract 2021;2:2633489520987828 [FREE Full text] [doi: 10.1177/2633489520987828] [Medline: 35601889]
18. Dix A. Human-computer interaction. In: Liu L, Özsu MT, editors. Encyclopedia of Database Systems. Boston, MA: Springer; 2009.
19. Maramba I, Chatterjee A, Newman C. Methods of usability testing in the development of eHealth applications: a scoping review. Int J Med Inform 2019 Jun;126:95-104. [doi: 10.1016/j.ijmedinf.2019.03.018] [Medline: 31029270]
20. Corrigan P. How stigma interferes with mental health care. Am Psychol 2004 Oct;59(7):614-625. [doi: 10.1037/0003-066X.59.7.614] [Medline: 15491256]
21. Marshall JM, Dunstan DA, Bartik W. Clinical or gimmickal: the use and effectiveness of mobile mental health apps for treating anxiety and depression. Aust N Z J Psychiatry 2020 Jan;54(1):20-28. [doi: 10.1177/0004867419876700] [Medline: 31552747]
22. Molloy A, Anderson PL. Engagement with mobile health interventions for depression: a systematic review. Internet Interv 2021 Sep 11;26:100454 [FREE Full text] [doi: 10.1016/j.invent.2021.100454] [Medline: 34621626]
23. Wu A, Scult MA, Barnes ED, Betancourt JA, Falk A, Gunning FM. Smartphone apps for depression and anxiety: a systematic review and meta-analysis of techniques to increase engagement. NPJ Digit Med 2021 Feb 11;4(1):20 [FREE Full text] [doi: 10.1038/s41746-021-00386-8] [Medline: 33574573]
24. Khademian F, Aslani A, Bastani P. The effects of mobile apps on stress, anxiety, and depression: overview of systematic reviews. Int J Technol Assess Health Care 2020 Dec 14;37:e4. [doi: 10.1017/S0266462320002093] [Medline: 33314997]
25. Bilsker D, Goldner EM, Anderson E. Supported self-management: a simple, effective way to improve depression care. Can J Psychiatry 2012 Apr;57(4):203-209. [doi: 10.1177/070674371205700402] [Medline: 22480584]
26. Murphy JK, Xie H, Nguyen VC, Chau LW, Oanh PT, Nhu TK, et al. Is supported self-management for depression effective for adults in community-based settings in Vietnam?: a modified stepped-wedge cluster randomized controlled trial. Int J Ment Health Syst 2020 Feb 12;14:8 [FREE Full text] [doi: 10.1186/s13033-020-00342-1] [Medline: 32071614]
27. Durst D, Lanh TH, Pitzel M. A comparative analysis of social work in Vietnam and Canada: rebirth and renewal. J Comp Soc Work 2010 Oct 01;5(2):77-88. [doi: 10.31265/jcsw.v5i2.63]
28. Lan NT, Hugman R, Giang TT. Social work education in Vietnam: the development, challenges and ongoing effort to indigenization in the past two decades. Soc Work Educ 2024 May 30;44(1):22-36. [doi: 10.1080/02615479.2024.2360056]

XSL•FO
RenderX

29. Murphy J, Nguyen VC, Chau L, O'Neil J, Hoi NV, Minas H. From innovation to scale: policy engagement for enhanced community-based depression care in Vietnam. In: Okpaku S, editor. Innovations in Global Mental Health. Cham, Switzerland: Springer; 2019.
30. Luong NK, La DC, Truong LV, Ha TT, La N, Minas H. Viet Nam. In: Minas H, editor. ASEAN Mental Health Systems. Jakarta, Indonesia: ASEAN Secretariat; 2016.
31. Minas H, Edington C, La N, Kakuma R. Mental health in Vietnam. In: Minas H, Lewis M, editors. Mental Health in Asia and the Pacific. Boston, MA: Springer; 2017.
32. Ng CH, Than PT, La CD, Van Than Q, Van Dieu C. The national community mental health care project in Vietnam: a review for future guidance. Australas Psychiatry 2011 Apr;19(2):143-150. [doi: 10.3109/10398562.2011.563308] [Medline: 21443393]
33. Vuong DA, Van Ginneken E, Morris J, Ha ST, Busse R. Mental health in Vietnam: burden of disease and availability of services. Asian J Psychiatr 2011 Mar;4(1):65-70. [doi: 10.1016/j.ajp.2011.01.005] [Medline: 23050918]
34. Proctor E, Silmere H, Raghavan R, Hovmand P, Aarons G, Bunger A, et al. Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda. Adm Policy Ment Health 2011 Mar;38(2):65-76 [FREE Full text] [doi: 10.1007/s10488-010-0319-7] [Medline: 20957426]
35. Chau LW, Murphy JK, Nguyen VC, Tran H, Minas H, Lam RW, et al. Conceptualizing and testing fidelity-adaptation in the context of developing a digital intervention for depression from an evidence-based in-person format. Research Square. Preprint posted online on November 15, 2024 [FREE Full text]
36. Nilsen P. Making sense of implementation theories, models and frameworks. Implement Sci 2015 Apr 21;10:53 [FREE Full text] [doi: 10.1186/s13012-015-0242-0] [Medline: 25895742]
37. Proctor EK, Bunger AC, Lengnick-Hall R, Gerke DR, Martin JK, Phillips RJ, et al. Correction: ten years of implementation outcomes research: a scoping review. Implement Sci 2023 Oct 25;18(1):53 [FREE Full text] [doi: 10.1186/s13012-023-01313-z] [Medline: 37880754]
38. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
39. Phi HN, Manh BX, Ngoc TA, Chau PT, Tho TQ, Nghia NT, et al. Psychometric properties of Vietnamese versions of the clinician-rated and self-reported quick inventory of depressive symptomatology and the patient health questionnaire. East Asian Arch Psychiatry 2023 Jun;33(2):65-70 [FREE Full text] [doi: 10.12809/eaap2258] [Medline: 37400228]
40. Le Hoang Ngoc T, Le MA, Nguyen HT, Vo HV, Le NQ, Tang LN, et al. Patient Health Questionnaire (PHQ-9): a depression screening tool for people with epilepsy in Vietnam. Epilepsy Behav 2021 Dec;125:108446. [doi: 10.1016/j.yebeh.2021.108446] [Medline: 34839244]
41. Nguyen TQ, Bandeen-Roche K, Bass JK, German D, Nguyen NT, Knowlton AR. A tool for sexual minority mental health research: the Patient Health Questionnaire (PHQ-9) as a depressive symptom severity measure for sexual minority women in Viet Nam. J Gay Lesbian Ment Health 2016;20(2):173-191 [FREE Full text] [doi: 10.1080/19359705.2015.1080204] [Medline: 27642381]
42. Torous J, Staples P, Shanahan M, Lin C, Peck P, Keshavan M, et al. Utilizing a personal smartphone custom app to assess the Patient Health Questionnaire-9 (PHQ-9) depressive symptoms in patients with major depressive disorder. JMIR Ment Health 2015 Mar 24;2(1):e8 [FREE Full text] [doi: 10.2196/mental.3889] [Medline: 26543914]
43. Suh H, Moon E, Park JM, Lee BD, Lee YM, Jeong HJ, et al. A validation study of mental health monitoring through a mobile application. Psychiatry Investig 2023 Jun;20(6):575-580 [FREE Full text] [doi: 10.30773/pi.2023.0047] [Medline: 37357673]
44. Kim S, Lee K. Screening for depression in mobile devices using Patient Health Questionnaire-9 (PHQ-9) data: a diagnostic meta-analysis via machine learning methods. Neuropsychiatr Dis Treat 2021 Nov;Volume 17:3415-3430. [doi: 10.2147/ndt.s339412]
45. BinDhim NF, Shaman AM, Trevena L, Basyouni MH, Pont LG, Alhawassi TM. Depression screening via a smartphone app: cross-country user characteristics and feasibility. J Am Med Inform Assoc 2015 Jan 17;22(1):29-34 [FREE Full text] [doi: 10.1136/amiajnl-2014-002840] [Medline: 25326599]
46. Lewis JR. Sample sizes for usability studies: additional considerations. Hum Factors 1994 Jun;36(2):368-378. [doi: 10.1177/001872089403600215] [Medline: 8070799]
47. Gill P, Stewart K, Treasure E, Chadwick B. Methods of data collection in qualitative research: interviews and focus groups. Br Dent J 2008 Mar 22;204(6):291-295. [doi: 10.1038/bdj.2008.192] [Medline: 18356873]
48. Gill P, Baillie J. Interviews and focus groups in qualitative research: an update for the digital age. Br Dent J 2018 Oct 05;225(7):668-672. [doi: 10.1038/sj.bdj.2018.815] [Medline: 30287965]
49. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
50. Carroll C, Patterson M, Wood S, Booth A, Rick J, Balain S. A conceptual framework for implementation fidelity. Implement Sci 2007 Nov 30;2(1):40 [FREE Full text] [doi: 10.1186/1748-5908-2-40] [Medline: 18053122]
51. Burnard P, Gill P, Stewart K, Treasure E, Chadwick B. Analysing and presenting qualitative data. Br Dent J 2008 Apr 26;204(8):429-432. [doi: 10.1038/sj.bdj.2008.292] [Medline: 18438371]

XSL•FO
RenderX

52. Krefting L. Rigor in qualitative research: the assessment of trustworthiness. Am J Occup Ther 1991 Mar;45(3):214-222. [doi: 10.5014/ajot.45.3.214] [Medline: 2031523]
53. Lincoln YS, Guba EG. Naturalistic Inquiry. Thousand Oaks, CA: SAGE Publications; 1985.
54. NVivo: leading qualitative data analysis software. Lumivero. URL: https://lumivero.com/products/nvivo/ [accessed 2019-12-05]
55. Arean PA, Hallgren KA, Jordan JT, Gazzaley A, Atkins DC, Heagerty PJ, et al. The use and effectiveness of mobile apps for depression: results from a fully remote clinical trial. J Med Internet Res 2016 Dec 20;18(12):e330 [FREE Full text] [doi: 10.2196/jmir.6482] [Medline: 27998876]
56. Cohn SP. Privacy and confidentiality in the nationwide health information network. National Committee on Vital and Health Statistics. 2006. URL: https://ncvhs.hhs.gov/wp-content/uploads/2006/06/Report-Privacy-Confidentiality-in-the-Nationwide-Health-Information-Network-NHIN-508.pdf [accessed 2022-03-25]
57. Vaishnav M, Javed A, Gupta S, Kumar V, Vaishnav P, Kumar A, et al. Stigma towards mental illness in Asian nations and low-and-middle-income countries, and comparison with high-income countries: a literature review and practice implications. Indian J Psychiatry 2023 Oct;65(10):995-1011 [FREE Full text] [doi: 10.4103/indianjpsychiatry.indianjpsychiatry_667_23] [Medline: 38108051]
58. Brian RM, Ben-Zeev D. Mobile health (mHealth) for mental health in Asia: objectives, strategies, and limitations. Asian J Psychiatr 2014 Aug;10:96-100. [doi: 10.1016/j.ajp.2014.04.006] [Medline: 25042960]
59. French B, Babbage C, Bird K, Marsh L, Pelton M, Patel S, et al. Data integrity issues with web-based studies: an institutional example of a widespread challenge. JMIR Ment Health 2024 Sep 16;11:e58432 [FREE Full text] [doi: 10.2196/58432] [Medline: 39284170]
60. Ethics guidelines for internet-mediated research. The British Psychological Society. 2021 Jun 7. URL: https://www.bps.org.uk/guideline/ethics-guidelines-internet-mediated-research [accessed 2025-06-03]
61. van Dijk J. What is the digital divide? In: The Digital Divide. Hoboken, NJ: John Wiley & Sons; 2019.
62. Statistics. International Telecommunication Union. URL: https://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx [accessed 2022-03-04]
63. Stawarz K, Preist C, Tallon D, Wiles N, Coyle D. User experience of cognitive behavioral therapy apps for depression: an analysis of app functionality and user reviews. J Med Internet Res 2018 Jun 06;20(6):e10120 [FREE Full text] [doi: 10.2196/10120] [Medline: 29875087]
64. Chau LW, Murphy JK, Nguyen VC, Tran H, Minas H, Lam RW, et al. Conceptualizing and testing fidelity-adaptation in the context of developing a digital intervention for depression from an evidence-based in-person format. Research Square. Preprint posted online on November 15, 2024 [FREE Full text] [doi: 10.21203/rs.3.rs-5375723/v1]
65. Murphy J, Goldner E, Corbett KK, Morrow M, Nguyen VC, Linh DT, et al. Conceptualizing depression in Vietnam: primary health care providers' explanatory models of depression. Transcult Psychiatry 2018 Apr 06;55(2):219-241. [doi: 10.1177/1363461517748846] [Medline: 29405843]
66. Torous J, Lipschitz J, Ng M, Firth J. Dropout rates in clinical trials of smartphone apps for depressive symptoms: a systematic review and meta-analysis. J Affect Disord 2020 Feb 15;263:413-419. [doi: 10.1016/j.jad.2019.11.167] [Medline: 31969272]
67. Lyles CR, Sarkar U, Osborn CY. Getting a technology-based diabetes intervention ready for prime time: a review of usability testing studies. Curr Diab Rep 2014 Oct;14(10):534 [FREE Full text] [doi: 10.1007/s11892-014-0534-9] [Medline: 25173689]
68. Baumeister H, Reichler L, Munzinger M, Lin J. The impact of guidance on internet-based mental health interventions — a systematic review. Internet Interv 2014 Oct;1(4):205-215. [doi: 10.1016/j.invent.2014.08.003]
69. Donker T, Kleiboer A. Special issue: e-health innovations for global mental health. Glob Ment Health (Camb) 2018 Feb 01;5:e5 [FREE Full text] [doi: 10.1017/gmh.2018.6] [Medline: 29507741]
70. Lecomte T, Potvin S, Corbière M, Guay S, Samson C, Cloutier B, et al. Mobile apps for mental health issues: meta-review of meta-analyses. JMIR Mhealth Uhealth 2020 May 29;8(5):e17458 [FREE Full text] [doi: 10.2196/17458] [Medline: 32348289]
71. Mohr DC, Lyon AR, Lattie EG, Reddy M, Schueller SM. Accelerating digital mental health research from early design and creation to successful implementation and sustainment. J Med Internet Res 2017 May 10;19(5):e153 [FREE Full text] [doi: 10.2196/jmir.7725] [Medline: 28490417]
72. Theobald S, Brandes N, Gyapong M, El-Saharty S, Proctor E, Diaz T, et al. Implementation research: new imperatives and opportunities in global health. The Lancet 2018 Nov;392(10160):2214-2228 [FREE Full text] [doi: 10.1016/s0140-6736(18)32205-0]
73. Bauer MS, Damschroder L, Hagedorn H, Smith J, Kilbourne AM. An introduction to implementation science for the non-specialist. BMC Psychol 2015 Sep 16;3(1):32 [FREE Full text] [doi: 10.1186/s40359-015-0089-9] [Medline: 26376626]
74. Chau LW, Murphy JK, Nguyen VC, Xie H, Lam RW, Minas H, et al. Evaluating the effectiveness and cost-effectiveness of a digital, app-based intervention for depression (VMood) in community-based settings in Vietnam: protocol for a stepped-wedge randomized controlled trial. PLoS One 2023 Sep 5;18(9):e0290328 [FREE Full text] [doi: 10.1371/journal.pone.0290328] [Medline: 37669289]

XSL•FO
RenderX

75. Murphy JK, Chau LW, Nguyen VC, Minas H, Anh DV, O'Neil J. An integrated knowledge translation (iKT) approach to advancing community-based depression care in Vietnam: lessons from an ongoing research-policy collaboration. BMC Health Serv Res 2024 Jan 27;24(1):142 [FREE Full text] [doi: 10.1186/s12913-023-10518-3] [Medline: 38279141]

Abbreviations

ASW: Antidepressant Skills Workbook
CBT: cognitive behavioral therapy
CHC: Commune Health Center
MOLISA: Ministry of Labour, Invalids, and Social Affairs
PHAD: Institute of Population, Health, and Development
PHQ-9: 9-item Patient Health Questionnaire
RCT: randomized controlled trial
SPC: Social Protection Center
SSM: supported self-management

Edited by A Mavragani; submitted 01.11.24; peer-reviewed by KW Tay, M Pulier; comments to author 05.05.25; revised version received 27.05.25; accepted 09.06.25; published 30.06.25.

Please cite as:
Chau LW, Murphy JK, Nguyen VC, Tran HN, Minas H, Lam RW, Hayashi K, Nguyen TTX, Krebs E, O'Neil J
Challenges and Mitigation Strategies in the Development and Feasibility Assessment of a Digital Mental Health Intervention for Depression (VMood): Mixed Methods Feasibility Study
JMIR Form Res 2025;9:e68297
URL: https://formative.jmir.org/2025/1/e68297
doi:10.2196/68297
PMID:

©Leena W Chau, Jill K Murphy, Vu Cong Nguyen, Hai Nhu Tran, Harry Minas, Raymond W Lam, Kanna Hayashi, Thi Thanh Xuan Nguyen, Emanuel Krebs, John O'Neil. Originally published in JMIR Formative Research (https://formative.jmir.org), 30.06.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Assessing the Fit of a Digitally Delivered National Diabetes Prevention Program Among Rural Living Adults: Qualitative Study

Gerit Wagner[1], MPH; Lyndsie M Koon[2], PhD; Patricia Smith[3], MS; Kameron B Suire[4], PhD; Mary Hastert[1], PhD; Joseph E Donnelly[1], EdD; Melissa D Olfert[5], DrPH; Paul Estabrooks[6], PhD; Anna M Gorczyca[1], PhD

[1]Division of Physical Activity and Weight Management, Department of Internal Medicine, University of Kansas Medical Center, Kansas City, KS, United States

[2]Research and Training Center on Independent Living, University of Kansas, Lawrence, KS, United States

[3]CU Anschutz Health and Wellness Center, University of Colorado Anschutz Medical Campus, Aurora, CO, United States

[4]Department of Kinesiology, Berry College, Rome, GA, United States

[5]School of Agriculture and Food Systems, West Virginia University, Morgantown, WV, United States

[6]Department of Health & Kinesiology, College of Health, University of Utah, Salt Lake City, UT, United States

Corresponding Author:
Anna M Gorczyca, PhD
Division of Physical Activity and Weight Management
Department of Internal Medicine
University of Kansas Medical Center
3901 Rainbow Boulevard
Suite 150, Mailstop 1073
Kansas City, KS, 66160
United States
Phone: 1 9135889077
Email: agorczyca@kumc.edu

Abstract

Background: Rural living adults are disproportionately affected by type 2 diabetes compared to their urban counterparts. The Centers for Disease Control and Prevention's National Diabetes Prevention Program (National DPP) is an evidence-based intervention that reduces the risk of type 2 diabetes through increased physical activity and modest weight loss, but overall reach remains limited, specifically in rural communities.

Objective: This qualitative study aimed to examine the fit of the National DPP delivered digitally using Zoom or Facebook to rural living adults at risk for type 2 diabetes.

Methods: Focus group scripts assessed the characteristics and perceptions of rural adults at risk for type 2 diabetes, infrastructure supports for implementation and sustainability, and external factors that could influence program fit. A reflexive thematic analysis was conducted separately on coded transcripts for each focus group. Themes were then deductively linked to the Practical, Robust Implementation and Sustainability Model domains.

Results: Two focus groups were conducted with 14 participants after participating in the National DPP for 6 months, delivered through Zoom (n=9) or Facebook (n=5). Participants highlighted positive relationships between Practical, Robust Implementation and Sustainability Model constructs related to participant characteristics (ie, value of health improvements, weight loss, and reduced medication dependence as primary motivators) and perceptions of compatibility (ie, content alignment with participant needs) as well as infrastructure (ie, digital platforms provided better access) with program success in reach and engagement. Conversely, both formats were negatively impacted by interruptions in internet connectivity. External factors, such as referral pathways from local health care providers, could improve program reach. When considering differences between implementation infrastructure, Zoom facilitated greater social engagement and accountability compared to Facebook.

Conclusions: This study identified contextual factors influencing the fit of digitally delivering the National DPP to rural living adults, including opportunities for using existing connections and health motivations to help improve acceptability, while tailoring curriculum, modality, and technology may improve appropriateness for rural populations.

Trial Registration: ClinicalTrials.gov NCT05387434; https://clinicaltrials.gov/study/NCT05387434

XSL•FO
RenderX

(JMIR Form Res 2025;9:e70406) doi:10.2196/70406

KEYWORDS

diabetes prevention program; rural; digital delivery; PRISM; qualitative; Practical, Robust Implementation and Sustainability Model

Introduction

Type 2 diabetes represents a significant public health concern, impacting the quality of life of those having the disease and accounting for US $412.9 billion (about US $1300 per person) through direct and indirect health care costs annually in the United States [1]. The prevalence of type 2 diabetes and prediabetes is 11.6% and 38% of the US population, respectively, and continues to rise [2,3]. To combat the increased prevalence of these conditions, in 2010, Congress authorized the Centers for Disease Control and Prevention (CDC) to implement the National Diabetes Prevention Program (National DPP), a large-scale, lower-cost version of the original 2002 Diabetes Prevention Program (DPP) [4]. The DPP is a behavioral weight loss program that emphasizes calorie reduction and increased physical activity (PA) to achieve a goal of 5%-7% weight loss over 6 months. The original DPP demonstrated a 58% decrease in the incidence of type 2 diabetes, as well as a 71% decrease in type 2 diabetes incidence in participants older than 60 years.

The reach of the National DPP, the largest translation of the original DPP, remains limited. As of 2022, there were only 2098 National DPP-recognized providers nationwide, and merely 3% of individuals diagnosed with prediabetes have participated in the program [5]. Substantial implementation barriers, including the inability to recruit high-risk populations coupled with the lack of sufficient financial reimbursement to build and sustain the program, have been reported [6]. Furthermore, access to and delivery of the National DPP is limited within communities where type 2 diabetes is more prevalent, including communities of lower socioeconomic status, rural counties [7], as well as communities of color, including Latinx and Indigenous communities, as well as older adults. Equal provision of programs such as the National DPP may mitigate rural or urban health disparities and address the disproportion in diabetes risk between different racial or ethnic and demographically located groups.

To address the limited reach of the program, the CDC currently recognizes delivery of the National DPP through several modalities including in-person, online (eg, asynchronous classroom with Lifestyle Coach interaction), distance learning (eg, telephone, Zoom [Zoom Video Communications], Skype [Microsoft Corp], FaceTime [Apple Inc], etc), or combinations of in-person, distance learning or online delivery [8]. Digital delivery of the National DPP, including online and distance learning platforms, eliminates the time and expense associated with lengthy travel to attend on-site programs, negating a significant barrier to participation for adults living in rural and challenging terrain [9]. Moreover, digital delivery of the National DPP may improve participation and retention of vulnerable populations [10,11], which are both important predictors of weight loss and reduced incidence of type 2 diabetes [6,12-15].

The US Cooperative Extension System (Extension) has been identified as a potential vehicle for the delivery of this program to rural communities [16]. Extension is a partnership between the US Department of Agriculture, land-grant universities, state, and local governments to improve research-based knowledge and application to communities. Extension uses nutrition and family consumer science agents in approximately 3200 offices across the United States as designated by the land-grant universities within each state. As of 2023, Extension delivered 20 National DPP programs across 16 states with a coordinated interest to expand access to the program [17]. Research indicates that distance learning delivery of the National DPP in rural areas by Extension agents is feasible and effective in achieving the program goals [18]. However, less is known about multilevel contextual barriers and facilitators of digital delivery of the National DPP in rural living adults. The Practical, Robust Implementation and Sustainability Model (PRISM) implementation framework was developed to improve the translation of evidence-based interventions into practice by identifying key contextual factors that are associated with program outcomes [19,20]. PRISM domains include (1) the intervention (organization and recipient perspectives), (2) the recipient (organizational and participant level characteristics), (3) implementation and sustainability infrastructure, and (4) external environment. The objective of this qualitative study was to use PRISM as the conceptual framework to examine the fit of the National DPP delivered digitally using Zoom or Facebook (Meta Platforms, Inc) to rural living adults at risk for type 2 diabetes.

Methods

Study Design

Focus groups were conducted in this cross-sectional qualitative study to gather information on a digitally delivered National DPP lifestyle change program for rural living residents by distance learning, where Extension personnel delivered live sessions over Zoom or online, asynchronously by a university research staff member via a private Facebook group (NCT 05387434). A comparison of intervention components in the pilot trial is provided in Table 1. Detailed methodology, feasibility, and effectiveness results of the pilot trial have been reported elsewhere [18]. Focus groups were chosen as the National DPP emphasizes group participation and social support [21]. Group dynamics encourage discussion and spontaneity that may not emerge in other methodologies, such as individual interviews. Further, focus group discussion helps highlight cultural norms across various groups that can help inform future research in rural-living adults with prediabetes [22].

XSL•FO
RenderX

Table 1. Overview of study components for a digitally delivered National Diabetes Prevention Program among rural living adults.

	Zoom	Facebook
Lifestyle coach	Extension agent	University research staff
Delivery modality	Zoom	Facebook
Curriculum	2016 PreventT2	2016 PreventT2
Contact	60-minute Zoom meeting	3 posts and direct messaging
Frequency	Weekly	Weekly
Duration	6 months	6 months
Self-monitoring device	Wearable and scale	Wearable and scale
Self-monitoring app	Fitbit	Fitbit

Ethical Considerations

All participants provided informed consent to participate in the original study through REDCap (Research Electronic Data Capture; Vanderbilt University) electronic consent and were reconsented to participate in the focus group. This pilot study was approved by the University of Kansas Medical Center institutional review board and registered on ClinicalTrials.gov (NCT05387434). Study personnel implemented appropriate safeguards to prevent unauthorized use or disclosure of personal health information and implemented safeguards to protect the confidentiality, integrity, and availability of protected health information. Participants were pseudonymized through the transcription service to reveal only the gender identity or role of the speaker, that is, “F” for a female participant, “M” for a male participant, or “Q” for the interviewer. In the original pilot study, participants were compensated US $20 for time and travel to attend each of the 3 outcome testing visits (baseline, 3 months, and 6 months) and allowed to keep the Fitbit (Google LLC) and wireless scales used for self-monitoring in the pilot study. No compensation was provided for participation in the focus groups.

Recruitment

In the original pilot study, members of the research team collaborated with participating Extension offices to recruit individuals via traditional media (ie, newspapers), social media, Extension-managed listservs, as well as distributing flyers at local businesses and physicians’ offices. Potential participants were directed to this study’s website, where they filled out an initial eligibility questionnaire. Once deemed preliminarily eligible, research staff met with potential participants via phone or Zoom to review and discuss the informed consent. Participation in the pilot study was required for participation in the focus group. Participants were offered the option to take part in focus groups about the acceptability and appropriateness of the National DPP at the end of the 6-month study via email. Purposive, convenient sampling was used with a minimum of 5 participants per focus group.

Reflexivity and Positionality

A crucial aspect of performing a reflexive thematic analysis is to value the subjectivity and background of the researchers [23]. With regards to the research question, the senior author, AMG, has implemented lifestyle interventions in rural populations for the past 9 years [24,25]. AMG is an experienced scientist in lifestyle behavior interventions, and combined with this practical lived experience, while conducting both focus groups to facilitate participants’ identifying supports and barriers in this formative work. Research participants were informed of the experience of the senior researcher’s background, which helped encourage participant responses in the focus group data collection. This may have resulted in a level of translation that would be representative of other similar rural populations.

Data Collection

Baseline participant characteristics were assessed in March and May 2021. Study data were collected and managed using REDCap tools hosted at the University of Kansas Medical Center. All participants provided informed consent via REDCap. Focus groups were conducted by delivery modality (5-9 participants/group) via Zoom. Focus groups took place in August and October 2021. For Zoom participants, the focus group was held at a typical intervention meeting time to minimize scheduling conflicts. All Zoom participants who reconsented took part in the focus groups (9/9; Figure 1). As the Facebook group operated asynchronously without a set meeting time, the research team distributed a scheduling poll to determine availability and scheduled the Facebook focus group when the maximum number of participants were available. Of the 12 Facebook participants who reconsented, 5 (42%) participated in the focus groups. The focus group questions are presented in Textbox 1. Each focus group lasted approximately 60 minutes and was recorded and then transcribed verbatim by a third-party transcription service.

XSL•FO
RenderX

Figure 1. Flowchart of participants from the pilot study engaging in focus groups and included in the analysis, assessing the fit of a digitally delivered National Diabetes Prevention Program. DPP: Diabetes Prevention Program.

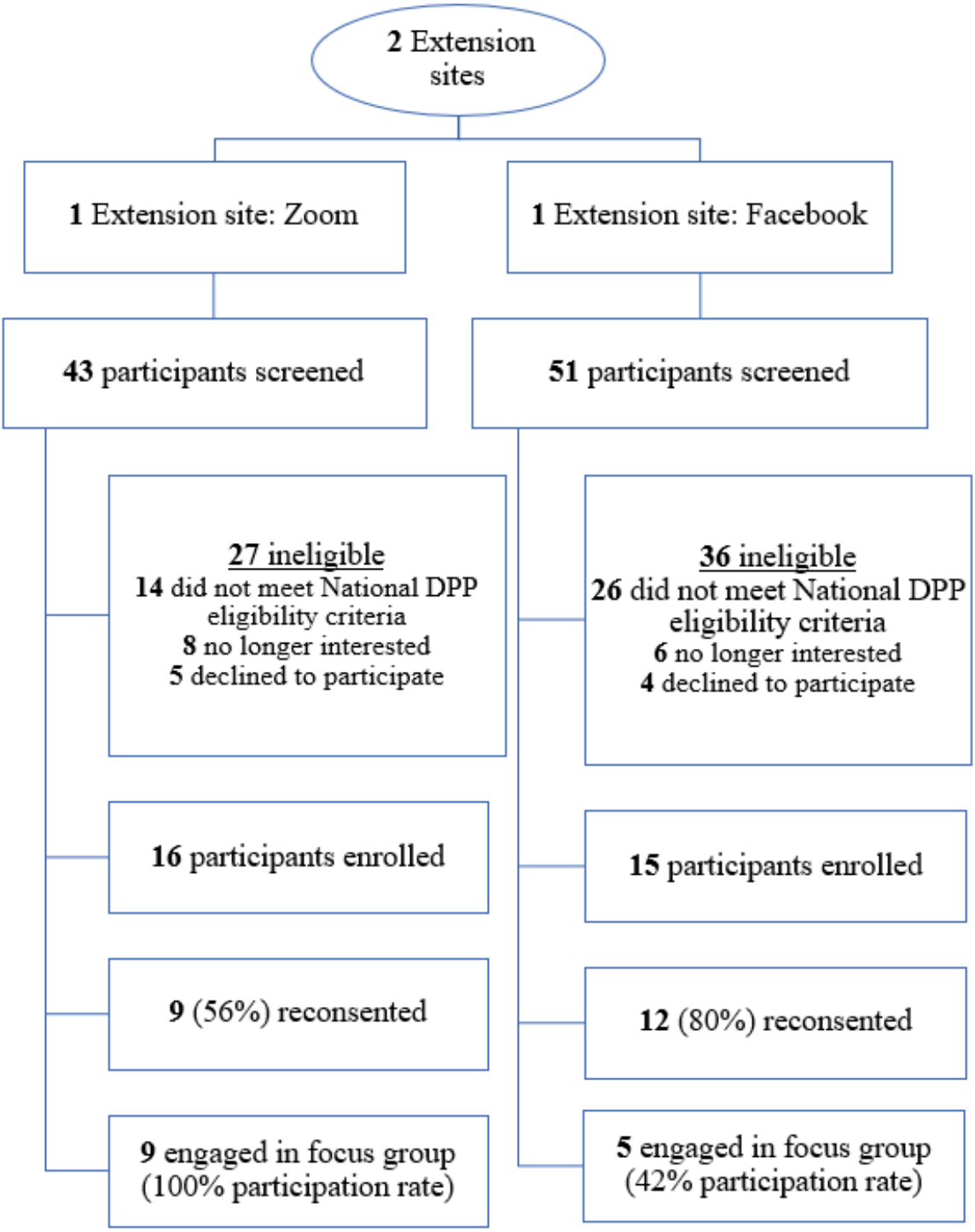

XSL•FO
RenderX

Textbox 1. Focus group guide to assess the fit of a digitally delivered national Diabetes Prevention Program among rural living adults.

Enrollment

- Why did you participate in this program?
- What information specific to rural communities could have been included?
- How could the intervention or intervention materials have been tailored specifically for rural communities?

Dietary changes

- What types of dietary changes did you make?
- What helped you or could have helped you make dietary changes?
- What were some barriers to changing your diet?
- What would help you maintain your healthier eating habits?

Physical activity

- How did you get in the prescribed physical activity?
- Where did you do the physical activity?
- What types of activities did you do?
- What barriers did you encounter to increase your physical activity?

Delivery modality

- How did you feel about using Facebook as the primary platform for the program? (Facebook only)
- How did you feel about using Zoom as the primary platform for the program? (Zoom only)
- Did you have any issues with internet connectivity?
- Did you feel that you could connect with other group members virtually?

Technology

- Do you see yourself interacting with other group members using the Fitbit community option?
- Do you see yourself tracking your diet using a smartphone or iPad application?
- Tell me about your experience using the provided scale

Other components

- Would you like to tell us anything else about your experience with the Diabetes Prevention Program?

Data Analysis Strategy

While the analysis was iterative in nature, it was conducted in six phases: (1) familiarization with the dataset; (2) coding; (3) generating initial themes; (4) developing and reviewing themes; (5) defining, refining, and naming themes; and (6) reporting [26]. For familiarization with the dataset, 3 researchers trained in qualitative analysis independently read through the focus group transcripts and familiarized themselves with participant responses, then constructed an initial codebook using MAXQDA software (VERBI Software, 2020). The codebook was refined through a second review to create a final codebook. Codes that addressed the primary aims were reviewed to develop themes. Themes, defined as patterns in the data that provide description and meaning to a qualitative dataset [27,28], were generated independently by the 3 initial researchers based on the primary aims of this study. Themes and subthemes were then discussed among all 3 researchers and were refined and defined until a consensus was met to align with PRISM domains. The COREQ (Consolidated Criteria for Reporting Qualitative Research) guidelines were followed (Multimedia Appendix 1) [29].

Results

Overview

Among the 31 participants enrolled in the pilot study (Zoom n=16; Facebook n=15), 14 (Zoom n=9; Facebook n=5) agreed to participate in the focus groups (Table 2). Notably, 93% (n=13) identified as female, 13% (n=2) identified as non-White and Hispanic or Latino, and 64% (n=9) possessed a college degree.

XSL•FO
RenderX

Table 2. Demographic characteristics of rural-living adults participating in focus groups to assess the fit of a digitally delivered National Diabetes Prevention Program.

	All pilot participants (n=31)	Focus group participants (n=14)	Zoom (n=9)	Facebook (n=5)
Age (years), mean (SD)	55.1 (12.8)	60.5 (9.3)	58.2 (10.6)	64.6 (3.8)
Sex, n (%)				
Male	6 (19)	1 (7)	0 (0)	1 (20)
Female	25 (81)	13 (93)	9 (100)	4 (80)
Racial or ethnic minority[a], n (%)	5 (16)	2 (13)	2 (20)	0 (0)
BMI (kg/m^2), mean (SD)	36.4 (7.8)	36.1 (9.4)	38.2 (5.4)	32.2 (13.1)
Education, n (%)				
High school graduate	2 (7)	2 (14)	2 (22)	0 (0)
Some college	5 (16)	3 (22)	2 (22)	1 (20)
Bachelor's degree	14 (45)	7 (50)	3 (33)	4 (80)
Graduate degree	6 (19)	2 (14)	2 (22)	0 (0)
Income (US $), n (%)				
<$40,000	6 (19)	4 (29)	3 (33)	1 (20)
$40,000-$79,999	9 (29)	4 (29)	2 (22)	2 (40)
>$80,000	15 (48)	6 (42)	4 (44)	2 (40)

[a]Non-White, Hispanic, Latino, other, or unknown ethnicity.

The finalized themes and subthemes were evaluated to describe the experience with the intervention for each group (Zoom and Facebook) and to compare between groups to provide a rich description of similarities and differences. Responses were categorized into four themes based on PRISM: (1) participant characteristics, (2) perceptions of program compatibility, (3) program delivery infrastructure, and (4) external factors influencing program success. The final set of themes and subthemes is presented in Table 3 with illustrative quotes.

XSL•FO
RenderX

Table 3. Themes and selected supporting quotes from rural living adults participating in focus groups to assess the fit of a digitally delivered National Diabetes Prevention Program.

I. Participant characteristics	
Weight loss was a common goal	• For me it was losing weight as a side which is a great one, but I do have several family members that have diabetes, and it caught my eye because by getting healthy, I can maybe possibly prevent diabetes. [Female, Zoom]
Improve their laboratories or health numbers	• I had been to the doctor and... my sugar was high and so she wanted to do an A1C check. And when she checked, my A1C was a little high. So she wasn't concerned, but I was. So I saw this advertised in the paper and decided that it was a way to be proactive about staying on top of that. [Female, Facebook]
Desire to reduce or eliminate dependence on prescription medications	• I was already on metformin in the morning and metformin at night when I got this. And now I'm down to one metformin at night and my blood sugar now is at – my A1c is 5.4. So I'm hoping that if I continue on this, I can get off of it completely. [Female, Zoom]
II. Perceptions of program compatibility	
More information on how to eat healthier	• I still think – I don't know. I was the one that said recipes were hard. The whole time, recipes were hard to find. You know?... So it'd be really cool if there was some kind of recipe book that went along with this, you know? [Female, Zoom]
Challenges of integrating the program's dietary recommendations into their daily lives	• I think when I looked at how long it takes to do the food prep and then I look at what's available in the grocery store, I kind of understand why people just grab the plastic sack and open it. [Female, Facebook]
Differences in accessibility and convenience of delivery modality	• For me personally Zoom has been so nice because I can still take care of my family and there were nights that we had Girl Scout meetings and I didn't have to miss a meeting or – I could do it wherever, whenever. Right now I'm sitting outside with kids. [Female, Zoom]
III. Program delivery infrastructure	
Recruited through a variety of mediums	• I believe that I originally found the ad for it in the newspaper. [Female, Zoom]
Health care providers could be better used in the recruitment process	• Before I signed up for this, when I got the email, I went and talked to [my doctor]... I said, Have you ever heard about this? And he pointed and there was a flyer on his cabinet... And he goes, Is it this one? And I kind of looked and I said, Yeah, I'm pretty sure it is. And he goes, No because I've never had anybody go through that. And so I guess I'm thinking here I am somebody that is already on metformin. Why wouldn't you have promoted that with me? So I think you need to work with the doctors in rural areas more. [Female, Zoom]
Differences in digital literacy impact self-monitoring behaviors	• I'm not that good with the cell phone. [Female, Zoom]
Digital self-monitoring apps increased participant accountability and motivation	• Having to actually write down and record how many fruits and vegetables you eat is always eye-opening. I mean, I think I eat a lot of fruits and vegetables. But I eat more today than I did six months ago because I had to write them down and I had to account for each one. And if I eat more, then my family eats more too. [Female, Facebook]
Use of self-monitoring apps in their daily lives	• I used MyFitnessPal because I've used it before and I log – I have logged every single thing I've eaten for the last six months I guess, and it communicates with Fitbit. [Male, Facebook]
Detailed knowledge of physical activity habits	• And having the accountability of how many steps. Before, I used my phone. Well, if you don't take your phone everywhere, then you don't get all the steps. And so this is, again, a tool that can maybe make you more accountable. And I want to look into some of the other features that the Fitbit has that I haven't used like how well are you sleeping and some of that kind of stuff. [Female, Facebook]
Social connections and accountability	• Zoom: I did like the contact of seeing everybody. It kind of felt like you were more connected to everybody. [Female, Zoom] • Facebook: Yes, I posted [to Facebook], but I don't think I felt there was a real connection. [Female, Facebook]
IV. External factors influencing program success	
Various barriers to physical activity	• I found excuses it was really hot. I just don't adjust to the heat as well anymore, and I know exercise helps that, but that and then we had mosquitoes this year like crazy, and that literally put on so much spray every day. [Female, Facebook]

XSL•FO
RenderX

I. Participant characteristics	
Strategies to overcome physical activity barriers	• I got my bicycle in June and that's really been good for me mentally because I can just get on it and go by myself and just zone out. So that was good for me. And then I did get the thing to hook it to this winter so I can pedal and watch TV at the same time or read a book, whatever, so. [Female, Zoom]
Technical challenges	• Zoom: If [Lifestyle Coach] was at home and the wind and stuff, there was some internet issues that way. [Female, Zoom] • Facebook: I don't know in rural areas using the technology platform. I know that the smaller the town, maybe the less comfortable they feel using it. Here in Salina, I thought it was great. But it depends on who we're trying to reach. I mean, I'm sure there's lots of people that the technology is not a turnoff or not a downside. [Female, Facebook]

Participant Characteristics

Overview

Participants noted a variety of reasons for joining this study, characterizing many of their specific health needs and motives. Motives to participate in the program were similar across both groups and included weight loss, improving health metrics such as cholesterol and glycated hemoglobin (HbA_{1c}), reducing medication dependence, and reducing the risk of disease progression (ie, type 2 diabetes).

Weight Loss Was a Common Goal Among Participants

Weight loss was often linked to their broader health objectives. Many participants recognized that losing weight could help them improve their overall health, reduce the risk of developing chronic diseases, and possibly eliminate the need for certain medications. They viewed the National DPP as an opportunity to achieve sustainable weight loss, which would in turn lead to better health outcomes.

Participants Wanted to Improve Their Laboratory or Health Numbers

Many participants were motivated to enroll in the program to improve their health metrics, particularly laboratory results such as HbA_{1c} levels and cholesterol. They expressed concern about their current health status and wanted to take proactive steps to avoid more serious health issues, such as a diabetes diagnosis. Participants mentioned how their doctors had warned them about rising HbA_{1c} levels, prompting them to join the program to better manage their blood glucose.

Participants Were Driven by the Desire to Reduce or Eliminate Their Dependence on Prescription Medications

Some were already taking medication for conditions such as high blood pressure or diabetes, and hoped that by participating in the program, they could lower their dosage or stop taking the medication altogether. This goal of reducing medication use was a significant motivator for many, as they viewed it as a marker of improved health and well-being.

Perceptions of Program Compatibility

Overview

Participants discussed how the content and delivery modalities of the program aligned with the needs of rural living adults at risk for diabetes, including how the content of CDC's Prevent T2 curriculum met their stated needs and opportunities to further strengthen program content for future interventions. Participants from both groups found the curriculum valuable but desired more structure and guidance with topics such as recipes and meal planning. Further, participants noted differences in how the different delivery modalities accommodated their schedules, lifestyles, and social needs.

Participants Desired More Information on How to Eat Healthier

While the program provided some guidance, participants felt that more detailed and structured information would have been beneficial. This included more recipes, guidance on managing portion sizes, and strategies for incorporating healthier foods into their daily routines. Some participants expressed the need for more practical resources, such as a recipe book or meal planning guides, to help them incorporate more fruits and vegetables into their diets. The lack of clarity around certain dietary concepts, such as understanding carbohydrates, was also highlighted as a barrier to making informed food choices.

Participants Expressed Challenges Integrating the Dietary Recommendations Into Their Daily Lives

Many found it difficult to make healthier food choices consistently, especially when faced with limited time or access to fresh ingredients. The shift toward fresh produce and away from processed foods was a significant change for many, requiring a considerable adjustment in shopping and meal preparation, and budget. Participants also noted that dietary guidance was sometimes at odds with family needs. One challenge mentioned was the responsibility as the primary household food preparer, needing to satisfy the dietary needs and preferences of family members. Others noted the inconvenience of preparing healthy foods as a challenge in long-term adherence to dietary changes. Furthermore, participants highlighted how eating out and social gatherings conflicted with the program's recommendations, as participants found it difficult to maintain their healthy eating habits in these settings.

Participants Noted Differences in Accessibility and Convenience of Delivery Modality

Zoom

Participants generally found Zoom a highly convenient platform for participating in the program. It allowed them to join sessions

XSL•FO
RenderX

from home, making it easier to integrate the meetings into their daily schedules. The ability to participate from anywhere, such as while caring for children or handling other responsibilities, was particularly appreciated. However, a few participants mentioned occasional internet connectivity issues that sometimes disrupted the experience.

Facebook

Participants generally found the Facebook group flexible and convenient, allowing them to engage with the program at their own pace. The ability to access materials and participate in discussions asynchronously was appreciated, particularly by those with busy schedules. However, some participants missed the immediacy of real-time interaction, which they expressed would have enhanced their engagement and accountability.

Program Delivery Infrastructure

Overview

Participants discussed their experiences with the program's infrastructure, including perceptions of the recruitment methods used and opportunities for improving the reach of recruitment efforts. Further, participants discussed how the technological infrastructure, including provided hardware and software apps, influenced their health behaviors and social interactions, as well as how this same technology presented unintended barriers to their participation in the program.

Participants Were Recruited Through a Variety of Media

Participants were recruited for the program through various channels, including newspaper advertisements, newsletters from local Extension services, word of mouth, and emails. Some participants mentioned discovering the program by chance, such as reading about it in a local paper or hearing about it from a neighbor. Despite the variety of recruitment methods, some felt the outreach could have been more extensive to reach a wider audience.

Health Care Providers Could Be Better Used in the Recruitment Process

Some participants expressed frustration that health care providers were not as involved in the recruitment process as they could have been. Several participants noted that although they discussed their health issues, such as rising HbA_{1c} levels, with their doctors, the program was not recommended to them directly. In some cases, participants found out about the program independently and mentioned it to their doctors afterward, who were then supportive but had not been proactive in suggesting the program. The consensus was that health care providers could play a more active role in promoting such programs, given their professional authority and influence on patients' health decisions.

Differences in Digital Literacy Impacted Self-Monitoring Behaviors

Some participants expressed confusion about using the apps for intended self-monitoring purposes, while others mentioned how they could tailor the apps to track specific activities or set personalized goals, including step counts and protein or calorie goals, making the tools more relevant to their unique health goals.

Digital Self-Monitoring Apps Increased Participant Accountability and Motivation

Participants mentioned the Fitbit and MyFitnessPal apps improved their ability to track steps, monitor progress, and receive reminders that kept them motivated to meet their daily PA goals. The real-time feedback provided by these tools acted as a constant motivator, encouraging participants to stay active and adhere to their exercise routines.

Participants Discussed the Use of Self-Monitoring Apps in Their Daily Lives

Many participants appreciated how straightforward it was to log their activities and monitor their progress. The synchronization between the Fitbit wearable device, the Fitbit smartphone app, and MyFitnessPal was particularly valued, as it allowed participants to track both PA and dietary intake in 1 place, simplifying the self-monitoring process.

Participants Valued the Detailed Insights Apps Provided Into Their PA

Features such as step counting, heart rate monitoring, and calorie expenditure allowed participants to gain a deeper understanding of their personal behaviors. This detailed tracking helped participants set more specific goals and measure their progress over time, which was crucial for maintaining their commitment to the program.

Social Connections and Accountability Felt Different Among Participants

Zoom

The visual and real-time interaction on Zoom significantly contributed to a sense of engagement among participants. The requirement to join live sessions and interact with others encouraged participants to stay on track with the program. Many noted that seeing each other's faces and hearing voices fostered a supportive environment. The structured environment of scheduled meetings also provided a sense of discipline, which many participants found beneficial for their progress in the program.

Facebook

Engagement in the Facebook group varied, with some participants actively contributing while others were more passive by observing. The lack of real-time meetings or structured schedules made it easier for some participants to disengage. Nonetheless, those who did participate valued the ongoing support and found the platform to be a useful tool for maintaining accountability, especially when paired with tools such as Fitbit for tracking PA. However, the absence of regular, live interaction was seen as a barrier to sustained engagement and accountability for some. Some participants expressed discomfort with posting publicly within the private Facebook group, leading to less frequent and less personal interactions. Others noted that while the group allowed for some level of connection, it did not foster the same sense of community and

XSL•FO
RenderX

support that could have been achieved through in-person or distance learning modalities.

External Factors Influencing Program Success

Overview

Participants discussed how external influences shaped their ability to successfully engage with the program. Environmental barriers, modality-specific challenges, and other logistical constraints created barriers to sustained behavior change.

Various Barriers to PA Were Identified

Barriers included environmental factors such as heat, pests such as mosquitoes, and limited access to exercise facilities. Some participants mentioned that they did not feel comfortable exercising in public places such as gyms, while others found it difficult to maintain a regular exercise routine due to busy schedules or physical limitations, such as recovering from surgery. These barriers often hindered participants' ability to engage in consistent PA.

Participants Strategized to Overcome External PA Barriers

Many adapted their routines to fit their circumstances, such as walking indoors during hot weather or using home exercise equipment. Some participants took creative approaches, such as walking in circles around their house or using local parks and tracks for outdoor exercise.

Different Technical Challenges Experienced by Delivery Modality

Zoom

While participants generally adapted well to the Zoom platform, a few did experience technical challenges, primarily related to internet connectivity. These issues were often minor and were typically resolved quickly, allowing participants to continue engaging with the program. Despite these occasional technical difficulties, Zoom was seen as a manageable and effective platform for delivering the intervention.

Facebook

Minimal technical challenges were reported in using Facebook for the program. Most participants found the platform familiar and easy to navigate. However, the asynchronous nature of the interaction sometimes led to delayed responses and a less dynamic exchange of ideas, which some participants found limiting.

Discussion

Principal Findings

This study provides important insights into the appropriateness of program recruitment, curriculum, digital delivery, and technology usage for self-monitoring in the context of the National DPP. For acceptability, rural living adults thought rural health care providers could play a larger role in program awareness and referrals. For appropriateness, there was a desire for more structured nutrition digital resources in the curriculum. Distance learning using Zoom fostered group communication and accountability, while online delivery using Facebook hindered participant social interaction. Self-monitoring weight and PA using wireless scales and wearable fitness devices helped reinforce behavior change and increase accountability to the program goals. Overall, results indicate considerations for the implementation of the National DPP using digital delivery strategies for rural living adults.

Participants reported that acceptability may be improved through greater health care providers' awareness of the National DPP to enable a referral process for patients at risk for type 2 diabetes. Low rates of health care provider referrals have been a longstanding barrier to implementation of the National DPP, which is even more detrimental in rural communities [30]. Active recruitment strategies, which use collaborations with the medical community for participant referral, have the potential to improve National DPP recruitment [8] beyond that of traditional passive recruitment, which relies on potential participants to initiate contact with the National DPP. Traditional passive recruitment methods often lead to a population comprised predominantly of highly educated women [31] and lacking racial diversity [32]. Leveraging active recruitment methods, including point-of-care or population health management, may lead to greater participation in general and that of underrepresented populations in diabetes prevention trials [33]. Previous research has found that providers using the electronic health record were more likely to refer to the National DPP, which may provide a starting point for community organizations initiating conversations with clinical entities for recruitment [34]. Further, participants did not note that the program being delivered by an Extension agent negatively impacted program acceptability, suggesting that program quality remains consistent when delivered by Extension. However, the absence of explicit questions regarding the impact of the facilitating group's lifestyle coach may limit the strength of this conclusion.

Despite the National DPP's overall focus on diet quality, participants expressed a desire for structured guidance surrounding nutrition-related changes. The 2021 PreventT2 curriculum underwent updates emphasizing "whole foods and healthy eating patterns," promoting "greater flexibility in allowing participants to tailor approaches to healthy eating and weight loss support" [35]. These updates align with professional consensus advocating for the individualization of eating plans and patterns to enhance adherence to healthy behaviors [36]. Despite the focus on overall dietary quality over specific benchmarks, focus group participants voiced a need for additional guidance and specific resources to facilitate dietary and food selection changes. However, the National DPP does not mandate that lifestyle coaches possess nutrition credentials, limiting their capacity to offer evidence-based clarifications or resources to support these changes. Recent evidence suggests the PreventT2 curriculum does not support meaningful changes in diet quality [37]. This presents a dilemma: participants seek more structured information, while the professional consensus suggests that rigid guidelines are ineffective in fostering long-term behavior change, a conflict difficult to resolve within the National DPP framework [36,38].

XSL•FO
RenderX

Our results suggest that differences in effectiveness between distance learning and online may, in part, be due to increased social engagement and accountability with the distance learning modality. A recent analysis of all National DPP participants through December 31, 2019, found greater weight loss for distance learning (4.7%) compared to online (2.6%) [39]. However, online, asynchronous delivery of the National DPP may be associated with lower costs than other modalities [40]. Using digital delivery, including both distance learning and online modes, may overcome barriers specific to rural communities in accessing effective lifestyle interventions, such as the lower availability of programs, transportation, and other caretaking duties. Organizations should weigh the increased favorability and higher rate of weight loss associated with distance learning delivery against the convenience and accessibility that online delivery can provide.

There was high acceptability in using the wireless scale and wearable fitness tracker to self-monitor, which helped participants' accountability and motivation to engage in PA and monitor body weight. Self-monitoring weight and PA are cornerstone health behaviors targeted in the National DPP. Wearable devices may empower participants to engage in PA; however, the efficacy of the device alone as a behavioral strategy is unknown [41,42]. This study provided both the scale and fitness trackers to participants, which may not be financially appropriate for all programs. Furthermore, there are sociodemographic disparities in wearable ownership that skew toward younger, educated adults living in urban areas [43]. Stakeholders interested in increasing the uptake of their National DPP should consider the provision of scales and fitness trackers, as rural adults in this study felt they were helpful.

The results of this study align with results from various demographic groups within the United States as well as DPPs present in other countries. In the United States, rural communities can exhibit substantial heterogeneity. This diversity in rural-living adults presents itself as a barrier in designing culturally appropriate materials for rural communities, as needs may differ with geographic location and sociodemographic characteristics [44]. However, the sentiments expressed by the participants in this study have been echoed across diverse demographic groups' participation in the National DPP as well as other DPPs internationally. Interventions conducted in California note the positive impact virtual platforms may offer in terms of increasing program reach [45,46]; however, concerns regarding the potentially negative impact of digital platforms on participant engagement and peer support remain. Additionally, a faith-based community intervention among African American participants noted the importance of self-monitoring through weight tracking and similar desires for more in-depth program materials, similar to the interest in expanded nutrition resources observed in our study [47]. Internationally, the effectiveness of DPPs is well documented [48], and a recent systematic review including studies from the United States, England, and Australia examining health care workers' perspectives on barriers and facilitators for referrals to DPPs matched the sentiments of participants in our study: there is a need for DPPs to further strengthen the role of health care providers in referring participants to these programs [49]. Literature regarding perceptions of individuals participating in digital DPPs, specifically, however, is less abundant. International digital DPPs, such as the Irish National Diabetes Prevention Programme and the National Health Service Healthier You*:* Digital Diabetes Prevention Programme in England, note findings similar to those noted within our study [50,51]. Participants in both programs have highlighted the role of health care providers in recruiting participants and have emphasized the need for greater provider engagement to encourage participation and program uptake. Further, participants acknowledged that even in digital settings, personalized support from lifestyle coaches was important for accountability [52]. Self-monitoring tools, such as tracking food intake, were also noted to be of importance for some participants. These findings underscore the shared challenges and opportunities in diabetes prevention efforts across diverse populations and settings.

Limitations

While our study provides important insights on the acceptability and appropriateness of digital delivery of the National DPP to rural adults, it is not without limitations. Given the increasing diversity of rural populations, this study may not reflect the experiences of other geographic areas due to the relative sample homogeneity and small sample (n=14). Second, the planned duration of the feasibility study was originally only 6 months. While the National DPP is traditionally a 12-month program, this study may not represent the perspective of individuals who experience the full year-long program. However, as the National DPP requires at least 16 sessions conducted within the first 6 months of the program and 6 sessions in the final 6 months, participants were exposed to most of the curriculum at the time of the focus groups [8]. Finally, conducting focus groups required participants to reconsent to the research process, which resulted in a relatively low response rate in the Facebook group. Therefore, results do not fully represent all online delivery modalities of the National DPP and are specific to Facebook.

Despite these limitations, this study fills an important gap in the perception of digital delivery to improve the reach of the National DPP to rural adults using an evidence-based implementation framework (PRISM). Furthermore, the use of a reflexive thematic analysis enhanced the ability to analyze findings within and between groups, specifically the acceptability of distance learning and online delivery modalities. Leveraging the perspectives of rural living adults provides robust information that can be used for developing and testing future implementation strategies to improve program accessibility.

Future Directions

To enhance the reach and impact of the National DPP, future research should explore ways to improve recruitment strategies, particularly by engaging health care providers in rural and underserved communities to actively refer high-risk individuals. Furthermore, integrating culturally tailored dietary and PA resources could improve the program's acceptability across diverse populations. Additionally, the effectiveness of emerging digital self-monitoring tools, such as wearable devices, should be assessed for their long-term impact on health outcomes and participant adherence, particularly in low-resource settings.

XSL•FO
RenderX

Finally, comparative research evaluating cost-effectiveness between different delivery modalities (eg, distance learning versus online) is essential to inform scalable, sustainable interventions that reduce health disparities in rural areas.

Conclusions

This study highlights the potential of digital delivery, through both distance learning and online modalities, to expand the reach of the National DPP to rural populations. While digital modalities offer significant advantages in accessibility and flexibility, differences in participant engagement, social support, and technology use across platforms must be addressed to enhance program effectiveness. The results suggest that tailoring recruitment strategies, particularly through health care provider involvement, and offering more structured dietary guidance could improve program acceptability and outcomes. Ultimately, these insights underscore the need for continued research to optimize digital delivery approaches and ensure equitable access to diabetes prevention resources in underserved communities.

Acknowledgments

AMG and GW conceived the premise of this paper. Focus group interviews were conducted by AMG, GW, and LMK conducted the deductive reflexive analysis. GW drafted the initial paper. All authors assisted in drafting and provided a critical review of this paper. All authors had final approval of the submitted and published version of this paper. The research team extends its gratitude to the participants in this study, as well as the extensive support from the administration and agents of K-State Research and Extension. This study was supported by the Kansas Center for Metabolism and Obesity Research (P20GM144269) and Clinical and Translational Science Award (CTSA, UL1TR002366).

Conflicts of Interest

None declared.

Multimedia Appendix 1
COREQ checklist.
[PDF File (Adobe PDF File), 482 KB - formative_v9i1e70406_app1.pdf]

References

1. Parker E, Lin J, Mahoney T, Ume N, Yang G, Gabbay R, et al. Economic costs of diabetes in the U.S. in 2022. Diabetes Care 2024;47(1):26-43. [doi: 10.2337/dci23-0085] [Medline: 37909353]
2. Centers for Disease Control and Prevention. National diabetes statistics report. In: Centers for Disease Control and Prevention. Atlanta, GA: U.S. Dept of Health and Human Services; 2022.
3. Lin J, Thompson TJ, Cheng YJ, Zhuo X, Zhang P, Gregg E, et al. Projection of the future diabetes burden in the United States through 2060. Popul Health Metr 2018;16(1):9 [FREE Full text] [doi: 10.1186/s12963-018-0166-4] [Medline: 29903012]
4. Knowler WC, Barrett-Connor E, Fowler SE, Hamman RF, Lachin JM, Walker EA, Diabetes Prevention Program Research Group. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. N Engl J Med 2002;346(6):393-403 [FREE Full text] [doi: 10.1056/NEJMoa012512] [Medline: 11832527]
5. Alva ML, Chakkalakal RJ, Moin T, Galaviz KI. The diabetes prevention gap and opportunities to increase participation in effective interventions. Health Aff (Millwood) 2022;41(7):971-979 [FREE Full text] [doi: 10.1377/hlthaff.2022.00259] [Medline: 35759735]
6. Ely EK, Gruss SM, Luman ET, Gregg EW, Ali MK, Nhim K, et al. A national effort to prevent type 2 diabetes: participant-level evaluation of CDC's National Diabetes Prevention Program. Diabetes Care 2017;40(10):1331-1341 [FREE Full text] [doi: 10.2337/dc16-2099] [Medline: 28500215]
7. Ariel-Donges AH, Gordon EL, Dixon BN, Eastman AJ, Bauman V, Ross KM, et al. Rural/urban disparities in access to the National Diabetes Prevention Program. Transl Behav Med 2020;10(6):1554-1558 [FREE Full text] [doi: 10.1093/tbm/ibz098] [Medline: 31228199]
8. Centers for Disease Control and Prevention. Centers for Disease Control and Prevention Diabetes Prevention Recognition Program: Standards and Operating Procedures. Atlanta, GA: United States Department of Health and Human Services; 2024.
9. Klitzman P, Armstrong B, Janicke DM. Distance as a predictor of treatment attendance in a family based pediatric weight management program in rural areas. J Rural Health 2015;31(1):19-26. [doi: 10.1111/jrh.12078] [Medline: 25040534]
10. Ritchie ND, Turk MT, Holtrop JS, Durfee MJ, Dickinson LM, Kaufmann PG. A virtual recruitment protocol promotes enrollment of underrepresented groups in a diabetes prevention trial. J Clin Transl Sci 2024;8(1):e26 [FREE Full text] [doi: 10.1017/cts.2024.11] [Medline: 38384920]
11. Ritchie ND, Turk MT. Enhancing access and impact of the Medicare Diabetes Prevention Program using telehealth: a narrative review. mHealth 2024;10:10 [FREE Full text] [doi: 10.21037/mhealth-23-37] [Medline: 38323146]

XSL•FO
RenderX

12. Ali MK, McKeever Bullard K, Imperatore G, Benoit SR, Rolka DB, Albright AL, et al. Reach and use of diabetes prevention services in the United States, 2016-2017. JAMA Netw Open 2019;2(5):e193160 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.3160] [Medline: 31074808]
13. Gruss SM, Nhim K, Gregg E, Bell M, Luman E, Albright A. Public health approaches to type 2 diabetes prevention: the US National Diabetes Prevention Program and beyond. Curr Diab Rep 2019;19(9):78 [FREE Full text] [doi: 10.1007/s11892-019-1200-z] [Medline: 31385061]
14. Delahanty LM, Pan Q, Jablonski KA, Aroda VR, Watson KE, Bray GA, Diabetes Prevention Program Research Group. Effects of weight loss, weight cycling, and weight loss maintenance on diabetes incidence and change in cardiometabolic traits in the Diabetes Prevention Program. Diabetes Care 2014;37(10):2738-2745 [FREE Full text] [doi: 10.2337/dc14-0018] [Medline: 25024396]
15. Perri MG, Limacher MC, Durning PE, Janicke DM, Lutes LD, Bobroff LB, et al. Extended-care programs for weight management in rural communities: the treatment of obesity in underserved rural settings (TOURS) randomized trial. Arch Intern Med 2008;168(21):2347-2354 [FREE Full text] [doi: 10.1001/archinte.168.21.2347] [Medline: 19029500]
16. Dewitt E, Ross T, Buckman C, Faris A, Norman-Burgdolf H. Leveraging clinical-community partnerships to deliver the National Diabetes Prevention Program through cooperative extension. ADCES in Practice 2023;11(3):8-13. [doi: 10.1177/2633559x231159379]
17. The case for diabetes prevention in extension- preventing diabetes is everyone's business. Cooperative Extension National Diabetes Prevention Program Working Group. 2023. URL: https://preventdiabetes.extension.org/ [accessed 2025-05-16]
18. Gorczyca AM, Washburn RA, Smith P, Montgomery RN, Koon LM, Hastert M, et al. Feasibility and comparative effectiveness for the delivery of the National Diabetes Prevention Program through cooperative extension in rural communities. Int J Environ Res Public Health 2022;19(16):9902 [FREE Full text] [doi: 10.3390/ijerph19169902] [Medline: 36011535]
19. Feldstein AC, Glasgow RE. A practical, robust implementation and sustainability model (PRISM) for integrating research findings into practice. Jt Comm J Qual Patient Saf 2008;34(4):228-243. [doi: 10.1016/s1553-7250(08)34030-6] [Medline: 18468362]
20. Glasgow RE, Trinkley KE, Ford B, Rabin BA. The application and evolution of the Practical, Robust Implementation and Sustainability Model (PRISM): history and innovations. Glob Implement Res Appl 2024;4(4):404-420. [doi: 10.1007/s43477-024-00134-6] [Medline: 39568619]
21. Kitzinger J. Qualitative research. Introducing focus groups. BMJ 1995;311(7000):299-302 [FREE Full text] [doi: 10.1136/bmj.311.7000.299] [Medline: 7633241]
22. Tausch AP, Menold N. Methodological aspects of focus groups in health research: results of qualitative interviews with focus group moderators. Glob Qual Nurs Res 2016;3:2333393616630466 [FREE Full text] [doi: 10.1177/2333393616630466] [Medline: 28462326]
23. Braun V, Clarke V. Supporting best practice in reflexive thematic analysis reporting in palliative medicine: a review of published research and introduction to the Reflexive Thematic Analysis Reporting Guidelines (RTARG). Palliat Med 2024;38(6):608-616 [FREE Full text] [doi: 10.1177/02692163241234800] [Medline: 38469804]
24. Gorczyca AM, Washburn RA, Ptomey L, Mayo MS, Sullivan DK, Gibson CA, et al. Weight management in rural health clinics: the Midwest diet and exercise trial. Contemp Clin Trials 2018;67:37-46 [FREE Full text] [doi: 10.1016/j.cct.2018.02.006] [Medline: 29454140]
25. Gorczyca AM, Washburn RA, Ptomey LT, Mayo MS, Krebill R, Sullivan DK, et al. Weight management in rural health clinics: results from the randomized midwest diet and exercise trial. Obes Sci Pract 2024;10(2):e753 [FREE Full text] [doi: 10.1002/osp4.753] [Medline: 38660371]
26. Braun V, Clarke V. One size fits all? What counts as quality practice in (reflexive) thematic analysis? Qual Res Psychol 2021;18(3):328-352. [doi: 10.1080/14780887.2020.1769238]
27. Vaismoradi M, Turunen H, Bondas T. Content analysis and thematic analysis: implications for conducting a qualitative descriptive study. Nurs Health Sci 2013;15(3):398-405. [doi: 10.1111/nhs.12048] [Medline: 23480423]
28. Braun V, Clarke V. A critical review of the reporting of reflexive thematic analysis in Health Promotion International. Health Promot Int 2024;39(3):daae049 [FREE Full text] [doi: 10.1093/heapro/daae049] [Medline: 38805676]
29. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
30. Ackermann RT, O'Brien MJ. Evidence and challenges for translation and population impact of the diabetes prevention program. Curr Diab Rep 2020;20(3):9 [FREE Full text] [doi: 10.1007/s11892-020-1293-4] [Medline: 32080770]
31. Estabrooks P, You W, Hedrick V, Reinholt M, Dohm E, Zoellner J. A pragmatic examination of active and passive recruitment methods to improve the reach of community lifestyle programs: the Talking Health trial. Int J Behav Nutr Phys Act 2017;14(1):7 [FREE Full text] [doi: 10.1186/s12966-017-0462-6] [Medline: 28103935]
32. Prasanna A, Miller HN, Wu Y, Peeler A, Ogungbe O, Plante TB, et al. Recruitment of black adults into cardiovascular disease trials. J Am Heart Assoc 2021;10(17):e021108 [FREE Full text] [doi: 10.1161/JAHA.121.021108] [Medline: 34431310]

XSL•FO
RenderX

33. Wilson K, Michaud T, Almeida F, Schwab R, Porter G, Aquilina K, et al. Using a population health management approach to enroll participants in a diabetes prevention trial: reach outcomes from the PREDICTS randomized clinical trial. Transl Behav Med 2021;11(5):1066-1077 [FREE Full text] [doi: 10.1093/tbm/ibab010] [Medline: 33677529]
34. Nhim K, Khan T, Gruss SM, Wozniak G, Kirley K, Schumacher P, et al. Primary care providers' prediabetes screening, testing, and referral behaviors. Am J Prev Med 2018;55(2):e39-e47 [FREE Full text] [doi: 10.1016/j.amepre.2018.04.017] [Medline: 29934016]
35. PreventT2 Curriculum and Handouts. Centers for Disease Control and Prevention. 2021. URL: https://www.cdc.gov/diabetes-prevention/php/lifestyle-change-resources/t2-curriculum.html [accessed 2025-05-23]
36. American Diabetes Association Professional Practice Committee. 3. Prevention or delay of diabetes and associated comorbidities: standards of care in diabetes-2024. Diabetes Care 2024;47(Suppl 1):S43-S51. [doi: 10.2337/dc24-S003] [Medline: 38078581]
37. Wilson HK, Castillo-Hernandez IM, Evans EM, Williams ER, Berg AC. Diet quality outcomes of a cooperative extension Diabetes Prevention Program. J Nutr Educ Behav 2023;55(10):734-742. [doi: 10.1016/j.jneb.2023.06.005] [Medline: 37480887]
38. Ritchie ND, Baucom KJ, Sauder KA. Current perspectives on the impact of the National Diabetes Prevention Program: building on successes and overcoming challenges. Diabetes Metab Syndr Obes 2020;13:2949-2957 [FREE Full text] [doi: 10.2147/DMSO.S218334] [Medline: 32903871]
39. Ely EK, Ng BP, Cannon MJ. Delivering the National Diabetes Prevention Program: assessment of outcomes in in-person and virtual organizations. J Diabetes Res 2023;2023:8894593. [doi: 10.1155/2023/8894593] [Medline: 37928892]
40. Smith DH, O'Keeffe-Rosetti M, Fitzpatrick SL, Mayhew M, Firemark AJ, Gruß I, et al. Costs and cost-effectiveness of implementing a digital diabetes prevention program in a large, integrated health system. Perm J 2022;26(3):74-82 [FREE Full text] [doi: 10.7812/TPP/22.029] [Medline: 36038966]
41. Wu S, Li G, Du L, Chen S, Zhang X, He Q. The effectiveness of wearable activity trackers for increasing physical activity and reducing sedentary time in older adults: a systematic review and meta-analysis. Digit Health 2023;9:20552076231176705 [FREE Full text] [doi: 10.1177/20552076231176705] [Medline: 37252261]
42. Tang MSS, Moore K, McGavigan A, Clark RA, Ganesan AN. Effectiveness of wearable trackers on physical activity in healthy adults: systematic review and meta-analysis of randomized controlled trials. JMIR mHealth uHealth 2020;8(7):e15576 [FREE Full text] [doi: 10.2196/15576] [Medline: 32706685]
43. Nagappan A, Krasniansky A, Knowles M. Patterns of ownership and usage of wearable devices in the United States, 2020-2022: survey study. J Med Internet Res 2024;26:e56504 [FREE Full text] [doi: 10.2196/56504] [Medline: 39058548]
44. Cruz JL, Luke DA, Ceballos RM, Ramanadhan S, Emmons KM. Reconceptualizing rurality: exploring community capital to identify distinct rural classes in the United States. SSM Popul Health 2025;29:101729 [FREE Full text] [doi: 10.1016/j.ssmph.2024.101729] [Medline: 39759383]
45. Loeb TB, Gholami M, Ramm K, Shedd K, Soetenga S, Jackson NJ, et al. Multilevel perceptions of the virtual delivery of the University of California diabetes prevention program on RE-AIM domains due to COVID-19 mandates. Front Public Health 2024;12:1327429 [FREE Full text] [doi: 10.3389/fpubh.2024.1327429] [Medline: 38525342]
46. Loeb TB, Ramm K, Gholami M, Shedd K, Soetenga S, Bhagat M, et al. Implementation lessons learned from the University of California's Diabetes Prevention Program initiative. BMC Public Health 2024;24(1):2777 [FREE Full text] [doi: 10.1186/s12889-024-20198-1] [Medline: 39394103]
47. Brown CW, Alexander DS, Ellis SD, Roberts D, Booker MA. Perceptions and practices of diabetes prevention among African Americans participating in a faith-based community health program. J Community Health 2019;44(4):694-703. [doi: 10.1007/s10900-019-00667-0] [Medline: 31065854]
48. Andrade CAS, Lovas S, Mahrouseh N, Chamouni G, Shahin B, Mustafa EOA, et al. Primary prevention of type 2 diabetes mellitus in the European union: a systematic review of interventional studies. Nutrients 2025;17(6):1053 [FREE Full text] [doi: 10.3390/nu17061053] [Medline: 40292470]
49. Haseldine C, Pallin JA, Kearney PM, Mc Hugh SM, Riordan F, Cotterill S, et al. Healthcare workers' perspectives on barriers and facilitators to referral to type 2 diabetes prevention programmes: a systematic review. BMJ Open 2025;15(3):e090105 [FREE Full text] [doi: 10.1136/bmjopen-2024-090105] [Medline: 40054870]
50. Rodrigues AM, Haste A, Penn L, Bell R, Summerbell C, White M, et al. Stakeholders' perceptions and experiences of the National Health Service Diabetes Prevention Programme in England: qualitative study with service users, intervention providers and deliverers, commissioners and referrers. BMC Health Serv Res 2020;20(1):307. [doi: 10.1186/s12913-020-05160-2] [Medline: 32293424]
51. Haseldine C, O'Donoghue G, Kearney PM, Riordan F, Kerins C, Kirby L, et al. Factors influencing participation in an online National Diabetes Prevention Programme: a qualitative study with attenders and educators. Diabet Med 2024;41(6):e15277. [doi: 10.1111/dme.15277] [Medline: 38150286]
52. Moussas-Alvarez D, Hawkes RE, Miles LM, Dack C, French DP. Service users' experiences of, and engagement with, a nationally implemented digital diabetes prevention programme. Br J Health Psychol 2025;30(1):e12787. [doi: 10.1111/bjhp.12787] [Medline: 39968862]

XSL•FO
RenderX

Abbreviations

CDC: Centers for Disease Control and Prevention
COREQ: Consolidated Criteria for Reporting Qualitative Research
DPP: Diabetes Prevention Program
Extension: US Cooperative Extension System
HbA1c: glycated hemoglobin
National DPP: National Diabetes Prevention Program
PA: physical activity
PRISM: Practical, Robust Implementation and Sustainability Model
REDCap: Research Electronic Data Capture

Edited by A Mavragani; submitted 20.12.24; peer-reviewed by M Ocean, R Hawkes; comments to author 21.03.25; revised version received 11.04.25; accepted 09.05.25; published 02.07.25.

Please cite as:
Wagner G, Koon LM, Smith P, Suire KB, Hastert M, Donnelly JE, Olfert MD, Estabrooks P, Gorczyca AM
Assessing the Fit of a Digitally Delivered National Diabetes Prevention Program Among Rural Living Adults: Qualitative Study
JMIR Form Res 2025;9:e70406
URL: https://formative.jmir.org/2025/1/e70406
doi:10.2196/70406
PMID:

©Gerit Wagner, Lyndsie M Koon, Patricia Smith, Kameron B Suire, Mary Hastert, Joseph E Donnelly, Melissa D Olfert, Paul Estabrooks, Anna M Gorczyca. Originally published in JMIR Formative Research (https://formative.jmir.org), 02.07.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

User-Driven Development of a Digital Behavioral Intervention for Chronic Pain: Multimethod Multiphase Study

Afra Selma Taygar[1], MSc; Sara Laureen Bartels[1,2], PhD; Rocío de la Vega[3,4], PhD; Ida Flink[5], PhD; Linnéa Engman[1,6,7], PhD; Suzanne Petersson[1,8], PhD; Sophie I Johnsson[9], MSc; Katja Boersma[7], PhD; Lance M McCracken[10], PhD; Rikard K Wicksell[1,6], PhD

[1]Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden
[2]Department of Psychiatry and Neuropsychology, Maastricht University, Maastricht, The Netherlands
[3]Department of Personality, Assessment and Psychological Treatment, Universidad de Málaga, Málaga, Spain
[4]Instituto de Investigación Biomédica de Málaga, Málaga, Spain
[5]Department of Social and Psychological Studies, Karlstad University, Karlstad, Sweden
[6]Pain clinic, Saint Göran Hospital, Stockholm, Sweden
[7]School of Behavioural, Social and Legal Sciences, Center for Health and Medical Psychology (CHAMP), Örebro University, Örebro, Sweden
[8]Department of Medicine and Optometry, Linnaeus University, Kalmar, Sweden
[9]Department of Molecular Medicine and Surgery, Karolinska Institutet, Stockholm, Sweden
[10]Department of Clinical Psychology, Uppsala University, Uppsala, Sweden

Corresponding Author:
Afra Selma Taygar, MSc
Department of Clinical Neuroscience
Karolinska Institutet
Nobels Väg 9
Stockholm, 17165
Sweden
Phone: 46 702698762
Email: afra.taygar@ki.se

Abstract

Background: Recent research shows that chronic pain affects 27% of the adult population. For many, pain significantly impairs quality of life and everyday functioning. Behavioral interventions have shown utility, but access remains limited. Digital health solutions can increase reach, but there is a need for user-friendly, feasible, and evidence-based digital interventions.

Objective: This study aimed to clarify how a digital behavioral intervention for people with chronic pain can be developed through a user-centered approach to address the needs and preferences of the target population.

Methods: This study used a multimethod approach involving end users, namely, patients with chronic pain and therapists, to develop prototypes for a digital behavioral intervention across 3 phases. In the preparation phase (phase 0), fictional patient personas (n=3) were created to represent the diversity of the target population while emphasizing transdiagnostic features across people with chronic pain. In the design phase (phase 1), qualitative data from focus groups with patients (n=5; aged 37-51 years; 4/5, 80% women; 2/5, 40% diagnosed with Ehlers-Danlos syndrome; 3/5, 60% either undiagnosed or uncertain about their diagnosis) and therapists (n=12 licensed psychologists; aged 29-64 years; 9/12, 75% women) were collected to explore end-user preferences for the intervention design and content. In the testing phase (phase 2), the initial full prototype of the digital intervention was piloted with patients (n=11; aged 36-58 years; 9/11, 82% women; with diverse diagnoses, including migraine, arthritis, fibromyalgia, complex regional pain syndrome, hypermobile Ehlers-Danlos syndrome, herniated disc, chronic fatigue syndrome, and 1/11, 9% cases of undiagnosed pain) and therapists (n=3 licensed psychologists; aged 36-58 y; 3/3, 100% women). The Consolidated Framework for Implementation Research was used to structure analyses of end-user feedback.

Results: On the basis of end-user input, a 6-week digital behavioral intervention for chronic pain was created. Focus groups highlighted the importance of accessibility and adaptability of the digital intervention, emphasizing the need for tailored content, flexibility (eg, contact with the therapist via asynchronous messaging, telephone, or video calls), and user-friendly design (eg, easy navigation between modules, short microsessions, and visualizations). Average weekly ratings (scale from 1=not at all to 7=very much) by patients during pilot-testing indicated that the intervention was helpful (mean range 4.27-5.45, SD range

XSL•FO
RenderX

1.20-2.20), enjoyable (mean range 3.81-4.81, SD range 1.12-2.08), and understandable (mean range 4.45-6, SD range 1.30-1.86), suggesting initial acceptability and usability of the intervention.

Conclusions: The results illustrated the utility of the patient personas when preparing, of the focus groups when designing, and of the end-user feedback when testing this new digital intervention for people with chronic pain. The findings indicated that the intervention is promising while also providing relevant end-user suggestions (eg, video content, text-to-speech function, and add-on modules) to guide further improvements.

(JMIR Form Res 2025;9:e74064) doi:10.2196/74064

KEYWORDS

chronic pain; digital therapeutics; behavioral intervention; development; end-user involvement

Introduction

Background

Chronic pain affects approximately 27% of adults [1], significantly impacting their daily life and general well-being [2]. Behavioral interventions building on, for instance, the fear-avoidance [3] and psychological flexibility models [4,5], aim to enhance resilience to pain and distress. These interventions have robust empirical support and are increasingly used [6-8]. However, access to evidence-based behavioral interventions remains low [9].

The evolution of digital solutions represents a paradigm shift, with the potential to maximize the accessibility of behavioral treatments, yet there is a scarcity of evidence-based digital treatments in regular health care [10]. To facilitate further development in the field, the innovation process should be concise and efficient [11]. It should be executed within an established *framework* to facilitate standardization, *described* in detail for transparency and quality assurance, and *evaluated* scientifically for data-driven decisions on how to proceed for further testing and implementation [12].

Existing research suggests that user involvement is essential to match digital health interventions to preferences and needs [13-16]. In addition, stakeholders should be consulted early on to facilitate successful and sustainable implementation [17]. Stakeholders include, for instance, *innovation facilitators* such as health care managers and IT developers and *end users* such as patients and health care professionals. In particular, end-user involvement provides insights into relevant needs and priorities [18,19]. Collaboration with end users to tailor the treatment is critical in the development phase to emphasize person-centeredness [20,21]. Furthermore, end-user engagement throughout the development process is essential for creating effective digital health interventions [22] and increasing ease of use [23].

Recently, the UK Medical Research Council recommended that the development of novel interventions involve individuals with lived experiences in each phase (ie, development, feasibility, evaluation, and implementation) to assure inclusivity, accessibility, and efficacy [24]. While an increasing number of studies have identified preferences and needs regarding digital interventions for chronic pain [25-27], end-user involvement is usually limited to a certain phase of the innovation process. How end users can be involved in the implementation and evaluation of digital behavioral interventions has been described [28,29], but clear descriptions of end-user engagement during the development phase remain limited. Thus, there is a need for studies clarifying how the development of novel digital interventions for chronic pain can apply a user-centered approach to facilitate replicability and establish standards for digital health innovation.

Objectives

Digital Behavioural Health for Chronic Pain (DAHLIA) is a user-centered multiphase project with two distinct yet related purposes: (1) to create an evidence-based digital health intervention for people with chronic pain and (2) to provide a robust and replicable process for user-centered development, evaluation, and implementation [30]. The overarching aim of this study was to clarify how a digital behavioral intervention can be developed through a user-centered approach to address the needs and preferences of the target population. Development included 3 phases: preparation (phase 0), design (phase 1), and testing (phase 2).

More specifically, the research questions were as follows:

1. Phase 0—preparation. How can patient personas be used to define the relevant patient characteristics, needs, and treatment targets of people living with chronic pain? How can a better understanding of patient personas guide the prototype development?
2. Phase 1—design. What are the preferences of end users for the design (ie, content, structure, and format) of this digital intervention as reported in end-user focus groups (patients and therapists)?
3. Phase 2—testing. How do end users (patients and therapists) perceive engaging with the first treatment prototype (version 1.0), and how can the digital intervention be further improved to meet needs and preferences?

Methods

Study Design

This study used a multimethod approach in the development of DAHLIA prototypes 1.0 and 2.0 and included 3 phases (0=preparation, 1=design, and 2=testing). This study is part of a larger project also containing evaluation and implementation phases [30], which will be presented elsewhere.

Throughout, *participants* refers to all individuals who took part in the study, including both patients and therapists. We used the term *patients* referring to people living with chronic pain

XSL•FO
RenderX

and *therapists* referring to people delivering the intervention. Both groups are considered *end users* of the digital intervention, reflecting their involvement in using and evaluating the intervention.

Intervention development was centered on end users (patients and therapists) and used complementary user engagement approaches, namely, patient personas (phase 0; see the Phase 0 [Preparation]: Preliminary Patient Characteristics, Needs, and Treatment Targets section), qualitative focus groups (phase 1; see the Phase 1 [Design]: User-Centered Design of the Digital Intervention section), and multimethod pilot-testing based on perceived end-user experience (phase 2; see the Phase 2 [Testing]: Piloting the Digital Intervention section). These development phases are presented in Figure 1 and further described in the following sections.

The development of the intervention was reported following the Guidance for the Reporting of Intervention Development (GUIDED) checklist [31]. The intervention was also described using the Template for Intervention Description and Replication checklist [32] (see the completed GUIDED and Template for Intervention Description and Replication checklists in Multimedia Appendix 1 [30-32]), as recommended by the GUIDED checklist.

Figure 1. Overview of the development phases (ie, 0=preparation; 1=design; 2=testing) of the DAHLIA digital behavioral intervention for people with chronic pain. The process followed a user-centered design approach involving end users (patients and therapists) throughout.

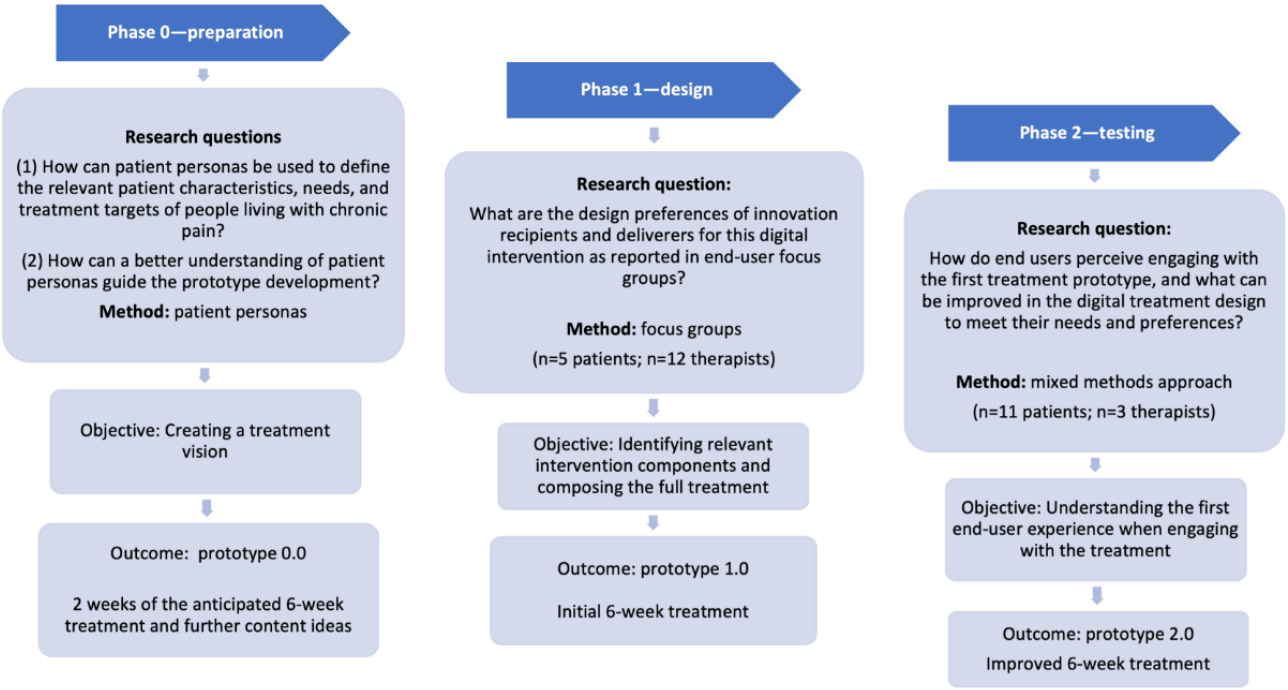

Ethical Considerations

This study was approved by the Swedish Ethical Review Authority (approval Dnr 2021-02437) and carried out in accordance with the Declaration of Helsinki. All participants provided informed consent before joining the study, including participation in the focus groups and pilot-testing. Participants also consented to the use of anonymized quotes and data for dissemination. Collected data were stored securely on encrypted servers and were accessible only to the research team. Participants received no compensation.

Phase 0 (Preparation): Preliminary Patient Characteristics, Needs, and Treatment Targets

To address the first research question, the initial design of the prototype was based on a theoretical framework and conceptual model, as well as a preliminary definition of the patient characteristics, needs, and relevant treatment targets based on patient personas.

Theoretical Frameworks and Conceptual Model

The DAHLIA treatment program is based on learning theory [33], with the fear-avoidance model [3] and the psychological flexibility model [4,5] integrated into a comprehensive conceptual model. The primary treatment objective is to increase resilience (ie, being able to sustain living a fulfilling life in the presence of distress) [34] to chronic pain and distress by improving behavioral self-management skills relevant to well-being and functioning [35].

Patient Personas

Personas, or fictional user profiles, aim to depict the target group of a treatment or product [36]. As the involvement of people with chronic pain may be challenging in the early stages of

XSL•FO
RenderX

development due to ethical or practical reasons, personas can be used as representations of the target group [37]. Personas are a design methodology, and their utility depends on the project aim.

In the DAHLIA project, 3 patient personas were used during the initial conceptualization and preparation of the treatment. Patient personas were used to discuss whether the treatment vision (eg, content, setup, and design) aligned with the needs and characteristics of the target population. The patient personas were designed to represent the heterogeneity (eg, gender, age, and pain conditions) of the target population while emphasizing transdiagnostic features across people living with chronic pain (eg, self-management strategies). The patient personas facilitated an inclusive design process and the identification of common user needs across different patient profiles to guide decisions on the structure and content of the digital intervention.

The patient personas were developed by clinical researchers (RKW, IF, KB, LMC, and SP) inspired by patient personas used in a previous study [38] and edited over several months (SLB and SIJ) until the project partners reached a consensus. The patient personas were informed by existing chronic pain research, particularly the survey of chronic pain in Europe by Breivik et al [39]. Each patient persona consisted of 4 main domains: sociodemographics, pain profile, health care, and personal needs and goals. One of the 3 patient personas used to guide the user-centered design and tailor the intervention content is presented in Multimedia Appendix 2, and the other 2 are presented elsewhere [30,37]. Through narrative synthesis, a summary of the patient persona's characteristics, key challenges in the development of the intervention, and potential ways to approach these challenges were identified, guiding discussions and decisions to refine the treatment content.

Intervention Structure for Prototype Version 0.0

The treatment prototype version 0.0 was prepared based on theoretical considerations and implications of the patient personas following consensus discussions in the project group (see the Results section). Moreover, a microlearning approach [40,41] was applied to organize treatment content into brief and frequent sessions, which was considered useful to prospective users experiencing challenges due to, for instance, dyslexia, fatigue, attention deficits, or reduced cognitive functioning. It was anticipated that the final treatment would include 6 modules, with 4 sessions per module, totaling 24 treatment modalities. Overall, a 6-week treatment duration was anticipated, with 1 module per week. The outcome of phase 0 was treatment prototype version 0.0, including 2 modules of the treatment, building the foundation for end-user input in phase 1.

Phase 1 (Design): User-Centered Design of the Digital Intervention

Insights from phase 0 (the preparation phase), including the initial understanding of patient needs based on patient personas, informed the development of the prototype in phase 1. The design of the digital health intervention (prototype version 1.0) was based on a preliminary understanding of patient needs and relevant treatment targets and guided by end-user input from focus groups.

End-User Focus Groups

To develop the full 6-week treatment, both prospective patients and therapists were involved to provide user input on their needs and preferences regarding the digital intervention. A qualitative study with 2 patient focus groups (n=5 in total) and 3 therapist focus groups (n=12 in total) was conducted. Notably, the initial research plan [30] was to involve 6 to 8 participants per focus group and conduct the discussions face-to-face. However, due to the COVID-19 pandemic, focus groups were conducted digitally, and the target size of each group was reduced to 3 to 5 participants per group based on previous research [42], especially for digital settings [43].

Participants and Recruitment

Inclusion criteria to participate in a *patient focus group* were (1) age of 18 to 65 years (working age); (2) pain duration of ≥3 months; (3) ability to communicate in Swedish; and (4) access to a computer, smartphone, and internet connection in a home environment. Individuals were excluded if they had serious psychiatric comorbidities (eg, risk of suicide).

Inclusion criteria for the *therapist focus group* were (1) being a licensed psychologist or psychotherapist with training in cognitive behavioral therapy (CBT), (2) fluency in Swedish, and (3) access to a computer or laptop with an internet connection. Experience treating patients with chronic pain was recommended but not required. No exclusion criteria were specified to increase external validity.

Eligible therapists received an email with a link to provide informed consent. Participants (patients and therapists) were recruited from 2 health care centers in 2 different regions of Sweden (Stockholm and Kalmar). For patient recruitment, flyers were distributed at clinics, and interested individuals scanned a QR code and were directed to a digital system (REDCap [Research Electronic Data Capture; Vanderbilt University] [44,45]) to register. Registered individuals were contacted by a research assistant (SIJ) to receive detailed information about the study. A clinical coordinator (SP, licensed psychologist) screened potential participants to check eligibility.

Materials and Procedures

Eligible participants were contacted via email with a link to provide informed consent and sociodemographic information (eg, age, sex, and occupation) in REDCap before the focus group. Therapists also reported their level of experience delivering psychological treatment (number of years), including for people with chronic pain. Patients were asked to complete questions specific to their pain condition.

The focus groups were conducted as 2-hour–long recorded video meetings (Microsoft Teams) held in the first half of 2022 and transcribed verbatim. The meetings were moderated by a research assistant (SIJ) with support from 2 psychology students. The focus groups were conducted in Swedish, professionally translated into English, and subsequently analyzed in English by nonnative speakers. The focus groups followed a semistructured format (see Multimedia Appendix 3 for the full interview guide) with 2 distinct objectives: to identify *general health needs* and generate user input on *the digital treatment*

XSL•FO
RenderX

content and design. This study focused on the user-centered input on the digital treatment, and user input on general health needs will be presented in another publication. To facilitate concrete feedback, participants were given access to the treatment before the focus groups and asked to focus on the first 2 modules of the treatment and the general treatment vision.

Data Analysis: Focus Groups

Sample characteristics were analyzed using descriptive statistics. A qualitative framework analysis [46] was used for the (qualitative) focus group data using the following steps: (1) *familiarization* with the data by listening to recordings and reading transcripts, (2) *identification* of the thematic framework by deciding whether to conduct an inductive or deductive approach, (3) *indexing* by creating a coding frame of the highlights from the data, (4) *charting* by placing themes into rows and columns in summary, and (5) *mapping and interpretation* by looking into similarities and dissimilarities between participants' responses.

The Consolidated Framework for Implementation Research (CFIR) [47] was used to guide the coding of the focus group data. The CFIR consists of 5 main domains (ie, intervention characteristics, outer setting, inner setting, individual characteristics, and implementation process) [47,48]. Phase 1 focused on the CFIR domain *intervention characteristics*, specifically on the design and content of the digital behavioral intervention, including the subdomains of evidence base, relative advantage, adaptability, complexity, and design of the innovation (which refers to the "digital intervention"; see specific definitions in Multimedia Appendix 4). In total, 2 independent researchers (AT and SLB) conducted the qualitative analysis, with a third researcher validating the outcome of the analysis by reviewing the final framework.

Phase 2 (Testing): Piloting the Digital Intervention

On the basis of feedback from the focus group participants in phase 1, the prototype was refined and prepared for testing in phase 2 to be carried out through piloting the digital intervention with end users. The objective of this phase was to test and identify areas for improvement of the digital treatment design (ie, content, structure, and format).

Participants and Recruitment

The inclusion and exclusion criteria for patients and therapists in phase 2 were consistent with the eligibility criteria of the focus groups (phase 1; see the Patient and Recruitment section). In addition, patients were excluded if they (1) had an injury or illness that required an immediate assessment or treatment, (2) had changes in prescribed medication in the previous 3 months or changes were expected in the following 3 months, or (3) had received CBT treatment during the previous 6 months. Participants (n=11 patients and n=3 therapists) were recruited from the same 2 regions (Stockholm and Kalmar).

Materials and Procedures

Overview

The DAHLIA treatment (prototype version 1.0) consisted of 4 self-guided microsessions per week for a total of 24 sessions delivered over 6 weeks. Moreover, patients had weekly contact with a therapist through a 30-minute phone or video call. The digital intervention was offered through the Swedish health care system implemented into a digital system called 1177, the national health care web platform in Sweden [49], in collaboration with health care providers from Region Stockholm and Kalmar and health care developers and digital designers in Region Kalmar and supported by the industry partner Inera for maintenance. The 1177 platform is a secure system that ensures confidentiality via Sweden's public e-identification systems (BankID, Freja eID, and Foreign eID).

The intervention was delivered via 1177's *Stöd och Behandling* (support and treatment) feature, which is designed to support digital care programs. While the platform provides a structured framework enabling health care providers to deliver digital care programs, the technical architecture imposes certain constraints, especially on the design and customization. These constraints include limited flexibility in the user interface design and multimedia services. Therefore, interactive features (eg, videos and exercises) could only be implemented through preexisting templates or external resources. In addition, as the platform does not support real-time communication (ie, video calls and phone calls), each region complemented the intervention by using its secure internal systems to conduct these interactions.

Evaluation of Engagement With Treatment

Every week, at the end of the contact with their therapists, patients were asked to evaluate the module based on how helpful, enjoyable, and comprehensible they perceived it to be by responding to a set of statements (eg, "I experienced this week's session as helpful/enjoyable/understandable") rated on a 7-point scale ranging from 0 (not at all) to 7 (very much). Furthermore, after completing the intervention, a semistructured exit interview (see Multimedia Appendix 5 for the full patient exit interview guide and Multimedia Appendix 6 for the full therapist exit interview guide) was conducted with each participant (ie, patients and therapists) by a researcher (LE, SP, SLB, and AST) to assess user experiences of engaging with prototype version 1.0, focusing on overall experience of the treatment (eg, Did the digital treatment interfere with your daily routines? Would you recommend this treatment to a friend with a similar condition?) and more specific questions on microsessions and weekly contact with their therapist.

In addition, therapists were interviewed about their experience of providing the intervention (eg, Was it easy to navigate the digital treatment? Was the frequency of communication with the patient adequate?). These evaluations and reflections from end users, including treatment patients and therapists, were intended to guide the researchers in revising the treatment's design (structure, content, and format). Overall, these evaluations were chosen to explore intervention and implementation success (see the Materials and Procedures section). The exit interviews were held in either Swedish or English, professionally translated into English when conducted in Swedish, and subsequently analyzed in English by nonnative speakers.

XSL•FO
RenderX

Data Analysis: Pilot-Testing

All quantitative data were analyzed using descriptive statistics (mean, SD, and range). The same stepwise approach as in phase 1 (Intervention Structure for Prototype Version 0.0 section) was used to analyze the qualitative data [46,47]. Specifically, the *implementation process* domain of the CFIR [47] was used as codes to analyze the qualitative data from patients' and therapists' weekly evaluations and exit interviews. For this analysis, the CFIR *reflecting and evaluating* construct within the *implementation process* domain was used to explore the end users' perceived intervention and initial implementation success (see specific definitions in Multimedia Appendix 7).

Results

The results of each phase (0=preparation, 1=design, and 2=testing) are presented in the following sections.

Phase 0 (Preparation): Relevant Patient Characteristics, Needs, and Treatment Targets for the Digital Intervention to Guide the Preparation of the Prototype (Version 0.0)

Considering the literature and clinical experiences, the group of researchers and clinicians found it helpful and sufficient to use 3 patient personas to describe patient characteristics, needs, and treatment targets. Across all patient personas, the four domains included (1) sociodemographics, such as educational level, work, family, background, social environment, and living location; (2) pain profile, including pain problems, impact of pain, pain behavior, and attitude toward the treatment; (3) health care, such as contact with health care, comorbid conditions, and medications; and (4) personal needs and goals, particularly those related to treatment. The synthesis of characteristics, needs, and treatment targets across all 3 patient personas resulted in a visual summary (Figure 2). Moreover, based on the patient personas, the group identified a number of key challenges and ways to approach these challenges in the design and testing phases, which are summarized in Table 1.

Figure 2. Summary of implications of patient personas on the development of the DAHLIA digital behavioral intervention for people with chronic pain.

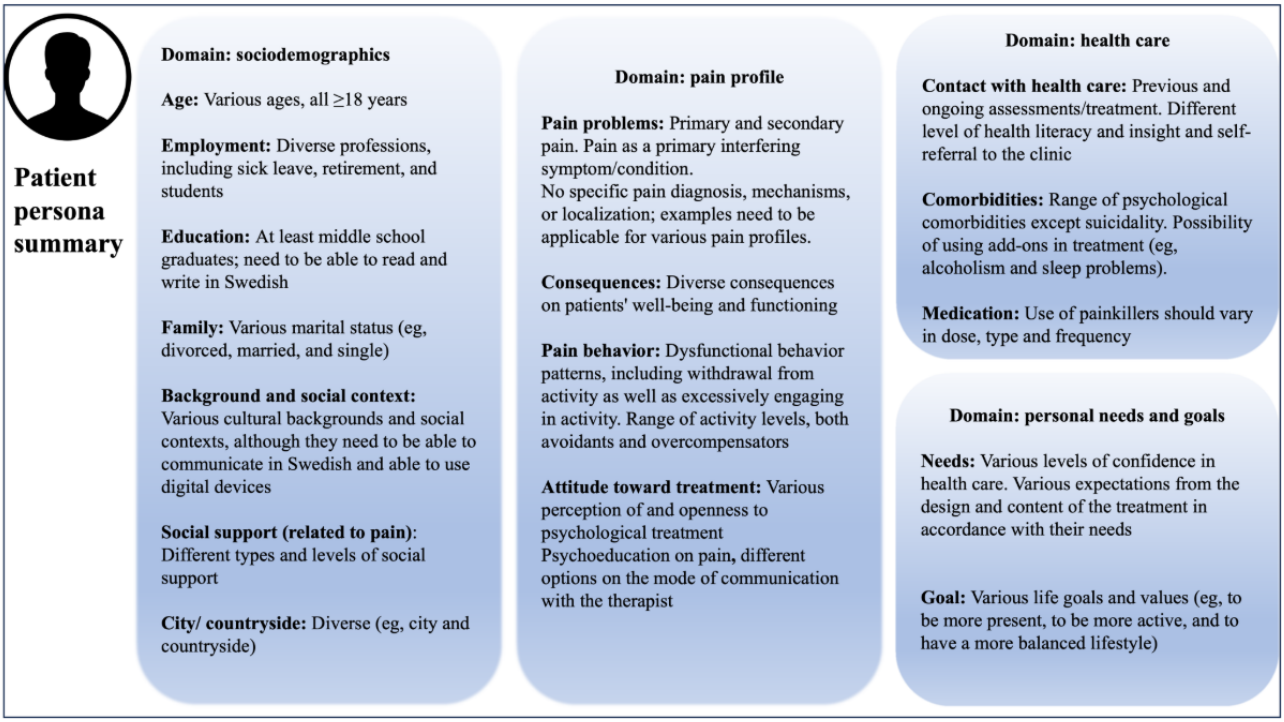

Table 1. Key challenges identified by using patient personas for the development of the DAHLIA digital behavioral intervention for people with chronic pain and potential ways to approach these challenges.

Key challenge	Approach
The chronic pain population is heterogeneous due to various characteristics, including age, gender, pain history, needs, comorbidities, abilities, and ethnicity.	The content needs to be written using inclusive language (eg, easy reading level) and feature diverse examples, the treatment needs to be tested in a heterogeneous sample, and subgroup analysis might be necessary to explore which specific characteristics influence treatment outcomes.
Digital literacy will vary significantly as some patients will be reluctant or not able to use technology.	A nondigital alternative and technology training in the onboarding phase (during enrollment) need to be offered.
The varying needs and expectations in this heterogenous population may require flexibility regarding treatment content.	The theoretical foundation of the treatment (see the Theoretical frameworks and conceptual model section) is transdiagnostic and a standardized prototype treatment and, therefore, can be assumed to benefit the population, but tailoring or add-on elements may be needed to provide support for patients with specific or more complex health challenges.

XSL•FO
RenderX

Phase 1 (Design): Preferences of End Users for the Design (ie, Content, Structure, and Format) of the Digital Intervention

Characteristics of the Participants

Patient Focus Groups

A total of 7 patients provided informed consent, resulting in 4 (57%) patients in the first focus group and 3 (43%) patients in the second focus group. Despite rescheduling both groups once, 1 patient per group was unable to join (reasons for dropout: sickness and physician's appointment). Consequently, 71% (5/7) of the patients participated in the focus groups (3/5, 60% in the first focus group and 2/5, 40% in the second focus group). Participants were, on average, aged 43.6 (SD 7.8; range 37-53) years, 20% (1/5) identified as men, and 80% (4/5) identified as women. The educational levels of the participants ranged from high school diploma to college or university degree (3/5, 60% high school diploma; 1/5, 20% ongoing university studies; 1/5, 20% college or university degree). All participants (5/5, 100%) reported having chronic pain, 60% (3/5) were unsure or had not received a specific pain diagnosis, and 40% (2/5) reported being diagnosed with Ehlers-Danlos syndrome (of whom 1/2, 50% reported a fibromyalgia diagnosis as well). Using a numerical rating scale from 0 to 10 [2], participants reported a current average pain level of 5.8 (SD 1.5; range: 4-8) and an average pain level during the previous week of 6.6 (SD 1.14; range 5-8). A summary of the characteristics of the patients is provided in Table 2.

Table 2. Sociodemographic and pain-related characteristics of the patient participants involved in the focus groups during the design phase of the DAHLIA digital behavioral intervention for people with chronic pain (N=5).

	Patient 1	Patient 2	Patient 3	Patient 4	Patient 5	Patient 6	Patient 7
Participation	Yes	Yes	Yes	Yes	Yes	Dropout	Dropout
Age (y)	40	37	53	37	51	47	__[a]
Gender	Woman	Man	Woman	Woman	Woman	Woman	—
Educational level	High school diploma	High school diploma	Ongoing university studies	High school diploma	College or university degree	High school diploma	—
Pain diagnosis	Ehlers-Danlos syndrome	Unsure or did not know	Undiagnosed pain	Ehlers-Danlos syndrome and fibromyalgia	Undiagnosed	Ehlers-Danlos syndrome	—
Current pain level (from 0 to 10)	6	4	5	6	8	6	—
Pain level during the previous week (from 0 to 10)	7	6	5	7	8	7	—

[a]No data available for patient 7.

Therapist Focus Groups

A total of 3 focus groups with 4 therapists each were conducted. In total, 12 therapists (n=8, 67% from Region Kalmar and n=4, 33% from Region Stockholm) provided informed consent and joined the focus groups (12/12, 100% retention rate). The therapists were, on average, aged 42.8 (SD 12.0; range 29-64) years, 25% (3/12) identified as men, and 75% (9/12) identified as women. All participants were licensed psychologists with, on average, 12.3 (SD 9.4; range 2-31) years of clinical experience and 9.5 (SD 9.8; range 0-28) years of experience working with patients with chronic pain.

Outcomes of the Patient and Therapist Focus Groups

The focus groups provided input on the treatment relevant to the CFIR *intervention* domain and its related constructs (evidence base, relative advantage, adaptability, complexity, and the design; see the specific definitions in Multimedia Appendix 4), which is presented in brief in the following sections.

Evidence Base of the Treatment

Therapists expressed that they perceived the content as relevant, valid, and evidence based. Patients did not have any input regarding the evidence-based content.

Relative Advantage of the Treatment

Patients and therapists reported different advantages and disadvantages in relation to other types or formats of treatment, which are presented in Table 3 in detail and summarized in this section. Some patients reported that group-based treatment had an advantage over individual treatment as it facilitated the exchange of experiences with others, whereas others endorsed individual treatment as the pace could be adjusted to the patients' needs. It was also mentioned that digital delivery may be challenging for older adults who struggle with technological devices. Furthermore, therapists reported that the digital self-help format with preexisting content and structure enabled the delivery to be more concise than standard treatment, that the microsession format was preferable over longer modules, and that the digital treatment format could be introduced to the patients earlier and cover a broader geographic area (ie, rural areas) than traditional approaches.

XSL•FO
RenderX

Table 3. Perceived relative advantages of the digital behavioral intervention for people with chronic pain as identified by patients (n=5) and therapists (n=12) in qualitative focus groups.

Participant group and subcategory—relative advantage	Example quotes
Patients	
Pros and cons of individual vs group-based treatment	• "I think that's a good thing: if someone has difficulties with something that might not be so easy to deal with in a group, then there might be an opportunity for private, individual sessions that." [Patient 2] • "It's always good to exchange experiences with others I think." [Patient 5]
Potential disadvantage—digital delivery (eg, older age and low socioeconomic status)	• "[Father of participant] was born in 50s', hates technology and everything surrounding it, doesn't even have a computer and refuses to get one. Is there any kind of technology help? What is required to do these programmes? Do you need a computer or is it enough to have a phone? How do you deal with those on welfare, if they can't afford a smartphone and computer?" [Patient 4]
ACT[a] preferred over traditional CBT[b]	• "I've tried CBT. I'm not a big fan of CBT, to be honest. I put more stock in ACT. Acceptance, which is something I've worked really hard on because I've been a bit black or white." [Patient 5]
Therapists	
Digital treatment enables therapists to provide knowledge to the patient in a more structured way than face-to-face or telephone-based treatment	• "I personally would have had a more difficult time if [I] met the patients face-to-face in this type of treatment without any support... Having a treatment program that you follow, I feel that the knowledge I have [then] about chronic pain is good enough." [Therapist 1] • "This kind of program [is not so bad], because there is built-in structure and the knowledge that the patient is getting, is what the patient needs so that it doesn't depend as much on my structure." [Therapist 4]
Digital treatment offers more training options to therapists (than currently available)	• "It is difficult to become good at something that requires a lot from the practitioner.... I think that a program like this can fulfil such a [training] function." [Therapist 4]
iCBT[c] for chronic pain currently not available	• "It's worth developing this because it's needed." [Therapist 2]
Microsession format potentially advantageous over longer iCBT sessions	• "It's very different from the regular internet-based CBT we [therapists] work with but [when] you get the patient to understand to go in ten minutes every day, then it becomes more of a habit." [Therapist 7]
Digital treatment could be offered earlier to patients than traditional therapy	• "Internet-based [treatment] is one of the first steps.... [If the system would] be able to [refer patients at] an earlier stage and in that case, we would be able to be helpful with patients that we don't see today." [Therapist 7]
Digital treatment could cover a wider area than traditional therapy	• "The possibility of practicing because I have the whole region as a 'catchment area.' It's really good to be able to reach out in this way [digitally]." [Therapist 5]

[a]ACT: acceptance and commitment therapy.

[b]CBT: cognitive behavioral therapy.

[c]iCBT: internet-based cognitive behavioral therapy.

Adaptability of the Treatment

In both patient and therapist focus groups, participants discussed the mode of communication for the weekly therapist-patient contact. While therapists emphasized the benefits of an asynchronous messaging function over a live chat function and expressed potential preference for phone calls over video calls, patients expressed a general preference toward video calls over phone calls or messages. The varying views and preferences imply a need for flexibility in the mode of communication.

Furthermore, therapists suggested that the digital treatment should be adaptable according to patients' level of knowledge and interest as some patients might have sufficient pain education whereas others would benefit from additional information. The therapists argued that the treatment length might also need to be adapted to symptom complexity; for instance, patients with longer and more complex pain and health issues might benefit from longer treatment periods, and the frequency of contact should be flexible, as highlighted by therapists. Finally, both patients and therapists proposed a text-to-speech function and subtitles in videos to better accommodate people with disabilities such as hearing, visual, or cognitive limitations. Details and supporting quotations can be found in Table 4.

XSL•FO
RenderX

Table 4. Adaptability aspects of the DAHLIA digital behavioral intervention for people with chronic pain as identified by patients (n=5) and therapists (n=12) in qualitative focus groups.

Participant group and subcategory—adaptability	Example quotes
Patients	
Flexibility in mode of communication; preference for video calls	• "There's someone who sees me. Once a week, I know I'll be seen." [Patient 3] • "I think you should be able to choose.... You might feel that's enough writing for today. I don't have anything to say but today I want to talk on the phone because I don't feel I can sit at the computer [and type a message]. So I can make the call while sitting in the car, taking the bus. But then I might feel like, now I want to see another human being [and I chose a video call]." [Patient 4]
Flexibility in engagement with treatment (ie, choosing certain exercises)	• "That maybe there's the possibility to adapt, if there's somebody who's having a very hard time, [then] you don't have to do all of this, but we can just do a specific part. To be able to adapt it to the individual a little bit because in my experience [of those who] have chronic pain, we are all different with different functions, different personalities, basic stages, other difficulties. I think it would do some good to be a bit flexible from person to person." [Patient 2]
Text-to-speech function and subtitles in videos	• "Is it possible to have the text read aloud, or will this be made possible?... text-to-speech and audio description or subtitling are a must I think." [Patient 4]
Inclusivity toward people with disabilities	• "Because if you do as much preparatory work as possible, e.g., you put in a glossary, you prepare for audio description and subtitling and text-to-speech and everything, because if you have a good foundation, the target group doesn't really matter because then you can just update." [Patient 4]
Mobile and computer versions	• "You can make one [version of the program] that's a mobile version but with even less text and even more in summary form, but still the same content." [Patient 4]
Provision of technical support if needed	• "But will they be any kind of technical help then [for those who need it]?" [Patient 4]
Paper-based version (if technology is not usable and) as a long-term resource	• "You [could] have some kind of physical book available with information and everything that you can go back to and read parts of.... You've undergone the treatment, you received the information, but five-ten years later the time may be ripe to put it to use and then you still have the book." [Patient 4]
Therapists	
Digital treatment needs to be adaptable to patients' abilities and disabilities (eg, linguistic and cognitive) and preferences	• "Pain patients often have difficulty with concentration and with reading a text. If an audio file could be added for each text, so that [the patient] can press [play] on it [and] there is a voice that says exactly the same things as in the text so [the patient] can both listen and somehow follow the text." [Therapist 10]
Digital treatment should be adaptable according to patients' level of knowledge and interest	• "If you [patient] want to delve deeper, then, maybe have something that you can activate and read on if you want more information about it... Maybe some patients already know what pain is and want to go directly do the exercises... make it more efficient." [Therapist 3]
Treatment length might need to be adapted to symptom complexity	• "I think that pain is very complex, and it can take time to implement behavioral changes... if you want to implement a proper behavioral change to something that has been going on for thirty years, six weeks may be a very short time.... It depends on which patient group you think should be included in this [treatment]. Is it really severe pain patients or is it primary care patients with a little milder pain who haven't had pain for so long? Then I think that then it [the 6-week duration] is quite reasonable." [Therapist 5]
No space limitations on patients' response to exercises (ie, word limit)	• "[The patient could] have the possibility of filling in more." [Therapist 10]
Flexibility in frequency of contact	• "With some patients, you maybe have many phone calls while with others you may have nothing in the meantime but a follow-up at the end or something in the middle and then at the end. It can be varied and timewise it is also a bit different." [Therapist 7]

XSL•FO
RenderX

Participant group and subcategory—adaptability	Example quotes
Flexibility in mode of communication (ie, telephone, video, and messaging)	• "Let them choose when we book an appointment: should I call you up or do you want to write in the chat function if you want some feedback on any particular part of it [the exercise]." [Therapist 2]
Reason to provide flexibility in mode of communication—scheduling (patients and therapists)	• "It's better [for one's own structure] to have a clear that this is the feedback day. This is when you can expect to get a response from me.... They know that there is a specific day and that we have a contract at the beginning that this is what the agreement looks like." [Therapist 4]
Reason to provide flexibility in mode of communication—time for preparation (therapists)	• "You have to be very aware of what [the patients] have worked on because it is difficult to go [into the treatment] and have video [the camera on] at the same [while talking online face-to-face with the patient]." [Therapist 8]
Reason to provide flexibility in mode of communication—therapists encourage use of the messaging function (therapists)	• "We always encourage people doing internet-based CBT to write to us in this messaging function.... There are quite a few who never write in that box no matter how much you encourage them. [Therapists could have personal contact regularly]." [Therapist 7]
Reason to provide flexibility in mode of communication—preference for phone calls (patients) and video not being superior to phone calls (therapists)	• "If you choose only chat function then it is very difficult... Video is also an option, if it works. You often waste a lot of time if it's not working, and then phone always works." [Therapist 3]
Asynchronous messaging (not live chat) function	• "I write answers and feedback based on what they [patients] wrote previously, not that it is a live chat." [Therapist 8]

Complexity of the Treatment

Concerning the complexity of the treatment, patient focus groups suggested that the amount of text in the digital treatment could be challenging for some individuals and should be reduced to improve feasibility and user-friendliness. In addition, bullet points and summaries could enhance clarity, and exercise instructions should be presented clearly for better understanding.

Therapists further reported that the treatment could be considered demanding for both patients and therapists.

Design of the Treatment

Within the CFIR construct "design," 2 subthemes were identified: "design elements," which refers to technology, branding, media, text, and aesthetics, and "content elements," which refers to reflections from participants on how the content or treatment elements delivered should be improved or what should be added or removed.

Design Elements of the Treatment

Figure 3 summarizes preferences and suggestions on the design of the digital intervention from patients and therapists.

XSL•FO
RenderX

Figure 3. Summary of preferences and suggestions for the design elements of the DAHLIA digital behavioral intervention for people with chronic pain as identified by the focus group participants (blue boxes=patients [n=5]; orange boxes=therapists [n=12]).

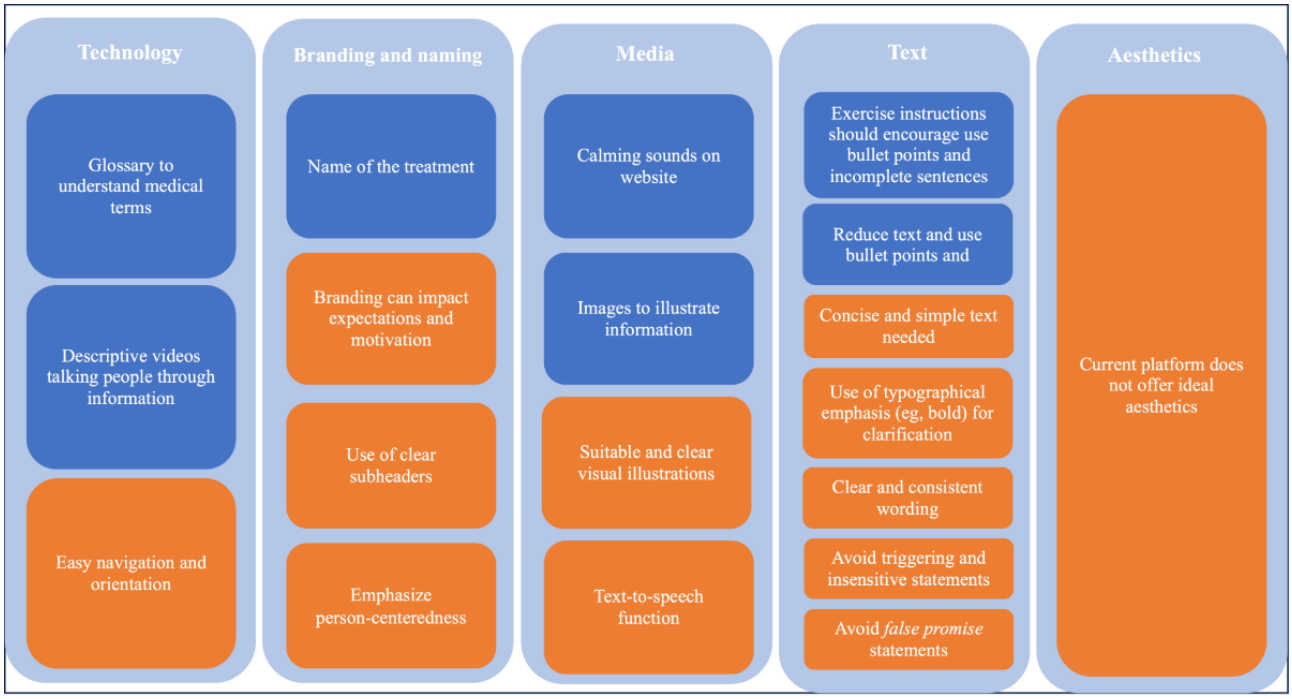

Content Elements of the Treatment

Table 5 presents preferences and suggestions from participants regarding the content of the digital intervention, which is also summarized in this section. The treatment program (prototype 0.0, with suggestions for the full version) was evaluated as well structured by therapists. Therapists suggested having a table of contents and summaries at the beginning of the treatment, including psychoeducation on pain, and inclusive and diverse examples. Moreover, therapists highlighted the importance of the content fitting diverse patient groups, potentially targeting behaviors such as pacing or overcompensating.

Patients suggested that the content, including exercises, should ideally be tailored to individuals' needs, such as involving sleep-related content for people who struggle with insomnia and assessing treatment expectations. Additional material to the treatment content could potentially be offered through external links and further readings. Patients suggested 1 or 2 sessions as a "booster" after the completion of the intervention period as a reminder and relapse prevention. Patients also mentioned the importance of content that focuses on enabling a meaningful life, for instance, focused on how to find things that bring feelings of happiness and contentment in life by identifying things that are important for people and targeting (mourning of) the loss of function and realization of the current capacity. Further details and example quotes are presented in Table 5.

XSL•FO
RenderX

Table 5. Preferences and suggestions from focus group participants (n=5 patients and n=12 therapists) regarding the structure and content elements of the DAHLIA digital behavioral intervention for people with chronic pain.

Focus group and subcategory—design (specifically the content)	Example quotes
Patients	
Sleep-related content	• "I'm thinking more about how to sleep, what you can do in bed, how to relieve things in the best way to get a better sleep." [Participant 3]
Several booster sessions	• "I would recommend two [booster sessions]." [Participant 3]
Asking patients about treatment expectations	• "[Talking about expectations] makes it easier to move on to the next step, because I'm open with what I expect and where I come from, what I've chosen not to do and how it's affected me.... And if you are honest and say that this is what I expect, then the person who meets me can be honest and say this is what we can provide you with." [Participant 1]
Focus on enabling a meaningful life	• "That's what I've been missing. Getting back a little more, meaning to live life instead of surviving it. To be able to find fun things to feel content, happy and satisfied." [Participant 2]
Focus on life values	• "I think it's important to reflect and look at what works for me in everyday life, and what it is that drives me in life. You may sometimes think it's really hard to have a family when dealing with pain, but at the same time, they have been a driving force. They are my motor. I also have my work, which I really enjoy, and I have an understanding employer and all my co-workers know that I am sometimes in pain. And I think it's important to take the time to stop and reflect [though these exercises] and realize, today is a bad day, but I can still make things work." [Participant 5]
Focus on (mourning) loss of function and realizing current capacity	• "You mourn yourself; you've passed away in some way.... You had a hard time finding your purpose again." [Participant 1]
Microsessions likely to be suitable	• "[Microsessions are] just enough. More than that would be too much. It then becomes like homework.... you can squeeze it in here and there... 10 minutes is easier to fit into life." [Participant 1]
Optional extra material	• "[The program is] good as it is but you might want the opportunity to be able to understand a bit more. You [could] link to a page where you can read more about [a certain topic]. There should be some opportunity to delve deeper." [Participant 4]
Body scans and meditative exercises	• "For each day, you might have the option of doing one of those body scans.... For those panic attacks where my body stops working. I couldn't walk for six months and [doing a body scan] was the only thing that got me to learn to walk again. I became aware of my body." [Participant 4]
Reuse of existing content sources	• "I've learned when doing these projects that it's unnecessary to make material that already exists." [Participant 4]
Therapists	
Structured program	• "It feels really good to finally have a structural treatment program." [Therapist 1]
Emphasis on exercise and skill training	• "[Therapists] train the [patients] to observe, and that is a skill to learn, [to] observe nuances as well. Not everything is [about] pain, it might also be other things, [e.g., tiredness]." [Therapist 10]
Mandatory reflections on exercises	• "[Currently,] you can write to your practitioner how you perceived [the exercise] and I think that is valuable information... maybe not leave it optional." [Therapist 1]

XSL•FO
RenderX

Focus group and subcategory—design (specifically the content)	Example quotes
Pros and cons of stating duration and word count per exercise	• "A detail that bothered me a lot, is that it says how long the exercises take, it takes 2 minutes to complete this one [exercise]. I don't understand the purpose of that. I think that each patient will be very individual.... It can still be good to have a short [text] 'here do a short reflection'... should write two sentences or two hundred." [Therapist 1] • "Super great thing that it says four minutes [to do this session] and [the patient might think] yes I have that [time right now], so I think that is good." [Therapist 10]
Microsessions	• "It's good that [the treatment] seems to be very brief.... This whole idea of taking small sections every day instead of longer pieces [like in other iCBT programs] might be a good adaptation for these patients who usually have much lower energy." [Therapist 5]
Importance of first exercise	• "It's good that [the treatment] takes values and goals into account early." [Therapist 5]
Table of contents and summaries	• "It would have been nice to have a bit more of a summary that sums up all these sections.... For the patient to keep track of all these things." [Therapist 5]
Fittin diverse patient groups (ie, pacing vs overcompensating)	• "The material seems to be about a patient [with more] fear of movement, but many of these patients are over compensators who are angry and struggling with their pain thinking 'I won't let the pain win' and they do too much." [Therapist 6]
Inclusive and diverse examples as part of exercises	• "[The patient] can press on [a symbol] and then [they] get more information about an exercise [and] what [they] can answer in this exercise. I think it would be possible to have a few more examples and maybe a few more pain-related examples.... Pain patients can sometimes have difficulty with generalizing. If I [as the therapist give an example about] pain in my knee, then a patient who has back pain will say 'but then this doesn't apply to me.' You often need to have a lot of variations of examples so that the [patients] will be able to recognize themselves." [Therapist 10]
Psychoeducation on pain	• "[The program should] have an explanation [on the] physiology of chronic pain to understand why [is the patient] doing this [program]. Many [patients] can be quite preoccupied with [the idea of] what is wrong with [their] body and finding [a] cause behind it. [the program needs to] explain that there is a very big difference between acute and chronic pain, they are driven by completely different mechanisms." [Therapist 9]
Asking patients to report limitations at the start of treatment	• "At the beginning of the treatment, the patient had to report 'Do I have any limitations, or have I been told that there is something I shouldn't or can't do?' [It is difficult as a therapist] having to play detective and find out and the patient can also have more concerns, but then you can discuss things." [Therapist 6]
Inclusion of more complex ACT[a] processes (eg, self as context)	• "Internet-based treatment is now mature enough to take a step further and include more [ACT] processes... maybe self as a context. It [this process] is really hard to grasp even as a therapist." [Therapist 6]

[a]ACT: acceptance and commitment therapy.

Summary of Key User Feedback

To enhance clarity and facilitate interpretation, a summary table was created to illustrate how example key user feedback was mapped to relevant constructs within the CFIR *intervention characteristics* domain (Table 6).

XSL•FO
RenderX

Table 6. Summary of selected user feedback from focus groups (n=5 patients and n=12 therapists) mapped to Consolidated Framework for Implementation Research (CFIR) constructs—example insights for intervention design.

CFIR construct	User feedback	Explanation
Evidence base	Content was relevant and valid	Evidence-supported content was valued
Relative advantage	Microsession format was suitable	The microsession format was seen as an advantage in comparison to longer sessions
Adaptability	Flexibility in mode of communication	Weekly patient-therapist contact should be adaptable to different communication methods (eg, video or phone calls and messaging)
Complexity	Desire for reduced text and use of bullet points and summaries	Simpler and more concise text was suggested, using bullet points and summaries to improve clarity
Design	Request for visual aids and audio options (eg, images, videos, and text-to-speech function)	The intervention can be more engaging and accessible through multimedia elements

Phase 2 (Testing): End Users' Perceptions of Engaging With the Treatment Prototype and Areas for Improvement

Overview

On the basis of the focus group findings, the intervention prototype version 1.0 was created by 3 researchers with clinical expertise (RKW, LE, and IF). The treatment consisted of 6 modules, including weekly contact with a therapist where patients and therapists together chose the mode of communication (phone or video call) as well as 2 booster sessions (after 2 and 4 months). Each module included 4 microsessions, resulting in a total of 24 microsessions. An overview of 1 module and a screenshot of an exercise are provided in Multimedia Appendix 8.

Each new module was enabled by the therapist after the previous module was completed; while it was suggested that patients complete 1 module per week, flexibility was allowed. Suggestions from the focus groups that were not implemented in the treatment prototype version 1.0 are presented in the End-User Suggestions for Further Treatment Improvement section.

Participant Flow and Characteristics

For phase 2 (testing), 15 patients expressed interest, of whom 4 (27%) were not eligible due to ongoing pain treatment or rehabilitation (n=3, 75%) or brain damage (n=1, 25%). Thus, 11 patients (n=8, 72% from Region Stockholm and n=3, 27% from Region Kalmar) provided informed consent and started the treatment. A total of 18% (2/11) of the participants dropped out, and 82% (9/11) of the patients completed all modules (see Figure 4 for details). Of the participants involved in the focus groups, 17% (2/12) of the therapists and no patients also took part in the pilot-testing, and the remaining pilot participants were newly recruited.

Patients were, on average, aged 46.9 (SD 6.9; range 36-58) years, 18% (2/11) identified as men, and 82% (9/11) identified as women. Educational levels ranged from high school diploma (3/11, 27%) to college or university degree (2/11, 18% ongoing university studies; 6/11, 55% college or university degree). At baseline, numerical rating scale scores (0-10) showed that the current pain level was, on average, 5.1 (SD 2.3; range 3-10) and the mean pain level in the previous week was 5.6 (SD 2; range 3-10). Patients had various diagnoses—migraine, chronic fatigue syndrome, herniated disc, arthritis, fibromyalgia, complex regional pain syndrome, and hypermobile Ehlers-Danlos syndrome—and 9% (1/11) of the participants did not know their pain diagnosis.

In total, 3 licensed psychologists delivered the treatment (n=2, 67% from Region Stockholm and n=1, 33% from Region Kalmar). The therapists were all women (3/3, 100%) and had a mean age of 47.6 (SD 11.06; range 36-58) years, with an average of 21.7 (SD 13.31; range 7-33) years of experience providing psychological therapy, 18.3 (SD 14.6; range 2-30) years of experience working specifically with people with chronic pain, and 12 (SD 9.84; range 1-20) years of experience using acceptance and commitment therapy.

XSL•FO
RenderX

Figure 4. Patient flow in the piloting of the DAHLIA digital behavioral intervention for people with chronic pain (phase 2).

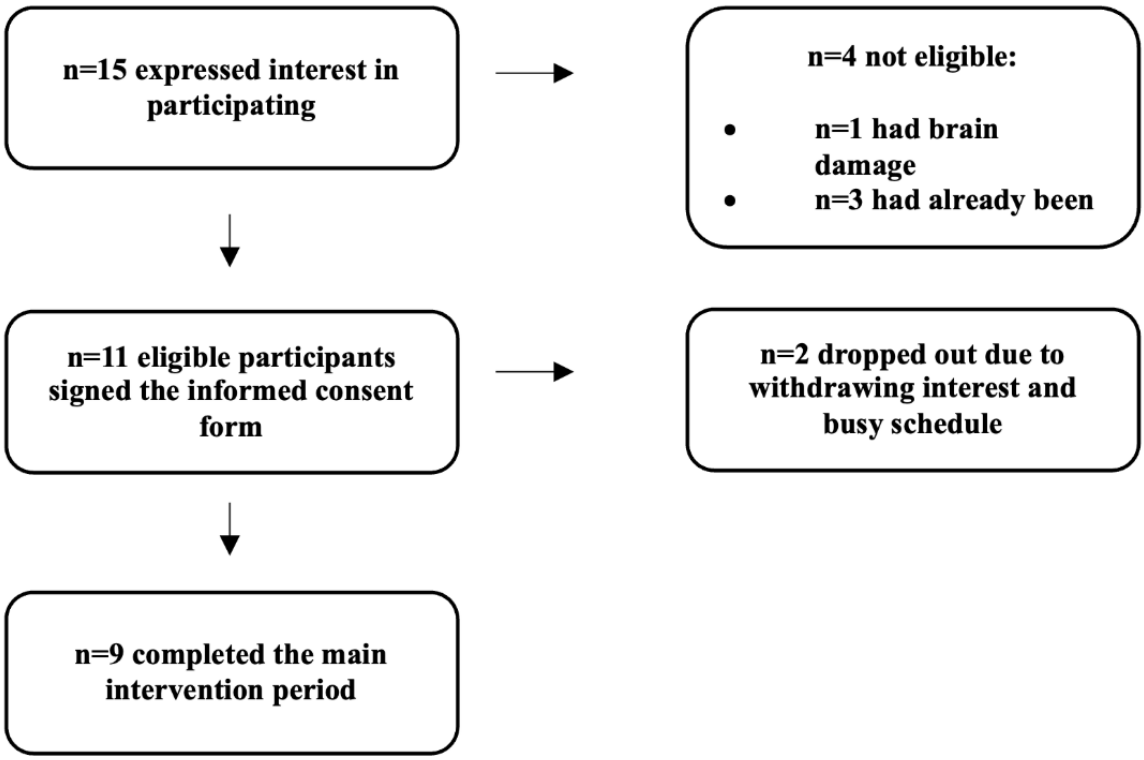

End-User Perceived Success of the Intervention

The success of the intervention refers to the direct reflection and evaluation of engaging with the intervention. As such, patients rated the modules as very helpful (mean 5.1, SD 1.69; range 1-7), moderately enjoyable (mean 4.3, SD 1.66; range 1-7), and very understandable (mean 5.4, SD 1.54; range 1-7). In addition, qualitative feedback was provided. Input varied, resulting in different suggestions for improvement (Table 7).

XSL•FO
RenderX

Table 7. Patient evaluations of the weekly modules during pilot-testing (phase 2) of the DAHLIA digital behavioral intervention for people with chronic pain, including ratings for helpfulness, enjoyability, and understandability, and selected example elaborations[a].

	Helpful		Enjoyable		Understandable	
	Rating, mean (SD; range)	Example quote	Rating, mean (SD; range)	Example quote	Rating, mean (SD; range)	Example quote
Module 1 (n=11)	4.27 (2.05; 1-7)	"...a good start to understand pain, however, a bit more theoretical"	4.81 (2.08; 1-7)	"...fun and motivating to continue"	5.45 (1.80; 2-7)	"...easy to understand"
Module 2 (n=9)	4.89 (2.20; 1-7)	"...already thought about the things"	4.56 (1.94; 1-7)	"...it is good that the module is difficult"	6 (1.30; 3-7)	"...instructions for the exercises were difficult to understand"
Module 3 (n=9)	5.27 (1.19; 3-7)	"...exercises led to a conversation with the therapist which in turn led to insights"	3.81 (1.53; 1-5)	"...the navigation in between exercises messy"	4.45 (1.86; 1-7)	"...highlights should be used for different exercises to make them more distinct"
Module 4 (n=9)	5.45 (2.01; 1-7)	"...difficulty with changing things in combination with the high load in life"	4.09 (1.75; 1-7)	"...satisfying to get new insights"	5.63 (1.36; 3-7)	"...more theoretical rather than practical"
Module 5 (n=8)	5.36 (1.20; 4-7)	"...a bit less concrete than earlier chapters"	4.63 (1.12; 3-7)	__[b]	5.45 (1.50; 2-7)	"...more difficult than the previous modules"
Module 6 (n=7)	5.36 (1.50; 2-7)	"...difficult and overwhelming due to the number of boxes that needed to be filled in each exercise"	3.90 (1.57; 2-6)	"...very difficult to navigate in between exercises"	5.54 (1.43; 2-7)	"...the exercises were massive"
Total	5.1 (1.69; 1-7)	—	4.3 (1.66; 1-7)	—	5.4 (1.54; 1-7)	—

[a]Perceived helpfulness, enjoyability, and understandability were scored on a scale from 1 (not at all) to 7 (very much).

[b]Not available.

Reflecting on the treatment as a whole during the exit interviews, patients considered the written material understandable and the text easily readable in terms of font size and formatting. A total of 4 sessions per week over 6 weeks were considered adequate, and the time needed to complete the sessions was acceptable.

Therapists evaluated the intervention as very beneficial for patients; they were overall satisfied and self-reported that the intervention was delivered as intended. All therapists expressed interest in delivering the intervention again in the future. The overall time investment to deliver the intervention was considered acceptable, whereas the time saved by delivering this treatment online was perceived as moderate. Details of patients' and therapists' perceived intervention success are provided in Table 8.

XSL•FO
RenderX

Table 8. Exit interview questions for patients (n=9) and therapists (n=3) on "reflecting and evaluating" the DAHLIA digital behavioral intervention for people with chronic pain following pilot-testing[a].

Question	Rating, mean (SD; range)	Example quotes
Patient feedback (n=9)		
"Did you experience the online treatment as helpful overall?"	5.75 (2.05; 1-7)	"I really think so. Mainly thanks to the sessions with my therapist. They were invaluable. Without them I am unsure of the value of the program." [Patient D]
"Did you experience the online treatment as meaningful overall?"	5.88 (1.72; 2-7)	__[b]
"Was the written material understandable?"	5.50 (1.30; 3-7)	"First chapters were a bit tricky to understand, talked to my therapist and it got better." [Patient F]
"Could you easily read the text in the treatment (ie, in terms of font size and formatting)?"	5.88 (1.80; 3-7)	"Font size and formatting were OK, but there could be bold and italic text formatting or different colors can be used to link parts together better. And [there should be] clickable links instead of having to go back and forth, that was messy! Extra difficult to do in the phone. Text-to-speech function would be an improvement." [Patient A]
"Was the number of sessions per week (4 sessions) adequate?"	5.50 (1.41; 4-7)	"A good number of sessions per week, and you could do them more or less thoroughly, you could adjust to what is needed for you, personally." [Patient B]
"Was the total number of sessions adequate?"	6.38 (0.91; 5-7)	"In general, six weeks felt adequate. During this specific period, it was a bit stressful. Don't think that it is possible to do it in a shorter amount of time. But think that 6-8 weeks is optimal." [Patient D]
"Was the time needed to complete the sessions acceptable?"	5.83 (1.16; 4-7)	"Sometimes it took a very long time, depending on how much writing I have done, or how much pain I have." [Patient C]
"Did microsessions influence your behavior in everyday life?"	4.88 (2.58; 1-7)	"[The treatment] made me think about how I do things and how I can improve." [Patient F]
"Did microsessions influence your emotions?"	4.75 (1.90; 2-7)	"[The treatment] made me think about things, in the back of my head." [Patient D]
"Did microsessions influence your thoughts?"	5.38 (1.30; 3-7)	"I become more aware, was able to see myself from a different perspective." [Patient C]
Therapist feedback (n=3)		
"Was the online treatment overall beneficial for your patients?"	6.33 (0.58; 6-7)	"For the majority of them it was very beneficial in different ways." [Therapist B]
"How satisfied are you with the intervention overall?"	5.67 (0.58; 5-6)	"I like that it is 6 weeks, quite easy to understand. Seems to be inspiring for the patient. The therapist has to talk to patient every week and has to be prepared for the questions." [Therapist A]
"Was the overall time investment to deliver the online treatment acceptable?"	5.67 (1.15; 5-7)	"It was well, and 6 weeks treatment is short, but also enough." [Therapist A]
"Did delivering this treatment online save you time?"	3.67 (2.08; 2-6)	"Documentation and contact with participants take time similar to that of which I had seen them in person. However, [there is] greater flexibility in the digital format." [Therapist B]
"Was the online treatment delivered as intended?"	7 (0; 7-7)	"I was lucky with my participants, they did what they were expected. It may not always be the case, everyone got the treatment in the intended time." [Therapist A]
"Would you deliver the intervention again in the future?"	7 (0), 7-7	—
"What facilitated you to deliver the intervention?"	N/A[c]	"The protocol and the documentations helped a lot, contact with the supervisors helped a lot." [Therapist B]
"What hindered you in delivering the intervention?"	N/A	"Nothing that I can think of, but it can be messy to set up time with the patients." [Therapist C]
"What aspects of the intervention need improvements?"	N/A	"In this stage, module 3 should be clarified. All of my patients got stuck there, and did not understand. It seems like it is unclear that active pain patients have recommendations about being active, but they are already too active." [Therapist A]

[a]Responses were scored on a scale from 1 (not at all) to 7 (very much).

[b]Not available.

XSL•FO
RenderX

[c]N/A: not applicable.

End-User Perceived Implementation Success

The success of the implementation refers to the engagement with the wider treatment process (beyond the direct engagement with the treatment). As such, patients found the information regarding the use of the digital platform clear, experienced few technical problems, and considered the navigation of the digital platform easy. The weekly communication with the therapist was evaluated as very helpful, and the weekly meeting with the therapist was perceived as easy to schedule. Therapist contact was considered a very motivating experience, and patients mentioned that they felt supported by their therapist.

Similarly, therapists did not report any technical problems and found the platform very easy to navigate and the support for delivering the intervention, namely, training, technical guidance, and supervision, sufficient. The frequency of communication with patients and the time per interaction were very acceptable. Details on the end-user perceived implementation success are provided in Table 9.

XSL•FO
RenderX

Table 9. Exit interview questions for patients (n=9) and therapists (n=3) on "reflecting and evaluating" the implementation of the DAHLIA digital behavioral intervention for people with chronic pain following pilot-testing (phase 2)a.

Question	Rating, mean (SD; range)	Example quotes
Patient feedback (n=9)		
"Did the online treatment interfere with your daily routines (work or other things)?"	3.88 (2.16; 1-7)	__[b]
"The treatment was delivered using a digital platform on 1177. Was the information about the digital 1177 platform clear?"	6.43 (0.53; 6-7)	"I was clearly guided during the briefing call before the intervention." [Patient H]
"Was it easy to navigate the digital 1177 platform?"	5.38 (1.59; 3-7)	"Yes, but a bit difficult to find your way back to earlier exercises." [Patient E]
"Did you experience any technical problems using 1177?"	0.88 (0.35; 0-1)	—
"Did you experience communicating with your health care professional as helpful overall?"	6.50 (0.92; 5-7)	"Absolutely, it was what I needed. Without it I do not know how much progress I would have made. Very meaningful!" [Patient D]
"Was it easy to schedule meetings with your health care professional?"	6.75 (0.70; 5-7)	"Very easy, my therapist was very helpful and flexible." [Patient C]
"Did you experience communicating with your health care professional as motivating?"	6.50 (0.92; 5-7)	"Yes, my therapist has really made me think about things and given me new ideas." [Patient F]
"Did you feel supported by your health care professional?"	6.50 (1.41; 3-7)	"Yes, in a way that I didn't expect. I didn't think that it would give so much and much of it was my therapist's merit." [Patient A]
"Would you consider the past 6 weeks 'ordinary'?"	4.50 (1.77; 1-6)	—
"Did anything unusual occur during the treatment period?"	2.38 (2.26; 1-6)	—
"Would you recommend this online treatment to a friend with a similar condition?"	6.38 (1.18; 4-7)	"Already talked to a friend and recommended her." [Patient F]
Therapist feedback (n=3)		
"Did you experience any technical problems using 1177?"	1 (0; 1-1)	—
"Was it easy to navigate 1177?"	7 (0; 7-7)	—
"Was the support for delivering the intervention (eg, training, technical guidance when issues arose, supervision) sufficient?"	7 (0; 7-7)	"I did not have any problems, but I felt very secure, it was not hard for me to receive help." [Therapist C]
"Was the frequency of communication with the patient acceptable?"	7 (0; 7-7)	"It would be very time-consuming with so many patients, e.g. in primary care. It is not very frequent for written feedback but when you speak with them it is a little intense." [Therapist A]
"Was the time per interaction (eg, phone call) acceptable?"	5.66 (1.15; 5-7)	"I thought 30 minutes [per interaction] would be tight, but on the contrary 30 minutes was perfect. Of course, some of them [the patients] exceeded but in general it was within 30 minutes." [Therapist B]
"Did you feel prepared to deliver this treatment?"	6.33 (1.15; 5-7)	—
"Would you recommend the intervention to a colleague?"	5.67 (1.53; 4-6)	"I think it is a bit too early to say, but I am also drafting a couple of colleagues so we can say that I am already recommending; the intervention needs further elaboration." [Therapist A]

[a]Responses were scored on a scale from 1 (not at all) to 7 (very much).

[b]Not available.

XSL•FO
RenderX

End-User Suggestions for Further Treatment Improvements

When developing and improving the treatment, not all end-user suggestions could be integrated due to practical and technical reasons. The research team prioritized adaptations that could be addressed given the available time and financial resources. For instance, creating video materials is time-consuming and will benefit from collaboration with communication experts. Similarly, no automatic text-to-speech function is currently available in the platform, and alternatives need to be explored prospectively. Thus, taking all end-user input together, the following additional suggestions for improvements emerged in this study and require further refinement and testing to meet user needs: (1) creating video content and adding pictures or graphics to make the treatment more visually appealing; (2) integrating a text-to-speech function; (3) creating and integrating add-on materials, for instance, methods to improve sleep; and (4) facilitating navigation in between exercises. Moreover, according to participants' views, it will be important to test whether the communication mode, namely, messaging only compared to phone or video calls, contributes to intervention and implementation success as user preferences differed. To illustrate how end-user feedback informed changes throughout the development of the DAHLIA intervention, Figure 5 provides a visual summary of feedback integration across prototypes.

Figure 5. Integration of end-user feedback (patients and therapists) across prototypes in the development process of the DAHLIA digital behavioral intervention for people with chronic pain, highlighting how end-user feedback informed design adaptations over time.

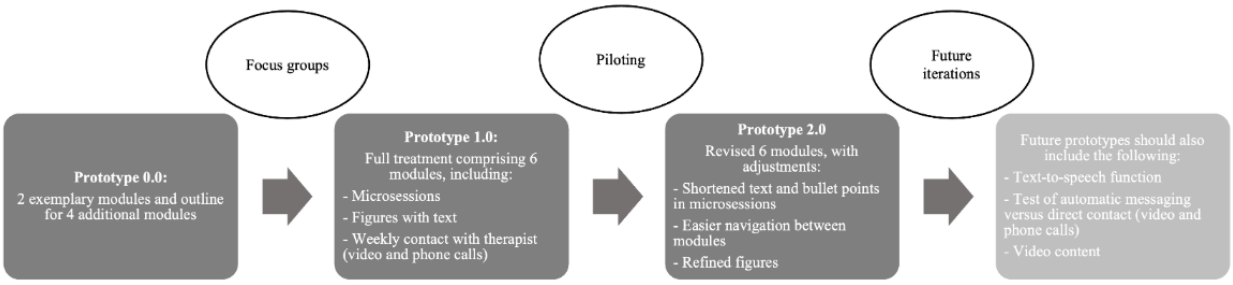

Discussion

Principal Findings

This study aimed to clarify how a digital behavioral intervention can be developed through a user-centered approach to address the needs and preferences of the target population and the specific user input on the DAHLIA prototype 1.0 as part of the multiphase DAHLIA project. This paper presents data from the first iteration, which is considered part of the development phase. The following iterations are included in the evaluation phase, will be used to assess the feasibility and preliminary efficacy of the treatment, and will be presented elsewhere. The primary objective of this study was to develop a user-centered and evidence-based digital behavioral intervention for individuals with chronic pain by clarifying how the intervention should be designed in terms of structure, content, and format to address the needs and preferences of the target population.

This study involved 3 phases: preparation, design, and testing. The user-centered approach during the preparation phase was built on fictional patient personas, resulting in a representation of heterogeneous patient characteristics, needs, and potential treatment targets that shaped the treatment vision. During the design phase, input from end users via focus groups provided insights on the treatment content, design, and structure, with an emphasis on flexibility and person-centeredness. In the testing phase, 11 patients and 3 therapists participated, and findings confirmed that a behavioral approach was considered suitable for individuals with chronic pain. In addition, digital delivery of the intervention using a microsession format combined with regular contact between therapists and patients was seen as beneficial by end users as the modules were overall perceived as helpful, understandable, and enjoyable. End users rated their experiences using the intervention as good or excellent, indicating initial satisfaction with both the treatment and wider implementation procedure, such as the use of digital platforms for delivery as well as research-related informed consent and data collection.

Finally, end users provided suggestions for further improving the design and content, such as a text-to-speech function, additional content for specific needs such as insomnia, and aesthetic adjustments through images and videos.

End-User Involvement: Complementary and Contradictory Views

Overview

Critically reflecting on the input provided by end users regarding the utility of the development process and the DAHLIA prototype 1.0 is important for outlining the next steps in the DAHLIA project, namely, the further treatment improvements, testing, and implementation [30]. While the existing literature does not provide specific information on the recommended level of involvement of end users [50], user engagement is generally encouraged [20,51] and is considered a cornerstone in health care innovation [52]. Although many studies have focused on a single end-user group, such as patients or therapists, Rosser and Eccleston [53] emphasize the utility of including health care providers alongside patients during the development of a novel treatment [54].

However, while combining views from different stakeholders may be beneficial, it can also result in challenges. In this study, end-user input varied, with both consensus and discrepancies between individuals and end-user groups regarding preferences in treatment design. For instance, some patients preferred group-based treatment, whereas others favored individual treatment as it provided opportunities to choose their own pace.

XSL•FO
RenderX

Moreover, patients mentioned that digital delivery may be suitable for some but can be a barrier to engage for specific groups such as older adults with chronic pain, which is in line with research highlighting that the digitalization of health care systems may result in the exclusion especially of those who are older and of a lower socioeconomic status [55]. These differing views emerged during phase 1 (design).

Importantly, patients' and therapists' preferences regarding the mode of communication differed. While patients generally expressed a preference for video calls over phone calls, therapists viewed phone calls as similarly useful as video calls and suggested that patients should be able to choose. This finding is in line with results from a systematic review aimed at examining the attitudes toward video- and phone-based telehealth [56]. Even though some studies presented no significant differences between patients' and health care providers' attitudes toward the mode of communication [57], others preferred video calls, and only a small number preferred phone calls [58]. This lack of consensus in the field underlines the importance of flexibility and adapting the mode of communication to patients', and potentially also therapists', preferences and needs. In the testing phase, therapists and patients decided together how to communicate, and the number of phone calls and video meetings was nearly equal, with phone calls being slightly more preferred. Future studies can empirically compare modes of communication to explore, for instance, whether dropout rates or user experiences vary based on this treatment feature. This flexibility was provided in phase 2, pilot-testing, where therapists and patients decided together on the mode of communication, aiming to incorporate flexibility and personalization.

Examining treatment content and design, both patients and therapists perceived a need to condense the text reduce the risk of the treatment being (too) challenging for some patients and provided helpful suggestions, such as the use of bullet points, overviews, or summaries. In addition, they highlighted the importance of incorporating images into the intervention to illustrate information and a text-to-speech function to enhance accessibility. This feedback led to the incorporation of bullet points, summaries, images, and audio files in the full prototype. While a built-in text-to-speech function is not supported by the 1177 platform, we plan to include this feature in future iterations through an additional solution in collaboration with IT developers. Overall, during the development and adaptation of a digital treatment, researchers face the challenge to balance "ideal" and "feasible" ideas as some intervention components may require resources (ie, time, technical features, funding, and expertise) that are not available.

In the testing phase, end-user experiences also varied between individuals. While therapists' ratings were rather similar, patients' experiences showed larger differences. In the weekly evaluations, where patients evaluated each module, and in the exit interviews, where they evaluated the intervention in general, they provided ratings that spanned almost the entire scale (from 1 to 7), indicating that individual differences may influence how the treatment is perceived. This finding points to a need to balance the end-user voices with practical requirements during the development process to optimize feasibility. Overall, the structure of this study, initially from conceptual preparation to collaborative design and, finally, to testing, enabled a gradual integration of feedback, allowing for tailoring the intervention to end-user needs.

Heterogeneity of the Target Population: One Size Does Not Fit All

A key challenge in designing a digital intervention is meeting varying individual needs [59]. Therefore, for the preparation phase of development, patient personas of different age groups, comorbidities, and pain experiences were created as potential representatives of the target population, aiming to amplify patient perspectives early on. Although seemingly useful, The application of patient personas in designing novel pain treatments is not standardized [37]. Groos et al [60] have suggested that different types of patient personas can be insightful for researchers while deciding the next steps of the development process. In the DAHLIA project, experiences from the preparation phase suggest that patient personas can also be helpful in the continuous refinement of the treatment or in tailoring the treatment to different patient groups to initially consider important characteristics such as disabilities, socioeconomic status, language skills, or cultural background that may reflect varying needs.

The heterogeneity in the chronic pain population is well known [1], and during the testing phase, patients with various pain profiles, such as migraine, chronic fatigue syndrome, herniated disc, arthritis, fibromyalgia, complex regional pain syndrome, and hypermobile Ehlers-Danlos syndrome, provided valuable input.

Therapists suggested that treatment should be tailored according to the patients' needs and highlighted the importance of providing inclusive and diverse examples as part of exercises within the treatment. In a consensus statement that aims to provide a practical guideline for researchers in the field of e–mental health, Seiferth et al [23] emphasized the necessity of considering the target group when deciding the structure (eg, exercises and division into modules) and the complexity of the content. Similarly, Evangelista et al [61] highlighted the significance of participatory involvement of those who might be considered minority groups during the design of an intervention to better determine and meet their needs.

To address the varied preferences and needs of people with chronic pain, the intervention was made flexible and adaptable in its content and structure. The intervention was delivered over a 6-week period through microsessions that participants could complete at their own pace. Participants were able to choose the mode of communication (eg, text, video call, and phone call) for the weekly contact with their therapists. The content and exercises were not specific to a single group with chronic pain (eg, patients on sick leave) but were instead developed with a transdiagnostic approach (based on acceptance and commitment therapy and CBT principles) to ensure that the content was relevant to a wider range of people with chronic pain. In future iterations of the DAHLIA intervention, we aim to incorporate optional add-on modules (eg, sleep-related content) tailored to specific needs.

XSL•FO
RenderX

Strengths and Limitations

By involving both patients and therapists in the development of the initial treatment version, this study adopted a user-centered approach that adapts the intervention to end users' needs and preferences. In addition, this study had a strong emphasis on empirical data and various methodologies, integrating qualitative feedback and quantitative ratings. Patient personas were used to create awareness of potential target populations and their diverse needs. The treatment was tested in heterogeneous samples, providing insights into applicability across various demographics. In addition, the integration of established frameworks such as the CFIR strengthened the design.

Several limitations need to be considered when interpreting the results. First, a larger sample of end users would have provided more information and potentially a more diverse group. There was a discrepancy between the number of participants involved in patient (n=5) and therapist (n=12) focus groups. The targeted sample size for patients could not be reached as interest was low, 2 patients dropped out, and the project period did not allow for further recruitment. Even though the number of dropouts was low in this study, the small number of noncompleters limited our ability to analyze potential differences in demographics or engagement. Future iterations with larger samples will allow for more robust dropout analyses. While more therapists than anticipated participated, the limited number of patients raised the risk that data saturation was not achieved, particularly given the goal of capturing a heterogeneous population. This remains a limitation, although the enrollment of 11 patients in the pilot test may have strengthened patient voices. Second, more female than male participants took part in the study. A total of 5 patients and 12 therapists took part in the focus groups, most of whom were female (n=4, 80% of the patients and n=9, 75% of the therapists). In the pilot phase, 11 patients and 3 therapists participated, and they were mostly female (n=9, 82% of the patients and n=3, 100% of the therapists), which is consistent with similar studies [26,62]. While this gender distribution aligns with the higher prevalence of chronic pain in women [63,64] and the female dominance in psychology and psychotherapy professions [65], this imbalance should be considered when interpreting the findings as it may have influenced the feedback and preferences on the intervention design and content. Future studies should include more male end users and explore potential gender differences in preferences and needs.

In addition, the narrow range of pain diagnoses may limit the generalizability of the results. While there was a diversity of pain diagnoses among patients in the pilot test, only 40% (2/5) of the focus group participants had received a specific diagnosis (Ehlers-Danlos syndrome; 3/5, 60% were undiagnosed or did not know the diagnosis). Although the inclusion criteria aimed to increase heterogeneity, this limited diagnostic representation may have constrained the variety of preferences and needs reflected and may result in an intervention that aligns more with the experiences of certain chronic pain populations. Previous research shows that people with different chronic pain conditions can vary in treatment expectations [66], highlighting the need for representativeness and inclusivity when tailoring interventions for different subgroups. Therefore, future iterations of the DAHLIA project should ensure that the intervention aligns with the needs of the broader pain population.

This limited diversity suggests that the findings may not fully represent a broader population, highlighting the need to examine inclusivity in future research. Incorporating diverse groups—considering gender identity, sexual orientation, age, race and ethnicity, diagnosis, and medical and psychiatric comorbidities—can provide broader end-user feedback. To address this limitation, future iterations of the DAHLIA project will implement a broader recruitment strategy by expanding outreach to hospitals and clinics across different regions of Sweden and leveraging social media and online platforms (eg, websites and podcasts). These efforts aim to increase sample size and enhance demographic diversity regarding gender, age, and pain diagnosis. In addition, while the intervention is currently available only in Swedish, it can be culturally adapted and delivered in additional languages to improve accessibility [67,68].

On the basis of the development and initial testing described in this paper, the next steps of the DAHLIA study will include further refinement building on the pilot results, followed by an iterative optimization study and a randomized controlled trial [30]. These upcoming phases will assess the DAHLIA intervention's feasibility, acceptability, and implementation potential in a larger scale. The insights from this study will inform the future phases, aiming to ensure that the intervention aligns with user preferences and needs.

Conclusions

This study aimed at presenting the user-centered development of a digital behavioral intervention for chronic pain, the DAHLIA prototype 1.0. The results illustrate the utility of patient personas when preparing, of focus groups when designing, and of end-user feedback when testing this new intervention. The findings indicated that the treatment holds promise and provided relevant end-user suggestions to guide further improvements.

Acknowledgments

The authors would like to thank Emelie Albertsson, Hanna Solstrand, Haya Al Sharaa, and Oriana Kolokotronis Johansson for their valuable contributions in transcribing the focus groups and Anna Norrenge and Julia Ström for their valuable contributions in designing the intervention for the 1177 platform. This work was funded by Afa Insurance (grant Dnr 190 252).

XSL•FO
RenderX

Data Availability

The datasets generated or analyzed during this study are not publicly available due to data privacy but are available from the corresponding author (AST) on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Guidance for the Reporting of Intervention Development checklist, a guideline for reporting intervention development studies, and Template for Intervention Description and Replication checklist.
[DOCX File , 23 KB - formative_v9i1e74064_app1.docx]

Multimedia Appendix 2
One of the 3 patient personas used in the development phase of the DAHLIA project (Swedish name: Göran).
[DOCX File , 542 KB - formative_v9i1e74064_app2.docx]

Multimedia Appendix 3
Semistructured focus group guide.
[DOCX File , 18 KB - formative_v9i1e74064_app3.docx]

Multimedia Appendix 4
Consolidated Framework for Implementation Research construct names and their definitions from the "Innovation Characteristics" (refers to the "digital intervention") domain applied in the user-centered development phase of the DAHLIA digital behavioral intervention for people with chronic pain in Sweden.
[DOCX File , 18 KB - formative_v9i1e74064_app4.docx]

Multimedia Appendix 5
Semistructured exit interview guide to evaluate the general feasibility and acceptability of the treatment for patients.
[DOCX File , 21 KB - formative_v9i1e74064_app5.docx]

Multimedia Appendix 6
Semistructured exit interview guide to evaluate the general feasibility and acceptability of the treatment for health care professionals.
[DOCX File , 19 KB - formative_v9i1e74064_app6.docx]

Multimedia Appendix 7
Consolidated Framework for Implementation Research construct names and their definitions from the Implementation Process domain applied in the piloting of the DAHLIA digital behavioral intervention for people with chronic pain in Sweden.
[DOCX File , 17 KB - formative_v9i1e74064_app7.docx]

Multimedia Appendix 8
Examples (screenshots) of the prototype of the DAHLIA digital behavioral intervention for people with chronic pain.
[DOCX File , 223 KB - formative_v9i1e74064_app8.docx]

References

1. Zimmer Z, Fraser K, Grol-Prokopczyk H, Zajacova A. A global study of pain prevalence across 52 countries: examining the role of country-level contextual factors. Pain 2022 Sep 01;163(9):1740-1750 [FREE Full text] [doi: 10.1097/j.pain.0000000000002557] [Medline: 35027516]
2. Breivik H, Borchgrevink PC, Allen SM, Rosseland LA, Romundstad L, Hals EK, et al. Assessment of pain. Br J Anaesth 2008 Jul;101(1):17-24 [FREE Full text] [doi: 10.1093/bja/aen103] [Medline: 18487245]
3. Vlaeyen JW, Crombez G, Linton SJ. The fear-avoidance model of pain. Pain 2016 Aug;157(8):1588-1589. [doi: 10.1097/j.pain.0000000000000574] [Medline: 27428892]
4. McCracken LM. Psychological flexibility, chronic pain, and health. Annu Rev Psychol 2024 Jan 18;75(1):601-624 [FREE Full text] [doi: 10.1146/annurev-psych-020223-124335] [Medline: 37585667]
5. Hayes SC, Villatte M, Levin M, Hildebrandt M. Open, aware, and active: contextual approaches as an emerging trend in the behavioral and cognitive therapies. Annu Rev Clin Psychol 2011 Apr 27;7(1):141-168. [doi: 10.1146/annurev-clinpsy-032210-104449] [Medline: 21219193]

XSL•FO
RenderX

6. Vowles KE, Witkiewitz K, Levell J, Sowden G, Ashworth J. Are reductions in pain intensity and pain-related distress necessary? An analysis of within-treatment change trajectories in relation to improved functioning following interdisciplinary acceptance and commitment therapy for adults with chronic pain. J Consult Clin Psychol 2017 Feb;85(2):87-98. [doi: 10.1037/ccp0000159] [Medline: 27991806]
7. Wicksell RK, Melin L, Lekander M, Olsson GL. Evaluating the effectiveness of exposure and acceptance strategies to improve functioning and quality of life in longstanding pediatric pain--a randomized controlled trial. Pain 2009 Feb;141(3):248-257. [doi: 10.1016/j.pain.2008.11.006] [Medline: 19108951]
8. Lai L, Liu Y, McCracken LM, Li Y, Ren Z. The efficacy of acceptance and commitment therapy for chronic pain: a three-level meta-analysis and a trial sequential analysis of randomized controlled trials. Behav Res Ther 2023 Jun;165:104308. [doi: 10.1016/j.brat.2023.104308] [Medline: 37043967]
9. Darnall BD. Psychological treatment for chronic pain: improving access and integration. Psychol Sci Public Interest 2021 Sep 20;22(2):45-51 [FREE Full text] [doi: 10.1177/15291006211033612] [Medline: 34541966]
10. Subbiah V. The next generation of evidence-based medicine. Nat Med 2023 Jan 16;29(1):49-58. [doi: 10.1038/s41591-022-02160-z] [Medline: 36646803]
11. Michie S, Yardley L, West R, Patrick K, Greaves F. Developing and evaluating digital interventions to promote behavior change in health and health care: recommendations resulting from an international workshop. J Med Internet Res 2017 Jun 29;19(6):e232 [FREE Full text] [doi: 10.2196/jmir.7126] [Medline: 28663162]
12. Dziallas M, Blind K. Innovation indicators throughout the innovation process: an extensive literature analysis. Technovation 2019 Feb;80-81:3-29. [doi: 10.1016/j.technovation.2018.05.005]
13. Yardley L, Spring BJ, Riper H, Morrison LG, Crane DH, Curtis K, et al. Understanding and promoting effective engagement with digital behavior change interventions. Am J Prev Med 2016 Nov;51(5):833-842. [doi: 10.1016/j.amepre.2016.06.015] [Medline: 27745683]
14. van Gemert-Pijnen JE, Nijland N, van Limburg M, Ossebaard HC, Kelders SM, Eysenbach G, et al. A holistic framework to improve the uptake and impact of eHealth technologies. J Med Internet Res 2011 Dec 05;13(4):e111 [FREE Full text] [doi: 10.2196/jmir.1672] [Medline: 22155738]
15. Schouten MJ, Derksen ME, Dekker JJ, Goudriaan AE, Blankers M. Preferences of young adults on the development of a new digital add-on alcohol intervention for depression treatment: a qualitative study. Internet Interv 2023 Sep;33:100641 [FREE Full text] [doi: 10.1016/j.invent.2023.100641] [Medline: 37559821]
16. de Beurs D, van Bruinessen I, Noordman J, Friele R, van Dulmen S. Active involvement of end users when developing web-based mental health interventions. Front Psychiatry 2017 May 03;8:72 [FREE Full text] [doi: 10.3389/fpsyt.2017.00072] [Medline: 28515699]
17. Baker TB, Gustafson DH, Shah D. How can research keep up with eHealth? Ten strategies for increasing the timeliness and usefulness of eHealth research. J Med Internet Res 2014 Feb 19;16(2):e36 [FREE Full text] [doi: 10.2196/jmir.2925] [Medline: 24554442]
18. Higgins KS, Tutelman PR, Chambers CT, Witteman HO, Barwick M, Corkum P, et al. Availability of researcher-led eHealth tools for pain assessment and management: barriers, facilitators, costs, and design. PAIN Reports 2018 Sep;3(7):e686. [doi: 10.1097/pr9.0000000000000686]
19. Slater H, Campbell JM, Stinson JN, Burley MM, Briggs AM. End user and implementer experiences of mHealth technologies for noncommunicable chronic disease management in young adults: systematic review. J Med Internet Res 2017 Dec 12;19(12):e406 [FREE Full text] [doi: 10.2196/jmir.8888] [Medline: 29233804]
20. Keogh E, Rosser BA, Eccleston C. e-Health and chronic pain management: current status and developments. Pain 2010 Oct;151(1):18-21. [doi: 10.1016/j.pain.2010.07.014] [Medline: 20674174]
21. Jessen S, Mirkovic J, Ruland CM. Creating gameful design in mHealth: a participatory co-design approach. JMIR Mhealth Uhealth 2018 Dec 14;6(12):e11579 [FREE Full text] [doi: 10.2196/11579] [Medline: 30552080]
22. Ledel Solem IK, Varsi C, Eide H, Kristjansdottir OB, Mirkovic J, Børøsund E, et al. Patients' needs and requirements for eHealth pain management interventions: qualitative study. J Med Internet Res 2019 Apr 01;21(4):e13205 [FREE Full text] [doi: 10.2196/13205] [Medline: 30877780]
23. Seiferth C, Vogel L, Aas B, Brandhorst I, Carlbring P, Conzelmann A, et al. How to e-mental health: a guideline for researchers and practitioners using digital technology in the context of mental health. Nat Mental Health 2023 Aug 07;1(8):542-554. [doi: 10.1038/s44220-023-00085-1]
24. Skivington K, Matthews L, Simpson SA, Craig P, Baird J, Blazeby JM, et al. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. BMJ 2021 Sep 30;374:n2061. [doi: 10.1136/bmj.n2061] [Medline: 34593508]
25. Stinson JN, Lalloo C, Harris L, Isaac L, Campbell F, Brown S, et al. iCanCope with Pain™: user-centred design of a web- and mobile-based self-management program for youth with chronic pain based on identified health care needs. Pain Res Manag 2014 Jan;19(5):257-265 [FREE Full text] [doi: 10.1155/2014/935278] [Medline: 25000507]
26. Ledel Solem IK, Varsi C, Eide H, Kristjansdottir OB, Børøsund E, Schreurs KM, et al. A user-centered approach to an evidence-based electronic health pain management intervention for people with chronic pain: design and development of EPIO. J Med Internet Res 2020 Jan 21;22(1):e15889 [FREE Full text] [doi: 10.2196/15889] [Medline: 31961331]

XSL•FO
RenderX

27. Jibb LA, Cafazzo JA, Nathan PC, Seto E, Stevens BJ, Nguyen C, et al. Development of a mHealth real-time pain self-management app for adolescents with cancer: an iterative usability testing study [formula: see text]. J Pediatr Oncol Nurs 2017 Apr 04;34(4):283-294. [doi: 10.1177/1043454217697022] [Medline: 28376666]
28. Wilson K, Bell C, Wilson L, Witteman H. Agile research to complement agile development: a proposal for an mHealth research lifecycle. NPJ Digit Med 2018 Sep 13;1(1):46 [FREE Full text] [doi: 10.1038/s41746-018-0053-1] [Medline: 31304326]
29. de Korte EM, Wiezer N, Janssen JH, Vink P, Kraaij W. Evaluating an mHealth app for health and well-being at work: mixed-method qualitative study. JMIR Mhealth Uhealth 2018 Mar 28;6(3):e72 [FREE Full text] [doi: 10.2196/mhealth.6335] [Medline: 29592846]
30. Bartels SL, Johnsson SI, Boersma K, Flink I, McCracken LM, Petersson S, et al. Development, evaluation and implementation of a digital behavioural health treatment for chronic pain: study protocol of the multiphase DAHLIA project. BMJ Open 2022 Apr 15;12(4):e059152 [FREE Full text] [doi: 10.1136/bmjopen-2021-059152] [Medline: 35428645]
31. Duncan E, O'Cathain A, Rousseau N, Croot L, Sworn K, Turner KM, et al. Guidance for reporting intervention development studies in health research (GUIDED): an evidence-based consensus study. BMJ Open 2020 Apr 08;10(4):e033516 [FREE Full text] [doi: 10.1136/bmjopen-2019-033516] [Medline: 32273313]
32. Hoffmann TC, Glasziou PP, Boutron I, Milne R, Perera R, Moher D, et al. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. BMJ 2014 Mar 07;348(mar07 3):g1687 [FREE Full text] [doi: 10.1136/bmj.g1687] [Medline: 24609605]
33. Cordier L, Diers M. Learning and unlearning of pain. Biomedicines 2018 Jun 05;6(2):67 [FREE Full text] [doi: 10.3390/biomedicines6020067] [Medline: 29874854]
34. Goubert L, Trompetter H. Towards a science and practice of resilience in the face of pain. Eur J Pain 2017 Sep 02;21(8):1301-1315. [doi: 10.1002/ejp.1062] [Medline: 28573783]
35. Gentili C, Rickardsson J, Zetterqvist V, Simons LE, Lekander M, Wicksell RK. Psychological flexibility as a resilience factor in individuals with chronic pain. Front Psychol 2019;10:2016 [FREE Full text] [doi: 10.3389/fpsyg.2019.02016] [Medline: 31551871]
36. Pruitt J, Adlin T. The Persona Lifecycle: Keeping People in Mind Throughout Product Design. Burlington, MA: Morgan Kaufmann Publishers; 2010.
37. Bartels SL, Taygar AS, Johnsson SI, Petersson S, Flink I, Boersma K, et al. Using personas in the development of ehealth interventions for chronic pain: a scoping review and narrative synthesis. Internet Interv 2023 Apr;32:100619 [FREE Full text] [doi: 10.1016/j.invent.2023.100619] [Medline: 37273935]
38. Gentili C, Zetterqvist V, Rickardsson J, Holmström L, Simons LE, Wicksell RK. ACTsmart - development and feasibility of digital acceptance and commitment therapy for adults with chronic pain. NPJ Digit Med 2020;3:20 [FREE Full text] [doi: 10.1038/s41746-020-0228-4] [Medline: 32128450]
39. Breivik H, Collett B, Ventafridda V, Cohen R, Gallacher D. Survey of chronic pain in Europe: prevalence, impact on daily life, and treatment. Eur J Pain 2006 May;10(4):287-333. [doi: 10.1016/j.ejpain.2005.06.009] [Medline: 16095934]
40. Rickardsson J, Zetterqvist V, Gentili C, Andersson E, Holmström L, Lekander M, et al. Internet-delivered acceptance and commitment therapy (iACT) for chronic pain-feasibility and preliminary effects in clinical and self-referred patients. Mhealth 2020 Jul;6:27 [FREE Full text] [doi: 10.21037/mhealth.2020.02.02] [Medline: 32632365]
41. Reyes AT. The process of learning mindfulness and acceptance through the use of a mobile app based on acceptance and commitment therapy: a grounded theory analysis. Issues Ment Health Nurs 2022 Jan 04;43(1):3-12. [doi: 10.1080/01612840.2021.1953652] [Medline: 34346278]
42. Jayasekara RS. Focus groups in nursing research: methodological perspectives. Nurs Outlook 2012 Nov;60(6):411-416. [doi: 10.1016/j.outlook.2012.02.001] [Medline: 22464693]
43. Haugland MJ, Rosenberg I, Aasekjær K. Collaborative learning in small groups in an online course - a case study. BMC Med Educ 2022 Mar 10;22(1):165 [FREE Full text] [doi: 10.1186/s12909-022-03232-x] [Medline: 35272639]
44. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381 [FREE Full text] [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
45. Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O'Neal L, et al. The REDCap consortium: building an international community of software platform partners. J Biomed Inform 2019 Jul;95:103208 [FREE Full text] [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]
46. Thompson D, Deatrick JA, Knafl KA, Swallow VM, Wu YP. A pragmatic guide to qualitative analysis for pediatric researchers. J Pediatr Psychol 2022 Sep 15;47(9):1019-1030. [doi: 10.1093/jpepsy/jsac040] [Medline: 35552750]
47. Damschroder LJ, Reardon CM, Widerquist MA, Lowery J. The updated Consolidated Framework for Implementation Research based on user feedback. Implement Sci 2022 Oct 29;17(1):75 [FREE Full text] [doi: 10.1186/s13012-022-01245-0] [Medline: 36309746]
48. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. Implement Sci 2009 Aug 07;4(1):50 [FREE Full text] [doi: 10.1186/1748-5908-4-50] [Medline: 19664226]

XSL•FO
RenderX

49. 1177 homepage. 1177. URL: https://www.1177.se/ [accessed 2025-06-16]
50. Elkes J, Cro S, Batchelor R, O'Connor S, Yu LM, Bell L, et al. User engagement in clinical trials of digital mental health interventions: a systematic review. BMC Med Res Methodol 2024 Aug 24;24(1):184 [FREE Full text] [doi: 10.1186/s12874-024-02308-0] [Medline: 39182064]
51. Ludden GD, van Rompay TJ, Kelders SM, van Gemert-Pijnen JE. How to increase reach and adherence of web-based interventions: a design research viewpoint. J Med Internet Res 2015 Jul 10;17(7):e172 [FREE Full text] [doi: 10.2196/jmir.4201] [Medline: 26163456]
52. Eysenbach G. Medicine 2.0: social networking, collaboration, participation, apomediation, and openness. J Med Internet Res 2008 Aug 25;10(3):e22 [FREE Full text] [doi: 10.2196/jmir.1030] [Medline: 18725354]
53. Rosser BA, Eccleston C. Smartphone applications for pain management. J Telemed Telecare 2011 Aug 15;17(6):308-312. [doi: 10.1258/jtt.2011.101102] [Medline: 21844177]
54. de la Vega R, Serrano-Ibáñez ER, Ruiz-Párraga GT, Palermo TM, Wicksell R, Fernández-Jiménez E, et al. Protocol for a multi-phase, multi-center, real-world, hybrid effectiveness-implementation study of a digital intervention for pediatric chronic pain co-designed with patients (Digital SPA). Digit Health 2023 Dec 19;9:20552076231219490 [FREE Full text] [doi: 10.1177/20552076231219490] [Medline: 38130799]
55. Tappen RM, Cooley ME, Luckmann R, Panday S. Digital health information disparities in older adults: a mixed methods study. J Racial Ethn Health Disparities 2022 Feb 07;9(1):82-92 [FREE Full text] [doi: 10.1007/s40615-020-00931-3] [Medline: 33415705]
56. Connolly SL, Ferris SD, Azario RP, Miller CJ. Patient and provider attitudes toward video and phone telemental health care during the COVID-19 pandemic: a systematic review. Clin Psychol (New York) 2024;31(4):488-503. [doi: 10.1037/cps0000226] [Medline: 40129719]
57. Békés V, Ménard CG, Schmitz N, Aafjes-van Doorn K. Therapist-reported differences between teletherapy via phone and via videoconferencing. Brain Sci 2023 Dec 14;13(12):1714 [FREE Full text] [doi: 10.3390/brainsci13121714] [Medline: 38137162]
58. Mirbaha S, Morgan A, Tang A, Smith-Turchyn J, Richardson J. Models of telehealth service delivery in adults with spinal cord injuries: scoping review. JMIR Rehabil Assist Technol 2023 Jun 29;10:e41186 [FREE Full text] [doi: 10.2196/41186] [Medline: 37384377]
59. Graham AK, Lattie EG, Powell BJ, Lyon AR, Smith JD, Schueller SM, et al. Implementation strategies for digital mental health interventions in health care settings. Am Psychol 2020 Nov;75(8):1080-1092 [FREE Full text] [doi: 10.1037/amp0000686] [Medline: 33252946]
60. Groos SS, Linn AJ, Kuiper JI, van Schoor NM, van der Velde N, van Weert JC. Combining user-centered design and behavioral theory to enhance health technologies: a personas-based approach for a primary-care based multifactorial falls risk assessment tool. Int J Med Inform 2024 Jun;186:105420 [FREE Full text] [doi: 10.1016/j.ijmedinf.2024.105420] [Medline: 38518678]
61. Evangelista L, Steinhubl SR, Topol EJ. Digital health care for older adults. The Lancet 2019 Apr;393(10180):1493. [doi: 10.1016/s0140-6736(19)30800-1]
62. Bostrøm K, Børøsund E, Varsi C, Eide H, Flakk Nordang E, Schreurs KM, et al. Digital self-management in support of patients living with chronic pain: feasibility pilot study. JMIR Form Res 2020 Oct 23;4(10):e23893 [FREE Full text] [doi: 10.2196/23893] [Medline: 33094734]
63. Sá KN, Moreira L, Baptista AF, Yeng LT, Teixeira MJ, Galhardoni R, et al. Prevalence of chronic pain in developing countries: systematic review and meta-analysis. Pain Rep 2019;4(6):e779 [FREE Full text] [doi: 10.1097/PR9.0000000000000779] [Medline: 31984290]
64. Tsang A, Von Korff M, Lee S, Alonso J, Karam E, Angermeyer MC, et al. Common chronic pain conditions in developed and developing countries: gender and age differences and comorbidity with depression-anxiety disorders. J Pain 2008 Oct;9(10):883-891 [FREE Full text] [doi: 10.1016/j.jpain.2008.05.005] [Medline: 18602869]
65. Gruber J, Mendle J, Lindquist KA, Schmader T, Clark LA, Bliss-Moreau E, et al. The future of women in psychological science. Perspect Psychol Sci 2021 May;16(3):483-516 [FREE Full text] [doi: 10.1177/1745691620952789] [Medline: 32901575]
66. Goossens ME, Vlaeyen JW, Hidding A, Kole-Snijders A, Evers SM. Treatment expectancy affects the outcome of cognitive-behavioral interventions in chronic pain. Clin J Pain 2005;21(1):18-26; discussion 69. [doi: 10.1097/00002508-200501000-00003] [Medline: 15599128]
67. Perera C, Salamanca-Sanabria A, Caballero-Bernal J, Feldman L, Hansen M, Bird M, et al. No implementation without cultural adaptation: a process for culturally adapting low-intensity psychological interventions in humanitarian settings. Confl Health 2020;14:46 [FREE Full text] [doi: 10.1186/s13031-020-00290-0] [Medline: 32684948]
68. Heim E, Mewes R, Abi Ramia JA, Glaesmer H, Hall B, Harper Shehadeh M, et al. Reporting cultural adaptation in psychological trials - the RECAPT criteria. Clin Psychol Eur 2021 Nov;3(Spec Issue):e6351 [FREE Full text] [doi: 10.32872/cpe.6351] [Medline: 36405678]

XSL•FO
RenderX

Abbreviations

CBT: cognitive behavioral therapy
CFIR: Consolidated Framework for Implementation Research
DAHLIA: Digital Behavioural Health for Chronic Pain
GUIDED: Guidance for the Reporting of Intervention Development
REDCap: Research Electronic Data Capture

Edited by A Mavragani; submitted 21.03.25; peer-reviewed by M Adnan, C Onah; comments to author 21.04.25; revised version received 12.05.25; accepted 02.06.25; published 08.07.25.

Please cite as:
Taygar AS, Bartels SL, de la Vega R, Flink I, Engman L, Petersson S, Johnsson SI, Boersma K, McCracken LM, Wicksell RK
User-Driven Development of a Digital Behavioral Intervention for Chronic Pain: Multimethod Multiphase Study
JMIR Form Res 2025;9:e74064
URL: https://formative.jmir.org/2025/1/e74064
doi:10.2196/74064
PMID:

©Afra Selma Taygar, Sara Laureen Bartels, Rocío de la Vega, Ida Flink, Linnéa Engman, Suzanne Petersson, Sophie I Johnsson, Katja Boersma, Lance M McCracken, Rikard K Wicksell. Originally published in JMIR Formative Research (https://formative.jmir.org), 08.07.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Online Tailored Decision Aid for Maternal Pertussis Vaccination in a Randomized Controlled Trial: Process Evaluation Study

Charlotte Anraad[1,2], PhD; Pepijn van Empelen[2], PhD; Robert AC Ruiter[1], PhD; Hilde M van Keulen[2], PhD

[1]Department of Work & Social Psychology, Faculty of Psychology and Neuroscience, Maastricht University, Maastricht, The Netherlands
[2]TNO Child Health, Netherlands Organization for Applied Scientific Research, Leiden, The Netherlands

Corresponding Author:
Charlotte Anraad, PhD
Department of Work & Social Psychology
Faculty of Psychology and Neuroscience
Maastricht University
P.O. Box 616
Maastricht, 6200MD
The Netherlands
Phone: 31 43 388 2222
Email: charlotte.anraad@maastrichtuniversity.nl

Abstract

Background: To promote informed decision-making and maternal pertussis vaccination (MPV) uptake, we systematically developed an interactive, web-based decision aid for pregnant users. Intervention reach (the percentage of participants in the intervention group who used the intervention), use (how much and how long those participants used the intervention), and acceptability (how positively they evaluated the intervention) are essential for it to be effective and should be reported to assess which intervention components may have been effective.

Objective: This is a process evaluation aiming to evaluate (1) the reach and (2) the use, and (3) the acceptability of the intervention.

Methods: We analyzed the reach and use of the intervention among participants in the intervention group of a randomized controlled trial (RCT) that assessed the effects of an online tailored decision aid in the form of a web app. Participants were recruited via social media and midwifery clinics and invited via email to use the intervention at 18 weeks of pregnancy. Reach was measured objectively by assessing the number of participants who visited the intervention at least once. Use of the intervention was logged and included time spent on the decision aid, the number of times clicked, pages visited, and answers given in interactive components. Data from the baseline survey (at <18 wk of pregnancy) were used to measure sociodemographics, informed decision-making, MPV uptake, and determinants of uptake. A posttest survey (20-22 weeks of pregnancy) was used to evaluate the acceptability of the decision aid. We report the findings descriptively and assess baseline differences between those who used versus those who did not use the intervention.

Results: Of the 586 participants in the intervention group, 463 (79%) reached the home page of the intervention. Intervention reach appeared higher among those in their first pregnancy (8.35% difference, P=.11), those recruited via their midwife rather than via social media (10.56% difference, P=.04), and those who had completed a higher educational level (7.35% difference, P=.06). On average, participants spent 4.25 (SD 4.39) minutes on the decision aid. Most participants used the decision aid once (56.2% of those who reached it, n=260) or twice (26.6%, n=123). The average number of clicks was 27.24 (SD 25.08) and varied widely. Regarding acceptability, participants evaluated the decision aid positively with an overall grade of 8.0 out of 10 (SD 1.01). In total, 38.9% (180/463) of participants who used the intervention indicated that the decision aid helped them with their MPV decision-making.

Conclusions: The reach of the decision aid was successful with 79%, and participants were very positive about the decision aid. The use of the intervention (eg, time spent on the intervention) leaves room for improvement and should be improved to maximize intervention effects.

(JMIR Form Res 2025;9:e50709) doi:10.2196/50709

XSL•FO
RenderX

KEYWORDS

decision aid; process evaluation; informed decision making; intervention mapping; vaccine hesitancy; maternal pertussis vaccination

Introduction

In December 2019, maternal immunization during pregnancy (ie, maternal pertussis vaccination [MPV]) was introduced into the National Immunisation Programme (NIP) in the Netherlands. The aim of MPV is to protect infants from pertussis in their first months of life [1]. MPV is given at 22 weeks of pregnancy. Upon introduction in 2019, the uptake of MPVs in the Netherlands was 70% in 2020 [2].

Getting an MPV is voluntary. A decision aid is a tool aimed at preparing people to make a decision [3]. Because decision aids have shown promising results for screening decisions, leading to greater knowledge and lower decisional conflict among users, we have developed an online decision aid for MPV decisions [4-6]. This is the first decision aid for MPV to our knowledge.

In an overarching study (preregistration [7]), we investigated the effects of the intervention on MPV uptake, informed decision-making (IDM), and determinants of uptake [8]. Results showed a positive effect on IDM, and no significant effect on MPV uptake [8]. However, the intervention consists of several components targeting different outcomes and is offered to a varied group of participants, and such complexity requires an evaluation beyond effects only [9]. Therefore, in this study, we aimed to gain a further understanding of the actual usage of the interventions and the extent to which intervention components contributed to the outcomes by means of a process evaluation. A process evaluation is defined as an "assessment of fidelity and quality of implementation, clarifying causal mechanisms and identifying contextual factors associated with variation in outcomes" [10]. A process evaluation can be used for the following three objectives: (1) it provides insight into the extent to which an intervention and its various components are used, to be able to identify effective components and mechanisms [9,11,12]; (2) it provides insight into the implementation of the intervention, thereby avoiding a type 3 error (concluding that an intervention is ineffective due to unsuccessful implementation), and (3) it provides recommendations for intervention improvement [11].

In this study, we focused on reach, use, acceptability, and satisfaction. Reach refers to the proportion of participants who use the interventions when it is offered to them (ie, all participants in the intervention group of the randomized controlled trial [RCT]). Use refers to the extent to which they will use it. Reviews about decision aids often do not report their reach and use [4,5]. This may be because many studies offer decision aids in a controlled environment (where participants use the decision aid at a research location), in which the reach is 100%. In an uncontrolled environment like in our study, studies have reported a reach between 60% and 65% [13,14]. We will assess whether reach and use are related to baseline characteristics and sociodemographics, because we expect that certain subgroups, such as those in doubt about MPV, may have a higher need for the use of the intervention.

Methods

Study Design

This study is part of an RCT to assess intervention effectiveness on IDM, MPV uptake, and determinants of MPV uptake [8]. The original protocol has been described previously [7], and the effect evaluation is described in another study [8]. Because we evaluated intervention reach and use, only participants from the intervention condition were included in this study (N=586). The target group of the study was women in the Netherlands who were less than 18 weeks pregnant at the time of inclusion. Other inclusion criteria were a command of the Dutch language and having given informed consent. For this study, only participants who were in the intervention group of the RCT were included. The study took place between November 2020 and April 2022. The RCT protocol has been described elsewhere in detail [8].

Ethical Considerations

This study was reviewed by the TNO institutional review board for human research (registration number 2018-01). All participants gave informed consent, and the data were pseudonymized. Participants received a voucher for the value of €5 (equivalent to approximately US $5.87 based on the average conversion rate of €1=US $1.174 applicable during the study period; European Central Bank), to thank them for their participation. All the data presented here cannot be linked to individuals.

Procedure

Participants were invited to participate via midwifery clinics (n=62) and social media (Facebook and Instagram [Meta]). Midwifery clinics were geographically located in different rural and urban areas in the Netherlands. Participants could enroll in the study if they were less than 18 weeks pregnant. We chose this cutoff point because this is generally before participants make an appointment to receive MPV. They could follow a website link in the invitation letter that directed them to a website with study information and an informed consent form. Informed consent meant giving permission to participate in the questionnaires and random assignment to the control or intervention condition, and permission to request vaccination status from Praeventis, which is the national vaccination register. Those who provided informed consent were assigned a unique participant code and were immediately directed to the online baseline questionnaire, which assessed sociodemographics, IDM, and psychological determinants of MPV uptake targeted in the decision aid. Those who did not fill out the baseline questionnaire at 17 weeks of pregnancy received a reminder via email. After filling out the baseline questionnaire, participants were randomized to either of the 2 conditions.

Participants who were recruited online were randomized into the intervention and control conditions. Participants recruited via midwifery clinics were randomized at the clinic level. Participants in the intervention condition received a link via

XSL•FO
RenderX

email to the intervention with their personal participant code at 18 weeks of pregnancy. They could visit the intervention as much as they wanted until they received an invitation and link via email to fill out the posttest measurement at 20 weeks of pregnancy. This is generally around the time that pregnant women make an appointment to receive MPV at 22 weeks, or might already have received the vaccination. At posttest measurement, IDM, decisional certainty, and determinants of MPV uptake were measured again, and a subjective evaluation of the intervention was performed. Those who had not filled out the posttest measurement at 21 weeks received a reminder via email. Participants who had, for whatever reason, already received MPV when filling out the posttest measurement were excluded because this may have influenced their answers.

A more extensive description of the procedure can be found in another study [8].

Intervention

The intervention is an online decision aid. Participants were sent an email with an invitation to use the decision aid at 18 weeks of pregnancy. The decision aid is a mobile-first web application that can be visited on any device or saved as an app on a mobile phone. The systematic development of the decision aid has been described in detail elsewhere [15], but we will provide some information on this below, as it is relevant to this study. We used Intervention Mapping [16], a systematic intervention planning framework, to develop the intervention. As part of these planning activities, we defined behavioral objectives and identified important, changeable behavioral determinants of vaccination uptake. Next, we matched these with behavior change techniques, which were then translated into practical applications. We used user-centered design methods to create the intervention, aiming to meet the needs and user preferences of the target group. We involved the target group in 4 iterations during the development process, with new participants in each round. Nielsen and Landauer [17] determined that 5-8 participants per iteration identified most usability issues. In all pretests, we involved pregnant women of diverse ages and backgrounds. The pretests aimed to get participants' feedback on the intervention's clarity, relevance, usability, and overall structure. In the initial pretest, a focus group consisting of 6 pregnant women was presented with a static intervention prototype. In the second iteration, 5 pregnant women individually used an interactive prototype of the intervention during think-aloud sessions. The third iteration featured a full, interactive version of the intervention, with 6 pregnant women using the intervention individually during a think-aloud session. The fourth iteration was a usability test with 4 low-literacy users in individual think-aloud sessions. After the final iteration, the intervention was tested further by members of the project group on various devices to ensure usability. Adjustments made based on each iteration are described in our other paper [15].

This resulted in a decision aid with the aim to actively inform users about MPV, help them process and weigh the information, and deliberate on what the information means for them in the light of their personal values. In addition, the decision aid aimed to reduce the barriers of talking about MPV to an important other or health care professional, and the barrier of making an appointment.

The decision aid consists of 3 components: an information component, a "my choice" component, and a "make an appointment" component. The information component stimulated active learning by providing information in 3 modes (video, text, and audio) and by providing knowledge questions with immediate feedback. Information topics included "How does MPV work?" "What is in the vaccine?" "What is whooping cough?" "Side effects" "Safety" "When can you not get MPV?" and practical information about MPV, and each page contained on average 94 words in text. The "my choice" component was divided into 3 subcomponents. The first, "test your knowledge," used active learning and feedback to provide the most basic and relevant information about the vaccine [18-21]. The second, "weighing pros and cons," aimed to improve decisional certainty by asking participants to fill out questions about potential considerations and then providing a tailored overview of the participants' answers using a decisional balance [22]. The questions in the "test your knowledge" and "weighing pros and cons" components can be found in the Results section and Figure 1. The third component is "prepare a conversation about the vaccine." In this chat-like conversational component, the participants prepared for a conversation with a significant other, indicating what they wanted to gain from a conversation with an important social referent or health care provider, and what their feelings, needs, and questions were regarding the MPV. At the end of the exercise, they could save an overview of their answers. Fourth, the "make an appointment" component aimed to decrease the barrier to getting an MPV by providing a postcode-based location finder for an MPV appointment. Participants could navigate through the web application using a menu at the bottom of the web page (Figure 2A). Figure 2 shows screenshots of the decision aid.

XSL•FO
RenderX

Figure 1. Flow diagram of the recruitment and response of study participants from in the intervention condition.

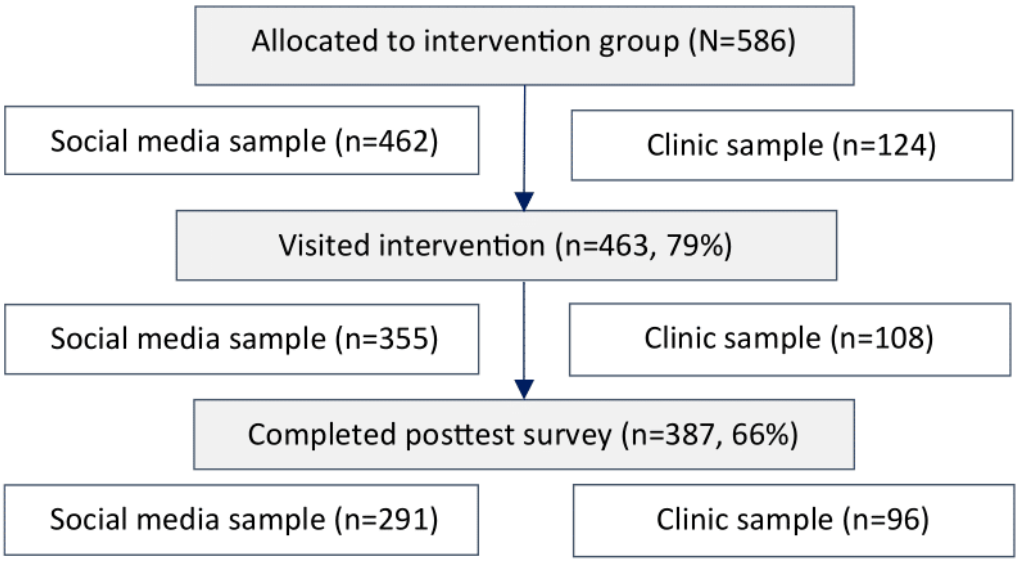

Figure 2. Screenshots of a selection of the web application components. (A) An overview of the information topics. (B) An example of a "test your knowledge" question and tailored feedback. (C) An overview page of the decisional balance with tailored pros and cons. (D) The "prepare a conversation" exercise. (E) The "make an appointment" page.

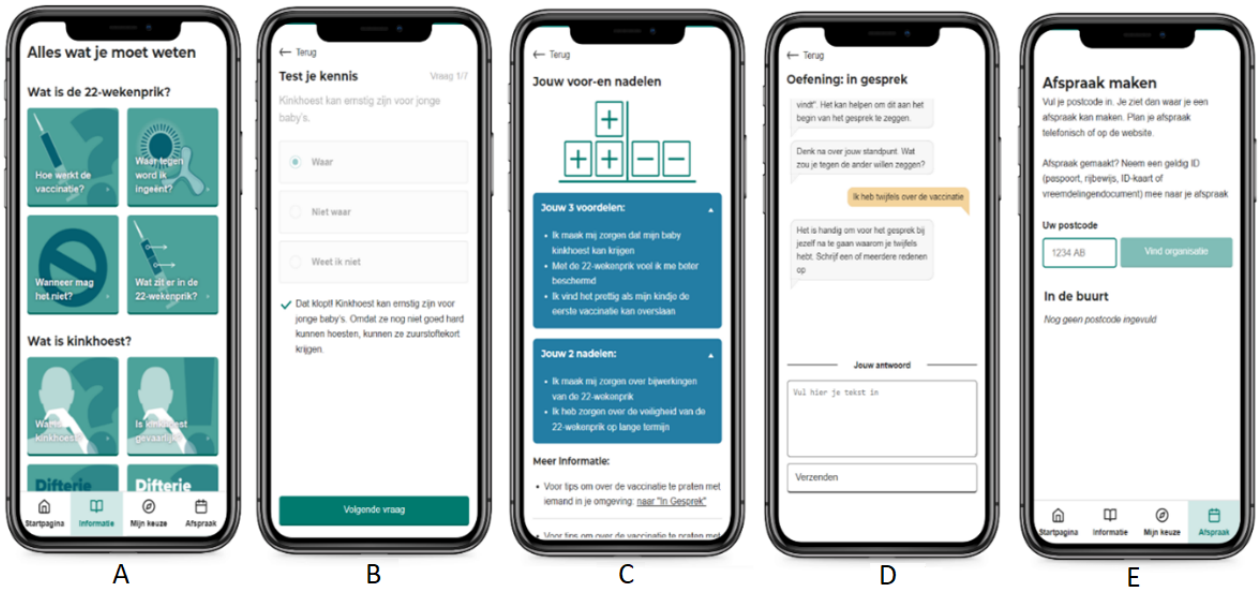

Measurements

Program Reach

Program reach is a measure of intervention participation that reflects the proportion of participants who used the intervention as a percentage of those who were offered the intervention (ie, the intervention group) [23]. Intervention reach was measured according to computer logs.

Program Use

Program use refers to how much the intervention was used by those who reached it. Oftentimes, this is conceptualized either as implementation or the delivered and received dose of the intervention, or both [23]. Because our specific intervention was designed in a way that participants could use the components that were most relevant to them, we will not focus on complete use of the intervention as an outcome. Instead, we will describe the use of the intervention, so it becomes clear which parts of the intervention have contributed to the effects [13]. Intervention use was measured according to computer logs (objectively) and in the survey (subjectively). Every time a participant used the intervention, a record was created linked to their assigned unique participant number. Every time a participant clicked on a page, a question, or a menu option, it was logged, creating a record of which components and pages were visited, as well as what answers were given to questions in interactive components.

Time spent on the intervention was calculated by taking the sum of the difference in time between each event that was logged for a participant. When a participant did not click on anything for more than 4 minutes, the session was considered ended. We chose this cutoff value because some pages take time to read, or to watch a video on, and inactivity for up to 4 minutes

XSL•FO
RenderX

is to be expected there. We measured the number of times participants visited the intervention by looking at how often they revisited the intervention after having not used it for at least 10 minutes.

Sociodemographics and Determinants of MPV Uptake

Sociodemographics and determinants of MPV were measured with a survey at baseline (<18 wks of pregnancy) to assess the relationship between reach and use with sociodemographics and determinants of MPV uptake.

Age, highest completed educational level, country of birth of the participant, and whether the participant had children and religious affiliation were measured at baseline. The educational level was classified as low (less than secondary or vocational education), intermediate (secondary and vocational education), or high (higher or university education). The country of birth was divided into 2 categories: the Netherlands or other.

Determinants of MPV were measured with a survey at baseline (<18 wks of pregnancy). To measure the determinants targeted in the intervention, we computed composite scores (Table 1). Determinants included MPV intention, attitude about MPV, perceived susceptibility and severity of pertussis for babies, beliefs about the safety and effectiveness of MPV, moral and injunctive norms, affect around MPV, and perceived control. To assess internal consistency, Cronbach alpha was used for scales consisting of more than 2 items, and the Pearson correlation coefficient (r) was used for scales consisting of 2 items. Negatively formulated items were reverse-coded (marked with an R in Table 1). The determinants and their measurements were based on an earlier study of the determinants of MPV uptake intention in the Netherlands [24].

XSL•FO
RenderX

Table 1. Overview of outcome measures, their scores or scales, and internal consistencies.

Measures and items	Score or scale	Cronbach alpha or Pearson *r*
Knowledge (7 items)	Sum score of correct answers. 0=low knowledge about MPV to 7=high knowledge about MPV (total number of correct answers)	__[a]
The MPV[b] is meant to protect the baby		
A painful arm is a common side effect of MPV		
Whooping cough is never serious for young babies (R)[c]		
The MPV only protects against whooping cough, and not against other diseases (R)[c]		
After getting MPV, the baby can skip their first vaccination after birth		
Whooping cough can be transmitted by coughing		
The MPV protects only my baby, and not me, against whooping cough (R)[c]		
Decisional certainty (3 items)	1=low decisional certainty to 5=high decisional certainty	0.93 (α)
It is clear to me what the best decision for me is regarding the MPV		
I feel certain about my decision about MPV		
I find it easy to make a decision about MPV		
MPV intention (3 items)	1=low intention of getting MPV to 5=high intention of getting MPV	0.98 (α)
I plan to get MPV		
I expect to get MPV		
It is probable that I will get MPV		
Attitude about MPV (4 items)		
I find MPV:	1=negative to 5=positive	0.93 (α)
very bad – very good		
very unimportant – very important		
very undesirable – very desirable		
very unnecessary – very necessary		
Perceived susceptibility of pertussis (1 item)		
The chance that my baby will get whooping cough (without the MPV) is ….	1=very small to 5=very big	—
Perceived severity of pertussis (2 items)		0.62 *(r)*
Whooping cough in a baby is ...	1=not serious to 5=very serious	
If your baby gets whooping cough, how likely do you think it is that it will have serious effects?	1= not likely at all to 5= very likely	
Beliefs safety of MPV (3 items)	1=very unsafe to 5=very safe	0.71 (α)
MPV has been sufficiently tested in pregnant women		
MPV is safe for pregnant women		
MPV is safe for the baby		
Beliefs effectiveness of MPV (3 items)	1=very ineffective to 5=very effective	0.78 (α)
MPV decreases the chance of whooping cough in babies		
Babies are better protected against whooping cough with MPV than without MPV		

XSL•FO
RenderX

Measures and items	Score or scale	Cronbach alpha or Pearson *r*
Knowledge (7 items)	Sum score of correct answers. 0=low knowledge about MPV to 7=high knowledge about MPV (total number of correct answers)	__[a]
There is less whooping cough in babies because of MPV		
Moral norm (2 items)	1=low moral norm to 5=high moral norm	0.59 *(r)*
I find it my responsibility to get vaccinated for whooping cough		
I find it the responsibility of pregnant women to get MPV		
Injunctive norms (2 items)	1=low injunctive norm to 5=high injunctive norm	0.44 *(r)*
I think my partner wants me to get an MPV		
I think my obstetric care provider wants me to get MPV		
Affect about MPV (5 items)		
About the MPV I feel:	1=negative to 5=positive	0.89 (α)
Comfortable		
Relaxed		
Angry (R)[c]		
Certain		
Scared (R)[c]		
Perceived control (5 items)		
I can manage the following situations:	1=low perceived control to 5=high perceived control	0.83 (α)
Assessing information about MPV		
Having a conversation about MPV		
Making a decision about MPV		
Making an appointment to get MPV		
Going to the Youth Health Centre to get an MPV		

[a]Not available.

[b]MPV: maternal pertussis vaccination.

[c]R: negatively formulated items were reverse-coded.

Program Acceptability and Satisfaction

Program acceptability was measured with a post-test survey at 20-22 weeks of pregnancy. We asked participants to evaluate the intervention with a grade (1=very bad and 10=excellent), as well as to rate the different components and characteristics like layout. In addition, participants were asked open-ended questions about which things they valued the most and the least about the decision aid, as well as what could be improved. Table 2 shows an overview of the measures with references. In addition, we added a measure of engagement, given the interactive nature of the intervention.

XSL•FO
RenderX

Table 2. Acceptability measurements.

Measures and items	Score or categories	Reference
Interest		
I found this online decision aid:		
Interesting	1 = completely disagree – 5 = completely agree	[25]
Usefulness		
I found this online decision aid:		
Useful	1 = completely disagree – 5 = completely agree	[25]
Ease of use		
I found this online decision aid:		
Easy to use	1 = completely disagree – 5 = completely agree	[26]
Trustworthiness		
I found this online decision aid:		
Trustworthy	1 = completely disagree – 5 = completely agree	[25]
Attitude toward application		
How good or bad did you find the following components (if you did not use the component, click not applicable)		[27]
The speed of the web app	1 = very bad – 5 = very good	
The lay-out	1 = very bad – 5 = very good	
The amount of text	1 = very bad – 5 = very good	
The videos	1 = very bad – 5 = very good	
The images	1 = very bad – 5 = very good	
The option to do a knowledge test	1 = very bad – 5 = very good	
The option to weigh the pros and cons about MPV[a]	1 = very bad – 5 = very good	
The option to prepare a conversation	1 = very bad – 5 = very good	
The option to find a location to get the MPV	1 = very bad – 5 = very good	
Elaboration		
How well did you read the information on the information pages?	1 = very superficially – 5 = very attentively	__[b]
Information evaluation		
Understandable (R)[c]		
I found the information in the decision aid:		
Difficult to understand	1 = completely disagree – 5 = completely agree	[25]
Relevance		
I found the information in the decision aid:		
Relevant	1 = completely disagree – 5 = completely agree	[25]
Quantity		
I found the information in the decision aid:		
Much	1 = completely disagree – 5 = completely agree	[25]
Sidedness		
I found the information in the decision aid:		
Independent	1 = completely disagree – 5 = completely agree	[13]
Engagement		
I found the information in the decision aid:		
The online decision aid made me think about MPV	1 = completely disagree – 5 = completely agree	—

XSL•FO
RenderX

Measures and items	Score or categories	Reference
Support		
I found the information in the decision aid:		
The online decision aid helped me in my decision about MPV	1 = completely disagree – 5 = completely agree	[13]
Recall		
I found the information in the decision aid:		
I can recall the online decision aid well	1 = completely disagree – 5 = completely agree	[13]
Recommend		
I found the information in the decision aid:		
I would recommend the online decision aid	1 = completely disagree – 5 = completely agree	[26]

[a]MPV: maternal pertussis vaccination.

[b]Not applicable.

[c]R: negatively formulated items were reverse-coded.

Analyses

Descriptive analyses (percentages, numbers, means, and SDs) were used to describe the intervention use and acceptability. We used chi-square and independent sample *t* tests to test for baseline differences between participants who visited the decision aid and those who did not reach it, despite being offered the intervention. In view of multiple testing, an effect was considered significant when $P<.004$ (Bonferroni corrected $\alpha=.05/13$ factors, to avoid a type-1 error [28]).

We also investigated the characteristics of the participants who used each component. We did this for the use of the "Information" component, "Weighing pros and cons," "Knowledge test," "Conversation preparation," "Appointment," and the use of videos (having watched at least one video). Here, we will also report averages and SDs and whether differences are statistically significant based on chi-square and independent sample t-tests, after Bonferroni correction. We checked if acceptability correlates with time spent on the intervention and education level.

Analyses were performed in R (version 4.1.2; R Core Team) [29].

Results

Sample Description

Figure 1 shows the flow diagram of participants in the intervention group. Dropout at the posttest survey (at 20-22 wks of pregnancy) was 33%.

Tables 3 and 4 show the sociodemographic characteristics of participants in the intervention condition (N=586). The mean age of the participants was 32.21 (SD 4.04) years (the average age of women in the Netherlands at the moment of giving birth was 31.7 years in 2021 [30]), about half of the participants (274/586, 46.8%) already had a child, and 97.6% (572/586) of participants were born in the Netherlands. The level of education of the participants was high, with 80% being highly educated compared with 60.9% in the general population among women of the average age of 32.32 (SD 4.04) years in our sample [31].

XSL•FO
RenderX

Table 3. Baseline characteristics (measured <18 wks of pregnancy) of the study sample, and for participants who visited the intervention versus those who did not (part 1).

Sociodemographic variables	Total sample intervention group (N=586), n (%)	DA visited at least once, (n=463), n (%)	DA not visited (n=123), n (%)	Chi-square (*df*)	*P* value
Recruitment channel				4.49 (1)	.03
Clinic sample	124 (21.16)	107 (23.11)	17 (13.82)		
Social media sample	462 (78.84)	356 (76.89)	106 (86.18)		
Has at least one child				2.02 (1)	.15
No	312 (53.24)	254 (54.87)	58 (47.15)		
Yes	274 (46.75)	209 (45.14)	65 (52.85)		
Country of birth				0.91 (1)	.34
Netherlands	572 (97.6)	450 (97.19)	122 (99.19)		
Other	14 (2.4)	13 (2.81)	1 (0.81)		
Highest education completed				4.06 (1)[a]	.04[a]
Low	4 (0.7)	4 (0.86)	0 (0)		
Intermediate	113 (19.3)	80 (17.28)	33 (26.83)		
High	469 (80)	379 (81.82)	90 (73.17)		

[a]Low and intermediate versus high educational levels were compared.

Table 4. Baseline characteristics (measured <18 wks of pregnancy) of the study sample, and for participants who visited the intervention versus those who did not (part 2)

Sociodemographic variables (scale)	Total sample intervention group (N=586), mean (SD)	DA visited at least once (n=463), mean (SD)	DA not visited (n=123), mean (SD)	*t* test[a] (*df*)	Cohen *d* (95% CI)	*P* value
Age (years)	32.21 (4.04)	32.22 (4.03)	32.18 (4.10)	–0.10 (189.51)	–0.01 (–0.21 to 0.19)	.92
Religion (1-7)	2.06 (1.29)	2.04 (1.29)	2.13 (1.31)	0.70 (189.22)	0.07 (–0.13 to 0.27)	.48
Intention to accept MPV[b] (1-5)	4.61 (0.73)	4.62 (0.70)	4.56 (0.83)	–0.81 (170.86)	–0.09 (–0.29 to 0.11)	.42
Attitude MPV (1-5)	4.61 (0.65)	4.63 (0.61)	4.54 (0.78)	–1.12 (163.84)	–0.13 (–0.33 to 0.07)	.26
Beliefs safety (1-5)	4.41 (0.81)	4.19 (1.01)	4.21 (1.01)	–2.63 (159.54)	–0.32 (–0.52 to –0.12)	<.01
Beliefs effectiveness (1-5)	4.58 (0.59)	4.61 (0.53)	4.46 (0.78)	–2.03 (152.91)	–0.26 (–0.46 to –0.06)	.04
Perceived severity (1-5)	4.38 (0.66)	4.40 (0.64)	4.34 (0.72)	–0.75 (175.55)	–0.08 (–0.28 to 0.12)	.45
Perceived susceptibility (1-5)	2.63 (0.80)	2.59 (0.77)	2.80 (0.86)	2.53 (178.42)	0.27 (0.07 to 0.47)	.01
Moral norm (1-5)	4.31 (0.82)	4.31 (0.81)	4.30 (0.88)	–0.12 (180.04)	–0.01 (–0.21 to 0.19)	.91
Knowledge (1-5)	4.68 (1.53)	4.70 (1.52)	4.62 (1.57)	–0.53 (187.15)	–0.06 (–0.25 to 0.14)	.60
Perceived control (1-5)	4.67 (0.48)	4.68 (0.46)	4.60 (0.55)	–1.49 (168.62)	–0.17 (–0.37 to 0.03)	.14
Injunctive norm (1-5)	4.09 (0.98)	4.14 (0.93)	3.91 (1.13)	–2.00 (167.86)	–0.23 (–0.43 to –0.03)	.05
Affect (1-5)	4.39 (0.78)	4.40 (0.77)	4.36 (0.82)	–0.46 (184.28)	–0.05 (–0.25 to 0.15)	.65
Decisional certainty (1-5)	4.39 (0.95)	4.39 (0.95)	4.40 (0.96)	0.13 (189.78)	0.01 (–0.19 to 0.21)	.90

[a]Independent sample *t* test.

[b]MPV: maternal pertussis vaccination.

Program Reach

In total, 463 out of 586 participants (79%) from the intervention group used the decision aid at least once. Tables 3 and 4 show baseline characteristics for the participants who did versus those who did not visit the intervention despite being offered the intervention (as all participants were in the intervention group of the RCT). Intervention reach appeared higher among those in their first pregnancy (7.71% difference, *P*=.15), those recruited via their midwife rather than via social media (9.29% difference, *P*=.03), and those who had completed a higher educational level (8.65% difference, *P*=.04). There were small differences in baseline scores for determinants of MPV uptake. Overall, determinants were slightly more positive toward MPV

XSL•FO
RenderX

among those who visited the intervention. The largest differences in reach of the intervention (difference between users and nonusers) were found for different levels of beliefs about safety (95% CI –0.515 to –0.115; *P*=.009) and perceived susceptibility (95% CI 0.07 to 0.47; *P*=.01), although these differences were not significant after Bonferroni correction.

Program Use

Table 5 shows the average (SD) time spent on the intervention, the number of times the intervention was visited, the number of clicks, and the number of pages and information pages visited, as well as how much attention the participants indicated having used the decision aid. On average, participants spent 4.25 (SD 4.39) minutes on the decision aid. Most participants used the decision aid once (56.2% of those who used it, 260/586) or twice (123/586, 26.6%). The average number of clicks was 27.24 (SD 25.08) and varied widely. The average number of visited components was 2.70 (SD 1.63), and the average number of information pages was 3.90 (SD 1.36). There was a correlation of 0.75 between the number of sections visited and the number of pages that participants looked at.

Table S1 in Multimedia Appendix 1 shows the number and percentage of participants who visited each component and information page of the intervention. Of the participants who clicked at least one page, 68.8% (319/463) visited at least one information page. The information pages about how MPV works, the side effects, and safety were the most visited information pages. The appointment component, where participants could use a postcode-based location finder, was visited by 53.1% (246/463) of the participants. Of the "my choice" components, the option to weigh pros and cons was the most used, it was completed by 32.6% (151/463) of the participants. A quarter of the participants (116/463, 25.1%) completed the knowledge test. Only 4 of 463 participants (0.8%) completed the conversation preparation.

We investigated the baseline characteristics of participants who used the different intervention components. There were no significant differences after Bonferroni correction.

Table 5. Program use: average (SD) time spent on the intervention, the number of clicks, and the number of pages and information pages visited (N=463).

	Mean (SD)	Range (minimum-maximum)
Time spent (mins)	4.25 (4.39)	>0-29.4
Number of times visited (a new "time" was defined by a revisit after at least 10 mins of no activity)	1.86 (1.48)	1-9
Clicks	27.24 (25.08)	1-167
Number of components visited out of the main components ("Homepage," "Information," "Appointment," and interactive "My choice" subcomponents: "Weighing pros and cons," "Knowledge test," "Conversation preparation")	2.70 (1.63)	1-6
Number of information pages visited	3.90 (1.36)	1-19

Acceptability

Overall, participants evaluated the intervention positively; they gave the intervention an average grade of 8 on a 10-point scale (1=very bad, 10=excellent). This score did not correlate with time spent on the intervention (*r*=.055). Table S2 in Multimedia Appendix 2 shows the mean and SD scores of the subjective evaluations. Mean scores were larger than 4 on a 5-point scale (1=not at all to 5=very much) for whether the decision aid was informative (mean 4.56, SD 0.83), trustworthy (mean 4.5, SD 0.73), useful (mean 4.45, SD 0.74), easy to use (mean 4.27, SD 0.9), and interesting (mean 4.23, 0.74). The speed of the web app, layout, text amount and images, and the option to weigh the pros and cons all received high mean scores (means>4) from participants. Videos, the knowledge test, the option to prepare a conversation, and the option to find a location to get MPV received scores between 3 and 4 and had larger SDs (>1.85). Participants indicated finding the information relevant (4.51). Opinions were divided about whether the information was much (mean 2.68, SD 1.14 with 1=not much and 5=very much), and independent (mean 3.71, SD 1.03). To the question of whether the decision aid helped the participant in their choice about MPV, 38.9% (107/275) answered "agree a bit" or "completely agree." The mean score was 2.96 (SD 1.38). In addition, 71% (192/270) answered "agree a bit" or "completely agree" to whether they would recommend the decision aid to others (mean score 3.85, SD 0.87).

Because our sample was relatively highly educated, we checked for differences in acceptability between the highest completed educational levels. The only significant difference (*P*<.05) was found in the evaluation of the knowledge test: low and intermediate-educated participants evaluated the knowledge test more positively (mean 4.28, SD 1.53) than high-educated participants (mean 3.81, SD 185).

In the open-ended questions in the survey, participants indicated enjoying the information, the interactive and visual elements such as the videos, and the option to weigh the pros and cons of MPV. When asked what they least enjoyed, some participants indicated that the decision aid was quite long. Answers showed that the needs for depth and amount of information varied among participants. Some indicated wanting more information, while others said it was already too much and quite difficult to follow. Some participants indicated finding it complex to navigate through the decision aid. Many participants also indicated wanting an easier way to make an appointment online. This could not be included in the decision aid because it is organized at the regional level in the Netherlands.

XSL•FO
RenderX

Discussion

Reach

This study describes the process evaluation of an online decision aid to promote IDM about MPV. We found that the reach of the intervention was adequate (79%) among our sample. This is slightly higher than in other studies about online decision aids. In other studies, a decision aid for HPV vaccinations was used by 62.8% of participants [13], and 64% of participants used a web-based intervention aimed at healthy dietary and physical activity behaviors [14]. Many studies about similar interventions did not report reach, making it difficult to compare our results. It should be noted here that our sample was selective, with highly educated people overrepresented, despite using 2 methods of recruitment (online recruitment and via midwifery clinics). Therefore, the intervention did not reach the intended diversity of users that we anticipated in the user-centered design. We will go into selection bias and recruitment challenges further under the Strengths and Limitations section.

Those recruited for the study by their midwife more often reached the decision aid than those recruited via social media. In addition, participants who scored higher at baseline on the belief that MPV is safe, and with lower perceived susceptibility of their baby getting pertussis, more often visited the intervention than those who did not believe MPV was safe and those with a higher perceived susceptibility of their baby getting pertussis. However, these differences were not significant after correction for multiple testing. A potential difference may be explained by those who believe that MPV is safe possibly have more faith in a decision aid made by a scientific institute. There were no other significant baseline differences between those who did and did not use the decision aid, although data indicated that use may be higher among those in their first pregnancy.

Use

Time spent on the intervention (mean 4.25, SD 4.39 minutes) was lower than in the study with the HPV decision aid (20 minutes; [13]). Other studies found similar times of use as we did [32,33]. Time spent on an online intervention, especially in a noncontrolled environment, is difficult to interpret because it is unclear whether participants are doing other things while using the intervention, or whether they spend all the time that is being logged on the intervention. We found a large variance (SD 4.39) in the time used for the intervention, indicating that some participants spent a very short time, and others a longer time on the intervention. How much time is spent on a decision aid may depend on different aspects, such as the total amount of content in the decision aid and the level of doubt about the decision experienced among users. The big difference between 20 minutes spent on the HPV decision aid and the 4.25 minutes spent on our decision aid indicates that the use of our decision aid leaves room for improvement, for example, by keeping the user engaged. In general, 4.25 minutes on average still gives a participant time to absorb some information. The average number of words on an information page was 94, given that Dutch adults read on average 202 words per minute [34]. In light of exercises such as “weighing the pros and cons,” more time is expected to be spent in order to give deliberated answers to the questions. However, there was a wide range of time spent on the intervention (0-29.4 minutes), indicating that some people engaged a lot more than others. A high correlation (r=0.75) between the number of pages visited and the time spent on the intervention indicated that those who visited more components of the intervention needed and took more time to absorb the information.

All components of the intervention were used by at least some participants. The most visited components of the decision aid were the information pages about how MPV works, side effects, and the safety of MPV, as well as the component where one could find a location to make an appointment to get MPV, and where the pros and cons of MPV could be weighed. The least used component was the conversation preparation about MPV, and the information pages about secondary issues, such as who is involved in MPV and whether MPV is painful. These outcomes are in line with what we know about decision-making about MPV from earlier studies. These showed that participants have a need for information about safety and side effects [35,36]. The low use of conversation preparation can be explained by that only a specific group of people with low decisional certainty and important others in the environment who potentially disagree with their stance have a need for conversation preparation. However, we did not find major differences in baseline characteristics between those who used specific components and those who did not. The use of videos was overall low. Furthermore, Readspeaker (having text read out loud) was used only by 5 participants and they were all highly educated. Because of the low use of Readspeaker and the low number of low-educated participants in our sample, it is difficult to draw conclusions about the use of this feature. We expected that Readspeaker would mostly be used by low-literacy people, people with dyslexia, and people with other reasons for wanting or not being able to read easily, such as impaired vision. Even though an application like Readspeaker is only useful for a limited group, it is important that decision aids are accessible and inclusive [37]. Because people have different preferred modes of receiving information [32], the fact that video and Readspeaker functions were only used by subgroups is not surprising.

Acceptability

Participants evaluated the decision aid in general positively, with an overall grade of 8.0 out of 10. They found the decision aid relevant and easy to use. Opinions varied about the videos, the conversation preparation, and the amount of text. This confirms that the needs among participants differ, and therefore some people would have liked more or less text, or did not evaluate all the components as useful. Of those who used the decision aid, 38.9% indicated that it helped them to an extent to decide about MPV and 71.1% would recommend it to others. High acceptability was also found in another study [13]. We believe that extensively involving the target group in the design of the decision aid and fine-tuning it to their needs contributed to high acceptability rates [15].

Participants evaluated the component with a postal code–based location finder for MPV with 3.53 on a scale of 1-5. Upon finding the right location using our decision aid, the procedure

XSL•FO
RenderX

for making the appointment varies per region of the Netherlands. This may have caused confusion, and unfortunately, it limits the options within the decision aid for making an appointment. In general, this emphasizes that making an appointment to get an MPV should be made as easy as possible. Furthermore, some participants in the subjective evaluation indicated that the navigation through the decision aid could be improved.

Online health interventions and decision aids are known to have limited reach among lower-educated populations [38,39]. However, the limited number of low-educated and intermediate-educated groups in our sample evaluated the decision aid just as positively as high-educated participants. It is important to research acceptability further among low-educated users, because this study did not include enough low-educated participants to draw firm conclusions on this.

Strengths and Limitations

A limitation of the study is that we had a selective sample in which highly educated women who were born in the Netherlands were overrepresented. High baseline levels of determinants of MPV uptake further indicate that our sample was subject to selection bias [8]. Caution should be exercised when interpreting the generalizability of this study's findings, meaning that the reach and use of the intervention are described for the specific population we were able to include in the study. This kind of selection bias is not uncommon in studies about vaccination uptake [40], but it is a problem because interventions are most needed among those in doubt and holding ambivalent beliefs about vaccinations. Although efforts should always be made to make decision aids and other online interventions as suitable as possible for a broad target group, some interventions could also be developed in parallel to meet the needs of specific subgroups, as one approach that fits all may not be realistic. For example, alongside the decision aid, we have developed an antenatal group-care intervention in which MPV is discussed [15]. Unfortunately, the effects and use of this intervention could not be included in the RCT due to COVID-19 regulations at the time of data collection.

In our sample, the attitude and intention toward MPV were already high at baseline. This indicates that the sample included relatively many people who were already positive about MPV, and perhaps more open to using a decision aid than people who are less positive about MPV. However, we found no differences in reach and use of the intervention in our sample based on attitude or intention toward MPV at baseline. In addition, our sample was far more highly educated (80% highly educated) than the average in the Netherlands (30% highly educated) [31]. This means that our conclusions about reach and use of the decision aid have limited generalizability to the general population, and low-educated people may need to be targeted specifically for the decision aid, and studies about similar interventions to reach them. This could, for example, be done using more outreaching methods of recruitment and dissemination, such as was done for the COVID-19 vaccination hotline in the Netherlands [33]. In addition, key figures in communities can be used to disseminate interventions [40].

We found no significant baseline differences regarding determinants of MPV uptake between those who did and did not use the decision aid upon being offered to do so. Because of multicollinearity between determinants and unequal group sizes, we did not perform a multivariate analysis to assess variance between those who did and did not use the intervention, but only assessed differences on the individual determinant level.

Conclusions

The decision aid had adequate reach, with 79% of the intervention group using the intervention to some extent. Efforts should be made to reach low-educated populations and those who are not already positive about MPV, as this could be improved. This could be done by using a recommendation to use the decision aid from a trusted healthcare professional, or by using a snowball recruitment method [41]. The use of the intervention varied across participants, indicating a variety of needs among the target group. Usage, such as time spent on the intervention, left room for improvement. Acceptability of the intervention was high, indicating that the user-centered design paid off.

Acknowledgments

This work was supported by ZonMW, the Netherlands Organization for Health Research and Development (grant number 522004003). No artificial intelligence was used in the writing of this manuscript.

Data Availability

The datasets generated during and/or analyzed during this study are not publicly available because participants did not consent to this, but they are registered at Zenodo (registration number 7661309) and are available upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Programme use: number and percentage of participants who visited the intervention pages per component among those who visited the decision aid at least once (N=463). The pages are ordered per component by the number of participants that visited them.

XSL•FO
RenderX

[PDF File (Adobe PDF File), 111 KB - formative_v9i1e50709_app1.pdf]

Multimedia Appendix 2
Subjective evaluation of the intervention at post test among participants in the intervention group who reported that they had visited the intervention to some extent (n=272).
[PDF File (Adobe PDF File), 98 KB - formative_v9i1e50709_app2.pdf]

References

1. van der Maas NAT, Mooi FR, de Greeff SC, Berbers GA, Spaendonck MAC, de Melker HE. Pertussis in the Netherlands, is the current vaccination strategy sufficient to reduce disease burden in young infants? Vaccine 2013;31(41):4541-4547. [doi: 10.1016/j.vaccine.2013.07.060] [Medline: 23933365]
2. van LE, Kamp L, Oomen P. Vaccinatiegraad en jaarverslag rijksvaccinatieprogramma Nederland 2020. RIVM Rijksinstituut voor Volksgezondheid en Milieu 2021. [doi: 10.21945/RIVM-2021-0011]
3. International patient decision aid standards (IPDAS) collaboration. IPDAS Voting Document. 2005. URL: http://www.ipdas.ohri.ca/IPDAS_First_Round.pdf [accessed 2025-03-15]
4. Stacey D, Bennett C, Barry M, Col N, Eden K, Holmes-Rovner M, et al. Decision aids for people facing health treatment or screening decisions. Cochrane Database Syst Rev 2011(10):CD001431. [doi: 10.1002/14651858.CD001431.pub3] [Medline: 21975733]
5. Bruel S, Leclercq T, Ginzarly M, Botelho-Nevers E, Frappé P, Gagneux-Brunon A. Patient decision aid in vaccination: a systematic review of the literature. Expert Rev Vaccines 2020;19(4):305-311 [FREE Full text] [doi: 10.1080/14760584.2020.1742111] [Medline: 32163307]
6. Vujovich-Dunn C, Kaufman J, King C, Skinner SR, Wand H, Guy R, et al. A systematic review and meta-analysis of effectiveness of decision aids for vaccination decision-making. Vaccine 2021;39(28):3655-3665. [doi: 10.1016/j.vaccine.2021.05.021] [Medline: 34052064]
7. Promoting informed decision making about maternal pertussis vaccination among pregnant women. CCMO. Research with human participants. URL: https://onderzoekmetmensen.nl/en/trial/25018 [accessed 2025-06-19]
8. Anraad C, van Empelen P, Ruiter RA, van Keulen H. Effects of an online tailored decision aid to promote informed decision making about maternal pertussis vaccination in the Netherlands: a randomized controlled trial. Vaccine 2023;41(49):7348-7358. [doi: 10.1016/j.vaccine.2023.10.068] [Medline: 37977943]
9. Moore GF, Audrey S, Barker M, Bond L, Bonell C, Hardeman W, et al. Process evaluation of complex interventions: medical research council guidance. BMJ 2015;350:h1258 [FREE Full text] [doi: 10.1136/bmj.h1258] [Medline: 25791983]
10. Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M, Medical Research Council Guidance. Developing and evaluating complex interventions: the new medical research council guidance. BMJ 2008;337:a1655 [FREE Full text] [doi: 10.1136/bmj.a1655] [Medline: 18824488]
11. Saunders RP, Evans MH, Joshi P. Developing a process-evaluation plan for assessing health promotion program implementation: a how-to guide. Health Promot Pract 2005;6(2):134-147. [doi: 10.1177/1524839904273387] [Medline: 15855283]
12. Vlemmix F, Warendorf JK, Rosman AN, Kok M, Mol BWJ, Morris JM, et al. Decision aids to improve informed decision-making in pregnancy care: a systematic review. BJOG 2013;120(3):257-266. [doi: 10.1111/1471-0528.12060] [Medline: 23145991]
13. Pot M, Paulussen TG, Ruiter RA, Mollema L, Hofstra M, Van Keulen HM. Dose-response relationship of a web-based tailored intervention promoting human papillomavirus vaccination: process evaluation of a randomized controlled trial. J Med Internet Res 2020;22(7):e14822 [FREE Full text] [doi: 10.2196/14822] [Medline: 32520718]
14. Kelders SM, Van Gemert-Pijnen JEWC, Werkman A, Nijland N, Seydel ER. Effectiveness of a web-based intervention aimed at healthy dietary and physical activity behavior: a randomized controlled trial about users and usage. J Med Internet Res 2011;13(2):e32 [FREE Full text] [doi: 10.2196/jmir.1624] [Medline: 21493191]
15. Anraad C, van Empelen P, Ruiter RAC, Rijnders M, van Groessen K, van Keulen HM. Promoting informed decision making about maternal pertussis vaccination: the systematic development of an online tailored decision aid and a centering-based group antenatal care intervention. Front Public Health 2024;12:1256337 [FREE Full text] [doi: 10.3389/fpubh.2024.1256337] [Medline: 38425460]
16. Bartholomew Eldredge LK, Markham CM, Ruiter RAC, Fernandez ME, Kok G, Parcel GS. Planning Health Promotion Programs: An Intervention Mapping Approach. San Francisco: Jossey-Bass & Pfeiffer Imprints, Wiley; 2016.
17. Nielsen J, Landauer TK. A mathematical model of the finding of usability problems. In: J Diabetes Sci Technol. 1993 Presented at: INTERCHI93: Conference on Human Factors in Computing; 1993 April 24 - 29; Amsterdam The Netherlands p. 206-213. [doi: 10.1177/19322968211018236]
18. Petty RE, Barden J, Wheeler SC. The elaboration likelihood model of persuasion: developing health promotions for sustained behavioral change. In: Emerging Theories in Health Promotion Practice and Research. Hoboken, NJ: Jossey-Bass/Wiley; 2009:185-214.

XSL•FO
RenderX

19. Bandura A. Social Foundations of Thought and Action: A Social Cognitive Theory. New York, NY: Prentice Hall; 1986.
20. Kazdin AE. Behavior Modification in Applied Settings: Seventh Edition. Long Grove, Illinois: Waveland Press; 2012.
21. Latham GP, Locke EA. New developments in and directions for goal-setting research. European Psychologist 2007;12(4):290-300. [doi: 10.1027/1016-9040.12.4.290]
22. Miller WR, Rollnick S. Motivational Interviewing: Helping People Change. New York: Guilford Press; 2012.
23. Forman J, Heisler M, Damschroder LJ, Kaselitz E, Kerr EA. Development and application of the RE-AIM QuEST mixed methods framework for program evaluation. Prev Med Rep 2017;6:322-328 [FREE Full text] [doi: 10.1016/j.pmedr.2017.04.002] [Medline: 28451518]
24. Anraad C, Lehmann BA, Visser O, van Empelen P, Paulussen TG, Ruiter RA, et al. Social-psychological determinants of maternal pertussis vaccination acceptance during pregnancy among women in the Netherlands. Vaccine 2020;38(40):6254-6266. [doi: 10.1016/j.vaccine.2020.07.047] [Medline: 32782189]
25. van Stralen MM, de Vries H, Mudde AN, Bolman C, Lechner L. Efficacy of two tailored interventions promoting physical activity in older adults. Am J Prev Med 2009;37(5):405-417. [doi: 10.1016/j.amepre.2009.07.009] [Medline: 19840695]
26. Stoyanov SR, Hides L, Kavanagh DJ, Zelenko O, Tjondronegoro D, Mani M. Mobile app rating scale: a new tool for assessing the quality of health mobile apps. JMIR Mhealth Uhealth 2015;3(1):e27 [FREE Full text] [doi: 10.2196/mhealth.3422] [Medline: 25760773]
27. Spook J, Paulussen T, Kok G, van Empelen P. Evaluation of a serious self-regulation game intervention for overweight-related behaviors ("Balance It"): a pilot study. J Med Internet Res 2016;18(9):e225 [FREE Full text] [doi: 10.2196/jmir.4964] [Medline: 27670222]
28. Sedgwick P. Multiple significance tests: the bonferroni correction. BMJ 2012;344(jan25 4):e509-e509. [doi: 10.1136/bmj.e509]
29. Core Team R. R: A language and environment for statistical computing. R Foundation for Statistical Computing. URL: https://www.R-project.org/ [accessed 2025-03-15]
30. The Netherlands in numbers: infographics inspired by 36 questions. Centraal Bureau voor de Statistiek. 2022. URL: https://longreads.cbs.nl/the-netherlands-in-numbers-2022/how-old-are-dutch-first-time-mothers-on-average/ [accessed 2024-06-06]
31. Centraal bureau voor de statistiek. CBS Open data StatLine. 2022. URL: https://opendata.cbs.nl/statline/portal.html?_la=nl&_catalog=CBS [accessed 2022-03-14]
32. Nguyen MH, Smets EMA, Bol N, Loos EF, Van Weert JCM. How tailoring the mode of information presentation influences younger and older adults' satisfaction with health websites. J Health Commun 2018;23(2):170-180 [FREE Full text] [doi: 10.1080/10810730.2017.1421729] [Medline: 29345531]
33. Preuhs K, Bussink-Voorend D, van Keulen HM, Wildeman I, Hautvast J, Hulscher M, et al. An online tailored COVID-19 vaccination decision aid for Dutch citizens: development, dissemination, and use. JMIR Form Res 2024;8:e56390 [FREE Full text] [doi: 10.2196/56390] [Medline: 39475719]
34. Trauzettel-Klosinski S, Dietz K, IReST Study Group. Standardized assessment of reading performance: the new international reading speed texts IReST. Invest Ophthalmol Vis Sci 2012;53(9):5452-5461. [doi: 10.1167/iovs.11-8284] [Medline: 22661485]
35. Anraad C, van Keulen H, Lehmann B, Mollema L, van Empelen P, Ruiter R. Kinkhoestvaccinatie tijdens de zwangerschap. TSG Tijdschr Gezondheidswet 2020;98(3):101-106. [doi: 10.1007/s12508-020-00266-8]
36. Wilson RJ, Paterson P, Jarrett C, Larson HJ. Understanding factors influencing vaccination acceptance during pregnancy globally: a literature review. Vaccine 2015;33(47):6420-6429 [FREE Full text] [doi: 10.1016/j.vaccine.2015.08.046] [Medline: 26320417]
37. Gilbert RM. Inclusive Design for a Digital World: Designing with Accessibility in Mind. New York City: Apress; 2019.
38. Krebs P, Prochaska JO, Rossi JS. A meta-analysis of computer-tailored interventions for health behavior change. Prev Med 2010;51(3-4):214-221 [FREE Full text] [doi: 10.1016/j.ypmed.2010.06.004] [Medline: 20558196]
39. Pot M, Paulussen TG, Ruiter RA, Eekhout I, de Melker HE, Spoelstra ME, et al. Effectiveness of a web-based tailored intervention with virtual assistants promoting the acceptability of HPV vaccination among mothers of invited girls: randomized controlled trial. J Med Internet Res 2017;19(9):e312 [FREE Full text] [doi: 10.2196/jmir.7449] [Medline: 28877862]
40. Vader S, Uiters E, van der Lucht F, Smits C, Kroese F, de Bruin M. [Vaccination willingness and educational level]. TSG 2022;100(Suppl 1):40-43 [FREE Full text] [doi: 10.1007/s12508-021-00317-8] [Medline: 34690537]
41. Sadler GR, Lee H, Lim RS, Fullerton J. Recruitment of hard-to-reach population subgroups via adaptations of the snowball sampling strategy. Nurs Health Sci 2010;12(3):369-374 [FREE Full text] [doi: 10.1111/j.1442-2018.2010.00541.x] [Medline: 20727089]

Abbreviations

IDM: informed decision-making
MPV: maternal pertussis vaccination
NIP: National Immunisation Programme

XSL•FO
RenderX

RCT: randomized controlled trial

Edited by A Mavragani; submitted 10.07.23; peer-reviewed by I Mesters, B Eshrati; comments to author 16.02.24; revised version received 07.06.24; accepted 16.01.25; published 08.07.25.

Please cite as:
Anraad C, van Empelen P, Ruiter RAC, van Keulen HM
Online Tailored Decision Aid for Maternal Pertussis Vaccination in a Randomized Controlled Trial: Process Evaluation Study
JMIR Form Res 2025;9:e50709
URL: https://formative.jmir.org/2025/1/e50709
doi:10.2196/50709
PMID:

©Charlotte Anraad, Pepijn van Empelen, Robert AC Ruiter, Hilde M van Keulen. Originally published in JMIR Formative Research (https://formative.jmir.org), 08.07.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Multidomain Behavioral Change Digital Coaching for Chronic Disease Management in Patients With Type 2 Diabetes: Framework Development and Preliminary Evaluation

Konstantina Kostopoulou[1], MS; Danae Lekka[1], MS; Aristodemos Pnevmatikakis[1], PhD; Nelina Angelova[1], MS; Panagiotis Stafylas[1,2], MD, PhD; Stefanos Tamouridis[3], MD, MS; Alexandra Bargiota[3], MD, PhD; Sofoklis Kyriazakos[1], PhD

[1]Innovation Sprint Srl, Brussels, Belgium
[2]Healthink, Thessaloniki, Greece
[3]University of Thessaly, Volos, Greece

Corresponding Author:
Konstantina Kostopoulou, MS
Innovation Sprint Srl
Clos Chapelle-aux-Champs 30
Bte. 1.30.30
Brussels, 1200
Belgium
Phone: 32 47729137
Email: kkostopoulou@innovationsprint.eu

Abstract

Background: Unhealthy lifestyle behaviors have been identified as a major cause of numerous health issues, with a steady global increase in their prevalence. Addressing this challenge requires comprehensive behavioral changes to promote the adoption of a sustainable healthier lifestyle. However, despite the prevalent need, cost-effective and successful digital coaching for health-related behavior change remains scarce.

Objective: This study aimed to present a holistic framework for designing, modeling, and executing behavior change strategies through a multiagent reasoning system that selected optimal digital coaching techniques based on individual assessments and integrated data-driven decision-making.

Methods: Behavioral change theories have been explored to design a multiagent system aimed at achieving sustainable lifestyle changes. This system selected behavior change techniques based on individual user assessments, prioritizing those with the strongest impact on key behavioral components. The framework incorporated evidence-based practices stemming from behavioral change science and integrated them into Healthentia's behavioral change coaching scheme. Healthentia, a certified software as a medical device, implemented this framework in its non-medical modules that aim for lifestyle behavioral change and wellbeing specifically for chronic disease management, serving as an eHealth solution that advances decentralized care by enabling remote monitoring, data-driven content selection, and personalized digital coaching that adjusts to patient progress and engagement patterns.

Results: This study explored the application of the Healthentia behavioral change coaching scheme in patients with type 2 diabetes. Behavioral attributes have been evaluated in 9 patients, yielding notable results in terms of fasting glucose dropping by an average of –17.3 mg/dL (Cohen d=1.5; P=.002), further underscored by a narrow 95% CI (–26.1 to –8.43), and in terms of weight and BMI, with mean reductions of –2.89 kg and –1.05 kg/m², respectively. These changes yielded large effect sizes (Cohen d approximately 1.05) and were statistically significant (P=.01). The positive outcomes were at least partly attributed to the personalized delivery of content, 71.66% (1125/1570) of which was well received by the patients.

Conclusions: Our study of this multiagent system, which was tested through simulated patient behavior and preliminary, limited behavior observations of patients with type 2 diabetes, promises improved health outcomes using personalized digital coaching strategies. Future directions include optimizing the multiagent selection process; further exploring the type 2 diabetes program; conducting an in-depth evaluation of its results, including glycated hemoglobin measurements; and expanding its applications to other chronic conditions.

XSL•FO
RenderX

(JMIR Form Res 2025;9:e73807) doi:10.2196/73807

KEYWORDS

behavior change; mobile health; mHealth; remote patient monitoring; RPM; digital therapeutics; DTx; software as a medical device; SaMD; Healthentia; artificial intelligence; AI

Introduction

Background and Significance

The escalating global prevalence of lifestyle diseases, such as obesity, diabetes, and cardiovascular conditions, has become a critical health challenge that demands innovative digital coaching to enable behavioral change. These chronic conditions, exacerbated by sedentary lifestyles and accompanied by poor dietary habits, insufficient physical activity, and unhealthy habits such as smoking, reflect broader societal patterns of behavior that compromise population health and increase premature mortality. These diseases combined are responsible for >70% of global deaths.

Traditional medical approaches focused solely on treatment have proven insufficient in addressing the root causes of these health challenges. A more comprehensive approach is necessary to address the complex interplay between individual behaviors, environmental factors, and systemic health determinants. Healthentia (product of Innovation Sprint) attempts to achieve this by integrating behavioral change digital coaching on top of remote standard-of-care monitoring offered to patients with chronic diseases, presenting a promising approach to targeting the underlying mechanisms that contribute to unhealthy lifestyle choices.

Technological Approach

Healthentia platform [1-4] is a class IIa–certified software as a medical device intended for (1) the collection and transmission of physiological data, including heart rate, blood pressure, oxygen saturation, and weight, directly to health care providers via automated electronic means, in combination with validated Internet of Things devices; (2) the visualization (patient-based dashboards) and the mathematical treatment of data (trends analysis and alerts) related to the monitored physiological parameters of patients with chronic diseases; (3) the transmission of patient's outcomes and outcome scores related to patient's health status, health-affecting factors, health-related quality of life, disease knowledge, and adherence to treatment through validated questionnaires; and (4) user (patient) interaction with a conversational digital coach for informative and motivational purposes, supporting patient telemonitoring, decision-making, and behavioral coaching.

Healthentia has 2 components: a mobile app for patients and a web portal for health care professionals. The platform can be customized for different programs and therapeutic areas with a collection of medical and nonmedical modules. The medical modules of the device are intended to support patient telemonitoring, decision-making, and digital coaching by collecting and transmitting physiological data through validated Internet of Things devices. They also enable data visualization and their mathematical treatment, transmission of patient outcomes through validated questionnaires, and interaction with a conversational digital coach. Further to the remote patient monitoring features of Healthentia, which are part of the medical modules under the CE-mark, other non-medical modules such as the smart services, drive digital therapeutics features that aim to improve patients' lifestyle and wellbeing using a novel, behavioral change–based digital coaching approach.

Conceptual Framework

To empower individuals to make sustainable lifestyle changes, we need to understand and modify the cognitive, motivational, and environmental factors that influence health behaviors. To accomplish this, the theoretical frameworks that apply to behavioral change are examined. An essential element of success for lifestyle changes is to achieve patient empowerment by training them in self-management of their condition, enhancing therapy adherence, and improving health outcomes. The use of smartphones, which have become essential daily tools, has emerged as an effective strategy for promoting self-management by providing access to health information and coaching [5]. In recent years, many mobile health apps delivering personalized digital coaching have emerged, offering cost-effective and convenient at-home support; however, most of them lack embedded evidence-based behavior change theories, limiting their impact on long-term, sustainable health outcomes [6].

The proposed framework aims to address multiple gaps within the existing landscape of digital coaching. By implementing multidomain behavioral change agents, the framework seeks to provide comprehensive support tailored to individual patient needs. Many current digital coaching systems lack real-time evaluations and personalized feedback, particularly in home-based settings. Immediate or close to immediate feedback, combined with emotional and psychological support, creates a framework that offers holistic care by addressing various aspects of well-being through agents and their multidimensional parameters. Furthermore, addressing the notable gap in seamless integration of digital coaching systems with established clinical pathways, the framework aligns the digital coaching system with existing clinical protocols, improving patient outcomes by enhancing the continuity of care.

Objectives and Hypothesis

In this study, we aimed to present a framework for personalized behavior change coaching (BCC), which integrates widely used theoretical models, and describe the implementation of this framework within the Healthentia platform for type 2 diabetes mellitus (T2DM) management. A detailed process of each phase of this framework has been described, establishing its conceptual and operational foundations. In addition, preliminary results from an ongoing pilot study have been shared to assess feasibility, engagement metrics, and initial clinical outcomes. While the preliminary results can provide suggestive trends and early indications, full validation of this framework will require more data collection and analysis. We hypothesized that

XSL•FO
RenderX

personalized digital coaching, rooted in behavioral science and aligned with an individual's unique beliefs, capabilities, and motivations, would demonstrate higher engagement and retention rates, potentially enhancing patient self-management strategies [7].

Methods

Overview

Developing, implementing, and evaluating behavior change techniques (BCTs) to change established behaviors can be challenging. A robust theoretical framework is essential for backing up any type of behavior change, as it enables predicting the desired outcomes, guides the digital coaching design by establishing the success criteria, and facilitates empirical evaluation and replication.

According to Michie et al [8] and the UK Medical Research Council, a strategic, systematic approach for designing behavioral change programs consisted of 3 phases, grounded in scientific understanding of behavior change mechanisms and moving from theoretical conceptualization to practical implementation [8]. First was the "theory" phase, where the theoretical basis of the BCC program was developed. The "modeling" phase hypothesized the behavioral determinants and identified possible targets and ways of changing them by testing the BCTs. The final one was the "experimental" phase, during which exploratory trials assessed the findings. This part of the paper analyzed these 3 phases and the way they were incorporated into our overall framework.

Theory Phase

Overview

To design behavioral change–based digital coaching, we looked for theories following a systematic approach, as there is evidence supporting that following theoretical constructs is more effective than relying solely on factors such as age or sex [8]. Although there exist many theories and models for designing behavioral change approaches, there is limited guidance on how to effectively implement them in practice.

Our systematic approach integrated 3 complementary theoretical frameworks. The *Behaviour Change Wheel (BCW)* was used to design the program based on unique behavioral, population, and setting characteristics [9,10], linking evidence-based intervention functions and digital coaching methods to a behavioral model at the core. This core model was the *capability, opportunity, motivation, and behavior (COM-B) model*, the second framework used, which helped us understand behaviors in terms of their determinants, further linking them to BCTs as their implementation strategies. The last framework that was integrated into our approach was the *belief-desire-intention (BDI)* framework, which was used to model the interactions between a coachee's multiple risk behaviors, a coach's decision-making process for selecting the most appropriate technique or content to achieve the optimum outcome, and the actual expected effect on the behavior.

All these frameworks had distinct yet complementary roles across the 3 phases introduced in the Overview section. In the theory phase, the COM-B model defined the behavioral determinants that had to be targeted. BCW mapped COM-B components to specific intervention functions, and the BDI framework established the foundation for modeling decision-making processes.

In the modeling phase, the COM-B model was the basis for user assessment, translating raw data into their behavioral constructs. BCW guided the selection and prioritization of BCTs, and finally, BDI provided the computational architecture for simulating different patient behaviors and coaching responses.

In the experimental phase, the COM-B model was again used as the patient assessment framework; however, it was also used to measure the real-time changes in behavioral determinants. BCW was used in a manner similar to the modeling phase, and the BDI framework drove the real-time selection and adaptation of coaching strategies.

More details of the modeling and experimental phases have been provided in the subsequent sections. A detailed explanation of the theoretical frameworks has been presented in the remainder of this section.

BCW and the COM-B Model

BCW synthesized 19 different frameworks into a single comprehensive approach of 2 layers: the intervention *function layer* comprised 9 functions (ie, education, persuasion, incentivization, coercion, training, enablement, modeling, environmental restructuring, and restrictions) and the *policy categories layer* comprised 7 categories, linked at their core with the COM-B model.

The COM-B model identified 3 essential components for behavior change: *capability* (physical and psychological), which evaluated the knowledge and skills required and the physical ability to perform; *opportunity* (social and physical), which identified environmental factors and social influences; and *motivation* (reflective and automatic), which reflected the conscious decision-making as well as the unconscious responses and emotional reactions [11]. BCW defined behavior as an interaction between the 3 necessary components of the COM-B model [9].

The BCW design process has 3 stages, as shown in Figure 1.

XSL•FO
RenderX

Figure 1. Design process. BCT: behavior change technique; BCW: Behaviour Change Wheel; COM-B: capability, opportunity, motivation, and behavior.

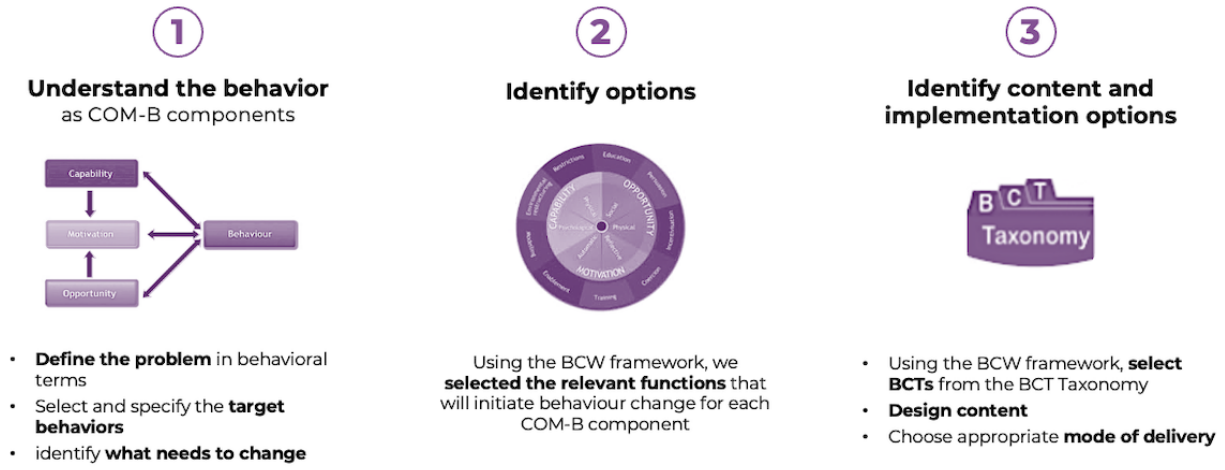

The COM-B components assessed behavior in terms of theoretical behavioral domains, identifying which human behavioral aspects could be influenced to achieve behavior change. These domains could then be targeted through multidimensional processes called "mechanisms of actions" [12], which drove the behavior change and allowed the understanding of "why" certain BCTs changed behaviors, giving us proof of effectiveness. The theoretical domains were addressed by implementing BCTs that addressed the components of ≥1 behaviors separately or simultaneously.

BCTs were identified as standardized digital coaching drawn from a validated taxonomy of 93 identified available techniques [13]. Literature indicated that the combination of multiple BCTs was generally more effective than their single, stand-alone use; therefore, we selected promising combinations based on the findings and tools from the Theory and Techniques of Behavior Change Project [14], effectively linking BCTs with their mechanisms of action.

Literature examples of combining "habit formation" with "self-monitoring" or "environmental restructuring" with "goal setting" showed a particular promise [15,16], as did "self-monitoring" with "goal setting" [17]. However, it is important to mention that the effectiveness of specific BCTs could vary based on the targeted behavior, the individual, and the context. The way and reason behind choosing a technique and the way we monitored their effectiveness to adjust them if needed have been explored further in the *Healthentia BCC* section.

BDI Framework

The last challenge was to find a theoretical paradigm to model behavioral change interactions. To determine how a coaching solution could decide which type of digital coaching should be used based on a desired targeted outcome, the theoretical paradigm of the BDI framework was used. The BDI paradigm was developed by Bratman [18] in the 1980s, explaining rational decision-making processes by modeling how agents selected and committed to specific actions based on their understanding of the environment and their desired outcomes.

The BDI framework consisted of three interconnected mental states:

1. Beliefs represented an agent's knowledge about the world
2. Desires articulated the agent's goals or objectives
3. Intentions described the planned actions to achieve those desires

This framework provided a structured approach toward understanding how rational agents made decisions, offering insights into goal-directed behavior by demonstrating how intentions emerged from an agent's existing beliefs and desires.

Healthentia BCC

Overview

The BCC cycle was the framework that we applied to our solution, integrating the theoretical frameworks mentioned in the Theory Phase section in terms of assessing, modeling, and optimizing behavior change digital coaching. "Digital coaching" is essentially the application of structured BCTs and their digital delivery through the conversational agent of the mobile app, as detailed subsequently.

The framework is depicted in Figure 2 and divided into 4 main steps: assess, predict, tailor, and run and adjust.

XSL•FO
RenderX

Figure 2. Behavior change coaching cycle.

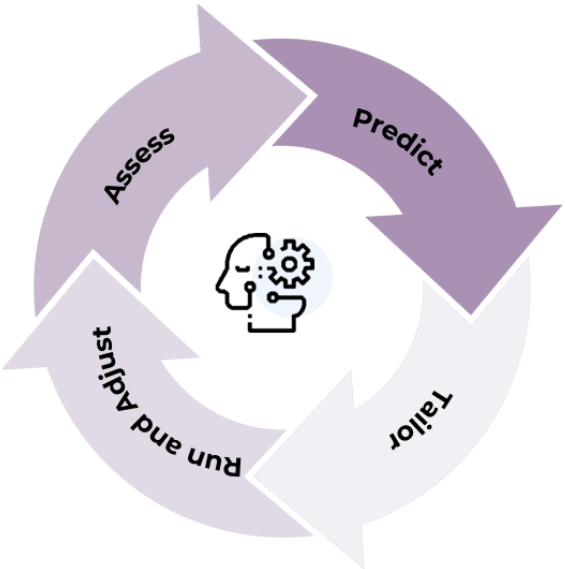

Step 1: Assess

In this step, reference patient data are collected and calculated to establish a baseline state A, providing insights into their behavior and health status and identifying their key behavioral determinants and characteristics. The data collected in this step might include clinical (health metrics and medical history), demographic (educational, marital, and environmental details), behavioral (baseline activity steps and intensity minutes), and psychosocial indicators (mental state, motivation, social support, and environmental factors).

After data collection, we proceeded to feature engineering, where the collected data are normalized and mapped to COM-B constructs, linking initial behavioral determinants to key domains, such as physical activity, sleep, nutrition, and self-care, and ensuring consistency and relevance for the subsequent modeling phase by effective feature structuring. It focused on the following constructs:

- Capability analysis: evaluating knowledge, skills, and physical attributes
- Opportunity analysis: assessing social and environmental factors
- Motivation analysis: understanding reflective and automatic motivations

Step 2: Predict

In this step, the individual feature vector from state A was used, and state B was predicted as the desired outcome after a 12-week digital coaching period.

Step 3: Tailor

In this step, based on the 2 states (A and B), a tailored digital coaching process was created, with the selection of domains and specific BCTs based on COM-B scores personalized to the patient.

Step 4: Run and Adjust

In this step, the BCTs were applied following the tailored digital coaching plan. The system continuously monitored patient behavior to detect deviations and reassessed the digital coaching plan if deviations exceeded acceptable thresholds.

Modeling Phase

Overview

In the theory phase, the theoretical frameworks of BCW, COM-B, and BDI were introduced, and the integration of these theoretical models into the Healthentia BCC framework was explained. Moving into the modeling phase, the methodology of how we modeled the Healthentia BCC framework and integrated it into our complete product offering has been described as follows.

Healthentia included predefined therapeutic program templates for lifestyle diseases, such as T2DM, heart failure, or chronic obstructive pulmonary disease, and BCC was provided in addition to the capabilities of a standard telemonitoring tool, addressing risk behaviors, such as poor diet, physical inactivity, inadequate sleep, emotional health challenges, and smoking, to improve outcomes. These templates were tailored to the individual needs as part of their personalized program journey.

Lifestyle Therapeutic Program Design

In Healthentia, the therapeutic programs designed for lifestyle diseases used a standard-of-care clinical protocol for remote monitoring, which consisted of collecting patient outcomes in the form of surveys and measurements or vital signs at a specified frequency. Each program was orchestrated by an open-source business process modeling and notation tool, which scheduled events and reminders while also streamlining the different phases of the program with all system interactions. The workflows navigated a patient through the requirements of each program phase for a specific timeline, beginning with the assessment phase, where all initial data collection occurred.

For each therapeutic program of a lifestyle disease, we identified the risk behaviors to be targeted using the BCW, as described in the BCW and COM-B Model section. This process helped us address and influence behavior change for all potential needs

XSL•FO
RenderX

of patients in the program, by identifying all suitable functions, BCTs, and related content.

Behavioral Change Digital Coaching Modeling Design

After having identified the targeted behaviors and BCTs for a specific therapeutic program, a way to initiate the system with a starting plan for real-time execution was needed. In this section, we describe how the BDI paradigm was used to simulate real-world health plans, where digital coaches worked alongside individuals to achieve their goals while considering human factors, such as capability, motivation, fatigue, and personal expectations.

The BCTs were the implementation strategies of digital coaching and could be classified into 3 categories in the system based on their mode of delivery. A BCT could take the form of content (article or media), such as "Instruction on how to perform a behavior." It could also be a systemic function or feature, such as "Discrepancy between current behavior and goal," where the system calculated the difference, as well as an interactive dialogue, such as "Problem-Solving," using the chatting feature.

For the reader's convenience, a comprehensive glossary of all BCTs mentioned in this paper, including definitions, implementation examples, and mode of delivery, has been provided in Multimedia Appendix 1.

First, a simulating platform was used to model BDI agents per behavioral domain, representing a rational decision-making entity to select the digital coaching and a coachee model to simulate behavioral responses based on the applied plan for different BCTs.

The 3 primary components of each BDI coach agent were as follows:

1. Beliefs:
   inclusion of the current state of the coachees' behaviors, performance factors, history, and environmental context
2. Desires:
   inclusion of the target behavioral outcomes, optimal COM-B states, and priorities
3. Intentions:
   selection of the appropriate BCTs, constructing a plan, and monitoring it throughout the digital coaching period

The development was initiated by a physical activity coach inspired by Taj et al [19], who simulated a health coaching system that aimed to promote physical activity. Coaching agents for other behavior domains (nutrition, sleep, and self-care) were designed similarly. Some BCTs that we included in this first BDI model were for beliefs (2.1: monitoring of behavior without feedback), desires (1.1 and 1.4: goal setting and planning), and intentions (1.6: discrepancy between current behavior and goal; 2.2: feedback on behavior; 10.3: nonspecific reward; 5.1: information about health consequences, 5.3: information about social and environmental consequences, and 5.6: information about emotional consequences; 6.1: demonstration of the behavior; and 4.1: instruction on how to perform a behavior).

A coachee model represented an individual receiving digital coaching and was modeled based on COM-B constructs, performance factors (eg, fatigue and recovery) across various behavioral domains, and response mechanisms to digital coaching. The modeling of the coachee behavior was inspired by Howlett et al [20], who examined how COM-B components influenced moderate to vigorous physical activity. In this study, COM-B components were mapped to the theoretical behavioral determinants and were measured through validated or custom questionnaires that were benchmarked against a behavioral outcome a week later. In summary, the coachee model depicted how BCTs impact an individual's COM-B components through direct capability enhancement, opportunity modification, and motivation strengthening.

The effectiveness of digital coaching was assessed and optimized to achieve the predicted behavioral outcome. By testing different combinations of BCTs, the system identified the most effective digital coaching plan for achieving the desired behavior changes. The simulation outputs included predicted outcomes of behaviors across key domains; recommended BCTs and their frequency of use; and generated insights into recovery capacity, resilience, and maintenance potential. These outputs were refined and stored as a starting plan for real-time execution.

BCC Initialization

Established models were ready to be fed into the BCC cycle, which provided personalized coaching for enrolled users. After the onboarding of a patient to a program that had behavioral coaching enabled, the process described in the theory phase was initialized by starting with a 2-week long *initialization* phase to collect and process data and assess the patient's capability, opportunity, and motivation (COM-B) required to build an understanding of the patient's baseline behaviors. This ensured that digital coaching was tailored to the individual, increasing the likelihood of success.

After the completion of the initialization, the iterative process of the BCC cycle started with steps 1 and 2, and the digital coaching phase kicked off with steps 3 and 4 with personalized coaching. Figure 3 presents the integration of the BCC cycle within the whole product through its connections with other functionalities of the system, such as the mobile app, the dialogue editor, and the executor or the process automation business process modeling and notation tool.

XSL•FO
RenderX

Figure 3. Healthentia behavior change coaching solution.

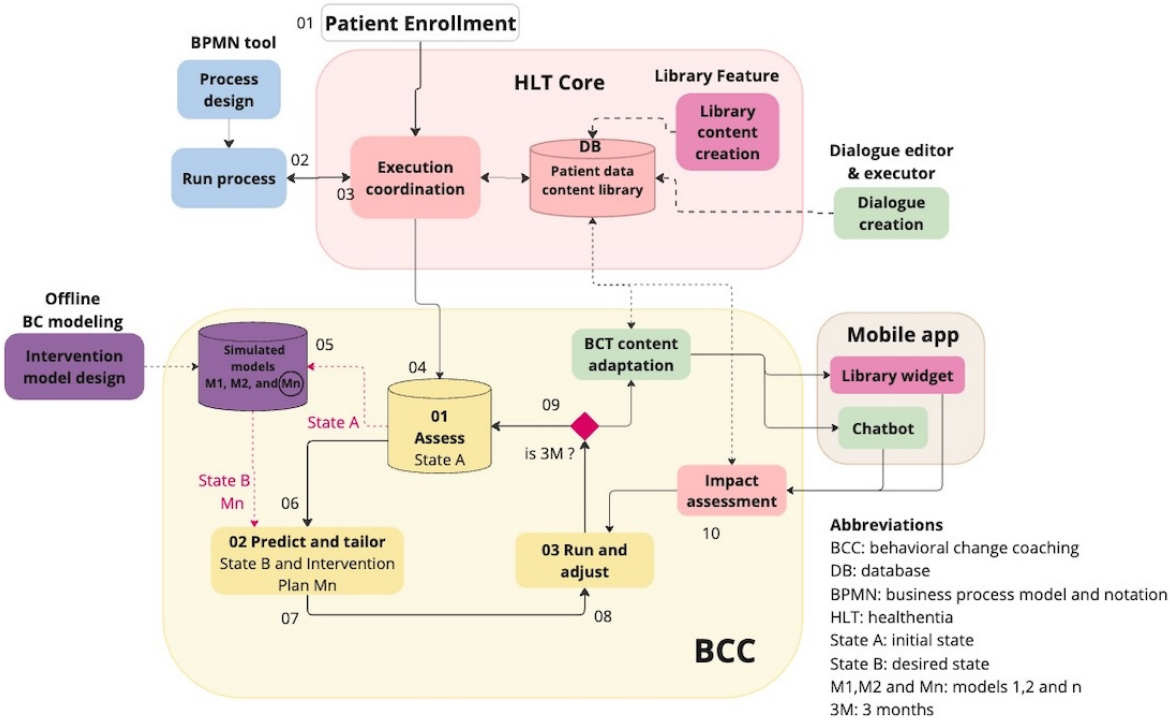

Digital coaching was delivered to the patients through an artificial intelligence–driven conversational agent, “the coach.” Its technical implementation used a rule-based engine embedded within the Healthentia ecosystem and operated autonomously without human intervention, using the BDI framework to select appropriate content and timing based on patient data. The coach could deliver scheduled educational content and respond adaptively to patient-initiated interactions.

It was designed by a multidisciplinary team, including behavioral scientists with expertise in digital health interventions, clinicians specializing in diabetes and cardiovascular disease management, and software engineers.

In practice, the coach is used to deliver the personalized content selected and to engage the users through dialogues. This method provides a more interpersonal and convincing approach to advising users, offers feedback on progress, and addresses barriers or facilitators to help achieve the desired progress toward the targeted behavior. It also allowed for interactive data collection, facilitating continuous adaptation of the digital coaching plan. The frequency and content of interactions between the patient and coach were dynamically adjusted based on each user’s digital coaching plan, needs, and preferences.

The coach was integrated into the mobile app, where users engaged with the delivered content in a chat-based interface. When the BCC system determined that specific content should be shared with the user, it was delivered via the coach as a notification. Users could tap on the notification to expand it, revealing a structured conversational flow that guided them through personalized health information, self-care tips, and digital coaching strategies.

Figure 4 shows an example of the conversational agent in the mobile app, where the coach provides information about diabetes mellitus complications. It illustrates the delivery of educational content in an engaging manner, aiming to encourage users to explore further details and assess the content delivered through the mobile app.

XSL•FO
RenderX

Figure 4. An example of the digital coach in the Healthentia mobile app, assisting a patient with diabetes through an interactive conversation.

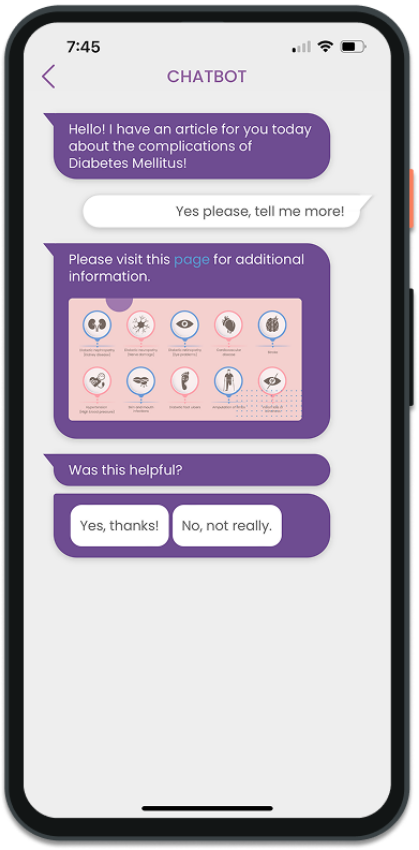

Experimental Phase

Study Design and Methods

A program for T2DM was co-designed with clinicians from the General University Hospital of Larissa leveraging BCC principles. This program followed a structured approach integrating behavior change digital coaching with remote clinical monitoring. It aimed to improve self-monitoring, motivation, and adherence, which in turn contributed to improved glycemic control and health outcomes. The objectives of this program were as follows:

- Primary clinical objective: to investigate the effect of the digital coaching on the glycated hemoglobin (HbA1c), as it was measured at the beginning and the end of the program
- Secondary clinical objective: to investigate the effect of digital coaching on blood glucose measurements using factors such as timely digital coaching and acceptance of recommendations
- Behavioral objectives: to investigate the effect of the digital coaching on weight (toward normal BMI range), physical activity (an increase in baseline step count and reaching at least 150 active min spent per week), and sleep (toward a duration between 6 and 8 hour)

Patients with T2DM who met the inclusion criteria and were being treated at the University Clinic of Endocrinology and Metabolic Diseases were recruited to the study. Each enrolled patient participated for a total duration of 14 weeks (2 control weeks of assessment and 12 weeks of behavior change digital coaching). During this period, each patient continued to be treated as usual, receiving standard medical care.

With regard to inclusion criteria, we are currently collecting data from a clinical study involving patients with T2DM, using the inclusion and exclusion criteria outlined in Textbox 1.

XSL•FO
RenderX

Textbox 1. Inclusion and exclusion criteria.

Inclusion criteria

- Glycated hemoglobin value >6.5% and access to a suitable mobile phone (smartphone) device that patients could adequately operate or a caregiver available to enter the patient's data into the app in case the patient was not familiar with it.
- Receiving one of the following treatments to manage their diabetes:
- Antidiabetic tablets
- Insulin or other injectable treatments
- A combination of the above

Exclusion criteria

- Unwillingness to have frequent capillary blood glucose measurements
- Refusal of written informed consent to participate in the study
- Currently using or having received a lifestyle intervention or modification related to diabetes mellitus within the past 12 months

The primary end point was the change in HbA_{1c} levels, while the secondary end points focused on the changes in physical activity, body weight and BMI, sleep habits, smoking cessation, dietary habits, health-related quality of life, and other laboratory parameters.

Following the BCC cycle of Figure 2, the program had an assessment phase and a coaching phase, which integrated the predict, tailor, and run and adjust steps.

The assessment phase had a dual goal:

1. To assess the patient by collecting baseline measurements and initial self-reports via behavioral questionnaires (see full list below, in the last paragraph of this section).
2. To empower the patient by introducing them to Healthentia for their diabetes program. Adherence was tracked by the system and was reported to the health care professional, who decided whether the patient was ready to progress to the behavior change digital coaching phase.

The coaching phase had the following goals:

- Personalized goal setting for steps, intensity minutes, and sleep based on the data collected from the assessment phase and the physician recommendations
- Continuous collection of patient data using a Fitbit tracker (Fitbit, Inc) and manual mobile app inputs
- Acting upon the collected data, the system both tailored the behavioral coaching and delivered it. The BCC system always monitored and evaluated data, prompting reminders and alerts and guiding the coach and its delivery content and timing.

While life-threatening situations were beyond the scope of Healthentia, the diabetes program added to the time-critical alert process for measurements by considering extreme blood glucose values.

Patient data were continuously collected and monitored throughout the BCC-driven diabetes program, including blood glucose measurements, weight measurements, questionnaire responses, and physical activity and sleep data. All data were logged and could be reviewed from Healthentia's mobile app.

During the first week of assessment, patients were asked to perform structured blood glucose measurements, as indicated by the Healthentia mobile app notifications. This not only provided a baseline for blood glucose levels but also familiarized patients with the app's notification system and self-reporting process. Throughout the coaching phase, patients continued to track blood glucose levels at predetermined times of the day with greater flexibility, using a pattern designed for patients not meeting glycemic targets, aiming for at least 6 measurements per week. Healthentia's BCC system monitored adherence, sending reminders when there were too few or not widely spread measurements. It also detected indications of hyperglycemia, encouraging patients to consult their physician if a single extreme value or consistently high measurements were encountered. The structured blood glucose measurements were repeated in the final week of the coaching phase.

Patients were expected to log into the app and enter their weight at least weekly during the coaching phase through a weight widget. Healthentia monitored weight values and sent notifications every Friday to patients who had not logged any values for the current week.

Physical activity and sleep data were collected using a Fitbit activity tracker. During assessment, the average daily steps were collected to establish each patient's baseline, which was then used to evaluate progress.

The program used 3 custom questionnaires (the COM-B questionnaire; Multimedia Appendix 2 [10]) and 9 validated ones (Fagerstrom Test for Nicotine Dependence, International Physical Activity Questionnaire, Problem Areas in Diabetes, Diabetes Knowledge Questionnaire-24, Diabetes Management Self-Efficacy Scale, EQ-5D, Mediterranean Diet Score, General Anxiety Disorder-7, and Patient Health Questionnaire-9). During the assessment phase, all questionnaires were first sent over the weekend of the first week to be answered during the second week. Reminders were sent periodically for unanswered questionnaires until they expired a week later. During the coaching phase, 3 questionnaires were repeated, offering updated information for the run and adjust step (Mediterranean Diet Score, General Anxiety Disorder-7, and Patient Health Questionnaire-9). During the final week of the coaching phase,

XSL•FO
RenderX

all questionnaires were resent, except the Fagerstrom Test, which was replaced by the Poststudy System Usability Questionnaire.

Coaching in Action

When the assessment period was completed, all the collected patient data were used to establish a baseline state A, highlighting their key behavioral determinants and characteristics mapped to COM-B constructs. Using expert knowledge and the modeling simulations described in the Methods section, a state B was established and used to prioritize behavioral domain needs by setting personalized goals and selecting the relevant BCTs and content associated with them (eg, tips and lessons).

The "coach," central to the run and adjust step, sustained patient engagement by delivering personalized content. It delivered three main types of content:

1. From 7 interactive dialogues on goals, 1 was selected and sent every morning to support patients in meeting their goals.
2. From 40 educational tips, 1 was selected and delivered every evening (except Sundays) to both support and educate the patients for improved disease management.
3. From 7 lessons, 1 was selected and sent each Sunday evening, offering more detailed educational material.

The content that could be delivered was dynamically adjusted based on each patient's needs and preferences. These messages were tailored based on the patient's progress and BCTs across 4 behavioral domains, including self-care, nutrition, physical activity, and sleep, as shown in Table 1.

Table 1. Classification of the content delivered to patients per type, domain, and function. The number of content items is shown in parentheses.

Content type and domain	BCTs[a]
Lessons (n=7)	
Self-care	4.1, 4.2, and 5.1
Tips (n=40)	
Self-care (n=1)	4.1 and 4.2
Self-care (n=9)	8.2, 7.1, and 15.1
Nutrition (n=4)	4.1 and 4.2
Nutrition (n=11)	8.2, 7.1, and 15.1
Physical activity (n=3)	4.1 and 4.2
Physical activity (n=3)	8.2, 7.1, and 15.1
Sleep (n=4)	4.1 and 4.2
Sleep (n=5)	8.2, 7.1, and 15.1
Interactive dialogues on goals (n=7)	
Self-care (n=2)	1.1-1.7, 2.1, 2.2, 2.6, and 2.7
Physical activity (n=3)	1.1-1.7, 2.1, 2.2, 2.6, and 2.7
Sleep (n=2)	1.1-1.7, 2.1, 2.2, 2.6, and 2.7

[a]BCT: behavior change technique.

Examples of some of the BCTs mentioned in Table 1 include the following:

- 4.1: instruction on how to perform a behavior (an article that demonstrated step-by-step how to monitor blood glucose and how to take proper medication [21] and an article that offered tips and guidance for better sleep patterns [22])
- 5.1, 5.3, and 5.6: information about health, social, environmental, and emotional consequences (small text tips were delivered as messages from the digital coach for each of these BCTs)

An example of an evidence-based tip on physical activity is shown in Figure 5. This and most of the physical activity content were based on the study by Colberg et al [23].

XSL•FO
RenderX

Figure 5. An example of a physical activity tip delivered to a patient with type 2 diabetes.

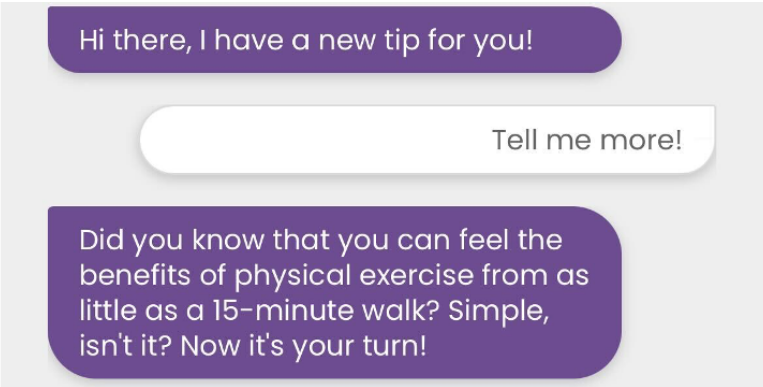

The full list of BCTs mentioned in Table 1 is reported in Multimedia Appendix 1. The chatbot interactions were purely text based at this point of the study. However, other media, such as audio or video, were planned to be included as content in the future.

Ethical Considerations

This open-label observational study, titled "Clinical Intervention Study for Diabetes Mellitus Type 2 Patients Based on Healthentia App," with an intraindividual control group, was approved by the General University Hospital of Larissa Scientific Committee on November 12, 2024 (protocol number 45937). The study is currently ongoing and implements the T2DM program.

Results

Study Overview and Demographics

The preliminary results from the "Clinical Intervention Study for Diabetes Mellitus Type 2 Patients Based on Healthentia App," which is still ongoing with new patients being enrolled continuously, are presented in this section. The target sample size of the study is >30 patients, which has not been reached; however, the already enrolled patients provided suggestive trends and early indications. Currently, 16 patients have been enrolled. Their demographic data in terms of sex, age, and BMI are provided in Table 2. Most (11/16, 69%) were female patients aged >60 years, characterized as obese-low weight category. Out of the 16 patients, 3 (19%) almost completed the intervention (≥10th week), 6 (38%) completed the intervention, and 5 (31%) had their final clinical data captured in the system.

XSL•FO
RenderX

Table 2. Distribution of participants by age group, sex, and weight category (N=16).

Characteristics	Participants, n (%)
Age group (y) and sex	
31-40	
Male	2 (12)
Female	1 (6)
41-50	
Male	0 (0)
Female	1 (6)
51-60	
Male	1 (6)
Female	2 (12)
61-70	
Male	1 (6)
Female	6 (38)
71-80	
Male	1 (6)
female	1 (6)
Weight category	
Normal to high	3 (19)
Overweight to low	1 (6)
Overweight to high	3 (19)
Obese to low	7 (44)
Obese to high	2 (12)

Coaching Outcomes and Statistical Analysis

The primary and secondary clinical objectives and behavioral objectives, described in the Study Design and Methods section, were assessed by 11 identified attributes. From these attributes, 6 were behavioral: weight, BMI, fasting glucose, sleep duration goal (as set by the physicians) achievement rate, steps walked, and minutes spent in moderate- or high-intensity activity. The sleep duration goal achievement was measured as the ratio of the days in a week the daily sleep duration goal was met. Intensity minutes were tracked weekly by Fitbit activity trackers and calculated as the sum of moderate-intensity minutes plus twice the number of high-intensity minutes. A total of 2 attributes were related to patient engagement, namely, the proportion of the content that patients completed (accessed to an acceptable degree) or accepted (provided positive feedback) out of the content delivered to them. The clinical objectives were assessed using 3 attributes, namely, HbA_{1c}, low-density lipoprotein, and estimated glomerular filtration rate.

In the preliminary results presented in this paper, we considered the 9 patients who had (almost) completed the intervention in terms of the behavioral and engagement attributes, while only 5 patients had provided final clinical data for the clinical attributes. Patients who decided to drop out were not included in our analysis. Note that more patients are being enrolled every week; we expect to have at least 30 patients complete the study. We aim to reduce sex bias in our future enrollments. In contrast, we do not expect the bias toward higher age and BMI to be alleviated because this is the typical population afflicted by T2DM, while the geographical bias (the city of Larissa and the surrounding rural areas) will be absolute.

To evaluate the evolution of key clinical, behavioral, and engagement-related attributes over the course of the intervention, we analyzed both the initial values and their variations between baseline (assessment) and intervention completion. Table 3 summarizes the statistics (mean and SD) of the baseline attribute values. The extent of the baseline attribute values is visualized in Figure 6.

XSL•FO
RenderX

Table 3. Baseline characteristics of measured attributes.

Attribute	Initial values, mean (SD)
Weight (kg)	93.4 (13.3)
BMI (kg/m^2)	33.2 (4.57)
Fasting glucose (mg/dL)	143 (23.1)
Sleep goal achievement rate	0.455 (0.192)
Steps, n	7530 (2120)
Intensity (min)	350 (477)
Completed content ratio	0.849 (0.121)
Accepted content ratio	0.730 (0.174)
HbA_{1c}[a] (%)	7.08 (1.13)
LDL[b] (mg/dL)	62.6 (9.53)
eGFR[c] (mL/min/1.73 m²)	105 (35)

[a]HbA_{1c}: glycated hemoglobin.

[b]LDL: low-density lipoprotein.

[c]eGFR: estimated glomerular filtration rate.

Figure 6. Visualization of the normalized extent of the baseline attribute values. To represent all ranges in the same graph, they are normalized in the way indicated by the multiplicative factors in the parentheses. eGFR: estimated glomerular filtration rate; HbA1c: glycated hemoglobin; LDL: low-density lipoprotein.

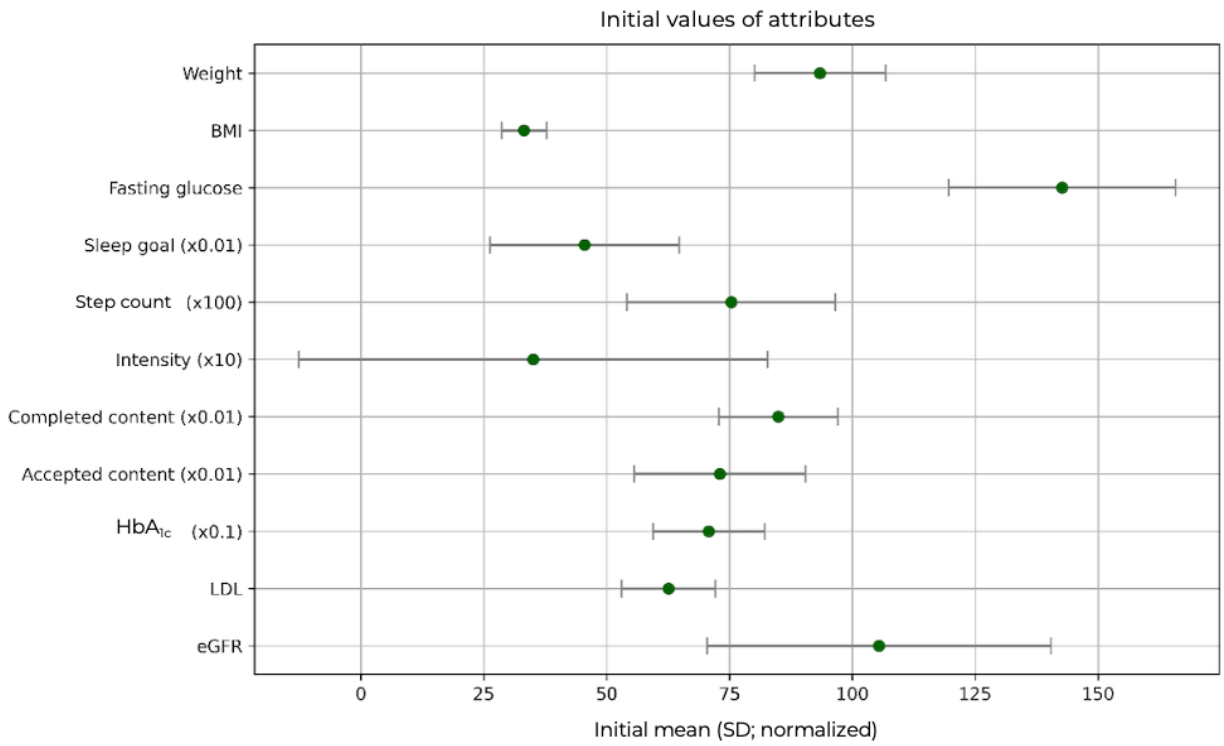

Table 4 summarizes the variations observed in the attributes during the intervention, providing their mean, SDs, and relevant inferential statistics. Given the relatively small sample size in this study, sole reliance on *P* values could be misleading. While *P* values indicated whether observed differences were statistically unlikely under a null hypothesis, they were highly sensitive to sample size and might fail to capture clinically meaningful effects in underpowered studies. To mitigate this, we also reported Cohen *d*, which quantified the standardized magnitude of change regardless of sample size, and 95% CI, which offered a range of plausible values for the mean difference.

XSL•FO
RenderX

Table 4. Postcoaching variation and effect size.

Attribute	Variation, mean (SD)	Cohen *d*	Variation (95% CI)	*P* value
Weight (kg)	–2.89 (2.75)	1.05	(–5.00 to –0.777)	.01
BMI (kg/m^2)	–1.05 (0.986)	1.06	(–1.80 to –0.288)	.01
Fasting glucose (mg/dL)	–17.3 (11.5)	1.50	(–26.1 to –8.43)	.002
Sleep goal achievement rate	–0.0106 (0.266)	0.0398	(–0.215 to 0.194)	.91
Steps, n	1890 (2560)	0.737	(–80.7 to 3860)	.06
Intensity (min)	150 (405)	0.371	(–161 to 461)	.30
Completed content ratio	–0.0998 (0.303)	0.329	(–0.333 to 0.133)	.35
Accepted content ratio	–0.125 (0.279)	0.447	(–0.339 to 0.0896)	.22
HbA_{1c}[a] (%)	–0.86 (1.26)	0.682	(–2.42 to 0.705)	.20
LDL[b] (mg/dL)	1.15 (7.81)	0.147	(–11.3 to 13.6)	.79
eGFR[c] (mL/min/1.73 m²)	–1.6 (9.91)	0.161	(–13.9 to 10.7)	.74

[a]HbA_{1c}: glycated hemoglobin.

[b]LDL: low-density lipoprotein.

[c]eGFR: estimated glomerular filtration rate.

These complementary metrics provide more nuanced insight into intervention effects. For instance, several attributes (eg, fasting glucose and BMI) demonstrated large effect sizes and narrow CIs even when the corresponding *P* values approached conventional significance thresholds. Thus, clinical interpretation was grounded not only in statistical significance but also in the magnitude and precision of observed effects. The effect size is plotted against the statistical significance in Figure 7.

Figure 7. Effect size (Cohen d) versus statistical significance (*P* value) of all attributes. The thresholds for statistical significance (red horizontal dashed line), moderate effect (yellow vertical dashed line), and high effect (blue vertical dashed line) are also shown.

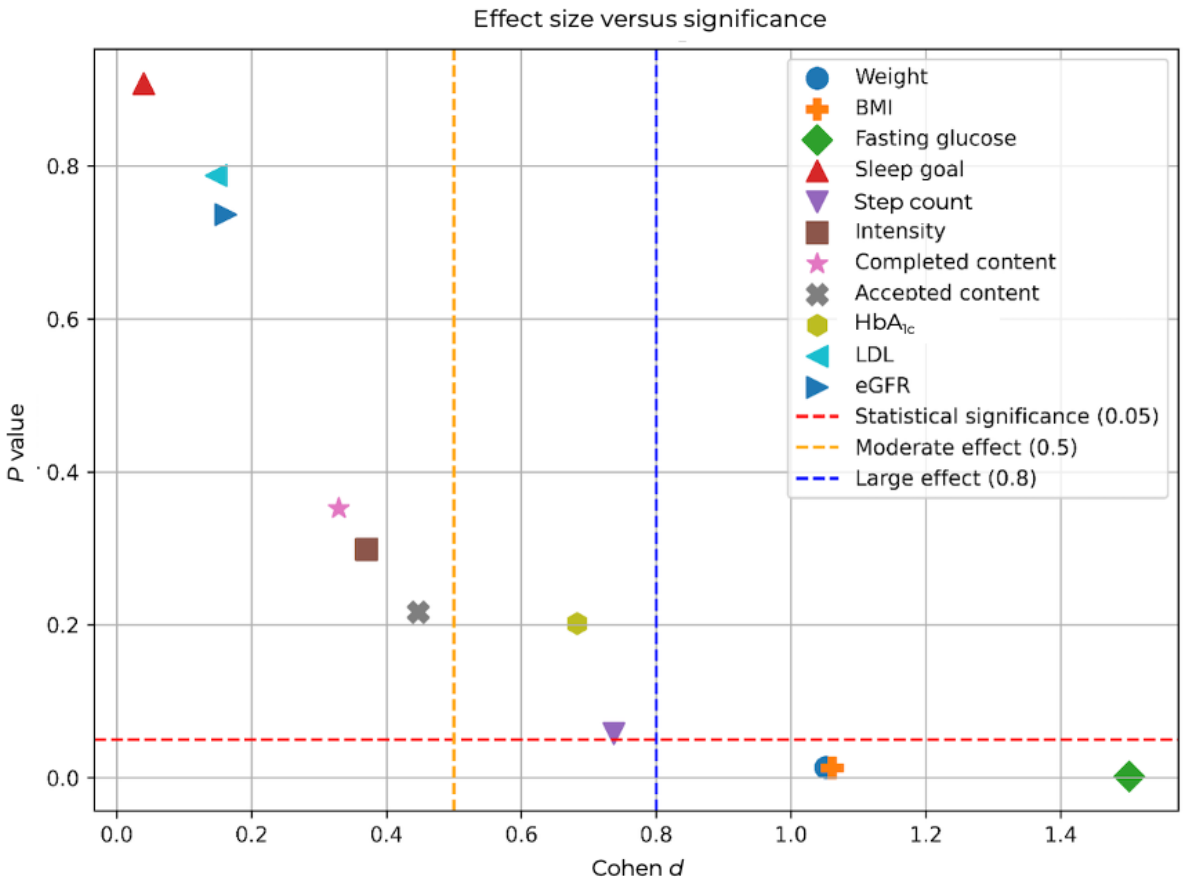

XSL•FO
RenderX

In addition to numerical trends, we categorized individual trajectories based on relative variation: patients with changes >0.5% were considered positively varying, those with changes <–0.5% as negatively varying, and the rest as fairly constant. The result is shown in Figure 8. For instance, 8 (89%) of 9 patients showed decreased fasting glucose, while sleep goal adherence was more evenly split across categories, reflecting variable behavioral compliance.

Figure 8. Individual patient trajectories based on relative variation for each attribute. The positive or negative variation can be a good result depending on the attribute. The good result bar is indicated with the letter "G." For most of the attributes, more patients exhibit a good variation. eGFR: estimated glomerular filtration rate; HbA1c: glycated hemoglobin; LDL: low-density lipoprotein.

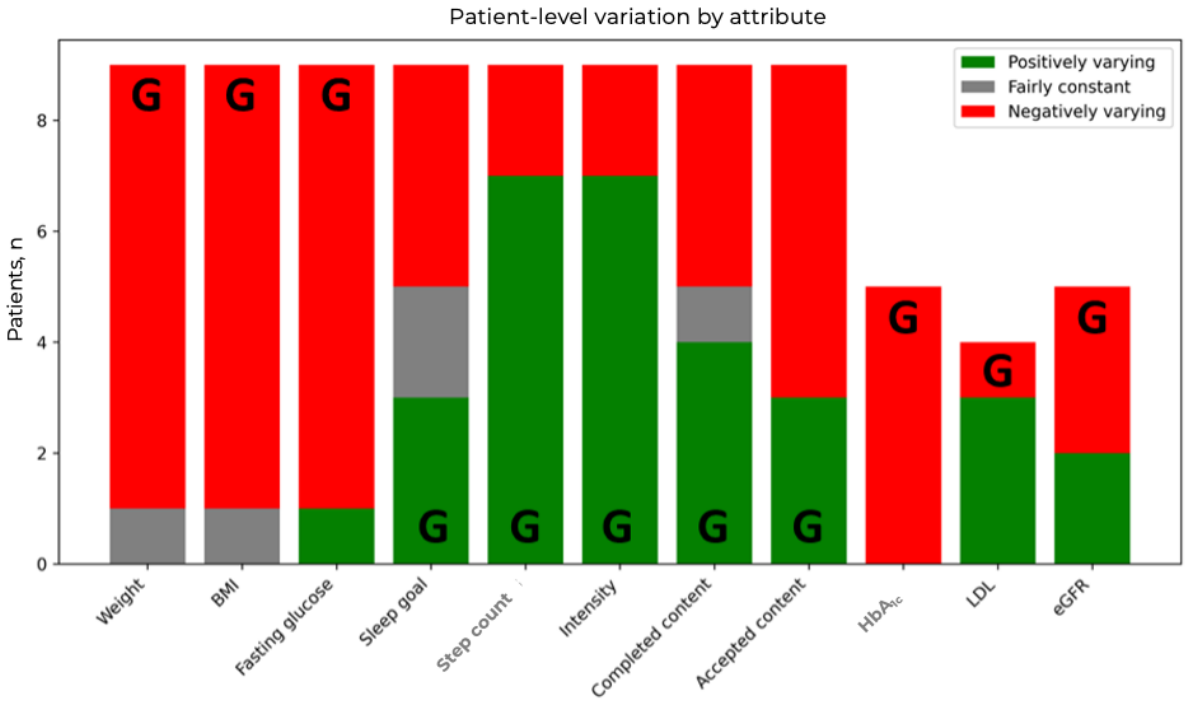

We anticipate favorable outcomes, driven at least in part by changes in patient behavior resulting from their engagement with the delivered content (refer to Table 1 for details). We consider both the volume of delivered content and its acceptance by them.

The patients could access or ignore the content items delivered to them and further provide feedback on them. Of the 1570 content items delivered thus far, 1310 (83.4%) had been accessed by the patients, and 1125 (71.7%) received positive feedback. These data are provided in Table 5 for the 3 content types and their domains.

Table 5. Content delivered (n=1510 content items), completed (n=1310), and positively received (n=1125) across different domains and types, showing approximate values.

Content type and domain	Delivered content items, n (%)	Completed content items, n (%)	Positively received content items, n (%)
Tips			
Nutrition	150 (9.55)	150 (11.45)	150 (13.33)
Physical activity	20 (1.27)	10 (0.76)	10 (0.89)
Self-care	200 (12.74)	150 (11.45)	150 (13.33)
Sleep	310 (19.75)	270 (20.61)	260 (23.11)
Dialogues			
Physical activity	320 (20.38)	250 (19.08)	180 (16.00)
Self-care	330 (21.02)	260 (19.85)	190 (16.89)
Sleep	130 (8.28)	120 (9.16)	100 (8.89)
Lessons			
Self-care	110 (7.01)	100 (7.63)	85 (7.56)

XSL•FO
RenderX

Tailored Coaching Approaches

To illustrate the tailored nature of the BCC system, we present 2 patient cases with different behavioral domain focus areas, demonstrating how content delivery was personalized to address specific needs.

In the first case study, 2U1J0 is a female patient aged 70 years, with a baseline HbA_{1c} of 6.6% and weight of 76 kg, that required a self-care–focused coaching. On the basis of the initial COM-B assessment results, she showed low capability and motivation scores for self-care behaviors.

The BCC system delivered a higher proportion of self-care–focused content to patient 2U1J0, the first case study, compared to the overall study population average:

- The patient received 45 (57%) interactive dialogues on self-care goals out of the 79 interactive dialogues overall, compared to a study-wide average of 33 (42%).
- The patient demonstrated high engagement with the content delivered, with a 92.9% (156/168) completion rate.

Patient 2U1J0 showed improvements in self-management behaviors and clinical outcomes including a +17% increase in the Diabetes Management Self-Efficacy Scale score, satisfying weight loss of 2 kg, HbA_{1c} reduction of –0.1%, and fasting glucose reduction of 18 mg/dL.

In the second case study, FGSQN is a female patient aged 66 years, with a baseline HbA_{1c} of 8.9% and a weight of 96 kg, who reported significant barriers to physical activity. Initial assessment revealed only 5318 average daily steps.

The BCC system for patient FGSQN prioritized physical activity content:

- The patient received 42 (50%) interactive dialogues on physical activity–related goals out of the 84 interactive dialogues overall compared to a study-wide average of 35 (42%).
- The patient demonstrated high engagement with the content delivered with 89% (150/168) completion rate.

Patient FGSQN showed improvements in self-management behaviors and clinical outcomes, including an increase of daily steps by 234 (4.4% increase) counts, modest weight reduction of 1 kg, impressive HbA_{1c} reduction by 2.1%, and fasting glucose reduction of 22 mg/dL.

Discussion

Principal Findings

The preliminary results from the ongoing T2DM study show potential, despite being at an early stage of the study with few patients having completed the intervention phase. Three attributes fall in the statistically significant and high-effect portion of the plane of Figure 7, with favorable variations:

1. Fasting glucose levels showed the largest relative change, dropping by an average of –17.3 mg/dL (Cohend =1.5; *P*=.002), further underscored by a narrow 95% CI –26.1 to –8.43.
2. Weight and BMI, with mean reductions of –2.89 kg and –1.05 kg/m², respectively. These changes yielded large effect sizes (Cohen d approximately 1.05) and were statistically significant (*P*=.01).

Two more attributes have favorable variations with moderate effects while not statistically significant:

1. Change in steps walked has a moderate effect (Cohen d=0.737) and was close to statistical significance (*P*=.06).
2. Change in HbA1c level has a moderate effect (Cohen d=0.682), albeit far from statistical significance (*P*=.20).

For the rest of the attributes, the variation has neither at least moderate effect nor statistical significance:

- Intensity minutes attribute has a large favorable average variation. Notably, it does not fall close to the top-left portion of the effect-significance plane of Figure 7.
- Sleep goal, low-density lipoprotein levels, and estimated glomerular filtration rate levels have nonfavorable variations. Notably, they constitute the top-left portion of the effect-significance plane of Figure 7.
- Completed and accepted content both dropped during the coaching phase. This is expected because patients can be fatigued by repeated content. We plan to address this by introducing more variations to the content per domain and intervention function.

The tailored nature of this framework is presented by analyzing 2 individual cases and their delivered content that can demonstrate different coaching experiences as they have different profiles. The first patient case is focused on improving self-care behaviors, while the second patient case is focused on physical activity–related behaviors.

We next consider the content delivered to the patients. Referring to Table 5, we note a tendency to select many more tips in the sleep domain and fewer tips in the physical activity domain compared to the other 2 domains; however, the predominance of sleep-related tips declined as the study progressed. All tips are usually accessed and well received by the patients, as are the weekly lessons. The interactive dialogues on goals show a decrease in acceptance. This is not pronounced for the sleep domain, where the least number of dialogues was delivered (90/457, 19.7%), but it is for the coping and the physical activity domains, where the delivered dialogues are almost doubled (170/457, 37.2% and 197/457, 43.1%, respectively).

The poor results in terms of physical activity variation (see seventh and eighth rows of Table 3) and the acceptance of the related goal dialogues (ie, lower than the average 85%) are noted. If this trend continues, we should consider a more even pacing of the interactive dialogues on goals, even when the goals are not met.

Limitations

The main limitation of the reported results is the limited number of patients enrolled and those who have already completed the study. This is being alleviated as the study progresses.

A second limitation stems from the inherent selection bias commonly associated with this type of study. Patients who can

XSL•FO
RenderX

be convinced to participate in and complete such a study are those predisposed to cooperate, adhere to instructions, and invest the necessary time to implement the behavioral changes as instructed. This, apart from the evident role it plays in the consistency of the study population, also affects the outcome of digital coaching in various ways. Most importantly, the participants who comply are usually the same patients who already display better disease control and have less room to express improvement (ie, already have HbA_{1c} or BMI close to target values). This fact could potentially limit the strength of the study results. For example, the population of this study consists predominately of female participants, pensioners, and former civil servants.

In addition to that, the encouraging initial results stemmed from a combination of digital coaching and other factors. These combinations include the following:

- The novelty effect arises from patients participating in a care process that differs from what they have previously experienced. The coaching content delivered is the cornerstone of this novelty. The activity trackers and mobile app conclude this novelty for the participants.
- The patients are on the more digitally literate side of the spectrum. This can be an indication of higher literacy in general. It is known that higher literacy can help in the self-management of chronic diseases. It can help in ingesting the coaching content delivered.
- The patients are affected by the feedback. This includes not only the recommendations but also the information on how they are doing, presented in the form of the plots in the different app widgets.

Overall, coaching content delivered plays a role in all 3 factors. Certainly, the exact extent of the improvement that can be attributed to the recommendations cannot be fully evaluated without the use of a control group.

A final limitation of the study is the patient-reported nature of the data collection process, which leaves room for data validity to be questioned, although the patients have no incentive to input false data. Nonetheless, the characteristics of the study population are also expected to change as the study matures, and steps could be taken to mitigate such bias.

Comparison With Prior Work

We anticipate that our ongoing study will eventually achieve better results than the study used as a reference [24], available under the German Clinical Trials Registry (DRKS00027392), which reported an average HbA_{1c} reduction of –0.9% (SD 1.1%; $P<.001$), body weight reduction of –4.3 (SD 4.5) kg, BMI reduction of –1.4 (SD 1.5) kg/m^2, and fasting glucose reduction of –10.8 (SD 23.4) mg/dL. This expectation stems from having established the technical equivalence of Healthentia to several other applications in the study by Kyriazakos et al [25]. In addition, in contrast to other studies, we are modeling the patients' behavior based on established theoretical models, such as COM-B, and using widely studied BCTs as methods for changing targeted behaviors tailored to individual needs. The studies using those mobile apps exhibited improvements in HbA_{1c} level, weight and BMI, and sleep duration. When the study matures, we expect improvement in all the reported attributes and the HbA_{1c} levels, and we can then safely compare the results of the reference study with the T2DM initiative.

Conclusions

This paper introduces an end-to-end solution that integrates science-backed, personalized digital coaching with remote patient monitoring to address the challenges of chronic disease management. By combining behavioral science with digital health technologies, the approach offers a structured intervention that could drive sustainable lifestyle changes. The implementation of our framework as a part of Healthentia's lifestyle behavioral change and wellbeing (non-medical modules) demonstrates how theoretical models can be used in practical digital health applications.

Our initial implementation of a T2DM study provides insights into the applicability and feasibility of the approach while also showing some promising but very preliminary results on the objectives, both clinical and behavioral. These include a reduction in BMI, a decrease in fasting glucose values, and a moderate reduction in HbA_{1c} levels for all patients. There is also a high engagement rate with personalized content, suggesting successful content selection.

As the T2DM program evolves and more patients complete the full program, we expect a clearer understanding of its short-term and long-term effectiveness, particularly regarding critical biomarkers, such as HbA_{1c}. These insights will allow us to further refine the methodology explored and validate the impact of the approach across various patient populations and contexts.

The potential impact of this framework extends beyond individual health outcomes, promising significant reductions in health care costs through fewer hospitalizations and complications. In addition, decentralized care delivery can help mitigate the broader societal and economic burden associated with chronic disease management. Future research will explore applications to other chronic conditions and optimization of digital BCC.

Acknowledgments

This study has been partially conducted in the context of HumAIne project, which has received funding from the European Union's Horizon Europe Research and Innovation program under grant agreement 101120218.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Authors' Contributions

KK contributed to conceptualization and project administration. DL was involved in conceptualization and software development. AP contributed to conceptualization, data curation, and visualization. NA participated in conceptualization. PS, ST, and AB were involved in the investigation. SK contributed to conceptualization and supervision. All authors participated in writing the original draft and in reviewing and editing the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Behavior change technique glossary and reference table.
[DOCX File , 30 KB - formative_v9i1e73807_app1.docx]

Multimedia Appendix 2
Capability, opportunity, motivation, and behavior model questionnaires.
[DOCX File , 28 KB - formative_v9i1e73807_app2.docx]

References

1. Kyriazakos S, Pnevmatikakis A, op den Akker H, Kostopoulou K. The role of big data and artificial intelligence in clinical research and digital therapeutics. In: Cesario A, D'Oria M, Auffray C, Scambia G, editors. Personalized Medicine Meets Artificial Intelligence: Beyond “Hype”, Towards the Metaverse. Cham, Switzerland: Springer; 2023:63-81.
2. Kyriazakos S, Pnevmatikakis A, Cesario A, Kostopoulou K, Boldrini L, Valentini V, et al. Discovering composite lifestyle biomarkers with artificial intelligence from clinical studies to enable smart eHealth and digital therapeutic services. Front Digit Health 2021 Sep 6;3:648190 [FREE Full text] [doi: 10.3389/fdgth.2021.648190] [Medline: 34713118]
3. op den Akker H, Cabrita M, Pnevmatikakis A. Digital therapeutics: virtual coaching powered by artificial intelligence on real-world data. Front Comput Sci 2021 Dec 16;3:750428. [doi: 10.3389/fcomp.2021.750428]
4. Pnevmatikakis A, Kanavos S, Matikas G, Kostopoulou K, Cesario A, Kyriazakos S. Risk assessment for personalized health insurance based on real-world data. Risks 2021 Mar 01;9(3):46. [doi: 10.3390/risks9030046]
5. Oh YJ, Zhang J, Fang ML, Fukuoka Y. A systematic review of artificial intelligence chatbots for promoting physical activity, healthy diet, and weight loss. Int J Behav Nutr Phys Act 2021 Dec 11;18(1):160 [FREE Full text] [doi: 10.1186/s12966-021-01224-6] [Medline: 34895247]
6. Mohan S, Venkatakrishnan A, Hartzler AL. Designing an AI health coach and studying its utility in promoting regular aerobic exercise. ACM Trans Interact Intell Syst 2020 May 30;10(2):1-30. [doi: 10.1145/3366501]
7. Direito A, Pfaeffli Dale L, Shields E, Dobson R, Whittaker R, Maddison R. Do physical activity and dietary smartphone applications incorporate evidence-based behaviour change techniques? BMC Public Health 2014 Jun 25;14:646 [FREE Full text] [doi: 10.1186/1471-2458-14-646] [Medline: 24965805]
8. Michie S, Johnston M, Francis J, Hardeman W, Eccles M. From theory to intervention: mapping theoretically derived behavioural determinants to behaviour change techniques. Appl Psychol 2008 Jul 08;57(4):660-680. [doi: 10.1111/j.1464-0597.2008.00341.x]
9. Michie S, van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. Implement Sci 2011 Apr 23;6:42 [FREE Full text] [doi: 10.1186/1748-5908-6-42] [Medline: 21513547]
10. Murphy K, Berk J, Muhwava-Mbabala L, Booley S, Harbron J, Ware L, et al. Using the COM-B model and Behaviour Change Wheel to develop a theory and evidence-based intervention for women with gestational diabetes (IINDIAGO). BMC Public Health 2023 May 15;23(1):894 [FREE Full text] [doi: 10.1186/s12889-023-15586-y] [Medline: 37189143]
11. Barker F, Atkins L, de Lusignan S. Applying the COM-B behaviour model and behaviour change wheel to develop an intervention to improve hearing-aid use in adult auditory rehabilitation. Int J Audiol 2016 Jul;55 Suppl 3:S90-S98 [FREE Full text] [doi: 10.3109/14992027.2015.1120894] [Medline: 27420547]
12. Schenk PM, Wright AJ, West R, Hastings J, Lorencatto F, Moore C, et al. An ontology of mechanisms of action in behaviour change interventions. Wellcome Open Res 2023;8:337 [FREE Full text] [doi: 10.12688/wellcomeopenres.19489.1] [Medline: 38481854]
13. Michie S, Richardson M, Johnston M, Abraham C, Francis J, Hardeman W, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013 Aug;46(1):81-95 [FREE Full text] [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
14. Johnston M, Carey RN, Connell Bohlen LE, Johnston DW, Rothman AJ, de Bruin M, et al. Development of an online tool for linking behavior change techniques and mechanisms of action based on triangulation of findings from literature synthesis and expert consensus. Transl Behav Med 2021 May 25;11(5):1049-1065 [FREE Full text] [doi: 10.1093/tbm/ibaa050] [Medline: 32749460]

XSL•FO
RenderX

15. Gardner B, Smith L, Lorencatto F, Hamer M, Biddle SJ. How to reduce sitting time? A review of behaviour change strategies used in sedentary behaviour reduction interventions among adults. Health Psychol Rev 2016;10(1):89-112 [FREE Full text] [doi: 10.1080/17437199.2015.1082146] [Medline: 26315814]
16. Harkin B, Webb TL, Chang BP, Prestwich A, Conner M, Kellar I, et al. Does monitoring goal progress promote goal attainment? A meta-analysis of the experimental evidence. Psychol Bull 2016 Feb;142(2):198-229 [FREE Full text] [doi: 10.1037/bul0000025] [Medline: 26479070]
17. Michie S, Abraham C, Whittington C, McAteer J, Gupta S. Effective techniques in healthy eating and physical activity interventions: a meta-regression. Health Psychol 2009 Nov;28(6):690-701. [doi: 10.1037/a0016136] [Medline: 19916637]
18. Bratman M. Intention, Plans, and Practical Reason. Cambridge, MA: Harvard University Press; 1987.
19. Taj F, Klein MC, van HA. Towards a generic framework for a health behaviour change support agentth International Conference on AgentsArtificial Intelligence, ICAART 2020. In: Proceedings of the 12th International Conference on Agents and Artificial Intelligence. 2024 Presented at: ICAART '20; February 22-24, 2020; Valletta, MT p. 311-318 URL: https://www.scitepress.org/Papers/2020/91735/91735.pdf [doi: 10.5220/0009173503110318]
20. Howlett N, Schulz J, Trivedi D, Troop N, Chater A. A prospective study exploring the construct and predictive validity of the COM-B model for physical activity. J Health Psychol 2019 Sep;24(10):1378-1391. [doi: 10.1177/1359105317739098] [Medline: 29172808]
21. Monitoring your blood sugar. Centers for Disease Control and Prevention (CDC). URL: https://www.cdc.gov/diabetes/diabetes-testing/monitoring-blood-sugar.html [accessed 2025-04-29]
22. Samson SL, Vellanki P, Blonde L, Christofides EA, Galindo RJ, Hirsch IB, et al. American Association of Clinical Endocrinology Consensus statement: comprehensive type 2 diabetes management algorithm - 2023 update. Endocr Pract 2023 May;29(5):305-340. [doi: 10.1016/j.eprac.2023.02.001] [Medline: 37150579]
23. Colberg SR, Sigal RJ, Fernhall B, Regensteiner JG, Blissmer BJ, Rubin RR, American College of Sports Medicine, American Diabetes Association. Exercise and type 2 diabetes: the American College of Sports Medicine and the American Diabetes Association: joint position statement. Diabetes Care 2010 Dec;33(12):e147-e167 [FREE Full text] [doi: 10.2337/dc10-9990] [Medline: 21115758]
24. Bretschneider MP, Klásek J, Karbanová M, Timpel P, Herrmann S, Schwarz PE. Impact of a digital lifestyle intervention on diabetes self-management: a pilot study. Nutrients 2022 Apr 26;14(9):1810 [FREE Full text] [doi: 10.3390/nu14091810] [Medline: 35565782]
25. Kyriazakos S, Pnevmatikakis A, Kostopoulou K, Ferrière L, Thibaut K, Giacobini E, et al. Benchmarking the clinical outcomes of Healthentia SaMD in chronic disease management: a systematic literature review comparison. Front Public Health 2024 Dec 24;12:1488687 [FREE Full text] [doi: 10.3389/fpubh.2024.1488687] [Medline: 39776481]

Abbreviations

BCC: behavior change coaching
BCT: behavior change technique
BCW: Behaviour Change Wheel
BDI: belief-desire-intention
COM-B: capability, opportunity, motivation, and behavior
HbA1c: glycated hemoglobin
T2DM: type 2 diabetes mellitus

Edited by J Sarvestan; submitted 12.03.25; peer-reviewed by P Dunn, M Adnan; comments to author 21.04.25; revised version received 12.05.25; accepted 22.05.25; published 09.07.25.

Please cite as:
Kostopoulou K, Lekka D, Pnevmatikakis A, Angelova N, Stafylas P, Tamouridis S, Bargiota A, Kyriazakos S
Multidomain Behavioral Change Digital Coaching for Chronic Disease Management in Patients With Type 2 Diabetes: Framework Development and Preliminary Evaluation
JMIR Form Res 2025;9:e73807
URL: https://formative.jmir.org/2025/1/e73807
doi:10.2196/73807
PMID:

©Konstantina Kostopoulou, Danae Lekka, Aristodemos Pnevmatikakis, Nelina Angelova, Panagiotis Stafylas, Stefanos Tamouridis, Alexandra Bargiota, Sofoklis Kyriazakos. Originally published in JMIR Formative Research (https://formative.jmir.org), 09.07.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License

XSL•FO
RenderX

(https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Developing and Integrating Digital Sources in an Accessible and Sustainable Online Platform for Adolescents and Young Adult Cancer Survivors: Collaborative Design Approach

Carla Vlooswijk[1], MSc; Sophia HE Sleeman[2], BSc; Jonas Pluis[2], BSc; Daphne Bakker[2], BSc; Lisanne de Groot[2], BSc; Eveliene Manten[2], PhD; Peter Heine[3], MSc; Olga Husson[4,5,6], PhD; Mies C van Eenbergen[1†*], PhD; Pieter Vandekerckhove[7*], PhD

[1]Department of Research and Development, Netherlands Comprehensive Cancer Organisation, Utrecht, The Netherlands
[2]Dutch AYA 'Young & Cancer' Care Network, Utrecht, The Netherlands
[3]Kanker.nl Foundation, Amsterdam, The Netherlands
[4]Department of Medical Oncology, The Netherlands Cancer Institute, Amsterdam, The Netherlands
[5]Department of Surgical Oncology, Erasmus University Medical Center, Rotterdam, The Netherlands
[6]Department of Public Health, Erasmus MC Cancer Institute, Erasmus University Medical Center, Rotterdam, The Netherlands
[7]Delft Centre for Entrepreneurship, Delft University of Technology, Delft, The Netherlands
[†]deceased
[*]these authors contributed equally

Corresponding Author:
Carla Vlooswijk, MSc
Department of Research and Development
Netherlands Comprehensive Cancer Organisation
Godebaldkwartier 419
Utrecht, 3511 DT
The Netherlands
Phone: 31 088 234 60 00
Email: c.vlooswijk@iknl.nl

Abstract

Background: Digital support for adolescent and young adult cancer survivors is fragmented and results in lacking a reliable overview of support services available to them. Collaborative design promises to integrate perspectives of diverse stakeholders and could help to develop an online platform, which increases access and has a long-term perspective. However, it has not yet been explored how collaborative design can be used more strategically to develop an online platform with these aims.

Objective: This study aimed to explore how a collaborative design approach could be applied to develop an online platform for adolescent and young adult cancer survivors, focusing on three key objectives: integrating existing resources, improving accessibility, and ensuring long-term financial sustainability.

Methods: In this action research study, we reflect on a collaborative design process to develop an online platform for adolescent and young adult cancer survivors. A quantitative questionnaire was sent out to adolescent and young adult cancer survivors and health care professionals. Stakeholders were actively engaged in stakeholder consensus meetings, and project management was carried out through monthly design meetings to facilitate coordination and decision-making of the development of the platform. Afterwards, a focus group was conducted among the project group to evaluate the collaborative design approach, analyzed using an inductive thematic approach.

Results: Through the collaborative design approach, several Dutch organizations collaborated to develop, enhance, and combine online services for adolescent and young adult cancer survivors and their relatives. A dedicated online “young and cancer” platform for adolescent and young adult cancer survivors was developed, which integrates different types of information tools and supportive interactive elements from different sources. The integration of different resources into one platform improves the access and user experience of adolescent and young adult cancer survivors when it comes to online support. Through the reflection about the collaborative process, three themes were identified: (1) value of stakeholder participation; (2) conditions for working with

XSL•FO
RenderX

adolescents and young adults with lived experience; and (3) collaboration between adolescents and young adults with lived experience and professionals from different backgrounds and organizations.

Conclusions: A collaborative design approach can be used to efficiently develop an online platform for adolescent and young adult cancer survivors. The collaboration among professionals, including online developers, researchers, and adolescents and young adults with lived experience facilitated a direct translation of insights into the platform, while the support of the national cancer platform provides long-term sustainability. This study highlights the importance of strategic stakeholder selection and the intense involvement of stakeholders through a collaborative design approach.

(JMIR Form Res 2025;9:e60897) doi:10.2196/60897

KEYWORDS

adolescents and young adults with cancer; online platform; internet; age-specific; collaborative design; peer support; patient involvement; design; teens; youth; development; quantitative questionnaire; patient experiences

Introduction

Increasingly, the internet and digital technologies hold a great promise to provide additional care and support for patients with and survivors of cancer [1]. A growing number of digital interventions have been developed and designed for patients with cancer living with and beyond a cancer diagnosis [2-7]. These emerging technologies aim to support, educate, and empower cancer patients and survivors to manage their own health through the use of information platforms, online support groups, and self-monitoring applications [8]. Despite the careful development of these digital solutions, significant challenges remain regarding their uptake, accessibility, and reach [9], which hinders their ability to make a long-lasting impact [5]. These solutions are often developed as stand-alone solutions without a long-term strategy or appropriate integration into the existing architecture and without continuous funding consideration.

Support for adolescents and young adults with cancer is particularly underdeveloped. adolescent and young adult cancer survivors are defined as individuals diagnosed with cancer between the ages of 18 and 39 years [10,11]. A cancer diagnosis at such a young age has a huge impact and causes physical, psychological, and social challenges that differ significantly from those of older adults and children [12]. Young cancer survivors struggle with various- and often age-related questions in several areas, such as fertility, family, children, employment, and establishing (romantic) relationships [13]. Zebrack et al [14] showed that adolescent and young adult cancer survivors, especially those treated in adult clinical settings, reported unmet needs in regard to age-appropriate websites. Given that 96% of the Dutch adolescents and young adults in the age of 25-35 years use the internet in their daily life [15], this young group appears to be a suitable population for the use of these online technologies. Therefore, there is a potential to provide age-specific information to adolescent and young adult cancer survivors in a digital manner.

In the Netherlands, disease-related, age-specific information and online services are scattered and could be found on several websites of different organizations. Firstly, the Dutch adolescent and young adult “Young and Cancer” Care Network initiated a public web page with age-specific topics and a secured online adolescent and young adult4 community in 2012. In this adolescent and young adult community, adolescent and young adult cancer survivors could express feelings and exchange information, which has been found helpful in coping with cancer [16-18]. This informative landing page and community have been the starting point for further development and implementation of digital adolescent and young adult services. To access live peer-to-peer meetings or events, adolescent and young adult cancer survivors have to surf to other websites, such as the patient organization for young patients with cancers (Stichting Jongeren en Kanker [SJK]) [19] or “walk-in houses” for everyone who gets impacted by cancer [20]. Another example is the applications developed to empower and support adolescent and young adult cancer survivors, such as the adolescent and young adult Match app, aiming to support adolescent and young adult cancer survivors when it comes to keeping in contact with their relatives after their cancer diagnosis [21]. The fragmentation of all these services results in adolescent and young adult cancer survivors lacking a reliable overview of support services available to them. Therefore, a solution would need to be financially sustainable, easily accessible, and integrate the existing sources.

An age-specific online platform for adolescent and young adult cancer survivors could meet those needs, reducing the complexity of navigating multiple sources and ensuring that content is relevant to their distinct needs. A dedicated platform for adolescent and young adult cancer survivors could bring together different types of information, tools, and interactive elements, elevating the experience and efficiency of information access and support. Therefore, the Dutch adolescent and young adult “Young & Cancer” Care Network and SJK set out to improve and enhance online services for adolescent and young adult cancer survivors in the Netherlands. The COMPRehensive health outcome and intervention research among patients with adolescent and young adult cancer (COMPRAYA) study provided funding and a stepping stone for developing an online “young and cancer” platform together with and for adolescent and young adult cancer survivors [22].

From a methodological point of view, it can be challenging to develop an online platform that is financially sustainable, accessible, and integrated in collaboration with different stakeholders. Collaborative design could be promising here as it is widely used and has been shown to improve commitment, support the development and implementation of digital health applications and provide more tailored solutions to users’ needs [23-26]. Collaborative design can involve stakeholders to various degrees in a collective creative process whereby

XSL•FO
RenderX

stakeholders are considered partners in the design process [23]. For instance, this approach involves creative workshops whereby diverse stakeholders collaborate in groups [27], and it can also involve interviews and observations [28]. Yet, it has been noted that overlooking long-term perspectives could potentially reduce the success of the implementation process [23]. Considering certain potential problems upfront in the innovation process can help to avoid costly mistakes in the future [29]. In this case, collaborative design could be strategically used by considering how to develop an online platform with a triple aim: (1) to integrate existing sources, (2) to make them more accessible, and (3) to have a financially sustainable perspective. To this end, we aimed to explore how a collaborative design approach could be applied to develop an online support platform for adolescent and young adult cancer survivors, focusing on these 3 key objectives.

Methods

Overview

This study follows an action research methodology [30,31], whereby we want to learn about the use of collaborative design methods through the development of the "young and cancer" platform for adolescent and young adult cancer survivors. Action research is ideally suited for studying innovation processes and working towards sustainable transformation as one generates research knowledge through practical experience [30]. We used a combination of quantitative and qualitative research methods.

Hosting the "Young and Cancer Platform"

The aim to develop the Dutch online "young and cancer" platform stems from a broader research project called COMPRAYA [22]. The aim of the platform was to provide age-specific information, support, and potential digital interventions for adolescent and young adult cancer survivors. One of the conditions was to combine existing information and services and create new age-tailored services in a sustainable online place, preferably integrated into an existing organization or platform, such as Kanker.nl [7]. Within the collaborative design approach, the stakeholders were involved in the strategic choice for future-proof embedding in an already existing platform. In this way, we tried to avoid later problems with implementation, embedding, and financial sustainability. The Kanker.nl platform was chosen as the most suitable place to integrate the adolescent and young adult platform into. Kanker.nl is a centrally initiated Dutch online platform (on average 1.077.313 visits per mo, all ages) and structurally funded by the Dutch Cancer Society (Koningin Wilhelmina Fonds [KWF]). It is dedicated to providing information, support, and resources for patients with cancer and cancer survivors in general [7].

Collaborative Design Approach

A collaborative design approach was a fundamental principle guiding the development of this online platform: nothing about adolescents and young adults, without adolescents and young adults [23]. Throughout the entire process, adolescent and young adult with lived experiences were at the forefront, from concept designing to prioritizing themes and decision-making.

The involvement matrix was used to show the roles of various stakeholders (adolescent and young adult cancer survivors, relatives, and health care professionals [HCPs]) throughout the developmental stages of diverse elements of the platform [32]. The involvement matrix is developed to help involve patient experts in research projects. It can help to give the collaboration of patient experts a central place in research, improve the collaboration, and report on collaboration in a systematic and clear manner. The roles are listener, co-thinker, advisor (gives [un]solicited advice), partner (works as an equal partner), and decision maker (takes initiative, [final] decision). A listener received information and updates, such as attending project meetings and being shown the information and tools that have been developed and will be available on the platform. A co-thinker is asked to give their opinion on already developed materials, for example, by testing online questionnaires and providing feedback on draft versions of tools and information for adolescent and young adult cancer survivors. An advisor is given (un)solicited advice in early stages, such as providing advice on themes and topics that need to be included and is given advice about prioritization of the implementation of the tools. A partner works as an equal partner, for example, writing materials or being a podcast host. A decision maker takes initiative and is in charge of final decisions.

The collaborative design activities are shown in Figure 1. This process consisted of a needs research using quantitative questionnaires stakeholders were actively engaged in 3 stakeholder consensus meetings, a review group provided feedback ad hoc on developed content, and project management was carried out through collaborative design meetings to facilitate coordination and decision-making. These activities are further detailed below.

XSL•FO
RenderX

Figure 1. A graphical overview of the timelines of the collaborative design approach for the online "young and cancer" platform, showing the involvement of stakeholders from December 2019 to the platform's launch in June 2023.

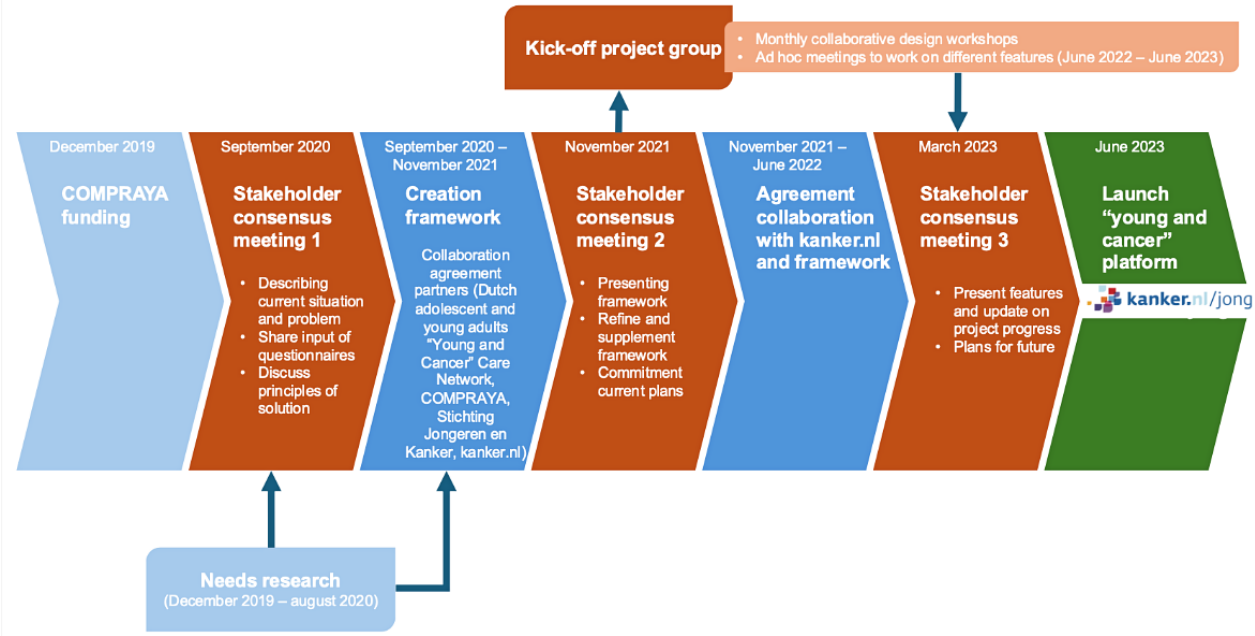

Needs Research

Results of 2 conducted questionnaires were analyzed to understand the online wishes and needs among adolescent and young adult cancer survivors and the perspective of HCPs on this matter [33]. adolescent and young adult cancer survivors were invited via an online community, newsletters of the Nederlandse Federatie van Kankerpatiëntenorganisaties (NFK; Dutch Federation of Cancer Patient Organizations) and social media platforms and participants of the health-related quality of life and late effects among SURVivors of cancer in Adolescence and Young Adulthood (SURVAYA) study [34]. HCPs were recruited via a mailing from the Dutch adolescent and young adult "Young & Cancer" Care Network. These HCPs either already provided some form of integral adolescent and young adult care or were willing to do so [35].

Collaborative Design Meetings

A project group coordinated the development of the "young and cancer" platform by taking part in monthly collaborative design meetings, once per 1 or 2 months, mostly online via Microsoft Teams. The project group consisted of adolescents and young adults with lived experience, a project leader, a project coordinator, a medical editor, and a business development manager. Within the monthly collaborative design meetings, the "young and cancer" platform was given shape through open discussion, exploring options, and taking into account principles of collaborative design, for example, equality [23]. In these meetings, regular evaluations of the approach, workload, and time investment were also conducted, with particular attention to the adolescents and young adults with lived experience who participated outside of their professional roles. This process was essential for ensuring continued alignment within the project group and to make sure that everyone was satisfied with the chosen approach and collaboration.

Stakeholder Participation

Overall, 3 stakeholder consensus meetings were organized with people with 8 different backgrounds (medical specialists, medical social workers, nurse practitioners, board members of patient organizations, adolescent and young adult cancer survivors, relatives of adolescent and young adult cancer survivors, and online specialists). These 3 online meetings, each lasting 90 minutes, were conducted via Microsoft Teams and organized at 3 points during the different phases of the platform development (Figure 1). The aim of these meetings was to inform relevant stakeholders and act as an advisory board and provide ideas and suggestions. In addition, the meetings were used to build consensus and create support for the "young and cancer" platform among relevant stakeholders who play an important role in referring adolescents and young adults to this online platform.

Review Group

An open call was extended to all adolescents and young adults and their relatives who were interested in providing feedback on the project group's output and written materials from the perspective of their lived experiences.

In total, 20 adolescents and young adults and relatives were recruited through various recruitment places (eg, consultation room and social media). Reviewers from HCPs' perspective were recruited via the stakeholder consensus meetings and asked to critically review the written materials for inaccuracies and supplement them with practical experiences. Researchers within the adolescent and young adult cancer field were asked to provide their input and feedback based on their research perspectives.

Evaluation of the Collaborative Design Process

A focus group with the project group was conducted online via Microsoft Teams on April 12, 2023, for approximately 90 minutes. The focus group was done in accordance with the

XSL•FO
RenderX

COREQ (Consolidated Criteria for Reporting Qualitative Research) guidelines for qualitative studies [36] (Multimedia Appendix 1). All project members of the project group were invited to participate by the project leader via email. Subsequently, the participants were contacted by the researcher and informed of the purpose of the focus group. All participants provided written informed consent to participate in the focus group. Before participation, participants self-reported sociodemographic and, if applicable, clinical information. The facilitator (CV), an independent trained female researcher, guided the focus group following a semistructured topic guide (Textbox 1).

Textbox 1. The topic guide of the focus group for evaluating the 'collaborative design approach within the project group dedicated to developing the online "young and cancer" platform.

Topic guide:

- Introduction
  - How is the division of roles and tasks within the project group?
  - What is your role in this project?
  - Why are you participating in this project?
  - When would you not have participated?
  - What is decisive? Time, money, people, and appreciation?
  - What is going well and what could be improved?
- Cooperation project group
  - Do you feel that you have influence on the project?
  - Is this (in)sufficient?
  - What is the added value of working in this project group?
  - Would you like to do this more often?
  - What helps you to do this?
- Online services for adolescents and young adults
  - When did the project succeed for you?
  - Which elements and issues are important to you in the final product (online platform for adolescents and young adults)?
  - What are the limiting or impeding factors?
  - Do you have tips or recommendations for others in similar projects?

Data Analysis

Quantitative analyses were performed using SAS (SAS Institute) software. Descriptive statistics included frequencies, means, SDs, and percentages. The focus group was recorded in Microsoft Teams, transcribed verbatim, and analyzed via an inductive thematic approach described by Braun and Clarke [37] using NVivo (Lumivero). Initially, the focus group was reviewed by the researcher, and all the information describing the evaluation of the collaborative process was highlighted (open coding). Subsequently, these codes were discussed and agreed upon with SS.

Ethics Approval

This project was approved by the Research Ethics Review Committee of the Tilburg School of Humanities and Digital Sciences (internal code: REDC 2019.104). Informed consent was obtained from all adolescent and young adult cancer survivors who participated in the questionnaire study and participants in the focus group. The adolescents and young adults with lived experience in our project were compensated with gift vouchers with approximately €50 (US $58.58) value per meeting.

Results

Needs Research

More than half (373, 63%) of the adolescent and young adult cancer survivors indicated that they wanted to find age-specific information and digital tools from various organizations converged into one website [33]. adolescent and young adult cancer survivors (43%) indicated that the priority should be on age-specific information on a website specifically designed for adolescent and young adult cancer survivors. Online peer contact was used by 18.4% of the adolescent and young adult cancer survivors. They mainly posted or responded on different platforms, such as Amazon Breast Cancer Support Group, patient organizations, their own websites, Facebook (Meta), Kanker.nl, Instagram (Meta), WhatsApp (Meta), and the adolescent and young adult community. More adolescent and young adult cancer survivors (38.3%) prefer to passively participate by reading online forums, blogs, discussion

XSL•FO
RenderX

platforms, or patient communities. However, 6.4% could not find a suitable website and 12.6% prefer a website where mostly younger people with cancer gather. Only a small number (6.9%) of adolescent and young adult cancer survivors posed online questions to HCPs via, for instance, a secured online environment of the hospital or general practitioner, email, or WhatsApp. While 41.4% of adolescent and young adult cancer survivors indicated that they would like to ask questions of their HCP online, but they have never done this yet. More extended analyses based on these questionnaires to assess the quantity of cancer-related internet use, content searched for, perceived impact of cancer-related internet use, and eHealth needs among adolescent and young adult cancer survivors are published elsewhere [33,38]. These results were also compared with the perceptions of HCPs [38].

Stakeholder Participation

Stakeholder Consensus Meeting 1

The project group presented the basic principles and scenarios for the suggested "young and cancer" platform and sought input from all stakeholders. Different scenarios were presented, such as the improvement or enrichment of websites of SJK and the Dutch adolescent and young adult "Young & Cancer" Care Network, where adolescent and young adult cancer survivors are familiar with finding this information (Multimedia Appendix 2), Other scenarios were using the existing platform of Kanker.nl. One of the scenarios even described building a completely new platform. The presented scenarios varying in terms of the amount of investment and complexity are required. Scenario 1 was "do nothing" and scenario 7 was building an entirely new, stand-alone platform, which required significant investment, short-term as well as long-term, and was more complex. A new channel integrated within an existing platform, Kanker.nl (a combination of scenarios 5 and 6), seemed to be the most reliable and optimal solution. This allows for the use of the infrastructure, maintenance, management, and overall development of the platform, while simultaneously creating a distinct look and feel for adolescent and young adult cancer survivors and presenting age-specific content. It could integrate existing sources, making the "young and cancer" platform more accessible and having a financially sustainable perspective. The existing features on Kanker.nl can be used and, where necessary, tailored for adolescent and young adult cancer survivors. A potential limitation of integrating the "young and cancer" platform into Kanker.nl is the existing principles and possible restrictions of the current platform, such as the available modals and infrastructure of its content management system.

Stakeholder Consensus Meeting 2

In total, 4 personas were used to provide a clear and concise way of ensuring everyone has a shared understanding of how the platform and features should function (Multimedia Appendix 3). These personas represent adolescent and young adult cancer survivors and their relatives, including different sexes (male and female), tumor types (Hodgkin lymphoma, lung cancer, leukemia, and breast cancer), different ages (eg, 21, 23, 30, and 36 y), different times since diagnosis (currently under treatment, diagnosed 2 y ago), and varying needs on the platform (such as reading blogs, connecting with peers, accessing information files, using health applications).

The stakeholders highlighted several key features they wanted from the new platform, including Google search results linking to the "young and cancer" platform, dedicated promotional materials, an online calendar of available peer activities, the ability to join secure online support groups, the option to indicate openness to contact with other peers, and the ability to search personal profiles and reach out to suitable peers.

Stakeholder Consensus Meeting 3

The project group provided an update on the project's progress and actively sought input and feedback from all stakeholders to gain a deeper understanding of their perspectives. They focused on refining, sharpening, and clarifying the plans to strengthen commitment to the current direction. To gain the most traffic to the new platform, the launch of the "young and cancer" platform was presented on June 15 June, 2023, during the SPACE4AYA congress for adolescent and young adult cancer survivors and their HCPs. Agreements on ongoing collaboration among stakeholders, COMPRAYA, SJK, Dutch "Young and Cancer" Care Network, and Kanker.nl ensure integration and further development of the "young and cancer" services in the future within the sustainable platform Kanker.nl.

Development of the Online "Young and Cancer" Platform

Various features were developed within the new "young and cancer" platform. An overview of these features can be found in Table 1. The primary target group of the platform is individuals with cancer who have received a cancer diagnosis between the ages of 18 and 39 years (adolescents and young adults). The secondary target group are relatives of adolescents and young adults and Young Adult Childhood Cancer Survivors, those diagnosed with cancer before the age of 18 years, who are now between 16 and 39 years of age and can struggle with similar age-specific problems as adolescent and young adult cancer survivors. The framework, written by the project group and validated by the stakeholders, formulated the following concept: create a future-proof environment with reliable up-to-date information, the possibility to contact peers and access to eHealth applications. The following conditions were formulated: reuse of existing content with the possibility of adding age-specific information, reuse of existing features with the possibility to tailoring them to adolescents and young adults, ensuring a future-proof platform, continuation of maintenance when COMPRAYA funding ends, using results of scientific research from COMPRAYA and making the developed services for adolescents and young adults also available to other target groups if appropriate, for example, people with rare cancer.

XSL•FO
RenderX

Table 1. The stakeholder involvement in the development process for the various features of the online "young and cancer" platform is illustrated using the involvement matrix.

Features online platform	Involvement matrix [32]				
Roles	Listener	Co-thinker	Advisor	Partner	Decision-maker
Design landing page	—[a]	—	—	1[b]	1
Profile modifications for adolescent and young adult cancer survivors		1	—	—	—
Age-specific information for adolescent and young adult cancer survivors and relatives					
Written information dossiers	—	2[c], 3[d],4[e]	—	1	—
Event calendar	1	—	—	1	—
Different types of content, for example, photography, visuals, ambassadors' stories, podcasts, and videos	—	4		1	1
Opportunities for social interaction					
Blogs	1	—	1	—	—
Online support groups for adolescent and young adult cancer survivors and their relatives	—	4	1	—	—
Peer finder	—	1,2	2	1	—
Health care professional on demand	—	4	1	3	—
Guide for reliable information and support	1	—	1	—	—
Library for resources and a specific app store	—	—	1	—	—

[a]Not Applicable.

[b]1: Project group.

[c]2: adolescent and young adult with lived experience and relative's reviewers.

[d]3: health care providers and research reviewers.

[e]4: Stakeholder consensus meetings.

Landing Page

The existing infrastructure of Kanker.nl was used to create a "landing page" for adolescents and young adults with an age-specific look-and-feel (including photography of young people) and offers the flexibility to vary in terms of design, navigation items, and content [39]. The users of the landing page can navigate to different age-specific features presented in their own unique tone of voice. The aim was to create an environment where adolescents and young adults feel seen and acknowledged, and to create an environment where adolescent and young adult cancer survivors could identify themselves with.

Age-Specific Information for Adolescents and Young Adult Cancer Survivors

Age-specific content has been created, resulting in information dossiers for adolescent and young adult cancer survivors: disease and complaints, treatment and care, finance and insurance, fertility and pregnancy (Multimedia Appendix 4), taking care of yourself, study and cancer, work and cancer, intimacy and sexuality, appearance, dealing with the people around you, emotions, and existential questions of life and palliative care. In addition, a separate information dossier for relatives, parents, partner, family, and friends was created. In addition to just text, both existing and newly developed interviews and podcasts related to adolescents' and young adults' cancer experiences were placed on the "young and cancer" platform. Furthermore, an events calendar with (online) activities for adolescent and young adult cancer survivors was placed. This calendar is linked to the calendar of SJK, which includes social gatherings, informational meetings, outdoor adventures, and other events that focus on connecting with peers.

Opportunities for Social Interaction

adolescent and young adult cancer survivors are able to read personal experiences of other (adolescent and young adult) cancer survivors and relatives in the blog section of Kanker.nl. They also have the opportunity to share their own experiences. A few blogs of adolescent and young adult cancer survivors and relatives have been highlighted, and an age-related filter will be put in place. The online support group, named "the young and cancer online support group," exists on Kanker.nl. Thereby, an online support group especially for relatives of adolescent and young adult cancer survivors has been created, called "young and cancer-relatives." Moderators (adolescents and young adults with lived experience and relatives of adolescent and young adult cancer survivors) have been appointed. The online support groups are open to registered members, allowing anyone with an account to read and participate. Visitors without an account can only view the conversation title unless the poster explicitly specifies that the content is fully public. To offer cancer survivors a safe and noncommittal way to get in touch with peers online, the "peer finder" was created. This is a tool in which patients with cancer can search for peers based on information from their personal profile, such as age, type of

XSL•FO
RenderX

cancer, stage of the disease, treatment, and (late) effects (Figure 2). Cancer survivors and relatives can get in touch with people in a similar situation and directly send private messages in a secure environment.

Figure 2. A screenshot of the "Peer finder": a digital tool available on the online "young and cancer" platform to help cancer survivors find and connect with peers with cancer.

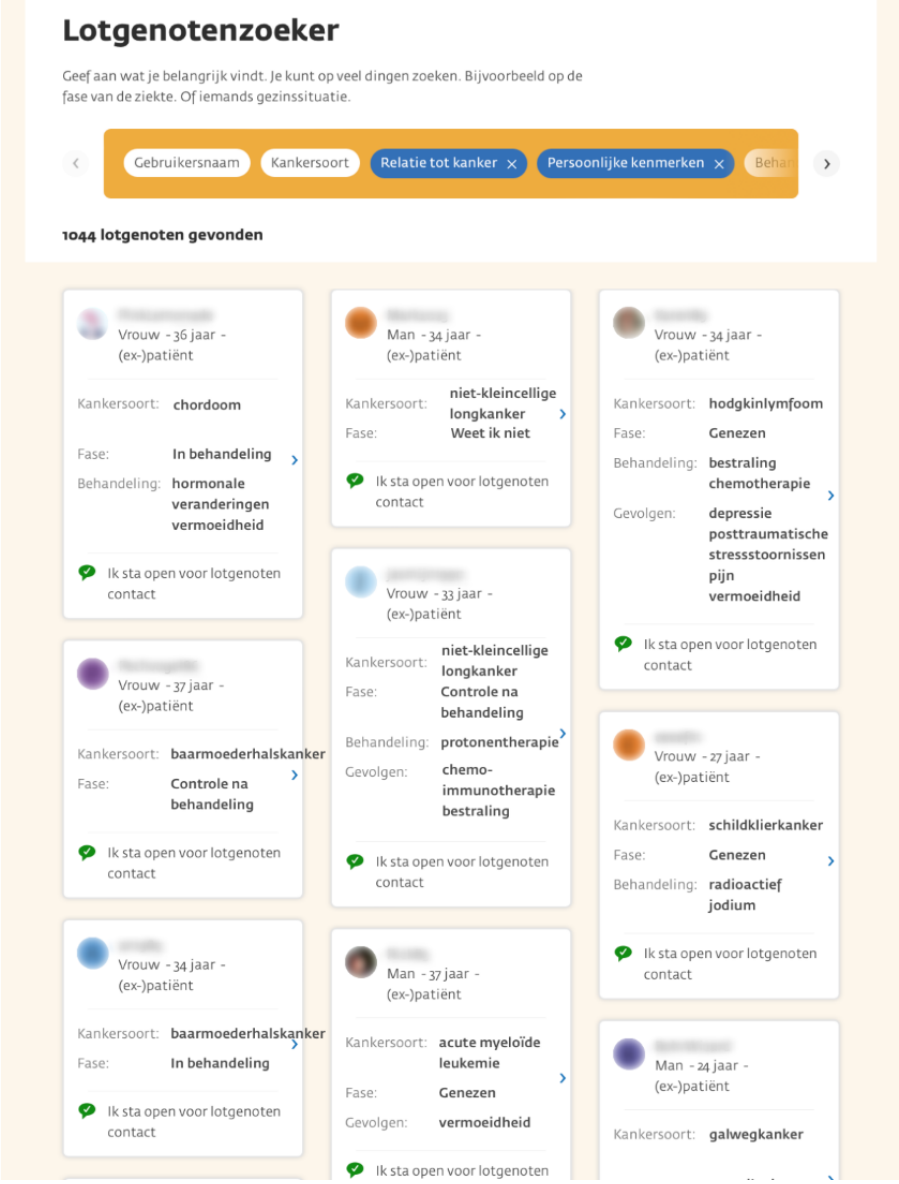

Health Care Professional on Demand

adolescent and young adult cancer survivors have the opportunity to (anonymously) pose online questions to HCPs with expertise in adolescent and young adult care. These questions and answers given by HCPs are publicly available to anyone. The HCP is an expert in the field, recognized by partner organizations, and preferably works in a multidisciplinary team within a network of colleagues/hospitals. Cooperation in this on-demand service from HCPs is on a voluntary (unpaid) basis.

Guide for Reliable Information and Support

This guide is for patients with cancer and cancer survivors who are searching for support or additional care, for instance, HCPs (physiotherapists, social workers) with expertise in treating and supporting adolescents and young adults.

Library for Resources and Specific App Store

The oncology app store consists of evidence-based and reliable applications. A couple of age-specific tools, including applications and decision aids can be found. These decision aids help adolescents and young adults in contemplating their treatment options and in preparing for discussion with HCPs.

Evaluation of the Collaborative Design Process

All project members of the project group participated in the focus group. The demographic characteristics of the focus group (mean 44, SD 15 years; n=7) are presented in Table 2.

XSL•FO
RenderX

Table 2. The demographic characteristics of the project group participated in the focus group (n=7).

Characteristics	Values
Sex, n (%)	
Male	5 (74)
Female	2 (29)
Current age (years)	
Range	22-66
Mean (SD)	44 (15)
Age at initial cancer diagnosis (years)	
Range	17-33
Mean (SD)	27 (8)
Living situation, n (%)	
Living alone	1 (14)
Living with partner	4 (57)
Living with children	2 (29)
Highest achieved level of education, n (%)	
Secondary education or less	1 (14)
Secondary vocational education	0 (0)
Applied university	2 (29)
University	4 (57)
Employment status, n (%)	
Student	1 (14)
Part-time work	3 (43)
Full-time work	2 (29)
Self-employed	1 (14)

The adolescents and young adults with lived experience in the project group considered the online platform for adolescent and young adult cancer survivors a success, as it effectively meets adolescent and young adult cancer survivors' online needs. It should be used as a first starting point in the search for information, help adolescent and young adult cancer survivors to gain awareness and find recognition on what might hit them along the way, prevent them from getting lost or stuck on their search, or reach a dead-end. Furthermore, they aimed to lay the groundwork for the platform, ensuring it is not a one-time initiative but an ongoing resource that continuously monitors and updates important information and support for adolescent and young adult cancer survivors. A positive spin-off observed from the collaborative design approach was that the texts and topics developed were not only valuable for adolescent and young adult cancer survivors but also had relevance for older cancer survivors.

The evaluation of the collaborative design consisted of three themes, which are described below, and corresponding quotes are available in Multimedia Appendix 5.

Value of Stakeholder Participation

The collaboration within a multistakeholder group avoided potential misunderstandings and missed opportunities during the development process. It enabled improvements in the information and services provided by the platform. The benefit of working with professionals from various organizations and expertise was the easy access to different expertise from different organizations and professionals. The understanding and inclusion of the perspectives of adolescents and young adults with lived experience was valuable for tailoring the online platform according to their needs and was important for validating the features.

Conditions for Working With Adolescents and Young Adults With Lived Experience

Working with adolescents and young adults with lived experience required attention to several key factors. Not only did adolescent and young adult cancer survivors feel it was important to be heard, but they also valued recognition and the assurance that their contributions were taken seriously and used in the development of future services. adolescent and young adult cancer survivors wanted to be treated as equal partners within the project, to be involved from the beginning of the project and not just seen as a "checkbox" to be ticked off once they have participated. They preferred an open atmosphere where everyone was free to express their thoughts and feelings. Their main motivation to participate in the project group was that they wanted to improve information provision and care for

XSL•FO
RenderX

future adolescent and young adult cancer survivors. However, a barrier to working as an adolescent and young adult with lived experience was the lack of financial compensation for their time and expertise, in contrast to other professionals. As a result, they often had to balance their work for the project group with other responsibilities, such as employment and child care. For this reason, the project leader felt responsible to not overburden them, by avoiding giving them additional time-consuming tasks, such as writing notes or producing materials. adolescent and young adult cancer survivors had to be capable of looking beyond their own situation and not basing their thoughts and actions solely on their own experiences. In addition, working in a project group with people with diverse (professional) backgrounds required attention to make jargon understandable for all involved.

Collaboration Between Adolescents and Young Adults With Lived Experience and Professionals From Different Backgrounds and Organizations

The project group expressed that they really enjoyed working in this way. Working with different organizations and professionals provided the advantage of a financial and future-proof perspective while also incorporating a wide range of knowledge and skills into the project group. It was sometimes difficult for stakeholders to keep their own organization adequately informed and engaged in the process, as they were representing their organization in this external and interorganizational project. Throughout the project, efforts were made to explore ways to keep the wider organization informed and connected to the “young and cancer” project.

This interorganizational approach also required adherence to certain preconditions and established procedures over which there may have been little to no control. For instance, fundamentally, the design directions were limited by the web-hosting organization Kanker.nl; the solution could technically not be a serious game or a virtual reality experience, it was rather limited towards a static website. This resulted in having to compromise between the original framework and the given limitations of the existing platform. Expectation management and communication were crucial in this context, as the project team had specific wishes and requirements that needed to be addressed.

Discussion

Principal Findings

The “young and cancer” platform is integrated into the existing Kanker.nl platform using a collaborative design approach. The centralized location improves access to and overview of available digital support resources. The integration of these resources improves both the experience and efficiency for adolescent and young adult cancer survivors. The evaluation of the collaborative process identified three themes: values of stakeholder participation, conditions for working with adolescents and young adults with lived experience, and collaboration between adolescents and young adults with lived experience and professionals from different backgrounds and organizations.

In this study, we described how we involved different stakeholders, and especially adolescent and young adult with lived experiences in various different ways along the collaborative development process. This contributes to the lack of literature about how patients are involved in the development of digital interventions [40,41]. A systematic review from McCann et al [2], showed that only 14 studies reported involving young people in the development process, and 8 studies had expert input or included a steering group during the design phase. Patient involvement included focus groups [42,43], interviews [44,45], and questionnaires [46,47]. It also included usability testing [48], think-aloud testing [42], heuristic evaluation [42], and collaboration with research partners who participated in one-day meetings [49]. These meetings included plenary sessions, small group discussions, and individual assignments. However, there is limited evidence on how these patients were precisely involved in more detail. In our study, we explain different ways in how patients can be involved to develop different parts through a combination of questionnaires and the involvement of adolescents and young adults with lived experiences who were part of the project group and involved in decision-making. The involvement of adolescents and young adults with lived experience in research poses a number of challenges. Effective communication between researchers and patients is crucial; however, it can be problematic when technical jargon is used or when there is a lack of clear explanations about what is expected of adolescent and young adult cancer survivors. Furthermore, it is challenging to engage a diverse group of adolescents and young adults with lived experience, which can represent a wide range of perspectives. This population often has competing responsibilities, such as young kids and a growing career, that limit their availability for research or co-design activities, particularly if they are also managing health issues. In addition, measuring and evaluating the impact and effectiveness of patient involvement can be difficult, leading to questions about the value of their contributions. Gaining insight into patient involvement challenges in the development of information and services for adolescent and young adult cancer survivors could provide valuable guidance to HCPs to overcome these challenges.

Lessons Learned

Collaboration between professionals, including online developers, researchers, and adolescents and young adults with lived experience, ensured a direct translation of insights into the platform. The involvement of professionals from various organizations in the project provided direct access to content creators for the platform, for instance, epidemiological research from the Integraal Kankercentrum Nederland (IKNL; Comprehensive Cancer Organisation) and technical expertise from Kanker.nl. An important aspect of working in a project group with various organizations is the discussion of mutual expectations and collaboration agreements before the start of the project. Project members working across different organizations were responsible for communicating project details and their collaborative design approach to their internal colleagues.

According to the project group, the involvement of adolescents and young adults with lived experience was considered very

XSL•FO
RenderX

important in the development of an online platform. Several key elements appeared to be important in involving adolescents and young adults with lived experience: an inclusive and open atmosphere during the collaborative process, emphasizing the principles of equality and creating a welcoming environment. adolescents and young adults with lived experiences indicated that they wanted to be involved from the initial stage of the project in order to be able to contribute to the development of the project design. This also contributed to the feeling that the patient's experience was important and valued, as well as equality and being a full partner in the project. Regular evaluation of the collaboration was essential to ensure continued alignment with the project among all members of the project group. It was valuable to involve different adolescents and young adults with lived experiences in the project in different ways and at different times during the development. In addition, the use of specialized terminology (jargon) could be a potential barrier to collaboration. Including a glossary of commonly used terms helped to bridge communication gaps, ensured that all the stakeholders spoke the same language, and facilitated a more inclusive and effective collaboration [50]. In light of these findings and also found in previous literature, such as the INVOLVE initiative from the United Kingdom [50], education for both patients and researchers could improve and guide collaboration between them. This should include an understanding of patient involvement, insights into the experiences of other patients and researchers regarding patient involvement (both facilitators and barriers), and practical skills for effective collaboration. But particularly when working in interorganizational settings, knowledge and understanding of each other's activities and organizations could be helpful.

In this study, adolescents and young adults with lived experience reported that there was no formal financial compensation available for their contribution to the project group, but they did receive informal financial compensation in the form of gift cards. This was also reported in a recent systematic review by Grace Fox et al [51], where only 25% reported offering financial compensation, mostly in the form of honoraria or stipend, salary, or gift cards [52]. In this study, compensation for their contributions as adolescents and young adults with lived experiences was seen as a meaningful form of acknowledgment and demonstrated that their efforts were recognized and valued. They also mentioned the financial responsibilities they have, such as paying for a mortgage, rent, and supporting their family, all while coping with the consequences of their cancer diagnosis and treatment. Although there is previous research and guidelines on compensating patients for their contributions to health research [50,52,53], not all organizations and professionals are used to this and do not include these costs in the project budgets [52]. In line with our study, barriers hindering researchers from providing financial compensation were the limited funding and the lack of institutional guidance [52]. Organizations should have clear policies on their approach to compensating patients for their involvement, budgeting for patient involvement should be included in the research protocol and professionals should have guidance on this.

Future Proof

Many developed interventions remain unused because they don't make it out of the pilot phase due to limited funding. The COMPRAYA funding has provided the same start and leads to the opportunity to establish this "young and cancer" platform. However, given the temporary nature of this funding, it is important to consider its future beyond the funding period. An important aspect for choosing and creating the "young and cancer" platform within the existing Kanker.nl platform was its assurance of continued development and maintenance of the platform, as it is continuously funded by KWF. Formal agreements among stakeholders COMPRAYA, SJK, Dutch "Young and Cancer" Care Network and Kanker.nl ensure integration and further development of the "young and cancer" services in the future. The collaboration between COMPRAYA, SJK, and Kanker.nl is essential for ensuring continuous involvement in the platform. Within this agreement, the unique expertise of each organization is used. Kanker.nl plays a crucial role by hosting the platform on its website, leveraging its technical expertise in building and maintaining digital resources. Meanwhile, COMPRAYA, SJK, and Dutch adolescent and young adult "Young & Cancer" Care Network contribute valuable insights and input based on their specialized knowledge and experience on adolescent and young adult cancer survivors and their relatives based on scientific knowledge.

Opportunities and Implications for the "Young and Cancer" Platform

As we continue to gain more insight into the (medical) care and support needs of adolescent and young adult cancer survivors and their relatives and how they might benefit from digital interventions, the information provision and services on the "young and cancer" platform need to be updated constantly. This includes exploring how to increase awareness of the online platform among adolescent and young adult cancer survivors. Insights from HCPs regarding their experiences with online services and the extent to which they inform adolescent and young adult survivors about this platform could also be helpful [54,55]. Conducting evaluation and assessing experiences and needs among visitors (both adolescent and young adult cancer survivors and their relatives) and HCPs will help identify what works well, what is lacking, and areas for improvement.

More research is needed to assess the use and impact of the "young and cancer" platform, focusing on the evolving needs and needs of underrepresented adolescent and young adult subgroups who might have different needs. For example, adolescents and young adults with an uncertain or poor prognosis [46], adolescents and young adults with low (health) literacy [46], adolescent and young adult cancer survivors with diverse cultural backgrounds, and Young Adult Childhood Cancer Survivors. More insight is needed into what already exists for these specific subgroups and to identify what is needed for customizing information and services to their specific needs. By continually refining the platform to meet adolescent and young adult cancer survivors' needs, we aim to maximize its impact and utility.

XSL•FO
RenderX

Limitations

The main limitation of this study was that there were no direct observations of the stakeholder consensus meetings and collaborative design meetings. Thereby, to evaluate the collaborative design, only project members were interviewed. Conducting interviews with a broader range of stakeholders with different backgrounds could enrich the findings and offer more interdisciplinary insights. The adolescents and young adults with lived experiences who were part of the project group were not representative of the adolescent and young adult population. They were all highly educated, and there was no diversity in terms of ethnicity. All had some experience of working as a patient representative, which can be seen as an advantage as they are familiar with the field and may have built up a network. However, the risk of having a "professional patient partner" is that their own patient experience may be sidelined and input become more similar to that of researchers. We have tried to minimize this by involving inexperienced adolescent and young adult cancer survivors at several points in time during the process, such as the stakeholder consensus meeting and by asking them to review the content of the platform.

Conclusion

The collaborative design approach helped to integrate online information and services for adolescent and young adult cancer survivors. The collaboration between professionals, including online developers, researchers, and adolescents and young adults with lived experience, facilitated a direct translation of insights into the platform, while the support of the national cancer platform provides long-term sustainability. This resulted in the development of a dedicated sustainable "young and cancer" platform. This study highlights the importance of strategic stakeholder selection and intense involvement of the stakeholders through a collaborative design approach. Knowledge and understanding of each other's activities and organizations can facilitate better communication and contribute to more integrated collaboration, especially when working on an interorganizational project.

Acknowledgments

The authors would like to thank all the adolescent and young adult cancer survivors for their participation in this study, and the Patient Reported Outcomes Following Initial Treatment and Long-Term Evaluation of Survivorship (PROFILES) registry group for the data collection support.

Data Availability

The datasets generated during and/or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
COREQ checklist.
[DOCX File , 26 KB - formative_v9i1e60897_app1.docx]

Multimedia Appendix 2
The various scenarios presented to stakeholders outlined different strategies for improving online information and services for adolescent and young adult cancer survivors, each varying in terms of investment and complexity.
[PPTX File , 90 KB - formative_v9i1e60897_app2.pptx]

Multimedia Appendix 3
Four personas representing adolescent and young adult cancer survivors and their relatives, covering various characteristics such as gender, tumor types, age, time since diagnosis, and platform needs.
[PPTX File , 8546 KB - formative_v9i1e60897_app3.pptx]

Multimedia Appendix 4
A screenshot of the age-specific content on the "young and cancer" platform, related to fertility and pregnancy.
[PNG File , 1976 KB - formative_v9i1e60897_app4.png]

Multimedia Appendix 5
(Sub)themes of the evaluation of the collaboration within the project group.
[XLSX File (Microsoft Excel File), 13 KB - formative_v9i1e60897_app5.xlsx]

References

XSL•FO
RenderX

1. Shaffer KM, Turner KL, Siwik C, Gonzalez BD, Upasani R, Glazer JV, et al. Digital health and telehealth in cancer care: a scoping review of reviews. Lancet Digit Health 2023;5(5):e316-e327 [FREE Full text] [doi: 10.1016/S2589-7500(23)00049-3] [Medline: 37100545]
2. McCann L, McMillan KA, Pugh G. Digital interventions to support adolescents and young adults with cancer: systematic review. JMIR Cancer 2019;5(2):e12071 [FREE Full text] [doi: 10.2196/12071] [Medline: 31368438]
3. Devine KA, Viola AS, Coups EJ, Wu YP. Digital health interventions for adolescent and young adult cancer survivors. JCO Clin Cancer Inform 2018;2:1-15 [FREE Full text] [doi: 10.1200/CCI.17.00138] [Medline: 30652583]
4. Doyle-Lindrud S. State of eHealth in cancer care: review of the benefits and limitations of eHealth tools. Clin J Oncol Nurs 2020;24(3):10-15. [doi: 10.1188/20.CJON.S1.10-15] [Medline: 32441698]
5. Changrani K, Chima S, Sharma A, Han G, Sharma A, McNamara M, et al. A systematic review of smartphone applications for cancer survivors. J Cancer Surviv 2024;18(6):1951-1973. [doi: 10.1007/s11764-023-01435-9] [Medline: 37700151]
6. Adam R, McMichael D, Powell D, Murchie P. Publicly available apps for cancer survivors: a scoping review. BMJ Open 2019;9(9):e032510 [FREE Full text] [doi: 10.1136/bmjopen-2019-032510] [Medline: 31575584]
7. Kanker.nl. URL: https://www.kanker.nl/ [accessed 2023-10-01]
8. WHO Organization. Using e-health and information technology to improve health. 2023. URL: https://www.who.int/westernpacific/activities/using-e-health-and-information-technology-to-improve-health [accessed 2025-02-20]
9. Szinay D, Jones A, Chadborn T, Brown J, Naughton F. Influences on the uptake of and engagement with health and well-being smartphone apps: systematic review. J Med Internet Res 2020;22(5):e17572 [FREE Full text] [doi: 10.2196/17572] [Medline: 32348255]
10. Ferrari A, Stark D, Peccatori FA, Fern L, Laurence V, Gaspar N, et al. Adolescents and young adults (AYA) with cancer: a position paper from the AYA working group of the European society for medical oncology (ESMO) and the European society for paediatric oncology (SIOPE). ESMO Open 2021;6(2):100096 [FREE Full text] [doi: 10.1016/j.esmoop.2021.100096] [Medline: 33926710]
11. Saloustros E, Stark DP, Michailidou K, Mountzios G, Brugieres L, Peccatori FA, et al. The care of adolescents and young adults with cancer: results of the ESMO/SIOPE survey. ESMO Open 2017;2(4):e000252. [doi: 10.1136/esmoopen-2017-000252] [Medline: 29018578]
12. Zebrack BJ. Psychological, social, and behavioral issues for young adults with cancer. Cancer 2011;117(10 Suppl):2289-2294 [FREE Full text] [doi: 10.1002/cncr.26056] [Medline: 21523748]
13. DeRouen MC, Smith AW, Tao L, Bellizzi KM, Lynch CF, Parsons HM, AYA HOPE Study Collaborative Group. Cancer-related information needs and cancer's impact on control over life influence health-related quality of life among adolescents and young adults with cancer. Psychooncology 2015;24(9):1104-1115 [FREE Full text] [doi: 10.1002/pon.3730] [Medline: 25611943]
14. Zebrack BJ, Block R, Hayes-Lattin B, Embry L, Aguilar C, Meeske KA, et al. Psychosocial service use and unmet need among recently diagnosed adolescent and young adult cancer patients. Cancer 2013;119(1):201-214 [FREE Full text] [doi: 10.1002/cncr.27713] [Medline: 22744865]
15. Internet access and internet activities; person characteristics. Statistics Netherlands. URL: https://opendata.cbs.nl/#/CBS/nl/dataset/84888NED/table [accessed 2023-10-03]
16. Kaal SE, Husson O, van Dartel F, Hermans K, Jansen R, Manten-Horst E, et al. Online support community for adolescents and young adults (AYAs) with cancer: user statistics, evaluation, and content analysis. Patient Prefer Adherence 2018;12:2615-2622 [FREE Full text] [doi: 10.2147/PPA.S171892] [Medline: 30584285]
17. Schiffman JD, Csongradi E, Suzuki LK. Internet use among adolescent and young adults (AYA) with cancer. Pediatr Blood Cancer 2008;51(3):410-415. [doi: 10.1002/pbc.21616] [Medline: 18506753]
18. Allison KR, Patterson P, Guilbert D, Noke M, Husson O. Logging On, Reaching Out, and Getting By: a review of self-reported psychosocial impacts of online peer support for people impacted by Cancer. Proc ACM Hum Comput Interact 2021;5(CSCW1):1-35. [doi: 10.1145/3449169]
19. Stichting Jongeren en Kanker. URL: https://www.jongerenenkanker.nl/ [accessed 2023-10-01]
20. Activiteiten. IPSO Centra voor leven met en na kanker. URL: https://ipso.nl/ [accessed 2023-10-01]
21. Sleeman SHE, Reuvers MJP, Manten-Horst E, Verhees B, Patterson P, Janssen SHM, et al. 'Let Me Know If There's Anything I Can Do for You', the development of a mobile application for adolescents and young adults (AYAs) with cancer and their loved ones to reconnect after diagnosis. Cancers (Basel) 2022;14(5):1178 [FREE Full text] [doi: 10.3390/cancers14051178] [Medline: 35267484]
22. Husson O, Ligtenberg MJL, van de Poll-Franse LV, Prins JB, van den Bent MJ, van Eenbergen MC, et al. Comprehensive assessment of incidence, risk factors, and mechanisms of impaired medical and psychosocial health outcomes among adolescents and young adults with cancer: protocol of the prospective observational COMPRAYA cohort study. Cancers (Basel) 2021;13(10):2348 [FREE Full text] [doi: 10.3390/cancers13102348] [Medline: 34068026]
23. Vandekerckhove P, de Mul M, de Groot L, Elzevier HW, Fabels B, Haj Mohammad S, et al. Lessons for employing participatory design when developing care for young people with cancer: a qualitative multiple-case study. J Adolesc Young Adult Oncol 2021;10(4):404-417 [FREE Full text] [doi: 10.1089/jayao.2020.0098] [Medline: 33185496]

XSL•FO
RenderX

24. van Gemert-Pijnen JEWC, Nijland N, van Limburg M, Ossebaard HC, Kelders SM, Eysenbach G, et al. A holistic framework to improve the uptake and impact of eHealth technologies. J Med Internet Res 2011;13(4):e111 [FREE Full text] [doi: 10.2196/jmir.1672] [Medline: 22155738]
25. Holmes L, Cresswell K, Williams S, Parsons S, Keane A, Wilson C, et al. Innovating public engagement and patient involvement through strategic collaboration and practice. Res Involv Engagem 2019;5(1):30 [FREE Full text] [doi: 10.1186/s40900-019-0160-4] [Medline: 31646001]
26. Vandekerckhove P, de Mul M, Bramer WM, de Bont AA. Generative participatory design methodology to develop electronic health interventions: systematic literature review. J Med Internet Res 2020;22(4):e13780. [doi: 10.2196/13780] [Medline: 32338617]
27. Vandekerckhove P, Timmermans J, de Bont A, de Mul M. Diversity in stakeholder groups in generative co-design for digital health: assembly procedure and preliminary assessment. JMIR Hum Factors 2023;10:e38350 [FREE Full text] [doi: 10.2196/38350] [Medline: 36787170]
28. Sanders EBN, Stappers PJ. Probes, toolkits and prototypes: three approaches to making in codesigning. CoDesign 2014;10(1):5-14. [doi: 10.1080/15710882.2014.888183]
29. Dolfsma W, van der Eijk R, Mahdi S. Front-loading the front-end of innovation. J Mod Proj Manag 2022;10(1):141-149 [FREE Full text]
30. Bradbury H, Waddell S, O' Brien K, Apgar M, Teehankee B, Fazey I. A call to action research for transformations: the times demand it. Action Research 2019;17(1):3-10. [doi: 10.1177/1476750319829633]
31. Coghlan D, Brydon-Miller M. The SAGE Encyclopedia of Action Research. London, United Kingdom: SAGE Publications Ltd; 2014.
32. Smits DW, van Meeteren K, Klem M, Alsem M, Ketelaar M. Designing a tool to support patient and public involvement in research projects: the involvement matrix. Res Involv Engagem 2020;6:30 [FREE Full text] [doi: 10.1186/s40900-020-00188-4] [Medline: 32550002]
33. Vlooswijk C, Husson O, Krahmer EJ, Bijlsma R, Kaal SEJ, Sleeman SHE, et al. Differences in internet use and eHealth needs of adolescent and young adult versus older cancer patients; results from the PROFILES registry. Cancers (Basel) 2021;13(24):6308 [FREE Full text] [doi: 10.3390/cancers13246308] [Medline: 34944928]
34. Vlooswijk C, Poll-Franse LVVD, Janssen SHM, Derksen E, Reuvers MJP, Bijlsma R, et al. Recruiting adolescent and young adult cancer survivors for patient-reported outcome research: experiences and sample characteristics of the SURVAYA study. Curr Oncol 2022;29(8):5407-5425 [FREE Full text] [doi: 10.3390/curroncol29080428] [Medline: 36005166]
35. SONCOS standardisation report 11 2023. Multi-disciplinary standardisation of oncology care in the Netherlands. URL: https://demedischspecialist.nl/sites/default/files/2023-06/soncos_normeringsrapport_11_english_version.pdf [accessed 2025-02-20]
36. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
37. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
38. van de Graaf DL, Vlooswijk C, Bol N, Krahmer EJ, Bijlsma R, Kaal S, et al. AYAs' online information and eHealth needs: a comparison with healthcare professionals' perceptions. Cancer Med 2023;12(2):2016-2026 [FREE Full text] [doi: 10.1002/cam4.5048] [Medline: 35879825]
39. Kanker.nl/jong. URL: https://www.kanker.nl/jong [accessed 2025-02-20]
40. Collins TE, Akselrod S, Altymysheva A, Nga PTQ, Banatvala N, Berlina D, Expert Contributors Group. The promise of digital health technologies for integrated care for maternal and child health and non-communicable diseases. BMJ 2023;381:e071074 [FREE Full text] [doi: 10.1136/bmj-2022-071074] [Medline: 37220916]
41. van Kessel R, Roman-Urrestarazu A, Anderson M, Kyriopoulos I, Field S, Monti G, et al. Mapping factors that affect the uptake of digital therapeutics within health systems: scoping review. J Med Internet Res 2023;25:e48000 [FREE Full text] [doi: 10.2196/48000] [Medline: 37490322]
42. Jones JK, Kamani SA, Bush PJ, Hennessy KA, Marfatia A, Shad AT. Development and evaluation of an educational interactive CD-ROM for teens with cancer. Pediatr Blood Cancer 2010;55(3):512-519 [FREE Full text] [doi: 10.1002/pbc.22608] [Medline: 20533523]
43. Huang JS, Dillon L, Terrones L, Schubert L, Roberts W, Finklestein J, et al. Fit4Life: a weight loss intervention for children who have survived childhood leukemia. Pediatr Blood Cancer 2014;61(5):894-900 [FREE Full text] [doi: 10.1002/pbc.24937] [Medline: 24436138]
44. Akard TF, Gilmer MJ, Friedman DL, Given B, Hendricks-Ferguson VL, Hinds PS. From qualitative work to intervention development in pediatric oncology palliative care research. J Pediatr Oncol Nurs 2013;30(3):153-160 [FREE Full text] [doi: 10.1177/1043454213487434] [Medline: 23632900]
45. Berg CJ, Stratton E, Giblin J, Esiashvili N, Mertens A. Pilot results of an online intervention targeting health promoting behaviors among young adult cancer survivors. Psychooncology 2014;23(10):1196-1199 [FREE Full text] [doi: 10.1002/pon.3526] [Medline: 24639118]

XSL•FO
RenderX

46. Knijnenburg SL, Kremer LC, Versluys AB, Braam KI, Mud MS, van der Pal HJ, et al. Evaluation of a patient information website for childhood cancer survivors. Support Care Cancer 2013;21(4):919-926. [doi: 10.1007/s00520-012-1604-7] [Medline: 23007883]
47. Beale IL, Kato PM, Marin-Bowling VM, Guthrie N, Cole SW. Improvement in cancer-related knowledge following use of a psychoeducational video game for adolescents and young adults with cancer. J Adolesc Health 2007;41(3):263-270. [doi: 10.1016/j.jadohealth.2007.04.006] [Medline: 17707296]
48. Stinson JN, Jibb LA, Nguyen C, Nathan PC, Maloney AM, Dupuis LL, et al. Development and testing of a multidimensional iPhone pain assessment application for adolescents with cancer. J Med Internet Res 2013;15(3):e51 [FREE Full text] [doi: 10.2196/jmir.2350] [Medline: 23475457]
49. Winterling J, Wiklander M, Obol CM, Lampic C, Eriksson LE, Pelters B, et al. Development of a self-help web-based intervention targeting young cancer patients with sexual problems and fertility distress in collaboration with patient research partners. JMIR Res Protoc 2016;5(2):e60 [FREE Full text] [doi: 10.2196/resprot.5499] [Medline: 27073057]
50. Briefing notes for researchers: involving the public in NHS, public health and social care research. INVOLVE. NIHR. 2012. URL: https://www.nihr.ac.uk/briefing-notes-researchers-public-involvement-nhs-health-and-social-care-research [accessed 2021-04-05]
51. Fox G, Lalu MM, Sabloff T, Nicholls SG, Smith M, Stacey D, et al. Recognizing patient partner contributions to health research: a systematic review of reported practices. Res Involv Engagem 2023 Sep 09;9(1):80 [FREE Full text] [doi: 10.1186/s40900-023-00488-5] [Medline: 37689741]
52. Fox G, Lalu MM, Sabloff T, Nicholls SG, Smith M, Stacey D, et al. Recognizing patient partner contributions to health research: a systematic review of reported practices. Res Involv Engagem 2023;9(1):80. [doi: 10.1186/s40900-023-00488-5] [Medline: 37689741]
53. Richards DP, Jordan I, Strain K, Press Z. Patient partner compensation in research and health care: the patient perspective on why and how. Patient Exp J 2018;5(3):6-12. [doi: 10.35680/2372-0247.1334]
54. Adduci A, Grampa P, Barretta F, Sironi G, Silva M, Luksch R, et al. Adolescents and young adults with cancer: surfing the web to know and manage their disease. EJC Paediatr Oncol 2023;1:100009 [FREE Full text] [doi: 10.1016/j.ejcped.2023.100009]
55. Lea S, Martins A, Fern LA, Bassett M, Cable M, Doig G, et al. The support and information needs of adolescents and young adults with cancer when active treatment ends. BMC Cancer 2020;20(1):697 [FREE Full text] [doi: 10.1186/s12885-020-07197-2] [Medline: 32723357]

Abbreviations

COMPRAYA: COMPRehensive health outcome and intervention research among patients with adolescent and young adult cancer
COREQ: Consolidated Criteria for Reporting Qualitative Research
HCP: health care professional
IKNL: Integraal Kankercentrum Nederland
KWF: Koningin Wilhelmina Fonds
NFK: Nederlandse Federatie van Kankerpatiëntenorganisaties
SURVAYA: health-related quality of life and late effects among SURVivors of cancer in Adolescence and Young Adulthood
SJK: Stichting Jongeren en Kanker

Edited by A Mavragani; submitted 24.05.24; peer-reviewed by F Horick; comments to author 26.08.24; revised version received 23.11.24; accepted 11.12.24; published 11.07.25.

Please cite as:
Vlooswijk C, Sleeman SHE, Pluis J, Bakker D, de Groot L, Manten E, Heine P, Husson O, van Eenbergen MC, Vandekerckhove P Developing and Integrating Digital Sources in an Accessible and Sustainable Online Platform for Adolescents and Young Adult Cancer Survivors: Collaborative Design Approach
JMIR Form Res 2025;9:e60897
URL: https://formative.jmir.org/2025/1/e60897
doi:10.2196/60897
PMID:

©Carla Vlooswijk, Sophia HE Sleeman, Jonas Pluis, Daphne Bakker, Lisanne de Groot, Eveliene Manten, Peter Heine, Olga Husson, Mies C van Eenbergen, Pieter Vandekerckhove. Originally published in JMIR Formative Research

XSL•FO
RenderX

(https://formative.jmir.org), 11.07.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Exploring the Impact of Online Mental Health Resources During the COVID-19 Pandemic on Lesbian, Gay, Bisexual, Transgender, Queer, and Questioning Adults Compared to Heterosexual Adults: Pretest-Posttest Survey Analyses

Natalia Ramos[1*], MPH, MD; Skylar Jones[1*], BA; Lily Zhang[2*], MS; Miriam Nuño[3*], PhD; Benita Ramsey[4*], MA, JD; Dannie Ceseña[5*], MPH; Alyssa Mireles[2], MPH; Kenneth Wells[1,6,7*], MPH, MD

[1]Department of Psychiatry and Biobehavioral Sciences, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, United States

[2]Research Center for Health Services and Society, Jane and Terry Semel Institute for Neuroscience and Human Behavior, University of California, Los Angeles, Los Angeles, CA, United States

[3]Department of Public Health Sciences, University of California Davis, Davis, CA, United States

[4]Rainbow Pride Youth Alliance, San Bernadino, CA, United States

[5]California LGBTQ Health Human Services Network, Carson, CA, United States

[6]Center for Health Services and Society, Jane and Terry Semel Institute for Neuroscience and Human Behavior, University of California, Los Angeles, Los Angeles, CA, United States

[7]Department of Health Policy and Management, Fielding School of Public Health, University of California, Los Angeles, Los Angeles, CA, United States

[*]these authors contributed equally

Corresponding Author:
Kenneth Wells, MPH, MD
Center for Health Services and Society
Jane and Terry Semel Institute for Neuroscience and Human Behavior
University of California, Los Angeles
760 Westwood Plaza, Suite 17-369b
Los Angeles, CA, 90024
United States
Phone: 1 3107943728
Email: kwells@mednet.ucla.edu

Abstract

Background: Lesbian, gay, bisexual, transgender, queer, and questioning (LGBTQ+) individuals faced greater mental health challenges during the COVID-19 pandemic than binary-gender heterosexual (non-LGBTQ+) adults. The Together for Wellness/Juntos por Nuestro Bienestar website with free well-being resources, developed during the COVID-19 pandemic with partner input, included LGBTQ+ resources. A pilot evaluation among adults (aged ≥18 years) found engagement with and use of the website 4 to 6 weeks before follow-up was associated with reduced (pretest-posttest) depression. Results for LGBTQ+ participants were not reported.

Objective: This study describes baseline depression, anxiety, and website engagement for LGBTQ+ compared with non-LGBTQ+ adults and pretest-posttest changes in depression and anxiety (the primary outcome).

Methods: Community partners invited health and social services providers, clients, and partners to visit the website and complete a survey app (Chorus Innovations) at baseline (September 20, 2021-April 4, 2022) and a 4- to 6-week follow-up (October 22, 2021-May 17, 2022). LGBTQ+ adults were compared to non-LGBTQ+ adults in demographics, website use, depression, and anxiety. Sensitivity analyses were adjusted for nonresponse (inverse probability weighting). Regression analyses identified predictors for reduction (pretest-posttest) in depression (2-item Patient Health Questionnaire [PHQ-2]) and anxiety (2-item Generalized Anxiety Disorder scale [GAD-2]).

Results: Of 315 adults who completed the baseline survey and 193 who completed the follow-up survey, 64 (20.3%) and 37 (19.2%), respectively, were LGBTQ+. At baseline, LGBTQ+ compared to non-LGBTQ+ adults had higher scores on the PHQ-2

XSL•FO
RenderX

(mean 2.4, SD 1.7 vs 1.3, SD 1.3; t_{294}=5.31; P<.001) and GAD-2 (mean 2.7, SD 1.7 vs 1.6, SD 1.5; t_{295}=4.96; P<.001) and more COVID-19 stressors (mean score 8.1, SD 4.4 vs 6.5, SD 4.0; t_{298}=2.8; P=.003). Before follow-up, LGBTQ+ adults had similar website use (P=.65) and likelihood to recommend the website to others (P=.26) compared to non-LGBTQ+ adults. LGBTQ+ adults had more reduction (pretest-posttest) in mean GAD-2 scores (−0.8, SD 2.0 vs 0.0, SD 1.2; t_{177}=−3.08; P=.002) and mean PHQ-2 scores (−0.7, SD 1.7 vs −0.1, SD 1.4; t_{180}=−2.16; P=.03) compared to non-LGBTQ+ adults. For LGBTQ+ adults, predictors of pretest-posttest decline (adjusting for nonresponse) in mean GAD-2 scores included visiting the website and using resources 4 to 6 weeks before (β=−1.95, 95% CI −3.20 to −0.70; P=.003); for decline in mean PHQ-2, visiting website/using resources had a trend as predictor that was not significant adjusting for nonresponse (β=-.94 (-2.00, 0.013), P=.09).

Conclusions: LGBTQ+ adults reported higher baseline depression, anxiety, and COVID-19 stressors than non-LGBTQ+ adults. Among LGBTQ+ but not among non-LGBTQ+ adults, higher website use was associated with reduced anxiety over time. Findings suggest that online resources may promote well-being for LGBTQ+ adults in pandemics.

(JMIR Form Res 2025;9:e67082) doi:10.2196/67082

KEYWORDS

digital mental health; depression; anxiety; prevention; COVID-19; lesbian, gay, bisexual, transgender, queer, and questioning; LGBTQ+; well-being resources

Introduction

Background

Lesbian, gay, bisexual, transgender, queer, and questioning (LGBTQ+) individuals experience significant inequities in mental health indicators. Research studies consistently demonstrate high rates of depression, anxiety, suicidal thoughts and behaviors, and nonsuicidal self-injury among LGBTQ+ adolescents and adults [1-3]. US data for the 7 years before the COVID-19 pandemic showed that LGBTQ+ adults aged 18 to 65 years had higher odds of alcohol and drug use disorder than their cisgender heterosexual peers [4]. Minority stress theory explains that these mental health inequities result from experiences of identity-linked stigma and discrimination—for example, employment discrimination, anti-LGBTQ+ legislation [5], and prejudiced statements—across social settings and over sustained periods [6,7]. Furthermore, LGBTQ+ individuals face unique barriers in mental health care settings, including identity disclosure [8,9]; low rates of health insurance coverage [10] difficulty finding well-trained, gender-affirming mental health care providers [11].

During the COVID-19 pandemic, US and international studies alike showed greater increases in stress, anxiety, and depression among LGBTQ+ individuals as compared to their non-LGBTQ+ peers, with variations observed among specific LGBTQ+ and gender-diverse subpopulations [12-17]. LGBTQ+ young adults in the United States had a disproportionate prevalence of major depressive disorder and generalized anxiety disorder [18]. These amplified mental health challenges faced by LGBTQ+ communities may be attributable to (1) the disproportionate burden of social stressors impacting already less resourced and more marginalized communities and (2) reduced access to social and community support networks [19] and health care resources known to buffer stressors among LGBTQ+ populations. Both general social support and identity-specific support help prevent anxiety and depression among gender-diverse individuals [20-22]. Unsurprisingly, data suggest that social isolation may have contributed to anxiety, stress, and fear during the pandemic [23].

Pandemic data reveal disproportionate mental health impacts across different gender identities and sexual orientations. During the pandemic, women reported experiencing more stress and a higher need for mental health support compared to men [24-26]. Gender-diverse individuals experienced more severe depression, anxiety, and stress than cisgender respondents in a cross-sectional Canadian study conducted early in the pandemic [27]. LGBTQ+ adults experienced greater psychological distress and peritraumatic stress than cisgender heterosexual adults [16]. Although most communities experienced lowered access to mental health care during pandemic shutdowns, LGBTQ+ individuals may have been especially impacted given the compounded barriers to care and a higher baseline need. Gender-diverse (eg, transgender, nonbinary, and genderqueer) individuals also faced unique stress related to poor access to gender-affirming care [27]. In addition, the COVID-19 pandemic may have disproportionately affected younger populations, which have a higher representation of LGBTQ+ individuals [19].

Digital Resources

Digital tools may serve as an important resource to support the well-being of LGBTQ+ individuals, who often face pronounced social stressors, isolation from supportive spaces, and limited care access both during periods of reduced socialization such as the COVID-19 pandemic and in their everyday lives. The demand for digital mental health services increased substantially worldwide during the pandemic given the strain on already overburdened mental health systems [28-30]. Digital resources providing health care services and social support took on many forms during the pandemic, including videoconferencing, educational websites, mobile health apps, SMS text messaging, and online chats (live or chatbots) [31]. Most digital technologies were web based, although supportive mobile SMS text messaging and online chats were also offered in multiple countries [19]. In Australia, efforts to integrate digital youth portals focused on mental illness prevention launched via existing brick-and-mortar youth mental health centers in an effort to expand access across rural areas and decrease barriers to care [20]. A systematic review of LGBTQ+-specific digital resources for youth during the pandemic found that both

XSL•FO
RenderX

structured and unstructured formal resources—such as telehealth and mobile apps—reduced mental illness symptoms despite needed improvement of mobile apps [22]. In addition, remote services may enhance access to LGBTQ+-affirming substance use treatment and harm reduction services [32].

Despite the growth in interest and demand for digital mental health resources, data on the effectiveness of digital interventions for mental health remain limited [30]. A 2016 systematic review of depression apps found that only 10% integrated evidence-based interventions such as cognitive behavioral therapy and behavioral activation [33]. Digital resources, of increased importance during the COVID-19 pandemic [34], often face limitations in their ability to be tailored to specific communities and cultural contexts [34,35]. To maximize the effectiveness of digital interventions, it is crucial to consider population-specific factors and the unique needs of users [34-36]. For LGBTQ+ individuals experiencing depression and anxiety symptoms, digital interventions can help address the emotional and behavioral impacts of stigma-related events [36]. Community-partnered development of digital resources has previously been proposed as a potentially effective model for creating accessible, culturally relevant mental health content that fosters strong community engagement and uptake [37].

The Together for Wellness/Juntos por Nuestro Bienestar Initiative

Together for Wellness/Juntos por Nuestro Bienestar (T4W/Juntos) is a free public website focused on mental health and well-being during the COVID-19 pandemic. Funded by the California Health Care Services Division of Behavioral Health, T4W/Juntos was developed through an academic-community partnership that focused on meeting the needs of marginalized and underresourced communities affected disproportionately by COVID-19, including LGBTQ+ communities. The site offered resources that were evidence informed or evidence based, developed in collaboration with diverse partner agencies and reviewed following the principles of community-partnered participatory research [38]. The details of the initiative's community-partnered approach and conceptual framework combining the technology acceptance model and vulnerable populations framework for engagement in healthy equity–focused prevention resources are described elsewhere [39-43]. The website offered resources on stress, grief, social connections, racial discrimination, COVID-19, and support for families, teachers, and older adults during this evaluation phase [42]. A total of 133 resources were offered across various digital formats, including websites, videos, YouTube links, PDF files, apps, and hotlines, with some content available in 13 languages.

A pilot evaluation recruited adults through community partner email invitations. Participants were invited to consent, engage with a research duplicate version of the public website, and fill out surveys at baseline and after 4 to 6 weeks. Previous analyses on the sample of all respondents found significant levels of COVID-19 stressors, depression, and anxiety for adults (both English and Spanish speaking), and greater use of the website was associated with a reduction in depression for the overall sample [43]. However, the previous analyses and reports did not present findings on LGBTQ+ participants separately or in relation to non-LGBTQ+ peers. The goals of this analysis were to describe LGBTQ+ participants in the T4W/Juntos pilot evaluation and compare their use of the website and depression (2-item Patient Health Questionnaire [PHQ-2]) and anxiety (2-item Generalized Anxiety Disorder scale [GAD-2]) levels to those of non-LGBTQ+ participants [44,45]. We hypothesized that, at baseline, LGBTQ+ participants would report higher levels of COVID-19–related stress, depression, and anxiety compared to non-LGBTQ+ respondents. We further hypothesized that LGBTQ+ participants would exhibit decreased depression and anxiety symptoms with greater use of the website before follow-up (as in the main analysis for the overall sample for depression). Other significant predictors of pretest-posttest decline in depression and anxiety were also explored.

Measuring Sexual Orientation and Gender Identity

Previous research on surveying LGBTQ+ identities has emphasized the importance of affirming language, recognizing the fluidity of identity over time, assessing comfort with disclosure, and highlighting the role of community support and gender-affirming services [46,47]. While previous studies have inconsistently captured identity-related factors, some literature suggests assessing gender and sexual orientation across 3 dimensions: identity, behavior, and attraction [47]. Few studies adhere to the 2022 National Institutes of Health (NIH) recommendations to capture at least 2 of the 3 key domains [48]. For gender identity, many federally funded initiatives use the NIH-recommended 2-step process, using separate questions to survey both *sex assigned at birth* and current *gender identity* [48]. However, the category of *sex assigned at birth* has its own limitations given that it is determined solely by the appearance of the infant's external genitalia at birth, overlooking biological variations in sex traits such as chromosomes, sex hormones, and internal anatomy [49]. Furthermore, some transgender or intersex individuals and community partners also find the term problematic [50,51].

Methods

T4W/Juntos Website Resources

As explained previously and in a previous publication, the website was developed with partner input from 11 agencies, including 2 serving LGBTQ+ populations, with feedback from 4 focus groups providing input on final design [42]. The resources were largely evidence informed, including mindfulness apps, as there are few randomized trials of online resources (eg, websites and apps) designed for specific populations or tested broadly across multiple population [52-55].

Evaluation Design

The website evaluation was conducted by having 11 California community-based agencies who signed participation agreements email invites to a prespecified number of participants (clients, community partners, and health and social services providers) on their existing lists. For those with large email lists, a random selection tool was provided. The number of invites was initially 80 to 100 per agency to enroll 30 to 40 participants, later expanded to 70 to have 300 to 350 eligible participants.

XSL•FO
RenderX

Participants followed a link (through a Chorus Innovations app) to review an information sheet and consent form, and those who agreed were sent a link to complete a baseline survey and review a research duplicate of the public website only for invited participants and then invited to complete a 4- to 6-week follow-up survey. The research duplicate website was also used by the inviting agencies to track enrollment.

The survey period was September 20, 2021, to April 4, 2022, for baseline and October 22, 2021, to May 17, 2022, for follow-up.

Ethical Considerations

This study was reviewed and approved by the University of California, Los Angeles Institutional Review Board for human participants (20-002163-AM-00008). All participants provided informed consent online and were advised that participation was voluntary. Data were deidentified, with contact information for follow-up and linking data stored on a secure server separately from study data. Informed consent was affirmed at follow-ups. For each event, participants received a US $25 e-gift card. The inviting agencies were not informed of the individuals participating.

Measures

Baseline

Demographics included age in years, race or ethnicity (American Indian, Native American, or Alaskan; Black, African American, or African; East Asian; South East Asian; South Asian; Hispanic, Latino, or Spanish origin; Middle Eastern; Pacific Islander; White, Caucasian, or European; unknown; or not stated), website language preference (English, Spanish, Cantonese, Vietnamese, Tagalog, Mandarin, Korean, Japanese, Russian, Farsi, Armenian, Arabic, Mixteco, other, unknown, or not stated), educational level (some high school or lower, high school graduate or equivalent, vocational education or certificate, some college, college graduate, or graduate school), and zip code [56-59]. In total, 2 mental health stigma items were included at baseline (5-point Likert agreement scale plus “don’t know”) [56].

In consultation with community partners, we selected two items to capture gender and sexual identities: (1) gender identity (female, male, genderqueer, questioning, trans man, trans woman, other, unknown, and prefer not to state) and (2) sexual orientation (straight, gay/lesbian, bisexual/pansexual, queer, not sexual/none, questioning, other, unknown, and prefer not to state). For our analysis, we categorized participants as LGBTQ+ if they identified with a nonbinary gender (female or male) or a nonheterosexual orientation. For analyses, we excluded individuals with missing data, including those who selected “prefer not to state” or “unknown,” to explicitly focus on those directly describing their gender and sexual identities. For the main analyses, we compared the LGBTQ+ participants to the non-LGBTQ+ participants, and for the sensitivity analyses (Multimedia Appendix 1), we compared 3 groups (LGBTQ+, female heterosexual, and male heterosexual participants) overall and in 2-way analyses. These analyses were designed to inform future studies comparing mental health and stressors across gender groups given evidence that gender identity can influence these outcomes [13,18,23,26,58,59].

Baseline and Follow-Up

Depression and anxiety were assessed using mean PHQ-2 and GAD-2 scores (range 0-6; score of ≥3 indicating moderate or greater [high] levels vs a score of <3 [low]), and for the sensitivity analyses as well as fitting predictors into main analyses, a PHQ-2 or GAD-2 score of ≥3 was also included as a predictor in regression analyses if it was more significant than either PHQ-2 or GAD-2 alone [44,45]. Behavioral health service use—including behavioral health hotline use and any behavioral health service use (emergency room, primary care, mental health care, substance use visits, hospitalization, and residential care)—was assessed for the previous 6 months at baseline and for the previous month at follow-up [56]. Pandemic-related stress at baseline was measured based on the overall impact of COVID-19 as assessed using 5 categories. We used an adapted version of the COVID-19 Stress Scale to capture related behavior changes (social distancing, isolation, caring for someone at home, working from home, not working, change in health care services, following media coverage on COVID-19, and changing travel plans) [60]. We used an adapted version of COVID-19 stressors [58] that captured having COVID-19; fear of acquiring or spreading COVID-19; worrying about others; stigma or discrimination; financial, food, and home insecurity; frustration, depression, anxiety, alcohol use, sleep, and sexual activity change; confusion about COVID-19; and other difficulties (not contributing to the greater good, having social and emotional support, and having financial support from others) [58,60,61]. We included an additional measure of racial and ethnic discrimination during the COVID-19 pandemic (6 items; 4-point Likert scale [62]). For COVID-19–related stressors, we used the total score and explored the individual stressors noted previously. We also included date of survey completion.

Website use and engagement at baseline were assessed using website satisfaction, relevance, and ease of use (mean score of 3 items), as well as comfort using the website (separate item). Each item used a 5-point Likert agreement scale plus a “don’t know” option. At follow-up, website use and engagement were assessed based on (1) any use in the previous month (yes or no); (2) a total score of use summed across all 6 categories of resources, each with 5 response options (“did not use,” “used some,” “used and valuable,” “used and very valuable,” and “used and would recommend to anyone”); and (3) whether they would recommend it to others (and, if so, the number of times they recommended it and which resource categories were recommended) [37,43,56].

Statistical Analyses

We conducted analyses for the LGBTQ+ participants compared to non-LGBTQ+ participants (those identifying as having a male or female gender and heterosexual) as main analyses. For sensitivity analyses (Multimedia Appendix 1), we included a 3-way comparison of LGBTQ+, female heterosexual, and male heterosexual groups with 2-way comparisons within the 3 groups (Tables S1-S4 in Multimedia Appendix 1). Chi-square tests were used for categorical variables, and *t* tests (2-tailed) were used for continuous variables to compare the 2 groups. To

XSL•FO
RenderX

describe the samples, we used means for continuous variables and counts and percentages for categorical variables. For preliminary bivariate analyses, the predictors tested included demographics, depression and anxiety, mental health stigma, COVID-19–related behaviors and stressors, and behavioral health service use (including adjustment for follow-up nonresponse). The final regression models included predictors significant at $P<.05$.

For website use in the month before follow-up, we used logistic regression for any use and linear regression for the total score of use across the 6 categories. We considered significant predictors of pretest-posttest difference in PHQ-2 and GAD-2 mean scores as our main analysis. We initially examined all measures as predictors in bivariate analyses, including predictors of follow-up nonresponse. We fit regression models using significant individual predictors, and when including predictors, retaining only those that remained significant through sequential testing to inform the final models. Regarding the final follow-up models on website use and mental health, for the sensitivity analyses, we applied inverse probability weighting to adjust for nonresponse predictors at baseline (age) and follow-up [63,64]. All main analyses were conducted for LGBTQ+ versus non-LGBTQ+ participants. As a sensitivity analysis, we conducted 3-way overall and separate analyses for LGBTQ+, female heterosexual, and male heterosexual groups, with results and tables provided in Multimedia Appendix 1. Statistical results appear in tables (and Multimedia Appendix 1) and in the text when cited or not tabled.

Results

Sample

Of the 315 respondents at baseline, 236 (74.9%) identified as heterosexual, 17 (5.4%) identified as gay or lesbian, 31 (9.8%) identified as bisexual or pansexual, 8 (2.5%) identified as queer, 2 (0.6%) selected the option “not sexual/none,” 2 (0.6%) identified as questioning, and 4 (1.3%) identified as “other.” Regarding gender identity, of the 315 respondents, 225 (71.4%) identified as female, 70 (22.2%) identified as male, and 15 (4.8%) identified as gender diverse (including genderqueer, questioning, trans man, trans woman, and “other”). In total, 4.8% (15/315) of the participants skipped the sexual orientation or gender identity items or selected “don’t know” or “prefer not to say”; these individuals were not included in the analyses, for a total sample for analysis of 300.

Baseline Characteristics

As shown in Table 1, LGBTQ+ participants were significantly younger than non-LGBTQ+ participants (mean age 33.1, SD 12.4 years vs 40.4, SD 13.6 years; $t_{298}=-3.88$; $P<.001$). Website language preference differed ($\chi^2_2=6.3$; $P=.04$), with LGBTQ+ participants having a higher preference for English (57/73, 78% vs 191/235, 81.3%) and a lower preference for Spanish (4/63, 6% vs 32/235, 13.6%) or other language (0% vs 12/235, 5.1%) than non-LGBTQ+ participants. LGBTQ+ and non-LGBTQ+ participants did not differ significantly in race or ethnicity ($\chi^2_3=3.2$; $P=.36$) and educational level ($\chi^2_4=0.8$; $P=.93$). LGBTQ+ participants had significantly higher rates of baseline depression (mean PHQ-2 score 2.4, SD 1.7 vs 1.3, SD 1.3; $t_{294}=5.31$; $P<.001$) and anxiety (mean GAD-2 score 2.7, SD 1.7 vs 1.6, SD 1.5; $t_{295}=4.96$; $P<.001$) than non-LGBTQ+ participants. LGBTQ+ participants had a higher use of behavioral health services than non-LGBTQ+ participants (33/41, 80% vs 82/146, 56.2%; $\chi^2_1=8.0$; $P=.005$; imputing for missing data: 33/48, 69% vs 82/181, 45.3%; $P=.004$) and a significantly higher number of COVID-19 stressors (mean score 8.1, SD 4.4 vs mean 6.5, SD 4.0; $t_{298}=2.8$; $P=.005$). There were no significant differences in mean website engagement (3 items; mean score 4.0, SD 0.7 vs 4.1, SD 0.7; $t_{288}=-1.02$; $P=.31$), comfort with website use ($\chi^2_4=3.3$; $P=.51$), or mean mental health stigma (mean score 3.0, SD 0.9 vs 2.8, SD 1.0; $t_{275}=1.48$; $P=.14$).

XSL•FO
RenderX

Table 1. Characteristics (baseline survey; September 2021-April 2022) of the California Together for Wellness/Juntos por Nuestro Bienestar website study participants—lesbian, gay, bisexual, transgender, queer, and questioning (LGBTQ+)[a] versus non-LGBTQ+[b] participants (N=300).

Baseline characteristic	Overall	Male or female heterosexual participants (n=236)	LGBTQ+ participants (n=64)	Statistic[c]		*P* value
				t test (*df*)	Chi-square (*df*)	
Age (y), mean (SD)	38.9 (13.7)	40.4 (13.6)	33.1 (12.4)	−3.88 (298)	__[d]	<.001
Gender, n (%)				—	59.5 (2)	<.001
Female	217 (72.3)	183 (77.5)	34 (53.1)			
Male	68 (22.7)	53 (22.5)	15 (23.4)			
Other	15 (5)	0 (0)	15 (23.4)			
Race, n (%)				—	3.2 (3)	.36
Black or African American	53 (18.4)[e]	40 (17.6)[f]	13 (21.3)[g]			
Hispanic, Latinx, or Spanish origin	125 (43.4)[e]	104 (45.8)[f]	21 (34.4)[g]			
White	79 (27.4)[e]	58 (25.6)[f]	21 (34.4)[g]			
Other	31 (10.8)[e]	25 (11)[f]	6 (9.8)[g]			
Educational level, n (%)				—	0.8 (4)	.93
Lower than high school	21 (7)[h]	16 (6.8)[i]	5 (7.9)[j]			
High school graduate	34 (11.4)[h]	28 (11.9)[i]	6 (9.5)[j]			
Some college	84 (28.2)[h]	68 (28.9)[i]	16 (25.4)[j]			
College	113 (37.9)[h]	88 (37.4)[i]	25 (39.7)[j]			
Graduate school	46 (15.4)[h]	35 (14.9)[i]	11 (17.5)[j]			
Language preferred for website use, n (%)				—	6.3 (2)	.04
English	250 (83.9)[h]	191 (81.3)[i]	59 (93.7)[j]			
Spanish	36 (12.1)[h]	32 (13.6)[i]	4 (6.3)[j]			
Other	12 (4)[h]	12 (5.1)[i]	0 (0)[j]			
PHQ-2[k] score, mean (SD)	1.6 (1.4)	1.3 (1.3)	2.4 (1.7)	5.31 (294)	—	<.001
PHQ-2 score, n (%)				—	22.4 (1)	<.001
High (≥3)	57 (19.3)[l]	32 (13.7)[m]	25 (40.3)[n]			
Low	239 (80.7)[l]	202 (86.3)[m]	37 (59.7)[n]			
GAD-2[o] score, mean (SD)	1.8 (1.6)	1.6 (1.5)	2.7 (1.7)	4.96 (295)	—	<.001
GAD-2 score, n (%)				—	23.4 (1)	<.001
High (≥3)	76 (25.6)[p]	45 (19.2)[m]	31 (49.2)[j]			
Low	221 (74.4)[p]	189 (80.8)[m]	32 (50.8)[j]			
Stigma score, mean (SD)	2.8 (1.0)	2.8 (1.0)	3.0 (0.9)	1.48 (275)	—	.14
Engagement score (3 items)[q], mean (SD)	4.0 (0.7)	4.1 (0.7)	4.0 (0.7)	−1.02 (288)	—	.31
Did not feel comfortable using the website, n (%)				—	3.3 (4)	.51
Strongly disagree	99 (34)[r]	74 (32.3)[s]	25 (40.3)[n]			
Disagree	105 (36.1)[r]	81 (35.4)[s]	24 (38.7)[n]			

XSL•FO
RenderX

Baseline characteristic	Overall	Male or female heterosexual participants (n=236)	LGBTQ+ participants (n=64)	Statistic[c]		*P* value
				t test (*df*)	Chi-square (*df*)	
Neither agree nor disagree	22 (7.6)[r]	19 (8.3)[s]	3 (4.8)[n]			
Agree	36 (12.4)[r]	30 (13.1)[s]	6 (9.7)[n]			
Strongly agree	29 (10)[r]	25 (10.9)[s]	4 (6.5)[n]			
Any service use for emotional, mental health, alcohol, or drug problems, n (%)				—	8 (1)	.005
Yes	115 (61.5)[t]	82 (56.2)[u]	33 (80.5)[v]			
No	72 (38.5)[t]	64 (43.8)[u]	8 (19.5)[v]			
Any service use for emotional, mental health, alcohol, or drug problems (imputed)[w], n (%)				—	8.3 (1)	.004
Yes	115 (50.2)[s]	82 (45.3)[x]	33 (68.8)[y]			
No	114 (49.8)[s]	99 (54.7)[x]	15 (31.3)[y]			
Total number of COVID-19–related behavior changes, mean (SD)	3.9 (2.0)	3.8 (1.9)	4.0 (2.2)	0.88 (298)	—	.38
Total number of COVID-19 stressors experienced, mean (SD)	6.9 (4.1)	6.5 (4.0)	8.1 (4.4)	2.8 (298)	—	.005

[a]This category captures those with a gender identity other than male or female or a sexual orientation other than heterosexual (excluding missing values or “don’t know” or “prefer not to state” response options).

[b]The non-LGBTQ+ category includes participants who identified as heterosexual in their sexual orientation *and* with a binary gender identity (either male or female). None of these participants disclosed a transgender or gender-diverse identity.

[c]Chi-square tests were used for categorical variables, and *t* tests were used for continuous variables to compare the 2 groups with and without follow-up response.

[d]Not applicable.

[e]n=288.

[f]n=227.

[g]n=61.

[h]n=298.

[i]n=235.

[j]n=63.

[k]PHQ-2: 2-item Patient Health Questionnaire.

[l]n=296.

[m]n=234.

[n]n=62.

[o]GAD-2: 2-item Generalized Anxiety Disorder scale.

[p]n=297.

[q]Items (ease of use, relevance of topics, and satisfaction) were averaged as mean engagement based on 5-point Likert scales.

[r]n=291.

[s]n=229.

[t]n=187.

[u]n=146.

[v]n=41.

[w]The inpatient or rehabilitation item in this set was assigned 0 if it was skipped but the item on other use was answered. The imputed version was for the sensitivity analysis.

[x]n=181.

[y]n=48.

XSL•FO
RenderX

Baseline to Follow-Up Response

The mean number of days between baseline and follow-up for the sample that completed both (185/300, 61.7%) was 34.6 (SD 10.8) and was a borderline predictor, although it was not statistically significant, for longer follow-up for LGBTQ+ participants (37/185, 20%; mean 37.5, SD 13.2 days) versus non-LGBTQ+ participants (148/185, 80%; mean 33.9, SD 10.2 days; $t_{183}=-1.81$; $P=.07$). For follow-up completion (October 2021-May 2022) by baseline sexual orientation and gender identity for the California T4W/Juntos study participants (N=300): for the LGBTQ+ sample (n=64, 23% total sample), there were 27 (42%) with no follow-up and 37 (58%) with follow-up data ($\chi^2_1=0.5$; $P=.48$). For Non-LGBTQ+ (overall n=236, 78.7% total sample) there were 88 (37.3%) with no follow-up data and 148 (62.7%) for having follow-up data ($\chi^2_1=0.5$, $P=.48$). The difference in follow-up response between LGBTQ+ and non-LGBTQ+ participants was not significant (27/64, 42% vs 88/236, 37.3%, respectively; $\chi^2_1=0.5$; $P=.48$). Among LGBTQ+ participants, at least one respondent in the follow-up sample selected each gender identity and sexual orientation response at baseline.

Follow-Up Descriptive Characteristics

At follow-up (Table 2), there was no significant difference between LGBTQ+ and non-LGBTQ+ participants in the percentage of those who visited the website or used the resources in the preceding 4 to 6 weeks (24/37, 65% vs 90/148, 60.8%; $\chi^2_1=0.2$; $P=.65$), the total score of the categories viewed or used (mean 12.9, SD 7.7 vs mean 11.6, SD 6.2; $t_{176}=1.1$; $P=.27$), or the percentage of those who recommended the website to others (19/27, 70% vs 91/148, 61.5%; $\chi^2_1=1.26$; $P=.26$). The total COVID-19 stressor score at follow-up was greater for LGBTQ+ participants than for non-LGBTQ+ participants (mean 8.5, SD 4.4 vs mean 6.6, SD 4.2; $t_{183}=2.42$; $P=.02$). For 2 weeks before follow-up, LGBTQ+ participants did not differ significantly from non-LGBTQ+ participants in depression (mean PHQ-2 score 1.8, SD 1.6 vs 1.4, SD 1.3; $t_{183}=1.77$; $P=.08$) or anxiety (mean GAD-2 score 2.1, SD 2.0 vs 1.6, SD 1.5; $t_{180}=1.41$; $P=.16$).

XSL•FO
RenderX

Table 2. Characteristics of the California Together for Wellness/Juntos por Nuestro Bienestar (T4W/Juntos) study participants—lesbian, gay, bisexual, transgender, queer, and questioning (LGBTQ+)[a] versus non-LGBTQ+[b] at survey follow-up (October 2021-May 2022; N=185).

Characteristic (follow-up)	Overall	Non–gender diverse and heterosexual participants (n=148)	LGBTQ+ participants (n=37)	Statistic[c]		*P* value
				t test (*df*)	Chi-square (*df*)	
Visited the T4W/Juntos **site or used the resources 4-6 wk before, n (%)**				__[d]	0.2 (11)	.65
Yes	114 (61.6)	90 (60.8)	24 (64.9)			
No	71 (38.4)	58 (39.2)	13 (35.1)			
Total score of T4W/Juntos resource categories viewed or used 4-6 wk before, mean (SD)	11.9 (6.5)	11.6 (6.2)	12.9 (7.7)	1.1 (176)	—	.27
Recommended the website to others (4-6 wk before), n (%)				—	1.3 (1)	.26
Yes	110 (59.5)	91 (61.5)	19 (51.4)			
No	75 (40.5)	57 (38.5)	18 (48.6)			
Number of times recommended[e], mean (SD)	3.0 (4.6)	3.0 (4.4)	3.1 (5.3)	−0.09	—	.93
Follow-up mental health variables						
PHQ-2[f] score, mean (SD)	1.5 (1.4)	1.4 (1.3)	1.8 (1.6)	1.77 (183)	—	.08
PHQ-2 score, n (%)				—	6.0 (1)	.01
High (≥3)	38 (20.5)	25 (16.9)	13 (35.1)			
Low	147 (79.5)	123 (83.1)	24 (64.9)			
PHQ-2 pretest-posttest score change, mean (SD)	−0.2 (1.5)	−0.1 (1.4)	−0.7 (1.7)	−2.16 (180)	—	.03
GAD-2[g] score, mean (SD)	1.7 (1.6)	1.6 (1.5)	2.1 (2.0)	1.41 (180)	—	.16
GAD-2 score, n (%)				—	2.2 (1)	.14
High (≥3)	39 (21.4)[h]	28 (19.2)[i]	11 (30.6)[j]			
Low	143 (78.6)[h]	118 (80.8)[i]	25 (69.4)[j]			
GAD-2 pretest-posttest score change, mean (SD)	−0.2 (1.4)	0.0 (1.2)	−0.8 (2.0)	−3.08 (177)	—	.002
Stigma score, mean (SD)	2.9 (1.0)	2.9 (1.0)	2.8 (1.1)	−0.88 (177)	—	.38
Any service use for emotional, mental health, alcohol, or drug problems, n (%)				—	5.2 (1)	.02
Yes	53 (43.8)[k]	37 (38.5)[l]	16 (64)[m]			
No	68 (56.2)[k]	59 (61.5)[l]	9 (36)[m]			
Any service use for emotional, mental health, alcohol, or drug problems (imputed)[n], n (%)				—	5.7 (1)	.02
Yes	53 (36.1)[o]	37 (31.4)[p]	16 (55.2)[q]			
No	94 (63.9)[o]	81 (68.6)[p]	13 (44.8)[q]			
Total number of COVID-19 stressors experienced, mean (SD)	7.0 (4.3)	6.6 (4.2)	8.5 (4.4)	2.42 (183)	—	.02

[a]This category captures those with a gender identity other than male or female or a sexual orientation other than heterosexual (excluding missing values or “don’t know” or “prefer not to state” response options).

[b]The non-LGBTQ+ category includes participants who identified as heterosexual in their sexual orientation *and* with a binary gender identity (either male or female). None of these participants disclosed a transgender or gender-diverse identity.

[c]Chi-square tests were used for categorical variables, and *t* tests were used for continuous variables to compare 2 groups with and without follow-up response.

XSL•FO
RenderX

[d]Not applicable.

[e]One responder with an outlier response (value of 1000) was excluded.

[f]PHQ-2: 2-item Patient Health Questionnaire.

[g]GAD-2: 2-item Generalized Anxiety Disorder scale.

[h]n=182.

[i]n=146.

[j]n=36.

[k]n=121.

[l]n=96.

[m]n=25.

[n]The inpatient or rehabilitation item in this set was assigned 0 if it was skipped but the item on other use was answered. The imputed version was for the sensitivity analysis.

[o]n=147.

[p]n=118.

[q]n=29.

Regarding the primary analyses, pretest-posttest reduction was significantly greater for LGBTQ+ participants than for non-LGBTQ+ participants in terms of both mean PHQ-2 score (−0.7, SD 1.7 vs −0.1, SD 1.4; t_{180}=−2.16; *P*=.03) and mean GAD-2 score (−0.8, SD 2.0 vs 0.0, SD 1.2; t_{177}=−3.08; *P*=.002). During the period before follow-up, LGBTQ+ participants reported greater use of behavioral health services than non-LGBTQ+ participants (16/26, 62% vs 37/96, 39%; χ^2_1=5.2; *P*=.02), including when imputing for missing data (16/29, 55% vs 37/118, 31.4%; χ^2_1=5.7; *P*=.02).

For the exploratory analyses, we examined which individual COVID-19 stressors differed significantly for LGBTQ+ and non-LGBTQ+ respondents at baseline and follow-up. At baseline, compared to non-LGBTQ+ participants, LGBTQ+ participants had (1) significantly greater depression during the COVID-19 pandemic (χ^2_1=4.0; *P*=.045); (2) increased alcohol or other substance use (χ^2_1=13.6; *P*<.001); (3) greater change in sexual activity (χ^2_1=13.6; *P*<.001); and (4) more financial support from family and friends (χ^2_1=9.2; *P*=.002), a potential indirect measure of financial difficulty or need. At follow-up, LGBTQ+ participants reported significantly greater (1) diagnoses of COVID-19 (χ^2_1=3.9; *P*=.05), (2) fear of giving COVID-19 to someone else (χ^2_1=10.0; *P*=.002), (3) frustration or boredom (χ^2_1=4.0; *P*=.047), (4) depression during the COVID-19 pandemic (χ^2_1=6.4; *P*=.01), (5) use of alcohol and other substances (χ^2_{10}=10.3; *P*=.001), (6) change in sexual activity (χ^2_1=4.6; *P*=.03), and (7) loneliness (χ^2_1=7.2; *P*=.007).

Regression Analyses: Final Predictors of Site Use, Depression, and Anxiety at Follow-Up

Final significant predictors of follow-up measures, with adjustment for follow-up nonresponse (inverse probability weighting) for both LGBTQ+ and non-LGBTQ+ participants are provided in Table 3. For LGBTQ+ participants, final predictors of any visit to or use of the website or resources in the 4 to 6 weeks before follow-up included (1) baseline PHQ-2 or GAD-2 score of ≥3, which was significant when adjusting for nonresponse (odds ratio [OR] 5.75, 95% CI 1.13-29.43; *P*=.03); and (2) total COVID-19–related behavior changes, which was significant without and with adjusting for nonresponse (without adjusting: OR 0.59, 95% CI 0.36-0.98; *P*=.04; with adjusting: OR 0.62, 95% CI 0.40-0.98; *P*=.04). Predictors for moderate or greater depression (PHQ-2 score≥3) at follow-up were (1) baseline PHQ-2 score of ≥3 (which was significant when adjusting for nonresponse: OR 7.44, 95% CI 1.23-45.19; *P*=.03) and (2) having visited the T4W/Juntos website or used the resources before follow-up had a trend as a predictor of lower likelihood of depression, which was not significant without or with adjusting for nonresponse (without adjusting: OR 0.22, 95% CI 0.04-1.28; *P*=.09; with adjusting: OR 0.25, 95% CI 0.04-1.54; *P*=.13). Predictors of higher GAD-2 score at follow-up with and without adjusting for nonresponse were White ethnicity (with adjusting: OR 23.31, 95% CI 1.80-301.34; *P*=.02) and baseline GAD-2 score of ≥3 (with adjusting: OR 55.76, 95% CI 3.30-943.16; *P*=.005).

XSL•FO
RenderX

Table 3. Final models for predictors of visits to or use of the website and 2-item Patient Health Questionnaire (PHQ-2) score of ≥3 or 2-item Generalized Anxiety Disorder scale (GAD-2) score of ≥3 at follow-up (FU) and pretest-posttest mean score changes for the California Together for Wellness/Juntos por Nuestro Bienestar (T4W/Juntos) participants (April 2021-May 2022)—lesbian, gay, bisexual, transgender, queer, and questioning (LGBTQ+)[a] and non-LGBTQ+[b] baseline (BA) and FU predictors tested.

Variable	Main analysis (unweighted)		Sensitivity analysis (IPW[c]; nonresponse)	
	OR[d] or β[e] (95% CI)[e]	*P* value	OR or β (95% CI)	*P* value
LGBTQ+ (not including the "don't know" response option; n=37)				
Predictors of any visit to or use of the website at follow-up[f]				
PHQ-2 or GAD-2 score of ≥3 (BA)	4.81 (0.92 to 25.27)[d]	.06	5.75 (1.13 to 29.43)[d]	.03
Total number of COVID-19–related behavior changes (BA)	0.59 (0.36 to 0.98)[d]	.04	0.62 (0.40 to 0.98)[d]	.04
Secondary outcomes				
Predictors of PHQ-2 score of ≥3 at follow-up[f]				
PHQ-2 score of ≥3 (BA)	6.62 (1.15 to 38.32)[d]	.04	7.44 (1.23 to 45.19)[d]	.03
Any visit to T4W/Juntos website or use of resources (FU)	0.22 (0.04 to 1.28)[d]	.09	0.25 (0.04 to 1.54)[d]	.13
Predictors of GAD-2 score of ≥3 at follow-up[g]				
White ethnicity (BA)	29.67 (2.28 to 385.57)[d]	.01	23.31 (1.80 to 301.34)[d]	.02
GAD-2 score of ≥3 (BA)	60.32 (3.35 to 1085.00)[d]	.005	55.76 (3.30 to 943.16)[d]	.005
Primary outcomes				
Predictor of PHQ-2 pretest-posttest mean score change[f]				
Any visit to T4W/Juntos website or use of resources (FU)	−1.16 (−2.28 to 0.05)[e]	.04	−0.94 (−2.00 to 0.13)[e]	.09
Predictors of GAD-2 pretest-posttest mean score change[g]				
Hispanic or Latinx ethnicity (BA)	−1.68 (−3.22 to −0.14)[e]	.03	−1.95 (−3.20 to −0.70)[e]	.003
Any visit to T4W/Juntos website or use of resources 4-6 wk before FU	−1.48 (−2.78 to −0.17)[e]	.03	−1.08 (−2.05 to −0.10)[e]	.03
Female and male (binary) heterosexual participants (n=148)				
Predictor of any visit to or use of the website at follow-up[h]				
Feeling comfortable using the website (BA)	0.43 (0.19 to 0.94)[d]	.03	0.44 (0.19 to 1.00)[d]	.05
Secondary outcomes				
Predictors of PHQ-2 score of ≥3 at follow-up[h]				
PHQ-2 score of ≥3 (BA)	5.20 (1.83 to 14.72)[d]	.002	5.26 (1.72 to 16.10)[d]	.003
Website engagement mean score (3 items[i]; BA)	0.55 (0.32 to 0.95)[d]	.03	0.55 (0.31 to 0.97)[d]	.04
Predictors of GAD-2 score of ≥3 at follow-up[j]				
GAD-2 score of ≥3 (BA)	21.21 (7.08 to 63.60)[d]	<.001	21.36 (6.34 to 71.94)[d]	<.001
Total score of T4W/Juntos resource categories viewed or used (FU)	1.09 (1.01 to 1.18)[d]	.03	1.12 (1.12 to 1.23)[d]	.007
Primary outcomes				
Predictors of PHQ-2 pretest-posttest mean score change[h]				

XSL•FO
RenderX

Variable	Main analysis (unweighted)		Sensitivity analysis (IPW[c]; nonresponse)	
	OR[d] or β[e] (95% CI)[e]	*P* value	OR or β (95% CI)	*P* value
Any visit to T4W/Juntos website or use of resources (FU)	−0.46 (−0.93 to 0.00)[e]	.05	−0.37 (−0.84 to 0.10)[e]	.12
Total number of COVID-19 stressors (BA)	−0.07 (−0.13 to −0.02)[e]	.009	−0.09 (−0.16 to −0.02)[e]	.02
Predictors of GAD-2 pretest-posttest mean score change[h]				
Age of 18-30 y (vs older; BA)	−0.58 (−0.97 to −0.19)[e]	.004	−0.55 (−0.95 to −0.16)[e]	.006
High school educational level or lower (BA)	−0.80 (−1.28 to −0.39)[e]	<.001	−0.86 (−1.36 to −0.36)[e]	<.001
Total number of COVID-19 stressors (BA)	−0.10 (−0.14 to −0.06)[e]	<.001	−0.10 (−0.16 to −0.05)[e]	<.001

[a]This category captures those with a gender identity other than male or female or a sexual orientation other than heterosexual (excluding missing values or "don't know" or "prefer not to state" response options).

[b]The non-LGBTQ+ category includes participants who identified as heterosexual in their sexual orientation *and* with a binary gender identity (either male or female). None of these participants disclosed a transgender or gender-diverse identity.

[c]IPW: inverse probability weighting for nonresponse predictors at BA and FU.

[d]OR: odds ratio.

[e]β coeffient.

[f]Analytical sample: n=36.

[g]Analytical sample: n=33.

[h]Analytical sample: n=144.

[i]Items (ease of use, relevance of topics, and satisfaction) were averaged as mean engagement based on 5-point Likert scales.

[j]Analytical sample: n=139.

For the non-LGBTQ+ sample (Table 3), the significant predictor of use of the website or resources before follow-up with and without adjusting for nonresponse was baseline comfort with using the website (without adjusting: OR 0.43, 95% CI 0.19-0.94; *P*=.03; with adjusting: OR 0.44, 95% CI 0.19-1.00; *P*=.05). The total score of website use before follow-up was a significant predictor of a GAD-2 score of ≥3 before follow-up with and without adjusting for nonresponse (without adjusting: OR 1.09, 95% CI 1.01-1.1.8; *P*=.03; with adjusting: OR 1.12, 95% CI 1.12-1.23; *P*=.007). Mean website engagement was a significant predictor of a PHQ-2 score of ≥3 before follow-up with and without adjusting for nonresponse (without adjusting: OR 0.55, 95% CI 0.32-0.95; *P*=.03; with adjusting: OR 0.55, 95% CI 0.31-0.97; *P*=.04).

Regarding the primary outcomes of pretest-posttest reduction (Table 3) for LGBTQ+, for mean PHQ-2 score, the significant predictor was having visited the T4W/Juntos website or used the resources 4 to 6 weeks before follow-up (β=−1.16, 95% CI −2.28 to 0.05; *P*=.04), but it was not significant at *P*<.05 when adjusting for nonresponse (β=−0.92, 95% CI −2.00 to 0.13; *P*=.09). Regarding pretest-posttest reduction in GAD-2 scores, significant predictors without and with adjusting for nonresponse included Hispanic or Latinx ethnicity at baseline (without adjusting: β=−1.68, 95% CI −3.22 to −0.14; *P*=.03; with adjusting: β=−1.95, 95% CI −3.20 to −0.70; *P*=.003) and having visited the website or used the resources before follow-up (without adjusting: β=−1.48, 95% CI −2.78 to −0.17; *P*=.03; with adjusting: β=−1.08, 95% CI −2.05 to –0.10; *P*=.03).

Regarding the main exploratory outcomes (Table 3) for non-LGBTQ+, for pretest-posttest change in PHQ-2 score, having visited the website or used the resources before follow-up had a trend as predictor of reduction in depression (β=−0.046, 95% CI −0.93 to 0.00; P=.05), which was not significant when adjusting for nonresponse (β=−0.37, 95% CI −0.84 to 0.10; P=.12). For pretest-posttest change in GAD-2 score, this website use variable was not in the final models given that it was not significant in the initial models (t_{142}=−0.56; *P*=.58), but rather younger age, lower education, and COVID-19 stressors were included as significant predictors (Table 3).

Discussion

Principal Findings and Comparison With Prior Work

This exploratory subanalysis focused on the experiences of LGBTQ+ adults (aged ≥18 years) in California and in comparison to non-LGBTQ+ adults who completed baseline and 4- to 6-week follow-up surveys after visiting a free mental well-being website during the COVID-19 pandemic.

LGBTQ+ participants exhibited greater baseline depression and anxiety than non-LGBTQ+ participants, consistent with previous literature on long-term LGBTQ+ mental health inequities [1-11] and COVID-19–associated mental health inequities [12-18,65]. In 3-way comparisons, LGBTQ+ participants had greater

XSL•FO
RenderX

baseline depression than female heterosexual participants and greater anxiety than both female and male heterosexual participants. LGBTQ+ participants reported more COVID-19 stressors and higher use of behavioral health services than non-LGBTQ+ participants at baseline and follow-up. Collectively, these findings support those of previous studies on minority stress among LGBTQ+ populations, suggesting that mental health inequities stem from social stressors such as discrimination and their impacts on internalized expectations, self-efficacy, and self-image [6,7]. To more fully understand this, we explored differences in individual COVID-19 stressors, finding that LGBTQ+ participants exhibited a higher number of stressors than non-LGBTQ+ participants in 4 areas at baseline (depression, increased alcohol use, change in sexual behavior, and increased financial support from others) and in 7 areas at follow-up (the first 3 mentioned previously plus being diagnosed with COVID-19, fear of giving it to someone, frustration or boredom, and loneliness). Such findings mirror those of studies illustrating disproportionate burden and may suggest specific areas of support needed for LGBTQ+ communities [20-22].

Regarding the main analyses, we found that LGBTQ+ participants had a greater pretest-posttest decline in depression and anxiety than non-LGBTQ+ participants, and in 3-way comparisons for the sensitivity analysis (Multimedia Appendix 1), LGBTQ+ participants had a greater reduction in anxiety than female heterosexual participants. Before follow-up, LGBTQ+ participants had a higher behavioral health service use than non-LGBTQ+ participants, which was significant when imputing for missing data and for LGBTQ+ participants compared to female heterosexual participants in 3-way comparisons. This suggests a greater reduction in anxiety over time for LGBTQ+ participants after initially presenting with higher anxiety levels while having a higher use of behavioral health services between baseline and follow-up.

To understand potential explanatory mechanisms, we used sequential regression analyses to find the best predictors of follow-up website use, depression and anxiety at follow-up, and changes over time in depression and anxiety. For LGBTQ+ participants, predictors of visiting the website or using website resources were depression or anxiety before baseline (which was significant when imputing for missing data) and total COVID-19–related behavior changes, suggesting that persons with higher need were more likely to visit the website or use the resources. Regarding primary analyses of predictors of pretest-posttest reduction in depression (PHQ-2), the main predictor was visiting the website before follow-up, which was significant in the main analysis but borderline significant when imputing for missing data. For reduction in anxiety (GAD-2), website use and Hispanic or Latinx ethnicity were significant predictors.

For non-LGBTQ+ participants, the main predictor of website use was comfort using the website, and total website use and engagement was associated with follow-up depression and anxiety. For pretest-posttest change in need, having visited the website was not significant at $P<.05$ for reduction in PHQ-2 scores, including when adjusting for missing data, and website use did not predict reduction in GAD-2 scores. Similarly, in the sensitivity analyses for female and male heterosexual participants separately (Multimedia Appendix 1), website use did not predict reduction in depression or anxiety. Thus, the association between reduction in anxiety and website use was primarily for LGBTQ+ adults.

Limitations

This evaluation focused on participants invited via emails from partnering agencies that participated in developing the website, which could mean that participants might feel more positive about the website. It is possible that those who agreed to participate may also have had higher digital (computer or internet) access or familiarity than others.

LGBTQ+ participants represented approximately 20% of our sample (64/315, 20.3%), which fits with data on US younger adult populations (millennials, born between 1981 and 1996 and younger) [66]. While consistent with general populations, the overall sample size of LGBTQ+ participants (64/315, 20.3%; 37/64, 58% with follow-up survey data) was modest, and most participants identified as female and heterosexual (225/315, 71.4% and 236/315, 74.9%, respectively). Three-way comparisons among LGBTQ+, male heterosexual, and female heterosexual participants were included for sensitivity analyses in Multimedia Appendix 1. The sample in this study was mainly binary in its distribution of gender identities, with only 4.8% (15/315) of the respondents identifying as gender diverse (including genderqueer, questioning, transgender men, transgender women, or other gender). Among millennials, approximately 12% identify as gender diverse (including transgender) [66]. Our gender identity data may have limitations, consistent with studies showing diversity in how transgender and nonbinary adults describe their gender identities [67] as some may not disclose due to concerns about privacy, discrimination, or other factors [48,50].

Furthermore, to capture LGBTQ+ identities, we used 2 measures: sexual orientation and gender identity, consistent with the NIH 2022 recommendations [48]. With partner input, we used one measure of gender identity, not including sex assigned at birth, and did not capture multiple dimensions of sexual orientation recommended by others (eg, romantic attraction, sexual attraction, and sexual behavior) [47]. For the analyses, we excluded participants who did not provide information on gender identity and sexual orientation—some of whom may have been LGBTQ+. However, all gender identity and sexual orientation survey categories assessed had at least one representative in the follow-up sample. For the main regression analyses exploring predictors of outcomes, the subsample size of LGBTQ+ individuals who had pre- and posttest data was modest (n=30-34), and while key findings about the effects of website use were statistically significant, it will be important to conduct larger studies in the future to confirm such effects in this context.

Other limitations include that the study used baseline and follow-up surveys without intervention comparison groups or randomization; thus, predictors of outcomes may not clearly represent causation. For example, the association between greater website use and reduced anxiety before follow-up could reflect the effects of website use or suggest that those with declining anxiety were more likely to use the website. As a

XSL•FO
RenderX

formative and exploratory study, we did not formally adjust for multiple comparisons, but for the hypothesized greater depression and anxiety and higher number of COVID-19 stressors for LGBTQ+ participants than for non-LGBTQ+ participants at baseline, significance ranged from P=.005 to P<.001, so it was significant with multiple comparisons. Regarding the association between website use and decline over time in mean depression and anxiety for LGBTQ+ participants, significance ranged from P=.04 to P=.09 for depression and P=.03 for anxiety; thus, adjusting for 2 comparisons, the results are considered exploratory to inform future research.

Public Health Implications

Overall, our findings showed that LGBTQ+ participants in our sample of community-recruited adults who visited a public well-being website and provided survey data had greater depression and anxiety and more COVID-19 stressors than non-LGBTQ+ participants. Over time, LGBTQ+ participants exhibited a greater reduction in anxiety than non-LGBTQ+ participants and, separately, than female heterosexual and male heterosexual participants. Specific factors such as COVID-19 stressors (eg, loss and financial strain) may be associated with increased baseline depression and anxiety in this population. Knowing specific risk factors and stressors may inform efforts to mobilize relevant resources to LGBTQ+ populations during pandemics. Consistent with this hypothesis, LGBTQ+ participants had a more consistent prediction of reduced anxiety over time associated with website use, which was not a predictor for non-LGBTQ+ participants. These findings, though exploratory, suggest that digital resources created with community partner input [38] and consistent with community-partnered participatory research, the technology acceptance model, and the vulnerable populations framework [38,41,43] may be helpful when tailored to LGBTQ+ populations—or potentially other communities facing great stressors or mental health symptoms. These are important areas for further research that may inform practice and policy strategies on digital interventions to support well-being in LGBTQ+ during crises such as COVID-19.

Acknowledgments

Melissa Brymer, PsyD, PhD, was the University of California, Los Angeles Together for Wellness/Juntos por Nuestro Bienestar colead and youth and young adult lead for website and evaluation. Sergio Aguilar-Gaxiola, MD, PhD, and Elliott Housman-Turrubiate, MA (University of California, Davis Health, Center for Reducing Health Disparities), led outreach to and engagement with Latino, Latina, and Latinx and farmworker communities and with community-based organizations in northern California. For Chorus Innovations, Armen Arevian, MD, PhD; Minhxuan Tran, MPH; Gail DiRaimondo, licensed clinical social worker; and Brian Lee (now consultant to UCLA) were leads for digital technology and coordinated stakeholder input for website development. Community-based organization and community partner leads for website and evaluation design input included Felica Jones (Healthy African American Families II); Melissa Hannah, MS (United Parents); Hang Nguyen (Boat People SOS); Lupita Rodriquez, MS, and Debra Oto-Kent, MPH (Health Education Council); Melinda Cordero-Barzaga (Vision y Compromiso); and Steven Kite (National Alliance on Mental Illness California). Funding was provided by the California Health Care Foundation (grant G-31202; 11/1/2020-9/30/2022), “CalHOPE Phase 1 Evaluation of Website Use” (for COVID-19 response; principal investigator [PI]: KW); a Federal Emergency Management Agency and Substance Abuse and Mental Health Services Administration grant that was managed by the California Mental Health Services Authority through the California Department of Health Care Services for COVID-19 crisis intervention and prevention (contract 700-FEMA-2021-UCLA-A2; 11/9/2020-6/30/2022), “CalHOPE Website Expansion” (PI: KW); a Federal Emergency Management Agency and Substance Abuse and Mental Health Services Administration grant that was managed by the California Mental Health Services Authority through the California Department of Health Care Services (contract 700-FEMA-2021-UCLA-A3; 2/10/2022-6/30/2022), “California Mental Health Equity and Resilience in Information Technology” (PI: KW); and the California Department of Health Care Services (contract 21-10458:A2; 5/11/2022-6/30/2024), “Together for Wellness/Juntos por Nuestro Bienestar (T4W/Juntos) Youth and Quality Improvement Expansion” (co-PIs: KW and Melissa Brymer).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding or statistician authors on reasonable request.

Authors' Contributions

NR and KW drafted the manuscript. LZ conducted the statistical analyses and obtained the data for the tables and Multimedia Appendix 1. SJ and MN provided statistical consultation and edits to the manuscript. DC and BR facilitated community-based organization and community partner input into the design, measures, and manuscript. AM coordinated community input and author collaboration. All authors reviewed the final manuscript.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
Sensitivity analyses (3-category comparisons for LGTBQ+, female heterosexual and male heterosexual).
[DOC File , 236 KB - formative_v9i1e67082_app1.doc]

References

1. Marshall A. Suicide prevention interventions for sexual and gender minority youth: an unmet need. Yale J Biol Med 2016 Jun;89(2):205-213 [FREE Full text] [Medline: 27354846]
2. Reisner SL, Greytak EA, Parsons JT, Ybarra ML. Gender minority social stress in adolescence: disparities in adolescent bullying and substance use by gender identity. J Sex Res 2015;52(3):243-256 [FREE Full text] [doi: 10.1080/00224499.2014.886321] [Medline: 24742006]
3. Valentine SE, Shipherd JC. A systematic review of social stress and mental health among transgender and gender non-conforming people in the United States. Clin Psychol Rev 2018 Dec;66:24-38 [FREE Full text] [doi: 10.1016/j.cpr.2018.03.003] [Medline: 29627104]
4. Rice CE, Vasilenko SA, Fish JN, Lanza ST. Sexual minority health disparities: an examination of age-related trends across adulthood in a national cross-sectional sample. Ann Epidemiol 2019 Mar;31:20-25 [FREE Full text] [doi: 10.1016/j.annepidem.2019.01.001] [Medline: 30792064]
5. Ramos N, Burgess A, Ollen E. The current status of sociopolitical and legal issues faced by lesbian, gay, bisexual, transgender, queer, and questioning youth. Adolesc Psychiatry (Hilversum) 2023 Nov;12(3):180-195 [FREE Full text] [doi: 10.2174/2210676611666211105120645] [Medline: 37064428]
6. Meyer IH. Prejudice, social stress, and mental health in lesbian, gay, and bisexual populations: conceptual issues and research evidence. Psychol Bull 2003 Sep;129(5):674-697 [FREE Full text] [doi: 10.1037/0033-2909.129.5.674] [Medline: 12956539]
7. Hendricks ML, Testa RJ. A conceptual framework for clinical work with transgender and gender nonconforming clients: an adaptation of the Minority Stress Model. Prof Psychol Res Pract 2012 Oct;43(5):460-467. [doi: 10.1037/a0029597]
8. Institute of Medicine, Committee on Lesbian, Gay, Bisexual, and Transgender Health Issues and Research Gaps and Opportunities, Board on the Health of Select Populations. The Health of Lesbian, Gay, Bisexual, and Transgender People: Building a Foundation for Better Understanding. Washington, DC: National Academies Press; 2011.
9. Coker TR, Austin SB, Schuster MA. The health and health care of lesbian, gay, and bisexual adolescents. Annu Rev Public Health 2010 Mar 01;31(1):457-477. [doi: 10.1146/annurev.publhealth.012809.103636] [Medline: 20070195]
10. Almeida J, Johnson RM, Corliss HL, Molnar BE, Azrael D. Emotional distress among LGBT youth: the influence of perceived discrimination based on sexual orientation. J Youth Adolesc 2009 Aug 24;38(7):1001-1014 [FREE Full text] [doi: 10.1007/s10964-009-9397-9] [Medline: 19636742]
11. Lyons HZ, Bieschke KJ, Dendy AK, Worthington RL, Georgemiller R. Psychologists' competence to treat lesbian, gay and bisexual clients: state of the field and strategies for improvement. Prof Psychol Res Pract 2010 Oct;41(5):424-434. [doi: 10.1037/a0021121]
12. Manning WD, Kamp Dush CM. COVID-19 stress and sexual identities. Socius 2022;8:23 [FREE Full text] [doi: 10.1177/23780231221105376] [Medline: 36467105]
13. Dal Santo T, Sun Y, Wu Y, He C, Wang Y, Jiang X, et al. Systematic review of mental health symptom changes by sex or gender in early-COVID-19 compared to pre-pandemic. Sci Rep 2022 Jul 06;12(1):11417 [FREE Full text] [doi: 10.1038/s41598-022-14746-1] [Medline: 35794116]
14. Grady A, Stinchcombe A. The impact of COVID-19 on the mental health of older sexual minority Canadians in the CLSA. BMC Geriatr 2023 Dec 07;23(1):816 [FREE Full text] [doi: 10.1186/s12877-023-04513-w] [Medline: 38062412]
15. Coulter RS, Sumetsky N, Marsh LN, Hoffman BL, Burke JG, Friedman MR, et al. How has the COVID-19 pandemic affected mental health inequities for sexual and gender minority adults? Ann LGBTQ Public Popul Health 2023 May 31;5(1):1-11. [doi: 10.1891/lgbtq-2021-0043]
16. Sharma AJ, Subramanyam MA. A cross-sectional study of psychological wellbeing of Indian adults during the COVID-19 lockdown: different strokes for different folks. PLoS One 2020 Sep 3;15(9):e0238761 [FREE Full text] [doi: 10.1371/journal.pone.0238761] [Medline: 32881946]
17. Suen YT, Chan RC, Wong EM. Effects of general and sexual minority-specific COVID-19-related stressors on the mental health of lesbian, gay, and bisexual people in Hong Kong. Psychiatry Res 2020 Oct;292:113365 [FREE Full text] [doi: 10.1016/j.psychres.2020.113365] [Medline: 32862107]
18. Freibott CF, Stein MD, Lipson SK. The influence of race, sexual orientation and gender identity on mental health, substance use, and academic persistence during the COVID-19 pandemic: A cross-sectional study from a national sample of college students in the healthy minds study. Drug Alcohol Depend Rep 2022 Jun;3:100060 [FREE Full text] [doi: 10.1016/j.dadr.2022.100060] [Medline: 35529429]
19. Vincke J, van Heeringen K. Confidant support and the mental wellbeing of lesbian and gay young adults: a longitudinal analysis. J Community Appl Soc Psychol 2002 May 02;12(3):181-193. [doi: 10.1002/casp.671]
20. Budge SL, Adelson JL, Howard KA. Anxiety and depression in transgender individuals: the roles of transition status, loss, social support, and coping. J Consult Clin Psychol 2013 Jun;81(3):545-557. [doi: 10.1037/a0031774] [Medline: 23398495]

XSL•FO
RenderX

21. Budge SL, Rossman HK, Howard KA. Coping and psychological distress among genderqueer individuals: the moderating effect of social support. J LGBT Issues Couns 2014 Mar 03;8(1):95-117. [doi: 10.1080/15538605.2014.853641]
22. Pflum SR, Testa RJ, Balsam KF, Goldblum PB, Bongar B. Social support, trans community connectedness, and mental health symptoms among transgender and gender nonconforming adults. Psychol Sex Orientat Gend Divers 2015 Sep;2(3):281-286. [doi: 10.1037/sgd0000122]
23. Pharr JR, Terry E, Wade A, Haboush-Deloye A, Marquez E, Nevada Minority Health And Equity Coalition. Impact of COVID-19 on sexual and gender minority communities: focus group discussions. Int J Environ Res Public Health 2022 Dec 21;20(1):93 [FREE Full text] [doi: 10.3390/ijerph20010050] [Medline: 36612372]
24. Flor LS, Friedman J, Spencer CN, Cagney J, Arrieta A, Herbert ME, et al. Quantifying the effects of the COVID-19 pandemic on gender equality on health, social, and economic indicators: a comprehensive review of data from March, 2020, to September, 2021. Lancet 2022 Jun 25;399(10344):2381-2397 [FREE Full text] [doi: 10.1016/S0140-6736(22)00008-3] [Medline: 35247311]
25. Peterson ZD, Vaughan EL, Carver DN. Sexual identity and psychological reactions to COVID-19. Traumatology 2021 Mar;27(1):6-13 [FREE Full text] [doi: 10.1037/trm0000283]
26. Chima C, Shalaby R, Lawal MA, Vuong W, Hrabok M, Gusnowski A, et al. COVID-19 pandemic: influence of gender identity on stress, anxiety, and depression levels in Canada. Trauma Care 2022 Jan 09;2(1):11-22 [FREE Full text] [doi: 10.3390/traumacare2010002]
27. Kidd JD, Jackman KB, Barucco R, Dworkin JD, Dolezal C, Navalta TV, et al. Understanding the impact of the COVID-19 pandemic on the mental health of transgender and gender nonbinary individuals engaged in a longitudinal cohort study. J Homosex 2021 Mar 21;68(4):592-611 [FREE Full text] [doi: 10.1080/00918369.2020.1868185] [Medline: 33502286]
28. Titov N, Staples L, Kayrouz R, Cross S, Karin E, Ryan K, et al. Rapid report: early demand, profiles and concerns of mental health users during the coronavirus (COVID-19) pandemic. Internet Interv 2020 Sep;21:100327 [FREE Full text] [doi: 10.1016/j.invent.2020.100327] [Medline: 32537424]
29. Piper S, Davenport TA, LaMonica H, Ottavio A, Iorfino F, Cheng VW, et al. Implementing a digital health model of care in Australian youth mental health services: protocol for impact evaluation. BMC Health Serv Res 2021 May 12;21(1):452 [FREE Full text] [doi: 10.1186/s12913-021-06394-4] [Medline: 33980229]
30. Taylor CB, Ruzek JI, Fitzsimmons-Craft EE, Sadeh-Sharvit S, Topooco N, Weissman RS, et al. Using digital technology to reduce the prevalence of mental health disorders in populations: time for a new approach. J Med Internet Res 2020 Jul 24;22(7):e17493 [FREE Full text] [doi: 10.2196/17493] [Medline: 32706665]
31. Li J. Digital technologies for mental health improvements in the COVID-19 pandemic: a scoping review. BMC Public Health 2023 Mar 01;23(1):413 [FREE Full text] [doi: 10.1186/s12889-023-15302-w] [Medline: 36859184]
32. Perri M, Schmidt RA, Guta A, Kaminski N, Rudzinski K, Strike C. COVID-19 and the opportunity for gender-responsive virtual and remote substance use treatment and harm reduction services. Int J Drug Policy 2022 Oct;108:103815 [FREE Full text] [doi: 10.1016/j.drugpo.2022.103815] [Medline: 35969920]
33. Huguet A, Rao S, McGrath PJ, Wozney L, Wheaton M, Conrod J, et al. A systematic review of cognitive behavioral therapy and behavioral activation apps for depression. PLoS One 2016;11(5):e0154248 [FREE Full text] [doi: 10.1371/journal.pone.0154248] [Medline: 27135410]
34. Figueroa CA, Aguilera A. The need for a mental health technology revolution in the COVID-19 pandemic. Front Psychiatry 2020 Jun 3;11:523 [FREE Full text] [doi: 10.3389/fpsyt.2020.00523] [Medline: 32581891]
35. Schueller SM, Hunter JF, Figueroa C, Aguilera A. Use of digital mental health for marginalized and underserved populations. Curr Treat Options Psych 2019 Jul 5;6(3):243-255. [doi: 10.1007/s40501-019-00181-z]
36. Burns MN, Montague E, Mohr DC. Initial design of culturally informed behavioral intervention technologies: developing an mHealth intervention for young sexual minority men with generalized anxiety disorder and major depression. J Med Internet Res 2013 Dec 05;15(12):e271 [FREE Full text] [doi: 10.2196/jmir.2826] [Medline: 24311444]
37. Arevian AC, O'Hora J, Jones F, Mango J, Jones L, Williams PG, et al. Participatory technology development to enhance community resilience. Ethn Dis 2018;28(Suppl 2):493-502 [FREE Full text] [doi: 10.18865/ed.28.S2.493] [Medline: 30202203]
38. Jones L, Wells K. Strategies for academic and clinician engagement in community-participatory partnered research. JAMA 2007 Jan 24;297(4):407-410. [doi: 10.1001/jama.297.4.407] [Medline: 17244838]
39. Goodsmith N, Moore EM, Siddiq H, Barceló NE, Ulloa-Flores E, Loera G, et al. Community-partnered development of a digital mental health resource website to support diverse communities during the COVID-19 pandemic. J Health Care Poor Underserved 2022;33(1):506-516 [FREE Full text] [doi: 10.1353/hpu.2022.0039] [Medline: 35153238]
40. Pavlov PA. Consumer acceptance of electronic commerce: integrating trust and risk with the technology acceptance model. Int J Electron Commer 2014 Dec 23;7(3):101-134. [doi: 10.1080/10864415.2003.11044275]
41. Gelberg L, Andersen RM, Leake BD. The behavioral model for vulnerable populations: application to medical care use and outcomes for homeless people. Health Serv Res 2000 Feb;34(6):1273-1302 [FREE Full text] [Medline: 10654830]
42. Wells K, Thames AD, Young AS, Zhang L, Heilemann MV, Romero DF, Together for Wellness/Juntos Collaborators and Writing Group. Engagement, use, and impact of digital mental health resources for diverse populations in COVID-19:

XSL•FO
RenderX

community-partnered evaluation. JMIR Form Res 2022 Dec 07;6(12):e42031 [FREE Full text] [doi: 10.2196/42031] [Medline: 36346902]
43. Venkatesh V, Davis FD. A theoretical extension of the technology acceptance model: four longitudinal field studies. Manag Sci 2000 Feb;46(2):186-204. [doi: 10.1287/mnsc.46.2.186.11926]
44. Kroenke K, Spitzer RL, Williams JB, Monahan PO, Löwe B. Anxiety disorders in primary care: prevalence, impairment, comorbidity, and detection. Ann Intern Med 2007 Mar 06;146(5):317-325. [doi: 10.7326/0003-4819-146-5-200703060-00004] [Medline: 17339617]
45. Kroenke K, Spitzer RL, Williams JB. The Patient Health Questionnaire-2: validity of a two-item depression screener. Med Care 2003 Nov;41(11):1284-1292. [doi: 10.1097/01.MLR.0000093487.78664.3C] [Medline: 14583691]
46. Mayer KH, Bradford JB, Makadon HJ, Stall R, Goldhammer H, Landers S. Sexual and gender minority health: what we know and what needs to be done. Am J Public Health 2008 Jun;98(6):989-995. [doi: 10.2105/AJPH.2007.127811] [Medline: 18445789]
47. Patterson JG, Jabson JM, Bowen DJ. Measuring sexual and gender minority populations in health surveillance. LGBT Health 2017 Apr;4(2):82-105 [FREE Full text] [doi: 10.1089/lgbt.2016.0026] [Medline: 28287877]
48. National Academies of Sciences, Engineering, and Medicine. Measuring Sex, Gender Identity, and Sexual Orientation. Washington, DC: The National Academies Press; 2022.
49. Raveenthiran V. Neonatal sex assignment in disorders of sex development: a philosophical introspection. J Neonatal Surg 2017;6(3):58 [FREE Full text] [doi: 10.21699/jns.v6i3.604] [Medline: 28920018]
50. Puckett JA, Brown NC, Dunn T, Mustanski B, Newcomb ME. Perspectives from transgender and gender diverse people on how to ask about gender. LGBT Health 2020 Sep 01;7(6):305-311 [FREE Full text] [doi: 10.1089/lgbt.2019.0295] [Medline: 32598211]
51. Alpert AB, Ruddick R, Manzano C. Rethinking sex-assigned-at-birth questions. BMJ 2021 May 24;373:n1261. [doi: 10.1136/bmj.n1261] [Medline: 34031025]
52. Lopez-Maya E, Olmstead R, Irwin MR. Mindfulness meditation and improvement in depressive symptoms among Spanish- and English speaking adults: a randomized, controlled, comparative efficacy trial. PLoS One 2019 Jul 5;14(7):e0219425 [FREE Full text] [doi: 10.1371/journal.pone.0219425] [Medline: 31276540]
53. Bégin C, Berthod J, Martinez LZ, Truchon M. Use of mobile apps and online programs of mindfulness and self-compassion training in workers: a scoping review. J Technol Behav Sci 2022;7(4):477-515 [FREE Full text] [doi: 10.1007/s41347-022-00267-1] [Medline: 36091081]
54. Champion L, Economides M, Chandler C. The efficacy of a brief app-based mindfulness intervention on psychosocial outcomes in healthy adults: a pilot randomised controlled trial. PLoS One 2018 Dec 31;13(12):e0209482 [FREE Full text] [doi: 10.1371/journal.pone.0209482] [Medline: 30596696]
55. Galla BM, O'Reilly GA, Kitil MJ, Smalley SL, Black DS. Community-based mindfulness program for disease prevention and health promotion: targeting stress reduction. Am J Health Promot 2015 Sep 01;30(1):36-41. [doi: 10.4278/ajhp.131107-quan-567]
56. Arevian AC, Jones F, Tang L, Sherbourne CD, Jones L, Miranda J. Depression remission from community coalitions versus individual program support for services: findings from community partners in care, Los Angeles, California, 2010–2016. Am J Public Health 2019 Jun;109(S3):S205-S213. [doi: 10.2105/ajph.2019.305082]
57. 2019-2020 stakeholder surveys for diverse communities and families. National Alliance on Mental Illness California. URL: https://namica.org/blog/2019-2020-stakeholder-surveys-for-diverse-communities-families-%EF%BB%BF/ [accessed 2021-06-02]
58. Elwood WN, Irvin VL, Liu B, Lee R, Breen N. Health-related influences of extending marital benefits to same-sex couples: results from the California Health Interview Survey. Fam Relat 2020 Oct 10;69(5):934-943. [doi: 10.1111/fare.12507]
59. Jans M, Viana J, Grant D, Cochran SD, Lee AC, Ponce NA. Trends in sexual orientation missing data over a decade of the California Health Interview Survey. Am J Public Health 2015 May;105(5):e43-e50. [doi: 10.2105/ajph.2014.302514]
60. Tambling RR, Russell BS, Park CL, Fendrich M, Hutchinson M, Horton AL, et al. Measuring cumulative stressfulness: psychometric properties of the COVID-19 stressors scale. Health Educ Behav 2021 Feb 11;48(1):20-28. [doi: 10.1177/1090198120979912] [Medline: 33307818]
61. Harkness A, Behar-Zusman V, Safren SA. Understanding the impact of COVID-19 on Latino sexual minority men in a US HIV hot spot. AIDS Behav 2020 Jul 16;24(7):2017-2023 [FREE Full text] [doi: 10.1007/s10461-020-02862-w] [Medline: 32300989]
62. Fisher CB, Tao X, Yip T. The effects of COVID-19 victimization distress and racial bias on mental health among AIAN, Asian, Black, and Latinx young adults. Cultur Divers Ethnic Minor Psychol 2023 Apr;29(2):119-131. [doi: 10.1037/cdp0000539] [Medline: 35389692]
63. Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. Stat Methods Med Res 2013 Jun 10;22(3):278-295. [doi: 10.1177/0962280210395740] [Medline: 21220355]
64. Li L, Shen C, Li X, Robins JM. On weighting approaches for missing data. Stat Methods Med Res 2013 Feb;22(1):14-30 [FREE Full text] [doi: 10.1177/0962280211403597] [Medline: 21705435]

XSL•FO
RenderX

65. Kamal K, Li JJ, Hahm HC, Liu CH. Psychiatric impacts of the COVID-19 global pandemic on U.S. sexual and gender minority young adults. Psychiatry Res 2021 May;299:113855 [FREE Full text] [doi: 10.1016/j.psychres.2021.113855] [Medline: 33721788]
66. Accelerating acceptance: a Harris poll survey of Americans' acceptance of LGBTQ people. GLAAD. URL: https://www.glaad.org/files/aa/2017_GLAAD_Accelerating_Acceptance.pdf [accessed 2025-05-29]
67. Wilson DB, Meyer IH. Nonbinary LGBTQ adults in the United States. UCLA School of Law. URL: https://williamsinstitute.law.ucla.edu/publications/nonbinary-lgbtq-adults-us/ [accessed 2025-05-28]

Abbreviations

GAD-2: 2-item Generalized Anxiety Disorder scale
LGBTQ+: lesbian, gay, bisexual, transgender, queer, and questioning
NIH: National Institutes of Health
OR: odds ratio
PHQ-2: 2-item Patient Health Questionnaire
T4W/Juntos: Together for Wellness/Juntos por Nuestro Bienestar

Edited by A Mavragani; submitted 01.10.24; peer-reviewed by K Phankitnirundorn; comments to author 11.03.25; revised version received 24.03.25; accepted 24.04.25; published 11.07.25.

Please cite as:
Ramos N, Jones S, Zhang L, Nuño M, Ramsey B, Ceseña D, Mireles A, Wells K
Exploring the Impact of Online Mental Health Resources During the COVID-19 Pandemic on Lesbian, Gay, Bisexual, Transgender, Queer, and Questioning Adults Compared to Heterosexual Adults: Pretest-Posttest Survey Analyses
JMIR Form Res 2025;9:e67082
URL: https://formative.jmir.org/2025/1/e67082
doi:10.2196/67082
PMID:

©Natalia Ramos, Skylar Jones, Lily Zhang, Miriam Nuño, Benita Ramsey, Dannie Ceseña, Alyssa Mireles, Kenneth Wells. Originally published in JMIR Formative Research (https://formative.jmir.org), 11.07.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Hand Motion Control Ability Between Young and Older Adults: Comparative Study

Jung-Soon Kim[1*], MS; Hui-Jun Kim[2*], BS; Minju Kim[2], MS; Sung-Hee Kim[3*], PhD

[1]Department of Artificial Intelligence, Dong-eui University, Busan, Republic of Korea
[2]Department of IT Convergence, Dong-Eui University, Busan, Republic of Korea
[3]Department of Industrial ICT Engineering, Dong-Eui University, Busan, Republic of Korea
[*]these authors contributed equally

Corresponding Author:
Sung-Hee Kim, PhD
Department of Industrial ICT Engineering
Dong-Eui University
Eomgwang-ro 176
Busan, 47340
Republic of Korea
Phone: 82 51 890 2366
Email: sh.kim@deu.ac.kr

Abstract

Background: Age-related differences in motor skills have been extensively studied, with growing interest in using behavioral data for cognitive assessment. Compared to traditional tools like the Mini-Mental State Examination or Cognitive Impairment Screening Test, behavior-based methods offer the advantage of shorter testing durations, less learning effects, and continuous data tracking. Hand movements, in particular, provide a practical way to gather motor performance data with fewer spatial constraints. This study aims to explore whether hand rotation movement can effectively distinguish age-related motor skill differences, with future applications potentially extending to cognitive assessments, including early detection of mild cognitive impairment.

Objective: This study investigates whether hand rotation movements can be used to distinguish 2 age groups, young adults (aged 20-29 years) and older adults (aged 65-80 years). We hypothesize that differences in hand motion control ability will exist between the 2 groups. In total, 7 hand motion measurement indicators related to single hand test indicators, time comparison indicators between rotations, and angle comparison indicators between rotations were defined to test this hypothesis, aiming to identify meaningful indicators for older adults experiencing normal aging before conducting experiments on patients with mild cognitive impairment or dementia.

Methods: A total of 68 participants, 39 older adults (aged 65-80 years) and 29 young adults (aged 20-29 years), all capable of normal arm, hand, and finger movements, participated in the experiment. Participants sat facing a webcam and were asked to perform hand rotation movements as quickly and accurately as possible with both hands for 10 seconds. They performed 3 trials with a 30-second break in between. For statistical verification, we set the significance level at .05 and analyzed the data using the generalized estimation equations model to assess the effects of the between-subject factor (age group: younger vs older) and the within-subject factors (hand: left vs right, and trials 1, 2, and 3).

Results: Among the 7 measured indicators, 3 (total rotation count, angle, and time) showed statistically significant differences between age groups. Younger participants performed more rotations (B=5.29, *P*=.002), demonstrated a greater range of motion (B=1334.37, *P*=.007), and completed the task in less time (B=0.99, *P*=.003), indicating age-related differences in upper limb motor function. Trial order also had a significant main effect on rotation count and angle. Trial 1 differed significantly from trials 2 and 3, while no difference was observed between trials 2 and 3, suggesting that trial 1 may reflect a practice effect.

Conclusions: The findings revealed that the older adult group demonstrated statistically significant differences compared to the young adult group in their ability to control hand rotation movements. A learning effect was observed across the 3 trials, suggesting that the first trial should be discarded for use as a stable measurement.

(JMIR Form Res 2025;9:e65224) doi:10.2196/65224

XSL•FO
RenderX

KEYWORDS

youth vs elders; hand gesture; hand motion control; older adults; young adults; hand movements

Introduction

Background

Recently, a growing number of studies have focused on digital biomarkers for the early detection and intervention in preclinical stages of mild cognitive impairment (MCI) and dementia due to their advantages, such as low cost, time efficiency, and high accessibility [1,2]. Traditional detection methods include imaging assessments (eg, positron emission tomography) and biological evaluations (eg, cerebrospinal fluid analysis), but these methods have limitations such as being invasive, inconvenient, time-consuming, expensive, and difficult to access [3]. Furthermore, while clinical and neuropsychological assessments remain key standards, they also have limitations, including long testing intervals to prevent learning effects [4], reliance on self-reporting, and interevaluator variability [5]. In contrast, digital biomarkers are defined as objective and quantifiable physiological and behavioral data collected through digital devices, enabling frequent or continuous evaluation with minimal disruption to daily life. This allows for objective, ecologically valid, and long-term follow-up in preclinical detection processes [6-9]. Developing new digital biomarkers that fulfill these requirements is important.

Research on digital biomarkers based on behavioral data has primarily focused on gait or hand movement data [10]. In particular, hand movement-based studies are gaining attention due to their relative advantages compared to gait, such as requiring less space [11,12], ensuring physical safety for participants (ie, gait poses a risk of falls), and using low-cost data measurement devices. However, previous studies have either measured imitation accuracy of predefined static hand movements using artificial intelligence (AI) [13] or collected dynamic hand movement data similar to daily activities using virtual reality (VR) devices [14,15]. These methods face limitations in overcoming the drawbacks of traditional biomarkers. Challenges include the occurrence of learning effects from repetitive presentation of fixed movements in repeated measurement processes, the high cost of VR devices, and psychological barriers stemming from the use of unfamiliar equipment. As a result, these methods lack sufficient validity verification required for their application as digital biomarkers in cognitive ability assessments.

To address these limitations, we used a hand rotation-based digital biomarker inspired by the correlation between upper limb motor skill capability and cognitive function. Numerous prior studies have demonstrated that cognitive impairment leads to a decline in upper limb motor abilities [16-19]. The hand rotation-based movement reflects hand motion control ability that has minimal learning effects [13] and can be measured by images with an AI-based measurement system to collect quantitative data, thereby minimizing interevaluator variability [20]. To quantitatively measure hand rotation data, several measures can be used, such as total rotation count, total rotation time, and symmetry [21], as well as derived indicators, such as changes in amplitude and variations in performance time. Before conducting experiments on patients with MCI or dementia, we investigated if the hand rotation-based digital biomarker could identify different age groups, based on the understanding that healthy aging induces neurochemical and structural changes in older adults [22], which consequently impacts their motor performance, leading to reduced motor control and coordination [23,24].

Objective

To achieve this, our objective is to validate the proposed hand rotation-based digital biomarker to distinguish differences between a young adult group and an older adult group. We used a mixed design to investigate the effects of age, hand dominance, and repeated trials on motor performance. The significance of this research lies in overcoming the economic and spatial limitations of existing behavior-based digital biomarkers and developing a more accessible digital biomarker. Additionally, this study provides direction and insights for future digital biomarker development.

Methods

Ethical Considerations

Ethical Approval and Participant Recruitment

The study was conducted with the approval of the Dong-eui University institutional review board (DIRB-202307-HR-E-20) to respect and protect the safety, well-being, and rights of the participants.

The selection of study participants was based on a similar study on hand motion measurement [25]. In this experiment, participants were divided into experimental and control groups, and the experiment involved repeatedly tapping finger-to-finger movements. This study evaluates hand movement control abilities, with the experimental group being older adults aged 65-80 years and the control group being young adults aged 20-29 years. Older adult participants were selected based on their cognitive status, ensuring they were not diagnosed with MCI or dementia and could perform daily arm, hand, and finger movements. A 2-sided test was conducted with a significance level of .05 and a power of over 70%. To achieve similar significant results, 31 young adults and 39 older adults were selected for evaluation in each group. Older adult participants were recruited through a public announcement at the Dasarang Cultural Complex Arts Center in Busan, which is a community welfare center operated by the local government. The young adult participants were selected from among the undergraduates at Dong-eui University through a flyer.

Informed Consent and Privacy Protection

Participants were fully informed about the purpose, procedures, and any potential risks or discomforts associated with the study. They were informed that video would be recorded and the data would be analyzed and used only for this study. They were also informed that they could withdraw from the study at any time

XSL•FO
RenderX

without any disadvantage. Written consent was obtained only from those who voluntarily agreed to participate. Participants' names were replaced with anonymized identifiers, and only minimal clinical information, such as age, gender, and physical characteristics, was collected.

The collected information and data were securely stored on a personal portable storage device (USB) and were immediately destroyed if a participant withdrew or upon completion of data analysis. The data were not used for any other purpose.

Rewards

The compensation for the experiment was a dietary supplement worth 20,000 KRW (US $14.72) per person. Additionally, participants were informed that they would receive the compensation even if they did not meet the normal range criteria for the Purdue Pegboard Test (PPT) or if they had to stop in the middle and could not complete the study.

Study Design

This study used a mixed-design method to investigate the effects of age, hand dominance, and repeated trials on motor performance. The experiment included both between-subject and within-subject factors. The between-subject factor was age group, with participants divided into 2 groups: younger adults (ages 20-29 years) and older adults (ages 65-80 years). This design allows for the examination of age-related differences in motor function. The within-subject factors included hand (right vs left) and trial (3 repeated trials). Each participant completed the task using both hands. The task was repeated 3 times (trial 1, trial 2, and trial 3) to assess potential changes in performance due to repetition, learning, or fatigue. Participants were asked to perform a hand rotation task, in which they rotated their hands for a total of 10 seconds. The outcome variables measured are explained in Table 1. These measures were collected across both hands and trials for each participant, providing a comprehensive assessment of motor function. By using a mixed design, this study allowed for the analysis of both between-subject effects (comparing younger and older participants) and within-subject effects (comparing performance across hands and trials), as well as their interactions. The design also enabled us to investigate whether performance improvements across trials, indicative of a learning effect, differed between the age groups.

Table 1. Measurements.

No	Measurements	Explain
1	Total rotation count	Total number of hand rotations in 10 seconds[a]
2	Total rotation time	Cumulative total of each rotation time[a]
3	Total rotation time change[b]	Sum of the amount of time change that occurs while rotating the hand
4	The number of rotation time changes	The number of times the time changes as the hand rotates[a]
5	Total rotation angle	Cumulative total of on rotation time
6	Total rotation angle change[c]	The sum of the angular changes that occur while rotating the hand[a]
7	The number of rotation angle changes	The number of times the angle changes as you rotate your hand[a]

[a]Rotation unit = 1 rotation.

[b]TCA: time of rotation change amount.

[c]ACA: angle of rotation change amount.

Statistical Analysis

Model Specification

Generalized estimating equations (GEE) were used to estimate causal models for panel data or across entire panels, particularly for multivariate variables that deviate from a normal distribution by applying the generalized linear model. GEE is a technique capable of handling repeated measures and time-series data, which are challenging for generalized linear models [26]. All statistical analyses were conducted using Python (version 3.10) on Google Colab (Ubuntu; version 22.04) with the following libraries: pandas (version 2.2.2; developed by the pandas community, under the NumFOCUS umbrella) for data manipulation, statsmodels (version 0.14.4; developed by the statsmodels community of contributors) for statistical modeling to fit GEE models, scipy (version 1.13.1; developed by the SciPy community, under the NumFOCUS umbrella) for additional statistical functions, and pingouin (version 0.5; developed by the Pingouin project team) for post hoc pairwise comparisons between trials to examine potential learning effects. By leveraging GEE, we effectively controlled for within-subject correlations inherent in repeated measures designs, and post hoc pairwise comparisons were performed to identify significant differences across trials. The application of Bonferroni adjustments ensured that the results remained robust against type I error due to multiple comparisons.

Data Exclusion

Before the main experiment, the PPT [27] was conducted to verify that there were no abnormalities in the arm, hand, and finger movements of the participants. Participants falling out of the range of the benchmark in Table 2 were excluded from the hand movement control ability experiment. To note, while recruiting, we informed participants that they should not have

XSL•FO
RenderX

been diagnosed with MCI. However, we did not administer an additional test, and participation was voluntary.

Table 2. Purdue Pegboard Test average scores by age group (n=158).

Age	Male, mean (SD)	Female, mean (SD)
20-29 years old	12.8 (2.9)	13.6 (0.9)
60 years and older	7.9 (1.7)	8.2 (1.8)

Overview of Hand Movement Measurement System

Defining Hand Rotation Behavior

Hand rotation is defined as having a rotation angle of 0° when the palm is facing forward toward the webcam system and a rotation angle of 180° when the back of the hand is facing forward. Hand rotation consists of 3 stages based on the change in hand rotation angle: increase, maintain, and decrease. One full rotation is defined as rotating 180° in one direction and then 180° in the opposite direction (Figure 1). The total rotation angle is 360°, and the video consists of 30 frames, with each frame lasting approximately 0.03 seconds.

Figure 1. Example of hand rotation for 360° over 30 frames. Hand rotation is defined from 0° (palm facing forward) to 180° (back of hand facing forward). Hand rotation consists of 3 phases: increasing, maintaining, and decreasing. One full rotation includes 180° in each direction (total 360°), measured over 30 frames (~0.03 s per frame).

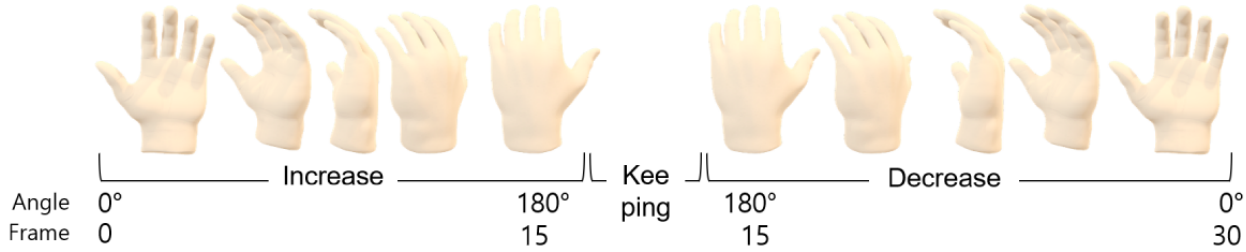

Figure 2. Experiment setting. This figure shows the experimental setup for the comparative study on hand movement control abilities between younger and older adults. In the experiment, participants sat facing the webcam, and upon the appearance of a "recording" sign on the screen, they were instructed to perform hand rotation movements with both hands as quickly and accurately as possible for 10 seconds. This procedure was repeated 3 times. The experiment for the older adult group was conducted at a welfare facility operated by the local government (Dasarang Cultural Complex Arts Center in Busan), while the experiment for the younger adult group was conducted with university students enrolled at Dong-eui University.

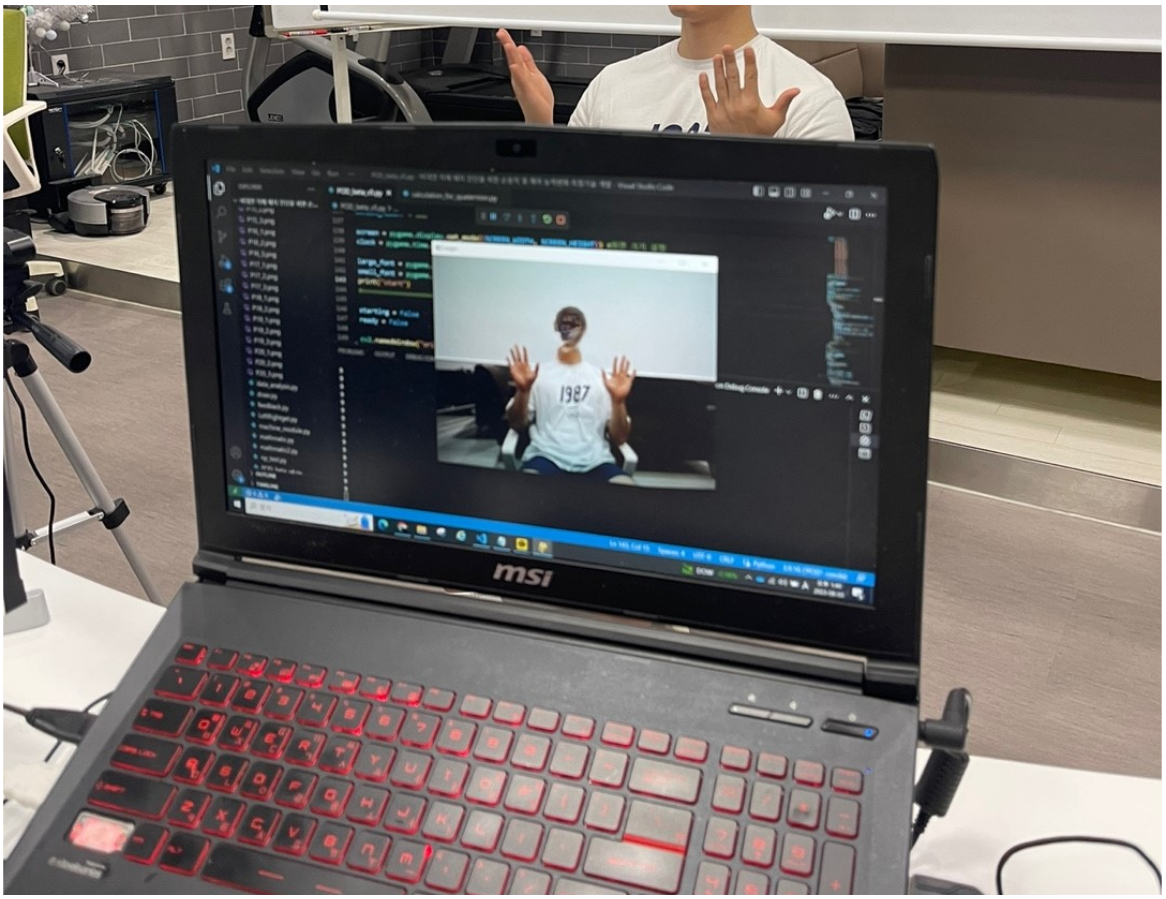

XSL•FO
RenderX

Measurement System

The main experiment used an AI-based video processing hand rotation measurement technology [20]. This system involves the participant sitting in front of a webcam and performing hand rotation movements in real time for 10 seconds (Figure 2). The video data were recorded, and the system recognized the hand skeleton to collect and analyze the data. Using real-time video information captured by the webcam, the system estimated the 3D coordinates (x, y, z) of 21 hand joints (landmarks), as shown in Figure 3, using the MediaPipe Hands model Application Programming Interface [28]. The extracted coordinates were used to calculate the position vector through vector operations between the tip of the thumb (landmark 4) and the wrist (landmark 0). The rotation angle of the hand is determined by using the change in the position vector's movement along the Z-axis over time (frames). The algorithm for calculating the hand's rotation angle involves inverse calculating the quaternion [29]. To measure 1 full rotation of the hand, the reversal of the movement direction of the aforementioned position vector is used. Since the sign (negative or positive) of the quaternion's slope (5 frames) is the same as the vector movement direction, 1 full rotation of the hand can be measured according to the sign reversal of the slope value. Ultimately, the landmark coordinates of the fingers, the video file captured by the webcam, and the 7 measurements explained in Table 1 (ie, total rotation count, total rotation time, total rotation time change, the number of rotation time changes, total rotation angle, total rotation angle change, and the number of rotation angle changes) were automatically saved to the computer.

Figure 3. Landmarks based on hand shape. Visualization of the 21 hand landmarks.

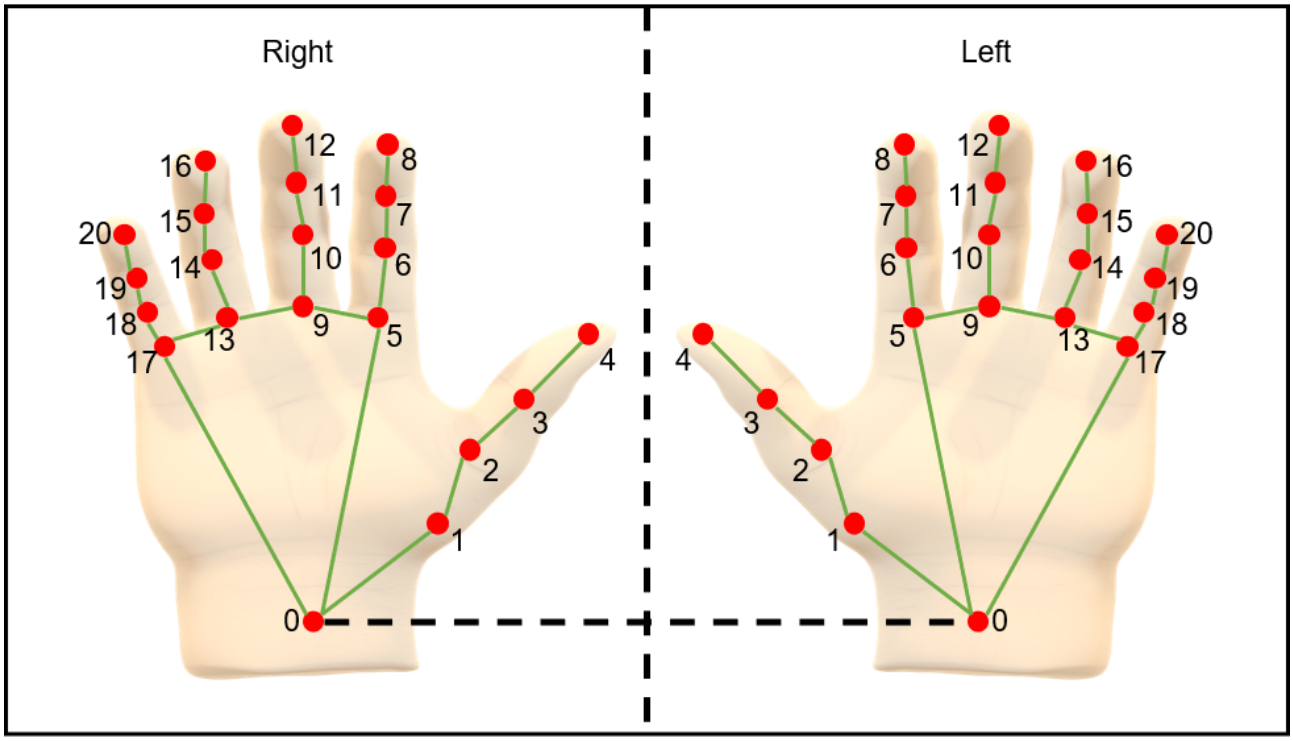

Experimental Procedure

This study conducted a comparative experiment on hand motion control ability between young adults and older adults. Prior to the main experiment, (1) a recruitment notice was posted to recruit study participants. On the day of the experiment, when the experimenter visited the experiment site, (2) the experimenter thoroughly explained the experiment to the participants using an explanation document and received a consent form, and (3) written consent for the provision of personal information was obtained from the participants. Then, (4) for the PPT (preliminary experiment), participants sat at a desk in a comfortable chair facing the researcher. Older adult participants were encouraged to use magnifying glasses if they had difficulty seeing the holes in the pegboard. Before the main experiment, participants practiced inserting 3-4 pins into the pegboard to correct any mistakes, and then the 30-second test was conducted. The experiment began when the researcher said "start" and ended after 30 seconds, with the number of pins inserted being recorded. This process is repeated 3 times. (5) For the hand motion control measurement action (main experiment), the researcher sat in front of a laptop, with the webcam facing the participant. The participant sat in a chair facing the webcam. They were instructed to perform hand rotation movements as quickly and accurately as possible with both hands for 10 seconds. The experimenter demonstrated the motion and provided a detailed explanation before the experiment began. Once the "recording" sign appeared on the screen, participants performed the hand rotation movements. This was repeated 3 times, and there was a 30-second break in between trials. (6) A survey was conducted to collect the demographic characteristics (gender, age, education level, etc). Young participants complete the survey via Google Forms themselves, while for older adult participants, the researcher reads the survey questions aloud, and their responses are recorded in Google Forms by the researcher. After finishing the survey, the participants received the reward.

XSL•FO
RenderX

Measurements

Based on the video-captured data, we have defined 7 measurements and automatically calculated them by the system (Table 1). The single-hand test indicator assesses wrist flexibility through rotational movements. The time comparison indicator between rotations evaluates changes in hand movement over time, and the angle comparison indicator between rotations assesses changes in hand angle over time. The single-hand test indicators include total rotation count, total rotation time, and total rotation angle. The time comparison indicators between rotations include total rotation time change and the number of rotation angle changes, while the angle comparison indicators between rotations include total rotation angle change and the number of rotation angle changes. Here, the time of rotation change amount represents the total time variation for performing 1 full rotation. Simply put, if the time of rotation change amount is negative, the rotation speed increases as the number of rotations increases; if positive, the rotation speed decreases. The angle of rotation change amount indicates the unit of change in rotation angle. If negative, the rotation angle performed by the participant decreases as the number of rotations increases; if positive, the rotation angle gradually increases [29].

Results

User Statistics

This study recruited a total of 70 participants (31 young adults and 39 older adults). In total, 2 young adults were excluded based on the PPT exclusion criteria, resulting in a final sample of 68 participants: 39 (57.35%) in the experimental group and 29 (42.65%) in the control group. The analysis results showed that the average age of the young adults (20-29 years old) was a mean of 22.75 (SD 1.83) years, and the average age of the older adults (65-80 years old) was a mean of 70.36 (SD 3.70) years. Based on the dominant hand, there were 6 left-handed individuals (20.69%) and 23 right-handed individuals (79.31%) among the young adults, while there was 1 (2.78%) left-handed individual and 44 (97.22%) right-handed individuals among the older adults. In terms of educational level, there were 3 (10.34%) highly educated individuals (university graduates or higher) among the young adults and 6 (15.38%) among the older adults (refer to Table 3). This demographic information provides a basis for evaluating the impact of variables such as age, dominant hand, and educational level on hand motion control ability.

Table 3. General demographic characteristics of participants (N=68).

	Control group	Experimental group
Age group	Young (20-29 years)	Older adults (65-80 years)
Gender		
Male, n (%)	17 (58.62)	5 (12.82)
Female, n (%)	12 (41.38)	34 (87.18)
Age (years), mean (SD)	22.75 (1.83)	70.36 (3.70)
Education, n (%)		
No formal education	—[a]	3 (7.69)
Elementary school	—	7 (17.95)
Middle school	—	13 (33.33)
High school	26 (89.66)	10 (25.64)
More than college degree	3 (10.34)	6 (15.38)
Dominant hand		
Left hand	6 (20.69)	1 (2.78)
Right hand	23 (79.31)	44 (97.22)
Purdue Pegboard Test score	13.31	10.67

[a]Not applicable.

Evaluation Outcomes

A GEE model was used to assess the effects of age group (younger vs older), hand (right vs left), and trials (1, 2, and 3). Table 4 shows the representative statistics for all the measurements that were recorded to give an overview of the results. Among the measures we have defined, only 3 measures (ie, total rotation count, total rotation time, and total rotation angle) have shown a significant effect on the main factors that we have designed. The analysis results are explained for the 3 measurements, including all the interaction effects, are provided in the main text in the following sections.

XSL•FO
RenderX

Table 4. Generalized estimating equations analysis techniques (N=68).

Variable and group	Mean	SD	Significance level, *P* value		
			Group	Hand	Group x Hand
Total rotation count					
Young	25.66	7.16	—[a]	—	—
Older adults	19.56	7.02	<.01	.24	.85
Total rotation time					
Young	8.77	1.53	—	—	—
Older adults	7.37	1.84	<.01	.18	.46
Total rotation time change					
Young	–.00	0.09	—	—	—
Older adults	–.03	0.14	.32	.33	.16
The number of rotation time changes					
Young	.11	2.09	—	—	—
Older adults	–.18	1.83	.76	.72	.34
Total rotation angle					
Young	4679.74	3209.04	—	—	—
Older adults	2718.68	1235.20	<.01	.61	.17
Total rotation angle change					
Young	–1.65	45.16	—	—	—
Older adults	–2.98	33.22	.75	.34	.07
The number of rotation angle changes					
Young	.19	3.46	—	—	—
Older adults	–.62	3.17	<.58	.50	.48

[a]Not applicable.

Total Rotation Count

For the total rotation count, the model revealed a significant main effect of age group. Younger participants performed significantly more full rotations compared to older participants (B=5.29, SE 1.67, *z*=3.16, *P*=.002). Specifically, younger participants completed 5.29 more rotations on average than their older counterparts, suggesting that age significantly influences task performance. A significant main effect of trial was also observed, indicating that performance improved across trials. Participants in trial 2 completed 1.77 more rotations compared to trial 1 (B=1.77, SE 0.59, *z*=3.00, *P*=.003), and participants in trial 3 completed 2.31 more rotations compared to trial 1 (B=2.31, SE 0.72, *z*=3.20, *P*=.001). These findings suggest a learning effect across trials, with participants demonstrating better performance in subsequent trials. The main effect of the hand was not significant (B=–1.00, SE 0.87, *z*=–1.15, *P*=.25), indicating that there was no significant difference between the performance of the right and left hands.

Finally, to explore the significant main effect of trial, post hoc pairwise comparisons were conducted between trial 1, trial 2, and trial 3 using Bonferroni-corrected *P* values to account for multiple comparisons. The comparison between trial 1 and trial 2 revealed a statistically significant difference in performance, with participants showing an improvement from trial 1 to trial 2 (t_{67}=–4.87, *P*<.0001, Bonferroni-adjusted *P*=.000021). Similarly, a significant difference was found between trial 1 and trial 3, indicating continued improvement (t_{67}=–4.85, *P*<.0001, Bonferroni-adjusted *P*=.000023). However, no significant difference was observed between trial 2 and trial 3 (t_{67}=–0.76, *P*=.45, Bonferroni-adjusted *P*>.99), suggesting that performance plateaued between these 2 trials. These results confirm a learning effect between trial 1 and the subsequent trials, with participants significantly improving their performance from the first to the second and third trials. However, no further improvement was observed between the second and third trials.

Total Rotation Angle

The model revealed a significant main effect of age group, with younger participants exhibiting a significantly greater range of motion compared to older participants (B=1334.37, SE 492.99, *z*=2.71, *P*=.007). On average, younger participants had a 1334-unit larger range of motion than older participants, indicating that age significantly influences performance in this task. There was no significant main effect of hand (B=–202.30, SE 140.47, *z*=–1.44, *P*=.15), indicating no substantial difference in range of motion between the right and left hands. The main

XSL•FO
RenderX

effect of the trial showed a trend toward significance for trial 2 compared to trial 1 (B=193.69, SE 101.86, z=1.90, P=.06), although this effect did not reach conventional levels of statistical significance. However, there was a significant increase in range of motion in trial 3 compared to trial 1 (B=295.09, SE 129.29, z=2.28, P=.02), suggesting that participants exhibited a larger range of motion in the third trial. For interaction effects, several interaction effects between age group, hand, and trial were also examined. The interaction between age group and hand was marginally different (B=470.23, SE 248.58, z=1.89, P=.07), suggesting that the effect of hand may differ by age group, though this did not meet the threshold for statistical significance. The interaction between age group and trial was significant for both trial 2 (B=736.10, SE 263.73, z=2.79, P=.005) and trial 3 (B=644.74, SE 317.01, z=2.03, P=.042). These results indicate that the younger participants exhibited a significantly larger increase in range of motion across trials compared to older participants, suggesting a learning effect for the younger group as they improved more across trials. Posthoc pairwise comparisons were conducted to further explore the significant main effect of trial, with Bonferroni-corrected P-values to account for multiple comparisons. The comparison between trial 1 and trial 2 revealed a statistically significant improvement in performance, with participants showing better results in trial 2 compared to trial 1 (t_{67}=–3.68, P<.001, Bonferroni-adjusted P=.0014, Hedges g=–0.19). This indicates a meaningful improvement from trial 1 to trial 2. A significant difference was also found between trial 1 and trial 3, with participants continuing to improve (t_{67}=–3.37, P=.0013, Bonferroni-adjusted P=.0038, Hedges g=–0.21). This suggests that participants maintained better performance in trial 3 compared to trial 1.

However, no significant difference was observed between trial 2 and trial 3 (t_{67}=–0.41, P=.68, Bonferroni-adjusted P≥.99, Hedges g=–0.01), indicating that performance plateaued between these 2 trials. These results confirm a learning effect between trial 1 and the subsequent trials, where participants significantly improved from trial 1 to both trial 2 and trial 3. However, no further improvement was observed between trial 2 and trial 3.

Total Rotation Time

The model revealed a significant main effect of age group, indicating that younger participants took significantly longer on the task compared to older participants. Specifically, younger participants had a mean increase of 1.40 seconds in total task time compared to older participants (B=1.40, SE 0.35, z=3.94, P<.001). This suggests that age significantly affects the time taken to complete the rotations.

There was no significant main effect of hand (B=–0.30, SE 0.23, z=–1.33, P=.18), indicating no meaningful difference in time between the right and left hands.

The main effect of the trial was marginally different for trial 2 compared to trial 1, with participants taking approximately 0.33 seconds longer in trial 2 (B=0.33, SE 0.19, z=1.69, P=.09). However, this effect did not reach the conventional level of statistical significance. Similarly, the effect of trial 3 compared to trial 1 was not significant (B=0.32, SE 0.25, z=1.28, P=.20), suggesting that the overall time taken did not significantly change between the trials.

Discussion

Principal Results

Research on digital biomarkers based on behavioral data has primarily focused on gait and hand movement data [10]. Among these, hand movement tasks have gained attention due to their minimal spatial requirements [11,12], absence of fall risk, and the use of low-cost measurement devices. This study investigated whether hand rotation—a task with these advantages—could effectively distinguish age-related differences in motor function.

Accordingly, we hypothesized that there would be significant differences between the young adult group and the older adult group. The results from the GEE analysis demonstrated clear distinctions between younger and older participants across the 3 measures: total rotation count, total rotation angle, and total rotation time. Younger participants consistently outperformed older participants, completing more rotations and exhibiting a greater range of motion, which highlights significant age-related differences in motor efficiency and flexibility. Although younger participants took slightly longer to complete the task, this may reflect a more controlled approach to performing the rotations. Importantly, all 3 measures were effective in distinguishing age groups, proving especially valuable in capturing the declines in motor performance associated with aging. Together, these measures provide a comprehensive profile of motor function, with total rotation count reflecting speed and efficiency, total rotation angle highlighting flexibility, and total rotation time capturing pacing strategies, all of which vary significantly between younger and older adults.

Detailed Interpretations and Comparison to the Literature

The age-related differences observed in hand rotation metrics likely reflect a combination of cognitive and physical factors. This biomarker, inspired by the link between upper limb motor skills and cognitive function, aligns with prior findings showing that cognitive impairment often accompanies motor declines. Tasks like hand rotation, which require controlled, repetitive movements, depend on cognitive functions such as coordination, spatial awareness, and consistent adjustments—areas that can be affected by age-related cognitive decline. Reduced cognitive capacity could thus contribute to the lower motor performance observed in the older adults group. At the same time, physical fitness factors, including muscle strength and joint flexibility, also decline with age and impact performance on repetitive motor tasks. Although the hand rotation task is low in exertion, it demands a baseline level of physical capability, particularly for maintaining consistent movements across trials. Therefore, this biomarker likely captures influences from both cognitive and physical domains, with declines in either potentially diminishing performance. Future research might benefit from a dual approach, examining both physical fitness and cognitive performance measures in conjunction with this motor biomarker to clarify their relative contributions. This approach could help

XSL•FO
RenderX

identify the extent to which cognitive versus physical decline drives the observed performance differences and potentially enhance the biomarker's specificity for early detection of age-related cognitive impairment.

Implications of the Learning Effect: First Trial as a Practice Trial

The results clearly demonstrated a significant learning effect across trials, particularly for younger participants, who showed substantial improvement from trial 1 to trial 2 and trial 3 in both full rotations and range of motion. The lack of a significant difference between trial 2 and trial 3 suggests that participants reached a performance plateau by the second trial. This indicates that trial 1 likely served as a familiarization or practice phase, during which participants adapted to the task demands. To improve the accuracy of future analyses and minimize the confounding effect of learning, it would be advisable to treat the first trial as a practice trial and focus analysis on trial 2 and trial 3 as representative of stable performance. By doing so, researchers can better isolate genuine performance differences between age groups, unconfounded by initial adaptation or learning.

Limitations

This study has several limitations. First, the sample size used in the study was limited, which may restrict the generalizability of the results. Future research should increase the sample size to enhance the reliability of the findings. Additionally, this study was cross-sectional, analyzing data at a single point in time. To more accurately understand changes associated with age, longitudinal studies that track changes over time are needed. As the first step in validating digital biomarkers, this study conducted experiments on younger and older adults through hand rotation movements prior to MCI and dementia diagnoses. Although we informed that we were recruiting participants without MCI, we did not conduct an additional screening test. In subsequent phases, we plan to collaborate with Dong-eui University Oriental Medicine Hospital to conduct cognitive tests on older adults diagnosed with MCI and dementia, as well as older adults without cognitive impairment. Through this, we aim to explore and analyze differences in physical, physiological, and gender factors. Furthermore, future research will increase the sample size to study differences between genders and age groups and continue to explore the potential of hand rotation movements as digital biomarkers.

Conclusions

This study confirmed that hand rotation metrics can effectively distinguish motor performance differences between younger and older adults. The older group showed statistically significant decreases in rotation frequency, angle, and duration compared to the younger group, indicating an age-related decline in motor control abilities. Additionally, a learning effect was observed during repeated trials, suggesting that excluding the first trial may enhance measurement stability in future assessments.

These findings demonstrate the potential of hand rotation as a digital biomarker for capturing age-related changes in motor performance and further suggest its applicability as an early detection tool for MCI and other early cognitive declines. This method can be implemented across various platforms—including webcams, wearable devices, and VR systems—enabling broad applications such as clinical diagnostic support, home-based self-assessment kiosks, and community health screening. In particular, it holds promise as an important assessment tool for the early identification of neurodegenerative diseases and for the development of tailored intervention programs.

Future studies should aim to further refine this hand rotation-based digital biomarker for broader applicability. This study has some limitations, including a relatively small and demographically homogeneous sample, which may limit the generalizability of the findings. Follow-up research involving larger and more diverse populations is essential to improve the generalizability of the findings. Moreover, applying this method to individuals with MCI and other cognitive impairments will help evaluate its effectiveness in detecting early signs of cognitive decline. Expanding the framework to include other motor tasks and integrating multimodal indicators—including cognitive, physical, and emotional factors—could contribute to building a comprehensive and personalized digital health monitoring system for early diagnosis and preventive care.

Acknowledgments

This work was supported by the innovative Human Resource Development for Local Intellectualization program through the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (IITP-2025-RS-2020-II201791).

Data Availability

The dataset generated and analyzed during this study is publicly available on GitHub [30]. The repository contains anonymized hand rotation movement data from younger and older adult participants, along with metadata such as participant group information and measurement variables (eg, rotation count, angle, and time).

Conflicts of Interest

None declared.

References

XSL•FO
RenderX

1. Youn BY, Ko Y, Moon S, Lee J, Ko SG, Kim JY. Digital biomarkers for neuromuscular disorders: a systematic scoping review. Diagnostics (Basel) 2021;11(7):1275 [FREE Full text] [doi: 10.3390/diagnostics11071275] [Medline: 34359358]
2. Motahari-Nezhad H, Fgaier M, Mahdi Abid M, Péntek M, Gulácsi L, Zrubka Z. Digital biomarker-based studies: scoping review of systematic reviews. JMIR Mhealth Uhealth 2022;10(10):e35722 [FREE Full text] [doi: 10.2196/35722] [Medline: 36279171]
3. Josephy-Hernandez S, Norise C, Han JY, Smith KM. Survey on acceptance of passive technology monitoring for early detection of cognitive impairment. Digit Biomark 2021;5(1):9-15 [FREE Full text] [doi: 10.1159/000512207] [Medline: 33615117]
4. Tombaugh TN. Test-retest reliable coefficients and 5-year change scores for the MMSE and 3MS. Arch Clin Neuropsychol 2005;20(4):485-503. [doi: 10.1016/j.acn.2004.11.004] [Medline: 15896562]
5. Piau A, Wild K, Mattek N, Kaye J. Current state of digital biomarker technologies for real-life, home-based monitoring of cognitive function for mild cognitive impairment to mild alzheimer disease and implications for clinical care: systematic review. J Med Internet Res 2019;21(8):e12785 [FREE Full text] [doi: 10.2196/12785] [Medline: 31471958]
6. Kaye J, Mattek N, Dodge HH, Campbell I, Hayes T, Austin D, et al. Unobtrusive measurement of daily computer use to detect mild cognitive impairment. Alzheimers Dement 2014;10(1):10-17 [FREE Full text] [doi: 10.1016/j.jalz.2013.01.011] [Medline: 23688576]
7. Popp Z, Low S, Igwe A, Rahman MS, Kim M, Khan R, et al. Shifting from active to passive monitoring of alzheimer disease: the state of the research. J Am Heart Assoc 2024;13(2):e031247 [FREE Full text] [doi: 10.1161/JAHA.123.031247] [Medline: 38226518]
8. Byun S, Lee HJ, Kim JS, Choi E, Lee S, Kim TH, et al. Exploring shared neural substrates underlying cognition and gait variability in adults without dementia. Alzheimers Res Ther 2023;15(1):206 [FREE Full text] [doi: 10.1186/s13195-023-01354-y] [Medline: 38012628]
9. Hantke NC, Kaye J, Mattek N, Wu CY, Dodge HH, Beattie Z, et al. Correlating continuously captured home-based digital biomarkers of daily function with postmortem neurodegenerative neuropathology. PLoS One 2023;18(6):e0286812 [FREE Full text] [doi: 10.1371/journal.pone.0286812] [Medline: 37289845]
10. Jang Y, Kim H, Kim S. Research trends and usability challenges in behavioral data-based cognitive function assessment. Electronics 2024;13(19):3830 [FREE Full text] [doi: 10.3390/electronics13193830]
11. bo H, Jeong G, Bae HY, soo S. Comparison of gait characteristics in young and old persons with GAITRite system analysis. PNF Movement 2003;1(1):33-42.
12. Kanko RM, Laende E, Selbie WS, Deluzio KJ. Inter-session repeatability of markerless motion capture gait kinematics. J Biomech 2021;121:110422. [doi: 10.1016/j.jbiomech.2021.110422] [Medline: 33873117]
13. Negin F, Rodriguez P, Koperski M, Kerboua A, Gonzàlez J, Bourgeois J, et al. PRAXIS: towards automatic cognitive assessment using gesture recognition. Expert Syst Appl 2018;106:21-35. [doi: 10.1016/j.eswa.2018.03.063]
14. Chua SIL, Tan NC, Wong WT, Allen JC, Quah JHM, Malhotra R, et al. Virtual reality for screening of cognitive function in older persons: comparative study. J Med Internet Res 2019;21(8):e14821 [FREE Full text] [doi: 10.2196/14821] [Medline: 31373274]
15. Park J, Seo K, Kim S, Ryu H, Choi H. Early screening of mild cognitive impairment through hand movement analysis in virtual reality based on machine learning: screening of MCI through hand movement in VR. J Cogn Interv Digit Health 2022;1(1):1. [doi: 10.58558/jcd.2022.1.1.1]
16. Aggarwal NT, Wilson RS, Beck TL, Bienias JL, Bennett DA. Motor dysfunction in mild cognitive impairment and the risk of incident Alzheimer disease. Arch Neurol 2006;63(12):1763-1769. [doi: 10.1001/archneur.63.12.1763] [Medline: 17172617]
17. Scherder E, Dekker W, Eggermont L. Higher-level hand motor function in aging and (preclinical) dementia: its relationship with (instrumental) activities of daily life--a mini-review. Gerontology 2008;54(6):333-341 [FREE Full text] [doi: 10.1159/000168203] [Medline: 18997468]
18. Ilardi CR, Iavarone A, La Marra M, Iachini T, Chieffi S. Hand movements in mild cognitive impairment: clinical implications and insights for future research. J Integr Neurosci 2022;21(2):67 [FREE Full text] [doi: 10.31083/j.jin2102067] [Medline: 35364655]
19. Wu Q, Chan JSY, Yan JH. Mild cognitive impairment affects motor control and skill learning. Rev Neurosci 2016;27(2):197-217. [doi: 10.1515/revneuro-2015-0020] [Medline: 26426886]
20. Kim HJ, Kim JS, Kim SH. Vision-based hand rotation recognition technique with ground-truth dataset. Appl Sci 2024;14(1):422 [FREE Full text] [doi: 10.3390/app14010422]
21. Liu F, Qiu K, Wang H, Dong Y, Yu D. Decreased wrist rotation imitation abilities in children with autism spectrum disorder. Front Psychiatry 2024;15:1349879 [FREE Full text] [doi: 10.3389/fpsyt.2024.1349879] [Medline: 38699453]
22. Seidler RD, Bernard JA, Burutolu TB, Fling BW, Gordon MT, Gwin JT, et al. Motor control and aging: links to age-related brain structural, functional, and biochemical effects. Neurosci Biobehav Rev 2010;34(5):721-733 [FREE Full text] [doi: 10.1016/j.neubiorev.2009.10.005] [Medline: 19850077]
23. Sorond FA, Cruz-Almeida Y, Clark DJ, Viswanathan A, Scherzer CR, De Jager P, et al. Aging, the central nervous system, and mobility in older adults: neural mechanisms of mobility impairment. J Gerontol A Biol Sci Med Sci 2015;70(12):1526-1532 [FREE Full text] [doi: 10.1093/gerona/glv130] [Medline: 26386013]

XSL•FO
RenderX

24. Maes C, Gooijers J, Orban de Xivry J, Swinnen SP, Boisgontier MP. Two hands, one brain, and aging. Neurosci Biobehav Rev 2017;75:234-256. [doi: 10.1016/j.neubiorev.2017.01.052] [Medline: 28188888]
25. Colella D, Guerra A, Paparella G, Cioffi E, Di Vita A, Trebbastoni A, et al. Motor dysfunction in mild cognitive impairment as tested by kinematic analysis and transcranial magnetic stimulation. Clin Neurophysiol 2021;132(2):315-322. [doi: 10.1016/j.clinph.2020.10.028] [Medline: 33450553]
26. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986;73(1):13-22 [FREE Full text] [doi: 10.1093/biomet/73.1.13]
27. Marvin K. Purdue Pegboard Test (PPT). Dostupno Stroke Engine. 2012. URL: https://strokengine.ca/en/assessments/purdue-pegboard-test-ppt/ [accessed 2024-08-06]
28. Lugaresi C, Tang J, Nash H, McClanahan C, Uboweja E, Hays M, et al. MediaPipe: A Framework for Building Perception Pipelines. 2019. URL: http://arxiv.org/abs/1906.08172 [accessed 2024-08-06]
29. Kim S, Kim M. Rotation representations and their conversions. IEEE Access 2023;11:6682-6699. [doi: 10.1109/access.2023.3237864]
30. GitHub - deudxlab/handcognition. URL: https://github.com/deudxlab/handcognition

Abbreviations

AI: artificial intelligence
GEE: generalized estimating equations
MCI: mild cognitive impairment
PPT: Purdue Pegboard Test
VR: virtual reality

Edited by A Mavragani; submitted 12.08.24; peer-reviewed by J Zhou, P Nohl-Deryk; comments to author 23.09.24; revised version received 17.10.24; accepted 23.04.25; published 21.07.25.

Please cite as:
Kim JS, Kim HJ, Kim M, Kim SH
Hand Motion Control Ability Between Young and Older Adults: Comparative Study
JMIR Form Res 2025;9:e65224
URL: https://formative.jmir.org/2025/1/e65224
doi:10.2196/65224
PMID:40690275

©Jung-Soon Kim, Hui-Jun Kim, Minju Kim, Sung-Hee Kim. Originally published in JMIR Formative Research (https://formative.jmir.org), 21.07.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Health Care Professionals' Perspectives on Technology Use in Urinary Care: Cross-Sectional Survey-Based Study

Nicole Zhang[1], RN, PhD; Ni Zhang[2], MA, MPH, PhD; Yun Wang[3], PhD; Rachel Wong[2], BS; Kaylin Ma[1], BSN

[1]Valley Foundation School of Nursing, College of Health and Human Sciences, San Jose State University, San Jose, CA, United States

[2]Department of Public Health and Recreation, College of Health and Human Sciences, San Jose State University, San Jose, CA, United States

[3]Biomedical Engineering Department, Charles W Davidson College of Engineering, San Jose State University, San Jose, CA, United States

Corresponding Author:
Nicole Zhang, RN, PhD
Valley Foundation School of Nursing
College of Health and Human Sciences
San Jose State University
1 Washington Sq
San Jose, CA, 95192
United States
Phone: 1 (408) 924 3131
Email: nicole.zhang@sjsu.edu

Abstract

Background: Urinary issues pose a significant burden on health care systems, necessitating innovative solutions to enhance patient care and alleviate the provider burden.

Objective: The objective of this study was to explore health care professionals' perceptions of urinary issues and assess their acceptance and readiness to adopt wearable and remote technologies for managing these conditions. The study aimed to identify the attributes and barriers influencing technology integration in clinical practice, using established theoretical frameworks, such as the Health Belief Model (HBM) and the Technology Acceptance Model (TAM).

Methods: A cross-sectional survey-based study was conducted. A structured survey questionnaire was administered online to a sample of 256 health care professionals recruited through social media and personal networks. The survey included both closed- and open-ended questions to gather data. Quantitative data were analyzed using descriptive statistics, Pearson correlation, and multiple regression.

Results: Quantitative analysis revealed strong correlations between belief agreement and factors such as health literacy (r=0.591, P<.001), the perceived burden (r=0.628, P<.001), device attributes (r=0.650, P<.001), and support services (r=0.622, P<.001). Multiple regression analysis identified that the perceived burden (β=.284, P=.01), device attributes (β=.371, P<.001), and integrating technology (β=.312, P<.001) are positively associated. The survey demonstrated strong internal consistency, with Cronbach α=.85, indicating high reliability in measuring health care professionals' perceptions of technology adoption.

Conclusions: Health care professionals' acceptance of technology in managing urinary issues is influenced by factors such as the perceived burden, device attributes, and the ease of integrating technology into existing workflows. Addressing barriers to technology adoption, providing comprehensive training and support, and prioritizing user-centered design are crucial for successful technology integration. Future research should focus on longitudinal studies and explore the perspectives of patients and other stakeholders to gain a more holistic understanding of technology integration in urological care.

(JMIR Form Res 2025;9:e73453) doi:10.2196/73453

KEYWORDS

internet research; survey study; urology; technology; urinary health technology

XSL•FO
RenderX

Introduction

Background

Urinary issues are pervasive in health care across multiple settings. Common urinary issues encountered by health care professionals include conditions such as urinary tract infection (UTI), urinary incontinence (UI), benign prostatic hyperplasia (BPH), and urinary retention (UR). The prevalence of these conditions is staggering. As of 2021, there were an estimated 4.49 billion cases of UTIs worldwide, representing a 66.45% increase since 1990 [1]. The age-standardized incidence rate (ASIR) was 5531.88 per 100,000 population, with women and older adult men experiencing the highest incidence. Approximately 20 million American women and 6 million American men experience UI at some point in their lives, with prevalence increasing with age. In women over 30 years old, stress incontinence affects 24%-45%. The significance of addressing urinary issues is multifaceted. Clinically, UTIs are the second-most common type of infection in older adults, potentially leading to serious complications, such as kidney failure and even death if left untreated. These conditions impose a heavy burden on patients, as well as caregivers, influencing workflow management in health care settings. Economically, the impact is substantial [2].

Despite the prevalence of urologic conditions, these conditions often remain underdiagnosed and undertreated, especially among older adults and marginalized populations. Timely detection and effective management are essential but are frequently hindered by the limitations in current monitoring methods. These current methods rely heavily on self-reporting and episodic clinical assessments. Advances in health technology, particularly wearable and remote devices, offer promising solutions to bridge these gaps and enable real-time noninvasive monitoring of urinary symptoms and behaviors. Understanding health care professionals' acceptance of these technologies is critical to their successful implementation in adoption and practice.

Although general technology integration in health care has received substantial attention, specific challenges in urinary care, such as underreporting, delayed diagnosis, and lack of real-time monitoring, remain under addressed. Urinary conditions, such as UI and lower urinary tract symptoms (LUTS), significantly affect the quality of life, particularly among older adults, yet they are often overlooked due to stigma and limited access to effective assessment tools [3]. Traditional care models rely heavily on self-reporting or episodic clinical assessments, which may fail to capture the dynamic nature of symptoms. Delay in identification and intervention can exacerbate morbidity and contribute to poor outcomes [4]. Emerging wearable and digital technologies offer a potential solution, enabling noninvasive, continuous monitoring to support earlier recognition and personalized interventions [5,6]. However, as of now, the adoption of these tools in urinary care remains limited. This study sought to explore health care professionals' readiness to integrate such technologies, with particular attention to the barriers and enablers specific to urinary care delivery.

The staggering prevalence of urinary issues underscores the critical need for innovative solutions in health care. As we find ourselves in an era of rapid technological advancement, the integration of technology into medical practice has become not just beneficial but also necessary, especially when confronting such widespread health concerns.

Technology integration in health care has emerged as a key driver in improving patient care and operational efficiency. With the increasing complexity of health care delivery, stand-alone systems are no longer sufficient to address the multifaceted challenges posed by conditions such as UTIs and UI. The integration of various health care technologies offers a holistic approach to patient care, enabling seamless communication between different systems and departments. In the context of urinary health, technological advancements have led to the development of innovative tools and approaches. By leveraging technology integration, health care professionals can streamline operations, reduce administrative burdens, and focus more on delivering high-quality patient care. This approach is essential in tackling the growing prevalence of urinary conditions and improving overall health care efficiency. Although this is the case, there are many aspects to consider regarding health care professionals' acceptability of technology in practice.

Factors Surrounding Technology Acceptance in Health Care Settings

The integration of digital technologies in health care, ranging from electronic health records (EHRs) to artificial intelligence (AI) systems, promises to enhance patient care, safety, and clinical efficiency. However, successful implementation depends heavily on the acceptance of these technologies by health care professionals, including nurses, physicians, and allied staff. Understanding the factors influencing their acceptance is critical for designing effective interventions and implementation strategies.

Individual-Level Factors

Health care professionals' willingness to adopt new technologies is often shaped by their individual traits. Personal innovativeness, self-efficacy, and readiness for change are strong predictors of technology acceptance. Studies have found that health care professionals with greater self-confidence in using digital tools are more inclined to adopt systems such as EHRs and telemedicine [7,8]. Conversely, psychological barriers, such as technophobia, resistance to change, and fear of reduced autonomy, can significantly hinder acceptance [9]. Addressing these internal barriers through education and peer support is a recurrent recommendation.

The perceived benefit of a system plays a pivotal role. If health care professionals believe a specific technology improves clinical workflow or patient outcomes, they are more likely to accept it [10,11]. This finding echoes across multiple studies, highlighting that practical, outcome-oriented value is central to user adoption.

Technology-Related Factors

A consistent theme in the literature is that ease of use and a user-friendly design facilitate acceptance. Technologies

XSL•FO
RenderX

perceived as intuitive and minimally disruptive have higher adoption rates [10,12]. These characteristics align with the core components of the Technology Acceptance Model (TAM), which is frequently used as a guiding framework in the studies reviewed. System failures, bugs, and unreliability are significant deterrents. For instance, a study found that technical issues, especially with AI-based decision support tools, causes distrust and disengagement. Alert fatigue, particularly in decision support systems, is cited as a recurring concern among clinicians [9].

Organizational Context

Adequate infrastructure is a foundational requirement. Without access to a stable internet connection, compatible hardware, and technical support, even the most promising technologies face poor uptake [13]. Targeted and ongoing training programs are crucial to equip users with the skills needed to operate new systems. Studies consistently recommend that training be context specific, iterative, and responsive to user needs [9,11]. The involvement of clinical staff in the design and rollout of technology enhances buy-in. Implementation strategies that include stakeholder engagement, local champions, and organizational incentives have been found to increase adoption success [13].

Professional Practice Impact

Technologies that align well with clinical workflows are more readily accepted. Systems that disrupt routines or increase the burden face greater resistance [11,13]. Effective technologies must therefore be designed with a deep understanding of clinical operations. Many health care professionals express concern that technology, especially AI and decision support systems, may diminish their clinical judgment or autonomy [9]. This is less of a concern with more administrative technologies, such as EHRs, but remains relevant across the digital health spectrum. Perceived effects on patient care, either positive (eg, efficiency, safety) or negative (eg, depersonalization, error risk), strongly influence acceptance. Health care professionals are more open to using technologies they believe enhance patient outcomes [14].

The Gap and How Technology Can Help

Technology offers promising solutions for managing urinary issues more effectively. Electronic alerts in EHRs have been shown to significantly reduce catheter-associated urinary tract infections (CAUTIs). Researchers have developed mobile uroflowmetry technology and urination log apps that use AI and acoustic analysis to assist in diagnosing urinary diseases [15]. Wearable sensing devices and Internet of Things (IoT) technologies can collect and analyze urination information for personal health care services and disease management. Overall, the use of easy-to-operate sensing technology is currently underused in addressing these common issues. Research in other domains, such as diabetes and cardiovascular care, has shown that health care professionals' acceptance of wearable technology hinges on usability, integration into workflows, and institutional support [16,17]. These lessons provide a valuable context for understanding how similar factors may influence wearable technology adoption in urinary care.

Although there is a growing body of literature on wearable health technologies, most studies emphasize clinical efficacy or patient-centered outcomes. Few focus on health care providers' acceptance, particularly for urinary applications. This study fills that gap by examining how nurses, physicians, and allied health staff perceive wearable and remote technologies for managing urinary issues, including their readiness to adopt such innovations into practice.

Our study addressed wearable and remote sensing technologies for urinary issues; however, we also addressed general health care technology exposure and acceptability. Specifically, in addition to urinary-specific technologies, the study also examined respondents' general exposure to digital health tools, such as remote monitoring apps, wearables, and telemedicine platforms, which could influence their openness to adopting similar tools for urinary care. This approach allowed us to gain a comprehensive understanding of health care providers' perceptions and readiness to adopt new technologies in their practice and the attributes of those technologies. The survey explored health care providers' exposure to various technologies, both specific to urinary care and in general health care settings. This helped us assess their familiarity with different technological solutions and how this exposure might influence their acceptance of new technologies.

Previous research has found that technology is more likely to be accepted if there are fewer financial constraints, appropriate training, proper integration into current systems, and adequate technical support [18]. However, barriers to technology adoption include cost, efficient implementation and integration, and effective training. Broader considerations, such as resistance to change and lack of workflow integration, can also impact the efficiency of work and technology adoption.

Addressing Gaps and Research Goals

This study aimed to fill gaps in previous research by investigating health care providers' perceptions of urinary issues and exploring the use of technology for managing these issues. The research questions are as follows:

- What is the level of acceptance and readiness among health care providers toward adopting health care technology for treating and managing urologic issues?
- What attributes and barriers influence the clinical integration and utility of health care technologies for urinary issue management?

By addressing these research questions through a cross-sectional survey, the study sought to understand the major barriers and facilitators identified by health care professionals across multiple health care settings. This understanding will help facilitate the adoption of new technology, ultimately improving patient care and reducing the burden of urinary issues on health care systems. Health care professionals are the gatekeepers of technology implementation in clinical settings. Their acceptance is essential for technology to be successfully integrated into workflows, adopted by patients, and sustained over time.

The research used a comprehensive survey designed to capture health care professionals' views across multiple health care settings. The survey was developed through a rigorous process,

XSL•FO
RenderX

beginning with an extensive literature review to identify key themes and gaps in current knowledge. This was followed by consultations with subject matter experts in urology, nursing, and health technology to ensure the survey's content validity.

Theoretical Underpinnings

In this study, we used TAM as a guiding framework and integrated constructs from the Health Belief Model (HBM) to better understand health care providers' attitudes, acceptance, and readiness toward adopting technologies for managing urologic issues. Both models provided complementary perspectives for examining behavioral drivers and perceived barriers to clinical technology integration.

TAM focuses on 2 primary constructs, perceived usefulness and perceived ease of use, as key predictors of technology adoption. These concepts aligned well with our exploration of how health care providers perceive and use technology in the management of urologic conditions. For example, their willingness to use health care technologies is often influenced by whether they believe these tools will improve efficiency, clinical outcomes, or patient care (perceived usefulness) and whether they feel the technology is easy to learn and integrate into existing workflows (perceived ease of use). These elements directly inform attitudes toward use and behavioral intentions, which were central to our investigation of acceptance and readiness.

To further deepen our analysis, we incorporated elements from the HBM, which emphasizes individuals' perceptions of susceptibility, severity, benefits, and barriers in relation to health behavior. Applying the HBM allowed us to explore how health care providers' beliefs about the importance of managing urologic issues may shape their motivation to adopt technology. For example, if a health care provider perceives urinary issues as a significant and underaddressed problem in clinical practice, they may be more motivated to adopt technologies they believe could mitigate these risks, especially if they believe such tools are beneficial and not overly burdensome. Generally, the HBM is helpful for framing beliefs around addressing health issues. This includes the modalities for addressing health issues, so this framework is helpful for assessing the facilitators of and barriers to technology adoption.

Items derived from TAM were evaluated as essential determinants of technology adoption, including perceived usefulness and perceived ease of use. These constructs are widely validated in studies examining the acceptance of health care technologies and are relevant for assessing health care professionals' readiness to engage in technology usage.

To complement the behavioral focus of TAM, key constructs from the HBM, including perceived susceptibility, perceived severity, perceived benefits, and perceived barriers, were included to assess health-related motivations and the perceived consequences of adopting or resisting technology.

By combining TAM and the HBM, we were able to capture both the technological factors that affect adoption (eg, usability and perceived value) and the health-related beliefs that drive motivation and behavior in clinical decision-making. This dual-framework approach offered a robust structure for evaluating the multifaceted nature of technology integration in urinary care. Our survey instrument was developed using this dual-framework approach and included additional items drawn from validated questionnaires These theories provided complementary perspectives on technology adoption in health care settings, offering a comprehensive framework for understanding and promoting the acceptance of urinary health technologies among both patients and health care professionals.

Methods

Study Design

A cross-sectional study was conducted using a structured survey questionnaire developed to collect data from a variety of health care professionals, including nurses, physicians, nurse practitioners, and other patient-facing professionals. The survey included both closed- and open-ended questions to capture a wide range of perspectives and experiences. Pilot testing was conducted with a small group of health care professionals to refine the survey's clarity, relevance, and ease of use. The final survey instrument covered a range of topics, including health care professionals' experiences with urinary issues, their attitudes toward various technologies for urinary management, perceived barriers to and facilitators of technology adoption, and demographic information. By administering this carefully crafted survey, we sought to obtain valuable insights into the current landscape of urinary care and the potential for technological interventions to improve patient outcomes and health care efficiency.

Recruitment

For this study, participants were recruited through a combination of social media platforms, specifically Reddit, and personal networks. The recruitment process aimed to gather a diverse sample representative of the target population. This form of recruitment could introduce bias into the sample and has certain limitations that will be described in greater detail in the *Limitations* section of this paper.

Reddit, a widely used online forum and community platform, was used for recruitment to reach a broad and varied demographic. The research team identified relevant subreddits that aligned with the survey's topic, ensuring a population with interests and experiences related to the study. Recruitment posts were created, which included a brief description of the survey's purpose, eligibility criteria, estimated time commitment, and a link to the survey. To maximize visibility and engagement, posts were shared in multiple subreddits, and follow-up comments were provided to address questions and encourage participation. In some cases, moderators of specific subreddits were contacted in advance to ensure that the study met subreddit guidelines and was not perceived as spam.

In addition to Reddit, recruitment was extended through the research team's personal networks, including professional and academic connections. Invitations were sent via email, direct messaging, and personal referrals to individuals who met the study's eligibility criteria. These communications contained details of the survey, emphasizing confidentiality, voluntary participation, and potential benefits. To promote a snowball

XSL•FO
RenderX

sampling effect, participants were encouraged to share the survey link with individuals in their networks who might be eligible to participate.

Inclusion and Exclusion Criteria

There were 286 responses to the survey. After cleaning the data and applying the inclusion and exclusion criteria, we ended up with 256 participants included in the resultant study. Inclusion criteria were (1) current or previous employment as a health care provider in a patient-facing health care setting (list of professions included in survey), (2) age 18 years or older, (3) proficiency in English, and (4) completion of most of the survey. We chose to seek the perspectives of multiple health care providers because we wanted to figure out whether technology implementation and perception and acceptability of technology could be related to various professional perceptions that may be related to the job title or role. We also wanted to create a comprehensive assessment of technology integration needs from different potential users. A screening question at the beginning of the survey confirmed participants' health care roles (eg, physician, nurse, allied health staff). Those who did not self-identify as health care professionals were screened out.

Ethical Considerations

The study was conducted in compliance with ethical guidelines and was approved by the Institutional Review Board (IRB) of San Jose State University (IRB protocol #24-087). Prior to recruitment and data collection, all study procedures were reviewed and determined to meet the ethical standards for research involving human participants. All procedures involving human participants were conducted in accordance with ethical standards. Participants provided informed consent prior to participation, and confidentiality was maintained throughout the study.

Data Collection

The survey was administered online using the Qualtrics survey platform, with an estimated completion time of 15 minutes per participant. Survey items were developed based on validated constructs from several surveys [19-21]. The instrument underwent expert review by clinicians and digital health researchers to ensure content validity. To support survey validation, the instrument was reviewed by a panel of 10 content experts in nursing, urology, and public health. Their feedback was used to refine question clarity, content relevance, and construct alignment, thereby enhancing content and face validity. Where appropriate, the survey drew upon existing validated instruments, such as the eHealth Literacy Scale (eHEALS) and others described in depth in the *Measurement* section [21,22]. A pilot test (n=10) was conducted to assess question clarity, item wording, and overall usability. The experts provided feedback on item relevance, wording, and alignment with theoretical constructs, supporting content and face validity. Based on their recommendations, several items were reworded for clarity, redundant questions were removed, and the response scale was refined. This process ensured that the instrument accurately captured the intended domains and was appropriate for the target population. Additionally, the pilot confirmed the feasibility and usability of the survey format, contributing to overall instrument validity prior to full-scale distribution.

Data Analysis

Statistical analyses were conducted using IBM SPSS Statistics. Descriptive statistics were used to summarize participant demographics and key survey variables. Prior to inferential testing, we assessed the normality of continuous variables using the Shapiro-Wilk test and visual inspection of histograms and quintile-quintile (Q-Q) plots. Most key variables were normally distributed, with 1 major variable, health literacy, present as not normally distributed. Despite this deviation from normality, we used parametric tests, specifically Pearson correlation and multiple linear regression, based on central limit theorem (CLT), which supports the use of parametric methods when sample sizes are sufficiently large, typically over 200 participants (N=256). We also assessed regression assumptions, including linearity and multicollinearity. All variance inflation factor (VIF) values were below 2.0, indicating no significant multicollinearity between predictors.

Next, to examine associations between health care providers' perceptions of technology. Pearson correlation analyses were conducted to examine the bivariate relationships between belief agreement and key independent variables. Additionally, simultaneous multiple regression was conducted to assess predictors of belief agreement, with independent variables and covariates. Cronbach α was computed to assess the internal reliability of survey subscales, with a threshold of .70 considered acceptable for reliability.

Measurement

The survey questions were developed using validated constructs, TAM and the HBM, to assess factors influencing the acceptance of technology by health care professionals. HBM variables were assessed using a validated questionnaire adapted from Champion and Skinner's study [19,20]. Several survey constructs were used to develop questions measuring the perceptions of innovation adoption, specifically looking at attributes of technology effectiveness and receptiveness of technology. Several specific Likert scale questions came from the Unified Theory of Acceptance and Use of Technology (UTAUT), as described by Venkatesh et al [20], and the Diffusion of Innovations framework by Rogers [21], which informed perceptions around innovation and adoption behaviors. To assess perceived usefulness and ease of use, which are central to technology adoption, items were adapted from TAM, originally developed by Davis [23]. Additionally, the survey incorporated elements from Shertzer's work [24] on self-efficacy which addresses resistance to change. To measure digital literacy and comfort with online health tools, items were adapted from eHEALS developed by Norman and Skinner [25]. Usability perceptions of health IT systems were captured using items adapted from the Customizable Health IT Usability Evaluation Scale developed by Yen et al [26]. For clarity, we defined each of the variables in the statistical analysis (Table 1).

XSL•FO
RenderX

Table 1. Variable definitions and construct triangulation.

Variable	Definition
Belief agreement	Health care professionals' agreement with the perception that technology is efficient and their confidence in its integration, patient acceptance, its transformative potential, and the value of investing in technological solutions for managing urinary issues
Health literacy	The extent to which health care professionals understand and apply information about urinary health and related technologies
Perceived burden	The level of difficulty or inconvenience health care professionals associate with adopting new urinary health and related technologies
Frequency	How often health care professionals encounter patients with urinary concerns or use urinary health and related technologies in practice
Bladder volume	The utility of the estimated or measured urine volume in the bladder, as assessed through various technologies
Device attributes	The specific features of urinary health and related technology that health care professionals consider important when evaluating its usefulness
Importance	The priority health care professionals assign to technology use in improving patient outcomes
Concern level	The degree of apprehension or hesitation health care professionals have regarding the adoption of urinary health and related technology
Patient receptiveness	The willingness and openness of patients to use urinary health and related technology as part of their health care routine
Integrating technology	The feasibility and ease with which health care professionals can incorporate technology into their workflow
Support services	The availability of technical assistance, training, and resources needed to facilitate the implementation of technology in clinical settings

Results

Reliability Analysis

To assess the internal consistency of the survey instrument, Cronbach α was calculated for the Likert scale questions. The overall survey instrument demonstrated strong internal consistency, with Cronbach α=.85, indicating good reliability. All individual constructs within the survey also showed good internal consistency, with Cronbach α values ranging from .80 to .90, supporting the reliability of the scale across multiple domains. This suggests that the survey items measuring health care professionals' perceptions of urinary issues and technology adoption were highly consistent in capturing the intended constructs.

Quantitative Results

Of the 256 respondents, 94 (36.7%) identified as male, 155 (60.5%) as female, 3 (1.2%) as nonbinary/third gender, and 4 (1.6%) as self-described or "preferred not to state." The mean age was 36.37 (SD 8.39, range 22-75) years. The majority of respondents were physicians (n=171, 66.8%), registered nurses (n=42, 16.4%), and allied health staff (n=17, 6.6%). The demographic data can be seen in Table 2.

XSL•FO
RenderX

Table 2. Participant demographics (N=256).

Category	Value[a]
Gender, n (%)	
Male	94 (36.7)
Female	155 (60.5)
Nonbinary/third gender	3 (1.2)
Self-described/prefer not to state	4 (1.6)
Occupation, n (%)	
Physician	171 (66.8)
Registered nurse	42 (16.4)
Allied health staff	17 (6.6)
Medical assistant	7 (2.7)
Medical technologist	6 (2.3)
Nurse practitioner	5 (2.0)
Licensed vocational nurse	2 (0.8)
Certified nurse assistant	1 (0.4)
Other	5 (2.0)
Age (years), mean (SD)	36.37 (8.39)
20-29	18 (7.0)
30-39	155 (60.5)
40-49	62 (24.2)
50-59	14 (5.5)
60-69	4 (1.6)
70-79	1 (0.4)
Other	2 (0.8)
Education, n (%)	
Master's degree	97 (37.9)
Bachelor's degree	93 (36.3)
Doctoral degree	26 (10.2)
Associate's degree	17 (6.6)
Technical degree or working training	16 (6.2)
Other or prefer not to say	7 (2.7)

[a]The sums of percentages might not add up to 100% due to rounding.

Pearson correlation analysis was conducted to examine the relationship between belief agreement and a range of independent variables related to health care providers' perspectives on urinary care and technology integration (Table 3). All variables demonstrated statistically significant positive correlations with belief agreement ($P<.001$), indicating that as each factor increases, so does the level of agreement with beliefs about the value and utility of managing urinary issues with technology.

XSL•FO
RenderX

Table 3. Correlation between independent variables and belief agreement.

Variable	r	Mean (SD)	*P* value
Health literacy	0.591	29.285 (6.890)	<.001
Perceived burden	0.628	73.549 (18.546)	<.001
Frequency	0.497	46.714 (8.108)	<.001
Bladder volume (monitoring perceptions)	0.611	31.042 (5.520)	<.001
Device attributes	0.650	46.403 (9.514)	<.001
Importance (of technology)	0.572	38.436 (10.614)	<.001
Concern level	0.351	30.895 (5.046)	<.001
Patient receptiveness	0.612	32.615 (7.912)	<.001
Integrating technology	0.479	25.321 (4.393)	<.001
Support services	0.622	30.823 (5.697)	<.001

The strongest correlation was observed between device attributes (r=0.650, *P*<.001) and belief agreement, suggesting that perceptions of favorable device features are strongly associated with greater belief in the value of urinary health and related technology. Similarly, perceived burden (r=0.628, *P*<.001) and support services (r=0.622, *P*<.001) were also highly correlated with belief agreement, indicating that health care providers who perceive urinary conditions as burdensome and who value supportive resources tend to hold stronger beliefs in the benefit of technology-based solutions.

Notably, patient receptiveness (r=0.612, *P*<.001) and bladder volume–monitoring perceptions (r=0.611, *P*<.001) were also strongly correlated with belief agreement, highlighting the importance of both patient attitudes and clinical relevance in shaping provider beliefs. Health literacy (r=0.591, *P*<.001) and the importance of urinary management (r=0.572, *P*<.001) showed moderate-to-strong correlations, indicating that more informed health care providers and those who place higher importance on managing urinary issues are more likely to agree with positive beliefs about technology use.

Frequency (r=0.497, *P*<.001) and integration of technology into practice (r=0.479, *P*<.001) were moderately correlated with belief agreement, while the concern level (r=0.351, *P*<.001) had the weakest, yet still significant, correlation.

Multiple regression analysis (Table 4) showed the dependence of belief agreement on the various predictors. The perceived burden (β=.284, *P*=.01), device attributes (β=.371, *P*<.001), and integrating technology (β=.312, *P*<.001) were positively associated.

XSL•FO
RenderX

Table 4. Multiple regression model of belief agreement (N=256).

Variable[a]	B[b] (SE)	β	t_{237} Test	*P* value
Health literacy	0.024 (0.046)	0.047	0.509	.61
Perceived burden	0.053 (0.020)	0.284	2.644	.01
Frequency	0.0 (0.034)	–0.001	–0.010	.99
Bladder volume	0.005 (0.052)	0.008	0.095	.92
Device attributes	0.13 (0.036)	0.371	3.573	<.001
Importance	–0.068 (0.035)	–0.216	–1.941	.05
Concern level	–0.038 (0.042)	–0.055	–0.899	.37
Patient receptiveness	0.053 (0.039)	0.121	1.362	.18
Integrating technology	0.236 (0.051)	0.312	4.579	<.001
Support services	0.044 (0.050)	0.076	0.888	.38
Female (vs male)	–0.181 (0.350)	–0.027	–0.516	.61
Registered nurse (vs physician)	0.328 (0.542)	0.037	0.605	.55
Licensed vocational nurse (vs physician)	1.182 (2.309)	0.024	0.512	.61
Medical assistant (vs physician)	0.857 (1.084)	0.038	0.791	.43
Nurse practitioner (vs physician)	–0.499 (1.655)	–0.014	–0.301	.76
Medical technologist (vs physician)	0.727 (0.992)	0.035	0.733	.46
Allied health staff (vs physician)	0.492 (0.662)	0.038	0.743	.46

[a]The reference group for gender is male and for occupation is physician.

[b]B: unstandardized regression coefficient.

Discussion

Principle Findings

The findings of this study highlight key factors influencing belief agreement regarding the implementation of technology. The bivariate analysis demonstrated strong correlations between belief agreement and multiple factors, including device attributes (r=0.650, *P*<.001), the perceived burden (r=0.628, *P*<.001), support services (r=0.622, *P*<.001), patient receptiveness (r=0.612, *P*<.001), bladder volume (r= 0.611, *P*<.001), health literacy (r= 0.591, *P*<.001), importance (r=0.572, *P*<.001), frequency (r=0.497, *P*<.001), integrating technology (r=0.479, *P*<.001), and the concern level (r=0.351, *P*<.001). Effect sizes, as interpreted using Cohen's conventions, ranged from medium to large. These relationships suggest that both individual and systemic factors play significant roles in shaping attitudes toward integrating bladder-monitoring technologies.

The relationship with the perceived burden demonstrated that health care professionals who experience a higher burden when managing urinary issues tend to support technology adoption more strongly. This makes sense; if urinary care is seen as challenging and time-consuming, tools that could ease this burden (eg, wearable bladder sensors) are more appealing. The correlation between bladder volume and belief agreement likely stems from the perceived need for accurate, noninvasive bladder monitoring. Those who frequently assess bladder volume may see the value in advanced technology for continuous or more accessible measurement. The strongest correlation in the bivariate analysis was the relationship with device attributes. This means that the more health care professionals perceive the device as having useful and effective features (eg, ease of use, accuracy, integration with existing workflows), the more they agree that technology is beneficial for urinary care. This indicates that the attributes of the device play a crucial role in acceptance. Regarding patient receptiveness, this indicates that if health care professionals feel that patients will be receptive toward technology, they are more likely to accept it. Finally, support services are essential. The bivariate analysis suggested that without adequate support, health care professionals are less likely to accept technology and will have concerns about failures of technology. The strongest relationships were found with device attributes, the perceived burden, and support services, which means that the design and usability of the device, the urgency of addressing urinary care burdens, and the availability of technical assistance are key determinants in acceptance.

The results of this study highlight several key predictors influencing belief agreement regarding the integration of health care technology. Notably, the perceived burden, device attributes, and the integration of technology into practice were all positively associated with belief agreement, suggesting that when clinicians perceive technology as less burdensome, more functional, and better integrated into workflows, they are more likely to align with positive beliefs about its utility and relevance.

The strongest association was observed with device attributes (β=.371, *P*<.001), emphasizing the critical role of usability, reliability, and functionality in shaping clinician acceptance. This finding supports the existing literature suggesting that

XSL•FO
RenderX

perceived usefulness and ease of use are foundational elements in technology adoption frameworks. When devices are perceived as practical and well designed, clinicians are more likely to view them as beneficial to patient care and professional practice. The integration of technology (β=.312, $P<.001$) was also a significant predictor, indicating that belief agreement is reinforced when technologies are seamlessly embedded in daily workflows rather than introduced as disruptive or isolated tools. This underscores the importance of implementation strategies that consider organizational readiness, change management, and workflow alignment in order to promote sustained adoption.

Comparison With Prior Work

This study builds upon the existing literature on the adoption and integration of technology in health care, particularly in the context of urological care. Prior research has highlighted the potential of technology to improve patient outcomes, enhance efficiency, and reduce the burden on health care professionals. However, these studies have also identified significant barriers to technology adoption. Consistent with previous findings, this study confirms that the perceived burden is a significant concern among health care professionals [14]. Research has consistently shown that poorly designed or implemented technologies can lead to cognitive overload, disrupted workflows, and decreased job satisfaction [22]. These frameworks emphasize the importance of perceived benefits, barriers, and self-efficacy in determining technology adoption. Prior studies have emphasized the critical role of training and support in facilitating technology integration. Effective training programs can enhance self-efficacy, minimize resistance to change, and improve the overall user experience [22,27]. This study reinforces these findings, highlighting the importance of integrating technology into existing systems and providing ongoing technical support. The positive association between device attributes and belief agreement is consistent with the research on technology acceptance. A user-friendly design, ease of use, and perceived usefulness are key factors that influence health care professionals' willingness to adopt new technologies [28]. This study acknowledges that resistance to change and lack of integration into workflows can inhibit technology integration. Prior work has explored strategies for overcoming resistance, such as involving end users in the design process, providing clear communication about the benefits of technology, and addressing concerns about job security [29]. Prior studies have also shown the impact of early detection and trend evaluation using technologies such as tomography, which have increased the detection of symptomatic stones. This study supports the continued use of advanced technologies to improve detection and contribute to a better understanding of the increased incidence of kidney stones.

Strengths

This study builds upon previous research by specifically examining health care professionals' perceptions of urinary issues. This focused approach enables a deeper understanding of the unique challenges and opportunities within this area of health care relating to implementation of technologies. By integrating both quantitative and qualitative data, the study offers a comprehensive perspective on the factors influencing technology adoption. The qualitative analysis provides valuable insights into health care professionals' experiences and viewpoints, complementing the statistical trends identified in the quantitative findings. These findings offer crucial guidance for clinicians, developers, and policymakers aiming to facilitate the adoption of these innovations.

Limitations

Although this study provides valuable insights into health care professionals' perceptions of urinary issues and the acceptability of technology for managing them, it is essential to acknowledge its limitations. There is a potential for sample bias. Participants were recruited through social media platforms (Reddit) and personal networks, which may introduce selection bias. This method might overrepresent individuals who are more tech savvy or active online or other personal factors, potentially skewing the results. The majority of respondents were physicians (66.8%), which limits the representation of other health care professionals, such as nurses and allied health staff, whose perceptions may differ.

Data were collected through a structured survey questionnaire, relying on self-reported perceptions and experiences. This method is subject to recall bias and social desirability bias, as participants may provide responses they believe are more acceptable or desirable. The study used a cross-sectional design, which captures data at a single point in time. This design limits the ability to establish causality or understand how perceptions and attitudes change over time. Longitudinal studies would be beneficial to track the evolution of technology acceptance and its impact on clinical practice. Finally, multiple regression analysis revealed several correlations between independent variables and belief agreement. However, correlation does not equal causation, and there may be other confounding variables not accounted for in the model.

Recommendations

To support the effective adoption of urinary health technologies, 3 key strategies emerge from this study. First, user-centered design must be prioritized. Technologies should be intuitive, reliable, and seamlessly integrated into existing clinical workflows to enhance usability and encourage clinician engagement. When devices are perceived as practical and easy to use, they are more likely to be adopted and sustained in practice.

Structured training and ongoing support were cited as critical by respondents, particularly for technologies that involve patient monitoring or data interpretation. Ensuring that clinical staff receive role-relevant training and have access to timely assistance may reduce resistance and increase confidence in use. Some strategies to bolster support are to establish comprehensive training and a robust support infrastructure. Initial onboarding, ongoing education, and easily accessible technical assistance are critical to building confidence among health care professionals and reducing resistance to new tools. These efforts can also promote consistent and effective use of technology over time.

Finally, implementation strategies must address the perceived burden of technology use by aligning new tools with existing

XSL•FO
RenderX

routines. Technologies that streamline tasks, reduce duplication, and enhance efficiency without increasing the provider burden are more likely to gain clinician acceptance and improve patient care outcomes. Aligning new tools with real-world clinical demands is essential to improving uptake in this specialized area.

Conclusion

This study provides important insights into health care professionals' perceptions of urinary care and the acceptability of wearable and remote technologies to manage such urinary conditions. Our findings highlight that belief agreement regarding the value of urinary health and related technology is significantly influenced by both individual perceptions and system-level factors. Specifically, the strongest correlates of belief agreement include device attributes, the perceived burden, and support services, with device usability, design, and functionality emerging as central to clinician acceptance. Notably, the perceived burden, device attributes, and integration of technology into workflows were also significant predictors of belief agreement in the regression model, emphasizing that practical implementation concerns remain critical for successful adoption.

Although other variables, such as health literacy, bladder volume, and patient receptiveness, were strongly correlated with belief agreement, they did not independently predict belief agreement in the multivariate model. This suggests that although these factors are influential, the most immediate drivers of acceptance are those tied to usability, burden relief, and ease of implementation.

Importantly, the study affirms that health care professionals are more likely to support the use of technology in urologic care when it is perceived as a meaningful solution to existing clinical burdens and when they feel confident in its practical integration. These findings align with prior literature and underscore the need for user-centered design, workflow-sensitive implementation strategies, and ongoing technical support and training.

The several limitations of this study, including the use of Reddit and personal networks for participant recruitment, may introduce selection bias, potentially overrepresenting those more comfortable with technology or engaged online. Additionally, the cross-sectional and self-reported nature of the data limits causal inference and may be subject to recall and social desirability bias. Future research should adopt longitudinal designs and include more diverse professional perspectives, including nurses and allied health staff, to better understand evolving attitudes over time.

Overall, this study offers actionable insights for developers, educators, and health system leaders seeking to promote the thoughtful integration of urinary health technologies that could be further explored in depth. Addressing the perceived burden, prioritizing user-friendly designs, and embedding technology into existing workflows will be critical for maximizing uptake and improving outcomes in urinary care.

Acknowledgments

This research was supported by funding from the California State University (CSU) STEM-NET grant series. The authors gratefully acknowledge the support of CSU STEM-NET, which made this study possible.

Conflicts of Interest

None declared.

References

1. He Y, Zhao J, Wang L, Han C, Yan R, Zhu P, et al. Epidemiological trends and predictions of urinary tract infections in the Global Burden of Disease study 2021. Sci Rep 2025 Feb 08;15(1):4702 [FREE Full text] [doi: 10.1038/s41598-025-89240-5] [Medline: 39922870]
2. Khandelwal C, Kistler C. Diagnosis of urinary incontinence. Am Fam Physician 2013 Apr 15;87(8):543-550 [FREE Full text] [Medline: 23668444]
3. Coyne KS, Sexton CC, Thompson CL, Milsom I, Irwin D, Kopp ZS, et al. The prevalence of lower urinary tract symptoms (LUTS) in the USA, the UK and Sweden: results from the Epidemiology of LUTS (EpiLUTS) study. BJU Int 2009 Aug;104(3):352-360. [doi: 10.1111/j.1464-410X.2009.08427.x] [Medline: 19281467]
4. Abrams P, Andersson K, Apostolidis A, Birder L, Bliss D, Brubaker L, members of the committees. 6th International Consultation on Incontinence. Recommendations of the International Scientific Committee: evaluation and treatment of urinary incontinence, pelvic organ prolapse and faecal incontinence. Neurourol Urodyn 2018 Sep;37(7):2271-2272. [doi: 10.1002/nau.23551] [Medline: 30106223]
5. Ben Arous M, Haddar I, Truong A, Ayena JC, Ouakrim Y, El Kamel L, et al. Non-invasive wearable devices for urinary incontinence detection—a mini review. Front Sens 2023 Nov 9;4:1279158. [doi: 10.3389/fsens.2023.1279158]
6. Nasrabadi MZ, Tabibi H, Salmani M, Torkashvand M, Zarepour E. A comprehensive survey on non-invasive wearable bladder volume monitoring systems. Med Biol Eng Comput 2021 Aug 14;59(7-8):1373-1402. [doi: 10.1007/s11517-021-02395-x] [Medline: 34258707]
7. Beglaryan M, Petrosyan V, Bunker E. Development of a tripolar model of technology acceptance: hospital-based physicians' perspective on EHR. Int J Med Inform 2017 Jun;102:50-61. [doi: 10.1016/j.ijmedinf.2017.02.013] [Medline: 28495348]

XSL•FO
RenderX

8. Garavand A, Aslani N, Nadri H, Abedini S, Dehghan S. Acceptance of telemedicine technology among physicians: a systematic review. Inform Med Unlocked 2022;30:100943. [doi: 10.1016/j.imu.2022.100943]
9. Lambert SI, Madi M, Sopka S, Lenes A, Stange H, Buszello C, et al. An integrative review on the acceptance of artificial intelligence among healthcare professionals in hospitals. NPJ Digit Med 2023 Jun 10;6(1):111 [FREE Full text] [doi: 10.1038/s41746-023-00852-5] [Medline: 37301946]
10. Gagnon MP, Orruño E, Asua J, Abdeljelil AB, Emparanza J. Using a modified Technology Acceptance Model to evaluate healthcare professionals' adoption of a new telemonitoring system. Telemed J E Health 2012 Jan;18(1):54-59 [FREE Full text] [doi: 10.1089/tmj.2011.0066] [Medline: 22082108]
11. Li J, Talaei-Khoei A, Seale H, Ray P, Macintyre CR. Health care provider adoption of eHealth: systematic literature review. Interact J Med Res 2013 Apr 16;2(1):e7 [FREE Full text] [doi: 10.2196/ijmr.2468] [Medline: 23608679]
12. Hu PJ, Chau PY, Sheng ORL, Tam KY. Examining the Technology Acceptance Model using physician acceptance of telemedicine technology. J Manag Inf Syst 2015 Dec 02;16(2):91-112. [doi: 10.1080/07421222.1999.11518247]
13. Gagnon M, Desmartis M, Labrecque M, Car J, Pagliari C, Pluye P, et al. Systematic review of factors influencing the adoption of information and communication technologies by healthcare professionals. J Med Syst 2012 Feb;36(1):241-277 [FREE Full text] [doi: 10.1007/s10916-010-9473-4] [Medline: 20703721]
14. Borges do Nascimento IJ, Abdulazeem H, Vasanthan LT, Martinez EZ, Zucoloto ML, Østengaard L, et al. Barriers and facilitators to utilizing digital health technologies by healthcare professionals. NPJ Digit Med 2023 Sep 18;6(1):161. [doi: 10.1038/s41746-023-00899-4] [Medline: 37723240]
15. Oh JK, Lee JY, Eun S, Park JM. New trends in innovative technologies applying artificial intelligence to urinary diseases. Int Neurourol J 2022 Dec;26(4):268-274 [FREE Full text] [doi: 10.5213/inj.2244280.140] [Medline: 36599335]
16. Jacob C, Sanchez-Vazquez A, Ivory C. Social, organizational, and technological factors impacting clinicians' adoption of mobile health tools: systematic literature review. JMIR Mhealth Uhealth 2020 Feb 20;8(2):e15935 [FREE Full text] [doi: 10.2196/15935] [Medline: 32130167]
17. Schomakers E, Lidynia C, Vervier LS, Calero Valdez A, Ziefle M. Applying an extended UTAUT2 model to explain user acceptance of lifestyle and therapy mobile health apps: survey study. JMIR Mhealth Uhealth 2022 Jan 18;10(1):e27095 [FREE Full text] [doi: 10.2196/27095] [Medline: 35040801]
18. AlQudah AA, Al-Emran M, Shaalan K. Technology acceptance in healthcare: a systematic review. Appl Sci 2021 Nov 09;11(22):10537. [doi: 10.3390/app112210537]
19. Champion VL, Skinner CS. The Health Belief Model. In: Glanz K, Rimer BK, Viswanath K, editors. Health Behavior and Health Education: Theory, Research, and Practice, 4th Ed. Hoboken, NJ: Jossey-Bass/Wiley; 2008:45-65.
20. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view. MIS Quarterly 2003;27(3):425. [doi: 10.2307/30036540]
21. Rogers EM. Diffusion of Innovations, Fifth Edition. New York; London; Toronto; Sydney: Free Press; 2003.
22. Yin P, Ou CX, Davison RM, Wu J. Coping with mobile technology overload in the workplace. Internet Res 2018 Oct 02;28(5):1189-1212. [doi: 10.1108/intr-01-2017-0016]
23. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly 1989 Sep;13(3):319. [doi: 10.2307/249008]
24. Shertzer Y. Self-efficacy and means-efficacy as a strategy for reducing resistance to change—transition from frontal learning to asynchronous learning. Appl Psychol Res 2025 Mar 12;4(1):2231. [doi: 10.59400/apr2231]
25. Norman CD, Skinner HA. eHEALS: the eHealth Literacy Scale. J Med Internet Res 2006 Nov 14;8(4):e27 [FREE Full text] [doi: 10.2196/jmir.8.4.e27] [Medline: 17213046]
26. Yen PY, Wantland D, Bakken S. Development of a customizable health IT usability evaluation scale. AMIA Annu Symp Proc 2010 Nov 13;2010:917-921 [FREE Full text] [Medline: 21347112]
27. Jalo H, Pirkkalainen H. Effect of user resistance on the organizational adoption of extended reality technologies: a mixed methods study. Int J Inf Manag 2024 Apr;75:102731. [doi: 10.1016/j.ijinfomgt.2023.102731]
28. Hussain A, Zhiqiang M, Li M, Jameel A, Kanwel S, Ahmad S, et al. The mediating effects of perceived usefulness and perceived ease of use on nurses' intentions to adopt advanced technology. BMC Nurs 2025 Jan 09;24(1):33 [FREE Full text] [doi: 10.1186/s12912-024-02648-8] [Medline: 39789568]
29. Holden RJ, Karsh B. The Technology Acceptance Model: its past and its future in health care. J Biomed Inform 2010 Feb;43(1):159-172 [FREE Full text] [doi: 10.1016/j.jbi.2009.07.002] [Medline: 19615467]

Abbreviations

AI: artificial Intelligence
eHEALS: eHealth Literacy Scale
EHR: electronic health record
HBM: Health Belief Model
IRB: Institutional Review Board
TAM: Technology Acceptance Model

XSL•FO
RenderX

UI: urinary incontinence
UTI: urinary tract infection

Edited by J Sarvestan; submitted 10.03.25; peer-reviewed by R Collier, K Mouloudj, RN Al-Jabali; comments to author 03.04.25; revised version received 30.05.25; accepted 08.07.25; published 24.07.25.

Please cite as:
Zhang N, Zhang N, Wang Y, Wong R, Ma K
Health Care Professionals' Perspectives on Technology Use in Urinary Care: Cross-Sectional Survey-Based Study
JMIR Form Res 2025;9:e73453
URL: https://formative.jmir.org/2025/1/e73453
doi:10.2196/73453
PMID:40705415

©Nicole Zhang, Ni Zhang, Yun Wang, Rachel Wong, Kaylin Ma. Originally published in JMIR Formative Research (https://formative.jmir.org), 24.07.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Evaluating the Construct Validity and Sensitivity to Change of the Klenico Depression Domain in Psychotherapeutic Inpatient Care: Instrument Validation Study

Stefan Reutimann[1], MSc; Jasmin Steiner[1], MSc; Noah Hübscher[1,2], MSc; Ulrich Voderholzer[3,4], Prof Dr Med; Adrian Meule[5], PhD; Mareike Augsburger[1], PhD

[1]University of Zurich Start-Up, Klenico Health AG, Zürich, Switzerland
[2]Department of Health Sciences and Technology, Swiss Federal Institute of Technology, Zurich, Switzerland
[3]Department of Psychiatry and Psychotherapy, Ludwig Maximilian University of Munich, Munich, Germany
[4]Schoen Clinic Roseneck, Prien am Chiemsee, Germany
[5]Department of Psychology, University of Regensburg, Regensburg, Germany

Corresponding Author:
Jasmin Steiner, MSc
University of Zurich Start-Up
Klenico Health AG
Förrlibuckstrasse 190
Zürich, CH-8005
Switzerland
Phone: 41 44 533 13 90
Email: jasmin.steiner@klenico.com

Abstract

Background: The accurate diagnosis of mental disorders, such as depression, requires comprehensive, valid, and reliable tools to ensure evidence-based treatments and effective outcome monitoring. Existing diagnostic practices often lack standardization, leading to missed comorbidities and variable diagnostic accuracy. The Klenico system is an innovative, web-based diagnostic tool that integrates patient self-reports with clinical validations by mental health professionals. This system covers a broad spectrum of mental disorders, including depression.

Objective: This research aimed to evaluate the psychometric properties of the Klenico Depression Domain (KDD), the component of the Klenico system that measures depressive symptomatology, in a real-world clinical setting. Specifically, the evaluation focused on the assessment of its construct validity, internal consistency, and sensitivity to change in symptom severity.

Methods: Anonymized data from 496 inpatients with mental disorders collected between 2019 and 2022 were analyzed. Patients completed the KDD alongside parts of the Patient Health Questionnaire (PHQ), Beck Depression Inventory (BDI-II), and Satisfaction With Life Scale (SWLS) at both admission and discharge. Internal consistency was measured using Cronbach α. Exploratory factor analysis was conducted to examine the factor structure. Construct validity was assessed via Pearson correlations with PHQ-9 and BDI-II, while divergent validity was tested against the PHQ Somatic Symptoms Scale (PHQ-15), PHQ–Generalized Anxiety Disorder-7, and SWLS. Sensitivity to change was evaluated using paired 1-tailed *t* tests, effect sizes, and repeated measures correlations.

Results: The KDD demonstrated excellent internal consistency (Cronbach α=0.91 at admission and 0.93 at discharge). Factor analysis revealed a 7-factor structure encompassing dimensions like “inadequacy,” “anhedonia,” and “self-hatred,” aligning with core depressive symptoms outlined in the *International Statistical Classification of Diseases, Tenth Revision*. The correlations with the convergent questionnaires PHQ-9 (r=0.68; P<.001) and BDI-II (r=0.70; P<.001) were high. While the KDD showed a moderate correlation with the divergent PHQ-15 (r=0.35; P<.001), it was more strongly associated with the divergent SWLS (r=–0.51; P<.001) and Generalized Anxiety Disorder-7 (r=0.51; P<.001). Sensitivity to change was high, with significant reductions in KDD scores for patients with improved symptoms (t_{27}=5.36, P<.001; Cohen d=0.79) and high repeated measures correlation with both the BDI-II (r=0.61; P<.001) and the PHQ-9 (r=0.59; P<.001).

Conclusions: The KDD shows promise as a reliable and valid instrument for diagnosing depression and monitoring treatment outcomes in psychotherapeutic settings. Its alignment with *International Statistical Classification of Diseases, Tenth Revision*

XSL•FO
RenderX

diagnostic criteria and sensitivity to symptom change underlines its potential utility. These findings highlight the Klenico system's potential to enhance clinical diagnostics by addressing current gaps in mental health care, thus improving diagnostic accuracy and consistency. Further research is recommended to validate its performance across different populations and settings.

(JMIR Form Res 2025;9:e50504) doi:10.2196/50504

KEYWORDS

depression diagnosis; Klenico system; Klenico Depression Domain; sensitivity to change; psychometric properties; construct validity; mental health assessment; clinical outcome monitoring

Introduction

Mental health disorders represent a broad spectrum of conditions with significant impacts on individuals' lives globally [1]. These disorders impair both personal and social functioning [2] and pose substantial challenges to public health systems [3]. For example, depression affects a significant portion of the European population, with a lifetime prevalence of 12% and an annual prevalence of about 7% [4,5], imposing considerable economic burdens of roughly €3000 (≈US $3549) per patient per year [6]. This highlights the critical need for qualitative diagnosis and treatment strategies across the spectrum of mental disorders. However, the diagnosis of mental disorders remains challenging due to common comorbidities among different disorders [7,8] and their complex structures. For instance, depression is a highly heterogeneous disorder characterized by a wide variety of different presentations and a broad range of symptoms [9].

In current clinical practice, the quality of diagnostic procedures is often compromised by a lack of standardization and comprehensiveness. Health professionals predominantly tend to rely on unstructured, open interviews for diagnosis and classification [10-12], which compromises validity and reliability [13,14], thereby impeding evidence-based diagnoses [15]. Standardized tools like structured interviews are rarely used in real settings, as they are considered too time-consuming [10,11]. Alongside structured interviews, self-report questionnaires have emerged as an additional standardized method of assessment, seeing more frequent application in routine diagnostic procedures. Although these questionnaires are valid and time-saving tools when implemented for their intended fields of use, such as screening, severity assessment, and course measurement, they do generally not cover the whole spectrum of symptoms [16]. Focusing solely on one particular disorder domain, such as depression, using a disorder-specific questionnaire may lead to missing critical comorbidities. Moreover, self-report questionnaires lack consistency, as they measure a wide variety of different constructs and show inhomogeneity in the represented diagnostic criteria. For instance, the most common depression self-report questionnaires revealed only a 39% similarity in their covered symptoms [17]. Therefore, self-assessments and their cutoff scores alone should not be relied on for diagnosis [18]. It is necessary to obtain further clarifications on comorbidities and symptom duration through structured diagnostic interviews [19]. Alongside initial diagnostics, there is a significant lack of standardized measurement techniques for assessing outcomes [20-22]. However, this step is indispensable for the precise determination of treatment success [23].

The described lack of quality standardization and comprehensiveness in clinical practice causes mental disorders to often be unrecognized or misdiagnosed in different settings [15,24-28]. For instance, studies comparing routine diagnoses made by unstructured, unstandardized diagnostic procedures with standardized diagnoses made by semistructured interviews—the gold standard in diagnostics—among psychiatric inpatients and outpatients found that depression was correctly identified in only half of the cases. The agreement for most other disorders was also limited or poor [14,15,29]. Additionally, comorbidity is often missed in routine diagnostics [30]; for example, the frequent comorbidity between depression and anxiety was detected in fewer than half of the cases using unstandardized routine diagnostics [29].

These inadequate diagnostic practices, among other possible reasons such as poor access to therapy [31], can lead to ineffective therapeutic interventions. This situation may culminate in unwarranted patient distress and increased demands on health care resources [32]. This is reflected by the fact that only 60% of diagnosed depression patients are provided with evidence-based therapy according to guidelines in primary care in Germany [33], and only one-third of mental health treatments in Spain meet minimal adequacy criteria in both specialty and general medical care [34], with similar findings in other countries and for other mental disorders [35-37].

To enable efficient and adequate treatments, diagnostic procedures need to assess the entire symptomatology of mental disorders in a standardized way, allow conclusions about individual symptoms, and be able to comprehensively record potential comorbidities by querying all relevant disorders. In addition, it is important to adequately measure the course of symptoms to evaluate the success of treatments [23]. However, they also have to be practical, efficient, user-friendly, and feasible in clinical practice due to time constraints [38].

Digital technologies offer a promising solution for the described situation. These tools not only facilitate the diagnostic process but also offer the capability for self-administered assessments, thereby reducing the burden on healthcare professionals and enhancing the accessibility of diagnostic services by overcoming conventional barriers. Through the application of adaptive testing, a wide array of symptoms can be efficiently evaluated, optimizing both time and comprehensiveness. Furthermore, digital tools can empower professionals in standardized diagnostics by supporting their decision-making processes.

To this end, the Klenico system was developed as a comprehensive, web-based diagnostic tool that covers the following mental disorder domains: anxiety disorders, disorders

XSL•FO
RenderX

of adult personality and behavior, eating disorders, mental and behavioral disorders due to psychoactive substance use, depression, mania, obsessive-compulsive disorder, reaction to severe stress and adjustment disorders, and psychotic disorders. Furthermore, the system includes aspects of autism spectrum disorder, attention-deficit/hyperactivity disorder, and dementia. Its items are directly derived from the diagnostic criteria of the *International Statistical Classification of Diseases, Tenth Revision* (*ICD-10*) [39]. The Klenico Depression Domain (KDD) investigated in this research covers all *ICD-10* depressive disorders under F32 and F33, as well as dysthymia (F34.1), providing a comprehensive representation of depressive symptomatology.

More specifically, Klenico consists of a web-based, adaptive self-report module with a total of 496 items, including a preceding screening as well as subsequent clinical validation in the form of a diagnostic interview conducted by mental health professionals. The detailed process of the Klenico system is illustrated in Figure 1. Lustig [40] provides more details on the development of the Klenico system. For clinical validation, the software creates an interactive interview guide based on responses to self-reports. This allows the professional to assess the symptomatology reported by the patient and reevaluate it externally through their expertise by asking questions and making observations. The combination of self-report and external assessment is crucial for standardized, multimodal diagnostics and serves as the ideal procedure. This integration facilitates efficient diagnoses of depression and other disorders that are easily implemented in everyday clinical practice and can relieve the burden on professionals while still maintaining a standardized and comprehensive diagnostic approach. Based on the results of the clinical validation, the system generates final automated suggestions for the most likely diagnoses, ensuring strict adherence to the diagnostic criteria outlined in the *ICD-10*. Moreover, the manifestations of the individual symptoms are clearly visually presented on a so-called symptom map, which enables targeted assessment at the symptom level and reveals important comorbidities. Furthermore, the assessments can be repeatedly performed throughout the course of treatment, whereby symptom change is made visible, thereby enabling treatment outcomes to be monitored.

Figure 1. Flowchart depicting the entire process of the Klenico system. ICD-10: International Statistical Classification of Diseases, Tenth Revision.

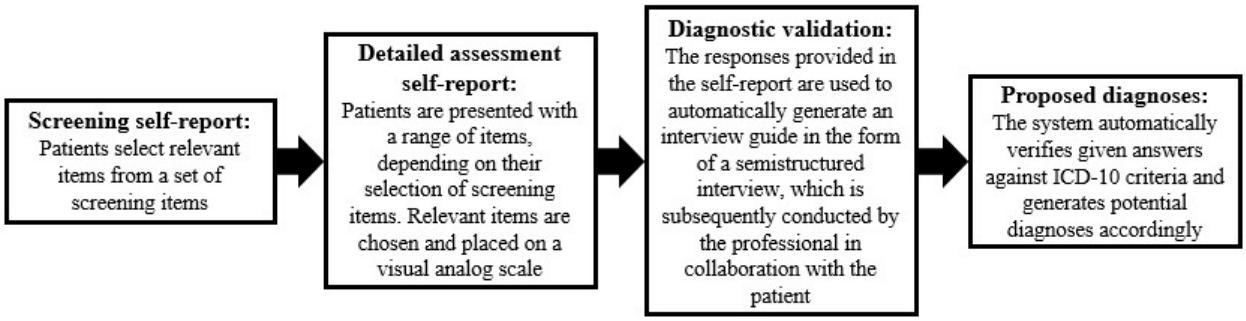

The validity of the Klenico self-report module was preliminarily confirmed on an outpatient sample in a previous study. The study showed that the diagnostic results from the Klenico self-report module aligned well with those obtained from the Structured Clinical Interview for *Diagnostic and Statistical Manual of Mental Disorders* (*DSM*), *Fourth Edition*, thus demonstrating its criterion validity. Furthermore, preliminary indications of its convergent and divergent validity were also established through the comparison with common self-report questionnaires [41]. The purpose of this assessment was to confirm these findings and further validate the Klenico system by laying the focus on its depression domain (KDD). Research was focused on the KDD's reliability, construct validity, and sensitivity to change by considering the measured severity of the symptomatology. This is of substantial importance for ensuring the quality of the diagnostic tool and thus enabling valid, adequate, and standardized diagnostics of depression and mental disorders in general.

Methods

Procedure

The data used in the present research were collected between 2019 and 2022 in a German psychotherapeutic inpatient clinic, which specializes in treating patients from a range of diagnostic groups. The clinic's standard diagnostic procedure was administered to all patients, which provided the foundation for the data used in this study. The Klenico system was added provisionally to these standard procedures as an extension to the usual care.

Upon admission, all inpatients received their routine diagnoses from the clinic based on an initial interview and received routine clinical self-report questionnaires (Beck Depression Inventory [BDI-II], Patient Health Questionnaire [PHQ]–9, PHQ-15, PHQ–Generalized Anxiety Disorder-7 [GAD-7], and Satisfaction With Life Scale [SWLS]) to further clarify their symptoms. Afterward, all patients were informed about the procedure and were asked to sign informed consent to allow their routine data and data from the Klenico assessment to be used for scientific studies and publications. Patients then received access to the Klenico self-report module.

Patients then completed the Klenico self-report module on a device of their choice (computer or tablet) and the clinics' routine self-report questionnaires within the first 2 weeks of their stay. The German version of the Klenico system was used. After completion, mental health professionals conducted the Klenico clinical validation module together with the patient in their next session. At discharge, the procedure was repeated while patients again were asked to answer the clinic's routine

XSL•FO
RenderX

self-questionnaires and the Klenico self-report module. Additionally, the Klenico clinical validation module was performed again. No diagnoses were made in the data collection based on the Klenico system.

Mental health professionals consisted of licensed psychotherapists, psychotherapists in training under supervision (with a minimum of 1 year of professional experience), and cotherapists. Professionals received an introductory session and a demonstration of the use of the Klenico system prior to their initial diagnostic validation. They were also provided with instructions on how to interpret and elucidate the results. The data were fully anonymized.

Measures

BDI-II Questionnaire

The BDI-II was part of the clinic's routine diagnostic procedure to evaluate depressive symptom severity and to serve as a measure of progress at discharge. The BDI-II [42] evaluates the severity of symptoms of depression in the past 2 weeks. It includes 21 items that are summed up to a maximum total score of 63. The BDI-II has good validity and reliability [43], including test-retest reliability [44], and is sensitive to symptom change [45]. The sum score was calculated. Cronbach α coefficient was 0.91 in this analysis.

PHQ-D Questionnaire

The PHQ-D [46] is a self-report instrument for recording mental disorders. The questionnaire demonstrates construct validity based on the number of doctor visits, psychosocial impairment, and incapacity to work, as well as excellent criterion validity (85% sensitivity and 70% specificity) when compared with Structured Clinical Interview for *DSM-IV, Axis I* diagnoses. Both professionals and patients show high acceptance of the PHQ-D, with a satisfaction rate greater than 90% [47]. The short form (option C according to the manual), which screens for depressive (PHQ-9), generalized anxiety (GAD-7), and somatoform (PHQ-15) symptoms, was used in the data collection as part of the clinic's routine diagnostic procedure at admission and discharge to evaluate symptom severity and to measure progress. The PHQ-9 has good test-retest reliability [48] as well as sensitivity to symptom change [49]. Sum scores of the PHQ-9 (maximum score=27), PHQ-15 (maximum score=30), and GAD-7 (maximum score=21) were calculated. Cronbach α coefficients for the PHQ-9, PHQ-15, and GAD-7 were 0.84, 0.78, and 0.83, respectively.

SWLS Questionnaire

The SWLS was part of the clinic's routine diagnostic procedure upon admission and is a 5-item questionnaire for measuring life satisfaction [50]. Items are summed up to a maximum score of 35, indicating extreme satisfaction. The SWLS has factorial validity as well as reasonable convergent and divergent validity indicated through correlations with other indicators of quality of life [51,52]. The sum score was calculated. Cronbach α coefficient was 0.86.

Klenico (Including the KDD)

The Klenico system is a web-based software designed to assist professionals in diagnosing mental disorders. It comprises 2 modules: a self-report followed by clinical validation. The general process of the Klenico system is depicted in Figure 1. Initially, the self-report module presents patients with 55 screening items, from which they select the relevant ones. Depending on the screening items selected, up to 496 items may then be adaptively presented to the patient for a more detailed assessment. The patient again selects the applicable ones and rates each previously selected item by dragging and dropping it on a visual analog scale (ranging from 0=low to 100=high) to assess symptom severity. Upon completion by the patient, the process transitions to the professional. The results from the self-report are graphically displayed on a symptom map, and an interview guide is generated for use in the subsequent clinical validation. This clinical validation involves a semistructured interview where the professional assesses the self-report from an external perspective. This includes posing questions based on the self-reported items, now reformulated for external inquiry, alongside additional external assessment items. These items are rated on a 4-point Likert scale. The system then compares the responses with *ICD-10* diagnostic criteria and provides the professional with suggested diagnoses.

In the data collection used for this assessment, the Klenico study version was used, in which all available 496 items are presented in the detailed assessment of the self-report, regardless of the previously selected screening items.

The KDD assessed in this assessment consists of a total of 31 items, with 29 items included in the self-report module. The clinical validation module reiterates these 29 items as external interview questions, with 2 additional observational items (on psychomotoric agitation and retardation) made by the professional. Means of the KDD for the clinical validation module were calculated where possible. For patients who solely answered the self-report module, means for the self-report were calculated consequently. Cronbach α coefficient was 0.91. Item names, means and SDs of the different KDD items, as well as an exemplary selection of the item formulations, are presented in Multimedia Appendix 1.

Sample

Fully anonymized data from 496 adult inpatients with mental disorders were used (337 female patients and 159 male patients). Data from discharge were available for 59 patients. Data from the Klenico clinical validation module were available for 445 patients at admission and 42 patients at discharge. The mean age was 38.9 (SD 15) years, and the mean duration of stay was 70.2 (SD 38.6) days. A total of 391 patients received a diagnosis of depression. The frequencies of the different depression diagnoses are reported in Table 1 (for a full overview of all given diagnoses, see Multimedia Appendices 2 and 3 [somatic diagnoses]).

XSL•FO
RenderX

Table 1. Overview of the given depression diagnosis with corresponding International Statistical Classification of Diseases, Tenth Revision (ICD-10) codes. The diagnoses were established through the initial interview as part of the routine diagnostic procedures followed at the clinic. Listed frequencies are relative to the number of total given depression diagnoses.

Depression diagnosis	*ICD-10* code	Frequency (n=391), n (%)
Mild depressive episode	F32.0	4 (1)
Moderate depressive episode	F32.1	75 (19.2)
Severe depressive episode without psychotic symptoms	F32.2	41 (10.5)
Severe depressive episode with psychotic symptoms	F32.3	1 (0.3)
Recurrent depressive disorder, current episode mild	F33.0	7 (1.8)
Recurrent depressive disorder, current episode moderate	F33.1	135 (34.5)
Recurrent depressive disorder, current episode severe without psychotic symptoms	F33.2	115 (29.4)
Recurrent depressive disorder, current episode severe with psychotic symptoms	F33.3	2 (0.5)
Recurrent depressive disorder, currently in remission	F33.4	1 (0.3)
Dysthymia	F34.1	10 (2.6)

Statistical Analysis

Data analysis was performed using RStudio software (version 2021.9.1.372; Posit, PBC). *P* values <.05 were considered statistically significant.

Exploratory factor analyses (EFA) were conducted to explore the underlying factor structure of the KDD using the *lavaan*, *SEMtools*, and *psych* packages. Kaiser Meyer Olkin (KMO) and Bartlett tests were used to test the adequacy of the factor analysis using the functions *cortest.bartlett* and *KMO*. Parallel analysis and fit estimation were used to identify the optimal number of factors using the functions *efaUnrotate* and *fa.parallel*. A maximum likelihood estimator with robust standard errors was applied. Oblique rotation was applied by using the *oblqRotate* function (the R code for the EFA can be found in Multimedia Appendix 4). The model fit was considered adequate when the comparative fit index >0.90, Tucker Lewis coefficient >0.90, standardized root mean square residual <0.10, and root mean square error of approximation <0.08 [53]. Items were assigned to a specific factor if they demonstrated a loading greater than 0.30 on that factor.

Convergent validity for the KDD was tested by performing Pearson correlation analysis with both the PHQ-9 and BDI-II, while divergent validity was assessed by comparison with the PHQ-15, GAD-7, and SWLS. An r value from >0.1 to <0.3 was considered to indicate a small effect, from ≥0.3 to <0.5 was considered to indicate a medium effect, and ≥0.5 was considered to indicate a large effect [54]. Bonferroni-adjustment was performed to correct for multiple testing.

Sensitivity to change of the KDD was examined by categorizing each patient into groups of either improved or unchanged symptoms based on changes in the PHQ-9 or BDI-II sum scores. KDD means for admission and discharge were compared for both improved and unchanged symptom groups by paired 2-tailed *t* tests. Effect sizes and standardized response means were calculated. Additionally, repeated measures correlations between the KDD and the BDI-II and between the KDD and the PHQ-9 were calculated using the *rmcorr* package [55]. Bonferroni adjustment was performed to correct for multiple testing.

A dropout analysis was conducted to compare patients who completed the Klenico system and the PHQ-9 and BDI-II questionnaires at discharge with those who did not complete these assessments. The analysis examined potential differences in demographic characteristics, questionnaire scores (KDD, PHQ-9, BDI-II, and SWLS), and length of stay using 2-tailed *t* tests or Wilcoxon rank sum tests. This approach aimed to identify any systematic biases between the groups.

Post hoc power analyses were conducted to assess the statistical power of the results (using the functions *pwr.r.test*, *pwr.t.test*, and *power.rmcorr*).

Ethical Considerations

Ethical approval was not required for the studies because according to German law, no ethical approval was required for the use of the Klenico system in routine diagnostic procedures. With reference to the Council Directive 93/42/EEC of the European Union and the according German legislation (Medizinproduktegesetz (MPG) § 23b), the test-related use of a medical device with certification (Conformité Européenne marking) does not require an ethics vote if it is carried out within the scope of its intended purpose [56]. Furthermore, the use of the Klenico system in routine diagnostic procedures falls within the scope of the responsibility of the attending physician (Berufsordnung der Ärztekammern §15). According to the ethics committee of the Ludwig Maximilian University Munich, such data collection with the aim of evaluating the performance of an accepted, evidence-based practice (quality assurance) does not require ethical approval. Finally, no ethical approval was required for analysis of the data, since retrospective studies using already available, anonymized data are exempt from needing ethics approval, according to the guidelines of the institutional review board at the Ludwig Maximilian University Munich. The study was conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study. Data were anonymized. No compensation was provided.

XSL•FO
RenderX

Results

Reliability

Excellent internal consistency was demonstrated by the KDD both at admission (Cronbach α=0.91) and at discharge (Cronbach α=0.93).

Exploratory Factor Analysis

A KMO value of 0.91 and a significant Bartlett test indicated eligibility for EFA. Parallel test and fit estimation favored a 7-factor solution (Multimedia Appendices 5 and 6). The 7-factor model revealed good fit (χ^2_{269}=382.912, comparative fit index=0.971, Tucker Lewis coefficient=0.950, standardized root mean square residual=0.025, and root mean square error of approximation=0.031). Based on the rotated loadings presented in Table 2, the 7 factors were interpreted and labeled as “inadequacy,” “sleep disturbance,” “affect,” “anhedonia,” “somatization,” “appetite decrease,” and “self-hatred.” The item “feelings of worthlessness” was cross-loaded onto 2 factors. Four items (“psychomotor agitation,” “psychomotor retardation,” “inner tension,” and “indecision”) did not load adequately on any factor.

XSL•FO
RenderX

Table 2. Exploratory factor analysis (EFA) factor structure with rotated loadings of the 31 items of the Klenico Depression Domain (KDD). All reported factor loadings are statistically significant at *P*<.05, unless otherwise specified.

	Inadequacy	Sleep disturbance	Affect	Anhedonia	Somatization	Appetite decrease	Self-hatred
Depressed mood	—[a]	—	–0.338[b]	–0.171	–0.239	—	—
Loss of interest	—	—	—	–0.229	–0.380[b]	—	—
Loss of pleasure	—	—	—	–1.111[b]	—	—	—
Loss of energy	—	—	—	–0.108	–0.723[b]	—	—
Tiredness	—	—	–0.274	—	–0.493[b]	—	—
Difficulty falling asleep	0.143	–0.350[b]	–0.135	—	–0.184	—	0.157
Disturbed sleep	—	–0.833[b]	—	—	—	—	—
Lack of restful sleep	0.220	–0.497[b]	–0.165	—	–0.221	—	—
Early morning awakening	—	–0.602[b]	—	—	—	—	—
Hypersomnia	0.163	0.147	—	—	–0.435[b]	—	—
Weight loss	—	—	—	—	—	0.455[b]	—
Loss of appetite	—	—	—	—	—	0.591[b]	—
Increased appetite	—	—	—	—	–0.416[b]	–0.241	—
Psychomotor agitation	—	—	—	—	—	—	—
Psychomotor retardation	—	—	—	—	0.150	—	—
Inner tension	—	–0.236	—	—	–0.210	—	—
Indecision	–0.264	–0.137	—	—	–0.228	—	—
Diminished ability to think	—	—	–0.194	—	–0.310[b]	—	—
Rumination	–0.233	–0.148	–0.301[b]	—	—	—	—
Feelings of hopelessness	—	—	–0.673[b]	—	—	—	—
Feelings of worthlessness	–0.395[b]	—	—	—	—	—	0.300[b]
Loss of self-esteem	–0.547[b]	—	—	—	—	—	—
Unreasonable feelings of guilt	—	—	—	—	—	—	0.816[b]
Unreasonable feelings of self-reproach	—	—	—	—	—	—	0.869[b]
Low self-esteem	–0.529[b]	—	—	—	—	—	0.184
Crying	–0.183	—	–0.263	—	—	—	0.128
Despair	—	—	–0.780[b]	—	—	—	—
Loss of affective reactivity	—	—	–0.186	–0.389[b]	—	—	—
Circadian rhythms: worse in the morning	—	–0.239	—	—	–0.210	—	—
Loss of libido	—	–0.161	—	—	–0.291	—	—
Suicidality	—	—	–0.297	—	—	—	—

[a]Not applicable.

[b]Factor loadings greater than 0.3.

Convergent and Divergent Validity

Regarding convergent validity, the KDD showed very high correlations with both the BDI-II (*r*=0.70; *P*<.001) and PHQ-9 (*r*=0.68; *P*<.001). Regarding divergent validity, moderate associations were apparent with the PHQ-15 (*r*=0.35; *P*<.001). In contrast, the GAD-7 (*r*=0.51; *P*<.001) and the SWLS (*r*=–0.51; *P*<.001) were strongly correlated with the KDD.

XSL•FO
RenderX

Sensitivity to Change

The KDD mean scores of the improved symptom group (n=28) demonstrated a significant difference (t_{27}=5.36, P<.001) between admission (mean 36.85, SD 16.43) and discharge (mean 23.90, SD 18.54), with a high effect size (Cohen d=0.79) and a standardized response mean of 1.01. The means of the other measurements can be found in Multimedia Appendix 7. Values of the unchanged symptom group could not be reliably calculated due to the small group size (n=4). A total of 27 patients were excluded from the analysis due to missing BDI-II and PHQ-9 scores or deteriorated symptoms. Additionally, the KDD revealed a high repeated measures correlation with both the BDI-II (r=0.61; P<.001) and the PHQ-9 (r=0.59; P<.001). The dropout analysis revealed significant differences between patients who completed the Klenico system and standard questionnaires at discharge and those who did not. Completers showed higher scores in the KDD, BDI-II, and SWLS, and a longer length of stay (see Multimedia Appendix 7 for detailed results). Post hoc analysis revealed statistical power above 0.80.

Discussion

Principal Findings

This research assessed the validity of the KDD by focusing on its construct validity, reliability, and sensitivity to change.

Regarding KDDs' underlying structure, the EFA suggested a 7-factor model (Table 2). The items on "psychomotor retardation" and "agitation" as well as for "inner tension," "indecision," "crying," "circadian rhythms: worse in the morning," "loss of libido," and "suicidality," did not adequately load on any of the factors.

The 7 factors of the KDD roughly reflect the 10 diagnostic criteria for depression as outlined in the classification of *ICD-10*. Table 3 summarizes which KDD factors reflect which *ICD-10* diagnostic criteria. The KDD factors "inadequacy," "sleep disturbance," "self-hatred," "anhedonia," "affect," and "appetite decrease" each reflect a single *ICD-10* criterion. The factor "somatization" maps on 3 criteria.

Table 3. Overview of the International Statistical Classification of Diseases, Tenth Revision (ICD-10) diagnostic criteria reflected by the Klenico Depression Domain (KDD) factors. Overview of the ICD-10 diagnostic criteria reflected by the KDD factors.

KDD factors	*ICD-10* diagnostic criteria
Inadequacy	• Loss of confidence or self-esteem
Sleep disturbance	• Sleep disturbance any type
Anhedonia	• Loss of interest or pleasure in activities that are normally pleasurable
Self-hatred	• Unreasonable feelings of self-reproach or excessive and inappropriate guilt
Affect	• Depressed mood
Somatization	• Decreased energy or increased fatiguability • Complaints or evidence of diminished ability to think or concentrate... • Change in appetite (decrease or increase) with corresponding weight change
Appetite decrease	• Change in appetite

These results indicate that the *ICD-10* criteria for depression are subject to a complex and heterogeneous structure composed of different symptom dimensions. This is underlined by the fact that the latent structure of depression is widely disputed [57]. A recent study on the factor structure of the *DSM-IV* major depression criteria showed comparable results to our findings, as it demonstrated a multidimensional symptom structure with 5 factors for the 24 criteria [58].

The multiple mappings of the "somatiziation" factor on the *ICD-10* criteria and the inadequate loading of some items can be explained by the sample-specific characteristics, the Klenico system, and the structure of depression.

The factors "somatization" and "appetite decrease" reflect those *ICD-10* criteria categorized as somatic symptoms in depression that have been assigned to a specific somatic factor in various previous studies [57]. This confirms that the heterogeneous group of somatic symptoms has an underlying common structure that is separate from the rest of the affective and cognitive symptoms of depression.

A possible explanation for the split into 2 separate somatic factors is that loss of appetite and the resulting weight loss are related to the overall severity of depression and may be determined by the severity of other symptoms, such as anhedonia, which is usually most pronounced in severely affected patients, potentially leading to a small variance and a possible floor effect [59]. Finally, the separation found between general somatic symptoms and sleep disturbance does not contradict the factorial validity of the KDD, since previous studies have shown that the 2 symptom dimensions impose separate factors [58].

Effects comparable to those at the factor level may also have occurred at the item level. The lack of significant factor loadings for "psychomotor agitation" and "retardation" may be due to their inclusion only in the clinical validation and not in the Klenico self-report module. These symptoms are more accurately assessed through external observation rather than self-report and are often more pronounced in severe depression [23,60], which can lead to reduced variability and a resulting

XSL•FO
RenderX

floor effect in the data. This smaller variance might hinder their ability to align clearly with a specific factor in the analysis. The same reasoning accounts for the items for "suicidality" and "loss of libido," both of which were just below the cutoff of 0.3—"suicidality" with loads just not on "affect" and "loss of libido" with loads just not on "somatization." Their prevalences are correlated with the severity of depression [59,61], and acute suicidality is mainly an indication for treatment in a psychiatric clinic. Since the present sample comes from a psychotherapeutic clinic in which suicidality is an exclusion criterion, it is to be expected that such patients were less present.

"Inner tension," "indecision," "circadian rhythms: worse in the morning," and "crying" did not load strongly on a single factor, instead showing distributed loadings across multiple factors. "Inner tension" aligns with both the "sleep disturbance" and "somatization" factors, indicating its role in exacerbating both sleep issues and physical symptoms. "Indecision" correlates with the "somatization" and "inadequacy" factors, highlighting its connection to cognitive impairments and reduced self-confidence typical in depressive episodes. "Circadian rhythms: worse in the morning" describes the symptom of morning mood lows in depression. Due to the wording, it is likely that this item is answered in a rather undifferentiated way and can therefore be associated by patients with both poor sleep and physical effects. "Crying" is a universal symptom that may not load strongly on any specific factor, likely due to its general association with various forms of distress across the depressive spectrum. It possibly associated with multiple factors such as "affect" and "self-hatred," reflecting its role as a common expression of emotional pain, sadness, and feelings of low worth. This pattern suggests these items intersect various dimensions of depression, making their specific factor alignment more complex. Overall, the findings confirm that depression is a multidimensional construct and suggest that the KDD is determined by the heterogeneous structure of the *ICD-10* criteria. Since the structure of the KDD faithfully mirrors the *ICD-10* criteria, this offers initial support for its factorial validity.

Concerning convergent validity, large associations were found between the KDD and the corresponding BDI-II and PHQ-9, as expected. Additionally, the KDD showed medium associations with divergent PHQ-15. Exceptions were the GAD-7 and SWLS, with high-range correlations.

Similar to the structure of depression, the factor structure for different diagnostic tools is widely debated. The number of factors found for the BDI-II ranges between 2 and 4 [44] and for the PHQ-9 ranges between 1 and 2 [62]. For both, various compositions of the different items loading on the different factors were found. However, although the KDD reflects the factor structure according to *ICD-10* criteria, it showed significantly high associations with these 2 instruments, thus indicating the measurement of a comparable construct of depression.

The results regarding divergent validity are not surprising, as comorbidity between depression and various anxiety disorders is highly frequent [63], and depressive symptoms show structural overlap with anxiety disorders [64]. This is further supported by other studies demonstrating strong intercorrelations between the GAD-7 and PHQ-9 [65], as well as significant correlations between the GAD-7 and BDI-II. [66]

Moreover, depressive disorders are strongly negatively associated with life satisfaction and subjective well-being [67]. The strong negative correlation with the SWLS aligns with theoretical expectations and highlights the substantial impairment in life satisfaction that accompanies depressive symptomatology.

The results therefore do not counteract the construct validity of the KDD but rather support its ability to capture the broader context of psychological distress and well-being.

Regarding sensitivity to change, the KDD was able to detect improvements in symptom severity according to PHQ-9 or BDI-II sum scores and showed strong associations with changes in the BDI-II and PHQ-9 sum scores. This demonstrates that the KDD is highly sensitive to symptom change and is therefore able to monitor treatment outcomes, similar to the commonly used self-report questionnaires.

Overall, the results show that the KDD is a reliable and valid instrument for the diagnosis of depression according to *ICD-10* criteria while being able to adequately recognize improvements in symptom severity. In contrast to routine diagnostic procedures, the KDD enables a state-of-the-art diagnostic procedure that includes both self-report and clinical validation by experts and is additionally able to assign comorbidities by covering all relevant mental disorders (n=12).

A major strength of this study is that the KDD was tested in a large clinical sample under real-world conditions. It could thus be demonstrated that the KDD measures adequately in patients with frequent comorbidities in clinical reality, providing real-life evidence of its validity.

Limitations and Outlook

A limitation of the study was the small sample size at discharge. Despite the fact that the post hoc power analysis has shown sufficient power for the *t* tests and the repeated measurement correlation, the findings regarding the sensitivity of change should be considered preliminary. Furthermore, the analysis of the sensitivity of change in the unchanged symptom group was not feasible due to the limited sample size and requires further research. In addition, the patients who used the Klenico tool at discharge exhibited greater symptom severity than those who did not. A reason for this higher representation may be that those more heavily burdened patients perceived greater relevance in completing the exit assessment, leading to higher motivation. The composition of the sample also imposed certain limitations. On the one hand, less severe cases were to be expected in the psychotherapeutic setting, and on the other hand, the design of the study could further lead to the possible exclusion of severe depression cases. This was shown by the fact that, for example, "severe depressive episode with psychotic symptoms" were extremely rare, but also "mild depressive episode" were relatively uncommon. Since the completion of the Klenico assessment requires a certain level of motivation and cognitive abilities, patients affected by severe concentration problems and anhedonia may have faced difficulties in

XSL•FO
RenderX

participation. These circumstances led to certain restrictions in the factor analysis, for example, the nonloading of the item "psychomotor retardation." Additionally, the sample exhibited a gender imbalance, with a majority (68%) being female. This mirrors the real-world scenario where women are more commonly affected by mental disorders compared with men [68]. However, this imbalance could also impact the outcome, given that different genders may exhibit variations in the manifestation of depressive symptoms [69]. Moreover, the age distribution of our participants, concentrated primarily in the younger adult segment and in the segment aged around 55 years, suggests a demographic skew that may have influenced the clinical profile observed. For these reasons, the use of the psychotherapeutic sample in this study might restrict the generalizability of the findings to broader populations, particularly individuals with severe depression. Therefore, it would be valuable to assess the functionality and generalizability of the KDD in different settings that involve patients with more severe symptoms and with a balanced gender and age distribution. This would provide a more comprehensive understanding of the measure's effectiveness across diverse populations.

This analysis focused on examining Klenico assessment of symptom severity, and Klenico-suggested diagnoses were not examined. Therefore, this should be added in future analyses. Finally, the current validation was limited to the German version of the Klenico system. Further validation studies are necessary to assess the system in other languages, such as English. While the EFA provided initial insights into the factor structure of the KDD, these results need to be confirmed through a confirmatory factor analysis. Additionally, the factor structures of other Klenico domains also require assessment to ensure the comprehensive validity of the system across different mental health conditions.

Conclusions

The KDD presents itself as an accurate tool for measuring depressive symptoms and a valuable tool for measuring depressive symptoms and monitoring treatment outcomes in psychotherapeutic settings. As an integral part of the Klenico system, it comprehensively covers the whole range of depression symptoms according to *ICD-10* and might therefore be a comprehensive alternative for the diagnosis and course measurement of depression. The current assessment represents an important next step in validating the psychometric properties of the Klenico system. Therefore, it may help to avoid misdiagnosis and incorrect therapies through standardized and evidence-based diagnostic procedures, thus reducing health care costs and minimizing the burden on the patients.

Acknowledgments

This work was funded by Klenico Health AG.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

SR was responsible for formal analysis, drafting the original manuscript, and reviewing and editing the manuscript. JS contributed to project administration, drafting the original manuscript, and reviewing and editing the manuscript. NH participated in drafting the original manuscript. MA oversaw supervision, validation, conceptualization, and reviewing and editing of the manuscript. UV contributed to the conceptualization and provided resources. AM handled data curation and participated in reviewing and editing the manuscript.

Conflicts of Interest

SR, NH, and MA were formerly employed by Klenico Health AG, while JS remains currently employed. We declare that this did not affect data collection or analysis. UV and AM report no conflicts of interest.

Multimedia Appendix 1
Item names, means and SDs of the different KDS items, as well as an exemplary selection of the item formulations.
[PDF File (Adobe PDF File), 142 KB - formative_v9i1e50504_app1.pdf]

Multimedia Appendix 2
Frequencies of ICD-10 diagnostic categories. Groups were formed by assigning each individual diagnosis of each patient to the corresponding disorder domain. Multiple diagnoses are possible. Listed frequencies are relative to the number of total given diagnoses. ICD-10: International Statistical Classification of Diseases, Tenth Revision.
[PDF File (Adobe PDF File), 126 KB - formative_v9i1e50504_app2.pdf]

Multimedia Appendix 3
The most frequent somatic diagnoses in the sample, in descending order.

XSL•FO
RenderX

[PDF File (Adobe PDF File), 96 KB - formative_v9i1e50504_app3.pdf]

Multimedia Appendix 4
R code for the exploratory factor analyses.
[DOCX File , 24 KB - formative_v9i1e50504_app4.docx]

Multimedia Appendix 5
Exploratory factor analysis fit estimations for different number of factors.
[PDF File (Adobe PDF File), 100 KB - formative_v9i1e50504_app5.pdf]

Multimedia Appendix 6
Scree plot of exploratory factor analysis factor estimation.
[PDF File (Adobe PDF File), 134 KB - formative_v9i1e50504_app6.pdf]

Multimedia Appendix 7
Means, SDs, medians, and IQRs of demographic parameters and standard questionnaires at admission for the whole patient sample, patients who were measured at discharge, and patients who were not measured at discharge.
[PDF File (Adobe PDF File), 138 KB - formative_v9i1e50504_app7.pdf]

References

1. Steel Z, Marnane C, Iranpour C, Chey T, Jackson JW, Patel V, et al. The global prevalence of common mental disorders: a systematic review and meta-analysis 1980-2013. Int J Epidemiol 2014;43(2):476-493 [FREE Full text] [doi: 10.1093/ije/dyu038] [Medline: 24648481]
2. Hirschfeld RMA, Montgomery SA, Keller MB, Kasper S, Schatzberg AF, Möller HJ, et al. Social functioning in depression: a review. J Clin Psychiatry 2000;61(4):268-275. [doi: 10.4088/jcp.v61n0405] [Medline: 10830147]
3. Christensen MK, Lim CCW, Saha S, Plana-Ripoll O, Cannon D, Presley F, et al. The cost of mental disorders: a systematic review. Epidemiol Psychiatr Sci 2020 Aug 18;29:e161 [FREE Full text] [doi: 10.1017/S204579602000075X] [Medline: 32807256]
4. Busch MA, Maske UE, Ryl L, Schlack R, Hapke U. Prävalenz von depressiver Symptomatik und diagnostizierter Depression bei Erwachsenen in Deutschland. Prevalence of depressive symptoms and diagnosed depression among adults in Germany: results of the German Health Interview and Examination Survey for Adults (DEGS1). Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz. Federal Health Gazette Health Research Health Protection 2013 May;56(5-6):733-739. [doi: 10.1007/s00103-013-1688-3] [Medline: 23703492]
5. Lim GY, Tam WW, Lu Y, Ho CS, Zhang MW, Ho RC. Prevalence of depression in the community from 30 countries between 1994 and 2014. Sci Rep 2018 Feb 12;8(1):2861 [FREE Full text] [doi: 10.1038/s41598-018-21243-x] [Medline: 29434331]
6. Eden JL, Konnopka A, König HH. Costs of depression in Germany - systematic review. Psychiatr Prax 2021 Sep;48(6):290-300. [doi: 10.1055/a-1399-4430] [Medline: 33773504]
7. Avenevoli S, Swendsen J, He JP, Burstein M, Merikangas KR. Major depression in the national comorbidity survey-adolescent supplement: prevalence, correlates, and treatment. J Am Acad Child Adolesc Psychiatry 2015 Jan;54(1):37-44.e2 [FREE Full text] [doi: 10.1016/j.jaac.2014.10.010] [Medline: 25524788]
8. Kessler RC, Chiu WT, Demler O, Merikangas KR, Walters EE. Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. Arch Gen Psychiatry 2005 Jun;62(6):617-627 [FREE Full text] [doi: 10.1001/archpsyc.62.6.617] [Medline: 15939839]
9. Fried EI, Epskamp S, Nesse RM, Tuerlinckx F, Borsboom D. What are 'good' depression symptoms? comparing the centrality of DSM and non-DSM symptoms of depression in a network analysis. J Affect Disord 2016;189:314-320 [FREE Full text] [doi: 10.1016/j.jad.2015.09.005] [Medline: 26458184]
10. Bruchmüller K, Margraf J, Suppiger A, Schneider S. Popular or unpopular? therapists' use of structured interviews and their estimation of patient acceptance. Behav Ther 2011;42(4):634-643. [doi: 10.1016/j.beth.2011.02.003] [Medline: 22035992]
11. Aboraya A. Use of structured interviews by psychiatrists in real clinical settings: results of an open-question survey. Psychiatry (Edgmont) 2009;6(6):24-28 [FREE Full text] [Medline: 19724758]
12. Anderson DA, Paulosky CA. A survey of the use of assessment instruments by eating disorder professionals in clinical practice. Eat Weight Disord 2013;9(3):238-241. [doi: 10.1007/bf03325075]
13. Miller PR. Inpatient diagnostic assessments: 2. interrater reliability and outcomes of structured vs. unstructured interviews. Psychiatry Res 2001;105(3):265-271. [doi: 10.1016/S0165-1781(01)00318-3]
14. Andreas S, Theisen P, Mestel R, Koch U, Schulz H. Validity of routine clinical DSM-IV diagnoses (Axis I/II) in inpatients with mental disorders. Psychiatry Res 2009;170(2-3):252-255. [doi: 10.1016/j.psychres.2008.09.009] [Medline: 19896721]

XSL•FO
RenderX

15. Shear MK, Greeno C, Kang J, Ludewig D, Frank E, Swartz HA, et al. Diagnosis of nonpsychotic patients in community clinics. Am J Psychiatry 2000;157(4):581-587. [doi: 10.1176/appi.ajp.157.4.581] [Medline: 10739417]
16. Serra F, Spoto A, Ghisi M, Vidotto G. Formal psychological assessment in evaluating depression: a new methodology to build exhaustive and irredundant adaptive questionnaires. PLoS One 2015;10(4):e0122131 [FREE Full text] [doi: 10.1371/journal.pone.0122131] [Medline: 25875359]
17. Newson JJ, Hunter D, Thiagarajan TC. The heterogeneity of mental health assessment. Front Psychiatry 2020;11:76 [FREE Full text] [doi: 10.3389/fpsyt.2020.00076] [Medline: 32174852]
18. Ferentinos P, Paparrigopoulos T, Rentzos M, Zouvelou V, Alexakis T, Evdokimidis I. Prevalence of major depression in ALS: comparison of a semi-structured interview and four self-report measures. Amyotroph Lateral Scler 2011;12(4):297-302. [doi: 10.3109/17482968.2011.556744] [Medline: 21428731]
19. Maurer DM. Screening for depression. Am Fam Physician 2012;85(2):139-144 [FREE Full text] [Medline: 22335214]
20. Zimmerman M, McGlinchey JB. Why don't psychiatrists use scales to measure outcome when treating depressed patients? J Clin Psychiatry 2008;69(12):1916-1919. [doi: 10.4088/jcp.v69n1209] [Medline: 19192467]
21. Hatfield DR, Ogles BM. Why some clinicians use outcome measures and others do not. Adm Policy Ment Health 2007;34(3):283-291. [doi: 10.1007/s10488-006-0110-y] [Medline: 17211715]
22. Jensen-Doss A, Haimes EMB, Smith AM, Lyon AR, Lewis CC, Stanick CF, et al. Monitoring treatment progress and providing feedback is viewed favorably but rarely used in practice. Adm Policy Ment Health 2018;45(1):48-61 [FREE Full text] [doi: 10.1007/s10488-016-0763-0] [Medline: 27631610]
23. DGPPN B, KBV , AWMF , Härter M, Schorr S, Schneider F, editors. S3-Leitlinie/Nationale Versorgungsleitlinie Unipolare Depression-Langfassung. S3 Guideline/National Care Guideline on Unipolar Depression-Long Version. Berlin, Heidelberg: Springer; Mar 10, 2017.
24. J. Rakofsky J, W. Dunlop B. The over-under on the misdiagnosis of bipolar disorder: a systematic review. CPSR 2015;11(4):222-234. [doi: 10.2174/1573400512666151117211523]
25. Vermani M, Marcus M, Katzman MA. Rates of detection of mood and anxiety disorders in primary care. Prim Care Companion CNS Disord 2011:27211. [doi: 10.4088/pcc.10m01013]
26. Takara K, Kondo T, Kuba T. How and why is autism spectrum disorder misdiagnosed in adult patients? - from diagnostic problem to management for adjustment -. Ment Health Fam Med 2015 Nov 30;11(02):73-88 [FREE Full text] [doi: 10.25149/1756-8358.1102011]
27. Wahl K, Kordon A, Kuelz KA, Voderholzer U, Hohagen F, Zurowski B. Obsessive-compulsive disorder (OCD) is still an unrecognised disorder: a study on the recognition of OCD in psychiatric outpatients. Eur Psychiatry 2010;25(7):374-377. [doi: 10.1016/j.eurpsy.2009.12.003] [Medline: 20627468]
28. Fusar-Poli L, Brondino N, Politi P, Aguglia E. Missed diagnoses and misdiagnoses of adults with autism spectrum disorder. Eur Arch Psychiatry Clin Neurosci 2022;272(2):187-198 [FREE Full text] [doi: 10.1007/s00406-020-01189-w] [Medline: 32892291]
29. Zander E, Wyder L, Holtforth MG, Schnyder U, Hepp U, Stulz N. Validity of routine clinical diagnoses in acute psychiatric inpatients. Psychiatry Res 2018;259:482-487. [doi: 10.1016/j.psychres.2017.11.004] [Medline: 29154169]
30. Zimmerman M. A review of 20 years of research on overdiagnosis and underdiagnosis in the Rhode Island methods to improve diagnostic assessment and services (MIDAS) project. Can J Psychiatry 2016;61(2):71-79 [FREE Full text] [doi: 10.1177/0706743715625935] [Medline: 27253697]
31. Herzog DP, Wagner S, Ruckes C, Tadic A, Roll SC, Härter M, et al. Guideline adherence of antidepressant treatment in outpatients with major depressive disorder: a naturalistic study. Eur Arch Psychiatry Clin Neurosci 2017;267(8):711-721. [doi: 10.1007/s00406-017-0798-6] [Medline: 28421334]
32. Stahnke B. A systematic review of misdiagnosis in those with obsessive-compulsive disorder. J Affect Disord Rep 2021;6:100231. [doi: 10.1016/j.jadr.2021.100231]
33. Trautmann S, Beesdo-Baum K. The treatment of depression in primary care. Dtsch Arztebl Int 2017;114(43):721-728 [FREE Full text] [doi: 10.3238/arztebl.2017.0721] [Medline: 29143731]
34. Fernández A, Haro JM, Codony M, Vilagut G, Martínez-Alonso M, Autonell J, et al. Treatment adequacy of anxiety and depressive disorders: primary versus specialised care in Spain. J Affect Disord 2006;96(1-2):9-20. [doi: 10.1016/j.jad.2006.05.005] [Medline: 16793140]
35. Wang PS, Aguilar-Gaxiola S, Alonso J, Angermeyer MC, Borges G, Bromet EJ, et al. Use of mental health services for anxiety, mood, and substance disorders in 17 countries in the WHO world mental health surveys. Lancet 2007;370(9590):841-850 [FREE Full text] [doi: 10.1016/S0140-6736(07)61414-7] [Medline: 17826169]
36. Wang PS, Demler O, Kessler RC. Adequacy of treatment for serious mental illness in the United States. Am J Public Health 2002;92(1):92-98. [doi: 10.2105/ajph.92.1.92] [Medline: 11772769]
37. Lora A, Conti V, Leoni O, Rivolta AL. Adequacy of treatment for patients with schizophrenia spectrum disorders and affective disorders in Lombardy, Italy. Psychiatr Serv 2011;62(9):1079-1084. [doi: 10.1176/ps.62.9.pss6209_1079] [Medline: 21885588]
38. Aboraya A, Nasrallah HA, Elswick DE, Ahmed E, Estephan N, Aboraya D, et al. Measurement-based care in psychiatry-past, present, and future. Innov Clin Neurosci 2018;15(11-12):13-26 [FREE Full text] [Medline: 30834167]

XSL•FO
RenderX

39. The ICD-10 classification of mental and behavioural disorders: clinical descriptions and diagnostic guidelines. World Health Organization. 1992. URL: https://apps.who.int/iris/handle/10665/37958 [accessed 2025-02-26]
40. Lustig L. Die Klenico-Items: theoretische Herleitung und empirische Fundierung eines neuen Ansatzes in der psychiatrischen Diagnostik. The Klenico Items: Theoretical Derivation and Empirical Foundation of a New Approach in Psychiatric Diagnostics. Zurich, Switzerland: University of Zurich; 2020.
41. Reutimann S, Milanovic D, Gulewitsch MD, Augsburger M. Preliminary validation of the Klenico diagnostic software self-report module through comparison with the diagnostic gold standard in an outpatient routine clinical sample. Health Psychol Behav Med 2023;11(1):2244576 [FREE Full text] [doi: 10.1080/21642850.2023.2244576] [Medline: 37663014]
42. Beck AT, Steer RA, Brown GK. Beck Depression Inventory (BDI-II). London, UK: Pearson; 1996.
43. Sprinkle SD, Lurie D, Insko SL, Atkinson G, Jones GL, Logan AR, et al. Criterion validity, severity cut scores, and test-retest reliability of the Beck Depression Inventory-II in a university counseling center sample. J Couns Psychol 2002;49(3):381-385. [doi: 10.1037/0022-0167.49.3.381]
44. Wang YP, Gorenstein C. Psychometric properties of the Beck Depression Inventory-II: a comprehensive review. Braz J Psychiatry 2013;35(4):416-431 [FREE Full text] [doi: 10.1590/1516-4446-2012-1048] [Medline: 24402217]
45. Meesters Y, Sijbrandij JJ, Visser E, de Beurs E. Sensitivity to change of the Beck Depression Inventory versus the inventory of depressive symptoms. J Affect Disord 2021;281:338-341 [FREE Full text] [doi: 10.1016/j.jad.2020.12.036] [Medline: 33341647]
46. Löwe B, Spitzer RL, Zipfel S, Herzog W. Gesundheitsfragebogen für patienten (PHQ-D). komplettversion und kurzform. testmappe mit manual, fragebögen, schablonen. 2. auflage. Health Questionnaire for Patients (PHQ-D). complete version and short form. test folder with manual, questionnaires, templates. 2nd edition. GESIS – Leibniz Institute for the Social Sciences. 2002. URL: https://search.gesis.org/publication/zis-L%C3%B6weSpitzerZipfelHerzog2002Gesundheitsfragebogen [accessed 2025-01-31]
47. Gräfe K, Zipfel S, Herzog W, Löwe B. Screening psychischer störungen mit dem “Gesundheitsfragebogen für Patienten (PHQ-D)“. Screening for mental disorders using the ‘Health Questionnaire for Patients (PHQ-D)’. Diagnostica 2004;50(4):171-181. [doi: 10.1026/0012-1924.50.4.171]
48. Zuithoff NP, Vergouwe Y, King M, Nazareth I, van Wezep MJ, Moons KG, et al. The Patient Health Questionnaire-9 for detection of major depressive disorder in primary care: consequences of current thresholds in a crosssectional study. BMC Fam Pract 2010 Dec 13;11(1):98 [FREE Full text] [doi: 10.1186/1471-2296-11-98] [Medline: 21144018]
49. Löwe B, Kroenke K, Herzog W, Gräfe K. Measuring depression outcome with a brief self-report instrument: sensitivity to change of the Patient Health Questionnaire (PHQ-9). J Affect Disord 2004;81(1):61-67. [doi: 10.1016/S0165-0327(03)00198-8] [Medline: 15183601]
50. Diener E, Emmons RA, Larsen RJ, Griffin S. The Satisfaction With Life Scale. J Pers Assess 1985 Feb;49(1):71-75. [doi: 10.1207/s15327752jpa4901_13] [Medline: 16367493]
51. Hinz A, Conrad I, Schroeter ML, Glaesmer H, Brähler E, Zenger M, et al. Psychometric properties of the Satisfaction With Life Scale (SWLS), derived from a large German community sample. Qual Life Res 2018 Jun;27(6):1661-1670. [doi: 10.1007/s11136-018-1844-1] [Medline: 29589212]
52. Glaesmer H, Grande G, Braehler E, Roth M. The German version of the Satisfaction With Life Scale (SWLS). Eur J Psychol Assess 2011;27(2):127-132. [doi: 10.1027/1015-5759/a000058]
53. Marsh HW, Hau KT, Wen Z. In search of golden rules: comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. Struct Equ Model 2004 Jul;11(3):320-341. [doi: 10.1207/s15328007sem1103_2]
54. Cohen J. Statistical Power Analysis for the Behavioral Sciences. London, United Kingdom: Routledge; 2013.
55. Bakdash JZ, Marusich LR. Repeated measures correlation. Front Psychol 2017;8:456 [FREE Full text] [doi: 10.3389/fpsyg.2017.00456] [Medline: 28439244]
56. Medical Devices Act. Federal Ministry of Justice and Consumer Protection. 2002. URL: https://tinyurl.com/443hfw8w [accessed 2025-06-213]
57. Bi Y, Wang L, Cao C, Fang R, Li G, Liu P, et al. The factor structure of major depressive symptoms in a sample of Chinese earthquake survivors. BMC Psychiatry 2021 Jan 28;21(1):59 [FREE Full text] [doi: 10.1186/s12888-020-02993-3] [Medline: 33509157]
58. van Loo HM, Aggen SH, Kendler KS. The structure of the symptoms of major depression: factor analysis of a lifetime worst episode of depressive symptoms in a large general population sample. J Affect Disord 2022 Jun 15;307:115-124 [FREE Full text] [doi: 10.1016/j.jad.2022.03.064] [Medline: 35367501]
59. Zimmerman M, Balling C, Chelminski I, Dalrymple K. Understanding the severity of depression: which symptoms of depression are the best indicators of depression severity? Compr Psychiatry 2018 Nov;87:84-88. [doi: 10.1016/j.comppsych.2018.09.006] [Medline: 30282058]
60. Angst J, Gamma A, Benazzi F, Ajdacic V, Rössler W. Does psychomotor agitation in major depressive episodes indicate bipolarity? evidence from the Zurich study. Eur Arch Psychiatry Clin Neurosci 2009;259(1):55-63. [doi: 10.1007/s00406-008-0834-7] [Medline: 18806921]

XSL•FO
RenderX

61. Williams K, Reynolds MF. Sexual dysfunction in major depression. CNS Spectr 2006 Aug;11(8 Suppl 9):19-23. [doi: 10.1017/s1092852900026729] [Medline: 16871134]
62. Lamela D, Soreira C, Matos P, Morais A. Systematic review of the factor structure and measurement invariance of the Patient Health Questionnaire-9 (PHQ-9) and validation of the portuguese version in community settings. J Affect Disord 2020;276:220-233. [doi: 10.1016/j.jad.2020.06.066] [Medline: 32697702]
63. Li Y, Shi S, Yang F, Gao J, Li Y, Tao M, et al. Patterns of co-morbidity with anxiety disorders in Chinese women with recurrent major depression. Psychol Med 2011 Nov 30;42(6):1239-1248. [doi: 10.1017/s003329171100273x]
64. Forbes MK, Sunderland M, Rapee RM, Batterham PJ, Calear AL, Carragher N, et al. A detailed hierarchical model of psychopathology: from individual symptoms up to the general factor of psychopathology. Clin Psychol Sci 2021;9(2):139-168 [FREE Full text] [doi: 10.1177/2167702620954799] [Medline: 33758691]
65. Villarreal-Zegarra D, Paredes-Angeles R, Mayo-Puchoc N, Arenas-Minaya E, Huarcaya-Victoria J, Copez-Lonzoy A. Psychometric properties of the GAD-7 (General Anxiety Disorder-7): a cross-sectional study of the Peruvian general population. BMC Psychol 2024 Apr 02;12(1):183 [FREE Full text] [doi: 10.1186/s40359-024-01688-8] [Medline: 38566138]
66. Johnson SU, Ulvenes PG, Øktedalen T, Hoffart A. Psychometric properties of the General Anxiety Disorder 7-Item (GAD-7) scale in a heterogeneous psychiatric sample. Front Psychol 2019;10:1713 [FREE Full text] [doi: 10.3389/fpsyg.2019.01713] [Medline: 31447721]
67. Gigantesco A, Fagnani C, Toccaceli V, Stazi MA, Lucidi F, Violani C, et al. The relationship between satisfaction with life and depression symptoms by gender. Front Psychiatry 2019;10:419 [FREE Full text] [doi: 10.3389/fpsyt.2019.00419] [Medline: 31258495]
68. Boyd A, van de Velde S, Vilagut G, de Graaf R, O'Neill S, Florescu S, et al. Gender differences in mental disorders and suicidality in Europe: results from a large cross-sectional population-based study. J Affect Disord 2015;173:245-254. [doi: 10.1016/j.jad.2014.11.002] [Medline: 25462424]
69. Winkler D, Pjrek E, Heiden A, Wiesegger G, Klein N, Konstantinidis A, et al. Gender differences in the psychopathology of depressed inpatients. Eur Arch Psychiatry Clin Neurosci 2004 Aug;254(4):209-214. [doi: 10.1007/s00406-004-0471-8] [Medline: 15309388]

Abbreviations

BDI-II: Beck Depression Inventory
DSM: Diagnostic and Statistical Manual of Mental Disorders
EFA: exploratory factor analyses
GAD-7: Generalized Anxiety Disorder-7
ICD-10: International Statistical Classification of Diseases, Tenth Revision
KDD: Klenico Depression Domain
KMO: Kaiser Meyer Olkin
PHQ: Patient Health Questionnaire
SWLS: Satisfaction With Life Scale

Edited by A Mavragani; submitted 04.07.23; peer-reviewed by V Sideropoulos, C Haag, B Guo; comments to author 31.01.24; revised version received 29.04.24; accepted 27.12.24; published 24.07.25.

Please cite as:
Reutimann S, Steiner J, Hübscher N, Voderholzer U, Meule A, Augsburger M
Evaluating the Construct Validity and Sensitivity to Change of the Klenico Depression Domain in Psychotherapeutic Inpatient Care: Instrument Validation Study
JMIR Form Res 2025;9:e50504
URL: https://formative.jmir.org/2025/1/e50504
doi:10.2196/50504
PMID:

©Stefan Reutimann, Jasmin Steiner, Noah Hübscher, Ulrich Voderholzer, Adrian Meule, Mareike Augsburger. Originally published in JMIR Formative Research (https://formative.jmir.org), 24.07.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Acceptability of an Intervention to Prevent Older Adult Mistreatment Among Family Caregivers to Persons With Dementia: Multimethod Pilot Study

Kylie Meyer[1], MSc, PhD; Wenxing Wei[2], BSN, MSW; Jeanine Yonashiro-Cho[3], PhD; Susanna Mage[4], MA, PhD; Sohee Kim[1]; Elliane Irani[1], RN, PhD; Christopher Burant[1], PhD; Zachary Gassoumis[3], PhD; Erin Gentry Lamb[5], PhD; Jaclene A Zauszniewski[1], RN-BC, PhD; Donna Benton[4], PhD

[1]Frances Payne Bolton School of Nursing, Case Western Reserve University, Cleveland, OH, United States

[2]Jack, Joseph, and Morton Mandel School of Applied Social Sciences, Case Western Reserve University, Cleveland, OH, United States

[3]Keck School of Medicine, University of Southern California, Los Angeles, CA, United States

[4]Leonard Davis School of Gerontology, University of Southern California, Los Angeles, CA, United States

[5]Department of Bioethics, School of Medicine, Case Western Reserve University, Cleveland, OH, United States

Corresponding Author:
Kylie Meyer, MSc, PhD
Frances Payne Bolton School of Nursing
Case Western Reserve University
10900 Euclid Ave
Cleveland, OH, 44106
United States
Phone: 1 2163681928
Email: knm77@case.edu

Abstract

Background: Older adult mistreatment occurs in many as one-half of dementia care partners. Psychological mistreatment is the most common form of older adult mistreatment by family caregivers and is known to create mental health morbidities among care recipients. The Knowledge and Interpersonal Skills to Develop Enhanced Relationships (KINDER) intervention is among the first older adult mistreatment prevention interventions focused on family caregivers. KINDER was designed to prevent psychological mistreatment of older adults. Caregivers found the initial asynchronous web-based version (KINDER 1.0) to be acceptable but expressed a desire to engage with other family caregivers. KINDER was revised to integrate 3 facilitated small group discussion sessions conducted by videoconference. This study examines the acceptability of a revised KINDER intervention. This research addresses the extent to which caregivers find a novel approach to older adult mistreatment prevention to be acceptable.

Objective: This study aims to evaluate the acceptability of the revised *KINDER* intervention.

Methods: The investigators conducted semistructured qualitative interviews with a purposive sample of family caregivers following participation in *KINDER* (N=11) and collected postintervention survey data (N=71). The qualitative interview codebook and survey questions were informed by the Theoretical Framework of Acceptability by Sekhon et al. Components of acceptability in this framework include affective attitude, burden, ethicality, intervention coherence, opportunity costs, perceived effectiveness, and self-efficacy at completing activities. Qualitative interviews were coded by 2 independent coders using a thematic analytic approach. Survey data were analyzed using frequencies and percentages.

Results: Of the 98 caregivers who attended *KINDER*, 71 (72%) completed satisfaction surveys. Caregivers reported high levels of overall satisfaction with *KINDER*; 80% (53/66) of participants reported they were “Very Satisfied” with the intervention, and 20% (13/66) indicated they were “Satisfied.” More than 80% of caregivers (56/69, 81%) rated the newly added group discussions as being “Very valuable.” Qualitative findings supported positive attitudes revealed in survey responses. Themes addressed (1) the interventions’ alignment with caregiver values (affective attitude, intervention coherence, ethicality), (2) beliefs about the effectiveness of the program (perceived effectiveness), (3) difficulty participating in the program relative to its perceived overall value (burden, opportunity cost, self-efficacy), and (4) recommendations to further improve the intervention.

XSL•FO
RenderX

Conclusions: These findings indicate that *KINDER* was well received among family caregivers, who reported high levels of satisfaction and positive feedback on its components. The addition of virtual group discussion sessions was particularly valued. The use of multiple data collection methods in this research provided a comprehensive understanding of caregiver experiences. This study contributes to current knowledge by demonstrating the acceptability of a novel intervention to prevent older adult mistreatment by family caregivers to persons with dementia. Future research should focus on testing the efficacy of *KINDER* and exploring its implementation in health and social service settings.

Trial Registration: ClinicalTrials.gov NCT05783102; https://clinicaltrials.gov/study/NCT05783102

(JMIR Form Res 2025;9:e73778) doi:10.2196/73778

KEYWORDS

family caregiving; dementia; elder mistreatment; intervention; acceptability; qualitative

Introduction

Background

Older adult mistreatment is defined by the World Health Organization (WHO) as "a single or repeated act, or lack of appropriate action, occurring within any relationship where there is an expectation of trust, which causes harm or distress to an older person" [1]. Among community-dwelling older adults with dementia, 80% receive informal care at home from family caregivers [2]. Despite providing essential care, caregivers, unfortunately, are among the most likely to engage in older adult mistreatment: An estimated one-third to over one-half of caregivers self-report mistreatment toward the care recipient with dementia [3,4]. Given the high prevalence of mistreatment of older adults by caregivers to persons with dementia, this population is an important group to target in interventions designed to prevent older adult mistreatment. However, little is known about evidence-based approaches to prevent older adult mistreatment by family caregivers, let alone the acceptability of such approaches to potential users. This study considers the acceptability of a psychoeducational intervention to prevent psychological mistreatment of older adults in dementia care contexts.

Psychological mistreatment of older adults accounts for the majority of mistreatment by caregivers and is the most common type of older adult mistreatment outside neglect [5,6]. Behaviors characterizing psychological mistreatment of older adults include verbal and nonverbal behaviors that may inflict emotional pain, fear, or distress in the older person, including actions like making threats, humiliating, or ignoring the older person [7]. Victims of psychological older adult mistreatment face an increased risk of anxiety, depression, and chronic disease [8,9]. Importantly, psychological mistreatment of older adults often co-occurs with other types of mistreatment, such as neglect and physical abuse [10]. With the number of persons with Alzheimer disease and related dementias projected to double by 2060, there is an urgent need to identify and develop effective strategies to prevent caregiver psychological mistreatment toward this population [11].

Knowledge and Interpersonal Skills to Develop Enhanced Relationships (*KINDER*) is a psychoeducational intervention developed to prevent psychological mistreatment by caregivers toward the person with dementia for whom they provide care. The intervention focuses on building healthy relationships between caregivers and their care recipients as a framework to prevent mistreatment, following similar successful approaches in the area of intimate partner violence [12,13]. After identifying barriers to participation in its initial pilot study, *KINDER* recently underwent an additional round of testing to determine the intervention's feasibility, preliminary efficacy, and acceptability [14]. Findings from the second pilot test indicate feasibility, including an 82% retention rate in a pre- and posttest pilot study, as well as preliminary efficacy, including statistically significant decreases in the occurrence of psychological mistreatment of older adults [15]. This study examines the acceptability of the *KINDER* intervention. "Acceptability" is a concept believed to drive adherence, engagement, and future adoption of behavioral intervention and is, therefore, a critical criterion for intervention translation [16].

Interventions to Prevent Older Adult Mistreatment Among Family Caregivers

Few interventions have been tested to prevent mistreatment by caregivers toward persons with dementia, regardless of mistreatment type [17,18]. Even less is known about the acceptability of these interventions for their intended audiences, which is a critical factor in determining participation adherence, implementation, and outcomes. In one of the earliest efforts to test the impact of an information-based supportive caregiver intervention on reducing older adult mistreatment, the *Strategies for Relatives* (*START*) program sought to mitigate caregiver mistreatment of older adults by addressing known mental health mistreatment risk factors affecting caregivers, such as depression, through a supervised manual-based coping program [19]. Despite positive findings of decreased anxiety and depression for caregivers, results from this study showed no differences between the intervention and control group participants on outcomes related to older adult mistreatment. Recently, the *Comprehensive Older Adult and Caregiver Help* (*COACH*) intervention, also aimed at reducing older adult mistreatment, revealed promising results [20]. Caregivers who participated in up to 12 weekly 1-on-1 supportive caregiver coaching sessions, which included topics like goal setting, experienced a statistically significant decrease in mistreatment behaviors compared with an information-only control group condition. Although such findings are promising for determining efficacious approaches to prevent older adult mistreatment in a care context, neither the *START* nor *COACH* program reported on intervention acceptability.

XSL•FO
RenderX

Similarly, the *RISE* intervention combines tailored support, including goal setting, with advocacy for individuals in the Adult Protective Service (APS) system. Consistent with its person-centered approach, at the discretion of the person who experienced mistreatment, the intervention allows for the inclusion of perpetrators of mistreatment, with goals that could include repairing the relationship [21]. Although acceptability data from the older adult who experienced mistreatment and caregivers involved in *RISE* have not been published, a recent study examining APS worker perceptions of *RISE* found that the program was described as being a valuable resource to address older adults' needs that fall outside of or otherwise complement what APS can provide [22]. Among caregivers who engage in mistreatment of older adults, little is known about their willingness to participate in interventions like *COACH* and *RISE*. These interventions nonetheless provide important examples of ongoing innovative approaches researchers are taking to mitigate older adult mistreatment.

The KINDER Intervention

KINDER Development

KINDER was developed to reduce the occurrence of psychological mistreatment of older adults in the context of family caregiving to persons with dementia by focusing on the relationship between the caregiver and care recipient. The goal of *KINDER* is both the primary *and* secondary prevention of psychological mistreatment of older adults; that is, the aim is 2-fold: to prevent mistreatment from occurring in the first place and reduce such behaviors if they are already occurring [17]. Although the intervention focuses on psychological mistreatment of persons with dementia by a caregiver, the content also addresses other forms of mistreatment, like neglect and physical mistreatment. This is important given the likelihood of more than one type of mistreatment occurring simultaneously within the care dyad [10]. To distinguish between the first and the more recent version of *KINDER*, we refer to the initial version as "*KINDER* 1.0" and the most recent version as "*KINDER*."

KINDER 1.0

Structure and Components

KINDER 1.0 was tested between March 2020 and September 2021. This first version was an 8-week, web-based intervention [14,23]. The intervention was delivered asynchronously to give caregivers flexibility with completing lessons. Caregivers received weekly reminders to log into the web-based portal to complete independent lessons. Content was developed based on existing literature about risk factors for low-quality care, including potentially harmful care and older adult mistreatment, related to the care relationship (eg, prior occurrence of mistreatment) [24], as well as findings from qualitative focus groups and 1-on-1 interviews with caregivers about their experiences managing relationship strain [25]. Topics included education about community resources and seeking help, self-care, common brain changes with dementia, responding to behavioral symptoms of dementia, recognizing older adult mistreatment, communicating with persons with dementia, managing difficult topics around safety and independence, managing mental health conditions, and coping with stress related to the care relationship. Each lesson in *KINDER* 1.0 included a brief story-based video informed by qualitative focus group findings, a reading that elaborated on topics from videos (eg, communicating with a person with dementia), a reading quiz, and a journal reflection. In addition, caregivers were advised each week to engage in a pleasant, self-care activity.

Summary of Findings From KINDER 1.0

Of the 27 caregivers who enrolled in *KINDER* 1.0, only 7 completed the intervention (ie, each lesson through Lesson 8) [14]. Although a low proportion of completers is not unusual for an asynchronously delivered intervention, the investigators sought to better understand low treatment fidelity [26]. In-depth, semistructured qualitative interviews were conducted with these caregivers to understand their experiences while participating in *KINDER* 1.0. Findings illustrated overall acceptability of the program among completers, who reported liking the program's authenticity and emphasis on self-care. Only one caregiver identified discomfort with the content, noting that she could not relate to depictions of caregivers becoming angry and that this made her feel uneasy. Technical problems with the web-based platform (eg, too complex of a display to track pleasant activities) were also identified as a likely barrier to participation. When asked how to improve the program beyond the technical challenges, caregivers expressed a desire to interact with other family caregivers during their participation in the program.

Revisions to KINDER 1.0

Based on caregivers' feedback on *KINDER* 1.0 and prior studies demonstrating the value of group-based interventions, *KINDER* 1.0 was revised as *KINDER* to include virtual synchronous and group-based components over an 8-week period [27,28]. We integrated 3 group-based discussion sessions held via videoconference into independent lessons. Small group discussion sessions occurred during Week 1, Week 4, and Week 8. Discussion sessions were facilitated by 2 members of the study team, including the intervention developers (a geropsychologist and a gerontologist) or a nurse interventionist who received training prior to the intervention on how to deliver the program with fidelity.

Participants received 2 weekly emails reminding them to complete independent lesson activities. Participants also received a telephone introduction from one of the facilitators prior to the start of the first discussion session (ie, Week 0) and 2 check-in calls during Week 3 and Week 6 to support adherence. Web content from *KINDER* was adapted into a workbook, available as a PDF and as a printed hard copy mailed to participants, depending on the caregiver's preference. Videos could be accessed by participants using a URL or a QR code found in their workbook. The first figure in the publication by Meyer et al [15] illustrates the structure of the revised intervention.

Purpose of This Study

This multimethod study aimed to evaluate the acceptability of the revised *KINDER* intervention among caregivers of persons with dementia to support future implementation. Evidence of acceptability is particularly important for the *KINDER* intervention, given the sensitive nature of the topics covered, which could deter future participation among caregivers. To

XSL•FO
RenderX

determine intervention acceptability, we administered satisfaction surveys to caregivers following their participation in the intervention and conducted in-depth qualitative interviews with program completers to elucidate multiple components of acceptability and understand opportunities for program improvement. Findings will inform future *KINDER* refinements and guide decisions about whether the intervention should proceed to efficacy testing.

Methods

Study Design

This study used a design incorporating posttest data collection only, including collection of qualitative survey data and qualitative interview data. Of the 151 caregivers registered to participate in *KINDER*, 98 caregivers completed the program when it was offered in a community-based service setting from April 2023 to March 2024 [15]. Enrollment in the feasibility study was optional; 72% (71/98) of caregivers who attended the program indicated interest in participating, of which 63% (45/71) were enrolled in a pre- and posttest feasibility study. In the posttest survey, participants were asked about the acceptability of the *KINDER* intervention via survey. Caregivers who were not enrolled in the pre- and posttest study also received an invitation to complete a satisfaction survey at the end of the program. Of the 53 caregivers who participated in *KINDER* outside the feasibility study, 64% (34/53) completed the satisfaction survey. Participants were eligible if they registered for and attended the *KINDER* intervention, which was open to family caregivers to persons with dementia. Although a sample size was identified for the pre- and posttest trial to evaluate preliminary efficacy, no sample size was predetermined for collection of acceptability data. Cohorts were conducted until the pre- and posttest sample size for enrollment was reached.

This study followed the model of feasibility by Bowen et al [29], wherein intervention acceptability is one component of feasibility; “acceptability” is the focus of this study [29]. Additional information on the pre- and posttest study design and additional feasibility data, such as preliminary efficacy, can be found in the study by Meyer et al [15]. Multimedia Appendix 1 illustrates the creation of the analytic sample for survey participants and describes differences in how caregivers received survey invitations, as well as distribution of incentives for caregivers enrolled in the pre- and posttest study and those who completed the *KINDER* program outside the trial.

Conceptual Model of Acceptability

Data collection tools to assess intervention acceptability were informed by the Theoretical Framework of Acceptability (TFA) by Sekhon et al [30]. This model includes 7 components of acceptability with intervention activities: affective attitude, burden, ethicality, intervention coherence, opportunity costs, perceived effectiveness, and self-efficacy at completing intervention activities. A strength of this model is that it is based on a systematic review of reviews on measuring acceptability within health care interventions and thus is comprehensive in its scope. A limitation of this model is that some of the identified components overlap heavily with other components of feasibility, such as usability and practicality, thus we also report on some aspects of feasibility that are strongly related to acceptability [29,31]. Definitions of each component can be found in Table 1.

Table 1. Knowledge and Interpersonal Skills to Develop Enhanced Relationships (KINDER) codebook definitions.

Code	Definition
Acceptability	A multifaceted construct that reflects the extent to which people delivering or receiving an intervention consider it to be appropriate, based on anticipated or experiential cognitive and emotional responses to the intervention
Affective attitude	How an individual feels about the intervention
Intervention coherence	The extent to which the participant understands the intervention and how it works
Perceived effectiveness	The extent to which the intervention is perceived as likely to achieve its purpose
Burden	The perceived amount of effort that is required to participate in the intervention
Opportunity cost	The extent to which benefits, profits, or values must be given up to engage in the intervention
Self-efficacy	The participant’s confidence that they can perform the behaviors required to participate in the intervention

Generally, these components reflect participants’ perceptions of the intervention’s appropriateness based on their cognitive understanding of and emotional experiences during the intervention. Using a multicomponent framework of acceptability can uncover complex and nuanced responses to an intervention, allowing participants to endorse certain aspects while recognizing opportunities for improvement in others. Incorporating necessary adjustments identified in the findings from an acceptability study can improve intervention feasibility and overall acceptance and has driven efforts to continually improve the *KINDER* program [30].

Data Collection

Survey Data

Quantitative data on intervention acceptability were collected through postintervention survey. Surveys were self-administered online using the Research Electronic Data Capture (REDCap) system and included 27 items that were developed for this study [32]. Items were broadly based on components of the TFA to address multiple aspects of acceptability and were iteratively developed through discussion with the investigator team, as a validated measure of acceptability using the TFA model does

XSL•FO
RenderX

not yet exist [33]. Nevertheless, we found that TFA components were applicable to caregiver experiences in the first feasibility study that relied on qualitative interviews [14] and found that an author-generated acceptability measure based upon the model was preferable to other validated acceptability measures that are narrower in scope, especially given the ability to triangulate with qualitative interviews [34]. Reflecting overall acceptability, caregivers were asked about their overall satisfaction with *KINDER* and how likely they would be to recommend the program to other caregivers. Response options were recorded on a 5-point Likert scale, ranging from "Strongly disagree" (0) to "Strongly agree" (4). Reflecting the components' "perceived effectiveness" and "opportunity costs" in the TFA model, caregivers also rated how valuable they found each component of the intervention to be, with response options including "Not at all valuable" (0), "Somewhat valuable" (1), and "Very valuable" (2). Detailed survey data about caregiver experiences while participating in the discussion sessions were collected, given these sessions were a newly added feature to the intervention. Caregivers rated their level of agreement with statements about the discussion sessions (eg, "I felt comfortable with sharing experiences with others") on a 5-point Likert scale, ranging from "Strongly disagree" (0) to "Strongly agree" (4). Items evaluated multiple components of the TFA, such as perceived effectiveness, as well as "burden" given the time demands of attending an synchronous discussion session. Survey questions did not require a response. Reponses could be modified until the survey was submitted.

Interview Data

Purposive sampling was used to identify 23 caregivers enrolled in the feasibility study to participate in a one-time, in-depth, semistructured qualitative interview following participation in *KINDER*. Although this sampling technique risks selection bias, we used this approach to support the inclusion of perspectives of caregivers from various genders, races, ethnicities, and kin-relationships to the care recipient. Of the 23 caregivers invited, 11 agreed to participate. Initial interviews were conducted by the first author (KM), a female PhD faculty member with prior experience leading qualitative interview studies and who co-delivered the intervention. Later interviews were administered by an undergraduate student research assistant (SK) trained by the principal investigator who had no prior relationship with participants. All interviews were 1-on-1 with participants, except one interview where with both the principal and research assistant interviewers attended for training purposes. Interviews were held over Zoom videoconferencing software and video recorded for accuracy. The interview guide used to evaluate the acceptability of the revised *KINDER* intervention aligns with the one previously used to evaluate the initial version of the program, *KINDER* 1.0 (see Meyer et al [14]). Topics included caregiver perceptions on how their participation affected care relationships, the perceived skills developed while participating, the most and least valuable components, any potential discomfort with intervention modules, barriers to participation, and recommendations on how to improve the program. Interviews lasted an average of 37.8 (SD 21.2) minutes.

Data Analysis

Survey data were analyzed using descriptive statistics, including frequencies, percentages, means, and measures of variance. Surveys were included even if responses to some items were missing. All analyses were conducted in Stata 18.0 [35]. Qualitative interview recordings were professionally transcribed before analysis. Transcripts were analyzed using a modified version of the thematic analytic approach by Braun and Clark [36]. This analytic approach was selected for its flexibility, including the ability to code both latent and manifest content. To establish credibility, a component from the criteria by Lincoln and Guba [37] to establish trustworthiness in qualitative research, transcripts were reviewed prior to applying the codebook [38]. Confirmability, another component of trustworthiness, was supported using a codebook that included clear definitions of each code [38]. Coding was primarily deductive and adhered to the 7 components of the TFA [30]. By using the TFA framework to inform survey questions, the codebook facilitated triangulation of method and study dependability. Two investigators, KM and WW, independently coded each interview using the codebook (see Table 1). Differences in coding were discussed until agreement was reached after each transcript was independently coded. The lead investigator reviewed excerpts under each code, identified key themes, and prepared a draft write-up of the findings. Data saturation was reached as coders identified no need to add new codes nor novel applications of codes. The resulting themes were validated by the second coder to ensure their accuracy, again supporting credibility in the analysis. All coding was conducted in NVivo 14 [39].

Ethical Considerations

The trial was registered at ClinicalTrials.gov and received approval from the Case Western Reserve University Institutional Review Board (STUDY20221333). Additional protocol information is available on ClinicalTrials.gov. Caregivers enrolled in the feasibility study provided written consent to participate in this research. For caregivers who were not enrolled in the feasibility study, a study information sheet was provided at the beginning of the online survey. All qualitative interview participants received an information sheet prior to their scheduled interview and provided verbal consent at the start of the interview. Participants were notified of the voluntary nature and estimated length of study activities, of the name of the principal investigator, of the purpose of study activities including the investigators' reasons for conducting interviews, and that their de-identified data would be retained indefinitely. Access to data was limited to IRB-approved study team members. Consolidated Criteria for Reporting Qualitative Research (COREQ) and Consolidated Standards of Reporting Trials (CONSORT) checklists are available in Multimedia Appendix 4 and Multimedia Appendix 5, respectively [40,41].

XSL•FO
RenderX

Results

Quantitative Findings

Sample Characteristics

Caregivers enrolled in the feasibility study (N=45) were, on average, 59 (SD 14.4) years of age (range: 34 to 87 years), and most indicated they were of the female sex (39/45, 89%). Slightly more than one-half (25/45, 57%) identified as white, followed by Asian (7/45, 16%) then "Other" or multiple races (5/45, 12%). Approximately one-third (15/45, 33%) of caregivers assisted a spouse, and a little over one-half (25/45, 56%) cared for a parent or parent-in-law. Most caregivers (33/45, 75%) had been in this role for at least 3 years. See the study by Meyer et al [15] for additional sample characteristics.

Satisfaction surveys were completed by a total of 71 of the 151 caregivers (47%) who registered to participate. In addition to the 37 caregivers (37/45, 82%) who completed the follow-up survey questions related to intervention acceptability as part of the feasibility study, another 34 caregivers who were not enrolled in the pre- and posttest study completed the satisfaction survey anonymously (34/53, 64%; see Multimedia Appendix 1 for additional information on completion rates). Demographic data are not available for these caregivers.

Survey Results

Caregivers reported high levels of overall satisfaction with *KINDER;* 80% (53/66) reported they were "Very Satisfied" with the intervention, and 20% (13/66) indicated they were "Satisfied." When asked about their level of satisfaction with individual program components, most components were rated as being "Very valuable." Group discussion sessions (56/69, 81%) and lesson readings (60/70, 86%) had the highest ratings, while the component most frequently rated as "Not valuable at all" was the reading quizzes given at the end of each independent lesson (5/70, 7%). The perceived value of all *KINDER* intervention components can be found in Figure 1. Group discussion sessions were also highly rated by caregivers, with most caregivers reporting they "Strongly agree" (46/68, 68%) or "Agree" (16/68, 24%) with the statement, "I enjoyed participating in group discussion." See Table 2 for complete results about the level of agreement with the various features of group discussion sessions.

Figure 1. Satisfaction survey results on the perceived value of KINDER intervention components (N=71).

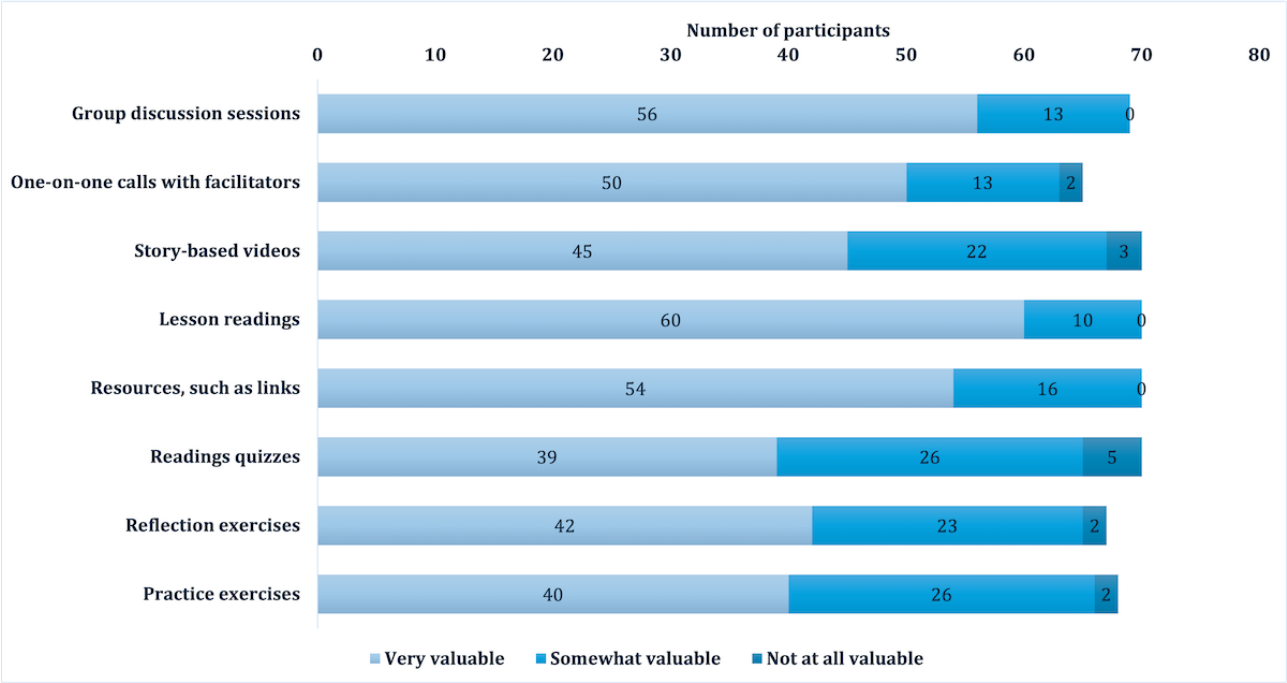

XSL•FO
RenderX

Table 2. Caregivers' level of agreement with features of group discussion sessions (N=71).

Group discussion features	Strongly disagree, n (%)	Disagree, n (%)	Neither agree nor disagree, n (%)	Agree, n (%)	Strongly agree, n (%)
The facilitators were knowledgeable about the topics covered.[a]	0 (0)	0 (0)	1 (1)	8 (11)	61 (87)
I felt comfortable sharing my experiences with other caregivers.[a]	0 (0)	1 (1)	5 (7)	10 (14)	54 (77)
I learned new ideas from other caregivers.[a]	0 (0)	1 (1)	4 (6)	21 (30)	44 (63)
I enjoyed participating in group sessions.[b]	0 (0)	0 (0)	6 (9)	16 (23)	46 (68)

[a]Missing 1 case (n=70).

[b]Missing 3 cases (n=68).

Qualitative Findings

Overview of the Themes

The thematic analysis revealed 4 core themes. The first theme focuses on caregivers' overall attitudes toward *KINDER*, including its alignment with their values. The second theme pertains to caregivers' beliefs about the effectiveness of the intervention at improving their caregiving relationships. The third theme captures the perceived difficulty with participating in the program relative to its perceived overall value. The final theme highlights opportunities the caregivers identified for refining future implementations of *KINDER*. Multimedia Appendix 2 presents the demographic characteristics of each caregiver. Multimedia Appendix 3 provides exemplars of each component of the TFA model by Sekhon et al [30]

Theme 1: Caregivers Reported a Positive Perception of the KINDER Intervention

The first theme characterizes caregivers' overall attitudes toward *KINDER*, including the extent to which they enjoyed participating and understood the intervention's purpose and how well the approach aligned with their values. Within the TFA model, this theme addresses the affective attitude, intervention coherence, and ethicality components of acceptability.

Caregivers Expressed a Positive Attitude Toward KINDER

Caregiver descriptions about their experiences with the intervention were overwhelmingly positive: **"**I loved the class," said one caregiver (CG 45). Another shared, "I think this program should be shared the minute that somebody is diagnosed with something" (CG 12). Similarly, when closing out their interview, a participant caring for her husband expressed her wishes for *KINDER* to continue:

> *I will be praying that God will bless your program and help it to really help a lot of people to have a more positive relationship in caregiving.* [CG 9]

The Intervention Appeared Coherent, and Caregivers Understood its Intended Purpose

Caregivers understood that *KINDER* was developed "to promote a healthier relationship between the care recipient and the caregiver" (CG 35) and encourage a "kinder, gentler, more understanding approach [to caregiving]" (CG 15). They were also cognizant that *KINDER* focused on relationship quality "to increase the quality of care that the people—the care recipients—get" (CG 9). Although the intervention did not directly indicate that one of its goals was to prevent older adult mistreatment, caregivers were aware that part of improving the care relationship also included preventing negative interpersonal interactions. When asked about *KINDER*'s purpose, one caregiver helping her mother responded that it was to "prevent as much hostile or negative interaction that can easily arise in interacting with a loved one" (CG 42).

Caregivers Felt KINDER Aligned With Their Values

Despite tackling difficult topics, responses from participants suggest the intervention aligned with their values. "I feel very positive about the approach...I agree with the approach and how to handle situations," said one participant (CG 35). Another caregiver particularly liked how the program emphasized that caregivers did not need to be "perfect" in their role. She felt that, in stressing this point, "it raised the standards of kindness, but it diminishes the standards of perfectionism" (CG 9).

Of particular interest to the investigators was whether the story-based videos depicting examples of relationship strain and potential mistreatment were consistent with caregivers' values and preferences. Most participants were enthusiastic about these videos. They felt the videos were relevant: "I loved the video material. I thought that was super helpful and you could relate" (CG 45). One caregiver even expressed that she could see herself in the videos: "So, that week seven video could've been me. The kitchen one. I do remember that" (CG 27). Although most caregivers felt that the videos were tasteful in their approach, a few caregivers expressed that the videos were a "little over the top" (CG 22). One participant remembered thinking to herself: "'Really? People can go there?'" (CG 12). Moreover, although there was awareness that the videos could be uncomfortable to watch, there was a general feeling that their value was worth some potential discomfort. The same participant who expressed initial surprise that caregivers "can go there" later recognized their worth during her interview, saying that the videos "remind you that you can go to a place if you're not ready, and I think by being ready you can avoid going there" (CG 12).

XSL•FO
RenderX

Theme 2: Caregivers Described the Intervention as Beneficial to Their Care Relationship

Participants perceived that *KINDER* helped them address their caregiving relationship and improved the quality of caregiving. Within the TFA model, this theme primarily relates to the "perceived effectiveness" component.

KINDER Facilitated a Perceived Reduction in Relationship Strain

Specifically, caregivers reported that *KINDER* helped them reframe their interactions with the care recipient, enabling them to better understand their behaviors and respond more appropriately:

> *So, really the KINDER program just really opened my eyes to the fact that I wasn't seeing him as he is today. I was reacting to him from old historical references instead of what's happening present day right here now.* [CG 35]

One caregiver, currently caring for his wife, reflected on how participating in the program would have benefited him when he cared for his late father. He described feeling anger toward his father and wondered whether attending a course like *KINDER* might have influenced his earlier approach to caregiving:

> *I was trying to be very helpful, but I was also quite angry with him. And, um, had I taken a course like this or been exposed in some way to, um, more effective caregiving, I think I would have approached him very differently.* [CG 15]

Caregivers Gained Skills to Help Manage Relationship Strain and Improve Care Quality

Several participants provided specific examples of how they applied the intervention's lessons to modify their caregiving approach. For example, one of the strategies taught in *KINDER* to reduce relationship strain is to "go with the flow." One caregiver helping her mother described how she has applied this mantra:

> *I used to think, my mom has to listen to me, and that's the way it's supposed to be. Now, I'm more relaxed. That's the way. I will go with the flow.* [CG 19]

Participants also applied other skills to reduce relationship tension with the care recipient, such as redirecting the care recipient's attention to manage the behavioral symptoms of dementia or cognitively reframing the situation:

> *Learning all those skills from the book and then, kinda like how to divert or think of positive side instead of drilling on the negative side of why I am taking care of this person. I think it's really helpful in building relationship with my parent.* [CG 14]

Similarly, another participant recognized the need to modify her interactions with her husband in response to changes in his cognition:

> *You know I said, "Oh, wait a minute, I've got to, you know, tone it down a little bit," and just [9] a different approach, and just some reminders about of all of that.* [CG 12]

Perhaps most interesting was the caregiver who shared a reflection exercise response, where they applied intervention content about intimidating body language to their care situation:

> *My week seven reflection was basically, I just wrote something very short, I wrote "Physical aggression that doesn't involve hitting, but that can be just as affecting. It causes more confusion and is likely more isolating. Even though she is very tough, it's not nice to make things scarier than they need to be for her, ever."* [CG 27]

This reflection reinforced an important message about relationship strain and psychological mistreatment from the intervention, highlighting where caregivers' body language can be a source of intimidation.

Theme 3: Caregivers Felt Participating in KINDER Was Worthwhile

The third theme explores whether the caregivers felt the time and effort required to participate in *KINDER* was justified. This theme aligns with the burden, opportunity costs, and self-efficacy components within the TFA model. Overall, caregivers reported participation in *KINDER* was worthwhile and not overly burdensome.

Caregivers Reported That KINDER Was "a Very Good Use of Time" (CG 15)

When asked whether there was content that should be removed or condensed, none of the caregivers had any suggestions:

> *I didn't really find anything that was just like so left field that I was like, "Oh, I don't need this." Not at all. I felt like every reflection, I had something to write and say.* [CG 27]

Echoing this statement, another caregiver said:

> *I think all of the chapters [of the workbook] are very helpful. I think the whole thing should be kept.* [CG 14]

Even one initially skeptical participant eventually came around to acknowledging that *KINDER* was worthwhile:

> *And I think when it was first going on, I was wondering, "What? Do I really need to be involved in this kind of stuff?" And the answer was of course. Resoundingly so. And so, I'm feeling, I would say, a lot closer to [care recipient name] now than I was three or four years ago.* [CG 15]

Caregivers Found That Participation Was Not Overly Burdensome

When caregivers reported any difficulty in participating, they felt that any burden imposed by the intervention was worth the effort. For example, some caregivers felt the independent lesson readings in the workbook could be overwhelming:

> *I will say it was hard to make the time to do the reading. And that's interesting because I was not in the total thick of it. If I had had to try to do that*

XSL•FO
RenderX

> *reading and watch the videos while I was in the total thick of it, I don't know if I would have been able to do it.* [CG 42]

On the other hand, sharing enthusiasm for workbook content, another caregiver said, "I'm never throwing this book away" (CG 29). Caregivers appreciated the email reminders sent each week to help keep them on track with readings and practice exercises:

> *I love your reminder every week to do your homework. That's a good reminder because we're so busy with other things.* [CG 19]

Caregivers reported that the regular reminders and the discussion meetings created a sense of "accountability," where, as one caregiver described it, "you're [39] just sort of on an island, reading a book. It's like, no, you're in touch with people that are inquiring as to these things" (CG 27).

Theme 4: Caregivers Identified Opportunities to Improve KINDER

The fourth theme describes insights caregivers offered about how *KINDER* could be adapted to better meet their needs. Although participants generally found the program worthwhile, interviews identified opportunities for improvement, particularly around increasing caregiver interaction and making participation more convenient. The following recommendations highlight key areas for improving the program's accessibility and engagement.

Recommendation 1: Provide Additional Opportunities for Caregivers to Connect

The primary recommendation to improve *KINDER*'s acceptability was to provide more opportunities for caregivers to connect with other cohort members. Several caregivers suggested increasing the number of group discussion sessions. "I think you should have more online meetings with everybody where they can discuss things," shared one caregiver (CG 22). Another participant felt that meeting once per week may have been more beneficial than the 3 meetings spread throughout the 8 weeks: "I feel like maybe once a week would be a better way to gain insight from others going through the same or similar situation" (CG 35).

Recommendation 2: Facilitate Opportunities for Interaction During the Discussion Sessions

Despite an overall appreciation for the group sessions, some caregivers felt sessions could have been more interactive. One caregiver shared that the groups "need more interaction between each other somehow, or a little group breakout room" (CG 27). Other ideas included encouraging caregivers to turn on their cameras during the group sessions and incorporating more icebreakers to help participants get to know each other better. One participant offered insights on improving session facilitation to promote more exchange, noting that when a single caregiver's concern took up a large part of the group session, it negatively affected their experience:

> *I feel for them and I'm sorry that they're experiencing that...But I do feel like that did go on a little bit long.* [CG 45]

Recommendation 3: Integrate Support to Make Participation More Convenient

Caregivers proposed ways to make participation more convenient. Some suggested sending text message reminders, in addition to emails, for independent lessons and discussion sessions: "I don't get on my email as much, but my text [messages]—I always have my phone" (CG 12). To make it easier to participate in independent lessons, another idea was to audio record the assigned readings so that caregivers could listen to lessons while doing other activities such as "during their commute" (CG 29). Another caregiver suggested delivering the program via an app or web-based platform where they could complete lesson activities online:

> *There needs to be just look at it on your phone...There's click here to enter it to keep a log, to use the website or whatever, to save your work.* [CG 42]

When asked about how to reconcile different needs and preferences of caregivers regarding the integration of technologies, like an app or website, one participant responded that the best solution was to offer choices and used a wheelchair ramp as an analogy:

> *People think, "Oh, let's see what the majority will use." Who gives a crap? If one person needs a wheelchair ramp, we need a wheelchair ramp.* [CG 42]

Discussion

Summary of Findings

This study aimed to assess the acceptability of the latest version of an intervention to prevent psychological mistreatment by family caregivers to persons with dementia. We are unaware of any other study that has reported on the acceptability of caregiver interventions developed to prevent older adult mistreatment, making this work a novel addition to the current literature. Findings support *KINDER*'s overall acceptability among caregivers to persons with dementia. Postintervention satisfaction survey responses reflect that caregivers would recommend the intervention to other caregivers and that caregivers found each component of the intervention valuable. One of the primary changes between the current version of *KINDER* and the previous version is the addition of 3 virtual group discussion sessions. Survey data show that caregivers had a positive experience with these discussion groups. Qualitative data support this finding, revealing that some caregivers would even be interested in additional group discussions throughout the 8-week program. By using multiple methods of data collection, this study offers a holistic understanding of how *KINDER* was perceived by caregiver participants. In so doing, findings not only support the intervention's acceptability but also highlight opportunities to continue improving caregivers' experiences during future program implementations.

Findings, particularly those from the qualitative interviews, demonstrate a positive affective attitude toward the program, alignment with values (ethicality), and the overall belief that

XSL•FO
RenderX

the demands of participating were not too high and mostly worthwhile (opportunity cost and burden). These findings align with those of the initial first web-based, asynchronous version of *KINDER* 1.0, where participants also completed qualitative interviews following the intervention [14]. Whereas the earlier study applied the TFA after conducting thematic analysis, this study integrated components of the framework into the coding process to more clearly evaluate their application to caregiver experiences with *KINDER* more closely [30].

One of *KINDER*'s strengths identified in both studies was the inclusion of story-based videos. Videos and visual examples have been found to enhance the experience of caregivers participating in interventions [31]. Prior studies of digital caregiver interventions have found that participants value videos portraying the "messiness" of caregiving [42]. A comparison of qualitative findings from the two feasibility studies of *KINDER* suggests that challenges related to participation were largely addressed. For example, in the earlier study (*KINDER* 1.0), caregivers reported technological issues affected their experiences. After simplifying the mode of delivery from a web-based delivery platform to a PDF or print workbook, qualitative interview data collected in the second study showed far fewer concerns about technology being a barrier to participation than in the first study. This suggests technology issues may have been mitigated with the modifications made to *KINDER* in the second study. Nevertheless, this concern still requires ongoing consideration in future studies, as nonresponders for interviews and surveys may have had different experiences than responders. Another obstacle reported by caregivers who participated in *KINDER* 1.0 was that the content could be overwhelming and they lacked time to complete independent lessons. However, fewer caregivers in the revised *KINDER* program identified this as a challenge, likely due to its regular check-in reminders, group discussion sessions, and flexible scheduling that allowed participants to "catch up" on independent learning activities when needed. Furthermore, we also found that group-based sessions recommended by caregivers in the earlier study enhanced participant experiences in the *KINDER* intervention. Caregivers, for the most part, felt connected to others in their cohort. Some even wished for more opportunities to connect. This finding is consistent with those from similar caregiver interventions that use a hybrid group-based and asynchronous delivery approach [42,43].

Comparison With Other Interventions to Support Caregiving Relationship Quality

There are several interventions focused on improving caregiving relationships, though not necessarily on prevention of older adult mistreatment, providing points of comparison with our findings. In the *AtThe Crossroads* caregiver intervention, researchers conducted a randomized controlled trial of a psychoeducational intervention for driving cessation among persons with dementia [44]. Like *KINDER*, this intervention emphasized the potential for varying priorities between the caregiver and person with dementia, particularly related to the balance of safety and autonomy. Findings revealed that caregivers in the intervention group felt more prepared to discuss driving cessation with their family member and were less worried about upsetting or angering the care recipient than those in the control condition. This aligns with findings from our study, wherein caregiver participants similarly felt more prepared to reframe interactions with the care recipient and reported enacting skills taught in the intervention to promote positive care relationships. Although the two interventions—*At The Crossroads* and *KINDER*—have distinctly different overall goals (driving cessation versus preventing older adult mistreatment), both interventions leverage existing interpersonal relationships to improve relational functioning through similar approaches.

Another intervention well-suited for comparison is the *Support, Health, Activities, Resources, and Education* (*SHARE*) intervention. This intervention is designed to reconcile care values between caregivers and those with dementia while creating a plan for future care based on these shared values [45]. Unlike *At a Crossroads* and *KINDER*, *SHARE* is an in-person intervention that includes both care partners and a trained counselor. However, like these other interventions, *SHARE* was delivered over 6 weekly sessions and explicitly addresses the reconciliation of care values, including safety and autonomy. Caregivers participating in *SHARE* reported satisfaction with the overall experience of participating, the counselor, the weekly sessions, and their relationship functioning with the care recipient postintervention [45]. Qualitative findings also affirm overall satisfaction, as well as the application of new skills, development of knowledge, and greater awareness of available resources [46]. Caregivers in *SHARE* reported similar challenges to those identified in the evaluation of *KINDER* 1.0, such as challenges with finding time to participate and difficulties navigating emotionally intense content. These similarities suggest that interventionists should be mindful of the potential for emotional distress among participants.

Opportunities to Improve the Current KINDER Intervention

Findings from this study indicate opportunities to improve *KINDER* and enhance intervention acceptability. First, caregivers expressed an interest in using technology to enhance convenience when participating, suggesting text message reminders in addition to email. This recommendation holds promise in terms of likely acceptability and feasibility, given that 82% of caregivers use a smartphone and recent research shows that caregivers find text-message delivered interventions to be highly acceptable [47,48]. However, although text messaging is a relatively simple technology integration, caution is needed when considering more complex digital intervention delivery models where the time and effort required for participants to learn these systems may outweigh their perceived benefits [31]. *KINDER* is unique in that its first iteration took place via a web-based portal, before it was revised to make it more accessible and easier for caregivers to participate. This raises a dilemma: To what extent should caregiver preferences for integration of digital technologies be heeded when prior applications demonstrate a lack of engagement?

Multiple factors should be considered when addressing this question and when interpreting study results. First, the context of intervention delivery for caregivers has changed tremendously in recent years and may offer an alternative explanation as to

XSL•FO
RenderX

why technology barriers were featured less often in the second feasibility study than the first one. Social distancing restrictions put in place during the COVID-19 pandemic accelerated the transition to digitally delivered programming, which may have increased caregivers' comfort with digital platforms [49]. Although the pandemic most likely accelerated acceptance and uptake of digital caregiver interventions given service organizations' efforts to provide technology support [50,51], this shift likely occurred within the broader context of already evolving social and cultural norms around technology adoption as society continues to acclimatize to emerging technologies [16]. If the online version of *KINDER* (*KINDER* 1.0) was offered today, it is plausible that it would achieve greater acceptance than when it was initially offered from 2020 to 2021.

Another consideration was illustrated by a study participant who revealed the importance of accommodating varying needs and preferences for intervention delivery: "If one person needs a wheelchair ramp, we need a wheelchair ramp." This statement aligns with calls to prioritize health equity when developing and testing digital solutions [52]. Based on our findings, we propose that, rather than optimizing an intervention's delivery approach to fit the needs of most caregivers, interventions should instead be offered in multiple formats to accommodate varying needs. This aligns with prior recommendations to offer digital materials in a variety of formats to meet different needs [53]. Beyond offering options for participation, in future studies, we intend to measure digital literacy among users to better distinguish between the acceptability of a digital intervention and users' level of digital literacy. Limited digital literacy can influence user acceptability and indicates a need for strategies targeted at digital literacy specifically to improve acceptability (eg, technology coaching to increase comfort with digital tools) [54]. Further, during intervention eligibility screening, we will identify the frequency at which limited access to digital devices and internet prevents participation. A recent scoping review identified inequitable internet access as a structural barrier to caregivers accessing community support services [55]. Overcoming inequitable broadband access is also a recommendation in the United States' 2022 National Strategy on Caregiving to enable easier access to digital caregiver supports [56].

Further, although caregivers expressed an interest in having more opportunities to connect with group members, some participants in the *KINDER* program voiced concerns about the perceived low level of engagement from others. For example, one caregiver suggested the facilitators encourage participants to turn on their cameras during group sessions so others could see their faces, and another suggested adding more icebreakers to kick off the group sessions. In their qualitative study of the *TeleSavvy* intervention, Kovaleva et al [42] identified a similar issue, where participants expressed bother when they noticed other participants were distracted. Low engagement among certain participants can undermine the experience of others, who may benefit less from group discussion and shared examples from others, making this a serious concern. At the same time, behaviors that may appear as disinterest could be driven by factors related to health equity, such as lacking additional support that would allow a caregiver to participate in *KINDER* without disruption. Indeed, one caregiver in this study described having to listen to one of the discussion sessions while picking up her daughter from school, rendering her unable to contribute to group conversations. We partially addressed this in *KINDER* by encouraging caregivers to keep their cameras on during sessions, while also reminding them that it is okay to go "off camera" when needed. Still, given the variability this creates for participants between cohorts, further consideration may be needed to understand how to address participant engagement in synchronous group-based caregiver interventions. Nevertheless, observer fidelity reports indicate that in 17 of 18 sessions where participant engagement in group sessions was assessed, observers "Agreed" or "Strongly agreed" that participants were engaged [15].

Limitations

This study has limitations. First, the initial qualitative interviews were led by the lead investigator, who also cofacilitated the intervention. As a result, participants' responses may have been affected by a social desirability bias. However, there was no apparent difference in participant attitudes in interviews conducted with another member of the research team and those who spoke with the lead investigator. In both instances, caregivers provided both criticism and praise of the *KINDER* program. Rapport with the lead investigator may also have facilitated trust so that participants felt more comfortable sharing negative feedback and how they applied what they learned to manage challenging care relationships [57]. Although sample representativeness is not an objective of qualitative studies, it is possible that results were affected by self-selection bias where caregivers who had more positive experiences agreed to be interviewed.

Another limitation is that this study conflated the concepts of usability and acceptability, recognized in the literature as distinct concepts. Although we applied the validated TFA to help evaluate our intervention's acceptability among caregivers, future studies should also include standardized measures of acceptability and usability that delineate these interconnected though distinct concepts to better inform intervention refinement [16,31]. Nevertheless, we used an established acceptability framework (TFA) to guide the qualitative interview data collection and analysis, which enabled us to evaluate various dimensions of intervention acceptability. At the time when this study was administered, there was no validated questionnaire available based on the TFA. Such a questionnaire is under development and will be a valuable tool to measure *KINDER*'s acceptability once validated. The use of a nonvalidated measure in this study weakens the generalizability of our quantitative findings and can undermine future replication efforts. Finally, demographic data were not available from participants who completed the satisfaction survey. Nevertheless, data from the online event registration form revealed no statistically significant difference in age, gender, race, ethnicity, educational attainment, nor financial difficulty of caregivers in the study compared with those registered to attend without being in the study [15].

Conclusion

This study advances current knowledge on interventions for caregivers of persons with dementia to prevent older adult

XSL•FO
RenderX

mistreatment by demonstrating the acceptability of a novel intervention, *KINDER*. Additional research is needed to test the efficacy of the intervention, including the plausibility of theoretical mechanisms that may drive a reduction of psychological EM, and to continue evaluating the implementation of the intervention within health and social service settings. Specifically, the next step for this research will involve conducting a fully powered randomized controlled trial as part of a Stage II efficacy trial in a controlled research setting [58]; in addition to measures of study outcomes, study participants randomized into the intervention arm will be asked about intervention acceptability in a postintervention survey to ensure ongoing acceptability. To prepare for effectiveness testing in applied settings, such as Area Agencies on Aging, we will interview service providers to identify how the *KINDER* intervention may need to be adapted to fit provider workflows and organizational processes [59]. Nevertheless, by using quantitative and qualitative methods to assess multiple components of acceptability, this study provides a robust picture of the acceptability of one of the few interventions designed to prevent EM in a family care context. If proven efficacious, the *KINDER* intervention has the potential to promote care that upholds the dignity and safety of both caregivers and persons with dementia [15]. However, this potential can only be met if the intervention is acceptable to users. Findings from this study indicate that KINDER's intended audience, family caregivers, find the intervention to be acceptable, though ongoing evaluation of acceptability of KINDER and similar programs is needed.

Acknowledgments

The project was supported by Faculty "Start Up" funds provided to KM from the Case Western Reserve University Frances Payne Bolton School of Nursing. The content is solely the responsibility of the authors and does not necessarily represent the official views of Case Western Reserve University.

Data Availability

Data from both studies described in this manuscript are not available due to the sensitive nature of topics discussed with participants.

Authors' Contributions

Conceptualization – KM, JYC, SM, ZG, JAZ, DB
Data curation – KM, SK
Formal analysis – KM, WW, DB
Investigation – KM, WW
Project administration – KM
Validation – JAZ, DB
Writing – original draft – KM, WW, JAZ, DB
Writing – review & editing – KM, WW, JYC, SM, SK, EI, CB, ZG, EL, JAZ, DB

Conflicts of Interest

None declared.

Multimedia Appendix 1
Creation of analytic sample for satisfaction surveys from program registrants.
[DOCX File , 139 KB - formative_v9i1e73778_app1.docx]

Multimedia Appendix 2
Demographic information for qualitative interviews (N = 11).
[DOCX File , 19 KB - formative_v9i1e73778_app2.docx]

Multimedia Appendix 3
Exemplars of each component of acceptability from the Theoretical Framework of Acceptability (TFA) model by Sekhon et al.
[DOCX File , 17 KB - formative_v9i1e73778_app3.docx]

Multimedia Appendix 4
COREQ checklist.
[PDF File (Adobe PDF File), 424 KB - formative_v9i1e73778_app4.pdf]

Multimedia Appendix 5
CONSORT 2010 checklist.
[PDF File (Adobe PDF File), 144 KB - formative_v9i1e73778_app5.pdf]

XSL•FO
RenderX

References

1. Abuse of older people. World Health Organization. 2024 Jun 15. URL: https://www.who.int/news-room/fact-sheets/detail/abuse-of-older-people#:~:text=The%20abuse%20of%20older%20people,distress%20to%20an%20older%20person [accessed 2025-07-19]
2. AARP, National Alliance for Caregiving. Caregiving in the United States 2020. AARP. 2020 May 14. URL: https://www.aarp.org/pri/topics/ltss/family-caregiving/caregiving-in-the-united-states/ [accessed 2025-07-19]
3. Orfila F, Coma-Solé M, Cabanas M, Cegri-Lombardo F, Moleras-Serra A, Pujol-Ribera E. Family caregiver mistreatment of the elderly: prevalence of risk and associated factors. BMC Public Health 2018 Jan 22;18(1):167 [FREE Full text] [doi: 10.1186/s12889-018-5067-8] [Medline: 29357866]
4. Pickering CEZ, Yefimova M, Maxwell C, Puga F, Sullivan T. Daily context for abusive and neglectful behavior in family caregiving for dementia. Gerontologist 2020 Apr 02;60(3):483-493 [FREE Full text] [doi: 10.1093/geront/gnz110] [Medline: 31425586]
5. Wiglesworth A, Mosqueda L, Mulnard R, Liao S, Gibbs L, Fitzgerald W. Screening for abuse and neglect of people with dementia. J Am Geriatr Soc 2010 Mar;58(3):493-500 [FREE Full text] [doi: 10.1111/j.1532-5415.2010.02737.x] [Medline: 20398118]
6. Browning WR, Yildiz M, Hernandez Chilatra JA, Yefimova M, Maxwell CD, Sullivan TP, et al. Mechanisms underlying the use of abusive and neglectful behaviors in dementia caregiving: the role of caregiver mental health. Res Gerontol Nurs 2024;17(5):227-236 [FREE Full text] [doi: 10.3928/19404921-20240808-01] [Medline: 39347758]
7. Pillemer K, Burnes D, Riffin C, Lachs MS. Elder abuse: global situation, risk factors, and prevention strategies. Gerontologist 2016 Apr;56 Suppl 2(Suppl 2):S194-S205 [FREE Full text] [doi: 10.1093/geront/gnw004] [Medline: 26994260]
8. Wong JS, Waite LJ. Elder mistreatment predicts later physical and psychological health: results from a national longitudinal study. J Elder Abuse Negl 2017;29(1):15-42 [FREE Full text] [doi: 10.1080/08946566.2016.1235521] [Medline: 27636657]
9. Mouton CP, Rodabough RJ, Rovi SLD, Brzyski RG, Katerndahl DA. Psychosocial effects of physical and verbal abuse in postmenopausal women. Ann Fam Med 2010;8(3):206-213 [FREE Full text] [doi: 10.1370/afm.1095] [Medline: 20458103]
10. Maxwell CE, Yefimova M, Pickering C. Co-occurrence of abuse and neglect as reported by dementia family caregivers. Innovation in Aging 2019;3(Supplement_1):S544 [FREE Full text] [doi: 10.1093/geroni/igz038.2002]
11. Alzheimer's Association. 2024 Alzheimer's disease facts and figures. Alzheimers Dement 2024 May;20(5):3708-3821 [FREE Full text] [doi: 10.1002/alz.13809] [Medline: 38689398]
12. Braithwaite SR, Fincham FD. Computer-based prevention of intimate partner violence in marriage. Behav Res Ther 2014 Mar;54:12-21. [doi: 10.1016/j.brat.2013.12.006] [Medline: 24463577]
13. Doss BD, Knopp K, Roddy MK, Rothman K, Hatch SG, Rhoades GK. Online programs improve relationship functioning for distressed low-income couples: results from a nationwide randomized controlled trial. J Consult Clin Psychol 2020 Apr;88(4):283-294. [doi: 10.1037/ccp0000479] [Medline: 32134290]
14. Meyer K, Gonzalez A, Benton D. Qualitative evaluation of family caregivers' experiences participating in knowledge and interpersonal skills to Develop Exemplary Relationships (KINDER): web-based intervention to improve relationship quality. JMIR Form Res 2023 Aug 22;7:e42561 [FREE Full text] [doi: 10.2196/42561] [Medline: 37606980]
15. Meyer K, Yonashiro-Cho J, Zauszniewski J, Burant C, Mosqueda L, Gassoumis Z, et al. A feasibility study of KINDER: an elder mistreatment intervention for family caregivers of persons living with dementia. J Elder Abuse Negl 2025;37(2):107-131. [doi: 10.1080/08946566.2025.2460835] [Medline: 39886850]
16. Perski O, Short CE. Acceptability of digital health interventions: embracing the complexity. Transl Behav Med 2021 Jul 29;11(7):1473-1480 [FREE Full text] [doi: 10.1093/tbm/ibab048] [Medline: 33963864]
17. Atkinson E, Roberto KA. Global approaches to primary, secondary, and tertiary elder abuse prevention: a scoping review. Trauma Violence Abuse 2024 Jan;25(1):150-165. [doi: 10.1177/15248380221145735] [Medline: 36636948]
18. Mikton C, Beaulieu M, Burnes D, Choo W, Herbst J, Pillemer K, et al. High time for an intervention accelerator to prevent abuse of older people. Nat Aging 2022 Nov;2(11):973-975 [FREE Full text] [doi: 10.1038/s43587-022-00301-0] [Medline: 37118090]
19. Cooper C, Barber J, Griffin M, Rapaport P, Livingston G. Effectiveness of START psychological intervention in reducing abuse by dementia family carers: randomized controlled trial. Int Psychogeriatr 2016 Jun;28(6):881-887 [FREE Full text] [doi: 10.1017/S1041610215002033] [Medline: 26652193]
20. Gassoumis ZD, Martinez JM, Yonashiro-Cho J, Mosqueda L, Hou A, Han SD, et al. Comprehensive Older Adult and Caregiver Help (COACH): A person-centered caregiver intervention prevents elder mistreatment. J Am Geriatr Soc 2024 Jan;72(1):246-257. [doi: 10.1111/jgs.18597] [Medline: 37791406]
21. Burnes D, Connolly M, Salvo E, Kimball PF, Rogers G, Lewis S. RISE: a conceptual model of integrated and restorative elder abuse intervention. Gerontologist 2023 Jul 18;63(6):966-973 [FREE Full text] [doi: 10.1093/geront/gnac083] [Medline: 35705108]
22. Burnes D, MacNeil A, Connolly M, Salvo E, Kimball PF, Rogers G, et al. A qualitative evaluation of the "RISE" elder abuse intervention from the perspective of adult protective services caseworkers: addressing a service system gap. J Elder Abuse Negl 2022;34(5):329-348. [doi: 10.1080/08946566.2022.2140321] [Medline: 36316963]
23. Meyer KN, McPherson N, Benton D. So, you want to build a program? Generations 2023 Apr;47(1):1 [FREE Full text]

XSL•FO
RenderX

24. Meyer KN, Glassner A, Lee K, Pickering CEZ, White CL. Conceptualizing how caregiving relationships connect to quality of family caregiving within the stress process model. J Gerontol Soc Work 2022;65(6):635-648 [FREE Full text] [doi: 10.1080/01634372.2021.2010855] [Medline: 34851796]
25. Meyer K, Rath L, Avent E, Benton D, Nash P, Wilber K. How do family caregivers of older adults cope with relationship strain? Aging Ment Health 2023;27(10):1990-1999 [FREE Full text] [doi: 10.1080/13607863.2023.2247353] [Medline: 37574858]
26. Richards D, Richardson T. Computer-based psychological treatments for depression: a systematic review and meta-analysis. Clin Psychol Rev 2012 Jun;32(4):329-342. [doi: 10.1016/j.cpr.2012.02.004] [Medline: 22466510]
27. Samia LW, Hepburn K, Nichols L. "Flying by the seat of our pants": what dementia family caregivers want in an advanced caregiver training program. Res Nurs Health 2012 Dec 21;35(6):598-609. [doi: 10.1002/nur.21504] [Medline: 22911130]
28. Meyer K, Glassner A, Norman R, James D, Sculley R, LealVasquez L, et al. Caregiver self-efficacy improves following complex care training: results from the Learning Skills Together pilot study. Geriatr Nurs 2022;45:147-152 [FREE Full text] [doi: 10.1016/j.gerinurse.2022.03.013] [Medline: 35447558]
29. Bowen DJ, Kreuter M, Spring B, Cofta-Woerpel L, Linnan L, Weiner D, et al. How we design feasibility studies. Am J Prev Med 2009 May;36(5):452-457 [FREE Full text] [doi: 10.1016/j.amepre.2009.02.002] [Medline: 19362699]
30. Sekhon M, Cartwright M, Francis JJ. Acceptability of healthcare interventions: an overview of reviews and development of a theoretical framework. BMC Health Serv Res 2017 Jan 26;17(1):88 [FREE Full text] [doi: 10.1186/s12913-017-2031-8] [Medline: 28126032]
31. Ottaviani AC, Monteiro DQ, Oliveira D, Gratão ACM, Jacinto AF, Campos CRF, et al. Usability and acceptability of internet-based interventions for family carers of people living with dementia: systematic review. Aging Ment Health 2022 Oct;26(10):1922-1932. [doi: 10.1080/13607863.2021.1975095] [Medline: 34511028]
32. Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O'Neal L, REDCap Consortium. The REDCap consortium: building an international community of software platform partners. J Biomed Inform 2019 Jul;95:103208 [FREE Full text] [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]
33. Sekhon M, Cartwright M, Francis JJ. Development of a theory-informed questionnaire to assess the acceptability of healthcare interventions. BMC Health Serv Res 2022 Mar 01;22(1):279 [FREE Full text] [doi: 10.1186/s12913-022-07577-3] [Medline: 35232455]
34. Weiner BJ, Lewis CC, Stanick C, Powell BJ, Dorsey CN, Clary AS, et al. Psychometric assessment of three newly developed implementation outcome measures. Implement Sci 2017 Aug 29;12(1):108 [FREE Full text] [doi: 10.1186/s13012-017-0635-3] [Medline: 28851459]
35. Stata Statistical Software: Release 17. StataCorp. 2021. URL: https://www.stata.com/stata17/ [accessed 2025-07-20]
36. Braun V, Clarke V. Using thematic analysis in psychology. Qualitative Research in Psychology 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
37. Lincoln YS, Guba EG. Naturalistic Inquiry. Newbury Park, CA: SAGE Publications Ltd; 1985.
38. Nowell L, Norris J, White D, Moules N. Thematic analysis. International Journal of Qualitative Methods 2017 Oct 02;16(1):1 [FREE Full text] [doi: 10.1177/1609406917733847]
39. NVivo 14. Lumivero. 2023. URL: https://lumivero.com/products/nvivo/ [accessed 2025-07-20]
40. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
41. Eldridge SM, Chan CL, Campbell MJ, Bond CM, Hopewell S, Thabane L, PAFS consensus group. CONSORT 2010 statement: extension to randomised pilot and feasibility trials. BMJ 2016 Oct 24;355:i5239 [FREE Full text] [doi: 10.1136/bmj.i5239] [Medline: 27777223]
42. Kovaleva M, Blevins L, Griffiths PC, Hepburn K. An online program for caregivers of persons living with dementia: lessons learned. J Appl Gerontol 2019 Feb;38(2):159-182. [doi: 10.1177/0733464817705958] [Medline: 28460557]
43. White CL, Barrera A, Turner S, Glassner A, Brackett J, Rivette S, et al. Family caregivers' perceptions and experiences of participating in the learning skills together intervention to build self-efficacy for providing complex care. Geriatr Nurs 2022 May;45:198-204 [FREE Full text] [doi: 10.1016/j.gerinurse.2022.04.012] [Medline: 35533583]
44. Stern RA, D'Ambrosio LA, Mohyde M, Carruth A, Tracton-Bishop B, Hunter JC, et al. At the crossroads: development and evaluation of a dementia caregiver group intervention to assist in driving cessation. Gerontol Geriatr Educ 2008;29(4):363-382 [FREE Full text] [doi: 10.1080/02701960802497936] [Medline: 19064472]
45. Whitlatch CJ, Heid AR, Femia EE, Orsulic-Jeras S, Szabo S, Zarit SH. The Support, Health, Activities, Resources, and Education program for early stage dementia: results from a randomized controlled trial. Dementia (London) 2019 Aug;18(6):2122-2139. [doi: 10.1177/1471301217743033] [Medline: 29171296]
46. Orsulic-Jeras S, Whitlatch CJ, Szabo SM, Shelton EG, Johnson J. The SHARE program for dementia: implementation of an early-stage dyadic care-planning intervention. Dementia (London) 2019 Jan 13;18(1):360-379. [doi: 10.1177/1471301216673455] [Medline: 27738110]

XSL•FO
RenderX

47. Perales-Puchalt J, Ramírez-Mantilla M, Fracachán-Cabrera M, Vidoni ED, Ellerbeck EF, Ramírez AS, et al. A text message intervention to support latino dementia family caregivers (CuidaTEXT): feasibility study. Clin Gerontol 2024;47(1):50-65. [doi: 10.1080/07317115.2022.2137449] [Medline: 36268684]
48. Lee S, Ory MG, Vollmer Dahlke D, Smith ML. Technology use among older adults and their caregivers: cross-sectional survey study. JMIR Aging 2024 May 01;7:e50759 [FREE Full text] [doi: 10.2196/50759] [Medline: 38717339]
49. Cuffaro L, Di Lorenzo F, Bonavita S, Tedeschi G, Leocani L, Lavorgna L. Dementia care and COVID-19 pandemic: a necessary digital revolution. Neurol Sci 2020 Aug;41(8):1977-1979 [FREE Full text] [doi: 10.1007/s10072-020-04512-4] [Medline: 32556746]
50. Masoud SS, Meyer KN, Martin Sweet L, Prado PJ, White CL. "We don't feel so alone": a qualitative study of virtual memory cafés to support social connectedness among individuals living with dementia and care partners during COVID-19. Front Public Health 2021;9:660144 [FREE Full text] [doi: 10.3389/fpubh.2021.660144] [Medline: 34055724]
51. O'Connor MK, Nicholson R, Epstein C, Donley T, Salant R, Nguyen AH, et al. Telehealth support for dementia caregivers during the COVID-19 pandemic: lessons learned From the NYU family support program. Am J Geriatr Psychiatry 2023 Jan;31(1):14-21 [FREE Full text] [doi: 10.1016/j.jagp.2022.08.005] [Medline: 36167652]
52. Lindeman DA, Kim KK, Gladstone C, Apesoa-Varano EC. Technology and caregiving: emerging interventions and directions for research. Gerontologist 2020 Feb 14;60(Suppl 1):S41-S49 [FREE Full text] [doi: 10.1093/geront/gnz178] [Medline: 32057082]
53. Premanandan S, Ahmad A, Cajander ?, Ågerfalk P, van Gemert-Pijnen L. Designing and evaluating IT applications for informal caregivers: scoping review. J Med Internet Res 2024 Oct 23;26:e57393 [FREE Full text] [doi: 10.2196/57393] [Medline: 39442167]
54. Cho H, Cho E, You SB, Sefcik JS, Hodgson NA, Demiris G. Understanding digital literacy of persons with dementia and their caregivers: a scoping review and an evolutionary concept analysis of empirical studies. J Appl Gerontol 2025 Jun 19:7334648251348703 [FREE Full text] [doi: 10.1177/07334648251348703] [Medline: 40537959]
55. Choi H, Reblin M, Litzelman K. Conceptualizing family caregivers' use of community support services: a scoping review. Gerontologist 2024 May 01;64(5):1. [doi: 10.1093/geront/gnad039] [Medline: 37022354]
56. The Recognize, Assist, Include, Support, and Engage (RAISE) Act Family Caregiving Advisory Council, The Advisory Council to Support Grandparents Raising Grandchildren. 2022 National Strategy to Support Family Caregivers. Administration for Community Living. 2022. URL: https://acl.gov/sites/default/files/RAISE_SGRG/NatlStrategyToSupportFamilyCaregivers-2.pdf [accessed 2025-07-20]
57. Prior MT. Accomplishing “rapport” in qualitative research interviews: empathic moments in interaction. Applied Linguistics Review 2017;9(4):1. [doi: 10.1515/applirev-2017-0029]
58. Onken LS, Carroll KM, Shoham V, Cuthbert BN, Riddle M. Reenvisioning clinical science: unifying the discipline to improve the public health. Clin Psychol Sci 2014 Jan 01;2(1):22-34 [FREE Full text] [doi: 10.1177/2167702613497932] [Medline: 25821658]
59. Hodgson N, Gitlin LN. Implementing and sustaining family care programs in real-world settings: Barriers and facilitators. In: Gaugler JE, editor. Bridging the Family Care Gap. Cambridge, MA: Academic Press; 2021:179-219.

Abbreviations

APS: Adult Protective Service
CONSORT: Consolidated Standards of Reporting Trials
COREQ: Consolidated Criteria for Reporting Qualitative Research
KINDER: Knowledge and Interpersonal Skills to Develop Enhanced Relationships
REDCap: Research Electronic Data Capture
TFA: Theoretical Framework of Acceptability
WHO: World Health Organization

Edited by J Sarvestan; submitted 11.03.25; peer-reviewed by SQ Yoong, M Troutman-Jordan; comments to author 05.05.25; revised version received 28.05.25; accepted 14.07.25; published 30.07.25.

Please cite as:
Meyer K, Wei W, Yonashiro-Cho J, Mage S, Kim S, Irani E, Burant C, Gassoumis Z, Lamb EG, Zauszniewski JA, Benton D Acceptability of an Intervention to Prevent Older Adult Mistreatment Among Family Caregivers to Persons With Dementia: Multimethod Pilot Study
JMIR Form Res 2025;9:e73778
URL: https://formative.jmir.org/2025/1/e73778
doi:10.2196/73778
PMID:

XSL•FO
RenderX

©Kylie Meyer, Wenxing Wei, Jeanine Yonashiro-Cho, Susanna Mage, Sohee Kim, Elliane Irani, Christopher Burant, Zachary Gassoumis, Erin Gentry Lamb, Jaclene A Zauszniewski, Donna Benton. Originally published in JMIR Formative Research (https://formative.jmir.org), 30.07.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Automating Colon Polyp Classification in Digital Pathology by Evaluation of a “Machine Learning as a Service” AI Model: Algorithm Development and Validation Study

David Beyer[1], BSc, MD; Evan Delancey[2], BSc, MSc; Logan McLeod[3], BSc, MSc

[1]Department of Lab Medicine and Pathology, University of Alberta, Edmonton, AB, Canada
[2]NGIS (Australia), Victoria, Australia
[3]Deptartment of Environmental Studies, University of Victoria, Victoria, BC, Canada

Corresponding Author:
David Beyer, BSc, MD
Department of Lab Medicine and Pathology
University of Alberta
5-411 Edmonton Clinic Health Academy
Edmonton, AB, T6G 1C9
Canada
Phone: 1 (780) 492 3111
Email: dbbeyer@ualberta.ca

Abstract

Background: Artificial intelligence (AI) models are increasingly being developed to improve the efficiency of pathological diagnoses. Rapid technological advancements are leading to more widespread availability of AI models that can be used by domain-specific experts (ie, pathologists and medical imaging professionals). This study presents an innovative AI model for the classification of colon polyps, developed using AutoML algorithms that are readily available from cloud-based machine learning platforms. Our aim was to explore if such AutoML algorithms could generate robust machine learning models that are directly applicable to the field of digital pathology.

Objective: The objective of this study was to evaluate the effectiveness of AutoML algorithms in generating robust machine learning models for the classification of colon polyps and to assess their potential applicability in digital pathology.

Methods: Whole-slide images from both public and institutional databases were used to develop a training set for 3 classifications of common entities found in colon polyps: hyperplastic polyps, tubular adenomas, and normal colon. The AI model was developed using an AutoML algorithm from Google’s VertexAI platform. A test subset of the data was withheld to assess model accuracy, sensitivity, and specificity.

Results: The AI model displayed a high accuracy rate, identifying tubular adenoma and hyperplastic polyps with 100% success and normal colon with 97% success. Sensitivity and specificity error rates were very low.

Conclusions: This study demonstrates how accessible AutoML algorithms can readily be used in digital pathology to develop diagnostic AI models using whole-slide images. Such models could be used by pathologists to improve diagnostic efficiency.

(JMIR Form Res 2025;9:e67457) doi:10.2196/67457

KEYWORDS

AI; AI models; applicability; artificial intelligence; automated detection; cancer; cancer screening; colon cancer; colon; colon cancer screening; colon polyp; detection; develop; development; digital pathology; effectiveness; imaging; machine learning; model; pathologist; pathology; screening; whole slide imaging

Introduction

Many important pathological diagnoses are made by expert pathologists’ careful examination of formalin-fixed paraffin-embedded tissue slides. Advances in digital microscopy have enabled large-scale digitization of glass slides at high resolution, and the adoption of whole-slide images (WSIs) for primary sign-out of pathology is increasing [1-3]. One of the main benefits of digital pathology is improved diagnostic efficiency, which is increasingly important as the field of pathology deals with increased volumes while also struggling

XSL•FO
RenderX

with the recruitment of new pathologists [4,5]. The combination of a decreased workforce coupled with an increase in the volume of cases secondary to an aging population means that pathologists need to become more efficient to meet future demand. Artificial intelligence (AI) tools applied to WSIs can be used to improve the efficiency of pathologists [6-8]. Advancements in slide scanners [9], which facilitate large-scale digitization of slides, have brought about a paradigm shift in pathology. Digitization not only enhances the efficiency of pathological examinations but also bridges the gap between conventional techniques and the ever-evolving field of AI. Ultimately, the creation of WSIs is the first step in incorporating AI into the field of pathology.

One advantage of the digitization of WSIs will be the creation of libraries of high-quality labeled training data for use with machine learning (ML) algorithms [10]. Recent developments in the fields of ML and AI, such as deep learning, for computer vision and object detection–related tasks [11-13] have led to a rapid uptake of the use of these tools in computational pathology research, where their utility has been widely recognized [14-17]. ML has traditionally required massive computational power and advanced knowledge of computer science and programming languages such as Python and R [18-20]. However, with the large-scale deployment of Machine Learning as a Service (MLaaS) platforms, these barriers to entry are minimized, allowing domain-specific experts (ie, pathologists and medical imaging professionals) to make use of advanced AI/ML tools [21]. “AutoML” algorithms and cloud-based ML platforms such as Amazon’s Sagemaker and Google’s VertexAI provide affordable, easy-to-access options that reduce overall costs by allocating centralized computer resources on demand to end users [22].

Pathologists are well-positioned with the expertise and tools required to build high-quality training datasets, which are the bedrock of effective AI models, and develop real-world uses for the production of ML models. Our project examined whether a small dataset for common colon polyp entities could be used to develop a robust and accurate ML model for diagnostic purposes using an AutoML model from Google’s Vertex AI. Colon polyps are a precursor to invasive carcinoma, and a high-volume sample is encountered in the pathology lab. As many jurisdictions use screening programs to detect and remove polyps for cancer surveillance, accurate and efficient pathology diagnosis is a key part of colon cancer screening programs [23,24]. As there are relatively few diagnostic entities for colon polyps, this area is well-suited to AI screening algorithms to assist pathologists in making a rapid diagnosis. This led to the formation of our project, aiming at evaluating an MLaaS model, trained on our own institutional data, to evaluate both the ease of model development as well as model performance when compared to pathologist interpretation.

Methods

Overview

Alberta Precision Laboratories has a robust digital pathology slide set, used primarily for teaching. This slide set contains images of previous cases (hematoxylin and eosin–stained histology slides) that have been scanned using an Aperio GT450. The bulk of our case images came from this dataset. In addition, to increase variability within our dataset, both for hematoxylin and eosin stain quality, as well as scanner type, we also used publicly accessible WSIs from Leeds University, University of Michigan, and the Cancer Imaging Archive [25-27]. Cases (n=494) were randomly divided into 3 allocations: training (75%), validation (10%), and test (15%, Table 1).

Table 1. Data allocation.

Image allocation	Number of images used, n	Percent of total
Training	1110 (370 cases)	75%
Validation	150 (50 cases)	10%
Test	222 (74 cases)	15%

In Table 1, cases and their associated images were split into 75% training, 10% validation, and 15% test. Slightly increased test cases were used (that the traditional 80/10/10 split) to better evaluate model performance, and to compensate for a relatively small training set.

We focused on 3 common entities seen in colon polyps: hyperplastic polyps, tubular adenomas, and normal colon. All cases were opened in Aperio Imagescope, and manual image/patch extraction was completed at 4× and 10× objective power (40× and 100× magnification, respectively), focusing on the most representative tumor/diagnostic areas. A total of one 4× extraction/tile and two 10× extraction/tiles were used per case, for a total of 1482 images (Table 1). Tiles were chosen by a pathologist to ensure that the most representative areas of the slide were used. The extraction was carried out using Aperio-Imagescope’s built-in image extraction tool, using the embedded International Color Consortium (ICC) profile and exported in .jpeg format. The ICC profile is a method of color normalization used by Aperio (and other digital pathology vendors) to ensure the image generated from the slide matches as closely as possible to the real-world color profile of the slide. Both the scanner and image profile used in this study are fully validated and accredited for clinical use.

Images in the dataset were rescreened by a blinded pathologist to ensure that the appropriate diagnostic label had been applied and to ensure that no images from test cases were present in either the training or validation data. A total of 3 users (2 pathologists and 1 pathology resident) were involved in the tile and review process to prevent bias in the tile selection. Instructions for tile selection were to select the most appropriate tile for the diagnosis. Only tiles with perfect consensus (3 of 3 agreed) labels were chosen to be used.

XSL•FO
RenderX

A single-label image classification model with 3 labels ("hyperplastic polyp," "tubular adenoma," and "normal") was then developed using Vertex AI and an autoML algorithm. Other model parameters were set to default to demonstrate a general yet easy to use model for nontechnical experts (model details are as follows: training method: "AutoML," Objective: "Image classification (single-label)," Data-split: "Manual," Budget: 48 node hours, Actual: 43.04 node hours, Training time: 5 h 33 min).

The AI model was tested on the "test" allocation of cases that were not part of the training or validation datasets. XRAI overlay [28] was used to view explainability and ensure that the algorithm was identifying the pertinent areas of the image (Figure 1). Model output was evaluated on a per-image basis, as opposed to a per-case basis. Results from the test data were used to evaluate precision and recall. VertexAI incorporates model evaluation and automatically generates both the precision-recall curve as well as precision-recall by threshold. An overall confidence threshold of 0.50 was chosen for the evaluation of our model.

Figure 1. (A) Normal colon sample image. (B) Normal colon sample image with XRAI (4) overlay. (C) Hyperplastic polyp sample image. (D) hyperplastic polyp sample image with XRAI overlay. Green intensity correlates with areas of increased positive probability, that is, image segments that contribute most strongly to a given class prediction.

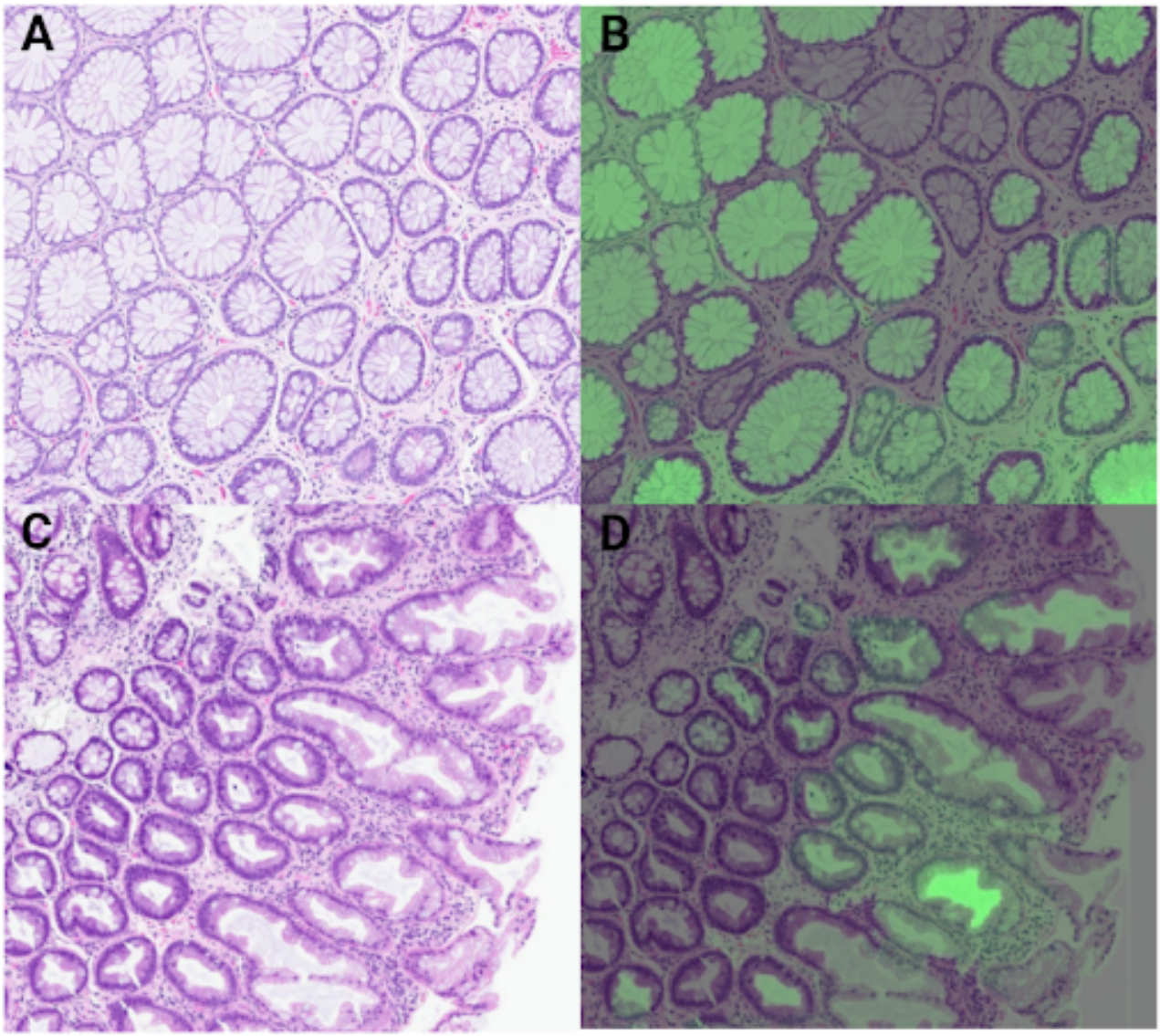

Ethical Considerations

Ethics approval and a waiver of consent were obtained for using deidentified case images from our institutional database. This study was approved by the Health Research Ethics Board of Alberta (HREBA.CC-23-0347).

Results

Using a CI of 0.5, the overall accuracy of the model on the test dataset was 98.4% (Table 2). Tubular adenomas and hyperplastic polyps were identified 100% of the time (66/66 and 48/48, respectively). Normal colon was identified with 97% accuracy (102/105, with 3/105 being misclassified as "hyperplastic"). Visual inspection of XRAI overlays demonstrates that the model

XSL•FO
RenderX

is identifying the pertinent areas (Figure 1). Results from the classification of the test data were used to calculate both recall as well as precision, in addition to a precision-recall by threshold curve (Figure 2), and an area under the curve value of 0.99.

Table 2. Tubular adenoma and hyperplastic polyps were identified 100% of the time (66/66 and 48/48, respectively). Normal colon was identified with 97% accuracy (102/105).

True label ↓/Predicted label →	Tubular adenoma, n (%)	Hyperplastic polyp, n (%)	Normal colon, n (%)
Tubular adenoma	66 (100)	0 (0)	0 (0)
Hyperplastic polyp	0 (0)	48 (100)	0 (0)
Normal colon	0 (0)	3 (3)	105 (97)

Figure 2. Precision-recall curve by threshold. Using a confidence value of 0.25 results in a precision of 98.7%, and recall of 100%. Using a confidence value of 0.5 results in both a precision and recall of 98.6%. Using a confidence value of 0.75 results in a precision of 100% and a recall of 97.3%. Precision sharply falls below a confidence threshold of 0.13. Recall sharply falls above a confidence threshold of 0.80. Our model output data are based on a confidence value of 0.50.

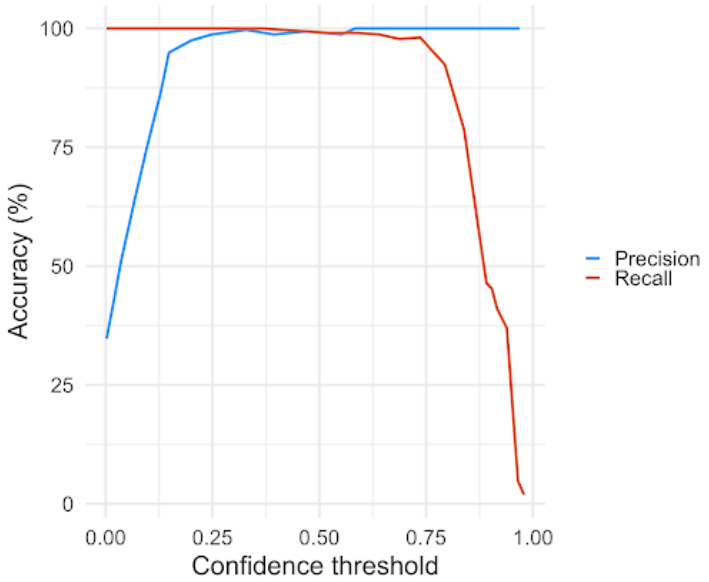

Discussion

Principal Findings

The integration of AI in the field of digital pathology is not merely a technological advancement; it is becoming a necessity driven by the contemporary challenges that the medical community faces. Here we have demonstrated that it is relatively easy to train your own AI algorithms on your own data, as a pathologist. As ML continues to revolutionize diagnostic pathology, our study demonstrates how simple off-the-shelf AI tools can readily be used to develop effective models for improving diagnostic efficiency. The recall and precision of the developed AI model in detecting colon polyps were remarkable, approaching 100%. This exceptional performance underscores the potential of ML to transform diagnostic pathology, especially given the rapid advancements in technology, availability, and cost-effectiveness of MLaaS platforms such as Google's VertexAI.

Previous AI applications in digital pathology, specifically in colorectal cancer screening, have shown great potential in increasing the efficiency and accuracy of diagnosis. For instance, Korbar et al [29] demonstrated that deep learning models could classify colorectal histology slides with high accuracy, bridging the gap between manual microscopic evaluations and automated assessments. Our study builds upon these foundations, offering further evidence for the efficiency and effectiveness of AI in this domain. One could say we have crafted a model that emulates the performance of an early-stage pathology resident, as the 3 entities used in this model are relatively simple; the difference is that this model took days to train, versus years for the average pathology resident. Nevertheless, there are certain limitations to consider.

First, our study had a limited sample size and did not include a complete range of diagnostic entities. With only tubular adenoma, hyperplastic polyp, and normal colon as labels, this model does not account for other critical entities like serrated adenomas, high-grade dysplasia, or carcinoma, an important histologic feature of high-risk polyps [30]; this would be a good area for future projects. AI models, particularly deep learning models, generally require large training datasets in order to generalize well in real-world scenarios [18,31]. A more extensive dataset might allow for more refined models that capture nuances and rare presentations of colon polyps, such as high-grade dysplasia. This limitation is significant, as evidenced by the emphasis on broad scope in successfully used deep neural networks for detecting colorectal cancer on WSIs [32].

XSL•FO
RenderX

Second, while our model shows promise, integrating it into real-world workflows would necessitate the inclusion of a more diverse range of clinical criteria beyond just the image data. We did not include certain clinical characteristics that would be important, like location (ie, ascending colon), which can affect diagnosis, especially for serrated lesions [23].

Third, the need to manually annotate our dataset means that we are effectively working with a "best-case scenario" data input. Real-world scenarios might present slides with artifacts, suboptimal staining, or overlapping tissues that could challenge the model's predictions [33]. The success of AI models in clinical settings largely depends on the quality and diversity of the input data. As we only used tiles with consensus between our 3 subject matter experts (2 pathologists and 1 pathology resident), this meant a small number of imperfect tiles were excluded. These "imperfect" samples, of course, exist in the real world, and these must be interpreted as well.

Fourth, our model misclassified a number of "normal" as "hyperplastic polyp," which appeared to be on normal colon where glands were not perfectly round. Likely inclusion of more "normal" images with different-shaped lumens would help tune the model for proper identification.

However, despite these limitations, our study exemplifies the promise of using ML as a service in histopathology. The near-perfect performance of our model in a controlled environment suggests that with adequate training data and a broader range of diagnostic entities, such platforms could soon play a significant role in augmenting the capabilities of pathologists.

Our project has shown that widely available "AutoML" algorithms, such as Google's Vertex AI, can be applicable to the medical field, specifically in digital pathology. Our training model was robust with a sensitivity approaching 100% and a specificity of 98%. This aligns with other authors' findings of success with AutoML algorithms in the medical field [31]. Such AI tools will help to improve the efficiency of front-line pathologists, allowing them to keep up with increasing demand and so helping to mitigate workforce constraints. For example, our model could be used to auto-generate a report, saving time for the pathologist in dictating/typing.

Conclusion

Overall, we show that cloud-based ML platforms can produce accurate models that are specific to the field of pathology and WSIs. Furthermore, they are easy to use, with only a relatively basic level of ML knowledge required. As the medical community strives for implementation of precision medicine that should lead to improved patient outcomes [34], integration of ML tools into the realm of histopathology may soon become indispensable. Future studies with larger datasets, diverse diagnostic entities, and more real-world scenarios are warranted to further elucidate the potential of AI in this domain.

Authors' Contributions

DB was involved in conceptualization, data curation, and writing of the initial draft. LM was involved in supervision, validation, and writing (review and editing). ED was involved in the validation and writing (review and editing).

Conflicts of Interest

None declared.

References

1. Louis DN, Feldman M, Carter AB, Dighe AS, Pfeifer JD, Bry L, et al. Computational pathology: a path ahead. Arch Pathol Lab Med 2016;140(1):41-50 [FREE Full text] [doi: 10.5858/arpa.2015-0093-SA] [Medline: 26098131]
2. Zarella MD, Bowman D, Aeffner F, Farahani N, Xthona A, Absar SF, et al. A practical guide to whole slide imaging: a white paper from the digital pathology association. Arch Pathol Lab Med 2019;143(2):222-234 [FREE Full text] [doi: 10.5858/arpa.2018-0343-RA] [Medline: 30307746]
3. Rizzo PC, Caputo A, Maddalena E, Caldonazzi N, Girolami I, Dei Tos AP, et al. Digital pathology world tour. Digit Health 2023;9:20552076231194551 [FREE Full text] [doi: 10.1177/20552076231194551] [Medline: 37654717]
4. Ho J, Ahlers SM, Stratman C, Aridor O, Pantanowitz L, Fine JL, et al. Can digital pathology result in cost savings? A financial projection for digital pathology implementation at a large integrated health care organization. J Pathol Inform 2014;5(1):33 [FREE Full text] [doi: 10.4103/2153-3539.139714] [Medline: 25250191]
5. Lujan G, Quigley JC, Hartman D, Parwani A, Roehmholdt B, Meter BV, et al. Dissecting the business case for adoption and implementation of digital pathology: a white paper from the digital pathology association. J Pathol Inform 2021;12:17 [FREE Full text] [doi: 10.4103/jpi.jpi_67_20] [Medline: 34221633]
6. Hanna MG, Reuter VE, Ardon O, Kim D, Sirintrapun SJ, Schüffler PJ, et al. Validation of a digital pathology system including remote review during the COVID-19 pandemic. Mod Pathol 2020;33(11):2115-2127 [FREE Full text] [doi: 10.1038/s41379-020-0601-5] [Medline: 32572154]
7. Schüffler PJ, Geneslaw L, Yarlagadda D, Hanna M, Samboy J, Stamelos E, et al. Integrated digital pathology at scale: a solution for clinical diagnostics and cancer research at a large academic medical center. J Am Med Inform Assoc 2021;28(9):1874-1884 [FREE Full text] [doi: 10.1093/jamia/ocab085] [Medline: 34260720]

XSL•FO
RenderX

8. Marletta S, Eccher A, Martelli FM, Santonicco N, Girolami I, Scarpa A, et al. Artificial intelligence-based algorithms for the diagnosis of prostate cancer: a systematic review. Am J Clin Pathol 2024;161(6):526-534. [doi: 10.1093/ajcp/aqad182] [Medline: 38381582]
9. Rizzo PC, Girolami I, Marletta S, Pantanowitz L, Antonini P, Brunelli M, et al. Technical and diagnostic issues in whole slide imaging published validation studies. Front Oncol 2022;12:918580 [FREE Full text] [doi: 10.3389/fonc.2022.918580] [Medline: 35785212]
10. Hanna MG, Singh R, Parwani AV. Whole slide imaging: applications in education. In: Whole Slide Imaging. Cham: Springer; 2022:95-103.
11. Bankhead P. Developing image analysis methods for digital pathology. J Pathol 2022;257(4):391-402 [FREE Full text] [doi: 10.1002/path.5921] [Medline: 35481680]
12. Ben Hamida A, Devanne M, Weber J, Truntzer C, Derangère V, Ghiringhelli F, et al. Deep learning for colon cancer histopathological images analysis. Comput Biol Med 2021;136:104730. [doi: 10.1016/j.compbiomed.2021.104730] [Medline: 34375901]
13. Rakha EA, Toss M, Shiino S, Gamble P, Jaroensri R, Mermel CH, et al. Current and future applications of artificial intelligence in pathology: a clinical perspective. J Clin Pathol 2021;74(7):409-414. [doi: 10.1136/jclinpath-2020-206908] [Medline: 32763920]
14. Acs B, Rantalainen M, Hartman J. Artificial intelligence as the next step towards precision pathology. J Intern Med 2020;288(1):62-81 [FREE Full text] [doi: 10.1111/joim.13030] [Medline: 32128929]
15. Janowczyk A, Madabhushi A. Deep learning for digital pathology image analysis: a comprehensive tutorial with selected use cases. J Pathol Inform 2016;7:29 [FREE Full text] [doi: 10.4103/2153-3539.186902] [Medline: 27563488]
16. Erickson BJ, Korfiatis P, Akkus Z, Kline TL. Machine learning for medical imaging. Radiographics 2017;37(2):505-515 [FREE Full text] [doi: 10.1148/rg.2017160130] [Medline: 28212054]
17. Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. Nat Med 2018;24(10):1559-1567 [FREE Full text] [doi: 10.1038/s41591-018-0177-5] [Medline: 30224757]
18. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521(7553):436-444. [doi: 10.1038/nature14539] [Medline: 26017442]
19. Campanella G, Hanna MG, Geneslaw L, Miraflor A, Werneck Krauss Silva V, Busam KJ, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat Med 2019;25(8):1301-1309 [FREE Full text] [doi: 10.1038/s41591-019-0508-1] [Medline: 31308507]
20. Ho C, Zhao Z, Chen XF, Sauer J, Saraf SA, Jialdasani R, et al. A promising deep learning-assistive algorithm for histopathological screening of colorectal cancer. Sci Rep 2022;12(1):2222 [FREE Full text] [doi: 10.1038/s41598-022-06264-x] [Medline: 35140318]
21. Ribeiro M, Grolinger K, Capretz MAM. MLaaS: machine learning as a service. 2015 Presented at: IEEE 14th International Conference on Machine Learning and Applications (ICMLA); 2015 December 09-11; Miami, FL, USA p. 896-902. [doi: 10.1080/15265161.2015.1028658]
22. Berg G. Image classification with machine learning as a service: - a comparison between Azure, SageMaker, and Vertex AI.: Linnaeus University; 2022. URL: https://urn.kb.se/resolve?urn=urn:nbn:se:lnu:diva-113829 [accessed 2025-05-16]
23. Bujanda L, Cosme A, Gil I, Arenas-Mirave JI. Malignant colorectal polyps. World J Gastroenterol 2010;16(25):3103-3111 [FREE Full text] [doi: 10.3748/wjg.v16.i25.3103] [Medline: 20593495]
24. Rex DK, Boland CR, Dominitz JA, Giardiello FM, Johnson DA, Kaltenbach T, et al. Colorectal cancer screening: recommendations for physicians and patients from the U.S. multi-society task force on colorectal cancer. Am J Gastroenterol 2017;112(7):1016-1030. [doi: 10.1038/ajg.2017.174] [Medline: 28555630]
25. Slide Library, Virtual Pathology at the University of Leeds. 2022. URL: http://www.virtualpathology.leeds.ac.uk/slides/library/ [accessed 2022-01-12]
26. Slide Library, University of Michigan Virtual Slide Box. 2022. URL: https://www.pathology.med.umich.edu/apps/slides/ [accessed 2022-01-12]
27. Slide Library, Histopathology imaging on TCIA (The Cancer Imaging Archive). 2022. URL: https://www.cancerimagingarchive.net/histopathology-imaging-on-tcia/ [accessed 2022-01-12]
28. Kapishnikov A, Bolukbasi T, Viégas FB, Terry M. XRAI: better attributions through regions. In: Ann Nutr Metab.: IEEE; 2019 Presented at: IEEE/CVF International Conference on Computer Vision (ICCV); 2019 October 27-02 November; Seoul, Korea (South). [doi: 10.1159/000508056]
29. Korbar B, Olofson AM, Miraflor AP, Nicka CM, Suriawinata MA, Torresani L, et al. Looking under the hood: deep neural network visualization to interpret whole-slide image analysis outcomes for colorectal polyps. 2017 Presented at: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); July 21-26, 2017; Honolulu, HI, USA p. 821-827.
30. Gschwantler M, Kriwanek S, Langner E, Göritzer B, Schrutka-Kölbl C, Brownstone E, et al. High-grade dysplasia and invasive carcinoma in colorectal adenomas: a multivariate analysis of the impact of adenoma and patient characteristics. Eur J Gastroenterol Hepatol 2002;14(2):183-188. [doi: 10.1097/00042737-200202000-00013] [Medline: 11981343]

XSL•FO
RenderX

31. Basu S, Mitra S, Saha N. Deep learning for screening COVID-19 using chest x-ray images. medRxiv 2020. [doi: 10.1101/2020.05.04.20090423]
32. Schwen LO, Schacherer D, Geißler C, Homeyer A. Evaluating generic autoML tools for computational pathology. Inform Med Unlocked 2022;29:100853. [doi: 10.1016/j.imu.2022.100853]
33. Komura D, Ishikawa S. Machine learning methods for histopathological image analysis. Comput Struct Biotechnol J 2018;16:34-42 [FREE Full text] [doi: 10.1016/j.csbj.2018.01.001] [Medline: 30275936]
34. Aeffner F, Zarella MD, Buchbinder N, Bui MM, Goodman MR, Hartman DJ, et al. Introduction to digital image analysis in whole-slide imaging: a white paper from the digital pathology association. J Pathol Inform 2019;10:9 [FREE Full text] [doi: 10.4103/jpi.jpi_82_18] [Medline: 30984469]

Abbreviations

AI: artificial intelligence
ICC: International Color Consortium
ML: machine learning
MLaaS: Machine Learning as a Service
WSI: whole-slide image

Edited by J Sarvestan; submitted 11.10.24; peer-reviewed by S Marletta, LR Guo; comments to author 15.11.24; revised version received 22.03.25; accepted 14.05.25; published 31.07.25.

Please cite as:
Beyer D, Delancey E, McLeod L
Automating Colon Polyp Classification in Digital Pathology by Evaluation of a "Machine Learning as a Service" AI Model: Algorithm Development and Validation Study
JMIR Form Res 2025;9:e67457
URL: https://formative.jmir.org/2025/1/e67457
doi:10.2196/67457
PMID:

©David Beyer, Evan Delancey, Logan McLeod. Originally published in JMIR Formative Research (https://formative.jmir.org), 31.07.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Mitigating the Negative Effects of Internet Browsing on Young People's Resilience and Outlook on Life Through Classic Grimms' Fairy Tales: Exploratory Randomized Controlled Study

Congcong Hou[1], MSc, PhD; Thomas Foscht[2], MSc, PhD; Barbara Duffek[3], MSc, PhD; Annisa Arigayota[4], BSc; Andreas Benedikt Eisingerich[4], MPhil, PhD

[1]Faculty of Commerce, Kyushu Sangyo University, Fukuoka, Japan
[2]Department of Marketing, University of Graz, Graz, Austria
[3]Robinson College of Business, Georgia State University, Atlanta, GA, United States
[4]Imperial College Business School, Imperial College London, London, United Kingdom

Corresponding Author:
Andreas Benedikt Eisingerich, MPhil, PhD
Imperial College Business School
Imperial College London
South Kensington Campus
London, SW7 2AZ
United Kingdom
Phone: 44 740333 ext 6597
Email: a.eisingerich@imperial.ac.uk

Abstract

Background: Internet browsing is a daily activity for many young people. However, how internet browsing affects young people's resilience and positive (vs negative) outlook on life remains largely unaddressed. Critically, how reading classical fairy tales may help mitigate the influence of internet browsing on resilience and foster a more positive rather than negative outlook on life has yet to be explored.

Objective: This study examines the influence of internet browsing on young people's resilience and positive (vs negative) outlook on life. Furthermore, this study aims to examine the potential mitigating effect of reading classical Grimms' fairy tales, such as Hansel and Gretel and Little Red Riding Hood, on the relationship between internet browsing and postgraduate students' resilience and outlook on life.

Methods: A randomized controlled study was conducted using a 2 (internet browsing vs no internet browsing) × 2 (reading a classical fairy tale vs no classical fairy tale) between-subjects design. All study participants (N=412) were postgraduate students and randomly assigned to one of the study's 4 conditions and answered a brief questionnaire, examining their resilience and positive versus negative outlook on life. To examine the potential mitigating effect of classical fairy tales on the relationship between internet browsing and resilience as well as positive versus negative outlook on life, we conducted an exploratory bootstrapping-based moderated mediation analysis with 5000 resamples.

Results: The results showed a significant moderating role of reading classical Grimms' fairy tales on the negative effect of internet browsing on postgraduate students' resilience and outlook on life. Specifically, when study participants browsed the internet, they reported a more positive outlook on life when they read a Grimms' fairy tale (read fairy tale: mean 5.46, SD 0.151 vs not read fairy tale: mean 3.01, SD 0.150, SE 0.213, 95% CI –2.860 to –2.024; P<.001). Furthermore, the results showed that when participants browsed the internet, they indicated significantly greater resilience when they read a Grimms' tale (mean 4.62, SE 0.179, 95% CI 4.271-4.976) than when they did not (mean 2.59, SE 0.179, 95% CI 2.243-2.945). In addition, an exploratory analysis demonstrated that the effect of internet browsing on outlook on life is mediated by resilience (effect 0.85, SE 0.17, 95% CI 0.52-1.20).

Conclusions: The findings of this study show that reading a classical Grimms' fairy tale, such as Hansel and Gretel or Little Red Riding Hood, helped mitigate the negative effects of internet browsing on postgraduate students' resilience and outlook on life.

Trial Registration: ISRCTN 16972408; https://www.isrctn.com/ISRCTN16972408

XSL•FO
RenderX

(JMIR Form Res 2025;9:e76770) doi:10.2196/76770

KEYWORDS

internet browsing; classical fairy tales; Brothers Grimm; resilience; outlook on life; randomized controlled study

Introduction

Overview on Internet Browsing, Young People's Resilience, and Outlook on Life

Internet browsing has the potential to be informative and enriching, but unfiltered use can also have negative psychological effects, especially on younger users who are still developing their sense of self and resilience. Young people who frequently browse news feeds and social media can be inundated with distressing headlines, crisis updates, and alarming videos [1]. This continuous exposure to negative events, often without the buffer of context or constructive discussion, can fuel anxiety and pessimism about the future. Moreover, various internet platforms often showcase carefully curated highlight reels, which can lead to unrealistic expectations and self-comparison [2]. Constantly measuring one's life against seemingly "perfect" snapshots can erode self-esteem and foster a negative self-image. That is, being bombarded with negative or unrealistic messages can create a sense that the world is filled with endless threats and personal inadequacies.

Numerous platforms on the internet, including social media, are strongly addictive, and young people may find themselves hooked [3,4]. Relying heavily on online communication can also limit meaningful face-to-face interactions, as in-person encounters often require young people to practice social skills such as reading body language, managing conflicts, and empathizing with peers. Furthermore, the internet can provide a form of instant escape [5,6], browsing videos, memes, or messaging apps, whenever stress or discomfort arises. Resilience, however, develops partly through experiencing and overcoming manageable stressors, such as small disagreements with friends or navigating unfamiliar situations. If young people rarely face these real-life challenges, or if they continuously escape them through digital means, they may struggle to build strong coping mechanisms.

Moreover, social media platforms as well as search engines and even various news sites often use algorithms that personalize content [7-9]. If a user frequently engages with negative, sensational, or critical information, the algorithm may feed them more of the same, creating a feedback loop of negativity [10]. People also naturally gravitate toward content that reaffirms their existing beliefs and feelings [11]. A young person feeling discouraged may seek out material that confirms a pessimistic worldview, reinforcing a negative outlook and limiting exposure to hopeful or solution-oriented perspectives. When the online environment consistently mirrors or amplifies negative thoughts, it can be difficult for young people to see alternative viewpoints or foster optimism [12,13]. This echo chamber effect can cement a sense of fatalism and hopelessness [14].

The internet provides an endless stream of entertainment and quick answers. As a result, some young users may become accustomed to immediate gratification [15,16]. This can weaken patience, perseverance, and the ability to tolerate frustration when tasks become challenging. Rapid scrolling through social media feeds, video clips, shorts, or headlines can condition the brain to seek quick hits of novelty and pleasure. This makes it harder to stay focused on sustained tasks, such as reading a book, thinking through a difficult problem, or engaging in a complex real-world project.

Resilience often involves the ability to concentrate on long-term goals and persevere despite difficulties. When young people are conditioned to expect instant results or quick solutions, they may be more prone to giving up in the face of real-world hurdles. In addition, the anonymity of the internet can embolden bullying behavior [17,18]. Young people who are targeted may experience heightened stress, anxiety, and feelings of isolation, which can diminish their ability to cope with adversity. Seeing peers' social activities in real time can also lead to feelings of exclusion or jealousy if they are not included [19-21]. This ongoing sense of missing out can erode self-esteem and overall life satisfaction. A vital component of resilience is a supportive social network. When online spaces become sources of bullying or pressure, it undermines that support and can negatively affect a young person's self-confidence and emotional well-being [22-25].

In addition, internet browsing may foster a largely passive role, scrolling through feeds, or watching videos, leading to fewer opportunities for skill-building or tangible accomplishments. The upshot is that many forms of digital engagement encourage consumption over creation. Young people may spend less time engaging in problem-solving, artistic pursuits, or physical activities that build self-efficacy and pride in achievement [26-28]. In this research, we define resilience as the capacity to adapt positively to adversity, major changes, or setbacks by managing emotional responses, maintaining persistence, and viewing challenges as opportunities for growth. Outlook on life is defined as a future-oriented psychological disposition characterized by positive anticipation, hope, and confidence that life events will unfold favorably. Thus, as part of this research, we aimed to answer the following research questions: To what extent, if at all, does internet browsing affect young people's self-expressed resilience and outlook (positive vs negative) on life. Furthermore, we aimed to answer how reading a classical Grimms' fairy tale may help mitigate the impact of internet browsing on young people's resilience and outlook on life. We discuss the potential influence of classical fairy tales in the next section.

Overview on Grimms' Classical Fairy Tales, Young People's Resilience, and Outlook on Life

Classical Brothers Grimm fairy tales, such as Hansel and Gretel or Little Red Riding Hood, have endured for centuries precisely because they speak to fundamental human experiences of challenge, danger, and ultimate survival [29]. Grimms' fairy tales often introduce universal anxieties (fear of abandonment,

XSL•FO
RenderX

danger from strangers, isolation in the woods) in a way that is contained within a story [30]. Young and older people who engage with these tales can process and explore fears vicariously [31,32]. Because the stories are set in a fantastical or once-upon-a-time realm, it becomes psychologically safer to confront these themes [33-35].

Furthermore, Grimms' classical fairy tales can affirm moral values such as bravery, kindness, and clever thinking in the face of wrongdoing. Hansel and Gretel teaches the importance of sticking together and helping one another, while Little Red Riding Hood underscores the risks of straying from the path or trusting too easily. These lessons can instill a sense of moral clarity that contributes to confidence and resilience in real-life choices [36]. Despite their dark elements, many classic tales promise hope in the form of a "happy ending" [37]. This element reassures readers that hardships, no matter how dire, can be resolved and that it is possible to emerge from difficult times intact and, indeed, stronger. As young people read or listen to these stories, they learn the language of emotions: fear, relief, joy, sadness, and hope. Building an emotional vocabulary helps them understand and express their own feelings, which is a cornerstone of emotional resilience [38,39].

Background on The Moderating Effect of Grimms' Fairy Tales on the Influence of Internet Browsing on Resilience and Outlook on Life

Internet browsing has become an integral part of daily life, especially for young people who have grown up with digital technology. Although it offers numerous benefits such as access to information, there are ways in which it might weaken resilience and contribute to a more negative outlook on life [40]. For instance, the internet provides an unending stream of news, much of which can be negative, focusing on crises, disasters, and conflicts [41-43]. Continuous exposure to such content can lead to anxiety, desensitization to violence, and a worldview that emphasizes the negative aspects of life especially for younger people [42-44]. Moreover, online interactions can expose young people to bullying [45,46], which can be more pervasive and relentless than traditional forms due to the anonymity and reach of the internet. This can significantly impact self-esteem, increase anxiety, and decrease resilience.

Platforms often showcase idealized or curated versions of others' lives, leading to unrealistic comparisons [47]. Young people might feel inadequate or that their lives are less fulfilling, which can foster envy, depression, and a negative self-image, undermining personal resilience [25,40,48]. In addition to this, the constant exposure to others' experiences can make individuals feel they are missing out on life, leading to dissatisfaction with their own lives and a diminished capacity to find contentment in everyday moments [49,50]. Moreover, the internet is a vast repository of information, which can lead to cognitive overload. This constant switching between tasks or pieces of information can reduce the ability to focus on and solve complex problems, potentially weakening resilience when faced with real-life challenges.

There may also be the temptation for and a tendency toward shallow browsing rather than deep learning or engagement [51,52], which might not foster the deep thinking or critical analysis necessary for building resilience through overcoming long-term or complex issues. Although the internet connects people globally, it can also lead to physical isolation, reducing face-to-face interactions that are crucial for developing social skills, empathy, and support networks that are foundational for resilience [52,53]. The nuances of human interaction, like reading body language or dealing with conflict in person, might be less practiced, leading to difficulties in handling real-life social complexities. The internet also frequently provides instant access to information, entertainment, and social validation, which can skew perceptions of effort versus reward in real life. This might lead to impatience or a lack of perseverance when faced with tasks that require sustained effort, thus affecting resilience [39].

Although classical Grimms' fairy tales include dark or frightening elements, these stories offer powerful lessons in overcoming adversity. For instance, Grimms' tales are known for their darker undertones, showing children in perilous situations. However, these narratives almost always culminate in liberation or redemption, conveying that hope can persist even when circumstances seem dire. Seeing peers (albeit fictional) succeed independently fosters the belief that one is capable of self-reliance and problem-solving in moments of difficulty. This narrative structure underlines that, no matter how overwhelming a problem appears, there is a path to resolution, fostering optimism and fortitude in readers. In doing so, fairy tales allow children to experience fear, tension, and relief in a controlled, fictional context. They learn that it is possible to feel frightened but still move forward, a concept vital to developing emotional regulation skills.

Furthermore, traditional fairy tales tend to have clear moral lines. Acts of kindness, bravery, or cleverness bring the protagonists closer to safety; malicious or deceitful behavior is penalized. These straightforward moral frameworks help children internalize the belief that virtuous behavior is both valued and effective. Though often overshadowed by more dramatic elements, moments of kindness or cooperation (eg, siblings helping one another) highlight the importance of empathy. Recognizing this promotes healthy social-emotional development in children. Young readers see characters who suffer, facing hunger, abandonment, or trickery, and learn that hardship does not define one's destiny.

Thus, although steeped in dark imagery, the ultimate message of these tales is one of empowerment: Young people can confront formidable challenges and still emerge triumphant. In a world that can sometimes feel overwhelming, reading these classic fairy tales can serve as a reminder that adversity may be part of the journey, but so, too, is the promise of a brighter ending.

Internet browsing can sometimes blur moral lines or leave young people feeling confused or jaded, particularly when exposed to ambiguous news events or negative online behaviors. Fairy tales, by contrast, often highlight moral clarity and the redemptive power of virtue. Classic Grimms' tales are known for containing darker elements than modern retellings. Although this can be unsettling, it also provides a nuanced understanding

XSL•FO
RenderX

that the world has dangers, yet these dangers can be navigated. Critically, the fairy tale characters' mistakes and subsequent growth can teach resilience by showing that errors are part of the learning process, a lesson that can be overshadowed by the instant gratification and perfection often depicted online. These tales typically conclude with a resolution that underscores the possibility of overcoming adversity. This sense of adversity being an integral part of life's story, and that it can, indeed, be mastered are crucial for developing an "I can handle this" mindset, particularly in a digital era that often leaves negative events unresolved or amplified.

Methods

Study Design

The study was registered in ISRCTN. The completed CONSORT-SPI (Consolidated Standards of Reporting Trials statement for social and psychological interventions) 2018 checklist is included in Multimedia Appendix 1. As part of this study, we used a 2 (internet browsing vs no internet browsing) × 2 (reading a classical fairy tale vs no classical fairy tale) between-subjects experimental design. Over the course of 2 weeks in spring 2025, posters and flyers were used to invite postgraduate university students to take part in a study on student well-being and daily activities. Participants were randomly allocated to 1 of the 4 study conditions using a random 1-4 number generator (condition 1: internet browsing + reading a classical fairy tale; condition 2: internet browsing + no classical fairy tale; condition 3: no internet browsing + reading a classical fairy tale; condition 4 [control group]: no internet browsing + no classical fairy tale). The study was conducted in a lab experimental setting on a university campus.

In condition 1, participants were invited to spend 20 minutes to freely browse the internet on any digital device they had with them (eg, smartphone, tablet, laptop). Extra care was taken to inform study participants that they could browse the internet very freely and visit any site they want and that absolutely no data were collected based on which sites they visited. Research assistants were trained and instructed to look at a phone themselves, so that participants did not feel observed during the study time period. After 20 minutes of free internet browsing, participants were invited and given 20 minutes to read either the classical fairy tale Hansel and Gretel or Little Red Riding Hood (the Brothers Grimm version; randomly assigned, so that some participants read Hansel and Gretel and others read the Little Red Riding Hood fairy tale). Finally, study participants completed a brief survey. More specifically, across all 4 conditions, participants answered a brief survey with Likert-scale measurement items (1=strongly disagree, 9=strongly agree) capturing participants' positive versus negative outlook on life as well as resilience. Detailed measurement items are listed in Table 1. The measurement items to capture resilience were adapted from previously published work [54,55]. We used reverse-coded measurement items as part of our measurement of resilience and outlook to reduce response bias. The measurement items used for this study were also adapted from prior published research [56,57]. Again, reverse-coded measurement items were used as part of the scale to reduce response bias. In addition, in condition 1 and condition 2, participants answered questions about whether they enjoyed browsing the internet and found it interesting as controls. Furthermore, as controls, participants in condition 1 and condition 3 answered questions about whether they enjoyed reading the fairy tale and whether they were familiar with the fairy tale that they had read. Across all 4 conditions, participants were asked whether they could guess the study's purpose, and none did. In condition 2, participants were invited to spend 20 minutes to freely browse the internet as in condition 1. After browsing the internet, participants in condition 2 completed the brief survey. In condition 3, participants were randomly allocated to either read Hansel and Gretel or Little Red Riding Hood and given 20 minutes to read the fairy tale and subsequently complete a brief survey as in condition 1. Finally, in condition 4 (control group), participants simply completed the brief survey.

XSL•FO
RenderX

Table 1. Study measurement items and reliabilities.

Variable and measures	Response, mean (SD)	Response, range	Statistical test result	AVE[a]	CR[b]
Positive versus negative outlook on life			0.93[c]	0.62	0.87
"I look forward to each new day in my life."	4.18 (1.85)	1-9			
"I do not feel hopeful that things in my life will work out for the best."[d]	4.25 (2.02)	1-9			
"I find it difficult to believe that things will be better in the future."[d]	4.17 (2.04)	1-9			
"I generally feel positive about my future."	4.16 (1.86)	1-9			
Resilience			0.94[c]	0.79	0.94
"I believe I can handle whatever life throws at me."	3.51 (2.16)	1-9			
"I feel I can manage and adapt to major changes or disruptions in my life."	3.51 (2.09)	1-9			
"I am discouraged by setbacks and find it difficult to move forward in life."[d]	3.47 (2.12)	1-9			
"I view difficulties in life as opportunities to learn and grow stronger."	3.35 (2.15)	1-9			
Enjoyment of internet browsing			0.76[e]	0.86	0.92
"I very much enjoyed browsing the internet just now."	3.80 (1.85)	1-9			
"I found browsing the internet just now to be interesting."	3.76 (1.88)	1-9			
Enjoyment of reading the fairy tale			0.79[e]	0.90	0.95
"I enjoyed reading the fairy tale."	4.42 (1.68)	1-9			
"I did not like the fairy tale very much."[d]	4.44 (1.75)	1-9			
Familiarity with the fairy tale read			0.84[e]	0.90	0.95
"I am familiar with the fairy tale that I just read."	3.89 (2.15)	1-9			
"I know the fairy tale that I just read very well."	3.74 (2.11)	1-9			

[a]AVE: average variance extracted.
[b]CR: composite reliability.
[c]Cronbach α.
[d]Reverse coded.
[e]Pearson *r*.

Prior to our main study, we conducted a pretest (n=175) that followed the same recruitment procedure as discussed for our main study and included a neutral distraction task for condition 2, condition 3, and condition 4 to test whether the total time spent in the study affected the results. Specifically, for condition 2, after 20 minutes of internet browsing, study participants were provided with sheets with grids of random consonants and vowels and told "for the next twenty minutes, please find and cross out all instances of the letter 'X' and 'Z' on the pages provided. Feel free to work at a steady pace. This is not a test of skill, so just do your best to stay focused." Hence, in condition 2, after 20 minutes of internet browsing, study participants completed the letter cancellation task for 20 minutes, followed by the survey, as discussed in the previous section. Unlike tasks such as puzzles, games, or reading neutral text, letter cancellation was chosen as a distraction task as part of the pretest because it is unlikely to affect resilience or outlook and the task is not harmful or frustrating. The pretest results showed that participants' expressed resilience did not differ between condition 2 that did include the filler task (n=30; mean 2.48) and the one that did not include the filler task (n=28; mean 2.40; t_{56}=0.18, *P*=.86). Similarly, the pretest results demonstrated that participants' outlook did not differ across condition 2 with (mean 2.93) and without (mean 2.91; t_{56}=0.05, *P*=.96) the filler task. Moreover, in the pretest of condition 3, after 20 minutes of fairy tale reading, study participants completed the letter cancellation task, followed by the survey. The pretest results demonstrated that resilience did not differ between condition 3 that included the filler task (n=28; mean 3.20) and condition 3 that did not include the filler task (n=29; mean 3.22; t_{55}=0.05, *P*=.96). The pretest results also indicated that outlook did not differ across condition 3 with (mean 4.21) and without (mean 4.10; t_{55}=0.18, *P*=.86) the filler task. Finally, in the pretest for condition 4, participants completed the letter cancellation task for 40 minutes, followed by the survey. Pretest results showed that participants'

XSL•FO
RenderX

resilience did not differ between condition 4 with (n=30; mean 3.27) and without (n=30; mean 3.38; t_{58}=0.22, *P*=.83) the filler task. Further, the pretest showed that participants' outlook did not differ between condition 4 with (mean 4.08) and without (mean 3.97; t_{58}=0.19, *P*=.85) the filler task. Thus, total time spent in the study did not impact the results, and no filler task was used for our main study to help minimize participants' time in the study. After the pretest, all participants were debriefed about the purpose of the study and about the purpose of letter cancellation as a time-filler.

Ethical Considerations

All participants were informed that the data collected will inform academic research, will not be used for other commercial purposes, and will be anonymized and treated with strict confidentiality and that they were free to stop and withdraw from the study at any time and without giving a reason. Participants were informed that the data from participants who stopped early will not be used as part of the study. We received ethical approval from Kyushu Sangyo University's ethics committee (#2024-0018). Each study participant was thanked and received a US $5 note as a token of gratitude for taking the time to participate in the study. Written consent forms were obtained from all study participants after they were offered an information leaflet and time (minimum of 24 hours) for consideration was allowed.

Participants and Procedure

Our study used a 2×2 between-subjects design to examine the effects of internet browsing and fairy tale exposure on outlook and resilience. A total of 412 participants were recruited, exceeding 100 per condition. A sensitivity analysis was conducted using GPower (version 3.1) for a between-subjects ANOVA with 4 groups. With a total sample size of 412, α of .05, and power (1-β) of .80, the study was sufficiently powered, and the design had adequate sensitivity to detect meaningful main and interaction effects [58]. That is, this sample size provided >99% power to detect medium-sized effects (Cohen *f*≥0.25), which aligns with effect sizes observed in prior digital intervention and narrative psychology research (eg, positive psychology interventions: *d*=0.40-0.60). For smaller effects (*f*=0.15), power remained high at 99%. Even very small effects (*f*=0.10) are detectable with 86% power, surpassing conventional power thresholds (≥80%). This ensured robustness for our primary analyses (ANOVA main effects and interactions) while accommodating plausible effect magnitudes in behavioral health contexts. More specifically, a total of 420 university postgraduate students took part in this study. During the data collection process, 8 participants decided to stop and withdraw from the study. The data from the participants who decided to withdraw were not used as part of the study analyses, resulting in an effective total number of 412 participants for the study (condition 1: n=103; condition 2: n=104; condition 3: n=103; condition 4: n=102).

Data Analysis

We examined whether items loaded significantly on intended factors with low cross-loadings [59]. As a next step, we examined the estimates for the average variance extracted (AVE), supporting convergent validity [60] (see Table 1). Moreover, we assessed the possibility that measurement errors can vary in magnitude across items and calculated and compared the AVE for all pairs of constructs to the squared correlation between the 2 constructs of interest [61]. The squared correlation between any pair of constructs was not greater than the respective AVE for each of the constructs in the pair, supporting discriminant validity. In addition, Cronbach α were generally high (Table 1). In order to examine the potential mitigating effect of classic Grimms' fairy tales on the relationship between internet browsing and resilience as well as outlook on life, we conducted a bootstrapping-based moderated mediation analysis with 5000 resamples (PROCESS model 8) [62].

Results

As evident in Table 2, when participants were allowed to browse the internet, they reported a more positive outlook on life if they read Grimms' fairy tales (read fairy tale: mean 5.46, SD 0.151 vs not read fairy tale: mean 0.3.01, SD 0.150, SE 0.213, 95% CI –2.860 to –2.024; *P*<.001). When participants did not browse the internet, this effect disappeared. A detailed comparison of the cell means is presented in Tables 2-4. Furthermore, the results showed that, when participants browsed the internet, they indicated significantly greater resilience when they read a Grimms' fairy tale (mean 4.62, SE 0.179, 95% CI 4.271-4.976) than when they did not (mean 2.59, SE 0.179, 95% CI 2.243-2.945). When participants did not browse the internet, this effect disappeared (read fairy tale: mean 3.37, SE 0.179, 95% CI 3.014-3.719 vs not read fairy tale: mean 3.26, SE 0.180, 95% CI 2.908-3.617).

Table 2. Effects of internet browsing and reading a fairy tale on outlook on life: cell means estimates for the dependent variable of outlook on life.

Condition	Mean	SE	95% CI
Internet browsing: no			
Read fairy tale: no	4.11	0.151	3.813-4.408
Read fairy tale: yes	4.177	0.151	3.881-4.473
Internet browsing: yes			
Read fairy tale: no	3.014	0.150	2.720-3.309
Read fairy tale: yes	5.456	0.151	5.160-5.753

XSL•FO
RenderX

Table 3. Detailed cell means comparisons for impact of internet browsing and reading a fairy tale on outlook on life.

Condition	Mean difference[a]	SE	*P* value	95% CI
Internet browsing: no				
Read fairy tale: no-yes	–0.067	0.214	.75	–0.487 to 0.353
Internet browsing: yes				
Read fairy tale: no-yes	–2.442[b]	0.213	<.001	–2.860 to –2.024
Read fairy tale: no				
Internet browsing: no-yes	1.096[b]	0.213	<.001	0.677 to 1.515
Read fairy tale: yes				
Internet browsing: no-yes	–1.279[b]	0.213	<.001	–1.698 to –0.860

[a]Based on estimated marginal means.

[b]The mean difference is significant at the .05 level.

Table 4. Univariate test results for influence of internet browsing and reading a fairy tale on outlook on life.

Internet browsing	Sum of squares (*df*)	Mean square	*F* (*df*)	*P* value
Internet browsing: no			0.098 (1, 408)	.75
Contrast	0.229 (1)	0.229		
Error	954.807 (408)	2.340		
Internet browsing: yes			131.855 (1, 408)	<.001
Contrast	308.658 (1)	308.568		
Error	954.807 (408)	2.340		
Read fairy tale: no			26.426 (1, 408)	<.001
Contrast	61.843 (1)	61.842		
Error	954.807 (408)	2.340		
Read fairy tale: yes			36.006 (1, 408)	<.001
Contrast	84.262 (1)	84.262		
Error	954.807 (308)	2.340		

To examine the potential moderating effect of Grimms' fairy tales on the relationship between internet browsing and resilience as well as young people's outlook on life, we conducted a bootstrapping-based moderated mediation analysis with 5000 resamples (PROCESS model 8) [62]. The analysis was conducted with outlook on life as the dependent variable, internet browsing as the independent variable, resilience as the mediator variable, and whether the participant read a fairy tale as the moderator variable. The sample size was 412 participants. Detailed results are presented in Tables 5-7.

XSL•FO
RenderX

Table 5. Moderated mediation regression results.

Predictor	B	SE	*t* (*df*)	*P* value	95% CI
Dependent variable: resilience[a]					
Constant	5.75	0.89	6.39 (408)	< .001	3.98 to 7.51
Internet browsing	–2.59	0.57	–4.57 (408)	< .001	–3.70 to 1.48
Fairy tale	–1.82	0.57	–3.21 (408)	.002	–2.94 to –0.70
Internet browsing × fairy tale	1.93	0.36	5.37 (408)	<.001	1.22 to 2.63
Dependent variable: outlook on life[b]					
Constant	4.98	0.68	7.37 (407)	< .001	3.65 to 6.31
Internet browsing	–2.33	0.42	–5.59 (407)	<.001	–2.15 to –1.51
Resilience	0.44	0.04	12.39 (407)	< .001	0.37 to 0.51
Fairy tale	–1.51	0.41	–3.65 (407)	.003	–2.32 to –0.69
Internet browsing × fairy tale	1.53	0.27	5.74 (407)	<.001	1.00 to 2.05

[a]R=0.38, R^2=0.14, mean squared error=3.32, $F_{3,408}$=22.36, *P*<.001.

[b]R=0.67, R^2=0.45, mean squared error=1.70, $F_{4,407}$=83.89, *P*<.001.

Table 6. Conditional effects of internet browsing on outlook on life at levels of fairy tale presence.

Fairy tale presence	Effect	SE	*t* (*df*)	*P* value	95% CI
Read fairy tale: no	–0.80	0.18	–4.37 (407)	<.001	–1.16 to –0.44
Read fairy tale: yes	0.73	0.19	3.87 (407)	<.001	0.36 to 1.09

Table 7. Conditional indirect effects of internet browsing on outlook on life via resilience.

Fairy tale presence[a]	Effect	BootSE	95% CI
Read fairy tale: no	–0.29	0.11	–0.52 to –0.08
Read fairy tale: yes	0.55	0.13	0.32 to 0.81

[a]Index of moderated mediation: index=0.85, BootSE=0.1724, 95% CI 0.52 to 1.20.

The regression results in Table 5 show that internet browsing and reading fairy tales enhanced resilience (B=1.93, SE 0.36; t_{408}= 5.37, *P*<.001, 95% bootstrap CI 1.22-2.63), which in turn strengthened positive outlook on life (B=0.44, SE 0.04; t_{407}=12.39, *P*<.001, 95% CI 0.37-0.51). Furthermore, the results in Table 5 show that the effect of internet browsing on outlook on life is mediated by resilience (effect 0.85, SE 0.17, 95% CI 0.52-1.20). Together, these results highlight the potentially important effect of internet browsing and reading of fairy tales on outlook on life via resilience.

Discussion

Principal Findings

This study shows that classical Grimms' fairy tales, such as Hansel and Gretel and Little Red Riding Hood, serve as profound tools for fostering resilience and nurturing a positive outlook on life among young people. Critically, these tales, although often dark and cautionary, embed timeless lessons through their narratives, characters, and symbolic struggles, which helped mitigate the negative effects of internet browsing on young people's resilience and outlook on life.

For instance, Grimms' tales often feature protagonists confronting dire circumstances. By witnessing characters navigate fear and emerge stronger, readers internalize the idea that hardship is surmountable and perseverance and adaptability can lead to triumph. Moreover, Grimms' fairy tales often follow a "darkness-to-light" arc, where suffering precedes redemption and crises are temporary. Thus, they can instill hope, suggesting that effort can lead to positive outcomes, a mindset critical for maintaining optimism. A happy ending, even if hard-won, reinforces optimism about overcoming life's trials and encourages readers to approach obstacles with ingenuity rather than despair. Such narratives reinforce the value of resourcefulness, a key component of resilience.

Comparison With Prior Work

Previous research notes the impact of internet browsing and the role of frequent news feed browsing as well as social media on anxiety and pessimism on life [1,2]. In addition, prior work has indicated the addictive nature of internet browsing and the potential dangers of personalized content and consumption of negative information on the internet [7-11]. In addition, previous research has shown the effects of various online content on young people's feelings of hopelessness [12,13,40]. Moreover,

XSL•FO
RenderX

prior work indicates the role of the internet in providing a constant stream of news about crises, disasters, and violence [41-43], which may have a significant impact on young people's lives [42-44]. Various social media platforms also showcase idealized and curated versions of others' lives, which can lead to reduced contentment and satisfaction with one's own life [47,49,50]. However, critical questions remain about the effect of random internet browsing on resilience and outlook on life of young people. To the best of our knowledge, this is the first study that examines the role of free internet browsing on postgraduate students' resilience and outlook on life. In addition to this, previous research has noted the role of classical fairy tales in addressing fundamental human experiences of challenge, danger, and survival [29,30], which may help young people with processing their own fears and worries in daily life [31,32,36]. Critical questions remain regarding the potential mitigating role of reading classical fairy tales in helping reduce the negative effects of internet browsing on resilience and outlook on life of young people. Thus, this study offers critical new insights about the impact of random internet browsing on the resilience of young people as well as their outlook (positive vs negative) on life. Furthermore, this study demonstrates the extent to which reading classical fairy tales may help mitigate the negative impact of internet browsing.

Limitations and Future Research

This study has a number of limitations that offer promising avenues for future research. Although we explored the role of Brothers Grimm fairy tales in mitigating the negative impact of internet browsing, additional research that examines the role of other classical or modern fairy tales is richly deserving. Moreover, we examined the role of internet browsing and reading classical fairy tales in affecting the resilience and outlook on life of postgraduate students. We invite future research to explore the effects of browsing the internet and reading classical fairy tales for study participants other than postgraduate students (eg, younger or older age groups). In addition, we note that, due to the cross-sectional measurement of both outlook and resilience after the manipulation, the causal mediating role of resilience cannot be confidently established. The current test of resilience as a mediator is exploratory in nature, and future research examining the effects of internet browsing and reading classical fairy tales on outlook and resilience over time is richly deserving. Moreover, previous research indicates that, in addition to resilience, other mediators may exist that help explain the effects of internet browsing on someone's outlook on life [62-67]. For instance, people may use the internet to gain access to inspiring information, helpful and enabling tools, counselling, or humorous content that may lift their mood, as well as connect with family and friends for enhanced feelings of group membership and communion and contributing positively to society [68-79]. Therefore, we invite additional research to examine how young people may use the internet as a tool and force for good and enhanced resilience as well as positive outlook on life. Furthermore, prior work notes the potential for digital solutions to positively affect mental health and well-being [80] but also indicates privacy concerns people may have when using various internet-based tools [81]. Future research shedding further light on these critical issues is richly deserving. In addition, prior work indicates the need for critical reading of some of the fairy tales and their various, more modern versions [82-85]. For instance, Snow White displays surprising strength and bravery when focusing on inner resilience and moral courage rather than mere physical action. Snow White survived years of psychological abuse and servitude under the jealous queen, yet retains her inherent kindness, optimism, and hope. This emotional resilience in the face of cruelty is a profound strength [86]. Moreover, Snow White chose survival in terror. When sentenced to death in the woods, she does not freeze nor despair. Snow White actively flees through a terrifying, dark, unknown forest, which is a genuinely brave act for anyone, especially a young girl (depicted as very young in the original tale). In addition, Snow White basically rebuilt her life from scratch, demonstrating a courage to start anew. Critically, maintaining compassion and hope in a world that has shown her cruelty requires significant inner fortitude and a special kind of bravery. It is the core of her identity that Snow White refuses to let the evil queen destroy. Through this angle, readers may see that she is a strong and resilient character in a deeply psychological way and that there can be a path to finding hope even in very dark times [87]. Sometimes the quietest forms of strength, such as resilience, hope, and kindness in the face of cruelty, are the most profound, yet the easiest to overlook [88].

By balancing caution with courage and individualism with community, these stories cultivate a resilient spirit and a belief in positive outcomes, proving that even the darkest forests can be navigated with perseverance and heart. Grimms' fairy tales are like ancient mirrors. They reflect both people's deepest fears and their capacity to outgrow them. What is fascinating is that resilience and optimism often bloom not despite the darkness in these tales but because of it. They do not sugarcoat life's terrors. Instead, they whisper "Yes, the world has witches and wolves...but also breadcrumb trails, cleverness, and kind woodsmen." That balance feels deeply human and endlessly relevant.

Summary

This study shows that reading classical Grimms' fairy tales such as Hansel and Gretel and Little Red Riding Hood can help mitigate the negative effect of internet browsing on young people's resilience and lead to a more positive versus negative outlook in their lives. Classical Grimms' fairy tales are more than escapism; they can be blueprints for resilience and optimism. Through adversity, resourcefulness, and hope, they equip readers with mental frameworks to confront life's challenges.

Many fairy tales end with hope, villains are defeated, families reunite, and protagonists emerge stronger. This narrative arc fosters optimism and a sense of agency, countering the helplessness or cynicism that can arise from exposure to negative online content (eg, doomscrolling, cyberbullying). Reading fairy tales also requires sustained attention and imagination, engaging the brain in ways that differ from the rapid, fragmented consumption of digital media. This deeper engagement may help strengthen empathy (by inhabiting characters' perspectives), enhance critical thinking (interpreting themes and morals), and

XSL•FO
RenderX

provide a mental slow space to reflect, contrasting with the dopamine-driven scroll cycles of the internet.

Conclusions

Reading classical fairy tales like Hansel and Gretel and Little Red Riding Hood can indeed serve as a counterbalance to some of the negative effects of internet browsing on young people's resilience and outlook. Grimms' fairy tales offer timeless lessons in resilience, morality, and hope, qualities that can mitigate the passive consumption, instant gratification, and emotional detachment sometimes exacerbated by internet browsing. Classic Grimms' fairy tales can be ancient little engines of resilience: They spark connection, reflection, and hope.

Acknowledgments

The authors thank all study participants for their time and efforts to take part in the research. The authors used DeepSeek (DeepThink R1) for parts of the writing's grammar and writing checks.

Data Availability

Study data are available from the authors upon reasonable request.

Authors' Contributions

CH was responsible for the methodology; data collection; formal analysis; investigation; and writing, reviewing, and editing the manuscript. TF was responsible for the methodology, formal analysis, and writing and editing the manuscript. BD was responsible for the methodology, formal analysis, investigation, and writing and reviewing the manuscript. AA was responsible for writing, reviewing, and editing the manuscript. ABE was responsible for conceptualization; methodology; formal analysis; investigation; writing, reviewing, and editing the manuscript; and supervision.

Conflicts of Interest

None declared.

Multimedia Appendix 1
CONSORT-SPI 2018 Checklist: Reporting Randomized Trials of Social and Psychological Interventions.
[DOCX File , 38 KB - formative_v9i1e76770_app1.docx]

References

1. Bauman S, Rivers I. Information Overload and Zoom Fatigue. In: Mental Health in the Digital Age. Cham, Switzerland: Springer International Publishing; 2023:109-126.
2. Hawi NS, Samaha M. The relations among social media addiction, self-esteem, and life satisfaction in university students. Social Science Computer Review 2016 Aug 10;35(5):576-586. [doi: 10.1177/0894439316660340]
3. Yu DJ, Wing YK, Li TMH, Chan NY. The impact of social media use on sleep and mental health in youth: a scoping review. Curr Psychiatry Rep 2024 Mar 08;26(3):104-119 [FREE Full text] [doi: 10.1007/s11920-024-01481-9] [Medline: 38329569]
4. Alonzo R, Hussain J, Stranges S, Anderson K. Interplay between social media use, sleep quality, and mental health in youth: a systematic review. Sleep Med Rev 2021 Apr;56:101414 [FREE Full text] [doi: 10.1016/j.smrv.2020.101414] [Medline: 33385767]
5. Alhabash S, Smischney TM, Suneja A, Nimmagadda A, White LR. So similar, yet so different: how motivations to use Facebook, Instagram, Twitter, and TikTok predict problematic use and use continuance intentions. Sage Open 2024 May 22;14(2):1. [doi: 10.1177/21582440241255426]
6. Jouhki H, Savolainen I, Sirola A, Oksanen A. Escapism and excessive online behaviors: a three-wave longitudinal study in Finland during the COVID-19 pandemic. Int J Environ Res Public Health 2022 Sep 30;19(19):12491 [FREE Full text] [doi: 10.3390/ijerph191912491] [Medline: 36231799]
7. Przybylski AK, Weinstein N. Digital screen time limits and young children's psychological well-being: evidence from a population-based study. Child Dev 2019 Jan 13;90(1):e56-e65. [doi: 10.1111/cdev.13007] [Medline: 29235663]
8. Twenge JM, Campbell WK. Associations between screen time and lower psychological well-being among children and adolescents: evidence from a population-based study. Prev Med Rep 2018 Dec;12:271-283 [FREE Full text] [doi: 10.1016/j.pmedr.2018.10.003] [Medline: 30406005]
9. Odgers CL, Jensen MR. Annual Research Review: adolescent mental health in the digital age: facts, fears, and future directions. J Child Psychol Psychiatry 2020 Mar 17;61(3):336-348 [FREE Full text] [doi: 10.1111/jcpp.13190] [Medline: 31951670]

XSL•FO
RenderX

10. Valkenburg PM, Meier A, Beyens I. Social media use and its impact on adolescent mental health: an umbrella review of the evidence. Curr Opin Psychol 2022 Apr;44:58-68 [FREE Full text] [doi: 10.1016/j.copsyc.2021.08.017] [Medline: 34563980]
11. Schweiger S, Cress U. How confidence in prior attitudes, social tag popularity, and source credibility shape confirmation bias toward Antidepressants and psychotherapy in a representative German sample: randomized controlled web-based study. J Med Internet Res 2019 Apr 23;21(4):e11081 [FREE Full text] [doi: 10.2196/11081] [Medline: 31012865]
12. Houghton S, Lawrence D, Hunter SC, Rosenberg M, Zadow C, Wood L, et al. Reciprocal relationships between trajectories of depressive symptoms and screen media use during adolescence. J Youth Adolesc 2018 Nov 25;47(11):2453-2467 [FREE Full text] [doi: 10.1007/s10964-018-0901-y] [Medline: 30046970]
13. Wollebæk D, Karlsen R, Steen-Johnsen K, Enjolras B. Anger, fear, and echo chambers: the emotional basis for online behavior. Social Media + Society 2019 Apr 09;5(2):1. [doi: 10.1177/2056305119829859]
14. Luzsa R, Mayr S. False consensus in the echo chamber: exposure to favorably biased social media news feeds leads to increased perception of public support for own opinions. Cyberpsychology: Journal of Psychosocial Research on Cyberspace 2021 Feb 16;15(1):3. [doi: 10.5817/CP2021-1-3]
15. Du J, Kerkhof P, van Koningsbruggen GM. Predictors of social media self-control failure: immediate gratifications, habitual checking, ubiquity, and notifications. Cyberpsychol Behav Soc Netw 2019 Jul;22(7):477-485. [doi: 10.1089/cyber.2018.0730] [Medline: 31295024]
16. Livingstone S, Helsper E. Gradations in digital inclusion: children, young people and the digital divide. New Media & Society 2007 Aug 01;9(4):671-696. [doi: 10.1177/1461444807080335]
17. Przybylski AK, Weinstein N. A large-scale test of the Goldilocks hypothesis. Psychol Sci 2017 Feb;28(2):204-215 [FREE Full text] [doi: 10.1177/0956797616678438] [Medline: 28085574]
18. Orben A, Przybylski AK. The association between adolescent well-being and digital technology use. Nat Hum Behav 2019 Feb 14;3(2):173-182. [doi: 10.1038/s41562-018-0506-1] [Medline: 30944443]
19. Beyens I, Pouwels JL, van Driel II, Keijsers L, Valkenburg PM. The effect of social media on well-being differs from adolescent to adolescent. Sci Rep 2020 Jul 01;10(1):10763 [FREE Full text] [doi: 10.1038/s41598-020-67727-7] [Medline: 32612108]
20. Shensa A, Escobar-Viera CG, Sidani JE, Bowman ND, Marshal MP, Primack BA. Problematic social media use and depressive symptoms among U.S. young adults: a nationally-representative study. Soc Sci Med 2017 Jun;182:150-157 [FREE Full text] [doi: 10.1016/j.socscimed.2017.03.061] [Medline: 28446367]
21. Verduyn P, Ybarra O, Résibois M, Jonides J, Kross E. Do social network sites enhance or undermine subjective well‐being? A critical review. Social Issues Policy Review 2017 Jan 13;11(1):274-302. [doi: 10.1111/sipr.12033]
22. Firth J, Torous J, Stubbs B, Firth JA, Steiner GZ, Smith L, et al. The "online brain": how the internet may be changing our cognition. World Psychiatry 2019 Jun 06;18(2):119-129 [FREE Full text] [doi: 10.1002/wps.20617] [Medline: 31059635]
23. Marciano L, Ostroumova M, Schulz PJ, Camerini A. Digital media use and adolescents' mental health during the Covid-19 pandemic: a systematic review and meta-analysis. Front Public Health 2021 Feb 1;9:793868 [FREE Full text] [doi: 10.3389/fpubh.2021.793868] [Medline: 35186872]
24. Heffer T, Good M, Daly O, MacDonell E, Willoughby T. The longitudinal association between social-media use and depressive symptoms among adolescents and young adults: an empirical reply to Twenge et al. (2018). Clinical Psychological Science 2019 Jan 29;7(3):462-470. [doi: 10.1177/2167702618812727]
25. Hunt MG, Marx R, Lipson C, Young J. No more FOMO: limiting social media decreases loneliness and depression. Journal of Social and Clinical Psychology 2018 Dec;37(10):751-768. [doi: 10.1521/jscp.2018.37.10.751]
26. Granic I, Morita H, Scholten H. Beyond screen time: identity development in the digital age. Psychological Inquiry 2020 Nov 05;31(3):195-223. [doi: 10.1080/1047840x.2020.1820214]
27. Uhls YT, Michikyan M, Morris J, Garcia D, Small GW, Zgourou E, et al. Five days at outdoor education camp without screens improves preteen skills with nonverbal emotion cues. Computers in Human Behavior 2014 Oct;39:387-392. [doi: 10.1016/j.chb.2014.05.036]
28. Bettelheim B. The Uses of Enchantment: The Meaning and Importance of Fairy Tales. London, England: Penguin Books, Limited; 1976.
29. O. Althobaiti MA. The evolution of European fairy tales: a comparative analysis of the Grimm Brothers and Hans Christian Andersen. ESJ 2023 Aug 31;19(23):43. [doi: 10.19044/esj.2023.v19n23p43]
30. Nikolajeva M. Reading for Learning: Cognitive Approaches to Children's Literature. Cambridge, MA: John Benjamins Publishing Company; 2014.
31. Zipes J. The Irresistible Fairy Tale: The Cultural and Social History of a Genre. Princeton, NJ: Princeton University Press; 2012.
32. Hohr H. Dynamic aspects of fairy tales: social and emotional competence through fairy tales. Scandinavian Journal of Educational Research 2000 Mar;44(1):89-103. [doi: 10.1080/713696665]
33. Tatar M. Enchanted Hunters: The Power of Stories in Childhood. New York, NY: W. W. Norton & Company; 2009.
34. Kidd DC, Castano E. Reading literary fiction improves theory of mind. Science 2013 Oct 18;342(6156):377-380. [doi: 10.1126/science.1239918] [Medline: 24091705]

XSL•FO
RenderX

35. Brown NW. The therapeutic use of fairy tales with adults in group therapy. Journal of Creativity in Mental Health 2008 Mar 04;2(4):87-96. [doi: 10.1300/j456v02n04_08]
36. Ramamurthy C, Zuo P, Armstrong G, Andriessen K. The impact of storytelling on building resilience in children: a systematic review. J Psychiatr Ment Health Nurs 2024 Aug 12;31(4):525-542. [doi: 10.1111/jpm.13008] [Medline: 38087837]
37. Heath MA, Sheen D, Leavy D, Young E, Money K. Bibliotherapy. School Psychology International 2005 Dec 01;26(5):563-580. [doi: 10.1177/0143034305060792]
38. Haven K. Story Smart: Using the Science of Story to Persuade, Influence, Inspire, and Teach. Exeter, England: Libraries Unlimited; 2014.
39. Haidt J. The Anxious Generation: How the Great Rewiring of Childhood is Causing an Epidemic of Mental Illness. London, England: Penguin Press; 2024.
40. Hörter K. “It Makes Me Sad When So Much of the Forest Burns Down!” Childhood and youth in the context of the climate crisis. The Psychoanalytic Study of the Child 2024 Jan 23;77(1):389-401. [doi: 10.1080/00797308.2023.2287383]
41. Schor M, Clayton S, Cardoso MCV, Fernandes M, Frota MA, Surkan PJ. ‘I am really scared for the future’: climate distress among Brazilian youth. Environmental Education Research 2024 Sep 28:1-22. [doi: 10.1080/13504622.2024.2409946]
42. Hughes B, Jones K. Young people's experiences of death anxiety and responses to the Covid-19 pandemic. Omega (Westport) 2024 Dec 17;90(2):656-673 [FREE Full text] [doi: 10.1177/00302228221109052] [Medline: 35715963]
43. Caporino NE, Exley S, Latzman RD. Youth anxiety about political news. Child Psychiatry Hum Dev 2020 Oct 25;51(5):683-698. [doi: 10.1007/s10578-020-00972-z] [Medline: 32100153]
44. Valkenburg PM, Peter J. Online communication among adolescents: an integrated model of its attraction, opportunities, and risks. J Adolesc Health 2011 Feb;48(2):121-127. [doi: 10.1016/j.jadohealth.2010.08.020] [Medline: 21257109]
45. Craig W, Boniel-Nissim M, King N, Walsh SD, Boer M, Donnelly PD, et al. Social media use and cyber-bullying: a cross-national analysis of young people in 42 countries. J Adolesc Health 2020 Jun;66(6S):S100-S108 [FREE Full text] [doi: 10.1016/j.jadohealth.2020.03.006] [Medline: 32446603]
46. Vogel EA, Rose JP, Okdie BM, Eckles K, Franz B. Who compares and despairs? The effect of social comparison orientation on social media use and its outcomes. Personality and Individual Differences 2015 Nov;86:249-256. [doi: 10.1016/j.paid.2015.06.026]
47. Mann RB, Blumberg F. Adolescents and social media: the effects of frequency of use, self-presentation, social comparison, and self esteem on possible self imagery. Acta Psychol (Amst) 2022 Aug;228:103629 [FREE Full text] [doi: 10.1016/j.actpsy.2022.103629] [Medline: 35661975]
48. Servidio R, Soraci P, Griffiths MD, Boca S, Demetrovics Z. Fear of missing out and problematic social media use: a serial mediation model of social comparison and self-esteem. Addict Behav Rep 2024 Jun;19:100536 [FREE Full text] [doi: 10.1016/j.abrep.2024.100536] [Medline: 38495391]
49. Gori A, Topino E, Griffiths MD. The associations between attachment, self-esteem, fear of missing out, daily time expenditure, and problematic social media use: a path analysis model. Addict Behav 2023 Jun;141:107633 [FREE Full text] [doi: 10.1016/j.addbeh.2023.107633] [Medline: 36753932]
50. Buzzai C, Filippello P, Costa S, Amato V, Sorrenti L. Problematic internet use and academic achievement: a focus on interpersonal behaviours and academic engagement. Soc Psychol Educ 2021 Jan 04;24(1):95-118. [doi: 10.1007/s11218-020-09601-y]
51. Zhang Y, Qin X, Ren P. Adolescents' academic engagement mediates the association between Internet addiction and academic achievement: the moderating effect of classroom achievement norm. Computers in Human Behavior 2018 Dec;89:299-307. [doi: 10.1016/j.chb.2018.08.018]
52. Waytz A, Gray K. Does online technology make us more or less sociable? A preliminary review and call for research. Perspect Psychol Sci 2018 Jul 14;13(4):473-491. [doi: 10.1177/1745691617746509] [Medline: 29758166]
53. Downey DB, Gibbs BG. Kids these days: are face-to-face social skills among American children declining? American Journal of Sociology 2020 Jan 01;125(4):1030-1083. [doi: 10.1086/707985]
54. Scheier MF, Carver CS. Optimism, coping, and health: assessment and implications of generalized outcome expectancies. Health Psychology 1985;4(3):219-247. [doi: 10.1037/0278-6133.4.3.219]
55. Scheier MF, Carver CS, Bridges MW. Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem): a reevaluation of the Life Orientation Test. Journal of Personality and Social Psychology 1994;67(6):1063-1078. [doi: 10.1037/0022-3514.67.6.1063]
56. Connor KM, Davidson JR. Development of a new resilience scale: the Connor-Davidson Resilience Scale (CD-RISC). Depress Anxiety 2003 Sep 02;18(2):76-82. [doi: 10.1002/da.10113] [Medline: 12964174]
57. Campbell-Sills L, Stein MB. Psychometric analysis and refinement of the Connor-davidson Resilience Scale (CD-RISC): validation of a 10-item measure of resilience. J Trauma Stress 2007 Dec;20(6):1019-1028. [doi: 10.1002/jts.20271] [Medline: 18157881]
58. Faul F, Erdfelder E, Lang A, Buchner A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior Research Methods 2007 May;39(2):175-191. [doi: 10.3758/BF03193146]
59. Anderson JC, Gerbing DW. Structural equation modeling in practice: a review and recommended two-step approach. Psychological Bulletin 1988;103(3):411-423. [doi: 10.1037//0033-2909.103.3.411]

XSL•FO
RenderX

60. Bagozzi RP, Yi Y. On the evaluation of structural equation models. JAMS 1988 Mar;16(1):74-94. [doi: 10.1007/bf02723327]
61. Fornell C, Larcker DF. Evaluating structural equation models with unobservable variables and measurement error. Journal of Marketing Research 1981 Feb;18(1):39. [doi: 10.2307/3151312]
62. Hayes AF. Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach. New York, NY: The Guilford Press; 2017.
63. Anto A, Basu A, Selim R, Foscht T, Eisingerich AB. Open-world games' affordance of cognitive escapism, relaxation, and mental well-being among postgraduate students: mixed methods study. J Med Internet Res 2024 Dec 17;26:e63760 [FREE Full text] [doi: 10.2196/63760] [Medline: 39689301]
64. Hartnup B, Dong L, Eisingerich AB. How an environment of stress and social risk shapes student engagement with social media as potential digital learning platforms: qualitative study. JMIR Med Educ 2018 Jul 13;4(2):e10069 [FREE Full text] [doi: 10.2196/10069] [Medline: 30006324]
65. Anto A, Asif RO, Basu A, Kanapathipillai D, Salam H, Selim R, et al. Exploring the impact of social media on anxiety among university students in the United Kingdom: qualitative study. JMIR Form Res 2023 Jun 16;7:e43037 [FREE Full text] [doi: 10.2196/43037] [Medline: 37327030]
66. Anto A, Basu A, Selim R, Zaman J, Kanapathipillai D, Salam H, et al. Exploring the facilitators and barriers to addressing social media's impact on anxiety within primary care: a qualitative study. BJGP Open 2024 Jan 25;8(2):BJGPO.2023.0190. [doi: 10.3399/bjgpo.2023.0190]
67. Lin Y, Tudor-Sfetea C, Siddiqui S, Sherwani Y, Ahmed M, Eisingerich AB. Effective behavioral changes through a digital mHealth app: exploring the impact of hedonic well-being, psychological empowerment and inspiration. JMIR Mhealth Uhealth 2018 Jun 15;6(6):e10024 [FREE Full text] [doi: 10.2196/10024] [Medline: 29907557]
68. El-Hilly AA, Iqbal SS, Ahmed M, Sherwani Y, Muntasir M, Siddiqui S, et al. Game on? Smoking cessation through the gamification of mHealth: a longitudinal qualitative study. JMIR Serious Games 2016 Oct 24;4(2):e18 [FREE Full text] [doi: 10.2196/games.5678] [Medline: 27777216]
69. Tudor-Sfetea C, Rabee R, Najim M, Amin N, Chadha M, Jain M, et al. Evaluation of two mobile health apps in the context of smoking cessation: qualitative study of cognitive behavioral therapy (CBT) versus non-CBT-based digital solutions. JMIR Mhealth Uhealth 2018 Apr 18;6(4):e98 [FREE Full text] [doi: 10.2196/mhealth.9405] [Medline: 29669708]
70. Eisingerich AB, Kretschmer T. In e-commerce, more is more. Harvard Business Review. 2008 Mar. URL: https://hbr.org/2008/03/in-e-commerce-more-is-more [accessed 2025-07-21]
71. Eisingerich AB, MacInnis DJ, Park CW. A theoretical framework exploring three foundational benefits of brand attachment. In: King C, Murillo E, editors. A Research Agenda for Brand Management in a New Era of Consumerism. Camberley, England: Edward Elgar Publishing; Sep 2023:161-175.
72. Eisingerich AB, Marchand A, Fritze MP, Dong L. Hook vs. hope: how to enhance customer engagement through gamification. International Journal of Research in Marketing 2019 Jun;36(2):200-215. [doi: 10.1016/j.ijresmar.2019.02.003]
73. Kadam PD, Chuan HH. Erratum to: Rectocutaneous fistula with transmigration of the suture: a rare delayed complication of vault fixation with the sacrospinous ligament. Int Urogynecol J 2016 Mar;27(3):505 [FREE Full text] [doi: 10.1007/s00192-016-2952-5] [Medline: 26811110]
74. Bell SJ, Auh S, Eisingerich AB. Unraveling the customer education paradox. Journal of Service Research 2017 Feb 15;20(3):306-321. [doi: 10.1177/1094670517691847]
75. Foscht T, Lin Y, Eisingerich AB. Blinds up or down? EJM 2018 Feb 02;52(3/4):476-498. [doi: 10.1108/EJM-10-2016-0576]
76. Lin Y, MacInnis DJ, Eisingerich AB. Strong anxiety boosts new product adoption when hope is also strong. Journal of Marketing 2020 Jul 10;84(5):60-78. [doi: 10.1177/0022242920934495]
77. Eisingerich AB, MacInnis DJ, Whan Park C. Do CSR efforts that focus on helping the environment influence brand purchase more than other forms of CSR? Journal of Business Research 2023 Nov;168:114240. [doi: 10.1016/j.jbusres.2023.114240]
78. Yang C, Hunhoff P, Lee Y, Abrell J. Social media activities with different content characteristics and adolescent mental health: cross-sectional survey study. J Med Internet Res 2025 Apr 28;27:e73098 [FREE Full text] [doi: 10.2196/73098] [Medline: 40294901]
79. Wu X, Liew K, Dorahy M. Trust, anxious attachment, and conversational AI adoption intentions in digital counseling: a preliminary cross-sectional questionnaire study. JMIR AI 2025 Apr 22;4:e68960 [FREE Full text] [doi: 10.2196/68960] [Medline: 40262137]
80. Eisenstadt M, Liverpool S, Infanti E, Ciuvat RM, Carlsson C. Mobile apps that promote emotion regulation, positive mental health, and well-being in the general population: systematic review and meta-analysis. JMIR Ment Health 2021 Nov 08;8(11):e31170 [FREE Full text] [doi: 10.2196/31170] [Medline: 34747713]
81. Herriger C, Merlo O, Eisingerich AB, Arigayota AR. Context-contingent privacy concerns and exploration of the privacy paradox in the age of AI, augmented reality, big data, and the internet of things: systematic review. J Med Internet Res 2025 May 14;27:e71951 [FREE Full text] [doi: 10.2196/71951] [Medline: 40367513]
82. Bacchilega C. Postmodern Fairy Tales: Gender and Narrative Strategies. Philadelphia, PA: University of Pennsylvania Press; 1997.
83. Zipes J. The Oxford Companion to Fairy Tales: The Western Fairy Tale Tradition from Medieval to Modern. Oxford, England: Oxford University Press; 2000.

XSL•FO
RenderX

84. Tatar M. Off With Their Heads!: Fairy Tales and the Culture of Childhood. Princeton, NJ: Princeton University Press; 1992.
85. Parsons LT. Ella evolving: Cinderella stories and the construction of gender-appropriate behavior. Children's Literature in Education 2004 Jun;35(2):135-154. [doi: 10.1023/b:clid.0000030223.88357.e8]
86. Frankl VE. Man's Search for Meaning: An Introduction to Logotherapy. Boston, MA: Beacon Press; 1984.
87. Frankl VE. The Doctor and the Soul: From Psychotherapy to Logotherapy, 3rd edition. New York, NY: Vintage Books; 1986.
88. Frankl VE. The Will to Meaning: Foundations and Applications of Logotherapy. New York, NY: Penguin/Plume; 1988.

Abbreviations

AVE: average variance extracted
CONSORT-SPI: Consolidated Standards of Reporting Trials statement for social and psychological interventions

Edited by A Mavragani; submitted 30.04.25; peer-reviewed by D Vuckovic, R Saar-Ashkenazy; comments to author 20.05.25; revised version received 04.06.25; accepted 15.07.25; published 06.08.25.

Please cite as:
Hou C, Foscht T, Duffek B, Arigayota A, Eisingerich AB
Mitigating the Negative Effects of Internet Browsing on Young People's Resilience and Outlook on Life Through Classic Grimms' Fairy Tales: Exploratory Randomized Controlled Study
JMIR Form Res 2025;9:e76770
URL: https://formative.jmir.org/2025/1/e76770
doi:10.2196/76770
PMID:40768252

©Congcong Hou, Thomas Foscht, Barbara Duffek, Annisa Arigayota, Andreas Benedikt Eisingerich. Originally published in JMIR Formative Research (https://formative.jmir.org), 06.08.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

User Experiences Among Patients and Health Care Professionals Who Participated in a Randomized Controlled Trial of E-nergEYEze, a Vision-Specific eHealth Intervention to Reduce Fatigue in Adults With Visual Impairment: Mixed Methods Study

Manon HJ Veldman[1,2], MD, PhD; Hilde PA van der Aa[1,2], MD, PhD; Hans Knoop[3], Prof Dr, PhD; Christina Bode[4], MD, PhD; Ger HMB van Rens[1], MD, Prof Dr, PhD; Ruth MA van Nispen[1,2], Prof Dr, PhD

[1]Amsterdam UMC, location Vrije Universiteit Amsterdam, Ophthalmology, Amsterdam, The Netherlands

[2]Quality of Care, Mental Health, Amsterdam Public Health, Amsterdam, The Netherlands

[3]Amsterdam UMC, location University of Amsterdam, Departments of Medical Psychology, Amsterdam Public Health research institute, Amsterdam, The Netherlands

[4]Department of Psychology, Health and Technology, University of Twente, Enschede, The Netherlands

Corresponding Author:
Manon HJ Veldman, MD, PhD
Amsterdam UMC
location Vrije Universiteit Amsterdam
Ophthalmology
De Boelelaan 1117
Amsterdam, 1081 HV
The Netherlands
Phone: 31 204444444
Email: m.veldman@amsterdamumc.nl

Abstract

Background: Fatigue is a common symptom occurring in individuals with visual impairment (VI). Feeling fatigued has a strong impact on an individual's well-being, with profound consequences. Cognitive and emotional functioning, social roles, and participation are negatively affected in severely fatigued individuals with VI. Therefore, we developed E-nergEYEze, a blended vision-specific eHealth intervention based on cognitive behavioral therapy and self-management to reduce fatigue severity in adults with VI.

Objective: We aimed to report the experience of patients and professionals with E-nergEYEze. To complement cost-effectiveness outcomes, the user experiences from both perspectives were considered relevant for a better understanding of the intervention uptake.

Methods: E-nergEYEze was studied in a randomized controlled trial. User experiences of participants with VI and severe fatigue (51/98, 52%; median age 58.0, IQR 53.0-65.0 years; female participants: 32/51, 63%), who were randomized to the intervention group, and professionals (n=11), who provided blended support, were evaluated. The Dutch Mental Health Care Thermometer questionnaire and a therapist evaluation were used and analyzed using mixed methods. A focus group meeting with social workers (4/7, 57%), a computer trainer (1/7, 14%), and psychologists (2/7, 29%) was held for more in-depth information. The eHealth platform provided data on user engagement from both perspectives.

Results: E-nergEYEze was completed by 63% (32/51) of patients for more than 80% of the module steps. Overall, results on user engagement showed that a median 89% (IQR 45%-100%) of all assigned module steps were completed, with all modules being completed by at least 50% (37/51) of the patients. Completion of the intervention was related to the presence of digital proficiency; having the appropriate expectations; content that matches personal preferences and life context; and the absence of impeding personal circumstances, mental health issues, or other concurrent rehabilitation programs. The intervention was given a median grade of 7.0 out of 10.0 (IQR 6.0-8.0), and 87% (39/45) of the patients reported that they would recommend E-nergEYEze to others. However, improvements in the frequency and quality of guidance were considered highly relevant. Professionals reported that E-nergEYEze required patients' self-efficacy, motivation, and digital skills; therefore, preselection was seen as essential. Professionals' affinity with eHealth was considered important to provide appropriate remote support.

XSL•FO
RenderX

Conclusions: eHealth provides treatment opportunities for individuals with VI for which guidance is considered highly relevant. During participation in E-nergEYEze, patients were engaged, internalized personally relevant topics, and made use of the benefits of eHealth. More attention to the suitability of patients and training of professionals for providing remote support is considered essential. These user experiences underlined the potential of E-nergEYEze to reduce fatigue severity in adults with VI and provided valuable insights to learn from and optimize E-nergEYEze.

Trial Registration: International Clinical Trials Registry Platform (ICTRP) NL7764; https://tinyurl.com/32b3xt74

(JMIR Form Res 2025;9:e53080) doi:10.2196/53080

KEYWORDS

visual impairment; fatigue; eHealth; internet-based on cognitive behavioral therapy; iCBT; user experiences

Introduction

Background

Fatigue is known to be an everyday problem for most individuals with visual impairment (VI) [1-4]. Consequences of fatigue on emotional and cognitive functioning, social roles, and participation in society have been reported to be problematic [5,6]. To address this difficulty, an intervention called E-nergEYEze was developed tailored to the unique aspects of individuals with vision loss who experience severe fatigue and investigated in a randomized controlled trial (RCT) [7]. E-nergEYEze is a blended eHealth intervention based on cognitive behavioral therapy (iCBT) and self-management. Professionals from low vision services offer online support and motivate patients to follow the intervention as intended. The goal is to teach patients how to cope with fatigue by focusing on understanding vision-specific fatigue, which makes them perceive it as more controllable and to reduce its consequences on daily life [5,8].

Considering the increasing demand for support in health care, eHealth offers a solution to reach a broader and larger audience. It can be offered independently of time and place, may stimulate patient empowerment, and requires less professional effort [9,10]. However, the challenge is to ensure that an intervention is consistent with patients' needs and is easy for individuals with VI to use [11-13]. User experiences are of essential importance in all development and testing steps of iCBT to identify factors that drive participant behavior or affect outcomes and to gain a deeper understanding of participant experiences and needs [14]. Moss-Morris et al [15] conducted a pilot-RCT of a blended iCBT for individuals with multiple sclerosis (MS) in which qualitative feedback was reported on content and technical difficulties. Muijs et al [16] conducted a self-help iCBT intervention to prevent and reduce fatigue in adults with diabetes and reported the relevance of user experiences, stating that content and features need to be further personalized to preferences and needs. Finally, Babbage et al [17] studied a self-management app for fatigue in individuals with MS. Users reported challenges with the quantity and complexity of the text-based content. Results confirmed the relevance of qualitative feedback and the perceived value of fatigue-reducing iCBT tools.

The perspective of professionals is equally important to consider, especially because Dijksman et al [18] reported that 30% of psychologists did not believe that blended care had advantages in effectiveness and time savings compared to treatment as usual. However, overall, professionals were positive toward blended care, with significant differences in favor of active users versus nonusers. Because guidance skills, time, and input could influence outcomes [15,19], educational activities regarding eHealth, patient perceptions, and liability are considered necessary before using a new blended care intervention [18].

Study Objectives

E-nergEYEze was designed in collaboration with all stakeholders (ie, patients, professionals, and researchers in low vision, fatigue, and eHealth). To maintain involvement in future iterations, this newly developed iCBT intervention was first pilot-tested in a usability and feasibility study by considering fidelity. Results showed a promising trend to reduce severity and impact of fatigue. The study methods were feasible, and the fidelity of the intervention protocol was suitable [20]. As the intervention was only studied partially to refine E-nergEYEze, we planned to further explore stakeholder experiences by offering the full intervention while focusing on its fatigue-reducing accessibility, content, and guidance. To complement the future cost-effectiveness outcomes, our aim was to evaluate the user experiences of the E-nergEYEze intervention from the perspective of patients and professionals.

Methods

The E-nergEYEze Protocol

This study was conducted according to the E-nergEYEze protocol [7].

Study Design and Participants

Recruitment with information letters was performed by low vision service organizations in the Netherlands (Royal Dutch Visio and Bartiméus) and additionally through social media, patient associations, and 2 companies that provide low vision aids. Patients with VI and severe fatigue were included (N=98) in a single-blinded RCT and randomized to the intervention group (n=51, 52%) or control group (n=47, 48%). Patients were screened on the inclusion criteria having a mild VI (logarithm of the minimum angle of resolution visual acuity 0.30-0.40 or peripheral vision loss of <45°) or worse caused by an ophthalmic disorder; experiencing severe fatigue (a score on the Checklist Individual Strength questionnaire's fatigue severity subscale of ≥35) [21]; understanding the Dutch language; having access to the internet; and being aged ≥18 years. Exclusion criteria were having severe cognitive impairment, as assessed with a 6-item

XSL•FO
RenderX

screener (the validated 6-item Mini-Mental State Examination, with a cutoff score of >3) [22] and receiving treatment or having received treatment in the past 12 months from a medical specialist for a comorbid disease that could explain the presence of fatigue (eg, MS, cancer, or psychiatric disorder). A process evaluation was performed in patients who received the intervention (n=51, 52%) and professionals (n=11) of low vision service organizations who provided support. Both social workers (n=7) and computer trainers (n=4) kept time records, and the eHealth platform (e-platform) provided data on user engagement. A mixed methods design was used.

Intervention

Patients were invited for a face-to-face session with a social worker and a computer trainer. Access was given to the e-platform Minddistrict [23] to follow an introduction module and 8 thematic modules with information and assignments on how to cope with fatigue based on cognitive behavioral therapy and self-management. The module themes and content are provided in Textbox 1.

Textbox 1. Module themes and content of the eHealth platform.

Introduction: Understanding vision-related fatigue

It provides psychoeducation on vision impairment (VI) and its association with fatigue, including perpetuating factors of chronic fatigue.

Module 1: Dealing with VI

It focuses on the emotional impact of vision loss, processing feelings of grief, and coping strategies.

Module 2: Formulating helpful fatigue-related beliefs

It helps to understand the effect of beliefs on fatigue and fatigue maintaining behavior with exercises to formulate helping cognitions and shifting attention away from fatigue.

Module 3: Graded activity program

It focuses on balancing activities followed by a graded activity program.

Module 4: Communication and social support

It informs on the understanding of the role of verbal and nonverbal communication in the association between VI and fatigue with exercises aimed at fostering independence from others, facilitating adjustments in social relationships, and cultivating assertiveness.

Module 5: Relaxation

It helps to recognize stress and how to release tension by relaxation exercises.

Module 6 (optional): Improving sleep

It concentrates on understanding the importance of a fixed sleep-wake cycle, obtaining a personalized optimal sleep-wake cycle, and useful strategies to sleep or stay awake.

Module 7 (optional): Work optimization

It creates insight in balancing work-related stressors and energy sources and offers exercises to improve self-efficacy and possibilities for work resumption.

Module 8: The future

It focuses on achieving personal goals for the future, recognizing signs of relapse, and experiencing healthy levels of fatigue.

Patients were able to log in to their computer, tablet, or smartphone at their preferred place. Content was represented in module steps: the steps within a module that contained information or assignments. The e-platform logged how often, what content, and for how long patients were active over a period of 12 months. Social workers gave online feedback on assignments, and psychologists (n=2) were available for peer consultation.

Outcome Measures

A survey was conducted using the Dutch Mental Health Care Thermometer (MHT) questionnaire [24] after 6 months by telephone to evaluate patients' experiences. This questionnaire consisted of yes or no questions with regard to dimensions specified in the subscales: appreciation of information, appreciation of decision-making, and appreciation of professional and treatment results. Internal consistency of the subscales ranges between 0.70 and 0.86 [25]. Researchers (PhD candidates) from Amsterdam University Medical Centers conducted the assessments by telephone with registration into data-entry software called Castor [26]. There was no relationship established between researchers and patients before the study commencement. The e-platform provided data on user engagement.

Blended care is a relatively new way of providing guidance for professionals of low vision services in the Netherlands. The exploration of the experience of social workers and computer trainers with blended care and with guiding patients who follow the E-nergEYEze modules enabled a deeper understanding of insight into the professional perspective toward the digital aspect and content of the E-nergEYEze intervention within this target group. Therefore, social workers (n=7) completed the therapist evaluation for each patient. This questionnaire included questions on their experience with the intervention, online guidance, and compliance of patients to the intervention.

XSL•FO
RenderX

Computer trainers (n=4) kept time records and evaluated the computer training. To complement the therapist evaluation outcomes, a focus group meeting was held with the involved social workers (n=4), computer trainer (n=1), and psychologists (n=2).

Data Analysis

Careful consideration was given to information related to whether the intervention was completed. We decided to distinguish between patients who completed most of the module steps (≥80%) and patients who discontinued earlier (<80%). Quantitative data were analyzed using descriptive statistics, including baseline fatigue severity. Data were not normally distributed; therefore, results were reported with medians and IQR. The Mann-Whitney *U* test was performed to examine group differences in continuous variables and the chi-square test for nominal and ordinal variables. Qualitative data were analyzed by applying a thematic approach in MaxQDA2020 (VERBI Software) [27], a method to identify, analyze, and report patterns within data. To strengthen the reliability and validity of the analyses, MV (medical doctor and PhD candidate) and HvdA (senior researcher) separately analyzed answers to open questions of 3 evaluation forms to create thematic codes for either perspective, after which 2 code trees were created. All evaluations were then coded by MV, and data were analyzed by interpreting patterns within and between themes. Coding, analyses, and results were discussed in detail with HvdA.

Results from the professional perspective formed the content of the focus group meeting which was organized with 3 moderators (RvN, HvdA, and MV). Interaction among the professionals was encouraged with the leading questions “What goes well?” and “What can be improved and how?” This session was audio-recorded and transcribed verbatim by MV. Data were coded in vivo and analyzed thematically, using the preexisting code tree as a guide, and returned to the professionals for additional comments.

Ethical Considerations

Ethics approval was obtained by the Medical Ethical Committee of the VU University Medical Center Amsterdam, the Netherlands (reference no. 2019/027, NL67802.029.18). The COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist was applied in this manuscript [28]. Data was anonymized during data collection and analysis. Verbal consent was given by all professionals to record and transcribe the focus group meeting. All participants signed an informed consent form to participate in the study. Participants did not receive compensation for participation in the study.

Results

Results reflect user experiences and user engagement of patients and professionals with the E-nergEYEze intervention.

Participant Characteristics

Demographics and clinical characteristics of social workers and patients are presented in Tables 1 and 2, respectively. None of the variables were statistically different between patient subgroups. Clinically relevant differences were noticed in the subgroup who discontinued the intervention, including longer duration of the VI, less paid work, a higher percentage of participants living alone, and a history of psychiatric illness and comorbidities.

Table 1. Demographic and clinical characteristics of social workers (n=7) participating in the E-nergEYEze intervention.

Characteristics	Values
Age (y), median (IQR)	53.5 (49.0-62.0)
Sex (female), n (%)	6 (86)
Professional experience (y), median (IQR)	23.0 (13.0-31)
Working with individuals with VI[a] (y), median (IQR)	12.0 (1.0-19)
Previous experience with eHealth (yes), n (%)	2 (29)

[a]VI: visual impairment.

XSL•FO
RenderX

Table 2. Demographic and clinical characteristics of patients with visual impairment participating in the E-nergEYEze intervention.

Characteristics	Total group (n=51)	>80% of module steps completed (n=32)	<80% of module steps completed (n=19)	*P* value
Age (y), median (IQR)	58.0 (53.0-65.0)	57.5 (53.0-66.5)	59.0 (52.0-65.0)	.91[a]
Duration of VI[b] (y), median (IQR)	26.0 (14.0-47.0)	20.0 (10.8-41.3)	35.0 (19.0-51.0)	.14[a]
Education (y), median (IQR)	12.0 (11.0-15.0)	11.5 (11.0-15.0)	12.0 (11.0-15.0)	.94[a]
Fatigue severity score[c], median (IQR)	42.0 (37.0-47.0)	43.5 (37.0-48.0)	40.0 (36.0-44.0)	.13[a]
Sex (female), n (%)	32 (63)	21 (66)	11 (58)	.58[d]
Best-corrected eye[e], n (%)				.16[d]
Blind	11 (22)	5 (16)	6 (32)	
Severe VI	3 (6)	3 (9)	0 (0)	
Moderate VI	26 (51)	15 (47)	11 (58)	
Mild VI	11 (22)	9 (28)	2 (10)	
Eyesight (subjective), n (%)				.81[d]
Blind	2 (4)	1 (3)	1 (5)	
Very bad	21 (41)	12 (37)	9 (47)	
Bad	20 (39)	13 (41)	7 (37)	
Fair	8 (16)	6 (19)	2 (10)	
Ophthalmic conditions[f], n (%)				—[g]
Maculopathy	17 (33)	12 (37)	5 (26)	
Retinopathy	17 (33)	9 (28)	8 (42)	
Optic neuropathy	14 (27)	11 (34)	3 (16)	
Anterior segment	4 (8)	2 (6)	2 (10)	
Other[h]	9 (18)	4 (12)	5 (26)	
Comorbidities, n (%)				.13[d]
None	28 (55)	21 (66)	7 (37)	
1	13 (25)	6 (19)	7 (37)	
≥2	10 (20)	5 (16)	5 (26)	
History of psychiatric illness, n (%)	12 (23)	6 (19)	6 (32)	.30[d]
History of cancer, n (%)	6 (12)	3 (9)	3 (16)	.55[d]
Hearing impairment[i], n (%)	16 (31)	10 (31)	6 (32)	.64[d]
Work status, n (%)				.44[d]
Paid work	18 (35.3)	13 (41)	5 (26)	
Volunteer work	14 (27.5)	7 (22)	7 (37)	
No work	19 (37.3)	12 (37)	7 (37)	
Living alone, n (%)	14 (27.5)	6 (19)	8 (42)	.07[d]
Optional module, n (%)				—
“Improving sleep”	37 (72)	22 (69)	13 (68)	
“Work optimization”	24 (47)	14 (44)	9 (47)	

[a]Denotes *P* values for the Mann-Whitney U test.

[b]VI: visual impairment.

[c]Fatigue severity score: severe fatigue ≥35 points.

XSL•FO
RenderX

[d]Denotes *P* values for the chi-square test.

[e]Best-corrected eye categories: blind: logarithm of the minimum angle of resolution (LogMAR) >1.30 or peripheral vision loss (<10°); severe vision impairment: LogMAR 1.30; moderate vision impairment: LogMAR 0.52 to 1.00 or peripheral vision loss (<30°); central visual field impairment or homonyme heminopia; mild vision impairment: LogMAR 0.30 to 0.40 or peripheral vision loss (<45°); and no visual impairment: LogMAR ≤0.22 or quadrantanopia.

[f]Patients can have one or more ophthalmic conditions at the same time.

[g]Not available.

[h]Other ophthalmic diseases: uveitis, infectious, scleritis, nystagmus, albinism, microphtalmia, evisceration, enucleation, trauma, iatrogenic, and acquired brain disease.

[i]Hearing impairment: subjective hearing being fair or bad.

Experiences With the E-nergEYEze Intervention: Quantitative Results

eHealth Platform

Results on user engagement showed that median 89% (IQR 45-100) of all assigned module steps were completed (Table 3). In total, 32 (63%) patients completed most module steps (Tables 2 and 3) and graded the intervention median 8.0 (IQR 7.0-8.0). Modules were completed by at least 50% (37/51) of the participants and in both subgroups 50% (26/51) of the patients were active on the e-platform after completing the assigned modules. The number of patients that used the e-platform via the web browser and app was equally distributed. We noticed that the adherence to complete modules dropped below 50% after module 2 or during module 3 in the subgroup that discontinued the intervention, with little use of diaries (Tables 4 and 5). The diary on sleep-wake rhythm was used considerably more often compared to other diaries (Table 5). In addition, patients sent median 7.5 (IQR 4-12) messages and social workers sent median 12 (IQR 11-15) messages to their patients.

Table 3. User engagement data from patients who followed the E-nergEYEze intervention within the eHealth platform (n=51)[a].

User engagement data	Total group (n=51), median (IQR)	>80% of module steps completed (n=32), median (IQR)	<80% of module steps completed (n=19), median (IQR)
Total number of module steps completed (%)	88.9 (44.7-100.0)	__[b]	—
Total time to complete the intervention (wk)	17.0 (11.0-24.0)	18.0 (13.3-25.0)	13.0 (6.0-24.0)
Time per module (wk)	2.8 (1.7-4.3)	2.3 (1.6-3.5)	4 (1.8-7.5)
Total time there was log-in activity (wk)	25.0 (18.0-38.0)	26.0 (19.8-46.0)	23.0 (14.0-33.0)
Total number of log-in times, n	35.0 (23.0-58.0)	49.5 (35.0-87.5)	22.0 (12.0-30.0)
Web browser use	16.0 (7.0-26.0)	20.0 (10.3-35.0)	9.0 (2.0-16.0)
Phone or tablet app use	15.0 (0.0-39.0)	22.5 (0.5-64.3)	6.0 (0.0-26.0)

[a]Three patients dropped out before registration on the eHealth platform.

[b]Not available.

XSL•FO
RenderX

Table 4. Modules use data from patients who followed the E-nergEYEze intervention within the eHealth platform (n=51)[a].

Modules used	Total group (n=51), n (%)		>80% of module steps completed (n=32), n (%)		<80% of module steps completed (n=19), n (%)	
	Opened at least once	Completed 100%	Opened at least once	Completed 100%	Opened at least once	Completed 100%
Introduction: fatigue	48 (94)	44 (86)	32 (100)	30 (94)	16 (84)	14 (74)
Module 1: dealing with VI[b]	48 (94)	48 (94)	32 (100)	32 (100)	16 (84)	16 (84)
Module 2: negative thoughts	45 (88)	43 (84)	32 (100)	32 (100)	13 (68)	11 (58)
Module 3: graded activity	43 (84)	38 (74)	32 (100)	32 (100)	11 (58)	8 (42)
Module 4: social contact	37 (72)	33 (65)	32 (100)	30 (94)	5 (26)	3 (16)
Module 5: relaxation	38 (75)	32 (63)	32 (100)	31 (97)	6 (32)	1 (5)
Module 6 (optional): sleep	27[c] (73)	22[c] (60)	24 (100)	22 (92)	3 (24)	1 (8)
Module 7 (optional): work	16[d] (67)	14[d] (58)	15 (100)	14 (93)	1 (11)	0 (0)
Module 8: future	30 (59)	25 (50)	29 (91)	25 (81)	1 (5)	0 (0)

[a]Three patients dropped out before registration on the eHealth platform.

[b]VI: visual impairment.

[c]Module 6 was chosen by 37 patients.

[d]Module 7 was chosen by 24 patients.

Table 5. Diary use data from patients who followed the E-nergEYEze intervention within the eHealth platform (n=51)[a].

Diaries	Total group (n=51)		>80% of module steps completed (n=32)		<80% of module steps completed (n=19)	
	Used at least once, n (%)	Number of times used, median (IQR)	Used at least once, n (%)	Number of times used, median (IQR)	Used at least once, n (%)	Number of times used, median (IQR)
Diary use	37 (72)	21.0 (13.0-42.0)	28 (87)	27.0 (20.0-54.0)	7 (37)	10.0 (4.0-18.0)
“My reaction”	35 (69)	5.0 (2.0-8.0)	26 (81)	6.5 (2.8-8.0)	7 (37)	4.0 (2.0-8.0)
“My helpful thoughts”	32 (63)	2.5 (1.0-5.8)	27 (84)	4.0 (2.0-6.0)	5 (26)	1.0 (1.0-2.5)
“My activity distribution”	31 (61)	4 (2.0-10.0)	27 (84)	5.0 (2.0-11.0)	4 (21)	1.5 (1.0-8.0)
“My activity build-up”	23 (46)	3.5 (1.0-11.0)	20 (62)	3.5 (1.3-11.0)	3 (16)	2.0 (1.0-1.0)
“My mental activity build-up” (optional)	18 (35)	2.0 (1.0-4.3)	16 (50)	2.0 (1.0-4.8)	2 (10)	1.0 (1.0-1.0)
“My social activity build-up” (optional)	17 (33)	2.0 (1.0-2.5)	14 (44)	2.0 (1.0-3.0)	3 (16)	1.0 (1.0-1.0)
“My relaxation”	26 (51)	4.5 (2.0-7.3)	25 (78)	5.0 (2.0-7.5)	1 (5)	3.0 (3.0-3.0)
“My sleep-wake rhythm” (optional)	18 (35)	12.5 (2.8-14.0)	17 (53)	13.0 (2.5-14)	1 (5)	3.0 (3.0-3.0)

[a]Three patients dropped out before registration on the eHealth platform.

Dutch MHT Questionnaire

E-nergEYEze was given a median grade of 7.0 (IQR 6.0-8.0) and patients put effort into following the intervention correctly with a median score of 8.0 (IQR 5.0-8.0). In total, 65% (30/46) of patients responded with “yes” to the question “Did you feel the intervention was the right approach for your problems or complaints.” The same number of patients (30/46, 65%) responded positively on 2 or more progress questions: “Did the intervention give you more control over your problems or complaints?” “Have you made sufficient progress as a result of the intervention?” “Are you better able to do things that matter to you as a result of the intervention?” and “Does the treatment help you cope better with people and situations, which you previously had problems with?” Some patients also indicated that they did not receive sufficient information about the

XSL•FO
RenderX

intervention (5/46, 11%) and they felt that they were not given enough information about the expected outcome (16/46, 35%). Overall, most patients (39/45, 87%) reported that they would recommend E-nergEYEze to others.

Therapist Evaluation

Social workers spent median 190 (IQR 95-281) minutes per patient on preparation, making appointments, face-to-face session, online feedback, optional midterm or end evaluations, and peer consultation. The face-to-face session took a median 60 (IQR 33-90) minutes and providing feedback took a median 10 (IQR 10-15) minutes, with a median 8 (IQR 5-10) times per participant. Most contact with patients was online (79%) and by telephone (14%), and a small amount was live 4%). Social workers rated the effort patients (31/51, 61%) put into following the intervention correctly at a median of 8.0 (IQR 6.0-9.0). The computer trainer spent a median 30 (IQR 15-30) minutes on computer training with e-platform instructions. Additional computer training (n=11) was provided to 8 (N=51, 16%) patients for a median 15 (IQR 10-45) minutes, by telephone (7/11, 64%), online (2/11, 18%), and live (2/11, 18%). In total, 87% (7/8) of patients who requested additional computer training continued the intervention. Professionals provided a median grade between 7.0 and 8.0 (IQR 5.0-8.0).

Experiences With the E-nergEYEze Intervention: Qualitative Results

Overview

The main themes derived from the qualitative data of the Dutch MHT questionnaire from the patient's perspective (n=46), the therapist evaluation by social workers (n=7), and the focus groups with professionals (n=7) are summarized in Textbox 2. There were themes mentioned by the 3 stakeholder groups (patients, social workers, and professionals), for example, "effect and experiences," 2 stakeholder groups, for example, "expectation, motivation, and time" that was mentioned by patients and social workers, or themes reflected by one group. Textbox 2 shows the themes reflecting the qualitative data of patients and professionals.

Textbox 2. Main themes derived from the qualitative data of patients and professionals regarding the E-nergEYEze intervention.

Patients (n=46)

- Expectation, motivation, and time
- Content
- Guidance
- Digital accessibility
- Effect and experiences

Social workers (n=7)

- Expectation, motivation, and time
- Mental well-being
- Intervention delivery
- Providing digital support
- Effect and experiences
- Implementation suggestions

Focus group meeting with professionals (n=7)

- Intervention delivery
- Providing digital support
- Effect and experiences
- Implementation suggestions

Expectation, Motivation, and Time

Patients and social workers mentioned topics related to the theme "expectation, motivation, and time." Feedback was given by patients (n=5, 11%) on the expectation of the intervention content and digital aspects: they would have liked more information upfront regarding module themes, the amount and length of the modules, and the eHealth design regarding the use of an online application. Some patients experienced a threshold to start the intervention (n=2, 4%) or had difficulties continuing the modules (n=1, 2%). Others mentioned that they had invested much time and effort to follow the modules (n=3, 6%) or would have liked more time per module (n=4, 9%). In total, 3 (6%) patients were of the opinion that the time investment was too great and 2 (4%) patients experienced difficulties attending E-nergEYEze in addition to having a job and mentioned a reduction in energy because of the increased awareness of fatigue being present. Social workers mentioned positive factors in participants' continuation: having the appropriate expectations, a positive attitude, and completing the modules in time.

XSL•FO
RenderX

Content

Patients reflected positively on the intervention's educational content, of which an important experience was being able to extract what was personally relevant to apply in daily life.

Assignments about how to distribute activities during the day, formulate helpful thoughts, apply relaxation exercises, avoiding sleep during the day, and optimize their work situation were reflected positively:

> *I put more attention on modules that were applicable and meaningful to me, namely: helping thoughts and distributing activities. I try to apply what I learned from these two modules in my daily life and I benefit from it.* [Male patient, 55 y]

There were difficulties with the amount and level of assignments and keeping diaries (n=8). In addition, patients (n=8, 17%) experienced insufficient progress because the content was either familiar or did not meet personal needs with regard to age, education level, or lifestyle. However, opinions varied on this topic—the intervention should be made more recognizable to the younger, socially active, and nondepressed people, or to more highly educated people as well as a better fit for older and less educated people. If the VI was already accepted, modules 1 and 2 seemed less compatible.

Guidance and Intervention Delivery

Patients were positive about the digital guidance (n=6, 13%) but also desired more frequent and substantive guidance (n=16, 35%). Illustrative examples included the following:

> *Consultation with the social worker could be better, there was limited response. I would have liked more interaction and shorter links with the social worker. For me, telephone contact would have helped.* [Male patient, 62 y]

> *Guidance was minimal and limited in content. I only received responses about being on track, but I felt the need to discuss whether I was succeeding and what my experiences were.* [Female patient, 60 y]

The computer training was experienced as supportive (n=2, 4%). However, some patients suggested that guidance was not necessary (n=3, 6%) or preferred the intervention in a face-to-face setting (n=4, 9%). Social workers mentioned that patients who gave no, evasive, or short answers, had no depth in their answers or those who were in need of supportive or empowering online feedback, were more likely to discontinue. Regarding intervention delivery, they experienced challenges with goal setting, providing online feedback, and remote support, especially with regard to assessing patients' needs for the correct intensity and depth of online contact. Computer trainers (n=4) evaluated that most training sessions went well.

Mental Well-Being

When impeding personal circumstances or mental health issues were at play, patients struggled with continuing the intervention, for example, a changing home situation or the presence of mental or physical complaints caused by another illness.

Digital Accessibility and Providing Digital Support

Patients reflected positively on the eHealth design (n=4, 9%), with the possibility to reread information at a later time (n=2, 4%), and the diary reminders (n=1, 2%). However, digital skills among patients varied. Some patients mentioned no problems with the digital aspect of the intervention (n=4, 9%) and provided positive feedback on the presence of read-aloud audio within modules (n=3, 6%). Others mentioned inconveniences such as small font size (n=4, 9%) and no ability to flip the contrast (n=1, 2%) and had difficulties with navigating (n=4, 9%) and providing answers on the web (n=5, 11%). There were also patients who indicated that they had insufficient digital skills (n=2, 4%). Digital accessibility was mentioned as a barrier to progress (n=4, 9%). One of the responses included the following:

> *I am handy with my smartphone, but I found the app inconvenient for answering questions because some buttons were not read out loud or had to be operated in an illogical way. This was a stumbling point for me when following E-nergEYEze.* [Female patient, 51 y]

Social workers wondered, with regard to providing digital support, to what extent in-depth feedback was expected, how this should be shaped, and how to deal with evasive online contact. This is illustrated by the following quote:

> *I continue to find it difficult to provide proper feedback because from what is written I find it difficult to get clear to what extent a participant has really mastered the material. You don't really get to a 'conversation' even when you do have more questions.*

In contrast, only a few social workers consulted available psychologists for help. A total of 3 (43%) social workers indicated that they felt more comfortable with face-to-face contact and were of the opinion that affinity with eHealth is key.

Effect and Experiences

Patients experienced a positive change because of the intervention (n=9, 20%), for example, they gained more insight and awareness of problems based on the various topics covered in the intervention with valuable and educational information and personally relevant modules provided positive change in dealing with fatigue in daily life. Examples illustrating these results include the following:

> *In the past I was concerned with my eyes, but now in a different way. I have learned to think more positively. I am very enthusiastic about the intervention and have benefited greatly from several components.* [Female patient, 60 y]

> *I liked the e-learning as a tool very much, because I could read things a second or third time and therefore the content was better retained (vs. if it was presented orally). I was surprised that Module 1 was providing so much insight about the impact of low vision.* [Female patient, 56 y]

XSL•FO
RenderX

To the question "What do you think needs to be improved to increase the intervention grade?" the following responses were given:

> *Content was good, but I needed more personal guidance and more practical help tailored to my needs.* [Female patient, 34 y]
>
> *That the social worker was more responsive to the answers a person gave by mirroring for insight. There is also even more to be gained if guidance would capitalize more on specific problems given in answers to questions.* [Female patient, 67 y]
>
> *The intervention has to be changed, it was too long and intense. I felt a barrier to get started and felt the training burdensome for someone with fatigue.* [Female patient, 58 y]

Social workers felt the intervention had added value for patients and was a meaningful way to reduce fatigue in adults with VI. E-nergEYEze compelled thought and reflection by requiring users to fill in answers, allowing them to reread information, and allowing them to work at self-set time and pace. Patients gained insight and were able to make changes, especially if digital skills were present. In the opinion of the professionals, patients absorbed information of the different modules well. This was reflected by the following quote:

> *I thought it was good that my patient, while she had fatigue complaints since childhood and had developed a certain lifestyle, was still motivated and able to make changes with the help of the modules; for example she had instructed her environment not to ask about fatigue anymore.*

Implementation Suggestions

Professionals indicated that motivation, self-discipline, and digital skills are a requirement to follow E-nergEYEze. Educational level must match content level for proper understanding and intended progress with clear expectations about the online aspect, intensity, and outcome of the intervention. Negative predictors were noted if patients had an unclear help request or personal goal, difficulties with the online aspect, and challenges with the content and diary assignments. In addition, it was felt necessary to exclude patients for whom mental issues or other rehabilitation needs play a prominent role. Therefore, critical preselection is important. Suggestions for improvements were made regarding prescheduled time blocks for feedback and peer consultation. Guidance should be approachable in a low-threshold manner with midterm and final evaluation. Evaluation moments could be used for the evaluation and reformulation of personal goals. It was discussed that fatigue requires attention early in the rehabilitation process, that E-nergEYEze is a fairly complete training, and that it provides neutral communication of information and tools to reduce severe fatigue.

Discussion

Principal Findings

This study evaluated the user experiences of the E-nergEYEze intervention, an iCBT intervention to reduce severe fatigue [14], from the perspective of patients with VI and professionals who used E-nergEYEze [7]. These firsthand experiences showed that patients completed a median of almost 90% of the module steps, almost two-thirds of the participants completed most of the intervention, and some were active up until 12 months from baseline. More than half of the patients were active on the e-platform at least 1 week after individually completing the last module step. This does not suggest that every patient finished all module steps but confirms that patients were actively engaged in the intervention and the potential of repetition offered by eHealth was noticed and applied to achieve long-term progress. Continuing E-nergEYEze was related to the presence of digital proficiency; having the appropriate expectation; content that matches personal preferences and life context; and the absence of impeding personal circumstances, mental health issues or concurrent other rehabilitation programs. However, it is challenging to find or develop an e-platform that is all-encompassing for the accessibility needs of individuals with VI.

E-nerEYEze was perceived as valuable to patients, especially personally relevant topics that triggered a positive change in dealing with fatigue. iCBT seems to be equally effective as face-to-face cognitive behavioral therapy [10], though we noticed an adherence drop in the subgroup that discontinued after module 2 or during module 3. Module 3, on graded activity, proved to be especially important in the reduction of fatigue [29,30]; however, this module is also relatively long and intensive, which may have caused patients to drop out. Adherence could be increased by motivating patients, embedding a midterm evaluation, or considering changing the order of modules. Reevaluating the difficulty of the content with professionals in cognitive behavioral therapy could be useful as well, for example, by reconsidering content of module 3 on the essence of physical activity to increase physical health and reduce fatigue. The examples of physical activities could be reconsidered in consultation with patient representatives to increase the feasibility of this module for individuals with VI and to lower the barrier to starting physical activities. The length of a module affects concentration and fatigue symptoms; however, a module does not have to be completed all at once. Practical alternatives to deal with length and intensity of content suggested by Babbage et al [17] were already implemented in the E-nergEYEze intervention, such as the use of voice-overs, read-aloud audio, and videos. We chose a blended design for E-nergEYEze design in consideration of the benefits of eHealth [9,10] and the more favorable outcomes that have been reported for guided iCBT versus unguided iCBT [31-33]. Our results showed that guidance was essential for patients; we underestimated the influence of and need for frequent and substantive support. Social workers were instructed to follow, motivate, and confirm patients' progress in feedback moments, without extensive conversations on the costs and benefits of eHealth; this was seen as a discrepancy with patients' needs.

XSL•FO
RenderX

Worm-Smeitink et al [34] suggested including more flexibility in the intensity of guidance, a suggestion that we echo with respect to remote support. Our social workers had no experience in providing online support, and some struggled with the remote nature of eHealth, especially because of the limited interaction. Online interaction calls on other skills; therefore, skill, time and input could have influenced the treatment effects [15,19]. Precisely because eHealth is a new mode of treatment in this target population, it is important to pay adequate attention to the training of professionals providing low vision services and arrange prescheduled guidance time and peer consultation moments. In accordance with Dijksman et al [18], our results indicated that professionals need more support in providing substantive remote support.

Strengths and Limitations of This Study

The major strengths of this study are presented in Textbox 3.

Textbox 3. Strengths of this study.

- User experiences of the E-nergEYEze intervention were studied in the intervention arm of a single-blinded randomized controlled design.
- Quantitative and qualitative data were combined for optimal representation of practice.
- Stakeholders contributed to the development and guidance of E-nergEYEze to secure a pragmatic design.
- Internal validity was ensured in compliance with the study protocol.
- User experiences were analyzed ahead of the cost-effectiveness results.

We considered it a strength that these user experiences were analyzed ahead of the cost-effectiveness results, which confirm an independent analysis and interpretation of data. E-nergEYEze was developed by various professionals and patients, based on evidence-based iCBT [35,36] and was pilot tested [20]. Patients gave positive feedback on improvements made based on the pilot study, reflecting the value of pilot testing before conducting an RCT. The content variety gave patients the possibility to personalize their treatment by applying coping strategies that appealed to them in daily life, as well as broadening their horizon outside their comfort zone. E-nergEYEze has been implemented in compliance with the study protocol [7], ensuring the internal validity of these results with respect to the intervention. However, the reliance of self-reported time records of professionals could limit actual or perceived validity. Another limitation that should be considered is the possibility of self-assessment bias, that is, the overestimation of one's own performance. The mixed method design included qualitative analysis; therefore, discussing reliability is challenging, but we have striven to be consistent in the analysis. Internal validity was enhanced by triangulation of research methods and verification of qualitative analysis. The sample size of this study was relatively small, as we zoomed in on the user experiences of the intervention group and the professionals who were involved. This could limit the generalizability of the results. In addition, attrition bias could have influenced the results, as 5 patients did not complete the Dutch MHT questionnaire.

Implications

Our results respond to the knowledge gap regarding user experiences with blended iCBT in adults with VI [37]. To determine whether a patient receiving low vision services is suitable for E-nergEYEze, the difference in patient characteristics between the subgroups that we investigated may be of clinical relevance. Patients who discontinued the intervention could have benefited from receiving more support, for which training of professionals is essential. Guided iCBT has increasingly been investigated, with varying guidance intensity [15,36,38,39], and it would be valuable to examine what frequency and type of guidance can best be offered in iCBT, as well as the amount of content and how to match the difficulty to the skills and opportunities of patients. Taking our results into consideration, individuals with VI who experience severe fatigue, who are digitally sufficient, and who are willing to receive remote treatment with the ability and motivation to work at self-set times are likely to benefit from the E-nergEYEze intervention. Individuals with impeding personal circumstances, mental health issues, or who are participating in other concurrent rehabilitation programs are less likely to benefit or more likely to discontinue the E-nergEYEze intervention.

Conclusions

The potential of the E-nergEYEze intervention to reduce fatigue severity in individuals with VI was underlined by user experiences. These results provide information to learn from and optimize E-nergEYEze into a useful, available, and affordable tool to reduce severe fatigue in individuals with VI. This research offers future prospects for the development of blended eHealth interventions for other conditions in this target group with specific needs regarding accessibility of e-platforms and guidance with an affinity for providing digital care.

Acknowledgments

This work was supported by the Dutch Organization for Health Research and Development (ZonMw program Inzicht; grant 60-00635-98-219) and the health care professionals of Royal Dutch Visio and Bartiméus. The authors thank their colleagues in the low vision research group for their help in collecting the data.

XSL•FO
RenderX

Data Availability

The nonidentifiable data and data collected directly from patients during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

RvN and HvdA conceived the study design, with extensive advice from GvR, HK, CB, and MK. MV executed the trial with substantive support from RvN and HvdA. MV and HvdA performed data analysis. MV wrote the manuscript. All authors revised the manuscript and approved it for submission.

Conflicts of Interest

None declared.

References

1. Schakel W, van der Aa HP, Bode C, Hulshof CT, van Rens GH, van Nispen RM. The economic burden of visual impairment and comorbid fatigue: a cost-of-illness study (from a societal perspective). Invest Ophthalmol Vis Sci 2018 Apr 01;59(5):1916-1923. [doi: 10.1167/iovs.17-23224] [Medline: 29677352]
2. Belenky G, Wesensten NJ, Thorne DR, Thomas ML, Sing HC, Redmond DP, et al. Patterns of performance degradation and restoration during sleep restriction and subsequent recovery: a sleep dose-response study. J Sleep Res 2003 Mar;12(1):1-12 [FREE Full text] [doi: 10.1046/j.1365-2869.2003.00337.x]
3. van den Berg SW, Gielissen MF, Custers JA, van der Graaf WT, Ottevanger PB, Prins JB. BREATH: web-based self-management for psychological adjustment after primary breast cancer--results of a multicenter randomized controlled trial. J Clin Oncol 2015 Sep 01;33(25):2763-2771. [doi: 10.1200/JCO.2013.54.9386] [Medline: 26169621]
4. Pöttgen J, Moss-Morris R, Wendebourg JM, Feddersen L, Lau S, Köpke S, et al. Randomised controlled trial of a self-guided online fatigue intervention in multiple sclerosis. J Neurol Neurosurg Psychiatry 2018 Sep;89(9):970-976. [doi: 10.1136/jnnp-2017-317463] [Medline: 29549193]
5. Schakel W, Bode C, van der Aa HP, Hulshof CT, Bosmans JE, van Rens GH, et al. Exploring the patient perspective of fatigue in adults with visual impairment: a qualitative study. BMJ Open 2017 Aug 03;7(8):e015023 [FREE Full text] [doi: 10.1136/bmjopen-2016-015023] [Medline: 28775181]
6. Bruijning JE, van Rens G, Fick M, Knol DL, van Nispen R. Longitudinal observation, evaluation and interpretation of coping with mental (emotional) health in low vision rehabilitation using the Dutch ICF Activity Inventory. Health Qual Life Outcomes 2014 Dec 24;12:182 [FREE Full text] [doi: 10.1186/s12955-014-0182-4] [Medline: 25539603]
7. Veldman MH, van der Aa HP, Bode C, Knoop H, Hulshof CT, Koopmanschap M, et al. E-nergEYEze, a vision-specific eHealth intervention based on cognitive behavioral therapy and self-management to reduce fatigue in adults with visual impairment: study protocol for a randomized controlled trial. Trials 2021 Dec 28;22(1):966 [FREE Full text] [doi: 10.1186/s13063-021-05935-w] [Medline: 34963472]
8. Knoop H, van Kessel K, Moss-Morris R. Which cognitions and behaviours mediate the positive effect of cognitive behavioural therapy on fatigue in patients with multiple sclerosis? Psychol Med 2012 Jan;42(1):205-213. [doi: 10.1017/S0033291711000924] [Medline: 21672300]
9. van Uden-Kraan CF, Drossaert CH, Taal E, Seydel ER, van de Laar MA. Participation in online patient support groups endorses patients' empowerment. Patient Educ Couns 2009 Jan;74(1):61-69. [doi: 10.1016/j.pec.2008.07.044] [Medline: 18778909]
10. Hedman E, Ljótsson B, Lindefors N. Cognitive behavior therapy via the internet: a systematic review of applications, clinical efficacy and cost-effectiveness. Expert Rev Pharmacoecon Outcomes Res 2012 Dec;12(6):745-764. [doi: 10.1586/erp.12.67] [Medline: 23252357]
11. Senjam SS, Manna S, Bascaran C. Smartphones-based assistive technology: accessibility features and apps for people with visual impairment, and its usage, challenges, and usability testing. Clin Optom (Auckl) 2021;13:311-322 [FREE Full text] [doi: 10.2147/OPTO.S336361] [Medline: 34866955]
12. Power C, Freire AP, Petrie H, Swallow D. Guidelines are only half of the story: accessibility problems encountered by blind users on the web. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2012 Presented at: CHI '12; May 5-10, 2012; Austin, TX p. 433-442 URL: https://dl.acm.org/doi/10.1145/2207676.2207736 [doi: 10.1145/2207676.2207736]
13. Ferati M, Vogel B, Kurti A, Raufi B, Astals DS. Web accessibility for visually impaired people: requirements and design issues. In: Proceedings of the 1st International Workshop, UsARE 2012, Held in Conjunction with ICSE 2012, Zurich, Switzerland, June 4, 2012 and Second International Workshop, UsARE 2014, Held in Conjunction with RE 2014 on Usability- and Accessibility-Focused Requirements Engineering. 2016 Presented at: UsARE '14; August 25, 2014; Karlskrona, Sweden p. 79-96 URL: https://link.springer.com/chapter/10.1007/978-3-319-45916-5_6
14. Wheeler S, Holloway I. Qualitative Research in Nursing and Healthcare. 3rd edition. London, UK: Wiley Blackwell Publishing; 2010.

XSL•FO
RenderX

15. Moss-Morris R, McCrone P, Yardley L, van Kessel K, Wills G, Dennison L. A pilot randomised controlled trial of an internet-based cognitive behavioural therapy self-management programme (MS Invigor8) for multiple sclerosis fatigue. Behav Res Ther 2012 Jun;50(6):415-421 [FREE Full text] [doi: 10.1016/j.brat.2012.03.001] [Medline: 22516321]
16. Muijs LT, de Wit M, Knoop H, Snoek FJ. Feasibility and user experience of the unguided web-based self-help app 'MyDiaMate' aimed to prevent and reduce psychological distress and fatigue in adults with diabetes. Internet Interv 2021 Sep;25:100414 [FREE Full text] [doi: 10.1016/j.invent.2021.100414] [Medline: 34401373]
17. Babbage DR, van Kessel K, Drown J, Thomas S, Sezier A, Thomas P, et al. MS energize: field trial of an app for self-management of fatigue for people with multiple sclerosis. Internet Interv 2019 Dec;18:100291 [FREE Full text] [doi: 10.1016/j.invent.2019.100291] [Medline: 31890637]
18. Dijksman I, Dinant GJ, Spigt M. The perception and needs of psychologists toward blended care. Telemed J E Health 2017 Dec;23(12):983-995. [doi: 10.1089/tmj.2017.0031] [Medline: 28556693]
19. van Kessel K, Wouldes T, Moss-Morris R. A New Zealand pilot randomized controlled trial of a web-based interactive self-management programme (MSInvigor8) with and without email support for the treatment of multiple sclerosis fatigue. Clin Rehabil 2016 May;30(5):454-462. [doi: 10.1177/0269215515584800] [Medline: 25952587]
20. Veldman MH, van der Aa HP, Knoop H, Bode C, Hulshof CT, van der Ham L, et al. Usability and feasibility of E-nergEYEze: a blended vision-specific E-health based cognitive behavioral therapy and self-management intervention to reduce fatigue in adults with visual impairment. BMC Health Serv Res 2023 Nov 16;23(1):1271 [FREE Full text] [doi: 10.1186/s12913-023-10193-4] [Medline: 37974261]
21. Vercoulen JH, Alberts M, Bleijenberg G. De Checklist Individual Strenght (CIS). Gedragstherapie 1999;32:131-136.
22. Callahan CM, Unverzagt FW, Hui SL, Perkins AJ, Hendrie HC. Six-item screener to identify cognitive impairment among potential subjects for clinical research. Med Care 2002 Sep;40(9):771-781. [doi: 10.1097/00005650-200209000-00007] [Medline: 12218768]
23. Ehealth platform with personal routes to change. Minddistrict. URL: https://www.minddistrict.com/en-gb [accessed 2025-05-29]
24. Kok I, van Wijngaarden B. Client valuation in the GGZ. Questionnaires and manual: GGZ Thermometer for adult care. Trimbos-instituut/GGZ Nederland Utrecht. 2003. URL: https://www.therapieland.nl/vragenlijsten/evaluatie/ggz-jeugdthermometer-ouders-lang-behandeling-kind [accessed 2025-05-29]
25. Jongerius J, Hull H, Derks J. Hoe scoort de verslavingszorg?: kwaliteitsbeoordeling door clie¨nten: een landelijk onderzoek. Hague, The Netherlands: Koninklijke Bibliotheek; 1994.
26. Castor electronic data capture. Castor. URL: https://castoredc.com [accessed 2024-05-29]
27. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
28. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec 16;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
29. Moss-Morris R, Sharon C, Tobin R, Baldi JC. A randomized controlled graded exercise trial for chronic fatigue syndrome: outcomes and mechanisms of change. J Health Psychol 2005 Mar;10(2):245-259. [doi: 10.1177/1359105305049774] [Medline: 15723894]
30. de Gier M, Picariello F, Slot M, Janse A, Keijmel S, Menting J, et al. The relation between cognitive-behavioural responses to symptoms in patients with long term medical conditions and the outcome of cognitive behavioural therapy for fatigue - a secondary analysis of four RCTs. Behav Res Ther 2023 Feb;161:104243 [FREE Full text] [doi: 10.1016/j.brat.2022.104243] [Medline: 36549190]
31. Karyotaki E, Efthimiou O, Miguel C, Bermpohl FM, Furukawa TA, Cuijpers P, Individual Patient Data Meta-Analyses for Depression (IPDMA-DE) Collaboration, et al. Internet-based cognitive behavioral therapy for depression: a systematic review and individual patient data network meta-analysis. JAMA Psychiatry 2021 Apr 01;78(4):361-371 [FREE Full text] [doi: 10.1001/jamapsychiatry.2020.4364] [Medline: 33471111]
32. Baumeister H, Reichler L, Munzinger M, Lin J. The impact of guidance on internet-based mental health interventions — a systematic review. Internet Interv 2014 Oct;1(4):205-215. [doi: 10.1016/j.invent.2014.08.003]
33. Richards D, Richardson T. Computer-based psychological treatments for depression: a systematic review and meta-analysis. Clin Psychol Rev 2012 Jun;32(4):329-342. [doi: 10.1016/j.cpr.2012.02.004] [Medline: 22466510]
34. Worm-Smeitink M, Janse A, van Dam A, Evers A, van der Vaart R, Wensing M, et al. Internet-based cognitive behavioral therapy in stepped care for chronic fatigue syndrome: randomized noninferiority trial. J Med Internet Res 2019 Mar 14;21(3):e11276 [FREE Full text] [doi: 10.2196/11276] [Medline: 30869642]
35. van der Aa HP, van Rens GH, Verbraak FD, Bosscha M, Comijs HC, van Nispen RM. Anxiety and depression in patients who receive anti-VEGF treatment and the usability and feasibility of e-mental health support: the E-PsEYE pilot study. Ophthalmic Physiol Opt 2021 Jul;41(4):808-819 [FREE Full text] [doi: 10.1111/opo.12837] [Medline: 34050550]
36. Menting J, Tack CJ, van Bon AC, Jansen HJ, van den Bergh JP, Mol MJ, et al. Web-based cognitive behavioural therapy blended with face-to-face sessions for chronic fatigue in type 1 diabetes: a multicentre randomised controlled trial. Lancet Diabetes Endocrinol 2017 Jun;5(6):448-456. [doi: 10.1016/S2213-8587(17)30098-0] [Medline: 28462869]

XSL•FO
RenderX

37. Denecke K, Schmid N, Nüssli S. Implementation of cognitive behavioral therapy in e-mental health apps: literature review. J Med Internet Res 2022 Mar 10;24(3):e27791 [FREE Full text] [doi: 10.2196/27791] [Medline: 35266875]
38. Abrahams HJ, Gielissen MF, Donders RR, Goedendorp MM, van der Wouw AJ, Verhagen CA, et al. The efficacy of Internet-based cognitive behavioral therapy for severely fatigued survivors of breast cancer compared with care as usual: a randomized controlled trial. Cancer 2017 Oct 01;123(19):3825-3834 [FREE Full text] [doi: 10.1002/cncr.30815] [Medline: 28621820]
39. Janse A, Worm-Smeitink M, Bleijenberg G, Donders R, Knoop H. Efficacy of web-based cognitive-behavioural therapy for chronic fatigue syndrome: randomised controlled trial. Br J Psychiatry 2018 Feb;212(2):112-118. [doi: 10.1192/bjp.2017.22] [Medline: 29436329]

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Research
iCBT: intervention based on cognitive behavioral therapy
MHT: Mental Health Care Thermometer
MS: multiple sclerosis
RCT: randomized controlled trial
VI: visual impairment

Edited by A Mavragani; submitted 25.09.23; peer-reviewed by E Molinari, S Haag; comments to author 27.08.24; revised version received 17.09.24; accepted 24.12.24; published 07.08.25.

Please cite as:
Veldman MHJ, van der Aa HPA, Knoop H, Bode C, van Rens GHMB, van Nispen RMA
User Experiences Among Patients and Health Care Professionals Who Participated in a Randomized Controlled Trial of E-nergEYEze, a Vision-Specific eHealth Intervention to Reduce Fatigue in Adults With Visual Impairment: Mixed Methods Study
JMIR Form Res 2025;9:e53080
URL: https://formative.jmir.org/2025/1/e53080
doi:10.2196/53080
PMID:

©Manon HJ Veldman, Hilde PA van der Aa, Hans Knoop, Christina Bode, Ger HMB van Rens, Ruth MA van Nispen. Originally published in JMIR Formative Research (https://formative.jmir.org), 07.08.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

What Matters Most to Veterans When Deciding to Use Technology for Health: Cross-Sectional Analysis of a National Survey

Bella Etingen[1,2,3], PhD; Bridget M Smith[1,4,5], PhD; Stephanie L Shimada[1,6,7,8], PhD; Stephanie A Robinson[1,6,9], PhD; Robin T Higashi[1,3], PhD; Ndindam Ndiwane[1,6], MPH; Kathleen L Frisbee[1,10], MPH, PhD; Jessica M Lipschitz[11,12], PhD; Eric Richardson[1,12,13], PhD; Dawn Irvin[1,4], BS; Timothy P Hogan[1,3,6], PhD

[1]eHealth Partnered Evaluation Initiative, VA Bedford Healthcare System, Bedford, MA, United States
[2]Research and Development Service, Dallas VA Medical Center, Dallas, TX, United States
[3]Peter O'Donnell Jr. School of Public Health, The University of Texas Southwestern Medical Center, Dallas, TX, United States
[4]Center of Innovation for Complex Chronic Healthcare, Edward Hines, Jr. VA Hospital, Hines, IL, United States
[5]Feinberg School of Medicine, Northwestern University, Chicago, IL, United States
[6]Center for Health Optimization and Implementation Research, VA Bedford Healthcare System, Bedford, MA, United States
[7]Department of Health Law, Policy, and Management, School of Public Health, Boston University, Boston, MA, United States
[8]Department of Population and Quantitative Health Sciences, University of Massachusetts Chan Medical School, Worcester, MA, United States
[9]The Pulmonary Center, School of Medicine, Boston University, Boston, MA, United States
[10]Office of Connected Care, Veterans Health Administration, Washington, DC, United States
[11]Brigham and Women's Hospital, Boston, MA, United States
[12]Department of Psychiatry, Harvard Medical School, Boston, MA, United States
[13]Center for Health Optimization and Implementation Research, VA Boston Healthcare System, Boston, MA, United States

Corresponding Author:
Bella Etingen, PhD
Research and Development Service
Dallas VA Medical Center
4500 S Lancaster Rd.
Dallas, TX, 75216
United States
Phone: 1 214 742 8387
Email: bella.etingen@va.gov

Abstract

Background: There is an increasingly diverse range of mobile apps and digital health devices available to help patients manage their health. Despite evidence for the effectiveness of such technologies, their potential has not been fully realized because adoption remains low. Such limited uptake can have direct implications for the intended benefits of these technologies.

Objective: This study aimed to understand what matters most to US military veterans when deciding whether to use digital health technologies (DHTs) such as mobile health apps or devices to manage their health and compare these factors between veterans with and without prevalent chronic physical and mental health conditions.

Methods: We conducted a cross-sectional analysis of survey data collected from a national sample of veterans who receive care from the Veterans Health Administration (VHA), which was predominantly gathered as part of the last wave of a larger longitudinal data collection effort.

Results: Among respondents (n=857), 86.7% (736/849) reported currently using or having previously used ≥1 devices to manage their health, and 78.4% (639/815) also reported using either VHA or non-VHA health apps. Considerations most frequently endorsed as "very important" by veterans when deciding whether to use DHTs included receiving secure messages from their health care team about DHTs, knowing data from DHTs would be used to inform their care, and receiving recommendations from providers to use DHTs. Conversely, considerations most frequently endorsed as "not at all important" included seeing information about DHTs on social media, having community support to use DHTs, and receiving encouragement from peers to use DHTs. Considerations did not significantly differ between veterans with or without prevalent chronic health conditions; however, a greater proportion of veterans with prevalent mental health conditions reported the following considerations to be

XSL•FO
RenderX

"very important:" seeing information about DHTs on social media, having community support to use DHTs, having other veterans encourage DHT use, and having help from family, friends, or other important people to use DHTs.

Conclusions: Understanding what matters most to patients when they are deciding to adopt a technology for their health can, and should, inform implementation strategies and other approaches to enhance health-related technology use. Our results suggest that, for veterans, recommendations from health care team members and knowing that the data from DHTs will be used in clinical care are more important than information from social media, community sources, or peers when deciding to use DHTs, although perceptions of importance regarding the latter may differ among patients with different conditions. Our findings suggest that communication from health care team members to patients, perhaps either in-person or electronically, could help encourage DHT adoption and use.

(JMIR Form Res 2025;9:e77113) doi:10.2196/77113

KEYWORDS

digital health technology; implementation; mobile health apps; patient preferences; veterans

Introduction

Mobile apps to support health management are proliferating and can support condition self-management, health behavior tracking, and health education [1-4]. Mobile health apps can be accessed and used on or with various technology platforms and digital health–related devices, including smartphones, tablet computers, and sensor technologies like smart watches. We refer to such mobile health apps and digital health–related devices as digital health technologies (DHTs).

Evidence suggests that DHT use can positively affect health behaviors and outcomes [5-9]. Accordingly, the Veterans Health Administration (VHA), the largest integrated health care system in the United States, is championing incorporation of DHTs into care. However, adoption and use of DHTs remain limited, which are among the most significant problems facing DHTs [10-17]. Prominent theoretical frameworks posit various constructs that can influence uptake of different technologies, including, but not limited to, perceptions of ease of use, usefulness, the opinions of others, and existing support for technology use, although the relative importance of such factors varies in different contexts [18-20].

For DHTs to realize their potential, efforts to bolster adoption are needed. In implementation research, there is increasing recognition of the importance of stakeholder-informed strategies to support intervention adoption [21-24], that is, it is key to understand what stakeholders (in this case, patients who use technology for health) think the most effective ways are for promoting adoption and use of interventions and to design and tailor implementation strategies accordingly [25].

Our objective in this evaluation was to understand what matters most to veterans who are known technology users when deciding whether to use DHTs and to compare perceptions between those with and without prevalent chronic physical and mental health conditions to inform the development of future implementation strategies for DHTs.

Methods

Design

In partnership with researchers from the VHA's Quality Enhancement Research Initiative Program, the VHA Office of Connected Care, responsible for the VHA's digital health strategy, launched the Veterans Engagement with Technology Collaborative (VET-C) cohort in 2017. A longitudinal data collection effort, the VET-C cohort combines veteran survey data collected over 3 timepoints with VHA administrative records [26]. The goal of the cohort is to support veteran engagement in the evaluation of VHA patient-facing technologies [26]. This manuscript reports the results of a cross-sectional analysis of data gathered predominantly as part of the third round of data collection efforts from the VET-C cohort to understand perspectives on what is most important to veterans when deciding whether to use DHTs, with the goal to inform implementation strategies that the VHA can leverage to support DHT adoption.

Participants

The VET-C cohort consists of a national sample of veterans who received care from 14 geographically distinct VHA facilities. These sites were selected purposefully to represent VHA facilities with high use of secure messaging supported by My HealtheVet (the VHA's web-based patient portal), diverse segments of the veteran population, hubs for VHA research, and experience with technology rollouts. Data collection did not focus on a specific clinical setting. Veterans invited to the cohort were active users of VHA health care and technology users, defined as owning a cellular phone and having sent ≥5 secure messages through My HealtheVet in the previous year [26,27].

VET-C surveys were collected between 2017-2018, 2019-2020, and 2021-2022. As noted above, data for the variables reported in this analysis were primarily gathered in the third round of VET-C data collection from veterans who had responded to the first 2 rounds of surveys. Information about the first 2 rounds of VET-C survey data collection is published elsewhere [26-28].

Data Collection

Our recruitment procedures comprised several steps. Veterans were first mailed a copy of the third round VET-C survey along with a cover letter and postage-paid reply envelope. A follow-up mailing was then sent to nonrespondents approximately 4 weeks after the initial mailing to facilitate response. The survey was designed iteratively based on veteran feedback and Office of Connected Care leadership input and included a combination of validated items and customized questions.

XSL•FO
RenderX

Measures

Veterans were asked to report their sex, age, race and ethnicity, highest level of education, relationship status, self-perceived physical and mental health status [29], extent of health care received through the VHA (compared to community settings), and typical travel time (in minutes) from home to their VHA primary care doctor.

We further asked veterans to report technology access and ownership, technology use [30], and whether they are mobile health app users and health-related digital device users. We also asked veterans to report how easy it is for them to communicate with their VHA health care team.

We collected information on perceptions of what is important when deciding whether to use technology for health by asking veterans to rate various considerations as “very important,” “somewhat important,” or “not at all important” when deciding whether to use DHTs to help them manage their health. Response options included considerations at multiple levels, including those related to the health care system, the community, the patient themselves, and an open-ended “other” response option where respondents could list and rate considerations not otherwise included among the closed-ended response options.

We supplemented survey data with administrative data to identify chronic physical and mental health conditions from the prior 5 years that are prevalent among veterans and to fill in missing demographic data as necessary. We also used administrative data to calculate a prior-year comorbidity index score using the hierarchical condition community score [31,32]. The hierarchical condition community score factors in a person’s diagnoses, eligibility for Medicare and Medicaid services, sex, and age and commonly ranges from 0.9-1.7, with scores greater than 1 often interpreted as indicating poorer health [31-33].

Analyses

We used descriptive statistics to characterize survey data, and chi-square analyses to compare perceptions of what veterans consider important when deciding whether to use technology for health, comparing those with and without prevalent chronic health conditions (ie, diabetes, hypertension, and chronic kidney disease) and prevalent mental health conditions (ie, depression, anxiety, and posttraumatic stress disorder). Analyses were conducted in Stata MP (version 14.2; StataCorp).

Ethical Considerations

This work was reviewed by relevant VHA Institutional Review Boards and designated as program evaluation for quality improvement purposes, exempting it from further oversight (Program Guide 1200.21) [34]. Returning a survey was considered consent to participate. Data were deidentified and labeled with a unique ID number. Veterans were not compensated for their participation.

Results

Response Rate

We invited 1373 veterans to complete a third-round survey. We removed 15 veterans from the denominator because their surveys were undeliverable or they had passed away. A total of 858 (63.2%) veterans returned a survey. We removed 1 veteran with unavailable health condition data from our analytic cohort.

Sample Description

As summarized in Tables 1 and 2, survey respondents were predominantly men (750/857, 87.5%), White (762/857, 88.9%), and older than 65 years of age (602/857, 70.2%). Over half (452/853, 53%) considered themselves early technology adopters. Most (736/849, 86.7%) reported currently using or previously having used ≥1 health-related digital devices and currently or previously using either VHA or non-VHA mobile health apps (639/815, 78.4%).

XSL•FO
RenderX

Table 1. Demographic of veteran respondents to the Veterans Engagement with Technology Collaborative cohort survey.

Demographic	n (%)
Age (n=857)	
Aged over 65 years	602 (70.2)
Aged 65 years or younger	255 (29.8)
Sex (n=857)	
Men	750 (87.5)
Women	107 (12.5)
Race (n=857)	
White	762 (88.9)
Black	55 (6.4)
Other	40 (4.7)
Ethnicity (n=857)	
Non-Hispanic	830 (96.9)
Hispanic	27 (3.3)
Relationship status (n=826)	
In a relationship	610 (73.8)
Not in a relationship	216 (26.2)
Education status (n=851)	
At least some college or vocational school	745 (87.5)
Less than some college or vocational school	106 (12.5)
Socioeconomic status (n=775)	
Not very hard to pay for basics[a]	586 (75.6)
Hard to pay for basics	189 (24.4)
Location of medical care (n=856)	
Mostly receives medical care in the VHA[b]	616 (72)
Does not mostly receive medical care in the VHA	240 (28)
Time to VHA primary care doctor's office (n=856)	
<60 min	726 (84.8)
>60 min	130 (15.2)
Perceived physical health status (n=853)	
Good or very good or excellent	596 (69.9)
Fair or poor	257 (30.1)
Perceived mental health status (n=855)	
Good or very good or excellent	694 (81.2)
Fair or poor	161 (18.8)
Ease of communication (n=850)	
Very easy or easy	664 (78.1)
Very difficult or difficult or neutral	186 (21.9)
Hierarchical condition community score (n=857), mean (SD)	1.7 (1.3)
Technology ownership (n=849)	
Desktop or laptop computer	793 (93.4)
Smartphone	761 (89.6)

XSL•FO
RenderX

Demographic	n (%)
Tablet computer	448 (52.8)
Technology adoption (n=853)	
Early technology adopter	452 (53)
Not an early technology adopter	401 (47)
Comfort using computers (n=852)	
Very comfortable or confident using computers	759 (89.1)
Not very comfortable or confident using computers	93 (10.9)
Self-reported mobile health app user? (n=815)	
Yes	639 (78.4)
No	176 (21.6)
Self-reported health-related digital device user? (n=849)	
Yes	736 (86.7)
No	113 (13.3)

[a]Basics refers to essential needs like food and heating and cooling.

[b]VHA: Veterans Health Administration.

Table 2. Health conditions of veteran respondents to the Veterans Engagement with Technology Collaborative cohort survey. Health conditions from only the prior 5 years are included.

Health condition (n=857)	n (%)
Hypertension	667 (77.8)
Osteoarthritis	453 (52.9)
Anxiety disorders	404 (47.1)
Diabetes	377 (44)
Depression	327 (38.2)
Chronic kidney disease	257 (30)
Ischemic heart disease	256 (29.9)
Asthma	204 (23.8)
Posttraumatic stress disorder	251 (29.3)
Chronic obstructive pulmonary disease	158 (18.4)
Peripheral vascular disease	136 (15.9)
Atrial fibrillation	130 (15.2)
Heart failure	92 (10.7)
Prostate cancer	65 (7.6)
Stroke	56 (6.5)
Traumatic brain injury	26 (3)
Acute myocardial infarction	25 (2.9)
Lung cancer	20 (2.3)
Colorectal cancer	13 (1.5)

Considerations When Deciding Whether to Use DHTs

Considerations most frequently endorsed as “very important” included receiving secure messages through the VHA’s My HealtheVet web-based patient portal from one’s health care team about DHTs (561/766, 73.2%), knowing that one’s health care team will use the information entered into the DHTs to inform one’s care (427/756, 56.5%), having one’s provider recommend DHTs (402/764, 52.6%), having somewhere to get help with issues using DHTs (389/760, 51.2%), and having other members of one’s health care team recommend DHTs (345/760, 45.4%). These considerations did not significantly

XSL•FO
RenderX

differ between veterans with or without prevalent chronic health conditions or prevalent mental health conditions.

The considerations most frequently endorsed as "not at all important" by veterans included the following: seeing information about DHTs on social media (533/756, 70.5%); having community support through Veteran Service Organizations, churches, libraries, or other organizations (501/754, 66.4); having other veterans encourage DHT use (430/758, 56.7%); having help from family, friends, or other important people to use DHTs (315/759, 41.6%); and receiving printed information about DHTs from the VHA (201/758, 26.5%). These considerations did not significantly differ between veterans with or without prevalent chronic health conditions. However, a greater proportion of veterans with (compared to without) prevalent mental health conditions reported the following considerations to be "very important": seeing information about DHTs on social media (those with mental health conditions: 42/428, 9.8%; those without mental health conditions: 19/328, 5.8%; χ^2_2=6.2; *P*=.05); having community support through Veteran Service Organizations, churches, libraries, or other organizations to use DHTs (with: 56/427, 13.1%; without: 25/327, 7.6%; χ^2_2=7.9; *P*=.02); having other veterans encourage DHT use (with: 70/429, 16.3%; without: 24/329, 7.3%; χ^2_2=15.9; *P*<.001); and having help from family, friends, or other important people to use DHTs (with: 106/431, 24.6%; without: 58/328, 17.7%; χ^2_2=6.4; *P*=.04). Please see Table 3 for more details.

XSL•FO
RenderX

Table 3. Perceptions of importance regarding factors that could impact decisions about the use of digital health technologies[a] among veteran respondents to the Veterans Engagement with Technology Collaborative cohort survey and comparisons of perceptions among veterans with and without prevalent chronic health conditions and prevalent mental health conditions (n=857).

Factor	Overall, n/N (%)	Respondent has ≥1 chronic health conditions[b] n/N (%)			Respondent has ≥1 mental health conditions[c] n/N (%)		
		Yes (713/857)	No (144/857)	*P*	Yes (481/857)	No (376/857)	*P*
Receiving MyHealtheVet secure messages from my VHA[d] health care team about them				.09			.26
Very important	561/766 (73.2)	474/641 (73.9)	87/125 (69.6)		325/435 (74.7)	236/331 (71.3)	
Somewhat important	158/766 (20.6)	133/641 (20.7)	25/125 (20)		81/435 (18.6)	77/331 (23.3)	
Not at all important	47/766 (6.1)	34/641 (5.3)	13/125 (10.4)		29/435 (6.7)	18/331 (5.4)	
Knowing that my VHA health care team members will use the information that I enter into them to inform my health care				.56			.19
Very important	427/756 (56.5)	355/631 (56.3)	72/125 (57.6)		254/428 (59.3)	173/328 (52.7)	
Somewhat important	248/756 (32.8)	205/631 (32.5)	43/125 (34.4)		132/428 (30.8)	116/328 (35.4)	
Not at all important	81/756 (10.7)	71/631 (11.3)	10/125 (8)		42/428 (9.8)	39/328 (11.9)	
Having my VHA doctor(s) recommend them to me				.24			.88
Very important	402/764 (52.6)	344/638 (53.9)	58/126 (46)		230/433 (53.1)	172/331 (52)	
Somewhat important	274/764 (35.9)	221/638 (34.6)	53/126 (42.1)		152/433 (35.1)	122/331 (36.9)	
Not at all important	88/764 (11.5)	73/638 (11.4)	15/126 (11.9)		51/433 (11.8)	37/331 (11.2)	
Having somewhere to get help if I am having issues using them				.82			.18
Very important	389/760 (51.2)	322/634 (50.8)	67/126 (53.2)		232/431 (53.8)	157/329 (47.7)	
Somewhat important	262/760 (34.5)	219/634 (34.5)	43/126 (34.1)		137/431 (31.8)	125/329 (38)	
Not at all important	109/760 (14.3)	93/634 (14.7)	16/126 (12.7)		62/431 (14.4)	47/329 (14.3)	
Having other members of my VHA clinical team recommend them to me				.11			.39
Very important	345/760 (45.4)	297/635 (46.8)	48/125 (38.4)		203/429 (47.3)	142/331 (42.9)	
Somewhat important	307/760 (40.4)	246/635 (38.7)	61/125 (48.8)		170/429 (39.6)	137/331 (41.4)	
Not at all important	108/760 (14.2)	92/635 (14.5)	16/125 (12.8)		56/429 (13.1)	52/331 (15.7)	
Having a video tutorial about how to use them				.57			.81
Very important	255/758 (33.6)	208/633 (32.9)	47/125 (37.6)		146/429 (34)	109/329 (33.1)	
Somewhat important	315/758 (41.6)	265/633 (41.9)	50/125 (40)		174/429 (40.6)	141/329 (42.9)	
Not at all important	188/758 (24.8)	160/633 (25.3)	28/125 (22.4)		109/429 (25.4)	79/329 (24)	
Having someone available to help me learn how to use them through in-person trainings				.43			.40
Very important	241/758 (31.8)	202/633 (31.9)	39/125 (31.2)		144/429 (33.6)	97/329 (29.5)	
Somewhat important	273/758 (36)	233/633 (36.8)	40/125 (32)		147/429 (34.3)	126/329 (38.3)	
Not at all important	244/758 (32.2)	198/633 (31.3)	46/125 (36.8)		138/429 (32.2)	106/329 (32.2)	
Receiving printed information about them from the VHA				.58			.37
Very important	231/758 (30.5)	197/632 (31.2)	34/126 (27)		139/427 (32.6)	92/331 (27.8)	
Somewhat important	326/758 (43)	271/632 (42.9)	55/126 (43.7)		179/427 (41.9)	147/331 (44.4)	
Not at all important	201/758 (26.5)	164/632 (25.9)	37/126 (29.4)		109/427 (25.5)	92/331 (27.8)	
Having help from family, friends, or other important people in my life to use them				.24			.04
Very important	164/759 (21.6)	143/633 (22.6)	21/126 (16.7)		106/431 (24.6)	58/328 (17.7)	
Somewhat important	279/759 (36.8)	226/633 (35.7)	53/126 (42.1)		146/431 (33.9)	133/328 (40.5)	
Not at all important	316/759 (41.6)	264/633 (41.7)	52/126 (41.3)		179/431 (41.5)	137/328 (41.8)	
Having other veterans encourage me to use them				.69			<.001

XSL•FO
RenderX

Factor	Overall, n/N (%)	Respondent has ≥1 chronic health conditions[b] n/N (%)			Respondent has ≥1 mental health conditions[c] n/N (%)		
		Yes (713/857)	No (144/857)	*P*	Yes (481/857)	No (376/857)	*P*
Very important	94/758 (12.4)	77/634 (12.1)	17/124 (13.7)		70/429 (16.3)	24/329 (7.3)	
Somewhat important	234/758 (30.9)	193/634 (30.4)	41/124 (33.1)		135/429 (31.5)	99/329 (30.1)	
Not at all important	430/758 (56.7)	364/634 (57.4)	66/124 (53.2)		224/429 (52.2)	206/329 (62.6)	
Having support in the community through my VSO[e], church, library, or other such organizations for using them				.37			.02
Very important	81/754 (10.7)	64/630 (10.2)	17/124 (13.7)		56/427 (13.1)	25/327 (7.6)	
Somewhat important	172/754 (22.8)	148/630 (23.5)	24/124 (19.4)		103/427 (24.1)	69/327 (21.1)	
Not at all important	501/754 (66.4)	418/630 (66.3)	83/124 (66.9)		268/427 (62.8)	233/327 (71.3)	
Seeing information about them on social media				.52			.05
Very important	61/756 (8.1)	48/631 (7.6)	13/125 (10.4)		42/428 (9.8)	19/328 (5.8)	
Somewhat important	162/756 (21.4)	134/631 (21.2)	28/125 (22.4)		98/428 (22.9)	64/328 (19.5)	
Not at all important	533/756 (70.5)	449/631 (71.2)	84/125 (67.2)		288/428 (67.3)	245/328 (74.7)	
Other				.67[f]			.34[f]
Very important	12/76 (15.8)	10/68 (14.7)	2/8 (25)		8/40 (20)	4/36 (11.1)	
Somewhat important	7/76 (9.2)	7/68 (10.3)	0/8 (0)		5/40 (12.5)	2/36 (5.6)	
Not at all important	57/76 (75)	51/68 (75)	6/8 (75)		27/40 (67.5)	30/36 (83.3)	

[a]VHA mobile health apps, non-VHA mobile health apps, or devices.

[b]Hypertension, diabetes, or chronic kidney disease.

[c]Anxiety, depression, or posttraumatic stress disorder.

[d]VHA: Veterans Health Administration.

[e]VSO: Veteran Service Organization.

[f]Fisher exact test *P* value.

Discussion

Principal Findings

Understanding what matters most to patients when deciding to use DHTs can inform the development of strategies to support their adoption and use. For veterans who receive VHA care and are known technology users, our results suggest that the most important considerations are tied to the health care system and include receiving encouragement from health care team members to use DHTs, knowing DHT data will be used in care, and having help available when using DHTs. Conversely, our results suggest that the least important considerations for veterans who use VHA care and are known technology users when deciding to use DHTs are rooted outside the health care system, including social media and encouragement and support from peers or other community members, although the veterans in our sample with prevalent mental health conditions did place more importance on these considerations. These insights can be leveraged to promote different technology platforms as part of health care services and inform efforts to ensure that implementation strategies deployed align with patient preferences and values. They also reflect the relative importance in this context of different theoretical constructs posited to influence the adoption of new technologies, including sources of social influence and the perceptions of others deemed important to an individual [18,19].

From our findings, outreach and recommendations from care team members were most frequently endorsed as very important to decisions about DHT use. These findings are consistent with previous literature that highlights the importance of health care team member encouragement to veteran use of technologies for health [12,14,35,36], and similar findings have also emerged in other published studies of other patient populations [37,38]. This trend suggests that implementation strategies intended to bolster veteran DHT adoption and use must reflect care team member involvement to be optimally impactful. This could include emphasizing care team member endorsement of specific technologies and their potential value to veterans or affirmation that care team members will use the data that veterans gather and share through these technologies to inform clinical practice, a consideration that has also been highlighted as important in literature about nonveteran populations [39].

Uniquely, our findings highlight different approaches that can be useful for outreach and recommendations and how such approaches may have differential impacts among segments of patient populations. In this sample of veterans who are known technology users, receiving secure messages about DHTs through the VHA's web-based patient portal was the

XSL•FO
RenderX

consideration most frequently endorsed as very important when deciding to use DHTs. The VHA already uses its web-based patient portal to disseminate health promotion information, including information about new technologies [12]. Using the messaging feature of a web-based patient portal may be an efficient, low-cost way for health care teams to proactively communicate with patients about technologies they think could be helpful to them.

Beyond health care team–related factors, the availability of technical support was the second most frequent consideration endorsed as very important when deciding to use DHTs. Previously published reviews have also identified difficulties using technology and tech support availability as factors influencing patient use [40,41]. Although our survey did not ask veterans whether they thought such support should be offered by the VHA, we see such help and support as services that the VHA could provide to address this need. Such services are evident in recent VHA initiatives like Veteran Health Resource Centers, which provide hands-on support within VHA facilities to veterans, their informal caregivers, and staff who are interested in using VHA technology platforms [42].

Unlike previously published studies suggesting that leveraging communities, peers, and social media may facilitate the adoption of technology for health in certain contexts [43-45], the veterans in our sample less frequently endorsed peer and community support and information shared through social media as important considerations. Although peer-, community-, and social media–based resources may represent efficient and low-cost approaches to facilitating DHT use, health care systems should work to understand patient preferences and perspectives before relying too heavily on such approaches, which may be best suited for use among certain segments of their patient populations. Patients with mental health conditions may be one such example, where strategies grounded in community and peer support are well-positioned to enhance reach and engagement with DHTs.

Limitations

The veterans in the VET-C cohort are known technology users. As such, our results may not generalize to the overall veteran population or other patient populations. This sample characteristic may also explain why receiving secure messages from one's health care team through the VHA's web-based patient portal was deemed such an important consideration. However, this finding parallels other findings regarding the importance that veterans place on health care team member recommendations when deciding whether to use DHTs [14,35]. Moreover, data from our sample of known technology users may not address considerations important to veterans who minimally or do not use technology when deciding to use DHTs. We must guard against overgeneralizing these findings so as not to inadvertently overlook the needs of patients for whom accessibility of care, and by extension DHTs, are formidable barriers. These groups may require alternative implementation strategies that more adequately reflect their situation [46,47]. Potentially high-impact future work could include replicating a survey similar to the one reported here with subsets of the veteran population who lack both ready access to DHTs and health care services. Cross-sectional survey data may also be influenced by recall, response, and social desirability biases. Finally, we cannot infer how our results may translate to actual technology adoption and use. Additional work is needed to develop and rigorously test implementation strategies to provide such insights.

Conclusions

Understanding what matters most to patients when deciding to use DHTs can inform strategies to enhance DHT use. This analysis represents an additional step towards foregrounding the perspective of patients as key stakeholders when designing implementation strategies to bolster adoption of health-related technology. Recommendations from health care team members and knowing their data will be used by their care team members in clinical care are more important considerations for veterans who use technology for health than information they may encounter through social media, community sources, or peers when deciding whether to adopt DHTs, although perceptions of importance regarding the latter may differ among patients with different conditions. Taken together, our results suggest that communication from health care team members to patients, perhaps either in-person or electronically, may be beneficial in many cases to promote DHT adoption and use. Future work should also examine these issues among veterans who may have a strong interest in using DHTs but are not yet DHT users and face various barriers to their adoption.

Acknowledgments

This work was supported by the US Department of Veterans Affairs, Office of Connected Care, and Office of Research and Development, Health Services Research and Development Service, Quality Enhancement Research Initiative Program (grant PEC 15-470 to TPH). JML was supported by an NIMH Mentored Patient-Oriented Career Development Award (grant K23MH120324). The views expressed in this article are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs or the United States Government.

Data Availability

The data used for this analysis are not permitted to leave the VHA firewall without a data use agreement. This limitation is consistent with other work based on VHA data. However, VHA data are made freely available to investigators behind the Department of Veterans Affairs' firewall with appropriate documentation.

XSL•FO
RenderX

Conflicts of Interest

None declared.

References

1. Xu W, Liu Y. mHealthApps: a repository and database of mobile health apps. JMIR Mhealth Uhealth 2015 Mar 18;3(1):e28 [FREE Full text] [doi: 10.2196/mhealth.4026] [Medline: 25786060]
2. Vaidya A. TechTarget. 2022. URL: https://www.techtarget.com/virtualhealthcare/feature/Key-Features-of-mHealth-Apps-Trends-in-Use [accessed 2023-12-14]
3. Jake-Schoffman DE, Silfee VJ, Waring ME, Boudreaux ED, Sadasivam RS, Mullen SP, et al. Methods for evaluating the content, usability, and efficacy of commercial mobile health apps. JMIR Mhealth Uhealth 2017 Dec 18;5(12):e190 [FREE Full text] [doi: 10.2196/mhealth.8758] [Medline: 29254914]
4. Boulos MNK, Brewer AC, Karimkhani C, Buller DB, Dellavalle RP. Mobile medical and health apps: state of the art, concerns, regulatory control and certification. Online J Public Health Inform 2014;5(3):229 [FREE Full text] [doi: 10.5210/ojphi.v5i3.4814] [Medline: 24683442]
5. Villinger K, Wahl DR, Boeing H, Schupp HT, Renner B. The effectiveness of app-based mobile interventions on nutrition behaviours and nutrition-related health outcomes: a systematic review and meta-analysis. Obes Rev 2019 Oct;20(10):1465-1484 [FREE Full text] [doi: 10.1111/obr.12903] [Medline: 31353783]
6. Scarry A, Rice J, O'Connor EM, Tierney AC. Usage of mobile applications or mobile health technology to improve diet quality in adults. Nutrients 2022 Jun 12;14(12):2437 [FREE Full text] [doi: 10.3390/nu14122437] [Medline: 35745167]
7. Serrano-Ripoll MJ, Zamanillo-Campos R, Fiol-DeRoque MA, Castro A, Ricci-Cabello I. Impact of smartphone app-based psychological interventions for reducing depressive symptoms in people with depression: systematic literature review and meta-analysis of randomized controlled trials. JMIR Mhealth Uhealth 2022 Jan 27;10(1):e29621 [FREE Full text] [doi: 10.2196/29621] [Medline: 35084346]
8. Treadwell JR, Reston JT, Rouse B, Fontanarosa J, Patel N, Mull NK. Automated-Entry Patient-Generated Health Data for Chronic Conditions: The Evidence on Health Outcomes. Rockville, MD: Agency for Healthcare Research and Quality; 2021.
9. Triantafyllidis A, Kondylakis H, Votis K, Tzovaras D, Maglaveras N, Rahimi K. Features, outcomes, and challenges in mobile health interventions for patients living with chronic diseases: a review of systematic reviews. Int J Med Inform 2019 Dec;132:103984. [doi: 10.1016/j.ijmedinf.2019.103984] [Medline: 31605884]
10. Holko M, Litwin TR, Munoz F, Theisz KI, Salgin L, Jenks NP, et al. Wearable fitness tracker use in federally qualified health center patients: strategies to improve the health of all of us using digital health devices. NPJ Digit Med 2022 Apr 25;5(1):53 [FREE Full text] [doi: 10.1038/s41746-022-00593-x] [Medline: 35469045]
11. Krebs P, Duncan DT. Health app use among US mobile phone owners: a national survey. JMIR Mhealth Uhealth 2015 Nov 04;3(4):e101 [FREE Full text] [doi: 10.2196/mhealth.4924] [Medline: 26537656]
12. Hogan TP, Etingen B, Lipschitz JM, Shimada SL, McMahon N, Bolivar D, et al. Factors associated with self-reported use of web and mobile health apps among US military veterans: cross-sectional survey. JMIR Mhealth Uhealth 2022 Dec 30;10(12):e41767 [FREE Full text] [doi: 10.2196/41767] [Medline: 36583935]
13. Reger GM, Harned M, Stevens ES, Porter S, Nguyen J, Norr AM. Mobile applications may be the future of veteran mental health support but do veterans know yet? A survey of app knowledge and use. Psychol Serv 2022 Aug;19(3):480-487. [doi: 10.1037/ser0000562] [Medline: 34081527]
14. Lipschitz J, Miller CJ, Hogan TP, Burdick KE, Lippin-Foster R, Simon SR, et al. Adoption of mobile apps for depression and anxiety: cross-sectional survey study on patient interest and barriers to engagement. JMIR Ment Health 2019 Jan 25;6(1):e11334 [FREE Full text] [doi: 10.2196/11334] [Medline: 30681968]
15. Erbes CR, Stinson R, Kuhn E, Polusny M, Urban J, Hoffman J, et al. Access, utilization, and interest in mHealth applications among veterans receiving outpatient care for PTSD. Mil Med 2014 Nov;179(11):1218-1222. [doi: 10.7205/MILMED-D-14-00014] [Medline: 25373044]
16. Yeager CM, Benight CC. If we build it, will they come? Issues of engagement with digital health interventions for trauma recovery. Mhealth 2018;4:37 [FREE Full text] [doi: 10.21037/mhealth.2018.08.04] [Medline: 30363749]
17. Birnbaum F, Lewis D, Rosen RK, Ranney ML. Patient engagement and the design of digital health. Acad Emerg Med 2015 Jun;22(6):754-756 [FREE Full text] [doi: 10.1111/acem.12692] [Medline: 25997375]
18. Dearing JW, Cox JG. Diffusion Of innovations theory, principles, and practice. Health Aff (Millwood) 2018 Feb;37(2):183-190. [doi: 10.1377/hlthaff.2017.1104] [Medline: 29401011]
19. Venkatesh V, Morris M, Davis G, Davis F. User acceptance of information technology: toward a unified view. MIS Quarterly 2003;27(3):425-478. [doi: 10.2307/30036540]
20. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly 1989 Sep;13(3):319. [doi: 10.2307/249008]

XSL•FO
RenderX

21. Powell BJ, Beidas RS, Lewis CC, Aarons GA, McMillen JC, Proctor EK, et al. Methods to improve the selection and tailoring of implementation strategies. J Behav Health Serv Res 2017 Apr;44(2):177-194 [FREE Full text] [doi: 10.1007/s11414-015-9475-6] [Medline: 26289563]
22. Powell BJ, McMillen JC, Proctor EK, Carpenter CR, Griffey RT, Bunger AC, et al. A compilation of strategies for implementing clinical innovations in health and mental health. Med Care Res Rev 2012 Apr;69(2):123-157 [FREE Full text] [doi: 10.1177/1077558711430690] [Medline: 22203646]
23. Knapp AA, Carroll AJ, Mohanty N, Fu E, Powell BJ, Hamilton A, et al. A stakeholder-driven method for selecting implementation strategies: a case example of pediatric hypertension clinical practice guideline implementation. Implement Sci Commun 2022 Mar 07;3(1):25 [FREE Full text] [doi: 10.1186/s43058-022-00276-4] [Medline: 35256017]
24. van Rooijen M, Lenzen S, Dalemans R, Beurskens A, Moser A. Stakeholder engagement from problem analysis to implementation strategies for a patient-reported experience measure in disability care: A qualitative study on the process and experiences. Health Expect 2021 Feb;24(1):53-65 [FREE Full text] [doi: 10.1111/hex.13147] [Medline: 33125177]
25. Powell BJ, Waltz TJ, Chinman MJ, Damschroder LJ, Smith JL, Matthieu MM, et al. A refined compilation of implementation strategies: results from the Expert Recommendations for Implementing Change (ERIC) project. Implement Sci 2015 Feb 12;10:21 [FREE Full text] [doi: 10.1186/s13012-015-0209-1] [Medline: 25889199]
26. Etingen B, Amante DJ, Martinez RN, Smith BM, Shimada SL, Richardson L, et al. Supporting the implementation of connected care technologies in the veterans health administration: cross-sectional survey findings from the Veterans Engagement with Technology Collaborative (VET-C) cohort. J Particip Med 2020 Sep 30;12(3):e21214 [FREE Full text] [doi: 10.2196/21214] [Medline: 33044944]
27. Martinez RN, Smith BM, Etingen B, Houston TK, Shimada SL, Amante DJ, et al. Health-related goal setting and achievement among veterans with high technology adoption. J Gen Intern Med 2021 Nov;36(11):3337-3345 [FREE Full text] [doi: 10.1007/s11606-021-06779-5] [Medline: 33963510]
28. Hogan TP, Etingen B, Lipschitz JM, Shimada SL, McMahon N, Bolivar D, et al. Factors associated with self-reported use of web and mobile health apps among US military veterans: cross-sectional survey. JMIR Mhealth Uhealth 2022 Dec 30;10(12):e41767 [FREE Full text] [doi: 10.2196/41767] [Medline: 36583935]
29. 36-item short form survey (SF-36). RAND Health Care. URL: https://www.rand.org/health-care/surveys_tools/mos/36-item-short-form.html [accessed 2022-11-19]
30. Cho AH, Arar NH, Edelman DE, Hartwell PH, Oddone EZ, Yancy WS. Do diabetic veterans use the Internet? Self-reported usage, skills, and interest in using My HealtheVet Web portal. Telemed J E Health 2010 Jun;16(5):595-602. [doi: 10.1089/tmj.2009.0164] [Medline: 20575727]
31. Pope GC, Kautter J, Ellis RP, Ash AS, Ayanian JZ, Lezzoni LI, et al. Risk adjustment of Medicare capitation payments using the CMS-HCC model. Health Care Financ Rev 2004;25(4):119-141 [FREE Full text] [Medline: 15493448]
32. Morgan RO, Petersen LA, Hasche JC, Davila JA, Byrne MM, Osemene NI, et al. VHA pharmacy use in veterans with Medicare drug coverage. Am J Manag Care 2009 Mar 16;15(3):e1-e8 [FREE Full text] [Medline: 19298095]
33. Yeatts JP, Sangvai D. HCC coding, risk adjustment, and physician income: what you need to know. Fam Pract Manag 2016;23(5):24-27 [FREE Full text] [Medline: 27626116]
34. Program guide 1200.21: VHA operations activities that may constitute research. Department of Veterans Affairs. URL: https://www.research.va.gov/resources/policies/ProgramGuide-1200-21-VHA-Operations-Activities.pdf [accessed 2025-07-23]
35. Hogan TP, Etingen B, McMahon N, Bixler FR, Am L, Wacks RE, et al. Understanding adoption and preliminary effectiveness of a mobile app for chronic pain management among US military veterans: pre-post mixed methods evaluation. JMIR Form Res 2022 Jan 20;6(1):e33716 [FREE Full text] [doi: 10.2196/33716] [Medline: 35049515]
36. Saleem JJ, Read JM, Loehr BM, Frisbee KL, Wilck NR, Murphy JJ, et al. Veterans' response to an automated text messaging protocol during the COVID-19 pandemic. J Am Med Inform Assoc 2020 Aug 01;27(8):1300-1305 [FREE Full text] [doi: 10.1093/jamia/ocaa122] [Medline: 32470974]
37. Cajita MI, Hodgson NA, Lam KW, Yoo S, Han H. Facilitators of and barriers to mHealth adoption in older adults with heart failure. Comput Inform Nurs 2018 Aug;36(8):376-382 [FREE Full text] [doi: 10.1097/CIN.0000000000000442] [Medline: 29742549]
38. Steeb T, Wessely A, Mastnik S, Brinker TJ, French LE, Niesert A, et al. Patient attitudes and their awareness towards skin cancer-related apps: cross-sectional survey. JMIR Mhealth Uhealth 2019 Jul 02;7(7):e13844 [FREE Full text] [doi: 10.2196/13844] [Medline: 31267978]
39. Kessel KA, Vogel MM, Kessel C, Bier H, Biedermann T, Friess H, et al. Mobile health in oncology: a patient survey about app-assisted cancer care. JMIR Mhealth Uhealth 2017 Jun 14;5(6):e81 [FREE Full text] [doi: 10.2196/mhealth.7689] [Medline: 28615159]
40. Palacholla RS, Fischer N, Coleman A, Agboola S, Kirley K, Felsted J, et al. Provider- and patient-related barriers to and facilitators of digital health technology adoption for hypertension management: scoping review. JMIR Cardio 2019 Mar 26;3(1):e11951 [FREE Full text] [doi: 10.2196/11951] [Medline: 31758771]

XSL•FO
RenderX

41. Borghouts J, Eikey E, Mark G, De Leon C, Schueller SM, Schneider M, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. J Med Internet Res 2021 Mar 24;23(3):e24387 [FREE Full text] [doi: 10.2196/24387] [Medline: 33759801]
42. Virtual health resource Ccnters. US Department of Veterans Affairs. URL: https://connectedcare.va.gov/virtual-health-resource-centers [accessed 2023-12-14]
43. Yingling LR, Brooks AT, Wallen GR, Peters-Lawrence M, McClurkin M, Cooper-McCann R, et al. Community engagement to optimize the use of web-based and wearable technology in a cardiovascular health and needs assessment study: a mixed methods approach. JMIR Mhealth Uhealth 2016 Apr 25;4(2):e38 [FREE Full text] [doi: 10.2196/mhealth.4489] [Medline: 27113680]
44. Newton RL, Carter LA, Johnson W, Zhang D, Larrivee S, Kennedy BM, et al. A church-based weight loss intervention in African American adults using text messages (LEAN study): cluster randomized controlled trial. J Med Internet Res 2018 Aug 24;20(8):e256 [FREE Full text] [doi: 10.2196/jmir.9816] [Medline: 30143478]
45. Schueller SM, Neary M, O'Loughlin K, Adkins EC. Discovery of and interest in health apps among those with mental health needs: survey and focus group study. J Med Internet Res 2018 Jun 11;20(6):e10141 [FREE Full text] [doi: 10.2196/10141] [Medline: 29891468]
46. Veinot TC, Mitchell H, Ancker JS. Good intentions are not enough: how informatics interventions can worsen inequality. J Am Med Inform Assoc 2018 Aug 01;25(8):1080-1088 [FREE Full text] [doi: 10.1093/jamia/ocy052] [Medline: 29788380]
47. Yao R, Zhang W, Evans R, Cao G, Rui T, Shen L. Inequities in health care services saused by the adoption of digital health technologies: scoping review. J Med Internet Res 2022 Mar 21;24(3):e34144 [FREE Full text] [doi: 10.2196/34144] [Medline: 35311682]

Abbreviations

DHT: digital health technology
VET-C: Veterans Engagement with Technology Collaborative
VHA: Veterans Health Administration

Edited by A Mavragani; submitted 08.05.25; peer-reviewed by N Wolkowicz, RN Al-Jabali; comments to author 02.06.25; revised version received 23.06.25; accepted 09.07.25; published 07.08.25.

Please cite as:
Etingen B, Smith BM, Shimada SL, Robinson SA, Higashi RT, Ndiwane N, Frisbee KL, Lipschitz JM, Richardson E, Irvin D, Hogan TP
What Matters Most to Veterans When Deciding to Use Technology for Health: Cross-Sectional Analysis of a National Survey
JMIR Form Res 2025;9:e77113
URL: https://formative.jmir.org/2025/1/e77113
doi:10.2196/77113
PMID:

©Bella Etingen, Bridget M Smith, Stephanie L Shimada, Stephanie A Robinson, Robin T Higashi, Ndindam Ndiwane, Kathleen L Frisbee, Jessica M Lipschitz, Eric Richardson, Dawn Irvin, Timothy P Hogan. Originally published in JMIR Formative Research (https://formative.jmir.org), 07.08.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Data Mining–Based Model for Computer-Aided Diagnosis of Autism and Gelotophobia: Mixed Methods Deep Learning Approach

Mohamed Eldawansy[1*], MSc; Hazem El Bakry[1*], PhD; Samaa M Shohieb[1*], Phd

Information Systems, Faculty of Computers and Information, Mansoura University, Mansoura, Egypt

[*]all authors contributed equally

Corresponding Author:
Samaa M Shohieb, Phd
Information Systems
Faculty of Computers and Information
Mansoura University
1 Elgomhoria St, El Mansoura Daqahlia
Mansoura, 35511
Egypt
Phone: 20 01142558595
Email: sm_shohieb@mans.edu.eg

Abstract

Background: Gelotophobia, the fear of being laughed at, is a social anxiety condition that affects approximately 6% of neurotypical individuals and up to 45% of those with autism spectrum disorder (ASD). This comorbidity can significantly impair the quality of life, particularly in adolescents with high-functioning ASD, where the prevalence reaches 41.98%. Accurate and automated detection tools could enhance early diagnosis and intervention.

Objective: This study aimed to develop a deep learning–based diagnostic system that integrates facial emotion recognition with validated questionnaires to detect gelotophobia in individuals with or without ASD.

Methods: The system was trained to identify ASD status using a balanced dataset of 2932 facial images (n=1466; 50% from individuals with ASD and n=1466; 50% from neurotypical individuals). The images were processed using the DeepFace library to extract facial features, which were then used as input for the deep learning classifier. After identifying ASD status, the same images were further analyzed using the pretrained DeepFace model to evaluate facial expressions for signs of gelotophobia. In cases where facial cues were ambiguous, the GELOPH<15> questionnaire, consisting of 15 items, was administered to confirm the diagnosis The system was fully implemented using the Python programming language. Deep learning models were developed using libraries such as PyTorch for training the multilayer perceptron classifier, while CUDA was used to accelerate computations on compatible graphics processing units. Additional libraries from the Python programming language, such as *scikit-learn*, *NumPy*, and *Pandas*, were used for preprocessing, model evaluation, and data manipulation. DeepFace was integrated using its Python application programming interface for facial recognition and emotion classification.

Results: The dataset comprised 2932 facial images collected from platforms such as Kaggle and ASD-related websites, including 1466 (50%) images of children with ASD and 1466 (50%) images of neurotypical children. The dataset was split into 2653 (90.48%) training samples and 279 (9.51%) testing samples, with each image contributing 100,352 extracted features. We applied various machine learning models for ASD identification. The system achieved an overall prediction accuracy of 92% across both training and testing datasets, with the multilayer perceptron model demonstrating the highest testing accuracy. The system successfully classified gelotophobia in cases where facial expressions were clear. However, in cases of ambiguous facial cues, the DeepFace model alone was insufficient. Incorporating the GELOPH<15> questionnaire improved diagnostic reliability and consistency.

Conclusions: This study demonstrates the effectiveness of combining deep learning techniques with validated diagnostic tools for detecting gelotophobia, particularly in individuals with ASD. The high accuracy achieved highlights the system's potential for clinical and research applications, contributing to the improved understanding and management of gelotophobia among groups considered socially vulnerable. Future research could expand the system's applications to broader psychological assessments.

XSL•FO
RenderX

(JMIR Form Res 2025;9:e72115) doi:10.2196/72115

KEYWORDS

social anxiety; emotion recognition; machine learning; facial expression analysis; neurodevelopmental disorders; GELOPH<15>; residual network with 50 layers; ResNet-50; artificial intelligence

Introduction

Background

Individuals with gelotophobia hold the belief that they are being laughed at. Over time, this belief may become internalized, resulting in heightened sensitivity to rejection, feelings of shame, and an increased tendency toward social avoidance. These observations highlight the complex relationship among laughter, social perception, and emotional regulation, particularly for those who struggle with interpreting subtle social cues. Laughter and humor are widely regarded as positive social tools that enhance interpersonal relationships and help individuals navigate life's challenges, often evoking joy and connection. However, it is important to recognize that laughter can also carry negative connotations, particularly when used to mock or ridicule others [1]. For some individuals, distinguishing between different types of laughter is challenging. They may perceive all laughter as negatively directed at them, interpreting it as a form of mockery or threat, which often leads to social withdrawal and isolation [2]. While many people can manage feelings such as embarrassment, anger, or discomfort in response to ridicule, individuals with social competence challenges frequently struggle to regulate these emotions. This difficulty may foster a persistent fear of social interactions, rooted in the belief that they are being mocked for perceived social flaws [3,4].

Autism Spectrum Disorder

Autism spectrum disorder (ASD) is a neurodevelopmental disorder that typically manifests in early childhood and persists throughout an individual's life. The term *spectrum* reflects the wide variation in the severity of symptoms and abilities among individuals with ASD. Symptoms commonly include difficulties in social interactions, communication impairments, repetitive behaviors, and restricted interests. These challenges can range from mild to severe, influencing the individual's ability to function in society [5]. ASD is diagnosed early, usually between 18 and 24 months, when signs such as delayed speech, limited social interactions, and unusual behavior patterns become apparent. In some cases, children may exhibit typical developmental patterns during the first year but later show regression, notably in social engagement and language skills [6]. According to a study by Harris [7], individuals with ASD often face difficulties in interpreting emotional expressions, which may hinder their ability to respond appropriately in social situations, leading to increased social withdrawal. Consequently, early diagnosis and intervention, including therapies such as applied behavior analysis, are crucial for improving outcomes. Recent advancements in data mining and machine learning are revolutionizing ASD diagnosis and management. Using deep learning models and facial expression recognition systems, the detection of subtle behavioral markers in individuals with ASD can be enhanced, facilitating earlier and more accurate diagnoses [8].

Gelotophobia: Fear of Being Laughed at

Gelotophobia, the fear of being laughed at, is often linked to social anxiety disorders and can profoundly impact an individual's psychological well-being. While laughter is generally perceived as a positive social signal, individuals with gelotophobia may interpret it as mockery or ridicule, even when no negative intent exists [9]. This heightened sensitivity to humor and social interactions is especially prevalent in individuals with ASD, who struggle to decode facial expressions and contextual cues in social environments [10]. Individuals with gelotophobia experience significant distress and may avoid social situations altogether to prevent perceived humiliation. This social withdrawal can exacerbate feelings of isolation and lead to further deterioration of mental health. The relationship between gelotophobia and autism is particularly significant because individuals with ASD often exhibit impaired ability to differentiate between various social cues, including laughter, making them more vulnerable to developing gelotophobia [11]. Research by Führ et al [12] suggests that approximately 45% of individuals with ASD are diagnosed with gelotophobia, highlighting the need for early detection and intervention. Unlike typical social anxieties, gelotophobia specifically affects the individual's ability to engage socially due to constant fear of being laughed at or mocked. This fear can be debilitating, leading to severe social withdrawal and heightened anxiety.

Link Between Gelotophobia and ASD

Gelotophobia is particularly prevalent in individuals with ASD due to their difficulties in understanding social cues and interpreting emotions. Studies show that individuals with ASD are more likely to misinterpret facial expressions, especially in humorous contexts, making them prone to developing gelotophobia [13]. For instance, they may interpret others' laughter as mockery even when not intended harmfully. This misinterpretation can trigger intense embarrassment, shame, and anxiety. The heightened prevalence of gelotophobia among individuals with ASD stems not only from challenges interpreting social cues but also from increased sensitivity to negative social stimuli. Research by Husseiny et al [14] indicates that up to 41.98% of adolescents with high-functioning autism experience moderate gelotophobia. This statistic underscores the importance of recognizing and addressing this comorbidity to improve individuals' quality of life.

Several studies have explored gelotophobia prevalence in individuals with ASD, indicating that it is a significant yet often overlooked aspect of autism. For example, research by Führ et al [12] found that individuals with ASD are more likely to develop gelotophobia due to limited ability to distinguish between benign and mocking laughter. These individuals, particularly with high-functioning autism, exhibited increased social anxiety, especially when perceiving themselves as ridicule targets. Ul Haque and Valles [15] emphasize the role of social

XSL•FO
RenderX

context in gelotophobia development, highlighting heightened sensitivity to social interaction nuances. Negative childhood experiences, such as bullying or exclusion, may exacerbate this sensitivity, leading gelotophobia to develop as a defense mechanism.

Research Gap and Study Objectives

While previous studies have examined the prevalence of gelotophobia in individuals with ASD and their challenges in interpreting social cues, objective and automated tools that integrate facial expression analysis with psychometric assessments remain lacking, limiting opportunities to improve early diagnosis and intervention. Current diagnostic methods often rely on subjective clinical observations, delaying identification and treatment. This study addresses the gap by proposing a novel deep learning–based model combining facial recognition (using DeepFace; Facebook AI Research) with the GELOPH<15> questionnaire, aiming to enhance accuracy and efficiency in detecting gelotophobia in children with ASD. This integrated approach offers a more objective and scalable solution to support clinical decision-making and improve patient outcomes.

Despite the known challenges of diagnosing gelotophobia and ASD via traditional clinical assessments and psychometric questionnaires, these methods rely heavily on subjective interpretation, often delaying timely diagnosis and intervention. Our study proposes an integrated approach combining DeepFace, a state-of-the-art facial recognition system that objectively analyzes subtle facial expressions, with the GELOPH<15> questionnaire, which captures social behaviors related to gelotophobia. This integration enables a comprehensive evaluation by leveraging behavioral self-reports and automated facial analysis.

Specifically, outputs from DeepFace, which detects nuanced facial cues indicative of gelotophobia and ASD traits, are used alongside questionnaire results to improve diagnostic accuracy. This method functions synergistically, enabling robust screening and early detection.

By applying a deep learning model with carefully selected hyperparameters, the study aimed to optimize prediction performance while ensuring generalizability across diverse populations. This artificial intelligence (AI)–powered diagnostic tool promises to reduce clinical burdens; provide objective assessment measures; and ultimately facilitate earlier, more effective interventions for children experiencing gelotophobia and ASD.

The Role of AI and Deep Learning in Early Diagnosis

With advances in AI and machine learning, early ASD diagnosis is increasingly feasible through facial pattern analysis. Neural networks detect facial expressions linked to neurological conditions. Data mining systems are now being developed to aid in the diagnosis of both gelotophobia and ASD. Ganesan et al [16] found that 41.98% of adolescents with high-functioning autism experienced moderate gelotophobia, with female individuals showing higher rates, potentially due to increased bullying exposure during adolescence. This highlights the ASD-gelotophobia connection, suggesting that individuals with ASD may be more vulnerable to fear of being laughed at.

Deep Learning Model for Detecting Gelotophobia and ASD

This paper introduces a deep learning model that analyzes facial features in children with ASD and typically developing children, aiming to achieve more accurate ASD identification. The model extracts facial characteristics and applies deep face analysis to predict gelotophobia. In addition, the GELOPH<15> questionnaire [17], consisting of 15 social behavior questions, has been developed as an interactive tool to detect gelotophobia.

Literature Review

Brauer and Proyer [1] applied a deep convolutional neural network (CNN) to analyze facial expressions in images. The Facial Expression Recognition 2013 dataset, consisting of thousands of labeled facial expression images, was used for training. The CNN was structured to recognize subtle differences in facial expressions, which can be challenging for children with autism to interpret. The high accuracy rate (89.2%) indicates that CNNs can effectively support the development of tools designed to teach facial recognition skills to children with autism. This has significant implications for educational and therapeutic applications, providing a technological supplement to traditional methods. Further research could involve real-time applications of this model in interactive settings, such as games or virtual reality environments, to enhance engagement and learning outcomes for children with autism.

Canestrari et al [2] compared multiple machine learning models using the Visual Geometry Group (VGG)-16 architecture, a deep CNN pretrained on a large image dataset. By adapting VGG-16 for ASD prediction, they tested models such as support vector machine, CNN, and Haar cascade classifiers. Achieving 90% accuracy suggests that deep learning models, particularly VGG-16, can be instrumental in early diagnosis and intervention planning for ASD. The comparative approach also helps identify which models are most effective in clinical settings. Future work might focus on integrating these models into diagnostic tools that clinicians can use during routine checkups, potentially combining image data with other behavioral indicators for a more comprehensive assessment. Similarly, Arru et al [18] explored visual behavior analysis for ASD identification, highlighting the role of computer vision techniques in supporting early diagnosis.

On the other hand, Eslami et al [17] proposed AD-Diagnet, which integrated an autoencoder for unsupervised feature learning with a single-layer perceptron for classification. The use of functional magnetic resonance imaging data from the Autism Brain Imaging Data Exchange dataset allowed for the exploration of neural activity patterns associated with ASD. With an 82% accuracy, this approach highlights the potential of combining neural network models for extracting meaningful features from complex brain imaging data. It underscores the role of advanced machine learning techniques in improving the accuracy of neurodevelopmental disorder diagnoses. Expanding the framework to include additional layers or integrating it with

XSL•FO
RenderX

other neural network architectures could further enhance accuracy. In addition, applying this model to other neuroimaging datasets could validate its generalizability and robustness.

The research by Leader et al [4] focused on eye-tracking data to assess visual attention in individuals with ASD. Unlike other studies that consider object semantics, this study specifically looked at the regions of the face that individuals with autism focus on during social interactions. The tendency of individuals with ASD to focus more on the mouth than the eyes provides crucial insights into the unique social processing characteristics of this population. This could inform the development of more targeted social skills training programs. Further studies could explore intervention methods aimed at encouraging eye contact in individuals with autism, as this may enhance social interaction and communication skills.

Moreover, Nunes et al [5] tracked eye movements in children during face recognition tasks, examining how different visual scanning patterns could be used to classify ASD. Machine learning models were developed to analyze these patterns. The findings, with an accuracy of 88.51%, suggest that eye movement patterns are a reliable marker for ASD. This could lead to the development of noninvasive, eye-tracking–based diagnostic tools that are easy to implement in various settings. Broader studies with larger, more diverse populations could enhance the generalizability of these findings. Moreover, integrating eye-tracking data with other behavioral and genetic data could improve the predictive power of ASD diagnostic tools.

Redfield et al [6] constructed a large-scale video dataset capturing social interactions in children to predict ASD. The analysis was conducted in 2 phases: detecting behaviors indicative of ASD and using these data for prediction through deep learning models. The use of real-world video data enhances the ecological validity of the findings, making the predictions more applicable to everyday settings. This study highlights the potential of video-based analysis for large-scale ASD screening. Future work should focus on enhancing deep learning models to accommodate diverse datasets, ensuring greater applicability across various cultural and demographic groups. In addition, developing real-time analysis tools that can be used in schools or clinics would be a significant advancement.

On the other hand, a study by Harris [7] reviewed various machine learning algorithms applied to diagnose ASD and learning disabilities using a dataset of infants' videos. The study emphasized behaviors such as gaze, smile, and vocalization. The ability to use simple behavioral markers for early diagnosis underscores the potential for early intervention, which can significantly improve developmental outcomes for children with ASD. Future studies could enhance the model's predictive accuracy by incorporating multimodal data, such as audio and physiological signals, alongside video analysis. Expanding the dataset to include a wider age range and more diverse populations would also be beneficial.

Alsaade and Alzahrani [8] proposed a transfer learning–based framework for facial recognition to improve autism detection. Transfer learning allowed the model to leverage preexisting knowledge from related tasks to enhance its performance. The framework's 91% accuracy indicates that transfer learning can significantly improve the detection of autism, especially in environments with limited data. This makes it a practical solution for early detection in domestic settings. Implementing this framework in mobile apps or smart home devices could provide parents and caregivers with accessible tools for monitoring developmental milestones and identifying early signs of autism.

Titze [9] used eye-tracking technology in virtual reality environments to distinguish between children with autism and typically developing children, focusing on how visual attention is allocated during social tasks. With an accuracy of up to 86%, the findings suggest that eye-tracking can be a powerful tool for diagnosing ASD. Furthermore, the use of virtual reality environments offers a controlled setting for assessing social attention in a way that mirrors real-world interactions. Scaling up the sample size and including external validation datasets could strengthen the reliability of the findings. Exploring the use of virtual reality headsets to make eye-tracking more immersive and accessible is another promising direction.

The study by Ruch [10] explored the prevalence of gelotophobia among adolescents with high-functioning ASD. Using psychometric assessments, the research measured the fear of being laughed at and analyzed demographic factors. The findings of a 41.98% prevalence rate highlight the importance of addressing emotional and psychological aspects in the treatment of ASD. This study also points to gender differences in the experience of gelotophobia, which could inform more tailored interventions. Further research could examine the underlying causes of gelotophobia in those with ASD and explore interventions that reduce social anxiety. Longitudinal studies tracking these adolescents into adulthood could provide insights into how gelotophobia evolves over time and its long-term impacts (Table 1).

XSL•FO
RenderX

Table 1. Summary of related work on autism spectrum disorder (ASD) detection using machine learning (ML) and other methods.

Study	Focus area	Dataset	Algorithm	Accuracy (%)	Age range	GELOPH<15> test
Ul Haque and Valles [15], 2018	Using smartphones or tablets to help children with autism recognize facial expressions	Kaggle Facial Expression Recognition 2013 dataset (web based)	DCNN[a]	89.2	Not specified	Not used
Ganesan [16], 2021	Classification of ASD in children using ML for facial analysis	Kaggle repository (web based)	VGG-16[b]	90	Not specified	Not used
Eslami et al [17], 2019	Developing ASD-DiagNet to classify ASD using fMRI[c] data	ABIDE[d] dataset (web based)	ASD-DiagNet	82	Not specified	Not used
Liu et al [19], 2016	Classifying ASD in children based on atypical face scanning patterns (eye movement data)	29 children diagnosed with ASD (local dataset)	SVM[e]	88.51	Not specified	Not used
Wu et al [20], 2021	Early ASD diagnosis using ML to analyze infant behaviors from video data	Interactive play session video dataset (web based)	Baseline ML classifier	82	Infants (0-3 y)	Not used
Akter et al [21] 2021 and Alcañiz et al [22], 2022	Transfer learning–based facial recognition framework for early ASD detection	Kaggle data repository (web based)	MobileNet-V1	92.1	Children (4-12 y)	Not used
Kats [23], 2022	Eye-tracking in virtual reality to distinguish ASD from typically developing children	Kaggle data repository (web based)	SVM	86	Children (3-12 y)	Not used
Husseiny et al [14], 2024	Examining gelotophobia prevalence and severity in adolescents with high-functioning ASD	Real-world data from Cairo Governorate, Egypt	EFA[f] and CFA[g]	76.88% variance explained, with excellent fit indices (χ^2=59.4 and RMSEA[h]=0.000)	Adolescents (13-18 y)	Used (after study)

[a]DCNN: deep convolutional neural network.
[b]VGG-16: Visual Geometry Group 16.
[c]fMRI: functional magnetic resonance imaging.
[d]ABIDE: Autism Brain Imaging Data Exchange.
[e]SVM: support vector machine.
[f]EFA: exploratory factor analysis.
[g]CFA: confirmatory factor analysis.
[h]RMSEA: root mean square error of approximation.

Study Contribution

This study developed a dual diagnostic framework aimed at the early detection of both ASD and gelotophobia. The proposed system achieved a high classification accuracy of 92% for ASD using a deep learning–based facial emotion recognition model (multilayer perceptron [MLP]). To improve the reliability of gelotophobia detection, especially in ambiguous cases, the framework integrated facial expression analysis using the DeepFace model with the GELOPH<15> psychometric questionnaire. This hybrid approach effectively combines objective facial analysis with subjective questionnaire evaluation, allowing simultaneous assessment of multiple psychological conditions. The study used a publicly available, balanced dataset of 2932 images, ensuring reproducibility and practical applicability. Furthermore, it offers a scalable and comprehensive tool to support early diagnosis and intervention planning for populations considered socially vulnerable. Finally, the framework lays a foundation for future research in expanding multimodal behavioral markers, such as voice patterns and body language.

Methods

Proposed Framework and Methods

This study proposed an integrated framework to predict gelotophobia in both individuals with and without autism using facial analysis and questionnaire data. The methodology combined deep learning DeepFace and residual network with

XSL•FO
RenderX

50 layers (ResNet-50), psychological assessment (GELOPH<15>), and a classification model (MLP). Figure 1 illustrates the overall workflow of the framework.

Figure 1. Workflow of the proposed gelotophobia prediction framework integrating DeepFace, GELOPH<15> questionnaire, and multilayer perceptron (MLP).

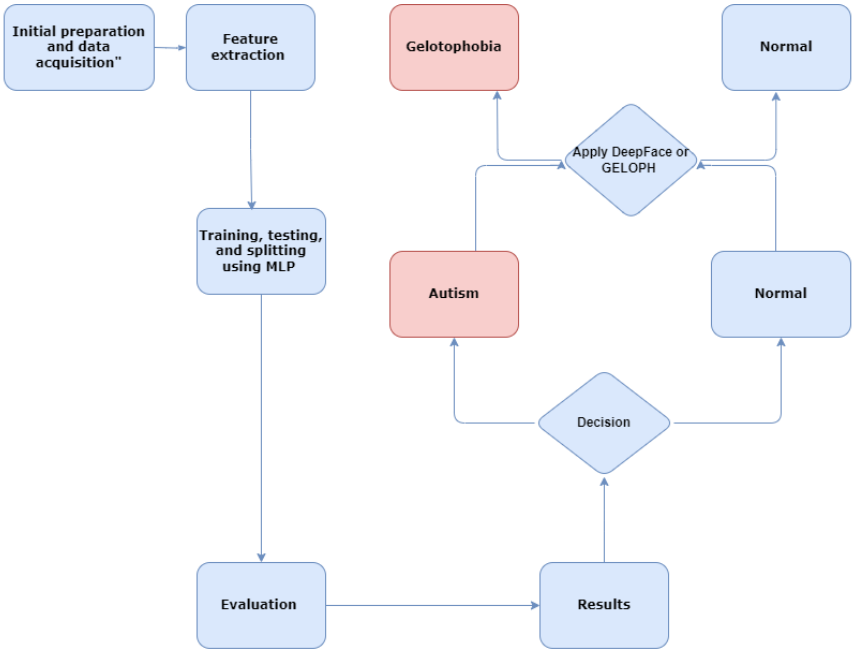

Dataset

This study examined facial photographs of children with ASD and typical children taken from the publicly available web-based Kaggle platform [11]. The dataset comprised 2932 facial photographs: half (n=1466, 50%) of the photographs showed children without ASD, while the other half (n=1466, 50%) depicted children with autism. This information was gathered from online resources, including ASD-related web pages and Facebook (Meta Platforms, Inc) pages. The training data contained 2653 (90.48%) records, each with 100,352 features (float64), and the test data had 279 (9.51%) records. The shape of the target vectors suggested a binary classification problem, as there were 2 distinct class types in both training and test sets.

Initial Preparation

The photographs were cleaned and cropped as part of the data preparation. The data had to be preprocessed before they could be used to train the deep learning model, as Piosenka gathered them from web-based sites. The face in the original image was automatically cropped by the dataset author. The dataset of 2932 photographs was divided into 279 (9.51%) photographs for testing and 2653 (90.48%) images for training. The normalization approach was used to scale, with all picture parameters being rescaled from (0, 255) to (0, 1) in the dataset.

Feature Extraction

Overview

A CNN design, ResNet-50, is an integral component of the residual network (ResNet) family. Because of its well-known deep architecture, ResNet succeeds in challenges involving picture recognition and classification. In particular, ResNet-50 is a 50-layer version of the ResNet architecture that is comparatively powerful and deep [24].

Key Features of the ResNet-50 Architecture

ResNet-50 is a deep learning model consisting of 50 layers, which include 1 convolutional layer, 16 residual blocks (each containing multiple convolutional layers), a global average pooling layer, and a fully connected (dense) layer. A major innovation of ResNet-50 is its use of residual blocks with identity shortcut connections that bypass certain layers, addressing the vanishing gradient issue commonly encountered during deep network training. Furthermore, the model uses a bottleneck design within the residual blocks to optimize computation, incorporating 3 convolutional layers: a 1×1 convolution to reduce dimensions, a 3×3 convolution for the core transformation, and a second 1×1 convolution to restore the original dimensions. With approximately 23 million parameters, ResNet-50 is more computationally efficient than deeper variants such as ResNet-101 and ResNet-152. This architecture has demonstrated high performance in tasks such as object detection, image segmentation, and image classification. ResNet-50 notably triumphed in the 2015 ImageNet competition, achieving significantly lower training errors than previous models. This model, presented in Figure 2, processed images and extracted features from them using the network. The extracted features, along with their corresponding labels, were stored in CSV files for later use with traditional machine learning algorithms. In addition, the label encoder was saved for decoding the labels in the future.

XSL•FO
RenderX

Figure 2. Residual network with a 50-layer architecture used for feature extraction.

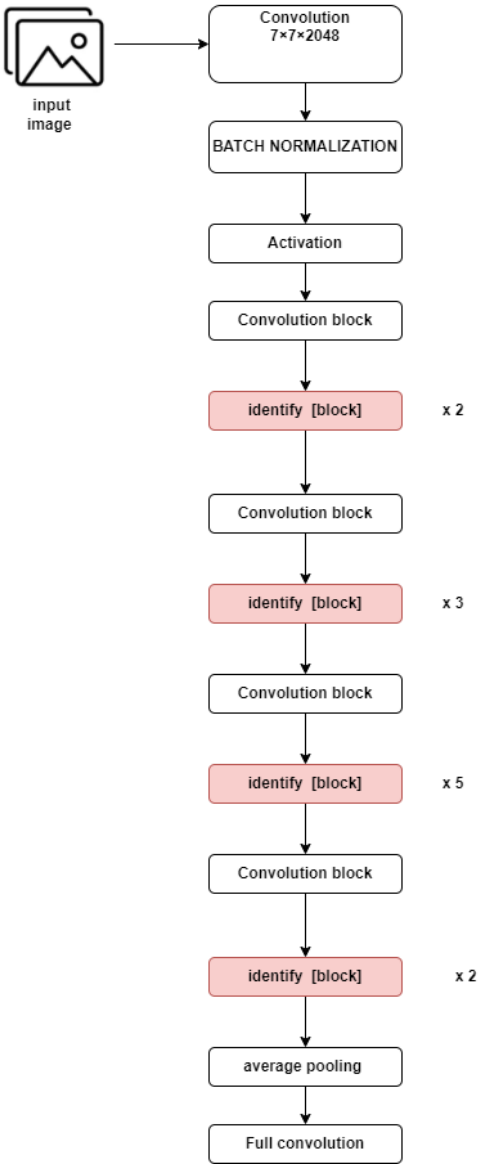

MLP Model

MLP was a kind of artificial neural network that was exploited in deep learning and machine learning. Because it was a feedforward neural network, there were no loops or feedback connections in the information flow, which proceeded from the input layer through hidden layers to the output layer.

Input Layer

The input features of the data were represented by the neurons (nodes) in this layer. The input values were given to each neuron, which represented a feature.

Hidden Layers

There might be 1 or more hidden layers in between the input and output layers. Every buried layer was made up of many neurons. To extract complicated patterns and representations from the input data, these neurons computed the data using activation functions and weights.

Weights

Weights were the parameters that the network learned during training. They verified how strongly neurons in various levels were connected to one another. The network learned to make predictions by varying these weights.

Activation Functions

Every neuron in the buried layers usually adds a weighted sum of its inputs to an activation function. The sigmoid rectified linear unit and hyperbolic tangent functions were examples of common activation functions. These functions provided the model with nonlinearity, which enabled it to discover complex links in the data.

Output Layer

Using the data acquired in the hidden layers, the output layer, which was the last layer, generated the model's predictions. The kind of problems being solved determines how many neurons were in this layer. For multiclass classification, there were numerous neurons, 1 for each class; however, for binary classification, there was 1 neuron that outputs a probability.

XSL•FO
RenderX

MLP Architecture and Configuration

The MLP model used in this study consisted of 2 hidden layers, each serving a distinct functional role in the feature learning process. The first hidden layer was composed of 128 neurons and was designed to capture high-dimensional and complex patterns from the input data. This larger capacity allowed the network to model intricate relationships that were often not linearly separable. The second hidden layer contained 64 neurons and was intended to refine the representations learned by the first layer. By reducing the dimensionality of the intermediate feature space, it retained only the most salient and relevant patterns necessary for effective classification. This hierarchical structure emulated human perceptual processes, where general features were recognized first, followed by more specific details. This model had 2 hidden layers, as shown in Figure 3. The MLP model was trained using the MLP classifier module from the scikit-learn library with the parameters provided in Table 2.

Figure 3. Architecture of the multilayer perceptron model used for autism spectrum disorder prediction.

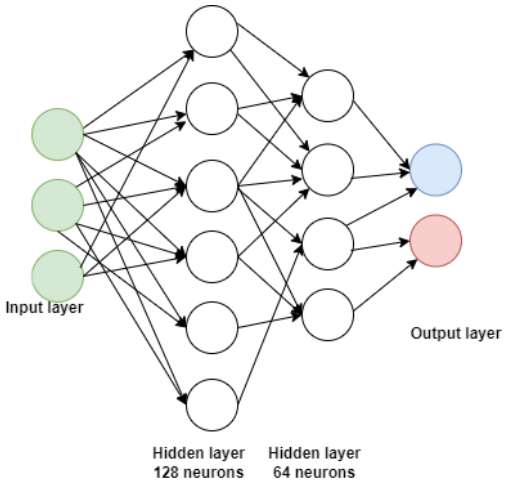

Table 2. Multilayer perceptron parameters.

Parameter	Description
Hidden layer sizes=(128, 64)	Two hidden layers with 128 and 64 neurons
Activation="relu"	ReLU[a] activation function for nonlinearity
Solver="adam"	Adam[b] optimizer for efficient training
Alpha=0.0001	L2 regularization coefficient to prevent overfitting
Batch size=64	Number of samples per gradient update
Learning rate="constant"	Static learning rate throughout training
Initial learning rate=0.001	Initial learning rate for weight updates
Maximum iterations=1000	Maximum number of iterations for convergence
Momentum=0.9	Momentum for SGD[c] (not used with Adam)
Nesterov momentum=true	Enables Nesterov momentum (if SGD is selected)

[a]ReLU: rectified linear unit.

[b]Adam: adaptive moment estimation.

[c]SGD: stochastic gradient descent.

The hidden layer sizes (128 and 64), learning rate (0.001), and batch size (n=64) were selected based on preliminary experiments that aimed at optimizing classification accuracy while avoiding overfitting. These values were also consistent with commonly adopted configurations in similar deep learning classification tasks.

Following training, the MLP model was used to classify the input images into autistic or nonautistic categories. The output of this classification was then passed to subsequent modules for gelotophobia analysis via facial expression recognition and questionnaire validation.

Confusion Matrix

Offering a tabular depiction of true and false values in test results, a confusion matrix is a useful tool for assessing categorization performance. The confusion matrix of the MLP model is provided in Figure 4.

XSL•FO
RenderX

Figure 4. Confusion matrix. FN: false negative (incorrectly classified as negative when they are actually positive); FP: false positive (incorrectly classified as positive when they are actually negative); TN: true negative (correctly classified negative instances); TP: true positive (correctly classified positive instances).

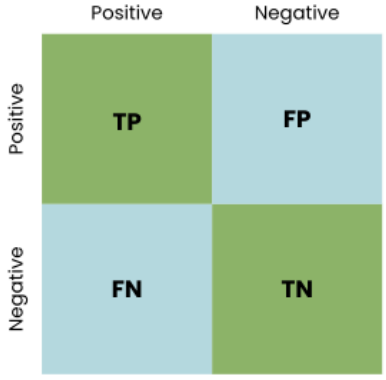

Accuracy, defined as the percentage of correct classifications, was calculated using the following formula:

Accuracy = (TP + TN) / (FP + FN + TP + TN) × 100% **(1)**

Specificity, which measured the model's ability to correctly identify children without autism, was computed as follows:

Specificity = TN / (TN + FN) × 100% **(2)**

Sensitivity, which assessed the model's capability to accurately identify children with autism, was determined as follows:

Sensitivity = TP / (TP + FP) × 100% **(3)**

In equations 1 to 3, TP represents true positive, FP represents false positive, TN represents true negative, and FN represents false negative. Specificity focused on correctly identifying normal children, while sensitivity focused on correctly identifying children with autism.

DeepFace

DeepFace is a lightweight and versatile deep learning framework designed for face recognition and facial attribute analysis. It abstracts the complexity of multiple state-of-the-art face recognition models, providing a unified interface for tasks such as face verification, identification, and emotion analysis. DeepFace supports several backbone models, including VGG-Face, Facenet, ArcFace, Dlib, and more, allowing flexible deployment depending on application needs. Among these, VGG-Face served as the default model, recognized for its robust performance in face recognition tasks. The VGG-Face architecture, as depicted in Figure 5, was based on the VGG-16 network and consisted of a deep stack of convolutional layers followed by fully connected layers, culminating in a softmax classifier trained on a large-scale dataset of celebrity faces.

Figure 5. Architecture of the Visual Geometry Group (VGG) Face model used in DeepFace. FC: fully connected; ReLU: rectified linear unit.

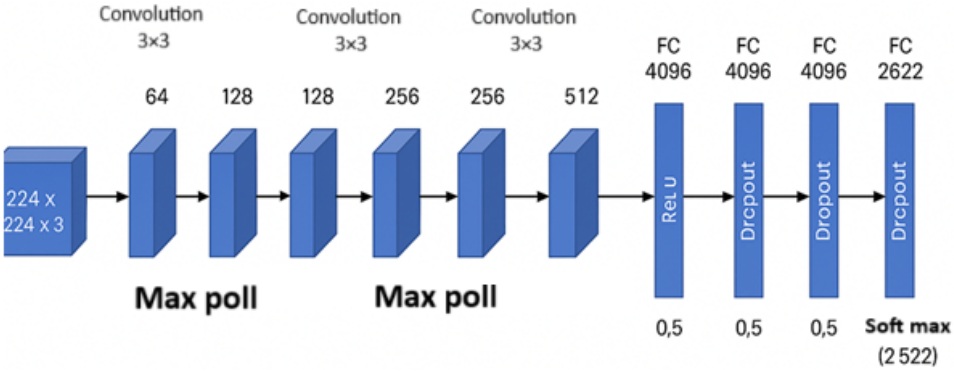

Integration of DeepFace and GELOPH<15> Outputs

In this study, a sequential approach was used to integrate the outputs of the MLP, DeepFace, and GELOPH<15> models for diagnosing ASD and predicting gelotophobia. First, facial images were fed into the MLP model to classify whether the individual had ASD. Then, the same images were analyzed by DeepFace to detect facial expressions linked to gelotophobia. If DeepFace identified emotions associated with gelotophobia, the prediction was further validated using the GELOPH<15> psychometric questionnaire, which assessed self-reported fear of being laughed at. This hybrid approach combined objective facial expression analysis with subjective questionnaire responses to improve the accuracy and reliability of gelotophobia detection, especially in ambiguous cases.

Ethical Considerations

This study was approved by the Faculty of Computers and Information Research Ethics Committee (approval code 2024 01 004). All procedures were conducted in accordance with the Declaration of Helsinki and followed ethical standards regarding data privacy and the responsible use of AI in health-related research. Facial image data used for autism detection were obtained from a publicly available, anonymized dataset on Kaggle, ensuring no personally identifiable information was used. The GELOPH<15> questionnaire was not directly

XSL•FO
RenderX

administered to participants in this study; it was used solely as a reference from previously validated research. No direct contact with human participants occurred, and all data were either publicly available or ethically approved under the research protocol.

Results

Overview

The experimental results were obtained using a PC equipped with an 11th-generation Intel Core i7-11800H control processing unit running at 2.30 GHz, 16 GB of RAM, and a 64-bit Windows 10 operating system. The proposed system and experiments were implemented using Python 3.9.7 (Python Software Foundation), along with the following libraries: scikit-learn 0.24.2, TensorFlow 2.9.2, and DeepFace 0.0.93. The results of our research to identify ASD and gelotophobia are presented in this section. In this analysis, we performed 10-fold cross-validation on the training data using an MLP classifier to evaluate its performance on predicting the target variable. The cross-validation process allowed us to assess the model's generalizability by dividing the data into 10 distinct subsets or folds, ensuring that each subset served as a validation set while the remaining data were used for training. For each fold, several evaluation metrics, including accuracy, precision, recall, and F_1-score, were calculated. The results for each fold are provided in Table 3, along with the average and SD across all folds.

The best model was then used for testing; the model's performance was assessed on the test data using classification metrics and a confusion matrix, which were used for further evaluation. The system achieved 92% accuracy, 92% precision, 93% recall, and 92% F_1-score. The confusion matrix is provided in Figure 6.

Table 3. Performance metrics of the multilayer perceptron classifier using 10-fold cross-validation for the detection of autism spectrum disorder and gelotophobia.

Fold	Accuracy	Precision	Recall	F_1-score
1	0.800752	0.800905	0.800752	0.800727
2	0.815789	0.81595	0.815789	0.815766
3	0.793233	0.794048	0.793233	0.79309
4	0.796226	0.796226	0.796226	0.796226
5	0.781132	0.784392	0.781132	0.780563
6	0.803774	0.804073	0.803774	0.80374
7	0.773585	0.774186	0.773585	0.773488
8	0.822642	0.82358	0.822642	0.822535
9	0.818868	0.81893	0.818868	0.818852
10	0.830189	0.830373	0.830189	0.830174
All folds, average (SD)	0.803619 (0.018339)	0.804266 (0.017908)	0.803619 (0.018339)	0.803516 (0.018427)

Figure 6. Confusion matrix.

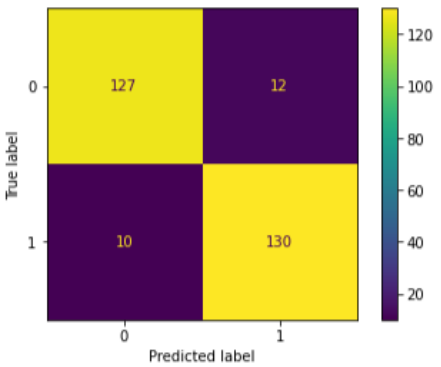

For identifying ASDs, we compared the proposed system with various machine learning models to validate its performance, including a random forest classifier, light gradient boosting machine classifier, K-nearest neighbors classifier, support vector classifier, logistic regression, stochastic gradient descent classifier, and MLP classifier, as shown in Figure 7. Each model was trained and tested to detect facial characteristics distinguishing children with autism from those without autism. Notably, the MLP model outperformed all other models, achieving the highest testing accuracy of 92%. In contrast, the K-nearest neighbors classifier delivered the lowest accuracy at 72%. The dataset was generated from various internet sources, leading to natural variability in age and image quality. Nevertheless, the MLP model demonstrated exceptional accuracy and robustness, as provided in Table 4.

XSL•FO
RenderX

Figure 7. Comparison of classification models for autism spectrum disorder identification based on facial features. The table summarizes each model's testing accuracy. LGBM: light gradient boosting machine; MLP: multilayer perceptron; RF: random forest; SGD: stochastic gradient descent.

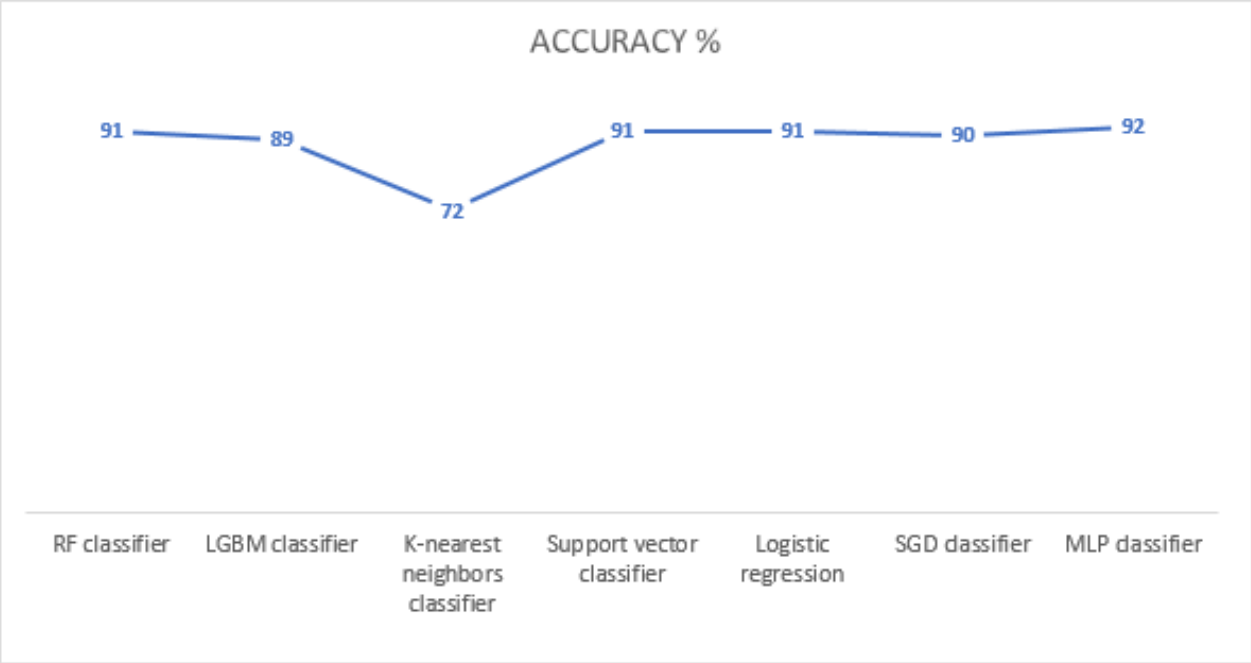

Table 4. Pretrained deep learning models.

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F_1-score (%)
RF[a] classifier	91	90	91	91
LGBM[b] classifier	89	88	91	89
K-nearest neighbors classifier	72	64	98	77
Support vector classifier	91	89	94	91
Logistic regression	91	89	94	91
SGD[c] classifier	90	92	88	90
MLP[d] classifier	92	91	92	92

[a]RF: random forest.
[b]LGBM: light gradient boosting machine.
[c]SGD: stochastic gradient descent.
[d]MLP: multilayer perceptron.

Evaluation Outcomes

A trained deep learning model called DeepFace was created by Facebook's AI research and was built on different deep neural network architectures for face analysis tasks, including emotion detection. CNNs were commonly used by DeepFace and related models to extract features from facial photographs [13]. DeepFace is a lightweight face recognition and facial attribute analysis framework. It wraps several models, such as *VGG-Face.* The model was based on the *VGG-16* CNN architecture (Table 5). Key features are as follows: the input was RGB (red, green, and blue) images of size 224×224; there were 13 convolutional layers, 5 max pooling layers, and 3 fully connected layers at the end; activation included rectified linear unit for all layers; and the final fully connected layer was softmax with 2622 classes (corresponding to 2622 people in the training set).

XSL•FO
RenderX

Table 5. DeepFace model training parameters.

Parameter	Value or description
Learning rate	Started at 0.001 (learning rate), decreased during training
Optimizer	SGD[a]
Momentum	0.9 (momentum)
Weight decay	5e-4 (0.0005; L2 regularization [ridge regression])
Batch size, n	256
Dropout	0.5 (dropout; in fully connected layers)
Training time	Trained on 2.6 M images for days using 4 GPUs[b]
Loss function	Softmax cross entropy

[a]SGD: stochastic gradient descent.

[b]GPU: graphical processing unit.

The deep learning model typically used ≥1 fully linked layers, often referred to as dense layers, to conduct classification after feature extraction. These layers took the characteristics that had been collected and used machine learning to associate them with particular emotions. Similar emotion identification challenges may be implemented with open-source deep learning frameworks such as PyTorch or TensorFlow. These frameworks provided the ability to design the architecture and equations for unique deep learning models, which may be used to construct and train models for a variety of tasks, including emotion detection.

Facial expressions indicating gelotophobia (the fear of being laughed at) included tense brows, forced smiles, avoiding eye contact, blushing, and nervous laughter, all of which reflected discomfort or the fear of being ridiculed. If a person's face exhibited these signs, we can predict using an algorithm whether they had gelotophobia or not.

GELOPH<15>

The GELOPH<15> test has proven to be a reliable tool for measuring gelotophobia among children and adolescents. In a study conducted with a sample of children and adolescents aged 11 to 16 years, the Danish version of the GELOPH<15> questionnaire demonstrated good psychometric properties, including high internal consistency and a 1D factor structure that closely aligned with the adult version. Despite the need for younger participants to receive assistance from teachers in filling out the items, the test was found to be reliable and easily understood by most participants. The study highlighted that the GELOPH<15> questionnaire can serve as a solid foundation for future research on gelotophobia, with its capacity to reliably measure the fear of being laughed at across various age groups. Furthermore, the GELOPH<15> questionnaire's consistency with the adult version indicated its applicability across different populations, offering a robust tool for studying gelotophobia in both children and adults [14].

The GELOPH<15> consists of 15 items rated on a 4-point Likert scale (1 = strongly disagree to 4 = strongly agree). A mean score of 2.5 or higher typically indicates the presence of gelotophobia. If participants strongly agree with at least 4 statements, this may also reflect a strong tendency toward gelotophobia (Multimedia Appendix 1 [25]).

Discussion

Principal Findings

This study aimed to explore the potential of facial emotion recognition in detecting ASD and gelotophobia using deep learning approaches. Our system achieved a high accuracy of 92%, surpassing many existing models in the field. Integration the DeepFace model with an MLP classifier allows for robust feature extraction from facial expressions, contributing to a reliable ASD diagnosis. In addition, we included the GELOPH<15> questionnaire to assess gelotophobia, a comorbid condition often observed in individuals with ASD. This dual-focus approach of detecting ASD while simultaneously considering the social anxieties related to being laughed at offers a more comprehensive tool than previous models that typically focus on ASD alone.

The proposed system serves as a powerful support tool for clinicians in diagnosing ASD and gelotophobia, enabling faster and more accurate assessment compared to traditional methods. With each use, AI continuously learns from new data, improving its performance over time and enhancing the reliability of clinical decisions. This makes the system easily integrable into existing clinical care pathways, reducing the burden on medical staff and improving patient outcomes through earlier and more comprehensive diagnosis.

Model Interpretability and Clinical Relevance

Although deep learning models, such as DeepFace and MLP, are often considered "black boxes," we have taken steps to enhance their interpretability for clinical use. Key facial expression features related to emotions such as fear, sadness, and social anxiety were identified as the most informative indicators for classifying ASD and gelotophobia. To assist clinicians, our system can be extended to include visualization tools, such as heat maps or feature importance scores (eg, using Shapley additive explanation or local interpretable model-agnostic explanation techniques), which highlight the facial regions or features most influential in the model's decisions. This transparency supports clinicians in understanding

XSL•FO
RenderX

and trusting the system's output, facilitating its integration into diagnostic workflows.

Comparison to Previous Work

Table 6 compares the performance of our model with that of several notable studies in the field, focusing on accuracy, methodology, and the disorders targeted. As shown in Table 6, our system outperforms many previous approaches, especially in accuracy, while also addressing gelotophobia, which is often overlooked in ASD detection (Table 7).

Table 6. Comparison to previous work.

Authors	Model or approach	Accuracy (%)	Disorder targeted	Notes
This study	DeepFace and MLP[a]	92	Autism or gelotophobia	High accuracy, combining facial features and deep learning
Ul Haque and Valles [15], 2018	DCNN[b]	82	Autism	Lower accuracy; focused on facial expression detection
Ganesan et al [16], 2021	ML[c]	84.5	Autism	Focused on facial features; less robust generalization
Eslami et al [17], 2019	Hybrid learning (fMRI[d])	90	Autism	High, but <92%; relies on fMRI, which limits practicality
Liu et al [19], 2016	ML on face processing	81.7	Autism	Insufficient accuracy for clinical use
Wu et al [20], 2021	ML from videos	89.2	Autism	Video-based study; strong results, but <92%
Canestrari et al [2], 2021	Survey or statistical analysis	N/A[e]	Gelotophobia (with ASD[f])	Explores gelotophobia in those who experience cyberbullying with parental attachment correlations
Leader et al [4], 2018	Statistical analysis (psychometrics)	N/A	Gelotophobia in individuals with ASD	Psychological assessment of gelotophobia in individuals with high-functioning ASD
Husseiny et al [14], 2024	Questionnaire based	N/A	Gelotophobia in teenagers with ASD	Descriptive statistics; not a classifier

[a]MLP: multilayer perceptron.
[b]DCNN: deep convolution neural network.
[c]ML: machine learning.
[d]fMRI: functional magnetic resonance imaging.
[e]N/A: not applicable.
[f]ASD: autism spectrum disorder.

Table 7. Comparison of the results of our system and existing models.

Study	Deficiency
Ul Haque and Valles [15], 2018	Focused only on autism detection using the Facial Expression Recognition 2013 dataset, which lacks diversity in expressions, age, ethnicity, and gender
Ganesan et al [16], 2021	Limited to diagnosing autism, with challenges such as low SVM[a] accuracy (approximately 65%) and dataset diversity affecting model generalizability
Eslami et al [17], 2019	Focused solely on autism detection using fMRI[b] data, with accuracy (82%) insufficient for clinical application, and issues with generalization
Nunes et al [5], 2019	Preliminary research, not yet integrated into clinical practice; lacks robust multimodal fusion development
Wu et al [20], 2021	Low accuracy in detecting behavioral cues such as vocalization (53%) and object focus (67%), limiting ASD[c] diagnosis reliability
Alsaade and Alzahrani [8], 2022	Transfer learning risks overfitting, especially on small or homogeneous datasets, affecting generalization
Ruch [10], 2009	Study limited by small sample size, lack of diversity, and reliance on self-reports, which may bias gelotophobia assessment in individuals with high-functioning ASD

[a]SVM: support vector machine.
[b]fMRI: functional magnetic resonance imaging.
[c]ASD: autism spectrum disorder.

XSL•FO
RenderX

Limitations

Despite promising accuracy, our study has several limitations. First, the dataset—sourced from publicly available 2D facial images—may not fully represent the diversity of facial expressions in children with ASD or gelotophobia, as demographic variables, such as age, gender, and cultural background, were not explicitly controlled. Second, while 10-fold cross-validation mitigates overfitting, the lack of an external validation cohort limits confidence in generalizability. Third, subtle emotional states (eg, fear and disgust) remain challenging for the current model to distinguish reliably, as these expressions often involve minimal facial muscle movements. Fourth, our approach relies solely on static images; real-world interactions are dynamic and multimodal (including voice, gesture, and context), which our system does not capture. Finally, although we propose advanced architectures (ResNet and EfficientNet) and multimodal inputs (eye-tracking and speech) as future directions, these were not implemented in this study.

Future Work

To address the current limitations, future efforts will focus on expanding and diversifying the dataset by collecting and annotating a larger, demographically varied sample of children with and without ASD and gelotophobia, followed by validating the model on independent cohorts from multiple institutions to ensure robustness. We also plan to experiment with more advanced model architectures such as deeper CNNs, including ResNet-50 and EfficientNet-B4, and lightweight models, such as MobileNet-V3, aiming to improve the extraction of subtle facial expressions. Furthermore, integrating multimodal data, including eye-tracking, speech prosody, and body language cues, will allow capturing dynamic emotional signals and reduce reliance on static images. The development of a real-time application prototype and conducting pilot studies in clinical settings will evaluate the system's usability, acceptability, and diagnostic utility. In addition, longitudinal studies are planned to investigate whether early emotion recognition markers can predict later social outcomes or responses to interventions. Finally, we will incorporate explainable AI techniques to interpret model decisions and implement fairness checks to detect and mitigate demographic biases.

Conclusions

This study demonstrated the promising potential of using facial expression recognition combined with deep learning techniques for the detection of ASD and gelotophobia. By leveraging the power of the DeepFace model to extract facial features, we achieved an accuracy of 92% using an MLP classifier. These findings highlight the significant role of facial expression analysis in identifying subtle emotional processing deficits in individuals with ASD and gelotophobia, both of which are often characterized by challenges in interpreting social cues.

The results emphasize that deep learning models, particularly those capable of processing facial expressions, can serve as valuable tools in autism diagnosis and the assessment of related comorbid conditions. This approach has the potential to aid clinicians by providing a noninvasive, reliable method for early detection and intervention.

However, further research is needed to refine the model, particularly in increasing its generalizability across diverse populations and enhancing its ability to detect a wider range of emotional expressions. In addition, exploring other advanced deep learning architectures, such as ResNet or EfficientNet, could further improve accuracy. The integration of multimodal data, including eye-tracking and voice analysis, might also provide a more comprehensive assessment of emotional processing in ASD and gelotophobia.

In summary, this study lays the groundwork for future advancements in the field of AI-based diagnostic tools for ASD and related conditions, paving the way for more efficient and accurate clinical practices.

Acknowledgments

This research did not receive any funding from any governmental, commercial, or nonprofit organization.

Data Availability

The datasets generated or analyzed during this study are available in the Kaggle repository [26].

Conflicts of Interest

None declared.

Multimedia Appendix 1
Fear of being laughed at (gelotophobia) questionnaire.
[DOCX File , 15 KB - formative_v9i1e72115_app1.docx]

References

1. Brauer K, Proyer RT. To love and laugh: testing actor-, partner-, and similarity effects of dispositions towards ridicule and being laughed at on relationship satisfaction. J Res Pers 2018 Oct;76(1):165-176. [doi: 10.1016/j.jrp.2018.08.008]

XSL•FO
RenderX

2. Canestrari C, Arroyo GD, Carrieri A, Muzi M, Fermani A. Parental attachment and cyberbullying victims: the mediation effect of gelotophobia. Curr Psychol 2021 Mar 24;42(19):16401-16412 [FREE Full text] [doi: 10.1007/s12144-021-01642-6]
3. Murray A, Mannion A, Chen JL, Leader G. Gaming disorder in adults with autism spectrum disorder. J Autism Dev Disord 2022 Jun 28;52(6):2762-2769 [FREE Full text] [doi: 10.1007/s10803-021-05138-x] [Medline: 34184139]
4. Leader G, Grennan S, Chen JL, Mannion A. An investigation of gelotophobia in individuals with a diagnosis of high-functioning autism spectrum disorder. J Autism Dev Disord 2018 Dec 3;48(12):4155-4166. [doi: 10.1007/s10803-018-3661-3] [Medline: 29971658]
5. Nunes LC, Pinheiro P, Pinheiro MC, Pompeu M, Filho MS, Comin-Nunes R, et al. A hybrid model to guide the consultation of children with autism spectrum disorder. In: Proceedings of the Research & Innovation Forum 2019. 2019 Presented at: RIIFORUM 2019; April 24-26, 2019; Rome, Italy URL: https://doi.org/10.1007/978-3-030-30809-4_38 [doi: 10.1007/978-3-030-30809-4_38]
6. Baio J, Wiggins L, Christensen DL, Maenner MJ, Daniels J, Warren Z, et al. Prevalence of Autism Spectrum Disorder Among Children Aged 8 Years - Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2014. MMWR Surveill Summ 2018 Apr 27;67(6):1-23 [FREE Full text] [doi: 10.15585/mmwr.ss6706a1] [Medline: 29701730]
7. Harris J. Leo Kanner and autism: a 75-year perspective. Int Rev Psychiatry 2018 Feb 18;30(1):3-17. [doi: 10.1080/09540261.2018.1455646] [Medline: 29667863]
8. Alsaade FW, Alzahrani MS. Classification and detection of autism spectrum disorder based on deep learning algorithms. Comput Intell Neurosci 2022 Feb 28;2022:8709145 [FREE Full text] [doi: 10.1155/2022/8709145] [Medline: 35265118]
9. Titze M. Gelotophobia: the fear of being laughed at. Humor 2009;22(1-2):27-48 [FREE Full text] [doi: 10.1515/HUMR.2009.002]
10. Ruch W. Fearing humor? Gelotophobia: the fear of being laughed at introduction and overview. Humor 2009;22(1-2):1-25 [FREE Full text] [doi: 10.1515/HUMR.2009.001]
11. Keates N, Waldock KE. Autistic people, gelotophobia, gelotophilia and katagelasticism: a narrative review. Rev J Autism Dev Disord 2023 May 03;12:40-49. [doi: 10.1007/s40489-023-00376-4]
12. Führ M, Proyer RT, Ruch W. Nordic Psychol 2012 Jul 11;61(2):62-73. [doi: 10.1027/1901-2276.61.2.62]
13. Yolcu G, Oztel I, Kazan S, Oz C, Palaniappan K, Lever TE, et al. Facial expression recognition for monitoring neurological disorders based on convolutional neural network. Multimed Tools Appl 2019 Nov 23;78(22):31581-31603 [FREE Full text] [doi: 10.1007/s11042-019-07959-6] [Medline: 35693322]
14. Husseiny MA, Al-Tihami A, Abdelrazek MM. Gelotophobia in high-functioning adolescents with autism spectrum disorder. J Psychol Counsel 2024;77(2):539-561 [FREE Full text] [doi: 10.21608/cpc.2024.371280]
15. Ul Haque MI, Valles D. A facial expression recognition approach using DCNN for autistic children to identify emotions. In: 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON). New York, NY: IEEE; 2018.
16. Ganesan S, Raju, Senthil J. Prediction of autism spectrum disorder by facial recognition using machine learning. Webology 2021 Sep;18(02):406-417 [FREE Full text]
17. Eslami T, Mirjalili V, Fong A, Laird AR, Saeed F. ASD-DiagNet: a hybrid learning approach for detection of autism spectrum disorder using fMRI data. Front Neuroinform 2019 Nov 27;13:70 [FREE Full text] [doi: 10.3389/fninf.2019.00070] [Medline: 31827430]
18. Arru G, Mazumdar P, Battisti F. Exploiting visual behaviour for autism spectrum disorder identification. In: Proceedings of the IEEE International Conference on Multimedia & Expo Workshops. 2019 Presented at: ICMEW 2019; July 8-12, 2019; Shanghai, China. [doi: 10.1109/ICMEW.2019.00123]
19. Liu W, Li M, Yi L. Identifying children with autism spectrum disorder based on their face processing abnormality: a machine learning framework. Autism Res 2016 Aug 01;9(8):888-898. [doi: 10.1002/aur.1615] [Medline: 27037971]
20. Wu C, Liaqat S, Helvaci H, Samson Chcung SC, Chuah CN, Ozonoff S. Machine learning based autism spectrum disorder detection from videos. In: Proceedings of the IEEE International Conference on E-Health Networking, Application & Services. 2021 Presented at: HEALTHCOM 2021; March 1-2, 2021; Shenzhen, China URL: https://doi.org/10.1109/healthcom.2021.9398924 [doi: 10.1109/healthcom49281.2021.9398924]
21. Akter T, Ali MH, Khan MI, Satu MS, Uddin MJ, Alyami SA, et al. Improved transfer-learning-based facial recognition framework to detect autistic children at an early stage. Brain Sci 2021 May 31;11(6):734 [FREE Full text] [doi: 10.3390/brainsci11060734] [Medline: 34073085]
22. Alcañiz M, Chicchi-Giglioli IA, Carrasco-Ribelles LA, Marín-Morales J, Minissi ME, Teruel-García G, et al. Eye gaze as a biomarker in the recognition of autism spectrum disorder using virtual reality and machine learning: a proof of concept for diagnosis. Autism Res 2022 Jan 22;15(1):131-145 [FREE Full text] [doi: 10.1002/aur.2636] [Medline: 34811930]
23. Kats Y. Integrated support of students with autism spectrum disorders and learning disabilities. In: Kats Y, Stasolla F, editors. Education and Technology Support for Children and Young Adults With ASD and Learning Disabilities. Hershey, PA: IGI Global; 2021.
24. Samson A. Humor(lessness) elucidated – sense of humor in individuals with autism spectrum disorders: review and introduction. Humor 2013;26(3):393-409. [doi: 10.1515/humor-2013-0027]

XSL•FO
RenderX

25. Ruch W, Proyer RT. The fear of being laughed at: individual and group differences in Gelotophobia. Humor 2008;21(1):47-67 [FREE Full text] [doi: 10.1515/humor.2008.002]
26. Autism_Image_Data. Kaggle. URL: https://www.kaggle.com/datasets/cihan063/autism-image-data [accessed 2025-05-29]

Abbreviations

AI: artificial intelligence
ASD: autism spectrum disorder
CNN: convolutional neural network
FP: false positive
FN: false negative
MLP: multilayer perceptron
ResNet: residual network
ResNet-50: residual network with 50 layers
TN: true negative
TP: true positive
VGG: Visual Geometry Group

Edited by J Sarvestan; submitted 04.02.25; peer-reviewed by A Elhady, A El Mouatasim, A Kocian, R Simeoli; comments to author 10.04.25; revised version received 22.05.25; accepted 17.06.25; published 13.08.25.

Please cite as:
Eldawansy M, El Bakry H, M Shohieb S
Data Mining–Based Model for Computer-Aided Diagnosis of Autism and Gelotophobia: Mixed Methods Deep Learning Approach
JMIR Form Res 2025;9:e72115
URL: https://formative.jmir.org/2025/1/e72115
doi:10.2196/72115
PMID:40802390

©Mohamed Eldawansy, Hazem El Bakry, Samaa M Shohieb. Originally published in JMIR Formative Research (https://formative.jmir.org), 13.08.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Early Diagnosis of Knee Osteoarthritis With a Natural Language Processing–Driven Approach Based on Clinician Notes: Development and Validation Study

Narathip Thanyakunsajja[1], MS; Kulsawasd Jitkajornwanich[2], PhD; Shan Xu[2], PhD; Donghee Shin[2], PhD; Pattama Charoenporn[1], PhD

[1]King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand
[2]Texas Tech University, Lubbock, TX, United States

Corresponding Author:
Kulsawasd Jitkajornwanich, PhD
Texas Tech University
2500 Broadway
Lubbock, TX, 79409
United States
Phone: 1 6824149164
Email: kulsawasd.jitkajornwanich@ttu.edu

Abstract

Background: Knee osteoarthritis (OA) is a common form of knee arthritis that can cause significant disability and affect a patient's quality of life. Although this disease is chronic and irreversible, the patient's condition can be improved and the progression of the disease can be prevented if the disease is diagnosed early and the patient receives appropriate treatment immediately. Therefore, the prediction of knee OA is considered one of the essential steps to effectively diagnose and prevent further severe OA conditions. Knee OA is commonly diagnosed by medical experts or physicians, and the diagnosis of OA is mainly based on patients' laboratory results and medical images, including x-ray and magnetic resonance images. However, diagnosis through such data is often time-consuming. Moreover, the diagnosis results can vary among physicians depending on their expertise. Previous studies mostly focused on using approaches, such as those involving artificial intelligence, to automatically detect knee OA through such data. However, these studies did not incorporate clinicians' or doctors' notes (text data) into the analysis, although these data involving reported symptoms and behaviors are already available and easier to collect and access than laboratory data and image data.

Objective: We propose a novel natural language processing–driven approach based on clinicians' or doctors' notes of patient-reported symptoms (text data only) for diagnosing knee OA.

Methods: The textual information from clinicians' or doctors' notes was first preprocessed using text analysis algorithms with respect to natural language processing. We then incorporated deep learning models, including convolutional neural networks, bidirectional long short-term memory (BiLSTM), and gated recurrent units. Lastly, a disease-specific standard questionnaire called WOMAC (Western Ontario and McMaster Universities Arthritis Index) was taken into account to improve the overall performance of the models.

Results: Our experiment included 5849 records (OA: 3455; non-OA: 2394). Before applying our WOMAC-based processing approach, the best-performing model was BiLSTM (area under the curve, 0.85; accuracy, 0.87; precision, 0.85; sensitivity, 0.95; specificity, 0.76; F_1-score, 0.90), and there was an improvement in the results with BiLSTM after applying our approach (area under the curve, 0.91; accuracy, 0.91; precision, 0.91; sensitivity, 0.94; specificity, 0.87; F_1-score, 0.93).

Conclusions: Our proposed method for predicting the occurrence of knee OA showed better performance than other conventional methods that use image data and statistical laboratory data. The findings indicate the feasibility of using text data (symptom descriptions reported by patients and recorded by doctors) to predict knee OA. Medical notes of symptom reports can be considered a valuable data source for predicting whether a particular knee is likely to experience OA progression.

(JMIR Form Res 2025;9:e64536) doi:10.2196/64536

XSL•FO
RenderX

KEYWORDS

computational communication; health informatics; natural language processing; artificial intelligence; text mining

Introduction

Background

Knee osteoarthritis (OA) is a long-term condition, and its cause is still unknown. The most common form of arthritis involves cartilage degradation or bone abnormalities [1-3]. Joint pain, stiffness, and limitation of knee movement are some of the primary symptoms of knee OA. Knee OA is a slow but aggressive disease, and as the damage to the cartilage and bones cannot be restored, a prevention protocol could be beneficial [4]. Knee OA affects 10%-15% of adults worldwide. In 2015, the World Health Organization (WHO) estimated that 9.6% of men and 18% of women aged 60 years or older had symptomatic knee OA. Knee OA symptoms affect mobility in 80% of cases, while 25% of people with symptomatic knee OA struggle to perform their daily activities. Knee OA can not only affect individuals both physically and psychologically but also create a burden on society and the economy [5]. In 2008, knee OA was ranked as one of the most expensive diseases to treat. The cost of hospital treatment for arthroplasty surgery has reached approximately US $40 billion worldwide [6]. Having said that, knee OA can be diagnosed at an early stage of the disease. Early detection and diagnosis are possible and crucial to overcome the issues associated with this highly disabling disease, as early-stage treatment can prevent further breakdown of the bone and cartilage [1,7].

Owing to the presence of big data and the advancement of artificial intelligence (AI), many researchers are using these to identify key factors influencing health and disease-related issues with the help of machine learning models. Furthermore, deep learning is one of the new and rising trends and opportunities in the field of machine learning for many studies in diverse domains. This is particularly true in the medical domain, where deep learning and natural language processing (NLP) could be used for medical text classification tasks [8,9]. This has led to the emergence of new research areas in medical text classification and summarization. Machine learning approaches for text classification can be roughly divided into 2 types: traditional machine learning approaches and deep learning–based approaches. Because traditional approaches have a limited ability to process large text data, which requires complex computations in the corpus, the performance of models is greatly affected. On the other hand, deep learning models have been shown to be highly effective in text classification [10]. Moreover, many deep learning models, such as recurrent neural networks (RNNs), long short-term memory (LSTM), convolutional neural networks (CNNs), and gated recurrent units (GRUs), are widely used in the field of text classification [11,12]. To use deep learning in NLP, we need a form of word representation to transform words into vectors, known as word embeddings. One of the most widely used methods for this is Word2Vec [13,14], which we adopted in our work.

In this study, we have proposed an NLP model to detect the occurrence of knee OA by incorporating deep learning models (GRUs, bidirectional long short-term memory [BiLSTM], and CNNs) based on the textual data of knee OA symptoms reported in doctors' notes. Moreover, we have performed a literature review. The methodology of the proposed sentiment analysis has been explained in the Methods section, and the results have been provided in the Results section. Finally, discussions, conclusions, and future work have been communicated in the Discussion section.

Literature Review

In the medical AI–related literature, deep learning has been used to help detect and diagnose diseases. Often, each study is differentiated by the type of data used, such as images and statistical data, including laboratory results. We have discussed several studies in this section related to the use of AI, machine learning, and deep learning for detecting and predicting knee OA based on both medical images and statistical data, including laboratory results.

In the study by Brahim et al [1], x-ray images of the knees of patients, together with machine learning algorithms, were used to develop a computer-aided diagnosis (CAD) system for detecting early knee OA. The machine learning algorithms used were naïve Bayes and random forest models, which were adopted for classification tasks. The approach used 1025 knee x-ray images from the public Knee Osteoarthritis Initiative database [15,16]. Yong et al [17] proposed a method to detect Kellgren/Lawrence (KL) grading [18], using an ordinal regression module (ORM) in neural networks. The module can predict the inclusion of images in multiple classes that follow a natural severity order, ranging from KL grades 0 to 4. They also used data from the Knee Osteoarthritis Initiative [15,16]. A total of 4130 pairs of knee images were labeled with KL grades and collected from the baseline cohort. These studies yielded good results, but they only considered x-ray images (data were obtained from x-ray machines). Halilaj et al [19] proposed a mixed effects model to differentiate clusters of knee OA progression. They used data from 1234 patients in the Knee Osteoarthritis Initiative, who were classified as being at high risk of developing knee OA based on age, BMI, and medical and occupational histories, to identify knee OA progression. They also developed models to predict the probability of belonging to a cluster using Least Absolute Shrinkage and Selection Operator (LASSO) regression. Lim et al [6] proposed a feed-forward neural network trained with standard back propagation as a deep learning model to detect the occurrence of knee OA using the statistics of patients, including medical and health-related behavioral information. They used data from 5749 subjects in the Korea National Health and Nutrition Examination Survey [20], which was performed by the Korea Centers for Disease Control and Prevention (KCDC). Wang et al [21] developed a model to assess the severity of radiographic knee OA and identify key risk factors for early intervention and treatment. To predict the KL grade, they used a LSTM model with an attention mechanism based on data from 518 patients in the Knee Osteoarthritis Initiative [15,16]. From these previous studies, we noticed that even though the statistical methods

XSL•FO
RenderX

were faster than the image methods, the high accuracy of the image methods yielded better results compared to the statistical methods. Recently, Tiulpin et al [22] developed a machine learning approach to predict the progression of structural knee OA, using raw radiographic image data, physical examination results, patient medical history, anthropometric information, and KL grade as features. They included data from the Knee Osteoarthritis Initiative (2711 subjects) [15,16] in the training set and data from the Multicenter Osteoarthritis Study (2129 subjects) [16,23] in the test set. The experiments combined both the image method and statistical method and yielded good performance. However, this approach considered x-ray images, which need to be retrieved from x-ray machines and take a long time to process.

Recently, the field of NLP has shown significant technological advances, especially in text classification involving assigning a label or class to a given text. Text classification has many practical applications, such as spam detection in email servers, news categorization, sentiment analysis, hate speech detection in social media, etc. It has also been applied to medical fields. For example, Liu et al [8] applied text classification to label words retrieved from social media big data as medical terms and categorized them accordingly.

Traditional text classification approaches are dictionary-based and involve basic machine learning methods [24]. Differentiating texts related to patients is challenging. This is because the texts used in the medical field are still unstructured data and can be complicated to process, especially when processed by basic machine learning approaches [25]. On the other hand, deep learning methods have shown strong performance in image classification as well as speech recognition. These methods have been widely used in the fields of NLP and text classification and have yielded good results [26]. CNNs and RNNs are among the most widely adopted deep learning models for text classification.

Jang et al [24] proposed a method to classify movie user reviews into positive, negative, and neutral categories using a sentiment analysis model that combines BiLSTM and CNN. The Internet Movie Database (IMDb) movie review dataset was used for this experiment. The model used a CNN to reduce the number of input features. The features extracted from the CNN were input into a BiLSTM model for the final classification. Qing et al [26] proposed a neural network approach to classify medical texts into one of four dataset categories: (1) clinical records from Traditional Chinese Medicine in the Classified Medical Records of Distinguished Physicians Continued Two (TCM), (2) an open inpatient medical records dataset from the China Conference on Knowledge Graph and Semantic Computing (CCKS), (3) a corpus of biomedical publication abstracts labeled for cancer hallmarks, and (4) a dataset of Activating Invasion and Metastasis (AIM). They employed a convolutional layer to extract features from sentences and used a bidirectional gated recurrent unit (Bi-GRU) for classification.

Many previous studies have focused on image and statistical data for knee OA prediction with machine learning and deep learning techniques. However, a limitation of image data is that they must be generated and retrieved through x-ray machines or magnetic resonance imaging (MRI) scanners. Moreover, statistical data have limitations associated with the integrity of the factors and variables related to patients.

Objectives

NLP or text classification has potential and could be beneficial in knee OA prediction. For the early detection of knee OA, we propose a novel approach based on an NLP or text classification method that uses only textual data, instead of x-ray or MRI image data and statistical data, and incorporates deep learning models, including GRU, BiLSTM, and CNN models. We used textual data in our research, which included clinicians' or doctors' notes on patient symptoms related to knee OA.

Methods

Ethical Considerations

Data permission was approved and granted by the study hospital and the associated institutions according to the institutional review board (IRB2023-736; develop language models for OA diagnosis). Regarding our dataset, no identifiable data of patients were collected in this study. The CONSORT (Consolidated Standards of Reporting Trials) (Multimedia Appendix 1), CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) (Multimedia Appendix 2), and STARD (Standards for Reporting of Diagnostic Accuracy Studies) (Multimedia Appendix 3) checklists [27-29] were followed, and training was provided in human ethics for conducting research in humans.

Resources

All models were trained using a MacBook Pro 2018 laptop having an Intel Core i5 quad-core CPU (2.3 GHz), an Intel Iris Plus 655 GPU, and a Macintosh disk storage (256 GB).

Study Design and Data Sources

The approach comprised 6 stages. First, deidentified data on patients' reported symptoms from clinicians' or doctors' notes were obtained from a hospital in Thailand (N=6737; knee OA: 3945, non-OA: 2792). No personal attributes or biospecimen data, such as laboratory results, were used. Second, data selection, filtering, cleaning, and preparation processes were performed for the data of patients who met the criteria of the experiments (N=5849; knee OA: 3455, non-OA: 2394). Third, text analysis and mining processes were performed on the datasets. Fourth, to enhance the efficiency of the model, we incorporated our text analysis and mining models with the WOMAC (Western Ontario and McMaster Universities Osteoarthritis Index) and specific concepts in our AI model. Fifth, to conform to reporting guidelines, we randomized our data into 3 groups: training, validation, and test sets. Finally, we applied and conducted comparison experiments with deep learning models (ie, CNNs, BiLSTM, and GRUs) to predict knee OA diagnosis.

Texts in the medical field mainly contain 3 types of words. The first category is standardized medical terminology, which refers to specific concepts in the medical field. The second category is user-based health terminology, such as "can't sleep" and "stomach pain." The third category includes words that are

XSL•FO
RenderX

considered noise, such as “because” and “so.” These words typically lack relevant health information and can cause interference [8]. Our dataset primarily combined the first 2 categories. However, the data in this work consisted of words describing symptoms, which can be varied (such as pain [“ ” in Thai] and ache [“ ” in Thai], where both words mean “pain”). Furthermore, the data may have some incorrect words in sentences. Therefore, one of the core steps in our methodology was to use a WOMAC-based processing approach to transform some words in sentences that are similar in the WOMAC corpus to medical terminology by applying the *SequenceMatcher* library [30]. We defined a cutoff score of similarity of 0.85 to replace old words with WOMAC words.

To conform to appropriate reporting guidelines, CONSORT 2010 [27] and CONSORT-EHEALTH [28] guidelines were applied in randomizing our data into 3 groups: training, validation, and test sets. Moreover, STARD 2015 [29] guidelines were used to ensure the quality and clarity of reporting of our diagnostic accuracy studies. Figure 1 illustrates the process in detail.

Figure 1. Schematic representation of our knee osteoarthritis prediction model, consisting of (A) textual raw symptom reports, (B) data cleaning, (C) text processing and analysis, (D) randomizing data (with respect to CONSORT [Consolidated Standards of Reporting Trials], CONSORT-EHEALTH [Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth], and STARD [Standards for Reporting of Diagnostic Accuracy Studies] checklists), (E) training models, and (F) classification. WOMAC: Western Ontario and McMaster Universities Osteoarthritis Index.

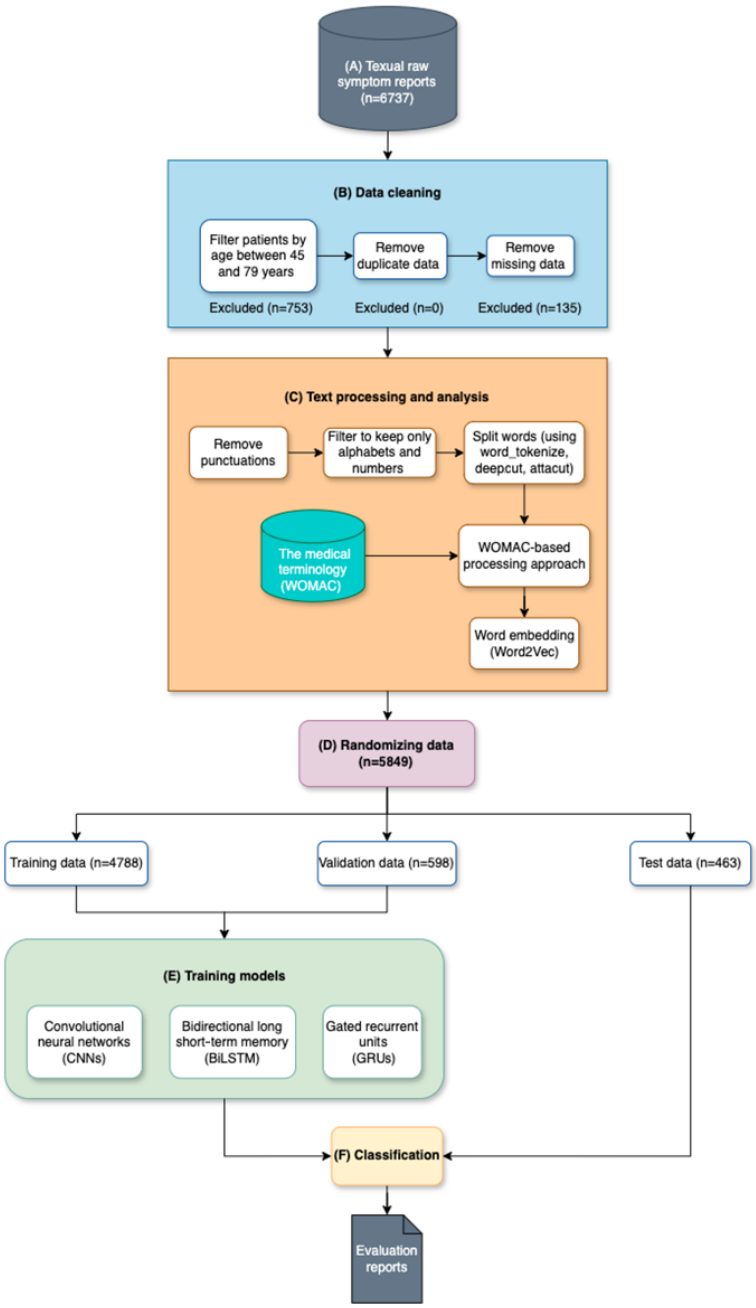

XSL•FO
RenderX

Data Preparation and Experiment Plan

Data Collection

The text reports of knee OA were provided by a hospital in Thailand and were collected by orthopedists between 2019 and 2021. Specifically, we focused only on textual knee OA reports on symptoms in the Thai language, and we considered only data from patients who first visited the hospital. To conduct fair experiments, we only considered patients who were new and had not been treated by orthopedists before, as we focused on patients who had not received treatment by orthopedists. We did not consider patients who had been treated previously, as their symptoms may not be clearly mentioned in their reports, which can degrade the accuracy of our model and predictions. The medical terminology (from the WOMAC corpus) was downloaded from the Royal College of Physiatrists of Thailand [31].

Data Cleaning

The data cleaning process involved selection, cleaning, filtering, and preparation. After obtaining medical text reports from the hospital, we investigated the data and excluded patient data that were incomplete or duplicate in order to keep only complete patient data. Patients aged between 45 and 79 years were considered for 2 reasons. First, many previous studies focused on data from patients aged between 45 and 79 years [1,17,19,21,22]. Second, in the Knee Osteoarthritis Initiative datasets [15,16], which are considered some of the standard resources to study the progression of OA and develop innovative approaches to treat this chronic disease, the focus is on patients aged 45-79 years.

Text Analysis and Processing

Data preprocessing was conducted. Punctuations were removed, and the texts were filtered to include only Thai alphabets and numbers. Moreover, word segmentation, tokenization, and correction of misspelled words were performed to ensure data quality. Next, we incorporated word embedding with the WOMAC and concepts. Each word was initially transformed into its embedding vector, with positional encoding and attention applied to assign weight to the words in each sentence. This yielded the final sentence vector.

Experiment Plan

We planned to use and experiment with 3 classification models, including CNN, BiLSTM, and GRU models, along with NLP components, which have been discussed in detail in the next section.

NLP and Deep Learning Techniques

In this section, we describe the NLP components used in our research, namely, word segmentation and Word2Vec, as well as deep learning models.

Word Segmentation

Since Thai sentences can be verbose and lead to ambiguity, word segmentation was considered before conducting further analysis of the text. The 3 most prevalent word segmentation approaches for Thai sentences are presented below.

Word_Tokenize

Word_Tokenize [32,33] is one of the functions in the PyThaiNLP library. Word_Tokenize uses the maximal matching algorithm, which involves dictionary-based maximum matching and uses the Thai character cluster. Results demonstrated that it is the fastest word tokenizer on the BEST 2010 [34] benchmark, with 71.18% accuracy [35,36].

DeepCut

DeepCut [37] is a technique built on a CNN that uses characters and character types as features. The goal of the model is to predict whether a character is the beginning of a word before word segmentation can be processed. The BEST 2010 corpus [34] was used for training when building the model. Results showed that the model had 97.8% precision, 98.5% recall, and 98.1% F_1-score [32].

AttaCut

Inspired by DeepCut, AttaCut [38] was designed to speed up the segmentation process. It achieves this goal by disabling neurons in DeepCut's layers, and at the end, its speed performance increased by 6 times. Results on the in-domain dataset revealed 89%-91% F_1-scores, and those on the out-domain dataset revealed 63%-81% F_1-scores [32].

Word2Vec

The well-known word embedding or word vector is an approach for representing documents and words. It is defined as a numeric vector input that allows words with similar meanings to have the same representation. Word2Vec is one of the models used to transform words and represent them as distributed vectors. It is commonly used as an initial step in predictive models for tasks related to semantics and information retrieval [24,39,40]. A neural network is then used to generate word embeddings as weights (by considering relationships between words and their surrounding context). Word2Vec generates word vectors using 2 main algorithms: continuous bag of words (CBOW) and skip-gram [41,42]. The CBOW and skip-gram architectures are shown in Figures 2 and 3, respectively.

XSL•FO
RenderX

Figure 2. Continuous bag of words model architecture: one of the algorithms inside Word2Vec to produce the word vector of the current word by considering surrounding words.

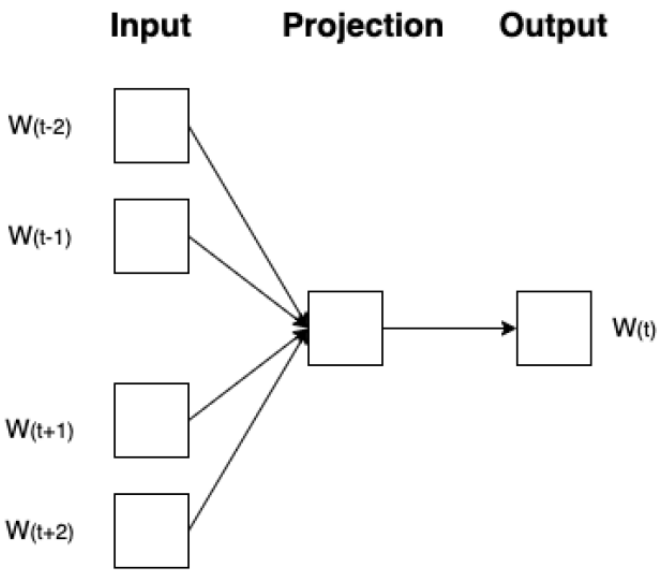

Figure 3. Skip-gram model architecture: one of the algorithms inside Word2Vec to produce word vectors for the surrounding words by considering the current word.

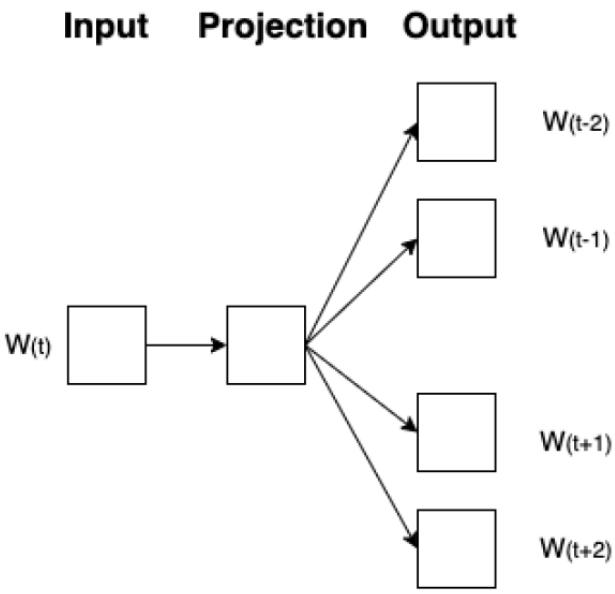

Skip-gram is slower than CBOW in training and has lower performance on frequent words, but it is better suited for rare words or phrases. Since our data included medical terminology, which can contain rare words and phrases, we focused on Word2Vec by using skip-gram.

The skip-gram model generates word-context pairs (w, c) from text data as input for a neural network.

$$(w, c) = (w_1, c_1), (w_2, c_2), \ldots, (w_n, c_n) \quad \textbf{(1)}$$

$$w_i \in V_w, c_i \in V_c$$

where V_w is the input vector (center word) and V_c is the output vector (context word) that represent the word and context vocabularies with sizes S_w and S_c, respectively. The goal of the skip-gram model is to learn the distribution of each word (based on context) by maximizing the probability of context words, as shown below.

XSL•FO
RenderX

where E and F refer to the parameter matrices of ($S_w \times d$) and ($S_c \times d$), respectively, and d refers to the dimensionality of the embedding vector space.

Deep Learning Models

As mentioned in the literature review section, deep learning methods have demonstrated strong performance in image classification and speech recognition. In recent years, they have also been widely adopted in NLP and have achieved good performance as well [26]. In NLP tasks, the deep learning models widely used include GRU, LSTM, and CNN [11]. Therefore, we focused on these deep learning models for the detection of knee OA.

GRU Model

GRU is a type of RNN specifically designed to address the vanishing gradient problem when capturing long-term dependencies in RNNs. It is also very efficient in processing sequential or time series data in NLP tasks. The structure of GRU consists of an update gate and a reset gate [26,43]. GRU uses gating mechanisms to control which information should be passed to the output or remain in the network. GRU model architecture is illustrated in Figure 4.

Figure 4. Gated recurrent unit model architecture: one of the recurrent neural network types. Its structure consists of input state (x_t), hidden state (h_t), output state (o_t), and update gate (z_t).

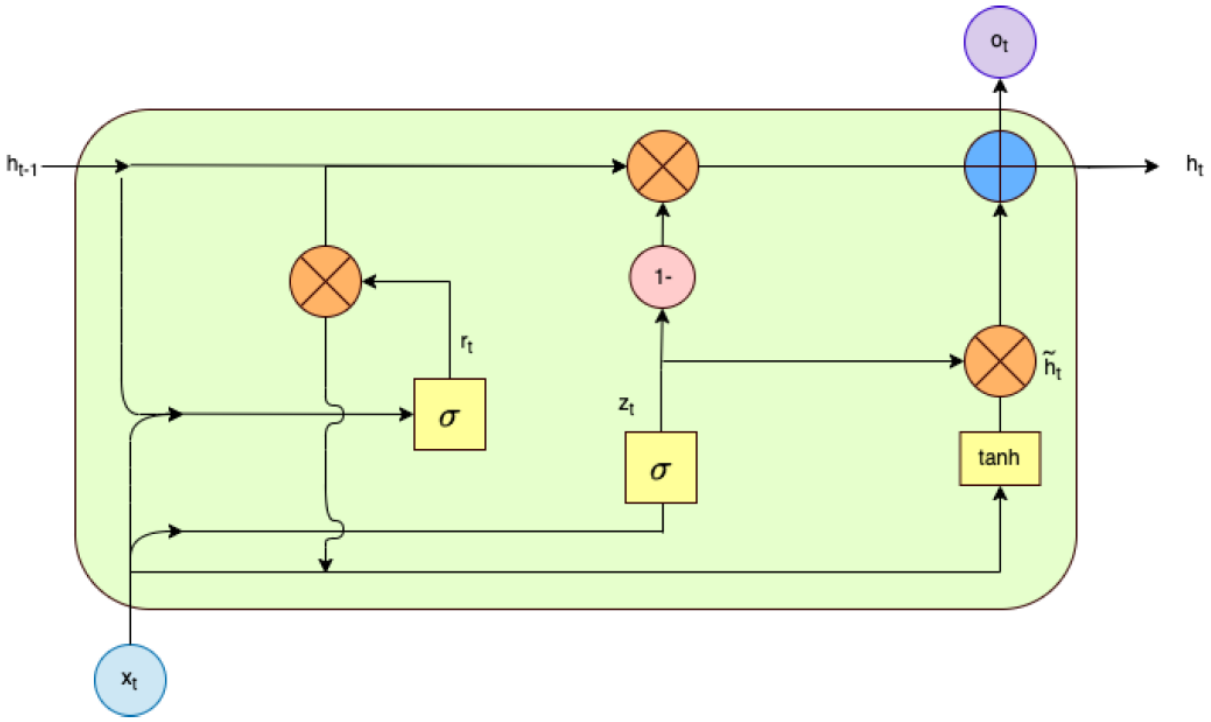

The GRU structure is illustrated in Figure 4. In the equations listed below, the input x_t (the sequence vector of the current word at time t) is the input for the hidden state h_t.

$$r_t = \sigma(W_r x_r + U_r h_{t-1} + b_r) \quad \textbf{(3)}$$

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \quad \textbf{(4)}$$

where the update gate and the reset gate are represented by z_t and r_t, respectively.The previous state is represented by h_{t-1}, while the candidate state is represented by at time t. The sigmoid and hyperbolic tangent functions are represented by σ and *tanh*, respectively. b_z, b_h, and b_r are all biases, whereas the operator is element-wise multiplication.

LSTM Model

LSTM is considered among the most widely adopted RNN types for sequential data-related tasks. It is designed to solve the problem of the traditional RNN model in capturing long-term dependencies, similar to GRU. The structure is shown in Figure 5.

XSL•FO
RenderX

Figure 5. Long short-term memory model architecture: one of the recurrent neural network types. Its structure consists of input state (x_t), hidden state (h_t), output state (o_t), forget gate (z_t), and memory cell (c_t).

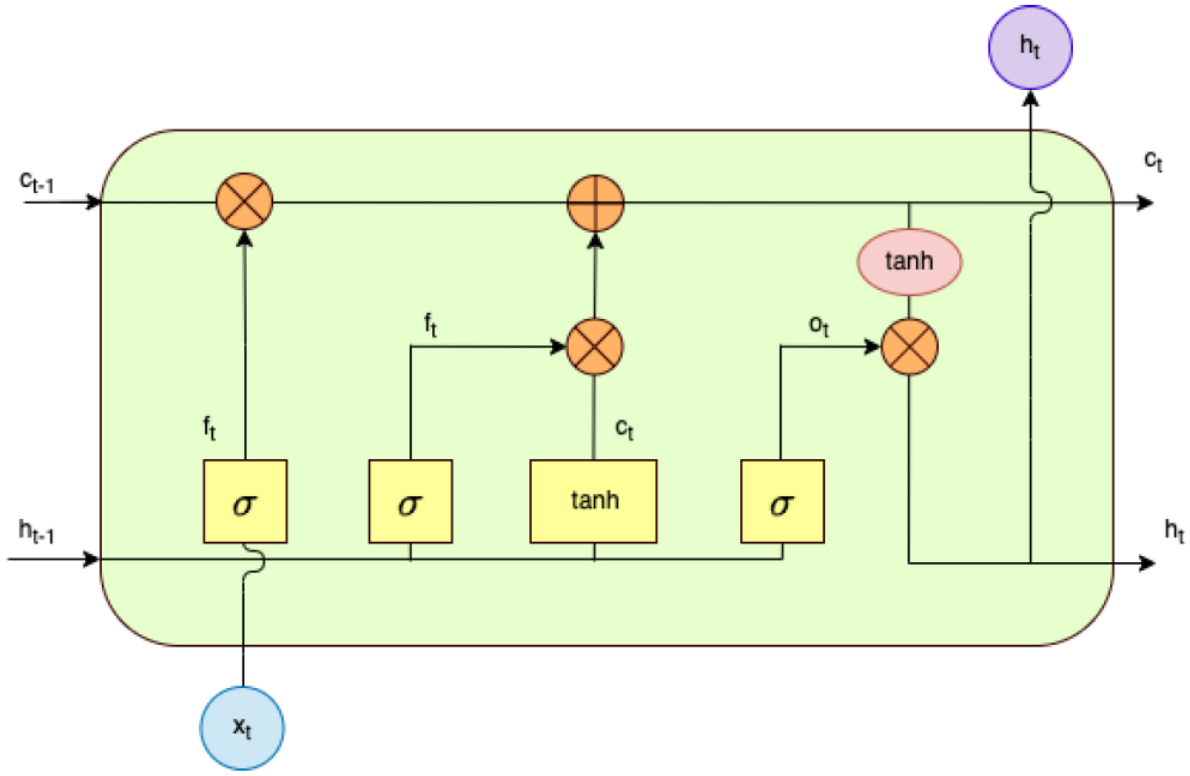

Often, RNNs face the vanishing gradient problem when trying to capture long-term dependencies. Consequently, LSTM is designed to solve this issue by using some internal gates [44]. The LSTM architecture consists of 3 gates (input gate, output gate, and forget gate) along with a memory cell [45]. The goal of the memory cell is to store information over long sequences by capturing previous words (c_{t-1}, h_{t-1}), with some being dropped (forgotten) by the forget gate, which regulates the flow of information in (x) and out (h) of the cell by the 3 gates. The input gate, forget gate, output gate, and cell output can be illustrated as follows:

$$f_t = \sigma(W^{(f)}x_t + U^{(f)}h_{t-1} + b^{(f)}) \quad \textbf{(7)}$$

$$i_t = \sigma(W^{(i)}x_t + U^{(i)}h_{t-1} + b^{(i)}) \quad \textbf{(8)}$$

$$o_t = \sigma(W^{(o)}x_t + U^{(o)}h_{t-1} + b^{(o)}) \quad \textbf{(9)}$$

$$c_t = f_t \boxtimes c_{t-1} + i_t \boxtimes tanh\,(W^{(c)}x_t + U^{(c)}h_{t-1} + b^{(c)}) \quad \textbf{(10)}$$

$$h_t = o_t \boxtimes tanh\,(c_t) \quad \textbf{(11)}$$

where the input is represented by $x_t \in R^d$ at time step t, while d is the feature dimension for each word. The element-wise sigmoid function and the element-wise dot product are represented by σ and ⊠, respectively. The memory cell represented by c_t aims to solve the problem of vanishing gradients, which facilitates the learning of long-term dependencies. The forget gate, represented by f_t, is used to decide what information to store in the memory cell, whereas the input gate and output gate, represented by i_t and o_t, respectively, aim to control the input and output of the memory cell, respectively. Finally, the output of LSTM is represented by h_t.

However, LSTM still has some limitations, since it can only read in 1 direction (forward direction). In this case, the results may show low classification performance because some contextual information from future words may be ignored by considering just 1 forward direction. Thus, Alharbi et al [14] proposed a solution to address the limitations of LSTM, which is called BiLSTM. BiLSTM can capture long-term dependencies from both the past and the future of a sequence (2 directions), which is important for understanding the full context of the sequence. As an extension of LSTM, the BiLSTM network processes both the forward and backward directions of a sequence to capture dependencies [24,46,47]. The BiLSTM model architecture is illustrated in Figure 6.

XSL•FO
RenderX

Figure 6. Bidirectional long short-term memory model architecture: one of the recurrent neural network types. Its structure consists of 2 long short-term memory parts: one processes input sequences in the forward direction, and the other processes input sequences in the backward direction. LSTM: long short-term memory.

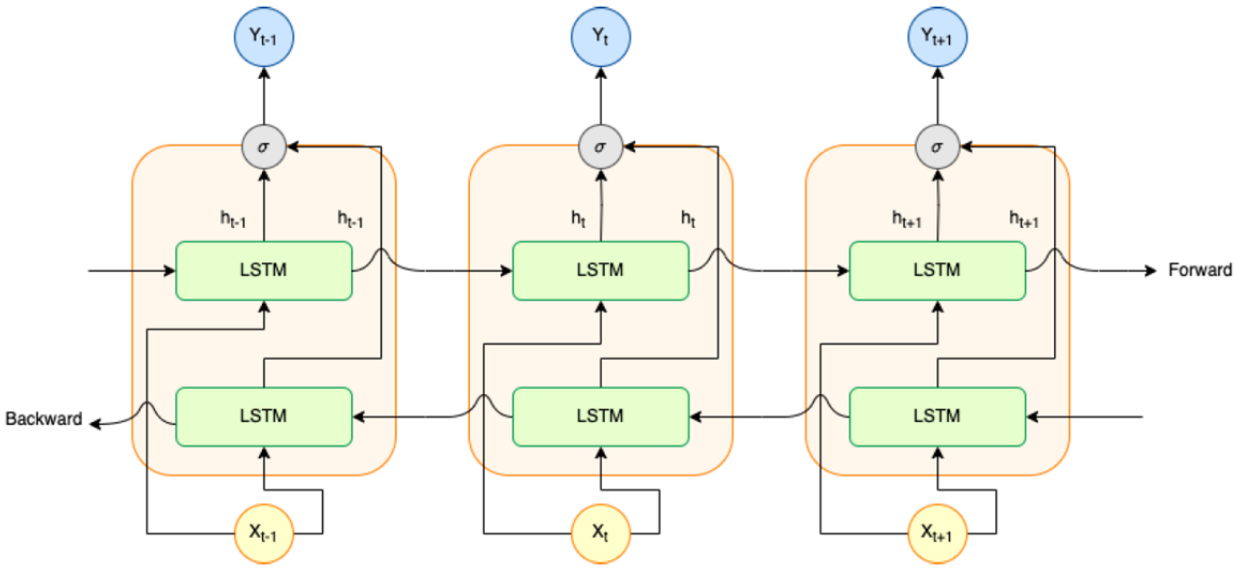

CNN Model

CNNs are generally developed for image classification. Currently, CNNs have shown strong performance in various NLP tasks, particularly in text classification. They are often considered appropriate for short text classification tasks [47-49]. A CNN model typically consists of several layers, which can be separated into 3 groups. First, the convolutional layer transforms the text data to obtain a feature map to detect the presence of a specific feature. Second, the pooling layer follows the convolutional layer, and it is designed to reduce the input data while preserving important information further. Third, the final stage of the CNN is the fully connected layer, which is responsible for using the features extracted in the previous layers for classification. The CNN architecture is illustrated in Figure 7.

Embedding layer: Let $w_i \in R^m$ be a word vector with m-dimensional representation associated with a word in a sentence. A sentence of length n is represented as follows:

$$w_{1:n} = w_1 \boxed{\times} w_2 \boxed{\times} \ldots \boxed{\times} w_n \quad \textbf{(12)}$$

where $\boxed{\times}$ denotes the concatenation operator.

Convolutional layer: The filter is represented by $x \in R^{h \times m}$, where h is the number of words in a window and m is the dimension of a word vector. A feature map c = [c_1, c_2, …, c_{n-h+1}] can be generated by:

$$c_i = f(x \cdot w_{i:i+h-1} + b) \quad \textbf{(13)}$$

where $b \in R$ denotes a bias term and f is a nonlinear function.

Pooling layer: The pooling operation was applied for the respective filter to select the most important feature from each feature map. We used the max pooling operation to take the maximum value in the feature map.

$$\boxed{\times} = max\ \{c\} \quad \textbf{(14)}$$

Fully connected layer: The selected features from the pooling layer are flattened into a single vector. It is passed to a fully connected layer to assign a probability for a specific label.

$$\boxed{\times}$$

XSL•FO
RenderX

Figure 7. Convolutional neural network model architecture. Its structure consists of an embedding layer (input vectors), a 1D convolutional layer to extract local patterns, a pooling layer to select the most important features from the convolutional layer, a fully connected layer to prepare for the classification, and an output layer to calculate the output from the fully connected layer to the final binary classification result.

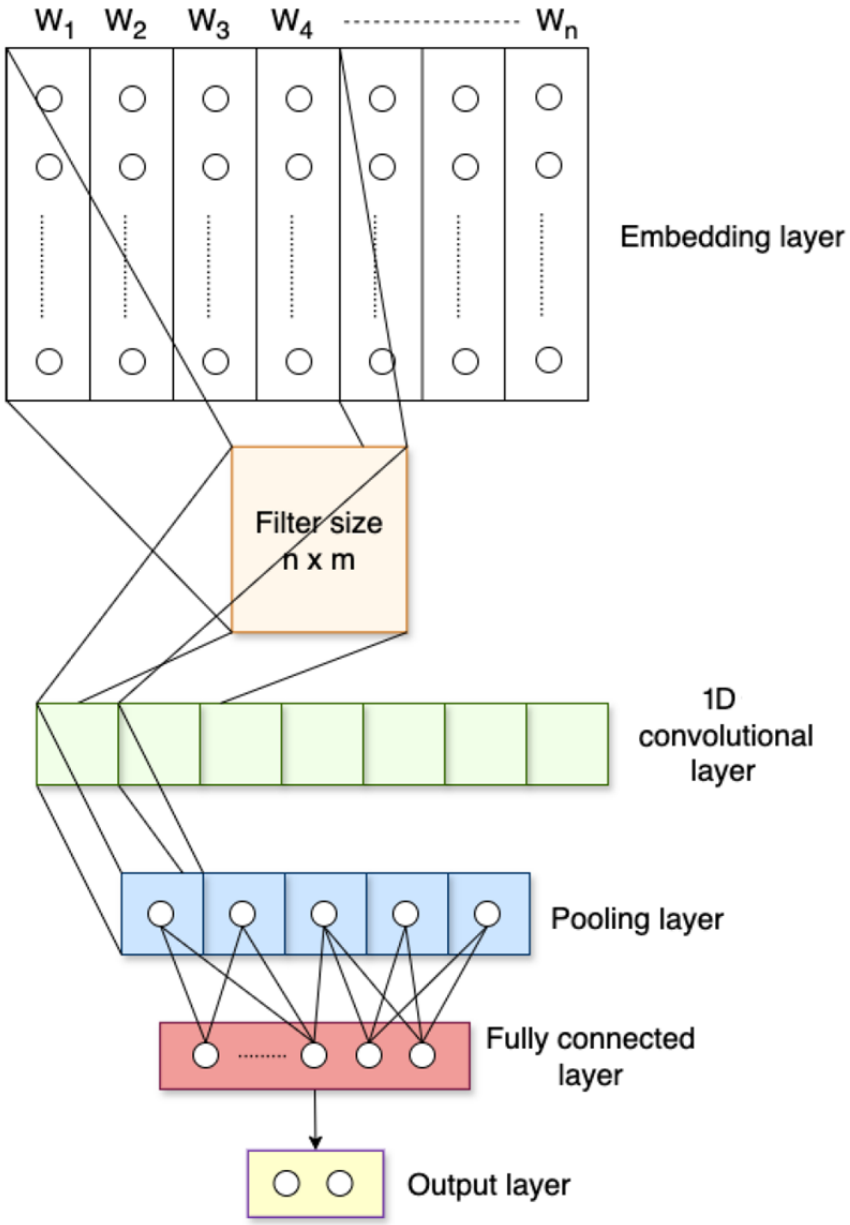

Algorithmic Details

According to the deep learning models mentioned in the previous subsections, the key procedures can be described as follows: (1) import a dataset of reported symptoms of diseases (that are in Excel files) to our program; (2) perform data cleaning in sentences from Excel files; (3) create a vocabulary set with indexes from sentences to transform the sentences into vectors using a tokenizer; (4) create a word embedding from unique vocabulary for pretraining by using Word2Vec (skip-gram approach); (5) identify the longest vector with the maximum number of words in a sentence and adjust all the vectors to the same length as the longest (since the word count of words in a sentence can vary, the size of each vector can be different); (6) pad such vectors with fewer words than the maximum one with zeroes; (7) separate training data, validation data, and test data; (8) build the proposed models with BiLSTM, GRU, and CNN; (9) train the models according to the range given by the hyperparameters; and (10) evaluate the models.

We present the pseudocode of our methods, in which algorithm 1 was designed for BiLSTM and GRU (Textbox 1) and algorithm 2 was designed for CNN (Textbox 2). For CNN, we used a 1D convolution layer, and the feature map was selected by max pooling in convolutional features with kernel sizes of 4, 6, and 8. We used dropout and rectified linear unit (ReLU) as activation functions. Lastly, we used the Softmax activation function in a fully connected layer.

XSL•FO
RenderX

Textbox 1. Pseudocodes of bidirectional long short-term memory and gated recurrent unit for knee osteoarthritis predictions.

Algorithm 1: Pseudocodes of bidirectional long short-term memory (BiLSTM) and gated recurrent unit (GRU)

Input: A collection of medical symptom reports for diseases in Excel files (Let D be one report and $S(i)$ be an ith sentence: $D = \{S(1), S(2), S(3), \ldots, S(i)\}, i \in N$)

Output: Prediction results/performance of BiLSTM and GRU

1: Import the reports of knee osteoarthritis data from Excel files into our program

2: while i in (1, n) do;

Select characters (that are Thai alphabets only) in the sentences

Tokenize by Word_Tokenize, DeepCut, AttaCut

Remove Thai stop words in sentences

Edit missing words with WOMAC (Western Ontario and McMaster Universities Osteoarthritis Index)-based processing approach (cutoff = 0.85)

end while

3: Create a set of vocabulary along with indexes by tokenizer from TensorFlow

4. Generate word embedding by Word2Vec (skip-gram approach)

vector_size = 128, epochs = 1000

5: Identify the max-length vector and adjust all vectors accordingly

6: Pad all smaller vectors with zeroes

7: Separate training, validation, and test datasets [training dataset: 80%, validation dataset: 10%, and test dataset: 10%]

8: Build the BiLSTM or GRU model

Set vocab_size, drop_out, BatchNomalization, and activation function parameters

9: while i <= epochs; // e = epoch numbers

execute the attention module;

import input vector;

Dropout; // (0.5)

BatchNormalization;

Update/optimize parameters with Adam optimization to minimize the loss function in the model

Verify the BiLSTM or GRU model using

validation data;

end while

10: Evaluate model by the test data;

Return the performance of BiLSTM and GRU

XSL•FO
RenderX

Textbox 2. Pseudocodes of convolutional neural network for knee osteoarthritis predictions.

Algorithm 2: Pseudocodes of convolutional neural network (CNN)

Input: A collection of medical symptom reports for diseases in Excel files (Let D be one report and $S(i)$ be an ith sentence: $D = \{S(1), S(2), S(3), \ldots, S(i)\}, i \in N$)

Output: The result of the prediction from CNN

1: Import the report of knee osteoarthritis data

2: while i in (1, n) do;

Select characters (that are Thai alphabets only) in the sentences

Tokenize by Word_Tokenize, DeepCut, AttaCut

Remove Thai stop words in sentences

Edit missing words with the WOMAC (Western Ontario and McMaster Universities Osteoarthritis Index)-based processing approach (cutoff = 0.85)

end while

3: Create a set of vocabulary along with indexes by tokenizer from TensorFlow

4. Generate word embedding by Word2Vec (skip-gram approach)

vector_size = 128, epochs = 1000

5: Identify the max-length vector and adjust all vectors accordingly

6: Pad all smaller vectors with zeroes

7: Separate training, validation, and test datasets [training dataset: 80%, validation dataset: 10%, and test dataset: 10%]

8: Build the CNN model

Set vocab_size, drop_out, BatchNormalization, and activation function parameters

9: while i <= epochs; // e = epoch numbers

execute the attention module;

import input vector;

convolutional layer; // kernel_size = 4, 6, 8

Dropout; // (0.5)

Max pooling;

Fully connected layer;

Update/optimize parameters with Adam optimization to minimize the loss function in the model

Verify the CNN model using validation data;

end while

10: Evaluate model by using test data;

Return the performance of the CNN

Evaluation Metrics

We evaluated our models by using 5 metrics: accuracy, sensitivity (or recall), specificity, precision, and F_1-score. These parameters are defined as follows:

Accuracy = (TP + TN) / (TP + FN + FP + TN) **(16)**

Sensitivity/recall = TP / (TP + FN) **(17)**

Specificity = TN / (TN + FP) **(18)**

Precision = TP / (TP + FP) **(19)**

F_1-score = (2 × [precision × recall]) / (precision + recall) **(20)**

where true positive (TP) is the total number of patients classified and tagged in the dataset as OA, true negative (TN) is the total number of patients classified and tagged in the dataset as non-OA, false negative (FN) is the total number of patients classified as non-OA but tagged in the dataset as OA, and false positive (FP) is the total number of patients classified as OA but tagged in the dataset as non-OA. Accuracy is the ratio of true results of both OA and non-OA among the total number of patients examined. Sensitivity is the proportion of patients classified as OA by the model among all OA-tagged data. Specificity is the proportion of patients classified as non-OA by the model among all non-OA–tagged data. Precision is the percentage of patients classified as OA by the model, and F_1-score is the average of recall and precision.

In summary, our contributions are as follows. First, our approach proposes to use clinicians' or doctors' notes (instead of x-ray or MRI images and statistical laboratory results), which contain

XSL•FO
RenderX

meaningful messages from patients who have the same symptoms, to help in diagnosis. Second, we used the power of text classification through NLP in diagnosis, which can be used in complement with other diagnostic methods (such as using medical images or biospecimen laboratory results) to enhance accuracy and achieve a timely diagnosis. Third, we introduced the high potential of textual analysis based on doctors' notes of symptoms to the field of knee OA as an alternative, which is new, and combined it with the WOMAC (dictionary/key terms for knee OA). Fourth, although our approach is performed on Thai datasets, the methodology can be adapted or repeated for diagnosing knee OA in other languages. Fifth, we combined and compared the text classification methods with deep learning models, including GRU, BiLSTM, and CNN, and at the end, we were able to identify the best combination in diagnosing knee OA in our approach. Sixth, to further tune our approach for better performance, we fine-tuned our method with our technique in adjusting misleading or wrong words by using real experts and physicians, as well as the Thai-verified version of the WOMAC.

Results

Records

To ensure data quality, we removed 888 (incomplete) records from 6737 knee OA records, leaving 5849 records containing complete patient reports. The dataset was separated into training, validation, and test sets comprising 4788, 598, and 463 records, respectively. We used the GRU, BiLSTM, and CNN models for knee OA prediction.

Hyperparameters of GRU, BiLSTM, and CNN

The hyperparameters were adjusted for each model used in the experiments, including the quantity of neurons in each layer, the choice of the activation function, the optimization method, and the regularization techniques. In selecting hyperparameters for our experiments, we referred to relevant research studies in the literature that demonstrated strong performance [10,14,44,48,50]. The hyperparameters of each model are shown in Table 1.

Table 1. Hyperparameters of gated recurrent unit, bidirectional long short-term memory, and convolutional neural network.

Hyperparameter[a]	GRU[b]	BiLSTM[c]	CNN[d]
Filter	__[e]	—	32
Kernel	—	—	4, 6, 8
Activation function (last layer)	ReLU[f] (Softmax)	ReLU (Softmax)	ReLU (Sigmoid)
Dropout	0.5	0.5	0.5
Epoch	50	50	50
Batch size	128	128	128
Loss function	categorical_crossentropy	categorical_crossentropy	binary_crossentropy
Optimizer/learning rate	Adam/0.0001	Adam/0.0001	Adam/0.0001
Regularization	Batch normalization	Batch normalization	Batch normalization

[a]BiLSTM: bidirectional long short-term memory.

[b]GRU: gated recurrent unit.

[c]CNN: convolutional neural network.

For GRU and BiLSTM, the weights were passed to the output layer with a Softmax function to diagnose knee OA. For CNN, the weights were passed to the output layer with a sigmoid function to diagnose knee OA. For all models, the hidden layer was trained and regularized using ReLU and batch normalization, respectively. To reduce the overfitting problem, we applied dropout [14], with a probability of 0.5 after each model while training with the Adam optimizer [10], using a learning rate of 0.0001 to optimize the model parameters. The experiments have two parts: (1) experiments that use the baseline approach and (2) experiments that use our WOMAC-based processing approach. In Tables 2 and 3, we present the experimental results of the models "before" processing the missing or edited words in sentences by using our WOMAC-based processing approach. The best-performing model was the BiLSTM model that used the tokenization of the AttaCut algorithm, with the following metrics: area under the curve (AUC), 0.85; accuracy, 0.87; precision, 0.85; sensitivity, 0.95; specificity, 0.76; and F_1-score, 0.90.

In Tables 4 and 5, we present the experimental results of the models "after" processing the missing or edited words in sentences by using our WOMAC-based processing approach. The best-performing model was the BiLSTM model that used the tokenization of the AttaCut algorithm, with the following metrics: AUC, 0.91; accuracy, 0.91; precision, 0.91; sensitivity, 0.94; specificity, 0.87; and F_1-score, 0.93.

XSL•FO
RenderX

Table 2. Results of the baseline approach (n=463) in terms of true positive, false positive, false negative, true negative, and runtime for each model.

Model	TP[a], n (%)	FP[b], n (%)	FN[c], n (%)	TN[d], n (%)	Runtime (s)
Word_Tokenize (n=463)					
BiLSTM[e]	248 (53.6)	36 (7.8)	26 (5.6)	153 (33.0)	389.01
GRU[f]	249 (53.8)	40 (8.6)	25 (5.4)	149 (32.2)	303.88
CNN[g]	252 (54.4)	43 (9.3)	22 (4.8)	146 (31.5)	138.72
DeepCut (n=463)					
BiLSTM	258 (55.7)	47 (10.2)	16 (3.5)	142 (30.7)	476.81
GRU	249 (53.8)	46 (9.9)	25 (5.4)	143 (30.9)	411.99
CNN	258 (55.7)	49 (10.6)	16 (3.5)	140 (30.2)	152.16
AttaCut (n=463)					
BiLSTM	260 (56.2)	46 (9.9)	14 (3.0)	143 (30.9)	467.10
GRU	251 (54.2)	40 (8.6)	23 (5.0)	149 (32.2)	387.65
CNN	248 (53.6)	35 (7.6)	26 (5.6)	154 (33.3)	146.02

[a]TP: true positive.
[b]FP: false positive.
[c]FN: false negative.
[d]TN: true negative.
[e]BiLSTM: bidirectional long short-term memory.
[f]GRU: gated recurrent unit.
[g]CNN: convolutional neural network.

Table 3. Results of the baseline approach (n=463) in terms of accuracy, precision, sensitivity, specificity, and F1-score for each model.

Model	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F_1-score (%)
Word_Tokenize					
BiLSTM[a]	0.86	0.85	0.92	0.77	0.89
GRU[b]	0.86	0.86	0.91	0.79	0.89
CNN[c]	0.87	0.87	0.91	0.91	0.89
DeepCut					
BiLSTM	0.86	0.85	0.85	0.94	0.75
GRU	0.85	0.84	0.83	0.91	0.84
CNN	0.86	0.85	0.85	0.94	0.75
AttaCut					
BiLSTM	0.87	0.85	0.95	0.76	0.90
GRU	0.86	0.86	0.92	0.79	0.89
CNN	0.87	0.88	0.91	0.81	0.89

[a]BiLSTM: bidirectional long short-term memory.
[b]GRU: gated recurrent unit.
[c]CNN: convolutional neural network.

XSL•FO
RenderX

Table 4. Results of the Western Ontario and McMaster Universities Osteoarthritis Index–based processing approach (n=463) in terms of true positive, false positive, false negative, true negative, and runtime for each model.

Model	TP[a], n (%)	FP[b], n (%)	FN[c], n (%)	TN[d], n (%)	Runtime (s)
Word_Tokenize (n=463)					
BiLSTM[e]	252 (54.4)	24 (5.2)	22 (4.8)	165 (35.6)	359.62
GRU[f]	256 (55.3)	36 (7.8)	18 (3.9)	153 (33.0)	294.49
CNN[g]	254 (54.9)	25 (5.4)	20 (4.3)	164 (35.4)	128.76
DeepCut (n=463)					
BiLSTM	251 (54.2)	25 (5.4)	23 (5.0)	164 (35.4)	429.84
GRU	250 (54.0)	27 (5.8)	24 (5.2)	162 (35.0)	367.58
CNN	258 (55.7)	36 (7.8)	16 (3.5)	153 (33.0)	129.69
AttaCut (n=463)					
BiLSTM	257 (55.5)	24 (5.2)	17 (3.7)	165 (35.6)	397.51
GRU	254 (54.9)	31 (6.7)	20 (4.3)	158 (34.1)	343.44
CNN	251 (54.2)	22 (4.8)	23 (5.0)	167 (36.1)	117.44

[a]TP: true positive.
[b]FP: false positive.
[c]FN: false negative.
[d]TN: true negative.
[e]BiLSTM: bidirectional long short-term memory.
[f]GRU: gated recurrent unit.
[g]CNN: convolutional neural network.

Table 5. Results of the Western Ontario and McMaster Universities Osteoarthritis Index–based processing approach (n=463) in terms of accuracy, precision, sensitivity, specificity, and F1-score for each model.

Model	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F_1-score (%)
Word_Tokenize					
BiLSTM[a]	0.90	0.91	0.93	0.87	0.92
GRU[b]	0.88	0.88	0.93	0.81	0.90
CNN[c]	0.90	0.91	0.92	0.87	0.92
DeepCut					
BiLSTM	0.89	0.88	0.94	0.81	0.91
GRU	0.89	0.90	0.91	0.86	0.91
CNN	0.90	0.91	0.92	0.87	0.91
AttaCut					
BiLSTM	0.91	0.91	0.94	0.87	0.93
GRU	0.89	0.89	0.93	0.84	0.91
CNN	0.90	0.92	0.92	0.88	0.92

[a]BiLSTM: bidirectional long short-term memory.
[b]GRU: gated recurrent unit.
[c]CNN: convolutional neural network.

XSL•FO
RenderX

Discussion

Overview

While conventional medical data, such as images and statistical or laboratory data, have been extensively used, the use of textual data through NLP for knee OA diagnosis is still a relatively new approach. As mentioned previously, we propose an NLP model to detect the occurrence of knee OA by incorporating deep learning models, including GRU, BiLSTM, and CNN, based on textual data of knee OA symptoms reported in doctors' notes. We discuss principal findings, comparison with prior work, strengths, limitations, and future work in the following subsections.

Principal Findings

A key finding in this study is that our method, using only textual data from doctors' notes for knee OA prediction, performed better than previous studies that used image data and statistical or laboratory data. Moreover, in our experiments, the models trained very quickly due to the nature of our textual data (short sentences), which, in turn, can reduce the model's runtime. We also found that the best tokenization algorithm and model for this experiment was BiLSTM combined with AttaCut tokenization.

Although CNNs can extract relevant local features in sentences where specific combinations of words (local features) are important, it was not designed to handle sequential data like our data. Therefore, CNNs may not effectively handle variable-length sequences. GRUs have a simpler architecture and train faster than BiLSTM. However, GRUs may not perform as well as BiLSTM on very complex sequence data. This may be because BiLSTM can address the issues of complex sequences or long-distance dependencies in sentences better, as shown in our experiments, and thus, BiLSTM had the best performance. The best-performing model "before" applying our WOMAC-based processing approach was BiLSTM (AUC, 0.85; accuracy, 0.87; precision, 0.85; sensitivity, 0.95; specificity, 0.76; F_1-score, 0.90), and after applying our approach, there was an improvement in the results with BiLSTM (AUC, 0.91; accuracy, 0.91; precision, 0.91; sensitivity, 0.94; specificity, 0.87; F_1-score, 0.93). Therefore, our WOMAC-based processing approach can increase the performance of models. A comparison of the baseline model and the model using the WOMAC-based processing approach is shown in Table 6.

Table 6. Comparison between the baseline approach and Western Ontario and McMaster Universities Osteoarthritis Index–based processing approach, and the improvement rate from baseline for each model.

Model	Baseline			WOMAC[a]-based processing approach			Improvement rate		
	F_1-score (%)	Macro avg[b] (%)	Weighted avg[c] (%)	F_1-score (%)	Macro avg[b] (%)	Weighted avg[c] (%)	F_1-score (%)	Macro avg[b] (%)	Weighted avg[c] (%)
Word_Tokenize									
BiLSTM[d]	0.89	0.88	0.89	0.92	0.90	0.90	0.03	0.02	0.01
GRU[e]	0.88	0.85	0.86	0.90	0.88	0.88	0.02	0.03	0.02
CNN[f]	0.89	0.86	0.87	0.92	0.90	0.90	0.03	0.04	0.03
DeepCut									
BiLSTM	0.89	0.85	0.86	0.91	0.88	0.89	0.02	0.03	0.03
GRU	0.88	0.84	0.85	0.91	0.89	0.89	0.03	0.04	0.04
CNN	0.89	0.85	0.86	0.91	0.89	0.90	0.02	0.04	0.04
AttaCut									
BiLSTM	0.90	0.86	0.87	0.93	0.91	0.91	0.03	0.05	0.04
GRU	0.89	0.86	0.86	0.91	0.88	0.89	0.02	0.02	0.03
CNN	0.89	0.86	0.87	0.92	0.88	0.89	0.03	0.02	0.02

[a]AUC: area under the curve.

[b]MAE: mean absolute error.

[c]Not applicable.

Comparison With Prior Work

In comparisons of our study with prior studies, we considered the primary data source used (image data or statistical data). Sources of data are presented in Multimedia Appendix 4. A detailed comparison of the model in our study with the models in previous studies is shown in Table 7. When compared to prior models that used image data [1,17,22], our model showed better accuracy [1,17] and a higher AUC [22]. When compared to prior models that used statistical data [2,6,19], our model showed higher accuracy, sensitivity, specificity, and AUC [2,6]. However, the AUC of our model was lower than that of the model in the study by Halilaj et al [19].

XSL•FO
RenderX

Table 7. Comparison of our model with models in previous studies.

Study (data source)	Number of patients	Accuracy (%)	AUC[a] (%)	MAE[b] (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F_1-score (%)
Brahim et al [1] (images)	1024	0.83	__[c]	—	—	0.87	0.81	—
Shen et al [2] (stats)	195	—	—	0.11	—	0.71	0.67	0.69
Halilaj et al [19] (stats)	3924	—	0.95	—	—	—	—	—
Yong et al [17] (images)	4796	0.88	—	0.33	—	—	—	—
Lim et al [6] (stats)	5749	0.72	0.77	—	—	0.67	0.73	—
Tiulpin et al [22] (images)	4928	—	0.79	—	—	—	—	—
Our study	5849	0.91	0.91	0.09	0.91	0.94	0.87	0.93

[a]AUC: area under the curve.

[b]MAE: mean absolute error.

[c]Not applicable.

Strengths

The overall performance of our model surpasses that of models in previous studies. Our proposed method introduces a novel approach to knee OA prediction research. Our research has 3 strengths. First, our model uses textual data, a valuable information source for AI learning. While textual data is critical, investigation and verification steps are also important. Second, since the nature of our dataset is textual and typically consists of short sentences, it does not require much processing time; hence, our model can be trained very quickly. Third, our patient dataset is larger than the datasets in previous studies (Table 7).

Limitations

A limitation of our method is that it currently works only with the Thai language. As a result, it requires local application programming interfaces (APIs) that can handle the Thai language for processing. However, these local APIs may not have a programmer community as large as that for English or global APIs, and therefore, they may not be as productive. Our work focused on using textual data only in our model, and thus, we did not incorporate other types of data, such as image data and statistical or laboratory data, into our model, which could have improved diagnostic performance.

Future Work

In the future, we aim to improve our proposed method by incorporating image data and statistical data to achieve even better knee OA prediction. Additionally, we want to extend our work to make the model more precise by predicting the stage of knee OA (there are 5 stages based on WOMAC scores, ranging from 0 to 4) [12,16,18]. Furthermore, our proposed method can be applied to other medical conditions or diseases, such as depression.

Conclusions

We proposed a method to predict the occurrence of knee OA by using clinicians' or doctors' notes. This method has better prediction performance than other conventional methods that use images or statistical laboratory data. We confirmed the possibility of using symptom texts reported by patients (recorded by doctors) to predict knee OA conditions. Moreover, we demonstrated that the text of symptom reports can be a valuable data source to predict or classify a disease (whether a particular knee will have OA progression or not). Furthermore, our method involving text data (mostly consisting of short sentences) can save a considerable amount of time compared with other methods that use image data and statistical laboratory data (patients must visit the hospital and undergo imaging or tests to generate such data).

Data Availability

The datasets used in this study are not publicly available, as they require institutional approval on a case-by-case basis from the relevant authorities. However, the contact information of the relevant authorities can be requested by contacting the corresponding author.

Conflicts of Interest

None declared.

Multimedia Appendix 1

XSL•FO
RenderX

CONSORT (Consolidated Standards of Reporting Trials) checklist.
[PDF File (Adobe PDF File), 66 KB - formative_v9i1e64536_app1.pdf]

Multimedia Appendix 2
CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) checklist.
[PDF File (Adobe PDF File), 1119 KB - formative_v9i1e64536_app2.pdf]

Multimedia Appendix 3
STARD (Standards for Reporting of Diagnostic Accuracy Studies) checklist.
[PDF File (Adobe PDF File), 195 KB - formative_v9i1e64536_app3.pdf]

Multimedia Appendix 4
Sources of data used according to the literature.
[PNG File , 57 KB - formative_v9i1e64536_app4.png]

References

1. Brahim A, Jennane R, Riad R, Janvier T, Khedher L, Toumi H, et al. A decision support tool for early detection of knee OsteoArthritis using X-ray imaging and machine learning: Data from the OsteoArthritis Initiative. Comput Med Imaging Graph 2019 Apr;73:11-18. [doi: 10.1016/j.compmedimag.2019.01.007] [Medline: 30784984]
2. Shen L, Yue S. A clinical model to predict the progression of knee osteoarthritis: data from Dryad. J Orthop Surg Res 2023 Aug 28;18(1):628 [FREE Full text] [doi: 10.1186/s13018-023-04118-4] [Medline: 37635226]
3. Farrow L, Ashcroft GP, Zhong M, Anderson L. Using artificial intelligence to revolutionise the patient care pathway in hip and knee arthroplasty (ARCHERY): Protocol for the development of a clinical prediction model. JMIR Res Protoc 2022 May 11;11(5):e37092 [FREE Full text] [doi: 10.2196/37092] [Medline: 35544289]
4. Chen C, Wu C, Chen CPC, Chung C, Chen S, Lee MS, et al. Predicting the risk of total hip replacement by using a deep learning algorithm on plain pelvic radiographs: Diagnostic study. JMIR Form Res 2023 Oct 20;7:e42788 [FREE Full text] [doi: 10.2196/42788] [Medline: 37862084]
5. Wang H, Luo P, Liu S, Liu Y, Zhang X, Li J. Effectiveness of internet-based telehealth programs in patients with hip or knee osteoarthritis: Systematic review and meta-analysis. J Med Internet Res 2024 Sep 30;26:e55576 [FREE Full text] [doi: 10.2196/55576] [Medline: 39348685]
6. Lim J, Kim J, Cheon S. A deep neural network-based method for early detection of osteoarthritis using statistical data. Int J Environ Res Public Health 2019 Apr 10;16(7):1281 [FREE Full text] [doi: 10.3390/ijerph16071281] [Medline: 30974803]
7. Nowinka Z, Alagha MA, Mahmoud K, Jones GG. Predicting depression in patients with knee osteoarthritis using machine learning: model development and validation study. JMIR Form Res 2022 Sep 13;6(9):e36130 [FREE Full text] [doi: 10.2196/36130] [Medline: 36099008]
8. Liu K, Chen L. Medical social media text classification integrating consumer health terminology. IEEE Access 2019;7:78185-78193. [doi: 10.1109/access.2019.2921938]
9. Helwan A, Azar D, Ozsahin DU. Medical Reports Summarization Using Text-To-Text Transformer. 2023 Presented at: 2023 Advances in Science and Engineering Technology International Conferences (ASET); February 20-23, 2023; Dubai, United Arab Emirates. [doi: 10.1109/aset56582.2023.10180671]
10. Wazrah AA, Alhumoud S. Sentiment analysis using stacked gated recurrent unit for Arabic tweets. IEEE Access 2021;9:137176-137187. [doi: 10.1109/access.2021.3114313]
11. Santur Y. Sentiment Analysis Based on Gated Recurrent Unit. 2019 Presented at: 2019 International Artificial Intelligence and Data Processing Symposium (IDAP); September 21-22, 2019; Malatya, Turkey. [doi: 10.1109/IDAP.2019.8875985]
12. Ashalatha B, T T, Geetha M, T S, Mittal M, K V. Predicting severity of osteoarthritis using recurrent convolutional neural networks (Rcnn) and medical imaging data. South Eastern European Journal of Public Health 2024 Sep 02:47-61. [doi: 10.70135/seejph.vi.894]
13. Al-Saqqa S, Awajan A. The Use of Word2vec Model in Sentiment Analysis: A Survey. In: AIRC '19: Proceedings of the 2019 International Conference on Artificial Intelligence, Robotics and Control. 2019 Presented at: 2019 International Conference on Artificial Intelligence, Robotics and Control; December 14-16, 2019; Cairo, Egypt. [doi: 10.1145/3388218.3388229]
14. Alharbi O. A deep learning approach combining CNN and Bi-LSTM with SVM classifier for Arabic sentiment analysis. International Journal of Advanced Computer Science and Applications 2021;12(6):0120618. [doi: 10.14569/ijacsa.2021.0120618]
15. Hampton T. Osteoarthritis initiative. JAMA 2004 Apr 28;291(16):1951. [doi: 10.1001/jama.291.16.1951-a]

XSL•FO
RenderX

16. Xin Teoh Y, Othmani A, Li Goh S, Usman J, Lai KW, Lai KW. Deciphering knee osteoarthritis diagnostic features with explainable artificial intelligence: A systematic review. IEEE Access 2024;12:109080-109108. [doi: 10.1109/access.2024.3439096]
17. Yong CW, Teo K, Murphy BP, Hum YC, Tee YK, Xia K, et al. Knee osteoarthritis severity classification with ordinal regression module. Multimed Tools Appl 2021 Jan 26;81(29):41497-41509. [doi: 10.1007/s11042-021-10557-0]
18. Helwan A, Azar D, Abdellatef H. An update on the knee osteoarthritis severity grading using wide residual learning. XST 2022 Sep 30;30(5):1009-1021. [doi: 10.3233/xst-221190]
19. Halilaj E, Le Y, Hicks JL, Hastie TJ, Delp SL. Modeling and predicting osteoarthritis progression: data from the osteoarthritis initiative. Osteoarthritis Cartilage 2018 Dec;26(12):1643-1650 [FREE Full text] [doi: 10.1016/j.joca.2018.08.003] [Medline: 30130590]
20. Kweon S, Kim Y, Jang M, Kim Y, Kim K, Choi S, et al. Data resource profile: the Korea National Health and Nutrition Examination Survey (KNHANES). Int J Epidemiol 2014 Feb;43(1):69-77 [FREE Full text] [doi: 10.1093/ije/dyt228] [Medline: 24585853]
21. Wang Y, You L, Chyr J, Lan L, Zhao W, Zhou Y, et al. Causal discovery in radiographic markers of knee osteoarthritis and prediction for knee osteoarthritis severity with attention-long short-term memory. Front Public Health 2020 Dec 18;8:604654 [FREE Full text] [doi: 10.3389/fpubh.2020.604654] [Medline: 33409263]
22. Tiulpin A, Klein S, Bierma-Zeinstra SMA, Thevenot J, Rahtu E, Meurs JV, et al. Multimodal machine learning-based knee osteoarthritis progression prediction from plain radiographs and clinical data. Sci Rep 2019 Dec 27;9(1):20038 [FREE Full text] [doi: 10.1038/s41598-019-56527-3] [Medline: 31882803]
23. Segal NA, Nevitt MC, Gross KD, Hietpas J, Glass NA, Lewis CE, et al. The Multicenter Osteoarthritis Study: opportunities for rehabilitation research. PM R 2013 Aug;5(8):647-654 [FREE Full text] [doi: 10.1016/j.pmrj.2013.04.014] [Medline: 23953013]
24. Jang B, Kim M, Harerimana G, Kang S, Kim JW. Bi-LSTM model to increase accuracy in text classification: Combining Word2vec CNN and attention mechanism. Applied Sciences 2020 Aug 24;10(17):5841. [doi: 10.3390/app10175841]
25. Yahia HS, Abdulazeez AM. Medical text classification based on convolutional neural network: a review. International Journal of Science and Business 2021;5(3):27-41 [FREE Full text]
26. Qing L, Linhong W, Xuehai D. A novel neural network-based method for medical text classification. Future Internet 2019 Dec 10;11(12):255 [FREE Full text] [doi: 10.3390/fi11120255]
27. Schulz KF, Altman DG, Moher D, CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomized trials. Ann Intern Med 2010 Jun 01;152(11):726-732 [FREE Full text] [doi: 10.7326/0003-4819-152-11-201006010-00232] [Medline: 20335313]
28. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of Web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126 [FREE Full text] [doi: 10.2196/jmir.1923] [Medline: 22209829]
29. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, STARD Group. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ 2015 Oct 28;351:h5527 [FREE Full text] [doi: 10.1136/bmj.h5527] [Medline: 26511519]
30. Wołk K, Marasek K. A Sentence Meaning Based Alignment Method for Parallel Text Corpora Preparation. In: Rocha Á, Correia A, Tan F, Stroetmann K, editors. New Perspectives in Information Systems and Technologies, Volume 1. Advances in Intelligent Systems and Computing, vol 275. Cham: Springer; 2014.
31. Jinks C, Jordan K, Croft P. Measuring the population impact of knee pain and disability with the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC). Pain 2002 Nov;100(1-2):55-64 [FREE Full text] [doi: 10.1016/s0304-3959(02)00239-7] [Medline: 12435459]
32. Arreerard R, Mander S, Piao S. Survey on Thai NLP Language Resources and Tools. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference. 2022 Presented at: Thirteenth Language Resources and Evaluation Conference; June 2022; Marseille, France.
33. Chormai P, Prasertsom P, Cheevaprawatdomrong J, Rutherford A. Syllable-based Neural Thai Word Segmentation. In: Proceedings of the 28th International Conference on Computational Linguistics. 2020 Presented at: 28th International Conference on Computational Linguistics; December 2020; Barcelona, Spain (Online). [doi: 10.18653/v1/2020.coling-main.407]
34. Boriboon M, Kriengket K, Chootrakool P, Phaholphinyo S, Purodakananda S, Thanakulwarapas T. BEST Corpus Development and Analysis. 2009 Presented at: 2009 International Conference on Asian Language Processing; December 07-09, 2009; Singapore. [doi: 10.1109/IALP.2009.76]
35. Phatthiyaphaibun W, Chaovavanich K, Polpanumas C, Suriyawongkul A, Lowphansirikul L, Chormai P, et al. PyThaiNLP: Thai Natural Language Processing in Python. arXiv. URL: https://arxiv.org/abs/2312.04649 [accessed 2025-06-12]
36. Theeramunkong T, Sornlertlamvanich V, Tanhermhong T, Chinnan W. Character cluster based Thai information retrieval. In: IRAL '00: Proceedings of the Fifth International Workshop on Information Retrieval with Asian Languages. 2000 Presented at: Fifth International Workshop on Information Retrieval with Asian Languages; September 30, 2000-October 1, 2000; Hong Kong China. [doi: 10.1145/355214.355225]

XSL•FO
RenderX

37. Seeha S, Bilan I, Sanchez LM, Huber J, Matuschek M, Schütze H. ThaiLMCut: Unsupervised Pretraining for Thai Word Segmentation. In: Proceedings of the Twelfth Language Resources and Evaluation Conference. 2020 Presented at: Twelfth Language Resources and Evaluation Conference; May 2020; Marseille, France.
38. Chormai P, Prasertsom P, Rutherford A. AttaCut: A Fast and Accurate Neural Thai Word Segmenter. arXiv. 2019. URL: https://arxiv.org/abs/1911.07056 [accessed 2025-06-12]
39. Shao Y, Taylor S, Marshall N, Morioka C, Zeng-Treitler Q. Clinical Text Classification with Word Embedding Features vs. Bag-of-Words Features. 2018 Presented at: 2018 IEEE International Conference on Big Data (Big Data); December 10-13, 2018; Seattle, WA, USA. [doi: 10.1109/bigdata.2018.8622345]
40. Jatnika D, Bijaksana M, Suryani A. Word2Vec model analysis for semantic similarities in English words. Procedia Computer Science 2019;157:160-167 [FREE Full text] [doi: 10.1016/j.procs.2019.08.153]
41. Hung PT, Yamanishi K. Word2vec skip-gram dimensionality selection via sequential normalized maximum likelihood. Entropy (Basel) 2021 Jul 31;23(8):997 [FREE Full text] [doi: 10.3390/e23080997] [Medline: 34441136]
42. Irsoy O, Benton A, Stratos K. Corrected CBOW Performs as well as Skip-gram. arXiv. 2020. URL: https://arxiv.org/abs/2012.15332 [accessed 2025-06-12]
43. Zulqarnain M, Ghazali R, Mohmad Hassim YM, Rehan M. Text classification based on gated recurrent unit combines with support vector machine. IJECE 2020 Aug 01;10(4):3734. [doi: 10.11591/ijece.v10i4.pp3734-3742]
44. Minaee S, Azimi E, Abdolrashidi A. Deep-Sentiment: Sentiment Analysis Using Ensemble of CNN and Bi-LSTM Models. arXiv. 2019. URL: https://arxiv.org/abs/1904.04206 [accessed 2025-06-12]
45. Otter DW, Medina JR, Kalita JK. A survey of the usages of deep learning for natural language processing. IEEE Trans Neural Netw Learning Syst 2021 Feb;32(2):604-624. [doi: 10.1109/tnnls.2020.2979670]
46. Xu G, Meng Y, Qiu X, Yu Z, Wu X. Sentiment analysis of comment texts based on BiLSTM. IEEE Access 2019;7:51522-51532. [doi: 10.1109/access.2019.2909919]
47. Yao L, Mao C, Luo Y. Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. BMC Med Inform Decis Mak 2019 Apr 04;19(Suppl 3):71 [FREE Full text] [doi: 10.1186/s12911-019-0781-4] [Medline: 30943960]
48. Guo B, Zhang C, Liu J, Ma X. Improving text classification with weighted word embeddings via a multi-channel TextCNN model. Neurocomputing 2019 Oct;363:366-374 [FREE Full text] [doi: 10.1016/j.neucom.2019.07.052]
49. Kim H, Jeong Y. Sentiment classification using convolutional neural networks. Applied Sciences 2019 Jun 07;9(11):2347. [doi: 10.3390/app9112347]
50. Salur MU, Aydin I. A novel hybrid deep learning model for sentiment classification. IEEE Access 2020;8:58080-58093. [doi: 10.1109/access.2020.2982538]

Abbreviations

AI: artificial intelligence
API: application programming interface
AUC: area under the curve
BiLSTM: bidirectional long short-term memory
CBOW: continuous bag of words
CNN: convolutional neural network
GRU: gated recurrent unit
KL: Kellgren/Lawrence
LSTM: long short-term memory
MRI: magnetic resonance imaging
NLP: natural language processing
OA: osteoarthritis
ReLU: rectified linear unit
RNN: recurrent neural network
WOMAC: Western Ontario and McMaster Universities Arthritis Index

XSL•FO
RenderX

Edited by A Mavragani; submitted 19.07.24; peer-reviewed by A Kimia, D Azar; comments to author 01.10.24; revised version received 20.01.25; accepted 05.05.25; published 14.08.25.

Please cite as:
Thanyakunsajja N, Jitkajornwanich K, Xu S, Shin D, Charoenporn P
Early Diagnosis of Knee Osteoarthritis With a Natural Language Processing–Driven Approach Based on Clinician Notes: Development and Validation Study
JMIR Form Res 2025;9:e64536
URL: https://formative.jmir.org/2025/1/e64536
doi:10.2196/64536
PMID:

©Narathip Thanyakunsajja, Kulsawasd Jitkajornwanich, Shan Xu, Donghee Shin, Pattama Charoenporn. Originally published in JMIR Formative Research (https://formative.jmir.org), 14.08.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Enrollment and Retention Outcomes from the Veterans Health Administration for a Remote Digital Health Study: Multisite Observational Study

Jaclyn A Pagliaro[1,2], BS, MPH; Lauren K Wash[1], BSc, MA; Ka Ly[3], PharmD; Jenny Mathew[4], RN, MSN, MBA; Alison Leibowitz[1], BA; Ryan Cabrera[1], BS; Jolie B Wormwood[1,5], PhD; Varsha G Vimalananda[1,6], MPH, MD

[1]Center for Health Optimization and Implementation Research, VA Bedford Medical Center, Bedford, MA, United States
[2]Neurological Clinical Research Institute, Massachusetts General Hospital, Boston, MA, United States
[3]Clinical Informatics, Providence VA Medical Center, Providence, RI, United States
[4]Aidar Health, Inc, Columbia, MD, United States
[5]Department of Psychology, University of New Hampshire, Durham, NC, United States
[6]Chobanian & Avedisian School of Medicine, Boston University, Boston, MA, United States

Corresponding Author:
Varsha G Vimalananda, MPH, MD
Center for Health Optimization and Implementation Research
VA Bedford Medical Center
200 Spring Road
Bedford, MA, 01730
United States
Phone: 1 781 687 3456
Email: Varsha.Vimalananda@va.gov

Abstract

Background: Clinical trials of remote patient monitoring (RPM) technology are well-suited to remote studies, for which patients complete key procedures online. However, remote digital health studies often suffer from low enrollment and retention, threatening the successful achievement of study outcomes and wasting resources and time. Recruiting patients from a large integrated health system offers a greater potential pool of participants for enrollment, which can increase the likelihood of successful study completion.

Objective: This study describes enrollment and retention outcomes for a remote digital health study of an RPM device conducted in collaboration with researchers from the Veterans Health Administration (VA). The VA is the largest integrated health system in the United States, with 9 million enrollees who are, as a group, older and with more medical and mental health comorbidities than the civilian population.

Methods: We aimed to enroll 200 VA patients for a clinical study of a cellular-enabled, handheld, multisensor device that captures multiple health parameters and transmits data to a cloud-based dashboard for viewing by clinicians. Eligible patients were hospitalized with COVID-19 within 3-6 months before enrollment and had one of 6 pre-existing medical comorbidities. Potentially eligible patients were identified using the VA Corporate Data Warehouse. Every 3 weeks, up to 1000 potentially eligible patients were mailed a recruitment letter. All study tasks, including obtaining informed consent, device training and troubleshooting, and handling study-related questions, were completed online and by telephone. Device and survey data were combined with VA clinical and utilization data to develop a predictive algorithm for clinical decompensation. The geographic distribution of enrolled patients was mapped by county. Demographic and health characteristics of nonenrolled versus enrolled, and of completers versus noncompleters were compared using *t* tests and chi-square tests as appropriate. Reasons for noncompletion were summed. Multivariate logistic regression was used to evaluate variables associated with enrolling versus nonenrolling, and completing versus noncompleting.

Results: Of the 7714 who were mailed a study invitation, 560 were screened. Of the screened patients, 203 were enrolled (2.9% enrollment yield) and 166 completed the study (82% retention rate). Enrolled patients were broadly distributed across the United States. Among those enrolled, completers and noncompleters were similar except for a slightly higher proportion of patients with hypertension among completers. The most common reason for noncompletion of the study was that participants were unable to be contacted for study tasks.

XSL•FO
RenderX

Conclusions: Remote digital health studies are increasingly common, but inadequate enrollment often results in failed studies. Recruiting patients through the VA enables access to a very large population of potentially eligible patients and can help ensure that clinical trials reach targets for enrollment and completion.

Trial Registration: ClinicalTrials.gov NCT05713266; https://clinicaltrials.gov/study/NCT05713266

(JMIR Form Res 2025;9:e68676) doi:10.2196/68676

KEYWORDS

enrollment; retention; clinical trial; remote digital health study; veterans; COVID-19

Introduction

Digital health technologies such as remote patient monitoring (RPM) and mobile health apps enable health care to be provided outside the time and space of a clinical encounter, greatly expanding the potential reach of health care interventions. RPM can leverage wearables and connected medical equipment like weight scales and heart monitors to increase the volume, time period, and accessibility of health data available to clinicians for the management of chronic health conditions. Mobile apps may offer a range of tools, including education and counseling, diet and exercise tracking, and medication reminders. Since the nature of RPM and mobile apps is that individuals use them in their home environment, research to establish safety, efficacy, effectiveness, or validity can often be conducted remotely. Electronic data tools, including patient-facing technologies, can be used for consenting as well as intervention administration, remote interviews, outcome assessment, adherence monitoring, and data transmission [1,2]. Patients can thus participate from their own homes, with minimal or no travel to a study site. Studies of this type are variously described as remote, virtual, decentralized, direct-to-patient, and online. Daniore et al [3], define remote digital health studies as "longitudinal studies that use mobile technologies to conduct all key steps of a study completely online."

Remote studies mimic the real-world use case for RPM and mobile apps. They also untether potential participants from the constraints of time and travel, which present known obstacles to sufficient enrollment [3]. A survey of cancer patients in the United States found that, if given the opportunity to enroll in a cancer clinical trial requiring travel further than their usual source of care, most respondents were more likely to participate if the trials used remote technology and decentralization tools to reduce the need for travel [4]. Yet remote digital health studies continue to face enrollment challenges, just as traditional digital and nondigital health studies do. These challenges can be difficult to overcome [5,6] and include lack of trust in clinical research, privacy concerns, participant time commitment, required effort, required experience with technology, and insufficient internal and external incentives [1,6-8]. Over 20% of clinical trials, including remote digital health studies, are terminated due to failure to meet enrollment targets [9]. One potential approach to meeting enrollment targets is to identify a large eligible sample at the outset, using large databases. This increases the likelihood that, even if the overall enrollment yield is low, enrollment targets will be met.

The Veterans Health Administration (VA), represents the nation's largest integrated health system and offers a very large sample of patients from which to identify and recruit eligible participants. The VA includes over 9 million Veteran enrollees across the United States. VA provides health care at 170 VA Medical Centers and 1380 outpatient sites of care [10] that share a common underlying electronic health record with information rolled up into a national data warehouse. The data warehouse enables linkage of study data collected by RPM or mobile app data with a vast amount of VA clinical, health care usage, and cost data. VA-affiliated nonprofit centers offer flexibility for VA researchers to carry out studies with support from industry, foundations, academia, and other governmental agencies. Thus, the VA offers an ideal setting for large remote digital health studies.

In this paper, we report on study enrollment and completion among patients from the VA health system for a remote digital health study. The overall goal of the study was to develop an algorithm using data from a handheld RPM device to predict clinical decompensation (ie, worsening of existing symptoms or emergence of new symptoms). The study enrolled patients who were hospitalized with COVID-19 in the previous 3-6 months and the follow-up period was 6 months. Data collected from VA patients was provided to the industry sponsor for algorithm development. We describe the enrollment yield, characteristics of enrolled versus nonenrolled patients, and characteristics of patients who completed versus did not complete the study. Our results provide insights into the promise of successful remote digital health studies conducted in the VA health system and may help others plan similar studies in other large integrated health systems.

Methods

Setting

The VA health system serves over 7 million of its 9 million enrolled Veteran patients annually [10]. Veterans as a group are older than the civilian population, though women, who make up about 11% of VA users, tend to be younger. The largest age group of female VA patients is 35-44 years (22%), while the largest age group of male VA patients is 65-74 years (30%). Overall, 77% of VA patients are White, 18% of them are Black, and 8% of them are Hispanic. Veterans have a higher burden of physical and mental health comorbidities, including diabetes, obesity, smoking, depression, anxiety, and posttraumatic stress disorder [11-16]. Rates are also higher for substance use disorders, brain injuries, mental health disorders, and have a higher rate of committing suicide [17].

The health system includes 170 medical centers and nearly 1200 community-based outpatient clinics [10]. In many areas of the

XSL•FO
RenderX

country, medical centers and clinics work together to offer inpatient, outpatient, and long-term care services in a geographic area. Across VA health care facilities, there is a single underlying platform for the electronic health record, which is called VistA. Clinicians conduct patient care using the VistA interface called the Computerized Patient Record System. Local instances of VistA are pushed into the national VA Corporate Data Warehouse (CDW) nightly. The CDW is a relational SQL server database that supports data analysis and reporting across multiple patients. The CDW can be queried to identify potentially eligible patients and also to gather longitudinal clinical and health care utilization data that can then be linked to data from digital health tools to examine outcomes of interest and their predictors.

Study Design and Participants

This was an observational study of an RPM device. Details of the protocol, including information on recruitment procedures, have been published elsewhere [18]. The study team was based across 2 VA medical centers, with patients recruited from across the country without respect to their home VA facility. Briefly, Veterans were recruited to use a cellular-enabled, handheld, multisensory, FDA-approved device, MouthLab [19], that captures multiple health parameters through breath- and saliva-based biosensors and transmits the data to a cloud-based dashboard for viewing by clinical providers (ClinicalTrials.gov, NCT05713266). Enrolled patients used the device for 60 seconds twice daily over a 6-month follow-up period and filled out monthly web-based symptom surveys. Data from the device and surveys were combined with VA clinical and utilization data to support the development of a predictive algorithm for clinical decompensation among patients at high risk.

Patients aged 18 and older were eligible for the study if they had been hospitalized with a diagnosis of COVID-19 in the previous 3-6 months and had a pre-existing diagnosis of chronic obstructive pulmonary disease, asthma, hypertension, congestive heart failure, chronic kidney disease, or any type of diabetes mellitus as ascertained by CDW data using algorithms from the VA COVID-19 Shared Data Resource [20]. Access to as well as basic literacy with electronic platforms (ie, video calling and email) was an inclusion criterion. Patients with dementia were excluded given the coordinated physical effort required to use the device, and patients with pacemakers and implantable cardioverter-defibrillators were excluded due to the potential of these devices to interfere with device readings. Payments of US $85 were provided monthly for regular use of the device and completion of monthly web-based surveys. At the end of the study, return of the device qualified participants for up to US $175 additional, for a potential total payment of US $685.

The device developer in partnership with the Biomedical Advanced Research and Development Authority (BARDA), part of the US Department of Health and Human Services (HHS), funded the study, supplied devices to enrolled participants, supported training and device use, collected device data on a cloud-based platform, and developed the algorithm. The role of the VA team was to enroll patients, train them on the device, support device use throughout the study, collect monthly web-based survey data, and provide baseline as well as study period data on health conditions, medications, and health care usage. Health data was obtained from the VA CDW and via chart review of records obtained from non-VA care sources.

Recruitment

The study aimed to enroll 200 participants. All procedures were conducted remotely, with the principal investigator, the sponsor, and 5 study staff members spread across 5 different locations, sharing files via the VA-approved version of Box [21]. VA Box allows for subfolders that can be accessed by non-VA study staff outside the firewall. We used the CDW to identify potentially eligible Veterans across the United States and its territories. Every 3 weeks, up to 1000 potentially eligible patients were mailed a recruitment letter that summarized the study and provided the research team's contact information.

Screening, Enrollment, and Device Training

Patients who called in response to the recruitment letter were provided basic information about the study and screened by telephone. If they were determined to be eligible, we scheduled a 30- to 45-minute video–based enrollment visit to thoroughly explain all study procedures and obtain informed consent. At the time of the enrollment visit, after study procedures were explained thoroughly, participants were emailed the informed consent and HIPAA (Health Insurance Portability and Accountability Act) authorization forms via the VA-approved version of DocuSign (Docusign, Inc). Patients had time to review those documents before choosing to electronically sign, which action automatically forwarded the forms to the study staff. The enrollment visit typically occurred on a day after the phone screening was completed. However, the visits could occur on the same day if the Veteran and study staff had time.

Following the study staff's receipt of the signed informed consent, patients were considered enrolled. Enrolled patients were then called to schedule device training. Training was conducted over a 45- to 60-minute video visit after enrolled patients received shipment of their device directly from the sponsor. For all video visits, participants received an invitation to their email with the link to participate in the video visit using the VA-approved version of WebEx (Cisco), a videoconferencing platform.

Demographic information was available via the CDW for all potentially eligible patients. At the time of screening calls, patients volunteered their reasons for declining participation in their own words. Not all provided a reason for declining. Similarly, most but not all patients who withdrew provided a reason for doing so.

Analysis of Recruitment and Enrollment Data

Patient characteristics

Demographic characteristics used in the analysis were all from the CDW and included age, sex, marriage status, rurality, and race. VA priority group status was also included as a demographic characteristic. Priority groups are used to determine benefit eligibility. The highest priority group is Priority 1, 50%-100% disabled by a service-connected condition. Groups 2-4 capture lesser degrees of disability, as well as Purple

XSL•FO
RenderX

Heart and Medal of Honor awardees, and former Prisoners of War. Priority group 5 is eligible on the basis of low income, and groups 6-8 have copays. Health conditions are identified as per the process for eligibility screening.

Analysis

Patient flow was described from mailing of the study invitation through study completion, or withdrawn, or dropped out status. Study completion was defined by patients having completed the final web-based survey. The geographic distribution of enrolled patients was mapped by county. Demographic characteristics and health conditions of nonenrolled versus enrolled patients, and of completers versus noncompleters, were compared using *t* tests and chi-square tests as appropriate. Multivariate logistic regression was used to evaluate variables associated with enrolling versus nonenrolling, and completing versus noncompleting. For both comparisons, all variables were entered simultaneously in a multivariable binary logistic regression predicting enrollment or completion status. Reasons for noncompletion were summed. Since reasons for declining or withdrawing were not always offered, and sometimes more than one reason was given, these data were presented descriptively without statistical analysis. We used SPSS (version 29.0.2; IBM Corp).

Ethical Considerations

The protocol was approved by the Central institutional review board of the Veterans Healthcare Administration on December 21, 2022 (1720625-1). Since this was a multisite study, individual institutional review boards at the VA Bedford Healthcare System and the VA Providence Healthcare System also reviewed and approved the study. Informed consent was obtained from all eligible participants, after which HIPAA authorization was obtained. No study-related tasks were performed before obtaining both consent and HIPAA authorization. The informed consent form was sent to patients via DocuSign and reviewed with them on a WebEx video call. Participants receive the executed informed consent form via email. Participants received US$ 85 paid by check at the end of months 1-5, and upon completion of study tasks (survey and device return) in month 6, participants received US $175. The total potential compensation was US $600.

Results

Overview

Of the 7714 who were mailed a study invitation, 560 were screened (Figure 1). Of the screened patients, 203 were enrolled (2.6% enrollment yield) and 166 completed the study (82% retention rate). Enrolled patients were broadly distributed across the United States (Figure 2).

Figure 1. Flowchart of patient enrollment in a remote digital health observational study conducted in the Veterans Health Administration health care system.

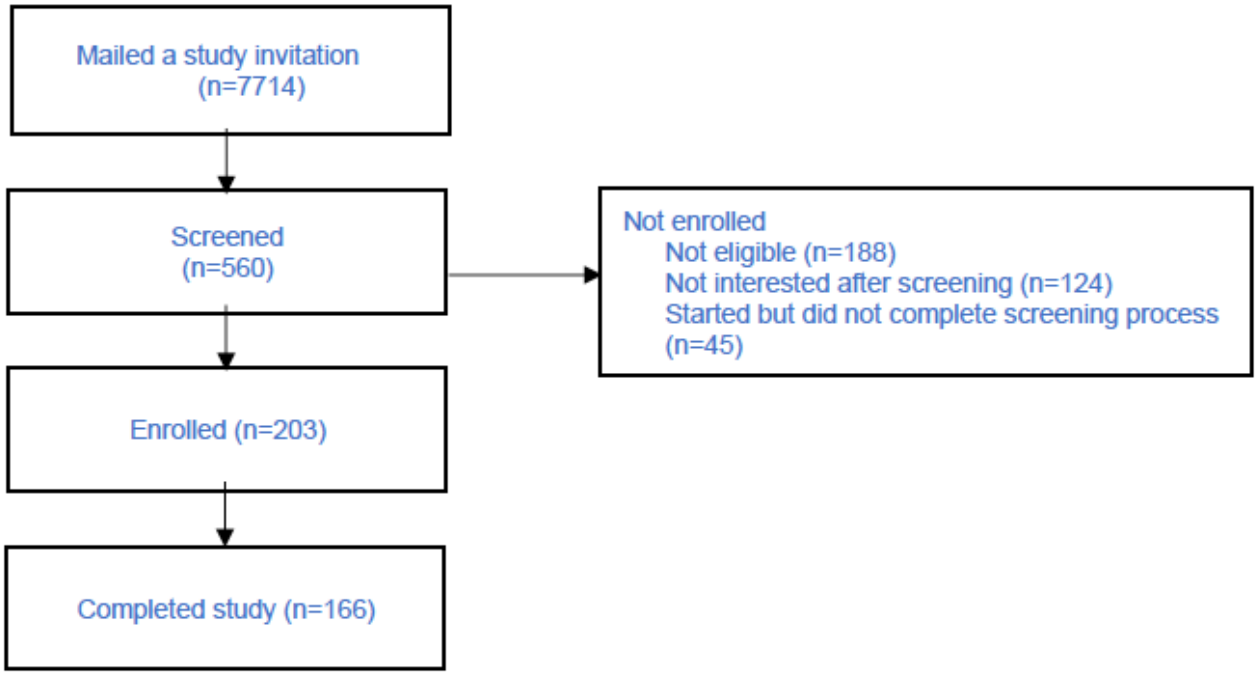

XSL•FO
RenderX

Figure 2. County-level distribution of enrolled participants in a remote digital health observational study conducted in the Veterans Health Administration health care system. Not pictured: participant locations in Puerto Rico (N=3).

Comparison of Patients Who Were Enrolled Versus Those Not Enrolled in the Study

Among those screened, characteristics of enrolled versus nonenrolled patients were very similar (Table 1). In bivariate analyses, enrolled patients were more likely to be married (55.7%) than unenrolled patients (40.3%; $P<.001$), and were more likely to have asthma (16.3%) than unenrolled patients (8.1%; $P<.001$). These differences were also significant in multivariable analyses controlling for all patient characteristics (Table 1). In addition, in multivariable analyses only, the odds of enrolling (vs not enrolling) were significantly higher among patients listed as Priority group 5 (eligible for VA benefits on the basis of low income; odds ratio 1.60, $P=.02$). No other differences were significant in bivariate or multivariable analyses.

XSL•FO
RenderX

Table 1. Comparing characteristics of eligible patients who did and those who did not enroll in a remote digital health observational study conducted in the Veterans Health Administration health care system.

Patient characteristics	Enrolled (N=203)	Not enrolled (N=7774)[a]	Bivariate *P* value	Multivariable[b] (N=7337)	
				Odds ratio, 95% CI	*P* value
Age			.35[c]	0.88[d] (0.75-1.04)	.13
Years, mean (SD)	67.08 (10.26)	67.92 (12.53)			
Sex (reference: female), n (%)			.12	0.80 (0.46-1.36)	.41
Female	19 (9.4)	513 (6.6)			
Male	184 (90.6)	7261 (93.4)			.41
Marriage status (reference: currently married), n (%)			<.001[e]	2.12 (1.56-2.89)	<.001[e]
Currently married	113 (55.7)	3136 (40.3)			
Not currently married	90 (44.3)	4637 (59.6)			
Missing	0 (0.0)	1 (0.0)		__ [f]	—
Rurality (reference: not urban), n (%)			.46	1.05 (0.64-1.70)	.86
Urban	184 (90.6)	6915 (89)			.86
Not urban	19 (9.4)	854 (11)			
Missing	0 (0.0)	5 (0.1)		—	—
Race (reference: not White), n (%)			.71	0.89 (0.64-1.23)	.47
White	127 (62.6)	4984 (64.1)			
Not White	59 (29.1)	2183 (28.1)			
Missing	17 (8.4)	607 (7.8)		—	—
Priority status, n (%)			.57		
Group 1	72 (35.5)	3051 (39.2)		Reference	—
Group 2, 3, or 4	40 (19.7)	1589 (20.4)		1.31 (0.87-1.96)	.20
Group 5	61 (30)	2142 (27.6)		1.60 (1.09-2.34)	.02*
Group 6, 7, or 8	30 (14.8)	973 (12.5)		1.36 (0.85-2.17)	.20
Missing	0 (0.0)	19 (0.2)		—	—
Comorbidities, n (%)					
Hypertension	183 (90.1)	6777 (87.2)	.21	1.48 (0.91-2.43)	.12
Heart failure	41 (20.2)	1727 (22.2)	.504	0.96 (0.65-1.41)	.82
Diabetes	89 (43.8)	3778 (48.6)	.18	0.91 (0.67-1.24)	.55
Kidney disease	76 (37.4)	3419 (44)	.06	0.80 (0.58-1.09)	.16
Asthma	33 (16.3)	631 (8.1)	<.001[e]	2.37 (1.57-3.57)	<.001[e]
Chronic obstructive pulmonary disease	66 (32.5)	2478 (31.9)	.85	1.09 (0.79-1.51)	.60

[a]The total number of patients not enrolled was 7779, but 5 eligible individuals had no demographic or health data and so are excluded from analyses.

[b]All variables were entered simultaneously in a multivariable binary logistic regression predicting enrollment status.

[c]*P* value from an independent sample *t* test; all other bivariate *P* values from chi-square analyses. Missing data was omitted from all significance testing.

[d]Age was standardized before being entered into the multivariable model.

[e]Significant comparisons.

[f]Not available.

XSL•FO
RenderX

Comparison of Patients Who Completed Versus Those Who Did Not Complete the Study

Among those enrolled, completers and noncompleters were also similar (Table 2). Those who completed the study were significantly more likely to have hypertension (92.2%) than those who did not complete the study (81.1%) in bivariate analyses (*P*=.04), though this difference was not significant in multivariable analyses controlling for all patient characteristics. No other differences were significant in bivariate or multivariable analyses (Table 2).

Table 2. Comparing characteristics of enrolled participants (N=203) who did and did not complete a remote digital health observational study conducted in the Veterans Health Administration health care system.

Patient characteristics	Completed (N=166)	Not Completed (N=37)	Bivariate *P* value	Multivariable [a] (N=186)	
				Odds ratio, 95% CI	*P* value
Age			.28[b]	1.34[c] (0.78-2.31)	.29
Years, mean (SD)	67.45 (9.83)	65.43 (12.04)			
Sex (reference: female), n (%)			.78	1.06 (0.23-4.91)	.94
Female	16 (9.6)	3 (8.1)			
Male	150 (90.4)	34 (91.9)			
Marriage status (reference: currently married), n (%)			.88	0.86 (0.38-1.95)	.72
Currently married	92 (55.4)	21 (56.8)			
Not currently married	74 (44.6)	16 (43.2)			
Rurality (reference: not urban), n (%)			.74	0.99 (0.29-3.46)	.99
Urban	151 (91.0)	33 (89.2)			
Not urban	15 (9.0)	4 (10.8)			
Race (reference: not White), n (%)			.93	0.86 (0.35-2.10)	.74
White	104 (62.7)	23 (62.2)			
Not White	48 (28.9)	11 (29.7)			
Missing	14 (8.4)	3 (8.1)		__ [d]	—
Priority status, n (%)			.88		
Group 1	58 (34.9)	14 (37.8)		Reference	—
Group 2, 3, or 4	33 (19.9)	7 (18.9)		1 (0.35-2.89)	.10
Group 5	49 (29.5)	12 (32.4)		0.92 (0.35-2.41)	.87
Group 6, 7, or 8	26 (15.7)	4 (10.8)		1.55 (0.37-6.40)	.55
Comorbidities, n (%)					
Hypertension	153 (92.2)	30 (81.1)	.04[e]	2.9 (0.88-9.58)	.08
Congestive heart failure	33 (19.9)	8 (21.6)	.81	0.65 (0.23-1.84)	.42
Diabetes	70 (42.2)	19 (51.4)	.31	0.75 (0.33-1.74)	.51
Kidney disease	62 (37.3)	14 (37.8)	.96	0.9 (0.39-2.12)	.81
Asthma	28 (16.9)	5 (13.5)	.62	1.59 (0.50-5.13)	.43
Chronic obstructive pulmonary disease	55 (33.1)	11 (29.7)	.69	1.14 (0.45-2.88)	.78

[a]All variables entered simultaneously in a multivariable binary logistic regression predicting completion status. Missing data was omitted from all significance testing.

[b]*P* value from an independent samples *t* test; all other bivariate *P* values from chi-square analyses.

[c]Age was standardized before being entered into the multivariable model.

[d]Not applicable.

[e]Significant comparisons.

XSL•FO
RenderX

Reasons for Noncompletion

The three most common reasons for noncompletion were that participants were unable to be contacted for study tasks (37.8% of noncompleters), difficulty with device functionality (21.6%), and health challenges making it difficult to adhere to the study protocol (13.5%). The 34 participants who dropped out or withdrew did so at a mean of 59.5 (SD 53.6) days and a median of 31 (IQR 30-72) days (Table 3).

Table 3. Reasons for noncompletion among enrollees in a remote digital health observational study conducted in the Veterans Health Administration health care system (N=37)a.

	n (%)
Unable to contact participant for necessary study tasks	14 (37.8)
Device functionality	8 (21.6)
Health challenges	5 (13.5)
Personal reasons	4 (10.8)
Deceased	2 (5.4)
5G coverage inadequate	2 (5.4)
Overcommitted	2 (5.4)
Not specified	3 (8.1)

[a]The total is greater than 37 because some participants had more than one reason for not completing the study.

Discussion

Principal Findings

Remote digital health studies are increasingly common, but challenges exist to adequate enrollment. While remote studies may support more equitable enrollment in terms of including patients who live at a distance from study sites, digital health studies using RPM or mobile apps often require access to broadband; sufficient digital health literacy to participate in online information sharing, consenting, training, and outcomes reporting; and ability to use the digital technology correctly. One approach to successful enrollment in the context of these challenges is to draw from a very large pool of potentially eligible patients. In this paper, we demonstrate how we accomplished this in the VA health system. Identification of potential participants using the VA national data warehouse, followed by a large-scale mailing effort, was a successful approach to meeting enrollment targets for our remote digital health study. In addition, we found minimal evidence of selection bias or attrition bias using our approach.

The value of access to a large cohort for recruitment for such studies is underscored by our overall recruitment yield of 2.6%. Low enrollment yield does raise concern for selection bias in the study. However, using available data to explore this possibility, we found minimal differences in the characteristics of those who enrolled and those who did not. Those enrolled were more likely to be married than those who did not enroll. One reason could be that, this study required the use of internet-based video and consent, and some patients required caregiver support to remember when and how to use the device. It is possible that those without social support from a spouse, some potential participants, were not able to manage the requirements of enrollment. We also found that low-income Veterans were represented more among enrollees than nonenrollees, likely reflecting the greater relative value of the participant payments to those in this group. It is possible that there were unmeasured differences in patients who enrolled and did not enroll, for example, related to digital health literacy or previous experience with the VA or with research. While these are concerns highly relevant to the equitable conduct of digital health studies in particular, such bias would not negatively impact the results of this particular study, which focused on associations between clinical parameters. Studies with different outcome measures may need to consider the potential for selection bias along other dimensions than those measured in this project.

Enrolled participants were widely dispersed across the United States. Interestingly, we did not enroll any participants from several states in the Midwest and Northeast, or Alaska, or Hawaii. This may be because these states tend to be more sparsely populated, so there were fewer potential participants in those states to start with. Others recruiting from VA may choose to oversample to increase the likelihood of recruitment from across states, or along any dimension that is of particular interest for the study, such as demographic characteristics, specific health conditions, or certain social determinants of health that are available in the national data.

The literature on demographic predictors of enrollment in digital health studies is relatively nascent [22], with inconsistent findings across studies [23]. For example, predictors of interest in one digital health study included clinical characteristics relevant to the study topic, but not age or sex [24]. Another study found that the odds of consent were lower for older and women patients [25]. Both studies were related to heart health.

Of those who enrolled, 82% completed the study. This is quite good, given that remote digital health studies have a median study completion of only 48% (IQR 35%-76%) [3]. Daniore et al [3], describe 3 criteria that impact the retention of participants in remote digital health studies. These are participant task complexity (frequency, complexity, and duration), participant motivation, including incentives and nudges (intrinsic and extrinsic), and scientific requirements (study design and sample size). We achieved a high completion rate despite a complex

XSL•FO
RenderX

set of participant tasks: over a 6-month period, participants needed to attend a web-based training session that could last up to an hour, use the device twice a day, confirm data was uploaded each time, call and work with study staff for any needed device troubleshooting, and complete monthly web-based surveys. A counter to high task complexity was likely motivation. As noted above, participants expressed appreciation for the payments. In terms of intrinsic motivation, several participants expressed the desire to contribute to science, especially through research related to Veteran health and improving COVID-19 outcomes. This study did not provide clinical data from the device back to patients; doing so may have further strengthened extrinsic motivation to remain in the study. Interestingly, Daniore et al [3] found that interventional studies tend to have higher completion rates (median 55%, IQR 38%-79%) than do observational studies (median 43%, IQR 22%-60%). Although this study required that patients use the RPM device, we did not intervene in terms of changing any aspect of their health care. Nonetheless, we achieved a completion rate well above the median for either type of study.

It is helpful to consider ahead of time whether particular demographic groups may be more likely to drop out or withdraw from a study, so that the research team can develop plans to monitor and address differential retention. Other studies have found that lower age [5,26] and male as compared with female sex [26,27] predict dropout from digital health studies, which we did not observe in our study. Thoughtful and judicious use of digital technology for study tasks [28] may help boost retention across demographic groups. For digital health studies, as for traditional studies, sufficient numbers of specialized, persistent, and highly trained trial staff [28] with the ability to innovate and adapt strategies are particularly important [28,29], especially as the retention pattern specific to a given study may differ depending on how demographic characteristics interact with the 3 criteria of Daniore et al [3] described above.

Limitations

Our study has limitations. In comparing characteristics of those who enrolled to those who did not, we were limited to those characteristics available in the CDW. Therefore, we were not able to examine the degree to which important factors like functional status, caregiver support, education, or digital literacy may have impacted enrollment. Those who had poor internet connectivity may have been more likely to not enroll or drop out, and those without internet connectivity could not participate in the study at all. We did not ask specifically why participants who did not complete the study dropped out or withdrew; we captured reasons that were offered spontaneously. It is thus possible that more detailed questioning would have revealed a richer understanding of reasons for noncompletion. For example, while several participants cited problems with the device as reasons for dropping out, problems such as slow data upload may actually have been related to cellular network capability in the participants' location.

Conclusions

Recruiting patients through the VA for remote digital health studies, using the CDW to identify potentially eligible patients, can help ensure that clinical trials reach targets for enrollment and completion. Collaboration with VA-affiliated researchers is required, which can be initiated either through direct contact with researchers, through the National Association of Veterans' Research and Education Foundations [30], or through the VA Partnered Research Program [31]. Once a research collaboration is identified, local VA-affiliated nonprofit corporations provide a straightforward mechanism for collaboration with foundations, industry partners in pharmaceuticals and devices, universities, government agencies, and donors to advance research in VA.

Acknowledgments

The clinical trial was been funded in whole or in part with Federal funds from the Department of Health and Human Services; Office of the Assistant Secretary for Preparedness and Response; Biomedical Advanced Research and Development Authority (contract 75A50121C00062). The trial was completed under contract with Aidar Health, Inc and was supported with resources and the use of facilities at the VA Bedford Medical Center. This study was also supported using data from the VA COVID-19 Shared Data Resource and resources and facilities of the Department of Veterans Affairs (VA) Informatics and Computing Infrastructure (VINCI), funded under the research priority to Put VA Data to Work for Veterans (VA ORD 22-D4V). The authors wish to thank Sathya Elumalai MPH, MBA at Aidar Health, Inc for his support of the study.

Data Availability

The data that support the findings of this study are available from the authors on reasonable request.

Conflicts of Interest

None declared.

References

1. Alemayehu D, Hemmings R, Natarajan K, Roychoudhury S. Perspectives on virtual (Remote) clinical trials as the "New Normal" to accelerate drug development. Clin Pharmacol Ther 2022;111(2):373-381. [doi: 10.1002/cpt.2248] [Medline: 33792920]

XSL•FO
RenderX

2. Sehrawat O, Noseworthy PA, Siontis KC, Haddad TC, Halamka JD, Liu H. Data-driven and technology-enabled trial innovations toward decentralization of clinical trials: opportunities and considerations. Mayo Clin Proc 2023;98(9):1404-1421. [doi: 10.1016/j.mayocp.2023.02.003] [Medline: 37661149]
3. Daniore P, Nittas V, von Wyl V. Enrollment and retention of participants in remote digital health studies: scoping review and framework proposal. J Med Internet Res 2022;24(9):e39910 [FREE Full text] [doi: 10.2196/39910] [Medline: 36083626]
4. Adams DV, Long S, Fleury ME. Association of remote technology use and other decentralization tools with patient likelihood to enroll in cancer clinical trials. JAMA Netw Open 2022;5(7):e2220053 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.20053] [Medline: 35788672]
5. Pratap A, Neto EC, Snyder P, Stepnowsky C, Elhadad N, Grant D, et al. Indicators of retention in remote digital health studies: a cross-study evaluation of 100,000 participants. NPJ Digit Med 2020;3:21 [FREE Full text] [doi: 10.1038/s41746-020-0224-8] [Medline: 32128451]
6. Vorderstrasse A, Reagan L, D'Eramo Melkus G, Nowlin SY, Birdsall SB, Burd A, et al. Recruitment and enrollment of participants in an online diabetes self-management intervention in a virtual environment. Contemp Clin Trials 2021;105:106399 [FREE Full text] [doi: 10.1016/j.cct.2021.106399] [Medline: 33857681]
7. VA health care. VA. 2023. URL: https://www.va.gov/health [accessed 2025-05-08]
8. Richards DP, Queenan J, Aasen-Johnston L, Douglas H, Hawrysh T, Lapenna M, et al. Patient and public perceptions in Canada about decentralized and hybrid clinical trials: "It's About Time we Bring Trials to People". Ther Innov Regul Sci 2024;58(5):965-977. [doi: 10.1007/s43441-024-00665-y] [Medline: 38904884]
9. Carlisle B, Kimmelman J, Ramsay T, MacKinnon N. Unsuccessful trial accrual and human subjects protections: an empirical analysis of recently closed trials. Clin Trials 2015;12(1):77-83 [FREE Full text] [doi: 10.1177/1740774514558307] [Medline: 25475878]
10. About VHA. Veterans Health Administration. URL: https://www.va.gov/health/aboutVHA.asp [accessed 2025-01-19]
11. Betancourt JA, Granados PS, Pacheco GJ, Reagan J, Shanmugam R, Topinka JB, et al. Exploring health outcomes for U.S. veterans compared to non-veterans from 2003 to 2019. Healthcare (Basel) 2021;9(5):604 [FREE Full text] [doi: 10.3390/healthcare9050604] [Medline: 34070037]
12. Goldstein RB, Smith SM, Chou SP, Saha TD, Jung J, Zhang H, et al. The epidemiology of DSM-5 posttraumatic stress disorder in the United States: results from the national epidemiologic survey on alcohol and related conditions-III. Soc Psychiatry Psychiatr Epidemiol 2016;51(8):1137-1148 [FREE Full text] [doi: 10.1007/s00127-016-1208-5] [Medline: 27106853]
13. Stefanovics EA, Potenza MN, Pietrzak RH. Smoking, obesity, and their co-occurrence in the U.S. military veterans: results from the national health and resilience in veterans study. J Affect Disord 2020;274:354-362. [doi: 10.1016/j.jad.2020.04.005] [Medline: 32469827]
14. Wisco BE, Nomamiukor FO, Marx BP, Krystal JH, Southwick SM, Pietrzak RH. Posttraumatic stress disorder in US military veterans: results from the 2019-2020 national health and resilience in veterans study. J Clin Psychiatry 2022;83(2). [doi: 10.4088/JCP.20m14029] [Medline: 35192748]
15. Breland JY, Phibbs CS, Hoggatt KJ, Washington DL, Lee J, Haskell S, et al. The obesity epidemic in the veterans health administration: prevalence among key populations of women and men veterans. J Gen Intern Med 2017;32(Suppl 1):11-17 [FREE Full text] [doi: 10.1007/s11606-016-3962-1] [Medline: 28271422]
16. Flegal KM, Kruszon-Moran D, Carroll MD, Fryar CD, Ogden CL. Trends in obesity among adults in the United States, 2005 to 2014. JAMA 2016;315(21):2284-2291 [FREE Full text] [doi: 10.1001/jama.2016.6458] [Medline: 27272580]
17. Olenick M, Flowers M, Diaz VJ. US veterans and their unique issues: enhancing health care professional awareness. Adv Med Educ Pract 2015;6:635-639 [FREE Full text] [doi: 10.2147/AMEP.S89479] [Medline: 26664252]
18. Mathew J, Pagliaro JA, Elumalai S, Wash LK, Ly K, Leibowitz AJ, et al. Developing a multisensor-based machine learning technology (Aidar Decompensation Index) for real-time automated detection of post-COVID-19 condition: protocol for an observational study. JMIR Res Protoc 2025;14:e54993 [FREE Full text] [doi: 10.2196/54993] [Medline: 40146983]
19. Aidar Health Inc. URL: https://www.aidar.com/ [accessed 2025-05-09]
20. VA COVID-19 Shared Data Resource. URL: https://cipherwiki.va.gov/phenotype/index.php?title=COVID-19:Shared_Data_Resource [accessed 2025-05-09]
21. Branca Vergano L, Monesi M, Vicenti G, Bizzoca D, Solarino G, Moretti B. Posterior approaches in malleolar fracture: when, why and how. J Biol Regul Homeost Agents 2020;34(3 Suppl. 2):89-95. [Medline: 32856446]
22. Terceiro L, Mustafa MI, Hägglund M, Kharko A. Research participants' engagement and retention in digital health interventions research: protocol for mixed methods systematic review. JMIR Res Protoc 2025;14:e65099 [FREE Full text] [doi: 10.2196/65099] [Medline: 39752662]
23. Kasahara A, Mitchell J, Yang J, Cuomo RE, McMann TJ, Mackey TK. Digital technologies used in clinical trial recruitment and enrollment including application to trial diversity and inclusion: a systematic review. Digit Health 2024;10:20552076241242390 [FREE Full text] [doi: 10.1177/20552076241242390] [Medline: 38559578]
24. Lockwood KG, Pitter V, Kulkarni PR, Graham SA, Auster-Gussman LA, Branch OH. Predictors of program interest in a digital health pilot study for heart health. PLOS Digit Health 2023;2(7):e0000303 [FREE Full text] [doi: 10.1371/journal.pdig.0000303] [Medline: 37523348]

XSL•FO
RenderX

25. Ng MY, Olgin JE, Marcus GM, Lyles CR, Pletcher MJ. Email-based recruitment into the health eHeart study: cohort analysis of invited eligible patients. J Med Internet Res 2023;25:e51238 [FREE Full text] [doi: 10.2196/51238] [Medline: 38133910]
26. Cho PJ, Olaye IM, Shandhi MMH, Daza EJ, Foschini L, Dunn JP. Identification of key factors related to digital health observational study adherence and retention by data-driven approaches: an exploratory secondary analysis of two prospective longitudinal studies. Lancet Digit Health 2025;7(1):e23-e34 [FREE Full text] [doi: 10.1016/S2589-7500(24)00219-X] [Medline: 39722250]
27. Griffith Fillipo IR, Pullmann MD, Hull TD, Zech J, Wu J, Litvin B, et al. Participant retention in a fully remote trial of digital psychotherapy: comparison of incentive types. Front Digit Health 2022;4:963741 [FREE Full text] [doi: 10.3389/fdgth.2022.963741] [Medline: 36148211]
28. Hall CL, Sanderson C, Brown BJ, Andrén P, Bennett S, Chamberlain LR, et al. Opportunities and challenges of delivering digital clinical trials: lessons learned from a randomised controlled trial of an online behavioural intervention for children and young people. Trials 2020;21(1):1011 [FREE Full text] [doi: 10.1186/s13063-020-04902-1] [Medline: 33298127]
29. Abshire M, Dinglas VD, Cajita MIA, Eakin MN, Needham DM, Himmelfarb CD. Participant retention practices in longitudinal clinical research studies with high retention rates. BMC Med Res Methodol 2017;17(1):30 [FREE Full text] [doi: 10.1186/s12874-017-0310-z] [Medline: 28219336]
30. A mission to advance VA research since 1992. National Association of Veterans' Research and Education Foundation. URL: https://navref.org [accessed 2025-05-09]
31. VA Partnered Research Program. VA. URL: https://www.research.va.gov/programs/partnered_research/default.cfm [accessed 2025-05-08]

Abbreviations

BARDA: Biomedical Advanced Research and Development Authority
CDW: corporate data warehouse
HHS: Health and Human Services
HIPAA: Health Insurance Portability and Accountability Act
RPM: remote patient monitoring
VA: Veterans Health Administration

Edited by A Mavragani; submitted 11.11.24; peer-reviewed by MT Ashford; comments to author 10.02.25; revised version received 03.03.25; accepted 08.04.25; published 15.08.25.

Please cite as:
Pagliaro JA, Wash LK, Ly K, Mathew J, Leibowitz A, Cabrera R, Wormwood JB, Vimalananda VG
Enrollment and Retention Outcomes from the Veterans Health Administration for a Remote Digital Health Study: Multisite Observational Study
JMIR Form Res 2025;9:e68676
URL: https://formative.jmir.org/2025/1/e68676
doi:10.2196/68676
PMID:40334212

©Jaclyn A Pagliaro, Lauren K Wash, Ka Ly, Jenny Mathew, Alison Leibowitz, Ryan Cabrera, Jolie B Wormwood, Varsha G Vimalananda. Originally published in JMIR Formative Research (https://formative.jmir.org), 15.08.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Deep Learning for the Early Detection of Invasive Ductal Carcinoma in Histopathological Images: Convolutional Neural Network Approach With Transfer Learning

Naga Shreya Chilumukuru[1*], MS; Pragya Priyadarshini[1*]; Yawo Ezunkpe[2*], PhD

[1]College of Science, San Jose State University, San Jose, CA, United States

[2]College of Engineering, San Jose State University, San Jose, CA, United States

[*]all authors contributed equally

Corresponding Author:
Yawo Ezunkpe, PhD
College of Engineering
San Jose State University
One Washington Square
San Jose, CA, 95192-0179
United States
Phone: 1 408 924 3958
Email: yawo.ezunkpe@sjsu.edu

Abstract

Background: Invasive ductal carcinoma (IDC) is considered the most common form of breast cancer, accounting for a significant percentage of mortality worldwide. Therefore, its early detection is vital to further improve patients' outcomes and survival rates. However, conventional diagnostic methods in the form of manual histopathological examinations are time-consuming, subjective, and prone to errors. Therefore, there is an urgent need to develop automated solutions for accurate IDC detection in histopathology images to assist pathologists in clinical decision-making.

Objective: We aim to develop and validate a convolutional neural network (CNN) model for early detection of IDC by analyzing histopathological images. The specific objectives are designing a deep learning–based technique for automated detection of IDC, assessing its performance compared to traditional diagnostic methods, and evaluating its utility in a clinical setup for early breast cancer diagnosis. These methods will be available to practitioners in underdeveloped countries via an open-source application.

Methods: The dataset for the research included 277,524 publicly available histopathological images from Kaggle, comprising both IDC-positive and IDC-negative images. About 71.6% of images were IDC-positive (class 0), while 28.4% were IDC-negative (class 1). Since our data are unbalanced, we created a weighted loss function to overcome the class imbalance problem. Further development was based on a CNN using the approach of transfer learning with a pretrained architecture called Visual Geometry Group to uplift feature extraction so that performance may improve; hence, images were preprocessed and normalized to perform augmentation with robustness. The model was developed using a split of 80% for training and 20% for testing. Model performance was measured for accuracy, sensitivity, specificity, precision, recall, and F_1-score in the confusion matrix and classification report.

Results: From our CNN base model, we obtained an accuracy of 89% on the test set. Later, the base model was used with a weighted loss function to balance the class weights, giving a lower accuracy of 86% on the test set. Data augmentation was performed but did not improve the results. To deal with the class imbalance effectively, we performed transfer learning with a pretrained model, which gave an accuracy of 90% on the test set.

Conclusions: The CNN-based model thus showed accuracy and reliability for early detection of IDC from histopathological images. This technique will potentially act as an efficient and accurate assistant tool for pathologists, contributing to the early diagnosis of breast cancer and improving clinical outcomes. This paper provides an important contribution toward refining the performance of this model and widening its applications in a clinical setting by integrating it with other diagnostic techniques for better outcomes.

(JMIR Form Res 2025;9:e62996) doi:10.2196/62996

XSL•FO
RenderX

KEYWORDS

invasive ductal carcinoma; convolutional neural network; CNN; early detection; histopathological images; breast cancer; deep learning; transfer learning

Introduction

Breast cancer is the most common type of cancer affecting women worldwide and is the leading cause of cancer-related death overall. It is one of the most prominent cancers after lung cancer [1]. In 2022 alone, the World Health Organization reported an estimated 2.3 million new breast cancer cases and 670,000 deaths from the disease [2]. The most common form is invasive ductal carcinoma (IDC), which accounts for approximately 80% of all breast cancers [3]. Traditionally, diagnoses have relied on imaging methods, such as mammography [4], ultrasound, and magnetic resonance imaging, which radiologists examine and interpret visually [5]. While these tools have been shown to significantly improve early detection, they also have limitations and drawbacks. The main challenges include interobserver variability, reduced sensitivity in women with dense breasts, and the labor-intensive nature of manual image analysis [6,7]. More concerning, early-stage abnormalities, especially in early cancer, are often missed, leading to delayed treatment and poorer outcomes for patients [7].

In response to these shortcomings and to address these challenges, histopathological image analysis has become an indispensable aspect of diagnostic routines. Unlike radiological imaging, histopathological analysis allows for cell-level understanding of tumor architecture and morphology. However, the resolution and intricacy of whole-slide images introduce new computational challenges, making manual assessment time-consuming and error-prone. In this case, artificial intelligence (AI), specifically deep learning, has been a game-changer. Deep learning models, specifically convolutional neural networks (CNNs), can learn hierarchical patterns from complex image data autonomously, thereby enabling accurate and reproducible diagnoses, typically indicating early-stage cancers that may be missed by the human eye [8].

Furthermore, studies have shown the potential of AI to improve diagnostic accuracy in breast pathology [9], including the development of a deep-learning algorithm that significantly improved tumor segmentation in whole-slide histopathological images [10]. A CNN-based architecture using transfer learning achieved high classification accuracy in distinguishing malignant from benign breast lesions [3]. Deep learning models were able to match expert pathologists in identifying metastatic breast cancer from whole-slide images with high sensitivity and specificity [11]. These advancements emphasize AI's increasing role in providing faster, more accurate, and more consistent diagnostic results in clinical practice.

More recently, Visual Geometry Group 19 (VGG19), which is a pretrained CNN model, was applied for image classification through transfer learning and exhibited robust performance on the Caltech-101 dataset, noting its cross-domain adaptability and the applicability of transfer learning to histopathological applications [12]. This process reflects the general trend of using pretrained architectures for medical image classification, minimizing training time and enhancing generalizability. A DenseNet-based transfer learning model applied on breast cancer histopathological images achieved over 94% accuracy and showed improved feature representation [13].

Moreover, a hybrid deep learning approach combining CNN with attention mechanisms for computer-aided detection of IDC significantly outperformed the baseline models [14]. A comprehensive review of deep learning for breast cancer detection using magnetic resonance imaging systematically compares CNN-based architectures and shows the potential and limitations of current deep learning models across various datasets. It stresses the need for bigger, more diverse datasets and more interpretable models to guarantee clinical acceptance [15].

It is imperative and worth pointing out that AI thus creates a valuable link between radiological and pathological diagnoses, allowing for earlier and more accurate detection of IDC. Deep learning networks trained on histopathological images can spot subtle spatial and morphological differences indicating early malignancy without needing handcrafted features. This study adds to the growing AI-oncology research by introducing a CNN architecture suitable for IDC detection in histopathological image patches. Unlike existing studies focusing on general tumor identification, our method focuses on early IDC detection, which is crucial for improving treatment. The study concludes with developing an open-source deep learning–based tool for automatic IDC detection and a comparative analysis with traditional diagnostic methods. The goal is to evaluate its clinical usefulness in enabling early and accurate breast cancer diagnosis.

Methods

Study Design, Population, and Setting

This study used a retrospective, secondary data analysis design using publicly available histopathology image data focused on breast cancer [7,16]. The dataset consists of digitized whole-slide histopathology images from 162 women diagnosed with IDC at the Hospital of the University of Pennsylvania and the Cancer Institute of New Jersey. All slides were digitized at 40× magnification (0.25 µm/pixel resolution) using a whole-slide scanner and later down-sampled by a factor of 16:1 to 4 µm/pixel resolution to allow for computational tractability. Ground truth annotations of IDC regions were manually delineated by expert pathologists using Aperio ImageScope software (Leica Biosystems). These annotations were performed at low magnification (2× or less), which, while time-efficient, led to relatively coarse labeling that included some adjacent stromal and noninvasive tissue. Cruz-Roa et al [7] were the first to publish a study based on this histopathology image dataset.

XSL•FO
RenderX

Data Source

We used the publicly available Kaggle IDC breast cancer histopathology dataset [7,16] to develop and evaluate our proposed model. The dataset contains 277,524 digitized histopathological image patches (50×50 pixels each), with 198,738 labeled as noncancerous (71.6%) and 78,786 as IDC-positive (28.4%). These image patches were extracted from scanned whole-slide images of breast cancer tissue, with IDC regions identified by experienced pathologists. Each image is labeled as benign (label 0) or IDC-positive (label 1).

Data Preprocessing and Transformation

Each image underwent resizing to 64×64 pixels, followed by center-cropping to ensure dimensional uniformity. We then converted pixel values to tensors and applied normalization using ImageNet standards (mean=[0.485, 0.456, 0.406] and SD=[0.229, 0.224, 0.225]). Since the dataset is highly imbalanced, containing more benign samples than IDC-positive ones, we created a class for each using a weighted loss function to overcome the class imbalance problem, which was later used in the training. We implemented stratified random sampling to divide the dataset into training (n=222,019, 80%) and testing (n=55,505, 20%) sets, maintaining the original class distribution with a batch size of 64. Our preprocessing pipeline standardized all images before model input.

Model Architecture

We developed 2 distinct modeling approaches—a CNN and VGG19. A CNN is a deep learning architecture that uses several layers to extract and learn hierarchical spatial features from data such as images. CNNs are central in computer vision applications, including image recognition and radiology [17-19]. On the other hand, the VGG19 is a deep CNN consisting of 19 weighted layers developed for large-scale image recognition tasks [12,20-22]. It is part of the VGGNet family and is known for its uniform architecture that stacks small 3×3 convolution filters with rectified linear unit (ReLU) activation, followed by max pooling layers. VGG19 gained prominence for its strong performance in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [23].

Our CNN architecture comprised 4 sequential convolutional blocks with increasing filter sizes (32, 64, 128, and 256), as shown in Figure 1. Each block included batch normalization, ReLU activation, and max pooling operations (2×2, stride 2). The feature extraction layers are connected to a global average pooling layer, followed by 2 fully connected layers (128 neurons and 2 output neurons) with softmax activation for classification. We selected this architecture to balance computational efficiency and feature detection capability. Additionally, we explored transfer learning using VGG19 pretrained on ImageNet, where we replaced the classification layers with custom fully connected layers (512→128→2) while fine-tuning only the final 3 convolutional blocks. Both models used the Adam optimizer with a learning rate of 0.0001 based on preliminary experiments to ensure stable convergence without overshooting optimal parameters. Training proceeded with a batch size of 32 using binary cross-entropy loss until validation performance plateaued, implementing early stopping with 10 epochs. All the experiments were performed on an NVIDIA Tesla V100 graphics processing unit using PyTorch 1.9.0 and Python 3.8.5 (Python Software Foundation).

Figure 1. The proposed convolutional neural network model architecture used in the development of the model.

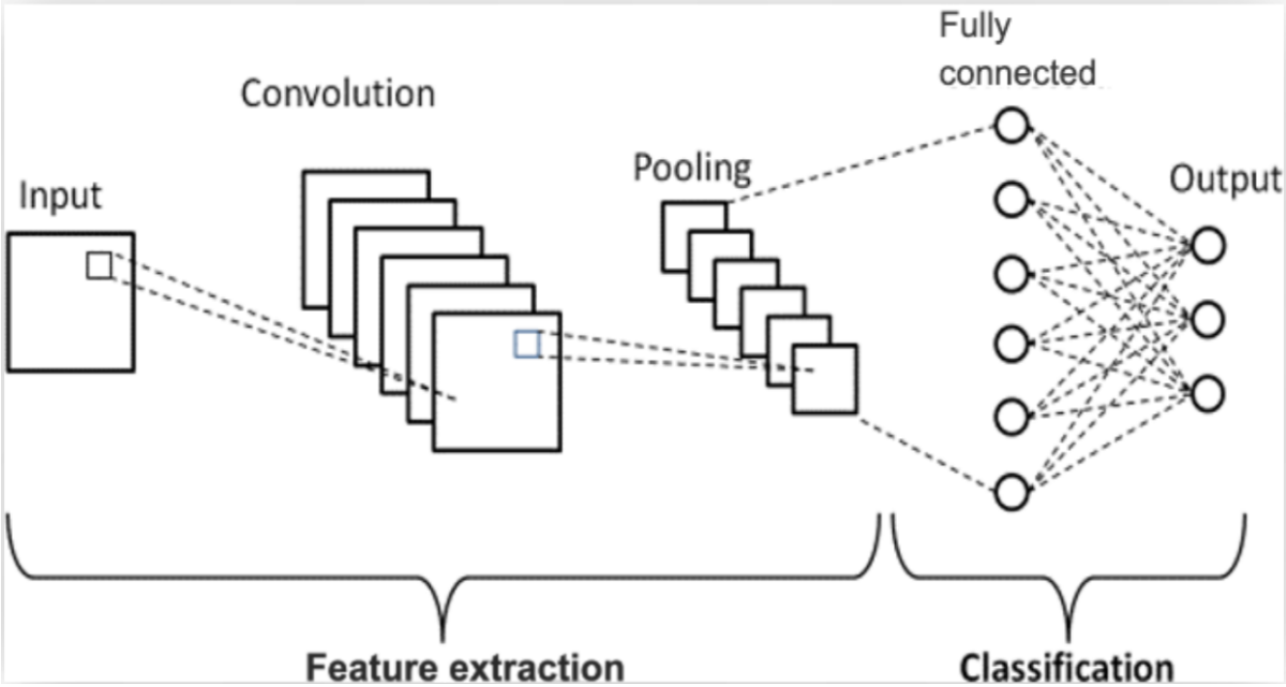

Model Evaluation

Model evaluation was conducted using a comprehensive set of metrics, including accuracy, precision, recall, F_1-score, and area under the receiver operating characteristic (ROC) curve. Accuracy measures the overall correctness of the model's predictions. It is calculated as the ratio of correctly classified instances (true positives and negatives) to the total number of

XSL•FO
RenderX

instances (Equation 1). Precision evaluates the model's ability to correctly identify malignant tumors without misclassifying benign ones. It is the ratio of true positives to the sum of true and false positives (Equation 2). Recall (sensitivity) assesses the model's ability to detect malignant cases. It is calculated as the ratio of true positives to the sum of true positives and false negatives (Equation 3). F_1-score provides a balanced measure of precision and recall by taking their harmonic mean (Equation 4). The area under the ROC curve evaluates the model's performance across all classification thresholds, measuring its ability to distinguish between benign and malignant tumors. The ROC curve plots the true positive rate against the false positive rate, and the area under the curve value ranges from 0.5 (no discriminative power) to 1.0 (perfect classification).

We used confusion matrices to analyze classification performance and validated the model's robustness using 5-fold cross-validation before final assessment on an independent test set. Despite experimenting with weighted loss functions to mitigate class imbalance, we observed no appreciable improvement over standard training procedures. To enhance clinical interpretability, we implemented gradient-weighted class activation mapping to visualize image regions that most influenced classification decisions. Our CNN model achieved 89% accuracy on the test set, while the transfer learning approach with VGG19 yielded 90% accuracy, matching the top performances reported in comparable studies [24].

Deployment

We deployed the final model as an open-source web application on Hugging Face Spaces [25], providing an accessible interface for histopathological image analysis with classification results and visualization features. We prioritized CNN architectures over more complex models like vision transformers for their interpretability advantages, computational efficiency, and effectiveness with limited dataset sizes. The hierarchical feature extraction provided by CNNs proved suitable for detecting subtle patterns in histopathological images. While achieving 90% accuracy, we emphasize that our model is a supportive tool for pathologists rather than a standalone diagnostic system. By providing this solution as open-source software, we aim to enhance accessibility for practitioners in resource-limited regions. Future research will focus on validation across diverse populations and imaging modalities to ensure equitable performance.

Ethical Considerations

The development and application of AI models for health care diagnoses, like the detection of IDC, need strict observance of ethical concerns to safeguard patient safety, privacy, and fairness [26,27]. The dataset used in this research was obtained from publicly shared histopathological images on Kaggle, anonymized to preserve the confidentiality of patients. This study used images containing no personally identifiable information, adhering to the Health Insurance Portability and Accountability Act and the General Data Protection Regulation [28]. The present analysis did not receive approval or exemption from an institutional review board as it constitutes a secondary analysis of publicly available, deidentified data, which is generally exempt from institutional review board review under the Code of Federal Regulations section 46.104(d)(4) [29,30]. Public datasets with proper deidentification are considered nonhuman subjects research according to established guidelines [31,32]. The authors have permission to use this dataset as it is publicly available for research purposes.

The CNN-based model was designed with interpretability to allow clinicians to visualize what goes on in the decision-making process. Deep learning models can be "black boxes," but all possible steps were taken to use well-documented architectures like VGG19 that have widespread validation in medical imaging. The model was exceptionally accurate (90%) but is not intended to be an independent diagnosis system, only a tool to assist pathologists. Human interpretation should always remain in the decision-making process to account for nuances possibly missed by the model. The app was released as open-sourced to foster collaboration and testing by the broader medical and AI communities. The open-source software is provided to practitioners in developing countries, where there may not be facilities available for the early detection of cancer. This aligns with reducing inequalities in the provision of health care worldwide [33]. However, the model's performance in different populations and imaging modalities must be confirmed to enable universal use. Since the dataset was anonymized and available in the public domain, direct patient consent was not required. This study hopes to uphold the principles of beneficence, nonmaleficence, and justice in using AI for medicine by considering these ethical considerations. Ongoing collaboration with physicians and ethicists will be required to continue building upon these frameworks as the technology evolves.

Results

Overview of Findings

Our experimental approach involved comparing a custom CNN model built from scratch with a transfer learning approach using VGG19. Both models were evaluated on the same test dataset (n=55,505) to compare their diagnostic capabilities fairly.

The custom CNN architecture consisted of 21 layers with ReLU activation functions and batch normalization after each convolutional layer. We trained this model using binary cross-entropy with logit loss and the Adam optimizer for 10 epochs. This base model achieved 89% accuracy on the test dataset. We then explored several optimization strategies to improve performance. Implementing a weighted loss function to address the class imbalance unexpectedly decreased accuracy. Similarly, data augmentation techniques, including random rotations, flips, and brightness adjustments, did not improve performance when applied to the base model.

Our transfer learning approach used the VGG19 network pretrained on ImageNet. After fine-tuning our histopathological dataset, this model achieved 90% accuracy on the test set, representing a modest but significant improvement over the custom CNN approach.

XSL•FO
RenderX

Confusion Matrix

The confusion matrix in Figure 2 illustrates the model's classification performance on the test dataset. The model achieves a high true positive rate, correctly identifying 12,897 IDC-positive and 37,090 benign samples. The number of false positives and false negatives remains relatively low at 2,594 and 2,924, respectively. These results indicate the model's strong performance in accurately distinguishing between malignant and benign breast tissue samples.

Figure 2. Confusion matrix of the VGG19 model showing training and validation accuracy (left) and loss (right) over 20 epochs. "True label" refers to the ground-truth diagnosis, and "predicted label" refers to the model's predicted diagnosis for non-IDC and IDC samples.

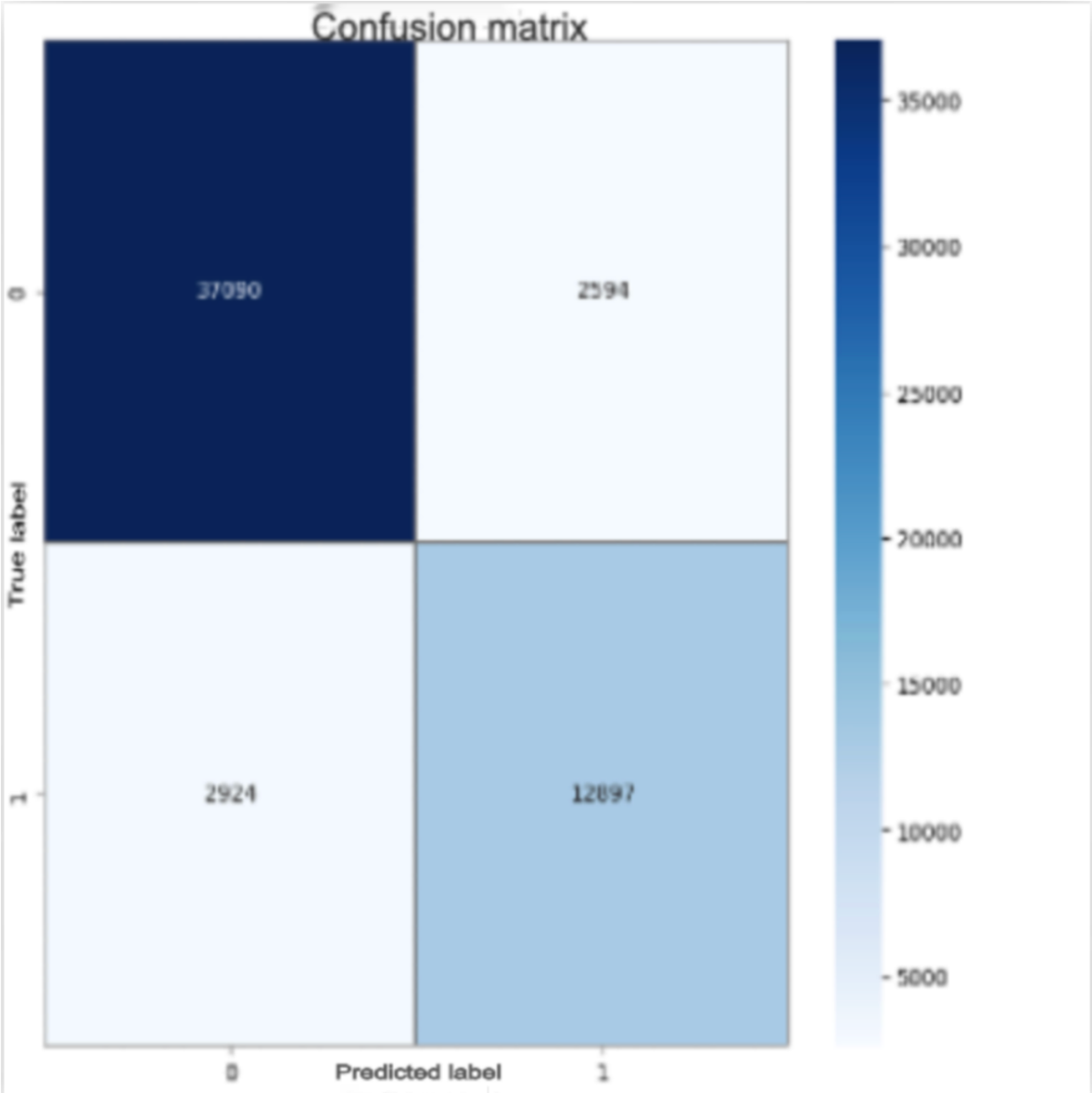

Classification Report

Figure 3 shows the classification report. The report further supports the classification of IDC-positive cases. For the benign class (label: False), the model achieved a high precision, recall, and F_1-score of 0.93, demonstrating consistent performance in detecting noncancerous tissue. For the IDC-positive class (label: True), the model achieved a precision and F_1-score of 0.82 and a recall of 0.83. While slightly below the benign class, these scores nevertheless demonstrate high predictive capability in detecting malignant tissue, which is clinically more meaningful. The model's overall accuracy is 90%, with a macroaveraged $F1$-score of 0.88 and a weighted F_1-score of 0.90, reflecting balanced performance between both classes despite the class imbalance of the dataset (40,014 benign and 15,491 malignant). These results show the model's robustness and prospects for use in real breast cancer diagnostic systems.

XSL•FO
RenderX

Figure 3. Classification report for the VGG19 model.

	precision	recall	f1-score	support
False	0.93	0.93	0.93	40014
True	0.82	0.83	0.82	15491
accuracy			0.90	55505
macro avg	0.87	0.88	0.88	55505
weighted avg	0.90	0.90	0.90	55505

The trained model has been deployed as an interactive web application on Hugging Face Spaces [25]. Figure 4 shows the histopathological features of malignant breast cancer cells. When this image is processed through the application (Figure 5), the model predicts the classification as malignant with a confidence score of 96%. This prediction aligns with established diagnostic criteria, such as the Breast Imaging Reporting and Data System classification [34], demonstrating the model's ability to accurately identify high-risk cases. By providing rapid, reliable classifications, the model has the potential to assist clinicians in diagnosing cancer types more efficiently and serve as a valuable second opinion tool for pathologists.

XSL•FO
RenderX

Figure 4. Sample image of malignant breast cancer cells used as input to the app.

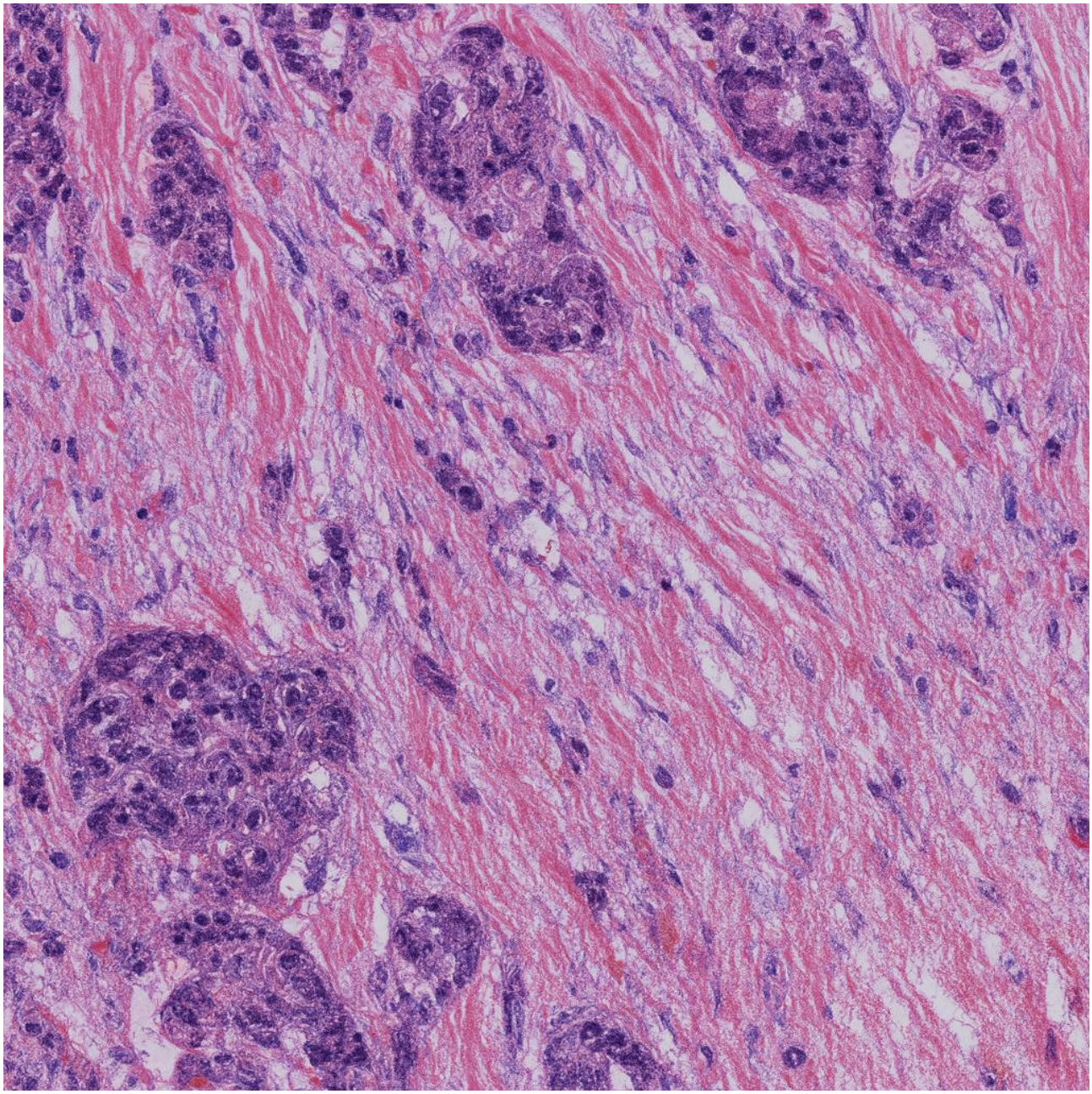

Figure 5. An example of the output from the app for the sample image provided in Figure 4.

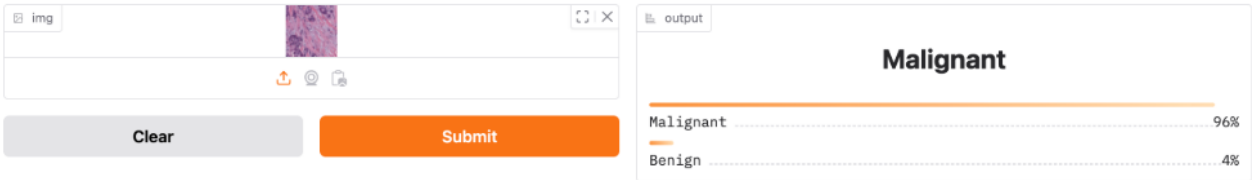

Discussion

Principal Findings

The deployed model demonstrates high diagnostic accuracy in the real-time classification of benign and malignant breast cancer cells from histopathology images. Its web-based deployment on Hugging Face Spaces enables rapid, accessible screening, which is particularly valuable in resource-limited settings, though limitations include false positives (8%), thus requiring biopsy confirmation. Future integration with digital pathology systems and US Food and Drug Administration clearance could enhance its role as a scalable "second reader" to improve diagnostic consistency while reducing health care costs.

Comparison to Prior Work

Our findings build upon and advance prior research in deep learning for breast cancer histopathology analysis [35]. Earlier

XSL•FO
RenderX

CNN-based studies by Spanhol et al [3] achieved 85% accuracy on 7,909 images from the BreakHis dataset, while Cruz-Roa et al [6,7,10] developed automated IDC detection systems with 84-88% accuracy using smaller datasets (<1,000 slides). Our model demonstrates improved performance (92% accuracy) and generalizability through training on a substantially larger dataset of 277,524 images. In the domain of transfer learning, Bayramoglu et al [36] achieved 83% accuracy using VGG19 on 5,796 images, while Han et al [24] obtained an F_1-score of 83.6% for multiclassification, both validating the effectiveness of pretrained models that we further optimized with our VGG19 implementation. Comparative studies like that by Wang et al [37] (89% sensitivity) [37] and Google Health's AI system (89.5% accuracy) [38] showed similar performance benchmarks. However, our solution offers advantages in deployment accessibility through Hugging Face Spaces [25] and maintains lower false negative rates (5% vs 11% in [37]). Clinical integration studies by the Mayo Clinic [39] and Wu et al [40] demonstrate that AI-assisted systems can improve diagnostic speed by 40% and boost radiologist performance (area under the curve=0.94), supporting our design approach of using AI as a "second reader" with interpretable heatmaps. Our web-based deployment aligns with the vision that emphasized the importance of scalable and generalizable deep learning systems in clinical settings [41]. While their system targeted structured electronic health record data, our platform provides an image-based interface that also promotes accessibility, reproducibility, and potential integration with existing clinical workflows. While these advances are significant, challenges noted in prior work, including class imbalance effects [42] and unresolved ethical concerns about data sharing [30,31], remain relevant to our implementation and highlight areas for future improvement. Our work extends the field through its unprecedented scale (35× larger training set than [3,6,36]), practical deployment, and balanced performance metrics while acknowledging persistent limitations that require ongoing research attention.

Strengths and Limitations

One of the major strengths of this study is the use of a large, publicly available histopathology image dataset, which provided a robust foundation for model training and validation. The application of transfer learning, particularly with a pretrained VGG19 model, significantly improved the classification performance and helped mitigate the class imbalance challenge. Given the high accuracy achieved, the model holds promise as a supportive diagnostic tool for pathologists in clinical practice.

However, the study has several limitations. Although the model demonstrated strong performance in detecting IDC, the dataset may not fully capture the wide variability in breast cancer pathology. Factors such as differences in staining protocols, image resolution, and slide preparation techniques can impact the model's generalization. Additionally, while the model performed well on the test set, its effectiveness in real-world clinical environments—where imaging conditions and patient demographics may differ—has yet to be validated.

Another important limitation is the lack of patient demographic information, such as age, ethnicity, or clinical background, within the current dataset. This absence restricts our ability to assess the model's fairness and generalizability across diverse patient populations. Furthermore, the study was conducted without direct collaboration with clinicians, which limited opportunities for incorporating clinical context or interpretation into the modeling process.

Conclusion and Future Work

This work used a CNN model for binary classification of a medical image dataset. First, the model was trained without any adjustments for class imbalance, which resulted in an accuracy of 89% on the validation set after 10 epochs. While reasonable, this model could have been biased due to the class imbalance between the 2 categories: healthy and unhealthy images. To deal with this, a weighted loss function was added to give higher weight to the minority class. This weighted loss–trained model slightly improved accuracy and reduced validation loss; hence, the imbalance was managed better. On the other hand, this gave a fluctuating graph for validation accuracy across epochs, probably indicating issues such as overfitting or further hyperparameter tuning. This problem was resolved by performing transfer learning with the VGG19 network pretrained on the ImageNet dataset. The improved process gave an accuracy of 90% on the test set.

The innovation of this work relies on a systematic approach to handling class imbalance in medical image classification, integrating a weighted loss function along with transfer learning. Whereas standard CNN-based systems may face imbalanced datasets resulting in biased prediction, this paper shows how assigning higher importance to the minority class through weighted loss can yield superior performance. Further, incorporating transfer learning using a pretrained VGG19 model outperforms classification in terms of reduced overfitting compared with the simpler baseline CNN model. The contribution of this work is optimizing deep learning models for medical imagery on training stability and fluctuations in validation. Moreover, this study explores the possibility of deploying a trained model on edge devices to bridge the gap between research and real-world applications in medical diagnostics.

The primary novelty of this paper lies in developing our proposed open-source, public application that facilitates quick patient screening and prioritization, particularly for high-risk patients in numerous cases. The app allows medical practitioners across the globe to quickly perform a preliminary assessment for detecting breast cancer [25].

Potential improvements to enhance the project could be using models like generative adversarial networks to generate synthetic data to augment the training set or explore other creative tasks like image-to-image translation. Our approach is based on incorporating transfer learning using a pretrained VGG19 model for optimal performance on labeled histopathological data. Future work can explore the integration of hybrid methods that include self-supervised learning. Self-supervised learning models have demonstrated the potential to learn robust features by recovering spatial context in medical imaging data [43]. These methods can enhance feature generalization, especially in

XSL•FO
RenderX

cases where datasets are limited or domain adaptation is necessary. We could implement k-fold cross-validation to ensure the results are robust and not dependent on the training-test split. Exploring the deployment of the trained model on edge devices, such as mobile phones or embedded systems, could be a practical direction for real-world applications, especially for retail or inventory management, and using visual recognition can be done in the future.

Acknowledgments

The authors thank the College of Engineering at San José State University for supporting interdisciplinary collaboration. The authors would also like to sincerely thank the anonymous referees and editorial team for carefully reading the manuscript and providing invaluable suggestions. The authors would like to seize the opportunity to acknowledge Mr. Ankith Indra Kumar for his involvement in the draft of this work at its early stage, and thank him for his willingness and devotion at the beginning

Data Availability

The data used in this study and that support the findings of this work are available publicly on the Kaggle from the Kaggle's public dataset [16].

Conflicts of Interest

None declared.

References

1. Ahmed M, Ali M, Roy J, Ahmed S, Ahmed N. Breast cancer risk prediction based on six machine learning algorithms. 2020 Presented at: 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering; December 16-18, 2020; Gold Coast, Australia. [doi: 10.1109/csde50874.2020.9411572]
2. Breast cancer. World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/breast-cancer [accessed 2024-03-13]
3. Spanhol F, Oliveira L, Petitjean C, Heutte L. Breast cancer histopathological image classification using convolutional neural networks. 2016 Presented at: 2016 International Joint Conference on Neural Networks; July 24-29, 2016; Vancouver, Canada. [doi: 10.1109/ijcnn.2016.7727519]
4. Melek A, Fakhry S, Basha T. Spatiotemporal mammography-based deep learning model for improved breast cancer risk prediction. 2023 Presented at: 45th Annual International Conference of the IEEE Engineering in Medicine and Biology Society; July 24-27, 2023; Sydney, Australia. [doi: 10.1109/embc40787.2023.10340602]
5. NCCN. Breast cancer screening and diagnosis guidelines version 2. National Comprehensive Caner Network. URL: https://www.nccn.org/guidelines/guidelines-detail?category=2&id=1421 [accessed 2025-07-24]
6. Cruz-Roa A, Gilmore H, Basavanhally A, Feldman M, Ganesan S, Shih NN, et al. Accurate and reproducible invasive breast cancer detection in whole-slide images: a deep learning approach for quantifying tumor extent. Sci Rep 2017 Apr 18;7:46450 [FREE Full text] [doi: 10.1038/srep46450] [Medline: 28418027]
7. Cruz-Roa A, Gilmore H, Basavanhally A, Feldman M, Ganesan S, Shih N. Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks. 2014 Presented at: Medical Imaging 2014: Digital Pathology; February 15-20, 2014; San Diego, California. [doi: 10.1117/12.2043872]
8. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017 Dec;42:60-88. [doi: 10.1016/j.media.2017.07.005] [Medline: 28778026]
9. Artificial intelligence in health care: promises and challenges. IDB. URL: https://publications.iadb.org/en/publications/english/viewer/Artificial-Intelligence-in-Health-Care-Promises-and-Challenges.pdf [accessed 2025-07-29]
10. Cruz-Roa A, Basavanhally A, Gonzalez F, Gilmore H, Feldman M, Ganesan S. Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks. 2014 Presented at: Medical Imaging 2014: Digital Pathology; February 15-20, 2014; San Diego, California. [doi: 10.1117/12.2043872]
11. Manjunathan N, Gomathi N, Muthulingam S. Early detection of breast cancer using machine learning. 2023 Presented at: 2023 International Conference on Sustainable Computing and Smart Systems; June 14-16, 2023; Coimbatore, India. [doi: 10.1109/icscss57650.2023.10169777]
12. Bansal M, Kumar M, Sachdeva M, Mittal A. Transfer learning for image classification using VGG19: Caltech-101 image data set. J Ambient Intell Humaniz Comput 2023;14(4):3609-3620 [FREE Full text] [doi: 10.1007/s12652-021-03488-z] [Medline: 34548886]
13. Wakili MA, Shehu HA, Sharif MH, Sharif MHU, Umar A, Kusetogullari H, et al. Classification of breast cancer histopathological images using DenseNet and transfer learning. Comput Intell Neurosci 2022;2022:8904768 [FREE Full text] [doi: 10.1155/2022/8904768] [Medline: 36262621]

XSL•FO
RenderX

14. Phalaagae P, Zungeru A, Yahya A, Sigweni B, Rajalakshmi S. A hybrid CNN-LSTM model with attention mechanism for improved intrusion detection in wireless IoT sensor networks. IEEE Access 2025;13:57322-57341. [doi: 10.1109/ACCESS.2025.3555861]
15. Adam R, Dell'Aquila K, Hodges L, Maldjian T, Duong TQ. Deep learning applications to breast cancer detection by magnetic resonance imaging: a literature review. Breast Cancer Res 2023 Jul 24;25(1):87 [FREE Full text] [doi: 10.1186/s13058-023-01687-4] [Medline: 37488621]
16. Mooney P. Breast histopathology images. Kaggle. 2019. URL: https://www.kaggle.com/datasets/paultimothymooney/breast-histopathology-images/data [accessed 2024-02-10]
17. Yamashita R, Nishio M, Do RKG, Togashi K. Convolutional neural networks: an overview and application in radiology. Insights Imaging 2018 Aug;9(4):611-629 [FREE Full text] [doi: 10.1007/s13244-018-0639-9] [Medline: 29934920]
18. Li Z, Liu F, Yang W, Peng S, Zhou J. A survey of convolutional neural networks: analysis, applications, and prospects. IEEE Trans Neural Netw Learning Syst 2022 Dec;33(12):6999-7019. [doi: 10.1109/tnnls.2021.3084827]
19. Chen C, Mat Isa NA, Liu X. A review of convolutional neural network based methods for medical image classification. Comput Biol Med 2025 Feb;185:109507. [doi: 10.1016/j.compbiomed.2024.109507] [Medline: 39631108]
20. Mascarenhas S, Agarwal M. A comparison between VGG16, VGG19 and ResNet50 architecture frameworks for image classification. In; 2021 Presented at: 2021 International Conference on Disruptive Technologies for Multi-Disciplinary Research and Applications; November 19-21, 2021; Bangalore, India p. 96-99. [doi: 10.1109/centcon52345.2021.9687944]
21. Rajinikanth V, Joseph Raj AN, Thanaraj KP, Naik GR. A customized VGG19 network with concatenation of deep and handcrafted features for brain tumor detection. Applied Sciences 2020 May 15;10(10):3429. [doi: 10.3390/app10103429]
22. Dey N, Zhang Y, Rajinikanth V, Pugalenthi R, Raja NSM. Customized VGG19 architecture for pneumonia detection in chest X-rays. Pattern Recognit Lett 2021 Mar;143:67-74. [doi: 10.1016/j.patrec.2020.12.010]
23. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet large scale visual recognition challenge. ArXiv. Preprint posted online on September 1, 2014 .
24. Han Z, Wei B, Zheng Y, Yin Y, Li K, Li S. Breast cancer multi-classification from histopathological images with structured deep learning model. Sci Rep 2017 Jun 23;7(1):4172 [FREE Full text] [doi: 10.1038/s41598-017-04075-z] [Medline: 28646155]
25. FastBreastCancerAI. Hugging Face. URL: https://huggingface.co/spaces/itahir/FastBreastCancerAI [accessed 2025-02-05]
26. Chen RJ, Wang JJ, Williamson DFK, Chen TY, Lipkova J, Lu MY, et al. Algorithmic fairness in artificial intelligence for medicine and healthcare. Nat Biomed Eng 2023 Jun;7(6):719-742 [FREE Full text] [doi: 10.1038/s41551-023-01056-8] [Medline: 37380750]
27. Introducing AI fairness 360. IBM. URL: https://research.ibm.com/blog/ai-fairness-360 [accessed 2025-07-29]
28. General data protection regulation (GDPR). GDPR.eu. URL: https://gdpr.eu/tag/gdpr/ [accessed 2025-07-29]
29. 46.104 exempt research. Code of Federal Regulations. URL: https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-A/part-46/subpart-A/section-46.104 [accessed 2025-07-24]
30. Nebeker C, Torous J, Bartlett Ellis RJ. Building the case for actionable ethics in digital health research supported by artificial intelligence. BMC Med 2019 Jul 17;17(1):137 [FREE Full text] [doi: 10.1186/s12916-019-1377-7] [Medline: 31311535]
31. Larson DB, Magnus DC, Lungren MP, Shah NH, Langlotz CP. Ethics of using and sharing clinical imaging data for artificial intelligence: a proposed framework. Radiology 2020 Jun;295(3):675-682. [doi: 10.1148/radiol.2020192536] [Medline: 32208097]
32. Thorogood A, Dalpé G, Knoppers BM. Return of individual genomic research results: are laws and policies keeping step? Eur J Hum Genet 2019 Apr;27(4):535-546 [FREE Full text] [doi: 10.1038/s41431-018-0311-3] [Medline: 30622328]
33. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health 2019 Oct;1(6):e271-e297 [FREE Full text] [doi: 10.1016/S2589-7500(19)30123-2] [Medline: 33323251]
34. BI-RADS ACR breast imaging reporting and data system. American College of Radiology. URL: https://www.acr.org/Clinical-Resources/Clinical-Tools-and-Reference/Reporting-and-Data-Systems/BI-RADS [accessed 2025-07-24]
35. Li Q, Li W. Using deep learning for breast cancer diagnosis-technical report. Hong Kong: Chinese University of Hong Kong; 2017.
36. Bayramoglu N, Kannala J, Heikkila J. Deep learning for magnification independent breast cancer histopathology image classification. 2016 Presented at: 23rd International Conference on Pattern Recognition; December 4-8, 2016; Cancun, Mexico. [doi: 10.1109/icpr.2016.7900002]
37. Wang D, Khosla A, Gargeya R, Irshad H, Beck AH. Deep learning for identifying metastatic breast cancer. ArXiv. Preprint posted online on June 18, 2016 [FREE Full text] [doi: 10.1007/s40615-015-0191-y] [Medline: 27294767]
38. Improving breast cancer screening with artificial intelligence. Google for Health. URL: https://health.google/caregivers/mammography/ [accessed 2025-07-29]
39. A.I. software evaluation and development for breast cancer detection in tomosynthesis scans. Mayo Clinic. 2023. URL: https://www.mayo.edu/research/clinical-trials/cls-20532539 [accessed 2025-07-29]

XSL•FO
RenderX

40. Wu N, Phang J, Park J, Shen Y, Huang Z, Zorin M, et al. Deep neural networks improve radiologists' performance in breast cancer screening. IEEE Trans Med Imaging 2020 Apr;39(4):1184-1194 [FREE Full text] [doi: 10.1109/TMI.2019.2945514] [Medline: 31603772]
41. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med 2018;1:18 [FREE Full text] [doi: 10.1038/s41746-018-0029-1] [Medline: 31304302]
42. Reza M, Ma J. Imbalanced histopathological breast cancer image classification with convolutional neural network. 2018 Presented at: 14th IEEE International Conference on Signal Processing; August 12-16, 2018; Beijing, China p. 12-16. [doi: 10.1109/icsp.2018.8652304]
43. Chen L, Bentley P, Mori K, Misawa K, Fujiwara M, Rueckert D. Self-supervised learning for medical image analysis using image context restoration. Med Image Anal 2019 Dec;58:101539 [FREE Full text] [doi: 10.1016/j.media.2019.101539] [Medline: 31374449]

Abbreviations

AI: artificial intelligence
CNN: convolutional neural network
IDC: invasive ductal carcinoma
ReLU: rectified linear unit
ROC: receiver operating characteristic
VGG19: Visual Geometry Group 19

Edited by A Mavragani; submitted 06.06.24; peer-reviewed by T Issa, HS Das, J Li, S Tayebi Arasteh, M Aria; comments to author 24.09.24; revised version received 06.12.24; accepted 10.07.25; published 21.08.25.

Please cite as:
Chilumukuru NS, Priyadarshini P, Ezunkpe Y
Deep Learning for the Early Detection of Invasive Ductal Carcinoma in Histopathological Images: Convolutional Neural Network Approach With Transfer Learning
JMIR Form Res 2025;9:e62996
URL: https://formative.jmir.org/2025/1/e62996
doi:10.2196/62996
PMID:

©Naga Shreya Chilumukuru, Pragya Priyadarshini, Yawo Ezunkpe. Originally published in JMIR Formative Research (https://formative.jmir.org), 21.08.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Characteristics Associated With the Use of the Mindfulness Meditation App Headspace in a Large Public Health Deployment: Cross-Sectional Survey Study

Judith Borghouts[1], PhD; Elizabeth V Eikey[2,3], PhD; Cinthia De Leon[1], MPH; Stephen M Schueller[4,5], PhD; Margaret Schneider[6], PhD; Nicole A Stadnick[7,8,9], MPH, PhD; Kai Zheng[4], PhD; Dana B Mukamel[1], PhD; Dara H Sorkin[1], PhD

[1]Department of Medicine, University of California, Irvine, Irvine, United States
[2]Herbert Wertheim School of Public Health and Human Longevity Science, University of California, San Diego, San Diego, United States
[3]The Design Lab, University of California, San Diego, San Diego, United States
[4]Department of Informatics, University of California, Irvine, Irvine, United States
[5]Department of Psychology, University of California, Irvine, Irvine, United States
[6]Department of Public Health, University of California, Irvine, Irvine, United States
[7]Department of Psychiatry, University of California, San Diego, San Diego, United States
[8]Dissemination and Implementation Science Center, UC San Diego Altman Clinical and Translational Research Institute, La Jolla, CA, United States
[9]Child and Adolescent Services Research Center, San Diego, CA, United States

Corresponding Author:
Judith Borghouts, PhD
Department of Medicine
University of California, Irvine
100 Theory
Irvine, 92617
United States
Phone: 1 9498240246
Email: jborghou@uci.edu

Abstract

Background: Mindfulness-based apps can be an effective and accessible resource for mental health support. However, little is known about their use outside of research settings and what user characteristics relate to app use.

Objective: This study aimed to examine the characteristics of people who decided to use, not use, or stop using Headspace within the context of a large-scale public deployment, which offered the mindfulness meditation app Headspace as a free mental health resource to community members.

Methods: Nearly 100,000 community members received Headspace. All members (N=92,311) received an email inviting them to complete a voluntary and uncompensated survey. In total, 2725 participants completed the survey. The 20-minute survey asked about the use of Headspace, user experience, mental health problems, mental health stigma, and mental health use. Logistic regression models were used to examine relationships between predictors and nonuse, past use, or current use of Headspace.

Results: Participants who were still using Headspace at the time of completing the survey (2076/2725, 76.18%) were more likely to experience mental health challenges and distress and make more use of other digital mental health resources (ie, online tools and connecting with people online) than people who were not using Headspace. In addition, current users of Headspace rated the app higher on user experience compared with past users. The most common reasons for abandoning Headspace were that people were already using other strategies to support their mental health (198/570, 34.7%), no longer needed Headspace (73/570, 12.8%), or did not think Headspace was useful (46/570, 8.1%).

Conclusions: Results indicate that a person's mental health challenges, a perceived need for support, and familiarity with digital resources were associated with continued use of Headspace. While the most common reason for not using Headspace was that people were already using other resources, it is important to consider the continuity of mental health support beyond these free programs for those who may not have easy access to other resources. We discuss potential implications of our findings for offering and using apps such as Headspace as a mental health resource, along with factors that influence engagement with this app.

XSL•FO
RenderX

(JMIR Form Res 2025;9:e73457) doi:10.2196/73457

KEYWORDS

Headspace; mental health; mindfulness; mobile health; mHealth; real-world use

Introduction

Background

Mental health is an increasing societal concern. The Centers for Disease Control and Prevention, for example, has formally recognized mental and psychological health as a major and global public health issue [1]. Around 1 in 5 adults in the United States are estimated to experience a mental illness every year, and a 2022 survey found that 90% of Americans believe that the United States is facing a mental health crisis [2].

Digital media can considerably increase access to mental health care. The development of digital mental health interventions (DMHIs) has accelerated in recent years due to a desire to find more scalable and cost-effective options that can overcome access barriers to care [3]. Examples of digital mental health mediums include internet-based therapy, as well as self-guided platforms such as websites and apps that provide psychoeducation, journaling tools, peer support, and mindfulness meditation. Mindfulness is a state of bringing nonjudgmental attention to the present experience. Mindfulness meditation is a practice that sets aside time to develop mindfulness, and mindfulness mediation apps tend to be some of the most popular and commonly downloaded DMHIs [4]. An analysis of 578 DMHIs found that mindfulness was the third most common feature, behind only psychoeducation and goal setting and habit, with 38% of products containing mindfulness features [5]. Systematic analyses and reviews showed that mindfulness meditation can have several mental health benefits [6,7]. Benefits include reducing stress, anxiety [8-10], depression [11,12], and psychological distress [13]. Thus, better understanding aspects about mindfulness meditation DMHIs provides useful information to the field.

DMHIs are beginning to be used as broad-scale supports for mental health, especially from a public health perspective where they can be made widely available. To increase awareness of and access to DMHIs, various initiatives by schools, workplaces, and other organizations have offered DMHIs to their community members free of charge [14,15]. DMHI companies themselves have also made their products available to specific populations or the public in times of need. For example, during the COVID-19 pandemic, some products were made freely available to help people cope with mental health concerns [16]. However, little is known about the characteristics of those who make use of these initiatives, as well as information about their engagement and experience with the DMHIs offered. In addition, few data are available about free-range, naturalistic DMHI use outside of research contexts. Research studies often provide compensation to complete study procedures, which may also impact use and engagement with DMHIs. Although these studies can demonstrate the potential and effectiveness of DMHIs [17,18], their effectiveness is ultimately related to their use, and previous studies researching mindfulness apps found a clear relationship between app use and improvements in mindfulness [6,19]. For instance, past research studies focusing specifically on the mindfulness meditation app Headspace with relatively small sample sizes [6,9] showed that Headspace was effective in improving depressive symptoms [9] and psychological distress [20] but only among study participants who used the app frequently. In addition, study participants received the app for free and were given recurring reminders to use the app [9] or daily surveys about their use of the app [20], which may have impacted engagement.

Engagement in real-world studies has shown considerable variation [21,22]. One study found that when metrics were available, research use was more than 4 times higher than real-world use across the same DMHI programs [23]. Unfortunately, data from broad deployments and real-world contexts are rarely available for others to learn from.

Objectives

This study draws from a California-based innovation program which offered the mindfulness meditation app Headspace for free to community members as a mental health resource. The evaluation of this program provides a unique opportunity to understand factors in the context of real-world use rather than a research environment. The aim of the study was to understand the characteristics of people who decide to use, stop using, or not use an app such as Headspace within the context of the program, as well as their experience with the app.

Gathering data in real-world deployments often requires balancing the views and priorities of various stakeholders. In our study, we worked closely with community members from participating counties and cities (referred to as sites throughout the rest of the paper) to develop the survey and select validated scales informed by both literature and the lived experiences of participants. This community-based approach fostered buy-in from sites to survey their communities and ensured the measures were relevant to the community, promoting participation and meaningful data collection. We were mindful of preserving the integrity of the psychometric properties of the scales and, where possible, used original scales. We discuss our rationale for cases where scales were adapted to the study context in the Survey Development section and report on internal consistency estimates in the Measures section to support their use in this study.

Methods

Program Description

The dissemination of Headspace was a part of the Help@Hand project, a state-wide innovation project in California that aimed to understand how DMHIs fit within the public mental health system of care. The project explored how technology could be used to reach people who are likely to go either unserved or underserved by traditional mental health care. The project was

XSL•FO
RenderX

funded and directed by multiple participating sites in California and overseen by the California Mental Health Service Authority.

To be eligible to enroll in the Headspace program, participants had to enter a zip code of the participating site where they lived, worked, or attended school at the time of enrollment. Participants did not have to be new users to Headspace, but they could not have an existing paid subscription at the time of enrollment. Participants enrolled via a specific web page launched by each site. After enrolling, participants had free access to Headspace up until the end of the program (see Table 1 for the different end dates of the programs across sites). Participants could enroll at any time during the program.

Table 1. Implementation launch and end dates of the Headspace Program in the 4 sites (Los Angeles County, San Mateo County, City of Berkeley, and Santa Barbara County). The first program launched in April 2020, and the last program ended in September 2023.

Site	Launch date	End date	Length of program
Los Angeles County	April 2020	February 2023	2 y and 11 mo
San Mateo County	September 2020	September 2021	1 y and 1 mo
City of Berkeley	October 2021	September 2023	2 y
Santa Barbara County	October 2021	September 2023	2 y

The Headspace program launch and end dates differed by site (Table 1), but the first program launched in April 2020 by Los Angeles County, and the last programs were launched in October 2021 by the City of Berkeley and Santa Barbara County.

Headspace Description

Headspace is a commercially available meditation app that aims to improve mental wellness. The app offers meditation and mindfulness exercises in both audio and video format. Its content is organized into 4 categories: meditate, sleep, move, and focus.

- Meditate offers a library of single meditations and courses consisting of a series of meditations that vary in length and topic. For example, it offers meditation exercises and courses focused on reducing anxiety and stress, as well as on improving mood and productivity. It also offers a beginner's course for users who are new to meditation.
- Sleep has guided sessions, meditations, and soundscapes intended for nighttime and to help with relaxation and falling asleep.
- Move has low-impact workout videos, such as yoga exercises, ranging from 10 to 30 minutes. The workouts are focused on incorporating mindfulness during movement.
- Focus offers music and meditation playlists that are intended to help focus during work.

The content is accessible through a mobile app, available on Android and Apple iOS, or the Headspace website, and the content is available in English, Spanish, French, German, and Portuguese.

Headspace was selected by the Help@Hand project after a vetting and approval process intended to assess the extent to which it met Help@Hand's needs (eg, by considering how Headspace addressed privacy and security of their product and whether content was offered in multiple languages). It was able to meet the necessary specifications for use in this project and was available in languages other than English. Its popularity and reputation were also attractive. Headspace has more than 70 million subscribers worldwide [24] and is one of the most downloaded mindfulness apps in the world [25]. As Headspace, together with the meditation app Calm, represents the majority (approximately 90%) of users of DMHIs [4], capturing characteristics of the use of Headspace can be valuable in contributing to a better understanding of DMHI use.

Survey Development

A community-based participatory approach was used to engage community members in the survey development. The research team initiated a work group with 9 key individuals from the 4 participating sites; the group met 5 times for 1 to 2 hours between January and July 2021. These key individuals included mental health advocates and project and program managers of behavioral health programs and site departments. During the sessions, work group members discussed and made decisions related to the survey method, instruments, and recruitment strategies. Members were given activities to provide feedback on the survey items both during and after the sessions.

Work group members provided multiple rounds of feedback on the draft surveys created by the research team. For example, concerns were raised about the tone of particular survey items and questions that may feel intrusive to participants. In response to this feedback, we supplemented 2 validated scales with additional survey items to balance positively worded items while maintaining items from validated scales that captured the necessary information. One item was added to the scale to measure loneliness: "How often do you feel connected with others?" Similarly, 1 item was added to the scale to measure distress: "During the last 30 days, about how often did you feel hopeful?" These additional items were included to enhance content balance but were not used in the calculation of loneliness and distress scores, respectively.

In addition, some survey items aimed at measuring mental health stigma were removed. The initial survey included measures of internalized stigma, resilience, and mental health treatment stigma based on key individual feedback from a 2019 in-person work group focused on stigma [26]. Following community members' input, the mental health treatment stigma measure was replaced with perceived stigma and stigma resistance measures.

In addition, work group members vocalized the importance of highlighting wellness and well-being, while also being representative of the communities to whom the survey would be sent. Work group members recommended commissioning

XSL•FO
RenderX

artwork from community members with lived mental health needs experiences, who were familiar with the diverse communities and could approach the task from a recovery perspective. Four art pieces were created by community members that incorporated themes of peer support, resilience, and multiformity. These art pieces were included in the emails sent out to participants to complete the survey (Figure 1 shows one of the art pieces).

The survey included questions related to people's use of Headspace and experience, their use of mental health resources, and their health and well-being. On the basis of participants' answers regarding their use of Headspace, different follow-up questions were asked (see Multimedia Appendix 1 for the survey instrument and the Measures section for a description of the survey measures).

Figure 1. One of the art pieces created by community members to include in recruitment emails.

Ethical Considerations

The study was approved by the University of California, Irvine, Institutional Review Board (20195406). The survey included a study information sheet that was reviewed and approved by the institutional review board. Respondents were able to download a copy of the sheet. All research data collected were stored securely and confidentially on a password-protected secure server. Program participants were made aware when signing up for the innovation program that their email address would be shared with the research team and that they would be contacted to take part in a survey. Completing or not completing the survey did not affect people's access to Headspace. Further details regarding the survey study are reported following the Checklist for Reporting Results of Internet e-Surveys guidelines [27] in Multimedia Appendix 2.

Study Design and Procedure

This paper reports on survey data collected between July 1, 2021, and October 21, 2022.

After obtaining institutional review board and key individual approvals, the survey was finalized and deployed on July 1, 2021. Everyone in the participating sites who had signed up for Headspace up until that point were sent a survey. After that, the survey was sent once per week to all new people who had signed up in the past week. Participants were sent up to 5 email reminders (sent once every 3 days) to complete the survey. Once participants completed a survey, their email address was removed from the reminder email list so they did not receive any further reminders.

The research team was granted access to a dashboard which contained email addresses of participants who had signed up for the Headspace program via their site. All enrolled participants were sent an email to complete a web-based survey on their use of Headspace and experience, their use of other mental health resources, and their health and well-being. The survey was distributed via REDCap (Research Electronic Data Capture, version 15.5.0; Vanderbilt University) [28,29], a secure, web-based software platform designed to support data capture for research studies. Participants had to self-report to be aged ≥18 years to be eligible to complete the survey but did not need to have used Headspace as we also wanted to be able consider "nonuse" within the program, that is, people who signed up for the program but did not use Headspace. The survey could be completed on a computer, mobile phone, or tablet; was available in English and Spanish; took approximately 20 minutes to complete; and completing the survey was voluntary and uncompensated.

The survey was sent to 92,311 people and received by 92,261 people (50 emails bounced back). A total of 3399 participants started the survey, resulting in a response rate of 3.68% (3399/92,261). A total of 2725 participants completed the survey (completion rate=2.95%). Survey responses were anonymous and not linked to respondents' email addresses. Only 1.03% (28/2725) of survey respondents completed the survey in Spanish; the other respondents completed the survey in English.

Measures

Overview

A description of each survey measure is given below. We primarily focus on dichotomous measures as these have been validated using standard cutoff points. Further details about the measures are included in Multimedia Appendix 3.

XSL•FO
RenderX

Length of Use of Headspace

Respondents were asked when they signed up for the Headspace program. Respondents could choose from "Less than a week ago," "2 weeks to 1 month ago," "2 to 6 months ago," "6 months to a year ago," and "Longer than a year ago."

Adoption of Headspace

Respondents were asked 1 multiple-choice question on whether they were using Headspace. Respondents could choose from the following answers: "Yes," "No, but I did use it in the past," or "No, and I never used it." On the basis of respondents' answers, we made a distinction between 3 types of users, which throughout the paper we call current users (who answered yes), past users (who used it in the past), and nonusers (who never used it).

Frequency of Use of Headspace

Respondents were asked 1 question on how frequently they used Headspace. While there is no gold standard on how to group frequency of use [30], past research has identified "active" users as individuals who use a mental health app daily [21], weekly or more [31], or monthly or more [32]. On the basis of the distribution of our sample, we grouped and dichotomized frequency of use into those who use it several times a week or more and those who use it less than that. Frequency of use was thus examined as a dichotomous measure as follows: 1 for "Daily/Several times a week" and 0 for "Several times a month/About once a month/I only used it once."

Mental Health Challenges

Respondents were asked 1 multiple-choice question on whether they had experienced any mental health challenges ("Have you or do you experience mental health challenges?"). Respondents could indicate that they had "not experienced any challenges," "had been diagnosed with a mental health condition," "had not been diagnosed but were experiencing challenges," or they could self-describe their experience. On the survey, it was explained that many different terms could be used to refer to these experiences, such as mental health challenge, emotional distress, and psychological disorder. The variable was operationalized as a dichotomous measure (1 for experiencing or having been diagnosed with a mental health challenge and 0 for not experiencing any challenges) in the analyses.

Distress

Psychological distress was measured using the Kessler Psychological Distress (K10) scale [33]. Respondents were asked to rate 10 statements related to how they had been feeling during the past 30 days (eg, "During the past 30 days, how often did you feel tired out for no good reason?"). The statements were rated on a 5-point Likert scale ranging from "None of the time" (score=1) to "All the time" (score=5), with a total sum score in the range of 10 to 50. A higher score indicates a higher level of psychological distress.

The internal consistency of this measure in this study was excellent (Cronbach α=0.91) [34]. The score was operationalized as a dichotomous measure in the analyses. In accordance with previous work [35-37], a cutoff score of >19 was used to categorize respondents as having distress.

Loneliness

Loneliness was measured using the 3-item Loneliness Scale [38], a shortened version of the University of California, Los Angeles Loneliness Scale [39]. Participants were asked to rate 3 statements related to loneliness on a 3-point Likert scale ranging from "Hardly ever" (score=1) to "Often" (score=3; eg, "How often do you feel left out?"), with a total sum score in the range of 3 to 9.

The internal consistency of this measure in this study was adequate (Cronbach α=0.78). The score was operationalized as a dichotomous measure in the analyses, and consistent with previous work [40], a cutoff score of ≥6 was used to categorize respondents experiencing loneliness.

Mental Health Stigma

Participants were asked to rate 4 statements related to mental health stigma (eg, "Being around people who don't have mental health challenges makes me feel out of place or inadequate") using a Likert scale ranging from "Strongly disagree" (score=1) to "Strongly agree" (score=5). The questions related to internalized stigma (1 item), perceived stigma (1 item), resilience (1 item), and stigma resistance (1 item). Items were taken from the Internalized Stigma of Mental Illness Inventory [41] for internalized stigma and stigma resistance, the Perceived Stigma subscale of the Depression Stigma Scale [42] for perceived stigma, and the Recovery Assessment Scale–Revised [43] for resilience. One item was used from each of these 4 subscales instead of all subscale items, based on input from the survey work group to keep the survey brief to mitigate response burden.

Items were chosen based on feedback to balance positively and negatively worded items, as well as key individual feedback from a 2019 in-person work group focused on stigma [26] and a subsequent factor analysis of responses from 6304 participants.

As the items measured different types of stigma and there was low internal consistency between stigma items (Cronbach α=0.58), the stigma items were analyzed separately and were not combined into a single score. This aligns with our community-based approach, which emphasized using measures that were both empirically grounded and meaningful to community members. Item scores were operationalized as dichotomous measures (1 for "Agree" and "Strongly agree" and 0 for "Neither agree nor disagree," "Disagree," or "Strongly Disagree").

Mental Health Care Use

Four dichotomous (yes or no) response items from the California Health Interview Survey [44] were used to identify whether respondents had used online tools other than Headspace for problems related to mental health, whether they had connected online with people with similar mental health problems, whether they had used online tools to connect to a mental health professional, and whether they had seen a mental health professional in the past 12 months.

User Experience

Respondents who indicated they had used Headspace were asked questions related to their experience using Headspace.

XSL•FO
RenderX

Participants were asked to rate 12 statements (eg, “I find Headspace useful in my daily life”) on a 5-point Likert scale ranging from “Strongly disagree” (score=1) to “Strongly agree” (score=5). Survey items were based on the Unified Theory of Acceptance and Use of Technology questionnaire [45], a questionnaire used to evaluate technology acceptance and adoption. The items showed high internal consistency (Cronbach α=0.91) and were combined as a single mean score that could range from 1 to 5. A higher score indicates a more positive experience with and opinion of Headspace. The score was operationalized as a dichotomous measure in the analyses, using a cutoff of >3 to constitute a positive experience [46].

Reasons for Abandonment and Nonuse and Intention to Use Headspace

Overview

Respondents who indicated they were not currently using Headspace were asked to indicate their reasons for not using Headspace. Respondents were instructed to “select all that apply” from a list of 14 options or give an answer in their own words. The list of reasons was based on common barriers to engagement from the literature regarding mental health technology [47], as well as common answers given by community members in prior studies of the Help@Hand project. The full list of options is included in Multimedia Appendix 1.

Respondents were also asked a multiple-choice question on whether they intended to use Headspace in the future. Respondents could choose from “Yes,” “No,” and “I’m not sure.”

Digital Literacy

Respondents were asked to rate one statement related to digital literacy (“I am confident using technology to look up information”) taken from the Mental Health Literacy Scale [48] on a 5-point Likert scale ranging from “Strongly disagree” (score=1) to “Strongly agree” (score=5).

Demographic Measures

Respondents’ age was collected as a categorical variable to adhere to project requirements, with answer options “18 to 25,” “26 to 59” and “>60.” Respondents’ age was operationalized as mutually exclusive dummy variables (0-1) for 18 to 25, 26 to 59, and >60 years.

Respondents’ gender identity was operationalized as mutually exclusive dummy variables (0-1) for man or male, woman or female, and another identity (the survey included options to select the following: transgender man; transgender woman; genderqueer, gender nonconforming, or nonbinary; questioning or unsure of gender; or they could self-describe their identity). Respondents’ race or ethnic identity was operationalized as mutually exclusive dummy variables (0-1) for non-Hispanic White, Hispanic or Latinx, Asian, Black or African American and another identity (the survey included options to select “American Indian, Native American, or Native Alaskan”; "Native Hawaiian or other Pacific Islander”; “Two or more races”; or they could self-describe their identity).

We first report on demographics and mental health symptoms of people who signed up. In the Results section, we explore the relationship between continued use of Headspace and demographics, prevalence of mental health problems, mental health care use, frequency of use of Headspace, user experience with Headspace, and mental health stigma. Finally, we report on respondents’ reasons for abandonment and nonuse of the app.

Participants

The majority (2238/2725, 82.13%) of survey respondents were aged between 26 and 59 years. In total, 288 (10.57%) respondents were aged between 18 and 25 years, and 199 (7.3%) were aged ≥60 years; 1983 (72.77%) respondents identified as woman or female. Of 2725 respondents, 1246 (45.72%) respondents identified as non-Hispanic White, 496 (18.2%) identified as Hispanic or Latinx, 421 (15.45%) identified as Asian, and 119 (4.37%) identified as Black or African American. In total, 2003 (N=2725, 73.5%) 7respondents indicated they experienced mental health challenges; 1462 (N=2725, 53.65%) respondents were likely to experience distress, and approximately half (1469/2725, 53.9%) of the respondents scored high on loneliness. The majority (2575/2725, 94.5%) of survey respondents used Headspace in English, 1.47% (40/2725) used it in Spanish, 0.88% (24/2725) used it in German, and 0.15% (4/2725) used the app in French. Further demographic characteristics per user group are shown in Table 2. On the basis of respondents’ answers on whether they were using Headspace, we made a distinction between 3 user groups:

1. Current users; respondents who indicated they were using Headspace at the time of the survey
2. Past users; respondents who indicated they were not using Headspace anymore at the time of the survey but had used it in the past
3. Nonusers; respondents who indicated they never used Headspace

Table 3 shows the demographics for the sites, based on county census data [49]. When comparing demographics of survey respondents with the county data, our survey sample had a higher proportion of women or female respondents, a lower proportion of Hispanic or Latinx respondents, and individuals with higher levels of education.

XSL•FO
RenderX

Table 2. Demographic characteristics of survey respondents overall and by user group.

Demographics	Current users (n=2076), n (%)	Past users (n=570), n (%)	Nonusers (n=79), n (%)	Total (N=2725), n (%)
Age (y)				
18-25	226 (10.9)	55 (9.6)	7 (8.9)	288 (10.6)
26-59	1717 (82.7)	464 (81.4)	57 (72.2)	2238 (82.1)
>60	133 (6.4)	51 (8.9)	15 (19)	199 (7.3)
Gender[a]				
Men	470 (22.6)	117 (20.5)	12 (15.2)	599 (22.0)
Women	1513 (72.9)	413 (72.5)	57 (72.2)	1983 (72.8)
Other gender identity	93 (4.5)	40 (7.0)	10 (12.7)	143 (5.2)
Race or ethnicity[a]				
Non-Hispanic White	965 (46.5)	243 (42.6)	38 (48.1)	1246 (45.7)
Hispanic or Latinx	384 (18.5)	100 (17.5)	14 (17.7)	498 (18.3)
Asian	296 (14.3)	113 (19.8)	12 (15.2)	421 (15.4)
Black or African American	92 (4.4)	27 (4.7)	0 (0)	119 (4.4)
Other race or ethnic identity	341 (16.4)	87 (15.3)	15 (19.0)	443 (16.3)
Highest education level				
Bachelor's degree	923 (44.5)	258 (45.3)	30 (38)	1211 (44.4)
Graduate or professional degree	747 (36)	228 (40)	29 (36.7)	1004 (36.8)
Some college with no degree	179 (8.6)	35 (6.1)	7 (8.9)	221 (8.1)
Associates degree	103 (5.0)	19 (3.3)	4 (5.1)	126 (4.6)
High school or less than high school	65 (3.1)	13 (2.3)	1 (1.3)	79 (2.9)
Language they used Headspace in				
English	2015 (97.1)	560 (98.2)	N/A[b]	2575 (94.5)
Spanish	36 (1.7)	4 (0.7)	N/A	40 (1.5)
German or French[c]	22 (1.1)	6 (1.1)	N/A	28 (1)

[a]See Multimedia Appendix 1 for all gender and race or ethnicity options.

[b]N/A: not applicable.

[c]Proportions of German and French users are merged given the small number of respondents in each category.

Table 3. Comparison of survey sample with county population sociodemographic characteristics. Our survey sample had a slightly higher proportion of women or female respondents, a lower proportion of Hispanic or Latinx respondents, and individuals with higher levels of education.

Demographics	Survey sample, n (%)	Aggregated county census data of Los Angeles, San Mateo, and Santa Barbara and City of Berkeley (%)
Gender		
Woman or female	1983 (72.77)	50.3
Race and ethnicity		
Non-Hispanic White	1246 (45.72)	39.3
Hispanic, Latino, Latina, or Latinx	498 (18.28)	33.1
Asian	421 (15.45)	18.8
Black or African American	119 (4.37)	5.4
Highest education level		
Bachelor's degree or higher	2215 (81.28)	48.8

XSL•FO
RenderX

Data Analysis

We analyzed the survey data using descriptive statistics (such as frequency counts) and inferential statistics. For each of the measures, we report proportions, means, and SD by user group. Chi-square tests were used to analyze differences between user groups in terms of age, gender, and race or ethnicity.

Logistic regressions were conducted for each outcome to control for site, length of use of Headspace, and covariates (age was added a covariate as there were significant differences in age between user groups) and to ensure that length of use of Headspace and any variations between sites did not affect the results. We report odds ratios (ORs) and 95% CIs. The analyses were performed using statistical software R (version 4.2.2) [50].

Total scores on loneliness, distress, and user experience with Headspace were only computed for respondents that completed ≥50% of the items on these respective scales. For respondents that completed ≥50% of the scale items, any missing data were imputed using mean imputation. Approximately 2% of participants had missing data imputed; <1% had a missing total score.

Results

Overview

Throughout this section, we report on the number of people who provided a particular survey answer and the total number of people who responded to the survey question.

Table 4 shows the number of respondents who were still using Headspace, were no longer using Headspace, and indicated they never used Headspace at the time of the survey. Approximately half of the respondents (1445/2725, 53.03%) had signed up for the Headspace program over a year ago. In total, 879 (32.26%) respondents had signed up between 2 months to a year ago and 398 (14.61%) had signed up <2 months ago.

Table 4. The number of people who were still using Headspace (current users), were no longer using Headspace (past users), and had never used Headspace (nonusers) at the time of completing the survey (N=2725).

User group	Sample, n (%)
Current users	2076 (76.18)
Past users	570 (20.921)
Nonusers	79 (2.9)

Differences in Adoption by Demographics

There was a significant association between age and adoption of using Headspace (χ^2_4=21.1; *P*<.001): older people (ie, aged ≥60 y) were more likely to have never used Headspace (15/79, 19%) than be a current user (133/2076, 6.41%) or past user (51/570, 8.9%; Table 5).

Although there was no association between adoption and the proportion of women or female respondents, or men or male respondents (χ^2_2=1.9; *P*=.37), there was a significant association between adoption and the proportion of people who identified with another gender identity or did not answer the question: nonusers were more likely to identify with another gender identity or not disclose their gender identity (10/79, 13%) than current users (93/2076, 4.48%) and past users (40/570, 7%; χ^2_4=16.6; *P*=.002). There were no significant race or ethnicity differences in adoption (χ^2_8=15.1; *P*=.06).

Table 5. Demographics by adoption of Headspace. People aged ≥60 years were more likely to be nonusers, and nonusers were more likely to identify with another gender identity or not disclose their gender identity.

	Age[a] (y), n (%)			Gender[b], n (%)			Race or ethnicity, n (%)				
	18-25	26-59	>60	Male	Female	Other identity	Non-Hispanic White	Latinx	Asian	Black	Other race or identity
Current users	226 (10.9)	1717 (82.7)	133 (6.4)	22.6	1513 (72.9)	93 (4.4)	965 (46.5)	384 (18.5)	296 (14.3)	92 (4.4)	339 (16.3)
Past users	55 (9.6)	464 (81.4)	51 (8.9)	117 (20.5)	413 (72.5)	70 (7)	243 (42.6)	100 (17.5)	113 (19.8)	27 (4.7)	87 (15.3)
Nonusers	7 (8.9)	57 (72.2)	15 (19)	12 (15.2)	57 (72.2)	10 (12.7)	38 (48.1)	14 (17.7)	12 (15.2)	0 (0)	15 (19)

[a]*P*<.001.

[b]*P*<.01.

Mental Health

Approximately 73.5% (2003/2725) of respondents indicated they currently experienced mental health challenges. Current users of Headspace were more likely to report having mental health challenges (1560/2076, 75.14%) than nonusers (40/79, 51%; OR 0.41, 95% CI 0.25-0.66, *P*<.001). Respondents who reported to experience distress were more likely to be current users (1184/2076, 57.03%) than nonusers (26/79, 33%; OR 0.44, 95% CI 0.27-0.72, *P*=.001). Figures 2 and 3 show the ORs and 95% CIs for outcomes among current users of Headspace compared with nonusers and past users, respectively.

XSL•FO
RenderX

Approximately 48.92% (1333/2725) of respondents scored high on loneliness, with no significant differences in loneliness between user groups (see Table S1 in Multimedia Appendix 4 for ORs).

Figure 2. Odds ratios and 95% CIs for outcomes among current users compared with nonusers. Current users were more likely to experience mental health challenges and distress and make more use of online tools than nonusers.

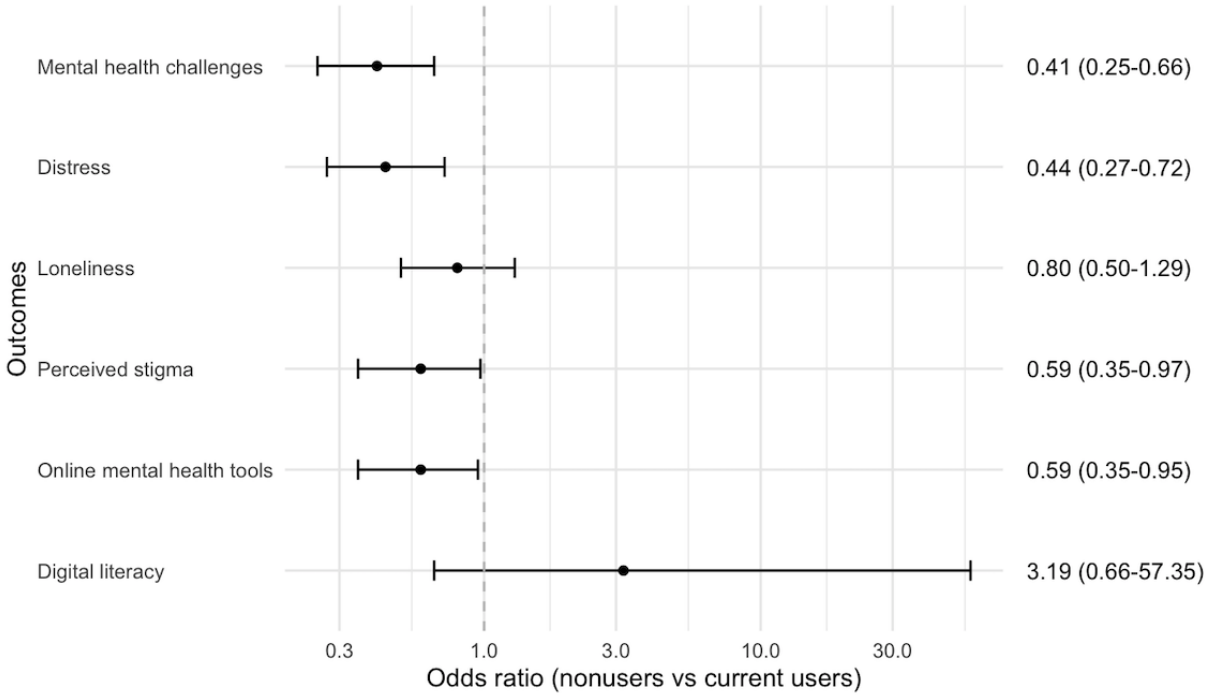

Figure 3. Odds ratios and 95% CIs for outcomes among current users compared with past users. Current users were more likely to rate Headspace high on user experience than past users.

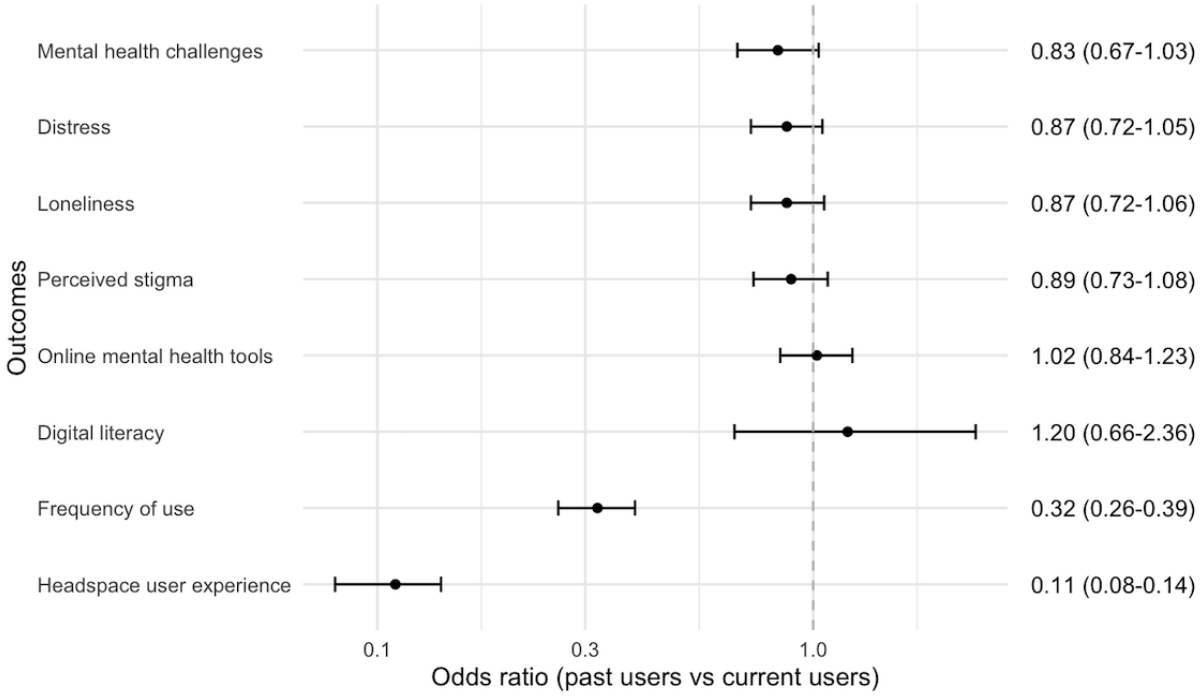

Mental Health Stigma

Respondents who agreed that most people believe that having mental health challenges is a sign of personal weakness (a perceived stigma measure) were more likely to be current users (904/2076, 43.55%) than nonusers (24/79, 30%; OR 0.59, 95% CI 0.35-0.97, *P*=.04). There were no differences on the other stigma items between users (see Table S2 in Multimedia Appendix 4 for ORs).

XSL•FO
RenderX

Mental Health Care Use

Respondents who had made use of online mental health tools were more likely to be current users (900/2076, 43.35%) than nonusers (23/79, 29%; OR 0.59, 95% CI 0.35-0.95, *P*=.04). Respondents who had connected with online with people that had mental health concerns were also more likely to be current users (609/2076, 29.34%) than nonusers (14/79, 18%; OR 0.52, 95% CI 0.26-0.94, *P*=.04).

There were no significant differences in use of professional mental health services or in using online tools to be referred to professional mental health services between user groups (see Table S1 in Multimedia Appendix 4 for ORs).

Digital Literacy

Most respondents (2620/2725, 96.15%) felt confident using technology to look up information, and there were no differences in digital literacy between user groups (see Table S2 in Multimedia Appendix 4 for ORs).

Frequency of Use of Headspace

When thinking about what best describes their past use of Headspace, current users indicated they used Headspace more frequently than past users did before they stopped using it (OR 0.32, 95% CI 0.26-0.39, *P*<.001). Approximately 63.1% (1310/2076) of current users indicated they used Headspace daily or several times a week, compared with 33.4% (190/570) of past users.

Headspace User Experience

Overall, experience with Headspace was positive (mean 4.10, SD 0.68), with current users more likely to report a positive Headspace user experience (mean 4.24, SD 0.58) than past users (mean 3.60, SD 0.77; OR 0.11, 95% CI 0.08-0.14, *P*<.001).

Only a third of respondents (901/2725, 33.06%) indicated they could get help from others with using Headspace if they had any difficulties, and this was true for current users (703/2076, 33.86%), past users (190/570, 33.86%), and nonusers (8/97, 8%).

Reasons for Not Using Headspace and Intention to Use Headspace

The most common reasons for not using Headspace anymore were that past users were already using other strategies and resources to support their mental health (198/570, 34.7%), they just wanted to try Headspace out (167/570, 29.3%), they no longer needed Headspace (73/570, 12.8%), or they felt Headspace was not useful (46/570, 8.1%).

The most common reasons for not using Headspace at all were that nonusers were already using other strategies and resources to support their mental health (19/79, 24%), they thought Headspace would not be useful (10/79, 18%), they could not find the time to use it (5/79, 6%), or they only wanted to use traditional mental health services (5/79, 6%).

Past users who were not currently using Headspace were more likely to indicate they intended to use Headspace in the future at 51.6% (294/570) compared with only 20% (19/97) of nonusers.

Discussion

The aim of this paper was to describe characteristics of people who decided to use, stop using, or not use Headspace within the context of a wide-scale deployment of Headspace for free to community members.

Demographics and Use of Headspace

Similar to other studies assessing meditation app users [51,52], most survey respondents were women or female and non-Hispanic White. Respondents' gender and racial or ethnic background was not related to adoption or continued use of Headspace in our study. This is in contrast to previous studies that found that females were more likely than other genders to engage with DMHIs [53] and use meditation [54].

However, age was found to be related to use. Older respondents (ie, those aged >60 y) were more likely to have not used Headspace. This is consistent with past research on DMHIs. Older adults can experience challenges with technology [55] and may be less comfortable using digital tools, which can impact DMHI use [56,57]. Such users might need additional support both to start using a DMHI and sustain long-term use. This relates to the increasing use of digital navigators to support these processes [58]. One evaluation of the impact of digital care navigation in routine care found a large increase in registration rates, especially when a digital navigator reached out on the day of referral [59]. In public health models, navigation might need to focus specifically on the types of users who need the most help, such as older adults.

Mental Health Symptoms and Use of Headspace

More than half of the respondents reported experiencing mental health challenges, distress, and loneliness, and respondents who experienced mental health challenges and distress were more likely to be current users than nonusers. Presence of mental health symptoms has been reported as a motivator to use DMHIs [53,60,61]. Therefore, it is possible that those with greater mental health concerns may be more likely to continue to use Headspace, potentially due to expected benefits. Indeed, a recent metareview showed that app-delivered mindfulness interventions can significantly improve perceived stress, anxiety, and depression [7]. These findings suggest a potential pattern worth exploring in a longitudinal study, as people may be more motivated to continue using an intervention if it has potential benefits for mental health challenges they are experiencing [47].

Headspace Experience and Use of Other Resources to Support Mental Health

Respondents overall rated Headspace highly on user experience, with current users indicating a more positive experience than past users. Prior studies on Headspace showed that a positive attitude toward use of the app enabled the use of Headspace [62].

Despite a positive experience with Headspace, 20.92% (570/2725) of respondents indicated they had already stopped using Headspace by the time they were surveyed. This is consistent with past work that has shown that nearly 1 in 4 people stop using DMHIs after only 1 use [63]. Given that such

XSL•FO
RenderX

use is likely not an adequate dose of the intervention, supporting sustained use may be important to drive interventions impacts. In our study, the most common reasons for not using or abandoning Headspace were that respondents were already using other strategies to support their mental health or no longer needed Headspace. Other work has highlighted that a positive impact on one's mental health can drive sustained use [32]; however, at some point when people have received sufficient benefits they are likely to stop using a DMHI. This type of "happy abandonment," where people no longer need the intervention was more common in our sample than those who did not find Headspace useful (73/570, 12.8% vs 46/570, 8.1%). Most studies do not receive information for those who stop using a DMHI, and although "happy abandonment" has been posited elsewhere [32], our study is one of the first of its kind to be able to quantify it in a sample of users.

People's use of other strategies and resources further highlights that although apps are often evaluated in isolation, people use a combination of various tools to support their health and well-being [64]. Each tool can have its own function within an individual's life, and rather than understanding each tool in isolation, it may be worthwhile to understand this ecosystem of tools. Future studies can explore how Headspace fits into existing resources people are already using, what motivates them to adopt or discontinue use, and how these resources complement each other. It is also important to better understand individuals' reasons behind switching between different resources, and the strengths and limitations of mobile apps in addressing more serious mental health challenges.

Current users were more likely to have made use of other online tools besides Headspace for their mental health, with almost half of current users having used another online tool. Past research has shown that increasing familiarity with DMHIs can impact user perceptions of DMHIs [65], and a past positive experience with digital health tools can increase likelihood to adopt other DMHIs in the future [66-68]. Our findings support this notion, suggesting that past experience and familiarity with digital tools may not only increase initial adoption but also facilitate continued use of apps such as Headspace. Notably, current users also used the app more frequently than other user groups, further underscoring the importance of early engagement in supporting continued use.

Need for Support

Only a third of respondents reported they had access to help if they had difficulties using the app, with no differences between current and past users. Though overall experience with Headspace was positive in our study, this lack of access highlights an important potential barrier, as technical issues are one of the most common barriers to using DMHIs [47]. People may be enthusiastic about a technology but be dissuaded from using it if they are experiencing technical barriers. This is another opportunity for digital navigators as discussed earlier. The efficiency model of human support [69] posits 5 potential failure points to using DMHIs including usability, engagement, fit, knowledge, and implementation. Overcoming technical barriers and usability issues has been found in other studies to be a main activity supporters engage in [69]. Studies should further explore what types of help people are interested in and design support programs to address these issues.

Implications and Future Directions

Overview

Mindfulness meditation apps can be a useful mental health resource to provide to communities. Our findings identified user characteristics associated with continued app use in the context of an innovation program. While the results of this study are exploratory, they can help generate hypotheses for future research and inform outreach strategies, especially within programs that aim to use DMHIs as a public mental health intervention. We outline several implications that may be important to consider.

Understand the Relationship Between Use and Mental Health Symptoms

Current users were more likely to experience mental health challenges and distress, which is a potential indication that people who need support are more likely to use it when options are made available. Future studies should seek to further elucidate the relationship between the use of Headspace and mental health symptoms. For example, it may be worthwhile to explore if continued use and engagement is associated with any improvement in mental health symptoms over time, and whether the severity of distress impacts perceived usefulness and ability to use meditation apps.

Engagement Metrics Alone Are Only Part of the Story

Despite continued concerns regarding poor engagement rates in real-world contexts of DMHI, future work should look beyond user engagement and understand reasons for not using an intervention. In our study, the number of people who stopped using Headspace because they no longer needed it outnumbered those who stopped using it because they did not find it useful. We should not assume that nonuse equals bad use. It is also possible that people might return to these tools over time if they found them useful. Indeed, we found that past users of Headspace were more likely to report an expectation of using Headspace in the future than those who never started using it (nonusers). Future studies could explore how and why people return to these tools over time, especially if longitudinal data are available.

Consider Ecologies of Care

Respondents used other mental health resources beyond Headspace, and the most common reason for not using Headspace was that they already had other strategies in place to support their mental health. Tools are often not used in isolation but exist within an ecosystem of tools used to manage one's mental health and well-being [64]. Furthermore, DMHIs may fit into a continuum of care that could include formal (ie, treatment) and nonformal (ie, social support and religious communities). Future work could further investigate how Headspace fits into people's existing toolbox of digital and non–digital mental health resources.

XSL•FO
RenderX

Limitations

Our study had considerable strengths, including the large sample and real-world deployment. It is worth considering these strengths in light of the study's limitations as well. A common issue with survey studies evaluating app use is reporting and response bias [51]. The use of Headspace was self-reported, and most of our survey respondents indicated they were current users, who may have had more positive opinions on Headspace and had more prevalent mental health challenges. Indeed, despite our large sample, we only received responses from 3% of the 92,311 users identified in the deployment. Our study sample differed from the overall population demographics of the participating sites, with more respondents having higher levels of education, a greater proportion identifying as women or female, and fewer identifying as Hispanic or Latinx. However, we cannot conclude if this was reflective of those who started using Headspace or merely those who completed our survey. Community members who did not sign up at all may have experienced additional barriers, such as comfort or experience with apps and access to required resources to use apps. Nevertheless, prior studies found similar trends with women [53,70] and the highly educated [71] being more likely to access DMHIs. Future program efforts may benefit from expanding outreach strategies to reach underrepresented groups, such as by featuring men in promotional materials [72] or developing culturally and linguistically appropriate materials to reach diverse communities.

Furthermore, our survey was launched after the program had started. Launch dates differed by site, and respondents had signed up for the program at different time points at the time of the survey launch. While adding length of use as a control variable did not affect any of the reported relationships between continued use and dependent variables, individuals who had signed up and stopped using Headspace a while ago at the time of administering the survey, never used Headspace, or had a bad experience may have been less likely to participate. In addition, the survey was cross-sectional, which limits our ability to track changes in outcomes over time. To address this limitation, follow-up surveys have been distributed to respondents to enable longitudinal analysis, which will be the focus of a future study.

Finally, the demographic survey items were designed according to project requirements: as required by California Code of Regulations (Cal. Code Regs. Tit. 9, § 3580.010), projects categorized and funded as innovative projects in California must adhere to specific data collection and reporting guidelines [73]. This restriction limited certain types of analyses. Age, for example, was asked as a categorical rather than continuous variable, which limits our ability to make conclusions on any correlations between age and use of Headspace. Some of the other measures may not conform to typical research measures; for example, validated stigma item scales were shortened based on stakeholder feedback to mitigate response burden and remove survey items that did not resonate with the community. However, our community-based participatory approach to survey development incorporated and centered the views of partners from the sites as well as those with lived-experience. Collecting data in real-world deployments often requires balancing the views and priorities of multiple invested parties.

Conclusions

To our knowledge, this is the first paper reporting the characteristics of app use with a large sample of potential and naturalistic users with the mindfulness app Headspace. Results showed that people who were still using Headspace were more likely to experience mental health challenges and distress and had made more use of other digital mental health resources (ie, online tools and connecting with people online) than people who were not using Headspace. In addition, current users of Headspace used the app more frequently than people who abandoned the app and rated the app higher on usefulness. The most common reasons for not using Headspace were that people were already using other strategies to support their mental health, could not find the time to use Headspace, or did not think Headspace would be useful. These results can be used to inform future outreach strategies and implementations of future digital mental health programs.

Acknowledgments

The authors would like to acknowledge the participating sites for their invaluable contributions to survey development and community members Amanda Kirk, Vanessa Hamill-Meeriyakerd, and Mimi Mier-Rosales for creating artwork to include in recruitment emails. This work was funded by the Help@Hand Project (agreement 417-ITS-UCI-2019), a project overseen by the California Mental Health Service Authority (CalMHSA). CalMHSA reviewed the manuscript to ensure its confidentiality. The information or content and conclusions presented here are those of the authors and should not be construed as the official position or policy of nor should any endorsements be inferred by the participating Help@Hand Counties or CalMHSA.

Data Availability

The survey instrument is included in Multimedia Appendix 1. The data collected and analyzed during the study are not publicly available as it was not a funding requirement, and study participants did not consent for their data to be shared publicly.

Conflicts of Interest

SMS has received consulting payments from Boehringer Ingelheim and Otsuka Pharmaceuticals and is a member of the Headspace Scientific Advisory Board, for which they receive compensation. SMS is an associate editor for *JMIR Meantal Health* at the time of this publication.

XSL•FO
RenderX

Multimedia Appendix 1
The survey instrument.
[DOCX File , 50 KB - formative_v9i1e73457_app1.docx]

Multimedia Appendix 2
Checklist for Reporting Results of Internet e-Surveys.
[DOCX File , 25 KB - formative_v9i1e73457_app2.docx]

Multimedia Appendix 3
The survey measures as defined in the analyses.
[DOCX File , 24 KB - formative_v9i1e73457_app3.docx]

Multimedia Appendix 4
Logistic regression tables showing odds ratios and 95% CIs for all measures.
[DOCX File , 24 KB - formative_v9i1e73457_app4.docx]

References

1. Mental health is a global public health issue. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/pcd/collections/Mental_Health_Is_a_Global_Public_Health_Issue.htm [accessed 2025-07-20]
2. Lopes L, Kirzinger A, Sparks G, Stokes M, Brodie M. KFF/CNN mental health in America survey. KFF. 2022 Oct 5. URL: https://www.kff.org/report-section/kff-cnn-mental-health-in-america-survey-findings/ [accessed 2025-07-20]
3. Torous J, Jän Myrick K, Rauseo-Ricupero N, Firth J. Digital mental health and COVID-19: using technology today to accelerate the curve on access and quality tomorrow. JMIR Ment Health 2020 Mar 26;7(3):e18848 [FREE Full text] [doi: 10.2196/18848] [Medline: 32213476]
4. Wasil AR, Gillespie S, Shingleton R, Wilks CR, Weisz JR. Examining the reach of smartphone apps for depression and anxiety. Am J Psychiatry 2020 May 01;177(5):464-465. [doi: 10.1176/appi.ajp.2019.19090905] [Medline: 32354266]
5. Camacho E, Cohen A, Torous J. Assessment of mental health services available through smartphone apps. JAMA Netw Open 2022 Dec 01;5(12):e2248784 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.48784] [Medline: 36576737]
6. Plaza I, Demarzo MM, Herrera-Mercadal P, García-Campayo J. Mindfulness-based mobile applications: literature review and analysis of current features. JMIR Mhealth Uhealth 2013 Nov 01;1(2):e24 [FREE Full text] [doi: 10.2196/mhealth.2733] [Medline: 25099314]
7. Gál É, tefan S, Cristea IA. The efficacy of mindfulness meditation apps in enhancing users' well-being and mental health related outcomes: a meta-analysis of randomized controlled trials. J Affect Disord 2021 Jan 15;279:131-142. [doi: 10.1016/j.jad.2020.09.134] [Medline: 33049431]
8. Creswell JD. Mindfulness interventions. Annu Rev Psychol 2017 Jan 03;68(1):491-516. [doi: 10.1146/annurev-psych-042716-051139] [Medline: 27687118]
9. Flett JA, Hayne H, Riordan BC, Thompson LM, Conner TS. Mobile mindfulness meditation: a randomised controlled trial of the effect of two popular apps on mental health. Mindfulness 2018 Oct 31;10(5):863-876. [doi: 10.1007/s12671-018-1050-9]
10. Kabat-Zinn J, Massion AO, Kristeller J, Peterson LG, Fletcher KE, Pbert L, et al. Effectiveness of a meditation-based stress reduction program in the treatment of anxiety disorders. Am J Psychiatry 1992 Jul;149(7):936-943. [doi: 10.1176/ajp.149.7.936] [Medline: 1609875]
11. Shapiro SL, Schwartz GE, Bonner G. Effects of mindfulness-based stress reduction on medical and premedical students. J Behav Med 1998 Dec;21(6):581-599. [doi: 10.1023/a:1018700829825] [Medline: 9891256]
12. Speca M, Carlson LE, Goodey E, Angen M. A randomized, wait-list controlled clinical trial: the effect of a mindfulness meditation-based stress reduction program on mood and symptoms of stress in cancer outpatients. Psychosom Med 2000;62(5):613-622. [doi: 10.1097/00006842-200009000-00004] [Medline: 11020090]
13. Baer RA, Smith GT, Hopkins J, Krietemeyer J, Toney L. Using self-report assessment methods to explore facets of mindfulness. Assessment 2006 Mar 01;13(1):27-45. [doi: 10.1177/1073191105283504] [Medline: 16443717]
14. Calm schools. Calm. URL: https://www.calm.com/schools [accessed 2025-07-20]
15. Stephen B. Headspace is now free for health care professionals because of the pandemic. The Verge. 2020 Mar 17. URL: https://www.theverge.com/2020/3/16/21181773/headspace-free-health-care-provider-public-health [accessed 2025-07-20]
16. Coping with COVID-19: free mental health apps. One Mind Psyberguide. URL: https://onemindpsyberguide.org/guide/resources/coping-with-covid-19/ [accessed 2025-07-20]
17. Himle JA, Weaver A, Zhang A, Xiang X. Digital mental health interventions for depression. Cogn Behav Pract 2022 Feb;29(1):50-59. [doi: 10.1016/j.cbpra.2020.12.009]

XSL•FO
RenderX

18. Garrido S, Millington C, Cheers D, Boydell K, Schubert E, Meade T, et al. What works and what doesn't work? A systematic review of digital mental health interventions for depression and anxiety in young people. Front Psychiatry 2019 Nov 13;10:759 [FREE Full text] [doi: 10.3389/fpsyt.2019.00759] [Medline: 31798468]
19. Kozlov E, Bantum E, Pagano I, Walser R, Ramsey K, Taylor K, et al. The reach, use, and impact of a free mHealth mindfulness app in the general population: mobile data analysis. JMIR Ment Health 2020 Nov 27;7(11):e23377 [FREE Full text] [doi: 10.2196/23377] [Medline: 33245289]
20. Flett JA, Conner TS, Riordan BC, Patterson T, Hayne H. App-based mindfulness meditation for psychological distress and adjustment to college in incoming university students: a pragmatic, randomised, waitlist-controlled trial. Psychol Health 2020 Sep 12;35(9):1049-1074. [doi: 10.1080/08870446.2019.1711089] [Medline: 32046499]
21. Baumel A, Muench F, Edan S, Kane JM. Objective user engagement with mental health apps: systematic search and panel-based usage analysis. J Med Internet Res 2019 Sep 25;21(9):e14567 [FREE Full text] [doi: 10.2196/14567] [Medline: 31573916]
22. Lattie EG, Schueller SM, Sargent E, Stiles-Shields C, Tomasino KN, Corden ME, et al. Uptake and usage of IntelliCare: a publicly available suite of mental health and well-being apps. Internet Interv 2016 May;4(2):152-158 [FREE Full text] [doi: 10.1016/j.invent.2016.06.003] [Medline: 27398319]
23. Baumel A, Edan S, Kane JM. Is there a trial bias impacting user engagement with unguided e-mental health interventions? A systematic comparison of published reports and real-world usage of the same programs. Transl Behav Med 2019 Nov 25;9(6):1020-1033. [doi: 10.1093/tbm/ibz147] [Medline: 31689344]
24. About headspace. Headspace. URL: https://www.headspace.com/about-us [accessed 2025-07-21]
25. Perez S. Meditation and mindfulness apps continue their surge amid pandemic. TechCrunch. 2020 May 28. URL: https://techcrunch.com/2020/05/28/meditation-and-mindfulness-apps-continue-their-surge-amid-pandemic/ [accessed 2025-07-21]
26. Sorkin DH, Mukamel D, Eikey E, Mark G, Schueller SM, Schneider M, et al. Conceptualizing and measuring mental health stigma. Program of Research in Translational Technology Enabling High Quality Care, University of California. 2020. URL: https://bpb-us-e2.wpmucdn.com/sites.uci.edu/dist/b/5702/files/2024/09/Conceptualizing-and-Measuring-Mental-Health-Stigma-Report.pdf [accessed 2025-07-21]
27. Eysenbach G. Improving the quality of web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34 [FREE Full text] [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
28. Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O'Neal L, et al. The REDCap consortium: building an international community of software platform partners. J Biomed Inform 2019 Jul;95:103208 [FREE Full text] [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]
29. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381 [FREE Full text] [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
30. Ng MM, Firth J, Minen M, Torous J. User engagement in mental health apps: a review of measurement, reporting, and validity. Psychiatr Serv 2019 Jul 01;70(7):538-544 [FREE Full text] [doi: 10.1176/appi.ps.201800519] [Medline: 30914003]
31. Khanna G. What are the best mental health apps in 2025? Appwrk IT Solutions. 2025 Jan 21. URL: https://appwrk.com/best-mental-health-apps [accessed 2025-08-05]
32. Kaveladze BT, Wasil AR, Bunyi JB, Ramirez V, Schueller SM. User experience, engagement, and popularity in mental health apps: secondary analysis of app analytics and expert app reviews. JMIR Hum Factors 2022 Jan 31;9(1):e30766 [FREE Full text] [doi: 10.2196/30766] [Medline: 35099398]
33. Kessler RC, Andrews G, Colpe LJ, Hiripi E, Mroczek DK, Normand SL, et al. Short screening scales to monitor population prevalences and trends in non-specific psychological distress. Psychol Med 2002 Aug;32(6):959-976. [doi: 10.1017/s0033291702006074] [Medline: 12214795]
34. Nunnally JC, Bernstein IH. Psychometric Theory. New York, NY: McGraw-Hill Companies; 1994.
35. Andrews G, Slade T. Interpreting scores on the Kessler Psychological Distress Scale (K10). Aust N Z J Public Health 2001 Dec;25(6):494-497 [FREE Full text] [doi: 10.1111/j.1467-842x.2001.tb00310.x] [Medline: 11824981]
36. Brugnera A, Zarbo C, Scalabrini A, Compare A, Mucci C, Carrara S, et al. Attachment anxiety, reflective functioning and well-being as predictors of burn-out and psychological distress among psychotherapists: a longitudinal study. Clin Psychol Psychother 2023 Jan 11;30(3):587-598 [FREE Full text] [doi: 10.1002/cpp.2823] [Medline: 36610037]
37. Heerde JA, Merrin GJ, Le VT, Toumbourou JW, Bailey JA. Health of young adults experiencing social marginalization and vulnerability: a cross-national longitudinal study. Int J Environ Res Public Health 2023 Jan 17;20(3):1711 [FREE Full text] [doi: 10.3390/ijerph20031711] [Medline: 36767076]
38. Hughes ME, Waite LJ, Hawkley LC, Cacioppo JT. A short scale for measuring loneliness in large surveys: results from two population-based studies. Res Aging 2004;26(6):655-672 [FREE Full text] [doi: 10.1177/0164027504268574] [Medline: 18504506]
39. Russell D, Peplau LA, Ferguson ML. Developing a measure of loneliness. J Pers Assess 1978 Jun;42(3):290-294. [doi: 10.1207/s15327752jpa4203_11] [Medline: 660402]

XSL•FO
RenderX

40. Steptoe A, Shankar A, Demakakos P, Wardle J. Social isolation, loneliness, and all-cause mortality in older men and women. Proc Natl Acad Sci U S A 2013 Apr 09;110(15):5797-5801 [FREE Full text] [doi: 10.1073/pnas.1219686110] [Medline: 23530191]
41. Ritsher JB, Otilingam PG, Grajales M. Internalized stigma of mental illness: psychometric properties of a new measure. Psychiatry Res 2003 Nov 01;121(1):31-49. [doi: 10.1016/j.psychres.2003.08.008] [Medline: 14572622]
42. Griffiths KM, Christensen H, Jorm AF, Evans K, Groves C. Effect of web-based depression literacy and cognitive-behavioural therapy interventions on stigmatising attitudes to depression: randomised controlled trial. Br J Psychiatry 2004 Oct;185:342-349. [doi: 10.1192/bjp.185.4.342] [Medline: 15458995]
43. Corrigan PW, Giffort D, Rashid F, Leary M, Okeke I. Recovery as a psychological construct. Community Ment Health J 1999 Jun;35(3):231-239. [doi: 10.1023/a:1018741302682] [Medline: 10401893]
44. CHIS 2020 adult public use files. UCLA Center for Health Policy Research. URL: https://healthpolicy.ucla.edu/our-work/public-use-files [accessed 2025-07-21]
45. Venkatesh V, Thong JY, Xu X. Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology. MIS Q 2012;36(1):157-178. [doi: 10.2307/41410412]
46. Odewumi MO, Yusuf OO, Oputa GO. UTAUT Model: intention to use social media for learning interactive effect of postgraduate gender in South-West Nigeria. Int J Educ Dev Using Inf Commun Technol 2018;14(3):239-251 [FREE Full text]
47. Borghouts J, Eikey E, Mark G, De Leon C, Schueller SM, Schneider M, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. J Med Internet Res 2021 Mar 24;23(3):e24387 [FREE Full text] [doi: 10.2196/24387] [Medline: 33759801]
48. O'Connor M, Casey L. The Mental Health Literacy Scale (MHLS): a new scale-based measure of mental health literacy. Psychiatry Res 2015 Sep 30;229(1-2):511-516. [doi: 10.1016/j.psychres.2015.05.064] [Medline: 26228163]
49. QuickFacts. United States Census Bureau. URL: https://www.census.gov/quickfacts/ [accessed 2025-07-21]
50. R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing. 2022. URL: https://www.r-project.org/ [accessed 2025-07-21]
51. Huberty J, Vranceanu AM, Carney C, Breus M, Gordon M, Puzia ME. Characteristics and usage patterns among 12,151 paid subscribers of the calm meditation app: cross-sectional survey. JMIR Mhealth Uhealth 2019 Nov 03;7(11):e15648 [FREE Full text] [doi: 10.2196/15648] [Medline: 31682582]
52. Militello L, Sobolev M, Okeke F, Adler DA, Nahum-Shani I. Digital prompts to increase engagement with the headspace app and for stress regulation among parents: feasibility study. JMIR Form Res 2022 Mar 21;6(3):e30606 [FREE Full text] [doi: 10.2196/30606] [Medline: 35311675]
53. Crisp DA, Griffiths KM. Participating in online mental health interventions: who is most likely to sign up and why? Depress Res Treat 2014;2014:790457 [FREE Full text] [doi: 10.1155/2014/790457] [Medline: 24804089]
54. Upchurch DM, Johnson PJ. Gender differences in prevalence, patterns, purposes, and perceived benefits of meditation practices in the United States. J Womens Health (Larchmt) 2019 Feb;28(2):135-142 [FREE Full text] [doi: 10.1089/jwh.2018.7178] [Medline: 30543475]
55. Borghouts J, Eikey EV, De Leon C, Schueller SM, Schneider M, Stadnick NA, et al. Understanding the role of support in digital mental health programs with older adults: users' perspective and mixed methods study. JMIR Form Res 2022 Dec 13;6(12):e43192 [FREE Full text] [doi: 10.2196/43192] [Medline: 36512387]
56. Kannisto KA, Korhonen J, Adams CE, Koivunen MH, Vahlberg T, Välimäki MA. Factors associated with dropout during recruitment and follow-up periods of a mHealth-based randomized controlled trial for Mobile.Net to encourage treatment adherence for people with serious mental health problems. J Med Internet Res 2017 Feb 21;19(2):e46 [FREE Full text] [doi: 10.2196/jmir.6417] [Medline: 28223262]
57. Lattie EG, Stiles-Shields C, Graham AK. An overview of and recommendations for more accessible digital mental health services. Nat Rev Psychol 2022 Feb 26;1(2):87-100 [FREE Full text] [doi: 10.1038/s44159-021-00003-1] [Medline: 38515434]
58. Perret S, Alon N, Carpenter-Song E, Myrick K, Thompson K, Li S, et al. Standardising the role of a digital navigator in behavioural health: a systematic review. Lancet Digit Health 2023 Dec;5(12):e925-e932. [doi: 10.1016/s2589-7500(23)00152-8]
59. Jaso-Yim B, Eyllon M, Sah P, Pennine M, Welch G, Schuler K, et al. Evaluation of the impact of a digital care navigator on increasing patient registration with digital mental health interventions in routine care. Internet Interv 2024 Dec;38:100777 [FREE Full text] [doi: 10.1016/j.invent.2024.100777] [Medline: 39410952]
60. Pruitt LD, Vuletic S, Smolenski DJ, Wagner A, Luxton DD, Gahm GA. Predicting post treatment client satisfaction between behavioural activation for depression delivered either in-person or via home-based telehealth. J Telemed Telecare 2019 Sep;25(8):460-467. [doi: 10.1177/1357633X18784103] [Medline: 29976097]
61. Toscos T, Coupe A, Flanagan M, Drouin M, Carpenter M, Reining L, et al. Teens using screens for help: impact of suicidal ideation, anxiety, and depression levels on youth preferences for telemental health resources. JMIR Ment Health 2019 Jun 21;6(6):e13230 [FREE Full text] [doi: 10.2196/13230] [Medline: 31228179]

XSL•FO
RenderX

62. Laurie J, Blandford A. Making time for mindfulness. Int J Med Inform 2016 Dec;96:38-50. [doi: 10.1016/j.ijmedinf.2016.02.010] [Medline: 26965526]
63. Perez S. Nearly 1 in 4 people abandon mobile apps after only one use. TechCrunch. 2016 May 31. URL: https://techcrunch.com/2016/05/31/nearly-1-in-4-people-abandon-mobile-apps-after-only-one-use/ [accessed 2025-07-21]
64. Burgess ER, Zhang R, Ernala SK, Feuston JL, De Choudhury M, Czerwinski M, et al. Technology ecosystems: rethinking resources for mental health. Interactions 2020 Dec 23;28(1):66-71. [doi: 10.1145/3434564]
65. Graham AK, Greene CJ, Kwasny MJ, Kaiser SM, Lieponis P, Powell T, et al. Coached mobile app platform for the treatment of depression and anxiety among primary care patients: a randomized clinical trial. JAMA Psychiatry 2020 Sep 01;77(9):906-914 [FREE Full text] [doi: 10.1001/jamapsychiatry.2020.1011] [Medline: 32432695]
66. Allan S, Bradstreet S, Mcleod H, Farhall J, Lambrou M, Gleeson J, EMPOWER Group, et al. Developing a hypothetical implementation framework of expectations for monitoring early signs of psychosis relapse using a mobile app: qualitative study. J Med Internet Res 2019 Oct 24;21(10):e14366 [FREE Full text] [doi: 10.2196/14366] [Medline: 31651400]
67. Berry N, Lobban F, Bucci S. A qualitative exploration of service user views about using digital health interventions for self-management in severe mental health problems. BMC Psychiatry 2019 Jan 21;19(1):35. [doi: 10.1186/s12888-018-1979-1] [Medline: 30665384]
68. Gunn J, Cameron J, Densley K, Davidson S, Fletcher S, Palmer V, et al. Uptake of mental health websites in primary care: insights from an Australian longitudinal cohort study of depression. Patient Educ Couns 2018 Jan;101(1):105-112. [doi: 10.1016/j.pec.2017.07.008] [Medline: 28739180]
69. Hernandez-Ramos R, Schueller SM, Borghouts J, Palomares K, Eikey E, Schneider M, et al. Evaluation of a pilot implementation of a digital cognitive behavioral therapy platform for isolated older adults in county mental health services. Implement Res Pract 2024 Oct 15;5:26334895241288571 [FREE Full text] [doi: 10.1177/26334895241288571] [Medline: 39431167]
70. Smail-Crevier R, Powers G, Noel C, Wang J. Health-related internet usage and design feature preference for e-mental health programs among men and women. J Med Internet Res 2019 Mar 18;21(3):e11224 [FREE Full text] [doi: 10.2196/11224] [Medline: 30882361]
71. Stevens R, Polk K, Merrill C, Feng F, Weiss M, Brosnan E. User experience design to enhance the effectiveness of mobile technologies for the treatment of mental health. In: Proceedings of the Systems and Information Engineering Design Symposium. 2018 Presented at: SIEDS 2018; April 27, 2018; Charlottesville, VA. [doi: 10.1109/sieds.2018.8374724]
72. Anderson C, Seff LR, Batra A, Bhatt C, Palmer RC. Recruiting and engaging older men in evidence-based health promotion programs: perspectives on barriers and strategies. J Aging Res 2016;2016:8981435 [FREE Full text] [doi: 10.1155/2016/8981435] [Medline: 27366330]
73. Cal. Code Regs. Tit. 9, § 3580.010 - Annual Innovative Project Report. Cornell Law School. URL: https://www.law.cornell.edu/regulations/california/9-CCR-3580.010 [accessed 2025-07-21]

Abbreviations

DMHI: digital mental health intervention
OR: odds ratio
REDCap: Research Electronic Data Capture

Edited by A Mavragani; submitted 04.03.25; peer-reviewed by E Bantum, P Matthews; comments to author 26.03.25; revised version received 13.04.25; accepted 27.06.25; published 22.08.25.

Please cite as:
Borghouts J, Eikey EV, De Leon C, Schueller SM, Schneider M, Stadnick NA, Zheng K, Mukamel DB, Sorkin DH
Characteristics Associated With the Use of the Mindfulness Meditation App Headspace in a Large Public Health Deployment: Cross-Sectional Survey Study
JMIR Form Res 2025;9:e73457
URL: https://formative.jmir.org/2025/1/e73457
doi:10.2196/73457
PMID:40844821

©Judith Borghouts, Elizabeth V Eikey, Cinthia De Leon, Stephen M Schueller, Margaret Schneider, Nicole A Stadnick, Kai Zheng, Dana B Mukamel, Dara H Sorkin. Originally published in JMIR Formative Research (https://formative.jmir.org), 22.08.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium,

XSL•FO
RenderX

provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Enhancing Engagement, Practice Integration, and Skill Learning in Mobile Technology–Delivered Interventions Using Human Support: Randomized Controlled Trial With Depressed College Students

Colleen S Conley[1], PhD; Brynn M DeLorenzo[1,2], PhD; Carol H Gonzales[1,3], PhD; Ian J Kahrilas[1,4], PhD; Jennifer Duffecy[5], PhD; Rebecca L Silton[1], PhD

[1]Psychology Department, Loyola University Chicago, Chicago, IL, United States
[2]Department of Psychiatry, Massachusetts General Hospital, Boston, MA, United States
[3]Saint Louis Children's Hospital, St. Louis, MO, United States
[4]Kythera Labs, Franklin, TN, United States
[5]Mood and Anxiety Disorders Program, Department of Psychiatry, University of Illinois Chicago, Chicago, IL, United States

Corresponding Author:
Colleen S Conley, PhD
Psychology Department
Loyola University Chicago
1032 W Sheridan Rd
Chicago, IL, 60660-1537
United States
Phone: 1 7735083603
Email: cconley@luc.edu

Abstract

Background: Among evidence-based mobile technology–delivered interventions (mTDIs), mindfulness apps such as Headspace have demonstrated numerous benefits. These benefits are particularly important for college students, who continue to face high rates of depression and psychological distress that are paired with insufficient mental health services to meet these needs. mTDIs offer scalable solutions to ameliorating mental health symptoms and may be able to help address this gap in limited access to mental health services for all populations. While mTDIs have great promise for maximizing reach, their utility can be hamstrung by low rates of user engagement and uptake. Thus, this study implemented 2 human support enhancements designed to boost user in-app engagement, practice integration into daily life (ie, sustainability), and app-related skill learning (ie, perceived benefits) in a sample of college students with depression who were granted full access to an mTDI (Headspace).

Objective: This randomized controlled trial evaluated the impact of two human support enhancements—(1) a one-time face-to-face orientation with or without (2) placement in a peer supportive accountability group—on self-reported and objectively captured mTDI engagement, practice integration, and skill learning among a sample of college students with depression.

Methods: Participants (n=123) authorized access to their recorded app use data, provided by Headspace. In addition, at the midpoint (1 mo), postintervention (2 mo), and follow-up (3 mo) assessments, participants self-reported on the extent to which they had used the app, how likely they were to continue using the app and related skills in the future, and the extent to which they learned skills and practiced these skills in their daily lives.

Results: Compared to participants who were simply given access to the app (37/123, 30.1%) without these enhancements, those who attended the orientation (86/123, 69.9%), regardless of additional random allocation to the peer supportive accountability group (48/123, 39%), demonstrated significantly greater mTDI engagement (ie, more minutes meditated [$F_{2,117}$=11.20; P<.001] and more sessions overall [$F_{2,117}$=15.00; P<.001]) and rated more favorably multiple aspects of practice integration (ie, more everyday mindfulness practice [$F_{2,72}$=6.20; P=.003] and greater likelihood of future mindfulness [$F_{2,71}$=7.42; P<.001]) and skill learning (ie, learning about mindfulness [$F_{2,73}$=6.02; P=.004], learning mindfulness skills [$F_{2,72}$=11.01; P<.001], and an increased awareness of thoughts and feelings [$F_{2,73}$=6.05; P=.004]), indicating potential implications for amplifying the benefits of mTDIs through increased user engagement.

XSL•FO
RenderX

Conclusions: The results of this study illustrate that an initial face-to-face orientation boosts mTDI engagement, enhances the integration of intervention skills into everyday life, and increases learning. Future work is needed to determine the active ingredients of the orientation as well as to narrow in on the optimal implementation of supportive accountability that might drive increased levels of engagement and the associated positive intervention benefits.

Trial Registration: Open Science Foundation (OSF) 3trzk; https://osf.io/3trzk

(JMIR Form Res 2025;9:e56963) doi:10.2196/56963

KEYWORDS

mobile app; digital intervention; human support; engagement; skill integration; skill learning; mindfulness; depression

Introduction

Background

Evidence-based technology-delivered interventions (TDIs), including mobile apps that promote wellness and mental health (mobile TDIs; mTDIs), have demonstrated great promise for improving emotional well-being and reducing psychological distress across various population samples, including college students [1,2]. Mindfulness apps in particular have been effective in reducing depression, anxiety, and stress and improving psychological well-being and life satisfaction [3]. Importantly, mindfulness mTDIs also promote the development of related skills, with participants reporting significant improvements in areas such as awareness, acceptance, nonjudgmental attitude, and focus on the present moment [4,5]. Thus, these tools offer an effective and accessible way to address the large treatment gap that exists between individuals who need mental health treatment and those who are receiving it, a discrepancy that is especially pronounced for college students with depression [6].

Beyond their clinical utility, mTDIs have broad appeal, especially among college students, who seem to prefer the convenience, immediacy, and perceived confidentiality [7] paired with the added benefit of 24-hour access, including availability after in-person clinic hours [1]. Although there are hundreds of mTDIs designed to address mental health problems, and these numbers are growing rapidly, many of these apps are not evidence-based [8,9]. Furthermore, clinical science grapples with translating the benefits often seen when evaluating mTDIs through research trials [10] into best practices for promoting real-world applications, often marked by low uptake and engagement [11]. Interventions might work well for a subset of people who engage with them but not for those who fall short of engagement or adherence standards [4]. Indeed, users cannot benefit from the behavioral changes facilitated by mTDIs if they do not engage with such programs and learn the necessary skills. To harness the potential benefits of mTDIs, it is an urgent priority to identify strategies and supports for fostering sustained engagement and integration into daily life so that users can ultimately learn and implement skills for well-being.

Although mTDIs are typically designed to be self-guided, they can be integrated with human support features to enhance user engagement and thus benefits. This study is part of a larger randomized controlled trial (RCT) that established the efficacy—specifically psychosocial benefits—of a mindfulness-based mTDI for college students experiencing depressive symptoms. Building on these findings, this study examines whether 2 human support enhancements—a face-to-face orientation and placement in small peer supportive accountability groups—(1) boosted *engagement* with the mindfulness-based mTDI, (2) cultivated increased *practice integration* into daily life, and (3) enhanced users' self-reported *skill learning*.

Engagement With Mobile Interventions: Challenges and Opportunities

Although mTDIs have great potential to deliver effective treatments to a broad population, such programs are marked by low rates of both uptake and sustained user engagement, particularly for self-guided treatments that involve lower levels of structure and prescriptive guidance [12,13]. Such digital approaches require users to be self-motivated to initiate and sustain their use independently [14]. Furthermore, the mental health symptoms driving users to seek out mTDIs, such as depressive symptoms (eg, reduced interest, energy, and concentration), may themselves interfere with mTDI engagement [15]. Unpacking patterns of user engagement along with studying intervention design strategies to boost and maintain engagement are critical steps for advancing the efficacy of evidence-based mTDI interventions.

Engagement has been defined as "a state of energy investment involving physical, affective, and cognitive energies directed toward a focal stimulus or task" [16,17]. For the purposes of this study, engagement is measured behaviorally via the number of use minutes and completed sessions recorded by the mTDI software. While there is ongoing debate regarding how to best define and measure engagement [18], tabulating app use data (including duration and frequency) remains one of the most common methods [19]. User data from mindfulness-based mTDIs generally show engagement rates that start high but gradually reduce over time, with precipitous drops within the first week of uptake [11]. Of concern, engagement is particularly low during self-guided follow-up study periods [20,21]; for example, Flett et al [4] found that college students used a mindfulness-based mTDI almost daily in the primary 10 days of the study, but less than half had *any* use in the subsequent "30-day extended use period". This sentiment is reflected in conversations with college student app users as well: despite almost 75% reporting some benefit from mTDIs, the same proportion engaged with them weekly or less frequently [7]. Such research highlights the discrepancy between interest in, and engagement with, mTDIs.

XSL•FO
RenderX

While emerging research in the area of digital mental health frequently recommends the integration of human support to enhance mTDI engagement, there remain many questions related to what type of support is needed; for example, there is great variability in the provider (eg, peers, unlicensed professionals, and licensed clinicians), the modality (eg, in-person contact, video visits, telephone calls, and text-based messaging), and the frequency of support provided. Further complicating this issue is that few studies provide a detailed description of the human support that was delivered or examine the role of such support in their study outcomes, making it difficult to draw larger conclusions across studies [22]. This study aims to address this gap by specifically examining the impact of 2 human support features—participation in a preintervention orientation session and placement in a peer supportive accountability group—on mTDI engagement, practice integration beyond the app, and skill learning.

Peer Supportive Accountability

Supportive accountability is a theoretical model that seeks to explain how human support can improve engagement with mTDIs [14]. The model posits that a user is likely to engage more with an mTDI when they are accountable to another person. Accountability includes knowing that actions or inactions with the intervention will be observed, and behavioral choices will have to be justified to someone. Ideally, the person a user is being accountable to will be viewed as legitimate and with beneficial knowledge to offer, and therefore the accountability process is viewed as supportive and noncoercive. A systematic review of 208 studies examining user engagement with mTDIs did not explicitly test the supportive accountability model, but it found that guided or supported interventions had higher engagement than self-guided interventions and that being able to connect with others through the mTDI increased engagement [15].

The supportive accountability theory has been tested in a number of formats, including telephone coaching from clinicians and peers as well as through digital formats [14,23]. Coaching delivered by peers is as effective at improving engagement as coaching delivered by licensed clinicians [24]. Peer-delivered support typically has been offered through embedded message boards that include components of visible goal setting and completion as well as methods to provide support and communication to increase goal completion (eg, “likes” and comments). Several studies across different populations have demonstrated increased mTDI use by participants with access to peer support features compared to completely self-guided participants [25-27]. Even in trials with no differences in engagement between groups or no self-guided comparison group, participants tend to rate peer supportive accountability as positive and helpful [24,28-30]. A systematic review of 24 peer-supported digital mental health interventions concluded that using peer support to deliver or supplement mTDIs is feasible, acceptable, and effective at improving engagement and psychological functioning [31]. Ultimately, incorporating supportive accountability features into mTDIs for depression has the potential to improve both engagement and mental health outcomes for a wide array of users.

Preintervention Orientation

Research has less commonly explored the ways in which one-time support interactions, such as an orientation to a TDI, might enhance engagement. Psychotherapy research has explored the benefits of engagement sessions, typically using motivational interviewing (MI) tools, to enhance engagement with traditional mental health services [32,33]. More recently, MI has been incorporated into TDIs as a potential engagement strategy. Such techniques allow users to interact with programs in a more intentional manner by setting expectations and goals for program use, encouraging the use of problem-solving and other skills, and normalizing challenges [34]. More specifically, using MI to support TDI users as they prepare for, and enact, behavioral changes is particularly encouraged to support program engagement [34]. Brief motivational interventions, such as through SMS text messages or telephone calls, are viewed positively by TDI users, promote program engagement, and even enhance initial program outcomes [35-38]. Similarly, Linardon and Fuller-Tyszkiewicz [13] found increased engagement for trials that had at least 1 opportunity for contact with the researchers (either a telephone interview or an in-person interview before enrolling) compared to trials where participants could enroll in a study on the web without contacting a researcher.

Some research has targeted motivation before users engage with the program, such as through a synchronous (eg, face-to-face or telephone) orientation session with research staff or through a module within the TDI. While some studies seem to incorporate such techniques into their methodological design [29,39], few have provided details about the content of these sessions or explicitly investigated their impact on engagement or other outcomes. Encouragingly, researchers have begun to recognize that such techniques may serve as an active component of treatment that can be investigated in and of itself; for example, Bur et al [40] conducted an RCT of a web-based self-help program for depression, exploring the effects on user engagement of different types of support, including elements with human contact (personal guidance and a preintervention diagnostic interview) and those without human contact (automated reminders and a preintervention MI module). At the end of the intervention, participants who received human support in the form of individualized guidance demonstrated both greater reductions in depressive symptoms and greater treatment adherence compared to the other support conditions [40]. Beyond building motivation for engagement, these preprogram touchpoints can help users navigate the technology of a new app as well as the large amounts of content that many TDIs provide, both of which can be overwhelming, particularly when users are also contending with mental health challenges such as depression [8]. Overall, more research is needed to better understand the role that preprogram interventions (eg, introductory, orientation, or engagement sessions) may have and their effects on subsequent program engagement and associated benefits.

Self-Guided Practice for Lasting Benefits

As with skills-based face-to-face treatments, the goal of mTDIs is to help users learn and practice skills so that they become

XSL•FO
RenderX

integrated into their daily lives over time. Notably, the completion of out-of-session exercises is uniquely linked to symptom improvement and treatment outcome across disorders [41]. In a similar vein, Schlosser et al [42] distinguished between TDI active use (eg, tasks or skill practices completed) and passive use (eg, time on the app that was not spent meaningfully engaging with other users or task completion) and found that only active use was linked to symptom improvement. This signals the critical importance of assessing not only objective engagement counts (eg, TDI use minutes and completed sessions) but also broader elements such as practice integration and skill learning.

Study Aims and Hypotheses

In a sample of college students with elevated depression, we examined the benefits of two mTDI enhancements—(1) a face-to-face orientation with or without (2) placement in a peer supportive accountability group—on three sets of outcomes: (1) mTDI engagement, (2) practice integration, and (3) skill learning. For the first set of outcomes—mTDI engagement—we hypothesized that those who were randomly allocated to receive the enhancements would demonstrate greater mTDI engagement—in terms of (1) mindfulness minutes and (2) total sessions completed—compared to those who were simply given access to the app without either of these enhancements. We examined whether participants in the peer supportive accountability group experienced incremental benefits in engagement, beyond those of the orientation. In an exploratory fashion, we examined whether there were differences among the 3 intervention groups on 2 other sets of self-report outcomes: For the second set of outcomes, practice integration, we specifically examined (3) "everyday mindfulness" and the likelihood of engaging in future mindfulness practices, whether (4) using the app, (5) engaging in mindfulness exercises on one's own, or (6) generally being more mindful in daily life. For the third set of outcomes, skill learning, we specifically examined (7) learning about mindfulness as a concept and practice, (8) learning mindfulness skills, and (9) being aware of one's thoughts and feelings.

Methods

Ethical Considerations

This RCT (registered on OSF before the start of analyses) was approved by the Loyola University Chicago Institutional Review Board (project number 2330), and all participants provided informed consent. Minimal identifying data were collected, and identifiers were removed from the dataset for analysis. Confidential (ie, identifying) data were stored on a university-supported, password-protected network drive and in locked filing cabinets to which only the research team had access. Participants were compensated monetarily (up to US $70) or with participant pool course credits (up to 10 credits) for the completion of study sessions.

Recruitment

Over 8 successive semesters, we recruited undergraduate students from a midsized Midwestern university using a multipronged voluntary response sampling strategy, including the psychology participant pool; cross-campus listserve emails; and flyers seeking students who self-identified as "down, sad, or distressed." Recruitment occurred at the beginning of each semester between 2017 and 2021. Eligible participants endorsed clinically significant levels of depressive symptoms as indicated by a score of ≥10 on a brief web-based screening using the Patient Health Questionnaire-8 (PHQ-8), which omits the suicidality item [43,44]. Individuals were excluded if they were engaged in psychotherapy at the start of the study, had regular practice of mindfulness or consistent use of the Headspace app within the past 6 months, or reported that they were unwilling to join the peer supportive accountability group if randomly allocated to this condition. As the broader project included electroencephalography assessments, participants were also excluded if they had a history of neurological conditions or head trauma (eg, concussions or seizures). Our sample size (10-27 students each semester) was determined by the number of participants who met study criteria each semester, paired with a consideration of resource constraints (ie, time, space, and funding) [45].

Intervention Groups and Procedures

Overview

Participants were randomly allocated to 3 nested groups: app as usual (App), app+orientation (App+O), and app+orientation and peer supportive accountability (App+OPS). Specifically, all participants received the base intervention of access to the app (App), all App+O and App+OPS participants (but not App participants) attended an orientation, and only the App+OPS participants were invited to the peer supportive accountability features (as described in the next subsection).

An unequal allocation procedure was used for randomization to ensure that a consistent set of 6 participants was assigned to each peer supportive accountability group (totaling to 8 groups over 8 semesters). Specifically, we performed a manual blind drawing to randomly allocate 6 participants to each App+OPS group and to randomly and evenly split the remaining participants among the App+O and App groups. This procedure yielded a final allocation ratio of 1.26:1.00:0.97. Participants completed 4 web-based self-report surveys providing demographic information before the intervention (within 2 wk before code activation) and assessing their experiences with the intervention at midpoint (1 mo after code activation), after the intervention (2 mo after code activation), and at 1-month follow-up (3 mo after code activation).

Intervention: Mobile Mindfulness App (Headspace)

Intervention participants received a 3-month code to access all content on Headspace [46], a web-based application and mobile app that delivers brief, guided mindfulness exercises as well as some additional wellness-related content (eg, "soundscapes," "sleepcasts," and "focus music"). The program includes courses that typically include 10 to 30 sessions that follow a particular theme (eg, handling sadness and managing anxiety). Headspace also includes single mindfulness meditation sessions, some of which are intended to be used in a specific situation (eg, difficult conversations) or as guided mindful activities (eg, mindful eating and mindful walking). Users can customize the length

XSL•FO
RenderX

of most sessions, ranging from shorter (1-5 min) to longer (10-30 min) options. Although participants were free to access any of the Headspace content, they were encouraged to engage with mindfulness exercises, particularly the "Basics" courses teaching foundational skills on mindfulness and meditation and the mental health–related meditations, such as those focused on psychological distress (eg, sadness, stress, anxiety, and "SOS" sessions) and positive well-being (eg, happiness and self-esteem).

Intervention: Orientation

The App+O and App+OPS groups (but not the App group) attended a 90-minute group orientation session wherein the first 3 authors reviewed study procedures, briefly described the principles and benefits of mindfulness, oriented participants to app features and content, helped them to activate their accounts with access codes, and provided recommendations for mTDI engagement. Although participants were encouraged to use the app as frequently and consistently as possible (ie, daily) and to use its mental health–focused content, no specific requirements were given, and use was ultimately self-guided. As recommended for mTDI engagement [34], elements of MI [47] were interspersed throughout the orientation session; for example, participants were encouraged to identify specific goals for their mTDI engagement and mindfulness practice, to consider their expectations and motivations for engaging in the program, to predict barriers to app use and engage in related problem-solving, and to reflect on approaches that have and have not been successful for them in the past in terms of establishing a new habit. Broader techniques such as eliciting participant thoughts and ideas, providing reflections, and using open-ended questions were also incorporated. At this point in the orientation, participants in the App+O group were dismissed, and for the duration of the study, they did not have additional contact with other study participants or the research staff beyond communication related to research assessments and compensation.

Intervention: Peer Supportive Accountability

Participants randomly allocated to the App+OPS group attended the aforementioned group orientation and stayed for an additional 20 to 30 minutes to review the rationale and procedures for the peer supportive accountability features. Researchers provided an overview of supportive accountability as well as the benefits of social sharing and of giving and receiving support. In cohorts of 6, App+OPS participants added each other as Headspace "buddies," which allowed them to share their progress. As time allowed, participants also set individual and group goals for their mindfulness practice (eg, daily use of the app) and for engagement with the peer supportive accountability group (eg, weekly posting in the web-based forum).

Participants in the App+OPS condition joined a private, closed Facebook group that served as a web-based forum. Participants were encouraged to post in the forum about their own successes and struggles with mindfulness practice; pose questions to the group; share information; and provide motivation, accountability, and support to other group members. Three to 5 times per week, research staff posted Headspace "mindful moments" (eg, "Imagine a world in which we witness thoughts without becoming them and experience feelings without being overwhelmed by them"), along with prompts encouraging participants to share their experiences with mindfulness and the app (eg, "What do you imagine? How would things be different?"). Twice per week, research staff posted user statistics, at the individual and group level (ie, number of sessions and minutes completed, session content, and days of use), along with an email digest that included the same user statistics and quotes from the web-based forum, examples of research findings demonstrating the benefits of mindfulness, and links to the web-based group and to Headspace.

In the third semester (and cohort) of the study, a face-to-face component was added to the App+OPS condition based on participant feedback. Participants met for 3 in-person group sessions approximately every other week (ie, 2, 4, and 6 wk after the orientation) for 45 to 60 minutes. Sessions did not include mindfulness exercises but rather focused on peer supportive accountability, following a general structure wherein research staff gave a brief introduction and reminder of the purpose of the group sessions, including any goals that were set in the previous session. With minimal direction from research staff, participants openly discussed their experiences with mindfulness (eg, barriers and successes), checked in with one another about progress, and fostered a sense of connectedness. During each meeting, a slide with suggested discussion topics was displayed for the group. At the end, research staff highlighted themes from the discussion and prompted participants to consider setting group and individual goals.

Measures

Social Identities

During the preintervention assessment, participants responded to a series of questions about different aspects of social identities, including age, racial and ethnic identity, gender, and sexual identity.

Engagement With mTDI

With participant consent, Headspace provided user data for each session, including date, time of day, module (eg, "Basics"), session number within the module (eg, session 1), and duration in minutes. Researchers calculated each participant's total minutes of mindfulness practice (mindfulness minutes) and total sessions of any type (total sessions). All meditation sessions (the vast majority of content in the app) as well as mindful activities (eg, guided mindful walking) were considered mindfulness practice. Sessions of advice, music ("focus music" and "sleep music"), and ambient noise and sounds ("sleepcasts" and "soundscapes") were not included in the totals of mindfulness minutes; however, they were included in total sessions to capture participants' broader engagement with the app.

Practice Integration: Self-Reported Mindfulness Practices and Likely Future Mindfulness Practice

To assess participants' practice integration or the use of mindfulness practice outside of the app, at the midpoint, postintervention, and 1-month follow-up assessments,

XSL•FO
RenderX

participants were asked about (1) their self-reported frequency of "nonguided, everyday mindfulness" (ie, self-guided practice) since starting the intervention. Participants rated how often they had practiced this on a Likert scale from 1 ("none: not at all") to 5 ("a lot: daily practice"). Participants were also asked how likely they were, *after* the study, to (2) use Headspace (not considering cost), (3) perform mindfulness exercises on their own, and (4) be more mindful in their everyday life. Participants rated each of these items on a scale ranging from 1 ("not likely") to 5 ("extremely likely").

Skill Learning

Finally, at the midpoint, postintervention, and 1-month follow-up assessments, participants were asked about their perceptions of the intervention and its benefits, adapting some items from the mindfulness intervention social validity questionnaire [48]. Participants were asked to consider their overall experience with the program and report the extent to which they felt they had, through the program, learned about mindfulness, learned mindfulness skills, and gained awareness of their thoughts and feelings. Participants rated each item on a scale ranging from 1 ("not at all true") to 5 ("extremely true").

Data Analysis Plan

In examining group effects, we used an intent-to-treat strategy, including all participants who were randomly allocated to 1 of the 3 intervention groups and for whom we had relevant data. For example, of the 120 participants with valid engagement data, a small subset (10/120, 8.3%; all in the App group) did not engage with the app; following an intent-to-treat approach, these participants were included in analyses. Furthermore, all App+OPS participants with valid data were included in the analyses, regardless of their level of engagement in the web-based forum or attendance at group meetings. All statistical analyses were conducted using SPSS software (version 28.0; IBM Corp) [49]. We report observed effect sizes in the results. Missing data were handled via case-wise deletion for each specific analysis.

To examine overall differences among the 3 groups across 3 time points (midpoint, postintervention, and 1-mo follow-up), we conducted group×time ANOVAs and interpreted the group effects. For outcomes with an overall effect of group (human support condition), we examined pairwise contrasts to identify any differential impact of the orientation alone versus the orientation plus peer supportive accountability.

Accounting for the effect of running multiple tests on type I error, we calculated a Bonferroni correction by dividing .05 by the total number of individual ANOVAs and consider both the standard significance threshold ($P<.05$) as well as the Bonferroni-adjusted significance threshold ($P<.0055$) in evaluating statistical significance. We also report effect sizes [50]. Specifically, we examined partial eta-squared values using the standards proposed by Cohen [51] for small ($\eta_p^2 \geq 0.01$), medium ($\eta_p^2 \geq 0.06$), and large ($\eta_p^2 \geq 0.14$) effects. We evaluated the distribution (skewness) of outcome variables at each time point, and we also ran Kolmogorov-Smirnov tests for each variable (Multimedia Appendix 2).

Results

Participant Characteristics

Participant recruitment yielded 123 undergraduate students (age: mean 19.1, SD 1.5 y), with an average PHQ-8 screening score of 14.1 (SD 3.4; range 10-22), randomly allocated to use the app as follows: 30.1% (37/123) App participants, 30.9% (38/123) App+O participants, and 39% (48/123) App+OPS participants (refer to Figure 1 for the CONSORT [Consolidated Standards of Reporting Trials] flow diagram; the results are presented according to the CONSORT-EHEALTH [Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth] checklist [Multimedia Appendix 1]). Of note, of the 37 participants in the App group, 51% (19/37) first completed a waitlist period, and these participants did not differ from the other 49% (18/37) of App participants who activated their Headspace access code at the same time as the rest of their cohort or from App+O or App+OPS participants on baseline depression ($F_{3,120}=0.30$; $P=.83$). Table 1 presents baseline demographic characteristics for the full sample and for participants randomly allocated to the 3 study groups, reflecting a higher proportion of women attending college and endorsing depression, but demonstrating no baseline group differences (refer to footnote a below Table 1).

XSL•FO
RenderX

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) flow diagram. App: app as usual; App+O: app+orientation; App+OPS: app+orientation and peer supportive accountability.

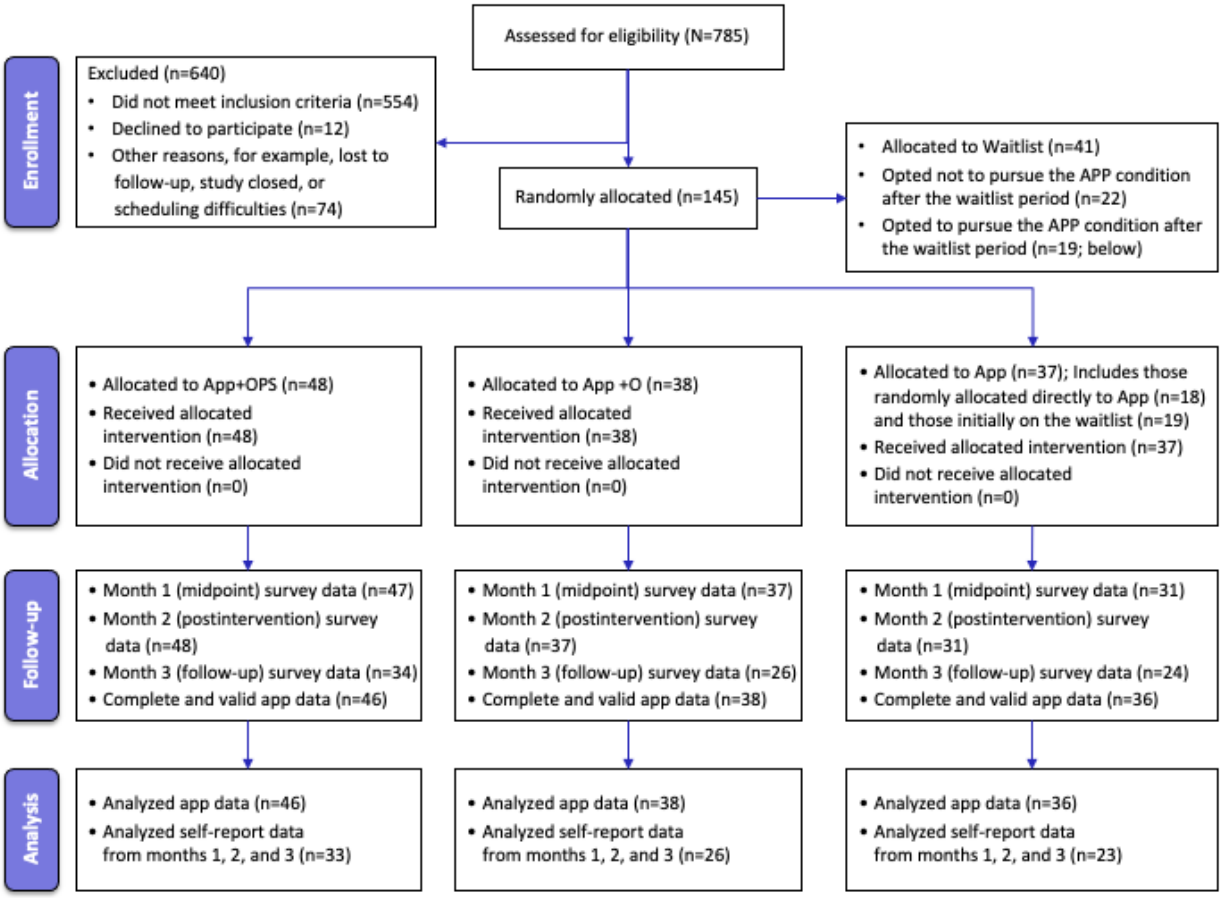

A few details shown in the CONSORT flow diagram are worth elaborating on. Three participants (n=1 in the App group and n=2 in the App+OPS group) did not have valid app-recorded engagement data due to data recording errors or access code malfunctions; therefore, they were excluded from analyses examining objective mTDI engagement. In addition, there were some missing data from surveys, as noted in the CONSORT flow diagram. Notably, participants who completed surveys at all 3 time points (82/123, 66.7%) did not differ from those who completed surveys at only 1 or 2 time points (41/123, 33.3%) in gender (χ^2_2=0.1; *P*=.96), sexual identity (χ^2_4=2.9; *P*=.58), racial and ethnic identity (χ^2_5=2.9; *P*=.71), or baseline depression scores (t_{120}=0.81; *P*=.42) but were more likely to be younger in age (t_{121}=2.77; *P*=.006) and a first-year student (χ^2_4=10.5; *P*=.03).

XSL•FO
RenderX

Table 1. Baseline demographic characteristics for the full sample and for participants randomly allocated to the 3 study groups[a].

	Total sample (n=123)	App+OPS[b] (n=48)	App+O[c] (n=38)	App[d] (n=37)
Age (y), mean (SD)	19.1 (1.5)	19.2 (1.3)	19.2 (1.8)	18.9 (1.1)
Gender, n (%)				
Woman	113 (91.9)	44 (1.7)	34 (89.5)	35 (4.6)
Man	7 (5.7)	4 (8.3)	1 (2.6)	2 (5.4)
Self-identified response (eg, nonbinary or transgender)	3 (2.4)	0 (0)	3 (7.9)	0 (0)
Sexual identity, n (%)				
Bisexual	23 (18.7)	5 (10.4)	9 (23.7)	9 (24.3)
Gay	4 (3.3)	3 (6.3)	1 (2.6)	0 (0)
Heterosexual	89 (72.4)	37 (77.1)	26 (68.4)	26 (70.3)
Lesbian	1 (0.8)	0 (0)	0 (0)	1 (2.7)
Self-identified response (eg, asexual, pansexual, or queer)	6 (4.9)	3 (6.3)	2 (5.3)	1 (2.7)
Racial and ethnic identity, n (%)[e]				
African American or Black	1 (0.8)	1 (2.1)	0 (0)	0 (0)
Asian or Asian American	19 (15.4)	5 (10.4)	6 (15.8)	8 (21.6)
Hispanic or Latiné	19 (15.4)	5 (10.4)	7 (18.4)	7 (18.9)
Native American or Pacific Islander	1 (0.8)	0 (0)	0 (0)	1 (2.7)
Non-Hispanic White	69 (56.1)	28 (58.3)	21 (55.3)	20 (54.1)
Multiracial or self-identified (eg, Middle Eastern)	14 (11.4)	9 (18.8)	4 (10.5)	1 (2.7)
Year in school, n (%)				
First-year undergraduate	69 (56.1)	26 (54.2)	20 (52.6)	23 (62.2)
Second-year undergraduate	30 (24.4)	13 (27.1)	12 (31.6)	5 (13.5)
Third-year undergraduate	12 (9.8)	4 (8.3)	2 (5.3)	6 (16.2)
Fourth-year undergraduate	10 (8.1)	3 (6.3)	4 (10.5)	3 (8.1)
Other (eg, accelerated nursing student)	2 (1.6)	2 (4.2)	0 (0)	0 (0)

[a]There were no significant differences across groups in age ($F_{2,120}$=0.35; P=.71) or other demographic data: gender (χ^2_4=8.0; P=.09), sexual identity (χ^2_8=8.6; P=.38), racial and ethnic identity (χ^2_{10}=11.7; P=.30), or year in school (χ^2_8=9.2; P=.33).

[b]App+OPS: app+orientation and peer supportive accountability.

[c]App+O: app+orientation.

[d]App: app as usual.

[e]Racial and ethnic identity categories were based on National Institutes of Health reporting guidelines from 2017, which reflected the categories identified by the US Office of Management and Budget standards for the classification of federal data on racial and ethnic identity.

Impact of Human Support Enhancements

The following subsections present the results for the impact of 2 human support enhancements (a face-to-face orientation with or without placement in a peer supportive accountability group) on the 3 categories of outcomes: (1) engagement with the mTDI (minutes and sessions), (2) practice integration (eg, everyday mindfulness practice beyond the mTDI), and (3) skill learning (eg, self-reported awareness, learning, and skills).

Engagement With mTDI

Table 2 presents group means and SDs and summarizes group effects (F, η_p^2, and P values) across the trial for each outcome. In addition, Figures 2-4 depict group means over time for each outcome. As shown in Table 2, there was an overall effect of group (human support condition) on cumulative mindfulness minutes ($F_{2,117}$=11.20, η_p^2=0.16; P=.001). Pairwise comparisons indicate that App participants had significantly lower mindfulness minutes than App+O (P=.005) and App+OPS (P<.001) participants, but mindfulness minutes did not significantly differ between App+O and App+OPS groups (P=.55; Figure 2A).

XSL•FO
RenderX

Table 2. Group means and SDs by time point—midpoint assessment (1 month after initiating app), postintervention assessment (2 months after initiating the app), and 1-month follow-up assessment (3 months after initiating the app, and 1 month after the intervention)—and group effects comparing participants randomly assigned to use a mindfulness app as usual (App; n=36), app+orientation (App+O; n=38), and app+orientation and peer supportive accountability (App+OPS; n=46).

Outcomes[a]	Group means (SDs)			ANOVA group effects		
	Midpoint	Postintervention	Follow-up	*F* test (*df*)	η_p^2	*P* value
Engagement with mobile technology–delivered intervention						
Cumulative mindfulness minutes				11.20 (2, 117)	0.161	*<.001*
App (n=36)[b]	55.0 (87.2)	85.4 (143.9)	98.6 (162.8)			
App+O (n=38)[c]	117.4 (97.6)	199.9 (181.0)	223.4 (212.3)			
App+OPS (n=46)[c]	147.8 (80.2)	251.7 (132.6)	260.5 (138.2)			
Cumulative any and all sessions				15.00 (2, 117)	0.204	*<.001*
App (n=36)[b]	7.22 (9.58)	10.83 (15.35)	12.67 (17.86)			
App+O (n=38)[c]	15.76 (10.11)	24.97 (17.31)	27.84 (20.95)			
App+OPS (n=46)[c]	19.28 (9.48)	31.46 (16.28)	32.91 (17.43)			
Practice integration (sustainability)						
Everyday mindfulness				6.20 (2, 72)	0.147	*.003*
App (n=21)[b]	1.48 (1.08)	1.33 (0.73)	1.95 (1.50)			
App+O (n=24)[c]	2.88 (1.42)	2.42 (1.44)	2.42 (1.59)			
App+OPS (n=30)[c]	2.33 (1.37)	2.87 (1.59)	2.73 (1.70)			
Future app (Headspace) use				0.11 (2, 72)	0.003	.89
App (n=21)	3.33 (1.24)	3.19 (1.29)	2.67 (1.32)			
App+O (n=23)	3.43 (1.20)	3.00 (1.41)	2.78 (1.38)			
App+OPS (n=31)	3.68 (0.98)	3.32 (1.14)	2.55 (1.09)			
Future mindful practice on own				3.10 (2, 73)	0.078	.05
App (n=21)[b]	2.81 (1.33)	2.86 (1.35)	3.00 (1.34)			
App+O (n=24)[c]	3.71 (1.08)	3.75 (1.15)	3.50 (1.29)			
App+OPS (n=31)[b,c]	3.16 (1.16)	3.39 (1.05)	3.32 (1.30)			
Future mindfulness in life				7.42 (2, 71)	0.173	*.001*
App (n=21)[b]	3.10 (1.18)	3.14 (1.35)	2.86 (1.46)			
App+O (n=24)[c]	4.13 (0.85)	3.96 (1.08)	4.04 (1.23)			
App+OPS (n=29)[c]	4.03 (0.82)	4.00 (0.85)	3.66 (1.08)			
Skill learning (perceived benefits)						
Learned about mindfulness				6.02 (2, 73)	0.142	*.004*
App (n=21)[b]	3.33 (0.97)	3.57 (1.17)	3.62 (1.02)			
App+O (n=24)[c]	4.04 (0.75)	4.00 (0.98)	4.29 (0.75)			
App+OPS (n=31)[c]	4.06 (0.73)	4.19 (0.65)	4.10 (0.70)			
Learned mindfulness skills				11.01 (2, 72)	0.234	*<.001*
App (n=21)[b]	3.24 (1.09)	3.43 (1.21)	3.67 (0.91)			
App+O (n=24)[c]	4.17 (0.70)	4.29 (0.75)	4.29 (0.86)			
App+OPS (n=30)[c]	4.13 (0.63)	4.33 (0.55)	4.17 (0.70)			

XSL•FO
RenderX

Outcomes[a]	Group means (SDs)			ANOVA group effects		
	Midpoint	Postintervention	Follow-up	*F* test (*df*)	η_p^2	*P* value
Awareness of thoughts and feelings				6.05 (2, 73)	0.142	*.004*
App (n=21)[b]	3.10 (1.18)	3.33 (1.24)	3.52 (1.17)			
App+O (n=24)[c]	4.04 (0.75)	3.96 (0.81)	4.17 (0.70)			
App+OPS (n=31)[c]	3.97 (0.84)	4.06 (0.81)	3.84 (0.86)			

[a]We evaluated the skewness of the data for the 9 outcome variables at each of the 3 time points across the whole sample (Multimedia Appendix 2). Only 11% (3/27) of the variables exceeded thresholds (−1 to 1) for symmetry of the normal distribution. The skewed distribution for these 3 variables is not anticipated to increase the risk for type I errors [52]. Of note, the distribution of these 3 variables were not skewed at the midpoint. Rather, the distribution of these variables became skewed over time. This data pattern indicates a reduction in app use over time, rather than an unexpected pattern of behavior. The Kolmogorov-Smirnov tests were significant (*P*<.05; Multimedia Appendix 2), indicating that the residuals were not normally distributed. However, the central limit theorem suggests that for larger sample sizes (≥100), the violation of normality does not present a significant concern. Furthermore, the sampling distribution of the observations may be normal, even if the residuals do not precisely follow a normal distribution. For partial eta squared, ηp2≥0.01 indicates a small effect, ηp2≥0.06 indicates a medium effect, and ηp2≥0.14 indicates a large effect. *P* values in italics meet the criteria for the Bonferroni-adjusted significance threshold (*P*<.0055).

[b,c]Pairwise contrasts with different footnote designators are significant after Bonferroni adjustment for multiple comparisons.

Similarly, there was an overall effect of group (human support condition) on cumulative any and all (mindfulness and nonmindfulness) sessions, $F_{2,117}$=15.00, η_p^2=0.20; *P*=.001. Pairwise comparisons indicate that App participants completed significantly fewer total sessions than App+O (*P*=.001) and App+OPS (*P*<.001) participants, but total sessions did not significantly differ between App+O and App+OPS groups (*P*=.37; Figure 2B).

Figure 2. Engagement with mTDI group means by time point (midpoint, postintervention, and 1-mo follow-up assessments) for participants randomly allocated to use the app without human support enhancement (app as usual [APP]; solid orange line), participants randomly allocated to attend a 1-time orientation (app+orientation [APP+O]; dashed green line), and participants randomly allocated to attend the orientation and be placed in a peer supportive accountability group (app+orientation and peer support [APP+OPS]; dotted blue line). Engagement outcomes include objectively captured: (A) cumulative mindfulness minutes, (B) cumulative any and all sessions. The y-axes reflect totals. mTDI: mobile technology–delivered intervention.

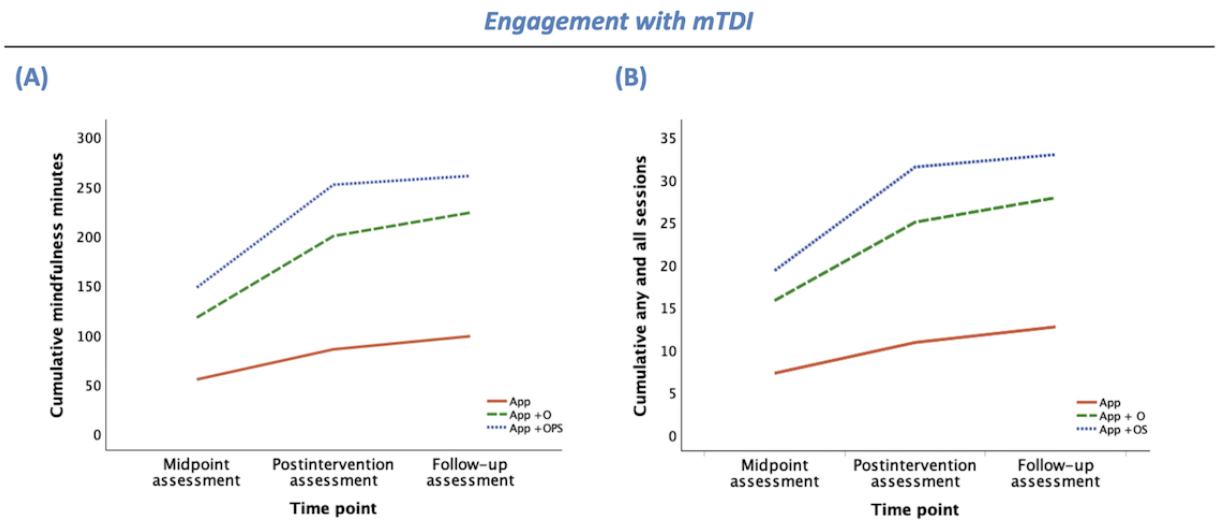

XSL•FO
RenderX

Figure 3. Practice integration (sustainability) group means by time point (midpoint, postintervention, and 1-mo follow-up assessments) for participants randomly allocated to use the app without human support enhancement (app as usual [APP]; solid orange line), participants randomly allocated to attend a 1-time orientation (app+orientation [APP+O]; dashed green line), and participants randomly allocated to attend the orientation and be placed in a peer supportive accountability group (app+orientation and peer support [APP+OPS]; dotted blue line). Practice integration outcomes include self-reported: (A) everyday mindfulness, (B) future app (Headspace) use, (C) future mindfulness practice on own, (D) future mindfulness in life. The y-axes reflect scale means (range 1-5).

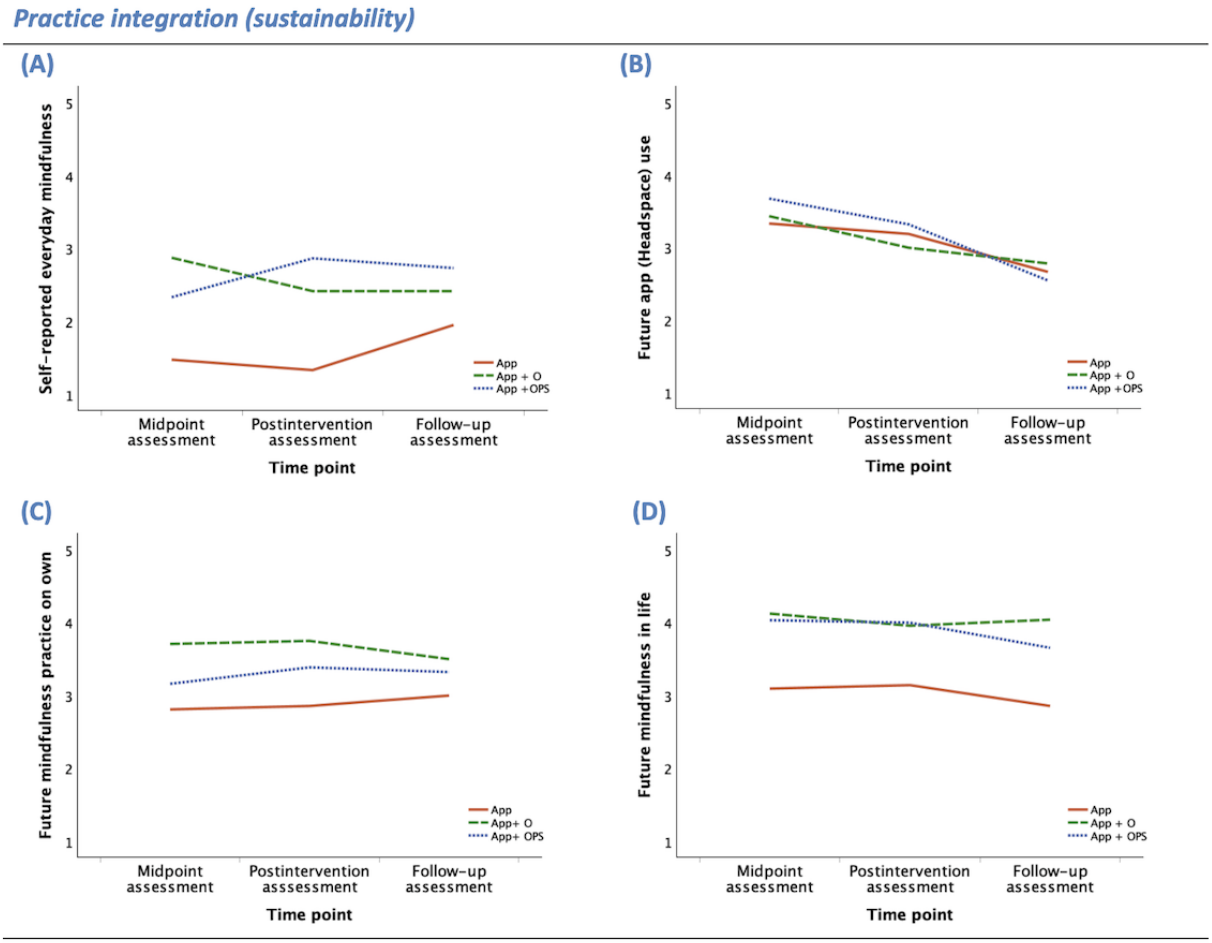

XSL•FO
RenderX

Figure 4. Skill learning (perceived benefits) group means by time point (midpoint, postintervention, and 1-mo follow-up assessments) for participants randomly allocated to use the app without human support enhancement (app as usual [APP]; solid orange line), participants randomly allocated to attend a 1-time orientation (app+orientation [APP+O]; dashed green line), and participants randomly allocated to attend the orientation and be placed in a peer supportive accountability group (app+orientation and peer support [APP+OPS]; dotted blue line). Skill learning outcomes include self-reported: (A) learning about mindfulness, (B) learning mindfulness skills, and (C) awareness of thoughts and feelings. The y-axes reflect scale means (range 1-5).

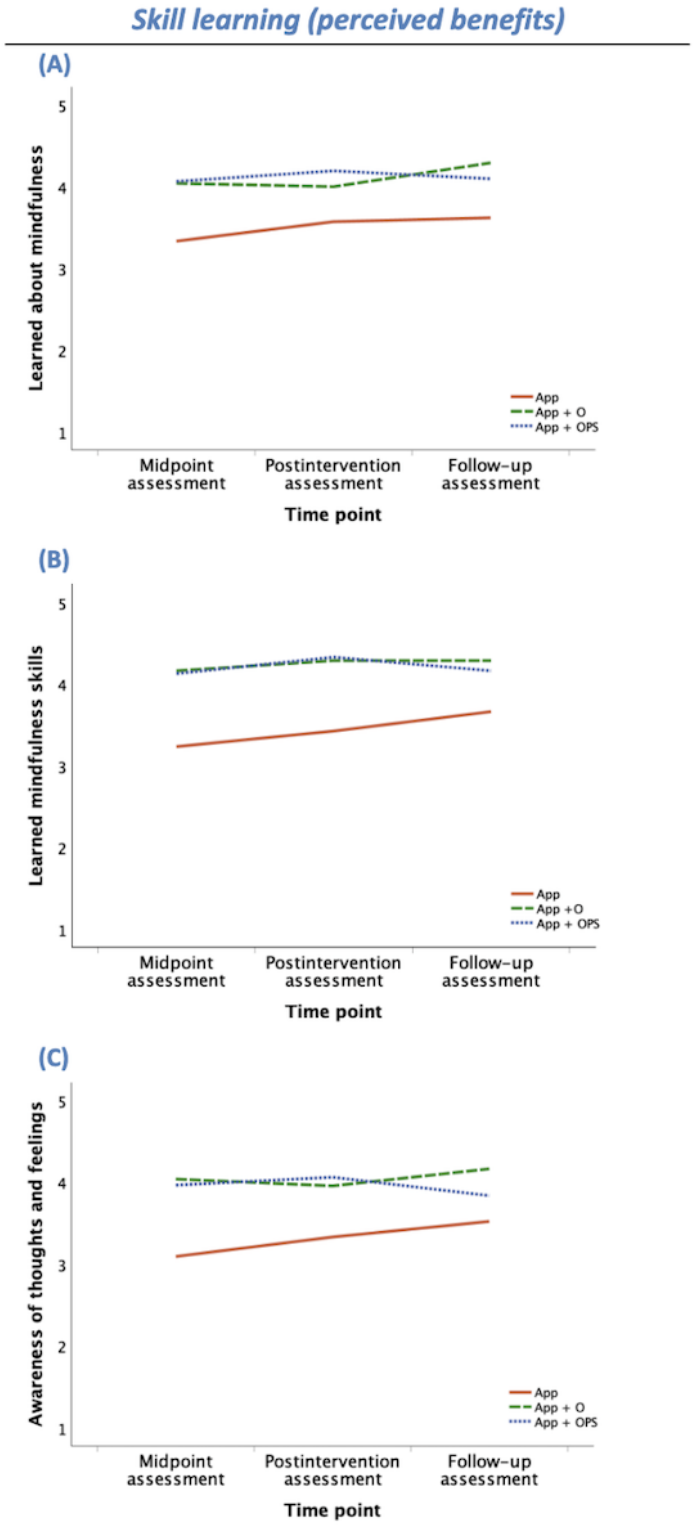

Practice Integration (Sustainability)

As shown in Table 2 (with means and SDs), there was an overall effect of group (human support condition) on self-reported everyday nonguided mindfulness practice, $F_{2,72}$=6.20, η_p^2=0.15; *P*=.003. Pairwise comparisons, with conservative Bonferroni correction, indicate that App participants engaged in significantly less nonguided mindfulness practice than App+O (*P*=.02) and App+OPS (*P*=.005) participants, but everyday nonguided mindfulness practice did not significantly differ between App+O and App+OPS groups (*P*=.99; Figure 3A).

XSL•FO
RenderX

There was no effect of group (human support condition) on self-reported likelihood to use the Headspace app in the future ($F_{2,72}$=0.11, η_p^2=0.003; *P*=.89; Figure 3B). Meanwhile, the effect of group on self-reported likelihood to practice mindfulness exercises on one's own in the future was right at the cutoff of statistical significance, with a medium effect size ($F_{2,73}$=3.10, η_p^2=0.08; *P*=.05). Pairwise comparisons indicate that App participants were significantly less likely to believe that they would engage in mindfulness exercises on their own in the future compared to App+O (*P*=.045) participants. App+OPS participants did not differ from App (*P*=.51) or App+O (*P*=.60) participants (Figure 3C).

There was an overall effect of group (human support condition) on self-reported likelihood of future mindfulness in everyday life ($F_{2,71}$=7.42, η_p^2=0.17; *P*=.001). Pairwise comparisons indicate that App participants were less likely to believe that they would engage in everyday mindfulness in the future than App+O (*P*=.002) and App+OPS (*P*=.006) participants. App+O and App+OPS groups did not differ (*P*=.99; Figure 3D).

Skill Learning (Perceived Benefits)

As shown in Table 2, there was an overall effect of group (human support condition) on the extent to which participants reported that they had learned about mindfulness through the intervention ($F_{2,73}$=6.02, η_p^2=0.14; *P*=.004). Pairwise comparisons indicate that App participants endorsed learning less about mindfulness than App+O (*P*=.01) and App+OPS (*P*=.007) participants, and App+O and App+OPS groups did not differ (*P*=.99; Figure 4A).

Similarly, there was an overall effect of group (human support condition) on the extent to which participants reported that they had gained mindfulness skills through this intervention ($F_{2,72}$=11.01, η_p^2=0.23; *P*<.001). App participants endorsed gaining mindfulness skills to a lesser extent than App+O (*P*<.001) and App+OPS (*P*<.001) participants, and App+O and App+OPS groups did not differ (*P*=.99; Figure 4B).

Finally, there was also an overall effect of group (human support condition) on the extent to which participants reported that they had gained awareness of their thoughts and feelings through this intervention ($F_{2,73}$=6.05, η_p^2=0.14; *P*=.004). App participants endorsed gaining this awareness to a lesser extent than App+O (*P*=.006) and App+OPS (*P*=.01) participants, and App+O and App+OPS groups did not differ (*P*=.99; Figure 4C).

Discussion

Summary of Findings

This study demonstrated that empirically driven human support enhancements can increase (1) engagement with an mTDI; (2) the integration of intervention practice into daily life; and (3) learning, including perceived intervention-based awareness and skills. Specifically, this study found that users randomly allocated to attend a 90-minute, face-to-face orientation session to the mTDI—whether with or without additional random allocation to a peer supportive accountability group—demonstrated significantly higher (1) intervention engagement, including total minutes of mindfulness content and total number of sessions overall; (2) integration of intervention practice into everyday life, both currently and anticipated in the future; and (3) learning about mindfulness and related skills and an awareness of thoughts and feelings. Among participants randomly allocated to attend an orientation, those who were further randomly allocated to peer supportive accountability groups exhibited higher mean scores on app use across the trial assessment points, but this was not a significant difference. Furthermore, participants randomly allocated to either of the human support–enhanced conditions generally demonstrated significant benefits for the integration of skills and perceived learning compared to those randomly allocated to use the app alone, but there were no significant differences between these 2 human support–enhanced groups. These findings highlight the varied benefits of including human support in otherwise self-guided digital mental health interventions, but they leave open important questions about the connections among human support, engagement, and broader benefits for daily-life integration and learning.

mTDI Engagement, Practice Integration, and Skill Learning: Understanding and Optimizing Connections

Overview

This study builds on the extensive evidence base demonstrating that mTDIs, specifically mindfulness-based mTDIs, have numerous benefits [1-3] and contributes to emerging research on how best to harness these benefits, both for increased engagement with the mTDI and for extended benefits into daily life. This study demonstrated broad benefits for human support on in-app engagement and real-life learning and applications of the intervention skills. These findings help advance our understanding of mTDI implementation science; yet, they also point to the need for continued research to better understand—as described in the following subsections—(1) the *active ingredients* of human support enhancements, including (2) *supportive accountability* and (3) the *optimal dosage* or level of mTDI engagement.

Active Ingredients of Human Support: What Promotes Success?

This study contributes to an emerging field of research attempting to better understand the active ingredients or mechanisms driving the benefits of human support. A recent review highlighted that including human support features that provide guidance (such as this study's orientation session) and foster social connectedness (initiated in this study's orientation session and deepened by the peer supportive accountability group) facilitate program engagement [15]. It is likely that human support serves, in part, to provide *reminders* for program engagement, which in turn enhance outcomes such as engagement, skill learning, and future intended practice [2]. However, it is notable that in this study, participants in the App+OPS group, who received frequent cues for mTDI engagement via the Facebook group, email digests, and in-person peer meetings, did not exhibit additive benefit beyond the orientation. These findings might indicate that receiving *personal, interactive guidance* at the outset was a critical

XSL•FO
RenderX

ingredient for engaging with a new mTDI, similar to other research [53]. Navigating the initial setup of an mTDI and exploring its content can be overwhelming for users, especially when lacking guidance and feedback [54]. Similarly, receiving this guidance from research staff early on may have enhanced the *credibility* of the mTDI, which has been linked to greater interest in downloading and using the mTDI [55] as well as continued engagement over time [56]. Having our team of trained interventionists conduct the orientation session may have been particularly impactful on building credibility because the mTDI may have been viewed as having expert endorsement [55].

In line with prior evidence on the benefits of brief motivational enhancement interventions for mTDI engagement [35-37], it is also plausible that the *content* of the orientation helped promote participant engagement with the mTDI. Overall, this study adds support for the utility of offering participants an initial orientation session to provide psychoeducation about the skills incorporated into the mTDI, build familiarity with the mTDI, provide guidance related to engagement, and foster motivation and connection. This represents an important step in examining and comparing the effects of different components of human support, although future research is needed to isolate the specific active ingredients of the orientation, including credibility, personalized guidance, motivational enhancement, and proactive efforts to address challenges and barriers. Within the larger field of digital mental health, it is increasingly recognized that human support features vary widely across studies, and few adequately describe what exactly these features entailed [22]. As a result, the effects of such enhancements have been mixed, with some studies finding benefits and others not, as reviewed by Bernstein et al [22]. Clarifying the impact of human support and the active ingredients of such support will also inform the answers to questions such as who should provide support (eg, peer, paraprofessional, or professional), when it should be provided (eg, once at the beginning or interspersed throughout the intervention), and the content or focus of the support (eg, providing psychoeducation, offering encouragement, evoking motivation, or helping with problem-solving). Ultimately, the answers to such questions are likely to be complex and will require future research exploring how individual differences interact with support enhancements.

The Impact of Supportive Accountability on mTDI Engagement

Although participants in the App+OPS group used the app more than those in the App+O group, this was not a significant difference. This finding suggests little additional benefit of the peer supportive accountability group above and beyond the orientation alone. However, there are many considerations when interpreting this finding. First, it is possible that the App+OPS group may have experienced other, intangible benefits beyond impact on objectively captured mTDI engagement; for example, participants in the App+OPS group overwhelmingly indicated that they enjoyed connecting with other students over shared experiences and generally found the group to be helpful and supportive [57]. This corroborates prior findings that peer supportive accountability is typically viewed positively by participants [24,29,41]. Second, the orientation allowed for some interaction with peers, including fellow participants as well as student members of the research team; therefore, App+O participants may have experienced some degree of peer supportive accountability as well. This study design is not fully sufficient to tease apart the impact of peer supportive accountability alone, and more research is needed on how best to harness its potential benefits.

Third, this study's findings may reflect the utility of integrating supportive accountability from a trained team of clinical *researchers*, without the additional benefit of adding supportive accountability from *peers*. Although all intervention groups, including those who used the mTDI without an orientation session (App group), had some degree of researcher contact across the trial (eg, for study consenting and assessment administration), the orientation session may have fostered a greater sense of participant accountability to the clinically trained research team, who facilitated the orientation. This is in line with reviews of mTDIs generally, which find that participants who enroll in mTDIs without any research contact show significantly higher attrition rates over time compared to participants who enroll through contact with research staff [13]. Those users who get started without researcher contact may not fully appreciate the effort and internal motivation that will be required to begin and sustain engagement over time [13], whereas researchers may be able to provide a more realistic picture to participants who do have contact with them. Therefore, it is critical to consider how human support enhancements may be implemented beyond the research context. Engagement with mTDIs is significantly improved by the structure and containment of research trials compared to engagement with the same programs in the "real world" [11]; therefore, finding ways to translate the benefits of an orientation session to the everyday mTDI user will be important.

This study's findings also raise questions about when, as well as how, to integrate human support. Our findings suggest that initial support (eg, an orientation) as users are beginning a new mTDI may be more beneficial for sustained engagement than continued support over time. Furthermore, with high levels of human support throughout the intervention, users could become reliant on external motivation and then decrease or stop engagement when the support is removed. In this study, the App+OPS group experienced a stark drop in human support, and likely supportive accountability, at the 2-month mark when the meetings ended and they stopped receiving email digests and posts in the web-based forum. Although app use tapered in all groups during the follow-up period, the decline was visually steepest for the App+OPS group. Though the pairwise contrasts between the App and App+O groups were not statistically significant, this pattern is worth investigating in future research. Perhaps, there is an optimal level of human support that, when tapered gradually, is conducive to sustaining mTDI engagement over the longer term.

Optimal Dosage of mTDI Engagement

Although certainly some level of engagement is needed to see benefits, the *optimal dosage* of mTDI use remains an open question. Research has recently begun to move beyond investigating the overall efficacy of mTDIs and toward exploring

XSL•FO
RenderX

the dose-response relationship between mTDI engagement and outcomes [58]. Often, there is an implicit assumption that more is better, but there is little understanding of the dose-response effect between mTDI engagement and outcomes and therefore the ideal amount of mTDI use [59,60]. More research is needed to better understand the complex relations between use and treatment response to promote optimal engagement and benefits.

There also remain questions about the reasons driving drop-offs in engagement after initial periods of use [61,62]. Declining mTDI use over time may not necessarily be indicative of a user's lack of interest in a program, lack of progress, or symptom severity. On the contrary, some users may have internalized the skills learned through the app or otherwise met their goals (eg, behavioral change and symptom improvement) such that they no longer needed to rely on the mTDI for support. Similar to attending psychotherapy sessions, mTDIs can be viewed as training wheels that provide users with psychoeducation and skill learning and practice. This, in turn, allows users to create personal well-being habits beyond the intervention, such as practicing mindfulness or other intervention skills in daily life—outside of an mTDI and beyond trial periods—as evidenced in this study. Thus, in some cases, the dwindling use of an mTDI after initial engagement might, in fact, be a marker of success. At a minimum, relying only on objectively measured app use through mTDI user data likely does not capture the full scope of individuals' engagement with the intervention content. Similar to this study, others have begun to incorporate additional self-report measures of skills practice outside of mTDIs and found that this is an important element in predicting benefits [60,63].

Limitations and Future Directions

This study used a rigorous RCT design with intent-to-treat analyses to examine multiple indicators of mTDI engagement and benefits, including objectively measured app use and self-reported practice integration and skill learning, through a 1-month follow-up period; yet, there are important limitations to address as well. First, although an RCT is the gold standard for establishing general intervention effects, other research is important for evaluating diverse intervention benefits (eg, dose-response effects, clinical utility, and social validity) as well as revealing processes and mechanisms of change. Future research should continue to explore various designs with a broader array of outcomes and more diverse samples to improve generalizability. Second, it is important to note that the onset of the COVID-19 pandemic occurred during the sixth semester of data collection, necessitating changes (eg, remote data collection, orientation, and App+OPS meetings) impacting cohorts 6 to 8. Third, although mTDIs provide methodological benefits by tracking in-app engagement and skills practice, data recording errors can occur; for example, 2.4% (3/123) of participants were excluded from mTDI engagement analyses due to glitches with their app-recorded use data. It is possible that other data-tracking errors could have occurred without the researchers' knowledge. Similarly, it is possible that participants could be inattentive, distracted, or asleep during some of the practices recorded by the app. Fourth, although the broad recruitment strategies and voluntary response sampling method have benefits for generalizability, they also yielded a sample consisting largely of women (reflecting both the higher proportion of women in college and their greater likelihood of meeting criteria for depression), which limits generalizability in other ways. Relatedly, not all participants completed the surveys assessing self-reported practice integration and skill learning, further limiting the generalizability of the findings.

Fifth, it is important to note that the App condition included participants who activated their Headspace access code at the same time as the rest of the cohort as well as a subset of participants who accessed their code *after* a 3-month waitlist period. Although analyses did not reveal any significant differences in mental health symptoms between App participants who could access Headspace immediately and those who waited 3 months, it is possible that combining these 2 groups may have affected the results of the App condition, particularly because prior research suggests that intervention engagement can be negatively impacted by time spent waiting for services [64].

Sixth, and related, the 3 randomly allocated groups were nested such that participants in all 3 groups received the base intervention (App), 2 of the nested groups attended an orientation (App+O and App+OPS), and 1 of these nested groups included additional peer supportive accountability features (App+OPS). Analyses tested for the incremental benefits of the 2 human support enhancements; yet, future research might benefit from examining the benefits of an app+peer supportive accountability condition without an orientation session.

Finally, this study's implementation of peer supportive accountability had some limitations that may have reduced the potential impact of the group on mTDI engagement and the integration of skills. The web-based forum included a "leaderboard" feature that displayed each member's use of the app from the current week and the trial period overall. This may have promoted external motivation, which is typically not as effective as intrinsic motivation for sustained engagement [15]. This type of feature also may function differently for individuals experiencing depression compared to those with other mental health symptoms. Given that self-criticism and negative self-concept are common among individuals with depression [65,66], having progress metrics shared with the larger group may have been perceived as discouraging rather than motivating. In addition, the use of Facebook for the online forum—selected for its social connection and privacy features—may have made engagement with the peer group more difficult because Facebook is becoming increasingly less popular among young adult participants [67]. Future studies should explore alternative platforms for web-based interaction among participants, ideally integrating all peer support features into the mTDI. Furthermore, future research should explore whether the benefits of human support, such as those demonstrated in this study, might also be harnessed with automated support, whether preprogrammed or assisted by interactive artificial intelligence. Beyond the more readily automatized elements such as providing information and reminders, research on the benefits of human support tends to emphasize more complex elements such as personalized prompts, interactive coaching, and professional guidance [2,22,34,53,54,68]. Notably, recent advances in artificial intelligence and machine learning might allow for the transfer

XSL•FO
RenderX

of some human support elements and benefits to automated programming [2]. Importantly, these developments raise crucial ethical and societal concerns that must be addressed [69].

Clinical Applications and Research Implications

This study's findings have important implications that move beyond foundational research on *whether* to use mindfulness-based mobile apps in clinical intervention work by investigating *how* best to maximize their benefits, such as through brief, face-to-face orientations and through small-group peer support networks. These results can be applied more broadly to mTDIs and also are likely to extend to other user populations; for example, the peer supportive accountability model implemented here—with the notable request for more face-to-face contact by the participants themselves—might be similarly relevant for mTDI users who share a common identity or life circumstance, such as high school students, employees, veterans, parents, or patients [25,26,28].

Boosting mTDIs with human support features has numerous applications in various settings and with diverse people who can serve to support engagement and thus enhance benefits; for example, many health care systems allow for providers within different departments (eg, behavioral health, primary care, and specialty clinics) to refer patients to mTDIs as a means of addressing lower-severity symptoms, supplementing preexisting mental health services, or bridging the patient to other levels of care, such as while they are on a waitlist [70]. Similarly, many employers now offer mental health mTDIs as part of their health insurance package [71], and there is emerging research on integrating mental health care into routine curricular and extracurricular settings on college campuses [72-74]. While mTDIs are increasingly incorporated into nontraditional settings, the typical approach is simply to provide individuals with access to the mTDI, leaving users to download and engage with the program on their own. The findings from this study suggest that evidence-based mTDIs can be even more effective and have a more powerful reach and impact if supplemented with human support; for example, professionals, paraprofessionals (without a mental health degree), or peers can serve as coaches to support engagement with evidence-based mTDIs through routine settings such as primary care clinics [75], workplaces [53], and college student academic services [76]. This study also suggests that an initial, one-time group introduction to the mTDI conducted by a knowledgeable source can produce a myriad of benefits and represents a cost-effective and scalable option for larger institutions; for example, hospitals and clinics, schools and community agencies, and corporate and nonprofit organizations can pair their offerings of mTDIs with regularly occurring orientation sessions that assist new users in accessing the program, demonstrate the app and its features, answer questions, and elicit motivation. This would allow many potential users to join at once, thus improving efficiency, and might also boost social connection and a sense of community—which in turn would promote mTDI use and well-being more broadly.

Conclusions

While mTDIs have great promise for increasing access to well-being interventions, relatively low rates of user uptake and engagement are current barriers to achieving the potential reach and impact of mTDIs. The results of this study illustrate that human support—particularly an initial face-to-face orientation session that provides psychoeducation, builds familiarity with the mTDI, and promotes motivation and connection—boosts mTDI engagement, enhances the integration of intervention practice into everyday life, and increases skill learning. Future work is needed to determine the active ingredients of the orientation (eg, credibility, personalized guidance, motivational enhancement, and proactive problem-solving) that might drive increased levels of engagement and the associated positive intervention benefits. Identifying optimal intervention dosage levels for evidence-based mTDIs also remains critical to fully harnessing the potential of these interventions and their role in improving societal health and well-being.

Acknowledgments

Special thanks to the following individuals, who collectively contributed to study design, data collection, data management, data analysis, and manuscript preparation: Noor Abdelfattah, JD; Maria Bandriwsky, MSGH; Alexandra Boutelle, BS; Marie Chamberlain, BS; Adriana Diviero, BS; Marcos Economides, PhD; Kerry Finnegan, BA; Axel Juarez; Hassan Khan, BS; Bradley Orzada; Kelly Polnaszek, PhD; Asnia Rafiq, BS; Andrew Rauch, MA; Genevieve Roth, MA; Margaret Schroer, MSN; Madeleine Suhs, BS; and Zuzu Trottier, BS. The authors are grateful for the contributions of the study participants to this research. This study was supported by internal research funding from Loyola University Chicago and in part by a subaward grant from M3EWB (Mechanisms Underlying Mind-Body Interventions and Measurement of Emotional Well Being), funded by NIH U24AT011281. Headspace did not provide any financial support to Loyola University Chicago.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

CSC: conceptualization (lead); formal analysis (supporting); funding acquisition (equal); investigation (lead); methodology (lead); project administration (lead); resources (equal); supervision (lead); original draft preparation (lead); manuscript review and editing (equal). BMD: conceptualization (supporting); data curation (equal); formal analysis (supporting); investigation (equal); methodology (supporting); project administration (equal); visualization (equal); original draft preparation (equal); manuscript review and editing (equal). CHG: data curation (equal); formal analysis (lead); investigation (equal); methodology (supporting);

XSL•FO
RenderX

project administration (equal); visualization (equal); original draft preparation (supporting); manuscript review and editing (supporting). IJK: data curation (supporting); formal analysis (supporting); investigation (supporting); project administration (supporting); visualization (equal); manuscript review and editing (supporting). JD: conceptualization (supporting); methodology (supporting); manuscript review and editing (supporting). RLS: conceptualization (supporting); formal analysis (supporting); funding acquisition (equal); investigation (supporting); methodology (supporting); project administration (supporting); resources (equal); supervision (supporting); original draft preparation (supporting); manuscript review and editing (equal). All authors commented on previous versions of the manuscript. All authors have read and approved the final manuscript.

Conflicts of Interest

JD is a consultant for Rebound Health, Chorus Health and MyAlloy. All other authors declare no conflicts of interest.

Multimedia Appendix 1
CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) checklist (V 1.6.1).
[PDF File (Adobe PDF File), 1003 KB - formative_v9i1e56963_app1.pdf]

Multimedia Appendix 2
Skewness values and SEs from the Kolmogorov-Smirnov test with Lilliefors correction for study outcome variables at 3 time points for the full study sample.
[PDF File (Adobe PDF File), 134 KB - formative_v9i1e56963_app2.pdf]

References

1. Conley CS, Gonzales CH, Huguenel BM, Rauch AA, Kahrilas IJ, Duffecy J, et al. Benefits of a technology-delivered mindfulness intervention for psychological distress and positive wellbeing in depressed college students: post-intervention and follow-up effects from an RCT. Mindfulness 2024 Jul 29;15(7):1739-1758. [doi: 10.1007/S12671-024-02398-3]
2. Linardon J, Cuijpers P, Carlbring P, Messer M, Fuller-Tyszkiewicz M. The efficacy of app-supported smartphone interventions for mental health problems: a meta-analysis of randomized controlled trials. World Psychiatry 2019 Oct 09;18(3):325-336 [FREE Full text] [doi: 10.1002/wps.20673] [Medline: 31496095]
3. Gál É, tefan S, Cristea IA. The efficacy of mindfulness meditation apps in enhancing users' well-being and mental health related outcomes: a meta-analysis of randomized controlled trials. J Affect Disord 2021 Jan 15;279:131-142. [doi: 10.1016/j.jad.2020.09.134] [Medline: 33049431]
4. Flett JA, Hayne H, Riordan BC, Thompson LM, Conner TS. Mobile mindfulness meditation: a randomised controlled trial of the effect of two popular apps on mental health. Mindfulness 2018 Oct 31;10(5):863-876. [doi: 10.1007/S12671-018-1050-9]
5. Kozlov E, Bantum E, Pagano I, Walser R, Ramsey K, Taylor K, et al. The reach, use, and impact of a free mHealth mindfulness app in the general population: mobile data analysis. JMIR Ment Health 2020 Nov 27;7(11):e23377 [FREE Full text] [doi: 10.2196/23377] [Medline: 33245289]
6. Key substance use and mental health indicators in the United States: results from the 2021 National Survey on Drug Use and Health (HHS publication no. PEP22-07-01-005, NSDUH Series H-57). Substance Abuse and Mental Health Services Administration. 2022. URL: https://www.samhsa.gov/data/sites/default/files/reports/rpt39443/2021NSDUHFFRRev010323.pdf [accessed 2025-05-29]
7. Kern A, Hong V, Song J, Lipson SK, Eisenberg D. Mental health apps in a college setting: openness, usage, and attitudes. Mhealth 2018 Jun;4:20 [FREE Full text] [doi: 10.21037/mhealth.2018.06.01] [Medline: 30050916]
8. Bautista J, Schueller SM. Understanding the adoption and use of digital mental health apps among college students: secondary analysis of a national survey. JMIR Ment Health 2023 Mar 22;10:e43942 [FREE Full text] [doi: 10.2196/43942] [Medline: 36947115]
9. Lattie EG, Kornfield R, Ringland KE, Zhang R, Winquist N, Reddy M. Designing mental health technologies that support the social ecosystem of college students. Proc SIGCHI Conf Hum Factor Comput Syst 2020 Apr;2020:1145 [FREE Full text] [doi: 10.1145/3313831.3376362] [Medline: 32656549]
10. Linardon J, Fitzsimmons-Craft EE, Brennan L, Barillaro M, Wilfley DE. Dropout from interpersonal psychotherapy for mental health disorders: a systematic review and meta-analysis. Psychother Res 2019 Oct;29(7):870-881. [doi: 10.1080/10503307.2018.1497215] [Medline: 30005586]
11. Baumel A, Muench F, Edan S, Kane JM. Objective user engagement with mental health apps: systematic search and panel-based usage analysis. J Med Internet Res 2019 Sep 25;21(9):e14567 [FREE Full text] [doi: 10.2196/14567] [Medline: 31573916]
12. Lattie EG, Cohen KA, Hersch E, Williams KD, Kruzan KP, MacIver C, et al. Uptake and effectiveness of a self-guided mobile app platform for college student mental health. Internet Interv 2022 Mar;27:100493 [FREE Full text] [doi: 10.1016/j.invent.2021.100493] [Medline: 35036331]

XSL•FO
RenderX

13. Linardon J, Fuller-Tyszkiewicz M. Attrition and adherence in smartphone-delivered interventions for mental health problems: a systematic and meta-analytic review. J Consult Clin Psychol 2020 Jan;88(1):1-13. [doi: 10.1037/ccp0000459] [Medline: 31697093]
14. Mohr DC, Duffecy J, Ho J, Kwasny M, Cai X, Burns MN, et al. A randomized controlled trial evaluating a manualized TeleCoaching protocol for improving adherence to a web-based intervention for the treatment of depression. PLoS One 2013;8(8):e70086 [FREE Full text] [doi: 10.1371/journal.pone.0070086] [Medline: 23990896]
15. Borghouts J, Eikey E, Mark G, De Leon C, Schueller SM, Schneider M, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. J Med Internet Res 2021 Mar 24;23(3):e24387 [FREE Full text] [doi: 10.2196/24387] [Medline: 33759801]
16. Eysenbach G. The law of attrition. J Med Internet Res 2005 Mar 31;7(1):e11 [FREE Full text] [doi: 10.2196/jmir.7.1.e11] [Medline: 15829473]
17. Nahum-Shani I, Shaw SD, Carpenter SM, Murphy SA, Yoon C. Engagement in digital interventions. Am Psychol 2022 Oct;77(7):836-852 [FREE Full text] [doi: 10.1037/amp0000983] [Medline: 35298199]
18. Torous J, Michalak EE, O'Brien HL. Digital health and engagement-looking behind the measures and methods. JAMA Netw Open 2020 Jul 01;3(7):e2010918 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.10918] [Medline: 32678446]
19. Milne-Ives M, Homer SR, Andrade J, Meinert E. Potential associations between behavior change techniques and engagement with mobile health apps: a systematic review. Front Psychol 2023 Sep 18;14:1227443 [FREE Full text] [doi: 10.3389/fpsyg.2023.1227443] [Medline: 37794916]
20. Cheung K, Ling W, Karr CJ, Weingardt K, Schueller S, Mohr DC. Evaluation of a recommender app for apps for the treatment of depression and anxiety: an analysis of longitudinal user engagement. J Am Med Inform Assoc 2018 Aug 01;25(8):955-962 [FREE Full text] [doi: 10.1093/jamia/ocy023] [Medline: 29659857]
21. Economides M, Martman J, Bell MJ, Sanderson B. Improvements in stress, affect, and irritability following brief use of a mindfulness-based smartphone app: a randomized controlled trial. Mindfulness (N Y) 2018 Mar 1;9(5):1584-1593 [FREE Full text] [doi: 10.1007/s12671-018-0905-4] [Medline: 30294390]
22. Bernstein EE, Weingarden H, Wolfe EC, Hall MD, Snorrason I, Wilhelm S. Human support in app-based cognitive behavioral therapies for emotional disorders: scoping review. J Med Internet Res 2022 Apr 08;24(4):e33307 [FREE Full text] [doi: 10.2196/33307] [Medline: 35394434]
23. Mohr DC, Duffecy J, Jin L, Ludman EJ, Lewis A, Begale M, et al. Multimodal e-mental health treatment for depression: a feasibility trial. J Med Internet Res 2010 Dec 19;12(5):e48 [FREE Full text] [doi: 10.2196/jmir.1370] [Medline: 21169164]
24. Lattie EG, Ho J, Sargent E, Tomasino KN, Smith JD, Brown CH, et al. Teens engaged in collaborative health: the feasibility and acceptability of an online skill-building intervention for adolescents at risk for depression. Internet Interv 2017 Jun;8:15-26 [FREE Full text] [doi: 10.1016/j.invent.2017.02.003] [Medline: 28584734]
25. Duffecy J, Sanford S, Wagner L, Begale M, Nawacki E, Mohr DC. Project onward: an innovative e-health intervention for cancer survivors. Psychooncology 2013 Apr 21;22(4):947-951 [FREE Full text] [doi: 10.1002/pon.3075] [Medline: 22438297]
26. Possemato K, Wu J, Greene C, MacQueen R, Blonigen D, Wade M, et al. Web-based problem-solving training with and without peer support in veterans with unmet mental health needs: pilot study of feasibility, user acceptability, and participant engagement. J Med Internet Res 2022 Jan 13;24(1):e29559 [FREE Full text] [doi: 10.2196/29559] [Medline: 35023846]
27. Tomasino KN, Lattie EG, Ho J, Palac HL, Kaiser SM, Mohr DC. Harnessing peer support in an online intervention for older adults with depression. Am J Geriatr Psychiatry 2017 Oct;25(10):1109-1119 [FREE Full text] [doi: 10.1016/j.jagp.2017.04.015] [Medline: 28571785]
28. Duffecy J, Grekin R, Hinkel H, Gallivan N, Nelson G, O'Hara MW. A group-based online intervention to prevent postpartum depression (Sunnyside): feasibility randomized controlled trial. JMIR Ment Health 2019 May 28;6(5):e10778 [FREE Full text] [doi: 10.2196/10778] [Medline: 31140443]
29. Duffecy J, Grekin R, Long JD, Mills JA, O'Hara M. Randomized controlled trial of Sunnyside: individual versus group-based online interventions to prevent postpartum depression. J Affect Disord 2022 Aug 15;311:538-547 [FREE Full text] [doi: 10.1016/j.jad.2022.05.123] [Medline: 35654284]
30. Ho J, Corden ME, Caccamo L, Tomasino KN, Duffecy J, Begale M, et al. Design and evaluation of a peer network to support adherence to a web-based intervention for adolescents. Internet Interv 2016 Nov;6:50-56 [FREE Full text] [doi: 10.1016/j.invent.2016.09.005] [Medline: 27722095]
31. Fortuna KL, Naslund JA, LaCroix JM, Bianco CL, Brooks JM, Zisman-Ilani Y, et al. Digital peer support mental health interventions for people with a lived experience of a serious mental illness: systematic review. JMIR Ment Health 2020 Apr 03;7(4):e16460 [FREE Full text] [doi: 10.2196/16460] [Medline: 32243256]
32. Dean S, Britt E, Bell E, Stanley J, Collings S. Motivational interviewing to enhance adolescent mental health treatment engagement: a randomized clinical trial. Psychol Med 2016 Jul 05;46(9):1961-1969. [doi: 10.1017/S0033291716000568] [Medline: 27045520]
33. Swartz HA, Zuckoff A, Grote NK, Spielvogle HN, Bledsoe SE, Shear MK, et al. Engaging depressed patients in psychotherapy: integrating techniques from motivational interviewing and ethnographic interviewing to improve treatment participation. Prof Psychol Res Pr 2007 Aug;38(4):430-439. [doi: 10.1037/0735-7028.38.4.430]

XSL•FO
RenderX

34. Lattie EG, Graham AK, Hadjistavropoulos HD, Dear BF, Titov N, Mohr DC. Guidance on defining the scope and development of text-based coaching protocols for digital mental health interventions. Digit Health 2019 Dec 16;5:2055207619896145 [FREE Full text] [doi: 10.1177/2055207619896145] [Medline: 31897306]
35. Gex KS, Mun EY, Barnett NP, McDevitt-Murphy ME, Ruggiero KJ, Thurston IB, et al. A randomized pilot trial of a mobile delivered brief motivational interviewing and behavioral economic alcohol intervention for emerging adults. Psychol Addict Behav 2023 May;37(3):462-474 [FREE Full text] [doi: 10.1037/adb0000838] [Medline: 35482647]
36. Hasin DS, Aharonovich E, Zingman BS, Stohl M, Walsh C, Elliott JC, et al. HealthCall: a randomized trial assessing a smartphone enhancement of brief interventions to reduce heavy drinking in HIV care. J Subst Abuse Treat 2022 Jul;138:108733 [FREE Full text] [doi: 10.1016/j.jsat.2022.108733] [Medline: 35131124]
37. John JC, Wang J, McNeill LH, Basen-Engquist K, Hoover DS, Daniel CR, et al. A mixed methods study on engagement and satisfaction with a digitally-enhanced pilot intervention among African American and Hispanic women. J Immigr Minor Health 2021 Oct 11;23(5):1011-1020 [FREE Full text] [doi: 10.1007/s10903-020-01095-2] [Medline: 33043411]
38. Venkatesan A, Rahimi L, Kaur M, Mosunic C. Digital cognitive behavior therapy intervention for depression and anxiety: retrospective study. JMIR Ment Health 2020 Aug 26;7(8):e21304 [FREE Full text] [doi: 10.2196/21304] [Medline: 32845246]
39. Nicholson J, Wright SM, Carlisle AM, Sweeney MA, McHugo GJ. The WorkingWell mobile phone app for individuals with serious mental illnesses: proof-of-concept, mixed-methods feasibility study. JMIR Ment Health 2018 Oct 25;5(4):e11383 [FREE Full text] [doi: 10.2196/11383] [Medline: 30361199]
40. Bur OT, Krieger T, Moritz S, Klein JP, Berger T. Optimizing the context of support to improve outcomes of internet-based self-help in individuals with depressive symptoms: protocol for a randomized factorial trial. JMIR Res Protoc 2021 Feb 02;10(2):e21207 [FREE Full text] [doi: 10.2196/21207] [Medline: 33528377]
41. Kazantzis N, Whittington C, Zelencich L, Kyrios M, Norton PJ, Hofmann SG. Quantity and quality of homework compliance: a meta-analysis of relations with outcome in cognitive behavior therapy. Behav Ther 2016 Sep;47(5):755-772. [doi: 10.1016/j.beth.2016.05.002] [Medline: 27816086]
42. Schlosser D, Campellone T, Kim D, Truong B, Vergani S, Ward C, et al. Feasibility of PRIME: a cognitive neuroscience-informed mobile app intervention to enhance motivated behavior and improve quality of life in recent onset schizophrenia. JMIR Res Protoc 2016 Apr 28;5(2):e77 [FREE Full text] [doi: 10.2196/resprot.5450] [Medline: 27125771]
43. Kroenke K, Strine TW, Spitzer RL, Williams JB, Berry JT, Mokdad AH. The PHQ-8 as a measure of current depression in the general population. J Affect Disord 2009 Apr;114(1-3):163-173. [doi: 10.1016/j.jad.2008.06.026] [Medline: 18752852]
44. Kroenke K, Spitzer RL. The PHQ-9: a new depression diagnostic and severity measure. Psychiatr Ann 2002 Sep 01;32(9):509-515. [doi: 10.3928/0048-5713-20020901-06]
45. Lakens D. Sample size justification. Collabra Psychol 2022;8(1):33267. [doi: 10.1525/collabra.33267]
46. Headspace: meditation and sleep. Headspace Inc. URL: https://apps.apple.com/us/app/headspace-mindful-meditation/id493145008 [accessed 2024-05-29]
47. Miller WR, Rollnick S. Motivational Interviewing: Helping People Change and Grow. 4th edition. New York, NY: The Guilford Press; 2023.
48. Worthen D, Luiselli JK. Social validity assessment and intervention evaluation of mindfulness education and practices with high school students. Mindfulness 2016 Dec 19;8(4):903-910. [doi: 10.1007/s12671-016-0664-z]
49. SPSS. Version 28.0. IBM Corp. URL: https://www.ibm.com/support/pages/downloading-ibm-spss-statistics-28 [accessed 2024-05-29]
50. Cohen J. The earth is round (P<.05). Am Psychol 1994;49(12):997-1003. [doi: 10.1037//0003-066X.49.12.997]
51. Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2nd edition. New York, NY: Routledge; 1988.
52. Knief U, Forstmeier W. Violating the normality assumption may be the lesser of two evils. Behav Res Methods 2021 Dec 07;53(6):2576-2590 [FREE Full text] [doi: 10.3758/s13428-021-01587-5] [Medline: 33963496]
53. Jesuthasan J, Low M, Ong T. The impact of personalized human support on engagement with behavioral intervention technologies for employee mental health: an exploratory retrospective study. Front Digit Health 2022 Apr 27;4:846375 [FREE Full text] [doi: 10.3389/fdgth.2022.846375] [Medline: 35574254]
54. Stiles-Shields C, Montague E, Lattie EG, Kwasny MJ, Mohr DC. What might get in the way: Barriers to the use of apps for depression. Digit Health 2017 Jun 08;3:2055207617713827 [FREE Full text] [doi: 10.1177/2055207617713827] [Medline: 29942605]
55. Lipschitz J, Miller CJ, Hogan TP, Burdick KE, Lippin-Foster R, Simon SR, et al. Adoption of mobile apps for depression and anxiety: cross-sectional survey study on patient interest and barriers to engagement. JMIR Ment Health 2019 Jan 25;6(1):e11334 [FREE Full text] [doi: 10.2196/11334] [Medline: 30681968]
56. Bernstein EE, Weingarden H, Greenberg JL, Williams J, Hoeppner SS, Snorrason I, et al. Credibility and expectancy of smartphone-based cognitive behavioral therapy among adults with body dysmorphic disorder. J Obsessive Compuls Relat Disord 2023 Jan;36:100781 [FREE Full text] [doi: 10.1016/j.jocrd.2023.100781] [Medline: 38313683]
57. Gonzales CH. The impact of peer supportive accountability on use of a mindfulness app in depressed college students: a mixed methods study. Loyola University. URL: https://ecommons.luc.edu/cgi/viewcontent.cgi?article=4980&context=luc_diss [accessed 2024-05-29]

XSL•FO
RenderX

58. Zeng Y, Guo Y, Li L, Hong YA, Li Y, Zhu M, et al. Relationship between patient engagement and depressive symptoms among people living with HIV in a mobile health intervention: secondary analysis of a randomized controlled trial. JMIR Mhealth Uhealth 2020 Oct 29;8(10):e20847 [FREE Full text] [doi: 10.2196/20847] [Medline: 33118956]
59. Conley CS, Raposa EB, Bartolotta K, Broner SE, Hareli M, Forbes N, et al. The impact of mobile technology-delivered interventions on youth well-being: systematic review and 3-level meta-analysis. JMIR Ment Health 2022 Jul 29;9(7):e34254 [FREE Full text] [doi: 10.2196/34254] [Medline: 35904845]
60. Weingarden H, Garriga Calleja R, Greenberg JL, Snorrason I, Matic A, Quist R, et al. Characterizing observed and effective behavioral engagement with smartphone cognitive behavioral therapy for body dysmorphic disorder: a methods roadmap and use case. Internet Interv 2023 Apr;32:100615 [FREE Full text] [doi: 10.1016/j.invent.2023.100615] [Medline: 36969390]
61. Molloy A, Anderson PL. Engagement with mobile health interventions for depression: a systematic review. Internet Interv 2021 Dec;26:100454 [FREE Full text] [doi: 10.1016/j.invent.2021.100454] [Medline: 34621626]
62. Pham Q, Graham G, Carrion C, Morita PP, Seto E, Stinson JN, et al. A library of analytic indicators to evaluate effective engagement with consumer mHealth apps for chronic conditions: scoping review. JMIR Mhealth Uhealth 2019 Jan 18;7(1):e11941 [FREE Full text] [doi: 10.2196/11941] [Medline: 30664463]
63. Furukawa TA, Horikoshi M, Fujita H, Tsujino N, Jinnin R, Kako Y, et al. Cognitive and behavioral skills exercises completed by patients with major depression during smartphone cognitive behavioral therapy: secondary analysis of a randomized controlled trial. JMIR Ment Health 2018 Jan 11;5(1):e4 [FREE Full text] [doi: 10.2196/mental.9092] [Medline: 29326098]
64. Wang J, Knitter AC, Staab EM, Beckman N, Araújo FS, Vinci LM, et al. Association between wait time and behavioral health appointment attendance across patient characteristics. Psychol Serv 2023 Nov;20(4):983-987. [doi: 10.1037/ser0000768] [Medline: 37141046]
65. Ehret AM, Joormann J, Berking M. Examining risk and resilience factors for depression: the role of self-criticism and self-compassion. Cogn Emot 2015 Dec 17;29(8):1496-1504. [doi: 10.1080/02699931.2014.992394] [Medline: 25517734]
66. Reed-Fitzke K. The role of self-concepts in emerging adult depression: a systematic research synthesis. J Adult Dev 2019 Jan 2;27(1):36-48. [doi: 10.1007/S10804-018-09324-7]
67. Anderson M, Faverio M, Gottfried J. Teens, social media and technology. Pew Research Center. 2023. URL: https://www.pewresearch.org/internet/2023/12/11/teens-social-media-and-technology-2023/ [accessed 2024-05-29]
68. Torous J, Lipschitz J, Ng M, Firth J. Dropout rates in clinical trials of smartphone apps for depressive symptoms: a systematic review and meta-analysis. J Affect Disord 2020 Feb 15;263:413-419. [doi: 10.1016/j.jad.2019.11.167] [Medline: 31969272]
69. Fiske A, Henningsen P, Buyx A. Your robot therapist will see you now: ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. J Med Internet Res 2019 May 09;21(5):e13216 [FREE Full text] [doi: 10.2196/13216] [Medline: 31094356]
70. Eberhart L, Seegan P, McGuire J, Hu H, Tripuraneni BR, Miller MJ. Attributes of provider referrals for digital mental health applications in an integrated health system, 2019-2021. Psychiatr Serv 2024 Jan 01;75(1):6-16. [doi: 10.1176/appi.ps.20220401] [Medline: 37494117]
71. Histon T. How to pick employee mental health apps. Kaiser Permanente®. URL: https://business.kaiserpermanente.org/insights/mental-health-workplace/mental-health-apps-workforce-wellness [accessed 2023-10-06]
72. Abelson S, Lipson SK, Eisenberg D. Mental health in college populations: a multidisciplinary review of what works, evidence gaps, and paths forward. In: Perna LW, editor. Higher Education: Handbook of Theory and Research. Volume 37. Cham, Switzerland: Springer; 2022:133-238.
73. Conley CS, Durlak JA, Kirsch AC. A Meta-analysis of universal mental health prevention programs for higher education students. Prev Sci 2015 May 7;16(4):487-507. [doi: 10.1007/s11121-015-0543-1] [Medline: 25744536]
74. Upsher R, Nobili A, Hughes G, Byrom N. A systematic review of interventions embedded in curriculum to improve university student wellbeing. Educ Res Rev 2022 Nov;37:100464. [doi: 10.1016/j.edurev.2022.100464]
75. Graham AK, Greene CJ, Kwasny MJ, Kaiser SM, Lieponis P, Powell T, et al. Coached mobile app platform for the treatment of depression and anxiety among primary care patients: a randomized clinical trial. JAMA Psychiatry 2020 Sep 01;77(9):906-914 [FREE Full text] [doi: 10.1001/jamapsychiatry.2020.1011] [Medline: 32432695]
76. Shack M, Broner SE, Conley CS. Facilitators and barriers to mentor-mentee relationships in a mental health-app intervention. Loyola University. URL: https://ecommons.luc.edu/cgi/viewcontent.cgi?article=1940&context=ures [accessed 2024-05-26]

Abbreviations

App: app as usual
App+O: app+orientation
App+OPS: app+orientation and peer supportive accountability
CONSORT: Consolidated Standards of Reporting Trials
MI: motivational interviewing
mTDI: mobile technology–delivered intervention
PHQ-8: Patient Health Questionnaire-8
RCT: randomized controlled trial

XSL•FO
RenderX

TDI: technology-delivered intervention

Edited by A Mavragani; submitted 01.03.24; peer-reviewed by S Dix, M Hecht; comments to author 19.08.24; revised version received 20.11.24; accepted 23.04.25; published 25.08.25.

Please cite as:
Conley CS, DeLorenzo BM, Gonzales CH, Kahrilas IJ, Duffecy J, Silton RL
Enhancing Engagement, Practice Integration, and Skill Learning in Mobile Technology–Delivered Interventions Using Human Support: Randomized Controlled Trial With Depressed College Students
JMIR Form Res 2025;9:e56963
URL: https://formative.jmir.org/2025/1/e56963
doi:10.2196/56963
PMID:

©Colleen S Conley, Brynn M DeLorenzo, Carol H Gonzales, Ian J Kahrilas, Jennifer Duffecy, Rebecca L Silton. Originally published in JMIR Formative Research (https://formative.jmir.org), 25.08.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Supporting Police Well-Being Through an Adaptive Shift Management System: Co-Design Study

Olumuyiwa Temitope Ayorinde[1], PhD; Huseyin Dogan[1], PhD; Festus Fatai Adedoyin[1], PhD; Nan Jiang[1], PhD; Fiona Bitters[2], MST; Sara Dempsey[2]

[1]Department of Computing and Informatics, Faculty of Science and Technology, Bournemouth University, Poole, United Kingdom

[2]Hampshire Constabulary, Hampshire and Isle of Wight, United Kingdom

Corresponding Author:
Olumuyiwa Temitope Ayorinde, PhD
Department of Computing and Informatics
Faculty of Science and Technology
Bournemouth University
Fern Barrow
Poole, BH12 5BB
United Kingdom
Phone: 44 7401803031
Email: ayorindeo@bournemouth.ac.uk

Abstract

Background: Police personnel work under challenging conditions commonly associated with complex shift patterns, unpredictable last-minute changes, and high stress levels, with shift work identified as the major contributor to police personnel health and well-being challenges. These challenges negatively impact their mental well-being, physical health, and job performance, leading to potential health concerns such as fatigue, poor sleep, long-term physical disabilities, anxiety, and poor work-life balance. Existing digital interventions fail to address the needs of shift workers due to focusing solely on conventional 9-to-5 schedules. This gap highlights the need for tailored interventions that incorporate shift management systems into health and well-being applications to support emergency service personnel.

Objective: This study aimed to co-design a shift management system that can be incorporated into a well-being app tailored to address the health and well-being challenges caused by shift work in police personnel.

Methods: This study used interactive management methodology combined with user-centered design and cocreation to facilitate 6 co-design workshops with a diverse group of stakeholders. Each session was structured around idea generation, structural prioritization, and iterative prototyping. User-centered design principles such as persona mapping, scenario walk-throughs, and structured feedback exercises were integrated into the workshop sessions to ensure that the system met diverse user needs. Data were analyzed through participatory feedback and thematic analysis, which allowed for continuous iteration and prioritization of system features based on stakeholder inputs.

Results: Participants highlighted the need for a shift management system capable of managing complex and variable shift schedules with real-time adaptability and support for work-life balance. Thematic analysis revealed shift management challenges such as limited flexibility in accommodating schedule changes and issues managing rotating shift patterns. In response to these identified challenges, a prototype was developed that included features such as bulk creation and modification of shift schedules, shift customization, and visualization tools for monitoring and identifying shift trends and reusable patterns for efficiency. This study demonstrated that integrating a schedule management system into a well-being app could provide personalized support based on users' shift schedules. The integration showed significant potential in supporting the health and well-being of police personnel.

Conclusions: The co-designed shift management system demonstrated strong feasibility and high acceptability among the participants. The integration of shift scheduling into a well-being app can provide tailored support across several domains such as nutrition, hydration, sleep, and physical activity. This combination shows promise in providing a sustainable approach to enhancing the health, well-being, and work-life balance of personnel working in high-stress occupations such as policing.

(JMIR Form Res 2025;9:e69986) doi:10.2196/69986

XSL•FO
RenderX

KEYWORDS

shift work; police personnel; stress; health; police well-being; digital health; shift management

Introduction

Background

Police work is widely recognized as one of the most stressful occupations, characterized by exposure to violence, trauma, death, heavy workloads, and irregular hours [1]. These stressors often lead to high rates of psychological stress, including anxiety and posttraumatic stress disorder [2]. High levels of stress among police officers have been linked to poor sleep quality, fatigue, malnutrition, long-term physical disabilities, sedentary lifestyles, impaired cognitive function, increased risk of cardiovascular disease, poor work-life balance, less family time, and mental health issues [3]. Police officers working afternoon and night shifts reportedly experience more stress due to the physically challenging situations and traumatic events they encounter, with some resorting to frequent alcohol consumption as a coping mechanism, which can lead to substance abuse issues [4,5]. Research has shown that UK police officers experience alcohol misuse at higher rates than their counterparts in continental Europe [5].

Sleep disorders are prevalent among police officers, with 54% of officers reporting issues, particularly those working night shifts [6]. Poor sleep quality and fatigue have been found to impair judgment, increasing the likelihood of mistakes that could endanger the public and the officers themselves [7]. The cumulative impact of these stressors is significant and often leads to lower job satisfaction and a higher likelihood of officers leaving the force among shift workers compared to non–shift workers [8].

Unlike conventional work environments, police officers operate within a highly disciplined structure, requiring officers to adhere to schedules that often involve last-minute changes, extended hours, or operational unpredictability. For these reasons, shift work increases the effect of occupational stressors in policing [1,6,9,10]. Research has suggested that shift workers face greater occupational stress and injury risk and reduced quality of life [11,12]. A well-being survey in partnership with the National Police Wellbeing Service in the United Kingdom highlights key challenges in the police workforce, including high workloads, emotional demands, and declining emotional energy at its lowest since 2020 [8]. These findings highlight the need for digital interventions that can address negative well-being effects of shift scheduling on police personnel. Digital solutions offer advantages such as real-time adaptability, interoperability, scalability, and cost-effectiveness, which makes them suitable for supporting personnel working under unpredictable and dynamic work schedules [13].

Digital technologies can be used to design solutions for health and well-being interventions that increase access to health information, promote healthy behavior, and improve quality of life [14]. Although digital technologies have increasingly been used to support physical, mental, and social health [15], current digital health and well-being applications often fail to meet the needs of police workers. Many existing mobile health (mHealth) apps are designed around standard, traditional 9-to-5 work schedules, making them unadaptable to professions characterized by last-minute changes, extended work hours, and certain schedule unpredictability [2]. Existing solutions often fail to account for the complex relationship among irregular shift schedules, frequent and unpredictable shift changes, and their profound impact on the overall health and well-being of these workers [2]. Buckingham et al [16] highlighted the frustration of police officers toward mHealth apps that do not cater to the realities of their working condition.

Given the unique pressures, high levels of occupational stress [12], sleep disturbances [6], and well-being risks [3,10] associated with shift work, there is a growing need for digital health interventions (DHIs) that not only accommodate operational demands but also promote long-term well-being. While many mHealth apps offer general stress or lifestyle support, few incorporate real-time shift data to personalize their recommendations, representing a critical gap in current digital health and well-being interventions.

This study sought to bridge this gap by cocreating a shift management system designed with input from police personnel that can be integrated into a health and well-being app. This system will support real-time adaptability and last-minute schedule changes and offer personalized and tailored well-being prompts based on police workers' actual schedules. The integration of operational schedules and proactive support can significantly reduce occupational stress, improve mental health and physical well-being, and enhance job performance, ultimately contributing to public safety.

This study advocates for the role of a digital shift management system tailored to police personnel to reduce the health and well-being challenges associated with shift and irregular work schedules. It makes several innovative contributions: first, it highlights the well-being challenges associated with unique police work schedules; second, it demonstrates the application of the interactive management (IM) cocreation methodology in developing a digital shift management system; and, third, it presents a case for integrating a shift scheduling system with a mobile well-being app as a novel strategy to support real-time adaptability and health promotion.

Literature Review

Police work is commonly viewed as a high-risk and stressful occupation associated with several stressors that negatively impact police personnel's physical and mental health [1]. Police stressors can be divided into 3 categories: administrative and organizational pressures, physical and psychological threats, and a lack of internal and external support [12,17]. Administrative and organizational stressors include judicial demands, limited decision-making authority, public criticism, and work-family conflicts [1,17]. Physical and psychological threats arise from exposure to violence, death or injuries, high-speed chases, traumatic incidents, and night shifts [1,12]. Finally, strained relationships with colleagues and supervisors,

XSL•FO
RenderX

inadequate departmental support, and political pressure contribute to the overall lack of support [12,18,19].

Research suggests that administrative stressors are reported more frequently than physical threats [20]. The 5 most common stressors experienced by police officers include managing family disputes, dealing with felonies, understaffing, quick decision-making, and working with underperforming colleagues. Stressors such as using force, witnessing the death of a colleague or a child, or being responsible for killing someone on duty are not common but are highly rated as they tend to cause profound psychological trauma [1,17,20]. Female officers report lack of support as their main stressor [20], whereas their male counterparts identify operational duties as the main source of their stress [6,20,21]. Purba and Demou [22] highlighted the impact of occupational stressors such as lack of support, working long hours, high work demand, and administrative and organizational pressure [23], which have been significantly associated with negative mental health outcomes among police personnel [21].

Occupational stress is increasingly recognized as a significant concern, with well-documented negative effects on employees' health and wellbeing, including a heightened risk of cardiovascular diseases [24]. In the 2023 to 2024 financial year, >14,500 police officers across 43 forces in England, Wales, and Scotland were absent from work due to stress, depression, anxiety, or posttraumatic stress disorder [25]. Police forces in Dorset and Cleveland reported alarmingly high rates of work-related mental health absences, with 343% and 526% increases, respectively [25].

Cartwright and Roach [26] reported that work absence due to mental health issues among officers nearly doubled to approximately 8.82% in the last 10 years [27]. Worryingly, 39% of those who took mental health absent leave subsequently required an additional period of absence, with the primary causes of these absences being stress, trauma, and psychological health issues. Gender differences were observed, with 75% of female officers reporting worsened mental health challenges due to their work compared to their male counterparts [26]. According to the Police Federation of England and Wales [28], 43% of officers reported finding their job extremely stressful, whereas 82% reported having experienced stress, anxiety, or several challenges with their health and well-being. High workload (60%), poor work-life balance (52%), and working in shift patterns (38%) were identified as the top 3 contributors to work-related health and well-being challenges reported by police staff [29].

While general occupational stressors are prevalent, shift work, particularly night shifts, has been identified as the leading contributor to negative physical and mental challenges among emergency services such as the police force [10,12,17]. Policing commonly involves day or night working time occurring in a fixed, rotating, or irregular shift schedule. Police workers with long and irregular shift patterns are at higher risk of burnout and emotional exhaustion than those with fixed shift schedules [10,30].

According to the 2023 National Police Wellbeing Survey, 43.5% of police officers working shifts reported experiencing insufficient sleep compared to 32.7% of non–shift workers. Subsequently, 25.5% of shift-working officers reported experiencing disturbed sleep versus 22.4% of non–shift workers [8]. Similarly, police staff working shifts also reported high rates of disturbed sleep (24.2%) and insufficient sleep (38.8%) compared to their non–shift-working counterparts with 16.4% and 24.8%, respectively [8]. This aligns with the findings of Fekedulegn et al [31], who reported that 54% of police officers experienced poor sleep quality, with a rate of 69% among night shift workers [6,31]. A longitudinal study involving 464 officers found that night shift workers took sick leave at the rate of 4.37 sick days per 10,000 working hours, which is nearly 3 times the sick leave rate of day shift workers, with 1.55 sick days per 10,000 working hours. The risk of absence due to sickness was 182% higher for night shift officers and 129% higher among officers who were overweight [31].

Extended absences of ≥3 days were 65% more common for night shift workers and 50% more common for afternoon shift officers than for their day shift worker counterparts [31]. These statistics show the accumulating physical toll associated with shift work, particularly when combined with occupational stress [4,6,12,17,31]. Operational work culture further amplifies these issues. The Police Federation Pay and Morale Survey 2022 report found that 33% of officers felt pressured to work long hours, and 47% reported working >48 hours per week [29]. Alarmingly, 38% of officers were unable to take their full annual leave entitlement, and 61% had ≥2 rest days canceled [29]. The practice of using annual leave to recover from illness was reported by 33% (physical ill-health) and 42% (psychological ill-health) of workers, whereas 67% admitted to attending work while unwell [29]. According to the 2023 Pay and Morale Survey report, 1 in 5 officers experienced >10 roster changes in a single year, largely due to unpredictable demands [29]. These frequent changes contribute to mental strain and work-life disruption, further compounding their stress.

Key insights from research on shift scheduling has shown that police officers prefer rapidly forward-rotating shift systems with at least 16 hours of rest between shifts despite these schedules offering fewer days off than compressed systems [32]. This preference highlights the importance of shorter consecutive workdays, reduced weekly hours, and adequate rest periods in promoting the health and well-being of officers [32]. These findings challenge claims that shift workers prefer compressed systems with more days off [33,34], highlighting the diversity of individual preferences. Age, current work hours, and previous experiences with shift schedules significantly influence these preferences [34]. Compared with officers with regular shift schedules, those working shifts experience high levels of fatigue, work overload, and reduced task control, resulting in lower job satisfaction and a higher likelihood of leaving the force [8,17].

These widely varied preferences highlight the need for tailored development of digital shift management systems that can address the mental health challenges associated with shift work, offering more personalized and effective solutions [32,35]. Despite the growing use of DHIs to address health and lifestyle challenges, many digital interventions are not tailored to the needs of shift workers [36]. Due to the continually changing shift pattern in policing, the effectiveness of these digital

XSL•FO
RenderX

interventions is often limited in mitigating work-related stress, depression, and anxiety among emergency service personnel [37].

A study exploring the views of UK police personnel identified a significant shortcoming in existing digital health and well-being interventions, which is their lack of personalization for shift-based working patterns. These interventions are often misaligned with the operational realities of police work, particularly due to their limited adaptability to variable and irregular shift schedules [3]. For example, the Headspace mindfulness app [37] provides mindfulness content but does not adapt it to users' shift patterns. Similarly, the Sleepfit app [38], which focuses on sleep health, lacks real-time integration with varying shift schedules. The Healthier Outcomes at Work app [39] addresses workplace stress management but fails to consider the unique demands of shift work. The Healthy Minds Innovation app [40] provides mindfulness and meditation support but does not integrate the intervention with the users' work schedules. A similar scenario occurs with the #SWPMoveMore Challenge intervention [41], which encourages physical activity but fails to incorporate shift schedules into its data tracking and recommendations. Although these interventions target different aspects of health and well-being, their lack of shift-specific personalization limits their effectiveness for the target population. This represents a key limitation for populations such as police officers and emergency workers with unpredictable and irregular work patterns.

Research on improving shift scheduling and fatigue management among police workers highlights several recommendations to ensure health, safety, and performance [42], such as involving employees in scheduling decisions, maintaining regular work hours, and ensuring that any shift changes rotate forward rather than backward [11,42]. These recommendations aim to reduce the adverse effects of shift work on both the physical and mental well-being of police workers. By implementing these recommendations, police workers and other workers in similar high-stress environments can benefit from effective and sustainable shift scheduling systems that enhance their overall well-being and performance [12,30,42].

mHealth apps can play a crucial in addressing mental health and stress-related concerns among emergency service personnel [43]. In a public engagement study conducted by the National Health Service England, the stakeholders emphasized that effective DHIs must be engaging, accessible, and tailored to users' needs, with features such as gamification, rewards, personalization, and responsive feedback channels identified as crucial for sustainable user engagement and satisfaction [43,44]. Despite the increasing use of health and well-being apps, many digital interventions fail to address the unique needs of shift-working officers. Mental health issues such as stress, depression, and anxiety remain prevalent due to existing apps often suffering from low adherence, insufficient customization for irregular schedules, and weak user engagement due to failure in accommodating fluctuating work patterns [45,46].

Marston et al [47] proposed a collaborative approach for developing DHIs specifically for emergency services by bringing together researchers, emergency staff, stakeholders, designers, and developers. They recommended not only establishing a framework for developing and implementing DHIs but also ensuring interoperability and consistent use of these apps across different emergency service organizations [47,48]. There is evidence suggesting that personalized, on-demand digital tools frequently outperform physical therapy sessions in managing work-related stress [49], yet many current DHIs lack the behavioral tailoring to users' routines, adaptative feedback responses to real-time user data, and shift-specific features such as rota or schedule changes necessary to meet the evolving needs of emergency personnel.

Kerr et al [50] identified key concerns among employees using DHIs for managing stress—health impact, privacy, autonomy, identity, and accountability could further hinder the adoption of DHIs [3,48]. While most of the employees in the aforementioned study were open to using digital stress management tools, they strongly preferred options that integrated seamlessly into their daily routines and prioritized data privacy. Weerasekara and Smedberg [2] found that users preferred stress management interventions that involved self-help modules, personalized activities, and peer support [51]. Howe et al [49] suggested developing digital interventions that reduce work-related stress, such as by integrating a variety of stress relief content, including personalization for users, and introducing stress-reducing strategies such as longer break periods [49]. Tailoring these interventions to individuals' experiences, needs, and preferences was identified as extremely essential for engagement. Building on these preferences, participants in previous studies have suggested integrating peer support; expert guidance; and gamification features such as social recognition, badges, mini games, and group challenges to boost user adherence and engagement [44,52]. Finally, participants have emphasized that privacy, security, and confidentiality are essential requirements for digital interventions [3,50,53].

To bridge these gaps, future digital health and well-being interventions for shift workers should be developed using inclusive and user-centered approaches that involve diverse end users [44]. They should also integrate adaptive shift management features that help users plan their activities, nutrition, exercise, sleep, and recovery around irregular schedules, thereby promoting a sustainable work-life balance [32,42]. Adherence to ethical design principles; transparent data privacy practices; and maintenance of active partnerships among emergency services, researchers, and developers will allow for iterative refinement and robust evaluation of developed digital health and well-being interventions [54]. The next step is to organize cocreation workshops with serving police staff and other important stakeholders to gather detailed, shift-specific requirements, thereby ensuring the development of future DHIs reflect the real-world demands of shift work.

XSL•FO
RenderX

Methods

Research Design

Overview

This study used a qualitative research approach using a series of IM workshops to gather insights on the scheduling needs and preferences of police workers. This iterative, participatory method enabled in-depth exploration of the participants' experiences with existing shift management systems and how they affect their well-being. This approach enabled the researchers to probe participants for responses, encourage collaborative discussion, and uncover the challenges associated with shift-based police work. Thematic analysis (TA), nominal technique, and interpretive structural modeling were used to analyze and structure the data collected.

Population and Sample

Participants were selected using a combination of purposeful and convenience sampling. Individuals directly involved in shift pattern coordination, operational management, frontline policing, and digital system design were targeted. This approach ensured that the perspectives of those most involved in implementing shift scheduling systems were represented. The participants were affiliated with a regional police force in the southern region of the United Kingdom. The sample comprised a chief superintendent of police, a police operations manager, an expert in human-computer interaction (HCI), a digital health expert, an academic researcher, and a shift management administrator who manages existing shift scheduling systems. This group represented a cross-section of technical, operational, and strategic expertise necessary for effective system co-design.

Purposive sampling ensured the inclusion of participants with specialist knowledge, whereas convenience sampling accommodated practical constraints related to availability [55,56]. Although broader representation was limited by logistical and ethical constraints, the selected participants were well suited to the aims of this study, representing a variety of roles; levels of responsibility; and those who could provide meaningful, experience-based feedback on system requirements and implementation feasibility. All participants were invited to join interactive co-design workshops depending on availability and relevance to specific development stages. Meetings were held in person or via web-based platforms (eg, via Microsoft Teams) depending on participant availability and logistical considerations.

Instruments

The primary instrument used for data collection was a semistructured discussion guide iteratively developed to support each of the 6 co-design workshops conducted. Early sessions focused on requirement gathering, terminology refinement, and system expectations, whereas later sessions addressed feature testing; interface usability; and refinement of core components such as bulk shift creation, interactive visualization, and preconfigured shift patterns (eg, gold and silver shift pattern template).

Five of the 6 meetings were conducted virtually via Microsoft Teams, recorded using the platform's built-in functionality, and transcribed using Microsoft Teams' automatic transcription feature. The session with the shift schedule meeting (meeting 3) was held in person and was recorded using a mobile recording device and manually transcribed by the research team.

The discussion guide was iteratively refined to reflect the evolving objectives of the project. It enabled (1) open-ended exploration of contextual challenges such as inflexible scheduling and well-being concerns; (2) targeted feedback on design elements, including shift visualization tools, terminology, and shift rota flexibility; and (3) prioritization of feature requirements through structured prompts and facilitated consensus-building exercises.

Trustworthiness

To ensure the trustworthiness of the qualitative process and findings [55], several strategies were used: (1) pilot-testing of the IM method with internal stakeholders to refine the guide and facilitation approach; (2) member checking during workshops, including real-time clarification and postsession return of professionally transcribed data to key participants for validation; and (3) triangulation of participant feedback across multiple sessions and user groups to identify consistent patterns and diverse perspectives.

Structure of the Co-Design Process

This study adopted a cocreation and user-centered design (UCD) approach structured around six 60-minute IM-facilitated workshops. The workshops integrated UCD principles such as persona mapping, scenario walk-throughs, and structured feedback exercises.

Meetings 1, 2, 4, 5, and 6 involved a core group of stakeholders, including a chief superintendent of police, a police operations manager, an expert in HCI, a digital health expert, and an academic researcher. These sessions focused on initial scoping, terminology development, interface functionality, technical feasibility, contextual constraints, user workflow, feature walk-throughs, usability feedback, and validation of the system.

Meeting 3 involved the operational shift scheduling unit responsible for coordinating shift patterns across departments. This involved demonstration of how the existing police scheduling system operates and provided a deep dive into management of complex shift patterns, the combination of multiple shift types, and the prepopulation of base shift pattern templates (eg, gold or silver shift patterns) across different policing departments.

Meeting 6 served as a final validation session focusing on implementation readiness and final design insights. This was followed by nominal group techniques and interpretive structural modeling. This process supported the prioritization of user-identified needs and clarified dependencies between proposed features.

XSL•FO
RenderX

IM Session

Overview

IM is a methodology that facilitates collaborative decision-making to tackle complex problems [57]. IM promotes communication and shared understanding among groups of stakeholders or participants to develop a deep understanding of a problem, explore meaningful solutions, and collectively agree on a defined course of action to solve the problem [58]. Unlike traditional focus groups, IM involves the identification of domain experts, consensus building, iterative idea development, and prioritization of the solution [59]. It is especially effective in solving problems associated with high complexity, diverse stakeholder viewpoints, and evolving constraints, making it well suited for addressing the multifaceted challenges surrounding police shift management and system design.

The phases of IM are outlined in the following sections.

Planning Phase

The planning stage involved collaborative work with key stakeholders to develop a shared understanding of the current system and the problem space. Activities included defining the scope, identifying affected actors, and cocreating guiding questions for the workshops. This ensured that workshop content addressed real-world operational challenges such as inflexible scheduling, system usability, and well-being implications [58].

Workshop Phase

The workshop sessions were structured around the IM framework's 3 core dimensions: context, content, and process. Each session began with a contextual overview provided by the facilitator to ground participants in the session's objectives and revisit insights from previous sessions [58]. This approach ensured continuity and a cumulative design process. Participants were then guided through idea writing exercises, each triggered by carefully framed, open-ended questions relevant to the project's aims.

Idea Writing

Participants exchanged ideas, refined them through discussion, and engaged in nominal group technique sessions. Participants collaboratively edited and clarified problem statements and then assigned priority ratings to the ideas, helping the group identify the most critical needs. The top-rated ideas were transformed into design objectives and used to construct interpretive structural models [59]. This modeling process helped identify interdependencies between ideas, clarifying design priorities and informing system architecture decisions.

Follow-Up Phase

Insights generated during each session were synthesized into actionable requirements and integrated into subsequent design iterations. Although no separate follow-up interviews were conducted, participant feedback was captured in the sessions and validated through ongoing collaborative engagement [58]. When new challenges emerged, such as departmental shift variation or ambiguous terminology, they were integrated into planning for the next workshop, maintaining the recursive and adaptive nature of the IM methodology.

Cocreation Process

Cocreation has been linked to several methods, such as co-design and coproduction, as well as design approaches such as design thinking. Cocreation is suggested to be also known as participatory design [60] and has been used to develop several digital health care interventions. Early involvement of users and other major key stakeholders in the design, development, and evaluation of digital health solutions has proven to be an important factor in the success of these interventions [61]. Interventions developed using cocreation have been reported to result in improved health outcomes [62]; improved mental health [63]; enhanced quality of life [64]; and enhanced overall well-being consisting of social, psychological, and physical aspects [65]. Research has shown that the potential actors involved in a cocreation process are the end users, stakeholders (mostly referred to as the nonacademic stakeholders), and academic researchers. The end users are the population that the intervention is targeted to, and the stakeholders are groups that are concerned with the implementation of the interventions, whereas the academic researchers conduct the research [65]. Involving end users in the cocreation of interventions has been reported to increase satisfaction, use, and effectiveness of the intervention due to the end user participation in creating solutions targeted to their problems and needs [66,67].

UCD Approach

UCD is an approach for developing interventions that involves and prioritizes the end users in every stage of the solution development phase. The involvement of users in every step of the development phase ensures high quality and user acceptance of the developed interventions [68]. UCD involves the iterative process of understanding the user needs and requirements through research, designing a prototype to obtain user feedback, refining the design based on user feedback, testing and evaluating the design, and monitoring the intervention after launching [69]. Recently, UCD has been used to develop health-related interventions targeted at different health issues, such as pain management, health and well-being, and mental health [70].

This research combines innovative and design approaches in developing the shift management system (Figure 1 [71]) by exploring UCD techniques in the design phase of the cocreation process. This cocreated intervention was collaboratively developed and targeted at police workers to address stress arising from shift work.

XSL•FO
RenderX

Figure 1. Cocreation process with embedded user-centered design principles.

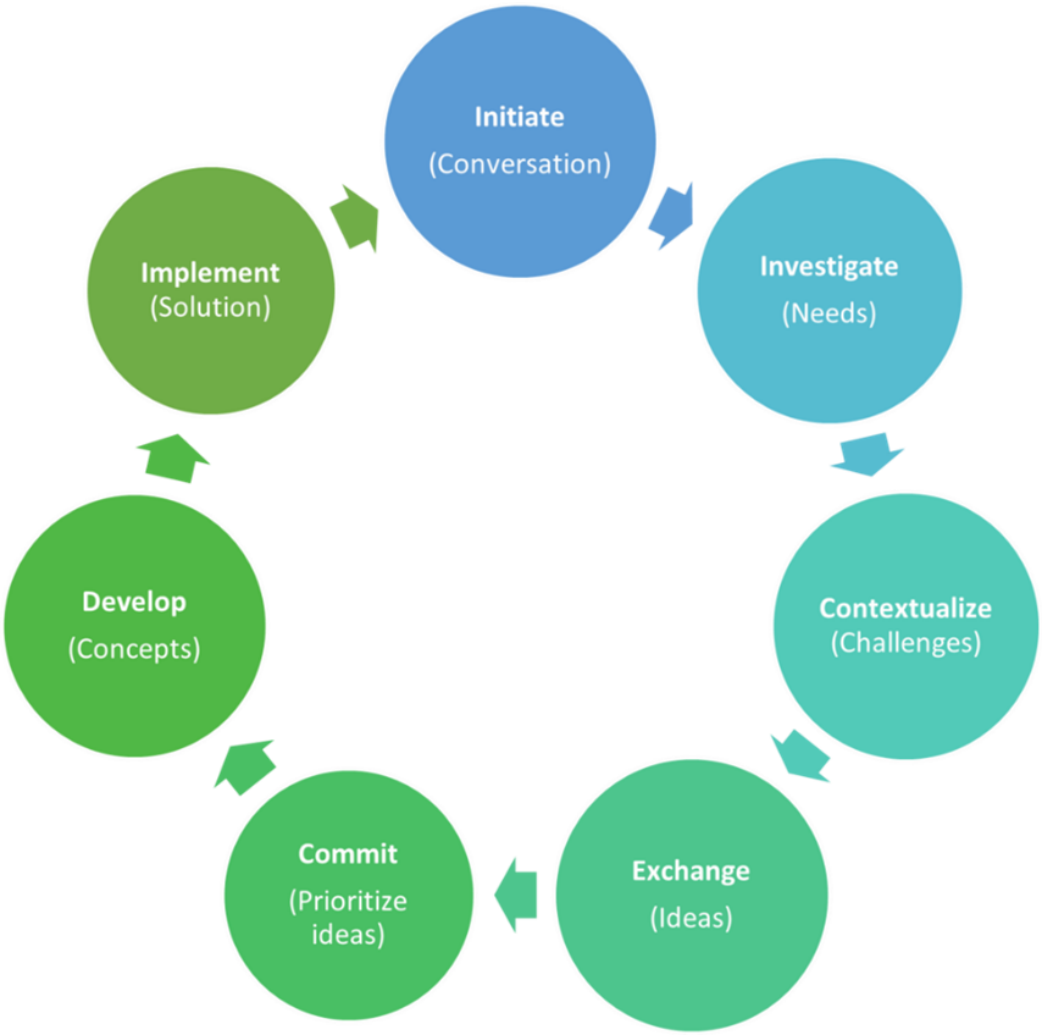

TA Process

TA was used to identify, analyze, and interpret patterns known as themes in qualitative data. TA uses data collected from participants' perspectives, views, behavior, and experiences to identify patterns or themes regarding how the participants feel and think [72]. TA can be used to analyze datasets involving few participants and datasets from a large number of participants [73]. TA has proven to be very efficient in analyzing a wide range of data collected through qualitative techniques such as workshops, focus groups, observations, interviews, and story completion [72,74].

Before conducting the coding process for the TA, the meeting minutes were transcribed to ensure an accurate representation of the agreed discussions. After transcription, keywords were identified from the data, and codes were developed from these keywords, which were subsequently grouped into themes. To ensure a thorough understanding of the qualitative data, the transcribed documents were read multiple times, allowing the research team to gain familiarity with the content and context of the discussions [74]. This iterative reading process helped refine the coding process and ensured that the themes accurately reflected the core insights from the cocreation sessions (see Multimedia Appendix 1 for more details).

The process led to the identification of 57 codes in the initial coding process, which were further refined to 15 unique codes after merging similar codes to remove duplications. These codes were derived from structured trigger questions posed during the cocreation sessions (Table 1).

The identified codes captured trends and patterns in the data, and after thorough synthesis, 5 overarching themes were identified (Figure 2) [70].

XSL•FO
RenderX

Table 1. Trigger questions and codes for creating themes from the cocreation sessions.

Trigger question	Codes for creating themes from participant responses
What are the main factors contributing to stress and well-being challenges caused by shift work and shift management?	Lack of shift flexibility to reduce stress, shift work–related stress, lack of support for work-life balance, issues managing rotating shifts, inconsistent shift scheduling, inadequate and inflexible shift pattern creation system for large teams (eg, 2000 users), disruption of personal life due to shift work, and varying shift start and finish times.
How should the system be designed to accommodate diverse user needs while ensuring ease of use?	Develop several detailed personas representing different user roles with unique schedules. These personas are used to guide design decisions to ensure that the functionalities meet the tailored needs of the user group. One proposed approach involves using the cyclelike approach of creating and managing shift patterns while ensuring easy navigation for users. The interface should be interactive and responsive while also reducing the number of steps required to complete a task.
Which key features are essential for efficient and effective shift schedule management?	Bulk creation of shift schedules; flexible modification of schedules; shift type monitoring through visualization; ease of creation and management of complex shift patterns; ability to merge rotating schedules; identification of shift trends and distribution through visualization; and simplified interactive process of creating, editing, and managing shift patterns.
What enhancements can be made to improve the flexibility and customization of shift patterns to better meet scheduling needs?	Daily shift time customization to accommodate varying personal and professional obligations, manual editing of individual days within a shift pattern for quick adjustments, creation of custom shift patterns with multiple finishing times within a single pattern, creation of shift patterns from existing base patterns, setup of recurring shift patterns with flexible cycles for users with regular nonidentical weekly or monthly schedules, and notification to alert users of upcoming shifts.
How can consistency and appropriate terminology enhance the understanding of the shift scheduling system?	Replacing shift terms with more understandable and familiar vocabulary, implementing consistent terminology to reduce confusion and improve communication, ensuring that shift patterns begin on a Monday to align with the traditional workweek, and use of the 24-h time format across the entire system to enhance clarity.

Figure 2. Thematic map of core themes from the interactive management–facilitated workshops.

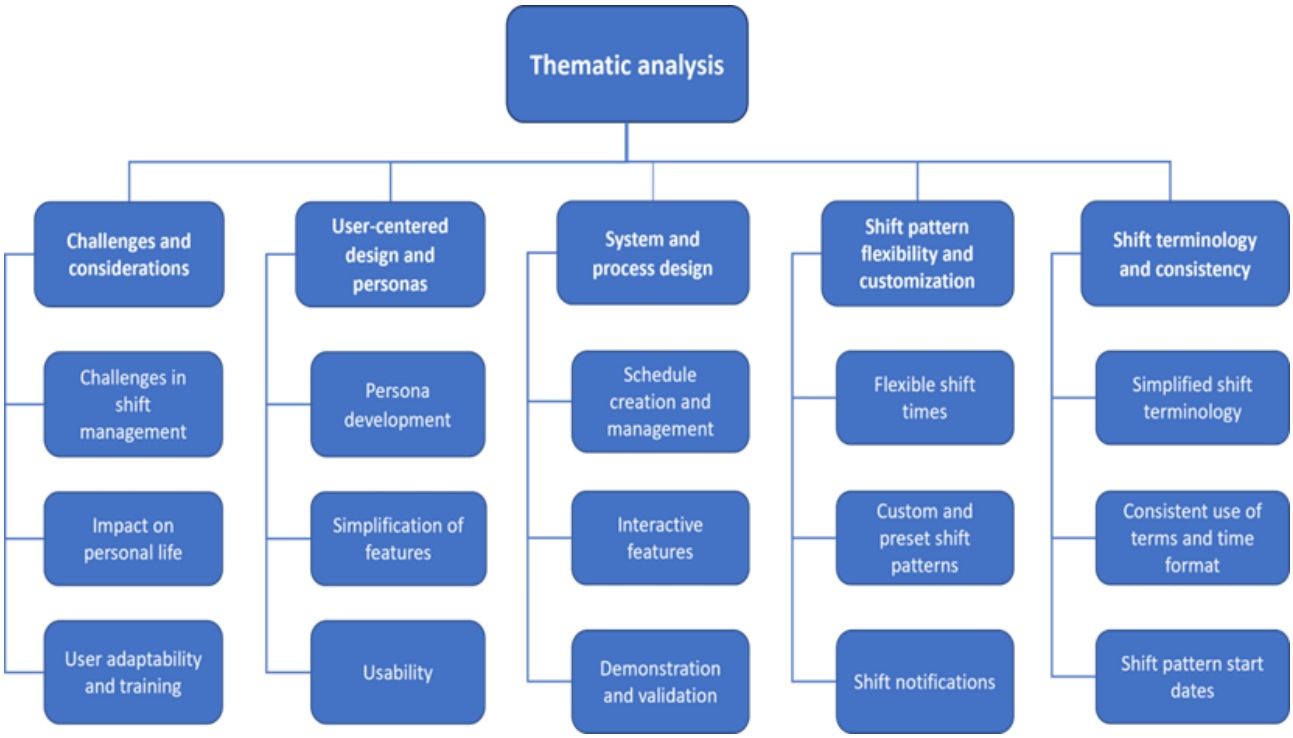

Ethical Considerations

This study received ethics approval from Bournemouth University Research Ethics Committee (approval 58206). All participants, including those authors who took part in the co-design sessions, were provided with a participant information sheet. This document outlined the study's purpose, procedures, data handling practices, and confidentiality provisions and the participants' rights, including voluntary participation and the ability to withdraw at any time without consequence. Informed consent was obtained in writing from all participants after they were provided with a detailed participant information sheet. Following the ethical standards of the university, all participant data were anonymized during transcription and securely stored on password-protected and encrypted devices. Anonymized data may be made available for future ethically approved research in compliance with the ethical agreement, including use in presentations, reports, or publications.

XSL•FO
RenderX

Results

What Are the Main Factors Contributing to Stress and Well-Being Challenges Caused by Shift Work and Shift Management?

Participants were asked about the main factors contributing to stress and well-being challenges caused by shift work and shift management. The findings revealed several key stressors: a lack of shift flexibility to reduce stress, stress directly related to shift work, insufficient support for achieving a work-life balance, difficulties in managing rotating shifts, and inconsistent shift scheduling. In addition, participants highlighted the inadequacy and inflexibility of shift pattern creation systems, particularly for large teams (eg, 2000 users), along with the disruption of personal life due to shift work and varying shift start and finish times. These factors collectively contribute to stress and well-being challenges among shift workers (Table 2). A few quotes from the sessions include the following (*M* indicates "meeting" and *P* indicates "participant"):

> *Rotating shifts are difficult to manage, especially when we have to combine multiple patterns.* [M3P1]
>
> *Poorly designed shift patterns harm personal life and work performance.* [M1P5]
>
> *Flexible or predictable scheduling system would help maintain a better work-life balance.* [M1P4]
>
> *Struggle with the complexity of software made for large teams, especially when everyone has different schedules.* [M1P2]

Table 2. Identified stressors associated with shift work among police personnel.

Theme	Description	Quotes	Percentage
Scheduling constraints and fatigue risks	Lack of adequate rest and recovery time due to an irregular shift structure	• "Current system has limited flexibility and inconsistent scheduling." [M1P3] • "I would appreciate reminders to take breaks when I've been doing back-to-back shifts." [M5P5] • "The system should suggest breaks when we've worked several heavy shifts." [M6P5] • "The system should be able to warn me if I'm scheduled to work too many night shifts in a row." [M6P4]	30.8
Insufficient support for work-life balance	Disruption of personal routines, relationships, and mental recovery caused by unpredictable shift schedules	• "Flexible or predictable scheduling system would help maintain a better work-life balance." [M1P4] • "Poorly designed shift patterns harm personal life and work performance." [M1P5]	15.4
Difficulty managing rotating shifts	Challenges in managing and combining multiple rotating shift patterns	• "Rotating shifts are difficult to manage, especially when we have to combine multiple patterns." [M3P1] • "We should simplify things by using cycles instead of overly complex shift patterns." [M4P1]	15.4
Incompatibility with real-world shift demands	Scheduling systems fail to accommodate diverse shift end times or departmental needs	• "We have officers with varying finish times, one pattern doesn't fit all." [M4P3] • "It's hard to understand systems designed to manage shifts for thousands of staff with varying start and end times." [M1P1]	15.4
Lack of clarity and an ineffective scheduling support tool	Stress caused by unclear guidance and a lack of user-friendly scheduling tools	• "There should be a simple way to describe different shift options and how we use them." [M6P1] • "The system should give me clear instructions, like how to start a pattern on a certain day and repeat it." [M6P2] • "Users will benefit from being able to see trends that show recurring scheduling issues." [M4P5]	23.1

How Should the System Be Designed to Accommodate Diverse User Needs While Ensuring Ease of Use?

Participants were asked how the system should be designed to accommodate diverse user needs while ensuring ease of use. The findings suggest that the system should incorporate several detailed personas representing different user roles with unique schedules. These personas are intended to guide design decisions, ensuring that the system's functionalities meet the tailored needs of various user groups. In addition, the design should follow a cyclelike approach for creating and managing shift patterns, ensuring easy navigation for users. The interface should also be interactive and responsive and minimize the number of steps required to complete tasks, facilitating a more efficient and user-friendly experience. A few quotes from the sessions include the following:

> *I think it would help to have a model or table that links user personas with shift patterns.* [M2P1]
>
> *We should create 5 to 10 fictional personas based on demographic data and real shift schedules.* [M2P5]
>
> *I would like to see a list of example characters with typical shift patterns to guide system use.* [M2P3]

XSL•FO
RenderX

We should simplify things by using cycles instead of overly complex shift patterns. [M4P1]

Personas have been reported to improve the identification of user needs and enhance quicker solution development [75]. Personas were used during the design phase to ensure that the proposed digital intervention met the needs of the diverse user groups (Figure 3).

Figure 3. Sample persona showing stakeholders' needs, motivations, and frustrations.

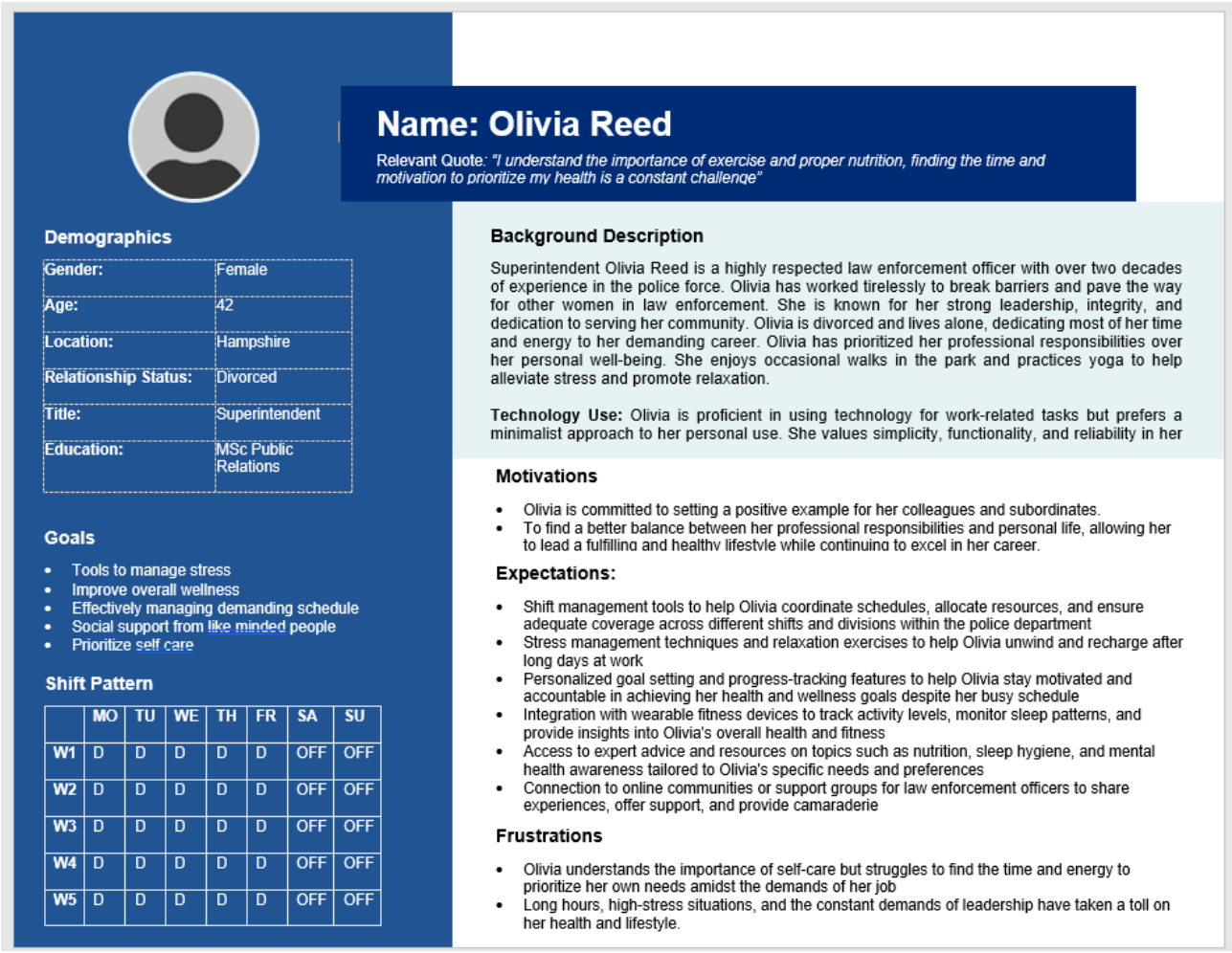

Which Key Features Are Essential for Efficient and Effective Shift Schedule Management?

Participants were asked to identify the key features essential for efficient and effective shift schedule management. The findings highlight several critical features, including the ability to create shift schedules in bulk and modify them flexibly. Monitoring shift types through visualization, simplifying the creation and management of complex shift patterns, and merging rotating schedules were also noted as important. In addition, the ability to identify shift trends and distributions via visualization, along with a simplified and interactive process for creating, editing, and managing shift patterns, were deemed essential to optimizing shift management. A few quotes from the sessions include the following:

Users should be able to set different shift times for different days within the same pattern. [M4P4]

I want to input my entire schedule in one go like a bulk pattern, not break it into separate parts. [M5P2]

It would be nice to merge two patterns together and use them as one ongoing schedule. [M5P2]

Interactive visual tools would make it easier for me to track progress over time. [M4P1]

The key features in Table 3 were derived using nominal group techniques (Table 4). Five participants independently ranked each proposed feature on a scale of 1 (least important) to 5 (most important). The scores were then summed to determine relative importance.

XSL•FO
RenderX

Table 3. Key ideas generated during the cocreation sessions.

Idea number	Idea description
1	Bulk creation of shift schedules
2	Flexible modification of shift schedules
3	Shift type monitoring through visualization
4	Ability to merge rotating schedules
5	Daily shift time customization to accommodate varying personal and professional obligations
6	Possibility for shift patterns to contain a combination of several shift types
7	Manual editing of individual days within a shift pattern for quick adjustments
8	Creation and management of complex shift patterns
9	Setup of recurring shift patterns with flexible cycles
10	Notification to alert users of upcoming shifts
11	Replacing shift terms with more understandable and familiar vocabulary
12	Ensuring that the shift pattern begins on a Monday to align with the traditional workweek
13	Use of the 24-h time format across the entire system to enhance clarity
14	Creation of a schedule with multiple finishing times within a single shift pattern
15	Prepopulation of the base shift pattern
16	Ability of the created shift pattern to be reused and modified to create a new shift pattern

Table 4. Participant prioritization scores for user-generated ideas.

Idea number	Participant 1 score (1-5)	Participant 2 score (1-5)	Participant 3 score (1-5)	Participant 4 score (1-5)	Participant 5 score (1-5)	Total score
1	4	4	4	2	4	18
2	3	3	2	3	4	15
3	2	2	2	3	3	12
4	3	2	2	3	3	13
5	4	3	2	3	4	16
6	5	4	3	2	4	18
7	4	5	2	2	4	17
8	4	1	3	4	3	15
9	4	3	3	3	4	17
10	5	1	1	5	2	14
11	4	5	1	1	1	12
12	1	5	3	2	2	13
13	1	5	1	3	2	12
14	3	2	1	4	2	12
15	3	3	2	4	3	15
16	3	4	4	2	4	17

What Enhancements Can Be Made to Improve the Flexibility and Customization of Shift Patterns to Better Meet Scheduling Needs?

Participants were asked about the potential enhancements that can improve flexibility and customization of shift patterns to better meet scheduling needs. The findings indicated that daily shift time customization, allowing for adjustments to accommodate varying personal and professional obligations, is essential. Manual editing of individual days within a shift pattern for quick changes, creation of custom shift patterns with multiple finishing times within a single pattern, and building shift patterns from existing shift patterns were highlighted (Figure 4). In addition, setting up recurring shift patterns with flexible cycles for users with nonidentical weekly or monthly schedules and providing notifications to alert users of upcoming shifts were

XSL•FO
RenderX

suggested as key improvements. A few quotes from the sessions include the following:

> *I want to be able to edit days in shift pattern, including rest days and specific shift types.* [M4P5]
>
> *Sometimes there is need to manually change just one shift within a recurring cycle.* [M5P4]
>
> *We have officers with varying finish times, one pattern doesn't fit all. Update the design to accommodate flexible shift time.* [M4P3]
>
> *It would help if the system recommended finish times for late and night shifts based on past inputs which can help with notification.* [M4P2]
>
> *It's hard to understand systems designed to manage shifts for thousands of staff with varying start and end times.* [M1P1]

Figure 4. Customization and flexibility requirements in shift scheduling.

How Can Consistency and Appropriate Terminology Enhance the Understanding of the Shift Scheduling System?

Participants indicated that replacing shift terms with more understandable and familiar vocabulary, as well as implementing consistent terminology throughout the system, significantly enhanced their ability to understand the shift management process. By reducing confusion and improving communication, participants felt more confident navigating the system. They also emphasized the importance of aligning shift patterns with the traditional workweek (starting on a Monday) and using a 24-hour time format, which further increased clarity and made it easier for them to manage their shifts effectively. Figure 5 shows how these changes can contribute to smoother operations and better communication across the system. A few quotes from the sessions include the following:

> *The system should use 24-hour time for consistency.* [M2P5]
>
> *We [police] usually begin our patterns on a Monday, so that should be the default.* [M4P3]
>
> *I would prefer more accurate labels instead of terms like "early," they're too vague, modify the term.* [M2P1]

XSL•FO
RenderX

Figure 5. System understanding through terminology and time format consistency.

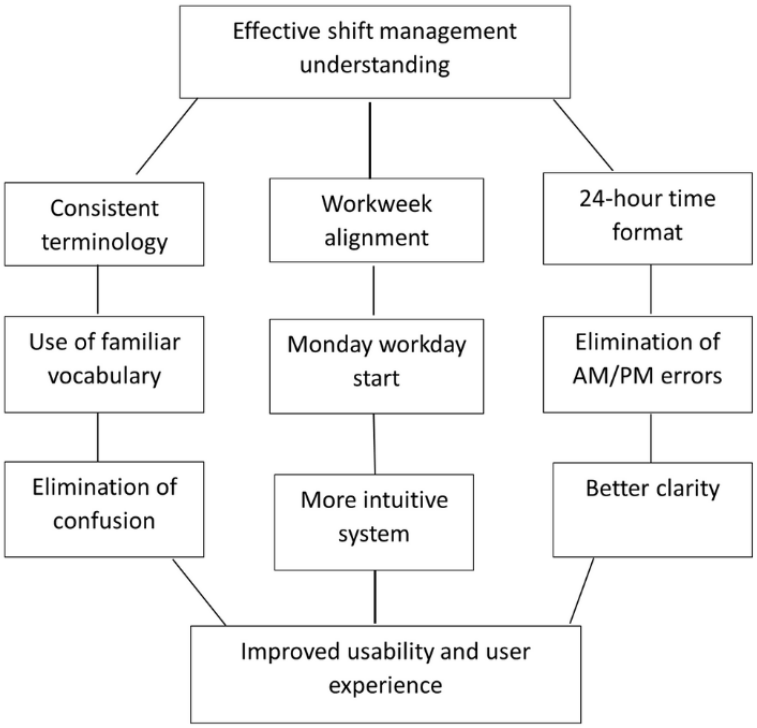

Results of the IM Workshop Session

The IM co-design workshops were conducted virtually via Microsoft Teams to promote participants' convenience and were recorded for transcription and analysis. To facilitate idea generation, participants were guided using trigger questions specifically focused on improving the flexibility, usability, and effectiveness of shift scheduling systems to promote well-being support for police personnel. Participants' ideas were recorded individually during the idea writing phase, and ideas were subsequently refined and discussed as a group and clarified collectively. One of the workshops was conducted on-site with members of the operational shift scheduling unit for policing departments.

Prioritization of ideas followed through nominal group techniques, allowing for categorization and construction of interpretive structural modeling.

The following trigger questions were used to prompt idea generation: (1) what enhancements can be made to improve the flexibility and customization of shift patterns? (2) How can shift management tools better accommodate personal and operational scheduling needs? (3) What features would improve communication and understanding of shift schedules?

Table 3 shows the ideas and participants' responses generated using the trigger questions listed in Table 1. Table 5 shows that participants highlighted effective and flexible shift scheduling as an important requirement for a system capable of supporting the well-being of police personnel. The IM sessions also identified customization of schedules, shift monitoring, and schedule visualization as being crucial to the shift management system.

Table 5. Categorization of user-generated ideas into thematic groups.

Category	Ideas
Effective and flexible shift scheduling	1, 2, 7, 8, and 14
Customization of shift schedule	4, 5, 6, and 9
Shift monitoring and visualization	3 and 10
Consistency and terminology	11, 12, and 13
Pattern reusability	15 and 16

The nominal technique was used to allow participants to prioritize the top 5 requirements generated from the ideation phase, as shown in Table 5. The participants ranked the ideas from 1 (most important) to 5 (least important). This revealed that enabling bulk creation of shift schedules, creating reusable shift patterns, combining several shift types into a pattern, setting up recurring shift patterns with flexible cycles, and allowing for manual editing of daily shifts were ranked the highest among the participants.

According to Table 4, ideas 1, 6, 7, 9, and 16 were ranked as the top 5 requirements based on the total score.

XSL•FO
RenderX

On the basis of the prioritised ideas presented in Table 4, objective statements were formulated, that can be used to develop defined system requirements. These statements are outlined in the section below and were used to construct an interpretive structural model illustrated in Figure 6. The figure shows the categorization of similar objective statements into boxes, and the logical relationships between them.

Figure 6. Interpretive structural model of prioritized system features.

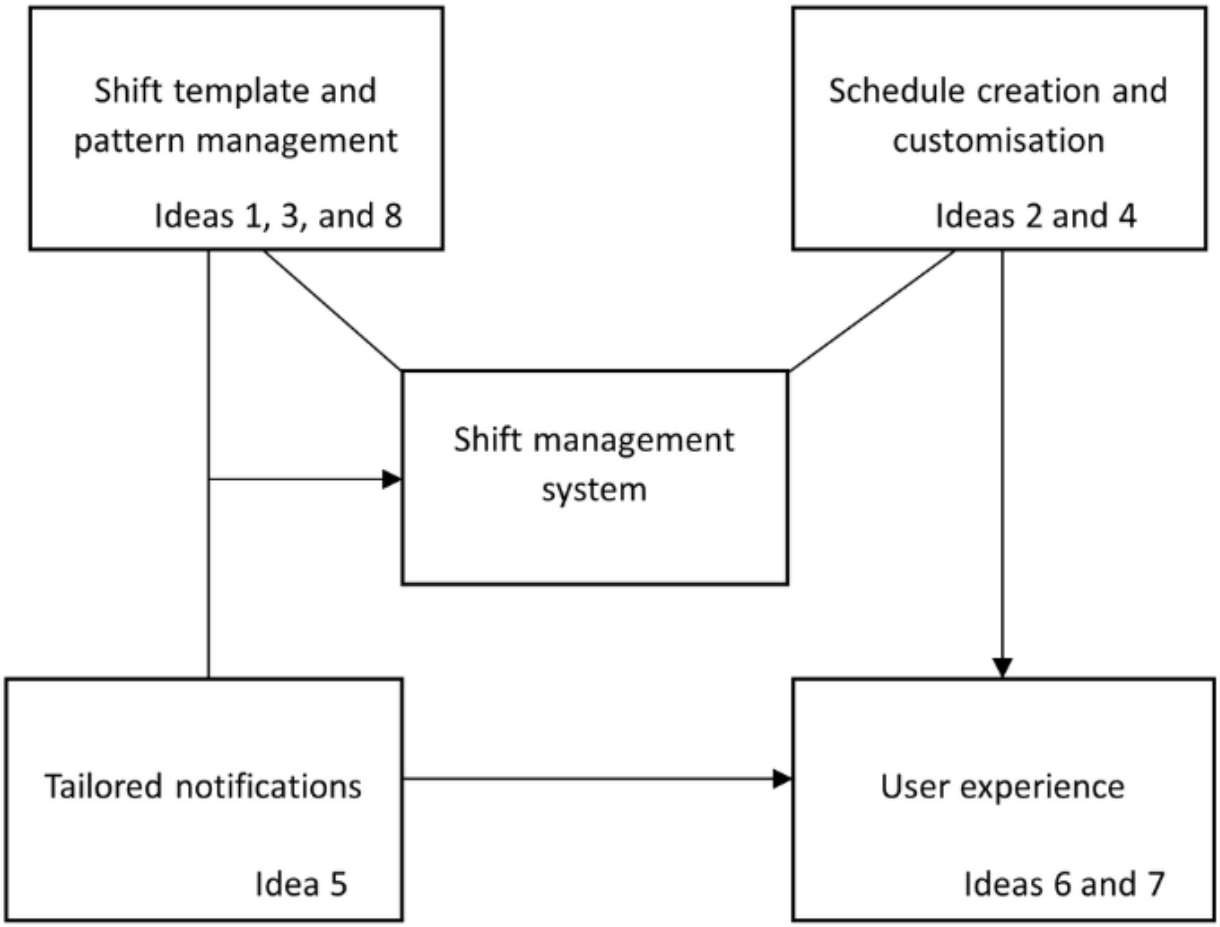

Interpretive Structural Modeling

Following prioritization, the top ideas were transformed into objective statements suitable for interpretive structural modeling (Figure 6). These statements represented the system requirements necessary to address user-identified needs: (1) enable reusable shift pattern templates with recurring cycle support to reduce manual effort and accommodate routine schedules; (2) allow for the bulk creation and deployment of shift schedules for efficiency across teams or departments; (3) support complex shift patterns by enabling the combination of multiple shift types within a single cycle; (4) allow for customization and manual editing of individual shift days to accommodate personal needs and real-time changes; (5) integrate notification features to remind users of upcoming shifts or updates; (6) use intuitive, standardized terminology to enhance user understanding and minimize confusion; (7) incorporate visual tools to monitor shift distribution, balance workloads, and identify scheduling issues; and (8) allow for the merging of rotating schedules to enable flexible, continuous coverage models.

Prototype Design

First Prototype Iteration

The prototype of the shift management system revealed several limitations that showed its ineffectiveness. One of the major issues was the lack of proper visualization, which made it difficult for users to identify trends and patterns in the various types of shifts (Figure 7A). This absence of clear, visual representation made it difficult to understand workload distribution and detect when a user is likely to be stressed from shift workload. In addition, the prototype did not account for varying finishing times of shifts, which is crucial for accommodating the diverse personal and professional schedules of police workers. Another significant shortcoming was the inability to merge multiple rotating shift pattern (Figure 7B), an important feature for efficiently managing schedules with complex, overlapping work times. The process of adding new shifts also involved too many steps, making the system time-consuming to use (Figure 7C).

XSL•FO
RenderX

Figure 7. First iteration of the shift management system.

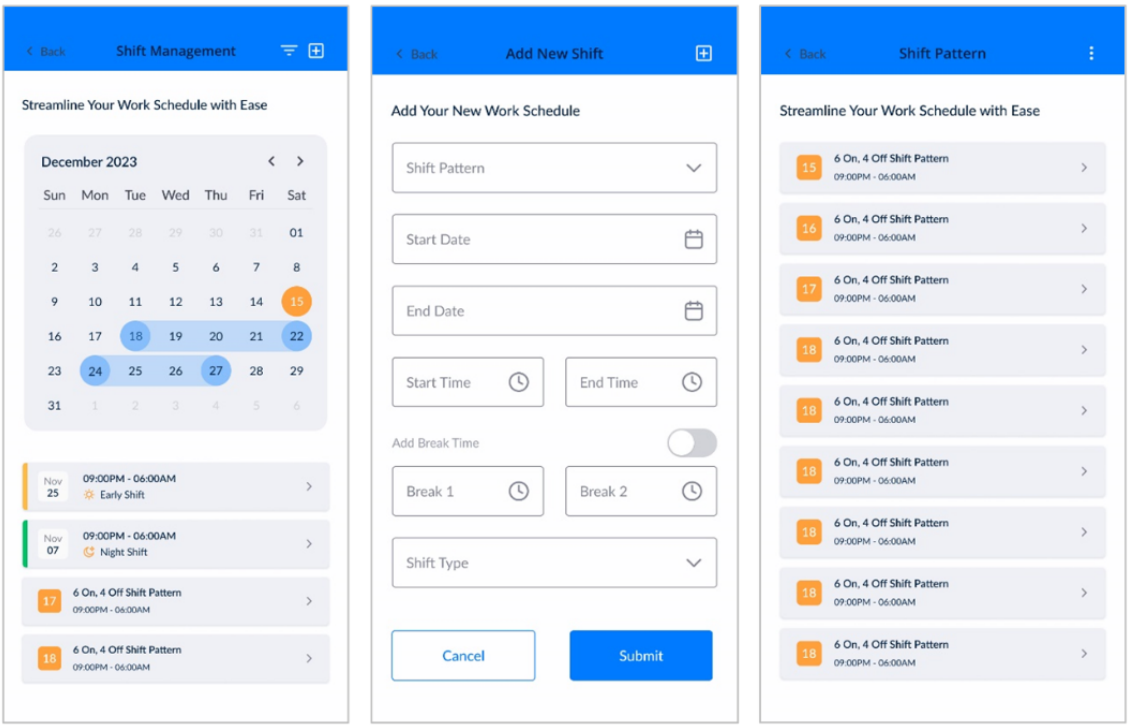

Second Prototype Iteration: Refined Interface of the Shift Management System

The second iteration of the shift management system was developed to address the issues identified in the initial prototype. One of the key improvements of this iteration was enhanced visualisation, which allowed users to more clearly identify trends in shift types and workload distribution, thereby supporting better decision-making (Figures 8A, 8B, and 8C).

Figure 8. Refined iteration of the shift management system.

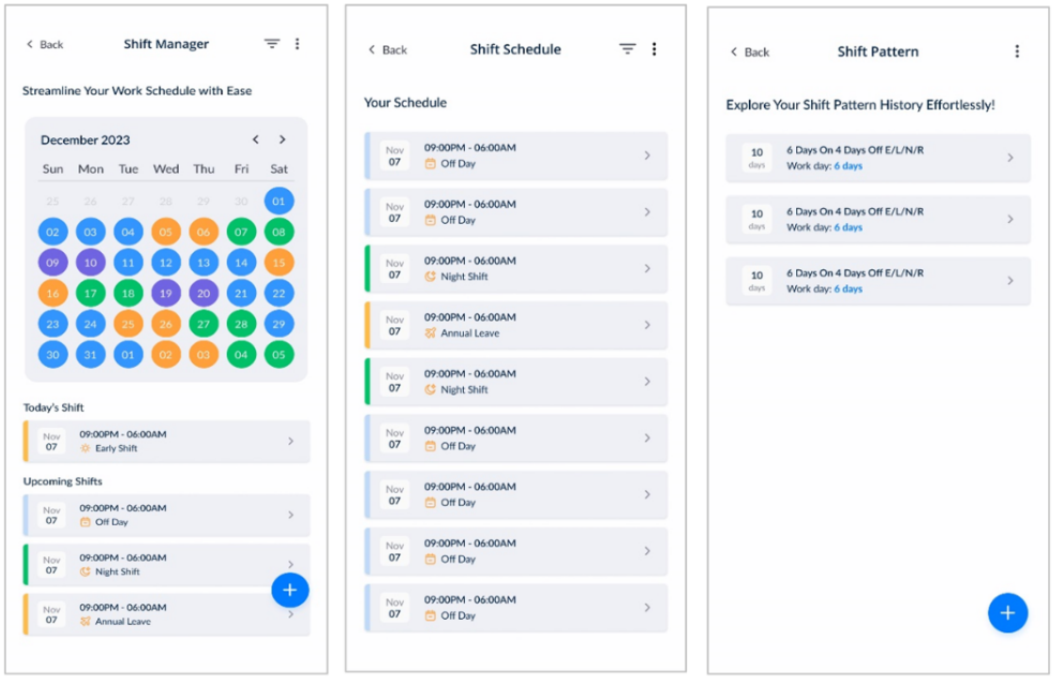

The system introduced the pre-population of base shift patterns, which allowed users to reuse and modify existing patterns, significantly reducing the time and effort required to create new schedules. This feature, combined with the ability to adapt and modify previously created patterns to generate new ones provided additional flexibility and efficiency for users managing dynamic shift schedules (Figures 9A and 9B). The system also introduced functionality to merge multiple rotating shift schedules, which helped to streamline the management of complex team schedules (Figure 9C).

XSL•FO
RenderX

Figure 9. Base shift pattern reusability features.

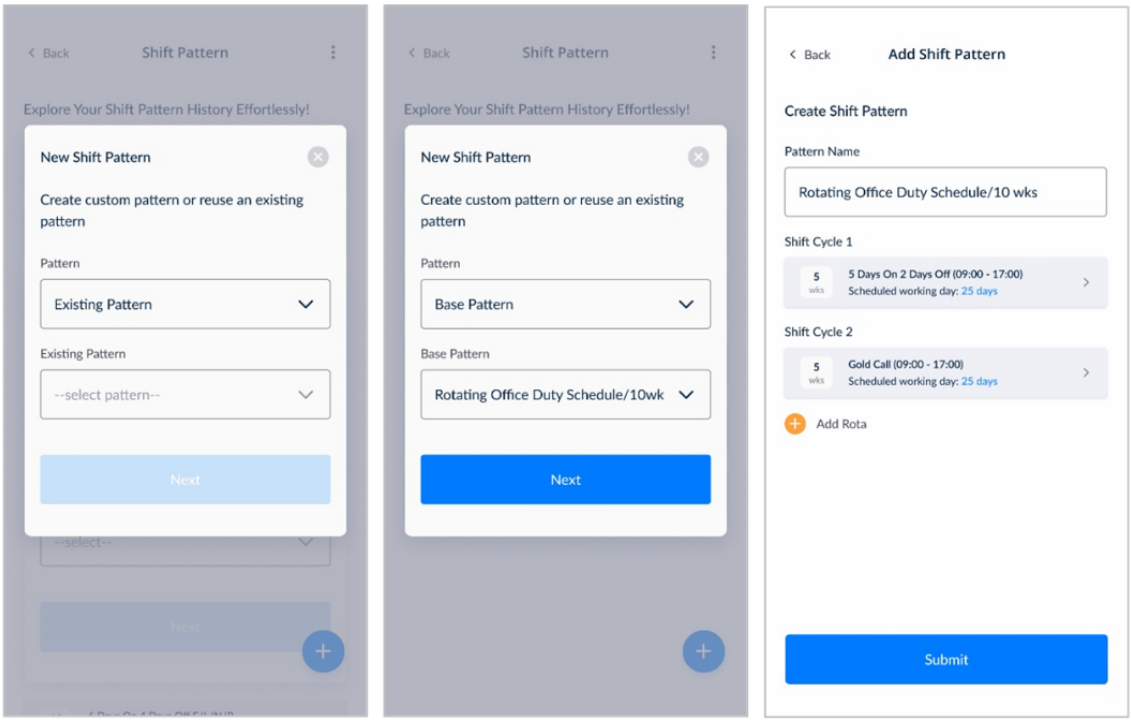

Further enhancements included simplifying the process of adding new shifts, thereby reducing the number of steps and improving user-friendliness (Figure 10A). In addition, the system was updated to accommodate varying shift end times, providing more flexibility and customisation options for users with nonstandard or extended work hours (Figures 10B and 10C).

Figure 10. Shift type variability and finishing time customization.

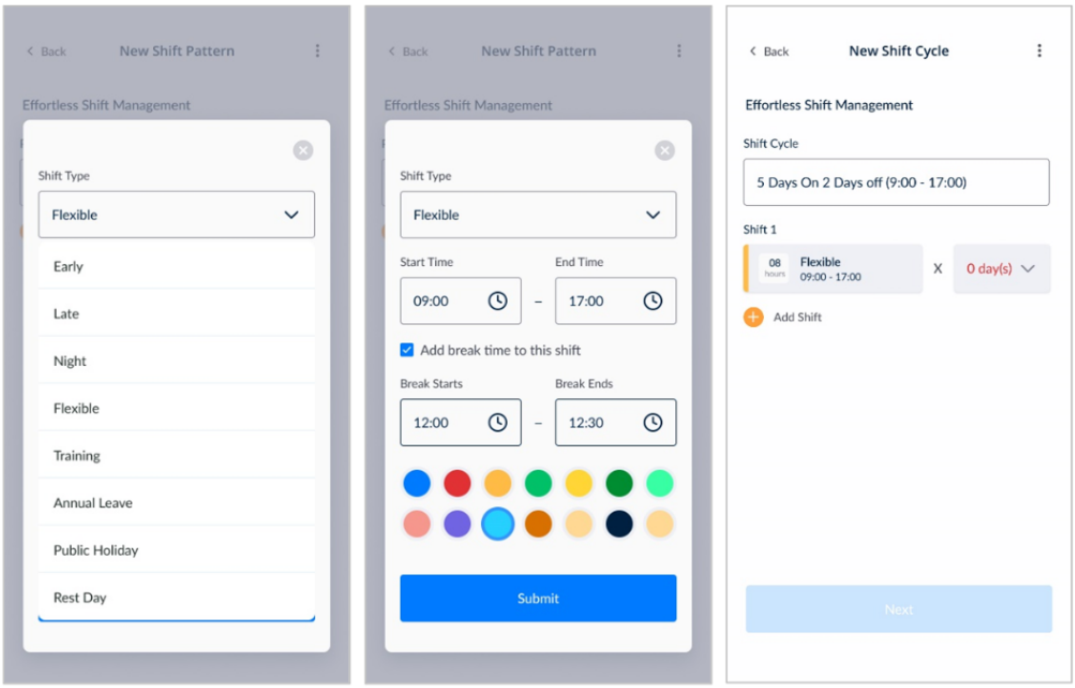

Finally, the iteration introduced consistent vocabulary and terminology to improve communication and reduce confusion. *Shift* was defined as a single working period (individual shift), *shift cycle* as a combination of shift and nonshift days over a working week, and *shift pattern* as one or more shift cycles distributed across a specified number of weeks. This standardisation in terminology helped users better understand and navigate the system (Figures 11A, 11B, and 11C).

Table 6 presents the functionality difference between the first and second iteration. The enhancements made the second iteration effective and user-friendly for the shift management system.

XSL•FO
RenderX

Figure 11. Recurring shift pattern design.

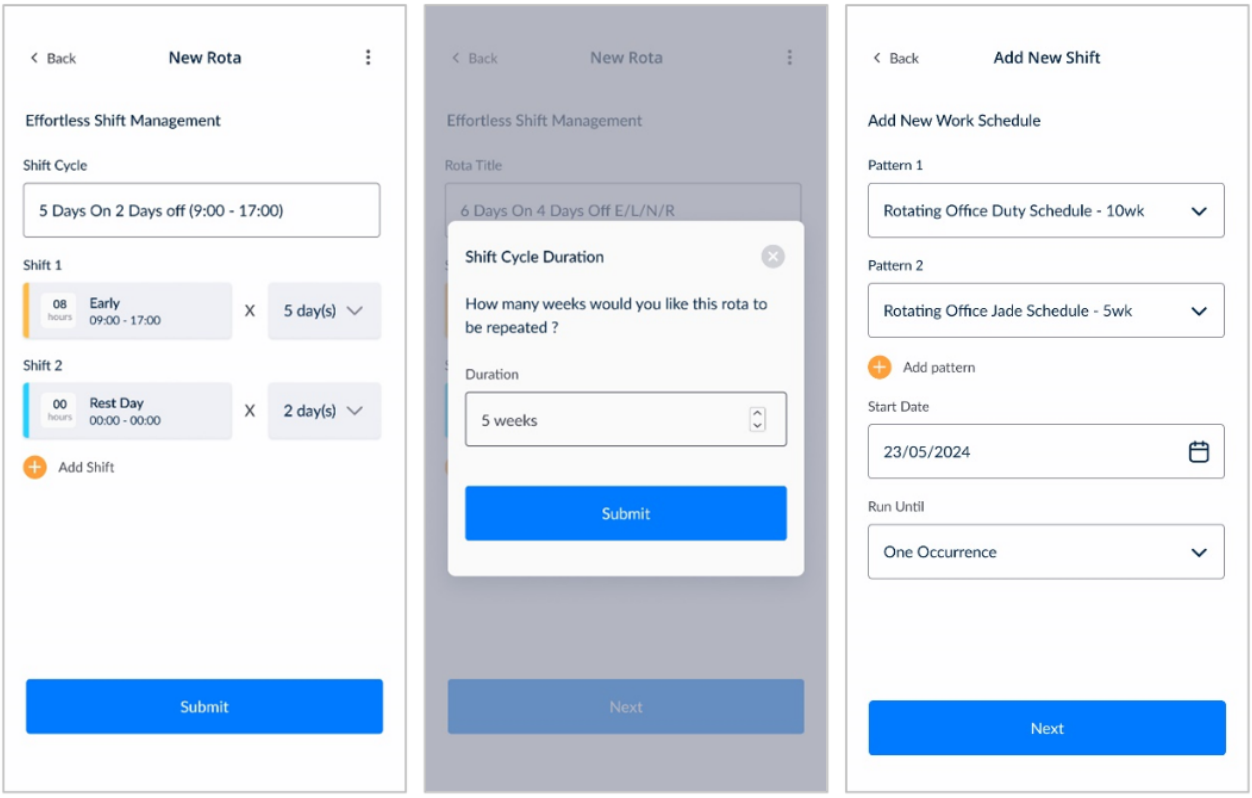

Table 6. Comparison of first and second iteration features of the shift management system.

Feature	First iteration	Second iteration
Shift scheduling	Basic, single shift type allocation	Bulk shift creation
Shift time customization	Fixed times	Daily time customization
Merging rotating schedules	Absent	Merging of multiple rotating schedules
Shift visualization	Limited and not color coded	Color-coded visualization
Notification	Absent	Notification for upcoming shifts added
Terminology	Unclear	Standardized terminology (“shift,” “shift cycle,” and “shift pattern”)
Preconfigured templates	Absent	Prepopulated base shift patterns and reusability of existing patterns
Recurring shift cycles	Absent	Recurring shift cycles with flexible cycle setup
Multiple finishing times	Absent	Supported within a single shift pattern
Shift creation process	Complex and requiring many steps	Simplified shift creation process
Time format	12-h format	24-h format
Start of shift patterns	Random start days	Shift pattern start aligned with Monday (traditional workweek)

In summary, the co-design process led to a substantial evolution of the prototype, resulting in a more refined and operationally aligned shift management system. The second iteration incorporated several key improvements that directly addressed the complex scheduling needs of police personnel. The iterated design included the introduction of bulk shift creation, flexible recurring shift cycles, merging of multiple rotating schedules, and daily shift time customization. The system also adopted a 24-hour time format, aligned shift pattern start dates to Mondays to reflect the traditional workweek, and introduced color-coded visualization and standardized terminology to enhance usability and clarity.

In addition, the integration of shift notifications and the reuse of preconfigured base patterns further increased the tool’s practicality and responsiveness to operational needs. The resulting prototype demonstrated a feasible and user-informed approach to digital shift management capable of supporting broader well-being goals for police personnel when integrated within a tailored DHI.

XSL•FO
RenderX

Discussion

Principal Findings

This study co-designed a digital shift management system tailored to the unique needs of UK police officers through 6 IM-facilitated co-design workshops. Key system features identified included flexible shift scheduling, visualization tools, preconfigured templates, daily time customization, and support for recurring patterns and last-minute changes. The resulting prototype reflected improved usability and operational alignment based on user feedback. This study highlights the value of UCD and participatory methods in enhancing the relevance and acceptance of digital tools. These findings highlight the importance of integrating shift management tools with digital health platforms. This integration would enable personalized well-being recommendations that support adequate hydration, nutrition, and physical activity; proper sleep quality; and sustainable work-life balance based on individual shift patterns.

The ideation results from the cocreation sessions revealed several key features that are important in developing an effective shift management system aimed at addressing the stress and workload challenges faced by police officers. These findings were generated from discussions with police workers and other stakeholders, providing experiences, scenarios, and insights into the demands of managing complex shift schedules.

Primary concerns identified included the lack of shift management features in digital health and well-being applications and the lack of flexibility in current shift management systems. These issues are particularly challenging for unconventional working hours and in handling last-minute changes such as extended hours or sudden shift adjustments, which significantly contribute to stress and work-life imbalance among police workers [3,16]. Functionality such as the bulk creation of shift schedules and flexible modification options emerged as critical for ensuring efficient management of large or block shift schedules and allowing for quick adjustments to accommodate personal or professional obligations [20]. Participants further stressed the need for daily shift time customization, which enables officers to balance their commitments with professional responsibilities. In addition, the manual editing of individual days within a shift pattern allowed for rapid adjustments without the need to recreate the entire schedule, thereby minimizing scheduling-related stress [49].

The ability to monitor shift types through visualization and the option to merge rotating schedules were also identified as important features. Regarding shift pattern design, participants recommended that shift patterns contain a combination of several shift types and that shift patterns with flexible cycles be offered. This would help manage the natural complexity associated with shift work and reduce disruptions that contribute to stress. Participants also highlighted the importance of the system being intuitive and user-friendly [70,76]. This includes using understandable and familiar vocabulary such as *shift*, *shift cycle*, and *shift pattern* and ensuring that shift patterns begin on a Monday to align with traditional workweeks. Such improvements would make the system more accessible and reduce the cognitive load required to interpret shift schedules. Furthermore, using a 24-hour time format across the system was suggested to ensure clarity and consistency in shift management [3].

Finally, the cocreation sessions pointed to the benefit of prepopulating the system with common base shift patterns that can be reused and modified, significantly reducing the time and effort required for creating new shift schedules. The ability to create schedules with multiple finishing times would accommodate the varying shift needs of officers. In addition, notifications to alert police personnel of upcoming shifts would help reduce last-minute scheduling surprises and associated stress, ensuring better communication and preparedness. Integrating planned or scheduled work breaks within each shift would be beneficial to police personnel in reducing fatigue, improving their focus, and promoting their overall health and well-being, thereby improving performance and job satisfaction [8,32,77,78].

Overall, these findings highlight the need for a shift management system that prioritizes flexibility, clarity, and real-time adaptability to accommodate the unpredictable nature of police work. The system must allow for the seamless input and modification of shift schedules, enabling officers to manage surprise schedule changes and long hours easily. This capability is essential to reduce the stress caused by the complex and demanding nature of shift work. The shift management system should both address operational challenges and play a critical role in enhancing officer well-being by providing real-time notifications and recommendations that reflect these sudden adjustments [77].

Evaluating the Relationship With Previous Findings

The findings of this study align with the lack of shift schedule personalization in existing health and well-being interventions for emergency workers such as police officers [3]. Studies on existing mHealth tools have revealed that, while apps such as Healthier Outcomes at Work [39], the #SWPMoveMore Challenge intervention [41], Headspace [37], Sleepfit [38], and Healthy Minds Innovation [40] address various aspects of health, they are not tailored to the specific needs of shift workers. These apps offer crucial interventions, such as workplace stress management, mindfulness, sleep health, and physical activity support, but lack integration with real-time shift schedules, thus making them limited in meeting the unique demands of populations such as police officers and emergency workers who experience unpredictable and irregular work patterns.

Compared with existing literature on digital stress interventions, the shift management tool codeveloped in this study shows alignment with the findings of Weerasekara and Smedberg [2], who emphasize user preferences for personalized and flexible digital solutions. By incorporating UCD principles [68] and IM [58], this study contributes to research advocating for participatory design [60] as a means of increasing the acceptability and usability of mHealth tools. The emphasis on adaptive scheduling, modular functionality, and simplified workflows reflects existing best practices for improving digital engagement and reducing occupational stress [79], fatigue [80], and depression in the workplace [81] through digital platforms [46,49]. Notably, the inclusion of domain experts in shift

XSL•FO
RenderX

coordination, digital health, and HCI enhanced the quality of design decisions and ensured alignment with real-world operational constraints.

Limitations

This study has several limitations. First, the relatively small participant sample size may limit the generalizability of the findings beyond the immediate police force context. Second, despite the diversity of professional roles represented in the co-design workshops, broader input from frontline personnel across multiple forces would have further enhanced the design applicability. Third, no long-term deployment or efficacy testing was conducted within this study phase, meaning that impacts on stress or job performance remain unmeasured. Finally, the IM workshops, though highly structured and effective for consensus building, required substantial facilitation and planning efforts that may not be feasible in all contexts.

Conclusions

This study co-designed a shift management system whose effectiveness and impact can be more thoroughly evaluated and used when integrated into a broader health and well-being application tailored to the needs of shift workers. This study concluded that developers and designers should create DHIs that are tailored to the needs of professions with irregular shift schedules to ensure real-time adaptability and personalization. Organizations specifically within emergency services should encourage the adoption of health and well-being solutions that promote employees' health, job satisfaction, and quality of life. Finally, more collaborations among health experts, frontline personnel, and technology experts should be encouraged to ensure that digital health solutions are relevant and sustainable in addressing the challenges associated with shift work.

Acknowledgments

The authors acknowledge the Department of Computing and Informatics at Bournemouth University, which supported this study.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary materials.
[DOCX File , 23 KB - formative_v9i1e69986_app1.docx]

References

1. Violanti JM, Charles LE, McCanlies E, Hartley TA, Baughman P, Andrew ME, et al. Police stressors and health: a state-of-the-art review. Policing 2017 Nov;40(4):642-656 [FREE Full text] [doi: 10.1108/PIJPSM-06-2016-0097] [Medline: 30846905]
2. Weerasekara M, Smedberg Å. Exploration of user needs and design requirements of a digital stress management intervention for software employees in Sri Lanka: a qualitative study. BMC Public Health 2023 Mar 27;23(1):566 [FREE Full text] [doi: 10.1186/s12889-023-15480-7] [Medline: 36973795]
3. Swanston E, Pulman A, Dogan H, Murphy J, Bitters F. Scoping the need for a tailored mHealth app to improve health and well-being behavioral transformation in the police: exploring the views of UK police workers via web-based surveys and client meetings. JMIR Form Res 2021 Aug 31;5(8):e28075 [FREE Full text] [doi: 10.2196/28075] [Medline: 34463625]
4. Galbraith N, Boyda D, McFeeters D, Galbraith V. Patterns of occupational stress in police contact and dispatch personnel: implications for physical and psychological health. Int Arch Occup Environ Health 2021 Mar 12;94(2):231-241. [doi: 10.1007/s00420-020-01562-1] [Medline: 33044570]
5. Stevelink SA, Pernet D, Dregan A, Davis K, Walker-Bone K, Fear NT, et al. The mental health of emergency services personnel in the UK Biobank: a comparison with the working population. Eur J Psychotraumatol 2020 Sep 23;11(1):1799477 [FREE Full text] [doi: 10.1080/20008198.2020.1799477] [Medline: 33062207]
6. Fekedulegn D, Burchfiel CM, Charles LE, Hartley TA, Andrew ME, Violanti JM. Shift work and sleep quality among urban police officers: the BCOPS study. J Occup Environ Med 2016 Mar;58(3):e66-e71 [FREE Full text] [doi: 10.1097/JOM.0000000000000620] [Medline: 26949891]
7. Riedy SM, Fekedulegn D, Vila B, Andrew M, Violanti J. 0179 poor sleep quality increases odds of on-duty injuries in policing abstract. Sleep 2020;43(Issue Supplement_1):A71. [doi: 10.1093/sleep/zsaa056.177]
8. Graham L, Plater M, Brown NB. National police wellbeing survey 2023 results. Oscar Kilo. URL: https://www.oscarkilo.org.uk/media/7027/download?inline [accessed 2024-10-15]
9. Scholarios D, Hesselgreaves H, Pratt R. Quality of working time in the police: the experience of shift extensification for officers and staff. In: Isidorsson T, Kubisa J, editors. Job Quality in an Era of Flexibility: Experiences in a European Context. New York, NY: Routledge; 2018:54-80.
10. Peterson SA, Wolkow AP, Lockley SW, O'Brien CS, Qadri S, Sullivan JP, et al. Associations between shift work characteristics, shift work schedules, sleep and burnout in North American police officers: a cross-sectional study. BMJ Open 2019 Dec 01;9(11):e030302 [FREE Full text] [doi: 10.1136/bmjopen-2019-030302] [Medline: 31791964]

XSL•FO
RenderX

11. Violanti JM, Fekedulegn D, Andrew ME, Charles LE, Hartley TA, Vila B, et al. Shift work and long-term injury among police officers. Scand J Work Environ Health 2013 Jul;39(4):361-368 [FREE Full text] [doi: 10.5271/sjweh.3342] [Medline: 23503596]
12. Ma CC, Andrew ME, Fekedulegn D, Gu JK, Hartley TA, Charles LE, et al. Shift work and occupational stress in police officers. Saf Health Work 2015 Mar;6(1):25-29 [FREE Full text] [doi: 10.1016/j.shaw.2014.10.001] [Medline: 25830066]
13. Murray E, Hekler EB, Andersson G, Collins LM, Doherty A, Hollis C, et al. Evaluating digital health interventions: key questions and approaches. Am J Prev Med 2016 Nov;51(5):843-851 [FREE Full text] [doi: 10.1016/j.amepre.2016.06.008] [Medline: 27745684]
14. Blandford A. HCI for health and wellbeing: challenges and opportunities. Int J Hum Comput Stud 2019 Nov;131:41-51. [doi: 10.1016/j.ijhcs.2019.06.007]
15. Smits M, Kim CM, van Goor H, Ludden GD. From digital health to digital well-being: systematic scoping review. J Med Internet Res 2022 Apr 04;24(4):e33787 [FREE Full text] [doi: 10.2196/33787] [Medline: 35377328]
16. Buckingham SA, Williams AJ, Morrissey K, Price L, Harrison J. Mobile health interventions to promote physical activity and reduce sedentary behaviour in the workplace: a systematic review. Digit Health 2019;5:2055207619839883 [FREE Full text] [doi: 10.1177/2055207619839883] [Medline: 30944728]
17. Gavin P, Porter CN. Understanding the impact of organisational and operational stressors on the mental health of police officers in Ireland. Police Pract Res 2024 Jun 07;26(1):90-100. [doi: 10.1080/15614263.2024.2364247]
18. Johnson LM, Watson D, Pino NW. Policing in Trinidad and Tobago: officers' experiences with stress. Soc Sci J 2020 Apr 14;60(3):432-446. [doi: 10.1080/03623319.2020.1744952]
19. Baek H, Han S, Seepersad R. The impact of social support and occupational stress on burnout in the Trinidad and Tobago police service. Police Q 2021 Jul 26;25(3):310-337. [doi: 10.1177/10986111211036007]
20. Violanti JM, Fekedulegn D, Hartley TA, Charles LE, Andrew ME, Ma CC, et al. Highly rated and most frequent stressors among police officers: gender differences. Am J Crim Justice 2016 Dec 23;41(4):645-662 [FREE Full text] [doi: 10.1007/s12103-016-9342-x] [Medline: 28260848]
21. Angehrn A, Fletcher AJ, Carleton RN. "Suck it up, buttercup": understanding and overcoming gender disparities in policing. Int J Environ Res Public Health 2021 Jul 18;18(14):7627 [FREE Full text] [doi: 10.3390/ijerph18147627] [Medline: 34300078]
22. Purba A, Demou E. The relationship between organisational stressors and mental wellbeing within police officers: a systematic review. BMC Public Health 2019 Oct 15;19(1):1286 [FREE Full text] [doi: 10.1186/s12889-019-7609-0] [Medline: 31615479]
23. Acquadro Maran D, Magnavita N, Garbarino S. Identifying organizational stressors that could be a source of discomfort in police officers: a thematic review. Int J Environ Res Public Health 2022 Mar 21;19(6):3720 [FREE Full text] [doi: 10.3390/ijerph19063720] [Medline: 35329405]
24. Kivimäki M, Kawachi I. Work stress as a risk factor for cardiovascular disease. Curr Cardiol Rep 2015 Sep;17(9):630 [FREE Full text] [doi: 10.1007/s11886-015-0630-8] [Medline: 26238744]
25. Record 14,508 officers were signed off with poor mental health in the past year. Police Oracle. URL: https://www.policeoracle.com/article-library/record-14508-officers-were-signed-off-with-poor-mental-health-in-the-past-year/ [accessed 2025-04-10]
26. Cartwright A, Roach J. The wellbeing of UK police: a study of recorded absences from work of UK police employees due to psychological illness and stress using freedom of information act data. Policing 2020;15(2):1326-1338 [FREE Full text] [doi: 10.1093/police/paaa033]
27. Quin E. Trends in absences owing to psychological illnesses in 'frontline' and 'response' officers in England and Wales. Police Pract Res 2024 Jul 12:763-771 [FREE Full text] [doi: 10.1080/15614263.2024.2375500]
28. Pay and morale survey 2023 – headline report February 2024. Police Federation of England and Wales. 2024. URL: https://www.polfed.org/media/19164/pay-and-morale-survey-report-2023.pdf [accessed 2025-04-10]
29. Pay and morale survey 2022 – headline report December 2023. Police Federation of England and Wales. URL: https://www.polfed.org/media/18245/pay-and-morale-2022_headline-report.pdf [accessed 2025-04-28]
30. Vila B, Morrison GB, Kenney DJ. Improving shift schedule and work-hour policies and practices to increase police officer performance, health, and safety. Police Q 2002 Mar 01;5(1):4-24. [doi: 10.1177/109861102129197995]
31. Fekedulegn D, Burchfiel CM, Hartley TA, Andrew ME, Charles LE, Tinney-Zara CA, et al. Shiftwork and sickness absence among police officers: the BCOPS study. Chronobiol Int 2013 Aug;30(7):930-941 [FREE Full text] [doi: 10.3109/07420528.2013.790043] [Medline: 23808812]
32. Kecklund G, Eriksen CA, Akerstedt T. Police officers attitude to different shift systems: association with age, present shift schedule, health and sleep/wake complaints. Appl Ergon 2008 Sep;39(5):565-571. [doi: 10.1016/j.apergo.2008.01.002] [Medline: 18281011]
33. McGettrick KS, O'Neill MA. Critical care nurses--perceptions of 12-h shifts. Nurs Crit Care 2006 Jun 20;11(4):188-197. [doi: 10.1111/j.1362-1017.2006.00171.x] [Medline: 16869525]
34. Knauth P. Extended work periods. Ind Health 2007 Jan;45(1):125-136 [FREE Full text] [doi: 10.2486/indhealth.45.125] [Medline: 17284884]

XSL•FO
RenderX

35. Amendola KL, Weisburd D, Hamilton EE, Jones G, Slipka M. An experimental study of compressed work schedules in policing: advantages and disadvantages of various shift lengths. J Exp Criminol 2011 Sep 24;7(4):407-442. [doi: 10.1007/S11292-011-9135-7]
36. Nunes F, Ribeiro J, Braga C, Lopes P. Supporting the self-care practices of shift workers. In: Proceedings of the 17th International Conference on Mobile and Ubiquitous Multimedia. 2018 Presented at: MUM '18; November 25-28, 2018; Cairo, Egypt p. 71-81 URL: https://dl.acm.org/doi/10.1145/3282894.3282914 [doi: 10.1145/3282894.3282914]
37. Xu HG, Tuckett A, Kynoch K, Eley R. A mobile mindfulness intervention for emergency department staff to improve stress and wellbeing: a qualitative study. Int Emerg Nurs 2021 Sep;58:101039. [doi: 10.1016/j.ienj.2021.101039] [Medline: 34333332]
38. Shriane AE, Vincent GE, Ferguson SA, Rebar A, Kolbe-Alexander T, Rigney G. Improving sleep health in paramedics through an app-based intervention: a randomised waitlist control pilot trial. BMC Public Health 2024 Sep 03;24(1):2395 [FREE Full text] [doi: 10.1186/s12889-024-19823-w] [Medline: 39227826]
39. Ravalier JM, Wainwright E, Smyth N, Clabburn O, Wegrzynek P, Loon M. Co-creating and evaluating an app-based well-being intervention: the HOW (Healthier Outcomes at Work) social work project. Int J Environ Res Public Health 2020 Nov 24;17(23):8730 [FREE Full text] [doi: 10.3390/ijerph17238730] [Medline: 33255460]
40. Pace TW, Zeiders KH, Cook SH, Sarsar ED, Hoyt LT, Mirin NL, et al. Feasibility, acceptability, and preliminary efficacy of an app-based meditation intervention to decrease firefighter psychological distress and burnout: a one-group pilot study. JMIR Form Res 2022 Jun 08;6(6):e34951 [FREE Full text] [doi: 10.2196/34951] [Medline: 35675115]
41. Oliver H, Thomas O, Copeland RJ, Hesketh I, Jukes M, Chadd K, et al. Proof of concept and feasibility of the app-based '#SWPMoveMore challenge': impacts on physical activity and well-being in a police population. Police J 2021 Jun 24;95(1):170-189. [doi: 10.1177/0032258x211024690]
42. Satterfield BC, Van Dongen HP. Occupational fatigue, underlying sleep and circadian mechanisms, and approaches to fatigue risk management. Fatigue Biomed Health Behav 2013 Jun 14;1(3):118-136. [doi: 10.1080/21641846.2013.798923]
43. Koh J, Tng GY, Hartanto A. Potential and pitfalls of mobile mental health apps in traditional treatment: an umbrella review. J Pers Med 2022 Aug 25;12(9):1376 [FREE Full text] [doi: 10.3390/jpm12091376] [Medline: 36143161]
44. Digital health technologies in mental health: insights report. Health Innovation Network. URL: https://healthinnovationnetwork.com/wp-content/uploads/2024/03/Digital-Health-Technologies-in-Mental-Health-Insights-Report.pdf [accessed 2024-12-15]
45. Moe-Byrne T, Shepherd J, Merecz-Kot D, Sinokki M, Naumanen P, Hakkaart-van Roijen L, et al. Effectiveness of tailored digital health interventions for mental health at the workplace: a systematic review of randomised controlled trials. PLOS Digit Health 2022 Oct 21;1(10):e0000123 [FREE Full text] [doi: 10.1371/journal.pdig.0000123] [Medline: 36812547]
46. Linardon J, Fuller-Tyszkiewicz M. Attrition and adherence in smartphone-delivered interventions for mental health problems: a systematic and meta-analytic review. J Consult Clin Psychol 2020 Jan;88(1):1-13. [doi: 10.1037/ccp0000459] [Medline: 31697093]
47. Marston HR, Hadley R, Pike G, Hesketh I. Police J 2020 Jul 06;94(3):372-387. [doi: 10.1177/0032258x20937327]
48. Abernethy A, Adams L, Barrett M, Bechtel C, Brennan P, Butte A, et al. The promise of digital health: then, now, and the future. NAM Perspect 2022;2022:956 [FREE Full text] [doi: 10.31478/202206e] [Medline: 36177208]
49. Howe E, Suh J, Bin MM. Design of digital workplace stress-reduction intervention systems: effects of intervention type and timing. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. 2022 Presented at: CHI '22; April 29-May 5, 2022; New Orleans, LA p. 1-16 URL: https://dl.acm.org/doi/10.1145/3491102.3502027 [doi: 10.1145/3491102.3502027]
50. Kerr JI, Naegelin M, Benk M, V Wangenheim F, Meins E, Viganò E, et al. Investigating employees' concerns and wishes regarding digital stress management interventions with value sensitive design: mixed methods study. J Med Internet Res 2023 Apr 13;25:e44131 [FREE Full text] [doi: 10.2196/44131] [Medline: 37052996]
51. Chaudhry BM, Islam A, Matthieu M. Correction: toward designs of workplace stress management mobile apps for frontline health workers during the COVID-19 pandemic and beyond: mixed methods qualitative study. JMIR Form Res 2022 Jan 25;6(1):e36769 [FREE Full text] [doi: 10.2196/36769] [Medline: 35077389]
52. Wright EM. Evaluation of a web-based holistic stress reduction pilot program among nurse-midwives. J Holist Nurs 2018 Jun;36(2):159-169. [doi: 10.1177/0898010117704325] [Medline: 29172933]
53. de Korte EM, Wiezer N, Janssen JH, Vink P, Kraaij W. Evaluating an mHealth app for health and well-being at work: mixed-method qualitative study. JMIR Mhealth Uhealth 2018 Mar 28;6(3):e72 [FREE Full text] [doi: 10.2196/mhealth.6335] [Medline: 29592846]
54. Global strategy on digital health 2020-2025. World Health Organization. URL: https://www.who.int/docs/default-source/documents/gs4dhdaa2a9f352b0445bafbc79ca799dce4d.pdf [accessed 2025-04-28]
55. Weyant E. Research design: qualitative, quantitative, and mixed methods approaches, 5th edition. J Electron Resour Med Libr 2022 Mar 14;19(1-2):54-55. [doi: 10.1080/15424065.2022.2046231]
56. Palinkas LA, Horwitz SM, Green CA, Wisdom JP, Duan N, Hoagwood K. Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. Adm Policy Ment Health 2015 Sep;42(5):533-544 [FREE Full text] [doi: 10.1007/s10488-013-0528-y] [Medline: 24193818]

XSL•FO
RenderX

57. Janes FR. Interactive management. In: Inllis K, Gregory A, Mears-Young BR, Ragsdell G, editors. Critical Issues in Systems Theory and Practice. Boston, MA: Springer; 1995:51-60.
58. Warfield JN, Cárdenas AR. A Handbook of Interactive Management Internet. 2nd edition. Palm Harbor, FL: Ajar Publishing Company; 2002.
59. Dogan H, Henshaw M, Urwin EN. A 'soft' approach to TLM requirements capture to support through-life management. In: Proceedings of the 3rd International Conference on Knowledge Science, Engineering and Management. 2009 Presented at: KSEM '09; November 25-27, 2009; Vienna, Austria p. 458-469 URL: https://link.springer.com/chapter/10.1007/978-3-642-10488-6_44 [doi: 10.1007/978-3-642-10488-6_44]
60. Stock C, Dias S, Dietrich T, Frahsa A, Keygnaert I. Editorial: how can we co-create solutions in health promotion with users and stakeholders? Front Public Health 2021;9:773907 [FREE Full text] [doi: 10.3389/fpubh.2021.773907] [Medline: 34957026]
61. Yardley L, Morrison L, Bradbury K, Muller I. The person-based approach to intervention development: application to digital health-related behavior change interventions. J Med Internet Res 2015 Jan 30;17(1):e30 [FREE Full text] [doi: 10.2196/jmir.4055] [Medline: 25639757]
62. Brown MC, Araújo-Soares V, Skinner R, Glaser AW, Sarwar N, Saxton JM, et al. Using qualitative and co-design methods to inform the development of an intervention to support and improve physical activity in childhood cancer survivors: a study protocol for BEing Active after ChildhOod caNcer (BEACON). BMJ Open 2020 Dec 21;10(12):e041073 [FREE Full text] [doi: 10.1136/bmjopen-2020-041073] [Medline: 33371034]
63. Mantovani N, Pizzolati M, Gillard S. Engaging communities to improve mental health in African and African Caribbean groups: a qualitative study evaluating the role of community well-being champions. Health Soc Care Community 2017 Jan 05;25(1):167-176. [doi: 10.1111/hsc.12288] [Medline: 26439601]
64. Sweeney JC, Danaher TS, McColl-Kennedy JR. Customer effort in value cocreation activities: improving quality of life and behavioral intentions of health care customers. J Serv Res 2015 Mar 10;18(3):318-335. [doi: 10.1177/1094670515572128]
65. McColl-Kennedy JR, Hogan SJ, Witell L, Snyder H. Cocreative customer practices: effects of health care customer value cocreation practices on well-being. J Bus Res 2017 Jan;70:55-66. [doi: 10.1016/j.jbusres.2016.07.006]
66. Leask CF, Sandlund M, Skelton DA, Altenburg TM, Cardon G, Chinapaw MJ, GrandStand, Safe StepTeenage Girls on the Move Research Groups. Framework, principles and recommendations for utilising participatory methodologies in the co-creation and evaluation of public health interventions. Res Involv Engagem 2019 Jan 9;5(1):2 [FREE Full text] [doi: 10.1186/s40900-018-0136-9] [Medline: 30652027]
67. De Koning J, Crul M, Wever R. Models of co-creation. In: Proceedings of the 2016 International Conference on Service Design Geographies. 2016 Presented at: ServDes '16; May 24-26, 2016; Copenhagen, Denmark URL: https://www.researchgate.net/publication/303541138_Models_of_Co-creation [doi: 10.4135/9781473988866]
68. De Vito Dabbs A, Myers BA, Mc Curry KR, Dunbar-Jacob J, Hawkins RP, Begey A, et al. User-centered design and interactive health technologies for patients. Comput Inform Nurs 2009;27(3):175-183 [FREE Full text] [doi: 10.1097/NCN.0b013e31819f7c7c] [Medline: 19411947]
69. Elliott A, Lang S, Truby H, Brennan L, Gibson S. Tackling the challenge of treating obesity using design research methods: a scoping review. Obes Rev 2022 Mar 20;23(2):e13360. [doi: 10.1111/obr.13360] [Medline: 34545669]
70. Saparamadu AA, Fernando P, Zeng P, Teo H, Goh A, Lee JM, et al. User-centered design process of an mHealth app for health professionals: case study. JMIR Mhealth Uhealth 2021 Mar 26;9(3):e18079 [FREE Full text] [doi: 10.2196/18079] [Medline: 33769297]
71. Michalik G. The co-creation process. In: Michalik G, editor. Co-Creation Mindset: Eight Steps Towards the Future of Work. Cham, Switzerland: Springer; 2023:83-97.
72. Clarke V, Braun V. Thematic analysis. J Posit Psychol 2016 Dec 09;12(3):297-298. [doi: 10.1080/17439760.2016.1262613]
73. Fugard AJ, Potts HW. Supporting thinking on sample sizes for thematic analyses: a quantitative tool. Int J Soc Res Methodol 2015 Feb 10;18(6):669-684. [doi: 10.1080/13645579.2015.1005453]
74. Naeem M, Ozuem W, Howell K, Ranfagni S. A step-by-step process of thematic analysis to develop a conceptual model in qualitative research. Int J Qual Methods 2023 Nov 08;22:62. [doi: 10.1177/16094069231205789]
75. Salminen J, Jung SG, Chowdhury S, Sengün S, Jansen B. Personas and analytics: a comparative user study of efficiency and effectiveness for a user identification task. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. 2020 Presented at: CHI '20; April 25-30, 2020; Honolulu, HI p. 1-13 URL: https://dl.acm.org/doi/10.1145/3313831.3376770 [doi: 10.1145/3313831.3376770]
76. Phythian R, Birdsall N, Kirby S, Cooper E, Posner Z, Boulton L. Developments in UK police wellbeing: a review of blue light wellbeing frameworks. Police J 2022 Mar 02;95(1):24-49. [doi: 10.1177/0032258x211073003]
77. Local action on health inequalities: workplace interventions to improve health and wellbeing. Public Health England. URL: https://assets.publishing.service.gov.uk/media/5a7df336e5274a2e8ab44de4/Briefing5a_Workplace_interventions_health_inequalities.pdf [accessed 2024-10-15]
78. Oliver H, Thomas O, Neil R, Moll T, Copeland RJ. Stress and psychological wellbeing in british police force officers and staff. Curr Psychol 2022 Nov 16;42(33):29291-29304. [doi: 10.1007/S12144-022-03903-4]

XSL•FO
RenderX

79. Porter CN, Gavin P. ‘Constant pressures from all angles’: understanding the mental health of UK police officers. Police J 2024 Sep 04;98(2):306-323. [doi: 10.1177/0032258x241280217]
80. Caldwell JA, Caldwell JL, Thompson LA, Lieberman HR. Fatigue and its management in the workplace. Neurosci Biobehav Rev 2019 Jan;96:272-289 [FREE Full text] [doi: 10.1016/j.neubiorev.2018.10.024] [Medline: 30391406]
81. Demou E, Hale H, Hunt K. Understanding the mental health and wellbeing needs of police officers and staff in Scotland. Police Pract Res 2020;21(6):702-716 [FREE Full text] [doi: 10.1080/15614263.2020.1772782] [Medline: 33408592]

Abbreviations

DHI: digital health intervention
HCI: human-computer interaction
IM: interactive management
mHealth: mobile health
TA: thematic analysis
UCD: user-centered design

Edited by A Mavragani; submitted 22.12.24; peer-reviewed by H Wei, T Nascimento ; comments to author 09.04.25; revised version received 30.04.25; accepted 12.06.25; published 28.08.25.

Please cite as:
Ayorinde OT, Dogan H, Adedoyin FF, Jiang N, Bitters F, Dempsey S
Supporting Police Well-Being Through an Adaptive Shift Management System: Co-Design Study
JMIR Form Res 2025;9:e69986
URL: https://formative.jmir.org/2025/1/e69986
doi:10.2196/69986
PMID:

©Olumuyiwa Temitope Ayorinde, Huseyin Dogan, Festus Fatai Adedoyin, Nan Jiang, Fiona Bitters, Sara Dempsey. Originally published in JMIR Formative Research (https://formative.jmir.org), 28.08.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Designing Digital Mental Health Interventions to Meet the Needs of Older Adolescents: Qualitative Interview and Group Discussion Study

Rachel Kornfield[1,2], PhD; Sarah Alexandra Popowski[1,2], BA; Emily Tack[1,2], BA; Jack Svoboda[1,2], BA; Miguel Herrera[3], MEd; Theresa Nguyen[4], LCSW; Ashley Arehart Knapp[2,3], PhD; David Curtis Mohr[1,2], PhD; Jonah Meyerhoff[1,2], PhD

[1]Department of Preventive Medicine, Northwestern University, Chicago, IL, United States

[2]Center for Behavioral Intervention Technologies, Northwestern University, Chicago, IL, United States

[3]Department of Psychiatry and Behavioral Sciences, Northwestern University, Chicago, IL, United States

[4]Mental Health America, Alexandria, VA, United States

Corresponding Author:
Rachel Kornfield, PhD
Department of Preventive Medicine
Northwestern University
750 N. Lake Shore Drive, 10th Floor
Chicago, IL, 60611
United States
Phone: 1 312 503 6813
Email: rachel.kornfield@northwestern.edu

Abstract

Background: Anxiety and depression are common in adolescents, but adolescents are often uninterested in formal mental health treatments or are unable to access them. Digital interventions can be delivered at scale to bridge critical gaps in mental health care but must address the needs and preferences of adolescents.

Objective: This study aims to conduct qualitative research involving adolescents aged 18 years to inform both the design of digital mental health interventions for adolescents broadly and new features and refinements to incorporate in an automated SMS text messaging intervention, Small Steps SMS, that was originally designed for young adults.

Methods: We recruited non–treatment-engaged older adolescents who were aged 18 years, lived in the United States, and had experienced depression or anxiety. In total, 12 participants were recruited through social media advertising and online self-screeners hosted by Mental Health America, a mental health advocacy organization. For 24 days, participants answered researcher prompts and engaged with one another in an asynchronous online discussion group, with a new discussion prompt released every 3 days. In parallel, partway through the discussion group, participants received interactive messages from Small Steps SMS, an automated SMS text messaging intervention that delivers daily dialogues supporting mental health self-management. Questions in the discussion group pertained to mental health challenges, help-seeking attitudes, perceptions of Small Steps SMS, and ways the program and other digital mental health interventions could meet the needs of older adolescents. A subset of participants (n=4, 33%) also completed interviews to elaborate on their responses. Thematic analysis was applied to transcripts of the discussion group and interviews to characterize user needs and design priorities when making Small Steps SMS and similar interventions available to adolescents.

Results: Participants reported factors that contributed to their experience of mental health symptoms, including the transition from adolescence to adulthood, fears that the world is unstable and their futures are uncertain, and ineffective use of social media to cope with symptoms. Participants were proud of their generation's mental health acceptance but also observed a generational divide in mental health stigma and literacy that could impede seeking help from parents and other adults. Participants appreciated that Small Steps SMS allowed them to pursue mental health self-management conveniently and independently. They suggested that the program and similar interventions address adolescent-specific challenges and facilitate intergenerational communication about mental health. They also recommended possible ways to increase engagement through peer-to-peer communication, gamification, and greater explanation of self-management strategies.

XSL•FO
RenderX

Conclusions: Major life transitions affected adolescent participants' mental health needs and preferences for digital mental health tools. While interactive automated messaging programs have the potential to support self-management in this population, program content and features should be adapted to adolescents' needs.

(JMIR Form Res 2025;9:e68950) doi:10.2196/68950

KEYWORDS

depression; anxiety; adolescents; qualitative data; SMS text messaging; digital mental health; user-centered design

Introduction

Background

Adolescence refers to the developmental stage between childhood and adulthood, defined by the World Health Organization as spanning ages 10 to 19 years [1]. Adolescence features unique biological developmental changes (ie, puberty), high attunement to social information, heightened emotionality, and evolution of social roles and self-concept [2-4]. These changes make adolescents susceptible to the emergence of mental health symptoms [5]. In addition, adolescent mental health has been affected by recent trends and events, including increases in screen time and social media use and the COVID-19 pandemic and related lockdowns [6,7]. Prevalence of past-year major depressive episodes among individuals aged between 12 and 17 years nearly doubled between 2005 and 2019, increasing from 8.7% to 15.7% and then further increased to 20.1% in 2021 [8-10]. In 2020, 20.5% of adolescents reported clinically elevated anxiety symptoms [11].

Despite high levels of mental health symptoms, few adolescents access mental health services. Typically, adolescents seek informal help from trusted friends and family members before turning to formal sources, such as medical professionals or school resources [12-14]. Among adolescents affected by recent major depressive episodes, only 40.6% received relevant treatment [10], and treatment rates for anxiety disorders were similarly low [15]. This reflects common barriers to help-seeking, such as stigma, which may be particularly salient to socially developing youth [4,16]. Other common barriers include costs, inadequate insurance, limited availability of services, low mental health literacy, negative beliefs about mental health services, and preferences for self-reliance [17-20]. Many states require parental consent for minors to pursue mental health treatment [21], which poses an additional barrier when adolescents anticipate or experience stigma from parents or guardians or prefer to pursue treatment independently [22,23].

There is a growing consensus that the unmet mental health needs of adolescents represent a public health crisis [24,25]. Untreated mental health symptoms in adolescence can lead to distress and increase the risk of poor academic performance, noncompletion of high school or college, psychiatric problems later in life, and suicide [26-29]. Early intervention can help adolescents manage symptoms and build coping skills, potentially changing the trajectory of mental health conditions into adulthood [30].

Digital mental health interventions (DMHIs) represent a potential solution to address the unmet mental health needs of adolescents. Most adolescents have access to and familiarity with digital devices, with 95% of US teenagers having access to a smartphone in 2022 [31]. DMHIs are generally well-received by adolescents, who report a preference for DMHIs that offer flexibility, self-reliance, opportunities for communication with others, and privacy [32]. Examples of existing DMHIs include mobile phone apps, therapeutic video games, and mental health–focused web programs [32,33]. These DMHIs have used many approaches, including coping skill education, mindfulness training, mental health coaching, self-monitoring and self-tracking, and social support [32,34,35]. Some DMHIs have been designed with scalability in mind, such as by delivering interventions in online spaces where adolescents seek mental health information and through communication channels that are already highly used by young people, such as messaging apps, texting, and social media [36-38].

Despite the potential alignment between DMHIs and the needs and preferences of adolescents, most DMHIs to date have focused on adults [39-41]. A systematic overview by Lehtimaki et al [32] concluded that DMHIs for adolescents demonstrate benefits for managing anxiety, depression, and stress when compared to nonactive control groups (receiving no treatment or wait-listed for treatment), but they did not find differences relative to active controls. They also found that completion rates among the reviewed interventions varied widely, ranging between 10% and 94%. Low engagement and high attrition are common issues in trials of DMHIs [35], and these issues are even more pronounced in the real world [42], reflecting the loss of structure and support outside of trials and a less motivated population of users [43]. To meaningfully address adolescents' unmet mental health needs, there is a need for DMHIs that are effective as well as engaging in the context of adolescents' day-to-day lives.

Objectives

To ensure digital mental health tools have potential for real-world engagement and uptake, the digital mental health community increasingly uses human-centered design methods from the human-computer interaction field [43-45]. These methods engage end users to understand their needs, elicit their design ideas, gather iterative feedback on DMHI prototypes, and establish usability, with the goal of creating tools that are engaging and useful to real-world users.

Consistent with such a focus, this paper applies a qualitative human-centered design approach to understand how DMHIs can be designed and adapted to better address the needs and preferences of older adolescents who are not in formal mental health treatment. We focused on those not currently engaged in treatment to inform the design of DMHIs that can help close treatment gaps. We seek to provide insights into the mental health challenges faced by these older adolescents and their

XSL•FO
RenderX

preferences concerning help seeking and self-management to understand how these needs and contexts might be addressed by design approaches and features for DMHIs. We solicit participants' ideas for automated DMHIs broadly and for potential features and revisions to an existing interactive automated SMS text messaging intervention, Small Steps SMS, a program that was originally designed for young adults [46-48]. Automated interventions delivered via communication channels regularly used by young people (such as SMS text messaging) have the potential to expand access to mental health support on a large scale and at low cost. Alongside asking open-ended questions through an online discussion group, we deployed a 3-week version of the Small Steps SMS program as a technology probe [49], offering users experience of using a possible solution to generate ideas and allow specific feedback on the self-management content and interactions within the program. The program provides psychoeducation and supports applying several mental health self-management strategies that draw on evidence-based psychotherapies, such as cognitive behavioral therapy and acceptance and commitment therapy. Participants' feedback will inform potential modifications to Small Steps SMS (eg, content revisions and new features) that can be prototyped and further assessed by end users.

Methods

Overview

To facilitate participation in study activities despite the sensitive and stigmatized nature of mental health issues, we used an asynchronous remote community (ARC) approach [50]. Originating in the human-computer interaction field, ARC allows participants to share their thoughts, opinions, experiences, and preferences anonymously within a private discussion forum and build on one another's contributions. The ARC method seeks to build a comfortable context for active participation and ongoing interaction among participants that fosters a sense of community and collaborative ideation [51]. Recommendations emphasize groups that are relatively small (eg, 10-20 individuals) to facilitate rapport building and participation [52]. Participants may also complete independent activities to gain familiarity with technological solutions and prototypes, informing and deepening their responses within the ARC group [51]. In line with such an approach, participants in our study used the Small Steps SMS program beginning partway through the discussion group, receiving daily interactive SMS text messages. A subset of participants also completed semistructured phone interviews after the discussion group and the Small Steps SMS messaging concluded.

Participants

All recruitment and study activities were fully remote, allowing individuals to engage from home or a location of their choice on their personal device. Participants in our study were recruited through paid advertisements targeting US-based users of social media platforms, namely Facebook (Meta Platforms, Inc) and Instagram (Meta Platforms, Inc), and through free self-screening surveys for depression and anxiety (ie, the Patient Health Questionnaire-9 [53] or Generalized Anxiety Disorder-7 scale [54]) hosted on the website of Mental Health America, a national mental health advocacy organization that serves largely young, non–treatment-seeking individuals with mental health symptoms [55]. The advertisement text described that participants could "Join a research study on technology and mental health." Clicking the attached link led to a landing page with further details about the study and an option to continue to the eligibility survey. Potential participants who completed the eligibility survey were invited to enroll in our study if they self-reported being US residents, experiencing depression or anxiety ("Have you ever had an experience of depression or anxiety even if it was not diagnosed by a professional?"), having sufficient English language ability to participate in study activities, having access to a personal smartphone, and being aged 18 years. This focus on individuals aged 18 years allowed us to capture challenges around the transition from high school into early adulthood and to gather informed consent directly from participants who were legal adults. Exclusion criteria included being engaged in formal mental health treatment; visual, hearing, or motor impairments that would restrict individuals in completing study activities; self-reported serious mental illness (psychotic disorder or bipolar disorder); or suicidal ideation with plan and intent. Potential participants were required to provide a valid US mobile phone number on which they could receive SMS text messages. Individuals who completed screening were provided with a list of resources for accessing 24/7 mental health support, if needed (eg, suicide prevention hotline and crisis text line).

Procedure

Data were collected between August 2023 and September 2023 through an online discussion group, parallel deployment of the Small Steps SMS program, and follow-up interviews.

Discussion Group

Participants were registered for an online discussion group based on ARC methods [50]. The discussion group was hosted on the FocusGroupIt [56] website. Participants were asked to submit text-based responses to prompts posted by the researchers every 3 days for 24 days (8 prompts in total). Prompts asked about participants' mental health experiences, treatment and self-management needs, and perspectives on the design of digital tools for mental health (Multimedia Appendix 1). To encourage active dialogue and collaborative thinking, participants were also asked to reply to each other's responses.

Responses in the discussion group were visible to all study participants; however, participants were identified to one another only by generic pseudonyms that were automatically assigned upon registration to the website (ie, Participant 1, Participant 2, etc), and under which all their comments were posted. Participants in the discussion group agreed to follow a code of conduct that advised them to engage others with respect and avoid disclosing personally identifying information or details about methods of suicide or self-harm. Responses were monitored daily for compliance with the code of conduct. Research staff were supervised by a member of the authorship team who is a licensed clinical psychologist. The team had a risk management protocol in place, such that a participant sharing any information signaling they were a risk to themselves or others would prompt a team member to contact the participant

XSL•FO
RenderX

to administer risk assessments and provide resources, if needed. No such risks emerged; therefore, no follow-up was conducted. Text-based data from the discussion group were automatically saved by the website and downloaded by the researchers at the conclusion of the study.

Small Steps SMS Program

Participants in our study were also enrolled in a truncated version of the Small Steps SMS program following the third discussion group prompt. They received messages for 3 weeks. Within the group, participants were asked to reflect on their experiences with, impressions of, and suggestions for the Small Steps SMS program and similar programs.

Small Steps SMS is an automated text messaging intervention to support self-management of depression and anxiety symptoms [46-48,57]. Small Steps SMS was designed through a human-centered design process encompassing elicitation activities, co-design workshops, prototype usability testing, and interviews [46-48], leading to a focus on sustaining engagement through (1) delivering diverse interaction types (eg, peer stories, reflection questions, psychoeducation, and action prompts) and (2) supporting low-stake experimentation with a variety of evidence-based self-management strategies (eg, thought challenging, behavioral activation, valued living, and self-compassion). Dialogues are interactive, with a subset of messages asking participants to respond with their feedback, choices, and experiences. Figures 1 and 2 show example interactions with Small Steps SMS. The number of daily messages sent to each participant was dependent on user responses, such that participants who were more engaged with the program received more messages. Participants were not required to respond to messages from Small Steps SMS and received no compensation for engaging with the program. They could send the word “stop” at any time to end all messages. Responses to the program were monitored daily for indications of safety risks. No such risks emerged, and no follow-up was conducted.

Figure 1. Example of interactive SMS text messaging dialogue with Small Steps SMS focused on behavioral activation.

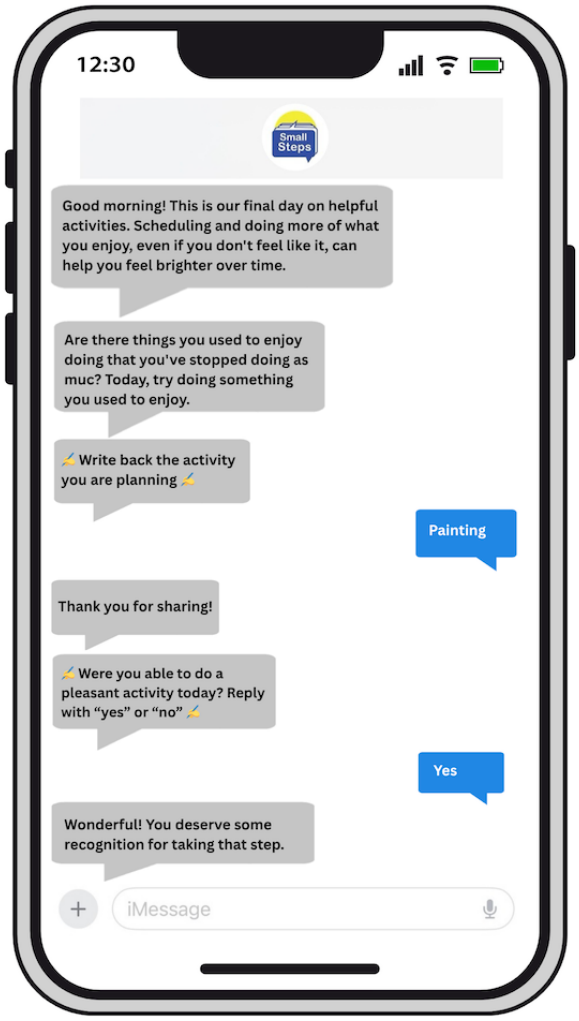

XSL•FO
RenderX

Figure 2. Example of interactive SMS text messaging dialogue with Small Steps SMS focused on self-compassion.

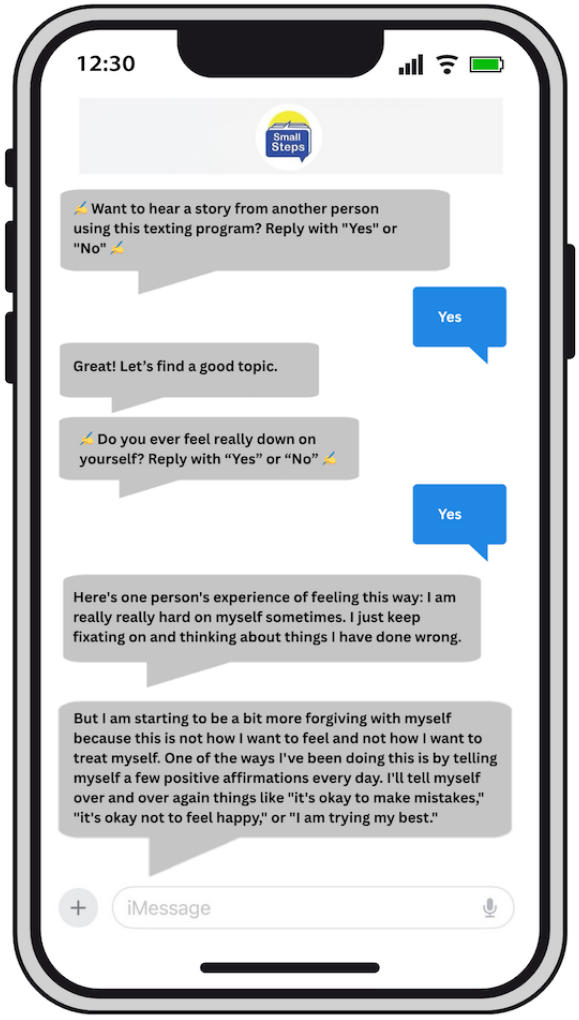

Interviews

Participants were invited to complete optional semistructured interviews with researchers after discussion group activities concluded. The main purpose of the interviews was to allow participants to share detailed feedback on the researcher prompts and Small Steps SMS that may not have been covered in the discussion group responses. Interviews lasted approximately 20 to 25 minutes and were conducted via the Zoom teleconferencing platform (Zoom Communications, Inc). Participants could turn their video on or off based on their preference. Interviews were audio recorded and transcribed.

Ethical Considerations

All study procedures were approved by the Institutional Review Board at Northwestern University (STU00211168). Participants provided informed consent before the commencement of study activities. They were informed that they could skip any questions they liked or end their participation. All study data were deidentified before analysis. Participants were compensated for research activities. In the discussion group, responding to each researcher-posted discussion prompt was compensated with US $7. For each prompt, responding to at least one other participant was compensated with US $2. Thus, participants could earn up to US $9 per prompt and US $72 total for participating in the discussion group, which was delivered in the form of a gift card upon completion of the study. Participants could earn an additional US $8 to complete an interview (in the form of a gift card).

Data Analysis

Deidentified transcribed data from the discussion group and interviews were subjected to thematic analysis [58]. Such an approach allows for inductive identification of themes and corresponds to the lack of established theory related to the research topic [59]. The lead author, who is an experienced mixed methods researcher, trained the other 3 coders, who were research staff, in qualitative methods. These 4 coders, who are also authors of this paper, first became familiar with the data by reading the discussion group and interview transcripts. They performed open coding to identify preliminary codes and met to discuss and prioritize these codes, guided by each code's relevance to the research questions. Prioritized codes and their definitions were captured in a shared codebook. The coders then applied the codebook to overlapping transcripts using the qualitative data analysis software Dedoose (SocioCultural Research Consultants, LLC) and subsequently met to discuss and resolve coding discrepancies and refine and consolidate the codebook. Approximately 4 overlapping coding rounds were completed before discussions ceased to yield codebook revisions. Three of the coders then divided the uncoded

XSL•FO
RenderX

transcripts and applied the final codebook. Key themes were identified that encompassed the coded data; these were discussed and refined with the larger authorship team, and excerpts were selected to illustrate each theme.

Results

Overview

In total, 47 individuals initiated the screening process. Of these, 22 (47%) individuals were excluded because they did not meet 1 or more of the eligibility criteria. Excluded individuals were not aged 18 years (n=11, 50%), were in mental health treatment (n=8, 36%), reported serious mental illness (n=6, 27%), or reported severe suicidality (n=2, 9%). In total, 4 (9%) of the 47 individuals did not complete the screening process. A total of 21 (45%) individuals were deemed eligible and invited to read and sign the consent form, of whom 17 (81%) did. Of these 17 individuals, 14 (82%) completed the required steps to enroll in our study, and, of these individuals, 12 (86%) engaged in the study's discussion group and are included in our analyses.

Table 1 presents participant demographics, treatment history, and symptom severity. Participants were mostly women, were racially and ethnically diverse, and half (6/12, 50%) of them reported previous mental health treatment. The current symptoms of both anxiety and depression were moderate, on average.

Table 1. Demographics, mental health history, and symptom levels of participants aged 18 years included in qualitative human-centered design activities to improve digital mental health interventions.

Characteristic	Discussion group (n=12)	Interview (n=4)
Gender, n (%)		
Woman	8 (67)	3 (75)
Man	3 (25)	1 (25)
Nonbinary	1 (8)	0 (0)
Race, n (%)		
African American	3 (25)	2 (50)
Asian	3 (25)	1 (25)
White	3 (25)	1 (25)
Multiracial	3 (25)	0 (0)
Ethnicity, n (%)		
Hispanic or Latinx	4 (33)	1 (25)
Not Hispanic or Latinx	8 (67)	3 (75)
Mental health treatment history, n (%)		
Previously consulted a mental health professional	6 (50)	1 (25)
Previously received a prescription for a mental health condition	3 (25)	1 (25)
Depression and anxiety symptom scores, mean (SD)		
PHQ-9[a]	13.42 (4.25)	13.00 (5.60)
GAD-7[b]	12.83 (3.79)	13.25 (5.06)

[a]PHQ-9: 9-item Patient Health Questionnaire.

[b]GAD-7: 7-item Generalized Anxiety Disorder.

On average, participants responded at least once to 5.92 (SD 2.43) of the 8 researcher-posted prompts. They responded an average of 18.08 (SD 11.54) times (including direct responses to prompts and replies to one another), contributing an average of 1238 (SD 839.79) words across all responses.

Our study included adolescents with any previous experiences of depression or anxiety; however, most participants reported mental health symptoms that were current. Participants' symptoms varied in duration; some had experienced chronic and persistent symptoms lasting months or years, while for others, symptoms had only manifested recently. Participants identified several key factors that they associated with their mental health challenges, including the life transition from adolescence to young adulthood, uncertainty in their own future, and the belief that the world is unstable. Participants also shared their perspectives on the influence of technology in mental health, including how technology can be both helpful and harmful for mental health, and its impact on mental health literacy and stigma. Finally, participants shared their perspectives on DMHIs and how these tools, including Small Steps SMS, can be optimized to address the specific mental health challenges they face and to better engage them.

Factors Influencing Mental Health

Participants reported experiencing a major life transition in late adolescence and shared that certain factors encountered during this transition may have contributed to the worsening of their

XSL•FO
RenderX

mental health. One factor highlighted was the pressure to manage an overwhelming number of tasks, all at once, which included school-related responsibilities, commitments to extracurriculars, and employment. This was described by one participant as follows:

> *I feel like things that affect this generation is like stress, mainly I feel like there is a lot of expectations being pushed. And you try to do your very best, and sometimes people want to work, but we're also expected to do that. And sometimes you have extracurriculars and so, you have to do all that.* [P6]

Participants also reported feeling overwhelmed by a growing number of responsibilities after graduating high school and transitioning to college, as stated by one participant as follows:

> *So, coming to college is a really big culture shock because you're so—you're responsible for everything. Everything's on you now. Everything bad that happens, you have to ask yourself, oh, was it my fault? Did I miss something?* [P7]

Participants indicated feeling scared, exhausted, and unmotivated in response to changes in their daily routines, which they perceived as rapid and drastic. Overall, despite acknowledging that the transition from adolescence to adulthood is characterized by both positive and negative aspects, they described their own experiences to be jarring. One participant reflected as follows:

> *I graduated this summer and now that it's almost fall I'm feeling weird about how life is going to be so different now because it was basically the same thing for so many years of my life. I wish I had more time to adjust to that before I'm shoved into "real life." It's kinda scary to be honest.* [P1]

Challenges in transitioning from adolescence to adulthood also extended to those who did not enroll in college:

> *One of the biggest challenges I'm currently facing is with moving out to live on my own and whether or not I want to move on to college.* [P2]

Feeling socially excluded during the transition from adolescence to adulthood was identified as a significant trigger for symptoms of depression and anxiety, and many participants reported struggling with these types of social challenges:

> *The biggest challenge in my life is adjusting to university with a very much lively social scene I'm not used to. It's hard adapting right now where everyone seemingly already have found their crowd of people to rely on.* [P11]

Similarly, a participant considered themself to have greater social challenges than their peers, partly because they were a commuter student. They shared the following:

> *I feel pretty left out while everyone seems to be getting together really well. And that spiralled into me being sad and replaying a bunch of moments from last year that I still haven't moved on from yet...I start feeling out of place, feeling very tired, and my body feels very irritable.* [P5]

Many participants also identified uncertainty about the future as a factor that contributed to their experiences of anxiety and depression. Participants worried about their ability to successfully manage both short- and long-term goals:

> *The biggest thing in life that's making me depressed and anxious is my future. I'm worried about what if I can't make it through my courses and I fail. Will I be able to make a living?...Will I ever be able to actually find someone with the way dating culture is these days? Basically everything about the future and it's exhausting. I'd just like to be able to live my life in the moment and enjoy it instead of constantly worrying about basically everything and not feeling so tired and not like doing anything.* [P1]

Finally, participants identified the belief that they were coming of age in an unpredictable world as a contributor to their mental health issues. For example, one participant shared the following:

> *With the state of the world these days, it's really hard to think positively about the future.* [P9]

Participants' fears that the world is unpredictable were based, in part, on their experiences of growing up during the COVID-19 pandemic. They believed the pandemic had long-lasting effects on their mental health, disrupting the transition from adolescence to adulthood, which was explained by one participant as follows:

> *...a lot of people in this young generation are stuck in their pre pandemic life where they were carefree and dependent on their parents. Counseling therapy should consider the fact that the teens from this generation need time to adjust to their new age and responsibilities.* [P8]

Overall, participants associated the transition from adolescence to adulthood with the experience of depressive and anxious symptoms. These challenges were further exacerbated by participants' perceptions of an unpredictable world and an uncertain future.

Perceptions of the Role of Technology in Mental Health

Overview

Participants perceived technology to be deeply embedded in their daily lives and often described it as a source of entertainment and distraction. For many participants, technology offered an escape from daily stressors and was considered a useful coping tool; however, some relayed that overreliance on technology could reinforce negative habits (eg, self-isolation) and worsen mental health symptoms. Despite reporting ambivalence regarding the impact of technology on mental health, participants believed that technology played a significant role in enhancing mental health awareness and reducing stigma, particularly within their own generation.

The Dual Impact of Technology on Mental Health: Helpful and Harmful

Participants reported relying on technology as a coping tool when dealing with acute mental health challenges. Their digital devices were described as a means of escape when they were

XSL•FO
RenderX

feeling mentally unwell, allowing them to seek social support and connection. One participant shared the following:

> *Technology connects us to people all over the world and gives people a sense a community, which I think is important for mental health or even just for having people to reach out to when you need them.* [P2]

Most participants discussed having used the internet as a tool to learn more about their mental health symptoms and others' experiences. Some had also disclosed their experiences of depression and anxiety online to others, sought to understand available treatment and support options, or used mobile phone apps for mental health and well-being.

Some individuals also reported that social media, specifically, served as a useful coping tool for social anxiety, and attributed its usefulness to social norms that sanction the use of cell phones in public. One participant described the following:

> *It's a good distraction sometimes when you're, I guess, alone. Like, let's say you're in a public space, and you feel alone. You can just go on social media and just be in your own pot, and just scroll through it or whatever.* [P4]

Many participants recognized that technology has a dual impact on mental health, noting that it has the potential to not only improve but also worsen symptoms of depression and anxiety. Participants reported beliefs that using technology as a mental health coping tool could encourage avoidant behaviors and potentially exacerbate symptoms. P10 summarized this concern, sharing that they frequently turned to technology to escape their problems, but technology did not address the root of these problems. They summarized that technology "hurts me more than it helps, I end up spending too much time on reels and ignoring the issues completely."

Similarly, one participant described the following:

> *I think I fell victim to using technology as a distraction all the time. But even I know that technology, like social media doesn't fix everything and can even make it feel worse.* [P7]

One participant also questioned whether using technology in place of traditional social engagement could have unintended negative consequences:

> *But then, on the other side, I think it can be very harmful because of the things that people use technology for. Social media can be very harmful. And I feel like some people can use technology as a way to avoid people.* [P7]

While participants recognized that technology could temporarily help them escape their problems, they recognized its potential to serve as an unhealthy coping mechanism that could prevent them from addressing the root causes of their mental health issues and engaging in more effective treatment strategies.

Reducing Mental Health Stigma Using Technology

Participants strongly advocated for mental health awareness and resources, emphasizing that mental health concerns should be taken seriously and that access to treatment should be readily available. They reported observing a generational gap in mental health awareness, believing that younger generations possess a greater understanding of mental health issues compared to older generations. One participant shared the following:

> *I think mental health issues are so normalized with younger people compared to older people...I think younger people are more understanding and open to talking about mental health issues because they've been exposed to it more often.* [P7]

Participants attributed generational differences in mental health awareness largely to the rise of the internet, and specifically social media, stated by one participant as follows:

> *People my age grew up with social media and the Internet as it was forming, so naturally we were a lot more exposed to mental health issues (both good and bad) than those who are older.* [P2]

Thus, many participants felt that technology had benefited their generation by raising mental health awareness, which enabled them to openly discuss mental health issues with others and seek help when needed; however, many participants reported fears that their mental health issues would be dismissed or invalidated by members of older generations. One participant described the following:

> *Older generations, in my experience, tend to brush off any negative emotions or feelings as being weak or made up for attention. This has to do with the lack of research and widespread knowledge of mental health issues at the time. So, now, we see a lot of older people tend to say they're fake simply because it's newer information, and they still need time to adjust to it being the norm.* [P9]

Similarly, a participant stated the following:

> *Younger people may just be told they're being dramatic or oversensitive or to just get over it or deal. Or that it will pass, you'll grow out of it etc.* [P1]

Lower mental health awareness among older generations was perceived to be a barrier to mental health treatment by participants, in part because adolescents often must go through a parent, teacher, or other adult to access mental health resources. One participant shared some of their concerns about discussing mental health treatment with a parent:

> *Well, personally, I would feel uncomfortable because my dad would be like, why are you doing that? What do you need it for? He's just gonna ask a lot of questions that I'm just gonna pause it, and don't know the answer to any of it.* [P4]

Some participants viewed technology as an opportunity to bridge gaps in mental health awareness and facilitate intergenerational conversations on mental health topics; however, those who lacked confidence in involving adults in their mental health decision-making were more likely to report an interest in seeking resources independently. Overall, participants were interested in mental health solutions that could help them treat symptoms of depression and anxiety without stigma or judgment.

XSL•FO
RenderX

Participants' preferences for how a digital tool might meet their needs are described in the next section.

SMS Text Messaging as a Tool for Mental Health Support

Endorsement of the Program

Adolescents in our study expressed interest in digital tools for mental health self-management, and most engaged actively with the Small Steps SMS program. All participants responded at least once to the program, with the average number of responses over the 3-week program being 28.66 (SD 25.71).

In the discussion group, participants shared that they perceived the Small Steps SMS program positively, describing it as convenient, accessible, and familiar due to its use of SMS text messaging. One participant explained as follows:

> *I definitely think the fact that it's a texting program is a factor that would make people my age sign up because it's low pressure and easy to fit in when you can.* [P1]

Participants believed that Small Steps SMS could be easily integrated within their daily routine because they could engage with it on their own time and in their own space. Others shared that SMS text messaging reduced some of the anxiety and vulnerability that can arise face-to-face. One participant stated the following:

> *It also feels like there's much less pressure because it's a texting program and not face to face/physically talking to someone.* [P3]

One participant also provided their perspective on the sense of privacy that texting provides:

> *I think a text service would be very helpful for when people are in public and they're dealing with a mental health issue or if they are in places where they don't want their mental health to be aired out for everyone to see.* [P7]

Addressing Adolescent-Specific Issues in Program Content

While participants held favorable perceptions of Small Steps SMS, they believed that the program could be improved for older adolescents. Participants thought that Small Steps SMS and other similar programs should tailor their content to the major life changes that their age group faces, which they had described in response to earlier prompts (eg, including the transition to college, moving away from home, or joining the workforce). One participant stated the following:

> *...adding more content regarding the life changes some people in the group might be going through and how to get through the mental barriers they are going through with the changes.* [P12]

Similarly, one participant emphasized the importance of including content tailored to the transition to college, which they identified as a particularly challenging life change:

> *I think it would be really helpful, especially in the first few months of college, when you're still trying to get to know people. You're still trying to figure out how to be an adult, how to be on your own, how to deal with homesickness, and stuff like that.* [P7]

Aside from major life transitions, some participants highlighted how parents, teachers, and other adults can impact their ability to seek help for mental health symptoms and advocated for a program that could help them navigate these intergenerational challenges. For example, one participant imagined a program that could help younger generations communicate their mental health needs to older generations:

> *For people my age, I feel like it would be more geared towards helping them find support and being comfortable talking about it with their parents.* [P12]

Reimagining Engagement Strategies for Adolescent Users

In addition to adapting the content within SMS text messages, participants identified ways that Small Steps SMS and related programs could improve the user experience and increase engagement among adolescents. These suggestions included gamification, opportunities for peer-to-peer communication, and more scaffolding of self-management strategies.

Several participants thought that Small Steps SMS and similar self-management interventions could benefit from the inclusion of game-like components. For example, one participant stated the following:

> *I think a fun and interactive way to make a tool for people my age (if I had the skill to make it ha ha) would be a mobile game where you do tasks to get points and earn rewards...After a certain amount of points, you get to pick a charity to donate some money to. It could also do e-gift cards as rewards as well.* [P3]

Participants theorized that the addition of more tangible rewards would increase engagement among adolescents. One participant provided additional guidance on how these types of rewards could be incorporated within programs that suggest daily self-management tasks:

> *I think a tool that people my age would like would be an app or game that interests them and gives them in-game rewards for different things they do. This would be able to ask them different prompts about how they're feeling while not making it all about mental health. With this, it can give them the opportunity to be honest with how they're feeling while also making sure that the thing that they're using isn't fully about how they're feeling.* [P12]

Overall, the possible addition of gamified elements was identified as an opportunity for DMHIs to incorporate an element of fun and allow individuals to see beyond their mental health issues while also addressing symptoms.

Moreover, participants thought the program could benefit from the addition of peer-to-peer communication. Many participants indicated a strong interest in communicating with other adolescents who experienced similar challenges. One participant explained the following:

XSL•FO
RenderX

> *I think that people my age would like being in a program like Small Steps because they can talk with people near their age going through the same life changes as they are and can experience it together and we can help each other through the mental health part of things.* [P12]

Participants particularly valued opportunities to connect over similar hobbies and interests. They saw these connections as a pathway to build community and reduce the stigma of mental health issues.

Finally, some participants thought that digital self-management interventions, such as Small Steps SMS, should incorporate more technical guidance as well as detailed instructions that break down how to use each psychological strategy introduced in the content. One participant stated the following:

> *[Mental health interventions] should have an easy-to-use interface with clear directions on how to use it...I'd add in videos or something that would give step by step guides to some of the techniques.* [P1]

Similarly, a participant shared the following:

> *I agree with adding more direction to the messages. I feel the messages could be a bit more substantial.* [P6]

Thus, presenting program content alongside gamified tasks, integrating peer-to-peer communication, and greater support for learning psychological strategies were seen as ways to increase engagement and satisfaction among these older adolescents. Participants noted that such adaptations could improve Small Steps SMS and also endorsed such strategies as broadly applicable to a range of automated DMHIs that support adolescents in self-managing mental health symptoms.

Discussion

Principal Findings

This study sought to understand how to design and adapt DMHIs to address the mental health needs identified by older adolescent participants. We focus on automated interventions, including the Small Steps SMS program, which delivers self-management support through daily interactive SMS text messages. Through an online discussion group and series of interviews, non–treatment-engaged participants aged 18 years identified key mental health challenges and barriers to effective coping that are tied to their transition from adolescence to young adulthood and discussed ways that DMHIs could address them. The transition from high school to college or the workforce was regarded as highly influential for participants' mental health. They also viewed their mental health symptoms as related not only to being young but also to being young in the present moment, pointing to the perception that the world is becoming increasingly unstable and inhospitable. They highlighted the complex influence of technology and social media on young people, contending that the impact of these technologies on an individual's mental health depends on how they are used, making them either beneficial or harmful. As a potential solution to address their mental health concerns, participants endorsed Small Steps SMS as helpful, describing its automated approach to be a nonstigmatizing and low-burden way to learn self-management strategies while avoiding disclosure and possible dismissive reactions from parents and other adults. SMS text messaging (as a medium) was viewed as convenient and familiar. They suggested several changes to Small Steps SMS and considerations for DMHIs generally, highlighting that these tools should address the specific challenges faced by older adolescents, including the disruption while transitioning out of high school and navigating intergenerational conversations about mental health. They also thought that programs could be structured to further engage them through game-like elements, peer-to-peer contact, and more detailed guidance on self-management practices.

Comparison With Prior Work

Given the extent of unmet mental health needs in the United States across demographic groups, many DMHIs are designed to be suitable across various individuals in need [39-41]; however, it is increasingly recognized that programs may be more appealing and potentially more effective when designed for narrower segments of the population [60,61]. This can be achieved either through programs that are targeted (ie, making them available to certain segments of the population) or tailored (ie, where content or components are delivered based on users' assessed characteristics) [61]. When targeting and tailoring DMHIs, designers must weigh the potential benefits of focusing on particular constituent groups against the resources required to do so [62]. Past work has developed programs aligned with the needs of racial and ethnic minoritized groups [61], those with limited resources or specific challenges (eg, homelessness and trauma) [63,64], and age ranges, including some that focus on adolescents [32]. Subsequently, we interpret our findings in relation to the larger body of work on the mental health challenges faced by adolescents and discuss approaches to adapting digital tools to address these challenges.

For young people, past work has focused on developing programs that reflect their distinct mental health needs, including through considerations of mental health literacy, differences in technology use, and settings for disseminating and implementing digital mental health tools (eg, schools, pediatric care, or public services such as libraries) [41,65,66]. Our findings are consistent with this work in suggesting that our participants wanted program content to address the specific concerns and challenges faced by their age group. Participants particularly wanted validation and guidance related to the transition from adolescence to adulthood, which was experienced as highly overwhelming and disruptive to their mental health. Corresponding to this finding, previous literature on late adolescence has also described rapid and disorienting transitions in social standing and from familial dependence to independence [3,67].

Furthermore, participants wanted tools that they could access independently, partly reflecting their perception of older generations as less aware and accepting of mental health challenges. Participants often thought that their mental health problems would be trivialized because of their age and avoided help-seeking through parents or other adults of older generations. Relatedly, some recent work has proposed a "prevalence

XSL•FO
RenderX

inflation hypothesis," wherein exposure to well-intended mental health awareness campaigns on social media and in schools and other settings could drive young people's inaccurate self-diagnoses or potentially exacerbate their symptoms [68]. While our findings cannot speak to the validity of this hypothesis, they suggest that our adolescent participants perceive skepticism about their mental health concerns to be widespread, and they find such views invalidating and stigmatizing. To avoid unsupportive responses to mental health disclosures, they were interested in accessing treatment on their own, although some also wanted support in raising or discussing mental health issues with parents or guardians. Some related work has explored technology as a tool for brokering intergenerational communication, including facilitating parent-child shared experiences, conversations about health topics or emotions between parents and children, in-the-moment parenting support, and self-monitoring family functioning [69,70]. Opportunities may also include dyadic interventions (ie, those that help the adolescent while also offering parallel guidance for the parent in how to support their child with a mental health issue), providing opportunities to rehearse mental health disclosure and build communication skills, and educational approaches that target parents' mental health stigma.

Our findings also suggest that these participants see their generation's acceptance of mental health issues as an important strength. In this regard, strength-based interventions, which are conceived to emphasize and understand the experiences, values, perspectives, and strengths of systemically marginalized people [71], may be useful for adolescents (who tend to identify closely with their generation). Participants noted that they are united not just as young people but as young people coming of age after the COVID-19 pandemic who are witnessing historic levels of social and technological change as they enter adulthood. Some strength-based interventions focus on individuals' strengths, such as one digital tool designed for transition-age youth on the autism spectrum that supports users in identifying their strengths and expressing those strengths to others [72]. Other strength-based approaches can build an empowering collective narrative (eg, around experiences as refugees) [73]. Helping adolescents claim and take pride in shared traits, such as openness and acceptance, may be a way to support adolescents through a sense of community and collective action. This may be particularly important given that participants felt their age group was misunderstood and disparaged.

Our data also speak to the deep enmeshment of technology in participants' daily lives and the dual role of technology in mental health. Potential negative effects of technology on the mental health of adolescents have been of great public concern, with the US Surgeon General recently issuing a warning about the negative effects of social media use in adolescence [74]. The extent and nature of these effects have been debated, as the research literature has yielded complex and sometimes contradictory findings [75-77]. However, less work has examined how adolescents themselves perceive and balance the potential benefits and harms of technology. Our findings suggest considerable nuance in how these older adolescent participants think about and use technology. Among our participants, technology was often used as a coping tool, sometimes successfully. Many described how much they had benefited from the social awareness and destigmatization that the internet had fostered related to mental health, and their regular use of mobile phones was key to achieving continuous reach of DMHIs, facilitating engagement in mental health self-management. However, they also recognized important harms of technology, particularly social media. Often used to cope with mental health symptoms, this use could backfire and lead to loss of productivity, avoidance, and isolation. Importantly, participants did not view all uses of technology as harmful but acknowledged variability across types of technology (eg, social media vs texting) and motivations for use (eg, avoidance vs seeking mental health information). This is consistent with past work describing distinctions between passive use of social media (eg, browsing feeds), which may have negative effects on mental health, whereas more active and targeted use (eg, direct communication with close ties) may have benefits [78,79]. Participants drew on their deep experience using apps and social media sites to suggest engagement strategies that might better sustain their use of DMHIs, emphasizing peer-to-peer communication and gamification—approaches which have also been endorsed by adolescents in previous literature [32,33,80].

Implications for Automated Messaging Interventions

These findings suggest specific modifications and new features for Small Steps SMS and related DMHIs. Participants' feedback suggests that interactive SMS text messaging is a promising pathway to reach those in their age range and also that content should be extended and refined, and new engagement strategies should be considered. As far as content, programs for older adolescents could be refined to directly address topics, such as the disruptions in completing high school and moving on to college or the workforce, intergenerational conversations about mental health, and redefining one's relationship with social media. Adaptations to Small Steps SMS may include new modules or tailored examples and information when introducing psychological strategies (eg, existing "help-seeking" content could be expanded to provide guidance on approaching parents for mental health support). As participants valued peer-generated content, these issues may also be addressed through crowdsourced peer stories from other adolescents showing how they used a psychological strategy to address common challenges such as isolation, feeling overwhelmed, or invalidation. Participants also suggested new features for Small Steps SMS and related DMHIs. Some wanted detailed, step-by-step instructions on using psychological strategies, which might take the form of supplementary interactive guidance that can be launched at the user's request (eg, through a text-based command such as "example" or "expand"). Others highlighted direct peer-to-peer communication, which could be facilitated through a web-based asynchronous discussion group parallel to the texting program or potentially through the program connecting peers to one another for direct text chat, as has been used in other texting interventions [81]. Participants also suggested game-like elements. In the context of Small Steps SMS, gamification could involve adding points or other rewards that can be collected by completing the daily tasks or activities suggested by the program or achieving streaks (eg, completing

XSL•FO
RenderX

3 behavioral activation tasks over 3 days). Before evaluating adolescent-targeted versions of the program, new content and prototypes of new features can be refined through feedback from users as part of our ongoing human-centered design process.

Limitations

Adolescence spans an age range from 10 to 19 years [1], and adolescents typically progress through profound changes in physical maturity, social roles, sense of self and identity, and emotional and intellectual development [2-5,67,82]. Our study focused on a relatively narrow subgroup of older adolescents—those who are aged 18 years. We prioritized direct access to participants who are legal adults because parental involvement (eg, parental consent) can act as a barrier to help-seeking and potentially to mental health research participation [22,23]. Consequently, our study is limited in what we can conclude about the needs of younger adolescents. While we asked participants to reflect on their recent experiences in high school and comment on how younger adolescents may receive the Small Steps SMS program, we note the importance of further research prioritizing those aged <18 years.

In addition, although sample sizes are often small in human-centered design work [83] and small group sizes for online discussion groups are recommended to increase participants' comfort and engagement [84], we acknowledge that using additional methods and recruiting a larger sample may be helpful to triangulate findings related to the mental health challenges adolescents face and their relationship with technology. Our qualitative findings can suggest potential hypotheses and theories to be evaluated using quantitative methods [59]. Finally, while we recruited a diverse group of participants in regard to race and ethnicity, their recruitment online and the study's focus on technology-based solutions may have contributed to overrepresenting those with interest in and access to technologies. Given the study's emphasis on SMS text messaging and the exposure to the Small Steps SMS program, it is also possible that participants were predisposed to come up with messaging-related solutions at the expense of other design directions.

Conclusions

Adolescents experience high levels of mental health issues but have limited access to treatment. We engaged older adolescents to better understand how to design digital mental health tools to address their mental health challenges and priorities. Participants emphasized that their mental health had been negatively and severely impacted by stressful life transitions and expressed pessimism about the future. They characterized a complex role played by technology, with both negative and positive effects on mental health. Technology, especially social media, could contribute to mental health issues, but the internet was also seen as a potential avenue for seeking support and reducing stigma among people of their generation. Given their regular use of smartphones and texting, participants were receptive toward automated SMS text messaging programs for mental health self-management. They appreciated that such programs provide a low-burden way to access support without the need to disclose their mental health issues to others, such as parents. Participants recommended that DMHIs include content relevant to the challenges of late adolescent life transitions (eg, high school to college, high school to workforce, and familial dependence to independence) and voiced desires for game-like engagement features and opportunities to communicate with peers. Scalable digital mental health tools, such as Small Steps SMS, may help to reduce the mental health treatment gap for adolescents if designed to meet adolescents' unique needs.

Acknowledgments

The authors were supported by grants from the National Institute of Mental Health (K01MH125172, R34MH124960, P50MH119029, K01MH121854, and K08MH128640). In addition, a gift from the Microsoft AI for Accessibility program to the Center for Behavioral Intervention Technologies (in part) supported this work [85].

Data Availability

The datasets generated or analyzed during this study are not publicly available due to the sensitive nature of text-based data and the potential risk of reidentification, but deidentified data are available from the corresponding author on reasonable request.

Conflicts of Interest

RK has accepted consulting fees and honoraria from CHESS Health and Jimini Health. DCM has accepted honoraria and consulting fees from Boehringer Ingelheim, Optum Behavioral Health, Centerstone Research Institute, Woebot Health, and the One Mind Foundation; royalties from Oxford Press; and has an ownership interest in Adaptive Health, Inc. JM has accepted consulting fees from Boehringer Ingelheim and Shirley Ryan AbilityLab. All other authors declare no conflicts of interest.

Multimedia Appendix 1
Online discussion group prompts used to collect feedback from older adolescent participants on digital mental health intervention design.
[DOCX File , 34 KB - formative_v9i1e68950_app1.docx]

XSL•FO
RenderX

References

1. Mental health of adolescents. World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/adolescent-mental-health [accessed 2024-07-08]
2. Casey BJ, Jones RM, Levita L, Libby V, Pattwell SS, Ruberry EJ, et al. The storm and stress of adolescence: insights from human imaging and mouse genetics. Dev Psychobiol 2010 May 10;52(3):225-235 [FREE Full text] [doi: 10.1002/dev.20447] [Medline: 20222060]
3. Sawyer SM, Azzopardi PS, Wickremarathne D, Patton GC. The age of adolescence. Lancet Child Adolesc Health 2018 Mar;2(3):223-228. [doi: 10.1016/S2352-4642(18)30022-1] [Medline: 30169257]
4. Somerville LH. Special issue on the teenage brain: sensitivity to social evaluation. Curr Dir Psychol Sci 2013 May 01;22(2):121-127 [FREE Full text] [doi: 10.1177/0963721413476512] [Medline: 24761055]
5. Rapee RM, Oar EL, Johnco CJ, Forbes MK, Fardouly J, Magson NR, et al. Adolescent development and risk for the onset of social-emotional disorders: a review and conceptual model. Behav Res Ther 2019 Dec;123:103501. [doi: 10.1016/j.brat.2019.103501] [Medline: 31733812]
6. Panchal U, Salazar de Pablo G, Franco M, Moreno C, Parellada M, Arango C, et al. The impact of COVID-19 lockdown on child and adolescent mental health: systematic review. Eur Child Adolesc Psychiatry 2023 Jul 18;32(7):1151-1177 [FREE Full text] [doi: 10.1007/s00787-021-01856-w] [Medline: 34406494]
7. O'Keeffe GS. Social media: challenges and concerns for families. Pediatr Clin North Am 2016 Oct;63(5):841-849. [doi: 10.1016/j.pcl.2016.06.009] [Medline: 27565362]
8. Results from the 2019 national survey on drug use and health: detailed tables. Center for Behavioral Health Statistics and Quality. 2019. URL: https://catalog.data.gov/dataset/2019-nsduh-annual-national-report [accessed 2024-07-10]
9. Daly M. Prevalence of depression among adolescents in the U.S. from 2009 to 2019: analysis of trends by sex, race/ethnicity, and income. J Adolesc Health 2022 Mar;70(3):496-499 [FREE Full text] [doi: 10.1016/j.jadohealth.2021.08.026] [Medline: 34663534]
10. Major depression. NIMH Information Resource Center. URL: https://www.nimh.nih.gov/health/statistics/major-depression [accessed 2022-07-15]
11. Madigan S, Racine N, Vaillancourt T, Korczak DJ, Hewitt JMA, Pador P, et al. Changes in depression and anxiety among children and adolescents from before to during the COVID-19 pandemic: a systematic review and meta-analysis. JAMA Pediatr 2023 Jul 01;177(6):567-581 [FREE Full text] [doi: 10.1001/jamapediatrics.2023.0846] [Medline: 37126337]
12. Husky M. Help-Seeking. In: Levesque RJ, editor. Encyclopedia of Adolescence. Cham, Switzerland: Springer; 2016:1-8.
13. Leavey G, Rothi D, Paul R. Trust, autonomy and relationships: the help-seeking preferences of young people in secondary level schools in London (UK). J Adolesc 2011 Aug 16;34(4):685-693. [doi: 10.1016/j.adolescence.2010.09.004] [Medline: 20952053]
14. Sawyer MG, Borojevic N, Ettridge KA, Spence SH, Sheffield J, Lynch J. Do help-seeking intentions during early adolescence vary for adolescents experiencing different levels of depressive symptoms? J Adolesc Health 2012 Mar;50(3):236-242. [doi: 10.1016/j.jadohealth.2011.06.009] [Medline: 22325128]
15. Merikangas KR, He J, Burstein M, Swendsen J, Avenevoli S, Case B, et al. Service utilization for lifetime mental disorders in U.S. adolescents: results of the National Comorbidity Survey-Adolescent Supplement (NCS-A). J Am Acad Child Adolesc Psychiatry 2011 Jan;50(1):32-45 [FREE Full text] [doi: 10.1016/j.jaac.2010.10.006] [Medline: 21156268]
16. Radez J, Reardon T, Creswell C, Lawrence PJ, Evdoka-Burton G, Waite P. Why do children and adolescents (not) seek and access professional help for their mental health problems? A systematic review of quantitative and qualitative studies. Eur Child Adolesc Psychiatry 2021 Mar 21;30(2):183-211 [FREE Full text] [doi: 10.1007/s00787-019-01469-4] [Medline: 31965309]
17. Aguirre Velasco A, Cruz IS, Billings J, Jimenez M, Rowe S. What are the barriers, facilitators and interventions targeting help-seeking behaviours for common mental health problems in adolescents? A systematic review. BMC Psychiatry 2020 Jul 11;20(1):293 [FREE Full text] [doi: 10.1186/s12888-020-02659-0] [Medline: 32527236]
18. Gulliver A, Griffiths KM, Christensen H. Perceived barriers and facilitators to mental health help-seeking in young people: a systematic review. BMC Psychiatry 2010 Dec 30;10:113 [FREE Full text] [doi: 10.1186/1471-244X-10-113] [Medline: 21192795]
19. Roberts M, Jones J, Garcia L, Techau A. Adolescents' perceptions of barriers and facilitators to engaging in mental health treatment: a qualitative meta-synthesis. J Child Adolesc Psychiatr Nurs 2022 May 04;35(2):113-125. [doi: 10.1111/jcap.12354] [Medline: 34609034]
20. Sylwestrzak A, Overholt CE, Ristau KI, Coker KL. Self-reported barriers to treatment engagement: adolescent perspectives from the national comorbidity survey-adolescent supplement (NCS-A). Community Ment Health J 2015 Oct 19;51(7):775-781. [doi: 10.1007/s10597-014-9776-x] [Medline: 25326732]
21. Kerwin ME, Kirby KC, Speziali D, Duggan M, Mellitz C, Versek B, et al. What can parents do? A review of state laws regarding decision making for adolescent drug abuse and mental health treatment. J Child Adolesc Subst Abuse 2015;24(3):166-176 [FREE Full text] [doi: 10.1080/1067828X.2013.777380] [Medline: 25870511]

XSL•FO
RenderX

22. Cavazos-Rehg P, Min C, Fitzsimmons-Craft EE, Savoy B, Kaiser N, Riordan R, et al. Parental consent: a potential barrier for underage teens' participation in an mHealth mental health intervention. Internet Interv 2020 Oct;21:100328 [FREE Full text] [doi: 10.1016/j.invent.2020.100328] [Medline: 32528858]
23. Logan DE, King CA. Parental facilitation of adolescent mental health service utilization: a conceptual and empirical review. Clin Psychol Pract 2001;8(3):333. [doi: 10.1093/clipsy/8.3.319]
24. AAP-AACAP-CHA declaration of a national emergency in child and adolescent mental health. American Academy of Pediatrics. URL: https://www.aap.org/en/advocacy/child-and-adolescent-healthy-mental-development/aap-aacap-cha-declaration-of-a-national-emergency-in-child-and-adolescent-mental-health/ [accessed 2024-08-12]
25. Wiederhold BK. The escalating crisis in adolescent mental health. Cyberpsychol Behav Soc Netw 2022 Mar 01;25(2):81-82. [doi: 10.1089/cyber.2022.29237.editorial] [Medline: 35072540]
26. Boden JM, Fergusson DM, Horwood LJ. Anxiety disorders and suicidal behaviours in adolescence and young adulthood: findings from a longitudinal study. Psychol Med 2007 Mar 17;37(3):431-440. [doi: 10.1017/S0033291706009147] [Medline: 17109776]
27. Copeland WE, Shanahan L, Costello EJ, Angold A. Childhood and adolescent psychiatric disorders as predictors of young adult disorders. Arch Gen Psychiatry 2009 Jul;66(7):764-772 [FREE Full text] [doi: 10.1001/archgenpsychiatry.2009.85] [Medline: 19581568]
28. Mojtabai R, Stuart EA, Hwang I, Eaton WW, Sampson N, Kessler RC. Long-term effects of mental disorders on educational attainment in the National Comorbidity Survey ten-year follow-up. Soc Psychiatry Psychiatr Epidemiol 2015 Oct 17;50(10):1577-1591 [FREE Full text] [doi: 10.1007/s00127-015-1083-5] [Medline: 26082040]
29. Owens M, Stevenson J, Hadwin JA, Norgate R. Anxiety and depression in academic performance: an exploration of the mediating factors of worry and working memory. Sch Psychol Int 2012 Feb 07;33(4):433-449. [doi: 10.1177/0143034311427433]
30. McGorry PD, Mei C. Early intervention in youth mental health: progress and future directions. Evid Based Ment Health 2018 Dec;21(4):182-184 [FREE Full text] [doi: 10.1136/ebmental-2018-300060] [Medline: 30352884]
31. Vogels EA, Gelles-Watnick R, Massarat N. Teens, social media and technology 2022. Pew Research Center. URL: https://www.pewresearch.org/internet/2022/08/10/teens-social-media-and-technology-2022/ [accessed 2024-10-05]
32. Lehtimaki S, Martic J, Wahl B, Foster KT, Schwalbe N. Evidence on digital mental health interventions for adolescents and young people: systematic overview. JMIR Ment Health 2021 May 29;8(4):e25847 [FREE Full text] [doi: 10.2196/25847] [Medline: 33913817]
33. Liverpool S, Mota CP, Sales CM, Čuš A, Carletto S, Hancheva C, et al. Engaging children and young people in digital mental health interventions: systematic review of modes of delivery, facilitators, and barriers. J Med Internet Res 2020 Jul 23;22(6):e16317 [FREE Full text] [doi: 10.2196/16317] [Medline: 32442160]
34. Farrer L, Gulliver A, Chan JK, Batterham PJ, Reynolds J, Calear A, et al. Technology-based interventions for mental health in tertiary students: systematic review. J Med Internet Res 2013 May 27;15(5):e101 [FREE Full text] [doi: 10.2196/jmir.2639] [Medline: 23711740]
35. Grist R, Porter J, Stallard P. Mental health mobile apps for preadolescents and adolescents: a systematic review. J Med Internet Res 2017 May 25;19(5):e176 [FREE Full text] [doi: 10.2196/jmir.7332] [Medline: 28546138]
36. Kruzan KP, Fitzsimmons-Craft EE, Dobias M, Schleider JL, Pratap A. Developing, deploying, and evaluating digital mental health interventions in spaces of online help- and information-seeking. Procedia Comput Sci 2022;206:6-22 [FREE Full text] [doi: 10.1016/j.procs.2022.09.081] [Medline: 37063642]
37. Czyz EK, Arango A, Healy N, King CA, Walton M. Augmenting safety planning with text messaging support for adolescents at elevated suicide risk: development and acceptability study. JMIR Ment Health 2020 May 25;7(5):e17345 [FREE Full text] [doi: 10.2196/17345] [Medline: 32160150]
38. Dobias ML, Morris RR, Schleider JL. Single-session interventions embedded within Tumblr: acceptability, feasibility, and utility study. JMIR Form Res 2022 Jul 25;6(7):e39004 [FREE Full text] [doi: 10.2196/39004] [Medline: 35877163]
39. Gan DZ, McGillivray L, Han J, Christensen H, Torok M. Effect of engagement with digital interventions on mental health outcomes: a systematic review and meta-analysis. Front Digit Health 2021;3:764079. [doi: 10.3389/fdgth.2021.764079] [Medline: 34806079]
40. Lattie EG, Stiles-Shields C, Graham AK. An overview of and recommendations for more accessible digital mental health services. Nat Rev Psychol 2022 Mar 26;1(2):87-100 [FREE Full text] [doi: 10.1038/s44159-021-00003-1] [Medline: 38515434]
41. Stiles-Shields C, Gustafson E, Lim P, Bobadilla G, Thorpe D, Summersett Williams FC, et al. Pre-implementation determinants for digital mental health integration in Chicago pediatric primary care. J Pediatr Psychol 2025 Jan 01;50(1):86-95. [doi: 10.1093/jpepsy/jsae058] [Medline: 39101560]
42. Cohen KA, Schleider JL. Adolescent dropout from brief digital mental health interventions within and beyond randomized trials. Internet Interv 2022 Mar;27:100496 [FREE Full text] [doi: 10.1016/j.invent.2022.100496] [Medline: 35257001]
43. Mohr DC, Lyon AR, Lattie EG, Reddy M, Schueller SM. Accelerating digital mental health research from early design and creation to successful implementation and sustainment. J Med Internet Res 2017 May 10;19(5):e153 [FREE Full text] [doi: 10.2196/jmir.7725] [Medline: 28490417]

XSL•FO
RenderX

44. Agapie E, Haldar S, Poblete SG. Using HCI in cross-disciplinary teams: a case study of academic collaboration in HCI-health teams in the US using a team science perspective. Proc ACM Hum Comput Interact 2022 Nov 11;6(CSCW2):1-35. [doi: 10.1145/3555610]
45. Lyon AR, Brewer SK, Areán PA. Leveraging human-centered design to implement modern psychological science: return on an early investment. Am Psychol 2020 Dec;75(8):1067-1079 [FREE Full text] [doi: 10.1037/amp0000652] [Medline: 33252945]
46. Kornfield R, Meyerhoff J, Studd H, Bhattacharjee A, Williams JJ, Reddy M, et al. Meeting users where they are: user-centered design of an automated text messaging tool to support the mental health of young adults. Proc SIGCHI Conf Hum Factor Comput Syst 2022 May;2022:329 [FREE Full text] [doi: 10.1145/3491102.3502046] [Medline: 35574512]
47. Kornfield R, Stamatis CA, Bhattacharjee A, Pang B, Nguyen T, Williams JJ, et al. A text messaging intervention to support the mental health of young adults: user engagement and feedback from a field trial of an intervention prototype. Internet Interv 2023 Dec;34:100667 [FREE Full text] [doi: 10.1016/j.invent.2023.100667] [Medline: 37746639]
48. Meyerhoff J, Nguyen T, Karr CJ, Reddy M, Williams JJ, Bhattacharjee A, et al. System design of a text messaging program to support the mental health needs of non-treatment seeking young adults. Procedia Comput Sci 2022;206:68-80 [FREE Full text] [doi: 10.1016/j.procs.2022.09.086] [Medline: 36388769]
49. Hutchinson H, Mackay W, Westerlund B, Bederson B, Druin A, Plaisant C, et al. Technology probes: inspiring design for and with families. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2003 Presented at: CHI '03; April 5-10, 2003; Ft. Lauderdale, FL p. 17-24 URL: https://dl.acm.org/doi/10.1145/642611.642616 [doi: 10.1145/642611.642616]
50. MacLeod H, Jelen B, Prabhakar A, Oehlberg L, Siek K, Connelly K. Asynchronous remote communities (ARC) for researching distributed populations. In: Proceedings of the 10th EAI International Conference on Pervasive Computing Technologies for Healthcare. 2016 Presented at: PervasiveHealth '16; May 16-19, 2016; Cancun, Mexico p. 1-8 URL: https://dl.acm.org/doi/10.5555/3021319.3021320 [doi: 10.4108/eai.16-5-2016.2263322]
51. Zakeresfahani A, Krawczyk M, Ibrahim Z, Modi F, Siek K. Taking stock of studies using asynchronous remote communities (ARC): a scoping review. Interact Comput 2025:022 [FREE Full text] [doi: 10.1093/iwc/iwaf022]
52. Dunbar JC, Connelly CL, Maestre JF, MacLeod H, Siek K, Shih P. Considerations for using the Asynchronous Remote Communities (ARC) method in health informatics research. In: Proceedings of the 2017 International Conference on American Medical Informatics Association Annual Symposium Workshop on Interactive Systems in Health Care. 2017 Presented at: AMIA '17; November 4, 2017; Washington, DC URL: https://www.researchgate.net/publication/320096467_Considerations_for_using_the_Asynchronous_Remote_Communities_ARC_method_in_health_informatics_research
53. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Oct;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
54. Spitzer RL, Kroenke K, Williams JB, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
55. Kruzan KP, Meyerhoff J, Nguyen T, Mohr DC, Reddy M, Kornfield R. "I wanted to see how bad it was": online self-screening as a critical transition point among young adults with common mental health conditions. Proc SIGCHI Conf Hum Factor Comput Syst 2022 May;2022:328 [FREE Full text] [doi: 10.1145/3491102.3501976] [Medline: 35531062]
56. Focus groups made easy. FocusGroupIt. URL: https://focusgroupit.com/ [accessed 2025-05-29]
57. Meyerhoff J, Beltzer M, Popowski S, Karr CJ, Nguyen T, Williams JJ, et al. Small Steps over time: a longitudinal usability test of an automated interactive text messaging intervention to support self-management of depression and anxiety symptoms. J Affect Disord 2024 Jan 15;345:122-130. [doi: 10.1016/j.jad.2023.10.119] [Medline: 37866736]
58. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
59. Edmondson AC, Mcmanus SE. Methodological fit in management field research. Acad Manage Rev 2007 Oct;32(4):1246-1264. [doi: 10.5465/amr.2007.26586086]
60. Spanhel K, Balci S, Feldhahn F, Bengel J, Baumeister H, Sander LB. Cultural adaptation of internet- and mobile-based interventions for mental disorders: a systematic review. NPJ Digit Med 2021 Aug 25;4(1):128 [FREE Full text] [doi: 10.1038/s41746-021-00498-1] [Medline: 34433875]
61. Ellis DM, Draheim AA, Anderson PL. Culturally adapted digital mental health interventions for ethnic/racial minorities: A systematic review and meta-analysis. J Consult Clin Psychol 2022 Oct;90(10):717-733. [doi: 10.1037/ccp0000759] [Medline: 36227330]
62. Nittas V, Daniore P, Chavez SJ, Wray TB. Challenges in implementing cultural adaptations of digital health interventions. Commun Med (Lond) 2024 Jan 05;4(1):7 [FREE Full text] [doi: 10.1038/s43856-023-00426-2] [Medline: 38182750]
63. Khanna MS, Carper M. Digital mental health interventions for child and adolescent anxiety. Cogn Behav Pract 2022 Feb;29(1):60-68. [doi: 10.1016/j.cbpra.2021.05.003]
64. Schueller SM, Glover AC, Rufa AK, Dowdle CL, Gross GD, Karnik NS, et al. A mobile phone-based intervention to improve mental health among homeless young adults: pilot feasibility trial. JMIR Mhealth Uhealth 2019 Jul 02;7(7):e12347 [FREE Full text] [doi: 10.2196/12347] [Medline: 31267980]

XSL•FO
RenderX

65. Knapp AA, Cohen K, Nicholas J, Mohr DC, Carlo AD, Skerl JJ, et al. Integration of digital tools into community mental health care settings that serve young people: focus group study. JMIR Ment Health 2021 Aug 19;8(8):e27379 [FREE Full text] [doi: 10.2196/27379] [Medline: 34420928]
66. Clarke AM, Kuosmanen T, Barry MM. A systematic review of online youth mental health promotion and prevention interventions. J Youth Adolesc 2015 Jan;44(1):90-113. [doi: 10.1007/s10964-014-0165-0] [Medline: 25115460]
67. Arnett JJ. Emerging adulthood: a theory of development from the late teens through the twenties. Am Psychol 2000;55(5):469-480. [doi: 10.1037/0003-066x.55.5.469]
68. Foulkes L, Andrews JL. Are mental health awareness efforts contributing to the rise in reported mental health problems? A call to test the prevalence inflation hypothesis. New Ideas Psychol 2023 Apr;69:101010. [doi: 10.1016/j.newideapsych.2023.101010]
69. Bank SE, Løkkeberg SE, Del Busso LA, Larsson M. A scoping review of digital interventions to improve communication between children and their parents. Research Square Preprint published on February 14, 2024 [FREE Full text] [doi: 10.21203/rs.3.rs-3910659/v1]
70. Theofanopoulou N, Antle AN, Slovak P. "They don't come with a handbook": exploring design opportunities for supporting parent-child interaction around emotions in the family context. Proc ACM Hum Comput Interact 2024 May 26;8:132. [doi: 10.1145/3637409] [Medline: 39015409]
71. Silverman DM, Rosario RJ, Hernandez IA, Destin M. The ongoing development of strength-based approaches to people who hold systemically marginalized identities. Pers Soc Psychol Rev 2023 Aug 12;27(3):255-271. [doi: 10.1177/10888683221145243] [Medline: 36632745]
72. Genova HM, Haas M, Chen YL, Elsayed HE, McGrath RE, Smith MJ. Development and adaptation of a strength-based job interview training tool for transition age youth on the autism spectrum using community engaged methods. Front Psychiatry 2023 Sep 14;14:1098334 [FREE Full text] [doi: 10.3389/fpsyt.2023.1098334] [Medline: 37779620]
73. Hughes G. Finding a voice through 'The Tree of Life': a strength-based approach to mental health for refugee children and families in schools. Clin Child Psychol Psychiatry 2014 Jan;19(1):139-153. [doi: 10.1177/1359104513476719] [Medline: 23444030]
74. Social media and youth mental health. US Department of Health and Human Services. URL: https://www.hhs.gov/surgeongeneral/reports-and-publications/youth-mental-health/social-media/index.html [accessed 2025-05-29]
75. Hilty DM, Stubbe D, McKean AJ, Hoffman PE, Zalpuri I, Myint MT, et al. A scoping review of social media in child, adolescents and young adults: research findings in depression, anxiety and other clinical challenges. BJPsych Open 2023 Aug 11;9(5):e152 [FREE Full text] [doi: 10.1192/bjo.2023.523] [Medline: 37563766]
76. McCrory A, Best P, Maddock A. The relationship between highly visual social media and young people's mental health: a scoping review. Child Youth Serv Rev 2020 Aug;115:105053. [doi: 10.1016/j.childyouth.2020.105053]
77. Odgers CL, Jensen MR. Annual research review: adolescent mental health in the digital age: facts, fears, and future directions. J Child Psychol Psychiatry 2020 Mar 17;61(3):336-348 [FREE Full text] [doi: 10.1111/jcpp.13190] [Medline: 31951670]
78. Webster D, Dunne L, Hunter R. Association between social networks and subjective well-being in adolescents: a systematic review. Youth Soc 2020 May 14;53(2):175-210. [doi: 10.1177/0044118X20919589]
79. Burke M, Kraut RE. The relationship between Facebook use and well-being depends on communication type and tie strength. J Comput Mediat Comm 2016 Jul 26;21(4):265-281. [doi: 10.1111/jcc4.12162]
80. Gkintoni E, Vantaraki F, Skoulidi C, Anastassopoulos P, Vantarakis A. Promoting physical and mental health among children and adolescents via gamification-a conceptual systematic review. Behav Sci (Basel) 2024 Jan 29;14(2):52 [FREE Full text] [doi: 10.3390/bs14020102] [Medline: 38392455]
81. Ybarra ML, Prescott T, Mustanski B, Parsons J, Bull SS. Feasibility, acceptability, and process indicators for Guy2Guy, an mHealth HIV prevention program for sexual minority adolescent boys. J Adolesc Health 2019 Oct;65(3):417-422 [FREE Full text] [doi: 10.1016/j.jadohealth.2019.04.025] [Medline: 31277991]
82. Coleman L, Coleman J. The measurement of puberty: a review. J Adolesc 2002 Oct;25(5):535-550. [doi: 10.1006/jado.2002.0494] [Medline: 12234559]
83. Caine K. Local standards for sample size at CHI. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. 2016 Presented at: CHI '16; May 7-12, 2016; San Jose, CA p. 981-992. [doi: 10.1145/2858036.2858498]
84. Gordon AR, Calzo JP, Eiduson R, Sharp K, Silverstein S, Lopez E, et al. Asynchronous online focus groups for health research: case study and lessons learned. Int J Qual Methods 2021 Feb 09;20:160940692199048 [FREE Full text] [doi: 10.1177/1609406921990489] [Medline: 35185443]
85. Innovation and AI for accessibility. Microsoft. URL: https://www.microsoft.com/en-us/accessibility/innovation [accessed 2025-05-29]

Abbreviations

ARC: asynchronous remote community
DMHI: digital mental health intervention

XSL•FO
RenderX

Edited by A Mavragani; submitted 18.11.24; peer-reviewed by A Van Meter; comments to author 13.04.25; revised version received 02.06.25; accepted 05.07.25; published 29.08.25.

Please cite as:
Kornfield R, Popowski SA, Tack E, Svoboda J, Herrera M, Nguyen T, Knapp AA, Mohr DC, Meyerhoff J
Designing Digital Mental Health Interventions to Meet the Needs of Older Adolescents: Qualitative Interview and Group Discussion Study
JMIR Form Res 2025;9:e68950
URL: https://formative.jmir.org/2025/1/e68950
doi:10.2196/68950
PMID:

©Rachel Kornfield, Sarah Alexandra Popowski, Emily Tack, Jack Svoboda, Miguel Herrera, Theresa Nguyen, Ashley Arehart Knapp, David Curtis Mohr, Jonah Meyerhoff. Originally published in JMIR Formative Research (https://formative.jmir.org), 29.08.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Feasibility and Acceptability of an App-Based Intervention Aimed at Improving Maternal Health Literacy Regarding Infant Play and Development: Mixed Methods Study

Fiona Bennin[1], PhD; Shane A Norris[1,2], PhD; Alessandra Prioreschi[1], PhD

[1]SA MRC/ Wits Developmental Pathways for Health Research Unit, Department of Paediatrics, Faculty of Health Sciences, School of Clinical Medicine, University of the Witwatersrand, Johannesburg, South Africa

[2]School of Human Development and Health, University of Southampton, Southampton, United Kingdom

Corresponding Author:
Fiona Bennin, PhD
SA MRC/ Wits Developmental Pathways for Health Research Unit, Department of Paediatrics
Faculty of Health Sciences, School of Clinical Medicine
University of the Witwatersrand
Faculty of Health Sciences
York Road, Parktown
Johannesburg, 2193
South Africa
Phone: 27 11 717 2038
Email: fionabennin@gmail.com

Abstract

Background: Allowing infants access to unstructured, unrestricted play in their home environment is imperative for increasing healthy movement behaviors and, therefore, developmental outcomes. Interventions should equip mothers to provide opportunities for infant play as early as possible. Evaluating such interventions is necessary to understand the feasibility for scale-up and implementation in specific contexts. Furthermore, the appropriateness and relevance of standardized outcome measures in different ethnic and socioeconomic contexts should be determined to ensure validity.

Objective: This study aimed to (1) test the feasibility and acceptability of an intervention aimed at improving maternal health literacy regarding infant play and development and (2) determine participants' understanding of the study outcome questionnaires.

Methods: This mixed methods study was nested within the Play Love and You (PLAY) study, a randomized controlled trial (PACTR202202747620052) designed to promote infant development. Mothers assigned to the PLAY study intervention arm at 6 months post partum (n=68) received telephone or in-person check-ins and assessments every 2 months and health literacy intervention content and resources (videos and infographics) delivered via a mobile app every week. Feasibility was assessed by monitoring appointment attendance (adherence) and frequency of access to the content via the app. Acceptability was explored using a questionnaire and 2 participant focus group discussions (FGDs) at the end of the study (12 months post partum). The FGDs also included questions exploring the participants' understanding of the 4 study outcome measure questionnaires administered at both 6 and 12 months.

Results: In total, 68 participants were enrolled in the study at 6 months, of whom 17 (25%) attended the FGDs (n=8, 47% in FGD 1 and n=9, 53% in FGD 2). A total of 79% (49/62) of the participants completed the acceptability questionnaire. The health literacy content was found to be highly acceptable based on qualitative and quantitative data. Most acceptability questions had 98% (48/49) positive answers. Participants enjoyed learning about active infant play and developmental milestones and how to make recycled toys. Over 80% of participants (62/68, 91%) attended the 12-month exit appointment. Most of the participants (47/62, 76%) could access the intervention content over the 12 months of the PLAY study, and of those, 60% (28/47) looked at content more than once a week, and 11% (5/47) did so every day. Less than a quarter (10/47, 21%) only looked at the content sporadically. Access was impacted by technical difficulties attributed to using inconsistent external service providers.

Conclusions: This study was found to be acceptable to participants and feasible in this setting. The high acceptability of the intervention content and belief that other mothers would benefit from it suggests potential for effectiveness in similar communities. However, the feasibility of app-based interventions relies on consistent and low-cost management of digital tools in low-resource settings.

XSL•FO
RenderX

(JMIR Form Res 2025;9:e76517) doi:10.2196/76517

KEYWORDS

acceptability; feasibility; mobile health; mHealth; maternal health literacy; mothers; infants; play and development

Introduction

Background

Due to the importance and vulnerability of the first 1000 days of an infant's life, it is necessary to start developing a range of healthy behaviors from birth [1]. The numerous negative side effects of an inactive, sedentary lifestyle can take years to present; however, there is evidence of an association between increased physical activity in infants (aged 1-12 months) and improved motor and cognitive development, as well as adiposity in infancy and later life [2,3]. In the early months of life, allowing infants access to unstructured, unrestricted play in their home environment is imperative for development and for encouraging increased physical activity [4-6]. Therefore, it is imperative to focus interventions on these formative months to encourage healthy movement behaviors as early as possible.

Previous research suggests that improving an individual's health literacy has potential to positively impact their self-efficacy, resulting in a positive effect on their health outcomes [7]. Similarly, when mothers are provided with the skills required to understand and implement parenting-related information, this can increase their motivation levels and ability to cope with their infants' demands, resulting in favorable infant health outcomes [8]. Therefore, it has been hypothesized that increasing caregiver health literacy is key to caregivers developing a deeper understanding of the importance of play and development and increased self-efficacy and motivation, resulting in greater participation in play in the early months and the years to follow.

In real-world settings, determining the acceptability and feasibility of an intervention is important to understand the practicality and suitability of an intervention for scale-up and implementation in specific contexts [9]. Feasibility is usually measured retrospectively to provide an explanation as to why certain aspects of the study were a failure or a success [9]. Specific indicators that can be used to measure feasibility include retention (how many participants were retained in the study), intervention fidelity (the extent to which an intervention was implemented as designed), acceptability (how well the intervention was received by the target group), adherence (the extent to which the individuals followed the intervention as prescribed), recruitment methods, and user engagement [10]. Furthermore, certain outcome measures developed in higher-income countries may not be culturally appropriate to certain ethnic and socioeconomic groups in other settings, which makes it important to determine whether the measures used in a study are appropriate, relevant, and understood by the target group [10].

With the increase in access to mobile phones in low- and middle-income countries, there has been an increased potential to encourage positive health behavior change and outcomes at a community level making use of digital interventions [11]. Conversely, there are significant resource barriers to implementing mobile health (mHealth) interventions in low-resource settings that impact the feasibility of these interventions. Barriers include a lack of electricity to charge devices (due to power outages at health facilities), limitations in hardware and software requirements, and limited access to affordable data or Wi-Fi [12-14]. Furthermore, when end users can access digital information, some individuals do not have the health literacy or language skills needed to understand and use health information appropriately [15,16].

However, there are certain facilitators that aid in intervention feasibility. To overcome challenges with understanding health information, studies conducted in similar settings suggest using simple, lay language together with images, illustrations, and interactive content and including oral explanations for those who cannot read [15,17]. Health content creators should also use local languages to convey health information and have a good understanding of the local context [15]. According to Sekhon et al [18] (and the definition used in this study), for an intervention to be acceptable, individuals need to like the content, and it should fit with their value system and not take up too much time or take the place of other important daily activities. The intervention should also be coherent and perceived to be effective while improving self-efficacy [18].

Objectives

The Play Love and You (PLAY) study is a randomized controlled trial (RCT) of an intervention designed to promote infant development through encouraging maternal self-efficacy using behavioral feedback and supportive microinterventions [19]. This study was nested within the PLAY study and fell under one of its secondary objectives, which was to test the efficacy, feasibility, and acceptability of providing intervention content and increasing opportunities for early learning through play to promote health literacy in mothers (in a low-resource setting) [19]. Findings from this study were to be used to guide the development of a high-fidelity intervention implemented at scale.

The specific objectives of this study were as follows:

1. To test the feasibility and acceptability of an intervention aimed at improving maternal health literacy regarding infant play and development
2. To determine participants' understanding of the study outcome questionnaires' content considering the context in which they live

Methods

Study Setting and Participants

The study described in this paper is a substudy of a larger RCT called the PLAY study. As per the PLAY study protocol, mother-infant pairs were recruited to the study within a month following delivery from 2 community clinics, Lilian Ngoyi and Chiawelo. Both clinics are primary clinics linked to Chris Hani

XSL•FO
RenderX

Baragwanath Hospital in Soweto, South Africa. They both have midwife obstetrics units as well as postnatal clinics, which provide antenatal and postnatal care, respectively [19]. Soweto is an urban *township* located in the southwest of Johannesburg, with a population of approximately 1.9 million people [20]. As of 2023, approximately 9.9% of the Soweto population comprised children aged <6 years [20]. Participants were screened and enrolled in the study at their postnatal checkup 2 days after birth. Mothers were eligible for recruitment if (1) they were aged ≥18 years with an infant aged <1 month (2) of whom they were the primary caregiver and (3) were living in Soweto for the duration of the study. Mothers with both primigravida or multigravida birth experiences were included in the study. Participants were ineligible if infants were born before term. Each pair was individually sequentially randomized into either a control or intervention group using simple 1:1 randomization generated by Stata (version 17; StataCorp) at enrollment. The intervention lasted for 12 months, and assessments were made at enrollment and when the infants were aged 2, 4, 6, 8, 10, and 12 months. Recruitment ran from August 2022 to June 2024. The full details of the PLAY study are described elsewhere [19]. Regarding this substudy, a health literacy intervention was then implemented with the PLAY study intervention group from when the infants were aged 6 months, which constituted enrollment for this substudy. Exit assessments were done at 12 months. Follow-up took place at 8 and 10 months. All health literacy intervention content and delivery methods were initially developed and piloted with a community advisory group (CAG), and examples of the developed content were found to be acceptable [21].

Ethical Considerations

Ethics approval was obtained from the University of the Witwatersrand Human Research Ethics Committee (substudy M220886 under main study M210846 and M220217). Participant information forms were provided to each participant to read, discuss with, and have any questions answered verbally by a research assistant (RA). Written informed consent was then obtained from all participants for taking part in the focus group discussions (FGDs) and questionnaires and for each FGD to be recorded. Once enrolled in the study, participants were given a unique trial identification number, which was used throughout the study to protect participant privacy and confidentiality. Participant identifier information was kept in a separate project database to maintain anonymity. All participants were reimbursed for travel costs and provided with refreshments at FGDs and hospital-based follow-ups.

Study Design

This substudy used mixed methods to describe and explore the acceptability and feasibility of the health literacy intervention among the PLAY study intervention group between 6 and 12 months. Mothers randomly assigned to the intervention group (n=68) received telephone or in-person check-ins and assessments conducted at 6, 8, and 10 months and referrals to health care facilities and community services if necessary (standard of care) [19], as well as the intervention content and resources delivered via an HTML5 mobile app deployed to run on the participants' standard-size smartphones. The app was downloaded and set up on the participants' smartphones at enrollment. Content was updated weekly so that each participant received content specific to their children' ages (based on age in months). The health literacy content included infographics and videos showing examples of toys that could be made from recycled material at home, as well as age-appropriate activities and games for infants to encourage active play and development. Informative content such as monthly developmental milestones and information and suggestions regarding safe active play in the home were included. Examples of the video and infographic content are shown in Figures 1 and 2, respectively. This content was codeveloped with the CAG and designed to be specifically contextually relevant for this community [21]. All infographics and video content were provided in English based on CAG recommendations [21]. The initial in-person appointment and enrollment provided a general check-in with the participants to discuss the content and assist with any technical difficulties they may have had when accessing the content through the app. At the 8- and 10-month appointments, specific questions were asked about the infant's play and development, and the content was discussed again.

XSL•FO
RenderX

Figure 1. Screenshot of an example of app health literacy video content.

Figure 2. Screenshot of an example of app health literacy infographic content.

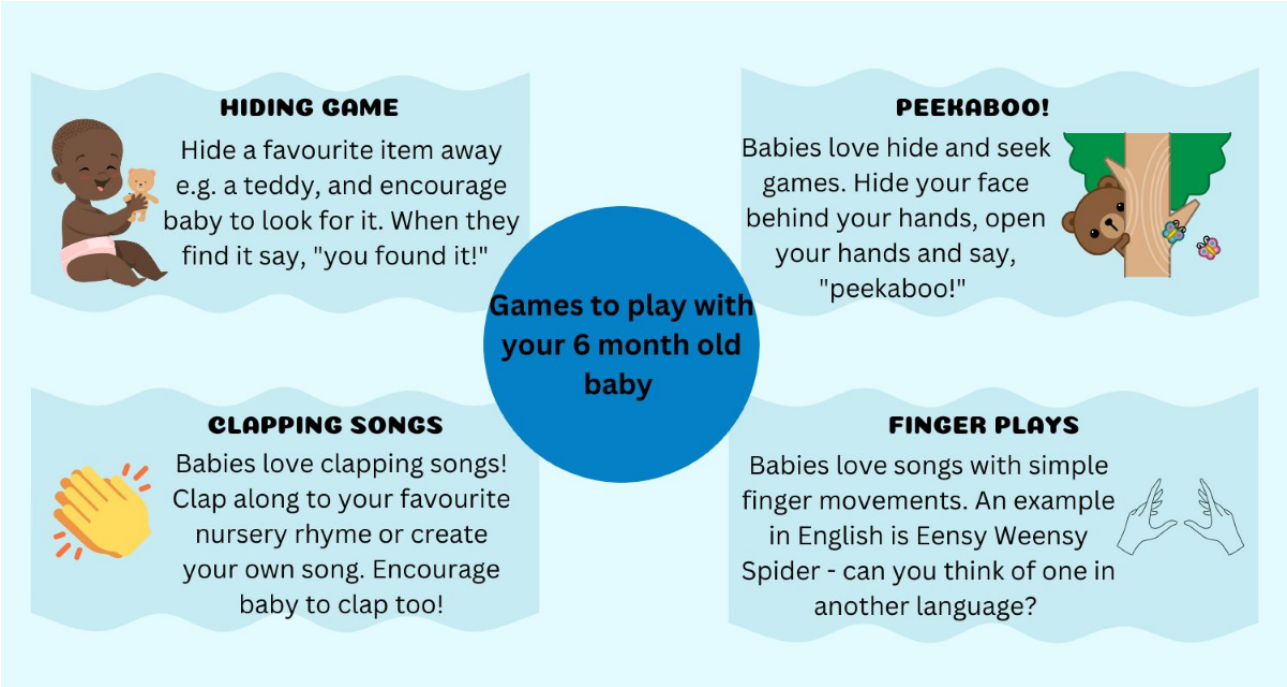

Data Collection

Quantitative Data

Overview

All quantitative data were collected by the PLAY study team's RAs using the REDCap (Research Electronic Data Capture; Vanderbilt University) software. The participants' age, educational level, and income status were collected via a questionnaire at initial PLAY study enrollment (within 1 month of delivery). The acceptability questionnaire was administered to all intervention group participants as part of the 12-month exit assessment from October 2023 to June 2024 by an independent RA to reduce bias. All participants who could not attend in-person 12-month exit assessments were traced and followed up on in their homes.

Feasibility

Participant contact (adherence) data were extracted from REDCap to determine the number of participants who attended the 6-, 8-, 10-, and 12-month appointments. As attrition of >20% has the potential to introduce significant bias to the study findings, the intervention was considered feasible if 80% of those participants enrolled in the substudy at 6 months attended

XSL•FO
RenderX

the 12-month exit appointment [22]. Finally, questionnaire data on compliance were used to determine how many participants could access content via the app and how frequently they accessed content over the full 12 months of the PLAY study. As this could not be disaggregated to include only access to content from 6 to 12 months, the feasibility of this substudy alone could not be accurately measured using these data, which were included for interest and discussion.

Acceptability

The acceptability questionnaire, consisting of 22 questions, was compiled by the study team and included questions that captured the acceptability of the health literacy intervention using an ordinal scale. This questionnaire focused specifically on health literacy acceptability to distinguish it from the overall acceptability of the larger study. The acceptability questions were based on the constructs developed by Sekhon et al [18], namely, affective attitude, burden, ethicality, intervention coherence, opportunity costs, perceived effectiveness, and self-efficacy. These constructs were also used to inform the development of the intervention content and, therefore, allowed for continuity of measuring acceptability before and after the study [18,21]. Figure 3 [18] provides a more detailed description of each construct.

Figure 3. The theoretical framework of acceptability comprising 7 component constructs by Sekhon et al [18].

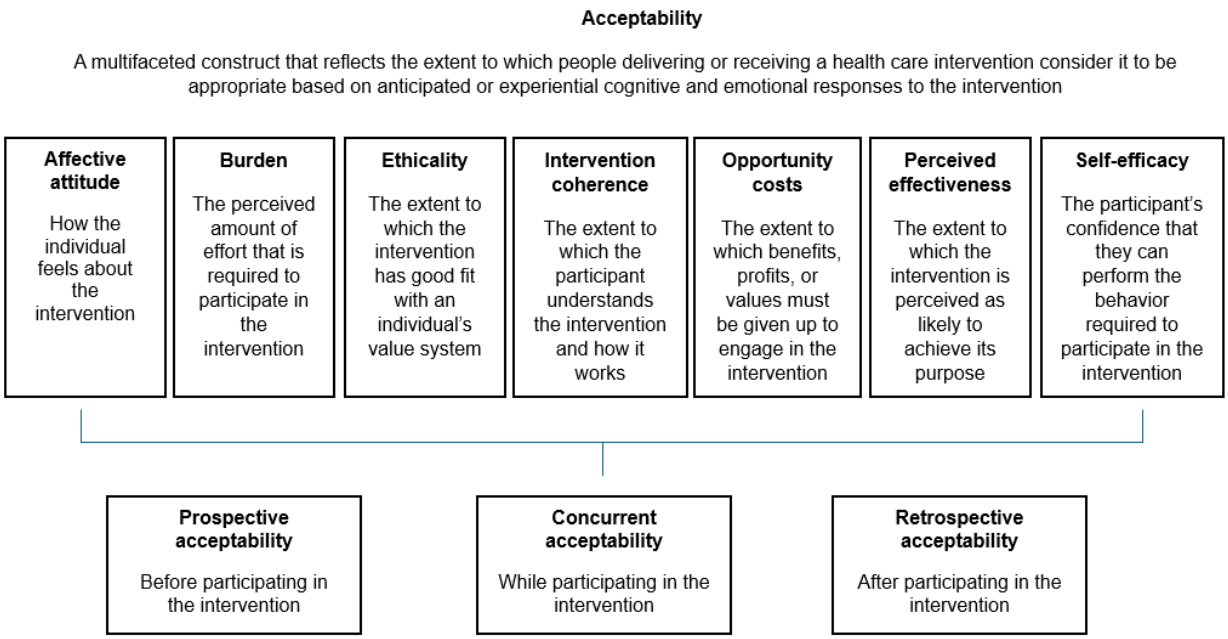

Qualitative Data

Two qualitative FGDs were conducted by experienced RAs in December 2023 (FGD 1; n=8) and February 2024 (FGD 2; n=9) with intervention group participants who had completed the intervention (n=17). Participants who had completed the acceptability questionnaire by December 2023 were invited to attend FGD 1. This was repeated again for FGD 2; however, those attending the second FGD were required to have also attended the 10-month visit by February 2024 to ensure that we could assess the acceptability of that visit properly.

The FGD guide expanded on the questions asked in the quantitative questionnaire to allow for a better understanding of acceptability of the intervention content. The FGD guide also included questions exploring the participants' understanding of the 4 study outcome measure questionnaires administered at both 6 and 12 months, namely, the Health Literacy Questionnaire [23], Affordances in the Home Environment for Motor Development–Infant Scale [24], and Assessing Children's Sedentary and Active Pursuits Questionnaire and Maternal Beliefs Questionnaire [6,25]. The Health Literacy Questionnaire is a standardized questionnaire measuring health literacy according to the following categories: (1) feeling understood and supported by health care providers, (2) having sufficient information to manage the child's health, (3) actively managing the child's health, (4) social support for health, (5) appraisal of health information, (6) ability to actively engage with health care providers, (7) navigating the health care system, (8) ability to find good health information, and (9) understanding health information well enough to know what to do [23]. The Affordances in the Home Environment for Motor Development–Infant Scale is an instrument measuring the affordances available for motor development in the child's environment, asking questions on availability of physical space and variety of stimulation and play materials in the home [24]. The Maternal Beliefs Questionnaire explores the mother's beliefs, attitudes, and intentions regarding their infant's physical activity, specifically regarding physical activity knowledge, views on children's physical activity, physical activity optimism, self-efficacy for promoting physical activity, future expectations regarding the infant's physical activity and television viewing, and floor play concerns [6,25]. The Assessing Children's Active and Sedentary Pursuits Questionnaire has two parts: mothers are asked to report (1) the amount of time (minutes) their infant spends in various activities on an average day and (2) whether their child has access to or is likely to have access to age-specific

XSL•FO
RenderX

toys and equipment (such as balls, push toys, and bicycles) within the following year [6,25]. As most of the participants were not first-language English speakers and the questions were primarily asked in English (and translated for further clarity if needed), they were asked to discuss any barriers to understanding the questions in each questionnaire. Furthermore, participants were asked to discuss the relevancy of each questionnaire in relation to their context.

Both FGDs were conducted at Chris Hani Baragwanath Hospital in Soweto, Johannesburg, by staff trained in qualitative research data collection. Although questions were asked in English, participants could speak in their home language, and conversations progressed in the vernacular. This was later translated into English (if the response was in a different language) and transcribed for analysis.

Statistical Analysis

Descriptive statistics (frequencies and percentages for categorical data and medians and SDs for continuous data) were used to describe the sociodemographic data of the participants, as well as the quantitative acceptability and feasibility data. Similarly, descriptive statistics summarized how many participants attended each appointment and how many participants could access the app. The number of participants who attended the 6-, 8-, 10-, and 12-month follow-up visits was calculated (Figure 4). For each acceptability question, the 2 lowest (1 and 2) and 2 highest (4 and 5) Likert scale scores were combined, resulting in a *low*, *neutral*, or *high* result.

Figure 4. Intervention group attendance (and reasons for withdrawal) at 6-, 8-, 10- and 12- month follow-up appointments.

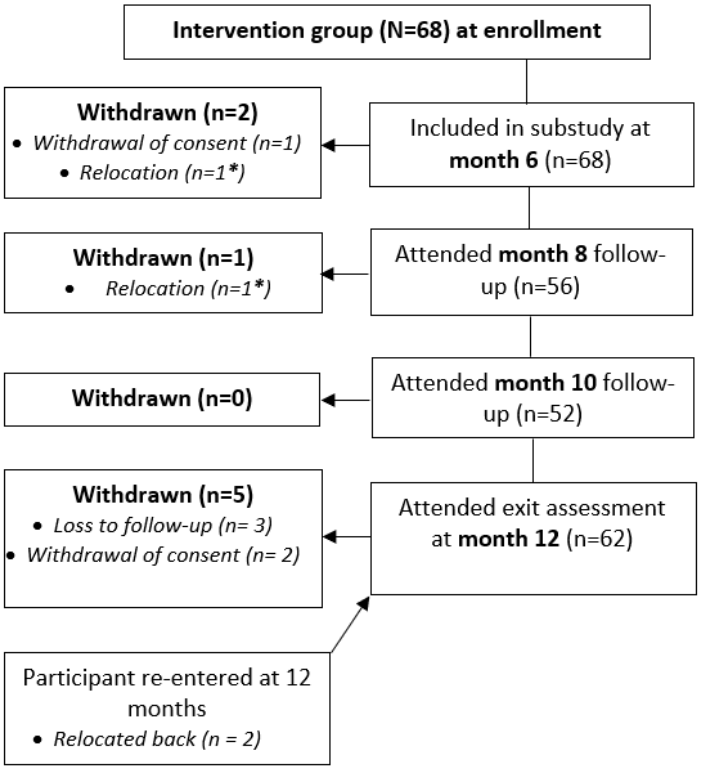

The qualitative data were analyzed by 2 researchers (FB and AP) using thematic network analysis [26]. Meaningful statements and ideas were coded and then grouped into themes, which were summarized and discussed in relation to the original research question. FGD 1 data were coded and classified into loose themes linked to the acceptability constructs. These findings led to the development of the guide for FGD 2, focusing mostly on providing further insights into the themes already established, as well as exploring other themes that were not discussed in depth in FGD 1, such as the validity of the outcome measure questionnaires. The data from FGD 2 were then analyzed using the themes from FGD 1, allowing for new themes to emerge. The codes for the themes from the 2 FGDs were then combined or adjusted accordingly to create new themes and subthemes. At each stage, the analysis was iterative, with researchers double coding the data as well as cross-coding the analyses. When disparities existed, discussions were held, and themes were collapsed or adjusted until agreement was reached.

Results

Feasibility

Of the 207 participants screened and invited to take part in the PLAY study, 36 (17.4%) did not complete baseline and randomization, resulting in 171 (82.6%) participants being enrolled in the PLAY study. Of these 171 participants, 87 (50.9%) and 84 (49.1%) were randomized into the intervention group and control group, respectively, after completing the PLAY study enrollment assessment. Of the 87 participants in the PLAY study intervention group, 68 (78%) attended the 6-month assessment, which was the enrollment visit for this substudy. The number of participants who attended the 6-, 8-,

XSL•FO
RenderX

10-, and 12-month follow-up visits for this substudy is summarized in Figure 4.

Participants' access to the intervention content is shown in Table 1. Most of the participants (47/62, 76%) could access the intervention content over the 12 months of the PLAY study. Of these, 60% (28/47) looked at content more than once a week, and 11% (5/47) looked at it every day, yet just under a quarter (10/47, 21%) only looked at the content a few times.

Table 1. Intervention group access to intervention content over 12 months (N=47).

Characteristic	Participants, n (%)
Looked at content every day	5 (11)
Looked at content more than once a week	28 (60)
Looked at content more than once a month	4 (9)
Only looked at content a few times or never	10 (21)

However, despite most participants having access to the content, the access issues experienced by some participants were largely linked to issues beyond the researchers' control, such as relying on external stakeholders to upload content and update participant profiles in a timely manner, as well as issues with the software, which prevented certain videos from being downloaded despite the content being mobile data free. The study RAs tried to overcome content access challenges by (1) attempting to troubleshoot any software issues when physically with the participants at follow-ups and, if not successful, (2) going through any missed content with the participants at follow-up visits.

Acceptability

Of the 62 participants who attended the 12-month exit assessment, 49 (79%) completed acceptability questionnaires. The average age of those who completed the acceptability questionnaire was 29 (SD 5.8) years. Multimedia Appendix 1 shows the results of the acceptability questionnaire. Overall, the questions (such as content ethicality), received very positive responses, with most having a 98% positive result. However, the responses regarding the effort it took to engage with the play and development information on the app were not as positive (37/49, 76%; Figure 5). Furthermore, only 51% (25/49) of the participants always used the information presented to them when deciding which activities to do with their infants (Table 2). Reasons for not engaging with the play and development intervention (opportunity cost) were a participant not having enough time (1/49, 2%) and another participant losing their phone (1/49, 2%; Multimedia Appendix 2). Further details regarding affective attitude, intervention coherence, perceived effectiveness and opportunity cost are shown in Multimedia Appendices 1-4.

XSL•FO
RenderX

Figure 5. Acceptability questionnaire findings for Burden category in intervention group (N=49).

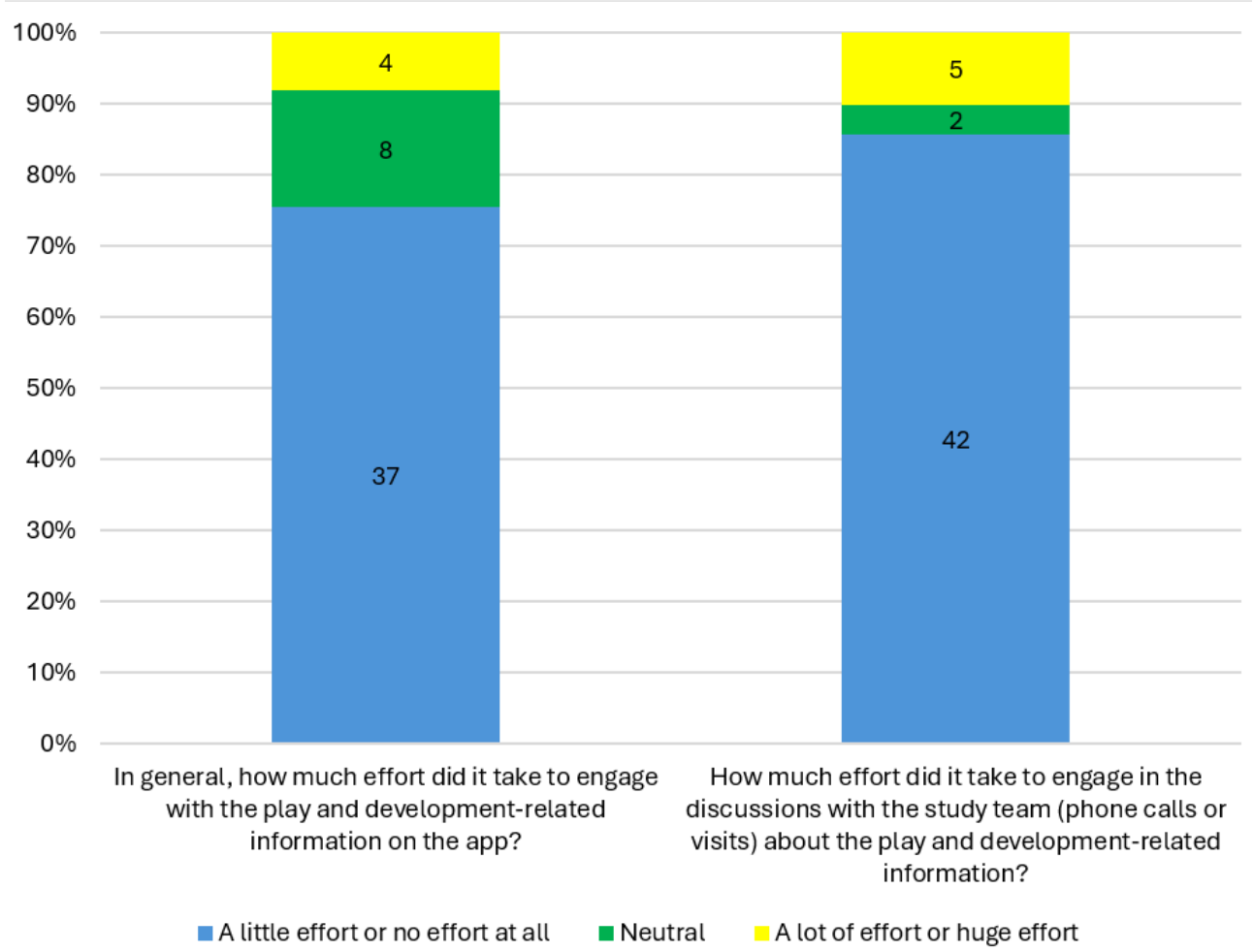

Table 2. Acceptability questionnaire findings for self-efficacy category in intervention group (N=49).

How often did you use the play and development information presented to you when deciding what activities to do with your baby?	Participants, n (%)
Usually or always	25
Sometimes	23
Hardly ever or never	1

FGD Results

Overview

The combined average age of the participants who attended the focus groups was 30 (SD 5.1) years. In total, 13% (8/62) of the participants attended FGD 1, and 15% (9/62) of the participants attended FGD 2. The themes and subthemes that emerged from the FGDs are discussed in the following sections.

Affective Attitude

Participants spoke about what they liked or did not like regarding the intervention content, as well as which aspects they enjoyed the most. Overall, the participants particularly liked the content on making toys and promoting play and development, which was presented in both video and infographic format.

Making Toys

The participants enjoyed how the content gave them the possibility of their infants having toys even when participants could not afford to buy them. As the content showed the participants how to make toys out of recycled material (eg, a shaker out of a toilet roll holder and rice), they appreciated that the toys were low cost and easy to make:

> *It was exciting because we got to learn how to create toys out of nothing. I learned how to make a shaker for a child using a bottle with beads or sand or maybe using eggshells. That turned out to be something that the child can use to shake. We got to know that if we are not able to afford buying toys, we can make them ourselves.* [FGD 2]

XSL•FO
RenderX

> *It could be a box and maybe coke plastic bottle and then they would help demonstrate how to create your own toy instead of going to buy one. They were accommodating those who do not have money so that they can also be able to make the choice themselves.* [FGD 1]

> *We are all going to say the same things. What I like the most about the app is about making the toys.* [FGD 2]

Participants expressed enjoying the video content, which showed them step-by-step instructions for making certain toys:

> *I liked the videos because they would teach us on how to make a toy. They would show you what you needed to make that particular toy.* [FGD 1]

The participants demonstrated a sense of pride in knowing that they received information and ideas from the study, despite one participant talking about being ridiculed by others for using recycled material to make the toys:

> *...There is another child where I am staying [who] likes buying juices packaged in a cell phone like bottle and then when they are done drinking, they would give me those bottles so I put beads inside and then I burned the opening to close it up and then when the child shakes it would make sounds. The child enjoyed it a lot and people would laugh at me and then I would tell them to not laugh at me because they had no idea where I got the information about making this toy.* [FGD 1]

Promoting Play and Development

Participants enjoyed the content on promoting play and development, which encouraged them to play with their children and engage in activities that promoted development:

> *I made different colourful shapes and put them for them [the infant] to choose. Whichever shape they would pick I will tell them if maybe it is a circle. I would encourage them to pick whichever shape I would tell them to.* [FGD 2]

> *I liked the information especially about how to encourage the child to sit or crawl.* [FGD 1]

Some participants reported that the content helped them learn about the importance of play and how to spend time actively playing with their children. Singing and dancing activities were particularly popular:

> *This Play Study has helped me to know that I should have time for my child and play with him. Not just playing to pass time but I can spend maybe one hour and it must be a routine that we do that every day.* [FGD 1]

> *I also liked songs such as "head, shoulders, knees and toes." We would sing while showing them by pointing their hands, shoulder, knees and toes. I would also sing the song in IsiZulu. We usually sing in front of a mirror and it would excite my child even though they were not understanding what it means. Sometimes I would also point at other body parts while telling them what they are.* [FGD 2]

> *What I liked the most was when I got to sing for my child and I could see that they would start dancing because they were enjoying it.* [FGD 2]

Burden, Coherence, and Opportunity Cost

For those participants who could access the intervention content, all of them found it easy to access and use, with no one reporting any problems with access and usability:

> *I think it was user friendly and simple to understand, depending on the information that I was looking for and I was able to access the information. Sometimes I would refer to the YouTube links for the videos.* [FGD 1]

When discussing how often the participants engaged with the intervention content, many participants reported that there were other competing priorities, such as work and attending to older children, that would limit their time on the app. This resulted in most participants engaging with the content ad hoc, when they had time:

> *For me it was finding time to go through the app because I go to work and when I get back, I still need to spend time with the children. On weekends I would be busy with house chores. I hardly had time to go into the app and that was my challenge.* [FGD 2]

> *It will be sometimes (going through the app) when I am just sitting alone, I would say let me take the phone and go through the app and read the contents.* [FGD 2]

Furthermore, most participants found the content easy to understand; however, those who struggled were those who had other tasks or children vying for their attention, therefore limiting their ability to concentrate and comprehend the information:

> *I do not think there is anything I came across that was difficult. Every information that was there was simple for me to understand. If there was information that I did not understand at the end there will be a video or some pictures which would help me to understand better what they were talking about. For each and every information you would read about there would be a picture attached to it explaining to you that this is what we are talking about.* [FGD 2]

> *...Most of the time I am left alone at home because children go to school. I would be on the app but my mind would be wandering somewhere. I could be reading and understanding but my mind would not be in the same space...I know that there is a child in the house and have to keep an eye on them but because there are a lot of things going on and there is no one to talk to. I would tell myself that I need to be focused on the child but my mind would not be there.* [FGD 2]

XSL•FO
RenderX

Perceived Effectiveness

Overall, the participants wanted the content to be accessible to all mothers in their community and were positive about the effectiveness of the content, particularly with regard to its impact on their health literacy and their infants' development.

Infant Development

While many of the participants were aware of the major developmental milestones that children should be reaching at certain ages, they found that the intervention content provided examples of activities to assist their infants in reaching these milestones. The intervention content also assisted in knowing what was considered *normal* with regard to their infants' development. Furthermore, the participants gained understanding of the importance of play for promoting development:

> *We knew that he can hold your child to help them to learn to stand or walk on their own for a little while but with the app, you become aware of the milestones and how to assist your child reach them.* [FGD 1]

> *I think playing with your child helps with their brain development and stimulating them. At first, I would just think that my son is just playing and I did not know that while he is playing, the development process is also taking place.* [FGD 1]

Health Literacy

Many of the participants reported an increase in their own knowledge of play and development and that their immediate and extended families gained more knowledge as a result of engaging with and sharing the intervention content. The participants felt that the alignment with the information provided by the clinics gave validity to the intervention content and made it seem trustworthy:

> *Another thing that I have found very important is the fact that I can always reference between the study and the road to health card, that I know what I am doing is right, it is not something that was created from nowhere. So, the fact that there is a backup of the information that I have received.* [FGD 2]

> *My eldest daughter is 21 years old and she is the one who made those toys because she would also read through the app. She is studying towards teaching qualification, so this was helpful for her as well. She would play most of the games with the child and with the other sibling. The app was not just helping me but all of us at home.* [FGD 1]

Some of the participants discussed how the intervention content helped develop their self-efficacy and gave them a sense of independence, which was particularly important for those who did not have support from partners or immediate family:

> *Another thing that I have just picked up from our sister who was speaking now, when you are a first-time mom and you do not have a mother, there are fathers who are hands on and are able to offer you support. But if you do not have that support structure and you know nothing and you have to teach yourself everything, even with me they just showed me once and then I was able to do everything for myself afterwards.* [FGD 1]

The participants reported that the intervention content was particularly helpful for those who could not read in teaching them activities to encourage their infants to reach their developmental milestones and the average timelines for each milestone:

> *I would like to add on that part of people who cannot read...If I am a new mother and do not know anything about the baby, the app shows me step by step. They mentioned the developments expected from the child month to month and if they are not doing what they were supposed to be doing they will tell you to do this and this to encourage the baby. It also shows that your baby will do this and that and this is how you should respond as the mother, you do not have to go to the clinic every time.* [FGD 2]

Effect of Assessments on Health Literacy

Change in Health Literacy and Behavior Following Questionnaires About Play

While the FGDs sought to determine whether the participants understood the questions asked during the study assessments, it was found that just being asked questions about play and development as part of the assessment questionnaires had an impact on the participants' behavior, thoughts, and decision-making regarding play environments and resources or toys.

The participants found the assessments relatable and often wanted to find out more about the assessment topic after the session was completed. Many learned more about the stages of development and what behavior was appropriate at the different stages from questionnaires about milestone attainment:

> *I would say the importance of the milestones. That is the one thing that I can take away for 10 months because immediately after you guys called you asking about if the child can see small objects, can you tell to see a pin and so forth. Some of the questions were eye opening because I never used to take notice of that but after the phone call, I realised that there are few milestones that I have not looked into but now the phone call came, and I wanted to see if my child could do those things you were asking us about.* [FGD 2]

Play Environments

Many of the participants reported that the questions encouraged them to reflect on their home and outdoor environments, particularly related to where their infant could play safely. The participants started to think about how their children could move around actively while being supervised and safe depending on their own home circumstances:

> *You remember those questions where they asked that, when the child is playing outside, is there anything that they can hold onto for support. I did not realise that when the child is playing outside, I need to guard*

XSL•FO
RenderX

them. They also asked if there are steps and how big they were. Those questions were very eye opening. I was now able to know what things to look out for when the child is playing to know whether they are in a friendly or danger zone. [FGD 2]

For me those questions made me feel somehow because she would be asking if I prefer my child playing or carrying them on my back. At that time, I would then think to myself that if I carry my child on my back, he will not be free to play, but if I put him on the floor, he will be free to play whatever he wants to. For me those questions were relevant because I got to understand that I do not need to confine my child in a small space because of being scared that they were get hurt while playing. Eventually the child will grow and go outside to play. [FGD 2]

Resources or Toys

Many of the participants discussed how the questions encouraged them to think about what types of toys their infants were playing with and when they should play with different types of toys according to their stage of development. Some did not know that they could introduce toys at an early age, whereas others learned more about the quality of toys and how they could impact their infants' development in different ways:

For me I can say I was not aware at that time because when the questions were asked, I thought my child was still very young to know about toys. I had not even bought him one toy because I thought he was still young; he cannot play with toys but only with us. [FGD 2]

It was eye opening in a sense that toys are not the same and they do not offer the same qualities in terms of the child's development. Like when you are asked about the book, whether the child has a book. My child has a book, and it is called a Peekaboo Book. The book itself ask about animals, who is hiding behind? It flips open, peekaboo and then it will show you that animal. That was eye opening because firstly the child gets to hear the sounds of the peekaboo that you are saying, and the child can see the colours and animals. For me those kinds of toys felt like they were more educational. As much as they were playful, they were more educational. It was better than having a toy that does not do anything, and it is dark, you do not know. Maybe if it was brighter the child would want it... [FGD 2]

Discussion

Principal Findings

Overview

This study aimed to determine the feasibility and acceptability of an intervention intending to improve maternal health literacy regarding play and development and to determine the participants' understanding and perceived relevance of the study outcome questionnaires. Recent studies have found the provision of content through mobile apps to be successful in supporting and improving maternal health during the peri- and postnatal periods [27-29]. However, there is an increasing awareness that, for mHealth interventions to improve maternal and child health, they need to be accessible, culturally appropriate, and relevant to the social and physical setting [27,30]. This study adds to the emerging body of knowledge exploring the feasibility and acceptability of this type of intervention in a low- to middle-income country [31,32]. While the intervention was found to be acceptable and feasible, some factors such as access to information on the mobile app were less so due to software-specific issues and preference for information in video format.

Feasibility

As >80% of the participants (62/68, 91%) attended the 12-month exit assessment, according to our definition, this study was found to be feasible. When exploring the feasibility of the intervention content and delivery, there were several factors that impacted the participants' ability to access the play and development content on the app linked to issues with the app software and access to mobile data–free content. These barriers are not limited to this study, and similar problems have been reported in existing maternal or child mHealth technologies (using SMS text messages) available in South Africa such as MomConnect [33], which have been reported as helpful but inconsistent in delivery [16]. Other external factors such as phone sharing, losing phones or having phones stolen, and changing of phone numbers were apparent but expected. These are common problems impacting adherence and delivery of mHealth interventions in similar low-resource settings [16,34,35]. Despite these difficulties, more than half (28/47, 60%) of the participants engaged with the content more than once a week. Other mHealth (SMS text message) studies in low-resource settings have found this frequency of content delivery to be acceptable among users [36]. For those who engaged less frequently, other priorities such as work and looking after other children often took preference. This was similarly noted in the prototyping work with the CAG that informed this study [21].

Study follow-up attendance was found to be relatively high despite the decrease observed in the 8- and 10-month follow-ups. Reasons for the low attrition could be attributed to the RAs developing an effective track-and-trace system, where all uncontactable participants were visited at home. From here, the RAs would make contact again with the participants or be provided with the participants' location or contact details by a family member or neighbor. Using a telephone to conduct the 10-month follow-up could have had an impact on adherence due to the factors regarding phone use discussed previously. Finally, as there was a focus on collecting data for the outcome analysis time points, this led to higher attendance rates at both the 6- and 12-month follow-up time points. It is also important to note that, even though participants may not have attended each follow-up visit, they were still engaging with the weekly content via the app.

Acceptability

Overall, there was a high level of acceptability of the intervention among the participants. The least acceptable

XSL•FO
RenderX

components of the intervention (that were still considered very highly acceptable at 37/49, 76%) regarded how often the participants used the information when deciding on what activities to do and the effort it took to engage with the information. In the questionnaire, the participants reported that they preferred meeting in person, and no participants reported that they preferred engaging with the infographics. This correlates with findings from the preliminary prototyping work with the CAG, which found that 90% of the group preferred team discussions and videos compared to the infographics when asked in a questionnaire but were more favorable toward the pictures during the discussion [21]. Unfortunately, due to limited resources for producing high-quality videos in this pilot study, the researchers had to provide a combination of videos and infographics to ensure the validity of the information shared. Similarly, while some participants liked reading the content information, some reported that they particularly liked the videos as they gave them step-by-step instructions that were easy to follow. This has also been reported in other low-resource settings where low (health) literacy levels can result in misinformation and limit understanding of written materials such that oral instructions or other traditional communication (eg, radio and television) are deemed more acceptable and feasible [16]. While other exploratory work in a similar setting found that women were skeptical about receiving information that would not elicit any change in behavior due to constraints in their environment [16], participants in this study were very positive about the intervention and the change it could achieve, particularly when used alongside the Road to Health card and information provided by health facilities. This can largely be attributed to the content being so contextually relevant as a result of involving a small group of mothers from this setting (the CAG) in testing the initial prototype, which was then used to create our final intervention content [21]. Furthermore, aligning the content with government-provided information played an important role in increasing the intervention's acceptability, the participants' ability to trust the information, and their ability to use it alongside the resources they already had. This has been observed in text-based mobile interventions in a similar context, where the provision of trusted health information resulted in positive health outcomes [37].

The participants found the questionnaires easy to understand, but this was largely impacted by having an RA who was able to translate any difficult words into the participants' home language if required. While participants found the questionnaires relatable to their context, an interesting finding was how the questionnaires did not just serve as outcome measures but acted as a means of improving knowledge regarding play, development, and the environment. This was largely due to participants being required to consider whether their environments were conducive to their children playing safely and discussing the importance of toy use at different ages to encourage development. This phenomenon is an extension of the Hawthorne effect (which occurs when individuals modify their behavior in response to being observed or knowing that they are part of an experiment), also known as measurement bias or "measurement reactivity" [38]. Measurement bias is commonly observed when repeating the same questionnaires (response shift) and when the act of completing a questionnaire results in new beliefs and behaviors regarding the research topic [39]. However, the Hawthorne effect usually disappears in well-designed RCTs, as would be expected in this study [38].

Potential for Effectiveness

The high acceptability of the intervention content and belief that other mothers would benefit from it provides great potential for observing effectiveness in similar communities. As there was such a positive response from the participants regarding being able to make their own toys and learn new skills, in practice, this ability would allow other mothers in similar settings to have access to these resources, therefore encouraging and equipping them to foster more play and, subsequently, further development in their infants. The participants reported an increase in knowledge of infant play and development and self-efficacy after engaging with the content, which supports other studies highlighting a positive association between health literacy and self-efficacy [40,41]. As self-efficacy has been found to be a mediator between maternal health literacy and positive parenting behaviors (ie, encouraging participation in infant play) [8], this links positively to the overall trial study objective of encouraging maternal self-efficacy using different microinterventions. Testing feasibility and acceptability is an important first step as it provides important guidelines when scaling up studies or implementing these interventions in practice.

Study Strengths and Weaknesses

Using mixed methodology helped improve the validity of the study findings, where insight into the quantitative findings could be challenged or consolidated through the qualitative FGDs. While there were technical issues related to delivering the app in this study (which can be easily rectified), the content of the app and overall intervention were found to be highly acceptable to the participants and feasible in this setting. This study's main limitations were linked to the feasibility of using inconsistent external service providers to deliver the content via a prototype mobile app, which resulted in technical issues (despite the app being mobile data free). To resolve this, the researchers have created an independent app containing the resource content, which counteracts the issues encountered in this feasibility study and can be implemented at scale. With regard to measuring feasibility, the health literacy intervention was provided from infant age of 6 to 12 months, but the larger study provided other content from 0 to 12 months. Therefore, we could not determine what content was received by each individual and when. However, purposively recruiting FGD participants who had received the entire 6- to 12-month intervention, including app content and 10-month follow-up attendance, allowed us to have a better, more in-depth understanding of the acceptability of the intervention provided specifically during those months. It is important to note that only FGD 2 included participants who had attended the 10-month follow-up visit. An additional FGD with participants who attended the 10-month follow-up may have provided further insight on acceptability from those who received the entire intervention. While it is assumed that most participants used Android phones (due to affordability), this was not documented, which would have been helpful to

XSL•FO
RenderX

determine whether there were different user experiences between Apple iOS and Android devices.

Implications for Future Research and Implementation

These findings suggest that interventions that provide meaningful, clear, and easily accessible information on play and development in infants would be useful and well received by mothers and caregivers in similar settings and show potential for effectiveness in improving health literacy regarding play and development. Given the digital limitations in this population, researchers and policy makers need to develop interventions that provide information through a variety of platforms to enable equitable access for all.

Conclusions

This study provides insight into the acceptability and feasibility of an intervention aimed at improving maternal health literacy regarding play and development in a low-resource setting. While the intervention was considered feasible within this setting and highly acceptable to participants, barriers related to external management of the app resulted in some issues with accessibility. Importantly, developing the intervention content with the guidance of community members resulted in very high acceptability and potential for effectiveness.

Acknowledgments

The authors would like to acknowledge Mr Ngomeni Mchunu (research assistant) and Ms Ngoyi Monyepote (research assistant) for conducting the focus group discussions and to Ms Ngoyi Monyepote for administering the 12-month acceptability questionnaire. This work was supported by the National Institute for Health and Care Research (NIHR; using the United Kingdom's Official Development Assistance funding) and Wellcome (222007/Z/20/Z) under the NIHR-Wellcome Global Health Research Partnership. The views expressed in this paper are those of the authors and not necessarily those of Wellcome, the NIHR, or the Department of Health and Social Care. The authors declare that no generative artificial intelligence was used in any portion of manuscript writing.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The 12-month acceptability questionnaire (intervention group) detailing affective attitude cost findings (N=49).
[PNG File , 56 KB - formative_v9i1e76517_app1.png]

Multimedia Appendix 2
The 12-month acceptability questionnaire (intervention group) detailing intervention coherence findings (N=49).
[PNG File , 55 KB - formative_v9i1e76517_app2.png]

Multimedia Appendix 3
The 12-month acceptability questionnaire (intervention group) detailing perceived effectiveness findings (N=49).
[PNG File , 55 KB - formative_v9i1e76517_app3.png]

Multimedia Appendix 4
The 12-month acceptability questionnaire (intervention group) detailing opportunity cost findings (N=49).
[PNG File , 50 KB - formative_v9i1e76517_app4.png]

References

1. Indrio F, Dargenio VN, Marchese F, Giardino I, Vural M, Carrasco-Sanz A, et al. The importance of strengthening mother and child health services during the first 1000 days of life: the foundation of optimum health, growth and development. J Pediatr 2022 Jun;245:254-6.e0. [doi: 10.1016/j.jpeds.2022.03.001] [Medline: 35259398]
2. Carson V, Lee EY, Hewitt L, Jennings C, Hunter S, Kuzik N, et al. Systematic review of the relationships between physical activity and health indicators in the early years (0-4 years). BMC Public Health 2017 Nov 20;17(Suppl 5):854 [FREE Full text] [doi: 10.1186/s12889-017-4860-0] [Medline: 29219090]
3. Eichner-Seitz N, Pate RR, Paul IM. Physical activity in infancy and early childhood: a narrative review of interventions for prevention of obesity and associated health outcomes. Front Endocrinol (Lausanne) 2023;14:1155925 [FREE Full text] [doi: 10.3389/fendo.2023.1155925] [Medline: 37293499]

XSL•FO
RenderX

4. Lee RL, Lane S, Brown G, Leung C, Kwok SW, Chan SW. Systematic review of the impact of unstructured play interventions to improve young children's physical, social, and emotional wellbeing. Nurs Health Sci 2020 Jun;22(2):184-196. [doi: 10.1111/nhs.12732] [Medline: 32358875]
5. Kracht CL, Redman LM, Casey PH, Krukowski RA, Andres A. Association between home environment in infancy and child movement behaviors. Child Obes 2021 Mar;17(2):100-109 [FREE Full text] [doi: 10.1089/chi.2020.0319] [Medline: 33471594]
6. Prioreschi A, Brage S, Hesketh KD, Hnatiuk J, Westgate K, Micklesfield LK. Describing objectively measured physical activity levels, patterns, and correlates in a cross sectional sample of infants and toddlers from South Africa. Int J Behav Nutr Phys Act 2017 Dec 22;14(1):176 [FREE Full text] [doi: 10.1186/s12966-017-0633-5] [Medline: 29273035]
7. Osborn CY, Paasche-Orlow MK, Bailey SC, Wolf MS. The mechanisms linking health literacy to behavior and health status. Am J Health Behav 2011;35(1):118-128 [FREE Full text] [doi: 10.5993/ajhb.35.1.11] [Medline: 20950164]
8. Lee JY, Murry N, Ko J, Kim MT. Exploring the relationship between maternal health literacy, parenting self-efficacy, and early parenting practices among low-income mothers with infants. J Health Care Poor Underserved 2018;29(4):1455-1471 [FREE Full text] [doi: 10.1353/hpu.2018.0106] [Medline: 30449757]
9. Proctor E, Silmere H, Raghavan R, Hovmand P, Aarons G, Bunger A, et al. Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda. Adm Policy Ment Health 2011 Mar;38(2):65-76 [FREE Full text] [doi: 10.1007/s10488-010-0319-7] [Medline: 20957426]
10. Teresi JA, Yu X, Stewart AL, Hays RD. Guidelines for designing and evaluating feasibility pilot studies. Med Care 2022 Jan 01;60(1):95-103 [FREE Full text] [doi: 10.1097/MLR.0000000000001664] [Medline: 34812790]
11. Cassidy S, Okwose N, Scragg J, Houghton D, Ashley K, Trenell MI, et al. Assessing the feasibility and acceptability of changing health for the management of prediabetes: protocol for a pilot study of a digital behavioural intervention. Pilot Feasibility Stud 2019;5:139 [FREE Full text] [doi: 10.1186/s40814-019-0519-1] [Medline: 31788325]
12. Ginsburg AS, Tawiah Agyemang C, Ambler G, Delarosa J, Brunette W, Levari S, et al. mPneumonia, an innovation for diagnosing and treating childhood pneumonia in low-resource settings: a feasibility, usability and acceptability study in Ghana. PLoS One 2016;11(10):e0165201 [FREE Full text] [doi: 10.1371/journal.pone.0165201] [Medline: 27788179]
13. van Heerden A, Sen D, Desmond C, Louw J, Richter L. App-supported promotion of child growth and development by community health workers in Kenya: feasibility and acceptability study. JMIR Mhealth Uhealth 2017 Dec 05;5(12):e182 [FREE Full text] [doi: 10.2196/mhealth.6911] [Medline: 29208588]
14. Meyer AJ, Armstrong-Hough M, Babirye D, Mark D, Turimumahoro P, Ayakaka I, et al. Implementing mHealth interventions in a resource-constrained setting: case study from Uganda. JMIR Mhealth Uhealth 2020 Jul 13;8(7):e19552 [FREE Full text] [doi: 10.2196/19552] [Medline: 32673262]
15. Brooke-Sumner C, Petersen I, Asher L, Mall S, Egbe CO, Lund C. Systematic review of feasibility and acceptability of psychosocial interventions for schizophrenia in low and middle income countries. BMC Psychiatry 2015 Feb 12;15:19 [FREE Full text] [doi: 10.1186/s12888-015-0400-6] [Medline: 25886524]
16. Klingberg S, Motlhatlhedi M, Mabena G, Mooki T, Verdezoto N, Densmore M, et al. "Must you make an app?" A qualitative exploration of socio-technical challenges and opportunities for designing digital maternal and child health solutions in Soweto, South Africa. PLOS Glob Public Health 2022;2(12):e0001280 [FREE Full text] [doi: 10.1371/journal.pgph.0001280] [Medline: 36962834]
17. Muke SS, Shrivastava RD, Mitchell L, Khan A, Murhar V, Tugnawat D, et al. Acceptability and feasibility of digital technology for training community health workers to deliver brief psychological treatment for depression in rural India. Asian J Psychiatr 2019 Oct;45:99-106 [FREE Full text] [doi: 10.1016/j.ajp.2019.09.006] [Medline: 31557716]
18. Sekhon M, Cartwright M, Francis JJ. Acceptability of healthcare interventions: an overview of reviews and development of a theoretical framework. BMC Health Serv Res 2017 Jan 26;17(1):88 [FREE Full text] [doi: 10.1186/s12913-017-2031-8] [Medline: 28126032]
19. Prioreschi A, Pearson R, Richter L, Bennin F, Theunissen H, Cantrell SJ, et al. Protocol for the PLAY Study: a randomised controlled trial of an intervention to improve infant development by encouraging maternal self-efficacy using behavioural feedback. BMJ Open 2023 Mar 07;13(3):e064976 [FREE Full text] [doi: 10.1136/bmjopen-2022-064976] [Medline: 36882258]
20. Soweto, Johannesburg. South African History Online. 2011. URL: https://sahistory.org.za/place/soweto-johannesburg [accessed 2025-08-18]
21. Bennin F, Theunissen H, Norris SA, Prioreschi A. Determining the perceived acceptability of an intervention designed to improve health literacy around developmentally appropriate play during infancy, with a community advisory group of mothers, in Soweto, South Africa. PLOS Glob Public Health 2024;4(8):e0002233. [doi: 10.1371/journal.pgph.0002233] [Medline: 39208036]
22. Dumville JC, Torgerson DJ, Hewitt CE. Reporting attrition in randomised controlled trials. BMJ 2006 Apr 22;332(7547):969-971 [FREE Full text] [doi: 10.1136/bmj.332.7547.969] [Medline: 16627519]
23. Osborne RH, Batterham RW, Elsworth GR, Hawkins M, Buchbinder R. The grounded psychometric development and initial validation of the Health Literacy Questionnaire (HLQ). BMC Public Health 2013 Jul 16;13:658 [FREE Full text] [doi: 10.1186/1471-2458-13-658] [Medline: 23855504]

XSL•FO
RenderX

24. Caçola PM, Gabbard C, Montebelo MI, Santos DC. Further development and validation of the Affordances in the Home Environment for Motor Development-Infant Scale (AHEMD-IS). Phys Ther 2015 Jun;95(6):901-923. [doi: 10.2522/ptj.20140011] [Medline: 25524875]
25. Hnatiuk J, Salmon J, Campbell KJ, Ridgers ND, Hesketh KD. Early childhood predictors of toddlers' physical activity: longitudinal findings from the Melbourne InFANT Program. Int J Behav Nutr Phys Act 2013 Nov 05;10:123 [FREE Full text] [doi: 10.1186/1479-5868-10-123] [Medline: 24188589]
26. Attride-Stirling J. Thematic networks: an analytic tool for qualitative research. Qual Res 2001 Dec 01;1(3):385-405 [FREE Full text] [doi: 10.1177/146879410100100307]
27. Silwanah AS, Suriah S, Jafar N, Areni IS, Yusuf RA. Content and features of Mobile health (mHealth) for mother and child nutrition in the first 1000 days of life (family based intervention): a systematic review. Nutr Health 2024 Dec;30(4):655-670. [doi: 10.1177/02601060241265550] [Medline: 39043216]
28. Yadav P, Kant R, Kishore S, Barnwal S, Khapre M. The impact of mobile health interventions on antenatal and postnatal care utilization in low- and middle-income countries: a meta-analysis. Cureus 2022 Jan;14(1):e21256 [FREE Full text] [doi: 10.7759/cureus.21256] [Medline: 35186540]
29. Knop MR, Nagashima-Hayashi M, Lin R, Saing CH, Ung M, Oy S, et al. Impact of mHealth interventions on maternal, newborn, and child health from conception to 24 months postpartum in low- and middle-income countries: a systematic review. BMC Med 2024 May 15;22(1):196 [FREE Full text] [doi: 10.1186/s12916-024-03417-9] [Medline: 38750486]
30. Kabongo EM, Mukumbang FC, Delobelle P, Nicol E. Explaining the impact of mHealth on maternal and child health care in low- and middle-income countries: a realist synthesis. BMC Pregnancy Childbirth 2021 Mar 09;21(1):196 [FREE Full text] [doi: 10.1186/s12884-021-03684-x] [Medline: 33750340]
31. Short VL, Bellad RM, Kelly PJ, Washio Y, Ma T, Chang K, et al. Feasibility, acceptability, and preliminary impact of an mHealth supported breastfeeding peer counselor intervention in rural India. Int J Gynaecol Obstet 2022 Jan;156(1):48-54 [FREE Full text] [doi: 10.1002/ijgo.13599] [Medline: 33454986]
32. Atukunda EC, Siedner MJ, Obua C, Musiimenta A, Ware NC, Mugisha S, et al. Evaluating the feasibility, acceptability, and preliminary efficacy of SupportMoms-Uganda, an mHealth-based patient-centered social support intervention to improve the use of maternity services among pregnant women in rural southwestern Uganda: randomized controlled trial. JMIR Form Res 2023 Mar 02;7:e36619 [FREE Full text] [doi: 10.2196/36619] [Medline: 36862461]
33. What is MomConnect? South African National Department of Health. URL: https://www.health.gov.za/momconnect/ [accessed 2025-08-18]
34. Hurt K, Walker RJ, Campbell JA, Egede LE. mHealth interventions in low and middle-income countries: a systematic review. Glob J Health Sci 2016 Sep 01;8(9):54429 [FREE Full text] [doi: 10.5539/gjhs.v8n9p183] [Medline: 27157176]
35. Aranda-Jan CB, Mohutsiwa-Dibe N, Loukanova S. Systematic review on what works, what does not work and why of implementation of mobile health (mHealth) projects in Africa. BMC Public Health 2014 Feb 21;14:188 [FREE Full text] [doi: 10.1186/1471-2458-14-188] [Medline: 24555733]
36. Soepnel LM, McKinley MC, Klingberg S, Draper CE, Prioreschi A, Norris SA, et al. Evaluation of a text messaging intervention to promote preconception micronutrient supplement use: feasibility study nested in the healthy life trajectories initiative study in South Africa. JMIR Form Res 2022 Aug 18;6(8):e37309 [FREE Full text] [doi: 10.2196/37309] [Medline: 35980731]
37. Coleman J, Eriksen J, Black V, Thorson A, Hatcher A. The mobile alliance for maternal action text message-based mHealth intervention for maternal care in South Africa: qualitative user study. JMIR Hum Factors 2020 Jun 29;7(2):e14078 [FREE Full text] [doi: 10.2196/14078] [Medline: 32459628]
38. Berkhout C, Berbra O, Favre J, Collins C, Calafiore M, Peremans L, et al. Defining and evaluating the Hawthorne effect in primary care, a systematic review and meta-analysis. Front Med (Lausanne) 2022;9:1033486 [FREE Full text] [doi: 10.3389/fmed.2022.1033486] [Medline: 36425097]
39. Miles LM, Elbourne D, Farmer A, Gulliford M, Locock L, McCambridge J, et al. Bias due to MEasurement Reactions In Trials to improve health (MERIT): protocol for research to develop MRC guidance. Trials 2018 Nov 26;19(1):653 [FREE Full text] [doi: 10.1186/s13063-018-3017-5] [Medline: 30477551]
40. Walters R, Leslie SJ, Polson R, Cusack T, Gorely T. Establishing the efficacy of interventions to improve health literacy and health behaviours: a systematic review. BMC Public Health 2020 Jun 30;20(1):1040 [FREE Full text] [doi: 10.1186/s12889-020-08991-0] [Medline: 32605608]
41. Hess Engström A, Flink M, Lindblom S, von Koch L, Ytterberg C. Association between general self-efficacy and health literacy among stroke survivors 1-year post-discharge: a cross-sectional study. Sci Rep 2024 Mar 27;14(1):7308 [FREE Full text] [doi: 10.1038/s41598-024-57738-z] [Medline: 38538651]

Abbreviations

CAG: community advisory group
FGD: focus group discussion
mHealth: mobile health

XSL•FO
RenderX

PLAY: Play Love and You
RA: research assistant
RCT: randomized controlled trial
REDCap: Research Electronic Data Capture

Edited by A Mavragani; submitted 25.04.25; peer-reviewed by T Schumacher, C Baxter; comments to author 28.05.25; revised version received 27.06.25; accepted 06.07.25; published 04.09.25.

Please cite as:
Bennin F, Norris SA, Prioreschi A
The Feasibility and Acceptability of an App-Based Intervention Aimed at Improving Maternal Health Literacy Regarding Infant Play and Development: Mixed Methods Study
JMIR Form Res 2025;9:e76517
URL: https://formative.jmir.org/2025/1/e76517
doi:10.2196/76517
PMID:40906520

©Fiona Bennin, Shane A Norris, Alessandra Prioreschi. Originally published in JMIR Formative Research (https://formative.jmir.org), 04.09.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Quality and Reliability of Adolescent Sexuality Education on Chinese Video Platforms: Sentiment-Topic Analysis and Cross-Sectional Study

Lan Wang[1], BM; Xiantao Shu[2], Prof Dr; Jianmei Huang[1], BM; Weiqian Yan[3], BSN; Duo Zhao[3], BM

[1]Department of Obstetrics and Gynecology, Huai'an 82 hospital, Huai'an, China

[2]Medical School of Yangtze University, Jingzhou, China

[3]Department of Orthopedics, The First People's Hospital of Guannan: Lianyun, Lianyungang, China

Corresponding Author:
Duo Zhao, BM
Department of Orthopedics
The First People's Hospital of Guannan: Lianyun
North Side of Xigang Avenue / East Side of Taizhou North Road, Guannan County
Lianyungang, 222500
China
Phone: 86 15298647736
Email: 78742728@qq.com

Abstract

Background: Adolescence is a critical period for lifelong health, which makes access to accurate and comprehensive sexuality education essential. As video platforms become a primary source of information for adolescents, the quality of their content significantly impacts their physical and mental health.

Objective: This study aimed to evaluate the quality, reliability, understandability, and actionability of adolescent sexuality education videos on major Chinese platforms (Bilibili, TikTok or Douyin, and Kwai), analyze associated user comment sentiment and topics, identify predictors of quality and reliability, and provide recommendations.

Methods: A cross-sectional analysis was conducted (April 2025) on the top 100 comprehensively ranked comprehensive sexuality education videos (N=300 total) retrieved from each platform using the keyword 青春期性教育 (“adolescent sexuality education”). Videos were assessed using the Global Quality Score, modified DISCERN, and Patient Education Materials Assessment Tool (PEMAT-U/A), with interrater reliability assessed via Cohen κ. A corpus of over 49,000 user comments underwent sentiment analysis (fine-tuned RoBERTa) and topic modeling (BERTopic, yielding 29 topics grouped into 6 themes). Statistical analyses included Kruskal-Wallis *H* tests, Spearman correlations, and stepwise linear regressions (SPSS [version 27.0]; $P<.05$).

Results: Video quality and reliability were moderate on Bilibili and TikTok but generally poor on Kwai. Content from verified sources (physicians, educators, and institutional media) demonstrated superior quality and stability compared to highly variable content from individual media (the predominant source type, especially on Kwai; 87/100, 87%). Paradoxically, Kwai exhibited the highest user engagement despite the lowest quality scores. Understandability (PEMAT-U) was consistently the strongest positive predictor for both quality (Global Quality Score, final model adjusted R^2=0.383, β=0.485) and reliability (modified DISCERN, final model adjusted R^2=0.209, β=0.319). Actionability (PEMAT-A) and video duration were also significant positive predictors. Understandability scores (PEMAT-U) were generally high (approximately 69%), while actionability scores (PEMAT-A) were moderate to low (33%-50%). Sentiment analysis revealed that comments were predominantly neutral (35,372/49,680, 71.2%), with negative comments (9141/49,680, 18.4%) significantly outweighing positive ones (5167/49,680, 10.4%). Key discussion themes identified included sources of knowledge acquisition, sexual safety and prevention, physiology, and sexual health and practices.

Conclusions: While online video platforms offer accessible channels for adolescent sexuality education in China, the current content is often of moderate-to-poor quality, with questionable reliability and limited actionability. Understandability is paramount, but high engagement does not necessarily correlate with high quality or reliability, potentially amplifying misinformation. To effectively empower youth, critical steps include enhancing content quality by adhering to evidence-based frameworks like the International Technical Guidance on Sexuality Education; strengthening platform accountability through improved verification

XSL•FO
RenderX

and algorithms; and promoting user media literacy. These measures aim to foster a healthier and more equitable future for Chinese adolescents, helping to achieve goals related to sexually transmitted infections and pregnancy prevention and promoting more open societal attitudes toward sexuality.

(JMIR Form Res 2025;9:e77100) doi:10.2196/77100

KEYWORDS

adolescent sexuality education; videos quality; Bilibili; TikTok; Kwai; Global Quality Score; GQS; modified DISCERN; mDISCERN; Patient Education Materials Assessment Tool-Audiovisual; PEMAT-A/V; information reliability; China; Spearman correlation analysis; stepwise regression analysis; topic modeling; sentiment analysis

Introduction

Background

Adolescence represents a critical developmental window for understanding sexuality, health, and relationships, shaped by complex sociobiological interactions [1]. Comprehensive sexuality education (CSE), grounded in human rights and gender equality principles as outlined by the International Technical Guidance on Sexuality Education (ITGSE) framework [2], is vital for equipping young people for safe, fulfilling lives [3]. Increasingly, adolescents turn to readily accessible online sources, particularly video platforms, for sexual health information [4,5]. This digital shift presents both significant opportunities for wide-reaching health communication and considerable challenges regarding content quality, accuracy, and impact.

Despite the potential of digital platforms for advancing sustainable development goals (SDGs) related to health (SDG 3), education (SDG 4), and gender equality (SDG 5), the quality and reliability of online CSE content are major global concerns. Consistent with ITGSE warnings, adolescents navigating the abundant digital landscape are susceptible to misinformation and biased perspectives, often amplified by unverified content creators (lacking platform-validated authentication of declared qualifications or professional status). Current digital content frequently lacks inclusivity, overlooking the specific needs and experiences of diverse youth, including lesbian, gay, bisexual, transgender, queer, two-spirit, quantum+ individuals and those with disabilities, groups who already face elevated risks of violence and barriers to Sexual and Reproductive Health information and care [6-8]. Furthermore, the typical short-video format, often prioritizing user engagement metrics over pedagogical substance, may fail to deliver the depth and progressive learning structure advocated by CSE frameworks such as ITGSE. The widespread availability of potentially harmful content, such as pornography, poses additional risks to healthy adolescent development, while the circulation of inaccurate reproductive health information can negatively impact health outcomes and trust in health care systems [9].

These challenges are particularly pronounced in China. The country boasts of the world's largest adolescent population [10]; however, it faces significant deficiencies in formal sex education infrastructure, including a scarcity of trained educators and age-appropriate materials. This gap is reflected in the low levels of sexual and reproductive health knowledge among young people [10,11]. This knowledge deficit is not trivial; it directly correlates with a range of severe public health issues, such as increased risks of adolescent pregnancies, the spread of sexually transmitted infections (STIs) and HIV, and heightened vulnerability to sexual violence. Specific vulnerable groups, such as rural left-behind children, encounter unique difficulties [12], and disparities exist between urban and rural adolescents in relevant knowledge [4], making them more susceptible to these negative health outcomes. The prevalent parental reluctance to openly discuss sexual topics further exacerbates these problems [13], leaving adolescents without necessary guidance and support when confronting sexual health challenges. In this context, online video platforms, such as Bilibili, Douyin, and Kuaishou have become primary information channels for young people [4]. While existing research has assessed the quality of online health videos for various other medical conditions [14-21], a systematic evaluation of adolescent CSE video quality, reliability, and user feedback on these major platforms—where millions of adolescents seek information—is conspicuously lacking, despite its clear public health significance.

This Study

Addressing this critical knowledge gap, this study aimed to comprehensively evaluate adolescent sexuality education video content disseminated via Bilibili, TikTok, and Kwai in China. Grounded in the ITGSE framework's tripartite foundation (knowledge, attitudes, and skills) [2], we assessed video quality (Global Quality Score [GQS]), reliability (mDISCERN), understandability (Patient Education Materials Assessment Tool-Understandability [PEMAT-U]), and actionability (Patient Education Materials Assessment Tool-Actionability [PEMAT-A]) using validated instruments and expert raters. Departing from solely content-focused analyses, we innovatively integrated computational methods by performing sentiment analysis and topic modeling (BERTopic) on a large corpus of associated user comments (approximately 49,000) to capture audience reception and discourse. By synthesizing evidence from video assessments, user feedback analysis, and platform comparisons, this research sought to identify current strengths, critical shortcomings, and strategic opportunities for enhancing digital CSE approaches to effectively empower adolescents.

Methods

Data Collection and Screening

This study used a cross-sectional approach. Between April 1 and April 4, 2025, new accounts were created on Bilibili, TikTok, and Kwai to perform video searches using the precise Chinese term 青春期性教育 (*qīngchūnqī xìngjiàoyù*,

XSL•FO
RenderX

"adolescent sexuality education"). For each platform, the 100 top-ranked videos according to the platform's internal algorithm were initially gathered for evaluation (Figure 1). Videos were included if they met three conditions: (1) the content addressed adolescent sexuality education, (2) the language used was Mandarin Chinese, and (3) the video length fell between 20 seconds and 20 minutes. Content was excluded based on several factors: (1) irrelevance to the research topic, (2) duplication or high similarity to other selected videos, (3) function as commercial advertising, (4) presentation as news reporting, (5) use of languages other than Chinese, (6) absence of audio or severely poor sound quality, and (7) uploads originating from unverified platform users. Application of these criteria resulted in a final sample of 300 videos, with an equal count of 100 videos derived from each of the 3 platforms (Bilibili, TikTok, and Kwai). The video content was independently reviewed and verified by 2 researchers. The subsequent analysis was intentionally restricted to these top 100 videos per platform, a decision informed by existing methodological studies, which indicate that videos ranking lower typically add minimal value and do not substantially change overall findings in comparable platform content research [22,23].

Figure 1. Flowchart for searching adolescent sex education videos.

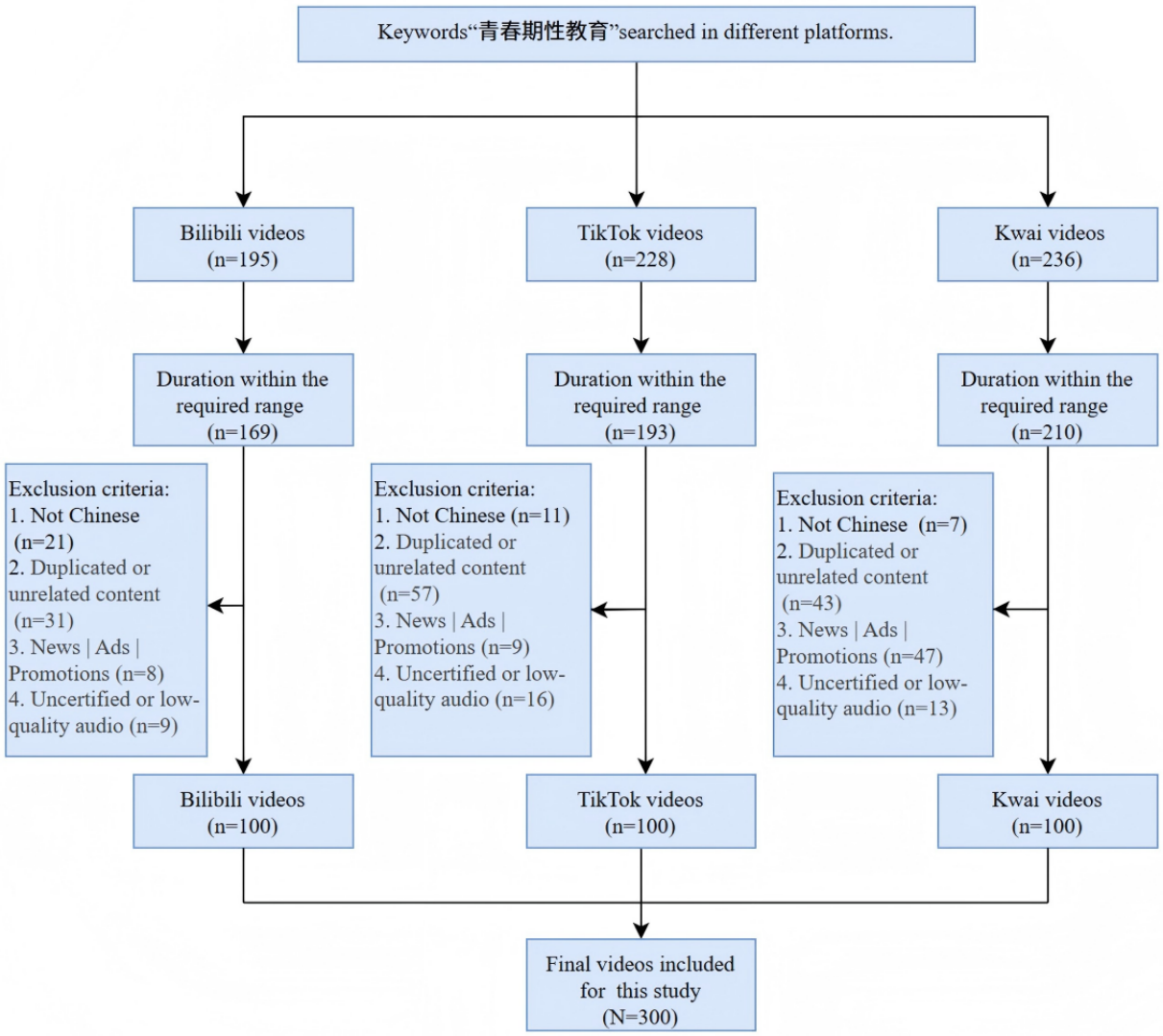

Data Collection Content

Key characteristics were documented for every video incorporated in the final sample. These systematically recorded parameters included the source or affiliation of the uploader, the video's runtime (duration), and several indicators of user interaction, such as counts for likes, favorites (also termed collections), shares, comments, along with the uploader's total number of followers. This dataset was organized and stored using Microsoft Excel software (Multimedia Appendix 1).

To enable systematic analysis, a video classification scheme was implemented. Each platform was assigned a numerical identifier (Bilibili=1, TikTok=2, and Kwai=3). On the basis of the principal subject matter, videos were sorted into five thematic content groups: (1) knowledge (covering aspects such as the evolution and current landscape of adolescent sexuality

XSL•FO
RenderX

education, its significance, physiological and psychological development during puberty, educational policies, and contraception use), (2) prevent sexual assault, (3) sexual behavior, (4) AIDS, and (5) other sexually transmitted diseases (STDs). Furthermore, the originators of the content were classified into four source types: (1) individual media (persons operating independently), (2) institutional media (entities, such as official bodies or established media companies), (3) physician (inclusive of psychologists, practitioners of modern medicine, and practitioners of traditional Chinese medicine), and (4) educator.

Web scraping techniques yielded an initial dataset of 86,647 user comments from 3 Chinese social media platforms: Bilibili, Douyin (TikTok), and Kwai. These raw data underwent a standardized cleaning protocol involving normalization, noise removal, and elimination of duplicate entries, resulting in a final analytic corpus containing roughly 49,000 comments. Concurrently, for the purpose of supervised model development, a distinct subset originating from 1999 raw comments was curated. After applying identical preprocessing steps, this subset was partitioned using stratified sampling into training (1699/1999, 85%) and validation (300/1999, 15%) datasets.

Quality Assessment Tools

GQS

Ranging from 1 to 5 points, with higher scores indicating superior quality [24], this metric evaluates accuracy, authority, completeness, and educational value.

Modified DISCERN Tool

Comprising 5 questions adapted from the original DISCERN (Decision Instrument for Screening Consumer Ehealth Resources aNd tools) tool, it is used to assess the reliability of health information [17], evaluating clarity, relevance, traceability, scientific basis, and impartiality. Each dimension is scored 1 point for “yes” and 0 points for “no” (Multimedia Appendix 2).

PEMAT

The PEMAT tool for audiovisual materials (PEMAT-A/V) with an automated scoring form assesses the understandability and actionability of videos [25].

Statistical Analysis

All statistical analyses were performed using SPSS software (version 27.0; IBM Corp). Key variables were summarized descriptively, presenting medians alongside IQRs. To examine variations between platforms concerning several metrics (such as video duration, likes, and collections), the nonparametric Kruskal-Wallis *H* test was applied. Relationships between video quality assessments and user engagement metrics were evaluated via Spearman rank-order correlation. Furthermore, predictive models were constructed through the application of stepwise linear regression techniques. Drawing figures with R (version 4.4.2; R Foundation for Statistical Computing), the degree of agreement between the 2 expert raters for quality scores was quantified using Cohen κ statistic. For all inferential statistical tests conducted, the threshold for significance was established at $P<.05$.

For classifying comment sentiment polarity, we used the hfl/chinese-roberta-wwm-ext model [26] after fine-tuning. Implementation relied on the transformers library [27], with training conducted over 3 epochs using a 3e-5 learning rate. The optimal model iteration was chosen based on maximizing the weighted F_1-score obtained on the designated validation dataset. To ensure the reliability and consistency of performance findings, the complete training and assessment cycle was executed 10 times independently, each initiated with a unique random seed (eg, 42, 13, 95, etc). Final performance measures were subsequently presented as means accompanied by their SDs.

Latent thematic structures within the comment data were identified using the BERTopic framework [28], which operated on document vector representations precalculated via the text2vec-base-chinese model [29]. This process incorporated dimensionality reduction via uniform manifold approximation and projection [30], set with parameters n_components=10 and min_dist=0.0, succeeded by density-based clustering using Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) [31], requiring a minimum cluster size of 180 comments. The resulting topic clusters were characterized by representative keywords derived from a class-based term frequency-inverse document frequency vectorization strategy, incorporating both single words and 2-word phrases (1-2 n-grams). Implementation of these core methodologies relied upon the Python programming language and its related ecosystem for scientific computation.

Ethical Considerations

This study strictly adhered to research ethics guidelines. This study constitutes a secondary analysis of publicly available data. All data were sourced from openly accessible videos on Bilibili, TikTok, and Kwai platforms. The research did not involve any clinical records, human biological samples, or animal experimentation data. The study exclusively used publicly visible video content from these platforms. No personally identifiable information or private user data were collected, processed, or analyzed. All data were anonymized to prevent identification of specific individuals. As this study did not involve direct contact with or intervention among platform users, and only used publicly available anonymized data, it was exempt from ethics review board approval in accordance with international ethical standards and institutional review board regulations. Since the research data were entirely obtained from public platforms and did not compromise user privacy, obtaining participant consent was not required.

Results

Videos Characteristics

A summary of the general features of the analyzed videos can be found in Tables 1 and 2. The nonparametric Kruskal-Wallis *H* test served to identify overall variations among Bilibili, TikTok, and Kwai regarding several key indicators (such as duration, likes, collections, shares, and comments). Findings from this test indicated that the platforms differed significantly concerning both user engagement measures and the assessed

XSL•FO
RenderX

scores for content quality and reliability. Therefore, the Dunn test was used for subsequent pairwise comparisons between the platforms to pinpoint specific differences. To control for the increased probability of type I errors arising from multiple comparisons, adjusted *P* values were calculated. Table 3 provides a detailed account of these post hoc comparison results.

As illustrated in Tables 4 and 5, which detail source and content distributions along with video features, individual media emerged as the most frequent source type overall. Creators operating independently contributed nearly two-thirds (199/300, 66.3%) of all videos analyzed across the Bilibili, TikTok, and Kwai platforms. This dominance by individual creators was especially pronounced on Kwai, where they accounted for 87% (87/100) of the content. Ranking significantly behind individual contributions were videos sourced from physicians (55/300, 18.3%) and educators (29/300, 9.7%), while institutional media provided the fewest videos (17/300, 5.7%). Distinct differences were observed between platforms regarding source prevalence. TikTok featured the highest concentration of content from physicians (33/100, 33%). Strikingly, the Kwai platform sample contained no videos produced by users identified as educators.

An examination of content themes revealed that "knowledge" concerning adolescent sexuality education was the most commonly encountered category; this theme accounted for 74% (74/100), 65% (65/100), and 75% (75/100) of the videos sampled from Bilibili, TikTok, and Kwai, respectively. Materials discussing "sexual behavior" also represented a substantial area of focus in the analyzed content.

Post hoc analyses revealed substantial differences between Bilibili and TikTok; specifically, the videos by Bilibili showed significantly elevated medians for runtime of 262 seconds (IQR 159-497), likes of 2345 (IQR 477-9354), collections of 1251 (IQR 183-3531), shares of 397 (IQR 72-1096), and comments of 189 (IQR 36-686) when contrasted with TikTok's content (all $P<.001$). When comparing Bilibili and Kwai, videos on the Bilibili platform maintained a significantly longer median duration ($P<.001$), although they garnered significantly fewer shares compared to those on the Kwai platform ($P<.05$). A comparison between TikTok and Kwai indicated that TikTok content received significantly less user engagement in terms of median likes of 121 (IQR 30-1066), collections of 40 (IQR 8-317), shares of 40 (IQR 9-500), and comments of 8 (IQR 2-51), all $P<.001$. However, the median video duration for these 2 platforms did not differ statistically (adjusted $P=.542$).

Assessments of quality and reliability revealed platform-specific characteristics. Bilibili videos registered a median GQS of 3 (IQR 3-4) alongside a median modified DISCERN (mDISCERN) score of 3 (IQR 2-3). In comparison, TikTok videos presented median scores of 3 (IQR 2-3) for GQS and 2 (IQR 2-3) for mDISCERN. The GQS scores for Bilibili and TikTok did not exhibit a statistically significant difference (adjusted $P=.052$), implying similar overall quality levels between these 2 platforms. However, when compared against Kwai, with a median GQS of 2 (IQR 2-3) and median mDISCERN of 2 (IQR 2-2), Bilibili content achieved significantly superior scores on both GQS and mDISCERN (both $P<.001$). A significant disparity in GQS was also identified between TikTok and Kwai ($P<.001$), but their reliability scores (mDISCERN) were statistically comparable (adjusted $P=0.200$). The consistency between the 2 independent assessors for the quality evaluations was measured via Cohen κ. Applying the benchmarks suggested by Landis and Koch [32] (κ>0.8: excellent; 0.6-0.8: substantial; 0.4-0.6: moderate; <0.4: poor), the obtained κ coefficients for GQS (κ=0.779) and mDISCERN (κ=0.786) both fall within the range indicating substantial interrater agreement.

Evaluation with the PEMAT-U indicated that video understandability was uniformly high across the studied platforms. Bilibili with a median of 69% (IQR 62%-77%), TikTok with a median of 69% (IQR 67%-77%), and Kwai with a median of 69% (IQR 63%-77%), all exhibited similar median scores with no statistically significant interplatform variation detected. In contrast, actionability (PEMAT-A) scores were considerably lower, ranging from moderate for Bilibili with a median of 50% (IQR 0%-67%) to low for TikTok with a median of 33% (IQR 33%-67%) and Kwai with a median of 33% (IQR 33%-59%), although these platform differences also failed to reach statistical significance. Taken together, these results indicate that while most videos were readily understandable, they generally lacked sufficient elements to empower viewers toward practical application or health-related decision-making, revealing a substantial need for improvement in content actionability.

XSL•FO
RenderX

Table 1. Characteristics of the videos on different platforms.

Variable	Bilibili (n=100), median (IQR)	TikTok (n=100), median (IQR)	Kwai (n=100), median (IQR)
Duration (s)	262 (159-497)	100 (63-179)	93 (55-160)
Likes	2345 (477-9354)	121 (30-1066)	7646 (375-39,400)
Collections	1251 (183-3531)	40 (8-317)	1011 (92-4402)
Shares	397 (72-1096)	40 (9-500)	960 (138-4206)
Comments	189 (36-686)	8 (2-51)	270 (15-3098)
Followers	43,000 (5258-226,250)	7272 (1356-127,750)	129,500 (25,000-793,250)
GQS[a]	3 (3-4)	3 (2-3)	2 (2-3)
mDIS[b]	3 (2-3)	2 (2-3)	2 (2-2)
PEMAT-U[c] (%)	69 (62-77)	69 (67-77)	69 (63-77)
PEMAT-A[d] (%)	50 (0-67)	33 (33-67)	33 (33-59)

[a]GQS: Global Quality Score.
[b]mDIS: modified DISCERN.
[c]PEMAT-U: Patient Education Materials Assessment Tool-Understandability.
[d]PEMAT-A: Patient Education Materials Assessment Tool-Accountability.

Table 2. Kruskal-Wallis H test across different platforms.

	Kruskal-Wallis *H* test (*df*)	Asymptotic significance[a]
Duration (s)	72.911 (2)	<.001
Likes	58.326 (2)	<.001
Collections	58.529 (2)	<.001
Shares	44.959 (2)	<.001
Comments	58.046 (2)	<.001
Followers	19.511 (2)	<.001
GQS[b]	41.343 (2)	<.001
mDIS[c]	20.672 (2)	<.001
PEMAT-U[d]	0.939 (2)	0.625
PEMAT-A[e]	2.159 (2)	0.34

[a]*P* values rounded to 3 decimal places. Values below .001 are reported as *P*<.001.
[b]GQS: Global Quality Score.
[c]mDIS: modified DISCERN.
[d]PEMAT-U: Patient Education Materials Assessment Tool-Understandability.
[e]PEMAT-A: Patient Education Materials Assessment Tool-Accountability.

XSL•FO
RenderX

Table 3. Dunn test across different platforms.

Comparison	1[a]-2[b] (Z)	1-2 (unadjusted *P* value)[c]	1-2 (adjusted *P* value)	1-3[d] (Z)	1-3 (unadjusted *P* value)	1-3 (adjusted *P* value)	2-3 (Z)	2-3 (unadjusted *P* value)	2 - 3 (adjusted *P* value)
Duration (s)	6.634	<.001	<.001	7.973	< .001	<.001	1.338	.181	.542
Likes	5.879	<.001	<.001	–1.282	.2	.6	–7.161	<.001	<.001
Collections	6.919	<.001	<.001	0.633	.527	.99	–6.286	<.001	<.001
Shares	4.057	<.001	<.001	–2.595	.009	.028	–6.652	<.001	<.001
Comments	6.178	<.001	<.001	–0.771	.441	.99	–6.95	<.001	<.001
Followers	2.101	.036	.107	–2.315	.0206	.0619	–4.415	<.001	<.001
GQS[e]	2.379	.017	.052	6.363	<.001	<.001	3.984	<.001	<.001
mDIS[f]	2.685	.007	.022	4.52	<.001	<.001	1.835	.067	.2
PEMAT-U[g]	0.613	.54	.99	0.957	.339	.99	0.344	.731	.99
PEMAT-A[h]	–0.999	.318	.955	0.435	.664	.99	1.433	.152	.456

[a]Bilibili.
[b]TikTok.
[c]Data are reported to 3 decimal places. *P* values less than 0.001 are indicated as *P*<001.
[d]Kwai.
[e]GQS: Global Quality Score.
[f]mDIS: modified DISCERN.
[g]PEMAT-U: Patient Education Materials Assessment Tool-Understandability.
[h]PEMAT-A: Patient Education Materials Assessment Tool-Accountability.

Table 4. Characteristics of the videos across sources.

Variables	Individual media (n=199)	Institutional media (n=17)	Physician (n=55)	Educator (n=29)
Duration (s), median (IQR)	136 (74-246)	210 (103-573)	90 (50-162)	199 (110-258)
Likes, median (IQR)	1803 (142-16,000)	286 (61-12,341)	458 (38-2359)	2328 (427-5574)
Collections, median (IQR)	532 (58-3313)	92 (11-2568)	200 (12-685)	825 (60-2535)
Shares, median (IQR)	435 (44-1877)	215 (24-5658)	147 (13-853)	374 (35-1341)
Comments, median (IQR)	89 (8-895)	18 (2-241)	17 (2-114)	175 (15-600)
Followers, median (IQR)	41,000 (1802-304,000)	94,000 (2428-825,500)	34,000 (2429-233,000)	58,000 (4997-110,500)
GQS[a], median (IQR)	3 (2-3)	3 (3-4)	3 (3-4)	4 (3-4)
mDIS[b], median (IQR)	2 (2-3)	3 (2-3)	2 (2-3)	3 (2-3)
PEMAT-U[c] (%), median (IQR)	69 (62-77)	77 (69-85)	69 (67-77)	69 (69-81)
PEMAT-A[d] (%), median (IQR)	33 (0-50)	67 (33-67)	50 (33-67)	50 (13-67)

[a]GQS: Global Quality Score.
[b]mDIS: modified DISCERN.
[c]PEMAT-U: Patient Education Materials Assessment Tool-Understandability.
[d]PEMAT-A: Patient Education Materials Assessment Tool-Accountability.

XSL•FO
RenderX

Table 5. Percentage distribution of adolescent sex education videos by source and content on different platforms.

Variable and subcategories	Bilibili, %	TikTok, %	Kwai, %
Source			
Individual media	65	47	87
Institutional media	6	8	3
Physician	12	33	10
Educator	17	12	0
Content			
Knowledge	74	65	75
Prevent sexual assault	7	1	3
Sexual behavior	14	21	18
AIDS	3	9	3
Sexually transmitted disease	2	4	1

Video Quality and Reliability Assessments

Platform comparisons highlighted differing quality profiles based on GQS assessments (Figure 2A). Content on Bilibili frequently scored between 3 and 4, denoting relatively high quality. TikTok's videos typically clustered around a score of 3, representing moderate quality. In contrast, Kwai content primarily ranged between 2 and 3, signifying comparatively lower quality. Reliability evaluations using mDISCERN (Figure 2B indicated moderate reliability for both Bilibili (scores concentrated between 2 and 3) and TikTok (median score of 2). However, Kwai presented a pattern skewed toward lower mDISCERN scores, suggesting inferior reliability overall. Synthesizing these results, while the overall quality difference between Bilibili and TikTok approached but did not reach statistical significance based on previous tests (adjusted P=0.052), the score distribution profile by Bilibili suggested stronger performance. Furthermore, Bilibili consistently demonstrated superior reliability compared to the other platforms. Kwai emerged as the platform with the lowest performance regarding both content quality and reliability metrics.

Examining scores based on the source of the video (Figure 2C and D), materials produced by institutional media, physicians, and educators typically clustered at the upper end of the GQS, signifying more dependable and higher-quality content. In contrast, while videos from individual media registered a median GQS of 3 (IQR 2-3) (Figure 2C), their distribution was notably bimodal, pointing to substantial variation in quality standards among these independent creators. With respect to reliability as measured by mDISCERN (Figure 2D), both institutional media and educators attained a median score of 3, outperforming the median of 2 recorded for the remaining source types.

When examining results based on content categorization (Figure 2E), videos covering knowledge, sexual assault prevention, sexual behavior, and AIDS typically achieved moderate GQS scores. In contrast, materials centered on STDs generally received evaluations corresponding to fair or lower-quality levels. Regarding reliability assessments (mDISCERN; Figure 2F), content specifically addressing AIDS distinguished itself by attaining relatively higher reliability scores compared to videos in the other thematic groups.

XSL•FO
RenderX

Figure 2. Global Quality Score (GQS) and modified DISCERN (mDISCERN) scores. (A) Comparison of GQS across different platforms. (B) Comparison of mDISCERN scores across different platforms. (C) GQS scores across different video sources. (D) mDISCERN scores across different video sources. (E) GQS scores across different content categories. (F) mDISCERN scores across different content categories..

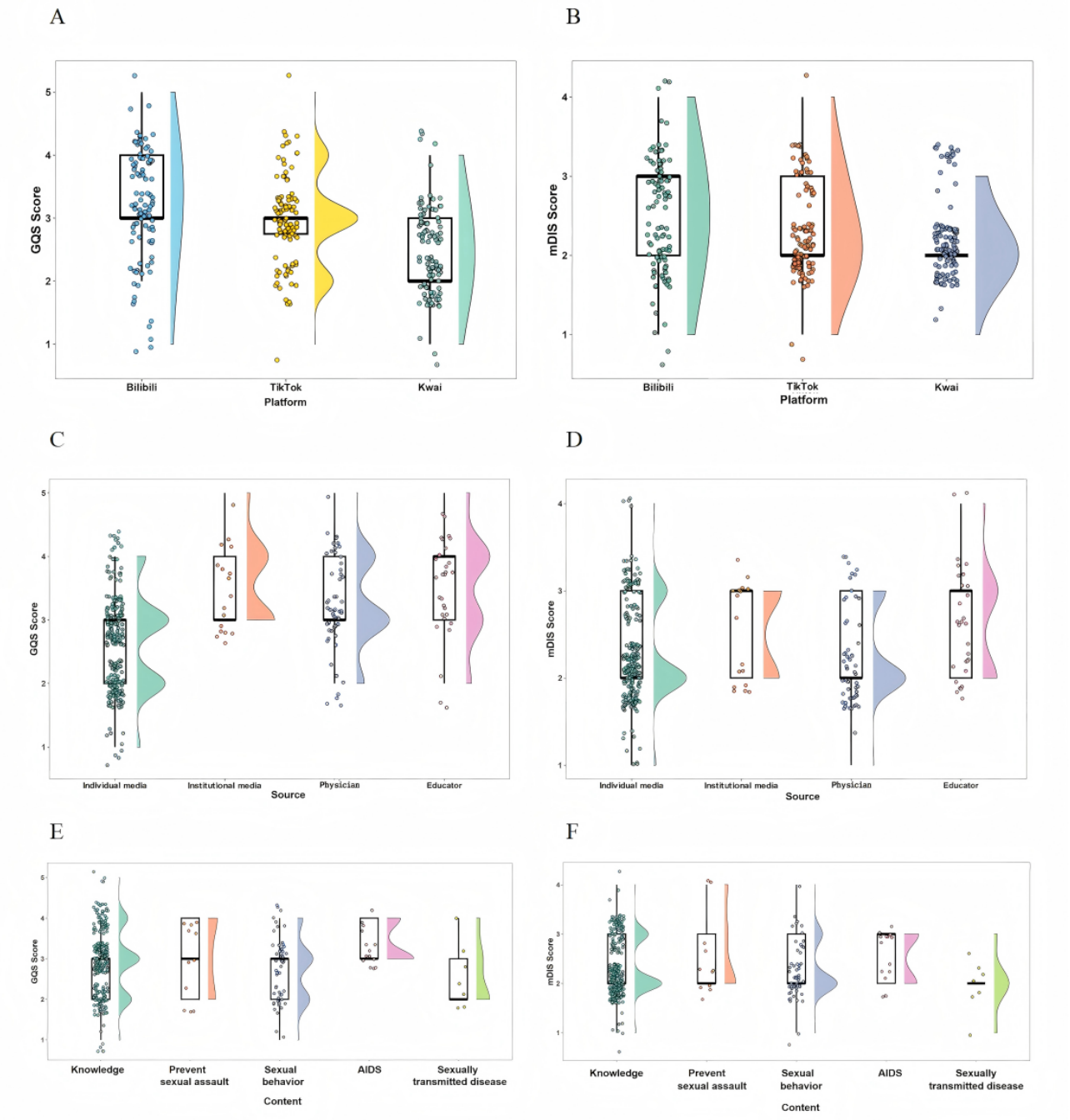

Video Understandability and Actionability

When PEMAT-U (understandability) scores were analyzed by platform (Figure 3A), a uniform median score of 69% was observed for Bilibili, TikTok, and Kwai. Despite identical medians, score distributions varied; TikTok's scores clustered more tightly at the higher end, implying greater consistency in understandability, while Bilibili displayed a more polarized distribution, possibly due to its diverse range of content creators. Kwai, whose content primarily originates from individual media, often without professional expertise, registered moderate understandability scores but demonstrated notably deficient performance regarding actionability (Figure 3B).

Analysis based on video source categorization yielded divergent profiles for understandability and actionability (Figure 3C and D). Notably, institutional media outperformed other sources across both metrics. Conversely, while exhibiting considerable spread in understandability scores, content originating from individual media demonstrated particularly weak actionability; PEMAT-A scores for this group clustered at lower values, with a median of just 33%. However, materials created by physicians and educators showed less score dispersion, suggesting more consistent results in terms of both conveying comprehensible information and providing actionable guidance.

XSL•FO
RenderX

When analyzed by thematic area, most videos demonstrated proficient levels of understandability (PEMAT-U), as shown in Figure 3E. However, significant variability was observed within the "knowledge" content category; despite a median PEMAT-U of 69%, the scores ranged broadly from a minimum of 33% to nearly 100%, indicating inconsistent effectiveness in information presentation for this theme. Assessments of actionability (PEMAT-A; Figure 3F implied) revealed generally insufficient performance. Content focused on STDs was particularly inadequate in this regard, while materials covering other topics typically displayed only moderate actionability.

Figure 3. Comparative analysis of video understandability and actionability. (A) Patient Education Materials Assessment Tool-Understandability (PEMAT-U) scores of videos across platforms. (B) Patient Education Materials Assessment Tool-Accountability (PEMAT-A) scores of videos across platforms. (C) PEMAT-U scores of videos across sources. (D) PEMAT-A scores of videos across sources. (E) PEMAT-U scores of videos across content categories. (F) PEMAT-A scores of videos across content categories.

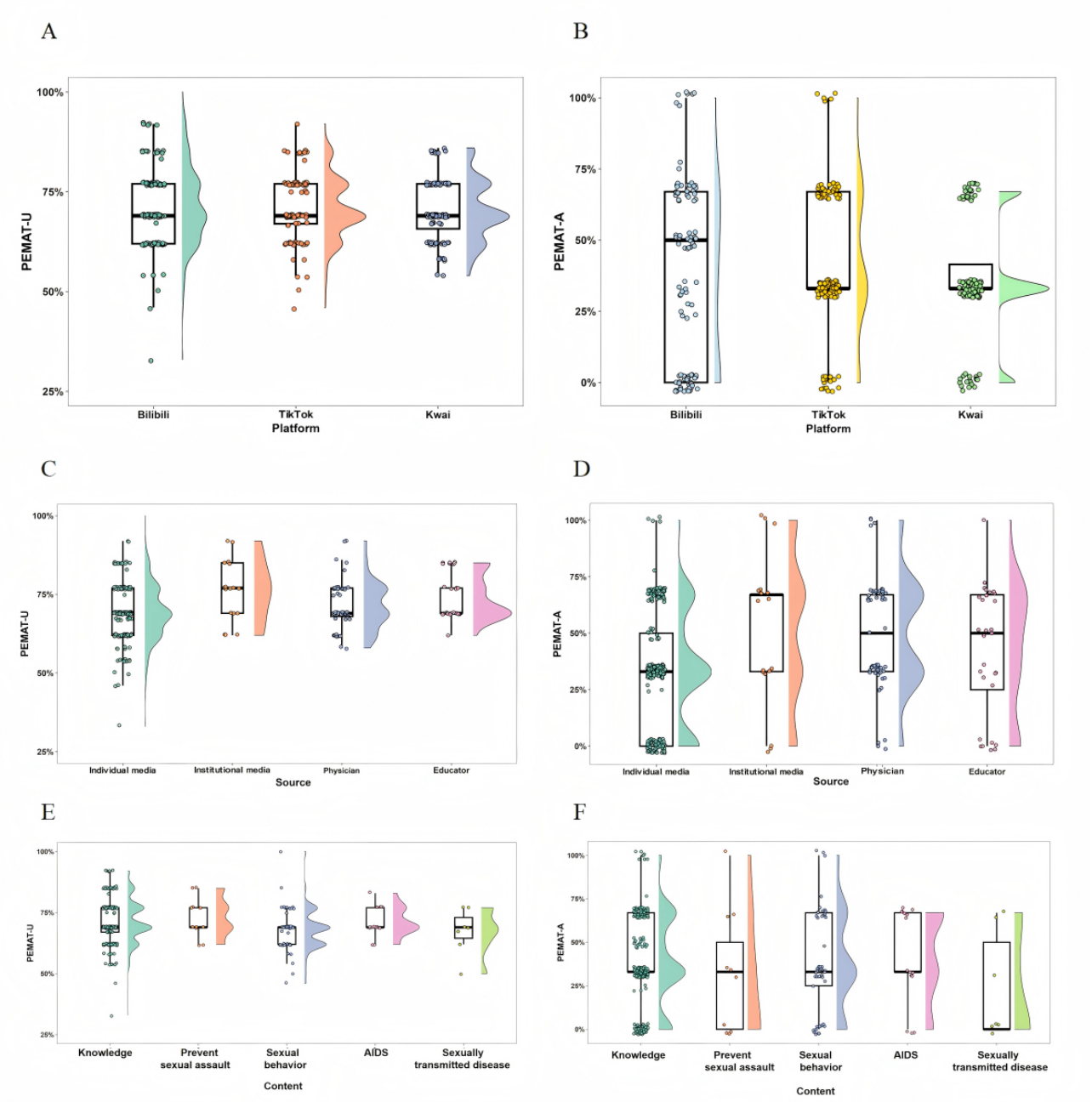

Correlation and Stepwise Regression Analysis

Further investigation into the interplay among variables involved Spearman rank correlation to evaluate connections between quality metrics (GQS and mDISCERN), user engagement levels, and video attributes; these associations are depicted in Figure 4. Results indicate a strong positive correlation between comments and likes, collections, and shares (r=0.83-0.94). A robust positive correlation was also observed between follower count and engagement metrics (r=0.56-0.64). Collectively, these indicators reflect content popularity and user engagement. The correlation between GQS and mDISCERN scores was r=0.46, with a significance level of P<.01, demonstrating a moderate positive relationship. Video duration showed correlations of r=0.29 with mDISCERN and r=0.33 with GQS.

XSL•FO
RenderX

Video understandability (PEMAT-U) exhibited a strong positive correlation with quality (GQS, r=0.53) and a moderate positive correlation with reliability (mDISCERN, r=0.34). Video actionability (PEMAT-A) showed a moderate positive correlation with GQS (r=0.30), but only weak positive correlations with mDISCERN and PEMAT-U (r=0.17 and r=0.15, respectively).

A stepwise linear regression procedure was implemented to ascertain key predictors for the GQS applied to adolescent sexuality education videos. This multistep process, detailed in Table 6, sequentially added variables maximizing explanatory contribution. The resultant final model (model 3) achieved statistical significance (F_3=62.790, P<.001), accounting for 38.3% of the GQS variance (adjusted R^2=0.383). Table 7 provides comprehensive details regarding the coefficients, significance levels, CIs, and collinearity checks for this concluding model. Findings confirmed positive and significant predictive roles for understandability (PEMAT-U: β=0.485, t_{296}=10.317, P<.001, two-tailed), actionability (PEMAT-A: β=0.232, t_{296}=5.040, P<.001, two-tailed), and video duration (β=0.186, t_{296}=3.979, P<.001, two-tailed) in determining GQS. The significance of these positive associations was further supported by the 95% CIs for their unstandardized coefficients (B), which excluded 0: PEMAT-U (0.036-0.053), PEMAT-A (0.004-0.010), and duration (0.000-0.001). Moreover, diagnostic tests confirmed low multicollinearity among these predictors (tolerance>0.93, VIF<1.08), indicating stable coefficient estimations.

Stepwise linear regression was additionally used to ascertain significant determinants of video reliability, quantified via the mDISCERN score. The progression of model construction and specifics of the final coefficients are documented in Tables 8 and 9. The culminating 3-predictor model demonstrated overall statistical significance (F_3=27.353, P<.001) and accounted for 20.9% of the variance observed in mDISCERN scores (adjusted R^2=0.209). Table 8 details the final model parameters, confirming significant positive relationships between mDISCERN and all included predictors: understandability (PEMAT-U: β=0.319, P<.001; 95% CI for B [0.014-0.027]), duration (β=0.242, P<.001; 95% CI for B [0.000-0.001]), and actionability (PEMAT-A: β=0.126, P=.016; 95% CI for B [0.000-0.005]). On the basis of standardized coefficients, PEMAT-U exerted the most substantial relative influence, succeeded by duration, and lastly PEMAT-A. Low multicollinearity among these variables (all VIF<1.08) affirmed the robustness of the coefficient estimates.

XSL•FO
RenderX

Figure 4. Correlation analysis. GQS: Global Quality Score; PEMAT-A: Patient Education Materials Assessment Tool-Accountability; PEMAT-U: Patient Education Materials Assessment Tool-Understandability;mDIS:modified DISCERN scores.

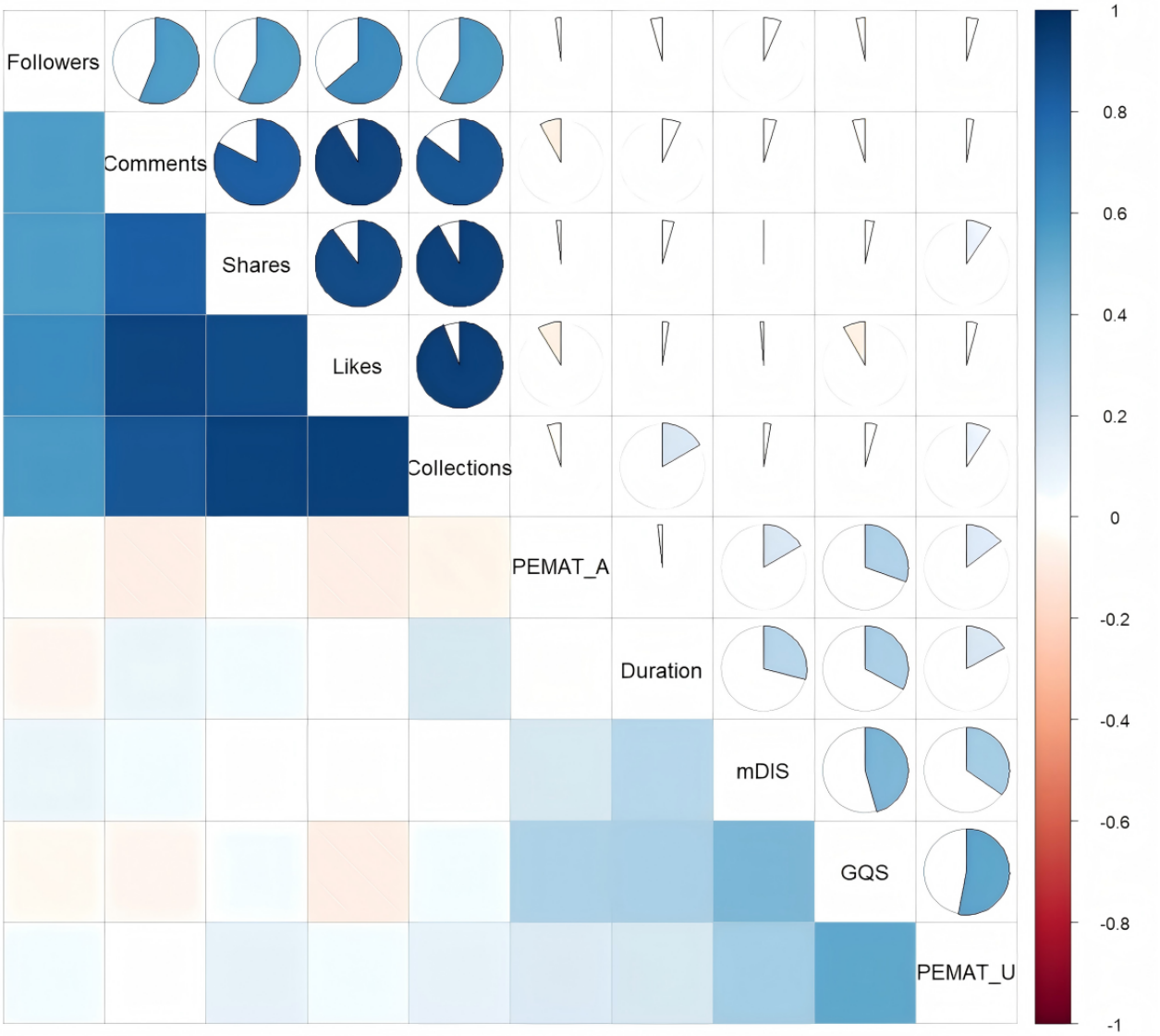

Table 6. Stepwise regression analysis (Global Quality Score).

Model	Predictors included	R^2	Adjusted R^2	SE of the estimate	*F* change	*P* value	ANOVA	
							F test(df)	*P* value
1	Constant and PEMAT-U[a]	0.31	0.307	0.677	133.674	<.001	133.674 (1,298)	<.001
2	Constant, PEMAT-U, and PEMAT-A[b]	0.356	0.352	0.655	21.472	<.001	82.165 (2,297)	<.001
3	Constant, PEMAT-U, PEMAT-A, and duration	0.389	0.383	0.639	15.833	<.001	62.790 (3,296)	<.001

[a]PEMAT-U: Patient Education Materials Assessment Tool-Understandability.

[b]PEMAT-A: Patient Education Materials Assessment Tool-Accountability.

XSL•FO
RenderX

Table 7. Stepwise regression coefficients, statistical significance, and collinearity assessment (Global Quality Score).

Model and predictors	Unstandardized coefficients		Standardized coefficients	*t* test (df)	*P* value	Collinearity statistics	
	B (95% CI)	SE	β			Tolerance	VIF[a]
Model 1							
(Constant)	−0.743 (−1.365 to −0.122)	0.316	__[b]	−2.354 (298)	0.019	—	—
PEMAT-U[c]	0.051 (0.043 to 0.060)	0.004	0.556	11.562 (298)	<.001	1.000	1.000
Model 2							
(Constant)	−0.800 (−1.401 to −0.198)	0.306	—	−2.615 (297)	0.009	—	—
PEMAT-U	0.049 (0.040 to 0.057)	0.004	0.527	11.207 (297)	<.001	0.981	1.019
PEMAT-A[d]	0.007 (0.004 to 0.009)	0.001	0.218	4.634 (297)	<.001	0.981	1.019
Model 3							
(Constant)	−0.690 (−1.279 to −0.100)	0.300	—	−2.302 (296)	0.022	—	—
PEMAT-U	0.045 (0.036 to 0.053)	0.004	0.485	10.317 (296)	<.001	0.933	1.072
PEMAT-A	0.007 (0.004 to 0.010)	0.001	0.232	5.040 (296)	<.001	0.976	1.025
Duration	0.001 (0.000 to 0.001)	0.000	0.186	3.979 (296)	<.001	0.949	1.054

[a]VIF: variance inflation factor.

[b]Not available.

[c]PEMAT-U: Patient Education Materials Assessment Tool-Understandability.

[d]PEMAT-A: Patient Education Materials Assessment Tool-Accountability.

Table 8. Stepwise regression analysis (modified DISCERN).

Model	Predictors included	R^2	Adjusted R^2	SE of the estimate	*F* change	*P* value	ANOVA	
							F test (df)	*P* value
1	Constant and PEMAT-U[a]	0.15	0.147	0.525	52.66	<.001	52.66 (1,298)	<.001
2	Constant, PEMAT-U, and duration	0.202	0.196	0.509	19.147	<.001	37.507 (2,297)	<.001
3	Constant, PEMAT-U, duration, and PEMAT-A[b]	0.217	0.209	0.505	5.827	0.016	27.353 (3,296)	<.001

[a]PEMAT-U: Patient Education Materials Assessment Tool-Understandability.

[b]PEMAT-A: Patient Education Materials Assessment Tool-Accountability.

XSL•FO
RenderX

Table 9. Stepwise regression coefficients, statistical significance, and collinearity assessment (modified DISCERN).

Model and predictors	Unstandardized coefficients		Standardized coefficients	*t* test (df)	*P* value	Collinearity statistics	
	B (95% CI)	SE	β			Tolerance	VIF[a]
Model 1							
(Constant)	0.550 (0.068-1.032)	0.245	__[b]	2.246 (298)	.025	—	—
PEMAT-U[c]	0.025 (0.018-0.032)	0.003	0.388	7.257 (298)	<.001	1.000	1.000
Model 2							
(Constant)	0.649 (0.179-1.119)	0.239	—	2.718 (297)	.007	—	—
PEMAT-U	0.022 (0.015-0.029)	0.003	0.338	6.371 (297)	<.001	0.955	1.048
Duration	0.001 (0.000-0.001)	0.000	0.232	4.376 (297)	<.001	0.955	1.048
Model 3							
(Constant)	0.631 (0.164-1.097)	0.237	—	2.661 (296)	.008	—	—
PEMAT-U	0.021 (0.014-0.027)	0.003	0.319	5.988 (296)	<.001	0.933	1.072
Duration	0.001 (0.000-0.001)	0.000	0.242	4.584 (296)	<.001	0.949	1.054
PEMAT-A[d]	0.003 (0.000-0.005)	0.001	0.126	2.414 (296)	.016	0.976	1.025

[a]VIF: variance inflation factor.

[b]Not available.

[c]PEMAT-U: Patient Education Materials Assessment Tool-Understandability.

[d]PEMAT-A: Patient Education Materials Assessment Tool-Accountability.

Topic Modeling and Sentiment Analysis

Validation on a separate dataset confirmed the stability and efficacy of the sentiment analysis model used. Robustness checks, involving 10 independent training and evaluation runs initiated with distinct random seeds, produced mean performance metrics indicative of strong reliability: a mean accuracy of 0.8070 (SD 0.0113) and a mean weighted F_1-score of 0.8017 (SD 0.0112) were achieved for the classification task. When deployed on the primary corpus (approximately 49,000 comments), the model classified most comments as neutral (35372/49680, 71.2%). Negative comments formed the next largest group (9141/49680, 18.4%), while positive comments were the least common category (5167/49680, 10.4%).

Application of the BERTopic algorithm to the comment corpus yielded 29 interpretable thematic clusters related to public perceptions and understanding of adolescent sexuality education, covering 87.3% (43371/49680) of the analyzed text. An additional, residual 12.7% (6309/49680) of comments fell into a miscellaneous category generated by the underlying bidirectional encoder representations from transformers (natural language processing) model for unclassifiable data, which was excluded from subsequent thematic analysis. These 29 meaningful topics were subsequently assigned to 6 overarching themes (Table 10). Examination of the prevalence associated with these topics offers direct insight into the subjects commanding the greatest public attention or dialogue. For instance, the considerable combined presence of topics within theme 3 (sources of knowledge acquisition) clearly signaled substantial public discussion and interest surrounding the channels—such as school, family, and online platforms—through which sexuality education is, or should be, delivered. Similarly, theme 2 (sexual safety education) emerged as significant, driven partly by the high individual prevalence of topic 0 (6857/43371, 15.8%), which focused on assault prevention and communication, thereby underscoring the public's strong concern with safety and risk avoidance strategies.

XSL•FO
RenderX

Table 10. Themes related to public perceptions of adolescent sex education, along with corresponding topics. (N=43,371).

Theme and topic	Keywords (top 10)	Comments, n (%)
Theme 1: adolescent physical development and physiological changes awareness		
Topic 3	menstruation, toilet, sanitary napkin, underwear, condom, clean, bathing, urethra, aunt[a], vagina	3018 (7.0)
Topic 4	cm, height, average, 18cm, 15cm, 5cm, per capita, length, gōngfēn[b], 14cm	2043 (4.7)
Topic 7	foreskin, retract, operation, to retract and expose, sensitive, expose, to have surgery, erection, the third kind, the second kind	1027 (2.4)
Topic 15	length, erection hardness, how long, size, dd[c], skill, sexual endurance, measure, backstab, length and girth	977 (2.2)
Theme 2: sexual safety education—contraception, sexual harassment prevention, and self-protection		
Topic 0	to resist, sexual assault, masculinity, lack, violence, rape, assault, victim, feminization, to talk about sex	6857 (15.8)
Topic 6	girl, elder sister, gal, little sister, a girl, měinǚ[d], pretty girl, respect for women, little girl, cross-dressing	1951 (4.5)
Topic 24	baby, hug, understood, sweetie, child, let me see, to force, to feel distressed for, got it, stay happy every day	467 (1.1)
Topic 26	to do well, protect, stay safe, be sure to protect, university student, measure, how old, protection, little angel, life	481 (1.1)
Topic 27	to go back, driving, to return home, to be late, send, close the door, hurry up, good person, to walk up onto, return first	508 (1.2)
Theme 3: sources of sexual knowledge acquisition—school-based, family-based, and online information channels		
Topic 2	biology lesson, biology teacher, second year of junior high school, first year of junior high school, lecture, to attend class, homeroom teacher, separate, grade 6, course	3153 (7.3)
Topic 5	to send out, in the group, to send to, what's up?, this book, classic, official, emmm[e], account, author	2173 (5.0)
Topic 9	popular science video, play, great, serious, excellent, to like, video, platform, detailed, awesome	1803 (4.2)
Topic 11	report, little child, reported by parents, parents and children, series, children and parents, children, complaint, count on, not teach	1769 (4.1)
Topic 19	mobile phone, computer, assignment, confiscate, play games, grades, to play games, National College Entrance Examination, to be addicted to, my little brother	949 (2.2)
Topic 21	self-taught talent, self-study, to figure out by oneself, study casually, study hard, focus on studying, study well, allow me, untaught, to acquire knowledge	551 (1.3)
Topic 23	to scroll and find, I browse, scroll, front page, to scroll and find a video, explosion, reflect, can scroll, check in, without reason or cause	426 (1.0)
Theme 4: sexual health, masturbation practices (benefits and risks), and impacts of pornography exposure		
Topic 1	sexual abstinence, desire, one week, moderate, to give up, one month, week, spirit, three times, harm	3736 (8.6)
Topic 8	Black person, bring, introduce, foreigner, foreign country/countries, Africa, to enter a country, take action, contagious, student studying abroad	998 (2.3)
Topic 18	to watch a movie, not watch, love to watch, TL:DR[6] (too long didn't read), watch too much, careful, fortunately, these words, take a look at a~, plot	1028 (2.4)
Topic 25	examine, traditional Chinese medicine (TCM), expert, treatment, doctor, regulate, experienced TCM doctor, to see a doctor, go see (a doctor) at the hospital, take medicine	710 (1.6)
Theme 5: sexual orientation, gender identity, and respect for diversity		
Topic 13	elder brother, guys, gay male, master, cute, unbelievable, my elder brother, invincible, eldest brother, driver	1145 (2.6)
Topic 16	have a meal, abnormal, review, while, devil, energy, music, to be at work, strawberry, public	1176 (2.7)

XSL•FO
RenderX

Theme and topic	Keywords (top 10)	Comments, n (%)
Topic 17	Acquired Immunodeficiency Syndrome, AIDS, infect, hpv, virus, transmit, test, male homosexuality, route, 'hiv	879 (2)
Topic 20	homosexuality, same-sex, group, heterosexuality, sexual orientation, discrimination, legal, proportion, gay man, gene	963 (2.2)
Theme 6: common psychological stressors and emotional regulation in adolescence		
Topic 10	responsible, exactly, responsibility, I will be responsible, department, don't believe, irresponsible, meet, believe men, don't understand	833(1.9)
Topic 12	joke, to joke, story, profile picture, can't understand, funny, to be funny, talented person, hee hee[g], upward	1601 (3.7)
Topic 14	magician, to be in love, deity, something years old, twenty, Beijing, twenty-something years old, mǔtāi[h], under, never held	1241 (2.9)
Topic 22	scary, terror, fear, whimper, really?, number, terrified, terrible, monster, unknown	544 (1.2)
Topic 28	love, Chinese, woc[i], real-name, have had, always have, lovely, and, jiao, me	364 (0.8)

[a]"姨妈" is a common colloquial term or euphemism for menstruation in mainland China, literally meaning "aunt."

[b]"公分" (gōngfēn) is a common, slightly more colloquial term for cm in Chinese, equivalent to "cm."

[c]"dd" is a common Chinese internet slang term or abbreviation, originating from "didi" (meaning "younger brother"), which is commonly used as a euphemism for penis.

[d]"měinǚ" is very commonly used in mainland China as a general term of address for young or middle-aged women, sometimes not necessarily referring to the person's outstanding appearance.

[e]"Emmm" is not a standard Chinese word, but rather mimics the sound of thinking, hesitation, or a slight pause, commonly found in online chat and spoken conversation.

[f]TL:DR is internet slang, meaning "save [data] flow."

[g]"Hee hee" is an onomatopoeia representing laughter, often carrying connotations of a chuckle, a silly giggle, or a slightly sly and mischievous undertone.

[h]The pinyin "mǔtāi" (for the characters 母胎, literally means "mother's womb") is often used in the internet slang phrase "mǔtāi solo" (母胎solo), meaning "(someone who has been) single since birth" or "has never dated."

[i]"Woc" is Chinese internet slang, usually a euphemistic substitute or abbreviation for the expletive "wǒ cào" (我操), used to express surprise, shock, dissatisfaction, or sometimes admiration, similar to interjections in English.

Discussion

Principal Findings

Knowledge

The ITGSE framework posits that CSE focused on "knowledge" should use a spiral, developmental curriculum covering 8 core conceptual areas to ensure learners progressively build essential understanding. Ideally, high-quality online videos would emulate this educational strategy. However, persistent gaps in sexual knowledge within target groups are evident. For example, data reported for 2023 by the Centers for Disease Control and Prevention showed a significant, concurrent drop in condom use among American youth, even as sexual activity and pregnancy rates fell, suggesting possible deficits in practical knowledge or risk perception [31]. Content available through popular digital platforms often fails to address these gaps comprehensively, frequently prioritizing topics perceived as high engagement (such as birth control methods) at the expense of fundamental CSE elements, such as STI prevention or discussions surrounding consent. This selective coverage occurs despite significant global health challenges, illustrated by recent Joint United Nations Program on HIV/AIDS (2024) statistics indicating 9.3 million people with HIV (including 43% of affected children) were without necessary medical care [6]. However, online video resources relevant to adolescent sexuality seldom offer thorough coverage of HIV prevention, thereby failing to fully use their potential for crucial public health messaging and service linkage in this critical area.

Attitudes

The ITGSE framework conceptualizes the "attitudes" component of sexuality education as fundamentally embedded within principles of human rights and gender equity. Therefore, optimal educational materials should purposefully confront harmful societal conventions and aim to diminish stigmatization. However, in practice, much online video content deviates from this ideal, often perpetuating established stereotypes and failing to address the distinct circumstances and requirements of marginalized groups, such as young people who identify as lesbian, gay, bisexual, transgender, queer+ or those with disabilities. The importance of tackling such disparities is highlighted by empirical evidence; for example, US data indicate Black female adolescents face a gonorrhea risk 8.8 times higher than their White counterparts, and young women with disabilities encounter double the likelihood of sexual assault compared to peers without disabilities [33]. By largely overlooking or failing to engage with these deep-seated structural inequalities, the capacity of current video content to foster significant attitudinal shifts and contribute to social change is substantially limited.

XSL•FO
RenderX

Skills

The ITGSE framework identifies "skills" development—encompassing areas such as interpersonal dialogue, assertion of boundaries (refusal), and accessing health care—as a key CSE component. While online video formats readily lend themselves to illustrating concrete actions (eg, demonstrating correct condom use), their utility often diminishes when addressing more intricate abilities, such as negotiation tactics or the cultivation of critical thinking about media messages. A relevant illustration of potential deficits in skills or adoption involves long-acting reversible contraceptives, including options such as intrauterine devices. These methods, despite proven high effectiveness, experience low global uptake [8,34]. This underuse is particularly striking given the comparatively modest use rates reported for condoms (approximately 38%) and oral contraceptive pills (approximately 27%) among young people in lower-resource settings [8,34]. Such discrepancies highlight an opportunity for well-crafted video interventions. By using lucid, engaging, and easily digestible presentation styles, videos could potentially narrow the divide between awareness of highly effective contraceptive options and their actual adoption or consideration by adolescents.

A correlation between platform publisher verification protocols and content quality is apparent; verified creators (eg, doctors) are prevalent on Bilibili and TikTok but rare on Kwai, contributing to the lower quality and trustworthiness of Kwai. The variations noted in video source distributions plausibly stem, at a minimum, from differing approaches to publisher verification across the platforms. TikTok, for example, mandates stricter validation for health care practitioners and imposes rigorous controls over medical advertising and pharmaceutical marketing. Kwai, conversely, typically requires only fundamental real-name confirmation for its individual content posters. Intriguingly, Kwai registered the most substantial user interaction metrics, notwithstanding its inferior content quality. This discrepancy suggests the potent effect of market-oriented algorithms prioritizing engagement, a dynamic risking magnification of inaccurate or false information. Approximately one-tenth of the examined videos propagated factual errors or misleading statements (eg, asserting human papilloma virus does not affect males or that condoms fail to prevent HIV transmission). Such findings highlight a profound digital-age hazard—readily available, unmediated online content can expose adolescents to significantly harmful viewpoints.

One particularly notable outcome of this investigation was identifying institutional media as the origin for 5.7% (17/300) of the analyzed videos. We believe this finding to be unique, as previous studies focusing on health information dissemination for other conditions via video platforms have not typically reported such a presence. This observation lends itself to a 2-fold interpretation regarding the state of adolescent sexuality education in China. First, it could indicate a heightened focus by governmental bodies on this area as a key public health issue. Alternatively, the very necessity for media channel involvement, coupled with the relatively small percentage observed, could paradoxically highlight persistent shortcomings or voids within the established educational systems tasked with providing CSE. Sexuality education requires careful instructional methods, focusing crucially on fostering attitude and perspective changes—often the most formidable barrier—beyond simply conveying information. Active parental engagement, including self-education and age-calibrated strategies, is also vital for an optimal learning environment. Internationally, China's provision significantly lags behind many Western nations, such as France, with its notably thorough curriculum [35]. Our analysis found that approximately 30% of the reviewed videos alluded to these shortcomings within the Chinese context. This reflects awareness that prevalent societal taboos ("tán sè xìng biàn") discourage open conversation, engendering widespread ignorance, reducing risk management abilities, and ultimately leading to avoidable negative consequences.

Studies confirm that while comprehensive and web-based sexuality education can effectively lower teenage birth rates, they typically exert only a marginal influence on altering actual sexual practices [11,36]. A significant complication for effective delivery is the notably low participation of health care providers, who are ideally positioned as sexual health advocates. This study highlights this specific gap, identifying a strikingly small percentage of physicians among video creators—a finding inconsistent with patterns noted in other scholarly work [23,35,37]. Supporting this, the 2016 clinical report by the American Academy of Pediatrics acknowledged the limited involvement of pediatricians in this domain. [38].

The low engagement of physicians on internet platforms is concerning, as it not only reflects a deficiency in the supply of educational content but also potentially exacerbates the knowledge disparity among adolescents in accessing accurate, professional sexual health information. This disparity may be influenced by socioeconomic or geographic factors, such as economic or regional differences in how frequently urban and rural children visit physicians, directly affecting their opportunities to obtain sexual health knowledge through traditional channels. If physicians, as sexual health advocates, fail to fully leverage internet platforms, adolescents who rely on online resources for information, especially those in medically underserved areas, may be more susceptible to inaccurate or incomplete sex education content, thereby widening health information inequities.

Future efforts should strive to rectify this imbalance, focusing on first, fostering unified home and school sexuality education strategies to ensure the dissemination of foundational knowledge, and second, and more critically, actively exploring how social media platforms can help correct this disparity. This could involve platform mechanisms that encourage more physicians to publish high-quality sexual health content and grant it greater visibility. By integrating physicians more deeply into educational settings (eg, "physicians-in-schools" programs or online expert lectures), young people's understanding of sexuality and reproductive well-being can be significantly bolstered.

Correlation and Stepwise Regression Analysis of Video Quality, Reliability, and Video Parameters

Correlation analysis revealed that a higher number of followers generally correlated with increased comments, shares, and likes,

XSL•FO
RenderX

which aligned with social media platform dissemination patterns. A moderate positive correlation (r=0.46) was observed between GQS and mDISCERN, suggesting that user-perceived quality and reliability change in parallel to some extent, though they are distinct concepts. Video duration also showed some association with information disclosure and quality scores. Longer videos may contain more information (higher mDISCERN), thus leading to higher-quality ratings (GQS). It is possible for a well-produced and easily understandable video (high GQS) to be deemed less reliable (low mDISCERN) due to an unclear source or lack of evidence. Understandability (PEMAT-U) serves as a crucial link, vital for enhancing both. A critical takeaway is the substantial potential for enhancing actionability across all identified content domains. Consequently, content developers are advised to focus on improving the incorporation of practical, actionable steps within knowledge-centric videos—most notably for subjects such as preventing sexual assault, contraception use, and STD and AIDS prevention—aiming to elevate both the user experience and the overall educational outcomes.

In concert, understandability (PEMAT-U), actionability (PEMAT-A), and duration were identified as significant factors predicting video quality (GQS), with understandability demonstrating the most substantial influence. While these same 3 variables also held predictive value for video reliability (mDISCERN), the model's overall capacity to explain variance was weaker in this case, and the specific impact of actionability was marginal (P=.02). The stepwise regression technique used offers benefits, such as automated variable selection, ease of use, suitability for initial exploration, and the possibility of achieving model parsimony. Nevertheless, despite its frequent application, this methodology possesses inherent constraints. A key limitation is its tendency to yield a model that is statistically optimal for the given dataset ("locally optimal") but may not align with the most theoretically sound or practically relevant model. Therefore, interpreting and extending the results derived from the final regression models generated via this approach warrants a degree of prudence.

Topic Modeling and Sentiment Analysis

Sentiment analysis of user comments revealed that discussions surrounding the sexuality education videos were generally neutral, mainly focusing on nonemotional factual exchange or discussion about the educational process, possibly also including user uncertainty or contradictory feelings. Notably, negative comments (9141/49,680, 18.4%) significantly outnumbered positive comments (5167/49,680, 10.4%). This may reflect users' concerns, dissatisfaction, or criticism regarding sensitive topics (such as safety, risks, prevention, etc), information quality, or personal experiences. Overall, the sentiment landscape highlights the challenges and potential anxieties faced by adolescent sexuality education in the digital space.

Topic modeling of over 43,000 comments revealed the aspects that most resonated with the audience. Users highly focused on the sources and methods of obtaining sexuality education (theme 3), and strongly focused on sexual safety and prevention (theme 2, especially sexual assault prevention [topic 0]). Discussions about physical development (theme 1) and sexual practices and health (theme 4, including masturbation and pornography impact) were also prominent, suggesting that these areas may lack sufficient information or be controversial. Although the proportion was lower, diversity and identity (theme 5) and psychological stress (theme 6) were also part of the discussion. The overlap between themes indicates that these topics are interconnected in public discussion.

Linking Engagement and Feedback

While the present regression models did not specifically test correlations between topic frequency and platform-level user interaction, broader results suggest that content quality and reliability may not align with audience engagement. Substantial interaction on platforms such as Kwai, often from lower-quality creators, may stem from a focus on provocative or less moderated themes (eg, physical development, sexual behavior, etc). Conversely, higher quality, reliable content from verified sources addresses more subtle CSE dimensions (eg, consent, diversity, and prevention). Consequently, their nuance and depth may result in lower mass interaction on platforms where algorithms prioritize engagement over content attributes.

Strengths and Limitations

This study analyzed adolescent sexuality education videos on major Chinese short-video platforms. Assessment by experts in medicine and education used GQS, mDISCERN, and PEMAT tools, focusing on video quality, reliability, clarity, and actionability. The study also innovatively integrated sentiment analysis and topic modeling of user comments to gauge audience reception and experience. Findings yield actionable recommendations for improving video content quality.

However, this study faces several limitations. First, the exclusive focus on Chinese-language content inherently limits the cross-cultural generalizability of our findings. Second, analyzing only the top 100 videos may introduce platform algorithm bias. Third, the applicability and validity of GQS and mDISCERN for this video genre require further validation. Finally, videos from institutional media and professional educators were underrepresented in the sample. Future research should address these limitations.

Despite screening the top 100 videos by comprehensive ranking to identify mainstream content, it is crucial to acknowledge the potential influence of platform algorithmic characteristics on video visibility and ranking. Douyin's algorithm, known for favoring short, visually impactful content, likely prioritizes videos aligning with these features, potentially leading to an underrepresentation of longer, more in-depth discussions. The algorithm by Kuaishou, while also emphasizing short-form video, incorporates a strong social-gifting and localized community focus, which can amplify content that drives immediate user interaction and caters to specific community preferences. In contrast, the algorithm by Bilibili prioritizes community engagement and is more compatible with medium-to-long-form content, potentially yielding a different video distribution.

This inherent algorithmic preference or bias can impact the generalizability of our findings regarding CSE information across the entire platform ecosystem and may introduce a

XSL•FO
RenderX

representational bias in the analyzed content. The phenomenon of “high engagement, low quality” is often exacerbated by these mechanisms. Algorithms optimize for metrics such as views, likes, and shares, which do not inherently correlate with factual accuracy or educational quality. This can lead to a feedback loop where engaging but potentially misleading content gains prominence. Future research should adopt differentiated sampling strategies that account for these varying algorithmic behaviors to more comprehensively capture the online adolescent sexual health information landscape.

Suggestions

This research provides valuable insights for enhancing online adolescent sexuality education videos. Drawing lessons from the success of evidence-based frameworks such as ITGSE [6], key recommendations emphasize improving content clarity (PEMAT-U) and actionability (PEMAT-A) through concrete guidance and building reliability via transparent sourcing, evidence citation, and neutrality. Incorporating diverse voices (eg, youth with disabilities and sexual or gender minorities) and considering appropriate video duration are vital for inclusivity and comprehensive information delivery. Video platforms should collaborate with authoritative institutions and experts to strengthen content and publisher verification and optimize recommendation algorithms to prioritize evidence-based, high-quality materials over simple engagement metrics, potentially offering credibility assessment tools. Fostering media and information literacy among educators and users is crucial for critical evaluation of online information and safeguarding health. Future research should further explore strategies for evaluating and improving the digital sexuality education ecosystem.

Conclusions

This investigation assessed the informational quality of adolescent sexuality education videos on China’s Bilibili, TikTok, and Kwai platforms. Key findings showed quality disparities. Bilibili and TikTok offered moderate-quality content, while Kwai had typically lower-quality content. Notably, videos from authenticated physicians and educators consistently achieved higher quality and reliability. This reflects a growing disparity in platform quality standards and professionalism. Entertainment-focused platforms such as Kwai are less advisable for reliable educational content compared to platforms such as Bilibili, which feature more academic material requiring critical user evaluation.

Directions for future research include exploring methods to improve platform-specific quality benchmarks and enhancing viewer media literacy for online health information. A challenge remains in reconciling professional credibility with broad appeal and accessibility. Video platforms hold significant, often unrealized, potential to transform adolescent sexuality education by featuring inclusive, skills-focused, evidence-based content.

Despite existing deficiencies, aligning with established guidelines, such as ITGSE, offers a clear path for advancement. These efforts can empower youth through quality education, enabling platforms to contribute to a healthier and more equitable future for Chinese adolescents, particularly by addressing goals related to STI education, pregnancy prevention, and fostering a more open societal attitude toward sexuality in China.

Data Availability

All data generated or analyzed during this study are included in this published paper and its multimedia appendices. The raw video and comment data analyzed are publicly available on the Bilibili, Douyin (TikTok), and Kwai platforms, subject to their terms of service and content availability.

Authors' Contributions

LW conceived and designed the study. JH, WY, and XS were responsible for reviewing and scoring the videos. DZ and LW collected and analyzed the data. LW prepared all tables and DZ prepared all figures. LW wrote the original draft. DZ reviewed the manuscript and provided critical revisions to the intellectual content. All the authors contributed to manuscript writing and editing and approved the final draft for submission.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Dataset on adolescent sex education.
[XLSX File (Microsoft Excel File), 43 KB - formative_v9i1e77100_app1.xlsx]

Multimedia Appendix 2
Quality assessment tools.
[DOCX File , 14 KB - formative_v9i1e77100_app2.docx]

References

XSL•FO
RenderX

1. Long JR, Damle LF. Adolescent sexuality. Obstet Gynecol Clin North Am 2024 Jun;51(2):299-310. [doi: 10.1016/j.ogc.2024.02.004] [Medline: 38777485]
2. International technical guidance on sexuality education. UNESCO Digital Library. URL: https://unesdoc.unesco.org/ark:/48223/pf0000260770_chi [accessed 2025-03-28]
3. Plesons M, De Meyer S, Amo-Adjei J, Casanova JV, Chipeta E, Jones N, et al. Protocol for a multi-country implementation research study to assess the feasibility, acceptability, and effectiveness of context-specific actions to train and support facilitators to deliver sexuality education to young people in out-of-school settings. Sex Reprod Health Matters 2023 Dec;31(2):2204043 [FREE Full text] [doi: 10.1080/26410397.2023.2204043] [Medline: 37436364]
4. Yang J, Zhu W. Investigation on adolescent sexual knowledge among primaryschool students in Hangzhou City. Chin Prevent Med J 2022;34(5):530-533 [FREE Full text] [doi: 10.19485/j.cnki.issn2096-5087.2022.05.022]
5. Döring N. [Sexuality education on the internet : from Dr. Sommer to Dr. Google]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2017 Sep;60(9):1016-1026. [doi: 10.1007/s00103-017-2591-0] [Medline: 28741187]
6. 2024 global AIDS report — the urgency of now: AIDS at a crossroads. Joint United Nations Programme on HIV/AIDS. 2024 Jul 22. URL: https://www.unaids.org/en/resources/documents/2024/global-aids-update-2024 [accessed 2025-04-10]
7. Bishop MD, Mallory AB, Gessner M, Frost DM, Russell ST. School-based sexuality education experiences across three generations of sexual minority people. J Sex Res 2021;58(5):648-658 [FREE Full text] [doi: 10.1080/00224499.2020.1767024] [Medline: 32486928]
8. Starrs AM, Ezeh AC, Barker G, Basu A, Bertrand JT, Blum R, et al. Accelerate progress-sexual and reproductive health and rights for all: report of the Guttmacher-Lancet Commission. Lancet 2018 Jun 30;391(10140):2642-2692. [doi: 10.1016/S0140-6736(18)30293-9] [Medline: 29753597]
9. John JN, Gorman S, Scales D, Gorman J. Online misleading information about women's reproductive health: a narrative review. J Gen Intern Med 2025 Apr;40(5):1123-1131. [doi: 10.1007/s11606-024-09118-6] [Medline: 39511120]
10. Zhao R, Zhang L, Fu XX. Sexual and reproductive health related knowledge, attitude and behavior among senior high school and college students in 11 provinces and municipalities of China. Chin J Public Health 2019;35(10):1330-1338. [doi: 10.11847/zgggws1124531]
11. Hu Z, Fu Y, Wang X, Zhang H, Guo F, Hee J, et al. Effects of sexuality education on sexual knowledge, sexual attitudes, and sexual behaviors of youths in China: a cluster-randomized controlled trial. J Adolesc Health 2023 Apr;72(4):607-615. [doi: 10.1016/j.jadohealth.2022.11.006] [Medline: 36604206]
12. Wang H, Jia X. Investigation on common sexual problems and sexual behavior of left-behind children during puberty. Chin J Sch Health 2020;41(5):676-679 [FREE Full text]
13. Guo H, Yi Y. Study on the current situation and influencing factors of family adolescent sex education: based on survey data from Guangzhou and Zhanjiang cities. China Sex Sci 2022;31(12):157-160.
14. Chen Y, Wang Q, Huang X, Zhang Y, Li Y, Ni T, et al. The quality and reliability of short videos about thyroid nodules on BiliBili and TikTok: cross-sectional study. Digit Health 2024 Oct 07;10:20552076241288831 [FREE Full text] [doi: 10.1177/20552076241288831] [Medline: 39381823]
15. Zhao X, Yao X, Sui B, Zhou Y. Current status of short video as a source of information on lung cancer: a cross-sectional content analysis study. Front Oncol 2024 Nov 22;14:1420976 [FREE Full text] [doi: 10.3389/fonc.2024.1420976] [Medline: 39650058]
16. Liu Z, Chen Y, Lin Y, Ai M, Lian D, Zhang Y, et al. YouTube/ Bilibili/ TikTok videos as sources of medical information on laryngeal carcinoma: cross-sectional content analysis study. BMC Public Health 2024 Jun 14;24(1):1594 [FREE Full text] [doi: 10.1186/s12889-024-19077-6] [Medline: 38877432]
17. Uzun O. Assessment of reliability and quality of videos on medial epicondylitis shared on YouTube. Cureus 2023 Apr 07;15(4):e37250 [FREE Full text] [doi: 10.7759/cureus.37250] [Medline: 37168186]
18. Ferhatoglu MF, Kartal A, Ekici U, Gurkan A. Evaluation of the reliability, utility, and quality of the information in sleeve gastrectomy videos shared on open access video sharing platform YouTube. Obes Surg 2019 May;29(5):1477-1484. [doi: 10.1007/s11695-019-03738-2] [Medline: 30706318]
19. Shackleford M, Horvath A, Repetto M, Thi A, Twells R, Sanders M, et al. An analysis of oral contraceptive related videos on TikTok. AJOG Glob Rep 2024 Jun 12;4(3):100364 [FREE Full text] [doi: 10.1016/j.xagr.2024.100364] [Medline: 39077679]
20. Stoddard RE, Pelletier A, Sundquist EN, Haas-Kogan ME, Kassamali B, Huang M, et al. Popular contraception videos on TikTok: an assessment of content topics. Contraception 2024 Jan;129:110300. [doi: 10.1016/j.contraception.2023.110300] [Medline: 37802460]
21. Zhang J, Yuan J, Zhang D, Yang Y, Wang C, Dou Z, et al. Short video platforms as sources of health information about cervical cancer: a content and quality analysis. PLoS One 2024 Mar 08;19(3):e0300180 [FREE Full text] [doi: 10.1371/journal.pone.0300180] [Medline: 38457419]
22. Mueller SM, Hongler VN, Jungo P, Cajacob L, Schwegler S, Steveling EH, et al. Fiction, falsehoods, and few facts: cross-sectional study on the content-related quality of atopic eczema-related videos on YouTube. J Med Internet Res 2020 Apr 24;22(4):e15599 [FREE Full text] [doi: 10.2196/15599] [Medline: 32329744]

XSL•FO
RenderX

23. Zheng S, Tong X, Wan D, Hu C, Hu Q, Ke Q. Quality and reliability of liver cancer-related short Chinese videos on TikTok and Bilibili: cross-sectional content analysis study. J Med Internet Res 2023 Jul 05;25:e47210 [FREE Full text] [doi: 10.2196/47210] [Medline: 37405825]
24. Kyarunts M, Mansukhani MP, Loukianova LL, Kolla BP. Assessing the quality of publicly available videos on MDMA-assisted psychotherapy for PTSD. Am J Addict 2022 Nov;31(6):502-507 [FREE Full text] [doi: 10.1111/ajad.13325] [Medline: 36000281]
25. PEMAT tool for audiovisual materials (PEMAT-A/V). Agency for Healthcare Research and Quality. URL: https://www.ahrq.gov/health-literacy/patient-education/pemat-av.html [accessed 2025-03-15]
26. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: a robustly optimized BERT pretraining approach. ArXiv Preprint posted online on July 26, 2019 [FREE Full text] [doi: 10.48550/arXiv.1907.11692]
27. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. HuggingFace's transformers: state-of-the-art natural language processing. ArXiv. Preprint posted online on October 9, 2019 2025 [FREE Full text] [doi: 10.48550/arXiv.1910.03771]
28. Grootendorst M. BERTopic: neural topic modeling with a class-based TF-IDF procedure. ArXiv. Preprint posted online on March 11, 2022 2025 [FREE Full text] [doi: 10.48550/arXiv.2203.05794]
29. Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using Siamese BERT-Networks. ArXiv. Preprint posted online on August 27, 2019 2025 [FREE Full text] [doi: 10.48550/arXiv.1908.10084]
30. McInnes L, Healy J, Melville J. UMAP: uniform manifold approximation and projection for dimension reduction. ArXiv. Preprint posted online on February 9, 2018 2025 [FREE Full text] [doi: 10.48550/arXiv.1802.03426]
31. Centers for Disease Control and Prevention. 2024 Aug 06. URL: https://www.cdc.gov/yrbs/dstr/index.html [accessed 2025-04-10]
32. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159-174. [Medline: 843571]
33. Lewinger S. Educating ideal neoliberal citizens: discourses of agency and responsibility in comprehensive sexuality education. Sex Reprod Health Matters 2023 Dec;31(1):2267203 [FREE Full text] [doi: 10.1080/26410397.2023.2267203] [Medline: 37887091]
34. Apter D. Contraception options: aspects unique to adolescent and young adult. Best Pract Res Clin Obstet Gynaecol 2018 Apr;48:115-127. [doi: 10.1016/j.bpobgyn.2017.09.010] [Medline: 29032945]
35. Cui N, Lu Y, Cao Y, Chen X, Fu S, Su Q. Quality assessment of TikTok as a source of information about mitral valve regurgitation in China: cross-sectional study. J Med Internet Res 2024 Aug 20;26:e55403 [FREE Full text] [doi: 10.2196/55403] [Medline: 39163110]
36. Mark ND, Wu LL. More comprehensive sex education reduced teen births: quasi-experimental evidence. Proc Natl Acad Sci U S A 2022 Feb 22;119(8):e2113144119 [FREE Full text] [doi: 10.1073/pnas.2113144119] [Medline: 35165192]
37. Zhang R, Zhang Z, Jie H, Guo Y, Liu Y, Yang Y, et al. Analyzing dissemination, quality, and reliability of Chinese brain tumor-related short videos on TikTok and Bilibili: a cross-sectional study. Front Neurol 2024 Oct 18;15:1404038 [FREE Full text] [doi: 10.3389/fneur.2024.1404038] [Medline: 39494168]
38. Breuner CC, Mattson G, Committee on Adolescence, Committee on Psychosocial Aspects of Child and Family Health. Sexuality education for children and adolescents. Pediatrics 2016 Aug;138(2):e20161348. [doi: 10.1542/peds.2016-1348] [Medline: 27432844]

Abbreviations

CSE: comprehensive sexuality education
GQS: Global Quality Score
ITGSE: International Technical Guidance on Sexuality Education
mDISCERN: modified DISCERN
PEMAT: Patient Education Materials Assessment Tool
SDG: sustainable development goal
STD: sexually transmitted disease
STI: sexually transmitted infection

XSL•FO
RenderX

Edited by A Mavragani; submitted 07.05.25; peer-reviewed by M Shackleford, SB Guo; comments to author 06.06.25; revised version received 18.06.25; accepted 09.07.25; published 05.09.25.

Please cite as:
Wang L, Shu X, Huang J, Yan W, Zhao D
Quality and Reliability of Adolescent Sexuality Education on Chinese Video Platforms: Sentiment-Topic Analysis and Cross-Sectional Study
JMIR Form Res 2025;9:e77100
URL: https://formative.jmir.org/2025/1/e77100
doi:10.2196/77100
PMID:

©Lan Wang, Xiantao Shu, Jianmei Huang, Weiqian Yan, Duo Zhao. Originally published in JMIR Formative Research (https://formative.jmir.org), 05.09.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Mental Health Care Provider Experiences of Remote Measurement-Based Care Rollout in an Urban Safety-Net Psychiatry Department: Three-Site Mixed Methods Hypothesis-Generating Implementation Study

Ana M Progovac[1,2,3], PhD; Carl Fulwiler[1,2], MD, PhD; Margaret Lanca[1,2], PhD; Daisy Wang[1], BS; Kate Zona[1,2], PhD; Norah Mulvaney-Day[1], PhD; Rajendra Aldis[1,2], MD; H Stephen Leff[1,2], PhD; Philip Wang[1,2], MD, DrPH; Lisa C Rosenfeld[1,2], MD, MPH

[1]Department of Psychiatry, Cambridge Health Alliance, Cambridge, MA, United States
[2]Department of Psychiatry, Harvard Medical School, Boston, MA, United States
[3]Institute for Excellence in Health Equity, NYU Langone Health, New York, NY, United States

Corresponding Author:
Ana M Progovac, PhD
Institute for Excellence in Health Equity
NYU Langone Health
180 Madison Ave
8th Floor
New York, NY, 10016
United States
Phone: 1 6465013450
Email: ana.progovac@nyulangone.org

Abstract

Background: Measurement-based care (MBC), including remote MBC, is increasingly being considered or implemented for mental health treatment and outcomes monitoring in routine clinical care. However, little is known about the health equity implications in real-world practice or the impact on patient-provider relationships in lower-resource systems that offer mental health treatment for diverse patients.

Objective: This hypothesis-generating study examined the drivers of MBC implementation outcomes, the implications for health equity, and the impact of MBC on therapeutic alliance (TA). The study was conducted 1 year after the implementation of remote MBC at 3 outpatient adult clinics in a diverse, safety-net health system.

Methods: This explanatory sequential mixed methods study used quantitative surveys and qualitative focus groups with mental health care providers. Repeated surveys were first used to understand mental health care provider experiences over a 6-month period, at least 1 year after MBC implementation. Surveys were analyzed to refine focus group prompts. Six mental health providers participated in repeated surveys over 6 months, after which the same 6 providers and 1 additional mental health provider took part in focus groups.

Results: Surveys revealed stable acceptability and utility ratings, concerns that MBC was not equally benefiting patients, little endorsement that MBC improved TA, and slightly decreasing feasibility scores. In focus groups, mental health care providers shared concerns about the acceptability, appropriateness, feasibility, and equity of processes for collecting MBC data. These providers had less first-hand experience with sharing and acting upon the data but still voiced concerns about the processes for doing so. TA both impacted and was impacted by MBC in positive and negative ways. The potential drivers of the findings are discussed using qualitative data.

Conclusions: More than 1 year after the implementation of remote MBC for mental health, mental health care providers had enduring concerns about its implications for health equity as well as its bidirectional relationship with TA. These findings suggest that further study is needed to identify system-level strategies to mitigate potential negative effects of real-world MBC implementations on health equity, particularly in low-resource settings with diverse populations.

(JMIR Form Res 2025;9:e71570) doi:10.2196/71570

XSL•FO
RenderX

KEYWORDS

measurement-based care; Collect-Share-Act model; therapeutic alliance; implementation science; safety net; mental health services

Introduction

Background

Measurement-based care (MBC) is an evidence-based approach that uses repeated collection of standardized, quantitative data to track progress on care goals and inform treatment planning [1,2]. In the mental health field, many types of measures can be used, including symptom screeners (eg, Patient Health Questionnaire-9), quality of life measures (eg, World Health Organization Quality of Life Brief Version), and functioning scales (eg, Practical and Social Functioning Scale). Regardless of the measure used, practicing MBC implies not only capturing trends in health data but also using these data in session to understand how the patient is doing and make shared decisions about next steps in care.

MBC, which has been used across a variety of mental health settings, has been linked to benefits such as improved health outcomes [3-6] and proposed as a means to reduce health disparities [7-9]. MBC may lead to improved outcomes compared to treatment as usual because it allows for more standardized communication between patients and health care providers [4]. Furthermore, MBC has the potential to enhance therapeutic alliance (TA; also referred to as "working alliance"), which is often considered a key mechanism for treatment success [10,11]. TA refers to the shared emotional bond, understanding of treatment goals, and conceptualization of the tasks needed to address these goals shared by patient and mental health care provider [12]. Potential mechanisms for MBC's positive impact on health disparities include reducing certain forms of bias in clinical judgment as well as improving health systems' abilities to identify differential needs and intervention effectiveness among patient subpopulations [4,13].

Despite its promise, MBC adoption in mental health remains low [4,14-18]. A survey of >800 psychologists published in 2007 who were members of the American Psychological Association found that 62.9% used no validated outcome measures in their work with patients [16]. An evaluation of >300 American psychiatrists published in 2008 revealed that a majority never or rarely used standardized scales, while just 6.5% used scales almost all the time [18]. As recently as 2018, a study of survey responses from >500 mental health counselors, marriage and family therapists, and social workers who were members of their respective professional societies found that the needle had not moved, with 61.5% never using validated outcome measures [17].

The Collect-Share-Act (CSA) model was developed as a framework to support mental health care provider adoption of MBC [19]. CSA breaks MBC into 3 distinct stages, each of which must be followed in sequence before progression to the next. In the *Collect* stage, mental health care providers are tasked with introducing MBC to patients and explaining the rationale for its use, then measuring the agreed-upon outcome repeatedly. In the *Share* stage, mental health care providers are asked to share and reflect on the measurement results with patients. In the *Act* stage, mental health care providers are directed to assess the patient's longitudinal trajectory and collaboratively determine next steps.

Most research on the use of MBC in mental health care has focused on relatively homogeneous patient populations with lower-acuity conditions, typically receiving care in small, stand-alone practices [20]. Furthermore, most work linking MBC with reductions in health disparities remains theoretical [21]; and national experts agree that there is no consensus about whether MBC, as currently designed and implemented, can assist with targeting and reducing mental health care disparities [22].

The paucity of data from higher-need, more diverse health systems raises the question of whether existing MBC literature applies to safety-net settings. The data that do exist suggest that MBC uptake is lower among older versus younger, racial and ethnic minority versus White, male versus female, severely versus nonseverely ill, and publicly versus privately insured individuals [3,23,24]. As patients with different demographic or diagnostic profiles may encounter group-specific challenges at any of the CSA stages of MBC, understanding whether and where these difficulties arise is essential for helping mental health care providers and mental health systems ensure access to the evidence-based practice of MBC for all patients.

Similarly, the relationship between MBC and TA in diverse real-world settings remains speculative in nature [22]. Evidence from randomized controlled trials and meta-analyses suggests that a strong TA between patients and their psychiatric providers improves mental health outcomes and that using MBC enhances patient-provider communication and hence shared decision-making [25-28]. However, to date, little is known about whether MBC enhances TA for patients from minoritized groups or whether strong TA can be leveraged to increase patient engagement with MBC. Demonstrating that MBC has a positive impact on TA among mental health care providers working with diverse patient populations could enhance buy-in among safety-net clinicians. Conversely, exploring the challenges that MBC poses to TA in these settings may allow mental health clinicians to anticipate and more effectively address barriers as well as remain committed to the evidence-based approach to care.

Objectives

In this explanatory sequential mixed methods study, we elicited feedback 1 year after the centralized implementation of an MBC tool called the Computerized Adaptive Test for Mental Health (CAT-MH). The study included 7 mental health care providers working with patients in an urban, ambulatory psychiatry department. Given the relative dearth of research exploring the on-the-ground realities of implementing MBC in safety-net health systems, the findings from this rich, small-scale study were intended to be hypothesis generating. Our goal was to better understand key phenomena that may drive disparate MBC uptake and varying real-world impact of MBC in safety-net settings. Specifically, we explored whether mental health care

XSL•FO
RenderX

providers found using MBC in these settings acceptable (ie, whether they liked using it with patients), appropriate (ie, whether they felt that it was relevant and compatible with their practice and patient population), and feasible (ie, whether they found it easy to implement in their practice) [29]. On the basis of unanswered questions in existing literature and our prior studies, we also sought to clarify how these implementation phenomena are perceived within each CSA stage and how MBC impacts health equity and TA.

Methods

Overview

This explanatory sequential mixed methods hypothesis-generating study used 3 repeated surveys over 6 months to understand provider experiences with MBC implementation and inform the focus group guide for the subsequent discussions. An overview of the study design is presented in Figure S1 in Multimedia Appendix 1 [19,30]. The results are jointly interpreted in the Discussion section, where potential avenues for further research are also presented.

Participants and Setting

This study was conducted in a diverse US urban safety-net system and training institution oriented toward team-based and psychosocial approaches to care. Patients receiving mental health services in this system were predominantly publicly insured, and many experienced chronic conditions that led to ongoing engagement in psychiatric care over the course of years.

All psychologists, social workers, psychiatrists, and psychiatric nurse practitioners who provided direct patient care at 3 adult-serving clinics were eligible for participation in the study. These included a general outpatient clinic, an addiction-focused clinic, and a transitional clinic that served individuals moving from inpatient to outpatient psychiatric care. Additional details on the demographic and diagnostic groups that each of these clinics served are provided in a prior publication [31].

These clinics were selected for inclusion in our study because they had been the sites of a centralized effort to roll out MBC in outpatient psychiatry in the year prior (2021). The providers at these clinics were expected to use session time to review MBC data with patients to assess their clinical progress and make shared decisions regarding next steps in care. The providers who were eligible for this study had at least 1 full year of experience with MBC after the implementation start date.

The MBC rollout involved the use of the CAT-MH, which has been validated against structured clinical interviews and shown to have high reliability and validity [32-34]. The CAT-MH draws from an item bank of nearly 1500 questions and uses multidimensional item response theory to assess 8 symptom domains [35]: depression, anxiety, mania and hypomania, psychosis, substance use disorder risk, posttraumatic stress disorder, attentional disorders, and a pilot quality-of-life module. Overall, the CAT-MH is reported to take about 2 minutes per domain to complete. The CAT-MH was delivered asynchronously through an app- or web-based electronic patient portal before an outpatient visit, with a minimum interval of 8 weeks between assessments. Module scores, including a severity indicator, were available on the electronic patient portal and electronic medical record. Individual item responses were not available for review once the survey was complete. Providers could graph the CAT-MH results using the electronic health record to see changes over time, although patients could not independently view their trend lines. Patient completion rates for MBC assessments in the first 6 months of implementation were 38% in the general outpatient clinic, 28% in the addiction-focused clinic, and 17% in the transitional clinic [31]. More detailed information on MBC completion trends is available elsewhere [31].

A key motivation for our study, which began in 2022 (a full year after the MBC rollout), was to understand how providers were experiencing MBC. Exposure to MBC came through 1-on-1 clinical work with patients, case discussions in weekly team meetings, and sporadic department-wide presentations (eg, grand rounds), some of which continued into early 2022. As implementing MBC is a complex process, we anticipated that it would take time for providers and patients to adjust to new workflows and that providers would have a better assessment of how these processes were impacting their work with patients, especially with regard to health equity, only after sufficient time had elapsed.

Ethical Considerations

This study received institutional review board approval from Cambridge Health Alliance (approval number CHA-IRB-20-21-6). Consistent with institutional policies at Cambridge Health Alliance, providers were not compensated financially for their participation in the study; however, they were permitted to block time in their work schedule to participate in the study. All participants gave written informed consent via DocuSign and were informed that study data would be shared anonymously.

Research Team

The research team was led by a health services researcher with PhD training (AMP) and a psychiatrist with an MD degree (LCR). It also included 3 other psychiatrists with MD degrees (CF, RA, and PW), 3 psychologists with PhD degrees (ML, KZ, and HSL), another PhD-trained health services researcher (NM-D), and an undergraduate research assistant (DW). Study leads AMP and LCR were trained in quantitative and mixed methods research in health services and implementation science. The study leads were women; the rest of the research team included 4 men and 4 women. At the time of the study, all members of the research team were employed by the safety-net health system as clinicians, researchers, clinical leaders, or part-time research assistant interns. Although the researchers had passing professional contact with study participants, only the study leads (AMP and LCR) and the research assistant (DW)—who had no supervisory roles related to participants—had direct contact with participants, knew their identities, or handled their identifiable data throughout the study duration. The study team had previously announced ongoing MBC evaluation efforts to all providers at the eligible sites, and these goals were explained again during recruitment.

XSL•FO
RenderX

Recruitment

All mental health care providers at the 3 sites were invited to participate via emails sent to each team. The study leads (AMP and LCR) also attended team meetings to introduce the study. The goals of the study, along with its voluntary nature, were reiterated in individual consent forms and in optional additional individual informed consent meetings.

Measures and Focus Groups

Repeated Surveys

Providers completed 3 identical 10- to 15-minute electronic surveys (via Google Forms) in May, July, and September 2022. The surveys included closed- and open-ended questions about MBC acceptability, adoption, feasibility, utility, equity, sustainability, and impact on TA (refer to the Results section for the survey items and question sources). Structured items were informed by the conceptual model of implementation research developed by Proctor et al [29], the implementation outcome measures formulated by Weiner et al [36], and the Working Alliance Inventory–Therapist version [37,38].

Qualitative Focus Groups

Three 1-hour web-based focus groups, each including 2 to 3 providers, were conducted between October and November 2022. They were led by the study leads (AMP and LCR) and designed to elicit narratives around key survey findings related to acceptability, appropriateness, feasibility, and equity within the CSA stages [19] as well as the impact of MBC on TA. The focus groups were recorded and transcribed verbatim (audio only). Transcripts were not returned to participants for correction but were quality checked by the first author (AMP). The study leads (AMP and LCR) compiled field notes after each session. The focus group guide was pilot-tested within the study team. No focus groups were repeated. Data saturation and informational richness [39] were prioritized via the longitudinal, explanatory mixed methods study design with key informant providers.

Mixed Methods Analysis

Quantitative Data Analysis to Inform Focus Groups

Survey data were collected via Google Forms. Although means and SDs were calculated using Google Sheets, quantitative data were not used for hypothesis testing. Instead, the researchers qualitatively assessed the distribution of individual responses, with no formal assessment of statistical significance. Descriptive assessment of survey responses across all 3 rounds informed the focus group guide. Examples of how individual items were qualitatively assessed are shown in Figure S2 in Multimedia Appendix 1, and how patterns of responses were used to inform the focus group guide are shown in Table S1 in Multimedia Appendix 1 (both of which are described in detail in the Data Integration subsection).

Qualitative Focus Group Data Analysis

Qualitative transcripts and memos were reviewed to generate preliminary thematic summaries and inform code tree development. Subsequently, the study leads (AMP and LCR) and the research assistant (DW) developed a code tree and code book with input from key study team members. The final code tree consisted of both deductive (a priori) codes based on the focus group guide and inductive codes, highlighting important emerging concepts (Table S2 in Multimedia Appendix 1).

Focus Group Coding, Including Valence Coding

Valence coding helps summarize and compare the influence of constructs on implementation outcomes [40]. Two primary coders (first author AMP and research assistant DW) coded focus group excerpts by applying up to 2 codes each for up to 4 structured levels in a Google Sheet (refer to Table S2 in Multimedia Appendix 1 for additional definitions). All excerpts were assigned a level 1 code (CSA or TA). For excerpts coded as a CSA stage at level 1, subsequent levels were as follows: level 2—acceptability, appropriateness, feasibility, or equity; and level 3—a valence code (−1, 0, +1, or N/A) to indicate how data from each excerpt impacted the implementation outcomes. For excerpts coded as TA at level 1, subsequent levels were as follows: level 2—MBC's impact on TA or TA's impact on MBC; and level 3—a valence code (−1, 0, +1, or N/A) to indicate the direction of impact (positive or negative).

Level 4 codes captured CSA subdomains, new inductive codes, and detail regarding the nature of MBC's relationship to TA. Codes were only applied where relevant (ie, could be left blank). To finalize the code tree and definitions, sample excerpts were also coded and reviewed by a third coder (senior author LCR), and additional input was obtained from the qualitative team during regular research meetings. For final coding, the 2 primary coders independently coded all excerpts; identified areas of coding incongruence; and resolved discrepancies through discussion, where possible. The third coder (senior author LCR) helped resolve remaining coding discrepancies via consensus. Final consolidated coding sheets contained up to 2 codes per excerpt.

Data Integration

The goal of this explanatory sequential mixed methods hypothesis-generating study was to generate a richer understanding (rather than to test hypotheses) of MBC implementation challenges at each CSA stage and how these challenges impact health equity and TA. Data integration occurred at 2 time points. First, provider surveys informed the focus group guide, including the selection of specific CSA substeps around which focus group prompts would be generated, the decision to include questions about the health equity implications at each CSA stage, and the decision to include targeted TA prompts to understand the direction and nature of the relationship between MBC and TA. Second, the qualitative data were used to explain and expand upon the observed quantitative findings.

Strategies to Ensure Data Trustworthiness

Our study incorporated 8 out of 10 recommended strategies for ensuring trustworthiness in qualitative research [41]. We used all 3 recommended strategies to ensure credibility, both recommended strategies to enhance transferability, and both recommended strategies to enhance dependability. While we only formally applied 1 of the 3 recommended strategies to ensure confirmability (ie, peer debriefing), our focus groups

XSL•FO
RenderX

drew inspiration from group member checking, in that we shared brief summaries of group-level findings from the quantitative data with focus group participants. This allowed the opportunity for providers to react to these group-level findings and offer contrasting perspectives during the focus groups.

Results

Participant Characteristics

Of the 26 eligible mental health care providers, 7 (27%) chose to participate in the study. Of these 7 participants, 1 (14%) chose to participate only in the focus group. Participants included psychologists, social workers, psychiatrists, and psychiatric nurse practitioners at 3 outpatient adult clinics. They were primarily female, White, and aged between 25 and 54 years; all provided care in English (Table 1). Participants had at least 1.5 years of experience in their roles at the organization before enrolling in this study, which meant that they were all involved in the MBC rollout for the full year before our study began.

Table 1. Participant characteristics (n=6[a]).

Characteristics	Participants, n (%)
Gender	
Men	1 (17)
Women	5 (83)
Nonbinary	0 (0)
Primary languages spoken at home	
English	6 (100)
Primary languages used in patient care	
English	5 (83)
English and Spanish	1 (17)
Race	
White	6 (100)
Ethnicity	
Non-Hispanic	6 (100)
Duration of employment at organization (years)	
<5	3 (50)
≥5	3 (50)
Specialization	
Clinical psychology or psychiatry[b]	3 (50)
Social work	3 (50)

[a]Of the 7 participants, 1 (14%) did not wish to provide demographic information.

[b]Psychiatry includes physicians and nurse practitioners.

Quantitative Surveys

All survey questions, question sources, and results from all 3 rounds of the quantitative surveys are presented in Table 2. All responses were on a Likert scale from 1 (completely disagree) to 5 (completely agree), unless otherwise noted.

Acceptability scores largely remained stable across the 3 rounds (round 1: mean 3.0, SD 0.7; round 2: mean 3.5, SD 0.6; and round 3: mean 3.3, SD 0.5). In examining adoption, only 1 (17%) of the 6 mental health care providers reported “never” reviewing CAT-MH results with patients over time, and providers generally reviewed CAT-MH results with patients only when new data were available (round 1: 4/5, 80%; round 2: 3/6, 50%; and round 3: 5/6, 83%). Mental health care providers were mostly evenly split between not reviewing CAT-MH results at all before visits (round 1: 2/5, 40%; round 2: 3/6, 50%; and round 3: 3/6, 50%) and spending 1 to 5 minutes reviewing these results (round 1: 3/5, 60%; round 2: 3/6, 50%; and round 3: 3/6, 50%). By round 3, most of the mental health care providers (4/6, 67%) reported spending 1 to 5 minutes during the patient visit reviewing CAT-MH results, whereas in round 1 there was a more even split across categories.

XSL•FO
RenderX

Table 2. Quantitative provider feedback on the implementation of the Computerized Adaptive Test for Mental Health (CAT-MH) implementation based on 3 surveys administered every 2 months between May and September 2022[a].

Domains and survey questions	Round 1 (n=5)	Round 2 (n=6)	Round 3 (n=6)
Acceptability, mean (SD)			
"I like reviewing CAT-MH data with patients"	3.0 (0.7)	3.5 (0.6)	3.3 (0.5)
Adoption, n (%)			
"I review CAT-MH results with patients"			
"Never"	0 (0)	1 (17)	0 (0)
"When there are new results"	4 (80)	3 (50)	5 (83)
"Only when the patient brings it up"	1 (20)	2 (33)	1 (17)
"At regular intervals"	0 (0)	0 (0)	0 (0)
"I spend approximately __ minutes reviewing CAT-MH results for a patient before the visit"			
0	2 (40)	3 (50)	3 (50)
1-5	3 (60)	3 (50)	3 (50)
5-10	0 (0)	0 (0)	0 (0)
>10	0 (0)	0 (0)	0 (0)
"I spend approximately __ minutes reviewing CAT-MH results for a patient during the visit"			
0	1 (20)	1 (17)	1 (17)
1-5	2 (40)	5 (83)	4 (67)
5-10	2 (40)	0 (0)	1 (17)
>10	0 (0)	0 (0)	0 (0)
Utility, mean (SD)			
"I trust the data generated by CAT-MH"	2.8 (0.8)	3.0 (0.9)	2.8 (0.4)
"CAT-MH data increases my confidence in a patient's diagnosis"	2.8 (0.8)	3.8 (0.4)	3.5 (0.8)
"CAT-MH data helps me and my patients set priorities for our visits"	2.8 (1.1)	3.5 (0.5)	3.3 (1.2)
"CAT-MH data impacts my treatment recommendations"	2.6 (1.3)	3.5 (0.5)	3.2 (1.0)
"Reviewing CAT-MH data during visits helps me and my patients collaborate on treatment decisions"	3.0 (1.2)	3.2 (0.8)	3.3 (0.8)
"CAT-MH data helps me assess whether a patient is responding to treatment"	2.8 (1.3)	3.3 (0.8)	2.7 (1.2)
Equity, mean (SD)			
"I believe that CAT-MH is benefitting all of my patients equally"	1.4 (0.5)	1.3 (0.5)	1.5 (0.5)
Therapeutic alliance, mean (SD)			
"Using CAT-MH has helped me feel like I better understand my patients"	2.6 (0.9)	3.0 (1.3)	3.0 (0.6)
"Using CAT-MH has helped me and my patients develop a common understanding of our treatment goals"	2.4 (1.1)	2.8 (1.2)	2.8 (1.0)
"CAT-MH has helped me and my patients build mutual trust"	1.6 (0.5)	2.3 (0.8)	2.3 (0.8)
Feasibility, mean (SD)			
"Discussing CAT-MH results with patients is doable in the time we have for visits"	3.6 (0.5)	3.5 (0.5)	3.3 (0.8)
Sustainability, mean (SD)			
"I believe that, over time, the benefits of using CAT-MH will outweigh the challenges of learning to use a new tool"	3.2 (0.4)	3.7 (0.5)	3.3 (0.5)
"I expect that, over time, incorporating CAT-MH into my day to day work will become: 1 (a lot harder) to 5 (a lot easier)"	3.4 (0.5)	3.8 (0.4)	3.7 (0.5)

[a]Unless otherwise noted, all Likert scale responses ranged from 1 (completely disagree) to 5 (completely agree), with a score of 3 (neither agree nor disagree) representing a neutral response. These data were assessed qualitatively to inform and refine the qualitative focus group guide and not for hypothesis testing; therefore, no formal statistical analysis is included here. Questions related to acceptability and feasibility were adapted from the

XSL•FO
RenderX

Acceptability of Intervention and Feasibility of Intervention measures [36]. Questions related to therapeutic alliance were adapted from the Working Alliance Inventory–Therapist version: "I feel like I really understand [patient name]," "[Patient name] and I have a common perception of his/her goals," "[Patient name] and I have built a mutual trust." The Working Alliance Inventory is described in Horvath and Greenberg [37].

The utility questions revealed that mental health care provider trust in CAT-MH data was largely neutral over time (round 1: mean 2.8, SD 0.8; round 2: mean 3.0, SD 0.9; and round 3: mean 2.8, SD 0.4). By survey rounds 2 and 3, these providers generally agreed that CAT-MH data increased their confidence in patient diagnoses (round 2: mean 3.8, SD 0.4; and round 3: mean 3.5, SD 0.8), helped them set priorities during visits (round 2: mean 3.5, SD 0.5; and round 3: mean 3.3, SD 1.2), impacted treatment recommendations (round 2: mean 3.5, SD 0.5; and round 3: mean 3.2, SD 1.0), and facilitated collaboration with patients on treatment decisions (round 2: mean 3.2, SD 0.8; and round 3: mean 3.3, SD 0.8). Except in round 2, these providers were unlikely to agree that CAT-MH data helped them assess patients' response to treatment (round 1: mean 2.8, SD 1.3; and round 3: mean 2.7, SD 1.2).

Regarding questions about equity, mental health care providers consistently reported that the CAT-MH was not benefiting all patients equally (round 1: mean 1.4, SD 0.5; round 2: mean 1.3, SD 0.5; and round 3: mean 1.5, SD 0.5). For TA-related questions—which explored whether CAT-MH data helped these providers better understand their patients, develop a common understanding of treatment goals, and build mutual trust—responses averaged between 1.6 (SD 0.5) and 3.0 (SD 0.6) and generally trended upward over time.

Feasibility scores decreased slightly over the 3 rounds (round 1: mean 3.6, SD 0.5; round 2: mean 3.5, SD 0.5; and round 3: mean 3.3, SD 0.8). Mean sustainability scores, as measured by the mental health care providers' sense that benefits would outweigh challenges and that workflows would become more manageable over time, were stably neutral to positive, ranging from 3.2 (SD 0.4) to 3.8 (SD 0.4).

Data Integration

The study leads organized survey findings in a heat map to identify priority areas for the focus group guide (Figure 1). Figure S2 in Multimedia Appendix 1 shows how the study leads examined the distribution of individual item responses over time to flag responses that generally signaled "good" (white), "neutral" (light gray), or "bad" (dark gray) implementation outcomes. "Bad" response profiles, particularly when consistent over time, were prioritized for exploration in focus groups.

Table S1 in Multimedia Appendix 1 provides a high-level description of how the simplified quantitative results were used to select specific CSA subdomains for the focus group guide as well as to generate the questions about equity and TA that emerged from the survey responses. Sample focus group prompts are also presented in Table S1 in Multimedia Appendix 1. Key focus group guide decisions made based on these data included prioritizing equity prompts for each CSA stage, narrowing to CSA subdomain questions with the most relevance to the study context and quantitative findings, and eliciting narratives about the impact of MBC on TA.

Data integration also occurred during qualitative data analysis by incorporating equity as an implementation outcome for valence analysis at each CSA stage (rather than as a stand-alone theme) and coding TA as a stand-alone domain (ie, a level 1 factor). The first decision was made because provider responses indicated unique equity challenges at each CSA stage, and we felt that this was important to disaggregate in valence coding. The second decision was made because the relationship between MBC and TA emerged as more complex and overarching than we felt was captured within the existing CSA stages or implementation outcomes.

Figure 1. Heat map of quantitative survey results to inform qualitative focus groups. The heat map represents the authors' qualitative assessment of quantitative survey results, including distributions of answers within each survey round and perceived distance from "ideal implementation" (eg, in the case of adoption). "Good" response profiles were coded in white, "bad" implementation outcome profiles were coded in dark gray, and neutral response profiles were coded in light gray. Each row corresponds to responses for the survey questions listed in Table 2, in order of appearance.

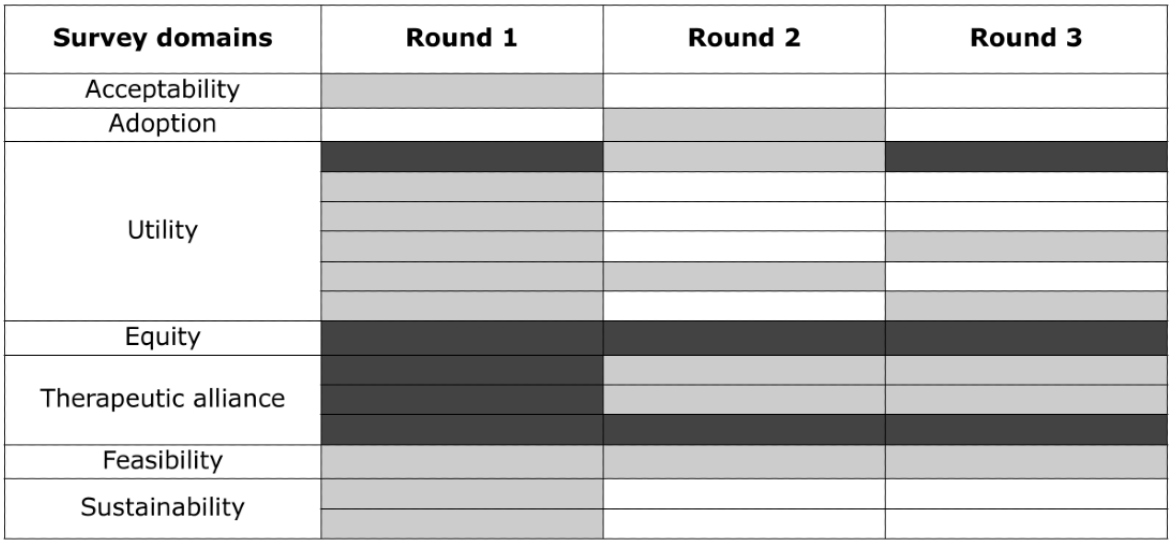

XSL•FO
RenderX

Qualitative Focus Groups

Valence Coding Results

Valence coding results are presented by focus group domain. CSA stages of MBC and implementation outcomes (acceptability, appropriateness, feasibility, and equity) are presented in Table 3. The TA domain and its subdomains (MBC impacts TA and TA impacts MBC) are presented in Table 4. Providers gave the most feedback about the *Collect* stage (83 excerpts), followed by *Share* (68 excerpts) and *Act* (28 excerpts).

Within *Collect*, feedback by outcome was mixed, but the overall valence was negative (83 excerpts; mean valence −0.77, SD 0.53), largely driven by very negative ratings for feasibility (mean valence −0.89, SD 0.31) and equity (mean valence −0.96, SD 0.20); ratings for acceptability (mean valence −0.64, SD 0.70) and appropriateness (mean valence −0.39, SD 0.72) were less negative.

Share had a relatively neutral valence (68 excerpts; mean valence 0.03, SD 0.91). Provider experiences mostly revealed instances of neutral to positive acceptability (mean valence 0.29, SD 0.90) and appropriateness (mean valence 0.19, SD 0.85) paired with lower feasibility (mean valence −0.33, SD 0.91). Although there were experiences pointing to equity barriers in the *Share* stage (mean valence −1.00, SD 0.00), they were fewer in number than in the *Collect* stage (3/68 excerpts, 4% for *Share* vs 26/83 excerpts, 31% for *Collect*).

Table 3. Qualitative valence coding results by Collect-Share-Act (CSA) domain from focus groups and implementation outcome[a].

CSA domains from focus groups	Acceptability	Appropriateness	Feasibility	Equity	Overall[b]
Collect, mean (SD)	−0.64 (0.70)	−0.39 (0.72)	−0.89 (0.31)	−0.96 (0.20)	−0.77 (0.53)
Collect Excerpts (n=83), n (%)	11 (13)	18 (22)	28 (34)	26 (31)	83 (100)
Share, mean (SD)	0.29 (0.90)	0.19 (0.85)	−0.33 (0.91)	−1.00 (0.00)	0.03 (0.91)
Share Excerpts (n=68), n (%)	21 (31)	26 (38)	18 (26)	3 (4)	68 (100)
Act, mean (SD)	1.00 (0.00)	0.25 (0.89)	−0.07 (1.00)	−1.0 (0.0)	0.11 (0.96)
Act Excerpts (n=28), n (%)	4 (14)	8 (29)	14 (50)	2 (7)	28 (100)

[a]Valences are from −1 to 1, with −1 being a maximally negative valence and 1 being a maximally positive valence.

[b]Overall means and SDs were calculated using all valence data within each qualitative focus group domain (eg, all *Collect* excerpts).

Table 4. Qualitative valence coding results by for therapeutic alliance (TA) domain from focus groups[a].

TA domain valence results	MBC[b] impacts TA	TA impacts MBC	Overall[c]
Values, mean (SD)	−0.23 (0.87)	0.13 (0.83)	−0.13 (0.86)
Excerpts (n=30), n (%)	22 (73)	8 (27)	30 (100)

[a]Valences are from −1 to 1, with −1 being a maximally negative valence and 1 being a maximally positive valence.

[b]MBC: measurement-based care.

[c]Overall means and SDs were calculated combining both categories of TA domain results (eg, MBC impacts TA and TA impacts MBC)

Providers had neutral to favorable attitudes toward their experiences within the *Act* stage (28 excerpts; mean valence 0.11, SD 0.96), with generally positive acceptability and appropriateness valences (mean 1.00, SD 0.00; and mean 0.25, SD 0.89, respectively). The feasibility valence was neutral to negative (mean −0.07, SD 1.00), while 2 (7%) of the 28 excerpts had a negative equity valence (mean −1.00, SD 0.00).

There were 30 excerpts related to TA (mean valence −0.13, SD 0.86), which revealed instances of MBC impacting TA (n=22, 73%) as well as of TA impacting MBC (n=8, 27%). When mental health care providers shared examples of MBC impacting TA in this safety-net setting, experiences were mixed, but overall resulted in a net negative valence (mean valence −0.23, SD 0.87). Experiences of TA impacting or influencing MBC were also mixed; yet, they resulted in an overall net positive valence (mean valence 0.13, SD 0.83).

Drivers of Valence Coding Results by Domain

Table 5 presents representative quotations; and the subsections that follow summarize the factors contributing to positive, negative, or neutral valence within each domain.

XSL•FO
RenderX

Table 5. Representative quotations from qualitative analysis of provider focus groups by domain and implementation outcome.

Domains and implementation outcomes	Representative quotations
Collect	
Acceptability	• "And the person essentially was like, I'm not filling out any more of those. So that was a drastic response." [Participant 2] • "He sent a message saying that he would not be filling this out because he had too many experiences of filling questionnaires out, and they're not used, and it's a waste of his time." [Participant 4]
Appropriateness	• "Probably because of the acuity of the patients that I'm meeting with, it doesn't feel as relevant of a conversation...[later]...it's not a top priority compared to the other things that I'm trying to accomplish, which is mostly stabilization." [Participant 2]
Feasibility	• "Patients don't know what it is. They, they get a lot of surveys, especially from primary care. Um. And yeah, so that's, there's a, a lot of that education that happens when, when a patient brings it up or if, if I bring [Computerized Adaptive Test for Mental Health] up." [Participant 5]
Equity	• "I've had some folks from non [...] US western cultural backgrounds who have gotten some strange scores on the perceived—the, um, how are they calling the psychosis module?...And I do think some of it is just culturally they, their answers are very normative, um, within their upbringing and, and not within kind of the, I think some of the bias in the, in the questions." [Participant 5] • "I have to agree with [Participant 3] that it feels like the goals are a little skewed. Um, you know, uh, helpful for us helping a health care system treat a whole population, um, may be important, but it's hard to figure out how you, how that translates individually patient to patient." [Participant 6]
Share	
Acceptability	• "I had not had on my radar any symptoms of psychosis for this person. And the measurement...came up as high risk or severe something. So I was able to ask more specific questions about that and actually determined that there was a level of paranoia and was more attuned to potential psychotic symptoms. If I had not seen the CAT-MH, I wouldn't have gone down that road of assessing that with them, and then a couple months later, the person actually expressed suicidal ideation and was hospitalized. So it was helpful to have the measurements to talk through more of the symptoms and also because there was kind of a lack of affect around what the person was reporting. The CAT-MH helped me recognize the possible severity of the symptoms." [Participant 2] • "I've also had a couple strong experiences that stick out in my mind of what I perceive as like, harm that has occurred, um, from taking the test. And those have mostly been in patients with PTSD and I'm, I'm curious if [Participant 7] and [Participant 5] have had those experiences too. But, um, one who was very distraught because her PTSD score was not high enough and so it invalidated her experience." [Participant 1]
Appropriateness	• "Um, and the, the test results of the CAT-MH, as far as I can tell, like, they, they're somewhat black-and-white as far as like what goes into the ADHD diagnostic category, um, versus, you know, attentional issues that are anxiety based or whatever. And I, and that I think maybe could reinforce like a more black-and-white understanding of where this could come from for the patient." [Participant 3]
Feasibility	• "I'm gonna put a plug in for, uh, as a psychiatrist, I...there's not enough time in your visits. I'm just putting it out there. I mean there's patients who report it, psychiatrists report it, there's just not enough time." [Participant 7]
Equity	• "I think patients who may have more therapy experience...who have more experience thinking about their emotional experiences, like, are probably more likely to um, be able to talk about it." [Participant 2]
Act	
Acceptability	• "There have been a couple of instances where people's scores have improved and it's been like an opportunity for, kind of, some celebration, which I think is nice. And it's nice to have this sort of like, this objective feel, you know, to, to their improvement, um, in a way that I think that's probably been the most meaningful. I, um, you know, that it's been for, for my treatment relationships." [Participant 3] • "I would say it's generally not changed treatment plans." [Participant 5] • "I feel we're pretty far still [from realizing the full potential benefits of measurement-based care], speaking from my experience and my team's experience." [Participant 6]

XSL•FO
RenderX

Domains and implementation outcomes	Representative quotations
Appropriateness	• "He was in remission for both [conditions] and then finally we, his, he did his CAT-MH and it, you know, the ADHD piece came up and it was something we had talked about when I first started working with him, but it was just so, the last priority, and, and that was a nice reminder of um, you know...it, it led to some good conversations and some book recommendations and it's, it's been a topic that we've, we've been kind of delving into." [Participant 5]
Feasibility	• "We started a medication. Maybe like three months later, whatever it was, those scores lessened and we're like, okay, let's continue this medication. And it felt very, like, simple and clean." [Participant 1]
Equity	• "The patient population that uses the tool more, um, and, and therefore that like has a little bit more economic stability. Um, computer literacy. Um, generally more often White, more often middle class, more often identified, more often with their diagnosis and their, their mental health care. Um, those are the folks who I see as, like, uh, uh, benefiting more and also putting in the work, you know, to, to engage with it more." [Participant 3]
TA[a]	
MBC[b] impacts TA	• "I thought it was interesting because sort of like, um, as a...a possible kind of point of contention between us...I didn't want to introduce, like, a stimulant for instance." [Participant 3] • "I think there's been this sort of pivot away from relationship-based care...in community psychiatry that's been this sort of, like, painful decades-long thing...and I think that I have sort of seen these...quantitative, computer-based assessment triage tools as like a piece of that shift." [Participant 3]
TA impacts MBC	• "I have felt very much like I had to side on, like I and I do side on their side, but I felt like I had to join with them to kind of, like, protect them against this CAT-MH thing that happened to them. Um, and try to separate myself from that test." [Participant 1] • "One of my highest risk people I think ever...scored just so beautifully low on everything...We're getting the wrong score on the CAT-MH and I had a conversation with her and she said, 'Oh, it's easy to figure these out...I don't wanna get hospitalized.'...We had a good conversation about it...she's chronically suicidal, so she's always gonna score really high on those measures and she's like, hospitalization doesn't help." [Participant 5]

[a]TA: therapeutic alliance.

[b]MBC: measurement-based care.

Factors Contributing to Positive, Negative, or Neutral Valence Within Each Domain

Collect

Acceptability

Excerpts with positive acceptability valence often referenced mental health care providers' positive buy-in for engaging in MBC. Examples of excerpts with negative acceptability valence included mental health care provider reports of patient difficulty completing the measure or strong negative reactions to the measure, providers feeling that nothing they said to patients improved patient engagement in measure completion, providers questioning the validity of the measure, and providers' low MBC buy-in.

Appropriateness

Examples of excerpts with positive appropriateness valence included mental health care providers feeling that administration frequency was suitable. Examples of excerpts with negative appropriateness valence included concerns about MBC's fit across the system (eg, mental health care providers feeling that encouraging patients to engage in MBC was not a good use of their limited time with patients with complex needs, especially for those in crisis).

Feasibility

Examples of excerpts with negative feasibility valence included mental health care providers reporting patients with higher symptom burden having trouble with the length of the MBC measure. They also include concerns that the centralized MBC approach (ie, a multidiagnostic MBC symptom tool selected by clinical leadership, administered centrally via the patient portal, and at the same frequency for all patients as determined by the leadership) created confusion for patients and providers because patients simultaneously received other types of electronic surveys from the health system.

Equity

The mental health care providers in this study reported a number of observed barriers to the equitable implementation of MBC data collection for their patient populations, including varying health literacy and English language literacy; lack of availability in languages other than English and Spanish; exclusion of patients not using the electronic patient portal; and measure completion challenges for patients working multiple jobs, experiencing life stressors, living with greater functional impairment, or less actively engaged in their mental health treatment. These providers shared salient examples of the CAT-MH, and MBC in general, not being a good fit for patients from non-Western cultural backgrounds; for example, one mental health care provider shared that a patient became very

XSL•FO
RenderX

upset after receiving scores indicating high psychosis-related symptom burden during a period of acute grief after the death of multiple family members. The provider felt that the score may have been influenced by factors related to how this patient experienced grief and loss in their culture, rather than being reflective of psychopathology; however, the patient was lost to follow-up, and the provider wondered whether it was because of the MBC results. Some mental health care providers theorized that MBC might help improve equity at the population health level but were not sure how equity "translate[d] individually patient to patient."

Share

Acceptability

Examples of excerpts with positive acceptability valence included instances in which discussing results together helped the mental health care provider identify a new clinical concern. One provider described being prompted by unexpectedly high CAT-MH scores to "ask more specific questions," subsequently realizing that the patient may have psychosis symptoms. The combination of the high severity scores and the patient's "lack of affect around what [they were] reporting" gave the provider the clues needed to explore a new diagnosis. Examples of excerpts with negative acceptability valence often included instances of scores not matching a patient's subjective experience. Some mental health care providers expressed concerns that patients were harmed when their MBC results showed lower severity scores than their subjective experience, which they perceived as invalidating. Mixed acceptability (neutral valence) excerpts were also common.

Appropriateness

Examples of excerpts with positive appropriateness included that some mental health care providers felt that patients' ability to view results via the patient portal before their visit was a good fit with their practice, allowing patients to formulate questions before the discussion about the scores. The providers in this study had mixed reports (neutral valence) about numeric scores, which they felt, for example, could sometimes be interpreted as too definitively diagnostic or pathologizing. They found it challenging to find the right moment to introduce a discussion of the results and felt conflicted about whether using the visit time to do so was valuable, given other urgent priorities.

Feasibility

Some mental health care providers with shorter appointments (ie, psychiatrists and psychiatric nurse practitioners) liked having only summary CAT-MH scores available because it increased the feasibility of reviewing data during visits. Other mental health care providers (mostly therapists) felt that not having individual item responses made it harder for them to interpret the scores with their patients. The providers in this study reported numerous examples where interpreting MBC data together was feasible, particularly within an established care relationship. Negative or mixed drivers of feasibility included factors complicating communication about the measure and results, including that (1) patients routinely received other electronic survey requests and were sometimes confused about which "survey" their mental health care providers were referencing, (2) mental health care providers did not always know whether patients had viewed their scores, (3) providers sometimes forgot to bring up new scores during visits even when they were available, and (4) providers and patients sometimes disagreed on the relevance of the results or the amount of time that should be allotted to discussing them during the visit.

Equity

Mental health care providers reported that it was harder to make time during visits to discuss CAT-MH results for patients experiencing more life stressors, medical complexity, or challenges arriving to appointments on time.

Act

Acceptability

Mental health care providers expressed mixed acceptability in the *Act* domain, and all reported that the implementation effort seemed far from realizing the full potential benefits of MBC. In the rare instances where providers did report using MBC data to help make treatment decisions, their experiences generally revealed positive acceptability; for example, a provider described the act of appraising patient progress via repeated measure administration as creating an opportunity to jointly celebrate patient progress that may otherwise have been missed. Most providers in the study reported that their treatment plans did not change based on the MBC results.

Appropriateness

Appraising the data together to contextualize and prioritize treatment goals was seen as an appropriate application of MBC. One mental health care provider described how a patient felt validated when the CAT-MH flagged their longtime concerns about possible attention-deficit/hyperactivity disorder and how the patient and provider were able to discuss the results and agree to focus first on more urgent clinical needs and revisit the symptoms of inattention once these were under better control.

Feasibility

The most positive feasibility examples in the *Act* domain included instances in which mental health care providers could use the CAT-MH results to make decisions around adjusting medications, for example, when the CAT-MH showed symptom improvement after starting a medication and facilitated a decision to continue the selected medication. Others reported more mixed experiences (neutral scores), for example, when the results did not align with patient reports, which made it more challenging to decide whether any adjustments to treatment were needed.

Equity

Mental health care providers who reported concerns about equity in the *Act* domain shared that while they could see the potential benefit of MBC for equity at the system level (eg, by guiding the investment of scarce resources into programs that served identified patient needs), they did not see evidence that MBC was reducing health inequities for their individual patients. In large part, this was due to the fact that engaging with MBC was a prerequisite to benefiting from it, and the people most likely to engage with the CAT-MH seemed to be those from more

XSL•FO
RenderX

privileged backgrounds and thus in less need of care optimization.

TA Analysis

MBC Impacting TA

Our analysis indicated a bidirectional relationship between MBC and TA that was both positive and negative. MBC positively influenced TA when MBC strengthened the bond between patient and clinician (eg, by highlighting when to celebrate patient progress) and when MBC helped patients and their mental health care providers jointly prioritize certain mental health concerns over others in the short and long term (eg, agreeing to prioritize other mental health conditions before possible attention-deficit/hyperactivity disorder). On the flip side, MBC negatively influenced TA when mental health care providers saw quantitative assessments as a "pivot" away from relationship-based care; when it was difficult to reach consensus between provider and patient about the accuracy, meaning, or implications of the MBC results; and when patients had very strong, negative reactions to the process of measure completion or the results themselves, with their frustration sometimes directed toward providers. Some mental health care providers reported having to "talk down" patients and highlighted instances in which they had to make significant efforts to preserve or repair TA after a rupture occurred due to invalidating results. Some clinicians also noted that taking the time to encourage CAT-MH completion or discuss the results came at the cost of time for other therapeutic and relational aspects of care.

TA Impacting MBC

Mental health care providers shared examples in which strong TA helped mitigate potential negative patient experiences with MBC, although whether this translated into increased or decreased uptake of MBC varied by case. Where mental health care providers had strong TA with patients, they reported being able to navigate the *Share* and *Act* stages, even when the results seemed inaccurate; for example, a provider for a patient considered high risk recognized that the measure scores were inappropriately low and discussed with the patient their reasons for underreporting symptoms (namely the patient's wish to avoid hospitalization). In this case, strong TA helped the clinician overcome MBC's limitations.

Participant Suggestions for Adaptations to MBC Implementation (No Valence Applied)

Textbox 1 presents mental health care providers' suggestions for improvement (excerpts without valence). When mental health care providers in our focus groups suggested MBC adaptations that were entirely speculative, they received appropriate level codes and an "N/A" valence. In the *Collect* stage, mental health care providers suggested that patients might complete the CAT-MH at higher rates if they could choose a subset of questions in their assigned modules (this suggestion was not implementable within the current workflow). Relevant to *Collect* and TA, mental health care providers also suggested that MBC might build trust more effectively if introduced specifically when diagnostic clarity was needed, rather than routine symptom progress monitoring only. Mental health care providers hypothesized that the *Share* stage of MBC might be more acceptable if scores were not shared with the patient until the visit time, allowing providers to facilitate interpretation and discussion. They also suggested that ensuring equity at the *Share* stage might require standardizing when the results are discussed in patient encounters (eg, at the start of every visit). Most providers in the study stated that they were not changing existing treatment plans in response to MBC data, in accordance CSA's *Act* stage. Mental health care providers felt that they would feel more confident deciding when to change treatment plans based on MBC data if they were given with decision support tools, such as shared decision-making scripts incorporating the MBC results; concrete options for changing treatment plans within existing system and patient constraints; data about how their treatment decisions impacted patient outcomes; and MBC target goals for specific populations where it was likely to have the biggest impact.

Textbox 1. Participants' suggestions for improvement (excerpts without valence).

- "I would love to know, like, is my prescrib[ing]...am I making the right decisions? And I'm sure that analytics could be used in some, you know, in some way to inform me about that." [Participant 3]
- "I, I, you know, I wonder if it would build trust if it was something I introduced at a moment where there was a need for diagnostic clarification...There are times where patients are asking me and they're wanting to know like, what do you think? And could this be this? Or, you know, we're trying to clarify the diagnosis together. And I'll say...It could help us if you filled that out and then we could talk about it afterwards." [Participant 6]
- "Instead of having the whole range of different diagnoses, what if we picked one of the ones that we have—you know—our patients who are suffering with the most, they're, they're...that we're not seeming to really affect much change. And what if we...honed in on just one piece and then, you know, would that make it easier for us to use...easier for the, those patients who need it to use, and for us to see different results?" [Participant 6]
- "Um, so I think if we screen for it, we need to have tools and things to help them when they screen positive for those things...And so I'm feeling, this is separate from CAT-MH, but also I guess related to measurement-based care in some ways maybe. So wanting to have the resources in place if we're gonna ask the questions so it's not further traumatizing." [Participant 1]

XSL•FO
RenderX

Discussion

Principal Findings

MBC literature has highlighted its potential and observed benefits, but there has been little documentation of real-world MBC implementations in underresourced health systems serving patients from diverse backgrounds and clinical profiles. This pragmatic mixed methods study adds to the literature by inviting feedback from a group of mental health care providers with >1 year of MBC experience who work with a heterogeneous group of patients with high needs in a safety-net setting.

By beginning with sequential surveys over 6 months and ending with focus groups that were deeply informed by the patterns that emerged from the surveys, the mixed methods used in this study allowed the data collection and interpretation to yield a gestalt understanding of MBC experiences of mental health care providers that could not have been obtained from quantitative or qualitative approaches alone. While the quantitative surveys were helpful in understanding the ongoing barriers and high-level perceptions of mental health care providers in the system, they could not answer why some markers of implementation outcomes did not improve—and in some cases worsened—as providers gained experience with MBC. Similarly, the strength and persistence of the converging survey responses pertaining to health equity influenced our decision to ask about health equity concerns at each CSA stage, which in turn influenced our ability to present a richer view of health equity in MBC than may have been possible in prior literature. Using the qualitative focus groups to explore the findings about TA from the quantitative data helped us uncover a richer understanding of the bidirectional relationship between TA and MBC, which we had not seen discussed in prior literature.

One key finding from our study was that the CSA model, which was originally developed with individual patient-provider dyads in mind, has limitations when applied at the health system level. Unlike individual practitioners, health systems tend to rely on centralized, coordinated processes to implement MBC. In our setting, department leadership decided on a single measure that would be sent to all patients, and administrative support staff sent the measure via the patient portal at a set time before upcoming visits. While these decisions were intended to streamline MBC data collection and hence improve feasibility for mental health care providers, they ultimately precluded some of the original CSA model's recommended *Collect* substeps from being performed; for example, providers could not discuss why they chose the MBC measure for the specific patient or involve the patient in selecting measures relevant to their treatment plan, both of which are identified in the CSA model as ideal practices. Instead, they were limited to introducing MBC as an innovative approach to care and explaining why the specific measure (ie, the CAT-MH) had been selected by the organization.

Our study also highlights the ways in which a centralized approach to MBC implementation may have mixed impacts on its acceptability. First, some of the mental health care providers in our focus groups identified moral injury associated with collecting data that quantified their patients' distress when they felt that they did not have clear guidance on how to adjust treatment in response, especially given their patients' very real resource constraints.

Second, to comply with the 21st Century Cures Act [42], our organization's MBC workflow shared results directly and instantaneously with patients via the patient portal. While prior studies indicate that patients generally prefer immediate release of most test results [43], our focus groups revealed that some patients had strong negative reactions to the MBC results they had viewed on their own—reactions that their mental health care providers then had to mitigate during appointment times. One possible solution to this problem would be to administer MBC measures immediately before a session, with the patient in the office waiting room. This approach would comply with the 21st Century Cures Act while also giving clinicians the opportunity to guide the initial conversation about the measure results.

Finally, because our system chose a single MBC measure that drew from an extremely large question bank, individual item responses were not available in the electronic health record. While implementation leaders chose this design to facilitate IT feasibility [44], clinicians' inability to review which items their patients endorsed likely impacted the ways in which they could be expected to engage with and respond to the CAT-MH results. Future studies should assess whether mental health care providers in similar settings who are presented with data from simpler MBC measures have more success with discussing and acting on the data in collaboration with their patients.

Drawing from these observations, our study indicates that researchers seeking to use the CSA model to guide MBC implementation in mental health systems may need to add components to the model that capture where clinic- or organizational-level factors can better support patients and care providers at each stage. Our work suggests that efforts to centralize the process in the service of feasibility may unintentionally undermine MBC acceptability for patients and mental health care providers alike. Anticipating where centralization may add friction and mitigating these effects are likely necessary to facilitate successful MBC implementation in contexts such as ours.

Another key finding from our study was that MBC may exacerbate health disparities, which contrasts with the foundational premise that MBC has the potential to enhance mental health equity [4]. Our study participants observed that MBC was being implemented unevenly across patient subgroups (as reflected in the survey data), which seemed to be driven largely by barriers to the *Collect* stage (as reflected in the focus group data). The stated drivers of unequal implementation echo those observed in patient-reported outcomes literature, namely language barriers; functional impairment driven by disease burden; differences in health and technology literacy and access; and differences by patient age, race and ethnicity, or gender [23,45,46]. In other words, because completing the CAT-MH was a necessary step before the data could be used to adjust treatment, the patients who could not or would not complete it—who tended to be marginalized across ≥1 domains—were unable to benefit from the treatment approach.

XSL•FO
RenderX

These observations are critical, given the value that implementation leaders in diverse safety-net settings place on offering equitable, culturally relevant mental health care [47-49]. Our findings, which align with prior research that emphasizes the importance of developing measures and treatment approaches that are culturally appropriate [21,50], suggest that a direct focus on practical implementation and equity may be needed in future MBC research. However, acting on these findings may be increasingly difficult as health systems that are subject to value-based payments face pressure to develop centralized MBC processes that limit patient-mental health care provider dyads from finding the most appropriate measure or choosing to abstain when the selected measure is not a good fit. These areas warrant further investigation.

Finally, while prior evidence suggests that MBC can enhance TA, our findings suggest that the relationship between MBC and TA is more complex and not wholly positive. Strong baseline TA seemed to enhance mental health care providers' ability to proceed through the CSA stages, including collecting and assessing the accuracy of the data, working with patients to resolve potential discrepancies between the data and the patient's narrative report, making sense of the data, and deciding how to proceed. However, when MBC was perceived as a threat to the TA, the clinicians in our study prioritized preserving or repairing their relationship with the patient over collecting MBC data and applying the results to treatment. In addition, mental health care providers were more enthusiastic about MBC when the data confirmed their preexisting diagnostic formulations. Conversely, the providers we spoke to tended to trust their clinical judgment over MBC data when the two were discordant. Furthermore, when the data differed from the patient's own explanatory model for their illness, mental health care providers perceived MBC as burdensome, sometimes leading to short- or long-term complications for TA. Bearing this complex relationship in mind, future implementation efforts may benefit from directly addressing the tension MBC can cause for TA, and future implementation research can assist these efforts by studying specific strategies to address this tension. Therapeutic ruptures may strengthen TA in the long run [51], but time-constrained safety-net mental health care providers would benefit from training on how to prepare for the ruptures that MBC may cause as well as institutional support to ensure time and opportunity for repair.

Strengths, Limitations, and Future Directions

This small-scale hypothesis-generating study had many strengths. First, we identified challenges and potential adaptations at each CSA stage in a diverse safety-net setting. Second, we intentionally included equity as a level 2 valence code, a critical implementation outcome that is often undercharacterized in implementation research. Third, we recruited mental health care providers with at least 1 year of real-world safety-net MBC experience and studied their responses longitudinally via sequential mixed methods to intentionally prioritize "informational richness" [39]. Specifically, we used purposeful sampling of "information-rich" cases [39], selecting mental health care providers from a health system that had implemented MBC 1 year prior and devoted significant resources to ongoing training and quality improvement activities in an effort to improve MBC uptake. Therefore, all participants in the study had experienced the same training and implementation activities as a group over this time period (emphasis on sampling for highly overlapping context, experience, and role in each case). In this sense, informational richness was prioritized a priori [39] with both the purposive sampling method and the use of the repeated surveys before the focus groups. In other words, the repeated quantitative data were used to hone the focus group questions over time to ensure maximum relevance for the study participants.

Although our focus group sample size was small (n=7), it was sufficient for phenomenological studies of lived experience, where 3 to 6 participants are generally considered adequate [39]. As the mental health care providers in our study not only participated in MBC trainings and gained experience with MBC through their 1-on-1 patient care but also engaged in ongoing team meetings and grand rounds related to MBC over this time period, any provider who participated in the study was aware of not just their own experience of MBC implementation but also that of their colleagues. This type of broad understanding of the phenomena emerged during focus groups, where open dialogue was encouraged. By balancing pragmatic data collection that minimized participant burden with a focus on information richness from mental health care providers who had a great deal of experience with the specific phenomena in an urban safety-net mental health clinic, this study adds significant value to directions for future research.

This study also had several limitations. Although we did not intend to test hypotheses, the relatively small study sample may not have been representative of the full group of mental health care providers in our department. Providers who participated were more likely to be female (5/6, 83%) compared to those who did not participate (13/19, 68%). In addition, among participants, 50% (3/6) were clinical psychologists or psychiatrists or psychiatric nurse practitioners and 50% (3/6) were social workers, compared to 42% (8/19) and 58% (11/19), respectively, among those who did not participate. While we did not gather feedback about reasons for nonparticipation from each eligible care providers, clinician burnout was noted to be prevalent at the time of the study. Given this context, it is possible that the clinicians who did not participate had lower morale, which could mean that the feedback we gathered was more positive than it would have been with broader representation. Of note, our response rate of 27% (7/26) is in line with prior research about specialist participation in web-based surveys [52]. Nevertheless, scaling up similar study designs would likely require efforts to reduce barriers to clinician participation even further. With greater levels of participation, future studies could assess whether the care settings or patient populations with whom different mental health care providers work correlated in any way to their response to MBC implementation.

Among the clinicians who did participate, there was considerably more feedback on the *Collect* stage than the *Share* or *Act* stages of MBC. This was likely because relatively few patients completed >1 CAT-MH. To the extent that we were able to gather feedback on the *Share* stage, some of the participants in our study identified the fact that individual item

XSL•FO
RenderX

responses were not visible in the electronic health record as a limitation in the implementation design, although for others this was noted to increase feasibility.

Another potential limitation is that the study leads, AMP and LCR, worked in the same organization as the participants. In theory, this may have suppressed people's willingness to participate in the study or share candid feedback about their experience with the measure. However, these effects were likely mitigated by the fact that the researchers interacting with the potential and enrolled participants had no supervisory relationship with them. Moreover, the recruitment and informed consent processes were designed to highlight the steps being taken to ensure that participant feedback remained anonymous to anyone in leadership or with a supervisory relationship with participants.

Finally, the study context may have impacted findings in a way that is important to consider. The health system where this study was conducted attracts mental health care providers who have a psychodynamic orientation, which may not lend itself as easily to MBC as other more structured treatment modalities. Health care providers in our system are also drawn to working with diverse patients as well as those considered disadvantaged, who may require additional coaching to overcome cultural and educational barriers to filling out written screeners as part of care. Furthermore, some of the challenges we identified may have been specific to the chosen MBC measure and implementation design: a computer-adaptive tool that was sent out through a centralized, system-wide messaging platform and administered remotely on patients' own electronic devices. As such, the directions for future research that we identified may be more relevant for care settings and implementation designs that share some of these characteristics.

Conclusions

Our study found that implementing MBC in safety-net health systems poses significant challenges that warrant further study and likely need to be addressed to ensure that all patients benefit equally from this evidence-based treatment approach. Our qualitative analysis highlighted unanticipated challenges at each CSA stage, for which health systems have limited guidance at this time. These were particularly evident in the *Collect* stage, although the mental health care providers in our study had less experience with the *Share* and *Act* stages, limiting the depth of feedback in these areas.

The rich data from this mixed methods analysis also highlighted the negative as well as positive ways in which MBC can impact TA. Our results point to a need for more research and clearer applied strategies about how to prioritize TA amid MBC implementation so that uptake is enhanced, and unintended psychological harm to patients is avoided.

Finally, our findings revealed bidirectional concerns regarding equity, namely that MBC may be inequitably implemented in safety-net settings and may also exacerbate rather than reduce health disparities among patient subgroups.

These findings, which are meant to uncover important areas for further attention and study, are timely in the current policy environment. Interest in psychiatric MBC is growing, and reimbursement is increasingly being tied to collecting and demonstrating improvements in patient-reported outcomes. Future research should explore which tools and how much customizability will improve MBC adoption in safety-net systems as well as which components of each CSA stage can be centralized to improve MBC feasibility without compromising equity or TA. Moving forward, it will be essential to address existing knowledge gaps about how to implement MBC equitably to avoid disproportionately impacting both the financial and ethical imperatives of safety-net health systems.

Acknowledgments

This study received funding support from the Cambridge Health Alliance Department of Psychiatry.

DW was affiliated with Cambridge Health Alliance at the time of the study and is now affiliated with Ragon Institute of Mass General Brigham, Massachusetts Institute of Technology, and Harvard University in Cambridge, MA.

HSL was affiliated with Cambridge Health Alliance and Harvard Medical School at the time of the study, and he is now retired.

PW was affiliated with Cambridge Health Alliance at the time of the study and is now affiliated with Brigham and Women's Hospital and Harvard T H Chan School of Public Health, which are both in Boston, MA. He is also still affiliated with Harvard Medical School.

Data Availability

The focus group guide is available from the corresponding author on reasonable request. The datasets generated or analyzed during this study are not publicly available, in accordance with the participant research protections at our institution. Limited, deidentified data summaries or tools may be available from the corresponding author on reasonable request and subject to institutional policies on research protections.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
Supplementary Figure S1 (Explanatory Sequential Study Design of Provider Experiences of MBC), Supplementary Figure S2 (Example of Individual Item Analysis Used to Generate Heatmap of Quantitative Survey Results), Supplementary Table S1 (Applying Patterns from Quantitative Findings to Narrow and Refine Focus Group Guide Approach), and Supplementary Table S2 (Qualitative Coding Tree Used for Qualitative Coding of Focus Group Data).
[DOCX File , 348 KB - formative_v9i1e71570_app1.docx]

References

1. Scott K, Lewis CC. Using measurement-based care to enhance any treatment. Cogn Behav Pract 2015 Feb;22(1):49-59 [FREE Full text] [doi: 10.1016/j.cbpra.2014.01.010] [Medline: 27330267]
2. Dowrick C, Leydon GM, McBride A, Howe A, Burgess H, Clarke P, et al. Patients' and doctors' views on depression severity questionnaires incentivised in UK quality and outcomes framework: qualitative study. BMJ 2009 Mar 19;338:b663. [doi: 10.1136/bmj.b663] [Medline: 19299474]
3. Boswell JF, Hepner KA, Lysell K, Rothrock NE, Bott N, Childs AW, et al. The need for a measurement-based care professional practice guideline. Psychotherapy (Chic) 2023 Mar;60(1):1-16. [doi: 10.1037/pst0000439] [Medline: 35771518]
4. Fortney JC, Unützer J, Wrenn G, Pyne JM, Smith GR, Schoenbaum M, et al. A tipping point for measurement-based care. Psychiatr Serv 2017 Feb 01;68(2):179-188. [doi: 10.1176/appi.ps.201500439] [Medline: 27582237]
5. Guo T, Xiang YT, Xiao L, Hu CQ, Chiu HF, Ungvari GS, et al. Measurement-based care versus standard care for major depression: a randomized controlled trial with blind raters. Am J Psychiatry 2015 Oct;172(10):1004-1013. [doi: 10.1176/appi.ajp.2015.14050652] [Medline: 26315978]
6. Lambert MJ, Whipple JL, Kleinstäuber M. Collecting and delivering progress feedback: a meta-analysis of routine outcome monitoring. Psychotherapy (Chic) 2018 Dec;55(4):520-537. [doi: 10.1037/pst0000167] [Medline: 30335463]
7. Tepper MC, Ward MC, Aldis R, Lanca M, Wang PS, Fulwiler CE. Toward population health: using a learning behavioral health system and measurement-based care to improve access, care, outcomes, and disparities. Community Ment Health J 2022 Nov;58(8):1428-1436 [FREE Full text] [doi: 10.1007/s10597-022-00957-3] [Medline: 35352203]
8. Cook BL, Hou SS, Lee-Tauler SY, Progovac AM, Samson F, Sanchez MJ. A review of mental health and mental health care disparities research: 2011-2014. Med Care Res Rev 2019 Dec 07;76(6):683-710. [doi: 10.1177/1077558718780592] [Medline: 29877136]
9. Merino Y, Adams L, Hall WJ. Implicit bias and mental health professionals: priorities and directions for research. Psychiatr Serv 2018 Jun 01;69(6):723-725. [doi: 10.1176/appi.ps.201700294] [Medline: 29493411]
10. Stubbe DE. The therapeutic alliance: the fundamental element of psychotherapy. Focus (Am Psychiatr Publ) 2018 Oct;16(4):402-403 [FREE Full text] [doi: 10.1176/appi.focus.20180022] [Medline: 31975934]
11. Wampold BE, Flückiger C. The alliance in mental health care: conceptualization, evidence and clinical applications. World Psychiatry 2023 Feb 14;22(1):25-41 [FREE Full text] [doi: 10.1002/wps.21035] [Medline: 36640398]
12. Ardito RB, Rabellino D. Therapeutic alliance and outcome of psychotherapy: historical excursus, measurements, and prospects for research. Front Psychol 2011;2:270 [FREE Full text] [doi: 10.3389/fpsyg.2011.00270] [Medline: 22028698]
13. Kilbourne AM, Beck K, Spaeth-Rublee B, Ramanuj P, O'Brien RW, Tomoyasu N, et al. Measuring and improving the quality of mental health care: a global perspective. World Psychiatry 2018 Feb;17(1):30-38 [FREE Full text] [doi: 10.1002/wps.20482] [Medline: 29352529]
14. Lewis CC, Boyd M, Puspitasari A, Navarro E, Howard J, Kassab H, et al. Implementing measurement-based care in behavioral health: a review. JAMA Psychiatry 2019 Mar 01;76(3):324-335 [FREE Full text] [doi: 10.1001/jamapsychiatry.2018.3329] [Medline: 30566197]
15. Resnick SG, Hoff RA. Observations from the national implementation of Measurement Based Care in Mental Health in the Department of Veterans Affairs. Psychol Serv 2020 Aug;17(3):238-246. [doi: 10.1037/ser0000351] [Medline: 31058519]
16. Hatfield DR, Ogles BM. Why some clinicians use outcome measures and others do not. Adm Policy Ment Health 2007 May;34(3):283-291. [doi: 10.1007/s10488-006-0110-y] [Medline: 17211715]
17. Jensen-Doss A, Haimes EM, Smith AM, Lyon AR, Lewis CC, Stanick CF, et al. Monitoring treatment progress and providing feedback is viewed favorably but rarely used in practice. Adm Policy Ment Health 2018 Jan;45(1):48-61 [FREE Full text] [doi: 10.1007/s10488-016-0763-0] [Medline: 27631610]
18. Zimmerman M, McGlinchey JB. Why don't psychiatrists use scales to measure outcome when treating depressed patients? J Clin Psychiatry 2008 Dec;69(12):1916-1919. [doi: 10.4088/jcp.v69n1209] [Medline: 19192467]
19. Barber J, Resnick SG. Collect, share, act: a transtheoretical clinical model for doing measurement-based care in mental health treatment. Psychol Serv 2023;20(Suppl 2):150-157. [doi: 10.1037/ser0000629] [Medline: 35201811]
20. DeSimone J, Hansen BR. The impact of measurement-based care in psychiatry: an integrative review. J Am Psychiatr Nurses Assoc 2024 Apr;30(2):279-287. [doi: 10.1177/10783903231177707] [Medline: 37264644]
21. Connors EH, Arora PG, Resnick SG, McKay M. A modified measurement-based care approach to improve mental health treatment engagement among racial and ethnic minoritized youth. Psychol Serv 2023;20(Suppl 1):170-184 [FREE Full text] [doi: 10.1037/ser0000617] [Medline: 35130008]

XSL•FO
RenderX

22. Barber J, Childs AW, Resnick S, Connors EH. Leveraging measurement-based care to reduce mental health treatment disparities for populations of color. Adm Policy Ment Health 2025 Jan 15;52(1):172-178. [doi: 10.1007/s10488-024-01364-4] [Medline: 38489017]
23. Swanson LM, Patel PD, Montalva R, Bullard KH, Parikh SV, Dalack GW. Enhanced adoption of measurement-based care in a psychiatry outpatient clinic after high-reliability process changes. Psychiatr Serv 2023 Apr 01;74(4):423-426. [doi: 10.1176/appi.ps.202200032] [Medline: 36164773]
24. Liu FF, Cruz RA, Rockhill CM, Lyon AR. Mind the gap: considering disparities in implementing measurement-based care. J Am Acad Child Adolesc Psychiatry 2019 Apr;58(4):459-461. [doi: 10.1016/j.jaac.2018.11.015] [Medline: 30926072]
25. Flückiger C, Del Re AC, Wampold BE, Horvath AO. The alliance in adult psychotherapy: a meta-analytic synthesis. Psychotherapy (Chic) 2018 Dec;55(4):316-340. [doi: 10.1037/pst0000172] [Medline: 29792475]
26. Brattland H, Koksvik JM, Burkeland O, Klöckner CA, Lara-Cabrera ML, Miller SD, et al. Does the working alliance mediate the effect of routine outcome monitoring (ROM) and alliance feedback on psychotherapy outcomes? A secondary analysis from a randomized clinical trial. J Couns Psychol 2019 Mar;66(2):234-246. [doi: 10.1037/cou0000320] [Medline: 30702322]
27. de Jong K, Conijn JM, Gallagher RA, Reshetnikova AS, Heij M, Lutz MC. Using progress feedback to improve outcomes and reduce drop-out, treatment duration, and deterioration: a multilevel meta-analysis. Clin Psychol Rev 2021 Apr;85:102002. [doi: 10.1016/j.cpr.2021.102002] [Medline: 33721605]
28. Carlier IV, Meuldijk D, Van Vliet IM, Van Fenema E, Van der Wee NJ, Zitman FG. Routine outcome monitoring and feedback on physical or mental health status: evidence and theory. J Eval Clin Pract 2012 Feb;18(1):104-110. [doi: 10.1111/j.1365-2753.2010.01543.x] [Medline: 20846319]
29. Proctor E, Silmere H, Raghavan R, Hovmand P, Aarons G, Bunger A, et al. Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda. Adm Policy Ment Health 2011 Mar;38(2):65-76 [FREE Full text] [doi: 10.1007/s10488-010-0319-7] [Medline: 20957426]
30. Creswell JW. A Concise Introduction to Mixed Methods Research. Thousand Oaks, CA: Sage Publication; 2015.
31. Aldis R, Rosenfeld LC, Mulvaney-Day N, Lanca M, Zona K, Lam JA, et al. Determinants of remote measurement-based care uptake in a safety net outpatient psychiatry department as part of learning health system transition. Learn Health Syst 2024 Jun;8(Suppl 1):e10416 [FREE Full text] [doi: 10.1002/lrh2.10416] [Medline: 38883875]
32. Gibbons RD, Alegría M, Cai L, Herrera L, Markle SL, Collazos F, et al. Successful validation of the CAT-MH Scales in a sample of Latin American migrants in the United States and Spain. Psychol Assess 2018 Oct;30(10):1267-1276 [FREE Full text] [doi: 10.1037/pas0000569] [Medline: 29792502]
33. Graham AK, Minc A, Staab E, Beiser DG, Gibbons RD, Laiteerapong N. Validation of the computerized adaptive test for mental health in primary care. Ann Fam Med 2019 Jan;17(1):23-30 [FREE Full text] [doi: 10.1370/afm.2316] [Medline: 30670391]
34. Gibbons RD, Kupfer DJ, Frank E, Lahey BB, George-Milford BA, Biernesser CL, et al. Computerized adaptive tests for rapid and accurate assessment of psychopathology dimensions in youth. J Am Acad Child Adolesc Psychiatry 2020 Nov;59(11):1264-1273 [FREE Full text] [doi: 10.1016/j.jaac.2019.08.009] [Medline: 31465832]
35. Gibbons RD, Weiss DJ, Frank E, Kupfer D. Computerized adaptive diagnosis and testing of mental health disorders. Annu Rev Clin Psychol 2016;12:83-104. [doi: 10.1146/annurev-clinpsy-021815-093634] [Medline: 26651865]
36. Weiner BJ, Lewis CC, Stanick C, Powell BJ, Dorsey CN, Clary AS, et al. Psychometric assessment of three newly developed implementation outcome measures. Implement Sci 2017 Aug 29;12(1):108 [FREE Full text] [doi: 10.1186/s13012-017-0635-3] [Medline: 28851459]
37. Horvath AO, Greenberg LS. Development and validation of the working alliance inventory. J Couns Psychol 1989 Apr;36(2):223-233. [doi: 10.1037/0022-0167.36.2.223]
38. Tichenor V, Hill CE. A comparison of six measures of working alliance. Psychother Theory Res Pract Train 1989;26(2):195-199. [doi: 10.1037/h0085419]
39. Palinkas LA, Horwitz SM, Green CA, Wisdom JP, Duan N, Hoagwood K. Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. Adm Policy Ment Health 2015 Sep;42(5):533-544 [FREE Full text] [doi: 10.1007/s10488-013-0528-y] [Medline: 24193818]
40. Damschroder LJ, Lowery JC. Evaluation of a large-scale weight management program using the consolidated framework for implementation research (CFIR). Implement Sci 2013 May 10;8:51 [FREE Full text] [doi: 10.1186/1748-5908-8-51] [Medline: 23663819]
41. Ahmed SK. The pillars of trustworthiness in qualitative research. J Med Surg Public Health 2024 Apr;2:100051. [doi: 10.1016/j.glmedi.2024.100051]
42. Dworkowitz A. Provider obligations for patient portals under the 21st Century Cures Act. Health Affairs. 2022 May 16. URL: https://www.healthaffairs.org/content/forefront/provider-obligations-patient-portals-under-21st-century-cures-act [accessed 2025-01-10]
43. Steitz BD, Turer RW, Lin CT, MacDonald S, Salmi L, Wright A, et al. Perspectives of patients about immediate access to test results through an online patient portal. JAMA Netw Open 2023 Mar 01;6(3):e233572 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.3572] [Medline: 36939703]

XSL•FO
RenderX

44. Rosenfeld LC, Tepper MC, Leff SH, Wang D, Zhang A, Tian L, et al. Pre-implementation patient, provider, and administrator perspectives of remote measurement-based care in a safety net outpatient psychiatry department. Learn Health Syst 2025 Apr 23;9(2):e10472 [FREE Full text] [doi: 10.1002/lrh2.10472] [Medline: 40247894]
45. Nguyen H, Butow P, Dhillon H, Sundaresan P. A review of the barriers to using Patient‐Reported Outcomes (PROs) and Patient‐Reported Outcome Measures (PROMs) in routine cancer care. J Med Radiat Sci 2020 Aug 19;68(2):186-195 [FREE Full text] [doi: 10.1002/jmrs.421]
46. Long C, Beres LK, Wu AW, Giladi AM. Patient-level barriers and facilitators to completion of patient-reported outcomes measures. Qual Life Res 2022 Jun;31(6):1711-1718. [doi: 10.1007/s11136-021-02999-8] [Medline: 34533759]
47. Office of the Surgeon General (US), Center for Mental Health Services (US), National Institute of Mental Health (US). Mental Health: Culture, Race, and Ethnicity: A Supplement to Mental Health: A Report of the Surgeon General. Rockville, MD: Substance Abuse and Mental Health Services Administration (US); 2001.
48. Cook BL, Zuvekas SH, Carson N, Wayne GF, Vesper A, McGuire TG. Assessing racial/ethnic disparities in treatment across episodes of mental health care. Health Serv Res 2014 Feb;49(1):206-229 [FREE Full text] [doi: 10.1111/1475-6773.12095] [Medline: 23855750]
49. Interian A, Lewis-Fernández R, Dixon LB. Improving treatment engagement of underserved U.S. racial-ethnic groups: a review of recent interventions. Psychiatr Serv 2013 Mar 01;64(3):212-222. [doi: 10.1176/appi.ps.201100136] [Medline: 23203442]
50. Adams LB, Baxter SL, Lightfoot AF, Gottfredson N, Golin C, Jackson LC, et al. Refining Black men's depression measurement using participatory approaches: a concept mapping study. BMC Public Health 2021 Jun 22;21(1):1194 [FREE Full text] [doi: 10.1186/s12889-021-11137-5] [Medline: 34158011]
51. Safran JD, Muran JC, Eubanks-Carter C. Repairing alliance ruptures. Psychotherapy (Chic) 2011 Mar;48(1):80-87. [doi: 10.1037/a0022140] [Medline: 21401278]
52. Cunningham CT, Quan H, Hemmelgarn B, Noseworthy T, Beck CA, Dixon E, et al. Exploring physician specialist response rates to web-based surveys. BMC Med Res Methodol 2015 Apr 09;15:32 [FREE Full text] [doi: 10.1186/s12874-015-0016-z] [Medline: 25888346]

Abbreviations

CAT-MH: Computerized Adaptive Test for Mental Health
CSA: Collect-Share-Act
MBC: measurement-based care
TA: therapeutic alliance

Edited by A Mavragani; submitted 21.01.25; peer-reviewed by A Secor, J Rash, CY Chen; comments to author 02.03.25; revised version received 15.05.25; accepted 06.06.25; published 05.09.25.

Please cite as:
Progovac AM, Fulwiler C, Lanca M, Wang D, Zona K, Mulvaney-Day N, Aldis R, Leff HS, Wang P, Rosenfeld LC
Mental Health Care Provider Experiences of Remote Measurement-Based Care Rollout in an Urban Safety-Net Psychiatry Department: Three-Site Mixed Methods Hypothesis-Generating Implementation Study
JMIR Form Res 2025;9:e71570
URL: https://formative.jmir.org/2025/1/e71570
doi:10.2196/71570
PMID:

©Ana M Progovac, Carl Fulwiler, Margaret Lanca, Daisy Wang, Kate Zona, Norah Mulvaney-Day, Rajendra Aldis, H Stephen Leff, Philip Wang, Lisa C Rosenfeld. Originally published in JMIR Formative Research (https://formative.jmir.org), 05.09.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Quantifying Spatial Shadow Zones and Their Association With Hospital Falls in Acute Care Unit: Real-Time Location System Observational Study

Yen-Pin Chen[1,2], MD, PhD; Yi-Chun Chen[3], PhD; Chen-Liang Lin[4], PhD; Chien-Yu Chi[2], MD; Yi-Ying Chen[2], MD; Bey-Jing Yang[2], RN, MSN; Chien-Hua Huang[2], MD, PhD

[1]Department of Emergency Medicine, National Taiwan University Hospital Hsin-Chu Branch, Hsinchu, Taiwan
[2]Department of Emergency Medicine, College of Medicine, National Taiwan University Hospital, Taipei, Taiwan
[3]Guidercare Limited, Taipei, Taiwan
[4]Biomedical Technology and Device Research Labs, Industrial Technology Research Institute, Hsinchu, Taiwan

Corresponding Author:
Chien-Hua Huang, MD, PhD
Department of Emergency Medicine
College of Medicine
National Taiwan University Hospital
No.7, Zhongshan South Road
Zhongzheng District
Taipei, 100
Taiwan
Phone: 886 23123456
Email: chhuang5940@ntu.edu.tw

Abstract

Background: Hospital falls represent a persistent and significant threat to safety within health care systems worldwide, impacting both patient well-being and the occupational health of health care staff. While patient falls are a primary concern, addressing fall risks for all individuals within the health care environment remains a key objective. Caregiver visibility and spatial monitoring are recognized as crucial considerations in mitigating fall-related incidents.

Objective: This study aimed to investigate the association between the percentage of spatial shadow zone, defined as areas within an acute care unit unvisited by mobile workstations for prolonged periods, and the incidence of hospital falls and intensive care unit (ICU) transfers.

Methods: This retrospective observational study was conducted in a 400-square-meter acute care unit of a tertiary hospital for over 210 days. An ultrawideband real-time location system was deployed to continuously track mobile workstations' spatial coverage. Spatial shadow zones were defined as areas unvisited by mobile workstations for 60 continuous minutes. The primary outcome was hospital falls; the secondary outcome was ICU transfers. Multivariable logistic regression analysis, adjusted for patient-to-nurse ratio and day of week, was used to examine the association between the percentage of spatial shadow zone and these outcomes. Sensitivity analyses were performed by varying the spatial dilation distance (1-4 meters) and temporal shadow zone thresholds (15-90 minutes).

Results: During this study's period, 8 hospital falls and 89 ICU transfers occurred. Real-time location system validation indicated a mean positional error of 0.346 (SD 0.282) meters. In multivariable regression, a higher percentage of spatial shadow zone was significantly associated with an increased odds of hospital falls (odds ratio 1.02, 95% CI 1.01 to 1.03, $P<.001$). Conversely, a higher percentage of spatial shadow zone was associated with decreased odds of ICU transfer (odds ratio 0.99, 95% CI 0.99 to 0.99, $P<.001$). Sensitivity analyses demonstrated consistency of the association between spatial shadow zones and falls across varying parameter settings.

Conclusions: This study provides novel evidence for a significant positive association between the percentage of spatial shadow zones and hospital falls, underscoring the critical role of caregiver visibility in fall prevention. The findings suggest that proactively minimizing spatial shadow zones through optimized hospital design, workflow strategies, and technology-enabled monitoring may be a valuable approach to enhance patient safety and reduce hospital falls in acute care settings.

XSL•FO
RenderX

(JMIR Form Res 2025;9:e75697) doi:10.2196/75697

KEYWORDS

indoor positioning systems; IPS; patient safety; health care; ultra-wideband; UWB

Introduction

Background

Hospital falls constitute a persistent and significant threat to both patient safety and the occupational well-being of health care staff within global health care systems, an issue documented for decades [1-3]. Research indicates that the anticipated increase in hospital fall incidence, driven by the expanding global aging population [4,5], will likely exacerbate health care expenditures [6]. Mitigating hospital falls remains a paramount concern for health care institutions worldwide. Numerous factors contribute to the occurrence of these incidents, spanning patient-specific characteristics to environmental hazards and organizational protocols [4,7,8]. Within this multifaceted landscape, visibility, an extrinsic factor related to the physical environment and staff awareness, has emerged as a critical area of investigation.

Existing literature underscores the influence of unit layouts and spatial configurations on caregiver visibility and, consequently, patient safety. Some studies have focused on how ward design impacts nurses' and other caregivers' cognitive load in maintaining patient surveillance [9,10]. These works suggest that certain spatial arrangements may inadvertently create blind spots or areas with reduced visibility, potentially hindering timely intervention and increasing fall risk. Conversely, other research has explored the implementation of decentralized or satellite nursing stations situated closer to patient rooms as a strategy to enhance visibility and responsiveness [6,11]. Furthermore, the role of mobile workstations or nursing carts has been considered as a means to improve caregiver proximity and visibility within patient care areas [12,13]. Regular patient rounding, a proactive nursing practice involving scheduled checks on patients, is also recognized as a potential intervention to improve visibility and address patient needs preemptively, although its effectiveness in significantly reducing falls remains a subject of ongoing investigation [6,14-16].

While existing research provides valuable insights into how environmental design and workflow impact visibility and fall prevention, a significant gap persists: the need for objective and continuous visibility assessment in hospitals. Current methods, often relying on manual observation or simulation studies [12,17], are limited. Manual observation, while informative, is susceptible to observer bias and struggles to capture comprehensive spatial-temporal data over extended periods. For example, 1 study improving visibility used a labor-intensive approach with observers manually logging staff positions and actions at set intervals across designated zones [12]. Another study used simulations to calculate the number of beds visible from nursing stations and then relied on paired observers to track staff frequency in patient rooms and nurse stations. Room occupancy was logged each time staff entered, and stopwatches were used to measure staff time spent in different areas during 4-hour observation shifts [17]. Such methodologies, while providing snapshots of activity, may not fully represent the dynamic and nuanced nature of staff movement and the resulting visibility patterns within a hospital unit over time.

To address these limitations and advance the understanding of visibility in fall prevention, this research introduces the application of the real-time location system (RTLS). RTLS technologies offer an objective means to track the real-time location and movement of equipment within indoor environments [18,19]. Compared to manual observation, RTLS provides continuous, granular data on spatial occupancy and movement patterns, enabling a more comprehensive and data-driven analysis of visibility dynamics. Prior studies have demonstrated the utility of RTLS in health care for diverse applications, including space use analysis and workflow optimization [20,21]. However, the potential of RTLS to precisely track the movement trajectories of mobile resources, such as nursing carts (satellite nursing stations), and to conduct fine-grained spatial analysis of caregiver activity spaces in relation to patient falls remains largely unexplored.

The primary objective of this study is to leverage an RTLS to analyze the movement trajectories and spatial coverage of health care staff within hospital patient rooms and correlate these patterns with the incidence of hospital falls. Specifically, we hypothesize that the percentage of unvisited areas (spatial shadow zones) within patient rooms over a defined period may be positively associated with fall rates. Furthermore, this research aims to identify an optimal interval for regular staff rounding (eg, 60 minutes or another empirically derived timeframe) that balances proactive patient monitoring with efficient resource allocation. By quantifying the relationship between the temporal dynamics of staff visibility, the spatial distribution of unvisited areas, and fall occurrence, this study seeks to contribute to the development of evidence-based strategies for enhancing patient safety and optimizing staff workflow in hospital settings.

Related Works

RTLS provides a means for continuously and in real-time determining the location of individuals or objects within indoor environments [18]. They have found increasing application within health care settings, primarily for the localization and tracking of medical equipment and personnel to enhance operational efficiency and optimize patient flow [19,22]. The selection of an appropriate RTLS technology is contingent upon the specific application requirements, with key considerations including positional accuracy, cost of deployment and maintenance, scalability, and potential for interference with medical devices [23]. Several RTLS technologies have been explored for health care applications, each with distinct characteristics. Radio-frequency identification (RFID) has been frequently investigated due to its relatively low cost; however, RFID systems typically offer room-level accuracy and are susceptible to interference with certain medical equipment [24].

XSL•FO
RenderX

Furthermore, achieving finer-grained location data with RFID often necessitates a denser infrastructure deployment, potentially increasing complexity and cost. Wireless Fidelity (Wi-Fi) based positioning leverages existing hospital infrastructure, as Wi-Fi networks are widely available in most health care facilities, thereby minimizing additional infrastructure investment. However, Wi-Fi positioning generally provides lower accuracy (in the range of 3-5 meters) and may require multiple access points to ensure adequate coverage and precision. Technologies such as ZigBee (Connectivity Standards Alliance) and Bluetooth Low Energy (Bluetooth Special Interest Group) offer advantages in terms of energy efficiency and tag miniaturization, but tend to exhibit lower positional accuracy compared to other modalities. Infrared systems can achieve high accuracy but are limited by line-of-sight requirements, restricting their applicability in complex and occluded hospital environments. Ultrasonic systems offer low cost and energy consumption but often suffer from poor scalability and challenges in tracking multiple moving objects concurrently [25].

In contrast to these alternatives, ultrawideband RTLS technology emerges as a particularly promising sensor choice for applications demanding high precision. Survey literature highlights ultrawideband's capacity for centimeter-level accuracy, good resistance to multipath propagation, low power consumption, compact tag size, and high operating frequencies that mitigate interference [22]. While potentially incurring higher upfront costs compared to some other RTLS options, the superior accuracy and real-time tracking capabilities of ultrawideband are crucial for applications requiring detailed spatial analysis. For instance, while RFID-based RTLS may be suitable for asset tracking where room-level accuracy suffices, precise monitoring of staff workflows and movement trajectories necessitates the finer spatial resolution afforded by ultrawideband. A systematic review of RTLS technologies in health care also underscores ultrawideband as one of the most accurate technologies, capable of providing tracking precision down to 20 centimeters [22].

Given the objective of this study to investigate the association between mobile workstation movement patterns, spatial shadow zones within a health care unit, and the incidence of patient falls, the precise tracking of mobile workstation trajectories and the granular spatial analysis of caregiver activity spaces are paramount. Therefore, ultrawideband RTLS technology was selected for the technological implementation of this research. Its capacity to provide centimeter-level accuracy and real-time location updates makes it uniquely suited to capture the nuanced spatial dynamics of caregiver movement and to reliably quantify the percentage of spatial shadow zones. Furthermore, established methodologies for validating RTLS accuracy, often using cumulative distribution function (CDF) analysis of the distance error between reported and actual positions [25,26], provide a framework for assessing the performance of the deployed ultrawideband system, as detailed in the subsequent Methods section.

Methods

Study Design and Settings

This study was conducted within the acute care unit of a tertiary hospital, specifically within a medical observation area dedicated to patients who have been stabilized following emergency treatment but require ongoing care while awaiting inpatient ward admission. The patient population within this medical observation area primarily comprised acute medical patients who have been stabilized following initial emergency treatment but require continued monitoring and care while awaiting transfer to an inpatient ward. This population presented with a diverse range of diagnoses commonly encountered in emergency observation units, including, but not limited to, respiratory ailments, cardiovascular conditions, sepsis, and complications related to cancer.

This study's setting comprised a 400-square-meter acute care unit characterized by a hybrid nurse station configuration. This configuration incorporated a central nursing station supplemented by mobile workstations to facilitate decentralized care delivery. These mobile workstations served as crucial nodes for clinical workflow, supporting a range of essential nursing tasks directly at the patient's bedside. Functionally, mobile workstations were used for documenting patient information and reviewing patient histories, accessing supplies, managing waste disposal, facilitating real-time medication administration, and performing other immediate nursing care activities. Each mobile workstation weighs approximately 75 to 120 kg and is equipped with a computer, medical devices, and a work surface, as well as storage space for clinical supplies and medications [13,27]. Medical coverage within the unit was provided continuously by resident physicians in 2 shifts around the clock. Nursing care was delivered by registered nurses working in 3 shifts, with approximate patient-to-nurse ratios of 7:1 during daytime shifts, 9:1 during evening shifts, and 10:1 during nighttime shifts. Data collection for both the RTLS and patient outcomes spanned a period of 210 days, from August 4, 2024, to March 1, 2025, encompassing a total of 30 weeks.

RTLS Infrastructure

An ultrawideband RTLS was deployed within the acute care unit to continuously monitor the location of mobile workstations (Figure 1). The RTLS infrastructure consisted of 6 ultrawideband anchors strategically positioned at fixed locations throughout the unit to ensure comprehensive spatial coverage. Ultrawideband tags were affixed to each mobile workstation. These tags emitted radio signals at a frequency of 1 Hz during periods of motion. To optimize battery life, the tag transmission rate was reduced to 1 signal per minute when the mobile workstation remained stationary. The ultrawideband anchors received signals from the tags, and this data was transmitted to a central server. The server used time-of-flight algorithms to calculate the 3D coordinates of each mobile workstation within the unit in real time.

XSL•FO
RenderX

Figure 1. Acute care unit layout and RTLS anchor placement. Depiction of the acute care unit study setting. The unit area, delineated in green, is virtually partitioned into a grid of 1-meter by 1-meter cells. Purple triangles indicate the locations of the 6 ultrawideband RTLS anchors deployed within the unit. Anchor placement was constrained by existing power outlet locations and medical building infrastructure regulations. RTLS: real-time location system.

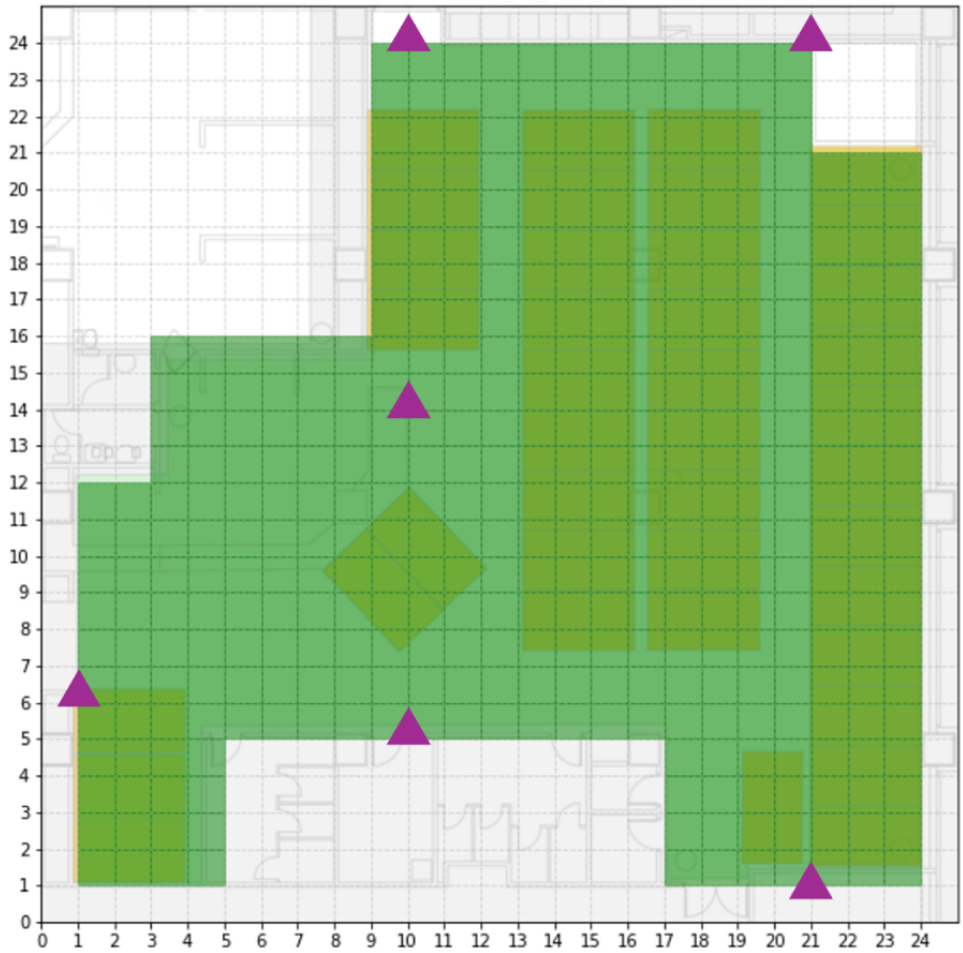

Outcomes

The primary outcome measure for this study was the incidence of hospital falls within the acute care unit. Hospital falls were defined as unintentional events resulting in a person coming to rest on the ground. Data on fall events were extracted from the hospital's incident reporting system. The secondary outcome measure was patient clinical deterioration requiring transfer to the intensive care unit (ICU). Data on these transfers were obtained from hospital electronic health records.

Spatial Data Processing and Analysis

To facilitate spatial analysis, the 400 square meter acute care unit was virtually partitioned into a grid of 1-meter by 1-meter square cells. For each mobile workstation, a "coverage area" was defined based on its real-time location. This coverage area was operationalized as the grid cell occupied by the mobile workstation, expanded by a spatial dilation of 2 grid cells in all cardinal and diagonal directions. This spatial dilation was implemented to approximate the visual and functional range of a caregiver in proximity to the mobile workstation. A dilation distance of 2 meters was selected, which is consistent with established literature defining the personal and social zones in proxemics theory and interpersonal distances [28,29]. Informed by Hall [29] distance theory, particularly the characteristics of the social and personal zones, this study operationalizes caregiver visibility by defining a coverage area around mobile workstations based on a spatial dilation of 2 meters. This 2-meter radius is intended to represent an approximate zone within which caregivers, when positioned with a mobile workstation, can reasonably be expected to maintain visual awareness and effectively respond to patient needs, balancing the need for proximity with practical considerations of workflow and spatial constraints in the acute care setting. We recognize that actual caregiver visibility and responsiveness in a dynamic acute care setting are influenced by a multitude of factors, including line-of-sight obstructions, physical barriers, and variability in clinical tasks. Thus, this 2-meter radius serves as a standardized operational definition for quantifying spatial coverage for the purposes of this specific research study. To assess the findings of this specific operationalization, sensitivity analyses were systematically conducted across a range of spatial dilation distances (1 to 4 meters), as detailed in the Sensitivity Analyses section.

RTLS Accuracy Validation

Before data collection, a validation procedure was conducted to assess the positional accuracy of the deployed RTLS within the acute care unit. A predefined route, representative of typical mobile workstation movement patterns within the unit, was

XSL•FO
RenderX

established and marked. Four ultrawideband tags, affixed to a mobile workstation, were then moved along this designated route at a controlled speed of approximately 0.5 to 1 meter per second. Simultaneously, location data was recorded by the RTLS. To quantify positional accuracy, the error distance for each recorded RTLS position was calculated by comparing it to the corresponding ground truth position along the predefined route. The distribution of these error distances was then visualized using a CDF plot.

Definition of Spatial Shadow Zones and Percentage of Spatial Shadow Zone

Spatial shadow zones were defined as grid cells within the acute care unit that were not included within the coverage area of any mobile workstation for a continuous period of 60 minutes. This 60-minute interval was selected as consistent with previous literature on rounding frequency [6,14]. The percentage of spatial shadow zone was calculated as the total area of spatial shadow zones within the unit, divided by the total area of the acute care unit (400 square meters). This percentage was calculated for each minute to represent the temporal dynamics of unvisited areas.

Sensitivity Analyses

To assess the robustness of the primary findings, sensitivity analyses were conducted by systematically varying key parameters in the spatial analysis. First, the spatial dilation distance defining the mobile workstation coverage area was varied across a range of 1 to 4 meters (1-meter, 2-meter, 3-meter, and 4-meter dilation distances were tested). Second, the temporal threshold for defining spatial shadow zones was examined across multiple durations, ranging from 15 to 90 minutes (specifically, periods of 15, 30, 45, 60, 75, and 90 minutes were evaluated). These variations were implemented to explore the sensitivity of the results to different operationalizations of caregiver visibility and unvisited areas.

In addition to parameter variations, a temporal sensitivity analysis was performed by dividing the total 210-day study period into 2 equal segments: the first 15 weeks and the subsequent 15 weeks. The primary analyses were then repeated separately for each of the 15 weeks. This temporal segmentation was conducted to investigate potential variations in the observed relationships over time, and to assess the consistency of the findings across different phases of the data collection period.

Statistical Analysis

Data processing and spatial analyses were conducted using Python (version 3.11.9; Python Software Foundation). Statistical analyses were performed using the R statistical software environment (version 4.3.1; R Foundation). Continuous variables were compared between independent groups using the independent samples *t* test (2-tailed). To assess the association between the percentage of spatial shadow zone and the dichotomous outcomes of hospital falls and ICU transfer, multivariable logistic regression analysis was used. In these regression models, the percentage of spatial shadow zone served as the primary independent variable. To account for potential confounding effects related to temporal variations in hospital activity and staffing, we adjusted for day of the week and patient-to-nurse ratio as covariates in the multivariable models. All statistical tests were 2-tailed, and statistical significance was defined as a *P* value less than .05.

Ethical Considerations

This study was granted an exemption from full review by the National Taiwan University Hospital Research Ethics Committee (NTUH-REC number: 202403045W; March 28, 2024). As this study involved retrospective analysis of anonymized, aggregated location data and did not involve direct interaction with human subjects, informed consent was not required and was not obtained. The study utilized aggregated, anonymized data, and no patient or staff identifying information was collected or analyzed, to make the data collected from the systems of that specific place as discrete as possible. Given the retrospective observational nature of the study and the lack of direct interaction with participants, no compensation was provided for participation.

Results

Study Population and Event Incidence

Over the 210-day study period, a total of 97 adverse events were recorded within the acute care unit. Among these, 8 were identified as hospital falls. Given the unit's 40-bed capacity and the 210-day observation period, the hospital fall rate was calculated as 1.0 per 1000 patient-days. Additionally, 89 patients experienced clinical deterioration necessitating transfer to the ICU.

Analysis of faller type revealed that 6 falls involved patients, 1 involved a patient's family member, and 1 involved a staff member. The temporal distribution of fall events indicated occurrences across various hours of the day: 7 AM (n=2), 9 AM (n=1), 10 AM (n=1), 11 AM (n=1), 4 PM (n=1), and 11 PM (n=2). The distribution of causes for ICU transfer was as follows: advanced respiratory support (n=45), circulatory support (n=35), cardiac arrest (n=4), neurological monitoring (n=3), acute coronary syndrome (n=1), and diabetic ketoacidosis (n=1). The temporal distribution of ICU transfers based on the nursing shift structure was: nighttime shift (n=16), daytime shift (n=24), and evening shift (n=49; Multimedia Appendices 1 and 2).

RTLS Positional Accuracy

To validate the accuracy of the RTLS, a dedicated accuracy assessment was performed. Data from 4 ultrawideband tags, collected along a predefined route, yielded 419 positional estimates (Figure 2). Descriptive statistics for the positional error, defined as the distance between the RTLS-estimated positions and the ground truth route, revealed a mean error of 0.346 meters, 95% CI 0.319 to 0.373. The CDF of positional errors is provided in Figure 3, visually illustrating the distribution of RTLS accuracy.

XSL•FO
RenderX

Figure 2. RTLS positional accuracy validation trajectory. Visualization of the RTLS positional accuracy validation procedure. The red line delineates the predefined ground truth route, representing a typical mobile workstation movement trajectory within the acute care unit. Colored dots represent the 419 positional estimates recorded by the RTLS from 4 ultrawideband tags affixed to a mobile workstation as it was moved along the predefined route. Yellow rectangles indicate the locations of the 40 patient beds within the unit. RTLS: real-time location system.

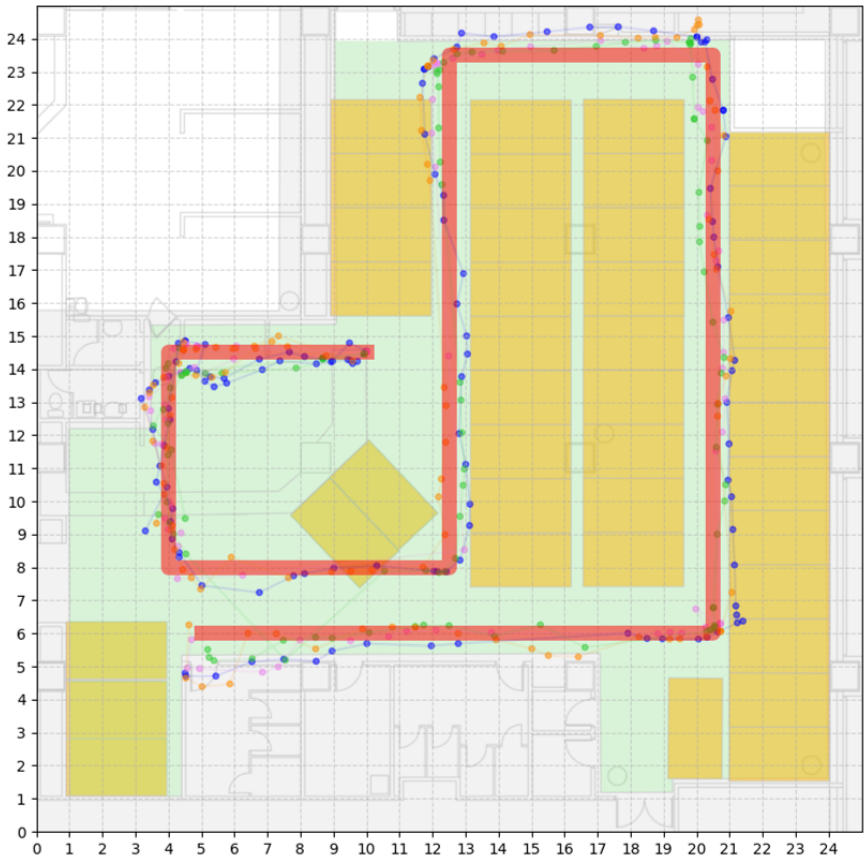

Figure 3. CDF of RTLS positional error distances. CDF plot illustrating the distribution of positional error distances from the RTLS accuracy validation. The x-axis represents the positional error distance in meters, and the y-axis represents the cumulative probability. The vertical dashed line at x=0.554 meters and the horizontal dashed line at y=0.8 indicate that 80% of the RTLS positional estimates were within an error distance of 0.554 meters or less. CDF: cumulative distribution function; RTLS: real-time location system.

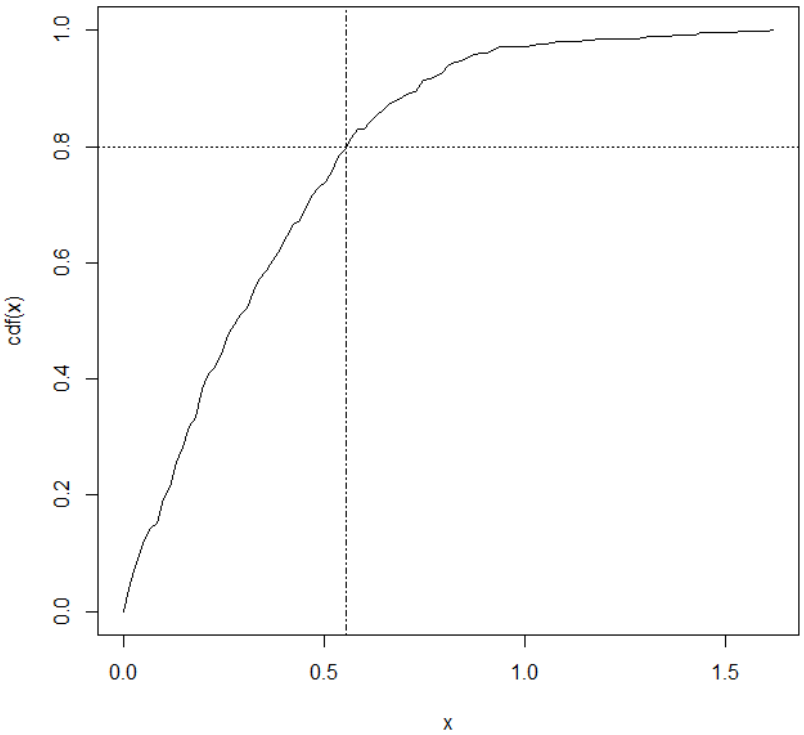

XSL•FO
RenderX

RTLS Data Quality Metrics

Analysis of RTLS data quality revealed a mean missing data rate of 0.376, 95% CI 0.369 to 0.383 per hour, after excluding periods of ultrawideband tag inactivity-induced data reduction. Furthermore, an evaluation of potential data artifacts identified instances of spurious positional displacement, characterized as isolated outlier points relative to preceding and subsequent positions more than 2.5 meters. Across a total of 47,230,357 recorded location data points, 106,564 points were classified as noise according to these criteria, representing an overall noise rate of approximately 0.226% (1 in 443 data points).

Spatial Shadow Zone and Clinical Outcomes

Descriptive statistics for the percentage of spatial shadow zone, stratified by fall events and ICU transfers, are presented in Table 1. The mean percentage of spatial shadow zone was numerically higher in instances of falls, exhibiting an average increase of 1.0 percentage point (95% CI 0.0 to 1.9), compared to nonfall events. When the analysis was restricted to patient falls only (excluding falls involving staff and visitors), the difference in mean percentage of spatial shadow zone remained numerically higher in fall events but was no longer statistically significant, demonstrating an average increase of 0.4 percentage points (95% CI −0.7 to 1.5). Conversely, the mean percentage of spatial shadow zone was statistically significantly lower in instances of ICU transfer, demonstrating an average decrease of 1.1 percentage points (95% CI 0.8 to 1.4), when compared to non-ICU transfer events. Independent samples *t* tests indicated a statistically significant difference in the mean percentage of spatial shadow zone between fall and nonfall events (*P*=.04) and between ICU transfer and non-ICU transfer events (*P*<.001).

Table 1. Percentage of spatial shadow zone by event occurrence.

Event	Percentage of spatial shadow zone (%), mean (SD)	*P* value
Overall (n=250,363)	76.8 (10.3)	__[a]
Fall events (n=470)	77.8 (10.9)	.04
Nonfall events (n=249,893)	76.8 (10.3)	.04
ICU[b] transfer events (n=4659)	75.8 (10.0)	<.001
Non-ICU transfer events (n=245,704)	76.9 (10.3)	<.001

[a]Not applicable.

[b]ICU: intensive care unit.

Multivariable Logistic Regression Analysis of Hospital Falls

The results of the multivariable logistic regression analysis examining the association between the percentage of spatial shadow zone and the odds of hospital falls are presented in Table 2. After adjusting for patient-to-nurse ratio and day of the week, the percentage of spatial shadow zone was significantly associated with an increased odds of hospital falls (odds ratio 1.02, 95% CI 1.01 to 1.03, *P*<.001). The coefficients for the day-of-week were not statistically significant (*P*>.05).

Table 2. Multivariable logistic regression analysis of hospital falls.

Predictor	Odds ratio (95% CI)	*P* value
Percentage of spatial shadow zone	1.02 (1.01 to 1.03)	.001
Patient-to-nurse ratio	0.84 (0.78 to 0.90)	<.001
Day of week (adjusted)	__[a]	>.05

[a]Not applicable.

Sensitivity Analyses Results

To assess the robustness of the primary finding regarding the association between the percentage of spatial shadow zone and hospital falls, sensitivity analyses were conducted by varying the spatial dilation distance and the temporal threshold for defining spatial shadow zones. Independent samples *t* tests were performed to compare the mean percentage of spatial shadow zone between fall and nonfall events across these varied parameters.

As shown in Table 3, the statistical significance of the difference in mean percentage of spatial shadow zone between fall and nonfall events varied depending on the spatial dilation distance and the temporal threshold for shadow zone definition. While the effect size was dependent on the choice of these parameters, statistically significant differences were observed across several parameter combinations. When the spatial dilation distance was reduced to 1 meter, statistically significant differences were observed across temporal thresholds ranging from 15 to 75 minutes. However, as the dilation distance increased, statistical significance was attenuated and lost at temporal thresholds of 60 minutes and above. With a dilation distance of 2 meters (the primary analysis parameter), statistical significance was observed at temporal thresholds of 15, 30, 45, and 60 minutes, but not at 75 or 90 minutes.

These sensitivity analyses suggest that the observed association between the percentage of spatial shadow zone and hospital falls was most robust within a temporal shadow zone threshold

XSL•FO
RenderX

of up to 45 minutes. Beyond these parameters, the statistical evidence for a significant association weakens.

In addition to varying spatial and temporal parameters for shadow zone definition, a temporal sensitivity analysis was conducted to assess the consistency of findings across different phases of this study's period. In the analysis of the first 15-week period, the mean percentage of spatial shadow zone was numerically higher for fall events (mean 78.5) compared to nonfall events (mean 77.6), but this difference was not statistically significant (*P*=.15, 95% CI −0.3 to 2.2). Similarly, for the subsequent 15-week period, while the mean percentage of spatial shadow zone remained numerically higher for fall events (mean 77.0) compared to nonfall events (mean 76.0), this difference also did not reach statistical significance (*P*=.14, 95% CI −0.3 to 2.4).

Table 3. Sensitivity analysis and 95% CIs for the difference in means of percentage of spatial shadow zone between nonfall and fall events across varying spatial dilation distances and temporal shadow zone thresholds.

Temporal shadow zone thresholds (minutes)	Dilation distance (meters), 95% CI			
	1	2	3	4
15	1.38 to 3.08[a]	2.67 to 6.27[a]	3.81 to 9.48[a]	4.45 to 11.47[a]
30	0.98 to 2.20[a]	1.54 to 4.12[a]	2.00 to 6.07[a]	2.40 to 7.43[a]
45	0.42 to 1.43[a]	0.51 to 2.65[a]	0.60 to 3.98[a]	0.79 to 4.97[a]
60	0.24 to 1.12[a]	0.03 to 1.90[a]	−0.17 to 2.78	−0.13 to 3.51
75	0.07 to 0.86[a]	−0.34 to 1.34	−0.82 to 1.83	−1.02 to 2.26
90	−0.07 to 0.66	−0.76 to 0.79	−1.49 to 0.96	−1.87 to 1.15

[a]Statistically significant.

Discussion

Principal Results

This study provides novel evidence for a significant positive association between the percentage of spatial shadow zone within an acute care unit and the incidence of hospital falls. Leveraging an RTLS for objective and continuous measurement of mobile workstation spatial coverage, we found that a higher percentage of unvisited areas over time was associated with substantially increased odds of fall events. This finding underscores the critical role of caregiver visibility and proactive patient monitoring in mitigating fall risk within hospital settings. Conversely, we observed an inverse relationship between the percentage of spatial shadow zone and the odds of ICU transfer.

The observed positive association between spatial shadow zones and hospital falls is congruent with the established understanding of visibility as a crucial extrinsic factor in patient safety and fall prevention [10]. Spatial shadow zones, as operationally defined in this study, represent periods and locations within the patient care environment where mobile workstations—and by extension, readily available caregivers—were absent for a continuous 60-minute interval. These extended periods of reduced caregiver presence could diminish opportunities for routine patient monitoring, proactive identification of evolving patient needs, and timely intervention to prevent potential falls. Our findings build upon and extend prior research that has emphasized the importance of unit layout design [17], nursing station configuration [12], and structured rounding practices [16] in shaping caregiver visibility and influencing patient outcomes. The odds ratio observed in our analysis suggests that interventions aimed at reducing spatial shadow zones might represent a potent strategy for enhancing fall prevention efforts.

The observed inverse relationship between spatial shadow zone and ICU transfers necessitates interpretation within the clinical context of this acute care unit. Specifically designed for patients in a transitional phase of postemergency stabilization, yet with preinpatient ward admission, this unit manages a patient population characterized by inherently dynamic and potentially labile conditions. While proactive rounding and consistent caregiver presence are generally posited to enhance patient monitoring, thereby potentially mitigating adverse events, they may not directly avert all instances of clinical deterioration requiring ICU escalation. Our interpretation is that in the temporal window preceding ICU transfer events, the exigency of deteriorating patient conditions could have likely prompted heightened nursing vigilance and intensified bedside care delivery. This increased demand for immediate, patient-centered interventions was typically observed to result in sustained caregiver presence at the patient's bedside, with mobile workstations consequently deployed and used near acutely ill individuals. This focused allocation of nursing resources and bedside workstation use would, in turn, reduce the percentage of spatial shadow zone within the unit during the 60-minute interval preceding ICU transfer. Under this interpretation, the diminished spatial shadow zone is not a protective factor preventing ICU transfer, but rather a potential marker of escalating patient acuity and the resultant, appropriately intensified nursing response characterized by concentrated bedside care.

Comparison With Prior Work

Several prior studies have investigated the impact of hourly rounding on patient fall rates, often using cohort study designs to analyze pre- and postimplementation fall rates [14-16]. These studies generally hypothesize that hourly rounding, by increasing caregiver presence and proactive patient interaction, would reduce the occurrence of spatial shadow zones within patient

XSL•FO
RenderX

units. Our sensitivity analyses, exploring varying temporal thresholds for shadow zone definition, suggest that a 45-minute interval may represent a more robust and potentially clinically relevant timeframe for observing significant associations with fall risk, particularly when considering the spatial dilation of mobile workstation coverage areas. While some research suggests that increasing rounding frequency beyond hourly intervals (eg, every 30 minutes or more frequently) does not necessarily yield further reductions in fall rates [16] and may indeed increase caregiver workload, our study focuses on the percentage of spatial shadow zone as a continuous metric rather than discrete rounding intervals. This approach offers a nuanced perspective on the temporal dynamics of caregiver visibility beyond simply the frequency of scheduled rounds.

Hospital fall rates, typically reported in the range of 2 to 8 falls per 1000 patient-days [8], are recognized as a relatively infrequent patient safety indicator, posing challenges for detecting statistically significant associations with interventions or risk factors, particularly in studies with limited sample sizes or observation periods [14]. Our temporal sensitivity analysis, segmenting the 210-day study period, indicated a consistent trend of numerically higher shadow zone percentages preceding fall events across both time segments, although statistical significance was not consistently achieved in these segmented analyses. This may reflect the inherent variability in fall occurrence and the need for larger datasets or longer observation periods to detect temporal relationships, especially for less frequent events.

The implementation of RTLS technology for objective spatial measurement represents a key methodological advancement in this study. Compared to prior work relying on manual observation or proxy measures of visibility, RTLS enables continuous and granular data collection, overcoming limitations inherent in traditional methods [12,17]. Our RTLS positional accuracy, with a mean error of 35 cm and approximately 70% of measurements within 45 cm of the ground truth, is slightly worse than previously reported accuracy for ultrawideband systems in ideal spatial settings, such as the 25 cm mean error and 70% within 40 cm reported by Silvia et al [25].

Limitations

There are several limitations in this study. This study was conducted within a single acute care unit in a tertiary hospital, potentially limiting the generalizability of our findings to other hospital settings with different patient populations, unit layouts, staffing models, or care protocols. The operational definition of spatial shadow zone, while informed by clinical considerations and practical feasibility, remains a specific metric, and future research could explore alternative definitions, temporal thresholds, and spatial parameters. Furthermore, while multivariable regression models adjusted for temporal factors such as day of the week, residual confounding by unmeasured or imperfectly measured variables cannot be entirely excluded. In particular, this analysis lacked granular data on individual-level characteristics, such as patient acuity, nurse experience, or visitor or staff health status, which may potentially affect the results. Finally, our outcome measures relied on hospital incident reporting systems and electronic health record data, which may be subject to potential underreporting, misclassification, or inconsistencies in data capture. This study does not include a power calculation or formal sample size justification for the number of observed fall events, given the relative rarity of this outcome, and raises concerns about the statistical power to detect meaningful associations.

Despite these limitations, this study provides evidence for the association between the percentage of spatial shadow zones and hospital falls. Moving forward, these findings have several implications for patient safety and health care operations. The quantifiable relationship between spatial shadow zones and hospital falls suggests that proactively minimizing unvisited areas within acute care units might be a valuable and targeted strategy for fall prevention. Hospital design and unit layout should prioritize maximizing caregiver visibility and minimizing spatial configurations that contribute to prolonged periods of reduced surveillance. Real-time monitoring of spatial shadow zones, enabled by RTLS technology, could serve as a proactive risk indicator, prompting targeted interventions during periods or in areas with elevated shadow zone percentages. Future research should focus on interventional studies to directly test the impact of strategies specifically aimed at reducing spatial shadow zones on hospital fall rates and other relevant patient outcomes. Multicenter studies encompassing diverse hospital settings and patient populations are needed to enhance the generalizability of these findings. Investigation into the inverse association with ICU transfers is crucial to elucidate its underlying mechanisms and clinical significance, potentially informing more nuanced and context-specific strategies for optimizing caregiver allocation and patient monitoring in acute care environments. Subsequent research should also explore the potential for leveraging RTLS technology to develop intelligent, real-time decision support systems, perhaps akin to a “fog of war” visualization, to proactively guide caregiver rounding and resource allocation, aiming to systematically and efficiently maintain low spatial shadow zone percentages without increasing caregiver workload. Future research should also address what can improve clinician efficiencies, decrease burnout, potentially serve as an early warning system for changes in patient acuity, and be integrated into existing health care environments.

Conclusions

Leveraging RTLS technology, this research objectively quantified spatial shadow zones and demonstrated their potential as a novel, quantifiable risk indicator associated with hospital falls. The findings underscore the importance of hospital design and workflow optimization strategies that prioritize maximizing caregiver spatial coverage and minimizing prolonged periods of unvisited patient areas, while giving a quantifiable metric tool that can assist in how the areas are more precisely monitored.

XSL•FO
RenderX

Acknowledgments

The authors gratefully acknowledge the contributions of Gemini 2.0 Flash (Google AI) and GPT-4o (OpenAI) for their assistance in enhancing the quality of this paper through proofreading and verification of grammatical and spelling accuracy. We also extend our sincere appreciation to Ms Sung, research assistant, for her support throughout this study in coordinating the installation and ongoing maintenance of the RTLS equipment within the medical unit. This work was supported by grants from the National Science and Technology Council, Taiwan (NSTC 113-2221-E-002-162); the Industrial Technology Research Institute, Taiwan; and National Taiwan University Hospital.

Authors' Contributions

YPC, CLL, and CHH contributed to this study's conceptualization and design. YPC, BJY, and CHH were responsible for the planning, coordination of research activities within the medical unit, and data collection. YPC, YCC, and CLL contributed to the planning and coordination of the real-time location system (RTLS) equipment installation, positioning algorithms, and data analysis algorithms. YPC, CLL, CHH, and YCC secured funding for the research. YPC, CYC, and YYC conducted the paper preparation and data analysis. YPC and CHH undertook the paper editing and final review. All authors have read and approved the final paper.

Conflicts of Interest

Author YCC is affiliated with Guidercare Limited. Guidercare Limited provided the RTLS equipment, system installation, and proprietary positioning algorithms used in this study at no cost to the research team. This in-kind contribution facilitated the technological implementation of this study. All other authors declare that they have no conflicts of interest.

Multimedia Appendix 1

Dynamic visualization of spatial shadow zones: fog of war analogy. This video illustrates the minute-by-minute evolution of spatial shadow zones over a 1-hour period. Areas in proximity to mobile workstations are visually illuminated, while shadow zones, representing areas unvisited for the defined temporal threshold, progressively darken, analogous to the "fog of war" concept in real-time strategy games.

[MP4 File (MP4 Video), 1123 KB - formative_v9i1e75697_app1.mp4]

Multimedia Appendix 2

Real-time dynamics of spatial shadow zones: a "fog of war" representation. This video demonstrates the temporal dynamics of spatial shadow zones within the acute care unit over a 60-minute period. Areas within the coverage radius of mobile workstations are depicted as "illuminated," while regions exceeding the temporal shadow zone threshold become progressively darker, visually mimicking the receding "fog of war" as caregiver presence wanes.

[MP4 File (MP4 Video), 1030 KB - formative_v9i1e75697_app2.mp4]

References

1. Morgan VR, Mathison JH, Rice JC, Clemmer DI. Hospital falls: a persistent problem. Am J Public Health 1985;75(7):775-777. [doi: 10.2105/ajph.75.7.775] [Medline: 4003655]
2. Min L, Tinetti M, Langa KM, Ha J, Alexander N, Hoffman GJ. Measurement of fall injury with health care system data and assessment of inclusiveness and validity of measurement models. JAMA Netw Open 2019;2(8):e199679 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.9679] [Medline: 31433480]
3. Jørgensen TSH, Hansen AH, Sahlberg M, Gislason GH, Torp-Pedersen C, Andersson C, et al. Nationwide time trends and risk factors for in-hospital falls-related major injuries. Int J Clin Pract 2015;69(6):703-709. [doi: 10.1111/ijcp.12624] [Medline: 25707918]
4. Taylor E, Hignett S. The SCOPE of hospital falls: a systematic mixed studies review. HERD 2016;9(4):86-109. [doi: 10.1177/1937586716645918] [Medline: 27240563]
5. Cigolle CT, Ha J, Min LC, Lee PG, Gure TR, Alexander NB, et al. The epidemiologic data on falls, 1998-2010: more older Americans report falling. JAMA Intern Med 2015;175(3):443-445 [FREE Full text] [doi: 10.1001/jamainternmed.2014.7533] [Medline: 25599461]
6. Gutierrez F, Smith K. Reducing falls in a definitive observation unit: an evidence-based practice institute consortium project. Crit Care Nurs Q 2008;31(2):127-139. [doi: 10.1097/01.CNQ.0000314473.72001.b4] [Medline: 18360143]
7. Evans D, Hodgkinson B, Lambert L, Wood J. Falls risk factors in the hospital setting: a systematic review. Int J Nurs Pract 2001;7(1):38-45. [doi: 10.1046/j.1440-172x.2001.00269.x] [Medline: 11811346]
8. Morris M, Webster K, Jones C, Hill A, Haines T, McPhail S, et al. Interventions to reduce falls in hospitals: a systematic review and meta-analysis. Age Ageing 2022;51(5):afac077 [FREE Full text] [doi: 10.1093/ageing/afac077] [Medline: 35524748]

XSL•FO
RenderX

9. Mokarami H, Eskandari S, Cousins R, Salesi M, Kazemi R, Razeghi M, et al. Development and validation of a nurse station ergonomics assessment (NSEA) tool. BMC Nurs 2021;20(1):83 [FREE Full text] [doi: 10.1186/s12912-021-00600-8] [Medline: 34059027]
10. Brewer BB, Carley KM, Benham-Hutchins M, Effken JA, Reminga J. Nursing unit design, nursing staff communication networks, and patient falls: are they related? HERD 2018;11(4):82-94 [FREE Full text] [doi: 10.1177/1937586718779223] [Medline: 29916273]
11. Fay L, Cai H, Real K. A systematic literature review of empirical studies on decentralized nursing stations. HERD 2019;12(1):44-68. [doi: 10.1177/1937586718805222] [Medline: 30387362]
12. Cai H. Out of sight, out of reach: Correlating spatial metrics of nurse station typology with nurses? communication and co-awareness in an intensive care unit. 2012 Presented at: Proceedings of the 8th International Space Syntax Symposium; January 06, 2012; Santiago, Chile p. 381-391 URL: http://sss8.cl/8039.pdf
13. Kao HC, Lin CJ, Lee YH, Chen SH. The effects of direction of exertion, path, and load placement in nursing cart pushing and pulling tasks: an electromyographical study. PLoS One 2015;10(10):e0140792 [FREE Full text] [doi: 10.1371/journal.pone.0140792] [Medline: 26485039]
14. Krepper R, Vallejo B, Smith C, Lindy C, Fullmer C, Messimer S, et al. Evaluation of a standardized hourly rounding process (SHaRP). J Healthc Qual 2014;36(2):62-69. [doi: 10.1111/j.1945-1474.2012.00222.x] [Medline: 23237186]
15. Meade CM, Bursell AL, Ketelsen L. Effects of nursing rounds: on patients' call light use, satisfaction, and safety. Am J Nurs 2006;106(9):58-70; quiz 70. [doi: 10.1097/00000446-200609000-00029] [Medline: 16954767]
16. Meade CM, Kennedy J, Kaplan J. The effects of emergency department staff rounding on patient safety and satisfaction. J Emerg Med 2010;38(5):666-674. [doi: 10.1016/j.jemermed.2008.03.042] [Medline: 18842381]
17. Fay L, Carll-White A, Schadler A, Isaacs KB, Real K. Shifting landscapes: the impact of centralized and decentralized nursing station models on the efficiency of care. HERD 2017;10(5):80-94. [doi: 10.1177/1937586717698812] [Medline: 28359162]
18. Alarifi A, Al-Salman A, Alsaleh M, Alnafessah A, Al-Hadhrami S, Al-Ammar M, et al. Ultra wideband indoor positioning technologies: analysis and recent advances. Sensors (Basel) 2016;16(5):707 [FREE Full text] [doi: 10.3390/s16050707] [Medline: 27196906]
19. Wichmann J. Indoor positioning systems in hospitals: a scoping review. Digital Health 2022;8:205520762210816. [doi: 10.1177/20552076221081696]
20. Fischer GS, Righi RDR, Ramos GDO, Costa CAD, Rodrigues JJ. ElHealth: using internet of things and data prediction for elastic management of human resources in smart hospitals. Eng Appl Artif Intell 2020;87:103285. [doi: 10.1016/j.engappai.2019.103285]
21. Marini G, Tag B, Goncalves J, Velloso E, Jurdak R, Capurro D, et al. Measuring mobility and room occupancy in clinical settings: system development and implementation. JMIR mHealth uHealth 2020;8(10):e19874 [FREE Full text] [doi: 10.2196/19874] [Medline: 33107838]
22. Bazo R, da Costa CA, Seewald LA, da Silveira LG, Antunes RS, Righi RDR, et al. A survey about real-time location systems in healthcare environments. J Med Syst 2021;45(3):35. [doi: 10.1007/s10916-021-01710-1] [Medline: 33559774]
23. Van Haute T, De Poorter E, Crombez P, Lemic F, Handziski V, Wirström N, et al. Performance analysis of multiple indoor positioning systems in a healthcare environment. Int J Health Geogr 2016;15:7 [FREE Full text] [doi: 10.1186/s12942-016-0034-z] [Medline: 26842830]
24. van der Togt R, van Lieshout EJ, Hensbroek R, Beinat E, Binnekade JM, Bakker PJM. Electromagnetic interference from radio frequency identification inducing potentially hazardous incidents in critical care medical equipment. JAMA 2008;299(24):2884-2890. [doi: 10.1001/jama.299.24.2884] [Medline: 18577733]
25. Silvia Z, Martina C, Fabio S, Alessandro P. Ultra wide band indoor positioning system: analysis and testing of an IPS technology. IFAC-PapersOnLine 2018;51(11):1488-1492. [doi: 10.1016/j.ifacol.2018.08.292]
26. Hsiao C, Sung Y, Lau S, Chen C, Hsiao F, Chu H, et al. Towards long-term mobility tracking in NTU hospital's elder care center. : IEEE; 2011 Presented at: 2011 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops); March 21-25, 2011; Seattle, WA, USA p. 649-654. [doi: 10.1109/percomw.2011.5766969]
27. Vanmeenen M, Hirt J, Malfait S, Möhler R. Comparing different scoping and mapping review methodologies: a practical example using the nursing mobile workstation. J Nurs Scholarsh 2024;56(6):802-814. [doi: 10.1111/jnu.13005] [Medline: 39039692]
28. Lu Y. Measuring the structure of visual fields in nursing units. HERD 2010;3(2):48-59. [doi: 10.1177/193758671000300205] [Medline: 21165869]
29. Hall ET. The Hidden Dimension. Hamburg Germany: Anchor; 1966.

Abbreviations

CDF: cumulative distribution function
ICU: intensive care unit

XSL•FO
RenderX

RFID: radio-frequency identification
RTLS: real-time location system
Wi-Fi: Wireless Fidelity

Edited by J Sarvestan; submitted 12.04.25; peer-reviewed by M Ries; comments to author 12.05.25; revised version received 22.06.25; accepted 13.08.25; published 10.09.25.

Please cite as:
Chen YP, Chen YC, Lin CL, Chi CY, Chen YY, Yang BJ, Huang CH
Quantifying Spatial Shadow Zones and Their Association With Hospital Falls in Acute Care Unit: Real-Time Location System Observational Study
JMIR Form Res 2025;9:e75697
URL: https://formative.jmir.org/2025/1/e75697
doi:10.2196/75697
PMID:

©Yen-Pin Chen, Yi-Chun Chen, Chen-Liang Lin, Chien-Yu Chi, Yi-Ying Chen, Bey-Jing Yang, Chien-Hua Huang. Originally published in JMIR Formative Research (https://formative.jmir.org), 10.09.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Impact of a Digital Digestive Health Program and Telehealth Visits in Socially Vulnerable Populations: Cohort Evaluation

Sanskriti Varma[1,2,3,4], MD; Alicen Black[4], RDN, MS; Erin Commons[4], MS, RDN; Pen-Che Ho[4], MS; Dena M Bravata[5], MD, MS; Hau Liu[4,6], MD, MBA, MPH

[1]Division of Gastroenterology, Massachusetts General Hospital, Boston, MA, United States
[2]Center for Neurointestinal Health, Massachusetts General Hospital, Boston, MA, United States
[3]Harvard Medical School, Boston, MA, United States
[4]Cylinder Health, Chicago, IL, United States
[5]Stanford Medicine, Stanford, CA, United States
[6]Department of Medicine, NYU Grossman School of Medicine, New York, NY, United States

Corresponding Author:
Hau Liu, MD, MBA, MPH
Cylinder Health
2045 W Grand Ave Ste B PMB 37767
Chicago, IL, 60612
United States
Phone: 1 888 200 5492
Email: hliu@cylinderhealth.com

Abstract

Background: Socially vulnerable populations have less access to quality gastrointestinal (GI) care. Digital telehealth services provided by GI-focused registered dietitian nutritionists (RDNs) and digestive health coaches (HCs) may improve digestive health outcomes by facilitating access to GI care and thereby reduce health care disparities among vulnerable populations.

Objective: The objectives of this study were to (1) evaluate the impact of a digital digestive health program on reducing GI symptoms among socially vulnerable populations and (2) assess whether telehealth visits with digital app use provide additional benefits in symptom reduction compared to digital app use alone among socially vulnerable populations.

Methods: A comprehensive digital digestive care program with optional telehealth visits with RDNs and HCs was provided to US employees of participating companies via their employee benefits. We enrolled participants in the program between 2022 and 2023 who tracked digestive symptoms multiple times at baseline and then over 90 days. We measured changes in GI symptoms from baseline to up to 3 months, comparing those who opted for telehealth visits with those who used the app only. We stratified participants by the median Social Vulnerability Index (SVI) to evaluate symptom improvement across socially vulnerable populations. Multivariable regressions adjusted for age, gender, race, BMI, and preexisting GI conditions.

Results: A total of 1656 participants met the inclusion criteria, of which 1362 (82%) scheduled at least one telehealth visit and 294 (18%) used only app-based resources. The majority (n=1417 86%) of participants saw GI symptom improvement, with an average reduction of 60% in symptom burden ($P<.001$). Participants who used telehealth services and the app had a reduction in symptoms 16% greater than that of app-only users ($P=.01$). High-SVI participants (ie, those with an SVI score above the median of 0.4, indicating greater social vulnerability) had a 22% greater reduction in GI symptom severity score than app-only high-SVI participants ($P=.04$).

Conclusions: Digital health solutions may be an important resource in advancing equitable access to quality GI care and addressing disparities among populations with high social vulnerability. Virtual telehealth visits with RDNs and HCs appear to be particularly beneficial in improving digestive symptoms in such populations.

(JMIR Form Res 2025;9:e70748) doi:10.2196/70748

KEYWORDS

telehealth; health equity; gastrointestinal diseases; digestive symptoms; digital health

XSL•FO
RenderX

Introduction

Chronic gastrointestinal (GI) diseases, including *Helicobacter pylori* infection, irritable bowel syndrome, pancreatitis, and celiac disease, are highly prevalent in the United States, accounting for high rates of mortality and cost that lead to significant burdens on health care systems, employers, and individual patients [1-3].

Despite this, less than 20% of individuals with abdominal pain, bloating, or diarrhea consult a health care provider for evaluation and management of these symptoms [4]. In this regard, telemedicine may offer increased access to GI specialty care and may also allow for standardized digital digestive care programs to provide effective long-term management options [5].

In addition to the high prevalence and economic burden of chronic GI diseases, considerable disparities exist in access to care across various populations. Patients with digestive diseases may experience gender, race/ethnicity, and economic disparities that lead to differential outcomes, including lower access to screening, lower education and health literacy, and poor food and nutrition [6-8]. For example, Black/African American patients face inequities in colon cancer screening rates, thought to be responsible for up to 42% of the disparity in colorectal cancer incidence and 19% of the disparity in mortality [9,10]. Historically vulnerable populations, including those with lower income, have low rates of telehealth adoption [11]. Structural factors contribute to disparities in telehealth access, including limited broadband access and lower rates of digital literacy [12].

We evaluated the impact of a digital app–based multimodal digestive health program (ie, virtual telehealth visits, app-based content, and tracking) on reducing GI symptoms in socially vulnerable populations. Telehealth visits with registered dietician nutritionists (RDNs) and health coaches (HCs) facilitate the building of trust and rapport between patients and their care team, develop new skills to manage GI issues, and encourage continuity of care. Thus, we sought to evaluate the improvement in chronic GI symptoms among participants who used telehealth visits with care from RDNs and HCs and app-based content compared to those who used app-based content alone.

Methods

Participants

The intervention has been described elsewhere [5,13]. Briefly, adult participants across the United States were provided access to Cylinder Health, a digital digestive chronic care program, through their employee benefits. Employer-approved marketing materials were used for outreach (eg, mailings and emails). Inclusion criteria included (1) preexisting GI diagnoses or symptoms; (2) enrollment in the program between January 1, 2022, and November 28, 2023; (3) symptoms recorded at baseline and on the app between days 30 and 90 after registration; and (4) symptom data recorded more than once. Access to the program was free for participants, who were not otherwise compensated for their participation.

Data Collection

Survey administration and symptom tracking were performed within the app. Participants completed an intake survey at baseline, including information on demographics and GI history (eg, previously diagnosed GI conditions and digestive symptoms over the last week). Each participant was assigned a score on the US Centers for Disease Control and Prevention's Social Vulnerability Index (SVI) based on their 5-digit zip code [14]. The SVI is presented as a percentile from 0 (least vulnerable) to 1 (most vulnerable) that allows geographies to be directly compared based on 16 measures across 4 categories per geography: socioeconomic status (rate of population below 150% of the federal poverty line, unemployed, high housing cost burden, no high school diploma, no health insurance), household characteristics (percentage aged 65 years or older, aged 17 years or younger, civilian with a disability, single-parent households, and English language proficiency), racial and ethnic minority status, and housing type and transportation (rate of multiunit structures, mobile homes, crowding, no vehicle, and group quarters). We collected information on racial group from the patients directly, so we did not use racial information from the SVI. To compare participants on the basis of their social vulnerability, we categorized participants as having high social vulnerability if their SVI score was greater than the median (0.4) and low social vulnerability if their SVI was less than or equal to the median. We then report data for groups stratified into those above or below the median SVI.

GI symptom tracking in the program has been described elsewhere [5,13]. Participants provided information on their digestive symptoms over the prior week and tracked symptoms via the app. Participants were asked to rate 9 common GI symptoms (abdominal pain, bloating, diarrhea, constipation, reflux, gas, nausea, vomiting, and loss of bowel control) on a 5-point scale: 0 (no symptoms), 1 (mild symptoms), 2 (moderate symptoms), 3 (severe symptoms), and 4 (very severe symptoms). Scores for each individual symptom and a composite symptom score (0 to 36, computed as the sum of each of the scores) were determined at baseline (prior to program engagement) and at the time of the last recorded symptom between days 30 and 90 of program participation.

Intervention

All participants used app-based content and were provided access to telehealth, coaching, and nutrition services. Participants who used the telehealth and app services ("telehealth+app") were compared to participants who used the app alone ("app-only," defined as all aspects of the program except the synchronous telehealth visits). At the time of this study, the app was available only in English; however, a real-time translation service was available to support telehealth visits in more than 200 languages. If participants used the app 3 or more times to track symptoms, they were considered frequent symptom trackers.

RDNs provided individualized, evidence-based, and culturally sensitive nutrition guidance [15-17]. Similarly, GI HCs provided individualized support with goal setting, lifestyle management (eg, stress reduction, mindfulness, healthy sleep, and physical activity), medication adherence, self-advocacy, self-monitoring,

XSL•FO
RenderX

coach-led cognitive behavioral therapy, and assistance with the app. The program delivered via the app provides each participant with a personalized care plan that includes targeted education on their conditions, symptoms, and supportive lifestyle interventions. If participants had any telehealth interaction with either an RDN or HC, we classified them as having used telehealth.

Statistical Analysis

We compared means and proportions at baseline for each of the participant characteristics (age, gender, race, BMI, and SVI) using the Student *t* test and chi-square test. We computed the change in the average GI symptom severity score from the baseline score to the latest recorded score for the 2 groups: telehealth+app and app only. The GI symptom severity score ranges from 0 to 36 (maximum 4 points per symptom for 9 GI symptoms).

To evaluate for clinically significant symptom changes during the program, we performed a logistic regression analysis to evaluate the probability that participants at presentation had multiple (2 or more) severe or very severe symptoms and then reported such symptoms as mild or none at program end, controlling for age, gender, race, baseline BMI, days between registration and last check, and baseline total symptom score. Although we did not control for program engagement (eg, symptom tracking), we did control for days between registration and last check, which is a proxy for tracking frequency.

To control for potential confounding factors, the average treatment effect for the treated (ATT) of the severity score change was computed through regression adjustment with the following covariates: age, gender, race, baseline BMI, baseline GI symptom severity scores, and number of days between program registration and the latest recorded score. We performed this analysis using the CAUSALTRT approach in SAS (SAS Enterprise Guide; version 8.4), which estimates the ATT based on the predicted symptom severity score change from generalized linear models fitted for the telehealth+app and the app-only groups separately. *P* values <.05 were considered statistically significant.

Ethical Considerations

Given that all data were deidentified and routinely collected as part of the condition management program, this protocol, based on secondary analyses of existing data, was considered exempt by the WCG institutional review board (VORD.00A; April 2023). Participants were provided access to the digestive health program via their employer benefits but were not otherwise compensated for participation.

Results

Participant Characteristics

Overall, 1656 participants completed the registration process and were enrolled in the study (Table 1). The average age of the participants was 43 (SD 12) years; 75% (n=1239) identified as female, 67% (n=1100) identified as White/Caucasian, 11% (n=187) as Asian/Pacific Islander, 6% (n=105) as Latino/Latina, 7% (n=120) as Black, and 9% (n=144) as being multiple/other races (Table 1). On average, participants were overweight (mean BMI 29, SD 7). The median SVI was 0.4 (SD 0.2). When stratified by median SVI (<0.4 vs >0.4), the high-SVI group (mean SVI 0.6, SD 0.1) was more likely to be female (n=632, 79% vs n=607, 71%) and less likely to be White/Caucasian (n=518, 64% vs n=582, 68%) than the low-SVI group (mean SVI 0.3, SD 0.1).

XSL•FO
RenderX

Table 1. Participant characteristics, stratified by Social Vulnerability Index (SVI)[a].

Characteristic	Total (N=1656)	High SVI (>0.4; n=805, 49%)	Low SVI (<0.4; n=851, 51%)	*P* value
Age (years), mean (SD)	43 (12)	43 (12)	43 (11)	.50
Gender, n (%)				<.001
Female	1239 (75)	632 (79)	607 (71)	
Male	408 (25)	166 (21)	242 (28)	
Prefer not to disclose	9 (1)	7 (1)	2 (0.2)	
Race, n (%)				<.001
African-American/Black	120 (7)	87 (11)	33 (4)	
Asian/Pacific Islander	187 (11)	61 (8)	126 (15)	
Latino/Latina	105 (6)	67 (8)	38 (5)	
Multiple/other	144 (9)	72 (9)	72 (8)	
White/Caucasian	1100 (67)	518 (64)	582 (68)	
SVI, mean (SD)	0.4 (0.2)	0.6 (0.1)	0.3 (0.1)	<.001
BMI (kg/m^2)				<.001
Overall, mean (SD)	29 (7)	29 (7)	28 (7)	
<18.5, n (%)	25 (2)	17 (2)	8 (1)	
≥18.5 to <25, n (%)	539 (33)	224 (28)	315 (37)	
≥25 to <30, n (%)	529 (32)	253 (31)	276 (32)	
≥30, n (%), n (%)	563 (34)	311 (39)	252 (30)	
Change in total symptom score, mean (SD)	–4.5 (4)	–4.7 (5)	–4.4 (4)	.10

[a]The SVI ranges from 0 to 1, with higher scores indicating higher risk for social vulnerability. The median score was 0.4 for this population.

Telehealth Use

The majority of participants (n=1362, 82%) had at least one telehealth visit with a member of their care team and were in the telehealth+app group. Just 18% (n=294) of participants used the app only. Among those participants in the telehealth+app group, 63% (n=859) were frequent symptom trackers. In contrast, among those participants in the app-only group, 43% (n=127) were frequent symptom trackers. There was no difference between the high- and low-SVI groups in the rate of frequent symptom trackers (n=478, 59.4% vs n=508, 59.7%; *P*=.90).

Symptom Improvement

The majority (n=1417, 86%) of participants reported improvement in their total GI symptom score, with an average reduction of 60% in their total symptom score (*P*<.001) relative to baseline. At baseline, the most commonly reported symptoms were gas (n=1545, 93%), bloating (n=1387, 84%), constipation (n=888, 54%), and abdominal pain (n=880, 53%) [5]. On average, participants reported a statistically significant improvement for all symptoms (*P*<.05 for each symptom) [5]. At baseline, 80% (n=1316) of participants reported moderate or severe GI symptom severity for at least one symptom. In contrast, at the end of the intervention, only 47% (n=784) of participants reported moderate or severe symptoms and 16% (n=270) reported no symptoms.

Telehealth+app participants' symptom reduction was 16% (0.7) greater than app-only participants (*P*=.01). High-SVI participants (median score >0.4) showed a 22% greater reduction in total GI symptom score among telehealth+app participants compared to app-only participants (*P*=.04). The clinical impact of telehealth visits in those with high social vulnerability was particularly notable in participants presenting with higher symptomatic acuity. High-SVI participants who were in the telehealth+app group and reported multiple severe or very severe digestive symptoms at baseline were significantly more likely to report no symptoms or only mild symptoms at the end of the program compared to the app-only group (odds ratio [OR] 3.6, 95% CI 1.15-11.26). In contrast, low-SVI participants who were in the telehealth+app group and reported multiple severe or very severe symptoms at baseline were not more likely than those in the app-only group to report symptom resolution or just mild symptoms (OR 0.3, 95% CI 0.09-1.12).

Overall, telehealth+app participants had significantly reduced GI symptom severity scores relative to app-only participants (mean change –0.65, 95% CI –0.13 to –1.17; *P*=.01). The ATT of the high-SVI group was over twice the ATT of the low SVI group (mean change –0.87, 95% CI –0.04 to –1.69; *P*=.04 vs mean change –0.39, 95% CI 0.18 to –0.96; *P*=.18).

XSL•FO
RenderX

Discussion

Principal Findings

This study had two key findings. First, participation in the digital digestive care management program was associated with a reduction in GI symptoms for all groups, including the entire population and those with both low and high SVI. We have previously reported similar findings showing that participants of all genders, races, and SVI strata experienced significant, comparable improvement in GI symptoms [13]. A digital GI care management program, including telehealth visits with RDN and HC support, can be an effective approach for improving debilitating GI symptoms. The strengths of this analysis include, first, the relatively large sample size of nearly 1700 participants followed longitudinally; the richness of the demographic data collected on participants (novel for digital health interventions), including race/ethnicity information and geographic location, which enabled detailed analyses of historically vulnerable populations; and use of the validated SVI to characterize social vulnerability among people with chronic digestive issues.

Second, the additional use of telehealth visits with RDNs and HCs significantly increased symptom reduction for all participants, especially so for participants with higher social vulnerability. This suggests that a multimodal digestive care management program, specifically one incorporating longitudinal synchronous human interaction in addition to digital tools, may be relevant to users with GI symptoms and could have an impact on improving health equity, particularly in those with high social vulnerability. Patients with higher social vulnerability often have poorer access to care, lower health literacy, and worse clinical outcomes; therefore, improving outcomes in this group is an active area of research and policy implementation [18]. Physical access to care is thought to be the most substantial obstacle to health equity, related to low physician supply in rural areas and lack of vehicle access in households of low socioeconomic status. Telehealth visits may bridge the gap by increasing access to culturally competent health care practitioners, as such access is crucial to meeting each individual's unique needs and improving outcomes with the increase in racial and ethnic diversity in the United States [19]. Finally, telehealth synchronous visits allow for regular follow-up and continuity of care, which is crucial for effective treatment, preventive care, psychosocial care, and chronic disease management, potentially reducing health care fragmentation [20,21]. It is plausible that the high-SVI participants in our cohort may have experienced these benefits of telehealth, ultimately contributing to the significant improvement shown with our telehealth program. Given the rise of artificial intelligence–based assistants for virtual care, future studies should evaluate whether the human touch experienced by participants in this study could be as effectively delivered by such novel technology. Notably, more participants in the telehealth+app group were frequent symptom trackers than those in the app-only group (63% vs 43%). Symptom tracking may be a marker for greater engagement in self-care, which could have contributed to the improved outcomes in the telehealth+app group. Prior studies in other clinical disciplines have demonstrated that symptom tracking is strongly associated with higher rates of adequate self-management practices that can improve quality of life [22], prevent clinical deterioration [23], and reduce preventable hospitalizations [24]. However, since there was no difference in the rate of frequent symptom trackers between the high- and low-SVI groups, the observed differences within the telehealth+app group by SVI are unlikely to be associated with symptom tracking frequency.

Future Work

These promising findings warrant additional study to further explore the role of digestive telehealth for vulnerable populations. Since participants self-selected to use telehealth plus the app versus using the app alone, it may have been that participants who used the app only had poorer internet access, digital literacy, patient activation, or other key predictors of health outcomes. Internet access may be a particularly important confounder that was not specifically evaluated in this study (although telehealth visits were available both via video and over the phone). Also, the participants with higher social vulnerability in this study had commercial insurance, and all were at least able to use the app or web browser version of the service; thus, they may have had more advantages than other populations with similar SVI scores. A future evaluation of this digital digestive care program for socially vulnerable populations with other forms of insurance (eg, Medicaid) could expand our understanding of the utility of a digital digestive care management program for low-income groups.

Limitations

Our study has four key limitations. First, given the design of this study, we did not have the ability to control for unmeasured confounding factors such as health motivation and literacy, symptom severity, access to broadband, and language barriers. Given that participants self-selected for the use of virtual care, it limits the ability to infer causality between engagement with the program and outcomes. Second, we did not assess types of additional care that participants may have been receiving, such as external primary care or specialty care. Third, the study population only included commercially insured adults, which may limit the generalizability of these results to other populations. However, our findings may be further amplified in populations at higher social vulnerability risk (eg, Medicaid or Medicare populations). Finally, this study was funded by Cylinder Health, which provided the service evaluated. Notably, the authors adhered to best practices for data analysis and interpretation and report results in accordance with the STROBE (Strengthening the Reporting of Observational studies in Epidemiology) guidelines (Multimedia Appendix 1).

Conclusions

A multimodal digital digestive care program that includes synchronous telehealth visits with RDNs and HCs can improve digestive symptoms in a diverse commercially insured population and may be a promising tool for increasing access and improving clinical outcomes for populations with high social vulnerability who have chronic GI symptoms. Future research should assess the effects of virtual digestive care programs in providing long-term effective care to those affected by health care disparities.

XSL•FO
RenderX

Acknowledgments

This study was funded by Cylinder Health.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Disclaimer

This study was presented at Food & Nutrition Conference & Expo 2024 (Minneapolis, October 2024) in abstract form.

Authors' Contributions

All authors reviewed, revised, and approved the final manuscript. PCH was responsible for study design and analytics. HL was responsible for the overall design of the study and project coordination. DMB was responsible for analytics design and manuscript drafting.

Conflicts of Interest

AB, EC, PCH, and HL are employees of Cylinder Health. SV and DMB are consultants for Cylinder Health.

Multimedia Appendix 1
STROBE checklist.
[PDF File (Adobe PDF File), 116 KB - formative_v9i1e70748_app1.pdf]

References

1. Sandler RS, Stewart WF, Liberman JN, Ricci JA, Zorich NL. Abdominal pain, bloating, and diarrhea in the United States: prevalence and impact. Dig Dis Sci 2000 Jun;45(6):1166-1171. [doi: 10.1023/a:1005554103531] [Medline: 10877233]
2. Peery AF, Crockett SD, Murphy CC, Lund JL, Dellon ES, Williams JL, et al. Burden and cost of gastrointestinal, liver, and pancreatic diseases in the United States: update 2018. Gastroenterology 2019 Jan;156(1):254-272.e11 [FREE Full text] [doi: 10.1053/j.gastro.2018.08.063] [Medline: 30315778]
3. Mathews SC, Izmailyan S, Brito FA, Yamal J, Mikhail O, Revere FL. Prevalence and financial burden of digestive diseases in a commercially insured population. Clin Gastroenterol Hepatol 2022 Jul;20(7):1480-1487.e7 [FREE Full text] [doi: 10.1016/j.cgh.2021.06.047] [Medline: 34217877]
4. Almario C, Ballal M, Chey W, Nordstrom C, Khanna D, Spiegel B. Burden of gastrointestinal symptoms in the United States: results of a nationally representative survey of over 71,000 Americans. Am J Gastroenterol 2018 Nov;113(11):1701-1710 [FREE Full text] [doi: 10.1038/s41395-018-0256-8] [Medline: 30323268]
5. Bravata D, Liu H, Colosimo M, Bullock A, Commons E, Pimentel M. Improvements in digestive symptoms after participation in an app-based chronic digestive disease management program: a prospective cohort evaluation. Cureus 2024 Aug;16(8):e66941. [doi: 10.7759/cureus.66941] [Medline: 39280538]
6. Sasegbon A, Vasant DH. Understanding racial disparities in the care of patients with irritable bowel syndrome: the need for a unified approach. Neurogastroenterol Motil 2021 May;33(5):e14152. [doi: 10.1111/nmo.14152] [Medline: 33835634]
7. Huang RJ, Shah SC, Camargo MC, Palaniappan L, Hwang JH. County rurality and socioeconomic deprivation is associated with reduced survival from gastric cancer in the United States. Gastroenterology 2020 Oct;159(4):1555-1557.e2 [FREE Full text] [doi: 10.1053/j.gastro.2020.05.006] [Medline: 32387539]
8. Ahuja A, Pelton M, Raval S, Kesavarapu K. Role of nutrition in gastroesophageal reflux, irritable bowel syndrome, celiac disease, and inflammatory bowel disease. Gastro Hep Adv 2023;2(6):860-872 [FREE Full text] [doi: 10.1016/j.gastha.2023.06.010] [Medline: 39130122]
9. Lansdorp-Vogelaar I, Kuntz KM, Knudsen AB, van Ballegooijen M, Zauber AG, Jemal A. Contribution of screening and survival differences to racial disparities in colorectal cancer rates. Cancer Epidemiol Biomarkers Prev 2012 May;21(5):728-736 [FREE Full text] [doi: 10.1158/1055-9965.EPI-12-0023] [Medline: 22514249]
10. Demb J, Gupta S. Racial and ethnic disparities in colorectal cancer screening pose persistent challenges to health equity. Clin Gastroenterol Hepatol 2020 Jul;18(8):1691-1693 [FREE Full text] [doi: 10.1016/j.cgh.2019.11.042] [Medline: 31786329]
11. Spaulding EM, Fang M, Commodore-Mensah Y, Himmelfarb CR, Martin SS, Coresh J. Prevalence and disparities in telehealth use among US adults following the COVID-19 pandemic: national cross-sectional survey. J Med Internet Res 2024 May 10;26:e52124 [FREE Full text] [doi: 10.2196/52124] [Medline: 38728070]

XSL•FO
RenderX

12. Bhagavathula AS, Aldhaleei WA. Bridging the telehealth divide: racial and ethnic disparities in Medicare telehealth usage highlights the need for equity. Telemed J E Health 2024 May;30(5):1272-1278. [doi: 10.1089/tmj.2023.0536] [Medline: 38215268]
13. Bravata D, Liu H, Colosimo MM, Bullock AC, Commons E, Pimentel M. Digital disease management programme reduces chronic gastrointestinal symptoms among racially and socially vulnerable populations. BMJ Open Gastroenterol 2024 Aug 28;11(1):e001463 [FREE Full text] [doi: 10.1136/bmjgast-2024-001463] [Medline: 39209334]
14. Social Vulnerability Index. Centers for Disease Control and Prevention, Agency for Toxic Substances and Disease Registry. URL: https://www.atsdr.cdc.gov/place-health/php/svi/?CDC_AAref_Val=https://www.atsdr.cdc.gov/placeandhealth/svi/index.html [accessed 2023-11-01]
15. Singh R, Salem A, Nanavati J, Mullin GE. The role of diet in the treatment of irritable bowel syndrome: a systematic review. Gastroenterol Clin North Am 2018 Mar;47(1):107-137. [doi: 10.1016/j.gtc.2017.10.003] [Medline: 29413008]
16. Fitzpatrick JA, Melton SL, Yao CK, Gibson PR, Halmos EP. Dietary management of adults with IBD - the emerging role of dietary therapy. Nat Rev Gastroenterol Hepatol 2022 Oct;19(10):652-669. [doi: 10.1038/s41575-022-00619-5] [Medline: 35577903]
17. Reznikov EA, Suskind DL. Current nutritional therapies in inflammatory bowel disease: improving clinical remission rates and sustainability of long-term dietary therapies. Nutrients 2023 Jan 28;15(3):668 [FREE Full text] [doi: 10.3390/nu15030668] [Medline: 36771373]
18. Schillinger D. Social determinants, health literacy, and disparities: intersections and controversies. Health Lit Res Pract 2021 Jul;5(3):e234-e243 [FREE Full text] [doi: 10.3928/24748307-20210712-01] [Medline: 34379549]
19. Hilty DM, Gentry MT, McKean AJ, Cowan KE, Lim RF, Lu FG. Telehealth for rural diverse populations: telebehavioral and cultural competencies, clinical outcomes and administrative approaches. Mhealth 2020;6:20 [FREE Full text] [doi: 10.21037/mhealth.2019.10.04] [Medline: 32270012]
20. Alazri M, Heywood P, Neal RD, Leese B. Continuity of care: literature review and implications. Sultan Qaboos Univ Med J 2007 Dec;7(3):197-206 [FREE Full text] [Medline: 21748104]
21. Amagai S, Vonesh E, Adams J, Luo Y. Closing the gap: addressing telehealth disparities across specialties in the sustained pandemic era. NPJ Digit Med 2024 Aug 21;7(1):217 [FREE Full text] [doi: 10.1038/s41746-024-01201-w] [Medline: 39164391]
22. Lizán L, Pérez-Carbonell L, Comellas M. Additional value of patient-reported symptom monitoring in cancer care: a systematic review of the literature. Cancers (Basel) 2021 Sep 15;13(18):4615 [FREE Full text] [doi: 10.3390/cancers13184615] [Medline: 34572842]
23. Lee K, Lennie T, Dunbar S, Pressler S, Heo S, Song E, et al. The association between regular symptom monitoring and self-care management in patients with heart failure. J Cardiovasc Nurs 2015;30(2):145-151 [FREE Full text] [doi: 10.1097/JCN.0000000000000128] [Medline: 24434828]
24. Riegel B, De Maria M, Barbaranelli C, Matarese M, Ausili D, Stromberg A, et al. Symptom recognition as a mediator in the self-care of chronic illness. Front Public Health 2022;10:883299 [FREE Full text] [doi: 10.3389/fpubh.2022.883299] [Medline: 35655456]

Abbreviations

ATT: average treatment effect for the treated
GI: gastrointestinal
HC: health coach
OR: odds ratio
RDN: registered dietitian nutritionist
STROBE: Strengthening the Reporting of Observational studies in Epidemiology
SVI: Social Vulnerability Index

Edited by A Mavragani; submitted 02.01.25; peer-reviewed by S Kumar, G Downer, T Liu; comments to author 21.04.25; revised version received 12.07.25; accepted 24.07.25; published 11.09.25.

Please cite as:
Varma S, Black A, Commons E, Ho PC, Bravata DM, Liu H
The Impact of a Digital Digestive Health Program and Telehealth Visits in Socially Vulnerable Populations: Cohort Evaluation
JMIR Form Res 2025;9:e70748
URL: https://formative.jmir.org/2025/1/e70748
doi:10.2196/70748
PMID:

XSL•FO
RenderX

©Sanskriti Varma, Alicen Black, Erin Commons, Pen-Che Ho, Dena M Bravata, Hau Liu. Originally published in JMIR Formative Research (https://formative.jmir.org), 11.09.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Youth-Centered Digital Infographic on Vaping Risks (What's in a Vape?): Mixed Methods Study

Kendra Nelson Ferguson[1*], PhD; Kyla Christianson[2*], BSc; Adrian Delgado[2*], BSc; Gina Martin[3*], PhD; Stephanie E Coen[4*], PhD; Laura Struik[2*], PhD

[1]Western University, London, ON, Canada
[2]University of British Columbia Okanagan, Kelowna, BC, Canada
[3]Athabasca University, Alberta, Canada
[4]University of Nottingham, Nottingham, United Kingdom
[*]all authors contributed equally

Corresponding Author:
Laura Struik, PhD
University of British Columbia Okanagan
1147 Research Rd
Kelowna, BC, V1V1V7
Canada
Email: laura.struik@ubc.ca

Abstract

Background: As youth engagement with traditional public health warnings declines, innovative strategies are needed. Visually compelling, youth-driven digital content such as interactive infographics may help bridge knowledge gaps, enhance risk perception, and support more informed decision-making. Despite this potential, limited research has assessed its effectiveness in conveying vaping-related harms to youth.

Objective: To address this gap, this study evaluated the impact of a codeveloped, youth-informed digital infographic (*What's in a Vape?*) on enhancing vaping education and improving youth understanding of vaping-related harms.

Methods: A convergent parallel mixed methods design was used to assess the impact of a youth-informed digital infographic. The infographic was created in collaboration with youth coresearchers and youth advisory councils to ensure relevance. Participants were recruited through community partners, school boards, and youth networks. By May 2024, we had enrolled 63 high school students aged 14 to 19 years (mean age 16.5, SD 1.2 years) primarily from Ontario and British Columbia. The survey evaluated baseline knowledge of vaping, engagement with the infographic, and postexposure perceptions on whether the content contributed to increased awareness or understanding of vaping.

Results: Data collection took place between April 2024 and May 2024. Quantitative analysis showed that 87% (55/63) of participants agreed that the infographic effectively communicated key information, and 86% (54/63) gained new knowledge about vaping. In addition, 73% (46/63) found that the infographic was presented in an easy and meaningful way, whereas 52% (33/63) indicated that they would definitely share it with others, reflecting strong engagement. However, over half (33/63, 52%) also found the amount of information excessive, and 17% (11/63) found it difficult to digest, indicating variation in youth information preferences. Thematic analysis of qualitative feedback revealed four key themes: (1) the visual content enabled gaining new insights into and knowledge of vaping, (2) the visual design had a positive impact on engagement with information, (3) sourced information enhanced the credibility of the infographic information, and (4) the digital design of the infographic made complex information more understandable. Qualitative insights contextualized and supported the quantitative findings, highlighting both benefits and areas for improvement.

Conclusions: This study demonstrates that youth-driven digital infographics may serve as useful health communication tools. Findings highlight the importance of peer-led design; evidence-based content; and interactive, visually compelling formats in enhancing youth comprehension and receptiveness to health messaging. By integrating youth feedback into development and prioritizing digital engagement, the infographic bridged knowledge gaps while reinforcing the credibility of its content. Variability in feedback about content overload suggests that future versions should consider more layered or modular designs. Results suggest that such approaches may complement broader public health strategies to curb youth vaping and inform future educational interventions. Continued research is warranted to assess long-term impacts on attitudes and behavior.

XSL•FO
RenderX

(JMIR Form Res 2025;9:e75694) doi:10.2196/75694

KEYWORDS

youth vaping; youth-centered digital infographic; health communication; mixed methods

Introduction

Background

Youth vaping remains a critical public health concern in Canada, with surveillance data showing a consistent rise since 2017 [1]. Although overall prevalence rates have remained relatively stable in recent years, the frequency of use has risen, signaling a shift toward habitual use and potential nicotine dependence among youth [2]. Vaping among youth has been linked to a range of well-documented health risks, including neurological, respiratory, and cardiovascular effects [2-6]. It is also associated with increased likelihood of substance use (eg, cannabis, cocaine, and heroin) [7-10] and higher rates of mood disorders such as anxiety and depression [7]. Despite these harms, many youths remain unaware of the risks, particularly those related to device design, heavy metal exposure, and addiction potential [11-14].

This gap in awareness emphasizes the need for clearer and more effective communication strategies. Rohde et al [15] found that youths recognized risks such as respiratory issues and device explosions but were less aware of the harms related to skin contact with e-liquids, exposure to diacetyl, cardiovascular damage, carcinogenic compounds, and addiction. Furthermore, Greiner Safi et al [16] found that uncertain language in youth-targeted public health messaging can lead to confusion or the dismissal of the information altogether, reinforcing the importance of clear and accessible messaging.

To improve message clarity and engagement, public health campaigns must prioritize effective, youth-aligned communication strategies. Visual aids such as infographics have been shown to simplify complex health information, reduce literacy barriers, and support knowledge retention [17], particularly when paired with direct messaging that can influence health behaviors [17]. Nonetheless, public health messaging has been slow to adopt these approaches, with many campaigns overlooking interactive digital tools despite their relevance to youth [18,19]. Studies by Rohde et al [15], England et al [20], and Stalgaitis et al [19] highlight the value of digital platforms in expanding the reach and relevance of health messaging.

Beyond format, actively involving youth in campaign development is increasingly recognized as essential for creating messages that resonate. Incorporating youth perspectives enhances the relevance, trust, and effectiveness of health communication [16]. For example, the Rethink Vape campaign [20] exemplifies the benefits of this approach, demonstrating that youth-centered campaigns lead to greater message retention and impact. Participants who engaged with the campaign demonstrated greater comprehension of vaping-related information, heightened risk perception, and stronger intentions to avoid vaping. As youth engagement with traditional public health warnings declines, incorporating youth-driven, visually compelling content with interactive digital strategies may be critical in addressing high vaping rates [16].

Given these findings, digital infographics codeveloped with youth offer a promising but underexplored avenue for bridging knowledge gaps, enhancing perceived relevance, and promoting informed decision-making. While large-scale campaigns such as The Real Cost [18] have demonstrated effectiveness, these efforts were primarily adult led and evaluated using broad exposure metrics rather than participatory design approaches. Even campaigns with participatory designs, such as Rethink Vape, focused on short-term engagement and lacked formal evaluation of individual infographic components [20].

Objectives

In contrast, this study assesses the effectiveness of a youth-informed, codeveloped digital infographic (ie, *What's in a Vape?*) created through an iterative design process with youth coresearchers. By delivering the infographic via an interactive web interface and assessing it using a convergent mixed methods approach, this study uniquely captured both perceived impact and experiential feedback. By integrating quantitative and qualitative findings, we offer a nuanced, youth-centered perspective on the role of digital infographics in vaping education.

Methods

Infographic Development

This project began with a poll conducted among 3 youth advisory councils, each composed of high school–aged youth (n=2, 67% in British Columbia and n=1, 33% in Ontario) to gather their questions about e-cigarettes and how they wanted them answered. *What's in a vape?* emerged as the most pressing unanswered question, and all councils advocated for an interactive, digital infographic to address it. The development of the infographic was led by a youth coresearcher (AD), an undergraduate student in the School of Nursing at the University of British Columbia Okanagan, who began by conducting a literature review on current evidence related to vape components and their health effects. Following this, AD engaged in collaborative sessions with the senior author (LS) to organize the findings into a mind map, visually structuring key information before transitioning to an infographic. Mind mapping served as an initial step to systematically categorize vaping-related risks, ensuring a logical and comprehensive infographic design. To transition this information to a digital, web-based infographic, AD collaborated with a design thinking researcher, who used Figma (Figma, Inc), a collaborative interface design tool, to incorporate graphic design and interactive digital strategies (eg, clickable links and illustrations), ultimately creating the final product. The infographic was hosted as a web-based interactive resource accessible via a unique hyperlink. This web-based format was chosen for its ease of access, scalability, and compatibility across devices commonly

XSL•FO
RenderX

used by young people (eg, smartphones, tablets, and laptops). Once the initial draft was completed, feedback was gathered from 2 high school youth coresearchers. Their input led to multiple refinements. A follow-up poll of the youth advisory councils was conducted to collect additional thoughts, opinions, and suggestions for improvement. After final revisions, the infographic was ready for fielding (Multimedia Appendix 1).

Study Design

This study used a mixed methods approach. Specifically, a convergent parallel design was used whereby both quantitative and qualitative data were collected simultaneously, analyzed as separate but complementary datasets, and converged for interpretation in the Discussion section [21]. Following independent analyses, the research team compared findings across both data types to identify areas of convergence and divergence. For example, descriptive statistics (eg, perceived knowledge gain and clarity of presentation) were examined alongside qualitative themes (eg, what knowledge was gained and what aspects supported clarity) to deepen interpretation. This integrative process allowed for a more nuanced understanding of how youth experienced and responded to the infographic.

A cross-sectional web-based survey, the Youth Vaping Information Project (VIP), was developed and distributed to assess perceptions of the interactive digital infographic—*What's in a Vape?*—which was embedded within the survey. This study was reported according to the Checklist for Reporting Results of Internet E-Surveys [22].

Ethical Considerations

Ethics approval for this study was obtained in February 2024 from the Human Research Ethics Boards at both the University of British Columbia (H23-03957) and Western University (124387). Upon completion of the survey, participants were emailed a CAD $10 (US $7.26) gift card as a thank you for their time.

Recruitment and Participants

The research team disseminated the Youth VIP recruitment poster to community partners (eg, health units and school boards), youth advisory councils, and youth coresearchers via email, requesting their support in circulating the poster to their networks. The poster included a QR code and web link directing potential participants to the Youth VIP survey on Qualtrics (Qualtrics International Inc). The survey began with a letter of information outlining study details and emphasizing that participation was voluntary and that participants could opt out at any time. This was followed by an eligibility questionnaire and assent or consent questions. Eligibility criteria were being a resident of Canada, currently enrolled in high school (grades 9-12), and fluent in English. Youth aged <16 years were required to obtain parental consent in addition to their own assent.

Recruitment followed a purposive, self-selection approach, primarily reaching youth with existing connections to school- or community-based programming. This strategy was intentionally chosen to engage youth who are more likely to interact with health messaging and digital platforms, aligning with the infographic's participatory development process and web-based delivery format. Given the focus on youth-driven content, it was important to include youth who could provide informed, critical feedback on the design and content of the infographic. However, we acknowledge that this approach may have introduced sampling bias as it likely excluded less engaged youth or those not involved in structured programming. As such, the sample may not be representative of all youth populations.

While the sample included participants from Ontario and British Columbia, data on school type (eg, public or private), rural or urban setting, or other contextual characteristics were not collected. Following demographic and behavioral questions, participants were shown the *What's in a Vape?* infographic, which was embedded in the Qualtrics survey platform. They were encouraged to take their time in reviewing the infographic to ensure that they could accurately respond to the subsequent questions. The survey included a hyperlink that directed participants to view the interactive version of the infographic on a separate web page before returning to complete the survey questions. This version included clickable sections and embedded visuals created using Figma.

The final sample consisted of 63 participants (n=26, 41% identified as boys, n=35, 56% identified as girls, and n=2, 3% preferred not to disclose) aged 14 to 19 years (mean age 16.5, SD 1.2 years). The sample was geographically concentrated in Ontario (n=34, 54%) and British Columbia (n=29, 46%), and 44% (28/63) of the participants identified as White.

Data Collection

The survey was developed by a multidisciplinary team (KNF, AD, and LS) specializing in health promotion, health behavior, health geography, and nursing. Questions were informed by knowledge translation principles [23] to ensure an accurate evaluation of the infographic. After drafting the survey, 2 high school youth coresearchers reviewed and refined the questions for youth relevance and project alignment. Following minor adjustments, members of 3 youth advisory councils piloted the survey for readability, functionality, and usability. The survey was finalized after further discussions and feedback from the councils. Data were collected between April 2024 and May 2024. Participants who provided consent or assent and parental consent (if aged <16 years) completed the survey, which took 20 to 30 minutes.

Survey

Measures

The survey questions captured participant demographics (ie, age, sex, gender, ethnicity, and socioeconomic status), and participants responded to a series of closed-ended questions assessing vaping behaviors and exposure to vape education. Questions related to vape use included the following—"Have you ever used a vape?" "Do you currently vape?" "Have you attempted to quit vaping?"—with *yes* or *no* response options. The latter 2 questions—"Do you currently vape?" and "Have you attempted to quit vaping?"—were only presented to participants who answered *yes* to "Have you ever used a vape?" Exposure to vape education was assessed using the following questions—"Have you attended any vape education

XSL•FO
RenderX

presentations at school?" and "Have you noticed any educational material related to vaping at school (eg, posters in the hall)?"—also using *yes* or *no* response options. After viewing the infographic, participants provided feedback in four areas: (1) intentions to share the infographic ("Will you share knowledge from this infographic with your family/friends?"; response range: *definitely not* to *definitely yes*), (2) effectiveness in addressing knowledge gaps ("The infographic tells me what I need to know about vapes"; response range: *strongly disagree* to *strongly agree*), (3) clarity and presentation ("The infographic is presented in an easily digestible and meaningful way"; response range: *strongly disagree* to *strongly agree*), and (4) information sufficiency (response options: *too much*, *the right amount*, or *not enough*).

To gain deeper insights into participants' perceptions and experiences with the infographic, the survey included 4 open-ended questions designed to elicit qualitative feedback. These questions were included after participants viewed the infographic and explored whether the content contributed to increased awareness or understanding of vaping as well as the overall appeal of the infographic, while also prompting participants to reflect on its design and content. Specifically, participants were asked the following: (1) "Does the infographic provide new knowledge? If yes, what did you learn?" (2) "What I liked about the infographic was..." (3) "What I did not like about the infographic was..." (4) "What would you change about the infographic?"

Quantitative Analysis

Descriptive statistics (frequencies and proportions) were calculated using Apple Numbers (version 3.6.2; Apple Inc), a spreadsheet application for Macintosh computers, which was used solely for basic data summarizations given the exploratory nature of the analysis. No inferential statistical tests were conducted.

Qualitative Analysis

Thematic analysis was conducted to explore the open-ended survey responses. Using an inductive approach, 2 researchers (AD and LS) independently reviewed and grouped responses into themes based on recurring ideas [24]. Any irrelevant responses (eg, *N/A*, *I don't know*, or *none*) were removed. Following this, AD and LS acted as "critical friends" [25] to challenge and refine theme interpretations. Key discussions led to revisions of theme names to ensure alignment with participant perspectives. For example, the theme *visual content enabled gaining new insights and knowledge into vaping* was initially titled *gaining new knowledge on vaping*, but after discussions, it was revised to further reflect the nuances of youth experiences. Once consensus was reached, the first author (KNF) reviewed the themes for plausibility and consistency, strengthening the study's rigor and trustworthiness.

Results

Overview

A narrative contiguous approach was implemented for reporting and integrating the mixed methods results. In this approach, quantitative findings are presented first, followed by qualitative findings, with each set of results outlined in different subsections [26]. Both datasets are then synthesized to inform the overall discussion.

Quantitative Results

Table 1 provides the demographic and behavioral characteristics of the sample (N=63). Of the participants, 46% (28/61) identified as male, and 54% (33/61) identified as female, and by gender, 43% (26/61) identified as boys, and 57% (35/61) identified as girls. Ethnic representation was predominantly White (28/61, 46%), followed by East Asian (15/61, 25%). Socioeconomic status varied, with 21% (13/63) never worrying about money, 44% (28/63) worrying only about extras, and 35% (22/63) having just enough or not enough to get by. Most participants rated their physical and mental health as very good (23/63, 37% and 24/63, 38%, respectively) or good (21/63, 33% and 19/63, 30%, respectively). Among participants who reported using a vape, 41% (26/63) indicated that they currently vaped. Of those current vapers, 65% (17/26) reported having attempted to quit. Social exposure to vaping was high, with 35% (22/63) indicating that many, most, or all of their friends vaped; 44% (28/63) reporting that some vaped; and 21% (13/63) reporting that none of their friends vaped. The most common exposure settings included school (44/63, 70%), peers or friends (29/63, 46%), and online or social media (26/63, 41%), with less frequent exposure at home and the workplace (4/63, 6% each). Just over half of the participants had attended vape education presentations at school (34/63, 54%) or had noticed vape education materials at school (33/63, 52%).

XSL•FO
RenderX

Table 1. Characteristics of the sample (N=63).

Variable	Values
Age (y), mean (SD; range)	16.5 (1.2; 14-19)
Sex, n/N (%)	
Male	28/61 (46)
Female	33/61 (54)
Missing	2/63 (3)
Gender, n/N (%)	
Boy	26/61 (43)
Girl	35/61 (57)
Missing	2/63 (3)
Race or ethnicity, n/N (%)	
Black	3/61 (5)
East Asian	15/61 (25)
Indigenous	7/61 (11)
South Asian	2/61 (3)
Southeast Asian	4/61 (7)
West Asian or Middle Eastern	2/61 (3)
White	28/61 (46)
Missing	2/63 (3)
Socioeconomic status, n/N (%)	
Never worry about money	13/63 (21)
Only worry about money for extras	28/63 (44)
Just enough or not enough money to get by	22/63 (35)
Physical health, n/N (%)	
Excellent	12/63 (19)
Very good	23/63 (37)
Good	21/63 (33)
Fair or poor	7/63 (11)
Mental health, n/N (%)	
Excellent	8/63 (13)
Very good	24/63 (38)
Good	19/63 (30)
Fair or poor	12/63 (19)
Vaping experience, n/N (%)	
Currently vape	26/63 (41)
Tried to quit vaping	17/26 (65)
Friends who vape, n/N (%)	
Many, most, or all	22/63 (35)
Some	28/63 (44)
None	13/63 (21)
Setting in which they were most exposed to vaping (participants could select all that applied), n/N (%)	
At school	44/63 (70)
Online (social media)	26/63 (41)

XSL•FO
RenderX

Variable	Values
With peers or friends	29/63 (46)
At home	4/63 (6)
At the workplace	4/63 (6)
Activities	11/63 (17)
Attended vape education presentations at school, n/N (%)	
Yes	34/63 (54)
No	29/63 (46)
Noticed vape education materials at school, n/N (%)	
Yes	33/63 (52)
No	30/63 (48)

Table 2 presents closed-ended responses to the infographic. Most participants (55/63, 87%) agreed or strongly agreed that the infographic effectively conveyed the necessary information. Regarding presentation, 73% (46/63) of the participants agreed or strongly agreed that the infographic was presented in an easy and meaningful way, whereas 17% (11/63) expressed disagreement or strong disagreement. When asked about the amount of information provided, 52% (33/63) felt that there was too much information, whereas 46% (29/63) believed that the amount was appropriate. A substantial proportion of respondents (54/63, 86%) reported gaining new knowledge from the infographic. In terms of sharing the infographic with family and friends, 52% (33/63) indicated that they would definitely do so.

Table 2. Closed-ended infographic-related responses (N=63).

Infographic-related questions	Participants, n (%)
"It tells me what I need to know"	
Strongly agree or agree	55 (87)
Neutral	4 (6)
Disagree or strongly disagree	4 (6)
"It is presented in an easy and meaningful way"	
Strongly agree or agree	46 (73)
Neutral	6 (10)
Disagree or strongly disagree	11 (17)
Amount of information	
Too much	33 (52)
The right amount	29 (46)
Not enough	1 (2)
Gained new knowledge	
Yes	54 (86)
No	9 (14)
Likely to share it with family and friends	
Definitely	33 (52)
Maybe	23 (37)
Probably or definitely not	7 (11)

Qualitative Results

Overview

Thematic analysis of open-ended survey responses revealed four main themes: (1) the visual content enabled gaining new insights into and knowledge of vaping, (2) the visual design had a positive impact on engagement with information, (3) sourced information enhanced the credibility of the infographic information, and (4) the digital design of the infographic made complex information more understandable. While open-ended survey responses were initially disaggregated by vaping status (ever vapers or never vapers), they were ultimately merged due

XSL•FO
RenderX

to substantial thematic overlap between the groups. Themes and illustrative quotes were consistent across both user types, suggesting similar patterns of infographic engagement and perceived knowledge gain. Each theme is explored through participant quotes from both ever and never vapers.

The Visual Content Enabled Gaining New Insights Into and Knowledge of Vaping

As reflected in the quantitative closed-ended questions, most of the participants reported gaining new knowledge from the infographic. Many participants noted that they had previously focused on the substances inside vapes rather than the materials used to construct them. One participant explained the following:

> *I did not know about the materials of a vape; we mostly learn about the substances inside rather than the carrier.* [Never vaper]

Others found the infographic useful in clarifying how e-cigarettes function, with one participant stating that it "provides an explanation for how e-cigarettes work" (ever vaper). Beyond understanding the mechanics of vaping devices, participants also reported learning about the overall health risks associated with vaping. Several noted that they were unaware of how vaping affects the respiratory system and academic performance, and how metals from vapes accumulate in vital organs such as the brain, liver, and kidneys. This is highlighted in the following quote—"[I learned that] inhaling chemicals and nicotine can damage lung function, leading to breathing difficulties" (ever vaper)—whereas another participant mentioned learning about how "aluminum particles are collected in the brain, liver, and kidneys" (never vaper). Along these lines, participants also reported increased awareness of the chemicals and metals in vape liquids and aerosols. One ever vaper put it as follows:

> *I learned that e-cigarette liquids contain nicotine, flavouring agents, and chemicals harmful to our health.*

Another participant echoed this sentiment:

> *I learned about the different harmful metals and substances found in vapes, which I didn't know of before.* [Never vaper]

It is apparent that the infographic provided valuable insights into both the components of vapes and the associated health risks for both ever and never vapers alike.

Positive Impact of Visual Design on Engagement With Information

Participants expressed an appreciation for the visual elements of the infographic, highlighting how the design effectively communicated essential information. The vibrant visuals helped capture attention, with one participant mentioning that "the infographic's vibrant colors and engaging design immediately caught [my] attention, making the information more accessible and enjoyable to explore" (never vaper). Similarly, another participant added that "the visual aspect made it appealing and more inviting to read" (ever vaper).

In addition to its appeal, participants found that the infographic's structure made the content easier to personalize and more accessible. One participant explained that it "allows users to customize the display of infographics to their personal preferences" (never vaper), whereas another participant said that "it was scientific, and had a diagram which helped me understand; it was also well organized" (ever vaper). The responses demonstrate that the visual aspects of the infographic were perceived to enhance both the accessibility of and engagement with the information.

Sourced Information Enhanced the Credibility of the Infographic Information

It was evident that youth participants valued the credibility and transparency of the information presented, particularly regarding the associated risks of vaping and the components of vapes. Pointing to the effective communication of health risks, one participant shared that "all the info about the harms is impactful and helps you to realize the effects of vaping on your body" (ever vaper).

Similarly, a never vaper commented on learning about "all the ingredients in e-cigarettes, all the damage they cause to the human body, the names of the ingredients and what they do." Another participant added the following:

> *I liked how the infographic contained information about the parts of a vape and also the harmful effects of vaping.* [Ever vaper]

Participants also discussed the importance of the infographic's transparency when it comes to the sources used and provided reference materials. This is explained in the following quote:

> *I like the source and reference materials of this information diagram the most so that I can learn more about electronic cigarettes.* [Ever vaper]

This was complemented by another participant, who highlighted the inclusion of data and statistics:

> *The inclusion of statistics and data adds credibility to the information presented in the infographic.* [Never vaper]

The Digital Design of the Infographic Made Complex Information More Understandable

Most participants recognized that the infographic effectively broke down complex information into digestible and understandable content. One participant shared the following:

> *[It was] easy to understand even though there was lots of new scientific information on how a vape works.* [Ever vaper]

Another participant echoed this sentiment by saying the following:

> *[It made] complex information easier to understand through data visualization.* [Never vaper]

The clear labeling and detailed presentation were helpful, as highlighted in the following quote:

XSL•FO
RenderX

> *I liked how it pointed to specific parts of the vape and broke it down. I liked how it was clearly labelled as well.* [Never vaper]

In addition, the infographic's visuals, such as diagrams and illustrations, were particularly appreciated for their ability to simplify the information presented. One participant shared the following—"the amount of diagrams provided to illustrate the main point of the texts [was helpful]" (ever vaper)—and another added the following—"I like that it uses charts and diagrams to display data, making it easier to understand and remember" (never vaper). These visual elements, combined with detailed information, made the content both accessible and engaging, helping participants understand the different aspects of vaping in a clear and concise manner.

Discussion

Principal Findings

This study explored whether a codeveloped, youth-informed digital infographic can enhance vaping education and strengthen understanding of vaping-related harms among youth. By integrating quantitative and qualitative data, this study provided comprehensive insight into how youth interact with and interpret vaping-related educational materials. To our knowledge, this is one of the first studies to examine the impact of a youth-created digital infographic on youth vaping, offering insights into how peer-designed digital materials can influence young people's knowledge and engagement with vaping-related education. Although the findings were presented using a narrative contiguous approach, deeper integration of the quantitative and qualitative data reveals a strong convergence between youth-reported outcomes and the content of their open-ended responses. Quantitative results established the prevalence of knowledge gain and infographic engagement, and qualitative findings offered illustrative context and nuance to these patterns.

The infographic's youth-centered and youth-created design may have enhanced its relevance and appeal, supporting previous research findings that peer-led interventions can be effective tools for disseminating vaping-related information [16,27]. The iterative process with youth allowed for the refinement of infographic concepts to better align with the needs and preferences of a youth audience, reflected in our qualitative findings related to visual content and digital design. As noted by England et al [20], concepts favored by adults are often dismissed by teenagers, emphasizing the need to prioritize youth input in health communication. Given that youth are experts in their own lives, they are well positioned to identify the information most relevant to them [20]. Our findings support the importance of incorporating youth voices and perspectives when developing public health education materials, thereby increasing their potential effectiveness at translating health information.

Findings indicate that participants appreciated the infographic's ability to simplify complex information, a sentiment echoed by 86% (54/63) of the participants, who reported gaining new knowledge after viewing the infographic. Qualitative responses substantiated this self-reported knowledge gain, with participants noting that the infographic effectively communicated technical facts by breaking down intricate details about vaping risks into digestible and understandable content without overwhelming the audience. Unlike traditional approaches that often oversimplify health information, resulting in vague and unhelpful messaging that may fail to engage or inform youth effectively [16,27], the infographic provided nuanced, evidence-based details in a structured and visually engaging format. This approach acknowledges that young audiences are capable of processing detailed health information when presented in a manner that does not sacrifice depth and accuracy. Rather than reducing content to simplistic messages, the digital format enabled a layered presentation of information, allowing youth to connect with key points. Research has shown that digital platforms, particularly those using interactive and visual elements, enhance youth engagement and comprehension by allowing them to explore content at their own pace [17,19]. Our findings reinforce existing research stating that youth are more responsive to health messaging delivered through digital media as such platforms align with their media consumption habits [18,19].

However, it is important to consider the interpretive limitations of these results. While participants' responses suggest strong engagement and perceived knowledge gain, these findings are based on self-reported impressions rather than validated knowledge or behavioral outcomes. This limits the extent to which we can infer actual understanding or educational impact. As such, the results should be interpreted as exploratory, offering insight into youth engagement rather than conclusive evidence of learning or behavior change.

The ability of the infographic to translate complicated health information also aligns with research demonstrating that visual and interactive educational tools improve knowledge, particularly for complex topics [17]. Participants stated that they valued the clarity and organization of the information, which is consistent with existing evidence that visually compelling materials enhance comprehension and accessibility [17,28,29]. This aligns with the 73% (46/63) of participants who agreed or strongly agreed that the infographic was presented in an easy and meaningful way, reinforcing that the visual design supported understanding across the sample. Specific design elements such as clear diagrams, attractive colors, and structured layouts were praised for making the content easier to digest, echoing studies that highlight the role of visual aids in improving the retention and clarity of health messages [17,29]. In addition, immersive features such as clickable links and illustrations further enhanced interest, demonstrating the potential of digital platforms to appeal to youth.

While most participants found the infographic engaging and informative, it is important to acknowledge that 52% (33/63) reported the amount of information to be excessive and 17% (11/63) did not find the content easily digestible. These findings suggest that, despite the overall positive sentiment regarding the infographic, some youths may have experienced cognitive overload or difficulty processing the information. This highlights the importance of striking a balance between depth and simplicity when designing digital health tools for young people. Future iterations may benefit from modular formats or

XSL•FO
RenderX

interactive layering that allows users to control how much information they engage with at one time, aligning better with diverse learning preferences and capacities.

Beyond design-related considerations, participants also valued the infographic's credibility and transparency, particularly the inclusion of data and references related to the health effects of vaping. This emphasizes the need for evidence-based, transparent messaging in educational resources as young audiences are often skeptical of health information that lacks clear scientific backing [16,27]. Due to the absence of clear public health messaging on the harms of e-cigarette use, youth are often exposed to mixed messages about vaping risks [20,30-32]. A reduced perception of risk, combined with a general underestimation of nicotine's addictive properties, can increase youth's susceptibility to vape marketing strategies [30], contributing to vaping initiation [15]. Thus, the inclusion of scientific sources strengthened the infographic's credibility among youth and may encourage them to seek out evidence-based materials when evaluating other forms of vaping information.

Strengths and Limitations

This study has several strengths. First, it contributes to the limited research on youth-centered vaping education by evaluating a digital infographic codeveloped with youth. By incorporating youth perspectives into its design, this study ensured that the educational material was relevant and engaging for the target audience (ie, youth). Second, using a mixed methods approach allowed for a more comprehensive understanding of how youth interact with and interpret vaping-related information. Including both ever vapers and never vapers provided valuable insight into differences in knowledge, perceptions, and responses to educational materials. Finally, the combination of both quantitative and qualitative responses provided a nuanced understanding of how youth interpret and perceive vaping-related messaging.

Despite these strengths, several limitations constrain the interpretive weight of our findings. The sample was geographically limited to Ontario and British Columbia and, due to the small sample size, is not representative of youth in those regions or reflective of the broader diversity of perspectives across Canada. Although efforts were made to include a range of participants, the demographic distribution does not capture the breadth of lived experiences across socioeconomic, cultural, and regional contexts. Contextual details such as school type (eg, public, private, or alternative), rural versus urban location, and other educational or community characteristics were not systematically collected. These gaps restrict our ability to assess how different environments may have shaped youth engagement with the infographic. As vaping prevalence, attitudes, and access to vaping products can vary across provinces, particularly between rural and urban settings [33,34], the findings should be interpreted as exploratory rather than generalizable.

This study also used a purposive self-selection sampling strategy, which primarily reached youth already engaged in school-based programs or youth advisory councils. This aligns with the participatory nature of the intervention, ensuring feedback from youth familiar with digital content and public health messaging. However, it likely introduced selection bias as these participants may differ from less engaged youth in terms of media literacy, health knowledge, or responsiveness to educational tools. As a result, the findings may overestimate perceived relevance or impact and cannot be extrapolated to disengaged or marginalized youth populations. Broader outreach strategies will be necessary to examine the infographic's effectiveness across different demographic and psychosocial contexts.

Perhaps most significantly, this study relied exclusively on self-reported perceptions to assess knowledge gain and the effectiveness of the infographic. Without validated pre- and postexposure testing or behavioral measures, we cannot determine whether participants' perceived understanding translated into actual knowledge uptake or shifts in intent to vape. These subjective impressions, although valuable for design feedback, are vulnerable to biases such as social desirability bias, confirmation bias, and imprecise recall. This substantially limits the strength of any claims regarding impact. Future research should incorporate validated, objective assessments of pre- and postexposure knowledge and behavioral intention measures to more rigorously evaluate the impact of youth-centered digital health interventions, including their potential for sustained knowledge retention and behavior change over time.

Finally, while ever vapers and never vapers were both included in the sample, subgroup comparisons were not conducted due to the small sample size. Qualitative responses were initially disaggregated but were ultimately merged due to thematic consistency. However, this decision may have obscured important differences in how youth with different vaping histories interpret and respond to prevention materials. The absence of stratified analyses in the quantitative results similarly limits insight into demographic or behavioral variability. Future studies with larger and more diverse samples should conduct subgroup comparisons (eg, by vaping status, ethnicity, gender, or geography) and control for key variables to better understand how targeted educational tools function across different youth populations.

Conclusions

This study suggests that youth-driven infographics may be useful educational tools for conveying vaping-related information and enhancing youth engagement with prevention messages. The findings highlight the potential value of participatory design, visual accessibility, and credibility in improving the relevance and clarity of digital health communication for young people. These findings are promising; however, the results should be interpreted with caution due to the small, nonrepresentative sample and reliance on self-reported perceptions rather than objective or behavioral outcomes. While the infographic appeared to support understanding of vaping-related harms among participants, further research is needed to assess its effectiveness with more diverse youth populations and to determine whether such tools can lead to sustained knowledge retention or behavior change over time.

XSL•FO
RenderX

Acknowledgments

This work was supported by a Canadian Cancer Society Emerging Scholar Award (grant 707156) and a Michael Smith Health Research British Columbia Scholar Award (grant SCH-2023-3195). The authors sincerely thank the youth advisory councils and youth coresearchers who contributed to developing the *What's in a Vape?* infographic. Their insights, feedback, and collaboration were invaluable in shaping this resource. The authors would also like to acknowledge Annie Cai and Cathy Yeung for their direct contributions to providing feedback and refining the final product. Their dedication to youth-centered knowledge translation made this project engaging and impactful. The authors also extend a special thank you to Adrianna Chong for their expertise in translating the infographic into a digital format, incorporating graphic design elements, and implementing interactive strategies to enhance accessibility and engagement. Their work was essential in bringing this infographic to life.

Authors' Contributions

All authors substantially contributed to the conception of the work, interpretation of the data, and development of this manuscript. All authors contributed to drafting or revising the manuscript, gave final approval of the version to be published, and agreed to be accountable for all aspects of the work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
What's in a Vape? infographic.
[DOCX File , 1864 KB - formative_v9i1e75694_app1.docx]

References

1. Hammond D, Reid JL. Trends in vaping and nicotine product use among youth in Canada, England and the USA between 2017 and 2022: evidence to inform policy. Tob Control 2025 Jan 02;34(1):115-118. [doi: 10.1136/tc-2023-058241] [Medline: 37940402]
2. Canadian Tobacco and Nicotine Survey (CTNS): summary of results for 2022. Government of Canada. 2022. URL: https://www.canada.ca/en/health-canada/services/canadian-tobacco-nicotine-survey/2022-summary.html [accessed 2025-06-14]
3. Consider the consequences of vaping. Government of Canada. URL: https://www.canada.ca/en/services/health/campaigns/vaping.html [accessed 2025-06-14]
4. Leslie FM. Unique, long-term effects of nicotine on adolescent brain. Pharmacol Biochem Behav 2020 Oct;197:173010 [FREE Full text] [doi: 10.1016/j.pbb.2020.173010]
5. Augenstein JA, Smaldone AM, Usseglio J, Bruzzese J. Electronic cigarette use and academic performance among adolescents and young adults: a scoping review. Acad Pediatr 2024 Mar;24(2):228-242 [FREE Full text] [doi: 10.1016/j.acap.2023.09.012] [Medline: 37748535]
6. Dearfield CT, Chen-Sankey JC, McNeel TS, Bernat DH, Choi K. E-cigarette initiation predicts subsequent academic performance among youth: results from the PATH Study. Prev Med 2021 Dec;153:106781 [FREE Full text] [doi: 10.1016/j.ypmed.2021.106781] [Medline: 34487749]
7. Grant JE, Lust K, Fridberg DJ, King AC, Chamberlain SR. E-cigarette use (vaping) is associated with illicit drug use, mental health problems, and impulsivity in university students. Ann Clin Psychiatry 2019 Feb;31(1):27-35 [FREE Full text] [Medline: 30699215]
8. Mehra VM, Keethakumar A, Bohr YM, Abdullah P, Tamim H. The association between alcohol, marijuana, illegal drug use and current use of E-cigarette among youth and young adults in Canada: results from Canadian Tobacco, Alcohol and Drugs Survey 2017. BMC Public Health 2019 Sep 02;19(1):1208 [FREE Full text] [doi: 10.1186/s12889-019-7546-y] [Medline: 31477067]
9. Westling E, Rusby JC, Crowley R, Light JM. Electronic cigarette use by youth: prevalence, correlates, and use trajectories from middle to high school. J Adolesc Health 2017 Jun;60(6):660-666 [FREE Full text] [doi: 10.1016/j.jadohealth.2016.12.019] [Medline: 28242187]
10. The adolescent brain and substance use. National Institute on Drug Abuse. 2024 Jun. URL: https://nida.nih.gov/research-topics/adolescent-brain-substance-use [accessed 2025-06-14]
11. Canadian drugs and substances strategy: overview. Government of Canada. URL: https://www.canada.ca/en/health-canada/services/substance-use/canadian-drugs-substances-strategy.html [accessed 2025-06-14]
12. DeVito EE, Krishnan-Sarin S. E-cigarettes: impact of e-liquid components and device characteristics on nicotine exposure. Curr Neuropharmacol 2018 May 01;16(4):438-459 [FREE Full text] [doi: 10.2174/1570159x15666171016164430]

XSL•FO
RenderX

13. Kosmider L, Sobczak A, Fik M, Knysak J, Zaciera M, Kurek J, et al. Carbonyl compounds in electronic cigarette vapors: effects of nicotine solvent and battery output voltage. Nicotine Tob Res 2014 Oct;16(10):1319-1326 [FREE Full text] [doi: 10.1093/ntr/ntu078] [Medline: 24832759]
14. Park JA, Crotty Alexander LE, Christiani DC. Vaping and lung inflammation and injury. Annu Rev Physiol 2022 Feb 10;84:611-629 [FREE Full text] [doi: 10.1146/annurev-physiol-061121-040014] [Medline: 34724436]
15. Rohde JA, Noar SM, Mendel JR, Hall MG, Baig SA, Ribisl KM, et al. E-cigarette health harm awareness and discouragement: implications for health communication. Nicotine Tob Res 2020 Jun 12;22(7):1131-1138 [FREE Full text] [doi: 10.1093/ntr/ntz194] [Medline: 31593586]
16. Greiner Safi A, Kalaji M, Avery R, Niederdeppe J, Mathios A, Dorf M, et al. Examining perceptions of uncertain language in potential e-cigarette warning labels: results from 16 focus groups with adult tobacco users and youth. Health Commun 2024 Mar 30;39(3):460-481 [FREE Full text] [doi: 10.1080/10410236.2023.2170092] [Medline: 36717390]
17. Kong HK, Zainab F, Turner AM, Bekemeier B, Backonja U. Trends in and effectiveness of infographics for health communication: a scoping review. Health Commun 2025 Feb 17;40(2):222-232 [FREE Full text] [doi: 10.1080/10410236.2024.2342595] [Medline: 38629448]
18. Zeller M. Evolving "the real cost" campaign to address the rising epidemic of youth e-cigarette use. Am J Prev Med 2019 Feb;56(2 Suppl 1):S76-S78 [FREE Full text] [doi: 10.1016/j.amepre.2018.09.005] [Medline: 30661529]
19. Stalgaitis CA, Djakaria M, Jordan JW. The vaping teenager: understanding the psychographics and interests of adolescent vape users to inform health communication campaigns. Tob Use Insights 2020 Jul 30;13 [FREE Full text] [doi: 10.1177/1179173x20945695]
20. England KJ, Edwards AL, Paulson AC, Libby EP, Harrell PT, Mondejar KA. Rethink vape: development and evaluation of a risk communication campaign to prevent youth e-cigarette use. Addict Behav 2021 Feb;113:106664 [FREE Full text] [doi: 10.1016/j.addbeh.2020.106664] [Medline: 33038677]
21. Edmonds WA, Kennedy TD. Convergent-parallel approach. In: An Applied Guide to Research Designs: Quantitative, Qualitative, and Mixed Methods. Thousand Oaks, CA: SAGE Publications; 2017.
22. Eysenbach G. Improving the quality of web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34 [FREE Full text] [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
23. Barwick MA. Knowledge Translation Planning Template. The Hospital for Sick Children. 2013. URL: https://www.sickkids.ca/en/learning/continuing-professional-development/knowledge-translation-training/knowledge-translation-planning-template-form/ [accessed 2025-06-14]
24. Auerbach C, Silverstein LB. Qualitative Data: An Introduction to Coding and Analysis. New York, NY: New York University Press; 2003.
25. Smith B, McGannon KR. Developing rigor in qualitative research: problems and opportunities within sport and exercise psychology. Int Rev Sport Exerc Psychol 2017 May 14;11(1):101-121 [FREE Full text] [doi: 10.1080/1750984x.2017.1317357]
26. Fetters MD, Curry LA, Creswell JW. Achieving integration in mixed methods designs-principles and practices. Health Serv Res 2013 Dec;48(6 Pt 2):2134-2156 [FREE Full text] [doi: 10.1111/1475-6773.12117] [Medline: 24279835]
27. Roditis ML, Dineva A, Smith A, Walker M, Delahanty J, D'lorio E, et al. Reactions to electronic nicotine delivery system (ENDS) prevention messages: results from qualitative research used to inform FDA's first youth ENDS prevention campaign. Tob Control 2020 Sep 10;29(5):510-515 [FREE Full text] [doi: 10.1136/tobaccocontrol-2019-055104] [Medline: 31506379]
28. Andrews JC, Netemeyer RG, Kees J, Burton S. How graphic visual health warnings affect young smokers' thoughts of quitting. J Mark Res 2014 Apr 01;51(2):165-183 [FREE Full text] [doi: 10.1509/jmr.13.0092]
29. Park SE, Tang L. How colour and visual complexity affect the evaluation of skin cancer infographics: an experiment study. J Vis Commun Med 2019 Apr;42(2):52-65 [FREE Full text] [doi: 10.1080/17453054.2019.1573633] [Medline: 31014148]
30. Ambrose BK, Rostron BL, Johnson SE, Portnoy DB, Apelberg BJ, Kaufman AR, et al. Perceptions of the relative harm of cigarettes and e-cigarettes among U.S. youth. Am J Prev Med 2014 Aug;47(2 Suppl 1):S53-S60 [FREE Full text] [doi: 10.1016/j.amepre.2014.04.016] [Medline: 25044196]
31. Amrock SM, Zakhar J, Zhou S, Weitzman M. Perception of e-cigarette harm and its correlation with use among U.S. adolescents. Nicotine Tob Res 2015 Mar;17(3):330-336 [FREE Full text] [doi: 10.1093/ntr/ntu156] [Medline: 25125321]
32. McKelvey K, Halpern-Felsher B. From tobacco-endgame strategizing to Red Queen's race: the case of non-combustible tobacco products. Addict Behav 2019 Apr;91:1-4 [FREE Full text] [doi: 10.1016/j.addbeh.2019.01.001] [Medline: 30642577]
33. Hammond D, Reid JL, Rynard VL, Fong GT, Cummings KM, McNeill A, et al. Prevalence of vaping and smoking among adolescents in Canada, England, and the United States: repeat national cross sectional surveys. BMJ 2019 Jun 20;365:l2219 [FREE Full text] [doi: 10.1136/bmj.l2219] [Medline: 31221636]
34. Canadian student alcohol and drugs survey. Government of Canada. URL: https://www.canada.ca/en/health-canada/services/canadian-student-tobacco-alcohol-drugs-survey.html [accessed 2025-06-14]

Abbreviations

VIP: Vaping Information Project

XSL•FO
RenderX

Edited by J Sarvestan; submitted 09.04.25; peer-reviewed by A Adekola, B Afolabi, N Baharom; comments to author 18.06.25; revised version received 26.06.25; accepted 31.07.25; published 11.09.25.

Please cite as:
Nelson Ferguson K, Christianson K, Delgado A, Martin G, Coen SE, Struik L
A Youth-Centered Digital Infographic on Vaping Risks (What's in a Vape?): Mixed Methods Study
JMIR Form Res 2025;9:e75694
URL: https://formative.jmir.org/2025/1/e75694
doi:10.2196/75694
PMID:

©Kendra Nelson Ferguson, Kyla Christianson, Adrian Delgado, Gina Martin, Stephanie E Coen, Laura Struik. Originally published in JMIR Formative Research (https://formative.jmir.org), 11.09.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Needs and Expectations Associated With an e–Mental Health Intervention for Reducing Somatoform, Anxiety, and Depressive Symptoms in Sexual and Gender Minority Adults: Qualitative Participative Study

Anna Gomes[1*], MSc; Angeliki Tsiouris[1*], Dr Phil; Florian Jung[1], MSc; Laura Rebecca Klein[1], MSc; Adina Kreis[1], MSc; Manfred E Beutel[1], Dr Med; Bernhard Strauss[2], Dr Phil; Madita Hoy[2], Dr Phil; Rüdiger Zwerenz[1], Dr Biol Hom

[1]Department for Psychosomatic Medicine and Psychotherapy, University Medical Center of the Johannes Gutenberg-University Mainz, Mainz, Germany
[2]Institute of Psychosocial Medicine, Psychotherapy and Psychooncology, University Hospital Jena, Jena, Germany
[*]these authors contributed equally

Corresponding Author:
Adina Kreis, MSc
Department for Psychosomatic Medicine and Psychotherapy
University Medical Center of the Johannes Gutenberg-University Mainz
Untere Zahlbacher Str. 8
Mainz, 55131
Germany
Phone: 49 6131 17 3376
Fax: 49 6131 17 5563
Email: adina.kreis@unimedizin-mainz.de

Abstract

Background: Sexual and gender minority individuals experience heightened risks of mental health disorders due to marginalization, discrimination, and inadequacies in health care.

Objective: This study aims to identify the needs and expectations concerning an e–mental health intervention designed for people who are lesbian, gay, bisexual, transgender, queer or questioning, intersex, asexual, or have other sexual orientation and gender identities (LGBTQIA+) to reduce somatoform, anxiety and depressive (SAD) symptoms.

Methods: A qualitative participative study was conducted, involving semistructured interviews (face-to-face and online) with 10 sexual and gender minority individuals experiencing SAD symptoms. Telephone interviews were conducted with 10 health care professionals (HCPs). This study was part of a participatory project, emphasizing cooperation with the LGBTQIA+ community. Data were analyzed through a deductive-inductive content analysis to derive categories of needs and expectations relevant for the development of an e–mental health intervention.

Results: Participants expressed a strong desire for the intervention to be inclusive, validating, and sensitive to the unique challenges faced by LGBTQIA+ people. Key themes included the need for information on the relationship between being queer and mental health; representation through case stories; psychoeducation; and exercises tailored to address minority stress, identity affirmation, and coping strategies. HCPs emphasized the importance of addressing the coming-out process, managing rejection, fostering self-acceptance, and including content on minority stress and its impact on mental health. Results of both interview groups highlighted the need for direct interaction with therapists or peer support, including both synchronous and asynchronous elements (eg, video calls and chat) based on nonheteronormative, sensitive therapeutic support, for example, avoiding preassumptions, using sensitive language, and reflecting possible trigger points.

Conclusions: This study underscores the need for e–mental health interventions tailored to a queer-sensitive and participatory approach. Interventions should incorporate comprehensive psychoeducation, interactive elements, content reflecting the lived experiences of LGBTQIA+ individuals with SAD symptoms, and the possibility to connect and exchange experiences with others facing similar challenges. Engaging with both LGBTQIA+ people and HCPs in the development process is essential to ensure the intervention's relevance, effectiveness, and acceptability.

(JMIR Form Res 2025;9:e65834) doi:10.2196/65834

XSL•FO
RenderX

KEYWORDS

sexual orientation; gender identity; sexual and gender minority; minority stress; qualitative methods; prevention; barriers to care; e–mental health; community-based participatory research

Introduction

Theoretical Background

Sexual and gender minority (SGM) individuals, encompassing nonheterosexual, transgender, and nonbinary populations, represent a demographic group that has historically experienced marginalization, mistreatment, pathologization, and neglect within a predominant cis-heteronormative society [1,2]. Existing inadequacies, such as microaggressions, are also barriers to health care for individuals who are lesbian, gay, bisexual, transgender, queer or questioning, intersex, asexual, or have other sexual orientation and gender identities (LGBTQIA+) [3-6]. Looking at LGBTQIA+ individuals with mental burden, it is important to acknowledge that identifying as an SGM individual is often associated with specific mental health risk factors that might not apply to cisgender heterosexual individuals [7,8]. The theory of minority stress [9] postulates that belonging to a minority group might be connected with distal factors, such as discrimination and victimization experiences, and proximal factors, such as fear of rejection, insecurity within the system, and internalized negative attitudes toward one's own identity [10]. Hence, health-related stereotype threats play a major role among SGM individuals and have a strong influence on their behavior. They are often associated with fear of physicians, delays in seeking mental health services, and poor self-reported mental health outcomes [11,12]. Experiencing stigma and discrimination elevates distress and limits effective coping strategies that interfere with positive development, health, and well-being in SGM individuals [13]. This is critical, as the currently limited empirical research on SGM individuals in Germany already shows a markedly increased risk of mental health disorders within this group. The risk of experiencing depression or burnout, for instance, is 3 times higher for members of the LGBTQIA+ community compared to the general population [14]. International research shows that lesbian, gay, bisexual, and transgender (LGBT) adolescents are 1.75 times more likely to experience symptoms of anxiety and depression than their cisgender heterosexual peers [15]. Finally, research suggests a higher risk for suicidal behaviors among LGBT populations [16]. Besides internalized homophobia and shame, significantly poorer mental health outcomes were identified as risk factors for suicide in LGBT populations [17].

E–mental health interventions promise effective ways to bridge treatment gaps [18-20]. Due to the described experiences within the health care system, many SGM individuals have a high affinity for online platforms to seek health-related information and support [21]. Internationally, a few clinical trials evaluated LGBTQIA+-specific e–mental health interventions with promising results. In the United States, a randomized controlled trial evaluating an online intervention revealed that LGBTQIA+ participants allocated to an expressive writing group or a self-affirmation intervention reported significant positive changes in several mental health parameters compared with participants in the control group [22]. In Germany, a treatment tailored to the needs of LGBTQIA+ individuals with mental disorder symptoms is not yet available. To the best of our knowledge, existing German e–mental health interventions mostly represent hetero-normative life realities in their content, so that persons outside these life realities often feel not sufficiently reflected. However, evidence is scarce, and research on the specific needs of LGBTQIA+ individuals regarding the content of an e–mental health intervention is crucial in order to develop an intervention that is effective in reducing symptoms of mental disorders and does not exclusively reproduce cis-heteronormative content.

In the development of online interventions, it is recommended to involve end users in the design process, a practice commonly applied in approaches such as user-centered design and cocreation. It leads to increased adherence and effectiveness of the intervention and is hence state of the art in intervention development [23,24]. Although there is consensus on the necessity of involving end users, various approaches with different levels of participation do exist [25,26]. For individuals who frequently encounter experiences of discrimination and marginalization, a high level of participation in the development has an even greater significance [27]. This guarantees an empowerment-oriented, sensitive approach with communication in all phases of research and development.

It is thus considered pivotal, in a first step, to capture the unmet needs and expectations of LGBTQIA+ people regarding an online intervention to ensure that the intervention aligns with the target group's needs. Using a participatory approach even at this stage could also lead to participants feeling welcome, validated, and understood—an experience that many LGBTQIA+ adults often lack not only in their social environments but also in the medical and therapeutic context.

Research Objectives

We do not know of any previous participative research studies that have investigated LGBTQIA+ individuals' expectations and needs regarding an e–mental health intervention. This study is part of the Queer-E–Mental Health (Queer-EMH) project, a bicentric research project carried out by the University of Mainz and the University of Jena that aims to design an effective e–mental health intervention for LGBTQIA+ individuals with mental burden in a participatory study design and establish an LGBTQIA+ advisory board. To tailor this intervention specifically to the needs of the target group, their specific needs and expectations were investigated. In terms of a user-centered development process, this study's objectives are as follows:

1. To determine attitudes, expectations, and potential barriers associated with psychological online interventions aiming at reducing somatoform, anxiety, depressive (SAD) symptoms in LGBTQIA+ individuals with mental burden
2. To obtain information and recommendations from health care professionals (HCPs) and counseling staff with experience in working with LGBTQIA+ people about what

XSL•FO
RenderX

an e–mental health intervention should contain and how it should be designed to reach the target group

Methods

Research Team and Reflexivity

This project was supported by an LGBTQIA+ advisory board, which was established at the beginning of the project and consisted of members of 10 German LGBTQIA+ associations. In a participatory setting, the LGBTQIA+ advisory board ensured that the study processes proceeded in a participant-friendly and queer-sensitive manner. In numerous sessions, work packages in the project (eg, recruitment procedures, wording of recruitment materials, development of questions for the interview guide, and data interpretation) were collaboratively developed and worked on.

One of the interviewers simultaneously worked as a coder. The interviews were conducted by researchers AT, AG, and Clara Pering at the study center in Mainz and by MH at the study center in Jena. All interviewers were female. No personal information about the interviewers (such as sexual orientation) was revealed to participants. There were no personal relationships between the interviewers and participants before the study. As clinical psychologists and experienced psychotherapy researchers, the interviewers are trained in qualitative research methods and the clinical picture of SAD disorders. Through their collaboration with our advisory board, they gained essential insights into relevant aspects of queer identity and living environments. A personal queer biography was not a prerequisite for the interviewers.

Ethical Considerations

Ethics approval for the study was obtained from the ethics committee of the Medical Association of the Federal State of Rhineland-Palatinate (registration number 2022-16876_1-NIS) and the ethics committee of the Jena University Hospital (registration number 2022-2855-Bef). All participants gave their informed consent on paper and were informed that participation was voluntary and that they could withdraw their consent at any time. The data were pseudonymized. Personal contact details of the participants were stored separately from interview data, inaccessible to third parties. There were no financial incentives for participants in this study. This study followed the COREQ (Consolidated Criteria for Reporting Qualitative Research) guidelines (Multimedia Appendix 1).

Recruitment and Participants

We conducted a qualitative study, with data collected via semistructured interviews with LGBTQIA+ individuals and telephone interviews with HCPs and counseling staff who regularly worked with LGBTQIA+ individuals with SAD symptoms (eg, consultation or psychotherapy). Using a cocreative, participatory approach, we ensured that the conceptualization of a queer-sensitive EMH intervention addresses the actual challenges perceived not only by LGBTQIA+ individuals but also by HCPs and counseling experts.

We recruited LGBTQIA+ individuals using various recruitment strategies. A central pillar of the recruitment was the Instagram project account (Meta Platforms, Inc) on which we published information about our study and referred to the possibility of participating. Besides social media recruitment, the cooperating LGBTQIA+ advisory board representatives distributed the study flyer and issued the call for participants within their networks (mainly via email). The study flyer contained information on the study procedure, inclusion criteria, and contact details for the study staff. Participants were included if they fulfilled the following inclusion criteria: (1) identified as LGBTQIA+ individuals, (2) reported SAD symptoms (currently or in the past), (3) were aged ≥18 years, (4) had sufficient German language skills to participate in an interview or focus group, and (5) were able to provide informed consent. Participants were excluded if they had severe mental burden (eg, severe depression and suicidal tendencies). In this case, the study staff would provide information on where to find professional support.

HCPs were recruited with the support of the LGBTQIA+ advisory board. HCPs interested in study participation reached out to the project team and were enrolled if they met the following inclusion criteria: (1) had experience counseling LGBTQIA+ individuals with mental disorders, (2) were aged ≥18 years, (3) had sufficient German language skills, and (4) provided informed consent.

Study Procedure

Interested individuals contacted the study team and were enrolled in the study following a telephone screening. During the telephone screening, participants were screened for the inclusion and exclusion criteria and informed about the aim and potential risks of the study (ie, exposure to topics that might trigger an emotional reaction, such as sadness, anxiety, or anger). With regard to symptomatology, participants were asked a yes or no question during the phone call about whether they were currently experiencing or had previously experienced SAD symptoms. Thus, participants self-referred to either the interview or 1 of the 4 focus groups that took place during the same time period. In this paper, we analyzed the interview results only.

A total of 51 interested persons contacted the research staff. We included participants in our study consecutively based on the date on which participants expressed their interest in the study. In addition, we took care to ensure representation of a diverse range of gender identities and sexual orientations among the participants, aiming to incorporate various subgroups within the LGBTQIA+ community into the study. One participant was excluded from the analyses of this study, as attitudes regarding online interventions were not discussed because the participant had already participated in the focus group earlier and provided feedback regarding the relevant questions in the focus groups. We stopped including participants after reaching the target of 10 participants. Because the qualitative study presented in this paper was only a small subproject within a larger project aimed at developing a concept for the corresponding intervention, the available personnel and time resources were limited, leading to the predetermined restriction on the number of interviews.

XSL•FO
RenderX

The interviews were conducted in a 2-person setting, either face-to-face in the study centers (Mainz and Jena) or online via a secure video platform. All interviews were audio recorded and subsequently transcribed and coded using MAXQDA (version 11.1.4; VERBI Software).

Interview Guide

The semistructured interview guide was developed in an iterative process in close cooperation with representatives of the LGBTQIA+ advisory board. In detail, the research staff collected relevant research topics and proposed specific questions and prompts, which were then presented to the advisory board and discussed in regard to their relevance and appropriateness (eg, in terms of sensitive language and potential triggers). The feedback process was repeated until the research staff and the representatives of the LGBTQIA+ advisory board agreed on the interview guide.

The interview started with an open-ended question about the participants' motivation to contribute to the study, followed by items querying the main topics (1) subjective mental illness concept, (2) medical and psychotherapeutic experiences, (3) attitudes and expectations toward e–mental health interventions, and (4) personal strengths and conclusion. While the primary objective of the interviews was to gather information on topic 3, the detailed introduction served to provide participants with time to ease into the conversation and establish a validating, understanding atmosphere from the outset, encompassing feelings and difficult experiences within and beyond the health care system. This was particularly crucial, as the interviewers, working in a clinic, were inherently part of the health care system, necessitating the dismantling of any preexisting reservations.

In this study, we analyzed and interpreted participants' responses to the main topic 3. The interview questions querying attitudes and expectations toward a queer sensitive online intervention are reported in Textbox 1. While this study examined not only content-related expectations but also aspects of presentation format and structural considerations, this paper focuses on presenting and discussing the findings related to content expectations. This decision was driven by the fact that presentation and structural requirements (eg, mode of delivery) did not differ significantly from those of other target groups of online interventions and have already been extensively researched, with a wealth of existing literature covering these aspects comprehensively [28,29].

Textbox 1. Interview questions querying attitudes and expectations of individuals who are lesbian, gay, bisexual, transgender, queer or questioning, intersex, asexual, or have other sexual orientations and gender identities (LGBTQIA+) toward a queer-sensitive online intervention.

Needs and expectations

- What would you like to see in an e–mental health intervention tailored to the needs of the LGBTQIA+ community?
- What are your expectations for an e–mental health intervention tailored to the needs of the LGBTQIA+ community?
- What advantages and disadvantages or caveats might be associated with an e–mental health intervention for members of the LGBTQIA+ community?

Qualitative data from the interviews were supplemented by sociodemographic data collected with a brief questionnaire, which was filled out by each participant before the interview started. The short questionnaire assessed gender identity, sexual orientation, year of birth, place of residence, marital or relationship status, highest education, and whether participants were currently on sick leave.

Qualitative Content Analysis

The coding guide was developed following an iterative deductive-inductive procedure by first deriving relevant categories from the literature (deductive) and then adding additional categories based on the data collected in the interviews (inductive). Each interview was coded by 2 persons and ratings were compared in terms of consensus. Deviating codes were discussed until consensus was reached. Frequencies of statements in the respective main and subcategories were analyzed descriptively. During the data analysis, the results were reviewed and interpreted together, with the LGBTQIA+ advisory board providing valuable support in contextualizing quotes, which added significant value to the interpretation of the findings.

Telephone Interviews With HCPs and Counseling Staff

In order to supplement attitudes toward and expectations from an e–mental health intervention from the perspective of LGBTQIA+ persons with symptoms of mental health disorders, we also conducted structured telephone interviews with 11 HCPs. The structured interview guide was developed in close cooperation with the LGBTQIA+ advisory board. The interview guide consisted of the following main topics: (1) perceived treatment needs of LGBTQIA+ individuals with SAD disorders, (2) LGBTQIA+ individuals in the German health care system, and (3) attitudes toward e–mental health interventions for LGBTQIA+ individuals. Similar to the interviews with LGBTQIA+ individuals, this paper analyzes responses by HCPs and counseling staff to the main topic 3. Detailed questions are presented in Textbox 2.

Interviews with HCPs were conducted by telephone and systematically documented in written bullet points by the interviewers. These bullet points were then transferred into the software MAXQDA and coded into categories. The qualitative content analysis of the data assessed in the telephone interviews with HCPs and counseling staff followed the same deductive-inductive approach and procedure as the data assessed in the face-to-face interviews.

XSL•FO
RenderX

Textbox 2. Interview questions querying the expectations and concerns of health care professionals and counseling staff regarding an online intervention for individuals who are lesbian, gay, bisexual, transgender, queer or questioning, intersex, asexual, or have other sexual orientations and gender identities (LGBTQIA+).

Needs and expectations

- What characteristics would an online intervention for LGBTQIA+ people with mental illness need to have for you to recommend it?
- What benefits might an online intervention provide for LGBTQIA+ people with mental illness?

Concerns and barriers

- What concerns do you personally have about e–mental health interventions?
- What concerns do you think your clients might have about e–mental health interventions?

Results

Research Objective 1: The Perspectives of LGBTQIA+ Individuals

Overview

The final sample consisted of 10 adults with symptoms of SAD disorders who identified themselves as LGBTQIA+. Table 1 displays sample characteristics. Participants were aged between 23 and 67 (mean 38.2, SD 12.75) years. Half (n=5, 50%) of the participants were single and had a diploma qualifying for university. Most (n=6, 60%) of the participants were currently on sick leave or retired. The interviews lasted for an average duration of 72.8 (SD 11.35) minutes.

XSL•FO
RenderX

Table 1. Sociodemographic characteristics of the sample of individuals who are lesbian, gay, bisexual, transgender, queer or questioning, intersex, asexual, or have other sexual orientations and gender identities (N=10).

Variable	Values
Age (y), mean (SD)	38.2 (12.75)
Gender or gender identity, n (%)	
Cisgender female	2 (20)
Cisgender male	1 (10)
Nonbinary and transmasculine	1 (10)
Nonbinary	2 (20)
Transgender female	3 (30)
Intersex female	1 (10)
Sexual and romantic orientation, n (%)	
Gay	2 (20)
Lesbian	2 (20)
Biromantic and asexual	1 (10)
Bisexual	1 (10)
Demisexual	1 (10)
Pansexual	1 (10)
Queer (pansexual)	1 (10)
Individual	1 (10)
Current relationship status, n (%)	
Single	5 (50)
In a relationship	3 (30)
Married or civil union	2 (20)
Highest education, n (%)	
No degree or lower secondary education diploma	3 (30)
General secondary education diploma	1 (10)
Diploma qualifying for university	5 (50)
University degree	1 (10)
Currently on sick leave	4 (40)
Retired	2 (20)

The previously described inductive-deductive procedure revealed 4 broad categories. Each category described a different level into which the expectations, barriers, and requirements regarding an online intervention for LGBTQIA+ individuals with SAD disorders, as raised in the interviews, could be classified: (1) content level, (2) linguistic level, (3) framework, and (4) other requirements. Figure 1 provides an overview of the number of codes assigned to the respective categories and subcategories. The subsequent paragraph presents participants' experiences categorized by those themes. We have put a special focus on the presentation of the category *content level*. The 3 remaining categories (linguistic level, setting, and other requirements) are subsequently briefly presented, and characteristics specific to the LGBTQIA+ population are highlighted. We have then presented complementary data drawn from the HCPs' interviews, highlighting similarities and differences in their respective perspectives.

XSL•FO
RenderX

Figure 1. Number of codes by category and subcategory in the qualitative content analysis of the interviews with individuals who are lesbian, gay, bisexual, transgender, queer or questioning, intersex, asexual, or have other sexual orientations and gender identities with symptoms of depression, anxiety, or somatoform disorders.

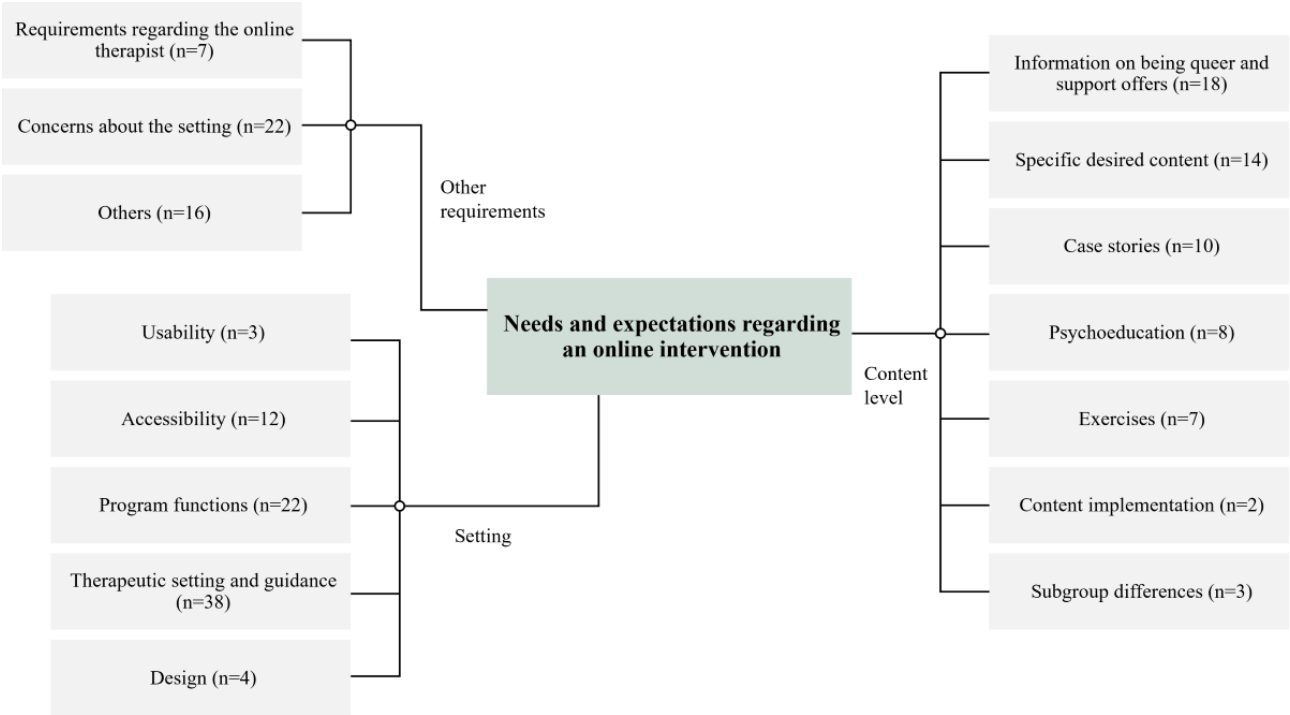

Content Level

Information About Being Queer and Support Offers (18 Codes)

On the basis of the feedback from most (7/10, 70%) of the participants, it became clear that integrating comprehensive information resources on queerness within the planned online self-help intervention is crucial. While the internet offers a wealth of information, distinguishing credible sources remains a challenge. For example, certain online content, such as definitions of specific terms, was reported as being discriminatory and emotionally harmful:

> *Because I...have my information since I was 14, that's when it started, that I collected information, but it must also be said that this collection has led to a lot of harmful contacts on the internet with harmful material, which is sexist, objectifying or simply represents hate crime somewhere. And I think that a self-help service like this, where all the information is brought together in written form with experiences, would be very helpful.*

In addition to the limited resources addressing critical topics, such as the impacts of hormone treatments during transitioning, participants reported a lack of easily accessible avenues for locating practical assistance. Participants expressed the need for a comprehensive directory of LGBTQIA+ friendly contacts, counseling services, support networks, outpatient facilities, and affirming medical practitioners and reported that this dearth of resources often perpetuates feelings of helplessness and isolation:

> *I was misgendered, I was not respected, I was not accepted. How do I deal with that? Where do I go for help? Where do I go for support? To put it bluntly, where can I have a good cry? Where can I get hope?*

Specific Desired Content (14 Codes)

A wide variety of content-related specific wishes and ideas was expressed by the participants (4/10, 40%), summarized under the category *specific desired content*. Significant attention revolved around the basic attitude with which the program and content should be designed. The participants expressed a deep desire to feel seen and acknowledged within the program. In this context, they referred to the predominantly hetero- and cis-normative society, in which they perceived a lack of acknowledgment and representation of LGBTQIA+ persons. The desire for validation and acceptance of one's own gender identity and sexual orientation as well as the normalization of the full spectrum of emotions associated with this aspect of personal identity was an important aspect that was repeatedly emphasized in all interviews. In addition, participants expressed strong interest in understanding the intersection between mental health challenges and the experience of being queer, including insights into how minority stress could potentially contribute to mental health disorders. Concerning therapeutic content, a recurring theme among many participants was to address their own emotions within the online intervention, with a focus on exploring, validating, and normalizing those feelings:

> *Include the emotional more....Emotions play a crucial role here, especially when it comes to one's own gender or sexual identity. This is inextricably linked with emotions. And to give space to the emotions and to confirm them...it is okay, if you are afraid of coming out, for example, or if you are angry that you are going through this experience, that it makes you sad, that is all valid at first. And then: That's where you*

XSL•FO
RenderX

> *can continue to work, so it's not like you're caught up in the emotions all the time, but you can work with that.*

There were also statements that related specifically to the therapeutic rationale. For example, one participant demanded clear strategies and tools, while another participant emphasized their need for the therapeutic orientation to go beyond standard psychotherapeutic approaches, such as cognitive behavioral therapy (CBT), which is currently most frequently used in the online context:

> *I think what would be important for me, for example, is not only to go to CBT, as useful as that is, but that doesn't work for every person. And maybe also to create offers, if that is somehow possible, that you can also control a little bit in which direction you want to go or what helps you. Because I often have the feeling that such offers are often very CBT-based. And as helpful as some exercises and some things are for me, I...often have the feeling...that often doesn't really work for me.*

Case Stories (10 Codes)

Many (4/10, 40%) interviewees expressed the desire to integrate personal case examples featuring individuals representing diverse life realities into the online intervention:

> *Or for example "I am gay. I came out at home. My parents kicked me out." Unfortunately, that still happens very often today. Where do I go then if I can't go to friends or family? And then that person just tells how they dealt with it. What did she do, where did she go? Did she go to friends' houses first, have a good cry, sleep on the couch or something? Or did she go to some counseling center and find someone to stay with? And when that is told as a personal experience, it gives you a bit of courage. And gives orientation.*

Furthermore, in the context of dealing with complex emotional responses, particularly regarding self-discovery and identity formation, the central role of these role models in normalizing a broad spectrum of intense emotions was underscored:

> *I find it helpful when other people describe their own emotions and experiences. Above all, that, this emotional aspect also comes into it, because of course finding one's identity is connected with a lot of emotion. That this is really described a bit...That you realize that you are not alone with these emotions. That you feel represented and also seen.*

Psychoeducation (8 Codes)

Concerning the participants' mental distress, a need for psychoeducation on mental health disorders was evident (6/10, 60%). This included not only information on the nature of mental illnesses, for example, their development, progression, associated symptoms, maintaining factors, and explanatory models, but also comprised the role of emotions, the interplay between mind and body, and issues such as suicidality. In addition, the intersections between mental symptomatology and being queer were highlighted as crucial points of interest:

> *I am currently thinking about whether it would be useful, if such an online offer also had short explanations about anxiety disorders, for example, social anxiety, such as a graphic, as my therapist drew, of how social anxiety develops,... perhaps that could help....*

It was emphasized several times that while it is important to address and explain connections between mental distress and being queer, it is also crucial to make a clear distinction between mental health conditions and sexual or romantic orientation or gender identity. This referred especially to the fact that, in any case, the impression that being an individual belonging to the LGBTQIA+ community would be equated with mental illness should be avoided, as this, in turn, would reproduce often experienced stigmatizations and prejudices:

> *[S]o to clearly separate sexual or romantic orientation and the clinical syndromes, because that blurs a lot. To be safe,...there is nothing wrong with me in terms of health.*

Exercises (7 Codes)

Specific suggestions for possible exercises were also raised in the interviews by the participants (5/10, 50%), mostly taken from personal experiences that the interviewees had and considered helpful. In addition to frequently mentioned mindfulness and relaxation exercises, gratitude exercises and questions reflecting oneself and one's own past were suggested. Another interviewed participant emphasized the importance of open-ended questions, which can provide an impulse to reflect on personal issues and thereby gain insights for oneself:

> *[O]r relaxation techniques to switch off and no longer be trapped in this circle of thoughts, concrete things that are also easily accessible and where you can sit down briefly and say, "I'm now doing the meditation or the exercise."*

Subgroup Differences (3 Codes)

As it was one of the study's objectives to determine whether it is realistic and feasible to develop a single intervention targeting the entire LGBTQIA+ community with SAD disorders or specific content is needed for different subgroups (eg, gender identity vs sexual orientation), these questions were also explored in the participant interviews (3/10, 30%). Surprisingly, few relevant differences between subgroups emerged. The differences mentioned were only related to information needs. It was noted that, for example, transgender individuals will benefit from information about the transition process, whereas this information is not relevant for cisgender individuals, and appropriate customization options should be available in the online intervention. Given the approach that the goal of the online intervention was to reduce psychological symptoms, the project was not considered too ambitious or unrealistic. Psychological symptoms related to minority stress (eg, arising through the feeling of being different) were always highlighted as a unifying element:

> *Then every special path is of course different...whether it is homosexuality or trans is a huge difference and to have a gender matched*

XSL•FO
RenderX

surgically—there are worlds in between.... But first of all, this common basic thing I can picture well: How do you deal with it when you have the feeling...I am different?

Setting and Other Requirements

Table 2 presents a comprehensive summary of the principal components classified under the 2 remaining primary categories, namely *other requirements* and *setting*. It is noteworthy that there was a strong emphasis on the importance of accessibility concerning the way of delivering e–mental health interventions. In addition, there was a frequent desire for direct interaction with a therapist, including both synchronous (eg, video calls) and asynchronous methods (eg, chat). None of the participants stated that they did not want to use an online intervention due to concerns about data security and anonymity.

Table 2. Key aspects related to the setting and other requirements for an online intervention for individuals who are lesbian, gay, bisexual, transgender, queer or questioning, intersex, asexual, or have other sexual orientations and gender identities with symptoms of depression, anxiety, or somatoform disorders, as identified in the qualitative interviews.

Category and subcategory[a]	Results
Setting	
Therapeutic setting and guidance (n=38)	• Desire for direct interaction with therapists or contact persons through video or chat
Program functions (n=22)	• Frequent requests for the opportunity to connect with others, for example, through a forum or mentorship program • Requirement to maintain a clear individualized path, for example, with graphical elements
Accessibility (n=12)	• Importance of maintaining a low threshold and good program accessibility, such as providing free access due to financial disparities, considering neurodivergent individuals or people with disabilities • Preservation of participant anonymity as a crucial component
Design (n=4)	• Emphasis on a modern and professional design • Inclusion of elements reflecting diversity • Preference for a minimalistic design
Usability (n=3)	• Requirement for a simple and clearly structured user interface
Other requirements	
Concerns about the setting (n=22)	• Concerns regarding the preservation of participant anonymity and highlighting the need for a safe space • Apprehensions about the lack of alignment with real-life situations • Need for the introduction of novel insights; avoidance of reiteration of already-known information
Requirements regarding the online therapist (in case of a guided intervention; n=7)	• High qualifications of online therapists • Taking a non–cis-heteronormative approach in therapy • Proficiency in queer issues and familiarity with the associated terminology (ie, inclusive, validating, and sensitive language use; awareness for preassumptions; and reflecting trigger points)
Others (n=16)	• Clear management of participant expectations, detailing program capabilities and limitations • Consideration of comorbidities in therapy planning • Guidance on the process of finding suitable therapy providers

[a]The numbers in parentheses indicate the absolute number of codes in each subcategory.

Research Objective 2: The Perspective of HCPs and Counseling Staff

This analysis focused exclusively on the aspects categorized under the *content level* section, as these elements are of particular significance to the scope of this paper. Table 3 displays sample characteristics.

XSL•FO
RenderX

Table 3. Sociodemographic characteristics of the health care professionals (N=10).

Variable	Values
Age (y), mean (SD)	39.7 (8.75)
Gender and gender identity, n (%)	
Cisgender female	2 (20)
Cisgender male	3 (30)
Nonbinary and transmasculine	1 (10)
Nonbinary	1 (10)
Transgender male	2 (20)
Intersex female	1 (10)
Sexual and romantic orientation, n (%)	
Gay	3 (30)
Bisexual	1 (10)
Heterosexual	1 (10)
Homosexual and panromantic	1 (10)
Androphile	1 (10)
Pansexual and polyamorous	1 (10)
Queer (all genders)	1 (10)
Other	1 (10)
Occupational status, n (%)	
Employed	6 (60)
Civil servant and self-employed	2 (20)
Employed and self-employed	2 (20)

Across all subcategories, HCPs provided significantly more content-related suggestions. They proposed specific exercises and introduced therapeutic topics that should be considered in the online intervention.

With regard to *information on queer identity and support services* (16 codes; 6/10, 60%), HCPs commonly emphasized the significance of the minority stress model [8,9] in the conception of content for the online intervention. In addition, from the HCPs' perspective, information should be provided on learned internalized hostilities, shame, doubts, and uncertainties in dealing with one's own sexual or gender orientation. Many topics previously raised by patients were also mentioned by the HCPs, such as clarification of the connection between queer identity, being stereotyped through othering and mental health, psychoeducation, and the presentation of networking opportunities. One topic not previously mentioned by patients but by HCPs was the discussion of attachment and attachment traumas, as this was perceived as important in counseling and psychotherapy.

Specific desired contents (32 codes; 10/10, 100%) of the intervention, according to the HCPs, primarily focused on managing the coming-out process (both within immediate family and in work or other social contexts) and dealing with the rejection of close family members and the associated emotional injuries. Other desired content included self-reflection on one's life trajectory and significant life events, a general examination of one's social relationships (including reflection on one's role within the family system), and coping with the ongoing conflict between closeness and distance. Similar to the patients, HCPs also frequently addressed coping with issues such as bullying, discrimination, and loneliness.

While patients expressed a strong desire for *case stories*, this aspect was less prominent in the HCP (3 codes; 3/10, 30%) interviews. Although the HCPs did mention the use of personal experiences or case examples as a potential means to demonstrate to participants that they are not alone and illustrate the possibility of corrective experiences, this topic held less prominence compared to the patient interviews.

All interviewed HCPs (10/10, 100%) placed a great emphasis on the category of *psychoeducation* (31 codes) and mentioned a large number of relevant points. In addition to the topics frequently mentioned by both patients and HCPs (eg, bio-psychosocial explanatory model and explanation of mental health disorders and their development), physicality and connections between body and psyche were particularly emphasized from the HCPs' point of view.

Compared to the patients, the HCPs reported significantly more concrete ideas of suitable *exercises* (19 codes; 7/10, 70%) which, in their view, could be integrated into the online intervention. The exercises mentioned included exercises not only on self-esteem and self-acceptance but also social skills training,

XSL•FO
RenderX

resilience, and skills and self-determination in the transition process.

In view of *differences among the various subgroups within the community* (9 codes; 8/10, 80%), the group of intersex individuals and the group of transgender persons were mentioned as subgroups that could potentially differ from the other groups in terms of their needs for the planned online intervention.

Regarding *language* (16 codes; 6/10, 60%), HCPs stressed the need for the language used in the online intervention to be both considerate and mindful of gender sensitivity. The users' identities and orientations should be respected, and this inclusivity should be reflected in the language, exemplified by the respectful use of preferred pronouns. In addition, the language should avoid excessive medical jargon or unnecessary complexity. The correct use of terminology related to specific LGBTQIA+ issues was also considered crucial.

Discussion

Principal Findings

As outlined by Pachankis [30], there is a pronounced need for studies to investigate whether and how existing evidence-based interventions need to be adapted to provide a queer-sensitive e–mental health intervention. To the best of our knowledge, this participative study is the first to provide insights into the specific needs, from both a patient's and HCP's perspective, regarding an online intervention aimed at improving mental health in LGBTQIA+ adults. This is an important prerequisite for developing a specific online intervention to reach the marginalized target group and address the supply gap.

In the interviews, across all categories, a clear desire became evident to be represented, seen, and accepted and to feel good about oneself with one's gender identity and sexual orientation. This is consistent with scientific findings indicating that SGM individuals frequently encounter discriminatory experiences in their daily lives and within the health care system [6]. It also aligns with recent research highlighting that aware, sensitive behaviors as well as community-oriented strategies are key resilience factors for transgender, nonbinary, and gender nonconforming individuals in coping with minority-stress [31].

In line with the minority stress model [9], where social support is considered a significant protective factor against negative mental effects of minority stress, our study also revealed a desire to connect and exchange experiences with others facing similar challenges. Studies also demonstrate the significant importance of engaging with others when it comes to the mental health of SGM groups. For example, a systematic review by Valentine and Shipherd [32] demonstrated that social support was central in protecting against mental health distress among transgender and gender nonconforming people in the United States, and the absence of social support was associated with poor outcomes. A study by Jacmin-Park et al [33] revealed that during the COVID-19 pandemic, social support networks available for SGM individuals had a buffering effect against depression that was 4 times stronger than for cisgender heterosexual individuals. They concluded that fostering connections and supporting the development of so-called "families of choice" appeared to be even more significant for SGM individuals than for cisgender heterosexual individuals in coping with crisis periods. This is presumably because SGM individuals more frequently encounter challenging and discriminatory experiences in their everyday interpersonal relationships. The internet makes it easy to contact other people, so it is not surprising that, for example, social media can hold particular significance for SGM youth and serve as a space where they can rely on a supportive online community and develop a positive SGM identity [34]. From this perspective, it is crucial, especially in the development of an online platform for SGM individuals, to offer a means of mutual exchange.

A significant topic that emerged from the interviews, which has received relatively little attention in the literature so far, is the desire for an online resource providing clear and credible information. The interviewed individuals expressed a wish for information not only about being queer itself and related topics (eg, information about transition processes) but also about mental health in general (eg, the development of depression) and the connections between these 2 areas (eg, the minority stress model). Individuals sometimes feel overwhelmed during online searches or may even encounter harmful content reproducing discriminatory experiences. The results are in line with a study in which interviews were conducted with SGM adolescents experiencing at least mild anxiety or depressive symptoms, aiming to explore how online mental health help-seeking interventions could be designed [35]. Among the interviewed adolescents, a strong need for informational resources became evident. When considering the results of our study, it is noteworthy that this need persists into adulthood and does not seem to be limited to adolescents.

The overall results align with the findings by Steinke et al [36], who investigated the needs of SGM youth regarding digital health interventions through interviews and focus groups. Their study highlights the desire for finding a supportive and validating environment and connecting with others who share similar experiences and the need for relevant and accurate information. Particularly, these aspects were consistently also emphasized by the adults interviewed in this study, indicating that these needs persist throughout the life span of LGBTQIA+ individuals. While the study by Steinke et al [36] did not explicitly interview individuals from the community with mental disorder symptoms, it nonetheless underscored the overarching challenge of maintaining mental well-being in the face of the substantial minority stress experienced by these individuals.

In the sample of HCPs, it was noticeable that they predominantly belonged to the SGM group themselves. This raises the question of the extent to which the identity of the treating individuals being perceived as part of the SGM group plays a role. This may establish a "safe space" from the beginning, where the likelihood of experiencing discrimination is reduced, facilitating quicker formation of a trusting therapeutic alliance. This suggests that it could be important to firmly integrate individuals from the target group into the research team. Further research on this topic and its implications for the design of online interventions is needed.

XSL•FO
RenderX

The perspective of HCPs working with SGM individuals with mental health issues, as additionally captured in this study, largely showed consistent results with the perspective of the affected individuals. Interestingly, unlike the SGM individuals with mental health challenges who were interviewed, identity figures in terms of case reports integrated in the online intervention seemed to be of less importance among the HCPs. By contrast, the relevance of attachment and attachment traumas was a higher priority for HCPs than patients. This is consistent with the results of a recent study, which showed that in addition to anxiety symptoms, attachment style, in particular, was a significant predictor of quality of life in the LGBT community, and nonsecure attachment styles were more often found in LGBT people than in heterosexual individuals [37].

The themes highlighted by HCPs partly align with the results of an interview study by Clark et al [38] in which mental health priorities for research were reported, for example, the importance of intimate relationships, stressors specific to SGM status, depression, and support for emotional resilience and well-being.

Regarding the differences between subgroups within the LGBTQIA+ community, only a few points relevant to an online intervention were addressed by the 2 surveyed groups. The intersex group is often neglected and not considered in corresponding offers, while transgender individuals who wish to undergo a gender reassignment surgery frequently face very specific challenges (eg, burdens in the transition process). Nevertheless, the overall intention to develop an intervention for all subgroups was reinforced, as there are many common and connecting elements in the psychological experiences across different subgroups. Therefore, it is crucial for the online intervention to focus on common psychological symptoms rather than subgroup affiliation.

Strengths and Limitations

To the best of our knowledge, our study is the first to specifically examine the perspectives of SGM adults with mental health challenges and those treating this target group regarding a potential online intervention using a participative approach. The overlapping yet sometimes complementary perspectives overall provide a comprehensive overview from different angles on the aspects that should be integrated into a corresponding online intervention to improve mental health. The interviewed individuals represent a wide spectrum of different sexual and romantic orientations as well as gender identities. A significant strength of this qualitative study is its participatory design, which goes beyond user-centered design and includes end users as active participants in all phases of the research and development process. As recommended, this study used a participatory approach from the outset [39], involving representatives from various German LGBTQIA+ organizations through the LGBTQIA+ advisory board. In addition, perspectives from individuals representing the future target audience (SGM individuals with psychological symptoms) and HCPs working with the target audience were captured in the early development phase of the intervention. Through the collaborative development of the interview guide and planning of study procedures with representatives from various LGBTQIA+ organizations and associations, it was ensured that all important questions were addressed in the interview and that the interviewed individuals felt safe during their participation. Feedback to the study team indicated that all points of contact with the study team during the study were perceived as a safe space for the participants. This sensitive approach was a prerequisite for not reproducing discriminatory experiences within the study but instead creating a secure atmosphere that allowed participants to openly discuss very personal and sometimes deeply distressing experiences during the interviews. By contrast, the collaboration with the LGBTQIA+ advisory board enabled the contextualization of quotes, providing significant added value to the interpretation of the results. Another strength of the study is that we were able to recruit interviewees from different subgroups of the LGBTQIA+ community for the study, which sets us apart from other studies that, for example, often do not take the perspective of intersex people into account [40]. Of course, it cannot be assumed that only one or a few participants speak for the entire subgroup to which they belong, as they can only report on their own personal experiences. Nevertheless, the diverse selection of study participants is an important prerequisite for the diversity of perspectives in the results.

There are also limitations to this study, primarily attributable to its qualitative nature. First, as is typical for qualitative studies, our sample is small and nonrepresentative due to self-selection. However, qualitative studies, which are less limited by resources, should aim for a precalculated larger sample size (eg, using the saturation calculation principle) and more representatives per subgroup, as is the case in this study [41]. Larger studies are necessary to conduct subgroup analyses and potentially highlight the different needs of various subgroups more effectively. Second, the severity of SAD symptoms was not assessed for study inclusion, and no external criteria beyond self-assessment were applied. This approach aimed to include both individuals currently experiencing symptoms and those with a history of such symptoms, allowing reflection on past support needs. However, this limits our ability to determine the actual severity of participants' symptoms, further compounded by the lack of a formal diagnosis requirement. This decision was made due to the significant gap in undiagnosed LGBTQIA+ individuals who nevertheless experience relevant symptomatology. Therefore, it is important to consider that the SAD symptoms are self-reported when interpreting the data. To minimize the imposition of predefined categories, participants were asked about their self-reported labels for gender identity and sexual orientation. While this approach complicates specific analyses of individual subgroups, it simultaneously emphasizes the individuality of each participant. Intersectionality (belonging to multiple minority groups within the LGBTQIA+ community or based on religion, ethnicity, socioeconomic status, etc) was not explicitly considered in this study, as is the case in many studies on online interventions—a limitation that has been highlighted in a recently published review on the extent of intersectionality in SGM interventions [42]. In agreement with the authors of the aforementioned study [42], we can state that guidelines for clinical practice and evaluation are needed to address intersectionality within interventions for SGM.

XSL•FO
RenderX

Conclusions

We argue that a queer and intersectional discrimination-sensitive approach to e–mental health care is needed, as the increased risk of psychological distress and illness can be linked to experiences of discrimination and violence based on sexual orientation and gender identity. At the same time, there are several barriers to accessing health care services. Because LGBTQIA+ individuals are a group that is open to digital media and digital media is used for networking and information, digital treatment options are particularly suitable for this population. When developing e–mental health interventions for specific target groups, it is important to include both those immediately affected and professionals working with the population to obtain an intervention that appeals to the target group and is likely to be used. The results of our qualitative participative study serve as a framework for an intervention concept. Besides topics that emerged as crucial for the planned intervention (eg, understanding symptoms in the context of minority stress and specific social challenges and mental needs of LGBTQIA+ individuals), this qualitative study highlighted the strong need for identification figures and the wish for direct interaction with therapists or contact persons within or paralleling the online intervention.

Acknowledgments

The authors thank the representatives of the following associations and organizations participating in the advisory board for their involvement in the project: Association for lesbian, gay, bisexual, trans*, intersexual and queer professionals in psychology (VLSP*); LGBT*IQ Coordination Office Wiesbaden; Women's Center Mainz; German Association for Transidentity and Intersexuality (dgti); German AIDS Service Organization; QueerNet Rheinland-Pfalz; Bundesverband Trans* (Federal Trans* Union); Intergeschlechtliche Menschen eV (Association of intersex individuals); Schwuguntia; and Queer Center Erfurt. They would like to express their sincere gratitude for the valuable contributions of the aforementioned organizations to the study design, collaboration on the interview guides, and support in interpreting qualitative data. The authors thank Clara Pering for assisting in data collection and for the ongoing support in conducting the study. In addition, the authors thank the research assistants Lou Doré, Juliane Grüner, and Lea Tasyuerek. Finally, the authors express their appreciation to all interviewed individuals for their willingness to engage in these personal interviews with the researchers.

This work was supported by the Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung; grants 01KG2212A and 01KG2212B). The funding sources had no role in the design, conduct, and reporting of this study or in the decision to submit the manuscript for publication.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

AT and AG were involved in conceptualization, writing the original draft, and methodology. RZ, BS, and MEB were involved in supervision. AT, AG, and MH were involved in formal analysis, conducting interviews and focus groups, and coding interview transcripts. MEB, BS, RZ, and MH were involved in writing the original draft. FJ, LRK, and AK were involved in reviewing and editing the manuscript. AT, AG, and RZ were involved in the investigation and methodology. RZ was involved in project administration.

Conflicts of Interest

None declared.

Multimedia Appendix 1
COREQ checklist.
[PDF File (Adobe PDF File), 281 KB - formative_v9i1e65834_app1.pdf]

References

1. Hutfless E. Zur Kritik der Heteronormativität - Queere Perspektiven in der Psychotherapie. Psychotherapie Forum 2019 May;23(1-2):44-50. [doi: 10.1007/s00729-019-0116-z]
2. Pöge K, Dennert G, Koppe U, Güldenring A, Matthigack EB, Rommel A. The health of lesbian, gay, bisexual, transgender and intersex people. J Health Monit 2020 Mar;5(Suppl 1):2-27 [FREE Full text] [doi: 10.25646/6449] [Medline: 35146279]
3. McNamara G, Wilson C. Lesbian, gay and bisexual individuals experience of mental health services - a systematic review. J Ment Health Train Educ Pract 2020 Jan;15(2):59-70. [doi: 10.1108/jmhtep-09-2019-0047]
4. Mitchell LA, Jacobs C, McEwen A. (In)visibility of LGBTQIA+ people and relationships in healthcare: a scoping review. Patient Educ Couns 2023 Sep;114:107828. [doi: 10.1016/j.pec.2023.107828] [Medline: 37301011]

XSL•FO
RenderX

5. Schürmann-Vengels J. Lebensrealitäten und Behandlungserfahrungen von LGBTQ + Personen. Psychotherapie Dialog 2025 Feb 25;26(01):64-67. [doi: 10.1055/a-2289-5874]
6. Ayhan CH, Bilgin H, Uluman OT, Sukut O, Yilmaz S, Buzlu S. A systematic review of the discrimination against sexual and gender minority in health care settings. Int J Health Serv 2020 Jan;50(1):44-61. [doi: 10.1177/0020731419885093] [Medline: 31684808]
7. Moagi MM, van Der Wath AE, Jiyane PM, Rikhotso RS. Mental health challenges of lesbian, gay, bisexual and transgender people: an integrated literature review. Health SA 2021;26:1487 [FREE Full text] [doi: 10.4102/hsag.v26i0.1487] [Medline: 33604059]
8. Pagliaccio D. Mental health disparities among sexual and gender minority students in higher education. J Am Coll Health 2024 Nov 12;73(5):1951-1962. [doi: 10.1080/07448481.2024.2404944] [Medline: 39533453]
9. Meyer IH. Prejudice, social stress, and mental health in lesbian, gay, and bisexual populations: conceptual issues and research evidence. Psychol Bull 2003 Sep;129(5):674-697 [FREE Full text] [doi: 10.1037/0033-2909.129.5.674] [Medline: 12956539]
10. Zeeman L, Sherriff N, Browne K, McGlynn N, Mirandola M, Gios L, et al. A review of lesbian, gay, bisexual, trans and intersex (LGBTI) health and healthcare inequalities. Eur J Public Health 2019 Oct;29(5):974-980 [FREE Full text] [doi: 10.1093/eurpub/cky226] [Medline: 30380045]
11. Aronson J, Burgess D, Phelan SM, Juarez L. Unhealthy interactions: the role of stereotype threat in health disparities. Am J Public Health 2013 Jan;103(1):50-56 [FREE Full text] [doi: 10.2105/AJPH.2012.300828] [Medline: 23153125]
12. Ojeda-Leitner D, Lewis RK. Assessing health-related stereotype threats and mental healthcare experiences among a LGBT sample. J Prev Interv Community 2019 Sep;49(3):251-265. [doi: 10.1080/10852352.2019.1654262] [Medline: 31549929]
13. Russell ST, Fish JN. Mental health in lesbian, gay, bisexual, and transgender (LGBT) youth. Annu Rev Clin Psychol 2016;12:465-487 [FREE Full text] [doi: 10.1146/annurev-clinpsy-021815-093153] [Medline: 26772206]
14. Kasprowski D, Fischer M, Chen X, de Vries L, Kroh M, Kühne S, et al. Geringere Chancen auf ein gesundes Leben für LGBTQI*-Menschen. DIW Wochenbericht. 2021. URL: https://www.diw.de/de/diw_01.c.810358.de/publikationen/wochenberichte/2021_06_1/geringere_chancen_auf_ein_gesundes_leben_fuer__lgbtqi_-menschen.html [accessed 2025-07-19]
15. National survey on LGBTQ youth mental health. The Trevor Project. 2019. URL: https://www.thetrevorproject.org/survey-2019/ [accessed 2025-07-19]
16. Skerrett DM, Kõlves K, De Leo D. Are LGBT populations at a higher risk for suicidal behaviors in Australia? Research findings and implications. J Homosex 2015;62(7):883-901. [doi: 10.1080/00918369.2014.1003009] [Medline: 25569508]
17. Skerrett DM, Kõlves K, De Leo D. Factors related to suicide in LGBT populations. Crisis 2016 Sep;37(5):361-369. [doi: 10.1027/0227-5910/a000423] [Medline: 27659515]
18. Gieselmann A, Böckermann M, Pietrowsky R. Internetbasierte Gesundheitsinterventionen. Psychotherapeut 2015 Jun;60(5):433-440. [doi: 10.1007/s00278-015-0038-3]
19. Andersson G, Cuijpers P, Carlbring P, Riper H, Hedman E. Guided internet-based vs. face-to-face cognitive behavior therapy for psychiatric and somatic disorders: a systematic review and meta-analysis. World Psychiatry 2014 Oct;13(3):288-295 [FREE Full text] [doi: 10.1002/wps.20151] [Medline: 25273302]
20. Andersson G, Titov N, Dear BF, Rozental A, Carlbring P. Internet-delivered psychological treatments: from innovation to implementation. World Psychiatry 2019 Feb;18(1):20-28 [FREE Full text] [doi: 10.1002/wps.20610] [Medline: 30600624]
21. Jia RM, Du JT, Zhao YC. Characteristics of the health information seeking behavior of LGBTQ+ individuals: a systematic review on information types, information sources and influencing factors. J Doc 2021 Jul;78(2):361-388. [doi: 10.1108/jd-03-2021-0069]
22. Pachankis JE, Williams SL, Behari K, Job S, McConocha EM, Chaudoir SR. Brief online interventions for LGBTQ young adult mental and behavioral health: a randomized controlled trial in a high-stigma, low-resource context. J Consult Clin Psychol 2020 May;88(5):429-444 [FREE Full text] [doi: 10.1037/ccp0000497] [Medline: 32271053]
23. Dominguez-Rodriguez A, De La Rosa-Gómez A. A perspective on how user-centered design could improve the impact of self-applied psychological interventions in low- or middle-income countries in Latin America. Front Digit Health 2022;4:866155 [FREE Full text] [doi: 10.3389/fdgth.2022.866155] [Medline: 35721795]
24. Vial S, Boudhraâ S, Dumont M. Human-centered design approaches in digital mental health interventions: exploratory mapping review. JMIR Ment Health 2022 Jun;9(6):e35591 [FREE Full text] [doi: 10.2196/35591] [Medline: 35671081]
25. Moore G, Wilding H, Gray K, Castle D. Participatory methods to engage health service users in the development of electronic health resources: systematic review. J Particip Med 2019 Feb;11(1):e11474 [FREE Full text] [doi: 10.2196/11474] [Medline: 33055069]
26. Behr S, Fenski F, Boettcher J, Knaevelsrud C, Hammelrath L, Kovacs G, et al. TONI - One for all? Participatory development of a transtheoretic and transdiagnostic online intervention for blended care. Internet Interv 2024 Mar;35:100723 [FREE Full text] [doi: 10.1016/j.invent.2024.100723] [Medline: 38370289]
27. Rustage K, Crawshaw A, Majeed-Hajaj S, Deal A, Nellums L, Ciftci Y, et al. Participatory approaches in the development of health interventions for migrants: a systematic review. BMJ Open 2021 Oct;11(10):e053678 [FREE Full text] [doi: 10.1136/bmjopen-2021-053678] [Medline: 34697122]

XSL•FO
RenderX

28. Phillips EA, Himmler SF, Schreyögg J. Preferences for e-mental health interventions in Germany: a discrete choice experiment. Value Health 2021 Mar;24(3):421-430 [FREE Full text] [doi: 10.1016/j.jval.2020.09.018] [Medline: 33641777]
29. Kegel F, Greve M, Schnell K, Stein M, Kolbe LM. Co-creative Vermittlungsplattformen für die psychische Gesundheitsversorgung. HMD Prax Wirtsch Inform 2022 Nov;59(6):1513-1526 [FREE Full text] [doi: 10.1365/s40702-022-00925-1] [Medline: 40478086]
30. Pachankis JE. The scientific pursuit of sexual and gender minority mental health treatments: toward evidence-based affirmative practice. Am Psychol 2018 Dec;73(9):1207-1219 [FREE Full text] [doi: 10.1037/amp0000357] [Medline: 30525805]
31. Läufer P. Genderbezogener Stress: Befunde einer empirisch quantitativen Studie zur Weiterentwicklung eines Messinstruments für Coping mit Minderheitenstress. In: Langer PC, O'Donnoko N, editors. Transgeschlechtlichkeit und Psychoanalyse. Bielefeld, Germany: Transcript Verlag; 2024.
32. Valentine SE, Shipherd JC. A systematic review of social stress and mental health among transgender and gender non-conforming people in the United States. Clin Psychol Rev 2018 Dec;66:24-38 [FREE Full text] [doi: 10.1016/j.cpr.2018.03.003] [Medline: 29627104]
33. Jacmin-Park S, Rossi M, Dumont L, Lupien SJ, Juster RP. Mental health and social support of sexual and gender diverse people from Québec, Canada during the COVID-19 crisis. LGBT Health 2022 Apr;9(3):151-160. [doi: 10.1089/lgbt.2021.0255] [Medline: 35290754]
34. Brinkman L, Francot R. Developing a resilient sexual and gender minority identity online: the importance of social media for youth before coming out. In: Woodman D, editor. LGBT+ Communities: Creating Spaces of Identity. London, UK: IntechOpen; 2023.
35. Smith CO, Wecht S, Odenthal K, Escobar-Viera CG, Radovic A. Optimizing online tools to support sexual and gender minority youth with depressive or anxiety symptoms: qualitative study. J Adolesc Health 2023 Mar;72(3):S65-S66. [doi: 10.1016/j.jadohealth.2022.11.135]
36. Steinke J, Root-Bowman M, Estabrook S, Levine DS, Kantor LM. Meeting the needs of sexual and gender minority youth: formative research on potential digital health interventions. J Adolesc Health 2017 May;60(5):541-548 [FREE Full text] [doi: 10.1016/j.jadohealth.2016.11.023] [Medline: 28108088]
37. Kardasz Z, Gerymski R, Parker A. Anxiety, attachment styles and life satisfaction in the polish LGBTQ+ community. Int J Environ Res Public Health 2023 Jul;20(14):6392 [FREE Full text] [doi: 10.3390/ijerph20146392] [Medline: 37510624]
38. Clark KD, Capriotti MR, Obedin-Maliver J, Lunn MR, Lubensky ME, Flentje A. Supporting sexual and gender minority health: research priorities from mental health professionals. J Gay Lesbian Ment Health 2019 Dec 17;24(2):205-221. [doi: 10.1080/19359705.2019.1700865]
39. Noergaard B, Sandvei M, Rottmann N, Johannessen H, Wiil U, Schmidt T, et al. Development of a web-based health care intervention for patients with heart disease: lessons learned from a participatory design study. JMIR Res Protoc 2017 May;6(5):e75. [doi: 10.2196/resprot.7084] [Medline: 28526674]
40. MacKenzie D, Huntington A, Gilmour JA. The experiences of people with an intersex condition: a journey from silence to voice. J Clin Nurs 2009 Jun;18(12):1775-1783. [doi: 10.1111/j.1365-2702.2008.02710.x] [Medline: 19646120]
41. Guest G, Namey E, Chen M. A simple method to assess and report thematic saturation in qualitative research. PLoS One 2020;15(5):e0232076 [FREE Full text] [doi: 10.1371/journal.pone.0232076] [Medline: 32369511]
42. Huang YT, Ma YT, Craig SL, Wong DF, Forth MW. How intersectional are mental health interventions for sexual minority people? A systematic review. LGBT Health 2020 Jul;7(5):220-236. [doi: 10.1089/lgbt.2019.0328] [Medline: 32412864]

Abbreviations

CBT: cognitive behavioral therapy
COREQ: Consolidated Criteria for Reporting Qualitative Research
Queer-EMH: Queer-E-Mental Health
HCP: health care professional
LGBT: lesbian, gay, bisexual, and transgender
LGBTQIA+: lesbian, gay, bisexual, transgender, queer or questioning, intersex, asexual, and other sexual orientations and gender identities
SAD: somatoform, anxiety, and depressive
SGM: sexual and gender minority

XSL•FO
RenderX

Edited by A Mavragani; submitted 30.08.24; peer-reviewed by C Grov, K Nwosu, NJ Goel; comments to author 26.01.25; revised version received 31.03.25; accepted 25.05.25; published 12.09.25.

Please cite as:
Gomes A, Tsiouris A, Jung F, Klein LR, Kreis A, Beutel ME, Strauss B, Hoy M, Zwerenz R
Needs and Expectations Associated With an e–Mental Health Intervention for Reducing Somatoform, Anxiety, and Depressive Symptoms in Sexual and Gender Minority Adults: Qualitative Participative Study
JMIR Form Res 2025;9:e65834
URL: https://formative.jmir.org/2025/1/e65834
doi:10.2196/65834
PMID:40938641

©Anna Gomes, Angeliki Tsiouris, Florian Jung, Laura Rebecca Klein, Adina Kreis, Manfred E Beutel, Bernhard Strauss, Madita Hoy, Rüdiger Zwerenz. Originally published in JMIR Formative Research (https://formative.jmir.org), 12.09.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Semaglutide and Tirzepatide in a Remote Weight Management Program: 12-Month Retrospective Observational Study

Rebecca Richards[1], CPsychol, PhD; William Lunt[1], BSc; Michael Whitman[1], BSc; Giulia Spaltro[1], CPsychol; Rachel Hall[1], BMBCh, MA, MSc

Second Nature, London, United Kingdom

Corresponding Author:
Rebecca Richards, CPsychol, PhD
Second Nature
483 Green Lanes
London, N13 4BS
United Kingdom
Phone: 44 2034880769
Email: becky@secondnature.io

Abstract

Background: Obesity affects >890 million adults worldwide, and traditional lifestyle interventions often lack long-term success. While glucagonlike peptide-1 receptor agonists (GLP-1RAs) have shown strong weight loss outcomes, access to specialist care is limited by cost and capacity.

Objective: This study evaluated the effectiveness, feasibility, acceptability, and potential cost-effectiveness of a 12-month remote GLP-1RA–supported weight management program, comparing outcomes between tirzepatide and semaglutide.

Methods: This retrospective analysis included 339 participants (n=278, 82% women) who completed a 12-month remote weight management program using either tirzepatide (n=209, 61.7%) or semaglutide (n=130, 38.3%) between February and June 2024. The program combined medication, app-based behavioral support, coaching from registered dietitians and nutritionists, and clinical oversight. It featured 5 phases with evidence-based behavior change techniques, monthly monitoring, and safety protocols. Primary outcomes were mean weight change and proportions achieving ≥10% and ≥15% weight loss. Secondary outcomes included behavior changes, side effects, acceptability, feasibility, and estimated cost-effectiveness compared to National Health Service care.

Results: Mean weight change at 12 months was −22.9 kg (–22.1% of baseline weight, SD 8%; $P<.001$) in the tirzepatide cohort and −18.1 kg (–17.1% of baseline weight, SD 8.1%; $P<.001$) in the semaglutide cohort. Achievement of ≥10% weight loss occurred in 95.2% (199/209) of participants using tirzepatide and 83.1% (108/130) of participants using semaglutide, whereas ≥15% weight loss was achieved by 83.7% (175/209) and 56.2% (73/130) of the participants, respectively. The proportion of inactive participants (no weekly exercise) decreased substantially in both cohorts (tirzepatide: 31/209, 14.8% to 14/209, 6.7%; semaglutide: 29/130, 22.3% to 7/130, 5.4%; $P<.001$). Side effects decreased significantly over the 12-month period, with participants who reported no side effects increasing from 41.6% (87/209) to 60.3% (126/209; $P<.001$) in the tirzepatide cohort and from 53.8% (70/130) to 67.7% (88/130) in the semaglutide cohort ($P=.02$), whereas common initial side effects, including constipation, nausea, and fatigue, showed significant reductions ($P<.001$). Economic modeling suggested a 60% to 70% cost saving compared to specialist weight management services and a 10% to 60% cost saving compared to primary care in the National Health Service.

Conclusions: This real-world evaluation demonstrates that remotely delivered, GLP-1RA–supported weight management programs can achieve weight loss outcomes that align closely with clinical trial results while potentially reducing health care costs by 10% to 70% compared to traditional UK services. Both the tirzepatide and semaglutide cohorts exceeded clinically significant weight loss thresholds with acceptable safety profiles and positive behavior changes. These findings support the feasibility and effectiveness of digital delivery models for expanding access to specialist obesity treatment within resource-constrained health care systems, with outcomes that compare favorably to pharmacological intervention alone.

(JMIR Form Res 2025;9:e81912) doi:10.2196/81912

XSL•FO
RenderX

KEYWORDS

obesity; weight management; glucagonlike peptide-1 receptor agonist; GLP-1RA; semaglutide; digital health; remote health care delivery; telemedicine; behavioral intervention; weight loss

Introduction

Background

Worldwide, >2.5 billion adults are overweight. Of this number, >890 million, or approximately 1 in 10 adults, are living with obesity [1]. Obesity is defined as a chronic, systemic illness that is underpinned by excess adiposity that can lead to severe organ dysfunction and life-threatening complications (eg, stroke, cardiovascular disease, type 2 diabetes, depression, and certain types of cancer) [1,2]. In the United Kingdom, obesity is expected to cost the National Health Service (NHS) nearly £12 billion (US $16.1 billion) in 2025, with total costs (NHS, individual, and wider societal) projected to reach >£73 billion (US $98.1 billion) [3,4].

While lifestyle interventions have historically formed the foundation of obesity treatment, their long-term effectiveness varies considerably, with more than half of lost weight typically regained within 2 years [5]. Following the success of glucagonlike peptide-1 receptor agonists (GLP-1RAs) in the treatment of type 2 diabetes mellitus, the Semaglutide Treatment Effect in People With Obesity (STEP) trials have demonstrated significant potential for this medication in weight management [6,7]. Subsequent trials of semaglutide, tirzepatide, and retratutide have demonstrated weight reductions of 14.9%, 20.9%, and 24.2% over 68 to 88 weeks, respectively. However, significant weight regain occurs following treatment discontinuation [8-11]. In both the STEP-1 and STEP-4 (semaglutide) trials and the SURMOUNT-4 (tirzepatide) trial, participants regained approximately two-thirds of the lost weight within 1 year of discontinuing treatment [9,12,13]. To improve efficacy in obesity care and reduce the risk of weight regain, the National Institute for Health and Care Excellence (NICE) recommended tirzepatide for indefinite use in obesity management, in conjunction with lifestyle interventions (NICE Technology Appraisal TA1026) [14].

In line with the NICE Technology Appraisal TA1026, the NHS Long Term Plan, and NHS England's interim commissioning guidance for tirzepatide implementation [14-16], there is growing pressure on primary care, as well as in specialist weight management services (SWMSs), to prescribe GLP-1RAs. Combined with existing waiting lists, limited budgets, and reduced staffing, both primary and secondary care have struggled to meet the mandated 90- and 180-day NICE implementation timelines [16]. NHS England is commissioning a national pilot Behavioural Support for Obesity Prescribing service to support tirzepatide prescribing—as mandated by the NICE guidelines—but the impact and utility of this service vary significantly among integrated care boards (ICBs) [16].

Evidence demonstrates that behavior change, alongside GLP-1RA medication, delivers greater weight loss over a 12-month period than medication alone [17]. Combined behavioral and pharmacological approaches have achieved a 19.1% reduction in body weight as compared to the 14.9% observed in pharmacological clinical trials [8,17]. Furthermore, initial UK-based cost modeling suggests that remotely delivered services are more scalable and significantly less expensive when delivered in primary care [14]. Following recent commissioning guidance changes, the evaluation of how remote services integrate with existing weight management services to deliver tirzepatide is essential.

Objectives

This 12-month retrospective observational analysis examined the effectiveness, feasibility, acceptability, and potential cost-effectiveness of Second Nature's remotely delivered weight management program supported by GLP-1RAs. This study also aimed to compare the outcomes of semaglutide and tirzepatide within a remote care setting.

Methods

For the purpose of reporting this evaluation, we have followed and used NHS England's evaluation toolkit checklist [18].

Participants

This service evaluation was conducted within Second Nature's commercial digital health platform, which operates as a private weight management service (independent of NHS infrastructure) in the United Kingdom. Participants were recruited via various digital marketing channels (eg, Google, Facebook, and Instagram) or through word-of-mouth referrals. No specific demographic or geographic parameter was targeted. All participants were enrolled via self-referral completing online eligibility questionnaires via the Second Nature website and app.

In July 2025, retrospective data were obtained for all participants who (1) began either the semaglutide- or tirzepatide-supported intervention during the period from February 2024 to June 2024; (2) completed the intervention; and (3) provided responses to the 1-, 3-, 6-, and 12-month surveys. All participants paid between £229 (US $310) and £299 (US $400) per month depending on the dose of tirzepatide or semaglutide prescribed.

Adults aged 18 to 75 years with weight-related comorbidities were eligible, with BMI thresholds of ≥35 kg/m^2 for tirzepatide and ≥30 kg/m^2 for semaglutide (tier 3 referral criteria for BMI: 30.0-34.9 kg/m^2). BMI thresholds were reduced by 2.5 kg/m^2 for participants from South Asian, Chinese, other Asian, Middle Eastern, Black African, or African-Caribbean backgrounds [14,19].

Users had to possess access and demonstrate capability to use smartphone or tablet devices to be considered eligible for participation. Several exclusion criteria for both medications were established: individuals with current or previous eating disorder diagnoses; those who were pregnant, nursing, or actively seeking to become pregnant; participants with known allergies to any medication components or excipients; and those

XSL•FO
RenderX

currently using specific medications (including diabetes treatments or drugs associated with the aforementioned exclusion conditions). Further exclusions applied to individuals presenting with any of the following medical conditions: thyroid cancer (current or historical), active malignancy, inflammatory bowel conditions (including Crohn's disease and ulcerative colitis), celiac disease, chronic malabsorption syndrome, pancreatitis, hepatic impairment, renal disease, cardiac failure, multiple endocrine neoplasia type 2, gallbladder disorders, or diabetic retinopathy.

As we have previously reported attrition—which, for a private intervention, is usually heavily weighted toward costs [17]—we wanted to focus on showing real-world effectiveness and medication tolerability and attaining the largest sample size for the most accurate effect size analysis. To do this, we self-selected participants who both completed the program and provided survey responses at months 1, 3, 6, and 12.

Intervention

This intervention builds on the previously described Second Nature digital program, which was developed by an in-house multidisciplinary team following NICE guidance for obesity management and behavior change principles [14,17,20-23]. The program consists of 5 phases (prepare, adapt, grow, progress, and momentum) spanning ≥68 weeks, delivering evidence-based behavior change techniques through daily educational content covering nutrition, medication management, sleep, mental well-being, and physical activity [24]. The web and mobile platform integrates self-monitoring tools for weight, sleep, and activity tracking via connected devices, whereas health coaches provide personalized support through one-to-one messaging, establish individualized weight and activity goals, and proactively engage participants who have not accessed the app for ≥4 days. Participants receive automated feedback through weekly progress graphs and health change summaries, with mandatory monthly surveys and weight measurements prompted by the app's notification system.

Medication

Participants chose between semaglutide or tirzepatide at enrollment, with medication-specific protocols and educational resources provided through the Second Nature app. Tirzepatide users received more intensive monitoring (weekly surveys and 3-5–day pharmacist intervention) compared to semaglutide users (monthly monitoring and 7-day intervention) due to increased side effect profiles [25]. Safety monitoring included monthly pharmacist-reviewed digital surveys, Bluetooth-connected weight tracking, and a tiered adverse event protocol with health coach response within 24 hours and pharmacist escalation within 4 hours for urgent concerns.

Effectiveness Measures

The primary outcome was the mean weight change, recorded between days 330 and 420 (approximately months 11 to 13) after program initiation, assessed in both kilograms and as a percentage of baseline weight. Secondary measures included the proportion of participants achieving ≥10% and ≥15% weight loss, along with monthly weight change data corresponding to each survey point for both pharmacological interventions.

Participants were provided with wireless weighing scales to share weight data with Second Nature. To ensure accuracy, they were instructed to weigh themselves on a firm, flat surface in the morning after using the restroom and at consistent weekly intervals. A validation algorithm accepted only readings within defined thresholds based on previous measurements and the time elapsed since the last entry, helping filter out anomalies and monitor for sudden changes in weight. If irregular readings were detected—such as when someone else in the household used the scale—participants received email notifications, and the data were not stored.

Cost Measures

We have previously reported the cost-effectiveness of a semaglutide-supported intervention compared to NHS SWMSs [17]. Given that tirzepatide is now recommended for use in primary care and specialist service settings, we conducted a cost comparison of the tirzepatide-supported intervention against both NHS SWMSs and nonspecialist primary care weight management services. Cost data for both services were obtained from NICE Technology Appraisal TA1026 resource impact templates [14].

A range of confidential costs for Second Nature's program were used, including the lower-intensity, lower-cost private service (evaluated in this study) and a higher-intensity model assuming that NHS delivery would require delivery to more complex patients based on the NHS interim commissioning guidance (BMI of >40 kg/m^2 and 4 comorbidities for the phase 1 cohort) [14]. To maintain commercial confidentiality, costs were presented as percentage reductions rounded to 10% intervals using relative cost methodology. We computed percentage variations between the intervention and primary and secondary care standard approaches outlined in the NICE health technology evaluation 14 and tirzepatide Technology Appraisal documents [14,26]. All costs were standardized to per-participant amounts over a 12-month treatment period. The analysis focused exclusively on direct intervention costs excluding potential long-term health benefit savings.

Cost comparisons were conducted using simple cost analysis between the tirzepatide-supported intervention and both NHS standard care models (SWMSs and nonspecialist primary care). Key parameters included 6.2% total eligibility (NICE default), 30% digital service uptake, and 75% medication initiation rate based on internal data from Second Nature's NHS services [17,26]. On the basis of these parameters, 1395 participants per 100,000 population were expected to enroll. Percentage cost differences were calculated as ([comparator cost − intervention cost]/comparator cost) × 100 to determine potential savings. Real-world GBP costs were calculated as number of participants (1395) × lower-range cost and number of participants (1395) × higher-range cost to provide an expected cost of delivery. This expected cost of delivery was then multiplied by the expected cost saving percentage (highest range number × lowest cost saving and lowest range number × highest cost saving) to provide an expected saving in GBP.

XSL•FO
RenderX

Behavior Changes

Web-based surveys administered at months 1, 3, 6, and 12 of the program assessed lifestyle-related behavioral modifications, encompassing cooking frequency, fruit and vegetable consumption, and physical activity levels.

Meal preparation behaviors were examined using the following question—"How many times do you cook each week on average?"—with responses documented as weekly cooking session numbers.

Daily vegetable and fruit consumption was evaluated using the following question—"How many different vegetables or fruits do you eat a day on average?"—with numerical responses indicating the average number of distinct items consumed daily.

Physical activity assessment used the following question—"How many times do you exercise each week on average?"—requiring participants to report their average weekly exercise session frequency.

Acceptability: Participant Experience

Free-text responses to the monthly survey question—"How have you found the last few weeks of the program?"—underwent content analysis. MW and WL coded and organized responses into higher-level categories for each question at every time point. These analytical frameworks underwent refinement following protocols established in previous research, resulting in 9 higher-level categories that encompassed key themes within participants' experiences [21]. To capture the complexity of participant experiences, individual responses could be classified across multiple themes. For instance, a participant's free-text response of *Ok* to "How have you found the last few weeks on the program?" would be categorized as a *neutral experience*.

Feasibility: Side Effects

Mandatory web-based check-in surveys were emailed to participants at 1, 3, 6, and 12 months to gather medication side effect information. A drop-down menu allowed for the selection from 15 possible side effects, *None*, or *Other*. Selecting *Other* required a mandatory free-text description.

Statistical Analysis

Analyses were conducted using the R statistical software (R Foundation for Statistical Computing). For effectiveness, we calculated mean weight change and SD and evaluated statistical significance using paired 2-tailed *t* tests ($P<.05$) plus proportions achieving ≥10% and ≥15% weight loss thresholds. Statistical analysis included descriptive statistics for baseline characteristics and primary weight change outcomes.

Monthly behavior changes, side effects, and participant experiences were analyzed as proportions of total survey respondents. For each behavioral subcategory, we conducted 2-proportion *z* tests comparing the proportion of participants in that specific category at each follow-up time point (months 3, 6, and 12) against the baseline proportion (month 1). Separate analyses were conducted for each medication group (tirzepatide and semaglutide). Given the multiple subcategory comparisons within each behavioral domain, we applied the Bonferroni correction to control the familywise error rate.

Ethical Considerations

This evaluation met Health Research Authority and Medical Research Council service evaluation criteria [26]—it assessed current care using existing, anonymized data from an established intervention; did not involve any change to standard practice; and did not involve randomization or a control group. The Health Research Authority explicitly states that service evaluation is considered part of usual professional practice and is exempt from research ethics committee oversight [27]. This study analyzed anonymized, non-NHS data that could not identify individuals, thus not requiring ethics approval (Multimedia Appendix 1); however, the data were not initially collected with a view to use them in this research. Participants were not compensated for taking part and consented to research use of anonymized data through acceptance of the privacy policy at registration. This evaluation adhered to the General Data Protection Regulation, maintaining participants' right to request data deletion.

Results

Baseline Characteristics

A total of 339 participants were included in the analysis. Of these, 82% (278/339) were women, and the average age was 48.4 (SD 10.7) years. The mean baseline BMI was 36.7 (SD 7.3) kg/m^2 (Table 1).

In the tirzepatide cohort (209/339, 61.7%), 82.8% (173/209) of the participants were women, with an average age of 48.2 (SD 10.6) years. The mean baseline BMI was 36.4 (SD 6.8) kg/m^2 (Table 1).

In the semaglutide cohort (130/339, 38.3%), 80.8% (105/130) of the participants were women. The average age was 48.8 (SD 10.9) years. The mean baseline BMI was 37.0 (SD 7.9) kg/m^2 (Table 1).

Table 1. Baseline characteristics for all participants, participants in the tirzepatide cohort, and participants in the semaglutide cohort (N=339).

Characteristic	All participants	Tirzepatide cohort participants (n=209)	Semaglutide cohort participants (n=130)
Age (y), mean (SD)	48.4 (10.7)	48.2 (10.6)	48.8 (10.9)
Female sex, n (%)	278 (82)	173 (82.8)	105 (80.8)
Baseline weight (kg), mean (SD)	103.8 (23.6)	103.5 (22.6)	104.1 (25.3)
Baseline BMI (kg/m^2), mean (SD)	36.7 (7.3)	36.4 (6.8)	37.0 (7.9)

XSL•FO
RenderX

Effectiveness

Weight Change

The mean weight change observed in the tirzepatide cohort (209/339, 61.7%) over the 12 months was −22.9 (SD 9.9; *P*<.001) kg, or −22.1% (SD 8%) of participants' starting weight. A total of 95.2% (199/209) of the participants achieved ≥10% weight loss, and 83.7% (175/209) achieved ≥15% weight loss (Figure 1 and Table 2).

In the semaglutide cohort (130/339, 38.3%), the mean weight change observed over the 12-month period was −18.1 (SD 12.9; *P*<.001) kg, or −17.1% (SD 8.1%) of participants' starting weight. A total of 83.1% (108/130) of the participants achieved ≥10% weight loss, and 56.2% (73/130) achieved ≥15% weight loss (Figure 1, Table 2).

Figure 1. Weight change percentage over 12 months (semaglutide vs tirzepatide).

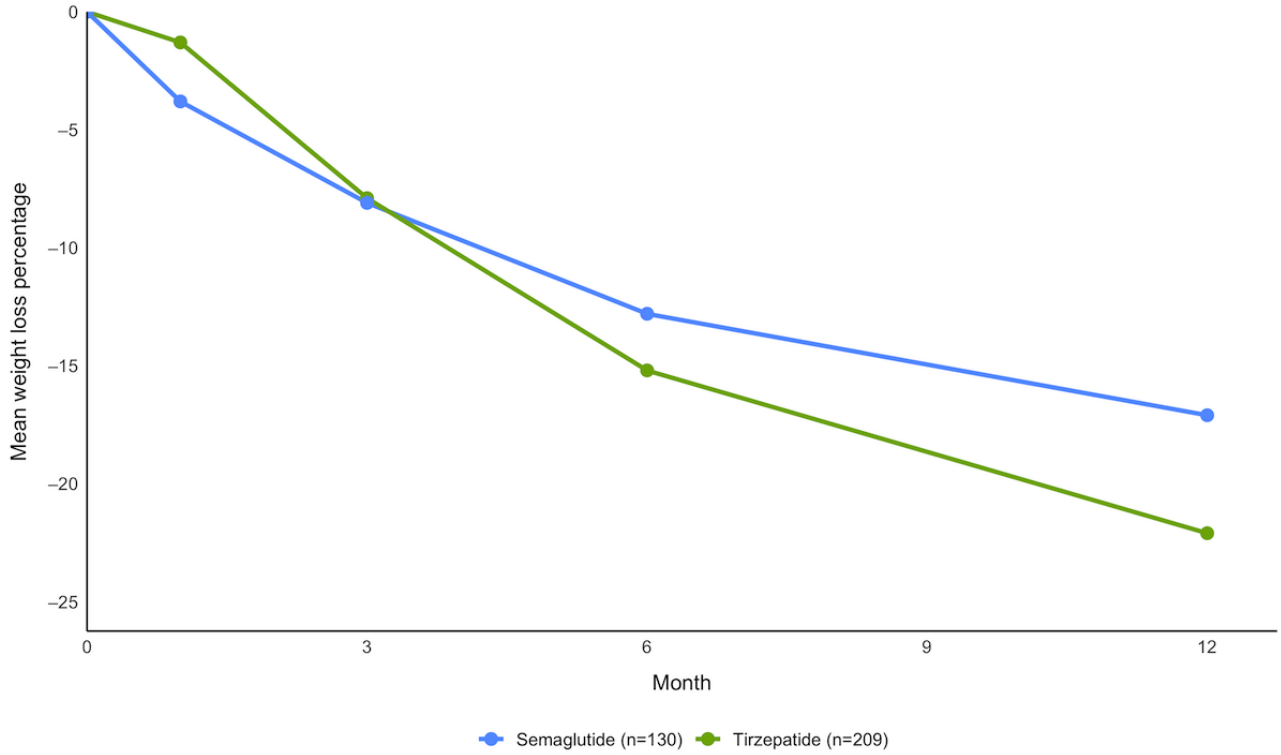

Table 2. Mean weight change percentage achieved in months 1, 3, 6, and 12 by molecule.

Month	Tirzepatide cohort (n=209; %), mean (SD)	Semaglutide cohort (n=130; %), mean (SD)
1	−1.3 (4.1)	−3.8 (14.1)
3	−7.9 (4.8)	−8.1 (14.2)
6	−15.2 (5.8)	−12.8 (14.6)
12	−22.1 (9.8)	−17.1 (16.7)

Cost Compared to Usual Care

The cost of delivering behavior change support in a secondary care SWMS (ie, delivered by a multidisciplinary team) was estimated by the external assessment group for the NICE health technology evaluation 14 guidance to be £1796 (US $2413.81) per patient per year [20]. We have previously evaluated that the Second Nature service could provide 60% to 70% cost savings when compared to SWMSs [17].

The cost of delivering behavior change support in a primary care or community-based, nonspecialist service has been estimated by the NICE to be between £868.21 (US $1166.87) and £1239.21 (US $1665.49) per patient per year [14,28]. Second Nature's model could deliver 10% to 60% cost savings compared to the NICE's primary care modeling.

For an expected 1395 enrollments per 100,000 persons, the total cost for delivering a service would be £1.2 to 1.7 million (US $1.6-$2.3 million), meaning that Second Nature's model could save between £120,000 (US $161,279; 10%) and £1,020,000 (US $1,370,870; 60%) per 100,000 individuals.

Behavior Changes

When asked about fruit and vegetable consumption, over half (114/209, 54.5%) of the participants in the tirzepatide cohort consistently reported consuming 3 to 4 different items per day over the 12-month period except for month 3 (Table 3). This trend was also observed in the semaglutide cohort; however,

XSL•FO
RenderX

this group consistently had a higher proportion of respondents (112/209, 53.6% of the tirzepatide cohort vs 73/130, 56.2% of the semaglutide cohort in month 1) and had no reduction in month 3.

Regarding exercise frequency, the proportion of participants in the tirzepatide cohort who were already exercising 3 to 4 times per week generally increased over the course of the study (from 73/209, 34.9% in month 1 to 90/209, 43.1% in month 12); however, this trend stopped in month 6. This trend was also observed in the semaglutide cohort (from 43/130, 33.1% in month 1 to 57/130, 43.8% in month 12). Across both cohorts, the percentage of respondents who never exercised each week also decreased, with the largest, statistically significant change observed in the semaglutide cohort (from 29/130, 22.3% in month 1 to 7/130, 5.4% in month 12; *P*<.001).

Across both cohorts, a decrease in cooking frequency was observed over the course of 12 months. In the tirzepatide cohort, the percentage of respondents who cooked every day decreased from 30.1% (63/209) in month 1 to 19.6% (41/209) in month 12, whereas in the semaglutide cohort, it decreased significantly from 41.5% (54/130) to 23.1% (30/130; *P*=.03). There was minimal change in those who never cooked across the 12 months; however, a substantial increase in the percentage of respondents who cooked *a few times a week* was observed in the tirzepatide cohort, changing from 24.4% (51/209) to 32.5% (68/209). In the semaglutide cohort, a similar trend was also observed, changing from 23.1% (30/130) in month 1 to 30% (39/130) in month 12.

Table 3. Responses to the lifestyle-related behavioral survey by month.

Category and subcategory	Month 1, n (%)		Month 3, n (%)		Month 6, n (%)		Month 12, n (%)	
	Tirzepatide cohort (n=209)	Semaglutide cohort (n=130)	Tirzepatide cohort (n=209)	Semaglutide cohort (n=130)	Tirzepatide cohort (n=209)	Semaglutide cohort (n=130)	Tirzepatide cohort (n=209)	Semaglutide cohort (n=130)
Fruit and vegetable consumption (items per day)								
0	0 (0)	3 (2.3)	0 (0)	1 (0.8)	2 (1)	1 (0.8)	0 (0)	1 (0.8)
1-2	36 (17.2)	17 (13.1)	50 (23.9)	23 (17.7)	44 (21.1)	26 (20)	40 (19.1)	29 (22.3)
3-4	112 (53.6)	73 (56.2)	97 (46.4)	69 (53.1)	114 (54.5)	74 (56.9)	114 (54.5)	72 (55.4)
≥5	61 (29.2)	37 (28.5)	62 (29.7)	37 (28.5)	49 (23.4)	29 (22.3)	55 (26.3)	28 (21.5)
Exercise (times per week)								
0	31 (14.8)	29 (22.3)	25 (12)	13 (10)	17 (8.1)	9 (6.9)	14 (6.7)[a]	7 (5.4)[b]
1-2	85 (40.7)	47 (36.2)	87 (41.6)	51 (39.2)	105 (50.2)	54 (41.5)	83 (39.7)	52 (40)
3-4	73 (34.9)	43 (33.1)	75 (35.9)	52 (40)	67 (32.1)	54 (41.5)	90 (43.1)	57 (43.8)
≥5	20 (9.6)	11 (8.5)	22 (10.5)	14 (10.8)	20 (9.6)	13 (10)	22 (10.5)	14 (10.8)
Cooking frequency								
Never	4 (1.9)	0 (0)	3 (1.4)	1 (0.8)	1 (0.5)	0 (0)	2 (1)	1 (0.8)
Less than once a week	2 (1)	2 (1.5)	2 (1)	5 (3.8)	7 (3.3)	6 (4.6)	5 (2.4)	1 (0.8)
Once a week	4 (1.9)	3 (2.3)	8 (3.8)	4 (3.1)	13 (6.2)	4 (3.1)	11 (5.3)	8 (6.2)
Few times a week	51 (24.4)	30 (23.1)	73 (34.9)	28 (21.5)	67 (32.1)[c]	41 (31.5)	68 (32.5)[d]	39 (30)
Nearly every day	85 (40.7)	41 (31.5)	77 (36.8)	58 (44.6)	81 (38.8)	55 (42.3)	82 (39.2)	51 (39.2)
Every day	63 (30.1)	54 (41.5)	46 (22)	34 (26.2)	40 (19.1)	24 (18.5)[e]	41 (19.6)[d]	30 (23.1)[d]

[a]*P*=.04.

[b]*P*<.001.

[c]*P*<.001.

[d]*P*=.03.

[e]*P*=.02.

XSL•FO
RenderX

Acceptability: Participant Experience

The participant responses were organized into 9 thematic categories: change in appetite, illness, intervention benefits, negative experience, neutral experience, other, other priorities, positive experience, and slow progress.

At month 1, across both cohorts, most participants reported positive experiences (130/209, 62.2% and 90/130, 69.2% for the tirzepatide and semaglutide cohorts, respectively). This declined significantly to an approximate third by month 12 (66/209, 31.6% and 51/130, 39.2% for the tirzepatide and semaglutide cohorts, respectively; *P*<.001). Negative experiences did not noticeably increase; however, neutral experiences significantly increased from 12.9% (27/209) in the tirzepatide cohort and 15.4% (20/130) in the semaglutide cohort at month 1 to 52.2% (109/209) and 50% (65/130) by month 12, respectively (Table 4; *P*<.001).

There was a trend in both change in appetite and intervention benefits in that, across both the semaglutide and tirzepatide cohorts, a reduction in reporting was observed. A total of 12.3% (16/130) of semaglutide users reported a change in appetite in month 1, but only 0.8% (1/130) did so in month 12. In total, 21.1% (44/209) of tirzepatide users reported this effect in month 1, but only 1.4% (3/209) did so in month 12. Both cohorts showed a significant reduction (*P*<.001). Intervention benefits decreased in reporting levels from 20% (26/130) to 3.1% (17/130) from month 1 to 12 (*P*<.001) in the semaglutide cohort, whereas in the tirzepatide cohort, it dropped from 12.4% (26/209) to 6.7% (14/209) over the same time frame.

Table 4. Percentage of cohort participants reporting different experiences throughout by program month.

Experience	Month 1, n (%)		Month 3, n (%)		Month 6, n (%)		Month 12, n (%)	
	Tirzepatide cohort (n=209)	Semaglutide cohort (n=130)	Tirzepatide cohort (n=209)	Semaglutide cohort (n=130)	Tirzepatide cohort (n=209)	Semaglutide cohort (n=130)	Tirzepatide cohort (n=209)	Semaglutide cohort (n=130)
Change in appetite	44 (21.1)	16 (12.3)	12 (5.7)[a]	10 (7.7)	7 (3.3)[a]	4 (3.1)	3 (1.4)[a]	1 (0.8)[a]
Illness	1 (0.5)	1 (0.8)	3 (1.4)	3 (2.3)	2 (1)	3 (2.3)	0 (0)	0 (0)
Intervention benefits	26 (12.4)	26 (20)	29 (13.9)	17 (13.1)	10 (4.8)	14 (10.8)	14 (6.7)	4 (3.1)
Negative experience	0 (0)	1 (0.8)	0 (0)	0 (0)	0 (0)	1 (0.8)	0 (0)	0 (0)
Neutral experience	27 (12.9)	20 (15.4)	41 (19.6)	24 (18.5)	68 (32.5)[a]	42 (32.3)[a]	109 (52.2)[a]	65 (50)[a]
Other	21 (10)	9 (6.9)	20 (9.6)	10 (7.7)	35 (16.7)	2 (1.5)	13 (6.2)	6 (4.6)
Other priorities	6 (2.9)	4 (3.1)	10 (4.8)	2 (1.5)	6 (2.9)	1 (0.8)	4 (1.9)	3 (2.3)
Positive experience	130 (62.2)	90 (69.2)	98 (46.9)[a]	70 (53.8)	83 (39.7)[a]	65 (50)[a]	66 (31.6)[a]	51 (39.2)[a]
Reduced effectiveness or slow progress	0 (0)	1 (0.8)	0 (0)	0 (0)	1 (0.5)	0 (0)	0 (0)	0 (0)

[a]*P*<.001.

Feasibility: Side Effects

In the first month for the tirzepatide cohort, the most commonly reported side effects were constipation (61/209, 29.2%), feeling sick (58/209, 27.8%), and feeling more tired than usual (45/209, 21.5%), whereas 41.6% (87/209) reported no side effects at this stage. After 12 months, 60.3% (126/209; *P*<.05) of participants reported no side effects, and there was a significant decrease in the prevalence of the most common side effects (constipation: 27/209, 12.9% and *P*<.001; feeling sick: 22/209, 10.5% and *P*<.001; feeling more tired than usual: 17/209, 8.1% and *P*<.001) but an increase in hair loss (20/209, 9.6%; *P*<.001; Table 5).

These effects were similarly observed in the semaglutide cohort, with constipation (41/130, 31.5%), feeling sick (41/130, 31.5%), and feeling more tired than usual (32/130, 24.6%) decreasing in prevalence. In month 12, a total of 67.7% (88/130) of the participants reported no side effects, 15.4% (20/130) reported constipation, 3.1% (4/130) reported feeling sick (*P*<.001), 3.1% (4/130) reported feeling more tired than usual (*P*<.001), and 5.4% (7/130) reported hair loss (Table 5).

XSL•FO
RenderX

Table 5. Percentage of cohort participants reporting side effects by molecule by survey month.

Side effect	Month 1, n (%)		Month 3, n (%)		Month 6, n (%)		Month 12, n (%)	
	Tirzepatide cohort (n=209)	Semaglutide cohort (n=130)	Tirzepatide cohort (n=209)	Semaglutide cohort (n=130)	Tirzepatide cohort (n=209)	Semaglutide cohort (n=130)	Tirzepatide cohort (n=209)	Semaglutide cohort (n=130)
Constipation	61 (29.2)	41 (31.5)	45 (21.5)	40 (30.8)	42 (20.1)	23 (17.7)	27 (12.9)[a]	20 (15.4)
Diarrhea	18 (8.6)	13 (10)	20 (9.6)	14 (10.8)	26 (12.4)	7 (5.4)	13 (6.2)	7 (5.4)
Feeling anxious	5 (2.4)	2 (1.5)	4 (1.9)	1 (0.8)	1 (0.5)	3 (2.3)	4 (1.9)	1 (0.8)
Feeling dizzy	12 (5.7)	6 (4.6)	7 (3.3)	5 (3.8)	2 (1)	2 (1.5)	6 (2.9)	5 (3.8)
Feeling more tired than usual	45 (21.5)	32 (24.6)	42 (20.1)	26 (20)	29 (13.9)	23 (17.7)	17 (8.1)[b]	4 (3.1)[b]
Feeling sick	58 (27.8)	41 (31.5)	59 (28.2)	46 (35.4)	42 (20.1)	14 (10.8)[b]	22 (10.5)[b]	4 (3.1)[b]
Hair loss	3 (1.4)	2 (1.5)	6 (2.9)	1 (0.8)	22 (10.5)[b]	4 (3.1)	20 (9.6)[b]	7 (5.4)
Headaches	16 (7.7)	20 (15.4)	9 (4.3)	18 (13.8)	5 (2.4)	3 (2.3)[b]	5 (2.4)	3 (2.3)[b]
Heartburn or indigestion	37 (17.7)	11 (8.5)	37 (17.7)	27 (20.8)	21 (10)	11 (8.5)	10 (4.8)[b]	7 (5.4)
No side effects	87 (41.6)	70 (53.8)	82 (39.2)	41 (31.5)[b]	92 (44)	70 (53.8)	126 (60.3)[b]	88 (67.7)
None of the above	4 (1.9)	3 (2.3)	4 (1.9)	1 (0.8)	4 (1.9)	3 (2.3)	5 (2.4)	2 (1.5)
Reactions in the injection site	10 (4.8)	2 (1.5)	12 (5.7)	0 (0)	25 (12)	1 (0.8)	11 (5.3)	2 (1.5)
Severe stomach pain	0 (0)	1 (0.8)	0 (0)	1 (0.8)	1 (0.5)	1 (0.8)	0 (0)	1 (0.8)
Tender or swollen stomach	2 (1)	1 (0.8)	1 (0.5)	2 (1.5)	2 (1)	1 (0.8)	0 (0)	0 (0)
Vomiting	5 (2.4)	2 (1.5)	10 (4.8)	7 (5.4)	6 (2.9)	3 (2.3)	5 (2.4)	1 (0.8)

[a]P=.002.

[b]P<.001.

Discussion

Principal Findings

This 12-month retrospective observational analysis examined the effectiveness, feasibility, acceptability, and potential cost-effectiveness of Second Nature's remotely delivered weight management program supported by GLP-1RAs. This study also aimed to compare the outcomes of semaglutide and tirzepatide within a remote care setting.

Users achieved substantial and clinically meaningful weight loss over 12 months, with participants in both the tirzepatide and semaglutide cohorts losing more weight over comparable periods than via pharmacological intervention alone [8-11]. In both cohorts, there was a significant decrease in the prevalence of the most common side effects—constipation, feeling more tired than usual, and feeling sick. However, there was an increase in hair loss. Accompanying this was a significant decrease in positive experiences and a significant increase in neutral experiences in both cohorts. In addition, reports of appetite changes and intervention benefits decreased markedly over the 12-month period.

Given the developing commissioning landscape of weight management within the NHS, our findings suggest that Second Nature's intervention could result in a 60% to 70% reduction in costs compared to secondary care SWMSs and up to a 60% reduction compared to primary care and community-based services. At a population level, this translates to potential savings of £120,000 to £1,020,000 (US $161,279-$1,370,870) per 100,000 individuals. These findings indicate that remotely delivered interventions could provide substantial cost savings for ICBs during commissioning.

Comparing between GLP-1RAs, the tirzepatide cohort achieved a mean weight change of −22.9 kg (−22.1% of starting weight, SD 8%), whereas those in the semaglutide cohort lost −18.1 kg (−17.1% of starting weight, SD 8.1%) over a 12-month period (Figure 1). These outcomes substantially exceed the clinically significant threshold of 10% weight loss, with 95.2% (199/209)

XSL•FO
RenderX

of tirzepatide participants and 83.1% (108/130) of semaglutide participants achieving this milestone.

Comparison With Previous Work

Weight Change Outcomes

Significant differences in study design limit direct comparisons with randomized controlled trials; however, the results remain consistent with both results of clinical trials and real-world data on GLP-1RA therapy. The tirzepatide group saw a −22.1% mean weight change, closely matching results from the SURMOUNT-1 and SURMOUNT-5 trials (−20.9% and −20.2%, respectively) [10,28]. The semaglutide group achieved −17.1% mean weight loss, aligning with STEP-1 and STEP-5 outcomes (−14.9% and −15.2%, respectively) [29,30]. Previous studies comparing between the 2 drugs have shown mixed results, with some real-world studies reporting lower weight change outcomes [31,32]. This evaluation highlights the added value of combining medication with behavioral support by providing similar weight outcomes from completers to the mean weight loss observed in trials such as STEP-3 and SURMOUNT-3 [22,33-36].

Behavioral Outcomes

Over the 12-month program, notable behavior changes were observed. While fruit and vegetable intake remained stable, daily cooking declined in both cohorts—significantly so in the semaglutide group (P=.03). These findings could support those of other studies showing reduced energy intake on GLP-1RA therapies [37,38]. Conversely, more participants began cooking a few times a week, possibly due to program guidance, with a substantial rise in the tirzepatide group.

Physical activity also improved, particularly among initially inactive participants. The proportion of participants reporting no weekly exercise fell from 14.8% (31/209) to 6.7% (14/209; P=.04) in the tirzepatide cohort and from 22.3% (29/130) to 5.4% (7/130; P<.001) in the semaglutide cohort. While GLP-1RA–induced weight loss can include muscle loss, increased activity suggests potential benefits for body composition and mobility [39].

Safety and Tolerability

The side effect profiles for both semaglutide and tirzepatide were consistent with clinical trial data, with constipation, nausea, and fatigue being most common [10,29,40]. Over time, side effects decreased significantly, with more participants reporting no side effects by month 12—increasing from 41.6% (87/209) to 60.3% (126/209; P<.001) in the tirzepatide cohort and from 53.8% (70/130) to 67.7% (88/130) in the semaglutide cohort. Common side effects such as constipation and nausea decreased significantly in both groups (P<.001). These results indicate that tolerability increases and that the common side effects are often resolved with time [30,41,42].

Notably, reports of hair loss increased with time—9.6% (20/209) of the tirzepatide cohort (P<.001) and 5.4% (7/130) of the semaglutide cohort. This has been previously observed, although not to this degree, with most studies observing approximately a 2% increase in reporting with time [43-45].

Clinical and Policy Implications

Following the NICE Technology Appraisal TA1026 and NHS England's phased rollout, ICBs are now required to prescribe tirzepatide in both primary and secondary care, whereas semaglutide remains limited to secondary care. The guidance mandates at least 9 months of wraparound care, including nutrition and behavior support, creating demand for scalable delivery models [14,16,46]. A remote service could provide cost savings of 60% to 70% in secondary care and 10% to 60% in primary care, mainly due to workforce efficiencies—although these figures are based on hypothetical service costs and modeled estimates [13,25,47].

Major implementation challenges remain, including long wait times and paused services due to capacity limits [46-48]. Integrating remote care within existing systems is essential to prevent duplication of costs. A supported *hub and spoke* model—a single point of access (hub) that is able to refer patients to the proper service (spokes; eg, tier 2-4 weight management services)—could enhance access and outcomes [48]. With remote care's scalability, this approach could quickly reach those most in need.

Limitations and Benefits

This 12-month study provides large-scale longitudinal evidence from 339 participants, offering substantial real-world data on remote GLP-1RA delivery effectiveness. It captures comprehensive outcomes across multiple domains, including effectiveness, safety, behavior changes, and economic impact, providing a holistic view of treatment success. Comparative insights between semaglutide and tirzepatide users reveal important differences in effectiveness and tolerability between these medications. Detailed documentation of side effects, behavioral modifications, and weight changes offers clinically relevant information for practitioners and patients considering this treatment approach.

This service evaluation exhibits self-selection bias that may limit the generalizability of the findings. Only participants who completed the full 12-month commercial intervention—recording weights and completing surveys at months 1, 3, 6, and 12—were included in the analysis. This focus on completers may overestimate treatment effectiveness as participants experiencing poor outcomes would be more likely to discontinue the service. Similarly, recruitment may limit results as the self-paying cohort likely represents a more motivated and affluent population than typical NHS patients, potentially overestimating treatment success rates. We also did not recruit for a control group, which makes addressing these selection biases, accounting for further confounding variables, and establishing true effect size difficult. The absence of clinical markers such as metabolic and cardiovascular data significantly restricts the ability to draw broader health conclusions beyond weight loss outcomes.

Future Research Directions

Building on this study's limitations, future research should include randomized controlled trials comparing remote and in-person delivery models, as well as long-term prospective studies to reduce bias and assess weight regain after treatment

XSL•FO
RenderX

ends. Further investigation into program intensity, behavioral support, service integration, and postmedication care is also warranted.

With tirzepatide now being prescribed through NHS primary care, real-world cost-effectiveness studies are needed. These should evaluate the financial implications of widespread implementation, including primary care capacity, monitoring requirements, and long-term health outcomes, to inform sustainable commissioning decisions and optimal care pathways.

Given evidence of significant weight regain after pharmacological treatment, studying the role of behavior change in mitigating this is crucial. Coupling such research with robust financial and health economic evaluations would help determine the long-term value of these interventions and inform NHS policy and commissioning decisions.

Conclusions

This real-world evaluation demonstrates that remotely delivered, GLP-1RA–supported weight management programs can achieve significant weight loss above and beyond corresponding medication-only clinical trial outcomes while potentially reducing health care costs by 60% to 70%. These findings have profound implications for health care delivery transformation, suggesting that digital health interventions can not only match but also surpass traditional care models in effectiveness while addressing the scalability challenges facing obesity treatment. The significant changes in weight, alongside positive behavior changes (eg, increases in cooking and exercise frequency) and changes in side effect occurrence (reduction overall over 12 months), indicate that integrated approaches combining pharmacological and behavioral support may be essential for optimizing long-term treatment success. This evidence supports a paradigm shift toward digitally enabled, behavior-focused care that could address the obesity epidemic at scale while reducing strain on traditional health care resources.

Acknowledgments

The authors thank the participants and the clinical and administrative teams at Second Nature Healthy Habits Ltd for delivering the service.

Data Availability

The datasets generated or analyzed during this study are not publicly available as participants did not consent to having their data shared but may be available from the corresponding author on reasonable request.

Conflicts of Interest

RR, WL, MW, GS, and RH are employees of Second Nature Healthy Habits Ltd.

Multimedia Appendix 1
Health Research Authority decision tool.
[PDF File (Adobe PDF File), 210 KB - formative_v9i1e81912_app1.pdf]

References

1. Obesity and overweight. World Health Organization. 2025 May 7. URL: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight [accessed 2025-07-01]
2. Rubino F, Cummings DE, Eckel RH, Cohen RV, Wilding JP, Brown WA, et al. Definition and diagnostic criteria of clinical obesity. Lancet Diabetes Endocrinol 2025 Mar;13(3):221-262. [doi: 10.1016/s2213-8587(24)00316-4]
3. Unhealthy numbers: the rising cost of obesity in the UK. Tony Blair Institute for Global Change. 2023 Nov 21. URL: https://institute.global/insights/public-services/unhealthy-numbers-the-rising-cost-of-obesity-in-the-uk [accessed 2025-07-01]
4. Bunting T, Bone J, Barber J. Why reducing obesity could help to tackle the UK's productivity problem. Nesta. 2025 Jul 2. URL: https://www.nesta.org.uk/blog/why-reducing-obesity-could-help-to-tackle-the-uks-productivity-problem/ [accessed 2025-08-23]
5. Hall KD, Kahan S. Maintenance of lost weight and long-term management of obesity. Med Clin North Am 2018 Jan;102(1):183-197 [FREE Full text] [doi: 10.1016/j.mcna.2017.08.012] [Medline: 29156185]
6. Kushner RF, Calanna S, Davies M, Dicker D, Garvey WT, Goldman B, et al. Semaglutide 2.4 mg for the treatment of obesity: key elements of the STEP trials 1 to 5. Obesity (Silver Spring) 2020 Jun 22;28(6):1050-1061 [FREE Full text] [doi: 10.1002/oby.22794] [Medline: 32441473]
7. Ghusn W, De la Rosa A, Sacoto D, Cifuentes L, Campos A, Feris F, et al. Weight loss outcomes associated with semaglutide treatment for patients with overweight or obesity. JAMA Netw Open 2022 Sep 01;5(9):e2231982 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.31982] [Medline: 36121652]
8. Wilding JP, Batterham RL, Calanna S, Davies M, Van Gaal LF, Lingvay I, et al. Once-weekly semaglutide in adults with overweight or obesity. N Engl J Med 2021 Mar 18;384(11):989-1002. [doi: 10.1056/nejmoa2032183]

XSL•FO
RenderX

9. Aronne LJ, Sattar N, Horn DB, Bays HE, Wharton S, Lin WY, et al. Continued treatment with tirzepatide for maintenance of weight reduction in adults with obesity: the SURMOUNT-4 randomized clinical trial. JAMA 2024 Jan 02;331(1):38-48 [FREE Full text] [doi: 10.1001/jama.2023.24945] [Medline: 38078870]
10. Jastreboff AM, Aronne LJ, Ahmad NN, Wharton S, Connery L, Alves B, et al. Tirzepatide once weekly for the treatment of obesity. N Engl J Med 2022 Jul 21;387(3):205-216. [doi: 10.1056/nejmoa2206038]
11. Jastreboff AM, Kaplan LM, Frías JP, Wu Q, Du Y, Gurbuz S, et al. Triple–hormone-receptor agonist retatrutide for obesity — a phase 2 trial. N Engl J Med 2023 Aug 10;389(6):514-526. [doi: 10.1056/nejmoa2301972]
12. Rubino D, Abrahamsson N, Davies M, Hesse D, Greenway FL, Jensen C, et al. Effect of continued weekly subcutaneous semaglutide vs placebo on weight loss maintenance in adults with overweight or obesity: the STEP 4 randomized clinical trial. JAMA 2021 Apr 13;325(14):1414-1425 [FREE Full text] [doi: 10.1001/jama.2021.3224] [Medline: 33755728]
13. Bechenati D, Castaneda R, Rivera R, Espinosa MA, Villamarin JL, Tama E, et al. The rebound effect—weight regain after tirzepatide discontinuation. Diabetes 2025 Jun;74(Supplement_1):1757 [FREE Full text] [doi: 10.2337/db25-1757-P]
14. Tirzepatide for managing overweight and obesity. National Institutes for Health and Care Excellence. 2024 Dec 23. URL: https://www.nice.org.uk/guidance/ta1026 [accessed 2025-08-23]
15. Fit for the future: 10 year health plan for England - executive summary. Department of Health & Social Care, United Kingdom Government. 2025. URL: https://www.gov.uk/government/publications/10-year-health-plan-for-england-fit-for-the-future/fit-for-the-future-10-year-health-plan-for-england-executive-summary [accessed 2025-08-23]
16. Interim commissioning guidance: implementation of the NICE technology appraisal TA1026 and the NICE funding variation for tirzepatide (Mounjaro®) for the management of obesity. National Health Service England. 2025. URL: https://tinyurl.com/yvne36kr [accessed 2025-08-23]
17. Richards R, Whitman M, Wren G, Campion P. A remotely delivered GLP-1RA-supported specialist weight management program in adults living with obesity: retrospective service evaluation. JMIR Form Res 2025 Jun 30;9:e72577 [FREE Full text] [doi: 10.2196/72577] [Medline: 40434299]
18. Guide to reporting, action planning, communication and dissemination. NHS Evaluation Toolkit. URL: https://nhsevaluationtoolkit.net/resources/guide-to-reporting-communicating-and-acting/ [accessed 2025-08-14]
19. Semaglutide for managing overweight and obesity. National Institute for Health and Care Excellence. 2022. URL: https://www.nice.org.uk/guidance/ta875/documents/129 [accessed 2025-07-01]
20. Digital technologies for delivering multidisciplinary weight-management services: early value assessment. National Institute for Health and Care Excellence. 2023. URL: https://www.nice.org.uk/guidance/indevelopment/gid-hte10007 [accessed 2025-07-01]
21. Richards R, Wren GM, Campion P, Whitman M. A remotely delivered, semaglutide-supported specialist weight management program: preliminary findings from a retrospective service evaluation. JMIR Form Res 2023 Dec 28;7:e53619 [FREE Full text] [doi: 10.2196/53619] [Medline: 38153780]
22. Idris I, Hampton J, Moncrieff F, Whitman M. Effectiveness of a digital lifestyle change program in obese and type 2 diabetes populations: service evaluation of real-world data. JMIR Diabetes 2020 Jan 20;5(1):e15189 [FREE Full text] [doi: 10.2196/15189] [Medline: 31958064]
23. Kar P, Goward C, Whitman M, Davies M, Willner T, Shaw K. Engagement and effectiveness of digitally enabled behavioural change support for people living with type 2 diabetes. Pract Diabetes 2020 Oct 07;37(5):167-72a. [doi: 10.1002/pdi.2295]
24. Michie S, van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. Implement Sci 2011 Apr 23;6(1):42 [FREE Full text] [doi: 10.1186/1748-5908-6-42] [Medline: 21513547]
25. Karrar HR, Nouh MI, Nouh YI, Nouh MI, Khan Alhindi AS, Hemeq YH, et al. Tirzepatide-induced gastrointestinal manifestations: a systematic review and meta-analysis. Cureus 2023 Sep;15(9):e46091 [FREE Full text] [doi: 10.7759/cureus.46091] [Medline: 37908927]
26. Is my study research? National Health Service Health Research Authority. URL: http://www.hra-decisiontools.org.uk/research/ [accessed 2025-07-01]
27. Scott AM, Kolstoe S, Ploem MC, Hammatt Z, Glasziou P. Exempting low-risk health and medical research from ethics reviews: comparing Australia, the United Kingdom, the United States and the Netherlands. Health Res Policy Syst 2020 Jan 28;18(1):11 [FREE Full text] [doi: 10.1186/s12961-019-0520-4] [Medline: 31992320]
28. Digital technologies for delivering multidisciplinary weight-management services: early value assessment. National Institute for Health and Care Excellence. URL: https://www.nice.org.uk/guidance/hte14/resources/resource-impact-template-excel-13197081133 [accessed 2025-07-01]
29. Wilding JP, Batterham RL, Davies M, Van Gaal LF, Kandler K, Konakli K, et al. Weight regain and cardiometabolic effects after withdrawal of semaglutide: the STEP 1 trial extension. Diabetes Obes Metab 2022 Aug 19;24(8):1553-1564 [FREE Full text] [doi: 10.1111/dom.14725] [Medline: 35441470]
30. Garvey WT, Batterham RL, Bhatta M, Buscemi S, Christensen LN, Frias JP, et al. Two-year effects of semaglutide in adults with overweight or obesity: the STEP 5 trial. Nat Med 2022 Oct 10;28(10):2083-2091 [FREE Full text] [doi: 10.1038/s41591-022-02026-4] [Medline: 36216945]

XSL•FO
RenderX

31. Aronne LJ, Horn DB, le Roux CW, Ho W, Falcon BL, Gomez Valderas E, et al. Tirzepatide as compared with semaglutide for the treatment of obesity. N Engl J Med 2025 Jul 03;393:26-36. [doi: 10.1056/nejmoa2416394]
32. Rodriguez PJ, Goodwin Cartwright BM, Gratzl S, Brar R, Baker C, Gluckman TJ, et al. Semaglutide vs tirzepatide for weight loss in adults with overweight or obesity. JAMA Intern Med 2024 Sep 01;184(9):1056-1064. [doi: 10.1001/jamainternmed.2024.2525] [Medline: 38976257]
33. Wadden TA, Bailey TS, Billings LK, Davies M, Frias JP, Koroleva A, et al. Effect of subcutaneous semaglutide vs placebo as an adjunct to intensive behavioral therapy on body weight in adults with overweight or obesity: the STEP 3 randomized clinical trial. JAMA 2021 Apr 13;325(14):1403-1413 [FREE Full text] [doi: 10.1001/jama.2021.1831] [Medline: 33625476]
34. Wadden TA, Chao AM, Machineni S, Kushner R, Ard J, Srivastava G, et al. Tirzepatide after intensive lifestyle intervention in adults with overweight or obesity: the SURMOUNT-3 phase 3 trial. Nat Med 2023 Nov 15;29(11):2909-2918 [FREE Full text] [doi: 10.1038/s41591-023-02597-w] [Medline: 37840095]
35. Ard JD, Hong YR, Foster GD, Medcalf A, Nadolsky S, Cardel MI. Twelve-month analysis of real-world evidence from a telehealth obesity-treatment provider using antiobesity medications. Obesity (Silver Spring) 2024 Dec 30;32(12):2246-2254. [doi: 10.1002/oby.24169] [Medline: 39478294]
36. Is coming off semaglutide slowly the key to preventing weight regain? European Association for the Study of Obesity. 2024 May 11. URL: https://easo.org/is-coming-off-Semaglutide-slowly-the-key-to-preventing-weight-regain/ [accessed 2025-07-01]
37. Blundell J, Finlayson G, Axelsen M, Flint A, Gibbons C, Kvist T, et al. Effects of once-weekly semaglutide on appetite, energy intake, control of eating, food preference and body weight in subjects with obesity. Diabetes Obes Metab 2017 Sep 05;19(9):1242-1251 [FREE Full text] [doi: 10.1111/dom.12932] [Medline: 28266779]
38. Heise T, DeVries JH, Urva S, Li J, Pratt EJ, Thomas MK, et al. Tirzepatide reduces appetite, energy intake, and fat mass in people with type 2 diabetes. Diabetes Care 2023 May 01;46(5):998-1004 [FREE Full text] [doi: 10.2337/dc22-1710] [Medline: 36857477]
39. Neeland IJ, Linge J, Birkenfeld AL. Changes in lean body mass with glucagon-like peptide-1-based therapies and mitigation strategies. Diabetes Obes Metab 2024 Jun 27;26(S4):16-27. [doi: 10.1111/dom.15728]
40. Mishra R, Raj R, Elshimy G, Zapata I, Kannan L, Majety P, et al. Adverse events related to tirzepatide. J Endocr Soc 2023 Feb 09;7(4):bvad016 [FREE Full text] [doi: 10.1210/jendso/bvad016] [Medline: 36789109]
41. Niu K, Fan M, Gao W, Chen C, Dai G. Adverse events in different administration routes of semaglutide: a pharmacovigilance study based on the FDA adverse event reporting system. Front Pharmacol 2024 Jun 3;15:1414268 [FREE Full text] [doi: 10.3389/fphar.2024.1414268] [Medline: 38887555]
42. Ryan DH, Lingvay I, Deanfield J, Kahn SE, Barros E, Burguera B, et al. Long-term weight loss effects of semaglutide in obesity without diabetes in the SELECT trial. Nat Med 2024 Jul 13;30(7):2049-2057. [doi: 10.1038/s41591-024-02996-7] [Medline: 38740993]
43. Buontempo MG, Santos BT. Exploring the hair loss risk in glucagon‐like peptide‐1 agonists: emerging concerns and clinical implications. Acad Dermatol Venereol 2025 Jan 24;39(2):263-264. [doi: 10.1111/jdv.20512]
44. Godfrey H, Leibovit-Reiben Z, Jedlowski P, Thiede R. Alopecia associated with the use of semaglutide and tirzepatide: a disproportionality analysis using the FDA adverse event reporting system (FAERS) from 2022 to 2023. J Eur Acad Dermatol Venereol 2025 Feb 24;39(2):e153-e154. [doi: 10.1111/jdv.20197] [Medline: 38925559]
45. Jairath R, Abreu D, Mann C. Semaglutide and hair loss: a real-world analysis. J Am Acad Dermatol 2024 Sep;91(3):AB309. [doi: 10.1016/j.jaad.2024.07.1228]
46. Local weight management services. National Health Service Cheshire and Merseyside. URL: https://www.cheshireandmerseyside.nhs.uk/your-health/prescribing/statements/mounjaro-tirzepatide/local-weight-management-services/ [accessed 2025-07-01]
47. Mozaffarian D, Agarwal M, Aggarwal M, Alexander L, Apovian CM, Bindlish S, et al. Nutritional priorities to support GLP-1 therapy for obesity: a joint Advisory from the American College of Lifestyle Medicine, the American Society for Nutrition, the Obesity Medicine Association, and The Obesity Society. Am J Clin Nutr 2025 Jul;122(1):344-367. [doi: 10.1016/j.ajcnut.2025.04.023] [Medline: 40450457]
48. Capehorn M, Johansson E, Davies A, Evans J, Godbeer F, van Hest N, et al. Cost-effectiveness of tirzepatide versus liraglutide (both adjunct to a reduced-calorie diet and increased physical activity) in patients with obesity or overweight from a UK perspective. In: Proceedings of the ISPOR Europe 2024. 2024 Presented at: ISPOR Europe 2024; November 17-20, 2024; Barcelona, Spain URL: https://www.ispor.org/docs/default-source/euro2024/ee789cem-tzp-vs-lira-ukjohansson-e146000-pdf.pdf?sfvrsn=e80125cf_0/> [doi: 10.1016/j.jval.2024.10.1069]

Abbreviations

GLP-1RA: glucagonlike peptide-1 receptor agonist
ICB: integrated care board
NHS: National Health Service
NICE: National Institute for Health and Care Excellence

XSL•FO
RenderX

STEP: Semaglutide Treatment Effect in People With Obesity
SWMS: specialist weight management service

Edited by A Mavragani; submitted 05.08.25; peer-reviewed by J Antoun, N Eid; comments to author 11.08.25; revised version received 18.08.25; accepted 19.08.25; published 12.09.25.

Please cite as:
Richards R, Lunt W, Whitman M, Spaltro G, Hall R
Semaglutide and Tirzepatide in a Remote Weight Management Program: 12-Month Retrospective Observational Study
JMIR Form Res 2025;9:e81912
URL: https://formative.jmir.org/2025/1/e81912
doi:10.2196/81912
PMID:

©Rebecca Richards, William Lunt, Michael Whitman, Giulia Spaltro, Rachel Hall. Originally published in JMIR Formative Research (https://formative.jmir.org), 12.09.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

TruVox Web-Based Software for Vocal Pitch Training in Transgender Women: Development and Single-Session Evaluations

Sam R Weese[1], BS; Mary E Wilkens[1], BS; Om Jadhav[1]; Xiangyi Wang[1], BS; Ansh Bhanushali[1]; Tyler DiLoreto[1]; Reyna Kozel[1], BS; Renee L Gustin[1], MS; Tara McAllister[2], PhD; Victoria Sue McKenna[1,3], PhD; Vesna Dominika Novak[1], PhD

[1]University of Cincinnati, Cincinnati, OH, United States

[2]Department of Communicative Sciences and Disorders, New York University, New York, NY, United States

[3]School of Communication Sciences and Disorders, University of Central Florida, Orlando, United States

Corresponding Author:
Vesna Dominika Novak, PhD
University of Cincinnati
2600 Clifton Avenue
Cincinnati, OH, 45221
United States
Phone: 1 5135562946
Email: novakdn@ucmail.uc.edu

Abstract

Background: Transgender people often experience distress due to a mismatch between their gender and the way their voice is perceived (eg, transgender women with low pitch), which significantly reduces their mental health and quality of life. This is especially a problem for transfeminine people and can be reduced with gender-affirming voice training (GAVT), but such training is often inaccessible due to factors such as price and geographical constraints.

Objective: We aim to improve the limited availability of GAVT by developing and testing a free web-based software platform (named TruVox; University of Cincinnati) that would combine real-time feedback about the user's voice with structured vocal pitch exercises for transfeminine people.

Methods: The current publicly accessible TruVox prototype focuses on vocal pitch training with 5 structured exercises that provide real-time pitch visualizations as well as supporting videos and text. It was tested in 2 evaluation stages: initial remote usability evaluations and a later single-session in-person evaluation with 21 transfeminine participants under the supervision of 2 researchers. In remote evaluations, participants reported bugs and usability issues that were iteratively addressed. In the in-person evaluation, participants tested the final software prototype and filled out the System Usability Scale, then performed 10 repetitions of different exercises to gauge performance improvement with practice. They also filled out the Intrinsic Motivation Inventory for each exercise.

Results: The System Usability Scale score had a mean of 79.8 (SD 12.8) on a 100-point scale, Intrinsic Motivation Inventory scores were high (eg, interest/enjoyment over 11/14), and exercise performance significantly improved in all but 1 exercise (P values ranging from below .001 to .095). As qualitative feedback, participants requested to be able to use the software without much preparation and suggested several desirable future features, such as performance tracking and goal-setting.

Conclusions: While the pitch training module should not be considered a complete GAVT package, TruVox represents a promising foundation for further GAVT software because it was perceived as usable and motivating and allowed participants to improve their exercise performance. To our knowledge, TruVox is the first GAVT software that combines real-time voice visualization with structured exercises, and this study represents the first quantitative human subjects evaluation of GAVT software. In the future, TruVox will be expanded with additional modules such as resonance training, then tested in longer-term trials.

(JMIR Form Res 2025;9:e73841) doi:10.2196/73841

XSL•FO
RenderX

KEYWORDS

transgender health; exercise technologies; serious games; voice training; usability; user experience; mobile phone

Introduction

Voice-Gender Incongruence and Gender-Affirming Voice Training

Transgender and gender diverse people (hereinafter abbreviated as transgender people) experience a significantly lower quality of life than the general public [1], often due to poor mental health and low self-esteem associated with gender dysphoria [2]. One cause of dysphoria is voice-gender incongruence: distress because the person's voice does not match the way their gender is perceived by themselves and others—for example, transgender women with low vocal pitch. Such voice-gender incongruence negatively affects quality of life for all transgender people [3,4]. However, it seems to be particularly an issue for transfeminine people (ie, transgender women and nonbinary people who were assigned male at birth) because hormone replacement therapy, which is commonly used by transgender people, does not affect transfeminine voices at all [5,6].

Voice-gender incongruence can be reduced using surgery, which can increase average voice pitch and perceived femininity in transgender women [7,8] as well as decrease average voice pitch in transgender men [9]. However, surgery is expensive, patient satisfaction is inconsistent [10], and there can be negative effects on other acoustic measures such as loudness and frequency range [7,8]. As a result, many transgender people are unaware that voice surgery is an option, and many of those who are aware of it would prefer a more noninvasive method [6].

The state-of-the-art noninvasive method for transgender voice modification is gender-affirming voice training (GAVT), which involves structured exercises that target different voice aspects and can be practiced over time to habituate a new speech pattern. Such GAVT is commonly provided by speech-language pathologists (SLPs) or transgender people with singing or music backgrounds and has been found to improve subjective self-perception of voice [11,12], others' perception of voice [13], and objective acoustic measures such as pitch, resonance, and intonation [12,14-16]. While GAVT was historically mostly done in person, teletherapy via Zoom (Zoom Video Communications) and other platforms has become increasingly popular [17], increasing accessibility.

Nonetheless, GAVT accessibility remains limited for many transgender people due to geographical constraints and a shortage of experts qualified for GAVT [18,19]. Even when GAVT is available, it is generally expensive (often over US $100/hour in the United States), requires many sessions (usually 15-25 visits [16]), is frequently not covered by health insurance [20,21], and is limited by psychological barriers such as anxiety, shame, and perceived provider bias.

Software Tools for GAVT

Given the limited accessibility of GAVT, there is an important opportunity for software that could help transgender people work on their voice by providing structured exercises, voice feedback, and other information. The software could be used as "homework" between expert-led GAVT sessions or even as a basic independent tool when human GAVT providers are not available. Such software is already used in other voice and communication training protocols besides gender affirmation [22-24] and is used for other conditions such as attention-deficit/hyperactivity disorder [23], obesity [25], and depression [26], so it should be applicable to GAVT. Indeed, such software has been praised as a promising research direction by the World Professional Association for Transgender Health [27]. Furthermore, transgender people are generally open to software solutions for other goals, such as improving adherence to HIV prevention strategies [28] and reducing minority stress [29].

The idea of science-backed GAVT software has existed since at least 2018 [30], and GAVT providers already use generic apps such as pitch monitoring tools (not originally intended for transgender people) to augment GAVT sessions, with positive results [31]. Additionally, several studies have studied the attitudes and needs of transgender people regarding GAVT software [6,30,32-35]. These studies have found positive attitudes and identified desirable features of such software: for example, it should provide visual feedback about the user's voice (eg, a pitch graph) while the user talks, combine it with exercises that are tailored to the user's voice goals, and track the user's improvement over time. Unfortunately, the state of the art in GAVT software is limited, and existing tools have not achieved widespread acceptance.

To our knowledge, the first science-backed GAVT app was Project Spectra [36], which implemented only limited functions before development stalled indefinitely. Another app, Attuned [37], consists primarily of text and video, with a pitch tracker added recently but not combined with any structured exercises, and no user studies have been conducted. Other than these 2 science-backed apps, some basic GAVT apps, such as EvaF.app and Christella VoiceUp, exist but have not been systematically evaluated and do not appear broadly accepted by transgender people [6,38]. Additionally, a few studies have explored other aspects of voice software for transgender people unrelated to GAVT (eg, automated gender classification and speaker verification [39] and automated voice changers for internet-based communication [40]), but these are not directly relevant to the current work.

Contribution of This Paper

This paper presents a prototype of TruVox: a free web-based GAVT platform that combines real-time visual voice feedback with structured voice exercises specifically for transfeminine people. While it is in the long term intended to target multiple aspects of voice, it currently focuses primarily on vocal pitch. Though transgender people correctly emphasize that excessive focus is placed on pitch in GAVT [6,30,36,41], pitch is nonetheless one of the main contributors to voice gender perception [42,43] and is easier to understand by laypeople than other contributors such as resonance [6,41]. It was therefore considered an appropriate first step, given the limited state of

XSL•FO
RenderX

the art, with a vocal resonance module to be added later. Furthermore, while transmasculine people are also interested in GAVT, we chose to initially focus on transfeminine voices because the voice is affected by hormone replacement therapy in transgender men but not transgender women [5,6], as transgender women appeared most interested in GAVT in our previous work [6,17], and as feminizing GAVT has been more thoroughly studied than masculinizing GAVT [27,44,45], making software development easier.

TruVox is currently publicly accessible [46], and its source code will be made public shortly. A much earlier prototype was presented as a conference paper and poster, but included fewer than half the current features and was only evaluated with limited qualitative interviews [47,48]. Prior papers on Project Spectra and Attuned focused extensively on the process of community-based design [36,37] and critical discourse analysis [38] of GAVT software, but provided only limited software features and did not conduct systematic evaluations. Though that research is valid and important, our paper aims to provide a complementary view by presenting specific pitch exercises and supporting materials, then evaluating them with transfeminine users to quantify perceived usability, users' intrinsic motivation, and exercise performance in a single-session study. It represents, to our knowledge, the first systematic evaluation of GAVT software that incorporates real-time feedback about the user's voice, and serves as a basis for future multisession trials of such software that could also evaluate long-term outcomes such as perceived voice femininity.

Researcher Positions and Terminology

The principal investigator, coauthor VDN, is a transgender femme who experienced GAVT as a client and has a background in serious games for motor rehabilitation. VSM, MEW, RLG, and TM are cisgender women and SLPs with experience in transgender research. SRW, OJ, XW, AB, TD, and RK have backgrounds in computer science or engineering and no prior experience with transgender research; RK is a transgender woman, SRW did not disclose their gender identity, and the others are cisgender men.

This paper uses identity-first rather than person-first language due to our team's and participants' preferences. We respect others' rights to their own preferences.

Methods

Overview

This section is divided as follows. We first present key features of the TruVox software as of January 2025. We then present a first remote evaluation that was carried out to gauge overall usability and identify significant usability issues. Finally, we present a later in-person evaluation that was carried out to gauge overall usability, users' intrinsic motivation in different exercises, and users' performance in the exercises.

Key Software Features

Overview

TruVox is compatible with most standard computer browsers, but does not currently have a mobile version. It is intended for speakers of American English, though many aspects are expected to apply to other dialects or languages. The main page [46] shows a welcome message, offers a tour, and lets users navigate further to pitch training, pitch and volume training, assessment, help, videos, and an about page.

The vocal pitch calculation algorithm used throughout TruVox is based on code from University College London [49]. It was validated during development by first playing pure tones at different frequencies, then playing recordings of the Rainbow Passage (a standard text in speech analysis) read by 16 transfeminine people before and after GAVT (obtained via our university's GAVT Clinic). Mean pitches calculated by TruVox were compared to those calculated by Praat (University of Amsterdam), a popular open-source software package for speech analysis [50]. For pure tones, the difference between the tone and TruVox's mean pitch was below 5 Hz. For Rainbow Passage readings, the mean difference between mean pitch in Praat and TruVox was 15.4 (SD 10.7) Hz, but this was found to be mainly due to Praat having better background noise rejection, so the algorithm was retained for this version of TruVox. Notably, all pitch calculations are done on the user's computer, with no audio data sent through the internet. This allows TruVox to even function if the user is not connected to the internet, and was requested in our prior interview study [6] as a method of protecting user privacy.

Pitch Exercises

As its key feature, TruVox has 5 pitch exercises that all use the same core pitch visualization graph. The graph plots the user's current and recent pitch values as a function of time as the user talks. It simultaneously shows a target pitch curve for the user to try to match, with the target type depending on the exercise. If the user's pitch is close enough to the target pitch, the target curve changes color to indicate success. A start or stop button allows voice recording to be activated or deactivated, and a retry button clears the graph; these buttons are present in all exercises.

The exercises, listed roughly from easiest to hardest, are Constant, Chanting, Stair, Human Curve, and Heteronyms. They are described below in more detail than in many human-computer interaction papers due to the limited state of the art in GAVT software.

The Constant exercise (Figure 1) is relatively simple: it asks the user to match a constant target pitch (presented as a horizontal line) while making any desired vocalizations. The target is 200 Hz by default (a stereotypically "feminine" pitch [16]) but can be adjusted with a slider, and a listen button allows the user to play a pure tone at the selected target. Users can also enable text that they can read during the exercise: they can select from several standard texts (Rainbow Passage, Grandfather Passage, and Harvard Sentences —all standard texts used in speech training and analysis) or type or upload their own text. The intended goal is for users to maintain an average pitch "near" the target pitch while reading, using the real-time display to adjust their voice.

XSL•FO
RenderX

Figure 1. Screenshot of the Constant exercise, with elements annotated. The (1) graph shows pitch in hertz, and the user tries to match the (2) target pitch with (3) their own pitch curve. The target pitch curve changes color when the user's pitch is close to the target. The user can (4) restart or (5) start or stop the exercise at any time, and can adjust the target pitch with a (6) slider. Next to the slider, (7) a button plays a pure tone at the target pitch for reference. The user can make whatever vocalizations they want, but can optionally (8) select from several different texts that are then (9) displayed below the graph. If the user selects a text with multiple sentences, they can use the (10) previous or next buttons to move between sentences.

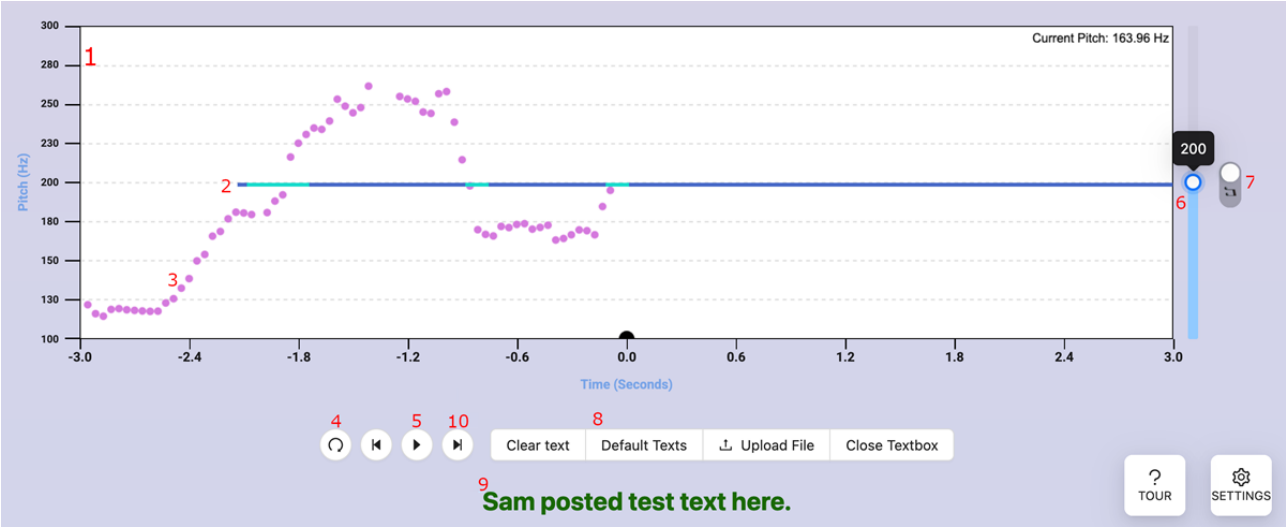

The Chanting exercise also asks the user to match a constant pitch, but does so in 3 stages. The user is first instructed to hum at a constant target pitch (again 200 Hz by default, presented as a horizontal line). Once the user is sufficiently close to the target for 1 second, the exercise automatically advances to the second stage, where the user is shown a short phrase and instructed to chant it at the target pitch. When done, the graph turns blank and the user is instructed to say the same phrase at a normal conversational pace. Once the user finishes, the exercise displays the mean pitch during the final (no-feedback) stage and allows the user to restart the exercise. Users can cycle between phrases using next or previous buttons. The exercise has 4 manually selectable "difficulty levels," which differ according to the presented phrases: the easiest level has resonant, easy-to-chant phrases (eg, "Mary made me mad") while the hardest level has nonresonant phrases that are harder to chant (eg, "No-one knew Tim too well"). Similar exercises are described in GAVT literature [44], and specific phrases are taken from the literature as well [51]. This exercise is intended to help the user carry over their target pitch from simple vocalizations (easier to do at a target pitch) to connected speech.

The Stair exercise (Figure 2) is more challenging in that it asks the user to match 5 consecutive pitch targets: a low, medium, high, medium, and low pitch. By default, these are set to 110, 155, and 200 Hz to move from a stereotypically "masculine" pitch to a stereotypically "feminine" pitch [16] and back, but both the low and high targets can be adjusted with the slider, with the medium target automatically set midway between low and high. Again, the user can make any desired vocalization, but a button optionally displays various premade 5-syllable phrases to be spoken for the exercise, with each syllable at a different pitch target. The idea for this exercise was gained during our previous interview study, which, among other things, asked transgender people to describe exercises they experienced in GAVT [6]. Its purpose is to train vocal flexibility and prosody.

XSL•FO
RenderX

Figure 2. The Stair exercise, which asks the user to match a series of 5 pitch targets. As in the Constant exercise, the user's pitch is by default shown in pink, while the target pitch curve is shown in blue. The target pitch is adjustable with the slider on the right. A show or hide button below the display optionally shows a 5-syllable phrase for the user to speak while matching the 5 pitch targets, and users can move between phrases with next or previous buttons. See Figure 1 for detailed annotation of basic elements.

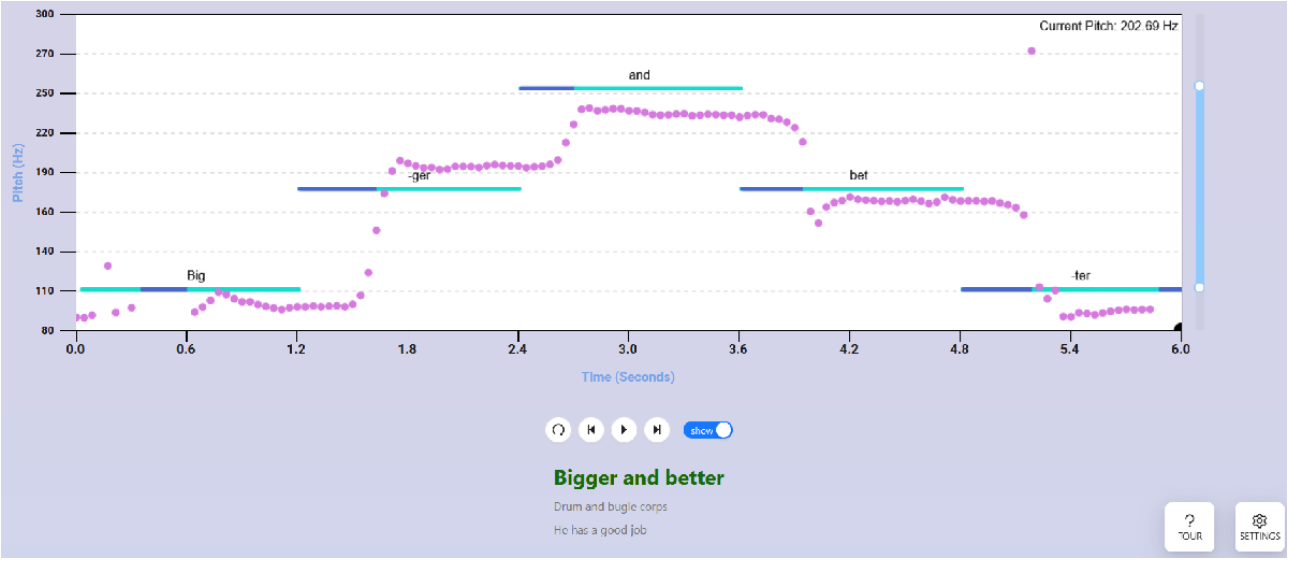

The Human Curve exercise (Figure 3) asks the user to match the pitch curve of a previously recorded human speaker ("vocal model") speaking a 2- to 5-syllable phrase (manually selectable). A listen button below the graph lets users play the recording of the vocal model speaking the current phrase, and next or previous buttons let users cycle between phrases. At this time, 9 vocal models can be selected from the settings menu; they are all cisgender women and native speakers of Mainstream American English. Vocal models 1-7 are aged 20-30 years, model 8 is in her early 50s, and model 9 is in her late 60s. Eight models are White, while the remaining one is biracial (African-American and White). This exercise was developed to train natural feminine intonation patterns through imitation.

Figure 3. The Human Curve exercise, which asks the user to match the pitch curve of a human "model" while speaking a target phrase. As in the Constant exercise, the user's pitch is by default shown in pink while the target pitch curve is shown in blue. A listen button (which looks like headphones) below the display lets the user listen to the actual speech recording the pitch curve is derived from, and users can move between phrases with next or previous buttons. See Figure 1 for detailed annotation of basic elements.

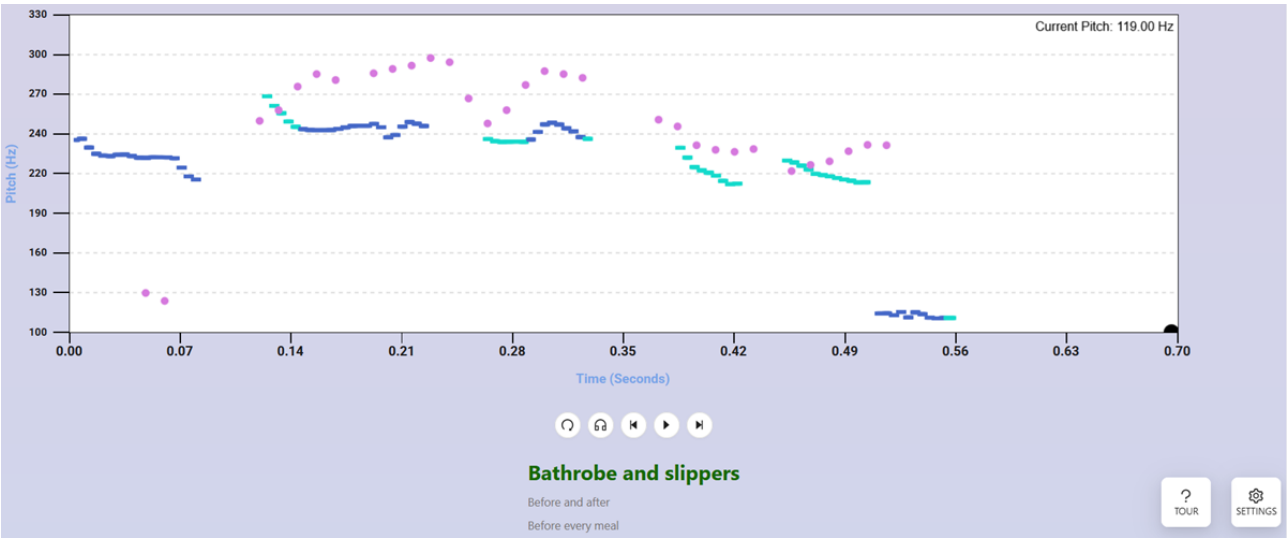

The Heteronyms exercise differs from the others in that the user is not expected to precisely match a target pitch or even intensely practice the exercise—instead, it is intended to teach users about upward and downward intonation because upward intonation is correlated with perceived voice femininity [27,43]. The exercise does this by presenting pairs of heteronyms: pairs of words that are spelled the same but pronounced differently, with stress on the first versus the second syllable. An example heteronym is the word "console." If the stress is placed on the first syllable, it is a noun; if the stress is placed on the second syllable, it is a verb. TruVox presents the word below the pitch graph and capitalizes the stressed syllable. The pitch graph then shows a pitch curve of a vocal model speaking the word; these models are the same as those in the Human Curve exercise, and the user can again listen to the recording with the listen button. Users can switch between the 2 variants of the word with a button. Additionally, the user can switch to a sentence display. In this case, TruVox presents an example sentence using that

XSL•FO
RenderX

word, the pitch graph shows a pitch curve of the vocal model speaking the sentence, and users can switch between the 2 variants of the word. Example sentences for the word "console" are "They bought a new CONsole" and "The doctor will conSOLE his patient." The idea for this exercise was again gained during our previous interview study [6].

In all 5 exercises, a settings button next to the pitch graph allows users to modify several options: the pitch range (adjustable between 50-600 Hz), time scale (from 2.5 to 15 s), pitch display unit (Hz or musical notes), color theme (default, dark, or colorblind), and the speed with which the pitch indicator moves from left to right (from very slow to very fast). There is also a toggle to automatically start an exercise when speech is detected by the microphone (as opposed to starting manually with the start or stop button), with an adjustable volume threshold for starting. Finally, an "advanced features" setting allows rudimentary performance tracking to be activated, which was used in the in-person evaluation (presented later). Below each exercise is a "How to use" section that provides text instructions for the exercise, an embedded demonstration video (see next section), and a description of the settings.

Other Features

While the pitch exercises represent the core part of TruVox at this time, several other features are available.

The Pitch and Volume module contains the same Constant and Stair exercises described above, with the graph again showing pitch on the y-axis. However, the graph now also shows the volume (loudness) of each recorded sound sample. Each sample is marked as quiet, medium, or loud, and the user can select whether they want to visually indicate the volume of each sample with color (blue=quiet, purple=medium, and red=loud), dot size (small=quiet and large=loud), or both. This feature was added because participants in our previous interview study stated that they often have trouble simultaneously speaking loudly and maintaining their desired pitch [6]. It has not been extensively developed at this time and serves mainly as a demonstration of visualizing multiple voice aspects simultaneously.

The videos module contains 3 videos to help users warm up their voice (breathing, resonance, and vocal stretching) as well as demonstration videos of all 5 pitch exercises. The warmup videos were recorded with a professional voiceover and partially show coauthor RLG demonstrating warmup motions. The demonstration videos were recorded and narrated by coauthor VDN using screen sharing functionality in Zoom, with the TruVox display taking up most of the screen and video of VDN visible in the top right. The more amateur nature of demonstration videos is because they were rerecorded several times during development as features were added and visual elements were changed.

The assessment module first asks the user to hold an /i/ sound for 5 seconds, then read the Rainbow Passage. It then calculates the user's mean pitch in both vocalizations. Further assessment features (eg, pitch range and resonance) are planned for future versions.

Finally, the help module contains background information about voice as well as text descriptions of different modules and embedded exercise demonstration videos.

Evaluation Summary

TruVox was evaluated in single-session studies carried out both remotely and in person, with the remote evaluations carried out in 2 consecutive stages. Details of the evaluations are presented in the next 2 sections (first remote and then in-person), and Table 1 summarizes key characteristics of each evaluation. All evaluations used a single-arm design where all participants in the evaluation experienced the same study protocol with the same software.

Table 1. Key characteristics of consecutive TruVox evaluations.

	Remote evaluation		In-person evaluation
	First stage	Second stage	
Data collection time-frame	Spring 2024	Fall 2024	November 2024 to January 2025
Goal	Identify issues and gather feedback	Verify usability and gather feedback	Verify usability, user motivation, and exercise performance
Outcome measures	Qualitative feedback only	SUS[a], qualitative feedback	SUS, Intrinsic Motivation Inventory, exercise error
Participants	13 completed	8 screened, 6 completed	26 screened, 21 completed
Overall protocol	Participants use the software alone and are then interviewed	Participants use software alone, give open-ended written feedback, and fill out the SUS	Participants use software in the laboratory, fill out the SUS, perform 10 repetitions of each exercise to measure exercise error, and periodically fill out the Intrinsic Motivation Inventory
Main statistical tests	None	SUS over 70?	Is SUS over 70? Is exercise error lower in later repetitions?
Already published?	Yes [47,48]	No	No

[a]SUS: System Usability Scale.

XSL•FO
RenderX

Remote Evaluation

Overview

The goal of the remote evaluation was to identify key usability issues to be addressed and to gauge overall usability in a quick and convenient manner without requiring participants to come to our laboratory. It was carried out in 2 stages.

First Stage

The first evaluation stage was carried out in spring 2024 with an early version of TruVox that only had 3 pitch exercises (Constant, Stair, and Human Curve), a less polished visual presentation, no assessment module, no videos, no introduction page (the current help page acted as the default introduction page), and only 7 of the current 9 vocal models for the Human Curve exercise.

This prototype was first presented to 5 transfeminine participants. A link to the web page and an informed consent form were provided electronically, and participants were asked to interact with the software for 15-20 minutes on their own computer. After that, participants engaged in a brief interview with coauthor VDN to discuss usability issues and general impressions. The results of this stage were purely qualitative and were published in the prior conference paper [47], so this stage is not described further.

Additionally, the prototype was presented to 8 transfeminine participants to specifically gauge their reactions to the 7 vocal models in the Human Curve exercise. Methods and results were presented as a conference poster [48]; in brief, participants generally had positive reactions but requested a broader range of ages and dialects, some transfeminine vocal models, and some models with lower pitches.

Second Stage

The second stage was carried out in fall 2024 and involved a version of TruVox similar to the one described in the Key Software Features section—no major functional changes were made after this stage, only usability improvements such as reducing unnecessary space, rerecording videos, editing the help page text, and rearranging elements on the pages.

Eight people were recruited for this study via a Discord (Discord Inc) server for transgender people in Ohio and a Health Communication Disorders Clinic at the University of Central Florida. Two did not complete the protocol (did not return the forms), leaving 6 participants: 5 transgender women and 1 transfeminine xenogender participant. All 6 were native speakers of American English and were aged 21, 26, 29, 30, 33, and 34 years. Three reported previous GAVT experience with a human expert. The sample size was chosen because a simple guideline suggests that 5 participants are sufficient to identify most usability issues in first prototypes [52].

Participants were again electronically sent a link to TruVox and an informed consent form. They were given a list of 14 website components (introduction, tour, Constant exercise, settings menu, Stair exercise, Chanting exercise, Human Curve exercise, Human Curve—specifically vocal models, Heteronyms exercise, assessment page, help page and help sections, videos, about page, and any additional comments) and asked to provide open-ended written comments about each component. They could test the components in any order they desired and could write down their thoughts either while testing each component or at the end. Furthermore, they were asked not to consider the Pitch and Volume exercises as they were incomplete. Similarly, they were told that resonance exercises would be available in a future version because vocal resonance is considered important by transfeminine people [6,41], and we expected that participants would criticize TruVox if resonance was not acknowledged. Finally, they were told that performing the exercises may cause vocal strain or fatigue or psychological discomfort (also noted on the consent form) and that they should feel free to take a break or discontinue participation if this occurs.

After providing open-ended feedback about the components, participants filled out the System Usability Scale (SUS) [53] to gauge the perceived usability of TruVox as a whole. SUS consists of 10 items that are all 5-point Likert scales and are combined into a single score by first reversing the negative items, then scaling the sum from 0 to 100. As a common SUS benchmark, a score of 70 is commonly considered “acceptable” for a first prototype [54]. Thus, the SUS score was the primary outcome measure, and we planned to continue iterative development and usability testing with additional participants if TruVox did not reach an SUS score of 70 in this stage.

Participants were not monitored while using the software or filling out the questionnaires, though a single reminder was sent if they did not return the forms within 72 hours. Blank copies of the informed consent form and full questionnaire provided to participants are available via the reference citation in the Data Availability section.

In-Person Evaluation

Participants

Study advertisements for the in-person evaluation were distributed to 2 Discord servers and a Facebook group for transgender people in southwest Ohio, a GAVT clinic run by University of Cincinnati Health (which coauthor RLG is affiliated with), the student mailing list of the University of Cincinnati College of Engineering and Applied Sciences (which coauthor VDN is affiliated with), and student and employee LGBTQ+ (lesbian, gay, bisexual, transgender, queer, and others) groups at the University of Cincinnati. Each advertisement included a brief description of this study’s purpose and procedure, a screenshot of TruVox (but no web link), emphasized that participants would receive gift cards but would be unlikely to see voice improvements due to participation, and directed potential participants to contact VDN by email. Readers were also encouraged to forward the advertisement to others who might be interested. Once potential participants contacted VDN by email, VDN answered any questions and verified if participants met inclusion or exclusion criteria (aged 18+ years, assigned male at birth, identify as transgender, native American English speaker, have sufficient hearing and vision, have no diagnosed speech, language, or hearing problems besides voice-gender dysphoria, no neurological issues with larynx, and no prior voice feminization surgery). Computer or internet

XSL•FO
RenderX

literacy was not used as a formal inclusion or exclusion criterion, though it was likely an implicit one. If participants fulfilled the requirements, they were scheduled for a session.

A total of 26 people contacted VDN about this study. One did not meet the inclusion criteria, while 4 stopped responding to emails before scheduling a session, resulting in 21 participants who all completed the in-person session. Further, 16 were transgender women (she/her or she/they pronouns), 3 transfeminine nonbinary or transgender femmes (she/they pronouns), 1 nonbinary (they/them pronouns), and 1 uncertain of gender (he/him pronouns). They were aged a mean of 30.2 (SD 8.0) years, minimum of 19 years and, maximum of 42 years. Furthermore, 19 were White, while 2 were biracial. A sample size of 21 is higher than or similar to most formative studies of GAVT software and transgender-related software, which mostly use 4-20 participants [6,30,31,33-35]. The first participant completed their session in late November 2024, while the last participant completed it in late January 2025.

All participants were native speakers of American English; 20 reported their dialect as Midwestern or Cincinnatian, while 1 reported it as Standard American English. Dialect data were collected because our previous work suggested that the suitability of GAVT exercises may vary between dialects [6]. When asked about previous experience with voice training, 12 reported having attended GAVT sessions with a professional, 7 reported no professional GAVT but some self-guided GAVT, 1 reported no GAVT but previous singing training, and 1 reported no experience. This information was collected because we expected that reactions to GAVT software would depend on prior GAVT experience [6].

Study Protocol

The sessions were carried out in a quiet room on the University of Cincinnati campus by coauthors VDN and MEW, with VDN present at all sessions and MEW present at 16 of 21 sessions. TruVox was presented on a laptop (Lenovo LEGION Y920 Notebook with 17.3” screen), with the participant’s voice input to TruVox via the laptop’s built-in microphone. Throughout the sessions, a video of the screen was also taken using screen capture software, together with voice recordings, so that any usability issues could be easily tracked. To increase confidentiality, all video and audio were recorded locally to the laptop and not automatically backed up to a cloud server.

The participant filled out a demographics questionnaire (see previous subsection) while audio and video recordings were started. They were also explicitly made aware that vocal strain or fatigue or psychological discomfort may occur due to performing the exercises, and that they were free to take breaks or discontinue participation if necessary. Throughout the session, the experimenters also monitored participants for discomfort, occasionally suggested they take breaks to reduce vocal fatigue, and provided disposable water bottles if needed.

In the first half of the session, the participant completed a usability evaluation. If MEW was present, she conducted this evaluation while VDN left the room for the duration of the evaluation unless she was temporarily needed for technical support; if MEW was absent, VDN conducted this evaluation. As in the remote evaluation, participants were given a list of website components and were encouraged to test TruVox in approximately that sequence, though they could deviate if desired. As in the remote evaluation, they were told not to test the Pitch and Volume module and that a resonance module was in development. They were encouraged to provide open-ended spoken comments and were regularly prompted for feedback with general open-ended questions (eg, “What did you think?”). If participants asked for clarification about how to do something, the researchers told them to try and find it through the information available in TruVox, because we wanted to see if TruVox was usable without expert guidance. Participants were further instructed not to perform too many repetitions of each exercise during this stage, with the explanation that there would be a formal exercise component afterward. At the end of the usability evaluation, participants completed the SUS [53].

The second half of the session was conducted by VDN, with MEW (if present) providing support. In the second half, the participant completed 10 repetitions of each of 7 exercises based on Constant, Chanting, Stair, and Human Curve exercises. Each exercise was first demonstrated 1-2 times by VDN, and the participant then performed the repetitions with no further familiarization. These exercises were:

- Constant—/i/: participants held an /i/ sound at a target frequency of 200 Hz in the Constant exercise for 5 seconds.
- Constant—Rainbow Passage: participants read the second sentence of the Rainbow Passage (“The rainbow is a division of white light into many beautiful colors.”) at a target frequency of 200 Hz in the Constant exercise.
- Chanting: participants completed the Chanting exercise at level 1 and target frequency of 200 Hz with the phrase “Mary made me mad.”
- Stair—/i/: participants held an /i/ sound at target frequencies of 110, 155, and 200 Hz in the Stair exercise, with the duration of each repetition being 6 seconds. Automatic start was enabled for both Stair exercises to help participants better match the timing.
- Stair—bigger and better: participants said the phrase “bigger and better” at the same target frequencies in the Stair exercise, with each syllable at a different target (“Big-ger and bet-ter”) and the total duration being 6 seconds.
- Human Curve—model 1: participant said the phrase “bathrobe and slippers” while trying to match the pitch curve of vocal model 1 at maximum pitch indicator speed. Automatic start was disabled, and participants were given the mouse so they could start the exercise and listen to the model recording at their own pace; this preference was identified in pilot trials.
- Human Curve—self-selected model: the participant previewed all the vocal models and was told to select one they liked, with no further specific clarification. They could not select model 1 again. They then performed the same exercise as the previous one with the new vocal model.

In each exercise, a repetition was redone if the participant clearly made a mistake (eg, coughing or laughing midrepetition). The exercises were performed in a fixed order as described above; while we considered randomization, we ultimately chose an order that roughly corresponds to increasing exercise difficulty

XSL•FO
RenderX

because participants may not be able to perform the harder exercises at all without experience with simpler ones. A target of 200 Hz was used as a stereotypical "feminine" pitch [16]. We originally considered providing a personalized target for each participant, but eventually chose not to do so because there is not yet a clear rule for setting participant-specific goals, and this may make data analysis more complex. We also considered a lower target in the gender-neutral range because this is sometimes recommended to reduce vocal strain in early GAVT stages [27], but decided against it because participants in pilot trials did not show major strain.

Participants filled out the Intrinsic Motivation Inventory (IMI—same 8-item version as in our previous work [55,56]) 4 times: after the Constant exercises, after Chanting, after the Stair exercises, and after the Human Curve exercises. Each time, they were instructed to rate their experience with that exercise type (eg, both Human Curve exercises). The IMI was selected because our previous studies found that it can identify differences in subjective reactions to somewhat similar tasks. After completing the last exercise, participants were given the US $20 gift card, asked for any final feedback, and thanked for their participation.

Blank informed consent forms and questionnaires are provided via a reference citation link in the Data Availability section. The same link also contains a copy of the experimenters' in-person session checklist that was written for internal use and is published unedited.

Data Analysis

SUS scores were compared to the benchmark score of 70 using a 1-sample *t* test (2-tailed). For the IMI, the 8 items were converted to 4 scales: interest/enjoyment, effort/importance, perceived competence, and pressure/tension (2 items per scale). These 4 IMI scales were then compared between the 4 exercise types using repeated-measures ANOVA. As there were 4 IMI ANOVA, the Benjamini-Hochberg procedure [57] with a base α of .05 was used to reduce the false discovery rate within these 4 tests. ANOVA that were significant after the Benjamini-Hochberg procedure were analyzed with post hoc Holm-Sidak tests.

As a measure of exercise performance, the built-in TruVox code was used to calculate the mean absolute difference between the target pitch values and the participant's pitch values over the course of each repetition. For Stair and Human Curve, 1 repetition was defined as being from the time recording was started to the time it ended automatically. For Constant /i/, it was defined as 5 seconds from the time the target pitch indicator reached the participant's pitch indicator. For Constant Rainbow Passage, it was defined from the time the target pitch indicator reached the participant's pitch indicator to just before the end of the spoken sentence. For Chanting, it was defined as the final phase of the exercise (when visual feedback was removed). This mean absolute difference is hereafter called the "exercise error" and reported in hertz. Each participant's mean exercise error was calculated over the first 5 repetitions and over the second 5 repetitions of an exercise. Participants' mean exercise errors were then compared between the first 5 repetitions and the second 5 repetitions of an exercise using 2-tailed paired *t* tests (or signed rank tests in the case of normality violations) to see if the errors decreased. The Benjamini-Hochberg procedure [57] with a base α of .05 was used to reduce the false discovery rate within the 7 paired *t* tests (for 7 exercises). Similar metrics (mean absolute difference between target and actual value) are used to measure improvement in fields such as limb motor learning [58].

Participant statements and experimenter observations are reported descriptively. To analyze participant statements during the sessions, audio (but not video) recordings were uploaded to a secure password-protected university server in a folder accessible only to VDN, MEW, and VSM, who then listened to the recordings, made notes of salient topics, and compared their impressions. No systematic qualitative analysis method (eg, thematic analysis) was used because participant comments were generally not extensive and the topics tended to repeat between participants. Following the analysis, audio recordings were removed from the server, and both audio and video recordings were backed up on 2 local hard drives in VDN's laboratory for long-term storage.

Finally, we note that improvements in perceived voice femininity were not evaluated because these are commonly evaluated after a full GAVT regimen of 10+ sessions [12,13,16,27] and are thus beyond the scope of a single-session feasibility study.

Ethical Considerations

All procedures were approved by the Institutional Review Board of the University of Cincinnati (#2023-0831), and were conducted per the principles of the Declaration of Helsinki. Upon the participant's arrival, the purpose and procedure of the study were explained, and the participant signed an informed consent form. All participants were explicitly made aware of the recordings and how they would be stored and analyzed, as well as how other data would be analyzed and shared in a deidentified manner. Each participant received a US $20 Amazon gift card upon completing the evaluation.

Results

Remote Evaluation

As mentioned, the results of the first stage were largely qualitative and are described in prior publications [47,48]. In the second stage, the 6 participants reported SUS scores of 95, 60, 80, 77.5, 82.5, and 62.5. We thus proceeded to in-person testing after addressing significant usability issues identified in this stage.

One key usability issue listed by the 2 participants with lower SUS scores was insufficient explanation of website components for unsupervised use. For example, participants requested more tour steps, more mouseover popups for different elements, and more emphasis on optional features that were not obvious (eg, ability to automatically start the exercise or slow down or speed up the horizontal pitch indicator movement). In some cases, these components were not explained at all; in other cases, the explanations were present in either the videos or help page, but were not found by participants. This issue was echoed by the other participants.

XSL•FO
RenderX

As another issue, the participant who gave an SUS score of 60 noted that the Stair exercise "at default settings … is stressful and feels like I'm failing, impacting my desire to do this exercise negatively." She similarly noted that the Human Curve exercise "feels impossible – no matter what I tried (including adjusting the Pitch Indicator Speed in Settings) I could not get my pitch to match the recording at all." She continued to say that the Human Curve exercise "doesn't feel discouraging to try like the Stair Exercise did at first, just feels ill-tuned and impossible to complete since the timing is not adjusted and feels super awkward to try and get that right." She noted that the Heteronyms exercise "suffers from the same issues as the Human Curve exercise - feels incredibly difficult to find the right timing to match the graph and vocal model." When queried further, the participant clarified that the challenge lay in both matching the timing (eg, intonation in the Human Curve) and matching the specific target pitch (eg, the high pitch of the default vocal model 1 in the Human Curve), and that she did not feel able to adjust the settings to an achievable target. As a possible solution, she suggested that "the site could benefit from automatically adjusting speed and pitch settings in the Human Curve and Heteronyms exercises based on the phrase and voice model rather than putting that adjustment on the trainee". The other participant who gave a low score had similar concerns regarding timing, noting that they "seem to speak words too fast for us to get a massive bit of usage out of this tool." Other participants generally agreed with these issues but appeared to be able to address them by adjusting the settings.

As other significant issues, participants mentioned a lack of optimization for different screen resolutions, a lack of video polish, a lack of volume control in the sound playback, an inconsistent vocal model module, a bug in the pitch calculation, and a bug with some buttons not working consistently. These issues were iteratively addressed during and after this evaluation stage.

In-Person Evaluation

Quantitative Results

The mean SUS score across all participants was 79.8 (SD 12.8), with the lowest value being 45 and the second-lowest being 65. The 1-sample *t* test showed that the score was higher than the benchmark 70 (t_{20}=28.5, *P*<.001).

IMI results for different exercises are shown in Table 2. ANOVA found that:

- Interest/enjoyment did not significantly differ between exercises (P=.26).
- Effort/importance differed between exercises (P<.001). Effort/importance in the Constant exercise was lower than in all other exercises (P<.01 for all 3 comparisons), and effort/importance in Chanting was lower than in Stair and Human Curve (P<.05 for both comparisons). Perceived competence differed between exercises (P=.007). Competence was lower in the Human Curve exercise than in Constant and Chanting exercises (P<.03 for both comparisons).
- Pressure/tension did not differ between exercises (P=.92).

Table 2. Intrinsic Motivation Inventory results for the 4 exercise types in the performance evaluation, presented as mean (SD) across 21 participants.

	Constant, mean (SD)	Chanting, mean (SD)	Stair, mean (SD)	Human Curve, mean (SD)
Interest/enjoyment	12.0 (2.0)	12.7 (1.4)	12.0 (2.3)	11.5 (2.9)
Effort/importance	11.2 (2.0)	12.2 (1.8)	12.9 (1.3)	12.8 (1.4)
Perceived competence	10.6 (2.5)	10.4 (3.0)	9.8 (2.9)	8.7 (3.3)
Pressure/tension	5.4 (2.4)	5.3 (2.7)	5.7 (3.6)	5.5 (3.5)

Exercise performance results are shown in Table 3. Further, 6 of 7 exercises showed a significant reduction in exercise errors (*P*<.05) between the first 5 and second 5 repetitions. When choosing a preferred model (other than model 1) in the Human Curve exercise, 4 participants chose model 2, 1 chose 3, 4 chose 4, 3 chose 5, 2 chose 6, 5 chose 7, 2 chose 8, and none chose 9. It is possible that model 9 was never selected because she is in her late 60s and thus more than 20 years older than the oldest participant.

XSL•FO
RenderX

Table 3. Exercise errors in the first 5 repetitions and the second 5 repetitions of the 7 exercises, presented as mean (SD) across 21 participants. *P* values are the results of tests comparing exercise errors between the first 5 and the second 5 repetitions.

Exercise type	Exercise error (Hz)		*P* value
	First 5 repetitions	Second 5 repetitions	
Constant—/i/	5.3 (4.3)	4.1 (4.0)	.005[a]
Constant—Rainbow Passage	26.8 (13.6)	20.8 (12.3)	.02[a]
Chanting	6.1 (5.1)	5.0 (5.0)	.003[a]
Stair—/i/	18.5 (9.3)	13.8 (7.2)	<.001[a]
Stair—bigger and better	13.4 (5.4)	12.4 (5.3)	.095
Human Curve—model 1	47.5 (23.7)	40.2 (17.8)	<.001[a]
Human Curve—preferred model	57.8 (32.7)	50.0 (28.4)	.003[a]

[a]Significant *P* value after Benjamini-Hochberg correction.

Qualitative Results

As in the remote evaluation, participants mainly raised usability issues regarding the ability to use the different exercises unsupervised. For example, opening the help page, 1 participant commented, “nobody’s going to read all that,” and requested more introductory components that would guide participants to exercises appropriate to their skill level. Similarly, 3 participants requested an overall introductory video that would introduce them to the website and explain which exercises to use. Notably, several participants stated that they did not know what the target pitch should be and requested more information about what a “feminine” pitch was. For example, 1 participant who had not done GAVT explicitly stated “I don’t have a target voice yet” and asked “how do I know what pitch to go for?” This target information was originally intentionally not emphasized due to concerns that users may be demotivated by knowing they are not in a “feminine” pitch range [6].

The Constant, Stair and Chanting exercises were generally well-received. In Stair, participants primarily expressed frustration regarding matching the timing of pitch changes; those who did not view the help page or demo videos often requested the ability to change pitch indicator speed or add an automatic start function. Additionally, 1 participant suggested an optional “delay” where the exercise would wait approximately half a second after pressing start before beginning. With Chanting, participants primarily had trouble understanding how to advance through the phases and what to do in each phase. For example, they stated that the change in phases should be more obvious (some participants did not notice it) and requested clarification about whether they should continue chanting in the final phase, whether they should speak freely at the target pitch, or whether they should speak freely at their normal pitch. Finally, when performing the Stair exercise in the second half of the session, several participants who had previously undergone GAVT noted that the lowest pitch target was difficult for them to achieve post-GAVT and that these targets should be either raised overall or adjusted for each participant.

The Human Curve and Heteronyms exercises received more mixed opinions. In Human Curve, participants generally understood the concept but experienced several challenges. Some participants had difficulty achieving the vocal models’ higher pitches, while others could reach the target pitch but could not match the cadence of the phrase. Some questioned the utility of matching a specific model’s pitch curve, with one commenting, “I don’t need to sound like someone else.” A few participants requested a larger range of vocal models, who were all cisgender women and mostly White. The Heteronyms exercise received the most mixed opinions: while it was intuitively understood by some participants, others either did not understand that the goal was not to perfectly match the target pitch curve or did not understand that they could toggle between 2 intonations or between words and phrases. One participant commented, “I don’t understand what the point of this is” regarding Heteronyms, with others expressing similar confusion.

Concerning exercise demo videos, participants generally found them informative but had mixed opinions regarding presentation. For example, several participants were happy to have the video-in-video of coauthor VDN to emphasize that the project is transgender-led, but others found it distracting. Participants also requested the videos to be more concise so that they would require minimal time to glean information. This was not a concern with the warmup videos, which were received positively with little criticism.

Concerning missing features, participants primarily requested a resonance training module, a more detailed assessment function (with more outputs besides mean pitch as well as some characterization of their voice concerning masculine or feminine ranges), the ability to play back voice recordings in exercises, better exercise performance metrics, more information about what a feminine or masculine pitch would be, and a “goal setting” module. Participants had different ideas for a goal-setting module, but the most common suggestions were a combination of automated voice analysis and questions asked by the software (eg, “how much do you know about voice?” and “how do you want to sound?”) that would then suggest specific exercises and targets or vocal models.

Concerning exercise performance, participants often noticed the exercise error shown on the screen as part of the advanced features; while the error was intended for the experimenter (who

XSL•FO
RenderX

wrote it down on paper), participants had several reactions to it. Most commonly, participants expressed satisfaction when their exercise error decreased, disappointment if it increased, and sometimes asked the experimenter to show them the paper error tracking form so they could see if they had improved over time. Several participants explicitly tried changing their vocal strategies to reduce the displayed error. Notably, 1 participant expressed disappointment when she achieved an exercise error of 0.6 Hz in the first repetition of Chanting, stating that "now I'm definitely not going to get better."

Several participants compared TruVox favorably to Voice Tools, a smartphone app by DevExtras (available on the Google Play store) that tracks pitch and indicates associated gender. Participants primarily noted that they appreciated the structured exercises and that they liked that TruVox did not "gender" their pitch. For comparison, Voice Tools has a blue-colored range for "masculine" pitch and a pink-colored range for "feminine" pitch. While TruVox does use blue and pink in its base color scheme (Figures 1-3), it does not use them to convey femininity information, and different colors are used in dark and colorblind schemes. Other apps were not brought up during sessions, and Voice Tools was likely mentioned because it is used as part of SLP-guided GAVT at the University of Cincinnati Health GAVT clinic, a participant recruitment location.

Discussion

Principal Results

The in-person evaluation found encouraging results. First, the mean SUS was 79.8 (SD 12.8), which is approximately a B on a scale from A (best) to F (worst) [54] and significantly surpasses our initial goal of 70. The worst scores were obtained on the items "I found the website unnecessarily complex" and "I needed to learn a lot of things before I could get going with this website," which is likely related to qualitative observations of participants being overwhelmed by the amount of information or having trouble finding the information they needed.

Second, IMI scores were generally good (Table 2—interest/enjoyment and effort/importance consistently over 11 of 14, pressure/tension consistently below 6 of 14). While IMI results are difficult to benchmark because there are many different variants of the IMI, they can be compared to our previous studies with the exact same 8-item IMI. In a study of a 2-player game, for example, we found that, on average, interest/enjoyment was about 10 and effort/importance was about 11 [55]. In a study of a resource management scenario with difficulty adaptation [56], we found interest or enjoyment values around 10, effort/importance around 12, competence around 9, and pressure/tension around 10. Thus, participants found the TruVox exercises enjoyable, challenging, and not very stressful. Stair and Human Curve required more effort than Constant and Chanting, which was also supported by qualitative observations: several participants commented that those exercises felt harder. As a result, participants also felt less competent in the Human Curve. This was unsurprising because natural progression through GAVT exercises also tends to begin with easier exercises, similar to Constant and Chanting.

Third, participants' performance significantly improved throughout 10 exercise repetitions in all exercises except Stair—bigger and better (Table 3). This is a promising result because it indicates that participants can learn the exercise and improve at it with a limited number of repetitions. In our opinion, the lack of statistical significance (P=.095) in one exercise is not very concerning because the other exercises do show improvement, there was a numerical difference in the expected direction between the first 5 and second 5 repetitions, and there are several reasons (discussed in the Study Limitations section) why performance may not show consistent improvement. The significant decreases in exercise error (ranging from about 1 Hz to 8 Hz—Table 3) are admittedly difficult to contextualize because no other GAVT software has been quantitatively evaluated, and standard SLP-guided GAVT regimens do not quantify errors on the level of individual exercise repetitions or even individual sessions. However, studies of human vocal pitch perception suggest that a pitch difference of 3-8 Hz in normal speech would be noticeable to typical adults [59,60], so the improvements observed during most of our exercises should be perceptible. As another comparison, a typical increase in mean pitch due to voice feminization surgery or a full GAVT regimen (over 10+ weekly sessions) is 25-30 Hz [7,8,12,16], which is higher than the improvements seen in our session and would be expected to generalize across all of a speaker's utterances. It is possible that long-term practice with the software would indeed lead to generalized voice improvements (eg, perceived as more feminine), but this is impossible to estimate from single-session results because classic GAVT regimens do not quantitatively track performance on a per-session basis and thus cannot be used for comparison. Nonetheless, while a positive single-session result does not guarantee long-term improvement, we do feel that a negative result here would have indicated that the exercises are not very suitable for users.

Overall, the evaluation found TruVox to be usable and motivating and showed that participants were able to improve their performance in the exercises over time. While a full version of the software would need to target several voice aspects besides pitch and would be tested over a much longer period, the current evaluation suggests that the TruVox pitch module is a promising foundation that can be built upon for future technology-aided GAVT.

Next Development Steps

Participants noted several possibilities for software improvement, and several other desirable features of GAVT software are already known from previous interview studies of transgender people [6,30,32-35]. In the future, we will thus expand TruVox with such features and recommend them for other GAVT software as well.

Notably, several participants requested a goal-setting module, and such a module may be developed based on prior work on individualized goal setting in GAVT [35]. Alternatively, it may be based on voice training software for applications outside GAVT, which has already used intelligent adaptation rules to set goals based on user performance [61]. We will also add user accounts and link the goal-setting module to them so that users'

XSL•FO
RenderX

goals can be stored between sessions and periodically reevaluated. The accounts will also let users visualize their voice improvements over time, which represents a potentially strong motivator [6]. The desire to track improvement was also observed qualitatively during our in-person evaluation, where participants kept track of their performance and commented on improvements. As an additional feature, we will also increase the diversity of vocal models. For example, transgender women and nonbinary people of different ages will be recruited to serve as models, and models with a larger variety of dialects will be recruited because model dialects are known to be important in GAVT [48].

We will also develop modules to target voice aspects besides pitch. As a key feature, transfeminine people often consider resonance to be more important than pitch [6,41]. We are thus developing a resonance module based on a previous laboratory-based software implementation by coauthor TM [62], and a preliminary paper about this module was recently accepted to a conference [63]. However, this preliminary study [63] suggests that a resonance module requires more explanation because quantitative metrics of resonance, such as formant frequencies [62], are not as easily understood as pitch metrics like musical notes. Users also do not know how to change their vocal resonance and need more instruction to effectively use the module—our current implementation advises users to watch about 10 minutes of instruction videos before exercising, which may conflict with the observed desires to use the software with minimal preparation. These issues will be carefully considered when further developing the resonance module and other modules, such as the currently incomplete Pitch and Volume module. That module was requested in our prior interview study [6] because transgender people often have trouble simultaneously maintaining a target pitch and target loudness, but simultaneous visualization and training of multiple aspects of a motion is a significant challenge in motor learning [64]. Similarly, our pilot resonance module tests [63] suggest that simultaneously visualizing pitch and resonance may be helpful, but this would again need to be carefully designed to avoid overwhelming users.

Separately from exercise modules, we will also improve the underlying acoustic processing algorithms. As mentioned, our current pitch calculation algorithm (based on code from University College London [49]) achieved about a mean 15.4 (SD 10.7) Hz difference from Praat [50], a popular benchmarking software. This was due to worse background noise rejection and resulted in occasional pitch estimation errors that confused users. In the future, we will thus replace the current pitch calculation algorithm with other algorithms and compare them to Praat to identify algorithms that achieve good pitch estimation, specifically in transfeminine voices—similar to coauthor VSM's previous work that compared Praat to custom algorithms for dysphonic voices [65]. Similar work will be done to identify the most robust algorithms for the calculation of other measures, such as formant frequencies. Based on our team's experience with remote GAVT, we believe that these algorithms will need to take the specifics of GAVT into account: many standard background noise rejection algorithms, such as those used in Zoom videoconferencing software, would not be suitable because they filter out many sounds made during GAVT exercises, such as humming and pitch glides. In addition to this noise reduction work, we will instruct users to use TruVox in quiet spaces and may add simple algorithms to warn users if excessive background noise is detected.

Finally, once multiple modules have been implemented, they will need to be combined into a complete software package that can help deliver GAVT over longer durations. If the software is used as a supplement to SLP-guided training as suggested by other authors [31], this would be relatively easy, as the SLP could periodically reevaluate the user's voice and suggest appropriate exercises and settings. However, if the software is used as an independent GAVT tool for people who cannot or do not want to engage in professional training, as suggested by some of our past interviewees [6], it would need additional features. For example, it would need to present exercises and materials appropriate to the user's skill level without overwhelming users with too many possibilities, and it may need a way to monitor the voice for signs of vocal strain [6,27]. Furthermore, such software would need to avoid influencing transgender voices in undesirable ways—for example, it should not make the voice conform to undesirable gender norms [38]. Nonetheless, in the long term, such efforts may lead to a complete software package that can help increase the accessibility of GAVT and consequently improve the quality of life for transfeminine people.

Study Limitations

As our study is a formative evaluation of prototype software, several methodological choices were made due to convenience. Additionally, because the study represents (to our knowledge) the first quantitative evaluation of GAVT software, several methodological choices had to be made based on protocols from adjacent fields such as limb motor learning. Below, we describe some key study limitations.

Participant Recruitment and Involvement of Developers in Evaluations

Due to convenience sampling, our evaluation was conducted with mainly White participants with Midwestern dialects and a relatively narrow participant age range (19-42 years). Future studies should also include older participants and a more diverse sample to examine, for example, the effects of technological literacy on perceived usability, vocal model preferences (as preliminarily investigated in our prior work [48]), and effects of participant dialect on exercise performance, as this would allow the software to benefit a broader segment of the target population. Previous papers have particularly noted that GAVT software tends to reproduce ideals of gender as White and able-bodied, so we consider a more racially diverse sample critical for the future [38].

Similarly, for convenience in formative evaluations, the head of the development team (VDN) also served as an in-person evaluator, which introduces potential bias because she may have intentionally or unintentionally discouraged negative feedback—for example, due to participants not wanting to disappoint a developer. We originally also made this choice for easier technical support in case of unexpected bugs, as well as

XSL•FO
RenderX

representation of the target population (VDN is a transgender femme). However, having a transgender developer act as an evaluator may have introduced further bias, as participants may be even less willing to disappoint a developer who belongs to the same marginalized population. In the future, we recommend having such evaluations conducted by transgender people who are unaffiliated with the development team, who identify themselves to participants as independent evaluators, and do not show an overly "medicalized" view of GAVT due to documented distrust that some transgender people have toward clinicians [6,38,41].

Evaluation of Exercise Performance

The exercise error metric (absolute difference between actual and target pitch) was selected based on our prior arm motor learning work [58] but may be suboptimal for voice exercises. For instance, if participants in the Human Curve exercise vocalize the same pitch curve faster or slower than the target curve, this would be considered a large error even though it would be highly desirable from an exercise performance perspective. We nonetheless used this simple metric because it is easily understood and still showed significant improvements in exercise performance, but future studies may consider other metrics that are more robust to, for example, differences in timing.

Similarly, to evaluate improvements in exercise performance, we compared the mean exercise error from the first 5 exercise repetitions to the error from the second 5 repetitions. However, exercise error does not consistently decrease due to various factors: for example, some participants exhibited vocal fatigue in later repetitions, while other participants tried out a new exercise strategy in later repetitions that did not work and consequently resulted in higher errors. Again, we used this simple comparison because it nonetheless showed significant improvements, but future studies may consider other metrics.

Fixed Order of Exercises and Fixed Pitch Targets

Performing the exercises in a fixed order from approximately easiest to hardest almost certainly influenced results. For example, while the Stair—bigger and better exercise was described by several participants as harder than the Stair—/i/ exercise, participants exhibited lower exercise errors in the bigger and better exercise (~12 Hz) and did not show significant differences between the first 5 and second 5 repetitions of the bigger and better exercise. This is likely because they had already performed the Stair—/i/ exercise and were able to transfer those skills to the second Stair exercise. Similarly, exercise errors would have likely been even higher in the later exercises if participants had not already performed the earlier, easier exercises. Fatigue likely also influenced results, as participants were more likely to complain about throat soreness and ask for water and breaks later in the session, potentially leading to worse exercise performance. We nonetheless believe that this is not a critical weakness, as actual GAVT would realistically introduce easy exercises before advanced ones.

Related to that issue, using fixed pitch targets (eg, 200 Hz) not adapted to the participant's goals or abilities likely also influenced results. For example, some participants found the upper or lower targets in the Stair exercise difficult to reach. The fixed targets were chosen for simplicity, but future studies may consider personalizing the targets based on an initial voice assessment and the participant's abilities and interests.

Other Considerations

Concerning the usability evaluation (which achieved an SUS score of 79.8), it is worth noting that our participants performed the evaluation almost completely "blind": the experimenter described the concept and provided a list of components to test, but did not provide further guidance. A higher usability score may be achieved if the website is introduced by an SLP during regular GAVT sessions, as recommended by the World Professional Association for Transgender Health [27] and as done with generic voice apps in GAVT studies [31]. However, at the same time, not all transgender people would be likely to use GAVT software with SLP supervision [6], and having the software be usable independently is likely beneficial even for those who would combine it with SLP guidance.

Additionally, while all our participants felt that GAVT had value, not all of them were actively interested in improving their pitch: for example, several had already completed a GAVT regimen with an SLP and did not desire further pitch improvement, and at least 1 stated that she was more interested in improving her resonance than her pitch. Instead, these participants primarily wanted to help evaluate software that may help other transgender people. Therefore, SUS results may change if the study were conducted only with transgender people who are at the stage of GAVT, where they are improving their pitch. As an ad hoc test of this, we divided participants into those who had experienced GAVT with a professional (12 participants) and those who had not (9 participants) and used a *t* test to compare SUS scores between groups. SUS scores were 76.5 (SD 14.2) in those with professional GAVT versus 84.3 (SD 9.7) in those without (t_{19}=−1.42, P=.15); while this is not significant in our small sample, we believe it may be worthy of investigation in the future. Such differences between participants may also affect outcomes other than SUS scores: for example, exercise error in the first 5 repetitions of the Constant—/i/ exercise (first exercise in the performance evaluation) was 4.1 (SD 3.0) Hz in those with professional GAVT versus 6.8 (SD 5.4) Hz in those without (t_{19}=−1.49, P=.15), indicating potential effects of prior GAVT experience on exercise performance.

Future Clinical Trials of GAVT Software

Once our software has been further expanded into a complete GAVT package (see the Next Development Steps section), it should ideally be tested in a clinical trial to evaluate its long-term effectiveness. The primary outcome metric of such a trial would be self-reported or externally perceived changes in voice femininity, which serve as primary subjective outcomes of many GAVT trials [12,13,16,27]. Such changes are commonly evaluated at the end of a full regimen of more than 10 sessions and were thus considered beyond the scope of the current first study.

As TruVox is, to our knowledge, the first systematically evaluated GAVT software, a clinical trial would not be able to compare it to other software specifically for GAVT—Project

XSL•FO
RenderX

Spectra [36] and Attuned [37] have only limited functionality, while EvaF.app and Christella VoiceUp do not appear to be broadly accepted by transgender people [6,38]. In our opinion, a reasonable first clinical trial would use TruVox as daily "homework" between weekly SLP-guided sessions and compare it to either standard homework (eg, written instructions for exercises) or to a generic voice analysis app as carried out in prior GAVT work [31]. In addition to comparing changes in voice femininity, such a clinical trial should also collect qualitative feedback via semistructured interviews, as this would provide insight into, for example, what motivates GAVT clients to practice at home. Furthermore, the trial should be conducted by a researcher who was not involved in TruVox development to reduce experimenter bias. We are currently piloting such a clinical trial of the current pitch module and the aforementioned resonance module [63] over a shorter duration (4 sessions) to identify issues that need to be addressed before a longer trial.

Simultaneously, we do not believe that a clinical trial is strictly necessary before TruVox can be used by transgender people. Many transgender people already use YouTube videos, generic voice analysis tools, and other unvalidated GAVT software (eg, EvaF.app) as part of their GAVT [6,38] despite potential safety concerns, and TruVox could be immediately used by transgender people and SLPs however they desire. We are thus preparing for a public announcement and simultaneous open-source code release. A public release will also allow us to conduct observational studies of, for example, how transgender communities use GAVT software without SLP supervision or how SLPs integrate such software into their GAVT regimens.

Extension to Transmasculine or Nonbinary Voice and Other Languages

Our current work focused on GAVT for transfeminine people for multiple reasons. Nonetheless, we acknowledge that transmasculine people are also interested in GAVT and that the reasons that led us to start with transfeminine voice are also the reasons why transmasculine voice has historically been neglected in GAVT and general research [44,45,66]. Several key components of TruVox (eg, real-time feedback about pitch, resonance, and loudness) are likely applicable to voice masculinization because transmasculine people have also expressed interest in them [6]. At the same time, a version of TruVox for transmasculine voice cannot be created by simply "reversing" the current feminizing exercises, as masculinizing GAVT exercises differ significantly from feminization exercises [45]. Nonetheless, we believe that first steps can be taken by reusing some of the exercises that would likely be applicable (eg, Constant and Stair exercises with lower target pitches), preparing supplementary material targeted at transmasculine users (eg, videos about the reused exercises presented by a transmasculine speaker), and simultaneously developing new exercises based on both academic and community sources about transmasculine voice.

Additionally, while our current work did include some nonbinary participants, TruVox is intended to train a "feminine" voice, and a future transmasculine version would likely train a "masculine" voice. However, interviews of nonbinary people suggest that they may not be interested in a "binary" voice—they may prefer to, for example, target a "neutral" vocal gender, adopt some masculine and some feminine voice features, or have the ability to vary their voice depending on the situation [6]. Such GAVT could, in the long term, be supported with a version of TruVox that includes both feminizing and masculinizing exercises, supporting materials that provide suggestions to nonbinary users, and potentially a goal-setting module (already discussed) that would be responsive to the needs of nonbinary users.

Finally, TruVox currently primarily targets native speakers of American English. While many elements of TruVox would likely apply to other dialects and languages, several changes would also be needed to support those languages and dialects. Obviously, the vocal models would need to be rerecorded, and different phrases would need to be selected. Different phrases may even be needed for different dialects, as participants in our prior interview study anecdotally suggested that phrases used in GAVT may have different intonation in, for example, British English [6] and thus may not be suitable for exercises such as Heteronyms that depend on specific intonation. More broadly, vocal indicators of gender are known to be different in other languages [67], so the optimal target values (eg, the 200 Hz) may be different for non-English speakers. However, to the best of our knowledge, there is little systematic research on the transferability of GAVT techniques between languages, and GAVT in other languages, such as German, does use similar exercises [66], so it is currently unclear how much GAVT software would need to be tailored to other languages.

Conclusions

While it currently only offers pitch exercises, TruVox was perceived as usable and motivating by transfeminine people. Additionally, exercise performance improved over the course of 10 repetitions in all but 1 exercise, indicating that participants were able to improve their skills in these exercises with practice. As qualitative feedback, participants wanted to be able to use the software without much preparation and suggested several desirable future features, such as performance tracking and goal-setting. Based on these positive results, we believe that the current version of TruVox represents a promising GAVT software foundation that will be expanded with modules such as resonance training in the future. To our knowledge, it is the first GAVT software that combines real-time voice visualization with structured exercises, and the in-person evaluation represents the first quantitative human subjects evaluation of GAVT software. In the long term, such software may provide a free and convenient way for transgender people to access GAVT, thus reducing their voice-gender incongruence and consequently improving their mental health. Additionally, it may be broadly beneficial for anyone who wants to change the sound of their voice.

XSL•FO
RenderX

Acknowledgments

Research reported in this publication was supported by the National Institute of Deafness and Other Communication Disorders of the National Institutes of Health under (R21DC021537; approximately 90% of financed costs) and by a Collaborative Pilot Grant from the University of Cincinnati Office of Research (approximately 10% of financed costs). The content is solely the responsibility of the authors and does not necessarily represent the official views of the funders. Additionally, the authors thank Mel Cody for their advice throughout this study and Katelyn Reid, MA, for voiceover in the vocal warmup videos.

Data Availability

Deidentified data from the in-person evaluation, all blank informed consent forms and questionnaires used in the study, and the experimenter checklist for in-person sessions (originally intended for internal use) are publicly available on Zenodo [68].

Authors' Contributions

TM, VSM, and VDN conceptualized this study. MEW, RLG, VSM, and VDN worked on the methodology. MEW, VSM, and VDN investigated this study and assisted with its formal analysis. SRW, OJ, XW, AB, TD, RK, and VDN helped with the software. SRW, MEW, and VDN visualized this paper. VSM and VDN handled the resources. SRW, OJ, VSM, and VDN curated the data. SRW, MEW, VSM, and VDN wrote the original draft. All authors reviewed and edited the writing while VDN supervised. VSM and VDN administered the project. TM, VSM, and VDN acquired the funding for this paper.

Conflicts of Interest

Authors SRW, OJ, XW, AB, TD, RK, and VDN are also developers of TruVox, with VDN also acting as an evaluator in this study.

References

1. Nobili A, Glazebrook C, Arcelus J. Quality of life of treatment-seeking transgender adults: a systematic review and meta-analysis. Rev Endocr Metab Disord 2018 Sep;19(3):199-220 [FREE Full text] [doi: 10.1007/s11154-018-9459-y] [Medline: 30121881]
2. Dhejne C, Van Vlerken R, Heylens G, Arcelus J. Mental health and gender dysphoria: a review of the literature. Int Rev Psychiatry 2016;28(1):44-57. [doi: 10.3109/09540261.2015.1115753] [Medline: 26835611]
3. Hancock AB, Krissinger J, Owen K. Voice perceptions and quality of life of transgender people. J Voice 2011 Sep;25(5):553-558. [doi: 10.1016/j.jvoice.2010.07.013] [Medline: 21051199]
4. Iwamoto SJ, Defreyne J, Rothman MS, Van Schuylenbergh J, Van de Bruaene L, Motmans J, et al. Health considerations for transgender women and remaining unknowns: a narrative review. Ther Adv Endocrinol Metab 2019;10:2042018819871166 [FREE Full text] [doi: 10.1177/2042018819871166] [Medline: 31516689]
5. Bultynck C, Pas C, Defreyne J, Cosyns M, den Heijer M, T'Sjoen G. Self-perception of voice in transgender persons during cross-sex hormone therapy. Laryngoscope 2017 Dec;127(12):2796-2804. [doi: 10.1002/lary.26716] [Medline: 28730692]
6. Bush EJ, Krueger BI, Cody M, Clapp JD, Novak VD. Considerations for voice and communication training software for transgender and nonbinary People. J Voice 2024 Sep;38(5):1251.e1-1251.e20. [doi: 10.1016/j.jvoice.2022.03.002] [Medline: 35382956]
7. Chang J, Brown SK, Hu S, Sivakumar G, Sataluri M, Goldberg L, et al. Effect of Wendler glottoplasty on acoustic measures of voice. Laryngoscope 2021 Mar;131(3):583-586. [doi: 10.1002/lary.28764] [Medline: 32598037]
8. Kelly V, Hertegård S, Eriksson J, Nygren U, Södersten M. Effects of gender-confirming pitch-raising surgery in transgender women: a long-term follow-up study of acoustic and patient-reported data. J Voice 2019 Sep;33(5):781-791. [doi: 10.1016/j.jvoice.2018.03.005] [Medline: 30077418]
9. Bultynck C, Cosyns M, T'Sjoen G, Van Borsel J, Bonte K. Thyroplasty type III to lower the vocal pitch in trans men. Otolaryngol Head Neck Surg 2021 Jan;164(1):157-159. [doi: 10.1177/0194599820937675] [Medline: 32600117]
10. Song TE, Jiang N. Transgender phonosurgery: a systematic review and meta-analysis. Otolaryngol Head Neck Surg 2017 May;156(5):803-808. [doi: 10.1177/0194599817697050] [Medline: 28349733]
11. Mills M, Stoneham G, Georgiadou I. Expanding the evidence: developments and innovations in clinical practice, training and competency within voice and communication therapy for trans and gender diverse people. Int J Transgenderism 2017 Jun 09;18(3):328-342. [doi: 10.1080/15532739.2017.1329049]
12. Chadwick KA, Coleman R, Andreadis K, Pitti M, Rameau A. Outcomes of gender-affirming voice and communication modification for transgender individuals. Laryngoscope 2022 Aug;132(8):1615-1621. [doi: 10.1002/lary.29946] [Medline: 34787313]
13. Hancock A, Helenius L. Adolescent male-to-female transgender voice and communication therapy. J Commun Disord 2012;45(5):313-324. [doi: 10.1016/j.jcomdis.2012.06.008] [Medline: 22796114]
14. Hirsch S. Combining voice, speech science and art approaches to resonant challenges in transgender voice and communication training. Perspect ASHA SIGs 2017 Jan;2(10):74-82. [doi: 10.1044/persp2.sig10.74]

XSL•FO
RenderX

15. Kim H. Vocal feminization for transgender women: current strategies and patient perspectives. Int J Gen Med 2020;13:43-52 [FREE Full text] [doi: 10.2147/IJGM.S205102] [Medline: 32104050]
16. Hancock AB, Garabedian LM. Transgender voice and communication treatment: a retrospective chart review of 25 cases. Int J Lang Commun Disord 2013 Jan;48(1):54-65. [doi: 10.1111/j.1460-6984.2012.00185.x] [Medline: 23317384]
17. McKenna VS, Gustin RL, Hobek AL, Howell RJ, Dickinson TE, Shanley SN, et al. Factors related to treatment attendance for patients seeking gender-affirming voice therapy. Am J Speech Lang Pathol 2023 Jan 11;32(1):216-233. [doi: 10.1044/2022_AJSLP-22-00004] [Medline: 36584326]
18. Matthews J, Olszewski A, Petereit J. Knowledge, training, and attitudes of students and speech-language pathologists about providing communication services to individuals who are transgender. Am J Speech Lang Pathol 2020 May 08;29(2):597-610 [FREE Full text] [doi: 10.1044/2020_AJSLP-19-00148] [Medline: 32320622]
19. Roblee CV, Mendes C, Horen SR, Hamidian Jahromi A. Remote voice treatment with transgender individuals: a health care equity opportunity. J Voice 2022 Jul;36(4):443-444 [FREE Full text] [doi: 10.1016/j.jvoice.2022.01.012] [Medline: 35165003]
20. DeVore EK, Gadkaree SK, Richburg K, Banaszak EM, Wang TV, Naunheim MR, et al. Coverage for gender-affirming voice surgery and therapy for transgender individuals. Laryngoscope 2021 Mar;131(3):E896-E902. [doi: 10.1002/lary.28986] [Medline: 32776558]
21. Dubey I, Stryker SD, Madzia JL, Pickle S, Dion GR, McKenna VS. Insurance coverage for gender-affirming voice and communication services: an in-depth analysis in a midwestern US state. Perspect ASHA SIGs 2024 Jun 03;9(3):490-511. [doi: 10.1044/2024_persp-23-00240]
22. Bartlett RS, Carpenter AM, Chapman LK. A systematic review of adherence strategies for adult populations in speech-language pathology treatment. Am J Speech Lang Pathol 2022 May 10;31(3):1501-1516. [doi: 10.1044/2022_AJSLP-21-00255] [Medline: 35320678]
23. Boulton KA, Hilton M, Sutton E, Guastella AJ. Apps and digital resources for child neurodevelopment, mental health, and well-being: review, evaluation, and reflection on current resources. J Med Internet Res 2025 Jan 01;27:e58693 [FREE Full text] [doi: 10.2196/58693] [Medline: 39742455]
24. Kim S, Song M, Jo Y, Jung Y, You H, Ko M, et al. Effect of voice and articulation parameters of a home-based serious game for speech therapy in children with articulation disorder: prospective single-arm clinical trial. JMIR Serious Games 2023 Oct 11;11:e49216 [FREE Full text] [doi: 10.2196/49216] [Medline: 37819707]
25. Gjestvang C, Kalhovde JM, Mauseth Tangen E, Clemm H, Haakstad LAH. Impact of in-person and mobile exercise coaching on psychosocial factors affecting exercise adherence in inactive women with obesity: 20-week randomized controlled trial. J Med Internet Res 2025 Feb 25;27:e68462 [FREE Full text] [doi: 10.2196/68462] [Medline: 39999434]
26. Duan C, Zhang C, Xu H, Tao J, Yu J, Zhang D, et al. Internet-based cognitive behavioral therapy for preventing postpartum depressive symptoms among pregnant individuals with depression: multicenter randomized controlled trial in China. J Med Internet Res 2025 Mar 04;27:e67386 [FREE Full text] [doi: 10.2196/67386] [Medline: 40053801]
27. Davies S, Papp VG, Antoni C. Voice and communication change for gender nonconforming individuals: giving voice to the person inside. Int J Transgenderism 2015 Nov 16;16(3):117-159. [doi: 10.1080/15532739.2015.1075931]
28. MacDonell KK, Wang B, Phanuphak N, Janamnuaysook R, Srimanus P, Rongkavilit C, et al. Optimizing an mHealth intervention to improve uptake and adherence to HIV pre-exposure prophylaxis in young transgender women: protocol for a multi-phase trial. JMIR Res Protoc 2022 May 19;11(5):e37659 [FREE Full text] [doi: 10.2196/37659] [Medline: 35587370]
29. Bauermeister J, Choi SK, Bruehlman-Senecal E, Golinkoff J, Taboada A, Lavra J, et al. An identity-affirming web application to help sexual and gender minority youth cope with minority stress: pilot randomized controlled trial. J Med Internet Res 2022 Aug 1;24(8):e39094. [doi: 10.2196/39094]
30. Ahmed A. Trans competent interaction design: a qualitative study on voice, identity, and technology. Interact Comput 2018;30(1):53-71. [doi: 10.1093/iwc/iwx018]
31. Hawley JL, Hancock AB. Incorporating mobile app technology in voice modification protocol for transgender women. J Voice 2024 Mar;38(2):337-345. [doi: 10.1016/j.jvoice.2021.09.001] [Medline: 34706847]
32. Smith C, Gregory C, Bryant L. Utilizing virtual reality for gender-affirming voice training: surveying the attitudes and perspectives of potential consumers. Int J Lang Commun Disord 2024;59(3):976-990. [doi: 10.1111/1460-6984.12968] [Medline: 37929613]
33. Millman N, Timmons Sund L, Johns M, Elliott A, Silverstein E, Bensoussan Y. Gauging stakeholder perspectives on artificial intelligence in gender-affirming voice care. Laryngoscope 2023 Jul;133(7):1698-1705. [doi: 10.1002/lary.30406] [Medline: 36177824]
34. Philippone M. User-Centered Design for Gamified Voice and Communication App: Applying UCD to Gender-Affirming Communication Services. Washington, D.C: George Washington University; 2022.
35. Povinelli K, Zhu HH, Zhao Y. Beyond the “industry standard”: focusing gender-affirming voice training technologies on individualized goal exploration. arXiv Preprint posted online on Feb 21, 2025 [FREE Full text] [doi: 10.1145/3706598.3713430]
36. Ahmed AA, Kok B, Howard C, Still K. Online community-based design of free and open source software for transgender voice training. Proc ACM Hum-Comput Interact 2021 Jan 05;4(CSCW3):1-27. [doi: 10.1145/3434167]

XSL•FO
RenderX

37. Alter IL, Chadwick KA, Andreadis K, Coleman R, Pitti M, Ezell JM, et al. Developing a mobile application for gender-affirming voice training: a community-engaged approach. Laryngoscope Investig Otolaryngol 2024 Dec;9(6):e70043. [doi: 10.1002/lio2.70043] [Medline: 39679168]
38. Ahmed AA, Kim L, Hoffmann AL. ‘This app can help you change your voice’: authenticity and authority in mobile applications for transgender voice training. Converg Int J Res into New Media Technol 2022 Apr 11;28(5):1283-1302. [doi: 10.1177/13548565221079459]
39. Netzorg R, Cote A, Koshin S, Garoute K, Anumanchipalli G. Speech after gender: a trans-feminine perspective on next steps for speech science and technology. 2024 Presented at: INTERSPEECH 2024; September 1-5, 2024; Kos Island, Greece. [doi: 10.21437/interspeech.2024-2269]
40. Povinelli K, Zhao Y. Springboard, roadblock or "crutch"?: How transgender users leverage voice changers for gender presentation in social virtual reality. 2024 Presented at: Proc 2024 IEEE Conf Virtual Real 3D User Interfaces 2024; 16-21 March, 2024; Orlando, FL. [doi: 10.1109/vr58804.2024.00057]
41. Huff AZ. Modern responses to traditional pitfalls in gender affirming behavioral voice modification. Otolaryngol Clin North Am 2022 Aug;55(4):727-738. [doi: 10.1016/j.otc.2022.05.001] [Medline: 35752493]
42. Holmberg EB, Oates J, Dacakis G, Grant C. Phonetograms, aerodynamic measurements, self-evaluations, and auditory perceptual ratings of male-to-female transsexual voice. J Voice 2010 Sep;24(5):511-522. [doi: 10.1016/j.jvoice.2009.02.002] [Medline: 19766444]
43. Leung Y, Oates J, Chan SP. Voice, articulation, and prosody contribute to listener perceptions of speaker gender: a systematic review and meta-analysis. J Speech Lang Hear Res 2018 Feb 15;61(2):266-297. [doi: 10.1044/2017_JSLHR-S-17-0067] [Medline: 29392290]
44. Boddicker O, Kominsky R. Behavioral management for masculinization of voice and communication across the gender spectrum. In: Courey MS, Rapoport SK, Goldberg L, Brown SK, editors. Voice and Communication in Transgender and Gender Diverse Individuals. Cham: Springer; Jul 19, 2023:135-154.
45. Block C. Making a case for transmasculine voice and communication training. Perspect ASHA SIGs 2017 Jan;2(3):33-41. [doi: 10.1044/persp2.sig3.33]
46. TruVox. University of Cincinnati. 2025 Jan 31. URL: https://ceas5.uc.edu/transvoice/ [accessed 2025-08-12]
47. Wang X, Weese S, McAllister T, McKenna V, Novak V. Design and usability evaluation of a web-based pitch control training app for transgender women. HCI Int 2024 Posters (2024) 2024;2114:224-234. [doi: 10.1007/978-3-031-61932-8_27] [Medline: 39439900]
48. Wilkens M, Novak V, McKenna V. Perspectives on using vocal models to guide pitch training in transfeminine patients Poster presentation. 2024 Presented at: Fall Voice Conference; October 24-26, 2024; Phoenix, AZ.
49. Huckvale M. Web tools for speech. 2025. URL: https://www.speechandhearing.net/laboratory/webtools.php [accessed 2025-08-12]
50. Boersma P, Weenink D. PRAAT, a system for doing phonetics by computer. Glot Int 2001;5:341-345 [FREE Full text]
51. Stemple JC, Roy N, Klaben BK. Clinical Voice Pathology: Theory and Management. San Diego, CA: Plural Publishing Inc; 2020.
52. Macefield R. How to specify the participant group size for usability studies: a practitioner's guide. JUX 2009;5(1):34-45 [FREE Full text]
53. Brooke J. SUS: a "quick and dirty" usability scale. In: Jordan PW, Thomas B, Weerdmeester BA, McClelland AL, editors. Usability Evaluation in Industry. London, United Kingdom: Taylor and Francis; 1996.
54. Lewis J, Sauro J. Item benchmarks for the System Usability Scale. J Usability Stud 2018;13(3):158-167 [FREE Full text]
55. Darzi A, Novak D. Automated affect classification and task difficulty adaptation in a competitive scenario based on physiological linkage: an exploratory study. Int J Hum Comput Stud 2021 Sep;153:102673 [FREE Full text] [doi: 10.1016/j.ijhcs.2021.102673] [Medline: 34092990]
56. Novak VD, Hass D, Hossain MS, Sowers AF, Clapp JD. Effects of adaptation accuracy and magnitude in affect-aware difficulty adaptation for the Multi-Attribute Task Battery. Int J Hum Comput Stud 2024 Mar;183:103180. [doi: 10.1016/j.ijhcs.2023.103180]
57. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B 1995;57(1):289-300. [doi: 10.1016/s0166-4328(01)00297-2]
58. Batson J, Kato Y, Shuster K, Patton J, Reed K, Tsuji T, et al. Haptic coupling in dyads improves motor learning in a simple force field. 2020 Presented at: 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society; July 20-24, 2020; Online. [doi: 10.1109/embc44109.2020.9176261]
59. Heller Murray ES, Stepp CE. Relationships between vocal pitch perception and production: a developmental perspective. Sci Rep 2020 Mar 03;10(1):3912. [doi: 10.1038/s41598-020-60756-2]
60. Smith D, Patterson R. The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age. J Acoust Soc Am 2005;118(5):3177-3186. [doi: 10.1121/1.2047107]
61. Byun TM, Campbell H, Carey H, Liang W, Park TH, Svirsky M. Enhancing intervention for residual rhotic errors via app-delivered biofeedback: a case study. J Speech Lang Hear Res 2017 Jun 22;60(6S):1810-1817 [FREE Full text] [doi: 10.1044/2017_JSLHR-S-16-0248] [Medline: 28655050]

XSL•FO
RenderX

62. Kawitzky D, McAllister T. The effect of formant biofeedback on the feminization of voice in transgender women. J Voice 2020 Jan;34(1):53-67. [doi: 10.1016/j.jvoice.2018.07.017] [Medline: 30174221]
63. McAllister T, Eagen C, Shan Y, Traver P, Harel D, Park T, et al. Web-based application for real-time biofeedback of vocal resonance in gender-affirming voice training: design and usability evaluation. 2025 Aug Presented at: INTERSPEECH 2025; August 17-21, 2025; Rotterdam, Netherlands.
64. Sigrist R, Rauter G, Riener R, Wolf P. Augmented visual, auditory, haptic, and multimodal feedback in motor learning: a review. Psychon Bull Rev 2013 Feb;20(1):21-53. [doi: 10.3758/s13423-012-0333-8] [Medline: 23132605]
65. Llico AF, Shanley SN, Friedman AD, Bamford LM, Roberts RM, McKenna VS. Comparison between custom smartphone acoustic processing algorithms and Praat in healthy and disordered voices. J Voice (forthcoming) 2023 Sep (forthcoming) [FREE Full text] [doi: 10.1016/j.jvoice.2023.07.032]
66. Azul D, Neuschaefer-Rube C. Voice function in gender-diverse people assigned female at birth: results from a participant-centered mixed-methods study and implications for clinical practice. J Speech Lang Hear Res 2019 Sep 20;62(9):3320-3338. [doi: 10.1044/2019_JSLHR-S-19-0063] [Medline: 31518520]
67. Pépiot E. Male and female speech: A study of mean f0, f0 range, phonation type and speech rate in parisian French and American English speakers. 2014 Presented at: Proc Int Conf Speech Prosody; 20-23 May, 2014; Dublin, Ireland p. 305-309. [doi: 10.21437/speechprosody.2014-49]
68. TruVox web-based software for vocal pitch training in transgender women: Development and single-session evaluations. Zenodo. 2025. URL: https://zenodo.org/records/16390648 [accessed 2025-08-12]

Abbreviations

GAVT: gender-affirming voice training
IMI: Intrinsic Motivation Inventory
LGBTQ+: lesbian, gay, bisexual, transgender, queer, and others
SLP: speech-language pathologist
SUS: System Usability Scale

Edited by J Sarvestan; submitted 12.03.25; peer-reviewed by Q Zhang; comments to author 30.05.25; revised version received 19.06.25; accepted 31.07.25; published 15.09.25.

Please cite as:
Weese SR, Wilkens ME, Jadhav O, Wang X, Bhanushali A, DiLoreto T, Kozel R, Gustin RL, McAllister T, McKenna VS, Novak VD
TruVox Web-Based Software for Vocal Pitch Training in Transgender Women: Development and Single-Session Evaluations
JMIR Form Res 2025;9:e73841
URL: https://formative.jmir.org/2025/1/e73841
doi:10.2196/73841
PMID:

©Sam R Weese, Mary E Wilkens, Om Jadhav, Xiangyi Wang, Ansh Bhanushali, Tyler DiLoreto, Reyna Kozel, Renee L Gustin, Tara McAllister, Victoria Sue McKenna, Vesna Dominika Novak. Originally published in JMIR Formative Research (https://formative.jmir.org), 15.09.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Online Yoga Pilot Intervention for Black Women at High Cardiovascular Risk: Internet-Based Recruitment and Engagement

Candace Crosby Johnson[1*], RN, MPH, PhD; Pascaline Ezouah[2*], MBA, MPH, LSSGB

[1]Nursing, College of Health Sciences, Lawrence Technological University, Southfield, MI, United States

[2]Ezouah Consulting, LLC, Atlanta, GA, United States

[*]all authors contributed equally

Corresponding Author:
Candace Crosby Johnson, RN, MPH, PhD
Nursing
College of Health Sciences
Lawrence Technological University
21311 Civic Center, B209
Southfield, MI, 48075
United States
Phone: 1 248 204 3581
Email: cjohnso11@ltu.edu

Abstract

Background: Disproportionately adverse heart health outcomes in Black women, characterized by high metabolic syndrome prevalence, underscore the need for innovative, accessible interventions. Digital health strategies, particularly web-based yoga videos, show promise for engaging this high-risk group in health-promoting behaviors.

Objective: This study aimed to evaluate the feasibility and acceptability of a web-based yoga intervention for community-dwelling Black women, providing preliminary data to inform a larger, mixed methods study on reducing cardiometabolic risks.

Methods: In this 4-week pilot study, grounded in Pender's Health Promotion Model, 28 participants engaged in daily online health education and yoga activities through YouTube videos. Using Fitbit trackers, electronic blood pressure monitors, and web-based logs, the study measured metabolic syndrome risk factors and sedentary behavior. Participant experiences were further explored through postintervention focus groups aiming to contextualize the intervention's impact.

Results: We enrolled 28 women, with a completion rate of 79% (22/28), demonstrating successful recruitment and retention. Participants were an average age of 43.3 years with a mean BMI of 40.9 kg/m^2, indicating a high-risk group for metabolic syndrome. Engagement with 2 or more intervention components were significantly correlated with study completion (χ^2_1=7.14, *P*=.008). Specifically, viewing over one-half of the instructional videos (χ^2_1=4.39, *P*=.04) and daily blood pressure monitoring (χ^2_1=5.67, *P*=.02) were key to participant adherence. The intervention was well-received, with 95% (19/20) of survey respondents finding it satisfactory and suitable. Technology use was high, with all participants having access to the internet, 96% (27/28) owning smartphones, and 53% (15/28) having a YouTube account prior to the study. Recruitment was effectively conducted online, primarily via Facebook and a university newsletter, each accounting for 39.3% (11/28) of participants. The qualitative focus group data unveiled 4 major themes: (1) accountability, emphasizing the shift toward self-prioritization and collective health responsibility; (2) increased awareness, highlighting enhanced understanding of health behaviors and metabolic syndrome risks; (3) health benefits, noting observed improvements in blood pressure and stress levels; and (4) unanticipated stressors, identifying external factors that challenged engagement. These insights underscore the intervention's multifaceted impact, from fostering health awareness to navigating external stressors.

Conclusions: This pilot study demonstrated the feasibility and acceptability of a culturally tailored, online yoga intervention among community-based, Black women at high risk for metabolic syndrome, showing promising engagement and potential health benefits. The high rates of participation and completion highlight the intervention's acceptability and the potential for digital platforms to facilitate health behavior changes in high-risk populations. The qualitative findings reveal critical insights into the psychological and social dynamics influencing health behavior change, suggesting the importance of addressing both individual

XSL•FO
RenderX

and communal barriers to improve intervention efficacy. Future research should further explore these dynamics in larger, more diverse cohorts to substantiate the intervention's potential in reducing cardiometabolic risks.

(JMIR Form Res 2025;9:e41221) doi:10.2196/41221

KEYWORDS

web-based interventions; digital health; mHealth; self-management; eHealth; health promotion, yoga; mind-body therapies, complementary therapy; African Americans; Blacks; women's health; metabolic syndrome

Introduction

Metabolic Syndrome

Metabolic syndrome is a predisease classification for individuals who may have a combination of factors that predispose them to diabetes, hypertension, and cardiovascular disease, including sedentary lifestyle, high-stress work and family dynamics, and dietary patterns [1]. Metabolic syndrome risk factors include elevations in the following biophysical markers: blood pressure, fasting blood glucose, cholesterol, triglycerides, and waist circumference. Metabolic syndrome is associated with a 5-fold increased risk for type 2 diabetes, a chronic condition linked to cardiovascular disease, the leading cause of death in US adults [2]. Among US racial/ethnic female subgroups, prediabetes and hypertension are highest in Black women, 40% of whom experience metabolic syndrome. Overall, cardiovascular disease morbidity and mortality are disproportionately highest in Black women when compared with other US subgroups [3].

To manage the risk presented by metabolic syndrome, it is recommended that optimal BMI, blood pressure, and blood glucose levels be maintained through lifestyle approaches such as increasing physical activity and exercise [4,5]. Black women experience comparably high obesity and physical inactivity rates when compared with other US subgroups [6]. Notwithstanding, Black women have difficulties implementing physical activity regimens due to lack of awareness and training, which influence their behavior management strategies [7].

Multiple Risk Factor Interventions

Nonpharmacological behavioral modification interventions reduce the burden of chronic diseases associated with metabolic syndrome. Research demonstrates that multiple risk factor interventions—programs that impact more than one symptom of metabolic syndrome—can lower blood pressure levels, BMI, and waist circumference in high-risk populations [8]. Behavioral interventions that address lifestyle modifications are recommended in Black patients with slightly elevated blood pressures as they can prevent or delay worsening disease [9]. When developing behavioral interventions, content should include caveats for stress management to help individuals stay engaged with lifestyle modifications [10]. Additionally, incorporating behavior change techniques such as self-monitoring and eliciting social support for exercise further aid with maintaining lifestyle behavioral changes [11]. Behavioral interventions incorporating culturally adapted, participant-involved strategies have been effective over usual care in weight management studies involving Black women [12].

Mind-body therapies that use spinal manipulation such as yoga have been shown to be efficacious in lowering blood pressure and improving lipid profiles, yet to date, there are a shortage of such interventions that reduce metabolic syndrome risk factors in high-risk Black populations [13]. Research suggests mind-body therapies such as mindfulness and yoga offer promise as a therapeutic treatment for cardiovascular disease risk [14]. Mindfulness meditation, a mental component of yoga practice, may improve health outcomes in Black communities. Mind-body therapies are feasible and acceptable interventions that result in positive health outcomes among Black individuals, particularly when cultural adaptations are applied [15]. Remotely delivered yoga interventions including those with postures, breathing techniques, and meditation exercises were shown to be generally safe and feasible for implementation in diverse populations with and without chronic conditions [16].

Novel interactive eHealth—digital health—educational approaches are being explored for patient empowerment and self-management of chronic diseases as they provide comprehensive, accessible content [17]. Remote, web-based interventions have demonstrated evidence for a positive, moderate effect on physical activity [18], showing promise for improving self-management of obesity, a major contributor to metabolic syndrome. Research using mobile devices suggests that this method is effective for influencing physical activity behaviors [19]. Mobile devices can be used to send information to users, gather information from users, and foster peer-to-peer connections [20]. Black women are more likely to be smartphone-dependent and consume more mobile- and internet-based health information than other US ethnic/racial counterparts [21]. When it comes to tracking their health, Black women are no more likely than other racial/ethnic subgroups to use their mobile devices to track their weight, diet, or exercise; however, the likelihood of engaging with eHealth strategies increases when they have one or more chronic conditions or have gone through a significant health change in the past year [21].

Challenges in Addressing Metabolic Syndrome in Black Women

Despite the efficacy of mind-body therapies like yoga in managing cardiovascular risk factors, there is a noted shortage of interventions specifically designed to reduce metabolic syndrome risk factors among high-risk Black populations. This gap highlights the need for culturally tailored interventions that are both acceptable and feasible for Black women, whose engagement with digital health strategies is often contingent upon existing chronic conditions or significant health changes.

XSL•FO
RenderX

This study aimed to assess the feasibility, engagement, and preliminary impacts of a culturally tailored, online yoga intervention, YogicDance, on metabolic syndrome risk factors and related psychological health outcomes in Black women aged 35 years to 50 years at high cardiovascular risk, addressing the unique needs and barriers faced by this population while evaluating its effectiveness at improving both metabolic syndrome risk factors and psychological health outcomes. Specifically, it evaluates participant enrollment and participation rates, their interaction with the intervention's technological components, and their perceptions of their satisfaction with the intervention as well as its suitability, ease of use, and helpfulness.

Methods

Study Design

The Health Promotion Model (HPM), outlined in Figure 1, is the framework that guided the foundational constructs to be explored in this pilot study. The HPM combines concepts from social cognitive theory [22] and expectancy-value theory [23] to facilitate researchers' understanding of the "background factors" [24] that influence health-promoting behaviors. The HPM has been used to explore the impact of health promotion strategies on physical activity [25], in older women [24], in Black women [26] and in Black women at risk for chronic disease [27]. The HPM elicits input from the participants regarding their individual characteristics and experiences as well as behavior-specific cognitions and affect including benefits and barriers related to intervention engagement. This culturally tailored physical activity and stress management intervention sought to address the complexities of the Black female experience with regards to accessing health information and engaging in health behaviors that ultimately reduce poor clinical health outcomes. The study was designed with selected ecological factors such as low access to physical activity resources, unsafe neighborhood environments, and cultural sensitivity in mind.

Figure 1. Health Promotion Model adapted for the YogicDance (YD) intervention study. PA: physical activity; BP: blood pressure; SBW: strong Black woman.

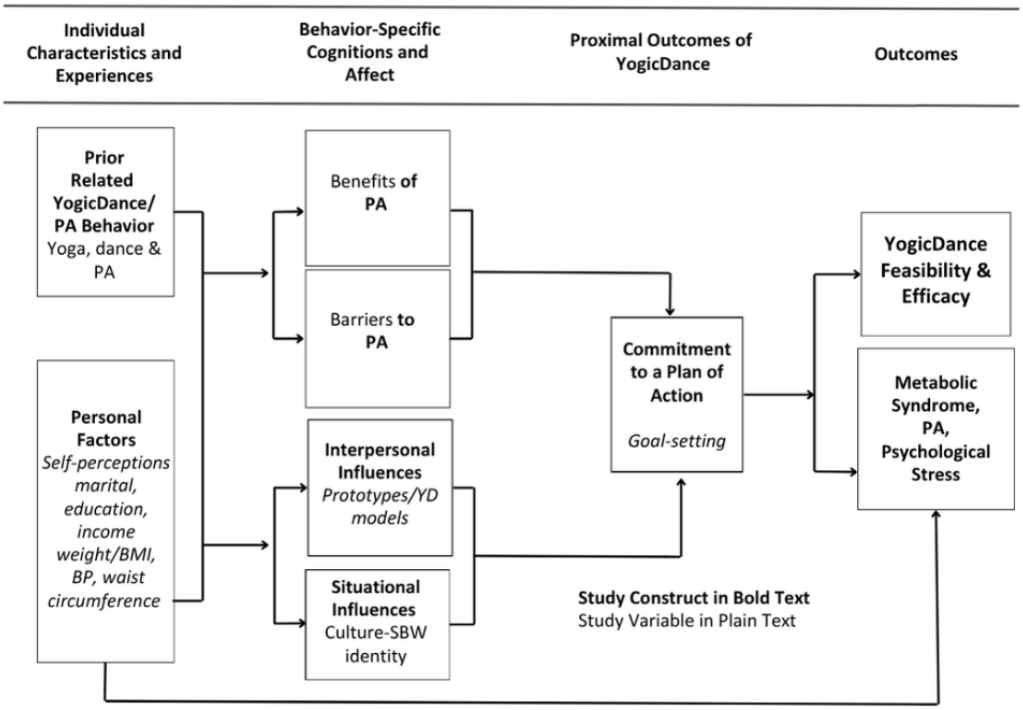

A mixed methods approach was selected to capture the participants' holistic perspectives on the acceptability of the YogicDance intervention. This study used an explanatory sequential mixed methods design [28]. Initially, quantitative data were collected and analyzed to measure the metabolic syndrome risk factors and physical activity levels of the participants using tools such as Fitbit trackers and electronic blood pressure monitors. Following this, qualitative data were gathered through focus groups to explore participants' experiences, barriers, and benefits related to the intervention. The qualitative findings were used to provide deeper insights and context to the quantitative results, allowing for a comprehensive understanding of the intervention's impact. This approach enabled the researchers to integrate quantitative outcomes with qualitative insights, thereby enhancing the overall interpretation of the study findings. A mixed methods approach was implemented because the complex determinants of physical activity in Black women warrant corroboration to give researchers deeper, contextual meaning to the broad, descriptive correlates assessed in this study. Focus group data provided contextual understanding of the relationships among study variables found through quantitative data measures [29]. A joint visual display that integrates the collection and analysis of both the quantitative and qualitative data was used to help enhance the clarity of the data integration in the final interpretation (see Figure 2).

XSL•FO
RenderX

Figure 2. YogicDance explanatory sequential study design. REDCap: Research Electronic Data Capture.

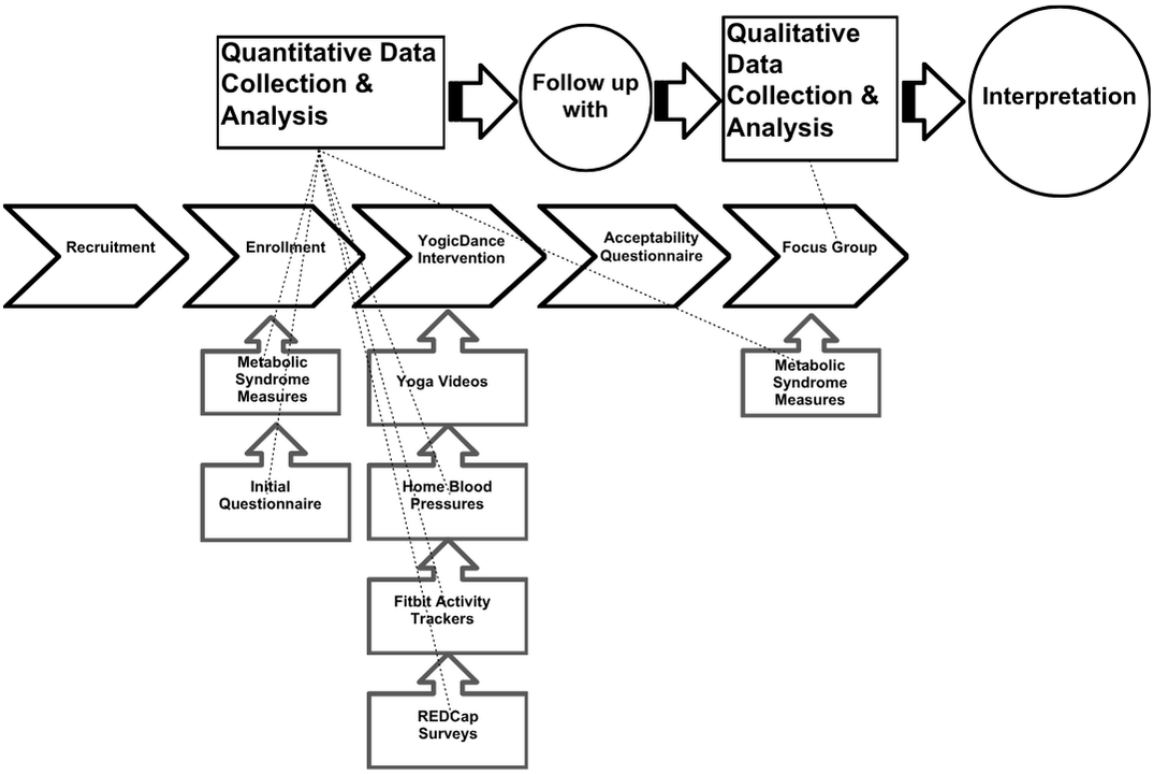

Eligibility Criteria

Eligibility screenings took place initially by telephone or in-person meetings in settings acceptable to and agreed upon by the interested individual and the research team. The inclusion criteria for the YogicDance study were as follows: Black woman (self-identified volunteers), aged 35 years through 50 years (valid ID); overweight (BMI >25 kg/m^2), abdominally obese (waist circumference >88 cm), available at-home internet and desktop or laptop screen access (research team observation), email access (confirmation email), able to navigate the YogicDance website (per YouTube video link access), and able to increase physical activity (per a "no" response to all questions on the Physical Activity Readiness Questionnaire+ [PAR-Q+]) [30]. Additionally, a health care provider's clearance was required for interested women to proceed further with the study if they were extremely obese (BMI >40 kg/m^2); had blood pressure measures above 140 mm Hg (systolic blood pressure) or 90 mm Hg (diastolic blood pressure); answered "yes" to one or more of the 7 PAR-Q+ questions; or reported taking medications or being under a health care provider's care for a diagnosis of diabetes, hypertension, or any other health condition.

The age range of 35 years through 50 years was chosen because this is an age group at high risk for the indicators seen in metabolic syndrome but who may not yet have exhibited cardiovascular disease, thus allowing an opportunity for prevention of disease. Individuals in this age range also tend to have busy lives with competing family and work responsibilities that contribute to identification with stress and sedentary lifestyles. YogicDance integrated the aforementioned strategies as YouTube-based eHealth education videos, cultural tailoring using Black models or educators, self-monitoring via commercially available activity trackers, and stress management using yogic breathing and movement techniques. Interested women were excluded from participating in the study if they were currently undergoing physical therapy, had symptoms of an untreated musculoskeletal injury, were housebound because of physical disability, walked with assistance or assistive devices, or were unable to secure a health care provider's clearance to participate in the YogicDance study. Interested individuals who met all the inclusion criteria (as observed or assessed by the study team) were consented into the study after having the consent form explained to them and exhaustively responding to questions about study participation. Screened women who met all of the initial inclusion criteria were scheduled for an in-home visit. During the (in-home) baseline data collection, the study team also confirmed that the potential participant had an internet connection and access to a working smartphone. In addition, the team observed that the individual had sufficient space to carry out the YogicDance movements with a clear view of a computer screen.

Recruitment

Participants were recruited using a variety of methods including printed institutional review board (IRB)–approved fliers, university campus TV monitors, and online newsletter advertisements containing research team contact information and depicting relevant images of Black women of varied body sizes. Study personnel made appearances at community health fairs and at hair salons to recruit Black women. Given the nature of this internet-based intervention, digital versions of the IRB fliers were also distributed on internet-based marketplaces such as Facebook and YouTube (see Multimedia Appendix 1). Because Blacks underparticipate in clinical trials, recruitment

XSL•FO
RenderX

and attrition are significant problems for researchers. Although using a small sample size carries the risk of making a type II error, a sample of 30 participants was deemed appropriate to detect intervention effects. As approximately 40% of Black adult women have metabolic syndrome, we screened 51 individuals to reach our target enrollment of 30 at-risk participants.

Intervention Details

Study recruitment and enrollment occurred on a rolling basis until the target enrollment was reached (see Figure 3 and Multimedia Appendix 2). The focus groups were held in 2 trusted, community-based locations on and around the university campus. Focus groups were used to assess the feasibility of the intervention, explore barriers and benefits to participating in YogicDance, and explore the contextual meaning and experiences with yoga in the lives of the participants during the study. Each participant was given two US $25 gift cards (at pre- and postintervention) as remuneration for their time and study involvement.

Enrolled study participants were assigned a Fitbit (Fitbit Inc) activity tracker and Omron (Omron Corporation) blood pressure monitor and were assisted by the research team in calibrating these monitoring devices to be used in the study. Immediately following calibration, the team assisted the participant with installing the device software and downloading the companion apps from Fitbit, Omron, and YouTube. To protect the participants' confidentiality, each participant was also given a card with passwords to the aforementioned website accounts, and assigned identification numbers were used to protect the participants' privacy. Baseline physical activity status (sedentary vs physically active) was determined using the International Physical Activity Questionnaire (IPAQ). To determine baseline metabolic syndrome risk, the research team collected metabolic syndrome–related anthropometric (weight and waist circumference), physiologic (systolic and diastolic blood pressures), and fasting blood (glucose and lipid profile) measures. Participants were also instructed to watch a series of introductory, safety-related, health education and goal-setting videos prior to beginning YogicDance physical activities. Additionally, participants were asked not to alter their diets or other health-related activities to reduce study intervention confounders.

During the 4 weeks of the intervention, participants watched YogicDance physical training videos and performed 10 minutes to 30 minutes of physical activity with their choice of 3 video training modules available on the YouTube website. YogicDance instructors included 2 women. The first was a registered nurse and yoga instructor (White American, 64 years old) with 500 hours of yoga training and additional certifications in breath-centered yoga as well as yoga for cardiovascular disorders, osteoporosis, and arthritis. Her 40 years of yoga instruction included experience in teaching yoga to community-based Black women with and at risk for cardiometabolic disease. She developed and instructed the participants in gentle, restorative yogic breathing and postures (asanas) in chair and mat yoga formats. The second was a certified physical education instructor and physical trainer (Black, 42 years old) with 20 years of training experience with Black female clientele who provided education and instructional training in warm-up and cool-down stretching using light resistance bands and moderate aerobic activity. All programs began with a warmup and concluded with a cool-down session. YogicDance instructors provided instruction on how to safely execute yoga and resistance training activities, with specific guidance for effective breathing, timing, and body positioning. Immediately following YogicDance practice, participants were asked to measure their blood pressures using the Omron electronic monitor as well as heart rate. Participants were also provided a paper booklet for recording their blood pressures in writing, to help them with recall in the event they are not able to complete online-based surveys administered using the Research Electronic Data Capture (REDCap) platform immediately following YogicDance practice.

Weekly during the intervention, study staff contacted participants by telephone to communicate any questions; concerns; problems; or reports of pain, injury, or adverse events. The research team helped the participant assess their pain or injury, and any participant reporting physical distress or acute symptoms such as intense pain, numbness, tingling, burning, swelling, redness, and tenderness was encouraged to make an appointment with their health care provider. Participants also had the opportunity to contact the research team after business hours and on weekends during the study to report any adverse events, pain, or injury.

Quantitative Phase

Data Collection

Baseline Measures

Baseline physical activity status (sedentary vs physically active) was determined using the IPAQ [31]. To determine baseline metabolic syndrome risk, the research team collected metabolic syndrome–related anthropometric (weight and waist circumference), physiologic (systolic and diastolic blood pressures), and fasting blood (glucose and lipid profile) measures. Baseline data collection involved assessing various parameters to understand the participants' health status. The data collection included several measures and tools, which are described in the following sections.

Demographic Characteristics Survey

Age, education level, household income, marital status, and employment status were collected using a standardized questionnaire developed by the principal investigator (PI). The frequencies and durations of walking, moderate activity, and vigorous activity over the past week were assessed using the IPAQ.

Participant Prior Technology Use Survey

A standardized, PI-developed questionnaire assessed technology use prior to the intervention as well as internet-based recruitment methods (eg, Facebook, email fliers, University listserv fliers) used to recruit the study participants.

XSL•FO
RenderX

YogicDance Study Engagement

Engagement metrics included the total number of study activities completed, videos watched, Fitbit worn, web-based surveys taken, and blood pressures recorded.

YogicDance Study Acceptability Survey

This PI-developed survey included participant feedback on the screening process and overall acceptability of the intervention.

Data Analysis

Descriptive statistics were used to calculate recruitment, refusal, attrition, and retention rates as well as reasons for refusal. Descriptive statistics were used to describe the sample and YogicDance intervention acceptability, including calculating means, standard deviations, and ranges for the continuous variables and counts with frequencies for the categorical variables. Chi-square analyses were used to explore significant relationships among categorical study engagement variables. These study activities included watching YouTube videos, wearing Fitbit activity trackers, taking home-based blood pressures, and completing web-based activity logs. The levels of significance were set at an α of .05 for all statistical analyses.

Quantitative data were analyzed using SPSS (Version 25; IBM Corp). Although this was a pilot study, examining potential differences to determine if there are trends toward improvements in metabolic syndrome risk factors and physical activity levels was important for determining effect sizes. In addition, means and standard deviations from this study will be used to conduct a more precise power analysis for a future larger randomized controlled trial. Pre-post intervention differences in metabolic syndrome risk factors and physical activity outcomes are reported elsewhere.

Qualitative Phase

Data Collection

The focus group participants were a subset of those who participated in the intervention (see Figure 3). Specifically, 17 of the 22 participants who completed the intervention also participated in the postintervention focus groups. These focus groups aimed to assess the feasibility of the intervention and explore barriers and benefits to participation. Two prescheduled focus groups were held in a cultural center and in a classroom located on the university campus. Bus and parking accommodations were provided, and childcare was offered to participants. We provided emailed weblinks to participants wishing to join the focus groups by Skype or GoogleMeet. An interview guide was used to ensure clarity and to avoid redundancy in the open-ended questions. Although an interview guide was used to elicit behavior-specific cognitions and affect of the participants, the open-ended questions and reflective interviewing skills of the researcher encouraged communication. The researcher was able to further prompt the participants regarding the benefits and barriers to engagement with the YogicDance intervention.

Qualitative focus group results provide a composite picture of the YogicDance intervention experience. Analyses of the interview data revealed that they were commensurate with the behavior-specific cognitions and affect construct of the HPM (Figure 1) as we sought to identify the benefits and barriers to YogicDance study participation. Hierarchical language frequency patterns associated with study engagement and self-monitoring were identified during qualitative analyses of transcript data.

Data Analysis

Cognitive and affective dimensions of participant engagement were measured by exploring semistructured focus group transcripts and open-ended acceptability survey responses to identify themes associated with study engagement and completion. Focus group interviews were transcribed and analyzed using NVivo 11 Pro for Windows (QSR International) to identify word frequency and other familial patterns related to benefits and barriers to study engagement. The HPM (Figure 1), by Pender, informed the initial identification of behavior-specific cognitions and affect related to intervention feasibility, namely the benefits and barriers to study participation. Source materials were coded for themes using NVivo queries for searching data and organized into hierarchical nodes that were then visualized into data maps interpreted by researchers. The researchers iteratively discussed and compared the coding of the data with emerging themes. The researchers also used member-checking, in which several participants reviewed the results and validated that our identified themes were representative of the experience of participating in the study.

We used deductive content analysis as outlined by Elo and Kyngäs [32] to analyze the focus group discussions. This choice was informed by our mixed methods design, aiming to integrate qualitative insights with quantitative findings for a comprehensive understanding of physical activity determinants in Black women. For the discussions, 17 participants contributed to 2 focus groups, which were audio-recorded and lasted approximately 60 minutes each. Transcriptions and field notes taken during these sessions underwent a rigorous coding process to identify recurrent themes, which were then categorized and analyzed for frequency. The analysis was conducted independently by 2 researchers, including the principal investigator and a qualitative expert in physical activity among Black women, ensuring trustworthiness and reliability through their iterative discussion and consensus. This collaborative analysis process allowed for the triangulation of themes, enriching our quantitative data with contextual depth and supporting the convergence of findings. Trustworthiness was further enhanced by adhering to established qualitative research guidelines, including credibility, transferability, dependability, and confirmability, through methodical data handling, analysis, and interpretation.

Ethical Considerations

IRB approval was obtained for study forms and procedures from the Virginia Commonwealth University IRB (number HM20004679). No other approvals were needed for study completion. All participants provided written informed consent. Our study used an open, web-based, e-survey design that received IRB approval, including a thorough informed consent process detailing the survey's duration, data handling practices, and purpose. To ensure the protection of participant data against

XSL•FO
RenderX

unauthorized access, robust data protection measures were implemented. The survey development phase included extensive pretesting to assess both usability and technical functionality, optimizing the participant experience. We actively recruited participants through a multifaceted advertising campaign, leveraging social media, email newsletters, and targeted banner ads, aiming to attract a wide-ranging demographic to our convenience sample.

The survey administration was structured to enhance participant engagement and data integrity, featuring adaptive questioning and item randomization to mitigate response biases. Voluntary participation was encouraged with incentives, and multiple safeguards, including the use of cookies and IP address monitoring, were in place to prevent duplicate submissions. Analytically, we addressed the potential nonrepresentativeness of our sample through statistical corrections, analyzing both complete and partially completed questionnaires to extract comprehensive insights. By following the CHERRIES (Checklist for Reporting Results of Internet E-Surveys) checklist, our e-survey methodology was transparent, secure, and methodologically sound, providing valuable data on YogicDance intervention engagement while ensuring participant privacy and data integrity.

For the focus groups, audio-recordings were made to capture the discussions accurately. These recordings were transcribed verbatim by the research team. No video recordings were made to ensure participant comfort and confidentiality. The audio files and transcripts were stored securely on password-protected computers accessible only to the research team. Physical copies of the transcripts were stored in locked cabinets within the research facility. Access to the audio recordings and transcripts was strictly limited to the principal investigators and authorized research team members involved in the data analysis.

Results

Participants Description

We enrolled 28 women (Figure 3) using primarily Facebook (11/28, 39%) and a university-wide, e-newsletter (11/28, 39%), and 22 women (age: mean 43.3, SD 4.69 years) completed the study. Of the 22 women who completed the study, 20 (90%) completed the postintervention acceptability survey. Of the 22 study completers, 17 (77%) participated in the postintervention focus groups. Of the women, 50% (11/22) possessed 3 of the 5 aforementioned metabolic syndrome risk factors, with a mean of 2.5 (SD 1.6) metabolic syndrome risk factors in the 28 enrolled women and a mean of 2.6 (SD 1.6) metabolic syndrome risk factors in the 22 study completers demonstrating similarities between the 2 groups (Table 1). Of the 28 women enrolled, 5 (18%) were physically active, 11 (39%) were married, and 24 (89%) were college-educated. The mean BMI for the enrolled sample was 40.9 (SD 8.7) kg/m^2. The enrolled women (N=28) identified 13 (SD 7.4) lifestyle stressors at study baseline, while the study completers (n=22) identified 11 (SD 7.2) lifestyle stressors at study baseline. Regarding technology, all 28 participants had in-home access to a desktop, tablet, or laptop computer (Table 2). Of the 28 participants, 27 (96%) had "smart" mobile phones, and 25% (7/28) had Fitbit accounts already established prior to the YogicDance study. Over one-half of the participants (15/28, 53%) had a YouTube account prior to the study.

Figure 3. YogicDance (YD) participant flow diagram.

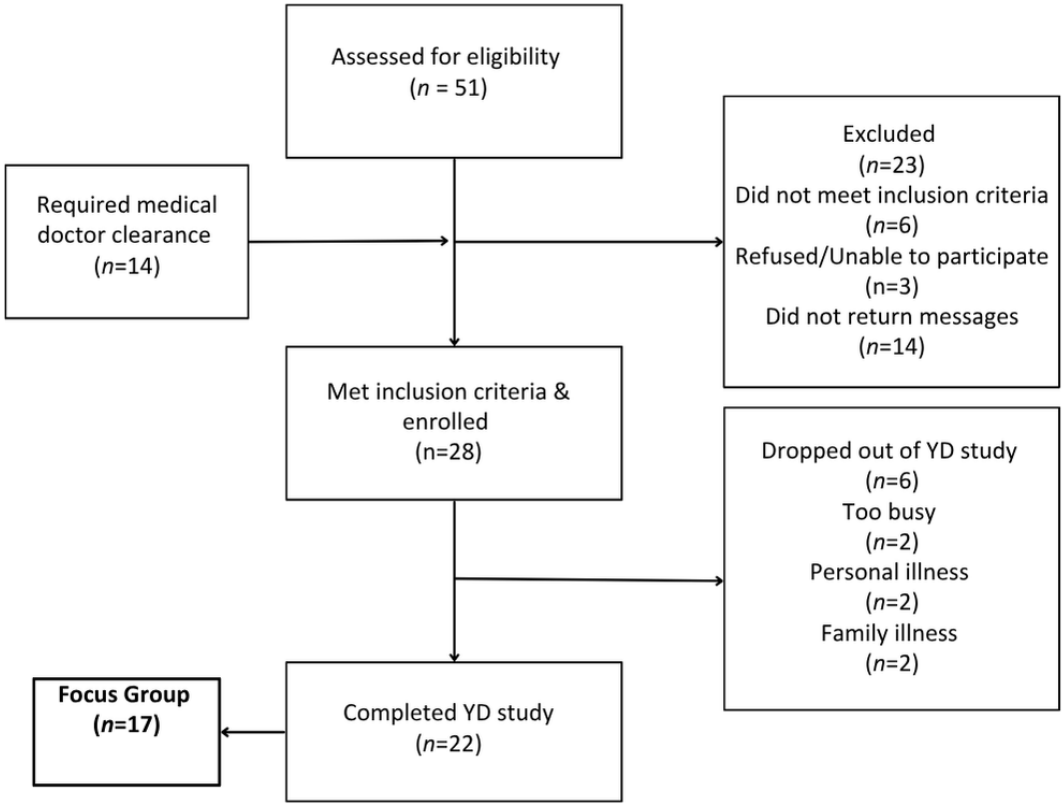

XSL•FO
RenderX

Table 1. Participant characteristics.

Characteristics	Total sample (n=28)	Study completers (n=22)
Demographics		
Age (years), mean (SD)	43.0 (4.5)	43.3 (4.7)
BMI (kg/m^2), mean (SD)	40.9 (8.7)	40.9 (8.6)
Marital status, n (%)		
Single, never married	10 (36)	10 (46)
Divorced, separated, widowed	5 (18)	3 (14)
Married, partnered	11 (39)	9 (41)
Highest level of education, n (%)		
Some college/associate's degree	4 (14)	4 (18)
Bachelor's degree	10 (36)	9 (41)
Master's degree	9 (32)	7 (32)
Doctorate	2 (7)	2 (9)
Yearly income range (US $), n (%)		
10,000-19,999	1 (4)	1 (5)
20,000-29,999	2 (7)	1 (5)
30,000-39,999	5 (18)	4 (18)
40,000-49,999	6 (21)	5 (23)
>50,000	12 (43)	11 (50)
Number of metabolic syndrome risk factors (from a total of 5), mean (SD)	2.5 (1.1)	2.6 (1.2)
Number of social supports, mean (SD)	1.6 (1.6)	1.7 (0.5)
Number of lifestyle stressors, mean (SD)	13.1 (7.4)	11.4 (7.2)
Activity Status, n (%)		
Sedentary-No	3 (11)	4 (18)
Sedentary-Yes	25 (89)	18 (82)
Metabolic syndrome, n (%)		
Low: <3 metabolic syndrome characteristics	13 (46)	11 (50)
High: ≥3 metabolic syndrome characteristics	15 (54)	11 (50)

XSL•FO
RenderX

Table 2. Participant technology access and internet recruitment (N=28).

Variables	Results, n (%)
Prior technology access	
Internet	28 (100)
Smartphone	27 (96)
Laptop, desktop, tablet	28 (100)
DVD player	27 (96)
Skype	13 (46)
Prior activity tracker	2 (7)
Prior online Fitbit account	7 (25)
Prior YouTube account	15 (54)
Prior in-home BP[a] monitor	7 (25)
Internet recruitment methods	
Study Facebook page	11 (39)
University Facebook page	0 (0)
University online newsletter	11 (39)
Gmail flier	3 (11)
Google Voice message	1 (4)
Print flier	2 (7)

[a]BP: blood pressure.

Engagement With the YogicDance Intervention

This pilot study supported the feasibility of a larger study in its ability to identify and recruit eligible subjects. Completing the study was associated with being engaged in 2 or more of the 4 main aspects of the intervention including watching the videos, taking daily blood pressures, completing the web-based surveys, and wearing the Fitbit trackers >3.5 days per week (χ^2_1=7.14, *P*=.008; Table 3). Watching over one-half of the instructional videos was associated with completing the study (χ^2_1=4.39, *P*=.04), as was taking at-home blood pressures (χ^2_1=5.67, *P*=.02). Study completion was not associated with merely completing the online surveys alone (χ^2_1=3.60, *P*=.06) or wearing the Fitbit tracker only (χ^2_1=0.392, *P*=.53).

Table 3. YogicDance study engagement (n=22).

Study engagement measures	Quantitative results	
	Likelihood-ratio, χ^2 (*df*)	*P* value
Total study activities completed	7.14 (1)	.008
Videos watched	4.39 (1)	.04
Fitbit worn	0.39 (1)	.53
Web-based surveys taken	3.60 (1)	.058
Blood pressure taken	5.67 (1)	.02

Feasibility and Acceptability of YogicDance Intervention

Acceptability and ease-of-use of the components of YogicDance intervention were evaluated using survey data. Overall, the study participants who completed postintervention acceptability surveys (n=20) found participating in the YogicDance intervention satisfying (19/20, 95%) and following along with the YogicDance video practice suitable (19/20, 95%; Table 4). Nearly all (19/20, 95%) of the women found the overall screening process suitable. In addition, 70% (14/20) of the participants found study researchers coming into their home to do baseline measurements suitable, while 85% (17/20) of the participants thought visiting the urban university campus for postintervention study measures suitable. Furthermore, 80% (16/20) of the women who completed the acceptability survey thought the web-based survey questions were clear, while 95% (19/20) of them found the survey platform, REDCap, to be easy to use. With regards to self-monitoring devices, 70% (14/20) and 95% (19/20) of the women found the blood pressure monitor and the Fitbit activity tracker, respectively, to be easy to use. Though 80% (16/20) of the women found the YogicDance

XSL•FO
RenderX

videos to be useful, only 50% (10/20) of the women found the YouTube website to be easy to use. In focus group discussions, participants shared how the use of Fitbits raised awareness about physical activity levels, sleep patterns, and the impact of exercise on blood pressure. In focus group discussions, participants expressed initial enthusiasm about participating in the yoga study and were eager to try something new and improve their health. Some participants were motivated by personal health concerns or a desire to reduce stress. Participants appreciated the immediate feedback on their health metrics, which motivated some to increase their activity levels. Exemplar quotes that represent sentiments of the collective focus group are provided. Pseudonyms are provided so that readers can understand how representative the qualitative data are.

Table 4. YogicDance (YD) study acceptability (n=20).

Acceptability survey measure	Quantitative results, n (%)	Exemplar quotes
Screening process suitable	19 (95)	None
Researchers coming to home suitable	14 (70)	“Having the researchers come to my home was convenient and put me at ease.” [Jasmine]
Poststudy visit to campus was suitable	17 (85)	“Visiting the campus for post-study measures was well-organized and professional.” [Jasmine]
Survey questions clear	16 (80)	“The survey questions were clear and easy to understand.” [Aisha]
Easy-to-use web surveys	19 (95)	“The web surveys were very user-friendly and simple to navigate.” [Jasmine]
Easy-to-use BP[a] monitor	14 (70)	“The blood pressure monitor was a bit tricky at first, but overall, it was a useful tool.” [Aisha]
Easy-to-use Fitbit tracker	19 (95)	“I really enjoyed using the Fitbit; it encouraged me to be more active.” [Ebony]
YD video practice suitable	19 (95)	“Following along with the YogicDance videos was suitable and enjoyable.” [Aisha]
Helpful health education videos	18 (90)	“The health education videos were very helpful and informative.” [Aisha]
YouTube suitable and easy to use	10 (50)	“Using YouTube was a bit cumbersome due to the login process.” [Jasmine]
Social aspects of YouTube satisfying	17 (85)	“The social aspects of YouTube were satisfying, though I wished for more interaction options.” [Keisha]
YouTube website helpful	16 (80)	“YouTube was helpful for accessing the videos, but more variety would have been nice.” [Keisha]
Yoga instructor suitable	15 (75)	“The yoga instructor was excellent and made the sessions engaging.” [Shanice]
Physical trainer suitable	14 (70)	“The physical trainer provided clear instructions that were easy to follow.” [Tanisha]
Chair Yoga Program satisfying	14 (70)	“The Chair Yoga Program was satisfying and perfect for my needs.” [Jasmine]
Floor Mat Program satisfying	15 (75)	“I enjoyed the Floor Mat Program. It was a great workout.” [Ebony]
Participation in study satisfying	19 (95)	“Overall, participating in this study was a very satisfying experience.” [Jasmine]

[a]BP: blood pressure.

Benefits of and Barriers to YogicDance Intervention

Of the 22 study completers, 17 attended the postintervention, semistructured focus group interviews. The following 4 important themes emerged from the qualitative analyses: accountability, increased awareness, health benefits, and unanticipated stressors (Table 5).

XSL•FO
RenderX

Table 5. Themes and subthemes derived from YogicDance qualitative data (n=17).

Theme	Related word frequency	Subtheme	Exemplar quotes
Accountability	Accountability (n=14), accountable (n=11)	• Black women feel they need to be accountable to others before themselves; they put themselves last. • Study participation provided opportunities for self-accountability via self-monitoring devices; however, there were missed opportunities for accountability via social support. • The social support aspect of practicing yoga is missed in the home-based, video watching experience.	• "...and you're always putting everyone else first and making sure the house is standing the way that it's supposed to be. ...we're neglecting ourselves and it's easier for me to tell you 'girl if you don't take care of yourself…' and if my daughter sneezes the wrong way it's like we've gotta make a doctor's appointment or if my husband's leg is bothering [him] we gotta get you to the doctor. All the while, I'm probably the sickest one in the house..." [Ebony] • "...as far as accountability, I like that it's [the intervention], a situation that I'm obligated to and I have to do it despite whatever else I have going on, so when I heard about Black women, health, [I thought] I'm there." [Jasmine] • "I need accountability in a group in regards to eating healthy and working out. I tried it by myself...but when it comes to trying to eat right and working out I can't do what I want. So I need people that will be able to encourage me and work together towards that type of goal." [Tamika]
Increased awareness	Awareness (n=9), aware (n=5)	• Tailored health education videos provided informative, enlightening explanations regarding cardiometabolic health and safe yoga practice. • Self-monitoring technology revealed negative health patterns and informed health-promoting behaviors.	• "I fairly enjoyed the educational videos, surprisingly. They were very informative. Sometimes when you are presented with medical or scientific information it can be quite overwhelming but it was all presented in very layman terms." [Aisha] • "I wanted to participate in the yoga study which actually explains yoga because I needed to learn how to relax, de-stress, meditate..." [Jasmine] • "I was used to being aware of where my blood pressure is, but taking it daily gave me a really good baseline of what I should be looking for; so it was extremely helpful." [Jasmine] • "...I really enjoyed it because it encouraged me to walk more, like if I took my daughter to the park, I was encouraged to walk around while she was playing instead of just sitting there. Or even at work it encouraged me to take steps..." [Monique]
Health benefits	Beneficial (n=10), benefits (n=8), benefit (n=7)	• Study engagement was perceived to be a step toward lowering risk of high blood pressure and diabetes, especially in those with family history. • Home-based videos and self-monitoring devices and logs provided opportunities for participants to see the immediate positive impact of intervention activities on health outcomes.	• "For me it was definitely a wake-up of how I need to make time for myself. Um and the ten-fifteen minutes or the fifteen to twenty the variety um of the videos that were offered really made me think about how much time I waste not taking care of myself. Um fifteen minutes you'll sit and watch a movie for an hour two three you'll sit there binge on something for hours and not take care of yourself so for me it was a lot of introspection that any amount of activity of any period of day you can take the time to do for yourself..." [Shanice] • "The breathing exercises were awesome for that. I didn't know how much that would… I mean I heard people say take deep breaths and stuff like that but actually focusing on my breathing and focusing on a specific way to do it as the instructor taught us. I really had to focus on that and it took my mind off of other things. Also the breathing itself helped to just relax my mind and body. It was a new experience to learn how effective just breathing could be in reducing stress. Before, I never paid much attention to how I was breathing, but this study taught me the importance of mindful breathing and how it can significantly impact my mental state and overall well-being." [Ebony] • "The reason why I wanted to be a part of this study is because I do currently have high blood pressure...I'm trying to figure out how to reduce that and I was really hoping that I would learn some tricks and tips from the yoga. I do have a family history of high blood pressure, diabetes, and I don't want to go too much further down that road." [Jasmine] • "I have a family history of high blood pressure, diabetes, and heart disease. And I also have an interest in yoga." [Ebony] • "It was really beneficial to me to understand my blood pressure. Since within the last year and a half I was recently diagnosed with high blood pressure. It was interesting to me to see the change of my blood pressure based on exercise, to see how I was feeling before and after..." [Aisha] • "And I loved having to do my blood pressure a couple of times because I could see the impact that even walking for twenty minutes or thirty minutes a day was influencing it. So it helped to be more aware of the steps that I have to take on a consistent basis so that I don't have to take any kind of medicine." [Keisha]

XSL•FO
RenderX

Theme	Related word frequency	Subtheme	Exemplar quotes
Unanticipated stressors	Stress (n=6), stressed (n=8), stressful (n=9), distress (n=8), stresses (n=8), stressors (n=9)	• Work, family, and community-level stressors influenced study engagement. • YouTube could have been used for social support. • Videos were time-consuming and could have had more variety.	• "...I think the added stress that sometimes Black women and Black people have is that societal pressure that comes from things that happen to our community...when I was taking this survey, it was right after Philando Castile and Alton Sterling happened. And when you work with people of other races, it's really difficult for them to understand how someone you never met, their death impacts you in a very personal way. And I was stressed when I took that survey." [Keisha] • "...and I got to a point where-I have my normal everyday stresses, and I'm doing what I can for the Black community, and there's no more that I can do so I'm just going to go ahead and turn the TV off. That combined with Donald Trump running for president, I stopped watching the news." [Monique] • "In reference to the YouTube video website, it was kind of cumbersome. ...if I had time to go back and do an additional video or so to go back and look through the links to see which one I wanted. So say if I watched the informational video to go back or for the resistance I had to go back into the emails that I had and find them. So that was a little cumbersome. But again, I understand because of privacy." [Jasmine] • "I think YouTube is good because I think so many people use YouTube and it's very accessible. But I think again because it was the same 4 videos it was really hard to be like 'you go, girl!' But if it was more like 'I just did that one,' or 'I'm on this one' or if it got harder then I probably would have wanted some social support." [Jasmine] • "I just think if it was more videos it might have taken off with that piece." [Tanisha] • "I've mentioned this before and it's been said, after I went through all the breathing, the endurance, the chair exercise, the floor exercise, then I was looking for something to be different the next week. But you were like oh we just recycle those and I went back and sort of got stuck in the chair exercise after that." [Latoya]

Accountability

Self- and group accountability for participating in health-promoting behaviors were seen by the women as both a benefit and a barrier to YogicDance engagement. The women identified with being accountable to many others in their lives, often putting themselves and their health last in priority and not being accountable to their own health promotion. Conversely, participating in a research study targeting Black women was seen as a way of holding themselves accountable for their own health and providing them opportunities for group accountability. Participation in the focus group gave them an opportunity to express the phenomenon of putting the needs of others before their own. Participants were interested in components of the intervention that encouraged them to hold themselves accountable for healthy behaviors. The women highlighted their desire to be involved in group physical activities like yoga, hoping for opportunities to be accountable to a group with similar health goals.

Increased Awareness

Study participants considered the intervention videos to be educational and enlightening. The women highlighted their desires to learn more about their own health in a way that was straightforward and easy to understand. They also described the benefits of increased health awareness. The study prompted self-reflection on personal health habits and lifestyle choices. Participants became more aware of their need to prioritize self-care and manage stress effectively.

In addition to the videos, the women found the physical activity and blood pressure self-monitoring devices helpful for revealing their own negative health patterns (ie, low activity levels and elevated blood pressures). Daily self-monitoring increased awareness of these unhealthful patterns and provided real-time impetus for positive health behavior change.

Health Benefits

Participants provided positive feedback on the yoga and breathing exercises, with many finding them relaxing and beneficial. Breathing exercises and yoga practices were highlighted as beneficial for reducing stress and improving mental clarity. Several participants mentioned the mental health benefits of focusing on breathing and relaxation techniques. The women considered engagement in the yoga intervention as a means for addressing their family histories of cardiometabolic disease. They believed, by engaging with the yoga intervention, their health would improve. Women were interested in learning more about the benefits of yoga and in experimenting with the positive health impacts of yoga practice. Participants' tracking of blood pressures following daily YogicDance practice provided opportunities for immediate gratification, as women often noted immediate beneficial impacts on their own blood pressures following yoga videos.

XSL•FO
RenderX

Unanticipated Stressors

Participants noted difficulties in maintaining daily routines and adhering to the study requirements due to various life stressors. Despite these challenges, many participants expressed a sense of obligation to complete the study and not let the group down. Work and community-level stressors played a role in the participants' study engagement. The participants resonated with having a sense of personal responsibility for their local Black communities, and they identified with psychological archetypes of being strong for and supportive of their communities in general. As a result, perceived responsibilities to their communities were woven into the personal lives of the participants. The women provided examples of the stressors that resulted from their feelings of responsibility to others on the community level and the ways these feelings impacted their engagement with the YogicDance intervention. Participants exemplified stressors that emerged during the intervention, which took place during the summer before a US presidential election. They described community-level stressors related to escalating news reports of community violence and political upheaval.

Other unanticipated stressors participants encountered during the intervention had to do with their restricted use of YouTube. Prior to beginning the intervention, participants were asked to use researcher-provided usernames and login passwords to pre-assigned email accounts with affiliated YouTube accounts. They were asked to login using YouTube video links emailed to their newly assigned email accounts and not their own personal YouTube accounts. These measures were taken to protect participant identities and to preserve their rights to privacy and confidentiality. Some participants struggled with certain exercises, particularly those involving resistance bands, and suggested the need for more tailored modifications. When participants were encouraged to share their thoughts on the effectiveness of the YouTube platform itself, they provided meaningful feedback regarding their desire for the intervention to have more variety than the few restorative mat and chair yoga programs being offered. Additionally, the participants wanted more opportunities to show social support to one another. Participants provided mixed feedback on the educational videos: Some found them informative and well presented, while others felt they were too long and sometimes difficult to engage with. Suggestions for improvement included shorter, more varied videos and more interactive content.

Discussion

Benefits to Engagement

Our study highlights the successful implementation and positive outcomes of the YogicDance intervention, demonstrating the feasibility and acceptability among Black women at high risk for metabolic syndrome. The interpretation of the quantitative and qualitative data collected and analyzed from this explanatory sequential mixed methods study determines the quantitative results that require further explanation. Key findings include significant engagement with the intervention's components, positive feedback on the technological aspects used for delivering the intervention, and overall satisfaction with the program's content and delivery method, with a significant majority of participants completing the intervention and the associated acceptability surveys and focus groups. Quantitative data showed high levels of engagement, with 95% of participants completing the postintervention acceptability survey. Qualitative data supported these findings, as participants expressed appreciation for the mental health benefits of focusing on breathing and relaxation techniques; the structured, culturally tailored health education; and the ease of use of digital platforms, which facilitated engagement. These results underscore the potential of culturally tailored, internet-delivered, yoga-based interventions to support lifestyle changes and address metabolic syndrome risk factors in this population [13].

Recruitment and Engagement Strategies

Our recruitment strategies leveraged digital platforms, which proved effective at engaging Black women in health interventions, reflecting participants' preferences for online interactions. This preference aligns with prior research indicating the effectiveness of web-based strategies at engaging Black women in health interventions [33]. Participants appreciated the convenience and accessibility of online recruitment, which included Facebook and a university-wide e-newsletter, contributing to the successful enrollment and retention of participants. A trial using Facebook and text messages with culturally relevant content increased participation, social support, and self-reported physical activity in a study on physical activity differences [34]. The use of culturally relevant images and messaging, along with the convenience and cost-saving aspects of online recruitment, contributed to the successful enrollment and retention of participants in an internet-based survey study [35].

Technological Engagement

Technological engagement was a critical component of the study, with participants actively using web-based surveys, instructional videos, Fitbit trackers, and electronic blood pressure monitors. Participants highlighted the benefits of digital engagement tools, such as instructional videos and Fitbit trackers, which provided a sense of accountability and motivation. The Fitbit and blood pressure monitor were generally well received for their ability to provide real-time health data. Issues with the accuracy and usability of these devices were noted, indicating a need for better onboarding and support. Quantitative data showed that 95% of participants found the web-based surveys and Fitbit trackers easy to use, underscoring the effectiveness of digital methods for engaging participants. The integration of these self-monitoring tools facilitated engagement and provided valuable insights into participants' health behaviors. Research demonstrates the effectiveness of web-based surveys for engaging participants and providing unique insights [35], particularly in cancer screening studies where online participants were more likely to follow screening recommendations [36]. Additionally, online survey completers older than 50 years typically had higher education and income [37]. In a study focusing on physical activity, web-based monitoring was preferred over paper journals [37], highlighting the effectiveness of digital methods for engaging participants. The integration of self-monitoring

XSL•FO
RenderX

tools and health education via digital platforms not only facilitated engagement but also provided insights into the participants' health behaviors and attitudes toward managing their cardiovascular risk [38-41].

Acceptability of YogicDance

The acceptability of the YogicDance intervention was high, with participants reporting satisfaction with the social aspects of YouTube, the instructional content, and the overall intervention. Quantitative results indicated that 95% of participants found the intervention suitable and satisfying. However, challenges with the YouTube platform's accessibility suggest the need for more user-friendly and accessible digital tools [42,43]. Despite these challenges, the intervention's impact on self-monitoring behaviors, particularly blood pressure monitoring, underscores the value of real-time feedback for motivating health behavior change and managing cardiovascular risk [44,45].

Barriers to Engagement

Despite these positive findings, participants identified several barriers to engagement, including lifestyle stressors and the complexity of accessing certain technological features. Many participants faced challenges such as time management issues, family obligations, and personal health problems that interfered with their ability to fully participate. Technical issues, such as problems with the blood pressure monitor and Fitbit, also posed barriers. Some participants found the online format less engaging and struggled with self-motivation. Existing research highlights the importance of addressing barriers like stress and time constraints due to family and caretaking responsibilities in efficacious physical activity diabetes prevention lifestyle interventions [45,46]. Participants' feedback indicated that issues with the YouTube platform's navigation impeded their ability to engage fully and suggested a need for improved usability and additional social support features. This feedback suggests the need for interventions to better use social media's features to enhance social support and engagement. Culturally competent care for minority women emphasizes the role of social context in health decisions, including the use of mind-body therapies [13].

Recommendations for Future Studies

Previous research has shown the significance of social support in encouraging participation in physical activities like YogicDance [47]. Barriers such as family-related stress and lack of social support, especially among low-income Black women, underscore the need for increased social support to mitigate cardiovascular disease risks [48]. Although the health education videos and YouTube platform were found helpful, participants recommended improvements like providing all educational content upfront, breaking videos into shorter segments, and offering more flexibility in how participants engage with the social media or social support platform features for future interventions [49,50]. More in-person or interactive online sessions were also recommended to enhance engagement and accountability. Enhancing social support features on platforms like YouTube could foster a sense of community and accountability, particularly beneficial for Black women's health [51,52]. Specifically, enhancing social support features and simplifying mobile technology use could further improve engagement and outcomes [53]. Additional recommendations were made for better onboarding for technological tools and clearer instructions for exercises. Participants also suggested incorporating more varied and advanced routines to cater to different fitness levels. These insights are crucial for refining future interventions to better meet the needs and preferences of Black women at high cardiovascular risk.

Limitations

A key limitation of our study is the potential for selection bias, as consistent, at-home internet access was an intervention prerequisite. This may have resulted in a sample with higher income stability and better broadband access than lower-income populations. Additionally, the higher education levels of our participants may limit the generalizability of our findings to women of varying socioeconomic backgrounds.

Conclusion

Our mixed methods study demonstrates the potential of the YogicDance intervention to engage Black women in physical activity and self-monitoring behaviors aimed at reducing cardiovascular risk. Using an explanatory sequential design approach, we gathered preliminary effect sizes for critical quantitative health biomarkers such as blood pressure, fasting blood glucose, waist circumference, and blood lipids, contributing to the groundwork for future research in this area. The qualitative insights from focus groups provide a deeper understanding of the unique benefits and barriers faced by participants, emphasizing the need for culturally tailored, user-friendly, digital tools to enhance engagement. Future research should build on these findings, addressing identified barriers such as lifestyle stressors and technical issues and leveraging digital platforms' strengths to improve health outcomes for this high-risk population.

Acknowledgments

This research was supported by an internal grant from Virginia Commonwealth University (VCU) School of Nursing. Its contents are solely the responsibility of the author and do not necessarily represent the official views of VCU. Use of the Research Electronic Data Capture (REDCap) platform was made possible by VCU's Clinical & Translational Science Award (CTSA; UL1TR000058 from the National Center for Research Resources) and the Center for Clinical and Translational Research (CCTR) Endowment Fund of the Virginia Commonwealth University. The authors would like to thank mentor, Dr. Faye Belgrave; videographer, Omari Johnson; physical trainer, Melanie N Bates; and yogi, Susan Turnage, for their contributions to this research.

XSL•FO
RenderX

The authors acknowledge the use of OpenAI's ChatGPT, a generative artificial intelligence (AI) tool, to address the comments and suggestions provided by the editors and reviewers during the revision process of this manuscript. ChatGPT was used to assist with refining responses to reviewer comments, generating textual content for revisions, and improving the clarity and coherence of the manuscript. The authors take full responsibility for the accuracy and integrity of the content presented in the final version of the manuscript. In compliance with JMIR's AI usage policy, we have documented the extent of AI involvement and retained original transcripts of ChatGPT interactions as supplementary materials (Multimedia Appendix 3).

Data Availability

The data sets generated or analyzed during this study are presented in the main manuscript. Additional supporting files, which include de-identified data, are available upon reasonable request. Researchers interested in accessing these files should contact the corresponding author, Dr. Candace Johnson, at cjohnso11@ltu.edu or cmc8k@virginia.edu, providing a brief description of the intended use of the data. Access will be granted to qualified researchers, ensuring adherence to ethical guidelines and data privacy regulations.

Authors' Contributions

CJ, LT, and PE contributed to writing the original draft. CJ contributed to conceptualization and writing, review, and editing of the manuscript. LT contributed to research design, statistical analyses, and project administration.

Conflicts of Interest

None declared.

Multimedia Appendix 1
YogicDance study components. This image depicts the key components of the YogicDance study, which aimed to engage African-American/Black women at high cardiovascular risk. The study included internet-based recruitment, community-based engagement, use of activity trackers, health monitoring, health education, and internet-based surveys. These components were designed to enhance engagement, monitor health, and educate participants to reduce cardiometabolic risks.
[PNG File , 180 KB - formative_v9i1e41221_app1.png]

Multimedia Appendix 2
YogicDance Facebook recruitment flier. This image displays the YogicDance research recruitment flier used on a clinical trials Facebook page of an urban university. The flier targets African-American women, featuring models from the study's demographic. It details the study's focus on beginner-level chair yoga, mat yoga, and light resistance training, and provides contact information for interested participants. The recruitment was conducted from summer 2015 to fall 2016, aiming to engage the community in a home-based research study.
[PNG File , 387 KB - formative_v9i1e41221_app2.png]

Multimedia Appendix 3
Documented Extent of AI Involvement in this Manuscript.
[DOCX File , 13 KB - formative_v9i1e41221_app3.docx]

References

1. Grundy SM. Metabolic syndrome: a multiplex cardiovascular risk factor. J Clin Endocrinol Metab 2007 Feb;92(2):399-404. [doi: 10.1210/jc.2006-0513] [Medline: 17284640]
2. Samson SL, Garber AJ. Metabolic syndrome. Endocrinol Metab Clin North Am 2014 Mar;43(1):1-23. [doi: 10.1016/j.ecl.2013.09.009] [Medline: 24582089]
3. Shrestha SS, Honeycutt AA, Yang W, Zhang P, Khavjou OA, Poehler DC, et al. Economic costs attributable to diabetes in each U.S. state. Diabetes Care 2018 Dec 10;41(12):2526-2534 [FREE Full text] [doi: 10.2337/dc18-1179] [Medline: 30305349]
4. Mosca L, Benjamin EJ, Berra K, Bezanson JL, Dolor RJ, Lloyd-Jones DM, American Heart Association. Effectiveness-based guidelines for the prevention of cardiovascular disease in women--2011 update: a guideline from the American Heart Association. J Am Coll Cardiol 2011 Mar 22;57(12):1404-1423 [FREE Full text] [doi: 10.1016/j.jacc.2011.02.005] [Medline: 21388771]
5. Khan RJ, Gebreab SY, Sims M, Riestra P, Xu R, Davis SK. Prevalence, associated factors and heritabilities of metabolic syndrome and its individual components in African Americans: the Jackson Heart Study. BMJ Open 2015 Nov 01;5(10):e008675 [FREE Full text] [doi: 10.1136/bmjopen-2015-008675] [Medline: 26525420]

XSL•FO
RenderX

6. Eyler AE, Wilcox S, Matson-Koffman D, Evenson KR, Sanderson B, Thompson J, et al. Correlates of physical activity among women from diverse racial/ethnic groups. J Women's Health Gend Based Med 2002 Apr;11(3):239-253. [doi: 10.1089/152460902753668448] [Medline: 11988134]
7. Holmes ME, Ekkekakis P, Eisenmann JC. The physical activity, stress and metabolic syndrome triangle: a guide to unfamiliar territory for the obesity researcher. Obes Rev 2010 Jul;11(7):492-507. [doi: 10.1111/j.1467-789X.2009.00680.x] [Medline: 19895413]
8. Uthman OA, Hartley L, Rees K, Taylor F, Ebrahim S, Clarke A. Multiple risk factor interventions for primary prevention of CVD in LMIC: a Cochrane Review. Glob Heart 2017 Sep;12(3):199-208.e8 [FREE Full text] [doi: 10.1016/j.gheart.2016.03.639] [Medline: 27452771]
9. Flack JM, Sica DA, Bakris G, Brown AL, Ferdinand KC, Grimm RH, International Society on Hypertension in Blacks. Management of high blood pressure in Blacks: an update of the International Society on Hypertension in Blacks consensus statement. Hypertension 2010 Nov;56(5):780-800. [doi: 10.1161/HYPERTENSIONAHA.110.152892] [Medline: 20921433]
10. Kirkendoll K, Clark PC, Grossniklaus D, Igho-Pemu P, Mullis R, Dunbar SB. Metabolic syndrome in African Americans: views on making lifestyle changes. J Transcult Nurs 2010 Apr;21(2):104-113 [FREE Full text] [doi: 10.1177/1043659609357636] [Medline: 20220030]
11. Greaves CJ, Sheppard KE, Abraham C, Hardeman W, Roden M, Evans PH, IMAGE Study Group. Systematic review of reviews of intervention components associated with increased effectiveness in dietary and physical activity interventions. BMC Public Health 2011 Feb 18;11:119 [FREE Full text] [doi: 10.1186/1471-2458-11-119] [Medline: 21333011]
12. Kong A, Tussing-Humphreys LM, Odoms-Young AM, Stolley MR, Fitzgibbon ML. Systematic review of behavioural interventions with culturally adapted strategies to improve diet and weight outcomes in African American women. Obes Rev 2014 Oct;15 Suppl 4(0 4):62-92 [FREE Full text] [doi: 10.1111/obr.12203] [Medline: 25196407]
13. Johnson CC, Sheffield KM, Brown RE. Mind-body therapies for African-American women at risk for cardiometabolic disease: a systematic review. Evid Based Complement Alternat Med 2018;2018:5123217 [FREE Full text] [doi: 10.1155/2018/5123217] [Medline: 29681975]
14. Haider T, Sharma M, Branscum P. Yoga as an alternative and complimentary therapy for cardiovascular disease: a systematic review. J Evid Based Complementary Altern Med 2017 Apr;22(2):310-316 [FREE Full text] [doi: 10.1177/2156587215627390] [Medline: 26787730]
15. Biggers A, Spears CA, Sanders K, Ong J, Sharp LK, Gerber BS. Promoting mindfulness in African American communities. Mindfulness (N Y) 2020 Oct;11(10):2274-2282 [FREE Full text] [doi: 10.1007/s12671-020-01480-w] [Medline: 33584869]
16. James-Palmer A, Anderson EZ, Daneault J. Remote delivery of yoga interventions through technology: scoping review. J Med Internet Res 2022 Jun 06;24(6):e29092 [FREE Full text] [doi: 10.2196/29092] [Medline: 35666562]
17. Kuijpers W, Groen WG, Aaronson NK, van Harten WH. A systematic review of web-based interventions for patient empowerment and physical activity in chronic diseases: relevance for cancer survivors. J Med Internet Res 2013 Feb 20;15(2):e37 [FREE Full text] [doi: 10.2196/jmir.2281] [Medline: 23425685]
18. Richards J, Hillsdon M, Thorogood M, Foster C. Face-to-face interventions for promoting physical activity. Cochrane Database Syst Rev 2013 Sep 30;2013(9):CD010392. [doi: 10.1002/14651858.CD010392.pub2] [Medline: 24085592]
19. Fanning J, Mullen SP, McAuley E. Increasing physical activity with mobile devices: a meta-analysis. J Med Internet Res 2012 Nov 21;14(6):e161 [FREE Full text] [doi: 10.2196/jmir.2171] [Medline: 23171838]
20. Fogg BJ, Allen E. 10 uses of texting to improve health. 2009 Presented at: 4th International Conference on Persuasive Technology; April 26-29, 2009; Claremont, CA. [doi: 10.1145/1541948.1541997]
21. Fox S, Duggan M. Tracking for Health. Pew Research Center. 2013 Jan 28. URL: https://www.pewresearch.org/internet/2013/01/28/tracking-for-health/ [accessed 2025-07-18]
22. Bandura A. Social cognitive theory in cultural context. Applied Psychology 2002 Apr 04;51(2):269-290. [doi: 10.1111/1464-0597.00092]
23. Feather NT. Expectations and Actions: Expectancy-Value Models in Psychology. London, England: Routledge; 1982.
24. Walker SN, Volkan K, Sechrist KR, Pender NJ. Health-promoting life styles of older adults: comparisons with young and middle-aged adults, correlates and patterns. ANS Adv Nurs Sci 1988 Oct;11(1):76-90. [doi: 10.1097/00012272-198810000-00008] [Medline: 3140723]
25. Brady B, Nies MA. Health-promoting lifestyles and exercise: a comparison of older African American women above and below poverty level. J Holist Nurs 1999 Jun;17(2):197-207. [doi: 10.1177/089801019901700207] [Medline: 10633652]
26. Ahijevych K, Bernhard L. Health-promoting behaviors of African American women. Nurs Res 1994;43(2):86-89. [Medline: 8152943]
27. Jefferson VW, Melkus GD, Spollett GR. Health-promotion practices of young black women at risk for diabetes. Diabetes Educ 2000;26(2):295-302. [doi: 10.1177/014572170002600210] [Medline: 10865595]
28. Creswell J, Creswell J. Research Design: Qualitative, Quantitative, and Mixed Methods Approaches, sixth edition. New York, NY: SAGE Publications; 2022.
29. Frith H. Research Methods for Managers. In: Gill J, Johnson P, editors. Data collection: Asking questions and collecting data. London: Routledge; 2007:141-165.

XSL•FO
RenderX

30. Warburton DER, Bredin S, Jamnik VK, Gledhill N. The Physical Activity Readiness Questionnaire for everyone (PAR-Q+) and electronic Physical Activity Readiness Medical Examination (ePARmed-X+). Health & Fitness Journal of Canada 2011;4(2):1 [FREE Full text] [doi: 10.14288/hfjc.v4i2.103]
31. Craig CL, Marshall AL, Sjöström M, Bauman AE, Booth ML, Ainsworth BE, et al. International Physical Activity Questionnaire: 12-country reliability and validity. Medicine & Science in Sports & Exercise 2003;35(8):1381-1395. [doi: 10.1249/01.mss.0000078924.61453.fb]
32. Elo S, Kyngäs H. The qualitative content analysis process. J Adv Nurs 2008 Apr;62(1):107-115. [doi: 10.1111/j.1365-2648.2007.04569.x] [Medline: 18352969]
33. Staffileno BA, Zschunke J, Weber M, Gross LE, Fogg L, Tangney CC. The feasibility of using Facebook, Craigslist, and other online strategies to recruit young African American women for a web-based healthy lifestyle behavior change intervention. J Cardiovasc Nurs 2017;32(4):365-371. [doi: 10.1097/JCN.0000000000000360] [Medline: 27428356]
34. Joseph RP, Keller C, Adams MA, Ainsworth BE. Print versus a culturally-relevant Facebook and text message delivered intervention to promote physical activity in African American women: a randomized pilot trial. BMC Women's Health 2015 Mar 27;15:30 [FREE Full text] [doi: 10.1186/s12905-015-0186-1] [Medline: 25886945]
35. Mahalingaiah S, Cheng JJ, Winter MR, Rodriguez E, Fruh V, Williams A, et al. Multimodal recruitment to study ovulation and menstruation health: internet-based survey pilot study. J Med Internet Res 2021 Apr 16;23(4):e24716 [FREE Full text] [doi: 10.2196/24716] [Medline: 33861203]
36. Hochheimer CJ, Sabo RT, Krist AH, Day T, Cyrus J, Woolf SH. Methods for evaluating respondent attrition in web-based surveys. J Med Internet Res 2016 Nov 22;18(11):e301 [FREE Full text] [doi: 10.2196/jmir.6342] [Medline: 27876687]
37. Kelfve S, Kivi M, Johansson B, Lindwall M. Going web or staying paper? The use of web-surveys among older people. BMC Med Res Methodol 2020 Oct 08;20(1):252 [FREE Full text] [doi: 10.1186/s12874-020-01138-0] [Medline: 33032531]
38. Nyenhuis SM, Balbim GM, Ma J, Marquez DX, Wilbur J, Sharp LK, et al. A walking intervention supplemented with mobile health technology in low-active urban African American women with asthma: proof-of-concept study. JMIR Form Res 2020 Mar 11;4(3):e13900 [FREE Full text] [doi: 10.2196/13900] [Medline: 32159520]
39. Auxier B, Anderson M. Social media use in 2021. Pew Research Center. 2021 Apr 07. URL: https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/ [accessed 2024-03-18]
40. Health topics. Pew Research Center. 2011 Feb 01. URL: https://www.pewresearch.org/internet/2011/02/01/health-topics-4/ [accessed 2024-03-18]
41. Jabour SM, Page A, Hall SF, Rodriguez L, Shields WC, Alvanzo AA. Information and communication technologies interest, access, and use: cross-sectional survey of a community sample of urban, predominantly Black women. J Med Internet Res 2018 Aug 14;20(8):e248 [FREE Full text] [doi: 10.2196/jmir.9962] [Medline: 30108036]
42. Leigh JW, Gerber BS, Gans CP, Kansal MM, Kitsiou S. Smartphone ownership and interest in mobile health technologies for self-care among patients with chronic heart failure: cross-sectional survey study. JMIR Cardio 2022 Jan 14;6(1):e31982 [FREE Full text] [doi: 10.2196/31982] [Medline: 35029533]
43. Chu P, Gotink RA, Yeh GY, Goldie SJ, Hunink MGM. The effectiveness of yoga in modifying risk factors for cardiovascular disease and metabolic syndrome: a systematic review and meta-analysis of randomized controlled trials. Eur J Prev Cardiol 2016 Feb;23(3):291-307. [doi: 10.1177/2047487314562741] [Medline: 25510863]
44. Wexler R, Elton T, Pleister A, Feldman D. Barriers to blood pressure control as reported by African American patients. J Natl Med Assoc 2009 Jun;101(6):597-603. [doi: 10.1016/s0027-9684(15)30947-0] [Medline: 19585931]
45. Schaeffer K. 5 facts about how Americans use Facebook, two decades after its launch. Pew Research Center. 2024 Feb 02. URL: https://www.pewresearch.org/short-reads/2024/02/02/5-facts-about-how-americans-use-facebook-two-decades-after-its-launch/ [accessed 2024-03-18]
46. Tenfelde SM, Hatchett L, Saban KL. "Maybe black girls do yoga": a focus group study with predominantly low-income African-American women. Complement Ther Med 2018 Oct;40:230-235. [doi: 10.1016/j.ctim.2017.11.017] [Medline: 30219456]
47. Dall TM, Zhang Y, Chen YJ, Quick WW, Yang WG, Fogli J. The economic burden of diabetes. Health Aff (Millwood) 2010 Feb;29(2):297-303. [doi: 10.1377/hlthaff.2009.0155] [Medline: 20075080]
48. Mosca L, Hammond G, Mochari-Greenberger H, Towfighi A, Albert MA. Fifteen-year trends in awareness of heart disease in women. Circulation 2013 Mar 19;127(11):1254-1263. [doi: 10.1161/cir.0b013e318287cf2f]
49. Johnson CC, Taylor AG, Anderson JG, Jones RA, Whaley DE. Feasibility and acceptability of an internet-based, African dance-modified yoga program for African-American women with or at risk for metabolic syndrome. J Yoga Phys Ther 2014;4:1 [FREE Full text] [doi: 10.4172/2157-7595.1000174] [Medline: 25593785]
50. Dlugonski D, Martin TR, Mailey EL, Pineda E. Motives and barriers for physical activity among low-income Black single mothers. Sex Roles 2016 Dec 2;77(5-6):379-392. [doi: 10.1007/s11199-016-0718-7]
51. Divens LL, Chatmon BN. Cardiovascular disease management in minority women: special considerations. Crit Care Nurs Clin North Am 2019 Mar;31(1):39-47. [doi: 10.1016/j.cnc.2018.11.004] [Medline: 30736934]
52. Kalinowski J, Idiong C, Blackman-Carr L, Cooksey Stowers K, Davis S, Pan C, et al. Leveraging the Black Girls Run web-based community as a supportive community for physical activity engagement: mixed methods study. JMIR Form Res 2023 Sep 07;7:e43825 [FREE Full text] [doi: 10.2196/43825] [Medline: 37676722]

XSL•FO
RenderX

53. Sequi-Dominguez I, Alvarez-Bueno C, Martinez-Vizcaino V, Fernandez-Rodriguez R, Del Saz Lara A, Cavero-Redondo I. Effectiveness of mobile health interventions promoting physical activity and lifestyle interventions to reduce cardiovascular risk among individuals with metabolic syndrome: systematic review and meta-analysis. J Med Internet Res 2020 Aug 31;22(8):e17790 [FREE Full text] [doi: 10.2196/17790] [Medline: 32865503]

Abbreviations

CHERRIES: Checklist for Reporting Results of Internet E-Surveys
HPM: health promotion model
IPAQ: International Physical Activity Questionnaire
IRB: institutional review board
PAR-Q+: Physical Activity Readiness Questionnaire+
PI: principal investigator
REDCap: Research Electronic Data Capture

Edited by A Mavragani; submitted 19.07.22; peer-reviewed by N Jiwani, SQ Yoong; comments to author 04.12.23; revised version received 21.03.24; accepted 13.08.24; published 17.09.25.

Please cite as:
Johnson CC, Ezouah P
Online Yoga Pilot Intervention for Black Women at High Cardiovascular Risk: Internet-Based Recruitment and Engagement
JMIR Form Res 2025;9:e41221
URL: https://formative.jmir.org/2025/1/e41221
doi:10.2196/41221
PMID:

©Candace Crosby Johnson, Pascaline Ezouah. Originally published in JMIR Formative Research (https://formative.jmir.org), 17.09.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Digital Psychosocial Service (Sui App) for Arabic-Speaking Refugees in Switzerland: Development and Cultural-Contextual Adaptation Using a User-Centered and Participatory Approach

Rilana Tanja Stoeckli[1], MSc; Thomas Berger[1], PhD; Monia Aebersold[2], MSc; Viktoria Zoellner[2], MSc; Farhad Haji[2]; Muriel Hunziker[1], MSc; Beatriz Jesus Ferreira[1], MSc; Michel Hosmann[2], MSc; Sebastian Burchert[3], Dipl-Psych; Jessica Wabiszczewicz[3], MSc; Christine Knaevelsrud[3], PhD; Eva Heim[4], PhD

[1]Department of Clinical Psychology and Psychotherapy, Institute of Psychology, University of Bern, Bern, Switzerland

[2]Swiss Red Cross, Bern, Switzerland

[3]Department of Clinical Psychological Intervention, Freie Universität Berlin, Berlin, Germany

[4]Department of Psychology, University of Lausanne, Lausanne, Switzerland

Corresponding Author:
Rilana Tanja Stoeckli, MSc
Department of Clinical Psychology and Psychotherapy
Institute of Psychology
University of Bern
Fabrikstrasse 8
Bern, 3012
Switzerland
Phone: 41 316845419
Email: rilana.stoeckli@unibe.ch

Abstract

Background: Upon arriving in host countries, forcibly displaced people face psychological, cultural, as well as sociostructural challenges. Access to mental health and psychosocial support (MHPSS) remains limited, affecting both refugees and host country structures. Digital services offer promise in addressing these challenges, given their potential for scalability and accessibility. Despite the increasing use of digital MHPSS, cultural and contextual adaptation remains insufficiently documented, requiring systematic documentation.

Objective: This study aims (1) to assess the psychosocial needs of newly arrived refugees in Switzerland and identify potential digital support solutions and (2) to participatorily develop and culturally adapt the Sui app to address those needs for Arabic-speaking refugees.

Methods: We used a 2-phase adaptation process, structured using the RECAPT (Reporting Cultural Adaptation in Psychological Trials) framework to ensure systematic documentation. The preparation phase (2019-2021) included a desk review, needs assessment interviews, and conceptualization discussions. The development phase (2021-2022) included iterative development loops with a user advisory board (UAB) and a beta test, followed by final adjustments.

Results: The desk review provided 5 key insights guiding development: mental health stigma, health literacy, sociostructural aspects, adherence to digital interventions, and task-shifting approaches. Findings from the needs assessment with 22 asylum care interviewees, 2 intercultural interpreters, and 4 target group interviewees confirmed the importance of integrating sociostructural and psychological factors. Through conceptualization discussions with a consultancy agency and 4 UAB (n=9) meetings, the scope of the Sui app was drafted: 9 relevant, everyday life topics (eg, asylum process, housing, and work), 5 psychological topics (eg, stress, sleep, and emotion regulation), and 1 peer-guided chat support feature. A multilevel translation process, involving 5 Arabic speakers from different countries, ensured linguistic accessibility. This allowed the UAB to focus on surface, content, and delivery during 19 development loops. The app's content was delivered in various formats, including text, illustrations, video testimonies, and audio exercises. The beta test revealed high acceptability, with most users engaging with the app several times a week. However, technical challenges such as slow loading times hindered the full exploration of the features. Participants valued the peer support function, highlighting the importance of faster response times and personalized messages. Based on the beta test, technical and content refinements were made to prepare the app for a future quantitative evaluation.

XSL•FO
RenderX

Conclusions: Through engagement with stakeholders and adherence to the RECAPT framework, we carefully considered the cultural and contextual circumstances of Arabic-speaking refugees newly arrived in Switzerland. The iterative participatory development process, spanning a preparation and development phase, ensured usability, accessibility, and relevance, while highlighting challenges in technical implementation and peer support. This study contributes to the existing knowledge of mental health and the needs of refugees and provides insights for future cultural adaptation of MHPSS interventions.

(JMIR Form Res 2025;9:e59905) doi:10.2196/59905

KEYWORDS

e–mental health; cultural adaptation; psychosocial support; digital MHPSS; refugees; participatory research; peer support; Arabic

Introduction

Psychosocial Impact of Forced Displacement

Forced displacement, which happens primarily in low- and middle-income countries (LMICs), poses urgent humanitarian challenges worldwide. The number of individuals fleeing armed conflict, persecution, or environmental threats has increased to more than 108 million worldwide [1]. This mass movement imposes substantial challenges on the sociostructural dynamics on the side of host countries [2]. On the side of the affected individuals, the impact is apparent on several levels of their well-being.

Refugees and other forcibly displaced people may encounter psychological difficulties, including posttraumatic stress disorder (PTSD), anxiety, depression, psychosomatic disorders [3-8], and low quality of life [9,10]. Premigration traumatic events that influence mental health [11] are further compounded by life-threatening events occurring on the migration route [12]. Finally, in countries of destination, sociostructural impediments and cultural differences may deteriorate the potentially vulnerable mental health state in which refugees find themselves [13-16]. In the literature, these postmigration living difficulties (PMLDs) were found to have an impact on mental health beyond the effect of premigration events [14,17-22] or at least increase mental health difficulties [23-25]. In the ADAPT (adaptation and development after persecution and trauma) model, Silove [16] postulates the 5 core pillars—“safety and security,” “bonds and networks,” “justice,” “roles and identities,” and “existential meaning”—as a holistic approach to explain the interplay of past and ongoing challenges affecting psychosocial well-being of displaced people. Restoring those core pillars interdependently is explained to be crucial to promoting psychosocial recovery.

Care Gap

Switzerland currently counts over 220,000 people in the asylum sector, including refugees and asylum seekers, with more than 20,000 from Syria alone. Recent asylum claims continue to come from 2 Arabic-speaking countries, Syria and Algeria, both of which rank among the top 5 countries of origin [26]. Disconcertingly, the critical need for psychological and sociostructural support is largely unmet, both in LMICs and in higher-income countries [2,27,28], such as Switzerland [29,30]. Specialists in Switzerland estimate that around 50% of refugees experience trauma-related psychological issues, with a large proportion untreated [30]. The limited access to essential mental health and psychosocial support (MHPSS) services is often caused on 2 levels. On the structural level, there is, for example, in Switzerland, a lack of mental health specialists working in the field of refugee mental health [30] and an immense overload weighing on social workers [31]. The latter results in fewer individuals being able to transition away from social welfare [32]. In addition, there is a lack of interpreters’ cost coverage [33], trained interpreters, childcare, and transportation (eg, to therapy) possibilities [28]. On the individual’s level, fear of stigma, language barriers, financial worries, and lack of self-perception of the mental health state are mentioned as barriers to psychological care [34-36]. For these reasons, the broad population of refugees, all of whom are facing major life changes, does not receive adequate psychosocial support [29]. More low-threshold psychosocial interventions are required, targeting the ADAPT model’s core pillars, PMLDs, and symptoms of mental distress to meet refugees’ psychological needs while taking into account their current life circumstances [16,25,37]. Swiss specialists have specifically recommended improving access to low-threshold services for individuals still living in asylum facilities [30].

Scalable Approaches

The World Health Organization has developed community-based and accessible interventions (PM+ and SH+) to reduce symptoms of depression, anxiety, and PTSD [38]; to alleviate PMLDs [39]; to prevent mental disorders [40]; and to improve mental well-being [41]. A transdiagnostic treatment approach to reach refugees at the symptom level (the Common Elements Treatment Approach) was proven to be effective in 2 LMICs [42].

In Switzerland, while several low-threshold short interventions for distressed refugees have been documented, only a few have been tested through randomized controlled trials (RCTs) [43]. However, to our knowledge, no service has been evaluated that addresses the lack of social assistance within the context of psychosocial support. Given these persistent challenges in providing psychological and sociostructural support for refugees, exploring innovative solutions is imperative.

Emerging digital interventions have gained current and prospective research priority in humanitarian MHPSS settings [44]. Digital interventions have been shown to improve mental health difficulties such as depression, anxiety, posttraumatic symptoms, functioning, well-being, and quality of life among minority communities with large effect sizes [45]. A meta-analysis focusing on LMICs reported moderate to large effects of digital interventions on depression and anxiety, and a small impact on quality of life [46]. Digital mental health

XSL•FO
RenderX

interventions offer the opportunity to overcome barriers such as limited structural or geographical access to specialists, language barriers, and fear of stigma [2,47]. This is particularly relevant for asylum seekers who, for example, in Switzerland, often reside in centers with limited access to urban areas [48], but usually have access to a smartphone [49,50]. Moreover, studies on internet-based interventions treating various mental disorders report comparable effectiveness as face-to-face interventions [51,52], with a conceivable advantage in cost-effectiveness [53] and a large potential for scalability [54].

The Lancet Commission on Global Mental Health has encouraged task-shifting interventions, including the involvement of nonspecialists in service provision, and advocates adopting digital platforms to scale up support for broader populations of people affected by mental ill health [55]. Internet-based interventions tend to be more effective [56] and have greater adherence [57] when they include human guidance, as opposed to unguided interventions. Interestingly, the professional education of the guide does not seem to be a critical factor [58,59]. The World Health Organization's culturally adapted digital intervention Step-by-Step for treating depression trained nonspecialists to guide participants with weekly contact and demonstrated moderate effects in an RCT with Syrian refugees in Lebanon [60,61]. Using the task-shifting approach presents the opportunity to train peers who speak the target group's first language and share similar experiences [62].

Cultural and Contextual Adaptation

Studies focusing on culturally adapting psychological interventions aim to increase the acceptability and improve health outcomes for underserved target groups [63,64]. Cultural adaptation studies have mostly used "top-down" approaches in which pre-existing psychological interventions were adapted [65]. In contrast, "bottom-up" processes incorporate cultural aspects from the outset when developing new interventions [65].

A challenge of cultural adaptation remains the definition of the term. Resnicow and colleagues [66] explain cultural adaptation in 2 dimensions. On the observable dimension, so-called surface adaptations are made in order to increase acceptance. This includes visuals, audio, language, and delivery formats. In the second dimension, the deep structure is adapted. This includes a deep understanding of the social, cultural, historical, environmental, and psychological constructs that influence the health behavior of the target group. In the case of refugees, these constructs can be strongly influenced by various psychosocial pillars [16]. The term "culture," therefore, does not only comprise the individual's upbringing background (such as language, ethnicity, traditions, and country of birth). Culture is also dynamically shaped by recent experiences (eg, traumatic events during migration and interaction with the host community) and the current surroundings (eg, living in asylum centers, access to health care, and availability of digital devices).

Thus, we highlight the sensitivity to context as part of the cultural adaptation. In a Swiss study, for example, Syrian refugees most frequently reported social, cultural, and structural difficulties [67]. These included long waiting times for asylum decisions, travel restrictions, and challenges with diploma recognition, all of which are shaped by Switzerland's specific legal and asylum system. While other European countries also present postmigration challenges, each country has distinct characteristics in its asylum and integration systems [68], requiring a context-specific understanding and adaptation of support services.

The benefits of cultural adaptation in psychological interventions are discussed in the literature. Some research has found superior effects of culturally adapted interventions compared to nonadapted interventions [69], whereas the extent of adaptation may play a role [70]. Other research suggests that the effects of psychological interventions do not depend on cultural adaptation [46,71,72]. In any case, it remains unclear what cultural adaptations of interventions have precisely entailed.

Various classification systems for reporting cultural adaptation to compare interventions have been developed for research [69-71,73]. Considering the inconsistent and varied classification methods, the RECAPT (Reporting Cultural Adaptation in Psychological Trials) framework was developed by a large consortium of researchers [65]. A growing body of studies adhering to standardized reporting criteria could enhance the replicability, comparability, and transparency of adapted interventions in the field of cultural clinical psychology [74].

Objectives

In this paper, we report on (1) the assessment of needs and potential digital solutions for the mental well-being of recently arrived refugees in Switzerland and (2) the participatory development and culture-sensitive adaptation of the Sui app. Due to resource constraints, the first language version was specifically developed in Arabic, given the significant presence of Arabic-speaking refugees and asylum seekers in Switzerland. "Sui" stands for self-help, support, and information (in German: *Selbsthilfe*, *Unterstützung*, and *Information*), and the app is provided by the Swiss Red Cross (SRC).

To address the challenges of PMLDs that appear directly after arriving in Switzerland, we focused on refugees who have arrived in Switzerland less than 5 years ago. Our approach draws on participatory methodologies, which have been proven to be effective in tailoring health interventions for migrants [75]. Participation intends to establish sustainable partnerships and trust between the research and the target community. Participatory research empowers the voices of potential end users throughout the development and design process [76]. In software development, user-centered approaches are used as a form of participatory design and are characterized by iterative loops collaborating with end users [61,77]. We used this approach to deeply empathize with the needs and preferences of end users and adapt the service accordingly [61].

We plan to adapt the Sui app later to accommodate additional languages spoken by refugee communities. To approach inclusivity across different refugee backgrounds, we emphasize the importance of including context in the development and lay focus on bottom-up processes.

XSL•FO
RenderX

Methods

Overview

The Sui app was developed based on a user-centered, participatory, and mixed methods approach to provide a scalable, low-threshold MHPSS service for asylum seekers and refugees who have recently arrived in Switzerland. In the following, the term "refugees" is used inclusively to refer to both asylum seekers and individuals granted refugee status unless otherwise specified.

Below, we describe the procedure of this paper in 2 major phases: *preparation* and *development* (see Figure 1). The documentation of cultural adaptation followed the RECAPT criteria by Heim et al [65]. See Multimedia Appendix 1 for the details on Chapter A: Set-Up; Chapter B: Formative Research; Chapter C: Intervention Adaptation; and Chapter D: Measuring Outcomes. In RECAPT, formative research is conducted in an iterative fashion before and during the process of adaptation. Formative research included a short desk review, as well as a broad stakeholder consultation at different stages. Interviews were analyzed using Mayring's content analysis [78]. The documentation table, Chapter C (Intervention Adaptation) in Multimedia Appendix 1, was established from the beginning of the project to increase transparency and keep track of decisions made during the development process.

Figure 1. User-centered development process of the Sui app. RCT: randomized controlled trial; UAB: user advisory board.

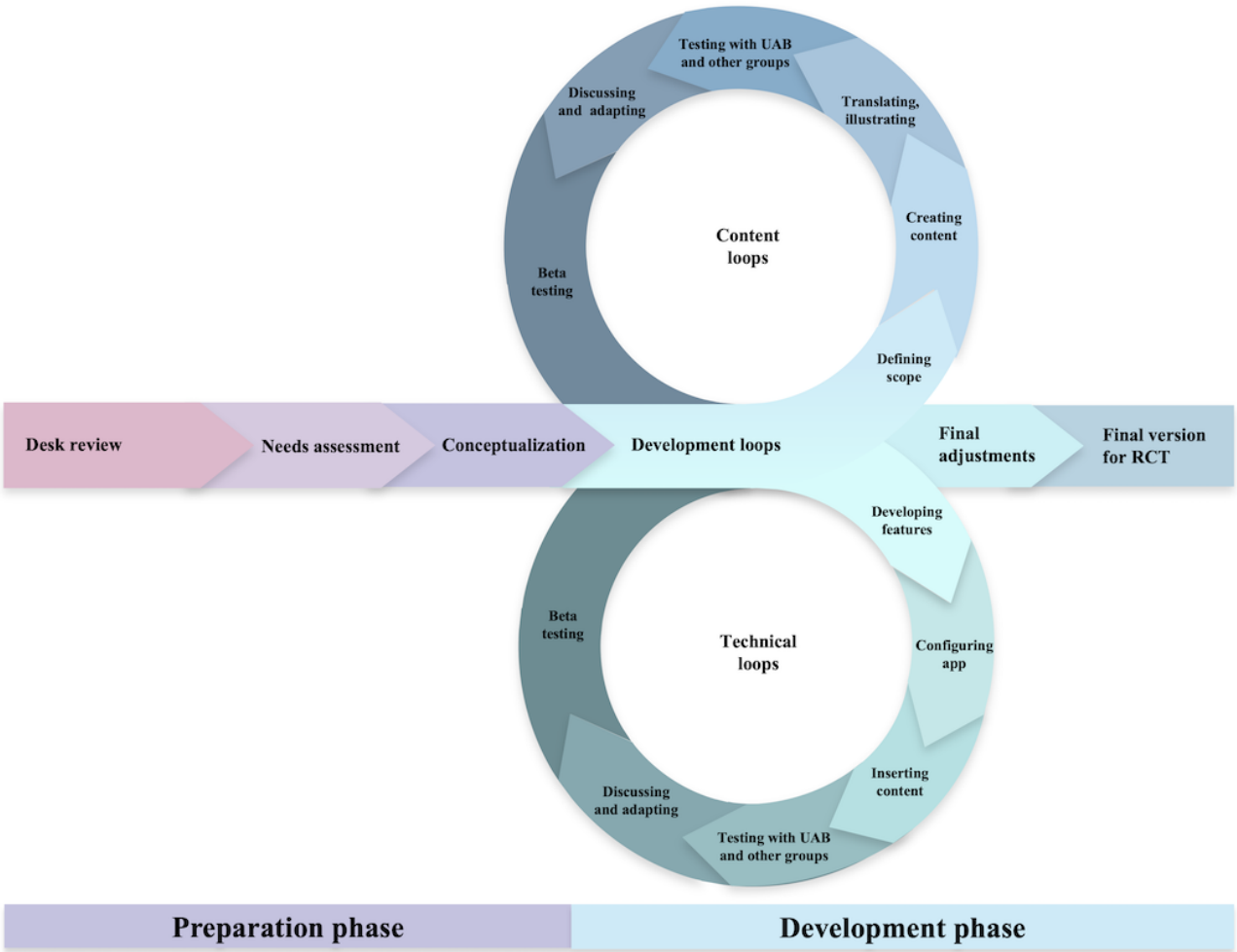

Preparation Phase

Desk Review

First, a desk review of existing literature, based on the recommendation by Greene et al [79], was conducted. The aim was to rapidly gain an overview of the mental health and care situation of and for refugees in host countries and Switzerland in particular. The purpose of this desk review was to inform the development of the interview guideline for the needs assessment and the potential benefit of a digital mental health intervention. Specific questions that were addressed included the following: (1) What is the mental health situation of refugees? (2) What is the current mental health care situation? (3) What are the barriers for refugees to accessing mental health services? (4) What are the opportunities and risks of digital mental health interventions for the care of refugees?

To gather relevant literature, the ResearchGate database was searched for review articles and meta-analyses in German and English related to the mental health of refugees. The search included terms such as "mental health refugees Switzerland," "psychosocial support asylum seekers," "cultural adaptation digital interventions," and "access to mental health care."

XSL•FO
RenderX

The search was not exhaustive but was conducted with awareness of bias risks. The retrieved articles were screened for relevance based on their ability to address one or more of the key research questions. Findings from the selected literature were summarized under key thematic insights, grouping similar results together. The summarized insights were compiled into an unpublished internal report in December 2019.

A concise version of the desk review findings is provided in Multimedia Appendix 1 under Criterion 6, where additional literature was continuously documented throughout the process. In the Results section, we included only one key example source per main insight to maintain conciseness.

Needs Assessment

The needs assessment aimed to gain insights into the psychosocial circumstances of newly arrived refugees in Switzerland (within approximately 5 years of arrival) and explore the feasibility of a potential digital care tool. For this purpose, semistructured interviews were conducted by MA, RTS, and MH with psychotherapists, social workers, marketing and innovation professionals, as well as with residential care and health care professionals working in asylum contexts (asylum care interviewees [ACIs]), recruited within the SRC and associated organizations from the SRC network. Moreover, we included interviews with intercultural interpreters (ICIs) who live in Switzerland and also work for the SRC.

Additionally, we interviewed 4 people from the first language version target group (target group interviewees [TGIs]). The aim was to gain an initial insight into their perspective without striving for data saturation, as further engagement with the target group was already planned. The TGIs were informed about the opportunity to participate by their social counselor, who works for an SRC social service. A professional ICI translated the preparatory information as well as the interviews. For details on the interview guidelines, see Criterion 10 under Chapter D: Measuring Outcomes of Multimedia Appendix 1.

The recorded interviews were processed according to the summarizing content analysis described by Mayring [78], which is suitable for recognizing central statements. Additionally, Mayring [78] encourages researchers to consider the study's macro context, including the cultural specialities of the research object. This aligns well with the culture-sensitive approach we followed with the RECAPT's framework. In preparation of a similar cultural adaptation study for a digital intervention for Albanian immigrants in Switzerland [80], Shala et al [81] conducted a qualitative analysis according to Mayring [78] to evaluate the target group's explanatory models. The needs assessment can be regarded as an explanatory model of psychosocial circumstances of refugees.

The first author (RTS), with an educational background in clinical psychology, analyzed the data. The analysis followed the subsequent five steps: (1) determination of units to be analyzed (recorded and transcribed interviews), (2) paraphrasing according to Z1-Z4 rules (reduction to essential and bundling of similar statements), (3) constructing the reduced statements into a category system, (4) re-evaluating the category system with the reduced material, and (5) finally interpreting the final categories.

To streamline and condense the content emerging from ACI interviews for this study, only categories with 5 or more codes from separate ACIs are reported with their corresponding frequencies. However, since the interviewees' backgrounds were very heterogeneous, we considered all interview data for the conceptualization (for the entire collection of categories from ACI interviews, see Criterion 6 under Chapter B in Multimedia Appendix 1). The interviews with ICIs and TGIs were less structured, and the low number of interviews does not allow for the meaningful use of frequencies. However, the codes are reported in a summarized version. The needs assessments lasted from November 2019 until March 2020.

Conceptualization

In collaboration with a software consultancy agency, we deliberated on the desk review and needs assessment findings. With their help, the core team formulated the technical requirements and scoped out the project with agreed focus areas for the planned digital service. We used iterative design thinking strategies such as creating user personas, brainstorming, paper prototyping, and testing digital mock-ups [82]. Following this, we evaluated potential approaches to software development. The meetings were not transcribed verbatim. Instead, we drafted thought processes on paper and documented the key findings in Chapter C (Intervention Adaptation) of Multimedia Appendix 1.

To further refine the conceptualization and validate concept drafts, we established an Arabic-speaking user advisory board (UAB) representing the first language version target group. Members were recruited through outreach by the SRC, local aid organizations, and personal contacts using direct invitations and flyers. An ICI (coauthor FH), a key member of the core team, played a crucial role in forming and coordinating the UAB. He also acted as an intermediary between the UAB and the research team (see Criterion 2: Teams and Roles in Multimedia Appendix 1). The UAB discussions were held in Arabic and were also not transcribed verbatim. FH translated key discussion points following small group discussions, while another core team member documented them in protocols. These records were later added to the documentation table in Chapter C (Intervention Adaptation) of Multimedia Appendix 1.

The conceptualization findings were derived from prioritized topics from the needs assessment, their alignment with desk review findings, input from the UAB, and final consensus within the core team. Decisions were made while considering technical feasibility and scalability, in collaboration with the consultancy agency. The conceptualization was conducted between March 2020 and February 2021.

Development Phase

Development Loops

In the development phase, the content and design were created according to the concept constructed during the conceptualization. The technical environment, as well as the content, including the design along with the text, was developed

XSL•FO
RenderX

in loops involving various stakeholders: the core team, the asylum-information expert group, the translation group, the UAB, the Swiss Refugee Council, social services from the SRC, mental health experts, a design agency, a copywriter agency, a team from the Freie Universität Berlin, a software development agency and selectively other people. Previous cultural adaptation processes for digital mental health interventions followed a similar iterative development process, including translation, surface, content, and delivery or technical adaptations [61,80,83].

Based on the conceptualization, the chapters' content was developed collaboratively. Each chapter was revised or partially written by respective experts (mental health experts, asylum-information experts, social services, and legal services). A professional interpreter translated the first version of each chapter, which was then revised by the translation group. This group was formed to cover understanding across various Arabic dialects. It consisted of 1 person each from Palestine, Tunisia, Egypt, Yemen, and Syria (2 women and 3 men). FH consolidated the feedback into a final translation.

The finalized translation was then revised in UAB meetings on surface, content, and technical revisions. Moreover, the UAB received homework (eg, reading texts or listening to audio exercises) to prepare for the meetings. Similar to the conceptualization meetings, development loop meetings for each chapter of the Sui app were conducted in Arabic, both in full-group and small-group discussions. Core statements of each small group were reported and translated at the end of the meetings by FH. These summaries were documented in protocols and later added to the Intervention Adaptation documentation table in Chapter C of Multimedia Appendix 1. The development loops took place between April 2021 and June 2022.

Finalization

In a beta test, participants completed an initial user experience (UX) test in a face-to-face meeting, tested the app for 4 weeks, and then took part in a face-to-face semistructured interview about the Sui app and optimal peer support at the end. The participants were recruited with the help of a local social counseling organization and the extended UAB and SRC networks. They had no prior involvement in developing the Sui app.

To evaluate how a contact person behind the app could be advantageous, a first concept for digital peer support was tested in the beta test with the aim of further developing it for the subsequent evaluation. This rough concept was based on the experiences of previous scalable interventions for refugees, such as Step-by-Step [84], Doing What Matters in Times of Stress [85], PM+ [86], and other internet-based treatments for depression and loneliness [87,88].

The SRC trained 2 Arabic-speaking people (beta test peers) with a migration background to provide a weekly message exchange with the app users. They were recruited following personal recommendations by the UAB. The training included getting familiar with the content of the Sui app and practicing writing support messages with standardized text templates, thereby using skills like active listening, paraphrasing, validating suffering, normalizing psychological symptoms of stress, and suggesting adequate subchapters of the app. They were also introduced to monitoring use behavior (completed subchapters) and activity (use time). Additionally, they had a supervision exchange with an Arabic-speaking psychotherapist. Findings from the beta test were used to understand the content and technical processes, develop a sensible structure for the motivational messages, understand the technological use in practice, evaluate the amount of support, and get an idea of what reactions to expect from participants.

The UX test, as well as the qualitative interviews, was conducted by 2 master's students and coauthors (MH and BJF) and interpreted by FH from Arabic to German simultaneously. The beta test peers were interviewed in German, also with a semistructured interview at the end of the test period. For details on the UX test and interview guidelines, see Criterion 10 under Chapter D: Measuring Outcomes in Multimedia Appendix 1. The interviews were transcribed, and analyses were conducted following Mayring's [78] content analysis with the same process as described for the needs assessment.

For the finalization, decisions and adjustments were applied based on the findings from the beta test. The peer concept was revised; technical and content app changes were made. The app was prepared to be used in a subsequent quantitative evaluation testing the effectiveness of Sui as a peer-guided or unguided service against a waitlist control group. The beta tests and the finalization lasted from July until October 2022.

Ethical Considerations

Participants in the needs assessment (ACIs, ICIs, and TGIs) were invited to interviews verbally or in writing. They were given a short explanation of the planned project on digital mental health for trauma-affected refugees from all backgrounds living in Switzerland. Upon agreement, an interview was scheduled, and verbal consent was obtained for participation and audio recording, with recordings deleted poststudy. TGIs received a CHF 50 (approx US $55) supermarket voucher, while ACIs and ICIs were not incentivized. Ethical review was not sought, as participation was voluntary and noncritical.

For the conceptualization and development loops, UAB participants provided informed consent in Arabic, agreeing to the use of their summarized feedback in this paper. They received symbolic compensation of CHF 100 (approx US $110) supermarket vouchers. This research was approved by the local ethics committee of the Faculty of Human Sciences of the University of Bern in Switzerland (2021-09-00003).

Before the beta test, participants received study information in Arabic and sufficient time for review. They provided written consent for participation and audio recordings, which were deleted poststudy. The beta test was approved by the same local ethics committee (2022-06-00001) and conducted from June to August 2022.

All data remain fully anonymous, with no names, initials, or acronyms appearing in this paper or multimedia appendices.

XSL•FO
RenderX

Results

Preparation Phase

Desk Review

The desk review yielded 5 valuable key insights that contributed to shaping the further research questions. These were subsequently discussed in interviews with ACIs, ICIs and TGIs.

First, mental health issues often seem to carry a social stigma, leading many refugees to avoid seeking help due to fears of social consequences or beliefs that their condition is untreatable [89]. Addressing this stigma could potentially enhance the effectiveness of mental health interventions aimed at this population. Second, refugees exhibit diverse health literacies, understandings of illness, help-seeking behaviors, and treatment needs [90]. This suggests that mental health interventions may be most effective when tailored to align with the beliefs, values, and needs of the recipients. Third, interventions for refugees may need to extend beyond solely addressing mental health symptoms to consider the broader context of their living conditions [15]. It is conceivable that addressing only symptoms may not necessarily lead to an overall improvement in the quality of life for refugees. Fourth, the effectiveness of digital mental health interventions seems to hinge significantly on participants' adherence and compliance, which are influenced by perceived benefits [91]. Thus, implementing various strategies to enhance adherence could have a positive impact on the effectiveness of interventions. Fifth, adopting a task-shifting approach shows promise for alleviating stress symptoms and improving refugees' psychosocial well-being [55].

Moving forward, it was important to further explore the implications of these findings for implementing a digital mental health service aimed at improving the care situation of refugees in Switzerland, considering realistic opportunities and risks associated with such projects.

Needs Assessment

Findings of ACI Interviews

The results from the needs assessment with 22 ACI interviews are presented in 2 sections. First, the prevailing status characterizing the psychological state of refugees in Switzerland is described. Table 1 lists the psychosocial problems of refugees living in Switzerland, as identified by ACIs. Second, the potential of a digital tool to address the existing issues is elaborated. Table 2 illustrates the primary aspects highlighted by ACIs regarding the potential of a digital psychosocial support service. Subsequently, the summarized results from interviews with 2 ICIs (1 from Syria and 1 from Lebanon) and 4 TGIs (2 Syrians and 2 Syrian or Palestinians) are reported. Each interview lasted for approximately 60 minutes.

Table 1. Results on the current psychosocial problem situation in refugees residing in Switzerland: frequency of coding references (≥5) from the needs analysis found in semistructured interviews with asylum care interviewees working with refugees in psychological, medical, social, and residential care (N=22).

Categories	ACIs[a], n (%)
Main psychosocial problems	
Sleep disturbances	9 (41)
Asylum situation (decision, rights, family reunification, and traveling)	9 (41)
Physical pain or tension as an expression of psychological burden	8 (36)
Occupational integration	8 (36)
Social inclusion	6 (27)
Housing	5 (23)
Obstacles to healing	
Asylum situation	13 (59)
Feelings of shame	7 (32)
Fear of being categorized as "crazy"	6 (27)
Resources	
Often strong resilience	11 (50)
Daily structure and activities	8 (36)
Social network	6 (27)
Feeling of being needed (identity)	6 (27)

[a]ACIs: asylum care interviewees.

XSL•FO
RenderX

Table 2. Results on the potential of a digital psychosocial support service for refugees living in Switzerland: frequency of coding references (≥5) from the needs analysis found in semistructured interviews with asylum care interviewees working with refugees in psychological, medical, social, and residential care (N=22).

Categories	ACIs[a], n (%)
Concerns and obstacles toward digital tools	
No relationship and trust building	7 (32)
Lacking sensitivity of the digital tool toward emergencies and triggers	5 (23)
Lack of individuality	5 (23)
Potential use of a digital service	
Bridging tool (preparation, filling)	13 (59)
Support for ongoing therapy	9 (41)
Most people have a smartphone, and digitalization is future-oriented	8 (36)
Contact person behind the tool would be advantageous	8 (36)
No replacement for existing face-to-face services	7 (32)
Tool as a source for professionals and relatives of the affected	6 (27)
Emergency plan (advice)	5 (23)
Stabilization as primary goal	5 (23)
Digitizable content	
Psychoeducation	16 (73)
Psychoeducation on PTSD[b] symptoms	8 (36)
Information, explanations on sociostructural everyday aspects, and integration	14 (64)
Housing	5 (23)
(Mental) health system	5 (23)
Occupational or educational integration	5 (23)
Standardized exercises	
Body focused	14 (64)
Activation of existing resources	12 (55)
Connection to social life and activities	9 (41)
Asylum-related legal information	7 (32)
Important technical aspects to consider	
Balance between tailored and generic use	9 (41)
Anonymity: as priority and possibility to reach people	9 (41)
Advice on the delivery formats	
Simple design and diverse formats	15 (68)
Images and illustrations	7 (32)
Example stories	7 (32)
Short texts	5 (23)
Videos	5 (23)
Advice for psychological language within the tool	
Resource-oriented psychological language: normalizing, empowering, validating, taking seriously, hopeful, patient, humorous, listening, positive, and asking	10 (45)
Avoid stigmatized terms	7 (32)
Advice for better accessibility	
Participation of the target group in development and dissemination	9 (41)
Access through symptomatology	8 (36)

XSL•FO
RenderX

Categories	ACIs[a], n (%)
Access to information about everyday (asylum) life	6 (27)

[a]ACIs: asylum care interviewees.

[b]PTSD: posttraumatic stress disorder.

Psychosocial Problems Reported by ACIs

In the interviews, the most prominent psychosocial challenges mentioned to be occurring in refugees were sleep disturbances, the overall asylum situation (including the asylum decision, process, and rights), physical pain or tension, occupational integration, social inclusion, and housing. Physical pain was frequently reported to serve as an expression of psychological burden in the target population.

Apart from being a psychosocial challenge, the asylum situation was mentioned as one main hindering factor for refugees to start a healing process or seek treatment since the focus is often on the structurally problematic situation. Other than that, feelings of shame and fear of being categorized or stigmatized as “crazy” were mentioned to be obstacles for people to express the need for help or seek treatment.

ACIs frequently mentioned that they perceive a resilience in many refugees that is difficult to describe but serves as a vital resource. In addition, a daily structure and activities, a social network and the feeling of being needed were mentioned as further concrete resources that contribute to improving the well-being of refugees.

Potential of a Digital Tool Evaluated by ACIs

Concerns of ACIs were that a digital tool could not build up a relationship or trust, which they perceive as essential in traditional psychotherapy. Another concern was that a tool could not detect emergencies or individual triggers and not react adequately. Similarly, they stressed that a digital tool cannot address individuality (eg, individual questions) properly.

However, a digital service was seen to be useful in approaching the present care gap. “The potential of the digital tool can lie in addressing the existing treatment gap, acting as a bridging, or filling additional service,” one ACI said. It was also envisioned as a tool supporting ongoing therapy, allowing clients to repeat information or exercises. In any case, a contact person behind the tool was considered advantageous for the users. According to the ACIs, a primary goal of the tool should be mental stabilization. Moreover, a digital tool was perceived as future-oriented, as most people own a smartphone. It was also recognized that the tool could benefit not only the target group but also professionals and relatives seeking information on psychosocial support for those affected. Nevertheless, it was emphasized that the digital tool cannot replace traditional face-to-face services.

Regarding the tool’s content, consistent suggestions were psychoeducation, particularly on symptoms related to PTSD. Typical psychotherapy exercises, especially body-focused exercises, were also deemed digitizable and effective. ACIs highlighted the importance of activating the individual’s existing resources (eg, asking about personal strengths). Along with psychological content, many ACIs suggested including sociostructural information typically provided by social services. Those include information about housing (eg, how to find an appropriate apartment), about the Swiss (mental) health system (eg, rights, existing services, and access), and occupational integration (eg, how to find a job or start an education). Additionally, a connection to “real” social life (eg, lists of communities, social clubs, and free activities) was considered indispensable. Besides providing information on everyday life, ACIs suggested integrating asylum-related legal information (eg, differences in residency status and family reunification).

Technical considerations of the tool involved striking a balance between generic use and individual tailoring. Anonymity was said to be paramount to ensure safety and to mitigate the fear of stigmatization toward psychological problems. Thanks to this anonymity, more people could be reached.

A consensus emerged on delivering content in a concise and straightforward manner, incorporating various formats, such as images, illustrations, example stories, short texts, and videos. It was advised that communication should adopt a resource-oriented approach: normalizing, empowering, validating, taking seriously, being hopeful, patient, humorous, listening, being positive, and asking. It was also stressed that the use of often stigmatized terms (eg, “patient” and “psychological disease”) should be avoided, even though they can vary across cultures.

Collaboration with the target group in both development and dissemination was strongly advocated. Furthermore, ACIs advised focusing on individual psychological symptoms as an entry point to the tool. Providing information on everyday life challenges that typically occur after migration can serve as another entry point to the tool.

Findings of ICI Interviews

The 2 ICIs reported similar psychological issues and PMLDs as the ACIs did. The reported problems included fear of being stigmatized as “crazy,” sleep disturbances, general mistrust, a lack of awareness regarding treatment options, worries about family left back home, residency status, and occupational integration challenges. One ICI explained that mental disorders are perceived as intrinsic to life and require time to pass. The other ICI stated that refugee traumatization should not be pathologized as a mental illness but rather viewed as adverse experiences that require strategies and possibly medication. Stigmatization, linguistic barriers, and difficulties in occupational and cultural integration were identified as obstacles to healing, while rapid networking within refugee communities and individual initiatives were recognized as resources. A digital support service should incorporate sociostructural assistance, such as links and addresses to social support resources, success stories, information on family reunification and other asylum-related topics, and general challenges encountered when

XSL•FO
RenderX

arriving in Switzerland. Psychological content, including "dealing with mental disorders," breathing exercises and their efficacy, and self-help techniques for crises, was also highlighted. The primary recommendation emphasized the significance of ensuring anonymity.

Findings of TGI Interviews

The main additional findings from the 4 interviews with TGIs were that they all use smartphones regularly, mainly to maintain contact with relatives, translate daily tasks, and access several consumer apps. Concerning psychological issues, all TGIs expressed that they will never be able to forget what they have survived and voiced concerns about the well-being of their family members left behind and have tried to reunite them in Switzerland. Some TGIs spoke about physical concerns that have occurred since their migration, alongside psychological challenges, such as stress, guilt, and shame. The encountered PMLDs among TGIs were as diverse as described by the ACIs, encompassing difficulties with language acquisition, residency status, occupational integration, waiting times, experienced discrimination, and navigating Swiss legal frameworks. Despite these challenges, all TGIs were able to identify positive aspects of their lives in Switzerland, citing factors like overall health, natural surroundings, happy people, tranquility, and being treated well. Reported resources included being grateful, having religious faith, fulfilling parental duties, ongoing efforts to reunite with family members, and receiving support from friends, family, or social counselors. Regarding the idea of a new support app, their imagination was limited to the knowledge of existing consumer apps and leisure activities. Consequently, they were rather unclear about the potential benefits of such a digital service.

Conceptualization

Overview

The UAB (n=9) participating in the conceptualization and subsequent development loops consisted of 8 participants and the key intermediary (FH) from Syria. The group was intentionally diverse in age (aged 19-55 years), ethnicity (4 Kurdish and 5 Arabic), educational background, and gender balance (4 women and 5 men). Four conceptualization meetings were conducted with the UAB, each lasting approximately 180 minutes.

The results from the conceptualization include decisions on technological aspects, the scope of the content, and its delivery formats. Additionally, a rough concept of peer-to-peer support to be used in the app is presented.

Technology

The digital service was decided to be delivered in the form of a cross-platform app that behaves like a native app, can be downloaded on iOS and Android devices, and is easily installable and updatable. At the time of making this decision, compared to progressive web apps, this format was seen as more accessible, robust, and better applicable on iOS devices for installation and receiving push notifications. Thanks to the compatible requirements, a mutually beneficial collaboration was agreed upon with the Department of Clinical Psychological Intervention at Freie Universität Berlin: Their "DIRECT" software can be used to create custom apps that can be adapted to the respective needs of the project and are specifically suitable for research in the field of mental health [92].

Scope of the Content

Regarding the content, the project team decided to combine sociostructural information with simple psychological well-being tools. The importance of focusing on sociostructural challenges in the asylum context was a key finding from the needs assessment (see above). Therefore, 9 chapters to provide sociostructural information and 5 psychological chapters were identified. The development of those chapters is described below. Following the recommendations of the ACIs, additional explanatory chapters were developed: national emergency contacts, introduction to the app, feedback on the app, and introduction of the storyline (for an overview, see Textbox 1). All content is made available directly after account setup since the target group and their needs are heterogeneous.

XSL•FO
RenderX

Textbox 1. Overview of the content of the Sui app. The Sui app is divided into 2 main sets of chapters, "Information" and "Well-Being," and contains an additional "General" set.

Information chapters

- Asylum Process
- Family Reunification
- Finances
- Health Promotion
- Health System
- Housing
- Residence Status
- Social Living
- Work and Education

Well-being chapters

- Stress
- Sleep
- Resources
- Chronic Pain
- Emotion Regulation
- Audio Exercises

General chapters

- Emergency Information
- Feedback
- Introduction
- Sui's Neighborhood

Delivery Formats

Emerging from the needs assessment's results, various delivery methods were decided to be used: simple texts, illustrations, stories, animations, explanation videos, video testimonials of refugees talking about personal experiences, audio exercises, list exercises, details of contact centers, and links to websites. We decided to create fictional characters to tell stories that our target group can identify with.

Peer Support

Due to significant concerns regarding security complexity, including the structured moderation of messages, we decided to abstain from implementing a feature that allows direct communication between users (eg, forum exchange). Nonetheless, the recommendation for a designated contact person within the app was strong. Given the demonstrated efficacy of guided formats in previous internet-based mental health intervention studies, we chose to integrate a guided version of the app. Furthermore, drawing insights from experience with prior peer-to-peer programs, we capitalized on the benefits of providing support through individuals sharing similar migration backgrounds with the target group (such as a shared language, cultural backgrounds, life experiences, and current challenges).

Development Phase

Development Loops

Overview

The previously formed UAB (see the Conceptualization section) was involved in 6 meetings for design and surface discussions (prototyping and voting) and 13 meetings for content revisions (feedback and inputs on each chapter). Technical tests were conducted concurrently with content evaluations. The meetings lasted for an average of 210 minutes.

The reported results from the development loops only include a summary of the findings from the iterative development process. In the following, the results of this iterative adaptation process for design, language, and each chapter of the Sui app are described. Refer to Chapter C (Criteria 7, 8, and 9) in Multimedia Appendix 1 for a comprehensive description of the Sui app's content, surface, and the associated adaptation and decision-making process. Over time, we divided the UAB into smaller groups so that discussions could be held in their native language (Kurdish or Arabic) and at different paces.

Illustrations

The bird "Sui" was introduced as the mascot of the app. Its name and color were accepted by all expert groups and the

XSL•FO
RenderX

UAB, and it was preferred over a customizable avatar. The storyline around the app's content introduces 12 fictional protagonists living together in an apartment block in Switzerland. They are from diverse sociodemographic backgrounds and countries, not only Arabic-speaking, to expand the app to other target groups at a later date. Their representation was evaluated by the UAB and by representatives of each of the characters' ethnic groups. Each image was revised by the UAB, along with the corresponding content. Criticized features were adapted by the design agency and included, for example, lighter and more diverse skin colors, the use of a hijab, hairstyles, generally happier faces, change of clothes, or change of names.

Language

Originally produced in German, the written content was proofread by a Swiss copywriter agency to ensure simplicity and adherence to "plain language" principles. This involved minimizing complexity and shortening sentences, consistent with the recommendations from expert interviews (as reported above). We also found that the use of plain language facilitates subsequent translation and reading comprehension for people with low levels of education. Technical terms are explained in the glossary of the app and as a tooltip feature within the text. Arabic is an official language in more than 20 countries, including many countries that face political conflict and economic crises that often lead to the displacement of people. We, therefore, wanted to use an understandable version devoid of specific regional dialects. The translation group opted for a "Levantine Arabic+" that would be widely comprehensible across the Arabic-speaking population. We supplemented the initial Middle Eastern Levantine translation, done by a professional translator, with additional synonymous words suggested by the translation group. This approach aimed to achieve cross-national understanding. The UAB then made revisions and suggested minor changes. The intercultural translator integrated all feedback into a final version, focusing on an analogous translation and following the same principle as in the German version of offering the simplest possible language. This multilevel translation process was necessary to allow the UAB to understand every sentence and focus on the content rather than the language specifics. In addition, after UAB discussions, we agreed to add both a male and female Arabic version, as the second person pronoun ("you") in Arabic distinguishes between two genders. The preferred form of address is set during account creation in the app.

Audio Exercises

Audio Exercises, featured in all psychological chapters and consolidated in one overview section, are drawn from several sources. However, most stem from the trauma-sensitive yoga founded by trauma therapist Härle [93-95]. The UAB deemed all exercises appropriate but generally preferred short durations. They particularly favored breathing exercises. Two voices, male and female, were chosen through a voting process involving the UAB and the translation group. Both female and male users hear both voices equally often. However, Arabic speakers will be addressed in either the female or male version of "you" based on preferences set during account creation.

Information Chapters

The 9 chapters on sociostructural information were developed based on existing first information material in collaboration with respective professionals, such as SRC social services, the Swiss Refugee Council, and an additional asylum expert group. The UAB revised each chapter. Because of the federal Swiss system, the depth of information was restricted to a national level since providing detailed information for each of the 26 cantons would have been impossible to implement simply. The chapters consist mainly of text-based step-by-step explanations but also include illustrations, video testimonies from refugees, and explanation videos.

The main results from the UAB meetings were as follows: housing covers everything important about the topic and includes valuable templates. However, they suggested incorporating more elements that convey hope for success into the content. For the "Work and Education" chapter, recommendations included adding more information on the education system, a job consultancy tool, a job platform, and instructions for CVs and application letters. Regarding the 3 asylum chapters, "Asylum Process," "Residence Status," and "Family Reunification," they emphasized that highlighting accuracy due to prevalent misinformation is important. "Social Living" was deemed clear and important, with suggestions to list free activities, new events, and information about language courses. The "Health System" chapter should clarify where people can turn to for specific, how they can access care, and provide information on pregnancy and contraception equally for men and women. The chapter "Finances" was first considered too complex because several systems were not known in the UAB (eg, 3-pillar system, old-age and survivors' insurance). It was therefore expanded with information and was rated important, with suggestions for including money-saving ideas and differentiating between asylum social assistance and general social assistance. In the "Health Promotion" chapter, the UAB underscored the importance of avoiding any implications suggesting universal tobacco or alcohol consumption among refugees. Furthermore, they made recommendations on how healthy nutrition is possible despite a low budget.

Well-Being Chapters

The 5 chapters on psychological well-being were developed based on existing materials described below and adapted by the first author (RTS) according to the recommendations from the needs assessment. They were then reviewed by psychotherapists and researchers in the field of clinical psychology before the UAB gave feedback. The chapters consist of psychoeducation, written exercises, audio exercises, illustrations, and video testimonies from refugees.

The principal outcomes of the UAB revisions to each chapter were as follows: "Stress," partly inspired by the DWM self-help guide [96,97], was considered important, mainly to understand the sources, to distract from, and maybe speak to someone about the experienced stress. "Sleep," adapted from the content of a self-help book developed for traumatized refugees [98], includes sleep hygiene tips, body-focused exercises, and an introduction to sleep rituals. The UAB criticized using only

XSL•FO
RenderX

smartphone-based exercises and recommended integrating exercises that can be carried out without a smartphone. The chapter "Resources" integrates planning activities, gratitude journaling, and strength activation, primarily inspired by the self-help book mentioned above [98]. It also features 2 subchapters on identifying values and staying present, derived from the DWM self-help guide [97]. Although the UAB deemed this chapter "not very important," they expressed overall appreciation for its content and expanded the list of activities. "Emotion Regulation" aims to manage intense emotions using the "problem-solving therapy" by Nezu et al [99], which was tested in an e–mental health intervention for depression [100]. The shortened and adapted version of the Sui app focuses only on the first 2 noncognitive steps of "stopping" and "slowing down" before acting. It uses a traffic stoplight metaphor and presents 3 case examples of Sui's neighborhood residents, whose relatable everyday life situations were appreciated by the UAB. For the "Chronic Pain" chapter, based on 2 self-help books [101,102], the UAB advised changing the psychoeducational texts into shorter and longer versions and rated the exercises as recommendable.

Finalization

Beta Test Setting

The Sui app was tested by 9 beta test participants, all of whom completed both the UX test and posttest qualitative interview. The group included 5 women and 4 men from Syria (mean age 28.44, SD 10.57 years), who had been living in Switzerland for 1 to 5 years. Of the 9 beta test participants, 5 were randomly assigned to receive weekly peer support during the testing period. The test peers were male and female (n=2, aged 29 and 33 years) and refugees in Switzerland coming from Syria and Egypt.

The results from the beta test, including the UX tests as well as the interviews, are summarized in the following. The UX tests averaged 45 minutes, whereas the interviews averaged approximately 60 minutes.

App Satisfaction

Most of the participants used the app several times a week and said that they were motivated to use the app. The functionality and structure of the app were unanimously rated as logical and simple. However, all participants complained about slow loading times during the interview and UX test. This led to reported problems accessing content or links within the app and app crashes. Most participants rated the illustrations as beautiful, appropriate, and compatible with the content. The illustrated protagonists were rated to be realistic, and participants could visually relate to them. Most participants rated the Arabic language as clear and understandable.

Half of the participants highlighted the chapters "Social Living," "Housing," and "Health System" to be helpful. Some participants rated "Sleep" to be good, particularly the exercises for falling asleep. However, some rated the included sleep hygiene tips to be superficial. Additionally, "Stress" and "Chronic Pain" were perceived as useful. Most participants appreciated the audio exercises, while a few had not used them. Some participants wished for more links, concrete realization examples, or more sociostructural chapters, for example, on drugs, forced marriage, or old-age and survivors' insurance.

Satisfaction With Peer Support

With regard to peer support, beta test participants desired a quick response rate in the chat (6-72 hours). The messages should be formulated professionally but also in everyday language and adapted to the user. Support was also requested for technical issues. Some said that they would like to know some basic facts about the peer (eg, age, country of origin, and length of time in Switzerland) to be sure that they speak with a human and not with a computer. Having a peer companion was reported to enhance to motivation to use the app.

The test peers agreed that a 1-week response time was too long and should be shortened. They reported that, particularly for technical inquiries, they would suggest a response time of 48 hours. One of the test peers stated that most questions revolved around technical issues. Both stated that they would have preferred to write the messages more individually, and the use of text templates should be reduced, to provide the feeling of being a real person instead of automatic answers. The prior training was valued by both peers and rated to be important as a preparation.

Additionally, both agreed that continuous supervision and feedback on their work was important to seek help and reduce fears. Both highlighted that their messages motivated the users to use the app, and both tried to show their assigned participants that they are not alone.

Final Adjustments

Final adjustments on technical features and content were applied according to these beta test results before releasing a version for a subsequent evaluation. Loading times were improved, and the introductory chapter was expanded with explanations of features that were not self-explanatory (eg, explanation of the chat, the chapter sets "Information" and "Well-Being," and the "Audio Exercises" chapter). For the peer concept, the peers' general response time was reduced to 48 hours, and more intensive support for the peers regarding questions was introduced. The text templates were not reduced but rather complemented and rephrased to sound less "robotic," as they were to remain standardized for comparability in the trial. The peers' first messages will provide a more personal yet anonymous introduction, and peers will include their observations from the user behavior monitoring in their messages. Figure 2 displays screenshots of the final product.

XSL•FO
RenderX

Figure 2. Screenshots of the final Sui app in German and Arabic.

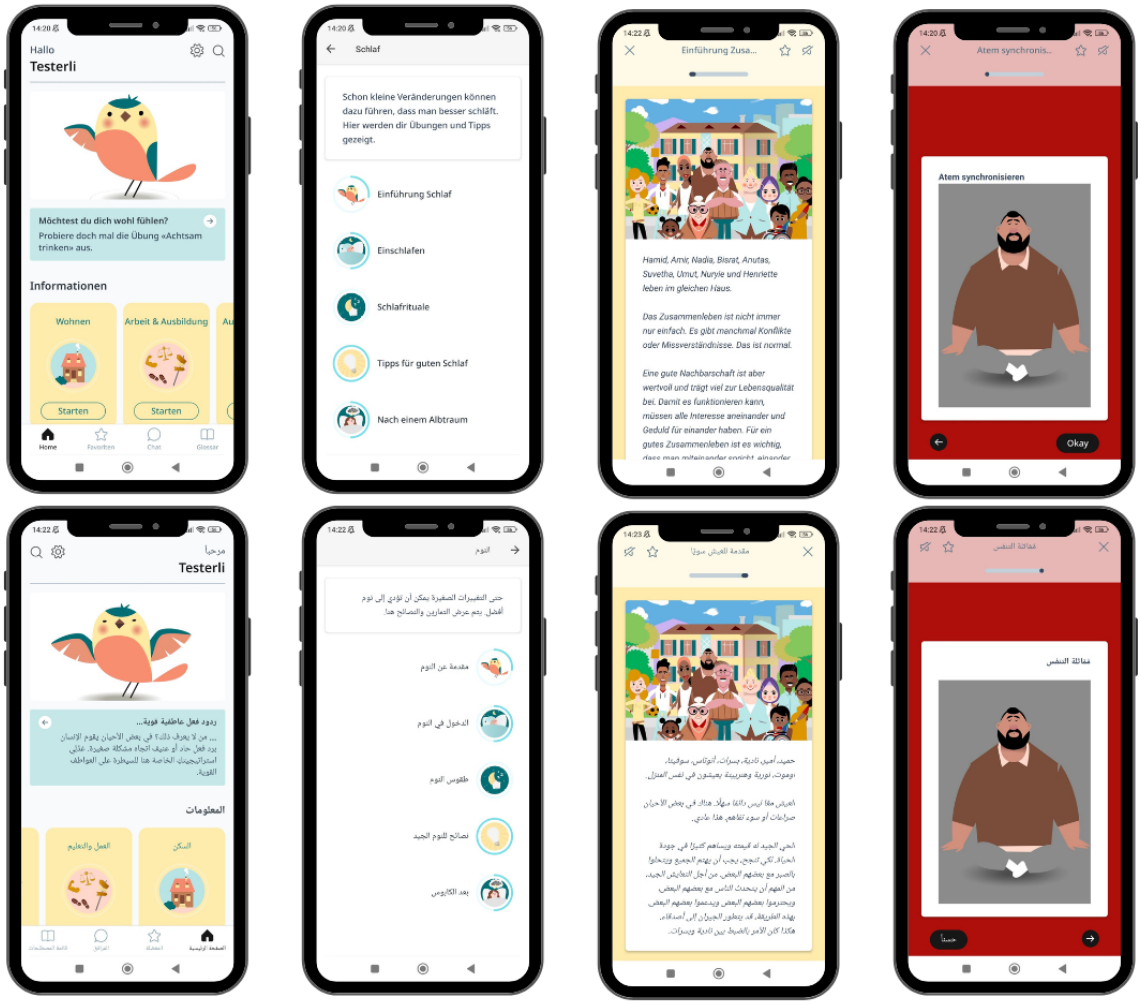

Discussion

Principal Findings

In line with our objectives of (1) assessing the psychosocial needs of newly arrived refugees and identifying potential digital solutions and (2) developing and culturally adapting the Sui app in a participatory manner, we proceeded in two phases. During the preparation phase, a brief desk review and semistructured interviews revealed that newly arrived refugees in Switzerland face both psychological challenges (eg, sleep disturbances, stigma, and emotional distress) and sociostructural barriers (eg, navigating asylum procedure, securing housing, and occupational integration). These insights informed a first concept draft of the app. During the development phase, iterative consultations with Arabic-speaking advisors and expert stakeholders followed, resulting in the Sui app's structure: 9 sociostructural chapters, 5 psychological chapters, and 4 general chapters. During these iterations, each chapter was refined (eg, simplified, reduced, and supplemented with examples), while content was translated into "Levantine Arabic+" with male or female pronoun options. Illustrations and stories, featuring the bird "Sui" as a mascot and 12 diverse characters, were added alongside audio exercises and video testimonials. Following a 4-week beta test involving 9 Arabic-speaking participants (5 of whom received peer support), the app was found to be highly acceptable: most participants used the app multiple times per week, particularly the chapters "Social Living," "Housing," and "Health System," and they appreciated the peer-guided chat messages. However, slow loading times limited full engagement. We improved technical performance, added explanatory texts, and optimized peer support through faster responses and personalized messages, resulting in a more user-friendly Sui app ready for further evaluation.

Our finding that both psychological distress and sociostructural barriers coexist is consistent with prior research showing that PMLDs often exacerbate mental health symptoms beyond premigration trauma alone [13-16,23-25,89]. By embedding sociostructural alongside psychological chapters, we aligned the Sui app with the ADAPT model. This model posits that disrupted safety and security, bonds and networks, justice, roles and identities, and existential meaning interact to exacerbate distress, underscoring the need to address both psychological and sociostructural domains concurrently [16]. Furthermore, this consideration of context follows the recommendation that cultural adaptation should encompass not only background factors such as language or traumatic events, but also current concerns and stressors [61,103]. In contrast, existing smartphone-based MHPSS tools for refugees have focused predominantly on symptom reduction without integrating PMLD-related content [104].

XSL•FO
RenderX

Adhering to RECAPT guidelines [65], we combined surface adaptations with deep-structure adaptations, as has been done in other recent digital mental health interventions for immigrants [80,83,105]. The early integration of stakeholders across both preparation and development phases corresponds to a bottom-up and user-centered approach, which likely enhanced content and surface acceptability. This approach is supported by a review and guidelines on user engagement in digital mental health [45,106]. The well-perceived peer-guided chat aligns with broader global mental health strategies promoting task-shifting and human support in digital interventions [55,62], and with the successful implementation of other peer-provided mental health programs [39-41]. Participants' preference for more personalized messages suggests that the standardized peer support should allow for flexibility to preserve authenticity while maintaining quality and consistency. Future research should examine the extent of personalization required to optimize user engagement and acceptability. Technical constraints (eg, slow loading times) were noted to be one of the main barriers to engagement in digital mental health studies [107], underscoring the value of beta testing predeployment. The intensive multistakeholder process required substantial time, financial resources, and coordination, echoing Shala et al [80]. However, the dual-phase design (preparation and development phase) created a modular core that should streamline future adaptation to other refugee language groups with similar needs in Switzerland, enhancing the potential scalability.

Limitations

Various limitations of this study have to be considered. The focus on the participatory approach was placed on the development of the service, whereby the science-specific parts were not involved. These would have included the understanding of study information, the self-assessments, or the recruitment materials.

In the preparation phase, only 4 representatives of the target group were interviewed. As a result, we only had a limited subjective view of refugees at the beginning. It could have been advantageous to include a bigger population of TGIs in the preparation phase to receive a more subjective impression of the refugee experience in Switzerland. Additionally, the focus on digital support solutions in the needs assessment interviews may have influenced responses. Due to these reasons, its findings should not be generalized, but rather should be seen as initial insights.

Another limitation was that in initial UAB meetings, discussions were translated simultaneously, which led to a fragmented flow of conversation and, thereby, frustration among the participants. This process was changed as soon as concrete drafts of the content chapters were available. From then on, discussions were held in smaller groups, in Arabic or Kurdish, and each group presented the main discussion points and remarks in summary. The documentation of the discussions was, therefore, limited to these summaries in the end. A further limitation was that during the development loop discussions, some of the chapters were discussed multiple times with the UAB, whereas other chapters were only revised once due to time limitations.

Additionally, the beta test faced technical malfunctions, including long loading times and nonfunctional push notifications, such as those intended to indicate new messages from the test peer. Given a tight time plan, these difficulties could not be resolved in advance, significantly affecting the beta test results, as participants struggled to engage with the app's content and primarily asked technical questions during their interactions with peers. This potentially prevented them from addressing other relevant issues. Furthermore, we did not assess quantitative metrics in the beta test, which limits our understanding of engagement patterns and barriers to adoption, which future studies should address by integrating a detailed tracking.

Another limitation was that the TGIs, the UAB, and the beta test included only participants from Syria, although we aimed to adapt the app for Arabic-speaking refugees from various countries. While different nationalities were included in the translation process to ensure broad linguistic accessibility, a more diverse participant group for user feedback would have strengthened the representativeness of adaptation needs.

While we used Mayring's [78] qualitative content analysis to structure the data, thematic analysis might have provided a more flexible framework for capturing emerging themes and patterns across discussions. Future studies may consider applying thematic analysis to allow for a more interpretive approach when analyzing participatory data, also considering that there exists a larger amount of literature on thematic analysis with refugees [90].

These limitations could have led to a bias in several interpretations and decisions made during the development process.

Conclusions and Implications

Forced displacement disrupts both psychological well-being and sociostructural stability. By addressing both psychological and sociostructural challenges, the Sui app represents a novel approach to digital psychosocial support for newly arrived refugees. Its task-shifting approach, that is, training peers to provide support via chat, can further improve engagement and effectiveness. Our iterative, participatory process ensured that content was not only linguistically accessible but also strongly responsive to refugees' everyday realities and cultural understanding. Despite encountering challenges such as technical issues during the beta test, our findings affirm the potential of digital MHPSS services and the use of trained peers in bridging care gaps for marginalized populations. By thoroughly documenting our adaptation process, this study strengthens the evidence base for culturally adapting psychological interventions [65].

Based on our findings, several implications arise. First, digital MHPSS tools for refugees should include sociostructural information alongside psychological modules, as this dual approach concurrently better reflects refugees' realities and enhances cultural relevance. Second, engaging target group representatives and other stakeholders from project inception through development ensures that content, language, and delivery format remain acceptable. Funders and implementers

XSL•FO
RenderX

should allocate time and resources specifically for this multilevel adaptation. Third, training nonclinical peers to provide culturally appropriate support via chat can improve acceptability. Future research should examine which aspects of peer support most strongly drive positive outcomes. Finally, technical robustness must be prioritized through systematic beta testing and ongoing maintenance to optimize user engagement.

Acknowledgments

The project around the Sui app is multidisciplinary and has included multiple stakeholders who have essentially contributed to the development of the app and the subsequent quantitative evaluation.

The authors would like to thank all advisory groups, including the experts on asylum-related topics, the Swiss Refugee Council, the Swiss Red Cross social services and psychology experts involved, the linguistic experts, the software consultancy agency, the design agency, and last but not least, our immensely valuable user advisory board. We would also like to thank our beta test peer team, which consisted of Pirko Selmo as supervisor and DH (who does not wish to use their full name) and Zaher Al Abbar as test peers. Our thanks go to all the providers of open-access resources for refugees on which we could base our psychological content. Furthermore, we are grateful to Dagmar Härle, who kindly provided most of the audio exercises from her trauma-sensitive yoga approach and helped create a color-coded concept for their implementation. In the context of the exercises, we also thank Costa Zbinden for his advice, making his recording studio available and processing the audios, and Hadil Shurbahji and MI (who does not wish to use their full name) for their voices.

The overall project was funded by the Humanitarian Foundation of the SRC. The authors used OpenAI's ChatGPT to refine the manuscript's language and enhance its readability. Its use was strictly limited to improving clarity and language quality. Following this, the authors reviewed and verified all content to ensure accuracy. Additionally, Grammarly was used to perform a grammar check on the manuscript.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

The authors variably contributed to this study, conducting the desk review (MA), creating or editing content for the app (RTS, TB, VZ, MA, FH, and EH), participating in discussion and decision-making processes (RTS, TB, MA, VZ, FH, MH, and EH), and executing the beta test and analysis (MH and BJF). We had a very fruitful technological collaboration in which we were able to draw on extensive expertise in the field and were continuously supported with patience (SB, JW, and CK) to make technological decisions (MH and RTS).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Completed template for documenting cultural adaptations of psychological interventions.
[DOCX File , 130 KB - formative_v9i1e59905_app1.docx]

References

1. Global trends: forced displacement in 2022. United Nations High Commissioner for Refugees. URL: https://www.unhcr.org/global-trends-report-2022 [accessed 2025-08-20]
2. Sijbrandij M, Acarturk C, Bird M, Bryant RA, Burchert S, Carswell K, et al. Strengthening mental health care systems for Syrian refugees in Europe and the Middle East: integrating scalable psychological interventions in eight countries. Eur J Psychotraumatol 2017;8(Suppl 2):1388102 [FREE Full text] [doi: 10.1080/20008198.2017.1388102] [Medline: 29163867]
3. Blackmore R, Boyle JA, Fazel M, Ranasinha S, Gray KM, Fitzgerald G, et al. The prevalence of mental illness in refugees and asylum seekers: a systematic review and meta-analysis. PLoS Med 2020;17(9):e1003337 [FREE Full text] [doi: 10.1371/journal.pmed.1003337] [Medline: 32956381]
4. Henkelmann JR, de Best S, Deckers C, Jensen K, Shahab M, Elzinga B, et al. Anxiety, depression and post-traumatic stress disorder in refugees resettling in high-income countries: systematic review and meta-analysis. BJPsych Open 2020;6(4):e68 [FREE Full text] [doi: 10.1192/bjo.2020.54] [Medline: 32611475]
5. Hoell A, Kourmpeli E, Salize HJ, Heinz A, Padberg F, Habel U, et al. Prevalence of depressive symptoms and symptoms of post-traumatic stress disorder among newly arrived refugees and asylum seekers in Germany: systematic review and meta-analysis. BJPsych Open 2021;7(3):e93 [FREE Full text] [doi: 10.1192/bjo.2021.54] [Medline: 33938425]

XSL•FO
RenderX

6. Nesterko Y, Jäckle D, Friedrich M, Holzapfel L, Glaesmer H. Prevalence of post-traumatic stress disorder, depression and somatisation in recently arrived refugees in Germany: an epidemiological study. Epidemiol Psychiatr Sci 2019;29:e40 [FREE Full text] [doi: 10.1017/S2045796019000325] [Medline: 31364579]
7. Patanè M, Ghane S, Karyotaki E, Cuijpers P, Schoonmade L, Tarsitani L, et al. Prevalence of mental disorders in refugees and asylum seekers: a systematic review and meta-analysis. Glob Ment Health (Camb) 2022;9:250-263 [FREE Full text] [doi: 10.1017/gmh.2022.29] [Medline: 36618716]
8. Turrini G, Purgato M, Ballette F, Nosè M, Ostuzzi G, Barbui C. Common mental disorders in asylum seekers and refugees: umbrella review of prevalence and intervention studies. Int J Ment Health Syst 2017;11:51 [FREE Full text] [doi: 10.1186/s13033-017-0156-0] [Medline: 28855963]
9. Al-Smadi AM, Tawalbeh LI, Gammoh OS, Ashour AF, Alshraifeen A, Gougazeh YM. Anxiety, stress, and quality of life among Iraqi refugees in Jordan: a cross sectional survey. Nurs Health Sci 2017;19(1):100-104. [doi: 10.1111/nhs.12323] [Medline: 28058754]
10. Leiler A, Bjärtå A, Ekdahl J, Wasteson E. Mental health and quality of life among asylum seekers and refugees living in refugee housing facilities in Sweden. Soc Psychiatry Psychiatr Epidemiol 2019;54(5):543-551. [doi: 10.1007/s00127-018-1651-6] [Medline: 30580381]
11. Mesa-Vieira C, Haas AD, Buitrago-Garcia D, Roa-Diaz ZM, Minder B, Gamba M, et al. Mental health of migrants with pre-migration exposure to armed conflict: a systematic review and meta-analysis. Lancet Public Health 2022;7(5):e469-e481 [FREE Full text] [doi: 10.1016/S2468-2667(22)00061-5] [Medline: 35487232]
12. Giacco D. Identifying the critical time points for mental health of asylum seekers and refugees in high-income countries. Epidemiol Psychiatr Sci 2019;29:e61 [FREE Full text] [doi: 10.1017/S204579601900057X] [Medline: 31576800]
13. Aragona M, Pucci D, Mazzetti M, Geraci S. Post-migration living difficulties as a significant risk factor for PTSD in immigrants: a primary care study. Ital J Public Health 2012;9(3):e7525. [doi: 10.2427/7525]
14. Carswell K, Blackburn P, Barker C. The relationship between trauma, post-migration problems and the psychological well-being of refugees and asylum seekers. Int J Soc Psychiatry 2011;57(2):107-119. [doi: 10.1177/0020764009105699] [Medline: 21343209]
15. Kaltenbach E, Schauer M, Hermenau K, Elbert T, Schalinski I. Course of mental health in refugees – a one year panel survey. Front Psychiatry 2018;9:352 [FREE Full text] [doi: 10.3389/fpsyt.2018.00352] [Medline: 30123145]
16. Silove D. The ADAPT model: a conceptual framework for mental health and psychosocial programming in post conflict settings. Intervention 2013;11(3):237-248. [doi: 10.1097/wtf.0000000000000005]
17. Chen W, Hall BJ, Ling L, Renzaho AM. Pre-migration and post-migration factors associated with mental health in humanitarian migrants in Australia and the moderation effect of post-migration stressors: findings from the first wave data of the BNLA cohort study. Lancet Psychiatry 2017;4(3):218-229. [doi: 10.1016/S2215-0366(17)30032-9] [Medline: 28161455]
18. Drescher A, Kiselev N, Akhtar A, Acarturk C, Bryant RA, Ilkkursun Z, et al. Problems after flight: understanding and comparing Syrians' perspectives in the Middle East and Europe. BMC Public Health 2021;21(1):717 [FREE Full text] [doi: 10.1186/s12889-021-10498-1] [Medline: 33849507]
19. Gorst-Unsworth C, Goldenberg E. Psychological sequelae of torture and organised violence suffered by refugees from Iraq: trauma-related factors compared with social factors in exile. Br J Psychiatry 1998;172:90-94. [doi: 10.1192/bjp.172.1.90] [Medline: 9534840]
20. Hajak VL, Sardana S, Verdeli H, Grimm S. A systematic review of factors affecting mental health and well-being of asylum seekers and refugees in Germany. Front Psychiatry 2021;12:643704 [FREE Full text] [doi: 10.3389/fpsyt.2021.643704] [Medline: 33815176]
21. Schick M, Zumwald A, Knöpfli B, Nickerson A, Bryant RA, Schnyder U, et al. Challenging future, challenging past: the relationship of social integration and psychological impairment in traumatized refugees. Eur J Psychotraumatol 2016;7:28057 [FREE Full text] [doi: 10.3402/ejpt.v7.28057] [Medline: 26886484]
22. Schweitzer RD, Brough M, Vromans L, Asic-Kobe M. Mental health of newly arrived Burmese refugees in Australia: contributions of pre-migration and post-migration experience. Aust N Z J Psychiatry 2011;45(4):299-307. [doi: 10.3109/00048674.2010.543412] [Medline: 21303193]
23. Gleeson C, Frost R, Sherwood L, Shevlin M, Hyland P, Halpin R, et al. Post-migration factors and mental health outcomes in asylum-seeking and refugee populations: a systematic review. Eur J Psychotraumatol 2020;11(1):1793567 [FREE Full text] [doi: 10.1080/20008198.2020.1793567] [Medline: 33312450]
24. Jannesari S, Hatch S, Prina M, Oram S. Post-migration social-environmental factors associated with mental health problems among asylum seekers: a systematic review. J Immigr Minor Health 2020;22(5):1055-1064 [FREE Full text] [doi: 10.1007/s10903-020-01025-2] [Medline: 32430778]
25. Miller KE, Rasmussen A. War exposure, daily stressors, and mental health in conflict and post-conflict settings: bridging the divide between trauma-focused and psychosocial frameworks. Soc Sci Med 2010;70(1):7-16. [doi: 10.1016/j.socscimed.2009.09.029] [Medline: 19854552]
26. Asylstatistik 2024. State Secretariat for Migration, Switzerland. URL: https://www.sem.admin.ch/dam/sem/de/data/publiservice/statistik/asylstatistik/2024/stat-jahr-2024-kommentar.pdf [accessed 2025-08-20]

XSL•FO
RenderX

27. Fuhr DC, Acarturk C, McGrath M, Ilkkursun Z, Sondorp E, Sijbrandij M, et al. Treatment gap and mental health service use among Syrian refugees in Sultanbeyli, Istanbul: a cross-sectional survey. Epidemiol Psychiatr Sci 2019;29:e70 [FREE Full text] [doi: 10.1017/S2045796019000660] [Medline: 31727205]
28. Satinsky E, Fuhr DC, Woodward A, Sondorp E, Roberts B. Mental health care utilisation and access among refugees and asylum seekers in Europe: a systematic review. Health Policy 2019;123(9):851-863 [FREE Full text] [doi: 10.1016/j.healthpol.2019.02.007] [Medline: 30850148]
29. Maier T, Schmidt M, Mueller J. Mental health and healthcare utilization in adult asylum seekers. Swiss Med Wkly 2010;140:w13110 [FREE Full text] [doi: 10.4414/smw.2010.13110] [Medline: 21104473]
30. Müller F, Roose Z, Landis F, Gianola G. Psychische Gesundheit von traumatisierten Asylsuchenden: Situationsanalyse und Empfehlungen. Interface. 2018 Jun. URL: https://www.interface-pol.ch/app/uploads/2022/03/Be_Asylsuchende_psychische_Gesundheit_d.pdf [accessed 2025-08-20]
31. Höglinger D, Rudin M, Guggisberg J. Analyse zu den Auswirkungen der Reduktion der Fallbelastung in der Sozialberatung der Stadt Winterthur. BASS, Switzerland. 2021 May. URL: https://www.buerobass.ch/fileadmin/Files/2021/2021_Reduktion_Falllast_Winterthur_Schlussbericht_DE.pdf [accessed 2025-08-20]
32. Davolio ME, Smith RSN, Gehrig M. Auswirkungen der Falllastreduktion in der Sozialhilfe auf Ablösequote und Fallkostenntschleunigung zahlt sich aus. Rev Suisse Trav Soc 2019:31-51. [doi: 10.5169/SEALS-855356]
33. Oetterli M, Niederhauser A, Pluess S. Ist-Analyse von psychosozialen Behandlungs-und Betreuungsangeboten für traumatisierte Personen im Asyl-und Flüchtlingsbereich. Interface/Evaluanda. 2013 Jun. URL: https://www.interface-pol.ch/app/uploads/2022/11/Be_Psychosoziale_Angebote_BFM.pdf [accessed 2025-08-20]
34. Bartolomei J, Baeriswyl-Cottin R, Framorando D, Kasina F, Premand N, Eytan A, et al. What are the barriers to access to mental healthcare and the primary needs of asylum seekers? A survey of mental health caregivers and primary care workers. BMC Psychiatry 2016;16(1):336 [FREE Full text] [doi: 10.1186/s12888-016-1048-6] [Medline: 27686067]
35. Kiselev N, Morina N, Schick M, Watzke B, Schnyder U, Pfaltz MC. Barriers to access to outpatient mental health care for refugees and asylum seekers in Switzerland: the therapist's view. BMC Psychiatry 2020;20(1):378 [FREE Full text] [doi: 10.1186/s12888-020-02783-x] [Medline: 32680485]
36. Shannon PJ, Wieling E, Simmelink-McCleary J, Becher E. Beyond stigma: barriers to discussing mental health in refugee populations. J Loss Trauma 2014;20(3):1-16. [doi: 10.1080/15325024.2014.934629]
37. Schick M, Morina N, Mistridis P, Schnyder U, Bryant RA, Nickerson A. Changes in post-migration living difficulties predict treatment outcome in traumatized refugees. Front Psychiatry 2018;9:476 [FREE Full text] [doi: 10.3389/fpsyt.2018.00476] [Medline: 30356791]
38. de Graaff AM, Cuijpers P, Twisk JWR, Kieft B, Hunaidy S, Elsawy M, STRENGTHS Consortium. Peer-provided psychological intervention for Syrian refugees: results of a randomised controlled trial on the effectiveness of Problem Management Plus. BMJ Ment Health 2023;26(1):e300637 [FREE Full text] [doi: 10.1136/bmjment-2022-300637] [Medline: 36789918]
39. Spaaij J, de Graaff AM, Akhtar A, Kiselev N, McDaid D, Moergeli H, STRENGTHS Consortium. The effect of a low-level psychological intervention (PM+) on post-migration living difficulties – results from two studies in Switzerland and in the Netherlands. Compr Psychiatry 2023;127:152421 [FREE Full text] [doi: 10.1016/j.comppsych.2023.152421] [Medline: 37708580]
40. Purgato M, Carswell K, Acarturk C, Au T, Akbai S, Anttila M, et al. Effectiveness and cost-effectiveness of Self-Help Plus (SH+) for preventing mental disorders in refugees and asylum seekers in Europe and Turkey: study protocols for two randomised controlled trials. BMJ Open 2019;9(5):e030259 [FREE Full text] [doi: 10.1136/bmjopen-2019-030259] [Medline: 31092670]
41. Tol WA, Leku MR, Lakin DP, Carswell K, Augustinavicius J, Adaku A, et al. Guided self-help to reduce psychological distress in South Sudanese female refugees in Uganda: a cluster randomised trial. Lancet Glob Health 2020;8(2):e254-e263 [FREE Full text] [doi: 10.1016/S2214-109X(19)30504-2] [Medline: 31981556]
42. Murray LK, Haroz E, Dorsey S, Kane J, Bolton PA, Pullmann MD. Understanding mechanisms of change: an unpacking study of the evidence-based common-elements treatment approach (CETA) in low and middle income countries. Behav Res Ther 2020;130:103430 [FREE Full text] [doi: 10.1016/j.brat.2019.103430] [Medline: 31780251]
43. Müller F, Schwegler C. Niederschwellige Kurzinterventionen für psychisch belastete Asylsuchende und Flüchtlinge. Bundesamt für Gesundheit BAG. URL: https://www.bag.admin.ch/de/gesundheitsversorgung-fuer-asylsuchende [accessed 2025-08-20]
44. Tol WA, Le PD, Harrison SL, Galappatti A, Annan J, Baingana FK, et al. Mental health and psychosocial support in humanitarian settings: research priorities for 2021-30. Lancet Glob Health 2023;11(6):e969-e975 [FREE Full text] [doi: 10.1016/S2214-109X(23)00128-6] [Medline: 37116530]
45. Ellis DM, Draheim AA, Anderson PL. Culturally adapted digital mental health interventions for ethnic/racial minorities: a systematic review and meta-analysis. J Consult Clin Psychol 2022;90(10):717-733. [doi: 10.1037/ccp0000759] [Medline: 36227330]

XSL•FO
RenderX

46. Karyotaki E, Miguel C, Panagiotopoulou OM, Harrer M, Seward N, Sijbrandij M, et al. Digital interventions for common mental disorders in low- and middle-income countries: a systematic review and meta-analysis. Glob Ment Health (Camb) 2023;10:e68 [FREE Full text] [doi: 10.1017/gmh.2023.50] [Medline: 38024798]
47. Andersson G, Titov N. Advantages and limitations of internet-based interventions for common mental disorders. World Psychiatry 2014;13(1):4-11 [FREE Full text] [doi: 10.1002/wps.20083] [Medline: 24497236]
48. Romer A, Rezzonico L, Della Torre L, Omlin E. Country report: Switzerland. Asylum Information Database (AIDA), European Council on Refugees and Exiles (ECRE). 2022. URL: https://asylumineurope.org/wp-content/uploads/2023/06/AIDA-CH_2022-Update.pdf [accessed 2025-08-20]
49. Alencar A, Kondova K, Ribbens W. The smartphone as a lifeline: an exploration of refugees' use of mobile communication technologies during their flight. Media Cult Soc 2018;41(6):828-844. [doi: 10.1177/0163443718813486]
50. Maitland CF, Xu Y. A social informatics analysis of refugee mobile phone use: a case study of Zaaatari Syrian Refugee Camp. SSRN Electron J 2015:10. [doi: 10.2139/ssrn.2588300]
51. Andersson G, Titov N, Dear BF, Rozental A, Carlbring P. Internet-delivered psychological treatments: from innovation to implementation. World Psychiatry 2019;18(1):20-28 [FREE Full text] [doi: 10.1002/wps.20610] [Medline: 30600624]
52. Carlbring P, Andersson G, Cuijpers P, Riper H, Hedman-Lagerlöf E. Internet-based vs. face-to-face cognitive behavior therapy for psychiatric and somatic disorders: an updated systematic review and meta-analysis. Cogn Behav Ther 2018;47(1):1-18 [FREE Full text] [doi: 10.1080/16506073.2017.1401115] [Medline: 29215315]
53. Catarino A, Harper S, Malcolm R, Stainthorpe A, Warren G, Margoum M, et al. Economic evaluation of 27,540 patients with mood and anxiety disorders and the importance of waiting time and clinical effectiveness in mental healthcare. Nat Mental Health 2023;1(9):667-678. [doi: 10.1038/s44220-023-00106-z]
54. Muñoz R, Bunge EL, Chen K, Schueller SM, Bravin JI, Shaughnessy EA, et al. Massive open online interventions: a novel model for delivering behavioral-health services worldwide. Clin Psychol Sci 2015;4(2):194-205. [doi: 10.1177/2167702615583840]
55. Patel V, Saxena S, Lund C, Thornicroft G, Baingana F, Bolton P, et al. The Lancet Commission on global mental health and sustainable development. Lancet 2018;392(10157):1553-1598. [doi: 10.1016/S0140-6736(18)31612-X] [Medline: 30314863]
56. Moshe I, Terhorst Y, Philippi P, Domhardt M, Cuijpers P, Cristea I, et al. Digital interventions for the treatment of depression: a meta-analytic review. Psychol Bull 2021;147(8):749-786. [doi: 10.1037/bul0000334] [Medline: 34898233]
57. Musiat P, Johnson C, Atkinson M, Wilksch S, Wade T. Impact of guidance on intervention adherence in computerised interventions for mental health problems: a meta-analysis. Psychol Med 2022;52(2):229-240. [doi: 10.1017/S0033291721004621] [Medline: 34802474]
58. Baumeister H, Reichler L, Munzinger M, Lin J. The impact of guidance on Internet-based mental health interventions — a systematic review. Internet Interv 2014;1(4):205-215. [doi: 10.1016/j.invent.2014.08.003]
59. Leung C, Pei J, Hudec K, Shams F, Munthali R, Vigo D. The effects of nonclinician guidance on effectiveness and process outcomes in digital mental health interventions: systematic review and meta-analysis. J Med Internet Res 2022;24(6):e36004 [FREE Full text] [doi: 10.2196/36004] [Medline: 35511463]
60. Cuijpers P, Heim E, Abi Ramia J, Burchert S, Carswell K, Cornelisz I, et al. Effects of a WHO-guided digital health intervention for depression in Syrian refugees in Lebanon: a randomized controlled trial. PLoS Med 2022;19(6):e1004025 [FREE Full text] [doi: 10.1371/journal.pmed.1004025] [Medline: 35737665]
61. Burchert S, Alkneme MS, Bird M, Carswell K, Cuijpers P, Hansen P, et al. User-centered app adaptation of a low-intensity e-mental health intervention for Syrian refugees. Front Psychiatry 2018;9:663 [FREE Full text] [doi: 10.3389/fpsyt.2018.00663] [Medline: 30740065]
62. Cohen F, Yaeger L. Task-shifting for refugee mental health and psychosocial support: a scoping review of services in humanitarian settings through the lens of RE-AIM. Implement Res Pract 2021;2:2633489521998790 [FREE Full text] [doi: 10.1177/2633489521998790] [Medline: 37089989]
63. Healey P, Stager ML, Woodmass K, Dettlaff AJ, Vergara A, Janke R, et al. Cultural adaptations to augment health and mental health services: a systematic review. BMC Health Serv Res 2017;17(1):8 [FREE Full text] [doi: 10.1186/s12913-016-1953-x] [Medline: 28056967]
64. McCleary JS, Horn TL. Processes for culturally adapting behavioral health interventions for people with refugee backgrounds: a scoping review. Am J Community Psychol 2024;73(1-2):250-266. [doi: 10.1002/ajcp.12709] [Medline: 37853807]
65. Heim E, Mewes R, Abi Ramia J, Glaesmer H, Hall B, Harper Shehadeh M, et al. Reporting cultural adaptation in psychological trials - the RECAPT criteria. Clin Psychol Eur 2021;3:e6351 [FREE Full text] [doi: 10.32872/cpe.6351] [Medline: 36405678]
66. Resnicow K, Baranowski T, Ahluwalia J, Braithwaite RL. Cultural sensitivity in public health: defined and demystified. Ethn Dis 1999;9(1):10-21. [Medline: 10355471]
67. Kiselev N, Pfaltz M, Schick M, Bird M, Pernille H, Sijbrandij M, et al. Problems faced by Syrian refugees and asylum seekers in Switzerland. Swiss Med Wkly 2020;150:w20381 [FREE Full text] [doi: 10.4414/smw.2020.20381] [Medline: 33105021]

XSL•FO
RenderX

68. Feinstein S, Poleacovschi C, Drake R, Winters LA. States and refugee integration: a comparative analysis of France, Germany, and Switzerland. J Int Migr Integr 2022;23(4):2167-2194. [doi: 10.1007/s12134-021-00929-8]
69. Hall GCN, Ibaraki AY, Huang ER, Marti CN, Stice E. A meta-analysis of cultural adaptations of psychological interventions. Behav Ther 2016;47(6):993-1014. [doi: 10.1016/j.beth.2016.09.005] [Medline: 27993346]
70. Harper Shehadeh M, Heim E, Chowdhary N, Maercker A, Albanese E. Cultural adaptation of minimally guided interventions for common mental disorders: a systematic review and meta-analysis. JMIR Ment Health 2016;3(3):e44 [FREE Full text] [doi: 10.2196/mental.5776] [Medline: 27670598]
71. Spanhel K, Balci S, Feldhahn F, Bengel J, Baumeister H, Sander LB. Cultural adaptation of internet- and mobile-based interventions for mental disorders: a systematic review. NPJ Digit Med 2021;4(1):128 [FREE Full text] [doi: 10.1038/s41746-021-00498-1] [Medline: 34433875]
72. Cuijpers P, Karyotaki E, Reijnders M, Purgato M, Barbui C. Psychotherapies for depression in low- and middle-income countries: a meta-analysis. World Psychiatry 2018;17(1):90-101 [FREE Full text] [doi: 10.1002/wps.20493] [Medline: 29352530]
73. Bernal G, Sáez‐Santiago E. Culturally centered psychosocial interventions. J. Community Psychol 2006;34(2):121-132. [doi: 10.1002/jcop.20096]
74. Heim E, Knaevelsrud C. Standardised research methods and documentation in cultural adaptation: the need, the potential and future steps. Clin Psychol Eur 2021;3:e5513 [FREE Full text] [doi: 10.32872/cpe.5513] [Medline: 36405674]
75. Rustage K, Crawshaw A, Majeed-Hajaj S, Deal A, Nellums L, Ciftci Y, et al. Participatory approaches in the development of health interventions for migrants: a systematic review. BMJ Open 2021;11(10):e053678 [FREE Full text] [doi: 10.1136/bmjopen-2021-053678] [Medline: 34697122]
76. Cargo M, Mercer SL. The value and challenges of participatory research: strengthening its practice. Annu Rev Public Health 2008;29:325-350. [doi: 10.1146/annurev.publhealth.29.091307.083824] [Medline: 18173388]
77. Human-centered design. U.S. General Services Administration. URL: https://digital.gov/topics/human-centered-design [accessed 2025-08-28]
78. Mayring P. Qualitative Inhaltsanalyse: Grundlagen und Techniken. 12th revised edition. Weinheim: Julius Beltz GmbH & Co KG; 2015.
79. Greene MC, Jordans MJD, Kohrt BA, Ventevogel P, Kirmayer LJ, Hassan G, et al. Addressing culture and context in humanitarian response: preparing desk reviews to inform mental health and psychosocial support. Confl Health 2017;11:21 [FREE Full text] [doi: 10.1186/s13031-017-0123-z] [Medline: 29163666]
80. Shala M, Morina N, Burchert S, Cerga-Pashoja A, Knaevelsrud C, Maercker A, et al. Cultural adaptation of Hap-pas-Hapi, an internet and mobile-based intervention for the treatment of psychological distress among Albanian migrants in Switzerland and Germany. Internet Interv 2020;21:100339 [FREE Full text] [doi: 10.1016/j.invent.2020.100339] [Medline: 32983906]
81. Shala M, Morina N, Salis Gross C, Maercker A, Heim E. A point in the heart: concepts of emotional distress among Albanian-speaking immigrants in Switzerland. Cult Med Psychiatry 2020;44(1):1-34. [doi: 10.1007/s11013-019-09638-5] [Medline: 31209652]
82. Design thinking (DT). Interaction Design Foundation. URL: https://www.interaction-design.org/literature/topics/design-thinking#the_five_stages_of_design_thinking-6 [accessed 2025-08-20]
83. Aeschlimann A, Heim E, Hoxha A, Triantafyllidou V, Killikelly C, Haji F, et al. Cultural adaptation of an internet-based self-help app for grieving Syrian refugees in Switzerland. BMC Public Health 2024;24(1):3048 [FREE Full text] [doi: 10.1186/s12889-024-20507-8] [Medline: 39501191]
84. Harper Shehadeh MJ, Abi Ramia J, Cuijpers P, El Chammay R, Heim E, Kheir W, et al. Step-by-step, an e-mental health intervention for depression: a mixed methods pilot study from Lebanon. Front Psychiatry 2019;10:986 [FREE Full text] [doi: 10.3389/fpsyt.2019.00986] [Medline: 32116815]
85. Mediavilla R, McGreevy KR, Felez-Nobrega M, Monistrol-Mula A, Bravo-Ortiz M, Bayón C, RESPOND Consortium. Effectiveness of a stepped-care programme of internet-based psychological interventions for healthcare workers with psychological distress: study protocol for the RESPOND healthcare workers randomised controlled trial. Digit Health 2022;8:20552076221129084 [FREE Full text] [doi: 10.1177/20552076221129084] [Medline: 36211795]
86. Spaaij J, Kiselev N, Berger C, Bryant RA, Cuijpers P, de Graaff AM, et al. Feasibility and acceptability of Problem Management Plus (PM+) among Syrian refugees and asylum seekers in Switzerland: a mixed-method pilot randomized controlled trial. Eur J Psychotraumatol 2022;13(1):2002027 [FREE Full text] [doi: 10.1080/20008198.2021.2002027] [Medline: 35126880]
87. Seewer N, Skoko A, Käll A, Andersson G, Luhmann M, Berger T, et al. Efficacy of an internet-based self-help intervention with human guidance or automated messages to alleviate loneliness: a three-armed randomized controlled trial. Sci Rep 2024;14(1):6569 [FREE Full text] [doi: 10.1038/s41598-024-57254-0] [Medline: 38503870]
88. Berger T, Hämmerli K, Gubser N, Andersson G, Caspar F. Internet-based treatment of depression: a randomized controlled trial comparing guided with unguided self-help. Cogn Behav Ther 2011;40(4):251-266. [doi: 10.1080/16506073.2011.616531] [Medline: 22060248]

XSL•FO
RenderX

89. Hassan G, Ventevogel P, Jefee-Bahloul H, Barkil-Oteo A, Kirmayer LJ. Mental health and psychosocial wellbeing of Syrians affected by armed conflict. Epidemiol Psychiatr Sci 2016;25(2):129-141 [FREE Full text] [doi: 10.1017/S2045796016000044] [Medline: 26829998]
90. Byrow Y, Pajak R, Specker P, Nickerson A. Perceptions of mental health and perceived barriers to mental health help-seeking amongst refugees: a systematic review. Clin Psychol Rev 2020;75:101812. [doi: 10.1016/j.cpr.2019.101812] [Medline: 31901882]
91. Turrini G, Purgato M, Acarturk C, Anttila M, Au T, Ballette F, et al. Efficacy and acceptability of psychosocial interventions in asylum seekers and refugees: systematic review and meta-analysis. Epidemiol Psychiatr Sci 2019;28(4):376-388 [FREE Full text] [doi: 10.1017/S2045796019000027] [Medline: 30739625]
92. Digital Research Creation Tool. Freie Universität Berlin. URL: https://www.ewi-psy.fu-berlin.de/psychologie/einrichtungen/cmhds/direct/index.html [accessed 2025-08-20]
93. Härle D. TSY® Traumasensitives Yoga. Institut für Traumatherapie. URL: https://www.trauma-institut.eu [accessed 2025-08-12]
94. Härle D. Körperorientierte Traumatherapie: Sanfte Heilung mit traumasensitivem Yoga. Germany: Junfermann Verlag; 2015.
95. Härle D. Yoga traumasensitiv unterrichten: Affektregulation, Selbstwirksamkeit und Körperwahrnehmung verbessern. Austria: Asanger; 2019.
96. Epping-Jordan JE, Harris R, Brown FL, Carswell K, Foley C, García-Moreno C, et al. Self-Help Plus (SH+): a new WHO stress management package. World Psychiatry 2016;15(3):295-296 [FREE Full text] [doi: 10.1002/wps.20355] [Medline: 27717271]
97. Doing What Matters in Times of Stress: An Illustrated Guide. Geneva: World Health Organization; 2020. URL: https://apps.who.int/iris/handle/10665/331901 [accessed 2025-08-20]
98. Sonne für die Seele, Ein Selbsthilfebuch für traumatisierte Flüchtlinge: Informationen und Übungen, die helfen können traumatische Erfahrungen zu bewältigen. Nürnberg, Germany: TraumaHilfeZentrum Nürnberg e.V; 2017. URL: https://tinyurl.com/2326e2dt [accessed 2025-08-20]
99. Nezu AM, Nezu CM, D'Zurilla TJ. Problem-Solving Therapy: A Treatment Manual. New York: Springer; 2012.
100. Bur OT, Krieger T, Moritz S, Klein JP, Berger T. Optimizing the context of support of web-based self-help in individuals with mild to moderate depressive symptoms: a randomized full factorial trial. Behav Res Ther 2022;152:104070 [FREE Full text] [doi: 10.1016/j.brat.2022.104070] [Medline: 35306266]
101. Haines S. Pain is Really Strange. Philadelphia, PA: Singing Dragon; 2015.
102. von Wachter M. Chronische Schmerzen: Selbsthilfe und Therapiebegleitung, Orientierung für Angehörige und konkrete Tipps und Fallbeispiele. Berlin: Springer; 2012.
103. Hinton DE, Jalal B. Guidelines for the implementation of culturally sensitive cognitive behavioural therapy among refugees and in global contexts. Intervention 2014;12:78-93. [doi: 10.1097/wtf.0000000000000069]
104. El-Haj-Mohamad R, Nohr L, Niemeyer H, Böttche M, Knaevelsrud C. Smartphone-delivered mental health care interventions for refugees: a systematic review of the literature. Glob Ment Health (Camb) 2023;10:e6 [FREE Full text] [doi: 10.1017/gmh.2022.61] [Medline: 36843879]
105. Böttche M, Kampisiou C, Stammel N, El-Haj-Mohamad R, Heeke C, Burchert S, et al. From formative research to cultural adaptation of a face-to-face and internet-based cognitive-behavioural intervention for Arabic-speaking refugees in Germany. Clin Psychol Eur 2021;3(Spec Issue):e4623 [FREE Full text] [doi: 10.32872/cpe.4623] [Medline: 36405676]
106. Seiferth C, Vogel L, Aas B, Brandhorst I, Carlbring P, Conzelmann A, et al. How to e-mental health: a guideline for researchers and practitioners using digital technology in the context of mental health. Nat Mental Health 2023;1(8):542-554. [doi: 10.1038/s44220-023-00085-1]
107. Borghouts J, Eikey E, Mark G, De Leon C, Schueller SM, Schneider M, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. J Med Internet Res 2021;23(3):e24387 [FREE Full text] [doi: 10.2196/24387] [Medline: 33759801]

Abbreviations

ACIs: asylum care interviewees
ADAPT: adaptation and development after persecution and trauma
DIRECT: Digital Research Creation Tool
ICIs: intercultural interpreters
LMICs: low- and middle-income countries
MHPSS: mental health and psychosocial support
PMLDs: postmigration living difficulties
PTSD: posttraumatic stress disorder
RCT: randomized controlled trial
RECAPT: Reporting Cultural Adaptation in Psychological Trials

XSL•FO
RenderX

SRC: Swiss Red Cross
TGIs: target group interviewees
UAB: user advisory board
UX: user experience

Edited by A Mavragani; submitted 25.04.24; peer-reviewed by A AL-Asadi, A Doğan Kuday, N Kiselev, K Spanhel, P Nascimento; comments to author 05.12.24; revised version received 02.03.25; accepted 23.06.25; published 17.09.25.

Please cite as:
Stoeckli RT, Berger T, Aebersold M, Zoellner V, Haji F, Hunziker M, Jesus Ferreira B, Hosmann M, Burchert S, Wabiszczewicz J, Knaevelsrud C, Heim E
A Digital Psychosocial Service (Sui App) for Arabic-Speaking Refugees in Switzerland: Development and Cultural-Contextual Adaptation Using a User-Centered and Participatory Approach
JMIR Form Res 2025;9:e59905
URL: https://formative.jmir.org/2025/1/e59905
doi:10.2196/59905
PMID:

©Rilana Tanja Stoeckli, Thomas Berger, Monia Aebersold, Viktoria Zoellner, Farhad Haji, Muriel Hunziker, Beatriz Jesus Ferreira, Michel Hosmann, Sebastian Burchert, Jessica Wabiszczewicz, Christine Knaevelsrud, Eva Heim. Originally published in JMIR Formative Research (https://formative.jmir.org), 17.09.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Patient-Reported Experiences With Long-Term Lifestyle Self-Monitoring in Heart Disease: Mixed Methods Study

Mayra Goevaerts[1,2], PhD; Nicole Tenbült - Van Limpt[1,2], MSc; Willem J Kop[3], Prof Dr; Hareld Kemps[1,2], Prof Dr, MD; Yuan Lu[1], PhD

[1]Department of Industrial Design, Eindhoven University of Technology, Eindhoven, The Netherlands

[2]Department of Cardiology, Máxima Medisch Centrum, Eindhoven, The Netherlands

[3]Center of Research on Psychological Disorders and Somatic Diseases, Department of Medical and Clinical Psychology, Tilburg University, Tilburg, The Netherlands

Corresponding Author:
Mayra Goevaerts, PhD
Department of Industrial Design
Eindhoven University of Technology
Groene Loper 3
Eindhoven, 5612 AE
The Netherlands
Phone: 31 040 247 9111
Email: w.f.goevaerts@tue.nl

Abstract

Background: Lifestyle behaviors strongly predict cardiovascular morbidity and mortality, emphasizing the need for strategies that support sustained lifestyle changes in patients with cardiac disease. Digital health solutions, including wearables, mobile apps, and chatbots, enable self-monitoring of lifestyle behaviors but often face challenges with engagement and usability. While self-monitoring systems can increase awareness and accountability, maintaining user engagement remains crucial for their effectiveness in promoting behavior change and long-term improvements.

Objective: This study evaluated patient experiences with a lifestyle monitoring system combining a web application, health watch, and chatbot. We explored facilitators of and barriers to long-term adherence and assessed the impact of self-monitoring on lifestyle awareness and behavior change in patients with cardiac disease.

Methods: We conducted a mixed methods study with patients who used an eHealth platform for self-monitoring lifestyle behaviors during 1 year following an invasive cardiac procedure. This study included 100 patients (mean age 61.6, SD 10.4 y; n=88, 88% male) comprising both completers (n=57, 57%) and dropouts (n=43, 43%). Patients engaged in quarterly phone interviews and questionnaires and completed an end-of-study questionnaire. Completers participated in a structured evaluation interview; dropouts provided a reason for discontinuation. Quantitative and qualitative data analyses focused on usability, long-term adherence facilitators and barriers, lifestyle awareness, and behavior change.

Results: Patients completed 157 quarterly questionnaires (n=145, 92.4% by completers and n=12, 7.6% by dropouts) and 217 phone interviews (n=171, 78.8% with completers and n=46, 21.2% with dropouts). In total, 77 patients (of whom n=54, 70% were completers and n=23, 30% were dropouts) completed end-of-study questionnaires, and 98% (56/57) of completers participated in the evaluation interviews. Completers reported higher perceptions of the platform's usefulness, ease of use, and satisfaction (P<.001 in all cases) than dropouts. Dropout reasons linked to self-monitoring (34/43, 79%) included high self-report burden and dissatisfaction with the chatbot, poor overall usability experience, health watch technical challenges causing frustration, limited perceived usefulness, mental stress from self-monitoring, and low motivation. Key facilitators of long-term engagement included routine formation, structured reminders, and minimal effort associated with the wearable. Barriers included repetitive chatbot questions (causing cognitive burden) and technical issues with the health watch. Self-monitoring increased lifestyle awareness among completers, particularly regarding physical activity (25/56, 45%) and nutrition (29/56, 52%), with smaller effects for sleep quality (7/56, 13%) and mental stress (1/56, 2%). It facilitated behavior change in physical activity and nutrition (16/56, 29% each) and sleep quality (4/56, 7%) but not in mental stress. Adaptive personalization, mobile accessibility, and real-time feedback could improve adherence.

XSL•FO
RenderX

Conclusions: Fostering routine formation while minimizing patient burden through personalized, flexible, and adaptive features is important for sustained patient engagement in eHealth monitoring systems. Enhancing relevance and usability while reducing complexity and technical barriers can optimize digital health tools and promote lasting behavior change.

International Registered Report Identifier (IRRID): RR2-10.1186/s12872-023-03222-x

(JMIR Form Res 2025;9:e76978) doi:10.2196/76978

KEYWORDS

cardiac rehabilitation; self-monitoring; wearable electronic devices; chatbots; behavior change

Introduction

Background

Cardiovascular disease remains a leading cause of morbidity and mortality worldwide [1], constituting a significant burden on health care systems [2] and highlighting the need for effective secondary prevention strategies. While (elective) medical interventions such as coronary artery bypass graft (CABG) surgery; percutaneous coronary intervention (PCI); radiofrequency catheter ablation; and valve surgery, including transcatheter aortic valve implantation, improve short-term outcomes (ie, reduce symptom burden and enhance short-term quality of life) [3-5], long-term patient-reported outcomes rely heavily on patients' ability to sustain healthy lifestyle behaviors [6-8]. However, maintaining healthy lifestyle behaviors over time remains challenging as many patients struggle with adherence to recommended behaviors such as regular physical activity, balanced nutrition, adequate sleep, and stress management [9]. Therefore, there is a growing interest in strategies that promote sustained lifestyle change after a cardiac intervention.

Digital health solutions such as wearable devices, mobile apps, and automated chatbot systems have emerged as promising tools for supporting patients with cardiac disease in self-monitoring their lifestyle behaviors with the aim of facilitating long-term behavior change beyond traditional cardiac rehabilitation [10]. While these technologies enhance engagement and adherence during the initial recovery period [11], long-term adoption remains a challenge [12]. Many interventions are associated with high dropout rates, yet the factors influencing sustained use over time are not completely understood. While continuous self-monitoring can enhance motivation [13], it may also lead to burden or disengagement, particularly if usability or system design issues are not adequately addressed [14]. A deeper understanding of these dynamics is crucial for optimizing digital interventions to promote long-term adherence [15] and integration into routine clinical care.

To effectively understand and support sustained behavior change, it is essential to examine not only the clinical efficacy of digital interventions but also the lived experiences of users, along with the facilitators of and barriers to long-term engagement. Previous studies have predominantly focused on short-term adherence and clinical outcomes [11], leaving limited insight into how patients with cardiac disease interact with self-monitoring platforms over extended periods, how usability and system design influence engagement, and whether continuous tracking supports or hinders long-term lifestyle behavior change. These gaps highlight the need for a more comprehensive understanding of how digital self-monitoring can be optimized for long-term adherence. Digital interventions have mainly focused on physical activity and diet, whereas interventions on other lifestyle factors such as sleep and stress management are less well studied [11] despite their well-documented impact on cardiovascular outcomes [16].

Therefore, we developed a lifestyle monitoring platform designed to balance accuracy, usability, engagement, and comprehensive tracking of key health behaviors over time. To achieve these goals, we carefully balanced the monitoring protocol to minimize patient burden while ensuring accurate insights into lifestyle behaviors based on the self-regulation theory [13], which emphasizes that individuals can manage behaviors effectively through goal setting, self-monitoring, and feedback. The platform was iteratively designed and refined through testing with both end users and experts in the field to enhance usability and engagement. We evaluated dropout and adherence patterns in patients with cardiac disease who were referred for or had undergone a cardiac intervention [17], and the results showed that over half of the patients used the system sustainably for 1 year following the intervention. These findings indicate that long-term lifestyle self-monitoring is feasible but that efforts should be made to increase patient engagement.

Objectives

In this study, a mixed methods approach was used to identify factors that are related to long-term adherence results by evaluating the experiences of both completers and dropouts using both interviews and questionnaires. Specifically, this study investigated how patients experienced the usability of the system and how this differed between completers and dropouts. Second, key facilitators of and barriers to sustained self-monitoring were explored via interviews with completers of the study and reasons for discontinuation from dropouts. In addition, this study explored how completers perceived the impact of continuous self-monitoring on lifestyle awareness and long-term behavior change while also examining their overall experiences after the cardiac intervention to gain deeper insights into their recovery process and its potential impact on engagement. We also evaluated the system design choices to derive key learnings that can inform the development of future solutions. By integrating mixed methods data from both dropouts and completers, this study provides novel insights to optimize future digital health interventions for sustained engagement in long-term monitoring of lifestyle factors.

XSL•FO
RenderX

Methods

Study Design and Setting

This mixed methods study was part of a larger research project, the Cardiovascular Research Opting for New Applications (CARE-ON) project, which aimed to evaluate a custom-built integrated eHealth system combining a wearable sensor, a chatbot, and a dashboard for self-monitoring lifestyle behaviors in patients with cardiac disease. The project was carried out at the Máxima Medical Center (Eindhoven and Veldhoven, the Netherlands) in collaboration with the Eindhoven University of Technology, Department of Industrial Design. As part of the CARE-ON project, a 1-year prospective observational clinical trial was conducted from November 2021 to July 2024 to evaluate dropout, adherence, usability, and effects of the system on lifestyle behavior for patients referred for or who had recently undergone a cardiac intervention (a detailed description of the protocol is available elsewhere [17]).

Study Equipment

The eHealth self-monitoring system consisted of 3 modules (Figure 1). The first module was a rule-based chatbot, which was accessible on both mobile and desktop, via which the patients received prompts to provide lifestyle data via a set measurement scheme (Figure 1). The chatbot was developed to capture habitual longitudinal self-report lifestyle data using mobile ecological momentary assessments with predefined questions and answer options. The chatbot focused on three domains, with questions based on validated tools: (1) sleep quality, based on the Consensus Sleep Diary [18] (only in the morning), (2) daily mental stress, based on the Positive and Negative Affect Schedule [19] (3 times per day), and (3) nutritional intake, using a 24-hour dietary recall method divided into 5 recall periods per day and integrated into the chatbot. This module was supported by a dietary database based on the Dutch Food Composition Database [20]. The development of the nutrition module is described in detail elsewhere [21].

The patients received prompts in 4-week cycles: 7 measurement days over 2 weeks distributed across 3 to 4 days per week (2 to 3 weekdays and 1 weekend day) within a 4-week cycle. On designated measurement days, patients received 5 prompts per day, with the timing of these prompts personalized according to their individual daily routines. A more detailed overview of the chatbot modules, including sample question flows, is provided in Multimedia Appendix 1. Screenshots are provided in Figure 2.

Figure 1. Illustration of the eHealth platform components and the measurement scheme for continuous lifestyle monitoring via chatbot prompts and health watch use.

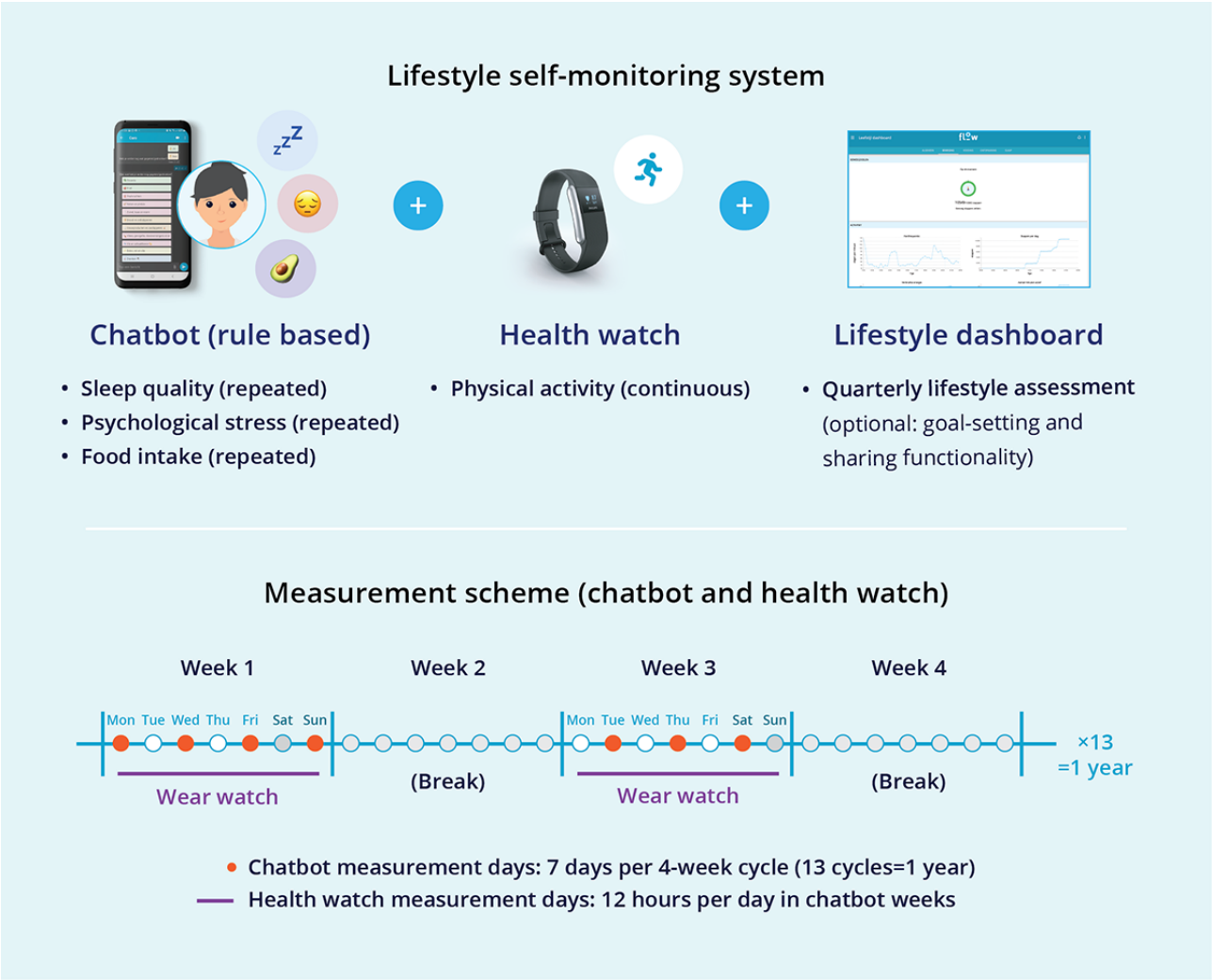

XSL•FO
RenderX

Figure 2. Screenshots of the rule-based chatbot—(A) questions related to sleep quantity and quality, (B) questions on psychological stress using a Likert scale, and (C) food categories available for users to select from when reporting food intake.

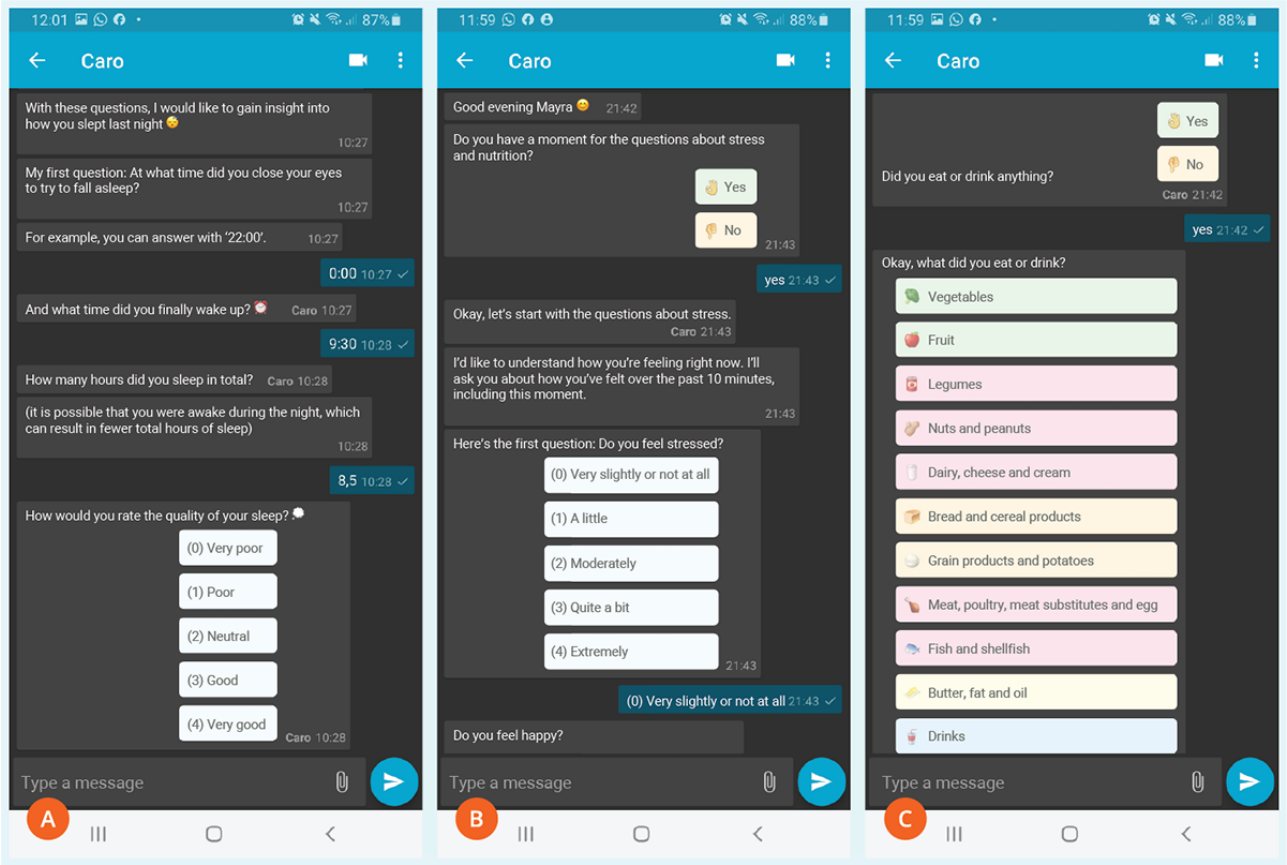

The second module was an integrated health watch (Figure 1), with which physiological parameters (ie, steps, activity counts and type, heart rate, respiration rate, energy expenditure, sleep time, cardiovascular fitness index, and watch wear time) were continuously measured. Patients were asked to wear the health watch in the same weeks in which they received chatbot prompts.

The third module was a platform where the collected data were visualized (Figure 3). In addition, lifestyle self-assessment questionnaires were made available quarterly via the platform, offering a point-in-time evaluation (that additionally included smoking behavior) and lifestyle advice based on the questionnaire scores. Patients could also use the platform to optionally set and monitor goals and share data with relatives or caregivers.

The patients were asked to at least use the chatbot and the health watch and fill in the quarterly lifestyle self-assessment questionnaires during 1 year for self-monitoring of their lifestyle behaviors.

XSL•FO
RenderX

Figure 3. Screenshot of the desktop eHealth platform integrated with a health watch—lifestyle dashboard available to patients displaying health watch data and a standardized goal as visualized lifestyle feedback.

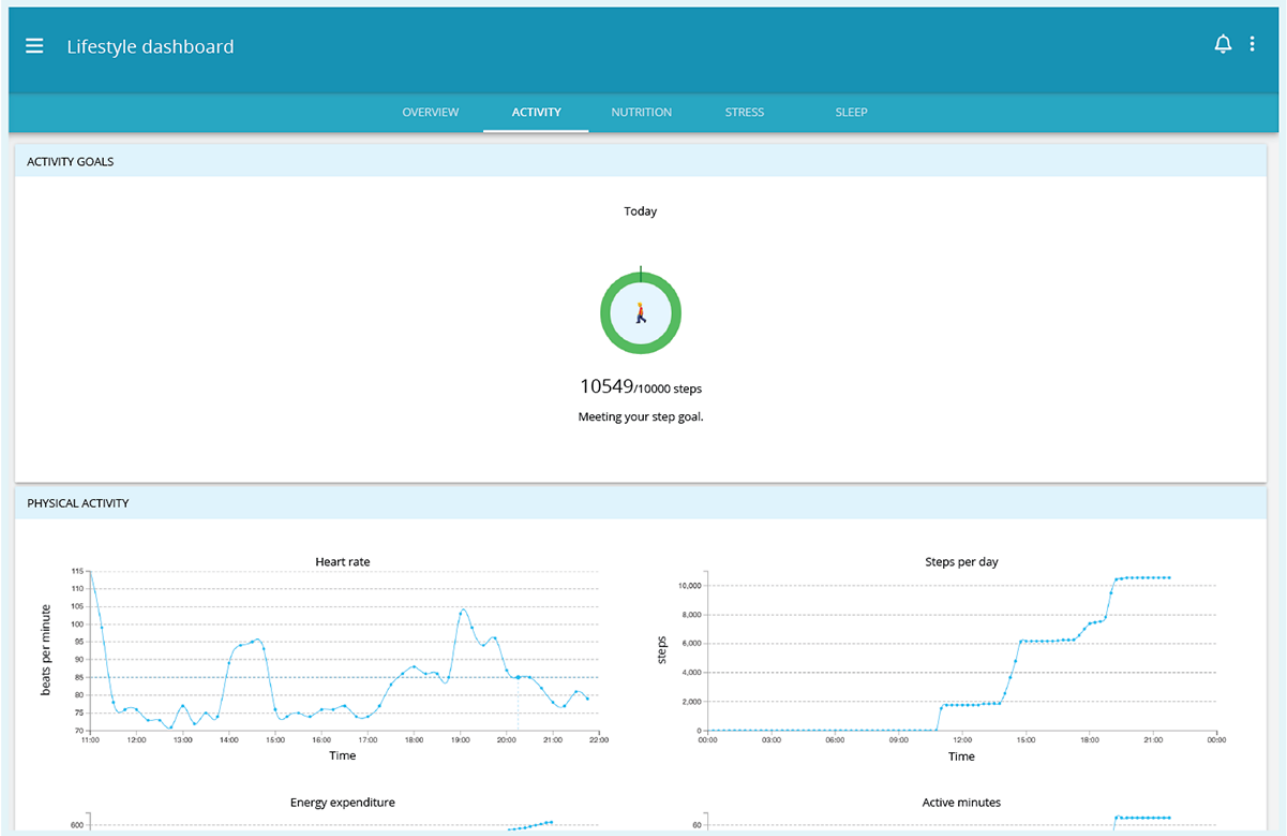

Ethical Considerations

The trial protocol was reviewed and approved by the institutional review board of the Máxima Medical Center, and the clinical trial was registered in the Dutch Trial Register (Landelijk Trial Register; registration NL9861). The CARE-ON clinical trial was conducted according to the Declaration of Helsinki. All patients provided written informed consent before study entry, including for the collection and analysis of the data used in this mixed methods study. Participant data were pseudonymized and stored securely to ensure confidentiality. No financial compensation was provided, although participants received a health watch on loan during the study period and access to the self-monitoring platform. The study protocol and the main results of the quantitative evaluation, including dropout and engagement over time, adherence, and usability, are described elsewhere [17,22]. The results are reported according to the Reporting of Studies Conducted Using Observational Routinely Collected Health Data [23] guidelines, which extend the Strengthening the Reporting of Observational Studies in Epidemiology statement, with the corresponding checklist provided in Multimedia Appendix 2.

Recruitment and Study Procedures

The complete trial comprised a sample of 100 patients who were referred for or had recently undergone CABG; a (fractional flow reserve–guided) PCI; radiofrequency catheter ablation; or valve surgery, including transcatheter aortic valve implantation. Patients were recruited either when they were scheduled for the cardiac intervention or during their clinical admission following the procedure at the Máxima Medical Center. In cases in which only a diagnostic procedure (ie, fractional flow reserve test) was performed and no intervention was deemed necessary, patients were still included because of the initial treatment intent and their comparable clinical context. Patients were asked to participate by the study nurse either via a phone consultation or during a hospital visit after being referred for a clinical intervention by their treating physician.

After providing informed consent, patients received instructions on how to use the lifestyle self-monitoring platform as part of an intake procedure conducted by the researcher and nurse specialist. In addition to the baseline intake procedure, the research nurse or researcher contacted patients via quarterly phone interviews every 3 months (at 3, 6, 9, and 12 months) to inquire about medical events and gather feedback on the self-monitoring system. At the end of the study year, an evaluation interview was conducted with patients who completed the full year (completers). Patients who discontinued participation (dropouts) were contacted to provide a reason for dropout. At each of these time points (baseline, 3, 6, 9, and 12 months or at the time of dropout), patients were asked to complete quarterly lifestyle and other research-related questionnaires via the eHealth platform.

The interview questions of both the quarterly phone interviews and the evaluation interview were entered into the data collection software in advance. During the interviews, the researcher systematically asked the questions in the same order for all patients and directly transcribed their responses into the software. The phone interviews lasted between 5 and 10 minutes. The evaluation interview (the interview guide can be found in

XSL•FO
RenderX

Multimedia Appendix 3) was conducted in person, and patients were interviewed individually (sometimes with relatives) at the hospital. Each interview lasted between 30 and 90 minutes. No audio recordings or verbatim transcripts were made as responses were documented in real time based on predefined questions, ensuring consistency across interviews and allowing for systematic data collection and analysis.

Health care providers did not have access to the collected data, and patients did not receive coaching based on their lifestyle data from the study nurse or researcher. Patients did receive structured chatbot notifications and reminders when a quarterly questionnaire was due.

Table 1 provides an overview of the data collection time points in addition to the lifestyle data captured via the self-monitoring system.

Table 1. Study setup and data collection time points.

Measurement	Study time point				
	Baseline—T0	Study year			
		3 mo	6 mo	9 mo	1 y (end)
All study participants					
Quarterly phone interviews (measurement A) with study nurse		✓	✓	✓	
Quarterly questionnaire (measurement B)—digital questionnaire via the platform		✓	✓	✓	✓
End-of-study questionnaire (measurement C)—digital questionnaire via the platform (qualitative and quantitative)		✓[a]	✓[a]	✓[a]	✓
Dropouts only					
Reasons for dropout (measurement D)—direct inquiry by study staff (nonmandatory)		✓[a]	✓[a]	✓[a]	✓[a]
Study completers only					
Evaluation interview (measurement E)—structured interview with study staff (qualitative)					✓

[a]At the point of study dropout.

Outcome Measures

Table 2 provides a more detailed overview of the outcome measures and their rationales, including the standard questions, questionnaires, and themes (particularly for the evaluation interview) used in this study.

XSL•FO
RenderX

Table 2. Outcome measures and rationales used in this mixed methods study.

Data source (qualitative or quantitative)	Rationale	Standard questions, questionnaires, and themes
Quarterly phone interviews (qualitative; measurement A)	Aimed at capturing usability challenges, system strengths and weaknesses, and potential enhancements, ensuring that patient experiences could be used to inform platform refinements	• During these interviews, patients were asked to (1) report their experiences with the system in general, (2) report whether they experienced (technical) difficulties while using the platform, and (3) identify both strengths and weaknesses of the system.
Quarterly questionnaire (qualitative; measurement B)	Aimed at capturing usability challenges, system strengths and weaknesses, and potential enhancements, ensuring that patient experiences could be used to inform platform refinements	• Open-ended question via the platform: "From your perspective, how can we improve the platform?"
End-of-study questionnaire (quantitative; measurement C)	To assess key TAM[a] [24] components, including perceived usefulness (ie, the extent to which the system was seen as beneficial for health behavior change), ease of use (ie, the perceived effort required to interact with different platform components, such as the chatbot, dashboard, and health watch), and satisfaction with platform use	• Perceived usefulness—one statement: "The feedback I receive from the patient platform helps me improve my health behavior." Responses were collected using a 5-point Likert scale (1="strongly disagree," 2="disagree," 3="neutral," 4="agree," and 5="strongly agree"). • SEQ[b]—three SEQs [25] were used in this study: (1) answering chatbot questions, (2) collecting movement data using the health watch, and (3) viewing lifestyle data in the lifestyle dashboard. Responses were collected using a 7-point scale (where 1 indicated "very difficult" and 7 indicated "very easy"), with an additional option for nonuse. • CSAT[c] [26]—two questions based on the CSAT: (1) "How satisfied are you with this platform for the monitoring of lifestyle behavior?" and (2) "How satisfied are you with the ease of use and appeal of the platform?" Responses were collected with a scoring of 1 for "extremely dissatisfied," 2 for "somewhat dissatisfied," 3 for "neither satisfied nor dissatisfied," 4 for "somewhat satisfied," and 5 for "extremely satisfied."
Dropout reasons (quantitative and qualitative; measurement D)	To gain deeper insights into the factors influencing disengagement	• Direct inquiry by study staff. Providing this information was not mandatory.
Evaluation interview (qualitative; measurement E)	A structured evaluation interview explored patients' experiences with posttreatment recovery, lifestyle changes, and self-monitoring behaviors (eg, facilitators of and barriers to long-term use), as well as their engagement with the system, including its usability (also per module) and perceived burden. This approach enabled a comprehensive assessment of the system's feasibility, acceptability, and potential impact on long-term behavior change.	• Post–cardiac intervention experiences: to contextualize engagement with lifestyle monitoring, we examined how patients' recovery experiences influenced their use of the system. • Effect of lifestyle self-monitoring on lifestyle awareness and behavior change: patients were asked to report on whether self-monitoring using the system had an effect on their lifestyle awareness and whether it helped them reflect on and eventually adjust their lifestyle behaviors in accordance with the self-regulation theory [13]. • System usability and engagement: we investigated the overall usability of the system and how patients engaged with it. We evaluated key platform components, including chatbot interactions, wearable usability, feedback mechanisms, and goal-setting and sharing features. In addition, perceived effort and time investment regarding the proposed monitoring protocol were assessed using the NASA-TLX[d] [27] scale.

[a]TAM: technology acceptance model.

[b]SEQ: single ease question.

[c]CSAT: customer satisfaction score.

[d]NASA-TLX: NASA Task Load Index.

XSL•FO
RenderX

Data Analysis

Data analysis was structured around the study objectives, integrating quantitative statistical approaches and qualitative analyses. We describe the results based on 5 emerging themes.

The first theme is *system adoption and continued engagement*. This theme uses data from the end-of-study questionnaires (measurement C)—descriptive statistics were used to summarize central tendencies and variability. Independent-sample *t* tests (2-tailed), combined with the Cohen *d* to measure effect sizes, were used to assess group differences.

The second theme is *reasons for dropout: insights from patients who discontinued using the monitoring system*. Descriptive statistics were used to examine dropout reasons (measurement D), with frequencies and percentages calculated for each emerging theme to summarize reported reasons for discontinuation.

The third theme is *facilitators of and barriers to long-term engagement in self-monitoring of health behaviors*. Qualitative analyses involved an inductive coding approach (thematic analysis [28]) to identify key facilitators of and barriers to long-term self-monitoring of lifestyle behaviors drawing from all qualitative data sources (measurements A, B, D, and E). The final themes were structured into a coherent narrative to provide insights into facilitators of and barriers to long-term engagement.

The fourth theme is *effect of long-term lifestyle self-monitoring on lifestyle awareness and behavior change*. This theme aimed to examine the effects of long-term self-monitoring on lifestyle awareness and behavior change among completers. A structured analysis was conducted using interview data (measurement E) aligned with the self-regulation theory.

The fifth theme is *evaluation of usability of the lifestyle monitoring system design*. A structured approach was used to analyze patient-reported feedback on system usability, incorporating insights from all qualitative data sources (measurements A, B, D, and E). Responses were categorized based on predefined usability aspects, including system design and component usability (chatbot; health watch; dashboard, which primarily displays patients' lifestyle data, but also includes the quarterly questionnaires, a goal-setting module [optional], and a sharing functionality [optional]). The NASA Task Load Index scores (measurement E), assessing the perceived burden related to self-monitoring, were analyzed quantitatively to assess the perceived burden of the monitoring protocol across interview participants.

Where relevant, qualitative findings were integrated with study outcomes to explore potential links among reported experiences, dropout, and adherence data from the system. In addition, when applicable, a deductive coding approach was used to structure qualitative findings within predefined themes.

Results

Study Sample

Of the 100 patients included in this study (mean age 61.6, SD 10.4 y; n=88, 88% male; n=45, 45% having undergone PCI; n=55, 55% having undergone another intervention), 57 (57%) completed the full study year, and 43 (43%) dropped out of the study. The demographic and clinical characteristics of the total sample, completers, and dropouts are described elsewhere [22]. In summary, no substantial differences were found between the groups with regard to sociodemographic characteristics, clinical history, and lifestyle factors at baseline.

Throughout the study, data were collected using five different methods: (1) quarterly phone interviews (measurement A); a total of 217 phone interviews were completed (n=46, 21.2% with dropouts and n=171, 78.8% with completers); (2) quarterly questionnaires (measurement B); a total of 157 quarterly questionnaires were administered (n=12, 7.6% to dropouts and n=145, 92.4% to completers); (3) end-of-study questionnaires (measurement C); a total of 77 patients (n=54, 70% completers; n=23, 30% dropouts) completed the end-of-study questionnaire; (4) reasons for dropout; of the 43 patients who dropped out, one died during the study; of the remaining 42 patients, n=41 (98%) provided a reason for their withdrawal; and (5) evaluation interviews: a total of 98% (56/57) of the completers took part in the evaluation interview. One completer was unable to participate in the evaluation interview for medical reasons.

System Adoption and Continued Engagement (Theme 1)

To examine differences in factors contributing to system adoption and sustained engagement in relation to the technology acceptance model, we analyzed end-of-study questionnaire responses from both completers and dropouts focusing on perceived usefulness and ease of use as key determinants of engagement. Figure 4 presents box plots comparing perceived usefulness, ease of use, and engagement-related usability ratings between the 2 groups. Background shading reflects interpretative ranges for each scale—satisfaction and agreement scores (1-5) follow a customer satisfaction score framework, where higher values (green) indicate positive experiences, midrange scores (gray) represent neutral responses, and lower values (red) indicate dissatisfaction. Ease of use scores (1-7) are visualized based on the single ease question scale, where green represents high ease of use, yellow indicates moderate difficulty, and red reflects increasing difficulty levels.

With respect to perceived usefulness, completers were significantly more likely to perceive the platform as beneficial, reporting that the feedback helped them improve their health behaviors ($P<.001$; $d=0.75$).

Regarding ease of use (chatbot, health watch, and dashboard), completers rated system usability significantly higher than dropouts, with chatbot ease of use showing the largest difference between groups ($P=.02$; $d=0.61$). Similar trends were observed for the health watch and dashboard, indicating that ease of interaction played a role in sustained engagement.

Finally, regarding platform satisfaction, completers reported significantly higher satisfaction with the platform than dropouts ($P<.001$; $d=1.09$). In addition, perceived platform attractiveness was lower among dropouts ($P=.01$; $d=0.63$), suggesting that visual and design aspects might have influenced engagement.

XSL•FO
RenderX

Figure 4. Box plots comparing perceived usefulness, overall satisfaction, platform attractiveness, and ease of use (chatbot, health watch, and dashboard), between completers and dropouts.

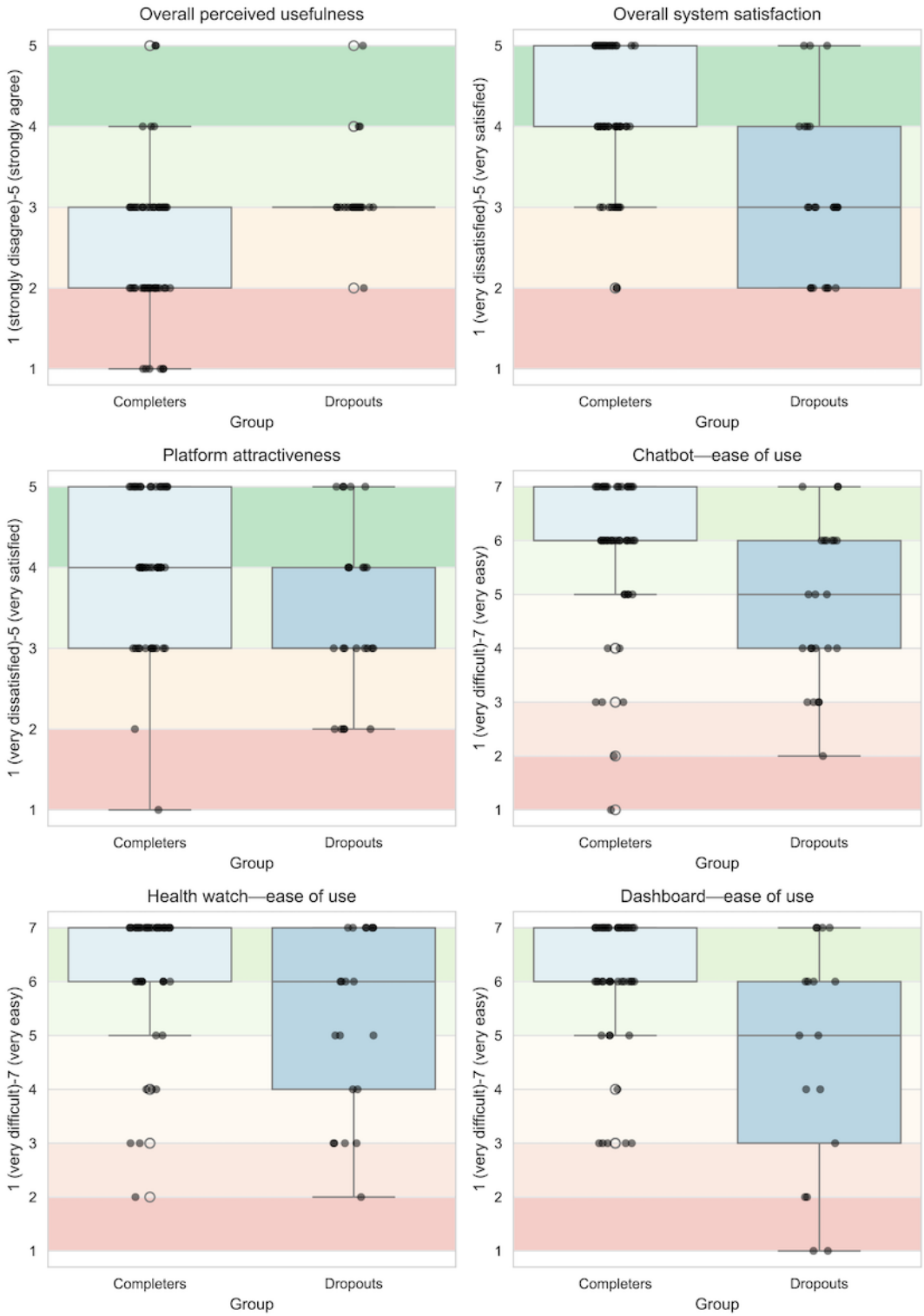

Reasons for Dropout: Insights From Patients Who Discontinued Use of the Monitoring System (Theme 2)

Overview

In total, 43% (43/100) of the patients dropped out of the clinical trial and discontinued use of the self-monitoring system. On the basis of the quantitative analyses, we previously reported that the main reasons for dropout were high participation burden (22/43, 51%), including technical issues with the health watch and dissatisfaction with the chatbot, and personal reasons (19/43, 44%) [22]. In addition, one patient withdrew for undisclosed reasons, and another dropped out due to death. In this analyses, we identified the main reasons that were explicitly linked to the use of the self-monitoring platform (34/43, 79%) to clarify the

XSL•FO
RenderX

underlying causes of discontinuation of use of the monitoring system.

The most frequently cited reason directly related to the self-monitoring platform was *burden of repeated self-reporting and dissatisfaction with the chatbot*, reported by 29% (10/34) of the patients. This was followed by *general usability issues across all modalities*, reported by 15% (5/34) of the patients. *Technical issues that led to frustration and demotivation* were cited by 9% (3/34) of the patients, and another 9% (3/34) identified *limited perceived usefulness* of the system as their primary reason for dropout. Among the personal reasons, 18% (6/34) of the patients reported that *self-monitoring caused mental stress*, whereas 15% (5/34) of the patients cited *low motivation*. The 6 key themes relevant to discontinuation of long-term monitoring of health behaviors are described in detail in the following sections.

Burden of Self-Reporting and Dissatisfaction

The chatbot was experienced as inefficient or requiring too much effort by 29% (10/34) of the patients, contributing to dissatisfaction and dropout. Patients found answering the same questions multiple times a day tedious, with some describing it as a routine that became a chore. A patient noted the following:

> *I'm also tired of having to answer the same questions three times a day, overkill.*

Difficulties in correcting input errors, excessive notifications, and the inability to reuse previous responses (eg, for repeated meals) added to their frustration. Some found the system too time-consuming for proper completion, whereas others struggled with its lack of flexibility, such as the inability to accurately log a specific diet. Over time, certain questions became irritating, further diminishing motivation to continue.

Usability Issues Across All Modalities

A subset of patients (5/34, 15%) described perceived poor usability of the self-monitoring platform across a multitude of components of the monitoring system. These usability concerns limited their engagement and ultimately contributed to dropout. One patient expressed it as follows:

> *If everything had worked properly, I would have worn the watch longer and used the chatbot more often.*

Technical Issues With the Health Watch Causing Frustration and Demotivation

Technical problems primarily involved synchronization issues and dissatisfaction with the health watch due to ergonomic concerns or malfunctions. A total of 9% (3/34) of the patients cited recurring synchronization failures as a key reason for dropout. These issues, unrelated to user error, stemmed from external system reliability problems, causing frustration and disengagement. One patient expressed their frustration stating the following:

> *I have not been able to sync for a week. These are all practical issues. My charger is acting up again. These things do not make it easy to keep up.*

Perceived Lack of Usefulness

A subset of patients (3/34, 9%) questioned the added value of the self-monitoring system for their personal health management. When perceived benefits were unclear or not directly experienced, motivation to continue declined, leading some to disengage from the study. Patients stated the following:

> *I was already aware of my lifestyle, so I did not need this.*
>
> *I realized I do not actually want to change my lifestyle.*

Self-Monitoring Causing Mental Stress

In total, 18% (6/34) of the patients reported feeling overwhelmed by the repeated reminders and notifications, which created pressure to engage and added to their daily stress levels. Some (2/34, 6%) became overly self-critical or anxious due to continuous self-monitoring. A patient noted the following:

> *It became an obsession; I kept checking my heart rate and panicked whenever I saw an unusual reading.*
>
> *I'm too focused on the medical aspects of the watch, and that does not benefit my health.*

Others (4/34, 12%) felt that the frequent prompts made their illness too present in daily life, making it harder to focus on living normally. One patient put it as follows:

> *There are too many reminders of my illness when I just want to live.*

Low Motivation

Low motivation was cited by 15% (5/34) of the patients, with some stating that they were not engaged from the start or frequently forgot about the study components. One patient mentioned the following:

> *I find it hard to motivate myself to do this. I forget the watch, the questionnaires, and I'm not good at keeping up with the chatbot.*

These responses suggest that personal motivation and perceived benefit played a role in dropout.

Facilitators of and Barriers to Long-Term Engagement in Self-Monitoring of Health Behaviors (Theme 3)

Understanding the factors that influence long-term engagement with digital self-monitoring is critical for optimizing usability and adherence. We analyzed open-text responses from quarterly phone interviews, questionnaires (quarterly and at the end of the study), reasons for dropout, and the evaluation interviews to identify key facilitators and barriers. Table 3 presents the overarching themes: usability and technical aspects, motivation and perceived usefulness, psychological factors, lifestyle behavior change, external and social influences, and customization and adaptability. The facilitators highlight features that encouraged continued participation, such as minimal-effort tracking, perceived health benefits, and habit formation. Conversely, the barriers reflect obstacles that led to discontinuation of use of the monitoring system, including technical frustrations, psychological burden, and lack of perceived relevance.

XSL•FO
RenderX

Table 3. Factors influencing adherence to the self-monitoring system.

Theme	Facilitators	Barriers
Usability and technical aspects	• Minimal effort required for wearables (passive data collection) • Perceived ease of use of some components (dashboard and step tracking) • System reliability when functioning properly • Integrated within the care platform, ensuring security without the need for multiple apps	• Technical frustrations (synchronization failures and inaccurate readings) • Repetitive chatbot questions • Issues with usability, reminder preferences, and navigation • Battery life concerns with wearables • Discomfort (eg, skin irritation) • Already owning a better functioning device that could not be integrated • Limited abilities of the watch
Motivation and perceived usefulness	• Perceived benefit of self-monitoring for health insights • Positive reinforcement from tracking progress • Goal setting motivation (for some patients)	• Low perceived personal relevance • No immediate feedback or tailored insights • Lack of motivation to engage with the system
Psychological factors	• Habit formation over time (routine integration) • Sense of accountability and self-awareness	• Psychological burden from continuous tracking (stress and anxiety) • Self-monitoring fatigue (overwhelming number of reminders) • Frustration with tracking unexpected fluctuations (eg, heart rate spikes)
Lifestyle behavior change	• Increased awareness of physical activity and nutrition • Some behavior modifications based on self-monitoring	• Minimal impact of self-monitoring on stress awareness and management • Limited engagement with lifestyle goal-setting features
External and social influences	• Support from the study team and research nurse • Social support from partner and social environment	• Competing priorities (work and personal life distractions) • Interruptions due to holidays, travel, or daily schedules • Preference for wearing a different-looking watch for special occasions
Customization and adaptability	• Flexibility in wearable use (passive tracking) • Structured chatbot reminders (eg, notifications) • The ability to configure and adjust the timing of chatbot prompt notifications to align with individual routines	• Limited chatbot personalization (eg, saving meals)

Habit formation and reminders supported chatbot adherence, whereas boredom, frustration, and forgetfulness were associated with discontinuation. The health watch provided a sense of security, but technical issues and inaccurate measurements sometimes led to frustration. Data reliability issues such as heart rate outliers, not registering certain activities, and missing step counts affected patient trust in the system. The practicality of adherence was influenced by technical challenges (eg, synchronization issues and notification failures) and usability concerns (eg, log-in barriers, skin irritation, and already owning another smartwatch). However, adherence was also supported by a sense of control over heart rate monitoring, the feeling of being observed and accountable, and the availability of study team support when needed.

Effect of Long-Term Lifestyle Self-Monitoring on Lifestyle Awareness and Behavior Change (Theme 4)

Overview

This section examines the experiences of study completers who engaged in self-monitoring of lifestyle behaviors using the monitoring system throughout the year following their cardiac intervention. Their postintervention recovery experiences were explored first. Understanding this process is crucial as it can influence their engagement with lifestyle changes and the sustainability of healthy behaviors over time. Furthermore, based on the evaluation interviews, the findings provide insights into how self-monitoring influenced lifestyle awareness and sustained behavior change framed within the context of the self-regulation theory.

Postintervention Recovery: Physical and Mental Adaptation in Daily Life

To better understand patients' engagement with the self-monitoring system, their postintervention recovery experiences were explored. While most reported full physical recovery, mental adjustment varied. To capture these differences, recovery experiences were categorized into physical and mental dimensions, as shown in Table 4. For physical recovery, 3 categories were defined: fully recovered (patients whose physical health and functionality returned to preintervention levels or who felt fully returned to normal), improving or partially recovered (those who made noticeable progress but were not yet fully recovered), and no improvement

XSL•FO
RenderX

or deterioration (patients with no improvement or a decline in physical condition). A parallel framework was applied to mental recovery, with categories labeled as accepting (patients who came to terms with their new reality and showed emotional stability), improving (those gradually adapting and showing mental health improvement), and struggling (those facing ongoing emotional or psychological challenges after the cardiac intervention). These categories provided insights into how patients managed their recovery and how it related to their engagement in self-monitoring.

Patients who reported full physical recovery (38/56, 68%) exhibited the highest engagement with both the chatbot and health watch, with only 31% (12/38) showing low chatbot adherence and 11% (4/38) showing low health watch adherence. In contrast, those who experienced no improvement or deterioration (6/56, 11%) were more likely to disengage, with 67% (4/6) showing low chatbot adherence and, similarly, 67% (4/6) showing low health watch adherence. A similar trend was observed in mental recovery—patients who fully accepted their condition (37/56, 66%) demonstrated more balanced adherence patterns, whereas those struggling (5/56, 9%) had the highest rates of low chatbot adherence (3/5, 60%), but lower rates of low health watch adherence (1/5, 20%). Notably, chatbot adherence appeared more affected by recovery status than health watch adherence, suggesting that passive data collection was easier to maintain than active chatbot interaction.

Patients' diagnosis and the type of cardiac intervention were associated with recovery experiences. Patients recovering from myocardial infarction, particularly those treated with PCI, generally reported better recovery and higher adherence. In contrast, those experiencing persistent dyspnea, often following valve surgery, showed lower adherence. While most patients (38/56, 68%) reported full physical recovery, many still noted reduced fitness, attributing it to aging and overall health decline. Slower recovery was more common among patients who had undergone valve surgery, whereas a small group (6/56, 11%) saw no improvement or deterioration, often adjusting their daily activities by slowing their pace and taking more breaks.

Mental recovery showed a more complex pattern of results (Table 4). Increased awareness of their condition led some patients to experience anxiety, heightened sensitivity to bodily sensations, and emotional strain. While most ultimately accepted their situation and adapted, others remained cautious, avoiding activities due to fear of overexertion. A small group (5/56, 9%) struggled emotionally, exhibiting reduced activity, forgetfulness, and changes in self-perception. Younger patients appeared to have more difficulty adjusting mentally than older patients, suggesting that age might play a role in psychological adaptation. Many described experiencing a "turning point" in their recovery, after which they moved toward acceptance, regaining confidence and emotional stability. While physical recovery was often seen as a more straightforward process, mental adaptation required more time and emotional adjustment.

XSL•FO
RenderX

Table 4. Overview of post–cardiac intervention experiences of the interviewed patients (completers).

Postintervention experience	Diagnosis	Intervention or procedure	Adherence level	Example quotes
Physical recovery				
Fully recovered: 38 (68%); mean age 62.9 (SD 9.7) y; 35 (92%) male	Most common: MI[a] (n=15, 40%); other diagnoses: AF[b], AP[c], atrial flutter, dyspnea, arrhythmia, and no symptoms	Most common: PCI[d] (n=17, 45%); other interventions or procedures: CABG[e], FFR[f], RFCA[g], valve surgery, and other	Chatbot: 12 (31%) low, 13 (34%) moderate, and 13 (34%) high; health watch: 4 (11%) low, 6 (42%) moderate, and 28 (74%) high	• "I do not see myself as a heart patient. It was a technical malfunction [related to the heart] and that has been resolved."
Improving or partially recovered: 12 (21%); mean age 61.4 (SD 11.9) y; 9 (75%) male	Most common: dyspnea (n=7, 58%); other diagnoses: AF and AP	Most common: valve surgery (n=7, 58%); other interventions or procedures: CABG, PCI, and RFCA	Chatbot: 4 (33%) low, 4 (33%) moderate, and 4 (33%) high; health watch: 1 (8%) low, 3 (25%) moderate, and 8 (67%) high	• "It feels like the brakes are on [walking with resistance] when I go for a walk."
No improvement or deterioration: 6 (11%); mean age 63.7 (SD 8.2) y; 5 (83%) male	Most common: AP (n=2, 33%) and MI (n=2, 33%); other diagnoses: dyspnea and no symptoms	Most common: PCI (n=4, 67%); other interventions or procedures: CABG and valve surgery	Chatbot: 4 (67%) low, 1 (17%) moderate, and 1 (17%) high; health watch: 1 (17%) low, 1 (17%) moderate, and 4 (67%) high	• "Before the treatment, I felt better than I do now. My physical fitness has deteriorated significantly, and that is unpleasant."
Mental recovery				
Accepting: 37 (66%); mean age 63.3 (SD 9.3) y; 33 (89%) male	Most common: MI (n=10, 27%); other diagnoses: AF, AP, atrial flutter, dyspnea, arrhythmia, and no symptoms	Most common: PCI (n=16, 43%); other interventions or procedures: CABG, FFR, RFCA, and valve surgery	Chatbot: 13 (35%) low, 12 (32%) moderate, and 12 (32%) high; health watch: 4 (11%) low, 6 (16%) moderate, and 27 (73%) high	• "I'm doing very well. I came here by bike. You can feel tired, but you just push through. Just keep your head up and keep going. I do not want to be part of the older generation yet."
Improving or adjusting: 14 (25%); mean age 61.9 (SD 12.8) y; 12 (86%) male	Most common: MI (n=5, 36%); other diagnoses: AF, AP, dyspnea, and no symptoms	Most common: PCI (n=6, 43%); other interventions or procedures: CABG, RFCA, valve surgery, and other	Chatbot: 4 (29%) low, 5 (36%) moderate, and 5 (36%) high; health watch: 1 (7%) low, 4 (29%) moderate, and 9 (64%) high	• "My daily life has not changed much, but mentally it has." • "I'm not playing team sports yet. There is a voice in my head saying, 'be careful.'"
Struggling: 5 (9%); mean age 60.0 (SD 4.74) y; 4 (80%) male	Most common: dyspnea (n=2, 40%) and MI (n=2, 40%); other diagnosis: AP	Most common: PCI (n=2, 40%) and valve surgery (n=2, 40%); other interventions or procedures: CABG	Chatbot: 3 (60%) low, 1 (20%) moderate, and 1 (20%) high; health watch: 1 (20%) low, 0 (0%) moderate, and 4 (80%) high	• "I do everything with less effort, because I fear that it might happen again."

[a]MI: myocardial infarction.

[b]AF: atrial fibrillation.

[c]AP: angina pectoris.

[d]PCI: percutaneous coronary intervention.

[e]CABG: coronary artery bypass graft.

[f]FFR: fractional flow reserve.

[g]RFCA: radiofrequency catheter ablation.

Lifestyle Awareness and Behavior Change Through Long-Term Self-Monitoring

In this section, we examine how patients engaged with their data, reflected on their progress, and translated insights into lasting lifestyle changes, aligning with self-regulation theory, which emphasizes the role of self-monitoring in enhancing awareness, setting goals, and adapting behavior. Table 5 summarizes patients' experiences with self-monitoring of key lifestyle behaviors, including physical activity, nutrition, sleep,

XSL•FO
RenderX

and stress management. For each lifestyle behavior, 2 areas of experiences are described: *awareness and reflection* and *behavior change and adaptation*.

The self-monitoring system primarily enhanced lifestyle awareness, particularly regarding physical activity, nutrition, and, to a lesser extent, sleep, making patients more conscious of their movement patterns, food choices and portion sizes, and sleep quality, respectively. However, sustained behavior change varied depending on individual motivation, preexisting habits, and the perceived need for change. While some patients increased their physical activity and adopted healthier eating habits, others used self-monitoring more for validation than for modification. Notably, stress self-monitoring had the least impact, with only 1 patient reporting increased awareness, and no evidence of behavior change in this area was found. The findings underscore the role of self-monitoring in self-regulation—while it can support reflection and goal-directed behavior, its effectiveness depends on individual engagement, personal circumstances, and the cognitive burden of active self-reporting.

Table 5. Impact of self-monitoring on lifestyle awareness and behavior change—patient-reported experiences and outcomes.

Lifestyle behavior and experiential dimension	Patients (n=56), n (%)	Impact of self-monitoring on lifestyle awareness	Example quotes
Physical activity			
Awareness and reflection	25 (45)	• Increased awareness of movement patterns • Monitored steps and heart rate	• "I was aware of my steps and movement, and that worked well for me."
Behavior change and adaptation	16 (29)	• Motivated to move more • Increased intentional activity choices • Sustained change in daily activity levels	• "If I was at 3000 steps, I made sure to add another 3000." • "Walking felt less intense over time, which was a good thing."
Nutritional intake			
Awareness and reflection	29 (52)	• Conscious of food choices and portion sizes • Awareness of nutritional balance	• "Through the chatbot, I became more aware of what I eat and how much I eat."
Behavior change and adaptation	16 (29)	• Reduced unhealthy food intake • More deliberate food selection • Long-term changes in eating habits, such as switching to healthier alternatives	• "I swapped bread for dairy, and I drink a little less wine now." • "I eat smaller portions because I have to log them; it stays in my mind."
Sleep quality			
Awareness and reflection	7 (13)	• Greater awareness of sleep patterns and rest quality	• "Seeing my sleep graph helped me realize when my rest was disturbed."
Behavior change and adaptation	4 (7)	• Adjusted bedtime habits to improve sleep duration	• "I started going to bed earlier and making sure I got enough sleep."
Mental stress levels			
Awareness and reflection	1 (2)	• Awareness of general mood states	• "When I have to fill in the mood states, I do think about it."
Behavior change and adaptation	0 (0)	—[a]	—

[a]Not applicable.

Evaluation of Usability of the Lifestyle Monitoring System Design (Theme 5)

Overview

Beyond usability and engagement, this study examined how specific design choices shaped user experiences. Table 6 summarizes each design choice, intended function, user feedback from completers and dropouts, and recommendations for improvement. These insights provide practical guidance for optimizing digital health interventions to enhance long-term adherence and usability.

User engagement varied, with accessibility, personalization, and technical reliability as key factors. On the basis of patient feedback, thorough onboarding to all system components is recommended to facilitate use and enhance overall user experience. The chatbot was motivational but felt repetitive and complex, requiring adaptive personalization. The health watch provided valuable insights but faced technical issues. Quarterly

XSL•FO
RenderX

assessments and information on the dashboard were informative ("I checked if I was being 'green' [green refers to a healthy food choice, as indicated by a green traffic light score]") but would benefit from a mobile version and, for the dashboard, more personalized real-time feedback. The goal-setting functionality had low engagement, requiring simpler guidance ("Setting a new goal was complicated"). The data sharing functionality was unused, needing clearer benefits and privacy assurances.

Table 6. Impact of design choices on usability and engagement.

Design feature (modality and engagement of completers)	Intended function	Positive user feedback	Negative user feedback	Effectiveness and recommendations
Chatbot interactions (mobile; mandatory component; engaged completers: n=57, 100%)	Support structured self-monitoring and lifestyle reflection	It was valued for its ease of use, usefulness, motivational impact, and ability to create awareness. In addition, habit formation, such as integrating it into daily routines (eg, notification timing aligned with personal schedule), was appreciated.	It was perceived as repetitive, complex (specifically with respect to nutrition), and difficult to correct, leading to reduced engagement over time. Users occasionally forgot to respond to prompts, and stress-related questions felt redundant.	Needs adaptive personalization to maintain engagement.
Health watch (device; mandatory component; engaged completers: n=57, 100%)	Passive lifestyle tracking (eg, heart rate, activity, and sleep)	It was valued for its heart rate insights, step tracking, motivational impact, sleek design, ease of use, and overall usefulness.	Frustration arose from technical issues, including synchronization failures, measurement inaccuracies, and short battery life.	Ensure technical reliability and offer alternative device options.
Quarterly lifestyle self-assessment via the dashboard (desktop web application; mandatory component; engaged completers: n=57, 100%)	Self-assessment tool for cardiovascular risk behaviors and providing feedback via the dashboard	It was valued for its immediate results and comprehensiveness, as well as the (sometimes strict) feedback and advice provided at the end of the questionnaires.	Limited accessibility due to desktop-only version. Questionnaires were lengthy but manageable.	Improve accessibility by offering mobile version.
Lifestyle dashboard (desktop web application; engaged completers: n=32, 58%)	Encourage behavior change through reflecting on self-monitoring insights	It was valued for useful nutritional advice (indication of healthiness per food item) and other health insights.	Limited accessibility (desktop only) and absence of real-time feedback reduced usability.	AI[a]-driven dynamic feedback could enhance engagement. Users expressed a desire to compare their health status with normative data for their patient group. Improving accessibility through a mobile app is recommended.
Goal-setting functionality (optional; desktop web application; engaged completers: n=4, 7%)	Allow users to set and track personal lifestyle goals	Gave motivation to set own goals or positively adjust goals (eg, increased steps).	Limited engagement due to lack of guidance and complexity.	Offer clearer instructions and integrate automated, personalized goal recommendations to improve usability.
Sharing functionality (optional; desktop web application; engaged completers: n=0, 0%)	Enable users to share collected lifestyle data via the platform with relatives or health care professionals	__[b]	No engagement due to lack of guidance, lack of perceived benefit, and privacy concerns.	Provide clearer use cases, privacy assurances, and user education on potential benefits.

[a]AI: artificial intelligence.

[b]Not applicable.

Perceived Effort of the Monitoring Protocol

The perceived effort of adhering to the proposed monitoring protocol was assessed through a NASA Task Load Index question on required effort among the interviewed patients (n=56). In total, 75% (42/56) of the patients indicated that only low effort was required to adhere to the protocol of the study, whereas 18% (10/56) of the patients indicated that it required medium effort, and 7% (4/56) indicated that it required high effort. Overall, the system's protocol was perceived as relatively easy to integrate into patients' daily lives because of its user-friendly interface, ease of integration into daily routines, immediate feedback, and accessibility on multiple devices. However, the perceived burden was significant for some patients, primarily related to the sustained effort required, the complexity and length of the questionnaires, occasional technical issues, and challenges in integrating certain components (such as the chatbot) into their routines. While the system was generally well received, these burdens impacted long-term engagement and satisfaction for some users.

XSL•FO
RenderX

Discussion

Principal Findings

Overview

This study provides insights into the use of a long-term digital lifestyle self-monitoring system in routine cardiac care. By evaluating patients' experiences over 1 year after a cardiac intervention, we identified key factors influencing engagement and dropout. Completers reported significantly higher perceived usefulness, ease of use, and satisfaction than dropouts, highlighting the importance of a positive user experience. Dropouts commonly cited dissatisfaction with the chatbot, technical issues with the health watch, and the burden of self-reporting, whereas completers found the protocol to be low effort. Facilitators of adherence included routine formation, structured reminders, and the ease of using the wearable. Barriers included cognitive load from repetitive chatbot questions, persistent technical problems, and physical or emotional recovery challenges. Patients experiencing poorer recovery were more likely to disengage. While self-monitoring raised awareness, especially regarding physical activity and nutrition, it had limited influence on actual behavior change, suggesting a need for improved system features and support.

Engagement and Dropout: Balancing Usefulness, Burden, and Technical Trust

Engagement with digital self-monitoring tools depends on perceived usefulness, usability, emotional impact, and technical reliability. In this study, patients who completed the monitoring year were more likely to find the system valuable, motivating, and easy to use, findings that aligned with the technology acceptance model and findings of previous research [29,30] emphasizing that perceived usefulness and ease of use are key to sustained engagement. In contrast, dropouts often perceived the system as redundant or frustrating, particularly when they already considered their lifestyle to be healthy and saw limited added value in self-monitoring.

Burden and psychological strain also contributed to dropout. Excessive notifications, repetitive chatbot dialogues, and the perceived obligation to self-report created monitoring fatigue. For some patients, frequent reminders of illness heightened stress or anxiety rather than offering support, underscoring previous research indicating that self-monitoring can be empowering for some but distressing for others [31,32].

Technical frustrations, particularly with health watch reliability and chatbot inefficiencies, were another major barrier to long-term engagement. Unlike interface usability issues, which users can often adapt to, recurring connectivity failures eroded trust and motivation. Consistent with previous research [33,34], our findings suggest that confidence in digital health tools depends on the availability of accurate, real-time data. When data are lost or inconsistent, user trust declines. These issues highlight the importance of seamless technical integration, intuitive troubleshooting, and transparent communication about system limitations supported by user education on how to interpret and act on health data.

Personalization and Design Considerations for Sustained Engagement

To support long-term use, future systems should adopt adaptive strategies such as personalized chatbot interactions, flexible reminder schedules, and predictive analytics. These approaches can help reduce monitoring fatigue and emotional resistance while enhancing perceived relevance, value, and trust in the system. Personalization in particular emerged as a promising strategy to increase engagement and reduce dropout. Evidence shows that health chatbots tailored to user profiles through personalized content, interfaces, or communication styles improve user satisfaction and dialogue quality [35-37]. This may include aligning messages with a patient's medical history, goals, or preferences; adapting interface elements such as avatars or notification frequency; and adjusting language to suit varying levels of health literacy or emotional tone. Granting users control over message timing or tone may further enhance comfort and reduce stress [38]. Personalized coaching interventions have shown positive effects on adherence across health domains [39]. Similarly, a chatbot could learn a user's preferred communication style [40] or detect behavioral lapses such as reduced physical activity and respond with tailored encouragement or barrier exploration. In practice, this might involve an initial intake procedure to understand the patient's context followed by ongoing adaptation based on user behavior over time. Personalization not only increases perceived relevance and satisfaction but may also enhance adherence and long-term engagement.

In addition to personalization, system design features played a critical role in shaping engagement in this study. Passive tracking was a strong facilitator, but recurring technical issues, particularly with the health watch, often undermined these benefits. While scheduled chatbot interactions helped some users establish healthy routines, others found the repetitive nature of dialogues demotivating. Furthermore, the lack of real-time feedback limited the platform's ability to maintain user interest. Previous research [41] suggests that dynamic, personalized feedback supports adherence, reinforcing the need for artificial intelligence–driven personalization that adapts system behavior to individual needs and preferences. These findings highlight the importance of thoughtful, user-centered design that integrates reliability, adaptability, and relevance to support long-term digital health engagement.

Post–Cardiac Intervention Recovery and Engagement

Although this was not a primary focus, our study showed that the patients' recovery trajectory influenced engagement. Most completers reported full physical recovery especially after PCI, whereas others faced ongoing symptoms or limited improvement. Interestingly, rapid recovery (eg, after PCI) can sometimes lead to complacency as patients may perceive the intervention as a definitive fix, reducing motivation for continued self-care or cardiac rehabilitation [42]. In contrast, those undergoing more invasive procedures such as CABG often face a more gradual and challenging recovery, which can increase engagement with structured rehabilitation programs to rebuild physical capacity [43]. Nevertheless, we found no significant differences in adherence by intervention type. In our

XSL•FO
RenderX

study, recovery experiences rather than the procedure itself were more strongly associated with engagement. These differences suggest that perceived need for continued care influences engagement with digital tools.

Not only physical but also mental recovery varied across patients—while many recovered well, some struggled emotionally. Differences in psychological recovery have also been reported to influence engagement. Research consistently indicates that emotional factors such as stress, anxiety, and perceived burden negatively impact long-term adherence to lifestyle changes [39,44]. More specifically, patients experiencing postintervention anxiety may avoid self-monitoring to reduce stress, whereas those who reach a state of acceptance are more likely to perceive monitoring as beneficial rather than burdensome [45]. Thus, building on the earlier personalization theme, support strategies should also address individual physical and psychological recovery trajectories. Adaptive interventions such as flexible logging schedules and motivational feedback can support users through difficult periods and improve adherence. Tailoring the system to the patient's recovery stage, emotional readiness, and capacity allows for more individually tailored, effective support in cardiac rehabilitation.

Self-Monitoring and Behavior Change: A Self-Regulation Perspective

Self-monitoring has been shown to be a tool for self-regulation [46], increasing patients' awareness of their lifestyle habits, particularly regarding physical activity and nutrition. In our study, completers indeed reported more positive behavior change in these areas. However, the system had limited impact on stress management. Patients under psychological strain often saw tracking as burdensome, which aligns with our previous findings [22] showing that depressive symptoms and lower mental quality of life predict low adherence to self-monitoring interventions. This is also consistent with findings suggesting the necessity of incorporating psychological support and adaptive personalization to maintain engagement among populations vulnerable to emotional distress [32,47]. This reflects the need for integrated psychological support and mental health–sensitive design in digital interventions. Features such as goal setting, timely feedback, and mental health tailoring can foster more sustainable behavior change.

Limitations

One key limitation is the absence of more thorough dropout perspectives as interviews were only conducted with completers. Including those who disengaged could have enriched the understanding of barriers such as low perceived benefit, poor usability, or external life factors. Future research should aim to include interviews with both completers and dropouts to gain a more comprehensive understanding of the factors influencing participation and sustained use.

Another limitation is the lack of audio recordings or verbatim transcripts as responses were documented in real time based on predefined questions by a single interviewer. The absence of verbatim records may have limited the depth of qualitative insights and may have hindered the ability to capture nuanced expressions or spontaneous remarks. While the approach facilitated efficient data processing and reduced the potential for subjective interpretation during transcription, it may have introduced interviewer bias. In addition, the structured interview approach may have constrained the emergence of unexpected insights, whereas recall bias could have influenced participants' reflections on experiences that occurred weeks or months earlier. Moreover, participants may have been influenced by social desirability bias.

The study sample may not fully represent the broader population of patients with cardiac disease as participants were relatively homogeneous. As a result, the findings may have limited generalizability to other groups, such as those with lower digital literacy or differing health backgrounds.

Conclusions

This study identified critical factors that influence long-term use of digital lifestyle self-monitoring systems in cardiac care. Completers found the platform more useful, more user-friendly, and less effortful than dropouts, who cited chatbot dissatisfaction and health watch technical issues. While self-monitoring increased awareness, particularly regarding physical activity and nutrition, its impact on behavior change was limited. To maintain engagement, systems must reduce burden; support routine formation; and offer personalized, psychologically attuned experiences. Ultimately, by improving usability, minimizing complexity and technical barriers, enhancing personalization, and providing psychological support, digital health tools can be optimized for long-term engagement, potentially promoting sustained positive health behavior change in cardiac rehabilitation and beyond.

Acknowledgments

The authors are grateful for the time and effort of all patients who participated in the CARE-ON clinical trial and provided the data used in this study. Funding for MG was provided by the Dutch Heart Foundation (grant 2019B015). The authors used the generative artificial intelligence tool ChatGPT by OpenAI [48] only for language editing of the Discussion section of this manuscript.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to inclusion of sensitive information, but they are available from the corresponding author upon reasonable request.

XSL•FO
RenderX

Authors' Contributions

MG led the conceptualization, with support from NT-VL, YL, HK, and WJK. Data curation was performed by MG, with NT-VL providing support. MG conducted the formal analysis. HK acquired funding, with WJK and YL providing support. MG and NT-VL conducted the investigation equally, with MG taking the lead. MG developed the methodology, supported by YL. Project administration was led by MG, with equal contributions from NT-VL and support from HK. Resources were provided by NT-VL and MG equally. Supervision was provided equally by WJK, YL, and HK. Validation was led by YL. MG was responsible for visualization. The original draft of the manuscript was written by MG, with support from YL. Review and editing of the manuscript were led by YL, with support from HK and WJK.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Sample Flows of the Rule-Based Chatbot.
[DOCX File , 23 KB - formative_v9i1e76978_app1.docx]

Multimedia Appendix 2
The Reporting of Studies Conducted Using Observational Routinely Collected Health Data (RECORD) statement (extension of the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement) Checklist.
[DOCX File , 27 KB - formative_v9i1e76978_app2.docx]

Multimedia Appendix 3
Structured Interview Guide.
[DOCX File , 19 KB - formative_v9i1e76978_app3.docx]

References

1. Cardiovascular diseases (CVDs). World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) [accessed 2024-09-02]
2. Timmis A, Aboyans V, Vardas P, Townsend N, Torbica A, Kavousi M, ESC National Cardiac Societies. European Society of Cardiology: the 2023 Atlas of Cardiovascular Disease Statistics. Eur Heart J 2024 Oct 07;45(38):4019-4062. [doi: 10.1093/eurheartj/ehae466] [Medline: 39189413]
3. Dimagli A, Spadaccio C, Myers A, Demetres M, Rademaker-Havinga T, Stone GW, et al. Quality of life after percutaneous coronary intervention versus coronary artery bypass grafting. J Am Heart Assoc 2023 Nov 21;12(22):e030069 [FREE Full text] [doi: 10.1161/JAHA.123.030069] [Medline: 37982221]
4. Straiton N, Jin K, Bhindi R, Gallagher R. Functional capacity and health-related quality of life outcomes post transcatheter aortic valve replacement: a systematic review and meta-analysis. Age Ageing 2018 May 01;47(3):478-482. [doi: 10.1093/ageing/afx203] [Medline: 29377982]
5. Gupta D, Vijgen J, Potter TD, Scherr D, Van Herendael H, Knecht S, et al. Quality of life and healthcare utilisation improvements after atrial fibrillation ablation. Heart 2021 Aug;107(16):1296-1302 [FREE Full text] [doi: 10.1136/heartjnl-2020-318676] [Medline: 33952593]
6. Booth JN, Levitan EB, Brown TM, Farkouh ME, Safford MM, Muntner P. Effect of sustaining lifestyle modifications (nonsmoking, weight reduction, physical activity, and mediterranean diet) after healing of myocardial infarction, percutaneous intervention, or coronary bypass (from the REasons for Geographic and Racial Differences in Stroke Study). Am J Cardiol 2014 Jun 15;113(12):1933-1940 [FREE Full text] [doi: 10.1016/j.amjcard.2014.03.033] [Medline: 24793668]
7. Chow CK, Jolly S, Rao-Melacini P, Fox KA, Anand SS, Yusuf S. Association of diet, exercise, and smoking modification with risk of early cardiovascular events after acute coronary syndromes. Circulation 2010 Feb 16;121(6):750-758. [doi: 10.1161/circulationaha.109.891523]
8. Giannuzzi P, Temporelli PL, Marchioli R, Maggioni AP, Balestroni G, Ceci V, GOSPEL Investigators. Global secondary prevention strategies to limit event recurrence after myocardial infarction: results of the GOSPEL study, a multicenter, randomized controlled trial from the Italian Cardiac Rehabilitation Network. Arch Intern Med 2008 Nov 10;168(20):2194-2204. [doi: 10.1001/archinte.168.20.2194] [Medline: 19001195]
9. De Bacquer D, Astin F, Kotseva K, Pogosova N, De Smedt D, De Backer G, EUROASPIRE IVV surveys of the European Observational Research Programme of the European Society of Cardiology. Poor adherence to lifestyle recommendations in patients with coronary heart disease: results from the EUROASPIRE surveys. Eur J Prev Cardiol 2022 Mar 11;29(2):383-395. [doi: 10.1093/eurjpc/zwab115] [Medline: 34293121]

XSL•FO
RenderX

10. Turan Kavradim S, Özer Z, Boz Ö. Effectiveness of telehealth interventions as a part of secondary prevention in coronary artery disease: a systematic review and meta-analysis. Scand J Caring Sci 2020 Sep 20;34(3):585-603. [doi: 10.1111/scs.12785] [Medline: 31747080]
11. Wongvibulsin S, Habeos EE, Huynh PP, Xun H, Shan R, Porosnicu Rodriguez KA, et al. Digital health interventions for cardiac rehabilitation: systematic literature review. J Med Internet Res 2021 Feb 08;23(2):e18773 [FREE Full text] [doi: 10.2196/18773] [Medline: 33555259]
12. Meyerowitz-Katz G, Ravi S, Arnolda L, Feng X, Maberly G, Astell-Burt T. Rates of attrition and dropout in app-based interventions for chronic disease: systematic review and meta-analysis. J Med Internet Res 2020 Sep 29;22(9):e20283 [FREE Full text] [doi: 10.2196/20283] [Medline: 32990635]
13. Kanfer FH, Hagerman S. A model of self-regulation. In: Halisch F, Kuhl J, editors. Motiv Intent Volition. Heidelberg, Germany: Springer; 1987:293-307.
14. Whitelaw S, Pellegrini DM, Mamas MA, Cowie M, Van Spall HG. Barriers and facilitators of the uptake of digital health technology in cardiovascular care: a systematic scoping review. Eur Heart J Digit Health 2021 Mar;2(1):62-74 [FREE Full text] [doi: 10.1093/ehjdh/ztab005] [Medline: 34048508]
15. Singh B, Olds T, Brinsley J, Dumuid D, Virgara R, Matricciani L, et al. Systematic review and meta-analysis of the effectiveness of chatbots on lifestyle behaviours. NPJ Digit Med 2023 Jun 23;6(1):118 [FREE Full text] [doi: 10.1038/s41746-023-00856-1] [Medline: 37353578]
16. Ambrosetti M, Abreu A, Corrà U, Davos CH, Hansen D, Frederix I, et al. Secondary prevention through comprehensive cardiovascular rehabilitation: from knowledge to implementation. 2020 update. A position paper from the Secondary Prevention and Rehabilitation Section of the European Association of Preventive Cardiology. Eur J Prev Cardiol 2021 May 14;28(5):460-495. [doi: 10.1177/2047487320913379] [Medline: 33611446]
17. Goevaerts WF, Tenbült-van Limpt NCCW, Kop WJ, Birk MV, Liu Y, Brouwers RWM, et al. Adherence to a lifestyle monitoring system in patients with heart disease: protocol for the care-on prospective observational trial. BMC Cardiovasc Disord 2023 Apr 17;23(1):196 [FREE Full text] [doi: 10.1186/s12872-023-03222-x] [Medline: 37069506]
18. Carney CE, Buysse DJ, Ancoli-Israel S, Edinger JD, Krystal AD, Lichstein KL, et al. The consensus sleep diary: standardizing prospective sleep self-monitoring. Sleep 2012 Feb 01;35(2):287-302 [FREE Full text] [doi: 10.5665/sleep.1642] [Medline: 22294820]
19. Tran V. Positive Affect Negative Affect Scale (PANAS). In: Gellman MD, editor. Encyclopedia of Behavioral Medicine. New York, NY: Springer; 2013:1708-1709.
20. Dutch Food Composition Database (NEVO). RIVM. URL: https://www.rivm.nl/en/dutch-food-composition-database [accessed 2025-02-11]
21. Liu Y, Goevaerts WF, Birk MV, Kemps H, Lu Y. Development of a conversational dietary assessment tool for cardiovascular patients. In: Proceedings of the 9th IFIP WG 13.2 International Working Conference on Human-Centered Software Engineering. 2022 Presented at: HCSE '22; August 24-26, 2022; Eindhoven, The Netherlands p. 179-190 URL: https://link.springer.com/chapter/10.1007/978-3-031-14785-2_12 [doi: 10.1007/978-3-031-14785-2_12]
22. Goevaerts WF, Tenbült-Van Limpt NCCW, Brouwers RWM, Goossen SA, Liu Y, Birk MV, et al. Adherence to a lifestyle monitoring system in cardiac care: results of the Care-On clinical trial. Eur J Prev Cardiol 2025 May;32(Supplement_1):494 [conference abstract; updated results in manuscript under review]. [doi: 10.1093/eurjpc/zwaf236.494]
23. Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, RECORD Working Committee. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. PLoS Med 2015 Oct;12(10):e1001885 [FREE Full text] [doi: 10.1371/journal.pmed.1001885] [Medline: 26440803]
24. Park YE, Tak YW, Kim I, Lee HJ, Lee JB, Lee JW, et al. User experience and extended technology acceptance model in commercial health care app usage among patients with cancer: mixed methods study. J Med Internet Res 2024 Dec 18;26:e55176 [FREE Full text] [doi: 10.2196/55176] [Medline: 39693615]
25. Sauro J, Dumas JS. Comparison of three one-question, post-task usability questionnaires. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2009 Presented at: CHI '09; April 4-9, 2009; Boston, MA p. 1599-1608 URL: https://dl.acm.org/doi/10.1145/1518701.1518946 [doi: 10.1145/1518701.1518946]
26. What is customer satisfaction score? (+ how to measure CSAT). Zendesk. URL: https://www.zendesk.nl/blog/customer-satisfaction-score/ [accessed 2025-01-06]
27. Braarud P. Investigating the validity of subjective workload rating (NASA TLX) and subjective situation awareness rating (SART) for cognitively complex human–machine work. Int J Ind Ergon 2021 Nov;86:103233. [doi: 10.1016/j.ergon.2021.103233]
28. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
29. Brewer LC, Kumbamu A, Smith C, Jenkins S, Jones C, Hayes SN, et al. A cardiovascular health and wellness mobile health intervention among church-going African Americans: formative evaluation of the FAITH! app. JMIR Form Res 2020 Nov 17;4(11):e21450 [FREE Full text] [doi: 10.2196/21450] [Medline: 33200999]

XSL•FO
RenderX

30. Sengupta A, Beckie T, Dutta K, Dey A, Chellappan S. A mobile health intervention system for women with coronary heart disease: usability study. JMIR Form Res 2020 Jun 03;4(6):e16420 [FREE Full text] [doi: 10.2196/16420] [Medline: 32348270]
31. Kelley C, Lee B, Wilcox L. Self-tracking for mental wellness: understanding expert perspectives and student experiences. Proc SIGCHI Conf Hum Factor Comput Syst 2017 May 02;2017:629-641 [FREE Full text] [doi: 10.1145/3025453.3025750] [Medline: 28920106]
32. Cole-Lewis H, Ezeanochie N, Turgiss J. Understanding health behavior technology engagement: pathway to measuring digital behavior change interventions. JMIR Form Res 2019 Oct 10;3(4):e14052 [FREE Full text] [doi: 10.2196/14052] [Medline: 31603427]
33. Freene N, van Berlo S, McManus M, Mair T, Davey R. A behavioral change smartphone app and program (ToDo-CR) to decrease sedentary behavior in cardiac rehabilitation participants: prospective feasibility cohort study. JMIR Form Res 2020 Nov 03;4(11):e17359 [FREE Full text] [doi: 10.2196/17359] [Medline: 33141091]
34. Brewer LC, Jones C, Slusser JP, Pasha M, Lalika M, Chacon M, et al. mHealth intervention for promoting hypertension self-management among African American patients receiving care at a community health center: formative evaluation of the FAITH! hypertension app. JMIR Form Res 2023 Jun 16;7:e45061 [FREE Full text] [doi: 10.2196/45061] [Medline: 37115658]
35. Kocaballi AB, Berkovsky S, Quiroz JC, Laranjo L, Tong HL, Rezazadegan D, et al. The personalization of conversational agents in health care: systematic review. J Med Internet Res 2019 Nov 07;21(11):e15360 [FREE Full text] [doi: 10.2196/15360] [Medline: 31697237]
36. Miller HN, Voils CI, Cronin KA, Jeanes E, Hawley J, Porter LS, et al. A method to deliver automated and tailored intervention content: 24-month clinical trial. JMIR Form Res 2022 Sep 06;6(9):e38262 [FREE Full text] [doi: 10.2196/38262] [Medline: 36066936]
37. Hungerbuehler I, Daley K, Cavanagh K, Garcia Claro H, Kapps M. Chatbot-based assessment of employees' mental health: design process and pilot implementation. JMIR Form Res 2021 Apr 21;5(4):e21678 [FREE Full text] [doi: 10.2196/21678] [Medline: 33881403]
38. Fan H, Poole MS. What is personalization? Perspectives on the design and implementation of personalization in information systems. J Organ Comput Electron Commer 2006 Jan;16(3-4):179-202 [FREE Full text] [doi: 10.1080/10919392.2006.9681199]
39. Jakob R, Harperink S, Rudolf AM, Fleisch E, Haug S, Mair JL, et al. Factors influencing adherence to mHealth apps for prevention or management of noncommunicable diseases: systematic review. J Med Internet Res 2022 May 25;24(5):e35371 [FREE Full text] [doi: 10.2196/35371] [Medline: 35612886]
40. Aggarwal A, Tam CC, Wu D, Li X, Qiao S. Artificial intelligence-based chatbots for promoting health behavioral changes: systematic review. J Med Internet Res 2023 Feb 24;25:e40789 [FREE Full text] [doi: 10.2196/40789] [Medline: 36826990]
41. Olawade DB, Wada OZ, Odetayo A, David-Olawade AC, Asaolu F, Eberhardt J. Enhancing mental health with artificial intelligence: current trends and future prospects. J Med Surg Public Health 2024 Aug;3:100099. [doi: 10.1016/j.glmedi.2024.100099]
42. Liu Z, Zhang G, Liang X, Qin D. Effect of a patient health engagement (PHE) model on rehabilitation participation in patients with acute myocardial infarction after PCI: a study protocol for a randomized controlled trial. Trials 2024 Nov 22;25(1):786 [FREE Full text] [doi: 10.1186/s13063-024-08643-3] [Medline: 39574197]
43. Douma ER, Kop WJ, Kupper N. Associations between psychological factors and adherence to health behaviors after percutaneous coronary intervention: the role of cardiac rehabilitation. Ann Behav Med 2024 Apr 11;58(5):328-340 [FREE Full text] [doi: 10.1093/abm/kaae008] [Medline: 38431284]
44. Tadas S, Coyle D. Barriers to and facilitators of technology in cardiac rehabilitation and self-management: systematic qualitative grounded theory review. J Med Internet Res 2020 Nov 11;22(11):e18025 [FREE Full text] [doi: 10.2196/18025] [Medline: 33174847]
45. Vaccarino V, Badimon L, Bremner JD, Cenko E, Cubedo J, Dorobantu M, ESC Scientific Document Group Reviewers. Depression and coronary heart disease: 2018 position paper of the ESC working group on coronary pathophysiology and microcirculation. Eur Heart J 2020 May 01;41(17):1687-1696 [FREE Full text] [doi: 10.1093/eurheartj/ehy913] [Medline: 30698764]
46. de Ridder DT, de Wit JB. Self-regulation in health behavior: concepts, theories, and central issues. In: de Ridder DT, de Wit JB, editors. Self‐Regulation in Health Behavior. Hoboken, NJ: John Wiley & Sons; 2006:1-23.
47. Egilsson E, Bjarnason R, Njardvik U. Usage and weekly attrition in a smartphone-based health behavior intervention for adolescents: pilot randomized controlled trial. JMIR Form Res 2021 Feb 17;5(2):e21432 [FREE Full text] [doi: 10.2196/21432] [Medline: 33481750]
48. ChatGPT. URL: https://chatgpt.com [accessed 2025-06-25]

Abbreviations

CABG: coronary artery bypass graft

XSL•FO
RenderX

CARE-ON: Cardiovascular Research Opting for New Applications
PCI: percutaneous coronary intervention

Edited by A Mavragani; submitted 13.05.25; peer-reviewed by L Magoun, K Coninx; comments to author 10.06.25; revised version received 01.07.25; accepted 03.07.25; published 17.09.25.

Please cite as:
Goevaerts M, Tenbült - Van Limpt N, Kop WJ, Kemps H, Lu Y
Patient-Reported Experiences With Long-Term Lifestyle Self-Monitoring in Heart Disease: Mixed Methods Study
JMIR Form Res 2025;9:e76978
URL: https://formative.jmir.org/2025/1/e76978
doi:10.2196/76978
PMID:

©Mayra Goevaerts, Nicole Tenbült - Van Limpt, Willem J Kop, Hareld Kemps, Yuan Lu. Originally published in JMIR Formative Research (https://formative.jmir.org), 17.09.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Stakeholder Perspectives on mHealth Technologies to Prevent Sitting-Acquired Pressure Injuries in Long-Term Care Facilities: Mixed Methods Study

Tamara Vos-Draper[1], OTR, PhD; Erin Vinoski Thomas[2], PhD; Emily Graybill[2], PhD; Melissa Morrow[3], PhD; Kathleen A Jordan[4], BS; Pamela R Manley[5], PhD; Sharon E Sonenblum[4], PhD

[1]Program of Occupational Therapy, College of Pharmacy, University of Minnesota, Minneapolis, MN, United States

[2]Center for Leadership in Disability, School of Public Health, Georgia State University, Atlanta, GA, United States

[3]School of Health Professions, Center for Health Promotion, Performance, and Rehabilitation Research, University of Texas Medical Branch, Galveston, TX, United States

[4]Nell Hodgson Woodruff School of Nursing, Emory University, Atlanta, GA, United States

[5]School of Public Health, Georgia State University, Georgia State University, Atlanta, GA, United States

Corresponding Author:
Tamara Vos-Draper, OTR, PhD
Program of Occupational Therapy
College of Pharmacy
University of Minnesota
321 Johnston Hall
101 Pleasant St. S.E.
Minneapolis, MN, 55411
United States
Phone: 1 507 533 5125
Email: vosdraper@gmail.com

Abstract

Background: Adults with Alzheimer disease (AD) or Alzheimer disease and related dementias (ADRD) who require a wheelchair to accommodate disease-associated decline in mobility are at elevated risk for pressure injuries. More than half of residents in long-term care (LTC) facilities in the United States experience AD or ADRD. In LTC facilities, bed-based technologies exist to facilitate pressure injury prevention efforts, but similar technologies have not yet been widely evaluated to address sitting-related pressure injuries.

Objective: This study aimed to determine preliminary design inputs from care providers for technology to address sitting-related pressure injury prevention in LTC settings. Specifically, we sought to (1) understand the types and use of sitting-related equipment used in LTC for residents with AD or ADRD, (2) identify challenges faced by nurses and other caregivers when repositioning seated residents, and (3) understand care provider preferences for features of future sitting-related feedback technologies designed to facilitate effective and timely repositioning.

Methods: Surveys (n=30) and semistructured interviews (n=9) of administrative and direct care providers in LTC facilities were administered. Survey results were summarized, and we used thematic qualitative analysis of interview responses to develop themes around challenges experienced by care providers and their perceptions about how technologies could facilitate the prevention of sitting-related pressure injuries.

Results: Survey respondents endorsed using many sitting surfaces for LTC residents with memory loss, such as padded reclining chairs, bedside or dining chairs, and wheelchairs with cushions. All indicated that shared equipment is provided by the facility, and 43% of respondents reported having access to a seating specialist at their facility. Sitting time was typically up to 12 hours per day. Themes related to pressure injury prevention in the LTC context, specific to those with memory loss, included (1) barriers to repositioning seated residents vary with the degree of memory loss, (2) care providers are aware of guidelines and policies around the 2-hour repositioning schedule, and (3) care providers are interested in technologies that have relative value over added burden. Care providers expressed interest in mobile health (mHealth) technologies that provide automatic repositioning in later stages of memory loss, delivery of cues for residents with mild memory loss to encourage independent repositioning, and tools to monitor resident sitting and pressure-related outcomes.

XSL•FO
RenderX

Conclusions: These findings highlight the complexity of addressing the repositioning needs of seated LTC residents with AD or ADRD using mHealth technologies due to changes as the disease progresses. mHealth technologies should encourage more independence by residents experiencing milder memory loss, with increasing automaticity in repositioning residents in later stages. Both approaches could potentially minimize care provider burden in repositioning seated residents throughout the day. Design, development, and implementation of technologies should carefully weigh benefit versus burden to care providers and residents and continue to engage with them for feedback as development progresses.

(JMIR Form Res 2025;9:e59590) doi:10.2196/59590

KEYWORDS

pressure injuries; Alzheimer disease; dementia; wheelchair use; user-centered design; long-term care; prevention; assistive technology; monitoring; feedback

Introduction

Background

Pressure injuries in long-term care (LTC) settings occur at high rates (3.4%-32.4%), and people with mobility decline and memory loss related to Alzheimer disease (AD) or Alzheimer disease and related dementias (ADRD) are among the populations at highest risk [1,2]. Furthermore, 58% of LTC facility residents in the United States have AD or ADRD [3]. The International Pressure Injury Prevention Guidelines [4] recommend repositioning at least every 2 hours in bed and while sitting to reduce the impact of prolonged pressure, which contributes to skin and tissue damage [4]. Health care facilities in the United States are motivated to reduce the prevalence of pressure injuries to meet standards of care and because their payments can be reduced if pressure injuries develop during the care they provide [5]. In the United States, the Centers for Medicare and Medicaid require systematic reporting of pressure injuries in LTC facilities, providing motivation to implement strategies to minimize their occurrence. To facilitate adherence to turning residents at least every 2 hours in bed, some facilities have implemented bed-based technologies that either automatically adjust the surface to manage pressure or provide auditory or visual feedback to caregivers [6-8]. However, bed-based technologies do not address sitting-related pressure from wheelchair use. As residents with AD or ADRD experience increasing mobility loss, they sit for longer periods, putting them at high risk for developing pressure injuries.

While some mobile health (mHealth) app options are now available to wheelchair users that provide cues or alerts to manage pressure, these technologies target independent wheelchair users living in the community [9,10]. Little is known about the use of monitoring or feedback technologies to facilitate caregiver efforts to reposition seated residents in residential care centers. Technologies to facilitate caregiving in the AD and ADRD population tend to focus more on case management, caregiver support, care pathways, or monitoring and tracking of the person with AD or ADRD for safety related to falls or wandering, and do not specifically address pressure injury risk [11,12]. This study is part of a multidimensional project spanning the development and usability of 2 mHealth systems that monitor and provide real-time feedback with the goal of increasing in-seat movement to prevent sitting-acquired pressure injuries [13-15]. One type of feedback under development focuses on a display showing pressure distribution at the seat-cushion interface and uses cues to remind the user to redistribute pressure throughout the day [13]. Pressure injury prevention is mutifaceted [4] and these technologies target user-driven actions to mitigate pressure at the seat interface. Pressure mapping at the seat interface is commonly used by occupational therapists (OTs) and physical therapists (PTs) during seating assessments and provides color-based depictions of the magnitude of pressure between a person and their sitting surface across a grid of at least 256 sensing areas [16]. The display of pressure distribution appears similar to a weather map and depicts areas of high pressure with yellow or red colors versus blue or green for areas of lower pressure. The development process for enhancing these mHealth systems has incorporated user-centered design [17,18] and their development was guided by behavior change frameworks [19].

The long-term goal of this project is to potentially adapt these mHealth systems for caregiver use in a setting where pressure injuries continue to be prevalent and to engage key stakeholders early in the design process to better understand their needs. Key stakeholders include LTC care providers who might have a direct impact on sitting, repositioning, and managing pressure injury prevention, such as certified nursing assistants (CNAs) and other nursing staff, and OTs and PTs. Findings from this research will identify important design criteria for developing and implementing such systems in an LTC setting, specifically to address the unique needs of residents with AD or ADRD.

Research Questions

Research questions centered on understanding caregiver perspectives on pressure injury prevention during sitting including the use of feedback technologies:

1. In LTC settings, what factors impact the provision of care for preventing pressure injuries in wheelchair users with AD or ADRD?
2. What are direct care provider perceptions about the use of feedback technologies to facilitate repositioning residents in their wheelchairs?
3. What considerations are there for the design and development of wheelchair-based feedback technologies for caregiver use in the LTC setting?

Methods

This cross-sectional study included a combination of survey and semistructured interviews of a small group of stakeholders who provide or oversee care in LTC settings. Stakeholders

XSL•FO
RenderX

included nurses, CNAs, social workers, administrators, OTs, and PTs.

Ethical Considerations

The study activities were reviewed and received approval from the appropriate institutional review boards (IRBs) at Georgia State (H22270), Georgia Institute of Technology (H21280), and the University of Minnesota (STUDY00016896). Each IRB determined that while the study qualified for exempt status, it met all institutional ethical standards and complied with relevant federal regulations for conducting the surveys and interviews. The 21-item checklist for the Standards for Reporting Qualitative Research was used for reporting guidance [20]. Informed consent was waived for respondents, and an information sheet describing the study and how data would be deidentified and used was provided and agreed to by participants before completing the survey or interview. All participant identifying information was stored in secure IRB-approved locations, accessible only by approved study staff. Data were deidentified before analysis and reporting of results. Survey participants were not compensated. Interview participants received compensation for their time.

Survey Data Collection

Design and Administration

Two separate surveys were developed for this study. The first (survey A) was developed by researchers with expertise in wheelchair seating and pressure injury prevention from the Georgia Institute of Technology and Georgia State University, in collaboration with the University of Minnesota. The survey questions related to seating and mobility equipment used by residents in LTC, including those with AD or ADRD. The web-based survey was distributed via a web link using Qualtrics, a survey platform. Participants were asked to respond to questions related to characteristics of residents and care providers in their LTC facilities. For full details, refer to survey A in Multimedia Appendix 1.

The second survey (survey B) was developed by study team members with expertise in wheelchair seating and pressure injury prevention at the University of Minnesota in collaboration with partners at Georgia Institute of Technology, following the administration of survey A. The questions from survey A were refined to gather additional information about knowledge of feedback technologies for preventing pressure injuries, specifically related to awareness and receptiveness to various wheelchair-based monitoring technologies. Survey B included questions about which technologies the respondents were familiar with, which they thought might be most helpful to caregivers in preventing pressure injuries, and which technologies they would use if available. Survey B was administered using the University of Minnesota's Research Electronic Data Capture system. Research Electronic Data Capture is a secure, web-based software platform designed to support data capture for research studies [21]. See survey B in Multimedia Appendix 1 for full details about the items in the survey, including annotations indicating where the survey differed from survey A.

Survey Participants

Survey A was made available to all LTC facilities in the state of Georgia, United States, that provide care to adults with AD or ADRD. The recruitment flyer for survey A was emailed individually to an administrator in all LTC facilities that provide care to adults with AD or ADRD. A follow-up phone call was made to each facility to confirm receipt of the survey link. An invitation to complete survey B was distributed by email to a convenience sample of rehabilitation care providers in LTC centers in Minnesota, United States, that were affiliated with the occupational therapy program as education sites. A follow-up email was sent after 2 weeks in an attempt to improve the response rate. Respondents were not offered remuneration for either survey A or B.

Survey Data Analysis

Survey responses and demographic variables for both surveys A and B were summarized using descriptive statistics for each to report the frequency of responses. Because there was no experimental design requiring inferential comparisons, descriptive analysis was the most appropriate to use.

Interview Data Collection

Participants

Interview participants were recruited from LTC facilities that provide care to adults with AD or ADRD within a 90-minute radius of a large metropolitan area. The targeted LTC facilities were provided with a recruitment flyer by email and in person, along with a follow-up phone call to confirm receipt of the recruitment flyer. The sampling frame included direct care providers (nurses and CNAs), members from the administration, and other stakeholders, such as OTs and PTs, who are directly involved in the care of persons with AD or ADRD who use wheelchairs and live in skilled care facilities. The inclusion criteria for prospective participants are presented in Textbox 1.

Textbox 1. Inclusion criteria.

Inclusion criteria

- Aged ≥18 years
- Work directly with patients who have Alzheimer disease or Alzheimer disease and related dementia or are a key stakeholder that may interact with the patients or systems
- Work at a skilled care facility or other facility that provides regular care to patients with Alzheimer disease or Alzheimer disease and related dementia and who use wheelchairs

XSL•FO
RenderX

For participating in the study, interviewees received a US $50 gift card, and study sites were compensated with a US $250 gift card for their involvement.

Interview Questions

Semistructured interviews were conducted to identify and understand the gaps in necessary care related to pressure injury prevention for wheelchair users with AD or ADRD in LTC. During interviews, participants were asked to discuss their respective roles, describe a typical workday, state knowledge of their employer's existing weight shift policies, as well as to provide feedback on the utility and feasibility of implementing technologies designed to reduce the incidence of pressure injuries. The interview questions were developed by the research team to probe perspectives around the use of mHealth technologies specifically related to pressure injury prevention and can be reviewed in Multimedia Appendix 1.

To solicit feedback on the feasibility of personal-use technologies to prevent pressure injuries, interview participants were shown 2 short videos. The videos demonstrated 2 different in-seat tracking and pressure management systems that are commercially available. The systems shown were the Sensoria Mat (Sensoria Health Inc) and the Sensomative Wheelchair (Sensomative GmbH), both thin 40 cm × 40 cm fabric mats with embedded force sensors that can be placed under or inside of the wheelchair cushion and communicate with their respective apps to provide feedback about weight shifts with the intent to increase body movement and weight shifting. These novel, personal-use technologies were not originally designed for wheelchair users living with ADRD but may offer features that could be useful for this population. The videos were specifically designed to provide examples of the types of technologies that are available for the general wheelchair-user population, rather than to elicit feedback only on these specific products.

Interview Data Analysis

The interview data were recorded, transcribed verbatim, and then analyzed using a thematic analysis approach. Thematic analysis is a qualitative technique that identifies patterns in raw interview data with the aim of developing overarching themes [22]. To store, manage, and analyze the interview data, we used NVivo, a qualitative software platform by QSR International.

The first step in our thematic analysis was data familiarization. Each interview transcript was perused several times for familiarization, as well as to observe any meanings and patterns in the data. During data familiarization, analytic memos were maintained in NVivo to reflect on participant statements and meanings. Next, open coding, the process of creating initial codes, was completed. An example of an initial code was *prevention*, a code that refers to participant descriptions of any response, effort, intervention, or resource aimed at preventing pressure injuries (eg, weight shifts). The prevention code also referred to strategies staff used to assist residents in keeping their skin healthy (eg, skin creams).

After the open coding process, a codebook containing all the codes inductively created in step 2 was developed. Codes that generated the greatest number of references (ie, chunks of coded interview data) were highlighted. For instance, the code *prevention* contained 162 references. See Multimedia Appendix 1 for an illustration of the codebook.

The third step in our thematic analysis was to create in vivo codes using NVivo. In vivo codes are excerpts reflecting the participants' own words. The excerpt, "I usually try to shift them every two hours," is an example of an in vivo code. In our thematic analysis, step 3 resulted in nearly 600 in vivo codes. The fourth step was to associate select in vivo codes—codes that reflect a recurring pattern in the data—with the initial codes we created in step 2. Step 4 was a process of grouping or merging recurring patterns into themes.

Results

Survey Responses

There were a total of 30 respondents between survey A (n=22) and survey B (n=8). Demographic characteristics include respondents' role in their facility, experience providing care to adults with AD or ADRD, their age, race, gender, and caregiver-to-resident ratio in the facility they work in (Table 1). Survey A included administrators, certified nursing and personal care assistants, and nurses, while survey B included primarily rehabilitation therapy staff and 1 nurse. All respondents of both surveys had >3 years of experience, and most respondents on survey A (n=13, 59%) and survey B (n=6, 75%) had at least 11 years of experience. All participants of the surveys were aged >30 years and predominantly female. The caregiver-to-resident ratio was similar between groups, with most participants of both surveys indicating a ratio of 1:6 to 1:10 in their facilities.

XSL•FO
RenderX

Table 1. Characteristics of survey and interview participants from long-term care facilities.

Characteristics	Survey A (n=22), n (%)[a]	Survey B (n=8), n (%)[b]	Interviews (n=9), n (%)
Role			
Certified nursing assistant	2 (11)	0 (0)	5 (56)
Personal care assistant	2 (11)	0 (0)	0 (0)
Registered nurse	2 (11)	1 (12)	0 (0)
Licensed practical nurse	2 (11)	0 (0)	0 (0)
Administrator	8 (42)	0 (0)	3 (33)
Operational director	1 (5)	0 (0)	0 (0)
Social worker	0 (0)	0 (0)	1 (11)
Occupational therapist or physical therapist	0 (0)	7 (88)	0 (0)
Experience (y) in Alzheimer disease or Alzheimer and related disease care			
3-5	3 (16)	0 (0)	N/A[b]
6-10	3 (16)	2 (25)	N/A
≥11	13 (68)	6 (75)	N/A
Age (y) of participants			
30-39	4 (24)	5 (62)	N/A
40-49	2 (12)	2 (25)	N/A
50-59	8 (47)	0 (0)	N/A
≥60	3 (18)	1 (12)	N/A
Race			
Black or African American	3 (18)	0 (0)	N/A
Multirace	1 (6)	0 (0)	N/A
White	12 (71)	8 (100)	N/A
Sex			
Female	14 (82)	7 (88)	N/A
Male	3 (18)	1 (12)	N/A
Caregiver-to-resident ratio in the facility			
1:6-1:10	9 (47)	7 (88)	N/A
1:11-1:15	8 (42)	1 (12)	N/A
>1:15	2 (11)	0 (0)	N/A

[a]n=19 for role, experience, and caregiver-to-resident ratio and n=17 for age, race, and sex.

[b]Not applicable.

The typical daily sitting time for the residents in their facilities and the types of surfaces used are reported in Tables 2 and 3, respectively. The 22 responses from survey A indicate that residents in their facilities sit for 0 to 6 hours per day in a wheelchair (n=13, 59%) and for 0 to 6 hours per day on other surfaces (n=15, 68%), resulting in the potential cumulative of up to 12 hours per day for some residents who may be included in both groups. About half (n=11, 50%) of the responses indicated that residents sit 7 to 12 hours or >12 hours per day.

XSL•FO
RenderX

Table 2. Care provider response for typical daily sitting time for long-term care residents with Alzheimer disease or Alzheimer disease and related dementias.

Resident characteristic	Sitting for 0-6 h, n (%)	Sitting for 7-12 h, n (%)	Sitting for >12 h, n (%)
All residents in wheelchairs (n=22)	13 (59)	7 (32)	4 (18)
All residents on other surfaces (n=22)	15 (68)	2 (9)	3 (14)
Mild memory loss (n=8)	2 (25)	4 (50)	2 (25)
Moderate memory loss (n=8)	0 (0)	7 (88)	1 (12)
Advanced memory loss (n=8)	1 (12)	5 (62)	2 (25)

Table 3. Types of sitting surfaces used in long-term care facilities for residents with Alzheimer disease and Alzheimer disease and related dementias, as reported by care providers, and methods of procurement.

Sitting surface characteristics	Survey A (n=22), n (%)	Survey B (n=8), n (%)
Nonwheelchair sitting surfaces[a]		
Geriatric “geri” chair	11 (50)	6 (75)
Recliner	9 (45)	8 (100)
Nonreclining bedside chair	0 (0)	5 (62)
Dining room chair	6 (30)	0 (0)
Broda chair	3 (15)	0 (0)
Sofa	2 (10)	0 (0)
Glider or lounge chair	1 (5)	0 (0)
Wheelchair-based equipment[a]		
Wheelchair (all types)	22 (100)	0 (0)
Manual wheelchair	0 (0)	8 (100)
Power wheelchair	0 (0)	7 (88)
Power scooter	0 (0)	4 (50)
Mobility device provision[b]		
Personal purchase	17 (77)	8 (100)
Facility owned (shared)	5 (23)	8 (100)
Seat cushions[a]		
Air filled	7 (32)	7 (88)
Gel	17 (77)	8 (100)
Foam	12 (55)	8 (100)
Honeycomb matrix	2 (9)	5 (62)
Specialty	0 (0)	6 (75)
Memory foam	0 (0)	3 (38)
Cushion provision[a]		
Self-purchase	10 (45)	3 (38)
Facility owned (shared)	22 (100)	8 (100)
Seating specialist recommended	10 (45)	3 (38)

[a]Respondents were allowed to select >1 response in this category on the survey.

[b]In Survey A, participants selected the percentage of residents in their facility for each category.

Survey B responses narrowed in on sitting time for residents at different stages of memory loss, reporting that most residents in their facilities across all stages of memory loss sit for 7 to 12 hours per day (Table 2). Survey B did not distinguish between sitting in a wheelchair and sitting on other surfaces, so 7 to 12

XSL•FO
RenderX

hours can be assumed to represent cumulative sitting time per day.

Both surveys indicated that mobility-related sitting equipment and nonmobile chairs were used in their facilities. Nonmobile sitting equipment was most selected including geriatric chairs and recliners. Wheelchairs were selected by all respondents as available and used in their facilities. The 8 responses in survey B narrowed in on the type of mobility device, and they indicated that in addition to manual wheelchairs, power wheelchairs (n=7, 88%) and scooters (n=4, 50%) were also commonly used in their facilities (Table 3). All (n=8, 100%) respondents of survey B indicated that sitting-related equipment included shared equipment provided by the facility and privately purchased by the resident (n=8, 100%). Survey A responses indicated that more often, mobility devices were privately purchased by the resident (17/22, 77%). Respondents on both surveys indicated that a range of skin protection seat cushion options were provided by the facility, with some residents opting to purchase their own (Table 3). Less than half of respondents endorsed having access to a seating specialist at their facilities in survey A (n=10, 45%) and survey B (n=3, 38%).

Survey B respondents (5/8, 62%) reported a perception that those in more advanced stages of memory loss were more susceptible to sitting-related pressure injuries (Table 4).

Table 4. Care provider perceptions about stages of memory loss most susceptible to pressure injuries in long-term care facilities (survey B; n=8).

Stage of memory loss	Respondents, n (%)
Mild	1 (12)
Moderate	2 (25)
Advanced	5 (62)

In survey B with 8 respondents, familiarity with mHealth technologies by care providers was most favorable for self-adjusting surfaces (n=7, 88%) and pressure mapping (n=6, 75%). Perceptions about which type of technology could be most helpful for caregivers were most favorable for self-adjusting surfaces (n=6, 75%) followed by wetness detection (n=4, 50%). Finally, regarding which technology the care provider would want to use if it were available to them, the most favorable response was for monitoring repositioning over time (n=6, 75%) and pressure mapping (n=4, 50%) (Table 5).

Table 5. Care provider perceptions about technologies to prevent pressure injuries in skilled nursing facilities (n=8). Respondents were able to select >1 technology type.

Sitting-related technology type	Familiar with technology, n (%)	Helpful for caregivers, n (%)	Would use if available, n (%)
Monitor repositioning over time	0 (0)	2 (25)	6 (75)
Pressure map chair	6 (75)	0 (0)	4 (50)
Self-adjusting surfaces	7 (88)	6 (75)	3 (38)
Timers or reminders	4 (50)	3 (38)	3 (38)
Wetness detection	3 (38)	4 (50)	3 (38)
Alerts or alarms to detect a problem	1 (12)	1 (12)	1 (12)

Themes From Interviews

The qualitative sample (n=9) included 3 (33%) administrators, 1 (11%) social worker, and 5 (56%) direct care providers (ie, CNAs). Three overarching themes were developed that addressed the study's research questions.

Repositioning Needs Vary Widely in Residents of LTC With AD or ADRD

Participants described a wide range of physical functioning and mobility among residents living with ADRD in their LTC settings. From being "bed-bound," to using wheelchairs to ambulating unassisted, participants noted a wide range of physical functioning that includes some residents performing activities of daily living on their own and others requiring assistance with transferring, repositioning, toileting, and grooming. One participant noted, "Some of them can get up, but we have to assist them with transferring." Another participant observed, "Some just want to sit in their [wheel]chair and not be bothered." Another direct care worker observed, "There are some residents that don't get up at all."

In terms of range in time spent sitting per day, a participant shared, "The majority of our residents range kind of in the middle."

Staff encouraged residents with higher functioning and mobility to engage in physical activities, such as dancing, playing games, and other group activities. For lower-functioning residents, staff were available to assist with transferring, walking, performing activities of daily living, and encouraging residents who were physically able to participate in recreational activities, such as bingo, bowling, and playing cards, to name a few. One participant noted, "We do music over there, so when they're dancing, they're...moving and most of them love dancing." Residents with lower mobility or physical functioning participate in more passive activities, such as watching movies, assembling blocks and puzzles, or matching and folding socks. A participant

XSL•FO
RenderX

who was a social worker stated, “Falls are the biggest concern we have for this population.” On the basis of her experience, a direct care worker recounted, “After they fall, they’re scared to move, scared to get up. They don’t want you touching them.” Another direct care worker advised, “We need to stabilize them, just so they do not fall.”

Repositioning Is Problematic for the Highest-Risk Residents With AD or ADRD

All 9 participants in this study, most of whom are direct care workers, acknowledged a corporate policy of repositioning residents every 2 hours throughout the day. One participant stated, “We change them as often as we can, cause...sitting in that chair is not feasible for the skin.” A participant, who was an administrator, stated, “Our policy is for staff to make sure they turn...at a minimum of every two hours and more frequently if needed for the resident’s wellbeing.”

Due to the high turnover of direct care workers in the LTC industry, a problem that was exacerbated during the COVID-19 pandemic, performing weight shifts around the clock to reposition residents is an ongoing challenge. In her role as a social worker, one participant made the following observations:

> *Typically, our goal is to get everyone out of bed every day, and they have the right to refuse. If someone is in bed enough for pressure ulcers, [that is] a constant concern. In speaking for our memory care unit...we have tight staffing...we get everybody out of bed, and the activities and all of that. If someone is in the bed enough for pressure ulcers to be constant concern, they typically are not aware enough to respond to a lot of prompts.*

As a preventive, one participant noted, “You have different paddings...and you have to know where to place them.” Most participants in the study indicated they have an adequate staffing pattern in their facilities to ensure weight shifts are being performed on a periodic basis. Examples include the following quotes:

> *There’s always someone to help them get up.*
>
> *We have always tried to sustain a higher staffing pattern.*
>
> *We have staff sufficient to be able to help that.*

It should be noted that 2 (22%) participants in the study who provided direct care to residents described feeling overwhelmed, at times, adhering to existing weight shift policies. Direct care workers in the study described daily challenges they encounter when attending to residents with cognitive impairment who need assistance repositioning. The following quotes exemplify the challenges direct care workers face in their efforts to reposition residents with ADRD, particularly wheelchair users:

> *A lot of my residents don’t want to be touched.*
>
> *They don’t want that range of motion.*
>
> *Some of them are very stubborn and they don’t want to do it.*
>
> *Some of them get a little agitated.*
>
> *It’s hard to explain to them, because they’re not gonna understand.*
>
> *You gonna get hit by some of ‘em, because it’s gonna scare them.*
>
> *Some of them just wanna sit in their chair and not be bothered.*

According to our sample of participants, residents with mild cognitive impairment tend to be more cooperative and often will help in repositioning themselves. A direct care worker offered the following example: “You say turn, leg over leg, and they move over.” The chief operating officer of one facility stated, “They may not always understand why, but they understand that we need them to move.” Another participant stated, “We haven’t had any in a long time that refuse to turn for you.”

Overall, participants in the study acknowledged and understood the significance of periodic weight shifts and repositioning and described a breadth of experiences in preventing pressure injuries in their resident populations. Most participants indicated that pressure ulcers were not a major problem in their facilities.

Innovative Technologies Can Help and Hinder Care

Residents living with AD or ADRD, who spend most of their waking hours in wheelchairs, require repositioning and frequent weight shifts throughout the day to prevent pressure injuries. In this vein, participants described a variety of techniques they have used to assist residents at risk for developing pressure injuries. For instance, one participant stated she makes “sure they have...good footrest to rest the leg, because mostly the sores [are] in the back sometimes.” Another participant stated, “Sometimes we put a pad or cushion on the chair to make it much softer; sometimes...we put a pad there to kind of ease them off that pressure point.” Another participant, a director of nursing, offered the following insight: “We have a restorative team that...helps with passive range of motion...to get those joints moving.”

In response to the video demos, participants offered a range of opinions on the utility and feasibility of these specific technologies and the generic concepts for use in their facilities. Overall, the participants were impressed with the technologies. “I think this is amazing,” remarked one participant. Another participant noted, “I think that would be an amazing asset to a memory care unit in assisted living.” A CNA participant offered the following insight: “That would be good. Save us from pulling or pushing.”

Due to advanced cognitive impairment in some residents, other participants were a bit skeptical of the technologies’ utility. For instance, 1 (11%) participant concluded, “There are some residents who can participate and benefit from that, [but] it will be a small percentage.”

Another respondent had the following opinion: “I’m not saying this is hard; I just think it’s too much for them. It’s way too coordinated for them.”

Offering a similar opinion, another participant observed, “For our patient population that wouldn’t be effective...they wouldn’t

XSL•FO
RenderX

know to do it, from that part of getting up" (referring to the part of the video that shows an alert for user to do a pressure relief)."

Despite the range of opinions, all participants endorsed that, while the sitting-related feedback technologies could be independently managed by a small percentage of residents with higher functioning, the feedback may be too complex for those with more advanced executive function impairment. As a consensus, participants felt that the technologies would be much more suitable for direct care workers (eg, CNAs) to encourage periodic weight shifts. In addition to direct care workers, they responded that sitting-related feedback technologies might also be useful for informal caregivers (eg, family members) to help their loved ones increase in-seat movement.

Discussion

Primary Findings

Our findings shed light on unique considerations that may impact the balance between helpful and burdensome when implementing assistive technologies for use by caregivers in LTC facilities with residents who have AD or ADRD, specifically to facilitate repositioning needs during sitting. While caregivers were interested in feedback tools to help them be more efficient or effective in repositioning residents, successful implementation may be challenging due to barriers around available time to learn or use a new technology in their daily workflow. Caregivers also brought awareness to the challenge of repositioning residents with AD or ADRD who experience agitation or combativeness in later stages. These behavioral challenges require a different approach than simply providing reminders that it is time to reposition someone or visual feedback about their pressure distribution. Thus, the assistive technologies for managing sitting-related pressure presented to caregivers in this study only partially address the caregiving needs for people with AD or ADRD in LTC settings.

Due to the ongoing shortage of direct care workers in LTC facilities and 58% of LTC residents having a diagnosis of AD or ADRD, the resulting workload of the currently employed direct care workers and participants increased and they noted that reminders to reposition residents could help with their required monitoring [5,23]. However, even with feedback about pressure or delivery of reminders, completing the required repositioning remains a challenge for already burdened caregivers who, in interviews, acknowledged the need for frequent repositioning of residents but lacked time or resources to do so. Thus, future implementation of feedback technologies will need to carefully consider this barrier, where technology could be perceived as both helpful and futile if the burden of care is not addressed adequately.

Technologies that offer automatic repositioning (eg, self-adjusting cushions [24]) could potentially help reduce caregiver workload by partially managing pressure relief for residents who cannot physically reposition themselves. Furthermore, higher functioning residents with mild memory loss could potentially reposition themselves independently if the feedback technologies are appropriately adapted for their cognitive limitations and if it is safe for the resident to do so [25]. While care providers did not endorse currently using any type of sitting-related technologies, they leaned on a diverse set of strategies for ensuring regular repositioning or activities to increase overall movement among the residents. These strategies were effective with residents with mild cognitive decline who tended to be more cooperative when provided with cues or instructions. However, residents with advanced cognitive decline were more likely to be resistant to movement and, in general, may be difficult to provide care for due to increased agitation, combativeness, and resistance. As a result, residents with more significant cognitive decline may spend more time sitting without pressure relief, resulting in an increased risk of pressure injuries. Understanding the end-user needs is critical for the successful implementation of any new technologies. While we only included a small number of participants and a limited number of care facilities, an important consideration arose—the added impact of cognitive decline as AD or ADRD advances. This creates a layer of complexity around how caregivers approach repositioning residents in LTC.

In the realm of LTC facilities, the integration of technology, particularly within memory care units, holds the potential to enhance both resident well-being and operational efficiency [26]. Various technologies, such as physiological monitoring systems, emergency detection systems, and those designed to foster social interactions, have been used over the years. However, a review study by Chan et al [27] revealed that certain technologies, such as position sensors aimed at reducing falls, have proven ineffective in LTC settings [27]. Despite fall detection systems having a history dating back to the 1970s, their adoption in nursing homes remains varied and limited. Similarly, pet robots and robotic seals, intended to manage behavioral and psychological symptoms of dementia, were found to be ineffective [27]. Technologies designed to promote movement, including the Leaf Patient Monitoring System (Smith+Nephew Leaf* System), M.A.P.TM Continuous Bedside Pressure Mapping System (CBPM; Wellsense), and ForeSite Intelligent Surface (XSensor), have seen limited evaluation in LTC, with Leaf being an exception, demonstrating acceptance [8] and improved turn compliance [28].

Regarding the use of interface pressure map (IPM) feedback as a potential strategy, while nursing staff indicated the pressure feedback was helpful in their efforts to reposition patients, IPM used on beds in the hospital setting did not yield significant changes in pressure injury outcomes in a geriatric or internal medicine ward, as noted by Gunningberg et al [29]. However, when IPM was used by OTs and PTs to evaluate and select seating cushion needs in LTC, the incidence of pressure injuries was reduced [30]. The challenges of implementing technology in LTC environments are evident, and the transition from adoption to implementation to tangible outcomes in pressure injury reduction is particularly arduous. Few products have achieved success, underscoring the importance of involving key stakeholders early in the development process to enhance the likelihood of success [31].

Overall, results from these analyses confirmed and expanded our understanding of the needs of LTC caregivers who care for residents with AD or ADRD in terms of managing sitting-related pressure. Future research could identify how feedback

XSL•FO
RenderX

technologies could support residents with mild memory loss to more independently manage pressure and support those with higher acuity, including those admitted with pressure injuries, by offering technologies, such as automated or self-adjusting seat cushions. As the population grows older and caregiver shortages increase across the United States, the use of appropriate technologies may increase effectiveness and efficiency in preventing pressure injuries.

Limitations

The participants in this study provided consistent feedback for developers of sitting-related technologies; however, the study has a few limitations. Survey response was low and may reflect limited interest in the topic of managing pressure injuries related to sitting in the greater context of other care provided by the participants. It may be a result of staffing barriers following the demands the pandemic placed on direct care providers. Survey response may have improved with remuneration for time to complete the survey. Surveys A and B differed slightly in their presentation of items and level of detail, making it difficult to combine the responses. We did not gather facility-level data, so it is not known how many respondents were from a particular facility, or how many total facilities were represented across respondents. While the purpose for structuring the surveys in this way was to gather preliminary information, future surveys and interviews across larger samples could include a more in-depth exploration within and across facilities because they can differ in culture and policy.

Implications for Practice

The aging population in the United States is increasing, acuity is increasing, and caregiver availability is decreasing. Innovative technologies could meet this gap by helping caregivers be more effective and efficient in adhering to the repositioning needs of residents in LTC. Preventing pressure injuries in patients with AD or ADRD, particularly those in the more advanced stages, can improve their quality of life. Effective sitting-related feedback technologies may have the potential to reduce the cost burden created by sitting-acquired pressure injuries.

Conclusions

Our findings confirm that the care providers in the LTC settings we engaged with are interested in sitting-related feedback technologies that can improve their efficiency and effectiveness. While care providers indicated that they understand the clinical guidelines and their facility's procedures for repositioning and they have access to seating professionals and appropriate pressure-relieving seat cushions, finding time to reposition residents is one of their primary barriers. Thus, future implementation should consider nursing burden, the cost-benefit balance, and the demands of learning and using new tools. Stakeholders' strongest interest in adopting sitting-related technology was in systems that automate pressure redistribution in some way, show how pressure is distributed on a chair or wheelchair, or provide feedback via reminders that someone needs to be repositioned.

In the LTC context, care providers may prefer technologies that reduce their care burden in some way versus offering them reminders to perform a task that they may not have time to complete, even though they have the knowledge, skills, and access to resources to help them perform the task.

Acknowledgments

This research was supported by the National Institutes of Health's National Center for Advancing Translational Sciences (grant) and the National Institute on Aging (grants and). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health's National Center for Advancing Translational Sciences or the National Institute on Aging.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview questions, survey A items, and survey B items.
[DOCX File , 249 KB - formative_v9i1e59590_app1.docx]

References

1. Anthony D, Alosoumi D, Safari R. Prevalence of pressure ulcers in long-term care: a global review. J Wound Care 2019 Nov 02;28(11):702-709. [doi: 10.12968/jowc.2019.28.11.702] [Medline: 31721671]
2. Sprigle S, McNair D, Sonenblum S. Pressure ulcer risk factors in persons with mobility-related disabilities. Adv Skin Wound Care 2020;33(3):146-154. [doi: 10.1097/01.asw.0000653152.36482.7d]
3. -. 2023 Alzheimer's disease facts and figures. Alzheimers Dement 2023 Apr 14;19(4):1598-1695. [doi: 10.1002/alz.13016] [Medline: 36918389]

XSL•FO
RenderX

4. Haesler E. Prevention and Treatment of Pressure Ulcers/injuries: Clinical Practice Guideline. New York, NY: European Pressure Ulcer Advisory Panel, National Pressure Injury Advisory Panel, Pan Pacific Pressure Injury Alliance; 2019.
5. Skilled nursing facility quality reporting program - specifications for percent of residents or patients with pressure ulcers that are new or worsened (NQF #0678). RTI International. 2016. URL: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/NursingHomeQualityInits/Downloads/SNF-QRP-Measure-Specifications_October-2016.pdf [accessed 2025-05-29]
6. Meaume S, Marty M. Pressure ulcer prevention using an alternating-pressure mattress overlay: the MATCARP project. J Wound Care 2018 Aug 02;27(8):488-494. [doi: 10.12968/jowc.2018.27.8.488] [Medline: 30086253]
7. Scott RG, Thurman KM. Visual feedback of continuous bedside pressure mapping to optimize effective patient repositioning. Adv Wound Care (New Rochelle) 2014 May 01;3(5):376-382 [FREE Full text] [doi: 10.1089/wound.2013.0519] [Medline: 24804157]
8. Yap TL, Kennerly SM, Ly K. Pressure injury prevention: outcomes and challenges to use of resident monitoring technology in a nursing home. J Wound Ostomy Continence Nurs 2019;46(3):207-213 [FREE Full text] [doi: 10.1097/WON.0000000000000523] [Medline: 31083063]
9. Van Loo L. Sensing pressure. The Free Library. 2021. URL: https://www.thefreelibrary.com/Sensing+Pressure.-a0652092386 [accessed 2025-05-29]
10. Wu YK, Liu HY, Kelleher A, Pearlman J, Cooper RA. Evaluating the usability of a smartphone virtual seating coach application for powered wheelchair users. Med Eng Phys 2016 Jun;38(6):569-575. [doi: 10.1016/j.medengphy.2016.03.001] [Medline: 27079179]
11. Kim E, Baskys A, Law AV, Roosan MR, Li Y, Roosan D. Scoping review: the empowerment of Alzheimer's Disease caregivers with mHealth applications. NPJ Digit Med 2021 Sep 07;4(1):131 [FREE Full text] [doi: 10.1038/s41746-021-00506-4] [Medline: 34493819]
12. Yousaf K, Mehmood Z, Awan IA, Saba T, Alharbey R, Qadah T, et al. A comprehensive study of mobile-health based assistive technology for the healthcare of dementia and Alzheimer's disease (AD). Health Care Manag Sci 2020 Jun;23(2):287-309. [doi: 10.1007/s10729-019-09486-0] [Medline: 31218511]
13. Ahad N, Sonenblum SE, Davenport MA, Sprigle S. Validating a wheelchair in-seat activity tracker. Assist Technol 2022 Sep 03;34(5):588-598 [FREE Full text] [doi: 10.1080/10400435.2021.1893237] [Medline: 33617402]
14. Jordan K, Vos-Draper T, Morrow M, Sonenblum S. The usability of two mobile health assistive technologies for wheelchair-related in-seat movement and pressure. J Rehabil Assist Technol Eng 2023 Nov 14;10:20556683231211808 [FREE Full text] [doi: 10.1177/20556683231211808] [Medline: 38028632]
15. Olney CM, Vos-Draper T, Egginton J, Ferguson J, Goldish G, Eddy B, et al. Development of a comprehensive mobile assessment of pressure (CMAP) system for pressure injury prevention for veterans with spinal cord injury. J Spinal Cord Med 2019 Nov 31;42(6):685-694 [FREE Full text] [doi: 10.1080/10790268.2019.1570437] [Medline: 30702395]
16. ISO/TR 16840-9:2015: wheelchair seating: part 9: clinical interface pressure mapping guidelines for seating. International Organization for Standardization. 2015. URL: https://www.iso.org/standard/65198.html [accessed 2025-05-29]
17. Horvath KJ, Ecklund AM, Hunt SL, Nelson TF, Toomey TL. Developing Internet-based health interventions: a guide for public health researchers and practitioners. J Med Internet Res 2015 Jan 23;17(1):e28 [FREE Full text] [doi: 10.2196/jmir.3770] [Medline: 25650702]
18. Yardley L, Morrison L, Bradbury K, Muller I. The person-based approach to intervention development: application to digital health-related behavior change interventions. J Med Internet Res 2015 Jan 30;17(1):e30 [FREE Full text] [doi: 10.2196/jmir.4055] [Medline: 25639757]
19. Rosenstock IM, Strecher VJ, Becker MH. Social learning theory and the Health Belief Model. Health Educ Q 1988 Sep 04;15(2):175-183. [doi: 10.1177/109019818801500203] [Medline: 3378902]
20. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014 Sep;89(9):1245-1251 [FREE Full text] [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]
21. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381 [FREE Full text] [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
22. Braun V, Clarke V. Thematic Analysis: A Practical Guide. Thousand Oaks, CA: Sage Publication; 2022.
23. Heiks C, Sabine N. Long term care and skilled nursing facilities. Dela J Public Health 2022 Dec;8(5):144-149 [FREE Full text] [doi: 10.32481/djph.2022.12.032] [Medline: 36751604]
24. Arias-Guzman S, Call K, Laskin JJ, Oberg C, Jellum S, Russon M, et al. Effects on tissue oxygenation in the gluteal region with a smart pressure redistribution wheelchair cushion compared to a pressure relief manoeuvre. Disabil Rehabil Assist Technol 2025 Feb 18;20(2):468-475 [FREE Full text] [doi: 10.1080/17483107.2024.2391976] [Medline: 39155408]
25. Turmell M, Cooley A, Yap TL, Alderden J, Sabol VK, Lin JA, et al. Improving pressure injury prevention by using wearable sensors to cue critical care patient repositioning. Am J Crit Care 2022 Jul 01;31(4):295-305. [doi: 10.4037/ajcc2022701] [Medline: 35773199]

XSL•FO
RenderX

26. Kennerly SM, Sharkey PD, Horn SD, Zheng T, Alderden J, Sabol VK, et al. Characteristics of nursing home resident movement patterns: results from the TEAM-UP trial. Adv Skin Wound Care 2022 May 01;35(5):271-280 [FREE Full text] [doi: 10.1097/01.ASW.0000822696.67886.67] [Medline: 35195085]
27. Chan DK, Chan LK, Kuang YM, Le MN, Celler B. Digital care technologies in people with dementia living in long-term care facilities to prevent falls and manage behavioural and psychological symptoms of dementia: a systematic review. Eur J Ageing 2022 Sep 15;19(3):309-323 [FREE Full text] [doi: 10.1007/s10433-021-00627-5] [Medline: 36052197]
28. Yap TL, Horn SD, Sharkey PD, Zheng T, Bergstrom N, Colon-Emeric C, et al. Effect of varying repositioning frequency on pressure injury prevention in nursing home residents: TEAM-UP trial results. Adv Skin Wound Care 2022;35(6):315-325. [doi: 10.1097/01.asw.0000817840.68588.04]
29. Gunningberg L, Sedin IM, Andersson S, Pingel R. Pressure mapping to prevent pressure ulcers in a hospital setting: a pragmatic randomised controlled trial. Int J Nurs Stud 2017 Jul;72:53-59. [doi: 10.1016/j.ijnurstu.2017.04.007] [Medline: 28460263]
30. Brienza DM, Kelsey S, Karg P, Allegretti A, Olson M, Schmeler M, et al. A randomized clinical trial on preventing pressure ulcers with wheelchair seat cushions. J Am Geriatr Soc 2010 Dec;58(12):2308-2314 [FREE Full text] [doi: 10.1111/j.1532-5415.2010.03168.x] [Medline: 21070197]
31. Engelsma T, Jaspers MW, Peute LW. Considerate mHealth design for older adults with Alzheimer's disease and related dementias (ADRD): a scoping review on usability barriers and design suggestions. Int J Med Inform 2021 Aug;152:104494 [FREE Full text] [doi: 10.1016/j.ijmedinf.2021.104494] [Medline: 34015657]

Abbreviations

AD: Alzheimer disease
ADRD: Alzheimer disease and related dementias
CNA: certified nursing assistant
IPM: interface pressure map
IRB: institutional review board
LTC: long-term care
mHealth: mobile health
OT: occupational therapist
PT: physical therapist

Edited by A Mavragani; submitted 16.04.24; peer-reviewed by H Owen; comments to author 16.10.24; revised version received 06.01.25; accepted 01.04.25; published 17.09.25.

Please cite as:
Vos-Draper T, Vinoski Thomas E, Graybill E, Morrow M, Jordan KA, Manley PR, Sonenblum SE
Stakeholder Perspectives on mHealth Technologies to Prevent Sitting-Acquired Pressure Injuries in Long-Term Care Facilities: Mixed Methods Study
JMIR Form Res 2025;9:e59590
URL: https://formative.jmir.org/2025/1/e59590
doi:10.2196/59590
PMID:

©Tamara Vos-Draper, Erin Vinoski Thomas, Emily Graybill, Melissa Morrow, Kathleen A Jordan, Pamela R Manley, Sharon E Sonenblum. Originally published in JMIR Formative Research (https://formative.jmir.org), 17.09.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Exploring the Impact of a Remote Monitoring System for Palliative and End-of-Life Care (CARE-PAC): Mixed Methods Feasibility Study

Roma Maguire[1], BN (Hons), MSc, PhD; Lisa McCann[1*], BSc, PhD; Claire Singleton[2*], BSc (Hons); Paul Perkins[3*], MA (Hons), MB BCh, MRCP, Dip Pall Med, MA; Ollie Minton[4*], PGcert Meded, BSc, MBBS, MSc, PhD; Nicola McCann[5*], BSc, Pg Dip Cancer Nursing, MSc; Emma Longford[6*], BSc, MBChB, MA; Alistair Dermot McKeown[7*], BSc Med Sci (Hons), MBChB; Kimberley Kavanagh[2*], BSc, PhD; Morven Miller[1*], BA (Hons), MSc, PhD

[1]Department of Computer and Information Sciences, University of Strathclyde, Glasgow, United Kingdom
[2]Department of Mathematics and Statistics, University of Strathclyde, Glasgow, United Kingdom
[3]Sue Ryder, Leckhampton Court Hospice, Cheltenham, United Kingdom
[4]Sussex Cancer Centre, University Hospitals Sussex NHS Foundation Trust, Brighton, United Kingdom
[5]NHS Lanarkshire, Bothwell, United Kingdom
[6]Wirral Hospice St John's, Wirral, United Kingdom
[7]Prince and Princess of Wales Hospice, Glasgow, United Kingdom
[*]these authors contributed equally

Corresponding Author:
Roma Maguire, BN (Hons), MSc, PhD
Department of Computer and Information Sciences
University of Strathclyde
Office 726 Livingston Tower
26 Richmond Street
Glasgow, G1 1XQ
United Kingdom
Phone: 44 141 548 3189
Email: roma.maguire@strath.ac.uk

Abstract

Background: In the United Kingdom, access to and the quality of palliative and end-of-life care (PEOLC) vary widely. In the final months of life, many patients face avoidable accident and emergency (A&E) visits and hospital admissions, driven by gaps in out-of-hours support and poorly coordinated care. This not only increases stress for patients and carers but also places avoidable strain and cost on the National Health Service (NHS). There is an urgent need for more compassionate, person-centered models that support people to remain at home, improve their quality of life (QoL), and reduce unnecessary use of acute services.

Objective: This study aimed to explore the usability, user experiences, and impact of the digital dyadic remote monitoring Care and Support System for Patients and Carers (CARE-PAC) for patients in the last year of life, their informal carers, and health professionals involved in their care.

Methods: Patients and informal carers were recruited to use CARE-PAC for up to 12 weeks. A mixed methods approach was used. Quantitative methods included the use of validated QoL scales and the System Usability Scale (SUS). Paired QoL and usability data were analyzed using the Wilcoxon (Pratt) signed-rank test, while unpaired usability data were analyzed using the Wilcoxon rank-sum test. Qualitative methods involved short catch-up calls, in-depth interviews, and focus groups conducted using topic guides informed by the domains of the Non-adoption, Abandonment, Scale-up, Spread, and Sustainability (NASSS) framework. Data were analyzed thematically.

Results: CARE-PAC was implemented across 5 UK clinical sites with 26 participants (13 patient-carer dyads). No significant changes were observed in patients' total QoL scores; however, significant improvements were seen in the "overall QoL" and "social" domains, alongside a significant decline in the "physical" domain. Carers showed no significant changes across total or domain-specific QoL scores. Usability was rated highly by patients (mean 87.9, SD 12.4) and carers (mean 94.7, SD 3.8), indicating an excellent user experience. Health care professionals (HCPs) reported lower usability scores (mean 63.6, SD 15.6), falling

XSL•FO
RenderX

below average but above the threshold for poor usability. Thematic analysis of qualitative data gathered via catch-up calls (all patient-carer dyads), in-depth interviews (2 patients-2 carers), and 4 focus groups/1 interview (12 HCPs) identified 4 key themes: impact on care experiences, reflections and satisfaction, implementation challenges, and future directions.

Conclusions: CARE-PAC is a usable, feasible, and acceptable remote monitoring and support system for patients in the last year of life, their carers, and HCPs. It enables real-time identification of needs and has shown positive impacts on the QoL of both patients and carers. These findings support the need for further research to evaluate its effectiveness at scale and explore pathways for wider implementation in PEOLC.

(JMIR Form Res 2025;9:e69394) doi:10.2196/69394

KEYWORDS

PEOLC; palliative and end-of-life care; CARE-PAC; Care and Support System for Patients and Carers; remote monitoring; usability

Introduction

Palliative and End-of-Life Care

The World Health Organization defines palliative care as an approach that improves the quality of life (QoL) of patients with life-limiting illnesses and their families by preventing and relieving suffering through early identification, assessment, and treatment of pain and other physical, psychosocial, and spiritual problems. It encompasses both palliative care, which can be provided alongside curative treatment, and end-of-life care, which is focused on the final phase of life [1].

There remains a significant unmet demand for palliative and end-of-life care (PEOLC) across many health systems. In the United Kingdom, over 600,000 people die each year, with 90% estimated to benefit from PEOLC—yet, access and quality vary widely. It is estimated that around 100,000 people who could benefit from palliative care die each year without receiving it [2]. This gap in provision has profound implications for both patients and their informal caregivers. For patients, inadequate access to timely and appropriate palliative care often results in poor symptom management, increased psychological distress, and diminished QoL during a highly vulnerable period [3]. Informal caregivers, typically family members or close friends, also experience substantial emotional and physical burdens. The lack of adequate support can lead to increased stress and anxiety, feelings of isolation, and difficulty coping during caregiving and in bereavement [4].

In addition to the human cost, there are substantial financial implications. The final year of life is widely recognized as one of the most resource-intensive periods for health services. In the United Kingdom, over 80% of health care spending in the final year of life is allocated to hospital care, while only 11% is directed toward primary and community health care and less than 4% to hospice care [5]. Although home-based interventions offer substantial savings to the health care system and improvements in patient and caregiver outcomes [6], inadequate support in the community often leaves patients and families with no alternative but to seek help through emergency routes, leading to avoidable accident and emergency (A&E) visits and hospital admissions [7]. These reactive and costly interventions are frequently a consequence of gaps in proactive, coordinated palliative care in the home or community setting [8].

Can Technology Improve PEOLC Services?

Digital technologies have the potential to support, strengthen, and scale up palliative care worldwide [9]. The most common digital health interventions implemented in palliative care are videoconferencing, electronic health records, and phone calls, with the most typical usage of technologies being education, symptom management, decision-making, information provision, and management and communication [10]. eHealth apps have been shown to have promise in promoting equal, individualized care; facilitating accessibility and patient participation in palliative care settings; and promoting feelings of safety and security, while contributing to a more sustainable and efficient use of health care resources [11]. That said, digital technologies should be used to enhance—not replace—face-to-face interactions [12], ensuring that PEOLC services continue to deliver compassionate, person-centered care [13].

Development of a Remote Monitoring and Support System for PEOLC

The Care and Support System for Patients and Carers (CARE-PAC) was co-designed over 2 stages (Figure 1) with patients, carers, and PEOLC health care professionals (HCPs). The result was a remote monitoring and supportive care system tailored to the specific needs of this population (Figure 2) that includes 3 connected components:

- Patient and Carer app: a mobile app for smartphones or tablets that enables users to (1) complete a brief daily symptom questionnaire, (2) submit a 3-weekly QoL and outcomes questionnaire, and access to an electronic library (e-library) featuring (1) web links to trusted websites aligned with questionnaire topics and (2) local service information (eg, helplines, support groups)
- Clinician dashboard: a secure, web-based platform accessible on any device that allows HCPs to (1) register and manage participants, (2) monitor daily symptom reports, and (3) receive and respond to real-time alerts
- Clinical algorithm: an embedded decision support tool that analyzes incoming data and flags concerning symptoms or issues, automatically generating alerts to HCPs via the clinician dashboard

A key strength of CARE-PAC is its dyadic nature through the inclusion of informal carers as active participants within the system, alongside patients themselves. This recognizes carers' crucial role in supporting patients receiving PEOLC. By

XSL•FO
RenderX

monitoring and supporting carers' well-being, alongside that of patients', CARE-PAC aims to help prevent carer crises that can lead to avoidable hospital admissions and to ensure more sustainable care at home.

Figure 1. Co-design and development of the CARE-PAC app. CARE-PAC: Care and Support System for Patients and Carers.

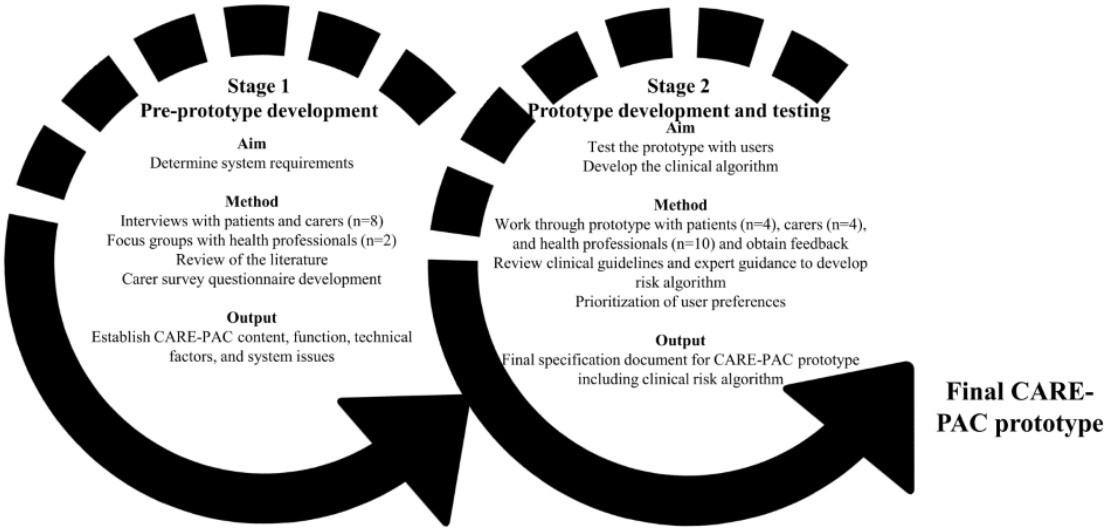

Figure 2. Final version of CARE-PAC: Care and Support System for Patients and Carers.

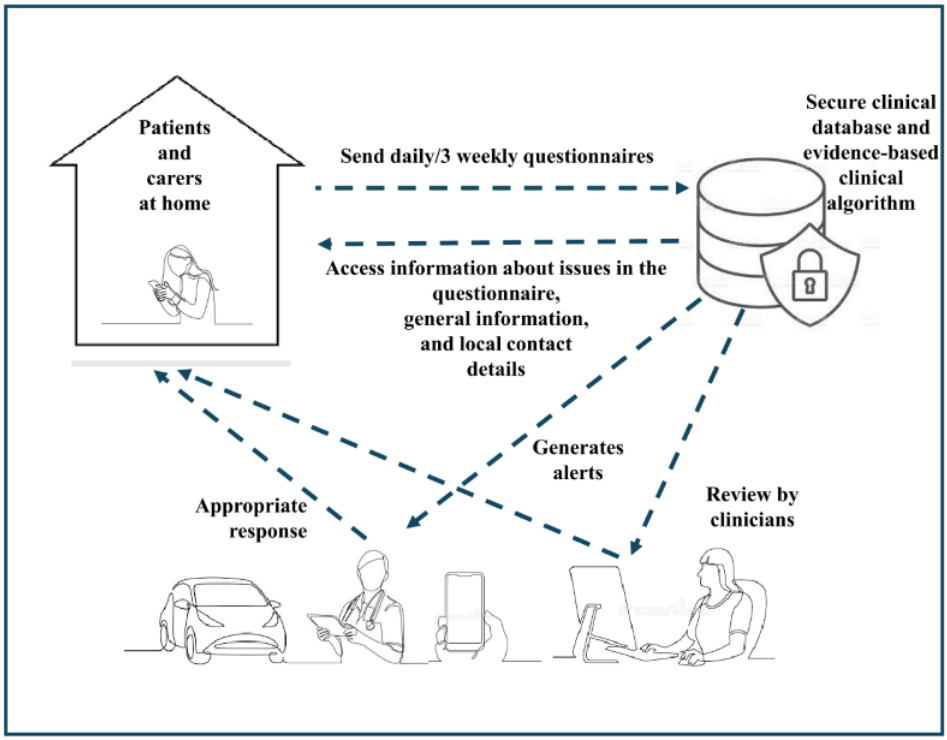

Aim of This Study

This study aimed to explore the usability, user experiences, and impact of CARE-PAC on the QoL, a digital dyadic remote monitoring and supportive care system for patients receiving PEOLC, their informal carers, and HCPs involved in their care.

Methods

Participants and Settings

Patient, informal carer, and HCP eligibility criteria for the study were set (Textboxes 1-3). The study was conducted in 5 clinical sites (3 hospices, 1 hospital, and 1 community setting) across Scotland and England. All participants were recruited from these 5 clinical sites. Patients and their informal carers were identified, approached, and recruited by HCPs involved in the study at these sites.

XSL•FO
RenderX

Textbox 1. Patient participation eligibility criteria.

Eligibility criteria:

- Diagnosed with one or more of the following illnesses: advanced cancer (any diagnosis) and advanced chronic respiratory diseases, such as chronic obstructive pulmonary disease (COPD), pulmonary fibrosis, or bronchiectasis
- Aware of their advanced diagnosis
- Registering one or more indicators on the Supportive Palliative Care Indicators Tool (SPICT), including both general and disease-specific indicators, as completed by a member of the multidisciplinary team
- Receiving palliative care services
- Aged 18 years or over
- Have an informal carer who is also willing to participate in the study
- Able to provide informed consent
- Deemed to be physically, psychologically, and cognitively fit to participate in the study, as confirmed by a member of the multidisciplinary clinical team
- Able to read, write, and understand English and have a basic understanding of computer and audiovisual software (for data and outcome data collection purposes)
- If possible, have access to a PC, laptop, tablet, smartphone, or other equipment that can run the Care and Support System for Patients and Carers (CARE-PAC) system and support audiovisual communication software

Textbox 2. Informal carer participation eligibility criteria.

Eligibility criteria:

- Identified by the patient (who meets the eligibility criteria listed in Textbox 1) as their key carer
- Aged 18 years or over
- Able to provide informed consent
- Able to read, write, and understand English and have a basic understanding of computer and audiovisual software (for data and outcome data collection purposes)
- If possible, have access to a PC, laptop, tablet, or smartphone or other equipment that can run the Care and Support System for Patients and Carers (CARE-PAC) system and support audiovisual communication software

Textbox 3. Health care professional participation eligibility criteria.

Eligibility criteria:

- A member of the multidisciplinary team caring for patients in any of the diagnostic groups identified based on the eligibility criteria listed in Textbox 1
- Working in or across any clinical setting that provides care to patients in any of the diagnostic groups identified based on the eligibility criteria listed in Textbox 1
- Have access to a PC, laptop, or other equipment that can support audiovisual communication software and a have basic understanding of computer and audiovisual software

Ethical Considerations

The study was approved by the North-West Preston Research Ethics Committee (reference 21/NW/0060) and sponsored by the University of Strathclyde. Research and development (R&D) approval was granted by each clinical site. The HCPs themselves provided informed consent to participate to the study's research fellow. All participants (patients, informal carers, and HCPs) received both written and verbal information about the study, along with a hands-on demonstration of CARE-PAC, and were provided with an opportunity to ask questions before providing written informed consent.

To help address the digital divide—such as disparities in access to internet-enabled devices—participants without the necessary technology were offered smartphones with SIM cards on loan. To further support inclusion, in-person training was provided to all individuals (patients, carers, and HCPs), including guided demonstrations using a training version of the system to build user confidence and familiarity. Participants also received tailored user manuals (for patients, carers, and HCPs), which featured step-by-step instructions, annotated screenshots, frequently asked questions (FAQs), and contact details for further support. These resources aimed to promote equitable engagement with the digital intervention by addressing not only access but also skills, knowledge, and confidence.

XSL•FO
RenderX

Day-to-Day Use of CARE-PAC

During the study, patients and their carers were invited to use CARE-PAC for up to 12 weeks, while HCPs at each clinical site participated throughout the study period.

Patients and Carers

Through the app, every day, patients and carers were prompted to complete a daily questionnaire. Patients completed the Integrated Palliative Care Outcome Scale (IPOS) [14], while carers completed the Digital Carer Daily Assessment (DiCDA), co-developed during stage 1 of CARE-PAC's design. Responses were automatically uploaded in real time to a secure clinical database accessible to the health care team. The daily questionnaire could only be submitted once per day.

Every 3 weeks, patients and carers were prompted to complete outcome measures—specifically, a QoL questionnaire [15,16] and the System Usability Scale (SUS) [17]. These were accessed immediately following the daily questionnaire and remained available for 2 consecutive days before becoming inactive until the next 3-week cycle.

At any time, and as frequently as required, participants could access the e-library.

At the end of their involvement in the study, patients and carers were instructed to delete the app from their device.

HCPs

HCPs at each clinical site established local protocols for incorporating daily checks of the clinician website into their routine workflow. These daily reviews enabled them to:

- Access participants' daily data reports to remotely monitor patients and carers, identify emerging symptom trends, and prioritize the clinical workload accordingly.
- View system-generated alerts, initiate appropriate clinical interventions, and document actions taken directly within the clinician website.

Outcome Data Collection and Synthesis

Given the advanced stage of disease in the study population, an attrition rate of approximately 30% was anticipated [18]. To mitigate potential data loss due to disease progression or death, outcome data were collected on a regular 3-weekly schedule. This approach ensured that all participants got an opportunity to contribute their experiences and outcomes using mixed methods (see Table 1), irrespective of their duration in the study.

Table 1. Mixed data collection methods across the study population: patients, carers, and HCPs[a].

Data collection method	Patient frequency	Carer frequency	HCP frequency
Quantitative data collection			
QoL[b] questionnaire	3-weekly	3-weekly	N/A[c]
SUS[d]	3-weekly	3-weekly	Start and end
Qualitative data collection			
Catch-up calls	3-weekly	3-weekly	N/A
In-depth interviews	End of participation	End of participation	N/A
Focus groups	N/A	N/A	End of study

[a]HCP: health care professional.
[b]QoL: quality of life.
[c]N/A: not applicable.
[d]SUS: System Usability Scale.

Quantitative Data Collection and Synthesis

Quality-of-Life Measurements

Patients' QoL was measured using the 21-item McGill Quality of Life-Expanded (MQoL-E) [15], which assesses the following domains: physical, psychological, social support, existential/spiritual, cognition, health care, environment, and the feeling of being a burden. Carers' QoL was measured using the 17-item Quality of Life in Life-Threatening Illness – Family Carer Version 3 (QOLLTI-F v3), a companion questionnaire to the MQoL-E that assesses 7 domains: environment, patient condition, caregiver's own condition, caregiver's outlook, QoL, relationships, and financial worries [16].

Both questionnaires were specific for this study's patient and carer population and were deemed relevant, easy to understand, and reflective of important aspects of the QoL, making their completion meaningful and valuable [19]. The questionnaires typically take 10-20 minutes to complete, with the completion time often decreasing with increased familiarity. Local principal investigators approved the use of these QoL questionnaires for inclusion in the study.

System Usability Assessment

Usability was assessed using the SUS, a quick 10-item tool designed to measure users' subjective perceptions of a system's usability [17], proven to be both reliable and valid in assessing system learnability and usability [20]. A SUS score of 68 or higher is considered above average, while a SUS score above 85/below 50 indicates excellent/poor user experience, respectively [20]. An adjective rating scale has been shown to correlate extremely well with total SUS scores and can be used to help understand how a total score translates into a judgement of overall usability [21].

XSL•FO
RenderX

Qualitative Data Collection and Synthesis

Catch-Up Calls

All patients and carers received a catch-up phone call every 3 weeks from a research team member (author MM), lasting 5-10 minutes. Participants were asked a series of questions about their experiences with CARE-PAC. These conversations were audio-recorded using an encrypted digital device, and responses were extracted from the recordings, with participants' views or verbatim quotes being documented in an anonymized, secure Microsoft Word document for each call.

Interviews and Focus Groups

At the end of participation, longer interviews were conducted with willing patients and carers, while all HCPs were invited to participate in focus groups. These interactions followed a topic guide based on the Non-adoption, Abandonment, Scale-up, Spread, and Sustainability (NASSS) framework [22], addressing topics such as perceptions of the technology, advantages and disadvantages of the system, and its impact/benefit. Both interviews and focus groups were audio-recorded using encrypted devices, transcribed verbatim, and reviewed for accuracy by a research team member prior to analysis.

Data Analysis

Data from all 5 study sites were collated, and unknown participant IDs or duplicate entries were removed. If a question was answered multiple times on the same day, the most up-to-date response was taken and the others removed.

Quantitative Data Analysis

QoL Questionnaires and the System Usability Scale

For all questions the Wilcoxon (Pratt) signed-rank test was used as the statistical test of differences. This test was used as the differences analyzed were paired as responses at the first and the last completion related to the same participant. The Pratt specification of the Wilcoxon test was used as there were some differences of zero that were of importance. In a standard Wilcoxon signed-rank test, these zero difference responses would be discarded. For the HCPs' SUS analysis, the Wilcoxon rank-sum test was used as baseline and endline responses were not paired.

Quality of Life

To examine differences in responses between the QoL questionnaire first and last completion, participants were included if they had at least 2 fully completed responses recorded on separate days. The difference in responses from first to last completion were examined at an individual question level, at a domain level using a domain score, and an overall QoL score level. To ensure a consistent QoL scale by domain, responses to some of the questions within each QoL questionnaire had to be transposed such that the new response was 10 minus the original response, which ensured that a response of 10 was the best QoL for each question. Within the patient QoL questionnaire, this related to 2 questions from the "physical" domain, 5 questions from the "thoughts and feelings" domain, and 1 question from the "social" domain. Within the carer QoL questionnaire, this related to 1 question in the "patient condition" domain, 1 question in the "your own condition" domain, 2 questions in the "relationships" domain, and 1 question in the "financial worries" domain.

System Usability Scale

Total SUS scores were calculated by first converting individual question responses to a consistent point system. This was done by subtracting 1 from the original score for odd-numbered questions and taking 5 minus the original score for even-numbered questions. Next, for each individual, their new responses to the 10 questions were summated, and the summed value was multiplied by 2.5. This gave a single total SUS score for each individual, which could take the value 0-100.

Mean (SD) values of overall and first and last SUS scores were calculated to examine differences in responses between the total SUS first and last completion score, and participants were included if they had at least 2 fully completed responses recorded on separate days.

Qualitative Data Analysis

Data from the catch-up calls, in-depth interviews, and focus groups were analyzed using thematic analysis, a method of identifying, analyzing, organizing, describing, and reporting themes found within a dataset [23]. NVivo software was used to analyze the data from in-depth interviews and focus groups. The initial coding framework was based on the interview/focus group topic guides, and data from the 2 groups (patients-carers and HCPs) were analyzed separately. Common overarching themes were identified in both the patient-carer and HCP data. However, there were instances where patients did not comment on specific subthemes highlighted by HCPs, as these were often related to the practicalities of conducting the research within clinical practice—an area outside the patients' experiences.

Data analysis, coding, and thematic findings from the catch-up calls, in-depth interviews, and focus groups were conducted by one of the research team members (MM) using a systematic thematic analysis approach. To ensure the rigor and validity of the findings, the emerging themes and subthemes were regularly presented to the wider research team during collaborative meetings. These discussions provided an opportunity to reflect on the data, challenge interpretations, and ensure that the analysis remained grounded in the participants' accounts. This collaborative review process facilitated a shared understanding of the data and supported the refinement of themes to ensure they accurately represented the findings. Through this iterative process, the research team concluded that thematic saturation had been reached, enhancing confidence in the robustness and credibility of the analysis.

Results

Overview

In total, 13 patient-carer dyads participated across the 5 clinical sites. Two participants, one patient, and one carer (unrelated) borrowed a study smartphone for the duration of their participation in the study.

XSL•FO
RenderX

Participants

The patient population consisted of 13 individuals (n=10, 76.9%, male; n=3, 23.1%, female) aged between 50 and 74 years. Most had a cancer diagnosis, including lung cancer (n=6, 46.2%), prostate cancer (n=3, 23.1%), pancreatic cancer (n=1, 7.7%), gastric cancer (1, 7.7%), and sarcoma (1, 7.7%), with 1 (7.7%) patient having a noncancer condition. The majority of carers were spouses (n=10, 76.9%), with others being siblings (n=2, 15.4%) or a parent (n=1, 7.7%).

The carer population comprised 13 individuals (n=11, 84.6% female; n=2, 15.4%, male), with ages ranging from 30 to 49 years.

The HCP group included 7 participants: 4 (57.1%) palliative care consultants and 3 (42.9%) research nurses or coordinators.

Quantitative Results

Quality of Life

Table 2 shows the completion and inclusion rates for the QoL questionnaire among patients and carers.

All QoL questions were answered on every completed attempt. Participants with at least 2 fully completed responses on separate days were included in the analysis of change over time, which examined individual questions, domain-level scores, and overall QoL scores.

At an individual question level, there were significant differences in patients' and carers' responses to some questions within their QoL questionnaires (Table 3).

Table 2. Completion and inclusion rates for the QoLa questionnaire among patients and carers.

Group	Total participants, N	Completed at least once, n (%)	Completed only once, n (%)	Completed ≥2 times, n (%)	Highest number of completions, n; participants, n (%)	Included in first vs last analysis, n (%)
Patients	13	11 (84.6)	3 (23.1)	8 (61.5)	5; 2 (15.4)	8 (61.5)
Carers	13	9 (69.2)	2 (15.4)	7 (53.8)	5; 1 (7.7)	7 (53.8)

[a]QoL: quality of life.

Table 3. Questions from the patient and carer QoL[a] questionnaires that had significantly different responses.

Questionnaire and domains	Question	Significance (%)	Direction of change
MQoL-E[b] questionnaire (patients)			
Overall QoL	Considering all parts of my life (eg, physical, emotional, social, spiritual, and financial) over the past 2 days (48 hours), the quality of my life was: Response: from "very bad" (0) to "excellent" (10)	5	Improved
Physical	Over the past 2 days (48 hours), being physically unable to do the things I wanted was: Response: from "not a problem" (0) to "a huge problem" (10)	5	Declined
Thoughts and feelings	Over the past 2 days (48 hours), I felt that the amount of control I had over my life was: Response: from "not a problem" (0) to "a huge problem" (10)	10	Declined
Social	Over the past 2 days (48 hours) communication with the people I care about was: Response: from "difficult" (0) to "very easy" (10)	5	Improved
Social	Over the past 2 days (48 hours), I felt my relationships with the people I care about were: Response: from "more distant than I would like" (0) to "very close" (10)	5	Improved
Social	Over the past 2 days (48 hours), I felt supported: Response: from "not at all" (0) to "completely" (10)	5	Improved
QOLLTI-F v3[c] questionnaire (carers)			
Your outlook	Presently, I feel that my life has meaning: Response: from "very little meaning" (0) to "very much" (10)	10	Declined

[a]QoL: quality of life.

[b]MQoL-E: McGill Quality of Life-Expanded.

[c]QOLLTI-F v3: Quality of Life in Life-Threatening Illness – Family Carer Version 3.

XSL•FO
RenderX

Regarding domains within each of the QoL questionnaires, a domain score was calculated for each participant as the mean of responses to questions in the domain. The patient group had significantly different domain scores from first to last completion for the domains: improved "overall quality of life" domain score, improved "social" domain score, and decreased "physical" domain score. The carers had no significant differences in domain scores.

The total QoL score was calculated as the mean of the domain scores for each participant. Neither the patients nor carers had significantly different overall QoL scores from first to last response.

System Usability Scale

Completion Rates

The SUS questionnaire was completed at least once by 9 (69.2%) of 13 patients and 10 (76.9%) of 13 carers. Furthermore, 2 (15.4%) carers and 3 (23.1%) patients completed the questionnaire only once, while the maximum number of completions was 5, achieved by 2 (15.4%) carers and 2 (15.4%) patients. All SUS items were completed on every occasion by all participants.

Participants were included in the analysis of change over time if they completed at least 2 fully completed SUS questionnaires on separate days. This reduced the sample size to 6 patients and 8 carers.

The SUS questionnaire was also administered to HCPs at baseline and endline. In total, 23 HCPs completed the baseline questionnaire, although 1 (4.3%) HCP omitted 5 items (questions 3, 5, 6, 8, and 9). The endline questionnaire was fully completed by 18 (78.3%) HCPs.

Summary of SUS Scores

Table 4 shows a summary of SUS completion and scores across participant groups.

Table 4. Summary of SUS[a] completion and scores across participant groups.

Group	Sample size, n/N (%)	SUS completion range	SUS score, mean (SD)	Highest number of completions, n; participants, n (%)	Significant change over time
Patients	9/13 (69.2)	1-5	87.9 (12.4)	5; 2 (15.4)	Yes (P<.10)
Carers	10/13 (76.9)	1-5	94.7 (3.8)	5; 2 (15.4)	No
HCPs[b]	Baseline: 23 (78.3) Endline: 18 (78.3)	1 (baseline and endline)	63.6 (15.6)	N/A[c]	No (P=.77)

[a]SUS: System Usability Scale.

[b]HCP: health care professional.

[c]N/A: not applicable.

The combined first and last SUS scores for patients and carers indicated excellent usability of the CARE-PAC app. Based on the adjective rating scale associated with SUS scores, carers rated the app as extremely usable (mean 94.7, SD 3.8), and patients also rated it as highly usable (mean 87.9, SD 12.4).

For HCPs, the mean SUS score across baseline and endline responses was 63.6 (SD 15.6), slightly below the industry-average benchmark of 68. This fell between the SUS adjective ratings of "OK" and "good." A comparison of baseline and endline scores using an independent Welch t-test showed no statistically significant change in HCP usability scores (P=.77).

Qualitative Outcomes

Themes Identified

Overarching themes and subthemes from patient, carer, and HCP qualitative data (catch-up calls and longer interviews/focus groups) were identified (Table 5).

The overarching themes (impact on care, satisfaction and reflections, challenges in project implementation, and future directions) were common across both patients and carers, with a few exceptions when subthemes were only identified by HCPs. Therefore, the themes from all participants (patients, carers, and HCPs) were reported together. Quotes for each subtheme, including identifiers to denote the group to which each participant belonged, can be found in Multimedia Appendix 1.

XSL•FO
RenderX

Table 5. Themes and subthemes from patient, carer, and HCP[a] qualitative data.

Overarching theme	Subthemes
Impact on care experiences	• Proactive monitoring and early detection • QoL[b], empowerment, and sense of worth • Increased sense of reassurance and security • Usefulness of reputable information
Reflections and satisfaction	• Personal reflections • Perceived added value
Challenges in project implementation	• Technical issues • Project population • Recruitment • Resources, workload, and staffing
Future directions	• Sustainability • Suggestions for improvement

[a]HCP: health care professional.

[b]QoL: quality of life.

Impact on Care Experiences

This theme related to the added value that CARE-PAC had on experiences of care delivery and receipt. Subthemes identified were proactive monitoring and early intervention, impact on QoL and sense of worth, increased sense of reassurance and security, and usefulness of reputable information.

Proactive Monitoring and Early Intervention

Patients, carers, and HCPs agreed that CARE-PAC enables faster, more effective symptom sharing and earlier intervention. Patients felt more responsible for tracking and understanding their symptoms, while HCPs valued the ability to spot changes early, benefitting health care generally through reducing the need for more costly interventions to address symptoms at a more advanced stage.

Impact on QoL and Sense of Worth

Participation in the study was reported to positively impact patients' and carers' QoL and sense of worth. Using the CARE-PAC app helped them reflect on their well-being, feel more in control, and contribute meaningfully to future care, which was also recognized and valued by HCPs.

Increased Sense of Reassurance and Security

Patients, carers, and HCPs felt CARE-PAC provides a sense of security and peace of mind, reassuring users that someone was monitoring them and helping patients and carers feel less alone.

Usefulness of Reputable Information

The CARE-PAC e-library was valued by patients and carers as a convenient and trustworthy source of information, offering symptom-specific resources and local service contacts in one place. Although not all users were aware of it, those who accessed it found it helpful.

Reflections and Satisfaction

This theme captured participants' reflections on the personal significance of CARE-PAC and its broader impact on the palliative care community, highlighting the sense of satisfaction derived from using it. Two subthemes were identified, personal reflections and perceived added value.

Personal Reflections

Participants consistently reported that CARE-PAC is easy to use, making it easy to integrate into daily routines and professional practice. Patients and carers expressed a sense of loss when the study ended, reflecting its value in their lives. Although all groups recognized the growing importance of technology in health care and saw CARE-PAC as a timely and necessary innovation, they also acknowledged potential drawbacks, such as digital exclusion, increased anxiety, and reminders of illness. Despite these concerns, patients, carers, and HCPs all valued their involvement in the study, with many expressing gratitude for the opportunity to contribute.

Perceived Added Value

Patients, carers, and HCPs valued CARE-PAC's ability to flag clinically concerning symptoms early, enabling timely interventions. The inclusion of carers' perspectives was seen as a major strength, offering insight into both the patient's condition and the carer's well-being. CARE-PAC also helped improve communication between patients and carers, occasionally prompting meaningful discussions. HCPs appreciated being able to track symptom trends over time, helping them better prepare for patient contact and prioritize care based on clinical needs.

Challenges in App Implementation

Although most challenges were identified by HCPs, patients also highlighted some difficulties. Four subthemes emerged: technical issues; project population; recruitment; and resources, workload, and staffing.

Technical Issues

HCPs found that some parts of the clinician website are not user friendly, and patients identified awkward sections within the daily questionnaire.

XSL•FO
RenderX

Study Population

The digital divide and lack of confidence in using technology were seen as barriers. HCPs noted that some individuals are generally less engaged with health care. Late referrals meant many patients were too unwell to participate by the time the study was introduced to them. The dyadic design (patient and carer) posed recruitment challenges, especially when a carer was unavailable or hesitant or due to discomfort with the term "carer."

Recruitment

HCPs found the research-related recruitment process time-consuming. They also experienced frustration with low referrals from other teams and acknowledged a tendency to gatekeep by not encouraging unsure patients to participate.

Resources, Workload, and Staffing

Although participation and managing alerts were manageable, HCPs raised ongoing concerns about limited staffing, time, and resources to support project delivery.

Future Directions

Building on the positive experiences of CARE-PAC, participants suggested future directions focusing on 2 main themes, sustainability and system improvements.

Sustainability

All participant groups emphasized the need for appropriate resourcing to ensure the long-term sustainability of CARE-PAC. HCPs highlighted the potential for a "snowball effect," where increased use demonstrates value and encourages further adoption. CARE-PAC's applicability to a wider range of palliative care conditions was seen as key to broader use. Identifying the optimal time to introduce CARE-PAC—when patients and carers can meaningfully engage—was considered vital. HCPs also recommended embedding CARE-PAC into routine care to ease implementation by removing additional research-related tasks.

Suggestions for Improvements

Participants proposed greater personalization of CARE-PAC, such as customizing alert timings, notification methods, and report frequency. Patients and carers expressed a desire to add context to daily reports, track symptoms over time, and include relevant notes (eg, pain medication) to enhance understanding and reduce unnecessary follow-ups. A key improvement suggested was integration across all health care teams, as patients felt frustrated that not all their care providers could access the clinician platform.

Discussion

Principal Findings

CARE-PAC is a practical and impactful remote monitoring and support system that holds significant value for the evolving field of digital PEOLC. Patients, carers, and HCPs consistently reported positive experiences and offered meaningful recommendations to guide future implementation and long-term sustainability.

Comparison With Prior Work

The CARE-PAC feasibility study contributes to the growing body of work on digital palliative care systems. Comparing this study to prior work in the field reveals several key similarities and areas of advancement.

Symptom Monitoring, Early Detection, and Timely Intervention

Digital monitoring systems in palliative care have been shown to support early detection of escalating symptoms, enabling prompt clinical intervention and improving the QoL and patient-reported outcomes [24,25]. Platforms such as MyPal and similar oncology-based tools have demonstrated benefits, including better access to clinicians, faster responses, and more efficient care, particularly in rural settings [26-28]. Although CARE-PAC shares these strengths, it also introduces a novel dyadic approach by including carers in the monitoring process. This allows for proactive identification of carer concerns and well-being, enabling health care teams to respond to issues that, if left unaddressed, could lead to crisis situations, such as carer burnout and unplanned and unnecessary hospital admissions.

Patient and Carer Engagement

Engaging people nearing the end of life in research remains a challenge—often due to HCPs' hesitancy to initiate these conversations, based on assumptions about the vulnerability of patients and families and concerns about placing an additional burden on them [29,30]. Ethical and practical concerns, such as challenges in gaining consent, avoiding undue distress, and balancing care needs with research demands, have also been cited as barriers [31-33]. CARE-PAC implemented established existing strategies to enhance user engagement [34,35], including involving patients and carers in co-design and testing to ensure the digital intervention is meaningful, usable, and relevant. The overwhelmingly positive feedback from patients, carers, and HCPs involved in CARE-PAC challenges these assumptions, demonstrating that meaningful involvement at this stage of life is both feasible and ethically justifiable.

Feedback loops have previously been used as a means of encouraging engagement with a digital system [27,28]. CARE-PAC provides 2 forms of feedback: an overview of patient and carer daily reports over time, enabling remote monitoring by health professionals, and a clinical risk algorithm that generates real-time alerts when submitted reports indicate clinically concerning issues. Patients and carers reported feeling a heightened sense of security and reassurance knowing that CARE-PAC would alert health care teams to any clinically concerning reports. This perceived value is reflected in the high compliance rates for completing the daily questionnaires—72.9% for patients and 65.5% for carers—which surpassed the completion rates for outcome measures (QoL and SUS), which had no immediate relevance or impact on their care. Future iterations of CARE-PAC will introduce a third feedback loop, enabling patients and carers to view their own symptom reports over time. This added feature has the potential to further enhance user engagement by promoting reflection and self-awareness.

XSL•FO
RenderX

Usability and Feasibility

Although digital health systems offer significant promise for remote monitoring in palliative care, usability and feasibility remain critical as poor design can hinder user adherence. Challenges, such as intrusive physical monitoring, privacy concerns, and rigid symptom reporting, persist [36,37]. However, mobile phone-based systems have shown strong acceptability—particularly in rural settings—when supported by user-centered design and insights from user experience research [36,38]. CARE-PAC's mixed methods approach—integrating quantitative data with qualitative feedback—provides a comprehensive view of the user experience. In contrast with other systems [37,39], improvements in usability scores over time suggest CARE-PAC becomes more intuitive and user friendly with continued use.

Strengths and Limitations

A key challenge in measuring experiences and outcomes in PEOLC is that patients may report a positive care experience despite poor clinical outcomes [40]. This challenge was evident in the CARE-PAC patient population, where individuals—despite receiving timely interventions, effective symptom management, and support—often experienced ongoing health deterioration due to the progressive nature of their condition.

Strengths

Co-design With Stakeholders

Engaging people receiving PEOLC in research can be challenging due to concerns about burdening vulnerable individuals, symptom severity, and misconceptions about the value of such research. Additionally, limited research infrastructure and funding in palliative care often restrict opportunities for meaningful inquiry. However, many patients and carers are motivated to participate, seeing it as a way to improve care for others in the future. CARE-PAC embraced this willingness by actively involving patients, carers, and HCPs in the co-design of the system. This collaborative approach ensured that the final prototype was not only technically functional but also aligned with the lived realities, priorities, and needs of intended users.

By designing the evaluation to capture data at 3-week intervals, the study ensured that participants' experiences were included even if they withdrew or died before the study's end. This approach reflects a compassionate and pragmatic design that acknowledges the realities of PEOLC. During the feasibility study, patients and carers consistently expressed gratitude for the opportunity to participate—reinforcing that, even in the context of advanced illness, people value meaningful engagement in research and can have positive experiences doing so.

Multiple Perspectives in Evaluation

One of CARE-PAC's distinctive strengths is its inclusion of multiple perspectives in both its design and evaluation. The dyadic model—incorporating feedback and inputs from both patients and their informal carers—is a unique and innovative feature. Informal carers play a vital yet often undersupported role in delivering PEOLC, and their involvement in CARE-PAC acknowledges the multidimensional nature of caregiving. Including carers allowed for the identification of both patient condition and carer well-being, enabling HCPs to better understand the overall situation and potentially intervene to prevent carer burnout. Although clinicians noted some limitations in addressing carer-specific alerts due to the lack of background information, they saw value in these insights and suggested triaging carer needs through existing support services. Patients and carers also appreciated the real-time responsiveness of CARE-PAC, reporting that symptom data entered into the system influenced the support and treatment they received. The study's design/methods also ensured that outcome data were collected even in cases of attrition, maximizing the utility of participants' contributions regardless of duration of involvement.

Multisite Implementation

CARE-PAC was implemented across 5 clinical sites in Scotland and England, encompassing a range of care settings, including hospices, hospitals, and community services. The consistency of positive experiences across these diverse settings suggests that CARE-PAC has potential for broader generalizability. Patients and carers, regardless of location or diagnosis, engaged with the system in meaningful ways and reported similar benefits, supporting its potential scalability.

Limitations

Sample Limitations

Small Sample Size and Limited Diagnoses

The study involved a small sample of 13 patients (12 of whom had a cancer diagnosis) and 13 carers. Such a limited cohort restricts the generalizability of the findings and prevents the identification of statistically significant trends. The predominance of cancer diagnoses also limits insight into the system's use across the wider palliative care population, including those with nonmalignant conditions.

Gatekeeping and Selection Bias

HCPs acknowledged challenges in recruiting participants and admitted to a degree of gatekeeping during the recruitment process. As a result, those who enrolled in the study may not fully represent the broader palliative care population. Participants were also likely to be more comfortable with technology, which may have positively skewed findings related to usability and acceptance. More broadly, difficulties in recruitment may reflect a well-recognized challenge within UK palliative care: the tendency for conversations about terminal illness and end-of-life care to occur too late or not at all [41]. These delays can prevent timely identification of individuals who would benefit from supportive interventions, such as CARE-PAC.

Dyadic Requirement

The requirement for participants to enroll as a dyad (patient and carer) introduced further limitations. This criterion reduced the pool of eligible participants, particularly where patients lacked an identifiable carer or where the carer did not see themselves in that role. Additionally, the dyadic model posed a dependency: when one member of the pair withdrew from the project—on

XSL•FO
RenderX

2 occasions—it forced the other, who may have wished to continue, to also withdraw.

Health Professional Involvement

Only small clinical teams from each site were involved in the study, making it difficult to assess how the system would perform at scale within wider, more complex health care systems.

Duration of Participation

The 12-week participation period, although appropriate and respectful of the health trajectories of people receiving PEOLC, limited the ability to evaluate the system's long-term usability, sustainability, and impact. This time frame was practical, given anticipated symptom progression and potential attrition, but feedback indicated that some participants would have preferred to continue using CARE-PAC beyond 12 weeks. Several participants expressed a sense of loss when the study ended. A more flexible, patient-led participation period could be considered in future studies to better reflect user preferences and provide richer longitudinal data.

Data Collection and Evaluation Gaps

Impact of Alerts and Clinical Response

Although HCPs were required to acknowledge and close alerts via the clinician interface, there was no evaluation of these alerts, the interventions they triggered, or the resulting clinical outcomes. In addition, although patients and carers anecdotally reported that alerts prompted responsive care, these accounts were not formally assessed. Future research should include a structured evaluation of alert management and its clinical impact.

Quality-of-Life Data

The QoL dataset collected was small, limiting the conclusions that can be drawn about CARE-PAC's impact on the QoL. Without robust data, improvements can only be cautiously interpreted.

Future Implications

To build on the early findings of CARE-PAC and address identified limitations, future research should aim to recruit a larger and more diverse participant sample, including individuals with a broader range of life-limiting conditions to evaluate the impact of CARE-PAC on patient and carer outcomes and cost-effectiveness of the system. Allowing patients and carers to engage either jointly or independently, and removing fixed end dates to give participants greater control over the duration of system use, may enhance flexibility and real-world relevance. Expanding clinician involvement across a variety of professional disciplines and health care settings will also be essential for assessing scalability, usability, and integration within routine clinical practice. Additionally, a formal evaluation of the clinical interventions triggered by CARE-PAC alerts is needed to better understand their impact on care quality, safety, health care costs, and patient and carer outcomes.

A significant insight from this study relates to the tension between research-related tasks and sustainable clinical implementation. Integrating CARE-PAC into standard models of care—rather than embedding it within research-only contexts—will be essential to its long-term viability. The requirement for tasks such as repeated consent visits and the distribution of information sheets created additional burdens that may hinder uptake in routine settings. These findings reflect wider challenges in palliative care research, where balancing research protocols with time-pressured, resource-limited clinical environments is often difficult. It is also notable that although several pilot studies on digital monitoring in palliative care have been published, few have reported on longer-term implementation, suggesting ongoing barriers to sustained integration.

CARE-PAC advances existing work in this area by offering a unique combination of features: high compliance; real-time data sharing; integrated, tailored information provision; and early responses from health care teams. Importantly, it also incorporates informal carers not only in their support roles but also as individuals with their own needs and well-being considerations. Its focus on early detection, proactive clinical intervention, and crisis prevention signals a shift toward a more anticipatory and comprehensive model of palliative care. Moreover, subthemes identified by HCPs around resourcing and triage point to the practical realities of embedding remote monitoring within overstretched services. These insights will be vital in shaping future models of care delivery.

To support sustainable implementation at scale, future development of CARE-PAC should prioritize 2 key areas. First, the interface for HCPs should be refined to ensure it presents the most relevant information in a concise, accessible, and user-friendly format, reducing information overload and supporting rapid decision-making. Second, CARE-PAC should be fully interoperable with existing electronic health record systems, allowing information captured via CARE-PAC to be viewed and acted upon within the digital platforms already used by the multidisciplinary care team. These improvements will facilitate seamless integration into everyday clinical workflows and enhance CARE-PAC's appeal to providers. Broader deployment across a range of care settings—including community, hospice, and hospital services—will also be important to test CARE-PAC's adaptability and utility in diverse real-world contexts. Ultimately, digital innovations like CARE-PAC have the potential to significantly improve the experience of PEOLC by supporting proactive, person-centered, and coordinated care for both patients and the people close to them.

Conclusion

The co-design and feasibility testing of CARE-PAC demonstrates promising potential to enhance the experiences of patients receiving PEOLC and their informal carers. Reported benefits include an improved QoL, earlier detection of symptoms enabling timely interventions, increased reassurance, and better access to tailored information and support. These findings underscore the value of digital health solutions in addressing some of the complexities and uncertainties of PEOLC, while supporting a more proactive, person-centered approach.

Nevertheless, several implementation challenges must be addressed. Recruitment was difficult, given the frailty and

XSL•FO
RenderX

declining health of many eligible participants, and future delivery will require dedicated resources to support ongoing use in clinical practice. As health and social care services face increasing pressure due to rising demand and limited capacity, digital innovations like CARE-PAC offer a timely and scalable opportunity to strengthen palliative care provision. With appropriate investment and integration into routine systems of care, such approaches could help ensure that patients and informal carers receive responsive, compassionate, and coordinated support as they navigate the final stages of life.

Acknowledgments

This study was partly funded by a charitable donation, and the feasibility study was sponsored by the University of Strathclyde. The research team gratefully acknowledges the invaluable contributions of the patients, informal carers, and health care professionals within all phases of the study and across all participating clinical sites. Their time, insights, and willingness to share their experiences were essential to the successful co-design and development of the Care and Support System for Patients and Carers project and prototype.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

All authors contributed to the conceptualization, writing, and editing of the paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Participant quotes supporting qualitative themes.
[DOCX File , 24 KB - formative_v9i1e69394_app1.docx]

References

1. Palliative care. World Health Organization. 2023. URL: https://www.who.int/europe/news-room/fact-sheets/item/palliative-care [accessed 2025-08-19]
2. Research briefing - palliative and end of life care. UK Parliament Post. 2022. URL: https://post.parliament.uk/research-briefings/post-pn-0675/ [accessed 2025-08-19]
3. Johansson T, Pask S, Goodrich J, Budd L, Okamoto I, Kumar R, et al. Time to Care: Findings from a Nationally Representative Survey of Experiences at the End of Life in England and Wales. Research report. London, UK: Marie Curie; 2024.
4. Kallis G, Hodge G, Wheat H, Oh TM, Pearce S. Exploring the challenges experienced by patients and families using palliative and end-of-life care services: a qualitative focus group study. Palliat Support Care 2025 Mar 03;23:e70 [FREE Full text] [doi: 10.1017/S1478951525000057] [Medline: 40026034]
5. Cummins L, Julian S, Georghiou T, Kumar G, Scobie S. Public Expenditure in the Last Year of Life. Research report. London, UK: Marie Curie; 2025.
6. Luta X, Ottino B, Hall P, Bowden J, Wee B, Droney J, et al. Evidence on the economic value of end-of-life and palliative care interventions: a narrative review of reviews. BMC Palliat Care 2021 Jun 23;20(1):89 [FREE Full text] [doi: 10.1186/s12904-021-00782-7] [Medline: 34162377]
7. The palliative care crisis: inadequate access, care and outcomes in the face of rising demand. Cisley Saunders International. 2025. URL: https://csiweb.pos-pal.co.uk/csi-content/uploads/2024/11/Palliative-Care-Crisis.pdf [accessed 2025-08-19]
8. Hodge G, Kallis G, Oh TM, Wheat H, Pearce S. Exploring perceived barriers to palliative and end of life care provision in South-West England: bringing together the perspectives of professionals, patients, and families. Front Sociol 2024;9:1488688 [FREE Full text] [doi: 10.3389/fsoc.2024.1488688] [Medline: 39902145]
9. Accelerating digital health transformation in Europe: a two-year progress report. World Health Organization. 2024. URL: https://tinyurl.com/ypn5zahy [accessed 2025-08-19]
10. Finucane AM, O'Donnell H, Lugton J, Gibson-Watt T, Swenson C, Pagliari C. Digital health interventions in palliative care: a systematic meta-review. NPJ Digit Med 2021 Apr 06;4(1):64 [FREE Full text] [doi: 10.1038/s41746-021-00430-7] [Medline: 33824407]
11. Widberg C, Wiklund B, Klarare A. Patients' experiences of eHealth in palliative care: an integrative review. BMC Palliat Care 2020 Oct 14;19(1):158 [FREE Full text] [doi: 10.1186/s12904-020-00667-1] [Medline: 33054746]

XSL•FO
RenderX

12. Phongtankuel V, Adelman RD, Reid MC. Mobile health technology and home hospice care: promise and pitfalls. Prog Palliat Care 2018;26(3):137-141 [FREE Full text] [doi: 10.1080/09699260.2018.1467109] [Medline: 30505077]
13. Mills J. Digital health technology in palliative care: friend or foe? Prog Palliat Care 2019 Aug 13;27(4):145-146 [FREE Full text] [doi: 10.1080/09699260.2019.1650420]
14. Murtagh FE, Ramsenthaler C, Firth A, Groeneveld EI, Lovell N, Simon ST, et al. A brief, patient- and proxy-reported outcome measure in advanced illness: validity, reliability and responsiveness of the Integrated Palliative care Outcome Scale (IPOS). Palliat Med 2019 Sep;33(8):1045-1057 [FREE Full text] [doi: 10.1177/0269216319854264] [Medline: 31185804]
15. Cohen SR, Russell LB, Leis A, Shahidi J, Porterfield P, Kuhl DR, et al. More comprehensively measuring quality of life in life-threatening illness: the McGill Quality of Life Questionnaire - Expanded. BMC Palliat Care 2019 Oct 31;18(1):92 [FREE Full text] [doi: 10.1186/s12904-019-0473-y] [Medline: 31672131]
16. Cohen R, Leis AM, Kuhl D, Charbonneau C, Ritvo P, Ashbury FD. QOLLTI-F: measuring family carer quality of life. Palliat Med 2006 Dec 01;20(8):755-767 [FREE Full text] [doi: 10.1177/0269216306072764] [Medline: 17148530]
17. Brooke J. SUS: a "quick and dirty" usability scale. In: Jordan PW, Thomas B, Weerdmeester BA, McClelland AL, editors. Usability Evaluation in Industry. London, UK: Taylor and Francis; 1996.
18. Oriani A, Dunleavy L, Sharples P, Perez Algorta G, Preston NJ. Are the MORECare guidelines on reporting of attrition in palliative care research populations appropriate? A systematic review and meta-analysis of randomised controlled trials. BMC Palliat Care 2020 Jan 09;19(1):6 [FREE Full text] [doi: 10.1186/s12904-019-0506-6] [Medline: 31918702]
19. Axelsson L, Alvariza A, Carlsson N, Cohen S, Sawatzky R, Årestedt K. Measuring quality of life in life-threatening illness - content validity and response processes of MQOL-E and QOLLTI-F in Swedish patients and family carers. BMC Palliat Care 2020 Mar 25;19(1):40 [FREE Full text] [doi: 10.1186/s12904-020-00549-6] [Medline: 32213170]
20. Sauro J. A Practical Guide to the System Usability Scale: Background, Benchmarks & Best Practices. Denver, CO: Measuring Usability LLC; 2011.
21. Bangor A, Kortum P, Miller J. Determining what individual SUS scores mean: adding an adjective rating scale. J Usability Stud 2009 May 1;4(3):114-123 [FREE Full text]
22. Greenhalgh T, Maylor H, Shaw S, Wherton J, Papoutsi C, Betton V, et al. The NASSS-CAT tools for understanding, guiding, monitoring, and researching technology implementation projects in health and social care: protocol for an evaluation study in real-world settings. JMIR Res Protoc 2020 May 13;9(5):e16861 [FREE Full text] [doi: 10.2196/16861] [Medline: 32401224]
23. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2008 Jul 21;3(2):77-101 [FREE Full text] [doi: 10.1191/1478088706qp063oa]
24. Calton B, Rabow M, Branagan L, Dionne-Odom J, Parker Oliver D, Bakitas M, et al. Top ten tips palliative care clinicians should know about telepalliative care. J Palliat Med 2019 Aug;22(8):981-985 [FREE Full text] [doi: 10.1089/jpm.2019.0278] [Medline: 31237467]
25. Mao S, Liu L, Miao C, Wang T, Chen Y, Jiang Z, et al. Electronic symptom monitoring for home-based palliative care: a systematic review. Palliat Med 2024 Sep;38(8):790-805 [FREE Full text] [doi: 10.1177/02692163241257578] [Medline: 38835179]
26. Maramis C, Karamanidou C, Schera F, Kiefer S, Koumakis L, Marias K. Using electronic patient reported outcomes to foster palliative cancer care: the MyPal approach. 2019 Presented at: IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE); October 28-30, 2019; Athens, Greece p. 405-409. [doi: 10.1109/bibe.2019.00079]
27. Scarfò L, Karamanidou C, Doubek M, Garani-Papadatos T, Didi J, Pontikoglou C, et al. MyPal ADULT study protocol: a randomised clinical trial of the MyPal ePRO-based early palliative care system in adult patients with haematological malignancies. BMJ Open 2021 Nov 02;11(11):e050256 [FREE Full text] [doi: 10.1136/bmjopen-2021-050256] [Medline: 34728446]
28. Bonsignore L, Bloom N, Steinhauser K, Nichols R, Allen T, Twaddle M, et al. Evaluating the feasibility and acceptability of a telehealth program in a rural palliative care population: TapCloud for palliative care. J Pain Symptom Manage 2018 Jul;56(1):7-14 [FREE Full text] [doi: 10.1016/j.jpainsymman.2018.03.013] [Medline: 29551433]
29. White C, Hardy J. What do palliative care patients and their relatives think about research in palliative care? A systematic review. Support Care Cancer 2010 Aug;18(8):905-911 [FREE Full text] [doi: 10.1007/s00520-009-0724-1] [Medline: 19705165]
30. Middlemiss T, Lloyd-Williams M, Laird BJ, Fallon MT. Symptom control trials in patients with advanced cancer: a qualitative study. J Pain Symptom Manage 2015 Nov;50(5):642-649.e1 [FREE Full text] [doi: 10.1016/j.jpainsymman.2015.05.009] [Medline: 26031710]
31. Keeley PW. Improving the evidence base in palliative medicine: a moral imperative. J Med Ethics 2008 Oct;34(10):757-760 [FREE Full text] [doi: 10.1136/jme.2007.022632] [Medline: 18827110]
32. LeBlanc TW, Wheeler JL, Abernethy AP. Research in end-of-life settings: an ethical inquiry. J Pain Palliat Care Pharmacother 2010 Sep;24(3):244-250 [FREE Full text] [doi: 10.3109/15360288.2010.493579] [Medline: 20718645]
33. Beaver K, Luker K, Woods S. Conducting research with the terminally ill: challenges and considerations. Int J Palliat Nurs 1999 Jan;5(1):13-17 [FREE Full text] [doi: 10.12968/ijpn.1999.5.1.9926]

XSL•FO
RenderX

34. Voorheis P, Petch J, Pham Q, Kuluski K. Maximizing the value of patient and public involvement in the digital health co-design process: a qualitative descriptive study with design leaders and patient-public partners. PLOS Digit Health 2023 Oct;2(10):e0000213 [FREE Full text] [doi: 10.1371/journal.pdig.0000213] [Medline: 37878566]
35. Brotherdale R, Berry K, Branitsky A, Bucci S. Co-producing digital mental health interventions: a systematic review. Digit Health 2024 Apr 25;10:20552076241239172 [FREE Full text] [doi: 10.1177/20552076241239172] [Medline: 38665886]
36. Cafazzo JA, Leonard K, Easty AC, Rossos PG, Chan CT. The user-centered approach in the development of a complex hospital-at-home intervention. Stud Health Technol Inform 2009;143:328-333 [FREE Full text] [Medline: 19380956]
37. Steindal SA, Nes AAG, Godskesen TE, Holmen H, Winger A, Österlind J, et al. Advantages and challenges of using telehealth for home-based palliative care: systematic mixed studies review. J Med Internet Res 2023 Mar 13;25:e43684 [FREE Full text] [doi: 10.2196/43684] [Medline: 36912876]
38. Bolzani A, Ramsenthaler C, Hodiamont F, Burner-Fritsch IS, Bausewein C. Monitoring of palliative care symptoms and concerns in specialized palliative home care using an electronic version of the Integrated Palliative care Outcome Scale (Palli-MONITOR): protocol for a mixed-methods study. BMJ Open 2021 Jun 01;11(6):e042266 [FREE Full text] [doi: 10.1136/bmjopen-2020-042266] [Medline: 34078632]
39. Pavic M, Klaas V, Theile G, Kraft J, Tröster G, Guckenberger M. Feasibility and usability aspects of continuous remote monitoring of health status in palliative cancer patients using wearables. Oncology 2020;98(6):386-395 [FREE Full text] [doi: 10.1159/000501433] [Medline: 31336377]
40. NHS Improving Quality. Measuring experience of care in end-of-life care: an overview. NHS England. 2018. URL: https://www.england.nhs.uk/north/wp-content/uploads/sites/5/2018/06/Measuring_experience_in_care_in_end_of_life_care_-_An_overview.pdf [accessed 2025-08-19]
41. Ambitions for palliative and end of life care: a national framework for local action 2021-2026. NHS England. 2021. URL: https://www.england.nhs.uk/wp-content/uploads/2022/02/ambitions-for-palliative-and-end-of-life-care-2nd-edition.pdf [accessed 2025-08-19]

Abbreviations

A&E: accident and emergency
CARE-PAC: Care and Support System for Patients and Carers
e-library: electronic library
HCP: health care professional
MQoL-E: McGill Quality of Life-Expanded
NASSS: Non-adoption, Abandonment, Scale-up, Spread, and Sustainability
NHS: National Health Service
PEOLC: palliative and end-of-life care
QoL: quality of life
QOLLTI-F v3: Quality of Life in Life-Threatening Illness – Family Carer Version 3
SUS: System Usability Scale

Edited by A Mavragani; submitted 28.11.24; peer-reviewed by H Du, P Cook; comments to author 04.04.25; revised version received 23.07.25; accepted 19.08.25; published 23.09.25.

Please cite as:
Maguire R, McCann L, Singleton C, Perkins P, Minton O, McCann N, Longford E, McKeown AD, Kavanagh K, Miller M
Exploring the Impact of a Remote Monitoring System for Palliative and End-of-Life Care (CARE-PAC): Mixed Methods Feasibility Study
JMIR Form Res 2025;9:e69394
URL: https://formative.jmir.org/2025/1/e69394
doi:10.2196/69394
PMID:

©Roma Maguire, Lisa McCann, Claire Singleton, Paul Perkins, Ollie Minton, Nicola McCann, Emma Longford, Alistair Dermot McKeown, Kimberley Kavanagh, Morven Miller. Originally published in JMIR Formative Research (https://formative.jmir.org), 23.09.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

User-Reported Mechanisms of Change on a Suicide Prevention Website: Single-Arm Pragmatic Trial

Martina Fruhbauerova[1], PhD; David Huh[2], PhD; Ursula Whiteside[2,3], PhD

[1]Department of Psychology, University of Kentucky, Lexington, KY, United States

[2]Now Matters Now, Bellevue, WA, United States

[3]Department of Psychiatry and Behavioral Sciences, School of Medicine, University of Washington, Seattle, WA, United States

Corresponding Author:
Ursula Whiteside, PhD
Now Matters Now
1915 140TH AVE NE
STE D2 PMB1012
Bellevue, WA, 98005
United States
Phone: 1 206 880 1549
Email: ursulawhiteside@gmail.com

Abstract

Background: Digital platforms can serve as effective interventions for individuals in crisis, those with limited access to mental health resources, or those who prefer web-based support over in-person care. NowMattersNow.org, a web-based platform grounded in dialectical behavior therapy, has been shown to reduce suicidal thoughts and negative emotions. However, little is known about the specific mechanisms that drive these improvements. Identifying the active ingredients that contribute to its effectiveness will help optimize its impact.

Objective: This study examined the reasons users reported behind reductions in suicidal thoughts and negative emotions after visiting NowMattersNow.org. Specifically, this study sought to determine which reported reasons were associated with greater versus lesser improvements and whether these changes differed across specific subgroups.

Methods: In this single-arm pragmatic trial, data were collected from 3185 respondents who completed a 6-item retrospective survey while visiting NowMattersNow.org. The survey assessed changes in suicidal ideation and emotional distress (ie, intensity upon entering the site vs at the time of survey completion), reasons the website was helpful, and basic nonexclusive demographic information. Cross-tabulations were used to examine the most commonly endorsed reasons for finding the website helpful, while longitudinal regression analyses assessed the statistical significance of changes in suicidal ideation and emotional distress.

Results: The majority of participants reported experiencing suicidal thoughts (n=2309, 72.5%) and negative emotions (n=2745, 86.2%) upon arriving at the website, with 52.4% (n=1211) and 55.6% (n=1527) of these individuals, respectively, experiencing reductions in suicidal thoughts and negative emotions after engaging with the site. Regarding the primary aims of the study, the most frequently cited reason for finding NowMattersNow.org helpful was “I learned something” (n=668, 21%), followed by “It distracted me” (n=544, 17.1%) and “I felt less alone” (n=414, 13%). These were also the top 3 reasons reported by LGBTQI individuals, those endorsing alcohol or opioid problems, and those experiencing unusual experiences, though the order varied across groups. Among participants who experienced the largest reduction in suicidal ideation (a 4-point decrease), the most common reasons cited were “It distracted me” (n=5, 29.4%), “I felt less alone” (n=3, 17.6%), and “I felt cared for” (n=3, 17.6%). Similarly, for those with the largest reduction in negative emotions (a 4-point decrease), the most frequently endorsed reasons were “It distracted me” (n=3, 23.1%), “I felt less alone” (n=3, 23.1%), and “I felt cared for” (n=2, 15.4%).

Conclusions: The findings suggest that NowMattersNow.org is an accessible, scalable digital intervention that shows promise for reducing suicidal ideation and emotional distress, particularly in vulnerable populations. Key elements, such as fostering social connectedness, distraction, and educational content, appear to be critical components of its effectiveness, indicating that web-based self-help tools like NowMattersNow.org can provide short-term management of suicidal thoughts and negative emotions.

(JMIR Form Res 2025;9:e71136) doi:10.2196/71136

XSL•FO
RenderX

KEYWORDS

suicide; dialectical behavior therapy; digital mental health; web-based interventions; suicide prevention; digital accessibility internet; help-seeking behavior; behavior therapy; crisis intervention

Introduction

Background

Suicide is a leading cause of death in the United States and represents a major public health crisis [1]. In 2022, over 13.2 million individuals in the United States seriously thought about suicide, 1.6 million attempted suicide, and 49,000 people died by suicide [2]. Many of these individuals had seen a health care provider in the month prior to their death [3,4]. However, many providers may not have the opportunity to intervene due to several factors: many individuals are uncomfortable discussing their suicidal thoughts with providers [5,6], suicidal ideation can fluctuate rapidly [7,8], and some suicide attempts are entirely unplanned [9-11]. Importantly, people experiencing suicidal thoughts often turn to the internet for support and information, with use tending to rise as suicidal thoughts intensify [12-15]. These findings reinforce the need for readily accessible and effective digital resources for those reaching out during critical moments. Indeed, researchers are calling for suicide-specific web-based tools for short-term management of suicidal thoughts and negative emotions [16].

Despite significant progress in developing empirically supported psychological treatments, a large proportion of people with mental health issues still does not receive these interventions. Closing this treatment gap is crucial. Web-based resources present scalable and accessible solutions that have the potential to extend help to a broader population and improve access to necessary mental health support. For example, a meta-analysis of 66 randomized controlled trials examining app-supported smartphone interventions for mental health problems found these platforms to be highly effective [17]. Specifically, smartphone interventions significantly outperformed control conditions in reducing depressive symptoms, generalized anxiety, stress, social anxiety, and general psychiatric distress, while also improving quality of life and positive affect. Importantly, these smartphone-based interventions were similar in their effectiveness to traditional active treatments, such as face-to-face or computerized therapy.

Several systematic reviews and meta-analyses have shown promising results for web-based suicide-specific prevention strategies as well. One systematic review and meta-analysis found that self-guided digital interventions aimed directly at reducing suicidal ideation were effective immediately after the intervention [18]. Another review indicated that, although there was not yet enough evidence to classify web-based cognitive behavioral therapy (CBT), mindfulness-based CBT, and dialectical behavioral therapy as evidence-based treatments for suicidal ideation and self-harm, digital interventions could be a safe and acceptable alternative for those unable or unwilling to engage in face-to-face therapy [19]. A separate systematic review and meta-analysis using individual participant data supported the efficacy of digital CBT in reducing suicidal ideation across different ages, genders, and histories of suicide attempts, though data on its effects on actual suicide attempts were limited [20]. Another review of web-based and mobile apps designed to prevent suicidal ideation and self-harm showed that digital interventions were linked to reductions in suicidal ideation scores postintervention [21]. However, one meta-analysis suggested that while internet- and mobile-based interventions had a positive effect on reducing suicidal ideation compared to usual treatment, this effect was not significant when compared to active control groups [22].

Similarly, a systematic review of mobile apps (such as iBobbly, Virtual Hope Box, BlueIce, and Therapeutic Evaluative Conditioning) found positive outcomes for individuals at high risk for suicide or self-harm, including reduced depression, psychological distress, and self-harm, and improved coping self-efficacy [23]. However, none of these apps significantly reduced suicidal ideation compared to a control condition. Finally, another systematic review of mobile- and internet-based interventions for self-harm showed that these interventions were generally well-accepted and associated with reductions in suicidal thoughts and behaviors (STBs) compared to treatment as usual [22]. However, their efficacy was not significant when compared to trials with active control groups. In summary, these systematic reviews and meta-analyses suggest that digital interventions can be a safe, acceptable, and potentially effective option for reducing STBs, particularly for individuals who may not access traditional in-person treatments. However, their effectiveness compared to active control conditions remains mixed, highlighting the need for further research to better understand the contexts in which these digital approaches are most effective.

Dialectical behavior therapy (DBT) is a well-established, evidence-based treatment for suicidal behaviors [24], making it a promising candidate for adaptation into scalable and accessible web-based interventions. Originally developed for individuals with borderline personality disorder who experience chronic suicidal thoughts and who engage in nonsuicidal and suicidal self-injury [25], DBT is grounded in the biosocial theory [25]. This theory posits that borderline personality disorder arises from the interaction between biologically predetermined emotional vulnerability (characterized by heightened emotional reactivity, intensity, and slow return to baseline) and invalidating environments, which can range from inattentive caregivers to physical or sexual abuse. This interplay impairs the individual's ability to trust, understand, and regulate emotions, leading to maladaptive behaviors that serve as escape mechanisms—suicidal behavior being the most extreme form.

Given DBT's strong emphasis on targeting STBs, digital adaptations hold significant potential for expanding access to evidence-based suicide interventions. However, there is limited evidence supporting its effectiveness for managing STBs through internet-based platforms. Still, some digital tools incorporating DBT principles have already demonstrated positive outcomes. For instance, CALMA, a DBT-based app

XSL•FO
RenderX

designed to reduce nonsuicidal and suicidal self-directed violence, has exhibited reductions in suicidal ideation, suicide plans, suicidal gestures, thoughts about nonsuicidal self-injury, and nonsuicidal self-injury itself from pre- to postintervention [26]. Similarly, an internet-delivered DBT skills training program for suicidal individuals with heavy episodic drinking showed significant reductions in suicidal ideation, alcohol use, and emotion dysregulation over 4 months [27]. Additionally, DBT Coach, a mobile app aimed at improving skills generalization through digital DBT skills coaching, was found to reduce subjective distress and urges to self-harm following app use [28]. Furthermore, other studies have indicated that delivering DBT through web-based platforms or apps is feasible, accessible, acceptable, and safe [29], and is generally well-received by participants [30]. In contrast, a pragmatic randomized clinical trial challenged the effectiveness of web-based DBT skills training [31]. Among adult outpatients with frequent suicidal ideation, those who received an invitation to skills training had approximately 30% higher risk of self-harm compared to those receiving usual care. This increased risk may be attributed to the trial's pragmatic design, where only 30% to 40% of participants actively engaged with the offered interventions. Among those who did engage, "active" participation was minimal; fewer than 10% viewed content beyond the brief introductory video, and only 2% sustained engagement for more than 9 months. These findings suggest that the unfavorable results may stem from underutilization of the service rather than the intervention's ineffectiveness, highlighting the importance of addressing barriers to sustained engagement in web-based treatments. Although the evidence for internet-based DBT interventions for STBs is still limited and mixed, several digital tools incorporating DBT principles have shown promising outcomes in reducing suicidal ideation, self-harm, and related symptoms, highlighting their potential for bridging the gap in immediate assistance for individuals at risk for suicide.

In addition to digital DBT interventions, there has been a rise in self-help websites aimed at making DBT resources more widely available, such as dbtselfhelp.com, dialecticalbehaviortherapy.com, dbt.tools, or dbtbites.com. While these websites hold significant potential for expanding access to web-based DBT resources, they lack empirical evidence supporting their effectiveness [32]. One exception is NowMattersNow.org, which was designed to support individuals experiencing STBs, as well as their providers. This website is grounded in DBT principles and features personal stories of successful coping and recovery from suicidal thoughts and painful emotions. In their study evaluating the utility of the website [16], 70% of survey respondents reported suicidal ideation upon entering the website. The results indicated a statistically significant reduction in suicidal thoughts, with an average decrease of 0.21 points (d=0.13) on a 1 to 5 scale, and a reduction in negative emotions of 0.32 points (d=0.21). Importantly, individuals with high suicidal ideation (ratings of 4 or 5) experienced a reduction of 6.4% points (d=0.13), while those with high negative emotions saw a 10.9% point reduction (d=0.23). Despite the small effect sizes, these reductions are still clinically significant, as even modest improvements in suicidal ideation and negative emotions, along with access to crisis management tools, could potentially impact individuals' ability to manage their distress and prevent suicidal behaviors.

Given the promising outcomes associated with NowMattersNow.org and other digital interventions aimed at suicide prevention, it raises important questions about the active ingredients and content features that drive their effectiveness. Understanding the mechanisms behind the reductions in suicidal ideation and negative emotions observed on NowMattersNow.org is crucial for optimizing its impact. Some potential mechanisms have been proposed in the suicidology field that may explain the effectiveness of such interventions. For example, "thwarted belongingness," or the sense of disconnection from others, is a well-recognized risk factor for suicidal ideation in major suicide theories [33,34]. This construct also aligns with the biosocial theory [25], in which chronically invalidating environments contribute to interpersonal difficulties, leaving individuals feeling isolated and disconnected. The videos on NowMattersNow.org, which feature individuals sharing their personal journeys through suicidal ideation and recovery, may counteract this disconnection, fostering a sense of connection and reducing feelings of isolation. Additionally, the putative mechanism of "caring contacts [35]"—the only intervention shown to have reduced suicide deaths in clinical trials [36,37]—is thought to operate through strengthening social connection [38]. In this way, NowMattersNow.org may help visitors feel less alone, which could partially explain the observed reductions in suicidal ideation and negative emotions. However, further investigation is needed to identify which features of the website contribute most effectively to these outcomes. Finally, it is important to examine whether high-risk and marginalized populations, who are often underrepresented in digital mental health research, benefit from interventions like NowMattersNow.org. If effective, such platforms could serve as a critical resource for reaching individuals who might otherwise have limited access to evidence-based suicide prevention support.

This Study

The next step in evaluating the effectiveness of the NowMattersNow.org website is to determine what features of the website are the most useful and effective in order to expand and refine the content so that the website's effectiveness can be improved even further, potentially saving lives. Additionally, discerning what users find helpful would also provide opportunities for future content personalization. For example, if some individuals benefit most from feeling less isolated or are encouraged to call a crisis line, the website could be tailored in the future to meet these specific needs. As such, this study was designed to extend the previous work of Whiteside et al [16] and determine the reasons for reductions in suicidal thoughts and negative emotions after visiting the NowMattersNow.org website. We hypothesized that participants who reported decreased suicidal thoughts and negative emotions after using the website would primarily attribute their improvement to feeling "less alone" from the content provided. We also aimed to investigate which reported reasons were associated with a greater or lesser magnitude of improvements in suicidal ideation and negative emotions. Due to the limited literature on the mechanisms through which websites make an

XSL•FO
RenderX

impact, we did not have a specific hypothesis regarding which reasons would be associated with varying degrees of improvement. Finally, we investigated whether the changes in suicidal ideation and negative emotions differed across specific subgroups. As this was an exploratory aim, we did not generate a priori hypotheses for these analyses.

Methods

Study Design

This study used a single-arm pragmatic trial design, leveraging quality improvement data from the NowMattersNow.org web-based platform. The website and survey were not originally created for research purposes, nor were they designed for any specific study. Participants were real-world users of the site, with natural variations in how they engaged with the content. This approach allowed for a more ecological and generalizable assessment of the website's impact, reflecting typical user interactions.

Procedures

Participants for this study were recruited through the NowMattersNow.org website. The survey was designed to activate after a specific duration of engagement with the site: 1 minute on mobile devices, and initially 3 minutes on desktop and tablet devices, which corresponds to the average length of one short video. However, due to limitations in the technology at the time, the survey could only be set to trigger after 1 minute of engagement on mobile devices. Unfortunately, this created discrepancies in the activation times between mobile and desktop users. In 2015, the activation time for desktop and tablet users was extended to 8 minutes—approximately the length of 2 short videos or 1 longer video—to encourage deeper engagement with the content before prompting the survey. While we could confirm that participants remained on a page long enough to trigger the survey, we could not track their total time spent across multiple pages due to the absence of cookies; this limitation was implemented to protect user privacy, given the sensitive nature of the website's content. Once the survey was activated, users were required to complete it before proceeding to other site features, although they could close the survey window using a button located in the upper right corner. After a survey was displayed—regardless of whether it was completed or not—participants would not be prompted again for at least 24 hours. We did not track how often the survey appeared without being completed. Prior to the launch of the study and data collection, the survey was rigorously tested by developers at Civilization [39] to ensure functionality and usability. Throughout this process, our only interaction with participants was through their survey responses, with no direct contact before or after submission.

NowMattersNow.org

The website aims to equip individuals experiencing suicidal thoughts and health care providers with DBT skills, serving as a valuable resource for supporting suicidal patients. The long-term objective is for providers to use these skills in their own lives and teach them to patients, thereby helping to prevent and manage suicidal crises. Consequently, the website was developed with health care providers in mind as a tool to improve care for suicidal patients. To illustrate how individuals can receive support, visitors might search for DBT skills and find real-life examples of how others have successfully used these skills to cope with distressing emotions, offering both guidance and validation. Additionally, the site provides access to a variety of crisis helplines tailored to different groups, helping users connect to appropriate support based on their needs. Many users also report feeling comforted by hearing personal recovery stories shared by others who have faced similar challenges. The NowMattersNow.org landing page, visible during the study period, reassures visitors with the message, "Have you had suicidal thoughts? Problems that felt unsolvable? You are in excellent company—we've been there." Further details on the origins and development of the NowMattersNow.org website can be found in the original study [16].

Measures

NowMattersNow.org routinely collects survey data during user sessions to improve quality and assess the website's impact on various populations. The survey consists of 7 items developed using HTML elements within WordPress and triggered by JavaScript. It is not publicly advertised and does not require a user login, ensuring confidentiality. The survey begins with a header stating, "We'd appreciate your confidential feedback to improve our site."

Users are prompted to retrospectively evaluate their "Intensity of negative emotions" and "Intensity of suicidal thoughts" on a 5-point scale from 1 to 5, ranging from "Not at all" to "Completely overwhelming." The survey captures both baseline and postwebsite ratings within a single session, asking participants to report their emotional state at 2 time points: "When you entered this site" (baseline) and "What level are they now?" (postwebsite use). As such, participants were asked to retrospectively report their emotional state before arriving at the website and then rate their current state after engaging with the site. Since the survey is completed in one sitting, participants who do not finish it would have incomplete responses. Additionally, users can retake the survey after 24 hours. These survey items were adapted from DBT skills "diary cards," which patients use to track suicidal thoughts, mood, and related behaviors.

Following these assessments, users are asked to identify what aspects of NowMattersNow.org they found helpful, selecting the top reason for feeling at least a little better. The options provided include: "It distracted me," "I felt less alone," "I learned something," "I called/texted a crisis line," "I saw that others could recover," "I felt cared for," "Time passed," "It wasn't helpful," and "Other." No follow-up question was included to further explore the "Other" responses provided by participants.

The survey also collects basic demographic information using nonexclusive checkboxes to identify users who identify as native or Indigenous, LGBTQI, or who have experienced alcohol or opioid problems, or unusual experiences. The term "unusual experiences" was not further defined for participants, but it was intended to capture a broad range of psychological phenomena

XSL•FO
RenderX

(eg, hallucinations and mania) in everyday language, avoiding the use of diagnostic terminology.

Although survey responses were not assessed for completeness prior to submission (ie, incomplete responses were allowed), only participants who provided complete responses on their suicidal thoughts and negative emotions were included in the analysis (ie, if users did not select the top reason for feeling at least a little better, and they were still included in the analyses and their reason was coded as "No response"). Once users clicked the "submit" button, their responses became final. Further details about the study measures can be found in the original paper [16].

Ethical Considerations

This study was approved by the Allendale Institutional Review Board (Old Lyme, Connecticut; IRB #: 20180223W1, Title: "Review of NowMattersNow.org Survey Data"). This study used previously collected quality improvement data from the NowMattersNow.org website. As the data were recorded in a manner that ensured participants could not be identified, either directly or through linked identifiers, the study qualified for an exemption from the HHS human participant regulations (under 45 CFR 46.101(b)(4)). Since no protected health information was collected as part of the quality improvement process, the study met the criteria for exemption, and informed consent was not required. Participants were given the option to skip questions or close the window if they did not wish to participate, ensuring their ability to opt out. To maintain privacy and confidentiality, IP address snippets were removed, and each participant was assigned a unique identification number. The final deidentified dataset was stored securely on a password-protected Google Drive. No compensation was provided to participants for taking the survey on the NowMattersNow.org platform.

Participants

This study used a convenience sample derived from data from 5539 survey responders who visited NowMattersNow.org, collected from May 14, 2020, to September 12, 2023. During this period, there were no major modifications to the website; only minor updates, such as the addition of short videos, may have occurred. However, because web designers and content developers did not systematically track these changes, we do not have records detailing what modifications were made or when they took place. Overall, it is reasonable to assume that all participants were exposed to the same fundamental content throughout the study period. Additionally, we did not track how long survey respondents remained on the website beyond completing the survey.

In cases where multiple survey responses originated from the same IP address, only the first response was retained to ensure that the sample represented unique individuals. Due to this approach, some individuals may have been assigned different IP addresses if they accessed the website from different locations, while others may have shared the same IP address if they were using a communal network (eg, within the same facility). We focused our analysis on the subset of 3374 survey respondents who did not identify as a mental health care provider. Our final analytic sample included 3185 survey respondents who answered 4 questions regarding suicidal ideation and negative emotions (baseline and postwebsite). The final dataset was stored in Google Drive, to which authors MF, DH, and UW had access.

Analytic Approach

First, preliminary analyses were conducted to characterize the degree of missing data and whether participant characteristics differed when outcome data were complete versus incomplete. These descriptive analyses informed how to address missing data in subsequent inferential analyses. Second, longitudinal regression analyses using generalized estimating equations evaluated whether the magnitude and statistical significance of user-reported changes in suicidal ideation and negative emotion varied for specific subgroups, including those who (1) identified as Native or Indigenous, (2) identified as LGBTQI, (3) endorsed alcohol or opioid misuse, and (4) endorsed unusual experiences (eg, psychotic symptoms). An exchangeable working correlation matrix in conjunction with cluster robust standard errors accounted for the correlation of repeated measures (baseline and post) within participants.

The following equation represents the model for the suicidal ideation outcome:

$$\text{SuicidalIdeation}_{ti} = b_0 + b_1\text{Indigenous}_{ti} + b_2\text{LGBTQI}_{ti} + b_3\text{AlcoholOpioidMisuse}_{ti} + b_4\text{UnusualExperiences}_{ti} + b_5\text{Time}_{ti} + b_6(\text{Indigenous}_{ti} \times \text{Time}_{ti}) + b_7(\text{LGBTQI}_{ti} \times \text{Time}_{ti}) + b_8(\text{AlcoholOpioidMisuse}_{ti} \times \text{Time}_{ti}) + b_9(\text{UnusualExperiences}_{ti} \times \text{Time}_{ti}),$$

Where t indexes time point and i indexes individual. The predictors were coded as follows: time (0=baseline, 1=postwebsite), Indigenous identity (0=no, 1=yes), LGBTQI identity (0=no, 1=yes), alcohol or opioid misuse (0=no, 1=yes), and unusual experiences (0=no, 1=yes). This parameterization evaluated all of the subgroup effects in a single regression model and accommodated participants who belonged to multiple subgroups. The equation for the negative emotions outcome was parameterized similarly. Thus, all of the subgroup characteristics assessed in the survey (ie, Indigenous identity, LGBTQI identity, alcohol or opioid misuse, and unusual experiences) were included as covariates in each outcome analysis.

Third, we evaluated cross-tabulations of user-reported reasons why NowMattersNow.org was helpful by (1) the baseline severity of suicidal ideation or negative emotions (1=no symptoms, 2 to 3=moderate symptoms, and 4 to 5=high symptoms). This categorization was based on a core tenet of DBT—the diary card [25]. While the DBT manual does not explicitly define these categories, we drew on the manual's guidance, as well as our clinical judgment (2 authors are clinicians) and lived experience, to define these symptom severity ranges. Additionally, (2) we examined the change in suicidal ideation or negative emotions from baseline to postwebsite (improved, did not change, or worsened).

Statistical significance was evaluated based on the conventional threshold of $P<.05$.

XSL•FO
RenderX

Results

Preliminary Analyses

Of the 3374 eligible survey responses from nonproviders, 3185 (94.4%) had complete data on suicidal ideation and negative emotions (baseline and postwebsite), and 189 (5.6%) had incomplete outcome data. Participants with complete outcome data were more likely to identify as LGBTQI (n=1034, 32.5% vs n=24, 12.7%; χ^2_1=32.4; *P*<.001) and more likely to endorse having experienced unusual experiences (n=1054, 33.1% vs n=36, 19%; χ^2_1=16.1; *P*<.001) than those with incomplete outcome data. There was no statistically significant difference between respondents with complete versus incomplete outcome data with respect to indigenous identity (n=147, 4.6% vs n=10, 5.3%; χ^2_1=0.2; *P*=.67) or endorsement of alcohol or opioid problems (n=286, 9% vs n=10, 5.3%; χ^2_1=3.0; *P*=.08).

The outcome analyses focused on the 3185 responses with complete outcome data, as this increased the representation of at-risk subgroups (ie, LGBTQI, and endorsement of potential psychotic symptoms). In the final analytic sample, 32.5% (n=1034) identified as LGBTQI, 5% identified as Native or Indigenous, 33.1% (n=1054) endorsed having experienced unusual experiences, and 9% (n=286) reported alcohol or opioid problems. As noted in the Analytic Approach section, the 4 participant characteristics assessed in the survey were included as covariates in the outcome analyses, so subgroup effects in the outcome analyses were adjusted for other participant characteristics.

Descriptive Analyses: Changes in Suicidal Ideation and Negative Emotions

Table 1 summarizes the intensity of and changes in suicidal thoughts and negative emotions, as well as reasons for NowMattersNow.org being helpful within-participant subgroups and overall across all respondents (n=3185).

At the time of entry, 2309 (72.5%) participants endorsed at least some suicidal ideation, defined as a rating greater than 1 on a 5-point scale (1=not at all; 5=completely overwhelming). Similarly, 2745 (86.2%) participants reported experiencing at least some degree of negative emotions.

Following engagement with NowMattersNow.org, 1211 (38%) participants reported improvements in the intensity of suicidal thoughts; when considering only those who initially endorsed suicidal ideation (n=2309), 52.4% (n=1211) experienced a reduction. Overall, there was a 0.5-point reduction in the intensity of suicidal thoughts, from an average of 3.0 at baseline to 2.5 after viewing the website (1=not at all, 5=completely overwhelming). There was a greater reduction in suicidal ideation among (1) those who identified as LGBTQI (vs non-LGBTQI; estimate –0.21, 95% CI –0.28 to –0.14; *P*<.001), (2) those who endorsed alcohol or opioid problems (vs those without alcohol or opioid problems; estimate –0.13, 95% CI –0.24 to –0.02; *P*=.02), and (3) those who endorsed unusual experiences (vs those without unusual experiences; estimate –0.11, 95% CI –0.18 to –0.04; *P*=.001).

With respect to negative emotions, 1527 (47.9%) participants reported improvements in the intensity of negative emotions; among those who initially endorsed negative emotions, 55.6% (n=1527) experienced a reduction. Further, respondents reported a 0.62-point reduction from an average of 3.66 at baseline to 3.04 after viewing the website. There was a greater reduction in negative emotions among (1) those who identified as LGBTQI (vs non-LGBTQI; estimate –0.15, 95% CI –0.22 to –0.08; *P*<.001) and (2) those who endorsed unusual experiences (vs those without unusual experiences; estimate –0.09, 95% CI –0.17 to –0.02; *P*=.01).

XSL•FO
RenderX

Table 1. Descriptive summary of the intensity of and change in suicidal thoughts and negative emotions, and reasons endorsed for NowMattersNow.org being helpful: overall and by participant characteristics (n=3185)[a].

	Native or Indigenous (n=147)	LGBTQI (n=1034)	Alcohol or opioid problem (n=286)	Unusual experiences (n=1054)	Overall (n=3185)
Intensity of suicidal thoughts, mean (SD)					
Previsit	3.22 (1.58)	3.47 (1.43)	3.38 (1.41)	3.18 (1.50)	3.03 (1.54)
Postvisit	2.59 (1.43)	2.79 (1.40)	2.70 (1.38)	2.55 (1.41)	2.51 (1.43)
Change in suicidal thoughts, n (%)					
Improved	59 (40.1)	492 (47.6)	141 (49.3)	455 (43.2)	1211 (38)
No Change	85 (57.8)	514 (49.7)	132 (46.2)	566 (53.7)	1882 (59.1)
Worsened	3 (2)	28 (2.7)	13 (4.5)	33 (3.1)	92 (2.9)
Intensity of negative emotions, mean (SD)					
Previsit	3.86 (1.38)	3.98 (1.22)	3.97 (1.17)	3.83 (1.31)	3.66 (1.41)
Postvisit	3.08 (1.35)	3.24 (1.23)	3.25 (1.24)	3.13 (1.31)	3.04 (1.37)
Change in negative emotions, n (%)					
Improved	76 (51.7)	573 (55.4)	157 (54.9)	551 (52.3)	1527 (47.9)
No Change	65 (44.2)	417 (40.3)	102 (35.7)	443 (42)	1495 (46.9)
Worsened	6 (4.1)	44 (4.3)	27 (9.4)	60 (5.7)	163 (5.1)
Perceived helpfulness of NowMattersNow.org, n (%)					
Helpful (specific and nonspecific)	126 (85.7)	855 (82.7)	243 (85)	885 (84)	2555 (80.2)
Not helpful	10 (6.8)	82 (7.9)	19 (6.6)	443 (42)	218 (6.8)
No Response	11 (7.5)	97 (9.4)	24 (8.4)	97 (9.2)	412 (12.9)
What about NowMattersNow.org was helpful? n (%)					
I learned something[b]	32 (21.8)[c]	155 (15)[d]	49 (17.1)[e]	190 (18)[d]	668 (21)[c]
It distracted me[b]	28 (19)[e]	215 (20.8)[c]	50 (17.5)[c]	205 (19.4)[c]	544 (17.1)[e]
I felt less alone[b]	16 (10.9)[d]	161 (15.6)[e]	45 (15.7)[d]	151 (14.3)[e]	414 (13)[d]
Time passed	9 (6.1)	100 (9.7)	34 (11.9)	103 (9.8)	252 (7.9)
Other	9 (6.1)	56 (5.4)	21 (7.3)	67 (6.4)	238 (7.5)
It wasn't helpful	10 (6.8)	82 (7.9)	19 (6.6)	72 (6.8)	218 (6.8)
I called/texted a crisis line	13 (8.8)	88 (8.5)	19 (6.6)	83 (7.9)	197 (6.2)
I felt cared for	9 (6.1)	50 (4.8)	15 (5.2)	48 (4.6)	127 (4)
I saw that others could recover	10 (6.8)	30 (2.9)	10 (3.5)	38 (3.6)	115 (3.6)
No response	11 (7.5)	97 (9.4)	24 (8.4)	97 (9.2)	412 (12.9)

[a]The sample sizes across the 4 subgroups do not add up to the overall n=3185 since respondents could endorse a single, multiple, or no subgroups.

[b]Top 3 reasons.

[c]Most frequently endorsed reason.

[d]Third most frequently endorsed reason.

[e]Second most frequently endorsed reason.

Main Outcome Analyses

With respect to the perceived helpfulness of NowMattersNow.org, 80.2% (n=2555) of all respondents reported that the website was helpful. Overall, the most common frequently endorsed reason why NowMattersNow.org was helpful was that "I learned something" (n=668, 21%), followed by "It distracted me" (n=544, 17.1%), and "I felt less alone (n=414, 13%). These were also the top 3 reasons across all subgroups, although "It distracted me" and "I felt less alone" were the top 1 and 2 reasons, respectively, among those identifying as LGBTQI, those that endorsed alcohol or opioid problems, and those that endorsed unusual experiences.

Tables 2 and 3 summarize the reasons for NowmattersNow.org being helpful by (1) the intensity of baseline symptoms (suicidal thoughts or negative motions) and (2) the change in symptoms (improved, no change, worsened). Overall, 38% (n=1211) of

XSL•FO
RenderX

all respondents reported a reduction in suicidal thoughts, and 47.9% (n=1527) reported a reduction in negative emotions after visiting the website.

Among the participants with high suicidal thoughts at baseline, 54.1% (n=759) improved after viewing the website, with the top 3 reasons being “It distracted me” (n=190, 25%), “I felt less alone” (n=128, 16.9%), and “I learned something” (n=110, 14.5%). Among the participants with moderate suicidal thoughts at baseline, 49.8% (n=452) improved after viewing the website, with the top 3 reasons being “I felt less alone” (n=105, 23.2%), “I learned something” (n=100, 22.1%), and “It distracted me” (n=96, 21.2%). Among the participants with no suicidal thoughts at baseline, 39.8% (n=349) reported that they still “learned something,” and 14.7% (n=129) reported that it was helpful in “other” ways.

Table 2. The frequencies of reasons endorsed for why NowmattersNow.org was helpful overall and categorized by the intensity of BL[a] suicidal thoughts and the change in suicidal thoughts (n=3185).

	No BL suicidal thoughts (1), n (%)			Moderate BL suicidal thoughts (2-3), n (%)				High BL suicidal thoughts (4-5), n (%)				Overall, n (%)			
	No change (n=857)	Worsened (n=19)	Overall (n=876)	Improved (n=452)	No Change (n=401)	Worsened (n=54)	Overall (n=907)	Improved (n=759)	No Change (n=624)	Worsened (n=19)	Overall (n=1402)	Improved (n=1211)	No Change (n=1882)	Worsened (n=92)	Overall (n=3185)
I learned something	347 (40.5)[b]	2 (10.5)[c]	349 (39.8)[b]	100 (22.1)[d]	60 (15)[d]	7 (13)[c]	167 (20.1)[d]	110 (14.5)[c]	41 (6.6)	1 (5.3)	152 (10.8)	210 (17.3)[c]	448 (23.8)[b]	10 (10.9)[c]	668 (21)[b]
It distracted me	86 (10)	6 (31.6)[b]	92 (10.5)	96 (21.2)[c]	79 (19.7)[b]	4 (7.4)	179 (21.6)[b]	190 (25)[b]	81 (13)[c]	2 (10.5)[c]	273 (19.5)[b]	286 (23.6)[b]	246 (13.1)[c]	12 (13)[d]	544 (17.1)[d]
I felt less alone	78 (9.1)	1 (5.3)	79 (9)	105 (23.2)[b]	47 (11.7)[c]	9 (16.7)[d]	161 (19.4)[c]	128 (16.9)[d]	46 (7.4)	0 (0)	174 (12.4)[c]	233 (19.2)[d]	171 (9.1)	10 (10.9)[c]	414 (13)[c]
Time passed	30 (3.5)	3 (15.8)[d]	33 (3.8)	34 (7.5)	44 (11)	4 (7.4)	51 (6.2)	73 (9.6)	59 (9.5)	5 (26.3)[b]	136 (9.7)	107 (8.8)	133 (7.1)	12 (13)[d]	252 (7.9)
Other	129 (15.1)[d]	0 (0)	129 (14.7)[d]	13 (2.9)	29 (7.2)	2 (3.7)	31 (3.7)	23 (3)	42 (6.7)	0 (0)	65 (4.6)	36 (3)	200 (10.6)	2 (2.2)	238 (7.5)
It wasn’t helpful	23 (2.7)	0 (0)	23 (2.6)	4 (0.9)	32 (8)	12 (22.2)[b]	48 (5.8)	7 (0.9)	136 (21.8)[d]	4 (21.1)[d]	147 (10.5)	11 (0.9)	191 (10.1)	16 (17.4)[b]	218 (6.8)
I called/texted a crisis line	7 (0.8)	2 (10.5)[d]	9 (1)	21 (4.6)	20 (5)	4 (7.4)	26 (3.1)	84 (11.1)	58 (9.3)	1 (5.3)	143 (10.2)	105 (8.7)	85 (4.5)	7 (7.6)	197 (6.2)
I felt cared for	22 (2.6)	1 (5.3)	23 (2.6)	36 (8)	13 (3.2)	1 (1.9)	38 (4.6)	40 (5.3)	12 (1.9)	2 (10.5)[c]	54 (3.9)	76 (6.3)	47 (2.5)	4 (4.3)	127 (4)
I saw that others could recover	36 (4.2)	1 (5.3)	37 (4.2)	14 (3.1)	17 (4.2)	2 (3.7)	20 (2.4)	36 (4.7)	9 (1.4)	0 (0)	45 (3.2)	50 (4.1)	62 (3.3)	3 (3.3)	115 (3.6)
No response	99 (11.6)[c]	3 (15.8)[d]	102 (11.6)[c]	29 (6.4)	60 (15)[d]	9 (16.7)[d]	108 (13)	68 (9)	140 (22.4)[b]	4 (21.1)[d]	212 (15.1)[d]	97 (8)	299 (15.9)[d]	16 (17.4)[b]	412 (12.9)

[a]BL: baseline.

[b]Most frequently endorsed reason.

[c]3rd most frequently endorsed reason.

[d]2nd most frequently endorsed reason.

XSL•FO
RenderX

Table 3. The frequencies of reasons endorsed for why NowmattersNow.org was helpful overall and categorized by the intensity of BL[a] negative emotions and by the change in negative emotion (n=3185).

	No BL negative emotion (1), n (%)			Moderate BL negative emotion (2-3), n (%)				High BL negative emotion (4-5), n (%)				Overall, n (%)			
	No change (n=400)	Worsened (n=40)	Overall (n=440)	Improved (n=347)	No change (n=300)	Worsened (n=86)	Overall (N=733)	Improved (n=1180)	No change (n=795)	Worsened (n=37)	Overall (N=2012)	Improved (n=1527)	No change (n=1495)	Worsened (n=163)	Overall (n=3185)
I learned something	164 (41)[b]	10 (25)[b]	174 (39.5)[b]	119 (34.3)[b]	83 (27.7)[b]	20 (23.3)[b]	222 (30.3)[b]	198 (16.8)[c]	69 (8.7)	5 (13.5)[c]	272 (13.5)[c]	317 (20.8)[d]	316 (21.1)[b]	35 (21.5)[b]	668 (21.1)[b]
It distracted me	7 (17.5)	7 (17.5)[d]	14 (3.2)	45 (13)[d]	53 (17.7)[d]	9 (10.5)	107 (14.6)[d]	277 (23.5)[b]	115 (14.5)[c]	2 (5.4)	394 (19.6)[b]	322 (21.1)[b]	204 (13.6)[c]	18 (11)	544 (17.1)[d]
I felt less alone	25 (6.3)	4 (10)	29 (6.6)	65 (18.7)[d]	29 (9.7)	9 (10.5)	103 (14.1)[c]	227 (19.2)[d]	52 (6.5)	3 (8.1)	282 (14)	292 (19.1)[c]	106 (7.1)	16 (9.8)	414 (13)[c]
Time passed	6 (1.5)	3 (7.5)	9 (2)	14 (4)	30 (10)[c]	12 (14)[c]	56 (7.6)	102 (8.6)	81 (10.2)	4 (10.8)	187 (9.3)	116 (7.6)	117 (7.8)	19 (11.7)	252 (7.9)
Other	78 (19.5)[d]	2 (5)	80 (18.2)[d]	25 (7.2)	29 (9.7)	5 (5.8)	59 (8)	36 (3.1)	62 (7.8)	1 (2.7)	99 (4.9)	61 (4)	169 (11.3)	8 (4.9)	238 (7.5)
It wasn't helpful	7 (1.8)	3 (7.5)	10 (2.3)	3 (0.9)	12 (4)	9 (10.5)	24 (3.3)	16 (1.4)	161 (20.3)[d]	7 (18.9)[d]	184 (9.1)	19 (1.2)	180 (12)	19 (11.7)[c]	218 (6.8)
I called/texted a crisis line	2 (0.5)	3 (7.5)	5 (1.2)	10 (2.9)	7 (2.3)	3 (3.5)	20 (2.7)	99 (8.4)	68 (8.6)	5 (13.5)[c]	172 (8.5)	109 (7.1)	77 (5.2)	11 (6.7)	197 (6.2)
I felt cared for	9 (2.3)	0 (0)	9 (2.2)	26 (7.5)	12 (4)	0 (0)	38 (5.2)	69 (5.8)	11 (1.4)	0 (0)	80 (4)	95 (6.2)	32 (2.1)	0 (0)	127 (4)
I saw that others could recover	15 (3.8)	3 (7.5)	18 (4.4)	14 (4)	12 (4)	6 (7)	32 (4.4)	50 (4.2)	13 (1.6)	2 (5.4)	65 (3.2)	64 (4.2)	40 (2.7)	11 (6.7)	115 (3.6)
No response	58 (14.5)[c]	5 (12.5)[c]	63 (15.3)[c]	26 (7.5)	33 (11)	13 (15.1)[d]	72 (9.8)	106 (9)	163 (20.5)[b]	8 (21.6)[b]	277 (13.8)[d]	132 (8.6)	254 (17)[d]	26 (16)[d]	412 (12.9)

[a]BL: baseline.

[b]Most frequently endorsed reason.

[c]3rd most frequently endorsed reason.

[d]2nd most frequently endorsed reason.

Among the participants with high negative emotions at baseline, 58.6% (n=1180) improved after viewing the website, with the top 3 reasons being "It distracted me" (n=277, 23.5%), "I felt less alone" (n=227, 19.2%), and "I learned something" (n=198, 16.8%). Among the participants with moderate negative emotions at baseline, 47.3% (n=347) improved after viewing the website, with the top 3 reasons being "I learned something" (n=119, 34.3%), "I felt less alone" (n=65, 18.7%), and "It distracted me" (n=45, 13%). Among the participants with no negative emotions at baseline, 39.5% (n=174) reported that they still "learned something," and 18.2% (n=80) reported that it was helpful in "other" ways.

Tables 4 and 5 summarize participants' reasons for finding NowMattersNow.org helpful, categorized by the magnitude of change in their suicidal ideation and negative emotions. For participants who experienced the largest reduction in suicidal ideation (ie, a 4-point decrease), the most common reasons cited were "It distracted me" (n=5, 29.4%), "I felt less alone" (n=3, 17.6%), and "I felt cared for" (n=3, 17.6%), with 17.6% (n=3) of participants not responding. Among those who showed smaller reductions in suicidal ideation (ie, a 1-, 2-, or 3-point decrease), the top reasons were similar: "It distracted me," "I felt less alone," and "I learned something." For the few participants who reported the greatest increase in suicidal ideation (ie, a 4-point increase), one indicated they "called/texted

XSL•FO
RenderX

a crisis line," while the other reported that "time passed." Participants with smaller increases in suicidal ideation (ie, a 1-, 2-, or 3-point increase) provided varied responses, which are detailed in Table 4. Among participants who reported no change in the magnitude of suicidal ideation (ie, a 0-point change), 23.8% (n=448) cited "I learned something" as helpful, 15.9% (n=299) did not respond, and 13.1% (n=246) noted "It distracted me."

Similarly, for those who experienced the largest reduction in negative emotions (ie, a 4-point decrease), 23.1% (n=3) reported "It distracted me," 23.1% (n=3) "I felt less alone," and 15.4% (n=2) "I felt cared for," with 21.3% (n=2) not responding. Participants who reported smaller reductions in negative emotions (ie, a 1-, 2-, or 3-point decrease) cited similar reasons, including "It distracted me," "I felt less alone," and "I learned something." Among participants with the largest increase in negative emotions (ie, a 4-point increase), 2 reported they "called/texted a crisis line," 2 indicated "time passed," 1 mentioned "It distracted me," and 1 stated, "It wasn't helpful." For those with smaller increases (ie, a 1-, 2-, or 3-point increase), responses were varied and are detailed in Table 5. For participants with no change in negative emotions (ie, a 0-point change), 21.1% (n=316) noted "I learned something," 17% (n=254) did not respond, and 13.6% (n=204) mentioned "It distracted me."

Table 4. The frequencies of reasons endorsed for why NowMattersNow.org was helpful, categorized by the magnitude of change in suicidal thoughts (n=3185).

	Change in suicidal thoughts, n (%)								
	4-point decrease (–4)	3-point decrease (–3)	2-point decrease (–2)	1-point decrease (–1)	No Change (0)	1-point increase (+1)	2-point increase (+2)	3-point increase (+3)	4-point increase (+4)
It distracted me	5 (29.4)	13 (14.8)	81 (23.5)	187 (24.5)	246 (13.1)	9 (13)	3 (16.7)	0 (0)	0 (0)
I felt less alone	3 (17.6)	19 (21.6)	63 (18.3)	148 (19.4)	171 (9.1)	7 (10.1)	3 (16.7)	0 (0)	0 (0)
I learned something	1 (5.9)	10 (11.4)	64 (18.6)	135 (17.7)	448 (23.8)	7 (10.1)	1 (5.6)	2 (66.7)	0 (0)
I called/texted a crisis line	2 (11.8)	6 (6.8)	38 (11)	59 (7.7)	85 (4.5)	3 (4.3)	2 (11.1)	1 (33.3)	1 (50)
I saw that others could recover	0 (0)	12 (13.6)	12 (3.5)	26 (3.4)	62 (3.3)	3 (4.3)	0 (0)	0 (0)	0 (0)
I felt cared for	3 (17.6)	10 (11.4)	22 (6.4)	41 (5.4)	47 (2.5)	4 (5.8)	0 (0)	0 (0)	0 (0)
Time passed	0 (0)	5 (5.7)	30 (8.7)	72 (9.4)	133 (7.1)	9 (13)	2 (11.1)	0 (0)	1 (50)
Other	0 (0)	1 (1.1)	7 (2)	28 (3.7)	200 (10.6)	0 (0)	2 (11.1)	0 (0)	0 (0)
It wasn't helpful	0 (0)	0 (0)	2 (0.6)	9 (1.2)	191 (10.1)	15 (21.7)	1 (5.6)	0 (0)	0 (0)
No response	3 (17.6)	12 (13.6)	25 (7.3)	57 (7.5)	299 (15.9)	12 (17.4)	4 (22.2)	0 (0)	0 (0)

Table 5. The frequencies of reasons endorsed for why NowMattersNow.org was helpful, categorized by the magnitude of change in negative emotions (n=3185).

	Change in negative emotions, n (%)								
	4-point decrease (–4)	3-point decrease (–3)	2-point decrease (–2)	1-point decrease (–1)	No Change (0)	1-point increase (+1)	2-point increase (+2)	3-point increase (+3)	4-point increase (+4)
It distracted me	3 (23.1)	17 (18.9)	90 (19.5)	212 (22)	204 (13.6)	11 (9.2)	5 (14.7)	1 (25)	1 (16.7)
I felt less alone	3 (23.1)	14 (15.6)	84 (18.2)	191 (19.8)	106 (7.1)	11 (9.2)	4 (11.8)	1 (25)	0 (0)
I learned something	1 (7.7)	16 (17.8)	92 (20)	208 (21.6)	316 (21.1)	29 (24.4)	5 (14.7)	1 (25)	0 (0)
I called/texted a crisis line	1 (7.7)	6 (6.7)	40 (8.7)	62 (6.4)	77 (5.2)	8 (6.7)	1 (2.9)	0 (0)	2 (33.3)
I saw that others could recover	0 (0)	5 (5.6)	29 (6.3)	30 (3.1)	40 (2.7)	6 (5)	5 (14.7)	0 (0)	0 (0)
I felt cared for	2 (15.4)	9 (10)	37 (8)	47 (4.9)	32 (2.1)	0 (0)	0 (0)	0 (0)	0 (0)
Time passed	0 (0)	5 (5.6)	35 (7.6)	76 (7.9)	117 (7.8)	16 (13.4)	1 (2.9)	0 (0)	2 (33.3)
Other	0 (0)	0 (0)	12 (2.6)	49 (5.1)	169 (11.3)	4 (3.4)	4 (11.8)	0 (0)	0 (0)
It wasn't helpful	0 (0)	0 (0)	3 (0.7)	16 (1.7)	180 (12)	13 (10.9)	4 (11.8)	1 (25)	1 (16.7)
No response	3 (23.1)	18 (20)	39 (8.5)	72 (7.5)	254 (17)	21 (17.6)	5 (14.7)	0 (0)	0 (0)

XSL•FO
RenderX

Discussion

Principal Findings

This study builds upon our previous research supporting NowMattersNow.org for the temporary relief of suicidal ideation and negative emotions [16]. Here, we present findings from a concise user-experience questionnaire completed by visitors to NowMattersNow.org. In addition to assessing changes in suicidal thoughts and negative emotions during website use, the questionnaire also explored users' perceptions of the site's helpfulness. This paper aimed to examine the reasons behind reductions in suicidal thoughts and negative emotions reported after using the site. Our hypothesis—that participants who experienced reductions in suicidal thoughts and negative emotions would attribute this improvement primarily to "feeling less alone"—was mostly supported. Over half of the participants who reported suicidal ideation (n=1211, 52.4%) or negative emotions (n=1527, 55.6%) upon arriving at the site experienced reductions in these symptoms after engaging with NowMattersNow.org. Specifically, over half of those with high baseline suicidal thoughts experienced improvement, citing "It distracted me," "I felt less alone," and "I learned something" as the top reasons for the website's helpfulness. Similarly, almost half of the participants with moderate baseline suicidal thoughts also showed improvement; for them, "I felt less alone" was the most frequently endorsed reason, followed by "I learned something" and "It distracted me." While feeling less alone was not the most common reason cited by those with the highest levels of suicidal ideation, it was the second most helpful factor and ranked as the most helpful for those with moderate suicidal ideation. These findings align with leading theories of suicide [33,34], which propose that social disconnection is a key driver of suicidal ideation. Thus, it is expected that individuals who feel less alone after visiting NowMattersNow.org would experience a decrease in the intensity of their suicidal thoughts. The site's focus on sharing others' recovery stories likely fosters a sense of community and connection, allowing visitors to feel understood and less isolated. This sense of shared experience may be a core "active ingredient" of NowMattersNow.org, suggesting that digital mental health tools aimed at suicide prevention might benefit from emphasizing social connectedness and community support.

For reductions in negative emotions, more than half of individuals with high baseline negative emotions reported improvement, with "It distracted me," "I felt less alone," and "I learned something" as the main reasons they found the website helpful. Similarly, nearly half of individuals with moderate baseline negative emotions improved, citing "I learned something," "I felt less alone," and "It distracted me" as the top reasons for the website's helpfulness. These patterns mirror those observed among participants with high and moderate levels of suicidal ideation, supporting the idea that fostering social connectedness can reduce not only suicidal thoughts but also negative emotions more broadly. An important point is that distraction during moments of crisis—when experiencing intense suicidal ideation or negative emotions—is a key skill taught in DBT to help patients manage acute distress [25]. The NowMattersNow.org website appears well-suited to fulfill this function for its users, providing an immediate coping tool. Additionally, participants with high or moderate levels of suicidal ideation or negative emotions who reported improvement also listed "I learned something" among their top reasons for the site's helpfulness. This suggests that NowMattersNow.org is not only a valuable support tool but also an educational resource that may empower individuals with knowledge that alleviates distress. These findings highlight 3 potential active ingredients for reducing suicidal ideation and negative emotions: promoting social connectedness, offering distraction, and providing educational content. By incorporating these elements, NowMattersNow.org may be effectively addressing the core needs for individuals in distress.

The observed reductions in suicidal ideation (0.5-point reduction) and negative emotions (0.62-point reduction) in this study correspond to percentage decreases of 12.5% and 15.5%, respectively. While these changes are statistically significant, they are relatively modest, and it remains an open empirical question whether such reductions are clinically meaningful. However, the authors believe that even these modest reductions could be clinically significant, especially considering that they were achieved after only one session. Although we cannot precisely determine how much time participants spent on the website, it is reasonable to assume that most engaged with it in a single sitting. A 12.5% reduction in suicidal thoughts and a 15.5% reduction in negative emotions could represent a shift from a crisis mode to a state where individuals can more readily use coping skills or seek additional help, potentially preventing maladaptive behaviors. For context, research on session-to-session anxiety and depression symptom reductions in therapies like CBT reports decreases ranging from 4% to 8% [40,41]. Therefore, a 12.5% reduction in suicidal ideation following just one visit to the website could have meaningful implications for the user's emotional state and overall well-being.

This study also explored which reported reasons were most strongly associated with the largest improvements in suicidal ideation and negative emotions. Given the limited research on the specific mechanisms through which websites exert their effects, we did not have a prespecified hypothesis about which factors would be linked to different levels of improvement. Participants who reported the largest reductions in suicidal ideation cited "It distracted me," "I felt less alone," and "I felt cared for" as the top reasons the website was helpful. Those with moderate improvements in suicidal ideation cited similar reasons, including distraction, feeling less alone, and learning something. The top 3 reasons for reductions in negative emotions closely mirrored those for suicidal ideation, with distraction, social connection, and learning cited as key contributors to improvements. These findings suggest that fostering social connection, offering distraction, and providing educational content may not only reduce suicidal ideation and negative emotions overall but may also lead to the most significant improvements. Prioritizing these elements in digital mental health resources like NowMattersNow.org could, therefore, maximize their impact and effectiveness for individuals in distress.

XSL•FO
RenderX

This study also gathered demographic information on vulnerable populations, including LGBTQI individuals, American Indian or Alaska Native individuals, those with alcohol or substance use issues, and those reporting unusual experiences. While participants in the full sample showed an average reduction of half a point in suicidal ideation intensity and more than half a point in negative emotions, certain groups experienced even greater benefits. LGBTQI-identifying individuals reported more significant reductions in both suicidal ideation and negative emotions compared to non-LGBTQI individuals. Similarly, those with alcohol or opioid use issues showed greater reductions in suicidal ideation than those without these issues. Additionally, individuals who reported unusual experiences experienced greater decreases in both suicidal ideation and negative emotions than those who did not report such experiences. The reasons participants provided for these reductions were consistent with those reported across the full sample, often reflecting themes such as distraction ("It distracted me"), connection ("I felt less alone"), and learning ("I learned something"). These findings highlight NowMattersNow.org's potential to effectively support vulnerable groups by addressing unique challenges that conventional mental health resources may overlook. The site's content appears to resonate particularly well with individuals who identify as LGBTQI, have substance use issues, or experience unusual mental health symptoms, likely because it fosters connection, validation, and immediate coping strategies that reduce distress. This suggests that digital interventions tailored to the needs of high-risk populations can serve as accessible, impactful resources, offering a crucial support option for those who might otherwise remain isolated in their distress.

In summary, this study sheds light on the mechanisms behind the effectiveness of NowMatttersNow.org in reducing suicidal ideation and negative emotions, with 80% (n=2555) of respondents indicating that the website was helpful. Key features identified by visitors included promoting social connectedness, offering distraction, and providing educational content. The inclusion of personal stories from individuals with lived experience may improve users' sense of connection and support. This finding aligns with research suggesting that adolescents, in particular, benefit from suicide prevention narratives that feature peer experiences and coping strategies [42]. Future studies should aim to collect other demographic data, such as age, to determine how it moderates the perceived helpfulness of specific website features. This study highlights the potential for NowMattersNow.org to serve as a scalable, accessible, and immediate resource for reducing suicidal ideation and negative emotions, especially for those who face barriers to traditional mental health care, including individuals from vulnerable populations. While challenges such as technical issues [43] and the complexity of web-based interactions [44] exist—often due to many websites not being designed with users in mind [45]—digital websites like NowMattersNow.org should not be viewed as replacements for in-person care. Instead, they may help bridge the gap when face-to-face support is unavailable or augment ongoing in-person treatment. Providers recommending these resources should avoid offering them as standalone interventions and ensure they are paired with psychotherapy whenever possible. However, using these tools as interim support for patients on waitlists is appropriate. While reliance on web-based tools as the sole form of treatment is discouraged, individuals should not be deterred from exploring them independently when other care options are inaccessible or undesired. Continued research is important to better understand the needs of individuals seeking digital mental health support, as well as to improve user-friendliness, ensuring that these resources are comprehensible and accessible to individuals from all backgrounds.

Strengths and Limitations

These findings should be interpreted within the context of the study's strengths and limitations. A key strength of this study was the large sample size (n=3185), which afforded higher statistical power to evaluate overall and subgroup differences in suicidal ideation and negative emotions. A related limitation of large sample sizes is the need to temper the interpretation of the observed reductions in suicidal ideation and negative emotions, as small effects are more likely to reach statistical significance with a large sample size. Specifically, the average reductions in suicidal ideation and negative emotions were approximately a half-point on the 5-point scales, which were relatively modest. These analyses adjusted for a limited number of characteristics associated with suicide risk (eg, alcohol or opioid misuse), but we were unable to adjust for other relevant risk factors (eg, age).

This study used a single-arm pragmatic trial design, using quality improvement data from the NowMattersNow.org web-based platform, with permission to use the data for research purposes. By adopting a real-world approach, this study allowed for natural variations in how users interacted with the site, rather than imposing rigid use protocols. This is in contrast to traditional experimental designs, where standardized engagement and control groups are the norm. As a result of its pragmatic nature, certain elements commonly found in randomized controlled trials were not present in this study. For instance, there was no pre- or postintervention assessment using validated questionnaires, and the outcomes were evaluated retrospectively. Additionally, due to the real-world setting of the trial, no control group was included for comparison. These factors, while limiting the ability to draw definitive causal conclusions, are in line with the pragmatic design's emphasis on ecological validity and generalizability. The strength of this pragmatic trial lies in its real-world context: it was conducted with actual users of the NowMattersNow.org website, who interacted with the platform in their own way, reflecting typical use patterns. This design closely mirrors how digital interventions are used in practice, providing valuable insights into the website's potential impact on those at risk of suicide, as it would occur in everyday settings outside the constraints of a clinical trial. The findings, therefore, have strong implications for understanding how such interventions can be applied in the broader community, offering actionable insights for optimizing digital suicide prevention resources.

A significant limitation is that user-experience data were gathered from individual users at a single point in time, and the lack of a control group for comparison. As such, it is possible that the observed improvements were influenced by distraction

XSL•FO
RenderX

alone or merely the passage of time, which could lead to similar reductions in suicidal ideation and negative emotions. Additionally, we cannot determine whether the reported improvements were sustained beyond the brief timeframe covered by the survey. Nonresponse bias is another concern that may affect the generalizability of these findings. Individuals who viewed the survey but chose not to complete it, as well as those who never encountered the survey, may differ in their experiences and responses, particularly if they did not have positive experiences with the website. Another limitation of the study is the use of the term "unusual experiences," which was not explicitly defined for participants. As a result, the interpretation of this term was left to individual interpretation, and it is unclear how participants construed this option. As such, the exact nature of "unusual experiences" within the context of this study remains unknown. Future studies should provide a clearer definition of this term and explore participants' interpretations to gain a more precise understanding of the experiences being referred to. Further, limitations related to IP address tracking should be acknowledged. To avoid duplicate responses, records with the same IP snippet were removed, which means that participants sharing a communal network (eg, within the same facility) may have been unintentionally excluded. Conversely, individuals who accessed the website from different locations may have been assigned different IP addresses, potentially leading to multiple records from the same participant. These issues highlight challenges in accurate participant counts and representative sampling in web-based research.

The sensitive nature of the survey topic (suicide) may have influenced participation; some individuals, particularly those experiencing self-stigma or discrimination, might have been less likely to participate, while others who found the topic highly relevant may have been more inclined to engage, including those at a higher risk for suicide. Furthermore, participants rated their suicidal thoughts and negative emotions before and after using the website within a single survey. While this approach was intended to encourage participation, it raises the possibility of recall bias, which could impact the accuracy of reported reductions in distress. Although the survey demonstrates face validity, other forms of validity have not been assessed, and it remains unclear whether the respondents accurately reflect other individuals experiencing suicidal thoughts. Future randomized controlled trials are crucial to determine the causal relationships between website use and its features and reported improvements. Additionally, longitudinal studies are necessary to assess the long-term impact of the website on users' mental health, allowing for insights into the sustainability of any observed benefits. Finally, researchers should investigate whether personal narratives do, in fact, lead to improved social connectedness.

Conclusions

This study improves our understanding of the mechanisms by which NowMattersNow.org provides relief from suicidal ideation and negative emotions. With 80% (n=2555) of respondents finding the website helpful, the research identified critical features that contribute to its effectiveness, namely, through social connectedness, distraction, and educational content. The inclusion of personal narratives from individuals with lived experience appears to resonate deeply with users, fostering a sense of connection that aligns with existing literature emphasizing the importance of social support in suicide prevention. Furthermore, our findings suggest that NowMattersNow.org may be particularly helpful for vulnerable populations, such as LGBTQI individuals and those with substance use issues, highlighting the platform's role in addressing unique challenges often overlooked by conventional mental health resources. It is crucial that this resource not be introduced via asynchronous messaging from an unknown provider. When offered by a current care provider, we recommend it be reviewed together during the initial encounter and revisited in follow-up sessions to avoid a delivery style associated with increased self-harm in a previous pragmatic clinical trial [31]. As digital interventions like NowMattersNow.org continue to evolve, it is vital to conduct further research that includes diverse demographic data to optimize their design and functionality. By prioritizing user needs and improving accessibility, NowMattersNow.org can serve as an immediate resource for those seeking self-help, including those who may encounter challenges in accessing conventional mental health services, offering crucial resources when direct, in-person options are not feasible. NowMattersNow.org can be a resource for individuals facing barriers to traditional mental health care, potentially bridging the gap when in-person support is unavailable and ensuring that individuals from all backgrounds can find meaningful support.

Acknowledgments

The authors would like to gratefully acknowledge the great number of volunteer hours that resulted in this work. To conduct this research and complete this manuscript, NowMattersNow.org received a grant from Four Pines Fund, a family foundation deeply connected to suicide prevention following the loss of their son. MF's efforts on this project were partially supported by the National Institute of Mental Health (grant 1F31MH135623). The authors attest that there was no use of generative artificial intelligence technology in the generation of text, figures, or other informational content of this manuscript.

Data Availability

The datasets generated and analyzed in this study are not publicly available, as participants did not provide consent for their data to be shared beyond quality improvement efforts and research conducted by NowMattersNow.org, but are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Authors' Contributions

All authors contributed to the study's conception and design. DH cleaned the data and performed the analysis. UW supervised the project. MF wrote the first draft of the manuscript. All authors commented on and contributed to the draft. All authors read and approved the final manuscript.

Conflicts of Interest

NowMattersNow.org is a free public resource. It is also a 501c3 nonprofit in the United States. MF, DH, and UW are contract staff who worked on this manuscript. There is no conflict of interest, but there are cointerests of better access to support for people at risk for suicide. Authors reported no equity ownership, profit-sharing agreements, royalties, patents, or research or other grants from private industry or closely affiliated nonprofit funds. No authors received support from pharmaceutical companies.

References

1. Suicide prevention. Centers for Disease Control and Prevention. 2023. URL: https://www.cdc.gov/suicide/index.html [accessed 2025-02-20]
2. Suicide data and statistics. Centers for Disease Control and Prevention. 2023. URL: https://www.cdc.gov/suicide/facts/data.html?CDC_AAref_Val=https://www.cdc.gov/suicide/suicide-data-statistics.html [accessed 2025-02-20]
3. Luoma JB, Martin CE, Pearson JL. Contact with mental health and primary care providers before suicide: a review of the evidence. Am J Psychiatry 2002;159(6):909-916 [FREE Full text] [doi: 10.1176/appi.ajp.159.6.909] [Medline: 12042175]
4. Ahmedani BK, Simon GE, Stewart C, Beck A, Waitzfelder BE, Rossom R, et al. Health care contacts in the year before suicide death. J Gener Intern Med 2014;29(6):870-877 [FREE Full text] [doi: 10.1007/s11606-014-2767-3] [Medline: 24567199]
5. Hom MA, Stanley IH, Podlogar MC, Joiner TE. "Are you having thoughts of suicide?" Examining experiences with disclosing and denying suicidal ideation. J Clin Psychol 2017;73(10):1382-1392. [doi: 10.1002/jclp.22440] [Medline: 28085200]
6. Husky MM, Zablith I, Alvarez Fernandez V, Kovess-Masfety V. Factors associated with suicidal ideation disclosure: results from a large population-based study. J Affective Disord 2016;205:36-43. [doi: 10.1016/j.jad.2016.06.054] [Medline: 27400193]
7. Kleiman EM, Turner BJ, Fedor S, Beale EE, Huffman JC, Nock MK. Examination of real-time fluctuations in suicidal ideation and its risk factors: results from two ecological momentary assessment studies. J Abnorm Psychol 2017;126(6):726-738. [doi: 10.1037/abn0000273] [Medline: 28481571]
8. Hallensleben N, Spangenberg L, Forkmann T, Rath D, Hegerl U, Kersting A, et al. Investigating the dynamics of suicidal ideation. Crisis 2018;39(1):65-69. [doi: 10.1027/0227-5910/a000464] [Medline: 28468557]
9. Whiteside U, Richards JA, Fruhbauerova M, Shokouhaghaei G. Conceptual model of and intervention development for unplanned suicide attempts. Perm J 2024;28(3):117-129 [FREE Full text] [doi: 10.7812/TPP/24.035] [Medline: 39104244]
10. Simon TR, Swann AC, Powell KE, Potter LB, Kresnow M, O'Carroll PW. Characteristics of impulsive suicide attempts and attempters. Suicide Life Threat Behav 2001;32(1 Suppl):49-59. [doi: 10.1521/suli.32.1.5.49.24212] [Medline: 11924695]
11. Deisenhammer EA, Ing C, Strauss R, Kemmler G, Hinterhuber H, Weiss EM. The duration of the suicidal process. J Clin Psychiatry 2008;70(1):19-24. [doi: 10.4088/jcp.07m03904]
12. Mok K, Jorm AF, Pirkis J. Who goes online for suicide-related reasons? Crisis 2016;37(2):112-120. [doi: 10.1027/0227-5910/a000366] [Medline: 27232427]
13. Mars B, Heron J, Biddle L, Donovan JL, Holley R, Piper M, et al. Exposure to, and searching for, information about suicide and self-harm on the internet: prevalence and predictors in a population based cohort of young adults. J Affective Disord 2015;185:239-245 [FREE Full text] [doi: 10.1016/j.jad.2015.06.001] [Medline: 26150198]
14. Padmanathan P, Biddle L, Carroll R, Derges J, Potokar J, Gunnell D. Suicide and self-harm related internet use. Crisis 2018;39(6):469-478 [FREE Full text] [doi: 10.1027/0227-5910/a000522] [Medline: 29848080]
15. Biddle L, Derges J, Goldsmith C, Donovan JL, Gunnell D. Online help for people with suicidal thoughts provided by charities and healthcare organisations: a qualitative study of users' perceptions. Soc Psychiatry Psychiatr Epidemiol 2020;55(9):1157-1166 [FREE Full text] [doi: 10.1007/s00127-020-01852-6] [Medline: 32157324]
16. Whiteside U, Richards J, Huh D, Hidalgo R, Nordhauser R, Wong AJ, et al. Development and evaluation of a web-based resource for suicidal thoughts: NowMattersNow.org. J Med Internet Res 2019;21(5):e13183 [FREE Full text] [doi: 10.2196/13183] [Medline: 31045498]
17. Linardon J, Cuijpers P, Carlbring P, Messer M, Fuller-Tyszkiewicz M. The efficacy of app-supported smartphone interventions for mental health problems: a meta-analysis of randomized controlled trials. World Psychiatry 2019;18(3):325-336 [FREE Full text] [doi: 10.1002/wps.20673] [Medline: 31496095]
18. Torok M, Han J, Baker S, Werner-Seidler A, Wong I, Larsen ME, et al. Suicide prevention using self-guided digital interventions: a systematic review and meta-analysis of randomised controlled trials. Lancet Digital Health 2020;2(1):e25-e36 [FREE Full text] [doi: 10.1016/S2589-7500(19)30199-2] [Medline: 33328037]

XSL•FO
RenderX

19. Stefanopoulou E, Hogarth H, Taylor M, Russell-Haines K, Lewis D, Larkin J. Are digital interventions effective in reducing suicidal ideation and self-harm? A systematic review. J Mental Health 2020;29(2):207-216. [doi: 10.1080/09638237.2020.1714009] [Medline: 31989852]
20. Büscher R, Beisemann M, Doebler P, Micklitz HM, Kerkhof A, Cuijpers P, et al. Digital cognitive-behavioural therapy to reduce suicidal ideation and behaviours: a systematic review and meta-analysis of individual participant data. Evid Based Ment Health 2022;25(e1):e8-e17 [FREE Full text] [doi: 10.1136/ebmental-2022-300540] [Medline: 36535686]
21. Witt K, Spittal MJ, Carter G, Pirkis J, Hetrick S, Currier D, et al. Effectiveness of online and mobile telephone applications ('apps') for the self-management of suicidal ideation and self-harm: a systematic review and meta-analysis. BMC Psychiatry 2017;17(1):297 [FREE Full text] [doi: 10.1186/s12888-017-1458-0] [Medline: 28810841]
22. Arshad U, Farhat-Ul-Ain, Gauntlett J, Husain N, Chaudhry N, Taylor PJ. A systematic review of the evidence supporting mobile- and internet-based psychological interventions for self-harm. Suicide Life Threat Behav 2020;50(1):151-179 [FREE Full text] [doi: 10.1111/sltb.12583] [Medline: 31448847]
23. Melia R, Francis K, Hickey E, Bogue J, Duggan J, O'Sullivan M, et al. Mobile health technology interventions for suicide prevention: systematic review. JMIR mHealth uHealth 2020;8(1):e12516 [FREE Full text] [doi: 10.2196/12516] [Medline: 31939744]
24. DeCou CR, Comtois KA, Landes SJ. Dialectical behavior therapy is effective for the treatment of suicidal behavior: a meta-analysis. Behav Ther 2019;50(1):60-72. [doi: 10.1016/j.beth.2018.03.009] [Medline: 30661567]
25. Linehan M. Cognitive-Behavioral Treatment of Borderline Personality Disorder. New York: Guilford Press; 1993.
26. Rodante DE, Kaplan MI, Olivera Fedi R, Gagliesi P, Pascali A, Quintero PSJ, et al. CALMA, a mobile health application, as an accessory to therapy for reduction of suicidal and non-suicidal self-injured behaviors: a pilot cluster randomized controlled trial. Arch Suicide Res 2022;26(2):801-818. [doi: 10.1080/13811118.2020.1834476] [Medline: 33073745]
27. Wilks CR, Lungu A, Ang SY, Matsumiya B, Yin Q, Linehan MM. A randomized controlled trial of an internet delivered dialectical behavior therapy skills training for suicidal and heavy episodic drinkers. J Affective Disord 2018;232:219-228 [FREE Full text] [doi: 10.1016/j.jad.2018.02.053] [Medline: 29499504]
28. Rizvi SL, Hughes CD, Thomas MC. The DBT coach mobile application as an adjunct to treatment for suicidal and self-injuring individuals with borderline personality disorder: a preliminary evaluation and challenges to client utilization. Psychol Serv 2016;13(4):380-388. [doi: 10.1037/ser0000100] [Medline: 27797571]
29. Lakeman R, King P, Hurley J, Tranter R, Leggett A, Campbell K, et al. Towards online delivery of dialectical behaviour therapy: a scoping review. Int J Ment Health Nurs 2022;31(4):843-856 [FREE Full text] [doi: 10.1111/inm.12976] [Medline: 35048482]
30. Austin SF, Jansen JE, Petersen CJ, Jensen R, Simonsen E. Mobile app integration into dialectical behavior therapy for persons with borderline personality disorder: qualitative and quantitative study. JMIR Ment Health 2020;7(6):e14913 [FREE Full text] [doi: 10.2196/14913] [Medline: 32525488]
31. Simon GE, Shortreed SM, Rossom RC, Beck A, Clarke GN, Whiteside U, et al. Effect of offering care management or online dialectical behavior therapy skills training vs usual care on self-harm among adult outpatients with suicidal ideation: a randomized clinical trial. J Am Med Assoc 2022;327(7):630-638 [FREE Full text] [doi: 10.1001/jama.2022.0423] [Medline: 35166800]
32. Simon GE, Shortreed SM, Rossom RC, Beck A, Clarke GN, Whiteside U, et al. Effect of offering care management or online dialectical behavior therapy skills training vs usual care on self-harm among adult outpatients with suicidal ideation: a randomized clinical trial. J Am Med Assoc 2022;327(7):630-638 [FREE Full text] [doi: 10.1001/jama.2022.0423] [Medline: 35166800]
33. Van Orden KA, Witte TK, Cukrowicz KC, Braithwaite SR, Selby EA, Joiner TE. The interpersonal theory of suicide. Psychol Rev 2010;117(2):575-600 [FREE Full text] [doi: 10.1037/a0018697] [Medline: 20438238]
34. Klonsky ED, May AM. The three-step theory (3ST): a new theory of suicide rooted in the “Ideation-to-Action” framework. Int J Cogn Ther 2015;8(2):114-129. [doi: 10.1521/ijct.2015.8.2.114]
35. Motto JA. Suicide prevention for high-risk persons who refuse treatment. Suicide Life Threat Behav 2010;6(4):223-230. [doi: 10.1111/j.1943-278x.1976.tb00880.x]
36. Motto JA, Bostrom AG. A randomized controlled trial of postcrisis suicide prevention. Psychiatr Serv 2001;52(6):828-833. [doi: 10.1176/appi.ps.52.6.828] [Medline: 11376235]
37. Fleischmann A, Bertolote JM, Wasserman D, De Leo D, Bolhari J, Botega NJ, et al. Effectiveness of brief intervention and contact for suicide attempters: a randomized controlled trial in five countries. Bull World Health Org 2008;86(9):703-709 [FREE Full text] [doi: 10.2471/blt.07.046995] [Medline: 18797646]
38. Milner A, Spittal MJ, Kapur N, Witt K, Pirkis J, Carter G. Mechanisms of brief contact interventions in clinical populations: a systematic review. BMC Psychiatry 2016;16(1):194 [FREE Full text] [doi: 10.1186/s12888-016-0896-4] [Medline: 27277833]
39. Ellsworth M. Civilization. URL: https://builtbycivilization.com [accessed 2025-11-20]
40. Fruhbauerova M, Terrill DR, Semcho SA, Stumpp NE, McCann JP, Sauer-Zavala S, et al. Skill use mediates the within-person effect of the alliance on session-to-session changes in anxiety and depression in the unified protocol. J Mood Anxiety Disord 2024;5:100043 [FREE Full text] [doi: 10.1016/j.xjmad.2023.100043] [Medline: 38523702]

XSL•FO
RenderX

41. Southward MW, Sauer-Zavala S. Dimensions of skill use in the unified protocol: exploring unique effects on anxiety and depression. J Consult Clin Psychol 2022;90(3):246-257. [doi: 10.1037/ccp0000701] [Medline: 34914417]
42. Braun M, Till B, Pirkis J, Niederkrotenthaler T. Effects of suicide prevention videos developed by and targeting adolescents: a randomized controlled trial. Eur Child Adolesc Psychiatry 2023;32(5):847-857 [FREE Full text] [doi: 10.1007/s00787-021-01911-6] [Medline: 34817663]
43. Wilks CR, Yin Q, Zuromski KL. User experience affects dropout from internet-delivered dialectical behavior therapy. Telemed J E-health 2020;26(6):794-797 [FREE Full text] [doi: 10.1089/tmj.2019.0124] [Medline: 31502945]
44. Haime Z, Kennedy L, Grace L, Cohen R, Derges J, Biddle L. The journey of engaging with web-based self-harm and suicide content: longitudinal qualitative study. JMIR Infodemiol 2024;4:e47699 [FREE Full text] [doi: 10.2196/47699] [Medline: 38546718]
45. Torous J, Larsen ME, Depp C, Cosco TD, Barnett I, Nock MK, et al. Smartphones, sensors, and machine learning to advance real-time prediction and interventions for suicide prevention: a review of current progress and next steps. Curr Psychiatry Rep 2018;20(7):51-51. [doi: 10.1007/s11920-018-0914-y] [Medline: 29956120]

Abbreviations

CBT: cognitive behavioral therapy
DBT: dialectical behavior therapy
STB: suicidal thought and behavior

Edited by A Mavragani; submitted 16.01.25; peer-reviewed by CY Chen, C Br; comments to author 25.02.25; revised version received 28.04.25; accepted 19.08.25; published 25.09.25.

Please cite as:
Fruhbauerova M, Huh D, Whiteside U
User-Reported Mechanisms of Change on a Suicide Prevention Website: Single-Arm Pragmatic Trial
JMIR Form Res 2025;9:e71136
URL: https://formative.jmir.org/2025/1/e71136
doi:10.2196/71136
PMID:

©Martina Fruhbauerova, David Huh, Ursula Whiteside. Originally published in JMIR Formative Research (https://formative.jmir.org), 25.09.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

eHealth Applications Improve Glycemic Control in Patients With Diabetes: Randomized Controlled Trial

Junjie Huang[1,2], PhD; Claire Chenwen Zhong[1,2], PhD; Siu Hin Wong[1,2], BSc; Chung Yi Lo[1,2], MA; Man Kin Yim[1,2], MMedSc; Martin C S Wong[1,2,3,4,5], MD

[1]The Jockey Club School of Public Health and Primary Care, Faculty of Medicine, Chinese University of Hong Kong, Hong Kong, China (Hong Kong)
[2]Centre for Health Education and Health Promotion, Faculty of Medicine, The Chinese University of Hong Kong, Hong Kong, China (Hong Kong)
[3]The School of Public Health, Peking University, Beijing, China
[4]The School of Public Health, The Chinese Academy of Medical Sciences and The Peking Union Medical Colleges, Beijing, China
[5]The School of Public Health, Fudan University, Shanghai, China

Corresponding Author:
Martin C S Wong, MD
The Jockey Club School of Public Health and Primary Care
Faculty of Medicine
Chinese University of Hong Kong
Room 407, 4/F, Postgraduate Education Centre, Prince of Wales Hospital, 30-32 Ngan Shing Street, Shatin, N.T.
Hong Kong
China (Hong Kong)
Phone: 852 2606 3500
Email: wong_martin@cuhk.edu.hk

Abstract

Background: The eHealth app introduced a chronic disease management module to enhance the efficiency of clinical consultations and facilitate self-health management. However, there was limited information on health outcomes after using the module.

Objective: This study aimed to examine the health outcomes of individuals who used the health management module of an individualized electronic application compared to those who did not.

Methods: A randomized controlled trial was conducted with 165 participants, with 82 participants assigned to the control group and 83 participants assigned to the intervention group. Randomization was done via a computer randomizer to evaluate the impact of the eHealth chronic disease management module installation on clinical outcomes such as blood pressure, hemoglobin A1c (HbA1c), renal function tests, estimated glomerular filtration rate, and urine albumin/creatinine ratio. Data were collected at baseline and at follow-up visits at 4 and 8 months. Student *t* tests and chi-square tests were performed to analyze the difference between the intervention and control groups and examined the potential impact of the use of the eHealth chronic disease management module on various health outcomes.

Results: In total, 161 participants were included in the analysis, with an average age of 66.58 (SD 9.75) and 66.49 (SD 8.45) years in control and intervention group respectively. After 4 months, the intervention group showed better glycemic control, with significantly lower mean HbA1c levels (mean 6.76%, SD 0.64%) compared to the control group (mean 7.09%, SD 0.82%, P=.007). Also, more participants in intervention group achieved optimal HbA1c levels (n=58, 73.4%; P=.004) compared to the control group (n=36, 49.3%) in month 4. App usage had significantly decreased when comparing the usage after 4 months (mean 1.88 points, SD 0.81 points) and month 8 (mean 1.39 points, SD 0.72 points; P<.001). The results indicated better glycemic control for participants using the module in a relatively shorter period of time, and app adherence was the key for the continuous optimal glycemic control.

Conclusions: These findings support the potential of the module for clinical application in patients with suboptimal glycemic control. The long-term benefit of the module may be affected by the compliance of participants to the module.

Trial Registration: Chinese Clinical Trial Register ChiCTR2500108895; https://www.chictr.org.cn/showprojEN.html?proj=214865

(JMIR Form Res 2025;9:e67761) doi:10.2196/67761

XSL•FO
RenderX

KEYWORDS

randomized controlled trial; hemoglobin A1c; eHealth chronic disease management module; glycemic control; diabetes

Introduction

The primary goal of eHealth is to leverage information and communication technology to improve health care outcomes [1]. This includes promoting healthy behaviors, supporting self-management of chronic diseases, and facilitating communication between patients and health care providers [2,3]. The Food and Drug Administration (FDA) defined a mobile medical app as an app used as an accessory to a regulated medical device [4]. The FDA released guidance to regulate mobile medical apps and ensured the safety of these apps for the user [5]. By using eHealth apps, patients can access crucial information, monitor their physical and mental well-being, and make informed health care decisions [6-10]. Some studies have demonstrated the positive impact of eHealth apps on specific conditions, such as diabetes self-management [11]. The use of mobile apps is associated with improved attitudes toward disease self-management, as well as significant improvements in hemoglobin A1c (HbA1c) levels, diabetes knowledge, and self-care behaviors [11,12].

In Hong Kong, the Food and Health Bureau and the Hospital Authority have developed an electronic health record sharing system (eHRSS) and an eHealth app. The eHRSS enables the sharing of electronic health records between public and private health care providers [13]. The number of beneficiaries from this computerized platform was numerous, including private hospitals and clinics, the Department of Health, primary care providers and clinicians, nursing colleagues, allied health professionals, and pathologists [14]. The health management module of eHealth was subsequently launched in July 2021 to enhance patient engagement and self-monitoring. This module enables users to record and monitor vital health indicators such as blood pressure, blood glycemic level, and heart rate, and provides reminders for regular measurements [15]. Additionally, patients can download and share their data or compile personal health reports for reference by health professionals and their family members. Hence, this may enhance the efficiency of clinical consultations and facilitate health management by patients and family members.

Given the growing role of eHealth in chronic disease management, this study seeks to determine whether the use of this module enhances patient engagement and self-monitoring, leading to better clinical outcomes compared to a control group without the app. Specifically, we conducted a randomized controlled trial to evaluate the effectiveness of the eHealth chronic disease management module in improving key health outcomes, including BMI, blood pressure, HbA1c, and urine albumin/creatinine ratio in patients with diabetes in a general outpatient clinic over a 1-year period.

Methods

Ethical Considerations

This study was approved by The Joint Chinese University of Hong Kong – New Territories East Cluster Clinical Research Ethics Committee (2022.264), Hong Kong SAR. All participants were informed of the voluntary and anonymous nature of the study. Written consent was obtained from all participants. Participants were given a document to explain the study's purpose and procedures. All data collected were anonymized to ensure that there was no personally identifiable information, encrypted, and stored electronically.

Study Design

This randomized controlled trial was conducted over a period of 1 year in 2024, involving 165 participants, with a follow-up time of 8 months. A total of 83 participants were assigned to the intervention group and 82 participants to the control group using a computer-generated randomization process. Patients were categorized into 2 groups: those who had adopted the health management module and those who had not. The intervention group received the eHealth app, which was installed on their phones with detailed instructions from our team members at baseline. Reminders were sent weekly to ensure the use of the app among participants in the intervention group. During the designated reporting period, 3 to 6 phone calls were conducted to remind participants for blood tests. The enrollment process adhered to the CONSORT (Consolidated Standards of Reporting Trials) guidelines [16].

Participants

This trial was performed in a designated general outpatient clinic with a high volume of patients with diabetes attending for follow-up. Patients' eHRSS registration status was documented in the clinical management system. Eligible participants included (1) adult patients aged 18 years or older, (2) patients with physician-diagnosed diabetes who had undergone diabetes mellitus complication screening before their installation of the eHRSS health management module, (3) regular visitors of the study clinic, and (4) patients diagnosed with type 2 diabetes. Exclusion criteria were the presence of medical conditions that impaired their ability to understand and participate in the study and previous use of the module.

Procedures

Eligible participants were recruited in the clinic and underwent a subsequent diabetes mellitus complication screening to assess the potential complications of diabetes. Social demographic information was collected, including information on the patients' lifestyles, such as dietary habits, physical activity, cigarette smoking, and alcohol consumption. Their clinical parameters were also assessed, including BMI, blood pressure levels, HbA1c levels, urinalysis (albumin/creatinine ratio), and the presence of diabetes complications.

XSL•FO
RenderX

Outcomes

Data were collected at baseline and at follow-up visits at 4 and 8 months. Recruited participants underwent assessments for various diabetes-related health parameters, including blood pressure, HbA1c, renal function tests, and urine albumin/creatinine ratio. Additionally, lifestyle data including dietary consumption, physical activity, cigarette smoking, and alcohol consumption were collected. Dietary consumption and physical activity were measured using a 4-point Likert scale from 1 (never) to 4 (every time), with higher scores representing healthier lifestyles. Information on diabetes-related complications (macrovascular complications such as coronary artery disease, peripheral arterial disease, and stroke and microvascular complications such as diabetic neuropathic, nephropathy, and retinopathy) and prescribed medications was also collected. App usage at 4 and 8 months were measured using a 6-point Likert scale from 0 (never) to 5 (every time), with higher scores representing higher levels of use.

Randomization

Block randomization was adopted in this study to ensure balanced group sizes throughout the trial using blocks of size 4. This approach allowed random assignment of participants within each block, helping to maintain equilibrium between treatment and control groups for potential confounding variables.

Statistical Analysis

Baseline characteristics were compared between the intervention and control groups. Student *t* tests for continuous variables and chi-square tests for categorical variables were performed. Each clinical parameter was analyzed between the intervention and control group to examine the potential impact of the use of the eHealth chronic disease management module on various health outcomes. All analyses were performed using SPSS version 26.0 (IBM Corporation). *P* values <.05 were considered statistically significant.

Results

Characteristics of Participants

A total of 165 patients were recruited, with 161 completing at least 1 follow-up (Figure 1). During the study, 4 participants were lost to follow-up, resulting in a default rate of 2.4%. In the end, 79 participants in control group and 82 participants in intervention group were included in the analysis. The average age of the participants was 66.58 (SD 9.75) years in the control group and 66.49 (SD 8.45) years in the intervention group, with no difference between groups. Details of patient sociodemographics are shown in Table 1.

At the 4-month follow-up visit, the intervention group exhibited better diabetes control, as indicated by a significant difference in HbA1c levels. The intervention group had a mean HbA1c of 6.76% (SD 0.64%), compared to 7.09% (SD 0.82%) in the control group (*P*=.007). However, for other outcome measures, such as renal function tests, estimated glomerular filtration rate, and urine albumin/creatinine ratio, no significant differences were observed between the 2 groups. Additionally, subsequent follow-ups did not reveal any significant differences (Table 2).

Figure 1. CONSORT (Consolidated Standards Of Reporting Trials) flow diagram.

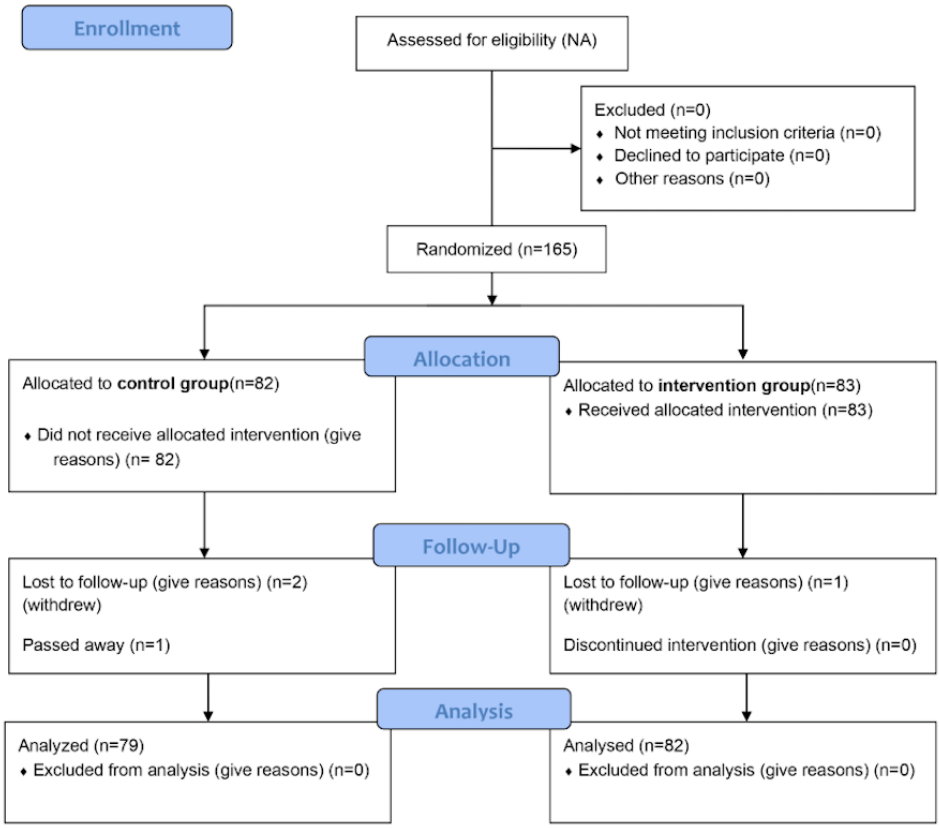

XSL•FO
RenderX

Table 1. Participant characteristics.

Demographics	Control	Intervention
Age (years), mean (SD)	66.58 (9.75)	66.49 (8.45)
Sex, n (%)		
Male	37 (46)	36 (44)
Female	43 (54)	46 (56)
Education level, n (%)		
Basic	32 (45)	28 (40)
Middle	35 (49)	36 (51)
High	4 (6)	6 (9)
Employment status, n (%)		
Employed	21 (33)	26 (40)
Unemployed	32 (51)	35 (54)
Housekeeper	10 (16)	3 (5)
Others	0 (0)	1 (2)
Government subsidy, n (%)		
Yes	35 (49)	33 (48)
No	36 (51)	36 (52)
Monthly household income (HKD)[a], n (%)		
<15,000	14 (74)	7 (50)
15,000-40,000	4 (21)	7 (50)
>40,000	1 (5)	0 (0)
Dietary consumption[b], mean (SD)	2.46 (0.55)	2.46 (0.55)
Physical activity[b], mean (SD)	2.43 (1.15)	2.65 (1.03)
Cigarette smoking status, n (%)		
Current smoker	7 (10)	5 (7)
Previous smoker	14 (20)	8 (12)
Never	49 (70)	56 (81)
Alcohol consumption, n (%)		
Current consumer	23 (33)	14 (20)
Previous consumer	4 (6)	4 (6)
Never	42 (61)	52 (74)
Chronic disease status, n (%)		
No chronic disease	1 (1)	6 (9)
One chronic disease	25 (35)	31 (45)
Multimorbidity	45 (63)	32 (46)
BMI (kg/m^2), mean (SD)	25.15 (4.33)	24.93 (4.88)
Height (cm), mean (SD)	161.67 (7.98)	160.24 (19.9)
Weight (kg), mean (SD)	65.61 (12.37)	66.32 (13.3)

[a]A currency exchange rate of HKD 1=US $7.80 is applicable.

[b]Measured on a 4-point Likert scale.

XSL•FO
RenderX

Table 2. Student's t tests analysis on the continuous variables between control and intervention groups

Continuous variables	Control, mean (SD)	Intervention, mean (SD)	*P* value	*t* value	Cohen *d*
SBP[a]	130.17 (18.11)	128.13 (13.54)	.79	0.43	0.128
DBP[b]	76.03 (10.14)	74.76 (9.05)	.81	0.417	0.132
SBP (month 4)	129.81 (14.09)	127.46 (10.82)	.25	1.148	0.187
SBP (month 8)	128.11 (14.03)	127.96 (13.09)	.95	0.067	0.011
DBP (month 4)	72.89 (11.32)	73.65 (9.28)	.66	–0.448	–0.073
DBP (month 8)	71.42 (9.05)	72.45 (9.47)	.5	–0.682	–0.111
HbA1c[c] (month 4)	7.09 (0.82)	6.76 (0.64)	.007	2.722	0.444
HbA1c (month 8)	6.83 (0.98)	6.89 (0.87)	.72	–0.365	–0.061
Sodium (mmol/L) (month 4)	139.54 (2.44)	139.46 (2.64)	.84	0.203	0.032
Sodium (mmol/L) (month 8)	139.71 (2.27)	138.96 (2.78)	.07	1.822	0.297
Potassium (mmol/L) (month 4)	4.20 (0.40)	4.29 (0.44)	.16	–1.401	–0.223
Potassium (mmol/L) (month 8)	4.17 (0.37)	4.31 (0.49)	.06	–1.904	–0.311
Urea (mmol/L) (month 4)	6.22 (2.06)	6.05 (1.64)	.56	0.578	0.088
Urea (mmol/L) (month 8)	6.65 (1.84)	6.14 (1.84)	.18	1.336	0.217
Creatinine (month 4) (μmol/L)	81.71 (26.97)	78.69 (22.97)	.45	0.761	0.121
Creatinine (month 8) (μmol/L)	86.75 (25.68)	82.87 (22.60)	.32	0.992	0.161
Urine albumin (month 4) (mg/d)	6.83 (15.99)	6.70 (20.29)	.97	0.037	0.007
Urine albumin (month 8) (mg/d)	8.10 (17.07)	7.07 (18.56)	.77	0.29	0.058

[a]SBP: systolic blood pressure (kPa).

[b]DBP: diastolic blood pressure (kPa).

[c]HbA1c: hemoglobin A1c (%).

In terms of the proportion of participants who achieved optimal glycemic control, the results indicated a significant difference in HbA1c levels at the 4-month follow-up between the intervention and control groups. Specifically, 75.95% (60/83) of the intervention group achieved an optimal result, compared to 58.75% (47/82) in the control group (P=.02). No significant differences were found for the other outcome parameters, such as estimated glomerular filtration rate and urine albumin/creatinine ratio, and no further differences were observed in the subsequent follow-up (Table 3).

App usage also declined over time, with the intervention group showing a mean usage of 1.88 points (SD 0.81 points) at the 4-month follow-up compared to 1.39 points (SD 0.72 points) at the 8-month follow-up (P<.001), suggesting a decrease in engagement over time (Table 4).

XSL•FO
RenderX

Table 3. Chi-square tests on the categorical variables between control and intervention groups.

Variable		Control group, n (%)	Intervention group, n (%)	Total, n (%)	*P* value	Effect size (Cramér V)
Sex (baseline)					.76	0.069
	Female	43 (54)	46 (56)	89 (55)		
	Male	37 (46)	36 (44)	73 (45)		
HbA1c[a] (baseline)					.86	0.016
	Optimal	41 (62)	43 (65)	84 (64)		
	Abnormal	25 (38)	23 (35)	48 (36)		
HbA1c (month 4)					0.004	0.234
	Optimal	36 (49)	58 (73)	94 (62)		
	Abnormal	37 (51)	21 (27)	58 (38)		
HbA1c (month 8)					.41	0.069
	Optimal	43 (58)	47 (66)	90 (62)		
	Abnormal	31 (42)	24 (34)	55 (38)		
SBP[b] (month 4)					.15	0.118
	Optimal	54 (74)	67 (85)	121 (80)		
	Abnormal	19 (26)	12 (15)	31 (20)		
DBP[c] (month 4)					>.99[d]	0
	Optimal	68 (93)	74 (94)	142 (93)		
	Abnormal	5 (7)	5 (6)	10 (7)		
SBP (month 8)					.66	0.036
	Optimal	59 (82)	67 (86)	126 (84)		
	Abnormal	13 (18)	11 (14)	24 (16)		
DBP (month 8)					>.99[d]	0
	Optimal	70 (97)	75 (96)	145 (97)		
	Abnormal	2 (3)	3 (4)	5 (3)		
Creatinine[e] (month 4)					>.99	0
	Optimal	63 (81)	64 (82)	127 (81)		
	Abnormal	15 (19)	14 (18)	29 (19)		
Creatinine (month 8)					.51	0.054
	Optimal	58 (74)	57 (80)	115 (77)		
	Abnormal	20 (26)	14 (20)	34 (23)		
eGFR[f] (month 4)					.34	0.08
	Optimal	54 (81)	65 (88)	119 (84)		
	Abnormal	13 (19)	9 (12)	22 (16)		
eGFR (month 8)					.37	0.077
	Optimal	48 (71)	52 (79)	100 (75)		
	Abnormal	20 (29)	14 (21)	34 (25)		
Sodium[g] (month 4)					.52	0.051
	Optimal	76 (95)	72 (9)	148 (93)		
	Abnormal	4 (5)	7 (9)	11 (7)		
Sodium (month 8)					.25	0.094

XSL•FO
RenderX

Variable		Control group, n (%)	Intervention group, n (%)	Total, n (%)	*P* value	Effect size (Cramér V)
	Optimal	77 (96)	65 (90)	142 (93)		
	Abnormal	3 (4)	7 (10)	10 (7)		
Potassium[h] (month 4)					.97	0.003
	Optimal	75 (94)	72 (92)	147 (93)		
	Abnormal	5 (6)	6 (8)	11 (7)		
Potassium (month 8)					>.99	0
	Optimal	73 (91)	66 (92)	139 (91)		
	Abnormal	7 (9)	6 (8)	13 (9)		
Urea[i] (month 4)					.76	0.024
	Optimal	57 (71)	59 (75)	116 (73)		
	Abnormal	23 (29)	20 (25)	43 (27)		
Urea (month 8)					.44	0.062
	Optimal	58 (73)	57 (79)	115 (76)		
	Abnormal	22 (28)	15 (21)	37 (24)		
Urine albumin[j] (month 4)					.93	0.009
	Optimal	46 (94)	51 (96)	97 (95)		
	Abnormal	3 (6)	2 (4)	5 (5)		
Urine albumin (month 8)					.53	0.063
	Optimal	42 (91)	52 (96)	94 (94)		
	Abnormal	4 (9)	2 (4)	6 (6)		

[a]HbA1c: hemoglobin A1c; optimal level <7%.
[b]SBP: systolic blood pressure; optimal level <140 kPa.
[c]DBP: diastolic blood pressure; optimal level <90 kPa.
[d]Outcome of Fisher test.
[e]Optimal creatinine levels: 52.1-91.9 µmol/L for women and 65.4-119.3 µmol/L for men.
[f]eGFR: estimated glomerular filtration rate; optimal level ≥60 mL/min/1.73 m^2.
[g]Optimal sodium level: 136-144 mmol/L.
[h]Optimal potassium level: 3.7-5.1 mmol/L.
[i]Optimal urea level: 2.1-7.1 mmol/L.
[j]Optimal urine albumin level: <30 mg/d.

Table 4. Student t test on app usage between month 4 and month 8 in the intervention group. App usage at 4 and 8 months was measured using a 6-point Likert scale from 0 (never) to 5 (every time), with a higher score representing a higher level of use.

Usage	Month 4	Month 8	*P* value
Intervention usage, mean (SD)	1.88 (0.81)	1.39 (0.72)	<.001

Discussion

Brief Summary

This study found a significant improvement in HbA1c levels among participants in the intervention group at the 4-month follow-up and no significant difference between participants in the intervention group and control group at the 8-month follow-up. This finding suggests that optimal glycemic control could be achieved for short period of time with app use. However, the compliance of users with the module could be a barrier to the long-term success of app use on glycemic control and other health outcomes.

Interpretations

This study showed short-term improvement of glycemic control, aligning with some previous studies. Tobacman et al [17] conducted a prospective cohort study and revealed a significant improvement in care service for the intervention group that had documentation of their own health records. Furthermore, better clinical outcomes were also observed from another randomized controlled trial study conducted by Bourgeois et al [18].

XSL•FO
RenderX

However, the phenomenon was not observed in month 8; this drop in the effect of the module could explain the decrease in medication adherence. Arshed et al [19] studied 23 studies on cardiovascular disease between 2000 and 2021 and revealed 26% of studies showed that mobile health interventions failed to show continuous impact on medication adherence. Pouls et al [20] conducted a systematic review and revealed that 17 out of the 29 randomized controlled trials included showed positive effects in medication adherence through use of SMS or interactive voice response, mobile apps, and phone calls as feedback, strategies to teaching, and facilitating communication or decision-making. The future of the management module could focus on developing features that facilitate increased adherence by patients, eventually improving clinical outcomes.

Comparisons to Existing Literature

These findings were compatible with results from previous studies [21]. Many studies have pointed out that the use of electronic health programs can help patients improve their health behaviors and achieve better disease control, which may be attributed to the ability of the eHealth chronic disease management module to enhance self-control and self-efficacy, such as illness recognition and the stimulation of self-reflection through continuous monitoring of health data [22].

However, the lack of significant improvements in other clinical parameters and the diminished effect at the 8-month follow-up point to challenges in sustaining long-term engagement with the app. Our study also found that user engagement with the app declines over time. Previous studies also pointed out that users' willingness to use the eHealth management module is often highest in the beginning stage, and the attrition rate may increase over time, potentially due to waning motivation, declining interest, and nonusage attrition [23,24]. This pattern may explain why the initial benefits observed at 4 months were not sustained at 8 months. Therefore, the sustainability of the apps' long-term impact and continuous use need to be taken into consideration [25].

Implications

Users of the eHealth app, particularly the eHealth management module, demonstrated better blood glucose control compared to nonusers. Older users showed better blood glucose control, despite the majority of app users being younger. Numerous factors, such as sociodemographic characteristics, medical conditions, duration of diabetes diagnosis, and medication, could have influenced the overall HbA1c outcomes. This study revealed no significant difference between the 2 groups in term of BMI, blood pressure, HbA1c, and urine albumin/creatinine ratio in patients with diabetes over a 1-year period. Developers of eHealth apps should focus on increasing the adherence of users to eHealth app usage and improving app features that may help patients to manage their weight, blood pressure, and chronic disease such as diabetes.

Limitation

This study demonstrated the positive health outcomes in glycemic control achieved by using the eHealth chronic disease management module. However, this study has some limitations. First, despite reminders being sent routinely to participants in intervention group, participants may not have fully adhered to the modules, leading to biased results. Second, the statistically significant difference in HbA1c levels between the intervention and control groups may not represent a clinical difference. Further follow-up and investigation in real clinical settings are needed to confirm the finding. Third, based on the outcomes of this study, longer follow-up periods—such as 12 or 24 months—are suggested in order to generate a comprehensive result. Fourth, individuals assessed for eligibility were not included in the study, as recruitment occurred in the clinic without maintaining formal records of those assessed during the eligibility screening process.

Conclusions

This randomized controlled trial demonstrated the potential short-term benefits of an eHealth chronic disease management module for improving diabetes control. However, the decline in app usage and lack of sustained improvement in other clinical parameters at the 8-month follow-up highlight the need for strategies to maintain user engagement over time. This study identified positive clinical outcomes from the module and the need for improvement in the compliance of patients. Future iterations of the eHealth program should incorporate features that enhance long-term user motivation, such as regular feedback, ongoing support, and educational interventions, particularly targeting older adults or those with limited technology proficiency. Ongoing evaluation and adaptation of the eHealth module will be critical to maximize its long-term impact on diabetes management.

Acknowledgments

This study was funded by the Hospital Authority of Hong Kong SAR (grant ITQ21-002).

Data Availability

The datasets used or analyzed during the current study are available from the corresponding author on reasonable request.

Authors' Contributions

MCSW, CCZ, and JH conceptualized and supervised the study. JH, CCZ, SHW, and CYL were responsible for data curation and formal analysis. JH, CCZ, SHW, CYL, and MKY drafted the manuscript. MCSW reviewed and revised the manuscript. All authors read and approved the final manuscript. MCSW and CCZ contributed equally as co-corresponding authors for this work.

XSL•FO
RenderX

Conflicts of Interest

None declared.

References

1. Wentzel M, Beerlage-de JN, Sieverink F. Co-Creatie eHealthboek: eHealth, Technisch Kunstje of Pure Veranderkunde?. 2014. URL: https://www.zorginnovatieboek.nl/wp-content/uploads/Cocreatie_eHealthboek.pdf [accessed 2025-09-03]
2. Vandelanotte C, Müller AM, Short CE, Hingle M, Nathan N, Williams SL, et al. Past, present, and future of eHealth and mHealth research to improve physical activity and dietary behaviors. J Nutr Educ Behav 2016 Mar;48(3):219-228.e1. [doi: 10.1016/j.jneb.2015.12.006] [Medline: 26965100]
3. Murray E. eHealth: where next? Br J Gen Pract 2014 Jul;64(624):325-326 [FREE Full text] [doi: 10.3399/bjgp14X680365] [Medline: 24982470]
4. Policy for device software functions and mobile medical applications. US Food and Drug Administration. 2019. URL: https://www.fda.gov/media/80958/download [accessed 2025-08-29]
5. Policy for Device Software Functions and Mobile Medical Applications. US Food and Drug Administration. 2022. URL: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/policy-device-software-functions-and-mobile-medical-applications [accessed 2025-08-29]
6. Free C, Phillips G, Watson L, Galli L, Felix L, Edwards P, et al. The effectiveness of mobile-health technologies to improve health care service delivery processes: a systematic review and meta-analysis. PLoS Med 2013;10(1):e1001363 [FREE Full text] [doi: 10.1371/journal.pmed.1001363] [Medline: 23458994]
7. Fiksdal AS, Kumbamu A, Jadhav AS, Cocos C, Nelsen LA, Pathak J, et al. Evaluating the process of online health information searching: a qualitative approach to exploring consumer perspectives. J Med Internet Res 2014 Oct 07;16(10):e224 [FREE Full text] [doi: 10.2196/jmir.3341] [Medline: 25348028]
8. Ross J, Stevenson F, Lau R, Murray E. Factors that influence the implementation of e-health: a systematic review of systematic reviews (an update). Implement Sci 2016 Oct 26;11(1):146 [FREE Full text] [doi: 10.1186/s13012-016-0510-7] [Medline: 27782832]
9. Henriquez-Camacho C, Losa J, Miranda JJ, Cheyne NE. Addressing healthy aging populations in developing countries: unlocking the opportunity of eHealth and mHealth. Emerg Themes Epidemiol 2014;11(1):136 [FREE Full text] [doi: 10.1186/s12982-014-0021-4] [Medline: 25642276]
10. Li J, Talaei-Khoei A, Seale H, Ray P, Macintyre CR. Health care provider adoption of eHealth: systematic literature review. Interact J Med Res 2013 Apr 16;2(1):e7 [FREE Full text] [doi: 10.2196/ijmr.2468] [Medline: 23608679]
11. El-Gayar O, Timsina P, Nawar N, Eid W. Mobile applications for diabetes self-management: status and potential. J Diabetes Sci Technol 2013 Jan 01;7(1):247-262 [FREE Full text] [doi: 10.1177/193229681300700130] [Medline: 23439183]
12. Zhou W, Chen M, Yuan J, Sun Y. Welltang - a smart phone-based diabetes management application - improves blood glucose control in Chinese people with diabetes. Diabetes Res Clin Pract 2016 Jun;116:105-110. [doi: 10.1016/j.diabres.2016.03.018] [Medline: 27321324]
13. Cheung N, Fung V, Wong WN, Tong A, Sek A, Greyling A, et al. Principles-based medical informatics for success--how Hong Kong built one of the world's largest integrated longitudinal electronic patient records. Stud Health Technol Inform 2007;129(Pt 1):307-310. [Medline: 17911728]
14. Wong ACK, Shum LHY, Chan FKK, Cheung CSK, Tong ELH, Cheung NT. Easy access to electronic health record sharing system (eHRSS) via clinical management system (CMS) extension on-ramp system by Hong Kong healthcare providers in private sectors 2016. Asia Pacific eHealth Conference 2025. URL: https://apehc.ehealth.org.hk/2016/Call-for-abstracts/1%20Easy%20access%20to%20Electronic%20Health%20Record%20Sharing%20System%20(eHRSS).pdf [accessed 2025-08-29]
15. Electronic health record sharing system (eHealth). GovHK. URL: https://www.gov.hk/en/residents/health/hosp/eHRSS.htm#:~:text=The%20Electronic%20Health%20Record%20Sharing,(eHR)%20for%20healthcare%20purposes [accessed 2025-08-29]
16. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of Web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126 [FREE Full text] [doi: 10.2196/jmir.1923] [Medline: 22209829]
17. Tobacman JK, Kissinger P, Wells M, Prokuski J, Hoyer M, McPherson P, et al. Implementation of personal health records by case managers in a VAMC general medicine clinic. Patient Educ Couns 2004 Jul;54(1):27-33. [doi: 10.1016/s0738-3991(03)00184-8]
18. Bourgeois FT, Simons WW, Olson K, Brownstein JS, Mandl KD. Evaluation of influenza prevention in the workplace using a personally controlled health record: randomized controlled trial. J Med Internet Res 2008 Mar 14;10(1):e5 [FREE Full text] [doi: 10.2196/jmir.984] [Medline: 18343794]
19. Arshed M, Mahmud AB, Minhat HS, Ying LP, Umer MF. Effectiveness of mHealth interventions in medication adherence among patients with cardiovascular diseases: a systematic review. Diseases 2023 Mar 01;11(1):41 [FREE Full text] [doi: 10.3390/diseases11010041] [Medline: 36975590]

XSL•FO
RenderX

20. Pouls BPH, Vriezekolk JE, Bekker CL, Linn AJ, van Onzenoort HAW, Vervloet M, et al. Effect of interactive eHealth interventions on improving medication adherence in adults with long-term medication: systematic review. J Med Internet Res 2021 Jan 08;23(1):e18901 [FREE Full text] [doi: 10.2196/18901] [Medline: 33416501]
21. Bassi G, Mancinelli E, Dell'Arciprete G, Rizzi S, Gabrielli S, Salcuni S. Efficacy of eHealth interventions for adults with diabetes: a systematic review and meta-analysis. Int J Environ Res Public Health 2021 Aug 26;18(17):8982 [FREE Full text] [doi: 10.3390/ijerph18178982] [Medline: 34501572]
22. van Olmen J. The promise of digital self-management: a reflection about the effects of patient-targeted e-Health tools on self-management and wellbeing. Int J Environ Res Public Health 2022 Jan 26;19(3):1360 [FREE Full text] [doi: 10.3390/ijerph19031360] [Medline: 35162383]
23. Knitza J, Simon D, Lambrecht A, Raab C, Tascilar K, Hagen M, et al. Mobile health usage, preferences, barriers, and eHealth literacy in rheumatology: patient survey study. JMIR Mhealth Uhealth 2020 Aug 12;8(8):e19661 [FREE Full text] [doi: 10.2196/19661] [Medline: 32678796]
24. Van der Mispel C, Poppe L, Crombez G, Verloigne M, De Bourdeaudhuij I. A self-regulation-based eHealth intervention to promote a healthy lifestyle: investigating user and website characteristics related to attrition. J Med Internet Res 2017 Jul 11;19(7):e241 [FREE Full text] [doi: 10.2196/jmir.7277] [Medline: 28698168]
25. Chiu TM, Eysenbach G. Stages of use: consideration, initiation, utilization, and outcomes of an internet-mediated intervention. BMC Med Inform Decis Mak 2010 Nov 23;10:73 [FREE Full text] [doi: 10.1186/1472-6947-10-73] [Medline: 21092275]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
eHRSS: electronic health record sharing system
FDA: Food and Drug Administration
HbA1c: hemoglobin A1c

Edited by A Mavragani; submitted 22.01.25; peer-reviewed by AN Ali, D Tilden; comments to author 14.03.25; revised version received 03.04.25; accepted 20.05.25; published 26.09.25.

Please cite as:
Huang J, Zhong CC, Wong SH, Lo CY, Yim MK, Wong MCS
eHealth Applications Improve Glycemic Control in Patients With Diabetes: Randomized Controlled Trial
JMIR Form Res 2025;9:e67761
URL: https://formative.jmir.org/2025/1/e67761
doi:10.2196/67761
PMID:

©Junjie Huang, Claire Chenwen Zhong, Siu Hin Wong, Chung Yi Lo, Man Kin Yim, Martin C S Wong. Originally published in JMIR Formative Research (https://formative.jmir.org), 26.09.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Health Care Providers' Perspectives of Clinical Decision Support Tools for Pediatric Sepsis in Bangladesh: Qualitative Study

Shamsun N Shaima[1,2*], MBBS, MPH; Alicia E Genisca[3,4*], MD; Md Tanveer Faruk[1,5], MBBS, MPH; Md Fakhar Uddin[1], MSS; Akash Saha[1], MSS; Nadia Sultana[1], MSS; Nidhi Kadakia[6], MD, MPH; Monique Gainey[7], MS, MPH; Elleen Kim[8], BA; Kikuyo Shaw[3], BS; Farzana Afroze[1], MBBS; Joan Chepngeno[9], BSc, MPH; Atin Jindal[10], MD; Sifat A Chowdhury[1], MBBS; Md Jobayer Chisti[1], MBBS, MMED, PhD; Adam C Levine[11], MD, MPH; Stephanie C Garbern[3], MD, MPH

[1]International Centre for Diarrhoeal Disease Research, Dhaka, Bangladesh
[2]Department of Nutritional Sciences, School of Graduate Studies, University of Toronto, Toronto, ON, Canada
[3]Department of Emergency Medicine, Warren Alpert Medical School of Brown University, Providence, RI, United States
[4]Department of Pediatrics, Warren Alpert Medical School of Brown University, Providence, RI, United States
[5]Department of Epidemiology and Community Health, University of North Carolina at Charlotte, Charlotte, NC, United States
[6]Department of Emergency Medicine, Yale School of Medicine, New Haven, CT, United States
[7]Virginia Tech Carilion School of Medicine, Roanoke, VA, United States
[8]Warren Alpert Medical School of Brown University, Providence, RI, United States
[9]Brown University, Providence, RI, United States
[10]Department of Internal Medicine, Warren Alpert Medical School of Brown University, Providence, RI, United States
[11]Division of Biology and Medicine, Brown University, Providence, RI, United States
[*]these authors contributed equally

Corresponding Author:
Alicia E Genisca, MD
Department of Emergency Medicine
Warren Alpert Medical School of Brown University
55 Claverick Street
2nd Floor
Providence, RI, 02903
United States
Phone: 1 401 444 9891
Fax: 1 401 444 4307
Email: alicia_genisca@brown.edu

Abstract

Background: Sepsis, a life-threatening condition resulting from a dysregulated immune response to infection, disproportionately affects children in low- and middle-income countries (LMICs). Children with sepsis in LMICs face high mortality rates, with early detection and clinical monitoring posing significant challenges to effective management. There is great potential for digital technologies, such as wearable biosensor devices and mobile health (mHealth) clinical decision support (CDS) tools, together referred to as clinical decision support systems (CDSSs), to enable closer monitoring and more prompt recognition of children at risk of advanced sepsis and death. However, little is known about the perceptions of health care providers (HCPs) regarding the introduction of new digital health tools for pediatric sepsis care in LMICs.

Objective: The objective of this study was to assess HCPs' understanding, perceptions, and recommendations regarding the design and implementation of digital CDSSs for pediatric sepsis care in Bangladesh.

Methods: Between February and May 2024, 18 individual semistructured in-depth interviews were conducted with HCPs (nurses and physicians) at 3 urban hospitals in Bangladesh. The data were transcribed, translated from Bangla to English, and analyzed using a framework matrix analysis approach. Participants were asked about familiarity with digital health tools, feedback on CDSS design, perceptions of the system's utility, and barriers and facilitators to use of similar tools in clinical settings in Bangladesh.

XSL•FO
RenderX

Results: Participants reported overall positive perceptions toward the potential implementation of a CDSS for pediatric sepsis care in Bangladesh. Some key priorities for the design of a CDSS were durability, reusability, cost considerations, reliability, and accuracy. Clinicians desired the CDS tool to also have customizable alarm parameters and include additional functions such as glucose monitoring. Many favored audio (ringtone) or visual (light) alarms to alert about changes in captured vital signs. HCPs believed that a CDSS could enhance patient care by allowing greater staff capacity to monitor patients, reducing management time, and aiding in faster clinical decision-making, with some suggesting it could lower mortality rates. Concerns regarding implementation included internet availability, affordability of the wearable devices, and trust in the CDSS outputs compared to expert clinician judgement.

Conclusions: The findings of this study highlight HCPs' perceptions toward the potential of wearable biosensor devices and CDS tools (CDSSs) for improving pediatric sepsis outcomes in LMICs and highlight the need to address implementation challenges to ensure the effective integration of CDSSs into health care systems.

(JMIR Form Res 2025;9:e73451) doi:10.2196/73451

KEYWORDS

global health; sepsis; pediatrics; digital health; implementation science; mobile health

Introduction

Sepsis is a life-threatening condition characterized by a dysregulated immune response to infection, which can lead to death if not promptly treated [1]. It is one of the leading causes of child mortality worldwide, with more than half of all sepsis cases worldwide occurring in children younger than 5 years [2,3], and remains a significant global health concern [4]. As of 2017, there were an estimated 2.9 million sepsis-related deaths among children younger than 5 years [5]. The burden of sepsis disproportionately affects children in low- and middle-income countries (LMICs) for a multitude of reasons, including lack of human, infrastructural, and material resources necessary to adhere to the current best practice protocols, which are usually based on standards developed in high-income countries.

The management of potentially septic children is challenging, as many children with infections do not progress to having sepsis, while those with sepsis can rapidly and unexpectedly decompensate, developing shock or multiorgan failure. The ability to provide close monitoring and risk stratify septic children is often only available in higher-resourced emergency and critical care wards in referral or specialized pediatric care facilities in LMICs [2,3]. Early anticipation of poor outcomes, including shock, is a mainstay of managing sepsis, as rapid treatment initiation significantly improves outcomes and can prevent or minimize the severity of multiple organ failure [3,4]. In high-income countries, diagnostic and prognostic tools, continuous monitoring systems, and advanced resources for resuscitation are typically recommended to optimize care. However, in LMICs, the feasibility of implementing recommended measures may be limited due to a lack of such resources, increasing the risk of severe complications [3,6]. In Bangladesh, a lower middle–income South Asian country, sepsis remains the leading cause of childhood mortality [7].

Digital health technologies, including wearable biosensor devices, artificial intelligence, and mobile health (mHealth) clinical decision support (CDS) tools, have great potential to mitigate the lack of critical care resources in LMICs by enabling closer monitoring of vital signs, allowing for more prompt recognition of children at risk of advanced sepsis and death [8]. Given the lack of consensus in pediatric sepsis care protocols, CDS tools could be beneficial in streamlining clinical knowledge and patient care. Current developments are addressing this challenge of defining optimal sepsis criteria and promoting its facilitation through CDS tools [9,10]. Recent studies have also demonstrated the ongoing evolution of CDS tools, which may now incorporate AI-driven algorithms and predictive models based on real-time physiological data, rather than solely relying on laboratory diagnoses [11,12]. These advancements further enhance the ability of CDS tools to detect and risk stratify patients with sepsis, enabling clinicians to provide care more promptly [13,14]. It is also recognized that existing tools for sepsis management are often designed for use by highly trained health care providers (HCPs) or rely on infrastructure, such as electronic health records, to function. However, in LMICs, clinicians have varying levels of pediatric sepsis awareness, infrastructural availability, and digital literacy [15]. The adaptation of new tools for sepsis care in LMICs therefore necessitates customization to fit various context-specific factors [10].

Understanding the perceptions of HCPs regarding digital health and CDS tools for pediatric sepsis care in LMICs is critical to ensure the successful implementation and adoption of such technologies and realize their immense potential [16-18]. This study aimed to assess HCPs' understanding and perceptions of digital and CDS tools for pediatric sepsis care in LMICs, in the context of the development of a novel clinical decision support system (CDSS) for pediatric sepsis care in Bangladesh.

Methods

Study Setting, Population, and Design

This study was conducted as part of a larger study called REMEDIES (Remote Electronic Monitoring and Decision Support for Improved Sepsis Care), which aims to develop a novel, wearable device-enabled mHealth CDSS for sepsis care in Bangladesh. The CDSS consists of a wearable biosensor device that collects real-time vital sign information and is linked to an mHealth CDS tool that can report vital signs and use machine learning algorithms to predict sepsis [19]. Questions

XSL•FO
RenderX

were included to gain feedback on a prototype version of this CDSS, as well as on digital technology more generally.

Eighteen semistructured in-depth interviews were conducted in person with HCPs at 3 hospitals in urban Bangladesh from February to May 2024. The inclusion criterion was being an HCP who had at least 1 year of experience providing care for pediatric patients with sepsis at 1 of the 3 study sites. Purposive selection was used to select potential participants for inclusion, ensuring diversity of clinical experience. The only exclusion criterion was declining to participate in the study. Three nurses and 3 physicians were interviewed at each of the 3 sites. Interviews were conducted in the local language, Bangla, at the participants' respective hospitals. The hospitals included were the International Centre for Diarrhoeal Disease Research, Bangladesh (icddr,b) Dhaka Hospital in Dhaka, Kushtia General Hospital (KGH) in Kushtia, and Institute of Child and Mother Health (ICMH) in Matuail (Figure 1). The icddr,b Dhaka Hospital is a private, nonprofit research hospital in Bangladesh's capital city, which treats about 200,000 patients yearly, over half of whom are pediatric patients younger than 5 years. The hospital has a team of 42 doctors, 82 nurses, and 150 support staff [20]. KGH is a 250-bed government hospital that provides care to approximately 83,000 patients annually, 16% of whom are younger than 5 years. ICMH in Dhaka is a maternal and neonatal specialty care hospital, with over 90 beds dedicated to the pediatric and neonatal special care units. It is one of the largest Special Care Newborn Units in Bangladesh [21].

The SRQR (Standards for Reporting Qualitative Research) checklist is provided in Multimedia Appendix 1.

Figure 1. Study site locations. The purple icon represents Kushtia General Hospital; orange represents the International Centre for Diarrhoeal Disease Research, Bangladesh; and blue represents the Institute of Child and Mother Health.

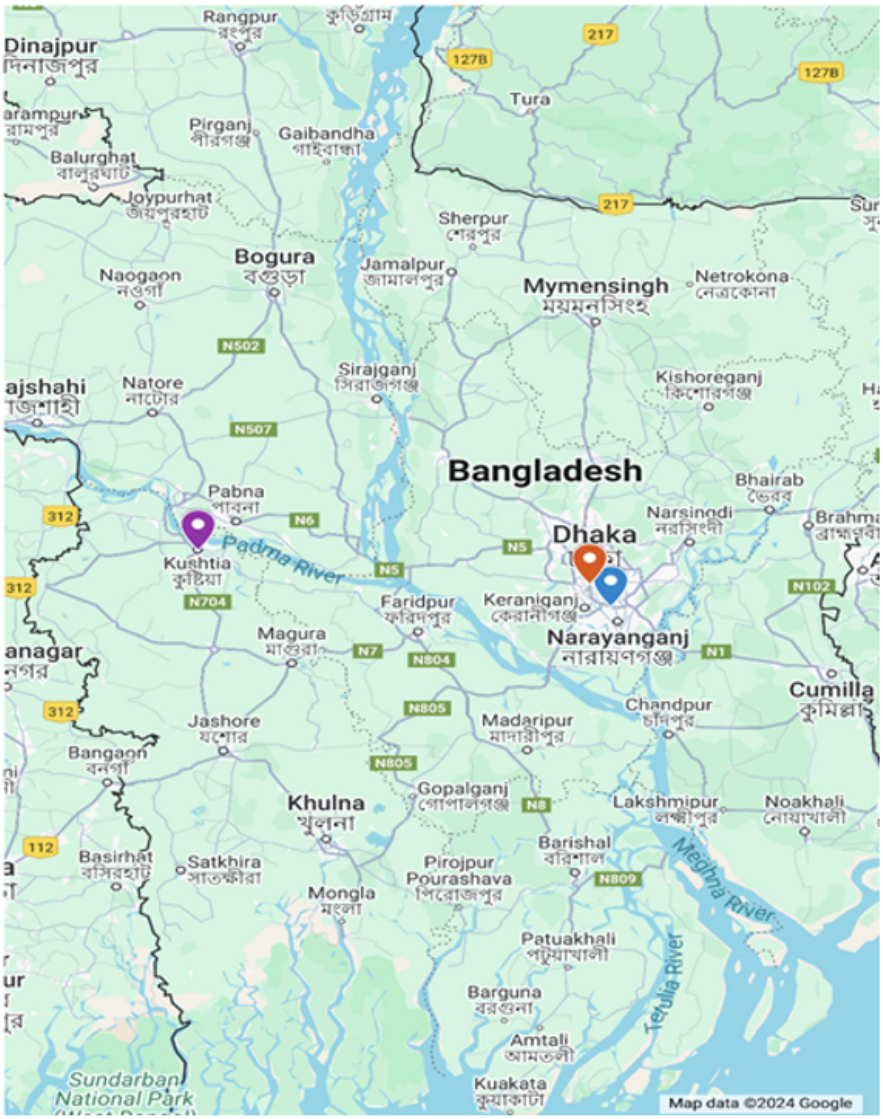

Study Training

The primary qualitative research team consisted of 2 physicians with pediatric and critical care experience and 2 anthropologists, who participated in a collaborative qualitative research theory and methodology training led by study investigators from icddr,b and Brown University from November 2023 until January 2024. Training included a 9-part lecture series on qualitative research as well as a practical component in which interviewers and the notetaker conducted mock interviews with other members of

XSL•FO
RenderX

the research staff to provide feedback on the interviewer's tone, provide alternative methods for asking questions, and prepare interviewers to troubleshoot typical challenges faced when conducting interviews. The first 2 participant interviews were conducted with other research staff observing in order to provide feedback after the interview concluded.

Data Collection

The interviews were designed to explore HCPs' experiences and challenges with pediatric sepsis management; awareness and perceptions of digital health tools, with a focus on their potential use in monitoring and treating pediatric sepsis; and recommendations and preferences regarding future implementation in Bangladesh and similar settings. Participant demographic information (eg, age and years of practice) was collected via a participant demographic questionnaire (Multimedia Appendix 2).

A comprehensive semistructured interview guide (Multimedia Appendix 3) was created to gather feedback on the development of the CDSS for pediatric sepsis care in Bangladesh. The guide consisted of several key sections, including participants' familiarity with digital tools, such as wearable biosensor devices and CDS tools, in both personal and professional contexts. During the interviews, participants were shown a wearable biosensor device prototype that linked to a prototype mHealth CDS tool, which captured live vital sign information. This was demonstrated to the participants using either the interviewer or a designated mannequin as a model. This allowed the participants to observe the real-time functionality of the wearable biosensor device and mHealth CDS tool displayed on a mobile device (Figure 2).

Figure 2. Flow schematic of the clinical decision support system. ECG: electrocardiography.

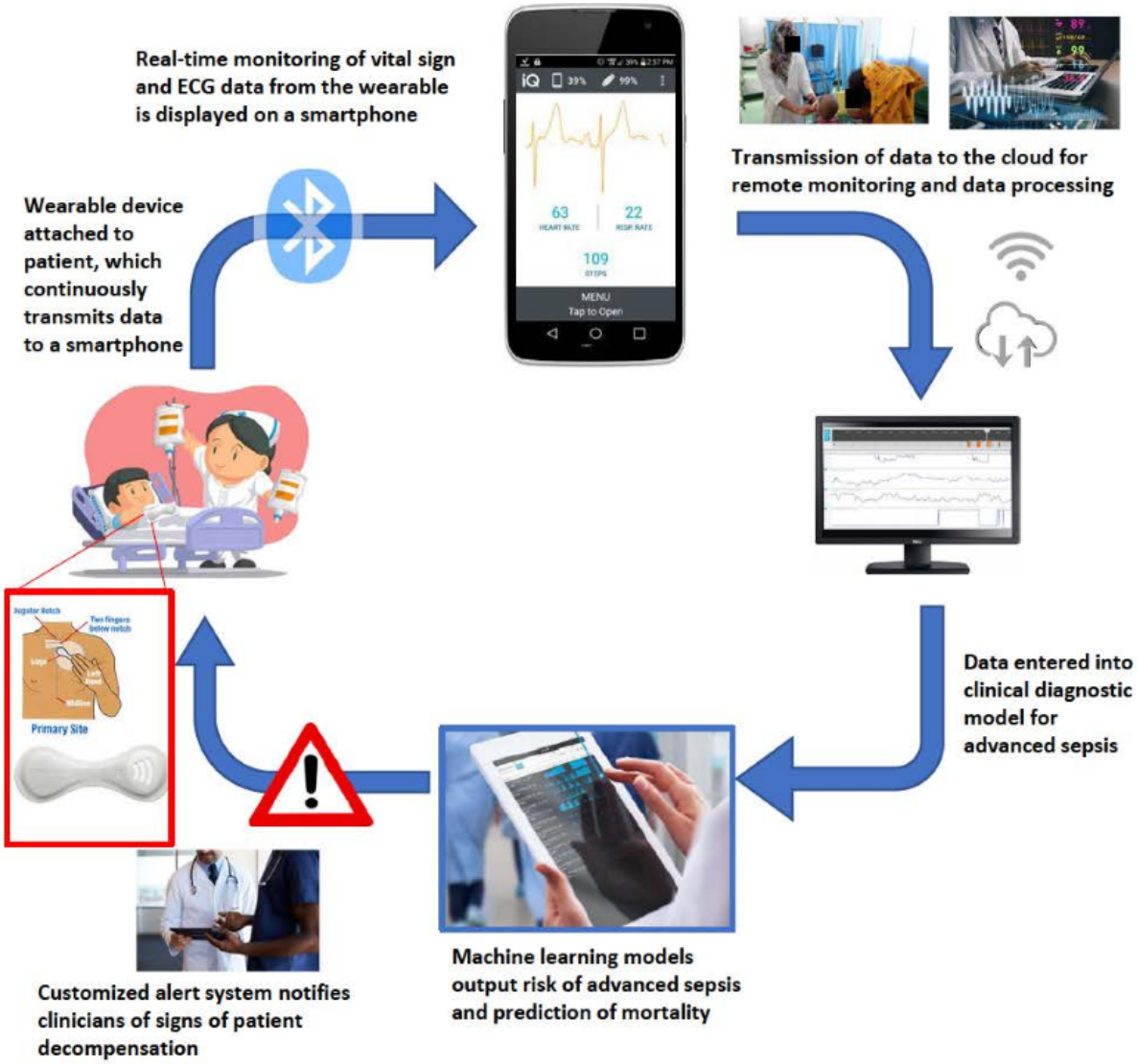

Data were then collected on the provider's opinions of the CDSS, including its potential benefits, challenges, considerations for implementation, and specific features and functionalities participants would find valuable, such as alert mechanisms and CDS capabilities. Each interview was conducted by a team of research staff and clinicians from icddr,b, who were fluent in the local language (Bangla) and English, and a notetaker was present at each interview to facilitate and clarify questions. The sessions were audio-recorded, transcribed, and translated into English for analysis. To mitigate participant fatigue, individuals were informed that they could take breaks at any time during the session, and culturally appropriate snacks were available to

XSL•FO
RenderX

support engagement throughout the interview process. Of the 18 interviews, only 3 ran longer than the proposed 90 minutes.

Debrief forms were completed by the interviewers and the notetaker after each session to summarize key points and assess data saturation. Data saturation was validated through a team-based consensus. Prior work has shown that data saturation can be satisfactorily obtained with 9-17 interviews [22], and our work analyzed 18 interviews. Data quality was assured by reviewing all English transcripts for clarity of meaning with the interviewer team.

Data Analysis

Data were analyzed using a framework matrix approach. A framework matrix is a systematic method for qualitative data analysis that organizes data into a matrix of rows (cases) and columns (codes), allowing researchers to compare and contrast data across and within cases while preserving the context of each participant's views [23].

Six framework themes were analyzed: (1) familiarity with digital health tools and devices, (2) perceptions about CDSS use for patient monitoring among children with sepsis, (3) perceptions about CDSS use for clinical decision-making among children with sepsis, (4) barriers and facilitators to the adoption of a CDSS, (5) feasibility, and (6) CDSS functionality.

Each theme was further split into codes corresponding to individual interview questions, and each participant was assigned to a row. The research team had primary and secondary coders (7 in total) to complete the framework matrices and provide summaries for each. Frameworks were double coded in this way to flag discrepancies between coders, followed by the resolution of any disagreements through consensus to enhance credibility and dependability. This iterative and collaborative process helped ensure consistency and reduced individual bias in the analysis. Our team maintained a detailed audit trail documenting analytic discussions throughout the study for confirmability and validity. We provide descriptions of participant experiences and health care settings to allow assessment for applicability to similar clinical contexts.

The multidisciplinary research team, which included Bangladesh- and US-based physicians, research assistants, and anthropologists, engaged in ongoing reflexivity to recognize and balance the varied positionalities while interpreting the experiences of the Bangladeshi physician and nurse participants.

Basic descriptive statistics were used to calculate frequencies, percentages, medians, and IQRs.

Ethical Considerations

Ethical approval for this project was obtained from the icddr,b Institutional Review Board (IRB) (including the Research Review Committee and Ethical Review Committee) and the Rhode Island Hospital IRB. Participants were informed that personal data would not be recorded or stored. Each participant was assigned a study ID based on study location and their role. Transcribed and translated data from the participants were always stored in a password-protected folder on a secure study computer. Written consent in Bangla or English was obtained after participants were allowed to ask questions. Participants from KGH and ICMH were offered a monetary honorarium in Bangladeshi taka (1600 [US $13.14] for physicians and 800 [US $6.57] for nurses) for their participation in the study. Participants from icddr,b, who were not permitted to receive a monetary honorarium, received gift items such as a diary and pen.

Results

Participant Demographics

A total of 18 participants (9 nurses and 9 physicians) were invited to take part in the study, and 6 interviews were conducted at each facility. The majority of participants were female and aged 25-34 years. The median number of years of experience treating sepsis was 7 years. The demographic characteristics of the participants are shown in Table 1. Each interview lasted between 50 and 100 minutes.

XSL•FO
RenderX

Table 1. Demographic characteristics of the participants.

Characteristic	Value (N=18)
Age (years), n (%)	
18-24	1 (6.0)
25-34	9 (50.0)
35-44	4 (22.2)
45-54	4 (22.2)
Gender, n (%)	
Male	5 (27.8)
Female	13 (72.2)
Education, n (%)	
Nursing diploma	9 (50.0)
Bachelor of Medicine, Bachelor of Surgery (MBBS) degree	9 (50.0)
Master of Public Health (MPH) degree	7 (38.9)
Working status, n (%)	
Part time	1 (5.6)
Full time	17 (94.4)
Current position, n (%)	
Nurse	9 (50.0)
Doctor	9 (50.0)
Approximate number of children with sepsis treated per month, n (%)	
2-5	5 (27.8)
6-10	4 (22.2)
11-20	4 (22.2)
>20	5 (27.8)
Experience, median (IQR)	
Years at the current position	8.5 (6-11)
Years treating children with sepsis	7 (5-9.5)

Familiarity With Digital Health Tools and Devices

The majority of respondents reported familiarity with a range of digital devices, including smartphones, tablets, laptops, wearable devices (eg, smartwatches), and desktop computers. Additionally, most respondents reported being familiar with digital health devices used in health care workplaces, such as hospital computers, smartphones, ultrasound machines, and cardiac monitors, and digital tools for patient assessment, such as digital thermometers, scales, oximeters, glucometers, ventilators, and blood pressure monitors. However, some respondents also understood “digital” to include electronic devices that were not always classified as digital devices, such as infusion pumps, mechanical ventilators, and infant warmers.

> *At work we have cardiac monitors, infusion pumps, syringe pumps, digital thermometers, pulse oximeters and many more.* [Nurse from icddr,b]

> *I have to use different kinds of digitized machines in the hospital. First of all, when the babies come, we take the weight at first on [the] weight machine, which is digital. Then, we use [a] digital thermometer, digital glucometer, [and] check the saturation [with a] pulse oximeter, and we have a cardiac monitor through which we attach the baby with the leads.* [Physician from icddr,b]

> *Then there are devices. I use syringe pump, CPAP, then mechanical ventilator. Also warmer is one kind of device. These are the maximally used.* [Physician from ICMH]

When asked specifically about their familiarity with wearable biosensor devices, HCPs commonly identified pulse oximeters and traditional (ie, wired) bedside cardiac monitors as wearable digital devices used in clinical settings. Personal use or general awareness of smartwatches was also prevalent, although these were not mentioned in clinical use in health care settings. Awareness of wireless, wearable biosensor devices was mentioned by members of icddr,b in a prior phase of this study, although this was not mentioned by participants from the other study sites. Overall, although a notable number of participants expressed a lack of familiarity with wearable biosensor

XSL•FO
RenderX

technology in health care settings, the idea of a wearable biosensor device for monitoring vitals for clinical care was readily understood after a brief demonstration and was overall received positively.

> *Yes, just like the pulse oximeter [it] can be applied to hands and feet. Currently we use a pulse oximeter, something like this. It looks like you put it on your arm and leg. I mean, something like a pulse oximeter that would record my totals (vitals), that would be great.* [Nurse from KGH]

Perceptions About CDSS Use for Patient Monitoring Among Children With Sepsis

In terms of using a digital device for monitoring children with sepsis, overall, both nurses and physicians felt that a wearable biosensor device would be helpful in providing continuous vital sign monitoring. Several providers explained that by relying less on manual vital sign assessment, they would be able to provide more efficient care and the workload on health care workers could be decreased, especially during times of high patient load and personnel shortage. More importantly, wireless wearable biosensor devices were identified as a way to minimize patient discomfort related to monitoring, especially in children.

> *Non-invasive alternative [and] minimizing discomfort for children. Just like how pulse oximeters revolutionized oxygen level monitoring non-invasively, future innovations might similarly advance capabilities like blood gas or glucose monitoring, potentially transforming clinical practice.* [Physician from ICMH]

Furthermore, both physicians and nurses at each of the clinical sites emphasized their perception that the CDSS should be used primarily for monitoring purposes, and making a diagnosis (eg, for sepsis) would require additional information.

> *Diagnosis will require everything starting from history. Here it only seems like monitoring. In terms of monitoring, it will be very good, but there is much more in the diagnosis area that seems not dependent on the device.* [Physician from KGH]

Several nurses mentioned that, as a continuous, remote monitoring device, the CDSS would allow them to efficiently communicate a patient's status better by facilitating the early detection of deteriorating patients and prompting them to notify their physicians more rapidly. Many participants also emphasized the benefit of this type of technology over the current standard of care for monitoring in their workplaces.

> *This would enable us to inform the doctor promptly, and the doctor could also monitor the data. As nurses, we would be able to provide special care to these babies.* [Nurse from ICMH]

> *If the monitoring system is like this, we can do a lot of good with it.* [Nurse from KGH]

Additionally, many felt the CDSS would provide reassurance to the caretakers or families of patients, as the caretaker would know their child was being continuously monitored, especially in lower-resourced hospitals, where caretakers are often relied upon to notify staff of changes in the clinical condition.

> *Yes, it would be beneficial for doctors and nurses to counsel the patient's guardians, reassuring them that the monitoring machine is always connected to their child. This would provide peace of mind, knowing that medical staff are actively monitoring their condition. While ultimate outcomes are in Allah's hands, constant monitoring can alleviate mental distress for families.* [Nurse from ICMH]

Perceptions About CDSS Use for Clinical Decision-Making Among Children With Sepsis

Both groups of providers expressed their belief that the overall quality of care would be improved by digital devices, allowing them to identify sepsis earlier and promptly take appropriate action (eg, adjusting the course of treatment in response to information from digital tools).

> *In all likelihood, if this device is in front of me, it won't take more than a minute to complete everything by monitoring vital signs because every minute counts for sepsis.* [Physician from icddr,b]

> *It saves me time, and I can give that time to another patient.* [Nurse from icddr,b]

> *I can make an early diagnosis that the child has sepsis and start early sepsis management.* [Physician from icddr,b]

After demonstrating an example of a wearable biosensor device and mobile CDS tool, the providers across all hospital systems further detailed their perspectives on how such a CDSS would be used to make clinical decisions. Nurses focused more on how quicker notification of abnormal vitals will guide them to seek assistance from the physician, while physicians discussed how such a system can be used to prioritize more severe patients, facilitate early treatment initiation, and assist with referral decisions. Several physicians also highlighted the versatility of the wearable biosensor device, explaining how the device's vital outputs can also be used for clinical vital sign monitoring in other diseases besides sepsis.

> *As it is a continuous vital monitoring device, vitals are coming in. If I can do early identification then my patient care, which means patient categorization, will be easy. A patient with an abnormality in [their] vitals will be categorized beforehand. So, in that case, we will be able to identify sepsis in advance if it is going from sepsis to severe sepsis. We can identify first seeing those vitals. So, if we get the information from a single device from the individual patches of each patient, that's our early identification.* [Physician from icddr,b]

> *I can see that the patient is fine or deteriorating, and I can take action immediately.* [Physician from KGH]

> *I should prioritize that patient. There may be more opportunities for early intervention...I can ensure early intervention.* [Physician from ICMH]

XSL•FO
RenderX

Compared to providers at icddr,b, those at other sites, specifically physicians, seemed more cautious about the CDSS's implementation and indicated that providers should not over-rely on digital devices as they cannot replace clinical gestalt.

> *The number one caution to be taken is primarily [to] never trust a machine over a human. Most importantly, it is most important that my clinical knowledge and expertise is not a substitute or its replacement because the critical patient needs to be judged by my own clinical eye. Overemphasis [on] the device can harm someone's condition.* [Physician from ICMH]

> *Diagnosis will require everything starting from history. Here it only seems like monitoring. In terms of monitoring, it will be very good, but there is much more in the diagnosis area that seems not dependent on the device.* [Physician from KGH]

CDSS Use for Risk Stratification and Clinical Prediction

Respondents reported that the function of the CDSS to provide risk estimates of sepsis severity or mortality risk would help to quickly identify patients at the highest risk of death, inform family discussions regarding the patient's prognosis, and help clinicians allocate resources.

> *Being able to predict these conditions (MODS, shock) in advance is crucial. By predicting them early, we can monitor the likelihood of death in advance and prepare accordingly.* [Physician from ICMH]

> *If I can pick the deterioration very early, my mortality rate will be much lower. The deterioration will be reduced, I will be able to shift the child's treatment plan very soon and go to high management. The children will be saved faster.* [Physician from icddr,b]

There also appeared to be discrepancies in the understanding of the different functions of the device in terms of both monitoring and risk stratification. Some respondents, when asked specifically to give feedback about risk estimate predictions, appeared to understand this function as the same as monitoring. Additionally, some respondents reported doubts about whether they could trust the risk estimates provided by the CDSS.

> *If there is a risk of death, after looking at the tool, I have to [think] myself. It is not possible for me to completely trust the device alone for the patient.* [Physician from ICMH]

Barriers and Facilitators to the Implementation of a CDSS

Barriers

The key barriers to the adoption of a CDSS identified by participants included the following: need for specific training on how to use the devices, perceived high device cost, device maintenance challenges, high patient volume, limited personnel, and concerns about reliability.

Both nurses and physicians emphasized the need for proper training, infrastructure, and device availability and maintenance to successfully use the device in their respective workplace settings. Several highlighted the importance of first educating health workers on the advantages of the devices, suggesting that a greater understanding of the benefits of the device was important to support increased uptake, given a general lack of awareness of newer digital technologies in their workplaces.

> *At first, keep in mind that the doctors and nurses should be sensitized about the device. If they are aware of the perks of using this device, then they will definitely concentrate on using this device and have the best use of it.* [Physician from KGH]

The cost associated with the CDSS, including both initial purchase and ongoing maintenance, and whether additional financial burden would be placed on patients and their families, emerged as a prominent barrier. Several providers, noting that their hospital primarily serves low-income patients, expressed worries about whether the costs of the devices would be borne by the hospital or passed on to families, potentially adding financial strain. The importance of a realistic and sustainable financial model to ensure long-term uptake and use of these technologies was also stressed by participants.

> *Often, if the cost is high, the family cannot afford it. Some patients are too poor to afford NICU admission treatment. So, how will the cost of the machine be covered?* [Nurse from ICMH]

> *If the cost were lower and there was a system in place to provide more devices to the hospital, it would be beneficial for both us and the children. We could ensure that all children with sepsis receive one, allowing us to monitor their vital signs continuously.* [Nurse from ICMH]

> *It must be affordable because it is a government hospital where very poor people come to get services, [people] who think they don't have the capacity to get services outside.* [Nurse from KGH]

Most providers emphasized that training would be essential for both physicians and nurses before implementing new digital tools, particularly in terms of understanding how to properly handle and use the devices. Some respondents suggested that training a few individuals in advance, who could then train the rest of the staff (ie, train-the-trainer approach), would be the most effective approach. Others felt that simultaneous profession-specific training for both nurses and physicians would be more beneficial.

> *If everyone can use it, understand it properly or understand the monitoring properly, if everyone learns it or slowly gets used to it, ultimately the workload on me will be reduced. The workload on other doctors or colleagues who will come in my place or who are in other shifts will also be reduced. When the workload on a doctor is less, the quality of work improves.* [Physician from icddr,b]

> *If our seniors can provide us with sufficient training, then I believe it will definitely be helpful. We also*

XSL•FO
RenderX

> *have our sisters [nurses] here, and doctors as well, so if everyone can be trained, then it will be beneficial.* [Physician from icddr,b]

High patient load and limitations in terms of manpower were barriers identified by both physicians and nurses across the different hospitals. The providers reported that pediatric wards usually have a high patient volume and are busy, and therefore, managing and tracking devices could be challenging. One of the physicians noted that if manpower issues could be addressed, using digital devices would be feasible.

> *I would say this facility is patient based. We have uncountable patients here. Whereas our capacity in pediatric ward is 20 and in SCANU [is] 11. In that place, we always have more than 20 patients in our facility. This is the main barrier of using devices in our facility. In case of manpower, we have enough sisters [nurses]. We lack doctors, and I think that can be overcome, and thus digital device can be used. No problem.* [Physician from KGH]

In addition, there were concerns about whether the wearable biosensor devices could be reused for multiple patients and their risk of causing nosocomial infection.

> *I think it's better not to transfer the device from one patient to another because it was attached to their body. Doing this could increase the risk of spreading infections, especially since children have weaker immune systems. Reusing the device could spread germs, so it's best to use one device per child.* [Nurse from ICMH]

Many respondents discussed that maintenance of the devices could be a potential challenge, particularly for devices that require charging. A nurse from icddr,b shared that wearable biosensor devices might be prone to damage or contamination, for example, from vomit and stool, as children with sepsis and infectious diseases often experience these symptoms.

> *In icddr,b, I don't think there will be any constraints. But since it is a diarrheal hospital, there's a possibility that the device [will] get ruined with puke and stool. It can be a problem since the kids tend to throw up.* [Nurse from icddr,b]

Respondents also identified several barriers related to perceptions of a CDSS, including negative attitudes, resistance to change, a lack of trust in the new technology, and difficulties in integrating the device into daily clinical practice. One of the providers explained that acceptance of new devices could be a challenge, as providers are accustomed to traditional methods, and adapting to new technologies takes time. A few providers also felt that new tools would not be as reliable as their current standard of care and would necessitate that providers verify their accuracy.

> *If we are introduced with a new device, first of all, my challenge will be how to use that?! I will face difficulty about if I am properly using the tool and if I am understanding the indications of the tool. I mean the total use of the device and understanding the whole use of the tool will take some time at first. Second, acceptance. We are habituated with old things, and hence we take some time to accept new things.* [Physician from icddr,b]

Providers expressed concerns about the sustainability of the implementation of a CDSS, particularly once the research project support ends. They felt that, without follow-up or ongoing maintenance, devices may stop functioning properly over time. It was also emphasized that a lack of coordination and administrative commitment to maintenance and quality improvement could hinder the long-term use and effectiveness of a CDSS.

> *This lack of coordination, the administrative sincerity [commitment] to improve the quality of work is a serious problem. If we can't overcome it, then [why you are] spending a lot of money on …studies, I don't understand, if they can't benefit the people!* [Physician from KGH]

> *It will cost money. They [hospitals] have no money to buy medicine; how will they buy it? They may agree with the government. But when it ends later, it will cause problems. It will cause problems with quality; it will cause problems with availability.* [Physician from KGH]

Facilitators

Identified facilitators to the use of a CDSS for the management of children with sepsis were ease of integration with existing patient monitoring systems, ease of use, availability of infrastructure including electricity and reliable Wi-Fi, staff readiness, and perceived benefits compared to the current standard of care. One provider highlighted that the resources they were currently using to manage sepsis were not enough and that there was a need for better tools and technology. A nurse at ICMH added that the availability of phone chargers at each bed, along with Wi-Fi access throughout the hospital, supports the feasibility of adoption of a CDSS in their current workplace.

> *If there is a digital device, the less it can be handled the better. Now if there is a storage space, then the handling will be less. And here there is a charger point on the bed in every bed. There should not be much difficulty for them.* [Nurse from ICMH]

Another nurse further emphasized that providers would generally be welcoming to a CDSS that could improve care for children with sepsis.

> *Often, we may not notice a child's condition has deteriorated until we check again because we are also attending to other children. A device that detects when a child's condition worsens due to sepsis, just by being in contact with their body, would help us give that child immediate attention. This would improve the ability to monitor and respond to the needs of septic children more effectively.* [Nurse from ICMH]

In addition, providers reported that a CDSS could help improve workflow since other staff could use the devices and report updates to the physicians. The devices were also seen as valuable

XSL•FO
RenderX

for providing rapid, continuous, and remote updates on patients' conditions, helping with early detection and management, and enabling the monitoring of multiple children at once. One physician noted:

> *...actually, I'm going to be the only doctor working. When my mobile shows that, it is definitely very easy for me, comfortable for me, that I can see it in one mobile. But if multiple devices are connected, like connected with my sisters [nurses], if for some reason I went to the washroom or I went to pray or to a meeting or to dinner or lunch, then for any reason if the sister realizes that the patient is getting worse, she can immediately inform me. Then, I could also follow-up with the sisters, then there would be no restrictions on the follow-up of the children in my absence.* [Physician from icddr,b]

Feasibility of CDSS Implementation

Both nurses and physicians considered the CDSS feasible for implementation into their sepsis treatment protocol. They specifically highlighted the potential for the monitoring system to improve the quality of care, save time for providers, and improve workflow efficiency by reducing the number of required staff. It was also noted that the CDSS would alleviate nurses' anxiety related to patient monitoring by providing a reliable way to easily monitor patients' stability. HCPs also expressed that its wireless capacity supports integration by reducing entanglement of wires from traditional monitoring devices and lowering the risk of infection spread due to the remote aspect of wireless, wearable biosensor devices.

> *Yes, it will be convenient for work. When the doctor comes for rounds, they can start treatment immediately upon hearing from me. I'll connect the device via my phone and inform them about the child's condition. The doctor won't need to spend much time monitoring because they'll see what they need to see, and based on that, they can take their next steps.* [Nurse from KGH]

> *It's good. We won't have to touch the baby that much. We can see everything in the monitor, not having to touch the baby. Less touches mean less chances of infection.* [Nurse from KGH]

> *It will be very beneficial for all of us who work in the children's department, because as a nurse when I get updates on these things for a baby... as soon as I put on the monitor and know the baby's total (vitals), mentally I will be relieved that this baby is fine.* [Nurse from KGH]

However, there were reservations about its feasibility for initial integration with existing monitoring equipment and trust in its functionality, for which participants stressed the importance of training.

> *There will be no issues from doctor, nurse, or patient side. But currently I'm not sure if there will be any technical problem. I'm not sure if it will be well functioning all the time…I mean we will have to check if there's any technical drawback.* [Physician from KGH]

Preferences for CDS and Alerts

Both physicians and nurses expressed that the mHealth CDS tool would be most useful if it provided alerts for abnormal vital signs. Clinicians desired the mHealth CDS tool to accurately and reliably collect standard vital signs (such as heart rate, respiratory rate, blood pressure, and oxygen saturation), and have adjustable alarm parameters. Many favored audio (ringtone) or visual (light) alarms to alert about changes in captured vital signs. One provider cautioned that alarms should not be continuous or excessive to avoid alarm fatigue and disturbances to the patient.

> *Having an alarm system is beneficial. When the heart rate or saturation decreases, the alarm alerts us, allowing me to promptly inform the doctor upon arrival. The doctor assesses the patient, confirming if the baby is fine. This process is beneficial for us.* [Nurse from ICMH]

> *Continuous alarm can also disturb the child in both day and night. It can also create problems for the duty doctor or nurse, so the alarm should only beep during the deterioration.* [Physician from KGH]

Preferences for CDSS Functionality

Both nurses and physicians at multiple hospitals discussed the advantages of having a central monitoring system that allows multiple patient vitals to be displayed on a single monitor. A feature to connect multiple patient devices to a single mobile device or central system was noted to be particularly helpful in settings with limited staff resources. As one physician expressed:

> *If multiple devices can be connected to one phone, it will be much more helpful because normally I am the only doctor. I do not have 14 doctors, who will be connected for 14 patients.* [Physician from icddr,b]

Another physician from icddr,b described the advantage of being able to access patient data remotely on a mobile device, explaining:

> *It's a small device. I can follow up on my mobile. I can't go to every child's bed to see the monitors. I can follow up with my device while sitting... Every child's follow-up is in front of my eyes. I can notice while doing any work, so I can catch any change. It will be very easy for me.* [Physician from icddr,b]

Additionally, nurses and physicians underscored the need for additional features, such as blood pressure measurement, and suggested that the CDSS should also have the capacity to include laboratory biomarker monitoring and notifications such as blood count, bilirubin, glucose, and additional electrolyte levels, enabling them to monitor other parameters that could indicate sepsis. Nurses recommended the ability for health care staff to set alarm parameters and silence alarms or mark alarms as seen. Physicians also mentioned that the CDSS should be able to detect seizure activity and apnea, as well as assess capillary refill time and hydration status.

XSL•FO
RenderX

> *One is of course heart rate, [then]saturation, respiratory rate, and if there is temperature then it is very good. Sepsis children develop hypothermia more often, [so] these four are mandatory.* [Physician from ICMH]

> *The alarm must [have] respiratory rate, heart rate and saturation. Saturation is not there for now. [Also] temperature, of course, there should be an alarm if there is hypothermia.* [Physician from ICMH]

Physicians highlighted the importance of durability and data storage capacity, recommending that patient information could be archived for later review if needed.

Both nurses and physicians preferred a wireless wearable biosensor device that could operate without having to be plugged into an electrical outlet and could use alternative power sources, such as solar power, to ensure continuous monitoring. Nurses specifically noted the need for a small, easily cleaned, skin-friendly wearable biosensor device that is appropriate for infant and child use and that could be applied to multiple areas of the body without causing skin irritation in infants with delicate skin.

> *The wireless patch, on the other hand, is very nice, modern, and comfortable to use. Cables in monitors cause a lot of trouble, especially when a bubble CPAP line gets twisted with the monitor cables, [it] creates a clumsy and tangled situation. From a hygiene point of view, the wireless patch is much better. Although we wash the wires with alcohol and reuse them for patients, the wireless option is more hygienic and avoids the hassle of dealing with cables.* [Nurse from icddr,b]

> *The thing is that this should be skin friendly. That is, the patch and the placement that I am opening should be skin friendly, soft. The children should not feel [that the device is] too hard, and there should not be any irritation. This is my first point for comfortability because what we often see is skin rashes [on children] when we put monitor cords on them. They get red, they get irritated, [or] children get hurt. That's why being skin friendly is the most important thing for this.* [Physician from icddr,b]

Discussion

Primary Findings

The findings of this study underscore the potential impact and high perceived feasibility and acceptability of a CDSS (wearable biosensor device and mobile CDS tool) for pediatric sepsis management in low-resource settings, such as Bangladesh. The nurse and physician participants in this study expressed overall high optimism regarding the potential positive impact that a CDSS could have in alleviating the challenges they face in sepsis management. By providing real-time continuous vital sign monitoring, a major current limitation due to resource constraints, participants perceived that novel CDSS technologies would allow them to prioritize high-risk patients and initiate timely interventions, thereby reducing mortality rates. Additionally, participants noted that a CDSS could potentially reduce staff workloads, particularly during times of high patient volumes and in clinical settings with health care worker shortages, by reducing the need for frequent manual assessments and enhancing workflow efficiency.

Participants often understood the real-time monitoring and predictive risk scoring capabilities of this CDSS as a single function, making it important to clarify the distinction and provide dedicated training on functionality during implementation efforts. The providers acknowledged that real-time monitoring of vital signs can enhance clinical decision-making by allowing providers to prioritize deteriorating patients. Additionally, some participants identified the value of a predictive algorithm that can provide estimates of mortality risk and proactively alert providers to patients who have a high risk. Prior work on digital CDS tools in Bangladesh has demonstrated that prediction models derived from clinical features can outperform the current standard guidelines for clinical management in both accuracy and reliability [24]. Using mobile apps to deliver real-time decision support illustrates how data-driven risk assessments can be feasibly implemented by clinicians in a low-resource setting.

While feedback on the utility of digital technology for pediatric sepsis was largely positive, HCPs identified several essential conditions for successful integration of such tools in Bangladesh and other LMIC contexts. They emphasized that, given the existing gaps in sepsis protocol adherence and technology use, targeted training on device functionality and the importance of pediatric sepsis monitoring would be crucial. Nurses and physicians also recommended adapting the device to include features tailored to their specific needs, workflows, and infrastructure availability. They indicated that having centralized monitoring devices capable of monitoring multiple patients simultaneously, introducing alert systems customizable according to clinician preferences and patient needs, and having comprehensive data storage would enhance usability and acceptability. Concerns were raised about the device's affordability and durability, with clinicians stressing the importance of making it accessible, especially in government-funded hospitals that serve low-income patients.

Participants were overall highly engaged by the prospect of digital technology for pediatric sepsis care and thought creatively about solutions to overcome potential implementation barriers, such as suggesting solar panels to ensure an electrical supply to charge devices in settings where electricity is not continuously guaranteed. Additionally, participants mentioned that other priority features should be included in future wearable-based monitoring systems, such as electrolyte, glucose, and laboratory biomarker monitoring. While this is not feasible in the prototype design of the system this research team is developing, wearable biosensor devices capable of electrolyte monitoring through sweat and wearable ultrasound devices are rapidly expanding the possibilities available for monitoring through novel wearable technologies [25].

Comparison to Prior Work

Our findings support themes identified in previous research on CDS tools, particularly the importance of flexibility and

XSL•FO
RenderX

customization. One study highlighted the need to tailor CDS tools for pediatric sepsis to varying levels of medical training as well as local differences in culture, language, and resources [10]. Similarly, the participants in our study underscored the value of user customization and mentioned that the ability to tailor alert parameters at the level of individual providers would enhance workflow with the tool. The importance of administrative support to integrate novel digital technologies into existing systems was also a key finding. Other considerations for implementation include having offline functionality or solar-powered battery options to mitigate internet connectivity challenges and having a simple user interface.

Limitations

This study has several limitations that must be considered when interpreting the findings. First, apart from 1 participant, the health care workers who participated in the study did not have direct experience using wearable biosensor devices. As such, their perceptions and feedback may differ from those of providers who have hands-on experience with wearable biosensor devices and newer digital technologies.

Second, the study did not include other types of stakeholders, such as policy makers, hospital administrators, or other clinical personnel, whose perspectives and operational insights could provide a more comprehensive understanding of the device's potential impact on workflows and resource use.

Third, the hospitals included in the study are not representative of all health care facilities in Bangladesh, particularly those in more rural or remote areas where resource constraints and workflows may differ significantly. In LMIC settings, it is essential for any CDSS to function without relying on continuous power sources or internet connection and have the capacity to store data offline when necessary. Having a rechargeable and reusable wearable biosensor device and considering alternative power supplies, such as solar power, to ensure charging sources for mobile devices can improve the feasibility of implementing this system in low-resource settings. The generalizability of the findings to hospitals operating in these settings is limited, and the findings are likely most relevant to relatively high-resourced hospitals in urban regions of other LMICs. This research team is currently planning to conduct further preimplementation research in public, rural, and hard-to-reach hospitals in Bangladesh to gain a better understanding of implementation barriers and facilitators in more diverse settings in the country.

Lastly, the findings may not be widely applicable beyond the specific context of this study, given the unique organizational and cultural characteristics of the hospitals included. Broader applicability would require further validation in diverse health care settings, both within Bangladesh and in other low-resource contexts.

Future Directions

Studies to evaluate the clinical impact of digital CDS tools in LMICs are greatly needed to better understand the true value that these tools may provide over the standard of care and traditional clinical management guidelines and tools. More research is also needed regarding how clinicians, particularly in LMICs, understand clinical prediction models and what factors may influence their trust or mistrust of prediction models, particularly in an age of increasing prevalence of machine learning and artificial intelligence in clinical care.

Notably, there appeared to be challenges among nurse participants in differentiating the functions of vital sign monitoring from clinical risk prediction in the proposed tool demonstrated to them. This likely reflects a lack of familiarity with other CDS tools and prediction models in clinical care, as well as difficulty communicating abstract concepts, such as risk prediction, among general practice nurses with minimal or no formal statistical exposure in their training. Further research should explore how best to educate nurses and other health care workers in low-resource settings about clinical prediction models in clear and understandable terms in order to build confidence and trust in using CDSSs and prediction models for potentially improving the care of patients and supporting their clinical decision-making. It is also possible that these issues were compounded by participant fatigue, as the questions addressing this distinction were posed toward the end of the interviews.

Efforts to ensure financial accessibility were identified as critical to any plans to scale up the use of similar technologies in Bangladesh. This includes exploring sustainable funding models, such as subsidies, public-private partnerships, and integration into national health systems, to reduce cost barriers for health care facilities and patients. Without such measures, the adoption and widespread use of these tools may remain limited, undermining their potential to improve pediatric sepsis outcomes. Robust training programs for HCPs were also identified as crucial for the scale-up of these technologies, as they help build familiarity and confidence in using these technologies while addressing gaps in understanding predictive functions and decision-support capabilities.

Implementation science frameworks offer a valuable structure for guiding the design and evaluation of interventions. As this study represents the preimplementation phase, our goal was to first explore and characterize context-specific barriers and facilitators. Our future research will use the Consolidated Framework for Implementation Research 2.0.

Conclusion

While wearable biosensor devices and CDS tools were viewed as a promising solution for improving pediatric sepsis management in Bangladesh and other LMICs, careful planning and a nuanced understanding of the local context will be essential for the successful implementation of these technologies in the future. These tools have the potential to bridge critical gaps in early detection and timely intervention for pediatric sepsis, but their success hinges on addressing key factors, including adapting digital technologies to suit local clinical contexts and needs to ensure usability and effectiveness; customization of CDS features by end users; ensuring adequate infrastructure, particularly electrical supply; and providing profession-specific training on how the tools can be best used to support health care workers and patients.

XSL•FO
RenderX

Finally, engaging a broader range of stakeholders, including nurses, technicians, hospital administrators, and policy makers, will be critical in identifying and mitigating potential operational and systemic challenges during implementation. By adopting a holistic and inclusive approach, this model could inform the integration of similar digital health technologies across other low-resource settings, paving the way for scalable and sustainable advancements in global health.

Acknowledgments

We would like to thank the clinical, research, and administrative staff of icddr,b Dhaka Hospital, Institute of Child and Mother Health, and Kushtia Hospital for their support of this study and the providers for making this research possible. We would also like to express our gratitude to Dr Rochelle Rosen for her invaluable guidance and insightful advice throughout the course of this research.

Funding for data collection was provided through a grant from the US National Institutes of Health Fogarty International Center (R33TW012211). The funder had no role in the study design, data collection, or reporting processes. The project described was supported by Institutional Development Award Number U54GM115677 from the National Institute of General Medical Sciences of the National Institutes of Health, which funds Advance Rhode Island Clinical and Translational Research (Advance RI-CTR). Support was received from the qualitative team in the Biostatistics, Epidemiology, and Research Design core. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: FA (equal), MJC (equal), ACL (equal), SCG (equal)
Data curation: SNS (equal), AEG (equal), MTF (equal), AS (equal), NS (equal), MG (equal), EK (equal), KS (equal), SCG (equal), MFU (supporting)
Formal analysis: AEG (lead), SNS (equal), MTF (equal), SCG (equal), MG (equal), EK (equal), KS (equal), MFU (supporting)
Funding acquisition: MJC (equal), ACL (equal), SCG (equal)
Investigation: SNS (equal), MTF (equal), AS (equal), NS (equal)
Methodology: MJC (equal), ACL (equal), SCG (equal), MFU (supporting)
Project administration: SCG (lead), AEG (supporting), JC (supporting)
Resources: SCG (lead), SNS (equal), AEG (equal), MTF (equal), MFU (equal), AS (equal), NS (equal), NK (equal), FA (equal), AJ (equal), MJC (equal), ACL (equal)
Supervision: AEG (equal), SCG (equal)
Visualization: All authors
Writing – original draft: All authors
Writing – review & editing: AEG

Conflicts of Interest

None declared.

Multimedia Appendix 1
SRQR checklist.
[PNG File , 257 KB - formative_v9i1e73451_app1.png]

Multimedia Appendix 2
Participant demographic questionnaire.
[DOCX File , 14 KB - formative_v9i1e73451_app2.docx]

Multimedia Appendix 3
Interview guide.
[PDF File (Adobe PDF File), 226 KB - formative_v9i1e73451_app3.pdf]

References

1. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA 2016 Feb 23;315(8):801-810 [FREE Full text] [doi: 10.1001/jama.2016.0287] [Medline: 26903338]

XSL•FO
RenderX

2. Gotts J, Matthay M. Sepsis: pathophysiology and clinical management. BMJ 2016 May 23;353:i1585. [doi: 10.1136/bmj.i1585] [Medline: 27217054]
3. Rudd KE, Johnson SC, Agesa KM, Shackelford KA, Tsoi D, Kievlan DR, et al. Global, regional, and national sepsis incidence and mortality, 1990-2017: analysis for the Global Burden of Disease Study. Lancet 2020 Jan 18;395(10219):200-211 [FREE Full text] [doi: 10.1016/S0140-6736(19)32989-7] [Medline: 31954465]
4. Fleischmann C, Scherag A, Adhikari NKJ, Hartog CS, Tsaganos T, Schlattmann P, International Forum of Acute Care Trialists. Assessment of global incidence and mortality of hospital-treated sepsis. Current estimates and limitations. Am J Respir Crit Care Med 2016 Feb 01;193(3):259-272. [doi: 10.1164/rccm.201504-0781OC] [Medline: 26414292]
5. Weiss SL, Peters MJ, Alhazzani W, Agus MSD, Flori HR, Inwald DP, et al. Surviving Sepsis Campaign international guidelines for the management of septic shock and sepsis-associated organ dysfunction in children. Pediatr Crit Care Med 2020 Feb;21(2):e52-e106. [doi: 10.1097/PCC.0000000000002198] [Medline: 32032273]
6. Panchal AR, Bartos JA, Cabañas JG, Donnino MW, Drennan IR, Hirsch KG, Adult Basic and Advanced Life Support Writing Group. Part 3: adult basic and advanced life support: 2020 American Heart Association guidelines for cardiopulmonary resuscitation and emergency cardiovascular care. Circulation 2020 Oct 20;142(16_suppl_2):S366-S468 [FREE Full text] [doi: 10.1161/CIR.0000000000000916] [Medline: 33081529]
7. Rahman AE, Hossain AT, Siddique AB, Jabeen S, Chisti MJ, Dockrell DH, et al. Child mortality in Bangladesh - why, when, where and how? A national survey-based analysis. J Glob Health 2021 Sep 11;11:04052 [FREE Full text] [doi: 10.7189/jogh.11.04052] [Medline: 34552721]
8. Jimenez-Zambrano A, Ritger C, Rebull M, Wiens MO, Kabajaasi O, Jaramillo-Bustamante JC, et al. Clinical decision support tools for paediatric sepsis in resource-poor settings: an international qualitative study. BMJ Open 2023 Oct 24;13(10):e074458 [FREE Full text] [doi: 10.1136/bmjopen-2023-074458] [Medline: 37879683]
9. Bennet T. Novel pediatric sepsis criteria and clinical decision support tools. National Institutes of Health. URL: https://reporter.nih.gov/search/03LYshvMFEOAHItUQIo-0Q/project-details/10904921 [accessed 2025-09-04]
10. Sanchez-Pinto LN, Del Pilar Arias López M, Scott H, Gibbons K, Moor M, Watson RS, et al. Digital solutions in paediatric sepsis: current state, challenges, and opportunities to improve care around the world. Lancet Digit Health 2024 Sep;6(9):e651-e661 [FREE Full text] [doi: 10.1016/S2589-7500(24)00141-9] [Medline: 39138095]
11. Ramgopal S, Sanchez-Pinto LN, Horvat CM, Carroll MS, Luo Y, Florin TA. Artificial intelligence-based clinical decision support in pediatrics. Pediatr Res 2023 Jan;93(2):334-341 [FREE Full text] [doi: 10.1038/s41390-022-02226-1] [Medline: 35906317]
12. El-Kareh R, Sittig DF. Enhancing diagnosis through technology: decision support, artificial intelligence, and beyond. Crit Care Clin 2022 Jan;38(1):129-139 [FREE Full text] [doi: 10.1016/j.ccc.2021.08.004] [Medline: 34794627]
13. Vidrine R, Zackoff M, Paff Z, Seger B, Satterlee M, Buenaventura E, et al. Improving timely recognition and treatment of sepsis in the pediatric ICU. Jt Comm J Qual Patient Saf 2020 May;46(5):299-307. [doi: 10.1016/j.jcjq.2020.02.005] [Medline: 32201121]
14. Dewan M, Vidrine R, Zackoff M, Paff Z, Seger B, Pfeiffer S, et al. Design, implementation, and validation of a pediatric ICU sepsis prediction tool as clinical decision support. Appl Clin Inform 2020 Mar;11(2):218-225 [FREE Full text] [doi: 10.1055/s-0040-1705107] [Medline: 32215893]
15. de Souza DC, Paul R, Mozun R, Sankar J, Jabornisky R, Lim E, et al. Quality improvement programmes in paediatric sepsis from a global perspective. Lancet Child Adolesc Health 2024 Sep;8(9):695-706. [doi: 10.1016/S2352-4642(24)00142-1] [Medline: 39142743]
16. Neal SR, Sturrock SS, Musorowegomo D, Gannon H, Zaman M, Cortina-Borja M, et al. Clinical prediction models to diagnose neonatal sepsis in low-income and middle-income countries: a scoping review. BMJ Glob Health 2025 Apr 09;10(4):e017582 [FREE Full text] [doi: 10.1136/bmjgh-2024-017582] [Medline: 40204466]
17. Joshi M, Mecklai K, Rozenblum R, Samal L. Implementation approaches and barriers for rule-based and machine learning-based sepsis risk prediction tools: a qualitative study. JAMIA Open 2022 Jul;5(2):ooac022 [FREE Full text] [doi: 10.1093/jamiaopen/ooac022] [Medline: 35474719]
18. Odendaal WA, Anstey Watkins J, Leon N, Goudge J, Griffiths F, Tomlinson M, et al. Health workers' perceptions and experiences of using mHealth technologies to deliver primary healthcare services: a qualitative evidence synthesis. Cochrane Database Syst Rev 2020 Mar 26;3(3):CD011942 [FREE Full text] [doi: 10.1002/14651858.CD011942.pub2] [Medline: 32216074]
19. Garbern SC, Mamun GMS, Shaima SN, Hakim N, Wegerich S, Alla S, et al. A novel digital health approach to improving global pediatric sepsis care in Bangladesh using wearable technology and machine learning. PLOS Digit Health 2024 Oct 30;3(10):e0000634. [doi: 10.1371/journal.pdig.0000634] [Medline: 39475844]
20. Annual Report 2023: Solving public health problems through innovative scientific research. icddr,b. 2023. URL: https://www.icddrb.org/dmdocuments/Annual%20Report%202023.pdf [accessed 2025-09-04]
21. Annual Report 2019-2020. Institute of Child and Mother Health (ICMH). URL: https://icmh.org.bd/annual-reports/ [accessed 2025-09-04]
22. Hennink M, Kaiser BN. Sample sizes for saturation in qualitative research: a systematic review of empirical tests. Soc Sci Med 2022 Jan;292:114523 [FREE Full text] [doi: 10.1016/j.socscimed.2021.114523] [Medline: 34785096]

XSL•FO
RenderX

23. Gale NK, Heath G, Cameron E, Rashid S, Redwood S. Using the framework method for the analysis of qualitative data in multi-disciplinary health research. BMC Med Res Methodol 2013 Sep 18;13:117 [FREE Full text] [doi: 10.1186/1471-2288-13-117] [Medline: 24047204]
24. Levine AC, Gainey M, Qu K, Nasrin S, Sharif MB, Noor SS, et al. A comparison of the NIRUDAK models and WHO algorithm for dehydration assessment in older children and adults with acute diarrhoea: a prospective, observational study. Lancet Glob Health 2023 Nov;11(11):e1725-e1733 [FREE Full text] [doi: 10.1016/S2214-109X(23)00403-5] [Medline: 37776870]
25. Gao F, Liu C, Zhang L, Liu T, Wang Z, Song Z, et al. Wearable and flexible electrochemical sensors for sweat analysis: a review. Microsyst Nanoeng 2023 Jan 01;9(1):1 [FREE Full text] [doi: 10.1038/s41378-022-00443-6] [Medline: 36597511]

Abbreviations

CDS: clinical decision support
CDSS: clinical decision support system
HCP: health care provider
icddr,b: International Centre for Diarrhoeal Disease Research, Bangladesh
ICMH: Institute of Child and Mother Health
IRB: Institutional Review Board
KGH: Kushtia General Hospital
LMIC: low- and middle-income country
mHealth: mobile health

Edited by A Mavragani; submitted 13.03.25; peer-reviewed by P Usuemerai, C Ogbeta; comments to author 10.04.25; revised version received 21.08.25; accepted 28.08.25; published 26.09.25.

Please cite as:
Shaima SN, Genisca AE, Faruk MT, Uddin MF, Saha A, Sultana N, Kadakia N, Gainey M, Kim E, Shaw K, Afroze F, Chepngeno J, Jindal A, Chowdhury SA, Chisti MJ, Levine AC, Garbern SC
Health Care Providers' Perspectives of Clinical Decision Support Tools for Pediatric Sepsis in Bangladesh: Qualitative Study
JMIR Form Res 2025;9:e73451
URL: https://formative.jmir.org/2025/1/e73451
doi:10.2196/73451
PMID:41004258

©Shamsun N Shaima, Alicia E Genisca, Md Tanveer Faruk, Md Fakhar Uddin, Akash Saha, Nadia Sultana, Nidhi Kadakia, Monique Gainey, Elleen Kim, Kikuyo Shaw, Farzana Afroze, Joan Chepngeno, Atin Jindal, Sifat A Chowdhury, Md Jobayer Chisti, Adam C Levine, Stephanie C Garbern. Originally published in JMIR Formative Research (https://formative.jmir.org), 26.09.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Evaluating a Website on Learning Disorders for Parents and Learning Therapists: Observational Mixed Methods Study

Olga Hermansson[1], MSc; Paula Dümig[1], MA; Björn Witzel[1], MSc; Lior Weinreich[1], MSc; Susanne Volkmer[1], Dr, Dipl Psych; Gerd Schulte-Körne[1], Prof Dr; Kristina Moll[1], PD Dr

Department of Child and Adolescent Psychiatry, University Hospital, Ludwig-Maximilians-University Munich, Munich, Germany

Corresponding Author:
Olga Hermansson, MSc
Department of Child and Adolescent Psychiatry
University Hospital
Ludwig-Maximilians-University Munich
Nussbaumstrasse 5
Munich, 80336
Germany
Phone: 49 4400 56915
Email: olha.drobina@med.uni-muenchen.de

Abstract

Background: Between 5% and 15% of children worldwide have a specific learning disorder. This creates a substantial demand for information among both parents and professionals. LONDI (londi.de) is a German-language website that (1) provides evidence-based information on learning disorders and (2) offers a tool to search for relevant diagnostics and intervention measures for professionals (Help System). This paper reports the results of an evaluation study of the website.

Objective: The aim of the study was to (1) evaluate the website and (2) inform existing theories on technology acceptance and user satisfaction. The study was conducted using the RE-AIM framework for evaluating public health impact and the information system continuance intention framework.

Methods: This mixed methods observational study was conducted online from February 2023 to August 2023 in Germany. Parents of children with learning difficulties and learning therapists participated in a 1.5-hour online session in which they were guided through the website. A pre-post design was used to assess changes in participants' knowledge, attitudes, and self-efficacy. Additionally, two path models assessing the predicting factors of the intention to further use (ie, continuance intention) and the intention to recommend the website were tested. The first model was based on the widely used information system continuance intention framework and tested expectations confirmation, perceived usefulness, and website satisfaction as predictors of the continuance intention and the intention to recommend the website. The second model focused on usability, content perception, visual aesthetics, and satisfaction as predictors of the same outcome variables.

Results: A total of 77 parents and 73 learning therapists participated in the study. In the sample of learning therapists, age correlated negatively with usability opinion and website satisfaction. A 2-tailed *t* test revealed a significant increase in knowledge about learning disorders in both groups (parents: t_{76}=12.02, P<.001; learning therapists: t_{71}=7.03, P<.001). There was no change in attitudes and self-efficacy in parents ($F_{1,76}$=2.04, P=.14; Wilks lambda=0.95), but there was a significant change for learning therapists ($F_{1,68}$=15.83, P<.001; Wilks lambda=0.68) after using the website. A path analysis revealed that the intention to recommend the website can be included as an additional variable in the information system continuance intention framework. For the informational pages, content perception and visual aesthetics significantly predicted website satisfaction (R^2=0.59, $F_{3,143}$=69.06, P<.001), and content perception significantly predicted continuance intention (R^2=0.45, $F_{3,143}$=39.74, P<.001). For the Help System, usability was the only significant predictor of website satisfaction (R^2=0.45, $F_{2,67}$=28.16, P<.001), continuance intention (R^2=0.34, $F_{1,68}$=34.57, P<.001), and intention to recommend (R^2=0.21, $F_{1,68}$=19.02, P<.001).

Conclusions: The website has been evaluated positively and has proven useful for the target audience. Predictors of website acceptance and further use are contextual and depend on the website type.

(JMIR Form Res 2025;9:e68365) doi:10.2196/68365

XSL•FO
RenderX

KEYWORDS

website evaluation; public health education; internet research; RE-AIM framework; information system continuance intention

Introduction

Background

Specific learning disorder is defined as persistent impairment in at least one of the three learning domains: reading, spelling, and mathematics [1]. They are often referred to as dyslexia or dyscalculia, depending on the affected skill area. People affected by dyslexia typically have problems with reading accuracy and fluency and with spelling; people affected by dyscalculia have problems with understanding and processing numerical information, resulting in problems with memorizing arithmetic facts and with calculations [1,2]. Learning disorders are not due to other conditions, such as intellectual disability, neurological conditions, or lack of instruction [3]. Between 5% and 15% of children worldwide have a specific learning disorder, according to the Diagnostic and Statistical Manual of Mental Disorders (Fifth Edition; DSM-5) [1]. However, the prevalence of undiagnosed learning difficulties in at least one of the three learning domains (reading, spelling, and mathematics) is estimated to be much higher (eg, 23% in Germany and approximately 20% in the United States) [4,5]. The consequences of learning disorders are worrisome: Academic failure, psychiatric disorders, and difficulty participating in society are some of the many risks that such children face in later life [6-8].

Consequently, a question arises: Where do caretakers and professionals turn for reliable, evidence-based information about learning disorders? A recent study surveying parents of children with learning disorders in Germany revealed that only 54% of all parents receive sufficient psychoeducation and continuous support from learning therapists or educational or medical professionals following their child's diagnostic procedure [9]. Where professionals are concerned, difficulties include choosing the suitable diagnostic measure and subsequently finding the best intervention program based on the deficit profile of the child. Parents and mental health professionals often turn to the internet in search of medical information about children's mental health [10]. Naturally, there is a need to determine the goodness criteria for such websites and evaluate them to ensure that users get good-quality information and support.

Aiming to fill these gaps, the online platform LONDI was created by psychology and education professionals at the Clinic of the Ludwig-Maximilians-University of Munich and German Institute for International Pedagogical Research of Frankfurt [11]. LONDI offers user-specific, evidence-based information on learning disorders for different stakeholders: parents of affected children, learning therapists, school psychologists, teachers, and youth welfare workers. Apart from the information on learning disorders, it provides a unique tool for professionals (ie, learning therapists, school psychologists, and teachers) called the LONDI Help System (further referred to as HS). HS provides recommendations for evidence-based diagnostic tools and intervention programs based on a child's specific needs.

Previous studies evaluating websites have mainly focused on the website's content (eg, [12]) while often overlooking the website's usage and effectiveness and whether usage is maintained over time. This study addresses this issue by adopting the RE-AIM framework for website evaluation. This framework is frequently applied in public health research for evaluating design, dissemination, and implementation of health intervention programs [12-16]. The abbreviation "RE-AIM" stands for Reach (ie, the number, proportion, and representativeness of people that the initiative reaches), Effectiveness (ie, the impact of the initiative on public health), Adoption (ie, the number and proportion of agents who are willing to use the initiative, the delivery conditions), Implementation (ie, the percentage of process objectives that were achieved and whether the intervention has been delivered as intended), and Maintenance (ie, how well the initiative is sustained over time).

A website's *Reach* in intervention evaluations is typically assessed by the number and characteristics (eg, age) of the participants who take part in the intervention [17]. For instance, older participants rate visual aesthetics, satisfaction, perceived usefulness, and expectation confirmation more positive than younger participants, whereas there seems to be no age effect for usability [18,19]. The *Effectiveness* of a website is operationalized as any outcome intended to be changed as a result of the intervention [20]. *Implementation* is measured by how well the process objectives of the intervention are achieved and focuses on the important milestones in implementation delivery [20]. For an informational website, such a milestone is the change in the participant's knowledge, attitudes, and self-efficacy [21,22]. *Maintenance* refers to website usage over time [20]. This can be measured by assessing users' intention to further use a website (ie, continuance intention) and to recommend it. Building on the Technology Acceptance Model [23-25], the Information System Continuance framework [26,27], and previous evaluation studies [28-32], it is assumed that perceived usefulness and the confirmation of user expectations predict user satisfaction and continuance intention (ie, the intention to further use a website). Moreover, it is assumed that the same variables also predict the intention to recommend a website to others [33-35]. However, studies directly testing this assumption by including both outcomes (continuance intention and intention to recommend) in one model are currently missing.

In addition, previous research assessed the influence of *website characteristics* and found that content perception, usability, and visual aesthetics predict continuance intention and intention to recommend, either directly [32,35] or via satisfaction [19].

This study's aim was to evaluate the LONDI website for two of the website's target groups (ie, parents and learning therapists) based on 4 of the 5 RE-AIM dimensions (ie, *Reach*, *Effectiveness*, *Implementation*, and *Maintenance*). The *Adoption* dimension was not addressed, as the usage of the website by the participants was strictly guided. Although parents evaluated

XSL•FO
RenderX

the informational pages developed for parents, learning therapists evaluated the informational pages for learning therapists and the HS for diagnostics and for interventions. Moreover, we aimed to create two path models to assess the factors influencing users' intention to further use (ie, continuance intention) and the intention to recommend the website.

Research Questions and Hypotheses

The research questions for the evaluation study were formulated according to the following RE-AIM dimensions: *Reach*, *Effectiveness* and *Implementation*, and *Maintenance*.

Reach

What is the relationship between the characteristics of the users (age, experience) and their opinions about the website characteristics (usability, content, visual aesthetics, perceived usefulness, expectations confirmation, satisfaction, continuance intention)?

Hypothesis 1a (H1a) was that opinions about the website (content, visual aesthetic, expectations confirmation, perceived usefulness, satisfaction, continuance intention) improve with age and experience for both parents and learning therapists.

Hypothesis 1b (H1b) was that opinions about the website's usability will not depend on age or experience of the participants.

Effectiveness and Implementation

How does the use of the platform affect the knowledge, attitude, and self-efficacy of the target audiences?

Hypothesis 2 (H2) was that the knowledge, attitudes, and self-efficacy of parents and learning therapists will increase as a result of using the website.

Maintenance

How do the opinions of the target groups about the website characteristics influence their satisfaction, continuance intention, and intention to recommend the website?

To address this research question, two path models were developed and tested. In the Statistical Analysis subsection, the hypothesized and tested models (H3 and H4a and 4b) are provided, along with corresponding figures.

Additional Data

Additionally, participants' website ratings were compared with validated benchmarks, and open-ended qualitative data were gathered to complement our quantitative findings [36] (see the Statistical Analysis subsection for a detailed description).

Registration

This evaluation study was preregistered on Open Science Framework [37,38].

Methods

Study Design

The study is a mixed methods observational study with pre- and postuse comparisons. Participants were guided by our research assistants through the website in a single 1.5-hour online session conducted via the redConnect tool. They were asked to complete specific tasks, such as read information passages or insert diagnostic data into the HS to get recommendations for intervention programs. The sessions were conducted by the research assistant staff of the university clinic who received relevant training from the study authors.

Before the session, participants completed online questionnaires about their demographic data (eg, age range, country of origin, profession; see the Measures section) and their knowledge, attitudes, and self-efficacy regarding learning disorders. During the session, parents and learning therapists familiarized themselves with the contents of the informational pages developed for parents and learning therapists, respectively. In addition to the informational pages, therapists also used both HS and were asked to find diagnostic and intervention tools based on a given case report.

After the session, participants answered the same questionnaires as before the session on their knowledge, attitudes, and self-efficacy. In addition, they completed questionnaires on the website's usability, visual aesthetic, content perception, user satisfaction, perceived usefulness, expectations confirmation, intention to use, and intention to recommend the website.

Finally, participants also responded to an open-ended question: "What did you like about the page in particular and what would you like to improve?" This question was asked separately for the information pages, HS for diagnostics, and HS for interventions.

Eligibility and Recruitment

Recruitment for both target groups (therapists and parents) was conducted through parent and learning therapist associations in Germany, Austria, and Switzerland. These associations typically serve as the first point of contact for parents of children with learning problems to acquire information about their child's condition. They also function as accrediting institutions for learning therapists and provide professional training and certification. Thus, the members of these associations represent both target audiences for this study. For recruitment, we contacted the associations and asked them to send electronic study flyers to their members via their mailing lists. Flyers included information about the study aims and procedures as well as our contact email address. Participants who were interested in taking part were asked to send us an email in order to receive detailed study information and an informed consent form, which they were asked to sign and return. Once the written informed consent was received, the participants were contacted to schedule an online videocall session with one of our testers (see the Data Collection subsection for more details).

The inclusion criteria were as follows: Parents had to be a parent, a guardian, or otherwise closely participating in the upbringing of a child who either (1) had an official diagnosis of a learning disorder or (2) was experiencing severe difficulties in reading, spelling, or mathematics according to the school and the parents. Learning therapists had to be accredited in Germany, Austria, or Switzerland or be in the last semester of their respective accreditation course. The language of the website

XSL•FO
RenderX

and the study was German; thus, a common inclusion criterion for both groups was to be a native or advanced German speaker.

Ethical Considerations

The study was approved by the Ethics Committee of Ludwig-Maximilians University (LMU) of Munich on May 3, 2022 (project number 22-0300; valid until May 2, 2027). The participants signed primary informed consent forms, which included the description of the study design, its scientific basis, the study objective, conditions for participation, compensation, and information about personal data protection procedures according to the General Data Protection Regulation (GDPR) regulations. The consent form contained information about the length of the online survey that they would need to complete during the guided session. All participants received a voucher of €20 (US $23.28) upon completion of the study.

The study design was evaluated and approved by the LMU Data Protection Officer for its compliance with GDPR. The data protection measures adopted to ensure data privacy included anonymization of the data and use of GDPR-approved software for data collection (SoSci Survey, redConnect) and approved hardware for data storage (only local network computers at the LMU University Clinic).

Measures

Pre-Use

The following was assessed during the pre-use evaluation: (1) demographic data, (2) knowledge about learning disorders (the number of correct answers on a true/false test), (3) attitudes toward learning disorders, and (4) self-efficacy when dealing with learning disorders.

Demographic Data

For parents, the following information was collected: age, gender, country of residence, diagnosis of the child (or affected learning domain), and years of experience dealing with learning disorders.

For learning therapists, the following information was collected: age, gender, country of residence, and years of professional experience.

Knowledge About Learning Disorders

Knowledge increase is operationalized as the increase in the number of correct answers in the knowledge test about learning disorders. The questionnaire includes 25 statement questions for each target group (parents and therapists) with the answer options “Correct,” “Incorrect,” and “Not sure.” This questionnaire was developed by authors PD and OH based on the website’s content.

Attitudes Toward Learning Disorders

Attitudes are defined by Ajzen and Fishbein [39] as a valent assessment of objects, concepts, or actions. According to the authors, valent assessments of subjective norms, valent assessments of behaviors, and subjective control over a certain action can all predict behavior. For our evaluation, we adapted an existing questionnaire on the attitudes of teachers toward learning disorders [40]. The questionnaire contains 24 questions for parents and 22 questions for therapists. Questions are assessed on a 6-point Likert scale. The Cronbach α reliability for the parental questionnaire based on this study’s sample was 0.71 for both the pre and posttest. For learning therapists, the Cronbach α was 0.77 for the pretest and 0.81 for the posttest. The calculation was performed using SPSS version 29 (IBM Corp).

Self-Efficacy Dealing With Learning Disorders

Self-efficacy change is defined as an increase in the belief that an individual can successfully execute an activity [41]. We adapted the questionnaire from Höltge et al [42]. The self-efficacy questionnaire for parents includes subscales on self-efficacy for supporting the child with academic success or school, self-efficacy for supporting the child in social life, self-efficacy for teaching the child everyday skills, and self-efficacy for emotional support. The subscales for learning therapists include their self-efficacy for reading therapy, spelling therapy, and mathematics therapy. The questionnaire for parents contains 22 questions measured on a 6-point Likert scale, while the questionnaire for learning therapists contains 36 questions. The Cronbach α for the parental questionnaire was 0.94 for the pretest and 0.93 for the posttest. The Cronbach α for the learning therapists’ questionnaire was 0.93 for the pretest and 0.93 for the posttest.

Postuse

The following was assessed during the postuse evaluation: (1) knowledge about learning disorders (the number of correct answers on a true/false test), (2) attitudes toward learning disorders, (3) self-efficacy for dealing with learning disorders, (4) usability, (5) visual aesthetics, (6) content perception (only for the informational pages’ evaluation), (7) perceived usefulness, (8) expectations confirmation, (9) satisfaction with the website, (10) continuance intention, and (11) intention to recommend the website.

Questionnaires 1-3 for knowledge, attitudes, and self-efficacy were the same as those applied before website use (see the descriptions in the Pre-Use subsection).

Usability

The international standard for computer systems ISO 9241-11 defines usability as the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use. In this study, this was measured using the German-language adaptation of the System Usability Scale [43-45], a unidimensional questionnaire that contains 10 questions assessed on an 8-point Likert scale. Reliability of the questionnaire has been reported to range between a Cronbach α of 0.85 and 0.91 [44,45].

Visual Aesthetics

Visual aesthetics is a measure of the aesthetic appeal as a pleasant feeling that comes before reasoning about an object [19]. A frequently used measure is the VisAWI questionnaire, developed and validated by Thielsch and Salaschek [19,45]. In this study, the full version of the questionnaire was used, with 18 questions assessed on an 8-point Likert scale: 5 questions in the Simplicity subscale, 5 questions in the Versatility subscale,

XSL•FO
RenderX

4 questions in the Colorfulness subscale, and 4 questions in the Craftsmanship subscale. The Cronbach α reliability of the questionnaire ranges between 0.76 and 0.81 [45].

Content Perception

Content is defined in the draft of DIN EN ISO 9241-151 as a compilation of all content objects of a web user interface [43]. We used the questionnaire developed by Theilsch and Salaschek [45] that measures 3 aspects of content perception: liking, comprehensibility, and quality/benefit. Each of these aspects is represented by 3 questions assessed on an 8-point Likert scale (9 questions in total). The reported Cronbach α reliability ranges between 0.71 and 0.90 [45].

Perceived Usefulness, Expectations Confirmation, Satisfaction With the Website, and Continuance Intention

The perceived usefulness of an information system is defined as the cognitive belief about how useful a system is for a user. Expectations confirmation is a measure of whether the initial expectations of using an information system have been met. Satisfaction is an affective state resulting from the experience of using a website. Continuance intention describes the intention to use an information system beyond an initial acceptance phase, in which a consumer successfully decides to start using a product [26,27]. These variables are part of the factors measured by the Information Systems Continuance Framework in the respective questionnaire [26]. The German version of the questionnaire used in this study was adapted and validated by Budner et al [46]. The questionnaire includes 13 questions assessed on an 8-point Likert-scale. The Cronbach α reliabilities for the measures are as follows: 0.92 for perceived usefulness, 0.77 for expectations confirmation, 0.92 for satisfaction, and 0.85 for continuance intention [46].

Intention to Recommend the Website

Intention to recommend is measured using a single-item questionnaire that is a standard for the industry: the Net Promoter Score (NPS) [47]. It measures a customer's loyalty, which is defined as the willingness of a user to make an investment or personal sacrifice in the form of a recommendation. In our case, this is the intention of the participant (parent or therapist) to recommend the website. It is represented by one question assessed on an 11-point Likert scale (from 0 to 10).

The full-text questionnaires for all measures are available in Multimedia Appendix 1.

Data Collection

Data were collected online using the online survey tool SoSci Survey. All questionnaires were programmed into the software. Participants were provided with a link to access the questionnaires and were required to enter their individual participant number to ensure data anonymity. All questionnaires were completed during the guided online session with the tester. The responses were saved automatically by the software on its local server in Germany and later downloaded for analysis.

Statistical Analysis

To answer the REACH research questions ("What is the relationship between the characteristics of the users (age, experience) and their opinions about the website characteristics (usability, content, visual aesthetics, perceived usefulness, expectations confirmation, satisfaction, continuance intention)?"), Pearson correlations were calculated to analyze relationships between the users' age and experience dealing with learning disorders and their perceived usefulness of the website, website satisfaction, and continuance intention. The software used for the analysis was R version 4.4.0, and the function used was rcorr. For *P* value correction with the Bonferroni-Holm method, we used the function p.adjust (method "holm").

Additionally, we descriptively compared participants' ratings of all website characteristics with published benchmark scores reported in the original validation studies of the respective questionnaires. This comparison was not subjected to inference statistic testing but served as an indication whether the observed values were higher or lower than established normative values.

To answer the Effectiveness and Implementation research question ("How does the use of the platform affect the knowledge, attitude, and self-efficacy of the target audiences?"), a 2-tailed paired-samples *t* test was conducted to assess the knowledge change from pre- to postuse separately for parents and learning therapists. Since attitudes and self-efficacy are reported to be linked [21,22], these 2 outcomes were analyzed in one analysis using a repeated-measures multivariate analysis of variance (MANOVA), followed by post hoc *t* tests. The analysis was done using SPSS version 29 (IBM Corp).

Finally, to answer the Maintenance research question ("How do the opinions of the target groups about the website characteristics influence their satisfaction, continuance intention, and intention to recommend the website?"), path analyses were conducted for 3 hypothesized models. These analyses were done using R version 4.4.0 with functions sem and modindices for index modification.

The first hypothesized model was based on the information system continuance intention framework [26,27], with the modification of adding the intention to recommend the website as a second outcome variable in addition to continuance intention. H3 tested whether expectations confirmation predicted satisfaction and perceived usefulness; perceived usefulness predicted satisfaction, continuance intention, and intention to recommend; and satisfaction predicted continuance intention and intention to recommend (see Figure 1). This model was tested 3 times: once for the informational pages using the combined data from parents and learning therapists, once for HS for diagnostics, and once for HS for interventions (HS based on the learning therapists).

The second hypothesized model tested how website characteristics (content perception, usability, and visual aesthetics) predicted the continuance intention and the intention to recommend the website. Two separate models were developed: one for the informational pages and one for HS. This was done due to the varying purposes of these two parts of the website. The informational pages focus primarily on content, whereas HS is a search tool that provides links to other resources in the form of search results. Therefore, the model for the informational pages included content perception as a

XSL•FO
RenderX

predictor variable, which was not part of the model for HS (see Figures 2 and 3). In order to have acceptable power for the models [48], for every path in the model, there needed to be 10 participants in the sample. Therefore, we kept the number of hypothesized paths at 7 or 8 per model due to our sample size. For the informational pages, we hypothesized (H4a) that content perception, usability, and visual aesthetics predicted continuance intention both directly and indirectly through satisfaction (Figure 2).

For the HS model, we hypothesized (H4b) that usability and visual aesthetic predicted continuance intention and intention to recommend both directly and through satisfaction (Figure 3). We tested this model 2 times: once for HS for diagnostics and once for HS for interventions.

Due to technical errors during data collection, data were missing for 3 learning therapists for the tests of attitudes and self-efficacy change (H2) and predictors of website usage (H4a and 4b), resulting in varying degrees of freedom in the reported analyses.

In addition to the quantitative analysis, a qualitative content analysis was performed on the text input received from the open-ended questions. The aim of the content analysis was to complement and explain the quantitative assessment. For the qualitative analysis, a top-down approach was adopted, derived from existing theories on website assessment [23-35] and the chosen quantitative measures [40,42,45,46]. Thus, a coding scheme was developed in advance.

The analysis consisted of the following steps. The first step involved dividing the written feedback from the participants into the smallest communicative units, each with one clear message or statement. For instance, a complete answer ("The page is well-composed...a great introductory text for beginner learning therapists...You could add some links to local websites...Fonts could be larger.") would be divided into 4 statements: (1) "The page is well-composed..." (2) "a great introductory text for beginner learning therapists..." (3) "You could add some links to local websites..." and (4) "Fonts could be larger."

Each statement was then coded along 3 dimensions: (1) positive, negative, or neutral; (2) existing feature or new feature request; and (3) correspondence to one of the variables used in the quantitative evaluation. This way, the previous response would be coded as follows:

1. "The page is well-composed..."—positive, existing feature, usability
2. "a great introductory text for beginner learning therapists..."—positive, existing feature, content
3. "You could add some links to local websites..."—neutral, new feature request, content
4. "Fonts could be larger"—negative, existing feature, content

Multimedia Appendix 2 contains a summary of all categories developed during coding.

Figure 1. Hypothesized path model (hypothesis 3) based on the information system continuance intention framework to test the predictors of continuance intention and intention to recommend the website.

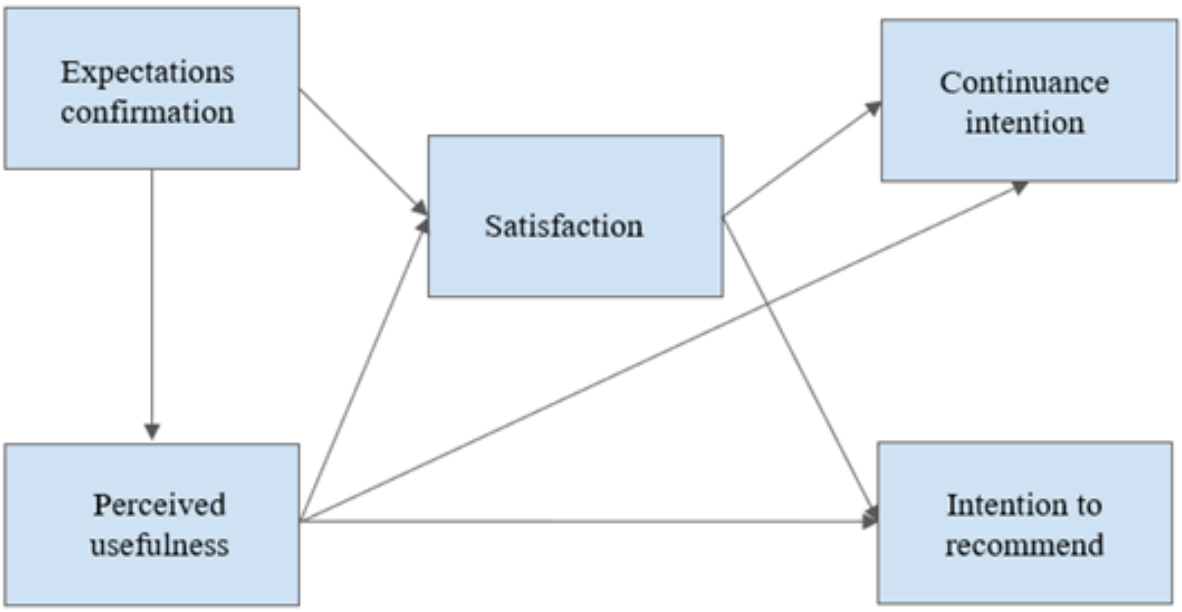

XSL•FO
RenderX

Figure 2. Hypothesized path model (hypothesis 4a) for the informational pages to test the predictors of satisfaction and continuance intention of parents and learning therapists.

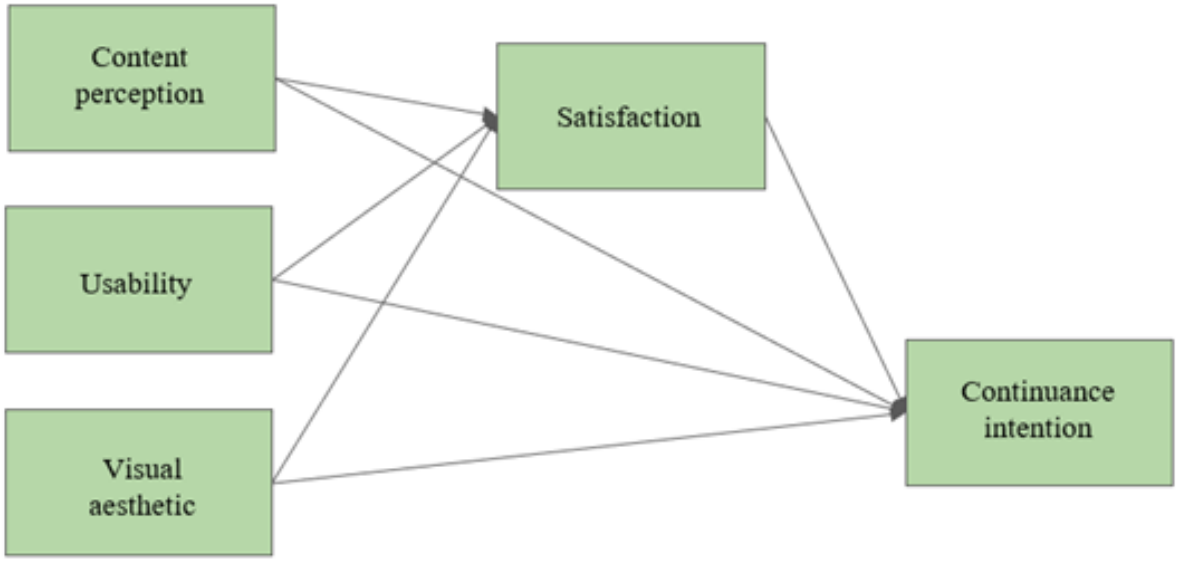

Figure 3. Hypothesized path model (hypothesis 4b) for the Help System (HS) tool to test the predictors of continuance intention and intention to recommend the website by learning therapists.

Results

Participant Statistics

The sample consisted of 150 participants: 77 parents and 73 learning therapists (see Table 1 for detailed demographic data).

XSL•FO
RenderX

Table 1. Demographic data (gender, country of residence, age, and experience with learning disorders for the 2 target groups: parents and learning therapists).

Demographic data	Parents (n=77), n (%)	Learning therapists (n=73), n (%)
Gender		
Female	73 (94)	68 (93)
Country of residence		
Germany	73 (94)	58 (79)
Austria	3 (3)	15 (20)
Italy	1 (1)	0 (0)
Age (years)		
25-34	0 (0)	8 (10)
35-44	29 (39)	18 (24)
45-54	42 (54)	29 (39)
55-64	5 (6)	12 (16)
≥65	1 (1)	6 (8)
Experience with learning disorders (years)		
Beginner	7 (9)	10 (13)
1-5 years	48 (62)	18 (24)
6-10 years	18 (23)	19 (26)
11-20 years	2 (3)	21 (28)
>20 years	2 (3)	5 (6)

Evaluation Outcomes

Benchmark Comparisons

For the informational pages, the raw scores for the website characteristics (ie, visual aesthetics, content perception, and usability) were higher than the benchmark or reference values for similar websites in German in both samples (see Table 2) [45]. For perceived usefulness, expectations confirmation, website satisfaction, and continuance intention, there are no benchmarks available; therefore, we used the results from the study by Budner et al [46] that adapted the questionnaire into German as reference values. For parents, the mean scores for perceived usefulness, expectations confirmation, website satisfaction, and continuance intention were all higher than the corresponding reference values reported by Budner et al [46]. For learning therapists, the mean scores for perceived usefulness and expectations confirmation were lower than the reference values, while satisfaction and continuance intention were higher than the reference values.

Table 3 shows the comparison between learning therapists' opinions about HS and the same reference values taken from Budner et al [46] and Thielsch and Moshagen [45]. The usability scores for the HS for diagnostics and interventions were within the benchmark range; perceived usefulness and satisfaction scores exceeded the benchmarks, while expectations confirmation and continuance intention were slightly lower than the benchmarks.

XSL•FO
RenderX

Table 2. Evaluation variables for the LONDI informational pages by participant type, in comparison with available benchmarks from other studies.

Variable (raw score range)	Parents (n=77), mean (SD)	Learning therapists (n=73), mean (SD)	Benchmark or reference value, mean (SD)
Visual aesthetic (1-7)	5.14 (1.22)	5.17 (1.01)	4.08 (1.26) [45]
Content perception (1-7)	6.12 (0.97)	6.16 (0.82)	3.69 (1.17) [45]
Usability (1-100)	82.85 (17.79)	86.52 (11.73)	80-90[a] [45]
Perceived usefulness (1-7)	5.84 (1.16)	5.03 (1.71)	5.22 (1.41) [46]
Expectations confirmation (1-7)	5.44 (1.44)	4.91 (1.51)	5.14 (1.09) [46]
Satisfaction (1-7)	5.94 (1.05)	5.94 (0.92)	5.56 (1.12) [46]
Intention to use (continuance intention)	5.87 (1.24)	5.74 (1.11)	5.50 (1.46) [46]
Intention to recommend (0-10)	9.12 (1.72)	8.86 (1.37)	N/A[b]

[a]Range 80-90 is referenced as "Good."
[b]N/A: not available.

Table 3. Evaluation variables for the LONDI Help System in comparison with available benchmarks from other studies.

Variable (raw score range)	Help System for diagnostics, mean (SD)	Help System for interventions, mean (SD)	Benchmark or reference value, mean (SD)
Usability (1-100)	54.02 (13.40)	53.91 (12.61)	50-80[a] [45]
Perceived usefulness (1-7)	5.24 (1.52)	5.29 (1.64)	5.22 (1.41) [46]
Expectations confirmation (1-7)	5.08 (1.67)	5.02 (1.73)	5.14 (1.09) [46]
Satisfaction (1-7)	5.84 (1.08)	5.60 (1.43)	5.56 (1.12) [46]
Intention to use (continuance intention)	5.34 (1.43)	5.40 (1.44)	5.50 (1.46) [46]
Intention to recommend (0-10)	8.29 (2.07)	8.10 (2.25)	N/A[b]

[a]Range 50-80 is referenced as "Acceptable."
[b]N/A: not available.

Relationship Between User Demographics and Website Opinions (Reach Dimension: H1)

The relationships between user demographics and website opinions were investigated using Pearson correlation matrices. Significant *P* values were corrected for multiple comparisons. For the informational pages, no significant relationship was found between parents' age or experience and their opinions about the website, whereas the age (but not experience) of learning therapists correlated negatively with their satisfaction with the informational pages (Table 4).

For the HS diagnostics, there was a low negative correlation between the learning therapists' age and their opinions on the tool's usability (see Table 5), suggesting that the older the therapists were, the more negative were their opinions on the tool's usability. Similarly, for the HS for interventions, the opinions about the tool's usability and satisfaction correlated negatively with learning therapists' age.

XSL•FO
RenderX

Table 4. Correlations exploring the relationship between age and opinions about the website of parents (n=77) and learning therapists (n=70) for the LONDI informational pages.

Variable	Visual aesthetic	Usability	Content perception	Satisfaction	Perceived usefulness	Expectations confirmation	Continuance intention
Parents							
Age							
r	–0.06	–0.00	–0.05	–0.12	–0.21	–0.13	–0.21
P value	.57	.97	.68	.31	.07	.24	.06
Experience							
r	0.11	0.12	0.06	0.00	0.09	0.08	0.01
P value	.35	.28	.59	.99	.45	.47	.95
Learning therapists							
Age							
r	–0.29	–0.23	–0.15	–0.32[a]	–0.17	–0.12	–0.22
P value	.01	.05	.21	.006	.14	.30	.05
Adjusted *P* value	__[b]	—	—	.04	—	—	—
Experience							
r	–0.03	–0.04	–0.12	–0.12	–0.05	–0.02	–0.04
P value	.77	.75	.30	.32	.69	.85	.76

[a]The correlation was significant at a significance level of .05 (2-tailed) for the adjusted *P* value.

[b]Not applicable.

XSL•FO
RenderX

Table 5. Correlations exploring the relationship between age and website opinions of learning therapists for the Help System (HS) for diagnostics (N=70).

Variable	Usability	Perceived usefulness	Expectations confirmation	Satisfaction	Continuance intention
HS for diagnostics					
Age					
r	–0.34[a]	–0.05	–0.18	–0.22	–0.15
P value	.004	.67	.14	.06	.22
Adjusted *P* value	.02	__[b]	—	—	—
Experience					
r	–0.15	–0.26[c]	–0.30[c]	–0.17	–0.16
P value	.20	.02	.01	.16	.17
Adjusted *P* value	—	.08	.05	—	—
HS for interventions					
Age					
r	–0.37[a]	–0.13	–0.24	–0.34[a]	–0.25
P value	.001	.27	.049	.003	.04
Corrected *P* value	.005	—	—	.01	—
Experience					
r	–0.20	–0.32	–0.30	–0.23	–0.20
P value	.09	.006	.01	.055	.09
Corrected *P* value	.27	.06	.08	.24	.27

[a]The correlation was significant at a significance level of .05 (2-tailed) for the adjusted *P* value.

[b]Not applicable.

[c]The correlation was not significant at a significance level of .05 (2-tailed) for the adjusted *P* value.

Knowledge, Attitudes, and Self-Efficacy Change (Effectiveness and Implementation Dimensions: H2)

The comparisons between pre- and posttest performance on the knowledge test about learning disorders using paired-samples *t* tests revealed a significant increase in correct responses in both user groups (parents and learning therapists). For parents, the mean number of correct responses increased from 14.35 (SD 3.47) to 19.06 (SD 2.68; t_{76}=12.02; Cohen *d*=1.37, 95% CI 1.05-1.67; *P*<.001). For learning therapists, the mean number of correct responses increased from 21.26 (SD 2.61) to 23.03 (SD 1.61; t_{71}=7.03; Cohen *d*=0.83, 95% CI 0.56-1.1; *P*<.001).

The MANOVA comparing the changes in both attitudes and self-efficacy pre- and postuse revealed no significant difference between these combined variables for parents ($F_{1,76}$=2.04, *P*=.14; Wilks lambda=0.95) but a significant change for learning therapists ($F_{1,68}$=15.83, *P*<.001; Wilks lambda=0.68). Post hoc tests showed that the average score of learning therapists' attitudes on a scale from 1 to 7 increased from 4.51 (SD 0.43) to 4.66 (SD 0.48; t_{69}=4.81; Cohen *d*=0.27, 95% CI 0.32-0.83; *P*<.001). The average score of the learning therapists' self-efficacy on a scale from 1 to 7 increased from 4.98 (SD 0.71) to 5.18 (SD 0.65; t_{69}=4.22; Cohen *d*=0.39, 95% CI 0.25-0.75; *P*<.001).

Predictors of Website Usage (Maintenance Dimension: H3, H4a, H4b)

Continuance Intention Framework

We conducted a path analysis to test the hypothesized model (H3; Figure 1) and to examine whether the information system continuance intention framework would hold with the added outcome variable (intention to recommend). The values were centered before each analysis to correct for multicollinearity using the mutate_all function in R.

For the informational pages, the analysis based on the combined sample (parents and therapists) yielded acceptable indices of fit [17,48-52]: χ^2_2=3.1, *P*=.21; comparative fit index (CFI)=0.99, standardized root mean squared residual (SRMR)=0.02, root mean square error of approximation (RMSEA)=0.06 (see Figure 4 for the updated confirmed path model).

For the HS for diagnostics, the initial model fit yielded a significant chi-square value (*P*=.02) and acceptable indices of fit: χ^2_1=3.06, *P*=.22; CFI=0.99, SRMR=0.02, RMSEA=0.09. For the HS for interventions, the initial model fit yielded a significant chi-square value (*P*=.04). The modification indices procedure suggested adding a direct path from expectations confirmation to continuance intention. The resulting model

XSL•FO
RenderX

yielded acceptable indices of fit: χ^2_1=1.6, *P*=.20; CFI=0.99, SRMR=0.01, RMSEA=0.09 (see Figure 5 for the updated confirmed path model).

Figure 4. Confirmed path model (hypothesis 3) for the informational pages based on the adapted information system continuance intention framework with intention to recommend as an additional outcome variable (LONDI informational pages, N=147). All *P*<.001.

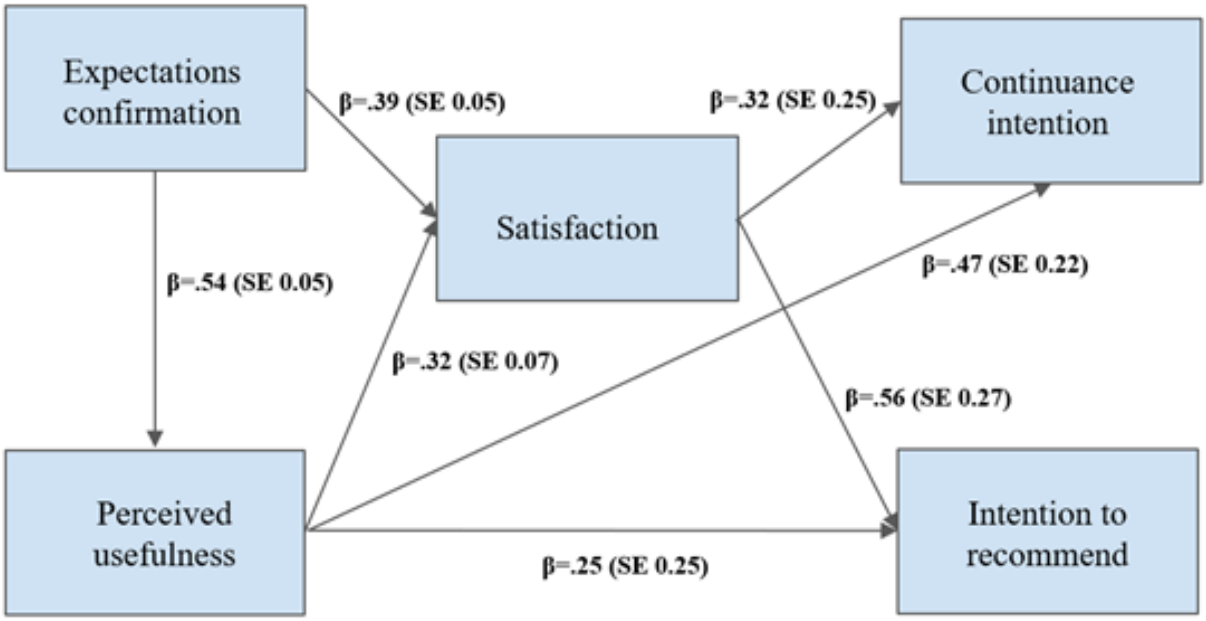

Figure 5. Confirmed path model for the LONDI Help System (N=70) for diagnostics (solid line) and intervention (blue dashed line) based on the adapted information system continuance intention framework with intention to recommend as an additional outcome variable. **P*<.001; ***P*=.01; ***all other *P* values: β=.27 (SE 0.10), *P*=.007, β=.27 (SE 0.13), *P*=.04), β=.21 (SE 0.10), *P*=.03), β=.15 (SE 0.10), *P*=.13.

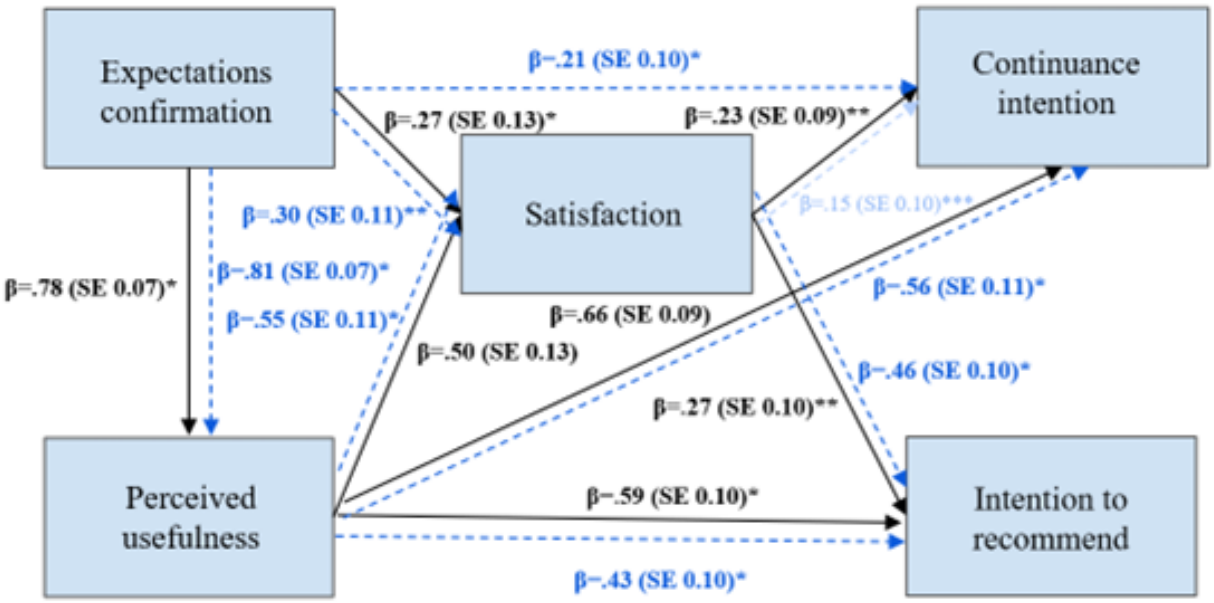

Website Characteristics

The second hypothesized model (H4a: Figure 2; H4b: Figure 3) for the prediction of continuance intention and the intention to recommend the website proved to be oversaturated for the available data and could not be tested in the hypothesized form. Therefore, the relationships were tested in a series of regression analyses.

For the informational pages (H4a; Figure 2), the following steps were conducted: (1) multiple regression analysis of the explanatory variables (content perception, visual aesthetic, and usability) on the outcome variable (satisfaction), (2) multiple

XSL•FO
RenderX

regression analysis of the explanatory variables (content perception, visual aesthetic, and usability) on the outcome variable (continuance intention), and (3) linear regression analysis of the explanatory variable (satisfaction) on the outcome variable (continuance intention).

The first step showed that visual aesthetic (β=.30, P<.001) and content perception (β=.49, P<.001) were significant predictors of website satisfaction (R^2=0.59, $F_{3,143}$=69.06, P<.001), whereas usability was not (β=.08, P=.36). The second step showed that content perception (β=.62, P<.001) was a significant predictor of continuance intention (R^2=0.45, $F_{3,143}$=39.74, P<.001), but visual aesthetics (β=.00, P=.94) and usability (β=.08, P=.42) were not. The last step revealed that satisfaction (β=.56, P<.001) was a significant predictor of continuance intention (R^2=0.31, $F_{1,145}$=64.29, P<.001). Figure 6 offers a graphic representation of the series of regressions.

For HS (H4b; Figure 3), the following steps were conducted: (1) multiple regression analysis of the explanatory variables (visual aesthetic and usability) on the outcome variable (satisfaction), (2) multivariate regression analysis of the explanatory variable (satisfaction) on the outcome variables (continuance intention and intention to recommend), and (3) multiple regression analysis of the explanatory variables (visual aesthetic and usability) on the outcome variables (continuance intention and intention to recommend).

For HS, the first step revealed that usability (β=.63, P<.001 for diagnostics; β=.44, P<.001 for interventions) but not visual aesthetic (β=.10, P=.30 for diagnostics; β=.17, P=.14 for interventions) was a significant predictor of satisfaction (R^2=0.45, $F_{2,67}$=28.16, P<.001 for diagnostics; R^2=0.27, $F_{2,67}$=12.63, P<.001 for interventions). The second step showed that satisfaction was a significant predictor of both continuance intention (β=.70, P<.001; R^2=0.49, $F_{1,68}$=66.31, P<.001 for diagnostics; β=.76, P<.001; R^2=0.58, $F_{1,68}$=94.75, P<.001 for interventions) and intention to recommend (β=.69, P<.001; R^2=0.47, $F_{1,68}$=62.06, P<.001 for diagnostics; β=.81, P<.001; R^2=0.65, $F_{1,68}$=128.7, P<.001 for interventions). The third step showed that usability was a significant predictor of both continuance intention (β=.58, P<.001; R^2=0.34, $F_{1,68}$=34.57, P<.001 for diagnostics; β=.58, P<.001; R^2=0.34, $F_{1,68}$=35.11, P<.001 for interventions) and intention to recommend (β=.47, P<.001; R^2=0.21, $F_{1,68}$=19.02, P<.001 for diagnostics; β=.38, P<.001; R^2=0.15, $F_{1,68}$=12.23, P<.001 for interventions). Figure 7 offers a graphic representation of the series of regressions for HS diagnostics (black solid lines) and HS interventions (green dashed lines).

Figure 6. Visual representation of regression analyses of content perception, visual aesthetic, and usability on satisfaction and of satisfaction on continuance intention for LONDI informational pages (H4a; N=147).

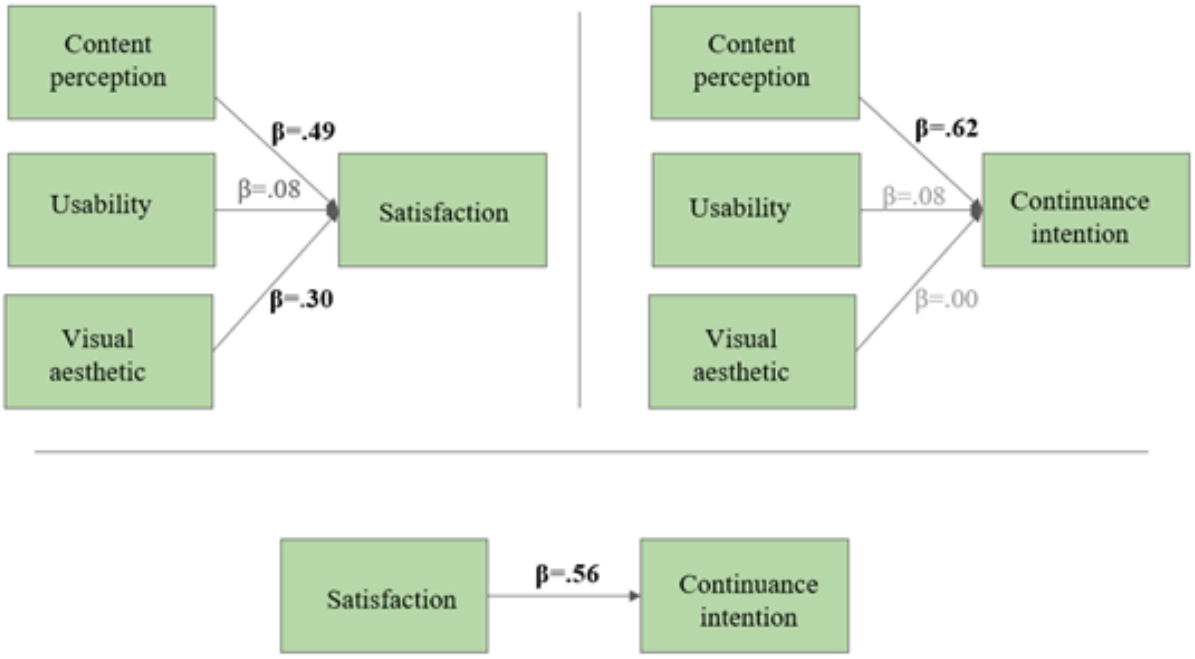

XSL•FO
RenderX

Figure 7. Visual representation of regression analyses of visual aesthetic and usability on satisfaction and of satisfaction on continuance intention and intention to recommend for LONDI Help System (HS) for diagnostics (solid line) and HS for interventions (green dashed line) for learning therapists (hypothesis 4b; n=70).

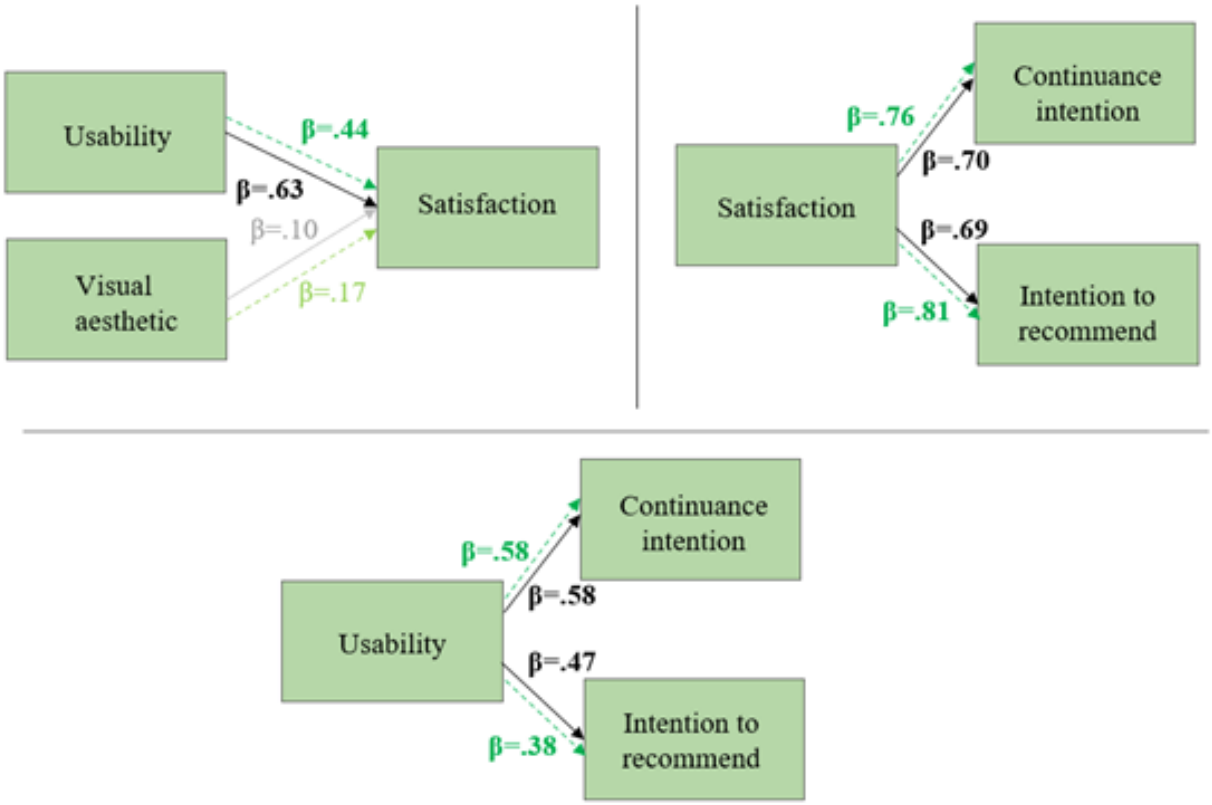

Qualitative Content Analysis

In the sample of parents, a total of 414 statements about the informational pages were coded and analyzed. In the sample of learning therapists, a total of 319 statements were coded for the informational pages, 211 statements were coded for HS for diagnostics, and a further 233 statements were coded for HS for interventions.

The interrater reliability for the content analysis coding was calculated based on the first 212 units of feedback and was found to be acceptable after the first round:

1. Positive/negative: κ=0.87 (95% CI 0.81-0.93)
2. Existing feature/new feature request: κ=0.81 (95% CI 0.72-0.90)
3. Variable: κ=0.92 (95% 0.86-0.97)

Among parents, one-half of the feedback regarding the informational pages was positive (222/414, 53.6% of all statements). A further 24.4% (101/414) of the statements were neutral, and 22% (91/414) were negative. Most statements concerned existing features and referred to the content of the pages (see Table 6 for a detailed overview). Of the 378 statements about content, 201 statements (53.2%) were positive. A recurring theme in the positive feedback was the usefulness of the information (49 mentions) and the clear structure of the text (20 mentions). Negative feedback and areas of improvement mainly concerned visual aesthetics, such as legibility and insufficient contrast of fonts against the background, text length, and irritating infographic features (35 mentions). Another criticism (18 mentions) was that the procedures described on the website did not always correspond to the parents' experiences (eg, getting support from the school or getting an official diagnosis). The most frequent new feature request was adding links with relevant information about different regions of Germany and Austria, since the school regulations regarding learning disorders vary across countries (15 mentions).

Among learning therapists, most of the feedback was again positive (194/319, 60.8% of all statements), concerned the existing features (256/319, 80.3%), and referred to the content of the pages (296/319, 92.8; see Table 6). Of the 296 statements about content, 181 (61.1%) statements were positive. The main positive issues raised included the well-structured content (41 mentions), high visual appeal (7 mentions), and ease of use of the website (8 mentions). The main criticism was that the information on the website was not perceived as useful by more experienced learning therapists, since they found it too basic (16 mentions). The new feature requests from learning therapists included adding downloadable questionnaires or survey sheets that can be used in their practice (11 mentions) and links to further professional resources (13 mentions).

The feedback given about HS by the learning therapists comprised 211 statements for the diagnostics section and 230 statements for the interventions section of the tool. Most feedback units were positive (140/211, 66.4% about diagnostics and 125/230, 54.3% about interventions). As in the case of informational pages, most feedback concerned the existing features and referred to the diagnostic and intervention measures suggested by HS (see Table 7). Approximately 20% (46/211, 21.8% for HS for diagnostics; 110/230, 47.8% for HS for

XSL•FO
RenderX

interventions) of all feedback concerned usability, and another 13.3% (28/211) to 15.7% (36/230) of the feedback mentioned the perceived usefulness of the tool.

Table 6. Content analysis coding categories for the informational pages feedback from parents (414 statements) and learning therapists (319 statements).

Coded statements	Informational pages, n (%)	
	Parents (n=414)	Learning therapists (n=319)
Positivity		
Positive statements	222 (53.6)	194 (60.8)
Negative statements	91 (22)	62 (19.4)
Neutral statements	101 (24.4)	63 (19.8)
Existing or new		
Existing	315 (76.1)	256 (80.3)
New	99 (23.9)	63 (19.8)
Variable		
Content	378 (91.3)	296 (92.8)
Visual aesthetics	3 (0.7)	1 (0.3)
Usability	16 (3.9)	6 (1.9)
Expectations confirmation	0 (0)	0 (0)
Satisfaction	11 (2.7)	5 (1.6)
Perceived usefulness	2 (0.5)	5 (1.6)
Continuance intention	0 (0)	0 (0)
Intention to recommend	1 (0.2)	1 (0.3)
Other	3 (0.7)	5 (1.6)

Table 7. Content analysis coding categories for the feedback on the Help System (HS) for diagnostics (211 statements) and HS for interventions (231 statements).

Coded statements	HS for diagnostics (n=211), n (%)	HS for interventions (n=230), n (%)
Positivity		
Positive statements	140 (66.4)	125 (54.3)
Negative statements	39 (18.5)	63 (27.4)
Neutral statements	32 (15.2)	42 (18.3)
Existing or new		
Existing	180 (85.3)	188 (81.7)
New	31 (14.7)	42 (18.3)
Variable		
Content	101 (47.9)	110 (47.8)
Visual aesthetics	0 (0)	0 (0)
Usability	46 (21.8)	46 (20)
Expectations confirmation	0 (0)	0 (0)
Satisfaction	21 (10)	17 (7.4)
Perceived usefulness	28 (13.3)	36 (15.7)
Continuance intention	6 (2.8)	5 (2.2)
Intention to recommend	0 (0)	1 (0.4)
Other	9 (4.3)	15 (6.5)

XSL•FO
RenderX

Among the 101 statements on the content for HS diagnostics, 58 (57.4%) were positive. Among the 46 usability statements, 33 (71%) statements were positive. Among the 110 statements on the content for HS interventions, 59 (53.6%) were positive. However, among the 46 statements on usability, only 17 (36%) were positive. The main criticisms were windows embedded within other windows, which resulted in two scroll bars on the side of the screen (8 mentions); the difficulty with getting to use the tool due to many preparatory steps (15 mentions); and overall complexity (10 mentions). The most prominent request for a new feature consisted of adding intervention programs and diagnostic measures to the databank that were not included (10 mentions).

Discussion

Principal Findings

The aim of this study was to evaluate the LONDI website for two of its target audiences—parents of children with learning difficulties and learning therapists—and to specify existing models for website acceptance. We therefore assessed parents' and learning therapists' opinions on key website characteristics, namely usability, visual aesthetics, content, perceived usefulness, expectations confirmation, continuance intention, intention to recommend, and open written feedback to complement our findings.

Overall Acceptance of the Website

Regarding the overall acceptance of the website, both the parents' and learning therapists' opinions on the website were comparable to or higher than the average benchmark ratings reported for similar websites in Germany. Exceptions were the criticisms concerning perceived usefulness and expectations confirmation of the informational pages given by the learning therapists and the usability of HS. The qualitative content analysis offered further insights into these negative ratings. Learning therapists described that the information provided on the website seems more suitable for therapists with less experience. With respect to the usability of the HS, answers suggested a relatively high level of complexity.

RE-AIM Framework

Furthermore, we assessed the Reach, Effectiveness and Implementation, and Maintenance dimensions of the RE-AIM framework and tested hypotheses regarding the influence of age and experience of the target audiences and their website opinions, knowledge, attitudes, and self-efficacy change and predictors of usage maintenance.

Reach

Our findings were not in line with previous research, which showed that older participants tend to rate websites more positively than younger participants, regardless of whether the website was dedicated to health topics [53]. For parents, we did not find a relationship between their age or experience and their website opinions. For learning therapists, a negative correlation was found between their age and their satisfaction with the informational pages. Furthermore, negative correlations between age and usability for both HS (diagnostics and interventions) and between age and satisfaction with the HS interventions were found. Thus, we were able to refute H1a and H1b.

One possible explanation for the inconsistent findings is the variation in complexity between the websites assessed. Research on adults' internet usage has shown that adults 45 years and older require simpler layouts; fewer mouse movements necessary to reach a goal on the website; and a simpler design overall to account for age-related changes in vision, hearing, and psychomotor coordination [54]. Given that 63% of our learning therapists were older than 45 years, their more negative ratings may result from the relatively high complexity level of the website's HS. A recent study by Hattink and colleagues [54] showed that simpler websites tend to be rated as more attractive and easier to use, which aligns with our findings.

In addition, the written feedback from the learning therapists provided further explanations for these negative correlations. Experienced therapists may have already established routines and preferred interventions, reducing their willingness to use new tools. Moreover, many noted that their preferred programs were not included due to the evidence-based inclusion criteria in both HS.

Together, these findings provide important information for future adaptations. Simplifying both HS and clarifying how programs are selected could improve usability and satisfaction.

Effectiveness and Implementation

Both groups showed increases in their knowledge about learning disorders after using the website. When it came to attitudes and self-efficacy, findings were less consistent. Although attitudes and self-efficacy increased in the sample of learning therapists, this was not the case in the sample of parents. Therefore, we can confirm H2 for learning therapists and partially confirm it for parents. The qualitative feedback provided by parents revealed several potential reasons why their attitudes and self-efficacy did not improve after using the website. Regarding attitudes, most parents in our sample already had at least 1 year of experience dealing with the learning disorders, which means that their core beliefs about the disorders have likely already been established. Furthermore, the qualitative responses highlighted the significance of parents' self-perceptions and their perceived role in their child's disorder. Several parents noted that realizing the disorder is not their fault and acknowledging their limited influence were particularly meaningful. Although the LONDI website does address these points, more content can be created on parents' self-regulation, and the definition of attitudes can be broadened to include many more aspects than the disorder itself. Unfortunately, the attitudes questionnaire assessed only parents' attitudes toward children and parent-child activities, entirely overlooking their attitudes toward themselves. Therefore, further investigation into the attitudes of parents in the context of learning disorders is needed.

Regarding self-efficacy, some parents' feedback indicated that the website offered theory but lacked actionable content. Parents requested more practical information, such as links to local authorities and real-life examples. Some felt the website idealized how support systems function, which may explain the lack of self-efficacy gains. Thus, future adaptations should

XSL•FO
RenderX

include content that better addresses the practical realities parents face.

Maintenance

Our analysis supported the information system continuance intention framework. Expectations confirmation and perceived usefulness remain reliable predictors of user satisfaction and continuance intention of the website. This study extended previous research by showing that the intention to recommend a website, as measured using the NPS score, is predicted by website satisfaction and perceived usefulness and might therefore be included in the information system continuance intention framework. Given that this study tested this relationship for the first time, future research would need to replicate these findings for websites with other scopes and characteristics.

Furthermore, our path analysis revealed that expectations confirmation is a direct predictor of continuance intention for HS for interventions but not for the informational pages nor HS diagnostics. In the case of HS for interventions, satisfaction is not a significant predictor of continuance intention, while expectation confirmation is. This difference may suggest that expectations confirmation plays a bigger role in user experience for online tools that have a more practical component. HS for interventions is more complex than HS for diagnostics: It requires more time to become familiar with the tool and manual input of diagnostic test results. Numerous meta-analyses done on the topic of the information system continuance intention framework have shown its effectiveness in different areas, such as mobile shopping, mobile health, online learning, and e-commerce [55-58]; however, none of these studies directly analyzed the role of involvement or level of activity of the system or tool in predicting continuance or recommendation intention. Thus, future research should investigate the degree of interactivity of an online tool and how this affects the predictors of continuance intention. Moreover, HS for interventions may be perceived by learning therapists as more important for their work, because many learning therapists do not conduct the diagnostic assessments themselves and focus solely on interventions. Therefore, expectations confirmation may play a bigger role in their continuance intention.

Regarding the website's characteristics such as usability, content, and visual aesthetic, the results revealed that different parts of the website predicted satisfaction and continuance intention differently. For the informational pages, the website's content and visual aesthetics were the only significant predictors of satisfaction and intention to use, while for the HS, a tool that has more interactive elements, only usability was essential. Thus, we could not confirm our hypothesis that all website characteristics will equally predict user satisfaction and continuance intention. Rather, these results suggested that the importance of the different website characteristics depends on the structure of the website itself. In line with this interpretation, previous studies have found that content and visual aesthetics were primarily important during the initial phase of forming a first impression of a website, whereas usability became more important as users interact with the website repeatedly and gain more in-depth knowledge of it [58,59]. Both HS are interactive tools for professionals that demand attention and in-depth understanding of how it works, whereas informational pages simply provide information in the form of text or graphics with few interactive elements. Therefore, future research needs to take into account that websites with different functionalities need to be assessed separately, considering their primary goal.

Strengths and Limitations

The study's main strength is its comprehensive scope, enabling us to assess key predictors of future user behavior and identify areas for further improvement. The website evaluation frameworks and questionnaires were chosen rigorously. We selected questionnaires that had already been validated in German to maximize the study's validity.

However, there are also limitations to consider when interpreting our results. Participants completed the questionnaires online in the presence of a tester, potentially inducing a social desirability bias.

Another limitation concerns the questionnaire used for assessing the participants' knowledge change. The same knowledge test was used pre- and postintervention, potentially creating a priming effect. This design was necessary due to content limitations. Regardless of this issue, the significant results suggest a positive change in knowledge after reading the platform's content.

Finally, the reference and benchmark values that we compared our website against were only available for informational websites and not for online data banks and search tools such as the HS. Thus, it is hard to say if HS in LONDI can truly compare with other websites.

Conclusions

Overall, the findings extended existing theories on website acceptance and provided evidence for the website's usefulness to the target audience investigated. Moreover, they offered valuable guidance for future improvements. Although the informational pages would benefit from more practical tips for parents, the HS tool needs substantial revisions to enhance its usability. Most importantly, the results highlighted the need for a deeper understanding of the target audience and its expectations.

Acknowledgments

This study was supported by the Federal Ministry of Education and Research of Germany (Bildungsministerium für Bildung und Forschung). We would like to extend our gratitude to the study assistants Rosalie Dengler, Jolina Cestonaro, and Valeriia Dubovikova and to our partners from the Leibniz Institute for Research and Information in Education.

Generative artificial intelligence (AI) was not used in the writing of this manuscript.

XSL•FO
RenderX

Data Availability

The datasets generated or analyzed during this study are available in the Open Science Framework (OSF) repository in the project "Phase I Evaluation of the information pages of londi.de, a Website for Learning Disorders for Parents and Learning Therapists" under Files [60].

Conflicts of Interest

None declared.

Multimedia Appendix 1
Full-text questionnaires in English and German.
[PDF File (Adobe PDF File), 681 KB - formative_v9i1e68365_app1.pdf]

Multimedia Appendix 2
Coding scheme and summary of coding categories for the qualitative content analysis.
[PDF File (Adobe PDF File), 3920 KB - formative_v9i1e68365_app2.pdf]

References

1. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition. Washington, DC: American Psychiatric Association Publishing; 2013.
2. Carroll JM, Holden C, Kirby P, Thompson PA, Snowling MJ, Dyslexia Delphi Panel. Toward a consensus on dyslexia: findings from a Delphi study. J Child Psychol Psychiatry 2025 Jul;66(7):1065-1076 [FREE Full text] [doi: 10.1111/jcpp.14123] [Medline: 39996533]
3. Gustafson S, Samuelsson S. Intelligence and dyslexia: implications for diagnosis and intervention. Scand J Psychol 1999 Jun;40(2):127-134. [doi: 10.1111/1467-9450.00109] [Medline: 10399495]
4. Moll K, Kunze S, Neuhoff N, Bruder J, Schulte-Körne G. Specific learning disorder: prevalence and gender differences. PLoS One 2014;9(7):e103537 [FREE Full text] [doi: 10.1371/journal.pone.0103537] [Medline: 25072465]
5. Wagner RK, Zirps FA, Edwards AA, Wood SG, Joyner RE, Becker BJ, et al. The prevalence of dyslexia: a new approach to its estimation. J Learn Disabil 2020;53(5):354-365 [FREE Full text] [doi: 10.1177/0022219420920377] [Medline: 32452713]
6. Korhonen J, Linnanmäki K, Aunio P. Learning difficulties, academic well-being and educational dropout: a person-centred approach. Learning and Individual Differences 2014 Apr;31:1-10. [doi: 10.1016/j.lindif.2013.12.011]
7. Carroll JM, Maughan B, Goodman R, Meltzer H. Literacy difficulties and psychiatric disorders: evidence for comorbidity. J Child Psychol Psychiatry 2005 May 21;46(5):524-532. [doi: 10.1111/j.1469-7610.2004.00366.x] [Medline: 15845132]
8. Goldston DB, Walsh A, Mayfield Arnold E, Reboussin B, Sergent Daniel S, Erkanli A, et al. Reading problems, psychiatric disorders, and functional impairment from mid- to late adolescence. J Am Acad Child Adolesc Psychiatry 2007 Jan;46(1):25-32. [doi: 10.1097/01.chi.0000242241.77302.f4] [Medline: 17195726]
9. Griepenburg C, Schuchardt K, Mähler C. [Psychoeducation for parents of children with learning disorders: a satisfaction study]. Prax Kinderpsychol Kinderpsychiatr 2021 May 12;70(4):298-315. [doi: 10.13109/prkk.2021.70.4.298] [Medline: 33977877]
10. Coughler C, M Burke S, Cardy JO. Analysis of the quality of online resources for parents of children who are late to talk. Autism Dev Lang Impair 2020 Apr 06;5:2396941520917940 [FREE Full text] [doi: 10.1177/2396941520917940] [Medline: 36381543]
11. Lernstörungen Online-Platform für Diagnostik and Intervention (LONDI). URL: https://londi.de/ [accessed 2025-05-27]
12. Glasgow RE, Harden SM, Gaglio B, Rabin B, Smith ML, Porter GC, et al. RE-AIM planning and evaluation framework: adapting to new science and practice with a 20-year review. Front Public Health 2019;7:64 [FREE Full text] [doi: 10.3389/fpubh.2019.00064] [Medline: 30984733]
13. de la Vega R, Ritterband L, Palermo TM. Assessing digital health implementation for a pediatric chronic pain intervention: comparing the RE-AIM and BIT frameworks against real-world trial data and recommendations for future studies. J Med Internet Res 2020 Sep 01;22(9):e19898 [FREE Full text] [doi: 10.2196/19898] [Medline: 32870158]
14. Glasgow RE, Dickinson P, Fisher L, Christiansen S, Toobert DJ, Bender BG, et al. Use of RE-AIM to develop a multi-media facilitation tool for the patient-centered medical home. Implement Sci 2011 Oct 21;6:118 [FREE Full text] [doi: 10.1186/1748-5908-6-118] [Medline: 22017791]
15. RE-AIM. URL: https://re-aim.org/ [accessed 2025-05-28]
16. Berreta K, Nguyen C, Stoner AM, Ridgeway L, Wilson A, Fadel N, et al. A RE-AIM analysis of a mental health app for undergraduate and medical students during the COVID-19 pandemic: a retrospective cross-sectional study. Int J Environ Res Public Health 2023 Jun 30;20(13):1 [FREE Full text] [doi: 10.3390/ijerph20136266] [Medline: 37444113]

XSL•FO
RenderX

17. Shen Y, Xu W, Liang A, Wang X, Lu X, Lu Z, et al. Online health management continuance and the moderating effect of service type and age difference: a meta-analysis. Health Informatics J 2022 Aug 17;28(3):14604582221119950 [FREE Full text] [doi: 10.1177/14604582221119950] [Medline: 35976977]
18. Vlachogianni P, Tselios N. Perceived usability evaluation of educational technology using the Post-Study System Usability Questionnaire (PSSUQ): a systematic review. Sustainability 2023 Aug 28;15(17):12954. [doi: 10.3390/su151712954]
19. Thielsch MT. Ästhetik von Websites – Wahrnehmung von Ästhetik und deren Beziehung zu Inhalt, Usability und Persönlichkeitsmerkmalen. Lengerich, Germany: Pabst Science Publishers; 2017.
20. de Moel-Mandel C, Lynch C, Issaka A, Braver J, Zisis G, Carrington MJ, et al. Optimising the implementation of digital-supported interventions for the secondary prevention of heart disease: a systematic review using the RE-AIM planning and evaluation framework. BMC Health Serv Res 2023 Dec 04;23(1):1347 [FREE Full text] [doi: 10.1186/s12913-023-10361-6] [Medline: 38049862]
21. Nordberg SS, Jaso-Yim BA, Sah P, Schuler K, Eyllon M, Pennine M, et al. Evaluating the implementation and clinical effectiveness of an innovative digital first care model for behavioral health using the RE-AIM framework: quantitative evaluation. J Med Internet Res 2024 Oct 30;26:e54528 [FREE Full text] [doi: 10.2196/54528] [Medline: 39476366]
22. Yau HK, Leung YF. The Relationship between Self-Efficacy and Attitudes towards the Use of Technology in Learning in Hong Kong Higher Education. 2018 Presented at: International MultiConference of Engineers and Computer Scientists 2018 (IMECS 2018); March 14-16, 2018; Hong Kong URL: http://www.iaeng.org/publication/IMECS2018/
23. Davis F. A technology acceptance model for empirically testing new end-user information systems: theory and results. MIT Libraries. 1986. URL: https://dspace.mit.edu/handle/1721.1/15192 [accessed 2025-08-11]
24. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly 1989 Sep;13(3):319-340. [doi: 10.2307/249008]
25. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view. MIS Quarterly 2003;27(3):425. [doi: 10.2307/30036540]
26. Bhattacherjee A. Understanding information systems continuance: an expectation-confirmation model. MIS Quarterly 2001 Sep;25(3):351. [doi: 10.2307/3250921]
27. Bhattacherjee A, Perols J, Sanford C. Information technology continuance: a theoretic extension and empirical test. Journal of Computer Information Systems 2015 Dec 11;49(1):17-26. [doi: 10.1080/08874417.2008.11645302]
28. Zhang J, Zhang M, Liu Y, Zhang L. What are the key drivers to promote continuance intention of undergraduates in online learning? A multi-perspective framework. Front Psychol 2023;14:1121614 [FREE Full text] [doi: 10.3389/fpsyg.2023.1121614] [Medline: 36910743]
29. Al Amin M, Muzareba AM, Chowdhury IU, Khondkar M. Understanding e-satisfaction, continuance intention, and e-loyalty toward mobile payment application during COVID-19: an investigation using the electronic technology continuance model. J Financ Serv Mark 2023 Jan 06;29(2):318-340. [doi: 10.1057/s41264-022-00197-2]
30. Tsai H, Lee Y, Ruangkanjanases A. Understanding the effects of antecedents on continuance intention to gather food safety information on websites. Front Psychol 2020;11:579322 [FREE Full text] [doi: 10.3389/fpsyg.2020.579322] [Medline: 33424693]
31. Wang T, Wang W, Liang J, Nuo M, Wen Q, Wei W, et al. Identifying major impact factors affecting the continuance intention of mHealth: a systematic review and multi-subgroup meta-analysis. NPJ Digit Med 2022 Sep 15;5(1):145 [FREE Full text] [doi: 10.1038/s41746-022-00692-9] [Medline: 36109594]
32. Wang L, Lew S, Lau S, Leow M. Usability factors predicting continuance of intention to use cloud e-learning application. Heliyon 2019 Jun;5(6):e01788 [FREE Full text] [doi: 10.1016/j.heliyon.2019.e01788] [Medline: 31198866]
33. Finn A, Wang L, Frank T. Attribute perceptions, customer satisfaction and intention to recommend e-services. Journal of Interactive Marketing 2009 Aug 01;23(3):209-220. [doi: 10.1016/j.intmar.2009.04.006]
34. Hosany S, Prayag G, Van Der Veen R, Huang S, Deesilatham S. Mediating effects of place attachment and satisfaction on the relationship between tourists' emotions and intention to recommend. Journal of Travel Research 2016 Nov 30;56(8):1079-1093. [doi: 10.1177/0047287516678088]
35. Byun DH, Finnie G. Evaluating usability, user satisfaction and intention to revisit for successful e-government websites. EG 2011;8(1):1. [doi: 10.1504/eg.2011.037694]
36. Henderson MSG, Andersen JA, Dinh JM, Ziegenfuss JY, Canterbury M, JaKa MM. In their own words: content analysis of open-ended survey responses for a more nuanced, local understanding of mental illness stigma. Issues Ment Health Nurs 2024 May;45(5):498-505. [doi: 10.1080/01612840.2024.2318603] [Medline: 38564780]
37. Phase I Evaluation of the information pages of londi.de, a Website for Learning Disorders for Parents and Learning Therapists. OSF Registries. 2023 Jun 23. URL: https://osf.io/qg5y2 [accessed 2025-05-28]
38. Phase I Evaluation of the help system for learning therapists at londi.de, a Website for Learning Disorders. OSF Registries. 2023 Jun 26. URL: https://osf.io/3x4hv [accessed 202-03-28]
39. Ajzen I, Fishbein M. Attitudes and the attitude-behavior relation: reasoned and automatic processes. European Review of Social Psychology 2011 Apr 15;11(1):1-33. [doi: 10.1080/14792779943000116]
40. Hannekum R, Schuchardt K, Mähler C. FRABU-LRS: Fragebogen zur Unterrichtsanpassung bei Schülerinnen und Schülern mit LRS. Diagnostica 2025 Jan;71(1):24-36. [doi: 10.1026/0012-1924/a000339]

XSL•FO
RenderX

41. Bandura A. Self-efficacy: the exercise of control. New York, NY: W H Freeman & Co; 1997.
42. Höltge L, Ehm J, Hartmann U, Hasselhorn M. Teachers' self-efficacy beliefs regarding assessment and promotion of school-relevant skills of preschool children. Early Child Development and Care 2017 May 04;189(2):339-351. [doi: 10.1080/03004430.2017.1323888]
43. Ergonomics of human-system interaction, Part 151: Guidance on World Wide Web user interfaces. ISO. 2008. URL: https://www.iso.org/standard/37031.html [accessed 2025-08-11]
44. Lewis JR. The System Usability Scale: past, present, and future. International Journal of Human–Computer Interaction 2018 Mar 30;34(7):577-590. [doi: 10.1080/10447318.2018.1455307]
45. Thielsch MT, Salaschek M. [Toolbox for user-experience evaluation of online health information]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2020 Jun;63(6):721-728. [doi: 10.1007/s00103-020-03142-7] [Medline: 32367206]
46. Budner P, Fischer M, Rosenkranz C, Basten D, Terlecki L. Information system continuance intention in the context of network effects and freemium business models: a replication study of cloud services in Germany. Transactions on Replication Research 2017 Dec;3(4):1-13. [doi: 10.17705/1atrr.00019]
47. Reichheld FF. The one number you need to grow. Harv Bus Rev 2003 Dec;81(12):46-54, 124. [Medline: 14712543]
48. Markus KA. Principles and practice of structural equation modeling by Rex B. Kline. Structural Equation Modeling: A Multidisciplinary Journal 2012 Jul 20;19(3):509-512. [doi: 10.1080/10705511.2012.687667]
49. Chen F, Curran PJ, Bollen KA, Kirby J, Paxton P. An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. Sociol Methods Res 2008 Jan 01;36(4):462-494 [FREE Full text] [doi: 10.1177/0049124108314720] [Medline: 19756246]
50. Hooper D, Coughlan J, Mullen M. Structural equation modelling: guidelines for determining model fit. Electronic Journal of Business Research Methods 2008;6(1):53-60. [doi: 10.21427/d7cf7r]
51. Cho G, Hwang H, Sarstedt M, Ringle CM. Cutoff criteria for overall model fit indexes in generalized structured component analysis. J Market Anal 2020 Sep 20;8(4):189-202. [doi: 10.1057/s41270-020-00089-1]
52. Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Structural Equation Modeling: A Multidisciplinary Journal 1999 Jan;6(1):1-55. [doi: 10.1080/10705519909540118]
53. Wagner N, Hassanein K, Head M. Computer use by older adults: a multi-disciplinary review. Computers in Human Behavior 2010 Sep;26(5):870-882 [FREE Full text] [doi: 10.1016/j.chb.2010.03.029]
54. Hattink B, Droes R, Sikkes S, Oostra E, Lemstra AW. Evaluation of the digital Alzheimer center: testing usability and usefulness of an online portal for patients with dementia and their carers. JMIR Res Protoc 2016 Jul 21;5(3):e144 [FREE Full text] [doi: 10.2196/resprot.5040] [Medline: 27444209]
55. Chauhan S, Kumar P, Jaiswal M. A meta-analysis of M-commerce continuance intention: moderating impact of culture and user types. Behaviour & Information Technology 2021 Jul 31;41(13):2905-2923. [doi: 10.1080/0144929x.2021.1960607]
56. Luceri B, (Tammo) Bijmolt T, Bellini S, Aiolfi S. What drives consumers to shop on mobile devices? Insights from a meta-analysis. Journal of Retailing 2022 Mar;98(1):178-196 [FREE Full text] [doi: 10.1016/j.jretai.2022.02.002]
57. Zhang M, Li S, Zhang Y. A meta-analysis of the moderating role of prior learning experience and mandatory participation on factors influencing MOOC learners? continuance intention. Australasian Journal of Educational Technology 2023;39(2):115-141 [FREE Full text] [doi: 10.14742/ajet.7795]
58. Thielsch MT, Blotenberg I, Jaron R. User evaluation of websites: from first impression to recommendation. Interacting with Computers 2013 Jun 04;26(1):89-102. [doi: 10.1093/iwc/iwt033]
59. Zhang P, von Dran GM. Satisfiers and dissatisfiers: a two-factor model for website design and evaluation. Journal of the American Society for Information Science 2000;51(14):1253-1268. [doi: 10.1002/1097-4571(2000)9999:9999<::aid-asi1039>3.0.co;2-o]
60. Drobina O, Dümig P, Witzel B, Weinreich L, Moll K, Schulte-Körne G. Phase I Evaluation of the information pages of londi.de, a Website for Learning Disorders for Parents and Learning Therapists. Open Science Framework. 2023 Mar 22. URL: https://osf.io/76xh3/ [accessed 2025-08-18]

Abbreviations

CFI: comparative fit index
DSM: Diagnostic and Statistical Manual of Mental Disorders (Fifth Edition)
GDPR: General Data Protection Regulation
HS: Help System
LMU: Ludwig-Maximilians University
MANOVA: multivariate analysis of variance
NPS: Net Promoter Score
RMSEA: root mean square error of approximation
SRMR: standardized root mean squared residual

XSL•FO
RenderX

Edited by A Mavragani; submitted 07.11.24; peer-reviewed by S Dipietro, J Tay; comments to author 08.05.25; revised version received 28.05.25; accepted 08.07.25; published 26.09.25.

Please cite as:
Hermansson O, Dümig P, Witzel B, Weinreich L, Volkmer S, Schulte-Körne G, Moll K
Evaluating a Website on Learning Disorders for Parents and Learning Therapists: Observational Mixed Methods Study
JMIR Form Res 2025;9:e68365
URL: https://formative.jmir.org/2025/1/e68365
doi:10.2196/68365
PMID:

©Olga Hermansson, Paula Dümig, Björn Witzel, Lior Weinreich, Susanne Volkmer, Gerd Schulte-Körne, Kristina Moll. Originally published in JMIR Formative Research (https://formative.jmir.org), 26.09.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Differential Diagnosis Assessment in Ambulatory Care With a Digital Health History Device: Pseudorandomized Study

Beth Healey[1], MBChB; Adrien Schwitzguebel[2], Dr med; Herve Spechbach[3], Dr med

[1]Emergency Department, University Hospitals of Geneva, Geneva, Switzerland

[2]AS Medicine du Sport, Hopital de la Providence, Neuchatel, Switzerland

[3]Department of Primary Care Medicine, University Hospitals of Geneva, Geneva, Switzerland

Corresponding Author:
Beth Healey, MBChB
Emergency Department
University Hospitals of Geneva
Clinique de Belle Idée
Geneva, 1226
Switzerland
Phone: 41 07375242292
Email: beth.healey@hcuge.ch

Abstract

Background: Digital health history devices represent a promising wave of digital tools with the potential to enhance the quality and efficiency of medical consultations. They achieve this by providing physicians with standardized, high-quality patient history summaries and facilitating the development of differential diagnoses (DDs) before consultation, while also engaging patients in the diagnostic process.

Objective: This study evaluates the efficacy of one such digital health history device, diagnosis and anamnesis (DIANNA), in assisting with the formulation of appropriate DDs in an outpatient setting.

Methods: A pseudorandomized controlled trial was conducted with 101 patients seeking care at the University Hospital Geneva emergency outpatient department. Participants presented with various conditions affecting the limbs, back, and chest. The first 51 patients were assigned to the control group, while the subsequent 50 formed the intervention group. In the control group, physicians developed DD lists based on traditional history-taking and clinical examination. In the intervention group, physicians reviewed DIANNA-generated DD reports before interacting with the patient. In both groups, a senior physician independently formulated a DD list, serving as the gold standard for comparison.

Results: The study findings indicate that DIANNA use was associated with a notable improvement in DD accuracy (mean 79.3%, SD 24%) compared with the control group (mean 70.5%, SD 33%; P=.01). Subgroup analysis revealed variations in effectiveness based on case complexity: low-complexity cases (1-2 possible DDs) showed 8% improvement in the intervention group (P=.08), intermediate-complexity cases (3 possible DDs) showed 17% improvement (P=.03), and high-complexity cases (4-5 possible DDs) showed 15% improvement (P=.92). The intervention was not superior to the control in low-complexity cases (P=.08) or high-complexity cases (P=.92). Overall, DIANNA successfully determined appropriate DDs in 81.6% of cases, and physicians reported that it helped establish the correct DD in 26% of cases.

Conclusions: The study suggests that DIANNA has the potential to support physicians in formulating more precise DDs, particularly in intermediate-complexity cases. However, its effectiveness varied by case complexity and further validation is needed to assess its full clinical impact. These findings highlight the potential role of digital health history devices such as DIANNA in improving clinical decision-making and diagnostic accuracy in medical practice.

Trial Registration: ClinicalTrials.gov NCT03901495; https://clinicaltrials.gov/study/NCT03901495

(JMIR Form Res 2025;9:e56384) doi:10.2196/56384

KEYWORDS

differential diagnosis; decision-making; artificial intelligence; machine learning; digital diagnosis; computer-assisted; digital assessments; hospital-outpatient clinics; emergency; clinical applications software; diagnosis; ambulatory care; digital health; pseudorandomized study; digital tools; digital technology; medical consultation; decision support; health informatics

XSL•FO
RenderX

Introduction

In modern health care, digital health history devices (DHHDs) are playing an increasingly significant role in improving the quality and efficiency of medical consultations. These devices offer standardized patient history summaries, which have the potential to enable more accurate differential diagnoses (DDs) by gathering comprehensive patient data before medical consultations. Research has demonstrated that DHHDs enhance clinical decision-making by reducing diagnostic errors and improving efficiency, allowing health care providers to focus on critical aspects of care rather than repetitive information gathering [1,2]. By streamlining the data acquisition process, DHHDs facilitate more informed decision-making, benefiting both patients and health care professionals [3].

Among these devices is "DIANNA" (diagnosis and anamnesis), which has undergone substantial improvements since its initial development. A previous randomized controlled trial provided early insights into the tool's impact on diagnostic accuracy and efficiency. Building on this foundation, DIANNA now includes a body pictogram feature to better select symptomatic areas, making it more intuitive and comprehensive for clinicians [4,5]. This enhanced version of DIANNA is designed to bridge the gap between patient-reported symptoms and clinical assessment, offering clearer data for health care professionals with the aim of improving the diagnostic process. The integration of artificial intelligence (AI) within DIANNA further augments its capabilities. AI features such as natural language processing algorithms, machine learning models, and predictive analytics provide valuable diagnostic support and assist health care professionals by suggesting DDs based on patient data [6,7].

Despite the promising advancements in DHHDs, significant gaps remain in the literature regarding their real-world efficacy across diverse clinical settings and their performance in larger patient populations. While prior studies have demonstrated that AI-driven tools can enhance diagnostic precision [8,9], research specifically focused on DHHDs such as DIANNA remains limited, particularly in assessing their long-term impact on clinical outcomes and their integration into routine practice.

This study aims to evaluate the efficacy of the updated version of DIANNA in supporting clinical diagnosis, building upon previous findings while addressing the knowledge gap surrounding its role in enhancing diagnostic accuracy and efficiency in real-world settings. By assessing its effectiveness in DD formulation, this research seeks to establish a clearer understanding of how DIANNA can assist health care professionals in making more informed clinical decisions and ultimately contribute to improving patient outcomes.

Methods

Study Design

This study used a single-center, unblinded, 1:1 pseudorandomized design to evaluate the efficacy of DIANNA. The first 50 patients recruited were assigned to the control group, while the subsequent 50 were allocated to the intervention group. The study did not require follow-up and no modifications were made to the protocol after the trial commenced. Throughout the trial, no significant technical or digital issues necessitated corrections.

Patient Population

The study enrolled adult patients who sought care at the University Hospital Geneva emergency outpatient department between February 2019 and June 2020. Recruitment was conducted through a combination of methods to ensure broad participation. Patients presenting with a qualifying complaint were invited to participate by health care providers, with eligibility criteria communicated at the time of their visit. This convenience sampling approach facilitated the inclusion of a diverse patient population seeking emergency care.

To enhance awareness, the study was also advertised through flyers posted within the emergency department and internal communication with health care providers. Interested patients who met the inclusion criteria were provided with detailed information about the study's objectives and the digitalized DIANNA tool, allowing them to voluntarily express their willingness to participate.

Patients were eligible if they presented with an injury or medical condition affecting the upper limb, trunk, or lower limb and were able to fully use DIANNA. Exclusion criteria included strictly dermatologic conditions, straightforward injuries (such as toe, hand, or ankle inversion injuries), urgent medical conditions, and patients unable to complete the DIANNA tool, including those with visual impairments, elderly individuals, and non-Francophones.

Pseudorandomization and Recruitment

Recruitment for the study occurred in 2 phases. The first 50 patients designated as the "control group" were recruited between February 2019 and April 2019. The subsequent 50 patients allocated to the "intervention group" were recruited across 2 distinct periods: November to December 2019 and June to July 2020. The recruitment process was influenced by several factors, including the availability of the coordinating researcher and the impact of the COVID-19 pandemic. Due to pandemic-related restrictions and disruptions, recruitment was paused during certain periods and resumed when feasible. Whenever a patient was identified by the emergency department software, the coordinating researcher assessed their eligibility for enrollment, with final confirmation made by the senior physician (SP). Recruitment continued until 50 patients were enrolled in each group, ensuring balanced representation for analysis and comparison. Any adaptations to the protocol due to the pandemic, if applicable, were not mentioned in the Methods section and should be clarified for transparency.

DIAANA Tool

DIAANA is supporting the diagnostic process by introducing a comprehensive body segmentation system with zone-selection pictograms, replacing traditional questions regarding symptom localization. This innovative approach involves an interactive questionnaire that patients complete prior to their consultation. Instead of a lengthy list of 269 questions, DIAANA uses multiple-choice formats to collect relevant data. Subsequently,

XSL•FO
RenderX

the software uses AI to analyze this information and suggest a list of potential diagnoses from a pool of 126 diagnostic entities.

The AI-driven reasoning within DIAANA emulates the decision-making process of a specialist physician to establish a DD. The resulting information is then seamlessly conveyed to the physician in a user-friendly format, which includes a concise patient history summary, highlighting pertinent elements from the questionnaire. In addition, the report provides a list of possible diagnoses, completed with emergency-level categorization, potential contributing factors, and first-line management recommendations. This innovative approach not only streamlines the diagnostic process but also enhances the quality of care provided to patients.

Intervention

In this study, 2 groups of resident physicians (RPs) were involved. The first group used the DHHD while the second group did not have access to this device. For patients in the intervention group, they were instructed to independently complete DIANNA using a touch screen tablet, guided by the coordinator, without external assistance. The DIANNA-generated summary was then printed and provided to the RP before the patient's consultation.

After the consultation, but before reviewing additional medical test results, the RP compiled a list of potential DDs from a digital diagnosis list without external support. The case was subsequently presented to the SP, who independently selected their own DD from the same digital list to establish a "gold standard" for comparison.

The SP's assessment was made without influence from the resident's evaluation or the DHHD report, ensuring an impartial evaluation. This setup allowed for a comparison between the 2 groups, one with the aid of the DHHD and one without, to assess the impact of this technology on the diagnostic process (Figure 1).

Figure 1. Study flowchart. RP: resident physician; SP: specialized physician.

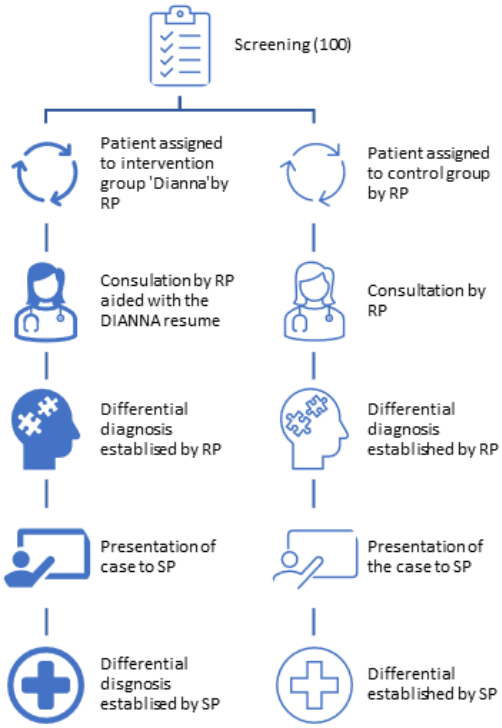

Outcomes

The outcomes of the study are listed in Textbox 1.

In summary, the study's comprehensive array of outcome measures aimed to assess the impact of DIANNA on diagnostic accuracy, consultation efficiency, patient satisfaction, and the perspectives of RPs. The novel stratification of DD accuracy based on case complexity added a unique dimension to the evaluation of DHHDs' performance in clinical practice.

XSL•FO
RenderX

Textbox 1. Primary and secondary outcome measures.

Primary outcomes

- The percentage of accurate differential diagnoses (DDs) established by resident physicians (RPs) using diagnosis and anamnesis (DIANNA) in comparison to the reference "gold-standard" assessments conducted by senior physicians.

Secondary outcomes

Secondary outcome measures encompassed a range of critical aspects, including:

- Consultation time: this outcome assessed the duration of patient-physician interactions, aiming to determine if the use of DIANNA impacted the efficiency of the consultation process.
- Patient satisfaction: patient satisfaction was evaluated on multiple dimensions, including:
  - Comprehensibility of questions (1-5 Likert scale): patients provided feedback on the clarity and ease of understanding the questions presented by DIANNA.
  - Accuracy in describing symptoms (percentage): patients rated the precision of how DIANNA described their symptoms.
  - Time taken to complete: this measure gauged the time efficiency of completing the DIANNA questionnaire.
  - Relevance to presenting complaint (percentage): patients assessed the extent to which DIANNA's questions were pertinent to their specific medical concern.
  - Preference for completion location and keeping summary (percentages): patients indicated their preference for completing the questionnaire at home and retaining a copy of the DIANNA summary for their records.
- Physician resident feedback: this outcome solicited feedback from RPs and included assessments of:
  - DIANNA's contribution to suggesting DDs (percentage): RPs indicated if DIANNA provided valuable insights and suggested DDs that they might have otherwise overlooked.
  - Exhaustiveness of the DD list: physicians assessed the completeness of the DD list generated by DIANNA.
  - Time savings (1-5 Likert scale): physicians rated the extent to which DIANNA expedited the diagnostic process on a Likert scale.
- Interest in future use in clinical practice (percentage): physicians expressed their willingness to incorporate DIANNA into their future clinical practice.
- Percentage of correct DDs according to case complexity: this outcome measure stratified the accuracy of DDs based on the complexity of the cases. Case complexity was defined by the number of DDs present in the "gold-standard" reference DDs, categorizing them into low complexity (1-2 DDs), intermediate complexity (3 DDs), and high complexity (4-5 DDs).

Statistical Analyses

The sample size for each group, consisting of 50 patients, was determined based on insights gained from our prior study [10]. Our rationale for this sample size calculation was as follows: we aimed to achieve 80% statistical power to detect a significant difference at the 5% level. Specifically, we sought to detect an increase in the primary outcome measure from an average of 59% correct diagnoses in the control group to 75% correct diagnoses in the experimental group. Given the study's robust design and expected low dropout rate, we determined that a total sample size of 100 patients was appropriate for this investigation.

For data analysis, we used descriptive statistics to characterize baseline characteristics. To assess group differences, we applied appropriate statistical tests. Table 1 explicitly states the statistical tests used to compare the demographic and clinical characteristics between the groups.

We used an intention-to-treat analysis, ensuring that all patients were included in the analysis. In cases of missing data, patients were not excluded; instead, analyses were conducted on available data.

The distribution of all variables was assessed using standard graphical analyses. As the distribution was determined to be parametric, we used the Student *t* test for all comparisons.

In addition, we conducted an analysis of covariance to adjust for potential confounders. This analysis focused on the primary outcome (use of DIANNA) and considered the number of diagnoses established by the SP, as well as other baseline characteristics where a *P* value of less than .2 was identified in the univariate analysis. The following covariates were included:

1. Global analysis, low complexity, and high complexity subgroups: number of diagnoses, chief resident's years of experience, and internal resident's evaluation.
2. Intermediate complexity subgroup: chief resident's years of experience and internal resident's evaluation (excluding the number of diagnoses).

Statistical significance was defined as $P<.05$.

All analyses were conducted using R v3.4.2 Portable, provided by the Free Software Foundation Inc. This analytical approach ensured a rigorous evaluation of study outcomes and allowed us to draw meaningful conclusions from the data.

XSL•FO
RenderX

Table 1. Baseline characteristics.

	DIANNA (n=50)	Control group (n=51)	*P* value
Age (years), mean (SD), range	34.1 (13), 18-65	38.5 (12), 17-63	.08
Male sex, n (%)	34 (68)	36 (71)	.95
Resident physician's practice (years), mean (SD), range	7.5 (3), 1-12	8 (3), 4-12	.36
Resident physician's skills (1-4) (years), mean (SD), range	2.7 (1), 1-4	2.4 (1), 1-4	.09
Senior physician's practice (years), mean (SD), range	6.9 (2), 1-11	6 (3), 1-10	.07
Senior physician's skills (1-4) (years), mean (SD), range	2.9 (1), 1-4	2.8 (1), 1-4	.64
Consultation time (years), mean (SD), range	3.7 (2), 1-10	3.2 (2), 2-11	.28
Initial complaint, n (%)			
Shoulder pain and trauma	5 (10)	6 (12)	>.99
Elbow pain	7 (14)	5 (10)	.73
Wrist and hand pain	9 (18)	7 (14)	.75
Back pain and trauma	2 (4)	5 (10)	.45
Pelvis pain	2 (4)	3 (6)	>.99
Knee pain and trauma	11 (22)	5 (10)	.16
Ankle trauma	4 (8)	3 (6)	.98
Foot trauma	3 (6)	9 (18)	.13
Soft tissue trauma and swelling	7 (14)	8 (16)	>.99
Case complexity (N of DD; years), mean (SD), range	2.6 (1), (1-5)	2.3 (1), 1-5	.23
Diagnoses found by resident physician (years), mean (SD), range	0.8 (0), 0.25-1	0.7 (0), 0-1	.13

Ethical Considerations

This study was approved by the Medical Ethics Committee of the University Hospitals Geneva in Switzerland (REQ-2017-00878). The study protocol was formally registered at ClinicalTrials.gov (NCT03901495) before recruitment commenced. Given the study's observational nature and the absence of any investigational drug or invasive procedures, an expedited ethical review process was conducted.

Since the study posed minimal risk to participants and did not involve any interventions beyond standard clinical procedures, oral informed consent was deemed appropriate. Each participant was provided with a concise written description of the study's aims, methods, and potential implications. The informed consent process ensured that participants understood the study and voluntarily agreed to participate before proceeding with DIANNA.

To safeguard patient privacy and maintain data confidentiality:

1. All collected data were anonymized, with no personally identifiable information linked to study records.
2. Patient responses and physician assessments were securely stored in an encrypted database accessible only to authorized research personnel.
3. The DIANNA tool operated on a secure platform without storing patient identifiers, ensuring compliance with data protection regulations.
4. Data analysis was conducted on deidentified datasets, minimizing risks of breaches in patient confidentiality.

Participants were not financially compensated for their involvement in this study, as the research was deemed noninterventional and posed no additional burden beyond routine medical care. However, participants were given the opportunity to retain a copy of their DIANNA summary for their own reference, which many found beneficial for future consultations.

Results

Overview

In this study, a total of 101 patients were screened and evenly divided into 2 groups: 50 in the intervention group and 51 in the control group. Importantly, no patients were lost to follow-up during the course of the study, ensuring a comprehensive dataset for analysis. The preintervention characteristics, including patient demographics, case complexity, and initial complaints, exhibited no significant differences between the 2 groups, establishing a strong baseline for comparison.

Demographics

Of the 101 patients screened, 50 were assigned to the intervention group and 51 to the control group. All were included in the study as none were lost to follow-up. Preintervention patient demographics, case complexity, and initial complaints did not differ between the groups (Table 1).

Differential Diagnosis

When assessing the primary outcome, the percentage of correct DDs, it was observed that the intervention group, which used DIANNA, outperformed the control group. The intervention

XSL•FO
RenderX

group achieved a mean of 79.3% accuracy in DDs, with a SD of 24%, while the control group had a mean accuracy of 70.5%, with a SD of 33%. This difference was statistically significant with a *P* value of .01, highlighting the effectiveness of DIANNA in improving the accuracy of DDs.

Furthermore, the analysis considered the impact of case complexity on DD accuracy. It was found that DIANNA consistently exhibited benefits across all levels of case complexity. Specifically, for low-complexity cases (1-2 DDs possible), the intervention group displayed an 8% improvement over the control group (*P*=.08). In intermediate-complexity cases (3 DDs), the improvement was even more pronounced at 17% (*P*=.03), and for high-complexity cases (4-5 DDs), the impact was substantial, though not statistically significant, at 15% (*P*=.91).

In addition to these findings, DIANNA demonstrated an overall DD accuracy of 72% for the entire cohort, with notably higher accuracy for low-complexity cases at 88%, followed by 84% for moderate-complexity cases, and 75% for high-complexity cases.

The accuracy of the intervention was not superior in comparison to the control group for low-complexity cases (*P*=.08) and for high-complexity cases (*P*=.92; Table 2).

In Table 2, the outcome variable for the multivariate analysis is the use of DIANNA. The following covariates were included in the analysis:

1. Global analysis, low complexity, and high complexity subgroups: number of diagnoses, chief resident's years of experience, and internal resident's evaluation.
2. Intermediate complexity subgroup: chief resident's years of experience and internal resident's evaluation (excluding the number of diagnoses).

Table 2. Percentage of correct differential diagnoses per group.

	DIANNA (n=50)		Control group (n=51)			*P* multivariate
	Mean (SD)	Range	Mean (SD)	Range	*P* univariate	
DD accuracy (%)	79.3 (24)	25-100	70.5 (33)	0-100	0.13	.01
Low complexity (1-2 DD)	88 (22)	50-100	79.7 (33)	0-100	0.26	.08
Moderate complexity (3 DD)	68.6 (22)	33.3-100	51.5 (27)	0-100	0.10	.03
High complexity (4-5 DD)	75 (26)	25-100	60 (30)	20-100	0.31	.91

Patients and Physicians' Satisfaction

The study assessed patient and physician satisfaction with DIANNA, demonstrating high levels of approval from both groups. Patients reported good overall satisfaction with the clarity and relevance of the questions, rating them 3.6 (SD 0.5) on a 5-point scale. In addition, 82% of patients expressed a desire to retain a summary of their DIANNA results, indicating its perceived value as a useful reference for future health care interactions.

Physician feedback further reinforced DIANNA's clinical utility. The tool was reported to aid in establishing DDs for 26% of RPs, while 82% of residents considered the DDs generated by DIANNA to be comprehensive. In addition, 70% of physicians expressed a willingness to incorporate DIANNA into their future clinical practice, citing its relevance and efficiency in the diagnostic workflow.

These findings suggest that DIANNA not only supports diagnostic accuracy but also enhances physician efficiency and patient engagement. The high satisfaction rates among both patients and physicians indicate that DIANNA has the potential to improve clinical decision-making and contribute to enhanced health care delivery in emergency outpatient settings (Table 3).

XSL•FO
RenderX

Table 3. Satisfaction outcomes.

Satisfaction outcomes	Values
AMHTD[a] suggested DD[b], mean (SD); range	
DD accuracy	81.6 (30); 0-100
Low-complexity index (1-2 DD)	88 (26.1); 0-100
Moderate-complexity index (3 DD)	84.3 (29.1); 0-100
High-complexity index (4-5 DD)	52.1 (31.3); 0-100
Patient's feedback	
I am able to answer correctly, mean (SD); range	3.9 (0.3); 3-4
Symptomatology accurately described, mean (SD); range	3.6 (0.5); 2-4
Length adequate, mean (SD); range	3.6 (0.5); 2-4
DIANNA[c] was relevant to my problem, mean (SD); range	3.5 (0.6); 1-4
Wish to fulfil DIANNA at home, n/N (%)	41/101 (41)
Wish to keep DIANNA summary, n/N (%)	82/101 (82)
Physician's feedback	
AMHTD helps finding DD, n/N (%)	13/50 (26)
DIANNA DD is exhaustive, n/N (%)	41/50 (82)
DD clearly presented, mean (SD); range	3 (0.9); 1-4
DIANNA is pertinent, mean (SD); range	3 (0.7); 2-4
DIANNA saves time, mean (SD); range	2.4 (0.7); 1-4
Would you use DIANNA in clinical practice?, n/N (%)	35/50 (70)

[a]AMHTD: automated medical history-taking devices.

[b]DD: differential diagnosis.

[c]DIANNA: diagnosis and anamnesis.

Discussion

Principal Findings

Our findings suggest that while DIANNA demonstrated significant potential in enhancing DDs, its effectiveness varied depending on case complexity. The results indicate that DIANNA improved diagnostic accuracy compared with traditional methods, particularly in intermediate-complexity cases, where a statistically significant improvement was observed. This highlights its potential role in enhancing clinical decision-making.

Furthermore, both patients and physicians expressed high levels of satisfaction with DIANNA. A noteworthy percentage of patients indicated a desire to retain a copy of their DIANNA summary for future reference, emphasizing its perceived value in enhancing patient involvement in their own care. However, despite these promising results, further validation is necessary to fully assess DIANNA's clinical impact, particularly in more complex cases where its effectiveness was less pronounced.

Comparison to Prior Work

This study builds upon previous research on automated medical history-taking devices and reinforces the growing recognition of DHHDs as valuable tools in modern health care. The introduction of a body pictogram feature within DIANNA (Figure 2) marks a notable advancement, streamlining the diagnostic process and improving data collection accuracy.

In line with our findings, the existing literature strongly supports the growing role of automated medical history-taking devices in health care. These tools have been shown to:

1. Enhance diagnostic accuracy by reducing physician bias and ensuring structured history-taking [11,12].
2. Improve patient engagement by allowing individuals to actively contribute to their medical history documentation [13].
3. Expedite data collection through structured, AI-assisted questioning that optimizes the consultation process [14].

Moreover, the integration of AI and machine learning into DHHDs represents a major advancement in health care. Unlike standalone AI diagnostic tools, DIANNA functioned as a complementary system, working in conjunction with clinical assessments conducted by chief residents. This hybrid human-AI collaborative model reflects a synergistic approach to health care decision-making, where technology supports but does not replace physician expertise.

XSL•FO
RenderX

Figure 2. Body pictograms feature in DIANNA.

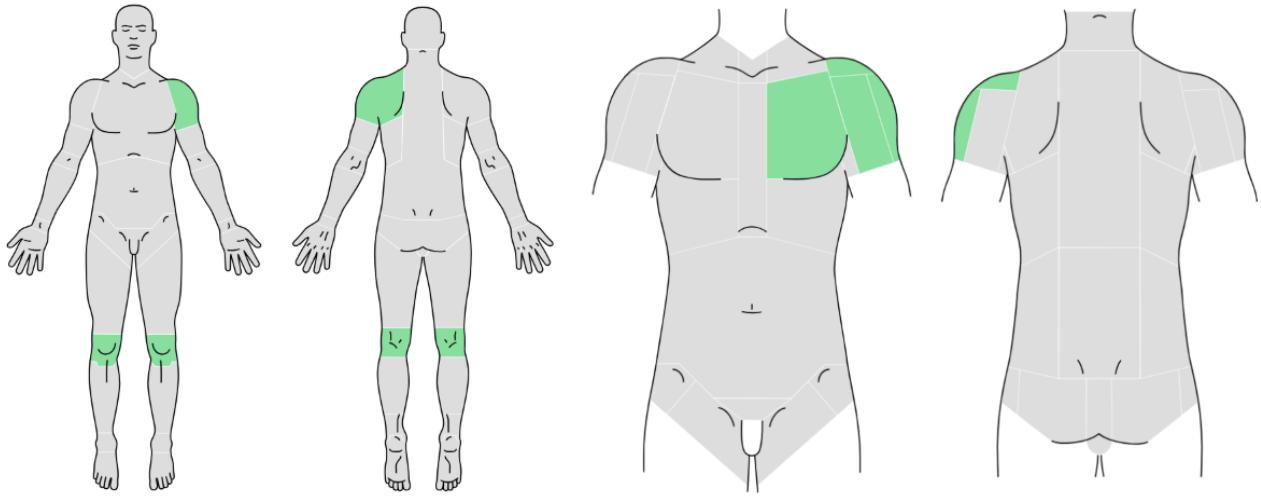

Strengths and Limitations

Overview

This study provides valuable insights into the evolution of DHHDs and the continued development of DIANNA. The enhanced version of DIANNA, featuring a body pictogram, represents a significant innovation in optimizing the diagnostic process and improving patient care. By providing structured and standardized history-taking, DIANNA facilitated symptom localization and enhanced DD accuracy, particularly in intermediate-complexity cases.

A major strength of this study is its hybrid human-AI collaboration model, where DIANNA functioned as a decision-support tool rather than an autonomous diagnostic system. This approach ensured that physician expertise remained central to decision-making, with AI augmenting rather than replacing clinical judgment. Furthermore, the study's real-world application in an emergency outpatient setting highlights its practical usability and feasibility in high-demand clinical environments. Both patients and physicians reported high levels of satisfaction, with many patients expressing interest in retaining a copy of their DIANNA summary for future reference.

Despite these strengths, several limitations must be acknowledged.

Single-Center Study

The study was conducted at a single health care center, which may limit the generalizability of its findings to other clinical settings with different patient populations, staffing structures, and medical infrastructures. While efforts were made to contextualize the results by comparing them with existing literature on AI-assisted history-taking tools, the study's applicability remains largely confined to settings like the University Hospitals Geneva emergency outpatient department. Future studies should address this limitation by conducting multicenter trials across diverse health care settings, ensuring that DIANNA's effectiveness is evaluated in varied clinical environments.

Small Sample Size

With only 101 patients included, the study's statistical power was limited, increasing the risk of type II errors, where real effects may have gone undetected. The relatively small sample size may also reduce the reliability of subgroup analyses, particularly for high-complexity cases, where larger datasets are necessary to draw meaningful conclusions. While resource constraints and feasibility considerations influenced the study design, expanding sample sizes in future studies will enhance statistical power and improve the ability to detect smaller effect sizes.

Pseudorandomized Study Design

Patients were assigned sequentially rather than through pure randomization, introducing the possibility of selection bias. Although pseudorandomization was a practical necessity, baseline characteristics were carefully assessed to ensure comparability between the control and intervention groups. However, some systematic differences may have persisted, potentially influencing the results. Future research should use fully randomized controlled trials to provide stronger causal inferences about DIANNA's impact and further eliminate potential bias in patient allocation.

Impact of COVID-19 on Recruitment

The study's recruitment process was affected by interruptions due to COVID-19–related restrictions, leading to variability in patient selection and flow. These disruptions may have introduced bias by altering the demographics and clinical characteristics of participants, as health care–seeking behaviors changed during the pandemic. Efforts were made to resume recruitment as soon as feasible while maintaining consistent eligibility criteria, but the possibility of pandemic-related effects on diagnostic patterns cannot be entirely excluded. Future studies should aim for uninterrupted recruitment periods or consider external factors influencing patient presentation when analyzing results.

Short-Term Outcomes Without Long-Term Follow-Up

The study focused primarily on immediate diagnostic accuracy and consultation efficiency, without evaluating long-term patient outcomes such as treatment adherence, health care usage, or

XSL•FO
RenderX

sustained clinical effectiveness. While this short-term focus was an intentional choice due to feasibility constraints, it limits the study's ability to determine whether AI-assisted history-taking translates into improved long-term patient care. Future research should incorporate longitudinal follow-up to assess whether DIANNA's benefits persist over time and contribute to better clinical outcomes and patient safety.

Use of a SP's Diagnosis as the Gold Standard

The study relied on a single SP's diagnosis as the gold standard for comparison, which introduces potential subjectivity in diagnostic validation. Although assessments were cross-referenced with clinical guidelines, interphysician variability could still influence diagnostic accuracy measurements. This potential bias highlights the need for multiple independent reviewers or consensus panels in future studies to improve reliability and objectivity in evaluating diagnostic accuracy.

Potential Learning Effect Among Physicians

Physicians using DIANNA may have become more familiar with its output over time, leading to improved diagnostic performance as they adapted to the tool. This learning effect could have inflated DIANNA's observed effectiveness, particularly in cases where physicians refined their ability to interpret and integrate AI-generated suggestions. Although pseudorandomization helped minimize bias, future studies should consider using crossover designs, where the same group of physicians rotates between using DIANNA and standard history-taking methods, to control for potential experience-related improvements.

Future Directions

Future research should focus on large-scale, multicenter trials to enhance the generalizability of DHHDs such as DIANNA across diverse health care settings. Beyond immediate diagnostic accuracy, studies should assess long-term patient outcomes, including treatment adherence and clinical effectiveness. Seamless integration with electronic health records will be essential for optimizing usability and supporting real-time decision-making.

Data security, privacy, and regulatory compliance must also be prioritized to ensure ethical implementation. Comparative studies between different DHHDs will help establish best practices and refine standardized guidelines. As AI technology continues to advance, DHHDs have the potential to improve diagnostic precision, enhance patient engagement, and streamline clinical workflows.

The future of health care will likely be shaped by hybrid models where AI supports, rather than replaces, physician expertise. Ensuring a balanced collaboration between AI-driven tools and clinicians will be key to maximizing their impact on patient care.

Conclusion

The findings of this study reinforce the transformative potential of DHHDs such as DIANNA in clinical practice. While AI-driven medical history tools are still evolving, their integration into hybrid health care models holds promise for enhancing diagnostic accuracy, improving efficiency, and strengthening patient-centered care. Future research should focus on longitudinal validation, multicenter trials, and integration with electronic health records to optimize the clinical utility of AI-assisted history-taking tools.

Acknowledgments

The authors acknowledge Beatriz Villars for her contribution to project administration and logistical support (Division of Primary Care Medicine, Department of Community Medicine, Primary Care and Emergency, Department of Medicine, Geneva University Hospitals). Generative artificial intelligence was used to assist with grammar and language refinement in this manuscript but did not influence the conceptual content, analysis, or conclusions of the paper. All intellectual contributions and interpretations remain those of the authors.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request to Adrien.schwitzguebel@gmail.com.

Authors' Contributions

BH contributed to writing–original draft, and visualization. HS contributed to conceptualization, methodology, investigation, supervision, writing – review & editing, data curation. AS contributed to formal analysis, software, writing – review & editing.

Conflicts of Interest

AS is a partner in the limited liability company that own the DIAANA DHHD.

References

1. Smith J. Evaluating the effectiveness of digital health history devices in clinical decision-making: a randomized controlled trial. J Med Inform 2021;28(4):452-460.

XSL•FO
RenderX

2. Nguyen T. The role of digital health technologies in improving healthcare delivery: a systematic review. Healthc Technol Lett 2020;7(5):1-9.
3. Brown A. Impact of digital health tools on healthcare efficiency: a review of the literature. J Health Inform 2022;29(3):55-63.
4. Jones L. Enhancing patient-centered care through digital health: a study on the usability of digital health history devices. J Digit Health 2023;9(2):101-109.
5. Kumar P. The integration of pictograms in digital health devices: improving patient interaction and symptom reporting. J Med Technol 2022;6(1):15-21.
6. Choi E. Predictive modeling in healthcare: a review of methods, techniques, and applications. J Healthc Eng 2021:1-1.
7. Lima R. Enhancing healthcare outcomes with artificial intelligence: current trends and future directions. AI in Healthcare 2022;1(2):65-78.
8. Harada Y, Tomiyama S, Sakamoto T, Sugimoto S, Kawamura R, Yokose M, et al. Effects of combinational use of additional differential diagnostic generators on the diagnostic accuracy of the differential diagnosis list developed by an artificial intelligence-driven automated history-taking system: pilot cross-sectional study. JMIR Form Res 2023;7:e49034 [FREE Full text] [doi: 10.2196/49034] [Medline: 37531164]
9. Kanazawa A, Fujibayashi K, Watanabe Y, Kushiro S, Yanagisawa N, Fukataki Y, et al. Evaluation of a medical interview-assistance system using artificial intelligence for resident physicians interviewing simulated patients: a crossover, randomized, controlled trial. Int J Environ Res Public Health 2023;20(12):6176 [FREE Full text] [doi: 10.3390/ijerph20126176] [Medline: 37372762]
10. Schwitzguebel AJ, Jeckelmann C, Gavinio R, Levallois C, Benaïm C, Spechbach H. Differential diagnosis assessment in ambulatory care with an automated medical history-taking device: pilot randomized controlled trial. JMIR Med Inform 2019;7(4):e14044 [FREE Full text] [doi: 10.2196/14044] [Medline: 31682590]
11. Semigran HL, Linder JA, Gidengil C, Mehrotra A. Evaluation of symptom checkers for self diagnosis and triage: audit study. BMJ 2015;351:h3480 [FREE Full text] [doi: 10.1136/bmj.h3480] [Medline: 26157077]
12. Zowawi H. Digital history-taking platformsvaluating diagnostic accuracy and user satisfaction. JAMA Network Open 2021;4(6):2345.
13. Yang Y, Henschke C, Richards B, Martin R. Engaging patients through digital history-taking: the future of AI in primary care. Ann Fam Med 2019;17(5):399.
14. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2017;2(4):230-243. [doi: 10.1136/svn-2017-000101] [Medline: 29507784]

Abbreviations

AI: artificial intelligence
DD: differential diagnosis
DHHD: digital health history device
DIANNA: diagnosis and anamnesis
RP: resident physician
SP: senior physician

Edited by A Mavragani; submitted 15.01.24; peer-reviewed by Y Hayat, A Hassan, R Bidkar, J Knitza, N Kolukula; comments to author 08.11.24; revised version received 13.03.25; accepted 08.04.25; published 01.10.25.

Please cite as:
Healey B, Schwitzguebel A, Spechbach H
Differential Diagnosis Assessment in Ambulatory Care With a Digital Health History Device: Pseudorandomized Study
JMIR Form Res 2025;9:e56384
URL: https://formative.jmir.org/2025/1/e56384
doi:10.2196/56384
PMID:40205939

©Beth Healey, Adrien Schwitzguebel, Herve Spechbach. Originally published in JMIR Formative Research (https://formative.jmir.org), 01.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Understanding Experiences of and Unmet Needs in Online Searches for Menopause Information: An Exploratory Survey

Erin Lucy Funnell[1*], BSc; Freya McConnell[1*], BA, MEng; Nayra A Martin-Key[1], PhD; Leyao Qian[2]; Kathryn Babbitt[1], BA; Sabine Bahn[1], MD, PhD

[1]Cambridge Centre for Neuropsychiatric Research, Department of Chemical Engineering and Biotechnology, University of Cambridge, Cambridge, United Kingdom

[2]Department of Psychology, University of Cambridge, Cambridge, United Kingdom

[*]these authors contributed equally

Corresponding Author:
Sabine Bahn, MD, PhD
Cambridge Centre for Neuropsychiatric Research
Department of Chemical Engineering and Biotechnology
University of Cambridge
Philippa Fawcett Drive
Cambridge, CB3 0AS
United Kingdom
Phone: 44 1223 334151
Email: sb209@cam.ac.uk

Abstract

Background: Menopause is a significant time in a woman's life, but only recently has there been an open discussion about it in the media, workplaces, and general society. With increasing frequency, women are using the internet to research menopause, making it essential that online sources provide safe, high-quality, and relevant information.

Objective: This study aimed to investigate the current state of the online information landscape for menopause from the perspective of information seekers, exploring (1) information-seeking behavior and (2) perceptions of online resources for menopause.

Methods: A 10- to 15-minute online survey was conducted asking about the respondents' use of and opinions about online resources specifically for menopause. We distributed the survey via social media, email, and word of mouth. Quantitative data were explored using means and frequencies. Group differences between menopausal groups were analyzed using chi-square, Fisher exact, or Kruskall-Wallis tests as appropriate. Qualitative data were analyzed using data-driven thematic analysis.

Results: Data from 627 participants were analyzed (early perimenopause: n=171, 27.3%, late perimenopause: n=125, 19.9%, postmenopause: n=262, 41.8%, and surgical menopause: n=69, 11%). The majority of respondents had used the internet as a source of information (581/627, 92.7%), with the internet being the first choice of information source (489/581, 84.2%). The most searched-for information online was about menopause symptoms (479/581, 82.4%), menopause treatment options (442/581, 76.1%), and self-help tips or strategies (318/581, 54.7%). The majority of participants trusted online information to some extent (615/627, 98.1%), with many also considering online information accurate to some extent (555/627, 88.5%). Many participants reported finding some but not all of the information they were looking for online (379/581, 65.2%). Thematic analysis revealed 10 themes related to information quality and accessibility and sought-after information (eg, symptom specifics, treatment, and nonformal management strategies). Analysis also indicated that information is lacking for several groups, including those in medically induced or surgical menopause.

Conclusions: The study showed that online informational resources are widely accessed and widely perceived as useful and trustworthy. However, it is crucial that the quality of online information is evaluated, especially considering the large number of users who rely on it as their first or only informational source. Online searches were usually performed to find information related to symptoms, treatment, and self-help recommendations, with differences in search behaviors observed across menopausal stages and groups, highlighting the need for tailored informational resources. Thematic analysis revealed gaps in the provision of online information both in terms of content and quality. Participants noted a lack of comprehensive symptom information, inadequate information for groups such as those experiencing medical or surgical menopause, and concerns about outdated content and a

XSL•FO
RenderX

lack of source transparency. Future research with more diverse samples is needed to better understand variations in online health information-seeking behaviors across groups.

(JMIR Form Res 2025;9:e75335) doi:10.2196/75335

KEYWORDS

menopause; perimenopause; online; internet; information

Introduction

Background

Menopause is defined as the time at which there has been no menstrual period for 12 months. Menopause may be naturally occurring or may be medically or surgically induced [1]. Natural menopause occurs due to the decline of ovarian follicular function, typically occurring between the ages of 45 and 55 years [2]. It is important to note that menopause can occur at any age, and the age at which a specific woman experiences menopause cannot be predicted.

Experiencing menopause before the age of 40 years is known as premature menopause and may be caused by several physical conditions, although in some women causal factors cannot be identified [2]. The perimenopause, or menopause transition, is the period before menopause where the frequency or duration of menstruation may change, and menopause symptoms may be present. On average, this occurs 4 years before menopause [3]. Menopause and perimenopause are associated with a wide range of both physical and mental health symptoms, the severity of which varies from person to person. Symptoms are wide-ranging, with some more well-known than others. Vasomotor symptoms, including hot flashes and night sweats, are common and are seen in as many as 80% of menopausal individuals [4]. The effects of menopause can cause huge difficulties in multiple domains of daily life, including professionally, personally, and with regard to health and well-being. As many as 31% of women have to take time off work due to their symptoms [5], with some having to give up work completely. In addition, 74% of women blame menopause for the breakdown of their marriage, and 67% of women report an increase in arguments and domestic abuse in relation to menopause [6].

It is clear that menopause is complex, potentially resulting in a variety of interpersonal problems, health difficulties, and poorer well-being. Menopause and the menopause transition require decisions to be made about when to reach out for help and which treatment intervention to use. Previous work has found that women feel unprepared for menopause and think that the early education of women and the wider society about menopause is lacking [7]. Research has also demonstrated that technology is perceived as a beneficial resource to support individuals facing menopause, being broadly used for 4 health-related purposes: connecting and sharing, information seeking, monitoring and reflection, and self-care [8]. Research also demonstrates the widespread use of the internet to gather information related to women's health, with data from the United Kingdom revealing that 71% of women choose to use Google to find information about menopause, and 69% of women use other search engines or blogs as an additional source of information [9]. This is compared to only 59% of women who consult their general practitioner (GP) or other health care professionals (HCPs) to obtain this information. This highlights the need for safe, relevant, comprehensive, and clinically accurate information to be easily found online [9]. This widespread use of the internet as a source of information is also observed specifically for information about menopause, with a reported use rate of 74% in previous work [10].

To date, research has been focused on which informational sources women are using, online or otherwise, with limited research into the specifics of the information being sought. Some previous research identified the need for information regarding symptoms and treatment information [11,12], as well as the risks of secondary health conditions, such as cancer [12]. To our knowledge, there have been no studies addressing the specific informational needs of menopausal women in the United Kingdom when searching online. This is important, as the Menopause Priority Setting Partnership Steering Group has highlighted the need for research to identify ways to help people prepare for the perimenopause and menopause, as well as to develop awareness of when to seek help, with one potential avenue to this improved preparation and knowledge being online resources [13].

Objectives

This study focuses on this knowledge gap, as without this information, the creation of useful online informational resources for menopausal women is not feasible. More specifically, the aims of this study were to (1) explore the information-seeking behavior of women in menopause and the menopause transition and (2) explore the perceptions of online informational resources for menopause. These insights will also be valuable in the design of content and delivery of future online resources. Relevancy of information is potentially as important as its quality, and what is classed as relevant may vary in women who are at different stages of menopause. Therefore, this study also considered the influence of specific menopausal stages on online searching behaviors, and perceptions of available online information resources.

Methods

Overview

The research team comprised a psychology and behavioral sciences undergraduate student (LQ), a master's-level chemical engineering and biotechnology student (FM), 2 research assistants (ELF and KB), a senior research associate (NAM-K), and a practicing psychiatrist and professor of neurotechnology (SB). All authors involved in the qualitative analysis identify as female. All authors involved in the qualitative analysis

XSL•FO
RenderX

received training in or have experience conducting thematic analysis.

The Checklist for Reporting Results of Internet E-Surveys for the current manuscript is provided in Multimedia Appendix 1.

Materials

A novel 10- to 15-minute online open survey was delivered using Qualtrics XM (Qualtrics International Inc). The survey consisted of approximately 77 questions; the flow of questions was adaptive, dependent on previous answers. All but 3 free-text questions were closed questions and had prespecified answer options, including single choice, multiple choice, and Likert scale formats. A single question is presented per page of the survey. All questions had a back button, meaning that participants could rereview and amend their answers if necessary. All closed questions were marked as required preventing participants from proceeding to the next question without entering an answer. Free-text questions were optional.

Questions were developed by FM with input from ELF and NAM-K, with some questions, particularly those related to sources of information and information sought, adapted from other similar studies [7,12,13]. All questions were reviewed and approved by a practicing psychiatrist (SB). All aspects of the study, including the consent form, participant information sheet, and questionnaire, were completed in the same online survey. The survey comprised quantitative and qualitative questions on sociodemographic characteristics (eg, age, gender, and ethnic group), menopause symptom presence and severity, and online informational source usage (ie, participants' experiences with and interest in using online informational sources for menopause). Questions about online informational source use focused on what type of information women search for online about menopause (eg, treatment options, health risks, or more general advice) and what type of source they feel most comfortable using. Motives behind using the internet to obtain information were explored, along with views on the currently available online resources. Usability testing (ie, testing all the adaptive logic, instruction clarity, response option functionality, and overall user experience) of the Qualtrics survey was performed by FM, ELF, and NAM-K. This process also provided an estimate of the time required to complete the survey, which was used to inform participants during the recruitment process. Responses to the consent form and survey were automatically saved by the Qualtrics software, and completeness checks (ie, percentage of the survey answered) were also completed by the survey software. For the current analysis, only questions related to sociodemographic characteristics and online informational source use were included. To see the survey questions, refer to Multimedia Appendix 2.

Participants

Recruitment was carried out between January 5, 2024, and January 15, 2024, using digital convenience sampling. Participants were required to be aged ≥18 years, living in the United Kingdom, not have been pregnant or breastfeeding in the past 12 months, and confirm that they were in menopause or the menopause transition and experiencing at least 1 related symptom (eg, irregular or absent periods, hot flushes, and night sweats). Participants were required to self-define which stage of menopause or the menopause transition they were in. Definitions of different menopause stages provided in the survey were based on the Staging of Reproductive Aging Workshop guidelines [14,15]. Definitions of menopause stages provided in the survey were as follows: early perimenopause (significant change in the length or regularity of menstrual bleed not due to pregnancy, breastfeeding, stress, or a medical condition), late perimenopause (no menstrual bleeding in the past 3-11 months not due to pregnancy, breastfeeding, stress, or a medical condition), postmenopause (no menstrual bleeding in the past 12 months not due to pregnancy, breastfeeding, stress, or a medical condition), and surgical menopause (no bleeding in the past 12 months due to hysterectomy with 1 or 2 ovaries retained, or another medical intervention).

Survey Distribution

The survey was disseminated using paid advertisements on social media, word of mouth, and emails sent to individuals who had previously participated in research carried out by the Cambridge Centre for Neuropsychiatric Research and had expressed interest in taking part in future studies. Recruitment materials specified that the study was focused on views of online resources for menopause. Paid advertisements via social media were delivered to specific populations based on self-reported gender and age to maximize the relevance of the recruitment materials to the target demographic.

Statistical Analysis

Quantitative data were analyzed in Excel (version 2407; Microsoft Office 365) and SPSS (version 29.0.1.1; IBM Corp). Comparisons between binary categorical variables were analyzed with the chi-square or Fisher exact test for low-frequency data (<5 occurrences). Effect sizes are reported as Cramér V (ϕ_c; small=0.10, medium=0.30, and large=0.50) [16]. For ordinal data with a Likert scale (ie, perceived accuracy and perceived trustworthiness), Kruskal-Wallis *H* tests were used, with post hoc Dunn tests subject to the Bonferroni correction for multiple comparisons conducted where appropriate.

The framework by Braun and Clarke [17] for thematic analysis was used for analysis of the open text data. A data-driven approach was adopted. Any free-text data that included "no" (or synonyms), "not applicable" (or synonyms), or responses that were considered ambiguous or nonspecific by both authors were labeled as not applicable and excluded from thematic analysis.

Before writing the codebooks, 1 author (ELF) read the open text data multiple times to gain familiarity with the data. During the codebook writing process, any responses that contained null data (eg, "no," "none," "N/A," or synonyms) were labeled as "not applicable" by ELF. The codebook was reviewed by a second author (NAM-K), who gave feedback to maximize the comprehensiveness of the codebook. Open text data were coded against the codebook under blinded conditions by 2 authors (FM and LQ) and reviewed by a third author (ELF). While data were being coded, free-text responses were reviewed independently by 2 authors (FM and LQ) to determine their

XSL•FO
RenderX

suitability for thematic analysis. Responses were labelled as "not applicable" and excluded if they met any of the following predefined criteria: (1) were off-topic or unrelated to the survey prompt or (2) were ambiguous, overly vague, or lacked sufficient detail for meaningful interpretation without avoiding the potential for assumptions to be made by the research team regarding the meaning (eg, "can't say" and "unsure"). Responses were removed only when both authors agreed that they met the exclusion criteria. As part of criteria for exclusion number 1, positive responses such as "I found enough..." (n=9) were excluded because, given the negatively framed prompt, similar positive experiences may have gone unreported and including only a few such responses could introduce inconsistency. The coding allocations were unblinded and compared, with any discrepancies being discussed until an agreement was reached. After the coding allocation was complete and finalized, the codes were organized into themes by 2 authors (FM and LQ) under blinded conditions. Each author reviewed the codes and grouped them into initial themes based on patterns and conceptual similarity. The theme lists generated were then compared, with any discrepancies being discussed (FM, LQ, ELF, and KB) until a final theme list was agreed upon. This process involved iterative discussions to refine the scope and definitions of each theme, ensuring that themes were clearly defined, distinct, and grounded in the data. Please note that the illustrative quotes provided as examples of any codes identified have not been revised in any way. Therefore, any spelling or grammatical errors are as provided by the participants.

Ethical Considerations

Ethics approval was awarded by the Cambridge Psychological Research Ethics Committee (PRE.2023.124). Participants were required to give informed consent to participate electronically before the main body of the survey began. Participation was completely voluntary. Data collected were anonymized using randomly generated participant IDs by the survey software and stored on password-protected devices. In exchange for participation, participants were invited to enter a draw for the chance to win one of three £50 (US $64.72) gift cards.

Results

Survey Uptake

A total of 1843 responses were recorded for the survey. In total, 774 participants consented to take part in the study, with 628 complete responses (ie, 100% completion rate). One response was omitted from further analysis because the respondent disclosed that they had given false demographic information, leaving a final sample of 627 participants (view rate: 627/1843, 34.02%, completion rate: 627/774, 81%, early perimenopause: 171/627, 27.3%, late perimenopause: 125/627, 19.9%, postmenopause: 262/627, 41.8%, and surgical menopause: 69/627, 11%).

Demographics

Full sociodemographic characteristics of the overall sample are reported in Table 1 (sociodemographic characteristics for each of the menopause stages can be found in Table S1 in Multimedia Appendix 3). The average age of participants was 52.40 (SD 4.67; range 31-71) years. The majority of the sample were White (594/627, 94.7%), identified as female (617/627, 98.4%), were in paid employment (446/627, 71.1%), and were married or in a civil partnership (379/627, 60.5%). Moreover, 48.8% (n=306) held at least an undergraduate degree, 42.4% (n=266) had an annual household income before tax of at least £45,001 (US $60,232.50), and 40.34% (n=253) of the participants reported having preexisting conditions with symptoms that may overlap with the symptoms of menopause.

XSL•FO
RenderX

Table 1. Sociodemographic characteristics (N=627)[a].

Sociodemographics	Respondents, n (%)
Gender	
Woman	617 (98.4)
Man	2 (0.3)
Nonbinary	1 (0.2)
Other	5 (0.8)
Prefer not to answer	2 (0.3)
Ethnicity	
Asian or Asian British	14 (2.2)
Black, Black British, Caribbean, or African	16 (2.6)
Mixed or multiple ethnic groups	3 (0.5)
White	592 (94.7)
Marital status	
Single	86 (13.7)
Married or civil partnership	379 (60.4)
Cohabiting	74 (11.8)
Separated	23 (3.7)
Divorced	41 (6.5)
Other	18 (2.9)
Prefer not to answer	6 (1)
Education	
Below GCSE[b] or equivalent	37 (5.9)
GCSEs, Scottish Higher, or equivalent	106 (16.9)
A-Levels[c], International Baccalaureate, or equivalent	133 (21.2)
Undergraduate degree	152 (24.2)
Postgraduate degree	154 (24.6)
Other	26 (4.2)
Prefer not to answer	19 (3)
Employment[d]	
Employed full-time	289 (46.1)
Employed part-time	149 (23.8)
Self-employed	71 (11.3)
Parental leave	26 (4.2)
Student	4 (0.6)
Retired	39 (6.2)
Volunteer	14 (2.2)
Unemployed	44 (7)
Prefer not to answer	11 (1.8)
Annual household income (before tax; £ [US $])	
<15,000 (US $20,226)	49 (7.8)
15,001-25,000 (US $20,227.35-$33,710.00)	69 (11)
25,001-35,000 (US $33,711.35-$47,194.00)	73 (11.6)

XSL•FO
RenderX

Sociodemographics	Respondents, n (%)
35,001-45,000 (US $47,195.35-$60,678.00)	73 (11.6)
45,001-55,000 (US $60,679.35-$74,162.00)	69 (11)
55,001-65,000 (US $74,163.35-$87,646.00)	49 (7.8)
65,001-75,000 (US $87,647.35-$101,130.00)	37 (5.9)
75,0001-85,000 (US $101,131.35-$114,614.00)	33 (5.3)
>85,001 (US $114,615.35)	78 (12.4)
Prefer not to answer	97 (15.5)

[a]Total percentage may exceed 100%, as participants were able to select >1 type of employment.

[b]GCSE: General Certificate of Secondary Education.

[c]A-Level: Advanced Level.

[d]Total percentage may exceed 100%, as participants were able to select >1 type of employment.

Use of the Internet

The vast majority of respondents had used the internet as a source of information (overall: 581/627, 92.7%, early perimenopause: 162/171, 94.7%, late perimenopause: 112/125, 89.6%, postmenopause: 242/262, 92.4%, and surgical menopause: 65/69, 94%). Of those who did not use the internet as a source of information (overall: 46/627, 7.3%), the most common reason for not using the internet provided was not knowing where to start looking (17/46, 40%; Table S2 in Multimedia Appendix 3). Figure 1 provides a summary of data related to the focus of searches for information, the perception of online information, satisfaction with online information, and perceived missing information.

Figure 1. Summary of data related to the most frequently searched for information, the perception of information available online, the perceived success in online information seeking, and missing information. Missing information is drawn from the qualitative data (N=394) and reflects the most frequently identified codes in the dataset. HRT: hormone replacement therapy.

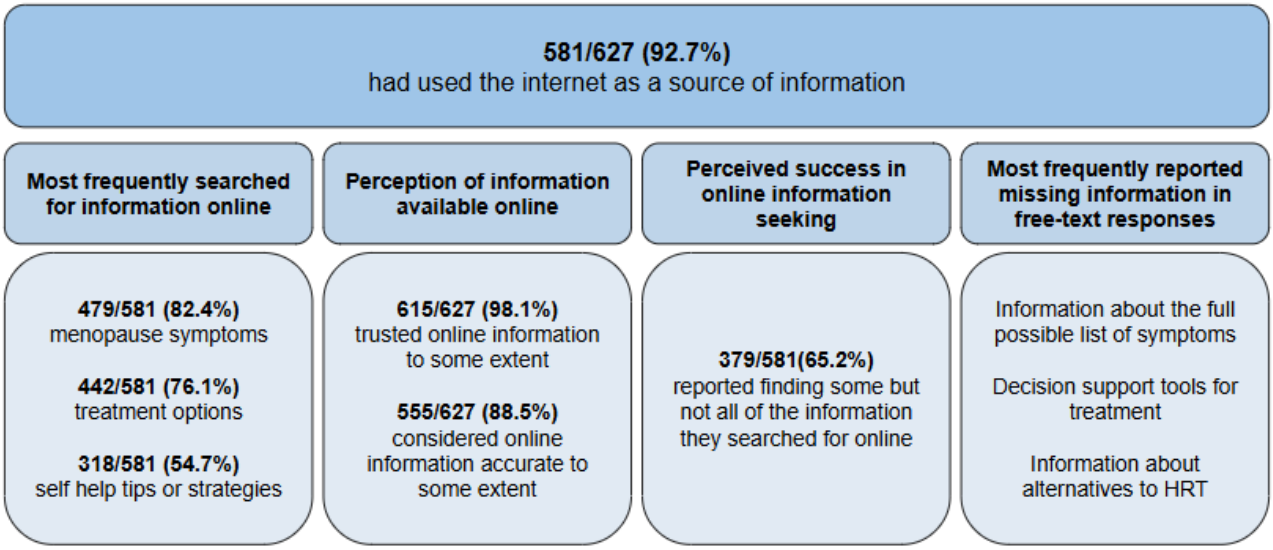

Of those who endorsed looking for information on the internet, the majority used the internet as their initial source of information (overall: 489/581, 84.2%, early perimenopause: 144/162, 88.9%, late perimenopause: 88/112, 78.6%, postmenopause: 204/242, 84.3%, and surgical menopause: 53/65, 82%). Table 2 provides a summary of informational resources accessed online. No group differences were observed regarding the type of information resources accessed. Informational resources on the internet were most frequently found via a Google Search (511/581, 88%).

XSL•FO
RenderX

Table 2. Information sources accessed online. Total percentages may exceed 100% because multiple information sources could be selected. Percentages were calculated by dividing the frequency of each code by the total number of responses (overall: n=581, early perimenopause: n=162, late perimenopause: n=112, natural menopause: n=242, and surgical menopause: n=65).

Information source	Overall, n (%)	Early perimenopause, n (%)	Late perimenopause, n (%)	Natural menopause, n (%)	Surgical and medical menopause, n (%)
NHS[a] online and NHS inform	447 (76.9)	130 (80.3)	82 (73.2)	180 (74.4)	55 (84.6)
Websites reporting other people's experiences	186 (32)	56 (34.6)	38 (33.9)	76 (31.4)	16 (24.6)
Websites suggesting in-person support options	103 (17.7)	27 (16.7)	35 (22.3)	40 (16.5)	11 (16.9)
Charities	262 (45.1)	80 (49.4)	53 (47.3)	96 (39.7)	33 (50.8)
Web sites aimed at HCPs[b]	143 (24.6)	43 (26.5)	24 (21.4)	56 (23.1)	20 (30.8)
Social media	359 (61.8)	98 (60.5)	80 (71.4)	145 (59.9)	36 (55.4)
Scientific literature	73 (12.6)	18 (11.1)	9 (8)	33 (13.6)	13 (20)
Online news sites with reports about menopause	205 (35.3)	55 (34)	37 (33)	86 (35.5)	27 (41.5)
Other	44 (7.6)	13 (8)	10 (8.9)	17 (7)	4 (6.2)

[a]NHS: National Health Service.

[b]HCP: health care professional.

Other Information Sources

Most online information seekers also sought information elsewhere (overall: 439/581, 75.6%), with a similar proportion between menopause stages (early perimenopause: 126/162, 77.8%, late perimenopause: 88/112, 78.6%, postmenopause: 174/242, 71.9%, and surgical menopause: 51/65, 78%).

Most of these had used a secondary source after using the internet as the primary source of menopause information (overall: 353/581, 60.8%, early perimenopause: 110/162, 67.9%, late perimenopause: 65/112, 58%, postmenopause: 139/242, 57.4%, and surgical menopause: 39/65, 60%; Table S3 in Multimedia Appendix 3). The most frequently cited reason for using a second source of information after the internet was to obtain as much information as possible (183/353, 51.8%), followed by wanting verification that the information found online was correct (164/353, 46.5%; Table S4 in Multimedia Appendix 3). A higher proportion of the surgical menopause group reported using a secondary source of information, as they could not find the information they were looking for online (9/39, 23%) compared to the early perimenopause group (8/110, 7.3%; χ^2_3=8; P=.045; ϕ_c=0.15). A larger proportion of the late perimenopause group (40/65, 62%) wanted verification that the information found online was correct and trustworthy, compared to the postmenopause group (55/139, 39.6%; χ^2_3=9.8; P=.02; ϕ_c=0.17). A larger proportion of the early perimenopause group considered the information available online overwhelming (45/110, 40.9%) than the postmenopause group (34/139, 24.5%; χ^2_3=8.4; P=.04; ϕ_c=0.15).

Of those who had not used a secondary informational source in addition to the internet, the most frequent reason for not doing so was not thinking symptoms were severe enough for further help-seeking (57/136, 41.9%) and that they found enough information online (55/136, 40.4%; Table S5 in Multimedia Appendix 3). The surgical menopause group had a higher proportion of individuals citing not using additional noninternet sources due to not trusting an HCP (3/14, 21%) compared to the postmenopause group (1/65, 2%; χ^2_3=11.9; P=.008; ϕ_c=0.27).

The source of information most frequently used other than the internet was a GP (353/439, 80.4%; Table S6 in Multimedia Appendix 3). A higher proportion of the surgical menopause group had contacted a gynecologist compared to any other menopause stage (Table 2; χ^2_3=28.6; P<.001; ϕ_c=0.26). Use of a support group was more common in women in the surgical menopause group (χ^2_3=8.6; P=.04; ϕ_c=0.14). The early perimenopause group endorsed consulting a pharmacist less than any other group (Table 2; χ^2_3=10.1; P=.02; ϕ_c=0.15).

Common Online Search Topics

Table 3 shows what information participants most frequently endorsed looking for online (Table S7 in Multimedia Appendix 3). Information about menopause symptoms (479/581, 82.4%), menopause treatment options (442/581, 76.1%), and self-help tips or strategies (318/581, 54.7%) were endorsed as the most frequent online searches. Of the 33 participants who endorsed searching for information about surgical menopause, 28 (86%) were in surgical menopause. Searches related to pregnancy were more frequent in the late perimenopause group (6/112, 5.4%) than in the postmenopause group (3/242, 1.2%; χ^2_3=8.2; P=.01; ϕ_c=0.12). In addition, a larger proportion of the late perimenopause group (21/112, 18.8%) had searched online for information about private care than the postmenopause group (21/242, 8.7%; χ^2_3=8.3; P=.04; ϕ_c=0.120). A larger proportion of the surgical menopause group had searched online for information about surgical menopause (28/65, 43%) compared to all other groups (early perimenopause: 2/162, 1.2%, late

XSL•FO
RenderX

perimenopause: 1/112, 0.9%, and postmenopause: 2/242, 0.8%; χ^2_3=191.1; *P*<.001; ϕ_c=0.57). A higher proportion of participants in the late perimenopause group (37/112, 33%) and postmenopause group had searched online for general advice for menopause (ie, "how to...;" 70/242, 28.9%) compared to the surgical menopause group (9/65, 14%; χ^2_3=8.9; *P*=.03; ϕ_c=0.12).

Table 3. Frequency of topics around menopause searched for (N=581).

Topic searched for	Overall, n (%)
Causes of menopause	69 (11.9)
Menopause treatment side effects	312 (53.7)
Health risks related to menopause	275 (47.3)
Pregnancy and menopause	16 (2.8)
Menopause consequences (eg, "difficulties with...")	283 (48.7)
Menopause symptoms	479 (82.4)
Menopause treatment options	442 (76.1)
Private care options	65 (11.2)
Support groups	117 (20.1)
Surgical menopause	33 (5.7)
General advice (ie, "how to...")	155 (26.7)
Self-help tips or strategies	318 (54.7)
Other	32 (5.5)

Regarding online searching for information about health-related risks related to menopause (Table S8 in Multimedia Appendix 3), the most searched for possible health risk was osteoporosis (188/275, 68.4%). A higher proportion of participants in the surgical menopause group had searched online for information about osteoporosis (32/36, 89%) than participants in the early (52/80, 65%) and the late perimenopause group (31/56, 55%; χ^2_3=12.1; *P*=.007; ϕ_c=0.21). A higher proportion of participants in the late perimenopause group (26/56, 46%) searched online for information about urinary tract infections than the early perimenopause (17/80, 21%) and postmenopause groups (23/103, 22.3%; χ^2_3=13.8; *P*=.003; ϕ_c=0.22).

The treatment information most frequently searched online (Table S9 in Multimedia Appendix 3) was about transdermal hormone replacement therapy (HRT; 289/442, 65.4%), followed by oral HRT (183/442, 41.4%), vaginal HRT (118/442, 26.7%), lifestyle changes (171/442, 38.7%), and testosterone (108/442, 24.4%). A higher proportion of individuals in the surgical menopause group had searched online for information about lubricants (15/47, 32%) than all other groups (early perimenopause: 13/132, 9.8%, late perimenopause: 7/89, 8%, postmenopause: 18/174, 10.3%; χ^2_3=20.1; *P*<.001; ϕ_c=0.213). A higher proportion of individuals in the late perimenopause group had searched online for information about complementary therapies (30/89, 34%) than the early perimenopause (18/132, 13.6%) and postmenopause (34/174, 19.5%) groups (χ^2_3=13.2; *P*=.004; ϕ_c=0.173).

The treatment side effects most frequently searched for online (Table S10 in Multimedia Appendix 3) were weight gain (187/312, 59.9%), mood changes (178/312, 57.1%), and risk of cancer (173/312, 55.5%). Group differences were observed for online searches about irregular bleeding (χ^2_3=17.8; *P*<.001; ϕ_c=0.24), with a higher proportion of participants in early (31/93, 33%) and late (21/58, 36%) perimenopause endorsing searching for online information about irregular bleeding than participants in postmenopause (22/130, 16.9%) or surgical menopause (2/31, 6%) groups. Participants in the early (26/93, 28%) and late (19/58, 33%) perimenopause groups were proportionally more likely to look for information about dizziness as a treatment side effect than those in the postmenopause (18/130, 13.9%) and surgical menopause (9/31, 29%; χ^2_3=13.9; *P*=.003; ϕ_c=0.21) groups.

In terms of general advice searched for online (Table S11 in Multimedia Appendix 3), the majority wanted information about how to ask for support generally (66/155, 42.6%) or information about how to talk to an HCP about menopause (65/155, 41.9%). No group differences were observed for general advice searching online.

Perceptions of the Quality of the Internet as a Source of Information

Most participants perceived information available in online sources as accurate to some extent (555/627, 88.5%), with the largest proportion considering the sources available online to be mostly accurate (304/627, 48.5%; Table S12 in Multimedia Appendix 3). There was a significant difference in the perceived accuracy of online information across menopause stages (H_3=8.827; *P*=.032). Post hoc tests revealed a significant difference in perceived accuracy between the postmenopause and surgical menopause groups (*P*=.043), with those in the postmenopause group considering the information available online to be of a higher accuracy.

XSL•FO
RenderX

Most participants trusted online sources of information to some extent (615/627, 98.1%), with the largest proportion reporting moderate trust (390/627, 62.2%; Table S12 in Multimedia Appendix 3). The online informational resources considered the most trustworthy were the National Health Service (NHS) or NHS Inform websites (490/627, 78.2%) and the websites of charities (401/627, 64%; Table S12 in Multimedia Appendix 3). No significant differences in trustworthiness were observed between menopause stages.

Most participants who had looked online for information related to menopause found some but not all the information they were looking for online (379/581, 65.2%), with only 4.5% (26/581) reporting that they had not found what they were looking for online (Table S3 in Multimedia Appendix 3). No differences were observed in perceived comprehensiveness between the menopause stages.

Qualitative Analysis

In total, 593 qualitative responses were provided. Of these, 199 (33.6%) were labeled as "not applicable" and removed. A final dataset of 394 (66.4%) responses was analyzed. Thematic analysis revealed 68 codes organized into 10 themes (Table S13 in Multimedia Appendix 3 provides the complete list of identified codes with their code frequencies).

Information Quality and Accessibility

Generally, perceptions of the quality of online resources expressed in free-text responses were negative. Participants who commented on information quality indicated that they perceived available information to be generic. In addition, there were concerns about the accuracy of the information available online, with it being perceived as outdated and lacking transparency regarding the source of the information presented. Some participants commented on their perception that some online content frame as menopause information was instead deceptive advertising ("I feel like I get sucked into social media ads targeted at 50+ year old women trying to sell me dubious supplements."). For some participants, this resulted in a sense of distrust of the information available online.

Regarding accessibility, participants commented on a perceived lack of research focused on menopause. There was also a request for improved dissemination of research online ("I'd like to be more familiar with research into the menopause.").

In addition, there were reports that information is inaccessible because it is not written in an easy-to-understand way ("I would like it all explained in a simpler way, as having brain fog made it hard for me to take it in.") or that websites are difficult to use. Similarly, some participants expressed that the information available online was overwhelming. As a potential solution, participants called for "One site that has all the information, instead of being sent from site to site," such as a centralized menopause information hub.

Treatment

Thematic analysis revealed a request for easier-to-understand information and guidance for treatment pathways and menopause interventions. Some participants specifically asked for decision-support tools, such as flowcharts ("Types of HRT—a nice flow-chart to help me choose"). This included a request for information about the safety of menopause management options, including long-term outcomes ("I have very little clue about what my options are together with what the side effects may be for me"). In addition, participants mentioned needing more information about the association between menopause treatments, such as HRT, and preexisting or comorbid conditions, such as cancer ("Specific info & treatments for someone who has had triple negative breast cancer"). Participants also noted the need for diversification of existing treatment guidelines to provide tailored information on menopause management for specific subgroups, such as those experiencing medically or surgically induced menopause ("There is no NICE guidance on risks of HRT post surgery for endometrial cancer and very little research").

Specifically regarding HRT, participants noted that there is misinformation related to HRT online ("Accurate information about the health benefits of [HRT] and removal of the hugely out of date and debunked 'scare stories' about HRT."), revealing a request for clearer and more balanced information. Free-text responses also requested more information related to specific types of HRT, such as testosterone ("More evidence of adding testosterone to the HRT options and it's benefits"), as well as alternatives to HRT.

Free-text responses also requested information regarding alternatives to HRT for those who cannot or do not wish to use HRT ("Treatment other than medication for people who can't use HRT"). Relatedly, other participants sought information about the consequences of not taking any menopause treatment (ie, letting menopause happen "naturally"; "What if you don't do anything at all and let it all happen naturally, what are the consequences of this").

Groups Lacking Information

Free-text data revealed a perception of information missing for certain groups, including HCPs ("Resources to take to your GP if they are not up to date on the menopause."), workplaces ("More online resources for employers to access."), individuals in medical or surgical menopause ("More recognition about surgical menopause and the impact of symptoms"), partners ("Menopause should be taught to both women and men. I had to make my husband read others experiences so that he understand my plight and the help I needed."), and men and boys ("Educate boys and men, it's important"), with requests for informational resources tailored to these groups to maximize their relevance.

Conversely, some participants stated they held a perception that the information online is tailored for specific groups (eg, women who may be considered "healthy" or not have any existing or comorbid conditions), which can make the information feel inaccessible or irrelevant ("Much seemed geared towards 'normal' healthy women—ie, not fat like me, or with pre-existing depression.").

Symptom Specifics

Participants also requested more information about symptoms of menopause. This included the complete list of menopausal symptoms, including "atypical" symptoms, such as dental

XSL•FO
RenderX

symptoms, dry eyes, and migraines ("finding concise information about the wide-ranging nature of symptoms is difficult", "Medical professionals/NHS websites providing full lists of symptoms as some quite common ones aren't included, e.g. itchy ears and bleeding gums."). Relatedly, free-text data included requests for more comprehensive information on some domains of menopause symptoms, such as urogenital and sexual symptoms ("More information about sex, painful sex") and mental health symptoms ("More information about the psychological symptoms and how to manage them. Most information is heavily swayed towards the physical symptoms. Not enough help or information about the psychological side.").

Beyond the need for a complete list of symptoms, requests were made about information to help understand the chronicity ("There's no information about what symptoms are permanent and what are transient.") and prevalence of these symptoms ("Perhaps more data rather than the 'you might experience' just to get some idea about how prevalent particular symptoms are.").

Participants also requested information on how to recognize or "spot" that you may be in perimenopause or menopause ("How to tell you are peri menopausal"), indicating a possible lack of clarity of how symptoms map onto the different menopause stages.

Health Care Access and Support (Including Offline)

Participants requested more online resources to provide recommendations of when it may be appropriate to seek formal support from an HCP ("When to see GP"). In the free-text data, participants also expressed the importance of having access to menopause specialists ("To be able to speak to a gynaecologist rather than a GP") and seeking information about how to gain a referral to such specialists ("How to refer oneself to specialist services"). In addition, participants requested recommendations on how to speak to an HCP about menopause, including how to self-advocate ("How to advocate for yourself when contacting a GP about menopause treatment"), noting the importance of accessing HCPs alongside online resources ("there should be medical professionals to provide more information, reassurance").

Other Support Outside of Formal Care

In terms of online information in relation to support outside of HCPs, participants in this study requested additional resources to guide them in gaining support in the workplace ("What are our rights re work support. How to get help when work becomes difficult [due] to symptoms"). Furthermore, support groups were mentioned as a useful support resource, with participants either expressing their benefits or requesting improved access to directories to find relevant support groups ("what I found most helpful was some facebook groups where women shared their experiences and stories as then there was always something you could relate to and understand how they'd managed or what advice they gave or had found etc.").

Lifestyle Support and Self-Management

Thematic analysis revealed an interest in online information regarding nonformal interventions to manage menopausal symptoms, including natural remedies, such as supplements and lifestyle changes ("naturally treat/support hormonal changes with the appropriate supplements"). In addition, some participants mentioned they wanted to see more online resources with information about food, diet, and exercise to manage symptoms of menopause or to lose weight ("Also more on exercising through menopause"). Free-text responses also included requests for reviews of supplements and other products specifically designed for use in menopause ("Reviews of natural supplements etc").

Menopause and Co-Occurring Conditions

Thematic analysis revealed requests for improved online information and guidance about the association between menopause and co-occurring conditions, including but not limited to chronic health conditions ("Specifications about how other chronic conditions can overlap with menopause"), psychological conditions (More advice about how the menopause can worse existing general anxiety disorder" and "[premenstrual dysphoric disorder] and it getting worse before menopause), and gynecologic conditions ("It is impossible to find any information for women who suffer with [polycystic ovary syndrome or PCOS] and how to manage this and menopause at the same time and they interact with each other. We are in desperate need of guidance and support for PCOS & Menopause").

Patient Empowerment

Thematic analysis also highlighted how online resources may be amended to empower patients. This included providing or enabling the sharing of narrative accounts in online spaces ("Personal stories of women coping and how they cope") to understand the experiences and successful management strategies of other individuals in a similar circumstance. In a similar vein, some participants requested that online resources provide reassurance, conveying the benefits for comfort and encouragement ("I think reassurance that you will get through it as it's quite daunting when you first go through it").

Perimenopause

Participants reported a perceived lack of online resources about perimenopause ("There isn't much about early premenopause either. it would be helpful if NHS or other credible UK healthcare providers shared information on this. How long should one wait before taking any measures? It is ok to start later? What are the window of opportunity for taking which actions?"), making it difficult to obtain information about this stage of the menopause transition.

Discussion

Principal Findings

The aims of this study were to explore the online information-seeking behavior of women in menopause and the menopause transition. In addition, the study sought to gain insights into perceptions of online resources for menopause information. Most participants in this study had used the internet as a source of menopause information. The internet seems to be not only highly accessible but also acceptable, with it being frequently used as the first point of information gathering. It

XSL•FO
RenderX

also seems that the internet is considered a high-quality resource for menopause information, with information perceived as mostly accurate and trustworthy. This is concerning, as reviews of online information for menopause have revealed varying quality [18]; however, given the pace with which websites can be updated, a more recent audit of quality and content is required. Reassuringly, the NHS website was the most frequently accessed resource for menopause information and thus is likely to be evidence based. However, qualitative analysis revealed a perception from the target audience for this resource that it may be considered out-of-date. Social media was the second most commonly accessed online resource for menopause information, revealing a considerable risk of misinformation. However, social media was not a resource that was frequently considered trustworthy. Therefore, this study highlights a potential need for guidance for those searching for menopause information online about how to differentiate high-quality, evidence-based content from potentially unreliable sources. Many participants had also sought information from other sources, with an NHS GP being the most frequently reported contact. Crucially, around 24.4% (142/581) of participants had only used the internet for information, not consulting any other information source. This further demonstrates the importance of validating the quality of information available online to tackle potential misinformation as well as maximizing the accuracy and utility of information found.

Information about menopause symptoms, menopause treatment options, and self-help tips or strategies was reported as most frequently looked for online. This reflects previous research investigating informational needs for menopausal individuals [10,11]. Other frequently searched topics were related to health, with participants endorsing having searched for information about menopause-related health risks as well as the interaction between menopause and other preexisting conditions. Interestingly, although information about menopause symptoms was the most commonly searched for information online, it seems that this information is also perceived to be not readily available. Many participants who provided a free-text response requested an improved provision of information online related to menopause symptoms, with there being a request for resources to provide a more extensive list of possible symptoms as well as information related to potential chronicity and duration of symptoms. Given the wide range of potential menopause symptoms and the distress that can accompany their onset [19], access to comprehensive information is vital. This is especially important in reducing the potential uncertainty or fear associated with the onset of unfamiliar or infrequently discussed menopause symptoms.

This study revealed that online search behavior varies considerably between menopausal stages. While the vast majority of participants, irrespective of menopause stage, made use of the internet as an informational resource, there were variations in the specific information that was searched for. For instance, those in the perimenopause searched online for information about general “how to” information and private care options, as well as complementary therapies. This may indicate that participants in this stage of menopause may have had poor experiences of formal care offered through the public sector and so were potentially motivated to seek information about alternative avenues for help and symptom management. This indicates the complexity that governs online searches for information about menopause, including but not limited to specific stages. Given this, future resources should aim to address the concerns and needs of different menopausal groups. This will involve in-depth co-design of materials, ensuring that a range of experiences and perspectives are captured and reflected in the resources generated. Beyond the menopausal stage, thematic analysis revealed a perception of a lack of information for other specific groups, such as those in medically induced or surgical menopause. Considering that surgical menopause is associated with poorer health-related quality of life outcomes [20] and more severe symptoms [21] compared to other menopausal groups it is crucial they are well-informed, but it appears that online information about menopause may not adequately address the unique needs of individuals who have undergone medically induced or surgical menopause. Therefore, there is a need to provide information that is relevant to the specific experiences of this subgroup.

Information about HRT was the most searched-for treatment information in this study, yet it is unclear whether these searches were a result of being offered HRT by an HCP, or if it was self-directed research. HRT is one of the most frequently prescribed treatments for the management of menopause symptoms, and therefore, these searches may indicate the goal of obtaining as much information as possible about this treatment option. However, there is evidence that some individuals have to self-advocate to access HRT and may be collecting relevant information to take to their HCP in this pursuit. Other than HRT, nearly a quarter of participants were looking for information about the use of testosterone for the management of menopause symptoms, with many free-text responses indicating an interest in this area. Testosterone is currently not licensed for use for menopause in the United Kingdom [22], with access to testosterone usually reserved for individuals presenting with reduced sex drive and when other HRT has not helped [23]. Although testosterone may not be widely prescribed, this study shows that there is obvious interest in its use. In fact, there was a 10-fold increase in the number of prescriptions of testosterone for women aged >50 years between 2015 and 2022 [24]. This increase may be associated with what has been described as the “Davina McCall effect,” with a sharp increase in the number of people receiving testosterone at the release of her first documentary about menopause [24]. A further increase was also observed after she released a second documentary that included information about testosterone treatment [24]. Apart from HRT, approximately 15% (68/442) of participants reported looking for information about antidepressants. According to the National Institute for Health and Care Excellence guidelines [23], if there is no independent diagnosis of major depressive disorder, prescription of antidepressants to treat the symptoms of low mood and anxiety related to menopause is not recommended. In practice, it has previously been reported that women experiencing low mood in menopause are prescribed selective serotonin reuptake inhibitors [25]. Prescriptions of antidepressants are seemingly perceived as inappropriate for menopause by those who have received them [26], and therefore, this search behavior may

XSL•FO
RenderX

reflect uncertainty regarding the suitability and efficacy of this treatment option for menopause. The differing types of treatment information sought is also reflected in free text responses, with participants demonstrating interest in information regarding alternative interventions to HRT. Overall, the identified online search behaviors for menopause treatment information highlight a significant gap in accessible, comprehensive resources available on and offline. Therefore, there is a need for detailed and reliable information on the wide range of therapeutic interventions available, which could both support the navigation of the complex landscape of menopause treatment and empower informed decision-making, both before and after contact with an HCP. In fact, this was something frequently requested by participants, with calls for improved information online and interest in decision-support tools, such as flowcharts, which may assist individuals in identifying the most appropriate intervention for them. Ideally, these resources would also include details of interventions or self-help strategies that can be used outside of formal health care settings, with free-text responses indicating interest in such an approach to manage menopause.

Frequently searched for treatment side effects online included weight gain, mood changes, and risk of cancer. Concerns about cancer related to menopause treatments are frequent, fueled by conflicting research publications about the risks associated with HRT. However, for most individuals, the benefits of HRT outweigh any potential risks [27], with HRT conferring both short-term symptom relief and longer-term protective health benefits [28]. Fears of HRT may also be propagated by HCPs who reportedly infrequently provide an opportunity to discuss HRT, including both potential risks and benefits [29]. Therefore, the internet may offer an effective modality for providing comprehensive information about HRT, particularly with regard to the risks associated with it. Thematic analysis revealed a perception that misinformation is spread online about HRT, potentially contributing to unnecessary confusion or distress when seeking formal care.

The type of general advice most sought via online searches was how to ask for support, both generally and from an HCP. Evidence indicates poor formal care experiences for menopause [26], with a high level of self-advocacy required. Therefore, increasing the skills needed to ask for help and advocate for support and health care is clearly needed, with the internet offering a convenient source to deliver this information. However, it seems that it may not yet be fulfilling this need, with free-text responses requesting that online resources include more advice and recommendations related to this. In addition, thematic analysis revealed requests for how to ask for support in the workplace and workplace rights. Previous evidence indicates that generally workplaces do not provide information about menopause [30], potentially necessitating online searching. Similarly, thematic analysis also indicated a perception that information provision is lacking for workplaces. Given that menopause is associated with women leaving the workforce [5], it is imperative that workplaces are well informed and prepared to support employees, and as such, high-quality online resources for both employees and employers are required.

Finally, this study demonstrates the importance of social connection in menopause. Free-text data asked for online resources to provide directories of support groups. Participants also stated the utility of support groups in online spaces, such as social media, to share information and experiences. This is similar to previous qualitative research in postmenopausal women, which found that support groups, including online support groups, enable knowledge sharing and reduce feelings of isolation [31]. Internet-based social networks have also been demonstrated to be beneficial for connecting with others who have a similar experience, particularly for individuals whose experience may be considered “atypical” [8]. Similarly, participants also requested narrative accounts of menopause, stating that these types of passive peer support would achieve similar benefits. This finding reflects other works that also found that information is used to better understand and validate menopause symptoms and experiences [32].

Strengths and Limitations

The findings from this study should be interpreted alongside several considerations. First, due to the characteristics of the study sample, extreme care must be taken when extrapolating these findings to the wider population. The participants in the study were disproportionately White (592/627, 95%), compared to the recorded national average of 81.7% in the 2021 census [33]. In addition, much of the sample endorsed a high educational achievement of at least an advanced-level certificate or equivalent. This is notable as education has a demonstrable influence on health literacy [34] and how one experiences and interacts with health care [35]. It is likely that different groups will differ in their use of and perspectives on online searching for menopause information, and the study’s findings will not reflect the perspectives from minority groups.

Second, there is a risk of selection bias due to the online recruitment method. Individuals who are familiar with using the internet, particularly those who possess the knowledge required to find health-related information, were more likely to see the study advertisements and choose to participate. This may have resulted in a sample that is not fully representative of the broader population. Individuals with limited internet access, lower digital literacy, or less experience with navigating online health resources will likely have been underrepresented. In addition, the recruitment source for individual participants was not tracked, and therefore, we are unable to analyze whether differences in recruitment pathways (eg, social media, professional networks, or word of mouth) may have influenced participant characteristics or responses. Therefore, we are unable to evaluate the risk of any potential recruitment bias. Consequently, a number of participants may have been recruited into this study after previously taking part in research conducted by the Cambridge Centre for Neuropsychiatric Research and indicating a willingness to be contacted for future studies. This may have introduced additional bias, as these individuals could differ from the general population in terms of research interest or health information–seeking behaviors.

Consequently, the findings may not reflect the use of the internet as an informational source for menopause in the wider population, with the behaviors and preferences of those who

XSL•FO
RenderX

are less comfortable with or have limited access to digital tools potentially differing significantly from the findings presented. For example, previous work has demonstrated that higher educational attainment and digital literacy are associated with higher rates of digital health information seeking [36].

Third, while the survey used in this study was reviewed and approved by a psychiatrist and provided a broad overview of online search behavior and informational needs related to menopause, some key aspects may have been overlooked, as it was not reviewed by a wider group of relevant stakeholders. Although we attempted to mitigate this limitation by including free-text fields where participants could share additional insights, we acknowledge that some key data may still not have been captured. Future studies could benefit from a more robust co-design process during survey development to ensure relevant data capture.

Finally, because the free-text question followed the questions with prespecified answer options, the responses may have been influenced by them.

Future Directions

Future research should aim to address the limitations in sample diversity present in this study. The overrepresentation of White, highly educated participants highlights the need for studies involving more ethnically and socioeconomically diverse populations to ensure findings are broadly applicable. Recruitment strategies that go beyond online platforms, as were used in this study, should also be explored to reduce selection bias and better capture the experiences of individuals who may have limited digital access or lower digital literacy.

Furthermore, given the apparent widespread reliance on the internet as a primary source of information about menopause, it is essential that research be undertaken to complete objective assessments of the quality and accuracy of these online resources. Such evaluations will help identify further gaps in information coverage or areas where resources could be improved and assess the risk of potential misinformation. This work is particularly important given the high levels of trust and perceived accuracy of online informational resources for menopause expressed by participants in this study.

Beyond research, this study also points to clear gaps in the available menopause-related information from the perspective of those seeking it and suggests opportunities to develop more tailored resources for underrepresented groups. Therefore, the current findings may be valuable for individuals or organizations involved in developing online or digital resources related to menopause information.

Conclusions

This study found that self-directed online searches for information about menopause are frequent, commonly being the first or potentially the only source of information sought by some women. Online searches are conducted principally to find information related to menopause symptoms and treatments. The information available online is widely trusted and generally perceived as accurate, with the websites of recognized health care services being perceived as most trustworthy. However, qualitative data indicate that information is perceived as generic, and there are several risk factors associated with distrust in information, including a lack of informational source transparency and perceived deceptive advertising practices. There are also apparent differences in the online search needs and behaviors between different menopausal groups. Future resource development should consider the role of individual characteristics and preferences, including the specific stage of menopause.

The participants in this study have highlighted some specific problems with the currently available online information about menopause and provided suggestions of what they would like to see changed. In particular, participants requested a more comprehensive and detailed list of all menopause symptoms and clearer guidance available on treatment options, including nonformal symptom management options. Internet users also seem to seek practical advice related to self-management of symptoms and on how to effectively communicate with HCPs and employers. In addition, the study revealed the potential for the internet to provide social support, both passively (ie, through narrative accounts) and actively (ie, through support networks), with this being cited as beneficial for reducing uncertainty about the menopausal experience.

Acknowledgments

This study was funded by the Stanley Medical Research Institute (grant 07R-1888).

Data Availability

The datasets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

FM, ELF, NAM-K, and SB conceived the study focus and materials. The recruitment was conducted by ELF. Quantitative analysis was performed by ELF. The thematic analysis was performed by FM and LQ as the first reviewers and ELF, and KB as the second reviewers. FM and ELF prepared the manuscript with revisions from LQ, NAM-K, KB and SB. All authors contributed to the paper and approved the submitted version.

XSL•FO
RenderX

Conflicts of Interest

Funding of this study was provided by the Stanley Medical Research Institute (grant 07R-1888). SB is cofounder and chief medical officer of Psyomics Ltd. SB holds shares in Psynova Neurotech Ltd and Psyomics Ltd. ELF and KB are paid consultants for Psyomics Ltd.

Multimedia Appendix 1
The Checklist for Reporting Results of Internet E-Surveys.
[DOCX File , 20 KB - formative_v9i1e75335_app1.docx]

Multimedia Appendix 2
Survey questions.
[DOCX File , 34 KB - formative_v9i1e75335_app2.docx]

Multimedia Appendix 3
Data tables.
[DOCX File , 75 KB - formative_v9i1e75335_app3.docx]

References

1. Surgical menopause: a toolkit for healthcare professionals. British Menopause Society. 2024. URL: https://thebms.org.uk/wp-content/uploads/2024/10/13-BMS-TfC-Surgical-Menopause-SEPT2024-D.pdf [accessed 2025-03-31]
2. Menopause. World Health Organization. 2024 Oct 16. URL: https://www.who.int/news-room/fact-sheets/detail/menopause [accessed 2025-03-31]
3. Menopause basics. Office on Women's Health. URL: https://womenshealth.gov/menopause/menopause-basics [accessed 2025-03-31]
4. Hickey M, Szabo RA, Hunter MS. Non-hormonal treatments for menopausal symptoms. BMJ 2017 Nov 23;359:j5101. [doi: 10.1136/bmj.j5101] [Medline: 29170264]
5. Women and Equalities Committee. Menopause and the workplace. United Kingdom Parliament. 2022. URL: https://publications.parliament.uk/pa/cm5803/cmselect/cmwomeq/91/report.html [accessed 2025-03-31]
6. Menopause puts final nail in marriage coffin. Balance. URL: https://www.balance-menopause.com/news/menopause-puts-final-nail-in-marriage-coffin/ [accessed 2025-03-31]
7. Harper JC, Phillips S, Biswakarma R, Yasmin E, Saridogan E, Radhakrishnan S, et al. An online survey of perimenopausal women to determine their attitudes and knowledge of the menopause. Womens Health (Lond) 2022;18:17455057221106890 [FREE Full text] [doi: 10.1177/17455057221106890] [Medline: 35758176]
8. Lopez EL, Felice MC, O'Kane AA. Using and appropriating technology to support the menopause journey in the UK. In: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. 2024 Presented at: CHI '24; May 11-16, 2024; Honolulu, HI. [doi: 10.1145/3613904.3642694]
9. Women's health strategy for England. Department of Health and Social Care United Kingdom Government. 2022. URL: https://www.gov.uk/government/publications/womens-health-strategy-for-england/womens-health-strategy-for-england [accessed 2025-03-31]
10. Martin-Key NA, Funnell EL, Benacek J, Spadaro B, Bahn S. Intention to use a mental health app for menopause: health belief model approach. JMIR Form Res 2024 Oct 16;8:e60434 [FREE Full text] [doi: 10.2196/60434] [Medline: 39412868]
11. Trudeau KJ, Ainscough JL, Trant M, Starker J, Cousineau TM. Identifying the educational needs of menopausal women: a feasibility study. Womens Health Issues 2011;21(2):145-152 [FREE Full text] [doi: 10.1016/j.whi.2010.10.001] [Medline: 21185735]
12. Hajesmaeel-Gohari S, Shafiei E, Ghasemi F, Bahaadinbeigy K. A study on women's health information needs in menopausal age. BMC Womens Health 2021 Dec 28;21(1):434 [FREE Full text] [doi: 10.1186/s12905-021-01582-0] [Medline: 34961491]
13. Nash Z, Christmas M, Gronlund T, Hickey M, Menopause Priority Setting Partnership Steering Group. Top ten menopause research priorities. Lancet 2025 Dec 21;404(10471):2535-2536. [doi: 10.1016/S0140-6736(24)02602-3] [Medline: 39709195]
14. Soules MR, Sherman S, Parrott E, Rebar R, Santoro N, Utian W, et al. Executive summary: stages of reproductive aging workshop (STRAW). Climacteric 2009 Jul 03;4(4):267-272. [doi: 10.1080/cmt.4.4.267.272]
15. El Khoudary SR, Greendale G, Crawford SL, Avis NE, Brooks MM, Thurston RC, et al. The menopause transition and women's health at midlife: a progress report from the Study of Women's Health Across the Nation (SWAN). Menopause 2019 Oct;26(10):1213-1227 [FREE Full text] [doi: 10.1097/GME.0000000000001424] [Medline: 31568098]
16. Cohen J. Statistical Power Analysis for the Behavioral Sciences. Milton Park, UK: Routledge; 1989.
17. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]

XSL•FO
RenderX

18. Pérez-López FR. An evaluation of the contents and quality of menopause information on the World Wide Web. Maturitas 2004 Dec 10;49(4):276-282. [doi: 10.1016/j.maturitas.2004.07.006] [Medline: 15531123]
19. Monteleone P, Mascagni G, Giannini A, Genazzani AR, Simoncini T. Symptoms of menopause - global prevalence, physiology and implications. Nat Rev Endocrinol 2018 Apr 2;14(4):199-215. [doi: 10.1038/nrendo.2017.180] [Medline: 29393299]
20. Bhattacharya SM, Jha A. A comparison of health-related quality of life (HRQOL) after natural and surgical menopause. Maturitas 2010 Aug;66(4):431-434. [doi: 10.1016/j.maturitas.2010.03.030] [Medline: 20434859]
21. Gibson-Helm M, Teede H, Vincent A. Symptoms, health behavior and understanding of menopause therapy in women with premature menopause. Climacteric 2014 Dec 17;17(6):666-673. [doi: 10.3109/13697137.2014.913284] [Medline: 24742007]
22. Testosterone replacement in menopause. British Menopause Society. 2022. URL: https://thebms.org.uk/wp-content/uploads/2022/12/08-BMS-TfC-Testosterone-replacement-in-menopause-DEC2022-A.pdf [accessed 2025-03-31]
23. Menopause: identification and management. National Institute for Health and Care Excellence. 2015. URL: https://www.nice.org.uk/guidance/NG23 [accessed 2025-03-31]
24. Connelly D. NHS testosterone prescribing in women rises ten-fold in seven years. Pharm J 2023 Feb 17 [FREE Full text] [doi: 10.1211/PJ.2023.1.174950]
25. Menopause and the workplace. Fawcett Society. URL: https://www.fawcettsociety.org.uk/menopauseandtheworkplace [accessed 2025-03-31]
26. Stevenson JC, Farmer RD. HRT and breast cancer: a million women ride again. Climacteric 2020 Jun;23(3):226-228. [doi: 10.1080/13697137.2020.1735797] [Medline: 32157921]
27. Genazzani AR, Monteleone P, Giannini A, Simoncini T. Hormone therapy in the postmenopausal years: considering benefits and risks in clinical practice. Hum Reprod Update 2021 Oct 18;27(6):1115-1150. [doi: 10.1093/humupd/dmab026] [Medline: 34432008]
28. Covello VT, Peters RG. Women's perceptions of the risks of age-related diseases, including breast cancer: reports from a 3-year research study. Health Commun 2002 Jul;14(3):377-395. [doi: 10.1207/S15327027HC1403_5] [Medline: 12186494]
29. Martin-Key NA, Funnell EL, Spadaro B, Bahn S. Perceptions of healthcare provision throughout the menopause in the UK: a mixed-methods study. npj Womens Health 2023 Dec 07;1:2. [doi: 10.1038/s44294-023-00002-y]
30. Beck V, Brewis J, Davies A. The remains of the taboo: experiences, attitudes, and knowledge about menopause in the workplace. Climacteric 2020 Apr 19;23(2):158-164. [doi: 10.1080/13697137.2019.1656184] [Medline: 31533494]
31. Aljumah R, Phillips S, Harper JC. An online survey of postmenopausal women to determine their attitudes and knowledge of the menopause. Post Reprod Health 2023 Jun 29;29(2):67-84 [FREE Full text] [doi: 10.1177/20533691231166543] [Medline: 36994487]
32. Emter F, Chavula C. Seeking control: how women evaluate and use menopause related information. In: Proceedings of the 2025 ACM SIGIR Conference on Human Information Interaction and Retrieval. 2025 Presented at: CHIIR '25; March 24-28, 2025; Melbourne, Australia. [doi: 10.1145/3698204.3716455]
33. Ethnic group, England and Wales: census 2021. Office for National Statistics. 2022. URL: https://www.ons.gov.uk/peoplepopulationandcommunity/culturalidentity/ethnicity/bulletins/ethnicgroupenglandandwales/census2021 [accessed 2025-03-31]
34. Sørensen K, Pelikan JM, Röthlin F, Ganahl K, Slonska Z, Doyle G, et al. Health literacy in Europe: comparative results of the European Health Literacy Survey (HLS-EU). Eur J Public Health 2015 Dec;25(6):1053-1058 [FREE Full text] [doi: 10.1093/eurpub/ckv043] [Medline: 25843827]
35. Smith SK, Dixon A, Trevena L, Nutbeam D, McCaffery KJ. Exploring patient involvement in healthcare decision making across different education and functional health literacy groups. Soc Sci Med 2009 Dec;69(12):1805-1812. [doi: 10.1016/j.socscimed.2009.09.056] [Medline: 19846245]
36. Jacobs W, Amuta AO, Jeon KC. Health information seeking in the digital age: an analysis of health information seeking behavior among US adults. Cogent Soc Sci 2017 Mar 13;3(1):1302785. [doi: 10.1080/23311886.2017.1302785]

Abbreviations

GP: general practitioner
HCP: health care professional
HRT: hormone replacement therapy
NHS: National Health Service

XSL•FO
RenderX

Edited by A Mavragani; submitted 01.04.25; peer-reviewed by A Adeoye, T Bezuglyy; comments to author 17.04.25; revised version received 26.06.25; accepted 27.06.25; published 01.10.25.

Please cite as:
Funnell EL, McConnell F, Martin-Key NA, Qian L, Babbitt K, Bahn S
Understanding Experiences of and Unmet Needs in Online Searches for Menopause Information: An Exploratory Survey
JMIR Form Res 2025;9:e75335
URL: https://formative.jmir.org/2025/1/e75335
doi:10.2196/75335
PMID:

©Erin Lucy Funnell, Freya McConnell, Nayra A Martin-Key, Leyao Qian, Kathryn Babbitt, Sabine Bahn. Originally published in JMIR Formative Research (https://formative.jmir.org), 01.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Low Risk Perception of Harm From Substance Use and Sexual Behaviors Among Online Help–Seeking Sexual and Gender Minoritized People in San Francisco, California: Cross-Sectional Survey

Jarett Maycott[1], MPH; Sean Arayasirikul[2], PhD

[1]San Francisco Department of Public Health, Center for Public Health Research, San Francisco, CA, United States

[2]The Legacy Center, Joe C Wen School of Population & Public Health, University of California, Irvine, Irvine, CA, United States

Corresponding Author:
Sean Arayasirikul, PhD
The Legacy Center
Joe C Wen School of Population & Public Health
University of California, Irvine
856 Health Sciences Quad
Irvine, CA, 92697-3957
United States
Phone: 1 949 824 2680
Email: sean.arayasirikul@uci.edu

Abstract

Background: Substance use and HIV epidemics have disproportionately affected sexual and gender minoritized (SGM) communities, with heightened risks among men who have sex with men (MSM) and transgender women of color due to intersecting challenges like poverty, mental health issues, and discrimination. Despite overall declines in substance use and sexual risk behaviors in the general population, these issues persist within SGM communities, exacerbated by stigma and systemic barriers to care. Digital health interventions have emerged as promising tools to address these disparities, offering accessible and stigma-reducing alternatives to traditional care, particularly effective among younger individuals and in underserved areas.

Objective: This study seeks to examine the social correlates of substance use and sexual risk perception among an online sample of help-seeking MSM and transgender women in San Francisco, California.

Methods: We recruited 409 help-seeking MSM and transgender women by using social media advertisements on Facebook, Instagram, and Grindr in 2022-2024. Participants provided informed consent and completed a baseline assessment.

Results: Utilization of testing resources for HIV and hepatitis was high among the participants (401/409, 98.04% and 360/409, 88.02%, respectively). Knowledge of HIV or other sexually transmitted infection health services was also high (379/409, 92.67%). Fewer participants (264/409, 64.55%) were knowledgeable about substance use–related services. Although many participants reported that using substances posed a high risk of harm, some perceived engaging in condomless sex, using prescription opioid drugs without a prescription, and using substances during sex as low risk (122/409, 29.83%, 41/409, 10.02%, and 60/409, 14.67%, respectively). Participants who reported experiencing unstable housing were more likely to report perceiving sharing needles (adjusted odds ratio [aOR] 7.20, 95% CI 1.99-27.80) and nonprescription opioid use (aOR 4.02, 95% CI 1.08-14.90) as low risk. Participants who reported an income below the federal poverty level were more likely to report perceiving sharing needles (aOR 6.35, 95% CI 1.84-23.40), prescription opioid use (aOR 2.89, 95% CI 1.32-6.18), and substance use during sex (aOR 2.29, 95% CI 1.14-4.48) as low risk. Participants who have not been tested for hepatitis in the past have 3.31 times the odds of perceiving prescription opioid use as low risk compared to counterparts who have been tested for hepatitis before (95% CI 1.36-7.68).

Conclusions: This study underscores the importance of social determinants in shaping low risk perception of the harm associated with substance use behaviors among online help–seeking SGM people in San Francisco. These systemic inequities structure participants' perceptions, access, and utilization of preventive and public health services. Our findings identify critical opportunities for outreach and preventative efforts needed to serve vulnerable populations.

(JMIR Form Res 2025;9:e81753) doi:10.2196/81753

XSL•FO
RenderX

KEYWORDS

substance use; digital health; sexual and gender minoritized community health; social media

Introduction

Substance use and HIV epidemics have disproportionately impacted sexual and gender minoritized (SGM) communities. The burden of substance use within SGM communities is disproportionately high and has increased precipitously during the COVID-19 pandemic [1]. Although there have been decreases in illicit substance use and sexual risk behaviors among the general population, associations between the two have persisted among SGM communities [1]. Men who have sex with men (MSM) of color and transgender women of color face heightened vulnerability to concurrent epidemics of poverty, poor mental health, discrimination, and substance use, each constituting a unique barrier to engagement in health care that perpetuates a cycle of neglect [2]. Researchers investigating factors associated with substance use and HIV found that Black MSM experienced an 8% decrease in the odds of reporting viral suppression for every unit increase in the substance abuse and mental illness symptoms [3]. A study in Massachusetts found a higher prevalence of homonegativity among MSM of color, suggesting that internalized stigma may be racialized [1]. Research on adherence to HIV treatment describes how HIV-related stigma and substance use stigma drive poor HIV care outcomes [4,5].

The confluence of these factors has created a unique environment for digital health interventions to become a vital tool for an evolving public health landscape. Digital health interventions may better serve SGM communities faced with complex decisions that limit their access to essential health services [6,7]. For example, researchers issued a self-administered survey to students and faculty in Oklahoma, Wyoming, and Texas, asking whether respondents had ever needed to visit a health care provider but did not in the last 6 months [8]. Their analysis found that SGM respondents experienced a higher likelihood of forgoing care compared to their non-SGM counterparts, suggesting that SGM communities disproportionately experience unmet health care needs [8]. Digital health interventions have been effective in addressing poor mental health among people living with HIV, especially for younger participants [9]. Digital health interventions have demonstrated potential in overcoming barriers rooted in complex stigma and isolation and mitigate stigma as a barrier to accessing health care [9].

Not only that but digital health interventions can also reach SGM populations where they are using online methods and having greater presence in online community spaces [10]. The AFFIRM study delivered cognitive behavioral therapy by using virtual conference technology to overcome traditional in-person barriers and was effective in improving the mental health of SGM youth [11]. Social networking apps have become digital spaces that lead to real-world connections for SGM communities. This can include strengthening community ties, forming relationships, and seeking sexual partners. Yet, these digital spaces can also facilitate sexual and substance use risk behaviors [12,13]. Studies have found that risk behaviors vary for SGM communities depending on the digital space or virtual platform [14,15]. This study seeks to examine the social correlates of substance use and sexual risk perception in an online sample of help-seeking MSM and transgender women in San Francisco, California. These data may inform public health strategies and future interventions to serve SGM communities at risk of and desiring help for substance use and HIV-related services.

Methods

Ethical Considerations

All procedures were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. The study protocol was approved by the institutional review board of the University of California, San Francisco (approval 20-33169) and covered secondary analyses without additional consent. Participants provided informed consent and were remunerated US $30 for completing the baseline assessment. Data for this secondary analysis were deidentified to protect participants' privacy and confidentiality.

Brief Description of the Parent Study

Health eNavigation 2.0 (or Health eNav 2.0) is a short, brief digital health intervention that connects participants to their own digital navigator through SMS text messaging to achieve behavior changes related to access and utilization of substance use and HIV prevention and treatment services. Participants were included in this study if they were interested in seeking help for substance use prevention or treatment and other related topics (eg, HIV/AIDS, mental health), were aged 18 years or older, resided in San Francisco, identified as a cisgender man who had sex with men or a transgender woman, were English-speaking, and had access to a smartphone. The digital navigator provided social support, motivational interviewing [16], and resource navigation and referrals to support participants' behavior change goals. The theoretical framework for this intervention has been described previously and applied to a sample of young people living with HIV to improve HIV care engagement [17].

Study Design and Recruitment

We recruited 409 participants online by using social media advertisements on Facebook, Instagram, and Grindr in 2022-2024. Advertisements sought out potential participants looking for help for substance use and interested in participating in the Health eNav 2.0 study. Interested potential participants clicked on the advertisement and were directed to a study interest website where they provided their contact information to study staff, and a staff person contacted them to schedule a remote enrollment visit via Zoom. During the enrollment visit, participants were screened for eligibility, and they provided informed consent electronically by using DocuSign. The enrolled participants completed a baseline assessment. They met with their digital navigator to cocreate a brief service plan

XSL•FO
RenderX

consisting of specific goals, objectives, and action steps related to their interest in the project. For example, a participant may be interested in using methamphetamine less. In conversation with their digital navigator, the participant may identify the consequences of their use, the reasons for their use, and actionable steps toward accessing treatment. After completion of the brief service plan, the digital navigator instructed participants on how to set up and use a password-protected, Health Insurance Portability and Accountability Act (HIPAA)-compliant text messaging platform. For the next 30 days, participants worked with their digital navigator over text messaging to enact their brief service plan. The digital navigator used motivational interviewing and social support to elicit change talk with participants during the intervention period. Participants completed an exit assessment at 1 month and a follow-up assessment at 3 months following exit.

Data Collection and Measures

Demographics

We measured age, race/ethnicity, socioeconomic status, housing stability, gender, and sexual orientation. Participants were asked to self-identify their race/ethnicity. Participants who indicated 2 or more races/ethnicities were coded as multiracial. Socioeconomic status was measured using self-reported categories of household income in the previous year. Categories included US $0-$10,000, US $10,001-$30,000, US $30,001-$50,000, US $50,001-$70,000, and higher than US $70,000. Using the 2022 federal poverty line, participants who reported US $10,000 or less were coded as below the federal poverty line [18]. Participants were asked to describe their housing situation. We measured housing as a dichotomous variable where participants who reported living in their own home or apartment was coded as having stable housing. Unstable housing included the following responses: living in a relative's home, in a group home, in a campus/dormitory housing, in a foster home, in a homeless shelter, or other. Participants self-reported their gender identity as cisgender man or transgender woman or other. Sexual orientation was measured with a self-reported question asking participants to select a response that best fit their sexual orientation between the following: gay/lesbian, bisexual, heterosexual or straight, and queer, pansexual, questioning, and/or other.

Knowledge and Utilization of Testing and Preventive and Treatment Services

The history of HIV testing was measured by asking participants, "Have you ever been informed of your HIV status (that is, whether or not you are HIV-positive) based on the result of an HIV test?" History of viral hepatitis testing was measured by asking participants, "Have you ever been informed of your viral hepatitis status (that means whether or not you are infected with the hepatitis virus) based on the result of a viral hepatitis test?" Participants' knowledge of HIV prevention and treatment services was assessed by asking, "Would you know where to go near where you live to see a health care professional regarding HIV/AIDS or other sexually transmitted health issues?" Participants' knowledge of substance use prevention and treatment services was assessed by asking, "Would you know where to go near where you live to see a health care professional regarding a drug or alcohol problem?" Responses for these items included yes, no, and do not know. We created dichotomous variables (yes/no) for these items, coding do not know as no.

Substance Use and Sex-Related Risk Perception

We measured participants' perception of the risk of people harming themselves when using specific substances and engaging in risky sexual behaviors. We measured participants' perception of risk related to sharing needles by asking, "What level of risk do you think people have of harming themselves physically if they share needles, syringes, or other injection equipment when using drugs once or twice a week?" This item was repeated to assess participants' perception of risk related to using nonprescription opioid drugs (eg, heroin, fentanyl), prescription opioid drugs without a prescription (eg, pain relievers), condomless sex, and substance use during sex. For example, participants' perception of risk related to substance use during sex was measured by asking, "What level of risk do you think people have of harming themselves if they have sex while high on drugs or under the influence of alcohol?" Responses included no risk, slight risk, moderate risk, great risk, and unknown risk. Low risk perception was coded as answering no, slight, or unknown risk. High risk perception was coded as answering moderate or great risk. These items are drawn from the National Survey of Drug Use and Health [19,20] and were adopted by the Substance Abuse and Mental Health Services Administration as reporting requirements for the evaluation of funded projects.

Statistical Analysis

This study is a cross-sectional analysis of baseline data that included social determinants of health and risk-related perceptions toward substance use behaviors and potential drivers of HIV risk. We used univariate statistics to describe the sample and variables of interest, including independent and dependent variables. Logistic regression was used to identify the social correlates of low risk perception of harm attributed to substance use and HIV-related behaviors. We built logistic regression models for each outcome controlling for age, gender, and sexual orientation. Analyses were performed using STATA software (version 17; STATA Corp, LLC).

Results

Table 1 describes participants' demographics, knowledge and utilization of testing and preventive services, and perceptions of harm attributed to substance use behaviors. Almost a third of participants (117/409, 28.61%) were aged 50 years or older, and 27.38% (112/409) were aged 30-39 years. The majority (327/409, 79.95%) identified as a cisgender man, and 20.05% (82/409) identified as a transgender women or other gender. Most participants identified as White or Latino/a/x/e racial/ethnic (169/409, 41.32% and 113/409, 27.63%, respectively). Approximately 17.60% (72/409) of the participants reported having unstable housing, and 17.36% (71/409) reported an annual income below the 2022 federal poverty line. Utilization of testing resources for HIV and hepatitis was high among our participants, with 98.04% (401/409) having previously been tested for HIV and 88.02%

XSL•FO
RenderX

(360/409) for hepatitis. Knowledge of HIV or other sexually transmitted infection health services was also high (379/409, 92.67%). However, participants were less knowledgeable about substance use–related services, with 64.55% (264/409) of the participants reporting they were aware of the resources.

XSL•FO
RenderX

Table 1. Sociodemographic characteristics, knowledge and utilization of testing and preventive resources, and risk perception among digital help–seeking men who have sex with men and transgender women in San Francisco, California, in 2021-2024 (N=409).

Characteristic	Values, n (%)
Demographics	
Age (years)	
18-29	103 (25.18)
30-39	112 (27.38)
40-49	77 (18.83)
50+	117 (28.61)
Race/ethnicity	
White	169 (41.32)
Latine	113 (27.63)
Asian and Pacific Islander and Native Hawaiian	50 (12.22)
Black	38 (9.29)
Multiracial/other	39 (9.54)
Gender	
Cisgender man	327 (79.95)
Transgender woman/other	82 (20.05)
Sexual orientation	
Bisexual	47 (11.49)
Gay/lesbian	282 (68.95)
Other	48 (11.74)
Straight/heterosexual	32 (7.82)
Housing stability	
Stable	337 (82.40)
Unstable	72 (17.60)
Socioeconomic status	
Above federal poverty line	338 (82.64)
Below federal poverty line	71 (17.36)
Knowledge and utilization of testing and preventive and treatment services	
History of HIV testing	
No	8 (1.96)
Yes	401 (98.04)
History of hepatitis testing	
No	49 (11.98)
Yes	360 (88.02)
Knowledge of where to access HIV or sexually transmitted infection services	
No	30 (7.33)
Yes	379 (92.67)
Knowledge of where to access substance use services	
No	145 (35.45)
Yes	264 (64.55)
Substance use and sexual risk perception	
Sharing needles	

XSL•FO
RenderX

Characteristic	Values, n (%)
High risk perception	397 (97.07)
Low risk perception	12 (2.93)
Nonprescription opioid drugs	
High risk perception	397 (97.07)
Low risk perception	12 (2.93)
Prescription opioid drugs without a prescription	
High risk perception	368 (89.98)
Low risk perception	41 (10.02)
Condomless sex	
High risk perception	287 (70.17)
Low risk perception	122 (29.83)
Substance use during sex	
High risk perception	349 (85.33)
Low risk perception	60 (14.67)

Overall, participants frequently reported that using substances posed high risk of harm. Approximately 97.07% (397/409) of the participants perceived sharing needles and using nonprescription opioid drugs as high risk, followed by prescription opioid drugs without a prescription, substance use during sex, and condomless sex (368/409, 89.98%, 349/409, 85.33%, and 287/409, 70.17%, respectively). Conversely, almost a third (122/409, 29.83%) perceived engaging in condomless sex as low risk. One in ten (41/409, 10.02%) perceived using prescription opioid drugs without a prescription as low risk, and 14.67% (60/409) perceived using substances during sex as low risk.

Table 2 describes the associations between social factors and low risk perception of harm attributed to substance use and HIV risk behaviors. We assessed the following social factors: race, housing stability, and socioeconomic status, history of receiving testing services, and knowledge of substance use and HIV-related services. Odds ratios and 95% CIs are presented, adjusted for age, gender, and sexual orientation.

Participants who reported experiencing unstable housing were more likely to report perceiving sharing needles (adjusted odds ratio [aOR] 7.20, 95% CI 1.99-27.80) and nonprescription opioid use (aOR 4.02, 95% CI 1.08-14.90) as low risk. Participants who reported an income below the federal poverty level were more likely to report perceiving sharing needles (aOR 6.35, 95% CI 1.84-23.40), prescription opioid use (aOR 2.89, 95% CI 1.32-6.18), and substance use during sex (aOR 2.29, 95% CI 1.14-4.48) as low risk. Participants who have not been tested for hepatitis in the past have 3.31 times the odds of perceiving prescription opioid use as low risk compared to counterparts who have been tested for hepatitis before (95% CI 1.36-7.68).

XSL•FO
RenderX

Table 2. Social correlates of low risk perception toward substance use and sexual risk among online help–seeking men who have sex with men and transgender women in San Francisco, California, in 2021-2024 (N=409).

Characteristic	Sharing needles, aOR[a] (95% CI)	Nonprescription opioids, aOR (95% CI)	Prescription opiods, aOR (95% CI)	Condomless sex, aOR (95% CI)	Substance use during sex, aOR (95% CI)
Race/ethnicity					
White	Ref[b]	Ref	Ref	Ref	Ref
Latine	3.97 (0.86-28.3)	0.92 (0.21-3.62)	0.74 (0.30-1.70)	0.63 (0.36-1.09)	0.86 (0.42-1.72)
Asian, Pacific Islander, and Native Hawaiian	1.90 (0.08-21.9	0.57 (0.03-4.09)	1.32 (0.43-3.62)	1.33 (0.67-2.60)	1.17 (0.47-2.74)
Black	1.50 (0.06-17.7)	0	0.55 (0.11-2.01)	0.17 (0.04-0.51)	0.15 (0.01-0.77)
Multiracial (non-Latino)/other	4.40 (0.44-44.8)	0.49 (0.02-4.07)	0.80 (0.20-2.55)	0.44 (0.17-1.03)	0.66 (0.20-1.84)
Housing stability					
Stable	Ref	Ref	Ref	Ref	Ref
Unstable	7.20 (1.99-27.8)	4.02 (1.08-14.9)	1.26 (0.52-2.86)	0.69 (0.36-1.28)	1.08 (0.48-2.27)
Socioeconomic status					
Above federal poverty line	Ref	Ref	Ref	Ref	Ref
Below federal poverty line	6.35 (1.84-23.4)	2.14 (0.51-7.77)	2.89 (1.32-6.18)	0.69 (0.36-1.25)	2.29 (1.14-4.48)
Has been tested for hepatitis					
Yes	Ref	Ref	Ref	Ref	Ref
No	2.82 (0.57-10.8)	2.94 (0.58-11.6)	3.31 (1.36-7.68)	1.12 (0.56-2.16)	0.96 (0.38-2.16)
Has been tested for HIV					
Yes	Ref	Ref	Ref	Ref	Ref
No	7.69 (0.36-60.7)	4.75 (0.21-39.4)	0.98 (0.05-6.65)	0.85 (0.12-3.95)	2.12 (0.29-10.5)
Knowledge of HIV/sexually transmitted infection services					
Yes	Ref	Ref	Ref	Ref	Ref
No	1.43 (0.08-8.27)	1.41 (0.07-8.71)	0.63 (0.10-2.31)	0.58 (0.21-1.39)	0.96 (0.27-2.66)
Knowledge of substance use–related services					
Yes	Ref	Ref	Ref	Ref	Ref
No	1.36 (0.38-4.56)	1.24 (0.31-4.29)	0.65 (0.29-1.36)	0.52 (0.32-0.84)	0.72 (0.38-1.30)

[a]aOR: adjusted odds ratio.

[b]Ref: reference.

Discussion

This study examines how SGM people seeking help for substance use–related services online perceive substance use risk and harms. We found that the highest proportion of participants perceived sharing needles and nonprescription opioid use as high risk, followed by prescription opioid drugs without a prescription, substance use during sex, and condomless sex. On the surface, these data demonstrate that participants are actively engaged in assessing risk and harms related to using specific substances and appropriately ascribe high risk to substances that have severe health consequences. For example, needle sharing is a mode of transmission for HIV acquisition [21], and use of nonprescription opioids like fentanyl could lead to life-threatening overdose [22]. Yet, for other behaviors like prescription opioid drug use without a prescription, condomless sex, and substance use during sex, more participants perceive these behaviors as less risky.

Our study finds that low risk perception is unevenly distributed, revealing that some people are more likely to perceive substance use behaviors as low risk. For example, although we found that while more than 95% of the participants perceived sharing needles as high risk, participants experiencing unstable housing and those with an income less than the federal poverty level are more likely to perceive sharing needles as low risk compared to their respective counterparts. Similarly, while most participants perceived nonprescription opioid drug use as high risk, people experiencing unstable housing were more likely to perceive this behavior as low risk.

XSL•FO
RenderX

These findings can inform how public health sectors ought to tailor interventions to reach communities who perceive substance use behaviors as low risk. We found that participants who reported an income below the federal poverty level were more likely to report perceiving prescription opioid use as low risk. Integrating substance use prevention and health education into employment programs and social welfare and other benefit programs may improve risk assessment and decision-making skills. For example, Bosk et al [23] describe opportunities for child welfare programs to integrate addiction treatment, education, and prevention among parents of families involved in the child welfare system. Additionally, we found that participants who have not been tested for hepatitis in the past are more likely to perceive prescription opioid use as low risk. Scaling up and enhancing viral hepatitis testing could act as a feasible conduit for health education about substance use disorder and prescription opioid use.

Our study is not without limitations. There is limited generalizability of our findings due to the cross-sectional study design and the unique sociohistorical environment and communities of the San Francisco Bay Area and its public health institutions. Although existing scientific literature suggests that social factors may influence substance use behaviors [24-26], our study cannot definitively establish causality or determine the direction of influence between these social factors and risk perception [27]. Historically, San Francisco has been a center of civil rights and public health movements for SGM communities—from HIV to harm reduction to transgender health to ending the syndemic of hepatitis C virus/HIV/sexually transmitted infections. As a result, not only have many SGM folks relocated to San Francisco to live in an environment less adverse to SGM people, the public health infrastructure has evolved to specifically meet the needs of SGM communities [28]. Despite these limitations, this study is unique in that the sample is comprised of SGM individuals seeking help online. Stigma can deter people from seeking help for substance use, mental health, and HIV services in-person [29], and creating entry points into the public health service system that interface with people seeking help online is critical to overcoming stigma and communities hardly reached by traditional methods [30]. Another limitation to our study may include how response categories were recoded to characterize low risk perception. Although many participants in our study reported knowledge of HIV/sexually transmitted infection services, more than a third were unaware of substance use services; reaching these individuals is critical to addressing gaps in substance use disorder service utilization, and increasing research find that it is important to reach people who use substances online [31,32]. Future research examining pre-exposure prophylaxis for HIV and sexually transmitted infections may be important factors in assessing sexual risk and vulnerability for people who use substances. Although our study examines perceptions of risk of harm to oneself related to substance use quantitatively, we understand that the dynamic processes of cognition and risk assessment in the lived experiences of vulnerable communities may perhaps be best investigated by employing qualitative methodologies. For example, phenomenology and repeated in-depth interviews over time may reveal critical insights into the lived experiences and meaning behind understanding risks related to substance use and decision-making.

This study underscores the importance of social determinants in shaping the low risk perception of harm associated with substance use behaviors among online help–seeking SGM people in San Francisco. These systemic inequities structure participants' perceptions, access, and utilization of preventive and public health services. Our findings identify critical opportunities for outreach and preventative efforts needed to serve vulnerable populations. Addressing the substance use and HIV syndemic requires public health interventions that seek to overcome structural barriers and leverage digital technology to meet people where they are. Although our study provides valuable insights, further research is needed to understand the influence of social determinants in digital spaces and facilitate the development of effective and inclusive interventions in the future.

Acknowledgments

The authors would like to thank all the participants in this study. This work was funded by the Substance Abuse and Mental Health Service Administration (SAMHSA) (award H79SP082077].

Data Availability

The datasets generated or analyzed during this study are not publicly available as they contain details that could be used to identify participants but are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

References

1. Batchelder AW, Foley JD, Wirtz MR, Mayer K, O'Cleirigh C. Substance use stigma, avoidance coping, and missed HIV appointments among MSM who use substances. AIDS Behav 2021 May;25(5):1454-1463 [FREE Full text] [doi: 10.1007/s10461-020-02982-3] [Medline: 32737816]

XSL•FO
RenderX

2. Babel RA, Wang P, Alessi EJ, Raymond HF, Wei C. Stigma, HIV risk, and access to HIV prevention and treatment services among men who have sex with men (MSM) in the United States: a scoping review. AIDS Behav 2021 Nov;25(11):3574-3604 [FREE Full text] [doi: 10.1007/s10461-021-03262-4] [Medline: 33866444]
3. Brown MJ, Serovich JM, Laschober TC, Kimberly JA. Age and racial disparities in substance use and self-reported viral suppression among men who have sex with men with HIV. Int J STD AIDS 2018 Oct;29(12):1174-1182 [FREE Full text] [doi: 10.1177/0956462418779663] [Medline: 29945540]
4. Stringer KL, Marotta P, Baker E, Turan B, Kempf M, Drentea P, et al. Substance use stigma and antiretroviral therapy adherence among a drug-using population living with HIV. AIDS Patient Care STDS 2019 Jun;33(6):282-293 [FREE Full text] [doi: 10.1089/apc.2018.0311] [Medline: 31166784]
5. Reif S, Wilson E, McAllaster C, Pence B. The relationship of HIV-related stigma and health care outcomes in the US deep south. AIDS Behav 2019 Oct;23(Suppl 3):242-250. [doi: 10.1007/s10461-019-02595-5] [Medline: 31317363]
6. Bogart LM, Ojikutu BO, Tyagi K, Klein DJ, Mutchler MG, Dong L, et al. COVID-19 related medical mistrust, health impacts, and potential vaccine hesitancy among Black Americans living with HIV. J Acquir Immune Defic Syndr 2021 Feb 01;86(2):200-207 [FREE Full text] [doi: 10.1097/QAI.0000000000002570] [Medline: 33196555]
7. Merrill JO, Rhodes LA, Deyo RA, Marlatt GA, Bradley KA. Mutual mistrust in the medical care of drug users: the keys to the "narc" cabinet. J Gen Intern Med 2002 May;17(5):327-333 [FREE Full text] [doi: 10.1046/j.1525-1497.2002.10625.x] [Medline: 12047728]
8. Tabler J, Schmitz RM, Charak R, Propst A. Forgone care among LGBTQ and Non-LGBTQ Americans during the COVID-19 pandemic: the role of health, social support, and pandemic-related stress. South Med J 2022 Oct;115(10):752-759 [FREE Full text] [doi: 10.14423/SMJ.0000000000001456] [Medline: 36191911]
9. Cao B, Bao H, Oppong E, Feng S, Smith KM, Tucker JD, et al. Digital health for sexually transmitted infection and HIV services: a global scoping review. Curr Opin Infect Dis 2020 Feb;33(1):44-50 [FREE Full text] [doi: 10.1097/QCO.0000000000000619] [Medline: 31789695]
10. Hsiang E, Offer C, Prescott M, Rodriguez A, Behar E, Matheson T, et al. Bridging the digital divide among racial and ethnic minority men who have sex with men to reduce substance use and HIV risk: mixed methods feasibility study. JMIR Mhealth Uhealth 2020 Apr 29;8(4):e15282 [FREE Full text] [doi: 10.2196/15282] [Medline: 32347807]
11. Craig SL, Leung VWY, Pascoe R, Pang N, Iacono G, Austin A, et al. AFFIRM Online: utilizing an affirmative cognitive-behavioral digital intervention to improve mental health, access, and engagement among LGBTQA+ youth and young adults. Int J Environ Res Public Health 2021 Feb 05;18(4):1541 [FREE Full text] [doi: 10.3390/ijerph18041541] [Medline: 33562876]
12. Young SD, Szekeres G, Coates T. The relationship between online social networking and sexual risk behaviors among men who have sex with men (MSM). PLoS One 2013;8(5):e62271 [FREE Full text] [doi: 10.1371/journal.pone.0062271] [Medline: 23658716]
13. Hoenigl M, Little SJ, Grelotti D, Skaathun B, Wagner GA, Weibel N, et al. Grindr users take more risks, but are more open to human immunodeficiency virus (HIV) pre-exposure prophylaxis: could this dating app provide a platform for hiv prevention outreach? Clin Infect Dis 2020 Oct 23;71(7):e135-e140 [FREE Full text] [doi: 10.1093/cid/ciz1093] [Medline: 31677383]
14. Chiasson MA, Parsons JT, Tesoriero JM, Carballo-Dieguez A, Hirshfield S, Remien RH. HIV behavioral research online. J Urban Health 2006 Jan;83(1):73-85 [FREE Full text] [doi: 10.1007/s11524-005-9008-3] [Medline: 16736356]
15. Klein H. Substance use and abuse among men using the Internet specifically to find partners for unprotected sex. J Psychoactive Drugs 2011;43(2):89-98 [FREE Full text] [doi: 10.1080/02791072.2011.587391] [Medline: 21858955]
16. Davidson R. Motivational issues in the treatment of addictive behaviour. Psychotherapy, Psychological Treatments and the Addictions 1996:173-188. [doi: 10.1017/cbo9780511722127.012]
17. Arayasirikul S, Trujillo D, Turner CM, Le V, Wilson EC. Implementing a digital HIV care navigation intervention (Health eNav): protocol for a feasibility study. JMIR Res Protoc 2019 Nov 08;8(11):e16406 [FREE Full text] [doi: 10.2196/16406] [Medline: 31702561]
18. Shrider E, Creamer J. Poverty in the United States: 2023. US Census Bureau. URL: https://www.census.gov/library/publications/2024/demo/p60-283.html [accessed 2025-09-22]
19. Reliability of key measures in the National Survey on Drug Use and Health. Substance Abuse and Mental Health Services Adminstration. URL: https://www.samhsa.gov/data/sites/default/files/2k6ReliabilityP/2k6ReliabilityP.pdf [accessed 2025-09-22]
20. SAMHSA. National Survey on Drug Use and Health: summary of methodological studies, 1971–2014. Substance Abuse and Mental Health Services Administration 2014 Nov:1-50. [Medline: 30199189]
21. Kipke MD, Unger JB, Palmer RF, Edgington R. Drug use, needle sharing, and HIV risk among injection drug-using street youth. Subst Use Misuse 1996 Jul;31(9):1167-1187. [doi: 10.3109/10826089609063971] [Medline: 8853236]
22. Han Y, Yan W, Zheng Y, Khan MZ, Yuan K, Lu L. The rising crisis of illicit fentanyl use, overdose, and potential therapeutic strategies. Transl Psychiatry 2019 Nov 11;9(1):282 [FREE Full text] [doi: 10.1038/s41398-019-0625-0] [Medline: 31712552]

XSL•FO
RenderX

23. Bosk EA, Paris R, Hanson KE, Ruisard D, Suchman NE. Innovations in child welfare interventions for caregivers with substance use disorders and their children. Child Youth Serv Rev 2019 Jun;101:99-112 [FREE Full text] [doi: 10.1016/j.childyouth.2019.03.040] [Medline: 32831444]
24. Tobin K, Davey-Rothwell M, Yang C, Siconolfi D, Latkin C. An examination of associations between social norms and risky alcohol use among African American men who have sex with men. Drug Alcohol Depend 2014 Jan 01;134:218-221 [FREE Full text] [doi: 10.1016/j.drugalcdep.2013.10.002] [Medline: 24210770]
25. Knerich V, Jones AA, Seyedin S, Siu C, Dinh L, Mostafavi S, et al. Social and structural factors associated with substance use within the support network of adults living in precarious housing in a socially marginalized neighborhood of Vancouver, Canada. PLoS One 2019;14(9):e0222611 [FREE Full text] [doi: 10.1371/journal.pone.0222611] [Medline: 31545818]
26. Rapier R, McKernan S, Stauffer CS. An inverse relationship between perceived social support and substance use frequency in socially stigmatized populations. Addict Behav Rep 2019 Dec;10:100188 [FREE Full text] [doi: 10.1016/j.abrep.2019.100188] [Medline: 31294075]
27. Wang X, Cheng Z. Cross-sectional studies: strengths, weaknesses, and recommendations. Chest 2020 Jul;158(1S):S65-S71. [doi: 10.1016/j.chest.2020.03.012] [Medline: 32658654]
28. Arayasirikul S, Turner C, Trujillo D, Sicro SL, Scheer S, McFarland W, et al. A global cautionary tale: discrimination and violence against trans women worsen despite investments in public resources and improvements in health insurance access and utilization of health care. Int J Equity Health 2022 Mar 03;21(1):32 [FREE Full text] [doi: 10.1186/s12939-022-01632-5] [Medline: 35241094]
29. Browne T, Priester MA, Clone S, Iachini A, DeHart D, Hock R. Barriers and facilitators to substance use treatment in the rural south: a qualitative study. J Rural Health 2016;32(1):92-101. [doi: 10.1111/jrh.12129] [Medline: 26184098]
30. Pretorius C, Chambers D, Coyle D. Young people's online help-seeking and mental health difficulties: systematic narrative review. J Med Internet Res 2019 Nov 19;21(11):e13873 [FREE Full text] [doi: 10.2196/13873] [Medline: 31742562]
31. Ranjit YS, Shrestha R, Copenhaver M, Altice FL. Online HIV information seeking and pre-exposure prophylaxis awareness among people who use drugs. J Subst Abuse Treat 2020 Apr;111:16-22. [doi: 10.1016/j.jsat.2019.12.010] [Medline: 32087834]
32. Davitadze A, Meylakhs P, Lakhov A, King EJ. Harm reduction via online platforms for people who use drugs in Russia: a qualitative analysis of web outreach work. Harm Reduct J 2020 Dec 09;17(1):98 [FREE Full text] [doi: 10.1186/s12954-020-00452-6] [Medline: 33298081]

Abbreviations

aOR: adjusted odds ratio
HIPAA: Health Insurance Portability and Accountability Act
MSM: men who have sex with men
SGM: sexual and gender minoritized

Edited by A Mavragani; submitted 02.08.25; peer-reviewed by G Rebchook; comments to author 02.09.25; revised version received 02.09.25; accepted 16.09.25; published 03.10.25.

Please cite as:
Maycott J, Arayasirikul S
Low Risk Perception of Harm From Substance Use and Sexual Behaviors Among Online Help–Seeking Sexual and Gender Minoritized People in San Francisco, California: Cross-Sectional Survey
JMIR Form Res 2025;9:e81753
URL: https://formative.jmir.org/2025/1/e81753
doi:10.2196/81753
PMID:

©Jarett Maycott, Sean Arayasirikul. Originally published in JMIR Formative Research (https://formative.jmir.org), 03.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Clinical System for Mood Disorder Care in Córdoba, Colombia: Participatory Design and Scenario-Based Usability Evaluation Study

Ever Augusto Torres-Silva[1*], MSc; Juan José Gaviria-Jiménez[1,2*], MSc; Eider Pereira-Montiel[1,3*], MSc; David Andrés Montoya-Arenas[4], MSc, PhD; José Fernando Flórez-Arango[5*], MSc, PhD

[1]Bioengineering and Microelectronics Research Group, School of Medicine, Universidad Pontificia Bolivariana, Medellín, Colombia

[2]Informed, School of Medicine, Universidad de Antioquia, Medellín, Colombia

[3]GI2B Research Group, Department of Exact and Applied Sciences, Instituto Tecnológico Metropolitano, Medellín, Colombia

[4]Psychology Research Group, School of Social Sciences, Universidad Pontificia Bolivariana, Medellín, Colombia

[5]Department of Population Health Sciences, Division of Health Informatics and Artificial Intelligence, Weill Cornell Medical College, New York, NY, United States

[*]these authors contributed equally

Corresponding Author:
Ever Augusto Torres-Silva, MSc
Bioengineering and Microelectronics Research Group
School of Medicine
Universidad Pontificia Bolivariana
Circular 1a 70-01
Medellín, 050031
Colombia
Phone: 57 604 448 83 88
Email: ever.torres@upb.edu.co

Abstract

Background: Mood disorders are among the leading causes of disability worldwide and present a growing public health concern. In Córdoba, Colombia, suicide rates have risen significantly in recent years, exposing structural gaps in mental health care delivery. Digital health solutions and telehealth interventions can expand access to early detection, referral, and monitoring of patients in underserved regions. However, their effectiveness depends on rigorous and diverse evaluations to ensure adoption and sustainability.

Objective: This study evaluated the usability of a clinical telehealth system for mood disorder care developed through participatory design, with emphasis on user-centered functionality and workload analysis.

Methods: The system was designed through 2 iterative development cycles, followed by a scenario-based usability evaluation. A functional Domain Ontology was constructed to prioritize 8 core functionalities, including telecounseling, a georeferenced institutional directory, hotline services, patient self-report tools, educational content, forums, and a population dashboard. Thirty participants representing patients, caregivers, clinical staff, and administrative personnel were recruited through convenience sampling. Usability was assessed through cognitive walk-throughs, the NASA (National Aeronautics and Space Administration) Task Load Index, and the Post-Study System Usability Questionnaire.

Results: A total of 34 usability sessions and 223 task-level workload assessments were conducted across 2 evaluation cycles. The system demonstrated high usability, with overall Post-Study System Usability Questionnaire scores of 2.2 in cycle 1 and 2.3 in cycle 2. Interfaces prioritized for patients and clinical staff achieved better evaluations (average 1.9-2.0) than administrative interfaces (average 3.0). Workload analysis indicated improvements between cycles, particularly for patient-centered tasks, with mental workload as the most significant source of cognitive demand. Twenty-three critical issues (9 system errors and 14 design flaws) were identified and corrected between cycles, leading to measurable usability gains.

Conclusions: The participatory and scenario-based approach facilitated early identification of usability challenges and supported iterative refinement of the system. Results suggest that the system is usable, acceptable, and effective in reducing workload for key user groups, particularly patients and clinicians. The findings reinforce the value of participatory methodologies in digital mental health and highlight the need to prioritize patient-facing interfaces. Future research should extend evaluations to mobile platforms and larger populations to support scalability and integration into regional mental health services.

XSL•FO
RenderX

(JMIR Form Res 2025;9:e58909) doi:10.2196/58909

KEYWORDS

user-centered design; user-computer interface; workload; mental health; mental health services; mood disorders; telehealth

Introduction

Recent studies indicate that mental health problems have reached epidemic proportions, with disorders such as depression, anxiety, bipolar disorder, trauma-related disorders, and stressors contributing significantly to the global burden of disease [1,2]. Epidemiological research reports that these conditions affect a large proportion of the world's population, often resulting in chronic disability and reduced quality of life. In parallel, there is growing evidence that high levels of alcohol consumption play a critical role in both precipitating mental health disorders and exacerbating symptoms in those already affected [3]. Despite the widespread prevalence and considerable individual and societal impact of these disorders, a treatment gap persists. Less than half of those with mental illness receive adequate care, a deficit attributed to stigma, limited financial resources, and deficiencies in existing health care infrastructures [4,5]. In the Colombian department of Córdoba, Simancas Fernández et al [6] conducted a review of 33 sources describing the mental health situation of the local population. Their findings evidenced a 33% increase in suicide cases between 2020 and 2022, identifying risk factors such as gender, depressive and anxiety symptoms, substance abuse, history of suicide, family dysfunction, forced displacement, and experiences of bullying [6]. Faced with the challenges posed by mental disorders, public policy makers and health professionals are increasingly advocating for the implementation of integrated care models. In this context, telehealth and the use of digital technologies emerge as key strategies to enhance remote detection and monitoring of patients by specialized health care personnel, aiming to improve access for populations in underserved areas, optimize patient follow-up, and reduce unnecessary travel, among other benefits [7].

An adequate participatory design of technological tools and high usability contributes to the improvement of the use of such strategies and enhances adherence, user satisfaction, and development and maintenance costs [8]. Participatory design consists of including the system's end users during the process. This participation is carried out in all design phases [9]. Usability evaluation of a product or system refers to the process of identifying how users complete activities and can meet the intended (usefulness) objectives while doing so easily and with a low level of errors [10]. There are multiple evaluation methods, including heuristic evaluation, scenario evaluation, and focus groups. These processes evaluate and apply rules to improve the usability of systems and tools both in the health software engineering and user-centered design area and in the care and clinical scenarios, for example, the design of electronic medical record systems, the design of mobile apps, and the adjustments and design of biomedical devices [11]. Recent studies support these approaches: in the development of the Kuamsha app for behavioral activation therapy, the authors implemented an iterative, user-centered agile design methodology involving adolescents in all design phases through focus groups, in-depth interviews, participatory workshops, and multicycle usability testing sessions to refine the interface, narrative elements, and functionality based on direct user feedback [12]. Similarly, the participatory design of a conversational artificial intelligence agent for mental health care incorporated focus groups with both users and psychotherapists during the early prototyping phases; usability and engagement were evaluated qualitatively through therapist feedback and patient-reported experiences after app use, which guided iterative improvements to ensure the system's integration into psychotherapy workflows [13].

Furthermore, recent work highlights that addressing quality, usability, and user engagement from the outset is essential for the long-term success of eHealth interventions. A cocreation-based framework, in which users, health care professionals, and developers collaborate as equal partners, has been proposed as an effective strategy to ensure usability, personalization, and sustained behavior change, particularly when artificial intelligence tools dynamically adapt content to user needs [14]. Building upon this evidence, this research aimed to evaluate the usability level of a system built through participatory design for detecting mood disorders and supporting telehealth care in mental health.

Methods

Study Design

The study follows a participatory design approach that required 2 iterations of development followed by a scenario-based evaluation. This methodology allows the identification of bottlenecks, additional system failures, ease of learning the tool, and the comprehensibility of the language, among others [15]. In the evaluation methodology by scenarios, the main functionalities and tasks included in the mental health care model of the Department of Córdoba, Colombia, were included. The Department of Córdoba is located in the northwestern region of Colombia, covering an area of approximately 25,020 km² (9660 square miles) and home to around 2 million people as of 2025 [16]. This model includes a system designed to provide a service for the timely detection of mood disorders, contact with mental health care personnel, and support for referring cases to the network of health care providers. At the time of the evaluation, the system had been built through participatory design and had not been released to end users. The study had three phases: (1) test scenario design and creation phase, (2) test user recruitment phase, and (3) test implementation phase (Figure 1).

XSL•FO
RenderX

Figure 1. Scenario-based usability evaluation flow applied in a participatory design study of a clinical information system for the care of people with mood disorders in Córdoba, Colombia (2022-2023). The diagram illustrates the sequence of steps, including scenario design, user recruitment, task assignment, observation, completion of PSSUQ and NASA-TLX questionnaires, system adjustments, and repeated testing across 2 development cycles. NASA-TLX: National Aeronautics and Space Administration Task Load Index; PSSUQ: Post-Study System Usability Questionnaire.

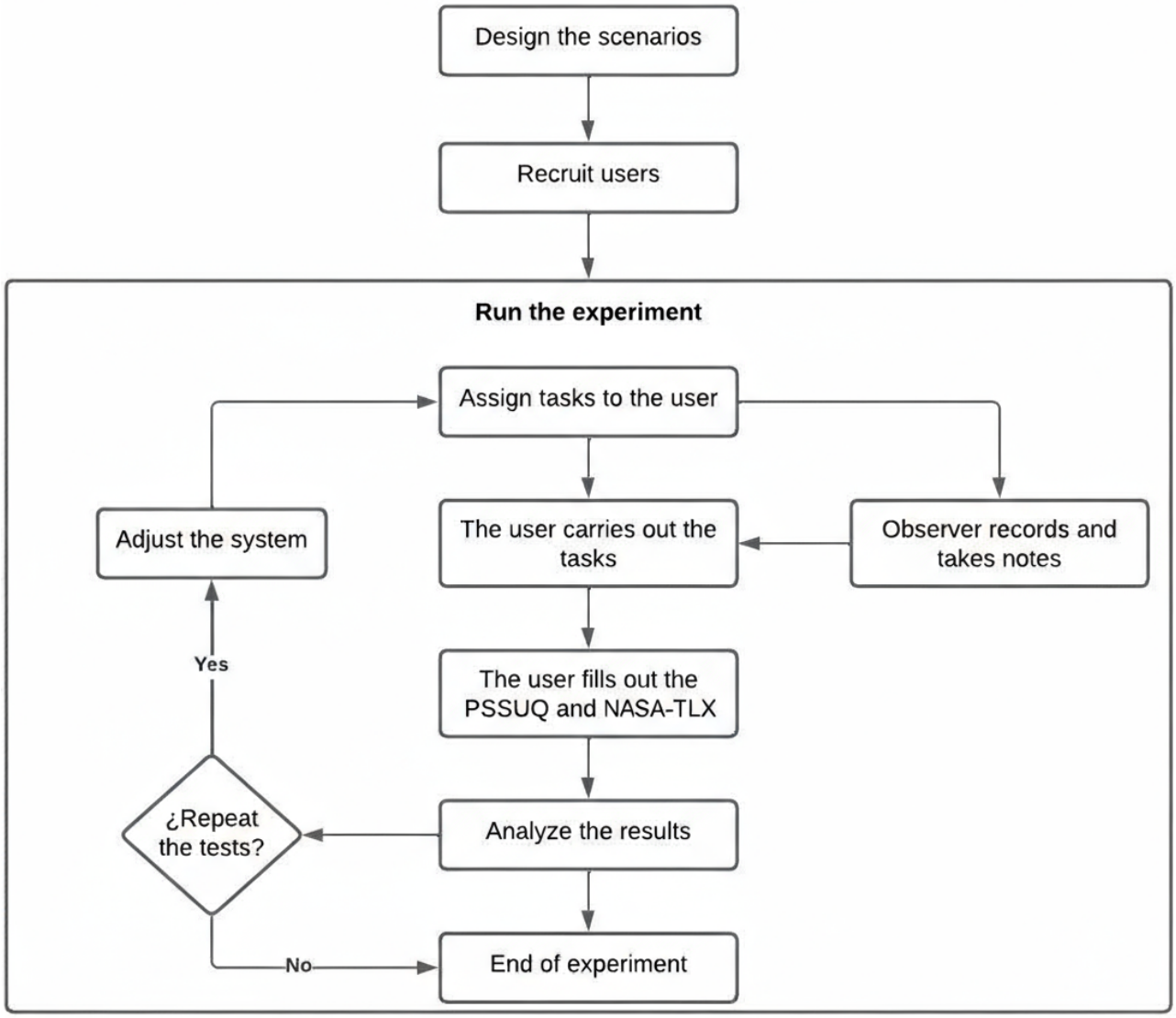

Scenario Preparation and Creation Phase

Functionalities and scenarios to be evaluated were selected based on the Knowledge Domain Ontology, following the GOMS (Goals, Objectives, Methods, Selection rules) methodology [11]. Two activities were carried out to construct the functional Domain Ontology (DO): (1) cocreation exercise with the project's coresearchers (16 psychologists, biomedical engineers, and communicators), in which the contributions of each of them were compiled on a collaborative platform; and (2) visits to governmental actors to identify the specific needs of the territory. This information was synthesized in Microsoft Excel. As a result, 8 basic functionalities of the system, containing 60 specific tasks, were found: (1) Telecounseling: the possibility of interacting with expert mental health personnel for patients and caregivers. (2) Directory of institutions: identification of georeferenced institutions that are part of the program. (3) Hotline: access through multiple communication channels to emergency mental health care. (4) Forum: the possibility of exchanging experiences among users of the mental health program. (5) Questions: compilation and display of frequently asked questions for quick reference. (6) Self-reported questionnaires: the possibility of completing scales, surveys, and screening and diagnostic tools reported to expert mental health personnel. (7) Information capsules: educational content about mental health and its care. (8) Population dashboard: display of active patients in the mental health program.

DO allows for prioritization of functionalities to be implemented through the frequency analysis of the potential users' participation in the tasks identified in the functional analysis described in the section "Scenario Preparation and Creation Phase." The DO identified main user groups: (1) Patients: individuals interested in receiving mental health services. Some of them formally join the program, which requires full identification of the individuals. These users have control over access to their own information. (2) Caregivers: individuals interested in getting mental health services for their care recipients. They may or may not be identified, and some

XSL•FO
RenderX

formally join the program, which requires full identification of the individuals. Access to patient information is limited. (3) Clinical staff: personnel trained at different levels of education to support the diagnosis, treatment, rehabilitation, and monitoring of individuals with mental health conditions. It ranges from community health workers to highly specialized personnel. (4) Administrative personnel: staff responsible for coordinating the service provision and performing administrative activities such as monitoring, billing, scheduling services, and managing demand, among others. (5) Departmental Health Secretariat: the governmental body responsible for directing, coordinating, and administering public health actions across the department, including the oversight and integration of a network of public and private health care service providers.

For the purposes of the study, the patient and caregiver groups were merged into a patient group since caregivers' tasks constituted a subset of patient tasks. Although the Departmental Health Secretariat was initially identified as a key stakeholder, it was not possible to include their participation in the study, as no officials were assigned to collaborate with the research activities.

User Recruitment Phase

The sample selection followed a convenience sampling. Potential users were recruited in 2 phases (September and October 2023) by the teaching team of the Faculties of Psychology of Universidad Pontificia Bolivariana at their sites in Medellín, Montería, and Palmira. The 3 types of roles (patient, clinical staff, and administrative personnel) with the greatest participation in the use of the mental telehealth system were included.

The inclusion criteria for the different types of participants were as follows: (1) *Patients and relatives* of patients older than 18 years (legal age in Colombia). (2) *Mental health care personnel* who were psychology or social work students in semesters of clinical internship (ninth semester of the degree or greater) or professionals in psychology or social work. (3) *Administrative or logistical* personnel with responsibilities in managing the population health program.

Potential users who did not have basic computer skills or who required support from others were excluded. After the initial call by the teaching team, the research team contacted the test participants to explain the evaluation procedure, and the synchronous testing spaces were coordinated through the Microsoft Teams platform. Participation was voluntary with verbal consent. The only data collected from individuals were those associated with the features of the type of user (age, gender, level of education, and occupation), excluding data that would allow their identification. The sessions were recorded on the Teams platform in case additional analysis was needed, and access was restricted to the research group.

Implementation Phase and Usability Evaluation by Scenarios

The scenario-based evaluation methodology was used, accompanied by cognitive walk-throughs observations, in which the user expressed his or her actions, thoughts, and interpretations of the tool aloud [17,18]. The usability tests were conducted individually, and the permissions and access to the respective screens and interfaces were verified at the beginning of each test. Moreover, a document with instructions and proposed tasks to be performed was shared with the users.

To understand different sources of tasks performance in the solution and provide detailed feedback on the need to improve the interface, we conducted task analysis, followed by usability analysis. After each scenario, each user filled out a Google Forms survey containing a questionary for task load measurement (the NASA [National Aeronautics and Space Administration] Task Load Index [NASA-TLX]) [19]. At the end of all the assigned scenarios, users filled out the Post-Study System Usability Questionnaire (PSSUQ) [20].

NASA Task Load Index

It is a self-administered questionnaire that seeks to measure the sources of workload and generates a load index that allows observing how a system affects the load of an individual when completing tasks during its use. This questionnaire has 2 parts: the first establishes the magnitude of each factor on a scale from 1 (low) to 10 (high), and the second makes a paired comparison to determine which factor is more important than another.

The factors contained in the scale are (1) Mental demand (M refers to the level of mental demand required by the task), (2) Physical demand (P refers to the level of physical demand), (3) Temporal demand (T refers to the pace imposed by the task), (4) Performance (P refers to the perceived performance in completing the task), (5) Effort (E refers to the level of effort and difficulty of the task), and (6) Frustration (F refers to how insecure, irritated, stressed, or annoyed one felt with the task) [19].

The NASA-TLX index is calculated as follows:

where is a constant indicating the number of factors, is the weighted loading source corresponding to a value assigned to each factor , corresponds to the number of times a factor is selected, and 15 is the number of possible comparisons between factors.

Poststudy System Usability Questionnaire

It is a self-managed questionnaire that seeks to measure the level of usability and satisfaction of users after interacting with interfaces or applications. Version 3 is composed of 16 items, which are scored on a Likert scale with 7 options, with 1 being Strongly Agree and 7 Strongly Disagree.

The global score of the scale is calculated with the average of all the scores obtained from the 16 statements or items. It also allows the calculation of three subscales, as follows: (1) System Usefulness (SYSUSE), which is the average of items 1- 6; (2) Information Quality (INFOQUAL), which corresponds to the average of items 7-12, and (3) Interface Quality (INTERQUAL), which is the average of items 13-15. These subscales provide more information and details about aspects of the software or system [20]. For the interpretation process, low scores refer to better levels of usability. According to Sauro and Lewis [21], the following score limits were considered adequate levels:

XSL•FO
RenderX

SYSUSE 2.80, INFOQUAL 3.02, INTERQUAL 2.49, and overall average 2.82.

A researcher (EPM) was present during all the tests through teleconference spaces in order to provide support and assistance without interfering in the evaluation activities. During the tests, the user evaluator reported the findings, and the accompanying researcher (EPM) recorded the application errors identified. Findings were consolidated in a project management platform (Azure DevOps), which served as input for the technical and development team of the evaluated mental health system. All the application errors identified during the testing scenarios were corrected for the second evaluation cycle.

Ethical Considerations

This study was reviewed and approved by the health research ethics committee of Universidad Pontificia Bolivariana, Medellín, Colombia (Minute 09 of 2022). All participants were informed of the objectives and procedures of the research prior to their involvement, and verbal informed consent was obtained before initiating the usability testing sessions. Participation was voluntary. To ensure privacy and confidentiality, no personally identifiable information was collected. Only general demographic characteristics (age, gender, level of education, and occupation) were recorded. Usability sessions were conducted via Microsoft Teams and recorded exclusively for research purposes, with access restricted to the research team. All data were anonymized prior to analysis. Participants did not receive any financial or in-kind compensation for their involvement. No identifiable images of participants are included in this manuscript or in supplementary materials.

Results

Scenario Preparation and Creation Phase

The summary of the scenarios created is shown in Table 1. Patients, clinical staff, and system administrators were identified as potential users. As shown, patients are the ones with the most types of interactions anticipated in the system.

XSL•FO
RenderX

Table 1. Summary of test scenarios by user role during the scenario-based usability evaluation of a participatory-designed clinical information system for mood disorder care in Córdoba, Colombia (2022-2023)[a].

Scenario	Scenario name	Statement
Patients		
1	Login	• You are a user who wants to register in the program, and you create an account in the system.
2	View directory	• You are a user of the system and want to see the list and location of the nearest mental health care providers.
3	Call to the hotline	• You are not feeling well today. You are having suicidal thoughts and want to contact the mental health hotline by telephone.
4	Emergency WhatsApp	• You are not feeling well today. You are having suicidal thoughts and want to contact the mental health hotline via WhatsApp (no need to type a message).
5	Make an appointment	• You want to make an appointment (psychology) of telecounseling by videoconference for today with the professional Eider Pereira.
6	Start the appointment	• You have an appointment scheduled for today and will attend by videoconference.
7	Data entry (scale)	• You are concerned about your mental health; you access the system and fill out the DASS-21 questionnaire.
8	See the information capsules	• You are a patient and want to know about depression in general. Your task is to find the information in the system.
9	Start forum post	• You are a user and want to make the following post in the forum: • Title: How I Feel Experience • Description: • Hello, my name is XXX. After seeing this space, I feel encouraged to share that many times I have felt down and stressed due to the large number of school assignments and exams. Studying for hours takes up a lot of time and prevents me from doing other activities I enjoy. • When exams are approaching, I tend to eat more, sleep less, and even bite my nails a lot. I sometimes feel like running away and leaving everything behind. My parents constantly remind me that I have to be the best and that sacrifices are necessary. They are very concerned about my academic performance. • I have few friends at school, and I would like to have more, but I am afraid they might not like me or want me to be part of their group. I don't know if others experience the same, but I often feel bored, unmotivated, and overwhelmed by the responsibilities I am expected to fulfill. • I know that stress affects me a lot, and I want to find a way to feel better. Has anyone gone through something similar? What do you recommend I do when I feel this way? • Thank you for your attention. • Category: Experiences
10	Forum comments	• You are a patient or user and would like to comment on the forum post called “Mis formas de afrontar la depresión” (My ways of coping with depression, in English). • Comment: Thanks for sharing it with the community
11	Questions and answers	• You are a user of the system, you have the following question about mental health and you ask it in the question section: What is depression?
Clinical staff		
1	View agenda	• You are a therapist and want to view the list of patients scheduled for today.
2	Register service	• You are tasked with moderating the mental health forum by approving or rejecting pending posts.
3	Moderate the forum	• You are tasked with moderating the mental health forum by approving or rejecting pending posts.

XSL•FO
RenderX

Scenario	Scenario name	Statement
4	Questions and answers	• You are going to answer the following user-submitted question pending approval: "What is depression?" • Answer: • Depression is an illness that affects your mental state, thereby altering your daily life. It is often characterized by persistent feelings of sadness and a loss of interest in activities such as eating, working, or studying, as well as disruptions to sleep patterns, including insufficient or nonrestorative sleep. These symptoms must persist for at least 4 weeks to be considered depression. Management of depression requires a multidisciplinary approach, which, depending on the patient's condition, may involve pharmacological treatment and/or psychotherapy. However, only a psychiatrist can establish the appropriate management plan, so I recommend scheduling a medical appointment with this professional. Additionally, I would like to understand why you are experiencing difficulties communicating with your mother. It is crucial to seek emergency care if you are unable to sleep at all or have thoughts that your life has no meaning. Please seek help from a psychiatric professional as soon as possible. (Excerpt from 1DOC3). • Category: Depression
Administrative		
1	Create the information capsule	• You are a provider leader in the system and want to update and add the following information in the capsule module: • Image: Any appropriate image can be placed here. • Title: How Do I Know If I Am Stressed? • Tags: Stress symptoms, Recommendations • Description: • Stress symptoms can affect your body, thoughts, feelings, and behavior. The most common symptoms associated with stress include headaches, muscle tension or pain, chest pain, fatigue, changes in sex drive, upset stomach, and sleep disturbances. If you have experienced more than one of these symptoms in the past week, it is very likely that you are experiencing signs of stress. Therefore, we encourage you to continue learning about stress management and to seek professional help if necessary. • Category: Stress
2	Manage care in the population board	• You are a population leader and must reschedule care for one of the patients.

[a]A total of 17 scenarios were defined: 11 for patients (eg, login, hotline contact, telecounseling, and forum use), 4 for clinical staff (eg, view agenda and moderate the forum), and 2 for administrative users (eg, population management and create information capsule).

User Recruitment Phase

A total of 30 individual users were recruited (14 patients, 8 people with a mental health care profile, and 8 people with administrative or logistical activities in the mental health care process). Some of the individuals participated in both evaluation cycles (Table 2).

Table 2. Demographic and educational characteristics of participants in the scenario-based usability evaluation of a participatory-designed clinical information system for mood disorder care in Córdoba, Colombia (2022-2023)[a].

	Average age (years)	Sex		Academic level			
		Male	Female	High school diploma	Technician	Bachelor's degree	Postgraduate
Patients	24.1	3	11	7	1	4	2
Clinical staff	27.4	1	7	5	0	2	1
Administrative personnel	27.0	1	7	4	0	1	3

[a]A total of 30 participants were included: 14 patients (average age of 24.1 years), 8 clinical staff (27.4 years), and 8 administrative personnel (27.0 years). The table details distribution by sex (male or female) and academic level (high school diploma, technical education, bachelor's degree, and postgraduate education).

Implementation Phase and Usability Evaluation by Scenarios

The evaluations were conducted via Microsoft Teams, where users shared their screen. The tests per user lasted approximately 45-90 minutes. Table 3 contains the consolidated number of participants by cycle and their respective roles in the evaluated system.

A total of 6 test groups were conducted (3 types of users in 2 cycles), including the 4 functionalities with the highest scores in the prioritization and impact process: login, providing mental

XSL•FO
RenderX

health care and monitoring (telecounseling), assistance to clinical staff in mental health emergencies (hotline), and population management (population dashboard).

All events were reported by the researcher (EPM) conducting the test. During the evaluations, 23 critical findings were identified that informed system improvements—9 were system errors and 14 were design errors. System errors refer to malfunctions in the application's functionality, whereas design errors were limited to issues within the user interface. Figure 2 provides an example of interface modifications based on these findings.

Table 3. Distribution of participants by test cycle in the scenario-based usability evaluation of a participatory-designed clinical information system for mood disorder care in Córdoba, Colombia (2022-2023)[a].

	Cycle 1	Cycle 2	Total (repeated)
Patients	8	7	15 (1)
Clinical staff	4	6	10 (2)
Administrative personnel	4	5	9 (1)
Total participants	16	18	34 (4)

[a]In total, 34 participations were recorded across 2 cycles (16 in cycle 1 and 18 in cycle 2), including patients, clinical staff, and administrative personnel. Four participants took part in both cycles.

Figure 2. Example of interface changes in the home screen of the clinical information system for mood disorder care, developed through participatory design and evaluated with a scenario-based usability approach in Córdoba, Colombia (2022-2023). The first version (A) offered limited access to information capsules, whereas the revised version (B), following 2 usability testing cycles, presented a redesigned layout highlighting educational resources on symptoms, depression, and mental health promotion.

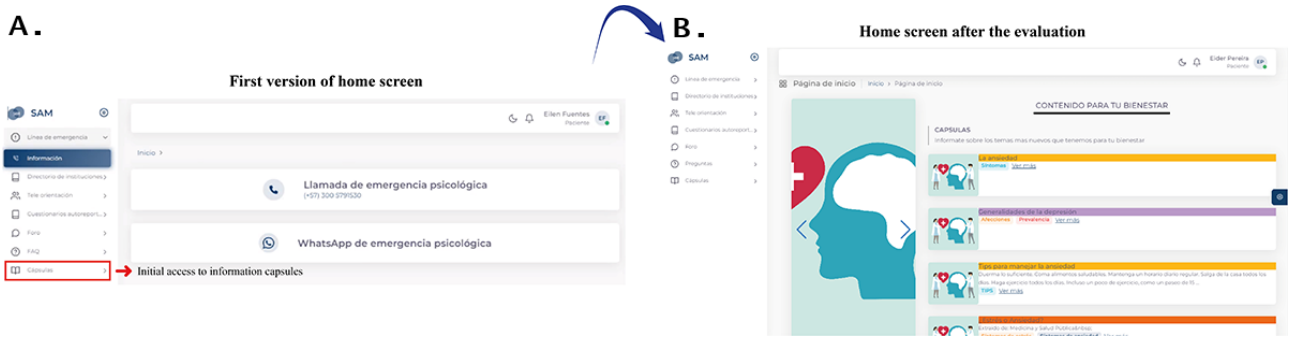

NASA-TLX Evaluations

A total of 223 individual task evaluations were carried out, where the workload was measured (112 and 111, respectively, in cycles 1 and 2); 165 (74%) were the patient role, 40 (17.9%) clinical staff, and 18 (8.1%) administrative personnel. Multimedia Appendix 1 contains the comparative changes between test cycles 1 and 2 for the scenarios evaluated for all types of users. Multimedia Appendix 1 shows the comparative summary of workload (NASA-TLX) by cycles, functionalities, and types of users. Table 4 shows further details of the components of the workload evaluation by the respective domains of effort.

XSL•FO
RenderX

Table 4. Results of workload scenarios (NASA [National Aeronautics and Space Administration] Task Load Index [NASA-TLX]) by domain type in the respective scenarios and test cycles for a participatory-designed clinical system for mood disorder care in Córdoba, Colombia (September-October 2023)[a].

Type of role and scenario	Cycle	Weighted average mental workload	Weighted Average physical workload	Average temporary workload	Weighted average performance workload	Weighted average effort	Weighted average frustration
Administrative							
1	1	12.8	0.8	6.8	21.8	10.5	18.0
1	2	11.2	0.2	10.8	8.2	6.6	4.2
2	1	10.0	0.3	5.3	10.5	9.8	4.0
2	2	11.8	0.8	12.6	13.4	12.2	11.0
Clinical							
1	1	9.8	5.8	12.5	1.5	4.8	0.8
1	2	4.0	2.8	5.5	9.3	3.5	1.3
2	1	17.8	4.3	12.5	24.3	15.8	1.0
2	2	19.0	3.8	10.3	10.8	13.0	5.3
3	1	13.3	0.5	9.3	19.0	6.3	0.5
3	2	15.0	9.0	10.3	11.7	13.0	2.7
4	1	20.3	0.3	9.8	16.8	12.0	2.0
4	2	19.0	3.0	7.5	14.3	8.0	7.7
Patient							
1	1	9.3	1.3	9.1	2.4	9.0	8.5
1	2	12.7	1.0	8.7	5.7	10.9	5.4
2	1	9.9	0.5	16.3	10.8	8.4	15.6
2	2	12.0	1.1	9.1	9.0	10.3	11.7
3	1	9.8	2.6	15.9	2.4	8.6	4.5
3	2	7.1	2.6	9.0	8.0	10.7	8.3
4	1	7.6	2.3	19.1	0.9	7.5	1.8
4	2	8.3	1.7	6.4	2.4	12.0	11.3
5	1	13.5	2.8	21.1	1.9	13.6	8.8
5	2	16.3	1.7	7.0	7.4	14.7	10.1
6	1	8.8	2.0	15.3	6.0	13.1	8.3
6	2	9.0	1.1	9.0	8.1	12.3	12.1
7	1	20.1	1.1	11.0	3.8	12.5	5.6
7	2	15.4	2.6	12.6	2.1	13.1	7.3
8	1	10.1	2.9	14.4	1.4	12.6	2.4
8	2	10.9	2.9	7.9	5.9	12.4	4.9
9	1	16.3	3.5	14.5	8.3	15.9	3.1
9	2	15.6	3.3	6.3	3.7	14.1	8.9
10	1	9.5	2.0	18.9	1.5	11.5	0.4
10	2	11.3	1.9	8.9	6.1	9.3	1.9
11	1	9.5	2.1	19.3	2.1	10.0	0.3
11	2	7.6	1.9	8.3	3.6	9.4	2.0

[a]Detailed breakdown of the workload components, as measured by the NASA-TLX Task Load Index, for the administrative, clinical, and patient user roles across 2 test cycles. Lower values correspond to lower perceived workload and better performance, and higher values correspond to higher perceived workload.

XSL•FO
RenderX

PSSUQ Evaluations

Thirty-four usability evaluations were obtained with the PSSUQ tool, which reports a level of usability in global terms by cycles that varied only from 2.2 to 2.3. The results by type of user and their corresponding measurement with the PSSUQ tool showed a tendency in both test cycles toward a better usability rating in the interfaces oriented to clinical and patient users. For example, in cycle 2, there was a difference of 2.0 and 1.9, respectively, versus a score of 3.0 in the administrative interfaces (Tables 5 and 6).

Table 5. Overall system usability results, measured by the PSSUQ, for 2 evaluation cycles of a participatory-designed clinical information system for mood disorder care in Córdoba, Colombia (2022-2023)[a].

	SYSUSE[b]	INFOQUAL[c]	INTERQUAL[d]	Overall
Cycle 1	2.1	2.3	2.2	2.2
Cycle 2	2.2	2.5	2.1	2.3

[a]The Post-Study System Usability Questionnaire scores are presented for each of the 2 test cycles. A low score indicates better usability and greater user satisfaction. The results are broken down into 3 subscales: SYSUSE, INFOQUAL, and INTERQUAL, along with an overall average.

[b]SYSUSE: System Usefulness.

[c]INFOQUAL: Information Quality.

[d]INTERQUAL: Interface Quality.

Table 6. Post-Study System Usability Questionnaire, broken down by user role and test cycle of a participatory-designed clinical information system for mood disorder care in Córdoba, Colombia (2022-2023)[a].

User role	Cycle 1				Cycle 2			
	SYSUSE[b]	INFOQUAL[c]	INTERQUAL[d]	Overall	SYSUSE[b]	INFOQUAL[c]	INTERQUAL[d]	Overall
Administrative	3.5	4.3	3.9	3.9	2.9	3.4	2.7	3.0
Clinical	1.3	1.5	1.3	1.4	2.2	2.1	1.7	2.0
Patient	1.7	1.7	1.9	1.7	1.7	2.2	2.0	1.9

[a]The results compare Post-Study System Usability Questionnaire (PSSUQ) scores across 2 test cycles for administrative, clinical, and patient users. Lower scores on the PSSUQ indicate a better level of usability. The scores are provided for the SYSUSE, INFOQUAL, INTERQUAL, and overall usability.

[b]SYSUSE: System Usefulness.

[c]INFOQUAL: Information Quality.

[d]INTERQUAL: Interface Quality.

Discussion

Principal Findings

This study evaluated the usability of a clinical telehealth system for mood disorder care developed through participatory design. The findings show that the system achieved usability scores (PSSUQ 2.2 and 2.3) and improvements in workload between evaluation cycles, particularly for patient-centered tasks, meeting the study's objective of assessing usability through a scenario-based methodology.

One of the challenges in systems evaluation is the granularity of evaluation at the task level and the order in which such evaluation is performed. The scenario-based evaluation approach allowed planning an evaluation strategy informed by an objective prioritization of functions and tasks. In addition, a separation of user groups was achieved, and common tasks that did not require additional users for validation were identified.

One of the challenges of system evaluation is the granularity of evaluation at the task level and the order in which such evaluation is performed. The scenario-based evaluation approach allowed planning an evaluation strategy informed by an objective prioritization of functions and tasks. In addition, a separation of user groups was achieved, and common tasks that did not require additional users for validation were identified.

The system can be considered to have good usability, given an overall PSSUQ score between 2.2 and 2.3 in cycles 1 and 2, considering that the lower the score, the better the performance. The reference value is 2.8 [20]. These findings suggest that the participatory design process helped anticipate user needs and improve the quality of the product before deployment However, it was also possible to identify elements that previous heuristic evaluations were unable to detect. This confirms how user participation in the design and early usability evaluation contribute to good system usability and help identify opportunities for system improvement before the system goes into production, optimizing the use of development resources. This is consistent with findings from previous studies showing that iterative and participatory usability evaluations with both clinicians and end users can significantly improve system navigation, content structure, and user satisfaction prior to system deployment [22]. This reinforces the importance of iterative usability evaluations early in the system development life cycle to ensure alignment with user needs and task requirements [23].

XSL•FO
RenderX

It is interesting to note how system usability scores were better for users with more functional and task representation. Patients rated the system higher than administrators. This can be explained by the prioritization in the development of interfaces informed by the DO, where development was carried out according to task prioritization. It should also be noted that the differences observed between groups can be attributed to the different types of tasks assigned to each group. Tasks prioritized for patients tended to be simpler and more user-centered, whereas tasks assigned to administrative and clinical users were more complex, involving greater cognitive demands. The differences in scores between cycles 1 and 2 cannot be established as statistically significant given the sample size. However, changes may be due to individual variability. A possible future evaluation method to address this could involve repeated measurements with the same users or increasing the number of evaluations. The number of evaluations may be limited in scenarios where the number of potential users is insufficient. Optional usability evaluations may be implemented in the future for all individuals registered in the mental health program.

When analyzing the sources of task workload, there is an improvement in the NASA-TLX workload index for the vast majority of tasks performed in the second version of the system, except for the forum modeling and population management tasks (associated with clinical users and system administrators). This can be explained by the higher priority given to system changes benefiting patient and family-type users and by the addition of interface elements that had not been subject to group discussion. The most significant source of workload is mental workload, as expected. Future research could focus on improving the information available on-screen to reduce cognitive demands. As anticipated, physical workload is minimal since the interaction does not require extensive clicking.

Limitations

This study has some limitations. The convenience sample size restricts the generalizability of findings, and the number of evaluations was insufficient to establish statistically significant differences between user groups. In addition, the evaluation of the synchronous interactive telehealth component was not carried out, as this would require a different experimental design. These limitations are consistent with findings from systematic reviews, which report that usability evaluations in mobile mental health interventions often rely on small sample sizes and primarily summative assessment methods [24]. In the future, the mobile interface of the mental health system will be evaluated, as it is considered a highly useful means of accessing and using the mental health system for end users (patients and relatives).

Conclusions

This study shows that combining participatory design with scenario-based usability evaluation enables early identification of usability challenges and iterative refinement of digital health systems. Beyond the specific case of Córdoba, Colombia, these findings have broader implications: engaging end users early, prioritizing patient-facing functionalities, and conducting iterative usability assessments can enhance adoption, optimize resources, and support sustainable models of mental health care in underserved regions.

Acknowledgments

This work is part of the project development of strategies for comprehensive mental health care for people during and after confinement caused by the state of emergency in the Department of Córdoba, financed with resources from the General System of Royalties Colombia. Call for proposals from the Science Technology and Innovation Fund of the General System of Royalties for the Strengthening of Regional Research and Development Capacities and initiatives for the Development and Transfer of Technology and Knowledge for Innovation aimed at addressing problems arising from the COVID-19 pandemic.

Data Availability

The datasets generated and analyzed during this study are not publicly available due to the sensitive nature of the data, which include recorded usability test sessions with a potential risk of participant identification. To protect participant privacy and confidentiality, access to these data is restricted. Deidentified data derived from questionnaires and usability evaluation forms may be made available from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Comparative summary of perceived workload (NASA-TLX) across 2 usability test cycles of a participatory-designed clinical information system for mood disorder care in Córdoba, Colombia (2022-2023). The results are shown separately for (A) administrative personnel, (B) clinical staff, and (C) patients, illustrating changes in mental workload associated with different tasks (eg, managing care dashboards, scheduling appointments, forum use, and accessing information capsules). Lower scores represent reduced workload after system adjustments. NASA-TLX: NASA Task Load Index.

[PNG File , 83 KB - formative_v9i1e58909_app1.png]

XSL•FO
RenderX

References

1. Rehm J, Gmel GE, Gmel G, Hasan OSM, Imtiaz S, Popova S, et al. The relationship between different dimensions of alcohol use and the burden of disease-an update. Addiction 2017;112(6):968-1001 [FREE Full text] [doi: 10.1111/add.13757] [Medline: 28220587]
2. Depression and other common mental disorders. Global health estimates. Institutional Repository for Information Sharing. Pan American Health Organization. 2017. URL: https://iris.paho.org/handle/10665.2/34006 [accessed 2025-09-16]
3. Vigo D, Thornicroft G, Atun R. Estimating the true global burden of mental illness. Lancet Psychiatry 2016;3(2):171-178. [doi: 10.1016/S2215-0366(15)00505-2] [Medline: 26851330]
4. Patel V, Saxena S, Lund C, Thornicroft G, Baingana F, Bolton P, et al. The Lancet Commission on global mental health and sustainable development. Lancet 2018;392(10157):1553-1598. [doi: 10.1016/S0140-6736(18)31612-X] [Medline: 30314863]
5. Thornicroft G, Chatterji S, Evans-Lacko S, Gruber M, Sampson N, Aguilar-Gaxiola S, et al. Undertreatment of people with major depressive disorder in 21 countries. Br J Psychiatry 2017;210(2):119-124 [FREE Full text] [doi: 10.1192/bjp.bp.116.188078] [Medline: 27908899]
6. Simancas Fernández M, Sánchez Novoa D, Uribe Urzola A, Ruiz González EP, Muñoz Argel M. Post–COVID-19 mental health context in Córdoba, Colombia: a documentary review. Ciencia Latina 2023;7(5):10571-10590. [doi: 10.37811/cl_rcm.v7i5.8680]
7. Lattie EG, Adkins EC, Winquist N, Stiles-Shields C, Wafford QE, Graham AK. Digital mental health interventions for depression, anxiety, and enhancement of psychological well-being among college students: systematic review. J Med Internet Res 2019;21(7):e12869 [FREE Full text] [doi: 10.2196/12869] [Medline: 31333198]
8. Shneiderman B, Plaisant C. Designing the user interface: strategies for effective human-computer interaction. Appl Ergon 1993;24(4):295. [doi: 10.1016/0003-6870(93)90471-k]
9. Sjöberg C, Timpka T. Participatory design of information systems in health care. J Am Med Inform Assoc 1998;5(2):177-183 [FREE Full text] [doi: 10.1136/jamia.1998.0050177] [Medline: 9524350]
10. Rubin J, Chisnell D. Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests. Hoboken: NJ: John Wiley & Sons; 2008.
11. Zhang J, Walji MF. TURF: toward a unified framework of EHR usability. J Biomed Inform 2011;44(6):1056-1067 [FREE Full text] [doi: 10.1016/j.jbi.2011.08.005] [Medline: 21867774]
12. Pozuelo JR, Moffett BD, Davis M, Stein A, Cohen H, Craske MG, et al. User-centered design of a gamified mental health app for adolescents in sub-saharan Africa: multicycle usability testing study. JMIR Form Res 2023;7:e51423 [FREE Full text] [doi: 10.2196/51423] [Medline: 38032691]
13. Danieli M, Ciulli T, Mousavi SM, Riccardi G. A conversational artificial intelligence agent for a mental health care app: evaluation study of its participatory design. JMIR Form Res 2021;5(12):e30053 [FREE Full text] [doi: 10.2196/30053] [Medline: 34855607]
14. Deniz-Garcia A, Fabelo H, Rodriguez-Almeida AJ, Zamora-Zamorano G, Castro-Fernandez M, Alberiche Ruano MDP, WARIFA Consortium. Quality, usability, and effectiveness of mHealth apps and the role of artificial intelligence: current scenario and challenges. J Med Internet Res 2023;25:e44030 [FREE Full text] [doi: 10.2196/44030] [Medline: 37140973]
15. US General Services Administration. Usability. Digital.gov. 2023. URL: https://www.usability.gov/how-to-and-tools/methods/scenarios.html [accessed 2023-09-24]
16. 2018 National population and housing census. Departamento Administrativo Nacional de Estadística (DANE). 2018. URL: https://www.dane.gov.co/index.php/estadisticas-por-tema/demografia-y-poblacion/censo-nacional-de-poblacion-y-vivenda-2018 [accessed 2025-04-25]
17. Nielsen J. Usability inspection methods. 1944 Presented at: CHI '94: Conference Companion on Human Factors in Computing Systems; April 24-28, 1994; Boston, MA p. 413-414. [doi: 10.1145/259963.260531]
18. Lewis C, Polson PG, Wharton C, Rieman J. Testing a walkthrough methodology for theory-based design of walk-up-and-use interfaces. 1990 Presented at: CHI '90: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems; April 1-5, 1990; Seattle, WA p. 235-242. [doi: 10.1145/97243.97279]
19. Hart SG, Staveland LE. Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. Adv Psychol 1988;52:193-183. [doi: 10.1016/s0166-4115(08)62386-9]
20. Lewis JR. IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use. In J Hum Comput Interact 1995;7(1):57-78. [doi: 10.1080/10447319509526110]
21. Sauro J, Lewis JR. Quantifying the User Experience: Practical Statistics for User Research. Amsterdam: Elsevier; 2012.
22. Wozney L, Baxter P, Newton AS. Usability evaluation with mental health professionals and young people to develop an Internet-based cognitive-behaviour therapy program for adolescents with anxiety disorders. BMC Pediatr 2015;15:213 [FREE Full text] [doi: 10.1186/s12887-015-0534-1] [Medline: 26675420]
23. Yen P, Bakken S. Review of health information technology usability study methodologies. J Am Med Inform Assoc 2012;19(3):413-422 [FREE Full text] [doi: 10.1136/amiajnl-2010-000020] [Medline: 21828224]
24. Inal Y, Wake JD, Guribye F, Nordgreen T. Usability evaluations of mobile mental health technologies: systematic review. J Med Internet Res 2020;22(1):e15337 [FREE Full text] [doi: 10.2196/15337] [Medline: 31904579]

XSL•FO
RenderX

Abbreviations

DO: Domain Ontology
GOMS: Goals, Objectives, Methods, Selection
INFOQUAL: Information Quality
INTERQUAL: Interface Quality
NASA-TLX: National Aeronautics and Space Administration Task Load Index
PSSUQ: Post-Study System Usability Questionnaire
SYSUSE: System Usefulness

Edited by A Mavragani; submitted 27.03.24; peer-reviewed by A Mortezapour; comments to author 13.11.24; revised version received 28.04.25; accepted 28.08.25; published 06.10.25.

Please cite as:
Torres-Silva EA, Gaviria-Jiménez JJ, Pereira-Montiel E, Montoya-Arenas DA, Flórez-Arango JF
Clinical System for Mood Disorder Care in Córdoba, Colombia: Participatory Design and Scenario-Based Usability Evaluation Study
JMIR Form Res 2025;9:e58909
URL: https://formative.jmir.org/2025/1/e58909
doi:10.2196/58909
PMID:41051853

©Ever Augusto Torres-Silva, Juan José Gaviria-Jiménez, Eider Pereira-Montiel, David Andrés Montoya-Arenas, José Fernando Flórez-Arango. Originally published in JMIR Formative Research (https://formative.jmir.org), 06.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Assessing a Community Health Worker-Facilitated, Digitally Delivered, Family-Centered Diabetes Management Program: Single-Arm Quasi-Experimental Study

Zenong Yin[1], PhD; Vanessa L Errisuriz[2], PhD; Heather Cuevas[3], PhD; Bertha E Flores[4], PhD; Laura Delfausse[5], MS; Christina Galvan[5], MS; Jing Wang[6], PhD; Chengdong Li[6], PhD; Renata Morfin[5], MS; Shiyu Li[7], PhD; Maysa Sapargeldiyeva[1], BA; Giliane Yza Muyna[1]; Minyu Zhang[5], PhD; Vanessa Sweet[8], MPH; Deborah Parra-Medina[9], PhD

[1]College of Health, Community and Policy, University of Texas at San Antonio, San Antonio, TX, United States
[2]Department of Public Health, Santa Clara University, Santa Clara, CA, United States
[3]School of Nursing, University of Texas at Austin, Austin, TX, United States
[4]School of Nursing, University of Texas Health Science Center San Antonio, San Antonio, TX, United States
[5]Latino Research Institute, University of Texas at Austin, Austin, TX, United States
[6]College of Nursing, Florida State University, Tallahassee, FL, United States
[7]School of Kinesiology, Louisiana State University, Baton Rouge, LA, United States
[8]Greater Austin YMCA, Austin, TX, United States
[9]Department of Family Medicine, University of Colorado Denver Anschutz Medical Campus, Aurora, CO, United States

Corresponding Author:
Deborah Parra-Medina, PhD
Department of Family Medicine
University of Colorado Denver Anschutz Medical Campus
130001 East 17th Place Suite CG009
Aurora, CO, 80045-2559
United States
Phone: 1 303 724 6347
Email: deborah.parra-medina@cuanschutz.edu

Abstract

Background: The high prevalence of type 2 diabetes (T2D) and associated complications disproportionately affect low-income Latino populations, who also experience disparities in diabetes self-management (DSM), including poor medication adherence, physical activity, diet, and glycemic control.

Objective: This study examined, through an academic-community partnership, the effectiveness of ¡Salud, Salud! (an evidence-based, family-centered diabetes self-management education and support [DSMES] program) on primary (glycemic control and quality of life) and secondary (social, psychological, and behavioral factors related to T2D management) outcomes among low-income Latino adults with T2D or prediabetes.

Methods: In total, 81 adults (mean age 48.90 years, SD 12.57; n=57, 70.4%, female; n=66, 81.5%, Latino) with T2D or prediabetes were enrolled in a 12-week, single-arm quasi-experimental study conducted in two Central Texas Young Men's Christian Association (YMCA) locations. ¡Salud, Salud! incorporated individual coaching by community health workers (CHWs), online family-centered DSMES training lessons, and a YMCA family membership. The delivery of ¡Salud, Salud! was supported and facilitated by digital technologies, including a dashboard to deliver intervention content and monitor participants' engagement in intervention activities. Outcomes measured at baseline and 12 weeks (ie, postintervention) included hemoglobin A1c (HbA1c); quality of life; anthropometrics; self-reported physical activity and diet; mindfulness; perceived stress; and diabetes-related knowledge, self-efficacy, and support. Participant engagement in program activities was assessed via four index variables that underlay multiple dimensions of influences on ¡Salud, Salud! uptake: family engagement and support, participation in self-management education, program support and facilitation, and participation in self-monitoring. Paired t-tests and McNemar chi-square tests were used to examine the change in outcomes from baseline to 12 weeks. The number of program activities

XSL•FO
RenderX

participants completed for each engagement index variable was converted to percentages to estimate the mean proportion of activities completed.

Results: In total, 48 (59.3%) participants completed the 12-week posttest. At the end of the program, participants demonstrated a marginally significant reduction in HbA1c (–0.30%, *P*≤.09) and a significant increase in participants reporting good-to-excellent health from baseline (n=19, 39.6%) to posttest (n=28, 58.3%; *P*≤.003). There were significant reductions in body weight (–1.30 kg, *P*=.02), body fat percentage (–1.26%, *P*=.01), perceived stress (–0.28, *P*=.02), added sugar intake (–2.15 teaspoons/day, *P*=.001), and time spent sedentary per week (–70.27 minutes, *P*=.003) from baseline to posttest. Mindfulness increased significantly (2.21, *P*=.01). Participant engagement in ¡Salud, Salud! varied, with participants exhibiting a high completion rate in program support and facilitation activities (88%) and a moderate-to-low completion rate in self-management training (66%), self-monitoring (56%), and family engagement and support (49%) activities.

Conclusions: ¡Salud, Salud! shows promising preliminary effects on key diabetes-related outcomes. Future research should investigate how to enhance participant engagement and optimize uptake of evidence-based T2D self-management practices among low-income Latino adults with diabetes.

(JMIR Form Res 2025;9:e79032) doi:10.2196/79032

KEYWORDS

health equity; academic-community partnership; community health workers; diabetes self-management; diabetes education; diabetes support

Introduction

The prevalence of type 2 diabetes (T2D) in the United States is 14.0% [1], while prediabetes affects 48% of US adults [2]. In 2017, the economic burden of diabetes reached US $327.2 billion, translating to an annual cost of US $13,240 per patient [3]. T2D accounts for 90%-95% of diabetes cases [4] and disproportionately affects low-income and racial/ethnic minority communities, particularly Hispanics/Latinos (referred to hereafter as Latinos)—the largest racial/ethnic minority group in the United States, representing 18% of the population. During 2013-2016, T2D prevalence was 19.8% in Latinos compared to 12.4% in non-Latino Whites [1], with recent data indicating that diabetes incidence has decreased in non-Latino Whites but not in other racial and ethnic groups [5].

Low-income Latino populations also experience significant disparities in diabetes self-management education and support (DSMES) outcomes [1,6]. These disparities manifest as lower diabetes medication adherence rates [7], reduced likelihood of meeting recommended physical activity and dietary guidelines [1], and poorer glycemic control. Notably, only 8.4% of Latinos in the Hispanic Community Health Study/Study of Latinos met the goals for glycemic control [8]. These challenges contribute to higher risks for diabetes complications and mortality compared to non-Latino Whites.

Multiple factors influence diabetes self-management (DSM) practices among low-income minority individuals with T2D and their families [9,10]. These factors can be categorized into three primary domains: personal and interpersonal, social determinants of health, and cultural and social barriers [11,12]. Key personal and interpersonal factors include DSM knowledge and skills, DSM efficacy, adherence to DSM behaviors, T2D-related stress and anger, and family and friend support. Additional factors include time conflicts with family responsibilities and use of digital technologies [13]. Access to DSMES is often impeded by social determinants, such as environmental and transportation barriers, lack of health insurance, language and culture incongruence with health care providers, concerns about documentation status, economic hardships, and limited health literacy. Cultural and social barriers also deter DSMES among low-income racial/ethnic minority populations, including constrained financial resources, acculturative stress, physically demanding occupations, family values, and culturally specific preferences for food and physical activity [14].

The persistent disparities in diabetes outcomes among Latinos, coupled with the unique cultural and social factors affecting the uptake of evidence-based DSM practices, highlight the need for innovative, culturally responsive approaches to diabetes management [15-17]. Family-centered interventions may be particularly relevant for Latino populations, given the cultural emphasis on *familismo*—the central role of family in daily life and health decisions [18]. Additionally, the integration of community health workers (CHWs) and digital technologies offers promising strategies to overcome social determinant–related barriers to care, while maintaining cultural concordance [13,19].

To address these needs, we developed and evaluated *¡Salud, Salud!*, an evidence-based, family-centered DSMES program that leverages both CHW facilitation and digital technologies. This program aimed to improve glycemic control and health-related quality of life (HRQoL) for people with T2D and prediabetes from low-income predominantly Latino families in Austin, Texas. The study evaluated the program's impact on primary outcomes (glycemic control and quality of life) and secondary outcomes (social, psychological, and behavioral factors related to DSM). We hypothesized that participants would demonstrate significant reductions in hemoglobin A1c (HbA1c) and improvements in self-reported general health following the 12-week intervention.

XSL•FO
RenderX

Methods

Study Design

A quasi-experimental, single-group, pretest-posttest design was used to evaluate the preliminary effectiveness of the ¡Salud, Salud! program in improving glycemic control and HRQoL. The study planned to recruit 98 dyads. Participant dyads consisted of one individual with T2D or prediabetes and a family member or fictive kin who lived with them or in the same geographic area. Participants who did not have a family member or fictive kin were permitted to enroll in the study. Findings related to the implementation of ¡Salud, Salud! have been published elsewhere [20].

Study Setting

The geographic focus for this study was the city of Austin in Travis County, Texas. In Travis County, 33.9% of the population is Latino, 7.8% Black, and 7.6% Asian, and 31.1% are speakers of a non-English language (most commonly Spanish). Social disadvantage and vulnerability in Austin are concentrated in an area known as the Eastern Crescent, a geographic area of the city that has experienced historical economic disinvestment that has shaped access to critical resources needed to maintain health for community residents. The Eastern Crescent has a high proportion of Latino residents (53.8%), and almost one-third live below the poverty level, lack personal transportation, and have no health insurance. The prevalence of obesity (40%) and diabetes (16%) in the Eastern Crescent is higher than in the county and state. Given the high concentration of economically disadvantaged Latinos with limited English proficiency in the Eastern Crescent, we focused our efforts on providing the program to this subpopulation. In consultation with our community partner, the Young Men's Christian Association (YMCA) of Austin, we identified two YMCA sites located within the Eastern Crescent to deliver the program.

Recruitment and Eligibility

Multiple strategies were used to recruit individuals with T2D or prediabetes into the program. Flyers were distributed and posted at various community organizations that served residents of the Eastern Crescent (eg, safety net health clinics, food banks, churches). The program was also promoted through the YMCA's e-newsletters, website, and social media platforms, such as Facebook. All flyers and promotional materials were available in English and Spanish and included a QR code linking to the YMCA's ¡Salud, Salud! program webpage. This webpage included a link to the study's online eligibility screening form hosted on REDCap, a secure web platform for data management. Interested individuals could either complete the eligibility screening form directly or provide their contact information (ie, name, telephone number, email address) if they preferred to speak with research staff. In such cases, research staff followed up over the phone to conduct eligibility screening. In addition, bilingual (English and Spanish) CHWs and research staff conducted in-person outreach by tabling at safety net clinics, community events, and YMCA sites. Using internet-enabled tablets, they informed attendees about the program and conducted onsite eligibility screenings in English or Spanish. All eligible individuals were referred to research staff to provide informed consent and complete baseline assessments.

Individuals were eligible if they reported that they had physician-diagnosed T2D or prediabetes, were at least 12 years old, had a smartphone, lived in Austin, and were willing to travel to program sites. Individuals were excluded from participation if they could not speak and read English or Spanish; were pregnant, nursing, or planning to become pregnant during the study; planned to relocate in the next 6 months; had a serious illness or unstable condition requiring supervision of diet and activity beyond T2D control; had a severe psychiatric illness that would prohibit following self-monitoring recommendations; or were unable or unwilling to complete survey instruments and assessment procedures. Participants were highly encouraged, but not required, to identify a family member, fictive kin, or community member to participate in the program with them as a support person. The support person needed to be at least 18 years old, have a smartphone, live in the same county as the participant, and be able and willing to complete survey instruments and assessment procedures. Enrollment was not limited by sex, sexual orientation, race, or ethnicity.

We aimed to recruit 98 participants with T2D or prediabetes. This sample size was calculated based on detecting a change in HbA1c at 12 weeks. Using data from a previous study showing a mean 0.1% HbA1c decrease (SD 0.29%) [21], we determined that 68 participants would provide 80% power with a 5% significance level, accounting for 30% attrition at the posttest.

Intervention

To promote the uptake and application of DSMES practices by participants and their support persons, the information-motivation-behavioral skills (IMB) model was used to guide development of ¡Salud, Salud! (see Figure 1 for the conceptual model). According to the IMB model, successful diabetes management depends on three factors: Individuals with diabetes must be (1) informed of the knowledge and benefits of DSM, (2) motivated by personal (eg, beliefs and commitment) and social (eg, family support) factors related to DSM, and (3) equipped with behavioral skills to engage in DSM [3,4]. ¡Salud, Salud! also uses evidence-based strategies and practices from the "Standards of Medical Care in Diabetes" [22] and the DSMES Toolkit [19]. Evidence-based strategies included the use of mobile health (mHealth) and CHWs (nonmedical public health workers who connect communities to health care and social service providers) to deliver the program and engage participants with T2D and their designated support persons. Table 1 shows the components of the ¡Salud, Salud! intervention.

XSL•FO
RenderX

Figure 1. ¡Salud, Salud! conceptual model.

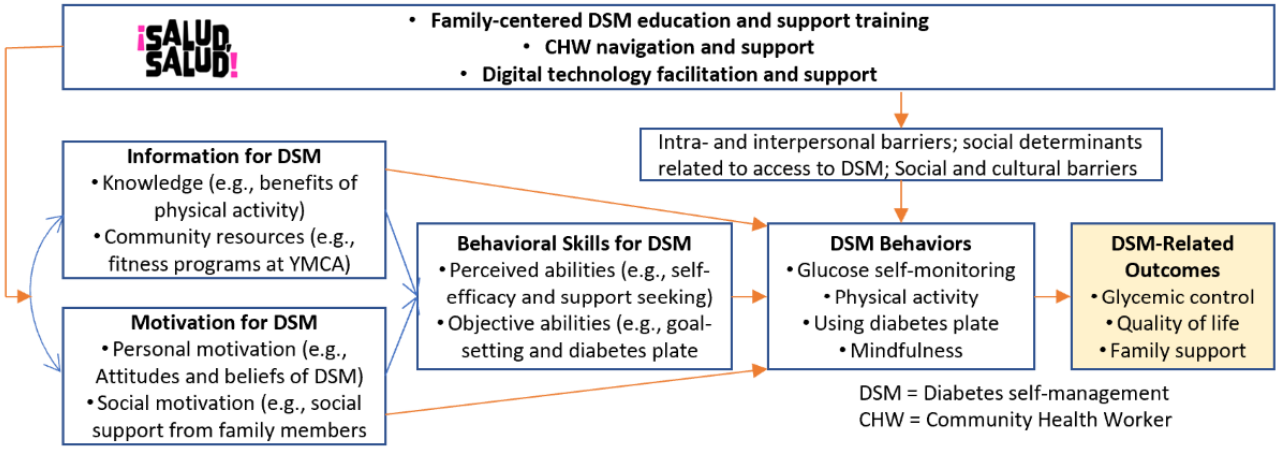

Table 1. ¡Salud, Salud! intervention components.

Component	Content and delivery
Family-centered DSMES[a] training	• Onsite, in-person program orientation for participants and their support persons with a CHW[b] • 1 in-person or virtual nutrition counseling session with a nutrition educator • 8 online, interactive DSMES training lessons for participants and their support persons to increase family support and uptake of DSM[c] practices
CHW navigation and support	• 12 weekly coaching sessions for support and problem solving with a CHW • Goal setting and action planning for DSM practices with a CHW
Digital technology facilitation and support for DSM via the Connected Health Dashboard	• Weekly online lesson delivery and lesson completion tracking • Text message delivery (motivation, tips, and program reminders) • Weekly self-monitoring survey delivery and tracking

[a]DSMES: diabetes self-management education and support.

[b]CHW: community health worker.

[c]DSM: diabetes self-management.

Participants and their designated support persons (when available) were introduced to the ¡Salud, Salud! program through a one-time, in-person orientation session held at a YMCA program site of their choosing. During this session, they met with their CHWs and learned about program goals and expectations. Participants were registered for a free 6-month YMCA family membership (ie, 2 adults and up to 4 children). The YMCA membership included access to all in-person and virtual YMCA programming, as well as a complementary personal training session. One week following the orientation, participants attended a one-on-one nutrition counseling session, which included an introduction to the American Diabetes Association's Diabetes Plate Method and individualized dietary counseling. The CHWs also introduced participants to the digital tools they would use during the program, including the YMCA mobile app, the Connected Health Dashboard [23], and the Insight Timer app [24]. Insight Timer is a popular meditation app that offers a wide variety of guided meditations, music tracks, and talks from mindfulness experts in various languages, including English and Spanish, to help users reduce stress, improve sleep, and enhance overall well-being [24].

The delivery of self-guided DSMES was facilitated by an online mHealth platform, the Connected Health Dashboard [23]. The dashboard is designed to allow participants the flexibility to view DSMES online lessons and perform self-monitoring activities at their convenience. All participants received eight narrated and interactive core lessons. The core lessons have been designed to provide strategies that engage the whole family in health practices that could positively impact participants' DSM and overall family health and well-being. A supplemental resource library of behavioral lessons was made available so that participants could receive additional training and support tailored to their individual behavioral change needs (eg, smoking, sleep, mindfulness). All content available on the dashboard could be viewed in English or Spanish, based on participant preference.

Each week, participants received a text message with a link to complete a brief self-monitoring survey to track the following behaviors: physical activity, healthy eating, mindfulness practice, and blood glucose monitoring. The dashboard linked to and displayed data collected from various digital sources (eg, lesson completion, surveys) so that participants and CHWs could monitor progress toward behavioral goals. The dashboard was also used to disseminate program text messages developed to provide informational and motivational support and enhance participant engagement with the dashboard.

XSL•FO
RenderX

CHWs provided personalized support to program participants via 11 weekly one-on-one encounters. These encounters were offered in person or remotely based on participants' preferences. During these encounters, participants selected and set SMART goals for two behavioral focus areas from the following: physical activity, healthy eating, mindfulness practice, and blood glucose monitoring. CHWs used motivational interviewing techniques to guide participants toward achieving these goals. CHWs were highly trained, having completed the CHW certification program established by the Texas Department of State Health Services [5], the Diabetes Paraprofessional Level 2 certification program established by the Association of Diabetes Care & Education Specialists [6], and a 3-day training (~24 hours) offered by study research staff about how to deliver the ¡Salud, Salud! intervention. Training topics covered by study research staff included motivational interviewing techniques, review of interactive online DSMES modules, and use of the Connected Health Dashboard.

Study Measurements

Assessment of study outcomes followed standardized data collection protocols conducted at baseline and immediately following completion of the 12-week program. Research assistants contacted participants over the phone and via text messages to schedule appointments to complete data collection in person at YMCA sites. Participants completed self-reported questionnaires related to quality of life, DSMES-related psychosocial factors and behaviors, and demographics. All questionnaires are available in English and Spanish, are reliable and valid, and have been used with Latino, Spanish-speaking, or economically disadvantaged populations. Standardized measurements of participant blood glucose levels and anthropometrics were also conducted in person by trained research assistants at YMCA sites. No incentives were provided for completing assessments. For the study, we included only data from participants with T2D or prediabetes.

Primary Outcomes

HbA1c Tests

Blood glucose levels were assessed using the A1CNow+, a certified method listed in the National Glycated Hemoglobin Standardization Program, that analyzes average blood glucose levels over a 3-month period [7]. Results are interpreted as follows: normal range (<5.7%), prediabetes (5.7%-6.4%), and T2D (≥6.5%).

Health-Related Quality of Life

Participants' HRQoL was measured with the four core items of the CDC Healthy Days HRQOL Survey, which is designed to assess an individual's perceived sense of well-being through a brief questionnaire [8]. Participants rated their general health on a scale from poor to excellent and reported the number of days over the past month during which they felt physically or mentally unhealthy. An "unhealthy days" summary score was calculated, which included the number of days (ranging from 0 to 30) that an individual felt either mentally or physically unhealthy (or both).

Secondary Outcomes

Body Composition

The Tanita Body Composition Analyzer SC-331S was used to measure participants' weight (to the nearest 0.1 kg) and body fat percentage via bioelectrical impedance analysis.

Dietary Intake

The 26-item NCI Dietary Screener Questionnaire (DSQ) [25] assessed diet quality. Participants responded to questions about the frequency and amount of selected food/drink intake over the past month, and estimates of intake values for fruits, vegetables, fiber, and added sugars were calculated.

Physical Activity

The 16-item Global Physical Activity Questionnaire (GPAQ) assessed the time participants spent in various physical activity intensities in the past week [26]. The total number of minutes per week spent sedentary and in moderate-to-vigorous leisure time physical activity was calculated.

Mindfulness

The 24-item Five Facet Mindfulness Questionnaire Short Form assesses five components of mindfulness: observing, describing, acting with awareness, nonjudging, and nonreactivity [27]. Participants rated the extent to which a variety of statements were true for them (eg, "I watch my feelings without getting carried away by them," "I make judgments about whether my thoughts are good or bad") on a 5-point Likert scale from 1 (never or very rarely true) to 5 (very often or always true). Items were summed, with higher scores indicating higher levels of mindfulness.

Stress

Participants reported their perceived stress via the 10-item Perceived Stress Scale (PSS-10), a survey that measures the extent to which life situations over the past month are viewed as uncontrollable, unpredictable, and stressful (eg, "In the past month, how often have you felt that things were going your way?", "How often have you been upset because of something that happened unexpectedly?") [28]. Participants rated items on a scale from 0 (never) to 4 (very often), and items were summed, with higher scores indicating higher levels of stress.

T2D Knowledge

An adapted version of the Diabetes Knowledge Test developed by Garcia et al [29] for Mexican American adults residing in South Texas was used to measure participants' T2D knowledge. The adapted measure consisted of 26 items related to general T2D knowledge about diet, physical activity, and blood glucose monitoring (eg, "Shaking and sweating are signs of high blood sugar," "The way I prepare my food is as important as the foods I eat," "Regular exercise will increase the need for insulin or other diabetic medication"). Participants responded to items with "yes," "no," or "I don't know," and we calculated the percentage of items answered correctly.

T2D Management Self-Efficacy

The 15-item Diabetes Management Self-Efficacy Scale assessed participants' confidence in their ability to engage in diabetes management activities (eg, "Avoid drinks that contain sugar,"

XSL•FO
RenderX

"Exercise three times per week," "Test blood glucose every day") [30]. Items were rated on a scale from 1 (not at all confident) to 4 (completely confident), and an average score was calculated.

T2D Support Satisfaction

Participants were asked to rate how satisfied they were with T2D management support provided by four potential sources (ie, family support person, other family member, YMCA staff, and others) on a scale from 1 (very dissatisfied) to 6 (very satisfied) [31]. An average satisfaction score across all four categories was calculated.

Participant Engagement

Four index variables (family engagement and support, participation in self-management education, program support and facilitation, and participation in self-monitoring) were created to track participants' and family support persons' engagement levels and uptake of key program activities. These index variables collectively represented the physical, social, and cognitive dimensions of influences on participants' engagement in ¡Salud, Salud! [32]. *Family engagement and support* was determined by three activities: having a support person, completing at least one online lesson with a support person, and attending the orientation meeting with a support person. *Participation in self-management education* was determined by the number of core online health lessons (eight lessons) completed by the participants. *Program support and facilitation* was determined by the number of encounters participants had with the CHWs (11 coaching sessions) and the nutrition educator (1 meeting). *Participation in self-monitoring* was determined by the number of weekly surveys completed by the participants. The number of activities completed for each index variable was converted to percentages to standardize the dose and facilitate interpretation of the results.

Statistical Analysis

Means (SDs) for continuous variables and frequencies and percentages for categorical variables were calculated for all participant characteristics and outcome measures. Paired *t*-tests were used to determine whether continuous primary and secondary outcome variables changed significantly from baseline to the 12-week assessment. McNemar chi-square tests were used to examine changes in the proportion of participants reporting good-to-excellent health from baseline to the 12-week assessment. Cohen d and ϕ estimated the effect sizes for continuous and categorical variables, respectively. The level of significance for all statistical tests was set at .05. Typically, adjustments to *P* values are applied to reduce type I errors in multiple comparisons. However, some statisticians argue that such adjustments may not be necessary in exploratory research, as they can decrease statistical power and increase the risk of type II errors, particularly in studies with small sample sizes [33]. In this study, sample sizes varied for each test due to the listwise deletion of missing values. Data were analyzed using IBM's Statistical Package for the Social Sciences, version 28.0.

Ethical Considerations

The study protocol was reviewed and approved by the Institutional Review Board at the University of Texas at Austin (approval number STUDY00002124). Written informed consent was obtained from all participants in either English or Spanish, depending on each participant's preferred language. Participants were provided with a copy of the consent form, which included detailed information about the study, data collection, and their rights as participants. Data collection training for all study staff emphasized the importance of protecting participants' privacy and maintaining confidentiality of data. Data collection, management, and quality assurance for this study were facilitated using tablets programmed with a Health Insurance Portability and Accountability Act (HIPAA)–compliant platform (ie, Connected Health) for participants and CHWs to track study activities, as well as REDCap, a secure data management system that ensures effective data collection, storage, retrieval, and quality control. All data collected in this study were deidentified, ensuring no identification of individual participants in the data or study was possible. As compensation for their participation, each participant received a 6-month family membership to the YMCA.

Results

Participant Characteristics

We distributed 2743 flyers and interacted with 3298 individuals through community outreach efforts. A total of 1798 individuals initiated the eligibility screener; however, only 363 (20.2%) completed it. Of those who completed the screener, 155 (42.7%) met the study's eligibility criteria. We enrolled 81 (52.3%) participants with T2D or prediabetes (mean age 48.90 years, SD 12.57; n=57, 70.4%, female; n=66, 81.5%, Latino) who completed the baseline assessment and orientation session. The primary reasons why those eligible did not enroll were loss of interest or our inability to reach them after multiple attempts to complete baseline assessments. The analysis included only those participants with T2D or prediabetes who completed at least one of the primary and secondary outcome measures at the 12-week assessment (n=48, 59.3% retention).

Baseline participant characteristics are displayed in Table 2. Participants were predominantly Latino, female, and economically disadvantaged. Although 28 (58.3%) of 48 participants preferred to use the Spanish language in daily conversations, 33 (68.8%) were born outside the United States. In addition, 36 (75%) participants had physician-diagnosed T2D, and 34 (70.8%) had obesity (BMI≥30). At baseline, 30 (62.5%) participants had a designated support person when they enrolled, and 24 (88.9%) of 27 participants reported that they received diabetes management–related support in at least 1 of 23 areas from their designated support person, while 24 (60%) of 40 participants received support in at least 1 area from other family members.

We examined differences in participant characteristics and primary outcomes at baseline between those who completed and those who did not complete the 12-week assessment (results not shown). Those who completed the 12-week assessment were less likely to report having difficulty paying their bills (n=15, 31.3%) compared to noncompleters (n=18, 54.5%; *P*<.05). No other significant differences were found.

XSL•FO
RenderX

Table 2. Participant characteristics at baseline (N=48).

Characteristics	Participants, n (%)
Race/ethnicity	
Latino	35 (72.9)
White	6 (12.5)
Black	2 (4.2)
Asian	3 (6.3)
Other	2 (4.2)
Female gender	35 (72.9)
Age≥50 years	23 (47.9)
Prefer the Spanish language	28 (58.3)
Currently not working	27 (56.3)
Difficulty paying bills	15 (31.3)
No health insurance	15 (31.3)
Less than high school education	30 (62.5)
Foreign born	33 (68.8)
Physician-diagnosed T2D[a]	36 (75.0)
Obesity[b]	34 (72.3)
Participated with a support person	30 (62.5)
Support in at least one area from support person[c]	24 (88.9)
Support in at least one area from other family member[d]	24 (60.0)

[a]T2D: type 2 diabetes.
[b]N=47.
[c]N=27.
[d]N=40.

Participant Engagement in Program Activities

Table 3 shows the participant engagement in, and uptake of, key program activities across the four index variables. The highest level of engagement was in program support and facilitation, with participants completing 88% of ¡Salud, Salud! encounter sessions with CHWs and the nutrition counselor. Engagement levels in the other three indices were below 70%.

Table 3. Participant engagement in program activities.

Type of engagement	Mean (SD), range of completed activities	Mean proportion of activities completed (%)
Family engagement and support (number of support activities completed by the support person; maximum 3)	1.48 (1.22), 0-3	49
Participation in self-management training (number of online health lessons completed; maximum 8)	3.48 (3.14), 0-8	66
Program support and facilitation (number of encounters with study staff; maximum 13)	11.54 (2.10), 4-13	88
Participation in self-monitoring (number of weeks in which weekly self-monitoring surveys completed; maximum 12)	7.34 (3.75), 0-12	56

Changes in Primary Outcomes

There was improvement in the primary outcomes from baseline to the12-week assessment. Although the reduction in HbA1c (–0.30%, $P \leq .09$) was marginally significant (Table 4), the proportion of participants self-reporting good-to-excellent health significantly increased (two-sided $P \leq .003$) from 19 (39.6%) at baseline to 28 (58.3%) at the 12-week assessment, with Somers D being 0.381 (SE 0.13), T being 2.93, and effect size (ϕ) being 0.39. Effect sizes were small for both outcomes.

XSL•FO
RenderX

Table 4. Comparisons of changes in the study's primary and secondary outcomes from baseline to 12-week assessment.

Outcomes	Value, mean (SD)		Change (SE; 95% CI) from baseline to 12-week assessment	Paired *t*-test results		Effect size (Cohen d)
	Baseline	12 weeks		*t* (*df*)	*P* value	
HbA1c[a] (%)[b]	7.33 (1.78)	7.04 (1.58)	–0.30 (0.17; –0.64 to 0.05)	–1.71(40)	.09	–0.27
Body weight (kg)[b]	89.51 (20.17)	88.21 (19.74)	–1.30 (0.55; –2.41 to –0.18)	–2.36(40)	.02	–0.37
Body fat (%)[b]	41.20 (6.47)	39.93 (6.58)	–1.26 (0.55; –2.37 to –0.16)	–2.31(40)	.01	–0.36
Diabetes knowledge[c]	16.94 (4.39)	17.89 (3.60)	0.94 (0.56; –0.19 to 2.08)	1.69(35)	.10	0.29
Diabetes management self-efficacy[d]	2.84 (0.64)	3.05 (0.58)	0.21 (0.12; –0.03 to 0.45)	1.77(38)	.08	0.28
Mindfulness[e]	30.66(5.89)	32.87(6.05)	2.21 (0.88; 0.44 to 3.98)	2.51(46)	.01	0.37
Perceived stress[e]	1.28 (0.76)	1.00 (0.79)	0.28 (0.12; –0.51 to –0.04)	–2.35(46)	.02	–0.34
Unhealthy days[f]	7.26 (9.46)	4.87 (5.80)	2.39(1.73; –1.11 to 5.90)	1.39(37)	.17	–0.23
Satisfaction with support[g]	4.70 (1.62)	4.99 (1.32)	–0.30 (0.34; –0.99 to 0.40)	–0.88(34)	.39	0.15
Sedentary time (minutes/week)[b]	287.44 (248.00)	217.17 (213.00)	–70.27 (23.87; –118.52 to –22.02)	–2.94 (40)	.003	–0.46
Leisure time moderate-to-vigorous activity (minutes/week)[h]	55.86 (95.79)	60.86 (91.63)	5.00 (22.64; –41.38 to 51.38)	0.22(28)	.82	0.04
Fiber intake (g/day)[i]	15.97 (2.61)	15.29 (2.44)	–0.68 (0.37; –1.41 to 0.06)	–1.85 (45)	.07	–0.27
Added sugar intake (teaspoons/day)[j]	13.76 (3.34)	11.61 (1.06)	–2.15 (0.46; –3.08 to –1.23)	–4.70(42)	<.001	–0.71
Vegetable intake (cups/day)[i]	1.32 (0.37)	1.25 (0.25)	–0.07 (0.06; –0.18 to 0.04)	–1.28(45)	.21	–0.19

[a]HbA1c: hemoglobin A1c.
[b]N=41.
[c]N=36.
[d]N=39.
[e]N=47.
[f]N=38.
[g]N=35.

Changes in Secondary Outcomes

Table 4 also shows changes in secondary outcomes from baseline to 12 weeks. Significant reductions in body weight ($P \leq .02$) and body fat percentage ($P \leq .01$) were observed. Increases in Diabetes Knowledge Test scores ($P=.10$) and Diabetes Self-Management Self-Efficacy Scale scores ($P=.08$) were marginally significant. PSS-10 scores significantly reduced ($P=.02$), and mindfulness scores significantly improved ($P=.001$). Although there were improvements in the number of physically and mentally unhealthy days during the past month (–2.39, SE 10.65) and in satisfaction scores for support across all support person categories (0.30, SE 2.00), these improvements were not statistically significant. Additionally, although the time spent in sedentary activities significantly reduced (–70 minutes/day, $P=.003$), no change in leisure time physical activity was observed. Finally, the added sugar intake significantly reduced ($P \leq .001$), and fiber intake marginally increased ($P \leq .07$). No change was reported in vegetable intake. Effect sizes were small for all secondary outcomes, except for a medium effect size for added sugar intake.

Discussion

Principal Findings

The multilevel and multidomain ¡Salud, Salud! program demonstrated promising but modest effects in supporting T2D self-management practices among predominantly low-income Latino adults with T2D or prediabetes. Our findings partially supported the study hypothesis, showing a marginally significant reduction in HbA1c (–0.30%, $P \leq .09$) and significant improvement in self-reported general health ($P \leq .003$) over the 12-week program. Importantly, participants demonstrated significant improvements in secondary outcomes, including body weight, psychological well-being, sedentary behavior, and added sugar intake, although the effect sizes were generally small.

To increase the uptake of T2D self-management practices, ¡Salud, Salud! was designed following national standards for DSMES that incorporated evidence-based strategies for individuals experiencing health disparities [34]. These strategies included focusing on lifestyle behavior changes, removing

XSL•FO
RenderX

barriers to physical activity, using telemedicine, providing support and facilitation by lay community educators, setting patient-centered goals, and implementing stress management [22]. Additionally, digital technologies were used to increase the program's accessibility and availability with timely and personalized coaching by trained CHWs [35]. Nonetheless, modest effects in glycemic control and weight loss among ¡Salud, Salud! participants are consistent with the literature on translation studies of evidence-based T2D management and prevention programs, such as Look AHEAD and the Diabetes Prevention Program [36,37]. The current literature demonstrates that participation in DSMES among people with T2D or prediabetes results in small-to-moderate reductions in HbA1c and body weight in in-person programs [38,39], while digitally delivered programs tend to produce small effects [13,40]. Furthermore, reductions in HbA1c and weight are smaller in studies with low-income minority participants compared to White participants [39,41]. For example, a 2020 systematic review reported a reduction of –0.24% in HbA1c among Latino participants in DSMES programs [42]. Additionally, moderate levels of program engagement in three of the four engagement index variables might have contributed to the smaller reduction in HbA1c among ¡Salud, Salud! participants. Previous research demonstrates that Latinos in low-to-medium-intensity DSMES interventions had similar levels of reduction in HbA1c [13]. Despite ongoing efforts at cultural tailoring and adaptation [43], the inequity in program effectiveness among socioeconomically disadvantaged minority populations is alarming [44] and requires innovations that address the social determinants of diabetes care to address upstream factors, such as living in deprived neighborhoods, competing priorities and time constraints, and access to affordable healthy foods—all of which influence T2D self-management at the individual level [45,46]. A 2024 systematic review of 106 randomized controlled trials of DSMES interventions among socially marginalized racial and ethnic minority populations in the United States noted that relatively few interventions target policy, community resources, or the physical environment, but those that do show more promise than DSMES alone [15].

Our findings regarding participants' quality of life and psychological well-being are noteworthy. The significant improvement in self-reported general health and psychological measures addresses a critical need, as individuals with T2D typically report a lower quality of life due to the burden associated with disease management [47]. These improvements may be especially meaningful for our socioeconomically disadvantaged participants, who often face additional stressors that can complicate T2D management [48,49].

Although participation in DSMES has contributed to improvements in T2D self-management knowledge, self-efficacy, and practices [17,38], the integration of family support in DSMES can have a greater impact on T2D care, with positive effects on diabetes-related support, self-efficacy, psychological well-being, and behaviors (eg, physical activity, diet) [18]. Although ¡Salud, Salud! was designed to be family centered, the program's effect on diabetes care–related social and psychological outcomes was mixed. We observed marginally significant improvements in DSM knowledge and self-efficacy, significant improvements in fiber intake and reduced time spent sedentary, but no improvement in physical activity or vegetable intake. Additionally, ¡Salud, Salud! participants reported marginally significant improvements in satisfaction with support in T2D care. Notably, for participants who reported receiving T2D care support from their support person or another family member at baseline, satisfaction levels were already high before enrolling in ¡Salud, Salud!, potentially contributing to the marginal effect of the intervention on satisfaction with support [50]. However, the level of family engagement and support in ¡Salud, Salud! (ie, activities completed by the support person) was less than 50%, suggesting a lack of instrumental family support related to T2D management. Therefore, it is critical to investigate how to enhance family informational and instrumental support in socioeconomically disadvantaged populations with T2D [9,18].

Participants were most engaged with the CHW encounters, indicating that our participants relied on human interaction more than on the self-guided mHealth components. In future studies, it might be advisable to increase digital literacy and navigation support through additional staff and CHW training and dedicated time when working with this population [32]. Although our CHWs were available to assist with the technology, their time with each participant was limited and more focused on goal setting and lifestyle modification support.

Strengths and Limitations

We attribute the promising impacts of ¡Salud, Salud! to the adoption of evidence-based strategies for DSMES [22,51], including the National Standards for Culturally and Linguistically Appropriate Services in Health and Health Care [52], to meet the needs of our study participants. Other strengths of the study included a theory-driven intervention model, community partnership with the YMCA that provided venues to reach the community, CHW facilitation and support that had the highest level of participant engagement, and program delivery combining digital technologies and personalized live coaching that are desired by socioeconomically disadvantaged populations [53].

Several important limitations should be considered in the interpretation of the study results. First, the short duration of the study (ie, 12 weeks) likely limited ¡Salud, Salud!'s impact on HbA1c and body weight, as DSMES interventions typically last 6-12 months and are often supplemented with maintenance programs [54,55]. Second, although the study sample was powered to test the study hypotheses, low enrollment at baseline and high participant dropout led to a smaller sample at posttest. Low participation [12] and high dropout rates [36,55] in clinical trials testing DSMES programs have been reported among socioeconomically disadvantaged participants. However, this limitation raises concerns about the reliability and generalizability of the findings, even though there were no differences between the completers and noncompleters in the study. Third, we were not able to evaluate the effect of family engagement on the study outcomes, given the low level of family participation and small sample size. Finally, ¡Salud, Salud! was not designed to address the social and financial needs that can influence the uptake of T2D care–related practices, such as

XSL•FO
RenderX

access to affordable food resources and lack of transportation to exercise at the YMCA [32,56].

Conclusion

The ¡Salud, Salud! program demonstrated promising but modest effects in improving T2D-related outcomes among predominantly low-income Latinos with T2D. Findings from this study reveal that cultural tailoring of DSMES is necessary but insufficient to optimize the uptake of evidence-based strategies and enhance participant engagement in T2D self-management practices. Future studies should broaden the search for effective strategies to reduce the barriers influencing the uptake of evidence-based DSM practices in low-resource populations.

Acknowledgments

This study was supported by a grant from the US Department of Health and Human Services Office of Minority Health (CPI2021003606). Artificial intelligence was not used in the preparation of this manuscript.

Data Availability

The datasets used and analyzed during the study are available from the corresponding author upon request.

Conflicts of Interest

None declared.

References

1. International Diabetes Federation. IDF Diabetes Atlas, Seventh Edition. Brussels, Belgium: International Diabetes Federation; 2015.
2. Liu J, Yi SS, Russo R, Mayer VL, Wen M, Li Y. Trends and disparities in diabetes and prediabetes among adults in the United States, 1999-2018. Public Health 2023 Jan;214:163-170 [FREE Full text] [doi: 10.1016/j.puhe.2022.10.021] [Medline: 36586345]
3. Meunier S, Coulombe S, Beaulieu M, Côté J, Lespérance F, Chiasson J, et al. Longitudinal testing of the information-motivation-behavioral skills model of self-care among adults with type 2 diabetes. Patient Educ Couns 2016 Nov;99(11):1830-1836 [FREE Full text] [doi: 10.1016/j.pec.2016.06.011] [Medline: 27373961]
4. Nelson LA, Wallston KA, Kripalani S, LeStourgeon LM, Williamson SE, Mayberry LS. Assessing barriers to diabetes medication adherence using the Information-Motivation-Behavioral skills model. Diabetes Res Clin Pract 2018 Aug;142:374-384 [FREE Full text] [doi: 10.1016/j.diabres.2018.05.046] [Medline: 29879495]
5. Health and safety code. Title 2. Health. Subtitle B. Health programs. Chapter 48. Promotoras and community health workers. Texas Department of Health. 2015. URL: https://statutes.capitol.texas.gov/Docs/HS/htm/HS.48.htm [accessed 2025-08-25]
6. Diabetes care and education specialist (DCES) certificate program. Association of Diabetes Care & Education Specialists. URL: https://www.adces.org/store/certificate-programs/detail/diabetes-care-and-education-specialist-certificate-program-(cpdces001) [accessed 2021-06-24]
7. Walicka M, Jozwiak J, Rzeszotarski J, Zonenberg A, Masierek M, Bijos P, et al. Diagnostic Accuracy of Glycated Haemoglobin and Average Glucose Values in Type 2 Diabetes Mellitus Treated with Premixed Insulin. Diabetes Ther 2019 Apr;10(2):587-596 [FREE Full text] [doi: 10.1007/s13300-019-0570-1] [Medline: 30734901]
8. Moriarty DG, Zack MM, Kobau R. The Centers for Disease Control and Prevention's Healthy Days Measures - population tracking of perceived physical and mental health over time. Health Qual Life Outcomes 2003 Sep 02;1:37 [FREE Full text] [doi: 10.1186/1477-7525-1-37] [Medline: 14498988]
9. Pamungkas RA, Chamroonsawasdi K, Vatanasomboon P. A Systematic Review: Family Support Integrated with Diabetes Self-Management among Uncontrolled Type II Diabetes Mellitus Patients. Behav Sci (Basel) 2017 Sep 15;7(3):1-2 [FREE Full text] [doi: 10.3390/bs7030062] [Medline: 28914815]
10. Schmitt A, Gahr A, Hermanns N, Kulzer B, Huber J, Haak T. The Diabetes Self-Management Questionnaire (DSMQ): development and evaluation of an instrument to assess diabetes self-care activities associated with glycaemic control. Health Qual Life Outcomes 2013 Aug 13;11:138 [FREE Full text] [doi: 10.1186/1477-7525-11-138] [Medline: 23937988]
11. Pillay J, Armstrong MJ, Butalia S, Donovan LE, Sigal RJ, Vandermeer B, et al. Behavioral Programs for Type 2 Diabetes Mellitus: A Systematic Review and Network Meta-analysis. Ann Intern Med 2015 Dec 01;163(11):848-860 [FREE Full text] [doi: 10.7326/M15-1400] [Medline: 26414227]
12. Tharakan A, McPeek Hinz E, Zhu E, Denmeade B, German J, Huang WA, et al. Accessibility of diabetes education in the United States: barriers, policy implications, and the road ahead. Health Aff Sch 2024 Aug;2(8):qxae097. [doi: 10.1093/haschl/qxae097] [Medline: 39206435]
13. Kerr D, Ahn D, Waki K, Wang J, Breznen B, Klonoff DC. Digital Interventions for Self-Management of Type 2 Diabetes Mellitus: Systematic Literature Review and Meta-Analysis. J Med Internet Res 2024 Jul 22;26:e55757 [FREE Full text] [doi: 10.2196/55757] [Medline: 39037772]

XSL•FO
RenderX

14. Zeh P, Sandhu HK, Cannaby AM, Sturt JA. Cultural barriers impeding ethnic minority groups from accessing effective diabetes care services: a systematic review of observational studies. Divers Equal Health Care 2014;11(1):9-33 [FREE Full text] [doi: 10.21767/2049-5471.100001]
15. Falk EM, Staab EM, Deckard AN, Uranga SI, Thomas NC, Wan W, et al. Effectiveness of Multilevel and Multidomain Interventions to Improve Glycemic Control in U.S. Racial and Ethnic Minority Populations: A Systematic Review and Meta-analysis. Diabetes Care 2024 Sep 01;47(9):1704-1712. [doi: 10.2337/dc24-0375] [Medline: 39190927]
16. Marquez I, Calman N, Crump C. A Framework for Addressing Diabetes-Related Disparities in US Latino Populations. J Community Health 2019 Apr;44(2):412-422. [doi: 10.1007/s10900-018-0574-1] [Medline: 30264184]
17. Ernawati U, Wihastuti TA, Utami YW. Effectiveness of diabetes self-management education (DSME) in type 2 diabetes mellitus (T2DM) patients: Systematic literature review. J Public Health Res 2021 Apr 14;10(2):1 [FREE Full text] [doi: 10.4081/jphr.2021.2240] [Medline: 33855427]
18. Busebaia TJA, Thompson J, Fairbrother H, Ali P. The role of family in supporting adherence to diabetes self-care management practices: An umbrella review. J Adv Nurs 2023 Oct;79(10):3652-3677. [doi: 10.1111/jan.15689] [Medline: 37150951]
19. Diabetes self-management education and support (DSMES) toolkit. Centers for Disease Control and Prevention. 2018. URL: https://www.cdc.gov/diabetes/dsmes-toolkit/index.html [accessed 2021-06-04]
20. Errisuriz VL, Parra-Medina D, Cuevas H, Flores BE, Delfausse L, Galvan C, et al. A Community-Academic Partnership Using Digital Solutions for Diabetes Self-Management. J Health Care Poor Underserved 2025;36(3S):77-89 [FREE Full text] [doi: 10.1353/hpu.2025.a967361] [Medline: 40820747]
21. Ockene IS, Tellez TL, Rosal MC, Reed GW, Mordes J, Merriam PA, et al. Outcomes of a Latino community-based intervention for the prevention of diabetes: the Lawrence Latino Diabetes Prevention Project. Am J Public Health 2012 Feb;102(2):336-342 [FREE Full text] [doi: 10.2105/AJPH.2011.300357] [Medline: 22390448]
22. American Diabetes Association. Standards of medical care in diabetes—2021. Diabetes Care 2021 Jan;44(Suppl 1):S1-S2 [FREE Full text] [doi: 10.2337/dc21-Sint] [Medline: 33298409]
23. Wang J, Zhang J. Adapting A Unified Electronic Health Record Usability Framework for Evaluation of Connected Health Care Technologies Linking Mobile Data. iProc 2015 Oct 27;1(1):e20 [FREE Full text] [doi: 10.2196/iproc.4477]
24. Insight timer. 2021. URL: https://insighttimer.com/ [accessed 2025-08-16]
25. National Cancer Institute Division of Cancer Control and Population Sciences. URL: https://epi.grants.cancer.gov/nhanes/dietscreen/scoring [accessed 2025-08-25]
26. Bull FC, Maslin TS, Armstrong T. Global physical activity questionnaire (GPAQ): nine country reliability and validity study. J Phys Act Health 2009 Nov;6(6):790-804. [doi: 10.1123/jpah.6.6.790] [Medline: 20101923]
27. Bohlmeijer E, ten Klooster PM, Fledderus M, Veehof M, Baer R. Psychometric properties of the five facet mindfulness questionnaire in depressed adults and development of a short form. Assessment 2011 Sep;18(3):308-320 [FREE Full text] [doi: 10.1177/1073191111408231] [Medline: 21586480]
28. Cohen S. Perceived stress in a probability sample of the United States. In: The Social Psychology of Health. Thousand Oaks, CA: SAGE Publications; 1988:31-67.
29. Garcia AA, Villagomez ET, Brown SA, Kouzekanani K, Hanis CL. The Starr County Diabetes Education Study: development of the Spanish-language diabetes knowledge questionnaire. Diabetes Care 2001 Jan;24(1):16-21 [FREE Full text] [doi: 10.2337/diacare.24.1.16] [Medline: 11194219]
30. Wang ML, Lemon SC, Welch G, Rosal MC. Development and validation of the Lifestyle Self-Efficacy Scale for Latinos with Diabetes (LSESLD). Ethn Dis 2013;23(4):428-435 [FREE Full text] [Medline: 24392604]
31. de la-Roche R, López-Serrano A, Barragán-Solís A, Arce-Arrieta E. User satisfaction at a Social Security Institute family medical center in Mexico City. Archivos en Medicina Familiar 2005;7(1):22-26 [FREE Full text]
32. Turnbull S, Lucas PJ, Hay AD, Cabral C. The role of economic, educational and social resources in supporting the use of digital health technologies by people with T2D: a qualitative study. BMC Public Health 2021 Feb 05;21(1):293 [FREE Full text] [doi: 10.1186/s12889-021-10325-7] [Medline: 33546661]
33. Althouse AD. Adjust for Multiple Comparisons? It's Not That Simple. Ann Thorac Surg 2016 May;101(5):1644-1645 [FREE Full text] [doi: 10.1016/j.athoracsur.2015.11.024] [Medline: 27106412]
34. American Diabetes Association. Abridged for Primary Care Providers. Clin Diabetes 2022 Jan;40(1):10-38 [FREE Full text] [doi: 10.2337/cd22-as01] [Medline: 35221470]
35. Fleming GA, Petrie JR, Bergenstal RM, Holl RW, Peters AL, Heinemann L. Diabetes digital app technology: benefits, challenges, and recommendations. A consensus report by the European Association for the Study of Diabetes (EASD) and the American Diabetes Association (ADA) Diabetes Technology Working Group. Diabetologia 2020 Feb;63(2):229-241 [FREE Full text] [doi: 10.1007/s00125-019-05034-1] [Medline: 31802144]
36. Ng BP, Ely E, Papali'i M, Cannon MJ. Delivering the National Diabetes Prevention Program: Assessment of Retention, Physical Activity, and Weight Loss Outcomes by Participant Characteristics and Delivery Modes. J Diabetes Res 2024;2024:8461704 [FREE Full text] [doi: 10.1155/2024/8461704] [Medline: 39165352]
37. Liss DT, Finch EA, Cooper A, Sheth A, Tejuosho AD, Lancki N, et al. One-year effects of a group-based lifestyle intervention in adults with type 2 diabetes: A randomized encouragement trial. Diabetes Res Clin Pract 2018 Jun;140:36-44 [FREE Full text] [doi: 10.1016/j.diabres.2018.03.030] [Medline: 29596950]

XSL•FO
RenderX

38. Odgers-Jewell K, Ball LE, Kelly JT, Isenring EA, Reidlinger DP, Thomas R. Effectiveness of group-based self-management education for individuals with Type 2 diabetes: a systematic review with meta-analyses and meta-regression. Diabet Med 2017 Aug;34(8):1027-1039 [FREE Full text] [doi: 10.1111/dme.13340] [Medline: 28226200]
39. Yang J, Xia Y, Sun Y, Guo Y, Shi Z, Cristina do Vale Moreira N, et al. Effect of lifestyle intervention on HbA1c levels in overweight and obese adults with type 2 diabetes across ethnicities: A systematic review and meta-analysis of randomized controlled trials. Diabetes Res Clin Pract 2023 May;199:110662 [FREE Full text] [doi: 10.1016/j.diabres.2023.110662] [Medline: 37028602]
40. Villegas V, Shah A, Manson JE, Tobias DK. Prevention of type 2 diabetes through remotely administered lifestyle programs: A systematic review. Contemp Clin Trials 2022 Aug;119:106817 [FREE Full text] [doi: 10.1016/j.cct.2022.106817] [Medline: 35691488]
41. Okeyo HM, Biddle M, Williams LB. Impact of Diabetes Self-Management Education on A1C Levels Among Black/African Americans: A Systematic Review. Sci Diabetes Self Manag Care 2024 Feb;50(1):87-95 [FREE Full text] [doi: 10.1177/26350106231213400] [Medline: 37994435]
42. Hildebrand JA, Billimek J, Lee J, Sorkin DH, Olshansky EF, Clancy SL, et al. Effect of diabetes self-management education on glycemic control in Latino adults with type 2 diabetes: A systematic review and meta-analysis. Patient Educ Couns 2020 Feb;103(2):266-275 [FREE Full text] [doi: 10.1016/j.pec.2019.09.009] [Medline: 31542186]
43. McCurley JL, Gutierrez AP, Gallo LC. Diabetes Prevention in U.S. Hispanic Adults: A Systematic Review of Culturally Tailored Interventions. Am J Prev Med 2017 Apr;52(4):519-529 [FREE Full text] [doi: 10.1016/j.amepre.2016.10.028] [Medline: 27989451]
44. Walker RJ, Strom Williams J, Egede LE. Influence of Race, Ethnicity and Social Determinants of Health on Diabetes Outcomes. Am J Med Sci 2016 Apr;351(4):366-373 [FREE Full text] [doi: 10.1016/j.amjms.2016.01.008] [Medline: 27079342]
45. Whittemore R. Culturally competent interventions for Hispanic adults with type 2 diabetes: a systematic review. J Transcult Nurs 2007 Apr;18(2):157-166 [FREE Full text] [doi: 10.1177/1043659606298615] [Medline: 17416718]
46. Fortmann AL, Savin KL, Clark TL, Philis-Tsimikas A, Gallo LC. Innovative Diabetes Interventions in the U.S. Hispanic Population. Diabetes Spectr 2019 Nov;32(4):295-301 [FREE Full text] [doi: 10.2337/ds19-0006] [Medline: 31798285]
47. Rubin RR, Peyrot M. Quality of life and diabetes. Diabetes Metab Res Rev 1999;15(3):205-218 [FREE Full text] [doi: 10.1002/(sici)1520-7560(199905/06)15:3<205::aid-dmrr29>3.0.co;2-o] [Medline: 10441043]
48. Anderson D, Christison-Lagay J. Diabetes self-management in a community health center: improving health behaviors and clinical outcomes for underserved patients. Clin Diabetes 2008;26(1):22-27 [FREE Full text] [doi: 10.2337/diaclin.26.1.22]
49. Verdecias N, McQueen A, Von Nordheim DA, Broussard DJ, Smith RE, Kreuter MW. Diabetes distress in a Medicaid sample: The role of psychosocial and health-related factors. J Diabetes Complications 2023 Jul;37(7):108495 [FREE Full text] [doi: 10.1016/j.jdiacomp.2023.108495] [Medline: 37156052]
50. Carpenter R, DiChiacchio T, Barker K. Interventions for self-management of type 2 diabetes: An integrative review. Int J Nurs Sci 2019 Jan 10;6(1):70-91 [FREE Full text] [doi: 10.1016/j.ijnss.2018.12.002] [Medline: 31406872]
51. Davis J, Fischl AH, Beck J, Browning L, Carter A, Condon JE, et al. 2022 National Standards for Diabetes Self-Management Education and Support. Sci Diabetes Self Manag Care 2022 Feb;48(1):44-59 [FREE Full text] [doi: 10.1177/26350106211072203] [Medline: 35049403]
52. National standards for Culturally and Linguistically Appropriate Services (CLAS) in health and health care: a blueprint for advancing and sustaining CLAS policy and practice. Office of Minority Health. 2013. URL: https://thinkculturalhealth.hhs.gov/assets/pdfs/EnhancedCLASStandardsBlueprint.pdf [accessed 2021-01-07]
53. Hamidi S, Gholamnezhad Z, Kasraie N, Sahebkar A. The Effects of Self-Efficacy and Physical Activity Improving Methods on the Quality of Life in Patients with Diabetes: A Systematic Review. J Diabetes Res 2022;2022:2884933 [FREE Full text] [doi: 10.1155/2022/2884933] [Medline: 35936392]
54. Chrvala CA, Sherr D, Lipman RD. Diabetes self-management education for adults with type 2 diabetes mellitus: A systematic review of the effect on glycemic control. Patient Educ Couns 2016 Jun;99(6):926-943 [FREE Full text] [doi: 10.1016/j.pec.2015.11.003] [Medline: 26658704]
55. Ely EK, Ng BP, Cannon MJ. Delivering the National Diabetes Prevention Program: Assessment of Outcomes in In-Person and Virtual Organizations. J Diabetes Res 2023;2023:8894593 [FREE Full text] [doi: 10.1155/2023/8894593] [Medline: 37928892]
56. Verdecias N. Feasibility of integrating social needs interventions in community-based diabetes programs. Diabetes 2024:678-P [FREE Full text] [doi: 10.2337/db24-678-P]

Abbreviations

CHW: community health worker
DSM: diabetes self-management
DSMES: diabetes self-management education and support
HbA1c: hemoglobin A1c

XSL•FO
RenderX

HRQoL: health-related quality of life
IMB: information-motivation-behavioral skills
mHealth: mobile health
PSS-10: 10-item Perceived Stress Scale
T2D: type 2 diabetes
YMCA: Young Men's Christian Association

Edited by A Mavragani; submitted 26.06.25; peer-reviewed by E Staab; comments to author 18.07.25; revised version received 06.08.25; accepted 11.08.25; published 06.10.25.

Please cite as:
Yin Z, Errisuriz VL, Cuevas H, Flores BE, Delfausse L, Galvan C, Wang J, Li C, Morfin R, Li S, Sapargeldiyeva M, Muyna GY, Zhang M, Sweet V, Parra-Medina D
Assessing a Community Health Worker-Facilitated, Digitally Delivered, Family-Centered Diabetes Management Program: Single-Arm Quasi-Experimental Study
JMIR Form Res 2025;9:e79032
URL: https://formative.jmir.org/2025/1/e79032
doi:10.2196/79032
PMID:

©Zenong Yin, Vanessa L Errisuriz, Heather Cuevas, Bertha E Flores, Laura Delfausse, Christina Galvan, Jing Wang, Chengdong Li, Renata Morfin, Shiyu Li, Maysa Sapargeldiyeva, Giliane Yza Muyna, Minyu Zhang, Vanessa Sweet, Deborah Parra-Medina. Originally published in JMIR Formative Research (https://formative.jmir.org), 06.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Collaborative Development of a Self-Tracking Assisted Psychotherapy Treatment Concept for Refugees With Complex Posttraumatic Stress Disorder: Participatory Action Research

Lisa Groenberg Riisager[1,2], PhD; Jakob Eg Larsen[3], PhD; Lotte Huniche[1], PhD; Thomas Blomseth Christiansen[4], MSc; Stine Bjerrum Moeller[1,2], PhD

[1]Department of Psychology, University of Southern Denmark, Odense M, Denmark
[2]Department of Multidisciplinary Trauma Treatment, Mental Health Services, Region of Southern Denmark, Middelfart, Denmark
[3]Department of Applied Mathemathics and Computer Science, Technical University of Denmark, Kongens Lyngby, Denmark
[4]Konsulent Blomseth, Hjoerring, Denmark

Corresponding Author:
Lisa Groenberg Riisager, PhD
Department of Psychology
University of Southern Denmark
Campusvej 55
Odense M, 5230
Denmark
Phone: 45 65501000
Email: lisagr@health.sdu.dk

Abstract

Background: Refugees are at high risk of severe mental health challenges due to exposure to war, torture, genocide, and childhood abuse. These experiences may lead to complex posttraumatic stress disorder (CPTSD), a condition that traditional treatments such as cognitive behavioral therapy and eye movement desensitization and reprocessing often struggle to treat adequately. Cultural complexity, limited relevance of standard interventions, and low adherence to therapeutic homework pose additional challenges. Self-tracking technologies offer a promising path for personalized mental health support in patients' everyday lives, but their integration into psychotherapy remains underexplored.

Objective: This study aimed to collaboratively develop a psychotherapeutic treatment concept for refugees with CPTSD by integrating a personalized, wearable self-tracking instrument, the One Button Tracker (OBT), into psychotherapy. The OBT allows patients to track subjective experiences in the moment they occur, offering a way to bridge therapy sessions and everyday life.

Methods: This study was conducted at a Danish trauma clinic specializing in treatment for refugees and veterans with posttraumatic stress disorder and CPTSD. A Participatory Action Research approach situated within the qualitative paradigm guided the process from November 2022 to April 2024. The codevelopment of the treatment concept involved therapists, patients, clinical psychology researchers, and human-computer interaction researchers (n=21). Qualitative data were gathered through patient interviews, therapist logbooks, and peer supervision sessions, and supplemented by self-tracking data from the OBT.

Results: Across 17 months, the team conducted 40 peer supervision sessions, 2 collaborative workshops, and 25 interviews with 9 patients who participated in therapy for 8 to 24 sessions. Self-tracking durations ranged from 22 to 366 days, covering 1 to 14 target phenomena per patient. The OBT was found to enhance patient engagement by supporting active symptom monitoring and reinforcing therapeutic interventions outside sessions. Therapists reported that the self-tracking data provided valuable insights into patients' lived experiences, supporting more personalized and context-sensitive interventions. The flexible use of the OBT also allowed patients to shift their focus from distressing symptoms to alternative coping strategies. Furthermore, the integration of self-tracking data strengthened the therapeutic alliance by improving communication and collaboration between patients and therapists. Some technical limitations affected data collection but did not substantially hinder the therapeutic process.

Conclusions: This is the first study to use a Participatory Action Research approach to codevelop a psychotherapeutic treatment concept integrating self-tracking technology for refugees with CPTSD. Findings indicate that the OBT may improve patient engagement, increase adherence to therapeutic tasks, and strengthen the therapeutic alliance. The treatment concept shows promise

XSL•FO
RenderX

as a transtheoretical and transdiagnostic approach, offering a flexible and personalized model for psychotherapy. Future research should refine the concept and examine its applicability across broader clinical contexts and populations.

(JMIR Form Res 2025;9:e66663) doi:10.2196/66663

KEYWORDS

wearables; mHealth; mobile health; mobile app; digital mental health; self-tracking; One Button Tracker; psychotherapy treatment; mental health; digital technology; digital intervention; mobile phone

Introduction

Background

As of 2023, the global refugee population has reached 30.5 million [1], with many refugees facing severe mental health challenges due to traumas such as war, torture, and childhood abuse [2,3]. These forms of repeated interpersonal trauma can precipitate the development of complex posttraumatic stress disorder (CPTSD) [4,5], a diagnosis recently included in the ICD-11 (International Statistical Classification of Diseases, Eleventh Revision). CPTSD extends the diagnostic framework of posttraumatic stress disorder (PTSD) by including disturbances in affect regulation, self-concept, and interpersonal relationships [3,4,6].

Although established psychotherapeutic approaches such as cognitive behavioral therapy (CBT), eye movement desensitization and reprocessing, and exposure therapy have shown effectiveness in treating CPTSD [7], their implementation is often inadequate in refugee populations. Much of the existing literature on psychotherapy for refugees focuses on PTSD [5], with findings often generalized to CPTSD without accounting for its distinct symptom profile and treatment needs. In particular, trauma-focused approaches may overwhelm patients when introduced prematurely, especially in the context of ongoing stressors, unstable living conditions, and complex grief [4]. Such mismatches often weaken therapeutic engagement [8,9], pointing to the need for treatment approaches that are specifically tailored to the lived realities of refugees with CPTSD.

Mental health professionals working with refugee populations also face specific challenges in addressing trauma. Karageorge et al [10] report that both refugee clients and staff express uncertainty about the value of focusing on traumatic events, with some providers experiencing secondary trauma themselves. Similarly, Peñuela-O'Brien et al [11] found that cultural and language barriers, along with strong emotional reactions from health care professionals, complicate trauma care, necessitating a flexible and person-centered approach.

While specialized treatments such as narrative exposure therapy have yielded promising outcomes [5,12], conventional psychotherapies often remain session-bound and do not fully engage with the everyday realities of refugee patients. Postmigration and socioeconomic stressors, as well as cross-cultural challenges, compound these challenges and can undermine patients' capacity to engage with therapy in daily life [13]. With between 16% and 38% of refugees meeting the criteria for CPTSD [4], there is a need for treatment models that are closely attuned to the complexity of patients' everyday contexts and the cultural dimensions of their trauma.

Personalized Approach to Self-Tracking

Digital mental health interventions increasingly aim to extend therapeutic support into patients' everyday lives. Among these, self-tracking technologies offer a means of collecting real-time data on thoughts, emotions, and behaviors, thereby enabling interventions that are more closely attuned to individual patterns and contexts [14]. While digital self-tracking draws on the long-standing clinical practice of self-monitoring, where patients record relevant experiences as part of therapeutic homework [15,16], its digital form transforms this activity into a continuous and integrated part of daily life [17].

Unlike passive sensing, which relies on device sensors to collect behavioral data without user input, active self-tracking requires patients to deliberately register their experiences [18,19]. This active engagement can increase awareness and agency, as patients learn to recognize symptom triggers and evaluate the effects of coping strategies between sessions [20,21]. Such data can serve as a concrete reference point in therapy sessions, offering therapists valuable insight into the patient's daily experiences that would otherwise remain inaccessible [16,22].

For therapists, the integration of self-tracking data can support more targeted, personalized interventions. Instead of relying solely on retrospective self-report or standardized questionnaires, therapists can engage with data that reflect lived experiences as they unfold [19]. Shared interpretation of these data in therapy sessions may also enhance collaboration and therapeutic alliance [18,23].

Despite its potential, research on self-tracking in psychotherapy remains limited. Existing studies have largely focused on smartphone apps or web-based interventions [24], with minimal exploration of how self-tracking might support face-to-face or blended care models. This gap is particularly striking concerning refugee populations affected by trauma. To date, no studies beyond our case study [20] and feasibility pilot study [17] have examined how self-tracking might be adapted to support psychotherapeutic treatment for refugees with CPTSD. Given the complex symptomatology and disrupted daily functioning often seen in this population, exploring self-tracking as a therapeutic tool in this context is both timely and necessary.

The One Button Tracker and Self-Tracking

To explore the integration of self-tracking in psychotherapy, this study used a research prototype of the One Button Tracker (OBT), a novel wearable self-tracking instrument that enables in-the-moment registration of a self-chosen subjectively experienced phenomenon relevant to the patient's mental health

XSL•FO
RenderX

challenges [17,20,25]. The OBT is designed with a single-button interface that enables patients to record subjective experiences with minimal effort or disruption to daily life [17,20,25-27]. This design ensures that self-tracking becomes a seamless part of the patient's daily routine, reinforcing engagement with the therapeutic process.

A central aspect of the OBT's application is its personalization. Therapists and patients collaboratively decide which specific phenomenon to track between sessions, tailoring the self-tracking practice to each patient's lived experience and current therapeutic focus [17,20]. This user-defined input differentiates the OBT from conventional symptom checklists or mobile apps that often offer fixed categories. As a form of real-time, patient-driven data collection, the OBT functions both as therapeutic homework and as an instrument for bridging experiences between therapy sessions [17]. It enables patients to become more aware of thoughts, emotional responses, and behavioral patterns as they unfold, while providing therapists with contextualized data that can inform and personalize therapeutic interventions [28].

The OBT is positioned within the methodological framework of ecological momentary assessment [29,30], allowing for in-the-moment data capture of naturally occurring phenomena. These data are visualized through a purpose-built interface and used during sessions to support reflection, pattern recognition, and collaborative decision-making between therapist and patient [17,20].

Participatory Action Research

Given the novelty of integrating wearable self-tracking technology into psychotherapy, especially within a culturally complex and clinically vulnerable population, a preliminary feasibility pilot study was conducted with approval from the Regional Committees on Health Research Ethics for Southern Denmark [17]. The study involved 2 patients and 2 therapists and focused primarily on the usability and integration of the OBT and the associated visualization tool into psychotherapeutic practice. Rather than assessing treatment outcomes, the aim was to explore initial engagement, practical feasibility, and interdisciplinary collaboration between human-computer interaction (HCI) specialists and clinical psychologists [17].

Building on these findings, the current study used a Participatory Action Research (PAR) design to codevelop a treatment concept in close collaboration with therapists, patients, and researchers. Through iterative cycles of planning, action, and reflection, the PAR approach allowed for continuous adaptation of the treatment method in response to real-world clinical challenges and patient feedback. This methodology supported the development of a psychotherapeutic approach that is both culturally responsive and grounded in the lived experiences of refugees with CPTSD [31-34].

Aim of This Study

This study aims to codevelop a psychotherapeutic treatment concept that integrates wearable self-tracking and the data visualization tool to personalize interventions for refugees with CPTSD. By focusing on the phenomenological and cultural experiences of individual patients, this approach seeks to bridge the gap between therapy sessions and the challenges refugees face in their daily lives, creating a more data-informed and personalized therapeutic approach.

Methods

Clinical Setting

This study was conducted at the Department of Multidisciplinary Trauma Treatment (DMTT) within the Mental Health Services in the Region of Southern Denmark. DMTT specializes in treating refugees and veterans enduring PTSD and CPTSD, following experiences from war, political persecution, and torture. The clinic uses a multidisciplinary treatment approach, where a team consisting of psychologists, psychiatrists, physiotherapists, nurses, and social workers collaboratively assesses and tailors treatment to each patient. Psychotherapy is conducted by psychologists and is primarily based on CBT and narrative exposure therapy. Psychologists receive monthly supervision. Treatment duration at DMTT is flexible, considering the patient's psychological, social, and somatic complexities [20]. If needed, patients have access to virtual interpreter services during sessions via an iPad (Apple Inc) app.

Design

This study, conducted from November 2022 to April 2024, used a PAR design. Central to the PAR process were iterative cycles of reflection and evaluation, driven by the experiences of both therapists and patients using the self-tracking instrument and data visualization tool within the therapy setting [35].

Peer supervision sessions were an integral part of both the iterative PAR process as well as the data collection. These sessions involved researchers and therapists collaborating to discuss the use of self-tracking technology in therapy. The agenda for each session was based on therapists' logbooks on the patient's therapy session, which documented their observations and experiences in applying the OBT and the data visualization tool. Through qualitative reports from these meetings, peer supervision sessions contributed to the development of the treatment concept.

To ensure transparency and quality in both design and reporting, we drew on the best practice principles for PAR as outlined by Smith et al [36]. These principles emphasize attention to power dynamics, iterative co-construction, and the inclusion of diverse stakeholder perspectives. Rather than following a conventional reporting checklist, this approach allows for a nuanced representation of the collaborative and evolving nature of PAR in a clinical context [36].

Given the focus on trauma within a vulnerable population, we engaged patients as indirect collaborators [37]. Their involvement was captured through interviews at the beginning, middle, and end of treatment, ensuring that their perspectives shaped the research process [38]. These patient insights not only informed the ongoing refinement of both the technology and treatment concept but also enhanced peer supervision discussions, helping therapists improve their use of the technology during therapy sessions.

XSL•FO
RenderX

In this study, qualitative data from interviews and observations were prioritized to capture participant experiences, while the self-tracking data played a descriptive role. This study's framework was defined by two key components: (1) the self-tracking instrument and data visualization tool, and (2) a preliminary model of the treatment concept as introduced in the pilot study.

Ethical Considerations

This study was approved by the Regional Committees on Health Research Ethics for Southern Denmark (ID: S-20210019 CSF) and conducted per the Danish Data Protection Act and the EU General Data Protection Regulation. Before participating in this study, patients provided both oral and written informed consent, while therapists gave oral consent. All participants received written and oral information about this study and provided informed consent before participation. The consent process made clear that data would be anonymized and used for research purposes only within the scope of this approved study. No additional consent was required for analysis, as the original consent covered both data collection and subsequent analysis. To protect participants' privacy and confidentiality, all data were anonymized at the point of collection and securely stored on institutional servers in line with the University of Southern Denmark's data management policies. Only authorized research team members had access to the data. Participants were not compensated for their involvement in this study. No identifiable individuals are depicted in any figures or supplementary materials.

Self-Tracking Instrument and Data Visualization Tool

The OBT is a wearable self-tracking instrument, a research prototype specifically designed for quick, in-the-moment registration of a subjectively experienced phenomenon (Figure 1). Equipped with a single button to enhance user-friendliness, the OBT prototype measures 41×31×12.5 mm, offering versatile wearability options, around the neck, in the pocket, or on the wrist, adjusting to individual preferences [17,20]. When the button on the OBT is pressed, it responds with vibrotactile feedback that matches the duration of the press, serving as a confirmation that the press and its length have been recorded. The OBT's internal storage logs the temporal data associated with each input event, including both the timestamp and duration of the button press. This design emphasizes user adaptability and discretion, minimizing intrusion while facilitating the collection of data of high temporal resolution. It can be operated without visual attention.

Figure 1. The One Button Tracker instrument shown in two different configurations: (1) worn with a neck strap and (2) worn with a wrist band.

The patient can choose to track 1 or 2 phenomena with the OBT, differentiated by single or double button presses, respectively. This is referred to as either a 1- or 2-press protocol. The tracking of subjectively experienced phenomena with the OBT is visualized with the data visualization tool. This involves the patient handing the OBT to the therapist, who then connects it to a desktop or laptop computer using a USB cable. This enables access to a specialized web-based visualization tool that provides a detailed, calendar-like overview of the data, as shown in Figure 2. It highlights the frequency and distribution of identified phenomena across different times of day and days, enhancing a data-driven collaborative analysis between therapist and patient [20].

XSL•FO
RenderX

Figure 2. A screenshot of the web-based calendar visualization tool displaying self-tracking data from patient P7. The bar chart at the top shows the total number of observations per day, while the line represents a 1-week moving average. The scatter plot depicts the first 105 of 252 days of observations (x-axis) and the times of the day the observations were made (y-axis). Each dot represents a single observation made. P: patient.

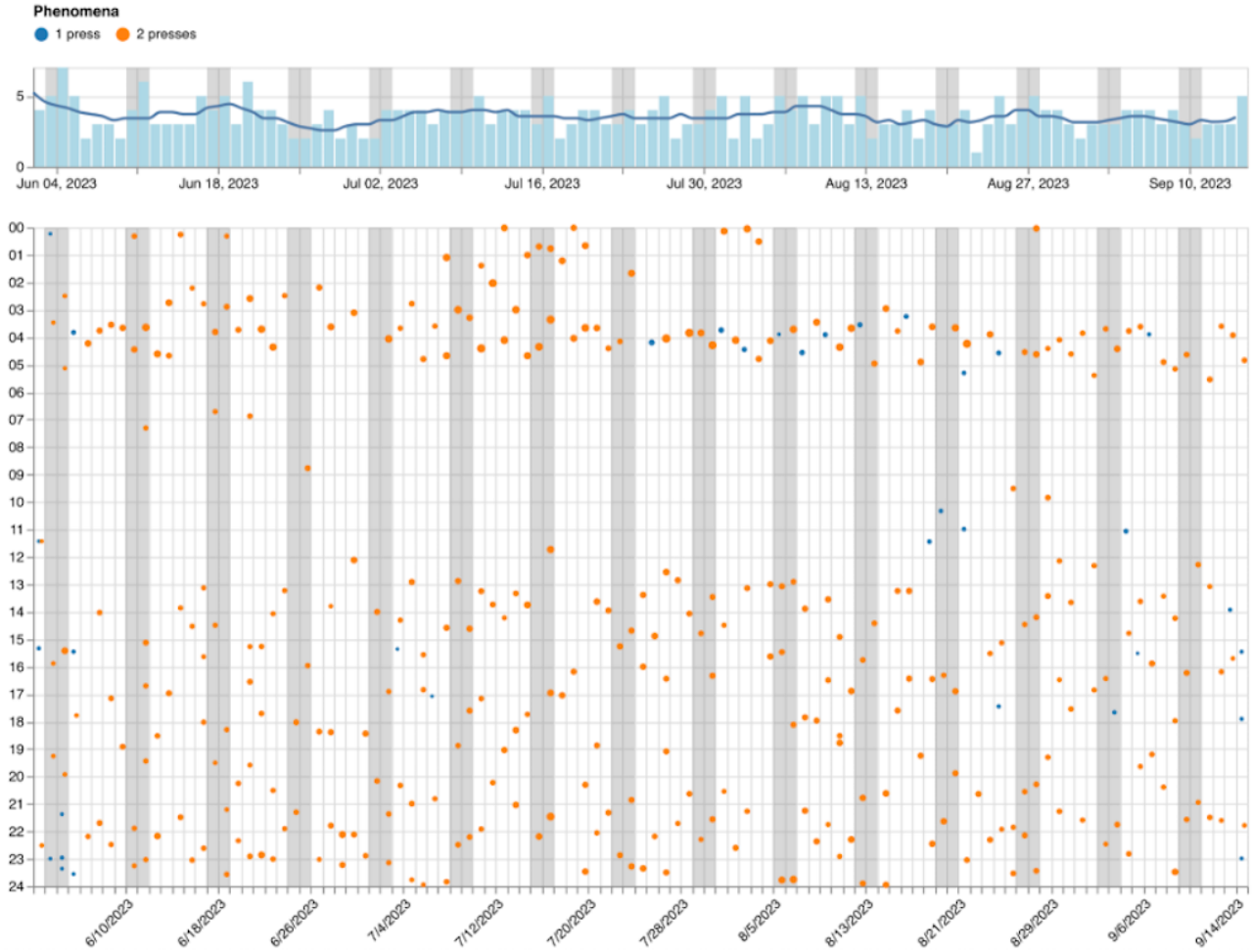

Self-Tracking Assisted Psychotherapy Treatment Concept

Drawing on the principles of the Quantified Self community [39], the OBT is conceptualized within psychotherapeutic treatment as a digital personal diary to track the occurrences of a subjectively meaningful phenomenon. A core aspect of this treatment approach is its focus on personalization, where patients actively choose what they wish to track based on their unique mental health challenges and therapeutic goals. The process of choosing a relevant phenomenon to track is introduced to the patient during the initial therapy sessions, where the rationale for using the OBT is thoroughly explained. An overview of the self-tracking-assisted psychotherapeutic process is presented in Figure 3, with key concepts detailed in Table 1.

XSL•FO
RenderX

Figure 3. Flowchart of the self-tracking-assisted psychotherapy process in the treatment concept. The approach involves the patient self-tracking a target phenomenon in daily life using the OBT instrument. Following the initial definition of the observation protocol, subsequent therapy sessions incorporate collaborative analysis and discussion of collected data by the patient and therapist. This informs the psychotherapeutic process and guides clinical interventions through an iterative process, allowing for redefinition of target phenomena and observation protocols as needed.

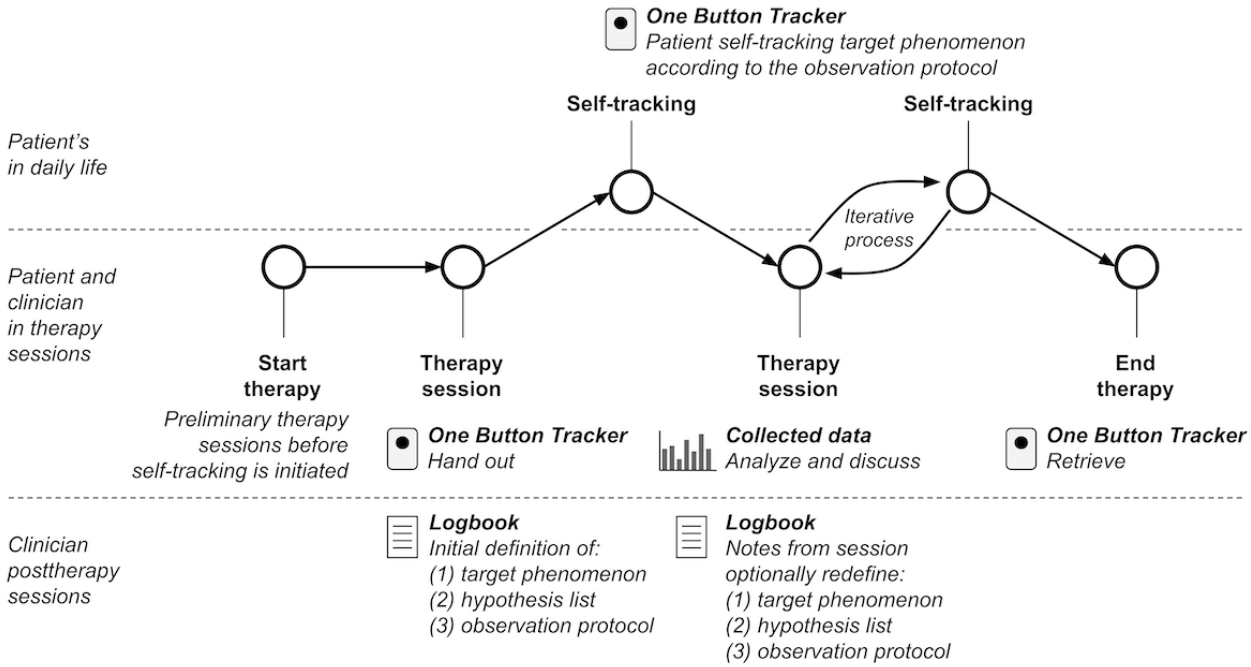

Table 1. Key concepts in the self-tracking-assisted psychotherapy treatment concept.

Key concept	Description
Target phenomenon	A distinct experience, such as a trigger, symptom, or behavior relevant to the patient's mental health challenges.
Hypothesis list	The patient's and the therapist's joint assumptions about the expected distribution of the occurrences of the target phenomenon and potential triggers.
Observation protocol	An agreement on how and when the patient presses the button on the OBT[a] and how to wear the OBT.
Evaluation of observation protocol	Assessing the tracking period between sessions, as well as the patient's ability to identify the target phenomenon, and the patient's experience with tracking.

[a]OBT: One Button Tracker.

In therapy sessions, patients and therapists collaborate to identify a subjectively significant target phenomenon that the patient chooses to track between sessions. Unlike traditional methods where tracking parameters are predefined through validated questionnaires or the therapist, this approach empowers patients to decide what is most relevant to them, fostering a more personalized and engaged tracking process. An observation protocol is collaboratively defined, outlining markers of the chosen phenomenon, whether somatic, mental, or situational, to ensure that patients can accurately recognize and track their selected target phenomenon.

To further align with patient expectations regarding the occurrence of target phenomena, a hypothesis list is developed. This list, reflecting the patient's interpretations of their experiences, is subsequently explored in the next session to examine the implications and insights derived from these hypotheses. As self-tracking progresses, each therapy session includes an evaluation of the observation protocol, ensuring that the tracking activity remains clinically focused and relevant to the patient's treatment goals. This iterative process is central to maintaining the personal relevance and clinical utility of the tracking activity, with in-session analysis of the data guiding any necessary adjustments.

Throughout therapy, this dynamic approach supports the continuous refinement of the target phenomena being assessed. The therapy sessions serve as a space where the tracking process is regularly revisited and updated as needed, always with the patient's input at the forefront. This ensures that the self-tracking activity not only aligns with but actively supports the evolving goals of the treatment.

Participants and Recruitment Procedure

This study involved 21 participants across 4 key groups: therapists, patients, clinical psychology researchers, and HCI researchers. First author LGR was the principal investigator of this study. Initially, 8 therapists from DMTT participated in this study, having been selected for their interest in the project. Over time, this number was reduced to 4 due to resignations and other commitments in the clinic. Additionally, the first author, LGR,

XSL•FO
RenderX

who is both a clinical psychology researcher and a trained psychologist, also participated as a therapist in this study. On the research and development side, 2 clinical psychology researchers (the first author, LGR, included) and 2 HCI researchers participated in this study.

On the patient side, 9 patients from DMTT were included through convenience sampling, based on the following eligibility criteria: being referred for psychiatric treatment, having refugee status, being aged 18 years or older, having a diagnosis of CPTSD according to the International Trauma Interview (Roberts N et al, unpublished data, 2019), and being proficient in Danish.

Data Collection

This study included both qualitative and quantitative data. The data collected included (1) therapy session attendance (ie, the number of therapy sessions patients attended while using the OBT); (2) patient interviews conducted before, during, and after treatment to explore patients' experiences with using the OBT and collaboratively analyze self-tracking data in sessions; (3) self-tracking data collected from the OBT to explore patient engagement and data patterns; (4) therapists' logbooks for each patient with observation protocols and accounts of interventions used in the therapy session to document the therapeutic process; (5) peer supervision meeting minutes with reflections on therapeutic strategies and adjustments discussed during supervision sessions; and (6) workshops (ie, contributions to the collaborative development of the treatment concept).

All interviews were carried out by the first author, LGR, except interviews with 1 patient for whom LGR acted as therapist. These interviews were conducted by coauthor SBM.

Research Process

This study was conducted from November 2022 to April 2024. The phases of the PAR process in this study are visualized in Figure 4. Phases and corresponding activities are described in the following.

Figure 4. Overview of the phases, participants, and activities of the PAR research process. HCI: human-computer interaction; OBT: One Button Tracker; PAR: participatory action research.

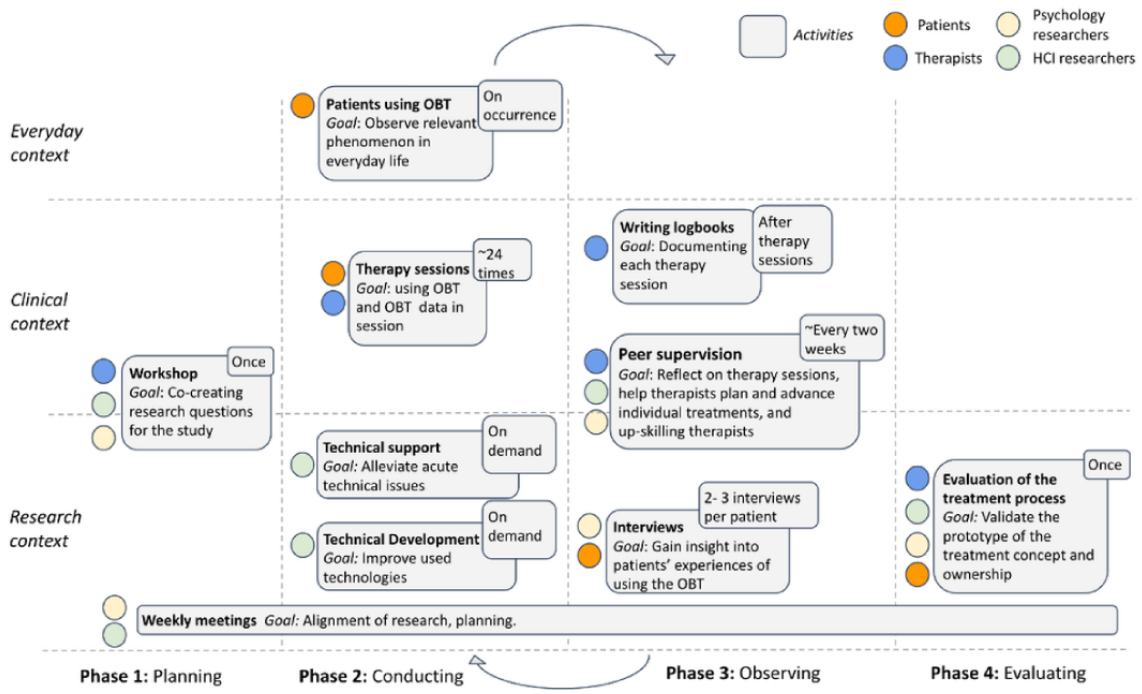

Phase 1: Planning

In November 2022, before patient enrollment, a 4-hour workshop was held to familiarize therapists with the research design, share findings from the pilot study, and collaboratively establish the objectives and research questions (RQs) of this study. Nine participants participated in the workshop, which included 5 therapists, 2 HCI researchers, and 2 clinical psychology researchers.

To begin the group discussion, therapists were introduced to the area of interest, specifically the role of the self-tracking instrument and data visualization tool in linking patients' daily lives with therapy sessions. They were then divided into 3 groups and spent 60 minutes discussing the use of the technology in psychotherapy. Discussions were audio-recorded, and themes were written on A1-sized posters. Group presentations subsequently synthesized these discussions into a poster with the final research themes and questions:

RQ1: ethical and practical considerations: How do therapists navigate ethical concerns when integrating the OBT into psychotherapy, and what practices ensure the protection of patient well-being?

RQ2: patient experiences with using the OBT: How do patients describe their experiences with the OBT in therapy, particularly regarding their personal engagement and self-awareness?

RQ3: setting and achieving therapeutic goals: How do therapists and patients collaboratively determine and reflect on the goals

XSL•FO
RenderX

for using the OBT in therapy, especially concerning trauma processing and emotional regulation?

RQ4: interpretation and use of self-tracking data: How do therapists interpret and integrate self-tracking data into therapy sessions, and how does this affect the therapeutic process?

RQs 1, 3, and 4 will be addressed in this paper, while RQ 2 will be addressed in a separate paper.

Phase 2: Conducting

Phases 2 and 3 were conducted iteratively from November 2022 to April 2024. In this phase, therapists integrated the use of the OBT and the data visualization tool into therapy, following the preliminary model of the treatment concept from the pilot study. Therapists decided when it was appropriate to introduce the OBT in treatment. They collaborated with patients in selecting a target phenomenon, and patients were then guided by therapists to integrate the use of the OBT into their daily lives. HCI researchers provided a worksheet to support therapists in this process (Multimedia Appendix 1).

Weekly meetings with the principal investigator and HCI researchers reviewed the use of the self-tracking instrument and data visualization tool in therapy, addressed technical issues, and discussed methodological insights. Throughout this study's period, HCI researchers offered technical support and made iterative improvements to the data visualization tool, based on feedback acquired and user needs identified, especially during the peer-supervision sessions.

Phase 3: Observing

After each session, therapists recorded key information in logbooks as part of the data collection. In the logbook, therapists outlined the key concepts of the therapy session, a brief statement on the therapy session, whether the self-tracking activity had led to insights for the patients, and any questions that might have come up on the use of the self-tracking technology. The information and questions from these logbooks were then used to outline the agenda for peer supervision sessions, ensuring that key reflections and issues were addressed. During peer supervision, therapists shared progress and challenges, which helped them enhance their skills in using self-tracking data to inform therapy. Discussions in these sessions focused on interventions, observation protocols, and session reflections.

In this phase, patient interviews served dual purposes: they provided insights into patient experiences and enabled patients to be included as cocreators. Insights from these interviews were shared with therapists during peer supervision sessions, contributing to the cocreation of the preliminary treatment concept.

Phase 4: Evaluating

This phase involved a 90-minute hybrid workshop with 4 remaining therapists, 2 HCI researchers, and the first author, LGR, to evaluate the findings from phase 3. The purpose was to evaluate the findings from phase 3 by collaboratively reviewing and refining the treatment concept based on preliminary, thematically organized data.

A web-based shared whiteboard facilitated real-time collaboration, allowing participants to categorize ideas and reflections into distinct themes. This collaborative evaluation process helped streamline and clarify key aspects of the treatment concept. The themes identified during the workshop included (1) introduction and start, (2) integration and engagement with the OBT, (3) communication and relationship building through data and the OBT, (4) therapeutic application and adaptation, (5) data utilization and ownership, and (6) patient agency and autonomy.

The analysis of these themes is presented in the Results section.

Data Analysis

Quantitative self-tracking data and qualitative data from workshops, logbooks, peer supervision sessions, and interviews were included in the collaborative development of the treatment concept.

For this paper, quantitative data provided by the OBT instruments were processed. Interactions with the instruments were categorized as single or double presses. A double press was defined as a second press occurring within 1500 milliseconds of the first; otherwise, it was defined as a single press. The total number of single and double presses was calculated, along with the average number of observations per day and per hour.

The qualitative data were analyzed by the first author, LGR, under the supervision of the coauthors, using the 6 phases of reflexive thematic analysis by Braun and Clarke [40]: familiarization with the data, coding, generating initial themes, reviewing themes, defining and naming themes, and reporting the findings. NVivo (version 12; Lumivero) software was used to assist in this analysis.

Results

Overview

Over 17 months, we conducted 40 peer supervision sessions, 2 workshops, and 25 patient interviews involving 9 patients whose treatment durations varied from 4 to 24 sessions. Most patients were female (6/9, 67%), with ages ranging from 22 to 63 years, resulting in a mean age of 42.8 years.

Self-Tracking Data

Table 2 presents an overview of the number of protocols, target phenomena, treatment duration, and self-tracking data for the 9 patients. The duration of OBT use varied, ranging from 22 to 366 days, with the total number of observations spanning from 37 to 4733, resulting in an average daily number of observations from 0.7 to 37.0. Patients tracked between 1 and 14 different target phenomena, equating to between 1 and 17 distinct observation protocols.

XSL•FO
RenderX

Table 2. Summary of patients in this study, detailing demographic information, treatment sessions, and self-tracking instrument usage. Days without data indicate periods where no data was collected due to technical issues, lost instrument, or lack of instrument access.

Participant	Gender	Age (years)	Total sessions with tracking	Total days with tracking	Days with data	Days without data	Number of observations	Daily average observations	Protocols used	Target phenomena tracked
P1[a]	F[b]	31	12	92	62	30	1276	20.6	4	3
P2	M[c]	63	17	112	87	25	347	4.0	4	4
P3	F	59	24	324	289	35	2038	7.1	17	14
P4	F	22	22	262	185	77	130	0.7	10	8
P5	F	32	13	366	278	88	980	3.5	8	8
P6	M	57	12	246	128	118	4733	37.0	5	5
P7	F	52	20	252	212	40	708	3.3	5	4
P8	M	47	8	175	58	117	188	3.2	1	1
P9	F	22	4	22	22	0	37	1.7	3	3

[a]P: patient.

[b]F: female.

[c]M: male.

Eight out of 9 patients had days without observations during the part of their treatment where the OBT was in use. These gaps were primarily due to technical issues (eg, battery problems), instances where the OBT was misplaced, or situations where the therapist forgot to return the tracking instrument at the end of a therapy session. In some cases, patients actively chose not to track on specific days, such as when they were on vacation (and therefore also wanted to take a break from their treatment) or when they had visitors whom the patients did not want knowing about their treatment.

Figure 5 illustrates the self-tracking data of 2 of the 9 patients (P6 and P7), selected to exemplify the variations in daily observations using 1-press and 2-press protocols. During their treatment spanning 8-9 months, the patients each tracked 2 distinct phenomena. In the monitoring category, these phenomena were a specific intrusive thought (IT) and a trigger of a problematic behavior, respectively. In the intervention category, an alternative coping strategy (ACS) and recording of a completed therapeutic exercise, respectively, were tracked. During this period, the phenomena of interest were adjusted multiple times to better align with their evolving therapeutic needs.

XSL•FO
RenderX

Figure 5. Daily observations, outages, and observation protocols for patients P6 (male, 56 years) and P7 (female, 52 years) from May to March the following year. The topmost row in each patient section displays the histogram of total daily observations, highlighting the frequency and distribution of data collection over time. Below the histograms, horizontal bars denote periods of outages due to technical issues, process changes, or adherence lapses, with different colors representing distinct types of interruptions. The subsequent heat maps illustrate the occurrence of specific phenomena, with the first heat map indicating phenomena that required a single press for recording and the second for those requiring 2 presses. Key changes in observation protocol and the types of phenomena (eg, BE for breathing exercise and ACS for alternative coping strategy) are annotated within the heat maps. ACS: alternative coping strategy; BE: breathing exercise; HA: high arousal; IT: intrusive thoughts; M: monitoring; MT: monitoring trigger; P: patient; RE: recorded exercise; SE: somatic empowerment.

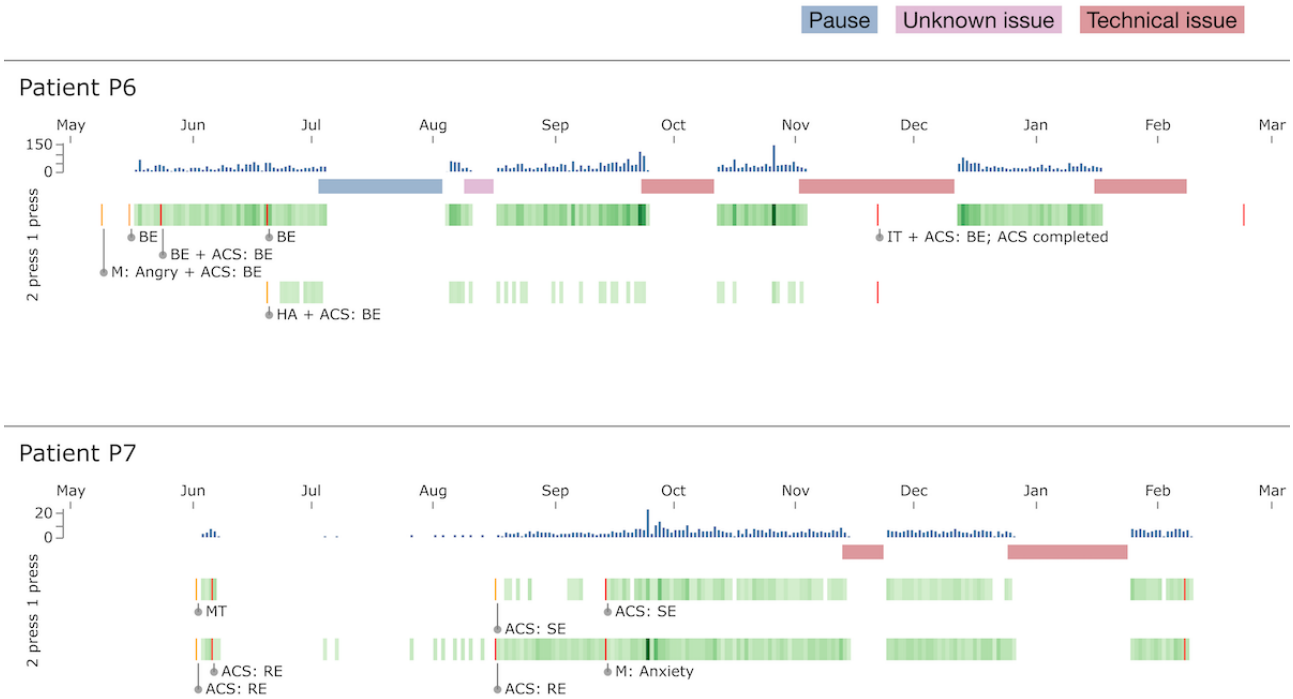

The data visualization shows a timeline of the tracked phenomena for both P6 (male, 57 years) and P7 (female, 52 years), using specific abbreviations. For example, phenomena such as breathing exercise (BE), ACS, high arousal, ITs, somatic empowerment, recorded exercise, monitoring trigger, and monitoring are shown. By analyzing these self-tracking activities, we can see how the patients engaged with the interventions and adapted their tracking practices based on their changing therapeutic goals throughout treatment.

Key Findings

The qualitative findings are presented along a timeline (Figure 6), structured around 3 distinct phases of the therapeutic process. As the context was found to shape how the OBT was used and experienced, the timeline is organized by phase, highlighting both the primary actor and the setting involved. This structure ensures that each theme is grounded in its specific context, offering a clear and coherent view of how the treatment unfolded over time.

XSL•FO
RenderX

Figure 6. A timeline illustrating key thematic findings across 3 phases of the treatment, highlighting the clinical utility of the technology at each therapeutic stage. OBT: One Button Tracker.

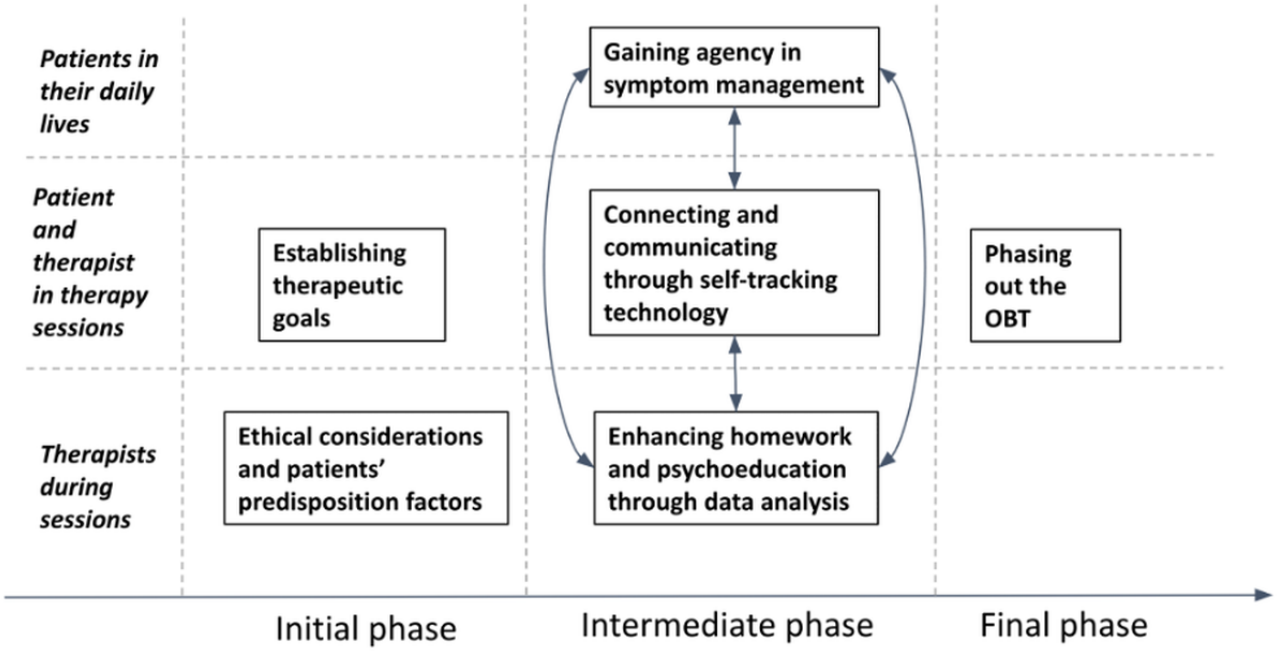

Initial Phase

Ethical Considerations and Patients' Predisposition Factors

Our final workshop highlighted that evaluating patients' predisposition factors was essential during the initial phase of integrating the OBT into therapy. This preliminary assessment enabled therapists to gauge each patient's capacity to manage the mental load associated with self-tracking. For instance, 2 patients declined participation, explaining that attending regular therapy sessions was already overwhelming, and the added strain of using the OBT would be too burdensome.

Our findings indicate that self-tracking technology is most relevant in psychotherapy for patients who already exhibit a motivation for self-help. In interviews, several patients described how beginning therapy and using the OBT encouraged them to take a more active role in managing their mental health. This impression was echoed in peer supervision, where therapists observed that the OBT appeared to cultivate a sense of responsibility, leading to greater patient engagement in the therapeutic process.

Logbook entries emphasized the importance of screening for depressive symptoms before introducing the OBT. In depressive states, patients often gravitate toward negative emotions or experiences, which can distort the focus of the self-tracking activity and diminish its therapeutic value. A drop in engagement or sparse data may signal that the patient is overwhelmed and should be addressed by the therapist. At the same time, these patterns make the OBT particularly valuable: by exposing subtle shifts in mood and behavior, it helps identify when patients are struggling.

For example, after 6 sessions, 1 patient showed signs of increased withdrawal, reflected in a noticeable drop in OBT interactions. By analyzing the self-tracking data, the therapist identified prolonged social isolation as a contributing factor to worsening depressive symptoms. In response, the therapist introduced psychoeducation and cognitive restructuring and repurposed the OBT to track positive thoughts about relatives and encourage social interactions.

At the same time, some concerns arose about the OBT's suitability for patients experiencing emotional instability or complex relational dynamics. One patient, who frequently cancelled sessions and exhibited overall emotional instability, found the OBT overwhelming, particularly when compounded by other life challenges. After 8 sessions, she and the therapist agreed to discontinue its use. Still, she reflected that the brief period of self-tracking had offered valuable insights into her maladaptive patterns, suggesting that it can still offer therapeutic benefits even when used briefly.

Establishing Therapeutic Goals

In peer supervision sessions, it became evident that the initial phase of therapy served as an exploratory period where therapists and patients collaboratively worked to identify and define target phenomena that aligned with the patient's broader treatment goals. This phase was essential for assessing whether tracking a specific phenomenon could reveal meaningful patterns or contextual insights into the patient's mental health challenges. Therapists emphasized the importance of choosing phenomena that patients could easily recognize and reliably track in everyday life. These were typically introduced and refined over 2 sessions, with attention to identifying precise mental or somatic markers to ensure relevance and accuracy.

Although therapists and patients agreed that self-tracking should serve treatment goals, logbook data showed that maintaining this alignment was difficult in practice. While some target phenomena were clearly defined at the outset, their connection to the overarching goals sometimes weakened over time, particularly when the target phenomena or observation protocols

XSL•FO
RenderX

were frequently adjusted. Defining phenomena with sufficient precision also remained a recurring difficulty. Both patients and therapists sometimes relied on abstract or vague terms such as “anxiety” or “anger,” which lacked the specificity needed for self-tracking to meaningfully inform therapeutic work. As a result, therapists found it difficult to work with stable, goal-aligned data throughout treatment.

This difficulty was less pronounced when the target phenomena involved concrete coping strategies, such as BEs or visualization techniques. Tracking these interventions provided a more structured and consistent therapeutic focus, likely because therapists were already familiar with these well-established interventions. This familiarity made it easier to guide patients in tracking their use, thereby maintaining a clearer link between the tracking activities and treatment goals. In contrast, identifying target phenomena that were personally relevant to patients, such as specific thoughts, emotions, or behaviors, was new to both therapists and patients. This novelty may have contributed to greater challenges in maintaining a clear therapeutic focus, as it required therapists to support patients in identifying and working with subjective experiences that were more diffuse and harder to define.

However, when tracking protocols that focused on alternative coping strategies without a clearly defined underlying problem, a disconnect sometimes emerged between the therapist and patient. As seen in logbooks, in these cases, patients were asked to track strategies such as BEs, but it remained unclear when and why they should be used. Without a shared understanding of the problem that the strategy was meant to address, the tracking activity lost direction and therapeutic relevance. These instances point to the importance of clearly defining both the problem and the specific context in which the coping strategies are to be applied. When this link is established, tracking coping strategies can become a meaningful and targeted part of the therapeutic process rather than a standalone activity.

The use of the OBT can also help bring such misalignment to light. As the use of the instrument requires clear definitions of what is being tracked, it can expose when interventions are being applied without a strong connection to the underlying issue. While the presence of a disconnect cannot be resolved by technology alone, the use of the treatment concept along with the OBT makes it more apparent for the therapist to notice such gaps. In some cases, this increased clarity enabled therapists to re-establish the connection between the problem and the coping strategy.

Intermediate Phase

Gaining Agency in Symptom Management

As therapy progressed, patients described in interviews a dynamic interaction with the OBT, which helped them to gain agency over distressing thoughts and maladaptive behaviors as they occurred. In the intermediate phase, data from interviews, logbooks, and peer supervision sessions indicated that most patients experienced a sense of agency in managing their mental health challenges, even among those who engaged minimally with the OBT. According to interviews, carrying the OBT and pressing the button during everyday activities was often guided by a clear intention and hope for change. For example, notes from peer supervision and logbook entries showed how 1 patient, a young woman, initially tracked binge eating episodes as her target phenomenon but later refined it to track the urge to binge eat instead. Over 4 sessions, she reported in an interview that she gained insights into her behavioral patterns and explained how pressing the button served as a mindful pause that allowed her to reassess her emotional impulses and disrupt her habitual response to emotional discomfort.

Several patients described in interviews that tracking coping strategies felt empowering and offered a way to take constructive action in moments of distress. Many used the OBT to shift attention away from ITs, helping them regain mental presence. A woman of Afghan descent, overwhelmed by catastrophic thoughts related to her health, described an immediate cognitive shift with a button press, comparing it to “turning off a light.” This simple yet impactful action helped her gain agency over her anxiety. In the interview, she described how pressing the button required a deliberate change in mindset, demonstrating how a button press facilitated a cognitive shift for her.

Interviews and peer supervision data further revealed that, for some patients, the vibrotactile feedback from the press developed into an intervention in its own right. One patient, for example, began pressing the button for extended periods—up to 43 seconds—and requested that the vibrotactile feedback be intensified to resemble an electric shock. This sensation, he explained in an interview, helped distract him during moments of intense anger.

In several interviews, patients described the vibrotactile feedback as part of an empowering feedback loop. They emphasized how the OBT’s immediate, physical response allowed them to feel understood without having to explain themselves. One patient remarked that the OBT did not ask “dumb questions,” describing it instead as a silent but supportive companion in everyday life.

Engagement and Psychoeducation Through Data Analysis

The iterative nature of the therapeutic process became particularly evident during the intermediate phase, as self-tracking data from the OBT was regularly discussed in sessions. According to therapists in peer supervision, the process initially involved a detailed data analysis, but gradually shifted toward establishing a baseline of patient engagement. One therapist, for example, described how a patient developed an observation protocol where they pressed the button once for anxiety-related tremors and twice for an ACS during the early stages of trauma-focused therapy. Although this phase of trauma-focused therapy typically involves a temporary increase in symptoms, the data showed that the patient’s symptoms did not escalate, offering reassurance to the therapist that the patient was coping well despite expectations of initial worsening.

Logbook entries further underscored the value of collaborative data analysis in identifying distressing contexts. In 1 case, a patient realized through her data that visits to her doctor triggered significant anxiety, which she traced back to the fact that her abusive ex-husband was registered at the same clinic.

XSL•FO
RenderX

This insight, discussed during a therapy session, led her to switch doctors to avoid the trigger. Another patient tracked the use of BEs in both calm and stressful situations and, through data analysis with the therapist, identified a link between job-related stress and worsening nightmares. Recognizing this pattern reinforced the importance of maintaining the BE as a coping strategy, helping the patient sustain the practice in daily life.

Data visualization also played a key role in psychoeducation. In interviews, some patients described a sense of satisfaction in seeing visual evidence of their progress. Being able to track improvements over time helped them better understand their mental health challenges and identify which interventions were most effective. One patient of Bosnian descent, for example, reported in an interview that he gamified his data by trying to increase the number of button presses linked to interventions rather than symptoms, which was an approach that symbolized his efforts to take more active control over his recovery.

At the same time, peer supervision sessions revealed several challenges related to the use of self-tracking data. Therapists noted that the level of detail in the visualizations could be difficult to manage in session. Small text sizes and complex graphical elements sometimes made it harder for both therapists and patients to engage meaningfully with the data. In 1 case, a patient became confused when the color scheme of the visualization changed due to a setup error, which disrupted his patient's connection to the data and led to reduced engagement in subsequent therapy sessions. Interviews also showed that some patients struggled to recall the context of individual data points, likely related to difficulties with autobiographical memory. One therapist described how the sheer density of the data made it difficult to work with in detail during therapy sessions.

Connecting and Communicating Through Self-Tracking Technology

Data from the OBT proved instrumental in fostering a shared understanding of symptom patterns, contextual triggers, and patient engagement, which appeared to positively influence the therapeutic alliance. As defined by Bordin [23], the therapeutic alliance consists of agreement on goals, collaboration on tasks, and an emotional bond between patient and therapist. Across interviews and peer supervision sessions, therapists and patients alike described how the integration of self-tracking data appeared to strengthen all 3 components of the alliance.

Engaging with self-tracking data often establishes a sense of partnership between therapists and patients. In interviews, patients emphasized the communicative value of the data, noting that it offered their therapists nonverbal insights into their daily experiences. For some, the data collected and the vibrotactile feedback functioned as a nonverbal connection, reinforcing their sense of being heard and supported, and contributing to an increased feeling of agency. Therapists, in turn, reported that the data helped them gain a clearer picture of when and where symptoms occurred, enabling more targeted interventions. This was particularly valuable when working with patients who were not fluent in Danish, as it reduced the pressure on verbal communication. In an interview, 1 patient explained that he did not have to try to express himself because the therapist could see how he was feeling through the data.

The perceived connection through the OBT offered a sense of continuity between sessions and a source of support during moments of isolation or distress. One woman, for example, described in an interview how the OBT served as a direct link to her therapist. Before using the instrument, she struggled with anxiety and tended to isolate herself as a coping strategy. The presence of the OBT changed this pattern by serving as a tangible reminder of therapy, helping her stay engaged and feel supported between sessions. For her, pressing the button became a way to feel heard without having to speak, a nonverbal affirmation of her therapeutic connection.

Sharing and collaboratively analyzing the data during therapy sessions also appeared to strengthen the therapeutic alliance. According to patient interviews, this process helped patients feel recognized and supported, and ensured that both the therapist and patient remained aligned on the goals and tasks related to the tracking activity. Peer supervision sessions provided additional examples of how the OBT enhanced collaboration in practice. One patient described pressing the OBT button to ensure that their experiences were communicated to the therapist, even without verbal explanation. This perception was reinforced when the therapist reviewed the data in session and invited reflection, creating a shared space of understanding and helping the patient feel seen.

Final Phase: Phasing Out the OBT

During the concluding phase of therapy, peer supervision discussions centered on how to support patients in transitioning away from the OBT. Therapists noted that some patients, having reached their treatment goals, often felt no further need for the OBT. For others, however, the OBT had taken on a symbolic role in the therapeutic process, making its removal more sensitive. To ease this transition, some therapists introduced the idea of a transitional object, often drawing on tools from occupational therapy, such as a fidget ring. The choice of tool depended on the aspect of the OBT that patients valued most, whether it was the vibrotactile feedback, the act of carrying the instrument, or the sense of support it represented.

Patient interviews revealed that letting go of the OBT was experienced as a challenge, especially for those who had integrated it into their daily routines. Some described the use of the OBT as an embodied habit, noting that it took 2-3 weeks both to adopt and to unlearn after returning the instrument at the end of treatment. Others described a sense of emotional loss, particularly the perceived connection to their therapist that had been reinforced through the OBT's responsive vibrotactile feedback. For some, these 2 aspects were intertwined: both the physical act of pressing the button and the emotional connection it symbolized were missed after the instrument was withdrawn.

One woman, for example, was given a rock with a symbolic meaning as a transitional object. In an interview, she shared that while the gesture was appreciated, she still missed the vibrotactile feedback from the OBT, which made her feel connected to her therapist. In addition, 2 patients expressed a

XSL•FO
RenderX

strong desire to retain access to their data, emphasizing the valuable insights they gained from using the OBT.

From the therapists' perspectives, discussed in peer supervision, these difficulties were partly attributed to the sudden absence of ongoing data collection and collaborative data analysis. These features of the OBT had supported both the therapeutic alliance and engagement throughout the treatment process, and their removal sometimes left a noticeable gap as therapy concluded.

Discussion

Principal Findings

Our study demonstrates how PAR methodology can be used to codevelop a psychotherapeutic treatment concept for refugees with CPTSD. Through iterative and collaborative processes involving therapists, patients, and researchers, we developed a culturally sensitive and personalized treatment approach in real-world clinical settings that integrates the OBT and the data visualization tool and can be used across psychotherapeutic modalities. Below, we discuss our findings about the 3 RQs that guided this study: ethical and practical considerations (RQ1), the setting and achievement of therapeutic goals (RQ3), and the interpretation and use of self-tracking data in therapy (RQ4).

Ethical and Practical Considerations of Self-Tracking in Psychotherapy

Our findings on RQ1 highlight several ethical and practical considerations for integrating the OBT and the data visualization tool into psychotherapy with refugees. One particularly interesting insight from our study is the variability in how patients engaged with the OBT, both in terms of duration of OBT use, ranging from 22 to 366 days, as well as the number of daily observations, ranging from an average of 0.7 to 37. Most of this variability is likely to be explained by the inherent flexibility in tracking different target phenomena. However, the variability could also be partly due to some patients experiencing technical issues with the OBT, and not entirely unexpected with the OBT being a research prototype. For example, gaps in data collection were observed due to technical issues such as battery problems, running out of memory space, as well as practical issues such as misplaced instruments or therapists forgetting to return the OBT to the patient at the end of the session. Although patients were generally able to re-engage with the OBT, these gaps in data collection highlight the need for further technical and practical adjustments to ensure more consistent tracking. Addressing these issues could enhance the reliability of self-tracking data, thereby making it a more dependable tool for therapeutic interventions. These disruptions raise important practical questions about the workflow for self-tracking technology in psychotherapy, as periods without data could hinder the therapeutic process.

From an ethical perspective, the variability in how different patients responded to self-tracking highly underscores the relevance of personalizing self-tracking to the individual needs of patients. For some, particularly those tracking distressing phenomena, the act of self-tracking seemed to be overwhelming sometimes. One patient, who was experiencing emotional disturbances, found the OBT to intensify her feelings of being overwhelmed, leading her to discontinue its use. Additionally, that was even though she reported gaining insights into her maladaptive behavioral patterns. This raises ethical concerns about the potential burden that self-tracking can impose on vulnerable patients who are struggling with emotional instability. Given that refugees with complex trauma often face severe mental health challenges [7], it is essential to ask whether self-tracking adds to these burdens or merely brings existing vulnerabilities to light. Our findings suggest that self-tracking likely reveals pre-existing challenges, though further research is needed to substantiate this and examine how the self-tracking technology can be applied in diverse patient populations.

However, for other patients, the adaptability of the OBT and the flexibility in selecting target phenomena were a strength. The ability to shift the focus from self-tracking a distressing phenomenon to more positive, alternative coping strategies seemed to give some patients a sense of control and allowed them to use therapeutic interventions in their daily lives. This highlights the strength of the personalization enabled by the OBT, aligning with previous research that emphasizes the empowering effect of active patient participation in mental health care [41]. Nevertheless, this variability in the use of the OBT across different patients also presents a challenge, as the flexible application of the OBT can make it difficult to establish a consistent approach for its use in therapy. This aligns with findings from a study by Peñuela-O'Brien et al [11], who found that health professionals often adapt their practices to accommodate the diverse needs of refugee patients, leading to a range of therapeutic approaches. Future research should explore how to find a balance between the flexibility in use of the OBT and the value creation seen from a therapeutic perspective.

Lastly, the role of peer supervision was instrumental in helping therapists navigate these ethical and practical complexities. Peer discussions allowed therapists to reflect on the best strategies for introducing the self-tracking technology into therapy, balancing the potential benefits with the risks of cognitive overload. This highlights the importance of ongoing professional support when integrating new technology into therapy, particularly when working with vulnerable populations. For the treatment concept, we recommend incorporating regular peer supervision as part of the concept to ensure ethical and effective use of the technology in diverse clinical settings.

Setting and Achieving Therapeutic Goals

Our findings reveal the complexities of using the OBT to align self-tracking activities with therapeutic goals, especially as the target phenomena shifted throughout therapy. This difficulty may stem from the novelty of using the OBT to track subjective experiences related to mental health challenges and integrating self-tracking data into therapy. One of the key challenges was the difficulty in operationalizing abstract emotional states such as "anxiety" and "anger" into discernible phenomena that could be recognized by the patient and tracked. As a result, some therapists and patients struggled to maintain consistent alignment between the self-tracking activity and the therapeutic goals, particularly when the phenomena being tracked changed

XSL•FO
RenderX

frequently. This reflects the broader challenges faced in psychotherapy with refugees, where therapists often deviate from standard protocols to accommodate the cultural and complex needs of patients, but there is insufficient transparency in how this adaptation is carried out [42]. Integrating self-tracking technology introduces a new dimension to therapy, requiring both therapists and patients to learn and adapt. The early stage of the treatment concept development and the iterative nature of the PAR methodology likely contributed to a steep learning curve, which could have complicated the establishment of clear, consistent therapeutic goals.

Moreover, the flexibility of the OBT to track both distressing phenomena and alternative coping strategies created opportunities for patient engagement but also posed challenges in maintaining a clear therapeutic focus related to the goal of the therapy. For some patients, tracking alternative coping strategies such as BEs was more easily operationalized and helped foster a sense of control over their mental health. However, others who frequently changed their target phenomena experienced confusion about the tracking activity, and the relation of this to therapeutic goals was not always clear for either the therapist or the patient. This variability in the use of the OBT highlights the need for therapists to continuously reassess and adapt the target phenomena to ensure they remain relevant to the patient's evolving therapeutic needs while maintaining a focus on the overall therapeutic goal.

At the same time, it is crucial to balance flexibility with consistency. Frequent changes to the target phenomena can interrupt the therapeutic process and confuse, making it more difficult for both patient and therapist to observe meaningful patterns and progress. Therefore, this study recommends limiting how often the target phenomena are changed and making shifts only when there is a clear therapeutic benefit. This method helps keep the therapeutic focus stable while still allowing for necessary adjustments. However, even when the target phenomena shifted frequently or the definitions became less clear, patients still gained insights into their mental health challenges, reinforcing the importance of collaborative goal-setting and self-tracking as an effective therapeutic instrument.

One possible explanation for why some treatment courses seem to have less focus on the overall therapeutic goal of the therapy, and instead an increased focus on the tracking activity itself, could be due to the design of the logbooks. Here, therapists were not explicitly asked to link tracking activities to specific therapeutic goals. This challenge was reinforced by the underuse of the hypothesis list in the logbooks, which may have contributed to a focus on tracking the phenomena themselves rather than linking them to broader therapeutic objectives. This suggests that further refinements in the design of the logbook and more structured guidance for therapists are needed to better connect the self-tracking activity to the overall therapeutic goal of the treatment.

Peer supervision sessions played a critical role in helping therapists navigate these challenges, providing room for supervision and reflection of the target phenomena and tracking protocols as therapy progressed. The iterative nature of the PAR methodology allowed therapists to collaboratively explore new approaches and refine the use of the OBT in line with patients' needs, demonstrating the importance of ongoing professional support when integrating self-tracking technology into therapy. For the treatment concept, we recommend clarifying the therapeutic goal before adapting or changing the target phenomenon to ensure that the tracking activity remains aligned with the overall direction of therapy.

Interpretation and Use of Self-Tracking Data

Our findings show that integrating self-tracking data into therapy sessions presents both opportunities and challenges for therapists and patients. The OBT's potential to positively influence patient engagement and offer a feasible and personalized way to track mental health phenomena was a key insight. Many patients felt empowered by taking an active role in tracking their mental health, which aligns with existing literature showing that active patient involvement increases satisfaction and feelings of empowerment [41]. For instance, the long-term engagement observed in our study, with patients tracking for up to 366 days, suggests the broad applicability of the OBT across diverse patient groups, irrespective of age or cultural background.

However, integrating self-tracking data into therapy sessions presented challenges for some therapists. One notable issue was that, despite the collaborative analysis of data during sessions, therapists reported that it was sometimes difficult to balance the dual attention required to interpret the data while remaining present and attending to the patient's immediate needs. This tension between the use of traditional therapeutic methods and the analysis of self-tracking data suggests that integrating digital health instruments into therapy may require more than simply adding data-driven components: it calls for a rethinking of how these tools interact with the flow of therapy. Therapists found that while the OBT provided valuable insights, it, unsurprisingly, required additional effort to manage alongside standard therapeutic practices.

We initially anticipated that the self-tracking data would lead to more systematic functional analysis, a core component of CBT, where specific behaviors and emotional triggers are mapped and addressed. However, in practice, some contextual cues associated with the data points were analyzed, but the comprehensive approach of functional analysis was not fully adopted. The self-tracking data served more as a navigational tool that helped therapists monitor patient engagement with the OBT and their target phenomena, as well as their overall progress. This finding aligns with previous research, which has noted that while digital tools can offer valuable data, the ability to integrate and analyze this data effectively in therapy remains a challenge [43]. It seems likely that clearer guidelines on how to integrate functional analysis with the self-tracking data provided by the OBT, as well as specific examples, might have facilitated more consistent adoption of this approach. The cognitive load of managing both the self-tracking data and the therapeutic conversation in-session, coupled with the developing nature of the treatment concept in this study, may also explain why comprehensive functional analyses were not fully used.

Notwithstanding these challenges, our findings also reveal the OBT's potential to foster deeper patient engagement in therapy.

XSL•FO
RenderX

By actively involving patients in the integration of the OBT and the self-tracking activity, the OBT empowered them to take ownership of their treatment, particularly through the tracking of personally meaningful phenomena. This aligns with the study by Kazantzis and Lampropoulos [44], who emphasize the importance of involving patients in the cocreation of therapeutic activities, such as homework, to enhance engagement and promote self-efficacy.

The added cognitive load of interpreting self-tracking data alongside ongoing therapeutic tasks might explain the underuse of certain tools in the logbook, such as the hypothesis list. To overcome this, future development work should aim to simplify and better integrate data analysis into routine therapeutic practices, allowing therapists to leverage self-tracking data more effectively to enhance the therapeutic process and patient outcomes.

Moreover, the vibrotactile feedback provided by the OBT upon button presses emerged as an unexpected but important therapeutic element. Some patients reported that the vibrotactile feedback from pressing the OBT button acted as a grounding technique, helping them manage distress in the moment. For these patients, the vibrotactile feedback created a sense of control over their symptoms, reflecting the findings from previous studies that users often form stronger connections with tangible, interactive devices such as jewelry, which can evoke more personal and intimate interactions compared to abstract digital interfaces [45]. This feature was also perceived as supporting nonverbal communication for some of the patients. By offering the patients a nonverbal means of expression, the OBT seemed to serve as a bridge between patient and therapist, allowing for deeper insights into the patient's mental state without the need for explicit verbalization. This reinforces the complementary role of technology in supporting traditional therapeutic communication methods, which, as Woodward et al [46] emphasize, can enhance accessibility and engagement in mental health interventions, particularly in contexts where traditional communication may be challenging.

One of the most significant contributions of the OBT was its potential to strengthen the therapeutic alliance between patients and therapists. Our findings suggest that the OBT positively affected all 3 components of the therapeutic alliance model by Bordin [23]: bond, goals, and tasks. By engaging patients in defining target phenomena and collaboratively analyzing the data, the OBT fostered a deeper sense of partnership between therapist and patient. This collaborative approach helped clarify therapeutic goals, while the use of the OBT for tracking specific phenomena ensured that both patient and therapist were aligned on the tasks of therapy. This aligns with previous studies emphasizing the importance of the therapeutic alliance in treating refugees, particularly the need for mutual understanding and trust in addressing complex trauma [10]. Similarly, Mohl et al [47] reported that patients who perceived a poor therapeutic alliance were more likely to drop out of treatment. Strengthening the alliance in this way appears to be crucial for maintaining engagement and facilitating therapeutic progress, particularly for patients with complex trauma histories.

Interestingly, therapists in our study often viewed the OBT as particularly useful in supporting homework activities between sessions. The real-world application of the OBT allowed therapists to gain insights into how patients were engaging with therapeutic interventions outside of the therapy session. This proved particularly useful in addressing the challenges posed by language barriers, as the OBT allowed patients to engage with therapeutic tasks in a nonverbal, tangible way, bypassing the need for complex verbal explanations. The ability of the OBT to act as an instrument to support homework rather than as a primary object of analysis during therapy sessions also suggests that self-tracking data may be best used to supplement, rather than dominate, the therapy session.

Limitations

Even with the challenges, this study delivered promising results. This study was conducted at DMTT, and while this is a strength in the PAR design, factors such as turnover of personnel, ongoing projects, and the decision to relocate the clinic during this study's period may have impacted both the iterative process of developing the treatment concept and the onboarding of therapists to this study. These circumstances likely affected the consistency of data collection and could have influenced therapist engagement with using the self-tracking technology in therapy sessions.

The participating patients could have been beyond interviews involved in the PAR framework; however, due to their vulnerable status and need to protect their privacy, we chose to involve them indirectly. Furthermore, physiotherapists were not included in the cocreation process, even though it became evident during this study that many interventions were related to the patients' physiotherapy treatments. Including physiotherapists could have enriched the cocreation process and provided more comprehensive insights into the treatment interventions.

There were technical issues related to the OBT's battery life. The OBT depends on obtaining a GPS signal to set the clock in the event the battery becomes depleted. If the OBT was not charged and restarted near a window, it could not receive a GPS signal, in some situations, causing the battery to drain prematurely due to the retries of trying to acquire a GPS signal.

Conclusions

This study sets out to codevelop and refine a self-tracking-assisted psychotherapeutic treatment concept for refugees with CPTSD. Building on a feasibility pilot study [17], we engaged patients, therapists, and researchers in a collaborative process to codevelop a treatment model that extends psychotherapy beyond the clinical setting.

Findings suggest that the OBT and the data visualization tool supported patient engagement by enabling in-the-moment self-tracking, creating agency, and strengthening the connection between daily experiences and therapy. Therapists reported that the self-tracking data provided detailed insights that enhanced personalization and informed clinical decision-making. The flexibility of the method allowed it to be adapted to individual needs and therapeutic goals.

XSL•FO
RenderX

By embedding lived experience into psychotherapy, this study shows the potential of momentary, low-burden self-tracking as a meaningful contribution to personalized care. Future research should explore this approach in other clinical contexts and assess how it can further support engagement, alliance, and continuity in care.

Acknowledgments

We would like to thank the patients, whose willingness to share their experiences has enriched our understanding and approach to treating complex posttraumatic stress disorder. We are equally grateful to the therapists whose expertise has been invaluable in cocreating the psychotherapeutic treatment concept. We also wish to acknowledge the Department of Multidisciplinary Trauma Treatments at the Mental Health Services in the Region of Southern Denmark for their support and collaboration. This study has received financial support from The Danish Victims Fund (20-610-00094). The execution, content, and results of the materials are the sole responsibility of the authors. The analysis and viewpoints that have been made evident from the materials belong to the authors and do not necessarily reflect the views of The Council of The Danish Victims Fund.

Data Availability

The datasets generated and analyzed during this study are not publicly available due to the sensitivity of the participant data and the small cohort size. All participants were refugees undergoing psychotherapy for complex posttraumatic stress disorder, and consent covered the use of anonymized data for this specific study only. Researchers may contact the corresponding author to request access to selected deidentified materials, where feasible. Such requests will be considered on a case-by-case basis per Danish data protection legislation.

Conflicts of Interest

The One Button Tracker was invented and developed by authors JEL and TBC.

Multimedia Appendix 1
Worksheet on questions for OBT treatment. OBT: One Button Tracker.
[DOCX File , 18 KB - formative_v9i1e66663_app1.docx]

References

1. Refugee data finder. UNHCR. 2022. URL: https://www.unhcr.org/refugee-statistics/ [accessed 2025-09-04]
2. Matheson C, Weightman E. A participatory study of patient views on psychotherapy for complex post-traumatic stress disorder, CPTSD. J Ment Health 2021;30(6):690-697. [doi: 10.1080/09638237.2020.1803229] [Medline: 32772598]
3. International classification of diseases for mortality and morbidity statistics (11th Revision). World Health Organization. 2022. URL: https://www.who.int/standards/classifications/classification-of-diseases [accessed 2025-09-04]
4. de Silva U, Glover N, Katona C. Prevalence of complex post-traumatic stress disorder in refugees and asylum seekers: systematic review. BJPsych Open 2021;7(6):e194 [FREE Full text] [doi: 10.1192/bjo.2021.1013] [Medline: 34649634]
5. Heide FJJT, Mooren TM, Kleber RJ. Complex PTSD and phased treatment in refugees: a debate piece. Eur J Psychotraumatology 2016;7(1):28687. [doi: 10.3402/ejpt.v7.28687] [Medline: 26886486]
6. Cloitre M. The "one size fits all" approach to trauma treatment: should we be satisfied? Eur J Psychotraumatol 2015;6:27344 [FREE Full text] [doi: 10.3402/ejpt.v6.27344] [Medline: 25994021]
7. Karatzias T, Murphy P, Cloitre M, Bisson J, Roberts N, Shevlin M, et al. Psychological interventions for ICD-11 complex PTSD symptoms: systematic review and meta-analysis. Psychol Med 2019;49(11):1761-1775. [doi: 10.1017/S0033291719000436] [Medline: 30857567]
8. Gruner D, Magwood O, Bair L, Duff L, Adel S, Pottie K. Understanding supporting and hindering factors in community-based psychotherapy for refugees: a realist-informed systematic review. Int J Environ Res Public Health 2020;17(13):4618 [FREE Full text] [doi: 10.3390/ijerph17134618] [Medline: 32604990]
9. Bernardsdóttir B, Sandahl H, Carlsson J, Mortensen EL, Palic S. Psychotherapy readiness domains as predictors of psychotherapy outcome in trauma-affected refugees. J Affect Disord 2023;325:248-255. [doi: 10.1016/j.jad.2022.12.130] [Medline: 36586606]
10. Karageorge A, Rhodes P, Gray R, Papadopoulos R. Refugee and staff experiences of psychotherapeutic services: a qualitative systematic review. Intervention 2017;15(1):51-69. [doi: 10.1097/wtf.0000000000000137]
11. Peñuela-O'Brien E, Wan MW, Edge D, Berry K. Health professionals' experiences of and attitudes towards mental healthcare for migrants and refugees in Europe: a qualitative systematic review. Transcult Psychiatry 2023;60(1):176-198 [FREE Full text] [doi: 10.1177/13634615211067360] [Medline: 34986056]
12. Tribe RH, Sendt K, Tracy DK. A systematic review of psychosocial interventions for adult refugees and asylum seekers. J Ment Health 2019;28(6):662-676. [doi: 10.1080/09638237.2017.1322182] [Medline: 28485636]

XSL•FO
RenderX

13. Li SSY, Liddell BJ, Nickerson A. The relationship between post-migration stress and psychological disorders in refugees and asylum seekers. Curr Psychiatry Rep 2016;18(9):82. [doi: 10.1007/s11920-016-0723-0] [Medline: 27436307]
14. Bond RR, Mulvenna MD, Potts C, O'Neill S, Ennis E, Torous J. Digital transformation of mental health services. npj Ment Health Res 2023;2(1):13 [FREE Full text] [doi: 10.1038/s44184-023-00033-y] [Medline: 38609479]
15. Korotitsch WJ, Nelson-Gray RO. An overview of self-monitoring research in assessment and treatment. Psychological Assessment 1999;11(4):415-425. [doi: 10.1037//1040-3590.11.4.415]
16. Cohen JS, Edmunds JM, Brodman DM, Benjamin CL, Kendall PC. Using self-monitoring: implementation of collaborative empiricism in cognitive-behavioral therapy. Cogn Behav Pract 2013;20(4):419-428. [doi: 10.1016/j.cbpra.2012.06.002]
17. Riisager LG, Huniche L, Larsen JE, Christiansen TB, Kring L, Palic S, et al. Fostering engagement using a wearable for self-tracking assisted psychotherapy with refugees diagnosed with complex PTSD: a feasibility pilot study. Front Psychiatry 2025;16:1250552 [FREE Full text] [doi: 10.3389/fpsyt.2025.1250552] [Medline: 40078522]
18. Oudin A, Maatoug R, Bourla A, Ferreri F, Bonnot O, Millet B, et al. Digital phenotyping: data-driven psychiatry to redefine mental health. J Med Internet Res 2023;25:e44502 [FREE Full text] [doi: 10.2196/44502] [Medline: 37792430]
19. de Angel V, Lewis S, White KM, Matcham F, Hotopf M. Clinical targets and attitudes toward implementing digital health tools for remote measurement in treatment for depression: focus groups with patients and clinicians. JMIR Ment Health 2022;9(8):e38934 [FREE Full text] [doi: 10.2196/38934] [Medline: 35969448]
20. Riisager LG, Christiansen TB, Moeller SB, Huniche L, Larsen JE. Snapping out of it: how a wearable for self-tracking assisted psychotherapy bridges the gap between thoughts and the world. 2024 Presented at: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems; May 11 - 16, 2024; Honolulu HI USA p. 1-7. [doi: 10.1145/3613905.3637112]
21. Dubad M, Elahi F, Marwaha S. The clinical impacts of mobile mood-monitoring in young people with mental health problems: the MeMO study. Front Psychiatry 2021;12:687270 [FREE Full text] [doi: 10.3389/fpsyt.2021.687270] [Medline: 34393850]
22. Aguilera A, Muñoz RF. Text messaging as an adjunct to CBT in low-income populations: a usability and feasibility pilot study. Prof Psychol Res Pr 2011;42(6):472-478 [FREE Full text] [doi: 10.1037/a0025499] [Medline: 25525292]
23. Bordin ES. The generalizability of the psychoanalytic concept of the working alliance. Psychother: Theory, Res Pract 1979;16(3):252-260. [doi: 10.1037/h0085885]
24. El-Refaay SMM, Toivanen-Atilla K, Crego N. Efficacy of technology-based mental health interventions in minimizing mental health symptoms among in immigrants, asylum seekers or refugees; systematic review. Arch Psychiatr Nurs 2024;51:38-47. [doi: 10.1016/j.apnu.2024.04.002] [Medline: 39034093]
25. van de Belt TH, de Croon A, Freriks F, Blomseth Christiansen T, Eg Larsen J, de Groot M. Barriers to and facilitators of using a One Button Tracker and web-based data analytics tool for personal science: exploratory study. JMIR Form Res 2022;6(3):e32704 [FREE Full text] [doi: 10.2196/32704] [Medline: 35230247]
26. Arendt ITP, Riisager LHG, Larsen JE, Christiansen TB, Moeller SB. Distinguishing between rumination and intrusive memories in PTSD using a wearable self-tracking instrument: a proof-of-concept case study. CBT 2021;14:e15. [doi: 10.1017/s1754470x2100012x]
27. Larsen JE, Eskelund K, Christiansen TB. Active self-tracking of subjective experience with a one-button wearable: a case study in military PTSD. arXiv Preprint posted online on Mar 9, 2017 [FREE Full text]
28. Hsin H, Fromer M, Peterson B, Walter C, Fleck M, Campbell A, et al. Transforming psychiatry into data-driven medicine with digital measurement tools. NPJ Digit Med 2018;1:37 [FREE Full text] [doi: 10.1038/s41746-018-0046-0] [Medline: 31304319]
29. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol 2008;4:1-32. [doi: 10.1146/annurev.clinpsy.3.022806.091415] [Medline: 18509902]
30. Schiepek G, Aichhorn W, Gruber M, Strunk G, Bachler E, Aas B. Real-time monitoring of psychotherapeutic processes: concept and compliance. Front Psychol 2016;7:604 [FREE Full text] [doi: 10.3389/fpsyg.2016.00604] [Medline: 27199837]
31. McTaggart R. Principles for participatory action research. Adult Education Quarterly 1991;41(3):168-187. [doi: 10.1177/0001848191041003003]
32. Baum F, MacDougall C, Smith D. Participatory action research. J Epidemiol Community Health 2006;60(10):854-857 [FREE Full text] [doi: 10.1136/jech.2004.028662] [Medline: 16973531]
33. Verweij L, Metsemakers SJJPM, Ector GICG, Rademaker P, Bekker CL, van Vlijmen B, et al. Improvement, implementation, and evaluation of the CMyLife digital care platform: participatory action research approach. J Med Internet Res 2023;25:e45259 [FREE Full text] [doi: 10.2196/45259] [Medline: 37713242]
34. Leask CF, Sandlund M, Skelton DA, Altenburg TM, Cardon G, Chinapaw MJM, GrandStand, Safe StepTeenage Girls on the Move Research Groups. Framework, principles and recommendations for utilising participatory methodologies in the co-creation and evaluation of public health interventions. Res Involv Engagem 2019;5:2 [FREE Full text] [doi: 10.1186/s40900-018-0136-9] [Medline: 30652027]
35. Kidd SA, Kral MJ. Practicing participatory action research. J Couns Psychol 2005;52(2):187-195. [doi: 10.1037/0022-0167.52.2.187]

XSL•FO
RenderX

36. Smith L, Rosenzweig L, Schmidt M. Best practices in the reporting of participatory action research: embracing both the forest and the trees 1Ψ7. Couns Psychol 2010;38(8):1115-1138. [doi: 10.1177/0011000010376416]
37. Klocker N. Participatory action research: the distress of (not) making a difference. Emotion, Space Society 2015;17:37-44. [doi: 10.1016/j.emospa.2015.06.006]
38. Speedlin S, Haberstroh S, Townsend C, Prasath P, McVay K. Participatory action research: strategies for implementation in counseling outcome research. J Prof Couns: Pract Theory Res 2021;49(1):21-33. [doi: 10.1080/15566382.2021.1949210]
39. Wolf GI, De Groot M. A conceptual framework for personal science. Front Comput Sci 2020;2:21. [doi: 10.3389/fcomp.2020.00021]
40. Braun V, Clarke V. Thematic Analysis: A Practical Guide. Thousand Oaks, California: SAGE; 2022:46-50.
41. Tambuyzer E, Van Audenhove C. Is perceived patient involvement in mental health care associated with satisfaction and empowerment? Health Expect 2015;18(4):516-526. [doi: 10.1111/hex.12052] [Medline: 23425015]
42. Wright A, Reisig A, Cullen B. Efficacy and cultural adaptations of narrative exposure therapy for trauma-related outcomes in refugees/asylum-seekers: a systematic review and meta-analysis. J Behav Cogn Ther 2020;30(4):301-314. [doi: 10.1016/j.jbct.2020.10.003]
43. Torous J, Michalak EE, O'Brien HL. Digital health and engagement-looking behind the measures and methods. JAMA Netw Open 2020;3(7):e2010918 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.10918] [Medline: 32678446]
44. Kazantzis N, Lampropoulos GK. Reflecting on homework in psychotherapy: what can we conclude from research and experience? J Clin Psychol 2002;58(5):577-585. [doi: 10.1002/jclp.10034] [Medline: 11967882]
45. Niemantsverdriet K, Versteeg M. Interactive jewellery as memory cue: designing a sound locket for individual reminiscence. 2016 Presented at: Proceedings of the TEI ?16: Tenth International Conference on Tangible, Embedded, and Embodied Interaction; February 14-17, 2016; Eindhoven, Netherland p. 532-538. [doi: 10.1145/2839462.2856524]
46. Woodward A, Burchert S, Barry AS, Broerse JE, Sondorp E, Bold A, et al. Scalability of digital psychological innovations for refugees: a comparative analysis in Egypt, Germany, and Sweden. SSM - Mental Health 2023;4:100231. [doi: 10.1016/j.ssmmh.2023.100231]
47. Mohl PC, Martinez D, Ticknor C, Huang M, Cordell L. Early dropouts from psychotherapy. J Nerv Ment Dis 1991;179(8):478-481. [doi: 10.1097/00005053-199108000-00005] [Medline: 1856710]

Abbreviations

ACS: alternative coping strategy
BE: breathing exercise
CBT: cognitive behavioral therapy
CPTSD: complex posttraumatic stress disorder
DMTT: Department of Multidisciplinary Trauma Treatment
HCI: human-computer interaction
ICD-11: International Statistical Classification of Diseases, Eleventh Revision
IT: intrusive thought
OBT: One Button Tracker
PAR: participatory action research
PTSD: posttraumatic stress disorder
RQ: research question

Edited by A Mavragani; submitted 19.09.24; peer-reviewed by S Mangelsdorf, C Baxter; comments to author 24.02.25; revised version received 25.06.25; accepted 02.07.25; published 07.10.25.

Please cite as:
Riisager LG, Larsen JE, Huniche L, Christiansen TB, Moeller SB
Collaborative Development of a Self-Tracking Assisted Psychotherapy Treatment Concept for Refugees With Complex Posttraumatic Stress Disorder: Participatory Action Research
JMIR Form Res 2025;9:e66663
URL: https://formative.jmir.org/2025/1/e66663
doi:10.2196/66663
PMID:

©Lisa Groenberg Riisager, Jakob Eg Larsen, Lotte Huniche, Thomas Blomseth Christiansen, Stine Bjerrum Moeller. Originally published in JMIR Formative Research (https://formative.jmir.org), 07.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted

XSL•FO
RenderX

use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Alongside Digital Wellness Program for Youth: Longitudinal Pre-Post Outcomes Study

Katherine Cohen[1*], MA; Andy Rapoport[1*], MPH; Elsa Friis[2], PhD; Shannon Hill[1], MPH; Sergey Feldman[2], PhD; Jessica Schleider[1], PhD

[1]Department of Medical Social Sciences, Feinberg School of Medicine, Northwestern University, Chicago, IL, United States

[2]Alongside Care, Seattle, WA, United States

[*]these authors contributed equally

Corresponding Author:
Katherine Cohen, MA
Department of Medical Social Sciences
Feinberg School of Medicine
Northwestern University
625 N Michigan Ave
21st Floor
Chicago, IL, 60611
United States
Phone: 1 8032575808
Email: katie.cohen@northwestern.edu

Abstract

Background: Youth are increasingly experiencing psychological distress. Schools are ideal settings for disseminating mental health support, but they are often insufficiently resourced to do so. Digital mental health tools represent a unique avenue to address this gap. The Alongside digital program is one such tool, intended as a universal prevention and early intervention. The platform includes social-emotional learning and self-help wellness features as well as an artificial intelligence–powered chatbot designed to build coping skills.

Objective: This evaluation aimed to examine the near-term impact of Alongside app use on students' self-reported mental health outcomes.

Methods: We conducted a nonrandomized pilot pragmatic evaluation leveraging anonymized user data. All data came from current Alongside users attending public middle and high schools in Texas and New Mexico, between 10 and 18 years old. Schools were actively engaged in partnership with Alongside and approved all procedures. Users were asked to complete mental health questionnaires upon app registration and at 1 and 3 months post registration. We conducted preregistered analyses as well as exploratory analyses to determine how symptoms changed over time and what factors (eg, demographic and app use) predicted changes in distress.

Results: Analyses revealed statistically significant within-person decreases in overall distress (Young Person's CORE; primary outcome) from baseline to 1 month with a small effect size (t_{42}=2.21, P=.03, r=0.34); however, there was no evidence that scores significantly decreased from baseline to 3 months (W=1821, n=85, P=.16). We found that at 3 months, identifying as part of the lesbian, gay, bisexual, transgender, queer, and questioning community predicted greater decreases in distress; otherwise, no demographic factors were significant predictors. In a nonregistered exploratory analysis of a subsample of users who reported elevated distress at baseline, decreases in distress were seen at both 1 month (W=128, n=20, P=.02, r=0.52) and 3 months (W=682, n=42, P=.004, r=0.45).

Conclusions: There may be short-term benefits associated with using the Alongside digital program. Further studies are required to determine potential preventative effects.

(JMIR Form Res 2025;9:e73180) doi:10.2196/73180

KEYWORDS

digital mental health; schools; youth; Alongside; LGBTQ

XSL•FO
RenderX

Introduction

Children and adolescents are increasingly experiencing psychological distress and mental health concerns; from 2009 to 2019, the proportion of youth in the United States who reported persistent feelings of sadness or hopelessness rose from 26% to 37% [1]. The COVID-19 pandemic exacerbated already concerning mental health outcomes in youth, with reports suggesting that minoritized youth (eg, homeless youth, LGBTQ [lesbian, gay, bisexual, transgender, queer, and questioning] youth, and youth of Color) experienced a disproportionately negative effect of pandemic stressors [2,3]. Unfortunately, access to needed mental health care for youth is limited by both structural (eg, financial) and attitudinal barriers (eg, stigma) [4]. Further, there are increasing calls to approach the youth mental health crisis from a population health perspective by teaching positive coping skills to all youth, which may serve to prevent subclinical distress from progressing to the point of needing intervention [5,6]. To address the rising rates of distress in youth, multilevel solutions are needed that can universally promote wellness while also linking students with additional needs to supports that can be accessed with minimal barriers or delays.

One of the most common settings where youth attempt to access mental health care is schools [7], likely due to schools being a unique setting in which certain barriers to access (eg, cost and location) are comparatively minimal. Additionally, schools are an ideal setting for universal prevention and early intervention efforts, as they are one of the only settings where the majority of a population (ie, youth) are found in one place, and the already educational nature of schools is consistent with providing psychoeducation surrounding mental well-being [8,9]. Students, particularly those who are traditionally underserved, are more likely to receive mental health support and wellness resources in schools independent of structural and policy challenges they may face externally, in their homes and communities [10]. Schools are overwhelmed by mental health needs among their students, however, and are often incapable of meeting the needs of all students who seek services or could benefit from preventative efforts [11,12]. Although some support and funding were made available during the pandemic to schools, those are likely not continuing [13]. To prevent worsening of outcomes and provide support as quickly as possible, schools need cost-effective solutions that do not strain an already overstretched system, while filling gaps that existing support cannot.

Digital mental health tools and interventions may be viable options for support that can add to an array of school-based services. Digital technology is widely used by young adults [14]. Many youth already use digital tools in efforts to aid their mental health and everyday distress, with more than 80% using an internet search to find mental health information [15]. Digital resources offer the opportunity for 24/7 support, meaning they can be accessed during acute moments of distress. Further, many students may be hesitant to ask for in-person support due to concerns around stigma; alternatively, they may wish to address issues themselves, as youth often report the desire for self-autonomy [4]. Critically, digital tools cannot and should not replace other school-based mental health supports, such as services delivered by a licensed school-based provider. Rather, they can be cost-effective solutions to deliver psychoeducational and skill-based content for universal or selective prevention. Considering the multitiered systems of support model, a widely used framework for categorizing and implementing school-based mental health supports, digital tools may be best used as “Tier 1” supports for all students, or “Tier 2” supports for students identified as at-risk for developing mental health disorders. Such tools need to be non-stigmatizing and promote student enrollment and participation in a way that is developmentally appropriate, flexible to reach them at the times and in the ways in which they need support, in a student’s native language, and fundamentally facilitate their sense of autonomy in promoting well-being.

The Alongside digital program was designed by experts in digital mental health, ed-tech, and education to leverage the benefits of digital mental health solutions and provide schools with a scientifically grounded and cost-effective method for bolstering student well-being without overburdening staff. Alongside aims to provide personalized social-emotional learning and self-help wellness support to students in 4th through 12th grades across the country and identify students in need of higher levels of support. Alongside includes features such as psychoeducational and mindfulness resources, a 24/7 chatbot that allows students to engage in personalized skill building, and student-initiated onramps to seek additional support. Our objective during this pilot pragmatic evaluation was to leverage anonymized user data to examine the near-term impact of the Alongside app use on student self-reported clinical and well-being outcomes. We tested the following preregistered hypotheses:

- Hypothesis 1: we hypothesized that completion of at least one chat within the app (outside an initial onboarding session) would be associated with significant decreases in overall distress, depression, and anxiety symptoms, as well as hopelessness and loneliness, and significant increases in mental health treatment expectancies at 1 month and 3 months post app registration. Overall distress was our primary outcome, while other measures were preregistered as secondary outcomes.
- Hypothesis 2: we hypothesized that users would show similar decreases in overall distress across sociodemographic groups.
- Hypothesis 3: we hypothesized that greater engagement in the app (ie, multiple chat sessions) would predict larger improvements in overall distress among users.

We also tested hypotheses that we did not preregister as exploratory: (1) we tested each hypothesis separately for users with elevated symptoms at baseline; (2) we tested whether users who showed clinically significant decreases in distress differed in their app use from users who did not show clinically significant decreases in distress; and (3) we tested whether LGBTQ users differed in their app use or baseline distress from non-LGBTQ users. We were interested in the potential differences between LGBTQ users and non-LGBTQ users for two reasons: (1) LGBTQ youth are more likely to experience mental health concerns such as depression and therefore may

XSL•FO
RenderX

differentially benefit from receiving mental health support [16]; (2) LGBTQ youth often face barriers receiving traditional care (eg, reluctance to disclose identity to parents and difficulty finding a culturally competent provider) and may be more likely to benefit from digital, self-help resources [17]. Exploratory analyses are described in more detail subsequently in the methods section.

Methods

Sample

This paper reports results from a nonrandomized, open, pragmatic evaluation using deidentified data. All participants were current Alongside users enrolled in public middle and high schools in Texas and New Mexico. Users were between the ages of 10 and 18 years. Upon app registration, users completed a set of questionnaires that asked about their mental health symptoms and help-seeking expectancies (additional details provided below). Approximately 1 month and 3 months after app registration, users were asked to complete the questionnaires again via an in-app notification. All questionnaires were available in Spanish and English. Participants were given a month to complete the questionnaires. Teachers and counselors at schools were allowed to answer student questions about the questionnaires or define words. There were no upper limits on sample size. Additionally, Alongside collected app use data (eg, which features of the app users engaged in, and how frequently) throughout the 3-month period. All data (ie, questionnaire data and usage data) were collected internally by Alongside as part of their ongoing efforts to evaluate and improve their program. The deidentified, user-level data was provided to researchers at Northwestern University for users who met the following criteria: (1) completed at least one chat within the app outside of the initial rollout session (ie, onboarding session) and (2) consented to allow their data to be used for research purposes. As all data were part of a completely anonymous evaluation, this analysis was deemed as nonhuman subjects research in consultation with the institutional review board at Northwestern University. This paper adheres to reporting standards for secondary data analyses [18].

Alongside

Overview

The Alongside app contains multiple features designed to provide personalized support for mental health. An initial prototype was developed in June 2022 based on over 50 in-depth interviews with students and school stakeholders. The first version of Alongside was piloted across the country during the 2022-2023 school year. Data from pilots and a 3-month collaborative codevelopment process with teens refined Alongside's theory of change and features.

At the time of this evaluation, Alongside provided personalized skill building through matching a mix of rule-based and highly regulated generative artificial intelligence (AI) chats grounded in Alongside's skill-building framework called EMPOWER: Engage (provide empathy and validation), Motivate (help the user understand how a skill helps them), Practical examples (learn through practical examples), Operationalize (apply the skill to their own life), Work on it (put the skill into action in real life), Evaluate (follow-up on how it goes), and Reinforce (celebrate progress and effort, not just the outcome). The EMPOWER framework was developed using concepts from motivational interviewing, common factors theory, and self-determination theory and was informed by the qualitative review and coding of over 5000 chats. Skills encompassed multiple theoretical approaches, including 3rd wave CBTs, including DBT, ACT, Mindfulness-based approaches, narrative therapy, social skills training, positive psychology, and solution-focused therapy. While the EMPOWER framework is scientifically grounded, development is ongoing, and the effectiveness of the framework has yet to be rigorously evaluated. Content covered within chats includes mental health coping, academic success, relationship skills, self-awareness, social awareness, responsible decision-making, and self-management. Several of these skill domains map onto social-emotional learning domains identified by CASEL. Chats are meant to be stand-alone supports, such that any one chat session is intended as a self-contained coping resource. Chats could be conceptualized as single-session interventions, which are theoretically grounded in the idea that meaningful insights can occur at any moment and change is possible within one encounter [19]. Therefore, chats are not only available once but are available one at a time.

In the version of the app that was implemented when the data from this evaluation was collected (between January and June 2024), the chatbot included over 120 rule-based and generative AI chat "modules" that aligned with over 50 common challenges students reported in pilot data (ie, building healthy friendships, boosting self-esteem, reducing anxiety). A screenshot with an example chat is presented in Figure 1. Alongside is a member of the EDSAFE AI Industry Council and takes a multifactor approach to ensuring the safe use of generative AI. At the time of this evaluation, generative AI was used to provide validating statements to students, such as "I understand it must be hard to feel so alone" to add personalization. Users could engage with chats in Alongside in over 30 languages. Students are aware that all engagement on Alongside is confidential unless they express a risk of harm to themselves, someone else, or report abuse or harassment. Below, we include additional information on Alongside's procedures in these instances. Outside of severe issues, students have autonomy to send a message or share a summary of their chat with their school counselor. Other activities available in the app include psychoeducational videos (eg, "The Power of Speaking Up: How Asking for Help Changed My Life"), a goal-setting tool, and a journaling tool. A mood-tracking feature was launched in February 2024. Additional user interface/user experience and content changes occurred during the evaluation period, including the creation of additional chat modules.

XSL•FO
RenderX

Figure 1. Example chat within the Alongside platform.

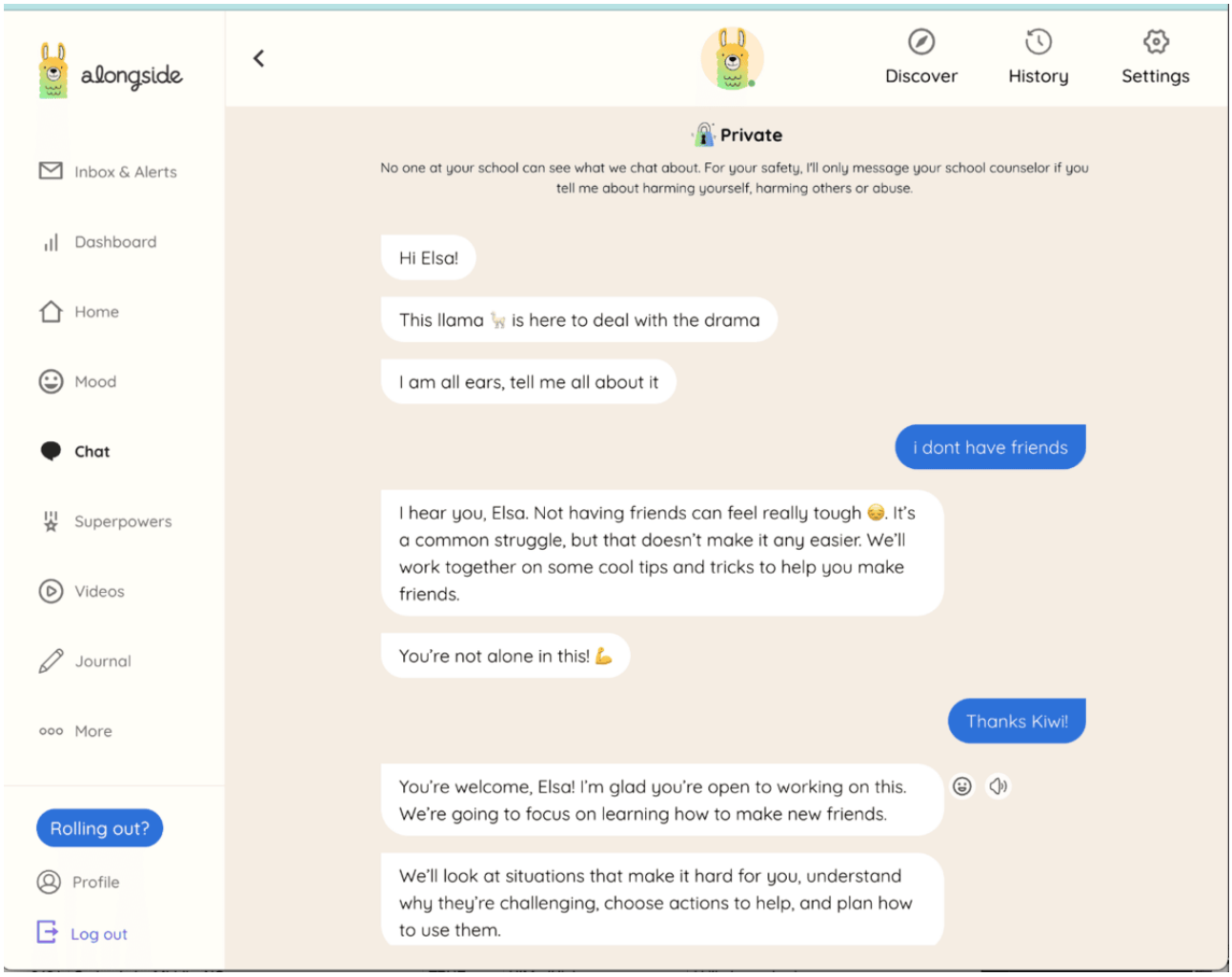

Risk Procedures

Alongside uses a proprietary machine learning safety monitoring system to screen all student free text inputs and identify any mention on self-harm, suicidal ideation, homicidal ideation, abuse or harassment. If potential risk is detected, the bot asks, "I am concerned you are at risk of serious harm, did I get that right?" The student either selects yes or no. If the student selects yes, they are asked, "Can you confirm what is happening?" Options include "thinking of hurting myself, "thinking of hurting someone else," "worried about someone hurting me," or "someone else needs help." If students say no, one of Alongside's clinical team members reviews the chat within 24 hours to ensure this was in fact a false trigger. If a student confirms a severe issue, the bot administers a self-reported severity measure such as the Columbia Suicide Severity Rating Scale, directs the student to call a 24/7 hotline such as 988, and walks the student through creating a safety plan based on their specific situation. Concurrently, an automated email and text message is generated to notify school staff of the concerning message and information gathered in the risk screening process. In addition, schools have access to a dashboard with the above information, allowing schools to easily track or refer back to important safety information. In addition to these standard procedures, if a student responded positively to "I have thought about harming myself" on the Young Person's CORE (YP-CORE) (as described below), the Alongside program automatically matched their user ID to their name/email and generated a real-time notification to the school reporting the student name, email, and response on the survey so the counseling team could follow up and provide support. School student wellness staff have access to a dashboard including deidentified data on app use, common challenges or topics students are discussing, as well as information on students with severe issues or who choose to reach out.

All data on Alongside is protected via industry-standard SSL-encryption. Deidentified data was transferred to Northwestern via a secure file transfer. A full explanation of Alongside's data security procedures can be found on the Alongside website. Northwestern stored the data in a secure, encrypted server.

Recruitment

During the fall of 2024, Alongside approached 4 schools that had purchased Alongside, who represented the typical school that partners with Alongside to share deidentified data for research purposes prior to rolling out the platform to students. One additional school was recruited from a district that had previously implemented Alongside in multiple schools and was open to providing Alongside to a school that had not previously used it in order to ensure there was representation from an Urban

XSL•FO
RenderX

district. All 5 middle and high schools agreed to partner with Alongside to provide the program to their students and share deidentified data for research purposes. Table 1 provides general information about each school as reported by the National Center for Education Statistics [20].

Table 1. Characteristics of the 5 middle and high schools that agreed to provide Alongside to students and share deidentified data.

School	Location	Type	Grades	Size	Urban/Rural	Percentage free or reduced lunch	Diversity[a]
School 1	New Mexico	Public	9th-12th	621	Remote town	100%	76% BIPOC[b]
School 2	Texas	Public	9th-12th	1157	Distant town	81%	90% BIPOC
School 3	Texas	Public	6th-8th	750	Suburban	86%	99% BIPOC
School 4	Texas	Public	6th-7th	561	Distant town	74%	90% BIPOC
School 5	Texas	Public	6th-8th	772	Large city	93%	94% BIPOC

[a]Percentage BIPOC is the percentage enrollment by non-White race/ethnicity, that is, the percentage enrollment of American Indian/Alaska Native, Asian, Black, Hispanic, Native Hawaiian/Pacific Islander, and 2 or more races.

[b]BIPOC: Black, Indigenous, and People of Color.

This study is a naturalistic implementation study and thus all procedures followed the typical roll-out procedures, Alongside uses across all schools. To implement Alongside, school counseling teams must complete a 45-minute orientation. Additionally, schools must determine parental consent/assent procedures prior to implementing Alongside. Each district has a protocol, largely determined by their superintendent, regarding whether parents must give active consent for their child to use Alongside, or whether parents may choose to opt out of providing consent for their child to use Alongside. All schools included in the current evaluation opted for a parental opt-out procedure for using Alongside and sharing deidentified data for research purposes. School 3 made the decision to exclude students younger than 13 years from using the app. All other schools made the decision that this process was sufficient per their school/district/state regulations. Letters were sent by the schools to all parents via their typical communication method (eg, email and Canvas) to provide information about Alongside, including a statement that deidentified data would be used for research purposes. Letters were sent in both English and Spanish. Parents were given the opportunity to opt their child out of accessing the Alongside program or opt out of deidentified data sharing at any time. No parents opted out of either.

Schools or districts that had active partnership agreements with Alongside but had not yet begun to implement the program with students by January 2024 were invited to participate in the research. All schools were recommended to roll out Alongside via a 20-minute teacher-led roll-out activity during class. Schools are provided with posters and cardboard cutouts to promote awareness of Alongside. Additionally, school counseling and wellness staff receive a monthly newsletter with topical suggestions. For example, in February, schools received a newsletter with recommended journal prompts, videos, and chat starters focusing on healthy dating relationships and teen dating violence prevention.

Ethical Considerations

Alongside adheres to all state and school district regulations concerning the use of AI in their product, topics that they cover, and parental consent procedures. Quantitative measures used in this study aligned with school universal screening procedures which is a service Alongside provides, therefore additional permission was sought to share deidentified data for research purposes. Parents were able to opt their children out of the Alongside platform for sharing deidentified data at any time. In addition, students were asked to provide consent to share their deidentified data with Northwestern University. Across all schools, 55.4% (1356/2448) of students consented to share deidentified data. Compensation was not provided to participants. All data was deidentified prior to being shared with the external Northwestern research team. As all data were part of a completely anonymous evaluation, this analysis was deemed as nonhuman subjects' research in consultation with the institutional review board at Northwestern University (STU00220768). There is no identification of participants in the text of this manuscript, figures/tables, images, or any related materials.

Measures

Demographics

Upon app registration, users were asked to report their age, sex (female, male, and prefer not to say), race (American Indian or Alaska Native, Asian, Black or African American, Hispanic or Latino, Native Hawaiian or Other Pacific Islander, or White) and whether they identified as someone who was part of the LGBTQ community (yes or no).

The YP-CORE

The YP-CORE was our preregistered primary outcome. This 10-item scale asks participants to indicate how often they experience symptoms of distress (eg, “I’ve felt edgy or nervous”) on a 5-point Likert scale ranging from 0 (not at all) to 4 (most or all of the time). Scores are summed and range from 0 to 40, with higher scores indicating greater levels of distress [21]. Among the users who completed the baseline and 1 month questionnaires, internal consistency (α) was .80 at baseline and .80 at 1 month. Among the users who completed the baseline and 3-month questionnaires, internal consistency (α) was .79 at baseline and .79 at 3 months.

XSL•FO
RenderX

Patient Health Questionnaire-2

The Patient Health Questionnaire-2 (PHQ-2) asks participants how often they've experienced symptoms of depression (eg, "Little interest or pleasure in doing things") over the last 2 weeks on a 4-point Likert scale ranging from 0 (not at all) to 3 (nearly every day). Scores are summed and range from 0 to 6, with higher scores indicating greater levels of depression [22]. Among the users who completed the baseline and 1 month questionnaires, internal consistency (α) was .69 at baseline and .72 at 1 month. Among the users who completed the baseline and 3-month questionnaires, internal consistency (α) was .69 at baseline and .74 at 3 months.

Generalized Anxiety Disorder Questionnaire-2

The Generalized Anxiety Disorder Questionnaire-2 (GAD-2) asks participants how often they've experienced symptoms of anxiety (eg, "Feeling nervous, anxious or on edge") over the last 2 weeks on a 4-point Likert scale ranging from 0 (not at all) to 3 (nearly every day). Scores are summed and range from 0 to 6, with higher scores indicating greater levels of anxiety [23]. Among the users who completed the baseline and 1 month questionnaires, internal consistency (α) was .87 at baseline and .67 at 1 month. Among the users who completed the baseline and 3-month questionnaires, internal consistency (α) was .81 at baseline and .81 at 3 months.

Beck Hopelessness Scale-4

The Beck Hopelessness Scale-4 (BHS-4) asks participants to indicate their current level of hopelessness through rating their agreement with statements (eg, "My future seems dark to me") on a 4-point Likert scale from 0 (absolutely disagree) to 3 (absolutely agree). Scores are summed and range from 0 to 12, with higher scores indicating greater levels of hopelessness [24]. Among the users who completed the baseline and 1 month questionnaires, internal consistency (α) was .83 at baseline and .78 at 1 month. Among the users who completed the baseline and 3-month questionnaires, internal consistency (α) was .84 at baseline and .86 at 3 months.

The University of California, Los Angeles Loneliness Scale-3

The UCLA (University of California, Los Angeles) Loneliness Scale-3 (ULS-3) asks participants to indicate how often statements reflecting loneliness (eg, "I feel left out") are descriptive of them on a 4-point Likert scale from 1 (never) to 4 (often). Scores are summed and range from 4 to 16, with higher scores indicating greater levels of loneliness [25]. Among the users who completed the baseline and 1 month questionnaires, internal consistency (α) was .77 at baseline and .80 at 1 month. Among the users who completed the baseline and 3-month questionnaires, internal consistency (α) was .70 at baseline and .82 at 3 months.

Expectancies

Participants' expectancies regarding mental health treatment were assessed using one item: "On a scale from 1 (not at all helpful) to 10 (extremely helpful), how helpful do you think individual therapy would be in reducing mental health problems, if you struggle/were to struggle with them?" [26].

App Usage Data

Alongside collected data over the 3-month period on the following usage metrics per-user: number of app opens, number of chat sessions started, number of chat sessions completed, number of chat messages sent, number of goals created, number of journal entries, number of videos watched, number of mood log entries, and total number of activities completed.

Data Analysis

The RStudio Statistical Program (R Foundation for Statistical Computing)) was used to complete data analyses [27]. For each statistical test, a P value of <.05 was considered statistically significant. Descriptive statistics for participant characteristics (eg, sex) as well as usage metrics (eg, number of chats completed) were reported in terms of means, SDs, and percentages. Cronbach α was calculated to evaluate the internal consistency of each quantitative measure at each time point. We used list wise deletion to exclude users who did not have complete data available for outcomes of interest. We additionally conducted either paired t tests or Wilcoxon signed-rank tests to determine whether demographic or clinical characteristics at baseline differed between participants who completed the 1 month assessment and participants who completed the 3-month assessment.

Hypothesis 1

To test hypothesis 1, we conducted either paired t tests or Wilcoxon signed-rank tests. We first tested the assumption of normality (via the Shapiro-Wilk normality test). If the assumption was met, we conducted a paired t test. If the assumption was not met, we conducted a Wilcoxon signed-rank test. We additionally calculated Cohen d effect sizes (with 95% CIs). We ran one set of tests comparing outcomes from baseline to the 1-month timepoint, and another set of tests comparing outcomes from baseline to the 3-month timepoint.

Hypotheses 2 and 3

To test hypotheses 2 and 3, we conducted multiple linear regressions. For hypothesis 2, the outcome variable was YP-CORE scores at the follow-up time-point, and predictor variables included baseline YP-CORE scores and demographic variables (age, race/ethnicity, sex, school, and whether or not they identified as a member of the LGBTQ community). We ran one model with YP-CORE scores at the 1-month timepoint as the outcome variable and one model with YP-CORE scores at the 3-month timepoint as the outcome variable. Race/ethnicity, sex, and LGBTQ status were dummy coded, with "White/Caucasian," "male," and "No" as the reference groups. School was dummy coded with School 1 as the reference group. For hypothesis 3, the outcome variable was YP-CORE scores at the follow-up timepoint, and predictor variables included baseline YP-CORE scores and user engagement variables. Each user engagement variable was tested individually; therefore, 10 models were run (total number of activity sessions, total number of videos watched, total number of chat sessions, total number of journal entries, and total number of goals set; 5 models for 1-month outcomes and 5 models for 3-month outcomes).

XSL•FO
RenderX

Additional Analyses

We additionally tested all hypotheses on a subsample of users who had clinically elevated distress at baseline (defined as YP-CORE of ≥14.1) [28]. We conducted tests to determine whether users who saw a clinically significant decrease in distress (defined as a decrease of ≥ 7.9) [27] differed in their usage data from users who did not see a clinically significant decrease in distress. We conducted tests to determine whether LGBTQ users differed in their usage data from non-LGBTQ users.

Preregistration

This analysis was preregistered on Open Science Framework [29]. There were several deviations from the preregistered analyses. We preregistered that we would run longitudinal analyses including all 3 timepoints (baseline, 1 month, and 3 months). However, we did not have sufficient data to run these analyses. Instead, we analyzed outcomes from baseline to 1 month, and baseline to 3 months. Additionally, we preregistered that we would analyze whether grade predicted changes in overall distress (hypothesis 2) and whether total number of minutes spent in the app predicted changes in overall distress (hypothesis 3). However, these data were unavailable. We did not preregister but chose to run the following exploratory analyses: (1) testing hypotheses separately for users with elevated symptoms at baseline; (2) testing whether users who showed clinically significant decreases in distress differed in their app use from users who did not show clinically significant decreases in distress; and (3) testing whether LGBTQ users differed from non-LGBTQ users in their baseline YP-CORE scores or app use metrics. The decision to add this set of analyses was made in light of emerging questions regarding who the app is most effective for, and which components drove effectiveness. Particularly, the decision to test whether LGBTQ users differed in their app use from non-LGBTQ users was made due to findings from hypothesis 2 as discussed in the results. Last, we preregistered that we would exclude participants younger than 12 years old, however the sample size was smaller than we anticipated. To increase the available sample size and thereby the likelihood of detecting true effects, we elected to include participants who were at least 10 years old.

Results

Participant Retention

Across the 5 schools, there were 3419 students who were eligible to use Alongside. There was a total of 3861 students at schools 1-5 (Table 1); however, school 3 excluded students aged 13 years or younger from this study, which removed 442 students from eligibility. A total of 2804 students signed into Alongside. Among these, 1356 consented to sharing deidentified data for research purposes. There were students who enrolled in the study on mobile, and the survey was not available on mobile; therefore, those students did not have the chance to participate. A total of 1114 students who consented to share deidentified data completed the baseline questionnaires upon registration; 301 (27%) of these students completed at least one chat outside of the onboarding session. A total of 199 students who consented to share deidentified data completed the 1-month questionnaires; 66 (33%) of these students completed at least one chat outside of the onboarding session. A total of 300 students who consented to share deidentified data completed the 3-month questionnaires; 116 (38%) of these students completed at least one chat outside of the onboarding session.

Participant Characteristics

A total of 66 users completed both the baseline and the 1-month questionnaires. A total of 116 users completed both the baseline and the 3-month questionnaires. A total of 26 users completed questionnaires at all 3 time points. Among users who completed both the baseline and the 1-month questionnaires, 97.01% (*n*=65) chose to complete the questionnaires in English. They ranged in age from 10 to 18 (mean 14.02, SD 2.05) years old. The majority were female (32/61, 52.46). The most commonly self-identified race/ethnicity was Hispanic or Latino (39/59, 66.10%). A total of 11 (18.33%) users identified as part of the LGBTQ community. Among users who completed both the baseline and the 3-month questionnaires, 91.45% (*n*=107) chose to complete the questionnaires in English. They ranged in age from 11 to 18 (mean 13.83, SD 2.17) years old. The majority were female (58/111, 52.25%). The most commonly self-identified race/ethnicity was Hispanic or Latino (78/107, 72.89%). A total of 15 (14.23%) users identified as part of the LGBTQ community. Table 2 includes detailed information on users' demographic characteristics.

XSL•FO
RenderX

Table 2. Demographic characteristics of users at each time point.

	Completed baseline and midline, n (%)	Completed baseline and endline, n (%)	Completed baseline, midline, and endline, n (%)
Language			
Spanish	2 (3.0)	10 (8.5)	1 (3.8)
English	65 (97.0)	107 (91.5)	25 (96.2)
Sex			
Female	32 (52.5)	58 (52.3)	10 (43.5)
Male	27 (44.3)	49 (44.1)	12 (52.2)
Prefer not to state	2 (3.3)	4 (3.6)	1 (4.3)
Race/Ethnicity			
American Indian or Alaska Native	2 (3.4)	1 (0.9)	1 (4.3)
Black or African American	5 (8.5)	16 (15.0)	1 (4.3)
White	12 (20.3)	7 (6.5)	5 (21.7)
Asian	1 (1.7)	4 (3.7)	1 (4.3)
Native Hawaiian or Other Pacific Islander	0 (0)	1 (0.9)	0 (0)
Hispanic or Latino	39 (66.1)	78 (72.9)	15 (65.2)
LGBTQ[a]			
Yes	11 (18.3)	15 (14.3)	5 (20.8)

[a]LGBTQ: lesbian, gay, bisexual, transgender, queer, and questioning.

Results from Wilcoxon and chi-square tests suggested that there were no significant differences in baseline demographic characteristics (age, sex, race/ethnicity, and LGBTQ status) between the users who completed both baseline and 1-month questionnaires and users who completed both baseline and 3-month questionnaires. Additionally, there was no evidence that the 2 groups differed in their baseline scores on the YP-CORE, PHQ-2, GAD-2, BHS-4, or ULS-3. There were 2 key differences between the 2 samples. (1) Users who completed baseline and 1-month assessments had significantly lower expectancies scores at baseline (mean 5.62, SD 2.92) compared with users who completed baseline and 3 month assessments (mean 6.71, SD 2.58, W=2572.5, P=.02); (2) there were significantly different proportions of students in each school who completed the 1-month versus the 3-month assessments (χ^2_4=50.64, P<.01). Figure 2 displays the different proportions of students from each school that are represented in the 2 samples.

XSL•FO
RenderX

Figure 2. Proportion of the sample that completed 1-month and 3-month surveys from each of the 5 schools.

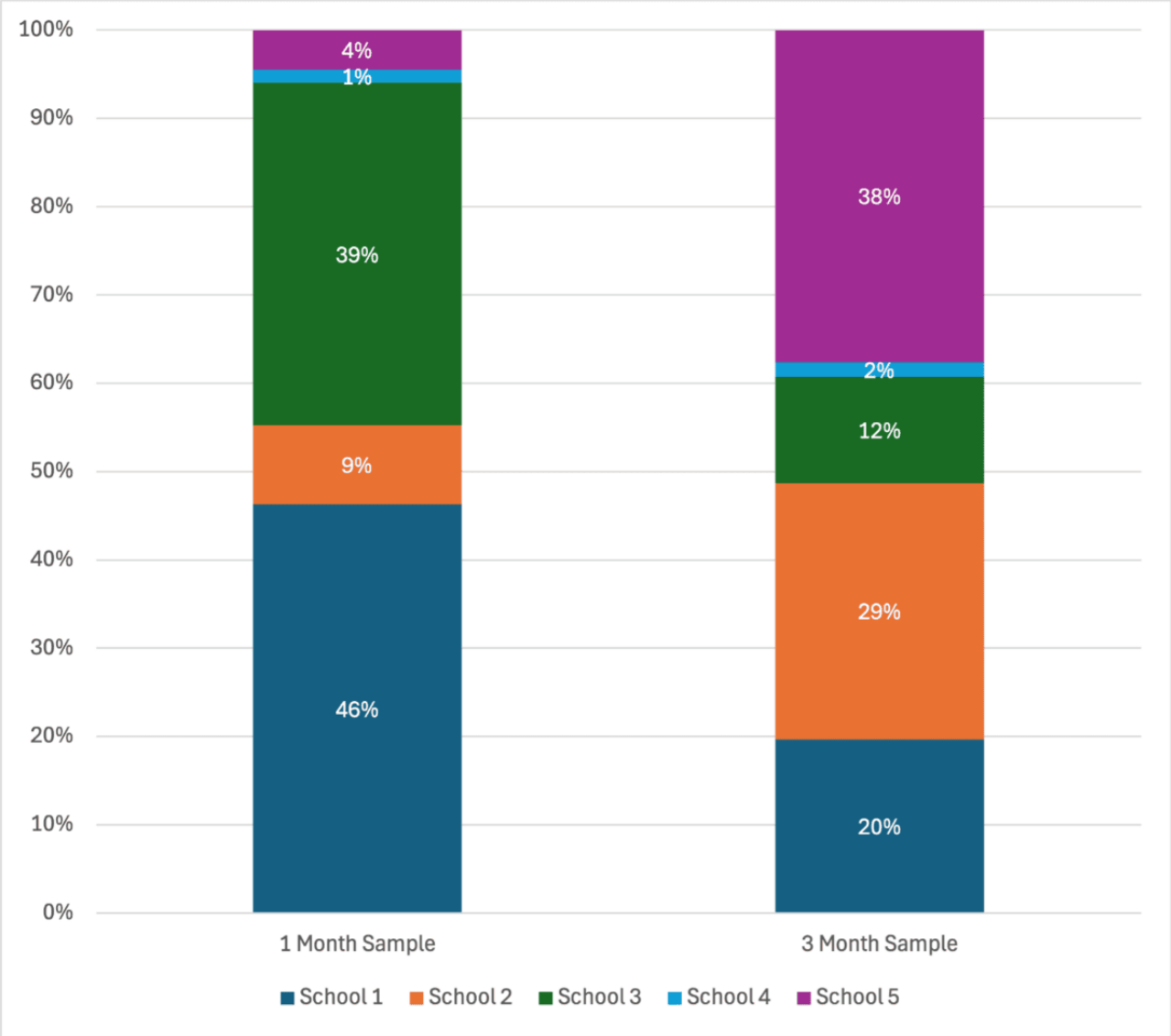

We additionally tested whether baseline YP-CORE scores differed between schools. Baseline YP-CORE scores were significantly different across schools among users who completed both baseline and 1-month questionnaires (F_1=8.41, P=.006), but not among users who completed both baseline and 3-month questionnaires (F_1=0.56, P=.46). Table 3 portrays baseline YP-CORE scores across schools for both samples.

Table 3. Example Young Person's CORE (YP-CORE) scores at baseline per school, for those who completed the 1- and 3-month surveys.

School	1-month data, mean (SD)	3-month data, mean (SD)
School 1	11.96 (6.59)	14.05 (7.53)
School 2	17.00 (10.39)	13.17 (6.48)
School 3	16.61 (6.16)	17.50 (7.79)
School 4	25.00 (N/A[a])	11.50 (4.95)
School 5	20.33 (5.86)	16.19 (8.45)

[a]N/A: not applicable.

Usage Metrics

Table 4 includes data on users' app use metrics. Results among the sample of users who completed both baseline and 3-month assessments encompass the entire 3-month period (not the period between 1 month and 3 months).

XSL•FO
RenderX

Table 4. Usage metrics across time points.

Usage metric	1-month data, mean (SD)	3-month data, mean (SD)
App opens	2.85 (2.85)	4.19 (4.44)
Chat sessions	3.09 (3.15)	3.33 (3.72)
Goals created	0.87 (1.48)	1.14 (1.85)
Journals	0.42 (1.09)	0.29 (0.92)
Videos	0.21 (0.67)	0.40 (1.88)
Mood log	0.12 (0.56)	0.17 (0.65)
Time spent chatting (seconds)	988.5 (1211.66)	1103 (1392.89)
Total activities	4.70 (5.13)	5.33 (6.99)
Chat messages sent	44.45 (44.42)	48.19 (54.63)
Chats completed	1.43 (1.21)	1.49 (1.69)

Hypothesis 1: Changes in Outcomes Across 3 Months

A paired *t* test showed statistically significant within-person decreases in YP-CORE scores from baseline (*n*=43, mean 14.33, SD 7.43) to 1 month (*n*=43, mean 12.42, SD 7.36; t_{42}=2.21, *P*=.03) with a small effect size (*r*=0.34). However, a Wilcoxon signed-rank test did not show statistically significant within-person decreases in YP-CORE scores from baseline (*n*=85, mean 14.67, SD 7.75) to 3 months (*n*=85, mean 14.05, SD 8.06; *W*=1821, *P*=.16).

Wilcoxon signed-rank tests did not show statistically significant within-person changes in PHQ-2 scores from baseline (*n*=50, mean 1.42, SD 1.42) to 1 month (*n*=50, mean 1.2, SD 1.55; *W*=264, *P*=.3) or from baseline (*n*=98, mean 1.5, SD 1.61) to 3 months (*n*=98, mean 1.37, SD 1.65; *W*=772, *P*=.29).

Wilcoxon signed-rank tests did not show statistically significant within-person changes in GAD-2 scores from baseline (*n*=53, mean 1.39, SD 1.63) to 1 month (*n*=53, mean 1.17, SD 1.44; *W*=220, *P*=.24) or from baseline (*n*=102, mean 1.69, SD 1.76) to 3 months (*n*=102, mean 1.52, SD 1.87; *W*=1174, *P*=.16).

A Wilcoxon signed-rank test did not show statistically significant within-person changes in BHS-4 scores from baseline (*n*=51, mean 2.98, SD 2.82) to 1 month (*n*=51, mean 3.0, SD 2.79; *W*=278, *P*=.75). However, a Wilcoxon signed-rank test showed statistically significant within-person decreases in BHS-4 scores from baseline (*n*=96, mean 3.62, SD 3.28) to 3 months (*n*=96, mean 2.99, SD 3.21; *W*=1530, *P*=.006) with a moderate effect size (*r*=0.32).

Wilcoxon signed-rank tests did not show statistically significant within-person changes in ULS-3 scores from baseline (*n*=49, mean 4.98, SD 1.76) to 1 month (*n*=49, mean 4.82, SD 1.73; *W*=295, *P*=.56) or from baseline (*n*=102, mean 5.19, SD 1.82) to 3 months (*n*=102, mean 4.88, SD 1.89; *W*=1400, *P*=.06).

Wilcoxon signed-rank tests did not show statistically significant within-person changes in expectancies scores from baseline (*n*=56, mean 5.46, SD 2.94) to 1 month (*n*=56, mean 5.79, SD 3.04; *W*=450, *P*=.6) or from baseline (*n*=103, mean 6.72, SD 2.62) to 3 months (*n*=103, mean 6.15, SD 2.99; *W*=1778, *P*=.10).

Hypothesis 2: Demographic Predictors of Change in YP-CORE

Results from linear regressions showed there was no evidence that sociodemographic factors (age, race/ethnicity, LGBTQ identity, sex, and school) significantly predicted changes in YP-CORE scores from baseline to 1 month. There was evidence that identifying as part of the LGBTQ community was a significant predictor of lower distress from baseline to 3 months (t_{57}=–2.07, *P*=.04). There was no evidence that any other sociodemographic factors predicted changes in YP-CORE scores from baseline to 3 months.

Hypothesis 3: App Use Predictors of Change in YP-CORE

Results from linear regressions showed there was no evidence that total number of activity sessions, videos watched, chat sessions, journal entries created, or goals set predicted changes in YP-CORE scores at 1 month or 3 months.

Exploratory Analyses

Elevated Subsample

In a subsample of users who had clinically elevated distress at baseline (defined as YP-CORE of ≥ 14.1) [25], a Wilcoxon signed-rank test showed statistically significant within-person decreases in YP- CORE scores from baseline (*n*=20, mean 20.6, SD 5.17) to 1 month (*n*=20, mean 16.9, SD 5.97; *W*=128, *P*=.02) with a large effect size (*r*=0.52). Additionally, a Wilcoxon signed rank showed statistically significant within-person decreases in YP-CORE scores from baseline (*n*=42, mean 20.95, SD 5.26) to 3 months (*n*=42, mean 17.71, SD 7.51; *W*=682, *P*=.004), with a moderate effect size (*r*=0.45).

A Wilcoxon signed rank showed statistically significant within-person decreases in GAD-2 scores from baseline (*n*=24, mean 2.38, SD 1.84) to 1 month (*n*=24, mean 1.63, SD 1.69; *W*=122, *P*=.03), however there did not appear to be statistically significant within-person changes in GAD-2 scores from baseline (*n*=44, mean 2.64, SD 1.79) to 3 months (*n*=44, mean 2.11, SD 2.03; *W*=398, *P*=.08).

A paired *t* test did not show statistically significant within-person decreases in BHS-4 scores from baseline (*n*=20, mean 4.9, SD

XSL•FO
RenderX

3.01) to 1 month (n=20, mean 4.05, SD 2.96; t_{19}=1.29, P=.21), however a Wilcoxon signed-rank test showed statistically significant within-person decreases in BHS-4 scores from baseline (n=40, mean 5.5, SD 3.39) to 3 months (n=40, mean 4.3, SD 3.45; W=504, P=.007) with a moderate effect size (r=0.32).

Results on other secondary outcomes (PHQ-2, ULS-3, expectancies) were not significant at 1 month or 3 months.

Results from linear regressions showed that, among users with clinically elevated distress at baseline, identifying as part of the LGBTQ community was a significant predictor of lower distress at 1-month (t_6=–3.04, P=.02) and at 3 months (t_{22}=–2.44, P=.02). There was no evidence that any other sociodemographic factors predicted changes in YP-CORE scores at 1 month or 3 months.

There was no evidence that total number of activity sessions, videos watched, chat sessions, journal entries created, or goals set predicted changes in YP-CORE scores at 1 month or 3 months.

Differences in Use Between Users Who Did and Did Not See Clinically Significant Decreases in Distress

We categorized users based on whether they showed clinically significant changes in distress. According to Twigg et al [28] the overall reliable change index value for the YP-CORE is 7.9 (including all age bands and genders). Among participants who completed both the baseline and 1-month YP-CORE, 13.95% (n=6) saw a clinically significant decrease in distress (scores decreased by 7.9 or more), 2.33% (n=1) saw a clinically significant increase in distress (scores increased by 7.9 or more), and 83.72% (n=36) did not see clinically significant changes (changes in scores were < 7.9 and > –7.9). Among participants who completed both the baseline and 3-month YP-CORE, 11.76% (n=10) saw a clinically significant decrease in distress, 10.59% (n=9) saw a clinically significant increase in distress, and 77.65% (n=66) did not see clinically significant changes.

At 1 month, users who saw clinically significant decreases in distress opened the app significantly more (n=6, mean 5, SD 3.22) compared with individuals who did not see clinically significant changes in distress (n=36, mean 2.94, SD 3.14, W=163, P=.04). Users who saw clinically significant decreases in distress also saved their mood more often (n=6, mean 0.83, SD 1.6) compared with individuals who did not see clinically significant changes in distress (n=36, mean 0.06, SD 0.33, W=141, P=.009), spent more seconds chatting (n=6, mean 1463, SD 1121) compared with individuals who did not see clinically significant changes in distress (n=36, mean 1045, SD 1417, W=173, P=.02), and sent more chat messages (n=6, mean 57.2, SD 24.8) compared with individuals who did not see clinically significant changes in distress (n=36, mean 47.8, SD 54.5, W=164, P=.045). No differences emerged for number of chat sessions, goals created, journal entries, videos watched, or chats completed.

At 3 months, users who had a clinically significant decrease in distress saved their mood significantly more (n=10, mean 0.7, SD 1.34) compared with uses who did not have clinically significant changes (n=66, mean 0.12, SD 0.57; $F_{2,\ 82}$=3.56, P=.03). No differences emerged for number of app opens, chat sessions, goals created, journal entries, videos watched, time spent chatting, total activities, chat messages sent, or chats completed.

Differences Between LGBTQ and Non-LGBTQ Users

There was no evidence that LGBTQ users differed in their baseline YP-CORE scores at in either the 1-month (W=136.5, P=.13) or 3-month samples (W=354, P=.16). At 1 month, LGBTQ users opened the app significantly more (n=11, mean 5.27, SD 4.92) compared with non-LGBTQ users (n=49, mean 2.49, SD 2.05; W=148, P=.02) and spent more time chatting in seconds (n=11, mean 1689, SD 1554) compared with non-LGBTQ users (n=49, mean 919, SD 1154; W=155, P=.03). At 1 month, no differences emerged for number of chat sessions, goals created, journal entries, videos watched, mood log entries, chat messages sent, or chats completed. At 3 months, LGBTQ users used the journal tool significantly more (n=15, mean 1.2, SD 2.08) compared with non-LGBTQ users (n=90, mean 0.18, SD 0.51; W=397, P<.01) At 3 months, no differences emerged for number of app opens, chat sessions, goals created, videos watched, mood log entries, time spent chatting, chat messages sent, or chats completed.

Discussion

Principal Findings

In light of the increased budget and resource limitations for student mental health and wellness in schools, innovative solutions are necessary. This paper reports findings from a pragmatic evaluation of Alongside, a novel digital mental health program designed to provide preventative mental health support and identify students who need more support while remaining cost-effective and scalable. We found evidence that overall distress (our primary outcome) decreased over 1 month among all Alongside users who completed at least one chat outside of the initial onboarding session, although decreases were not sustained 3 months post registration. Due to the timing of the 3-month follow-up correlating with end of the year exams, further controlled studies are necessary to disentangle potential timing effects. In a subsample of users who had elevated distress at baseline, however, findings were both stronger and sustained over time. Findings on secondary outcomes related to depression, anxiety, loneliness, and treatment expectancies were largely null, although some positive effects were seen in hopelessness among the full sample and both anxiety and hopelessness among the elevated subsample. While preliminary evidence suggests some positive results, it is critical that we contextualize them alongside null findings and consider areas for future work.

There are multiple reasons why decreases in overall distress may have been seen among the full sample at 1 month but not sustained at 3 months. First, the Alongside program may be most effective in the short-term for in-the-moment distress, but not currently effective for implementing long-term change. This does not negate the meaningfulness of short-term improvements. In school settings where resources are spread thin, it may prove better to provide easily accessible tools that promote short-term improvements, as opposed to no support at all for those who

XSL•FO
RenderX

have sub-clinical levels of distress (ie, those not in need of immediate intervention), which is the common approach currently. Second, a decrease in use of the Alongside program between 1 month and 3 months could explain why users did not see sustained benefits. When comparing app use between the sample that completed the 3-month questionnaires and the sample that completed the 1-month questionnaires, we see marginal differences; for example, the average number of app opens at 3 months (across the entire 3-month period) was 4.19, while the average number of app opens at 1 month was 2.85, a difference of only 1.34. We discuss the unclear relationship between app use and symptom change in more detail later in this discussion, but it is possible that a lack of benefit beyond 1 month could be attributed to low intervention uptake beyond 1 month. Last, there could be discrepancies between the sample of users who completed 1-month questionnaires and users who completed 3-month questionnaires. We found that the proportion of students from each school significantly differed between the 2 samples, and among the users who completed the 1-month questionnaires, distress at baseline significantly differed between users from different schools. At the same time, there was no evidence that baseline distress or demographic characteristics significantly differed between users who completed 1-month questionnaires and users who completed 3-month questionnaires. Still, it is possible that the practical limitations and implementation challenges that Alongside faced in recruiting users to complete questionnaires, discussed in more detail below, contributed to discrepancies between findings at 1 month and 3 months.

Contrary to our hypotheses, we did not see effects on any secondary outcomes related to depression, anxiety, loneliness, or treatment expectancies among the full sample. We caution against interpreting null results in this initial evaluation as a sign to abandon the Alongside program, particularly in light of other potential benefits on short-term distress or benefits among the subsample of users with elevated distress; rather, null findings may reflect that decreases in a largely subclinical sample may not be detectable with the study design. Our findings pointing to stronger effects among the clinically elevated subsample are unsurprising in light of previous research demonstrating larger effect sizes for school-based digital interventions among clinically elevated subsamples of students [30,31], even though Alongside was not specifically developed for this population. At the same time, the results from this evaluation are preliminary and must be interpreted with caution.

Our findings related to Hypothesis 2 suggest that LGBTQ users may see greater decreases in distress than non-LGBTQ users. Otherwise, there were no demographic predictors of changes in overall distress. LGBTQ users may have been more likely to benefit from using the Alongside program because LGBTQ youth often face unique barriers to accessing other mental health care [32]. We also found that LGBTQ users differed from non-LGBTQ users by using several app features more often than non-LGBTQ users. It is possible that the Alongside program is particularly well-suited for dissemination among LGBTQ students. However, these findings are exploratory due to the small sample size of LGBTQ students; replication is necessary to draw strong conclusions. It is also important to note that gender and sexual identity, can be fluid, particularly during adolescence; therefore, these subgroup findings reflect current self-identity. In future research, longitudinal studies can track both intervention effects and identity development over time.

Our findings related to hypothesis 3 suggest that there is no a relationship between users' degree of app use and users' change in overall distress. It is possible that one use may be enough for someone to benefit from, while users who repeatedly come back to a program may do so because they are not getting what they need. Individual differences may better explain the complicated relationship between app use and symptom change. At the same time, we found that users with clinically significant changes in distress used several app features more often than users who did not see clinically significant changes. It is possible that the relationship is not proportional (eg, greater use confers greater impact) but that there is a minimal level of use that can confer some impact, and the degree of impact is individually variable; additional testing is necessary. Alternatively, the strength of the relationship between use metrics and change in distress may exist but may be small enough that our limited sample size was not powered to detect it. In sum, we cannot discount that the degree of program use influences the degree of clinical impact to some extent, but it does not appear to be a particularly meaningful contributor.

Multiple limitations impact our ability to confidently interpret the findings of the current evaluation. First, this was not a randomized evaluation. Without a control group to compare to, we cannot confirm that any positive outcomes were due solely to use of the Alongside program. Simultaneously, we cannot determine whether there were preventative effects such that users who did not see decreases in distress may have otherwise seen increases in distress had they not had access to the program. Second, the primary and secondary outcomes were deficit-focused and did not include well-being measures. The dual-factor model of mental health [33] suggests that well-being and distress are distinct constructs; null findings related to distress do not preclude positive impacts on well-being, however, the current evaluation cannot elucidate the impact of the Alongside program on well-being. Last, Alongside faced a plethora of barriers to implementation that impacted the number of users who responded to questionnaires. One school had over 36% of students screen positive for thoughts of self-harm which contributed to hesitancy around pushing for follow-up assessments as the number of students identified overwhelmed the counseling staff. Additionally, the 3-month endline overlapped with state testing in all Texas schools, leading to difficulty promoting follow-up. Finally, one school asked students to return school devices before the 3-month endline and the survey was not available on personal devices. Only 26 users completed questionnaires at all 3 timepoints, therefore we could not reliably analyze change over time, including all timepoints. The timing of the study in the school year may also have contributed to low numbers of students who responded to questionnaires.

The current evaluation is strengthened by its pragmatic nature and preregistered analysis. The research-to-practice gap is maintained due to a lack of research on real-world

XSL•FO
RenderX

implementation, making findings from highly controlled research studies ungeneralizable to the constantly changing and complicated settings in which the intervention is intended to be deployed. By testing an intervention while it is actively being implemented in a nonresearch setting, it is more likely that findings will be reflective of the ultimate impact an intervention can have. Additionally, this analysis was preregistered, and all deviations from the preregistration are reported. Preregistration addresses the "file drawer" problem within the scientific community, wherein only statistically significant results are shared, and null findings are not published. Preregistration is particularly important when considering the lack of transparency that many commercial digital mental health tools maintain. Digital mental companies largely do not test their products through rigorous research at early stages of development. Transparent testing upfront is necessary for companies to then transparently improve their platforms.

This analysis highlights several opportunities for continued research. This study reports initial, small-scale findings, and replication is required. Additionally, randomized evaluations with large numbers of participants are necessary to reliably parse the effects of the Alongside program. Further evaluations may test future iterations of the Alongside program as it is updated to improve its features or add content that may directly influence secondary outcomes such as depression or anxiety. Additionally, hybrid effectiveness-implementation trials may evaluate the Alongside program as implemented via tier 2 for students who are experiencing early signs of mental distress.

Conclusions

Providing digital mental health tools through schools may help youth to access needed health supports and resources. In this nonrandomized pilot pragmatic evaluation of the Alongside digital mental health tool, we assessed the near-term impact of app use on the self-reported mental health outcomes of middle and high school students. We found that Alongside can confer some short-term benefits, but those benefits are not necessarily sustained in a universal sample. Additionally, the nonrandomized design, attrition, and implementation challenges limit causal inferences and generalizability. Future work may use a larger sample, randomized design, or longitudinal design to further understand when and for whom Alongside may be most effective. While this study takes a valuable step toward understanding digital mental health tools in schools, further validation is necessary.

Acknowledgments

This research was supported by a research grant from Alongside (principal investigator: JS). Alongside independently collected the data that were used in this analysis. Coauthors employed by Alongside (EAF and SF) were involved in determining research questions and writing the manuscript, but were not involved in the analysis. The analysis was independently conducted by Northwestern researchers (KAC, AR, SH, and JS). Outside of this work, Schleider and her research are supported by funding from the National Institutes of Health (DP5OD028123), the National Science Foundation (2141710), the Health Research and Services Administration (U3NHP45406-01-00), the Society of Clinical Child and Adolescent Psychology, Hopelab, the Child Mind Institute, Kooth, Understood, the Montana Department of Public Health and Human Services, the American Psychological Foundation, and the Florida Institute for Child Welfare. KAC is supported by funding from the National Institutes of Health (F31MH139185). The authors attest that there was no use of generative artificial intelligence technology in the generation of text, figures, or other informational content of this manuscript.

Data Availability

The datasets analyzed during this study are not publicly available due to the sensitive nature of the data and per request of the schools participating in the study. Data requests may be submitted to the corresponding author and may be granted pending school partner approval.

Conflicts of Interest

JLS has served on the Scientific Advisory Board for Walden Wise and the Clinical Advisory Board for Koko (unpaid), has received consulting fees from United Health and Woebot, and receives book royalties from New Harbinger; Oxford University Press; and Little, Brown Book Group. She is cofounder and chief scientific advisor for Mindly. No Mindly products were used or are referenced in this manuscript. KAC is a part-time employee at Lyra Health, Inc. EAF is an employee of Alongside. SF is an employee of Alongside.

References

1. Youth risk behavior survey. Center for Disease Control and Prevention. 2019. URL: https://www.cdc.gov/yrbs/dstr/pdf/YRBSDataSummaryTrendsReport2019-508.pdf [accessed 2025-08-29]
2. Fortuna LR, Brown IC, Lewis Woods GG, Porche MV. The impact of COVID-19 on anxiety disorders in youth: coping with stress, worry, and recovering from a pandemic. Child Adolesc Psychiatr Clin N Am 2023;32(3):531-542 [FREE Full text] [doi: 10.1016/j.chc.2023.02.002] [Medline: 37201965]
3. Gruber J, Prinstein MJ, Clark LA, Rottenberg J, Abramowitz JS, Albano AM, et al. Mental health and clinical psychological science in the time of COVID-19: challenges, opportunities, and a call to action. Am Psychol 2021;76(3):409-426 [FREE Full text] [doi: 10.1037/amp0000707] [Medline: 32772538]

XSL•FO
RenderX

4. Gulliver A, Griffiths KM, Christensen H. Perceived barriers and facilitators to mental health help-seeking in young people: a systematic review. BMC Psychiatry 2010;10:113 [FREE Full text] [doi: 10.1186/1471-244X-10-113] [Medline: 21192795]
5. Colizzi M, Lasalvia A, Ruggeri M. Prevention and early intervention in youth mental health: is it time for a multidisciplinary and trans-diagnostic model for care? Int J Ment Health Syst 2020;14:23 [FREE Full text] [doi: 10.1186/s13033-020-00356-9] [Medline: 32226481]
6. Purtle J, Nelson KL, Counts NZ, Yudell M. Population-based approaches to mental health: history, strategies, and evidence. Annu Rev Public Health 2020;41:201-221 [FREE Full text] [doi: 10.1146/annurev-publhealth-040119-094247] [Medline: 31905323]
7. Duong MT, Bruns EJ, Lee K, Cox S, Coifman J, Mayworm A, et al. Rates of mental health service utilization by children and adolescents in schools and other common service settings: a systematic review and meta-analysis. Adm Policy Ment Health 2021;48(3):420-439. [doi: 10.1007/s10488-020-01080-9] [Medline: 32940884]
8. Cohen KA, Ito S, Ahuvia IL, Yang Y, Zhang Y, Renshaw TL, et al. Brief school-based interventions targeting student mental health or well-being: a systematic review and meta-analysis. Clin Child Fam Psychol Rev 2024;27(3):732-806. [doi: 10.1007/s10567-024-00487-2] [Medline: 38884838]
9. Herrenkohl TI. Cross-system collaboration and engagement of the public health model to promote the well-being of children and families. J Soc Soc Work Res 2019;10(3):319-332. [doi: 10.1086/704958]
10. Fazel M, Hoagwood K, Stephan S, Ford T. Mental health interventions in schools 1: mental health interventions in schools in high-income countries. Lancet Psychiatry 2014;1(5):377-387 [FREE Full text] [doi: 10.1016/S2215-0366(14)70312-8] [Medline: 26114092]
11. Heinrich CJ, Colomer A, Hieronimus M. Minding the gap: evidence, implementation and funding gaps in mental health services delivery for school-aged children. Child Youth Serv Rev 2023;150:107023 [FREE Full text] [doi: 10.1016/j.childyouth.2023.107023] [Medline: 37261333]
12. Panchal N, Cox C, Rudowitz R. The landscape of school-based mental health services. In: Encyclopedia of Health Services Research. USA: SAGE Publications; 2022:6.
13. Hoover S. Moving upstream: opportunities in school mental health. In: Scanlan Center for School Mental Health Practice Briefs. Iowa City, Iowa, USA: University of Iowa, Scanlan Center for School Mental Health; 2023:1-12.
14. Curtis BL, Ashford RD, Magnuson KI, Ryan-Pettes SR. Comparison of smartphone ownership, social media use, and willingness to use digital interventions between generation Z and millennials in the treatment of substance use: cross-sectional questionnaire study. J Med Internet Res 2019;21(4):e13050 [FREE Full text] [doi: 10.2196/13050] [Medline: 30994464]
15. Pretorius C, Chambers D, Cowan B, Coyle D. Young people seeking help online for mental health: cross-sectional survey study. JMIR Ment Health 2019;6(8):e13524 [FREE Full text] [doi: 10.2196/13524] [Medline: 31452519]
16. Guz S, Kattari SK, Atteberry-Ash B, Klemmer CL, Call J, Kattari L. Depression and suicide risk at the cross-section of sexual orientation and gender identity for youth. J Adolesc Health 2021;68(2):317-323. [doi: 10.1016/j.jadohealth.2020.06.008] [Medline: 32680801]
17. Schueller SM, Hunter JF, Figueroa C, Aguilera A. Use of digital mental health for marginalized and underserved populations. Curr Treat Options Psych 2019;6(3):243-255. [doi: 10.1007/s40501-019-00181-z]
18. Swart E, Bitzer EM, Gothe H, Harling M, Hoffmann F, Horenkamp-Sonntag D, et al. A consensus German reporting standard for secondary data analyses, version 2 (STROSA-STandardisierte BerichtsROutine für SekundärdatenAnalysen). Gesundheitswesen 2016;78(S 01):e145-e160 [FREE Full text] [doi: 10.1055/s-0042-108647] [Medline: 27351686]
19. Schleider JL, Dobias ML, Sung JY, Mullarkey MC. Future directions in single-session youth mental health interventions. J Clin Child Adolesc Psychol 2020;49(2):264-278 [FREE Full text] [doi: 10.1080/15374416.2019.1683852] [Medline: 31799863]
20. The institute of education sciences. National Center for Education Statistics. URL: https://nces.ed.gov/ccd/schoolsearch/ [accessed 2025-08-29]
21. Twigg E, Barkham M, Bewick BM, Mulhern B, Connell J, Cooper M. The young person's CORE: development of a brief outcome measure for young people. Couns Psychother Res 2009;9(3):160-168. [doi: 10.1080/14733140902979722]
22. Richardson L, Rockhill C, Russo JE, Grossman DC, Richards J, McCarty C, et al. Evaluation of the PHQ-2 as a brief screen for detecting major depression among adolescents. Pediatrics 2010;125(5):e1097-e1103 [FREE Full text] [doi: 10.1542/peds.2009-2712] [Medline: 20368315]
23. Toussaint A, Hüsing P, Gumz A, Wingenfeld K, Härter M, Schramm E, et al. Sensitivity to change and minimal clinically important difference of the 7-item generalized anxiety disorder questionnaire (GAD-7). J Affect Disord 2020;265:395-401. [doi: 10.1016/j.jad.2020.01.032] [Medline: 32090765]
24. Perczel Forintos D, Rózsa S, Pilling J, Kopp M. Proposal for a short version of the Beck Hopelessness Scale based on a national representative survey in Hungary. Community Ment Health J 2013;49(6):822-830. [doi: 10.1007/s10597-013-9619-1] [Medline: 23756722]
25. Hughes ME, Waite LJ, Hawkley LC, Cacioppo JT. A short scale for measuring loneliness in large surveys: results from two population-based studies. Res Aging 2004;26(6):655-672 [FREE Full text] [doi: 10.1177/0164027504268574] [Medline: 18504506]

XSL•FO
RenderX

26. Schleider JL, Weisz JR. Parent expectancies and preferences for mental health treatment: the roles of emotion mind-sets and views of failure. J Clin Child Adolesc Psychol 2018;47(sup1):S480-S496. [doi: 10.1080/15374416.2017.1405353] [Medline: 29364729]
27. Allaire J. RStudio: Integrated Development Environment for R. 2012. URL: http://www.rstudio.com [accessed 2025-08-29]
28. Twigg E, Cooper M, Evans C, Freire E, Mellor-Clark J, McInnes B, et al. Acceptability, reliability, referential distributions and sensitivity to change in the young person's clinical outcomes in routine evaluation (YP-CORE) outcome measure: replication and refinement. Child Adolesc Ment Health 2016;21(2):115-123. [doi: 10.1111/camh.12128] [Medline: 32680371]
29. Evaluating the accessibility, engagement, and effectiveness of alongside digital mental health app components in middle and high schools. Open Science Framework. URL: https://osf.io/m8t6k [accessed 2024-05-29]
30. Osborn TL. Supplemental material for single-session digital intervention for adolescent depression, anxiety, and well-being: outcomes of a randomized controlled trial with Kenyan adolescents. J Consult Clin Psychol 2020;88(7):657. [doi: 10.1037/ccp0000505.supp]
31. Fitzpatrick OM, Schleider JL, Mair P, Carson A, Harisinghani A, Weisz JR. Project SOLVE: randomized, school-based trial of a single-session digital problem-solving intervention for adolescent internalizing symptoms during the coronavirus era. School Mental Health 2023;15(3):955-966. [doi: 10.1007/s12310-023-09598-7]
32. Choi KR, Wisk LE, Zima BT. Availability of LGBTQ mental health services for US Youth, 2014 to 2020. JAMA Pediatr 2023;177(8):865-867 [FREE Full text] [doi: 10.1001/jamapediatrics.2023.1473] [Medline: 37273225]
33. Antaramian SP, Scott Huebner E, Hills KJ, Valois RF. A dual-factor model of mental health: toward a more comprehensive understanding of youth functioning. Am J Orthopsychiatry 2010;80(4):462-472. [doi: 10.1111/j.1939-0025.2010.01049.x] [Medline: 20950287]

Abbreviations

AI: artificial intelligence
BHS-4: Beck Hopelessness Scale-4
GAD-2: Generalized Anxiety Disorder Questionnaire-2
LGBTQ: lesbian, gay, bisexual, transgender, queer, and questioning
PHQ-2: Patient Health Questionnaire-2
ULS-3: University of California, Los Angeles Loneliness Scale-3
YP-CORE: Young Person's CORE

Edited by A Mavragani; submitted 26.02.25; peer-reviewed by U Linstedt, N Seeman; comments to author 06.07.25; revised version received 28.07.25; accepted 14.08.25; published 08.10.25.

Please cite as:
Cohen K, Rapoport A, Friis E, Hill S, Feldman S, Schleider J
The Alongside Digital Wellness Program for Youth: Longitudinal Pre-Post Outcomes Study
JMIR Form Res 2025;9:e73180
URL: https://formative.jmir.org/2025/1/e73180
doi:10.2196/73180
PMID:

©Katherine Cohen, Andy Rapoport, Elsa Friis, Shannon Hill, Sergey Feldman, Jessica Schleider. Originally published in JMIR Formative Research (https://formative.jmir.org), 08.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Exploring Empowerment in Online Support Communities for People Living With Tic Disorders and Tourette Syndrome: Qualitative Survey Study of User Experiences

Ella C Ford[1,2], BSc, MSc; Neil S Coulson[1], PhD; E Bethan Davies[3,4], PhD

[1]Lifespan and Population Health, School of Medicine, The University of Nottingham, Nottingam, United Kingdom

[2]School of Humanities and Social Sciences, Leeds Beckett Univeristy, Leeds, United Kingdom

[3]NIHR MindTech HealthTech Research Centre, Institute of Mental Health, School of Medicine, The University of Nottingham, Nottingham, United Kingdom

[4]Clinical Neurosciences and Mental Health, Institute of Mental Health, School of Medicine, The University of Nottingham, Nottingham, United Kingdom

Corresponding Author:
Ella C Ford, BSc, MSc
School of Humanities and Social Sciences
Leeds Beckett Univeristy
University City Campus
Leeds, LS1 3HE
United Kingdom
Phone: 44 1138120000
Email: e.ford@leedsbeckett.ac.uk

Abstract

Background: People with tic disorders (TDs)—such as Tourette syndrome—report poorer quality of life compared to non-TD peers, and experience considerable difficulties, including societal stigmatization and barriers to accessing health care and evidence-based interventions. Peer support can help improve psychological outcomes, and online support communities (OSCs) are one way to access psychological support. Empowerment involves improving an individual's cognitive processes to increase their ability to assert control over their health condition. OSCs have been suggested to facilitate empowerment, but this has not yet been investigated in users of OSCs for TDs.

Objective: This study aimed to explore empowerment processes and outcomes present in OSCs as perceived and reported by users of OSCs for TDs living with a TD.

Methods: A web-based survey of current users of OSCs for TDs (n=39) was conducted in summer 2022. The survey included four free-text questions about the impact of using OSCs for TDs in relation to empowerment, such as how it has affected their interactions with health care professionals (HCPs), decisions about treatment, and their experiences of living with a TD. Survey responses were analyzed using deductive and inductive reflexive thematic analysis guided by an initial coding structure derived from the empowering processes and outcomes theoretical framework.

Results: Analysis of responses identified a range of empowering processes (eg, exchanging information, encountering emotional support, finding recognition, and sharing experiences) and outcomes (eg, being better informed, feeling more confident in the relationship with HCPs and in making treatment decisions, and enhanced well-being) as related to OSC use in people with TDs that were consistent with the theoretical framework. Additionally, the process of changing stereotypes and stigma, the outcomes of raising awareness, and "one size does not fit all" were identified. A small number of disempowering processes and outcomes were identified, notably the outcome of feeling less confident in the relationship with HCPs.

Conclusions: These findings contribute to evidence for empowerment processes and outcomes experienced by users of health OSCs generally and highlight unique aspects of empowerment for users of OSCs with TDs and how these affect their experiences. OSCs appear to be an important tool in improving mental well-being through validation and recognition from peers, related to the acknowledgment of feelings of imposterism. Using OSCs can expand knowledge through exchanging information and experiences they may not have otherwise had access to—increasing empowerment through improvements in self-management and confidence in treatment decisions. However, this can also decrease trust in HCPs and negatively impact relationships due to decreased hope of assistance and fear of stigmatization for using OSCs. The authors suggest that recommendations regarding the

XSL•FO
RenderX

use of OSCs are presented in self-management guidelines concerning coping strategies and the importance of peer support in living with tics.

(JMIR Form Res 2025;9:e66912) doi:10.2196/66912

KEYWORDS

empowerment; online health communities; Tourette syndrome; qualitative research; web-based survey; support community; tic disorder; Tourette; survey; peer support; emotional support; mental health

Introduction

Tic disorders (TDs)—such as Tourette syndrome (TS) and chronic tic disorder (CTD)—are neurodevelopmental conditions characterized by tics: rapid, persistent, involuntary movements or vocalizations [1]. TDs are more common in men than women [2], and symptoms tend to have their onset during childhood. If tics are present for more than 12 months, assessment for CTD (presence of motor or vocal tics) or TS (motor and vocal tics) can be considered. Global prevalence of these TDs is estimated to be 0.69% (vocal CTD), 1.65% (motor CTD), and between 0.50% and 0.77% (TS) [2,3]. Tics typically wax and wane in frequency, severity, and impact over time, with variable prognosis as individuals age [4,5]. For most people living with TDs, tics normally decline through adolescence and early adulthood, with remission possible—however, for others, tics continue into adulthood [4,5]. TDs are incurable, but evidence-based pharmacological and behavioral interventions can aid symptoms [6]. However, people with TDs consistently report difficulties accessing health care and appropriate interventions, including a perceived lack of awareness among health care professionals (HCPs) and a scarcity of available treatment options [7-9].

There are many physical, psychological, societal, and cultural challenges associated with living with tics. Compared to adults without TDs, those living with TDs report a lower perceived quality of life, heightened levels of psychological difficulties [10-12], and are more likely to attempt and die by suicide [13]. TDs, in particular TS, are widely misunderstood and stigmatized in society; people with TDs experience many stigmatizing consequences, including anticipated and actual discrimination, bullying, mistreatment, and ostracization [14]. Tics are often visible and witnessed by others, frequently resulting in unwanted attention and negative comments [15]. Although tics are involuntary, many people with TDs report being able to delay or suppress their tics [16] to “fit in” and avoid stigmatizing responses [17-22]. However, this can be painful [9] and may strengthen self-stereotyping beliefs about having TDs, such as feelings of inauthenticity around “faking” tics [14,17]. These experiences of interpersonal, community-level, and structural stigmatization can negatively impact an individual’s attitudes about their tics and lead to avoidance of others and social isolation [14].

Peer support may help mitigate the social exclusion that people living with TDs frequently report [10,23]. As theorized by the direct effects hypothesis of social support [24], peer support can help ease isolation and loneliness and improve psychological outcomes through fostering connections. Peer-led support groups typically involve sharing and receiving several facets of social support, such as emotional, appraisal, and informational support, with this support being based on experiential knowledge rather than that obtained from HCPs [25]. The expansion and increased access to internet-enabled technologies in everyday life have led to the creation of asynchronous text-based online support communities (OSCs) for those living with specific health issues, wherein individuals can access peer support [26]. Text-based OSCs are unique as they can be accessed regardless of time and location, unlike physical support groups [27], and often provide access to a backlog of user discussions, as well as the ability to actively engage by posting new content and engaging with content posted by other members. There has been considerable growth in the number of OSCs available to individuals living with long-term conditions (LTCs) in recent years [28,29]. Research regarding the impact of OSCs has been conducted with people living with a range of LTCs such as HIV, diabetes, and arthritis [29-31]. A web-based survey of users of OSCs for TDs—including people with TDs and caregivers—distinguished multiple mechanisms through which users and caregivers benefit from membership [32]. This included improved knowledge of their condition and common co-occurring issues, increased confidence and understanding around communication with HCPs, positive impacts on their psychological well-being and appraisal of having tics, and an increased sense of belonging. These benefits also appear to relate to empowering processes and outcomes.

Empowerment and social support are interconnected concepts, but have some important distinctions. Social support is generally theorized as the perception of care, assistance, and comfort from a defined social network with functions such as emotional and informational support and companionship [33]. Empowerment, on the other hand, is a complex, multifaceted concept that functions to improve a person’s ability to be more in control of their own life and their power in decision-making [34]. It is concerned with an individual’s cognitive processes around increasing their autonomy and decreasing powerlessness, increasing their awareness of and building their strengths [35]. Empowerment has been hypothesized to be facilitated through engagement with OSCs, with several potential positive outcomes for people living with LTCs, including protecting members’ self-identity, improving the self-management of their condition, and developing their confidence in their ability to improve relationships with HCPs [34,36,37]. The processes through which these outcomes are gained include social support elements, such as supporting and being supported by peers with their condition, finding recognition, and becoming better informed through the sharing of knowledge [30,34,36]. Though increasing support is an important facet to OSCs, they also act as a complement to health care expertise—enabling individuals to translate this first-hand experience into practical

XSL•FO
RenderX

self-management techniques, expanding their knowledge whilst providing informational, emotional, and social support [34]. This knowledge empowers individuals to have responsibility over their health and to make independent management decisions [38]. In short, this knowledge development is considered separate from social support, but this knowledge may be acquired through it [39]. The type and frequency of interactions with OSCs can differ [40] and have also been suggested to affect the presence of empowerment and associated outcomes in some, but not all, communities [41-43]. As well as at the individual level, empowerment can also occur collectively within OSCs, wherein individuals engage as a group to increase their social power and create change for their condition, such as through policy reform [44-46]. Collective empowerment has important potential implications, such as building a collective "voice" for action and subsequently improving support services and societal attitudes [22,45].

Recent literature suggests that empowerment may be a useful framework to explore the relationship between engagement in OSCs and improved psychosocial outcomes [34,44,47]. Increased perceptions of one's empowerment have been associated with several health-related outcomes, including improved quality of life and reduced condition-related impairment [31,48,49]. As OSCs have the potential to improve empowerment and psychosocial outcomes for individuals with TDs, they could be a useful tool to complement HCP care and aid self-management [34,50]. To date, research exploring empowerment processes and outcomes in individuals with TDs is limited: previous findings suggest OSCs are a space where these can be fostered and result in "offline" outcomes such as improved communication with HCPs and increased well-being [32,51]. Therefore, this study aimed to explore empowerment processes and outcomes present in OSCs as perceived and reported by users of OSCs for TDs to understand how OSC use may affect empowerment, particularly in relation to health care interactions and self-management.

Methods

Study Design

This cross-sectional study consisted of a web-based survey completed by current users of OSCs for TDs. A web-based survey approach was chosen as it allowed a "wide-angle" lens on the topic due to the ability to recruit a larger number of participants, as compared to more time-intensive qualitative approaches, from geographically dispersed locations, increasing the diversity of voices [52]. This approach may also appeal more to people living with stigmatized conditions such as TDs due to their felt anonymity [53].

A qualitative phenomenological approach within a constructivist research paradigm was appropriate for this study, as it focused on understanding participants' experiences and perceptions of participating in OSCs, enriched through broader social contexts and integration of existing theory [54]. The analysis was guided by the theoretical framework of empowerment in OSCs developed by van Uden-Kraan et al [36,55].

This study is reported in accordance with the SRQR (Standards for Reporting Qualitative Research; Multimedia Appendix 1) [56], the CHERRIES (Checklist for Reporting Results of Internet E-Surveys; Multimedia Appendix 2) [57], and the CROSS (Checklist for Reporting of Survey Studies; Multimedia Appendix 3) [58].

Participants and Recruitment

Recruitment took place over 6 weeks from June to July 2022. To be eligible for the study, participants needed to be aged 16 years or older. As the purpose was to investigate experiences of OSC use of people living with a TD, participants needed to be current users of at least one OSC for TDs and have a self-reported confirmed or suspected TD diagnosis or symptoms of a TD. A formal diagnosis was not required to participate due to the aforementioned barriers to health care, which include assessment and diagnosis [59,60]. Akin to the previous approach taken by authors [32], studies were advertised through charities for TDs (eg, Tourette's Action) and primarily placed directly on OSCs, specifically inclusive of or exclusively aimed at people living with TS or TDs. These OSCs were identified by using a mix of key terms (eg, "Tourette," "tics," and "support") across Facebook, Reddit, and Discord social networks or platforms. To be approached to share the study, the OSC needed to operate predominantly in English, allow advertisements of research studies to be posted or shared on the OSC, have had activity within the past 30 days (eg, users posting messages), and have ≥100 members [61]. Previous studies have used similar criteria to identify actively used OSCs to reach current users [49,62]. Both open (ie, no membership required to access OSC) and closed or private OSCs were targeted for recruitment. In total, 26 eligible OSCs were identified: four were public, and the study advert was posted directly onto the OSC. For the 22 closed or private OSCs, the first author (ECF) directly messaged the OSC's moderators to request permission to post the advert to the OSC: 13 moderators responded and provided permission to post the study to their OSC. In total, the study advert was shared on 17 distinct OSCs. Based on a previous similar study of the same population group, the target sample size was 76 participants [32]. Data collection ended in July 2022 due to time restrictions, as the research was conducted as part of a postgraduate degree.

Ethical Considerations

This study was reviewed and received ethical approval by the MSc Health Psychology ethics review panel at the University of Nottingham (MEDS4008-22-15) and was conducted in accordance with the ethical standards of the Declaration of Helsinki. All OSCs allowed advertisements of research studies to be shared, and permissions were sought and granted from moderators to advertise the study in closed or private OSCs. Participation was voluntary, and informed consent was obtained from all participants electronically through the web-based survey. Participants were not asked to provide personally identifiable information. Data collected was anonymous, deidentified, and stored securely. TDs are suggestible; talking, hearing, or reading about tics is known to increase premonitory urges and tics themselves [63]. Therefore, this potential risk was clearly stated in the participant information sheet, and prior

XSL•FO
RenderX

to the open-ended questions. Participants were informed that they could withdraw from the study at any time should they become uncomfortable. To further minimize potential distress, the participant information sheet and debrief contained a list of TS or TD organizations and crisis hotlines operating worldwide. No compensation was provided for participants.

Procedure

All advertisements provided brief information about the study and a hyperlink to the open web-based survey (hosted on Jisc Online Surveys), which took approximately 20 minutes to complete. Prior to going live, the survey was piloted by a person with a TD to ensure the questions were acceptable and understandable, to assess the usability and technical functionality of the survey, and to estimate the time to complete—with relevant changes undertaken in response to their feedback (eg, consolidating two items about passive and active use into one Likert scale measuring passive to active use). The survey consisted of 11 pages: the first 2 web pages consisted of information about the study, including the estimated length of time of the survey, information regarding data storage, investigator details, and the purpose of the study, followed by a consent form on the third page in which participants provided informed consent through selecting several mandatory tick-boxes. The next 7 pages consisted of the survey questions: first, participants answered several demographic questions about themselves (gender, age, and country of residence), whether they had a confirmed TD diagnosis, self-reported co-occurring conditions, and use of medication and behavioral therapies for TDs. Tic-related impairment was assessed through 2 items from the impairment subscale of the Yale Global Tic Severity Scale—self-report version [10]. The Online Peer Support Group Use Scale [64] was used to collect information regarding participants' use of OSCs for TDs, including how many OSCs for TDs they were members of, how long and how often they used OSCs, their satisfaction with using them, and their estimated use on a scale of passive-active engagement in the past week. Active use was defined as direct use, such as posting content, replying to posts, and commenting on others' responses; passive use was defined as indirect use, such as browsing, reading, viewing, and "liking" posts. Finally, participants were presented with 4 open-ended questions (Textbox 1; adapted from Holbrey and Coulson [31]) relating to participants' use of OSCs for TDs and how they may relate to empowerment. The final page debriefed participants. All items except open-ended questions were required to be completed to proceed with the survey. All demographic items had a "prefer not to say" option, except for the TD diagnostic question, as this was required to assess questionnaire eligibility. Participants could go back and revise responses if needed, and could provide their email to receive a link to come back to the survey later. This is automatically provided by Jisc Online Surveys; researchers did not have access to participants' email addresses. All participants were allocated a unique response number by Jisc Online Surveys.

Textbox 1. Open-ended questions used in the web-based survey.

- How has your use of online support communities (OSCs) affected your interactions with health care professionals about Tourette syndrome (TS)/tic disorders (TDs)/tics, if at all?
- How has your use of OSCs affected your decisions about treatment and management for your TS/TDs/tics, if at all?
- Can you explain if using the OSC has changed your perceptions and experience of living with TS/TD/tics?
- Do you have any further comments about your experiences using the OSC for TS/TDs/tics?

Data Analysis

Descriptive statistics were calculated, including the absolute and relative frequencies of categorical variables and the mean and SD of continuous variables. A reflexive thematic analytical approach, as outlined by Braun and Clarke [65], was used to analyze the qualitative responses to the 4 open-ended questions with a predominantly deductive orientation to data coding, involving several stages of familiarity, coding, and grouping codes. The authors (ECF and EBD) jointly created an initial coding framework based on the empowering processes and outcomes questionnaires and theoretical framework developed by van Uden-Kraan et al [36,55] (Multimedia Appendix 4).

This framework outlines seven overarching consequences of engagement with OSCs (eg, increased confidence in the doctor-patient relationship and enhanced social well-being) and five processes (eg, exchanging of information and encountering emotional support) through which these consequences occur [55]. A similar framework was used to analyze experiences of using OSCs for HIV [47] and prostate cancer [66]; therefore, this deductive orientation was considered useful to provide a lens through which to interpret the data and provide more depth to previous literature on empowerment within OSCs. Minor changes were made to the original wording of van Uden-Kraan et al [36] of some empowerment elements in the framework, such as changing references from "physician" to "health care professional," and references to "illness" to "condition" to reflect terminology more appropriate to TDs. In line with Braun and Clarke's [65] suggestions in conducting deductive thematic analysis, the authors worked from a "curious, open, and questioning" position. Throughout the analysis, authors reflected on the "fit" of the framework with the data, ensuring this was not limiting the interpretation of the data, exploring both how the framework was evidenced and how it was not, and in what ways. The authors developed and refined the codebook throughout the process and used inductive techniques in identifying and coding data that did not appear to map onto the framework.

Participants' responses were first anonymized and uploaded into the software NVivo (version 14; Lumivero). The first author (ECF) conducted the first two stages of analysis. This involved familiarization of the data through reading and rereading of

XSL•FO
RenderX

participants' responses. The data were then coded using the predefined empowerment outcomes and processes coding framework. Several samples of data were cross-checked by a second researcher (EBD). Data that did not appear to map onto these a priori codes were identified, reflected on, and discussed between the two researchers alongside the coding framework overall. Where the terminology used for processes and outcomes in the predefined framework appeared not to completely align with what was found in the data, minor adjustments to the labels within the coding framework were made. First, the outcome "feeling more confident about treatment" was changed to "feeling more confident about treatment decisions," to reflect that this included decisions not to have treatment. Second, the word "optimism" was removed from the phrasing of the "increased optimism and control over the future" code, as evidence of increased optimism was not found in participants' responses. Third, there appeared to be a conceptual overlap in the "enhanced self-esteem" and "improved acceptance of the condition" codes: these codes were combined, along with references to mental well-being, as there was considerable overlap between these concepts. This resulted in the "enhanced acceptance, self-esteem, and mental well-being" code. Negative aspects relating to some of the overarching codes were also found in the data. These were negatively coded from their positive counterpart (eg, "decreased confidence in the relationship with HCPs," as a counterpart to "increased confidence in the relationship with HCPs") and embedded in the results for the overall theme. Additionally, through in-depth discussion, the authors developed and assigned inductive codes to the data that did not appear to map onto this predefined coding framework, resulting in some themes outside of the original framework. The first author had previously performed an inductive thematic analysis of the data—the codes and themes derived from this were compared with those of the present analysis and found to be similar.

Positionality Statement

This study was part of ECF's MSc Health Psychology dissertation, supervised by other authors EBD and NSC. ECF has experience working in health care services. EBD has research expertise in the field of TDs and NSC in the field of OSCs. Both EBD and NSC have previously conducted research around empowerment, as well as OSCs for people living with TDs, and all authors have experience conducting qualitative survey research. No authors have lived experience of TDs, and therefore are outsiders to the topic; however, all authors have experience of using health OSCs. Authors consulted with a person with lived experience of TDs in the development of the survey and incorporated their input.

Results

Descriptive Statistics

A total of 58 participants responded to the web-based survey. Of these, 39 participants provided free-text answers to at least one (of the four) open-ended questions and were included in the analysis. This was markedly lower than the targeted sample size, potentially due to the optional nature of the open-ended questions and the limited time available to recruit, which may have limited the range of perspectives captured. Participants were mainly female (19/39, 49%), based in the United Kingdom (20/39, 51%), and had received a TD diagnosis (29/39, 74%; Table 1). The majority accessed an OSC for TDs once (10/39, 26%) or twice (8/39, 21%) daily, with 13 (33%) participants checking between three to five times daily (Table 1). Participants estimated that a mean of 46% of their time using OSCs was "active" use as opposed to "passive" use in the last week.

XSL•FO
RenderX

Table 1. Demographic makeup of the sample and their self-reported use of OSCs[a] for TDs[b].

Characteristic	Participants (n=39)
Age (years), mean (SD)	25.5 (14.2)
Sex, n (%)	
Female	19 (49)
Male	8 (21)
Nonbinary	10 (26)
Other[c]	2 (5)
Country of residence, n (%)	
United Kingdom	20 (51)
United States	11 (28)
Australia	4 (10)
Canada	2 (5)
Egypt	1 (3)
Italy	1 (3)
Diagnosis status, n (%)	
Diagnosed with a TD	29 (74)
TD symptoms and currently awaiting assessment	7 (18)
TD symptoms and currently not waiting for assessment	2 (5)
Prefer not to answer	1 (3)
Taking medication for TDs, n (%)	
Yes	9 (23)
No	30 (77)
Received behavioral therapy for TDs, n (%)	
Yes	7 (18)
No	31 (80)
Self-reported co-occurring conditions, n (%)	
Attention-deficit/hyperactivity disorder	13 (33)
Autism	9 (23)
Obsessive compulsive disorder	19 (49)
Anxiety	22 (56)
Depression	17 (44)
Functional neurological disorder	4 (10)
Other	5 (13)
Number of different OSC memberships, n (%)	
1	7 (18)
2	12 (31)
3	8 (21)
4	4 (10)
≥5	8 (20)
Reported length of OSCs membership, n (%)	
1-6 months	6 (15)
6-12 months	14 (36)

XSL•FO
RenderX

Characteristic	Participants (n=39)
1-2 years	9 (23)
≥2 years	10 (26)
Reported daily use of OSCs, n (%)	
1 time	10 (26)
2 times	8 (21)
3 times	4 (10)
4 times	4 (10)
5 times	5 (13)
≥ 6 times	8 (21)

[a]OSC: online support community.

[b]TD: tic disorder.

[c]Other sexes indicated in free text response were "trans man" (n=1) and "two spirit" (n=1).

Evaluation Outcomes

Overview

Analysis of participants' responses to the open-ended questions resulted in the following themes, which mapped onto the empowerment processes and outcomes framework used for analysis, as well as arising additional inductive themes. Negative aspects affecting these processes and outcomes are also highlighted and discussed. Quotes are provided verbatim.

Empowering Processes Within OSCs

Exchanging Information

Participants described OSCs as being reliable places to ask questions and exchange valuable information regarding living with TDs, as well as for signposting to resources for further support:

> *I feel like now I have a community that I know who to go to if I have problems and so they know who to avoid and how to deal with things.* [P16, 16 y, other]

They found the information provided in OSCs to be particularly relevant and accessible, compared to knowledge acquired elsewhere:

> *It's also been particularly useful to see how Tourette's [sic] can coincide with other disorders like ADHD [Attention Deficit Hyperactivity Disorder], as that kind of information is much harder to find in things like medical studies or articles.* [P37, 17 y, nonbinary]

For others, the information provided a substitute for accessing HCPs, as they were able to learn about their condition and self-management strategies whilst waiting for care:

> *I don't have access to professional treatment currently due to waiting lists, so I've learned tic redirection techniques from other people online.* [P7, 17 y, nonbinary]

However, some participants complained about the lack of information organization; one participant noted their misinformation concerns:

> *As someone working in Healthcare [sic], I do see a huge amount of misinformation.* [P28, 22 y, female]

Encountering Emotional Support

Participants placed considerable importance and appreciation on being supported by their peers, often over other sources of support:

> *The people online have treated me better then [sic] my own family, and they've supported me better then [sic] my family, they saved my life.* [P17, 16 y, other]

They valued the opportunity to be supported by people who shared their condition:

> *They are a great resource, especially for people who don't know anyone with tics in real life and don't have a lot of support.* [P36, 16 y, female]

Much of the emotional support described was in the form of reassurance relating to feelings of "imposter syndrome" such as being fraudulent:

> *I have imposter syndrome surrounding my disorders including my Tourette syndrome all the time. These communities have given me a lot of reassurance when I do feel this way however.* [P20, 16 y, male]

Some participants reported varying levels of support across different platforms:

> *Some are fantastic and make me feel better about myself every day. They help to relieve imposter syndrome and provide support daily. Not all communities offer the same support.* [P21, 20 y, female]

Finding Recognition

Being in OSCs allowed people with TDs to connect with likeminded others when offline options were not available—"There are no in person support groups in my area so without the online group, I would not have any contact with people like me" (P5, 39 y, male)—this was greatly valued and sometimes key to their mental health: "meeting people like me who completely understand was the key to my mental well-being, that no therapy could even come close to" (P33, 19

XSL•FO
RenderX

y, female). They described an increased understanding that they were not alone in living with a TD: "It's good to know theres [sic] more than just me out there" (P1, 18 y, nonbinary). Participants were reassured that others encountered comparable struggles—"It's just nice to have a space of like-minded people who have similar experiences to you" (P7, 17 y, nonbinary)—and had made similar choices regarding the management of their condition:

> *I felt empowered in this personal choice in Tourette's [sic] communities as others have made this same decision for their own reasons, which has helped me feel not alone in making the right decision medically for myself and my Tourette's.* [P19, 22 y, female]

Helping Others

Few participants described helping others directly, with many instead writing generally about the supportive and community nature of the OSCs. However, some spoke of giving back to others in their communities—"And in turn I'm able to help younger peers the same, and help them to understand that it's okay to be different. Because I understand how scary it can be when you first develop tics" (P19, 22 y, female)—and supporting others whom they deem need it most:

> *So many people have had much rougher times of it. I try to help where I can, especially if someone doesn't have the support network they need.* [P11, 53 y, male]

One respondent was an OSC moderator who saw helping others as part of this role:

> *As an admin, advocate, and mentor in the groups, I'm able to provide information to other members re [sic] where to access healthcare support.* [P23, 52 y, nonbinary]

Sharing Experiences

The process of sharing experiences was interwoven with several other empowerment processes and outcomes. This appeared particularly tied to relating to others' shared experiences—"I like to analyse [sic] to a great extent and have people's experiences I can relate to" (P3, 17 y, nonbinary)—or regarding the product or outcome of sharing experiences, rather than about this process directly. One example of this is how considering others' experiences increased knowledge and understanding of how to manage tics:

> *Participating in online communities has allowed me to see what types of therapy and medications people used, and how they've learned to cope day to day with tics.* [P37, 17 y, nonbinary]

One participant shared concern about their tics being triggered by others' sharing their experiences due to the influential nature of TDs:

> *I think that at the moment, the online support community has a negative impact because people's tics are made worse by reading descriptions of other people's tics.* [P9, 18 y, female]

Another commented regarding the negative comparison of their own experiences with others:

> *I sometimes feel jealous that so many people are in contact with doctors that have more of an understanding and bond with their patient so they're getting somewhere positive and I'm stuck in the same place I've been for the past two years.* [P30, 16 y, female]

Changing Stereotypes and Stigma

An inductive finding from the data related to changes in attitudes and stigma through participation in OSCs. Being exposed to a range of lived experiences through OSCs positively challenged beliefs they held regarding the stereotypes of the condition:

> *Through use of communities such as discord I have gained a better understanding of the people my age living with tics. Ability to think beyond the stereotype of Tourette's.* [P21, 20 y, female]

However, some perceived that their use of OSCs increased stigmatization from others to themselves and their community, which they reflected to be due to media coverage reporting links between social media and tics:

> *The impact that my online communities have on me is all positive, however I feel unable to talk to my healthcare professionals about this due to the recent 'TikTok causes Tourette's articles" [sic] (even though I know this is untrue).* [P33, 19 y, female]

This seems particularly salient in young people with TDs, with some respondents claiming this stigma is partly exacerbated by other OSCs users "faking" or misrepresenting their tic experiences in digital spaces:

> *I love the friends I have made but I've found that there are lots of kids who are too young to be on the app and they're pretending more and more to be a part of our community so we face a lot of fake claiming because of them.* [P2, 18 y, female]

Empowering Outcomes Arising From OSCs

Being Better Informed

Most participants expressed that their knowledge of TDs has increased through OSC participation:

> *It made me feel more confident and knowledgeable about my Tourettes [sic] Syndrome.* [P13, 18 y, female]

They reported an increased understanding of the "correct" terminology to describe their condition and its related concepts:

> *I know the academical terms for everything related to my condition.* [P4, 19 y, male]

Additionally, they described increased awareness of treatment options available and strategies for self-managing tics to use when waiting to access treatment:

> *Online support communities have helped me gain a clearer understanding of some techniques available to manage tics, e.g [sic] CBIT [Comprehensive Behavioral Intervention for Tics]/Habit reversal, meaning I could work on these things whilst awaiting for professional treatment.* [P28, 22 y, female]

XSL•FO
RenderX

Some felt they had become more informed about living with TDs from the OSC than from HCPs:

> *I have learned more about Tourette syndrome through the online communities than through the health care system, which allows me to be better informed on a day to day basis living with it.* [P39, 36 y, male]

Feeling More Confident in the Relationship With HCPs

Through this aforementioned knowledge acquisition, participants also felt increased confidence in explaining their condition to HCPs, for reasons such as understanding of medical terminology:

> *Better knowledge of vocabulary to describe my experiences and symptoms in a way that is relevant to a medical professional.* [P28, 22 y, female]

Participants reported feeling more equipped for appointments, such as by preparing questions to ask HCPs, and feeling confident in interpreting the information given by HCPs:

> *I'll know what to say, and what questions to ask. I'll understand more of what i'm [sic] being diagnosed with.* [P17, 16 y, other]

This confidence was extended to their ability to advocate for themselves and potentially challenge and educate their HCPs regarding their care:

> *I was encouraged by members of online support groups to do my own research and to be my own advocate when speaking with healthcare workers.* [P11, 53 y, male]

Some participants reported that their OSC use had aided them in receiving an appropriate diagnosis. Many participants had not yet had an HCP appointment but felt that their OSCs had adequately prepared them for when they do and had given them the confidence to seek this help:

> *Being part of communities has given me the confidence to seek help regarding my tics. I have yet to have meaningful interaction with a doctor but have made an appointment to do so in the future, because I felt more equiped [sic].* [P21, 20 y, female]

Feeling Less Confident in the Relationship With HCPs

In contrast, many participants described a decreased level of confidence in their relationships with HCPs. Some felt their increasing knowledge led to negative judgment from HCPs:

> *I can better understand what they're talking about but the more you've educated yourself on the topic the more judgemental [sic] doctors can get.* [P2, 18 y, female]

Participants also described how widespread negative health care experiences were shared on OSCs, such as HCPs disregarding their comfort, not being useful, not caring about the effects of treatments, or living with the condition and dismissing their concerns:

> *I've learnt how little drs [sic] know about tic disorders but they don't seem to realize or care how that affects us when they are "treating" us. So many people with tic disorders often get dismissed even by the so called tic specialists.* [P15, 17 y, nonbinary]

Lack of access was also commonly described:

> *It is very much a common knowledge in the community that it is difficult to get a diagnosis and to find knowledgeable professionals who can help.* [P8, 40 y, male]

These negative perceptions and experiences that have been shared across their OSCs appear to have decreased their confidence in the health care system:

> *I would say that it has opened up my eyes to how badly people tics and Tourette's get treated by the public/health care professionals.* [P29, 17 y, nonbinary]

Some participants expressed concerns about members being reliant on OSCs and peer support over HCPs, partly due to poor accessibility:

> *People use online communities instead of accessing appropriate healthcare (partly because they don't see it as necessary and partly because there aren't sufficient services).* [P9, 18 y, female]

Participants also shared concerns about disclosing to HCPs that they use OSCs. They reported that OSC use is stigmatized by HCPs, and that sharing this will increase judgment and accusations of "faking" their condition:

> *I feel as if telling them I am a part of the Tourette's [sic] community on tiktok [sic] will make them doubt the legitimacy of my experiences, as I also have a diagnosis of FND [functional neurological disorder].* [P33, 19 y, female]

Enhanced Acceptance, Self-Esteem, and Mental Well-Being

Participants described how participation in OSCs helped to foster a sense of pride and build their confidence around their condition. This was regularly expressed through decisions to limit the suppression of their tics:

> *I also have started suppressing less in situations where doing so would make me uncomfortable/make my tics worse ... I'm more comfortable in allowing others to see me tic, and not being as embarassed [sic] as I was about it.* [P7, 17 y, nonbinary]

They often described feeling validated through their involvement, with this helping relieve feelings of imposterism:

> *Firstly it showed me that I do have tics and has shown me I don't have to have complex tics all the time to still be valid as a ticker [sic].* [P25, 19 y, female]

They felt more able to be open about their condition and accept their diagnosis—"I don't feel like a freak anymore, I didn't even want to admit I had tics and the online community made me feel better about having them and being apart [sic] of the community" (P17, 16 y, other)—and increased confidence in their future living with tics:

XSL•FO
RenderX

> *It has helped me come to terms with the fact that I will be going through this for the rest of my life and it's ok to live with it.* [P8, 40 y, male]

For some, this acceptance was linked to the decision not to pursue treatment:

> *Online support communities have helped me to accept my tics more so I have decided not to actively treat my Tourette syndrome with medication or behavioural [sic] therapy which has had a positive impact on my life.* [P9, 18 y, female]

Additionally, participants often attributed improved mental well-being to their use of OSCs:

> *I'm very, very glad I found these spaces early on after my tics started worsening. I think that if I hadn't found these spaces, then I would be in a worse place mentally than I am now.* [P7, 17 y, nonbinary]

Feeling More Confident About Treatment Decisions

Many participants described feeling more confident about their treatment decisions, whether this be feeling more informed about treatment options and choosing to have treatment, or choosing not to pursue treatment to decrease tics and feeling confident in this decision. The group provided many members with information about treatments that allowed them to explore further options and take time to contemplate their decision:

> *In making me aware of treatment options, I have been able to weigh the pros and cons prior to meeting with healthcare providers thus keeping the decision from being an on the spot answer.* [P5, 39 y, male]

Support from others in the OSCs, and their shared experiences and opinions regarding treatment, increased their confidence in their decisions:

> *A lot of people automatically think we want to take medication when lots of us don't. not because we like or want our tics but because the medication can have such a huge impact on not only your tics but you as a person, many of us don't think it's worth it.* [P19, 22 y, female]

Many perceived having treatment as a personal decision arising from individual circumstances:

> *We agree that there's no "one size fits all" approach to treating tics as our experiences, where similar, also vary hugely. So it's finding treatments and coping mechanisms that for [sic] right for you personally. [P10, 35 y, nonbinary)*

A minority highlighted that they used the OSC to enhance their HCP's guidance, rather than supplement or replace medical opinion:

> *I don't let my online communities influence the desicions [sic] about my treatment made by healthcare professionals. However, if I read about a treatment that may help me I sometimes ask them about it and if they think it will help me.* [P33, 19 y, female]

Some participants, however, expressed a lack of confidence in treatment, partially due to widespread problems experienced and reported by others with the health care system:

> *I've come to realize other people within my community have had the same or similar issues involving the healthcare system. I feel as though my issues are something only I can take care of at this point.* [P6, 18 y, male]

Increased Control Over the Future

Participants described feeling more in control of their condition. They felt confident in making choices in the management of their TD, such as in their treatment decisions and in choosing not to suppress their tics:

> *Healthcare decisions ultimately lie with the individual. Not society, not the doctor. The person.* [P11, 53 y, male]

Participants reported gaining greater control in managing their condition by learning and using self-management strategies described within OSCs:

> *Online support communities have helped me to better deal with tic attacks, redirecting tics.* [P20, 16 y, male]

Some expressed preparation for hardships that may lie ahead of them. However, this was associated with a degree of decreased hope for some participants:

> *It made me aware of a lot of the hardships we can face before I ended up facing them myself. It has sometimes made me feel worse about my future, knowing the extent of the discrimination against us, but it's also made me feel more prepared to face + work around it.* [P7, 17 y, nonbinary]

Enhanced Social Well-Being

Respondents overwhelmingly described a sense of community and belonging through their OSC participation:

> *It's helped me become more self-accepting and has given me a stronger sense of belonging. I don't feel isolated as I'm part of an empathetic community.* [P23, 52 y, nonbinary]

Participants described feeling less alone, which often went hand-in-hand with acceptance:

> *It has made me realise [sic] I'm not on my [sic] and I feel validated me.* [P38, 48 y, female]

Some participants described having formed new social relationships through OSCs, which appeared to be an important part in their journey in living with tics:

> *I have my own group of crazy amazing people and I don't know what I would have done in the beginning with out [sic] them.* [P30, 16 y, female]

Raising Awareness

An inductive finding from the data suggested that a handful of participants felt OSCs had given users the opportunity and drive to talk about TDs and raise awareness of their condition: "It's

XSL•FO
RenderX

also made me more passionate about increasing awareness and education re TS and tic disorders" (P23, 52 y, nonbinary), both to help their peers—"but I trey [sic] feel amazing that I'm able to help so many people by talking and spreading awareness" (P30, 16 y, female)—and to work to change public perceptions of TDs: "I like to use my experiences to educate others on it which will make it easier to be ourselves in public" (P2, 18 y, female).

One Size Does Not Fit All

This inductive finding found that not all OSCs were regarded equally and varied in their setup, with participants voicing their concerns and wishes for more and better options. Numerous issues were raised; a primary problem appeared to be the perceived dominance of parents of children with TDs in many groups, which often led to feelings of ostracization in adults living with TDs:

> *The only downside I've experienced in online communities is that in communities with guardians of children and youth they can often be dismissive of adults with Tourette Syndrome as they do but [sic] want to acknowledge that they're [sic] child(ren) may not outgrow the disorder, as doctors still tell parents is a possibility which they then cling to. This can lead to Adults with Tourette Syndrome bring [sic] ostracized to some degree.* [P39, 36 y, male]

One participant described parents requiring more support, which can be demanding for the participant:

> *Most members are parents who have children with Tourettes [sic], so I don't feel I can find support as I end up too busy giving them support.* [P36, 31 y, male]

Additionally, one participant described how users with less "severe" tics do not use the OSCs as much as others:

> *We've found the ones who access (the most) it are mainly people who's [sic] tics and TS are moderate to severe. We have a few members who's [sic] tics are mild, but we find they don't interact as much.* [P10, 35 y, nonbinary]

One participant also complained that some groups are not very active:

> *I wish a lot more of the disocord [sic] communities specifically were more active. And they all seem to have issues.* [P19, 22 y, female]

Participants reported preferring a variety of OSCs for tics to find the best fit for themselves:

> *So many of the groups out there cater to parents of kids with TS or to particular subsets, finding a good match wasn't straightforward. I was overjoyed to eventually find a couple of groups that were more open and catered more to people with TS, regardless of age. I'm sure the groups I'm part of aren't everyone's cup of tea, either, so a broader set of communities to choose from would be a good thing.* [P11, 53 y, male]

Discussion

Principal Findings

This study aimed to explore empowerment-related processes and outcomes experienced by current users of OSCs for TDs, such as TS, expanding upon initial findings from research exploring experiences of using OSCs for TDs and social support mechanisms within them [32,51]. Thematic analysis of open-ended survey responses generated evidence of various empowerment processes and outcomes consistent with our empowerment framework [36], as well as additional inductive findings. Our findings align with previous research exploring OSCs for other LTCs such as HIV [47] and cancer [35], and how OSC use intertwines with health care experiences, group access, and systems [34], with this study generating findings specifically relating to those with TDs.

As in previous studies [32,51], many participants reported exchanging information as a key part of using OSCs for TDs and subsequently reported being better informed about many aspects of living with this LTC, such as improved understanding of treatment options and self-management of tics. Some participants reported learning more from their OSCs than through their health care systems. This reflects several studies highlighting patients' experiences with HCPs in primary care: in the United Kingdom, young patients and their families have reported that their general practitioners provided little information about tics and tic management [59]. Patients and their families often have many concerns and questions about their tics [67]. Considering HCPs may not have sufficient answers, particularly regarding rarer conditions like TDs, it is understandable that individuals will seek help elsewhere [7,68,69], including from peers with experiential knowledge through similar health-related experiences.

This improved knowledge and understanding from OSC use was often applied to their interactions with HCPs, as participants reported increased confidence in interacting with HCPs within doctor-patient consultations, such as through knowing what questions to ask, the perceived appropriate language to use, and confidence to explain and advocate for their needs. This supports research illustrating that higher levels of patient empowerment are associated with greater participation in health care interactions [46,70] and contribute to the limited literature around the implications of peer digital support on relationships with HCPs [71]. However, although some participants reported feeling encouraged by their OSCs to seek help, within this study, there appeared to be a collective distrust of HCPs. This could potentially be a result of TDs being stigmatized within society and misunderstood by HCPs, as well as systematic barriers preventing health care access [14,72]. In this study, many participants shared their frustrations and experiences with health care systems. In contrast, those having positive HCP interactions or little health care difficulties may not share their perspectives in digital spaces. These adverse experiences echo previous research identifying obstacles in accessing primary care and specialist secondary care for tics [7,59]. Reading about others' negative experiences and opinions and being confronted with negative aspects of their health condition and the health care

XSL•FO
RenderX

system can be disempowering within OSCs [31,55], and some participants in this study did report feeling less confident in HCPs and in their treatment decisions. This nuanced pattern of the impacts of OSCs on health care interactions and empowerment contrasts with the more uniformly positive impact reported in broader health OSC studies [71]. Though increased empowerment has generally been shown to improve HCP relationships [34,36,37], in a survey of OSC users for a range of health conditions, Audrain-Pontevia and Menvielle [70] found that empowerment was negatively associated with their commitment to their relationship with their physician. Petrič et al [73] also demonstrated a complex relationship between OSC use and patient empowerment in relationships with HCPs, wherein exchanging information with users positively impacted self-efficacy but increased dysfunctional competencies in the relationship.

Some participants also reported concerns about sharing knowledge they had gained through OSCs with HCPs and disclosing their OSC use, due to perceived negative attitudes of HCPs around patient use of social media relating to TDs. Rupert et al [74] have found that HCPs often have negative reactions to patients sharing knowledge gained from OSCs. In regard to TDs, this could be due to HCP concerns around the quality and accuracy of web-based information [75], as well as potential stigma relating to recent publicity around links between functional tic-like behaviors and social media, particularly within the COVID-19 pandemic [76]. Increased presentations of tic-like behaviors [77,78] have been suggested to result from TD-related social media content, with some characterizing this rise as mass social media–induced psychogenic illness [76,79,80]. Other researchers have raised concerns with this theory, as it could perpetuate stereotypes that tics are voluntary attention-seeking behaviors, and add negative scrutiny toward people with TDs [81] who already express self-doubt over the legitimacy of their presentation [17,51]. This appears to be reflected in participants' reported apprehensions about sharing their OSC use with HCPs. Mainly, women based in the United Kingdom shared these concerns, which may be due to media attention around concerns about social media and tics, particularly in the United Kingdom [82,83], and could also relate to gendered perceptions about tic presentation in women as being "attention-seeking" [17].

A key empowerment outcome from using OSCs was improved mental well-being, self-esteem, and acceptance of tics, which appeared to develop from access to emotional support and finding recognition from their peers. Similar to other studies around OSCs like mental health communities [84], OSCs were able to provide users with access to similar peers whom they often lack in their offline lives, helping reduce isolation and allowing them to form social connections. A previous thematic analysis of posts on one TD OSC demonstrated that people with TDs have reported concerns of "faking" their tics or that others believe so [51]. In this way, finding recognition and peer support was often seen as validating, as others understood their experiences and subsequently reduced their perceptions of being an "imposter." This draws parallels with research surrounding peer support in those with another neurodevelopmental condition, attention-deficit/hyperactivity disorder, who also reportedly experience fears of imposterism that are helped by validation and acceptance received through OSCs [85]. This feeling may be more pronounced in TDs, as many features of tics—such as their ability to be suppressed and hidden, their wax-and-waning nature, and inconsistency in expression—may contribute toward these feelings of imposterism, as well as stigmatizing misperceptions of TDs in society [14]. Feelings of imposterism and inauthenticity have been reported by young women with TS, in questioning themselves on whether their tics are genuine or exaggerated—due to their ability to suppress tics, overlapping with symptoms of co-occurring conditions, and gendered perspectives in society around illness—and subsequently impacted on their sense of self and identity [17]. People with TDs often report feeling different, "weird" or "abnormal" from others [17,86,87], and interacting with peers who may have experienced similar "othered" experiences may assist in changing how they make sense of their condition and identity and empower them to take control over their condition [17,86,87]. Furthermore, one way those with TDs cope with the condition is how they align it with themselves, either integrating it within or externalizing it to their identity [88].

Interacting with similar others enhanced a sense of group identity, and OSCs were reported to increase self-acceptance and, in turn, influence how they managed their tics—including empowering them in making decisions relating to treatment, tic expression, and suppression. Through unmasking and not suppressing their tics, people with TDs may reduce pain related to tic suppression [9] and may contribute toward changing the social perceptions around visible differences, like tics, to create a more positive social environment for acceptance and social support to promote adaptive coping [88].

An empowerment outcome inductively identified from participants' responses centered around raising awareness of TDs in society, to increase public understanding and decrease prejudice and discrimination toward people with TDs. This appears to map onto patient activism and collective empowerment [34]. Given that TDs are misunderstood and stigmatized in society [14], having a collective voice through OSCs might help in unifying marginalized individuals together and subsequently produce collective action to change societal attitudes. This patient activism could increase social change and improve service provision, although increased collective empowerment does not necessarily mean higher civic participation [44]. In interviews with people with an intellectual disability who used self-advocacy groups, Anderson and Bigby [89] reported that these groups helped increase their self-efficacy and confidence, including their ability to speak out for themselves and similar others, and to use their lived experiences as "expertise" in educating others about intellectual disability. As people with TDs may be another group marginalized and stigmatized within society, membership of TD groups could provide opportunities for creating individual and collective empowerment, which could counteract stigma and promote societal change.

Another inductive finding related to the suitability of different OSCs for different types of individuals using them, such as users who have tics or a TD themselves, or users who are parents or caregivers of children with TDs, with some users expressing

XSL•FO
RenderX

difficulties with finding an appropriate group. This suggests that empowerment processes and outcomes can vary depending on the OSC used and its characteristics. A previous study into post–COVID-19 condition highlighted the size of groups and their platform as impacting sense of community and symptom management [90], and moderation was deemed to be important in gaining positive outcomes from OSC use for family caregivers [91]. In this study, the presence of carers appeared to affect the perceived value gained from the groups for people with TDs. The prominence of parent-dominated groups reported by participants related to feelings of exclusion and ostracization from these OSCs, supporting findings from Perkins et al [32]. Adults with TDs have reported limited options for general age-appropriate social support for TDs, with support being more available for children and parents or caregivers [17,22]. This variability in membership of OSCs for TDs can relate to disempowerment [31]: people with TDs and their families have some similar but also contrasting support needs. For example, parents of children with TS may feel anxious and stressed by their child's symptoms and their impact on their daily life [92] and so require specific support for this. Therefore, it may be valuable to have distinct peer support options for people living with the condition and for those caring for them. Additionally, tailored support for subgroups could be important for addressing additional needs within specific groups—for example, having spaces for adults and women living with TDs [17,22].

Strengths and Limitations

This study contributes to the limited amount of evidence exploring OSC use for TDs, and how it can be useful as part of TD management and coping with the long-term nature of TDs, as well as how empowerment manifests in those with TDs. The study was novel in the use of an established empowerment framework to guide analysis and in asking participants to share their experiences in their own words.

One limitation of this study was the size and nature of the sample. Despite a recruitment drive over 6 weeks across a range of platforms, the sample was relatively small. Future studies would benefit from a longer sampling period and working with further patient organizations to aid recruitment. The sample itself was also homogeneous in some respects. First, most participants were based in the United Kingdom; therefore, findings may relate to issues mainly within the UK health care systems. Second, participants were mainly female or nonbinary, meaning male views were less represented in our findings. This gender ratio contrasts with TDs being more prevalent in male patients [2]: this may be due to gender differences in using social support for health-related purposes [93], or potentially in coping with TDs [88]. Third, the age of participants in this study was skewed quite young. This could be reflective of the higher prevalence of TDs in younger adults [4,5]. Research concerning how age may affect OSC use, particularly in relation to empowerment, is limited. One study [94] suggested that digital support had a more pronounced effect on positive coping affect among older women with cancer compared to younger adults. A more age-diverse sample could reveal differing effects among age groups. Though not the aim of qualitative constructivist research [95], the small size and elements of homogeneity in the sample limit generalization, and findings should be interpreted with this consideration. Demographic and socioeconomic factors (eg, ethnicity and income), which could potentially impact on health care experiences, were also not recorded [96]. TD diagnosis was self-reported, with most participants reporting a diagnosis, although this was not verified, and standardized measures were not used to assess current tic symptomology.

A significant limitation of the study is that we sought to recruit only participants who were current users of OSCs, who were self-selected. Despite some participants reporting negative experiences [34], this introduced a bias toward identifying positive outcomes of OSCs, as more empowered people are more likely to continue using OSCs [97]. People who have left OSCs are presumably more likely to have had negative experiences or to have found the groups disempowering or of no value. These perspectives and experiences are therefore likely to have been underrepresented in our sample. Though this does not detract from the mostly positive and helpful experiences reported by participants in this study, which is corroborated by other research on OSCs [34-36,50], experiences of empowerment in people who leave OSCs should be investigated in further research to more thoroughly determine the usefulness of OSCs for people living with TDs.

In regard to analysis, data often appeared to overlap across several empowerment processes and outcomes. This partially reflects the difficulty in fully deciphering responses due to the text-based survey design, as authors could not ask participants to expand on their answers, but also illustrates the difficulties of using a previously defined framework for deductive analysis. It is difficult to concretely separate actions or processes from their results or outcomes. Processes are that where participants explained taking part in forms of empowerment, such as exchanging information, and the outcomes are the consequences of this, such as being better informed. These are likely to overlap as participants, for example, discuss learning more about TDs from communities, and authors had to make judgments about this in the analysis. Nonetheless, the framework provided a useful lens to understand empowerment experiences in OSCs, and to further understanding of empowerment in people with TDs in particular, which has had limited exploration. Furthermore, the inductively formed themes appeared to highlight more of the unique experience of living with a TD than deductively derived themes, which often highlighted findings more broadly applicable to OSCs, though this contributed to key work in this area by demonstrating the similarities between diverse health communities.

Finally, the open-ended questions used likely elicited more responses about empowerment outcomes, such as health care experiences and treatment decisions, rather than empowerment processes such as helping others. Our questions also did not explicitly ask participants about collective empowerment or disempowerment outcomes and processes they had experienced within OSCs, meaning there may have been additional findings relating to these aspects had this been explicitly asked about.

Future Directions

A key perspective missing in research exploring OSCs for LTCs is from HCPs themselves: what do HCPs think about patients

XSL•FO
RenderX

bringing knowledge and perspectives they have learned from OSCs? Participants in this study reported concerns about discussing information gathered in OSCs with their HCPs. Previous research suggests that HCPs often react negatively to patients sharing information from OSCs [74]. This may be more pronounced in interactions with patients with TDs, as within the COVID-19 pandemic, HCPs observed a marked increase in functional tic-like behaviors presenting in specialist services [77,78], leading to doubts about the legitimacy of patients presenting with tics [76,79-81]. The etiology and treatment of functional tic-like behaviors are different from tics within TDs, but can understandably appear similarly in patients with TDs, so there may be a mismatch in understanding between patients and HCPs. It is important to understand the barriers to transparency and knowledge sharing in doctor-patient consultations, which can jeopardize the doctor-patient relationship; breaking down these barriers has the potential to improve patient knowledge, empowerment, and self-management [74].

Future research should also consider the perspectives of different types of TD OSC users. As TDs typically have onset in childhood, caregivers of children with TDs may use OSCs [32]. The empowering or disempowering processes and outcomes they experience through OSC use may have similarities and differences from users with TDs; exploring these may illuminate ways in which these platforms can be used to support caregivers in both supporting themselves and their relative or friend with a TD. Additionally, the perspectives of moderators or administrators of OSCs should be considered as the gatekeepers who arbitrate over the digital space. In this study, one participant reported that they were a moderator of a TD OSC and emphasized the importance of their role in supporting other users. Previous research has shown this is an empowering process [98]. Future research should investigate the role of moderators in OSCs concerning empowerment more specifically. Finally, this study focused on people who were active users of OSCs for TDs—in contrast, another topic to explore is former users of these spaces and what led them to leave or disconnect from OSCs. As this study highlighted several disempowerment processes, it may be that disconnecting from OSCs could be a disempowerment outcome arising from participation.

This research was cross-sectional and qualitative, subsequently only providing a snapshot into the experience of interest. Therefore, it may be useful to experimentally investigate how using OSCs impacts people living with TDs over time, for example, by capturing changes in empowerment, clinical outcomes, and mechanisms such as shame, self-esteem, and self-management while using OSCs, particularly as there is a lack of experimental research in this area [99]. This could also help to identify the mechanisms of action of the effects of OSCs on empowerment [100]. Further research may provide evidence of the effects found in this study and contribute to the understanding of how OSCs may help people living with TDs to manage their condition.

Conclusions

This study illustrates the empowering processes of OSC use and the empowering outcomes that their use can have on how people with TDs make sense of and live with long-term tics. This information could contribute to self-management guidelines concerning coping strategies and the importance of peer support in living with tics. It also emphasizes the relationship between OSC use and health care management and relationships with HCPs. HCPs must recognize their patients' experiences of using OSCs and be aware of patients' potential concerns about raising this with them.

Acknowledgments

We would like to thank all of the online support communities who advertised our study; Tourette's Action for their advertisement and support of the study, as well as Danni Phoenix-Kane for reviewing the content of the web-based survey prior to it going live, and to all participants who took their time to complete the study. Generative artificial intelligence was not used to generate any portion of the submitted manuscript. This work was cofunded by the National Institute for Health and Care Research (NIHR) MindTech HealthTech Research Center and the NIHR Nottingham Biomedical Research Center. The views expressed are those of the authors and not necessarily those of the National Health Service, the NIHR, or the Department of Health and Social Care. The sources of funding had no role in the study's design, data collection, data analysis, or interpretation.

Data Availability

The datasets generated or analyzed during this study are not publicly available as they include sensitive data, but are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization, methodology, writing—review and editing: ECF, NSC, EBD; Data curation and investigation: ECF; Formal analysis and writing—original draft: ECF and EBD; Supervision: NSC and EBD.

Conflicts of Interest

None declared.

Multimedia Appendix 1

XSL•FO
RenderX

Standards for Reporting Qualitative Research (SRQR) checklist.
[DOCX File , 22 KB - formative_v9i1e66912_app1.docx]

Multimedia Appendix 2
Checklist for Reporting Results of Internet E-Surveys (CHERRIES).
[DOCX File , 20 KB - formative_v9i1e66912_app2.docx]

Multimedia Appendix 3
Checklist for Reporting of Survey Studies (CROSS).
[DOCX File , 21 KB - formative_v9i1e66912_app3.docx]

Multimedia Appendix 4
Coding framework based on empowering processes and outcomes questionnaires developed by van Uden-Kraan et al (2009).
[DOCX File , 18 KB - formative_v9i1e66912_app4.docx]

References

1. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders. Washington, DC: American Psychiatric Publishing; 2013.
2. Knight T, Steeves T, Day L, Lowerison M, Jette N, Pringsheim T. Prevalence of tic disorders: a systematic review and meta-analysis. Pediatr Neurol 2012;47(2):77-90. [doi: 10.1016/j.pediatrneurol.2012.05.002] [Medline: 22759682]
3. Jafari F, Abbasi P, Rahmati M, Hodhodi T, Kazeminia M. Systematic review and meta-analysis of Tourette syndrome prevalence; 1986 to 2022. Pediatr Neurol 2022;137:6-16. [doi: 10.1016/j.pediatrneurol.2022.08.010] [Medline: 36182698]
4. Black KJ, Kim S, Yang NY, Greene DJ. Course of tic disorders over the lifespan. Curr Dev Disord Rep 2021;8(2):121-132 [FREE Full text] [doi: 10.1007/s40474-021-00231-3] [Medline: 34178575]
5. Hassan N, Cavanna AE. The prognosis of Tourette syndrome: implications for clinical practice. Funct Neurol 2012;27(1):23-27 [FREE Full text] [Medline: 22687163]
6. Chou C, Agin-Liebes J, Kuo S. Emerging therapies and recent advances for Tourette syndrome. Heliyon 2023;9(1):e12874 [FREE Full text] [doi: 10.1016/j.heliyon.2023.e12874] [Medline: 36691528]
7. Cuenca J, Glazebrook C, Kendall T, Hedderly T, Heyman I, Jackson G, et al. Perceptions of treatment for tics among young people with Tourette syndrome and their parents: a mixed methods study. BMC Psychiatry 2015;15:46 [FREE Full text] [doi: 10.1186/s12888-015-0430-0] [Medline: 25879205]
8. Malli MA, Forrester-Jones R. Stigma and adults with Tourette's syndrome: "Never Laugh at Other People's Disabilities, unless they have Tourette's—because how can you not?". J Dev Phys Disabil 2021;34(5):871-897. [doi: 10.1007/s10882-021-09829-2]
9. Taylor E, Anderson S, Davies EB. "I'm in pain and I want help": an online survey investigating the experiences of tic-related pain and use of pain management techniques in people with tics and tic disorders. Front Psychiatry 2022;13:914044 [FREE Full text] [doi: 10.3389/fpsyt.2022.914044] [Medline: 35990066]
10. Conelea CA, Woods DW, Zinner SH, Budman CL, Murphy TK, Scahill LD, et al. The impact of Tourette syndrome in adults: results from the Tourette syndrome impact survey. Community Ment Health J 2013;49(1):110-120. [doi: 10.1007/s10597-011-9465-y] [Medline: 22052430]
11. Hirschtritt ME, Lee PC, Pauls DL, Dion Y, Grados MA, Illmann C, Tourette Syndrome Association International Consortium for Genetics. Lifetime prevalence, age of risk, and genetic relationships of comorbid psychiatric disorders in Tourette syndrome. JAMA Psychiatry 2015;72(4):325-333 [FREE Full text] [doi: 10.1001/jamapsychiatry.2014.2650] [Medline: 25671412]
12. Evans J, Seri S, Cavanna AE. The effects of Gilles de la Tourette syndrome and other chronic tic disorders on quality of life across the lifespan: a systematic review. Eur Child Adolesc Psychiatry 2016;25(9):939-948 [FREE Full text] [doi: 10.1007/s00787-016-0823-8] [Medline: 26880181]
13. de la Cruz LF, Rydell M, Runeson B, Brander G, Rück C, D'Onofrio BM, et al. Suicide in Tourette's and chronic tic disorders. Biol Psychiatry 2017;82(2):111-118. [doi: 10.1016/j.biopsych.2016.08.023] [Medline: 27773353]
14. Pring K, Malli M, Hardy BW, Rapp SR, Storch EA, Mink JW, et al. Reframing stigma in Tourette syndrome: an updated scoping review. Eur Child Adolesc Psychiatry 2025;34(1):19-39. [doi: 10.1007/s00787-023-02332-3] [Medline: 38159134]
15. Buckser A. Before your very eyes: illness, agency, and the management of Tourette syndrome. Med Anthropol Q 2008;22(2):167-192. [doi: 10.1111/j.1548-1387.2008.00013.x] [Medline: 18717365]
16. Ueda K, Kim S, Greene DJ, Black KJ. Correlates and clinical implications of tic suppressibility. Curr Dev Disord Rep 2021;8(2):112-120 [FREE Full text] [doi: 10.1007/s40474-021-00230-4] [Medline: 34178574]
17. Coleman J, Melia Y. Me, my tics and I: an exploration of self-identity and its implications for psychological wellbeing in young women with Tourette's syndrome. J Dev Phys Disabil 2023;36:1-23. [doi: 10.1007/s10882-023-09911-x] [Medline: 37361460]

XSL•FO
RenderX

18. Cutler D, Murphy T, Gilmour J, Heyman I. The quality of life of young people with Tourette syndrome. Child Care Health Dev 2009;35(4):496-504. [doi: 10.1111/j.1365-2214.2009.00983.x] [Medline: 19638024]
19. Edwards KR, Mendlowitz S, Jackson E, Champigny C, Specht M, Arnold P, et al. A qualitative exploration of the experiences of children and adolescents with Tourette syndrome. J Can Acad Child Adolesc Psychiatry 2017;26(1):39-44 [FREE Full text] [Medline: 28331502]
20. Lee M, Mu P, Wang W, Wang H. 'Living with tics': self-experience of adolescents with Tourette syndrome during peer interaction. J Clin Nurs 2016;25(3-4):463-471. [doi: 10.1111/jocn.13074] [Medline: 26818371]
21. Lee M, Wang H, Chen C, Lee M. Social adjustment experiences of adolescents with Tourette syndrome. J Clin Nurs 2019;28(1-2):279-288. [doi: 10.1111/jocn.14564] [Medline: 29893437]
22. Malli MA, Forrester-Jones R, Triantafyllopoulou P. “Tourette’s Is a Lonely Place”: an interpretative phenomenological analysis of the personal experience and identity of adults with Tourette’s syndrome. J Dev Phys Disabil 2019;31(6):819-845. [doi: 10.1007/s10882-019-09676-2]
23. Smith H, Fox JRE, Trayner P. The lived experiences of individuals with Tourette syndrome or tic disorders: a meta-synthesis of qualitative studies. Br J Psychol 2015;106(4):609-634. [doi: 10.1111/bjop.12118] [Medline: 25721405]
24. Dennis C. Peer support within a health care context: a concept analysis. Int J Nurs Stud 2003;40(3):321-332. [doi: 10.1016/s0020-7489(02)00092-5] [Medline: 12605954]
25. Doull M, O'Connor A, Welch V, Tugwell P, Wells G. Peer support strategies for improving the health and well-being of individuals with chronic diseases. Cochrane Database Syst Rev 2017;2017(6):CD005352. [doi: 10.1002/14651858.cd005352.pub2]
26. Coulson NS, Buchanan H. The role of online support groups in helping individuals affected by HIV and AIDS: scoping review of the literature. J Med Internet Res 2022;24(7):e27648 [FREE Full text] [doi: 10.2196/27648] [Medline: 35881456]
27. Coulson NS, Smedley R, Bostock S, Kyle SD, Gollancz R, Luik AI, et al. The pros and cons of getting engaged in an online social community embedded within digital cognitive behavioral therapy for insomnia: survey among users. J Med Internet Res 2016;18(4):e88 [FREE Full text] [doi: 10.2196/jmir.5654] [Medline: 27113540]
28. Smedley RM, Coulson NS. A practical guide to analysing online support forums. Qual Res Psychol 2018;18(1):76-103. [doi: 10.1080/14780887.2018.1475532]
29. Allen C, Vassilev I, Kennedy A, Rogers A. Long-term condition self-management support in online communities: a meta-synthesis of qualitative papers. J Med Internet Res 2016;18(3):e61. [doi: 10.2196/jmir.5260] [Medline: 26965990]
30. Bartlett YK, Coulson NS. An investigation into the empowerment effects of using online support groups and how this affects health professional/patient communication. Patient Educ Couns 2011;83(1):113-119. [doi: 10.1016/j.pec.2010.05.029] [Medline: 20599338]
31. Holbrey S, Coulson NS. A qualitative investigation of the impact of peer to peer online support for women living with polycystic ovary syndrome. BMC Womens Health 2013;13:51 [FREE Full text] [doi: 10.1186/1472-6874-13-51] [Medline: 24341398]
32. Perkins V, Coulson NS, Davies EB. Using online support communities for tourette syndrome and tic disorders: online survey of users' experiences. J Med Internet Res 2020;22(11):e18099 [FREE Full text] [doi: 10.2196/18099] [Medline: 33141089]
33. Khademi K, Kaveh MH. Social support as a coping resource for psychosocial conditions in postpartum period: a systematic review and logic framework. BMC Psychol 2024;12(1):301 [FREE Full text] [doi: 10.1186/s40359-024-01814-6] [Medline: 38807228]
34. Johansson V, Islind AS, Lindroth T, Angenete E, Gellerstedt M. Online communities as a driver for patient empowerment: systematic review. J Med Internet Res 2021;23(2):e19910 [FREE Full text] [doi: 10.2196/19910] [Medline: 33560233]
35. Ziegler E, Hill J, Lieske B, Klein J, dem OV, Kofahl C. Empowerment in cancer patients: Does peer support make a difference? A systematic review. Psychooncology 2022;31(5):683-704. [doi: 10.1002/pon.5869] [Medline: 34981594]
36. van Uden-Kraan C, Drossaert C, Taal E, Seydel E, van de Laar M. Participation in online patient support groups endorses patients' empowerment. Patient Educ Couns 2009;74(1):61-69. [doi: 10.1016/j.pec.2008.07.044] [Medline: 18778909]
37. Wakefield D, Bayly J, Selman LE, Firth AM, Higginson IJ, Murtagh FE. Patient empowerment, what does it mean for adults in the advanced stages of a life-limiting illness: a systematic review using critical interpretive synthesis. Palliat Med 2018;32(8):1288-1304 [FREE Full text] [doi: 10.1177/0269216318783919] [Medline: 29956568]
38. Vainauskienė V, Vaitkienė R. Enablers of patient knowledge empowerment for self-management of chronic disease: an integrative review. Int J Environ Res Public Health 2021;18(5):2247 [FREE Full text] [doi: 10.3390/ijerph18052247] [Medline: 33668329]
39. Disler R, Appleton J, Smith T, Hodson M, Inglis S, Donesky D, et al. Empowerment in people with COPD. PI 2015:7. [doi: 10.2147/pi.s61195]
40. Carron-Arthur B, Ali K, Cunningham JA, Griffiths KM. From help-seekers to influential users: a systematic review of participation styles in online health communities. J Med Internet Res 2015;17(12):e271. [doi: 10.2196/jmir.4705] [Medline: 26627369]

XSL•FO
RenderX

41. Fullwood C, Chadwick D, Keep M, Attrill-Smith A, Asbury T, Kirwan G. Lurking towards empowerment: explaining propensity to engage with online health support groups and its association with positive outcomes. Comput Hum Behav 2019;90:131-140. [doi: 10.1016/j.chb.2018.08.037]
42. Smit D, Vrijsen JN, Broekman T, Groeneweg B, Spijker J. User engagement within an online peer support community (Depression Connect) and recovery-related changes in empowerment: longitudinal user survey. JMIR Form Res 2022;6(11):e39912 [FREE Full text] [doi: 10.2196/39912] [Medline: 36322110]
43. van Uden-Kraan CF, Drossaert CHC, Taal E, Seydel ER, van de Laar MAFJ. Self-reported differences in empowerment between lurkers and posters in online patient support groups. J Med Internet Res 2008;10(2):e18. [doi: 10.2196/jmir.992] [Medline: 18653442]
44. Atanasova S, Petric G. Collective empowerment in online health communities: scale development and empirical validation. J Med Internet Res 2019;21(11):e14392 [FREE Full text] [doi: 10.2196/14392] [Medline: 31746772]
45. Fayn M, des Garets V, Rivière A. Collective empowerment of an online patient community: conceptualizing process dynamics using a multi-method qualitative approach. BMC Health Serv Res 2021;21(1):958 [FREE Full text] [doi: 10.1186/s12913-021-06988-y] [Medline: 34511115]
46. Petrovčič A, Petrič G. Differences in intrapersonal and interactional empowerment between lurkers and posters in health-related online support communities. Comput Hum Behav 2014;34:39-48. [doi: 10.1016/j.chb.2014.01.008]
47. Mo PKH, Coulson NS. Are online support groups always beneficial? A qualitative exploration of the empowering and disempowering processes of participation within HIV/AIDS-related online support groups. Int J Nurs Stud 2014;51(7):983-993. [doi: 10.1016/j.ijnurstu.2013.11.006] [Medline: 24342449]
48. Mo PKH, Coulson NS. Online support group use and psychological health for individuals living with HIV/AIDS. Patient Educ Couns 2013;93(3):426-432. [doi: 10.1016/j.pec.2013.04.004] [Medline: 23669152]
49. Algtewi E, Owens J, Baker SR. Online support groups for head and neck cancer and health-related quality of life. Qual Life Res 2017;26(9):2351-2362 [FREE Full text] [doi: 10.1007/s11136-017-1575-8] [Medline: 28417218]
50. Anawade PA, Sharma D, Gahane S. Connecting health and technology: a comprehensive review of social media and online communities in healthcare. Cureus 2024;16(3):e55361 [FREE Full text] [doi: 10.7759/cureus.55361] [Medline: 38562335]
51. Soós MJ, Coulson NS, Davies EB. Exploring social support in an online support community for Tourette syndrome and tic disorders: analysis of postings. J Med Internet Res 2022;24(10):e34403 [FREE Full text] [doi: 10.2196/34403] [Medline: 36194454]
52. Braun V, Clarke V, Boulton E, Davey L, McEvoy C. The online survey as a qualitative research tool. Int J Soc Res Methodol 2020;24(6):641-654. [doi: 10.1080/13645579.2020.1805550]
53. Terry G, Braun V. Short but often sweet: the surprising potential of qualitative survey methods. In: Collecting Qualitative Data: A Practical Guide to Textual, Media and Virtual Techniques. Cambridge, England: Cambridge University Press; 2017.
54. Urcia IA. Comparisons of adaptations in grounded theory and phenomenology: selecting the specific qualitative research methodology. Int J Qual Methods 2021;20:16094069211045474. [doi: 10.1177/16094069211045474]
55. van Uden-Kraan CF, Drossaert CHC, Taal E, Shaw BR, Seydel ER, van de Laar MAFJ. Empowering processes and outcomes of participation in online support groups for patients with breast cancer, arthritis, or fibromyalgia. Qual Health Res 2008;18(3):405-417. [doi: 10.1177/1049732307313429] [Medline: 18235163]
56. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014;89(9):1245-1251 [FREE Full text] [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]
57. Eysenbach G. Improving the quality of web surveys: the checklist for reporting results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004;6(3):e34 [FREE Full text] [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
58. Sharma A, Duc NTM, Thang TLL, Nam NH, Ng SJ, Abbas KS, Jacqz-Aigrain, et al. A consensus-based checklist for reporting of survey studies (CROSS). J Gen Intern Med 2021;36(10):3179-3187 [FREE Full text] [doi: 10.1007/s11606-021-06737-1] [Medline: 33886027]
59. Marino C, Khan K, Groom MJ, Hall SS, Anderson S, Mcnally E, et al. Patients' experience of accessing support for tics from primary care in the UK: an online mixed-methods survey. BMC Health Serv Res 2023;23(1):788 [FREE Full text] [doi: 10.1186/s12913-023-09753-5] [Medline: 37488511]
60. Hall CL, Le Novere M, Murphy T, McNally E, Hollis C, Hunter R. Healthcare utilisation and costs associated with poor access to diagnosis and treatment for children and young people with tic disorders. BMJ Ment Health 2024;27(1):e301241 [FREE Full text] [doi: 10.1136/bmjment-2024-301241] [Medline: 39515845]
61. Voelpel SC, Eckhoff RA, Förster J. David against Goliath? Group size and bystander effects in virtual knowledge sharing. Hum Relat 2008;61(2):271-295. [doi: 10.1177/0018726707087787]
62. Mo PKH, Coulson NS. Developing a model for online support group use, empowering processes and psychosocial outcomes for individuals living with HIV/AIDS. Psychol Health 2012;27(4):445-459. [doi: 10.1080/08870446.2011.592981] [Medline: 21854088]
63. Efron D, Dale RC. Tics and tourette syndrome. J Paediatr Child Health 2018;54(10):1148-1153. [doi: 10.1111/jpc.14165] [Medline: 30294996]

XSL•FO
RenderX

64. Tan YT, Rehm IC, Stevenson JL, De Foe A. Social media peer support groups for obsessive-compulsive and related disorders: understanding the predictors of negative experiences. J Affect Disord 2021;281:661-672. [doi: 10.1016/j.jad.2020.11.094] [Medline: 33234279]
65. Braun V, Clarke V. Thematic Analysis: A Practical Guide. London: SAGE; 2022.
66. Campbell KA, Coulson NS, Buchanan H. Empowering processes within prostate cancer online support groups. IJWBC 2013;9(1):51-66. [doi: 10.1504/ijwbc.2013.051294]
67. Wu M, McGuire J. Chapter 2—psychoeducation about tic disorders and treatment. In: The Clinician's Guide to Treatment and Management of Youth with Tourette Syndrome and Tic Disorders. Oxford, England: Academic Press; 2018.
68. Ekhtiari H, Rezapour T, Aupperle R, Paulus M. Neuroscience-informed psychoeducation for addiction medicine: a neurocognitive perspective. Prog Brain Res 2017;235:239-264 [FREE Full text] [doi: 10.1016/bs.pbr.2017.08.013] [Medline: 29054291]
69. Ashtari S, Taylor A. Patients with rare diseases and the power of online support groups: implications for the medical community. JMIR Form Res 2023;7:e41610 [FREE Full text] [doi: 10.2196/41610] [Medline: 37707878]
70. Audrain-Pontevia A, Menvielle L. Do online health communities enhance patient-physician relationship? An assessment of the impact of social support and patient empowerment. Health Serv Manage Res 2018;31(3):154-162. [doi: 10.1177/0951484817748462] [Medline: 29280679]
71. Kjærulff EM, Andersen TH, Kingod N, Nexø MA. When people with chronic conditions turn to peers on social media to obtain and share information: systematic review of the implications for relationships with health care professionals. J Med Internet Res 2023;25:e41156 [FREE Full text] [doi: 10.2196/41156] [Medline: 37067874]
72. Groom M, Babbage C. Steering group policy brief: end the postcode lottery. The Insitute of Mental Health Nottingham. URL: https://institutemh.org.uk/images/Projects_and_studies/Projects_and_studies_images/Tourettes_Steering_Group/Tics_Healthcare_Policy_Brief_-_UoN.pdf [accessed 2025-06-24]
73. Petrič G, Atanasova S, Kamin T. Impact of social processes in online health communities on patient empowerment in relationship with the physician: emergence of functional and dysfunctional empowerment. J Med Internet Res 2017;19(3):e74 [FREE Full text] [doi: 10.2196/jmir.7002] [Medline: 28288953]
74. Rupert DJ, Moultrie RR, Read JG, Amoozegar JB, Bornkessel AS, O'Donoghue AC, et al. Perceived healthcare provider reactions to patient and caregiver use of online health communities. Patient Educ Couns 2014;96(3):320-326 [FREE Full text] [doi: 10.1016/j.pec.2014.05.015] [Medline: 24923652]
75. Treadgold BM, Coulson NS, Campbell JL, Lambert J, Pitchforth E. Quality and misinformation about health conditions in online peer support groups: scoping review. J Med Internet Res 2025;27:e71140 [FREE Full text] [doi: 10.2196/71140] [Medline: 40377972]
76. Martindale JM, Mink JW. The rise of functional tic-like behaviors: what do the COVID-19 pandemic and social media have to do with it? A narrative review. Front Pediatr 2022;10:863919 [FREE Full text] [doi: 10.3389/fped.2022.863919] [Medline: 35899132]
77. Heyman I, Liang H, Hedderly T. COVID-19 related increase in childhood tics and tic-like attacks. Arch Dis Child 2021;106(5):420-421. [doi: 10.1136/archdischild-2021-321748] [Medline: 33677431]
78. Pringsheim T, Ganos C, McGuire JF, Hedderly T, Woods D, Gilbert DL, et al. Rapid onset functional Tic-like behaviors in young females during the COVID-19 pandemic. Mov Disord 2021;36(12):2707-2713 [FREE Full text] [doi: 10.1002/mds.28778] [Medline: 34387394]
79. Giedinghagen A. The tic in TikTok and (where) all systems go: mass social media induced illness and Munchausen's by internet as explanatory models for social media associated abnormal illness behavior. Clin Child Psychol Psychiatry 2023;28(1):270-278. [doi: 10.1177/13591045221098522] [Medline: 35473358]
80. Müller-Vahl KR, Pisarenko A, Jakubovski E, Fremer C. Stop that! It's not Tourette's but a new type of mass sociogenic illness. Brain 2022;145(2):476-480 [FREE Full text] [doi: 10.1093/brain/awab316] [Medline: 34424292]
81. Conelea C, Bervoets J, Bethan Davies E, Varner K, Malli M, Jones D, et al. A call for caution: 'stop that' sentiments threaten tic research, healthcare and advocacy. Brain 2022;145(4):e18-e20. [doi: 10.1093/brain/awac028] [Medline: 35213691]
82. Kale S. ‘The unknown is scary’: why young women on social media are developing Tourette’s-like tics. The Guardian. 2021. URL: https://www.theguardian.com/media/2021/nov/16/the-unknown-is-scary-why-young-women-on-social-media-are-developing-tourettes-like-tics [accessed 2024-09-30]
83. Seale J. Britain’s Tourette’s mystery review—is Tik Tok really causing the boom in tics? The Guardian. 2022. URL: https://www.theguardian.com/tv-and-radio/2022/jul/19/britains-tourettes-mystery-review-tiktok-tik-attack-united-young-people-with-embarrassing-tics [accessed 2024-09-30]
84. Smith-Merry J, Goggin G, Campbell A, McKenzie K, Ridout B, Baylosis C. Social connection and online engagement: insights from interviews with users of a mental health online forum. JMIR Ment Health 2019;6(3):e11084 [FREE Full text] [doi: 10.2196/11084] [Medline: 30912760]
85. Eagle T, Ringland K. “You Can't Possibly Have ADHD”: exploring validation and tensions around diagnosis within unbounded ADHD social media communities. 2: ACM; 2023 Presented at: ASSETS '23: The 25th International ACM SIGACCESS Conference on Computers and Accessibility; October 22-25, 2023; New York, USA URL: https://dl.acm.org/doi/abs/10.1145/3597638.3608400 [doi: 10.1145/3597638.3608400]

XSL•FO
RenderX

86. Conelea CA, Woods DW, Zinner SH, Budman C, Murphy T, Scahill LD, et al. Exploring the impact of chronic tic disorders on youth: results from the tourette syndrome impact survey. Child Psychiatry Hum Dev 2011;42(2):219-242. [doi: 10.1007/s10578-010-0211-4] [Medline: 21046225]
87. Wadman R, Tischler V, Jackson GM. 'Everybody just thinks I'm weird': a qualitative exploration of the psychosocial experiences of adolescents with Tourette syndrome. Child Care Health Dev 2013;39(6):880-886. [doi: 10.1111/cch.12033] [Medline: 23363392]
88. Maxwell-Scott M, O'Keeffe F, Eccles FJR. Coping with tourette's syndrome: a meta-ethnography of individual and family perspectives. Psychol Health 2024:1-23 [FREE Full text] [doi: 10.1080/08870446.2024.2360126] [Medline: 38818804]
89. Anderson S, Bigby C. Self-advocacy as a means to positive identities for people with intellectual disability: 'We Just Help Them, Be Them Really'. J Appl Res Intellect Disabil 2017;30(1):109-120. [doi: 10.1111/jar.12223] [Medline: 26514344]
90. Mills F, Drury J, Symons C, Weston D, Amlôt R, Carter H. Online support groups, social identity, and the health and wellbeing of adults with long COVID: an interview study. J Community Appl Soc Psychol 2024;34(5):e2849. [doi: 10.1002/casp.2849]
91. Daynes-Kearney R, Gallagher S. Online support groups for family caregivers: scoping review. J Med Internet Res 2023;25:e46858 [FREE Full text] [doi: 10.2196/46858] [Medline: 38090796]
92. O'hare D, Eapen V, Grove R, Helmes E, Mcbain K, Reece J. Youth with tourette syndrome: parental perceptions and experiences in the Australian context. Aust J Psychol 2020;69(1):48-57. [doi: 10.1111/ajpy.12111]
93. Reevy GM, Maslach C. Use of social support: gender and personality differences. Sex Roles 2001;44(7-8):437-459. [doi: 10.1023/a:1011930128829]
94. Seçkin G. I am proud and hopeful: age-based comparisons in positive coping affect among women who use online peer-support. J Psychosoc Oncol 2011;29(5):573-591. [doi: 10.1080/07347332.2011.599361] [Medline: 21882936]
95. Varpio L, O'Brien B, Rees CE, Monrouxe L, Ajjawi R, Paradis E. The applicability of generalisability and bias to health professions education's research. Med Educ 2021;55(2):167-173. [doi: 10.1111/medu.14348] [Medline: 32779251]
96. Schmengler H, Cohen D, Tordjman S, Melchior M. Autism spectrum and other neurodevelopmental disorders in children of immigrants: a brief review of current evidence and implications for clinical practice. Front Psychiatry 2021;12:566368 [FREE Full text] [doi: 10.3389/fpsyt.2021.566368] [Medline: 33815159]
97. Sharma S, Khadka A. Role of empowerment and sense of community on online social health support group. ITP 2019;32(6):1564-1590. [doi: 10.1108/itp-09-2018-0410]
98. Coulson NS, Shaw RL. Nurturing health-related online support groups: exploring the experiences of patient moderators. Comput Hum Behav 2013;29(4):1695-1701. [doi: 10.1016/j.chb.2013.02.003]
99. Mills F, Drury J, Hall CE, Weston D, Symons C, Amlôt R, et al. A mixed studies systematic review on the health and wellbeing effects, and underlying mechanisms, of online support groups for chronic conditions. Commun Psychol 2025;3(1):40. [doi: 10.1038/s44271-025-00217-6] [Medline: 40089592]
100. Rayland A, Andrews J. From social network to peer support network: opportunities to explore mechanisms of online peer support for mental health. JMIR Ment Health 2023;10:e41855 [FREE Full text] [doi: 10.2196/41855] [Medline: 36853738]

Abbreviations

CHERRIES: Checklist for Reporting Results of Internet E-Surveys
CROSS: Checklist for Reporting of Survey Studies
CTD: chronic tic disorder
HCP: health care professional
LTC: long-term condition
OSC: online support community
SRQR: Standards for Reporting Qualitative Research
TD: tic disorder
TS: Tourette syndrome

Edited by A Mavragani; submitted 01.10.24; peer-reviewed by A Eisingerich, H Weingarden, R Zito; comments to author 26.03.25; revised version received 25.06.25; accepted 22.07.25; published 09.10.25.

Please cite as:
Ford EC, Coulson NS, Davies EB
Exploring Empowerment in Online Support Communities for People Living With Tic Disorders and Tourette Syndrome: Qualitative Survey Study of User Experiences
JMIR Form Res 2025;9:e66912
URL: https://formative.jmir.org/2025/1/e66912
doi:10.2196/66912
PMID:

XSL•FO
RenderX

©Ella C Ford, Neil S Coulson, E Bethan Davies. Originally published in JMIR Formative Research (https://formative.jmir.org), 09.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

ChatGPT-Based Chatbot for Help Quitting Smoking via Text Messaging: An Interventional Study

Lorien C Abroms[1], SCD; Christina N Wysota[2], PhD; Artin Yousefi[3], MSc; Tien-Chin Wu[1], MPH; David A Broniatowski[3], PhD

[1]Milken Institute School of Public Health, George Washington University, Washington, DC, United States
[2]Grossman School of Medicine, New York University, New York, NY, United States
[3]School of Engineering and Applied Science, George Washington University, Washington, DC, United States

Corresponding Author:
Lorien C Abroms, SCD
Milken Institute School of Public Health
George Washington University
950 New Hampshire St. NW
Washington, DC, 20052
United States
Phone: 1 3019801452
Email: lorien@gwu.edu

Abstract

Background: Large language model chatbots such as ChatGPT may be able to provide support to people who smoke cigarettes and are trying to quit.

Objective: This pilot study examined the feasibility and acceptability of integrating a specialized ChatGPT-based chatbot, BeFreeBot, into a smoking cessation text messaging intervention, BeFree. Chatbot fidelity was also examined.

Methods: Participants who smoked cigarettes in the previous 7 days were recruited from Amazon Mechanical Turk (N=23), enrolled in BeFree, and provided access to BeFreeBot. Surveys were administered at baseline and 4 weeks after enrollment to assess perceptions of BeFreeBot. Computer records of interactions between BeFreeBot and participants were also analyzed to assess participant engagement and adherence of BeFreeBot to its instructions. For the adherence analysis, transcripts were dual coded, and discrepancies were resolved by a third coder.

Results: Most participants (16/23, 70%) texted BeFreeBot with questions or concerns at least once. Participants sent 14.5 (SD 23.6) texts to BeFreeBot on average. Most participants were highly satisfied with BeFreeBot (13/18, 72%) and agreed that it was helpful for quitting (11/19, 58%). They also reported that the BeFreeBot responses were clear and easy to understand (16/17, 94%) and that they trusted responses from BeFreeBot (12/17, 71%). Most participants (17/19, 90%) reported trying to quit smoking for 1 day or longer, and 30% (7/23) self-reported no smoking in the previous 7 days. An analysis of transcripts of BeFreeBot responses (n=328) revealed that BeFreeBot functioned as instructed on most measures, with clear language (328/328, 100%), follow-up questions asked of participants (13/16, 81%), and recommendations to seek out professional counseling (13/16, 81%) or consider the use of Food and Drug Administration–approved medications (eg, nicotine replacement therapy; 14/16, 88%). Responses stayed on the topic of smoking cessation counseling (324/328, 98.8%) and did not include information that contradicted the US Preventive Services Task Force guidelines (328/328, 100%).

Conclusions: A specialized large language model chatbot integrated into an SMS text messaging program and accessed through SMS text message was found to be feasible and acceptable to smokers.

(JMIR Form Res 2025;9:e79402) doi:10.2196/79402

KEYWORDS

ChatGPT; large language model; chatbots; tobacco; smoking cessation; cigarettes; SMS text messaging; artificial intelligence

XSL•FO
RenderX

Introduction

Background

Tobacco use is the leading preventable cause of death, disability, and disease burden in the United States and worldwide [1]. Approximately 11.6% of adults in the United States are current smokers, and an estimated 70% of smokers want to quit [2]. The earlier in the life course an individual stops smoking, the greater the health benefits are for both the individual and society as a whole [1].

The US Preventive Services Task Force (USPSTF) states that effective smoking cessation treatments should offer counseling with social support and pharmacotherapy except where it is contraindicated (eg, pregnancy) [3]. There is a strong dose-response relationship between the intensity of smoking cessation counseling and its effectiveness, with counseling involving more sessions leading to higher quit rates [3].

Automated smoking cessation programs on mobile phones that use SMS text messaging have been found to be effective for smoking cessation and other health behaviors [4-7]. A recent meta-analysis of SMS text messaging programs for smoking cessation concluded that such programs double abstinence rates [5]. SMS text messaging programs for smoking cessation have been used by millions of smokers trying to quit [4].

Chatbots or computerized conversational agents have the potential to extend the capabilities of SMS text messaging programs and other digital interventions by providing interactive and humanlike personalized counseling on quitting [8-13]. Historically, chatbot conversations have been predetermined based on prewritten statements and if-then rules to create simulated conversations with users. One meta-analysis of chatbots for smoking cessation identified 6 trials of such chatbots and found that chatbot use by smokers was associated with a 66% increase in the odds of quitting [10]. The largest and best designed of these was a study by Perski et al [13] that examined the addition of a chatbot to a smoking cessation app and found that offering the chatbot+app increased engagement and improved smoking cessation–related outcomes over offering the app alone. However, most studies to date are limited by having small sample sizes and are of mixed quality. In addition, the chatbots studied predate the introduction of generative artificial intelligence (AI) chatbots [10-13].

With rapid and ongoing developments in generative AI, including with large language models (LLMs), the natural language capabilities of chatbots have significantly improved. LLM chatbots such as OpenAI's ChatGPT, Google's Gemini, and Meta's Llama allow for open-text queries and provide dynamic, natural language responses that are sensitive to context, nuance, and history and resemble a human conversation [14-16]. In 2025, a total of 34% of American adults and 58% of adults under 30 years of age in the United States, reported that they had used ChatGPT, with use rising from the previous year [17]. On April 1, 2024, the World Health Organization released a web-based chatbot based on OpenAI's ChatGPT, SARAH (Smart AI Resource Assistant for Health), which aims to assist in smoking cessation and modifying other health behaviors. SARAH, which has been evaluated for the content of its smoking cessation information [18], is available in 8 languages and has been used by over 40,000 people [19].

While these chatbots appear promising, challenges to their effective use may include chatbots providing information that is false or invented (ie, "hallucinated"). To date, a handful of studies have investigated LLM chatbots for smoking cessation [11,18,20,21]. One study used experts to rate motivational messages to quit smoking and found most LLM chatbot messages to be highly rated [21]. Another pilot study predating ChatGPT examined whether an app with an LLM chatbot feature could help in quitting smoking and showed promising results [20]. Finally, a study by our research team found that ChatGPT-based chatbots could provide reliable responses to queries on how to quit smoking [18]. Most responses to queries on quitting smoking had clear language (97.3%) and included a recommendation to seek out professional counseling (80.3%). Less common was the recommendation to consider using nicotine replacement therapy (NRT; 52.7%), provision of information on how to deal with cravings when quitting smoking (44.4%), or the recommendation to consider the use of non-NRT prescription drugs (14.1%). Specific queries that were most challenging for the chatbots included queries on how to quit smoking cold turkey, using gummies, or using hypnosis [18].

As LLM chatbots are increasingly used by the public, it is important to evaluate them as tools for smoking cessation and other types of health behavior change. Studies are needed to evaluate whether the instructions and materials provided to LLM chatbots serve as effective guardrails (eg, can chatbots stay on topic, maintain their roles as counselors, and adhere to prespecified guidelines regardless of the query). In addition, studies are needed that can explore how chatbots can be integrated into existing methods that smokers use for quitting and whether people are willing to engage with them. Ultimately, studies are needed to determine whether LLM chatbots can provide benefits to people trying to quit smoking.

Objectives

This study aimed to assess the feasibility and acceptability of offering a ChatGPT-based chatbot, BeFreeBot, to smokers in the context of an existing SMS text messaging smoking cessation program, BeFree. Key questions of interest for the study included whether people trying to quit smoking using an SMS text messaging program would engage with a chatbot (feasibility), how they would view the helpfulness of the chatbot in this context (acceptability), and whether the chatbot would respond to users as instructed (chatbot fidelity).

Methods

Overview

Participants were recruited in January 2025 and February 2025 by posting a Human Intelligence Task (HIT) in Amazon Mechanical Turk (MTurk) for smokers who were willing to review health messages. MTurk is a marketplace run by Amazon that connects workers willing to do tasks for businesses, researchers, and others. It allows workers to self-identify as being smokers (and having other characteristics) as part of its

XSL•FO
RenderX

registration survey. MTurk workers were able to access the HIT if they were over 18 years of age, lived in the United States, had an HIT approval rate of ≥85%, and self-identified as smokers in the MTurk registration system. In addition, the MTurk workers were eligible if they indicated on our eligibility survey that (1) they had smoked a cigarette in the previous 7 days and (2) they were willing to receive SMS text messages. No requirement was established regarding interest in quitting or motivation to quit. To prevent fraudulent enrollees, participants were excluded if they failed an attention-check question on the screener, a question designed to assess whether the participant was reading the question and unrelated to the subject of quitting smoking (eg, "Pick the response that is a vegetable").

Eligible MTurk workers provided consent and completed a 30-minute web-based baseline survey, during which they were required to set a quit date within the following week with a default date of Monday. Participants were then enrolled in the BeFree SMS text messaging program, which included instructions on accessing BeFreeBot. Participants completed follow-up surveys 4 weeks after baseline.

Ethical Considerations

This study was reviewed by the George Washington University Institutional Review Board (NCR235403) and deemed to be exempt. Participants provided written informed consent to be enrolled in the study via Qualtrics (Qualtrics International Inc). The consent form informed participants that conversations with the chatbot would be reviewed by study staff and would be monitored twice a week for the duration of the study. The surveys were stored on a secure, password-protected platform, and only authorized research staff had access to survey responses. Participants received a US $20 gift card for completing the baseline survey and US $25 for completing the 2- and 4-week follow-up surveys (the 2-week results are not presented in this paper). As a precautionary measure, chatbot conversations were reviewed by study staff 2 to 3 times per week for the presence of any alarming comments by the chatbot (eg, chatbot suggestions that may result in harm).

Intervention

BeFree

BeFree is an automated SMS text messaging program for smoking cessation and modeled on previous programs that have been found to double the chances of quitting, such as Text2Quit and SmokefreeTXT [5-7]. BeFree sends out 1 to 2 messages per day timed around each participant's chosen quit date and continuing for 30 days after enrollment. Message content includes an overview of the quitting process and triggers, behavioral tools and techniques to resist cravings, information on pharmacotherapy, information on how to maintain abstinence and respond to slipups, and general motivational messaging aimed at increasing self-efficacy. Users were given on-demand keywords for help with cravings ("CRAVE") and for relapse and resetting their quit date ("SMOKED").

To promote engagement with BeFreeBot (see the following section), for the first 7 days and then every 5 days after that, messages from BeFree ended by prompting the user to reach out and text BeFreeBot on a different phone number for more information on that subject. For example, one message from BeFree read the following: "BeFree: Remember, every day you're adding benefits for your health, wallet and your future. Text BeFreeBot and share your challenges: [Phone number]." At any point, participants had the option to end all messaging by texting "STOP" to either number.

BeFreeBot

BeFreeBot was a revised version of an earlier chatbot created by our team called BeFreeGPT [18]. BeFreeBot was created using a ChatGPT Plus account using GPT-4o-2024-11-20, which has training data up to October 2023, and the Assistants application programming interface [22]. The Assistants application programming interface feature allows for a simplified process of creating a chatbot. It allows for the provision of instructions to the chatbot in prose (vs code) and for the uploading of documents that form a knowledge base that has priority over the model's own existing knowledge base [23].

We provided instructions to the chatbot to act like a smoking cessation counselor, be empathetic, ask follow-up questions, and provide smoking cessation advice based on the documents provided in its knowledge base. Additional instructions to the chatbot included directions to introduce itself before answering the first question (ie, say, "Hi I'm BeFreeBot, an AI counselor to help you quit smoking..."), instructions to make responses less than 50 words long, and a directive to limit conversations to the topic of quitting smoking. The documents uploaded to its knowledge base were (1) the USPSTF recommendation statement *Interventions for Tobacco Smoking Cessation in Adults, Including Pregnant Persons* [3]; (2) *Clearing the Air: Quit Smoking Today*, a guidebook developed by the National Cancer Institute on how to quit smoking [24]; and (3) Food and Drug Administration (FDA)–approved and FDA-cleared cessation products [25]. Figure 1 shows a schematic of BeFreeBot's inputs and outputs.

Figure 1. Overview of BeFreeBot's inputs and outputs.

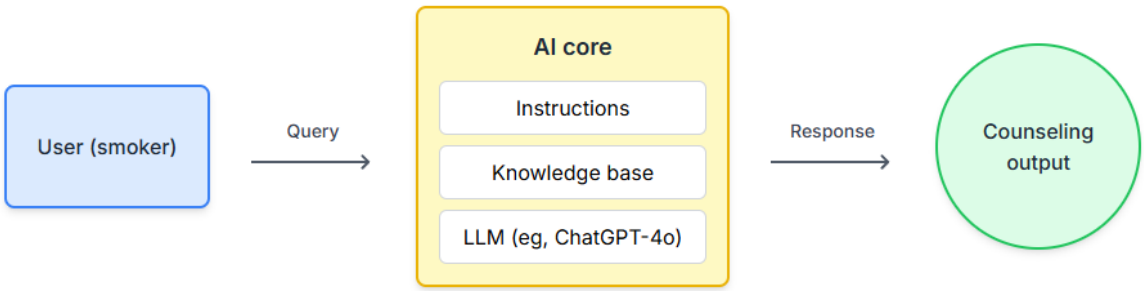

XSL•FO
RenderX

BeFreeBot differed from BeFreeGPT [18] in that it had additional instructions to increase its engagement with participants (eg, instructions to ask follow-up questions) and provide additional counselor-like empathy in its responses and a list of FDA-approved cessation products. In this study, participants could access BeFreeBot through SMS text messages and receive instant SMS text message responses in natural language to any query. This novel format was developed to simulate texting a live counselor who was available 24/7. At enrollment, BeFreeBot reached out to participants to introduce itself. In addition, at 2 weeks after enrollment, a check-in message was sent to each user from BeFreeBot that read the following: "Hi! It's BeFreeBot. Just checking in to hear how your quit attempt is going. Let me know if there are any challenges I can help you with?" Because of the experimental nature of BeFreeBot and the concern that interactions with it could be harmful to participants, all interactions between BeFreeBot and participants were reviewed by study staff twice a week for the duration of the trial to detect any potential problems (See Figure 2 for an example of messages with BeFree and BeFreeBot).

Figure 2. Example of a text message from (A) BeFree and (B) BeFreeBot.

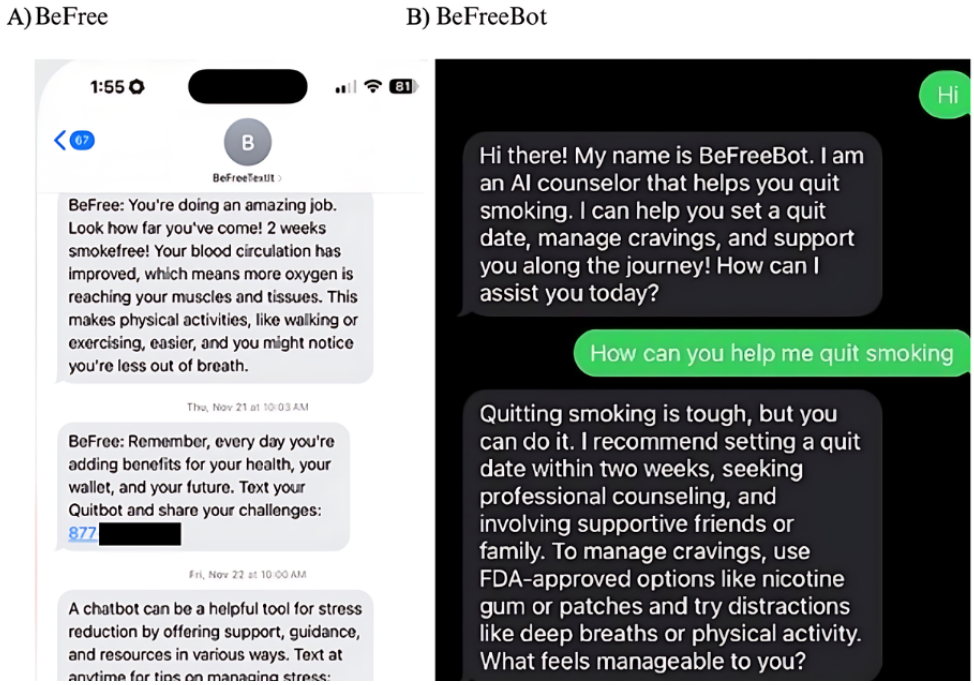

Measures

Overview

Demographic information (eg, age, gender, race and ethnicity, and educational level) and smoking characteristics were assessed on the baseline survey. Smoking characteristics included the average number of cigarettes smoked per day, whether they had had a past-year quit attempt (yes or no), nicotine dependence measured using the Fagerstrom Test for Nicotine Dependence [26], and an assessment of other tobacco products used.

Primary Outcome

The primary outcome for this study was BeFreeBot feasibility, which was operationalized as BeFreeBot engagement. Successful engagement was defined a priori as whether participants engaged with BeFreeBot by texting it a question or concern at least one time. The threshold of at least one text to the chatbot was chosen to reflect that this behavior, texting the chatbot, was somewhat complicated for users. Due to technical limitations of the system, users could not simply reply via text to the originating BeFree message to access the chatbot. Instead, they had to text the chatbot by texting *a different number*, requiring an additional step. Other metrics of engagement were the total number of texts sent to BeFreeBot by participants and the number of days of texting BeFreeBot. The data for these measures came from BeFreeBot computer records, which were accessed via the study account on Twilio (Twilio Inc), the SMS text messaging system.

Secondary Outcomes

A secondary outcome was the acceptability of BeFreeBot to participants. This outcome was assessed through participants' ratings of the program on the 4-week follow-up survey. This included indicators on a 5-point Likert scale of agreement (eg, 1="not at all"; 5="completely"), with statements about satisfaction with BeFreeBot, the helpfulness of BeFreeBot ("I think that BeFreeBot is competent and effective in helping people quit smoking"), their trust in responses from BeFreeBot, and whether they would recommend BeFreeBot to a friend. A rating of 4 or 5 was coded as "in agreement," and the percentage that agreed was calculated. Open-ended questions soliciting suggestions for improving the program were also asked to participants. One item also assessed the acceptability of the BeFree program overall (ie, the outgoing automated texts and the chatbot).

Exploratory Outcomes

Also of interest for exploratory purposes were smoking-related outcomes measured at the 4-week follow-up: smoking status

XSL•FO
RenderX

using 7-day point prevalence abstinence and past-month quit attempts, defined as not smoking for at least 24 hours because of trying to quit. In addition, changes in the number of cigarettes smoked per day from baseline to follow-up were measured.

Chatbot Fidelity

Finally, of interest was an assessment of chatbot fidelity, that is, whether the chatbot responses to participants were in accordance with its instructions. Measures of chatbot fidelity included whether the chatbot followed its instructions in the following ways: (1) introduced itself before answering the first question, (2) provided responses that were less than 50 words long, (3) offered responses in clear language, (4) showed markers of empathy in more than half of the responses to a participant, and (5) ended the responses to a question with a follow-up question to stimulate engagement. Related to smoking cessation guidelines [3], the chatbot was assessed for whether it (1) provided information to each participant in the course of the conversation on how to handle cravings using behaviors (eg, at least one strategy provided on how to handle cravings, such as by replacing smoking with deep breaths and physical activity, distracting the hands and mind, and changing the routine), (2) provided a recommendation to each participant in the course of the conversation to seek out professional counseling (eg, a physician or a quitline), (3) provided information to each participant on seeking out peer social support, (4) recommended to each participant to consider using NRT or non-NRT prescription medications, and (5) stayed on the topic of quitting smoking (even when participants brought up other topics). In addition, the chatbot responses were assessed for not providing misinformation, defined as advice on quitting smoking outside of USPSTF guidelines [3]. Finally, the chatbot was assessed for not having technical errors (eg, not providing a response at all or giving an error message).

Analysis

Means and proportions were calculated for outcomes including engagement (feasibility), acceptability, and smoking-related outcomes. Dichotomous smoking outcomes (eg, 7-day point prevalence abstinence) were calculated through self-report at 4 weeks and using missing data imputed as smoking. Responses to the open-ended item soliciting suggestions for improving the program were grouped thematically and summarized.

To conduct the analysis of BeFreeBot instruction adherence (fidelity), the BeFreeBot conversations for the 16 participants who engaged with BeFreeBot were exported from Twilio and analyzed across adherence items. Coders were trained to understand our measures of chatbot fidelity, most of which had been used in a previous study that examined chatbot adherence to the USPSTF guidelines [18]. Two independent coders (CW and AY) were instructed to read the complete conversation transcript for an individual and chatbot and then make inferences about the items of interest and code the items. This was done for all the aforementioned items except for the chatbot introducing itself and providing responses less than 50 words long. Where discrepancies arose, a third coder (TCW) served as a tiebreaker. Across 16 participant conversations and 9 items, the overall agreement was 86%, and the average κ was 0.68 (SD=0.23) (good or substantial agreement) [27]. The items related to whether the chatbot introduced itself (yes or no) and response length of <50 words were coded by only 1 coder (AY) as these items did not have a subjective component.

Results

Background

A total of 87 MTurk workers completed the eligibility survey. Of those 87 workers, 57 (66%) were ineligible (n=25, 44% because they had not smoked in the previous 7 days; n=24, 42% because they were not willing to receive SMS text messages; n=5, 9% because they were duplicates; and n=3, 5% because they did not pass the attention check question [categories not mutually exclusive]). After consenting, of the remaining 30 participants, 7 (23%) additional participants were excluded (n=6, 86% because they did not complete the baseline survey and enrollment procedures after consent and n=1, 14% because they had an invalid phone number), leaving 23 (77%) participants enrolled in the study. At 4 weeks, 83% (19/23) of the participants completed the follow-up survey.

Baseline Demographic and Smoking Characteristics

On average, participants were aged 50 (SD 12.4) years, with an approximately even distribution between male and female individuals and most having attended some college or higher (19/23, 83%; Table 1). Participants were largely of White race (20/23, 87%), and 13% (3/23) were of Hispanic origin. The average number of cigarettes smoked per day at baseline was 11.2 (SD 6.8), the average Fagerstrom Test for Nicotine Dependence score was 3.6 (SD 2.6), and just over a third of the participants (8/23, 35%) had made a quit attempt in the previous year. Other tobacco products used over the previous week included e-cigarettes (7/23, 30%) and, to a lesser extent, other combustible products (eg, cigars; 2/23, 9%) and smokeless products (2/23, 9%).

XSL•FO
RenderX

Table 1. Baseline demographic and smoking characteristics of the participants (N=23).

Characteristic	Values
Age (y), mean (SD)	50.0 (12.4)
Sex, n (%)	
Male	12 (52)
Female	11 (48)
Hispanic, Latino, or Spanish origin, n (%)	
Yes	3 (13)
No	20 (87)
Race, n (%)	
Black or African American	3 (13)
White	20 (87)
Other	0 (0)
Educational level, n (%)	
High school or lower	4 (17)
Some college or trade school	9 (39)
College graduate or higher	10 (43)
Employment status, n (%)	
Unemployed	2 (9)
Employed part time	2 (9)
Employed full time	19 (83)
Marital status, n (%)	
Single	5 (22)
Married	9 (39)
Other (living with partner, divorced, separated, or widowed)	9 (39)
Cigarettes per day, mean (SD)	11.2 (6.8)
FTND[a] score (0-10), mean (SD)	3.6 (2.6)
Smoking within 30 minutes of waking, n (%)	14 (61)
Quit attempt in the previous year, n (%)	8 (35)
Other tobacco products used in the previous week, n (%)	
e-cigarettes	7 (30)
Other combustible products (cigars, cigarillos, pipes, bidis, or hookah)	2 (9)
Smokeless products (dissolvable or chewed tobacco)	2 (9)

[a]FTND: Fagerstrom Test for Nicotine Dependence.

Participant Engagement, Perceptions, and Smoking-Related Outcomes

Most participants texted at least one question to the BeFreeBot (16/23, 70%; Table 2). In all, there were 661 text exchanges between BeFreeBot and the participants, with 333 (50.4%) texts sent to BeFreeBot by the participants, or 14.5 (SD 23.6) texts per participant on average, and 328 (49.6%) replies from BeFreeBot, or 14.3 (SD 23.0) replies per participant on average. Of the 16 participants who texted BeFreeBot at least once, 12 (75%) sent between 1 and 29 texts, and 4 (25%) sent between 30 and 96 texts. Text conversations with BeFreeBot spanned 3.9 (SD 4.4) days on average, with half of participants (8/16, 50%) engaging in conversation only during the first 2 weeks of enrollment in the program. In total, 9% (2/23) of the participants unsubscribed from the program using the keyword “STOP” before the end of the 4-week trial.

XSL•FO
RenderX

Table 2. Participant engagement, perceptions, and smoking-related outcomes (4 weeks; N=23).

	Values
Engagement	
Texted the chatbot at least once, n/N (%)	16/23 (70)
Number of texts sent to BeFreeBot, mean (SD; range)	14.5 (23.6; 0-96)
Responses sent by BeFreeBot, mean (SD; range)	14.3 (23.0; 0-92)
Days texting BeFreeBot, mean (SD)	3.9 (4.4)
Acceptability, n/N (%)	
Highly satisfied with the program overall	13/19 (68)
Highly satisfied with BeFreeBot	13/18 (72)
BeFreeBot was helpful	11/19 (58)
BeFreeBot responses were easy to understand	16/17 (94)
BeFreeBot could answer all questions	12/17 (71)
Would recommend BeFreeBot to a friend	13/18 (72)
Trusted responses from BeFreeBot	12/17 (71)
Preferred an AI[a] counselor over a human one	7/19 (37)
Smoking-related outcomes	
Self-report of no smoking in the previous 7 days, n/N (%)	7/19 (37)
Self-report of no smoking in the previous 7 days (missing imputed as smoking), n/N (%)	7/23 (30)
Change in cigarettes smoked per day, mean (SD)	4.5 (7.8)
Past-month quit attempt (yes), n/N (%)	17/19 (90)

[a]AI: artificial intelligence.

Participants generally reported overall satisfaction with the BeFree program (including BeFreeBot; 13/19, 68%) and with BeFreeBot specifically (13/18, 72%). With regard to BeFreeBot, most participants reported that the BeFreeBot responses were easy to understand (16/17, 94%), that they found the bot helpful for quitting smoking (11/19, 58%), that it could answer all questions about quitting (12/17, 71%), and that they would recommend the program to a friend (13/18, 72%). Most participants stated that they trusted the responses from BeFreeBot (12/17, 71%), although less than half endorsed the idea that they preferred an AI counselor over a human one (7/19, 37%). At the end of the study, 90% (17/19) of the participants reported a quit attempt in the previous month, and 30% (7/23) reported being abstinent from smoking for the previous 7 days. Participants reduced their smoking by 4.5 (SD 7.8) cigarettes per day on average. The most common suggestions for improving the program were enabling direct replies to texts to access BeFreeBot (eg, not requiring having to text a different number to access BeFreeBot) and having BeFreeBot have greater conversational depth and less small talk.

BeFreeBot Fidelity

An analysis of transcripts of BeFreeBot responses to participants revealed that, generally, BeFreeBot functioned as directed. Participant queries covered a variety of topics, including how to quit smoking; medications for quitting; how to handle specific stressors; and how to use meditation, food, or music to manage cravings. On the basis of the coding analysis, responses (n=328) used clear language (328/328, 100%), had markers of empathy (328/328, 100%), and promoted engagement with additional follow-up questions to participants (13/16, 81%, See Table 3). Most responses to participants included a recommendation to seek out professional counseling (13/16, 81%) or find peer support (11/16, 69%), recommended considering the use of FDA-approved medications (eg, NRT; 14/16, 88%), and provided information on how to manage cravings (13/16, 81%). Responses stayed on the topic of smoking cessation counseling (324/328, 99%) and did not include information that contradicted the USPSTF guidelines (328/328, 100%) or any alarming responses (eg, a response that would result in harm). Responses were mostly at the appropriate length of less than 50 words (309/328, 94%), and there were few technical errors (10/328, 3%). In the case of technical errors, these consisted of the chatbot replying “error” or not replying at all. One area in which instructions were not followed regarded BeFreeBot introducing itself to participants when first accessed. This occurred less than half (7/16, 44%) of the time.

XSL•FO
RenderX

Table 3. Results on whether BeFreeBot functioned as directed.[a]

	Instances, n/N (%)
Responses used clear language	328/328 (100)
Showed empathy in more than half of the responses to each participant[b]	16/16 (100)
Promoted engagement after answering the initial question[b]	13/16 (81)
Recommended professional counseling to each participant[b]	13/16 (81)
Recommended FDA[c]-approved medications to each participant[b]	14/16 (88)
Recommended seeking out peer social support to each participant[b]	11/16 (69)
Provided information on how to manage cravings to each participant[b]	13/16 (81)
Responses stayed on the topic of smoking cessation	324/328 (99)
No information that contradicted USPSTF[d] guidelines or no hallucinations	328/328 (100)
Introduced itself before answering the first question[b]	7/16 (44)
Responses were <50 words long	309/328 (94)
Responses had a technical error	10/328 (3)

[a]328 total BeFreeBot responses or messages across 16 participant conversations.

[b]Analyses conducted at the participant level.

[c]FDA: Food and Drug Administration.

[d]USPSTF: US Preventive Services Task Force.

Discussion

Principal Findings

ChatGPT and other LLM chatbots, while increasingly used by the public, have been understudied in the context of smokers trying to quit. In this study, we found that an LLM chatbot integrated into an SMS text messaging program and accessed through SMS text message was found to be feasible and acceptable to smokers for quitting. While participants were required to text a new number to access BeFreeBot, a somewhat cumbersome procedure, most participants (16/23, 70%) managed this and texted BeFreeBot at least once with a question. Participants generally expressed satisfaction with BeFreeBot (13/18, 72%), stating that it was easy to understand (16/17, 94%) and they trusted its responses (12/17, 71%). Quitting outcomes were promising, with most participants (17/19, 90%) making a quit attempt and almost one-third of the participants (7/23, 30%) reporting that they had quit.

The frequency of participant engagement with BeFreeBot is noteworthy given that texting an LLM chatbot is a novel format whose feasibility has not yet been assessed. Not only did we find that this format was feasible, but we also found high levels of engagement. While, on average, participants texted BeFreeBot 14.3 (SD 23.0) times, the range was considerable, with a couple of participants (2/23, 9%) texting BeFreeBot over 50 times (and receiving over 50 texts back). This level of engagement is encouraging for SMS text messaging programs as automated SMS text messaging programs generally offer short and formulaic forms of engagement (eg, texting the keyword “CRAVE” for a tip on managing cravings) [28]. Indeed, the chatbot’s ability to understand natural text and provide personalized responses to queries on topics such as specific stressors appears to have enabled a longer form of engagement on a given topic and possibly a deeper form of engagement than that with previous formulaic SMS text messaging programs. As program engagement has been linked to health behavior change [28,29], chatbots may be effective by boosting engagement. Studies are needed to further explore this possibility, including exploring the role that literacy may play in text-based chatbot engagement.

Our study also identified areas for design improvements. While text engagement was frequent, it was also short-lived. Conversations with BeFreeBot took place, on average, on 3.9 (SD 4.4) of 30 possible days of the study, with half of participants only engaging in conversation during the first 2 weeks after enrollment. This aspect of engagement could be improved by providing participants with additional prompts to engage with the chatbot and suggested conversation starters (eg, “Ask BeFreeBot about how to deal with your biggest trigger for smoking”). Indeed, we sent the last engagement prompt at 2 weeks, and such prompts could be extended in time for a longer period. In addition, the most common open-ended suggestion from participants was enabling direct replies to texts to access BeFreeBot (eg, not requiring having to text a different number to access BeFreeBot). Future efforts will be aimed at identifying a technical solution so that BeFree and BeFreeBot can be integrated into 1 phone number to improve the user experience. Indeed, as the technology is quickly evolving, new design possibilities will also be explored. These include possibilities that go beyond SMS text messaging and allow for the inclusion of generative images and videos and the use of voice instead or in addition to text, which might address literacy concerns.

XSL•FO
RenderX

Another finding of our study is that our chatbot, for the most part, adhered to our instructions and sent responses that followed them. An analysis of transcripts of the BeFreeBot responses to participants revealed that BeFreeBot stayed on the topic of quitting smoking even when participant queries veered outside of this topic. The chatbot also showed at least minimal levels of empathy and promoted engagement by asking follow-up questions, as well as adhering to recommendations consistent with USPSTF guidelines [3]. This preliminary finding is important given concerns published in the literature about the presence of hallucinations and misinformation in chatbot responses [30] and general concerns about the ethics of having a generative AI chatbot as a counselor or coach [31,32].

Indeed, to achieve the observed levels of adherence, our team dedicated considerable effort before the trial to validating, refining, and revising our instructions [18]. Still, some areas for improvement were identified, such as in following the direction to introduce itself during the first interaction with a participant, which the chatbot did less than half of the time. In the future, BeFreeBot instructions will be further refined to emphasize following them and ensure that they are followed for a greater variety of responses.

The strength of this study is that it represents one of the first to study a specialized ChatGPT-based chatbot with people who smoke and were willing to set a quit date. It is also the first study to our knowledge of a chatbot accessed through an automated SMS text messaging program. Results show that the chatbot and its SMS text messaging interface were feasible and mostly appreciated by smokers. These findings are important because ChatGPT has potential as a scalable tool in reaching the public with help for quitting smoking. Our findings suggest that such adaptations hold potential for supporting smoking cessation efforts.

The limitations of our study are that it was a pilot study with no control group. The small sample size implies that the chatbot only received a small number of queries, and the possibility remains that additional queries could have led the chatbot off topic or to diverge from its instructions in other ways. Future research that includes a larger sample, as well as a comparison group, will be able to test whether chatbots offer benefits for quitting beyond those of standard digital programs. Another limitation of this study is its generalizability. Apart from the small sample size, participants recruited on MTurk represent a specific type of employed person. By the nature of the MTurk work, workers likely have access to a PC and work at a desk. In our sample, participants were predominantly of White ethnicity and college educated. Future research is needed to demonstrate effects among a more diverse sample, including individuals with low levels of education and those with low literacy levels as these characteristics may impact engagement with a text-based program. While we created a crude metric for measuring chatbot adherence to instructions, especially in the measurement of empathy, more nuanced metrics could be developed. Furthermore, additional procedures could be put in place to ensure that the fidelity assessment team is independent from the development team, which in our pilot had some overlap and could introduce bias. Finally, our measure of acceptability could be improved by using a validated scale.

Conclusions

In conclusion, an LLM chatbot integrated into an SMS text messaging program and accessed through SMS text message was found to be feasible and acceptable to smokers. Future iterations of the program will be aimed at increasing engagement with BeFreeBot, facilitating conversations that span more days, and rigorously testing the program's efficacy in producing long-term benefits for quitting smoking. As the program is scaled to larger samples with more diverse backgrounds, a key consideration will be how to ensure that the LLM chatbot continues to provide evidence-based information and benefits to all groups of users.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

References

1. National Center for Chronic Disease Prevention and Health Promotion Office on Smoking and Health. The health consequences of smoking-50 years of progress: a report of the Surgeon General. National Library of Medicine. 2014. URL: https://www.ncbi.nlm.nih.gov/books/NBK179276/ [accessed 2025-05-29]
2. Cornelius ME, Loretan CG, Wang TW, Jamal A, Homa DM. Tobacco product use among adults — United States, 2020. MMWR Morb Mortal Wkly Rep 2022 Mar 18;71(11):397-405. [doi: 10.15585/mmwr.mm7111a1]
3. Krist AH, Davidson KW, Mangione CM, Barry MJ, Cabana M, Caughey AB, et al. Interventions for tobacco smoking cessation in adults, including pregnant persons: US preventive services task force recommendation statement. JAMA 2021 Jan 19;325(3):265-279. [doi: 10.1001/jama.2020.25019] [Medline: 33464343]
4. Dobson R, Whittaker R, Abroms LC, Bramley D, Free C, McRobbie H, et al. Don't forget the humble text message: 25 years of text messaging in health. J Med Internet Res 2024 Dec 17;26:e59888 [FREE Full text] [doi: 10.2196/59888] [Medline: 39689299]

XSL•FO
RenderX

5. Whittaker R, McRobbie H, Bullen C, Rodgers A, Gu Y, Dobson R. Mobile phone text messaging and app-based interventions for smoking cessation. Cochrane Database Syst Rev 2019 Oct 22;10(10):CD006611 [FREE Full text] [doi: 10.1002/14651858.CD006611.pub5] [Medline: 31638271]
6. Abroms LC, Boal AL, Simmens SJ, Mendel JA, Windsor RA. A randomized trial of Text2Quit: a text messaging program for smoking cessation. Am J Prev Med 2014 Sep;47(3):242-250 [FREE Full text] [doi: 10.1016/j.amepre.2014.04.010] [Medline: 24913220]
7. Cole-Lewis H, Kershaw T. Text messaging as a tool for behavior change in disease prevention and management. Epidemiol Rev 2010 Mar 30;32(1):56-69 [FREE Full text] [doi: 10.1093/epirev/mxq004] [Medline: 20354039]
8. Karekla M, Savvides SN, Gloster A. An avatar-led intervention promotes smoking cessation in young adults: a pilot randomized clinical trial. Ann Behav Med 2020 Oct 01;54(10):747-760. [doi: 10.1093/abm/kaaa013] [Medline: 32383736]
9. Abdullah AS, Gaehde S, Bickmore T. A tablet based embodied conversational agent to promote smoking cessation among veterans: a feasibility study. J Epidemiol Glob Health 2018 Dec;8(3-4):225-230 [FREE Full text] [doi: 10.2991/j.jegh.2018.08.104] [Medline: 30864768]
10. He L, Balaji D, Wiers RW, Antheunis ML, Krahmer E. Effectiveness and acceptability of conversational agents for smoking cessation: a systematic review and meta-analysis. Nicotine Tob Res 2023 Jun 09;25(7):1241-1250 [FREE Full text] [doi: 10.1093/ntr/ntac281] [Medline: 36507916]
11. He L, Basar E, Wiers RW, Antheunis ML, Krahmer E. Can chatbots help to motivate smoking cessation? A study on the effectiveness of motivational interviewing on engagement and therapeutic alliance. BMC Public Health 2022 Apr 12;22(1):726 [FREE Full text] [doi: 10.1186/s12889-022-13115-x] [Medline: 35413887]
12. Whittaker R, Dobson R, Garner K. Chatbots for smoking cessation: scoping review. J Med Internet Res 2022 Sep 26;24(9):e35556 [FREE Full text] [doi: 10.2196/35556] [Medline: 36095295]
13. Perski O, Crane D, Beard E, Brown J. Does the addition of a supportive chatbot promote user engagement with a smoking cessation app? An experimental study. Digit Health 2019 Sep 30;5:2055207619880676 [FREE Full text] [doi: 10.1177/2055207619880676] [Medline: 31620306]
14. Benke K, Benke G. Artificial intelligence and big data in public health. Int J Environ Res Public Health 2018 Dec 10;15(12):2796 [FREE Full text] [doi: 10.3390/ijerph15122796] [Medline: 30544648]
15. Biswas SS. Role of Chat GPT in public health. Ann Biomed Eng 2023 May 15;51(5):868-869. [doi: 10.1007/s10439-023-03172-7] [Medline: 36920578]
16. Ayers JW, Zhu Z, Poliak A, Leas EC, Dredze M, Hogarth M, et al. Evaluating artificial intelligence responses to public health questions. JAMA Netw Open 2023 Jun 01;6(6):e2317517 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.17517] [Medline: 37285160]
17. Sidoti O, McClain O. 34% of U.S. adults have used ChatGPT, about double the share in 2023. Pew Research Center. URL: https://www.pewresearch.org/short-reads/2025/06/25/34-of-us-adults-have-used-chatgpt-about-double-the-share-in-2023/ [accessed 2025-05-29]
18. Abroms LC, Yousefi A, Wysota CN, Wu T, Broniatowski DA. Assessing the adherence of ChatGPT chatbots to public health guidelines for smoking cessation: content analysis. J Med Internet Res 2025 Jan 30;27:e66896 [FREE Full text] [doi: 10.2196/66896] [Medline: 39883917]
19. S.A.R.A.H. smart AI resource assistant for health. World Health Organization. URL: https://www.who.int/campaigns/s-a-r-a-h [accessed 2025-05-29]
20. Bricker JB, Sullivan B, Mull K, Santiago-Torres M, Lavista Ferres JM. Conversational chatbot for cigarette smoking cessation: results from the 11-step user-centered design development process and randomized controlled trial. JMIR Mhealth Uhealth 2024 Jul 23;12:e57318 [FREE Full text] [doi: 10.2196/57318] [Medline: 38913882]
21. Calle P, Shao R, Liu Y, Hébert ET, Kendzor D, Neil J, et al. Towards AI-driven healthcare: systematic optimization, linguistic analysis, and clinicians' evaluation of large language models for smoking cessation interventions. Proc SIGCHI Conf Hum Factor Comput Syst 2024 May;2024:436 [FREE Full text] [doi: 10.1145/3613904.3641965] [Medline: 38912297]
22. Foundry models sold directly by Azure. Microsoft. URL: https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/models?utm_source=chatgpt.com&tabs=global-standard%2Cstandard-chat-completions [accessed 2025-06-19]
23. Azure OpenAI Assistants API (Preview). Microsoft Inc. URL: https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/assistants [accessed 2025-06-19]
24. Clearing the air. U.S. Department of Health and Human Services. 2025. URL: https://www.cancer.gov/publications/patient-education/clearing-the-air-pdf [accessed 2025-06-19]
25. Want to quit smoking? FDA-approved and FDA-cleared cessation products can help. US Food and Drug Administration. URL: https://www.fda.gov/consumers/consumer-updates/want-quit-smoking-fda-approved-and-fda-cleared-cessation-products-can-help [accessed 2025-06-19]
26. Weinberger AH, Reutenauer EL, Allen TM, Termine A, Vessicchio JC, Sacco KA, et al. Reliability of the Fagerström test for nicotine dependence, Minnesota Nicotine Withdrawal Scale, and tiffany questionnaire for smoking urges in smokers with and without schizophrenia. Drug Alcohol Depend 2007 Jan 12;86(2-3):278-282. [doi: 10.1016/j.drugalcdep.2006.06.005] [Medline: 16876968]

XSL•FO
RenderX

27. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22(3):276-282 [FREE Full text] [Medline: 23092060]
28. Abroms LC, Heminger CL, Boal AL, Van Alstyne JM, Krishnan N. Text2Quit: an analysis of user experiences with a mobile smoking cessation program. J Smok Cessat 2020 Mar;15(1):23-28 [FREE Full text] [doi: 10.1017/jsc.2019.22] [Medline: 33777241]
29. Kim SJ, Marsch LA, Brunette MF, Dallery J. Harnessing Facebook for smoking reduction and cessation interventions: Facebook user engagement and social support predict smoking reduction. J Med Internet Res 2017 May 23;19(5):e168 [FREE Full text] [doi: 10.2196/jmir.6681] [Medline: 28536096]
30. Alkaissi H, McFarlane SI. Artificial hallucinations in ChatGPT: implications in scientific writing. Cureus 2023 Feb;15(2):e35179 [FREE Full text] [doi: 10.7759/cureus.35179] [Medline: 36811129]
31. Tavory T. Regulating AI in mental health: ethics of care perspective. JMIR Ment Health 2024 Sep 19;11:e58493 [FREE Full text] [doi: 10.2196/58493] [Medline: 39298759]
32. Wang C, Liu S, Yang H, Guo J, Wu Y, Liu J. Ethical considerations of using ChatGPT in health care. J Med Internet Res 2023 Aug 11;25:e48009 [FREE Full text] [doi: 10.2196/48009] [Medline: 37566454]

Abbreviations

AI: artificial intelligence
FDA: Food and Drug Administration
HIT: Human Intelligence Task
LLM: large language model
MTurk: Amazon Mechanical Turk
NRT: nicotine replacement therapy
SARAH: Smart Artificial Intelligence Resource Assistant for Health
USPSTF: United States Preventive Services Task Force

Edited by A Mavragani; submitted 20.06.25; peer-reviewed by O Melamed, J Heffner, D Pócs; comments to author 03.07.25; revised version received 22.08.25; accepted 05.09.25; published 10.10.25.

Please cite as:
Abroms LC, Wysota CN, Yousefi A, Wu TC, Broniatowski DA
ChatGPT-Based Chatbot for Help Quitting Smoking via Text Messaging: An Interventional Study
JMIR Form Res 2025;9:e79402
URL: https://formative.jmir.org/2025/1/e79402
doi:10.2196/79402
PMID:

©Lorien C Abroms, Christina N Wysota, Artin Yousefi, Tien-Chin Wu, David A Broniatowski. Originally published in JMIR Formative Research (https://formative.jmir.org), 10.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Virtual Reality–Based mHealth App to Enhance Health Care Skills and Promote Health Equity by Empowering Health Care Professionals to Provide Empathetic, Compassionate, and Unbiased Patient Care Through Digital Experiential Learning: Qualitative Study

Dixit Bharatkumar Patel[1*], PhD; Mayank Bharatkumar Patel[2*], MS; Thomas Wischgoll[1], PhD; Yong Pei[3], PhD; Paul J Hershberger[4], PhD

[1]Department of Computer Science & Engineering, Wright State University, Dayton, OH, United States
[2]Fabricators Incorporated, Chattanooga, TN, United States
[3]Department of Computer Science, Kennesaw State University, Marietta, GA, United States
[4]Department of Family Medicine, Wright State University, Dayton, OH, United States
[*]these authors contributed equally

Corresponding Author:
Dixit Bharatkumar Patel, PhD
Department of Computer Science & Engineering
Wright State University
3640 Colonel Glenn Hwy
Dayton, OH, 45435
United States
Phone: 1 937 775 5131
Email: patel.837@wright.edu

Abstract

Background: Health care professionals' educational preparation and practices significantly influence care experiences and health outcomes. Deficient awareness of the impact of stereotypes, biases, prejudices, and social determinants of health (SDH) can lead to negative care experiences, strained health care professional-patient relationships, and health disparities. Addressing these challenges necessitates enhancing health care professionals' skills, including inclusive communication, cultural humility, recognition of SDH, and fostering empathy and compassion, promoting health equity and better care experiences.

Objective: This research aims to introduce a mobile health (mHealth) app designed using a digital experiential learning (DEL) approach to strengthen health care professionals' competencies, thereby improving patient care experiences and promoting health equity. The key objectives are to deliver essential health care skills, such as cultivating cultural humility, developing inclusive communication proficiencies, understanding the lasting impact of SDH, comprehending how implicit and explicit biases affect health outcomes, fostering compassionate and empathetic clinical attitudes, and promoting continuous professional development. Together, these aims advance patient-centered care and help reduce health disparities.

Methods: The mHealth app integrates virtual reality–based serious role-playing hypothetical scenarios and a life course module to provide health care professionals with immersive first-person learning experiences. The scenarios include a Syrian refugee with limited English proficiency and an African American pregnant woman with a history of opioid use disorder, each lasting ≈30 minutes to deliver essential health care skills. The pre- and postassessment questionnaires are integrated within the app to measure the learning outcomes of diverse professionals who voluntarily engaged with the app. Distinct hypotheses were formulated and evaluated, referring to specific questionnaire items to assess the app's impact on professionals' skills and attitudes.

Results: The mHealth app significantly enhanced health care professionals' skills and attitudes, including increased confidence, preparedness for patient interactions, awareness of the lasting impact of SDH, positive beliefs, reduced prejudice, improved perspectives on patient responsibility and external factors, as well as greater compassion and empathy. These outcomes, obtained through evaluating distinct hypotheses and analysis supported by multiple statistical approaches, including CI analysis, Cohen *d* effect sizes, odds ratios, 1-tailed paired *t* tests, and descriptive response distributions, directly align with the study's objectives,

XSL•FO
RenderX

fostering health equity and patient-centered care. Overall, the app effectively improved health care professionals' competencies, contributing to better care experiences and outcomes while promoting health equity.

Conclusions: This study delivers a comprehensive data-driven evaluation that validates the effectiveness of the mHealth app in enhancing health care professionals' skills and fostering health equity. It addresses challenges in health care education by delivering a scalable, accessible, and immersive learning platform. Supported by virtual reality technology and an experiential learning framework, the mHealth app presents a promising avenue to empower health care professionals with skills to provide patient-centered care in diverse, complex settings.

(JMIR Form Res 2025;9:e64147) doi:10.2196/64147

KEYWORDS

mHealth app; experiential learning; health care skills; health care; mobile health; serious games; health equity; virtual reality; role-playing games; health care professionals

Introduction

Background of Study

The Centers for Medicare & Medicaid Services NHE (National Health Expenditure) Fact Sheet highlights the growing demand for health care [1], emphasizing the need for equitable access. It is essential that individuals receive appropriate care without the influence of implicit or explicit biases stemming from sociodemographic factors such as race, ethnicity, sex, age, poverty, disability status, or geographic location [2].

Research shows that health care professionals are vulnerable to biases, which contribute significantly to health disparities and require urgent attention [3]. These biases, particularly toward those outside one's social circle, tend to form over time and are frequently negative in nature [4]. If left unaddressed, such biases can result in inequitable treatment of marginalized populations [5,6]. For example, Hoffman et al [7] found that African Americans may receive less pain medication compared with White Americans, highlighting how bias affects clinical decisions.

In addition to genetics and lifestyle choices, social determinants of health (SDH)—defined as the conditions in which people are born, grow, work, live, age, and engage in recreation—have a significant impact, accounting for 30% to 55% of health outcomes [8,9]. Health care providers who acknowledge their biases and apply skills such as perspective-taking and individuation can help reduce disparities [10-14]. Therefore, it is crucial to raise awareness among physicians, physician assistants, social workers, nurses, pharmacists, and other health care professionals about essential patient-centered skills. These include cultural humility, inclusive communication, understanding the impact of SDH, recognizing how bias affects health outcomes, and promoting empathy. Such competencies improve patient experiences, minimize biased health care encounters, and advance health equity.

Literature Review

In health care education, developing intellectual capacity and behavioral awareness is essential for fostering equitable and patient-centered care. Education plays a vital role not only in enhancing logical reasoning but also in shaping individuals' beliefs and social interactions, which are closely tied to how they perceive and respond to diverse patient populations. Hands-on learning experiences play a significant role in elevating an individual's competency level [15].

One foundational approach to hands-on education is experiential learning theory, introduced in 1984 by educational theorist Kolb [16]. Experiential learning theory integrates the learning styles inventory and the experiential learning cycle to emphasize that knowledge is created through the transformation of experience. Kolb's work [16], influenced by researchers such as John Dewey, Kurt Lewin, and Jean Piaget, highlights that learning involves grasping and transforming experiences, a concept especially relevant in training health care professionals to engage in self-awareness, perspective-taking, and bias reduction.

As illustrated in Figure 1, the experiential learning cycle, learners begin by engaging in direct experiences, reflect upon them, form abstract generalizations, and apply these insights through experimentation—fostering deeper understanding and skill acquisition [16]. Integrating this model into digital platforms empowers learners to apply knowledge in simulated environments and clinical scenarios.

XSL•FO
RenderX

Figure 1. Illustration of Kolb's experiential learning cycle [16], depicting the 4 stages of learning—Concrete Experience, Reflective Observation, Abstract Conceptualization, and Active Experimentation—central to the design of immersive health care education experiences.

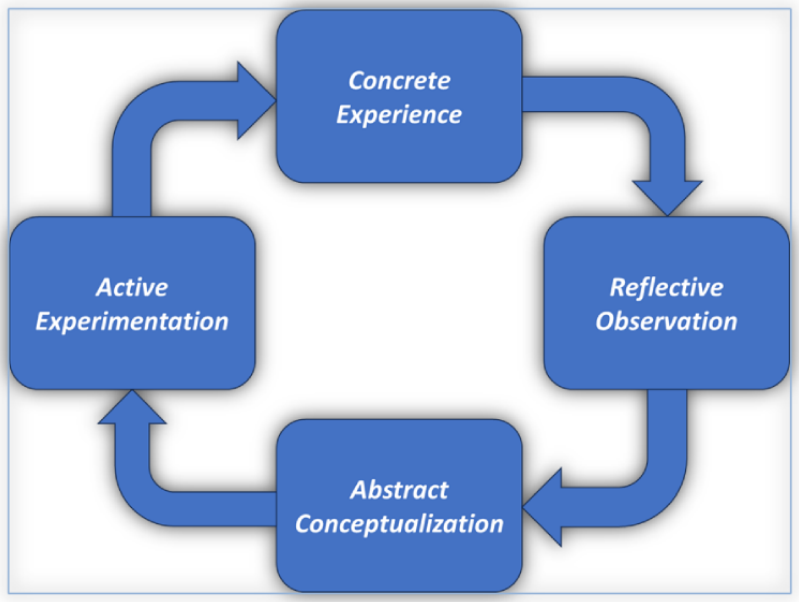

Hence, in this study, the "Digital Experiential Learning" (hereinafter referred to as DEL) term refers to an approach that harnesses digital technologies to create immersive and hands-on learning experiences. It integrates the principles of experiential learning [16], where learners actively engage with real-world tasks and reflect on their experiences, with digital tools and platforms to enrich the learning process. This approach uses various digital mediums such as simulations, virtual environments, interactive multimedia, and web-based collaborative platforms to facilitate exploration of complex concepts, development of practical skills, and acquisition of meaningful insights. The efficacy of DEL has been demonstrated in fields such as tourism education, where virtual field trips and simulations significantly enhanced learners' socioemotional engagement and content retention [17], and in sustainable horticulture and landscape management, where augmented and virtual reality (VR) technologies promoted team-based learning and environmental awareness within Kolb's experiential learning framework [18]. While DEL promotes active participation, problem-solving, and critical thinking, it also presents challenges such as access to technology and the need for digital literacy. Nonetheless, its dynamic and interactive nature enhances understanding and retention of knowledge across diverse educational settings.

Within this digital landscape, serious games have emerged as powerful educational tools. Serious games are digital games in which hidden or serious objectives are seamlessly integrated into the digital app [19]. These games embed learning objectives into engaging formats, providing motivational and learning experiences [19]. Thus, serious games are a powerful concept that can be used to deliver motivational and empathetic learning experiences for individuals within specific fields, including health care. Enhancing this notion further, role-playing games (RPGs) allow learners to embody characters in simulated settings, deepening empathy, intuition, and contemplation—particularly in clinical training [20,21]. Digital RPGs have shown success in teaching educational and social skills [22,23], and their adoption in occupational universities has improved standards in technology-enhanced learning [24]. Building on these concepts, Patel et al [25-27] have demonstrated the value of computer-supported expert-guided experiential and cognitive learning tools or mobile health (mHealth) apps for training health care professionals. These tools help cultivate nonverbal communication and behavior-modification skills crucial for interactions with individuals on the autism spectrum and in general health care contexts [25-27].

The evolution of technologies has accelerated, driving a shift from traditional learning to technology-based learning in higher education, including health care education [28]. Studies reveal that perceived usefulness, ease of use, and health benefits influence user adoption of mHealth apps [29]. Innovations like the integration of wearable devices and smartphones enable real-time health management and monitoring [30].

Several reviews have evaluated mHealth apps in health care delivery. Ameyaw et al [31] highlighted their effectiveness in maternal health—addressing anxiety, diabetes, weight management, and behavioral risks—though most implementations occurred in high-income regions, underscoring a need for global equity. Iribarren et al [32] found modest benefits of mHealth over standard care but emphasized the variability in app impact and the importance of rigorous design and evaluation methods. Lee et al [33] reviewed physicians' use of mobile apps for communication, education, monitoring, and clinical decision-making, identifying both advantages and challenges related to privacy and distraction. Similarly, Liu et al [34] found that hypertension-focused mHealth tools, while patient-centered, also supported provider education and chronic disease management. The PsyCovidApp, studied by Fiol-DeRoque et al [35], demonstrated effectiveness in reducing health care workers' stress during the COVID-19

XSL•FO
RenderX

pandemic—benefiting performance and well-being. Jong et al [36] explored nurses' smartphone use, finding improvements in access to medical information, alongside concerns about distraction and data privacy.

Additionally, a concept of "educational game design based on experiential learning theory" was introduced by Li et al [37], suggesting that incorporating experiential learning elements into educational games can enhance students' motivation and interest in learning. However, while experiential learning—often summarized as "learning by doing"—is a proven educational approach [16], direct implementation in high-risk, real-world health care settings may not be appropriate or safe [38]. Inadequate learning experiences and insufficient training can lead to suboptimal care and weaken provider-patient relationships. Therefore, developing structured digital tools that facilitate DEL offers a safer, immersive, and more effective educational experience—empowering health care professionals to build critical competencies before applying them in clinical practice.

Study Objectives and an Innovative mHealth App Solution

In today's evolving health care landscape, equipping professionals with essential skills—such as inclusive communication, cultural humility, empathy, and an understanding of enduring impacts of SDH, biases, and prejudices on health outcomes—are critical to improving care quality. Providing coaching tools and expert-guided experience enhances professional competence and supports equitable, patient-centered health care delivery. Innovative approaches to experiential learning help make these resources more accessible—regardless of location, schedule, or professional role—thus democratizing health care education and empowering both providers and communities.

To address these challenges, this study integrates experiential learning into digital platforms—leveraging VR and serious RPG features—to support the development of health care skills. The objectives include the following. (1) Facilitating the delivery or enhancement of essential health care skills. (2) Addressing the challenge of distributing effective and efficient educational mediums in the health care sector. (3) Maximizing learning outcomes by providing a first-person experiential learning approach. (5) Developing a scalable product that can be accessed and used on mobile devices at any time and from anywhere, with only an internet connection and no additional costs. (6) Reducing the reliance on real-life standardized patients and expert trainers for educational and clinical practices. (7) Promoting continuous professional development for health care professionals. (8) Enhancing patient-centered care by fostering confidence and preparedness for real-life patient interactions, positive beliefs, compassion, empathy, and nonprejudiced attitude toward a patient or an individual. (9) Improving patients' health outcomes and overall health care quality through immersive and interactive learning experiences. (10) Increasing awareness and understanding of the impact of SDH on patient care. (11) Addressing implicit biases and promoting cultural humility in health care delivery. (12) Creating a safe and controlled learning environment for practicing high-risk care scenarios.

In addition, the app strengthens core clinical competencies through realistic case simulations. Specifically, the educational content supports the development of the following learning outcomes:

- Deepened understanding of patients' lived experiences
- Recognition of how SDH intensifies patient challenges
- Awareness of implicit biases and their influence on clinical interactions
- Empathy-driven care, fostered through reflection on unique patient narratives

To achieve these objectives, an mHealth app [39,40] was developed based on DEL concepts (hereafter referred to as "the app" or "the mHealth app"). The app functions as a serious RPG and features 2 virtual case scenarios: one involving a Syrian refugee with limited English proficiency, and another centered on an African American pregnant woman with a history of opioid use disorder. These immersive simulations allow learners to practice clinical decision-making and empathetic interaction in a safe and controlled environment. The app also integrates an interactive life course module, adapted from content originally developed by CityMatCH and later adapted by Wright State University [41]. This feature guides users through the life journey of virtual patients, highlighting the long-term effects of SDH on health outcomes and clinical encounters. Embedded within the app are 2 assessment surveys designed to capture user feedback and evaluate learning outcomes—allowing for continuous improvement of the educational platform.

In contrast to earlier pilot efforts, this study includes a robust sample size of 240 participants and presents a detailed analysis of the app's design, backend workflow, and educational impact. Comprehensive survey data and hypothesis testing offer rich insights into user engagement and knowledge acquisition. Ultimately, this DEL-based mHealth app provides a scalable and innovative solution to health care education challenges—advancing both clinical skills and health equity through immersive, technology-enabled learning.

Methods

Ethical Considerations

This project was determined to be exempt by the Wright State University (IRB record #06659) and the Ohio State University (10-24-2018) Institutional Review Boards (IRBs). This study involved human participants and received a category 3b exemption from the Institutional Review Board of Wright State University. The study was supported in part by the Ohio Department of Medicaid (grant #60077777) as part of the Medicaid Equity Simulation Project, administered by the Ohio Colleges of Medicine Government Resource Center. Informed consent was obtained within the app prior to participation. Participant privacy and data confidentiality were strictly maintained throughout the study. All procedures were conducted in accordance with institutional guidelines. No compensation was provided to participants.

XSL•FO
RenderX

Study Approach

This research leverages recent advancements in serious games, VR, human-computer interaction, DEL, and mobile computing to achieve objectives. The study integrates a DEL approach into a VR-based serious RPG or mHealth app. This integration harnesses the multimodal capabilities of cost-effective Android and iOS smart devices, ensuring widespread accessibility and maximizing the delivery of health care skills and awareness.

Incorporating a first-person viewing experience within the DEL medium enhances the learning process by providing health care professionals and individuals with immersive and enduring knowledge acquisition. The use of VR-based hypothetical case scenarios featuring diverse virtual characters enables safe and risk-free learning experiments, eliminating potential harm to real-life patients during clinical practice.

A key strength of this approach is the incorporation of educational objectives within an entertaining gameplay mechanism. This combination facilitates self-driven and self-motivated learning, empowering learners to actively engage in the educational process. As a result, cognitive learning and behavioral change are promoted. Additionally, the inclusion of visual cues within the gameplay reduces the complexity of the learning mechanism, minimizing the need for extensive explanatory materials while ensuring an effective learning experience.

To assess the effectiveness of the DEL approach–based tools in the health care domain, the mHealth app incorporates pre- and postsurvey or assessment questionnaires within the gameplay. This facilitates an analysis of user learning performance and evaluation of the overall effectiveness of the DEL approach in health care education. Furthermore, this research seeks to advance health care education by providing an immersive, engaging, and effective learning platform that enhances health care professionals' skills and knowledge. The integration of innovative methodologies underscores the potential to transform health care education, delivering a novel and impactful learning experience.

mHealth App Design Framework and Implementation Strategy

Overview

The primary objective of this research is to design and develop a DEL approach-based mHealth app or a serious RPG that empowers and inspires health care professionals to acquire advanced health care experiential learning through the convenience of mobile platforms, specifically iOS and Android devices. By leveraging the capabilities of widely accessible and cost-effective smart devices like iPhones, iPads, Android phones, and tablets, the app maximizes the potential to effectively deliver health care skills and awareness.

The app design centers around 2 hypothetical case scenarios, each featuring a central virtual character or patient. The first case scenario revolves around Muhammad Ayoubi, a Syrian refugee with limited English proficiency (hereinafter referred to as Muhammad's case), while the second case scenario focuses on Ebony Williams, an African American pregnant woman with a history of opioid use disorder (hereinafter referred to as Ebony's case). These scenarios provide health care professionals and other learners with an engaging and immersive opportunity to experience a patient's clinical visit from a first-person perspective. By actively participating in these scenarios, learners can enhance their health care skills and contribute to the objectives of this research.

High-Level Block Diagram Illustrating the mHealth App's Workflow

To provide a clear overview of the mHealth app workflow, a high-level block diagram is illustrated in Figure 2. Upon launching the app, a session number (token) is obtained from the web server, which randomly assigns an identifier representing either case scenario (ie, Muhammad's case or Ebony's case) to the learner. The session number is essential to begin the app session; without it, the learner cannot proceed in the learning process. This is followed by the collection of demographic data, enabling analysis of participants' responses in relation to their demographic characteristics. Subsequently, the learner is introduced to the patient, providing a brief overview of their background and reason for the clinical visit based on the assigned case scenario.

During the patient introduction, the learner assumes the perspective of the health care provider, permitting a deeper and more immersive connection with the patient. To gauge the learner's initial impressions and emotional responses, a presurvey or assessment questionnaire is administered, capturing their thoughts on assuming the role of the health care provider and their perception of the introduced patient.

To further enhance the educational experience, the app incorporates a life course module focused on SDH specific to each virtual patient (Muhammad or Ebony). This module provides valuable insights into the patient's background, social context, and various factors that contribute to their overall health and well-being.

The app also features immersive VR-based role-playing scenes or sessions, carefully designed to encompass different aspects of health care practice. These scenes include the virtual patient's preparation for the clinical appointment, transportation to the health center, check-in experience at the facility, and the clinical encounter between the virtual patient and virtual health care provider. Moreover, specific tasks unique to each case scenario are incorporated, such as filling prescriptions in Muhammad's case and seeking medication-assisted treatment in Ebony's case. By engaging in these role-playing sessions, health care professionals can acquire advanced experiential learning in a safe and risk-free environment, avoiding any potential harm to real-life patients.

XSL•FO
RenderX

Figure 2. High-level block diagram of the backend architecture and user interaction flow within the mobile health app.

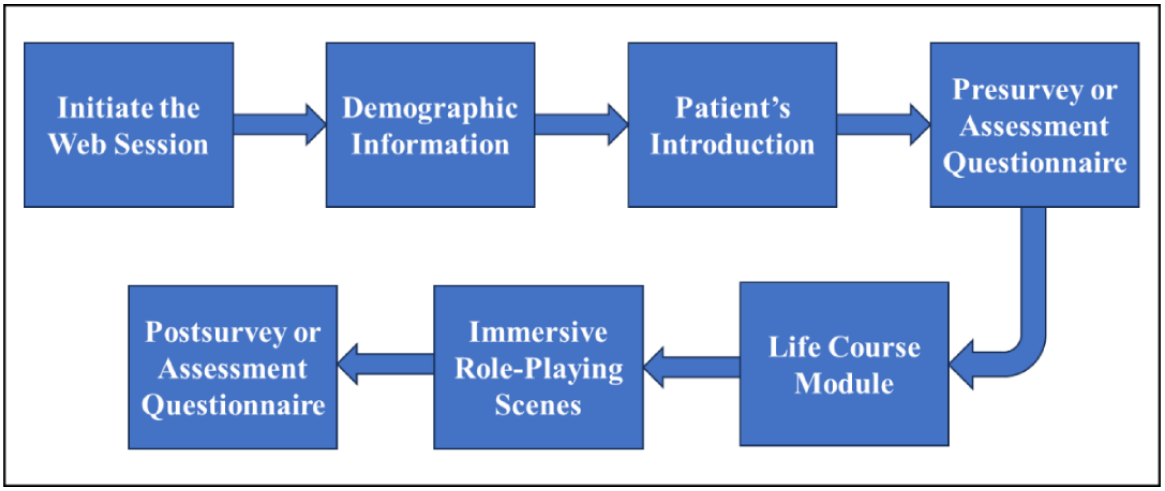

A crucial aspect of the app's design is the integration of hidden educational objectives within an entertaining gameplay mechanism. This combination ensures that the VR-based serious RPG serves as a self-driven and self-motivational learning tool, enabling learners to undergo cognitive learning and experience behavioral change. The incorporation of visual cues within the gameplay reduces the complexity of the learning process while still delivering an effective and engaging educational experience.

To evaluate the effectiveness of the mHealth app or serious RPG in the health care domain, pre- and postsurvey or assessment questionnaires are embedded within the gameplay. These questionnaires facilitate analysis of learner performance and provide insights into the overall impact of the approach in enhancing health care skills and education.

Thus, this study contributes to the advancement of health care skills and educational awareness by leveraging recent developments in serious games, VR, human-computer interaction, DEL, and mobile computing. The design and development of a VR-based serious RPG provides health care professionals with a transformative learning platform. By immersing learners in realistic virtual scenarios and enabling interaction with diverse virtual characters, the app facilitates effective acquisition and enhancement of essential health care skills.

Digital Experiential Learning Approach–Based mHealth App Design and Development Aspects

This section explores the intricate details of the app's design and development aspects. It outlines the key elements that shape the user experience and functionality, discussing the integration of various modules and features. Through an examination of the technical underpinnings and deliberate design considerations, valuable insights are provided into the creation of a seamless and immersive learning environment.

Seamless Integration of Virtual Environment and Virtual Characters

This study harnesses virtual environments and virtual characters to create an engaging and effective learning experience for health care professionals [39,40]. The integration of a virtual environment within the serious RPG served multiple purposes, including providing public education on medical diseases, patient treatment, and training for health care professionals. By immersing learners in real-world scenarios, this approach sustains active engagement with experiential learning modules, enabling them to relate to practical situations while developing essential academic and professional competencies.

Using lifelike virtual objects and intricate real-world representations, learners explored various clinical procedures, enhancing their understanding and proficiency in a risk-free environment. The Unity Game Engine, provided by Unity Technologies, played a pivotal role in developing and integrating the virtual world into the mHealth app, enabling the creation of realistic and immersive role-playing scenarios. Figure 3 shows examples of such scenarios.

In addition to the virtual environment, the integration of virtual characters further enriched the learning experience in the DEL platform. Virtual personas, including patients and peers, provided learners with focused and involved interactions, highlighting expressions and emotions that enhanced gameplay engagement. With well-developed storylines and virtual settings, learners could repeat gaming tutorials for clinical practices, professional communication, and cognitive skill development at their convenience.

Virtual characters played an active role in engaging and conversing with players during the hands-on experience, allowing for meaningful interactions that strengthened learners' empathy and compassion. Reallusion software tools were used to develop these virtual characters, which were later seamlessly integrated into the Unity Game Engine to create compelling story scenes as illustrated in Figure 3. The combination of the virtual environment and virtual characters elevated the overall learning experience and contributed to the app's success in empowering health care professionals with practical skills and compassionate attitudes toward patient care.

XSL•FO
RenderX

Figure 3. Visual representation of the simulated clinical environment, and animated patient characters featured in the app to facilitate immersive role-playing and empathy-based learning.

Integration of Interactive VR-Based Serious Role-Playing Sessions

An indispensable aspect of the mHealth app or serious game involves the integration of immersive role-playing scenarios that effectively engage and educate learners in the realm of clinical patient care [39,40]. Through these rigorous role-playing sessions, learners gain insights into patient experiences, the ramifications of SDH, and the implications of biases on health outcomes. Within these interactive sessions, learners assume the role of virtual health care providers and make consequential conversational choices that dictate the virtual patient's responses, behaviors, and health trajectories. This dynamic conversation flow is facilitated by using N-ary tree data structures and Unity C# scripts, ensuring a fluid and lifelike interaction with the virtual patients, thus delivering an authentic and compelling learning experience.

As learners progress through the clinical encounter module, they interact with individual virtual patients, Muhammad and Ebony, each presenting specific challenges and eliciting distinct responses, as highlighted in Figures 4 and 5. The user interface was meticulously designed to offer distinct conversational choices—Empathetic, Neutral, and Apathetic—presented as on-screen buttons. These choices trigger a spectrum of emotions and behaviors in the virtual patient, with their impacts portrayed through an array of cues, including facial expressions, gestures, dialogue captions, and even changes in heart rate. The choices for available responses were designed in a distinct color scheme and shown at a different position to clearly separate them from the captions provided throughout the entire interaction.

A key design decision was the inclusion of the heart rate indicator, an innovative addition that visually represents the patient's stress level at any given point, as illustrated in Figure 5A. The rationale behind this choice was to provide learners with immediate feedback on the emotional impact of their dialogue choices. For instance, when learners opt for the "Empathetic" choice, the virtual patient maintains direct eye contact, and the green-colored heart rate indicator in the upper left corner, as shown in Figure 5A, reflects a reassuringly normal range, indicating a sense of calm. Conversely, the "Neutral" choice, as demonstrated in Figure 5B, triggers a subtly upset response, with the patient's bowed head signifying concern. Here, the yellow-colored heart rate indicator, shown in Figure 5B, registers a moderately high range, signaling a heightened emotional state. Finally, the "Apathetic" choice, depicted in Figure 5C, triggers a notably distressed reaction, accompanied by a pronounced head bow. The patient avoids eye contact, and the red-colored heart rate indicator, shown in Figure 5C, signifies a significantly elevated stress level. The meticulous distinction in design between conversational buttons and subtitles ensures clarity in user interaction. The app's design

XSL•FO
RenderX

was guided by expert discussions that underlined the importance of facilitating a nuanced learning experience, reinforcing the significance of empathetic communication and compassionate, unbiased patient care.

Furthermore, the mHealth app encompasses supplementary role-playing sessions simulating the hurdles virtual patients face during clinic visits, including transportation predicaments and check-in procedures. By immersing themselves in these challenges from the virtual patient's perspective, learners foster heightened empathy and understanding, nurturing a compassionate disposition, and mitigating implicit/explicit biases.

Hence, the seamless integration of interactive VR-based serious role-playing sessions within the mHealth app advances the learning journey for health care professionals, empowering them to elevate their clinical practices and communicate with patients in a more empathic and compassionate manner. The infusion of these engaging role-playing scenarios significantly contributes to the overarching efficacy of the mHealth app.

Figure 4. Visual representation of interactive virtual reality–based serious role-playing scenarios embedded within the app. (A) Muhammad's case: a Syrian refugee with limited English proficiency, and (B) Ebony's case: an African American pregnant woman with a history of opioid use disorder.

Figure 5. (A) Outcome display when the health care provider selects an "Empathetic" response: virtual patient experiences improved health outcomes, and increased trust and engagement, reinforcing compassionate care approaches; (B) Outcome display from a "Neutral" choice selection by the health care provider, showing moderate patient engagement and incomplete resolution of health challenges—demonstrating the value of effective communication; and (C) Outcome display resulting from an "Apathetic" response, highlighting diminished patient health outcomes and disengagement—used to emphasize the negative effects of provider bias and lack of empathy.

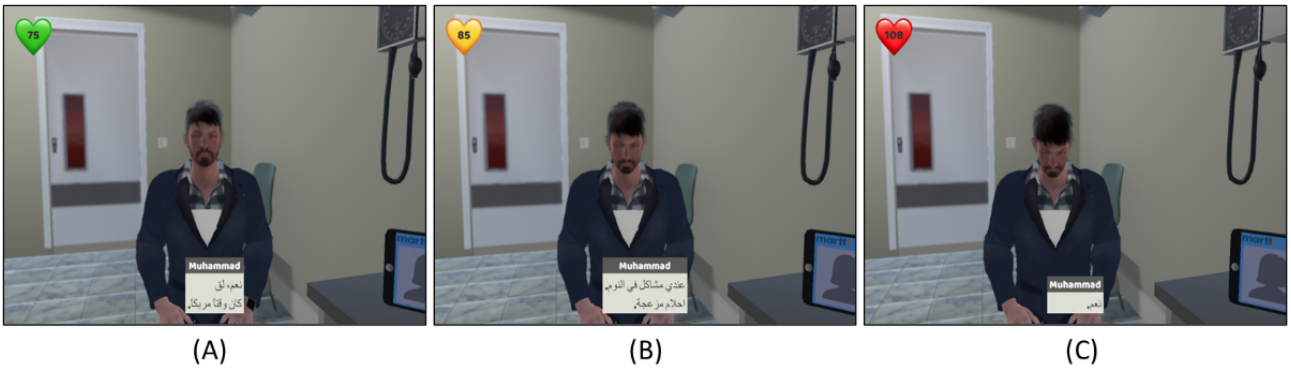

Integration of Immersive VR-Based Serious Role-Playing Modules

To craft a captivating and authentic clinical care practice experience for health care professionals, the immersive virtual reality–based serious role-playing modules were meticulously designed [39,40]. These modules are tailored for both Muhammad's and Ebony's cases, as illustrated in Figure 6A and Figure 6B, respectively. Within these modules, learners are completely immersed in a virtual environment, enabling them to interact with virtual elements and participate in clinical scenarios that provide firsthand experience in patient care practices.

XSL•FO
RenderX

Figure 6. (A) Illustrations of immersive virtual reality–based serious role-playing modules of the mHealth app design in "Muhammad's case," and (B) Illustrations of immersive virtual reality–based serious role-playing modules of the mHealth app design in "Ebony's case.".

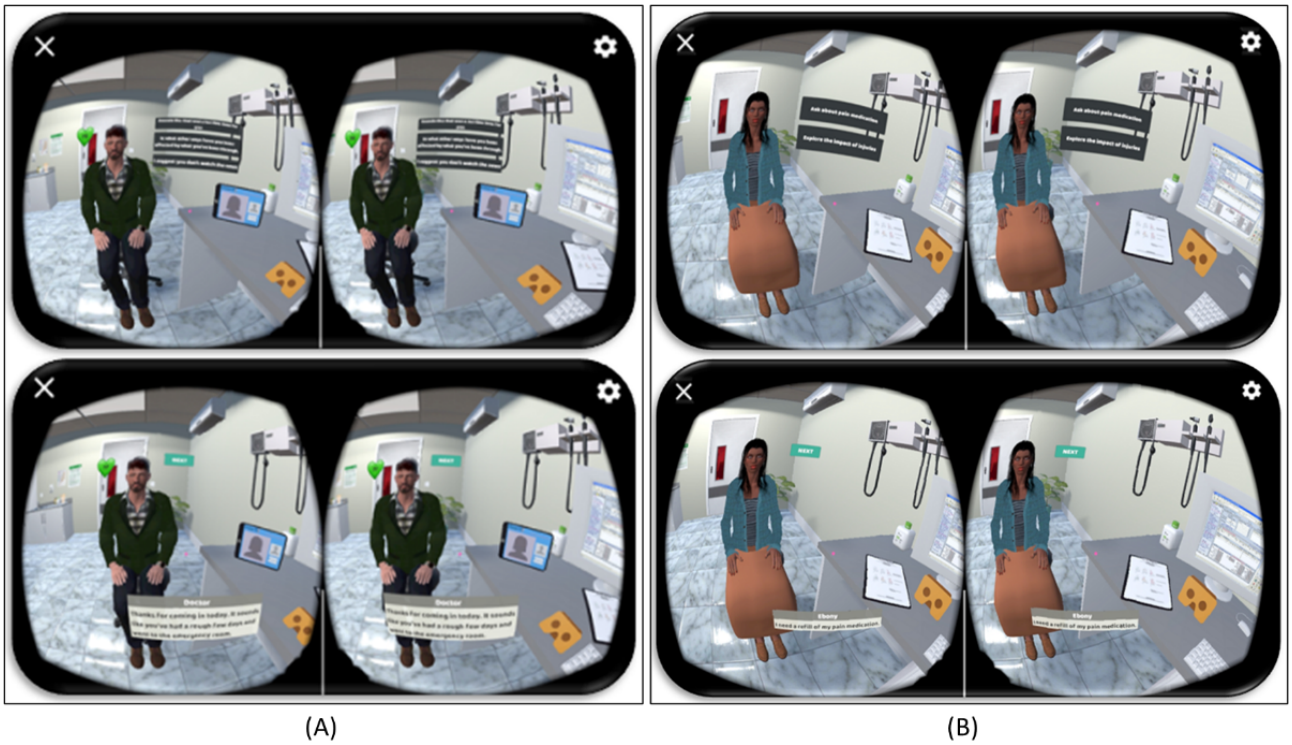

By offering a first-person viewing experience from the perspective of the virtual health care provider, the module encourages active participation and enhances experiential learning. The use of 360-degree viewing capabilities in the virtual health care environment further deepens the learners' involvement and improves their health care skills. During the immersive clinical encounter, learners are provided with a reticle pointer on the mobile device screen, enabling them to select conversational choices and interact with the virtual patient. The integration of the reticle pointer was accomplished using the Google VR software development kit, specifically the GvrReticlePointer prefab, in the app development process. However, the true power of these modules lies in their ability to evoke empathy and compassion. By immersing learners in the challenges of patient care, the learners acquire clinical skills and develop an understanding of the patient's perspective. Thus, the immersive VR-based serious role-playing modules bridge the gap between theory and practice, offering an immersive, experiential, and empathetic approach to health care education. They equip health care professionals with the skills and perspective necessary for patient-centered care in today's complex health care landscape.

Integration of Visual Cues in the mHealth App for Enhanced Learning

The assimilation of visual cues in the DEL mechanism plays a vital role in educational purposes, especially in health care–related learning tools [39,40]. In the mHealth app, visual cues serve as critical elements that help learners comprehend the various effects of their actions, interactions, and choices within the virtual case scenarios. These visual cues encompass a wide range of elements, including virtual objects, symbols, graphics, subtitles, thought bubbles, changes in virtual characters' behavior, and facial/body emotions. By providing learners with these visual clues, the app enhances their understanding of complex scenarios, encouraging active engagement and knowledge enrichment.

A dynamic feedback system is also integrated into the serious role-playing sessions to further enhance the learning experience. This system is intricately designed to deliver immediate visual and audio responses based on the conversational choices learners make during clinical encounters with virtual patients. When selecting from response options—Apathetic, Neutral, or Empathetic—the system reacts in real time to simulate realistic patient feedback. For instance, choosing the "Empathetic" option elicits a positive response from the virtual patient, characterized by maintained eye contact, relaxed body language, and a normalized heart rate. Simultaneously, learners receive reassuring responses from the virtual patient. In contrast, opting for the "Neutral" choice results in a more reserved patient response, with the virtual patient showing slight signs of concern, which results in a moderately high range of heart rate. The dialogue becomes less enthusiastic, reflecting the impact of a neutral approach on patient interaction. Finally, selecting the "Apathetic" option triggers an extremely high increase in heart rate and a visibly distressed reaction from the virtual patient, who avoids eye contact, displays physical discomfort, and responds with unease. The dialogue, in this case, conveys the impact of apathetic communication. This dynamic feedback mechanism serves several crucial purposes. First, it offers learners a direct link between their choices and patient responses, facilitating a deeper understanding of the consequences of their communication style. Learners can witness firsthand how empathetic communication fosters a positive patient-provider relationship, while apathetic

XSL•FO
RenderX

interactions can lead to patient distress. Second, the real-time nature of this feedback system provides an immersive and engaging learning experience, allowing learners to make immediate adjustments to their approach and observe the corresponding patient reactions. This iterative learning process promotes the development of essential health care skills, including empathetic communication, compassionate care, and the cultivation of unbiased attitudes toward patients. By simulating the complexities of real-world clinical encounters, this dynamic feedback system prepares health care professionals for a diverse range of patient interactions, contributing to the improvement of patient care and the promotion of patient-centric health care practices.

Figure 7 illustrates some of the distinct visual cues, including but not limited to heart rate indicators, conversational dialogue captions, 360-degree visual capability, thought bubbles, on-screen messages, conversational audio clips, and virtual characters' facial and body expressions, used in the app to provide learners with an immersive and effective learning experience. These visual cues serve as essential elements to enhance the assimilation of knowledge and encourage learners to develop vital skills required for health care professionals. Each of these visual cues offers specific benefits that contribute to an enriched learning process. The heart rate indicator, for instance, allows learners to gauge the emotional state of the virtual patient, aiding in understanding the impact of conversational choices. Facial and body expressions of virtual characters provide cues for emotional responses, fostering empathy and enhancing communication skills. Conversational dialogue captions enable learners to grasp the nuances of interactions, enhancing language proficiency and cultural sensitivity. The integration of 360-degree visual capability immerses learners in the virtual environment, promoting a holistic understanding of patient scenarios. Thought bubbles provide insights into the patient's thoughts, aiding in comprehension of underlying concerns. On-screen messages guide learners through complex scenarios, facilitating step-by-step learning. Conversational audio clips engage auditory learners, reinforcing effective communication strategies. Collectively, these visual cues not only facilitate the assimilation of knowledge but also encourage learners to develop essential skills required for health care professionals, enhancing a deeper understanding of patient-care dynamics and promoting a patient-centric approach.

Figure 7. Illustration of visual cues used within the mHealth app—such as body language indicators, facial expressions, thought bubbles, dialogue boxes, and guided instructions—designed to assist users in interpreting patient emotions and making informed decisions, and reinforcing nonverbal communication skills.

XSL•FO
RenderX

Integration of Life Course Module in the mHealth App to Explore Patients' Life Journeys

To enrich care experiences and foster empathy and compassion toward patients, the app incorporates a life course module [39,40]. This module delves into the patient's prior life experiences, especially those related to SDH, to provide learners with a comprehensive understanding of how these experiences impact health outcomes. By exploring both advantageous and disadvantageous life events of the patient and a privileged companion character, learners gain insights into the enduring effects of SDH on patient well-being. The life course module facilitates a connection between the learner and the patient, encouraging a deeper understanding of the individual's unique circumstances and fostering a more empathetic approach to patient care. Figure 8 illustrates instances of the life course module in the app, providing learners with firsthand experience of the patient's life journey and the challenges they have faced.

Figure 8. Visual representation of the integrated life course module within the mHealth app design showing virtual patient timelines and contextual influences of social determinants of health (SDH) across the lifespan.

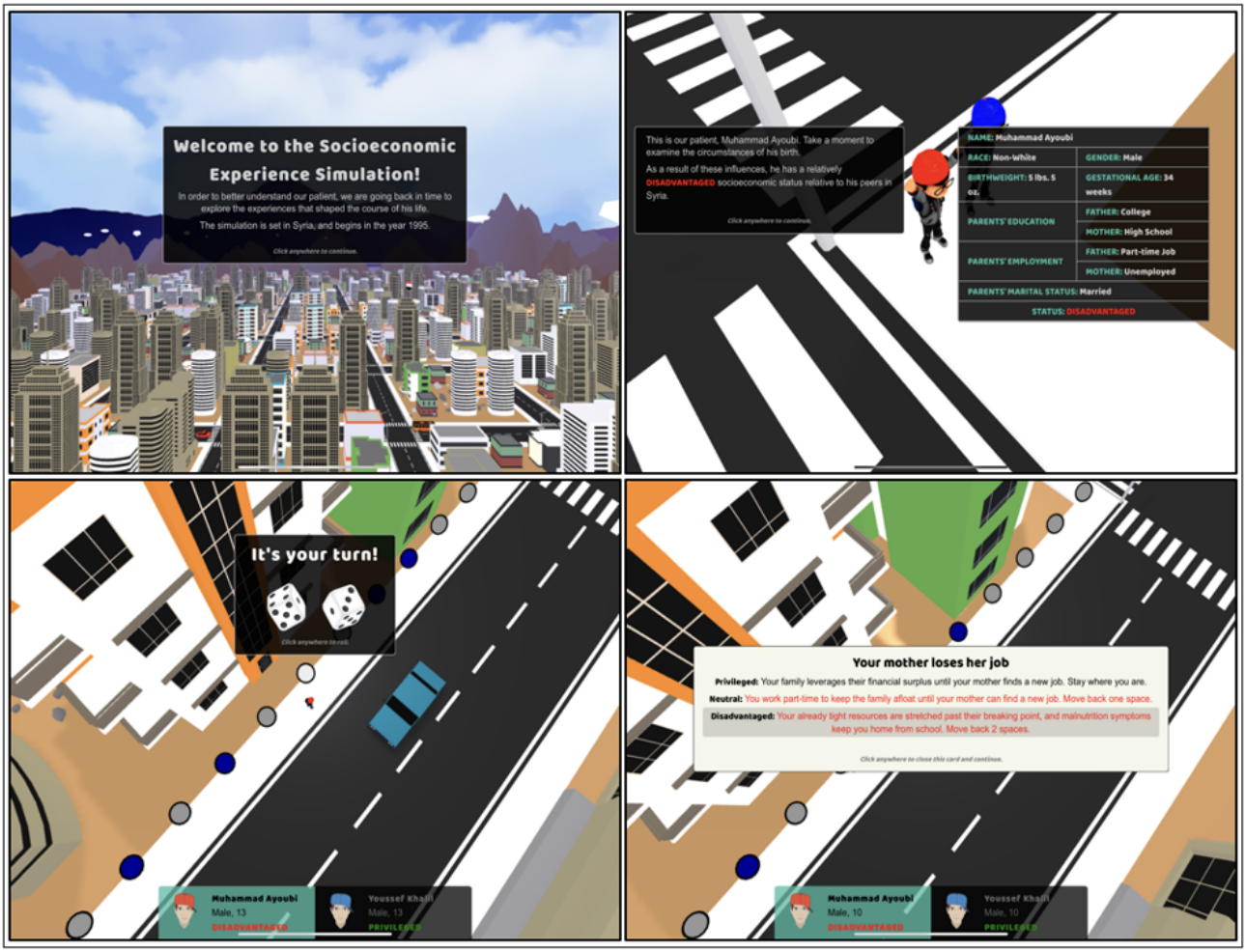

Incorporating the life course module adds value to the experiential learning process, equipping health care professionals with a more holistic view of patients' lives and the impact of social determinants on their health. The module serves as a compelling educational tool, conveying valuable learning experiences to the end users. Each virtual case scenario in the mHealth app features a tailored life course module, presenting past life experiences at different age instances of the main virtual character. These events are carefully chosen to evoke emotions and supplement the learning experience. The interactive life course module features a series of illustrations that represent significant life events of the main virtual characters, creating a powerful learning environment.

By exploring the past journey of the patients through the life course module, health care professionals gain a deeper appreciation for their patients' backgrounds, challenges, and resilience. This understanding strengthens the clinician-patient relationship, leading to improved patient care and reduced biases and health disparities. The life course module plays a vital role in instilling compassion and empathy in learners, enhancing the overall care experiences in health care settings.

Figure 8 vividly presents a series of visual illustrations drawn from the life course module, meticulously capturing pivotal events that have shaped the life journey of the main virtual character. Through these poignant depictions, learners are immersed in a poignant contrast between the life experiences of the main virtual character or patient and those of a privileged companion character. These carefully curated events, meticulously conveyed through visual cues, highlight the disparities and intricacies of various life stages, underscoring the influence of SDH on the patient's trajectory. This unique comparison prompts learners to intimately engage with the

XSL•FO
RenderX

patient's narrative, fostering a sense of empathy, compassion, understanding, and a patient-centric perspective in their approach to health care practice. This provides a direct comparison between the main character of the simulation and another person who was more fortunate in life to provide some additional background and illustrate some of the challenges the main character is facing.

Easy Accessibility, Availability, Affordability, Compatibility, and Scalability

Ensuring easy accessibility, availability, affordability, and scalability of the mHealth app or serious RPG is a cornerstone of this study. Recognizing the importance of self-paced learning in enhancing educational effectiveness, the app was designed and developed to be responsive across a range of devices. The layout is optimized to accommodate various aspect ratios commonly found in iOS and Android devices, ensuring consistent usability and functionality across iPhones, iPads, Android phones, and tablets. This design prioritizes compatibility across multiple platforms, thereby maintaining a seamless user experience and enhancing accessibility and engagement for health care professionals accessing the app for skill development, anytime and anywhere with an internet connection.

An integral aspect of the app's design philosophy is its affordability. The educational resource is intended to be accessible to all users, including health care professionals, without imposing any financial burden. Therefore, the app is made available to users at no additional cost, offering an affordable and equitable learning solution. Moreover, the scalability of the app is paramount in the overall design and implementation strategy. As the demand for health care education continues to grow, the app has been developed to be highly scalable, capable of supporting a diverse and expanding user base without compromising the quality of the learning experience.

By integrating essential elements such as accessibility, availability, affordability, and scalability, the resulting platform offers a comprehensive and user-friendly solution that empowers health care professionals to enhance their skills and knowledge effectively and efficiently. This approach also enables broader dissemination and increased reach across diverse user populations.

Tools and Technology

In this research, the major tools and technology used to develop the app were Unity and Reallusion Hub [42,43]. Unity, a powerful game engine, played a vital role in creating immersive VR scenarios and serious role-playing exercises. It provided real-time feedback during development, enabling lifelike simulations for enhanced learning experiences. Reallusion Hub, a comprehensive software suite for character creators, animation, and facial expression mapping, was instrumental in creating realistic avatars and enriching the simulated patient interactions within the app. The C# programming language was used to implement interactive functionalities and logic, adding dynamic elements to the app. The combination of these innovative tools and technology resulted in an innovative mHealth app or serious RPG that transformed health care professional education, offering a dynamic learning platform to empower professionals in their pursuit of improved patient care. Additionally, GitHub served as the collaborative version control platform, enabling multiple researchers and developers to work seamlessly on the project and track changes throughout the development process.

Results

The mHealth App's Distribution, Data Collection, and Recruitment Method

The mHealth app is conveniently distributed through the App Store and Google Play Store in the United States, ensuring its accessibility across a wide range of iOS and Android devices, including iPhones, iPads, Android phones, and tablets. This widespread availability enables reaching a diverse user base. To thoroughly evaluate the effectiveness of the app, developed based on the DEL approach, and to gain insights into the users' learning experiences, comprehensive pre- and postsurvey questionnaires have been integrated within the app. Table 1 shows 14 questions integrated into the assessment questionnaires of the app. Specifically, questions 1 to 9 are included in both the pre- and postsurvey or assessment questionnaires, while questions 10 to 14 are exclusive to the postsurveys or assessment questionnaires. For each question listed in Table 1, the "User Response Options or Levels" are presented as radio buttons, allowing learners to select one of the 5 distinct levels (ie, 1, 2, 3, 4, and 5) to indicate their response. These levels range from "low to high" or "never to always" or "strongly disagree to strongly agree."

XSL•FO
RenderX

Table 1. Assessment or survey questions used in the analysis.

Index	Assessment or survey question	User response options or levels	Expected user response options or levels (from 1 to 5)
1	With respect to having this individual as my next patient, the amount of ANXIETY I feel is:	1 (Low) to 5 (High)	Lower is better
2	With respect to having this individual as a patient, the amount of FRUSTRATION that I feel is:	1 (Low) to 5 (High)	Lower is better
3	With respect to having this individual as a patient, the amount of COMPASSION that I feel is:	1 (Low) to 5 (High)	Higher is better
4	If given a choice, instead of this patient I would prefer to see a different patient for a routine follow-up of a chronic health problem such as hypertension.	1 (Strongly disagree) to 5 (Strongly agree)	Lower is better
5	I expect that future encounters with this patient will be:	1 (Easy) to 5 (Difficult)	Lower is better
6	I believe that this patient is largely responsible for being in their current circumstances.	1 (Strongly disagree) to 5 (Strongly agree)	Lower is better
7	I believe that the circumstances in which this patient finds themself are largely beyond their control.	1 (Strongly disagree) to 5 (Strongly agree)	Higher is better
8	I try to hide any negative thoughts about patients like this to avoid negative reactions from others.	1 (Never) to 5 (Always)	Higher is better
9	I attempt to act in nonprejudiced ways toward patients like this because it is personally important to me.	1 (Never) to 5 (Always)	Higher is better
10	I believe going through this training experience will help me decrease any negative biases I may have toward patients like this.	1 (Strongly disagree) to 5 (Strongly agree)	Higher is better
11	I believe, going through this training experience will help me be a more understanding health professional.	1 (Strongly disagree) to 5 (Strongly agree)	Higher is better
12	This activity was an effective learning platform:	1 (Strongly disagree) to 5 (Strongly agree)	Higher is better
13	I will apply at least one new thing that I learned from this simulation experience in my clinical practice.	1 (Strongly disagree) to 5 (Strongly agree)	Higher is better
14	What will you take with you into your clinical practice as a result of this simulation experience?	Enter text…(Optional)	__[a]

[a]Not applicable.

At the outset of the users' engagement with the mHealth app, participants are guided to complete a presurvey, which functions as a baseline assessment to capture their initial understanding, expectations regarding health care interactions, and perceptions of the app's role. This instrument serves as a calibration tool, establishing a reference point for users' prior knowledge and attitudes. Upon completing all role-playing sessions within the app, participants are prompted to complete a postsurvey. While the presurvey establishes the baseline, the postsurvey evaluates the app's impact by measuring knowledge acquisition, skill development, and attitudinal changes resulting from the simulated clinical experiences. This dual-survey framework enables a comprehensive evaluation of the app's educational effectiveness. The comparative analysis of pre- and postsurvey responses provides insight into users' progression, including improvements in health care competencies, enhanced empathetic communication, and the adoption of patient-centered, unbiased care practices.

To collect diverse learning data, primarily comprising health care professionals, a multi-faceted data collection strategy was implemented. First, in-person training sessions were conducted at local hospitals, where participants were provided with mobile devices (ie, an iPhone, iPad, Android phone, or tablet)

XSL•FO
RenderX

preinstalled with the app, enabling direct engagement. This approach facilitated real-time feedback and allowed immediate resolution of any questions or concerns. Second, the app was made available through major digital platforms, specifically the App Store and Google Play Store, broadening the reach and enabling remote data collection. Additionally, a dedicated website with direct links to the app was developed, allowing users convenient access to the app page for easy download into their mobile devices.

In this study, 240 participants who voluntarily engaged with the mHealth app were evaluated based on their sustained participation throughout the sessions, demonstrating a clear willingness to contribute to the research. These participants were selected from a larger pool of volunteers who accessed the app both via web (ie, the App Store and Google Play Store) and in-person. Participants who were not evaluated could not successfully complete the entire app session, resulting in incomplete pre or postsurvey data. The unsuccessful completion of the entire session may have been due to various reasons, including technical issues like server failures, inappropriate internet connectivity, a decision to no longer participate, or simply testing the app. Hence, the analysis presented in this study corresponds to 240 participants who successfully completed the session. All descriptive and statistical calculations—including response counts, percentage derivations, CIs, effect size (Cohen *d*), odds ratios (ORs), and 1-tailed paired *t* test—were performed using Microsoft Excel 365 software (Microsoft Corporation, Redmond, Washington).

Among the 240 individuals who completed the mHealth app session, 207 (86.3%) individuals identified as female, 29 (12.1%) individuals identified as male, and 4 (1.7%) individuals identified as non-Binary. This demographic spread reflects a predominant female representation and highlights the need for further engagement strategies to support participation across gender identities.

Moreover, 36/240 (15.0%) individuals were aged between 11 and 20 years, 100/240 (41.7%) individuals fell within the 21-30 years of age, 39/240 (16.3%) individuals were aged between 31 and 40 years, 28/240 (11.7%) individuals were aged between 41 and 50 years, 25/240 (10.4%) individuals fell within the 51-60 years, 10/240 (4.2%) individuals were aged 61-70 years, and 2/240 (0.8%) individuals were in the age range of 71-80 years. This broad age spectrum highlights the inclusion of participants across generational cohorts, offering insight into how the mHealth app engaged learners at varied developmental stages and professional maturity levels.

In addition, the participant pool demonstrated a wide distribution of professional experience levels. Of the 240 individuals who completed the full app session, 95 (39.6%) individuals reported having less than one year of professional experience, followed by 47 (19.6%) individuals with 1 to 5 years, and 30 (12.5%) individuals with 6 to 10 years. A smaller subset included 20 (8.3%) individuals with 11 to 15 years of experience, while 3 equally sized groups—each with 9 (3.8%)—had between 16 and 30 years of experience, segmented into the 16- to 20-, 21- to 25-, and 26- to 30-year ranges. Finally, 21 (8.8%) individuals reported over 31 years of experience in their respective health care professions. This distribution underscores the app's relevance across a spectrum of clinical maturity—from early-career professionals to seasoned practitioners.

Furthermore, the participant pool comprises a diverse range of health care professionals, including 18/240 (7.5%) physicians, 75/240 (31.3%) nurses, 24/240 (10.0%) social workers, 8/240 (3.3%) nurse practitioners or other types of advanced practice registered nurses, 4/240 (1.7%) physician assistants, 2/240 (0.8%) dental hygienists, and 1/240 (0.4%) each of dentist, optometrist, physical therapist, occupational therapist, medical resident, clinical psychologist, pharmacist, and respiratory therapist. Additionally, 23/240 (9.6%) were nonclinical personnel, and 78/240 (32.5%) were individuals from other relevant health care professions. This diverse composition supports a robust analysis of how the mHealth app performs across a range of professional backgrounds.

The following section provides a distinct analysis to evaluate the effectiveness of the mHealth app, with all discussions pertaining to the question indices corresponding to the questions listed in Table 1. The distinctive nature of the questions necessitated careful consideration of whether higher or lower response levels indicated improvement, tailoring the result analysis to the specific learning objectives of each question.

Hypotheses Consideration for the Result Analysis

Motivation

In this research, 9 hypotheses were developed to evaluate the impact of the mHealth app on health care professionals' skills and attitudes toward patients. Each hypothesis aligns with a specific survey question from questions 1 to 13 listed in Table 1.

Hypothesis 1 (Corresponding to Question 1)

The mHealth app will lead to a reduction in participants' anxiety levels toward patients. It is hypothesized that the app's first-person VR scenarios and role-playing exercises would offer a risk-free environment for participants to practice patient interactions, contributing to decreasing anxiety levels.

Hypothesis 2 (Corresponding to Question 2)

The mHealth app will lead to a reduction in participants' frustration levels toward patients. It is hypothesized that the app's immersive learning experiences would provide participants with effective strategies for managing frustration and stress during patient care.

Hypothesis 3 (Corresponding to Question 3)

The mHealth app will lead to a statistically significant improvement in participants' compassion and empathy toward patients. It is hypothesized that the app's first-person VR scenarios and role-playing exercises will enhance participants' ability to understand patients' perspectives, mirror their feelings, and demonstrate greater empathy in their interactions.

Hypothesis 4 (Corresponding to Questions 4 and 5)

The mHealth app will positively influence participants' expectations of future patient encounters (corresponding to question 5) and their willingness to see the same patient again

XSL•FO
RenderX

for routine follow-ups (corresponding to question 4). It is hypothesized that the app's DEL approach will enhance participants' confidence and preparedness for real-life patient interactions, fostering positive expectations and increasing their willingness to see the same patient for follow-up care.

Hypothesis 5 (Corresponding to Questions 6 and 7)

The mHealth app based on the DEL approach will positively impact participants' beliefs about patient responsibility (corresponding to question 6) and their perceptions of the controllability of patients' circumstances (corresponding to question 7). It is hypothesized that the app's immersive learning experiences will raise awareness of the complexity of patient situations and promote a more empathetic and understanding approach in health care professionals.

Hypothesis 6 (Corresponding to Question 8)

Participants' tendencies to hide negative thoughts about patients will be positively impacted by the mHealth app. It is hypothesized that the app's immersive learning experiences will reinforce the importance of providing equitable and unbiased health care services, leading to reduced negative thoughts about patients.

Hypothesis 7 (Corresponding to Question 9)

The mHealth app based on the DEL approach will positively impact participants' commitment to act in nonprejudiced ways toward patients. It is hypothesized that the app's immersive learning experiences will reinforce the importance of providing equitable and unbiased health care services.

Hypothesis 8 (Corresponding to Questions 10 and 11)

Participants' engagement with the mHealth app will lead to a statistically significant increase in their belief that the training experience helps decrease negative biases (corresponding to question 10) and that it helps them become more understanding health professionals (corresponding to question 11). It is hypothesized that the immersive and experiential learning provided by the app will effectively challenge and reduce negative biases that health care professionals may have toward patients, while also promoting self-awareness and empathy.

Hypothesis 9 (Corresponding to Questions 12 and 13)

Participants will express a positive perception of the mHealth app as an effective learning platform (corresponding to question 12) and a strong intention to apply at least one new skill or insight gained from the simulation experience in their clinical practice (corresponding to question 13). It is hypothesized that the interactive and immersive learning experiences offered by the app will be positively received by participants, enhancing their educational experience and motivating them to apply new skills and insights in their professional practice.

By formulating these hypotheses, the research aims to establish specific expectations for the analysis of participants' responses and determine the impact of the app on achieving the intended learning outcomes and enhancing health care professionals' skills and attitudes toward patient care.

CI, Cohen d, and McNemar 2×2 Odds Ratios Analysis

CIs, Cohen d values, and McNemar 2×2 ORs were calculated to evaluate shifts in learner responses, using a 95% confidence level corresponding to a conventional alpha threshold of 0.05 (ie, α=0.05) for determining statistical significance. For each survey question 1-13, the change in mean score (pre vs post) with its standard deviation and 95% CI is provided. Additionally, for questions 1-9, Cohen d effect size for the standardized mean difference and a McNemar 2×2 OR (with 95% CI) based on a standard dichotomy of favorable versus nonfavorable responses are provided. As shown in Table 1, for questions 1, 2, 4, 5, and 6, favorable responses are user option levels 1-2 and nonfavorable responses are user option levels 3-5; whereas, for questions 3, 7, 8, and 9, favorable responses are user option levels 4-5 and nonfavorable responses are user option levels 1-3. Question 14 is analyzed separately due to its descriptive and optional nature.

The data indicate a measurable positive shift in learners' compassion toward patients, alongside reduced negative perceptions and reactions. These changes suggest the app supports skill development and promotes more empathetic and constructive engagement in clinical contexts. Participants also demonstrated greater awareness of SDH and health care biases, integrating this understanding into patient-centered care scenarios presented within the simulation.

As the analysis of individual questions proceeds, incorporating the Cohen *d* values and considering the limited response levels, it is important to acknowledge that the response scale in the survey had a finite range (ie, from 1 to 5). Participants may have already selected the optimal response level within this scale in the presurvey, leaving limited room for further showcasing an improvement in the postsurvey. This indicates that participants did not experience a decline in learning during the intervention. Moreover, engagement with the app itself inherently fostered knowledge enhancement, representing a significant positive outcome. Overall, the analysis highlights how the app provides a comprehensive learning experience that supports the development of emotional intelligence and cultural competence among health care professionals.

To illustrate, for question 1 (corresponding to hypothesis 1), presurvey mean 2.40 (SD 1.12; 95% CI 2.25-2.54) decreased to postsurvey mean 2.30 (SD 1.14; 95% CI 2.16-2.44). The narrower postsurvey CI and lower mean value of participants' responses compared with the presurvey indicate a reduced anxiety participants feel about encountering the presented patient as their next patient. The Cohen d effect size value is 0.08, reflecting the observed changes in participants' responses. Dichotomizing participants' responses into favorable (ie, user response levels 1-2) versus nonfavorable (ie, user response levels 3-5) yields OR 1.14 (95% CI 0.73-1.78), indicating a nonsignificant trend toward increased low-anxiety reporting.

Similarly, for question 2 (corresponding to hypothesis 2), presurvey mean 1.83 (SD 1.07; 95% CI 1.70-1.97) increased to postsurvey mean 2.08 (SD 1.14; 95% CI 1.94-2.23). The increment in the postsurvey CI and the higher mean value of participants' responses compared with the presurvey suggest a rise in frustration when managing the presented patient.

XSL•FO
RenderX

Participants may experience challenges or complexities while managing this patient, but the simulation enhances their ability to navigate such encounters. The Cohen *d* effect size value is 0.23, which demonstrates the impact of the intervention on participants' frustration levels. Dichotomizing participants' responses into favorable (ie, user response levels 1-2) versus nonfavorable (ie, user response levels 3-5) yields OR 0.51 (95% CI 0.31-0.83), indicating a significant decline in low-frustration reporting.

For question 3 (corresponding to hypothesis 3), presurvey mean 4.18 (SD 0.94; 95% CI 4.06-4.29) increased to postsurvey mean 4.60 (SD 0.73; 95% CI 4.51-4.69). The elevated postsurvey CI and higher mean value of participants' responses compared with the presurvey indicate a significant improvement in learners' compassion levels. The mHealth app effectively fosters a more empathetic response among participants, enhancing their ability to provide compassionate care to patients. The Cohen *d* effect size value is 0.50, highlighting the effect of the intervention on enhancing compassion among participants. Dichotomizing participants' responses into favorable (ie, user response levels 4-5) versus nonfavorable (ie, user response levels 1-3) yields OR 6.17 (95% CI 2.60-14.61), indicating a strong, significant increase in compassionate responses.

Moving on to question 4 (corresponding to hypothesis 4), presurvey mean 2.15 (SD 1.19; 95% CI 2.00-2.30) decreased to postsurvey mean 1.93 (SD 1.10; 95% CI 1.79-2.07). The reduction in the postsurvey CI and reduced mean value of participants' responses compared with the presurvey, signifies a decreased inclination among participants to see a different patient for routine follow-up, indicating greater comfort with managing such encounters. The Cohen *d* effect size value is 0.20, signifying the influence of the intervention on participants' preferences for different patients. Dichotomizing participants' responses into favorable (ie, user response levels 1-2) versus nonfavorable (ie, user response levels 3-5) yields OR 2.87 (95% CI 1.59-5.16), indicating a significant improvement in follow-up comfort.

For question 5 (corresponding to hypothesis 4), presurvey mean 2.87 (SD 0.98; 95% CI 2.74-2.99) decreased to postsurvey mean 2.74 (SD 1.00; 95% CI 2.61-2.87). The finer postsurvey CI and decreased mean value of participants' responses compared with the presurvey imply greater ease in dealing with future patient encounters, indicating improved confidence in managing patient interactions. The Cohen *d* effect size value is 0.13, suggesting changes in participants' expectations regarding future encounters. Dichotomizing participants' responses into favorable (ie, user response levels 1-2) versus nonfavorable (ie, user response levels 3-5) yields OR 1.85 (95% CI 1.16-2.96), indicating a significant gain in confidence for managing patient interactions.

Regarding question 6 (corresponding to hypothesis 5), presurvey mean 2.02 (SD 0.98; 95% CI 1.89-2.14) decreased to postsurvey mean 1.70 (SD 0.95; 95% CI 1.58-1.82). The reduction in postsurvey CI and decreased mean value of participants' responses compared with the presurvey suggests a change in participants' perspectives, indicating a decreased inclination to attribute patients' circumstances entirely to their own responsibility. The Cohen *d* effect size value is 0.33, indicating shifts in participants' perspectives on patient responsibility attribution. Dichotomizing participants' responses into favorable (ie, user response levels 1-2) versus nonfavorable (ie, user response levels 3-5) yields OR 2.94 (95% CI 1.70-5.10), indicating a significant shift toward reduced patient-blaming.

Conversely, for question 7 (corresponding to hypothesis 5), presurvey mean 3.51 (SD 1.02; 95% CI 3.38-3.64) increased to postsurvey mean 3.96 (SD 1.21; 95% CI 3.81-4.11). The elevated postsurvey CI and higher average value of participants' responses compared with the presurvey signify that a greater consensus among participants that patients face circumstances that are significantly beyond their control. This alignment with the statement suggests a prevailing belief that external factors play a substantial role in shaping the patients' situations. The Cohen *d* effect size value is 0.40, illustrating the influence of the intervention on participants' beliefs about external control. Dichotomizing participants' responses into favorable (ie, user response levels 4-5) versus nonfavorable (ie, user response levels 1-3) yields OR 6.00 (95% CI 3.07-11.72), indicating a robust improvement in external-factor recognition.

For question 8 (corresponding to hypothesis 6), presurvey mean 3.88 (SD 1.31; 95% CI 3.71-4.05) increased to postsurvey mean 4.03 (SD 1.31; 95% CI 3.86-4.20). The narrower postsurvey CI and higher average value of participants' responses compared with the presurvey suggest that the app successfully encourages participants to adopt a more empathetic tactic. Participants express a stronger commitment to suppressing negative thoughts and reactions, contributing to a more positive patient-provider interaction. The Cohen *d* effect size value is 0.11, pointing to changes in participants' efforts to suppress negative thoughts. Dichotomizing participants' responses into favorable (ie, user response levels 4-5) versus nonfavorable (ie, user response levels 1-3) yields OR 2.22 (95% CI 1.01-4.88), indicating a significant improvement in emotional regulation.

Regarding question 9 (corresponding to hypothesis 7), presurvey mean 4.71 (SD 0.70; 95% CI 4.62-4.80) increased to postsurvey mean 4.78 (SD 0.67; 95% CI 4.69-4.86). The narrower postsurvey CI and higher average value of participants' responses compared with the presurvey indicate that the app reinforces participants' commitment to acting in nonprejudiced ways. Participants exhibit an increased dedication to treating patients with equity and respect, reflecting positively on their interpersonal skills. The Cohen *d* effect size value is 0.10, reflecting the impact of the intervention on participants' commitment to nonprejudiced behavior. Dichotomizing participants' responses into favorable (ie, user response levels 4-5) versus nonfavorable (ie, user response levels 1-3) yields OR 1.50 (95% CI 0.42-5.32), indicating a nonsignificant trend toward improved equitable attitudes.

Together, the analyses of questions 1 to 9 confirm that the mHealth app or serious game fosters a more compassionate, empathetic, and inclusive health care environment. Participants' emotional responses and attitudes are positively influenced, contributing to improved patient-provider interactions and better patient outcomes. Additionally, responses to questions 10 to 13, which are included only in the postsurvey, further reinforce

XSL•FO
RenderX

the effectiveness of the app in fostering positive changes in learners' attitudes and behaviors.

For question 10 (corresponding to hypothesis 8), postsurvey mean 4.15 (SD 0.97; 95% CI 4.03-4.27) indicates a considerable proportion of participants agreed with the question's statement. The mean value of the participants' responses further emphasizes the positive impact of the simulation experience on reducing negative biases.

Similarly, for question 11 (corresponding to hypothesis 8), postsurvey mean 4.18 (SD 0.95; 95% CI 4.05-4.30) underscores a greater proportion of participants who agreed with the question's statement. The mean value of the participants' responses affirms the app's role in promoting empathy and understanding among health care professionals.

Regarding question 12 (corresponding to hypothesis 9), postsurvey mean 3.85 (SD 1.14; 95% CI 3.71-3.99) indicates a positive consensus among learners regarding the app's effectiveness as a learning platform. The mean value of the participants' responses supports the notion that interactive simulation fosters effective learning experiences.

Last, for question 13 (corresponding to hypothesis 9), postsurvey mean 4.07 (SD 0.94; 95% CI 3.95-4.19) demonstrates learners' commitment to applying the knowledge gained from the simulation in clinical practice. The mean value of the participants' responses further reinforces the app's potential to drive meaningful changes in real-world patient care.

Together, the analyses of questions 10 to 13 confirm that the app provides a comprehensive and impactful learning experience for health care professionals. It equips learners with essential skills, such as empathy, cultural humility, and self-awareness, which are vital for delivering patient-centered care. The app's positive influence on learners' emotional responses, attitudes, and professional development enhances the overall quality of patient care and strengthens the health care workforce's ability to provide inclusive and equitable services.

Analysis of Optional Descriptive Survey Question

Participants were invited to respond to an optional descriptive question (question 14) listed in Table 1 to gather qualitative insights into their experiences with the mHealth app's simulation. Table 2 presents a selection of verbatim responses from learners or participants for question 14, highlighting their improved self-awareness, compassion, and empathy toward patients. These responses provide valuable observations that illuminate the app's impact on their emotional connection with patients and their ability to empathize with patients' experiences.

The learners' feedback indicates that engaging with the simulation allowed them to develop a deeper sense of understanding and sensitivity to patients' emotions. By imagining how patients may feel and mirroring patients' feelings, participants demonstrated heightened emotional awareness and the capacity to connect with patients on a more acute level. This enhanced emotional connection is a crucial aspect of patient-centered care, as it fosters a stronger therapeutic alliance between health care professionals and their patients.

Furthermore, participants' responses suggest that the mHealth app effectively cultivated a nonprejudicial attitude among learners. By putting themselves in the patient's shoes and striving to understand the situation from the patient's perspective, participants demonstrated a commitment to embracing diverse patient experiences and perspectives. This newfound perspective-taking ability can break down barriers and disparities in health care, promoting equitable and inclusive care practices.

The qualitative feedback provided by participants in question 14 highlights the broader impact of the app beyond quantitative metrics. While the CI analysis and mean values offer valuable insights into learners' attitudes and self-awareness, the optional question delves into the emotional and empathetic transformations experienced by participants.

The collection of learners' responses in question 14 validates the effectiveness of the app in fostering a more empathetic and patient-centered approach among health care professionals. It also underscores the importance of integrating emotional intelligence and cultural competence into health care education and training. By developing emotionally intelligent health care professionals, the app empowers learners to deliver compassionate, patient-centered care and contribute positively to the overall health care experience.

This qualitative analysis provides a comprehensive understanding of the diverse ways in which the app influences learners' attitudes and emotional responses. These valuable insights complement the quantitative results, further affirming the mHealth app's potential to transform the health care landscape positively.

XSL•FO
RenderX

Table 2. A few of the participants' responses to the (an optional) survey question 14.

Index	Profession	Participant's learning experience or comments
1	Physician	"Directly engage social worker as part of the patient centered medical home"
2	Physician	"try to remember extenuting circumstances"
3	Physician	"I will better understand the implication of adverse life experiences."
4	Physician	"Try to understand a pt. from their perspective."
5	Physician	"forms in languages other than English"
6	Physician assistant	"I have learned to not be biased because you never know what people are going through."
7	Nurse	"More compassion!"
8	Nurse	"language barrier solutions"
9	Nurse	"understanding of what pt has to go through just to get to appointment"
10	Nurse	"use of simulation for my students concerning social determinants of health"
11	Nurse	"Being centered"
12	Nurse	"some patients circumstances are beyond their control"
13	Nurse	"to accomidate needs of patients. Work closely with soical workers to ensure proper followup"
14	Nurse	"Empathy"
15	Nurse	"Working to understand all aspects that go into what makes anpatient who they are"
16	Nurse	"Just how much behind thne individual can become with SDOH"
17	Nurse	"Qwait until the client broaches the topic and focus on the impact to functional life instead"
18	Nurse	"least 10 min rather then right away. let the patient calm down. we never know what someone else may be going through"
19	Nurse practitioner or other type of advanced practice registered nurse	"Look at the whole person."
20	Dental hygienist	"Remember to respond first with empathy and ask patient before doing anything."
21	Dental hygienist	"Not to judge anyone because you dont know their story"
22	Optometrist	"This worked well."
23	Social worker	"I appreciate the action orientedness of the clinic. The MD saw her eventhought she was late"
24	Social worker	"Take a step back use MI and try to understand the patients unique background"
25	Social worker	"ideration"
26	Social worker	"Taking the time to ask effective questions and understand why they came in."
27	Social worker	"Compassion and a non judgemental attitude. Always explore other options to help youe client"
28	Social worker	"People?s life circumstances are far beyond what meets the eye!"
29	Nonclinical personnel	"I will bring a better understanding and more empathy in my interactions."
30	Nonclinical personnel	"To identify a patient's issues rather than assuming based on previous encounters at other places."
31	Nonclinical personnel	"Be more understanding"
32	Nonclinical personnel	"The importance of asking questions"
33	Nonclinical personnel	"compassion and understanding is extremely important"
34	Other	"I will take a renewed reminder that people dont often choose their unfortunate circumstances."
35	Other	"putting myself in others shoes"

XSL•FO
RenderX

Index	Profession	Participant's learning experience or comments
36	Other	"patience"
37	Other	"to be more understanding"
38	Other	"always be understanding"
39	Other	"I do not know the daily things someone is going through and always treat everyone to the best of my ability."
40	Other	"I will not pass judgement on patients."
41	Other	"Asking the right questions and leading with support and understanding"
42	Other	"To act in a non prejudiced way towards all patients and help everyone no matter their circumstances"
43	Other	"The overall process of each step taken place during peoples appointments."
44	Other	"people can't always control their circumstances"
45	Other	"Treat every patient the same without any prejudice/bias. You never know what some patients have been through."
46	Other	"Before judging a patient i will try to put myself in their shoes"
47	Other	"I wont assume patients on opioids are drug seeking and i will try to advocate for patients"
48	Other	"I will try to understand each individual's background to help get them the treatment they need."
49	Other	"Be more open"
50	Other	"ls will allow as a healthcare professional to fully engage without bias and be able to further help the patient as a whole."

1-Tailed Paired t Test Analysis

To further explore the effectiveness of the mHealth app, a 1-tail paired *t* test analysis was conducted on questions 1 to 9, which were present in both the pre- and postsurvey questionnaires. An alpha level of 0.05 (ie, α=0.05) was used to determine statistical significance, meaning that *P* values less than .05 were considered indicative of meaningful improvement. As depicted in Table 3, the 1-tailed paired *t* test results revealed that questions 2, 3, 4, 5, 6, 7, 8, and 9 (corresponding to hypotheses 2, 3, 4, 5, 6, and 7, respectively) demonstrated statistically significant differences between pre- and postsurvey scores, with *P* values less than .05. This suggests that the app had a substantial positive impact on learners' responses to these questions, demonstrating improvements in various aspects, including confidence and preparedness for real-life patient interactions, positive beliefs, enhanced compassion, reduced negative thoughts, and a more nonprejudiced attitude toward patients. These findings further support the effectiveness of the app in enhancing the health care skills of learners.

Table 3. A 1-tail paired t test analysis between pre- and postassessment questionnaires or surveys for survey questions 1 to 9 (*P* values are rounded to nearest hundredth decimal).

Pre- and postsurvey question index	*P* value
Que-1	.11
Que-2	<.001
Que-3	<.001
Que-4	<.001
Que-5	.04
Que-6	<.001
Que-7	<.001
Que-8	.005
Que-9	.04

However, specifically for question 1 (corresponding to hypothesis 1), the 1-tailed paired *t* test results showed *P* values slightly greater than .05, indicating that the differences between pre- and postsurvey scores were not statistically significant at the 0.05 level. This outcome is likely influenced by the fact that the participants already gave near-optimal scores for question 1 in the presurvey, leaving limited room for measurable improvement. Nonetheless, the mHealth app showed positive

XSL•FO
RenderX

trends in this aspect, but due to the limited room for improvement, the result (for question 1 or hypothesis 1) was not strong enough to reach statistical significance in a tailed paired *t* test analysis.

Overall, the combination of CI analysis and the 1-tail paired *t* test provides valuable insights into the impact of the mHealth app on learners' experiences and skills. It demonstrates the app's effectiveness in enhancing learners' compassion, empathy, and nonprejudiced attitudes toward patients. The results suggest that the mHealth app can be a valuable tool in improving health care professionals' performance and providing better care experiences and outcomes for patients.

Additional Statistical Analysis

This section presents a comprehensive analysis of participants' responses to the survey questionnaires associated with the app, which is designed to enhance health care professionals' skills and promote health equity. The analysis of pre- and postsurvey data offers valuable insights into participants' learning experiences and the app's overall efficacy. Notably, as depicted in Table 4, for each survey question 1 to 9, 3 types of potential outcomes or responses were considered: (1) "learning improvement" or "positive response" which suggests the participant's improved or positive learning experience examined in the postsurvey response compared with the presurvey response, (2) "neutral response" which suggests the participant's neutral or no change learning experience examined in the postsurvey response compared with the presurvey response, and (3) "learning deterioration" or "negative response" which suggests the participant's deteriorating or negative learning experience examined in the postsurvey response compared with the presurvey response.

Table 4. Response counts based on the response type for survey questions 1 to 9.

Response type	Participants' response count								
	Que 1	Que 2	Que 3	Que 4	Que 5	Que 6	Que 7	Que 8	Que 9
Positive response or learning improvement	84	50	86	72	77	90	106	49	21
Neutral response	101	113	141	136	103	122	100	161	210
Negative response or learning deterioration	55	77	13	32	60	28	34	30	9

Exploring statistical findings and their implications provides insights into the app's capacity to cultivate compassion, empathy, and nonprejudiced attitudes toward patients. Furthermore, the app's role in enhancing health care skills and fostering a deeper understanding of patient perspectives is also examined. This comprehensive assessment illustrates the multifaceted nature of participants' learning behaviors, offering valuable guidance for health care education and practice.

Participants' advancement in learning was examined by categorizing responses into "positive" (indicating improvement) and "negative" (indicating deterioration) outcomes for survey questions 1 to 9. Neutral responses were intentionally excluded to calculate improvement rates based solely on participants who exhibited a change in their response level between the pre- and postsurvey on the 5-point scale (ie, user response levels or options). This approach focuses on visible shifts in attitudes and perceptions, while recognizing that it does not capture potential learning reinforcement among participants whose responses remained unchanged.

Notably, for question 3 (corresponding to hypothesis 3), 86/99 (87%) participants—excluding 141 neutral responses—demonstrated enhanced compassion and empathy toward patients. This underscores the mHealth app's success in nurturing emotional intelligence and effective care competencies among health care professionals.

Additionally, questions 6 and 7 (corresponding to hypothesis 5) revealed substantial growth, focusing on how participants perceive patient circumstances. For question 6, 90/118 (76.3%) participants demonstrated positive change, characterized by a reduced tendency to attribute blame to the patient for their situation. For question 7, 106/140 (75.7%) participants showed improvement, indicating heightened recognition that patients may be influenced by external, uncontrollable factors. These shifts reflect strengthened bias awareness and a deeper understanding of social determinants in clinical encounters, highlighting the mHealth app's effectiveness in enhancing these skills and attitudes.

Similarly, questions 1, 4, 5, 8, and 9 (corresponding to hypotheses 1, 4, 6, and 7) exhibited substantial learning improvement. Specifically, for question 1, 84/139 (60.4%) participants reported reduced anxiety toward patients. Question 4 showed that 72/104 (69.2%) participants felt more comfortable engaging the same patient for follow-up care. Question 5 reflected increased confidence in managing patient interactions, with 77/137 (56.2%) participants demonstrating positive change. For question 8, 49/79 (62%) participants reported stronger efforts to suppress negative thoughts, while question 9 revealed that 21/30 (70%) participants expressed a heightened commitment to providing nonprejudiced care. Collectively, these findings highlight the mHealth app's effectiveness in strengthening core aspects of health care professionalism, including emotional regulation, bias mitigation, and patient-centered attitudes.

Conversely, question 2 (corresponding to hypothesis 2) demonstrated a comparatively lower rate of learning improvement, with 50/127 (39.4%) participants showing positive change in their frustration levels toward patients, after excluding 113 participants who provided neutral responses. While the improvement percentage was modest relative to other outcomes, the simulation still delivered meaningful attitudinal shifts in how participants engaged with emotionally challenging

XSL•FO
RenderX

scenarios—underscoring its value in navigating emotionally charged patient interactions.

This analysis provides a holistic view of participants' learning improvement across various facets of the health care profession. The significant positive changes observed in key areas underscore the effectiveness of the mHealth app in fostering compassion, empathy, and nonprejudiced attitudes among health care professionals, contributing to improved patient care and health care outcomes.

Furthermore, an additional analysis was conducted to more comprehensively evaluate participants' learning improvement by combining positive (learning improvement) and neutral (or no change) responses versus negative (learning deterioration) responses across survey questions 1 to 9. For this analysis, neutral responses were grouped with positive ones to reflect any beneficial impact—even if a participant's response remained unchanged between the pre- and postsurvey. This approach provides a broader understanding of how the app contributed to learning, capturing both growth and knowledge reinforcement.

Neutral responses represent an important dimension of learning enhancement and should not be overlooked when interpreting participants' outcomes. In many cases, neutral responses occurred because participants selected optimal ratings—such as "Strongly agree" or "Strongly disagree" respectively to the question—in both the presurvey and postsurvey phases. This pattern suggests they entered the simulation with a solid foundation of knowledge or attitudes relevant to the measured construct. While these participants may not have shown numerical change, the mHealth app likely played a valuable role in reinforcing their understanding, leading to a consolidation of competencies that are essential for clinical practice.

Moreover, the design of the response scale, ranging from 1 to 5, inherently limits how participants can express marginal or qualitative gains. For individuals who already selected the highest possible rating in the presurvey, the postsurvey offered no room for upward movement—even if they did experience deeper insight, stronger commitment, or emotional growth. This structural constraint is important to consider, as it suggests that reported improvement rates may underrepresent the true impact of the simulation on highly engaged or experienced learners. As such, the neutral category includes individuals whose learning trajectory remained positive, even if not numerically reflected in postsurvey scores. Taken together, the following analysis underscores the app's positive impact on participants' learning experience, capturing both measurable improvement and reinforcement of existing knowledge.

Analyzing the results reveals that the mHealth app had a strong positive impact on participants' learning experience across all survey items. Notably, for question 3 and question 9—corresponding to hypotheses 3 and 7, respectively—227/240 (94.6%) participants and 231/240 (96.3%) participants demonstrated learning improvement, demonstrating the high effectiveness of the mHealth app in fostering compassion, empathy, and promoting nonprejudiced behavior among health care professionals toward patients.

Similarly, questions 4, 6, and 8—corresponding to hypotheses 4, 5, and 6—yielded strong outcomes, with participants showing marked learning enhancement. For question 4, 208/240 (86.7%) participants showed improvement by expressing reduced preference to avoid follow-up care with the same patient. Question 6 demonstrated positive change in 212/240 (88.3%) participants, indicating diminished tendencies to blame the patient for their circumstances and increased recognition of external influences. For question 8, 210/240 (87.5%) participants shifted away from masking negative thoughts, suggesting greater internal emotional regulation and growing comfort with expressing compassionate care authentically. These findings underscore the mHealth app's ability to support reflective attitude shifts across emotionally and ethically complex dimensions of clinical practice.

Questions 1, 2, 5, and 7—corresponding to hypotheses 1, 2, 4, and 5—also demonstrated substantial learning improvement. Specifically, for question 1 (reduced anxiety toward patients), 185/240 (77.1%) participants showed improvement. Question 2 (reduced frustration) saw gains in 163/240 (67.9%) participants, while question 5 (confidence in future patient encounters) showed progress in 180/240 (75.0%) participants. For question 7, 206/240 (85.8%) participants demonstrated increased recognition of external factors influencing patients' circumstances. Together, these findings reflect meaningful shifts in participants' ability to manage difficult emotions, anticipate future patient interactions more positively, and appreciate the broader social context of patient care—areas critical to effective, empathetic clinical practice.

Thus, the findings reflect the positive impact of the app on participants' learning experience, encompassing both learning improvement and consolidation of existing knowledge. Analyzing the users' responses provides valuable insights into the overall effectiveness of the app in advancing health care professionals' skills and attitudes, even for participants who may have already possessed some knowledge on the subject matter.

Additionally, Table 5 offers a comprehensive view of participants' overall response distributions across all levels (1 to 5) for both presurvey and postsurvey questions 1 through 9. This tabulation reinforces the effectiveness of the mHealth app by illustrating broader shifts in user sentiment and competence. It is important to clarify that Tables 4 and 5 represent distinct analytical approaches. Table 4 evaluates individual participants' directional changes, categorizing their responses as "positive (improvement), neutral (no change), or negative (deterioration)" based on pre- and postsurvey comparison. In contrast, Table 5 aggregates total counts obtained for each user response option or level (ie, 1 to 5), offering a macro-level snapshot of the entire cohort's scoring trends without pairing responses per participant. Together, these tables enrich the overall interpretation of learning impact, combining both individual-level transformations and collective score movements.

XSL•FO
RenderX

Table 5. Participants' response counts for each user response level or option for pre- and postsurvey questions 1 to 9.

User response option or level	Participants' response counts																	
	Pre-Q1	Post-Q1	Pre-Q2	Post-Q2	Pre-Q3	Post-Q3	Pre-Q4	Post-Q4	Pre-Q5	Post-Q5	Pre-Q6	Post-Q6	Pre-Q7	Post-Q7	Pre-Q8	Post-Q8	Pre-Q9	Post-Q9
1	62	77	125	99	4	4	98	113	25	27	90	132	10	14	20	21	4	4
2	69	59	57	60	11	0	50	63	49	70	76	67	20	19	25	18	0	0
3	72	69	40	54	32	12	62	40	108	90	58	29	93	40	28	23	11	9
4	26	25	9	16	85	56	17	16	49	44	12	6	72	57	58	49	31	19
5	11	10	9	11	108	168	13	8	9	9	4	6	45	110	109	129	194	208

Discussion

Principal Findings

This study evaluated a novel mHealth app built on a DEL approach aimed at enhancing health care professionals' competencies and fostering health equity. Among 240 participants, the app demonstrated statistically significant improvements in multiple skill domains—including compassion, empathy, inclusive communication, confidence in patient interactions, nonprejudiced attitudes, bias awareness, and understanding patients' perspectives. Evaluation of pre- and postsurvey responses using CI analysis, Cohen *d* effect sizes, ORs, 1-tailed paired *t* tests (α=0.05), and descriptive categorizations reinforcing the positive impact of the intervention.

The DEL approach's strength lies in immersing learners in interactive hypothetical virtual case scenarios using a first-person perspective. Participants experienced emotionally nuanced simulations designed to challenge their clinical assumptions and encourage reflective learning. The findings underscore the effectiveness of this approach in shifting learners' perspectives on SDH, patient responsibility attribution, and the relevance of cultural humility—all foundational to delivering equitable, patient-centered care. Moreover, the results demonstrate the significance of addressing biases and constraints in health care skills to foster health equity and improve health outcomes.

Moreover, the app's fully automated and self-paced nature allowed participants to engage without instructor supervision, increasing accessibility and flexibility for professional development. The results reflect substantial gains in cognitive understanding and effective learning domains—validating the app's potential as a scalable tool for health care education.

Comparison to Prior Work

Building upon initial pilot efforts [39,40], this study demonstrates significant advancements in both scale and analytical rigor. The prior version of the DEL-based mHealth app introduced foundational concepts through role-playing scenarios and a life course module; however, its scope was limited in terms of sample size, depth of evaluation, and data transparency. This study marks a prominent evolution, involving a robust sample of 240 health care professionals and using a comprehensive suite of statistical analyses of participants' responses for the pre- and postsurvey questionnaires, including 1-tailed paired *t* tests, ORs, CI estimations, and Cohen *d* effect sizes to rigorously assess the app's effectiveness and impact in the field.

In terms of educational design, this iteration features enhanced backend workflow integration, improved user interface reliability, and refined survey instruments aligned with clearly defined hypotheses. These upgrades enabled more accurate and meaningful tracking of learner outcomes, particularly in empathy, cultural humility, bias awareness, and clinical preparedness. Additionally, this study also offers a deeper exploration of app features and their role in shifting provider attitudes toward patient-centered care and equity.

Collectively, the expanded scale, methodological sophistication, and validated learning outcomes underscore the strong potential of this updated mHealth app to transform health care education. It moves beyond proof of concept and into evidence-based innovation, positioning this study as a scalable, self-directed resource capable of addressing systemic challenges in clinical training and professional development.

Strengths and Limitations

One of the primary strengths of this study lies in its integration of DEL approach into an mHealth platform, allowing health care professionals to engage with emotionally nuanced, interactive simulations in a self-directed, unsupervised manner. By offering a fully automated experience with realistic virtual case scenarios and first-person perspectives, the app fosters compassion, cultural humility, bias awareness, and reflective decision-making—all foundational to patient-centered care. Its accessibility across varying schedules and environments encourages continuous professional development, while the use of multiple analytical methods—including 1-tailed paired *t* tests, CI estimation, and effect size calculations—enhances the credibility and rigor of the evaluation. The statistically robust analysis of 240 participants adds further credibility to the intervention's effectiveness across diverse skill domains. The app's design—centered around virtual case-based simulations—created meaningful opportunities for learners to reflect on their beliefs, assumptions, and behaviors in patient care. The app's ability to deliver effective and cognitive learning at scale positions as a powerful, innovative tool for advancing health equity and transforming clinical education.

XSL•FO
RenderX

This study encountered a few limitations that could be addressed in future research. The mHealth app requires an internet connection to complete the simulation, which may limit accessibility for individuals with unreliable internet connectivity. This requirement necessitates further investigation into alternative delivery methods or offline capabilities to ensure broader access to the educational content. Additionally, the current version offers only 2 hypothetical case scenarios, which may not adequately capture the full spectrum of real-world health care challenges. Incorporating a wider range of cases with varied cultural, ethical, and clinical contexts would improve realism and learner engagement. Furthermore, the conversational context choices within the app are hardcoded, lacking the flexibility and adaptability of artificial intelligence and natural language processing technologies. Incorporating these technologies could significantly enhance the app's ability to respond dynamically to user inputs, offering a more personalized and engaging learning experience. These limitations underscore the need for future research and development efforts to address barriers to accessibility, diversify and enhance the realism of learning scenarios, and leverage advanced technologies to boost the app's effectiveness and user satisfaction.

Future Directions

Building on these findings, future work will focus on expanding the library of scenarios to represent a broader range of cultural contexts, clinical conditions, and ethical dilemmas. Optimizing system performance to reduce memory use and support offline access will increase usability across varied practice settings. Incorporating conversational intelligence via natural language processing and adaptive feedback mechanisms could transform the experience into a truly personalized learning environment.

Beyond technical upgrades, future DEL-based apps might integrate real-time expert guidance, augmented reality overlays, or gamified challenges to reinforce engagement and longitudinal learning. Similar virtual storylines or serious RPGs could be adapted for nursing, public health, or interprofessional education.

Ultimately, this research lays the groundwork for reimagining health care training through mobile experiential learning. By embedding equity-focused competencies into scalable platforms, such tools can contribute meaningfully to transforming how providers learn, reflect, and care—bringing us closer to a more empathetic, inclusive, and effective health care system.

Acknowledgments

This work was supported in part by the Ohio Department of Medicaid (grant #60077777) as part of the Medicaid Equity Simulation Project, administered by the Ohio Colleges of Medicine Government Resource Center. The overall goal of the Medicaid Equity Simulation project is to help reduce health disparities and improve patient care and health outcomes. The views expressed in this article are solely those of the authors and do not represent the views of the state of Ohio or federal Medicaid programs. This study involved human participants and was given a category 3b exemption by the Institutional Review Board of Wright State University.

Conflicts of Interest

None declared.

References

1. NHE fact sheet. US Centers for Medicare & Medicaid Services. URL: https://www.cms.gov/data-research/statistics-trends-and-reports/national-health-expenditure-data/nhe-fact-sheet [accessed 2024-05-07]
2. Mosadeghrad AM. Factors affecting medical service quality. Iran J Public Health 2014;43(2):210-220 [FREE Full text] [Medline: 26060745]
3. Hall WJ, Chapman MV, Lee KM, Merino YM, Thomas TW, Payne BK, et al. Implicit racial/ethnic bias among health care professionals and its influence on health care outcomes: a systematic review. Am J Public Health 2015;105(12):e60-e76. [doi: 10.2105/ajph.2015.302903]
4. Roberts SO, Rizzo MT. The psychology of American racism. Am Psychol 2021;76(3):475-487. [doi: 10.1037/amp0000642] [Medline: 32584061]
5. FitzGerald C, Hurst S. Implicit bias in healthcare professionals: a systematic review. BMC Med Ethics 2017;18(1):19 [FREE Full text] [doi: 10.1186/s12910-017-0179-8] [Medline: 28249596]
6. Dovidio JF, Fiske ST. Under the radar: how unexamined biases in decision-making processes in clinical interactions can contribute to health care disparities. Am J Public Health 2012;102(5):945-952. [doi: 10.2105/ajph.2011.300601]
7. Hoffman KM, Trawalter S, Axt JR, Oliver MN. Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites. Proc Natl Acad Sci USA 2016;113(16):4296-4301 [FREE Full text] [doi: 10.1073/pnas.1516047113] [Medline: 27044069]
8. Social Determinants of Health. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/socialdeterminants/index.htm [accessed 2024-05-07]
9. Social determinants of health. World Health Organization. URL: https://www.who.int/health-topics/social-determinants-of-health#tab=tab_1 [accessed 2024-05-07]

XSL•FO
RenderX

10. Chapman EN, Kaatz A, Carnes M. Physicians and implicit bias: how doctors may unwittingly perpetuate health care disparities. J Gen Intern Med 2013;28(11):1504-1510 [FREE Full text] [doi: 10.1007/s11606-013-2441-1] [Medline: 23576243]
11. Chapman MV, Hall WJ, Lee K, Colby R, Coyne-Beasley T, Day S, et al. Making a difference in medical trainees' attitudes toward Latino patients: a pilot study of an intervention to modify implicit and explicit attitudes. Soc Sci Med 2018;199:202-208 [FREE Full text] [doi: 10.1016/j.socscimed.2017.05.013] [Medline: 28532893]
12. Stone J, Moskowitz GB, Zestcott CA, Wolsiefer KJ. Testing active learning workshops for reducing implicit stereotyping of Hispanics by majority and minority group medical students. Stigma Health 2020;5(1):94-103 [FREE Full text] [doi: 10.1037/sah0000179] [Medline: 33134507]
13. Forscher PS, Mitamura C, Dix EL, Cox WT, Devine PG. Breaking the prejudice habit: mechanisms, timecourse, and longevity. J Exp Soc Psychol 2017;72:133-146 [FREE Full text] [doi: 10.1016/j.jesp.2017.04.009] [Medline: 29225371]
14. Devine PG, Forscher PS, Austin AJ, Cox WT. Long-term reduction in implicit race bias: a prejudice habit-breaking intervention. J Exp Soc Psychol 2012;48(6):1267-1278 [FREE Full text] [doi: 10.1016/j.jesp.2012.06.003] [Medline: 23524616]
15. Hasan Z. Enhancing student's hands-on learning experience through final year project in industry: implementations and challenges. Proc Int Conf Eng Educ 2009:29-33. [doi: 10.1109/iceed.2009.5490617]
16. Kolb DA. Experiential Learning: Experience as the Source of Learning and Development: FT Press; 2014.
17. He S, Zhang H. Exploring the impact of digital experiential learning on tourism education students: experimental study. J Hosp Tour Educ 2025;37(2):203-213 [FREE Full text] [doi: 10.1080/10963758.2025.2473017]
18. Kee T, Zhang H. Digital experiential learning for sustainable horticulture and landscape management education. Sustainability 2022;14(15):9116 [FREE Full text] [doi: 10.3390/su14159116]
19. Al Fatta H, Maksom Z, Zakaria M. Game-based learning and gamification: searching for definitions. Int J Simul Syst Sci Technol 2019;19:1-5. [doi: 10.5013/ijssst.a.19.06.41]
20. Libin A, Lauderdale M, Millo Y, Shamloo C, Spencer R, Green B, et al. Role-playing simulation as an educational tool for health care personnel: developing an embedded assessment framework. Cyberpsychol Behav Soc Netw 2010;13(2):217-224. [doi: 10.1089/cyber.2009.0040] [Medline: 20528282]
21. Rønning SB, Bjørkly S. The use of clinical role-play and reflection in learning therapeutic communication skills in mental health education: an integrative review. Adv Med Educ Pract 2019;10:415-425 [FREE Full text] [doi: 10.2147/amep.s202115]
22. Podjačevs P, Skorobogatova O. Computer roleplaying games as one of the most efficient methods of teaching and learning in higher education. Proc Int J Inf Technol Secur 2017:65-78 [FREE Full text]
23. Cuinas I, Sanchez MG. Role-playing games as a strategy to incorporate social skills in electrical engineering programs [education corner]. IEEE Antennas Propag Mag 2022;64(1):123-146. [doi: 10.1109/map.2021.3129689]
24. Staley A, Eastcott D. Computer-supported experiential learning (phase one - staff development). Res Learn Technol 2016;7(1):39-45. [doi: 10.1080/0968776990070107]
25. Patel DB, Ogletree X, Pei Y. Learning communication with autistic people with a mobile serious role-playing game. Proc Int Conf e-Soc Mobile Learn 2022:195-202. [doi: 10.33965/ES_ML2022_202202L025]
26. Patel DB, Ogletree X, Pei Y. Virtual expert guided serious role-playing game for learning communication with autistic people. IADIS Int J WWW/Internet 2022;20(1):62-73. [doi: 10.33965/ijwi_202220105]
27. Patel DB, Wischgol T, Pei Y, Castle A, Proulx A, Gainer D, et al. Computer-supported expert-guided experiential learning-based tools for healthcare skills. Electron Imaging 2023:1-11 [FREE Full text] [doi: 10.2352/ei.2023.35.1.vda-399]
28. Romli MH, Cheema MS, Mehat MZ, Md Hashim NF, Abdul Hamid H. Exploring the effectiveness of technology-based learning on the educational outcomes of undergraduate healthcare students: an overview of systematic reviews protocol. BMJ Open 2020;10(11):e041153 [FREE Full text] [doi: 10.1136/bmjopen-2020-041153] [Medline: 33234650]
29. Palos-Sanchez P, Saura JR, Rios Martin M, Aguayo-Camacho M. Toward a better understanding of the intention to use mHealth apps: exploratory study. JMIR Mhealth Uhealth 2021;9(9):e27021 [FREE Full text] [doi: 10.2196/27021] [Medline: 34499044]
30. Hong W. Advances and Opportunities of Mobile Health in the Postpandemic Era: Smartphonization of Wearable Devices and Wearable Deviceization of Smartphones. JMIR Mhealth Uhealth 2024 Jan 22;12:e48803 [FREE Full text] [doi: 10.2196/48803] [Medline: 38252596]
31. Ameyaw EK, Amoah PA, Ezezika O. Effectiveness of mHealth apps for maternal health care delivery: systematic review of systematic reviews. J Med Internet Res 2024;26:e49510 [FREE Full text] [doi: 10.2196/49510] [Medline: 38810250]
32. Iribarren SJ, Akande TO, Kamp KJ, Barry D, Kader YG, Suelzer E. Effectiveness of mobile apps to promote health and manage disease: systematic review and meta-analysis of randomized controlled trials. JMIR Mhealth Uhealth 2021;9(1):e21563 [FREE Full text] [doi: 10.2196/21563] [Medline: 33427672]
33. Lee M, Bin Mahmood ABS, Lee ES, Smith HE, Tudor Car L. Smartphone and mobile app use among physicians in clinical practice: scoping review. JMIR Mhealth Uhealth 2023;11:e44765 [FREE Full text] [doi: 10.2196/44765] [Medline: 37000498]

XSL•FO
RenderX

34. Liu F, Song T, Yu P, Deng N, Guan Y, Yang Y, et al. Efficacy of an mHealth app to support patients' self-management of hypertension: randomized controlled trial. J Med Internet Res 2023;25:e43809 [FREE Full text] [doi: 10.2196/43809] [Medline: 38113071]
35. Fiol-DeRoque MA, Serrano-Ripoll MJ, Jiménez R, Zamanillo-Campos R, Yáñez-Juan AM, Bennasar-Veny M, et al. A mobile phone-based intervention to reduce mental health problems in health care workers during the COVID-19 pandemic (PsyCovidApp): randomized controlled trial. JMIR Mhealth Uhealth 2021;9(5):e27039 [FREE Full text] [doi: 10.2196/27039] [Medline: 33909587]
36. de Jong A, Donelle L, Kerr M. Nurses' use of personal smartphone technology in the workplace: scoping review. JMIR Mhealth Uhealth 2020;8(11):e18774 [FREE Full text] [doi: 10.2196/18774] [Medline: 33242012]
37. Li Y. Educational game design based on experiential learning theory. 2021 Presented at: 9th International Conference on Information and Education Technology (ICIET); March 27-29, 2021; Okayama, Japan p. 190-193. [doi: 10.1109/iciet51873.2021.9419600]
38. Vozenilek J, Huff JS, Reznek M, Gordon JA. See one, do one, teach one: advanced technology in medical education. Acad Emerg Med 2004;11(11):1149-1154 [FREE Full text] [doi: 10.1197/j.aem.2004.08.003] [Medline: 15528578]
39. Hershberger PJ, Pei Y, Crawford TN, Neeley SM, Wischgoll T, Patel DB, et al. An interactive game with virtual reality immersion to improve cultural sensitivity in health care. Health Equity 2022;6(1):189-197 [FREE Full text] [doi: 10.1089/heq.2021.0128] [Medline: 35402778]
40. Patel DB, Pei Y, Vasoya M, Hershberger PJ. Computer-supported experiential learning-based tool for healthcare skills. IEEE Comput Grap Appl 2023;43(2):57-68. [doi: 10.1109/mcg.2023.3242921]
41. The national organization of urban MCH leaders. CityMatCH. URL: https://www.citymatch.org/ [accessed 2024-06-24]
42. Unity technologies. Unity. 2025. URL: https://www.unity.com [accessed 2025-07-20]
43. Reallusion hub. Reallusion. 2025. URL: https://www.reallusion.com/hub [accessed 2025-07-20]

Abbreviations

DEL: digital experiential learning
mHealth: mobile health
NHE: National Health Expenditure
OR: odds ratio
RPG: role-playing game
SDH: social determinants of health
VR: virtual reality

Edited by J Sarvestan; submitted 09.07.24; peer-reviewed by M Aksoy, M Teaford; comments to author 15.01.25; revised version received 11.06.25; accepted 26.08.25; published 10.10.25.

Please cite as:
Patel DB, Patel MB, Wischgoll T, Pei Y, Hershberger PJ
A Virtual Reality–Based mHealth App to Enhance Health Care Skills and Promote Health Equity by Empowering Health Care Professionals to Provide Empathetic, Compassionate, and Unbiased Patient Care Through Digital Experiential Learning: Qualitative Study
JMIR Form Res 2025;9:e64147
URL: https://formative.jmir.org/2025/1/e64147
doi:10.2196/64147
PMID:40864878

©Dixit Bharatkumar Patel, Mayank Bharatkumar Patel, Thomas Wischgoll, Yong Pei, Paul J Hershberger. Originally published in JMIR Formative Research (https://formative.jmir.org), 10.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

User Engagement With and Perceived Impact of a Digital Cognitive Training App on Cognition, Daily Functioning, and Mental Fitness: Secondary Analysis of Cross-Sectional Survey Data

Kelsey McAlister[1], PhD; Lara Baez[1], PhD; Anna Braunsdorf[2], BJ; Yana Yang[2], MSc; Dan Kessler[2], MBA; Jennifer Huberty[1], PhD

[1]Fit Minded, Inc, Phoenix, AZ, United States

[2]The Mind Company, San Francisco, CA, United States

Corresponding Author:
Kelsey McAlister, PhD
Fit Minded, Inc
2901 E Greenway Road
PO Box 30271
Phoenix, AZ, 85046
United States
Phone: 1 602 935 6986
Email: kelsey@fit-minded.com

Abstract

Background: Cognitive difficulties are common and can interfere with daily functioning. While digital cognitive training apps are widely used, few studies have examined whether personalized tools support perceived improvements in cognitive functioning, daily functioning, and overall mental fitness among general adult users.

Objective: The purpose of this secondary analysis was to explore the self-reported cognitive benefits of Elevate, a commercial, personalized cognitive training app developed to support cognitive functioning, as well as engagement with the app. We aimed to (1) describe demographics, engagement metrics, and self-reported improvements; (2) examine associations between app engagement and self-reported improvements in cognitive functioning skills directly targeted by the app; and (3) examine associations between app engagement and self-reported improvements in daily functioning and overall mental fitness as potential transfer effects of cognitive training.

Methods: Adult Elevate users (aged ≥18 years) who used the app at least twice in the previous 30 days completed a brief web-based survey on perceived cognitive, functional, and mental fitness improvements. Responses were linked to objective app use data, including total active weeks, mean active days per week, and mean time per day. Ordinal logistic regressions tested associations between engagement metrics and self-reported outcomes controlling for demographic variables. A Bonferroni correction was applied to adjust for multiple comparisons.

Results: A total of 3367 adult Elevate users were included. Participants were primarily middle-aged (mean 55, SD 16 y), White (2557/3336, 76.65%), and female (2184/3362, 64.96%), with 67.72% (2274/3358) holding at least a college degree. Using the app across more weeks was associated with a greater likelihood of reporting improvements in all cognitive skills (odds ratios [ORs] 1.0014-1.0027, 95% CI 1.0006-1.0036), several areas of daily functioning (eg, motivation and task efficiency; ORs 1.0014-1.0017, 95% CI 1.0006-1.0026), and overall mental fitness (OR 1.0021, 95% CI 1.0012-1.0031). More days of use per week were linked to improvement in math only (OR 1.15, 95% CI 1.09-1.22), whereas spending more time per day was associated with improvements in speaking, reading, math, motivation, personal progress, and mental fitness (ORs 1.02-1.04, 95% CI 1.01-1.06).

Conclusions: Greater use of the Elevate app was linked to self-reported improvements in cognitive skills, daily functioning, and overall mental fitness. These findings suggest that personalized, adaptive cognitive training apps such as Elevate may serve as scalable tools for enhancing everyday cognitive and functional well-being. Future research should use rigorous, longitudinal methods to confirm these effects and clarify which app features drive meaningful outcomes.

(JMIR Form Res 2025;9:e80027) doi:10.2196/80027

XSL•FO
RenderX

KEYWORDS

cognitive functioning; mobile health; real-world evidence; digital health; app-based intervention; mental performance

Introduction

Background

Cognitive difficulties such as challenges with memory, attention, and information processing are increasingly common, with estimates suggesting that 2 out of 3 Americans will experience some level of cognitive difficulty [1]. Even minor declines in cognitive abilities (eg, slower processing and reduced working memory) can impair everyday tasks such as managing finances and organizing schedules [2-4]. When these difficulties persist over time, they are associated with more than twice the risk of developing mild cognitive impairment or dementia within the next decade [5]. This risk increases as individuals age as 75% of older adults with ongoing cognitive concerns go on to develop clinical symptoms within 10 years [5]. These everyday difficulties are often intensified by stressors such as chronic stress, frequent multitasking, and digital overload, which place ongoing strain on attention and executive function [6,7]. For example, multitasking between multiple forms of screen media (eg, texting, watching television, and engaging with social media) has been associated with greater cognitive difficulties, including poorer working memory, reduced attentional filtering, and increased vulnerability to distraction [7]. As such, there is a growing need for scalable interventions that prevent cognitive difficulties and promote cognitive functioning in everyday life.

Mental fitness is a novel framework for assessing cognitive difficulties. Although it is an emerging concept that has not yet been empirically defined, we define mental fitness as a state of optimal cognitive and emotional functioning, supporting the ability to regulate emotions, think clearly, and apply knowledge effectively to navigate daily demands [8-13]. Rather than focusing solely on disease prevention or symptom reduction, mental fitness emphasizes the ongoing development of capacities that support real-world cognitive functioning and adaptability [1,11,13]. While related to constructs such as cognitive health and executive function, mental fitness is a dynamic, trainable construct that integrates both cognitive and emotional domains and is theorized to support everyday functioning. While mental fitness is gaining traction as an important construct, it lacks both a universally accepted definition and valid measurement approaches [8-11]. Nonetheless, this novel framing offers a useful theoretical foundation for evaluating how tools and practices may enhance cognitive functioning and prevent decline, especially in nonclinical or wellness-oriented populations.

A variety of interventions have been developed to prevent cognitive difficulties and support cognitive functioning as a core component of mental fitness, including lifestyle interventions (eg, physical activity, mindfulness, and stress management) and compensatory strategies (eg, note-taking and time management techniques) [14-16]. Lifestyle interventions may promote general brain health but typically do not involve targeted practice of specific cognitive skills [15,16]. Compensatory strategies help individuals work around cognitive difficulties but also do not target underlying cognitive abilities [14]. In contrast, cognitive training (ie, the active use and challenge of mental skills) is a promising pathway for maintaining or enhancing cognitive function, with evidence linking intellectually stimulating activities to lower risk of cognitive difficulties and clinical impairments (eg, dementia) [17-19].

A growing number of digital cognitive training tools aim to address cognitive difficulties and support functioning through brief exercises, often delivered via mobile apps targeting aspects of cognitive functioning such as attention, memory, and processing speed. While studies of research-based and commercial platforms have shown improvements in cognitive performance on specific trained tasks, research linking app-based cognitive training to related untrained tasks (ie, near-transfer effects) and real-world functioning (ie, far-transfer effects) is mixed [20,21]. Meta-analyses have shown that cognitive training tools may lead to improvements beyond the specific tasks trained. However, it remains unclear whether these gains translate into perceived improvements in real-world functioning or broader aspects of mental fitness, particularly among users outside of controlled research settings [20,22,23]. Furthermore, many studies evaluating the effectiveness of digital tools for cognitive decline have been conducted in specialized populations (eg, older adults and individuals with cognitive impairments), limiting generalizability to the broader adult population [24-27]. Although several commercial platforms have demonstrated effectiveness in real-world users [28-31], few studies have evaluated apps with more advanced technological features, such as a wide range of content coverage and adaptive personalization. These gaps point to the need for more research on commercial apps that incorporate more advanced features to determine their potential to support cognitive functioning in everyday life and overall mental fitness.

Objectives

The purpose of this secondary analysis was to explore app engagement and the self-reported cognitive benefits of Elevate, a commercial, personalized cognitive training app developed to support cognitive functioning. We aimed to (1) describe demographics, engagement metrics, and self-reported improvements; (2) examine associations between app engagement and self-reported improvements in cognitive functioning skills directly targeted by the app; and (3) examine associations between app engagement and self-reported improvements in daily functioning and overall mental fitness as potential transfer effects of cognitive training. This study offers novel insights by evaluating a personalized commercial cognitive training app in a real-world context, linking objective use data with self-reported outcomes, and exploring whether cognitive training supports broader mental fitness beyond task-specific performance. The findings will help clarify the potential of widely accessible platforms to serve as scalable tools for promoting everyday cognitive health in general adult populations.

XSL•FO
RenderX

Methods

Study Design and Participants

This study was a cross-sectional observational analysis involving secondary use of survey and app use data. Participants were adult users (aged ≥18 years) of Elevate, a cognitive training app, who met the following inclusion criteria: (1) ability to read and understand English, (2) being located within the United States, and (3) engagement with the app at least twice in the previous 30 days. Engagement was defined as completing any activity within the app, such as a game-based training exercise. Eligible users were identified via back-end data and invited via email (sent up to 2 times) to complete an optional survey assessing their perceptions of the Elevate app's impact on cognitive skills and everyday functioning. Participants who completed the survey were entered into a drawing to win a US $50 gift card; this incentive was offered by the company and was independent of the research protocol. The survey was administered between June 10 and 23, 2025. User engagement data were extracted from Elevate's back-end activity logs, spanning April 18, 2014, to June 23, 2025, and were linked to survey respondents to support exploratory analyses of perceived outcomes in relation to app engagement. Among 84,505 eligible users invited to participate via email, 4190 (4.96%) opened the survey, and 3367 (3.98%) completed it in full. This yielded an 80.36% (3367/4190) completion rate among users who accessed the survey, which, in addition to the response rate, is consistent with response patterns observed in large-scale, voluntary online surveys and still supports valid inferences given the sample size and analytic approach [32-34]. Only completed surveys were included in the analytic sample.

Ethical Considerations

The study protocol was reviewed by the Biomedical Research Alliance of New York Institutional Review Board (25-149-1708) and determined to be exempt under US federal guidelines for research involving human participants. Because the study met criteria for exemption, informed consent was not required. All data were deidentified before analysis to protect participant confidentiality.

Cognitive Training App

Elevate is a commercially available mobile app designed to enhance cognitive skills through personalized, curriculum-based training. Users find Elevate via app store searches (eg, Apple App Store and Google Play Store), digital advertisements, word of mouth, and media coverage. The app was developed by The Mind Company, with content created by an internal team of curriculum designers, cognitive scientists, and subject matter experts. Drawing on principles from cognitive science, Elevate delivers short, interactive exercises intended to strengthen real-world cognitive functioning rather than isolated task performance.

Training is organized across 5 primary skill domains: writing, speaking, reading, math, and memory. Writing modules target clarity, grammar, and vocabulary to improve written communication. Speaking modules focus on diction, pronunciation, and word retrieval to support clear and confident verbal expression. Reading modules address comprehension, processing speed, and vocabulary development. Math modules emphasize practical numeracy skills, including estimation, proportions, and financial calculations. Memory modules include tasks designed to improve short-term recall, sequencing, and synthesis. An adaptive learning algorithm adjusts difficulty based on user performance, enabling tailored progression. Immediate feedback is provided, and users can track their progress and compare their performance to peer benchmarks.

Self-Reported Improvements in Cognition, Daily Functioning, and Mental Fitness (Outcomes)

Survey questions were created specifically for this study and were not taken from existing standardized tools. The survey was developed by 2 doctoral-level researchers with expertise in behavior change, complementary health approaches, and digital health. The survey was designed to evaluate users' perceptions of how using the Elevate app affected their cognitive skills, daily functioning, and overall mental fitness. These topics were selected because they closely aligned with the app's training goals and the anticipated real-world benefits of cognitive training. The survey was web based (delivered using SurveyMonkey [SurveyMonkey Inc]) and took most participants 5 to 10 minutes to complete. All responses were anonymous. Participants completed 27 quantitative items assessing (1) perceived changes in specific cognitive skills targeted by the app (14 items) and (2) perceived changes in overall mental fitness and daily functioning since using Elevate (13 items). All items were rated on a 5-point Likert scale ranging from "Much worse" to "Much better." At the end of the survey, participants completed demographic questions, including age, gender, race and ethnicity, educational level, English-language fluency, and employment status. Although the survey included additional items, only a subset of items relevant to the aims of this study was analyzed and is reported in this paper. Table 1 summarizes the cognitive skills, daily functioning, and mental fitness domains examined in this study and the specific survey items used to assess each domain.

XSL•FO
RenderX

Table 1. Cognitive skills, daily functioning, and mental fitness domains and their corresponding survey items.

Domain	Survey item
Cognitive skills	
Writing	"Writing clearly and effectively"
Speaking	"Speaking clearly and confidently"
Reading	"Understanding what I read"
Math	"Doing mental math quickly and accurately"
Memory	"Remembering things when I need to"
Daily functioning	
Task efficiency	"Completing tasks efficiently"
Spending time intentionally	"Spending my time in ways that feel intentional and worthwhile"
Keeping up with responsibilities	"Keeping up with personal and work responsibilities"
Feeling motivated	"Feeling motivated to grow or improve"
Making personal progress	"Feeling a sense of making personal progress"
Overall mental fitness	
Mental fitness	"Overall mental fitness (ie, how sharp, calm, and knowledgeable you feel)"

App Engagement Data

Objective engagement data were obtained from Elevate's internal use tracking system starting from each user's first recorded activity and linked to survey responses. We examined 3 app engagement metrics. Total active weeks was defined as the number of weeks in which a user engaged with the app at least once. Mean active days per week was calculated as the average number of days in which users used the app at least once during each active week. Mean time per day (in minutes) was defined as the average duration of app use on active days.

Statistical Analysis

Descriptive statistics, including mean, SD, total counts, and the corresponding percentages, were calculated for demographic variables, app engagement variables, and self-reported improvement variables. A series of ordinal logistic regressions (cumulative link models [35]) were used to test the effects of the number of total active weeks, mean active days per week, and mean time per day on self-reported improvements in cognitive skills (writing, speaking, reading, math, and memory), daily functioning (task efficiency, spending time intentionally, keeping up with responsibilities, feeling motivated, and making personal progress), and overall mental fitness. The proportional odds assumption was tested for predictors of interest in each model (Multimedia Appendix 1). Each ordinal dependent variable was modeled separately. All models included the effects of covariates, including age, gender, race, ethnicity, educational level, and time since downloading Elevate, to reduce confounding. All predictors were retained in all models regardless of statistical significance to maintain comparability across analyses. All models were conducted on complete cases (ie, no missing predictor, outcome, or covariate data). Odds ratios (ORs) were calculated from the model estimates. To adjust for multiple comparisons across 11 ordinal regression models with 3 main predictors, a Bonferroni correction was applied. The α level was adjusted to α=.05/33=.0015, and statistical significance was evaluated at the corrected threshold [36]. All analyses were conducted in R (version 4.3.1; R Foundation for Statistical Computing) [37]. This study is reported in accordance with the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines, with the completed checklist provided in Multimedia Appendix 2.

Results

Descriptive Statistics of Demographic Variables

Users (N=3367) were, on average, aged 55 (SD 16) years, mostly White or European (2557/3336, 76.65%), non-Hispanic (3113/3341, 93.18%), and native English speakers (3063/3360, 91.16%), and most identified as women (2184/3362, 64.96%). Users were also highly educated (2274/3358, 67.72% had at least a bachelor's degree) and mostly employed (1481/3359, 44.09%) or retired (1153/3359, 34.33%; Table 2).

XSL•FO
RenderX

Table 2. Demographic characteristics of Elevate users.

Characteristic	Values
Age (y), mean (SD)	55 (16)
Race (n=3336), n (%)	
American Indian or Alaska Native	17 (0.5)
Asian or Asian American	202 (6.1)
Black or African American	266 (8)
MENA[a]	12 (0.4)
Multiracial	194 (5.8)
Native Hawaiian or Pacific Islander	5 (0.1)
White or European American	2557 (76.6)
Other	83 (2.5)
Gender (n=3362), n (%)	
Woman	2184 (65)
Man	1105 (32.9)
Gender diverse	29 (0.9)
Other	10 (0.3)
Declined to state	34 (1)
Hispanic ethnicity (n=3341), n (%)	254 (7.6)
English fluency (n=3360), n (%)	
“I am a native speaker of English”	3063 (91.2)
“I speak English fluently” (near native)	213 (6.3)
“I speak English well” (intermediate)	79 (2.4)
“I speak a little English” (basic or novice)	3 (0.1)
“I do not speak English”	2 (0.1)
Educational level (n=3358), n (%)	
11th grade or lower	9 (0.3)
High school diploma or GED[b]	183 (5.4)
Some college	465 (13.8)
2-y college or technical school degree	379 (11.3)
Bachelor’s degree or equivalent (eg, BA or BS)	1186 (35.3)
Graduate degree or equivalent (eg, MA, MD, or PhD)	1088 (32.4)
Other	48 (1.4)
Employment status (n=3359), n (%)	
Full-time employment	1481 (44.1)
Part-time employment	313 (9.3)
Student	71 (2.1)
Retired	1153 (34.3)
Unemployed	172 (5.1)
Other	169 (5)

[a]MENA: Middle Eastern or North African.

[b]GED: General Educational Development.

XSL•FO
RenderX

Descriptive Statistics of Engagement and Self-Reported Improvement Variables

Overall, users were highly engaged with the Elevate app as the average time since downloading was 1397 (SD 1281) days, which corresponds to 3.83 (SD 3.5) years. Users were active for an average of 120 (SD 130) weeks and an average of 5.42 (SD 1.42) days per week during active weeks and spent an average of 6.5 (SD 6.0) minutes per active day using the app (Table 3).

Table 3. Descriptive statistics of engagement variables.

Variable	Participants with data, n	Participants, mean (SD)
Time since downloading Elevate (d)	3367	1397 (1281)
Number of total active weeks	3361	120 (130)
Number of sessions per weeks	3361	5.42 (1.42)
Time spent per day (min)	3269	6.5 (6.0)

Most users reported improvement (ie, responding either "Better" or "Much better") across all cognitive skills (range 1915/3365, 56.9%-2534/3366, 75.4%), most daily functioning items (range 1626/3361, 48.3%-2932/3364, 87.3%), and overall mental fitness (2480/3364, 84.4%). Reports of deterioration (ie, "Worse" or "Much worse") were rare, ranging from 0.1% (3/3359) to 1.5% (53/3365) across all items (Table 4).

Table 4. Distribution of responses to self-report questions about Elevate experience.

Domain	Much worse, n (%)	Worse, n (%)	No change, n (%)	Better, n (%)	Much better, n (%)
Cognitive skills					
Writing (n=3362)	1 (0)	18 (0.5)	1181 (35.1)	1652 (49.1)	510 (15.2)
Speaking (n=3353)	2 (0.1)	26 (0.8)	1363 (40.7)	1538 (45.9)	424 (12.6)
Reading (n=3358)	3 (0.1)	9 (0.3)	1275 (38)	1618 (48.2)	453 (13.5)
Math (n=3366)	3 (0.1)	23 (0.7)	798 (23.8)	1741 (51.8)	793 (23.6)
Memory (n=3365)	3 (0.1)	50 (1.5)	1397 (41.5)	1559 (46.3)	356 (10.6)
Daily functioning					
Task efficiency (n=3360)	0 (0)	11 (0.3)	1398 (41.6)	1565 (46.6)	386 (11.5)
Spending time intentionally (n=3363)	1 (0)	34 (1)	983 (29.2)	1740 (51.7)	605 (18)
Keeping up with responsibilities (n=3364)	1 (0)	24 (0.7)	1713 (50.9)	1290 (38.3)	336 (10)
Feeling motivated (n=3361)	0 (0)	3 (0.1)	476 (14.2)	1962 (58.4)	920 (27.4)
Making personal progress (n=3362)	2 (0.1)	5 (0.1)	423 (12.6)	2019 (60.1)	913 (27.2)
Overall mental fitness					
Mental fitness (n=3364)	0 (0)	5 (0.1)	519 (15.4)	2265 (67.3)	575 (17.1)

Association Between Engagement Metrics and Improvements in Cognitive Skills

A series of ordinal logistic regressions was used to test the effects of the number of total active weeks, mean active days per week, and mean time per day on self-reported improvements in each of the cognitive skills, adjusting for age, gender, race, ethnicity, educational level, and time since downloading Elevate. After Bonferroni correction (α=.0015), a greater number of total active weeks was associated with a greater likelihood of reporting improvements in writing (OR 1.0027, 95% CI 1.0019-1.0036), speaking (OR 1.0022, 95% CI 1.0013-1.0030), reading (OR 1.0023, 95% CI 1.0014-1.0031), math (OR 1.0016, 95% CI 1.0007-1.0025), and memory (OR 1.0014, 95% CI 1.0006-1.0023). A greater number of mean active days per week was associated with a greater likelihood of reporting improvements in math (OR 1.15, 95% CI 1.09-1.22). Greater mean time per day on the app was associated with a greater likelihood of reporting improvements in speaking (OR 1.02, 95% CI 1.01-1.03), reading (OR 1.03, 95% CI 1.01-1.04), and math (OR 1.03, 95% CI 1.02-1.04; Table 5).

XSL•FO
RenderX

Table 5. Results from a series of ordinal logistic regressions conducted separately for each outcome (n=3160)a.

Outcome	Predictor								
	Number of active weeks			Mean active days per week			Mean time per day		
	b (SE)	OR[b] (95% CI)	*P* value[c]	*b* (SE)	OR (95% CI)	*P* value[c]	*b* (SE)	OR (95% CI)	*P* value[c]
Cognitive skills									
Writing	0.0027 (0.00045)	1.0027 (1.0019-1.0036)	<.001	−0.0011 (0.029)	1.00 (0.94-1.06)	.97	0.018 (0.0057)	1.02 (1.01-1.03)	.002
Speaking	0.0022 (0.00044)	1.0022 (1.0013-1.0030)	<.001	0.026 (0.029)	1.03 (0.97-1.09)	.37	0.020 (0.0057)	1.02 (1.01-1.03)	<.001
Reading	0.0023 (0.00044)	1.0023 (1.0014-1.0031)	<.001	0.050 (0.029)	1.05 (0.99-1.11)	.09	0.025 (0.0058)	1.03 (1.01-1.04)	<.001
Math	0.0016 (0.00044)	1.0016 (1.0007-1.0025)	<.001	0.14 (0.029)	1.15 (1.09-1.22)	<.001	0.029 (0.0061)	1.03 (1.02-1.04)	<.001
Memory	0.0014 (0.00044)	1.0014 (1.0006-1.0023)	.001	0.063 (0.029)	1.07 (1.01-1.13)	.03	0.017 (0.0056)	1.02 (1.01-1.03)	.002
Daily functioning									
Task efficiency	0.0014 (0.00044)	1.0014 (1.0006-1.0023)	.001	0.055 (0.029)	1.06 (1.00-1.12)	.06	0.015 (0.0063)	1.02 (1.00-1.03)	.02
Spending time intentionally	0.0016 (0.00044)	1.0016 (1.0007-1.0025)	<.001	−0.016 (0.029)	0.98 (0.93-1.04)	.58	0.017 (0.0063)	1.02 (1.00-1.03)	.008
Keeping up with responsibilities	0.0016 (0.00044)	1.0016 (1.0007-1.0025)	<.001	0.063 (0.029)	1.06 (1.01-1.13)	.03	0.0079 (0.0064)	1.01 (1.00-1.02)	.22
Feeling motivated	0.0017 (0.00046)	1.0017 (1.0008-1.0026)	<.001	−0.0088 (0.030)	0.99 (0.94-1.05)	.77	0.027 (0.0064)	1.03 (1.01-1.04)	<.001
Making personal progress	0.0011 (0.00046)	1.0011 (1.0002-1.0020)	.02	0.093 (0.030)	1.10 (1.03-1.16)	.002	0.027 (0.0065)	1.03 (1.02-1.04)	<.001
Overall mental fitness									
Mental fitness	0.0021 (0.00048)	1.0021 (1.0012-1.0031)	<.001	0.065 (0.032)	1.07 (1.00-1.14)	.04	0.042 (0.0066)	1.04 (1.03-1.06)	<.001

[a]All models adjusted for time since downloading the Elevate app, age, gender, race, ethnicity, and educational level.

[b]OR: odds ratio.

[c]The Bonferroni-corrected α on 33 tests was .0015.

Association Between Engagement Metrics and Improvements in Daily Functioning and Mental Fitness

A series of ordinal logistic regressions was used to test the effects of the number of total active weeks, mean active days per week, and mean time per day on self-reported improvements in each of the daily functioning and overall mental fitness outcomes, adjusting for age, gender, race, ethnicity, educational level, and time since downloading Elevate. After Bonferroni correction (α=.0015), a greater number of active weeks was associated with a greater likelihood of reporting improvements in task efficiency (OR 1.0014, 95% CI 1.0006-1.0023), spending time intentionally (OR 1.0016, 95% CI 1.0007-1.0025), keeping up with responsibilities (OR 1.0016, 95% CI 1.0007-1.0025), feeling motivated (OR 1.0017, 95% CI 1.0008-1.0026), and overall mental fitness (OR 1.0021, 95% CI 1.0012-1.0031). Mean active days per week did not predict any self-reported change in daily functioning outcomes or overall mental fitness. Greater mean time per day on the app was associated with a greater likelihood of reporting improvements in feeling motivated (OR 1.03, 95% CI 1.01-1.04), making personal progress (OR 1.03, 95% CI 1.02-1.04), and overall mental fitness (OR 1.04, 95% CI 1.03-1.06; Table 5).

XSL•FO
RenderX

Discussion

Principal Findings

The purpose of this secondary analysis was to explore app engagement and self-reported cognitive benefits of Elevate, a commercial, personalized cognitive training app developed to support cognitive functioning. We aimed to (1) describe demographics, engagement metrics, and self-reported improvements; (2) examine associations between app engagement and self-reported improvements in cognitive functioning skills directly targeted by the app; and (3) examine associations between app engagement and self-reported improvements in daily functioning and overall mental fitness as potential transfer effects of cognitive training. Most Elevate users reported perceived improvements across cognitive skills, daily functioning, and overall mental fitness. Users who engaged with the app across more total active weeks (ie, used the app at least once during the week) were more likely to report improvements in all cognitive skill areas. More frequent use within active weeks (ie, using the app on more days) was linked to perceived improvement in math, whereas spending more time per day on the app was associated with improvements in speaking, reading, and math. Greater weekly engagement was also tied to perceived gains in aspects of daily functioning, including task efficiency, intentional time use, staying on top of responsibilities, and motivation. In addition, spending more time per day on the app was linked to feeling more motivated and making personal progress. Both more weeks of use and longer daily sessions were associated with a higher likelihood of users reporting improvements in overall mental fitness.

Elevate users were primarily middle-aged (mean age 55, SD 16 years), with most identifying as women (2184/3362, 64.96%) and White individuals (2557/3336, 76.65%), and most reported a high level of education (2274/3358, 67.72% had a college degree or higher). These findings align with those of previous research indicating that middle-aged, well-educated adults, especially women, are frequent users of digital health and cognitive platforms [38,39]. Engagement data indicated consistent use across multiple weeks, and most users endorsed perceived improvements in cognitive skills, daily functioning, and overall mental fitness (range: 1626/3361, 48.3%–2932/3364, 87.2%). These findings could be attributed to the self-selected sample, which may have excluded dissatisfied or less engaged users, or to placebo or expectancy effects. Given that middle age may be an important time for cognitive health behaviors, the engagement patterns observed in our sample suggest that this demographic may be receptive to cognitive training interventions, although additional research is needed to determine the app's objective cognitive health benefits [40,41].

Elevate use was associated with perceived improvements in the core cognitive skills it was designed to target. More weeks of app use were associated with perceived improvements across all cognitive functioning domains, including writing, reading, speaking, math, and memory, which is in line with previous research showing that sustained engagement with cognitive training programs leads to improved cognitive functioning [42-44]. Engaging with the Elevate app on more days during an active week was associated with perceived improvement in math, whereas spending more time per day was linked to gains in speaking, reading, and math. These patterns may reflect differences in how cognitive skills are reinforced, with math potentially benefiting from frequent, distributed practice, whereas verbal skills may be more responsive to sustained, focused engagement that supports fluency and comprehension [45,46]. While previous studies of other commercial apps have reported gains in attention and processing speed [29,31], greater Elevate use was linked to self-reported improvements across a broad range of cognitive domains. Elevate's personalized training structure, which tailors session content to individual skill levels and goals, may contribute to these patterns by aligning challenge and repetition to individual needs, thereby fostering engagement in targeted domains. While these associations were significant, the effect sizes were relatively small, limiting practical significance. Future research using validated outcome measures and objective cognitive assessments is critical to determine whether these small self-reported improvements reflect meaningful cognitive gains or are influenced by user expectations and recall bias.

A greater number of weeks with any app use was associated with perceived improvements in daily functioning, including task efficiency, intentional time use, keeping up with responsibilities, and motivation. In addition, users who spent more daily time on the app were more likely to report feeling motivated and making personal progress. Engaging in core cognitive skill training has been linked to self-regulation skills, such as planning, sustained effort, and goal-directed behavior, that relate to everyday productivity and well-being [47,48]. Such perceived transfer effects are rarely examined in commercial cognitive training studies, which tend to focus more narrowly on task-specific performance metrics [29,31,49,50]. While one study of a commercial cognitive training app showed that using a cognitive training app more often was linked to better scores on related cognitive tests [51], this study extends this work by examining the effects of cognitive training on user perceptions of functional improvements. However, the practical significance of these small effects remains uncertain as the mechanisms underlying users' perceived benefits are unclear. Prior work in digital health has shown that psychological empowerment and enjoyment of app use can drive behavior change, pointing to potential mechanisms underlying users' perceived functional gains in this study [52]. Future research should use validated functional outcome measures and experimental designs to test these mechanisms and determine which app features most effectively promote meaningful improvements in daily functioning.

To our knowledge, this is the first study to show that greater use of a commercial cognitive training app is linked to self-reported improvements in overall mental fitness. Much of the existing research on commercial cognitive training apps focuses on performance-based outcomes over the subjective abilities to regulate emotions, think clearly, and apply knowledge effectively (ie, capacities that define mental fitness [8-13,28,30,49,50]). This study builds on existing research by showing preliminary evidence that users may perceive psychological benefits from cognitive training. In doing so, it

XSL•FO
RenderX

points to mental fitness as an outcome that merits scientific attention and the development of rigorous measurement tools. Future research should focus on creating and rigorously validating a standardized mental fitness assessment tool and also examine which aspects of Elevate's training model (eg, domain variety, adaptive difficulty, and perceived personal relevance) are most effective in promoting sustained improvements in users' mental fitness.

Strengths and Limitations

This study offers several notable strengths. First, it linked self-reported outcomes with objective, real-world app use data from a commercial cognitive training platform, enabling a more ecologically valid understanding of how such tools are experienced in naturalistic settings. Second, it moves beyond traditional performance-based outcomes to assess a broader set of perceived benefits, including daily functioning and overall mental fitness. To our knowledge, this is one of the first studies to capture user perceptions of mental fitness, a rising area of interest that encompasses clarity of thought, emotional regulation, motivation, and the ability to apply cognitive skills in everyday life. Third, by analyzing detailed engagement metrics, this study identified use patterns associated with perceived gains, helping inform hypotheses about dose-response relationships. Fourth, this study examined a general population sample of adults—rather than clinical or older adult populations—offering insights into how cognitive training may benefit individuals in everyday, nonclinical contexts. Finally, this study contributes novel insights into user engagement and perceived cognitive and functional benefits of a widely available, commercial cognitive training app, leveraging data from a large, real-world sample to explore how digital tools may support cognitive health and mental fitness in everyday life.

There are also limitations. First, the cross-sectional study design without a control condition precludes us from determining whether our findings reflect genuine benefits of Elevate use, natural improvement over time, or user expectations about the app's effectiveness. Future studies should use longitudinal experimental designs to better assess causal relationships between app engagement and outcomes. Second, this study relied entirely on self-reported outcomes, and the items used were developed specifically for this study rather than being drawn from validated instruments. Relatedly, the self-reported outcomes were based on the construct of mental fitness, which currently lacks an empirical definition and validated assessment tools. While we did assess perceived improvements in daily functioning domains and mental fitness relevant to everyday life, the absence of both objective outcome data and validated measurement tools limits the reliability and interpretability of these findings and could have led to measurement bias. Future research should incorporate validated cognitive and functional assessment tools, as well as behavioral or performance-based outcomes, to complement subjective reports. Fourth, this study did not include a control or comparison group, making it difficult to account for expectancy effects or alternative explanations for perceived change. Randomized controlled trials or matched comparison studies could help isolate the unique effects of the intervention. Fifth, the sample was predominantly composed of White, highly educated women, which limits the generalizability of the findings to more diverse populations. Finally, the self-selected sample was a relatively small proportion of the app's total user base, which could have led to biased or inflated results because more engaged or satisfied users may have been more likely to volunteer to share their experiences in the survey. Future studies should include broader recruitment strategies to reach more diverse user profiles, including lower-engagement or disengaged users, and ensure representation across demographic groups.

Conclusions

This study found that greater use of the Elevate app was associated with perceived improvements in a range of cognitive skills, including reading, writing, speaking, math, and memory, as well as various outcomes related to perceived daily functioning, including task efficiency, spending time intentionally, keeping up with responsibilities, feeling motivated, and making personal progress. Users also reported perceived gains in mental fitness, a multidimensional construct encompassing the ability to regulate emotions, think clearly, and apply knowledge effectively. Despite limitations such as unvalidated self-report measures and small effect sizes, these findings provide preliminary evidence that users see cognitive training apps as helpful not only for domain-specific skills but also for broader benefits. Future research using rigorous designs (eg, controlled trials) is needed to establish whether these perceptions reflect actual improvements in cognitive skills and daily functioning and determine which app features best support lasting cognitive and functional improvements.

Acknowledgments

This work was supported by The Mind Company.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to The Mind Company's policy for user data but are available from the corresponding author on reasonable request.

Authors' Contributions

KM contributed to conceptualization, methodology, formal analysis, visualization, writing—original draft, writing—review and editing, and final approval. LB contributed to conceptualization, formal analysis, writing—original draft, writing—review and editing, and final approval. AB contributed to conceptualization, resources, writing—original draft, writing—review and editing,

XSL•FO
RenderX

and final approval. YY contributed to conceptualization, data curation, writing—original draft, writing—review and editing, and final approval. DK contributed to conceptualization, resources, writing—original draft, writing—review and editing, and final approval. JH contributed to conceptualization, methodology, writing—original draft, writing—review and editing, final approval, and supervision.

Conflicts of Interest

AB, YY, and DK are employed by The Mind Company, and The Mind Company's app (Elevate) was the cognitive training app used in this study. JH is a paid consultant of The Mind Company, and KM and LB are employed by JH. To mitigate potential conflicts of interest, data analysis and interpretation were led by authors not employed by The Mind Company. The analysis plan was prespecified, and the results were interpreted collaboratively to ensure objectivity. The conflict of interest has been fully disclosed. Authors' employment status or salary are not dependent upon the results of their research.

Multimedia Appendix 1
Supplemental methods and results.
[DOCX File , 285 KB - formative_v9i1e80027_app1.docx]

Multimedia Appendix 2
STROBE checklist for cross-sectional studies.
[DOCX File , 22 KB - formative_v9i1e80027_app2.docx]

References

1. Hale JM, Schneider DC, Mehta NK, Myrskylä M. Cognitive impairment in the U.S.: lifetime risk, age at onset, and years impaired. SSM Popul Health 2020 Mar 31;11:100577 [FREE Full text] [doi: 10.1016/j.ssmph.2020.100577] [Medline: 32300635]
2. Corbo I, Favieri F, Forte G, Casagrande M. Decision-making under uncertainty in healthy and cognitively impaired aging: a systematic review and meta-analysis. Arch Gerontol Geriatr 2025 Feb;129:105643 [FREE Full text] [doi: 10.1016/j.archger.2024.105643] [Medline: 39369563]
3. Gray M, Gills JL, Glenn JM, Vincenzo JL, Walter CS, Madero EN, et al. Cognitive decline negatively impacts physical function. Exp Gerontol 2021 Jan;143:111164 [FREE Full text] [doi: 10.1016/j.exger.2020.111164] [Medline: 33232795]
4. Dubbelman MA, Jutten RJ, Tomaszewski Farias SE, Amariglio RE, Buckley RF, Visser PJ, et al. Decline in cognitively complex everyday activities accelerates along the Alzheimer's disease continuum. Alzheimers Res Ther 2020 Oct 29;12(1):138 [FREE Full text] [doi: 10.1186/s13195-020-00706-2] [Medline: 33121534]
5. Liew TM. Trajectories of subjective cognitive decline, and the risk of mild cognitive impairment and dementia. Alzheimers Res Ther 2020 Oct 27;12(1):135 [FREE Full text] [doi: 10.1186/s13195-020-00699-y] [Medline: 33109275]
6. Blum K. The impact of chronic stress on brain function and structure. Perspect Neurosci Psychiatry Open Access 2024;7(5):262-264 [FREE Full text] [doi: 10.47532/npoa.2024]
7. Parry DA, Le Roux D. "Cognitive control in media multitaskers" ten years on: a meta-analysis. Cyberpsychol J Psychosoc Res Cybersp 2021 Apr 26;15(2). [doi: 10.5817/CP2021-2-7]
8. Aidman E. Cognitive fitness framework: towards assessing, training and augmenting individual-difference factors underpinning high-performance cognition. Front Hum Neurosci 2020 Jan 14;13:466 [FREE Full text] [doi: 10.3389/fnhum.2019.00466] [Medline: 32009919]
9. Robinson PL. Conceptualising and measuring mental fitness. University of Wollongong. 2014. URL: https://ro.uow.edu.au/articles/thesis/Conceptualising_and_measuring_mental_fitness/27661959?file=50371740 [accessed 2025-07-02]
10. Bhat SG, Nagaraj M, Balentine C, Hogan T, Meier J, Prince H, et al. Assessing a structured mental fitness program for academic acute care surgeons: a pilot study. J Surg Res 2024 Mar;295:9-18 [FREE Full text] [doi: 10.1016/j.jss.2023.09.052] [Medline: 37956507]
11. Iorfino F, Varidel M, Capon W, Richards M, Crouse JJ, LaMonica HM, et al. Quantifying the interrelationships between physical, social, and cognitive-emotional components of mental fitness using digital technology. Npj Ment Health Res 2024 Jul 08;3(1):36 [FREE Full text] [doi: 10.1038/s44184-024-00078-7] [Medline: 38977903]
12. Prince M, Patel V, Saxena S, Maj M, Maselko J, Phillips MR, et al. No health without mental health. Lancet 2007 Sep 08;370(9590):859-877. [doi: 10.1016/S0140-6736(07)61238-0] [Medline: 17804063]
13. Keyes CL. Promoting and protecting mental health as flourishing: a complementary strategy for improving national mental health. Am Psychol 2007;62(2):95-108. [doi: 10.1037/0003-066X.62.2.95] [Medline: 17324035]
14. Kurtz MM. Compensatory strategies. In: Kreutzer JS, DeLuca J, Caplan B, editors. Encyclopedia of Clinical Neuropsychology. New York, NY: Springer; 2011.
15. Moreau D, Wiebels K. A precision-mapping approach to physical exercise interventions targeting cognitive function. Prog Brain Res 2024;283:231-253. [doi: 10.1016/bs.pbr.2023.12.001] [Medline: 38538190]

XSL•FO
RenderX

16. Dhahbi W, Briki W, Heissel A, Schega L, Dergaa I, Guelmami N, et al. Physical activity to counter age-related cognitive decline: benefits of aerobic, resistance, and combined training-a narrative review. Sports Med Open 2025 May 17;11(1):56. [doi: 10.1186/s40798-025-00857-2] [Medline: 40381170]
17. Livingston G, Huntley J, Sommerlad A, Ames D, Ballard C, Banerjee S, et al. Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. Lancet 2020 Aug 08;396(10248):413-446 [FREE Full text] [doi: 10.1016/S0140-6736(20)30367-6] [Medline: 32738937]
18. Mayor S. Mentally stimulating activities reduce risk of mild cognitive impairment, study finds. BMJ 2017 Jan 31;356:j526. [doi: 10.1136/bmj.j526]
19. Yates LA, Ziser S, Spector A, Orrell M. Cognitive leisure activities and future risk of cognitive impairment and dementia: systematic review and meta-analysis. Int Psychogeriatr 2016 Nov;28(11):1791-1806. [doi: 10.1017/s1041610216001137]
20. Tetlow AM, Edwards JD. Systematic literature review and meta-analysis of commercially available computerized cognitive training among older adults. J Cogn Enhanc 2017 Nov 9;1:559-575. [doi: 10.1007/s41465-017-0051-2]
21. Melby-Lervåg M, Redick TS, Hulme C. Working memory training does not improve performance on measures of intelligence or other measures of "far transfer": evidence from a meta-analytic review. Perspect Psychol Sci 2016 Jul;11(4):512-534 [FREE Full text] [doi: 10.1177/1745691616635612] [Medline: 27474138]
22. Nguyen L, Murphy K, Andrews G. A game a day keeps cognitive decline away? A systematic review and meta-analysis of commercially-available brain training programs in healthy and cognitively impaired older adults. Neuropsychol Rev 2022 Sep;32(3):601-630. [doi: 10.1007/s11065-021-09515-2] [Medline: 34251578]
23. Bonnechère B, Langley C, Sahakian BJ. The use of commercial computerised cognitive games in older adults: a meta-analysis. Sci Rep 2020 Sep 17;10(1):15276. [doi: 10.1038/s41598-020-72281-3] [Medline: 32943742]
24. Pedullà L, Brichetto G, Tacchino A, Vassallo C, Zaratin P, Battaglia MA, et al. Adaptive vs. non-adaptive cognitive training by means of a personalized app: a randomized trial in people with multiple sclerosis. J Neuroeng Rehabil 2016 Oct 04;13(1):88 [FREE Full text] [doi: 10.1186/s12984-016-0193-y] [Medline: 27716336]
25. Bonnechère B, Klass M, Langley C, Sahakian BJ. Brain training using cognitive apps can improve cognitive performance and processing speed in older adults. Sci Rep 2021 Jun 10;11(1):12313. [doi: 10.1038/s41598-021-91867-z] [Medline: 34112925]
26. Kirby ED, Beyst B, Beyst J, Brodie SM, D'Arcy RC. A retrospective, observational study of real-world clinical data from the Cognitive Function Development Therapy program. Front Hum Neurosci 2024 Dec 18;18:1508815 [FREE Full text] [doi: 10.3389/fnhum.2024.1508815] [Medline: 39743989]
27. Gary RA, Paul S, Corwin E, Butts B, Miller AH, Hepburn K, et al. Exercise and cognitive training intervention improves self-care, quality of life and functional capacity in persons with heart failure. J Appl Gerontol 2022 Feb;41(2):486-495 [FREE Full text] [doi: 10.1177/0733464820964338] [Medline: 33047625]
28. Hampshire A, Sandrone S, Hellyer PJ. A large-scale, cross-sectional investigation into the efficacy of brain training. Front Hum Neurosci 2019 Jul 9;13:221 [FREE Full text] [doi: 10.3389/fnhum.2019.00221] [Medline: 31338032]
29. Al-Thaqib A, Al-Sultan F, Al-Zahrani A, Al-Kahtani F, Al-Regaiey K, Iqbal M, et al. Brain training games enhance cognitive function in healthy subjects. Med Sci Monit Basic Res 2018 Apr 20;24:63-69 [FREE Full text] [doi: 10.12659/msmbr.909022] [Medline: 29674605]
30. Hagen BI, Landrø NI, Lau B, Koster EH, Stubberud J. Predictors of long-term improvement following cognitive remediation in a sample with elevated depressive symptoms. Front Psychol 2020 Sep 11;11:2232 [FREE Full text] [doi: 10.3389/fpsyg.2020.02232] [Medline: 33013583]
31. Savulich G, Thorp E, Piercy T, Peterson KA, Pickard JD, Sahakian BJ. Improvements in attention following cognitive training with the novel "Decoder" game on an iPad. Front Behav Neurosci 2019 Jan 21;13:2 [FREE Full text] [doi: 10.3389/fnbeh.2019.00002] [Medline: 30719000]
32. Fosnacht K, Sarraf S, Howe E, Peck LK. How important are high response rates for college surveys? Rev High Educ 2017;40(2):245-265. [doi: 10.1353/rhe.2017.0003]
33. Daikeler J, Bošnjak M, Manfreda KL. Web versus other survey modes: an updated and extended meta-analysis comparing response rates. J Surv Stat Method 2020 Jun;8(3):513-539. [doi: 10.1093/jssam/smz008]
34. Wu MJ, Zhao K, Fils-Aime F. Response rates of online surveys in published research: a meta-analysis. Comput Hum Behav Rep 2022 Aug;7:100206. [doi: 10.1016/j.chbr.2022.100206]
35. Christensen RH. ordinal: regression models for ordinal data. The Comprehensive R Archive Network. 2024 Aug 19. URL: https://cran.r-project.org/web/packages/ordinal/index.html [accessed 2025-07-02]
36. Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. BMJ 1995 Jan 21;310(6973):170. [doi: 10.1136/bmj.310.6973.170] [Medline: 7833759]
37. R: the R project for statistical computing. The Comprehensive R Archive Network. URL: https://www.r-project.org/ [accessed 2025-03-26]
38. Hegeman P, Vader D, Kamke K, El-Toukhy S. Patterns of digital health access and use among US adults: a latent class analysis. BMC Digit Health 2024;2(1):42. [doi: 10.1186/s44247-024-00100-0] [Medline: 39544227]

XSL•FO
RenderX

39. Lee J, Lim JM. Factors associated with the experience of cognitive training apps for the prevention of dementia: cross-sectional study using an extended health belief model. J Med Internet Res 2022 Jan 14;24(1):e31664 [FREE Full text] [doi: 10.2196/31664] [Medline: 35029540]
40. Antal BB, van Nieuwenhuizen H, Chesebro AG, Strey HH, Jones DT, Clarke K, et al. Brain aging shows nonlinear transitions, suggesting a midlife "critical window" for metabolic intervention. Proc Natl Acad Sci U S A 2025 Mar 11;122(10):e2416433122 [FREE Full text] [doi: 10.1073/pnas.2416433122] [Medline: 40030017]
41. Gates NJ, Rutjes AW, Di Nisio M, Karim S, Chong LY, March E, et al. Computerised cognitive training for maintaining cognitive function in cognitively healthy people in midlife. Cochrane Database Syst Rev 2019 Mar 13;3(3):CD012278 [FREE Full text] [doi: 10.1002/14651858.CD012278.pub2] [Medline: 30864746]
42. Kokubun K, Nemoto K, Yamakawa Y. Smartphone app for lifestyle improvement improves brain health and boosts the vitality and cognitive function of healthy middle-aged adults. Brain Behav 2024 May;14(5):e3500 [FREE Full text] [doi: 10.1002/brb3.3500] [Medline: 38685801]
43. Li Z, He H, Chen Y, Guan Q. Effects of engagement, persistence and adherence on cognitive training outcomes in older adults with and without cognitive impairment: a systematic review and meta-analysis of randomised controlled trials. Age Ageing 2024 Jan 02;53(1):afad247. [doi: 10.1093/ageing/afad247] [Medline: 38266127]
44. Belleville S, Cuesta M, Bier N, Brodeur C, Gauthier S, Gilbert B, et al. Five-year effects of cognitive training in individuals with mild cognitive impairment. Alzheimers Dement (Amst) 2024 Sep 06;16(3):e12626. [doi: 10.1002/dad2.12626] [Medline: 39246830]
45. Barzagar Nazari K, Ebersbach M. Distributed practice: rarely realized in self-regulated mathematical learning. Front Psychol 2018 Nov 20;9:2170. [doi: 10.3389/fpsyg.2018.02170] [Medline: 30524328]
46. Stine-Morrow EA, McCall GS, Manavbasi I, Ng S, Llano DA, Barbey AK. The effects of sustained literacy engagement on cognition and sentence processing among older adults. Front Psychol 2022 Jul 11;13:923795 [FREE Full text] [doi: 10.3389/fpsyg.2022.923795] [Medline: 35898978]
47. Seufert T. The interplay between self-regulation in learning and cognitive load. Educ Res Rev 2018 Jun;24:116-129. [doi: 10.1016/j.edurev.2018.03.004]
48. Lee LY, Healy MP, Fischer NL, Tong K, Chen AS, Sahakian BJ, et al. Cognitive flexibility training for impact in real-world settings. Curr Opin Behav Sci 2024 Oct;59:101413. [doi: 10.1016/j.cobeha.2024.101413]
49. Oki Y, Osaki T, Kumagai R, Murata S, Encho H, Ono R, et al. An 18-month multimodal intervention trial for preventing dementia: J-MINT PRIME Tamba. Alzheimers Dement 2024 Oct;20(10):6972-6983. [doi: 10.1002/alz.14170] [Medline: 39229900]
50. Pressler SJ, Martineau A, Grossi J, Giordani B, Koelling TM, Ronis DL, et al. Healthcare resource use among heart failure patients in a randomized pilot study of a cognitive training intervention. Heart Lung 2013;42(5):332-338 [FREE Full text] [doi: 10.1016/j.hrtlng.2013.05.001] [Medline: 23809197]
51. Osman AM, Jaffe PI, Ng NF, Kerlan KR, Schafer RJ. Transfer of learning: analysis of dose-response functions from a large-scale, online, cognitive training dataset. PLoS One 2023 May 17;18(5):e0281095 [FREE Full text] [doi: 10.1371/journal.pone.0281095] [Medline: 37195927]
52. Lin Y, Tudor-Sfetea C, Siddiqui S, Sherwani Y, Ahmed M, Eisingerich AB. Effective behavioral changes through a digital mHealth app: exploring the impact of hedonic well-being, psychological empowerment and inspiration. JMIR Mhealth Uhealth 2018 Jun 15;6(6):e10024. [doi: 10.2196/10024] [Medline: 29907557]

Abbreviations

OR: odds ratio
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by A Mavragani, J Sarvestan; submitted 02.07.25; peer-reviewed by A Eisingerich; comments to author 28.07.25; revised version received 01.08.25; accepted 21.09.25; published 10.10.25.

Please cite as:
McAlister K, Baez L, Braunsdorf A, Yang Y, Kessler D, Huberty J
User Engagement With and Perceived Impact of a Digital Cognitive Training App on Cognition, Daily Functioning, and Mental Fitness: Secondary Analysis of Cross-Sectional Survey Data
JMIR Form Res 2025;9:e80027
URL: https://formative.jmir.org/2025/1/e80027
doi:10.2196/80027
PMID:

XSL•FO
RenderX

©Kelsey McAlister, Lara Baez, Anna Braunsdorf, Yana Yang, Dan Kessler, Jennifer Huberty. Originally published in JMIR Formative Research (https://formative.jmir.org), 10.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Explainable AI-Driven Analysis of Radiology Reports Using Text and Image Data: Experimental Study

Muhammad Tayyab Zamir[1*], MPhil; Safir Ullah Khan[2*], PhD; Alexander Gelbukh[1], PhD; Edgardo Manuel Felipe Riverón[1], PhD; Irina Gelbukh[1], PhD

[1]Centro de Investigación en Computación (CIC), Instituto Politécnico Nacional (IPN), Ciudad de México, CDMX, Mexico

[2]Department of Cell Biology, Center for Research and Advanced Studies of the National Polytechnic Institute, Ciudad de México, CDMX, Mexico

[*]these authors contributed equally

Corresponding Author:
Alexander Gelbukh, PhD
Centro de Investigación en Computación (CIC)
Instituto Politécnico Nacional (IPN)
Avenida Juan de Dios Bátiz S/N, Gustavo A. Madero
Nueva Industrial Vallejo
Ciudad de México, CDMX, 07700
Mexico
Phone: 52 5518104587
Email: gelbukh@cic.ipn.mx

Abstract

Background: Artificial intelligence (AI) is increasingly being integrated into clinical diagnostics; yet, its lack of transparency hinders trust and adoption among health care professionals. The explainable artificial intelligence (XAI) has the potential to improve the interpretability and reliability of AI-based decisions in clinical practice.

Objective: This study evaluates the use of XAI for interpreting radiology reports to improve health care practitioners' confidence and comprehension of AI-assisted diagnostics.

Methods: This study used the Indiana University chest x-ray dataset containing 3169 textual reports and 6471 images. Textual data were being classified as either normal or abnormal by using a range of machine learning approaches. This includes traditional machine learning models and ensemble methods, deep learning models (long short-term memory network), and advanced transformer-based language models (GPT-2, T5, LLaMA-2, and LLaMA-3.1). For image-based classifications, convolutional neural networks, including DenseNet121 and DenseNet169, were used. Top-performing models were interpreted using XAI methods SHAP (Shapley Adaptive Explanations) and Local Interpretable Model-Agnostic Explanations to support clinical decision making by enhancing transparency and trust in model predictions.

Results: The LLaMA-3.1 model achieved the highest accuracy of 98% in classifying the textual radiology reports. Statistical analysis confirmed the model's robustness, with Cohen κ (k=0.981) indicating near-perfect agreement beyond chance. Both the chi-square and Fisher exact tests revealed a highly significant association between the actual and predicted labels ($P<.001$). Although the McNemar Test yielding a nonsignificant result ($P=.25$) suggests a balanced class performance, the highest accuracy of 84% was achieved in the analysis of imaging data using the DenseNet169 and DenseNet121 models. To assess explainability, Local Interpretable Model-Agnostic Explanations and SHAP were applied to the best-performing models. These models consistently highlighted that the medical-related terms such as "opacity," "consolidation," and "pleural" are clear indications for abnormal findings in textual reports.

Conclusions: The research underscores that explainability is an essential component of any AI systems used in diagnostics and is helpful in the design and implementation of AI in the health care sector. Such an approach improves the accuracy of the diagnosis and builds confidence in health workers, who in the future will use XAI in clinical settings, particularly in the application of AI explainability for medical purposes.

(JMIR Form Res 2025;9:e77482) doi:10.2196/77482

XSL•FO
RenderX

KEYWORDS

explainable AI; radiology; natural language processing; SHAP; LIME; artificial intelligence; Shapley Adaptive Explanations; Local Interpretable Model-Agnostic Explanations

Introduction

Background

Artificial intelligence (AI) has the potential to significantly improve diagnostic accuracy, efficiency, and reliability in radiology. Traditional radiology relies heavily on the expertise and subjective judgment of individual radiologists [1], which can be inconsistent and limited when processing large volumes of complex data [2]. AI methods can efficiently analyze vast textual and imaging datasets, but their adoption in clinical practice depends on transparency and interpretability, which are critical for building trust among health care practitioners [3]. This requirement has led to the development of explainable artificial intelligence (XAI), which focuses on both predictive performance and interpretability [4].

Radiology reports include textual narratives describing patient conditions and medical images such as x-rays, CT scans, or MRIs [5]. Most prior studies have analyzed either textual reports or images independently, and multimodal approaches remain limited and insufficiently interpretable. In this study, we analyze text and images separately, applying XAI techniques to each modality. This approach allows clinicians to understand the features driving AI predictions in a modality-specific manner, which is critical for accurate and trustworthy decision-making [6-9].

For textual analysis, natural language processing (NLP) methods extract clinically relevant information from radiology reports [5]. XAI applied to text highlights keywords or phrases influencing AI predictions, helping clinicians verify outputs [10-12].

For image analysis, convolutional neural networks (CNNs) identify critical visual features in radiology images [6-8]. XAI techniques such as Local Interpretable Model-Agnostic Explanations (LIME) or Grad-CAM (Gradient-Weighted Class Activation Mapping) applied to images provide interpretable visual explanations, allowing radiologists to assess the validity of AI predictions [13,14].

Analyzing text and images separately offers several advantages. It allows modality-specific insights, supports clinical education, identifies rare or abnormal findings, and provides independent evaluation of textual and imaging contributions to AI predictions. By making AI outputs interpretable for each modality, clinicians can validate AI suggestions against established diagnostic protocols, improving both confidence and clinical workflow integration [12-14].

In this research, we aim to evaluate the role of XAI models in improving the trustworthiness and interpretability of textual report classification and image analysis. We use NLP techniques by using the up-to-date LLaMA-3.1 language model to classify radiology reports into normal and abnormal categories, achieving 98% accuracy. Statistical validation showed strong model reliability, with a Cohen κ of 0.981 indicating perfect agreement, and both chi-square and Fisher exact test revealing a significant association ($P<.001$). For image analysis, we fine-tuned a DenseNet169 model for image analysis and resulting in an accuracy of 84%. By using these models, we have created a strong approach to interpret radiological information, benefiting from both text and image analysis strengths. In this study, the modalities were analyzed separately to allow clear modality-specific interpretability, which provides a robust foundation for future multimodal frameworks. Such frameworks can integrate textual and imaging features to more closely replicate the complementary reasoning process used by clinicians in practice. To enhance interpretability, we applied SHAP (Shapley Adaptive Explanations) and LIME on the best-performing textual and image models' outputs. These XAI methods provide insights into how features (words or image regions) influence the model's decisions, allowing clinical experts to assess whether the AI rationale aligns with clinical reasoning.

The primary objective of this study is to evaluate the role of XAI in enhancing transparency and trust in AI-assisted diagnostics, with a focus on radiology reports and chest x-ray images. Specifically, this study aims to (1) apply the state-of-the-art deep learning as well as transformer models to evaluate both text radiology reports and chest x-ray images separately on the Indiana University dataset, (2) use SHAP and LIME separately on the best performance models with text and image modalities, (3) confirm clinical relevance of the explanations (eg, the existence of such keywords as "opacity" and "consolidation" in text and the identification of pathologic areas on images), and (4) demonstrate explainability to improve clinical trust, interpretability, and the potential acceptance of AI in radiology.

Related Work

The use of AI in radiology has accelerated over the past few years, with AI being able to assist clinicians in making decisions by automatically interpreting images and analyzing the reports. The traditional radiology workflow relies on the experience and opinion of a radiologist; hence, it can lead to variability in diagnostics and even result in a delay in diagnosis [15].

Image Analysis in Radiology

Deep learning, particularly CNNs, has shown remarkable performance in detecting abnormalities in medical images. For example, CheXNeXt applied a DenseNet121-based CNN to identify pneumonia in chest x-rays, achieving accuracy comparable to radiologists [16]. Grad-CAM, an XAI technique for image data, generates heatmaps that highlight regions contributing most to model predictions. It has been widely applied in medical imaging tasks, including tuberculosis, breast cancer, and pneumonia detection, providing clinicians with insights into whether models focus on anatomically and diagnostically relevant areas [17].

XSL•FO
RenderX

Despite these advancements, most XAI methods for imaging focus on local explanations rather than providing a global understanding of model behavior. Moreover, few studies integrate image XAI with textual analysis, and clinical validation remains limited.

Textual Analysis in Radiology

Parallel to image analysis, NLP has advanced the interpretation of radiology reports. Early approaches, such as bag-of-words or TF-IDF (term frequency–inverse document frequency) combined with support vector machines (SVMs) and logistic regression (LR), could not capture contextual relationships [18]. Deep learning models, including recurrent neural networks (RNNs) and bidirectional long short-term memory networks (Bi-LSTMs), improved performance by modeling sequential dependencies in clinical narratives. The advent of transformer-based models, such as Clinical-BERT (Bidirectional Encoder Representations from Transformers) and Bio-BERT, further enhanced text understanding by leveraging pretraining on biomedical corpora [19,20]. The newest large language models (LLMs), including GPT-2, T5, LLaMA-2, and LLaMA-3.1, demonstrate strong generalization across medical domains with minimal fine-tuning, enabling accurate report classification and information extraction [20].

XAI has emerged to address the critical limitation of deep learning models: the lack of interpretability. SHAP assigns feature importance scores using cooperative game theory, while LIME provides local approximations for individual predictions [21,22]. SHAP has been applied to tasks such as sepsis prediction [23], whereas LIME has facilitated clinical validation of NLP models on pathology reports [24]. However, most studies provide explanations for individual predictions only, without offering a global understanding of model behavior.

Limitations and Research Gaps

Along with these advancements, the given approaches face several issues. The majority of XAI provide explanations for individual predictions but fail to provide a global perspective on model behavior. Most of the research focuses on text or images without combining both XAI methods. Finally, there is not much clinical validation of XAI systems yet because many studies focus on technical evaluation instead of how well they work in the real world and how well health care professionals accept them [25].

The existing studies reveal the gap in the lack of systematic application and validation of multiple XAI methods, both to textual and image-based radiology data, and the use of the same dataset, although SHAP and LIME, their comparative performance and compatibility with clinical knowledge are under discussion in an integrated environment. Besides, the results of new transformer models were released recently (such as LLaMA-3.1), when applied to clinical NLP problems, specifically explainability. Additionally, statistical validation is used to check how good a model is, which helps make sure that AI solutions for health care are reliable.

Methods

Study Design

This study followed a structured methodology comprising several steps, beginning with data acquisition and preprocessing, followed by model training for textual and image data separately, and concluding with the XAI technique to interpret the results. This study was conducted in line with the TRIPOD+AI (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis–Artificial Intelligence) guidelines (checklist provided in Multimedia Appendix 1).

Dataset Description

For this study, the dataset was sourced from the Indiana University chest x-ray collection made available in part by the National Library of Medicine's Open-I program [26-29]. The dataset contains both text-based clinical radiology reports and their associated imaging reports. The reports have been coded into normal and abnormal with the assistance of the MeSH (Medical Subject Headings) keyword. At the beginning of this work, the internal dataset consisted of 3955 samples of textual radiology reports. After a thorough data preprocessing, 3169 samples were left, with 2535 used for training and 634 for testing. For the image dataset, 7470 images were acquired in the first stage, but after duplicates and deficits in the sample designs were removed, 6471 samples were retained. These were further subdivided into 5177 images for training and 1294 images for testing. All images belong to a set of x-ray reports and are in PNG graphical format. A sample dataset is shown in Figure 1.

XSL•FO
RenderX

Figure 1. A chest x-ray image and its report with findings and impression.

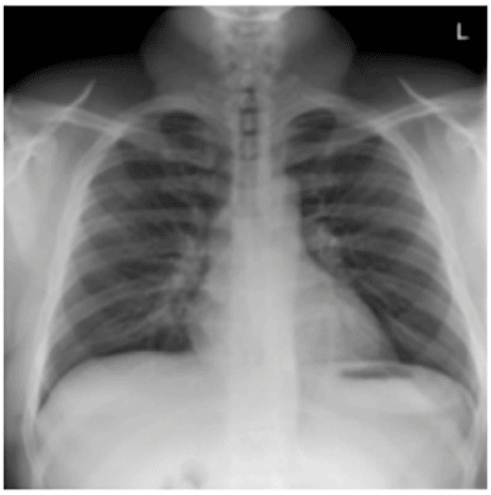

Findings: Cardiac and mediastinal counters are normal within limits. The lungs are clear. Bony structures are intact.

Mesh: Normal

Preprocessing

Preprocessing for Textual Reports

Processing of text data included removal of duplicate reports and then their tokenization, where individual sentences were broken down into smaller tokens that are words, with the help of the NLTK library. The next step performed on the text was to make it in lower case, to remove stop words using the NLTK English stop-word list, and lemmatization using WordNetLemmatizer to make sure that the data was preserved and enhanced for NLP tasks.

Preprocessing for Image Reports

For image data preprocessing, all the images were adjusted to 255×255 pixels by using bilinear interpolation to ensure standardization across the dataset. The pixel values were normalized in the range 0, 1 (binary) to enhance model efficiency, and repeated diagnostic reports were removed.

Textual Reports Classification Models

Several classification methods were used in this task on the textual resource [30], among them it is worth mentioning that SVM classifier is for classes separation, LR is where binary response is examined, naïve bayes is commonly used for classification with probabilities of being in certain class, AdaBoost is for boosting weak classifiers, decision trees is for the processing of decisions, k-nearest neighbor (KNN) is for working with classification by proximity, random forest is when there are large amounts of data, and multilayer perceptron uses neural networks for forecasting. To enhance the performance of machine learning models, we applied ensemble techniques, specifically hard and soft voting classifiers, by aggregating the predictions of multiple machine learning models. In the hard voting approach, each model casts a single vote for the predicted class label, and the final decision is made by majority voting. While soft voting, the classifier outputs class probabilities instead of discrete predictions, and these predictions are averaged, and the class with the highest average probability is selected for the final decision. Four deep learning architectures were put into place, in particular, RNNs aimed at learning inputs sequentially, as well as the Bi-LSTM model based on transformer, incorporating sequential learning and contextual embedding in actual classification of texts [31]. In this study, transformer models were used: BERT, RoBERTa (Robustly Optimized Bidirectional Encoder Representations from Transformers Pretraining Approach), ALBERT-base-v2 (A Lite Bidirectional Encoder Representations from Transformers), and DistilBERT (Distilled Version of Bidirectional Encoder Representations from Transformers) to analyze confidence levels in radiology reports. BERT has been effective in improving the detection of abnormalities by incorporating terminologies and contexts related to the medical field. RoBERTa and ALBERT base version 2 aided in the detection of preabnormalities by pretraining and optimizing power related to computational detection [32]. The availability of DistilBERT further enhanced the efficiency of the clinical decision-making system.

Strengths of various LLMs were also studied, with an emphasis on understanding and rating the degree of abnormalities that could be identified in radiological reports. The models included were XLNet, T5, GPT-2, LLaMA-2, and LLaMA-3.1. The XLNet model improved understanding of language using autoregressive pretraining, along with capturing bidirectional context, while the T5 model treated individual tasks of NLP as problems of converting text into text. GPT-2 was quite rapid in natural language generation, and LLaMA-2 and LLaMA-3.1 further advanced this capacity and trained on high volumes of data, performing language tasks of high complexity. Each model was fine-tuned with tailored hyperparameters as summarized in Table 1 to ensure fairness in evaluation rather than one-size-fits-all approaches.

XSL•FO
RenderX

Table 1. Summary of models, definitions, and hyperparameter configuration.

Model	Definition	Key parameters	Purpose of the study
Logistic regression	Linear model estimating probabilities using a logistic function.	Solver='liblinear', C=1.0, max_iter=1000	Baseline classifier for comparison.
Random forest	Ensemble of decision trees using bagging and feature randomness.	n_estimators=200, max_depth=20, criterion='gini'	Robust baseline to handle nonlinearities.
Decision tree	Tree-based model splitting data by features iteratively.	max_depth=15, min_samples_split=2	Simple interpretable model for benchmarking.
SVM[a]	Finds a hyperplane maximizing the margin between classes.	kernel='rbf', C=2.0, gamma='scale'	Strong baseline for high-dimensional text features.
KNN[b]	Classifies samples by majority vote of nearest neighbors.	n_neighbors=7, metric='minkowski'	Benchmark nonparametric learner.
AdaBoost	Boosting weak learners sequentially to minimize errors.	n_estimators=150, learning_rate=0.5	Boosting baseline for imbalanced classes.
MLP[c]	Feedforward neural network with multiple layers.	hidden_layers=(256,128), activation='relu', optimizer=Adam(lr=0.001), epochs=30	Captures nonlinear text or image relations.
Naïve bayes	Probabilistic model assuming feature independence.	MultinomialNB(alpha=1.0)	Classic text classification baseline.
Soft voting	Combines model probabilities for the final prediction.	Weights tuned based on validation accuracy	Ensemble stability.
Hard voting	Combines model class labels by majority voting.	Equal weight	Ensemble diversity.
CNN[d]	Convolutional layers capture spatial or textual patterns.	filters=128, kernel=3, optimizer=Adam(lr=1e-4), dropout=0.5	For image or text local feature extraction.
LSTM[e]	Sequential model capturing long-term dependencies.	hidden_units=128, dropout=0.3, optimizer=Adam(lr=1e-3), epochs=20	Sequential feature modeling in text.
Bi-LSTM[f]	Processes sequences forward and backward for a richer context.	hidden_units=128, dropout=0.3	Improved sequential modeling.
GRU[g]	Simplified LSTM with fewer parameters.	hidden_units=128, dropout=0.3	Efficient sequential modeling.
RNN[h]	Basic recurrent network for sequential data.	hidden_units=64, optimizer=Adam(lr=1e-3)	Baseline sequential learning.
BERT[i]	Transformer model pretrained on bidirectional context.	batch_size=16, lr=2e-5, epochs=3	Contextual embeddings for text.
RoBERTa[j]	Robustly optimized BERT variant.	lr=2e-5, batch_size=16, epochs=3	Better optimization for text classification.
DistilBERT[k]	Lightweight distilled version of BERT.	lr=3e-5, batch_size=32, epochs=4	Efficient transformer baseline.
ALBERT[l]	Parameter-sharing optimized BERT variant.	lr=2e-5, epochs=3	Memory-efficient transformer.
XLNet	Autoregressive transformer capturing bidirectional context.	lr=2e-5, batch_size=8, epochs=3	Strong baseline for sequential tasks.
GPT-2	Transformer-based generative language model.	lr=1e-5, epochs=3, adapter-tuning	Used for generative evaluation of reports.
T5	Text-to-text transformer for multiple NLP[m] tasks.	lr=3e-5, epochs=3, batch_size=16	For sequence-to-sequence modeling.
LLaMA-2	Lightweight LLM[n] optimized for research tasks.	LoRA/QLoRA fine-tuning, rank=8, α=16, lr=2e-4	Efficient fine-tuning on radiology text.
LLaMA-3.1	Latest LLaMA with enhanced reasoning.	QLoRA fine-tuning, rank=8, 4-bit quantization, lr=2e-4	Advanced LLM for domain text analysis.

[a]SVM: support vector machine.

[b]KNN: k-nearest neighbor.

[c]MLP: multilayer perceptron.

XSL•FO
RenderX

[d]CNN: convolutional neural network.
[e]LSTM: long short-term memory network.
[f]Bi-LSTM: bidirectional long short-term memory network.
[g]GRU: gated recurrent unit.
[h]RNN: recurrent neural network.
[i]BERT: Bidirectional Encoder Representations from Transformers.
[j]RoBERTa: Robustly Optimized Bidirectional Encoder Representations from Transformers Pretraining Approach.
[k]DistilBERT: Distilled Version of Bidirectional Encoder Representations from Transformers.
[l]ALBERT: A Lite Bidirectional Encoder Representations from Transformers.
[m]NLP: natural language processing.
[n]LLM: large language model.

Images Reports Classification Models

Three traditional classification algorithms were used on the analyzed image data; SVM was used for optimal class discrimination, and LR was applied for class probabilities estimation, with random forest assisting in 50,000 structures, limiting the over-fitting problem [26]. For deep learning, several models were fine-tuned [27], including ResNet50, ResNet152, EfficientNet B7, VGG16, DenseNet121, DenseNet169, and used their architecture for sharper feature extraction. In order to enhance the interpretability of model outputs, techniques of XAI LIME and SHAP were used [28] on the best model results for textual and image reports. With the help of such models, the reasons behind the AI classification were made clear to researchers and medical professionals, so that AI decisions would not contradict their clinical findings. These explanations were generated from input text and model output probabilities, helping to determine which specific terms triggered normal and abnormal findings. For image data, we use LIME on the pixel-level input features to generate heat maps highlighting the most significant region influencing the DenseNet169 model decisions. This visualization allowed us to verify whether the model focuses on clinically relevant areas of chest x-rays. It was shown that by incorporating image analysis in addition to text analysis, a powerful method was developed that takes the advantages of both modalities. Models developed within the unified framework significantly outperformed their individual counterparts in terms of diagnosis accuracy and interpretability on a diverse dataset of radiology reports. Such an approach demonstrates how XAI can bring the operational aspects of sophisticated AI models into a clinic and improve diagnosis and patient outcomes. Table 2 details the optimization strategies and hyperparameter configurations for image classification data to ensure fair and robust performance comparisons.

Table 2. Fine-tuning and hyperparameter settings for image models.

Model	Definition	Key hyperparameters
Logistic regression	Linear classifier predicting probabilities from weighted features.	C=0.1-10, Solver=lbfgs, Max iter=100
SVM[a]	Finds optimal margin separating classes in feature space.	C=0.1 Kernel=linear, Gamma=scale
Random forest	An ensemble of decision trees aggregating predictions.	Trees=100, Max depth=50, Min split=2
CNN[b]	Learns hierarchical spatial patterns from images.	Input=255×255, Batch=32, LR=2e-5, Epochs=10
ResNet50/152	Deep residual network with skip connections.	Pretrained, LR=2e-5, Batch=16, Epochs=3, Fine-tune last 3 layers
EfficientNet B7	Scales network depth, width, and resolution efficiently.	Pretrained, LR=2e-5, Batch=16, Epochs=30
VGG16	Sequential convolutional network with uniform architecture.	Pretrained, LR=2e-5, Batch=32, Epochs=30, Dropout=0.5
DenseNet121/169	Densely connected layers promote feature reuse.	Pretrained, LR=2e-5, Batch=16, Epochs=30, Dropout=0.5, Fine-tune last dense block

[a]SVM: support vector machine.
[b]CNN: convolutional neural network.

XAI Methods

XAI methods SHAP and LIME were applied to textual and image reports for the best-performing models to ensure transparency. This study did not yet implement a joint multimodal model; instead, each modality was analyzed independently to provide a foundation for future integration.

About LIME

To explain individual predictions, LIME uses an interpretable model that estimates the model locally.

LIME gives weight to particular words that have an impact on categorization for text. For images, it emphasizes the super-pixels that have a high impact on the prediction.

XSL•FO
RenderX

About SHAP

SHAP uses Shapley values of cooperative game theory to explain how each feature affects the outcome of any model. In the case of text classification, for example, normal and abnormal radiology reports, SHAP can be used to display how individual words can contribute to the estimated class. Positive values of SHAP increase the probability of the predicted class, whereas negative values move the prediction toward the alternative one. In the case of image-based models, SHAP can analyze the importance of regions of the image by assessing the variation in the probability of prediction when a region is masked.

Evaluation Metrics

The performance evaluation metrics for the classifiers for this work are evaluated on the basis of precision, recall, F_1-score, and accuracy.

Accuracy is the proportion of all correct predictions (true positives and true negatives) out of the total predictions.

$$\text{Accuracy}=(TP + TN)/ (TP + TN + FP + FN)$$

Where TP are the number of true positive classes, TN are the number of true negative classes, FP are the number of false positive classes, and FN are the number of false negative classes.

Recall is defined as the proportion of true positive predictions out of all actual positives.

$$\text{Recall}=TP/ (TP + FN)$$

Precision is the proportion of true positive predictions out of all positive predictions made.

$$\text{Precision}=TP/ (TP + FP)$$

F_1-score is the harmonic mean of precision and recall, balancing the two.

Statistical Tests

Different statistical tests were performed on the confusion matrices for best best-performing models for images and text data to validate the significant performance of the models.

Cohen κ test is used to measure the interrater agreement between the model predictions and ground truth, which can be computed as:

$$\text{Kappa}=(\text{observed agreement} - \text{expected agreement}) / (1 - \text{expected agreement}).$$

To calculate the significant relationship between actual labels and predicted labels, we used the chi-square test.

$$\text{Chi-square}=\text{sum of } [(\text{observed} - \text{expected})^2/ \text{expected}].$$

The Fisher exact test is an alternative to the chi-square test. It is used when the samples are smaller in size, and it calculates the probability of being as extreme as you can by considering the null hypothesis as true.

The McNemar test measures the performance of 2 models on the same dataset with the formula:

$(|b - c| - 1)^2 / (b + c)$, where b and c are counts for disagreements.

To statistically validate the performance of the best-performing models, multiple tests were applied to their confusion matrices. Cohen κ was used to measure agreement between the model predictions and the ground truth, accounting for agreement that could occur by chance, which is particularly important in datasets with class imbalance. Chi-square and Fisher exact tests assessed whether there was a significant association between predicted and actual labels, with Fisher Exact applied in cases of small sample sizes to ensure validity when some confusion matrix cells had low counts. The McNemar test was used to compare paired model predictions on the same dataset, focusing on the counts of disagreements to determine whether one model significantly outperformed another. These tests served as complementary validation to accuracy and F_1-scores, confirming consistency and ruling out random agreement. Their interpretive value is supportive rather than providing new insights.

Ethical Considerations

This study did not involve direct interaction with human participants. We used the public dataset Indiana University Chest x-ray dataset, which consisted of radiographic images and associated reports. The dataset documentation states that all of the data was completely anonymized before being made public, and additionally does not require any more ethical approval. Consequently, informed consent, compensation, and institutional review board approval did not apply to this study.

Results

Results for textual data and image data were calculated separately by applying different models.

Models Results for Textual Reports

The analysis of text-based radiology machine learning classifiers and ensemble methods revealed significant performance differences across the classifiers. In this research, we evaluate the various machine learning, deep learning, transformers, and LLMs for textual radiology reports performance as shown in Table 3. The performance evaluation metrics for the classifiers for this work were evaluated based on precision, recall, F_1-score, and accuracy.

XSL•FO
RenderX

Table 3. Classification algorithms' performance metrics for textual data.

Models	Precision (%)		Recall (%)		F_1-score (%)		Accuracy (%)
	Macro	Weighted	Macro	Weighted	Macro	Weighted	
Logistic regression	95	94	88	94	91	94	94
Random forest	93	93	85	92	88	92	92
Decision tree	83	88	83	88	83	88	88
SVM[a]	94	94	89	94	91	94	94
KNN[b]	75	86	83	81	77	82	81
AdaBoost	89	91	83	91	86	90	91
MLP[c]	91	93	90	93	91	93	93
Naïve bayes	93	92	85	92	88	92	92
Soft voting	96	95	91	95	93	94	95
Hard voting	96	95	89	95	92	94	95
CNN[d]	93	95	92	95	93	94	92
LSTM[e]	95	95	91	95	94	95	95
Bi-LSTM[f]	96	96	92	96	96	96	96
GRU[g]	94	95	91	95	92	96	95
RNN[h]	92	94	92	94	92	94	94
BERT[i]	92	95	94	95	93	95	95
RoBERTa[j]	96	96	96	96	96	96	96
DistilBERT[k]	93	96	94	95	93	95	95
ALBERT[l]-base	96	96	96	96	96	96	96
XLNET	94	96	95	96	95	96	96
GPT-2	93	96	94	96	94	96	96
T5	95	96	95	97	96	97	97
LLaMA-2	98	98	98	98	98	98	98
LLaMA-3.1	98	99	98	99	98	99	98

[a]SVM: support vector machine.
[b]KNN: k-nearest neighbor.
[c]MLP: multilayer perceptron.
[d]CNN: convolutional neural network.
[e]LSTM: long short-term memory network.
[f]Bi-LSTM: bidirectional long short-term memory network.
[g]GRU: gated recurrent unit.
[h]RNN: recurrent neural network.
[i]BERT: Bidirectional Encoder Representations from Transformers.
[j]RoBERTa: Robustly Optimized Bidirectional Encoder Representations from Transformers Pretraining Approach.
[k]DistilBERT: Distilled Version of Bidirectional Encoder Representations from Transformers.
[l]ALBERT: A Lite Bidirectional Encoder Representations from Transformers.

Soft voting and hard voting classifiers reported the highest precision of 96% and accuracy of 95%, thus the most reliable approaches within the current study scope. LR and SVM were also quite efficient, attaining scores of 94% for both. However, KNN performed the least with an accuracy of 81%. As a rule, the combined classifier approaches, in comparison to the singular classifiers, were more effective, giving improved precision, recall, and F_1-scores. From the deep learning perspective, the Bi-LSTM model was ranked first among the models with a macro precision of 96%, macro recall of 92%, macro F_1-score of 96%, and an accuracy of 96%. However, among the deep learning algorithms, the RNN model was the

XSL•FO
RenderX

least performing. For transformer-based models, RoBERTa and ALBERT-base achieved the best performance measures with macro precision, macro recall, macro F_1-score, and accuracy all at 96%. DistilBERT worked well but not quite as successfully, with a macro precision score of 93% and accuracy of 95%. Table 3 shows that LLaMA-3.1 performed the best with macro-precision, macro-recall, and macro F_1-score of 98%, weighted F_1-score 99% in addition to 98% accuracy. Our results outperform those reported by Wiggins et al [29], demonstrating improved accuracy and overall performance because LLaMA-3 has a larger capacity and optimized transformer architecture, which capture complex patterns and long-range dependencies. Its pretraining on a massive, diverse corpus, combined with domain-specific fine-tuning and careful hyperparameter optimization, enhances understanding of clinical terminology and report structure. Additionally, LLaMA-3 efficiently handles longer sequences, resulting in higher accuracy and stronger agreement with ground truth.

The LLaMA-2 followed with an accuracy of 98%, macro F_1-score of 98%, and weighted F_1-score 98%. T5 came next with 97% accuracy, while the performance of both XLNET and GPT-2 was on par with 96% accuracy. The above results suggest that the new approach goes a notch higher in performance than previous studies, making remarkable improvements in the accuracy and overall performance of the system.

Models' Results for Image Reports

To classify the image data for radiology reports, we applied machine learning and deep learning algorithms, and the results are shown in Table 4.

Table 4. Classification algorithms' performance metrics for radiology image data.

Models	Precision (%)		Recall (%)		F_1-score (%)		Accuracy (%)
	Macro	Weighted	Macro	Weighted	Macro	Weighted	
Logistic regression	66	67	65	66	65	66	66
SVM[a]	63	65	64	65	64	65	65
Random forest	55	59	54	59	56	59	59
CNN[b]	60	64	62	64	63	64	64
ResNet50	61	63	61	64	62	64	64
ResNet152	71	72	70	72	71	72	72
EfficientNet B7	61	64	61	62	63	64	64
VGG16	80	81	80	80	80	81	81
DenseNet121	81	84	82	83	82	84	84
DenseNet169	81	84	82	83	82	84	84

[a]SVM: support vector machine.

[b]CNN: convolutional neural network.

The machine learning classifiers for image data showed in Table 2 that LR outperformed SVM and random forest models; LR with an accuracy of 66%, macro precision of 66%, macro recall of 65%, and macro F_1-score of 65%. Among the deep learning classifiers, EfficientNet B7 achieved a notable accuracy of 72%, macro precision of 71%, macro recall of 70%, and macro F_1-score of 71% and obtained more results than ResNet50 and ResNet152. The highest performance was seen with DenseNet169 and DenseNet121, which reached an accuracy of 84%, macro precision of 81%, macro recall of 82%, and macro F_1-score of 82%, highlighting their superior capability in handling complex image data. In our study, the performance benchmarks for image analysis have been significantly improved compared to previous research, which reported an accuracy of 79.16% [27]. We fine-tuned the DenseNet169 on our specific radiology image dataset, allowing the model to adapt its feature representations to the unique characteristics of the data. Key hyperparameters, such as learning rate, batch size, number of frozen layers, and optimizer settings, were systematically tuned. We used some of the same models as the prior study but incorporated additional techniques such as random crops, flips, and color adjustments to enhance model generalization. These techniques have led to a substantial improvement in our results.

Statistical Analysis for Textural Reports

The classification performance of the top model LLaMA-3.1 was evaluated using a confusion matrix, as shown in Figure 2. The confusion matrix offers a comprehensive perspective of the model's predictions, particularly its ability to classify 2 target classes, normal and abnormal. Of 634 cases, 500 were classified as normal (true positive) and 114 as abnormal (true negative); only 20 samples were classified as missed. The overall metrics for models are calculated as shown in Figure 2.

XSL•FO
RenderX

Figure 2. Shown is the best confusion matrix for the best-performing model LLaMA-3.1 for textural radiology reports.

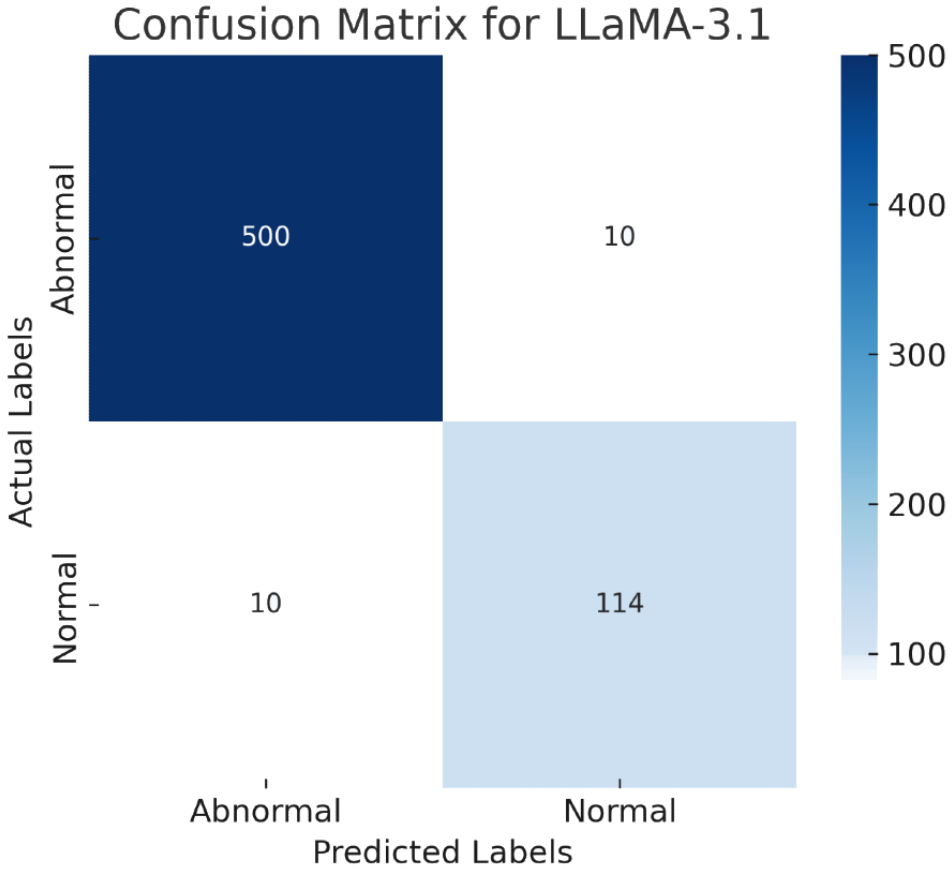

To evaluate whether the model errors are symmetrically distributed between false positives and false negatives, different statistical tests were applied as shown in Table 5, which include the McNemar test, chi-square test, Cohen κ test, and Fisher exact test. These tests provide complementary validation of the confusion matrix results. Cohen κ confirmed strong agreement beyond chance, while chi-square and Fisher exact indicated significant associations between predictions and ground truth. The McNemar test showed balanced errors between classes. While these results are consistent with the accuracy and F_1-scores, they do not provide independent evidence of clinical utility but rather confirm the robustness of the model's performance on this dataset.

Table 5. Statistical significance tests for top top-performing model LLaMA-3.1.

Test	Formula	Value	*P* value	Interpretation
McNemar test	$\chi^2 = ((\lvert b - c\rvert - 1)^2) / (b + c)$	__[a]	.62	0.25
Chi-square test	$\chi^2 = \Sigma((O - E)^2 / E)$	Very high (≫1000)	<.001	Strong dependence between predicted and actual classes
Cohen κ test	$\kappa = (Po - Pe) / (1 - Pe)$	$\kappa = 0.981$	—	Perfect agreement beyond chance
Fisher exact test	Based on hypergeometric probability	—	<.001	Significant association between actual and predicted

[a]Not available.

Table 5 provides a set of statistical tests used to evaluate the performance of the LLaMA-3.1 model in classifying normal and abnormal instances. The McNemar test, with a *P* value of .25, shows that there may be no statistically huge difference among the types of classification errors made by the model. The chi-square test yielded very high test statistics with a *P* value less than .001, strongly indicating a significant association between actual and predicted labels. In other words, the model predictions are highly aligned with ground truth labels.

Additionally, the Cohen κ score from Table 3, 0.981, demonstrated near-perfect agreement between the model prediction and actual labels. This reflects a very high level of consistency and reliability in the model classification performance. Finally, Fisher exact test, which is used in full for small sample sizes, also yielded a significant *P* value (<.001), further confirming the model's predictive performance is statistically significant. Overall, all 4 tests collectively suggest that LLaMA-3.1 is a highly accurate and dependable model for classifying normal and abnormal case reports.

XSL•FO
RenderX

Error Analysis for Image Reports

As illustrated in Figure 3, the DensetNet169 confusion matrix shows that the model correctly classifies 788 abnormal and 431 normal cases. However, it misclassifies 125 normal cases as abnormal (false positives) and 150 abnormal cases as normal (false negatives).

Figure 3. Shown is the best confusion matrix for the best-performing model, DenseNet169, for images of radiology reports.

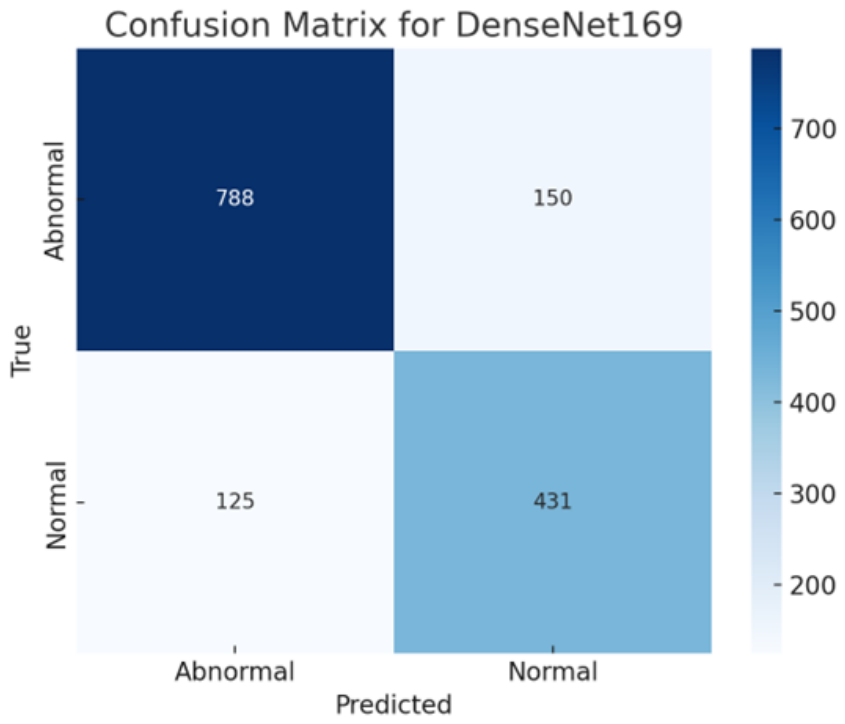

These results highlight the model's effectiveness in identifying abnormal instances with high precision.

Results for XAI

To enhance the interpretability of our AI models used for classifying both textual radiology reports and image data, we used XAI techniques LIME and SHAP on the best results models for textual and image test data. These techniques provide detailed justifications of the parts or regions that influenced the model's decisions, making the AI's classifications more transparent and easier to understand for researchers and medical experts**.**

In Figure 4, an interpretation analysis of LIME for a radiology sample report is given. The LIME model places a high probability of 0.76 that the report would be abnormal and 0.24 as to whether it would be normal. The result of this classification is critical in the clinic, where a wrong diagnosis may result in inappropriate treatment or a failure to diagnose a health issue. In blue, words such as spine, thoracic, chest, and degenerative have been included as the most influential features toward the abnormal prediction of the model. These terms are usually associated with structural or pathological observations, which substantiate the choice of the model and agree with the common radiological signs of abnormal results. Conversely, words such as "normal" and "no," in orange color, are related to the normal class. These words can indicate the lack of findings, which justifies the lower probability assigned to the normal class. The highlighted text at the bottom of the figure further refines how these words are distributed throughout the radiology report, offering a transparent view of how a model intercepts the context. This is essential for clinical explainability, allowing radiologists to verify whether the model predictions are grounded in medically relevant features.

XSL•FO
RenderX

Figure 4. LIME results for abnormal radiology report. LIME: Local Interpretable Model-Agnostic Explanations.

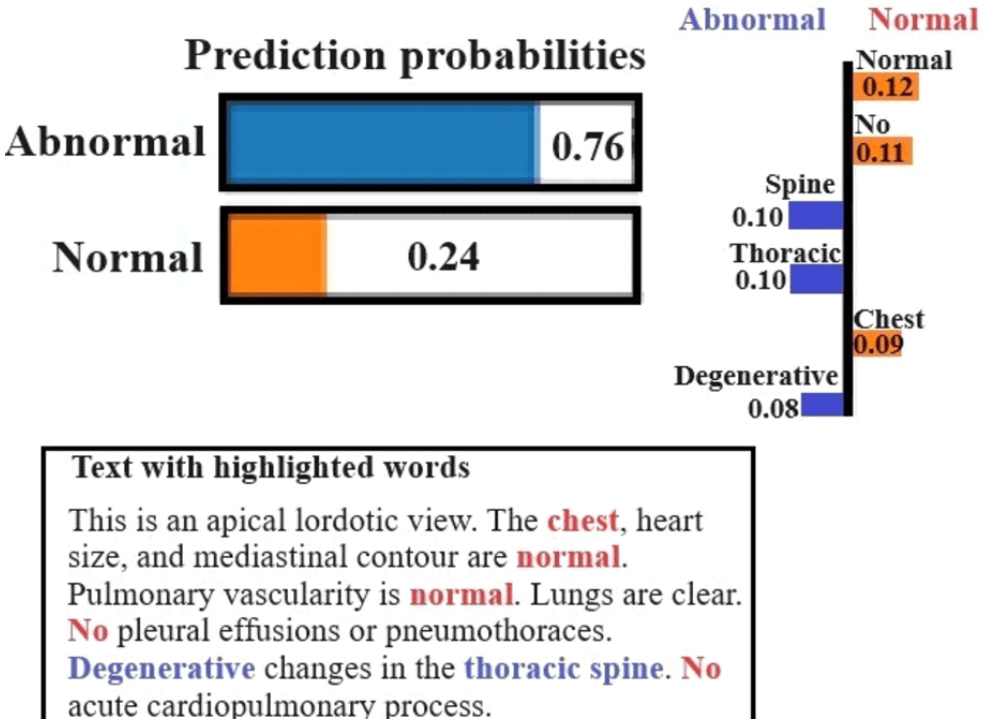

Figure 5 illustrates a radiology report interpretation where LIME assigns a high probability of 0.96 to the normal class, with only 0.04 probability for the abnormal class. This strongly indicates the model's confidence in classifying this report as a normal radiology finding. Words such as "normal," "no," and "acute" highlighted in orange significantly contributed to the normal classification. These terms repeatedly appeared in the report and reflect the absence of any abnormal findings. These are the typical language cues that radiologists use to confirm the normal case, and the model successfully aligns its prediction with them. Terms such as "tissues" are highlighted in blue, suggesting a minor contribution to the abnormal findings. The lower part of Figure 5 provides a visualization of how these words are distributed across the textual report. This helps interpret the reasoning behind the model decision, showing the model appropriate weight on reassuring normal findings while discounting potentially ambiguous terms. This level of interpretability is crucial in the clinical domain, as it allows health care professionals to audit the model's reasoning and ensure alignment with medical judgments. Figure 5 not only confirms the model accuracy but also validates its decision-making process in handling routine, noncritical radiology reports.

XSL•FO
RenderX

Figure 5. LIME results for a normal radiology report. LIME: Local Interpretable Model-Agnostic Explanations.

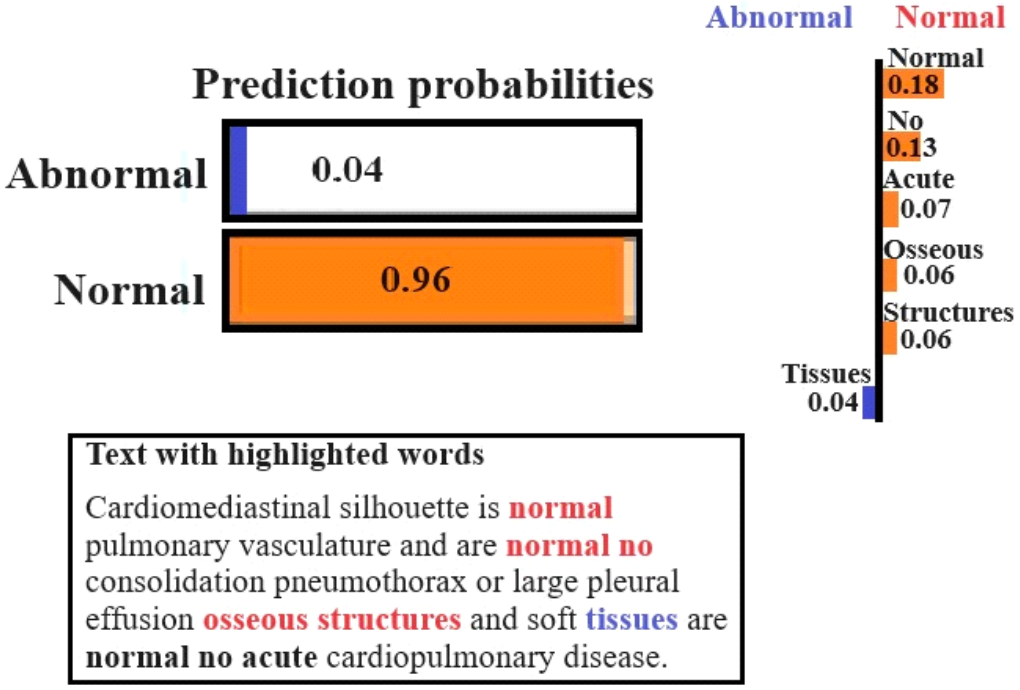

Figure 6 illustrates the contribution of different input features to the predictions of a model regarding normal versus abnormal medical reports. These plots serve as a powerful, interpretable tool, offering a fine-grained look at the model reasoning process for specific radiology reports. In Figure 6A, the SHAP plot shows that overall prediction is skewed toward the abnormal finding, with the base value of –0.888 indicating the model's confidence in this outcome. Keywords such as "pneumothorax" and "effusion" are highlighted in red, which indicates positively toward the abnormal classification. Meanwhile, the terms "lungs" and "heart" appear in blue, opposing the abnormal prediction. Despite the presence of some reassuring words, the presence of abnormality-related terms dominates, driving the model's final decision, and also the base value of prediction starts at around –0.9, indicating a strong likelihood of an abnormal report. In Figure 6B, the base value starts near 0, showing a more balanced prediction. SHAP base value starts at around –0.046, and the prediction is very close to 0. Features such as "lungs," "bilaterally cardiac," "vasculature," and "cardiopulmonary abnormality" also push the prediction for the normal, which are highlighted in blue. In contrast, terms such as "bony" and "abnormality" contribute toward the label abnormal, but not strongly enough to influence the outcome significantly.

XSL•FO
RenderX

Figure 6. SHAP results for (A) normal and (B) abnormal radiology reports. SHAP: Shapley Adaptive Explanations.

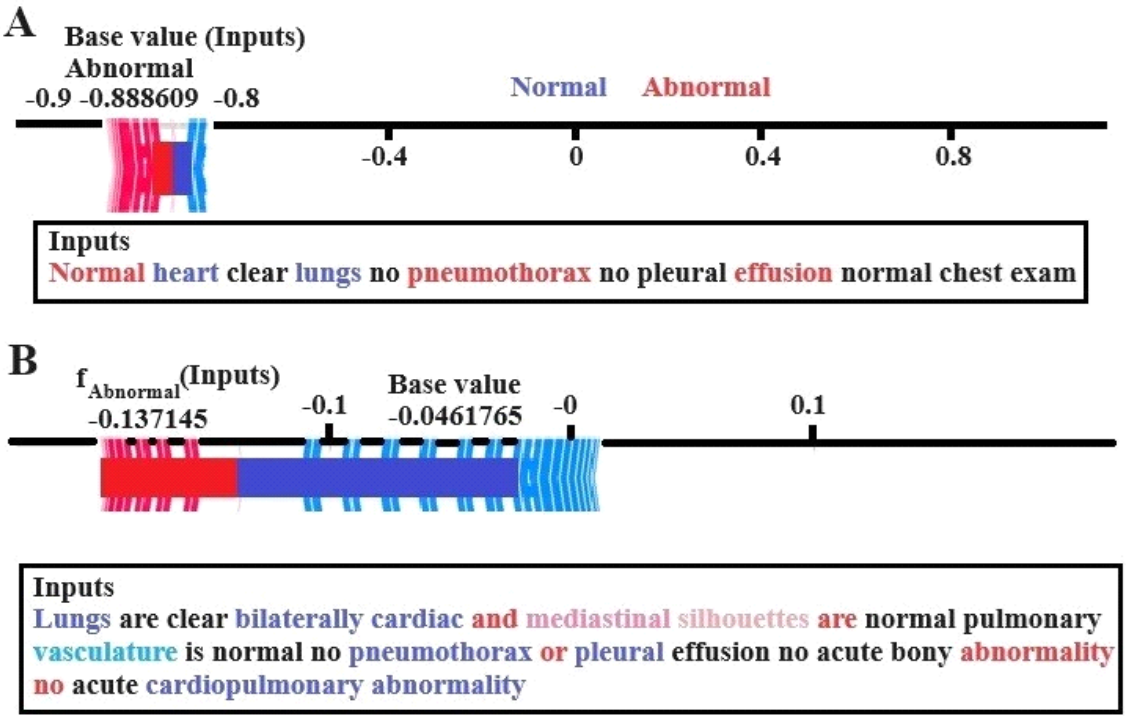

Figure 7 presents the LIME analysis, visualizing regions in medical images contributing to the prediction of normal and abnormal reports. The left side displays a normal image, where red and green areas represent parts of the image that the model considered significant for identifying a normal image. The red area may seem important, but the model correctly identifies its normal image. On the right-hand side, the abnormal image highlights regions contributing to the abnormal prediction, with more red areas indicating potential issues. The yellow outlines show the border of these influential areas. LIME visualizes these explanations by segmenting the image and showing how much impact the model decision has, providing insight into why the model made its prediction for normal and abnormal findings.

Figure 7. LIME explanation for the normal and abnormal image reports dataset. LIME: Local Interpretable Model-Agnostic Explanations.

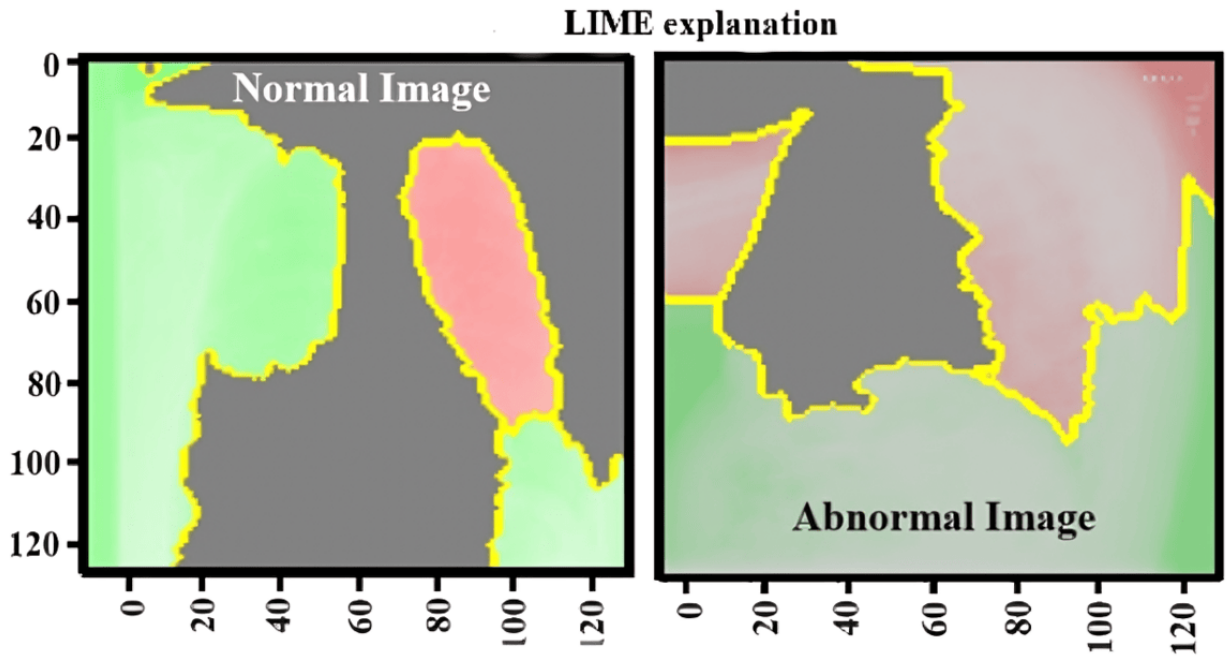

XSL•FO
RenderX

Enhancing Clinical Utility With XAI

SHAP and LIME were used to interpret model outputs; they serve other purposes, too. In clinical practice, radiologists often face ambiguous cases where features are overlapping, such as early signs of pneumonia and mild pulmonary edema. With SHAP, global and local feature importance gives radiologists the ability to view text or images as phrases, "increased opacity," and "no acute findings" that triggered the AI's decision. This indicates whether the model is leveraging medically applicable regions as opposed to irrelevant, misleading ones. SHAP, for example, helped confirm that model bias due to excess reliance on template phrases was not classified too heavily using them. LIME, on the other hand, gave out local explanations that were more intuitive by showing the words or pixels that were decisive for the model's decision, aiding in false positive or negative detection. The information allows radiologists to trust or contest AI's outputs, reducing uncertainty in diagnostics and increasing confidence in informed decisions, especially in nuanced situations where clinical discretion is pivotal.

Complementary Roles of SHAP and LIME

To enhance the transparency of model prediction, we used each SHAP and LIME, which serve complementary roles in explaining AI version behavior, while each methods' intention was to interpret complicated machine learning outputs.

SHAP is based on the game theory idea and provides both local and global explanations by calculating the contribution of each feature to the final prediction using Shapley values. It guarantees consistency and a clear mathematical basis, which makes it useful for knowing how a model behaves throughout the whole dataset and identifying the maximum influential functions as usual. LIME, in comparison, specializes in local interpretability by constructing an interpretable linear model around a single prediction to explain the results. It is especially intuitive and beneficial for explaining individual instances, specifically misclassifications or borderline selections, showing which functions contributed at maximum to that prediction.

Using both models allowed us to provide a comprehensive explanation of the model's decision-making process. SHAP summarized feature importance and overall model reliability, while LIME provided explanations for individual clinical predictions in a human-readable format. This dual approach enhances clarity and transparency, particularly for readers or clinicians who may not be familiar with the inner workings of the algorithm. Moreover, while recognizing prevailing trends, we acknowledge their validity in the context of real-world diagnostic applications.

Discussion

Principal Findings

This study benchmarked multiple machine learning, deep learning, and transformer models for radiology report and chest x-ray classification, with interpretable models. For textual radiology data, ensemble methods, particularly soft and hard voting classifiers [33], outperformed individual machine learning algorithms [34]. Clinically, this suggests that combining multiple models can produce more reliable predictions, reducing the likelihood of errors from a single classifier [35]. These results are consistent with previous research demonstrating the benefits of ensemble strategies for improving the robustness and performance of recognition systems [36].

For deep learning models, Bi-LSTM [37] efficiently captures sequential dependencies in radiology reports, allowing better contextual understanding of clinical narratives. Transformer-based models such as RoBERTa and ALBERT-base [38] also showed improved performance with domain-specific fine-tuning [39], highlighting that incorporating radiology-specific knowledge enhances both accuracy and interpretability. These models could help clinicians by efficiently summarizing complex report content and identifying subtle patterns, thereby reducing cognitive load and potential human error.

Among the evaluated models, LLaMA-3.1 showed the strongest overall performance. Its value lies not only in accuracy but in producing clinically relevant and interpretable predictions when paired with SHAP and LIME [40,41]. These models have the potential to assist in routine diagnostic tasks where high precision and interpretability are crucial, increasing clinicians' trust in AI-assisted decision-making. This emphasizes the need to adapt AI tools for specialized medical applications, ensuring their safety and effectiveness in clinical practice [42]. The future of AI in radiology also promises faster disease recognition, more personalized treatment recommendations, and improved patient outcomes [43].

Regarding imaging data, while traditional machine learning classifiers such as SVM [44] performed reasonably well, they could not match the performance of specialized CNN architectures such as EfficientNet B7 [45] and DenseNet169 [46]. DenseNet169 demonstrates the advantage of deep learning in extracting hierarchical and nuanced features from complex medical images, improving diagnostic accuracy.

SHAP and LIME [47] generated explanations that highlighted plausible clinical terms and regions. However, their relevance to practicing radiologists remains untested. Clinical trust cannot be assumed, and structured evaluations with radiologists are required to confirm whether these explanations enhance diagnostic confidence, efficiency, and decision-making.

Although our results demonstrate strong technical performance and interpretable output, their impact on real-world radiology workflows remains untested. In practice, clinical adoption depends not only on accuracy but also on whether AI explanations can integrate seamlessly into existing reporting systems, save time, or improve decision-making under time constraints. These aspects were beyond the scope of this study but are critical for clinical translation.

Limitations

We acknowledge several limitations in this work. First, the Indiana dataset used for training and evaluation is relatively small, dated, and lacks diversity in terms of patient demographics, imaging modalities, and reporting styles. These characteristics may bias both predictive performance and explainability outcomes. For example, the textual reports often follow templated or simplified structures, which could

XSL•FO
RenderX

artificially boost model accuracy while reducing robustness to more varied reporting styles. Similarly, the imaging data may not reflect the diversity of acquisition protocols or disease prevalence seen in modern radiology practice, limiting the transferability of both predictions and explanations. Consequently, while the dataset serves as a useful benchmark, its constraints reduce the applicability of our findings to contemporary, real-world clinical environments. Furthermore, the analysis relied entirely on a single dataset without cross-institutional or external validation, which limits the generalizability of the findings. Another important limitation is that text and images were analyzed separately rather than through a multimodal framework, which does not fully reflect the integrated nature of real diagnostic workflows. Additionally, although SHAP and LIME produced interpretable outputs, their clinical utility remains uncertain because they were not systematically validated with radiologists. In this study, we assumed that highlighted terms and image regions aligned with clinical reasoning, but we did not test whether clinicians found these explanations trustworthy, intuitive, or useful in practice. Without structured user validation, the degree to which XAI outputs improve diagnostic confidence or decision-making remains speculative. The inclusion of a large number of models provided broad benchmarking but lacked a sharply defined hypothesis, weakening the methodological focus. Another limitation is that our explainability analysis relied solely on SHAP and LIME. While these methods are widely used and model-agnostic, they primarily generate local explanations and are known to have limitations such as instability in feature attributions or sensitivity to parameter choices. The absence of complementary global or clinically validated approaches, such as Grad-CAM for imaging data or counterfactual explanations for text, may restrict the interpretive depth and clinical relevance of our findings. Finally, the statistical tests applied (Cohen κ, chi-square, Fisher exact, and McNemar) primarily confirmed results already reflected in accuracy and F_1-score metrics. While useful for robustness checks, they add limited new interpretive value, and their role in demonstrating clinical reliability should not be overstated.

Future Work

Future research will focus on addressing these limitations to strengthen the clinical value of XAI in radiology. Specifically, we aim to develop multimodal models that integrate textual reports with imaging data to better capture complementary diagnostic information and reflect real-world workflows. We will extend validation across external datasets from different institutions and explore approaches such as domain adaptation and federated learning to enhance cross-site robustness.

To confirm the practical value of AI explanations, we will conduct structured evaluations with radiologists through reader studies, workflow simulations, and integration into reporting systems. These studies will test whether XAI outputs reduce diagnostic uncertainty, improve decision-making compared to existing practices, and fit within the time constraints of clinical workflows. Additionally, usability testing will ensure that explanations are not only technically interpretable but also practical, efficient, and supportive of radiologists' routine tasks. As a part of future work, we will apply Grad-CAM to image models to generate global visual explanations that align with radiological regions of interest, integrated gradients to provide robust attribute scores, and counterfactual explanation techniques for textual reports to illustrate how minimal changes in phrasing alter predictions. We will test our models on larger and more diverse datasets such as MIMIC-CXR and CheXpert, which include broader patient demographics, varied imaging modalities, and more contemporary reporting styles. This will allow us to assess not only predictive accuracy but also the stability and clinical plausibility of SHAP and LIME explanations across heterogeneous clinical contexts.

Take Home Points

The key points are as follows:

- These types of techniques, especially modern deep learning and LLMs, increase the effectiveness of diagnosis in the case of textual and image-based radiological data.
- Using XAI for decision support using LIME and SHAP leads to a better understanding of the AI models and their outcome, which implies better clinical decisions.
- The work cites studies that have shown advanced performance improvements over previous work with large models such as LLaMA-3.1 that achieved high accuracy rates, demonstrating a strong appreciation for the use of fine-tuned models in the clinical domain.
- Some weaknesses of this study included the concerns of dataset representativeness, because of challenges of explaining modeling outcomes in real-world clinical practice with respect to the existing methods of XAI approaches, and future directions and improvement of the model are recommended.
- This study emphasizes that there is a need for such integration whereby the intention of those algorithms is about prediction rather than replacing the clinical judgment of the clinicians.

Conclusions

In conclusion, our study reveals that among machine learning classifiers for textual radiological data, ensemble methods such as soft voting and hard voting achieved the highest precision and accuracy at 96% and 95%, respectively, outperforming individual classifiers such as LR and SVM, which had accuracies of 94%. In contrast, KNN had the lowest performance with an accuracy of 81%. For image data, SVM led among traditional machine learning models with an accuracy of 66%, while DenseNet169 excelled among deep learning classifiers with an accuracy of 84%. Notably, fine-tuned LLMs such as LLaMA-3.1 reached a remarkable accuracy of 98%, with strong agreement (Cohen κ=0.981) and statistical validation (P<.001). The McNemar test (P=.25) confirmed balance classification, significantly surpassing previous benchmarks in textual data analysis. While these results indicate strong potential for improving diagnostic support, external validation and prospective clinical studies are required to confirm their effectiveness in real-world clinical settings.

XSL•FO
RenderX

Acknowledgments

The work was done with partial support from the Mexican Government through the grant A1-S-47854 of Consejo Nacional de Ciencia y Tecnologıa (CONACYT), Mexico, and grants 20241816, 20241819, and 20240951 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, México. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE (Instituto Nacional de Astrofísica, Óptica y Electrónica), Mexico, and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

Data Availability

The dataset is publicly available through the US National Library of Medicine Open-I platform and can be accessed under a public domain license. The code used in this study is publicly available online [48].

Conflicts of Interest

None declared.

Multimedia Appendix 1
TRIPOD+AI checklist.
[DOCX File , 31 KB - formative_v9i1e77482_app1.docx]

References

1. Najjar R. Redefining radiology: a review of artificial intelligence integration in medical imaging. Diagnostics 2023;13:2760. [doi: 10.20944/preprints202306.1124.v1]
2. Quan H, Li S, Zeng C, Wei H, Hu J. Big data and AI-driven product design: a survey. Appl Sci 2023;13(16):9433. [doi: 10.3390/app13169433]
3. 3 SV, Konda SR. Neural networks and explainable AI: bridging the gap between models and interpretability. Int J Comput Sci Technol 2021;5:163-176. [doi: 10.5281/zenodo.10779335]
4. Pesapane F, Tantrige P, De Marco P, Carriero S, Zugni F, Nicosia L, et al. Advancements in standardizing radiological reports: a comprehensive review. Medicina (Kaunas) 2023;59(9):1679 [FREE Full text] [doi: 10.3390/medicina59091679] [Medline: 37763797]
5. Pons E, Braun LMM, Hunink MGM, Kors JA. Natural language processing in radiology: a systematic review. Radiology 2016;279(2):329-343. [doi: 10.1148/radiol.16142770] [Medline: 27089187]
6. Borys K. Explainable AI in medical imaging: an overview for clinical practice. Eur J Radiol 2023;157:110554. [doi: 10.1016/j.ejrad.2023.110554]
7. Muhammad MD, Khan S, Raza M. A systematic review of explainable artificial intelligence in medical image analysis. J King Saud Univ Comput Inf Sci 2024;36(1):1-15. [doi: 10.1016/j.jksuci.2023.05.001]
8. E Ihongbe I, Fouad S, F Mahmoud T, Rajasekaran A, Bhatia B. Evaluating explainable artificial intelligence (XAI) techniques in chest radiology imaging through a human-centered lens. PLOS One 2024;19(10):e0308758 [FREE Full text] [doi: 10.1371/journal.pone.0308758] [Medline: 39383147]
9. de Vries BM, Zwezerijnen GJC, Burchell GL, van Velden FHP, Menke-van der Houven van Oordt CW, Boellaard R. Explainable artificial intelligence (XAI) in radiology and nuclear medicine: a literature review. Front Med (Lausanne) 2023;10:1180773 [FREE Full text] [doi: 10.3389/fmed.2023.1180773] [Medline: 37250654]
10. Houssein EH, Gamal AM, Younis EMG, Mohamed E. Explainable artificial intelligence for medical imaging systems using deep learning: a comprehensive review. Cluster Comput 2025;28(7):1-22. [doi: 10.1007/s10586-025-05281-5]
11. Saw SN, Yan YY, Ng KH. Current status and future directions of explainable artificial intelligence in medical imaging. Eur J Radiol 2025;183:111884. [doi: 10.1016/j.ejrad.2024.111884] [Medline: 39667118]
12. Bhati D, Neha F, Amiruzzaman M. A survey on explainable artificial intelligence (XAI) techniques for visualizing deep learning models in medical imaging. J Imaging 2024;10(10):239 [FREE Full text] [doi: 10.3390/jimaging10100239] [Medline: 39452402]
13. Aasem M, Javed Iqbal M. Toward explainable AI in radiology: ensemble-CAM for effective thoracic disease localization in chest x-ray images using weak supervised learning. Front Big Data 2024;7:1366415 [FREE Full text] [doi: 10.3389/fdata.2024.1366415] [Medline: 38756502]
14. Bazzano AN, Patel T, Nauman E, Cernigliaro D, Shi L. Optimizing telehealth for diabetes management in the deep south of the United States: qualitative study of barriers and facilitators on the patient and clinician journey. J Med Internet Res 2024;26:e43583 [FREE Full text] [doi: 10.2196/43583] [Medline: 37976468]
15. Langlotz CP. Will artificial intelligence replace radiologists? Radiol Artif Intell 2019;1(3):e190058 [FREE Full text] [doi: 10.1148/ryai.2019190058] [Medline: 33937794]

XSL•FO
RenderX

16. Rajpurkar P, Irvin J, Ball RL, Zhu K, Yang B, Mehta H, et al. Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS Med 2018 Nov;15(11):e1002686 [FREE Full text] [doi: 10.1371/journal.pmed.1002686] [Medline: 30457988]
17. Selvaraju RR, Cogswell M, Das A. Grad-CAM: visual explanations from deep networks via gradient-based localization. 2017 Presented at: Proceedings of the IEEE International Conference on Computer Vision (ICCV); October 22-29, 2017; Venice, Italy p. 618-626. [doi: 10.1109/iccv.2017.74]
18. Alsentzer E, Murphy J, Boag W, Weng WH, Jindi D, Naumann T. Publicly available clinical BERT embeddings. 2019 Presented at: Proceedings of the 2nd Clinical Natural Language Processing Workshop; June 7, 2019; Minneapolis, Minnesota, USA p. 72-78. [doi: 10.18653/v1/w19-1909]
19. Lee J, Yoon W, Kim S, Kim D, Kim S, So C, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020;36(4):1234-1240 [FREE Full text] [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
20. Peng C, Yang X, Chen A, Smith KE, PourNejatian N, Costa AB, et al. A study of generative large language model for medical research and healthcare. npj Digit Med 2023;6(1):210 [FREE Full text] [doi: 10.1038/s41746-023-00958-w] [Medline: 37973919]
21. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. 2019 Presented at: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems; 2017 December 4 - 9; Long Beach California USA p. 4768-4777. [doi: 10.5260/chara.21.2.8]
22. Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": explaining the predictions of any classifier. 2016 Presented at: KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13-17, 2016; San Francisco, CA p. 1135-1144. [doi: 10.1145/2939672.2939778]
23. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2020;2(1):56-67 [FREE Full text] [doi: 10.1038/s42256-019-0138-9] [Medline: 32607472]
24. Du M, Liu N, Song Q, Hu X. Towards explanation of DNN-based prediction with guided feature inversion. 2018 Presented at: KDD '18: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; August 19-23, 2018; London, United Kingdom p. 1358-1367. [doi: 10.1145/3219819.3220099]
25. Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. WIREs Data Min Knowl Discovery 2019;9(4):e1312 [FREE Full text] [doi: 10.1002/widm.1312] [Medline: 32089788]
26. Tang H, Hu Z. Research on medical image classification based on machine learning. IEEE Access 2020;8:93145-93154. [doi: 10.1109/access.2020.2993887]
27. 27 ZY, Du J, Guan K, Wang T. Multi-modal broad learning system for medical image and text-based classification. 2021 Presented at: 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); November 01-05, 2021; Mexico p. 3439-3442. [doi: 10.1109/embc46164.2021.9630854]
28. Shakhovska N, Shebeko A, Prykarpatskyy Y. A novel explainable AI model for medical data analysis. J Artif Intell Soft Comput Res 2024;14:121-137. [doi: 10.2478/jaiscr-2024-0007]
29. Wiggins WF, Kitamura F, Santos I, Prevedello LM. Natural language processing of radiology text reports: interactive text classification. Radiol Artif Intell 2021;3(4):e210035 [FREE Full text] [doi: 10.1148/ryai.2021210035] [Medline: 34350414]
30. Chen MC, Ball RL, Yang L, Moradzadeh N, Chapman BE, Larson DB, et al. Deep learning to classify radiology free-text reports. Radiology 2018;286(3):845-852. [doi: 10.1148/radiol.2017171115] [Medline: 29135365]
31. Bangyal WH, Qasim R, Rehman N, Ahmad Z, Dar H, Rukhsar L, et al. Detection of fake news text classification on COVID-19 using deep learning approaches. Comput Math Methods Med 2021;2021:5514220 [FREE Full text] [doi: 10.1155/2021/5514220] [Medline: 34819990]
32. Qasim R, Bangyal WH, Alqarni MA, Ali Almazroi A. A fine-tuned BERT-based transfer learning approach for text classification. J Healthcare Eng 2022;2022:3498123 [FREE Full text] [doi: 10.1155/2022/3498123] [Medline: 35013691]
33. Oliveira GPD, Fonseca A, Rodrigues PC. Diabetes diagnosis based on hard and soft voting classifiers combining statistical learning models. Braz J Biom 2022;40(4):415-427. [doi: 10.28951/bjb.v40i4.605]
34. Ajibade SM, Dayupay J, Ngo-Hoang DL, Oyebode OJ, Sasan JM. Utilization of ensemble techniques for prediction of the academic performance of students. J Optoelectron Laser 2022;41(6):48-54.
35. 35 IE, Preventza O, Hamilton BJI, Augoustides J, Beck A. ACC/AHA guideline for the diagnosis and management of aortic disease: a report of the American heart association/American College of Cardiology Joint Committee on Clinical Practice Guidelines. J Am Coll Cardiol 2022;80:e223-e393. [doi: 10.3410/f.742397675.793596406]
36. Nguyen D, Le M, Cung T. Improving intrusion detection in SCADA systems using stacking ensemble of tree-based models. Bulletin EEI 2022;11(1):119-127. [doi: 10.11591/eei.v11i1.3334]
37. Roy DK, Sarkar TK, Kamar SSA, Goswami T, Muktadir MA, Al-Ghobari HM, et al. Daily prediction and multi-step forward forecasting of reference evapotranspiration using LSTM and Bi-LSTM models. Agronomy 2022;12(3):594. [doi: 10.3390/agronomy12030594]
38. Sy CY, Maceda LL, Canon MJP, Flores NM. Beyond BERT: exploring the efficacy of RoBERTa and ALBERT in supervised multiclass text classification. IJACSA 2024;15(3):223-233. [doi: 10.14569/ijacsa.2024.0150323]

XSL•FO
RenderX

39. Amisse C, Jijón-Palma ME, Centeno JAS. Fine-tuning deep learning models for pedestrian detection. Bol Ciênc Geod 2021;27(2):e2021013. [doi: 10.1590/s1982-21702021000200013]
40. Chen J, Liu Z, Huang X, Wu C, Liu Q, Jiang G, et al. When large language models meet personalization: perspectives of challenges and opportunities. World Wide Web 2024;27(4):42. [doi: 10.1007/s11280-024-01276-1]
41. Karimi H. Greater entropy leads to more explicit referential forms during language production. Cognition 2022;225:105093. [doi: 10.1016/j.cognition.2022.105093] [Medline: 35305301]
42. Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. BMJ Qual Saf 2019;28(3):231-237 [FREE Full text] [doi: 10.1136/bmjqs-2018-008370] [Medline: 30636200]
43. van Leeuwen KG, de Rooij M, Schalekamp S, van Ginneken B, Rutten MJCM. Clinical use of artificial intelligence products for radiology in the Netherlands between 2020 and 2022. Eur Radiol 2024;34(1):348-354 [FREE Full text] [doi: 10.1007/s00330-023-09991-5] [Medline: 37515632]
44. Wang Q, Chen D, Li M, Li S, Wang F, Yang Z, et al. A novel method for petroleum and natural gas resource potential evaluation and prediction by support vector machines (SVM). Appl Energy 2023;351:121836. [doi: 10.1016/j.apenergy.2023.121836]
45. HS S, Sooda K, Rai BK. EfficientNet-B7 framework for anomaly detection in mammogram images. Multimed Tools Appl 2025;84(11):8995-9021. [doi: 10.1007/s11042-024-18853-1]
46. Ashwini A, Purushothaman K, Rosi A, Vaishnavi T. Artificial intelligence based real-time automatic detection and classification of skin lesion in dermoscopic samples using DenseNet-169 architecture. IFS 2023;45(4):6943-6958. [doi: 10.3233/jifs-233024]
47. Aldughayfiq B, Ashfaq F, Jhanjhi NZ, Humayun M. Explainable AI for retinoblastoma diagnosis: interpreting deep learning models with LIME and SHAP. Diagnostics (Basel) 2023;13(11):1932 [FREE Full text] [doi: 10.3390/diagnostics13111932] [Medline: 37296784]
48. Radiology-Reports. GitHub. URL: https://github.com/tayyabawan786/Radiology-Reports [accessed 2025-09-26]

Abbreviations

AI: artificial intelligence
BERT: Bidirectional Encoder Representations from Transformers
Bi-LSTM: bidirectional long short-term memory network
CNN: convolutional neural network
DistilBERT: Distilled Version of Bidirectional Encoder Representations from Transformers
Grad-CAM: Gradient-Weighted Class Activation Mapping
KNN: k-nearest neighbor
LIME: Local Interpretable Model-Agnostic Explanations
LLM: large language model
LR: logistic regression
MeSH: Medical Subject Headings
NLP: natural language processing
RNN: recurrent neural network
RoBERTa: Robustly Optimized Bidirectional Encoder Representations from Transformers Pretraining Approach
SHAP: Shapley Adaptive Explanations
SVM: support vector machine
TF-IDF: term frequency–inverse document frequency
TRIPOD+AI: Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis–Artificial Intelligence
XAI: explainable artificial intelligence

Edited by J Sarvestan; submitted 14.05.25; peer-reviewed by N Nanthasamroeng, SHA Faruqui; comments to author 14.07.25; revised version received 02.08.25; accepted 22.09.25; published 14.10.25.

Please cite as:
Zamir MT, Khan SU, Gelbukh A, Felipe Riverón EM, Gelbukh I
Explainable AI-Driven Analysis of Radiology Reports Using Text and Image Data: Experimental Study
JMIR Form Res 2025;9:e77482
URL: https://formative.jmir.org/2025/1/e77482
doi:10.2196/77482
PMID:40997754

XSL•FO
RenderX

©Muhammad Tayyab Zamir, Safir Ullah Khan, Alexander Gelbukh, Edgardo Manuel Felipe Riverón, Irina Gelbukh. Originally published in JMIR Formative Research (https://formative.jmir.org), 14.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Protocol

Assessing Minority Stress and Physiological Response Through Ecological Momentary Assessment and Sensors: Protocol for a Feasibility and Acceptability of the Stress and Heart Pilot Study

Dulce Urueta Tapia[1], BA; Heather L Corliss[1,2,3], MPH, PhD; Kang Hyuk Lee[1], BA, MA; Jerel P Calzo[1,2,3], MPH, PhD; Hee-Jin Jun[1,2,3], MPH, SCD

[1]San Diego State University Research Foundation, San Diego, CA, United States
[2]School of Public Health, San Diego State University, San Diego, CA, United States
[3]San Diego State University, Institute for Behavioral and Community Health, San Diego, CA, United States

Corresponding Author:
Hee-Jin Jun, MPH, SCD
School of Public Health
San Diego State University
5500 Campanile Drive
San Diego, CA, 92182
United States
Phone: 1 3393680406
Email: hjun@sdsu.edu

Abstract

Background: LGBTQ+ (lesbian, gay, bisexual, transgender, queer, and other diverse sexual and gender identities) young adults may experience discrimination based on their sexual and gender minority status, which results in LGBTQ+ population health disparities. Although the health effects of minority stress have been studied for more than 30 years, most research relies on retrospective and cross-sectional designs. These methods limit the ability to establish causal links and to capture the fluctuating and time-sensitive nature of stress responses. In contrast, ecological momentary assessment (EMA) and wearable sensors allow for real-time tracking of stress exposures and their resulting physiological effects, enhancing data accuracy and ecological validity.

Objective: The EMA and sensor protocol of the Stress and Heart Study was developed to capture real-time physiological responses to minority stress experiences among LGBTQ+ young adults. This pilot study aims to evaluate the acceptability and feasibility of the Stress and Heart Study protocol.

Methods: Participants who identified as LGBTQ+ between the ages of 18 and 30 years who reported experiencing LGBTQ+ discrimination in the past 30 days in a screening survey were invited to participate in a 2-week EMA and sensor study. Participants received 4 daily EMA surveys and one end-of-day (EOD) survey on their smartphones, assessing general and minority stress experiences, positive and negative emotional states, and substance use. Heart rate variability was continuously recorded using a wearable sensor. Upon completion, participants completed a short exit survey to evaluate their study experience and satisfaction.

Results: Twenty participants aged 18 to 27 (mean 21.7, SD 2.6) years were enrolled, representing diverse sexual orientations (8 lesbian/gay, 2 bisexual, 4 pansexual, and 6 queer), gender identities (10 cisgender, 3 transgender, and 7 non-binary), and racial/ethnic backgrounds (9 non-Hispanic [NH] White, 5 Latinx, 2 NH Black, and 4 NH Asian). Participants completed 89.4% (1001/1120) of EMA daily surveys and 92.1% (258/280) of EOD surveys. On average, participants wore the sensor for 74.6% (SD 24%) of the expected time (179/240 hours). The overall EMA daily and EOD surveys completion rate was 89.9% (1259/1400), with individual participation ranging from 70% (49/70) to 98.6% (69/70). A total of 85% (n=17) of participants reported wearing the sensor daily. In the exit survey, all participants indicated that the study's time commitment met their expectations. Additionally, 90% (n=18) of participants reported that the sensor was comfortable and that the EMA app was user-friendly, with appropriately timed questions.

Conclusions: This study supports the feasibility and acceptability of the Stress and Heart Study protocol using smartphones and wearable sensors to collect real-time data on minority stress experiences and physiological response. Further research is needed to validate the use of this protocol in larger observational and intervention studies aimed at addressing the adverse health impacts of minority stress among LGBTQ+ populations.

XSL•FO
RenderX

(JMIR Form Res 2025;9:e68733) doi:10.2196/68733

KEYWORDS

ecological momentary assessment; sensor; heart rate variability; stress; minority stress; sexual and gender minorities; acceptability; feasibility

Introduction

Cardiovascular disease (CVD), which includes conditions such as coronary heart disease, cerebrovascular disease, and deep vein thrombosis that can lead to heart attack or stroke, is the leading cause of morbidity and mortality in the United States [1]. Evidence suggests that certain populations, including racial and ethnic minorities, individuals with low socioeconomic status, and sexual and gender minority (SGM) populations (ie, those who identify as lesbian, gay, bisexual, transgender, queer, and other diverse sexual and gender identities [LGBTQ+]), may be at increased risk for CVD [2-7]. However, research comparing the occurrence of CVD between LGBTQ+ and heterosexual cisgender populations has yielded mixed findings. Some studies report higher rates of CVD among specific LGBTQ+ subgroups [8-13] while others find similar or even lower rates [9-11,13,14].

Despite these mixed findings regarding CVD outcomes, a growing body of evidence indicates that LGBTQ+ populations experience poorer cardiovascular health (CVH) across multiple domains. These disparities include both clinical and physiological indicators (eg, hypertension, diabetes, high cholesterol, obesity, and depression) and behavioral risk factors (eg, tobacco and alcohol use, low physical activity, and poor sleep), compared with their heterosexual and cisgender counterparts [4,9,13,15-17].

A primary cause attributed to LGBTQ+ disparities in CVH and CVD is minority stress. Minority stress refers to the chronic stressors that LGBTQ+ individuals often encounter due to their stigmatized sexual orientation or gender identity [18-20]. The minority stress model posits that distal stressors (eg, discrimination, rejection, victimization, non-affirmation of identity) impact proximal stress processes (identity concealment, expectations of rejection/mistreatment, internalized stigma), leading to negative psychological, behavioral, and physiological outcomes. Research suggests that minority stress experiences are related to poorer CVH and greater risk for CVD [4,9,21,22].

Only a few studies have linked minority stress to CVD [9,23,24]. For example, one study found that greater exposure to discrimination and violence was associated with increased self-reported CVD among gay and bisexual adult men compared with heterosexual and cisgender counterparts [9]. Another study identified an indirect association between minority stress and CVD risk factors via reduced sense of mastery [24]. Additionally, transgender individuals have been shown to experience higher rates of cardiovascular morbidity and mortality, influenced by self-reported minority stress and gender-affirming hormone therapy [23]. These findings underscore the need to investigate physiological mechanisms linking minority stress to CVD outcomes.

Stress negatively affects CVH by disrupting the autonomic nervous system, which comprises the sympathetic and parasympathetic branches [25]. The sympathetic system triggers the “fight or flight” response, releasing cortisol and adrenaline to elevate heart rate and blood pressure, while the parasympathetic system promotes relaxation and recovery [25]. A balanced interaction between these systems is essential for physiological homeostasis. Chronic stress can disrupt this balance, leading to autonomic dysfunction and increasing the risk of CVD [26,27].

Heart rate variability (HRV), a validated noninvasive marker of autonomic nervous system function, reflects the variation in time intervals between heartbeats [28-30]. Low HRV reflects diminished parasympathetic activity and heightened sympathetic arousal, signaling autonomic imbalance and impaired homeostasis [25]. Chronic stress is associated with lower HRV, which in turn has been linked to increased CVD risk [31-35]. Low HRV also indicates reduced capacity to respond to both internal (eg, inflammation and chronic pain) and external (eg, discrimination, loneliness, and poverty) stressors, further contributing to physiological strain [29].

Given that LGBTQ+ individuals experience greater chronic stress compared with heterosexual and cisgender individuals, they may be more susceptible to lower HRV. However, studies explicitly connecting minority stress and HRV in LGBTQ+ populations are rare. A literature review identified only a few studies that examined this relationship [36,37]. Huebner [36] conducted a lab-based study involving 134 LGB (lesbian, gay, and bisexual) adults (mean age of 25.8 years) and found that exposure to an anti-gay social stressor during a laboratory task resulted in heightened cardiovascular reactivity—increased heart rate and cortisol—and reduced HRV, suggesting that minority stress can trigger immediate physiological changes that may contribute to long-term CVD risk.

Rosati et al [37] conducted another lab-based study comparing 19 LGB and 20 heterosexual White adults’ (mean age of 35 years) physiological responses to an emotional stress task. The LGB participants showed higher HRV compared with the heterosexual participants, contradicting the hypothesis that LGB people are more susceptible to low HRV. However, the same study also found that the LGB participants had higher total peripheral resistance (TPR)—an indicator of blood vessel constriction and increased strain on the cardiovascular system. This pattern is referred to as the “cardiovascular conundrum” [38,39], a physiological state characterized by both high TPR [38] and high HRV [39], which may result from chronic exposure to minority stress. Typically, high TPR is associated with low HRV. Given the limited and conflicting findings, additional research is needed to clarify how minority stress affects HRV and its implications for CVH.

XSL•FO
RenderX

In addition to laboratory-based research on the effects of stress on HRV, much of the literature to date has relied on cross-sectional designs [40-42] or longitudinal cohort studies [43]. While valuable, such studies often have recall bias and may fail to capture dynamic, real-world autonomic responses. Ecological momentary assessment (EMA) addresses these issues by collecting real-time data in participants' natural environments, greatly reducing recall bias and improving ecological validity [44]. When combined with wearable sensors (eg, smartwatches), EMA enables noninvasive, high-frequency HRV monitoring [45-48]. Although the number of studies integrating EMA with wearable HRV is still relatively small, this area is rapidly expanding. Recent research has examined the association between momentary stress and HRV in daily life and generally found that increases in perceived stress are associated with lower HRV, although the magnitude of this relationship is often small. For example, Akbar et al [45], using a wrist-worn smartwatch, observed that HRV-derived stress peaks corresponded with self-reported stress during work and after-hours EHR tasks, and Martinez et al [47] reported that HRV features were significantly but weakly associated with perceived stress. Kim et al [46], using a chest-strap sensor, demonstrated a bidirectional pattern, where lower HRV predicted higher perceived stress, and stress predicted HRV decreases within 10 minutes.

These studies highlight that perceived stress is linked to short-term autonomic changes; yet, existing EMA + HRV research has largely focused on general daily stress and has not captured the excess minority stress uniquely experienced by LGBTQ+ populations. The Stress and Heart Study was designed to fill this gap by measuring both general daily stress and minority stress exposure in real time, alongside autonomic responses captured via noninvasive, wearable sensors. In this pilot study, we evaluate the feasibility and acceptability of this protocol in 20 LGBTQ+ young adults recruited from community settings, providing a foundation for future research on minority stress and cardiovascular health in real-world environments.

Methods

Project Overview

To achieve the study aims, we developed a methodological protocol to evaluate real-time assessments of general stress and minority stress experiences among SGM young adults based on minority stress theory [18-20]. We created an EMA questionnaire to measure stress experiences, mood, and substance use. Four focus group discussions were conducted with the target population to gather feedback to evaluate the face validity of the questionnaires and inform refinement of the questionnaires and the study procedures. We then recruited participants who agreed to participate in a 14-day EMA and sensor study (EMA/sensor study) to assess feasibility and acceptability. Participants self-reported momentary minority stress experiences, mood, and substance use using a smartphone-based app. Their HRV was measured using a wearable sensor.

Target Population and Eligibility

Eligible participants met the following inclusion criteria: (1) aged 18-30 years; (2) identify as SGM; (3) reporting minority stress experiences at least "sometimes" on 2 or more occasions in the past year due to their SGM status; (4) agree to participate in focus group discussion for 2-hours (for focus group participants); (5) have a smartphone (for EMA/sensor study participants); and (6) agree to visit the study site twice and participate in the 2-week EMA/sensor study (for EMA/sensor study participants).

Sample Size and Recruitment Procedures

We recruited 14 participants for the focus group and another 20 for the EMA/sensor study. There was no overlap in participants in the 2 samples. The focus groups were conducted via an internal review board (IRB)–approved Zoom (Zoom Video Communications) session and lasted 1 hour. Fourteen participants were sufficient to assess our measurement tools and gather feedback to refine our procedures. The inclusion of 20 participants in the EMA/sensor survey offered initial insights into the feasibility of recruitment and retention strategies for future scalable projects. The sample size was based on prior research with similar methods and practical budget constraints, including device costs and participant incentives [49]. With this sample size and study duration, we anticipated an 80% compliance rate, defined as completing 56 of 70 surveys over 14 days (about 2 weeks).

We used web-based and in-person methods to recruit participants, sharing our flyer with LGBTQ+ organizations in the San Diego area via email and in-person visits. The team also posted flyers on university bulletin boards and coffee shops frequented by students. The flyer outlined eligibility, compensation, and a US $50 random drawing for prescreening survey applicants via QR code or URL. After completing the prescreening, participants received a confirmation email regarding their availability for either the focus group Zoom session or the EMA/sensor study's initial training at a private, nonclinical research facility affiliated with a university located in San Diego. To facilitate recruitment, individuals interested in the focus group or the EMA/sensor study who provided contact information were entered into a drawing for a US $50 gift card. A separate drawing was held for the Focus group and the EMA/sensor study. One winner was selected for each drawing and was notified via email.

Selection of Baseline and EMA Measures

The study team met weekly from October 2020 to December 2022 to curate our assessment measures for both baseline and EMA data collection. Our weekly meetings focused on measure selections and the suitability of each measure for the baseline survey, the EMA questionnaire, or other aspects of data collection. Through a comprehensive literature review of articles indexed in MEDLINE (PubMed) and Google Scholar, we identified measures of SGM stress informed by the minority stress framework [18-20], as well as measures of mental health and substance use behavior. The search strategy included terminology related to methods (eg, EMA, sensor, smartwatch, and wristband), target population (eg, LGBTQ+, sexual

XSL•FO
RenderX

orientation, gender identity, sexual and gender minority), substance involvement (eg, smoking, alcohol, marijuana, and drug consumption), and mental health (eg, depression, anxiety and affective states) relevant to our study objectives. Additional studies were also identified by examining the references cited in the literature. Emphasis was placed on identifying EMA studies or studies that addressed or included our target population.

To create concise EMA items, we reviewed validated measures to assess their validity and reliability in our target population. Factor analysis studies were examined to identify items with strong correlations to the construct, and the highest factor loading items were included in our Baseline and EMA protocol. For example, to assess sexual orientation–related stress, we selected items from the sexual orientation microaggression inventory-short form [50], retaining those with the highest factor loadings, such as "Heterosexism: overreact when talking about negative experience" (0.708) and "Societal Disapproval: don't mind LGBQ people, they just shouldn't be so public" (0.722). Once selected, the wording was adjusted to reflect the real-time, momentary nature of the measures (eg, since your last prompt). Our final measures are presented in Table 1. Focus group discussions were held to evaluate the face validity and acceptability of the selected items through cognitive assessment, a method that explores how participants interpret and respond to items [51].

Table 1. Baseline, ecological momentary assessment (EMA), and end-of-day (EOD) survey measures used in the Stress and Heart Study of LGBTQ+ (lesbian, gay, bisexual, transgender, queer, and other diverse sexual and gender identities) young adults, 2021-2022.

Variable	Baseline	EMA (4/day)		EOD
		Before Modification	After modification	
Demographic characteristics				
Age	✓			
Race/ethnicity	✓			
Sexual orientation	✓			
Gender identity	✓			
Education	✓			
Marital status	✓			
Employment	✓			
Income	✓			
Mood and mental health				
Affective emotional states (I-PANAS-SF) [52]	✓	✓	✓	
Generalized Anxiety Disorder-7 [53]	✓			
PHQ-9 (Patient Depression Questionnaire) [54]	✓			
Stress				
Minority stress				
Everyday Discrimination Scale [55]	✓	✓	✓	
Daily Heterosexist Experience Questionnaire (DHEQ) [56]	✓			
Sexual Minority Adolescent Stress Inventory (SMASI) [57]	✓			
Nebraska Outness Scale [58]	✓			
Sexual Orientation Microaggression Inventory (SOMI-SF) [50]			✓	
General stress				
Ecological Momentary Assessment of Stressful Events [59]			✓	
Current Stress Level [60]			✓	
Perceived Stress Scale-4 [61]				✓
Substance use				
Alcohol use [62,63]	✓	✓	✓	
Cigarette smoking, e-cigarette [62,63]	✓	✓	✓	
Marijuana [62-64]	✓	✓	✓	
Prescription drugs [63]	✓			
Other drugs [63]				✓

XSL•FO
RenderX

Sensor Selection

For the sensor selection, we conducted a thorough search involving several approaches: reviewing previous studies to measure HRV, investigating commercially available sensors, analyzing comparison studies of sensors used for HRV measurement, and assessing the validity and reliability of HRV measurements [65,66]. Additionally, we considered ease of use, affordability based on the study budget, and compatibility with the EMA app. We evaluated the feasibility and desirability of different sensor types, including chest straps, wristbands, and pulse oximeters. Chest straps are known to be the most accurate way to measure HR, but they are often uncomfortable and burdensome for participants due to the need for chest electrodes [65,67]. Pulse oximeters are not currently designed for use by researchers overseeing multiple participants. On the other hand, wristbands can collect HR data continuously from the wrist. Given these considerations, we selected the Garmin Vivosmart 4 wristband [68,69]. This device offers several advantages, including accessing real-time data that matches EMA time frames. The Garmin software development kit provides real-time data via Bluetooth directly from the device to the EMA app on the phone and to ilumivu's servers [70]. The Vivosmart 4 is lightweight, easy to wear, and has a long battery life, making it suitable for ambulatory assessment of HRV. It provides real-time heart rate and interbeat interval data, which are easily transmitted and integrated with the incoming EMA data.

Focus Groups

The research team conducted 4 web-based focus group discussions between September 13 and October 15, 2021, to assess the validity of the candidate EMA measures and to enhance the survey's acceptability and feasibility by gathering participant feedback. The research team and participants reviewed and discussed each survey question, leading to improvements. Participants also provided feedback on the Garmin Vivosmart4 watch, including their thoughts on adherence and barriers to wearing it while completing 5 daily EMA surveys on their smartphones over 2 weeks.

The focus group meetings were conducted via Zoom, a HIPAA (Health Insurance Portability and Accountability Act)-compliant video conferencing platform. Participants received confirmation emails with the meeting details, including the date, time, Zoom link, and focus group information. An Informed Consent Form documented the participants' understanding of the study's purpose and procedures, and their consent to participate. Participants reviewed and signed the digital consent form via Qualtrics before the session began. During the meetings, a research assistant presented the purpose of the study and the scheduling of the EMA and end-of-day (EOD) survey questions. Participants received a US $30 online gift card, chosen according to their preference, as an incentive for their participation. Feedback from the focus group discussions led to several important items being revised or added. Details of the modification are summarized in Table 2.

Table 2. Modifications to ecological momentary assessment (EMA) and end-of-day survey items based on focus group feedback in the stress and heart study of LGBTQ+ (lesbian, gay, bisexual, transgender, queer, and other diverse sexual and gender identities) young adults, 2021-2022.

Survey questions and original item	New or revised item
Everyday Discrimination Scale [55]	
Characteristics attributable to discrimination	
N/A[a]	Add gender identity
N/A	Add religion
N/A	Add age
Place of discrimination	
N/A	Add LGBTQ community space
Substance use	
Reasons and motivations	
Boredom	Release/manage boredom
Peer pressure	Socialize/confirm socially
N/A	Add influenced by alcohol/other substance
End-of-day survey	
Daily device reminder	
N/A	Add at the end: "Please do not forget to charge your phone and watch and put the watch on in the morning."

[a]N/A: not applicable.

XSL•FO
RenderX

EMA/Sensor Study

Study Procedure

Eligible participants who agreed to take part in the 2-week EMA/sensor study—conducted between April 23 and November 29, 2022—were invited to the study office located at SDSURF. The study was conducted in a private office room within the research facility to ensure a quiet, comfortable, and confidential environment for participants. Upon the participant's arrival, a research assistant welcomed the participant and introduced the research team members. The assistant outlined the agenda for the first meeting and described the participant's involvement over the next 2 weeks. The participant was then guided to review thoroughly the Informed Consent Form before signing it. Once the participant reviewed and signed the Informed Consent Form, the meeting continued with the participant filling out the baseline survey.

Baseline Survey

The 15-minute baseline Qualtrics survey assessed demographic characteristics and validated measures of mood and mental health, minority stress, general stress, and substance use. A complete list of baseline measures is presented in Table 1.

Training

Following completion of the baseline survey, the research assistant reviewed the study timeline. Participants were informed that, over the 2-week period, they would complete 4 EMA surveys and one EOD survey daily. The brief EOD survey captured experiences related to stress and substance use not covered in the EMA prompts (Table 1 displays survey content). To improve response rates and data reliability, we used an interval-contingent schedule for EMA prompts, delivering surveys at consistent, predetermined times throughout the day [44,71].

The research assistant worked with each participant to customize the 2-week EMA schedule, which consisted of 10 weekdays and 4 weekend days (2 weekend days each week). The survey typically started about one hour after waking and continued until bedtime, with approximately 3-hour intervals between surveys. The first prompt began within an hour of waking, and the remaining 3 followed at roughly 3-hour intervals. Start times could vary between weekdays and weekends to accommodate participants' routines. For those with fixed commitments (eg, classes or work meetings), survey intervals were flexibly adjusted within a range of 2 to 4.5 hours. Across all participants, the first survey of the day occurred between 6 AM and 11:45 AM, the fourth survey between 4 PM and 9 PM, and the EOD survey—administered near bedtime—occurred between 9 PM and 12 AM.

After survey scheduling, a follow-up Zoom meeting was arranged for approximately 2 days after the survey start date to troubleshoot any issues. Participants then began training, which included downloading and syncing the mEMA Sense smartphone application to their personal device. The mEMA app, developed by ilumivu Inc, is available at no cost for iOS and Android platforms. It collects real-time survey data and connects via Bluetooth to the Garmin Vivosmart 4 wearable sensor, which continuously records physiological data such as heart rate and interbeat intervals for HRV estimation. Participants were guided through the process of syncing the wristband with the app and completing test EMA and EOD surveys. Figure 1 displays example screens from the mEMA app interface.

XSL•FO
RenderX

Figure 1. Screenshots of sample ecological momentary assessment (EMA) question and sensor data delivered via the mEMA smartphone app, 2021-2022.

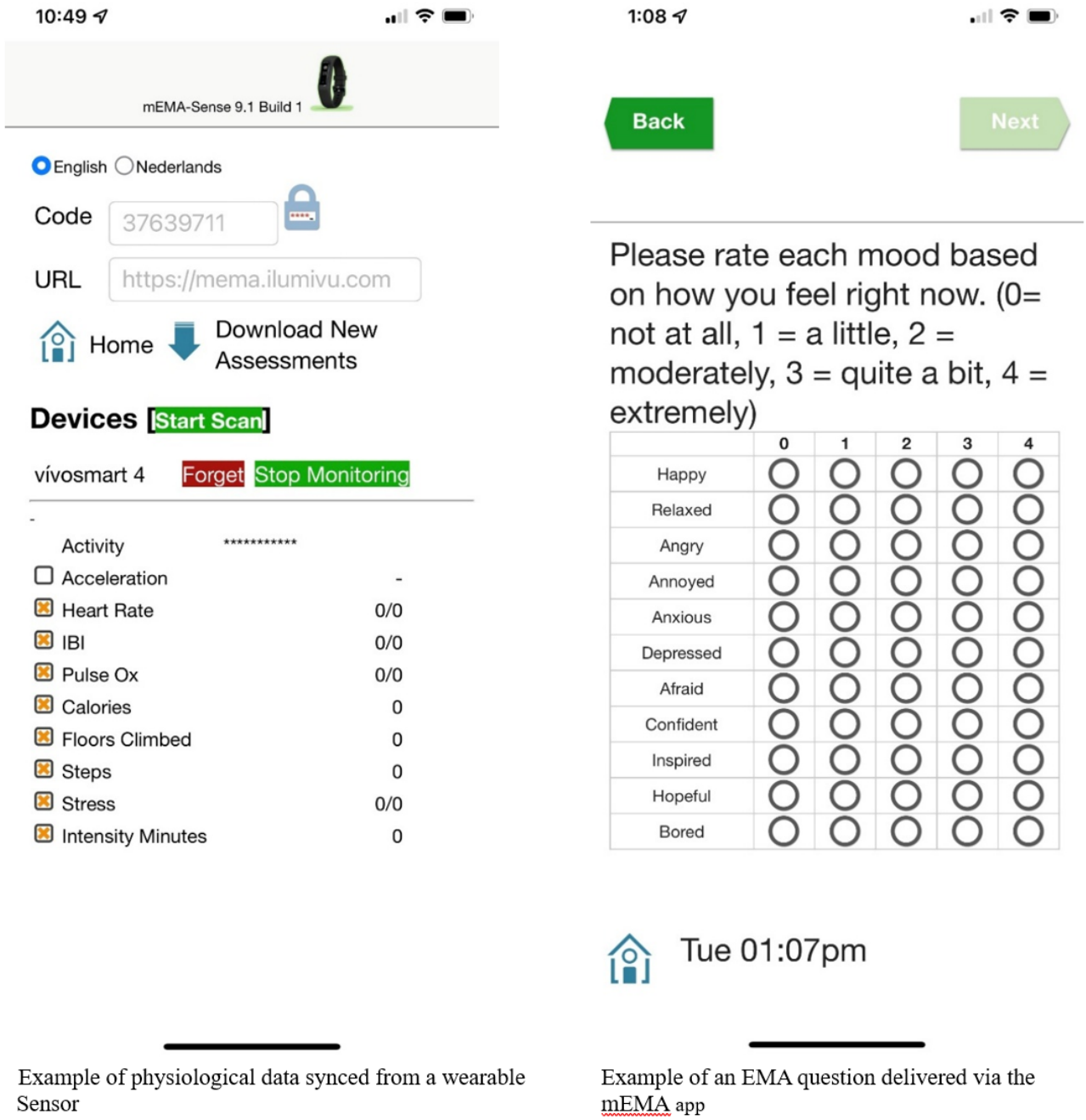

Example of physiological data synced from a wearable Sensor

Example of an EMA question delivered via the mEMA app

After completing the training session, an in-person exit visit was scheduled for participants to return their device and complete the exit survey. To ensure proper return of the device, participants were presented with a Device Loan Agreement, which they signed, stating that the device was for exclusive use in data collection during the study, was rented on a short-term basis, and must be returned within 7 days of the conclusion of their participation. Additionally, participants were asked if they preferred to receive daily reminders through the Remind app [72], a communication tool for sending messages. All participants agreed to receive a daily reminder text each morning to help them stay on track with their survey schedule. The Remind app was also used to communicate any issues participants encountered during the study period, including scheduling changes or other study-related matters.

The mEMA app was programmed to deliver EMA surveys at 4 participant-chosen times per day during waking hours. The first prompt began within an hour after waking and was followed by 5 push notifications at 5-minute intervals (5, 10, 15, 20, and 25 minutes). If the participant did not respond within 30 minutes, the survey was programmed to be closed and marked as a nonresponse. To minimize participant burden and reduce potential bias, several design features were implemented. EMA surveys were designed to be brief, taking no more than 3 minutes to minimize burden [73], and were limited to 5 times a day. A daily EOD survey—administered before bedtime—assessed stress and substance use experiences not covered in the EMA surveys (Table 1 displays content). After the 14-day data collection period, participants completed a brief exit survey including quantitative and open-ended questions on survey

XSL•FO
RenderX

clarity, difficulty, and overall satisfaction with the study protocol. The exit survey took less than 10 minutes to complete (Table S3 in Multimedia Appendix 1 shows details). Responses were analyzed to refine the EMA study protocol for future iterations.

EMA Survey Questionnaire and Modification

We established the EMA questionnaire based on a thorough literature review, discussions among our research team, and validation through focus group discussions. Several items we used were originally developed for general survey use; thus, we implemented some adaptations for the EMA context. Detailed questionnaire items and their response options are provided in Table S1 in Multimedia Appendix 1.

Initially, we used the Everyday Discrimination Scale [55] to measure minority stress. However, during the execution of the EMA study, we observed a significant limitation. Relying on the Everyday Discrimination Scale alone to assess minority stress provided an incomplete picture of the multifaceted SGM-related and general stressors encountered by participants. To measure participants' stress experiences more comprehensively, it was essential to supplement the measurement of minority stress with concurrent assessment of general stress, encompassing daily stressors. Additionally, other measures of minority stress were needed to capture the breadth of the minority stress experiences within this cohort.

Thus, the EMA/sensor study was divided into 2 phases due to the addition of new survey items. The first phase included 6 participants, and the second included 14 participants. Table 1 depicts specific modifications made between these 2 phases and Table S1 in Multimedia Appendix 1 shows details on item wording. Eight additional questions were selected from the sexual orientation microaggression inventory-short form in phase 2 to gauge minority stress related to participants' SGM status [50]. These questions addressed issues such as anti-gay attitudes and expressions, denial of sexuality, heterosexism, and societal disapproval. We also introduced a time categorization for stressful events, ranging from 0 to 30 minutes ago, 30 to 60 minutes ago, 1 to 2 hours ago, 2 to 3 hours ago, and 3+ hours ago, to enable stressful events to be measured at more proximal time points. Furthermore, 2 general stress measures were also added in phase 2. One measure asked participants to rate their current stress levels on a scale from not at all to extremely [60]. The other measure, the ecological momentary assessment of stressful events by the science of behavior change, encompasses inquiries about stress exposure, stressor type, timing, and intensity [59].

Safety Protocol

There is no evidence that repeated measures of stress introduce mental health issues or self-harm behaviors in studies on stress; however, given the high prevalence of mental health concerns among our target population and to minimize any potential harm associated with recurrent stress inquiries, our team implemented a safety protocol aimed at safeguarding the mental well-being of participants in our study.

Upon enrollment, participants received a comprehensive list of mental health resources, which contained information on local and national mental health resources, including the 988 Suicide and Crisis Lifeline. Additionally, a licensed clinical psychologist trained all research personnel who interacted with participants to manage adverse events. This training included instructions on how to implement the safety protocol and respond to different risk levels.

To identify participants who noted suicidal ideation (or other potential harmful situations we would need to respond to right away), we used a feature in the mEMA app that would send an immediate email notification when participants typed in a response to the question noted above. The automated email contained the date, time, and content of the response. Upon receipt of the email notification, the study staff would review the response for signs of intent to harm self or others. In the case that there was cause for concern, an email was sent to the participant asking for a Zoom meeting and informing the principal investigator and a supervising psychiatrist of the situation, along with additional information about the participant and the Zoom meeting.

During the meeting, the team members assessed suicide risk using the Columbia-Suicide Severity Rating Scale [74], gathering details on current thoughts, intent, access to means, protective factors, and past attempts. If risk is identified, the member will briefly consult their supervisor, per study protocol. While keeping the participant on Zoom, the team member mutes their audio/video and contacts the psychiatrist. Based on the psychiatrist's evaluation, the next steps fall into one of three categories. (1) Emergency: If emergency services are needed, the supervisor joins the Zoom briefly before calling for help while the staff member stays with the participant and engages in calming or distraction activities. (2) High risk (no Emergency): If the participant is high risk but stable, the staff member rejoins the meeting and completes a Safety Plan with them. (3) Low risk: The participant is offered to complete a Safety Plan [75,76], a tool designed for all risk levels to help individuals navigate suicidal feelings and urges, which contains a prioritized list of coping strategies and sources of support that can be used in times of distress. In all cases, the session ends with the participant being provided crisis resources.

We encountered 2 safety incidents during the study. The first incident occurred when a participant completed the EOD survey and responded to the question, "Is there anything else you'd like to report about your day?" and indicated that they had suicidal impulses but did not attempt. As soon as the study team became aware of this event, a team member spoke with the participant via phone to assess their current suicidality and offered the List of Resources for Emotional Distress. The next day, a study team member met with the participant via a Zoom meeting to assess their current mental health status and any suicidal thoughts. The participant stated that they did not have any current suicidality. Following this incident, we voluntarily paused the study and implemented our revised risk assessment protocol described above.

The second incident also involved a participant completing the EOD survey. In response to the same question that triggered the first event, the participant mentioned contemplating suicide without any specific plans or intent. Upon receiving this

XSL•FO
RenderX

notification, the study team promptly consulted the clinical psychologist and arranged a Zoom meeting with the participant. The following day, study team members met with the participant via Zoom and assessed the participant's current risk. Under the clinical psychologist's guidance, we determined that the participant's risk was moderate. We then completed the Safety Plan and provided the participant with the mental health resources list.

Ethical Consideration

All study procedures were reviewed and approved by the IRB at San Diego State University Human Research Protection Program (HS-2020-0016). All participants in the pilot study provided informed consent prior to enrollment, in accordance with IRB-approved protocols. Consent procedures for the EMA/sensor study ensured that participants understood the nature of the study, including the use of wearable sensors and EMA to collect real-time data on stress experiences and physiological responses. For any secondary analyses conducted using data from this study, the original informed consent included provisions for future use of deidentified data, and these analyses were also reviewed and approved by the IRB. No additional consent was required for the secondary use of data.

All data collected during the study were treated with strict confidentiality. Survey responses and physiological data were linked using participant ID codes and stored separately from identifying information. Data were deidentified prior to analysis, and only approved research personnel had access to the key linking IDs to identities. Electronic data were stored on secure, encrypted servers in compliance with institutional data security policies. No personally identifiable information was included in any publications or reports resulting from this research.

Participants could receive up to US $200 in gift cards for full participation in the study, which included attending the initial meeting, completing daily surveys over a 14-day period, returning the wearable device, and completing the exit survey. Compensation was prorated based on the level of participation, with tiered incentives tied to EMA survey completion rates. Participants received a US $40 gift card after the initial meeting, and up to US $160 upon study completion. Those who completed 50%-80% of the surveys received an additional US $30 gift card, and those who completed over 80% received an extra US $30. All participants returned the wearable devices, which were loaned under an agreement that outlined the return policy and potential withholding of incentives if devices were not returned.

Data Analysis

The feasibility and acceptability of the EMA/sensor study were assessed using participant response rates and responses to the exit survey. For feasibility, adherence was evaluated by tracking participants' compliance with the study protocol, including the percentage of EMA surveys completed and the number of days participants wore the sensor. We also examined prompt delivery success, overall study retention, and daily trends in response rates (eg, weekday vs weekend completion). Prompt-to-response time was calculated to assess the delay between signal and survey initiation, and survey completion times were compared between the 2 study phases due to the addition of EMA items. For acceptability, participants completed a structured exit survey, in which they rated the relevance, length, and frequency of the EMA surveys using 5-point Likert scales ranging from "completely disagree" to "completely agree." Additional questions assessed whether the study met their expected level of commitment, whether completing study activities caused added stress, and their willingness to participate in similar studies in the future. We also explored whether EMA response rates differed by demographic characteristics, including age, sex assigned at birth, race/ethnicity, sexual orientation, gender identity, income, and marital status.

Results

Demographics

The sample consisted of participants aged 18 to 27 (mean 21.7, SD 2.6) years and included a diverse range of gender identities, sexual orientations, and racial/ethnic backgrounds. Detailed demographic characteristics are presented in Table 3.

XSL•FO
RenderX

Table 3. Demographic characteristics of the ecological momentary assessment/sensor study participants of the Stress and Heart Study of LGBTQ+ (lesbian, gay, bisexual, transgender, queer, and other diverse sexual and gender identities) young adults (N=20), 2021-2022.

Category	Total sample (N=20)	Assigned male at birth (N=6)	Assigned female at birth (N=14)
Age (years), mean (SD)	21.7 (2.59)	21.2 (2.23)	21.9 (2.87)
Race/ethnicity, n (%)			
Non-Hispanic White	9 (45)	2 (33.33)	7 (50)
Latinx	5 (25)	2 (33.33)	3 (21.43)
Non-Hispanic Black	2 (10)	0 (0)	2 (14.29)
Non-Hispanic Asian	4 (20)	2 (33.33)	2 (14.29)
Sexual orientation , n (%)			
Lesbian/Gay	8 (40)	3 (50)	5 (35.71)
Bisexual	2 (10)	1 (16.67)	1 (7.14)
Pansexual	4 (20)	2 (33.33)	2 (14.29)
Queer	6 (30)	0 (0)	6 (42.86)
Gender identity , n (%)			
Cisgender man	4 (20)	4 (66.67)	0 (0)
Cisgender woman	6 (30)	0 (0)	6 (42.86)
Transwoman/Transfeminine	1 (5)	1 (16.67)	0 (0)
Transman/Transmasculine	2 (10)	0 (0)	2 (14.29)
Non-binary	7 (35)	1 (16.67)	6 (42.86)
Employment status, n (%)			
Employed full-time	5 (25)	2 (33.33)	3 (21.43)
Employed part-time	6 (30)	1 (16.67)	5 (35.71)
Student	9 (45)	3 (50)	6 (42.86)
Marital status , n (%)			
Single	13 (65)	5 (83.33)	8 (57.14)
Living with a partner	2 (10)	0 (0)	2 (14.29)
Not living with a partner	5 (25)	1 (16.67)	4 (28.57)
Income (US $) , n (%)			
Under 20,000	12 (60)	3 (50)	9 (64.29)
20,000-40,000	3 (15)	0 (0)	3 (21.43)
40,000-60,000	2 (10)	1 (16.67)	1 (7.14)
60,000-80,000	1 (5)	1 (16.67)	0 (0)
Prefer not to answer	2 (10)	0 (0)	1 (7.14)

Feasibility of EMA/Sensor Study

All 20 enrolled participants completed the EMA/sensor study, including the exit survey, with no attrition. A total of 1400 EMA prompts (5 per day × 14 days × 20 participants) were scheduled and successfully delivered. All prompts were sent as planned, with no delivery failures due to technical issues or participant-related factors. Participants completed 89.4% (1001/1120) of EMA surveys and 92.1% (258/280) of EOD surveys. A total of 18 (90.0%) out of 20 participants completed 80% or more of both EMA and EOD surveys. The average completion rate of all surveys was 89.9% (1259/1400), with a minimum of 70.0% (49/70) and a maximum of 98.6% (69/70). The average response rate across all participants ranged from 83.4% to 93.5% for each day of the week, with lower rates observed on weekends (86.1%) compared with weekdays (91.5%).

We also examined whether response rates varied by demographic characteristics, including age, sex assigned at birth, race/ethnicity, sexual orientation, gender identity, income, and marital status. There was a trend toward higher response rates among racial/ethnic minorities compared with non-Hispanic White individuals (P=.057) and among individuals assigned female at birth compared with those assigned male at birth (P=.139).

XSL•FO
RenderX

The median time to respond to the prompt signal was 0 minutes, and the average response time was 4.3 minutes. While more than two-thirds of participants responded immediately after receiving the prompt, in some instances, there were delays in response. We had programmed the survey window to close 30 minutes after opening using the mEMA app; however, upon inspection during data analysis, we discovered that the window did not consistently close as intended. In 6.6% (83/1259) of instances, responses were submitted more than 30 minutes after the prompt was issued. This issue appears to be due to a technical limitation of the software. On average, participants took 1 minute and 2 seconds to complete EMA surveys and 58 seconds to complete EOD surveys. Three additional questions were added to the EMA survey after the first 6 participants; as a result, the average completion time was shorter for the first 6 participants (54 seconds) compared with the last 14 participants (65 seconds). Participants wore the sensor for 74.6% (SD 24%) of the expected time (240 hours per person). Seventeen participants (85%) wore the sensor every day, while 3 participants missed at least one day.

Findings From the Exit Survey

All 20 participants completed exit surveys, with 90% agreeing that the mEMA app and sensor were easy to use. All participants found the EMA questions relevant, felt that the study met their expected commitment level, and expressed willingness to participate in a similar study in the future. Most agreed the survey length was appropriate, the time commitment manageable, and the study did not add stress, though one participant reported additional stress. Full details can be found in Figure 2, and the exit survey questionnaire is in Multimedia Appendix 1.

Figure 2. Participants' exit survey feedback on the feasibility and acceptability of the ecological momentary assessment (EMA)/sensor protocol, stress, and heart study of LGBTQ+ (lesbian, gay, bisexual, transgender, queer, and other diverse sexual and gender identities) young adults (N=20), 2021-2022.

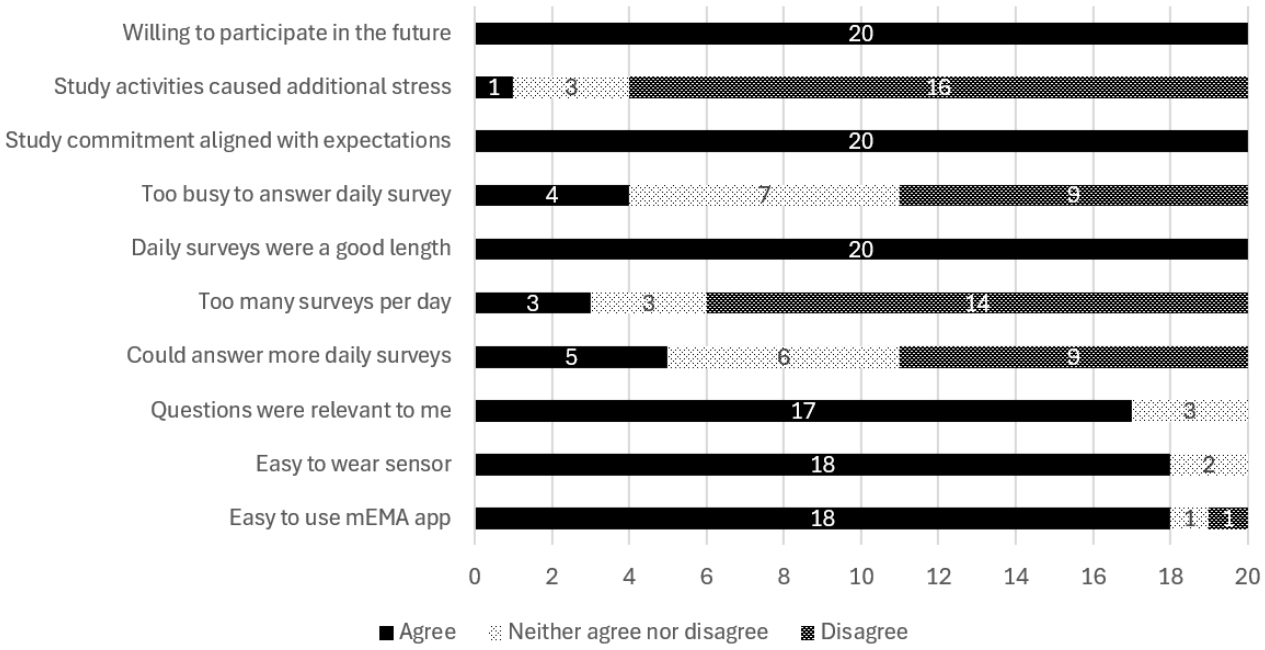

Discussion

Principal Findings

Our study aimed to assess the feasibility and acceptability of using smartphone-based EMA and wearable sensors to evaluate real-time measurements of minority stress experiences and physiological responses among SGM young adults. The high compliance of EMA survey responses and sensor wearing suggests the methodology works well in the target population and can be used for future research. Findings from the exit surveys indicate that participants perceived the study procedures as acceptable and feasible. Most participants reported that the frequency and timing of the EMA prompts and reminders were suitable for the study and their daily lives. In addition, the survey completion time was short, and participants reported that the questions were relevant to their daily experiences and easy to understand. While EMA prompts were programmed to expire 30 minutes after delivery, a technical issue with the mEMA app prevented the enforcement of this time limit. As a result, 6.6% of completed surveys were submitted after 30 minutes. These responses were retained, as they still fell within the broader intended window of assessing experiences since the last prompt. This limitation highlights the importance of verifying software functionality in future EMA studies. The study also encountered some safety concerns, but the study team responded promptly with the established safety protocol. Participants who expressed mental health issues were evaluated and found not to be at high risk for suicidality.

Smartphone-based EMA and wearable sensor studies can face challenges, such as participant burden, noncompliance, and reactivity to the protocol measures. Despite these potential challenges, our study demonstrated an excellent response rate of 89.4% (1001/1120) to the EMA surveys, higher than the average reported in other EMA studies [71,77,78]. In a meta-analysis of 477 EMA studies, these studies, on average, scheduled 6 assessments per day over 7 days, achieving a

XSL•FO
RenderX

compliance rate of 79% [71]. Studies with more frequent daily assessments tended to have shorter overall assessment periods. Compliance was not significantly influenced by the number of daily assessments, study design, or sample characteristics but was positively associated with financial incentives. To enhance compliance, we implemented a twofold incentive strategy, which has been used in previous studies [78,79]. In this approach, rewards were distributed after the baseline/training sessions and the exit survey. The reward amount was increased incrementally based on the participant's response rate. This suggests that this approach to delivering incentives might be an effective strategy for maintaining participant engagement in this target population. Additionally, proactive measures were taken to enhance compliance, including daily morning text message reminders [72] and ongoing communication with participants to address inquiries or concerns. Nonetheless, the small sample size of 20 participants makes it difficult to extrapolate the study's findings on compliance among all SGM young adults or subgroups.

Limitations

Wearable sensors also have challenges. Garmin vivosmart4 uses photoplethysmography (PPG), a noninvasive technology that uses a light source and photodetector at the skin's surface to measure volumetric variations of blood circulation [80-82]. PPG is comfortable and safe for continuous monitoring, which allows for continuous and real-time monitoring of physiological parameters such as heart rate and HRV [83]. However, the accuracy of PPG measurement can be affected by motion artifacts, the fit and position of the device on the body, variations in skin tone, and external factors such as ambient light, temperature, and pressure applied by the sensor [84-86]. The wearable device used in this study was developed for consumer use, not for research, and may not provide accurate heart rate measurements consistently. However, studies to test the validity of the measurement of heart rate with wearables, including previous versions of Garmin devices, have shown acceptable readings [87-89]. Sensor technology has been rapidly improving, and commercial devices are consistently being upgraded. As these advancements continue, the accuracy of these devices will improve, as will the accuracy and reliability of research findings.

Using fixed EMA times selected by participants may have introduced bias if certain time periods were consistently avoided, limiting variability in the data [44,90]. One possible consequence is an underreporting of stress experiences. While random survey timing can lessen this bias by adding unpredictability and better capturing transient events [44], 2 features of our design help mitigate the concern. First, our primary outcome, HRV, was continuously measured by wearable devices and is independent of EMA timing. We calculated the average HRV over several windows (eg, 30 or 60 minutes before each survey and the full interval between surveys); full details of this averaging method will appear in a forthcoming outcome paper. Second, our key EMA variable captured stress experiences since the last prompt, reducing sensitivity to exact survey timing. Future studies may consider randomized prompts within fixed windows (eg, ±30 minutes of a scheduled time) to reduce reactivity while maintaining structure and compliance.

In our study, sensor response rates were lower than EMA response rates. This difference likely stems from the existing use of cell phones versus the addition of a new wearable sensor. EMA surveys were administered through participants' cell phones, which are indispensable and rarely left behind, resulting in consistently high compliance. In contrast, sensors are a new device to be worn each morning and charged nightly. These additional steps can lead to occasional lapses in use, contributing to greater variability in sensor data collection. We used the ilumivu app for EMA and sensor measurements. For EMA, we monitored the response rate at the end of each day. However, we could only determine if additional data points were collected with the sensor. We could not ascertain the duration of wear until we downloaded and transferred the data into a MATLAB application [91]. During the data management phase, we developed an application to evaluate the sensor response rate daily throughout the study. We plan to use this application in our next study as an aid to improve participants' frequency of wearing the sensor.

Conclusion

This study adds to the literature by reporting on the development of an acceptable and feasible research approach that connects real-time minority stress experiences with real-time physiological responses by incorporating EMA methods and sensor technology among SGM young adults. Our goal was to develop a protocol for the study to understand the real-time association between stress experiences and physical health among SGM young adults and assess the feasibility and acceptability of our protocol.

Our findings highlight the feasibility and acceptability of this method. The results demonstrate the suitability of using smartphones and wearable sensors to capture real-time data for assessing the minority stress experiences and physical health of SGM young adults. Furthermore, the findings imply that real-time assessment of minority stress experiences using smartphone apps, along with real-time assessment of stress responses via wearable sensors, has the potential to be an effective tool for future stress interventions. Future research should consider extending the utility of this protocol to more diverse populations, for example, racial/ethnic minorities or economically disadvantaged populations that have health disparities linked to stress exposure.

Acknowledgments

We would like to thank all individuals who participated in this pilot study. Research reported in this publication was supported by the National Institute on Minority Health and Health Disparities of the National Institutes of Health (NIH) under Award Numbers S21MD010690 (San Diego State University (SDSU) HealthLINK Endowment) and U54MD012397 (SDSU HealthLINK Center). The NIH was not involved in the design, data collection, analysis, interpretation, or writing of this study. The content is

XSL•FO
RenderX

solely the responsibility of the authors and does not necessarily represent the official views of the NIH. We also acknowledge Paris DaSilva and Cameron Wadstrom for their valuable contributions as research assistants on this project.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
EMA, End-of-Day, and Exit Surveys Questionnaires.
[DOCX File , 42 KB - formative_v9i1e68733_app1.docx]

References

1. Martin SS. Correction to: 2024 heart disease and stroke statistics: a report of US and global data from the American heart association. Circulation 2024;149(8):e347-e913. [doi: 10.1161/cir.0000000000001247]
2. Borkowski P. Racial and socioeconomic determinants of cardiovascular health: a comprehensive review. Cureus 2024;16(5):e59497. [doi: 10.7759/cureus.59497]
3. Bonomo JA, Luo K, Ramallo JA. LGBTQ+ cardiovascular health equity: a brief review. Front Cardiovasc Med 2024;11:1350603 [FREE Full text] [doi: 10.3389/fcvm.2024.1350603] [Medline: 38510198]
4. Caceres BA, Streed CG, Corliss HL, Lloyd-Jones DM, Matthews PA, Mukherjee M, et al. Assessing and addressing cardiovascular health in LGBTQ adults: a scientific statement from the American heart association. Circulation 2020 Nov 10;142(19):e321-e332. [doi: 10.1161/cir.0000000000000914]
5. Cossío Aranda JE, Jain V, Figueiras-Graillet LM, Arias-Mendoza A, Cuéllar JL, Berni Betancourt A, et al. Toward a better understanding of cardiovascular risk in the transgender and gender-diverse community: a global call to action. Glob Heart 2024;19(1):27 [FREE Full text] [doi: 10.5334/gh.1300] [Medline: 38434153]
6. Satti DI, Chan JSK, Mszar R, Mehta A, Kwapong YA, Chan RNC, et al. Social determinants of health, cardiovascular health, and mortality in sexual minority individuals in the United States. J Am Coll Cardiol 2025;85(5):515-525. [doi: 10.1016/j.jacc.2024.11.026]
7. Zalabak A, Gantioque R. Cardiovascular health disparities among sexual and gender minority adults. J Nurse Pract 2025;21(2):105295. [doi: 10.1016/j.nurpra.2024.105295]
8. Sexual orientation differences in access to care and health status, behaviors, and beliefs: findings from the national health and nutrition examination survey, national survey of family growth, and national health interview survey. Centers for Disease Control and Prevention. Hyattsville, MD; 2022. URL: https://dx.doi.org/10.15620/cdc:115982 [accessed 2025-09-13]
9. Caceres BA, Brody A, Luscombe RE, Primiano JE, Marusca P, Sitts EM, et al. A systematic review of cardiovascular disease in sexual minorities. Am J Public Health 2017;107(4):e13-e21. [doi: 10.2105/ajph.2016.303630]
10. Caceres BA, Makarem N, Hickey KT, Hughes TL. Cardiovascular disease disparities in sexual minority adults: an examination of the behavioral risk factor surveillance system (2014-2016). Am J Health Promot 2019;33(4):576-585 [FREE Full text] [doi: 10.1177/0890117118810246] [Medline: 30392384]
11. Poteat T, Gallo LC, Harkness A, Isasi CR, Matthews P, Schneiderman N, et al. Influence of stress, gender, and minority status on cardiovascular disease risk in the Hispanic/Latino community: protocol for a longitudinal observational cohort study. JMIR Res Protoc 2021;10(5):e28997 [FREE Full text] [doi: 10.2196/28997] [Medline: 33955843]
12. Poteat TC, Divsalar S, Streed CG, Feldman JL, Bockting WO, Meyer IH. Cardiovascular disease in a population-based sample of transgender and cisgender adults. Am J Prev Med 2021;61(6):804-811 [FREE Full text] [doi: 10.1016/j.amepre.2021.05.019] [Medline: 34364725]
13. Sherman J, Dyar C, McDaniel J, Funderburg NT, Rose KM, Gorr M, et al. Sexual minorities are at elevated risk of cardiovascular disease from a younger age than heterosexuals. J Behav Med 2022;45(4):571-579 [FREE Full text] [doi: 10.1007/s10865-021-00269-z] [Medline: 35034218]
14. Tran NK, Lunn MR, Schulkey CE, Tesfaye S, Nambiar S, Chatterjee S, et al. Prevalence of 12 common health conditions in sexual and gender minority participants in the all of Us research program. JAMA Netw Open 2023;6(7):e2324969 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.24969] [Medline: 37523187]
15. Caceres B, Sharma Y, Doan D, Ravindranath R, Nguyen V, Ensari I, et al. Sexual identity differences in ideal cardiovascular health among cisgender adults in the all of us research program. Ann Behav Med 2025;59(1):kaaf032. [doi: 10.1093/abm/kaaf032] [Medline: 40338046]
16. Sharma Y, Bhargava A, Doan D, Caceres BA. Examination of sexual identity differences in the prevalence of hypertension and antihypertensive medication use among US adults: findings from the behavioral risk factor surveillance system. Circ Cardiovasc Qual Outcomes 2022;15(12):e008999. [doi: 10.1161/circoutcomes.122.008999]

XSL•FO
RenderX

17. Streed Jr CG. Correction to: assessing and addressing cardiovascular health in people who are transgender and gender diverse: a scientific statement from the American heart association. Circulation 2021;144(6):e136-e148. [doi: 10.1161/cir.0000000000001009]
18. Frost DM, Meyer IH. Minority stress theory: application, critique, and continued relevance. Curr Opin Psychol 2023;51:101579 [FREE Full text] [doi: 10.1016/j.copsyc.2023.101579] [Medline: 37270877]
19. Meyer IH. Minority stress and mental health in gay men. J Health Soc Behav 1995;36(1):38. [doi: 10.2307/2137286]
20. Meyer IH. Prejudice, social stress, and mental health in lesbian, gay, and bisexual populations: conceptual issues and research evidence. Psychol Bull 2003;129(5):674-697 [FREE Full text] [doi: 10.1037/0033-2909.129.5.674] [Medline: 12956539]
21. Caceres BA, Streed CG. Cardiovascular health concerns in sexual and gender minority populations. Nat Rev Cardiol 2021;18(4):227-228 [FREE Full text] [doi: 10.1038/s41569-021-00518-3] [Medline: 33510362]
22. Streed CG, Duncan MS, Heier KR, Workman TE, Beach LB, Caceres BA, et al. Prevalent atherosclerotic cardiovascular disease among veterans by sexual orientation. J Am Heart Assoc 2024;13(22):e036898. [doi: 10.1161/jaha.124.036898]
23. Defreyne J, Van de Bruaene LDL, Rietzschel E, Van Schuylenbergh J, T'Sjoen GGR. Effects of gender-affirming hormones on lipid, metabolic, and cardiac surrogate blood markers in transgender persons. Clin Chem 2019;65(1):119-134. [doi: 10.1373/clinchem.2018.288241] [Medline: 30602477]
24. Mereish EH, Goldstein CM. Minority stress and cardiovascular disease risk among sexual minorities: mediating effects of sense of mastery. Int J Behav Med 2020;27(6):726-736 [FREE Full text] [doi: 10.1007/s12529-020-09919-z] [Medline: 32734472]
25. Waxenbaum JA, Reddy V, Varacallo MA. Anatomy, autonomic nervous system. In: StatPearls. Treasure Island: StatPearls Publishing; 2023.
26. Shah AS, Lampert R, Goldberg J, Bremner JD, Li L, Thames MD, et al. Alterations in heart rate variability are associated with abnormal myocardial perfusion. Int J Cardiol 2020;305:99-105 [FREE Full text] [doi: 10.1016/j.ijcard.2020.01.069] [Medline: 32024598]
27. McEwen BS. Protective and damaging effects of stress mediators. N Engl J Med 1998;338(3):171-179. [doi: 10.1056/nejm199801153380307]
28. Thayer JF, Ahs F, Fredrikson M, Sollers JJ, Wager TD. A meta-analysis of heart rate variability and neuroimaging studies: implications for heart rate variability as a marker of stress and health. Neurosci Biobehav Rev 2012;36(2):747-756. [doi: 10.1016/j.neubiorev.2011.11.009] [Medline: 22178086]
29. Kim HG, Cheon E, Bai D, Lee YH, Koo B. Stress and heart rate variability: A meta-analysis and review of the literature. Psychiatry Investig 2018;15(3):235-245 [FREE Full text] [doi: 10.30773/pi.2017.08.17] [Medline: 29486547]
30. Villareal RP, Liu BC, Massumi A. Heart rate variability and cardiovascular mortality. Curr Atheroscler Rep 2002;4(2):120-127. [doi: 10.1007/s11883-002-0035-1] [Medline: 11822975]
31. McEwen BS. Protective and damaging effects of stress mediators: central role of the brain. Dialogues Clin Neurosci 2022;8(4):367-381. [doi: 10.31887/dcns.2006.8.4/bmcewen]
32. Kubota Y, Chen LY, Whitsel EA, Folsom AR. Heart rate variability and lifetime risk of cardiovascular disease: the atherosclerosis risk in communities study. Ann Epidemiol 2017;27(10):619-625.e2 [FREE Full text] [doi: 10.1016/j.annepidem.2017.08.024] [Medline: 29033120]
33. Kivimäki M, Steptoe A. Effects of stress on the development and progression of cardiovascular disease. Nat Rev Cardiol 2018;15(4):215-229. [doi: 10.1038/nrcardio.2017.189] [Medline: 29213140]
34. Hillebrand S, Gast KB, de Mutsert R, Swenne CA, Jukema JW, Middeldorp S, et al. Heart rate variability and first cardiovascular event in populations without known cardiovascular disease: meta-analysis and dose-response meta-regression. Europace 2013;15(5):742-749. [doi: 10.1093/europace/eus341] [Medline: 23370966]
35. Fang SC, Wu YL, Tsai PS. Heart rate variability and risk of all-cause death and cardiovascular events in patients with cardiovascular disease: a meta-analysis of cohort studies. Biol Res Nurs 2020;22(1):45-56. [doi: 10.1177/1099800419877442] [Medline: 31558032]
36. Huebner DM, McGarrity LA, Perry NS, Spivey LA, Smith TW. Cardiovascular and cortisol responses to experimentally-induced minority stress. Health Psychol 2021;40(5):316-325. [doi: 10.1037/hea0001067] [Medline: 34152785]
37. Rosati F, Williams DP, Juster R, Thayer JF, Ottaviani C, Baiocco R. The cardiovascular conundrum in ethnic and sexual minorities: a potential biomarker of constant coping with discrimination. Front Neurosci 2021;15:619171 [FREE Full text] [doi: 10.3389/fnins.2021.619171] [Medline: 34093107]
38. Brownlow BN, Williams DP, Kapuku G, Vasey MW, Anderson NB, Koenig J, et al. Ethnic differences in resting total peripheral resistance: a systematic review and meta-analysis. Psychosom Med 2020;82(6):548-560. [doi: 10.1097/psy.0000000000000820]
39. Hill LK. Ethnic differences in resting heart rate variability: a systematic review and meta-analysis. Psychosom Med 2015;77(1):16-25. [doi: 10.1097/psy.0000000000000133]
40. Altini M, Plews D. What is behind changes in resting heart rate and heart rate variability? a large-scale analysis of longitudinal measurements acquired in free-living. Sensors (Basel) 2021;21(23):7932 [FREE Full text] [doi: 10.3390/s21237932] [Medline: 34883936]

XSL•FO
RenderX

41. Föhr T, Tolvanen A, Myllymäki T, Järvelä-Reijonen E, Rantala S, Korpela R, et al. Subjective stress, objective heart rate variability-based stress, and recovery on workdays among overweight and psychologically distressed individuals: a cross-sectional study. J Occup Med Toxicol 2015;10(1):39 [FREE Full text] [doi: 10.1186/s12995-015-0081-6] [Medline: 26504485]
42. Ranganathan LB, Rajasundaram A, Kumar SKS. A cross-sectional study on the effect of stress on short-term heart rate variability and muscle strength among construction site workers. Int J Exp Res Rev 2024;44:1-10. [doi: 10.52756/ijerr.2024.v44spl.001]
43. Haag K, Hiller R, Peyk P, Michael T, Meiser-Stedman R, Fearon P, et al. A longitudinal examination of heart-rate and heart rate variability as risk markers for child posttraumatic stress symptoms in an acute injury sample. J Abnorm Child Psychol 2019;47(11):1811-1820 [FREE Full text] [doi: 10.1007/s10802-019-00553-2] [Medline: 31073881]
44. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol 2008;4(1):1-32. [doi: 10.1146/annurev.clinpsy.3.022806.091415] [Medline: 18509902]
45. Akbar F, Mark G, Prausnitz S, Warton EM, East JA, Moeller MF, et al. Physician stress during electronic health record inbox work: in situ measurement with wearable sensors. JMIR Med Inform 2021;9(4):e24014 [FREE Full text] [doi: 10.2196/24014] [Medline: 33908888]
46. Kim J, Foo JC, Murata T, Togo F. Reduced heart rate variability is related to fluctuations in psychological stress levels in daily life. Stress Health 2024;40(5):e3447. [doi: 10.1002/smi.3447] [Medline: 39032150]
47. Martinez GJ, Grover T, Mattingly SM, Mark G, D'Mello S, Aledavood T, et al. Alignment between heart rate variability from fitness trackers and perceived stress: perspectives from a large-scale in situ longitudinal study of information workers. JMIR Hum Factors 2022;9(3):e33754 [FREE Full text] [doi: 10.2196/33754] [Medline: 35925662]
48. Tazarv A. Personalized stress monitoring using wearable sensors in everyday settings. Annu Int Conf IEEE Eng Med Biol Soc 2021;2021:7332-7335. [doi: 10.1109/embc46164.2021.9630224]
49. Yao L, Yang Y, Wang Z, Pan X, Xu L. Compliance with ecological momentary assessment programmes in the elderly: a systematic review and meta-analysis. BMJ Open 2023;13(7):e069523 [FREE Full text] [doi: 10.1136/bmjopen-2022-069523] [Medline: 37438069]
50. Swann G, Bettin E, Ryan DT, Clifford A, Newcomb ME, Whitton SW, et al. The sexual orientation microaggression inventory short form (SOMI-SF): validation in three samples of racially/ethnically diverse sexual minority youth. Sex Res Social Policy 2023;20(2):589-598 [FREE Full text] [doi: 10.1007/s13178-022-00695-x] [Medline: 37637938]
51. Tourangeau R, Rasinski KA. Cognitive processes underlying context effects in attitude measurement. Psychol Bull 1988;103(3):299-314. [doi: 10.1037//0033-2909.103.3.299]
52. Thompson ER. Development and validation of an internationally reliable short-form of the positive and negative affect schedule (PANAS). J Cross-Cult Psychol 2007;38(2):227-242. [doi: 10.1177/0022022106297301]
53. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
54. Spitzer RL, Kroenke K, Williams JB. Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. primary care evaluation of mental disorders. patient health questionnaire. JAMA 1999;282(18):1737-1744. [doi: 10.1001/jama.282.18.1737] [Medline: 10568646]
55. Williams DR, Yan Y, Jackson JS, Anderson NB. Racial differences in physical and mental health: socio-economic status, stress and discrimination. J Health Psychol 1997;2(3):335-351 [FREE Full text] [doi: 10.1177/135910539700200305] [Medline: 22013026]
56. Balsam KF, Beadnell B, Molina Y. The daily heterosexist experiences questionnaire: measuring minority stress among lesbian, gay, bisexual, and transgender adults. Meas Eval Couns Dev 2013;46(1):3-25 [FREE Full text] [doi: 10.1177/0748175612449743] [Medline: 24058262]
57. Schrager SM, Goldbach JT, Mamey MR. Development of the sexual minority adolescent stress inventory. Front Psychol 2018;9:319 [FREE Full text] [doi: 10.3389/fpsyg.2018.00319] [Medline: 29599737]
58. Meidlinger PC, Hope DA. Differentiating disclosure and concealment in measurement of outness for sexual minorities: the Nebraska outness scale. Psychol Sex Orientat Gend Divers 2014;1(4):489-497. [doi: 10.1037/sgd0000080]
59. Change SOB. Ecological momentary assessment of stressful events. Cognitive Atlas. URL: https://www.cognitiveatlas.org/task/id/tsk_c6667oSpAnrBT/ [accessed 2025-09-13]
60. Park SH, Petrunoff NA, Wang NX, van Dam RM, Sia A, Tan CS, et al. Daily park use, physical activity, and psychological stress: a study using smartphone-based ecological momentary assessment amongst a multi-ethnic Asian cohort. Ment Health Phys Act 2022;22:100440. [doi: 10.1016/j.mhpa.2022.100440]
61. Warttig SL, Forshaw MJ, South J, White AK. New, normative, English-sample data for the short form perceived stress scale (PSS-4). J Health Psychol 2013;18(12):1617-1628. [doi: 10.1177/1359105313508346] [Medline: 24155195]
62. Collins RL, Parks GA, Marlatt GA. Social determinants of alcohol consumption: the effects of social interaction and model status on the self-administration of alcohol. J Consult Clin Psychol 1985;53(2):189-200. [doi: 10.1037//0022-006x.53.2.189] [Medline: 3998247]

XSL•FO
RenderX

63. Patrick ME, Schulenberg JE, O'Malley PM, Maggs JL, Kloska DD, Johnston LD, et al. Age-related changes in reasons for using alcohol and marijuana from ages 18 to 30 in a national sample. Psychol Addict Behav 2011;25(2):330-339 [FREE Full text] [doi: 10.1037/a0022445] [Medline: 21417516]
64. Lee CM, Neighbors C, Woods BA. Marijuana motives: young adults' reasons for using marijuana. Addict Behav 2007;32(7):1384-1394 [FREE Full text] [doi: 10.1016/j.addbeh.2006.09.010] [Medline: 17097817]
65. Stangl FJ, Riedl R. Measurement of heart rate and heart rate variability with wearable devices: a systematic review. In: Wirtschaftsinformatik 2022 Proceedings. 2022 Presented at: Wirtschaftsinformatik 2022; 2022-02-21; Nuremburg (online) p. 15.
66. Tung I, Balaji U, Hipwell AE, Low CA, Smyth JM. Feasibility and acceptability of measuring prenatal stress in daily life using smartphone-based ecological momentary assessment and wearable physiological monitors. J Behav Med 2024;47(4):635-646. [doi: 10.1007/s10865-024-00484-4] [Medline: 38581594]
67. Plews DJ, Scott B, Altini M, Wood M, Kilding AE, Laursen PB. Comparison of heart-rate-variability recording with smartphone photoplethysmography, polar H7 chest strap, and electrocardiography. Int J Sports Physiol Perform 2017;12(10):1324-1328. [doi: 10.1123/ijspp.2016-0668] [Medline: 28290720]
68. Hsu F, Huang HP. Heart rate accuracy from wearable wristbands. Labfront. 2022. URL: https://www.labfront.com/article/heart-rate-accuracy-wearables [accessed 2022-08-01]
69. Garmin Ltd.. Garmin Vivosmart 4 wearable device. Garmin. URL: https://www.garmin.com/en-US/p/605739/ [accessed 2020-12-01]
70. ilumivu. mEMA app mobile app. ilumivu. URL: https://ilumivu.com/solutions/ecological-momentary-assessment-app/ [accessed 2020-12-01]
71. Wrzus C, Neubauer AB. Ecological momentary assessment: a meta-analysis on designs, samples, and compliance across research fields. Assessment 2023;30(3):825-846 [FREE Full text] [doi: 10.1177/10731911211067538] [Medline: 35016567]
72. Remind I. Remind: school communication mobile app. 2022. URL: https://www.remind.com/ [accessed 2025-09-13]
73. Businelle MS, Hébert ET, Shi D, Benson L, Kezbers KM, Tonkin S, et al. Investigating best practices for ecological momentary assessment: nationwide factorial experiment. J Med Internet Res 2024;26:e50275 [FREE Full text] [doi: 10.2196/50275] [Medline: 39133915]
74. Posner K, Brown GK, Stanley B, Brent DA, Yershova KV, Oquendo MA, et al. The Columbia-suicide severity rating scale: initial validity and internal consistency findings from three multisite studies with adolescents and adults. Am J Psychiatry 2011;168(12):1266-1277 [FREE Full text] [doi: 10.1176/appi.ajp.2011.10111704] [Medline: 22193671]
75. Stanley B, Brown GK. Safety planning guide: quick guide for clinicians. Suicide Prevention Resource Center. 2008. URL: https://sprc.org/wp-content/uploads/2023/01/SafetyPlanningGuide-Quick-Guide-for-Clinicians.pdf [accessed 2022-04-30]
76. Stanley B, Brown GK. Safety plan template. Suicide Prevention Resource Center. 2008. URL: https://sprc.org/resources/stanley-brown-safety-plan/ [accessed 2025-09-27]
77. Davanzo A, D Huart D, Seker S, Moessner M, Zimmermann R, Schmeck K, et al. Study features and response compliance in ecological momentary assessment research in borderline personality disorder: systematic review and meta-analysis. J Med Internet Res 2023;25:e44853 [FREE Full text] [doi: 10.2196/44853] [Medline: 36920466]
78. Wen CKF, Schneider S, Stone AA, Spruijt-Metz D. Compliance with mobile ecological momentary assessment protocols in children and adolescents: a systematic review and meta-analysis. J Med Internet Res 2017;19(4):e132 [FREE Full text] [doi: 10.2196/jmir.6641] [Medline: 28446418]
79. Maher JP, Rebar AL, Dunton GF. Ecological momentary assessment is a feasible and valid methodological tool to measure older adults' physical activity and sedentary behavior. Front Psychol 2018;9:1485 [FREE Full text] [doi: 10.3389/fpsyg.2018.01485] [Medline: 30158891]
80. Allen J. Photoplethysmography and its application in clinical physiological measurement. Physiol Meas 2007;28(3):R1-R39. [doi: 10.1088/0967-3334/28/3/r01]
81. Alqaraawi A, Alwosheel A, Alasaad A. Heart rate variability estimation in photoplethysmography signals using Bayesian learning approach. Healthc Technol Lett 2016;3(2):136-142 [FREE Full text] [doi: 10.1049/htl.2016.0006] [Medline: 27382483]
82. Singh N, Moneghetti KJ, Christle JW, Hadley D, Plews D, Froelicher V. Heart rate variability: an old metric with new meaning in the era of using mHealth technologies for health and exercise training guidance. part one: physiology and methods. Arrhythm Electrophysiol Rev 2018;7(3):193-198 [FREE Full text] [doi: 10.15420/aer.2018.27.2] [Medline: 30416733]
83. Gambassi BB, Neves VR, Brito EZA, da Silva Fernandes DS, Sá CA, da Rocha Nogueira RM, et al. A validation study of a smartphone application for heart rate variability assessment in asymptomatic adults. Am J Cardiovasc Dis 2020;10(3):219-229 [FREE Full text] [Medline: 32923104]
84. Bent B, Enache OM, Goldstein B, Kibbe W, Dunn JP. Reply: matters arising 'Investigating sources of inaccuracy in wearable optical heart rate sensors'. NPJ Digit Med 2021;4(1):39 [FREE Full text] [doi: 10.1038/s41746-021-00409-4] [Medline: 33637842]
85. Ismail S, Akram U, Siddiqi I. Heart rate tracking in photoplethysmography signals affected by motion artifacts: a review. EURASIP J Adv Signal Process 2021;2021(1):1-19. [doi: 10.1186/s13634-020-00714-2]

XSL•FO
RenderX

86. Pollreisz D, TaheriNejad N. Detection and removal of motion artifacts in PPG signals. Mobile Netw Appl 2019;27(2):728-738. [doi: 10.1007/s11036-019-01323-6]
87. Chow HW, Yang CC. Accuracy of optical heart rate sensing technology in wearable fitness trackers for young and older adults: validation and comparison study. JMIR Mhealth Uhealth 2020;8(4):e14707 [FREE Full text] [doi: 10.2196/14707] [Medline: 32343255]
88. Fuller D, Colwell E, Low J, Orychock K, Tobin MA, Simango B, et al. Reliability and validity of commercially available wearable devices for measuring steps, energy expenditure, and heart rate: systematic review. JMIR Mhealth Uhealth 2020;8(9):e18694 [FREE Full text] [doi: 10.2196/18694] [Medline: 32897239]
89. Shcherbina A, Mattsson C, Waggott D, Salisbury H, Christle J, Hastie T, et al. Accuracy in wrist-worn, sensor-based measurements of heart rate and energy expenditure in a diverse cohort. J Pers Med 2017;7(2):3 [FREE Full text] [doi: 10.3390/jpm7020003] [Medline: 28538708]
90. Liao Y, Skelton K, Dunton G, Bruening M. A systematic review of methods and procedures used in ecological momentary assessments of diet and physical activity research in youth: an adapted STROBE checklist for reporting EMA studies (CREMAS). J Med Internet Res 2016;18(6):e151 [FREE Full text] [doi: 10.2196/jmir.4954] [Medline: 27328833]
91. MathWorks. MATLAB. MathWorks. URL: https://www.mathworks.com/products/matlab.html [accessed 2021-08-01]

Abbreviations

CVD: cardiovascular disease
CVH: cardiovascular health
EMA: ecological momentary assessment
EOD: end-of-day
HIPAA: Health Insurance Portability and Accountability Act
HRV: heart rate variability
IRB: internal review board
LGB: lesbian, gay, and bisexual
LGBTQ+: lesbian, gay, bisexual, transgender, queer, and other diverse sexual and gender identities
NH: non-Hispanic
PPG: photoplethysmography
SGM: sexual and gender minority
TPR: total peripheral resistance

Edited by A Mavragani; submitted 27.11.24; peer-reviewed by K Heron; comments to author 13.04.25; revised version received 31.07.25; accepted 28.08.25; published 15.10.25.

Please cite as:
Urueta Tapia D, Corliss HL, Lee KH, Calzo JP, Jun HJ
Assessing Minority Stress and Physiological Response Through Ecological Momentary Assessment and Sensors: Protocol for a Feasibility and Acceptability of the Stress and Heart Pilot Study
JMIR Form Res 2025;9:e68733
URL: https://formative.jmir.org/2025/1/e68733
doi:10.2196/68733
PMID:

©Dulce Urueta Tapia, Heather L Corliss, Kang Hyuk Lee, Jerel P Calzo, Hee-Jin Jun. Originally published in JMIR Formative Research (https://formative.jmir.org), 15.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Developing a Quick Isolation Bed Inquiry System During the COVID-19 Outbreak: User-Centered Design Approach Based on the Toyota Production System

Chien-Chung Lin[1*], MD; Jian-Hong Shen[2*], PhD; Shan-Li Chang[3*], BSc; Tai-Chih Kuo[4*], PhD; Hui-Ling Huang[5*], BSN; Cheng-Yi Peter Lin[6*], MBA; Hung-Meng Huang[7*], MD

[1]Department of Surgery and Orthopedic Surgery, Taipei City Hospital, University of Taipei, Taipei City, Taiwan
[2]Department of Finance, Chihlee University of Technology, New Taipei City, Taiwan
[3]Department of Energy and Refrigerating Air-Conditioning Engineering, National Taipei University of Technology, Taipei City, Taiwan
[4]Department of Biochemistry & Molecular Cell Biology, School of Medicine, Taipei Medical University, Taipei, Taiwan
[5]Department of Orthopedic Surgery, Taipei City Hospital, Taipei City, Taiwan
[6]Institute of Hospital & Health Care Administration, National Yang Ming Chiao Tung University, Taipei City, Taiwan
[7]Department of Otorhinolaryngology, Taipei City Hospital, Taipei City, Taiwan
[*]all authors contributed equally

Corresponding Author:
Chien-Chung Lin, MD
Department of Surgery and Orthopedic Surgery
Taipei City Hospital
University of Taipei
8F, Number 78, Section 1, Roosevelt Road
Zhongzheng District
Taipei City, 100
Taiwan
Phone: 886 912663169
Fax: 886 223567538
Email: ericdoctor@gmail.com

Abstract

Background: During the COVID-19 outbreak in May 2021, our hospital—designated as a specialized facility for severely infected patients—faced critical staff and resource shortages. The urgent need for efficient bed management to ensure timely admissions underscored the inefficiencies of the manual, phone-based allocation process, which averaged 454 seconds per query. Traditional IT solutions were not feasible due to time and cost constraints.

Objective: This study aims to design and implement a rapid, zero-cost Quick Isolation Bed Inquiry System that provides real-time bed information and enables timely admissions without requiring additional workforce or expense.

Methods: We conducted a 3-cycle pre-post quality improvement study guided by the Toyota business practice (TBP), an 8-step problem-solving framework. After clarifying the problem and constructing a value stream map, we identified bottlenecks. A user-centered solution was developed by leveraging an underutilized data export function in the hospital's bed inquiry platform. Using Microsoft Excel Visual Basic for Applications, we automated the filtering and display of relevant bed information. The primary outcomes were process time and number of steps; secondary outcomes included staff time savings and system accuracy.

Results: The baseline manual process required 25 steps and 454 seconds to complete a query. The new system reduced this to 3 steps and 12 seconds, representing a 97.4% gain in efficiency. Single-click execution generated 3 outputs (administrative PDF, large-screen display, and mobile version) in 4 seconds, with distribution to the hospital LINE group completed in 7 seconds. Reliability reached 100%, with continuous availability through virtual private network access. Development and debugging were completed within 3 days using only existing resources. Postpandemic, the system was adapted for general ward management with minimal modifications.

Conclusions: Applying TBP enabled the rapid development of a user-centered, zero-cost bed management tool by repurposing existing digital assets. The intervention markedly improved efficiency, reliability, and usability without additional staffing or

XSL•FO
RenderX

expenditure, providing a scalable model for agile health care systems operating under resource constraints. Future work will focus on deeper automation, such as application programming interface–based real-time updates, and on evaluating downstream impacts on patient flow and bed turnover.

(JMIR Form Res 2025;9:e67152) doi:10.2196/67152

KEYWORDS

autonomation; continuous improvement; information technology; quality improvement; health care IT; just-in-time; lean; problem solving; PDCA; Toyota business practice; Toyota production system; value stream map

Introduction

Problem Description

As of October 2020, over 37 million people worldwide had been diagnosed with COVID-19, and more than 1 million had died, reflecting a case-fatality rate of about 2.9% [1]. The pandemic's effects were especially severe in the United States, which accounted for roughly one-fifth of global infections and deaths [1]. It also exposed major weaknesses in hospital bed-management systems and highlighted that timely patient allocation was crucial for survival during surge periods [2,3].

This global crisis manifested locally in Taiwan in May 2021, when the country experienced a severe COVID-19 outbreak. In Taipei City, cases rose sharply from 27 in April to 2745 in May [4]. Our hospital, Taipei City Hospital Heping Branch, was designated as the capital region's primary COVID-19 treatment facility, managing 148 isolation beds for patients with COVID-19. This designation created an immediate surge in admission demand while the hospital simultaneously faced critical staffing and resource shortages.

The most urgent challenge was managing bed availability and rapidly assigning beds to patients, as mandated by the National Disease Control Bureau. Under normal conditions, bed allocation was managed manually by the Medical Affairs Office. The pandemic, however, exposed major inefficiencies in this approach. We lacked both sufficient personnel and a clear system to provide real-time updates on bed availability to hospital management. Bed control physicians, responsible for assignments, had to rely on a manual, time-consuming process of calling each ward individually to check availability. This method was inefficient and risked compromising patient safety and treatment quality due to delays.

Existing Knowledge

A comprehensive literature review revealed limited research on rapid bed allocation methods or real-time bed assignment platforms, particularly in emergency contexts [5-11]. The identified research gap included both the absence of rapid implementation frameworks for bed management systems during health care crises and the lack of real-time visibility tools capable of functioning under staffing shortages and high-stress conditions. Although commercial electronic bed management systems offer features such as real-time monitoring, predictive analytics, and integration capabilities [12-18], these typically require long implementation timelines and substantial resource investments, making them unsuitable for emergency deployment.

This gap reflects a broader challenge in health care IT: how to rapidly develop and deploy digital solutions during crisis conditions when traditional development cycles are not feasible [19-21]. The research question guiding this study is as follows: "Can structured problem-solving methodologies from manufacturing be effectively adapted to develop health care IT solutions under extreme time and resource constraints?"

Rationale

With no dedicated personnel or prior models to follow, a solution was required that could be developed rapidly using existing resources. In the absence of alternative theoretical frameworks in the scientific literature for rapid health care IT implementation during crises, we turned to the Toyota production system (TPS). We selected the Toyota business practice (TBP) framework of the TPS—a systematic 8-step problem-solving methodology—for 3 evidence-based reasons: (1) demonstrated success in health care quality improvement across various processes [22-27]; (2) emphasis on leveraging existing resources rather than requiring new investments [27-36]; and (3) proven applicability in our institution's previous digital health projects [30].

Objective

Therefore, the objective of this study is to provide a detailed case report on applying the structured 8-step TBP framework to rapidly design, develop, and implement a low-cost, user-centered bed inquiry system using only existing hospital resources. This study documents the iterative process of problem identification, root cause analysis, and solution development, offering a practical model for agile problem solving in health care crises.

Methods

Study Design

This quality improvement implementation study employed a pre-post design using TBP's structured problem-solving methodology across 3 iterative cycles. The study was conducted in accordance with implementation research principles and the SQUIRE (Standards for Quality Improvement Reporting Excellence) 2.0 guidelines, with particular emphasis on real-world effectiveness and the timely deployment of solutions [37-39].

Setting and Context

The study was conducted at Taipei City Hospital Heping Branch, a 411-bed public hospital designated as the primary COVID-19 treatment facility for Taipei City. During the study period, the

XSL•FO
RenderX

hospital managed 148 isolation beds across 4 wards (A9, A8, A7, and A6), with complex occupancy rules requiring gender-matched double rooms.

Methodological Framework: The Toyota Business Practice

The study was structured around the TBP, a formal 8-step problem-solving methodology derived from the core principles of the TPS [31-35]. This framework, also widely referred to in TPS literature as the A3 problem-solving process, provides a systematic and disciplined approach to identifying, analyzing, and resolving operational challenges [35,40]. The 8 steps are (1) clarify the problem, (2) break down the problem, (3) set a target, (4) analyze the root cause, (5) develop countermeasures, (6) implement countermeasures, (7) evaluate results and processes, and (8) standardize and share the successful approach [27,31,35].

Figure 1 outlines the TBP problem-solving process, highlighting the key elements at each stage.

Figure 1. Overview of the Toyota business practice problem-solving process, highlighting the key components of each step.

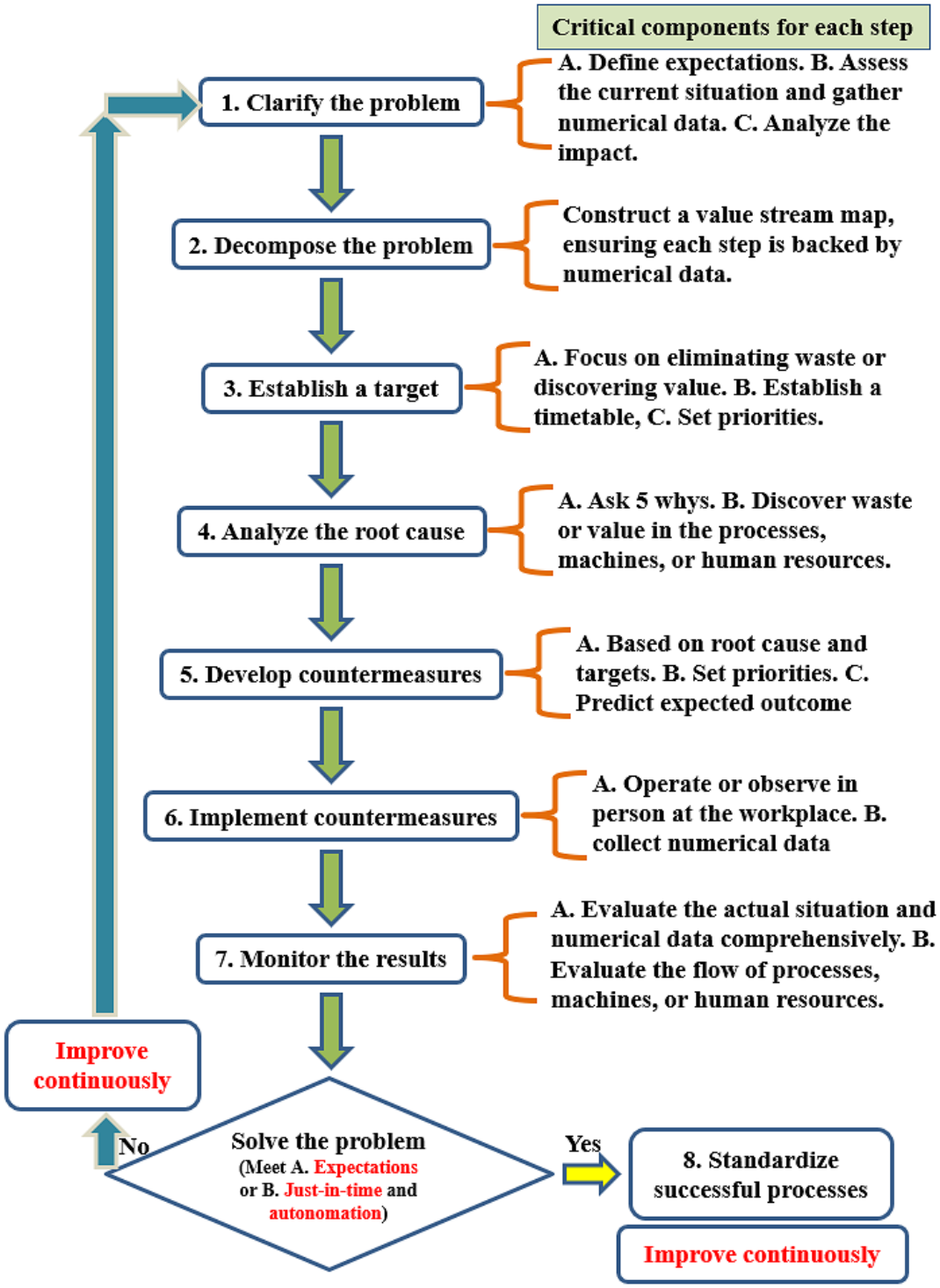

XSL•FO
RenderX

Ohno [32] consistently emphasized his strong belief in the concept of *gemba* (on-site work). Even after joining Toyota's top management, he continued to spend most of his time on the shop floor [32]. Consequently, it is fair to say that TPS is fundamentally centered on on-site management [32-35]. According to Ohno, work activities are classified based on their contribution to value into 3 categories: waste, non-value–added work (necessary but non-value–adding activities), and value-added work [32,33]. Waste refers to the use of resources without creating value. Non-value–added work, although necessary under current circumstances, does not generate value and is still considered a form of waste. Value-added work encompasses activities that directly enhance the original value [32,33]. TPS principles assert that a product or service is valuable only if it meets the needs and expectations of the user or customer [32-35]. Careful on-site observation is essential to clearly distinguish value-adding activities from wasteful actions.

To situate this methodology within the broader context of quality improvement science, it is important to recognize its structural alignment with the plan-do-check-act (PDCA) cycle, a foundational model for continuous improvement [35,40,41]. The TBP framework maps directly onto the PDCA cycle, demonstrating that it is a rigorous, cyclical quality improvement methodology rather than an ad hoc process [35,40]. The *plan* phase of PDCA is addressed by the first 5 steps of TBP, which emphasize deep problem understanding and planning. The *do*, *check*, and *act* phases correspond directly to TBP steps 6, 7, and 8, respectively. This alignment, highlighting the critical importance of the planning phase, reinforces TBP as a powerful tool for effective issue resolution. The focus on continuous improvement and the cyclical nature of TBP mirrors that of PDCA, enhancing the methodology's credibility and generalizability [35,40,41].

An Iterative, User-Centered Design Process

While TBP provided the overarching structure, the study's execution closely followed the principles of user-centered design [32,33,42,43]. User-centered design is a design philosophy and process that places the end user at the center of every development stage, focusing on their needs, workflows, and context to create tools that are not only functional but also highly usable and effective [42]. The 3 TBP cycles described in this paper can be formally interpreted as sequential phases of a comprehensive user-centered design process [32,33,42].

The study unfolded through the following user-centered design phases.

The First Cycle of TBP: Defining User Needs and Workflow Analysis

Step 1: Clarify the Problem

The initial cycle focused on empathy and user research, a cornerstone of user-centered design [42,43]. It involved direct on-site observation and collaboration with the primary users—the bed control physicians.

TPS Thinking 1

Problem identification begins with a need or expectation [32]. Once expectations are defined, direct observation and data collection at the workplace are essential to assess the current situation. A problem is identified when a gap exists between expectations and the current state [32,33].

Expectations

The hospital superintendent required immediate and accurate information on bed availability to ensure the prompt admission of patients with COVID-19. Bed control physicians—including the Vice Superintendent, Head of Internal Medicine, and Head of Surgery—were responsible for posting available bed information in the Critical Care Bed Coordination Group via the LINE messaging app (LY Corp) and for promptly assigning beds to patients.

Current Situation

After clarifying leadership's expectations, we followed TPS principles by conducting on-site observations and collecting real-time data, rather than relying on secondhand information or meetings. This direct assessment revealed that existing hospital procedures and systems were unable to provide timely bed availability information, highlighting a clear gap between expectations and actual performance.

Impact

Delays in bed availability information could compromise patient safety and treatment quality, making this a critical issue that must be addressed.

Step 2: Break Down the Problem

TPS Thinking 2

To break down the problem, it is essential to visit the workplace—either to perform tasks firsthand or to observe directly—while collecting relevant numerical data. This information is then used to create a value stream map (VSM) to visualize the problem [32,33,36].

Manual Process for Critical Care Bed Coordination

Initially, our hospital opened 2 wards (A9 and A8). Under emergency conditions, bed control physicians had to call each ward individually to check bed availability. As patient numbers increased, 3 wards (A9, A8, and A7) were opened within a week, eventually expanding to 4 wards (A9, A8, A7, and A6) with a total of 148 beds. Bed control physicians contacted nursing staff or clerks in each ward to count available beds, recorded the data, converted it into text, and posted it in the Critical Care Bed Coordination Group on the LINE messaging app.

It is important to note that hospital policy requires patients of the same gender (except couples) to share a double room in isolation wards. Therefore, when 1 bed in a double room is occupied, the patient's gender must be identified and recorded to ensure compliance when displaying the remaining vacant bed.

As we had no prior experience or information on querying bed availability, we conducted the search ourselves. Figure 2 illustrates that each step in the bed availability query process provided valuable insights for bed control physicians. Through hands-on operation and data collection, we gathered key numerical data, summarized in Multimedia Appendix 1. This

XSL•FO
RenderX

table highlights the shortest time required to retrieve available bed information, as observed by one author acting (C-CL) as a bed control physician. To address uncertainties in step 2, such as unanswered calls, we excluded non-value–added activities, including calls going unanswered, wait times exceeding 30 seconds, or the need to redial. Consequently, the data collected include only instances in which ward staff answered the call within 30 seconds. We also recorded the time required for ward staff to locate available beds on the TV screen and for the surgeon to document the vacant bed numbers.

Figure 2. Mapping the entire process of querying bed availability resulted in the creation of a value stream map.

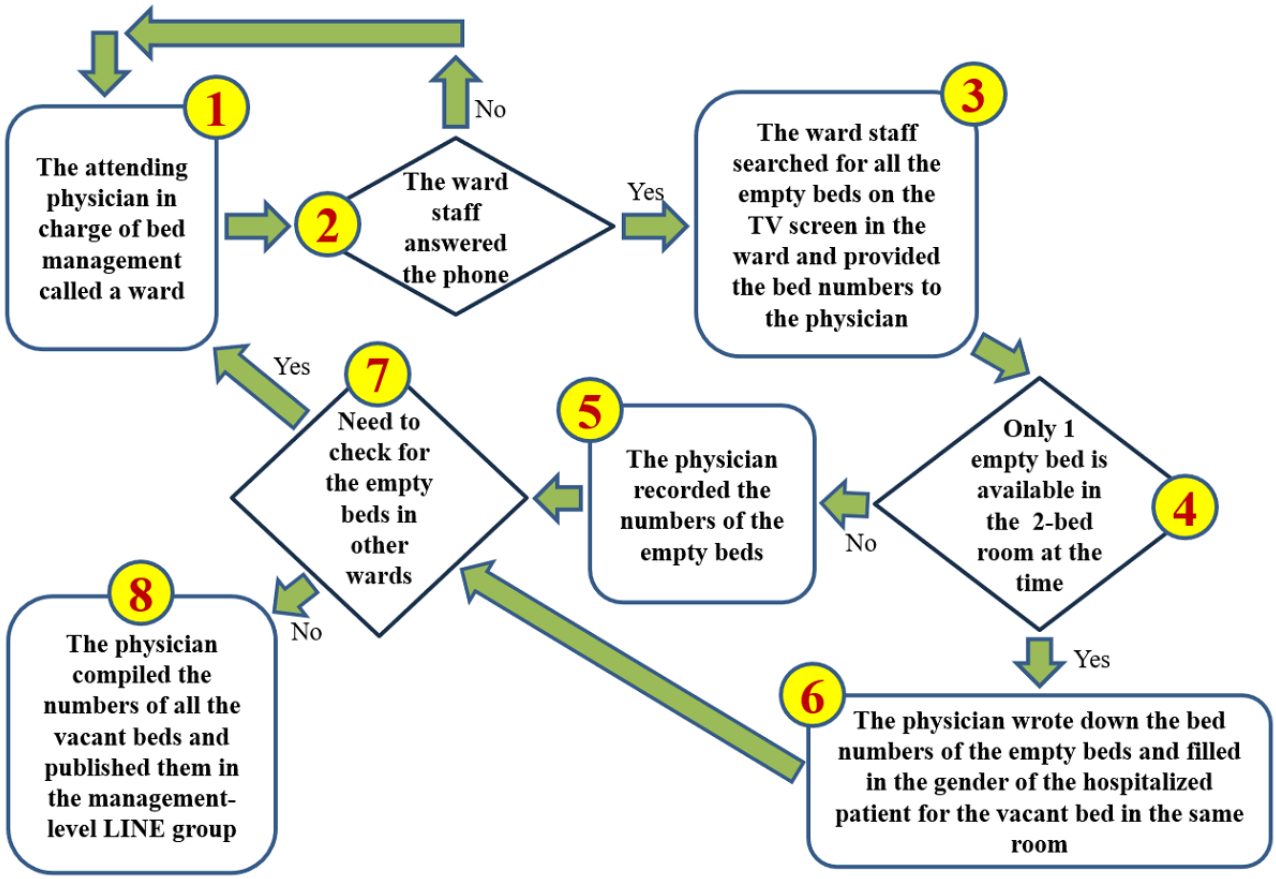

Multimedia Appendix 1 shows that the average shortest query time was 454 seconds (7 minutes and 34 seconds), which fell short of management's expectations for immediate bed availability to expedite patient assignments. To improve bed management, the hospital superintendent mandated that bed control physicians report bed availability to the Critical Care Bed Coordination Group at 8 AM, noon, 4 PM, and 8 PM daily. To comply with this directive, physicians had to begin checking available beds 20-30 minutes before each reporting time.

Our analysis, following step 1 (clarify the problem) and step 2 (break down the problem), clearly revealed why "current hospital procedures and systems were unable to provide timely bed availability information" relative to "leadership's expectations." Specifically, the hospital superintendent expected immediate and accurate information on bed availability to ensure the prompt admission of patients with COVID-19. Bed control physicians were responsible for posting available bed information in the Critical Care Bed Coordination Group via the LINE messaging app and for promptly assigning beds.

However, our on-site observations revealed a significant gap. The "current hospital procedures and systems were unable to provide timely bed availability information," primarily due to a manual, labor-intensive process involving the following steps: (1) Bed control physicians individually called each ward to check availability. (2) Ward staff manually searched for empty beds on their TV screens and verbally relayed the information. (3) Physicians recorded this information, converted it into text, and manually posted it. (4) For double rooms, the gender of existing patients had to be identified and recorded to ensure policy compliance for new admissions.

The extreme inefficiency of this manual process is highlighted in Multimedia Appendix 1, which shows that the average shortest query time for bed availability was 454 seconds (7 minutes and 34 seconds), even under ideal conditions (calls answered within 30 seconds). This fell far short of the "immediate" information expected by management. Moreover, the hospital superintendent mandated reports at 8 AM, noon, 4 PM, and 8 PM daily, requiring physicians to begin checking beds 20-30 minutes before each reporting time. This substantial preparation time underscores how the manual system failed to meet leadership's expectations for rapid, real-time updates. The process was not only slow but also prone to delays due to unanswered calls and staff busyness.

Attentive readers may note that many of our process evaluations were conducted using "best-case" or "shortest time" scenarios.

XSL•FO
RenderX

Multimedia Appendix 2 clarifies our methodology and explains the rationale for adopting these scenarios.

Step 3: Establish a Target

TPS Thinking 3

All problems arise from the existence of waste. Prioritize eliminating the most significant and impactful waste identified in the VSM, with a defined deadline for achieving the target [32,33].

Identified Issue and Improvement Goal

As shown in step 2 of Figure 2, the main issue arises from ward staff being too busy to answer calls, resulting in significant delays and inefficiencies. The goal was to reduce the number of phone calls taking more than 30 seconds to be answered by 50% within 3 days.

Step 4: Analyze the Root Cause

TPS Thinking 4

The root cause of operational problems is often attributable to waste within the process [32,33]. To identify this waste, TPS employs the "Five Whys" technique, which helps drill down to the root cause by repeatedly asking "why" [32,33].

Five Whys Analysis of Call Response Delays

As shown in Multimedia Appendix 3, the "Five Whys" method was applied to link the target with the VSM findings, enabling us to drill down into the key issue: "Why aren't calls answered promptly?" Different formulations of the question may lead to different root causes.

After applying the "Five Whys," it became clear that the root cause of the ward's inability to answer phone calls was a shortage of staff. Additionally, because priority was given to direct patient care, the head nurse could not ensure that someone in the ward would be available to answer the phone immediately.

Step 5: Develop Countermeasures

TPS Thinking 5

Countermeasures should be developed based on the root cause and goals, prioritizing those that address major issues or are simple to implement. Ideally, the problem solver should take direct action on-site [32,33].

Proposed Countermeasures and Implementation

To address the root cause and achieve our goals, we developed 2 countermeasures. First, assign ward staff to regularly check and report bed availability at 4 scheduled intervals or proactively notify available beds. Second, if no staff are available, the bed control physician should personally check bed availability at the nursing station. Successful implementation of the first countermeasure reduces the time physicians spend making follow-up calls and meets the hospital superintendent's expectations. During emergencies or when ward staff are unresponsive, the second countermeasure ensures immediate access to bed availability information. While implementing the first countermeasure, we simultaneously observed the situation and collected relevant numerical data.

Step 6: Implement Countermeasures and Observe at the Workplace

TPS Thinking 6

To verify results, it is necessary to return to the workplace where the problem originated, observe the implementation in practice, and collect numerical data [32,33].

Although the head nurse indicated that staff would be available at the nursing station to answer calls during the scheduled times, our implementation of the first countermeasure revealed a persistent issue identified in the VSM: frequent delays or failures of nursing and administrative staff in isolation wards to answer calls (Figure 2, step 2), particularly during evening shifts (data not shown). This forced bed control physicians to repeatedly call or contact other wards, creating a potential loop in steps 1 and 2. The delays were primarily due to reduced staffing during the evening shift (starting at 4:00 PM) and the absence of administrative personnel. Under these constraints, assigning staff specifically for bed availability checks was not feasible.

When the ward cannot answer the phone, the second countermeasure effectively locates available beds but has notable drawbacks. First, bed control physicians moving between nursing stations violates infection control protocols. Second, real-time updates on bed availability cannot be provided to administrators when changes occur. To address these issues, TPS principles were applied to optimize existing resources without increasing staff.

TPS Thinking 7

Prioritize process improvement and equipment optimization before considering an increase in workforce; adding staff should be the last resort [32,33].

To improve bed management efficiency, we conducted a detailed review of the VSM (Figure 2). We found that the primary method for checking bed availability involved ward staff searching for empty beds on the ward's TV screen (step 3). This indicated that the hospital's bed management system was already digitized, suggesting the presence of a database tracking bed occupancy and patient information.

Following this discovery, we consulted the Medical Affairs Office, which oversees admissions and discharges. They confirmed that the hospital information system (HIS) includes a bed inquiry platform containing essential inpatient data.

The Second Cycle of TBP: Usability Testing of Existing Digital Platform

Step 1: Clarify the Problem

Expectations

Bed control physicians expect the hospital bed inquiry platform to provide all information required by the authorities.

Current Situation

The hospital bed inquiry platform displays a menu (see Multimedia Appendix 4) and returns 761 bed records upon accessing the system. Each record includes 28 fields, such as bed number, patient name, medical record number, gender, admission date, discharge date, and attending physician. Of the

XSL•FO
RenderX

761 beds, 399 are for acute care, while the remaining beds comprise deactivated beds (n=215), postoperative recovery beds (n=6), dialysis beds (n=36), emergency/virtual observation beds, long-term respiratory care beds, psychiatric day care beds, and nursing home beds.

When reviewing available beds using the platform's built-in menu, 488 active beds are displayed, each with 28 fields. However, we only needed information for 148 beds, with just 3 relevant fields: bed number, medical record number (for occupancy), and gender. This demonstrates that the system provides unnecessary information.

The platform has a query function that allows bed availability checks by ward (see Multimedia Appendix 5). All combinations of the inquiry functions were tested. Currently, identifying double-occupancy rooms requires consulting a reference table. To find gender-specific beds, users must manually scan the 488 active beds, visually tracking the relevant data by moving the mouse pointer across the screen.

This inefficiency demonstrates that the current system does not meet physicians' expectations, highlighting a significant issue that needs to be addressed.

Impact

Delays in determining bed availability jeopardize patient safety and the quality of care.

Step 2: Break Down the Problem

Figure 3 illustrates the VSM, highlighting the most efficient method we identified for querying available beds hospital-wide. The times, measured in seconds, reflect the minimum duration required for a proficient physician to locate a single available bed. Steps E and H were the most time-consuming, taking 17 seconds to identify 1 bed. If multiple beds are available (eg, 10), the time required for steps E and H increases substantially.

Figure 3. The value stream map illustrates the built-in query function of the hospital bed inquiry platform, which is used to check bed availability across the hospital. The times shown (in seconds [s]) represent the fastest possible query duration for a physician proficient in computer operations, assuming only 1 available bed is found. These values indicate the minimum time required to complete the query process.

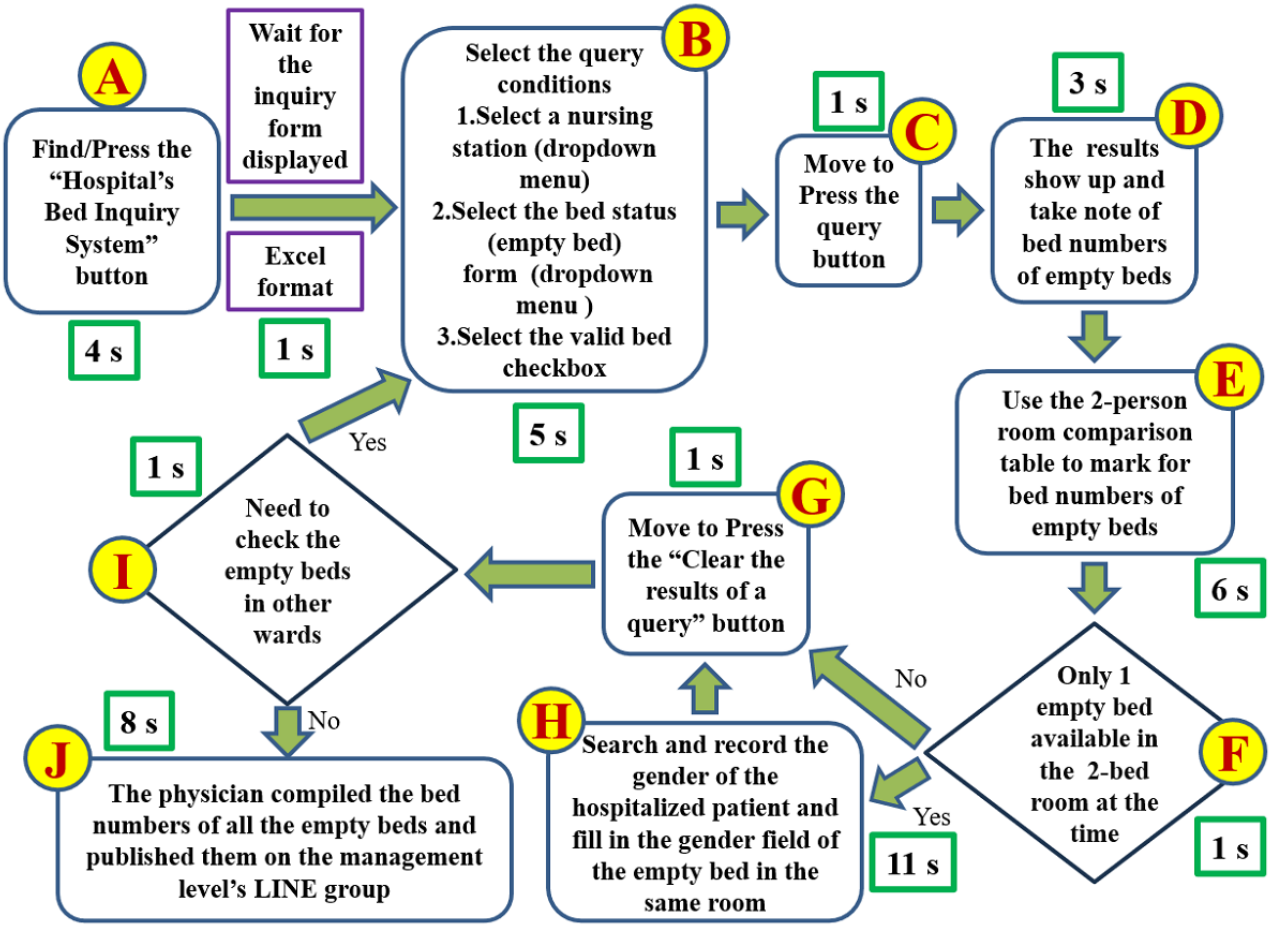

Step 3: Establish a Target

Our goal is to reduce the time spent on steps E and H by 50% within 3 days. The rationale is as follows: As shown in Figure 3, the VSM revealed that steps E and H were the most time-consuming elements in the existing hospital bed inquiry platform. Together, these steps took 17 seconds for a single bed, and the time increased substantially when multiple beds were available.

The rationale for targeting a 50% reduction in these steps was straightforward: they represented the major bottlenecks and sources of waste—manual effort, visual scanning, and reliance on external tables—within an already digitized but inefficient bed inquiry process. The goal was to make the system more

XSL•FO
RenderX

usable and efficient for physicians, moving closer to the "immediate and accurate information" expected by leadership.

A secondary subgoal was to reduce the total process time—from step A to step J—for a single bed in an isolation ward (42 seconds) by 50% within 6 days.

Step 4: Analyze the Root Cause

To identify double-occupancy rooms, we currently rely on a comparison table (Figure 3, step E). In collaboration with the Medical Affairs Office, we applied the "Five Whys" method to pinpoint the root cause (see Multimedia Appendix 6). The challenge arises because double-room designations depend on real-time factors, such as available equipment and ward conditions, preventing the establishment of a standardized guideline. Consequently, identifying double-occupancy rooms still requires referencing the double-room comparison table.

We applied the "Five Whys" method to determine the root cause of step H (data not shown). The root cause is that the hospital's information system cannot automatically identify the gender of the patient in the adjacent bed in double rooms.

Step 5: Develop Countermeasures

Based on the root cause analysis and established targets, 2 countermeasures were proposed. First, for step E, bed control physicians were tasked with memorizing the comparison table to expedite data comparison and potentially reduce processing time. Second, for step H, Medical Affairs Office staff were instructed to input the patient's gender into the bed management system during admission, resolving the gender mismatch issue in double-occupancy rooms. The second countermeasure involved enhancing physicians' proficiency in quickly searching bed numbers and patient gender in the system, further reducing task completion time. During implementation, we continuously monitored the process and collected relevant data to evaluate the effectiveness of these measures.

Step 6: Implement Countermeasures and Observe at the Workplace

Challenges in Using the Current Bed Management Process

After the actual operation, bed control physicians found the process excessively complicated. Checking the availability of a single isolation nursing station required at least 10 steps, involving multiple mouse clicks and selections, leading to eye and wrist strain. The method still relied on manually using the comparison table to determine and record gender, missing opportunities to leverage the computer system's advantages, such as speed, real-time processing, and automation. Even for experienced physicians, memorizing the double-room comparison table proved difficult. Consequently, no one opted to use these built-in functions for checking hospital-wide bed availability, highlighting the need for alternative solutions or improvements.

TPS Thinking 8

Problems arise in the workplace, and the solutions to these problems must also be found there [32].

The Third Cycle of TBP: Solution Development

Step 1: Clarify the Problem With the Hospital Bed Inquiry Platform's Built-In Features

Expectation

We expected to leverage the hospital's computer information system to enable faster, repeatable, and more immediate bed availability inquiries, addressing staff shortages and the urgent need for quick bed assignments.

Current Situation

The hospital bed inquiry platform lacked real-time bed availability features and administrative access, limiting its functionality to meet our needs. However, upon further examination, we discovered an export function (Multimedia Appendix 4) that generates a Microsoft Excel (.xls) file with 761 entries across 28 columns, referred to as the "Exported Raw Data .xls," as shown in Multimedia Appendix 7.

Step 2: Break Down the Problems in the Exported Raw Data .xls

Upon reviewing the exported data, we identified that most of the information was unnecessary. Of the 761 total entries, only 148 active bed numbers, provided by the Medical Affairs Office, were relevant. Similarly, of the 28 columns per entry, only 3 were pertinent: bed number, patient medical record number, and gender. The process for identifying vacant beds involved filtering the data down to these 148 entries and focusing on the 3 key columns. A bed was considered available if there was no associated medical record number. For 2-bed rooms, the gender of the admitted patient had to be considered, as only patients of the same gender can share a room. Figure 4 illustrates the VSM used to identify vacant beds from the exported raw data (.xls) across the hospital.

XSL•FO
RenderX

Figure 4. Value stream map for identifying vacant beds from the Exported Raw Data.xls file across the hospital. The strategy focuses on filtering out unnecessary information and retaining only the bed numbers of the 148 isolation beds, along with the patients' medical record numbers and gender. For double-occupancy rooms, if only 1 patient is present at the time of execution, the gender of the admitted patient must be considered when displaying the vacant bed number.

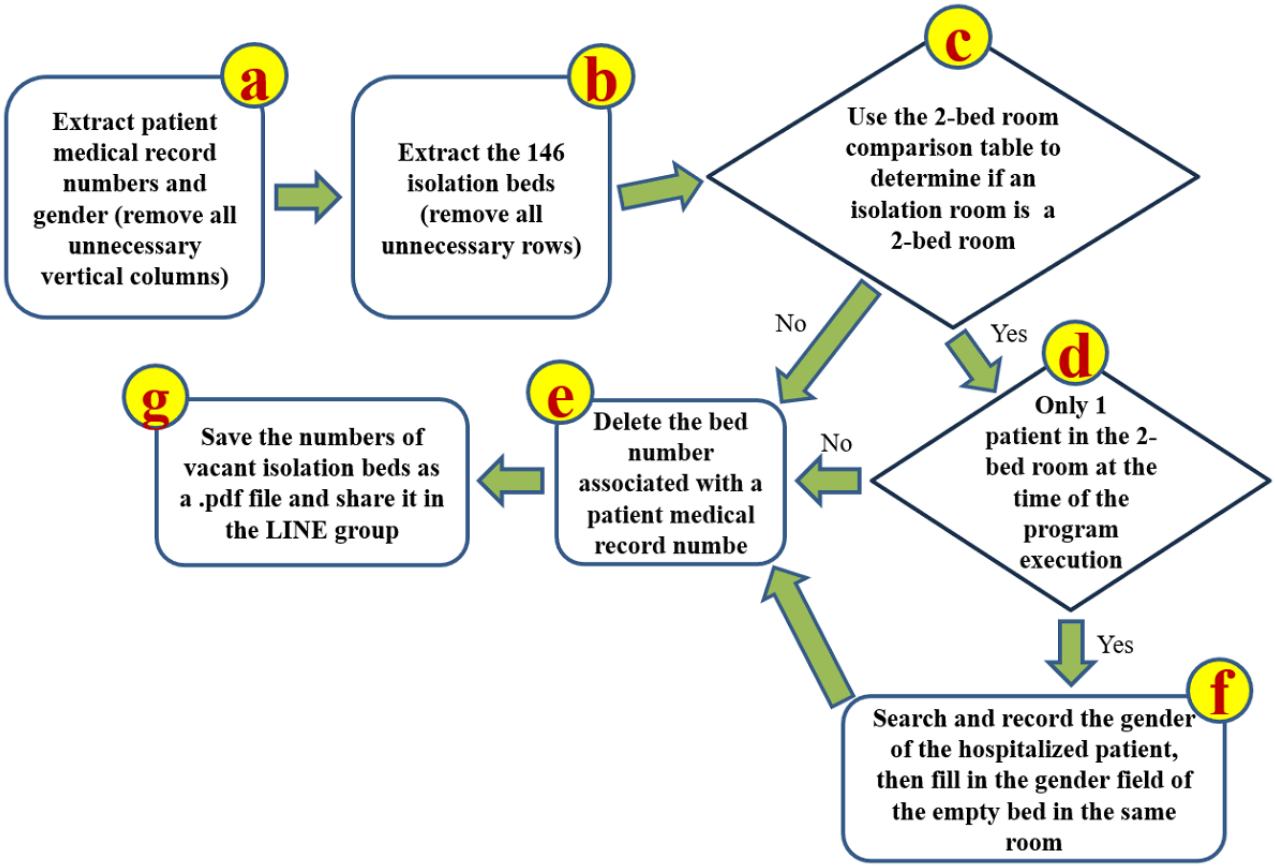

Step 3: Establish a Target

The goal is to develop a user-friendly Quick Isolation Bed Inquiry System within 3 days, leveraging data from the exported raw data (.xls) file. The system should allow anyone—not just physicians—to access real-time bed availability, enabling bed control physicians to quickly relay this information to hospital administrators.

Step 4: Analyze the Root Cause of Exported Raw Data .xls

The excess of irrelevant data in the exported file significantly hinders the efficient identification of available beds. Only the bed number, patient medical record number, and gender are considered essential and valuable for this process.

Steps 5 (Develop Countermeasures) and 6 (Implementing Countermeasures)

Based on the identified root cause and established targets, our countermeasure was clear: we needed to use the computer information system to eliminate unnecessary information. We explored 2 approaches: manual processing in Excel and automation using Visual Basic for Applications (VBA) to implement this plan through actionable steps.

The first countermeasure (manual processing in Microsoft Excel) involved the bed control physician manually handling the exported data file in Microsoft Excel. The steps were as follows:

- Delete the first 2 rows and unnecessary columns, retaining only the bed number, patient medical record number, and gender.
- Use the "Deactivated" column (cell N1) as a sorting header, then sort and remove all rows marked "Yes (Y)," leaving 488 usable beds.
- Filter these beds to retain only those from the A9, A8, A7, and A6 wards.
- Use the double-room comparison table to mark the gender of patients in double-occupancy rooms. If a room is partially occupied, the gender of the patient must be noted when displaying the vacant bed.
- Finally, sort by the medical record number column (cell I1) and remove all occupied beds. The remaining vacant beds are then recorded and shared in the LINE group.

The second countermeasure (automated process using VBA) utilized Microsoft Excel's VBA to automate the manual process. Drawing on prior experience developing a surgical scheduling system with VBA, we recognized the potential to streamline this task by creating a macro [30,44]. Before coding, we drafted an algorithm using the technique of constructing a VSM, converting the manual steps into a computer-executable workflow [44,45]. The algorithm in Figure 5 served as the blueprint for the VBA macro, automating the sorting, deletion, and comparison tasks of the first countermeasure.

XSL•FO
RenderX

During program development, we anticipated fluctuations in the number of isolation ward beds due to changes in patient numbers or nursing staff availability. To manage this, we stored the current 148 bed numbers in a separate file, Isolation Ward Bed Numbers.xlsx. Any future updates to bed numbers would require changes only to this file. An adjacent "Notes" column was used to provide key details about each bed. Double-occupancy rooms were marked with an asterisk (*), while beds equipped for dialysis were labeled "HD" (hemodialysis), as illustrated in Figure 6.

For further reference, Multimedia Appendix 8 provides supplementary explanations of key concepts related to TPS thinking and the TBP methodology.

Figure 5. An algorithm based on value stream mapping techniques was used to develop a Visual Basic for Applications macro in Excel to automate the manual processes of sorting, deleting, and comparing data.

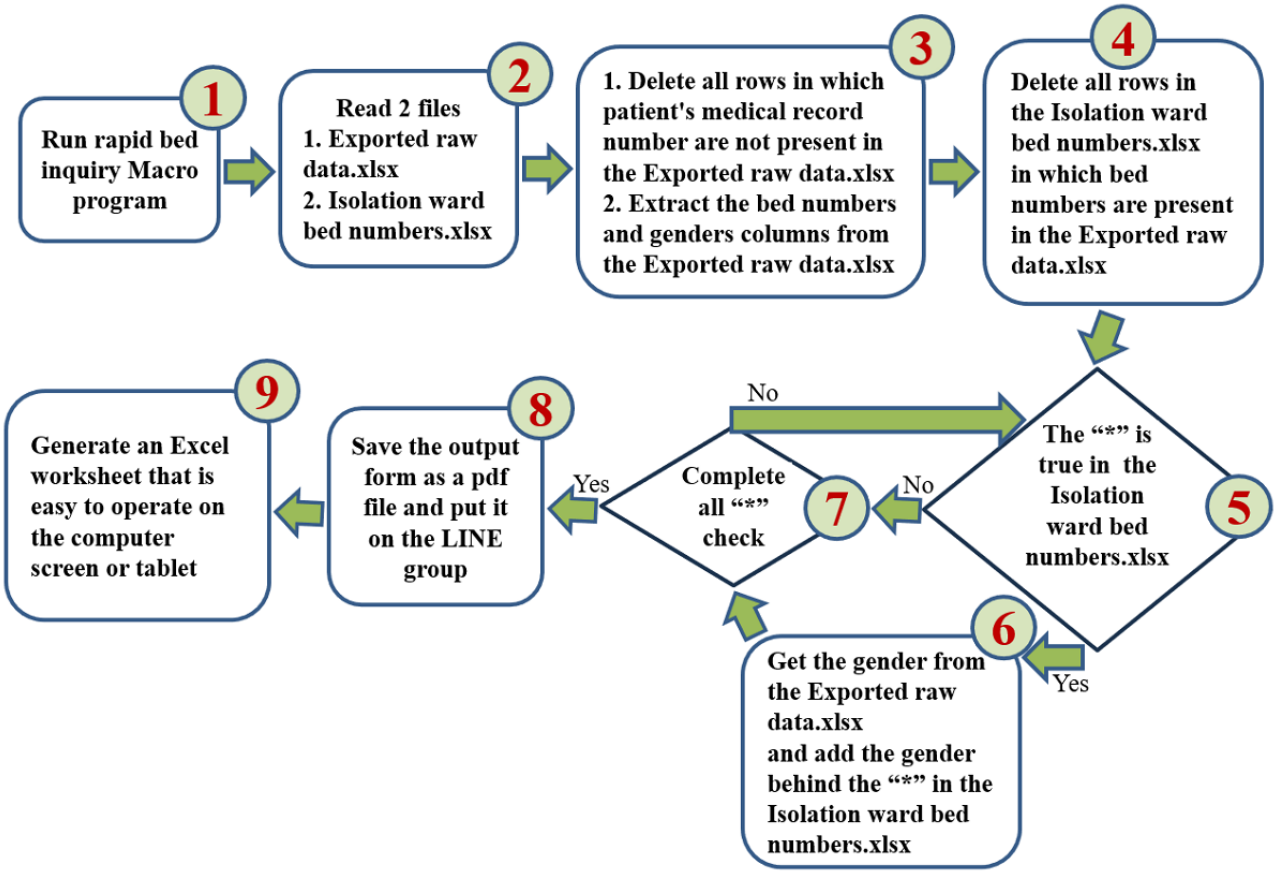

Figure 6. A simplified version of the Isolation Ward Bed Numbers.xlsx format. When the program runs, columns DE, GH, JK, and MN are merged into column AB. Column A lists the available bed numbers, while column B (the Notes column) records bed characteristics. For example, an asterisk (*) indicates a 2-bed room, "HD" denotes a dialysis-ready bed, and "INS" signifies that the patient in the adjacent bed has mental instability.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Bed number	Notes		Bed number	Notes		Bed number	Notes		Bed number	Notes		Bed number	Notes
2	HA9010	/HD		HA8111	/INS*		HA7020			HA6031	*		HICU17	
3	HA9020	/HD		HA8112	/INS*		HA7101	*		HA6032	*			
4	HA9030	/HD		HA8121	*		HA7102	*		HA6050				
5	HA9050	/HD		HA8122	*		HA7131	/HD*		HA6091	*			
6	HA9060			HA8131	/HD*		HA7132	/HD*		HA6092	*			
7	HA9070			HA8132	/HD*		HA7210	/HD		HA6101				
8	HA9080			HA8210			HA7221	*		HA6111				
9	HA9090			HA8320			HA7222	*		HA6121				

Measurable Outcomes and Evaluation

The system was evaluated with a focus on process performance, consistent with quality improvement principles and the TPS, in which data collection is used to measure improvement rather than for accountability [32,33,42,43]. Both quantitative and qualitative metrics were applied to assess the process, identify waste, and compare improvements across the 3 cycles: the manual phone-call process (cycle 1), the HIS built-in platform (cycle 2), and the final automated Microsoft Excel VBA system (cycle 3).

Key quantitative metrics included the time required to complete a query (measured in seconds) and the total number of steps involved. These were supplemented by qualitative observations and VSM to identify inefficiencies, such as unanswered calls or excessive mouse clicks.

The final automated system was evaluated against prespecified primary and secondary outcomes to confirm its effectiveness. Primary indicators included process efficiency, system reliability (100% successful queries with no errors), and user accessibility (available 24/7). Process efficiency was evaluated using 2

XSL•FO
RenderX

criteria: (1) each cycle had to reduce the total number of steps by at least one compared with the previous cycle, and (2) the total process time had to be reduced by at least 50%. Secondary indicators included staff time savings and 100% accuracy in reporting bed status. Collectively, these metrics confirmed the system's ability to provide a just-in-time, reliable, and user-friendly bed inquiry process.

Ethical Considerations

The project underwent administrative review and received approval from hospital leadership in line with institutional standards for operational improvement initiatives. Under applicable national regulations and institutional guidelines, formal ethical approval was not required because the work constituted a quality improvement activity focused on evaluating internal procedures, without involving direct interventions on human participants [46,47]. The study was conducted in compliance with the Personal Data Protection Act [46], particularly Chapter I, Article 6, which permits the use of health-related data for medical and public health purposes, provided that individuals cannot be identified. Taiwan's Constitutional Court, in a decision dated August 12, 2022, confirmed that Article 6, Paragraph 1 of the Personal Data Protection Act—which permits the use of health-related data for medical and public health purposes, provided that individuals cannot be identified—is constitutional [48]. The Court further found that the proviso in Article 6, Paragraph 1, Subparagraph 4 does not violate the principle of legal clarity or the principle of proportionality and is consistent with the protection of informational privacy under Article 22 of the Constitution [48]. In accordance with these regulations and the hospital's internal requirements, informed consent was not required (Multimedia Appendix 9) [47]. To ensure privacy, all information used was anonymized, and no individual can be identified in the figures or appendices included in this manuscript. The system processes only bed numbers and gender from an internal HIS export; no patient-identifiable information is stored or transmitted outside the hospital's secure network. No remuneration was provided, as the project did not involve the direct participation of human participants.

Results

Step 7: Evaluate the Results and the Process

Final System Functionality and Output Forms

Within 3 days of coding and verification, the Microsoft Excel macro was fully functional. Its interface was intentionally simplified to enhance accessibility, allowing users to press a single large button on the worksheet to automatically generate 3 distinct forms in approximately 4 seconds (discussed in the next section).

Final Interface

- PDF Form: Designed for upper management, this form is automatically exported in PDF and shared via the LINE group for quick reference (Figure 7).
- Large-Screen Form: Tailored for bed control physicians, this version is optimized for viewing on large screens such as computer monitors or tablets (Multimedia Appendix 10). The space below the bed numbers in the worksheet allows users to enter and store basic patient information and diagnoses for future reference.
- Smartphone Form: This mobile-friendly version is designed for bed control physicians to use on smartphones (Multimedia Appendix 11).

The completion of the isolation ward bed inquiry Microsoft Excel macro has greatly improved the efficiency of bed control physicians in querying available beds, resulting in an updated VSM that illustrates the enhanced workflow (Figure 8). After downloading the Exported Raw Data.xls, the macro generates 3 output forms in as little as 4 seconds (step B), and the fastest PDF report is sent to the LINE group in 7 seconds (step C). The macro also prevents duplicate file names by incorporating the execution date and time (down to the second) into each file name.

Figure 7. PDF form exported for upper management reference. An asterisk (*) indicates a 2-bed room, "HD" denotes a bed equipped for dialysis, and "INS" signifies that the patient in the adjacent bed has mental instability.

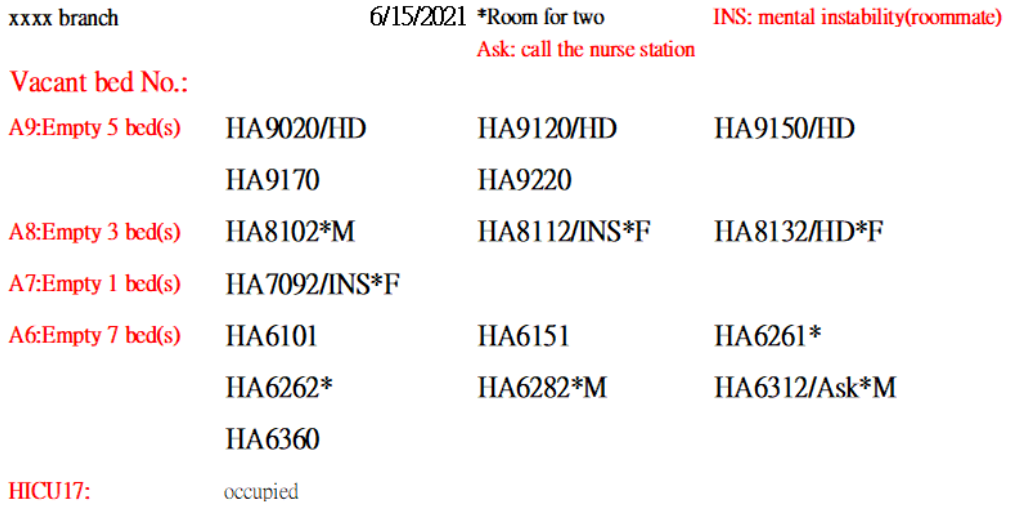

XSL•FO
RenderX

Figure 8. The completion of the isolation ward bed inquiry Excel macro has significantly streamlined the process for bed control physicians to quickly query available beds, resulting in an updated value stream map that reflects the improved workflow.

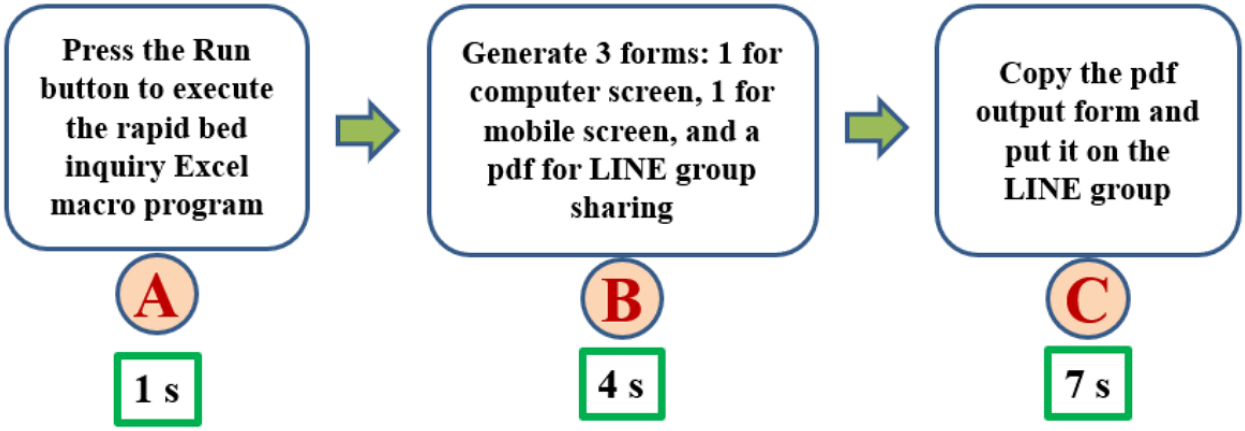

Step 8: Standardize and Share the Successful Process

The Microsoft Excel macro generates a PDF and 2 bed availability forms with a single click, providing an efficient solution that meets the needs of bed control physicians while fulfilling hospital management's expectations for quick and accurate information. The program was stored in a designated hospital system folder, allowing authorized personnel to run it remotely via a virtual private network at any time. It has also been adapted for Microsoft Office 365 and is publicly accessible (Multimedia Appendix 12).

Process Performance Improvements and Achievement of Outcomes

The 3-cycle TBP process yielded a dramatic improvement in the speed, reliability, and usability of the bed inquiry system. The final automated solution reduced the process time by 97.4% (n=442 seconds saved/N=454 baseline seconds) compared with the original manual method. A comparative analysis of performance metrics across the 3 cycles is presented in Table 1. The final system successfully met all prespecified primary and secondary project outcomes, as detailed in Table 2.

Table 1. Comparative performance of the bed-inquiry process across 3 Toyota business practice cycles. This table summarizes outcomes from the value stream maps for the initial manual process (cycle 1; Figure 2; times represent best-case scenarios), the HIS[a] built-in platform (cycle 2; Figure 3; tested by an experienced physician), and the automated Microsoft Excel VBA[b] macro–based system (cycle 3; Figure 8).

Performance metric	Cycle 1: Manual phone-call process	Cycle 2: HIS built-in platform	Cycle 3: Automated Excel VBA system
Primary workflow	Sequential phone calls to 4 wards	Multistep navigation through the HIS graphical user interface	Single-click execution of a Microsoft Excel macro
Number of steps	25 steps for 4 wards	34 steps for 4 wards	3 steps after data export
Time to complete query	454 seconds (4 wards)	144 seconds (4 wards)	12 seconds (4 wards)
Time reduction from baseline	Baseline	68.3% (n=310 seconds saved/N=454 baseline seconds)	97.4% (n=442 seconds saved/N=454 baseline seconds)
Key inefficiencies/notes	Unanswered calls, manual data recording, and verbal communication	Excessive data, manual cross-referencing for double rooms, and multiple mouse clicks and menu selections	None identified; process is fully automated
System reliability and accuracy	Low; prone to human error, delays, and inconsistent information	High accuracy but dependent on correct manual navigation	100% (automated, repeatable process minimizes human error)
User accessibility	Limited; dependent on staff availability at specific times	High; accessible via the hospital network but with usability barriers	24/7; remotely accessible via a virtual private network

[a]HIS: hospital information system.

[b]VBA: Visual Basic for Applications.

XSL•FO
RenderX

Table 2. Achievement of prespecified project outcomes by the final automated system.

Outcome category and indicator	Target	Result	Status
Primary			
Process efficiency (steps)	Reduction of at least one step per cycle	Cycle 1 (25 steps) to cycle 3 (3 steps)	Achieved
Process efficiency (time)	>50% reduction	97.4% (n=442 seconds saved/N=454 baseline seconds) reduction	Achieved
System reliability	100% successful queries	100% (automated process)	Achieved
User accessibility	24/7 availability	24/7 via a virtual private network	Achieved
Secondary			
Staff time savings	Implied reduction in physician time	Significant reduction in manual query time	Achieved
System accuracy	100%	100% (data directly from HIS[a] export)	Achieved

[a]HIS: hospital information system.

Improve Continuously

The Microsoft Excel macro can be adapted for various applications with minor modifications or additional functions. For example, incorporating ward codes allowed us to track availability and occupancy in each ward, and adding an attending physician's list enabled us to monitor how many patients each physician was responsible for in different wards (see Multimedia Appendix 13). This demonstrates how small adjustments can extend the macro's functionality, reflecting the principle of continuous improvement.

Discussion

Principal Findings

This study demonstrates that the structured TBP problem-solving process and TPS mindset facilitated the rapid creation of a Quick Isolation Bed Inquiry System, significantly improving bed allocation for patients with COVID-19. This user-centered system ensured timely admissions while meeting hospital requirements without the need for additional staff or expenses. The primary finding is the dramatic improvement in process efficiency. The final system reduced bed query time by 97.4% (from 454 seconds to 12 seconds) and streamlined the workflow from 25 manual steps to just 3 automated ones. A key insight emerged when the root cause of initial delays (staff shortages) could not be directly eliminated; the TPS mindset prompted a strategic pivot to leverage hidden value within the existing HIS, transforming a cumbersome data export function into an efficient, automated solution. This case provides a practical model of agile problem solving for resource-constrained health care environments. Opportunities for continuous improvement remain, including further automation through VBA integration and consolidation of manual steps, allowing for even greater time savings and resource optimization within the hospital's infrastructure.

Shared Problem-Solving Approaches in TPS and Health Care Systems

Outcomes are driven by actions, and actions arise from thinking. Although intangible, mindset shapes reasoning and direction, and is translated into tangible results through methods and tools [49-51].

TPS Thinking and Health Care Parallels

TPS is rooted in systematic problem solving [31-36,49-52]. The first step is to "go to the source" and confirm problems on-site, based on the principle that all problems stem from waste [32,33]. By identifying and eliminating waste, efficient and low-cost solutions can be achieved [32-35]. Similarly, in health care, illness represents the problem, diagnosis confirms the issue, and treatment addresses the root cause [51,52]. Just as TPS links waste to root causes, physicians identify the underlying pathology before intervening. Treatments that resolve or control disease are "good actions," paralleling TPS's problem-solving mindset [48,51,52].

The structured diagnostic process mirrors TPS: history taking gathers subjective data, physical examination provides direct observations, and laboratory or imaging studies supply objective measures. These steps reflect TPS's emphasis on observing the workplace and collecting accurate numerical data. Both approaches prioritize observation, root cause analysis, and targeted intervention [49-52].

Transforming Health Care Through TPS Principles

The similarity in thinking explains why health care can readily adopt TPS tools [23-27,30,49,51,52]. Virginia Mason Medical Center demonstrated this by applying TPS to achieve transformative results [52]. Patient-centered care in hospitals mirrors TPS's customer-focused philosophy, where success is defined by safe, effective, timely, affordable, and high-quality outcomes. Hospitals, therefore, must reduce costs, improve efficiency, and ensure quality, just as TPS strives for customer satisfaction in manufacturing [49,51,52].

Autonomation (automation with a human touch) ensures that machines or systems stop when abnormalities occur [27,32]. Hospitals apply this principle through medical devices that alarm when readings fall outside safety ranges and reporting systems that notify staff of abnormal test results. In our Isolation Ward Bed Numbers.xlsx program, debugging and error-flagging prevented mistakes and supported accurate bed allocation, embodying autonomation.

Just-in-time refers to delivering the right product at the right time [30,32]. In health care, this translates to timely medical services, exemplified by 24/7 emergency and surgical

XSL•FO
RenderX

departments. Continuous staff availability ensures appropriate care whenever needed. These practices align with TPS values of respect for people, teamwork, and continuous improvement.

Process Improvements by Eliminating Waste or Distilling Value

VSM is a lean tool for sequencing process steps, visualizing waste, and optimizing flow [32,33,36,49,51,52]. Three levels of VSMs were developed in this study, each highlighting waste in the bed inquiry process. The analysis revealed Toyota's classical 7 wastes [30,32,33]: waiting, transportation, motion, overprocessing, inventory, overproduction, and defects. Table 3 summarizes each type of waste and its manifestation in the hospital bed inquiry workflow.

Table 3. Application of Toyota's 7 wastes in the bed inquiry process.

Waste type	Definition in the Toyota production system	Example in the bed inquiry process
Waiting	Idle time when processes are delayed	Physicians waited up to 454 seconds for ward staff to answer phone calls before confirming bed status (Figure 2).
Transportation	Unnecessary movement of materials or information that is not essential to value-adding steps	Ward staff left phone stations to check monitors; physicians walked across wards to verify vacancies (Figure 2).
Motion	Excessive movement of people or equipment that does not add value to the product or service	Physicians performed repetitive clicks, navigated menus, and manually cross-referenced tables in the hospital's inquiry system (Figure 3).
Overprocessing	Performing more work or steps than necessary, overly complex tasks, or exceeding customer requirements	The workflow involved 25 manual steps in cycle 1 (Figure 2) and 34 steps in cycle 2 (Figure 3), far beyond what was required.
Inventory	Excess stock, materials, or data beyond what is useful	The database listed 761 beds, but only 448 were clinically available for patient use, leading to confusion and inefficiency
Overproduction	Producing more data, information, or services than needed	The system displayed 28 fields and 761 records, although only 3 fields and 148 records were essential.
Defects	Any product or service that fails to meet specifications, resulting in rework, scrap, or correction	Manual transcription produced error-prone records, increasing the risk of inaccurate or incomplete bed information; inefficiencies persisted in cycles 1 and 2.

We were able to develop the Quick Isolation Bed Inquiry System only after extracting value from the VSMs in Figures 2 and 3 and creating a new VSM, as shown in Figure 8. This followed the unsuccessful attempts to eliminate waste in the first and second cycles. The final automated system systematically removed these wastes, resulting in a streamlined, efficient, and error-proof workflow (Figure 8). This step-by-step process of extracting value and eliminating waste provides a clear demonstration of problem decomposition through VSM and illustrates how processes can be improved to achieve all prespecified outcomes. Moreover, these iterative cycles highlight the importance of learning from unsuccessful attempts; without the inefficiencies revealed in cycles 1 and 2, the breakthrough in cycle 3 could not have been achieved. This underscores TPS's core philosophy of continuous improvement, in which even failed experiments generate knowledge that guides the design of more effective and sustainable solutions.

The TPS-Driven Approach as an Agile Paradigm for Health Care IT

This study demonstrated that a rapid, zero-cost solution can succeed where traditional health care IT projects often fail. Conventional IT implementations are frequently hindered by high costs, prolonged rollouts, and limited acceptance from clinical users [12,18,53-55]. By leveraging existing software (Microsoft Excel) and internal expertise, we avoided financial barriers and external vendors, while the direct involvement of bed control physicians in the design ensured seamless adoption, in contrast to the resistance often seen in top-down systems. The system was developed in only 3 days, contrasting with the lengthy timelines typical of commercial platforms. Notably, the 3 TBP cycles paralleled Agile software development, each serving as a short "sprint" with rapid prototyping, user feedback, and incremental improvement [55-59]. This iterative approach emphasized collaboration between users and problem solvers and illustrated the Agile principle of "responding to change over following a plan" [55-58]. These findings not only highlight the adaptability of TBP as a structured form of Agile in health care but also provide a practical model for resource-constrained settings, reinforcing the broader discussion on lean and user-centered innovation.

Improve Continuously

After admissions of patients with COVID-19 ceased in August 2022, we easily adapted the Quick Isolation Bed Inquiry System's Microsoft Excel macro for general ward use with minor modifications. This adaptation, detailed in Multimedia Appendices 13 and 14, has significantly enhanced bed allocation transparency and maximized utilization. Multimedia Appendices 13 and 14 demonstrate that the system can display key bed-related metrics—such as patient load per attending physician and per ward, total inpatients per ward and department, available beds per ward, and next-day discharge information (patients per ward and per attending physician)—based on input files. Crucially, the macro's core program requires no modification for expansion. Simply inputting a file with physician names and their departmental affiliations (eg, surgical, internal medicine, or even the entire hospital) will display all relevant bed information for that specific group. This demonstrates the system's inherent scalability beyond infectious disease scenarios.

XSL•FO
RenderX

Impact on the Hospital's Bottom Line

Overview

While we did not quantify the direct financial "bottom line" impact in monetary terms, the implications for health care are significant.

Improved Patient Flow

Faster bed allocation allows patients requiring admission (especially severe COVID-19 cases) to be moved into appropriate care settings more quickly. This reduces wait times in emergency departments or holding areas, which can indirectly lead to better patient outcomes and shorter lengths of stay.

Optimized Resource Utilization

Efficient bed management ensures that beds are not left vacant unnecessarily, maximizing the hospital's capacity for patient care.

Reduced Staff Burnout and Increased Productivity

By reducing the time and mental strain staff spend on tedious bed inquiries, they can dedicate more time to direct patient care, consultations, and other critical duties. This boosts staff morale and overall productivity, helping to mitigate the impact of existing staff shortages. Although not a direct monetary saving, avoiding additional workforce costs was a key objective.

Implications for Health Care Management and Future Directions

Our results point to valuable directions for practical health care delivery and policy development. This study provides a compelling argument for empowering frontline innovators, demonstrating that clinicians and staff can develop highly effective solutions when equipped with appropriate methodological tools. This fosters a culture of bottom-up innovation, which can be more agile and responsive than traditional top-down management approaches [54,56,60]. For hospital IT departments, this case serves as a call to re-evaluate existing IT assets for "hidden value." Rather than focusing exclusively on procuring new systems, IT leaders should encourage periodic audits of existing platforms to identify underutilized functionalities—such as data export features—that can be leveraged for rapid, low-cost solution development. This also suggests a need for health care organizations to cultivate a "lean IT" culture that extends beyond clinical workflows into IT development and administrative problem solving. Such a culture would foster continuous improvement and provide staff with the training and autonomy to eliminate waste and distill value in digital processes.

To align with our commitment to continuous improvement and long-term system utility, we plan to conduct a comprehensive outcome evaluation of this innovation. Future efforts will focus on quantifying its impact on key performance indicators, including bed turnover rates, patient flow efficiency (eg, reduced wait times), and staff time savings. We will also administer user satisfaction surveys to gather both qualitative and quantitative feedback from bed control physicians, nursing staff, and administrators on usability, reliability, and workflow integration. In addition, we will evaluate the system's long-term sustainability, including performance and maintenance requirements. Together, these outcome measures will provide robust evidence of the system's effectiveness and enduring value, strengthening the case for impactful technology-enabled innovations in health care. Several avenues for advancement emerge from this work. Integrating the system with HIS through application programming interfaces, rather than relying on file exports, would enable true real-time updates. Incorporating machine learning algorithms could allow the prediction of bed availability based on historical patterns and current patient flow, shifting management from reactive to proactive. Furthermore, natural language processing could support voice-activated queries, further reducing user burden and enhancing accessibility.

Limitations

Information Security and System Sustainability

This paper has 5 limitations. First, the use of VBA macros introduces potential information-security risks, a well-recognized concern in IT governance. However, these risks were proactively mitigated within a robust security framework: the system operates exclusively on the hospital's secure intranet and is accessible only to authorized personnel via a virtual private network. Critically, the macro does not process, transmit, or store any patient-identifiable data outside institutional systems. These controls—together with standard institutional safeguards such as regular system patching and anti-malware protection—effectively mitigate the residual risks associated with macro-enabled files. This reflects a deliberate balance between urgent operational needs and security, offering a pragmatic model for rapid, frontline innovation within a controlled environment. Second, the long-term sustainability and compatibility of a VBA-based tool present a valid governance challenge, particularly amid ongoing Microsoft Office updates. We addressed this by framing the system's lifecycle in 2 stages.

Stage 1: Proven Prototype (Current State)

The current tool functioned as a robust and effective prototype, with its durability demonstrated by a straightforward migration from Microsoft Excel 2016 to Microsoft 365 that required only minor modifications. By providing the updated and fully functional code in Multimedia Appendix 12, we confirm its ongoing usability.

Stage 2: Blueprint for Enterprise Integration (Future State)

Having demonstrated substantial value and impact, the tool now serves as both a proof-of-concept and a clear blueprint for our IT department to develop a permanent, integrated solution within the hospital's enterprise infrastructure. This 2-stage approach addresses the long-term challenge by strategically transitioning the innovation from a rapid, tactical solution into a formal, sustainable system.

Third, as a single-center case study, our implementation was shaped by local institutional factors, such as an existing HIS export function and in-house familiarity with VBA. However, these factors do not limit the generalizability of the underlying

XSL•FO
RenderX

methodology. We argue that the core contribution lies not in the specific tool (VBA), but in the agile, resource-conscious framework. The fundamental prerequisite for our approach is an information system capable of exporting data to a standard spreadsheet format (eg, .xlsx, .csv)—a near-universal feature of modern hospital systems. The choice of VBA was one of convenience during an emergency, leveraging a built-in, zero-cost tool. The system's logic can be readily implemented in any modern programming environment capable of reading a spreadsheet file, including Python, R, C#, or Java. Thus, the approach is technology-agnostic and can be adapted to the existing technical expertise of any health care institution. The required skill is not "VBA expertise" but simply the ability to parse a data file—a straightforward task for any IT professional. This underscores that our problem-solving method is both highly accessible and transferable to a wide range of organizations facing similar challenges.

Fourth, while we demonstrated substantial improvements in process efficiency, this study did not measure the downstream impact on clinical or financial outcomes. We did not quantify changes in patient wait times, bed turnover rates, or overall length of stay, which represent important areas for future research.

Fifth, because most of the authors were also the system developers, a potential risk of evaluation bias exists. However, we contend that this risk was substantially mitigated by our study design and methodological approach. Our primary safeguard was the use of prespecified, objective, and quantitative metrics—specifically, query time and the number of process steps. These hard measures (eg, a 97.4% reduction, from 454 seconds to 12 seconds, in query time and a reduction from 25 steps to 3) are not susceptible to subjective interpretation. Furthermore, the dual implementer-evaluator role was a deliberate choice, aligning with established frameworks for rapid, real-world quality improvement, where this practice is common and often necessary. Our study's design is consistent with the SQUIRE 2.0 guidelines for reporting quality improvement [39], which accommodate projects in which implementers evaluate their own work. The iterative process also mirrors the plan-do-study-act (PDCA) cycle [41] and Agile methodology [55], both of which rely on tight feedback loops in which the development team is intrinsically involved in evaluation to facilitate agile problem solving. While independent validation could further strengthen the findings, our methodology represents a well-supported approach in quality improvement science, ensuring the credibility of our results in this urgent, real-world context.

Conclusions

By systematically applying the TPS mindset and the TBP framework, this study developed a zero-cost, user-centered Quick Isolation Bed Inquiry System. The intervention proved highly effective in a real-world hospital setting during the COVID-19 pandemic. All prespecified outcomes were achieved without incurring additional costs or requiring new staff, ensuring the timely admission of critically ill patients. This work offers a practical model for agile, bottom-up innovation in health care IT, demonstrating that leveraging underutilized functions within existing HIS can deliver rapid and impactful solutions, particularly in resource-constrained or crisis environments. Our findings contribute to the literature by providing a real-world example of applying TPS principles to overcome the high costs and lengthy implementation timelines typical of conventional health care IT projects. The system's adaptability and scalability for broader hospital use have been confirmed postpandemic. Future research should quantify downstream impacts on clinical outcomes, such as patient wait times and bed turnover rates, to further validate the value of this approach.

Data Availability

The data sets analyzed in this study are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: C-CL, J-HS, S-LC, H-LH, C-YPL, H-MH, T-CK
Data curation: J-HS, S-LC, C-CL
Formal analysis: J-HS, S-LC, C-CL
Investigation: J-HS, S-LC, C-CL
Methodology: C-CL, J-HS, H-LH, H-MH, T-CK
Project administration: C-CL, J-HS, S-LC, H-LH, C-YPL, H-MH
Software: J-HS, S-LC, C-CL
Supervision: H-MH, T-CK
Validation: J-HS, S-LC, T-CK, C-CL
Visualization: J-HS, S-LC, C-CL
Writing – original draft: C-CL, J-HS, S-LC, H-LH, C-YPL, H-MH
Writing – review & editing: C-CL, J-HS, S-LC, T-CK, H-LH, C-YPL, H-MH

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
Calculate the time spent (in seconds) querying the availability of vacant beds in wards A9, A8, A7, and A6, based on the value stream map shown in Figure 2.
[PDF File (Adobe PDF File), 189 KB - formative_v9i1e67152_app1.pdf]

Multimedia Appendix 2
Methodology and rationale for using "best-case" or "minimum-time" scenarios.
[PDF File (Adobe PDF File), 107 KB - formative_v9i1e67152_app2.pdf]

Multimedia Appendix 3
The "Five Whys" technique was used to identify the root cause of the problem: "Why are calls not answered promptly?".
[PNG File , 171 KB - formative_v9i1e67152_app3.png]

Multimedia Appendix 4
This is the menu screen of the hospital bed inquiry platform. Pressing the "Query" button displays 761 bed records, each containing 28 fields.
[PNG File , 237 KB - formative_v9i1e67152_app4.png]

Multimedia Appendix 5
The hospital bed inquiry platform's built-in query function allows users to check bed availability across different hospital wards.
[PNG File , 309 KB - formative_v9i1e67152_app5.png]

Multimedia Appendix 6
In collaboration with the Medical Affairs Office, we applied the "Five Whys" method to identify the root cause of the problem: "Why must we use the double-room comparison table to identify double-occupancy rooms?".
[PNG File , 91 KB - formative_v9i1e67152_app6.png]

Multimedia Appendix 7
A simplified version of the Exported Raw Data.xls file. Of the 28 available columns, only 3—highlighted in red—are required for use.
[PNG File , 720 KB - formative_v9i1e67152_app7.png]

Multimedia Appendix 8
The following section offers supplementary explanations of key concepts related to Toyota production system (TPS) thinking and the Toyota business practice methodology. It serves as background material for readers who may be less familiar with TPS.
[PDF File (Adobe PDF File), 224 KB - formative_v9i1e67152_app8.pdf]

Multimedia Appendix 9
Scope of human research cases exempt from ethics committee review.
[PDF File (Adobe PDF File), 126 KB - formative_v9i1e67152_app9.pdf]

Multimedia Appendix 10
This large-screen form is designed for use by bed-control physicians on computers or tablets. An asterisk (*) indicates a 2-bed room, “HD” denotes a bed equipped for dialysis, and “INS” signifies that the patient in the adjacent bed has mental instability.
[PNG File , 67 KB - formative_v9i1e67152_app10.png]

Multimedia Appendix 11
This smartphone form is designed for mobile use by bed-control physicians. An asterisk (*) indicates a 2-bed room, “HD” denotes a bed equipped for dialysis, and “INS” signifies that the patient in the adjacent bed has mental instability.
[PNG File , 18 KB - formative_v9i1e67152_app11.png]

Multimedia Appendix 12
The original Excel macro code for the Quick Isolation Bed Inquiry System has been adapted for compatibility with Microsoft Office 365.
[PDF File (Adobe PDF File), 529 KB - formative_v9i1e67152_app12.pdf]

XSL•FO
RenderX

Multimedia Appendix 13
With minor adjustments, the Excel macro provides versatility for various applications. For example, integrating ward codes allows efficient monitoring of bed availability and occupancy across different wards, while adding an attending physician list enables tracking of patient loads for each physician. In the Notes column, an asterisk (*) indicates a 2-bed room, "HD" denotes a dialysis-equipped bed, and "INS" signifies that the patient in the adjacent bed has mental instability.
[PNG File , 206 KB - formative_v9i1e67152_app13.png]

Multimedia Appendix 14
The Quick Isolation Bed Inquiry System's Excel macro can be easily adapted for a variety of applications with minor modifications or additional functions. By incorporating ward or department codes, we were able to track bed availability and occupancy in specific areas. Adding an attending physician list also allowed us to monitor the number of patients for whom each physician was responsible across different wards. Additionally, the system can generate statistics on the number of patients in each ward expected to be discharged the following day.
[PNG File , 284 KB - formative_v9i1e67152_app14.png]

References

1. Sen-Crowe B, Sutherland M, McKenney M, Elkbuli A. A closer look into global hospital beds capacity and resource shortages during the COVID-19 pandemic. J Surg Res 2021 Apr;260:56-63 [FREE Full text] [doi: 10.1016/j.jss.2020.11.062] [Medline: 33321393]
2. Phua J, Weng L, Ling L, Egi M, Lim C, Divatia JV, et al. Intensive care management of coronavirus disease 2019 (COVID-19): challenges and recommendations. The Lancet Respiratory Medicine 2020 May;8(5):506-517. [doi: 10.1016/s2213-2600(20)30161-2]
3. Grasselli G, Pesenti A, Cecconi M. Critical care utilization for the COVID-19 outbreak in Lombardy, Italy: early experience and forecast during an emergency response. JAMA 2020 Apr 28;323(16):1545-1546 [FREE Full text] [doi: 10.1001/jama.2020.4031] [Medline: 32167538]
4. Monthly Report. Taiwan Centers for Disease Control (CDC). URL: https://www.cdc.gov.tw/En/Report/MonthList/14567 [accessed 2025-07-17]
5. Allder S, Silvester K, Walley P. Managing capacity and demand across the patient journey. Clin Med (Lond) 2010 Feb;10(1):13-15 [FREE Full text] [doi: 10.7861/clinmedicine.10-1-13] [Medline: 20408298]
6. Allder S, Silvester K, Walley P. Understanding the current state of patient flow in a hospital. Clin Med (Lond) 2010 Oct;10(5):441-444 [FREE Full text] [doi: 10.7861/clinmedicine.10-5-441] [Medline: 21117373]
7. Belciug S, Gorunescu F. Improving hospital bed occupancy and resource utilization through queuing modeling and evolutionary computation. J Biomed Inform 2015 Feb;53:261-269 [FREE Full text] [doi: 10.1016/j.jbi.2014.11.010] [Medline: 25433363]
8. Guido R, Groccia M, Conforti D. An efficient matheuristic for offline patient-to-bed assignment problems. European Journal of Operational Research 2018 Jul;268(2):486-503 [FREE Full text] [doi: 10.1016/j.ejor.2018.02.007]
9. Mark TL, Howard JN, Misra S, Fuller L. Bed tracking systems: do they help address challenges in finding available inpatient beds? Psychiatr Serv 2019 Oct 01;70(10):921-926. [doi: 10.1176/appi.ps.201900079] [Medline: 31215354]
10. Hiragi S, Hatanaka J, Sugiyama O, Saito K, Nambu M, Kuroda T. Token economy-based hospital bed allocation to mitigate information asymmetry: proof-of-concept study through simulation implementation. JMIR Form Res 2022 Mar 04;6(3):e28877 [FREE Full text] [doi: 10.2196/28877] [Medline: 35254264]
11. Tello M, Reich ES, Puckey J, Maff R, Garcia-Arce A, Bhattacharya BS, et al. Machine learning based forecast for the prediction of inpatient bed demand. BMC Med Inform Decis Mak 2022 Mar 02;22(1):55 [FREE Full text] [doi: 10.1186/s12911-022-01787-9] [Medline: 35236345]
12. Schmidt PE, Meredith P, Prytherch DR, Watson D, Watson V, Killen RM, et al. Impact of introducing an electronic physiological surveillance system on hospital mortality. BMJ Qual Saf 2015 Jan;24(1):10-20. [doi: 10.1136/bmjqs-2014-003073] [Medline: 25249636]
13. Wong HJ, Caesar M, Bandali S, Agnew J, Abrams H. Electronic inpatient whiteboards: improving multidisciplinary communication and coordination of care. Int J Med Inform 2009 Apr;78(4):239-247. [doi: 10.1016/j.ijmedinf.2008.07.012] [Medline: 18786851]
14. Tahmasbi F, Toni E, Javanmard Z, Kheradbin N, Nasiri S, Sadoughi F. An overview of reviews on digital health interventions during COVID- 19 era: insights and lessons for future pandemics. Arch Public Health 2025 May 09;83(1):129. [doi: 10.1186/s13690-025-01590-8] [Medline: 40346715]
15. Percival J, McGregor C. An evaluation of understandability of patient journey models in mental health. JMIR Hum Factors 2016 Jul 28;3(2):e20 [FREE Full text] [doi: 10.2196/humanfactors.5640] [Medline: 27471006]
16. Bin KJ, Melo AAR, da Rocha JGMF, de Almeida RP, Cobello Junior V, Maia FL, et al. The impact of artificial intelligence on waiting time for medical care in an urgent care service for COVID-19: single-center prospective study. JMIR Form Res 2022 Feb 01;6(2):e29012 [FREE Full text] [doi: 10.2196/29012] [Medline: 35103611]

XSL•FO
RenderX

17. Hospital bed management systems market size report 2030. Grand View Research. San Francisco, CA: Grand View Research; 2023. URL: https://www.grandviewresearch.com/industry-analysis/hospital-bed-management-systems-market-report [accessed 2025-07-17]
18. What is bed management - benefits for healthcare Internet. The Access Group. London: The Access Group; 2025 Aug 28. URL: https://www.theaccessgroup.com/en-gb/health-social-care/software/bed-management/what-is-bed-management/ [accessed 2025-10-06]
19. Sittig DF, Wright A, Coiera E, Magrabi F, Ratwani R, Bates DW, et al. Current challenges in health information technology-related patient safety. Health Informatics J 2020 Mar;26(1):181-189 [FREE Full text] [doi: 10.1177/1460458218814893] [Medline: 30537881]
20. Cresswell K, Sheikh A. Organizational issues in the implementation and adoption of health information technology innovations: an interpretative review. Int J Med Inform 2013 May;82(5):e73-e86. [doi: 10.1016/j.ijmedinf.2012.10.007] [Medline: 23146626]
21. Slocum JD, Jelke D, Mai Q, Johnson JK, Nguyen MLT, Cong L, et al. User-centered design of trauma systems solutions for retriage of patients with injury: mixed methods study. J Med Internet Res 2025 Aug 27;27:e70846 [FREE Full text] [doi: 10.2196/70846] [Medline: 40862520]
22. Womack JP, Jones DT, Roods D. The Machine That Changed the World. New York, NY: Rawson Associates Scribner; 1990.
23. Teich ST, Faddoul FF. Lean management-the journey from toyota to healthcare. Rambam Maimonides Med J 2013 Apr;4(2):e0007 [FREE Full text] [doi: 10.5041/RMMJ.10107] [Medline: 23908857]
24. Mazzocato P, Savage C, Brommels M, Aronsson H, Thor J. Lean thinking in healthcare: a realist review of the literature. Qual Saf Health Care 2010 Oct;19(5):376-382. [doi: 10.1136/qshc.2009.037986] [Medline: 20724397]
25. D'Andreamatteo A, Ianni L, Lega F, Sargiacomo M. Lean in healthcare: a comprehensive review. Health Policy 2015 Sep;119(9):1197-1209. [doi: 10.1016/j.healthpol.2015.02.002] [Medline: 25737260]
26. Radnor ZJ, Holweg M, Waring J. Lean in healthcare: the unfilled promise? Social Science & Medicine 2012 Feb;74(3):364-371. [doi: 10.1016/j.socscimed.2011.02.011] [Medline: 21414703]
27. Lin C, Chueh J, Chen H, Chiu Y, Chu D. Applying the Toyota production system to decrease the time required to transport patients undergoing surgery from the general ward to the operating room and reviewing the essence of lean thinking. Front Med (Lausanne) 2022;9:1054583 [FREE Full text] [doi: 10.3389/fmed.2022.1054583] [Medline: 36569138]
28. Vest J, Gamm L. A critical review of the research literature on Six Sigma, Lean and StuderGroup's Hardwiring Excellence in the United States: the need to demonstrate and communicate the effectiveness of transformation strategies in healthcare. Implement Sci 2009 Jul 01;4:35 [FREE Full text] [doi: 10.1186/1748-5908-4-35] [Medline: 19570218]
29. Rose C, Nichols T, Hackner D, Chang J, Straube S, Jooste W, et al. Utilizing Lean software methods to improve acceptance of global eHealth initiatives: results from the implementation of the Basic Emergency Care app. JMIR Form Res 2021 May 26;5(5):e14851 [FREE Full text] [doi: 10.2196/14851] [Medline: 33882013]
30. Lin C, Shen J, Chen S, Chen H, Huang H. Developing a cost-effective surgical scheduling system applying Lean thinking and Toyota's methods for surgery-related big data for improved data use in hospitals: user-centered design approach. JMIR Form Res 2024 May 24;8:e52185 [FREE Full text] [doi: 10.2196/52185] [Medline: 38787610]
31. The power of process: Toyota mentoring prepares students to solve problems. Toyota Motor Corporation. URL: https://pressroom.toyota.com/the-power-of-process-toyota-mentoring-prepares-students-to-solve-problems/ [accessed 2024-08-17]
32. Ohno T. Toyota Production System: Beyond Large-Scale Production. Boca Raton, FL: CRC Press; 1988.
33. Rand G, Womack J, Jones D. Lean Thinking-Banish Waste and Create Wealth in Your Corporation. New York, NY: Free Press; Nov 1997:1148.
34. Imai M. Gemba Kaizen: A Commonsense Approach to a Continuous Improvement Strategy, 2nd Edition. Lake Mary, FL: McGraw Hill LLC; 2012.
35. Liker JK. The Toyota Way, Second Edition: 14 Management Principles from the World's Greatest Manufacturer. New York, NY: McGraw Hill LLC; 2021.
36. Rother M, Shook J. Learning to See: Value Stream Mapping to Add Value and Eliminate Muda. Boston, MA: Lean Enterprise Institute; 2018.
37. Proctor E, Silmere H, Raghavan R, Hovmand P, Aarons G, Bunger A, et al. Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda. Adm Policy Ment Health 2011 Mar;38(2):65-76 [FREE Full text] [doi: 10.1007/s10488-010-0319-7] [Medline: 20957426]
38. Mash R, Nyasulu J, Malan Z, Hirschhorn L. Understanding implementation research. Afr J Prim Health Care Fam Med 2025 Jun 03;17(2):e1-e7. [doi: 10.4102/phcfm.v17i2.4934] [Medline: 40459100]
39. Goodman D, Ogrinc G, Davies L, Baker GR, Barnsteiner J, Foster TC, et al. Explanation and elaboration of the SQUIRE (Standards for Quality Improvement Reporting Excellence) guidelines, V.2.0: examples of SQUIRE elements in the healthcare improvement literature. BMJ Qual Saf 2016 Dec;25(12):e7. [doi: 10.1136/bmjqs-2015-004480] [Medline: 27076505]
40. Sobek II, Durward K, Art S. Understanding A3 Thinking: A Critical Component of Toyota's PDCA Management System. London, UK: Taylor & Francis; 2008.

XSL•FO
RenderX

41. Taylor MJ, McNicholas C, Nicolay C, Darzi A, Bell D, Reed JE. Systematic review of the application of the plan-do-study-act method to improve quality in healthcare. BMJ Qual Saf 2014 Apr;23(4):290-298 [FREE Full text] [doi: 10.1136/bmjqs-2013-001862] [Medline: 24025320]
42. Gulliksen J, Göransson B, Boivie I, Blomkvist S, Persson J, Cajander ?. Key principles for user-centred systems design. Behaviour & Information Technology 2003 Nov;22(6):397-409. [doi: 10.1080/01449290310001624329]
43. Ray J, Finn EB, Tyrrell H, Aloe CF, Perrin EM, Wood CT, et al. User-centered framework for implementation of technology (UFIT): development of an integrated framework for designing clinical decision support tools packaged with tailored implementation strategies. J Med Internet Res 2024 May 21;26:e51952. [doi: 10.2196/51952] [Medline: 38771622]
44. Alexander M, Kusleika R. Excel 2016 Power Programming With VBA. Hoboken, NJ: John Wiley & Sons; 2016.
45. Beecher K. Computational Thinking: A Beginner's Guide to Problem-Solving and Programming. Swindon, UK: BCS Learning & Development Ltd; 2017.
46. Personal Data Protection Act. Taiwan Ministry of Justice. URL: https://law.moj.gov.tw/ENG/LawClass/LawAll.aspx?pcode=I0050021 [accessed 2024-05-10]
47. Scope of human research cases exempt from review by the ethics committee (Announcement No. 1010265075 by the Department of Health, Medical Affairs). Official Gazette of the Executive Yuan/Taipei Government. 2012 Jul 5. URL: https://tinyurl.com/z7fmr5cp [accessed 2024-05-10]
48. Case on the National Health Insurance Research Database [Judgment No. 111-Hsien-Pan-13]. Constitutional Court, R.O.C. (Taiwan). 2022 Aug 12. URL: https://cons.judicial.gov.tw/en/docdata.aspx?fid=5534&id=352245 [accessed 2025-09-18]
49. Jimmerson C. A3 Problem Solving for Healthcare: A Practical Method for Eliminating Waste. New York, NY: CRC Press; 2007.
50. Shook J. Managing to Learn: Using the A3 Management Process to Solve Problems, Gain Agreement, Mentor and Lead (Managing to Learn: Using the A3 Management ... Problems, Gain Agreement, Mentor, and Lead). Boston, MA: Lean Enterprise Institute; 2010.
51. Modig N, Åhlström P. This Is Lean: Resolving the Efficiency Paradox. Stockholm, Sweden: Rheologica Publishing; 2014.
52. Kenney C. Transforming Health Care: Virginia Mason Medical Center's Pursuit of the Perfect Patient Experience. Boca Raton, FL: CRC Press; 2010.
53. Kaplan B, Harris-Salamone KD. Health IT success and failure: recommendations from literature and an AMIA workshop. J Am Med Inform Assoc 2009;16(3):291-299 [FREE Full text] [doi: 10.1197/jamia.M2997] [Medline: 19261935]
54. Granja C, Janssen W, Johansen MA. Factors determining the success and failure of eHealth interventions: systematic review of the literature. J Med Internet Res 2018 May 01;20(5):e10235 [FREE Full text] [doi: 10.2196/10235] [Medline: 29716883]
55. Goodison R, Borycki EM, Kushniruk AW. Use of Agile project methodology in health care IT implementations: a scoping review. Stud Health Technol Inform 2019;257:140-145. [Medline: 30741186]
56. Abrahamsson P, Salo O, Ronkainen J, Warsta J. Agile software development methods: review and analysis. VTT Publications. 2002. URL: https://publications.vtt.fi/pdf/publications/2002/P478.pdf [accessed 2025-10-06]
57. Desai M, Tardif-Douglin M, Miller I, Blitzer S, Gardner DL, Thompson T, et al. Implementation of Agile in healthcare: methodology for a multisite home hospital accelerator. BMJ Open Qual 2024 May 27;13(2):e002764 [FREE Full text] [doi: 10.1136/bmjoq-2024-002764] [Medline: 38802269]
58. Meinert E, Milne-Ives M, Surodina S, Lam C. Agile requirements engineering and software planning for a digital health platform to engage the effects of isolation caused by social distancing: case study. JMIR Public Health Surveill 2020 May 06;6(2):e19297 [FREE Full text] [doi: 10.2196/19297] [Medline: 32348293]
59. Stoumpos AI, Kitsios F, Talias MA. Digital transformation in healthcare: technology acceptance and its applications. Int J Environ Res Public Health 2023 Mar 15;20(4):1-44 [FREE Full text] [doi: 10.3390/ijerph20043407] [Medline: 36834105]
60. Silic M, Back A. Shadow IT – a view from behind the curtain. Computers & Security 2014 Sep;45:274-283. [doi: 10.1016/j.cose.2014.06.007]

Abbreviations

HIS: hospital information system
PDCA: plan-do-check-act
SQUIRE: Standards for Quality Improvement Reporting Excellence
TBP: Toyota business practice
TPS: Toyota production system
VBA: Visual Basic for Applications
VSM: value stream map

XSL•FO
RenderX

Edited by J Sarvestan; submitted 04.10.24; peer-reviewed by N Gupta, DZQ Gan; comments to author 20.03.25; revised version received 26.07.25; accepted 22.09.25; published 17.10.25.

Please cite as:
Lin CC, Shen JH, Chang SL, Kuo TC, Huang HL, Lin CYP, Huang HM
Developing a Quick Isolation Bed Inquiry System During the COVID-19 Outbreak: User-Centered Design Approach Based on the Toyota Production System
JMIR Form Res 2025;9:e67152
URL: https://formative.jmir.org/2025/1/e67152
doi:10.2196/67152
PMID:

©Chien-Chung Lin, Jian-Hong Shen, Shan-Li Chang, Tai-Chih Kuo, Hui-Ling Huang, Cheng-Yi Peter Lin, Hung-Meng Huang. Originally published in JMIR Formative Research (https://formative.jmir.org), 17.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Quality of an Assistive Technology Web Application for Primary Care Physicians Serving Older Adults: Concurrent Mixed Methods Study

Elsa M Orellano-Colón[1*], MSc, PhD; Wency Bonilla-Díaz[2*], MPH; Radamés Revilla-Orellano[1*], MD; Jesús Mejías-Castro[3*], BAc; Abiel Roche-Lima[1*], PhD

[1]Medical Sciences Campus, University of Puerto Rico, San Juan, Puerto Rico
[2]Huertas College, Caguas, Puerto Rico
[3]University of Puerto Rico at Humacao, Humacao, Puerto Rico
[*]all authors contributed equally

Corresponding Author:
Elsa M Orellano-Colón, MSc, PhD
Medical Sciences Campus
University of Puerto Rico
PO Box 365067
San Juan, 00936-5067
Puerto Rico
Phone: 1 7877582525 ext 4200
Email: elsa.orellano@upr.edu

Abstract

Background: Older Latinos living in Puerto Rico experience significantly higher rates of functional disabilities (FDs; 1093/87,300, 27.8%) compared to older adults in the continental United States (755,685/57,913,200, 13.3%). While assistive technologies (ATs) can improve daily function and support aging in place, primary care physicians (PCPs), who are essential in addressing FDs resulting from chronic diseases, often lack knowledge about AT devices and services. The Mi Guía de Asistencia Tecnológica (MGAT; My Assistive Technology Guide) web application was empirically developed to address this gap by providing comprehensive information and videos about AT devices for older adults with functional difficulties in daily living activities.

Objective: This study aimed to assess the quality of MGAT among PCPs and describe their experiences using the app to increase access to AT for older Latinos.

Methods: A total of 10 PCPs participated in this usability project, receiving MGAT training before a 30-day implementation period. A concurrent mixed methods design was used, combining quantitative data from the User Version of the Mobile User Application Rating Scale (uMARS) and qualitative insights from semistructured individual interviews. The analysis included descriptive statistics and a directed content analysis.

Results: The MGAT received high overall objective quality ratings on uMARS (mean 4.06, SD 1.05). Among subdomains, information scored highest (mean 4.60, SD 0.51), followed by functionality (mean 4.20, SD 0.63), aesthetics (mean 4.00, SD 0.82), and engagement, which scored lowest (mean 3.34, SD 1.51). Subjective quality ratings were also favorable, with a mean score of 3.93 (SD 1.19), with recommending the app to others scoring the highest (mean 4.70, SD 0.48) and willingness to pay for the app the lowest (mean 3.11, SD 1.90). Perceived impact received the highest score across all domains (mean 4.82, SD 0.39), with behavior change scoring the highest (mean 5.82, SD 0) and awareness scoring the lowest (mean 4.60, SD 0.52). Qualitative findings revealed that PCPs found MGAT entertaining and interesting, but wanted more customization and interactive features to boost engagement. They appreciated its ease of use and navigation, but noted the need for a stable internet connection. While the design was visually appealing, improvements to the color scheme and element sizes were suggested. Participants valued the high-quality information relevant to older adults but desired more specialized content for medical professionals. They were likely to recommend MGAT, though cost opinions varied. Most importantly, MGAT increased awareness of patient needs, expanded AT knowledge, and positively influenced intentions to recommend AT, ultimately facilitating patient access to AT.

Conclusions: The high-quality and usefulness ratings suggest MGAT could be an effective tool for PCPs in managing older adults' FDs. Future research should evaluate the effectiveness of MGAT in managing FDs among older adults.

XSL•FO
RenderX

(JMIR Form Res 2025;9:e69645) doi:10.2196/69645

KEYWORDS

aged, assistive technology; Latino; mobile applications; physically disabled; primary care physicians; usability testing

Introduction

Assistive technology (AT) devices, such as jar openers, sock aids, and canes, play a significant role in enhancing the daily functioning, safety performance in daily activities, and social participation of older adults, ultimately contributing to their well-being and ability to live independently in their homes and community [1-6]. Despite the benefits of AT, a previous qualitative research study indicated that Latino populations, particularly older adults in Puerto Rico, have limited access to these devices due to a lack of awareness and information about the available AT options [7,8]. This access gap can lead to functional deterioration and adverse health outcomes, including increased hospitalizations, reduced quality of life, and higher health care costs [9,10]. Furthermore, primary care physicians (PCPs) in Puerto Rico, defined as licensed health care professionals who provide medical services in areas such as general medicine, internal medicine, family medicine, and geriatrics, may also lack awareness of the AT resources available to their patients. This is critical since information on acquiring AT is a key factor influencing its usage among older adults [11].

PCPs serve as the first point of contact in the health care system and play a crucial role in addressing the functional limitations faced by older adults with chronic diseases. Since functional ability is essential for healthy aging [12], it is crucial for PCPs to have comprehensive knowledge of AT devices and services. This knowledge enables PCPs to effectively support older adults in maintaining their function, independence, and quality of life, which are essential to healthy aging. In addition, it allows PCPs to make informed decisions about their patients' needs and facilitate access to appropriate resources [11].

The web application, Mi Guía de Asistencia Tecnológica (MGAT; My Assistive Technology Guide), was previously developed to address the information gap about AT for older Hispanics in Puerto Rico [13]. The content of the MGAT was determined based on the evidence of the need for AT devices in older Latinos living in Puerto Rico obtained from 2 quantitative descriptive studies [14,15]. The first quantitative study was initially conducted with a purposive sample of 60 independent-living Hispanic adults aged 70 years or older, residing in urban and rural areas of Puerto Rico. These individuals had physical functional disabilities (FDs) but no cognitive impairments [14]. Participants reported on approximately 50 AT devices for activities of daily living (ADLs) and instrumental activities of daily living, categorizing them into devices they used and those they did not use but were willing to consider using. A second cross-sectional study involved 211 independently living older Latinos aged 65 years or older, without cognitive impairments or FDs, who were randomly recruited from low-income communities in San Juan, Puerto Rico [15]. This study identified the most common physical disabilities using the Patient-Reported Outcomes Measurement Information System Physical Function Short Form. In addition, 2 qualitative studies, one with the purposive sample [14] and another with the randomly selected sample [7], explored barriers to AT device usage among 83 older Latino adults. Modifiable barriers identified in these studies informed both the selection of AT devices and the associated educational content provided for each device.

MGAT offers user-friendly access to detailed information about 94 AT devices in 8 categories of daily activities: mobility, self-care, bathing, dressing, food preparation, home management, medication management, and home safety. It incorporates visual aids and simple navigation to enhance its usability in providing information about AT devices to a variety of users.

MGAT has been studied to evaluate its quality and explore older Latinos with physical disabilities' experiences with it [16]. Using a mixed methods design, the study involved 12 participants who were trained to use the MGAT, then used it for 30 days, and finally provided feedback through quantitative ratings obtained from the Spanish version of the User Mobile Application Rating Scale (uMARS) [17] and semistructured interviews. The results showed that the MGAT was highly rated for both objective and subjective quality, with means of 3.99 and 4.13 (out of a maximum of 5), respectively. Qualitative findings indicated that the MGAT was accessible, usable, desirable, credible, useful, and valuable for increasing older Latinos' knowledge and autonomy regarding AT. Overall, the MGAT showed potential to improve AT awareness and adoption among older adults, addressing a significant barrier to AT use. This shows promise as a valuable resource for PCPs in managing FDs among older adults, but further research is needed to understand their perspectives on its usability and integration into clinical workflows. This gap highlights the importance of assessing PCPs' experiences with the MGAT, as they are the primary point of contact for older adults and play a critical role in promoting access to AT.

A mixed method research design that combines quantitative assessments using tools such as the uMARS [17] with qualitative insights from user interviews provides a comprehensive understanding of user experiences and perceptions [18]. This approach captures both objective usability issues and subjective perceptions to specifically address PCPs' unique needs around clinical decision support in AT. Mixed methods will provide comprehensive actionable insights that neither quantitative nor qualitative approaches alone could deliver, resulting in an app that meets both technical usability standards and seamlessly integrates into PCPs' demanding clinical workflows. Furthermore, this integrated approach not only identifies strengths but also highlights areas for improvement, ensuring that health applications like MGAT effectively support both health care providers and their patients in accessing AT resources.

XSL•FO
RenderX

This study aimed to assess the quality of the MGAT among PCPs and describe their experiences using it in a real-world setting in supporting consultations with older adults. By providing culturally relevant information on AT options, MGAT can help PCPs assess patients' needs more effectively and educate them about available solutions. However, evaluating the quality and usability of such health applications before widespread implementation is crucial to guide health care professionals in making informed recommendations [19,20].

Methods

Study Design

To assess the quality and user experiences with the MGAT, 10 PCPs were encouraged to use the web application for 30 days. At the end of this usage period, the researchers used a concurrent mixed method design, which consisted of simultaneously collecting quantitative and qualitative data within a single study phase (see Figure 1). We selected this design to achieve the study's aims, allowing the qualitative phase to enrich the quantitative results and thereby providing a comprehensive understanding of the quality and user experiences with the MGAT web application from the PCPs' perspectives. In this study, quantitative data were obtained measuring the quality of the MGAT web application using the uMARS tool [17]. Qualitative data were collected through semi-structured interviews conducted concurrently after participants responded to each item of the uMARS.

Figure 1. Integrated concurrent mixed methods design for MGAT (Mi Guía de Asistencia Tecnológica, My Assistive Technology Guide) evaluation.

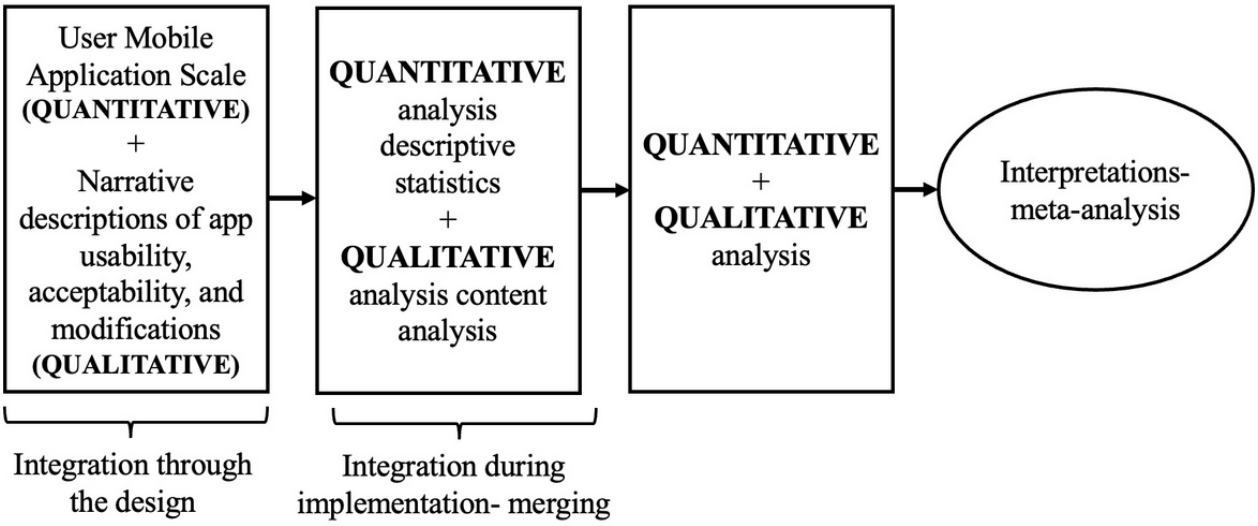

Integration of mixed methods occurred at 3 points, as described by Fetters et al (see Figure 1) [21]. First, integration through the design occurred during the data collection phase by collecting quantitative and qualitative data simultaneously in the same interview at the end of the MGAT web application trial period. The qualitative in-depth interview questions were informed by data provided in the quantitative section. Second, we used the merging approach for the integration during implementation. During the merging approach, we brought together quantitative and qualitative data to expand on the uMARS items rated with a 3 or less (disagree or strongly disagree). Finally, quantitative and qualitative data were integrated during interpretation through meta-inference using a narrative integration approach. This method allowed us to combine numerical data from the uMARS with thematic insights from interviews, providing a comprehensive understanding of the study's findings. By merging these data types, we gained a richer perspective on the usability and quality of the MGAT web application.

Karagianni's optimized honeycomb model [22], originally developed by Morville [23], was the theoretical perspective that guided data collection, analysis, interpretation, and discussion of the study results. The model was chosen due to its applicability to analyzing the user experience perspective during the application's design. It consists of 7 facets used to analyze the web application user experience perspective: findable, accessible, usable, desirable, credible, useful, and valuable. These facets are grouped into 3 parts, reasoning about how the user "Feel, Think and Use" the product by connecting the 7 facets.

Intervention

The MGAT web application had been developed in a previous study to increase community-living older Latinos' access to information about AT devices and services that could increase this population's functional performance in everyday activities [13]. The content included in the MGAT had been determined based on the following: (1) evidence of AT device needs of older Latinos living in Puerto Rico found in previous studies [14,15]; (2) the Matching Person and Technology Model [24], which considered the person, technology, and environmental factors found in previous studies that influenced the use of AT among older Latinos from poor communities (eg, AT with aesthetic appearance [umbrella cane] for overcoming the stigma associated with the use of mobility devices); and (3) evidence-based guidelines and recommendations for designing health applications for older adults reported recently in the literature [25-29]. The MGAT had been designed with two fundamental characteristics (see Figure 2). First, the app was

XSL•FO
RenderX

designed to be simple to use, regardless of the user's familiarity with mobile apps, making it accessible to everyone. Second, it allowed the user to effectively navigate the MGAT through 8 different areas of ADL to finally learn about the available AT devices used to compensate for the FDs that older adults might have in each daily living activity. The user could explore AT devices within 8 broad areas of ADL in which older adults might exhibit functional difficulties using buttons with photos and text to represent each area (mobility, self-care, bathing, dressing, meal preparation, home management, and medication management). Each of these areas contained subcategories of activities of each activity (see Table 1).

Figure 2. Screenshots of the MGAT (Mi Guía de Asistencia Tecnológica, My Assistive Technology Guide) web application.

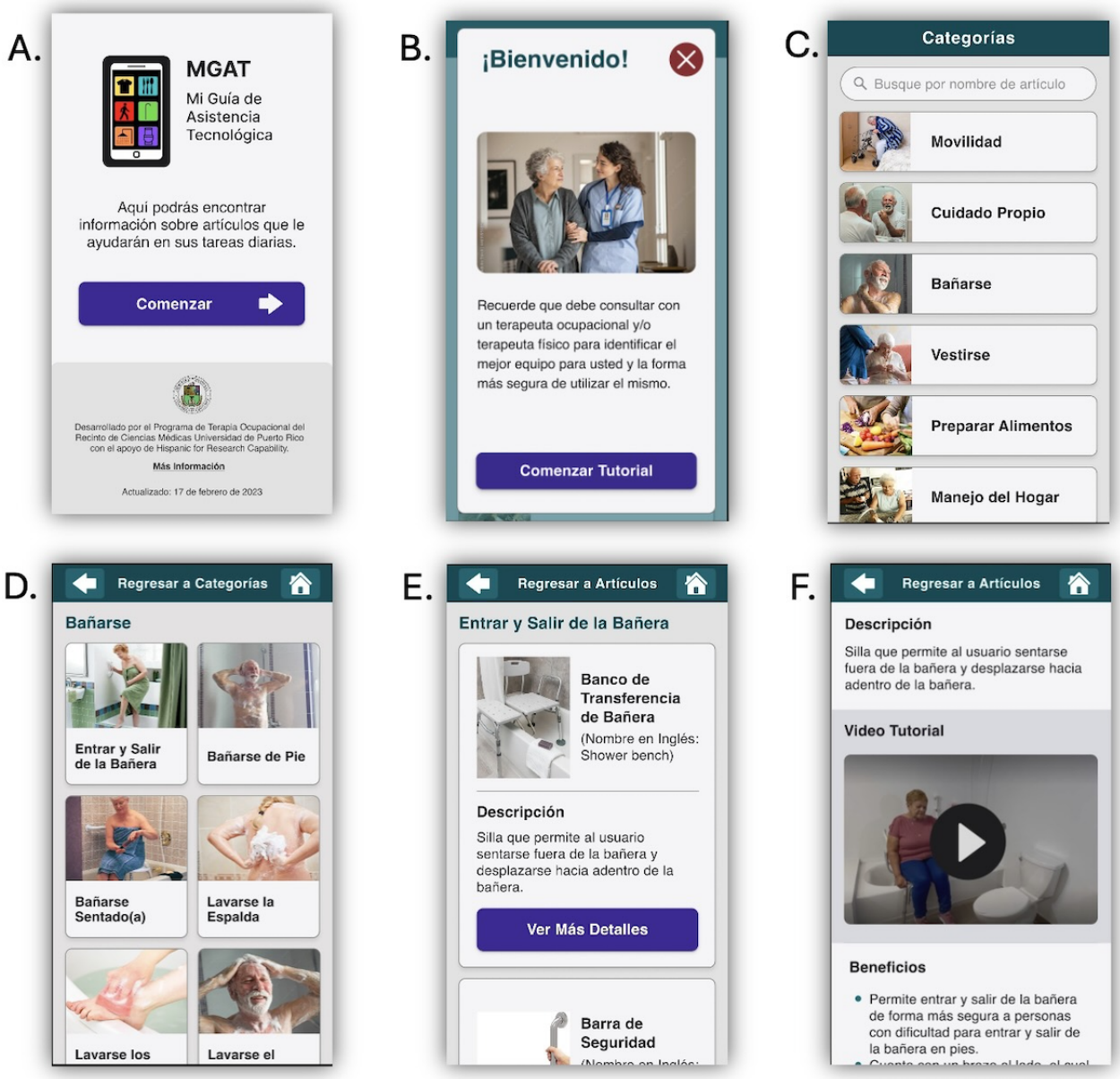

XSL•FO
RenderX

Table 1. Categories and subcategories of the daily activities included in the MGAT web application.

Categories and subcategories	Assistive technology devices included in the web application (N=94), n (%)
Mobility	
Getting in and out of bed	2 (2)
Sitting and standing from chairs	2 (2)
Walking	11 (13)
Getting in and out of a car	5 (5)
Sitting and standing from the toilet	6 (6)
Self-care	
Brushing teeth	3 (3)
Cutting fingernails	2 (2)
Cutting toe nails	2 (2)
Bathing	
Getting in and out of the shower	3 (3)
Bathing standing up	3 (3)
Bathing sitting down	3 (3)
Washing back	1 (1)
Washing feet	2 (2)
Washing hair	1 (1)
Dressing	
Putting on and taking off a shirt	1 (1)
Buttoning shirt	1 (1)
Putting on and taking off pants	2 (2)
Putting on and taking off socks	4 (4)
Putting on shoes	3 (3)
Meal preparation	
Using utensils	2 (2)
Pouring liquids	2 (2)
Cutting food	2 (2)
Opening jars	6 (6)
Opening cans	4 (4)
Opening bottles	2 (2)
Home management	
Reaching objects	2 (2)
Carrying objects	3 (3)
Cleaning bathroom	1 (1)
Cleaning floor	2 (2)
Opening doors	1 (1)
Medication management	
Opening medication containers	2 (2)
Cutting pills	1 (1)
Crushing pills	1 (1)
Organizing pills	2 (2)
Home safety	

XSL•FO
RenderX

Categories and subcategories	Assistive technology devices included in the web application (N=94), n (%)
Walking over carpets	2 (2)
Walking in dark areas	1 (1)
Medical alert systems	1 (1)

The MGAT web application features several key screens: an introductory screen that leads to the home page; a tutorial screen designed to guide users; a screen presenting categories of daily activities; a screen that lists specific activities within each selected category; a screen showcasing various assistive technology devices; and a final screen offering detailed information and instructional videos about these assistive technology devices.

In addition, the MGAT features instructional videos demonstrating older adults using AT devices. As an informational web application, the MGAT lacks interactive functionalities. It is designed to be compatible with any mobile device to enhance free access for users. While the MGAT offers valuable insights into AT devices, it does not guide the selection of the most appropriate device based on individual requirements. Consequently, all videos include a disclaimer: "The equipment shown in this video is only demonstrative. This video is solely for educational purposes. We recommend that you consult with your occupational therapist to evaluate your need for assistive technology devices."

Participants, Sampling, and Recruitment

Three researchers (an occupational therapist, a doctoral student in medicine, and a biology bachelor's degree student) received 24-hour training from the principal investigator to recruit, collect, and analyze data. The principal investigator is an occupational therapist with extensive experience conducting mixed method study designs. These researchers recruited a purposive sample of 10 PCPs who deliver services to older adults with physical FDs, following the sampling guidelines proposed by Hwang and Salvendy [30]. Their "10 ± 2 rule," based on a meta-analysis of usability studies, establishes this range as optimal for user testing, balancing statistical power with practical feasibility

This selection is further supported by Nielsen and Landauer's [31] problem discovery model, which demonstrates that a sample of 10 participants typically identifies 95% of usability issues. In health care-specific usability testing, Kushniruk and Patel [32] found that a participant range of 8-12 is sufficient for detecting major usability concerns in clinical information systems. Our study methodology incorporated task-based scenarios aligning with Kushniruk and Patel's [32] validated approaches for health care interface evaluation.

Furthermore, our findings demonstrated saturation in issue discovery, with no new critical usability concerns emerging in the final 2 participants, empirically confirming the adequacy of our sample size within this specific context. While larger samples may be required for broader qualitative research objectives, the chosen sample size effectively balances resource efficiency with comprehensive problem identification, ensuring a robust evaluation of the user interface.

Study sampling was carried out using purposive sampling, which involved an interactive process of selecting research subjects (based on their ability to elucidate a specific theme, concept, or phenomenon, in this case, the usability and acceptability of the AT web application). Individuals were recruited until the researchers perceived that no new themes or subthemes were emerging (ie, data saturation), a standard approach in qualitative research [33].

The inclusion criteria were (1) adults 21 years and older, and (2) being a PCP who provided health care to older Latinos with physical FD. PCPs could include general practitioners, family practice physicians, internal medicine physicians, or geriatricians. The exclusion criteria were PCPs who were 20 years or younger and those who did not provide health care services to older Latinos with physical function disabilities.

A total of 2 recruitment methods were used: (1) direct contact through PCPs' telephone numbers in Puerto Rico, obtained from publicly available internet-based contact information, and (2) snowball sampling procedures. During direct contact, researchers with experience in occupational therapy and general medicine called potential participants and provided the PCPs with a complete explanation of the purpose of the study, the study procedures, and that their participation in the study was completely voluntary. If the PCP expressed interest in participating in this study, the researchers assessed if the PCP met the 2 inclusion criteria and proceeded, with eligible participants, to schedule their convenient day, time, and place to conduct the consent form procedures. On the day of the consent form procedures, the researchers first met individually with each PCP in a private place at their preferred location (such as the clinic) to provide them with a complete explanation of the consent form, including the purpose, procedures, risks, and benefits of the study. The researchers gave PCPs the opportunity to ask any questions related to this study. After addressing all PCP concerns related to this study, the researchers asked them to sign the consent form and begin their participation in this study.

Data Collection Procedures

The researchers scheduled a baseline meeting with each PCP to fill out a sociodemographic questionnaire and provide them with training in using the MGAT. The download of the app took approximately 2 minutes, and the MGAT training took approximately 15 minutes. PCPs were trained using verbal instructions, demonstrations, and hands-on practice until they demonstrated competence in using the web application. Training ensured engagement and skill retention through real-time navigation, guided practice, and competency-based assessment. PCPs were required to demonstrate their ability to navigate and use the web application effectively. This included completing specific tasks and scenarios using the MGAT, such as identifying AT devices to compensate for difficulties in different

XSL•FO
RenderX

daily activities. Researchers conducted informal assessments during the training sessions to evaluate PCPs' understanding of the MGAT's functionalities and their ability to apply it in clinical settings. All PCPs had to successfully complete key tasks before integrating the web application into their practice. This structured approach ensured they were fully prepared to use the app effectively in patient care.

Afterward, PCPs were encouraged to use the MGAT app for 30 days while providing health care services to older adults with physical function difficulties performing ADL. At the end of the web application usage period, the researchers scheduled an individual face-to-face or videoconferencing meeting with each clinician in their preferred location and convenient time to collect quantitative data using the uMARS and qualitative data through in-depth semistructured interviews on the experience, usage, and acceptability of the MGAT. During the interview, the researchers also conducted follow-up questions concerning the uMARS items the PCPs rated with a score of three or less (some disagreement). The principal investigator (PI), an expert in mixed methods design, provided 8 hours of training to all members of the research team in conducting PCP recruitment and administration of the study measures. The data collection process was completed in one face-to-face or videoconferencing meeting. With the consent of PCPs the interviews were audio recorded with the consent of the PCPs, and field notes were taken to capture further details. The researchers produced verbatim transcripts of the audio-recorded interviews for subsequent analysis.

Assessment Tools

All assessment tools were conducted in Spanish and administered by the researchers of this study using an interview format on the convenient day of the PCP and the preferred location under the supervision of the PI after data collection training.

Sociodemographic Data Questionnaire

The team developed this questionnaire to collect the self-reported sociodemographic data of the following PCPs: (1) age, (2) sex, (3) medical specialty, (4) years of experience working with older adults with FDs, (5) average of older adults with FD attended monthly, (6) work setting, (7) frequency of app usage for the provision of health care services, (8) medical practice identifying and addressing FDs of older adults' FDs, and (9) types of assistive devices recommended to older adults with FD. This measure took approximately 10 minutes to complete and was administered before the MGAT usage trial period.

uMARS Spanish Version

The uMARS is a 29-item scale that provides comprehensive ratings of user experience and impressions of the mobile health application by assessing the quality of the application (objective and subjective) and perceived impact [34]. Each item has a custom wording that is appropriate for the aspect being assessed. Items use a common 5-point rating scale from 1 (inadequate) to 5 (excellent), such that higher scores represent a stronger impact of the app on that aspect of user cognition or potential behavior. The subjective quality and perceived impact of the MGAT were assessed on an individual basis per item. The English version of uMARS was cross-culturally adapted for the Spanish language using a 3-process design: cross-cultural adaptation, translation, and evaluation of statistical reliability and validity [17].

Interview Guide

The researchers in this study developed this guide to facilitate the discussion and collection of qualitative data from PCPs necessary to further assess the usability, acceptability, and modifications of the MGAT web application. The guide used open questions to obtain participants' feedback on their experience using the MGAT and their recommendations to improve its usefulness for PCPs. Interview questions included the following: "How adept do you think you are at using the app? Did you experience any difficulty when using the app? How did you cope with these issues? What are the strengths and weaknesses of the app? How can you improve the app? How useful is this app in your health service delivery system to older adults with FDs? Will you use this app as part of your health care delivery services? Why?" Furthermore, follow-up questions were asked to explain those items of the uMARS with a score of 3 or less. The interviews took an average of 20 minutes and were administered at the end of the MGAT usage trial period.

Data Analysis

Quantitative Data

Quantitative data from the sociodemographic questionnaire and uMARS were analyzed using descriptive statistics of the central tendency, presenting means (SDs) for continuous variables and frequencies with percentages for categorical variables. Microsoft Excel 2019 facilitated these statistical analyses.

Qualitative Data

A directed thematic content analysis was used to analyze the qualitative data obtained from semistructured interviews [35]. This directed qualitative content analysis was informed by the optimized honeycomb model to systematically analyze textual data through the predefined scales of uMARS. The analysis began with clear operational definitions for the objective quality, subjective quality, and perceived impact scales. The researchers deeply engaged with the data, identifying and highlighting text segments relevant to each scale and coding them accordingly. The research team held regular meetings to collaboratively interpret the findings. Data management was supported by QDA Miner Lite (version 3.0.2; Provalis Research), a qualitative data analysis software package.

To ensure rigor in the analysis, peer debriefing sessions were conducted with two researchers for each participant. In addition, an external audit was performed by a qualitative data analysis expert (the PI). This audit involved a comprehensive review of all interviews to confirm the interpretations and analyses carried out by the researchers.

Integration Through Analysis and Interpretation

After completing qualitative and quantitative analyses, the qualitative insights of individual interviews were merged with the quantitative findings of uMARS for a comprehensive

XSL•FO
RenderX

evaluation and comparison. A narrative integration approach was used, interweaving qualitative and quantitative results on a theme-by-theme or concept-by-concept basis to present a more comprehensive understanding of the experiences of the participants. Both quantitative and qualitative analyses were conducted in Spanish, with specific quotes chosen to illustrate each theme translated into English by a bilingual researcher for inclusion in the paper. This method ensures that the original meaning and context of the quotes are maintained during translation while making the findings accessible to an English-speaking audience. This version maintains the original meaning while using different phrasing and structure.

Ethical Considerations

The study was approved by the Institutional Review Board of the University of Puerto Rico, Medical Sciences Campus (protocol 2211060727). The study procedures complied with the Declaration of Helsinki. All PCP participants provided written informed consent before their enrollment in the study. The consent form detailed the study's purpose, procedures, potential risks and benefits, the voluntary nature of participation, and the right to withdraw at any time without penalty. Participants were given ample opportunity to ask questions and have their concerns addressed before signing the consent form. PCPs received compensation for their time and effort invested in data collection. Each PCP was compensated US $50 for their time invested in data collection at the beginning of the study (during the first visit) and US $50 at the end of the 30-day trial period of the MGAT web application, for a total of US $100. The images of older adults displayed within the MGAT web application were sourced from publicly available images obtained from Google. Since MGAT is a prototype developed solely for nonprofit research purposes, these images are used under the fair use doctrine, which permits using copyrighted materials without obtaining permission from the copyright owner for noncommercial research activities.

Results

Sample Characteristics

A total of 13 eligible participants were approached. Three participants refused to participate due to time constraints. A total of 10 participants completed the semistructured interviews and surveys after the trial period of using the MGAT web application. The trial period ranged from 34 to 63 days among participants. Table 2 describes the characteristics of the participants, illustrating a diverse group of primary health care providers with a mean age of 45.7 (SD 14.2) years, ranging from 29 to 62 years. Most participants were women (n=6, 60%). Regarding medical specialties, most of the participants were geriatricians or fellows (n=6, 60%), followed by those in internal medicine (n=3, 30%). The participants had an average of 17.3 (SD 13.7) years of service to older adults, ranging from 2 to 33 years. On average, PCPs provided medical services to 42.5 (SD 40.2) older adults with FDs per month. The distribution in the workplace showed that most PCPs worked in private clinics (n=7, 70%), university hospitals (n=3, 30%), and assisted living facilities (n=3, 30%).

XSL•FO
RenderX

Table 2. Sociodemographic characteristics of study participants.

Participants characteristics	Total (N=10)
Age (years)	
Mean (SD)	45.7 (14.2)
Range	29-67
Sex, n (%)	
Female	6 (60)
Male	4 (40)
Medical specialty	
Mean (SD)	3.93 (1.19)
Geriatrics or geriatric fellows, n (%)	6 (60)
Internal medicine, n (%)	3 (30)
Family medicine, n (%)	1(10)
Years of service to older adults	
Mean (SD)	17.3 (13.7)
Range	2-33
Older adult patients with FDs[a]seen by PCPs[b] per month	
Mean (SD)	42.5 (40.2)
Older adult patients with FDs seen by PCPs per month range (n)	20-150
Patients served using the MGAT[c] range (monthly)	0-30
Workplace, n (%)	
Private clinic	7 (70)
University hospital	5 (50)
Assisted living facility	3 (30)
University outpatient clinic	3 (30)
Primary health care clinic	1 (10)
Veteran hospital	1 (10)

[a]FD: functional disability.

[b]PCP: primary care physician.

[c]MGAT: Mi Guía de Asistencia Tecnológica (My Assistive Technology Guide).

Table 3 provides an overview of the self-reported characteristics of the study participants. It includes information on their sex, age, medical specialty, years of service to older adults with physical function disabilities, and their workplaces. The participants represent a diverse group in terms of sex, age, and professional background, working in various health care settings such as private clinics, university hospitals, outpatient clinics, primary health care clinics, assisted living facilities, and veteran hospitals. The years of service among participants range widely, reflecting a broad spectrum of experience in the field.

XSL•FO
RenderX

Table 3. A snapshot of participants' self-reported characteristics.

Participant ID (N=10)	Sex	Age (years)	Medical specialty	Years of service to older adults with physical function disabilities	Workplace
R1	Female	43	Internal medicine	15	University outpatient clinic
R2	Male	34	Internal medicine	3	Private clinic and primary health care clinic
R3	Female	54	Geriatrics	26	Private clinic, University hospital, and assisted living facility
R4	Male	67	Internal medicine	37	Private clinic and University hospital
R5	Female	29	Geriatric fellow	3	University outpatient clinic, University hospital, and veteran hospital
R6	Female	29	Geriatric fellow	3	University outpatient clinic and University hospital
R7	Female	62	Geriatrics	33	Private clinic
E2	Female	55	Geriatrics	25	Private clinic and University hospital
E3	Male	32	Geriatrics	2	Private clinic
W1	Male	52	Family medicine	26	Private clinic

Quantitative Findings From the uMARS

PCPs positively rated the objective quality of the MGAT web application, with a mean objective quality score of uMARS of 4.06 (SD 1.05; see Table 4). The quality, credibility, and quantity of information were rated the highest (mean 4.60, SD 0.51), while the engagement received the lowest rating (mean 3.34, SD 1.51). The subjective quality rating of MGAT was also moderately high (mean 3.93, SD 1.19), with the item "Recommend MGAT to others" rated highest (mean 4.70, SD 0.48) and the item "Willing to pay for the application" rated the lowest (mean 3.11, SD 1.90).

PCPs scored the highest on perceived impact of MGAT (mean 4.82, SD 0.39). The scores were similar in all aspects, indicating a consistent degree of high perceived impact of MGAT on PCP knowledge, attitudes, and intentions to use MGAT to provide AT services to their patients.

Table 4. Scores on the uMARS[a] scales and subscales for the MGAT web application.

uMARS scales and subscales	Score, mean (SD)
MGAT[b] objective quality	4.06 (1.05)
Engagement	3.34 (1.51)
Functionality	4.25 (0.71)
Aesthetic	4.27 (0.64)
Information	4.60 (0.51)
MGAT subjective quality	3.93 (1.19)
Recommend the app to others	4.70 (0.48)
Predicted frequency of use of the app in the next year	4.20 (0.92)
Willing to pay for the app	3.11 (1.90)
Overall MGAT rating	3.60 (0.70)
Perceived impact of MGAT	4.82 (0.39)
Awareness	4.60 (0.52)
Knowledge	4.80 (0.42)
Attitudes	4.90 (0.32)
Intention to change	4.90 (0.32)
Help-seeking	4.70 (0.48)
Behavior change	5.00 (0.00)

[a]uMARS: User Version of the Mobile Application Scale.

[b]MGAT: Mi Guía de Asistencia Tecnológica (My Assistive Technology Guide).

XSL•FO
RenderX

Qualitative Insights From Participant Feedback

In this section, a summary of the interview data results is reported, along with representative quotes that provide deeper insight into the reasons behind their scores on each uMARS scale: engagement, functionality, aesthetic, information, subjective quality, and perceived impact.

Engagement

PCPs agreed that the MGAT lacks clear customization options, which limited their engagement with the MGAT. This sentiment was expressed by a geriatrician who noted, “I didn’t make any changes to the MGAT because it doesn’t allow it” (E2). In addition, the MGAT was found to lack user interactivity, specifically in terms of input and feedback mechanisms. A geriatric fellow highlighted this issue, stating, “It does not allow me to interact with the application or provide feedback” (R6), which indicates a perceived gap in interactive features. Another geriatric fellow also pointed out the absence of notifications, commenting, “Regarding the alerts, I understood them as notifications, and that is not something the application is currently capable of doing” (R5).

However, all of the participants found the MGAT to be entertaining and interesting, and targeted to older Hispanics. A geriatrician stated, “I loved the part about where you can buy the equipment. I found the videos really cool, they made me smile. And the actors (older Hispanics) were professional” (E2). This PCP also found the MGAT videos appropriate to support older adults learning about AT use: “The application is very valuable because older adults learn with the demonstration.”

Functionality

PCPs found the application easy to use, learn, and navigate, as stated by a geriatrician: “The reality is that the application is quite easy to use. I think it is intuitive in that sense (R5).” However, some PCPs noted that scanning the QR code of the MGAT address was challenging for certain patients: “Many of the elderly, unfortunately, get frustrated when using the camera (on the phone) and cannot find the link. And pressing it is difficult for them (E3).” To address this challenge, an internal medicine physician (R2) applied various strategies, such as downloading the MGAT to patients’ cell phone home screens, teaching caregivers how to access and use the MGAT, or using the MGAT from the PCP computer screen.

Although performance received a slightly lower score, PCPs identified issues primarily related to the MGAT’s dependency on a reliable internet connection. This dependency resulted in slow logins or delays in loading MGAT screens and videos. As one internal medicine physician mentioned, “If there is no Internet, it doesn’t work. It has happened to me on the phone that it took me a long time to get help (from the MGAT) for the patient on the phone (R2).” This PCP addressed the internet connection issue by stating: “I downloaded the application on the computer and taught them how to navigate the page without having the problem of a bad signal.”

Aesthetics

Regarding aesthetics, the MGAT’s design was praised for its visual appeal. A geriatrician remarked, “I was surprised by the quality of the videos and the information (R5).” Another geriatrician appreciated the MGAT’s layout for diverse users by stating:

> *For me, it was a good way to divide it, doing it by needs, an easier way to access the options available, not just to a doctor but also for the patient to use it… The app has more options and a wider range of items, so if you were searching for a specific walker, you would be limiting yourself to that. In contrast, if you select that the limitation is mobility, you will see other options you might not have seen.* [R6]

However, participants provided several suggestions to enhance the MGAT’s aesthetics. Regarding visual appeal, a geriatrician suggested: “Maybe make it more attractive in terms of the application’s visibility… Maybe something more eye-catching in terms of aesthetics… more inviting to use. It looks a little bit analog” (E3). Another internal medicine physician recommended improving the MGAT color scheme: “I think sometimes you need to add more clarity... the contrast can be improved” (R4). In addition, 2 other PCPs recommended increasing the size of some of the MGAT’s elements, with a participant stating, “What could be done is to put the icons larger on the first page... because they need to see the image that is very good” (E3). In this regard, this geriatrician proposed the following strategy to improve visual access to the MGAT: “When you open it from the computer, it can be viewed more completely.”

Information Quality

The PCPs agreed that the quality of the information provided by the MGAT was high, as exemplified by this geriatrician’s comment: “I was surprised by the quality of the information” (R5). They also highlighted the relevance of the information to older adults. A family doctor stated, “It has much relevant information regarding what can be offered to patients to help them carry out their independent activities, as well as their safety and fall prevention” (W1). The positive reception of the MGAT by older adults further supports the relevance and appropriateness of the MGAT’s information for this demographic, as illustrated by this internal medicine physician's comment: “It helped patients clarify what equipment they need” (R1).

Regarding the quantity of information, the PCPs appreciated the conciseness of the MGAT, as it avoids overwhelming users with excessive text and facilitates ease of use and teaching. This was expressed by an internal medicine physician who said: “The good thing is that it’s short, meaning it’s not overwhelming to the point I would run out of time to show it or review it” (R1). PCPs also rated the comprehensiveness of the MGAT information highly. However, several PCPs suggested that the MGAT could benefit from more specialized information for medical professionals. Some recommended including specific *ICD-10 (International Statistical Classification of Diseases, Tenth Revision)* diagnostic codes from the Centers for Medicare & Medicaid Services (CMS), as mentioned by this geriatrician (E2):

> *Many problems we face are that when we place these orders for medical equipment so patients can benefit*

XSL•FO
RenderX

> *from their health plans, they ask us for specific codes. In other words, we have to dance to the tune of the health plan. In this regard, the application could also help us. For example, for a cane, what diagnosis code could be effective for getting it approved?* [E2]

Several PCPs also recommended adding the Healthcare Common Procedure Coding System codes for the devices covered by Medicare to the MGAT. A geriatrician expressed this idea by stating: "You should include, if possible, a more private section for the area of physicians that requires a subscription to access the part of the application that includes the codes (for AT devices)" (E3). Several PCPs also emphasized the importance of including links to CMS information to justify medical devices to insurance plans. A geriatrician stated:

> *It would be ideal if it took you to the CMS page because you know what? It tells you how to justify the equipment. As a geriatrician, I know how to justify it, but a general practitioner or an internist has no idea how to justify an electric wheelchair for someone who has the use of their hands but whose body does not function.* [R3]

In addition, 2 PCPs suggested that integrating a questionnaire within the MGAT could streamline the process for PCPs, making it more efficient and likely to increase the MGAT's usage. This was expressed by an internal medicine physician:

> *If the application provides a functional problem questionnaire linked to device recommendations, it will help increase the application's usage compared to a primary doctor simply asking random questions. This way, the doctor won't have to return to the application or search for another educational tool to make recommendations. Having both functions in one place is much better because it reduces the time needed for the task.* [R1]

Some PCPs also expressed the need for the MGAT to facilitate the transfer of information to patients by sending or printing the information directly from the web application: "Another thing is if it were possible to send a link or print information from the application for the patient to take with them" (R1).

Regarding visual information, participants agreed that the overall clarity of the MGAT photos and videos is satisfactory, as noted by a geriatrician: "I think the photos look clear" (E3). However, to improve the clarity of the visuals for older adults, this participant highlighted the strategy of accessing the MGAT on a computer rather than a mobile device.

A recurring theme was the need for audio instructions to accompany visual content. An internal medicine physician expressed that "The MGAT videos need audio to show them (older adults) how to do it (use the AT device)" (R2), suggesting that auditory cues could enhance older adults' understanding and retention of information. To further improve the visual explanation in the videos, a PCP also recommended adding subtitles, in addition to audio, stating, "not only narrative but also include the subtitles" (E3).

Subjective Quality

PCPs indicated a strong endorsement of the usability of MGAT, evidenced by a high likelihood of recommending it to other physicians. For example, a geriatrician considers the MGAT a valuable tool that can significantly benefit patients, implying a high probability of suggesting it to fellow doctors:

> *This app has great value because many times we are the first line of contact with the patient who truly needs it… so the more doctors who can have it, the more accessible assistive technologies will become to the community.* [E3]

Another internal medicine physician suggested that the MGAT benefits other medical specialists as well: "I believe that this is not only directed at primary care doctors but also at physiatrists and rheumatologists" (R4).

Similarly, the willingness of PCPs to use MGAT in the next 12 months, despite time constraints, suggests that the participants view the web application as a useful tool. An internal medicine physician stated, "I put the number 3 to 10 (times that I would use MGAT in the next 12 months) because I depend a lot on time. Sometimes, if I'm busy, I forget to use it" (R2).

The data showed mixed views on paying for the MGAT. Although some PCPs are willing to consider paying for it, others expressed concerns about the cost, particularly from the perspective of older Hispanics who might find it difficult to afford. A geriatrician pointed out: "The ideal situation would be for the application to be free so patients could access it" (E2). This feedback correlates with a lower mean rating of 3.2.

Perceived Impact

PCPs reported that the MGAT has significantly heightened their awareness of the functional needs of their patients' daily activities. An internal medicine physician mentioned:

> *Using the app and being able to show them to the patients has helped me become a little more aware and attentive to ask them more specific questions about their daily lives rather than if they have fallen or are wetting themselves. One asks about things that affect their daily activities, but not necessarily about whether they have problems brushing their teeth, bathing, dressing, putting on a shirt, and with this (the MGAT), I would come and say, 'Look, to button your shirt, this device helps you,' and I would show them the buttoner.* [R2]

All PCPs also expressed that, while they were previously familiar with common AT devices like walkers and canes, the MGAT has expanded their knowledge of other useful devices. A geriatrician commented:

> *Even though we doctors know these items exist, we are not 'up to date' with many of them. So, I learned about many items that I didn't know existed, and that could help my patients in their daily tasks. Thus, the MGAT helps us to give our patients more options.* [R6]

PCPs also expressed that the MGAT has significantly increased their attitudes and intentions to facilitate access to AT for their

XSL•FO
RenderX

patients. A geriatrician emphasized that the application is highly valuable because it makes them a crucial point of contact for patients needing AT. He stated, “Many times, we are the first line of contact with the patient who really needs it... the more doctors have it, the more accessible AT will be to the community” (E3).

The use of the MGAT also prompted changes in all PCPs’ behaviors regarding informing, demonstrating, and prescribing AT devices to their patients or informing their family members. PCPs found it beneficial to use the MGAT during consultations to visually demonstrate the AT devices, which helped make informed decisions about what might be useful for the patients. As a participant stated:

> *I showed them how to access it from their cell phone, but to speed up the interaction during the visit, I would say, “When you tap here on the cell phone, this will appear.” I would select any device and usually explain that there would be a description of the different devices they could use. Then, I would show them the video. For some, I would download the app to their cell phones and show them where they could find the devices. I would tell them that if they needed it, I could order it for them, and always mentioned that for the next time (follow-up visit), if they found something they needed, I could order it for them.* [E3]

A geriatrician talked about the value of the MGAT in increasing her patients’ quality of life.

> *My job is not only about health but also about quality of life. For them to be able to do gardening with something they wouldn’t consider… such as a gardening chair, and for me to be able to say, “Look, you can do it with this device,” it continues to give them a reason to be alive. … not just to be independent, but to enjoy life. Now with the device, they can dress themselves and go to the opera. … So, having things that give them independence is like an antidepressant.* [R3]

However, the limited time available during consultations was a significant barrier to using the MGAT. The high volume of patients and the need to prioritize immediate medical concerns often left little time for detailed discussions about assistive technology. A geriatrician mentioned:

> *In medical appointments, we sometimes have many patients scheduled, and although I would like to show them the video and all the assistive devices' instructions, I didn't have much time with them then. But I had already had the opportunity to see them beforehand, so I knew to whom it applied and to whom it didn't, so I gave them a brief summary of how to use it.* [R6]

PCPs provided several strategies to manage the time barrier associated with using the MGAT. One strategy was to provide patients access to the MGAT before the medical appointment, allowing them to familiarize themselves with the AT devices and identify their needs during the appointment. Another approach was introducing the MGAT to patients during the medical visit and asking them to identify their specific AT needs for the next follow-up appointment. Finally, several PCPs recommended including a screening of functional abilities in daily activities within the MGAT that could be completed by the patient, their family member, the receptionist, the clinic nurse, or the physician themselves as a strategy for more efficient use of the MGAT. For example, an internal medicine physician explained:

> *If the application provides you with a questionnaire on functional problems, where one marks the different issues the patient has, which is then linked to a recommendation for a type of assistive device, it will help increase the use of the application instead of a primary care doctor simply asking random questions... If both things are in one place, that's much better because you reduce the working time.* [R1]

Integration of Quantitative and Qualitative Data

Integrating uMARS ratings with qualitative feedback provides a comprehensive understanding of user experiences. For example, the low engagement score correlates with participant comments emphasizing the low engagement score correlates with participant comments that emphasize the lack of customization and interactive elements in the MGAT. Conversely, the higher functionality score aligns with user opinions regarding how easy the MGAT is to use, learn, and navigate.

Furthermore, integrating the uMARS Perceived Impact scores with qualitative feedback provides a comprehensive understanding of the app’s potential to influence knowledge, awareness, and behavior. The high mean score of 4.1 aligns with participants’ comments about increased knowledge and motivation for change.

However, the qualitative data also highlight the need for additional features and support to overcome barriers to sustained behavior change. This insight suggests areas for improvement in the app’s design and functionality to enhance its long-term impact on users’ health behaviors.

Discussion

Principal Findings

The primary objective of this study was to assess the quality of the MGAT web application among PCPs and to describe their experiences using it to facilitate access to AT services for older adults. This research is significant as it addresses a critical gap in the provision of AT services for older adults, particularly within Latino populations in Puerto Rico, who often face barriers to accessing necessary devices due to a lack of information and resources [7]. By equipping PCPs with this knowledge, the study aims to improve the provision of AT services, ultimately benefiting the daily functioning, well-being, and quality of life of older adults.

The results indicated that PCPs rated MGAT positively, with a high or moderately high mean score on all dimensions of uMARS, reflecting high perceived quality, usability, and impact

XSL•FO
RenderX

on their practice. While the MGAT provided high-quality content, engagement scores were influenced by the low ratings in the uMARS customization item (mean 1.40, SD 0.70) and interactive item (mean 2.20, SD 1.03). The absence of customization options and interactive elements, such as user input, feedback mechanisms, and notifications, was perceived as a limitation. The implications of the lower engagement score for the overall user experience of the MGAT suggest potential areas for improvement in user interaction and customization design. Interactive enhancements, such as allowing users to provide input, receive feedback, and receive notifications about new AT devices, along with customization features like selecting the most commonly prescribed AT devices, can improve engagement while maintaining high information quality.

Despite the overall perceived high quality of the MGAT, several key challenges emerged that impacted its usability, including reliance on a stable internet connection, the absence of audio descriptions for instructional videos on AT use, and the lack of current procedural terminology codes for AT devices, along with their medical justification. These challenges affect PCPs' ability to effectively use the application, as they may need to employ additional strategies to assist their patients, such as using the MGAT from a computer screen with a wired connection to the internet or spending extra time describing the use of the AT devices to the patients.

The variation in the use of the web application, ranging from 0 to 30 patients in a month among the 10 participating PCPs, highlights differences in engagement and feasibility within clinical workflows. This disparity suggests that while some PCPs were able to integrate the web application consistently into their practice, others faced barriers such as time constraints, competing clinical demands, or varying levels of familiarity with AT tools. Notably, one PCP did not use the MGAT web application at all, which underscores potential challenges in adoption.

These differences in usage may impact the generalizability of the findings, as the benefits and usability perceptions of the web application may be more reflective of those who used it frequently. Furthermore, the wide range in usage suggests that implementation strategies should consider tailored training, workflow integration support, and additional usability enhancements to encourage more consistent adoption among PCPs.

Comparison With Previous Work

The findings of this study are consistent with existing literature, which shows that health care providers are more likely to adopt technology that is intuitive, easy to navigate, perceived as useful, and encourages them to use it [36]. Similar to previous studies, this research highlights the need for user-friendly designs to ensure the effective use of health technology [37]. However, our study contributes new insights by highlighting specific areas for improvement, such as improving engagement and addressing technical issues, which are less frequently discussed in the literature.

Although there is a web-based AT tool, other than the MGAT, to improve access to AT information among people with disabilities—the Atvisor [38]—both tools represent 2 distinct approaches, each catering to the needs of different users. MGAT is specifically designed for older Latinos in Puerto Rico, providing user-friendly access to information on AT devices in various daily activities. Its simplicity and visual aids make it particularly accessible to older adults, enhancing awareness among PCPs about functional needs and available solutions. In contrast, Atvisor.ai is a clinical decision support system aimed at AT professionals and users in Israel, allowing personalized assessments based on individual needs and goals. This platform incorporates advanced features such as a matching algorithm to connect users with suitable products, facilitating online purchases and long-term follow-up. However, the strengths of MGAT lie in its ease of use and culturally relevant content, which is particularly beneficial for older adults who may be less familiar with technology.

Our observations regarding the differences between MGAT and Atvisor are based on an experiential, hands-on comparison, rather than an empirical research study. While MGAT's design prioritizes simplicity for PCPs requiring immediate access to AT information without complex decision-support features, Atvisor offers a more intricate system for environments where detailed evaluations and ongoing support are necessary. To avoid overclaiming, we acknowledge the need for future comparative studies to validate these distinctions empirically, particularly in terms of usability, user workload, and task efficiency.

This study holds significant value for PCPs who work with older adults. By providing a comprehensive resource for AT options, the MGAT can enhance patient care by allowing physicians to make informed recommendations tailored to their patients' functional needs, thus boosting the independence and quality of life of older adults. PCP recommendations for app improvements, such as integrating diagnostic codes and adding audio instructions, highlight areas where further development could increase the adoption and utility rates among health care providers.

Limitations and Future Research

This study acknowledges several limitations that may affect the interpretation of the results. The sample size was relatively small and consisted primarily of geriatricians, which may not fully represent the broader spectrum of PCPs. Furthermore, the study was conducted in a specific geographical and cultural context, which may limit the generalizability of the findings. The use of purposive sampling and snowball methods may introduce bias. These methods, while effective for identifying participants with specific experiences, can lead to a sample that is not representative of the broader population. We acknowledge this potential bias as a limitation of our study and suggest that future research consider more diverse sampling strategies to enhance generalizability. Future efforts should explore strategies to streamline app use within time-constrained environments, such as optimizing interface efficiency or integrating the tool into existing electronic health record systems. Future studies should also use a more diverse sample of health care providers with varied workload constraints to explore the usability of the app in different settings. Future iterations of the MGAT should

XSL•FO
RenderX

focus on addressing the technical challenges identified in this study and incorporate user-centered design approaches to optimize both usability and engagement, ensuring a more holistic user experience. Further studies could also explore the long-term impact of MGAT on functional outcomes of patients and investigate the effectiveness of the app in other cultural contexts. Finally, research could examine the integration of MGAT with other disability services in primary care to provide a more comprehensive solution for the management of FDs among older adults.

Conclusions

PCPs positively rated the MGAT web application, with a high perceived impact on their practice. MGAT shows promise as a valuable tool for PCPs in Puerto Rico, offering comprehensive and culturally relevant information on AT devices. Although technical challenges must be addressed, the positive reception of MGAT highlights its potential to improve the provision of assistive technology services for older adults. By improving PCP knowledge and awareness of AT options, MGAT can contribute to better patient care and improved outcomes for function, health, and quality of life for older adults.

Acknowledgments

This research was supported by the US Department of Education: Award S21MD001830 and partially supported by the Hispanics-In-Research Capability (HiREC)—National Institutes of Minority Health and Health Disparities (NIMHD) under Award Number S21MD001830, Center for Collaborative Research in Health Disparities (CCRHD-RCMI [U54-MD007600—NIMHD, and The Hispanic Alliance for Clinical and Translational Research (Alliance) supported by the National Institute of General Medical Sciences (NIGMS) National Institutes of Health under Award Number U54GM133807.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: EMO-C, WB-D, RR-O, JM-C
Methodology: EMO-C
Formal analysis: EMO-C, WB-D, RR-O, JM-C
Investigation (data gathering): WB-D, RR-O, JM-C
Writing – original draft: EMO-C
Writing – review & editing: AR-L

Conflicts of Interest

None declared.

References

1. Ben Mortenson W, Demers L, Fuhrer MJ, Jutai JW, Bilkey J, Plante M, et al. Effects of a caregiver-inclusive assistive technology intervention: a randomized controlled trial. BMC Geriatr 2018;18(1):97 [FREE Full text] [doi: 10.1186/s12877-018-0783-6] [Medline: 29669536]
2. Stanley R. Technology supports for community-dwelling frail older adults. ar 2015;6(1):41-49 [FREE Full text] [doi: 10.18357/ar.stanleyr.612015]
3. Yusif S, Soar J, Hafeez-Baig A. Older people, assistive technologies, and the barriers to adoption: A systematic review. Int J Med Inform 2016;94:112-116. [doi: 10.1016/j.ijmedinf.2016.07.004] [Medline: 27573318]
4. Chong N, Akobirshoev I, Caldwell J, Kaye HS, Mitra M. The relationship between unmet need for home and community-based services and health and community living outcomes. Disabil Health J 2022;15(2):101222. [doi: 10.1016/j.dhjo.2021.101222] [Medline: 34657829]
5. Wilson DJ, Mitchell JM, Kemp BJ, Adkins RH, Mann W. Effects of assistive technology on functional decline in people aging with a disability. Assist Technol 2009;21(4):208-217. [doi: 10.1080/10400430903246068] [Medline: 20066887]
6. Szanton SL, Leff B, Wolff JL, Roberts L, Gitlin LN. Home-based care program reduces disability and promotes aging in place. Health Aff (Millwood) 2016;35(9):1558-1563. [doi: 10.1377/hlthaff.2016.0140] [Medline: 27605633]
7. Orellano-Colón EM, Rivero-Méndez M, Ralat-Fonseca BN, Varas-Díaz N, Lizama-Troncoso M, Jiménez-Velázquez IZ, et al. Multilevel barriers to using assistive technology devices among older hispanics from poor and disadvantaged communities: the relevance of a gender analysis. Disabil Rehabil Assist Technol 2024;19(3):682-698. [doi: 10.1080/17483107.2022.2117427] [Medline: 36170426]
8. Prajapati G, Sharmila K. Difficulties experienced by older adults when not using assistive devices. Discov Soc Sci Health 2023;3(1) [FREE Full text] [doi: 10.1007/s44155-023-00037-1]

XSL•FO
RenderX

9. Al‐Oraibi S. Impact and economic assessment of assistive technology in care homes in Norfolk, UK. J Assist Technol 2012;6(3):192-201 [FREE Full text] [doi: 10.1108/17549451211261317]
10. Meiland F, Innes A, Mountain G, Robinson L, van der Roest H, García-Casal JA, et al. Technologies to support community-dwelling persons with dementia: A position paper on issues regarding development, usability, effectiveness and cost-effectiveness, deployment, and ethics. JMIR Rehabil Assist Technol 2017;4(1):e1 [FREE Full text] [doi: 10.2196/rehab.6376] [Medline: 28582262]
11. Waldron D, Layton N. Hard and soft assistive technologies: defining roles for clinicians. Aust Occup Ther J 2008;55(1):61-64. [doi: 10.1111/j.1440-1630.2007.00707.x] [Medline: 20887435]
12. Healthy ageing and functional ability internet. World Health Organization. Geneva; 2020. URL: https://www.who.int/news-room/questions-and-answers/item/healthy-ageing-and-functional-ability [accessed 2025-09-12]
13. Orellano-Colón EM, Ramos-Marichal AI, González-Crespo VR, Zeballos-Hernández BN, Ruiz-Márquez KN, Roche-Lima A, et al. Breaking barriers: the design and development of an assistive technology web app for older latinos with disabilities in daily activities. Technologies (Basel) 2024;12(11):232. [doi: 10.3390/technologies12110232] [Medline: 40400618]
14. Orellano-Colón EM, Rivero-Méndez M, Lizama M, Jutai JW. Assistive technology unmet needs of independent living older Hispanics with functional limitations. Disabil Rehabil Assist Technol 2018;13(2):194-200 [FREE Full text] [doi: 10.1080/17483107.2017.1300693] [Medline: 28326905]
15. Orellano-Colón EM, Suárez-Pérez EL, Rivero-Méndez M, Boneu-Meléndez CX, Varas-Díaz N, Lizama-Troncoso M, et al. Sex disparities in the prevalence of physical function disabilities: a population-based study in a low-income community. BMC Geriatr 2021;21(1):419 [FREE Full text] [doi: 10.1186/s12877-021-02362-z] [Medline: 34246238]
16. Orellano-Colón EM, Fernández-Torres A, Figueroa-Alvira N, Ortiz-Vélez B, Rivera-Rivera NL, Torres-Ferrer GA, et al. Empowering potential of the my assistive technology guide: exploring experiences and user perspectives. Disabilities (Basel) 2024;4(2):303-320. [doi: 10.3390/disabilities4020020] [Medline: 38962656]
17. Martin Payo R, Fernandez Álvarez MM, Blanco Díaz M, Cuesta Izquierdo M, Stoyanov S, Llaneza Suárez E. Spanish adaptation and validation of the mobile application rating scale questionnaire. Int J Med Inform 2019;129:95-99. [doi: 10.1016/j.ijmedinf.2019.06.005] [Medline: 31445295]
18. Johnson SG, Potrebny T, Larun L, Ciliska D, Olsen NR. Usability methods and attributes reported in usability studies of mobile apps for health care education: scoping review. JMIR Med Educ 2022;8(2):e38259 [FREE Full text] [doi: 10.2196/38259] [Medline: 35767323]
19. Deniz-Garcia A, Fabelo H, Rodriguez-Almeida AJ, Zamora-Zamorano G, Castro-Fernandez M, Alberiche Ruano MDP, WARIFA Consortium. Quality, usability, and effectiveness of mHealth apps and the role of artificial intelligence: current scenario and challenges. J Med Internet Res 2023;25:e44030 [FREE Full text] [doi: 10.2196/44030] [Medline: 37140973]
20. Manzano-Monfort G, Paluzie G, Díaz-Gegúndez M, Chabrera C. Usability of a mobile application for health professionals in home care services: a user-centered approach. Sci Rep 2023;13(1):2607 [FREE Full text] [doi: 10.1038/s41598-023-29640-7] [Medline: 36788261]
21. Fetters MD, Curry LA, Creswell JW. Achieving integration in mixed methods designs-principles and practices. Health Serv Res 2013;48(6 Pt 2):2134-2156 [FREE Full text] [doi: 10.1111/1475-6773.12117] [Medline: 24279835]
22. Karagianni K. Optimizing the UX honeycomb—A small amendment to the classic diagram hopefully improves its UX. URL: https://uxdesign.cc/optimizing-the-ux-honeycomb-1d10cfb38097 [accessed 2028-12-06]
23. Morville P. User experience design. Semantic Studios. URL: https://semanticstudios.com/user_experience_design/ [accessed 2004-06-21]
24. Scherer MJ. Living in the State of Stuck: How Technology Impacts the Lives of People with Disabilities. Cambridge, MA: Brookline Books; 2005.
25. Kasym M. How to make the older users love your product examples of UX design for seniors. Eleken. 2022. URL: https://www.eleken.co/blog-posts/examples-of-ux-design-for-seniors [accessed 2025-09-12]
26. Nielsen J. Usability for senior citizens: improved but still lacking. NielsenNormanGroup. 2013. URL: https://www.nngroup.com/articles/usability-seniors-improvements/ [accessed 2025-09-12]
27. Kane L. Usability for senior citizens: challengers and changes. NielsenNormanGroup. 2019. URL: https://www.nngroup.com/articles/usability-for-senior-citizens/(accessed [accessed 2024-07-08]
28. Bryant JA, David P. Solving for inclusive technology for older adults. AARP International. 2020. URL: https://www.aarpinternational.org/file%20library/aarpthejournal/bryant-david-2020aarpjournal.doi.10.26419-2fint.00045.025.pdf [accessed 2023-03-21]
29. Trzepla S. UX accessibility for elderly: 12 principles. UX Planet. 2019. URL: https://uxplanet.org/ux-accessibility-for-elderly-12-principles-9708289b6f78 [accessed 2025-09-12]
30. Hwang W, Salvendy G. Number of people required for usability evaluation. Communications of the ACM 2010;53(5):130-133 [FREE Full text] [doi: 10.1145/1735223.1735255]
31. Nielsen J, Landauer T. A mathematical model of the finding of usability problems. 1993 Presented at: CHI '93: Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems; April 24-29, 1993; Amsterdam, The Netherlands p. 206-213. [doi: 10.1145/169059.169166]

XSL•FO
RenderX

32. Kushniruk AW, Patel VL. Cognitive and usability engineering methods for the evaluation of clinical information systems. J Biomed Inform 2004;37(1):56-76 [FREE Full text] [doi: 10.1016/j.jbi.2004.01.003] [Medline: 15016386]
33. Kadam PD, Chuan HH. Erratum to: rectocutaneous fistula with transmigration of the suture: a rare delayed complication of vault fixation with the sacrospinous ligament. Int Urogynecol J 2016;27(3):505 [FREE Full text] [doi: 10.1007/s00192-016-2952-5] [Medline: 26811110]
34. Stoyanov SR, Hides L, Kavanagh DJ, Wilson H. Development and validation of the user version of the mobile application rating scale (uMARS). JMIR Mhealth Uhealth 2016;4(2):e72 [FREE Full text] [doi: 10.2196/mhealth.5849] [Medline: 27287964]
35. Kibiswa N. Directed qualitative content analysis (DQlCA): A tool for conflict analysis. TQR 2019;24(8):2059-2079 [FREE Full text] [doi: 10.46743/2160-3715/2019.3778]
36. Borges do Nascimento IJ, Abdulazeem H, Vasanthan LT, Martinez EZ, Zucoloto ML, Østengaard L, et al. Barriers and facilitators to utilizing digital health technologies by healthcare professionals. NPJ Digit Med 2023;6(1):161 [FREE Full text] [doi: 10.1038/s41746-023-00899-4] [Medline: 37723240]
37. Kumar S, Nilsen WJ, Abernethy A, Atienza A, Patrick K, Pavel M, et al. Mobile health technology evaluation: the mHealth evidence workshop. Am J Prev Med 2013;45(2):228-236 [FREE Full text] [doi: 10.1016/j.amepre.2013.03.017] [Medline: 23867031]
38. Ran M, Banes D, Scherer MJ. Basic principles for the development of an AI-based tool for assistive technology decision making. Disabil Rehabil Assist Technol 2022;17(7):778-781. [doi: 10.1080/17483107.2020.1817163] [Medline: 33275457]

Abbreviations

ADL: activities of daily living
AT: assistive technology
CMS: Centers for Medicare & Medicaid Services
FD: functional disability
ICD-10: International Statistical Classification of Diseases, Tenth Revision
MGAT: Mi Guía de Asistencia Tecnológica (My Assistive Technology Guide)
PCP: primary care physician
PI: principal investigator
uMARS: User Version of the Mobile Application Rating Scale

Edited by A Mavragani; submitted 04.12.24; peer-reviewed by S Ashraf, L Weinert, F Ricci, S Thongchoomsin; comments to author 17.02.25; revised version received 28.02.25; accepted 07.04.25; published 22.10.25.

Please cite as:
Orellano-Colón EM, Bonilla-Díaz W, Revilla-Orellano R, Mejías-Castro J, Roche-Lima A
Quality of an Assistive Technology Web Application for Primary Care Physicians Serving Older Adults: Concurrent Mixed Methods Study
JMIR Form Res 2025;9:e69645
URL: https://formative.jmir.org/2025/1/e69645
doi:10.2196/69645
PMID:

©Elsa M Orellano-Colón, Wency Bonilla-Díaz, Radamés Revilla-Orellano, Jesús Mejías-Castro, Abiel Roche-Lima. Originally published in JMIR Formative Research (https://formative.jmir.org), 22.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A “Papageno” Story Interview Suicide Prevention Intervention for Young Adults With Past-Month Suicidal Ideation: Uncontrolled Single-Group Pilot Study of Feedback and Acceptability

Jane Harness[1], DO; Kamara Gardner[2], MPH; Jessica L Schleider[3], PhD; Cheryl A King[1], PhD

[1]Department of Psychiatry, University of Michigan—Ann Arbor, Ann Arbor, MI, United States

[2]Department of University Health and Counseling, University of Michigan, Ann Arbor, MI, United States

[3]Department of Medical Social Sciences, Northwestern University Feinberg School of Medicine, Chicago, IL, United States

Corresponding Author:
Jane Harness, DO
Department of Psychiatry
University of Michigan—Ann Arbor
4250 Plymouth Rd.
Ann Arbor, MI, 48109
United States
Phone: 1 7347640231
Email: janeehar@umich.edu

Abstract

Background: Suicide continues to be a leading cause of death for young people, and the 2024 US National Strategy for Suicide Prevention has called for innovative approaches to suicide prevention. This strategy acknowledges the potential of the “Papageno” effect or the suicide preventive effect of stories of people who overcome suicide crises

Objective: This pilot study’s objectives were to develop filmed interviews of young adults sharing their personal story of overcoming suicide crises (“Papageno” stories) and to examine their acceptability and appropriateness in a sample of young adults with a past-month history of suicidal ideation.

Methods: In total, 6 filmed interviews (5 “Papageno” story interviews and 1 active control interview) were videotaped. Study participants were recruited from Instagram and community organizations. Interviewees and participants were between the ages of 18 and 24 years. To be eligible for the study, participants responded “yes” to the question, “In the past 1 month, have you wished you were dead or wished you could go to sleep and not wake up?” Participants provided demographic information and responded to a Brief Mood Introspection Scale. They were subsequently instructed to watch the video, which was followed by the Brief Mood Introspection Scale and questions about the acceptability and helpfulness of the video. They were also asked 5 free-response questions regarding their likes, dislikes, suggestions, and overall reactions to the filmed interviews and whether they would watch it again.

Results: We collected 3-4 participant responses per video for each of the 5 different interview videos (a total of 16 responses). The videos were found to be generally acceptable and appropriate, and no participants reported feeling less likely to reach out for help when needed. Qualitative feedback yielded recommendations to shorten the length of the videos as well as to include multiple interviewees in future videos so that the audience has a better chance to identify with the interviewee.

Conclusions: Filmed “Papageno” interviews were rated as generally acceptable and appropriate, reinforcing the need for more research examining their potential effects on proximal risk factors for suicide. Next steps include the incorporation of feedback to create a singular finalized video and a randomized controlled trial assessing the effects of the Papageno video.

(JMIR Form Res 2025;9:e71368) doi:10.2196/71368

KEYWORDS

youths; suicide prevention; Papageno interview; intervention; digital media; narration

XSL•FO
RenderX

Introduction

In 2022, suicide was the second leading cause of death for 20- to 24-year-old people and the third leading cause of death for 15- to 19-year-old people in the United States [1]. Public health suicide prevention approaches have involved the creation of guidelines for news and entertainment media due to known suicide contagion or social transmission effects. These guidelines [2-4] advise against showing suicide methods and recommend the encouragement of help-seeking and promotion of crisis resources. Suicide protective effects can occur when alternatives to suicide are modeled and shared in news media [5,6]. This protective effect is often referred to as the "Papageno effect" [7]. It is well known that protective factors are an important part of a suicide risk assessment and that an individual's reasons for living may modulate the threshold for acting on suicidal thoughts [8,9].

The White House introduced a National Strategy for Suicide Prevention in 2024 [10] due to the ongoing public health concern of suicide. Goal 1 of this plan (Establish effective, broad-based, collaborative, and sustainable suicide prevention partnerships) describes success in this domain as "foster[ing] creativity, new ideas and perspectives." Furthermore, they suggested collaborating with new industries including entertainment media. Goal 7 (Implement research-informed suicide prevention communication activities in diverse populations using best practices from communication science) referenced the "Papageno" effect in highlighting the potential beneficial effect of safe messaging strategies. For example, they mentioned Logic's song "1-800-273-8255," which prompted an increase in calls to the National Suicide Prevention Lifeline and a decrease in suicides around that time [11]. Objective 7.4 is "communicate stories of help, hope and healing using safe messaging strategies," and objective 7.6 is "engage news media, the entertainment industry, and schools of journalism and mass communication to encourage safe, accurate and responsible ... depictions of suicide and positive mental health coping skills." In summary, national leadership has acknowledged the urgent need for creative collaborations including leveraging the Papageno effect in suicide prevention strategies.

The literature about the effect of Papageno stories is mostly limited to news reporting [12] but suggests that sharing stories about crisis intervention may promote help-seeking behaviors in viewers [6]. However, one study of fictional "Papageno" stories in entertainment media included an experimental study of viewer response to *Elizabethtown* (2005), which includes a Papageno story. Compared with the 2 other movies that viewers were randomized to in this study (*Night Mother*—protagonist dies by suicide and *A Single Man*—protagonist dies from a heart attack), *Elizabethtown* was found to have a positive impact on viewer's life satisfaction in the group with lower baseline reasons for living [13]. However, this movie does not have a strong, intentional message of suicide prevention; thus, it remains possible that preventive effects could be amplified with a more targeted message. In keeping with social learning theory [14], which asserts that behaviors are emulated by observers, observing people overcome suicide crises may bring about those same coping, help-seeking, and problem-solving behaviors in viewers.

Furthermore, peer testimonial programs show promise as a suicide prevention strategy. For example, the "Sources of Strength" taped testimonials and accompanying curriculum improved youth knowledge and attitudes about suicide [15]. Participation in this multicomponent program was also associated with lower rates of suicide attempts at 3 months after the program, and youth participants adopted more favorable attitudes toward getting help for themselves or their friends [16]. Seeking help from adults at school was viewed as more acceptable and potentially helpful after participation in the "Sources of Strength" peer-led suicide prevention program compared to a waitlist control [17]. When peer-written testimonials of "hope, help, and strength" were delivered to a target population via text through the Stories of Personal Resilience in Managing Emotions program, they were found to be relatable, useful, intriguing, and likeable [18].

Scalable and accessible online programs, such as single session interventions (SSIs), have shown promising effects for youth mental health in spaces where youths may otherwise have not received any care at all [19]. An SSI including testimonials of older youths who shared their experience of using a growth mindset to cope during adversity found significant improvements in depression and anxiety symptoms compared to the control program as far out as 9 months after the session [19]. During the COVID-19 pandemic, when adolescents were randomized to a behavioral activation SSI, a growth mindset SSI, and supportive control, those randomized to the SSI conditions had decreased hopelessness after the intervention and 3 months after the intervention compared to those randomized to the supportive control [20]. Furthermore, social media users aged 13-44 (average age 17.93, SD 4.92) years flagged as being potentially in crisis who engaged in a 1-minute crisis response SSI had a greater decrease in hopelessness and increased likelihood of using the crisis resources provided compared to users who received the typical crisis response [21]. Online SSIs like this may fill an important gap for people who are on waiting lists for professional mental health care, and they have been shown to be helpful in decreasing hopelessness and increasing help-seeking.

Although suicide prevention organizations have produced hope-inspiring interviews with people who have come close to suicide [22,23], some of these publicly available videos have not been studied for their effects on viewers. For the public-facing videos that have been studied, the results are video-dependent. In a study comparing a music video called "Boys Don't Cry" to an excerpt from a documentary showing an individual being taught to increase his cognitive processing speed while learning table tennis, both videos had a positive effect on help-seeking intentions among Australian men (age ranging from 18 to 87 years). In a double-blind randomized controlled trial (RCT) in Austria of the "It Gets Better" videos of lesbian, gay, bisexual, transgender, and queer+ individuals overcoming challenges, researchers found small improvements in suicidal ideation for transgender and nonbinary individuals (age ranging from 14 to 22 years) [24]. They also found improvement in help-seeking intentions in the intervention

XSL•FO
RenderX

group (compared to the active control video about a healthy lifestyle). Another recent study demonstrated positive effects in Austria. When adolescents, ages 14-19 years, were randomized to watch a video of a 17-year-old telling his story of finding help versus an active control, those who watched the intervention video experienced a decrease in suicidal ideation and improvement in help-seeking intentions [25]. The study of similarly scalable and accessible story-based interventions in the United States for the young adult age group is needed. Additionally, further study of videos created specifically to highlight "Papageno" stories is needed. As such, we set out to film interviews with young adults with "Papageno" stories to determine their acceptability among a group of young adults with a past-month history of suicidal ideation. The social learning theory informs the intervention design because viewers may emulate the help-seeking behavior the interviewee used.

Methods

Ethical Considerations

An umbrella institutional review board (IRB) was approved for this project (HUM00208715). Creation of the filmed videos was exempted by the IRB (HUM00223570). The piloting of the filmed videos was approved by the IRB (HUM00230451). Informed consent was obtained for all participants as outlined below. Measures were taken to maintain privacy and confidentiality. Any identifying information needed for compensation purposes was collected in a survey separate from the consent and separate from the study questions. Participants were aware of the compensation, and this information was also included in the consent.

Intervention

We filmed 5 interviews of young adults, ages 18 to 24 years, with personal histories of "Papageno" stories. Interviewees were recruited from community organizations and met with the principal investigator (PI; JH) at least once to discuss the study and their stories before the filmed interview. They were given information related to how to safely share their story in accordance with the guidelines. They were created with the purpose of this research study and are not currently publicly available. Each interviewee was asked questions about what helped them during a time in their life when they came close to suicide. Additionally, the interviews focused on recovery, help-seeking, reasons for living, and hope for the future. Example follow-up questions include: "What was helpful during that time?" "What advice would you give to your younger self?" and "What advice would you give to people watching this video?" Interview questions were tailored in real time to guide the interviewee in telling their "Papageno" story and to highlight what was distinctly helpful to them during that time. Because each of their stories was unique, we did not use a standardized set of questions. A TIDieR (Template for Intervention Description and Replication) table outlining key components of the intervention can be found in Multimedia Appendix 1. An active control educational video about mental health resources was also created. All interview videos were between 20 and 35 minutes, and the role of the interviewer was to elicit the interviewee's Papageno story of hope and recovery. The visuals sometimes included shots of 2 people talking in a room, and at other times, the camera is focused solely on the interviewee. All of the videos followed the same structure of the individual first sharing their "Papageno" story and then discussion of what was helpful for them. Videos were filmed by a professional and edited by the PI. Video length was determined by the time needed for each interviewee to share their story and what was helpful during that time. Videos followed all guidelines for media with suicide content including not discussing suicide methods. Participants could have completed the survey at any point after watching the video, but they were encouraged to set aside time to do it all together in 1 sitting if possible.

Recruitment

Participants were recruited from social media (specifically Instagram) through targeted advertisements and mental health community groups via email. Inclusion criteria were ages 18-24 years and a past-month history of suicidal ideation. An exclusion criterion was the inability to participate in an online Qualtrics survey in English. For those recruited through social media, the advertisement was linked to a Qualtrics survey in which participants could read more about the study. If still interested, they could choose a time to review the consent over Zoom (Zoom Video Communications) with the PI. For those recruited from community groups, potential participants were given the same study information and directed to email the PI if they were interested in scheduling the consent process. During the Zoom meeting, the PI asked the participant two yes or no questions to ensure eligibility: (1) Are you between the age of 18 and 24 years? (2) In the past 1 month, have you wished you were dead or wished you could go to sleep and not wake up?

If the participant answered yes to both questions, the PI shared the link to the Qualtrics consent form via chat. The consent form was reviewed in real time together over Zoom, and the participant was given a chance to ask questions. After the participant verbally and electronically consented, they were sent a separate Qualtrics link to the study survey with the embedded video. Participants were compensated US $20 in the form of an e-gift card for each video that they reviewed for up to 5 videos. If participants wanted to respond to more than 1 video, they emailed the PI for the next video's Qualtrics link. Crisis numbers were provided during the recruitment process and in the study survey. Because participants needed to email the PI for the link to the next video survey, there was necessarily at least some time in between the viewings. Participants did not choose which video they received. There was no way for us to determine if the participants watched the full video with full attention, but they could not advance to the next survey page until at least half of the video length time had elapsed in case they decided to watch it at 2 times the original speed.

Measures

To test the viewer's mood before and after watching the filmed interview, a single-item question adapted from the Brief Mood Introspection Scale [26] was used. Participants were asked, "Overall my mood right now is –10=very unpleasant to 10=very pleasant." The rationale for including this scale was to capture any unintended induced dysphoria that could co-occur with

XSL•FO
RenderX

improvements in other measures such as hope for the future and help-seeking intentions.

To test filmed interview acceptability, the Program Feedback Scale (PFS) [27] and Acceptability of Intervention Measure (AIM) [28] were used. The PFS asks participants to rate agreement with 7 statements (eg, "I enjoyed the filmed interview") on a 5-point Likert scale (1=strongly disagree and 5=strongly agree). They were also asked the open-response questions: "What did you like about the filmed interview?" and "What would you change about the filmed interview?" The PFS was created to test SSIs and adapted from validated assessments intended for digital interventions [29]. The AIM asks participants to rate agreement with 4 statements (eg, "The filmed interview meets my approval") on a 5-point Likert scale (1=strongly disagree and 5=strongly agree).

To test filmed interview appropriateness, the Intervention Appropriateness Measure (IAM) [28] was used. The IAM asks participants to rate agreement with 4 statements (eg, "The filmed interview seems applicable") on a 5-point Likert scale (1=strongly disagree and 5=strongly agree). The AIM and the IAM have good structural validity and reliability, and each measure is sensitive to change in both directions.

Participants were asked the following 3 questions, with response options rated on a 5-point Likert scale (1=significantly less hopeful or understood or likely and 5=significantly more hopeful or understood or likely).

"Compared to before you watched this filmed interview, are you feeling" (1) more or less hopeful about the future? (2) more or less understood? or (3) more or less likely to reach out for help in the future?

Participants were prompted with the following to collect open-ended responses: (1) What did you like about the filmed interview? (from the PFS) (2) What would you change about the filmed interview? (from the PFS) (3) Please provide any additional comments or suggestions for us related to the filmed interview you watched. (4) What are your overall reactions to this filmed interview? (5) Would you watch this video again on your own time if you were going through a hard time? Why or why not?

Finally, to test participant identification with the interviewee, 8 items from Cohen Identification Scale [30] were used. Examples of items on this scale include: "I think I have a good understanding of the interviewee" and "While viewing the interview, I could feel the emotions the interviewee portrayed." Answers are rated on a 5-point Likert scale (1=strongly disagree and 5=strongly agree).

Statistical Analysis

We collected both quantitative and qualitative data. Because our sample was small (N=16 for the interview videos), we limited our analyses to descriptive statistics. To analyze the open-ended responses, we used content analysis [31] to highlight frequencies of codes for each of the 5 free-response questions. Authors JH and KG used an inductive approach and memoed to create a list of themes that were then consolidated into the final codebook through real-time consensus. Using the codebook, authors JH and KG coded each participant's response to each question. Where there were disagreements in coding, authors JH and KG met to reconcile the differences in real time. The codes and their frequencies as well as a representative quote were compiled to summarize the main findings of the qualitative data.

Results

Sample demographic characteristics can be found in Table 1. Because participants were able to provide feedback for up to 5 videos, the demographic information is reported per response for the 16 total responses for the intervention (Papageno story interview) videos. The sample size, therefore, was smaller than the number of responses. One video had 4 responses, and the remaining 4 videos had 3 responses each. All of the results we describe here must be interpreted in the context of having a very small sample size, which prohibits statistical testing; therefore, no meaningful effects can be assumed. Participants reported an average pre- to postmood change of +2.3 (SD 5.9; scale ranging from –10=very unpleasant to 10=very pleasant). For all of the questions involving a 5-point Likert scale (1=strongly disagree and 5=strongly agree), the averages of the responses were calculated and can be found in Table 2. For the measures of acceptability (the PFS and the AIM), response averages ranged from 3.75 to 4.5 (SD ranging from 0.48 to 1.18). For the measure of appropriateness (the IAM), response averages ranged from 3.94 to 4.13 (SD ranging from 0.44 to 0.77). For the measures of identification with the interviewee (questions from Cohen's Identification Scale), response averages ranged from 3.06 to 4.19 (SD ranging from 0.54 to 1.18). There were 3 instances of missing responses for the quantitative questions: 1 postintervention mood scale, 1 response to the question about whether the filmed interview would help other people their age, and 1 response to the question about whether they felt as if they were there while watching the filmed interview. Because missing data were rare, and this is an uncontrolled pilot study, we performed the statistical analyses without adjusting for the missing data.

Results regarding changes in future hopefulness, feeling understood, and the likelihood to reach out for help in the future are reflected in Table 3. No participants reporting feeling less likely to reach out for help in the future. In total, 4 responses indicated feeling somewhat less understood, but there were 0 responses for feeling significantly less understood. There was 1 response for feeling significantly less hopeful about the future, but this particular respondent also reported feeling significantly more likely to reach out for help in the future and said "I will love to watch more videos" when asked to provide additional comments or suggestions, showing that this response may have been an error. There were 0 responses for feeling somewhat less hopeful about the future.

Qualitative results for the 5 open-ended questions are recorded in Tables 4-8. When asked what they liked about the filmed interview, participants' responses were coded with the following codes (from most frequent to least): hopeful or helpful, knowledge, genuine, relatable or understandable, comprehensive, and simple. When asked what they would like

XSL•FO
RenderX

to change about the filmed interview, 7 of the 16 responses were coded as “nothing.” The most frequent recommendation was to shorten the length of the filmed interviews, with one respondent noting: “... Especially when I am at my lowest emotional points I do not think I would watch a 33 minute interview” and to include more stories or perspectives (n=4, 21.1% for each). In total, 3 responses had specific suggestions for the interviewer. One person asked for captions (which were available to be turned on, but they may have missed the button). When asked for additional comments or suggestions, 6 (46.2%) responses were coded as asking for more stories and perspectives. Additionally, the length of the videos came up again (n=2, 15.4%). A total of 2 responses were coded as “hopeful or helpful,” and 2 responses had specific advice about the visual elements of the filmed interview. Finally, 1 person mentioned again the desire for subtitles.

Table 1. Respondent characteristics (N=16).

Characteristics	Values
Age (years)	
Mean (SD)	22.3 (1.7)
Range	18-24
Sex, n (%)	
Male	3 (18.8)
Female	10 (62.5)
Nonbinary	2 (12.5)
Transgender male	1 (6.3)
Ethnicity, n (%)	
Hispanic	0 (0)
Non-Hispanic	16 (100)
Race, n (%)	
Asian	5 (31.3)
Black or African American	6 (37.5)
White	5 (31.3)

XSL•FO
RenderX

Table 2. Average scores for participant identification with the interviewee, filmed interview appropriateness, and filmed interview acceptability.

	Average (SD)[a]
Program Feedback Scale	
I enjoyed the filmed interview	3.94 (0.57)
I understood the filmed interview	4.50 (0.52)
Watching the filmed interview was easy to do	4.13 (0.72)
I was well-focused when I watched the filmed interview	4.06 (0.57)
I think the filmed interview would be helpful to other people my age	3.93 (0.88)
I would recommend this filmed interview to a friend going through a hard time	3.75 (1.18)
I agree with the filmed interview's message	4.31 (0.48)
Acceptability of Intervention Measure	
The filmed interview meets my approval	3.93 (0.59)
The filmed interview is appealing to me	3.81 (0.83)
I liked the filmed interview	4.06 (0.57)
I welcome the filmed interview	4.31 (0.60)
Intervention Appropriateness Measure	
The filmed interview seems fitting	4.06 (0.68)
The filmed interview seems suitable	4.06 (0.44)
The filmed interview seems applicable	4.13 (0.62)
The filmed interview seems like a good match	3.94 (0.77)
Cohen Identification Scale	
While viewing the filmed interview, I felt as if I was there	3.13 (1.06)
While viewing the filmed interview, I forgot myself and was fully absorbed	3.06 (1.18)
I was able to understand the events described by the interviewee in a manner similar to that in which the interviewee understood them	4.00 (0.82)
I think I have a good understanding of the interviewee	4.13 (0.62)
While viewing the interview, I could feel the emotions the interviewee portrayed	4.19 (0.75)
During viewing, I felt I could really get inside the filmed interviewee's head	4.00 (0.73)
At key moments in the interview, I felt I knew what the interviewee was going through	4.19 (0.54)
While viewing the interview, I wanted the interviewee to pursue their dreams	4.31 (0.60)

[a]1=strongly disagree to 5=strongly agree.

XSL•FO
RenderX

Table 3. Number of responses for "Compared to before you watched this filmed interview, are you feeling ..."

	Respondents, n (%)
Level of change in hope about the future	
Significantly less hopeful about the future	1 (6.3)
Somewhat less hopeful about the future	0 (0)
The same amount of hopefulness about the future	7 (43.8)
Somewhat more hopeful about the future	7 (43.8)
Significantly more hopeful about the future	1 (6.3)
Level of change in feeling understood	
Significantly less understood	0 (0)
Somewhat less understood	4 (25)
No change in whether you feel more or less understood	2 (12.5)
Somewhat more understood	7 (43.8)
Significantly more understood	3 (18.8)
Level of change in likelihood to reach out for help when needed	
Significantly less likely to reach out for help when needed	0 (0)
Somewhat less likely to reach out for help when needed	0 (0)
No change in the likelihood of reaching out for help when needed	4 (25)
Somewhat more likely to reach out for help when needed	9 (56.3)
Significantly more likely to reach out for help when needed	3 (18.8)

Table 4. Content analysis of qualitative data: what respondents liked about the filmed interview.

What did you like about the filmed interview?	Values, n (%)	Representative quote
Hopeful or helpful (effective or encouraging or engaging)	6 (24)	"I liked the fact that it the survivor does not just lament and unleash all her difficult moments as that would have affected my mood, but she went ahead to say how she battled [with it] to cope and thereby dropping some strong words of encouragement signifying [that I] shouldn't give up or let it way me down."
Knowledge (crisis resources or examples of ways people have found help)	5 (20)	"This film would not only give a person and advice, but hope as the message itself is coming from a Victim/survivor which makes it more realistic and the messages itself hopeful."
Genuine (candid or open or emotional or sincere)	5 (20)	"I appreciated how candid the interviewee was. His openness was palpable and it was nice to see how much better he was doing since his attempt. His messages at the end particularly stood out as being some of the interview's most emotional moments."
Relatable or understandable	5 (20)	"I found the interviewee to be relatable in her feelings and the way she described them!"
Comprehensive (detailed or holistic)	3 (12)	"I like that both systemic and personal barriers were addressed and the message that things are in flux and the environment influences a lot of how we feel."
Simple	1 (4)	"It was simple, understandable and effective."

XSL•FO
RenderX

Table 5. Content analysis of qualitative data: what respondents would change about the filmed interview.

What would you change about the filmed interview?	Values, n (%)	Representative quote
Nothing	7 (36.8)	"I don't think there is something to change."
Too long	4 (21.1)	"I did think the filmed interview was a bit long ..."
More stories or perspectives (not relatable or generalizable, include multiple people, include multiple reasons for living)	4 (21.1)	"Maybe more variety rather than just emphasis on personal relationships. When I am feeling very low, I'm convinced that nobody cares about me, so I don't think that alone could bring me back."
Interviewer (less and better questions or prompts)	3 (15.8)	"I feel like the parts of the interview that talked about suicide prevention advocacy took away from the individual story in a way, like it was performative, even though I know it's not. When I'm feeling bad, I think the personal stories are so much more impactful than hearing about efforts to keep lots of people alive."
Accessibility needs (captions)	1 (5.3)	"Captions would help because the sound was a little low."

Table 6. Content analysis of qualitative data: additional comments or suggestions from respondents.

Please provide any additional comments or suggestions for us related to the filmed interview you watched.	Values, n (%)	Representative quote
More stories or perspectives (not relatable or generalizable, include multiple people, include multiple reasons for living)	6 (46.2)	"Maybe the speaker could talk about how they brought up the topic of suicide/mental health with their social circles."
Too long	2 (15.4)	"To condense my previous suggestions, I think a shorter interview with a wider or more relatable breadth of experiences would be a lot more moving and beneficial."
Hopeful or helpful (effective or encouraging or engaging)	2 (15.4)	"I liked the whole journey story. Hearing about her struggles helped rationalize her solutions."
Visual (more visuals or videos add graphics or items of visual interest, camera focus)	2 (15.4)	"I would recommend a stronger camera focus on the individual, rather than two people in a room talking; individualizing the interviewee and emphasizing her common ground and character in relation to the viewer is important. I did not get much of a glimpse of her personality, which made the video seem impersonal."
Accessibility needs (captions)	1 (7.7)	"Providing subtitles would make it more accessible."

Table 7. Content analysis of qualitative data: respondents' overall reactions.

What are your overall reactions to this filmed interview?	Values, n (%)	Representative quote
Positive outcome or emotion (liked it or good or grateful or less impulsive or more reflective)	8 (34.8)	"My reactions were, for the most part, appreciative of the central message! His advice at the end stuck out to and touched me the most."
Wish for specific things to be different (content, ambiguity of religious lens, longing for more depth)	5 (21.7)	"I don't feel happier exactly, but I do feel more understood. I feel more just sad than angry or upset, which I think is good because it means I feel less impulsive and more reflective. I also thought the discussion of God and religion was interesting—it was neither religious nor non-religious, but I almost wish they had picked whether to base it in spirituality or not rather than trying to relate to everyone."
Understood (not alone, specific personal experiences, real)	4 (17.4)	"It was comforting to understand that other people could also feel what you are going through."
Hopeful or helpful (effective or encouraging or engaging)	3 (13)	"I really liked it and thought it was good to see that someone felt hope on the other side."
Knowledge (crisis resources or examples ways people have found help)	2 (8.7)	"I find it very knowledgeable but i would have love to here the deepest side of their stories."
Indifferent	1 (4.3)	"The film interview is reserved and more elaborate and its okay."

XSL•FO
RenderX

Table 8. Content analysis of qualitative data: respondents' thoughts about whether they would watch the video again.

Would you watch this video again on your own time if you were going through a hard time? Why or why not?	Values, n (%)	Representative quote
Yes	8 (53.3)	"Yes as it will improve my situation."
No	4 (26.6)	"Probably not, due to the length."
Maybe	3 (20)	"Not sure because its long."
Too long	3 (33.3)	"Given the sheer length of the interview, I don't think I'd watch it if I were going through a very hard time."
Relatable	2 (22.2)	"I would watch it again because it is like seeing your own struggles and recover through other people's lens."
I do not rewatch things	1 (11.1)	"Probably not. I think it's the sort of thing that would be helpful to watch once, but I very rarely re-watch anything. I'd consider re-watching it if I needed a nice little cry or moment of understanding."
Too close to own experience	1 (11.1)	"Maybe only some parts of it since other parts felt very close to my own experience, so the emotions from that time came up."
Not relatable	1 (11.1)	"No, because I don't resonate with her struggles. Her experience of recognition for academic giftedness actually made me a little embarrassed for my own attainments, which were limited in part due to my mental health spiraling pre- and post-COVID."
Helpful or reminder of tools	1 (11.1)	"Yes. I think it would remind me of some tools available."

When asked about overall reactions to the filmed interview, half (8/16) of the responses were coded as "positive outcome or emotion." In total, 5 responses had specific suggestions, 4 responses were coded indicating that the filmed interviews led to them feeling more understood, 3 responses were coded as "hopeful or helpful," 2 were coded as "knowledge," and 1 was coded as "indifferent." When asked if they would watch the video again if going through a hard time, over half (8/15) said yes, 4 said no, and 3 said maybe (1 respondent did not answer this question). Additional information provided was also coded with 3 responses indicating that the video was too long, 2 saying that they liked that it was relatable, 1 indicating that they do not rewatch things, 1 indicating that the video actually felt too close to their own experience, 1 indicating that it was not relatable, and 1 mentioning that it was a helpful reminder of tools.

In total, 3 responses were collected about the active control video to assess safety and acceptability; however, because of the small sample size, we cannot compare it to the "Papageno" videos. The average mood change from before to after the active control video was –0.67 (SD 0.58). All of the answers for the PFS were either agree or neutral. All of the participant answers from the AIM and IAM measuring acceptability and appropriateness were agree or neutral, except for 1 participant who responded "disagree" for "The filmed interview meets my approval." However, this respondent also stated "nothing" when asked, "What would you change about the filmed interview?" indicating some level of acceptability and appropriateness. Participants reported feeling either somewhat more understood or no change in feeling more or less understood. Participants reported feeling somewhat or significantly more likely to reach out for help when needed. In total, 2 participants reported feeling somewhat more hopeful about the future, and 1 participant reported feeling significantly less hopeful about the future. This participant responded, "I like that fact that an emergency number is displayed to reach out anytime someone is having a suicidal thoughts" to the question "What did you like about the filmed interview?" and "good" to "What are your overall reactions to this filmed interview?" When prompted, "Please provide any additional comments or suggestions for us related to the filmed interview you watched," one participant responded, "It is really helpful and I'm glad i watched it." When asked, "Would you watch this video again on your own time if you were going through a hard time? Why or why not?" participants responded, "Somehow", "Yes," and "Yes ... I will watch." The Cohen Identification Scale responses ranged from neutral to strongly agree, except for 1 participant who responded "disagree" to "While viewing the filmed interview, I felt as if I was there" and "While viewing the interview, I could feel the emotions the interviewee portrayed."

Discussion

Principal Findings

The filmed interviews we created of young adults sharing their "Papageno" stories were generally reported to be acceptable and appropriate by our sample of 18- to 24-year-old people with a past-month history of suicidal ideation. In their qualitative feedback, respondents recommended shortening the videos and including multiple perspectives combined in 1 video. These results will guide further revisions, leading to the creation of a singular shortened video including clips from all 5 interviewees. Additionally, as a result of the feedback received, we plan to emphasize focus on the interviewee (and minimize focus on the interviewer) and preserve elements that were reported to be helpful or inspire hope.

Although we did run descriptive statistics on the data collected, the sample size is too small to capture statistical significance, and inferential conclusions cannot be drawn. The average scores

XSL•FO
RenderX

on the acceptability and appropriateness scales were all higher than neutral or 3 (range 1-5), with the highest average from the PFS being "I understood the filmed interview" (mean 4.5, SD 0.52). The highest average from the AIM was "I welcome the filmed interview" (mean 4.31, SD 0.60). The highest average from the IAM was "The filmed interview seems applicable" (mean 4.13, SD 0.62). No respondents reported feeling less likely to reach out for help in the future after watching the filmed interview. Moreover, other than the 1 potentially erroneous response, no respondent reported feeling less hopeful about the future after watching the filmed interviews. These preliminary findings suggest that the Papageno story interview may show promise as an intervention for young adults with recent histories of suicidal ideation, but must be tempered by the limitations of the small sample size and uncontrolled pilot design.

The lowest average scores were for "I forgot myself and was fully absorbed" (mean 3.06, SD 1.18) and "I felt as if I was there" (mean 3.13, SD 1.06), with a score of 3 meaning "neutral." However, other measures of identification with the interviewee were 4 or above. Helping the viewer to feel more absorbed and as if they were there might help them to connect further with the story and imagine what will happen next, which is a key component of narrative transportation that can alter attitudes [32]. Our iterative video that will be put forward for an RCT is shortened and includes clips from all 5 interviewees in order to increase the likelihood that a viewer might identify with a story and thereby increase the likelihood of narrative transportation.

It is promising that over half (n=8) of the responses indicated that participants would watch the video again on their own time if they were going through a hard time (with an additional 3 responses indicating that they might). Of the 4 responses indicating that they would not watch it again, 2 indicated that this was due to its length, 1 simply said "No," and 1 cited that they were unable to relate to it. By combining multiple stories in 1 video, we are hopeful that more viewers would find something relatable about it. By shortening the length of the video, we are hoping that this will minimize any time commitment barrier and increase engagement.

Comparison With Prior Work

This uncontrolled single-group pilot study supported filmed "Papageno" story interviews' acceptability and appropriateness. The feedback received guided next steps in our pursuit toward a finalized singular intervention video combining clips from the original interviews, which could be tested for effectiveness in an RCT. A recent RCT of an "adolescent with past suicidal ideation, describing his story of getting help and recovering" showed that adolescents reported significantly lower suicidal ideation after watching the video [25]. Those who watched the video in that study also reported increased help-seeking intentions, which were maintained at 4-week follow-up. They also found a statistically significant decrease in favorable attitudes toward suicide (which is associated with increased suicide risk). After incorporating the feedback gathered from this study, the finalized intervention video we will test in an RCT (using similar methods) will differ because it will include all 5 interviewee perspectives in order to maximize the opportunity for viewer identification.

Compared with the Sources of Strength study [16], this study was a very small uncontrolled pilot from which conclusions about effectiveness or suicide outcomes cannot be made. Our study did not look at outcomes such as suicidal ideation or suicide attempts. Our study also was not an effectiveness study.

An Australian pilot study of a peer-to-peer video campaign called "Better Off With You," which involved videos challenging the notion of perceived burdensomeness, found no changes in participants' perceived burdensomeness, psychological distress, or help-seeking [33]. The campaign was tested in a community sample (n=157, age 20-65 years) compared to our study, which involved an eligibility criterion of past-month history of suicidal ideation. The qualitative interviews of participants indicated that people felt the campaign was safe and not distressing to people with lived experience. They concluded, "This result supports a meta-analysis that concluded that asking participants about suicide ideation or exposure to suicide-related content does not increase suicide risk" [34]. Future research is needed to understand which types of video content work best for which populations. A future direction might involve a study design where participants are allowed to choose the video they would like to watch by the preview thumbnail and short description or a study design that evaluates the real-world effectiveness of videos when presented via social media advertisements.

While each of the 5 videos we piloted featured only 1 interviewee, due to the feedback we received, the video we intend to put forward for an RCT will include all 5 perspectives, which will distinguish it from the aforementioned study. If the finalized video shows positive effects in the RCT, similar videos could be disseminated through social media or as a part of targeted interventions or therapies such as SSIs. A future direction in this field involves continued development and testing of various types of videos to further refine what elements make these videos more or less effective in their goal of suicide prevention.

Limitations

Although consistent with a small pilot study of intervention acceptability and feasibility, a limitation of this study is that we are unable to draw any conclusions regarding the effect of the Papageno video intervention, as we did not conduct an RCT. In addition, we are unable to generalize findings to the broader population of young adults because of the small sample size and sample homogeneity. Because this was a small pilot study, we were not able to obtain a representative sample. Recruitment via Instagram and community organizations may also invite bias. Although all participants had a past-month history of suicidal ideation, we were unable to further characterize their suicide risk. Recruitment of participants with higher suicide risk might necessitate alternative recruitment methods (as investing in clicking an Instagram advertisement or engaging with a message from a community organization demonstrates some amount of baseline help-seeking). Using only Instagram and not other social media platforms may also introduce a bias related to people who use or do not use Instagram. Future studies

XSL•FO
RenderX

are needed to determine the effects for individuals with both lower or higher suicide risk.

Another limitation is that participants were not queried at a follow-up time point, so any data reported are associated with the immediate postvideo time point. Because participants could watch multiple videos, it is possible that the sustained influence of the videos watched first could impact the feedback gathered from subsequent videos. At the same time, they may have found something uniquely helpful about subsequent videos.

Although missing data were rare, bias could have been introduced if people's responses would have diverged greatly from the average response, which is especially possible if they intentionally skipped the question.

Conclusions

The feedback received about these filmed interviews indicates that they are generally acceptable and safe. The qualitative feedback received, which pointed toward shortening the videos and including multiple perspectives, will be used to edit the 5 filmed interviews into 1 singular shortened intervention video to be put forward for an RCT. This pilot study adds momentum to the further study of "Papageno" stories shared in video format.

Acknowledgments

Funding for this work was provided by the American Foundation for Suicide Prevention Young Investigator Grant (YIG-0-062-21) and the Ouida Scholar Award at Michigan Medicine. The authors would also like to acknowledge Nathan Turner as the videographer for this project.

Conflicts of Interest

JH receives research funding through the American Foundation for Suicide Prevention, the Ouida Scholar Award (University of Michigan), and the Research Scouts (University of Michigan). CAK is a scientific advisor and stockholder with Oui Therapeutics and Vytyl Health Management. CAK is also on the Scientific Advisory Board of the American Foundation for Suicide Prevention. JLS serves on the Scientific Advisory Board for Walden Wise and the Clinical Advisory Board for Koko, has received consulting fees from United Health and Woebot, and receives book royalties from New Harbinger; Oxford University Press; and Little, Brown Book Group. JLS is co-founder and chief scientific advisor for Mindly. No Mindly products were used or are referenced in the present manuscript. JLS has received funding from the National Institutes of Health Office of the Director (DP5OD028123), the National Institute of Mental Health (R43MH128075), the Upswing Fund for Adolescent Mental Health, the National Science Foundation (2141710), the Health Research and Services Administration (U3NHP45406-01-00), the Society of Clinical Child and Adolescent Psychology, Hopelab, the Child Mind Institute, Alongside, Kooth, and the Klingenstein Third Generation Foundation.

Multimedia Appendix 1
TIDieR checklist for the video interventions.
[PDF File (Adobe PDF File), 81 KB - formative_v9i1e71368_app1.pdf]

References

1. Leading causes of death. WISQARS Leading Causes of Death Visualization Tool. 2024. URL: https://wisqars.cdc.gov/lcd/?o=LCD&y1=2022&y2=2022&ct=10&cc=ALL&g=00&s=0&r=0&ry=2&e=0&ar=lcd1age&at=groups&ag=lcd2age&a1=0&a2=199 [accessed 2025-08-24]
2. Reporting on suicide prevention. American Foundation for Suicide Prevention. 2021. URL: https://afsp.org/reporting-on-suicide-prevention/ [accessed 2021-03-08]
3. Preventing suicide: a resource for media professionals, update 2017. World Health Organization. 2017. URL: https://www.who.int/publications/i/item/WHO-MSD-MER-17.5 [accessed 2025-09-25]
4. Mishara B, Dargis L. Systematic comparison of recommendations for safe messaging about suicide in public communications. J Affect Disord 2019;244:124-154. [doi: 10.1016/j.jad.2018.09.031] [Medline: 30340101]
5. Niederkrotenthaler T, Voracek M, Herberth A, Till B, Strauss M, Etzersdorfer E, et al. Role of media reports in completed and prevented suicide: Werther v. Papageno effects. Br J Psychiatry 2010;197(3):234-243. [doi: 10.1192/bjp.bp.109.074633] [Medline: 20807970]
6. Jobes DA, Berman AL, O'Carroll PW, Eastgard S, Knickmeyer S. The Kurt Cobain suicide crisis: perspectives from research, public health, and the news media. Suicide Life Threat Behav 2010;26(3):260-271. [doi: 10.1111/j.1943-278x.1996.tb00611.x]
7. Domaradzki J. The werther effect, the papageno effect or no effect? A literature review. Int J Environ Res Public Health 2021;18(5):2396 [FREE Full text] [doi: 10.3390/ijerph18052396] [Medline: 33804527]
8. Malone KM, Oquendo MA, Haas GL, Ellis SP, Li S, Mann JJ. Protective factors against suicidal acts in major depression: reasons for living. Am J Psychiatry 2000;157(7):1084-1088. [doi: 10.1176/appi.ajp.157.7.1084] [Medline: 10873915]

XSL•FO
RenderX

9. Linehan MM, Goodstein JL, Nielsen SL, Chiles JA. Reasons for staying alive when you are thinking of killing yourself: the reasons for living inventory. J Consult Clin Psychol 1983;51(2):276-286. [doi: 10.1037//0022-006x.51.2.276] [Medline: 6841772]
10. National Strategy for Suicide Prevention. US Department of Health and Human Services (HHS). 2024 Apr. URL: https://www.hhs.gov/sites/default/files/national-strategy-suicide-prevention.pdf [accessed 2025-03-03]
11. Niederkrotenthaler T, Tran US, Gould M, Sinyor M, Sumner S, Strauss MJ, et al. Association of Logic's hip hop song "1-800-273-8255" with Lifeline calls and suicides in the United States: interrupted time series analysis. BMJ 2021;375:e067726 [FREE Full text] [doi: 10.1136/bmj-2021-067726] [Medline: 34903528]
12. Niederkrotenthaler T, Till B. Suicide and the media: from Werther to Papageno effects—a selective literature review. Suicid 2019;24(2). [doi: 10.5617/suicidologi.7398]
13. Till B, Strauss M, Sonneck G, Niederkrotenthaler T. Determining the effects of films with suicidal content: a laboratory experiment. Br J Psychiatry 2015;207(1):72-78. [doi: 10.1192/bjp.bp.114.152827] [Medline: 25906792]
14. Blood RW, Pirkis J. Suicide and the media. Part III: theoretical issues. Crisis 2001;22(4):163-169. [doi: 10.1027//0227-5910.22.4.163] [Medline: 11848660]
15. Brann KL, Baker D, Smith-Millman MK, Watt SJ, DiOrio C. A meta-analysis of suicide prevention programs for school-aged youth. Child Youth Serv Rev 2021;121:105826. [doi: 10.1016/j.childyouth.2020.105826]
16. Schilling EA, Aseltine RH, James A. The SOS suicide prevention program: further evidence of efficacy and effectiveness. Prev Sci 2016;17(2):157-166. [doi: 10.1007/s11121-015-0594-3] [Medline: 26314868]
17. Wyman PA, Brown CH, LoMurray M, Schmeelk-Cone K, Petrova M, Yu Q, et al. An outcome evaluation of the sources of strength suicide prevention program delivered by adolescent peer leaders in high schools. Am J Public Health 2010;100(9):1653-1661. [doi: 10.2105/AJPH.2009.190025] [Medline: 20634440]
18. Thiha P, Pisani AR, Gurditta K, Cherry E, Peterson DR, Kautz H, et al. Efficacy of web-based collection of strength-based testimonials for text message extension of youth suicide prevention program: randomized controlled experiment. JMIR Public Health Surveill 2016;2(2):e164 [FREE Full text] [doi: 10.2196/publichealth.6207] [Medline: 27829575]
19. Schleider J, Weisz J. A single-session growth mindset intervention for adolescent anxiety and depression: 9-month outcomes of a randomized trial. J Child Psychol Psychiatry 2018;59(2):160-170. [doi: 10.1111/jcpp.12811] [Medline: 28921523]
20. Schleider JL, Mullarkey MC, Fox KR, Dobias ML, Shroff A, Hart EA, et al. A randomized trial of online single-session interventions for adolescent depression during COVID-19. Nat Hum Behav 2022;6(2):258-268 [FREE Full text] [doi: 10.1038/s41562-021-01235-0] [Medline: 34887544]
21. Cohen K, Dobias M, Morris R, Schleider J. Improving uptake of mental health crisis resources: randomized test of a single-session intervention embedded in social media. J Behav Cogn Ther 2023;33(1):24-34. [doi: 10.1016/j.jbct.2022.12.001]
22. Moving America's Soul on Suicide (MASoS). 2021. URL: https://masosfilm.com/.? [accessed 2021-07-07]
23. Stories of hope and recovery. National Suicide Prevention Lifeline. 2025. URL: https://suicidepreventionlifeline.org/stories/.? [accessed 2025-09-25]
24. Kirchner S, Till B, Plöderl M, Niederkrotenthaler T. Effects of "It Gets Better" suicide prevention videos on youth identifying as lesbian, gay, bisexual, transgender, queer, or other sexual or gender minorities: a randomized controlled trial. LGBT Health 2022;9(6):436-446 [FREE Full text] [doi: 10.1089/lgbt.2021.0383] [Medline: 35575732]
25. Braun M, Till B, Pirkis J, Niederkrotenthaler T. Effects of suicide prevention videos developed by and targeting adolescents: a randomized controlled trial. Eur Child Adolesc Psychiatry 2023;32(5):847-857 [FREE Full text] [doi: 10.1007/s00787-021-01911-6] [Medline: 34817663]
26. Mayer JD, Gaschke YN. Brief Mood Introspection Scale. Biochem Pharmacol 1975;24(17):1639-1641. [doi: 10.1037/t06259-000]
27. Schleider JL, Mullarkey MC, Weisz JR. Virtual reality and web-based growth mindset interventions for adolescent depression: protocol for a three-arm randomized trial. JMIR Res Protoc 2019;8(7):e13368 [FREE Full text] [doi: 10.2196/13368] [Medline: 31290406]
28. Weiner BJ, Lewis CC, Stanick C, Powell BJ, Dorsey CN, Clary AS, et al. Psychometric assessment of three newly developed implementation outcome measures. Implement Sci 2017;12(1):108 [FREE Full text] [doi: 10.1186/s13012-017-0635-3] [Medline: 28851459]
29. Schleider JL, Dobias M, Sung J, Mumper E, Mullarkey MC. Acceptability and utility of an open-access, online single-session intervention platform for adolescent mental health. JMIR Ment Health 2020;7(6):e20513 [FREE Full text] [doi: 10.2196/20513] [Medline: 32602846]
30. Cohen J. Defining identification: a theoretical look at the identification of audiences with media characters. Mass Commun Soc 2001;4(3):245-264. [doi: 10.1207/S15327825MCS0403_01]
31. Hsieh H, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
32. Thomas VL, Grigsby JL. Narrative transportation: a systematic literature review and future research agenda. Psychol Market 2024;41(8):1805-1819. [doi: 10.1002/mar.22011]

XSL•FO
RenderX

33. Webb M, Carrotte ER, Flego A, Vincent B, Lee-Bates B, Heath J, et al. Safety, acceptability, and initial effectiveness of a novel digital suicide prevention campaign challenging perceived burdensomeness. Crisis 2023;44(3):189-197. [doi: 10.1027/0227-5910/a000840] [Medline: 35086355]
34. Blades CA, Stritzke WGK, Page AC, Brown JD. The benefits and risks of asking research participants about suicide: a meta-analysis of the impact of exposure to suicide-related content. Clin Psychol Rev 2018;64:1-12. [doi: 10.1016/j.cpr.2018.07.001] [Medline: 30014862]

Abbreviations

AIM: Acceptability of Intervention Measure
IAM: Intervention Appropriateness Measure
IRB: institutional review board
PFS: Program Feedback Scale
PI: principal investigator
RCT: randomized controlled trial
SSI: single session intervention
TIDieR: Template for Intervention Description and Replication

Edited by J Sarvestan; submitted 16.01.25; peer-reviewed by M Sinyor, V Carli, E Carrotte; comments to author 27.06.25; revised version received 09.07.25; accepted 22.07.25; published 23.10.25.

Please cite as:
Harness J, Gardner K, Schleider JL, King CA
A "Papageno" Story Interview Suicide Prevention Intervention for Young Adults With Past-Month Suicidal Ideation: Uncontrolled Single-Group Pilot Study of Feedback and Acceptability
JMIR Form Res 2025;9:e71368
URL: https://formative.jmir.org/2025/1/e71368
doi:10.2196/71368
PMID:

©Jane Harness, Kamara Gardner, Jessica L Schleider, Cheryl A King. Originally published in JMIR Formative Research (https://formative.jmir.org), 23.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Safe Patient Handling Musculoskeletal Injury-Prevention Smartphone App for Community Health Care Workers: Mixed Methods Feasibility Study

Alfred S Y Lee[1], PhD; Sarah Burrell[2], MBA; Sandy Courtnall[1], MSc; Shelly Wake[2]; Ryan E Rhodes[1], PhD

[1]Behavioural Medicine Laboratory, School of Exercise Science, Physical and Health Education, University of Victoria, Victoria, BC, Canada
[2]Island Health, Victoria, BC, Canada

Corresponding Author:
Alfred S Y Lee, PhD
Behavioural Medicine Laboratory
School of Exercise Science, Physical and Health Education
University of Victoria
PO Box 3010 STN CSC
University of Victoria
Victoria, BC, V8W 3N4
Canada
Phone: 1 (250) 472 5288
Email: alfredlee@uvic.ca

Abstract

Background: Safe patient handling is critical for reducing musculoskeletal injuries among health care workers; yet, community health care workers often face barriers such as limited access to training and real-time resources.

Objective: This study had three objectives: (1) provide detailed insights into the unmet needs of Island Health community health care workers with respect to safe patient handling resources and access to information, (2) translate those needs into a user-centered prototype of the Safe Patient Handling Musculoskeletal Injury-Prevention (SPH MSIP) smartphone app through an iterative co-design process, and (3) establish the acceptability and feasibility of SPH MSIP app to support community health care workers' safe patient handling practices using a mixed methods design.

Methods: A 3-phase participatory study was conducted. Phase 1 identified unmet safe patient handling needs through participatory meetings with 6 community health care workers, aligning with objective 1. Phase 2 involved developing the SPH MSIP app using co-design methods, integrating user feedback to address challenges such as guidance for high-risk tasks and intuitive design, addressing objective 2. Phase 3 evaluated the app's feasibility and acceptability, aligning with objective 3. The study recruited 28 participants who used the app for one month. A single-group mixed methods design was used, incorporating quantitative metrics such as recruitment (≥50%), retention (≥75%), and satisfaction (mean score ≥4). Qualitative feedback was gathered through small-group interviews to understand usability, usefulness, and integration into workflows.

Results: In phase 1, community health care workers identified barriers, including limited safe patient handling, refresher training, and isolation during tasks. In phase 2, the app was developed to address these safe patient handling needs, incorporating features like scenario-specific guidance for high-risk tasks. In phase 3, the app exceeded success criteria for recruitment, retention, and satisfaction, with participants highlighting its usefulness, usability, and adoption. Qualitative feedback emphasized the app's practical value as a real-time resource, particularly its step-by-step guidance and user-friendly design.

Conclusions: This study met its objectives, highlighting the SPH MSIP app's potential to address community health care workers' unmet safe patient handling needs and improve support in real time patient handling scenarios. While the findings suggest strong feasibility and acceptability, future research should focus on large-scale, extended effectiveness trials to evaluate the app's impact on reducing musculoskeletal injury rates and improving patient care outcomes.

(JMIR Form Res 2025;9:e70681) doi:10.2196/70681

KEYWORDS

ergonomics; mHealth; occupational safety; user-centered design; workplace injuries

XSL•FO
RenderX

Introduction

Background

Patient handling in health care involves lifting, transferring, repositioning, or mobilizing a patient's body [1]. These actions—particularly the forward flexion and twisting that often accompany them—place substantial compressive and shear loads on the lumbar and cervical spine [2,3]. Even under ideal conditions, the weight of an adult often exceeds health care professionals' lifting capacity [2]. Consequently, repeated exposure to high-load and awkward-posture tasks leads to acute sprains and strains of the low back, neck, and shoulders—the most common musculoskeletal injuries linked to patient-handling overexertion [4]. In 2020, the US Bureau of Labor Statistics reported approximately 10,510 workplace injuries in health care due to overexertion and bodily reactions [5]. At the Vancouver Island Health Authority (Island Health) in British Columbia, Canada, from April 1, 2021, to March 31, 2022, patient handling injuries accounted for 25% of all of Island Health's WorkSafeBC (employer work incident insurance) claims, at a cost of approximately US $3.0 million and a total of 21,286 worker time-loss days. The majority of these claims involved musculoskeletal sprains and strains resulting from unsafe patient-handling maneuvers, confirming that manual handling remains one of the leading injury mechanisms among Island Health's direct-care workforce.

Community health care workers provide personal care and daily-living support that routinely involves patient handling, often in clients' homes where access to safety resources and handling tools is limited [6]. They face multiple unmet needs related to safe patient handling, including concerns about the lack of refresher training and a general "get it done" attitude among co-workers that may discourage safe practices [7,8]. This lack of accessible safety resources and support, combined with the challenging environments in which the community health care workers work, further increases their vulnerability to injury. Addressing these gaps is essential to prevent musculoskeletal injuries among community health care workers and to reduce medical costs and the broader burden on health care systems.

A growing body of research shows that interventions can promote safe patient handling and reduce musculoskeletal injuries among health care workers [9], including community health care workers [10]. A systematic review by Wåhlin et al [9] identified 4 categories of successful interventions: the provision and training of assistive equipment (eg, mechanical body lifts), peer coaching programs, physical exercise initiatives, and cultivating a workplace culture with strong management support. Interventions that provide continuous support, rather than one-time training, are more effective [11]. Successful programs not only reduce injuries but also yield cost savings, with studies reporting reduced claims and recurrences after implementation [11-13]. Additionally, emerging technologies, such as real-time audio biofeedback [14] and vibrotactile feedback on patient handling [15], show promise for further enhancing safety practices. However, the practicality and cost-effectiveness of these technologies in real-world health care settings remain uncertain, and their development has not consistently involved feedback from health care workers themselves. Given these proven benefits, despite the constraints of in-person interventions, there is a strong rationale for developing an accessible and effective safe patient handling program tailored to support community health care workers. A digital health intervention, such as a mobile health (mHealth) app, could deliver real-time training, resources, and tools—overcoming logistical barriers and providing continuous support in diverse and often isolated work environments [16,17].

Existing apps such as eUlift [18], Safe Patient Handling [19], and Pain Point [20] illustrate the potential of digital tools, but do not address the realities of community health care workers who provide care in clients' homes, nor do they align with jurisdiction-specific musculoskeletal injury prevention policies; most also lack user-centered co-design with community health care workers and have not been evaluated in this population [10]. Despite extensive guidance and training programs, few tools are designed for community health care workers operating in clients' homes, where equipment is limited, workflows are mobile, refresher training is irregular, and peer norms can discourage best practice [7,8]. This creates a practical need for a low-burden, user-centered microcoaching application that provides context-specific guidance and resources within health-system constraints. This study addresses this gap.

This Study

Evidence-based interventions have demonstrated their efficacy in promoting safe patient handling [9,13], yet logistical and resource barriers often limit their accessibility, particularly for community health care workers in isolated settings [6]. Island Health, Canada, has established the musculoskeletal injury prevention team to promote safe patient handling practices, but its capacity is constrained by the team's small size, the large number of care workers, and the geographical spread of service areas. Despite successfully implementing guidelines, tools, and in-person training, these efforts are limited by costs and logistical challenges [21].

This study aimed to (1) provide detailed insights into the unmet needs of Island Health community health care workers with respect to safe patient handling resources and access to information, (2) translate those needs into a user-centered prototype of the Safe Patient Handling Musculoskeletal Injury-Prevention (SPH MSIP) smartphone app through an iterative co-design process, and (3) establish the acceptability and feasibility of the SPH MSIP app to support community health care workers' safe patient handling practices using a mixed methods design. To our knowledge, this is the first app to offer on-demand safe patient handling tools specifically for community health care workers in their work environments. The study also lays the groundwork for testing the hypothesis that on-demand digital coaching, through videos and support tools, can reduce the frequency and severity of musculoskeletal injuries, as well as related WorkSafeBC claims and costs.

XSL•FO
RenderX

Methods

Methodological Orientation and Operationalization

We adopted a pragmatist orientation to prioritize answers that are actionable for decision-making in this health-system context [22]. Accordingly, we used a mixed methods single-group feasibility design with prespecified progression criteria to judge whether a larger trial is warranted, and we paired brief quantitative signals with qualitative explanations to understand why those signals emerged [23,24].

In addition to pragmatism, this research adopted a Patient-Oriented Research approach [24]. Patient-oriented research spans information-sharing, consultation/collaboration, shared decision-making/full partnership, and patient-led models [24,25]. Since it increases relevance and acceptability, it enhances uptake and dissemination [24,26]. Aligned with this approach, we engaged musculoskeletal injury prevention coaches, app developers, health system decision-makers, and community health care workers with lived experience of musculoskeletal (patient-handling) injuries. This multidisciplinary team optimized the design and feasibility of the SPH MSIP app and supported co-design, implementation, and adoption.

The research methods were further guided by the IDEAS framework [27], a user-centered design process for developing and evaluating digital interventions. IDEAS is an acronym for its 4 overarching stages: Integrate, Design, Assess, and Share, encompassing 10 phases: empathize, specify, ground, ideate, prototype, gather, build, pilot, evaluate, and share [27]. This study specifically addressed the first 3 overarching stages of the IDEAS framework, encompassing several of the framework's phases: Integrate (ie, empathize, specify, and ground), Design (ie, ideate, prototype, gather, and build), and Assess (ie, pilot). It also emphasizes the integration of behavioral theory. The Technology Acceptance Model (TAM; [28-30] was selected to inform the development of research questions and guide data interpretation. The TAM [28] posits that perceived usefulness (belief that a technology enhances performance) and perceived ease of use (belief that it is effortless to use) influence attitudes, behavioral intentions, and adoption [29,30]. These constructs provide a practical framework for designing and evaluating user-centered tools [31], and are particularly relevant for assessing the SPH MSIP app. Widely applied in mHealth and app development [32,33], TAM has proven effective in identifying key factors that drive technology acceptance and sustained use in health care settings.

Study Phase Design

Consistent with our pragmatist and Patient-Oriented Research approach, and using IDEAS to structure activities, we organized 3 phases: phase 1 needs-finding with end users, phase 2 iterative co-design of the SPH MSIP prototype, and phase 3 a single-group mixed methods feasibility study with predefined progression criteria. Specifically, phase 1 merged theoretical insights with end user input to identify unmet safe-patient-handling needs among Island Health community health care workers (objective 1). Phase 2 used iterative co-design with the patient-oriented research team and the app developer to transform those needs into a functional SPH MSIP prototype (objective 2). Phase 3 assessed the prototype's feasibility and acceptability (objective 3) in a single-group mixed methods study [34], combining postintervention questionnaires with qualitative interviews and survey items to contextualize the numbers. While quantitative methods were the primary focus, qualitative data enriched the understanding of participants' experiences. To evaluate feasibility, the study established predefined success criteria, including an enrollment rate of ≥50% and a retention rate of ≥75%. Acceptability was measured using a participant satisfaction threshold of a mean score ≥4 points on a 7-point Likert scale.

Ethical Considerations

Ethical approval was obtained from the University of Victoria Human Research Ethics Board and harmonized ethics approval by Research Ethics BC (REBC#: H23-00322). All participants were adults and provided written, electronic informed consent before any procedures; participation was voluntary, and participants could withdraw at any time or skip any question without penalty. All questionnaire and interview responses were collected in a secure web-based survey platform and Zoom (Zoom Communications), respectively, immediately exported, and stored on a password-protected server in the Behavioural Medicine Laboratory at the University of Victoria. A unique study ID (not names or e-mail addresses) was assigned at first contact; the reidentification key is kept in a separate, encrypted file accessible only to the principal investigator and project coordinator. Audio files were deleted after verbatim transcription. Deidentified analytic files will be retained for 5 years, after which they will be permanently erased from the secure server. The use of the SPH MSIP app during client care was not required; participants were directed to follow employer phone-use and privacy policies and to access content outside client interactions where possible. The app contains training content only and does not collect audio, images, GPS, or client-identifying information. The app was installed on employer-issued work phones.

Phase 1: Patient-Oriented Research Team Building, Unmet Needs Insight, and Engagement

Eligibility Criteria of Participants

The first expressions of interest outlined the patient-oriented research team roles and responsibilities, and 6 employees were selected to participate as members. They consisted of 3 musculoskeletal injury prevention team coaches and 3 community health care workers. They were selected as per the following eligibility criteria. Musculoskeletal injury prevention team coaches qualified based on the following: (1) aged 19 years and older, (2) have a minimum of 12 months of active direct care employment with a Community Health Services (CHS) program at Island Health, (3) have experienced a workplace musculoskeletal injury, (4) are comfortable speaking and reading English, (5) have an Island Health smart phone, and (6) have been a musculoskeletal injury prevention team coach for a minimum of 12 months. The community health care workers qualified based on the following: (1) aged 19 years and older, (2) have a minimum of 12 months of active direct care employment with a CHS program at Island Health, (3) have

XSL•FO
RenderX

experienced a workplace musculoskeletal injury, (4) are comfortable speaking and reading English, (5) have an Island Health smart phone, and (6) have used an musculoskeletal injury prevention team coach in the past 12 months. No additional exclusion criteria were applied for these 2 groups of participants. The research team consisted of one professor with research experience in health psychology, feasibility research designs, and a project coordinator.

Procedure

The research team hosted 3 digital meetings (totaling 5.5 hours) with the patient-oriented research team. The purpose of meeting 1 was team building, providing an overview of the existing Island Health safe patient handling tools, and gathering initial feedback to better understand unmet needs related to safe patient handling resources and information access. This meeting encouraged engagement from all team members and allowed them to share insights and experiences. In meeting 2, the patient-oriented research team reviewed and selected specific safe patient handling tools for inclusion in the SPH MSIP app, discussing each tool's relevance and usability, while also providing suggestions for additional design and functionality features to ensure the app would effectively address identified needs. Based on the TAM [28], the research team focused on collecting feedback related to perceived usefulness (eg, how selected tools could enhance safety practices and reduce injury risks) and perceived ease of use (eg, identifying user-friendly formats for app features). This ensured that user insights guided design decisions, promoting greater acceptance of the app. Meeting 3 was dedicated to finalizing both the selection of safe patient handling tools and the unmet needs assessment. TAM [28] was again applied to prioritize features that enhanced user satisfaction, aligned with community health care workers' workflow, and addressed barriers to adoption, such as usability and resource accessibility. All meeting discussions were recorded, transcribed, and summarized in a report, which was submitted to Pathverse [35,36], the company responsible for developing the SPH MSIP app. This phase established a clear, user-informed foundation for the app's design, aligned with the priorities and insights of the patient-oriented research team.

Phase 2: App Development

In phase 2, a series of 4 meetings was conducted, during which the patient-oriented research team collaborated with Pathverse to coideate and codevelop the SPH MSIP app prototype. The team reviewed the selected safe patient handling tools in detail with the app developer, allowing for targeted discussions about integrating these tools into the app and addressing technical considerations. Feedback from this session guided initial design choices. Then, Pathverse incorporated patient-oriented research team feedback into the evolving prototype, presenting updated versions for further review and input. Using TAM [28,30] as a guiding framework, these successive cycles emphasized the enhancement of perceived usefulness and ease of use, while iterative design revisions focused on ensuring the app's compatibility with users' work environments. The collaborative approach prioritized the alignment of app features with user needs, fostering positive attitudes toward adoption and usage intentions. By the final meeting, the prototype had been refined and finalized, fully integrating the patient-oriented research team's insights and meeting the established goals for functionality and user experience. This collaborative approach ensured the app was optimized effectively, resulting in a practical, user-centered tool for safe patient handling support.

Phase 3: Feasibility Trial

Participants

Eligibility Criteria

Invitations were sent out to employees working in Island Health. The community health care workers qualified for participation based on the following eligibility criteria: (1) aged 19 years and older; (2) a minimum of 12 months of active direct care employment with a CHS program at Island Health; (3) preferably, have experienced a workplace musculoskeletal injury; (4) comfortable speaking and reading English; (5) own an Island Health smartphone; and (6) preferably, have used an musculoskeletal injury prevention team coach within the past 12 months. Community health care workers' musculoskeletal injury prevention team coaches were required to meet the same criteria, with the additional qualification of having served as a musculoskeletal injury prevention team coach for at least 12 months. No additional exclusion criteria were applied.

Procedure

The study took place from September to October 2023. Upon receiving consent, participants attended a 1-hour orientation and training session in September 2023, where they were introduced to the SPH MSIP app and prepared for the study. Participation was entirely voluntary, and participants were informed of their right to withdraw from the study at any time without consequence. If they chose to withdraw before completing the one-month study period, they could decide whether their data collected up to that point would be included. Participants were asked to use the SPH MSIP app for one month, from October 1 to October 31, 2023, on their work smartphones. The app provided them with on-demand, real-time videos and resources as a resource for safe patient handling in their work environments. At the end of the month, participants completed a follow-up questionnaire and participated in one of 4 small group interviews to discuss their experiences with the app. In recognition of their time, participants received in-kind wage compensation through their regular CHS programs, with additional reimbursement for travel and mileage expenses as needed. Each participant who completed the study also received an honorarium as a token of appreciation.

Intervention Materials

The SPH MSIP app was developed in response to the unmet needs identified in phase 1, serving as a comprehensive and user-friendly tool to support community health care workers in implementing safe patient handling practices. Guided by the TAM [28,30], the app incorporated 3 primary intervention components designed to address mechanisms of action related to perceived usefulness, perceived ease of use, and behavioral intention, ensuring the tool was practical, engaging, and compatible with the workflows of community health care workers.

XSL•FO
RenderX

Just-in-Time Training Videos

The app provided short instructional videos, each no longer than 1 minute and 30 seconds, offering real-time guidance for safe patient handling tasks. These videos emphasized key practices, such as completing a "Point of Care Patient Handling Risk Assessment" and demonstrating safe techniques like boosting in bed or using patient handling equipment. The videos were designed to enhance perceived usefulness by providing actionable, context-specific guidance, while also addressing perceived ease of use through intuitive accessibility. Videos were accessible through scrolling the main page or via a tailored algorithm, which guided users through a 3-question decision tree to recommend the most relevant video for their scenario. Additional tips, such as reminders to adopt the "STABLE" posture, were provided before task execution.

Standard Procedure and Safe Practice Documents

A dedicated resource section housed mobile-friendly PDF documents, including key guides like the "Pre-Standing Safety Check" and information on "Musculoskeletal Injuries Risk Factors." These materials provided written reinforcement of safe patient handling protocols, enabling community health care workers to review and confirm best practices in a clear, accessible format. This feature was designed to improve the perceived usefulness of the SPH MSIP app [30].

Frequently Asked Questions

The app included a robust frequently asked questions section developed collaboratively with the patient-oriented research team in phases 1 and 2. This feature addressed common questions on safe patient handling practices and processes, such as when to use a transfer belt or how to manage ergonomic hazards during client care tasks like donning compression stockings. This component enhanced perceived ease of use by proactively resolving potential barriers to app engagement and reinforced behavioral intention by ensuring the app aligned with users' practical concerns.

By integrating these TAM-guided features [28], the SPH MSIP app was designed to function as an all-in-one pocket resource, providing perceived usefulness through its comprehensive content, perceived ease of use via user-friendly navigation and personalization, and fostering behavioral intention to adopt and sustain app use. Detailed intervention materials are summarized in Table 1.

XSL•FO
RenderX

Table 1. Safe Patient Handling Musculoskeletal Injury-Prevention (SPH MSIP) app contents.

Intervention components	Examples
Just-in-time training videos (25 modules)	• One Person Transfer • One Person Slide and Turn Away • One Person Turn Towards • Two Person Head of the Bed Boost • Two Person Side of the Bed Boost • Two Person Kneeling on Bed Boost • Wheelchair Reposition: One Person Seated Scoot • Wheelchair Reposition: Two Person Raise and Roll • Wheelchair Reposition: Two Person Drawsheet Method • Using a Repositioning Sling • Using a Universal Sling • Using the Band Sling • Using a Hammock Sling Strap Configurations • Using the SARA Stedy • Using the SARA 3000
Standard procedure and safe practice documents (6 PDF resources)	• Bed Reposition: Slide to the side • STABLE: Tips to help reduce injuries • Point of Care Risk Assessment • Pre-Standing Safety Check
Frequently asked questions (33 questions and answers)	• Q: What is the recommended weight limit for manually lifting clients or parts of clients in ideal conditions? • A: 16 kgs (35lbs). The goal of safe client handling programs should be to eliminate all manual lifting whenever possible. • Q: Should a bed be height adjustable? • A: A client may need to be assessed by a Community Health Services clinician for a height adjustable bed to ensure workers can achieve safe working heights for all required tasks • Q: What is general safe working height for heavy work? • A: Heavy work requires large amounts of force. Generally a working height of 65-95 cm (23-37.5 inches) is considered safe for heavy work. Examples of heavy work could include: boosting, turning or repositioning clients. • Q: When do I use a transfer/gait belt? • A: Must be used during every manually assisted transfer of a client. Examples of these situations include but are not limited to: standby assist, one- or two-person transfer, wheelchair to chair, chair to toilet or commode, bed to wheelchair with or without a transfer pole, walking with or without a walking aid, when a patient is at risk of falls and repositioning in a wheelchair. • Q: How do I secure the transfer/fait belt on a client? • A: The belt should be fastened securely around the hips of the client. It should be snug but still comfortable. You should be able to place one or two finger widths snugly between the client and the belt when tight. • Q: When do I NOT use a transfer/gait belt? • A: A transfer/gait belt should never be used to “lift” a client. And these belts are NOT to be worn by care providers as a handhold for the client. • Q: What general ergonomic hazards when I assist clients to don medically prescribed compression stockings? • A: Donning of compression stockings is a hazard for workers due to the ergonomic risk factors involving: awkward postures, forces and static postures. • Q: What is equipment best practice for assisting clients to don medically prescribed compression stockings? • A: Clients who require assistance with donning and doffing of compression stockings when the compression is greater than 20mmHg should have appropriate donning and doffing equipment available and accessible for CHW use. • Q: What’s considered safe floor transitions? • A: Transitions between floor types are smooth with minimal gaps or changes in height. • Q: What are the recommendations for wheels and castors? • A: The type of wheels and castors attached to wheeled equipment must be appropriate for the type of flooring present to reduce ergonomic risk.

XSL•FO
RenderX

Measures

Quantitative Outcomes

Recruitment and Retention

Recruitment rate was calculated by dividing the number of participants who expressed interest in the study by the total number of spots available for the study. Retention was calculated by dividing the number of participants who completed the final questionnaire and exit interview by the number of participants who attended the prestudy orientation sessions.

Satisfaction, Acceptability, and Feasibility

A SPH MSIP app-specific TAM questionnaire was developed and adapted based on existing studies using TAM [28,30,37] to evaluate the participant experiences with the SPH MSIP app after the 1-month trial period. Satisfaction and feasibility were measured using a 7-item satisfaction and evaluation questionnaire on a 7-point Likert scale, ranging from 1 (Strongly Disagree) to 7 (Strongly Agree). Sample items were "Overall I enjoyed using the Safe Patient Handling MSIP Application," "If the SPH MSIP APP was available after this I would continue to use it," "I am open to using technology like the SPH Virtual Coach APP," and "It makes sense for me to use the SPH Virtual Coach APP in my work life."

Qualitative Outcomes

Qualitative data retrieval was completed using a small group interview-based approach (4 sessions) with participants over a 2-week period (November 2023). This data was used to further understand the quantitative outcomes. The interview was conducted using the digital conferencing platform Zoom and lasted approximately one hour. Interviews were conducted by 2 different interviewers. Both interviewers were trained and engaged in practice interviews prior to interviewing participants. Interviews were transcribed verbatim by multiple research assistants. These sessions explored TAM constructs [28,37], including (1) perceived usefulness/appropriateness/demand (eg, "Do you think this app is useful in reducing workplace MS Injury risk for yourself?"), (2) perceived ease of use (eg, "Are there certain features that made the app more usable? If yes, what are they?"), (3) perceived enjoyment/satisfaction (eg, "What, if anything, could contribute to making the app more enjoyable for you to use?"), (4) Intention to use/continue use (eg, "Can you see yourself using the app beyond this study?"), (5) actual use (eg, "How often did you use/look at the app in a week? Or in the month?"), (6) compatibility (eg, "Does using the app fit into your current work lifestyle?"), (7) attitude (eg, "Is there any apprehension you feel about using this app while working with clients?"), and (8) self-efficacy (eg, "How confident do you feel using an app like this in your workplace daily?"). The complete interview guide was included as Figure 1.

Figure 1. Screenshots of Safe Patient Handling Musculoskeletal Injury-Prevention (SPH MSIP) app.

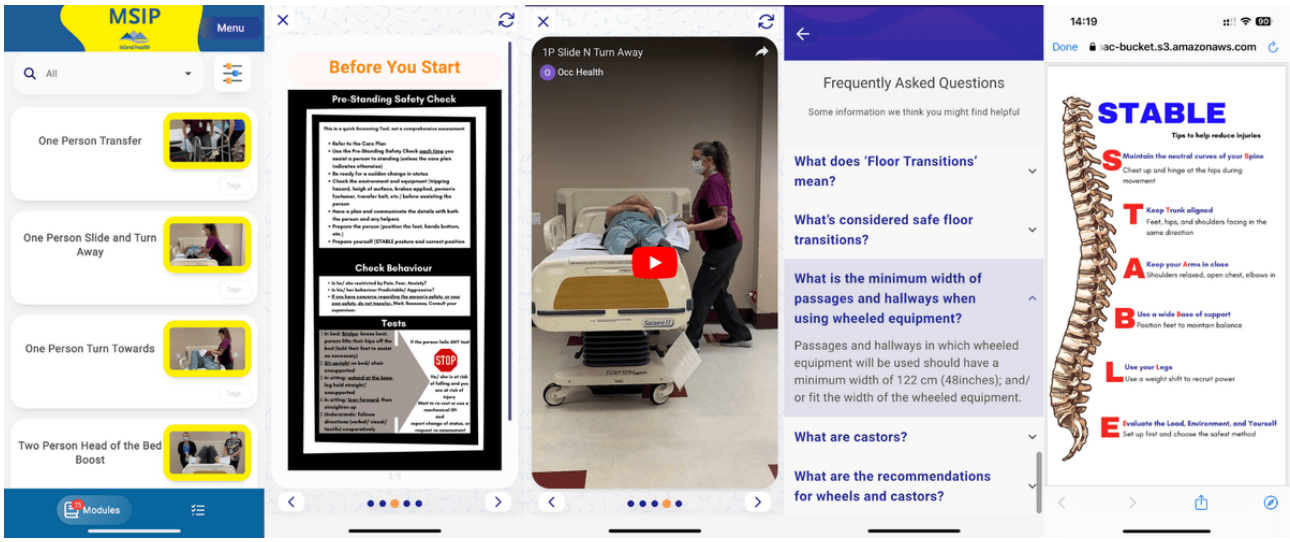

Data Analysis

Quantitative Outcomes

The rating system established for quantitative data analysis was based on the traffic light system from Lewis et al [38], which recommends that progression be based on all key feasibility criteria, with the decision ultimately being based on the worst-performing criterion. The AMBER (middle) zone may be divided into 2 cutoff points to indicate when minor or major amendments are necessary prior to proceeding [38].

Qualitative Outcomes

Interviews were audio-recorded, transcribed verbatim, and analyzed in Excel by AL and SW to identify common themes regarding the app's strengths and limitations. Using a constructivist approach [39], this analysis acknowledged that knowledge is coconstructed between researcher and participant, with meaning derived through interpretation rather than objective discovery. Following Braun and Clarke's [40] reflexive thematic analysis, AL and SW began by familiarizing themselves with the transcripts and open-ended responses, reading closely to capture initial impressions. They conducted initial coding to categorize data based on broad aspects such as

XSL•FO
RenderX

feasibility, acceptability, and intervention outcomes, grouping these codes into preliminary themes and subthemes. These preliminary groupings were then presented to SB, SC, and RR (supervisor) for feedback to refine the thematic structure. Following discussions, the team refined and finalized the themes, which captured participants' views on the app. Representative quotes were then selected to highlight these themes, providing concrete examples of the app's perceived strengths and limitations. This iterative process enabled a comprehensive understanding of participants' experiences, supported by both individual coding and collaborative reflection.

Results

Results of Phase 1

Based on phase 1 meetings, we identified key unmet needs among community health care workers related to safe patient handling and injury prevention, as well as specific feedback on safe patient handling tools that could enhance the SPH MSIP APP under development. Several major concerns and needs were expressed (Textbox 1 for details), including the high risk of injury due to client care tasks, frequent work in isolation at client homes, and a tendency to prioritize client needs over personal safety. Furthermore, gaps in safe patient handling knowledge and practices were noted, such as limited access to training on point-of-care risk assessment and challenges in obtaining refresher training on safe handling practices. Additionally, a "get it done" culture in the workplace, alongside inconsistent proficiency among leaders in safe handling techniques, emerged as a significant issue affecting community health care workers' safety and effectiveness. Notably, participants also highlighted specific high-risk patient handling activities that require additional support and clear guidance, including boosting clients in bed, assisting clients with compression stockings, providing care in low or bariatric beds, and assisting clients with walking in confined spaces. Essential equipment used in these tasks, such as slings, transfer belts, and slider sheets, requires proper usage training, and community health care workers expressed a need for accessible, step-by-step instructions. In addition, it was made clear by participants that the app needed to address practical questions that community health care workers often face, such as "How do I best use my body in this situation?" and "What is the best practice for this specific patient care task?" Overall, community health care workers emphasized that the app should provide easily accessible, concise, and practical information that could be accessed without disrupting client interactions.

XSL•FO
RenderX

Textbox 1. Phase 1 feedback from community health care workers.

Unmet needs identified by community health care workers in phase 1

- Concerns of general potential for client care and handling injury
- Concerns of injuries can happen fast
- Concerns of often working alone in clients' homes to provide care
- Gaps in knowledge regarding safe practice, safe assist weight, safe work expectations
- Gaps in knowledge and use of point-of-care risk assessments
- Concerns that co-workers become injured
- Concerns regarding "get it done" attitude
- Concerns of "problem solve on the spot; alone"
- Concerns of "even the nurse leaders aren't always proficient on safe patient handling and techniques"
- Community health care workers—very generous meeting client needs and sometimes push themselves beyond what is physically safe—putting the client first
- Concerns of community health care workers are in client's homes—a more intimate and personal environment
- Concerns of creating appropriate work boundaries in client's homes
- Concerns about no safe patient handling refresher training easily available for employees
- When have in person refresher training it is time consuming
- Access to the prehandling safety check tool in the moment
- Communication routes for reminders, refreshers, updated safe practice information that is easy to access and review quickly
- Top high risk patient handling activities:
  - Bed boosts
  - TED stockings/compression stockings
  - Bariatric bed care
  - Assisted walking clients with transfer belt
  - Showering clients in tight spaces
  - Low beds
  - Standing peri-care
- Top equipment used:
  - Slings
  - Slider sheets
  - Sara Stedy
  - Overhead lift systems
  - Floor lift systems
  - Hospital beds
  - Etac Turner (standing pivot disc)
  - EZ disc
  - Transfer belts
  - Transfer board
  - Bed Ladders (rare)
  - Sask-a-pole (transfer pole)
- Work Smartphones: I would just like our phones to actually work- better connectivity
- Worker can be some lack of comfort pulling out and using phone in front of the client (clients might be suspicious/ added confusion/ do not understand)
- Practice questions (concerns):

XSL•FO
RenderX

- How do I use this piece of equipment? Example: sara stedy
- How do I best use my body in this situation?
- What is best practice for this pt care and handlings situation?

- Training resources for teaching/co-working with client family members
- App must be easy to navigate
- Concerns about “language barriers” for employees/clients using the App. Some employees have hard time with English
- Need clear answers/ or clear video or clear explanation on the App that does not need to be reinterpreted or reclarified
- Easy to access information and resources and to maneuver through the information in a timely way. Simple and fast

Results of Phase 2

Building on the unmet needs identified in phase 1, phase 2 focused on designing a user-centered SPH MSIP APP tailored to community health care workers’ challenges in safe patient handling practices (Table 2). Collaborating with Pathverse, the app addressed critical gaps in safe patient handling knowledge, training accessibility, and the unique challenges of working in isolated client homes. The app provided clear, practical guidelines for high-risk tasks such as boosting clients in bed, assisting with compression stockings, and managing bariatric or low beds, as well as proper use of commonly used equipment like slings, transfer belts, and slider sheets. Tools such as point-of-care risk assessment and pre-standing safety check guides were also included to support safety and decision-making. The SPH MSIP APP improved training accessibility with mobile-friendly resources and concise, scenario-specific instructional videos. A tailored algorithm streamlined access to relevant materials, minimizing disruptions during client care. Although some tools and equipment (eg, Sara Stedy, Etac Turner) fell outside the project’s scope, most high-priority resources were successfully integrated, offering a reliable and practical resource to enhance community health care workers’ safe patient handling practices. Educational content (training videos, PDF, and equipment checklists) was adapted directly from Island Health’s Occupational Health and Safety safe-patient-handling resources. All materials were produced or reviewed by MSIP coaches, occupational health and safety specialists, and professionals to ensure they reflected standard device use and current workplace guidelines before being embedded in the app [41,42]. Screenshots of the SPH MSIP APP can be found in Figure 1.

XSL•FO
RenderX

Table 2. Phase 2 unmet needs addressed.

Unmet needs identified in phase 1 by the patient-oriented research team	How we addressed unmet needs in phase 2
Gaps in knowledge regarding safe practice, safe assist weight, safe work expectations	Provided SPH MSIP[a] app
Gaps in knowledge and use of point-of-care risk assessments	Included point-of-care risk assessments in SPH MSIP app
Concerns of 'problem solve on the spot; alone"	Provided SPH MSIP app
Concerns of "even the nurse leaders aren't always proficient on safe patient handling and techniques"	Provided SPH MSIP app and invitation to share experience with other colleagues and leaders
Concerns: Community health care workers are in client's homes—a more intimate and personal environment	Provided SPH MSIP app
Concerns of creating appropriate work boundaries in client's homes	Provided point-of-care risk assessments in SPH MSIP app
Concerns about no safe patient handling refresher training easily available for employees	Provided SPH MSIP app
When have in person refresher training it is time consuming	Provided SPH MSIP app
Access to the pre-handling safety check tool in the moment	Included Pre-handling Safety Check tool in SPH MSIP app
Communication routes for reminders, refreshers, updated safe practice information that is easy to access and review quickly	Provided SPH MSIP app
Top high risk patient handling activities: • Bed boosts. • TED stockings/compression stockings • Bariatric bed care • Assisted walking clients with transfer belt. • Showering clients in tight spaces • Low beds • Standing peri-care	Included information or guidelines for all activities in SPH MSIP app • Except standing peri-care and showering clients in tight spaces
Top equipment used: • Slings • Slider sheets • Sara Stedy • Overhead lift systems • Floor lift systems • Hospital beds • Etac Turner (standing pivot disc) • EZ disc • Transfer belts • Transfer board • Bed Ladders (rare) • Sask-a-pole (transfer pole)	Included information or guidelines for all equipment listed in SPH MSIP app • Except Etac Turner, Sara Stedy, EZ Disc, Transfer Board, Transfer pole
Work Smartphones: I would just like our phones to actually work—better connectivity	Ensured Island Health IM/IT and Pathverse provided best service for use of SPH MSIP app
Worker can be some lack of comfort pulling out and using phone in front of the client (clients might be suspicious/ added confusion/ do not understand)	Ensured resources in SPH MSIP app were short duration and easy to view
Practice questions (concerns): How do I use this piece of equipment? (Example: sara stedy). How do I best use my body in this situation? What is best practice for this pt care and handlings situation?	Provided resources in SPH MSIP app for practice questions in the videos, frequently asked questions, resources sections
Training resources for teaching/co-working with client family members	Provided SPH MSIP app and encouraged to share with colleagues for training purposes
App must be easy to navigate	Provided insights to Pathverse, the app development company
Need clear answers/or clear video or clear explanation on the app that does not need to be reinterpreted or reclarified	Provided best in case videos and resources in SPH MSIP app
Easy to access information and resources and to maneuver through the information in timely way. Simple and fast.	Provided SPH MSIP app without barriers to entry to Pathverse app platform once loaded on work smartphone

[a]SPH MSIP: Safe Patient Handling Musculoskeletal Injury-Prevention.

XSL•FO
RenderX

Results of Phase 3

Quantitative Outcomes

Recruitment and Retention

Thirty-two eligible staff members were invited, and 28 out of 32 (88%) participants enrolled—14 community health care workers and 14 community health care workers musculoskeletal injury prevention team coaches—from Port Alberni, Cowichan, Nanaimo, and Comox. Of those enrolled, 26 out of 28 (93%) participants completed all aspects of the study, with 2 completing the questionnaires but not the exit interview. The recruitment and retention rates exceeded the predefined success criteria (Table 3).

Table 3. Participant and peer mentor intervention feasibility and acceptability success criteria.

Outcome	Success criteria	Results
Enrollment rate, %	≥50	88
Retention rate, %	≥75	93
Satisfaction score (points), mean	≥4	>5.96

Satisfaction

Enjoyment of the SPH-MSIP app was high: 22 out of 28 (79%) participants agreed or strongly agreed that they enjoyed using it, 3 out of 28 (11%) participants somewhat agreed, and 3 out of 28 (11%) participants somewhat disagreed. This resulted in a mean score of 5.96 out of 7, significantly exceeding the predefined success criterion of a mean score of 4 (Table 3).

Acceptability and Feasibility

Participants expressed a strong overall willingness to use the SPH MSIP APP in the future, should it remain available after the study. A total of 23 (82%) participants agreed or strongly agreed with continuing to use the app. Openness to using similar technology in the future was even higher, with 26 participants (93%) agreeing or strongly agreeing. A majority—24 of 28 (86%) participants—felt that using the app in their future work life made sense, and 20 (71%) participants agreed or strongly agreed that it fit well into their workflow. Confidence in learning to use the app in the future was high, with 25 (89%) participants agreeing or strongly agreeing. Participants' belief in their ability to use the app was equally strong, with 25 (89%) participants agreeing or strongly agreeing that they had the necessary skills. There were no "strongly disagree" responses and only a small number of "disagree" responses across all items. Notably, one participant (1/28, 4%) disagreed with continuing to use the app if available, and one participant (1/28, 4%) disagreed with confidence about learning to use the app (Table 4).

Table 4. SPH MSIP app satisfaction and feasibility.

Questions	Strongly agree, n (%)[a]	Agree, n (%)[a]	Somewhat agree, n (%)[a]	Neither agree nor disagree, n (%)[a]	Somewhat disagree, n (%)[a]	Disagree, n (%)[a]	Strongly disagree, n (%)[a]	Mean (SD)
1. Overall, I enjoyed using the SPH MSIP app	11 (39)	11 (39)	3 (11)	—[b]	3 (11)	—	—	5.96 (1.23)
2. If the SPH MSIP app was available after this, I would continue to use it	15 (54)	8 (29)	3 (11)	—	1 (4)	1(4)	—	6.18 (1.25)
3. I am open to using technology like the SPH Virtual Coach app	15 (54)	11 (39)	1 (4)	—	1 (4)	—	—	6.36 (1.03)
4. It makes sense for me to use the SPH Virtual Coach app in my work life	12 (43)	12 (43)	3 (11)	1 (4)	—	—	—	6.25 (0.80)
5. I feel confident about learning to use the SPH Virtual Coach app	11 (39)	14 (50)	1 (4	1 (4)	—	1 (4)	—	6.14 (1.08)
6. I have the necessary skills to use the SPH Virtual Coach app	12 (43)	13 (46)	2 (8)	—	1 (4)	—	—	6.25 (0.89)
7. Using the SPH MSIP app is compatible with most aspects of my work life	4 (14)	16 (57)	5 (18)	2 (8)	1 (4)	—	—	5.71 (0.94)

[a]Percentages may not total 100% due to rounding.

[b]Indicates 0 responses (0%).

XSL•FO
RenderX

Qualitative Outcomes

Overview

Qualitative data were collected from 26 participants through small group interviews conducted over 4 sessions in a 2-week period. In general, the participants found the SPH MSIP APP to be an impactful tool in supporting their roles as community health care workers and musculoskeletal injury prevention coaches. The themes, subthemes, and selected quotes are presented in Table 5. Many participants described the app as a "game changer," emphasizing its ease of use, the relevance of its content, and its ability to enhance their confidence in executing care tasks safely and efficiently. They highlighted its effectiveness in bridging gaps in knowledge and its value as a resource for immediate reference, skill refreshment, and peer coaching. Nonetheless, participants also offered constructive feedback for improvement, including suggestions for enhancing app functionality, content comprehensiveness, and integration into their daily routines. Thematic analysis revealed 3 major themes—usefulness, usability, and adoption—each comprising several subthemes that encapsulate participants' feedback and reflections on the SPH MSIP app.

XSL•FO
RenderX

Table 5. Themes, subthemes, sub-subthemes, and selected quotes supporting the themes.

Main themes, subthemes, and sub-subthemes	Examples of quotations from participants
Usefulness	
Perceived effectiveness	
Pros	• It was nice to be able to review prior to coaching someone or when somebody was asking something, I was able to pull it out and show them what I meant; that made a really big difference. [SPA015] • It's easier for in the moment to show your peer or colleague that you're working with. As a coach before I'd have to go back to the office and email the how to or video or link. And so it's nice to have it right there. [SPA009] • It perhaps might remind you to be a little more present and mindful. These videos really kind of force you to pay more attention. Slow down a bit, be a little more mindful of all the little details that are being pushed in these videos that perhaps you might become complacent to [SPA008]
Cons	• (How does using the app compare to your experience using a human Musculoskeletal Injury Prevention team coach to support you in your workplace?) It's a great tool to use as support for a Musculoskeletal Injury Prevention team coach coming in, but not as a replacement for a coach. Having a human watching you and seeing where your body mechanics are is very valuable. [SPA025]
Content assessment	
Pros	• I really appreciate that the videos, were short and sweet. Sometimes we don't have a lot of time to go it was just right to the point [SPA020] • The video is really helpful. People tend to be visual. We only have a few minutes of our time just to check on that phone so we are able to imagine how we can do it. So, the video is more helpful for me. [SPA005] • The frequently asked questions I found very useful, they're pretty comprehensive [SPA019]
Cons	• (What type of content was missing from the app that you would like to see included?) Stocking application would've been a good one to have in there, like how to put on TED stockings properly [SPA004] • (What type of content was missing from the app that you would like to see included?) Transferring somebody onto a bath bench or a shower bench or a bathing chair. [SPA018] • It was weird to have the point of care risk assessment video and then the poster. people might just skip that to get to the point. Maybe it should just be completely together so that it’s repetitive that we do need to do our point of care risk assessment. One of the biggest things is that we could prevent injuries with just that simple tool. I appreciate it's there for every video, but instead of having the two or three little dots that you can slide through, just combine it all into one video. [SPA024]
Usability	
Perceived ease of use	
Pros	• I had one of our Musculoskeletal Injury Prevention team coaches who, she says she's not good with the cell phone. She is my age. It took five minutes with her and that's all she needed [SPA017] • I liked how straightforward the app was when I open it up, it is quick and bold and easy to read the different sections [SPA020] • It's been a lot easier than navigating than the SharePoint and the intranet pages. So, we used it a lot, when developing plans. So, it's been the ease of access that has been very nice. [SPA009]
Cons	• Is there a way that the videos can be saved so that we don't have to wait for them? [SPA006] • It'd be cool if the videos could be like embedded somewhere that we can access 'em right on our phone without having to access the internet, you couldn't load the videos if you didn't have internet [SPA004] • If there was a way to search FAQ'sa, I couldn't find a way to search for a keyword [SPA019]
Compatibility	
Pros	• We all use our phones for work already, so it's pretty easy to use the app because we do so much work on our phones, it's not that much of a leap to now jump on an app [SPA004] • I did show a client a video because of a way that she wanted me to adjust her in the wheelchair and I showed her that we aren't allowed to do it like that, and then I showed her how I was going to do it and she was super receptive to it and then everybody just started doing it like that cause I would show the next worker and the next worker, and then she would tell them this is how we're gonna do it now. [SPA004] • I'd like to use it as part of my toolset for coaching. Coaching the CHW's to check this video out quickly before we go in?’ they can prepare themselves for what they're doing. it makes it easier for me to explain where they could improve [SPA025]

XSL•FO
RenderX

Main themes, subthemes, and sub-subthemes	Examples of quotations from participants
Cons	• (Is there any apprehension you feel about using this app while working with clients?) I did explain to them why as well. So, they're a little bit more open to it after that. But the first reaction was, 'it's kind of rude being on your phone in my house' sort of thing. [SPA027] • I think the only apprehension I would have is probably time constraint, unfortunately, our time has become shorter and shorter with each client and the only issue I might have is that there might not be time to stop and look at your phone and see what's going on. [SPA007] • Some clients don't appreciate us being on our phones when we're in their homes [SPA004]
Adoption	
Actual use and frequency	
Pros	• (How often did you use/look at the app in a week?) I was about twice a week [SPA001] • We were using it daily. We're showing everybody we possibly could. We'd bring people down to show them. [SPA002] • I was doing quite a bit of coaching in the field around the time we got the app, and it was really helpful because there was no place to demonstrate anything to the CHWs before they came into the client's home. It was a really good way to show them the videos and make sure we are all on the same page in the driveway. Every time I coached, I would show the video. It was every day that I coached. [SPA019]
Intent to use	
Pros	• I would just like to say that you know, just continuing to use it as an education tool, especially with our new hires. I don't always have a room with a lift and can't go over every piece of part, like we, I can talk about all the slings, I can talk about how they're used, but to be able to have the video and have them pull it up in their own hands and talk about it in the motion [SPA003] • (Can you see yourself using the app beyond this study?) Yes - It only makes sense. It's literally like an app to make my job easier. [SPA018] • (Can you see yourself using the app beyond this study?) Yes - I think it was great for home support because it's a good refresher. we can go weeks when it's just, basic care, and then all of a sudden, we have a client who has an in-home overhead lift, full bed slider.It's good to look refresh on the app and just see, 'okay, this is how it's done,' [SPA001]
Preparedness and confidence	
Pros	• (What would help you feel more prepared to use the app in the workplace?) Advertise the app around the office [SPA011] • (What would help you feel more prepared to use the app in the workplace?) Each department sending out a look to the future kind of email. Like this will coming soon to our Island Health work phones. [SPA018] • (How confident do you feel using an app like this in your workplace daily? Scale of 1-5 (5 = extremely confident 1 = not confident at all) Five for confident over here. [SPA007, SPA022, SPA004, and SPA028]

[a]FAQ: frequently asked questions.

Usefulness

The theme of usefulness encompassed participants' perceptions of the app's effectiveness and the quality of its content, categorized into 2 subthemes: perceived effectiveness and content assessment.

Perceived Effectiveness

Participants described the SPH MSIP APP as "useful," "very useful," and "helpful," with many affirming that it "made a really big difference" in their work. The app was praised as a reliable, consistent, and readily accessible tool for reviewing proper techniques before executing tasks or coaching peers. One participant noted, "These videos really kind of force you to pay more attention, slow down a bit, and be a little more mindful of all the little details that are being emphasized." The app's role as a training tool for new hires and employees in CHS was particularly valued, especially by musculoskeletal injury prevention coaches. One coach shared, "It's easier for in the moment to show your peer or colleague that you're working with. As a coach before I'd have to go back to the office and email the how to or video or link. And so it's nice to have it right there." Despite these advantages, some participants emphasized that the app could not fully replace the role of a human musculoskeletal injury prevention team coach. They highlighted the unique value of having an experienced coach provide personalized feedback and guidance. One participant remarked, "It's a great tool to use as support for a Musculoskeletal Injury Prevention team coach coming in, but not as a replacement for a coach. Having a human watching you and seeing where your body mechanics are is very valuable."

Content Assessment

The app's videos received widespread appreciation, with participants praising their clarity and relevance. One participant explained, "The video is really helpful. People tend to be visual. We only have a few minutes of our time just to check on that phone so we are able to imagine how we can do it." Several participants also acknowledged the usefulness of the frequently asked questions; one participant commented, "The frequently

XSL•FO
RenderX

asked questions I found very useful, they're pretty comprehensive." However, participants identified areas for improvement, including the addition of videos on tasks such as applying TED stockings or transferring patients to shower benches, "Stocking application would've been a good one to have in there, like how to put on TED stockings properly" and "Transferring somebody onto a bath bench or a shower bench or a bathing chair."

Usability

The theme of usability highlighted participants' experiences with navigating and integrating the app into their workflows, focusing on 2 subthemes: perceived ease of use and compatibility.

Perceived Ease of Use

Most participants found the app intuitive and easy to navigate, with one musculoskeletal injury prevention team coach remarking, "I had one of our coaches who, she says she's not good with the cell phone. She is my age. It took five minutes with her and that's all she needed." Compared with existing resources like the Island Health intranet, the app was described as more accessible and user-friendly: "It's been a lot easier than navigating the SharePoint and the intranet pages. So, we used it a lot, when developing plans. So, it's been the ease of access that has been very nice." However, technical issues such as slow video loading and the need for an internet connection to access content were noted: "It'd be cool if the videos could be like embedded somewhere that we can access 'em right on our phone without having to access the internet, you couldn't load the videos if you didn't have internet."

Compatibility

The app was praised for its seamless integration into participants' existing work routines. Community health care workers noted that having the app on their work smartphones aligned well with their frequent use of these devices. As one participant observed, "We all use our phones for work already, so it's pretty easy to use the app because we do so much work on our phones, it's not that much of a leap to now jump on an app." However, some participants encountered challenges, such as concerns about appearing unprofessional when using phones in clients' homes ("I did explain to the client why as well. So, they are a little bit more open to it after that. But the first reaction was, 'it's kind of rude being on your phone in my house' sort of thing"). The majority of community health care worker participants enjoyed using the SPH MSIP APP. There was a high amount of adoption and usage among this group of community health care worker participants, who were engaged throughout and provided multiple suggestions for continued improvement of content and platform/delivery/experience.

Adoption

The theme of adoption examined the extent to which participants used the app, their intent to continue using it, and their preparedness and confidence.

Actual Use and Frequency

Participants reported using the app with varying frequency, ranging from daily use during coaching sessions to twice a week for quick reference. One coach explained, "It (the app) was a really good way to show them the videos and make sure we are all on the same page in the driveway. Every time I coached, I would show the video. It was every day that I coached."

Intent to Use

Participants expressed strong intent to continue using the app, describing it as a valuable resource for self-reflection, skill refreshment, and task preparation: "Yes - I think it was great for home support because it's a good refresher." When asked if they envisioned using the app beyond the study, responses were positive: "Yes - It only makes sense. It's literally like an app to make my job easier," accompanied by visual affirmations like thumbs up during group interviews.

Preparedness and Confidence

When asked what would help you feel more prepared to use the app in the workplace, some participants said it was already easy enough to use, and some recommended "Advertise the app around the office." Most participants felt confident using the app, with many rating their confidence as "5" on a scale of 1 to 5. They appreciated the app's role in instilling confidence not only in their own care practices but also in communicating these practices to clients and their families.

Discussion

Overview

This 3-phase formative study showed that a co-designed SPH MSIP smartphone app is both feasible and acceptable for community health care workers. We achieved high recruitment (28/32, 88%), excellent retention (26/28, 93%), and strong user satisfaction/usability (mean 5.96/7), while qualitative feedback highlighted the app's practicality, clarity, and seamless fit within everyday workflows. This study represents one of the few comprehensive efforts to develop and evaluate an app tailored specifically to the needs of community health care workers.

Principal Findings

Unmet Needs and Community Health Care Workers' Feedback

Our study revealed multiple unmet needs in safe patient handling resources for community health care workers, highlighting systemic barriers to safe handling practices. Key issues included the lack of accessible refresher training, inconsistent implementation of risk assessments, and workplace cultures that often prioritize efficiency over safety [43]. These findings align with previous research [44], which documented similar challenges in health care environments, where frontline workers face considerable strain due to inadequate support and resources [45]. Notably, the "get it done" attitude observed among coworkers mirrors findings from studies [8,46] emphasizing the disconnect between institutional safety policies and actual practices in health care settings.

This study adopted the Patient-Oriented Research approach, which was instrumental in addressing these barriers through a structured co-design process [24-26]. In this study, the Patient-Oriented Research team actively collaborated with

XSL•FO
RenderX

researchers and app developers to shape the SPH MSIP APP, contributing to decisions on content, functionality, and design. This inclusive approach ensured that the app was both practical and relevant, addressing the realities of the workers' diverse and often challenging environments [25]. The co-design process enabled the research team to incorporate feedback on essential features, such as scenario-specific guidance for high-risk tasks (eg, transferring patients) and intuitive navigation, directly addressing worker needs and enhancing the app's usability [47]. It also reflects a growing recognition in the literature that digital tools can bridge gaps in on-the-job training by offering immediate, context-relevant support [48,49]. From a TAM perspective [28], the co-design approach prioritized perceived usefulness and ease of use, both critical factors for adoption. For instance, the inclusion of step-by-step instructions for tasks like boosting patients in bed [50] directly addressed the perceived usefulness of the app by boosting workers' confidence in their handling practices. Likewise, the focus on a user-friendly interface and streamlined workflow integration reflected the importance of ease of use, ensuring that the app could be seamlessly adopted into daily routines.

By integrating the Patient-Oriented Research approach [24-26] and TAM constructs [28], our co-design process showed that engaging end users yields practical, user-centered solutions that are likely to be adopted. This echoes human-factors guidance that rigorous, iterative user-centered design enhances the effectiveness, efficiency, and satisfaction of mHealth technologies [51] and underscores the need for ongoing co-design [47], usability optimization, and feedback loops when scaling to other health regions.

Feasibility and Acceptability

The findings from this study suggest that the SPH MSIP APP is a feasible and acceptable intervention for community health care workers and contributes to the growing body of evidence supporting the utility of mHealth tools in health care settings. Consistent with previous studies [52,53], our feasibility results highlight the potential of digital interventions to address specific gaps in training and resource accessibility. Recruitment and retention rates exceeded expectations, and participants provided positive feedback regarding the app's ease of use, content relevance, and integration into daily workflows. The progression criteria established for this study—encompassing usability, satisfaction, and continued engagement—were consistently met, further underscoring the app's potential as a practical resource for community health care workers. These results echo findings in the broader literature that highlight the growing acceptance of mHealth interventions among health care providers [54,55], particularly when designed with user-centric principles [56].

The app's emphasis on real-time accessibility and user-friendly design was particularly appreciated by participants, many of whom noted its potential to enhance confidence in performing safe patient handling tasks. These features align closely with the TAM [28], particularly in terms of perceived ease of use and perceived usefulness. Participants found the app intuitive and valuable, suggesting that it effectively addressed some barriers to safe patient handling practices among community health care workers. However, the short duration of the study limits our understanding of whether these initial impressions would translate into sustained changes in behavior or reductions in musculoskeletal injuries. Previous research [57,58] has shown that the efficacy of digital interventions often diminishes over time without ongoing engagement strategies, suggesting a need for future iterations of the app to incorporate features such as in-app support from peers or coaches, personalized feedback, and periodic updates to maintain user interest [59]. These features would directly enhance the app's perceived usefulness and ensure that users remain engaged and motivated to incorporate safe handling practices into their routine [28,59]. Moreover, while high satisfaction rates are encouraging, they do not necessarily predict long-term adherence to best practices, highlighting the importance of further research to evaluate the app's impact on musculoskeletal injury outcomes in more complex, real-world settings. From a TAM perspective [28], while the app's ease of use and perceived usefulness were apparent, sustained use could depend on enhancing its relevance to users' evolving needs and providing mechanisms to reinforce behavior change [60,61]. This is where the app could benefit from additional evaluation to refine its features based on longer-term feedback, ensuring it remains useful and engaging beyond initial exposure [62].

While the app shows promise, its development and implementation also underscore the complexities involved in translating research into practice. For instance, the process of enabling the app on community health care workers' devices revealed structural barriers within Island Health, where restrictive IT policies limited the ease of deployment. This challenge aligns with the concept of perceived ease of use in TAM [28], which suggests that any technical barriers to usability can impede user acceptance. These barriers highlight the importance of considering organizational readiness and infrastructure when designing and scaling mHealth interventions [63]. This finding resonates with recommendations from the World Health Organization [64], which emphasize the need for alignment between technological innovations and institutional support systems to ensure successful implementation. Health authorities looking to implement similar interventions may need to account for potential technological and organizational challenges that can affect the perceived ease of use and, consequently, the overall success of the intervention [63,64].

Strengths, Limitations, and Future Directions

A key strength of this study is its rigorous, end-to-end co-design process: a multidisciplinary, patient-oriented research team of coaches, frontline workers, developers, and decision-makers guided every phase, ensuring the app matches the realities of community care settings. This collaborative approach translated into exceptional engagement and retention, and every prespecified usability and satisfaction criterion was surpassed. Moreover, the study delivers one of the first comprehensive evaluations of a safe-patient-handling m-health tool purpose-built for community health care workers, addressing a critical gap in the digital-safety literature.

This study also had several limitations that need to be acknowledged and addressed in future research. First, the small, self-selected sample may have introduced selection bias, as

XSL•FO
RenderX

those who volunteered may have been more technologically adept or motivated than the broader community health care workers population. To address this, future effectiveness studies should use more inclusive recruitment strategies, potentially incorporating random sampling methods to capture a more representative demographic, and include a control or comparison group (eg, randomized or quasi-experimental design) to enable stronger causal inference [65]. Second, the absence of blinded participatory meetings in phase 1 may have influenced the data collection process, as researchers' relationships with participants could have biased responses. Future studies should consider using independent facilitators for initial data collection phases to minimize potential biases and ensure the reliability of findings [66]. Furthermore, this feasibility trial did not assess the efficacy of the SPH MSIP app in reducing musculoskeletal injuries. A fully powered, longer-term randomized controlled trial is the logical next step to evaluate whether the app can lower musculoskeletal injury rates, lost-time claims, and related disability while enhancing patient-care outcomes. Future research should also focus on addressing the systemic barriers identified in this study, such as the lack of accessible, safe patient handling training and workplace cultural challenges, to ensure broader applicability and scalability of the intervention. By building on these findings and incorporating lessons learned, we can move closer to developing scalable, sustainable solutions that enhance safety and well-being for community health care workers in diverse settings. Additionally, this feasibility study used brief, adapted TAM items rather than a fully validated instrument tailored to community health care workers. This is not ideal for cross-study comparability, but it was a reasonable, low-burden choice for an early pilot focused on decision signals. Qualitative coding was conducted in Microsoft Excel. While acceptable for small feasibility datasets, Excel lacks features that enhance transparency and credibility. Future studies should consider using validated acceptance measures appropriate to this population, evaluate measurement properties in the target sample, and adopt dedicated qualitative software (eg, NVivo [Lumivero]) with a documented audit trail, multi-coder procedures with agreement checks, and reflexive memoing. Finally, the challenges associated with deploying the app on community health care workers' devices underscore the need for organizational support and streamlined IT processes. Collaborating with health authorities to address systemic barriers to technology adoption will be essential for scaling the intervention [67]. Future iterations of the app could also incorporate advanced features, such as personalized feedback, gamification elements, and real-time analytics, to enhance user engagement and effectiveness over time.

Conclusions

To conclude, this single-group feasibility study provides preliminary signals of feasibility and acceptability for the SPH MSIP APP to address unmet safe patient-handling needs among community health care workers. The app was well-received, with high levels of satisfaction and usability reported by participants, suggesting its potential as a practical tool for improving safe patient handling practices. While these findings are promising, they should be interpreted with caution, given the small, self-selected sample and the absence of a control or comparison group; TAM outcomes were used as low-burden, descriptive indicators rather than validated, mechanism-testing measures. Future effectiveness studies should recruit more broadly, use validated acceptance instruments and dedicated qualitative software, and use controlled or randomized designs with longer follow-up to evaluate effects on musculoskeletal injury rates, disability-related absences, and other implementation outcomes.

Acknowledgments

This study was funded by WorkSafeBC (grant RS2022-A131). The funder had no role in the study design; data collection, analysis, or interpretation; manuscript preparation; or the decision to submit for publication. The authors confirm that no generative artificial intelligence technology was used in the creation of the text, figures, tables, or other content of this manuscript. The authors have no financial relationships relevant to this article to disclose.

Data Availability

The datasets generated or analyzed during this study are available in the Open Science Framework (OSF) repository [68].

Authors' Contributions

SB, SC, SW, and RER contributed to the conceptualization, investigation, methodology, project administration, and provision of resources for the study. SB, SC, and SW were responsible for data curation, while ASYL conducted the formal analysis. Software development was carried out by SB, SC, SW, and ASYL. Validation was performed by ASYL, SB, SC, SW, and RER, and visualization was completed by ASYL and RER. Funding acquisition was undertaken by RER, who also supervised the project. ASYL prepared the original draft of the manuscript, and SB, SC, SW, and RER contributed to the review and editing. All authors read and approved the final version of the manuscript.

Conflicts of Interest

None declared.

References

XSL•FO
RenderX

1. Choi SD, Brings K. Work-related musculoskeletal risks associated with nurses and nursing assistants handling overweight and obese patients: a literature review. Work 2015;53(2):439-448. [doi: 10.3233/WOR-152222] [Medline: 26835850]
2. Jäger M, Jordan C, Theilmeier A, Wortmann N, Kuhn S, Nienhaus A, et al. Lumbar-load analysis of manual patient-handling activities for biomechanical overload prevention among healthcare workers. Ann Occup Hyg 2013;57(4):528-544. [doi: 10.1093/annhyg/mes088] [Medline: 23253360]
3. Menzel NN, Brooks SM, Bernard TE, Nelson A. The physical workload of nursing personnel: association with musculoskeletal discomfort. Int J Nurs Stud 2004;41(8):859-867. [doi: 10.1016/j.ijnurstu.2004.03.012] [Medline: 15476759]
4. Davis KG, Kotowski SE. Prevalence of musculoskeletal disorders for nurses in hospitals, long-term care facilities, and home health care: a comprehensive review. Hum Factors 2015;57(5):754-792. [doi: 10.1177/0018720815581933] [Medline: 25899249]
5. Nonfatal injuries and illnesses resulting in days off work among nurses up 291 percent in 2020. Washington, D.C: United States Bureau of Labor Statistics; 2020.
6. Franche R, Murray EJ, Ostry A, Ratner PA, Wagner SL, Harder HG. Work disability prevention in rural healthcare workers. Rural Remote Health 2010;10(4):1502 [FREE Full text] [Medline: 20964467]
7. Onazi O, Adejumo AO, Redwood L, Okorie O, Lawal O, Azuogu B, et al. Community health care workers in pursuit of TB: discourses and dilemmas. Soc Sci Med 2020;246:112756. [doi: 10.1016/j.socscimed.2019.112756] [Medline: 31954279]
8. O'Keeffe VJ, Tuckey MR, Naweed A. Whose safety? flexible risk assessment boundaries balance nurse safety with patient care. Saf Sci 2015;76:111-120. [doi: 10.1016/j.ssci.2015.02.024]
9. Wåhlin C, Stigmar K, Nilsing Strid E. A systematic review of work interventions to promote safe patient handling and movement in the healthcare sector. Int J Occup Saf Ergon 2022;28(4):2520-2532 [FREE Full text] [doi: 10.1080/10803548.2021.2007660] [Medline: 34789085]
10. Hegewald J, Berge W, Heinrich P, Staudte R, Freiberg A, Scharfe J, et al. Do technical aids for patient handling prevent musculoskeletal complaints in health care workers?-a systematic review of intervention studies. Int J Environ Res Public Health 2018;15(3):476 [FREE Full text] [doi: 10.3390/ijerph15030476] [Medline: 29522440]
11. Kurowski A, Gore R, Roberts Y, Kincaid KR, Punnett L. Injury rates before and after the implementation of a safe resident handling program in the long-term care sector. Saf Sci 2017;92:217-224 [FREE Full text] [doi: 10.1016/j.ssci.2016.10.012] [Medline: 34744312]
12. Lahiri S, Latif S, Punnett L, ProCare Research Team. An economic analysis of a safe resident handling program in nursing homes. Am J Ind Med 2013;56(4):469-478 [FREE Full text] [doi: 10.1002/ajim.22139] [Medline: 23203729]
13. Kurowski A, Pransky G, Punnett L. Impact of a safe resident handling program in nursing homes on return-to-work and re-injury outcomes following work injury. J Occup Rehabil 2019;29(2):286-294 [FREE Full text] [doi: 10.1007/s10926-018-9785-7] [Medline: 29785467]
14. Owlia M, Kamachi M, Dutta T. Reducing lumbar spine flexion using real-time biofeedback during patient handling tasks. Work 2020;66(1):41-51 [FREE Full text] [doi: 10.3233/WOR-203149] [Medline: 32417812]
15. Lieberman J, Breazeal C. TIKL: development of a wearable vibrotactile feedback suit for improved human motor learning. IEEE Trans Robot 2007;23(5):919-926. [doi: 10.1109/tro.2007.907481]
16. Källander K, Tibenderana JK, Akpogheneta OJ, Strachan DL, Hill Z, ten Asbroek AHA, et al. Mobile health (mHealth) approaches and lessons for increased performance and retention of community health workers in low- and middle-income countries: a review. J Med Internet Res 2013;15(1):e17 [FREE Full text] [doi: 10.2196/jmir.2130] [Medline: 23353680]
17. Zaidi S, Kazi AM, Riaz A, Ali A, Najmi R, Jabeen R, et al. Operability, usefulness, and task-technology fit of an mHealth app for delivering primary health care services by community health workers in underserved areas of Pakistan and Afghanistan: qualitative study. J Med Internet Res 2020;22(9):e18414 [FREE Full text] [doi: 10.2196/18414] [Medline: 32940612]
18. VIVES University of Applied Sciences, eUlift. 2019. URL: https://www.vives.be/en/vives-international [accessed 2025-10-10]
19. Public health. US Department of Veterans Affairs. 2019. URL: https://www.publichealth.va.gov/employeehealth/patient-handling/ [accessed 2025-10-10]
20. PainPoint – prevent musculoskeletal disorders (MSD) at work. Occupational Health Clinics for Ontario Workers. 2021. URL: https://www.ohcow.on.ca/resources/apps-tools-calculators/painpoint/ [accessed 2024-12-30]
21. Health, wellness and urgent support services. Island Health. 2024. URL: https://medicalstaff.islandhealth.ca/health-wellness-and-urgent-support-services [accessed 2025-10-10]
22. Johnson RB, Onwuegbuzie AJ. Mixed methods research: a research paradigm whose time has come. Educ Res 2004;33(7):14-26. [doi: 10.3102/0013189x033007014]
23. Hothersall SJ. Epistemology and social work: enhancing the integration of theory, practice and research through philosophical pragmatism. Eur J Soc Work 2019;22(5):860-870. [doi: 10.1080/13691457.2018.1499613]
24. Allemang B, Sitter K, Dimitropoulos G. Pragmatism as a paradigm for patient-oriented research. Health Expect 2022;25(1):38-47 [FREE Full text] [doi: 10.1111/hex.13384] [Medline: 34748689]
25. Sunderji N, Nicholas Angl E, Polaha J, Gao C. Why and how to use patient-oriented research to promote translational research. Fam Syst Health 2019;37(1):1-9. [doi: 10.1037/fsh0000405] [Medline: 30920259]

XSL•FO
RenderX

26. Sacristán JA. Patient-centered medicine and patient-oriented research: improving health outcomes for individual patients. BMC Med Inform Decis Mak 2013;13:6 [FREE Full text] [doi: 10.1186/1472-6947-13-6] [Medline: 23294526]
27. Mummah SA, Robinson TN, King AC, Gardner CD, Sutton S. IDEAS (Integrate, Design, Assess, and Share): a framework and toolkit of strategies for the development of more effective digital interventions to change health behavior. J Med Internet Res 2016;18(12):e317 [FREE Full text] [doi: 10.2196/jmir.5927] [Medline: 27986647]
28. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. Manag Inf Syst Q 1989;13(3):319-340. [doi: 10.2307/249008]
29. Tahar A, Riyadh HA, Sofyani H, Purnomo WE. Perceived ease of use, perceived usefulness, perceived security and intention to use E-filing: the role of technology readiness. J Asian Finance Econ Bus 2020;7(9):537-547. [doi: 10.13106/jafeb.2020.vol7.no9.537]
30. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view. Manag Inf Syst Q 2003;27(3):425-478. [doi: 10.2307/30036540]
31. Siebert JN, Gosetto L, Sauvage M, Bloudeau L, Suppan L, Rodieux F, PedAMINES Trial Group, PedAMINES Prehospital Group. Usability testing and technology acceptance of an mHealth app at the point of care during simulated pediatric in- and out-of-hospital cardiopulmonary resuscitations: study nested within 2 multicenter randomized controlled trials. JMIR Hum Factors 2022;9(1):e35399 [FREE Full text] [doi: 10.2196/35399] [Medline: 35230243]
32. AlQudah AA, Al-Emran M, Shaalan K. Technology acceptance in healthcare: a systematic review. Appl Sci 2021;11(22):10537. [doi: 10.3390/app112210537]
33. Marangunić N, Granić A. Technology acceptance model: a literature review from 1986 to 2013. Univ Access Inf Soc 2014;14(1):81-95. [doi: 10.1007/s10209-014-0348-1]
34. Doyle L, Brady AM, Byrne G. An overview of mixed methods research. J Res Nurs 2009;14(2):175-185. [doi: 10.1177/1744987108093962]
35. Liu S, La H, Willms A, Rhodes RE. A "No-Code" app design platform for mobile health research: development and usability study. JMIR Form Res 2022;6(8):e38737 [FREE Full text] [doi: 10.2196/38737] [Medline: 35980740]
36. Pathverse. Linkedin. 2021. URL: https://ca.linkedin.com/company/pathverse-inc [accessed 2025-10-10]
37. Askari M, Klaver NS, van Gestel TJ, van de Klundert J. Intention to use medical apps among older adults in the Netherlands: cross-sectional study. J Med Internet Res 2020;22(9):e18080 [FREE Full text] [doi: 10.2196/18080] [Medline: 32624465]
38. Lewis M, Bromley K, Sutton CJ, McCray G, Myers HL, Lancaster GA. Determining sample size for progression criteria for pragmatic pilot RCTs: the hypothesis test strikes back!. Pilot Feasibility Stud 2021;7(1):40 [FREE Full text] [doi: 10.1186/s40814-021-00770-x] [Medline: 33536076]
39. Lincoln YS, Guba EG. Naturalistic inquiry. 1st ed. California (USA), London: Sage Publications; 1985.
40. Braun V, Clarke V. Reflecting on reflexive thematic analysis. Qual Res Sport Exerc Health 2019;11(4):589-597. [doi: 10.1080/2159676x.2019.1628806]
41. Lee C, Knight SW, Smith SL, Nagle DJ, DeVries L. Safe patient handling and mobility: development and implementation of a large-scale education program. Crit Care Nurs Q 2018;41(3):253-263. [doi: 10.1097/CNQ.0000000000000204] [Medline: 29851674]
42. Nelson AL. Safe patient handling and movement: a practical guide for health care professionals. New Jersey, US: Springer Publishing Company; 2005.
43. Nelson AL, Collins J, Knibbe H, Cookson K, de Castro AB, Whipple KL. Safer patient handling. Nurs Manage 2007;38(3):26-32; quiz 32. [doi: 10.1097/01.NUMA.0000262923.39562.0e] [Medline: 17473792]
44. Andersen LL, Vinstrup J, Villadsen E, Jay K, Jakobsen MD. Physical and psychosocial work environmental risk factors for back injury among healthcare workers: prospective cohort study. Int J Environ Res Public Health 2019;16(22):4528 [FREE Full text] [doi: 10.3390/ijerph16224528] [Medline: 31731806]
45. Lee S, Faucett J, Gillen M, Krause N, Landry L. Factors associated with safe patient handling behaviors among critical care nurses. Am J Ind Med 2010;53(9):886-897. [doi: 10.1002/ajim.20843] [Medline: 20698021]
46. Stevens L, Rees S, Lamb KV, Dalsing D. Creating a culture of safety for safe patient handling. Orthop Nurs 2013;32(3):155-64; quiz 165. [doi: 10.1097/NOR.0b013e318291dbc5] [Medline: 23695761]
47. Song T, Yu P, Bliokas V, Probst Y, Peoples GE, Qian S, et al. A clinician-led, experience-based co-design approach for developing mHealth services to support the patient self-management of chronic conditions: development study and design case. JMIR Mhealth Uhealth 2021;9(7):e20650 [FREE Full text] [doi: 10.2196/20650] [Medline: 34283030]
48. Bosch J, Ludwig C, Fluch-Niebuhr J, Stoevesandt D. Empowerment for the digital transformation: results of a structured blended-learning on-the-job training for practicing physicians in Germany. Int J Environ Res Public Health 2022;19(20):12991 [FREE Full text] [doi: 10.3390/ijerph192012991] [Medline: 36293572]
49. Slovensky DJ, Malvey DM, Neigel AR. A model for mHealth skills training for clinicians: meeting the future now. Mhealth 2017;3:24 [FREE Full text] [doi: 10.21037/mhealth.2017.05.03] [Medline: 28736733]
50. Dockrell S, Hurley G. Moving and handling care of bariatric patients: a survey of clinical nurse managers. J Res Nurs 2021;26(3):194-204 [FREE Full text] [doi: 10.1177/1744987120970623] [Medline: 35251242]

XSL•FO
RenderX

51. Or CK, Holden RJ, Valdez RS. Human factors engineering and user-centered design for mobile health technology: enhancing effectiveness, efficiency, and satisfaction. In: Duffy VG, Ziefle M, Rau PLP, Tseng MM, editors. Human-Automation Interaction. Automation, Collaboration, & E-Services, vol 12. Switzerland: Springer, Cham; 2023.
52. Echelard J, Méthot F, Nguyen H, Pomey M. Medical student training in eHealth: scoping review. JMIR Med Educ 2020;6(2):e20027 [FREE Full text] [doi: 10.2196/20027] [Medline: 32915154]
53. Senbekov M, Saliev T, Bukeyeva Z, Almabayeva A, Zhanaliyeva M, Aitenova N, et al. The recent progress and applications of digital technologies in healthcare: a review. Int J Telemed Appl 2020;2020:8830200 [FREE Full text] [doi: 10.1155/2020/8830200] [Medline: 33343657]
54. Gagnon MP, Ngangue P, Payne-Gagnon J, Desmartis M. m-Health adoption by healthcare professionals: a systematic review. J Am Med Inform Assoc 2016;23(1):212-220 [FREE Full text] [doi: 10.1093/jamia/ocv052] [Medline: 26078410]
55. Early J, Gonzalez C, Gordon-Dseagu V, Robles-Calderon L. Use of mobile health (mHealth) technologies and interventions among community health workers globally: a scoping review. Health Promot Pract 2019;20(6):805-817. [doi: 10.1177/1524839919855391] [Medline: 31179777]
56. Bilal HSM, Amin MB, Hussain J, Ali SI, Razzaq MA, Hussain M, et al. Towards user-centric intervention adaptiveness: influencing behavior-context based healthy lifestyle interventions. IEEE Access 2020;8:177156-177179. [doi: 10.1109/access.2020.3026688]
57. Hershman SG, Bot BM, Shcherbina A, Doerr M, Moayedi Y, Pavlovic A, et al. Physical activity, sleep and cardiovascular health data for 50,000 individuals from the MyHeart counts study. Sci Data 2019;6(1):24 [FREE Full text] [doi: 10.1038/s41597-019-0016-7] [Medline: 30975992]
58. Pratap A, Neto EC, Snyder P, Stepnowsky C, Elhadad N, Grant D, et al. Indicators of retention in remote digital health studies: a cross-study evaluation of 100,000 participants. NPJ Digit Med 2020;3:21 [FREE Full text] [doi: 10.1038/s41746-020-0224-8] [Medline: 32128451]
59. Amagai S, Pila S, Kaat AJ, Nowinski CJ, Gershon RC. Challenges in participant engagement and retention using mobile health apps: literature review. J Med Internet Res 2022;24(4):e35120 [FREE Full text] [doi: 10.2196/35120] [Medline: 35471414]
60. Kim SD. Application and challenges of the technology acceptance model in elderly healthcare: insights from ChatGPT. Technologies 2024;12(5):68. [doi: 10.3390/technologies12050068]
61. Cho H, Chi C, Chiu W. Understanding sustained usage of health and fitness apps: incorporating the technology acceptance model with the investment model. Tech Soc 2020;63:101429. [doi: 10.1016/j.techsoc.2020.101429]
62. Bhattacharjya S, Cavuoto LA, Reilly B, Xu W, Subryan H, Langan J. Usability, usefulness, and acceptance of a novel, portable rehabilitation system (mRehab) using smartphone and 3d printing technology: mixed methods study. JMIR Hum Factors 2021;8(1):e21312 [FREE Full text] [doi: 10.2196/21312] [Medline: 33749608]
63. Bradway M, Carrion C, Vallespin B, Saadatfard O, Puigdomènech E, Espallargues M, et al. mHealth assessment: conceptualization of a global framework. JMIR Mhealth Uhealth 2017;5(5):e60 [FREE Full text] [doi: 10.2196/mhealth.7291] [Medline: 28465282]
64. From innovation to implementation: eHealth in the WHO European Region. Europe: World Health Organization; 2016.
65. Noor S, Tajik O, Golzar J. Simple random sampling. Int J Lang Educ 2022;1(2):78-82. [doi: 10.22034/ijels.2022.162982]
66. Page S, Persch AC. Recruitment, retention, and blinding in clinical trials. Am J Occup Ther 2013;67(2):154-161 [FREE Full text] [doi: 10.5014/ajot.2013.006197] [Medline: 23433269]
67. Zakerabasali S, Ayyoubzadeh SM, Baniasadi T, Yazdani A, Abhari S. Mobile health technology and healthcare providers: systemic barriers to adoption. Healthc Inform Res 2021;27(4):267-278 [FREE Full text] [doi: 10.4258/hir.2021.27.4.267] [Medline: 34788907]
68. Safe patient handling musculoskeletal injury-prevention smartphone application for community healthcare workers: mixed-methods feasibility study. Open Science Framework (Center for Open Science). URL: https://osf.io/x7ves [accessed 2025-10-15]

Abbreviations

CHS: Community Health Services
mHealth: mobile health
SPH MSIP: Safe Patient Handling Musculoskeletal Injury-Prevention
TAM: Technology Acceptance Model

XSL•FO
RenderX

Edited by J Sarvestan; submitted 31.12.24; peer-reviewed by C Or; comments to author 30.06.25; revised version received 29.07.25; accepted 19.09.25; published 31.10.25.

Please cite as:
Lee ASY, Burrell S, Courtnall S, Wake S, Rhodes RE
Safe Patient Handling Musculoskeletal Injury-Prevention Smartphone App for Community Health Care Workers: Mixed Methods Feasibility Study
JMIR Form Res 2025;9:e70681
URL: https://formative.jmir.org/2025/1/e70681
doi:10.2196/70681
PMID:

©Alfred S Y Lee, Sarah Burrell, Sandy Courtnall, Shelly Wake, Ryan E Rhodes. Originally published in JMIR Formative Research (https://formative.jmir.org), 31.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Quality Assessment of Large Language Model–Generated Medical Dialogue for Clinical Vignettes: Evaluation Study

Yasutaka Yanagita[1], MD, PhD; Daiki Yokokawa[1], MD, PhD; Shiichi Ihara[1], MD; Ryo Yoshida[1], MD; Yoshihide Okano[1], MD; Takanori Uehara[1], MD, PhD

Department of General Medicine, Chiba University Hospital, Chiba, Japan, Chiba, Japan

Corresponding Author:
Yasutaka Yanagita, MD, PhD
Department of General Medicine
Chiba University Hospital, Chiba, Japan
1-8-1, Inohana, Chuo-ku
Chiba, 260-8677
Japan
Phone: 81 43 222 7171 ext 6438
Fax: 81 43 224 4758
Email: y.yanagita@gmail.com

Abstract

Background: Traditional clinical vignettes, though widely used in medical education, often focus on prototypical presentations; require substantial time and effort to develop; and fail to represent patient diversity, the complexity of clinical conditions, patients' perspectives, and the dynamic nature of physician-patient interactions.

Objective: This study aimed to evaluate the quality of Japanese-language physician-patient dialogues produced by generative artificial intelligence (AI), focusing on their medical accuracy and overall appropriateness as medical interviews.

Methods: We created an AI prompt that included a specific clinical history and instructed the model to simulate a cooperative patient responding to the physician's questions to generate a physician-patient dialogue. The target diseases were those covered by the Japanese National Medical Licensing Examination. Each dialogue consisted of 25 turns by the physician and 25 by the patient, reflecting the typical volume of conversation in Japanese outpatient settings. Three internists independently evaluated each generated dialogue using a 7-point Likert scale across 6 criteria: coherence of the conversation, medical accuracy of the patient's responses, medical accuracy of the physician's responses, content of the medical history, communication skills, and professionalism. In addition, a composite score for each dialogue was calculated as the overall mean of these 6 criteria. Each dialogue was also examined for the presence of 5 essential clinical components commonly included in medical interviews: chief concern and clinical course since onset, physical findings, test results, diagnosis, and treatment course. A dialogue was considered to include a component only if all 3 evaluators independently confirmed its presence.

Results: The mean composite score was 5.7 (SD 1.0), indicating high overall quality. Mean scores for each criterion were as follows: coherence of the conversation, 5.9 (SD 0.9); medical accuracy of the patient's responses, 6.0 (SD 0.9); medical accuracy of the physician's responses, 5.6 (SD 1.1); content of medical history taking, 5.9 (SD 0.9); communication skills, 5.6 (SD 0.9); and professionalism, 5.5 (SD 1.1). Among the 5 clinical components assessed in each dialogue across 47 clinical cases, chief concern and clinical course were included in all 47 (100%) cases, physical findings in 15 (32%) cases, test results in 27 (57%) cases, diagnosis in 45 (96%) cases, and treatment course in 0 (0%) cases.

Conclusions: While physician oversight remains essential, it is feasible to efficiently create AI-generated educational materials for medical education that overcome the limitations of traditional clinical vignettes. This approach may reduce time and financial burdens, enhancing opportunities to practice clinical interviewing in settings that closely mirror real-world encounters.

(JMIR Form Res 2025;9:e80752) doi:10.2196/80752

KEYWORDS

ChatGPT; clinical vignettes; artificial intelligence; generative AI; medical education; physician-patient dialogue

XSL•FO
RenderX

Introduction

Artificial intelligence (AI) is critical for the advancement of medicine. In particular, generative AI, exemplified by large language models (LLMs), has the potential to fundamentally transform health care. Tools such as ChatGPT, developed by OpenAI, possess advanced conversational capabilities and broad applicability [1], with an increasing number of applications in the medical field. From the perspective of conversational capabilities, generative AI–based chatbots can provide rehabilitation guidance and mental health support to patients and assist health care professionals in patient management [2]. Generative AI is versatile enough to have passed the Japanese National Medical Licensing Examination [3]. Moreover, LLMS are now used to enter patients' medical histories and verify diagnostic accuracy in English and Japanese [4,5], indicating that the scope of generative AI applications is expected to expand further in the future.

The application of generative AI in medical education has attracted increasing interest. The use of generative AI to produce physician-patient dialogues has the potential to replicate disease-specific communication encountered in clinical practice and serve as a valuable educational resource for medical students.

A comparable and widely used educational tool in medical education research is the *clinical vignette* [6-8]. Vignettes are paper-based instructional resources that present a patient's medical history and key clinical information in an organized written format. They are considered highly effective in helping medical students understand disease concepts, symptomatology, and treatment pathways [9].

Exposure to various vignettes can deepen students' clinical understanding. However, traditional vignettes often lack diversity in case scenarios and tend to focus narrowly on diagnostic information, thereby failing to capture the complexity, ambiguity, and uncertainty that characterize real-world clinical practice [10]. In addition, the development of vignettes requires substantial time and effort. As clinical vignettes are generally written from the health care provider's perspective, they inherently fail to capture the patient's viewpoint and the dynamics of physician-patient interaction. Using generative AI such as ChatGPT to create dialogue-based educational materials that incorporate these missing elements is considered a promising approach. These educational materials offer several advantages. First, ChatGPT can generate a wide variety of dialogues tailored to different diseases and clinical scenarios, enabling repeated practice. Its use can reduce both the time and financial costs associated with producing educational content. Moreover, medical students can engage in clinical reasoning by organizing information while considering multiple differential diagnoses and can practice effective physician-patient communication, as would be required in actual clinical settings.

AI-generated dialogues incorporate the medical information necessary for diagnostic decision-making along with a variety of interactions that closely resemble those encountered in real clinical settings. This enables medical students to develop information-processing skills in authentic, context-rich scenarios. Through this process, students are expected to acquire the ability to identify clinically relevant information and develop a broader range of clinical competencies, including medical interviewing techniques, patient management skills, and effective communication strategies.

This study examined the feasibility of generating physician-patient dialogues using generative AI and evaluated the quality of the generated dialogues. Generative AI has already demonstrated the ability to generate differential diagnosis lists, illness scripts, and clinical vignettes based on medical information [11,12]. Accordingly, it is considered feasible to generate physician-patient dialogues with a high level of accuracy using generative AI.

In this study, we assessed the quality of dialogues generated in Japanese medical interviews, including their medical accuracy, and investigated their potential utility as educational materials in the context of medical education. By demonstrating the feasibility of easily generating diverse physician-patient dialogues using generative AI, it becomes possible to efficiently produce a large volume of educational content for use in medical education under the supervision of physicians with medical expertise. The appropriate application of AI technology is expected to increase the opportunities for medical students to engage with more practical and interactive content, thereby enhancing the effectiveness of medical education.

Methods

Overview

Physician-patient dialogues were generated using an LLM and were subjected to cross-sectional evaluation. To ensure relevance to medical education, 47 target diseases were selected to represent diseases that medical students were expected to learn. The selection was based on the content of the Japanese National Medical Licensing Examination [13] and finalized through discussion between a board-certified internist (YY) and a board-certified family physician (DY; Textbox 1). To evaluate the generated dialogues, we recruited 3 Japanese internists (SI, RY, and YO) who are actively involved in diagnostic practice at a university hospital and regularly manage a wide range of diagnostically challenging cases. These internists assessed the quality of the dialogues.

XSL•FO
RenderX

Textbox 1. List of 47 diseases used to generate and evaluate artificial intelligence–generated physician-patient dialogues.

Diseases

1. Gastroesophageal reflux disease
2. Functional dyspepsia
3. Esophageal achalasia
4. Crohn disease
5. Appendicitis
6. Pheochromocytoma
7. Primary aldosteronism
8. Cushing syndrome
9. Hashimoto disease
10. Subacute thyroiditis
11. Graves disease
12. Depression
13. Panic disorder
14. Gout
15. Pneumothorax
16. Myasthenia gravis
17. Trigeminal neuralgia
18. Parkinson disease
19. Alzheimer disease
20. Benign paroxysmal positional vertigo
21. Migraine
22. Cluster headache
23. Neurocardiogenic syncope
24. Intervertebral disk herniation
25. Insomnia
26. Bronchial asthma
27. Acute eosinophilic pneumonia
28. Rheumatoid arthritis
29. Systemic lupus erythematosus
30. Sjögren syndrome
31. Measles
32. Fibromyalgia
33. Gallstone attack
34. Acute cholecystitis
35. Unstable angina
36. COVID-19
37. Lumbar spinal stenosis
38. Thromboangiitis obliterans
39. Infectious mononucleosis
40. Streptococcal pharyngitis
41. Hepatitis
42. Transient ischemic attack

XSL•FO
RenderX

43. Iron deficiency anemia
44. Heatstroke
45. Acute pericarditis
46. Chronic obstructive pulmonary disease
47. Lung cancer

Production Environment

The analysis was performed on a terminal running Ubuntu (version 20.04.2 Long-Term Support; Canonical). The central processing unit was equipped with an AMD EPYC 7402*P* 24-core processor (Advanced Micro Devices, Inc), with 256 GB of main memory. An NVIDIA A100 graphics processing unit (NVIDIA Corp) with 40 GB of RAM was used for the computations. Python (version 3.6.9) and the OpenAI Python library (version 0.27.8; OpenAI) were used. The application programming interface used to generate the dialogues was the latest one available at the start of the study, gpt-4o-2024-11-20 [14], and the dialogues were generated on November 22, 2024.

Prompts to Be Entered Into the LLM

Referring to previous studies that used AI to generate vignettes and illness scripts [12], concise instructions specifying the desired output conditions were provided to the LLM as prompts (Figure 1). To enhance the completeness and coherence of the generated dialogues, the prompts included conditions such as the user supplying a specific vignette before the generation, refraining from using medical terminology, and responding cooperatively to the physician's questions. The dialogue length was determined with reference to the average volume of conversation in typical outpatient consultations (approximately 5-10 min) [15], and the LLM was instructed to generate 25 turns each from the physician and the patient.

XSL•FO
RenderX

Figure 1. A portion of the prompt used to generate physician-patient dialogues in Japanese using generative artificial intelligence (Top: Japanese, Bottom: English).

dialogue_1 = 状況：医療面接の文章化。ユーザが指示した症例（Vignette）を元に、医師と患者の会話を ChatGPT が生成する。
ユーザの役割：Vignette を提示する。
ChatGPT の役割：会話文の生成を行う。
ChatGPT の役割の詳細：以下の条件に則って生成しなさい。
目的：この会話を客観的に解説し、問診の仕方から診断に至るプロセスを学ぶ。
条件：
1. Vignette に沿った内容で会話を生成すること。
2. 会話はなるべく短い一文にすること
3. 医師の会話は医師：で書き出すこと。「」や…はつけない。
4. 患者の会話は患者：で書き出すこと。「」や"はつけない。
5. 患者の客観的な症状を述べさせるために家族を登場させても良い。特に患者が話す内容に信頼性がないときは、患者家族の発言を医師が聞き出すようにしなさい。患者家族：で書き出すこと。
6. すべての会話は 1 から順番に番号をふること。少なくとも 30 番までは患者が経験した症状やその特徴についての内容にしなさい。
7. 最低でも 50 番まで会話を行いなさい。
8. 患者は医師の質問に Vignette に沿った内容で患者として振る舞い、回答する。
9. 患者は医師に協力的に振る舞う。
10. 患者は、医師に病名を伝えられるまで、自分の診断名は知らない。よって、医師に診断名を伝えることはできない。
11. 患者は、医師に診察所見や検査結果を伝えることはできない。
12. 日本語で生成する。
13. 医師は同じ症状を呈する他の鑑別診断に関する質問も行う。それは結果的に陰性かもしれない。
14. 医師も診断を知らないため、質問内容が診断名と関係ないこともある。
15. 医師は患者の心理・社会に関わることも、必要があれば積極的に質問する。
16. すこし会話に雑談を加えなさい。症状を広げて聞き、関係ない話題が出ても許容しなさい。必ずしも Vignette に含まれる情報だけでなくても良い。
17. Vignette のうち、病歴、症状、診察、検査、診断を用いて生成しなさい。
18. 病歴と症状の文章量を最も増やしなさい。
19. 診断を伝えたら終了しなさい。
20. 方針の説明・経過は不要であり生成してはいけない。
21. 会話以外の文章を出力しない。
22. 治療薬は一般名を日本語で記載する。
23. 治療薬は日本で承認されているもののみ。
24. 患者は医学用語を使わない。一般的な用語を使用する。
25. 内容は教育的なものである必要がある。

dialogue_1 = Situation: Medical Interview Text. The user presents a specified case (Vignette), and ChatGPT generates a conversation between a physician and a patient based on it.
User role: Present the vignette.
ChatGPT role: Generate the dialogue.
ChatGPT should strictly follow the conditions below and not output anything else.
Objective: To learn the reasoning process—from questioning to diagnosis—by observing the physician's conversational approach and flow.
Conditions
1. Generate a conversation that follows the content of the vignette.
2. Keep each dialogue line as short and concise as possible.
3. Prefix physician's utterances with "Doctor:" and do not add quotation marks.
4. Prefix patient's utterances with "Patient:" and do not add quotation marks.
5. It is acceptable to include a family member in the dialogue to describe the patient's objective symptoms. When the patient's statements are unreliable, the physician should elicit information from the family member. Prefix such utterances with "Family:".
6. Start all conversations from "line 1" and continue sequentially. For at least the first 30 turns, the content should focus on the patient's experienced symptoms and their characteristics.
7. Continue the dialogue for at least 50 turns.
8. The physician's questions and the patient's answers must be consistent with the vignette.
9. The patient behaves cooperatively toward the physician.
10. The patient does not know their diagnosis until the physician informs them. Therefore, the patient cannot tell the physician the name of the disease.
11. The physician cannot convey the diagnosis to the patient.
12. The dialogue must be generated in Japanese.
13. The physician should also ask questions related to other possible differential diagnoses presenting with similar symptoms, even if those findings ultimately turn out to be negative.
14. Because the physician does not know the diagnosis, some of their questions may be unrelated to the actual diagnosis.
15. The physician will also ask questions related to the patient's psychological and social aspects when necessary.
16. Add a small amount of casual conversation. Broaden the discussion of symptoms and allow unrelated topics to appear naturally. The content does not have to be limited strictly to what is included in the vignette.
17. Use the vignette's information on medical history, symptoms, physical findings, tests, and diagnosis to generate the dialogue.
18. Give the greatest amount of detail and volume to the patient's history and symptoms.
19. End the dialogue once the diagnosis has been communicated.
20. Explanations of treatment plans or follow-up courses are unnecessary and must not be generated.
21. Do not output any text other than the dialogue itself.
22. Drug names should be written in Japanese using their International Nonproprietary Names.
23. Only use drugs that are approved for use in Japan.
24. The patient should not use medical terminology.
25. The content must have an educational purpose.

Evaluation of Dialogue

The presence of five key components commonly included in medical interviews was assessed: (1) chief concern and clinical course since onset, (2) physical findings, (3) test results, (4) diagnosis, and (5) treatment course. Three physicians independently reviewed each dialogue, and a dialogue was considered to contain all 5 components only if all 3 physicians confirmed their inclusion.

Next, criteria for assessing the quality of the medical interviews were developed. Six evaluation criteria were selected through discussions between a board-certified internist (YY) and a board-certified family physician (DY), with reference to the evaluation domains of the Mini-Clinical Evaluation Exercise (Mini-CEX) [16]: (1) coherence of the conversation, (2) medical accuracy of the patient's statements, (3) medical accuracy of the physician's statements, (4) quality of the physician's history taking, (5) communication skills, and (6) professionalism.

Criterion 1, coherence of the conversation, was evaluated as a linguistic criterion, focusing on the smoothness of the interaction between the physician and patient, the presence of inconsistencies, grammatical or typographical errors, and the logical relationship of the dialogue. Criteria 2 and 3, medical accuracy of the patient's and physician's statements, respectively, were assessed as clinical criteria by evaluating their consistency with the known clinical features of the respective diseases. Criterion 4, history taking, was assessed based on whether the physician elicited information regarding symptom characteristics and exacerbating or relieving factors. Criteria 5 and 6, communication skills and professionalism, respectively, were evaluated by assessing whether the physician

XSL•FO
RenderX

explored the patient's explanatory model and demonstrated respect, compassion, and empathy toward the patient (Table 1).

Following prior studies [12], we specifically employed 3 Japanese physicians affiliated with a university hospital, each of whom was involved in the supervision and education of medical students and residents, to evaluate the generated dialogues. Each of the 6 evaluation criteria was rated on a 7-point Likert scale, based on the perceived educational usefulness of the medical students. The scale was defined as follows: 1=not applicable at all or not useful—major overall revision required, 2=low usefulness—multiple major revisions needed, 3=limited usefulness—some valuable content but substantial revisions necessary, 4=moderate usefulness—both strengths and several areas for improvement, 5=generally useful—some revisions or adjustments desirable, 6=high usefulness—only minor adjustments possibly needed, and 7=extremely useful and complete—no further revisions required.

For each dialogue, the score of each evaluation criterion was calculated as the average of the ratings of the 3 evaluators. The composite score for the dialogue was then derived by averaging the scores across all 6 evaluation criteria and interpreted using the same 7-point Likert scale.

Table 1. Evaluation criteria and definitions used to assess artificial intelligence–generated physician-patient dialogues in medical interviews.

Evaluation criteria	Evaluation details
Coherence of the conversation	Assessed whether the dialogue between the physician and patient proceeded smoothly, with accurate grammar, no typographical or spelling errors, and overall linguistic clarity
Medical accuracy of the patient's statements	Evaluated whether the patient's utterances accurately reflected the typical onset patterns, symptoms, and clinical course associated with the relevant disease
Medical accuracy of the physician's statements	Assessed whether the physician's explanations and other statements were medically accurate and aligned with established clinical knowledge
Quality of the physician's history taking	Evaluated whether the physician asked about essential elements of the current illness, including symptom location, characteristics, severity, temporal course, contextual factors, aggravating and relieving factors, associated symptoms, and the patient's response to the symptoms
Communication skills	Assessed whether the physician conducted the interview in a way that facilitated open communication, explored the patient's explanatory model and psychosocial context, and confirmed the patient's understanding of the information discussed
Professionalism	Evaluated whether the physician demonstrated respect, compassion, and empathy toward the patient and whether a trusting therapeutic relationship was established

Ethical Considerations

This study did not involve human or animal participants, and therefore, ethics approval was not required.

Results

Using the gpt-4o-2024-11-20 model, physician-patient dialogues were generated for 47 clinical cases (Table 2). Among the 47 generated dialogues, clinical component 1, chief concern and clinical course since onset, was present in all 47 (100%) cases; clinical component 4, diagnosis, was included in 45 (96%) cases, and in each of these cases, the model accurately outputted the specified disease name, as instructed. In contrast, clinical component 2, physical findings, was included in 15 (32%) cases; clinical component 3, test results, was included in 27 (58%) cases; and clinical component 5, treatment course, was not included in any of the cases (0%). Regarding the quality of the medical interviews, the average score was 5.9 (SD 0.9) for coherence of the conversation, 6.0 (SD 0.9) for medical accuracy of the patient's statements, and 5.6 (SD 1.1) for medical accuracy of the physician's statements. The average score was 5.9 (SD 0.9) for quality of the physician's history taking, 5.6 (SD 0.9) for communication skills, and 5.5 (SD 1.1) for professionalism. The overall composite score, calculated as the mean of the 6 evaluation criteria, was 5.7 (SD 1.0).

A focused discussion was conducted among 5 physicians, 2 specialists (YY and DY), and 3 evaluators (SI, RY, and YO), centered on dialogues that were subject to point deductions to identify and clarify the specific issues present in the lower-rated dialogues. The results of this analysis are summarized in Table 3. For reference, one dialogue with a perfect average score of 7 and another that received a lower average score of 4 are provided in Multimedia Appendix 1 and Multimedia Appendix 2, respectively.

Table 2. Average scores for each of the 6 evaluation criteria used to assess artificial intelligence–generated physician-patient dialogues.

Evaluation criteria	Average score (SD)
Coherence of the conversation	5.9 (0.9)
Medical accuracy of the patient's statements	6.0 (0.9)
Medical accuracy of the physician's statements	5.6 (1.1)
Quality of the physician's history taking	5.9 (0.9)
Communication skills	5.6 (0.9)
Professionalism	5.5 (1.1)

XSL•FO
RenderX

Table 3. Problems identified for each evaluation criterion based on expert review of lower-rated artificial intelligence–generated physician-patient dialogues.

Evaluation criteria	Problems
Coherence of the conversation	• Responses to patient questions were omitted. • Although some expressions were unnatural, overall coherence was maintained.
Medical accuracy of the patient's statements	• Typical responses regarding aggravating and relieving factors were not provided.
Medical accuracy of the physician's statements	• A diagnosis was rendered, despite the interview being insufficient. • Diagnoses were finalized based on test results that were never mentioned as having been performed, indicating inappropriate or unjustified diagnostic reasoning. • The response is that it cures a disease that cannot be cured.
Quality of the physician's history taking	• Redundant questions were asked regarding information that had already been provided. • When multiple symptoms were present, it was unclear which symptom's aggravating or relieving factors were being discussed. • The structure and content of the medical interview were inadequate. • Additional questioning was conducted about symptoms not initially reported by the patient. • Unnecessary blood tests were included and described without justification.
Communication skills	• No attempts were made to confirm the patient's understanding of the information provided. • The patient's explanatory model was rarely elicited. • The patient's response lacked logical progression or flow.
Professionalism	• No appropriate responses were provided following expressions of anxiety. • The overall focus of the dialogue was limited to diagnostic questioning, with few utterances directed toward building rapport or fostering a therapeutic relationship.

Discussion

Principal Findings

This study aimed to generate physician-patient dialogue using generative AI and evaluate the prompt designs, the resulting outputs, and the overall quality of the dialogues. We created 47 dialogues based on diseases from the Japanese National Medical Licensing Examination and evaluated them using 6 criteria related to clinical communication. The dialogues consistently included the chief concern and clinical course and demonstrated high medical accuracy and coherence, while areas such as professionalism and inclusion of treatment information were less consistently addressed. The overall composite score was 5.7 (SD 1.0), indicating general usefulness with minor revisions required.

The analysis first focused on the outputs generated for 47 clinical cases and the corresponding prompt designs that produced them. Examination of the medical content essential for a clinical interview revealed that all dialogues included descriptions of the chief concern and clinical course since onset and that accurate diagnostic labels were presented in 45 (96%) of the 47 cases. In contrast, none of the dialogues included information regarding the treatment course. In designing the prompts, the number of dialogue turns was set at 25 for each participant (50 in total), considering the average consultation time in Japanese outpatient settings, which is approximately 5 minutes. As an exploratory extension, we tested longer dialogues (50 and 100 turns per speaker). Although brief mentions of treatment began to appear, these dialogues became increasingly verbose, with redundant differential diagnoses and questioning. This indicated a decline in dialogue quality and educational effectiveness, highlighting the need to identify an optimal dialogue length in prompt design. Although the generation of information related to the treatment course may be improved through more sophisticated prompt engineering [17,18] or future updates to generative AI models [19], prior studies have demonstrated that diagnostic reasoning and treatment management involve fundamentally distinct cognitive processes [20]. Therefore, separating these processes into distinct dialogue scenarios may be more educationally effective than integrating them into a single prompt.

The quality of the generated dialogues, produced in Japanese, was evaluated based on 6 criteria. Regarding criterion 1, coherence of the conversation, the dialogues were generally smooth and grammatically correct. This reflects the high-level natural language processing capabilities of generative AI and suggests its potential to replicate physician-patient interactions at a basic level of linguistic fidelity within the context of Japanese-language medical interviews. Regarding criterion 2, medical accuracy of the patient's statements, although there were instances in which the patient did not provide clear responses to the physician's questions, particularly in certain disease contexts, such ambiguity may reflect the nature of real-world clinical encounters. From an educational perspective, these instances may be valuable for simulating authentic dialogue dynamics. Criterion 3, medical accuracy of the physician's statements, received comparatively lower ratings than other criteria. This may be attributable to the perceived need for greater domain-specific precision, particularly when formulating clinical questions and providing medically accurate explanations to patients. This finding aligns with previous research on the limitations of generative AI in clinical reasoning [12]. Regarding criterion 4, quality of the physician's history

XSL•FO
RenderX

taking, in actual clinical practice, once the probability of a particular disease increases based on the patient's narrative, physicians typically engage in further in-depth inquiry. Generative AI appears to struggle with appropriately weighing clinical information and identifying which elements warrant further exploration; thus, at present, AI may be limited in its capacity to conduct history taking guided by probabilistic diagnostic reasoning. Regarding criterion 5, communication skills, the dialogues showed limited use of verbal cues such as acknowledgments or responses that convey an understanding of the patient's statements, and expressions of empathy toward patients' concerns were insufficient. In addition, there were a few attempts to elicit a patient's explanatory model, which is a critical step toward building rapport. Prompt design that encourages empathic and interactive responses may help address these limitations in future development. Regarding criterion 6, professionalism, although no ethically inappropriate expressions were identified, the dialogues generally lacked explicit demonstrations of patient-centered attitudes, respect, or empathic concern.

It is important to note that, in face-to-face medical interviews, nonverbal cues, such as facial expressions and nodding, play a substantially role in promoting patient satisfaction and emotional attunement [21,22], and the absence of such elements constitutes a fundamental limitation of text-based dialogue evaluations. Moreover, a small number of dialogues contained clinically inappropriate content, such as failure to provide justification for a diagnosis or incorrect assertion that a typically incurable disease could be cured. However, no ethically problematic or professionally inappropriate statements were identified in the dataset. Taken together, these results, along with the overall composite score of 5.7 (SD 1.0), suggest that, while physician oversight and revision are essential, generative AI–based dialogues may serve as a valuable educational resource for teaching clinical communication skills to medical students and early-stage trainees.

Because of their underlying architecture, LLMs exhibit inherent output variability, and identical prompts do not consistently yield identical responses. Although this randomness poses challenges in terms of reproducibility and control, it offers opportunities for prompt-based modulation, whereby carefully designed prompts can elicit diverse and contextually appropriate outputs [23]. In this respect, the ability to generate a wide range of physician-patient dialogues represents a particularly compelling and pedagogically valuable feature.

In this study, to facilitate the acquisition of fundamental structures in history taking, the AI was instructed to assume the role of a cooperative patient who provided clear and responsive answers. However, the simulated patient did not need to be restricted to cooperative profiles. It is also feasible to generate dialogues featuring patients with diverse communicative behaviors, such as those who are angry, uncommunicative, or unable to articulate clearly because of their underlying health conditions.

On the basis of our findings, we suggest that physician-patient dialogues generated by generative AI can be developed into high-quality educational materials with relatively minimal effort, provided that particular attention is paid to the medical accuracy of the physician's utterances and that supervising physicians revise the content as necessary. Given the capability of generative AI to rapidly produce medically relevant content, its application in medical education and clinical practice is expected to expand further in the coming years [24,25]. As demonstrated in previous studies, when generative AI is used in medical education [26], the ability of instructors to review and modify the generated content enables its use as a supplementary instructional resource. Considering the capacity of the model to generate dialogues tailored to a wide range of clinical scenarios, this approach, when used with appropriate oversight, could offer a highly flexible and scalable educational tool.

Limitations

This study had 3 primary limitations. First, the version of the generative AI used in this study was gpt-4o-2024-11-20, and the evaluation was based on the outputs generated on November 13, 2024. As the performance of generative AI models may evolve with future updates, periodic reevaluation is necessary. Moreover, as multiple LLMs are available, the quality and characteristics of dialogues generated by other models may differ from those evaluated in this study.

Second, there is currently no standardized method for prompt construction when interacting with generative AI, and the content of the input prompt can significantly influence the quality and nature of the output. In this study, the extent to which variations in the input conditions affect the generated dialogues was not examined systematically. Further investigations are warranted to clarify the impact of prompt design on output quality.

Third, this study was conducted in Japanese, and all prompts and generated dialogues were also in Japanese. Although prior research suggests a strong potential for adaptation to other languages, further validation is warranted.

Conclusions

In this study, physician-patient dialogues were generated using an LLM, that is, generative AI. These findings indicate that, although physician supervision remains essential, such dialogues can be easily developed into materials suitable for use in medical education. In particular, dialogues that incorporate the patient's perspective and the interactive elements of physician-patient communication could provide practical learning experiences that are difficult to achieve with conventional educational resources, suggesting a wide range of possible applications in medical education.

Moreover, the use of generative AI enables the efficient and large-scale creation of diverse educational content. This represents a significant advantage in terms of reducing the substantial time and effort traditionally required to develop medical instructional materials. By using this approach, medical students are expected to have more opportunities to learn in contexts that closely resemble actual clinical practice, thereby facilitating the acquisition of more practical medical interview skills.

XSL•FO
RenderX

Acknowledgments

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this paper.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

YY, DY, and TU designed and coordinated the study. YY and DY carried out data analysis and interpretation. YY, DY, and TU drafted the manuscript. SI, RY, YO, and TU revised the manuscript for important intellectual content. All authors read and approved the final manuscript and take responsibility for all aspects of the work, ensuring that any questions regarding accuracy or integrity are appropriately investigated and resolved. This work was supported by JSPS KAKENHI (grant JP25K20499).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Dialogue with a perfect average score (case 21: migraine).
[DOCX File , 18 KB - formative_v9i1e80752_app1.docx]

Multimedia Appendix 2
Dialogue with a lower average score of 4 (case 27: acute eosinophilic pneumonia).
[DOCX File , 17 KB - formative_v9i1e80752_app2.docx]

References

1. Mu Y, He D. The potential applications and challenges of ChatGPT in the medical field. Int J Gen Med 2024 Mar;Volume 17:817-826. [doi: 10.2147/ijgm.s456659]
2. Laymouna M, Ma Y, Lessard D, Schuster T, Engler K, Lebouché B. Roles, users, benefits, and limitations of chatbots in health care: rapid review. J Med Internet Res 2024 Jul 23;26:e56930 [FREE Full text] [doi: 10.2196/56930] [Medline: 39042446]
3. Yanagita Y, Yokokawa D, Uchida S, Tawara J, Ikusaka M. Accuracy of ChatGPT on medical questions in the national medical licensing examination in Japan: evaluation study. JMIR Form Res 2023 Oct 13;7:e48023 [FREE Full text] [doi: 10.2196/48023] [Medline: 37831496]
4. Fukuzawa F, Yanagita Y, Yokokawa D, Uchida S, Yamashita S, Li Y, et al. Importance of patient history in artificial intelligence-assisted medical diagnosis: comparison study. JMIR Med Educ 2024 Apr 08;10:e52674 [FREE Full text] [doi: 10.2196/52674] [Medline: 38602313]
5. Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. JAMA 2023 Jul 03;330(1):78-80 [FREE Full text] [doi: 10.1001/jama.2023.8288] [Medline: 37318797]
6. Coşkun Ö, Kıyak YS, Budakoğlu I. ChatGPT to generate clinical vignettes for teaching and multiple-choice questions for assessment: a randomized controlled experiment. Med Teach 2025 Feb;47(2):268-274. [doi: 10.1080/0142159X.2024.2327477] [Medline: 38478902]
7. Tremblay D, Turcotte A, Touati N, Poder TG, Kilpatrick K, Bilodeau K, et al. Development and use of research vignettes to collect qualitative data from healthcare professionals: a scoping review. BMJ Open 2022 Jan 31;12(1):e057095 [FREE Full text] [doi: 10.1136/bmjopen-2021-057095] [Medline: 35105654]
8. Wofford JL, Singh S. Exploring the educational value of clinical vignettes from the Society of General Internal Medicine national meeting in the internal medicine clerkship: a pilot study. J Gen Intern Med 2006 Nov;21(11):1195-1197 [FREE Full text] [doi: 10.1111/j.1525-1497.2006.00596.x] [Medline: 17026730]
9. Trullàs JC, Blay C, Sarri E, Pujol R. Effectiveness of problem-based learning methodology in undergraduate medical education: a scoping review. BMC Med Educ 2022 Feb 17;22(1):104 [FREE Full text] [doi: 10.1186/s12909-022-03154-8] [Medline: 35177063]
10. Evans SC, Roberts MC, Keeley JW, Blossom JB, Amaro CM, Garcia AM, et al. Vignette methodologies for studying clinicians' decision-making: validity, utility, and application in ICD-11 field studies. Int J Clin Health Psychol 2015;15(2):160-170 [FREE Full text] [doi: 10.1016/j.ijchp.2014.12.001] [Medline: 30487833]

XSL•FO
RenderX

11. Yanagita Y, Yokokawa D, Fukuzawa F, Uchida S, Uehara T, Ikusaka M. Expert assessment of ChatGPT's ability to generate illness scripts: an evaluative study. BMC Med Educ 2024 May 15;24(1):536 [FREE Full text] [doi: 10.1186/s12909-024-05534-8] [Medline: 38750546]
12. Yanagita Y, Yokokawa D, Uchida S, Li Y, Uehara T, Ikusaka M. Can AI-generated clinical vignettes in Japanese be used medically and linguistically? J Gen Intern Med 2024 Dec;39(16):3282-3289. [doi: 10.1007/s11606-024-09031-y] [Medline: 39313665]
13. Questions and answers for the 118th National Medical Examination. Ministry of Health, Labour, and Welfare Japan. 2024. URL: https://www.mhlw.go.jp/seisakunitsuite/bunya/kenkou_iryou/iryou/topics/tp240424-01.html [accessed 2025-10-27]
14. Model OpenAI API. OpenAI Platform. URL: https://platform.openai.com/docs/models#gpt-4o [accessed 2024-12-05]
15. Silverman J, Kurtz S, Draper J. Skills for Communicating with Patients. Boca Raton, FL: CRC Press; 2013.
16. Martinsen SS, Espeland T, Berg EA, Samstad E, Lillebo B, Slørdahl TS. Examining the educational impact of the mini-CEX: a randomised controlled study. BMC Med Educ 2021 Apr 21;21(1):228 [FREE Full text] [doi: 10.1186/s12909-021-02670-3] [Medline: 33882913]
17. Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. J Med Internet Res 2023 Oct 04;25:e50638 [FREE Full text] [doi: 10.2196/50638] [Medline: 37792434]
18. Liu X, Wang J, Sun J, Yuan X, Dong G, Di P, et al. Prompting frameworks for large language models: a survey. ArXiv Preprint posted online on November 21, 2023 [FREE Full text] [doi: 10.48550/arXiv.2311.12785]
19. Thirunavukarasu AJ, Ting DS, Elangovan K, Gutierrez L, Tan TF, Ting DS. Large language models in medicine. Nat Med 2023 Aug;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
20. Cook DA, Sherbino J, Durning SJ. Management reasoning: beyond the diagnosis. JAMA 2018 Jun 12;319(22):2267-2268. [doi: 10.1001/jama.2018.4385] [Medline: 29800012]
21. Collins LG, Schrimmer A, Diamond J, Burke J. Evaluating verbal and non-verbal communication skills, in an ethnogeriatric OSCE. Patient Educ Couns 2011 May;83(2):158-162. [doi: 10.1016/j.pec.2010.05.012] [Medline: 20561763]
22. Little P, White P, Kelly J, Everitt H, Gashi S, Bikker A, et al. Verbal and non-verbal behaviour and patient perception of communication in primary care: an observational study. Br J Gen Pract 2015 May 25;65(635):e357-e365. [doi: 10.3399/bjgp15x685249]
23. Lyu Q, Tan J, Zapadka ME, Ponnatapura J, Niu C, Myers KJ, et al. Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: results, limitations, and potential. Vis Comput Ind Biomed Art 2023 May 18;6(1):9 [FREE Full text] [doi: 10.1186/s42492-023-00136-5] [Medline: 37198498]
24. Karabacak M, Ozkara BB, Margetis K, Wintermark M, Bisdas S. The advent of generative language models in medical education. JMIR Med Educ 2023 Jun 06;9:e48163 [FREE Full text] [doi: 10.2196/48163] [Medline: 37279048]
25. Li Q, Qin Y. AI in medical education: medical student perception, curriculum recommendations and design suggestions. BMC Med Educ 2023 Nov 09;23(1):852. [doi: 10.1186/s12909-023-04700-8] [Medline: 37946176]
26. Han Z, Battaglia F, Udaiyar A, Fooks A, Terlecky SR. An explorative assessment of ChatGPT as an aid in medical education: use it with caution. Med Teach 2024 May;46(5):657-664. [doi: 10.1080/0142159X.2023.2271159] [Medline: 37862566]

Abbreviations

AI: artificial intelligence
LLM: large language model
Mini-CEX: Mini-Clinical Evaluation Exercise

Edited by A Mavragani; submitted 16.07.25; peer-reviewed by M Al-Agil, MA Roshani, S Mohamed Shaffi, S Banerjee, X Liang; comments to author 08.09.25; revised version received 18.09.25; accepted 15.10.25; published 03.11.25.

Please cite as:
Yanagita Y, Yokokawa D, Ihara S, Yoshida R, Okano Y, Uehara T
Quality Assessment of Large Language Model–Generated Medical Dialogue for Clinical Vignettes: Evaluation Study
JMIR Form Res 2025;9:e80752
URL: https://formative.jmir.org/2025/1/e80752
doi:10.2196/80752
PMID:

©Yasutaka Yanagita, Daiki Yokokawa, Shiichi Ihara, Ryo Yoshida, Yoshihide Okano, Takanori Uehara. Originally published in JMIR Formative Research (https://formative.jmir.org), 03.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly

XSL•FO
RenderX

cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Quantitative Assessment of Strabismus Using Cloud AI Computing: Validation Study

Junxian He[1,2*], BE; Jiawei Zhang[3*], BMed; Zheng Wang[4], PhD; Shrinivas Pundlik[1], PhD; Rui Liu[3,5,6], MD, PhD; Gang Luo[1], PhD

[1]Schepens Eye Research Institute of Mass Eye & Ear, Harvard Medical School, Boston, MA, United States
[2]Chongqing University, Chongqing, China
[3]Department of Ophthalmology, Eye & ENT Hospital of Fudan University, Shanghai, China
[4]School of Medicine, Jiaxing University, Jiaxing, China
[5]Key Laboratory of Myopia and Related Eye Diseases, Fudan University, Shanghai, China
[6]Shanghai Key Laboratory of Visual Impairment and Restoration, Shanghai, China
[*]these authors contributed equally

Corresponding Author:
Rui Liu, MD, PhD
Department of Ophthalmology
Eye & ENT Hospital of Fudan University
83 Fenyang Road, Xuhui District
Shanghai, 200031
China
Phone: 86 13764064088
Email: lratb1@aliyun.com

Abstract

Background: Strabismus measurement is essential in vision assessment and screening. It typically requires skilled clinicians or specialized equipment. Photographic strabismus measurement methods have value in terms of accessibility and convenience of use.

Objective: This study aimed to evaluate Eyeturn Cloud, a cloud-based artificial intelligence (AI) system for measuring strabismus angles based on eye images captured with smartphone cameras under cover test conditions.

Methods: The Eyeturn Cloud web app uses AI models to recognize eyes, eye lid, and iris, and then to segment iris precisely. It then computes strabismus based on ellipse fitting of the iris boundary and corneal reflection. The system was evaluated in patients (without glasses) with manifest strabismus and control participants. Clinicians measured eye deviations using the prism alternate cover test and also captured pictures of their eyes under alternate cover and unilateral cover conditions. The pictures were processed by Eyeturn Cloud.

Results: In total, 79 (mean age 11.9, SD 6.3 years; esotropia: n=15, exotropia: n=55, orthotropia: n=9) participants were enrolled; of which, data were available for 71 participants (8/79, 10.1% processing failure). The range of prism alternate cover test strabismus magnitude was from 78 base in to 78 base out prism diopters (PDs). A strong correlation was found between Eyeturn Cloud and clinical measurements (R^2=0.95; slope=0.91; P<.001). Bland-Altman analysis revealed that 95% limits of agreement between the 2 measurements were –20.2 to 14.6 PD. A repeatability test with 15 participants (4 photos each) found a 1.53 PD SD.

Conclusions: The cloud AI web app can compute strabismus angles reliably under alternate and unilateral cover conditions in clinical settings, and its potential for use in telehealth settings needs further evaluation.

(JMIR Form Res 2025;9:e79280) doi:10.2196/79280

KEYWORDS

strabismus; cloud computing; AI models; artificial intelligence; smartphone; web app

XSL•FO
RenderX

Introduction

Strabismus, a condition where the eyes are misaligned, can cause a wide range of visual dysfunctions in children [1,2] and adults [3]. According to the survey [4], strabismus not only negatively impacts activities of daily living but also causes self-concerns regarding appearance to others, even in most patients without diplopia. Current clinical assessments of strabismus are based on some form of cover testing with prisms to determine the presence and the magnitude of eye deviation. This procedure requires skilled and experienced clinicians using instruments such as a prism bar or synoptophore and also a high degree of cooperation from the patients for accurate assessments.

With the development of digital health care technology, various approaches have been developed for automating or simplifying the strabismus measurement process, including eye tracker–based systems [5-7], virtual reality–based systems [8], and smart glasses [9]. For vision screening purposes [10], some commercial screening devices such as Volk Eye Check, Spot Vision Screener, Plusoptix, iScreen, 2WIN, GazeLab [11], neos [12], and blinq are available. These devices are typically built on optoelectronic technologies and are relatively easier to use than conventional prisms. Some, for example, the blinq device, have shown very high sensitivity and specificity for strabismus detection [13]. However, these devices typically only raise red flags when concerning strabismus is detected and generally do not provide quantitative measures (with some exceptions, eg, 2WIN).

For the reasons of costs, accessibility, and sometimes operating difficulties, dedicated screening devices are not available in many settings, such as resource-limited regions, nursing homes, and in telehealth scenarios [14,15]. In these settings, photographic methods, which just need a picture of patients' eyes, may be preferable.

Approaches to measure strabismus based on a photograph include using dedicated cameras [16-18], or more recently, using a smartphone camera and flash to take a picture of the eyes and detect strabismus [19-24]. Other automated systems have used different technologies, such as virtual reality headsets or dedicated infrared eye trackers [25,26]. Strabismus assessment based on captured photos using web tools [27] and artificial intelligence (AI) models [28] has been developed. Photographic assessment of strabismus, via mobile apps, cloud computing, and AI algorithms, offers the advantages of accessibility and convenience.

Most photoscreening devices, including smartphone apps, are limited to measuring manifested strabismus, and intermittent patients with strabismus are likely to be missed. The ability to elicit and capture the eye movement based on covering 1 or both eye alternatingly remains a challenging problem for mobile apps and many other photoscreening devices. In fact, in our previous study of the Eyeturn app [19], a stand-alone photographic strabismus measurement tool, we indicated that the cover or uncover feature within the app required offline processing of the videos. In addition to the inability to effectively measure strabismus under cover or uncover or alternating cover (AC) mode, limitations of the prior stand-alone Eyeturn app include on-device model constraints (only lightweight models capable of running on mobile hardware could be used), uncorrected corneal-reflex or iris detection errors and lack of quality control, device variability that caused software mismatch with hardware, and inability to reprocess previously saved images (the app only worked with live camera capture).

To overcome this challenge, we have developed a smartphone app (cover test app) to enable automatic picture capture under various cover or uncover conditions [29] and an Eyeturn Cloud web app for strabismus angle calculation based on the uploaded pictures of the eyes. In this study, the cover test app is merely a facilitator for the precise capture of eye movements at the instant of uncovering (see demonstration video in Multimedia Appendix 1). The present cloud system addresses these gaps by using a different workflow for image capture in different cover test modes, which improves processing via various large computer vision models. The current cloud app can be used for retrospective reprocessing of archived images for recalculation and audit.

By separating photo capturing and image analysis, the Eyeturn Cloud web app was developed with the ultimate goal of improving accessibility and reducing dependence on specialized equipment, dedicated apps, or skilled operators for strabismus assessment. While the cover test app may be convenient, it is not strictly necessary to capture images for uploading to Eyeturn Cloud. One can use regular digital cameras or native phone camera apps to take photos, if the goal is to assess manifested strabismus (cover is not needed). Used together with either regular phone camera app or our dedicated cover test app [29], the proposed solution could be potentially used for vision screenings in resource-limited areas and for telehealth settings (such as virtual visits for prescreening or follow-up [14] and for vision rehabilitation [15]), where pictures taken by nonprofessional personnel could be uploaded for quantitative analysis and then reviewed by an ophthalmic expert at remote sites.

Together with other complementary apps for measuring various visual functions (eg, visual acuity and refractive error), Eyeturn Cloud may potentially help in enabling a more accessible and scalable solution in the management of strabismus. For patient subgroups such as children, a rapid, photographic-based method may improve cooperation compared to conventional examinations that require prolonged fixation [30]. Furthermore, obtaining objective, quantitative measurements is critical for monitoring disease progression or treatment response over time, which is a cornerstone of pediatric strabismus management [31]. Access to pediatric ophthalmology is uneven, with notable rural and underserved gaps. Contemporary screening recommendations endorse early detection but often lack scalable, quantitative tools deployable outside specialty clinics. By enabling photographic acquisition with cloud-based, standardized quantification, our system can provide actionable, reproducible measurements for triage and follow-up by nonspecialist personnel (eg, school nurses and primary care staff) and in community settings, potentially shortening time-to-referral and improving longitudinal care. Importantly, no specialized equipment or specialist presence is required. A

XSL•FO
RenderX

simple photograph suffices for measurement and can be acquired in routine environments without dedicated clinical instrumentation.

This study quantitatively evaluates the strabismus measurement approach using the cover test app for image capture and the AI-powered cloud computing paradigm for strabismus angle calculation. The primary objective was to evaluate the cloud-AI platform's effectiveness in accurately measuring strabismus angles. We hypothesize that the Eyeturn Cloud platform can provide quantitative measurements of strabismus that are highly consistent with the clinical gold standard of cover testing with prism. The primary outcomes for this validation are the coefficient of determination (R^2) from linear regression and the 95% limits of agreement from Bland-Altman analysis. A secondary outcome is the test-retest reliability of the platform's measurements.

Methods

Pipeline of Strabismus Calculation

The entire image processing is executed in a cloud app instance running in Amazon Web Services (AWS). As illustrated in Figure 1, the app includes 4 key steps to calculate eye deviation. First, using the YOLOv5-face (You Look Only Once) model [32], the eyes are detected, localized, and cropped. Next, using an hourglass network [33] trained with a synthesized eye image set [34], the iris is localized in cropped eye images. Using the center of the iris as a reference point, the Segment Anything Model (SAM) [35] is used to segment the visible iris region. Finally, an ellipse is fitted to the segmented iris. The center of the ellipse is considered as the precise location of the iris center. Combining with corneal reflection points, which are detected by simple thresholding, eye deviation can be calculated according to the Hirschberg method [16,20,36]. We did not train YOLOv5-face and SAM models. Pretrained versions were used with default parameters for inference.

Figure 1. Image processing pipeline of Eyeturn Cloud web app for eye deviation calculation. The processing includes 4 steps—eye localization, iris localization, iris segmentation, iris boundary fitting, and deviation calculation. SAM: Segment Anything Model; YOLO: You Look Only Once.

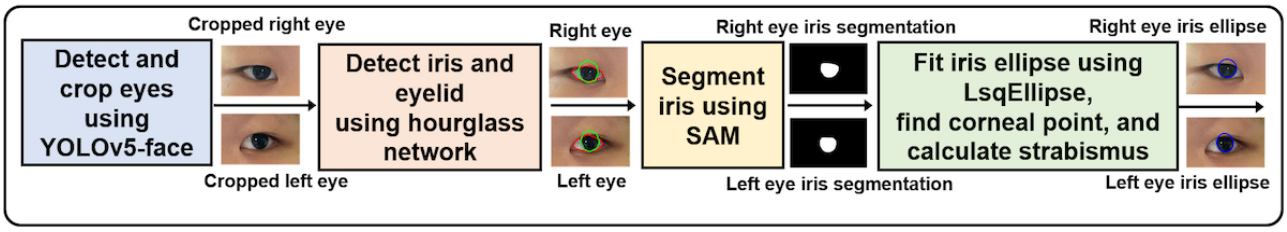

Iris Detection

Iris is detected using a coarse-to-fine approach. After obtaining the coordinates of the eyes, the left and right eye regions are cropped from the input image and resized to 72×120 pixels. The subimages are processed by the hourglass network for localizing the iris and eyelids. The network is a stacked-hourglass network trained on a synthetic eye image set [34]. Detected iris and eyelid landmarks are represented by polygons, which approximate the curves of those eye features. Based on our evaluation, the center of the iris polygon is not precise enough for strabismus calculation. Therefore, the iris center is then selected as the reference point for more precise segmentation using SAM [35], which segments the visible iris region pixel-wise. SAM generates 3 segmentation results ranked by probability. The result with the highest probability is not necessarily the best segmentation. Sometimes it corresponds to the pupil rather than the iris. To address this, we calculate the number of pixels within the area delineated by the iris segmentation results and exclude the pixels that are outside the eyelid polygon. The number of remaining pixels in each segmentation and shape are verified for reasonability. The closest match to the expected iris is chosen as the final segmentation.

Since the iris is partially occluded by eyelids in most cases, the center of the segmented iris is not the true iris center. Therefore, an ellipse is fitted to the boundary of the segmented iris. To achieve a better elliptical fit, we applied morphological opening to the segmented iris to smooth the boundary without significantly altering the area. Then, the segmentation edge of the iris was extracted, retaining only the curved lines on the left and right sides. Using the *LsqEllipse* package [37,38], an ellipse was fitted to these curved lines to represent the shape of the iris, and the center of the ellipse was regarded as the precise center of the iris.

The AI models for ocular landmark detection or segmentation were trained on diverse datasets from different image sources. Therefore, it is anticipated that the strabismus measurement method is hardware agnostic, and no retraining is needed for different acquisition hardware.

Cloud App: Eyeturn Cloud

The front-end of Eyeturn Cloud was developed using Cascading Style Sheets and JavaScript and was initiated with the Python *Flask* package. The Eyeturn Cloud website prompts users to upload a photo, and a "Process image" button appears once the upload is successful. Upon clicking the button, the back-end is activated to perform strabismus calculations (Figure 2). The back-end, developed in Python, was responsible for calculating the strabismus result based on the photo sent from the front-end. Once the calculation is completed, the cropped and processed eye images, along with the strabismus result, are displayed. The user could then verify the accuracy of the strabismus result by reviewing the processed eye images, with the estimated iris

XSL•FO
RenderX

ellipse and corneal light reflex points overlaid. The Eyeturn Cloud website is deployed on AWS. We selected a t2.2xlarge instance with 8 virtual central processing units to accelerate strabismus calculation and 32 GB RAM to accommodate the YOLOv5-face, High-Resolution Network, and SAM models. This instance was configured to run on Ubuntu 22.04 as a virtual server to ensure that the Eyeturn Cloud platform works continuously.

Figure 2. A screenshot of the Eyeturn Cloud interface when the web app was accessed from a phone. The uploaded picture shows that the participant covered his right eye with his hand unilaterally and then quickly uncovered. The ellipse fitting of iris is shown as a blue oval. If the fitting is obviously wrong, users should reject the result.

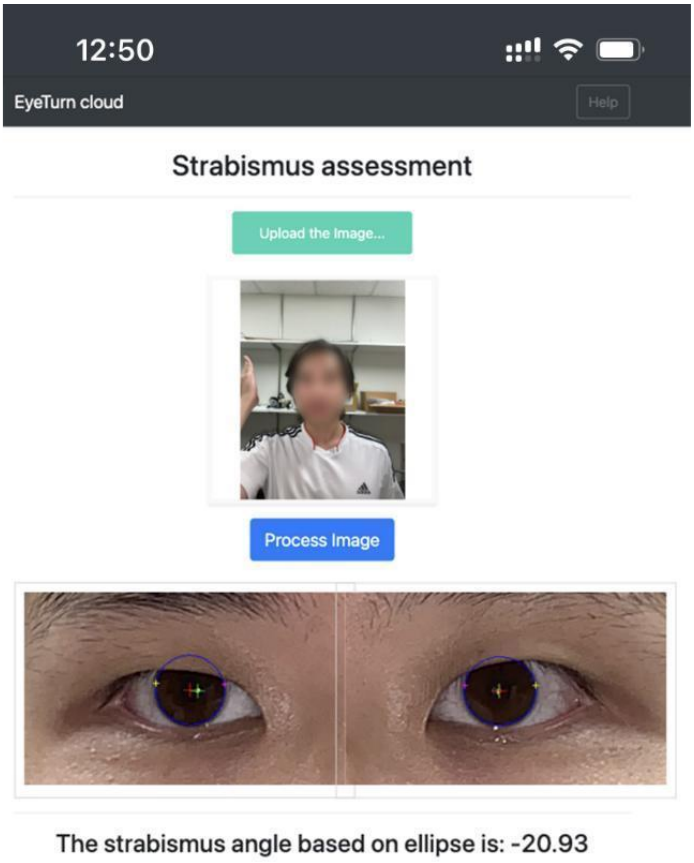

Cover Test App

A specialized app for capturing pictures during cover or uncover or AC tests was developed because the intermittent deviation needs to be computed at the precise moment, just after the cover is removed. Otherwise, there is a risk that the measured deviation via the app is different from the true deviation. The cover test app is based on the idea of detecting the frame when both eyes are simultaneously present [29]. The detection of eyes is done using custom-trained YOLOv5 deep learning model, capable of working in real time on contemporary iPhones. To develop the real-time eye detection model used in our data capture app (the cover test app), we created a custom dataset by collecting face images from public web sources and from our own captures. A total of 1013 eyes were manually annotated from these images. The dataset was split into training (810 eyes), validation (101 eyes), and testing (102 eyes) sets. We trained the model by fine-tuning the official YOLOv5 pretrained checkpoint on our custom dataset, using the source code with its default training parameters.

The procedure to perform cover or uncover or AC tests with the cover test app is similar to the conventional way. The examiner places an occluder in front of one eye, while the other eye fixates on the flash. The occluder is then removed away from the face by the examiner at an appropriate time, revealing both eyes. Once the app detects the presence of both eyes, it saves that frame, which can then be uploaded to Eyeturn Cloud for strabismus angle calculation. Since the app works in real time, there is very little delay in capturing the just-uncovered eye (approximately 33-66 milliseconds). In case of strabismus, the just uncovered eye is likely to be deviated, and this can be analyzed from the corneal reflection in the captured frame. Thus, the app facilitates measurement of intermittent strabismus, but it should be noted that Eyeturn Cloud does not have to be paired with this cover test app. It can process pictures captured through other ways; for instance, a frame extracted from a video.

Ethical Considerations

The clinical evaluation study was conducted in accordance with the tenets of the Declaration of Helsinki. The study was approved by the institutional review board of the Eye and ENT Hospital of Fudan University (Shanghai, China; approval: 2024126). Written informed consent was obtained from all participants prior to enrollment, including details about the study's purpose, procedures, potential risks and benefits, and the right to withdraw at any time without repercussions; for children, parental or guardian consent was obtained. Participant selection was not influenced by sex or age. Images were

XSL•FO
RenderX

processed locally on a computer (localhost) rather than uploaded to external cloud services like AWS during the study, minimizing data exposure. No identifiable information was shared beyond the research team. Participants did not receive any compensation for their participation. See Multimedia Appendix 2 for a completed checklist related to adherence to accepted guidelines for transparency, reproducibility, and methodological rigor in AI diagnostic studies. Data are not shared as per the conditions placed by the institutional review board to protect patient privacy.

Sample Size Calculation

From our previous study of the Eyeturn smartphone app [17], the SD of the difference between the app and clinical measurements was 4.1 prism diopters (PDs). Our preliminary analysis showed that our computation system can provide measurements with a reliability of ±3 PDs. Assuming twice the variability in this study and a type I error rate of 0.05 and 80% power, we arrive at a sample size of 64. Considering dropouts and capture errors, the calculated sample size was inflated further by 25% to arrive at a figure of 80 participants as the recruitment target. In total, 79 participants (mean age 11.9, SD 6.3; range 4-42 years) were enrolled from Eye and ENT Hospital of Fudan University (Shanghai, China).

Participant Recruitment

Patients with strabismus were recruited from the patient cohort treated by the author (RL) at their affiliated ophthalmology clinic. Inclusion criteria were a prior diagnosis of horizontal strabismus (constant or intermittent exotropia or esotropia) and no other visual impairments. Participants were eligible only if the best-corrected visual acuity (BCVA) in the worse eye was 20/40 or better (Snellen; 0.30 logMAR), using habitual correction or trial-frame refraction. Strabismic amblyopia was not an automatic exclusion if the BCVA threshold was met. In our cohort, only 3 participants had strabismic amblyopia, and they satisfied the BCVA better than or equal to 20/40 criterion. Therefore, they were included. As our current system is not sufficiently robust yet to process images of eye wearing glasses, patients with refractive errors higher than 1 diopter were excluded, concerning that they may not be able to fixate at the target properly without vision correction during the cover test. Cases of incomitant strabismus, such as paralytic strabismus, or those with a clinically significant vertical deviation (>5 PD) were also excluded. In addition, 9 participants without manifest strabismus were also enrolled to serve as control participants.

Study Procedures

Clinicians specialized in strabismus measurement performed the prism alternate cover test for near fixation (about 40 cm). In addition, they also took pictures under 3 cover conditions, AC, left eye cover-uncover (LC), and right eye cover-uncover (RC), using the cover test smartphone app as described earlier [29]. Those pictures were uploaded to the Eyeturn Cloud app, and the results are compared with prism alternate cover test measurements.

The image resolution used for this study was 3024 by 4032. The images were captured with a consumer iPhone 11 and 13, with autofocus at a distance of approximately 40 cm from the face.

We prioritized using the strabismus angle calculated from the AC images as the primary result in this study. If the AC image processing of a patient could not yield a result due to missing corneal reflection, SAM segmentation errors, or other issues, we selected the larger absolute value of the strabismus results from the LC or RC images as the final result.

Outcome Measures

The clinician's measurement was used as the ground truth, and the cloud computing results were analyzed by linear regression. A Bland-Altman analysis was also conducted to compare the physician's assessment and the Eyeturn Cloud results.

Results

Among the 79 enrolled participants (including 15 esotropia, 55 exotropia, and 9 orthotropia), strabismus was successfully calculated by the Eyeturn Cloud app for 71 participants. The processing failed for the remaining 8 (10.1%) participants due to failure to detect the corneal reflection for 1 participant and incorrect iris segmentation for 7 participants (Figure 3). The average age of the participants where processing failed was 11 (SD 5) years, which is similar to the age distribution of the entire sample 11.9 (SD 3.6) years. There were 4 male and 4 female participants, and an equal number of base in (43, 32, 77, and 40 PD) and base out (32, 77, 32, and 43 PD) deviations with processing failure. Since the study was conducted in China, the eye color was mostly uniform (deep brown). The failures did not seem to be related to demographic or clinical factors. This type of incorrect detection or segmentation was determined by visually examining the iris ellipse fitting overlaid on processed pictures.

In this study, we consider strabismus calculation larger than 104 PD to be an indication of incorrect image processing, since the highest 2 powers of the prism set used by the clinicians were 45 and 40 PD. With the 2 prisms used at the same time (1 prism on each eye), the largest strabismus would be 104 PD according to the lookup table of Cestari and Hunter [39]. None of the measurements in this study were above this threshold. The linear regression analysis and coefficient of determination are shown in Figure 4A, with a determination coefficient (R^2) of 0.95. The slope of the linear regression was close to 1 (0.91). When analyzed separately, the slopes of regression for base out and base in deviations were 0.908 and 0.904, respectively (since they were similar, the data were not separated). As the Bland-Altman plot showed in Figure 4B, the 95% limits of agreement was –20.2 to 14.6 PD.

Furthermore, image assessment based on different cover methods was compared using repeated measures ANOVA, which revealed a significant difference (P=.007; F_2=5.55). The averages of the 3 cover methods were 38.3 (SD 23.7), 30.8 (SD 17.2), and 32.1 (SD 17.8) PD for AC, LC, and RC, respectively. Bilateral cover test (based on AC images) revealed slightly but statistically significantly larger strabismus magnitudes (by about 6 to 8 PD) than unilateral cover test (based on LC and RC images). It is known that bilateral cover can better break

XSL•FO
RenderX

binocular fusion. In this analysis, 23 participants who had calculation results for all the AC, LC, and RC images were included (Figure 5). The other participants had at least 1 calculation missing. When analyzed separately, the regression line slopes for AC, LC, and RC were 0.9, 0.92, and 0.97, respectively.

To evaluate the repeatability of Eyeturn Cloud, another set of 15 participants (different than the ones reported earlier), with a mean age of 13.4 (SD 11.5) years (male: n=9 and female: n=6) and including 3 participants with exotropia, 3 with esotropia, and 9 with minimal deviation (<10 PD), were repeatedly photographed 4 times without covering. They were instructed to stare at the flash binocularly when pictures were taken. The variability across trials was calculated for each participant by subtracting the individual average. The SD of the variability was 1.53 PD. Therefore, the 95% CI (2σ) of the measurement was ±2.99 PD (Figure 6). This random noise level can be used as an indicator of the discrimination threshold of the system [40]. In addition, based on the image resolution, the minimum magnitude of phoria detectable (corresponding to 1 pixel for the image resolution of 3024×4032 used in the study) is about 2 PD, which is in a similar range as the measurement variability. Therefore, we rate the smallest amplitude of phoria detectable by the system to be ≈3PD.

Figure 3. Correct and incorrect examples of Eyeturn cloud processing. Upper panel: coarse iris and eyelid detection, middle panel: fine segmentation of iris, and lower panel: ellipse fitting of iris boundary. (A) A successful processing, (B) corneal reflection could not be detected because of many other reflections in the background, and (C) the segmentation and ellipse fitting of right eye iris are obviously incorrect; therefore, the calculation was rejected.

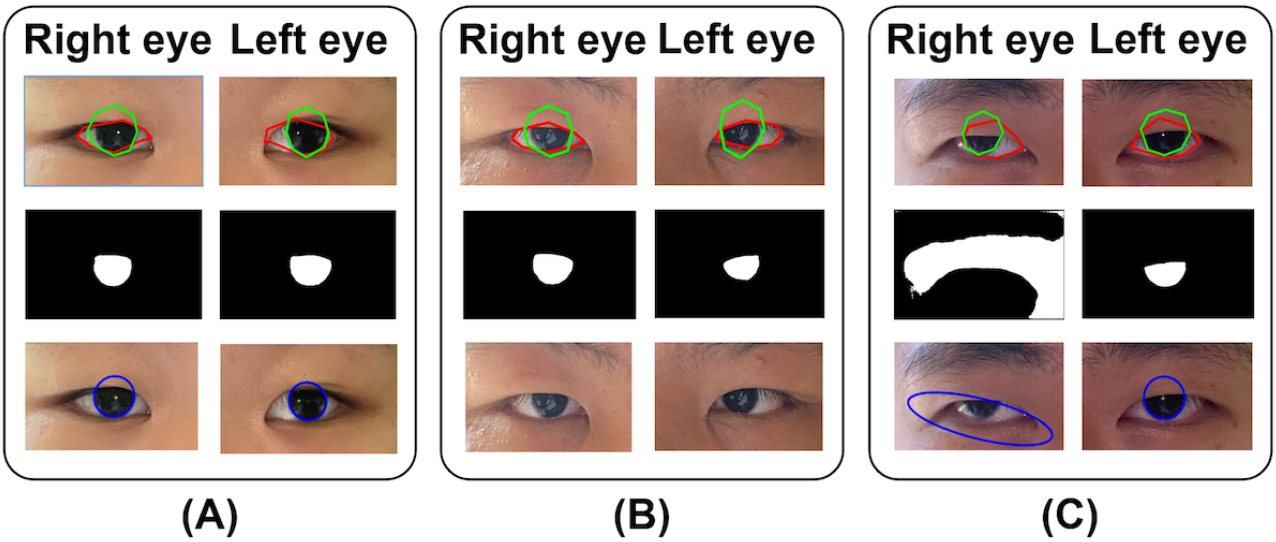

Figure 4. Correlation and agreement between photographic and clinical strabismus measurements. (A) Linear regression analysis comparing strabismus calculation results and physician measurements. (B) Bland-Altman plot for agreement between the strabismus detection system and the physician measurements. In both panels, data points derived from AC images (n=45) are shown as hollow circles, while points derived from unilateral cover (LC or RC) images (n=26) are shown as hollow triangles. AC: alternating cover; LC: left eye cover-uncover; PD: prism diopter; RC: right eye cover-uncover.

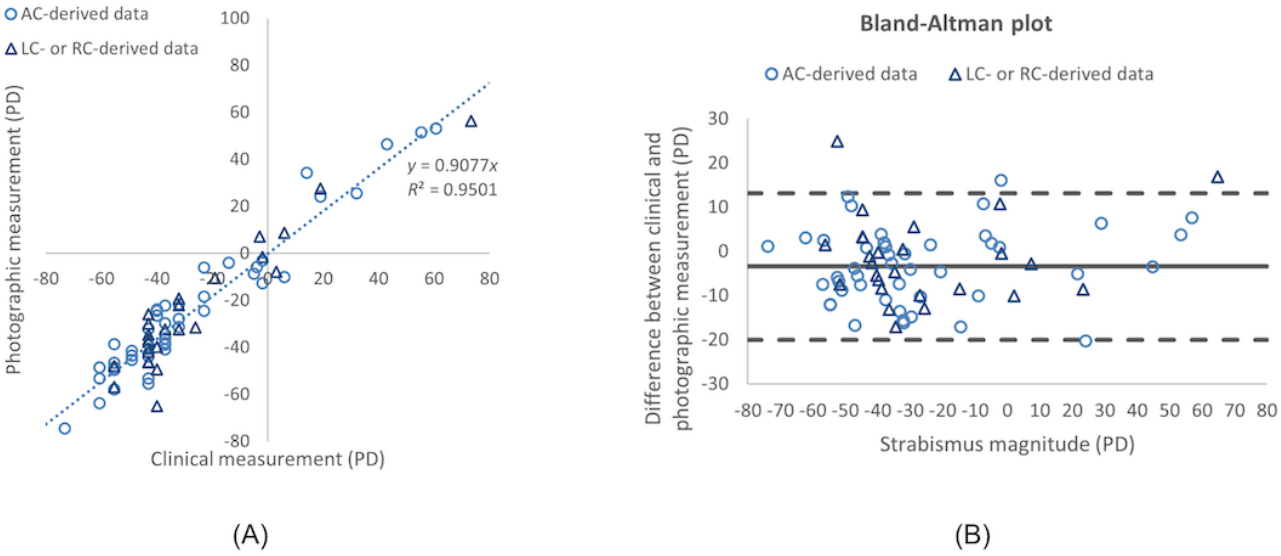

XSL•FO
RenderX

Figure 5. Each line in this graph represents the 3 strabismus calculation results based on AC, LC, and RC images, respectively. Bilateral covering (AC images) revealed larger strabismus than unilateral covering (LC and RC images). AC: alternating cover; LC: left eye cover-uncover; PD: prism diopter; RC: right eye cover-uncover.

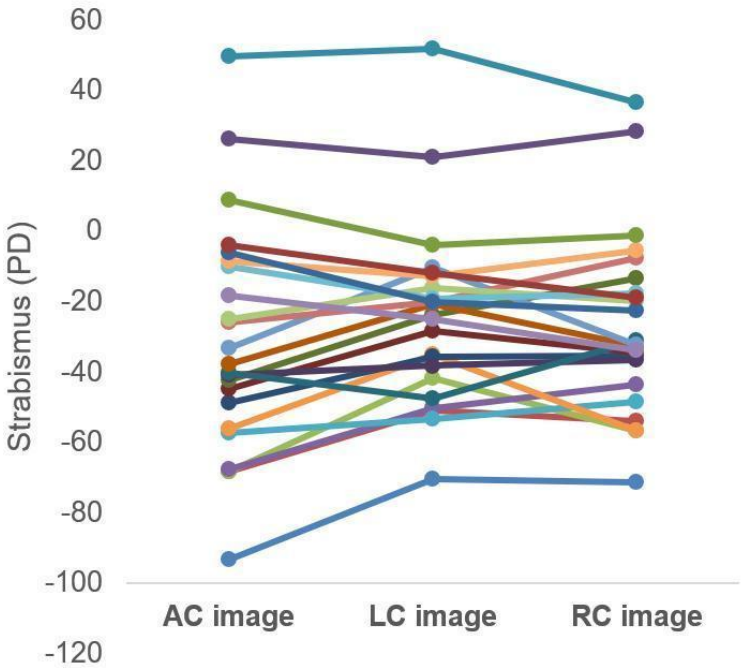

Figure 6. Repeatability of photographic assessment. Each participant (n=15) was photographed 4 times, and the residual errors after subtracting the individual's average were plotted here. Two dashed lines indicate ±2σ, which represent the 95% CI, ±2.99 PD. PD: prism diopter.

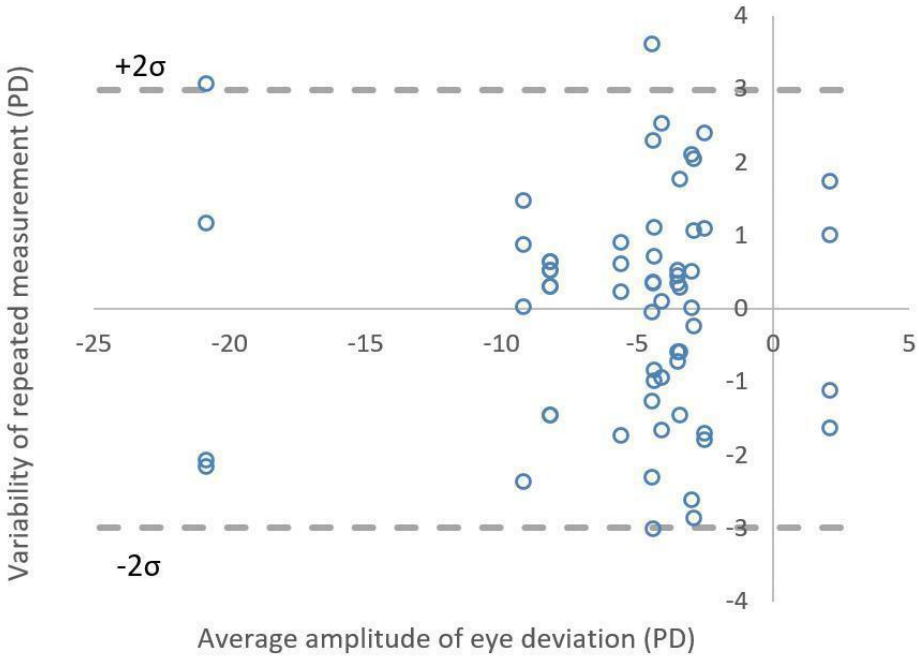

Discussion

Principal Findings

In this study, we developed and evaluated a web-based platform, Eyeturn Cloud, for strabismus angle calculation. We evaluated the level of agreement between the strabismus angles computed by Eyeturn Cloud based on uploaded pictures of patients' eyes and standard clinical measurements using prisms. The evaluation showed that the strabismus angles computed by the image-based cloud app had a strong linear correlation with clinical measurements (R^2=0.95; slope=0.91; Figure 4). The measures were obtained from different types of strabismus and for different types of cover conditions.

Comparison to Prior Work

While the limits of agreement (Figure 4B) may not seem to be small, 2 issues should be noted. First, the repeatability of prism measurement by clinicians may not be very high, depending on deviation amplitude, examiner skill, and patient factor. The Bland-Altman limits of agreement (–20.2 to +14.6 PD) for the

XSL•FO
RenderX

cloud app measurements are slightly worse than the interexaminer agreement limits (13 PD reported by de Jongh et al [41], or ±10 PD reported by Hatt et al [42], or ±11.7 PD reported by the Pediatric Eye Disease Investigator Group [43]) for large strabismus angles at near fixation. However, the limits of agreement for the Eyeturn Cloud app with prism testing compare favorably with other photographic strabismus assessment methods. For example, Strabocheck [27] evaluation reported 95% interval limit between –30.0 and 31.0 PD, and Garcia et al [23] reported agreement limits between –15.78 and 24.53 PD.

Second, the moderate error for small phoria of some control participants was due to the incorrect determination of esophoria versus exophoria, which leads to the incorrect sign of results. For instance, an exophoria of 3 PD was interpreted as an esophoria of –7PD. Thus, the disagreement became 10 PD. While it is still within the limits of agreement as mentioned earlier, the incorrect sign can be an error of clinical significance. There were 3 participants with orthophoria (3 PD base out, 6 PD base out, and 3 PD base in) misclassified in data analysis. One of the causes of this problem is that the visual axis is not precisely aligned with the pupil center because of the angle κ. Even when an eye is fixating on the flash, the corneal reflection is off the center nasally. In this study, we applied an offset of 4° angle κ in both eyes of participants with small deviation (<10 PD). As the angle κ of each individual eye was unknown, and because the current Eyeturn Cloud user interface does not have an option to indicate whether the uploaded image was taken with cover or not, or which eye was last uncovered if it was a cover test image, the deviated eye could not be identified simply by the larger offset. However, for relatively large deviations, base out and base in can be reliably determined based on the corneal reflection offset in the deviated eye. The current Eyeturn Cloud does not indicate base out or base in for users when the calculated deviation is smaller than 10 PD, in order to prevent misclassification in actual application. In this paper, the 3 participants with orthotropia were still counted misclassification just for the sake of data analysis. How to determine small esophoria and exophoria correctly is our future work. One possible approach can be through analyzing eye movement during the cover test based on videos. An alternative approach would be implementing a protocol where both eyes are captured monocularly (under unilateral cover conditions) as well as binocular condition (no cover). From the joint analysis of the 3 images, small-angle esotropia and exotropia could be classified with greater accuracy.

The proposed Eyeturn Cloud solution is similar to the Eyeturn smartphone app [19] in terms of the underlying approach of strabismus computation, although the landmarks within the eyes are now detected using AI algorithms. Compared to the 2WIN handheld photoscreening device that can provide quantitative strabismus measures, our approach is different in terms of imaging (visible vs infrared) and processing (iris vs pupil segmentation). Evaluation of the 2WIN portable device in 137 children with manifest strabismus and with dilated pupils showed R^2 values of 0.58 and 0.24 for esotropia and exotropia, respectively, when compared to alternate cover prism testing [21]. The maximum deviation included in that study was <40 PD. In addition, there are some differences between Eyeturn Cloud and previous web-based approaches that compute the strabismus angle from uploaded pictures. Compared to the Strabocheck website [27], which requires 3 pictures (left eye, right eye, and both eyes) and manually marking eye features, our approach can automatically assess strabismus based on a single picture, which can be captured under no cover, unilateral cover for 1 eye, or alternate cover conditions. Compared to the AI-based platform developed by Wu et al [28] for strabismus screening that detects the presence of strabismus from ocular pictures, Eyeturn Cloud provides a quantitative measurement of the eye deviation.

Strength

The Eyeturn Cloud platform enables the use of general digital devices, like smartphones and cameras, to capture and upload eye images for analysis by a web app in the cloud. This approach not only reduces the need for expensive, specialized equipment but also expands access to screening for a broader population. The platform's reliance on cloud infrastructure ensures that complex computations are handled remotely, eliminating the processing burden on local devices and allowing for batch processing of large datasets. This capability is particularly advantageous in large-scale screening programs, such as those conducted in schools, communities, or rural areas, where high-volume data processing must be processed efficiently. The ability to upload images offline and process them later when internet access is available is particularly crucial for rural sites.

While in this study, the pictures during unilateral cover or alternate cover testing were captured by a special-purpose cover test smartphone app, it should be noted that this app is not strictly necessary for capturing the appropriate pictures. Users could use any modern native camera apps (eg, live photo feature of iPhone camera app) or record a high-resolution (4K) video, and then, select the required frame when the eye is first visible after removal of the cover (see demonstration in Multimedia Appendix 1). In the future, the Eyeturn Cloud app could be updated to accept high-resolution videos and automatically select the appropriate frame based on the visibility of both eyes to compute strabismus angles. Furthermore, for constant strabismus assessment, cover is not needed. Thus, pictures taken with regular cameras can be processed by Eyeturn Cloud to yield valid results.

One potential advantage of cloud AI platforms is that they can facilitate the maintenance of digital archives of patient data, allowing ophthalmologists to monitor the progression of strabismus over time and adjust their treatment plans accordingly. Such a data-driven approach has the potential to enhance diagnostic accuracy and promote more personalized care, ultimately improving patient outcomes through timely interventions and continuous analysis. In addition, these AI platforms may support telemedicine-based screening and follow-up care [14,15], which is especially beneficial in areas lacking access to specialized ophthalmic equipment. This reduces the need for frequent patient visits and facilitates consistent remote monitoring [44]. Regarding screening, the Eyeturn smartphone app was previously used for strabismus screening in school children in a pilot study [45]. As the

XSL•FO
RenderX

underlying computation is conceptually similar, it is reasonable to expect that Eyeturn Cloud could also be potentially useful in such screening settings. However, the utility of Eyeturn Cloud in screening and telehealth settings is yet to be determined and is future work.

Limitations

There are still some unresolved issues with the current app. The cloud app was able to generate valid strabismus results for 71 patients, while the remaining 8 patients could not yield results. Among these, the results were invalid due to the absence of corneal light reflex points or incorrect iris segmentation by SAM. The issue of missing corneal light reflex points could potentially be resolved by selecting an appropriate environment for photography or by attempting multiple captures. On the other hand, eyes with glasses were excluded in this study, as the flash could interfere with the glasses causing spurious reflections, if caution was not taken during picture capturing. This could be solved in the future by more intelligent algorithms than simple thresholding methods.

In theory, our method could reliably measure ocular misalignment above 3 PD, but determining the directionality of misalignment is a challenge for photoscreening based on a single image. To mitigate sign instability near 0, the app withholds directionality when the estimated magnitude is <10 PD. Despite lack of directionality, strabismus magnitude measurement can still be useful information for vision screening.

There are also some limitations of this study. All of the pictures included in this study were all taken with our cover test app by clinicians in controlled settings rather than by lay people. This environment likely ensured optimal lighting, correct camera distance, and maximal patient cooperation. The usability, performance, and reliability of the Eyeturn Cloud platform when used by laypersons, such as parents at home for a telehealth consultation or by school nurses in a community screening program, remain unknown. In addition, pictures captured using other acquisition methods, such as native camera apps on other smartphones, have not been evaluated. Another limitation was that the number of base in deviation was higher than base out in our participants, due to the biased prevalence in the population [46], and future studies need to consider balancing the sample for more thorough analysis of app's performance for base out deviations. The study only evaluated images captured along primary gaze direction, and differentiating incomitant and concomitant strabismus and measuring the deviation patterns are future work. However, the findings of this study could guide future studies, where its utility for telehealth and screening applications could be fully evaluated.

Conclusions

The main novelty lies in the approach where we developed the cloud-AI app for strabismus angle computation based on a single picture, which can be acquired from different sources. While AI tools exist for strabismus screening from a single image [28], our app focuses on quantitative measurement of angular deviation. Our proposed solution was intended to address scenarios where a cover test is needed as well as to offer a possibility to assess photos captured with regular cameras without covering. Cover-testing mode is important for measuring patients with nonmanifest strabismus. The cloud-AI platform treats all image sources in the same way. Overall, the Eyeturn Cloud offers a convenient and accessible way of strabismus assessment. This study showed that quantitative assessment of strabismus with Eyeturn Cloud was repeatable and demonstrated a strong correlation with clinical measurements, although further work is required to improve its precision and address misclassifications of small-angle deviations. While the platform's architecture suggests potential for future application in vision screening and telehealth, its efficacy and reliability in these real-world settings, particularly when used by nonexperts with varied hardware, require rigorous future investigation before any such use can be recommended.

Acknowledgments

This research is supported in part by the China Medical Board (grant #24-558). Generative artificial intelligence was not used for manuscript generation.

Conflicts of Interest

SP and GL hold a patent related to strabismus measurement, which has been assigned to Mass Eye & Ear. Additionally, GL and SP serve as cofounders of EyeNexo, a startup company that develops smartphone apps for vision tests. However, EyeNexo is not involved in this study.

Multimedia Appendix 1
Demonstration video.
[MP4 File (MP4 Video), 9431 KB - formative_v9i1e79280_app1.mp4]

Multimedia Appendix 2
MAIC-10 brief quality checklist.
[PDF File (Adobe PDF File), 12 KB - formative_v9i1e79280_app2.pdf]

References

XSL•FO
RenderX

1. Robaei D, Rose K, Kifley A, Cosstick M, Ip J, Mitchell P. Factors associated with childhood strabismus: findings from a population-based study. Ophthalmology 2006;113(7):1146-1153. [doi: 10.1016/j.ophtha.2006.02.019] [Medline: 16675019]
2. Freeman AW, Nguyen VA, Jolly N. Components of visual acuity loss in strabismus. Vision Res 1996;36(5):765-774 [FREE Full text] [doi: 10.1016/0042-6989(95)00171-9] [Medline: 8762305]
3. Rowe F, VIS group UK. The profile of strabismus in stroke survivors. Eye (Lond) 2010;24(4):682-685. [doi: 10.1038/eye.2009.138] [Medline: 19521433]
4. Hatt SR, Leske DA, Kirgis PA, Bradley EA, Holmes JM. The effects of strabismus on quality of life in adults. Am J Ophthalmol 2007;144(5):643-647 [FREE Full text] [doi: 10.1016/j.ajo.2007.06.032] [Medline: 17707329]
5. Zou L, Tian T, Wygnanski-Jaffe T, Yehezkel O, Wang S, Moshkovitz A, et al. Effectiveness and repeatability of eye-tracking-based test in strabismus measurement of children. Semin Ophthalmol 2022;37(4):502-508. [doi: 10.1080/08820538.2021.2002918] [Medline: 34814794]
6. Economides JR, Dilbeck MD, Gentry TN, Horton JC. Ambulatory monitoring with eye tracking glasses to assess the severity of intermittent exotropia. Am J Ophthalmol 2023;250:120-129 [FREE Full text] [doi: 10.1016/j.ajo.2023.01.014] [Medline: 36681174]
7. Zhao Z, Li S, Wang J, Li X, Gao S. A strabismus widespread screening method based on wearable eye tracker. 2024 Presented at: IEEE BioSensors Conference (BioSensors); July 28-30, 2024; Cambridge, United Kingdom. [doi: 10.1109/biosensors61405.2024.10712701]
8. Mori DM, Kuchhangi A, Tame J, Cooper K, Hajkazemshirazi L, Indaram M, et al. Evaluation of a novel virtual reality simulated alternate cover test to assess strabismus: a prospective, masked study. Am J Ophthalmol 2025;269:266-272 [FREE Full text] [doi: 10.1016/j.ajo.2024.08.042] [Medline: 39245131]
9. Lee LC, Feng KM, Chuang PC, Chen YH, Chien KH. Preliminary data on a novel smart glasses system for measuring the angle of deviation in strabismus. Eye (Lond) 2023;37(13):2700-2706. [doi: 10.1038/s41433-023-02402-5] [Medline: 36702909]
10. Sopeyin A, Young BK, Howard MA. 2020 Evaluation of portable vision screening instruments. Yale J Biol Med 2021;94(1):107-114 [FREE Full text] [Medline: 33795987]
11. GazeLab. BCNINNOVA. URL: https://www.bcninnova.com/index.php?action=Gazelab [accessed 2025-10-11]
12. neos. machineMD. URL: https://www.machinemd.com/neos [accessed 2025-10-11]
13. Jost RM, Yanni SE, Beauchamp CL, Stager DR, Stager D, Dao L, et al. Beyond screening for risk factors: objective detection of strabismus and amblyopia. JAMA Ophthalmol 2014;132(7):814-820 [FREE Full text] [doi: 10.1001/jamaophthalmol.2014.424] [Medline: 24875453]
14. Bowe T, Hunter DG, Mantagos IS, Kazlas M, Jastrzembski BG, Gaier ED, et al. Virtual visits in ophthalmology: timely advice for implementation during the COVID-19 public health crisis. Telemed J E Health 2020;26(9):1113-1117 [FREE Full text] [doi: 10.1089/tmj.2020.0121] [Medline: 32408801]
15. Keilty M, Houston KE, Collins C, Trehan R, Chen Y, Merabet L, et al. Inpatient virtual vision clinic improves access to vision rehabilitation before and during the COVID-19 pandemic. Arch Rehabil Res Clin Transl 2021;3(1):100100 [FREE Full text] [doi: 10.1016/j.arrct.2020.100100] [Medline: 33363279]
16. Hasebe S, Ohtsuki H, Tadokoro Y, Okano M, Furuse T. The reliability of a video-enhanced Hirschberg test under clinical conditions. Invest Ophthalmol Vis Sci 1995;36(13):2678-2685. [Medline: 7499090]
17. de Almeida JDS, Silva AC, de Paiva AC, Teixeira JAM. Computational methodology for automatic detection of strabismus in digital images through Hirschberg test. Comput Biol Med 2012;42(1):135-146. [doi: 10.1016/j.compbiomed.2011.11.001] [Medline: 22119221]
18. Barry JC, Backes A. Limbus versus pupil center for ocular alignment measurement with corneal reflexes. Invest Ophthalmol Vis Sci 1997;38(12):2597-2607. [Medline: 9375579]
19. Pundlik S, Tomasi M, Liu R, Houston K, Luo G. Development and preliminary evaluation of a smartphone app for measuring eye alignment. Transl Vis Sci Technol 2019;8(1):19 [FREE Full text] [doi: 10.1167/tvst.8.1.19] [Medline: 30766761]
20. Luo G, Pundlik S, Tomasi M, Houston K. Using an automated Hirschberg test app to evaluate ocular alignment. J Vis Exp 2020(157):e60908. [doi: 10.3791/60908] [Medline: 32281979]
21. Racano E, Di Stefano G, Alessi S, Pertile R, Romanelli F. Validation of the 2WIN Corneal Reflexes App in children. Graefes Arch Clin Exp Ophthalmol 2021;259(6):1635-1642. [doi: 10.1007/s00417-020-05066-z] [Medline: 33404678]
22. Tenório Albuquerque Madruga Mesquita MJ, Azevedo Valente TL, de Almeida JDS, Meireles Teixeira JA, Cord Medina FM, Dos Santos AM. A mhealth application for automated detection and diagnosis of strabismus. Int J Med Inform 2021;153:104527. [doi: 10.1016/j.ijmedinf.2021.104527] [Medline: 34186433]
23. Garcia SSS, Santiago APD, Directo PMC. Evaluation of a Hirschberg test-based application for measuring ocular alignment and detecting strabismus. Curr Eye Res 2021;46(11):1768-1776. [doi: 10.1080/02713683.2021.1916038] [Medline: 33856941]
24. Raghuram A, Chinn R. Smartphone application: repeatability and validity in estimating manifest strabismus on children and adolescents. Invest Ophthalmol Vis Sci 2020;61(7):2132 [FREE Full text]
25. Vicini R, Brügger D, Grabe H, Abegg M. Automated measurement of strabismus angle using a commercial virtual reality headset. Klin Monbl Augenheilkd 2025;242(4):485-488. [doi: 10.1055/a-2466-0284] [Medline: 39694045]

XSL•FO
RenderX

26. Weber K, Rappoport D, Dysli M, Schmückle Meier T, Marks G, Bockisch C, et al. Strabismus measurements with novel video goggles. Ophthalmology 2017;124(12):1849-1856 [FREE Full text] [doi: 10.1016/j.ophtha.2017.06.020] [Medline: 28728924]
27. Azri C, Tomietto P, Marciano E, Bui-Quoc E. Evaluation of Strabocheck® as an objective measurement method of comitant horizontal strabismus in children undergoing surgery. Strabismus 2023;31(2):73-81. [doi: 10.1080/09273972.2023.2204886] [Medline: 37199169]
28. Wu D, Li Y, Zhang H, Yang X, Mao Y, Chen B, et al. An artificial intelligence platform for the screening and managing of strabismus. Eye (Lond) 2024;38(16):3101-3107. [doi: 10.1038/s41433-024-03228-5] [Medline: 39068250]
29. Luo G, He J, Chen G, Bhambhani G, Bachhav Y, Pundlik S. Measurement of strabismus using a mobile cover test app. ARVO 2025;66:1135 [FREE Full text]
30. Vijendran S, Alok Y, Kuzhuppilly NIR, Bhat JR, Kamath YS. Effectiveness of smartphone technology for detection of paediatric ocular diseases—a systematic review. BMC Ophthalmol 2025;25(1):323 [FREE Full text] [doi: 10.1186/s12886-025-04160-2] [Medline: 40448047]
31. Wallace DK, Christiansen SP, Sprunger DT, Melia M, Lee KA, Morse CL, et al. Esotropia and exotropia preferred practice Pattern®. Ophthalmology 2018;125(1):P143-P183. [doi: 10.1016/j.ophtha.2017.10.007] [Medline: 29108746]
32. Qi D, Tan W, Yao Q, Liu J. YOLO5Face: why reinventing a face detector. 2022 Presented at: European Conference on Computer Vision; October 23-27, 2022; Tel Aviv, Israel URL: https://link.springer.com/chapter/10.1007/978-3-031-25072-9_15
33. Newell A, Yang K, Deng J. Stacked hourglass networks for human pose estimation. 2016 Presented at: Computer Vision—ECCV; October 11-14, 2016; Amsterdam, The Netherlands URL: https://link.springer.com/chapter/10.1007/978-3-319-46484-8_29
34. Park S, Zhang X, Bulling A, Hilliges O. Learning to find eye region landmarks for remote gaze estimation in unconstrained settings. 2018 Presented at: ETRA '18: Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications; June 14, 2018; Warsaw, Poland URL: https://dl.acm.org/doi/10.1145/3204493.3204545
35. Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L. Segment anything. 2023 Presented at: IEEE/CVF International Conference on Computer Vision (ICCV); October 1-6, 2023; Paris, France. [doi: 10.1109/ICCV51070.2023.00371]
36. Eskridge JB, Wick B, Perrigin D. The Hirschberg test: a double-masked clinical evaluation. Am J Optom Physiol Opt 1988;65(9):745-750. [Medline: 3056019]
37. Hammel B, Sullivan-Molina N. bdhammel/least-squares-ellipse-fitting: initial release. Zenodo. 2019. URL: https://zenodo.org/records/2578663 [accessed 2019-02-27]
38. Halir R, Flusser J. Numerically stable direct least squares fitting of ellipses. 1998 Presented at: Proceedings of the 6th International Conference in Central Europe on Computer Graphics and Visualization; February 13, 1998; Pilsen (Plzen), Czech Republic p. 125-132.
39. Cestari DM, Hunter DG. Learning Strabismus Surgery: A Case-Based Approach. Ambler, PA: Lippincott Williams & Wilkins; 2012.
40. International Vocabulary of Metrology, Fourth edition—(VIM4 CD). Joint Committee for Guides in Metrology. 2021. URL: https://www.bipm.org/documents/20126/54295284/VIM4_CD_210111c.pdf [accessed 2025-10-28]
41. de Jongh E, Leach C, Tjon-Fo-Sang M, Bjerre A. Inter-examiner variability and agreement of the alternate prism cover test (APCT) measurements of strabismus performed by 4 examiners. Strabismus 2014;22(4):158-166. [doi: 10.3109/09273972.2014.972521] [Medline: 25360761]
42. Hatt SR, Leske DA, Liebermann L, Mohney BG, Holmes JM. Variability of angle of deviation measurements in children with intermittent exotropia. J AAPOS 2012;16(2):120-124 [FREE Full text] [doi: 10.1016/j.jaapos.2011.11.008] [Medline: 22525165]
43. Pediatric Eye Disease Investigator Group. Interobserver reliability of the prism and alternate cover test in children with esotropia. Arch Ophthalmol 2009;127(1):59-65 [FREE Full text] [doi: 10.1001/archophthalmol.2008.548] [Medline: 19139339]
44. Wong DS, Alsaif A, Bender L. The role of telemedicine in strabismus assessment: a narrative review and meta-analysis. Telemed J E Health 2024;30(8):e2240-e2255. [doi: 10.1089/tmj.2024.0115] [Medline: 38916770]
45. Cheng W, Lynn MH, Pundlik S, Almeida C, Luo G, Houston K. A smartphone ocular alignment measurement app in school screening for strabismus. BMC Ophthalmol 2021;21(1):150 [FREE Full text] [doi: 10.1186/s12886-021-01902-w] [Medline: 33765984]
46. Wang Y, Zhao A, Zhang X, Huang D, Zhu H, Sun Q, et al. Prevalence of strabismus among preschool children in eastern China and comparison at a 5-year interval: a population-based cross-sectional study. BMJ Open 2021;11(10):e055112 [FREE Full text] [doi: 10.1136/bmjopen-2021-055112] [Medline: 34667017]

Abbreviations

AC: alternating cover
AI: artificial intelligence

XSL•FO
RenderX

AWS: Amazon Web Services
BCVA: best-corrected visual acuity
LC: left eye cover-uncover
PD: prism diopter
RC: right eye cover-uncover
SAM: Segment Anything Model
YOLO: You Look Only Once

Edited by J Sarvestan; submitted 19.06.25; peer-reviewed by M Abegg, A Norouzpour; comments to author 08.08.25; revised version received 09.09.25; accepted 02.10.25; published 04.11.25.

Please cite as:
He J, Zhang J, Wang Z, Pundlik S, Liu R, Luo G
Quantitative Assessment of Strabismus Using Cloud AI Computing: Validation Study
JMIR Form Res 2025;9:e79280
URL: https://formative.jmir.org/2025/1/e79280
doi:10.2196/79280
PMID:

©Junxian He, Jiawei Zhang, Zheng Wang, Shrinivas Pundlik, Rui Liu, Gang Luo. Originally published in JMIR Formative Research (https://formative.jmir.org), 04.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Exploring Health Care Professionals' Perspectives on Using Video Feedback and Movement Analysis to Facilitate Physical Functioning for Older Adults Living at Home: Co-Design Approach

Jessica Olovsson[1], MSc; Mirjam Ekstedt[1,2], PhD; Cecilia Fagerström[1,3], PhD; Patrick Bergman[4], PhD; Sofia Backåberg[1], PhD

[1]Department of Health and Caring Sciences, Faculty of Health and Life Sciences, Linnaeus University, Växjö/Kalmar, Sweden

[2]Department of Learning, Informatics, Management and Ethics, Karolinska Institutet, Stockholm, Sweden

[3]Department of Research, Region Kalmar County, Kalmar, Sweden

[4]Department of Medicine and Optometry, Faculty of Health and Life Sciences, Linnaeus University, Kalmar, Sweden

Corresponding Author:
Jessica Olovsson, MSc
Department of Health and Caring Sciences
Faculty of Health and Life Sciences
Linnaeus University
Universitetsplatsen 1 Växjö
Växjö/Kalmar, 35195
Sweden
Phone: 46 470767003
Email: jessica.olovsson@lnu.se

Abstract

Background: Maintaining and motivating physical functioning among older adults has substantial health-related benefits, such as reducing the risk of falls and increasing the opportunities for independent living. Supporting preventive actions among older adults also has socioeconomic relevance. Previous studies have shown that digital tools involving video feedback can facilitate reflection and learning by encouraging active engagement.

Objective: This study aimed to explore health care professionals' experiences of using a video-based tool as part of the rehabilitation to facilitate physical functioning among older adults (aged ≥65 years) living at home.

Methods: An experience-based co-design approach was used, involving 20 health care professionals. Nine iterative workshops were conducted, followed by 9 group interviews held between 2022 and 2023. The data were analyzed using reflexive thematic analysis.

Results: The results from this study captured the experiences of health care professionals using a video-based tool to facilitate physical functioning in older adults living at home. The participants described focusing on supporting patient commitment, creating a shared language to enhance collaboration in the rehabilitation process, and navigating barriers to adopting the video-based tool in practice.

Conclusions: From the perspective of health care professionals, video feedback has the potential to improve movement performance in daily activities and may play a crucial role in providing motivation and promoting sustainable physical functioning among older adults. Clinical recommendations include training health care professionals to introduce video feedback in a patient-centered manner and using it to foster shared communication that promotes professional development and patient engagement. Further research is needed to assess the impact of video feedback on older adults' health outcomes and to identify strategies for implementation in complex rehabilitation needs.

(JMIR Form Res 2025;9:e73527) doi:10.2196/73527

KEYWORDS

physical functioning; eHealth; older adults; motivation; video feedback; rehabilitation

XSL•FO
RenderX

Introduction

As the proportion of older adults in the population increases, there is a growing need for interventions that effectively support their physical functioning [1]. Physical functioning is significantly associated with an individual's capacity to engage in activities of daily living. This becomes particularly important in older age, when maintaining adequate levels of physical functioning is vital for ensuring independence [2,3]. Remaining physically active despite advancing age is crucial for sustaining physical and mental health and has a preventive effect on major health conditions such as coronary artery diseases [4,5]. Regular physical activity can help prevent falls, which may otherwise lead to severe physical injuries such as fractures. This in turn may lead to fear of falling or social isolation. Early support from health care professionals, for example, rehabilitation staff, to regain physical function is therefore particularly important [6,7]. According to the World Health Organization, adults aged 65 years and above should complement their daily activities with individualized interventions that enhance balance, mobility, and strength [5,8]. Today, a large proportion of older adults reside at home. Older adults who receive home care report various barriers to physical activity, such as illness, a perception of being too old, and a fear of falling [9,10]. As people age, levels and patterns of physical activity often shift. Few people aged 65 years and above achieve the recommended levels of physical activity. According to Swedish national data, older adults spend approximately 10 hours per day sitting [11]. Relatives and close friends are essential in supporting older adults to maintain independence and physical function. In parallel, health care professionals in multidisciplinary teams play a crucial role in promoting these outcomes by providing specialized expertise and support [12]. Integrating physical activity into everyday life so that it becomes part of the daily routine is crucial for maintaining motivation [13,14].

Previous studies support the use of eHealth interventions to increase physical functioning [15,16]. Although eHealth interventions have been found to have positive outcomes and effects on physical activity among older adults, a gap in knowledge remains regarding their link to long-term health and behavioral changes [16,17]. Success factors for eHealth interventions that promote physical activity include support from health care professionals, self-monitoring, and social networking. Obstacles to using eHealth in health care include limited resources, the absence of organizational strategies, and a lack of knowledge [17,18]. This underlines the importance of acknowledging health care professionals' perspectives when developing and implementing eHealth interventions. Previous research shows that the motivation to learn new skills and remain active may be supported by interactive communication with health professionals [19]. Although both health care professionals and patients value eHealth in rehabilitation, their differing priorities (clinical functionality for professionals and social connectedness for patients) underscore the importance of designing solutions that address both emotional and functional motivation [20,21]. In a study by Kraaijkamp et al [22], health care professionals were found to perceive eHealth as being more complex for patients than for themselves. Therefore, tailored implementation strategies for eHealth interventions are required, with health care professionals' perspectives being essential to identifying and overcoming barriers. Even if eHealth interventions prove to be useful, implementation of work processes in health care that meet the patients' needs for support is challenging [23].

Research conducted in contexts outside of rehabilitation for older adults has demonstrated that video feedback can facilitate self-reflection by promoting active engagement and drawing attention to various aspects of movement. When combined with self-controlled feedback, where learners actively participate in the feedback process, video feedback tools have shown promising outcomes among younger participants [24-26]. Additionally, video feedback can enhance understanding of self-care management and serve as a valuable complement to traditional health care services [21]. However, despite these promising findings, the integration of interactive video feedback into rehabilitation at home remains rare. Furthermore, there is a notable gap in research examining the use of video feedback within rehabilitation settings for older adults, particularly from the perspective of health care professionals. Their insights are crucial for understanding practical implementation, perceived value, and potential barriers in clinical practice. This study aimed to explore health care professionals' experiences of using a video-based tool as part of rehabilitation to facilitate physical functioning among older adults living at home.

Methods

Study Design

The methodological approach was guided by the Medical Research Council's framework for developing and evaluating complex interventions [27]. The study adopted a participatory research design to engage stakeholders in developing the intervention. An experience-based co-design approach was applied [28], involving iterative workshops and group interviews with elements of think-aloud techniques [29]. The video feedback intervention was initially developed based on theoretical foundations and previous research. Following the Medical Research Council's framework, the primary focus of this study was to refine the intervention and retest the tentative program theory (Figure 1), identify key uncertainties, and adapt it to a new context (ie, rehabilitation at home for older adults). To address these steps, key stakeholders were engaged as co-design collaborators, including health care professionals, for example, physiotherapists, occupational therapists, and rehabilitation assistants involved in rehabilitation for older adults aged ≥65 years. The study was conducted as part of the larger research project *Confident in every step*, which was guided by a tentative program theory based on self-determination theory (SDT) [30]. The preliminary program theory includes (1) the patient and their resources (eg, knowledge, skills, and social connections); (2) the health care provider and their resources; (3) a digital tool enabling access to supportive information, learning, and feedback; and (4) the interactions, intentions, and capabilities of all parties to achieve health outcomes valued by older adults and supported by health care providers. These interactions are illustrated in Figure 1.

XSL•FO
RenderX

Figure 1. The preliminary program theory.

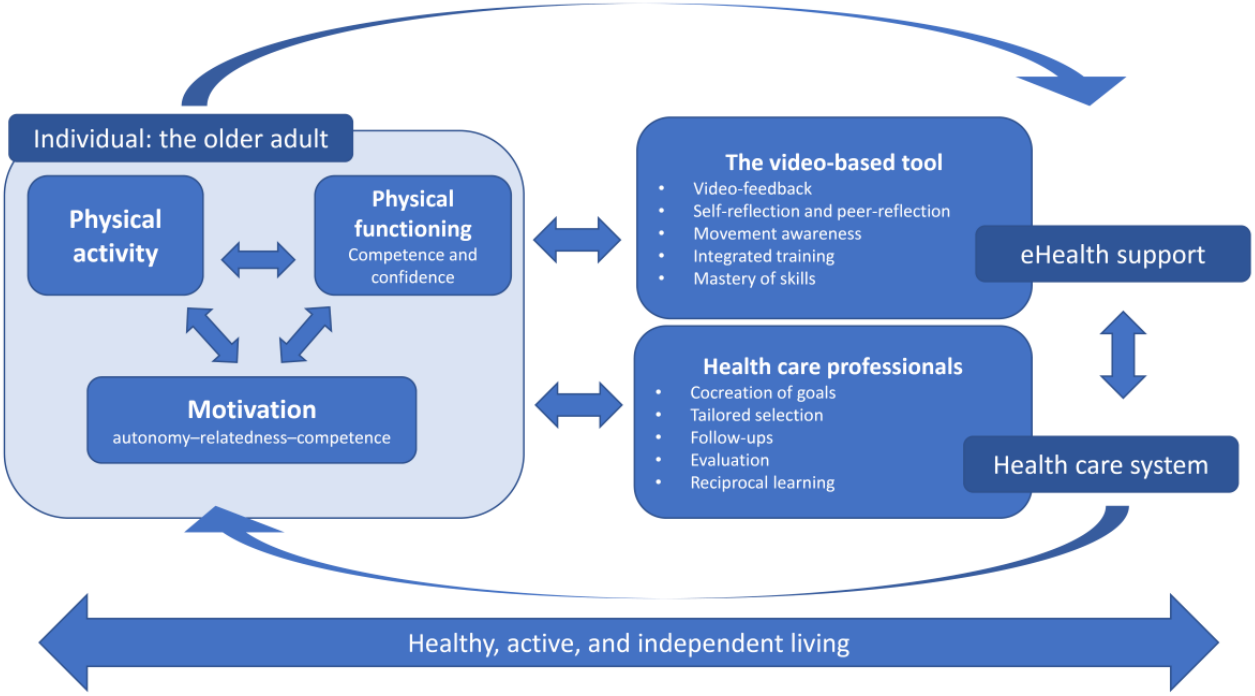

The Video-Based Tool

The video-based tool MOVEImprove (Savvy Knowledge Corporation) was used to support video feedback, structured analysis, and self-reflection on movement performance through guiding questions for each movement component. The platform was developed by researchers at the University of Calgary, and contents were developed by researchers at Linnaeus University. The video-based tool is compatible with mobile phones and tablets and allows users to record each other with the aim of improving physical skills. Assessments and reflections take place in pairs. In this study, health care professionals recorded patients performing exercises from 4 themes targeted at older adults: *balance for older adults*, *daily life movements*, *breathing exercises*, and *picking up from the floor*. The latter 2 themes were co-designed by researchers at Linnaeus University in earlier phases of the larger research project [21,31].

Data Collection

A total of 20 health care professionals from 3 municipal and regional primary care organizations in Sweden were recruited through convenience sampling. Participants involved in rehabilitation for older adults (Table 1) were invited face-to-face by representatives within their respective organizations.

In total, 9 workshops were conducted with 6 to 10 participants in each session; 2 were held digitally, whereas the rest took place at the participants' workplaces. Each organization participated in recurring workshops, followed by group interviews. Group interviews were semistructured and based on workshop outcomes. Co-design workshops and group interviews were conducted between 2022 and 2023, with 2 to 5 sessions per organization (Table 2).

XSL•FO
RenderX

Table 1. Demographic characteristics of participants (N=20).

Characteristic	Participants
Age (years), range	23-61
Gender, n (%)	
Women	18 (90)
Men	2 (10)
Professional experience (years), range	1-32
Professions, n (%)	
District nurse	2 (10)
Physiotherapist	9 (45)
Occupational therapist	4 (20)
Rehabilitation assistant	2 (10)
Medical officer for rehabilitation	2 (10)
Home care manager	1 (5)

Table 2. Data collection: workshops and subsequent group interviews.

Workshop (organization and format)	Participants by profession, n
Workshop 1 (organization A, in person)	1 MAR[a], 1 RA[b], 1 OT[c], and 2 PTs[d]
Workshop 2 (organization A, online)	1 MAR, 1 RA, 1 OT, and 2 PTs
Workshop 1 (organization B, in person)	1 OT and 4 PTs
Workshop 2 (organization B, online)	3 PTs
Workshop 1 (organization C, in person)	1 MAR, 2 OTs, 1 PT, and 1 HM[e]
Workshop 2 (organization C, in person)	1 OT, 2 PTs, 1 RA, 1 DS[f], and 1 HM
Workshop 3 (organization C, in person)	1 MAR, 2 OTs, 3 PTs, 1 RA, and 2 DSs
Workshop 4 (organization C, in person)	1 MAR, 2 OTs, and 2 PTs
Workshop 5 (organization C, in person)	1 MAR, 2 OTs, and 3 PTs

[a]MAR: medical officer for rehabilitation.
[b]RA: rehabilitation assistant.
[c]OT: occupational therapist.
[d]PT: physiotherapist.
[e]HM: home care manager.
[f]DS: district nurse.

The first author (JO), together with the last author (SB), facilitated all the workshops and group interviews. Each session, including interviews, lasted 60 to 180 minutes. All sessions were digitally recorded, and reflective notes taken by the research team supported and validated the data during analysis. During the iterative workshops, participants tested the video-based tool with each other and reflected on the usability of the tool, including reflections considering using it in a new context, that is, rehabilitation for older adults. Between sessions, health care professionals were provided with tablets to use the video-based tool with patients. Patient selection was based on the following inclusion criteria: aged 65 years or more, living at home, and receiving primary care or a newly implemented home care service.

Decisions were made in line with the Medical Research Council's framework by the research team and engaged stakeholders to review the progress and determine the next steps for the workshops and interviews. The co-design process resulted in a refined intervention, that is, a structured user guide for the video-based tool, including refined inclusion criteria of the patient population and a summary of identified areas of use. Through co-design with health care professionals, additional functionalities were introduced, encompassing comparative video analysis and slow-motion playback to improve the precision and clarity of movement execution when using the video-based tool.

Data Analysis

Workshop and interview recordings were transcribed and verified by the first author (JO). Data were analyzed using the reflexive thematic analysis by Braun and Clarke [32,33], following an inductive approach. Transcripts were organized

XSL•FO
RenderX

in NVivo (QSR International) to structure the data and support data management and traceability, whereas the thematic analysis was conducted manually in a step-by-step manner. Reflexive thematic analysis involved 6 steps. In step 1, the first author familiarized herself with the data through repeated readings. In step 2, initial codes were generated by the first author using open coding. In step 3, revised codes were collected. Theme development was discussed with coauthors (SB, ME, and CF), and subthemes were collaboratively refined into initial themes by the full research team (JO, SB, ME, CF, and PB). In step 4, overarching themes were reviewed and further discussed. The process involved iterative movement between quotations, coded segments, and the full dataset to ensure coherence and contextual grounding. Themes and subthemes were subsequently reviewed and refined to facilitate reflexive engagement with the data and to ensure that quotations had coherence with the original data. Step 5 involved defining and naming themes. Examples of the analysis process are presented in Table 3. Step 6 involved writing a coherent text that presents themes, including interpretation of the data [32,33].

Table 3. Examples of the reflective thematic analysis process.

Data extract	Code	Subtheme	Theme
"Well, it's been an eye-opener, because I might have felt that no, I don't want to use an iPad and such, because I felt that I was maybe a bit afraid that digitalization would distance me from the patient...or like my feeling from the start when we've talked about needing to digitalize more and so on, that it would mean I'm here with my screen and the patient is at home with their screen." [Participant 13]	Concern among health care providers that digital tools create distance	Navigating thresholds to adopting the tool	A new way of working
"I think where we filmed was more to build her confidence, or whatever you want to call it, to see that she actually managed it. There weren't many days in between, but there was still a clear improvement, and she got to see 'Well, I can actually do it.'" [Participant 19]	Film to build self-confidence	Clarifying sustainable movements and changes over time	Supporting patient commitment

Ethical Considerations

This study adhered to the principles of the Declaration of Helsinki [34] and was approved by the Swedish Ethical Review Authority (2023-00276-01). Participation was voluntary, and participants were informed that they could withdraw at any time without providing a reason. All participants received both verbal and written information about the study, including assurances that participation would not affect their current work situation or rehabilitation process. Written and verbal informed consent was obtained from all participants prior to data collection. Participants were made aware of the study's purpose, procedures, and their rights. Videos recorded in the clinic that included patients were deleted immediately by the responsible health care professionals to ensure privacy. All data were handled confidentially, and no identifiable information was retained or published. Participants did not receive any financial or material compensation for their participation. Data collection was conducted with attention to diversity in gender and cultural representation to ensure inclusivity and relevance.

Results

The analysis resulted in 3 overarching themes that described health care professionals' experiences of using a video-based tool to facilitate physical functioning in older adults within municipal and regional primary health care settings (Textbox 1 [32,33]).

Textbox 1. Overview of the 3 themes and 6 subthemes identified through the reflexive thematic analysis, as described by Braun and Clarke.

Supporting patient commitment

- Visualizing movement patterns and movement quality
- Clarifying sustainable movements and changes over time
- Enabling reflection and body awareness

Creating a shared language

- Facilitating collaboration with the patient
- Cocreation in the rehabilitation process

Navigating thresholds to adopting the tool

- Identifying the right patient
- Unfamiliarity: a new way of working

Supporting Patient Commitment

The video-based tool was perceived as a clinical pedagogical aid that enhanced patient commitment. Health care professionals described the video-based tool as a valuable method for illustrating movement quality to patients in a clear and accessible manner. Visual feedback enabled patients to observe their own movement patterns, fostering reflection and self-assessment.

XSL•FO
RenderX

However, responses varied: some patients were overly generous in their evaluations, while others were highly self-critical, highlighting the subjective nature of movement perception. The video-based tool was perceived as a significant advancement over traditional methods such as mirror-based observation, as the video feedback allowed for retrospective analysis and more objective evaluation of movement patterns. This facilitated detailed discussions between professionals and patients. It also allowed for objective measurements and the ability to review movements retrospectively. Therefore, it could serve as a potential evaluation resource, particularly useful for "before-and-after" comparisons in areas such as mobility, balance, and movement quality. Health care professionals described the video-based tool as a valuable instrument for documenting progress throughout the rehabilitation process. By enabling movement performance over time, it allowed patients to see how close they were to achieving their rehabilitation goals. This functionality supported professionals in tracking changes and making informed decisions about ongoing care. A participant reflected as follows:

> *"Well, I'm in a lot of pain, it doesn't work, I still can't walk." No, but a week ago you could only sit up for 15 seconds and now you're sitting here without holding on to anything. So, it can be an opportunity for them to see that for themselves.* [Participant 1]

By providing visual evidence of functional improvement, the tool contributed to strengthening patients' confidence in their physical capabilities:

> *I think where we filmed was more to build her confidence, or whatever you want to call it, to see that she actually managed it. There weren't many days in between, but there was still a clear improvement, and she got to see "Well, I can actually do it."* [Participant 19]

Health care professionals reported that the use of the video-based tool contributed to patients feeling acknowledged within the rehabilitation process. This sense of recognition was linked to the commitment demonstrated by professionals who actively engaged with the tool to support individualized care. The visual feedback provided by the tool was perceived as a means of affirming each patient's efforts and progress, thereby strengthening the therapeutic relationship. Some rehabilitation professionals were surprised by the potential of video technology in clinical practice, particularly in its ability to enhance patient engagement and interprofessional collaboration. However, they also emphasized the importance of identifying key facilitators to ensure successful implementation. These included effective time management, collaborative work in pairs, and strategic planning within the clinical team to address organizational challenges and sustainably integrate the tool into everyday practice. Health care professionals described the video-based tool as a valuable means of fostering patient reflection and enhancing body awareness. By enabling repeated observation and discussion of movement performance, the tool supported a deeper understanding of functional abilities and limitations. The visual feedback facilitated detailed reflection, allowing patients to recognize subtle aspects of their movements that might otherwise go unnoticed. Health care professionals noted that viewing recorded movements could evoke memories of previous physical capabilities, particularly in individuals recovering from injury. This process was believed to be instrumental in reconnecting patients with familiar movement patterns, thereby supporting the restoration of motor function. The tool also provided opportunities to revisit and reinforce therapeutic practices from earlier stages of life or previous rehabilitation experiences. A participant described recollecting an encounter with a patient, who stated,

> *"Well, this is exactly how I've done it before." And then she started. "This is how I used to exercise," and it felt like she found her muscle memory again.* [Participant 11]

Through this reflective process, the video-based tool contributed to active reflection and engagement in rehabilitation. Although the video-based tool was generally perceived as a valuable aid in enhancing body awareness and movement reflection, health care professionals also noted that its use could inadvertently lead to altered movement performance. The act of being filmed heightened patients' self-consciousness, prompting some to modify their movements in ways that were not representative of their habitual motor patterns. This performative response—driven by a desire to appear more competent or polished on video—was seen as a potential limitation, as it could decrease the authenticity of the movement performance.

Creating a Shared Language

Health care professionals found the video-based tool helpful in supporting learning and documenting patient information, particularly when multiple people were involved. This was evident in various contexts, including the patient-rehabilitation professional relationship. Health care professionals emphasized the importance of mutual understanding between professionals and patients in the rehabilitation process. The video-based tool was perceived as a valuable aid in fostering this collaboration, enabling clinicians to visually demonstrate specific movements, exercises, and progress. By showing tangible improvements, professionals could enhance patient collaboration. A participant shared the following example from clinical practice:

> *Well, look, this is how it was three weeks ago; now you [the patient] can actually do this, and you [the patient] can do it yourself. Now the caring staff only needs to stand by.* [Participant 12]

This visual feedback supported shared decision-making and strengthened the therapeutic alliance, allowing patients to take a more active role in their rehabilitation. The video-based tool was described as a facilitator of cocreation within the rehabilitation process, supporting professional collaboration. Health care professionals emphasized its value for internal learning and consensus building among rehabilitation staff, helping teams establish a shared understanding of patient progress and therapeutic strategies. The tool also served as a resource for external caregivers, such as home care staff, by visually demonstrating appropriate levels of assistance. This was seen as particularly important in preventing overassistance, which could inadvertently limit patient autonomy and rehabilitation outcomes. Observing how exercises were performed across different settings, whether in health care

XSL•FO
RenderX

facilities or at home, was considered especially beneficial during handovers. However, health care professionals noted that incompatible journal systems across organizations posed a barrier to seamless information sharing. The ability to store a digital history of movement performance was viewed as a key advantage, enabling professionals to track progress and support individualized care planning. A participant explained the following:

> *It's quite easy for a colleague to take over when there's a digital history, as they can look at the first and last videos and see what they've been working on, and what we should continue with if needed.* [Participant 1]

Navigating Thresholds to Adopting the Tool

Rehabilitation professionals encountered various challenges when integrating the video-based tool into clinical practice and described this as a process of overcoming obstacles. Establishing a therapeutic relationship was described as a key facilitator for the effective use of the video-based tool. As such, its implementation was typically deferred until after the initial patient contact, allowing time to build trust and assess suitability. Health care professionals identified that the tool was most beneficial for patients who could both articulate and reflect on their movement difficulties—an ability considered uncommon in older adult care settings. Cognitive capacity was also noted as an important factor, though views on this varied. Although some professionals regarded cognitive impairment as a barrier, others found that it did not necessarily preclude meaningful engagement with the tool. The video-based tool was particularly valued in cases involving frail older adults with limited energy, where repeated movement demonstrations were challenging. In such contexts, the ability to record and review performance across sessions enabled more detailed analysis and individualized care planning. The integration of video-based tools into rehabilitation practice was described as a shift that was novel and, at times, challenging. Health care professionals expressed concerns regarding data security and the ethical management of visual recordings within medical documentation. In addition, some professionals feared that digitalization might compromise the therapeutic relationship by creating an emotional or physical distance between clinician and patient. A participant reflected as follows:

> *Well, it's been an eye-opener, because I might have felt that no, I don't want to use an iPad and such, because I felt that I was maybe a bit afraid that digitalization would distance me from the patient...or like my feeling from the start when we've talked about needing to digitalize more and so on, that it would mean I'm here with my screen and the patient is at home with their screen.* [Participant 13]

Initial use of the tool also prompted safety-related adaptations. Health care professionals reported a need to remain physically close to patients or work in pairs, particularly during early sessions, to mitigate risks such as falls and to deal with the unfamiliarity of the method. Preconceived notions about older adults, such as discomfort with being filmed or limited digital literacy, were common but often challenged through practice. These findings highlight the importance of balancing innovation with relational and safety considerations in clinical implementation.

Discussion

Principal Findings

This study explored health care professionals' experiences with a video-based tool designed to facilitate physical functioning in older adults living at home. Participants highlighted its role in fostering patient engagement, facilitating a shared language in collaborative rehabilitation. The video-based tool was perceived as a motivational source that could enhance patients' confidence in movement. It was also perceived to improve interprofessional communication, particularly during handovers and collaborative planning, by demonstrating physical functioning and rehabilitation needs. Reflecting on video recordings enabled shared understanding of rehabilitation goals. The study's findings align well with SDT, as described by Ryan and Deci [30], which emphasizes the importance of 3 basic psychological needs in fostering intrinsic motivation and well-being: autonomy, competence, and relatedness. *Autonomy* is supported when individuals feel they are the origin of their actions. The video-based tool empowered older adults to engage with their rehabilitation in a way that felt meaningful. It also allowed older adults to observe and reflect on their own movements, giving them a sense of control and ownership over their rehabilitation. Seeing their own progress visually could help older adults set personal goals in the rehabilitation process. *Competence* is fostered when individuals feel capable of achieving desired outcomes. Participants noted that the video-based tool allowed patients to observe incremental improvements, providing clear evidence of progress. Capturing positive aspects of performance helped reinforce a sense of achievement, even in small steps. Conversely, when the participants did not achieve the desired outcomes, video feedback failed to enhance their sense of competence and may have adversely affected their perceived capability. Therefore, visual feedback could be especially powerful for older adults who struggle with self-efficacy due to age-related decline. *Relatedness* is fulfilled when individuals feel connected to others. Shared use of the video-based tool fostered a sense of partnership in the rehabilitation process, highlighting the importance of involving health care professionals in using video feedback with older adults. Reflecting on video recordings together could foster mutual understanding. The video-based tool also facilitated interprofessional communication, which supports collaborative rehabilitation. The findings from this study support the idea that use of the video-based tool in close collaboration with health care professionals can enhance intrinsic motivation in older adults by satisfying the core psychological needs outlined in the SDT. The findings also support the larger research project's preliminary program theory (Figure 1), suggesting that video feedback and movement analysis can promote sustained physical functioning and physical activity, which in turn promotes active and healthy aging. Furthermore, the results provide a deeper understanding of the preliminary program theory, offering potential for its refinement and expansion, particularly in relation to how video feedback can

XSL•FO
RenderX

support psychological aspects of rehabilitation, such as confidence and motivation.

Comparison With Previous Work

Video-based feedback has proven especially valuable for older adults by highlighting incremental improvements and reinforcing positive aspects of movement. Previous research indicates that aging is often accompanied by psychological stress, making visual confirmation of physical performance especially impactful and motivating [4,35,36]. The findings of this study highlight the importance of motivation in rehabilitation, particularly when progress is made visible through structured feedback [21,37]. Health care professionals emphasized the need to identify appropriate target groups for the tool. As more patients were included and positive outcomes were observed, the perceived applicability of the tool broadened. Although cognitive ability was considered relevant, cognitive impairment was not viewed as a barrier, which was seen as an encouraging stance. This contrasts with the frequent exclusion of cognitively impaired older adults from research, limiting the development of inclusive, evidence-based interventions [38]. By involving this population, the study contributes to more equitable rehabilitation research. eHealth interventions offer potential for promoting healthy aging but require thoughtful implementation [31,39]. The video-based tool showed promise for early detection of functional decline through structured movement analysis; however, the mechanisms underlying this effect and its comparative effectiveness remain unclear. Preventive interventions for community-dwelling older adults have yielded positive health outcomes [40]. The video-based tool that was refined in our study may contribute meaningfully to preventive strategies and support such efforts. However, challenges remain in implementing digital tools in rehabilitation, including concerns about data security, reimbursement, legal frameworks, and digital literacy among older adults [41]. Previous studies show that although physiotherapists value evidence-based digital tools for improving access and self-management, they also express concerns about reduced interaction and patient adherence. Successful implementation requires clear guidelines, training, and a user-friendly design [42]. While our study acknowledged similar findings, further investigation is warranted to explore the diversity and potential contradictions within these perspectives in the context of rehabilitation for older adults.

Strengths and Limitations

Credibility was enhanced through the inclusion of diverse samples from multiple settings, as elaborated in the criteria for qualitative research by Lincoln and Guba [43]. Workshops and interviews were conducted by professionals with different clinical backgrounds (physiotherapy and nursing). This diversity was expected to provide a range of perspectives. All interviews were transcribed verbatim and verified by the first author, with quotations used to support findings. Codes, subthemes, and themes were reviewed iteratively to enhance confirmability, including discussions of conceptual depth. Dependability was supported by consistent use of a semistructured interview guide and a single transcriber. The authors maintained reflective journals to document insights and ensure analytic consistency over time.

Transferability was addressed by clearly describing participants, data collection methods, and timing, allowing findings to be applied in similar rehabilitation contexts. Open-ended questions, think-aloud techniques, and repeated workshops promoted participant engagement. However, group interviews may have limited dissenting views due to hierarchical dynamics, although no participants were in dependent relationships. Interviews with multiprofessional groups may have influenced responses, which should be considered when interpreting the results. The study encompassed 3 workplaces, with varying numbers of workshops and intervals, which may have affected the depth of engagement with the video-based tool. Although most sessions were conducted in person, 2 were held digitally at participants' request, potentially limiting interactions with the research team.

Conclusions

From the perspective of health care professionals, experiences of using video feedback for older adults to support patient commitment, foster a shared language, and navigate thresholds for adopting the video-based tool into practice have shown potential to motivate enhanced movement performance and could therefore play an important role in sustainably facilitating physical functioning. However, its implementation requires considerations of ethical data management and potential influence on the therapeutic relationship. For patients, it may offer a way to engage more actively in their rehabilitation. Responses to being filmed ranged from improved confidence to altered movements, revealing the subjective nature of movement perception. Therefore, clinical recommendations include training health care professionals to introduce video feedback in a manner that is patient centered, emphasizing its role in supporting motivation rather than serving solely as an evaluative tool. Furthermore, video feedback should be used to establish a shared communicative framework that enhances professional development and facilitates constructive dialogue with patients regarding their rehabilitation progress. These insights offer a meaningful contribution to the preliminary tentative program theory, indicating that the mechanisms proposed, such as self-observation, reflection, and motivation, may be effective in rehabilitation for older adults. However, further research is warranted to investigate the effect and impact of video feedback on older adults' physical abilities and related health outcomes. Future studies should be conducted particularly in settings where older adults often face significant challenges in maintaining physical function and independence, such as during care transitions or when presenting with complex rehabilitation needs.

Acknowledgments

The authors extend their gratitude to the participating establishments for their valuable contributions. This study was funded by the Swedish Research Council for Health, Working Life, and Welfare (2020-01557) and the Swedish Foundation for International

XSL•FO
RenderX

Cooperation in Research and Higher Education (MG2020-8851). The study's design; data collection, analysis, and interpretation; manuscript writing; and the decision to submit for publication were carried out independently, without any influence from the funding agencies.

Data Availability

The data generated or analyzed during this study are not publicly available as participants did not consent to public availability. Anonymized data are available from the corresponding author on reasonable request (aggregated data in Swedish).

Authors' Contributions

JO contributed to conceptualization, data curation, formal analysis, investigation, validation, visualization, and writing—original draft. SB, ME, CF, and PB contributed to conceptualization, formal analysis, validation, writing—review and editing, and supervision, with SB also managing project administration and assisting with investigation. SB, ME, and CF contributed to funding acquisition.

Conflicts of Interest

None declared.

References

1. World report on ageing and health. World Health Organization. 2015 Sep 29. URL: https://www.who.int/publications/i/item/9789241565042 [accessed 2025-10-31]
2. Wang DX, Yao J, Zirek Y, Reijnierse EM, Maier AB. Muscle mass, strength, and physical performance predicting activities of daily living: a meta-analysis. J Cachexia Sarcopenia Muscle 2020 Feb;11(1):3-25 [FREE Full text] [doi: 10.1002/jcsm.12502] [Medline: 31788969]
3. Knapik A, Brzęk A, Famuła-Wąż A, Gallert-Kopyto W, Szydłak D, Marcisz C, et al. The relationship between physical fitness and health self-assessment in elderly. Medicine (Baltimore) 2019 Jun;98(25):e15984 [FREE Full text] [doi: 10.1097/MD.0000000000015984] [Medline: 31232930]
4. Cunningham C, O' Sullivan R, Caserotti P, Tully MA. Consequences of physical inactivity in older adults: a systematic review of reviews and meta-analyses. Scand J Med Sci Sports 2020 May;30(5):816-827. [doi: 10.1111/sms.13616] [Medline: 32020713]
5. WHO Guidelines on Physical Activity and Sedentary Behaviour. Geneva, Switzerland: World Health Organization; 2020.
6. Kumar A, Delbaere K, Zijlstra GA, Carpenter H, Iliffe S, Masud T, et al. Exercise for reducing fear of falling in older people living in the community: Cochrane systematic review and meta-analysis. Age Ageing 2016 May;45(3):345-352. [doi: 10.1093/ageing/afw036] [Medline: 27121683]
7. Sherrington C, Fairhall NJ, Wallbank GK, Tiedemann A, Michaleff ZA, Howard K, et al. Exercise for preventing falls in older people living in the community. Cochrane Database Syst Rev 2019 Jan 31;1(1):CD012424 [FREE Full text] [doi: 10.1002/14651858.CD012424.pub2] [Medline: 30703272]
8. Sherrington C, Fairhall N, Kwok W, Wallbank G, Tiedemann A, Michaleff ZA, et al. Evidence on physical activity and falls prevention for people aged 65+ years: systematic review to inform the WHO guidelines on physical activity and sedentary behaviour. Int J Behav Nutr Phys Act 2020 Nov 26;17(1):144 [FREE Full text] [doi: 10.1186/s12966-020-01041-3] [Medline: 33239019]
9. Chastin S, Gardiner PA, Harvey JA, Leask CF, Jerez-Roig J, Rosenberg D, et al. Interventions for reducing sedentary behaviour in community-dwelling older adults. Cochrane Database Syst Rev 2021 Jun 25;6(6):CD012784 [FREE Full text] [doi: 10.1002/14651858.CD012784.pub2] [Medline: 34169503]
10. Olanrewaju O, Kelly S, Cowan A, Brayne C, Lafortune L. Physical activity in community dwelling older people: a systematic review of reviews of interventions and context. PLoS One 2016;11(12):e0168614 [FREE Full text] [doi: 10.1371/journal.pone.0168614] [Medline: 27997604]
11. Niklasson J, Fagerström C, Backåberg S, Lindberg T, Bergman P. Daily activity patterns in older adults receiving initial support: the association between daily steps and sitting in bouts of at least 60 min. BMC Geriatr 2024 Jan 23;24(1):88 [FREE Full text] [doi: 10.1186/s12877-024-04681-3] [Medline: 38263077]
12. Mjøsund HL, Moe CF, Burton E, Uhrenfeldt L. Integration of physical activity in reablement for community dwelling older adults: a systematic scoping review. J Multidiscip Healthc 2020 Oct 29;13:1291-1315 [FREE Full text] [doi: 10.2147/JMDH.S270247] [Medline: 33154647]
13. Pollet S, Denison-Day J, Bradbury K, Essery R, Grey E, Western M, et al. A qualitative exploration of perceptions of a digital intervention to promote physical activity in older adults. J Aging Phys Act 2021 Jun 01;29(3):442-454 [FREE Full text] [doi: 10.1123/japa.2019-0484] [Medline: 33348316]

XSL•FO
RenderX

14. Pettersson B, Wiklund M, Janols R, Lindgren H, Lundin-Olsson L, Skelton DA, et al. 'Managing pieces of a personal puzzle' - older people's experiences of self-management falls prevention exercise guided by a digital program or a booklet. BMC Geriatr 2019 Feb 18;19(1):43 [FREE Full text] [doi: 10.1186/s12877-019-1063-9] [Medline: 30777026]
15. Mclaughlin M, Delaney T, Hall A, Byaruhanga J, Mackie P, Grady A, et al. Associations between digital health intervention engagement, physical activity, and sedentary behavior: systematic review and meta-analysis. J Med Internet Res 2021 Feb 19;23(2):e23180 [FREE Full text] [doi: 10.2196/23180] [Medline: 33605897]
16. Mönninghoff A, Kramer JN, Hess AJ, Ismailova K, Teepe GW, Tudor Car L, et al. Long-term effectiveness of mHealth physical activity interventions: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2021 Apr 30;23(4):e26699 [FREE Full text] [doi: 10.2196/26699] [Medline: 33811021]
17. Zaslavsky O, Roopsawang I, Chen AT. Promoting behavioral change in mobile health interventions for older adults: a scoping review. Res Gerontol Nurs 2020 Mar 01;13(2):102-116 [FREE Full text] [doi: 10.3928/19404921-20191031-01] [Medline: 31697395]
18. McGarrigle L, Todd C. Promotion of physical activity in older people using mHealth and eHealth technologies: rapid review of reviews. J Med Internet Res 2020 Dec 29;22(12):e22201 [FREE Full text] [doi: 10.2196/22201] [Medline: 33372894]
19. Albury C, Hall A, Syed A, Ziebland S, Stokoe E, Roberts N, et al. Communication practices for delivering health behaviour change conversations in primary care: a systematic review and thematic synthesis. BMC Fam Pract 2019 Aug 03;20(1):111 [FREE Full text] [doi: 10.1186/s12875-019-0992-x] [Medline: 31376830]
20. Ehn M, Johansson AC, Revenäs Å. Technology-based motivation support for seniors' physical activity-a qualitative study on seniors' and health care professionals' views. Int J Environ Res Public Health 2019 Jul 08;16(13):2418 [FREE Full text] [doi: 10.3390/ijerph16132418] [Medline: 31288398]
21. Strandberg S, Ekstedt M, Fagerström C, Backåberg S. Cocreation of a video feedback tool for managing self-care at home with pairs of older adults: remote experience-based co-design study. JMIR Form Res 2024 Oct 28;8:e57219 [FREE Full text] [doi: 10.2196/57219] [Medline: 39466305]
22. Kraaijkamp JJ, Persoon A, Aurelian S, Bachmann S, Cameron ID, Choukou MA, et al. eHealth in geriatric rehabilitation: an international survey of the experiences and needs of healthcare professionals. J Clin Med 2023 Jul 05;12(13):4504 [FREE Full text] [doi: 10.3390/jcm12134504] [Medline: 37445545]
23. Ekstedt M, Kirsebom M, Lindqvist G, Kneck Å, Frykholm O, Flink M, et al. Design and development of an eHealth service for collaborative self-management among older adults with chronic diseases: a theory-driven user-centered approach. Int J Environ Res Public Health 2021 Dec 30;19(1):391 [FREE Full text] [doi: 10.3390/ijerph19010391] [Medline: 35010652]
24. Backåberg S, Brunt D, Rask M, Gummesson C. Experiences of using a video-based learning model during a long-term process of movement awareness and learning – a hermeneutical study. Eur J Physiother 2019 Jul 05;23(1):41-47. [doi: 10.1080/21679169.2019.1635639]
25. Milajerdi HR, Katz L. The impact of order of performance on peer-to-peer learning of a lacrosse ball pick-up skill through video analysis with university students. In: Proceedings of the 9th International Performance Analysis Workshop and Conference & 5th IACSS Conference. 2022 Presented at: PACSS 2021; August 9-13, 2021; Virtual Event. [doi: 10.1007/978-3-030-99333-7_29]
26. Hillis TL, Milajerdi HR, Katz L. Using a structured interactive peer-to-peer video tool to facilitate learning in development level speed skaters. In: Proceedings of the 9th International Performance Analysis Workshop and Conference & 5th IACSS Conference. 2022 Presented at: PACSS 2021; August 9-13, 2021; Virtual Event. [doi: 10.1007/978-3-030-99333-7_23]
27. Skivington K, Matthews L, Simpson SA, Craig P, Baird J, Blazeby JM, et al. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. Int J Nurs Stud 2024 Jun;154:104705. [doi: 10.1016/j.ijnurstu.2024.104705] [Medline: 38564982]
28. Bate P, Robert G. Experience-based design: from redesigning the system around the patient to co-designing services with the patient. Qual Saf Health Care 2006 Oct;15(5):307-310 [FREE Full text] [doi: 10.1136/qshc.2005.016527] [Medline: 17074863]
29. Beatty PC, Willis GB. Research synthesis: the practice of cognitive interviewing. Public Opin Q 2007;71(2):287-311 [FREE Full text] [doi: 10.1093/poq/nfm006]
30. Ryan RM, Deci EL. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. Am Psychol 2000 Jan;55(1):68-78. [doi: 10.1037//0003-066x.55.1.68] [Medline: 11392867]
31. Backåberg S, Strandberg S, Freeman G, Katz L, Rafiei Milajerdi H, Wylant B, et al. Facilitating co-design among older adults in a digital setting: methodological challenges and opportunities. CoDesign 2024 Jul 09;21(1):118-135 [FREE Full text] [doi: 10.1080/15710882.2024.2372595]
32. Braun V, Clarke V. Thematic Analysis: A Practical Guide. Thousand Oaks, CA: SAGE Publications; 2021.
33. Braun V, Clarke V. Can I use TA? Should I use TA? Should I not use TA? Comparing reflexive thematic analysis and other pattern-based qualitative analytic approaches. Counsel Psychother Res 2020 Oct 18;21(1):37-47 [FREE Full text] [doi: 10.1002/capr.12360]
34. World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA 2013 Nov 27;310(20):2191-2194. [doi: 10.1001/jama.2013.281053] [Medline: 24141714]

XSL•FO
RenderX

35. Bergman YS, Bodner E. Aging anxiety in older adults: the role of self-esteem and meaning in life. GeroPsych 2022;35(4):196-201 [FREE Full text] [doi: 10.1024/1662-9647/a000295]
36. López-Novis I, Marques-Sule E, Deka P, Dobarrio-Sanz I, Klompstra L, Hernández-Padilla JM. Exploring physical activity experiences of community-dwelling oldest-old adults with chronic multimorbidity: a qualitative study. J Adv Nurs 2025 Apr;81(4):1953-1965. [doi: 10.1111/jan.16403] [Medline: 39152567]
37. Meredith SJ, Cox NJ, Ibrahim K, Higson J, McNiff J, Mitchell S, et al. Factors that influence older adults' participation in physical activity: a systematic review of qualitative studies. Age Ageing 2023 Aug 01;52(8):afad145 [FREE Full text] [doi: 10.1093/ageing/afad145] [Medline: 37595070]
38. Swan L, Kochovska S, Ries N, Gilmore I, Parker D, Sinclair C, et al. Strategies to improve research participation by older people with cognitive impairment: a systematic review. Gerontologist 2025 May 10;65(6):gnae188. [doi: 10.1093/geront/gnae188] [Medline: 39693265]
39. Alruwaili MM, Shaban M, Elsayed Ramadan OM. Digital health interventions for promoting healthy aging: a systematic review of adoption patterns, efficacy, and user experience. Sustainability 2023 Dec 02;15(23):16503. [doi: 10.3390/su152316503]
40. Bajraktari S, Sandlund M, Zingmark M. Health-promoting and preventive interventions for community-dwelling older people published from inception to 2019: a scoping review to guide decision making in a Swedish municipality context. Arch Public Health 2020;78:97 [FREE Full text] [doi: 10.1186/s13690-020-00480-5] [Medline: 33072316]
41. Estel K, Scherer J, Dahl H, Wolber E, Forsat ND, Back DA. Potential of digitalization within physiotherapy: a comparative survey. BMC Health Serv Res 2022 Apr 13;22(1):496 [FREE Full text] [doi: 10.1186/s12913-022-07931-5] [Medline: 35418069]
42. Arkkukangas M, Cederbom S. Movement toward an evidence-based, digital fall prevention future-perceptions from a physiotherapy perspective. Physiother Theory Pract 2023 Jan;39(1):128-136. [doi: 10.1080/09593985.2021.2005198] [Medline: 34791975]
43. Lincoln YS, Guba EG. Naturalistic Inquiry. Thousand Oaks, CA: SAGE Publications; 1985.

Abbreviations

SDT: self-determination theory

Edited by J Sarvestan; submitted 06.03.25; peer-reviewed by M Sandlund; comments to author 23.07.25; accepted 24.10.25; published 06.11.25.

Please cite as:
Olovsson J, Ekstedt M, Fagerström C, Bergman P, Backåberg S
Exploring Health Care Professionals' Perspectives on Using Video Feedback and Movement Analysis to Facilitate Physical Functioning for Older Adults Living at Home: Co-Design Approach
JMIR Form Res 2025;9:e73527
URL: https://formative.jmir.org/2025/1/e73527
doi:10.2196/73527
PMID:

©Jessica Olovsson, Mirjam Ekstedt, Cecilia Fagerström, Patrick Bergman, Sofia Backåberg. Originally published in JMIR Formative Research (https://formative.jmir.org), 06.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Clinician Perspectives on Incorporating Physical Activity and Sleep Prescriptions Using eHealth for Youth With Comorbid Psychiatric Disorders: Qualitative Focus Group Study

Alyssa M Button[1,2], PhD; James Slavet[3], PhD; Amanda E Staiano[2], MPH, PhD; April B Bowling[4], SCD

[1]Virginia Commonwealth University, Richmond, VA, United States
[2]Pennington Biomedical Research Center, Baton Rouge, LA, United States
[3]Marblehead Family Therapy and Wellness, Marblehead, MA, United States
[4]Merrimack College, North Andover, MA, United States

Corresponding Author:
April B Bowling, SCD
Merrimack College
315 Turnpike Street
North Andover, MA, 01845
United States
Phone: 1 978 837 5187
Email: bowlinga@merrimack.edu

Abstract

Background: Physical activity and sleep prescriptions are indicated for the treatment of psychiatric disorders among youth. However, there is limited clinical adoption of these practices. Exergaming (ie, games that require physical activity) is a feasible intervention to promote physical activity and sleep hygiene and is appealing to youth given their interest in video gaming. Integrating exergaming prescriptions into clinical mental health practices may offer an opportunity to expand access to these interventions, yet pragmatic considerations for adopting these programs are poorly understood.

Objective: This study aimed to gain feedback from practicing clinicians on adopting GamerFit, an app-based intervention that incorporates exergames, step and sleep tracking, and online coaching to promote physical activity and sleep, as a tool in treatment plans for youth aged 13 to 17 years with psychiatric disorders.

Methods: Mental health clinicians participated in 2 online focus groups. A semistructured interview collected information on perceptions of the importance of physical activity and sleep, considerations for using GamerFit with clients, and approaches for incorporating GamerFit into standard care. Qualitative analysis included a hierarchical thematic coding system of isolated quotes, with the structure, frequency, and interrelationships of the coded quotes used for analysis.

Results: All clinicians (8/8, 100%) endorsed physical activity and sleep prescriptions as important interventions, although they were not typically a focus of treatment. Clinicians reported varying levels of self-efficacy in encouraging physical activity goals (6/8, 75%) and, to a lesser extent, sleep hygiene (4/8, 50%). Most perceived eHealth approaches positively (7/8, 88%) and noted their appeal given the accessibility of this physical activity option via gaming (2/4, 50%). Clinicians were optimistic about the feasibility of using GamerFit; the exergame and health coaching aspects of GamerFit were perceived favorably (5/8, 62%). Clinicians desired to access app data in electronic health systems to incorporate in therapeutic sessions (4/8, 50%) and recommended using the app in residential settings with continued use at home (2/8, 25%). Clinicians expressed concern regarding the implementation of GamerFit with families with low technology literacy, noting that some patients would likely require parental assistance to help with reminders and technology use (1/8, 12%). Suggestions for improvement included a greater variety of exergames and features to increase adolescents' engagement (6/8, 75%). There was a considerable willingness to incorporate this technology into clinicians' clinical practices and a strong desire for insurance provisions to cover coaching and technological components (7/8, 88%).

Conclusions: Clinicians perceived GamerFit as a feasible and acceptable clinical approach to physical activity and sleep prescriptions for youth with psychiatric disorders. The remote delivery of this intervention was perceived to be of interest to patients and provided helpful guidance for clinicians who were short on time to address many important topics within limited session time frames.

XSL•FO
RenderX

(JMIR Form Res 2025;9:e71569) doi:10.2196/71569

KEYWORDS

mental health; health coaching; mobile health; mHealth; attention-deficit/hyperactivity disorder; ADHD; depression; autism

Introduction

Physical activity in adolescence is well known to promote optimal physical and mental health outcomes both immediately and over the long term [1-5]. Despite the benefits, most youth in the United States do not meet recommendations set forth by the *Physical Activity Guidelines for Americans, 2nd Edition* [6,7]. Youth with psychiatric disorders, who represent 1 in 5 children and adolescents in the United States per year [8], fall even further behind [9], contributing to a disproportionate risk of chronic disease [10]. Importantly, physical activity and sleep are bidirectionally associated, where physical activity contributes to improved sleep quality and duration, and vice versa [11]. Sleep disturbance rates are notably higher among youth with psychiatric conditions and likely exacerbate mental health symptoms [12], highlighting the value of treatments that can address both low physical activity and poor sleep simultaneously.

Youth with psychiatric disorders face unique barriers that reduce the likelihood of participating in physical activity programming, such as social isolation and peer exclusion, dislike of competition, preference for solitary activity, motor delays, dysregulation, medication side effects, and negative experiences with previous programming [13]. For youth with physical, cognitive, or self-regulatory challenges, additional parent or caregiver support and action may be required for engagement in physical activity [14]. There is a critical need for health programming to reduce chronic disease disparities in this population through the adaptation and use of behavioral health programming that overcomes these reported barriers by providing pragmatic support to parents and youth in multiple settings [14].

Physical activity interventions, in conjunction with traditional mental health treatment (ie, psychotherapy), have been shown to improve symptoms related to psychiatric conditions such as autism spectrum disorder [15], depression [16], anxiety [17], and attention-deficit/hyperactivity disorder [18]. Improvements in sleep similarly benefit a wide range of cognitive, affective, and behavioral outcomes in youth with psychiatric disorders [19]. For these reasons, mental health treatment providers are encouraged to use exercise as medicine through behavioral prescriptions [20] and to promote positive sleep hygiene [21]. Despite evidence of the importance of physical activity and sleep in mental health treatment, there are mixed reports of clinician uptake, with several studies documenting that physical activity interventions remain underused [22] and only a small proportion of clinicians providing recommendations that are consistent with scientific evidence and national guidance [23]. This may be due, in part, to a lack of clinician training and practical barriers for clients [24].

GamerFit is a theory-based exergaming (ie, gaming that requires physical exertion) and telehealth physical activity coaching intervention that has shown efficacy in promoting physical activity and sleep [25,26] and consequently in producing positive psychosocial outcomes (ie, self-regulation, affect, health related quality of life, and social support) among youth aged 13 to 17 years with psychiatric diagnoses [27]. It was adapted from the evidence-based GameSquad intervention designed for youth with overweight and obesity [25]. Adaptations were made to accommodate the needs of youth with heterogeneous and comorbid psychiatric diagnoses based on disability-informed empowerment theory [28]. This 12-week intervention uses a mobile app that incorporates telehealth coaching, exergaming challenges, and on-demand exercise videos to increase physical activity levels and improve sleep hygiene [25]. Exergames are video games that require exertion or physical activity from the user and may be more pleasant and enjoyable than traditional modes of physical activity based on user characteristics, including leisure time preferences [29]. GamerFit participants use an app that includes a menu of exergames to select (and then play the games on a separate video game console); progressive physical activity and sleep hygiene planning and tracking; step tracking that automatically syncs with a Fitbit (Fitbit Inc), which automatically uploads data to the app for self-monitoring and coach monitoring; real-time automated feedback; prerecorded videos using peer modeling to promote health behaviors; and preset motivational texting encouraging youth to engage in at least 60 minutes per day of moderate-to-vigorous physical activity and to achieve 8 to 10 hours per day of quality sleep. App users, with an accompanying caregiver, schedule telehealth sessions with a trained coach once weekly for 12 weeks, to set physical activity and sleep goals and to overcome barriers to achieving weekly goals. The coaches are trained in motivational and goal-setting principles and, importantly, do not need to have a specific license or background in mental health treatment or exercise physiology. The app and intervention were designed by the research team and developed by a third-party developer; it is not available in the public domain. This intervention offers a solution to overcome youth and family barriers to engagement in physical activity by providing structured, child-focused programming in the home setting and accountability with a telehealth coach.

The next step of this work was to translate GamerFit for clinical use to improve accessibility for youth with psychiatric disorders. To do so, the purpose of this study was to explore clinician feedback on incorporating GamerFit into ongoing mental health treatment, ultimately to help clinicians make physical activity and sleep prescriptions.

Methods

Participants

Practicing mental health clinicians were recruited via email listservs and word-of-mouth referrals to participate in one of two 75-minute focus groups to gather perceptions of the GamerFit intervention as an adjunct to their ongoing mental

XSL•FO
RenderX

health treatment practice. Participants were required to be currently practicing clinical pediatric mental health providers who were willing and able to participate in an English-language focus group. A convenience sample of participants was identified based on their clinical background and their response to recruitment attempts within the given 6-month time frame, and the remotely delivered groups were scheduled around the clinicians' availability. After the second focus group, analysis indicated thematic saturation; therefore, no additional participants were recruited.

Procedures

Focus groups were held via Zoom (Zoom Communications, Inc), a Health Insurance Portability and Accountability Act (HIPAA)–compliant web and videoconferencing platform, and were facilitated by trained study team members, including a clinical psychologist and a research scientist with experience delivering health promotion interventions for youth with psychiatric diagnoses. The interviewers were careful to avoid leading questions about the app and instead encouraged participants to share suggestions for improvement and practical considerations for implementation in clinical practice. Participants were presented with an overview of the GamerFit intervention components, which included a video demonstration of the user experience. Participants then asked questions about the GamerFit intervention. Next, a semistructured, empowerment theory-based instrument was administered with questions regarding (1) perceptions of using physical activity and sleep prescriptions in youth mental health treatment plans, (2) considerations for implementation of GamerFit in mental health treatment plans (ie, feasibility, acceptability, accessibility, and implementation fidelity), and (3) approaches for delivering physical activity and sleep prescriptions via the GamerFit intervention (Multimedia Appendix 1). These procedures did not change over the course of the study. The Standards for Reporting Qualitative Research guidelines [30] were followed in drafting and editing this manuscript (Multimedia Appendix 2).

Ethical Considerations

All study procedures were approved by the Pennington Biomedical Research Center Institutional Review Board (2021-071-PBRC GamerFit). Participants provided verbal consent to participate and be recorded during the session. All data in this study were deidentified, and indirect identifiers were coded to prevent the possibility of reidentification. Participants were provided with a US $50 gift card as compensation for their time.

Data Analysis

The focus group discussions were audiotaped, transcribed, and coded by study staff. Standard qualitative methods (ie, data immersion, clustering of preliminary categories, editing, exploring new categories and subcategories, and identifying major themes) were used. Qualitative analysis included (1) individual quotes isolated in transcripts; (2) a hierarchical coding system developed to organize the quotes to capture the full range and depth of participant responses; and (3) the structure, frequency, and interrelationships of the coded quotes to inform interpretation of results. The process included both inductive analysis (moving from facts to theory) and deductive analysis (applying theory to facts). Two independent reviewers reviewed each transcript line-by-line for patterns and themes. ABB developed the initial codebook, which was then further developed and applied by AMB. The team met to resolve coding disagreements, with AES acting as the tiebreaker, although this was not ultimately required.

Results

Overview

A total of 8 clinicians participated in the focus groups: 2 psychiatrists (MDs), 1 clinical psychologist (PhD), 1 psychiatric nurse practitioner, 1 licensed mental health clinician, and 3 licensed clinical social workers. All were licensed to practice in Massachusetts; 2 practiced in therapeutic school settings, 1 in a residential setting, 3 in clinic settings, and 2 in private practice. Six were women, 2 were men, 7 self-identified as White and 1 self-identified as a person of color.

Theme 1: Perspectives

Overall Importance of Incorporating Physical Activity and Sleep in Mental Health Treatment

Clinicians were prompted to discuss their perspectives on incorporating physical activity and sleep prescriptions in youth mental health treatment plans. Overall, participants endorsed physical activity and sleep prescriptions as important (8/8, 100%) but not necessarily as a focus of treatment. Clinicians generally recognized a relationship between sleep and physical activity, but they varied in their emphasis on physical activity and sleep during treatment sessions. One clinician who emphasized physical activity and sleep during sessions stated:

> *I spend quite a bit of time trying to explain to them that the most important piece of their treatment, that we're all a team...and that physical activity, sleep, and diet would be what I need them to work on first before I can even have an accurate impression of where to go with their treatment or get a baseline of what's going on with them.* [FG2, PT3]

Psychiatric diagnosis and prescribed medications were important considerations for clinicians when selecting physical activity or sleep as a target for treatment during sessions with youth:

> *I tend to talk a little bit more about physical activity just in terms of also medications that the youth may be on and thinking about the impact on physical health and weight...and potentially how that affects sleep and concerns for other medical conditions like sleep apnea.* [FG1, PT4]

> *I typically probably ask more about sleep than about physical activity partly because you know they're frequently coming in with crisis questions.* [FG2, PT4]

> *Even with anxiety or things like that, I find that teaching those coping strategies around a sleep routine tends to be more of a first point of intervention. And again, it's the population, versus depression where I will focus on sleep for sure, but*

XSL•FO
RenderX

> *I'm also doing that behavioral activation, and I try to include a lot of physical activity. So, I think it's contingent on the presenting problem.* [FG1, PT2]

Clinician Confidence in Discussing Health Behaviors During Treatment Sessions With Youth With Psychiatric Diagnoses

Many clinicians, but not all, reported feeling confident and prepared to discuss physical activity (6/8, 75%) and sleep (4/8, 50%) with clients. An emerging theme in this discussion was clinicians' desire for easy-to-apply clinical guidance. Several clinicians reported having basic training and handouts on these topics but expressed a desire for more guidance on helping youth implement health behavior changes:

> *I mean as far as maybe education on the importance, fairly comfortable, I could probably use more. But as far as implementing it right, that's a little more challenging.* [FG2, PT3]

Utility of Physical Activity and Sleep Prescriptions

While most clinicians reported that they did not use physical activity and sleep prescriptions, the majority (6/8, 75%) endorsed a desire to use them and expressed perceived clinical value. However, clarity is still needed for what these interventions should look like based on client age, developmental stage, and other individualized characteristics, as well as on how this guidance can fit within the time limitations of a clinical session:

> *It's actually never occurred to me to do sleep prescription or physical activity prescriptions, but I think it's a fantastic idea.* [FG2, PT5]
>
> *Sleep I feel like I have broad ranges for different age and development...whereas physical activity I'm like, "I don't know, walk for 10 minutes a day."* [FG1, PT2]
>
> *I don't do it. It would be a great idea to do. I think part of where I've come from is the time constraints I have kind of don't allow for some of it, which is unfortunate.* [FG2, PT3]

Use of Screens for Physical Activity

When it comes to engaging in physical activity via screens (ie, apps, exergames, and online coaching), clinicians initially expressed hesitation based on mixed feelings about adolescent screen use. Ultimately, clinicians expressed positive views on using screens to promote physical activity (7/8, 88%), noting their accessibility and acceptability among adolescents:

> *I'd rather see kids be on an exercise app or something that's related to their health than social media.* [FG1, P1]
>
> *This speaks to being in step with meeting them where they are. I think that this is, you know, video games and technology. If you speak the language and you can provide this method [of getting active], you've got a much better chance of getting them to get actively involved. And so I think that especially for physical activity, this is a great method.* [FG2, PT4]

Theme 2: Implementation of GamerFit During Mental Health Treatment

Feasibility

Regarding how realistic it is for clients to engage with the GamerFit intervention, clinicians were optimistic that it was doable but noted that it will likely require parental assistance for many youth (eg, reminders and technical support). Clinicians viewed health coaching as a positive aspect of this intervention that might help clients to be successful and noted that other aspects, such as the in-app journal and monitoring system, might be unlikely to be completed on a regular basis. Other feedback included suggestions for potential modifications based on psychiatric diagnoses and emphasis that some features may be especially suited for youth with certain disorders:

> *The availability of a coach or, or someone that's meeting with them weekly to, to really kind of work through that would be critical.* [FG2, PT6]
>
> *The barrier to entry might be a little higher for kids with a depressive diagnosis and they might need some extra support in getting started. I also think they would be really well suited to this intervention once they get started.* [FG1, PT4]

Acceptability and Appropriateness

Clinicians believed that the exergaming and coaching components of GamerFit would be most appealing to youth with psychiatric diagnoses. Clinicians indicated that age, gender, and socioeconomic status may impact youth engagement with the intervention. Clinicians also suggested that increased incentives and positive reinforcement, as well as more options for physical activity, would improve engagement. Finally, clinicians believed that the ability to engage with GamerFit at home was a major advantage for those who prefer privacy:

> *It's just probably more acceptable than some of the other things that we recommend…the availability of a coach…would be critical.* [FG2, PT6]
>
> *For me it would be kind of age, their diagnosis, whether or not there's a reward at the end, you know, whether they get any kind of benefit other than just the little trophy at the end.* [FG2, PT4]
>
> *Socioeconomic status...I see kids [without access to school and sport programming] buying in way more versus the kids [with lots of resources].* [FG1, PT2]

Theme 3: Approaches for Delivering Prescriptions via GamerFit

Incorporation Into Practice

Most clinicians (7/8, 88%) endorsed a willingness to incorporate the GamerFit intervention into their clinical practices. Clinicians expressed enthusiasm about being able to view client engagement with the app to facilitate discussion during sessions, as well as about having resources to share with children and families on the importance of physical activity and sleep, supported by a tangible tool to promote these health behaviors. They were also enthusiastic about the potential for this intervention to be particularly helpful to youth with autism

XSL•FO
RenderX

spectrum disorder as they become older and work toward becoming more independent in their health behaviors. An important concern mentioned was a lack of time during outpatient sessions to commit to covering all necessary topics such as health behaviors during regularly scheduled appointments. Other suggestions based on clinician experiences included integrating this intervention into school and residential settings, incorporating a group aspect to promote adherence and engagement, and including additional educational materials to share with clients and families about the role of physical activity and sleep in mental health treatment and how to navigate technology challenges. An additional desire described was advocacy for insurance provisions for coaching and technology coverage to improve family involvement and allow participants to continue the intervention at home after discharge:

> *I would love to be able to see this in my practice, as well as be able to access the data myself...and be able to talk directly about it...like, "Tuesdays are really hard for you. You don't get any steps on Tuesdays." Have those very specific conversations about troubleshooting.* [FG1, PT4]

> *You also want to have a fact sheet about the importance of physical exercise and sleep and then be able to step off from that and go, hey, you know...here's this thing that we are trying and again, can they, will they buy in?* [FG2, PT4]

> *[T]he older population of kids on the spectrum, you know we're looking for independent and we saw obesity and game addiction as critical issues that we were confronting all the time, so I think it's going to be advantageous.* [FG2, PT4]

> *[O]ur kids are in residential and so it seems like an easier intervention to implement in our setting.... I mean really easy, in fact, for us to implement and get almost every kiddo on board, that they would love it, and for us build it right into the rhythm of the day.* [FG1, PT1]

Reporting

Clinicians also provided guidance for preferred methods of accessing, sharing, and using patient data from the GamerFit intervention. Emails integration and incorporation of reporting components into the electronic health record system were identified as the easiest ways to adopt this intervention within clinical practices:

> *If I were to be able to have [it] sent to me weekly...tracking their progress, how much they're using it...that would help me kind of see both how effective it is but also kind of what aspects of it are most helpful.* [FG2, PT6]

> *One email with all the reports [for clients] instead of individual emails of each participant.* [FG1, PT3]

Discussion

Practicing mental health clinicians participated in 1 of 2 focus groups to review the GamerFit intervention and provide feedback on incorporating this program into ongoing mental health treatment to facilitate physical activity and sleep prescriptions. GamerFit was widely perceived as important, clinically feasible, appealing, and appropriate for the target population (ie, youth with common, often comorbid, psychiatric diagnoses). Clinicians identified suggested adaptations to improve clinical workflow and usability as well as suggestions to improve patient engagement.

Many clinicians reported that physical activity and sleep are important aspects of supporting optimal mental health and expressed a desire to incorporate health behavior counseling into their treatment practices. However, they did not feel confident in their training and abilities to incorporate robust eHealth-based physical activity and sleep interventions within sessions with clients. Moreover, there were mixed findings regarding the use of screens for health promotion: some clinicians viewed this as an opportunity to appeal to client interests in media and gaming and to improve accessibility, particularly for those without safe environments for other forms of physical activity, whereas others were more hesitant, given the already high levels of screen time in youth. These findings are consistent with other reports from mental health clinicians on the utility of eHealth in psychiatry, as described by Nicholas et al [31], who found enthusiasm for improving accessibility to treatment in a widely available manner that overcomes resource limitations and simultaneously reduces mental health stigma as a barrier to care. Furthermore, eHealth interventions offer clients the opportunity to access support and to engage with treatment in real time as needed.

In addition to concerns about sedentary behavior associated with screen time, ensuring client confidentiality and data privacy is essential in the design and implementation of eHealth programming. Transparency in data collection purposes and use, as well as established trust between the user, clinician, and the technology or platform itself, are required to overcome this obstacle [31]. eHealth interventions embedded within mental health treatment practices also offer opportunities to enhance real-time self-monitoring, extend clinicians' assessment of clients within their natural environment (eg, including caregiver involvement), and facilitate data processing to support personalized treatment plans and decision-making [32].

The next finding of this qualitative study was the perceived feasibility and acceptability of the GamerFit intervention in the clinical setting. Overall, the intervention was deemed to be practical, with some aspects perceived to be more engaging than others (eg, health coaching as a critical component, whereas in-app journaling for self-monitoring behavior goals was seen as less likely to be completed). Furthermore, suggestions were made for adapting the intervention based on client age and diagnoses, for example, linking in-app earned rewards with age-appropriate tangible rewards or using the sleep prescription to avoid oversleeping among some youth with depression. The biggest concerns about feasibility and acceptability were potential technical challenges and the likelihood that some clients would require support from parents or caregivers.

Results of this study will be used to help match future features with clinician and client needs (ie, integration into electronic medical records systems and an expanded selection of

XSL•FO
RenderX

exergames) [33]. eHealth interventions are widely effective in treating mental health disorders when used alongside psychotherapy, particularly when they are guided (vs unguided) [34], as is the case with the GamerFit intervention. Previous studies have shown that children with psychiatric disorders prefer eHealth interventions that use features such as videos, limited text, personalization options, opportunities to connect with others, and texting reminders [35], all of which are incorporated into the GamerFit intervention. Future iterations of GamerFit will likely provide more evidence on how to best support users with technological challenges, and the inclusion of a health coach is intended to overcome support needs that would typically be required of parents and caregivers. Caregivers are often resource depleted (ie**,** limited in time and energy) and constrained in taking on additional responsibilities. The GamerFit app may help overcome these limitations by applying external accountability and motivation through remote health coaching, thereby encouraging engagement in healthy physical activity and sleep practices.

Limitations of the study included the relatively homogenous sample of participants in the focus groups in terms of race and gender. However, this sample is representative of the demographic majority of the US psychology workforce [36]. An additional limitation is the potential of eHealth interventions to widen the divide in health disparities for those with limited digital literacy and device ownership [31]. Digital literacy and access to technology; data storage; Wi-Fi capabilities; compliance with wearing the Fitbit; and the cognitive abilities and planning skills needed to adhere to the scheduled appointments with the health coach and to complete the physical activity, sleep, and exergaming goals must be considered in the design and implementation of this treatment option.

This study facilitates the translation of a research-based intervention to real-world application for the alleviation of mental and physical health symptoms among a population at elevated risk via direct input from key collaborators (ie, the mental health clinicians). These findings will inform the adaptation and integration of GamerFit as an adjunct to clinical practice with specific considerations for the user (ie, clinician and client) experience, resources, and capabilities.

Acknowledgments

Funding was provided by the National Institutes of Health, Eunice Kennedy Shriver National Institute of Child Health and Human Development (R21HD106465). The funder was not involved in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: ABB, AES, JS
Data curation: AMB, ABB
Formal analysis: AMB, ABB
Funding acquisition: ABB, AES
Investigation: ABB, AES
Resources: AES
Software: ABB
Supervision: JS
Writing – original draft: AMB
Writing – review & editing: AMB, ABB, AES, JS

Conflicts of Interest

None declared.

Multimedia Appendix 1
Semistructured instrument for cross-sectional qualitative data collection across 2 remotely conducted focus groups of 8 mental health clinicians.
[DOCX File , 17 KB - formative_v9i1e71569_app1.docx]

Multimedia Appendix 2
Standards for Reporting Qualitative Research reporting checklist.
[DOCX File , 214 KB - formative_v9i1e71569_app2.docx]

References

XSL•FO
RenderX

1. Gillis BT, El-Sheikh M. Sleep and adjustment in adolescence: physical activity as a moderator of risk. Sleep Health 2019 Jun;5(3):266-272 [FREE Full text] [doi: 10.1016/j.sleh.2019.02.001] [Medline: 30928497]
2. Kandola A, Lewis G, Osborn DP, Stubbs BH, Hayes JF. Depressive symptoms and objectively measured physical activity and sedentary behaviour throughout adolescence: a prospective cohort study. Lancet Psychiatry 2020 Mar;7(3):262-271 [FREE Full text] [doi: 10.1016/S2215-0366(20)30034-1] [Medline: 32059797]
3. Carter T, Pascoe M, Bastounis A, Morres ID, Callaghan PP, Parker AG. The effect of physical activity on anxiety in children and young people: a systematic review and meta-analysis. J Affect Disord 2021 May 15;285:10-21 [FREE Full text] [doi: 10.1016/j.jad.2021.02.026] [Medline: 33618056]
4. Bland VL, Heatherington-Rauth M, Howe C, Going SB, Bea JW. Association of objectively measured physical activity and bone health in children and adolescents: a systematic review and narrative synthesis. Osteoporos Int 2020 Oct;31(10):1865-1894 [FREE Full text] [doi: 10.1007/s00198-020-05485-y] [Medline: 32542544]
5. Morelli C, Avolio E, Galluccio A, Caparello G, Manes E, Ferraro S, et al. Impact of vigorous-intensity physical activity on body composition parameters, lipid profile markers, and irisin levels in adolescents: a cross-sectional study. Nutrients 2020 Mar 11;12(3):742 [FREE Full text] [doi: 10.3390/nu12030742] [Medline: 32168929]
6. Physical activity guidelines for Americans, 2nd edition. U.S. Department of Health and Human Services. URL: https://health.gov/paguidelines/second-edition/pdf/Physical_Activity_Guidelines_2nd_edition.pdf [accessed 2024-11-11]
7. Merlo CL, Jones SE, Michael SL, Chen TJ, Sliwa SA, Lee SH, et al. Dietary and physical activity behaviors among high school students - youth risk behavior survey, United States, 2019. MMWR Suppl 2020 Aug 21;69(1):64-76 [FREE Full text] [doi: 10.15585/mmwr.su6901a8] [Medline: 32817612]
8. Bitsko RH, Claussen AH, Lichstein J, Black LI, Jones SE, Danielson ML, Contributor. Mental health surveillance among children - United States, 2013-2019. MMWR Suppl 2022 Mar 25;71(2):1-42 [FREE Full text] [doi: 10.15585/mmwr.su7102a1] [Medline: 35202359]
9. Mangerud WL, Bjerkeset O, Lydersen SI, Indredavik MS. Physical activity in adolescents with psychiatric disorders and in the general population. Child Adolesc Psychiatry Ment Health 2014 Jan 22;8(1):2 [FREE Full text] [doi: 10.1186/1753-2000-8-2] [Medline: 24450542]
10. Brady AM, Deighton J, Stansfeld S. Psychiatric outcomes associated with chronic illness in adolescence: a systematic review. J Adolesc 2017 Aug;59:112-123 [FREE Full text] [doi: 10.1016/j.adolescence.2017.05.014] [Medline: 28605691]
11. Master L, Nye RT, Lee S, Nahmod NG, Mariani S, Hale LB, et al. Bidirectional, daily temporal associations between sleep and physical activity in adolescents. Sci Rep 2019 May 22;9(1):7732 [FREE Full text] [doi: 10.1038/s41598-019-44059-9] [Medline: 31118441]
12. Palmer CA. Tired teens: sleep disturbances and heightened vulnerability for mental health difficulties. J Adolesc Health 2020 May;66(5):520-521 [FREE Full text] [doi: 10.1016/j.jadohealth.2020.01.023] [Medline: 32331621]
13. Hickingbotham MR, Wong CJ, Bowling AB. Barriers and facilitators to physical education, sport, and physical activity program participation among children and adolescents with psychiatric disorders: a systematic review. Transl Behav Med 2021 Oct 15;11(9):1739-1750 [FREE Full text] [doi: 10.1093/tbm/ibab085] [Medline: 34166515]
14. Bowling A, Blaine RE, Kaur R, Davison KK. Shaping healthy habits in children with neurodevelopmental and mental health disorders: parent perceptions of barriers, facilitators and promising strategies. Int J Behav Nutr Phys Act 2019 Jul 26;16(1):52 [FREE Full text] [doi: 10.1186/s12966-019-0813-6] [Medline: 31242904]
15. Toscano CV, Ferreira JP, Quinaud RT, Silva K, Carvalho HM, Gaspar JM. Exercise improves the social and behavioral skills of children and adolescent with autism spectrum disorders. Front Psychiatry 2022;13:1027799 [FREE Full text] [doi: 10.3389/fpsyt.2022.1027799] [Medline: 36620673]
16. Peng S, Fang Y, Othman AT, Liang J. Meta-analysis and systematic review of physical activity on neurodevelopment disorders, depression, and obesity among children and adolescents. Front Psychol 2022;13:940977 [FREE Full text] [doi: 10.3389/fpsyg.2022.940977] [Medline: 36533019]
17. Philippot A, Dubois V, Lambrechts K, Grogna D, Robert A, Jonckheer U, et al. Impact of physical exercise on depression and anxiety in adolescent inpatients: a randomized controlled trial. J Affect Disord 2022 Mar 15;301:145-153 [FREE Full text] [doi: 10.1016/j.jad.2022.01.011] [Medline: 35007642]
18. Seiffer B, Hautzinger M, Ulrich RW, Wolf S. The efficacy of physical activity for children with attention deficit hyperactivity disorder: a meta-analysis of randomized controlled trials. J Atten Disord 2022 Mar;26(5):656-673 [FREE Full text] [doi: 10.1177/10870547211017982] [Medline: 34041952]
19. Arns M, Kooij JJ, Coogan AN. Review: identification and management of circadian rhythm sleep disorders as a transdiagnostic feature in child and adolescent psychiatry. J Am Acad Child Adolesc Psychiatry 2021 Oct;60(9):1085-1095 [FREE Full text] [doi: 10.1016/j.jaac.2020.12.035] [Medline: 33556454]
20. Pedersen BK, Saltin B. Exercise as medicine - evidence for prescribing exercise as therapy in 26 different chronic diseases. Scand J Med Sci Sports 2015 Dec;25 Suppl 3:1-72 [FREE Full text] [doi: 10.1111/sms.12581] [Medline: 26606383]
21. Woodfield M, Butler NG, Tsappis M. Impact of sleep and mental health in adolescence: an overview. Curr Opin Pediatr 2024 Aug 01;36(4):375-381. [doi: 10.1097/MOP.0000000000001358] [Medline: 38747197]

XSL•FO
RenderX

22. Kleemann E, Bracht CG, Stanton R, Schuch FB. Exercise prescription for people with mental illness: an evaluation of mental health professionals' knowledge, beliefs, barriers, and behaviors. Braz J Psychiatry 2020;42(3):271-277 [FREE Full text] [doi: 10.1590/1516-4446-2019-0547] [Medline: 32130402]
23. Escobar-Roldan ID, Babyak MA, Blumenthal JA. Exercise prescription practices to improve mental health. J Psychiatr Pract 2021 Jul 28;27(4):273-282. [doi: 10.1097/PRA.0000000000000554] [Medline: 34398577]
24. Way K, Kannis-Dymand L, Lastella M, Lovell GP. Mental health practitioners' reported barriers to prescription of exercise for mental health consumers. Ment Health Phys Act 2018 Mar;14:52-60 [FREE Full text] [doi: 10.1016/j.mhpa.2018.01.001]
25. Bowling AB, Slavet J, Hendrick C, Beyl R, Nauta P, Augustyn M, et al. The adaptive GameSquad Xbox-based physical activity and health coaching intervention for youth with neurodevelopmental and psychiatric diagnoses: pilot feasibility study. JMIR Form Res 2021 May 14;5(5):e24566 [FREE Full text] [doi: 10.2196/24566] [Medline: 33988508]
26. Staiano AE, Beyl RA, Guan W, Hendrick CA, Hsia DS, Newton Jr RL. Home-based exergaming among children with overweight and obesity: a randomized clinical trial. Pediatr Obes 2018 Dec;13(11):724-733 [FREE Full text] [doi: 10.1111/ijpo.12438] [Medline: 30027607]
27. Bowling A, Button A, Beyl R, Slavet J, Daly T, Murray P, et al. Improving health behaviors and symptoms in youth with mental health disorders: the GamerFit RCT. Ment Health Phys Act 2025 Oct;29:100722 [FREE Full text] [doi: 10.1016/j.mhpa.2025.100722]
28. Perkins DD, Zimmerman MA. Empowerment theory, research, and application. Am J Comm Psychol 1995 Oct;23(5):569-579 [FREE Full text] [doi: 10.1007/bf02506982]
29. Glen K, Eston R, Loetscher T, Parfitt G. Exergaming: feels good despite working harder. PLoS One 2017;12(10):e0186526 [FREE Full text] [doi: 10.1371/journal.pone.0186526] [Medline: 29059227]
30. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014;89(9):1245-1251 [FREE Full text] [doi: 10.1097/acm.0000000000000388]
31. Nicholas J, Huckvale K, Larsen ME, Basu A, Batterham PJ, Shaw F, et al. Issues for eHealth in psychiatry: results of an expert survey. J Med Internet Res 2017 Mar 28;19(2):e55 [FREE Full text] [doi: 10.2196/jmir.6957] [Medline: 28246068]
32. Berrouiguet S, Perez-Rodriguez MM, Larsen M, Baca-García E, Courtet P, Oquendo M. From eHealth to iHealth: transition to participatory and personalized medicine in mental health. J Med Internet Res 2018 Jan 03;20(1):e2 [FREE Full text] [doi: 10.2196/jmir.7412] [Medline: 29298748]
33. Hendrikoff L, Kambeitz-Ilankovic L, Pryss R, Senner F, Falkai P, Pogarell O, et al. Prospective acceptance of distinct mobile mental health features in psychiatric patients and mental health professionals. J Psychiatr Res 2019 Mar;109:126-132 [FREE Full text] [doi: 10.1016/j.jpsychires.2018.11.025] [Medline: 30530207]
34. Bennett CB, Ruggero CJ, Sever AC, Yanouri L. eHealth to redress psychotherapy access barriers both new and old: a review of reviews and meta-analyses. J Psychother Integr 2020 Jun;30(2):188-207 [FREE Full text] [doi: 10.1037/int0000217]
35. Liverpool S, Mota CP, Sales CM, Čuš A, Carletto S, Hancheva C, et al. Engaging children and young people in digital mental health interventions: systematic review of modes of delivery, facilitators, and barriers. J Med Internet Res 2020 Jul 23;22(6):e16317 [FREE Full text] [doi: 10.2196/16317] [Medline: 32442160]
36. Data tool: demographics of the U.S. psychology workforce. American Psychological Association. URL: https://www.apa.org/workforce/data-tools/demographics [accessed 2024-11-11]

Abbreviations

HIPAA: Health Insurance Portability and Accountability Act

Edited by A Mavragani; submitted 21.01.25; peer-reviewed by KW Tay, F Dickerson; comments to author 11.08.25; accepted 15.10.25; published 07.11.25.

Please cite as:
Button AM, Slavet J, Staiano AE, Bowling AB
Clinician Perspectives on Incorporating Physical Activity and Sleep Prescriptions Using eHealth for Youth With Comorbid Psychiatric Disorders: Qualitative Focus Group Study
JMIR Form Res 2025;9:e71569
URL: https://formative.jmir.org/2025/1/e71569
doi:10.2196/71569
PMID:

©Alyssa M Button, James Slavet, Amanda E Staiano, April B Bowling. Originally published in JMIR Formative Research (https://formative.jmir.org), 07.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction

XSL•FO
RenderX

in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Generative AI Chatbot for Diabetes Management: Formative 2-Part Qualitative Study Using DTalksBot Involving Patients and Clinicians

Soyun Jeon[1], MA; Seolhee Lee[1], MA; Esther Hehsun Kim[1], MA; Jinsu Eun[2], MS; Kwangwon Lee[3], MS; Hajin Lim[1], PhD; Joonhwan Lee[1,3], PhD

[1]Department of Communication, Seoul National University, Seoul, Republic of Korea
[2]Department of Transdisciplinary Studies, Seoul National University, Seoul, Republic of Korea
[3]Interdisciplinary Program in Artificial Intelligence, Seoul National University, Seoul, Republic of Korea

Corresponding Author:
Joonhwan Lee, PhD
Department of Communication
Seoul National University
1, Gwanak-ro, Gwanak-gu
Seoul, 08826
Republic of Korea
Phone: 82 2 880 6450
Email: joonhwan@snu.ac.kr

Abstract

Background: Diabetes mellitus requires continuous self-management to prevent complications. Patients frequently rely on online resources and mobile apps for diabetes-related information; however, these often lead to information overload, limited personalization, and difficulty in navigation. Generative artificial intelligence (AI) chatbots may address these challenges by providing accessible, personalized, and responsive guidance.

Objective: This study aimed to explore the potential role of generative AI chatbots in diabetes management through a 2-part qualitative evaluation. Part 1 examined patients' information needs, user experiences, and expectations. Part 2 investigated specialists' perspectives on the practical utility of generative AI chatbots in supporting diabetes self-management. By incorporating perspectives from both patients and specialists, the study aimed to identify appropriate boundaries for the involvement of generative AI chatbots, reflecting the needs and expectations of both stakeholder groups.

Methods: This study was conducted using DTalksBot, a generative AI chatbot powered by GPT-4 (OpenAI) and enhanced with retrieval-augmented generation. In Part 1, we aimed to understand the experiences, needs, and expectations of patients with diabetes. To achieve this, 24 participants engaged in structured chatbot sessions, completed postinteraction surveys, and participated in in-depth interviews. Data were analyzed using thematic and content analysis to identify patterns in user queries and experiences. In part 2, we invited 4 family medicine specialists to assess the accuracy of DTalksBot's responses by reviewing conversation logs and to share expert insights on the future role of generative AI chatbots in diabetes management.

Results: In part 1, a total of 24 patients submitted a total of 643 questions, which were categorized into 4 primary themes: personalized health advice and guidance (n=281, 44.6%), complications and comorbidities (n=174, 27.1%), medication and treatment exploration (n=111, 17.3%), and mental health management and support (n=30, 4.7%). Patients emphasized the advantages of generative AI chatbots over traditional information sources, including faster access to reliable content, reduced cognitive burden, and the ability to comfortably discuss sensitive topics. In part 2, specialists recognized the generative AI chatbots' value in answering routine inquiries, but noted limitations in contextual accuracy, real-time data integration, and response personalization.

Conclusions: Generative AI chatbots showed promise as complementary tools for diabetes self-management by offering accessible, reliable, and tailored support. This formative evaluation provides empirical evidence on how generative AI chatbots can address patient information needs and complement existing health care resources. To maximize utility, future generative AI chatbots need to integrate real-time health data, enhance contextual relevance, and align with clinical workflows to ensure safety, trust, and broader applicability.

XSL•FO
RenderX

(JMIR Form Res 2025;9:e72553) doi:10.2196/72553

KEYWORDS

generative AI chatbot; large language models (LLMs); retrieval-augmented generation (RAG); diabetes mellitus; diabetes management; online health information seeking

Introduction

Background

Diabetes mellitus, a chronic condition caused by insufficient insulin production or usage [1-5], requires vigilant daily self-management to prevent complications such as cardiovascular disease, neuropathy, nephropathy, and retinopathy [1,3,6,7]. More than 85% of cases involve type 2 diabetes, which often improves with lifestyle modification, whereas type 1 diabetes usually requires a balanced diet and insulin therapy [2-5,8,9].

Because daily monitoring and lifestyle management occur largely outside clinical settings, individuals with diabetes require diverse and timely information [10-15]. They often search online more proactively than individuals with other chronic conditions, commonly using Google or Bing [16-18]. However, these search engines can overwhelm users with technical jargon, conflicting advice, and misinformation, highlighting the need for trustworthy digital support [18-21].

Most mobile apps and web-based platforms log quantitative self-monitoring data—blood glucose levels, dietary intake, and medication use—but rarely deliver personalized feedback or conversational support [9,12,22]. Their content is largely generic, and evidence for accuracy, clinical validity, and long-term benefit is limited [9,12]. Furthermore, the absence of real-time decision support has hindered adoption by both patients and clinicians [9]. Collectively, these limitations underscore the need for interactive and context-aware systems that provide individualized guidance.

To address these shortcomings, research on conversational interfaces has gained momentum [23-25]. Early rule-based chatbots, constrained by predefined scripts and decision trees, supported only narrowly defined tasks and offered limited flexibility or personalization, which reduced user engagement [23-28]. Recent advances in large language models have enabled generative artificial intelligence (AI) chatbots capable of interpreting complex queries and generating context-specific responses [29-34]. Systematic reviews have shown that such chatbots are being piloted across diverse health care domains, including chronic disease self-management, cancer care, and health behavior change. These systems often facilitate patient education, medication reminders, and emotional support, with generally high user acceptance but limited rigorous evidence of effectiveness [23,24,28]. Notably, a scoping review found that 30% of AI apps for chronic condition self-management focused on diabetes; however, most remain in early developmental stages and particularly lack support for emotional self-management, a critical component of comprehensive diabetes care [35]. Furthermore, a recent synthesis of conversational agents for chronic condition self-management highlighted that although such systems show promise in supporting patient engagement and self-care, their design and evaluation approaches are still characterized by conceptual fragmentation, indicating the need for more rigorous and multidisciplinary frameworks to enhance emotional and personalized support [36].

Beyond these design and evaluation challenges, unresolved safety risks further hinder the clinical deployment of generative AI chatbots. Insufficient transparency in training data and opaque inference mechanisms increase the risk of hallucinations—incorrect or misleading information presented as fact [31,37-45]. ChatGPT (OpenAI) and Gemini (Google) have provided incorrect guidance, particularly regarding medication-related inquiries, which underscores the need for cautious implementation and oversight in health care settings [37,46,47]. Consequently, some clinicians remain cautious about their use in clinical practice, whereas others view the technology as potentially viable if robust safeguards for safety and clinical reliability are implemented [31,44].

While interest in generative AI chatbots within health care is increasing, robust empirical studies for disease-specific implementations remain limited. Previous research has largely focused on general-purpose systems, leaving unanswered questions about whether generative AI chatbots can provide accurate, timely, and contextually relevant support to specific patient populations [48-54]. Furthermore, most existing studies assess either patient experience or clinician appraisal in isolation; accordingly, further research that integrates both stakeholder perspectives within a unified evaluative framework is warranted.

To address these gaps, we designed a 2-part qualitative study that incorporated both patient and specialist perspectives to explore the potential of generative AI chatbots in diabetes management. To ensure a safe and context-specific environment for interaction, we developed and used DTalksBot, a diabetes-specific generative AI chatbot powered by GPT-4 and enhanced with retrieval-augmented generation (RAG). By conducting interviews with both patients and specialists, we aimed to identify key insights into user perceptions, expectations, and the envisioned role of generative AI chatbots in diabetes management. The findings are expected to guide the safe and effective integration of generative AI chatbots into chronic disease care and provide actionable insights for clinicians considering AI-driven tools and for developers aiming to enhance chatbot design with improved contextual relevance, safety, and personalization.

Objectives

A 2-part evaluation was conducted to explore how generative AI chatbots could support diabetes management through 5 research questions (RQs): RQs 1-4 were investigated in part 1 with individuals with diabetes, whereas RQ 5 was explored in part 2 with family medicine specialists.

XSL•FO
RenderX

In part 1, the following RQs were investigated. First, what types of questions do patients frequently ask, and what information do they seek? This question was examined by analyzing conversation logs between participants and DTalksBot to identify recurring question types and information needs. Second, how do patients evaluate the overall user experience of DTalksBot? This question was addressed through postinteraction surveys in which patients rated usability, message credibility, emotional support, empathy, trust, and overall satisfaction. Third, how does interacting with DTalksBot compare with patients' use of online resources and consultations with health care providers? This question was explored by comparing interview narratives about generative AI chatbots with participants' accounts of their typical online searches and clinician consultations. Fourth, what are patients' expectations and preferences for generative AI chatbots in diabetes management? This question was investigated through thematic analysis of interview transcripts, focusing on desired functions and support features.

In part 2, the following RQ was explored: how do family medicine specialists evaluate DTalksBot's responses, and what role do they envision for generative AI chatbots in diabetes management? This question was studied by analyzing interviews for perceived utility, limitations, and future directions.

Methods

Study Design

Previous studies have emphasized the importance of qualitative inquiry in understanding how individuals navigate and use online health information effectively [18]. In the context of chronic disease management, where patients face distinctive challenges in accessing, interpreting, and applying information, qualitative approaches are particularly effective in capturing the complexity of patients' information needs and the contextual barriers they encounter [18]. Furthermore, as generative AI chatbots represent a novel and interactive medium for health communication, qualitative methods are especially suited to uncover the nuanced ways in which patients perceive, engage with, and form expectations of such systems. Therefore, this study adopted a qualitative research design to explore how generative AI chatbots could support diabetes management. Grounded in a constructivist paradigm, this approach enabled a deeper examination of how individuals with diabetes interact with generative AI chatbots and form expectations—insights that would be difficult to obtain through structured surveys or quantitative metrics.

In addition to qualitative methods, descriptive statistical analyses were also incorporated to summarize the overall user experience. Interviews and analysis were conducted by authors with expertise in human-computer interaction. None had previous relationships with participants. This study followed the Standards for Reporting Qualitative Research guidelines (Multimedia Appendix 1).

Participants

For RQ 1 through RQ 4, we recruited Korean adult patients with diabetes who were actively managing their condition through medication, insulin injections, or regular blood glucose monitoring. This group was selected because individuals engaged in active self-management were more likely to seek health information online and exhibit proactive health behaviors [55,56]. Recruitment took place entirely online, outside of clinical settings, via nationwide diabetes-related community platforms. Interested individuals who viewed the recruitment notice submitted an application form, and the research team contacted them individually to confirm eligibility and arrange the study session. The sample size was determined based on existing guidelines for achieving data saturation in qualitative research [57].

While this study focused on Korean adults, the prevalence of diabetes in South Korea was comparable with that in many Western countries [5]. Although Korean patients with type 2 diabetes tended to have lower BMI than their Western counterparts, they exhibited high rates of insulin resistance and cardiometabolic complications—partly due to carbohydrate-rich diets and increased central adiposity despite normal weight [5]. Consequently, dietary regulation played a critical role in diabetes care, and patients could benefit from continued, personalized guidance. These population-specific characteristics underscore the relevance of generative AI chatbots as practical tools for daily self-management in the Korean context.

Participants included both patients with type 1 and type 2 diabetes, but no distinction was made during recruitment and analysis. This decision reflected the study's focus on lifestyle-based self-management rather than disease-specific pathology. Type 1 diabetes generally begins earlier in life and requires lifelong insulin therapy due to autoimmune destruction of β-cells, whereas type 2 diabetes is associated with insulin resistance and lifestyle factors [1,3]. Despite these differences, both conditions required ongoing self-management, including blood glucose monitoring, dietary regulation, and physical activity [3]. Given these shared management needs, this study explored the potential of generative AI chatbots to support diabetes self-management in a broadly applicable context.

For RQ 5, Korean family medicine specialists with clinical expertise in treating diabetes were invited to provide expert evaluations. These participants were identified through professional referrals within the medical community. Similarly, a small number of expert participants were considered sufficient to yield meaningful insights, in line with previous recommendations for expert sampling in qualitative studies [57].

Procedure

Overview

This study was conducted between September 2023 and June 2024, encompassing chatbot development, participant recruitment, and data collection and analysis. The study consisted of two parts: (1) patients interacted with DTalksBot and then completed an evaluation, and (2) specialists reviewed conversation logs and provided insights on the potential role of generative AI chatbots in diabetes management. Part 1 lasted 60 minutes, whereas part 2 lasted 90 minutes. This 2-part structure was designed to integrate insights from both patients

XSL•FO
RenderX

and clinical experts to inform the future development of generative AI chatbots.

Part 1: Using and Assessing DTalksBot With Patients

Patients joined either in person or remotely and received a brief orientation. A presurvey collected demographic and health-related information, including diabetes type, duration, treatment methods, and typical sources of health information. The interaction with DTalksBot was divided into four 5-minute structured sessions (20 minutes total): session 1 allowed free exploration of diabetes-related topics; session 2 focused on daily management, including diet, exercise, and lifestyle habits; session 3 centered on medical treatment and emergencies; session 4 was open-ended for additional concerns.

Afterward, patients completed a postsurvey evaluating DTalksBot. Measures included usability [58], message credibility [59], perceived social support [60], empathy [61], trustworthiness [62], expectations [63], and satisfaction [64] (Multimedia Appendix 2). Follow-up semistructured interviews were then conducted to explore patients' experiences and expectations regarding generative AI chatbots. These interviews explored how patients compared them with conventional online resources and clinical consultations and perceived the role of generative AI chatbots in diabetes self-management.

Part 2: Analyzing Log Data and Assessing DTalksBot With Family Medicine Specialists

In the second part, family medicine specialists reviewed conversation logs from the patient interactions and interacted with DTalksBot to evaluate its performance. They qualitatively assessed the accuracy, context relevance, and clinical appropriateness of its responses, as well as its ability to handle high-risk scenarios safely and to remain within the scope of verified, diabetes-specific information based on current clinical guidelines. They also identified specific areas requiring improvement. Subsequently, semistructured interviews were conducted using a guide informed by previous research [29], focusing on the broader role of generative AI chatbots in diabetes management. These expert insights informed both the evaluation of DTalksBot and broader considerations for designing and implementing generative AI chatbots in diabetes management.

Developing the DTalksBot

Widely adopted generative AI chatbots, such as ChatGPT and Gemini, often exhibit hallucination and tend to deliver overly generalized responses that lack the specificity required for clinical use [31,38,45,65,66]. Although ongoing updates to large language models have reduced hallucination rates, these systems remain trained on potentially biased or unverified corpora, leaving unresolved challenges about data transparency and model explainability [25,31,45,67]. In high-stakes domains such as health care, where accuracy and contextual relevance are critical, these limitations can pose serious patient safety risks [31,37,38,43,44,46]. To mitigate these risks and ensure a safe experimental environment, we developed DTalksBot, a domain-specific generative AI chatbot specialized for diabetes management.

DTalksBot was developed by integrating GPT-4 with a RAG model to generate medically accurate and context-specific responses. While GPT-4 enabled natural language processing, the RAG model enhanced factual accuracy by retrieving information from a curated knowledge base. This database was built on validated sources, including the National Health Information Portal [68], the Korean Diabetes Association [69], and the Centers for Disease Control and Prevention [70], thereby ensuring evidence-based diabetes guidance [71].

The operational structure of DTalksBot was designed to maximize transparency and explainability while minimizing hallucinations. Reference materials were segmented into individual sentences and converted into vectors. When a user submitted a question, the system calculated the cosine similarity between the query and all sentence vectors, retrieved the top 5 most relevant sentences, and used GPT-4 to synthesize an accessible response. This design grounded each answer in verifiable sources and kept the reasoning traceable. By restricting outputs to verified diabetes-specific information and avoiding speculation, DTalksBot reduced hallucinations commonly observed in general-purpose generative AI chatbots [30,66,71-73].

Unlike general-purpose generative AI chatbots that answer any query, DTalksBot was restricted to diabetes-related topics, maintaining dialogue until the user chose to end the session and retaining previous conversational context to provide coherent follow-up responses. It declined to respond to unrelated queries—such as those about hair loss—thereby minimizing risk and maintaining thematic focus (Figure 1). When a question fell outside its curated knowledge base, DTalksBot prompted the user to pose an alternative question. The system prompt instructed DTalksBot to deliver concise yet supportive explanations with a brief rationale for each answer. It delivered condition-specific responses, such as dietary advice tailored to individuals with diabetes rather than generic public recommendations, and covered diverse aspects of diabetes management, including lifestyle modification, complication prevention, and general treatment guidance.

XSL•FO
RenderX

Figure 1. Example responses to patient queries about hair loss in the context of diabetes self-management. DTalksBot issues a safety-focused refusal (left), while ChatGPT provides a general answer (right).

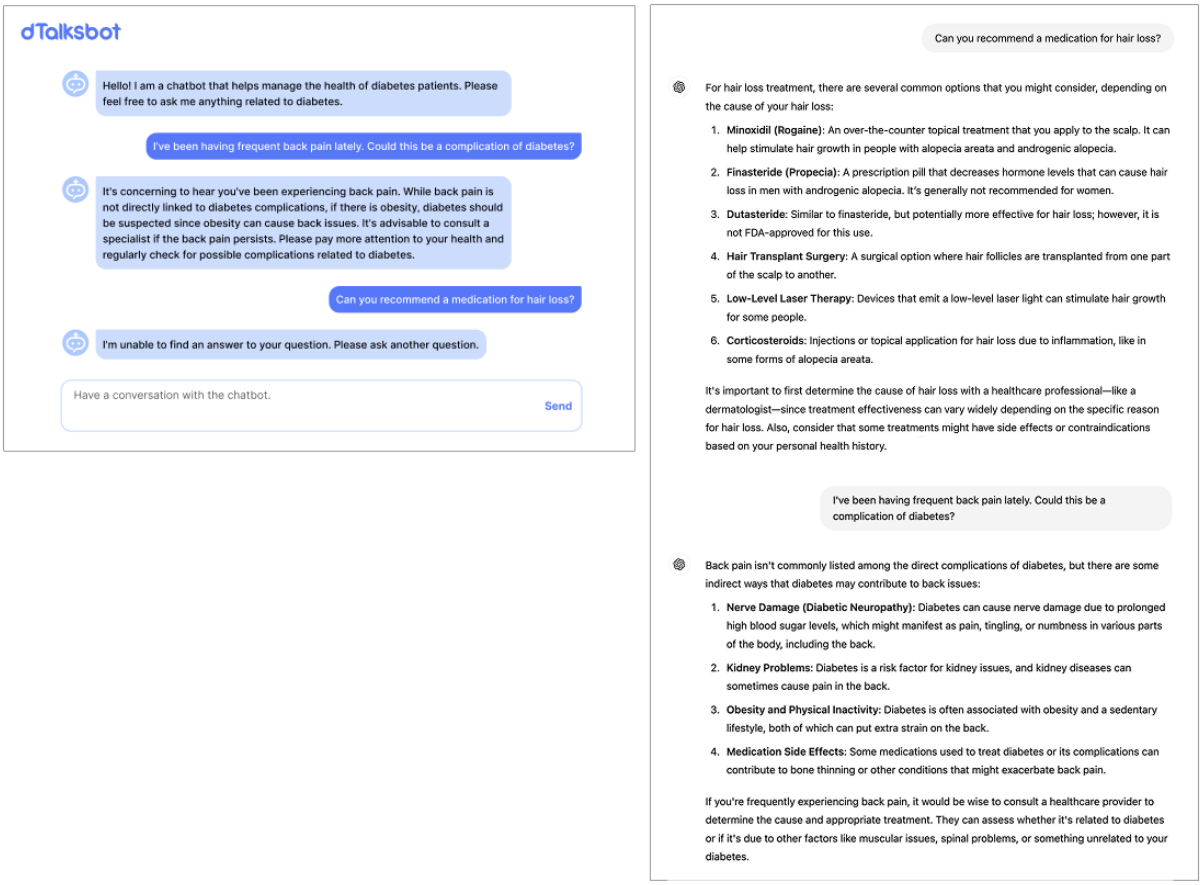

While DTalksBot provided comprehensive, evidence-based information, it intentionally excluded content on medication costs, insurance coverage, and investigational drugs in clinical trials, as these topics vary by individual circumstances and may pose safety concerns. It also did not offer medical diagnoses or prescriptions, instead directing users to seek clinician input for such needs. Furthermore, it did not integrate real-time biometric data, limiting its capacity for highly personalized recommendations. Nonetheless, its design bridged the gap between generic health information and the nuanced needs of patients with diabetes. This clinically bound scope-enabled patients and family medicine specialists to use the system in a safe, reliable, and clinically appropriate environment. Table 1 presents a comparison between DTalksBot and general-purpose generative AI chatbots. Example responses are provided in Multimedia Appendix 3.

Table 1. Key differences between general-purpose generative artificial intelligence chatbots (eg, ChatGPT and Gemini) and DTalksBot, a GPT-4–based chatbot with retrieval-augmented generation for diabetes management.

Feature	General-purpose generative AI[a] chatbots (eg, ChatGPT and Gemini)	DTalksBot
Domain specificity	Covers all topics, including nonmedical	Limited to diabetes-related topics
Knowledge source	Trained on large-scale web data with limited source verification	Based on verified medical sources
Response and reasoning	No explicit reference to sources; reasoning unclear	RAG[b] with validated sources and transparent reasoning
Personalization level	Generic advice for general users	Diabetes-specific advice
Safety for clinical use	Not clinically validated; may be inaccurate	Designed for safe diabetes management under clinical guidance

[a]AI: artificial intelligence.

[b]RAG: retrieval-augmented generation.

XSL•FO
RenderX

Ethical Considerations

The research was conducted strictly following ethical guidelines and received approval from the Institutional Review Board of Seoul National University (IRB 2312/004-016). Informed consent was obtained from all participants after explaining the study's purpose, procedures, and potential risks and benefits. Data were anonymized before analysis and securely stored. Each participant received 20,000 KRW (approximately US $15) as compensation.

Data Analysis

For RQ 1, thematic and content analysis were used to examine DTalksBot interaction logs. Thematic analysis identified recurring question types, while content analysis quantified their frequency. Related questions were grouped using an affinity diagram to structure patients' primary concerns. For RQ 2, descriptive statistics summarized user feedback. For RQ 3, thematic analysis of interviews explored how patients compared generative AI chatbots with traditional online information-seeking and clinician consultations. For RQ 4, thematic analysis examined patient expectations and preferences regarding the support provided by generative AI chatbots. For RQ 5, interviews with specialists were thematically analyzed to assess DTalksBot's accuracy, usability, and clinical alignment, and to gather expert insights on the potential role of generative AI chatbots in diabetes management.

Data Saturation

Data saturation was defined as the point at which no new codes or themes emerged during thematic analysis [74,75]. Previous studies suggested that saturation typically occurred within 12 to 20 interviews in relatively homogenous populations [57,75]. For the specialist interviews, we used purposive expert sampling, as even a small number of domain experts can yield rich qualitative insights [57].

The first, second, and third authors independently conducted open coding on an initial subset of transcripts to ensure analytic diversity. After comparing their code lists, they held consensus meetings to merge overlapping codes and resolve discrepancies, producing a preliminary codebook. This codebook was iteratively refined and applied to the entire data set; saturation was confirmed when 2 consecutive transcript batches produced no new codes.

Results

Participants' Demographic and Clinical Characteristics

A total of 25 patients with diabetes were initially recruited for part 1 of the study; however, 1 participant (P16) withdrew, resulting in a final sample of 24 patients (6 males, 18 females). More than half of the patients were in their 20s and 30s, and most had type 2 diabetes (n=15, 62.5%). The majority had been diagnosed within 5 years (n=15, 62.5%). Treatment approaches included medication (n=13, 54.2%) and insulin therapy (n=8, 33.3%). Patient participant characteristics are summarized in Table 2.

XSL•FO
RenderX

Table 2. Demographic and clinical characteristics of 24 patient participants in part 1.

Characteristics	Participants, n (%)
Age range (y)	
20-29	8 (33.33)
30-39	5 (20.83)
40-49	3 (12.5)
50-59	4 (16.67)
60-69	4 (16.67)
Sex	
Female	18 (75)
Male	6 (25)
Chronic conditions	
Diabetes	9 (37.5)
Diabetes and other	16 (66.67)
Types of diabetes	
Type 1	7 (29.17)
Type 2	15 (62.5)
Unknown	2 (8.33)
Duration of diabetes (y)	
<1	4 (16.67)
1-5	11 (45.83)
5-10	4 (16.67)
>10	5 (20.83)
Diabetes treatment	
Medication	13 (54.17)
Insulin injections	8 (33.33)
Medication and insulin injections	1 (4.17)
No treatment	2 (8.33)

Most patients primarily relied on clinician consultations (n=19) and internet searches (n=18) for health information. Use of mobile health apps was less common (n=6), and only a few patients had previous experience with generative AI chatbots (n=3). These results are presented in Table 3.

Table 3. Health information-seeking behaviors and the use of generative artificial intelligence chatbots among 24 patient participants in part 1.

Category	Counts, n
Usual sources of health information	
Clinician consultation	19
Internet search	18
Mobile health app	6
Use of generative AI[a] chatbots for health information seeking	
Yes (used)	3
No (not used)	21

[a]AI: artificial intelligence.

Additionally, 4 family medicine specialists participated in part 2 of the study (2 female, 2 male). Out of 4, 2 had more than 10 years of clinical experience. All were board-certified and specialized in diabetes care at academic medical centers. Specialist characteristics are detailed in Table 4.

XSL•FO
RenderX

Table 4. Demographic and clinical characteristics of 4 family medicine specialists in part 2.

Specialist	Age range (years)	Sex	Specialty	Clinical experience (y)
E1	30-39	Female	Family medicine	9
E2	40-49	Male	Family medicine	20
E3	30-39	Female	Family medicine	9
E4	40-49	Male	Family medicine	10

Types of Questions and Information Needs Identified From DTalksBot Interactions (RQ 1)

Overview

A total of 643 questions were collected from 24 patients with diabetes. Using an affinity diagram, we identified common question types and organized them into 4 themes. Although sessions 1 and 4 were open-ended and sessions 2 and 3 focused on specific topics, many queries ultimately reflected core concerns in diabetes management. Patients often asked similar questions across sessions regardless of the prompt. For example, questions about complications or treatment options appeared in both guided and free-form sessions. This consistency suggested that patients' information needs remained stable across contexts. Consequently, we integrated data from all 4 sessions for a more comprehensive analysis, thereby enabling inclusive thematic and content analysis and providing deeper insight into how patients prioritize aspects of diabetes self-management.

We identified four themes: (1) personalized health advice and guidance, including lifestyle and diet questions (n=281, 44.63%); (2) concerns about complications and comorbidities, reflecting proactive risk management (n=174, 27.06%); (3) medication and treatment exploration, covering conventional, alternative, and conflicting treatment advice (n=111, 17.26%); and (4) mental health management and support, addressing psychological and emotional aspects of diabetes (n=30, 4.67%). Representative questions and DTalksBot responses, organized by theme, are provided in Multimedia Appendix 4.

Theme 1: Personalized Health Advice and Guidance

Analysis of patient interactions with the DTalksBot revealed that patients frequently sought both general health advice and highly personalized guidance. Patients valued practical insights on daily self-management, such as dietary choices, exercise routines, and medication timing. Patients posed questions such as:

> *My blood sugar level slightly exceeds 220 after meals. Can you provide medical advice on this?* [P8]
>
> *I've been using an insulin pump for six years and my average blood sugar level has been stable at 6.3 for five years. Can I stop using the pump and switch to medication instead?* [P3]

These questions demonstrate that patients used generative AI chatbots not only for general knowledge but for responses tailored to their health profiles. By sharing personal health metrics, they sought feedback customized to their physiological conditions and daily routines, rather than receiving generic suggestions. This reflected a tendency to use generative AI chatbots as interactive tools for self-management, particularly when seeking guidance aligned with personal medical history and real-time data.

Theme 2: Concerns About Complications and Comorbidities

Many patients expressed concerns about diabetes-related complications and comorbidities. They frequently asked about the types, prevention, diagnosis, and treatment of complications. For instance:

> *How often should regular check-ups be conducted for managing complications?* [P1]

Several patients described symptoms and inquired whether they were related to complications. For example:

> *Is frequent eye redness related to diabetes complications?* [P5]
>
> *Sometimes my toes tingle. Is this related to my blood sugar levels?* [P12]

These questions illustrated that patients actively monitored bodily changes and sought to determine whether they indicated complications. This highlighted the role of generative AI chatbots in helping patients interpret symptoms and make informed self-management decisions.

In addition, some patients showed interest in how diabetes interacts with other chronic conditions. For instance:

> *Does hypertension affect blood sugar control? I'm wondering if treating hypertension will help with my diabetes.* [P8]
>
> *I have hepatitis B and diabetes. What complications should I be aware of when managing both conditions?* [P5]

These inquiries showed how patients with multiple chronic conditions seek guidance on how managing one condition affects another. This underscored the need for holistic, comorbidity-aware advice and highlighted the potential of generative AI chatbots to support proactive and whole-person diabetes care.

Theme 3: Medication and Treatment Exploration

Questions related to medication and treatment represented the third most frequently asked category, emphasizing patients' active efforts to verify, compare, and interpret information about therapies and health metrics. This theme reflected patients' interest in understanding the safety, efficacy, and relevance of various treatments—both medical and alternative—in the context of diabetes management. Many inquiries centered on

XSL•FO
RenderX

the credibility of nontraditional or supplemental treatments. For example:

> *I heard that drinking a lot of corn water can completely cure type 1 diabetes. Is that true?* [P17]
>
> *Is it true that drinking vinegar water helps with blood sugar management?* [P18]

Patients often turned to DTalksBot to assess the validity of unverified claims encountered online or recommended by nonmedical sources. These questions illustrated how generative AI chatbots can serve as a trusted checkpoint for evaluating non–evidence-based information in everyday contexts.

Additionally, some patients expressed concerns about conflicting professional guidance. For instance:

> *My doctor recommended that I wear a continuous glucose monitor and advised me to keep my blood sugar level below 6.3. However, many doctors overseas consider a level up to 7.0 as normal. Should I follow my doctor's recommendation of maintaining 6.3, or is 7.0 an acceptable target based on international standards?* [P3]

Such inquiries suggested that patients frequently encounter discrepancies in clinical recommendations and use generative AI chatbots to seek clarification. These interactions highlighted the potential of generative AI chatbots to mediate—not to override medical advice, but to help patients navigate uncertainty and make informed decisions amid conflicting opinions.

Theme 4: Mental Health Management and Support

The psychological and emotional burdens of managing diabetes were commonly reported, with many patients seeking strategies to maintain mental resilience and to cope with the unique stressors of their condition [76]. This theme illustrated a consistent need for mental health support as patients adjusted to lifestyle changes, dietary restrictions, and the broader impacts of diabetes on daily life.

Patients expressed curiosity about how daily stress affects blood sugar levels and overall diabetes outcomes. For example:

> *How does stress affect blood sugar levels?* [P8]
>
> *Can stress be a cause of diabetes?* [P13]

These questions reflected patients' desire to understand how stress might influence the onset or progression of diabetes.

Beyond general curiosity, some patients expressed concern about the behavioral consequences of stress, including emotional eating and loss of motivation. For example:

> *When I'm stressed, I find myself craving sweets. Can you recommend effective stress-relief methods for diabetes patients?* [P15]

Such inquiries underscored the need for practical, empathetic guidance to manage stress-related behaviors.

Others sought more specific coping strategies tailored to their psychological challenges. For instance:

> *How can I best manage stress related to living with type 1 diabetes?* [P14]
>
> *Please provide specific stress management strategies for individuals with diabetes.* [P23]

These questions highlighted the long-term mental strain associated with chronic disease management and a demand for personalized mental health support. This aligned with the concept of individualized care discussed in theme 1 and reinforced the value of emotional guidance in digital health interventions.

User Experience of DTalksBot (RQ 2)

Participants evaluated their experience with DTalksBot through a postinteraction survey using a 5-point Likert scale. Usability scored a mean of 4.06 (SD 0.78), reflecting that the system was generally easy to use. Message credibility received a mean of 4.08 (SD 0.86), suggesting that users trusted the accuracy of the information provided.

Emotional support was rated at 3.42 (SD 0.79), indicating a moderate level of perceived support, while empathy scored 3.08 (SD 0.93), reflecting a modest degree of emotional resonance. Trust in DTalksBot received a mean of 3.69 (SD 0.99), showing a moderate level of confidence in its performance.

Expectations were rated at 4.06 (SD 0.95), indicating generally high hopes for its utility. Satisfaction scored 3.94 (SD 0.88), suggesting that participants were largely satisfied with their experience. Overall, participants rated all dimensions above the midpoint, with relatively higher scores for usability, credibility, and expectations, and lower—but still positive—ratings for emotional support and empathy.

Comparison of Chatbot Interactions With Online Resources and Health Care Providers (RQ 3)

Overview

This section examined how chatbot-based information-seeking compared with using search engines or online communities, and how generative AI chatbot interactions contrasted with in-person consultations with health care providers. We focused on how perceptions of immediacy, credibility, and emotional safety differed across these sources.

Comparison of Online Health Information Seeking

Patients traditionally relied on search engines like Google or online communities to obtain health information. While these sources provided a broad range of experiences and resources, they also presented notable limitations. First, unlike generative AI chatbots, search engines typically require users to spend considerable time sifting through extensive lists of results and assessing each source for relevance and accuracy. For example:

> *On portal sites, related content for the searched term appears in a long list, taking a lot of time to find the necessary information, but the chatbot provides immediate answers to questions.* [P5]

These insights highlighted the efficiency and accessibility of generative AI chatbots. By delivering curated responses instantly, generative AI chatbots could reduce the cognitive and time burden typically involved in online searches.

XSL•FO
RenderX

Second, while online communities offered insights grounded in lived experiences, participants often found user-generated content to be inconsistent and difficult to verify. For instance:

> *In online communities, many people share their blood sugar management know-how based on their experiences, so I've gained a lot of information there. But sometimes, it's hard to tell if what someone says applies to everyone, so it would be helpful to verify it through the chatbot.* [P10]

This comparison suggested that generative AI chatbots were perceived as more dependable sources of information. Their consistent filtering of content helped minimize the effort required to evaluate diverse sources, positioning them as preferred tools for reliable guidance.

Additionally, the generative AI chatbots were appreciated for delivering condition-specific guidance. Since online portals predominantly catered to type 2 diabetes, patients with type 1 diabetes often felt underserved [3,8,9]. For instance:

> *There's so much varied information on portals, but the chatbot only gives me diabetes-related information, which makes it feel more tailored to what I actually need.* [P19]

> *There's so much information on Type 2 diabetes on the portal sites, but the chatbot gave me advice that actually fit my Type 1 needs.* [P14]

In summary, patients highlighted the limitations of traditional online resources, such as search engines and online communities, in their health information–seeking process. Search engines often required significant time and effort to filter through extensive and inconsistent content, while online communities provided unverified user perspectives. In contrast, the generative AI chatbots were valued for their ability to deliver accurate, professionally verified, and diabetes-specific information, offering a more efficient and reliable alternative tailored to individual needs.

Comparison of Conversations With Clinicians

Patients with diabetes typically visit clinicians every 2 to 6 months for consultations and medication prescriptions. Comparing these routine medical visits with interactions using generative AI chatbots revealed distinct strengths in each approach. About half of the participants noted little difference between the questions they asked clinicians and those posed to the generative AI chatbots, reporting similar responses. However, the other half identified distinct differences, especially regarding the complexity and personalization of their inquiries.

Patients often preferred generative AI chatbots to verify general information, clarify existing knowledge, or address questions they had forgotten to ask during consultations. Generative AI chatbots offered a more relaxed environment, allowing participants to ask questions freely without concerns about time constraints, social discomfort, or disrupting the clinicians' schedule. For instance:

> *I can ask the chatbot any question, anytime, without worrying about the doctor's schedule or other waiting patients.* [P15]

> *Sometimes I get nervous and forget to ask questions during my consultation, so it's nice to have the chatbot to ask follow-up questions I remember later.* [P24]

These insights suggested that patients viewed generative AI chatbots as valuable complements to medical visits, particularly for minor, follow-up, or lingering questions.

Patients also valued the freedom to raise sensitive or awkward questions that they might hesitate to raise with clinicians. For example:

> *Sometimes, due to time constraints, I can't ask all my questions, and some feel a bit awkward to bring up directly. With the chatbot, I can ask about various topics freely, even those that might feel uncomfortable in person.* [P20]

This nonjudgmental setting encouraged more open discussions, suggesting that chatbots could address the everyday uncertainties common in chronic care.

In contrast, patients preferred to leave complex, data-specific inquiries, such as interpreting laboratory results, adjusting medication, and modifying insulin dosages, to their clinicians. These interactions required clinical expertise and access to personal medical histories. For example:

> *Since my doctor has all my blood sugar records and a history of my medications, I feel more comfortable asking specialized questions related to my specific data.* [P20]

This contrast highlighted a complementary role for generative AI chatbots in diabetes management. While generative AI chatbots offered ongoing, accessible support between visits, high-stakes decisions remain within the domain of clinicians.

Overall, these findings indicated that although generative AI chatbots could not replace professional medical advice, they might effectively support daily diabetes management. By providing reliable, immediate health information and reducing the cognitive load of self-management, generative AI chatbots could facilitate a more balanced and holistic approach—reinforcing rather than replacing the clinicians' role.

Desired Roles and Perceived Risks of Generative AI Chatbots (RQ 4)

Overview

The interviews explored patients' expectations and desired roles for generative AI chatbots in diabetes management. As most patient participants had no previous experience with generative AI chatbots, this study offered a rare opportunity for direct, hands-on interaction with DTalksBot. Their feedback, grounded in actual interactions rather than theoretical concepts or superficial impressions, provided practical insights into both the potential roles and perceived limitations of these tools in diabetes management.

Evidence-Based and Reliable Responses

Patients consistently highlighted the value of generative AI chatbots as efficient tools for health information–seeking,

XSL•FO
RenderX

emphasizing their ability to provide reliable, specific, and actionable information. Trustworthiness and accuracy emerged as critical features, particularly considering the overwhelming and often conflicting nature of diabetes-related content online. For example:

> *The most important thing for a chatbot is providing accurate information.* [P7]
>
> *It provides only the information that I need, offering specific and reliable responses.* [P11]

Patients clearly preferred generative AI chatbots that provided accurate, reliable, and tailored health information, and expressed that these qualities directly influenced their sense of trust.

Patients also emphasized the importance of evidence-based recommendations to enhance credibility. For instance:

> *If the chatbot could summarize recent treatments, medications, or research findings and provide the sources, it would significantly increase trust.* [P1]

This feedback underscored a critical demand for medical accuracy and source transparency, suggesting that generative AI chatbots needed to be anchored in up-to-date and verifiable information to be perceived as trustworthy tools.

Actionable and Personalized Guidance

Beyond general information-seeking, patients preferred practical guidance that aligned with their daily routines and real-world circumstances. For example:

> *The chatbot's responses include easy-to-follow and tasty diet recommendations.* [P5]
>
> *The chatbot provided realistic management advice beyond what is typically found in books.* [P6]

Such comments reflected a desire for recommendations that could be easily integrated into daily life, particularly with respect to diet, exercise, and medication adherence.

A recurring theme in the interviews was the expectation for personalized and context-specific guidance. For example:

> *It would be great if the chatbot could understand my current health condition and provide personalized responses.* [P4]

These expectations highlighted the value of generative AI chatbots that went beyond generalized advice to deliver tailored and actionable guidance aligned with patients' unique health conditions and self-management goals.

Limitations and Perceived Risks

However, patients also voiced concerns about potential risks, particularly in areas requiring critical health decisions. For example:

> *Discussions about adjusting medications based on data should probably be done with a physician.* [P20]
>
> *Diagnosing and treating complications like ketoacidosis requires precision and should involve a physician.* [P22]

These comments reflected patients' pragmatic awareness of the limitations of AI tools. While generative AI chatbots could serve as supplementary resources, patients emphasized that they could not replace professional medical judgment, especially in situations that demand clinical expertise and high-stakes decision-making.

In summary, patients viewed generative AI chatbots as promising tools for efficient and accurate health information-seeking, particularly for clarifying complex or conflicting advice and delivering actionable guidance. However, they also recognized the importance of safeguards to ensure that generative AI chatbots served to complement rather than replace professional care, especially in contexts involving diagnostic accuracy and critical treatment decisions.

Expert Perspectives on the Role and Limitations of Generative AI Chatbots in Diabetes Management (RQ 5)

Overview

This section presents insights from family medicine specialists who qualitatively assessed the clinical relevance, accuracy, and limitations of DTalksBot in the context of diabetes management. Additionally, their perspectives offered professional reflections on the potential roles of generative AI chatbots in supporting patients' self-management and enhancing routine clinical workflows.

The Accuracy and Challenges of DTalksBot in Responding to Diabetes-Related Queries

DTalksBot was generally able to provide safe and appropriate guidance in high-risk scenarios requiring medical intervention. When asked about treatment adjustments, it consistently advised users to consult a clinician rather than provide a definitive answer. Specialists interpreted this approach as a strength, as it minimized the risk of misinformation and aligned with real-world clinical practice. Specialists noted that deferring to medical professionals in uncertain contexts reflects an appropriate and responsible strategy for health-focused generative AI chatbots.

> *For questions about diabetes medication, the chatbot consistently advised patients to consult their physician, which is the appropriate approach given the potential risks involved.* [E1]

However, DTalksBot lacked sufficient personalization in tailoring its responses to individual patient characteristics. While its dietary and lifestyle recommendations were generally evidence-based, specialists pointed out that it failed to consider comorbid conditions such as renal dysfunction or body weight variations that require modified guidance. For instance, although DTalksBot often encouraged a protein-rich diet, it failed to recognize the potential risks for those with kidney disease.

> *Typically, diabetic patients are advised to consume more protein than carbohydrates. However, for those with kidney dysfunction, protein intake must be restricted. The chatbot did not consider such individual variations.* [E2]

Finally, specialists observed that DTalksBot occasionally provided overly detailed or tangential information that reduced

XSL•FO
RenderX

the clarity and relevance of its responses. This tendency was particularly evident when DTalksBot elaborated on loosely related topics. Such overexplaining was seen as potentially confusing for patients, as it introduced unnecessary complexity into otherwise straightforward inquiries. One expert recalled a case in which a question about corn silk water triggered an unrelated discussion about smoking and hypoglycemia.

> *The chatbot could have simply said 'No' to a question about the effects of corn silk water, but instead it brought up smoking and hypoglycemia, which made the answer more confusing.* [E1]

In summary, DTalksBot demonstrated a relatively high level of accuracy and a cautious approach to clinical questions. However, its limitations in personalization and contextual precision pointed to areas in need of further refinement. These findings underscore the importance of developing generative AI chatbots that not only provide accurate medical information but also adapt to the diverse needs and characteristics of individual patients.

Potential of Generative AI Chatbots in Diabetes Management

After reviewing DTalksBot's interaction logs and using the system firsthand, the specialists also shared their views on how generative AI chatbots could be used in diabetes management. Specialists agreed that generative AI chatbots could play a meaningful role in supporting diabetes self-management by addressing routine questions and clarifying treatment-related information between medical visits. Emphasizing the importance of lifestyle management alongside medical treatments for patients with diabetes, E2 noted that continuous lifestyle adjustments are essential to effective diabetes care:

> *Lifestyle management is as crucial as medication for diabetes. A chatbot could provide ongoing support and answer questions between doctor visits, helping patients manage their condition more comfortably.* [E2]

This perspective suggested that generative AI chatbots could supplement routine care by offering consistent, reliable answers to questions about diet, exercise, and blood sugar control, thus reinforcing adherence to long-term management goals.

Furthermore, regarding alleviating provider workload, specialists viewed generative AI chatbots as useful for addressing lifestyle-related questions that might otherwise consume valuable consultation time. As E2 observed:

> *While we handle medication, the chatbot could manage lifestyle aspects like diet, creating a synergistic effect.* [E2]

This reflected a complementary division of roles, allowing clinicians to focus on clinical decisions while generative AI chatbots address everyday concerns. Such delineation could enhance consultation efficiency and promote better health outcomes through comprehensive and continuous support.

Additionally, specialists noted that generative AI chatbots could help organize patient inquiries and summarize their concerns to support continuity of care. As E1 suggested:

> *Many patients bring notebooks filled with questions they couldn't address during their last visit, but if the chatbot could help keep track of these inquiries, it would help me as a doctor review and address these concerns in the next appointment.* [E1]

By identifying patterns in patient queries, generative AI chatbots could support a more cohesive approach to care and could enrich consultations by offering patients and clinicians a clearer understanding of ongoing needs and challenges.

Challenges of Generative AI Chatbots in Diabetes Management

While specialists were cautiously optimistic about the generative AI chatbots' potential to enhance patient engagement, they noted that long-term behavior change—especially in areas such as diet and exercise—remains a significant challenge. E2 and E3 remarked:

> *Changing behavior through advice is difficult; thus, psychological well-being might be the main benefit of the chatbot.* [E2]

> *Behavior change through a chatbot is difficult. However, if the chatbot assists with practical tasks like scheduling doctor appointments or medication reminders, it could build rapport and aid in cognitive-behavioral therapy.* [E3]

These insights suggested that the primary role of generative AI chatbots might lie in maintaining motivation and supporting psychological well-being, rather than directly driving behavioral change. Future iterations could benefit from incorporating features that prioritize motivational support to foster long-term health engagement.

Specialists also noted limitations in the ability of generative AI chatbots to provide personalized advice. Due to the absence of patient-specific data, generative AI chatbots were not able to consistently account for critical variables like age, comorbidities, or treatment history, which were essential to individualized diabetes management. E3 highlighted this limitation, stating:

> *If an elderly patient uses the chatbot and it consistently advises that an average blood sugar level of 6.5 is normal, this could be dangerous. Important variables like age should be considered when providing health advice.* [E3]

These challenges suggested that while generative AI chatbots hold promise, further refinement was necessary for them to serve as effective tools for personalized diabetes management. Enhanced capabilities to interpret individual health metrics and respond to condition-specific factors could improve their utility across diverse patient populations.

Finally, specialists emphasized the importance of transparent data sources and explainable responses. They proposed that chatbots rely on curated, verified databases to ensure reliability and relevance in clinical contexts. As E4 commented:

> *ChatGPT has improved and provides accurate answers, but we don't know the data it was trained on, posing a risk of incorrect answers. The chatbot should be limited to a verified database like*

XSL•FO
RenderX

DTalksBot to ensure relevance and accuracy in a domestic context. [E4]

This insight underscored a critical direction for future development—enhancing transparency and data fidelity to foster patient trust and encourage integration of generative AI chatbots into routine diabetes management.

Discussion

Principal Results

This study explored how generative AI chatbots could support diabetes management by addressing 5 key RQs: identifying the types of questions patients frequently ask and the information they seek (RQ 1), evaluating patients' experiences with DTalksBot (RQ 2), comparing interactions with generative AI chatbots to online health information seeking and clinical consultations (RQ 3), investigating patients' expectations for the use of generative AI chatbots in diabetes management (RQ 4), and analyzing specialists' evaluations of DTalksBot and their perspectives on the broader role of generative AI chatbots in diabetes management (RQ 5).

As a part of RQ 1, we identified four primary themes reflecting the diverse needs of diabetes patients: (1) personalized health advice and guidance, (2) concerns about complications and comorbidities, (3) medication and treatment exploration, and (4) mental health management and support. These findings align with previous research, indicating that diabetes patients frequently seek online information about diet, exercise, complications, and medication [14,53]. In other words, the topics commonly explored through online searches also emerged frequently in DTalksBot interactions. However, patients tended to request more personalized information and emotional support from generative AI chatbots than from traditional online health-seeking behavior. Accordingly, future generative AI chatbots may benefit from incorporating not only personalized guidance and complication-related information but also stress management support as an integral component of diabetes management.

For RQ 2, patients generally perceived DTalksBot as usable, credible, and satisfying. High usability scores suggested that diabetes patients found the interface intuitive and accessible [27,28]. Strong message credibility further underscored the DTalksBot's ability to deliver reliable information, reinforcing participants' willingness to engage. While emotional support and empathy scores were moderate, this reflects a broader challenge for generative AI chatbots: difficulty in providing human-like empathy in health care contexts [32]. Nonetheless, enhancing empathetic conversation design could significantly influence chronic disease self-management behaviors [28].

For RQ 3, patients identified 2 key advantages of generative AI chatbots over traditional sources: mitigating information overload and providing a judgment-free environment. Patients described DTalksBot's responses as not only trustworthy—grounded in validated sources through its RAG-based architecture—but also practical and actionable, enabling them to apply health guidance immediately in their daily routines. This combination of immediacy, safety, and utility illustrates how generative AI chatbots, when built on reliable data and domain-specific design, can effectively reduce the cognitive burden associated with navigating fragmented and inconsistent online health resources [24]. Furthermore, patients felt more comfortable asking sensitive or awkward questions—such as those related to unverified folk remedies—which they might hesitate to raise with clinicians due to fear of embarrassment. This aligns with theme 3 in RQ 1, where several patients asked about alternative or nonstandard treatments—topics they perceived as inappropriate for clinical settings but freely discussed with the chatbot. Such open dialogue reinforces the value of generative AI chatbots as complementary tools to formal care by fostering patient empowerment and supporting holistic chronic disease management. This finding also echoes previous research showing that chatbots promote open disclosure by removing social barriers to communication [29,53,77].

In RQ 4, 2 key insights emerged. First, participants emphasized the need for transparent and evidence-based responses, suggesting that citing medical guidelines or recent research could enhance credibility. Second, patients emphasized a preference for practical, personalized guidance, especially for meal planning and blood sugar management, reinforcing the value of user-centered and context-specific recommendations. Nevertheless, patients remained cautious about relying on generative AI chatbots for critical health decisions. This underscores the need to clearly define the role of generative AI chatbots as adjuncts to—not replacements for—professional care, ensuring that complex decisions stay under clinician supervision.

In RQ 5, family medicine specialists highlighted the complementary role of generative AI chatbots in diabetes care. They noted that generative AI chatbots could handle routine lifestyle queries, freeing clinicians to prioritize complex clinical decisions. Specialists also appreciated the potential of generative AI chatbots to bridge communication gaps, enabling patients to revisit concerns outside formal consultations. However, limitations included insufficient personalization for comorbid patients and the absence of detailed emergency guidance. Experts further recommended that future generative AI chatbots proactively seek clarification for vague queries and direct users to authoritative resources when necessary.

Overall, both patients and specialists recognized generative AI chatbots as valuable adjuncts in daily diabetes management. Patients envisioned on-demand support for daily self-management, including diet, glucose control, and emotional well-being, while specialists underscored the potential of generative AI chatbots to enhance consultation efficiency by addressing routine lifestyle queries. Importantly, patients valued the ability to deliver curated and trustworthy information more efficiently than traditional online sources, emphasizing the importance of reliable data in AI-driven health communication. Nevertheless, both stakeholder groups agreed that the generative AI chatbots are not intended to replace professional medical judgment or individualized clinical decision-making.

While patients expressed a strong desire for individualized, real-time guidance, specialists raised concerns about the risks

XSL•FO
RenderX

of unsupervised hyperpersonalization—particularly in complex or ambiguous clinical contexts. These diverging perspectives underscore the need to define the appropriate boundaries of chatbot involvement. Rather than seeking to maximize chatbot autonomy, this study focused on identifying the types of support that can be safely handled by AI and those that require professional oversight. Conducted with individuals managing diabetes, this study not only demonstrates the feasibility of using generative AI chatbots in everyday health management but also offers foundational insights into the appropriate and inappropriate roles of generative AI chatbots in the context of patient-centered care. In doing so, it contributes to the early empirical groundwork necessary for establishing safe, effective, and ethically aligned roles for AI systems in chronic disease management.

Future Directions

Integrating generative AI chatbots into routine health care offers a promising opportunity to improve accessibility by providing timely, tailored, and reliable support. Although this study centers on diabetes, the proposed directions may be applicable to a broader range of chronic diseases requiring ongoing behavioral guidance. Drawing from this study's findings, we outline several design and research directions to enhance personalization, trustworthiness, emotional engagement, and clinical integration.

First, our findings demonstrated that domain-specific generative AI chatbots can mitigate many of the risks associated with opaque inference mechanisms and biased training data that characterize general-purpose systems [31,37-45]. As observed with DTalksBot, grounding responses exclusively in a curated, evidence-based knowledge base from verified medical sources enabled transparent reasoning and traceable information flow. This design minimized the potential for systematic bias and reduced the likelihood of disseminating inaccurate or misleading medical advice, a known concern in widely used chatbots such as ChatGPT and Gemini [37,46,47]. By restricting outputs to validated diabetes-specific content and clearly declining unsupported queries, DTalksBot illustrates how transparency in data provenance and scope control can enhance both safety and trustworthiness in AI-driven health communication.

Second, enhancing personalization through structured data integration may serve as a foundational design strategy. To tailor responses more effectively, future systems may benefit from integration with electronic health records (EHRs) and continuous glucose monitoring (CGM) data [78]. These integrations would enable generative AI chatbots to deliver personalized recommendations informed by historical health data. For instance, linking CGM data could enable chatbots to anticipate glucose fluctuations and offer proactive behavioral guidance, such as dietary and activity suggestions. Such features could be particularly beneficial for daily self-management, especially for patients with intensive insulin regimens.

While such data integrations enhance personalization, they also raise critical concerns regarding data privacy and user trust, which must be carefully addressed in future system design. Privacy-related behaviors are often shaped by privacy calculus theory, which posits that individuals weigh perceived benefits against risks when deciding whether to share personal data [79]. In practice, this can result in privacy paradoxes, where users express concern but still disclose data when they perceive high utility [80]. These dynamics suggest that trust and perceived control may influence user behavior more than technical safeguards alone. Accordingly, chatbot systems should prioritize not only data security, but also designs that promote user agency. One promising direction is the evaluation of data governance models—such as centralized versus decentralized architectures—that offer distinct trade-offs in transparency, control, and systemic risk [81].

Building on these considerations, caution is warranted regarding the extent of personalization and the privacy implications of data integration. While tailored recommendations can enhance user engagement, hyperpersonalized advice may pose clinical and regulatory risks. Instead, chatbots could provide guidance tailored to broader user groups (eg, age or comorbidity profiles) to balance relevance and safety. Additionally, compliance with regulations such as the Health Insurance Portability and Accountability Act in the United States and the General Data Protection Regulation in the European Union necessitates secure, transparent, and consent-driven data-sharing frameworks [82].

Third, ensuring the credibility and contemporaneity of chatbot responses requires robust information management and knowledge updating mechanisms. Advanced natural language processing and RAG models may enhance the chatbot's ability to interpret user input and generate relevant responses [71,72]. Future generative AI chatbots will need to integrate diverse health care databases encompassing diabetes and comorbidities, thereby supporting holistic assessments [83]. Additionally, dynamic learning architectures are necessary to ensure content accuracy. Continuous updates from verified medical sources, including treatment protocols and new research findings, could help maintain chatbot relevance [30,66,84]. Ideally, these updates would be validated through expert-reviewed backend systems [85].

Fourth, addressing the emotional and cognitive needs of patients is essential in chronic disease management. Incorporating psychological models into generative AI chatbot design could enhance emotional support and behavioral guidance. Techniques based on cognitive behavioral therapy, combined with sentiment analysis, could allow chatbots to recognize and respond empathetically to users' affective states [86,87]. While participants perceived the chatbot as moderately empathic, integrating such models could strengthen the chatbot not only as an information provider but also as a psychologically supportive companion [87].

Finally, generative AI chatbots are best positioned as collaborators within clinical workflows rather than autonomous decision-makers. Generative AI chatbots could alleviate the workload of health care providers by addressing routine patient inquiries, enabling clinicians to focus on more complex clinical decisions. Additionally, generative AI chatbots could act as intermediaries, collecting and summarizing patient inquiries between visits to improve consultation efficiency. Over time, generative AI chatbots may evolve into tools for long-term health management, tracking engagement, adherence, and health patterns while providing actionable insights [88]. To realize

XSL•FO
RenderX

this potential, it is essential to establish clearly defined frameworks for human-AI collaboration—particularly in chronic care settings, where continuity and coordination are critical. To fully integrate chatbots into clinical workflows, future research may benefit from defining appropriate scope boundaries—clarifying which types of user queries chatbots can safely address, and when they should defer to human clinicians. Establishing such thresholds is essential to ensure patient safety and preserve the integrity of professional care.

Limitations

Several limitations should be acknowledged. First, DTalksBot generated recommendations without access to real-time patient data. While it delivered general advice, future iterations may benefit from integrating real-time CGM data or EHR information to provide context-aware and patient-specific guidance.

Second, this study did not distinguish between type 1 and type 2 diabetes. Because these 2 groups differ in etiology, typical age of onset, and treatment requirements [1,3], subgroup analyses are needed to identify distinct information needs and interaction patterns.

Third, this study provides a cross-sectional perspective, which may not capture longitudinal trends. Longitudinal studies are required to determine whether engagement, satisfaction, and clinical impact are maintained over time and to observe how user expectations and system performance evolve with continued use.

Fourth, all participants were recruited in South Korea. Although diabetes self-management practice and digital health adoption are broadly similar to those in many other high-income countries, cultural differences, such as perceptions of mobile services, dietary practices, and levels of digital literacy, may influence the applicability of generative AI chatbots in other populations [5,89]. Future studies may benefit from examining the use of generative AI chatbots across more diverse sociocultural and health care contexts.

Conclusions

This study highlights the potential of generative AI chatbots to provide personalized guidance, trustworthy information, and holistic support for diabetes management. Both patients and specialists recognized these tools as valuable adjuncts to conventional health care, particularly for addressing routine inquiries, reinforcing self-management behaviors, and alleviating clinician workload. By demonstrating that a domain-specific generative AI chatbot grounded in verified medical sources can operate with transparency and minimal bias, this study offers a replicable model for safe AI deployment in health care.

Practical implications include guiding clinicians on integrating chatbots as complementary tools in chronic care workflows and informing developers on design strategies that enhance transparency, contextual relevance, and user trust. Future research would benefit from investigating the long-term impact of generative AI chatbots on clinical outcomes, explore integration with real-time biometric and EHR data, and evaluate their applicability across diverse chronic conditions and health care systems.

Acknowledgments

We thank all participants and the College of Social Sciences at Seoul National University for their support of this research.

This work was supported by the Seoul National University - Global Excellence Research Center establishment project and was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT; NO. RS-2021-II211343, Artificial Intelligence Graduate School Program [Seoul National University]). The funders had no involvement in the study design, data collection, analysis, interpretation of results, or the writing of the manuscript.

Generative AI tools were occasionally referenced for proofreading purposes. Their suggestions were used only as guidance, and all final wording, analysis, and interpretations were determined by the authors.

Data Availability

The data generated and analyzed during this study are not publicly available due to participant privacy concerns, but may be shared by the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: SJ, SL, EHK
Data curation: SJ, SL, EHK
formal analysis: SJ, SL, EHK
funding acquisition: SJ
investigation: SJ, SL, EHK
methodology: SJ, SL, EHK
project administration: SJ
software development and technical resources: JE, KL
Supervision: HL, JL

XSL•FO
RenderX

Validation: SJ, SL, EHK, HL, JL
Visualization: SL, EHK
Writing – original draft: SJ, SL, EHK
Writing – review and editing: HL, JL

Conflicts of Interest

None declared.

Multimedia Appendix 1
Standards for Reporting Qualitative Research (SRQR) checklist.
[DOC File , 54 KB - formative_v9i1e72553_app1.doc]

Multimedia Appendix 2
Survey Items for Evaluating the Chatbot.
[DOCX File , 19 KB - formative_v9i1e72553_app2.docx]

Multimedia Appendix 3
Screenshots and English translations of participants' chat logs.
[DOCX File , 4336 KB - formative_v9i1e72553_app3.docx]

Multimedia Appendix 4
Top 10 Representative Questions for Each Theme.
[DOCX File , 29 KB - formative_v9i1e72553_app4.docx]

References

1. Buse JB, Caprio S, Cefalu WT, Ceriello A, Del Prato S, Inzucchi SE, et al. How do we define cure of diabetes? Diabetes Care 2009;32(11):2133-2135 [FREE Full text] [doi: 10.2337/dc09-9036] [Medline: 19875608]
2. Zimmet PZ, Magliano DJ, Herman WH, Shaw JE. Diabetes: a 21st century challenge. Lancet Diabetes Endocrinol 2014;2(1):56-64. [doi: 10.1016/S2213-8587(13)70112-8] [Medline: 24622669]
3. Nathan DM. Diabetes: Advances in diagnosis and treatment. JAMA 2015;314(10):1052-1062. [doi: 10.1001/jama.2015.9536] [Medline: 26348754]
4. GBD 2021 Diabetes Collaborators. Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: a systematic analysis for the Global Burden of Disease Study 2021. Lancet 2023;402(10397):203-234 [FREE Full text] [doi: 10.1016/S0140-6736(23)01301-6] [Medline: 37356446]
5. Park SE, Ko S, Kim JY, Kim K, Moon JH, Kim NH, et al. Diabetes Fact Sheets in Korea 2024. Diabetes Metab J 2025;49(1):24-33 [FREE Full text] [doi: 10.4093/dmj.2024.0818] [Medline: 39828976]
6. Gregg EW, Cheng YJ, Srinivasan M, Lin J, Geiss LS, Albright AL, et al. Trends in cause-specific mortality among adults with and without diagnosed diabetes in the USA: an epidemiological analysis of linked national survey and vital statistics data. Lancet 2018;391(10138):2430-2440 [FREE Full text] [doi: 10.1016/S0140-6736(18)30314-3] [Medline: 29784146]
7. Pylypchuk R, Wells S, Kerr A, Poppe K, Harwood M, Mehta S, et al. Cardiovascular risk prediction in type 2 diabetes before and after widespread screening: a derivation and validation study. Lancet 2021;397(10291):2264-2274. [doi: 10.1016/S0140-6736(21)00572-9] [Medline: 34089656]
8. Kalantzi S, Kostagiolas P, Kechagias G, Niakas D, Makrilakis K. Information seeking behavior of patients with diabetes mellitus: a cross-sectional study in an outpatient clinic of a university-affiliated hospital in Athens, Greece. BMC Res Notes 2015;8:48 [FREE Full text] [doi: 10.1186/s13104-015-1005-3] [Medline: 25889724]
9. Holt R, DeVries J, Hess-Fischl A, Hirsch IB, Kirkman MS, Klupa T, et al. The management of type 1 diabetes in adults. A consensus report by the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD). Diabetes Care 2021;44(11):2589-2625. [doi: 10.2337/dci21-0043] [Medline: 34593612]
10. Linekin PL. Home health care and diabetes assessment, care, and education. Diabetes Spectr 2003;16(4):217-222 [FREE Full text] [doi: 10.2337/diaspect.16.4.217]
11. Diabetes ControlComplications Trial Research Group, Nathan DM, Genuth S, Lachin J, Cleary P, Crofford O, et al. The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. N Engl J Med 1993;329(14):977-986. [doi: 10.1056/NEJM199309303291401] [Medline: 8366922]
12. Kirk JK, Stegner J. Self-monitoring of blood glucose: practical aspects. J Diabetes Sci Technol 2010;4(2):435-439 [FREE Full text] [doi: 10.1177/193229681000400225] [Medline: 20307405]
13. Jamal A, Khan SA, AlHumud A, Al-Duhyyim A, Alrashed M, Bin Shabr F, et al. Association of online health information-seeking behavior and self-care activities among type 2 diabetic patients in Saudi Arabia. J Med Internet Res 2015;17(8):e196 [FREE Full text] [doi: 10.2196/jmir.4312] [Medline: 26268425]

XSL•FO
RenderX

14. Kuske S, Schiereck T, Grobosch S, Paduch A, Droste S, Halbach S, et al. Diabetes-related information-seeking behaviour: a systematic review. Syst Rev 2017;6(1):212 [FREE Full text] [doi: 10.1186/s13643-017-0602-8] [Medline: 29065919]
15. Zhang X, Foo S, Majid S, Chang Y, Dumaual HTJ, Suri VR. Self-care and health-information-seeking behaviours of diabetic patients in Singapore. Health Commun 2020;35(8):994-1003. [doi: 10.1080/10410236.2019.1606134] [Medline: 31303050]
16. Hughes A, Heydarian N, Gerardo D, Solis I, Morera O. Seeking health information and social support in the diabetes online community. Front Clin Diabetes Healthc 2021;2:708405 [FREE Full text] [doi: 10.3389/fcdhc.2021.708405] [Medline: 36994327]
17. Zhao YC, Zhao M, Song S. Online health information seeking among patients with chronic conditions: integrating the health belief model and social support theory. J Med Internet Res 2022;24(11):e42447 [FREE Full text] [doi: 10.2196/42447] [Medline: 36322124]
18. Lee K, Hoti K, Hughes JD, Emmerton L. Dr Google and the consumer: a qualitative study exploring the navigational needs and online health information-seeking behaviors of consumers with chronic health conditions. J Med Internet Res 2014;16(12):e262 [FREE Full text] [doi: 10.2196/jmir.3706] [Medline: 25470306]
19. Eysenbach G, Köhler C. How do consumers search for and appraise health information on the world wide web? Qualitative study using focus groups, usability tests, and in-depth interviews. BMJ 2002;324(7337):573-577 [FREE Full text] [doi: 10.1136/bmj.324.7337.573] [Medline: 11884321]
20. Geana MV, Daley CM, Nazir N, Cully L, Etheridge J, Bledowski C, et al. Use of online health information resources by American Indians and Alaska Natives. J Health Commun 2012;17(7):820-835 [FREE Full text] [doi: 10.1080/10810730.2011.650831] [Medline: 22642739]
21. Cline R, Haynes K. Consumer health information seeking on the internet: the state of the art. Health Educ Res 2001;16(6):671-692. [doi: 10.1093/her/16.6.671] [Medline: 11780707]
22. Liu PL. Understanding mHealth adoption and lifestyle improvement: an empirical test of the antecedents and mediating mechanisms. J Health Commun 2024;29(2):155-165. [doi: 10.1080/10810730.2024.2303641] [Medline: 38205945]
23. Safi Z, Abd-Alrazaq A, Khalifa M, Househ M. Technical aspects of developing chatbots for medical applications: scoping review. J Med Internet Res 2020;22(12):e19127 [FREE Full text] [doi: 10.2196/19127] [Medline: 33337337]
24. Oh YJ, Zhang J, Fang M, Fukuoka Y. A systematic review of artificial intelligence chatbots for promoting physical activity, healthy diet, and weight loss. Int J Behav Nutr Phys Act 2021;18(1):160 [FREE Full text] [doi: 10.1186/s12966-021-01224-6] [Medline: 34895247]
25. Sharma D, Kaushal S, Kumar H, Gainder S. Chatbots in healthcare: challenges, technologies and applications. : IEEE; 2022 Presented at: 4th International Conference on Artificial Intelligence and Speech Technology (AIST); 2022 December 09-10; Delhi, India p. 1-6. [doi: 10.1109/aist55798.2022.10065328]
26. Hussain S, Athula G. Extending a conventional chatbot knowledge base to external knowledge source and introducing user-based sessions for diabetes education. : IEEE; 2018 Presented at: 32nd International Conference on Advanced Information Networking and Applications Workshops (WAINA); 2018 May 16-18; Krakow, Poland p. 698-703. [doi: 10.1109/waina.2018.00170]
27. Gong E, Baptista S, Russell A, Scuffham P, Riddell M, Speight J, et al. My diabetes coach, a mobile app-based interactive conversational agent to support type 2 diabetes self-management: randomized effectiveness-implementation trial. J Med Internet Res 2020;22(11):e20322 [FREE Full text] [doi: 10.2196/20322] [Medline: 33151154]
28. Laymouna M, Ma Y, Lessard D, Schuster T, Engler K, Lebouché B. Roles, users, benefits, and limitations of chatbots in health care: rapid review. J Med Internet Res 2024;26:e56930 [FREE Full text] [doi: 10.2196/56930] [Medline: 39042446]
29. Palanica A, Flaschner P, Thommandram A, Li M, Fossat Y. Physicians' perceptions of chatbots in health care: cross-sectional web-based survey. J Med Internet Res 2019;21(4):e12887 [FREE Full text] [doi: 10.2196/12887] [Medline: 30950796]
30. Walker HL, Ghani S, Kuemmerli C, Nebiker CA, Müller BP, Raptis DA, et al. Reliability of medical information provided by chatgpt: assessment against clinical guidelines and patient information quality instrument. J Med Internet Res 2023;25:e47479 [FREE Full text] [doi: 10.2196/47479] [Medline: 37389908]
31. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
32. Chew HSJ. The use of artificial intelligence-based conversational agents (chatbots) for weight loss: scoping review and practical recommendations. JMIR Med Inform 2022;10(4):e32578 [FREE Full text] [doi: 10.2196/32578] [Medline: 35416791]
33. Andrew A. Potential applications and implications of large language models in primary care. Fam Med Community Health 2024;12(Suppl 1):e002602 [FREE Full text] [doi: 10.1136/fmch-2023-002602] [Medline: 38290759]
34. Aggarwal A, Tam CC, Wu D, Li X, Qiao S. Artificial intelligence-based chatbots for promoting health behavioral changes: systematic review. J Med Internet Res 2023;25:e40789 [FREE Full text] [doi: 10.2196/40789] [Medline: 36826990]
35. Hwang M, Zheng Y, Cho Y, Jiang Y. AI applications for chronic condition self-management: scoping review. J Med Internet Res 2025;27:e59632 [FREE Full text] [doi: 10.2196/59632] [Medline: 40198108]
36. Jiang Z, Huang X, Wang Z, Liu Y, Huang L, Luo X. Embodied conversational agents for chronic diseases: scoping review. J Med Internet Res 2024;26:e47134 [FREE Full text] [doi: 10.2196/47134] [Medline: 38194260]

XSL•FO
RenderX

37. Morath B, Chiriac U, Jaszkowski E, Deiß C, Nürnberg H, Hörth K, et al. Performance and risks of ChatGPT used in drug information: an exploratory real-world analysis. Eur J Hosp Pharm 2024;31(6):491-497. [doi: 10.1136/ejhpharm-2023-003750] [Medline: 37263772]
38. Au Yeung J, Kraljevic Z, Luintel A, Balston A, Idowu E, Dobson RJ, et al. AI chatbots not yet ready for clinical use. Front Digit Health 2023;5:1161098 [FREE Full text] [doi: 10.3389/fdgth.2023.1161098] [Medline: 37122812]
39. Kim JK, Chua M, Rickard M, Lorenzo A. ChatGPT and large language model (LLM) chatbots: The current state of acceptability and a proposal for guidelines on utilization in academic medicine. J Pediatr Urol 2023;19(5):598-604. [doi: 10.1016/j.jpurol.2023.05.018] [Medline: 37328321]
40. Temsah M, Aljamaan F, Malki KH, Alhasan K, Altamimi I, Aljarbou R, et al. ChatGPT and the future of digital health: a study on healthcare workers' perceptions and expectations. Healthcare (Basel) 2023;11(13):1812 [FREE Full text] [doi: 10.3390/healthcare11131812] [Medline: 37444647]
41. Wang L, Wan Z, Ni C, Song Q, Li Y, Clayton E, et al. Applications and concerns of chatgpt and other conversational large language models in health care: systematic review. J Med Internet Res 2024;26:e22769 [FREE Full text] [doi: 10.2196/22769] [Medline: 39509695]
42. Zeng J, Zou X, Li S, Tang Y, Teng S, Li H, et al. Assessing the role of the generative pretrained transformer (GPT) in alzheimer's disease management: comparative study of neurologist- and artificial intelligence-generated responses. J Med Internet Res 2024;26:e51095 [FREE Full text] [doi: 10.2196/51095] [Medline: 39481104]
43. Armbruster J, Bussmann F, Rothhaas C, Titze N, Grützner PA, Freischmidt H. "Doctor ChatGPT, Can You Help Me?" The patient's perspective: cross-sectional study. J Med Internet Res 2024;26:e58831. [doi: 10.2196/58831] [Medline: 39352738]
44. Yau JY, Saadat S, Hsu E, Murphy LS, Roh JS, Suchard J, et al. Accuracy of prospective assessments of 4 large language model chatbot responses to patient questions about emergency care: experimental comparative study. J Med Internet Res 2024;26:e60291 [FREE Full text] [doi: 10.2196/60291] [Medline: 39496149]
45. Ahmad MA, Yaramis I, Roy TD. Creating trustworthy LLMs: dealing with hallucinations in healthcare AI. Preprints 2023 Oct 26 [FREE Full text] [doi: 10.20944/preprints202310.1662.v1]
46. Lim ZW, Pushpanathan K, Yew SME, Lai Y, Sun C, Lam JSH, et al. Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. EBioMedicine 2023;95:104770 [FREE Full text] [doi: 10.1016/j.ebiom.2023.104770] [Medline: 37625267]
47. Hulman A, Dollerup OL, Mortensen JF, Fenech ME, Norman K, Støvring H, et al. ChatGPT- versus human-generated answers to frequently asked questions about diabetes: a Turing test-inspired survey among employees of a Danish diabetes center. PLoS One 2023;18(8):e0290773 [FREE Full text] [doi: 10.1371/journal.pone.0290773] [Medline: 37651381]
48. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
49. Xiao Z, Liao Q, Zhou M, Grandison T, Li Y. Powering an AI chatbot with expert sourcing to support credible health information access. 2023 Presented at: IUI '23: Proceedings of the 28th International Conference on Intelligent User Interfaces; 2023 March 27 - 31; Sydney NSW Australia p. 2-18. [doi: 10.1145/3581641.3584031]
50. Nazi ZA, Peng W. Large language models in healthcare and medical domain: a review. Informatics 2024;11(3):57 [FREE Full text] [doi: 10.3390/informatics11030057]
51. Wang D, Liang J, Ye J, Li J, Li J, Zhang Q, et al. Enhancement of the performance of large language models in diabetes education through retrieval-augmented generation: comparative study. J Med Internet Res 2024;26:e58041 [FREE Full text] [doi: 10.2196/58041] [Medline: 39046096]
52. Dao D, Teo J, Wang W, Nguyen H. LLM-powered multimodal AI conversations for diabetes prevention. : Association for Computing Machinery; 2024 Presented at: ICMR '24: International Conference on Multimedia Retrieval; 2024 June 10; Phuket Thailand p. 1-6. [doi: 10.1145/3643479.3662049]
53. Sagstad MH, Morken N, Lund A, Dingsør LJ, Nilsen ABV, Sorbye LM. Quantitative user data from a chatbot developed for women with gestational diabetes mellitus: observational study. JMIR Form Res 2022;6(4):e28091 [FREE Full text] [doi: 10.2196/28091] [Medline: 35436213]
54. Wu Y, Zhang J, Ge P, Duan T, Zhou J, Wu Y, et al. Application of chatbots to help patients self-manage diabetes: systematic review and meta-analysis. J Med Internet Res 2024;26:e60380 [FREE Full text] [doi: 10.2196/60380] [Medline: 39626235]
55. van Vugt M, de Wit M, Cleijne WHJJ, Snoek FJ. Use of behavioral change techniques in web-based self-management programs for type 2 diabetes patients: systematic review. J Med Internet Res 2013;15(12):e279 [FREE Full text] [doi: 10.2196/jmir.2800] [Medline: 24334230]
56. Elnaggar A, Ta Park V, Lee SJ, Bender M, Siegmund LA, Park LG. Patients' use of social media for diabetes self-care: systematic review. J Med Internet Res 2020;22(4):e14209 [FREE Full text] [doi: 10.2196/14209] [Medline: 32329745]
57. Jette D, Grover L, Keck C. A qualitative study of clinical decision making in recommending discharge placement from the acute care setting. Phys Ther 2003;83(3):224-236. [doi: 10.1093/ptj/83.3.224]
58. Brooke J. SUS - A quick and dirty usability scale. In: Usability Evaluation in Industry. London: Taylor and Francis; 1996.

XSL•FO
RenderX

59. Appelman A, Sundar SS. Measuring message credibility: construction and validation of an exclusive scale. Journalism & Mass Communication Quarterly 2015;93(1):59-79. [doi: 10.1177/1077699015606057]
60. Zimet GD, Dahlem NW, Zimet SG, Farley GK. The multidimensional scale of perceived social support. Journal of Personality Assessment 1988;52(1):30-41. [doi: 10.1207/s15327752jpa5201_2]
61. Charrier L, Rieger A, Galdeano A, Cordier A, Lefort M, Hassas S. The RoPE scale: a measure of how empathic a robot is perceived. : IEEE; 2019 Presented at: 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI); 2019 March 11-14; Daegu, Korea (South) p. 656-657. [doi: 10.1109/hri.2019.8673082]
62. Komiak, Benbasat. The effects of personalization and familiarity on trust and adoption of recommendation agents. MIS Quarterly 2006;30(4):941-960 [FREE Full text] [doi: 10.2307/25148760]
63. Mokmin NAM, Ibrahim NA. The evaluation of chatbot as a tool for health literacy education among undergraduate students. Educ Inf Technol (Dordr) 2021;26(5):6033-6049 [FREE Full text] [doi: 10.1007/s10639-021-10542-y] [Medline: 34054328]
64. Borsci S, Malizia A, Schmettow M, van der Velde F, Tariverdiyeva G, Balaji D, et al. The chatbot usability scale: the design and pilot of a usability scale for interaction with AI-based conversational agents. Pers Ubiquit Comput 2021;26(1):95-119. [doi: 10.1007/s00779-021-01582-9]
65. Yang R, Tan TF, Lu W, Thirunavukarasu AJ, Ting DSW, Liu N. Large language models in health care: development, applications, and challenges. Health Care Sci 2023;2(4):255-263 [FREE Full text] [doi: 10.1002/hcs2.61] [Medline: 38939520]
66. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. ACM Comput. Surv 2023;55(12):1-38. [doi: 10.1145/3571730]
67. Goodman RS, Patrinely JR, Stone CA, Zimmerman E, Donald RR, Chang SS, et al. Accuracy and reliability of chatbot responses to physician questions. JAMA Netw Open 2023;6(10):e2336483 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.36483] [Medline: 37782499]
68. National health information portal. Korea Disease Control and Prevention Agency (KDCA). 2025. URL: https://health.kdca.go.kr/healthinfo/ [accessed 2025-10-22]
69. General diabetes information. Korean Diabetes Association. 2025. URL: https://www.diabetes.or.kr/general/ [accessed 2025-10-22]
70. Diabetes homepage. Centers for Disease Control and Prevention (CDC). 2025. URL: https://www.cdc.gov/diabetes/index.html [accessed 2025-10-22]
71. Asai A, Wu Z, Wang Y, Sil A, Hajishirzi H. Self-RAG: learning to retrieve, generate, and critique through self-reflection. arXiv:2310.11511 2023. [doi: 10.5860/choice.189890]
72. Gao Y, Xiong Y, Gao X. Retrieval-augmented generation for large language models: a survey. Computer Science 2024 [FREE Full text] [doi: 10.48550/arXiv.2312.10997]
73. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V. Retrieval-augmented generation for knowledge-intensive NLP tasks. Computer Science 2021 [FREE Full text] [doi: 10.48550/arXiv.2005.11401]
74. Braun V, Clarke V. Using thematic analysis in psychology. Qualitative Research in Psychology 2008;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
75. Guest G, Bunce A, Johnson L. How many interviews are enough? Field Methods 2006;18(1):59-82. [doi: 10.1177/1525822x05279903]
76. Baptista S, Wadley G, Bird D, Oldenburg B, Speight J, My Diabetes Coach Research Group. Acceptability of an embodied conversational agent for type 2 diabetes self-management education and support via a smartphone app: mixed methods study. JMIR Mhealth Uhealth 2020;8(7):e17038 [FREE Full text] [doi: 10.2196/17038] [Medline: 32706734]
77. Lucas GM, Gratch J, King A, Morency L. It's only a computer: virtual humans increase willingness to disclose. Computers in Human Behavior 2014;37:94-100. [doi: 10.1016/j.chb.2014.04.043]
78. Afsaneh E, Sharifdini A, Ghazzaghi H, Ghobadi MZ. Recent applications of machine learning and deep learning models in the prediction, diagnosis, and management of diabetes: a comprehensive review. Diabetol Metab Syndr 2022;14(1):196 [FREE Full text] [doi: 10.1186/s13098-022-00969-9] [Medline: 36572938]
79. Dinev T, Hart P. An extended privacy calculus model for e-commerce transactions. Information Systems Research 2006;17(1):61-80 [FREE Full text] [doi: 10.1287/isre.1060.0080]
80. Jackson CB, Wang Y. Addressing the privacy paradox through personalized privacy notifications. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol 2018;2(2):1-25. [doi: 10.1145/3214271]
81. Shubina V, Holcer S, Gould M, Lohan ES. Survey of decentralized solutions with mobile devices for user location tracking, proximity detection, and contact tracing in the COVID-19 era. Data 2020;5(4):87 [FREE Full text] [doi: 10.3390/data5040087]
82. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. Nat Med 2019;25(1):30-36 [FREE Full text] [doi: 10.1038/s41591-018-0307-0] [Medline: 30617336]
83. Roca S, Lozano ML, García J, Alesanco. Validation of a virtual assistant for improving medication adherence in patients with comorbid type 2 diabetes mellitus and depressive disorder. Int J Environ Res Public Health 2021;18(22):12056 [FREE Full text] [doi: 10.3390/ijerph182212056] [Medline: 34831811]

XSL•FO
RenderX

84. Hung TKW, Kuperman GJ, Sherman EJ, Ho AL, Weng C, Pfister DG, et al. Performance of retrieval-augmented large language models to recommend head and neck cancer clinical trials. J Med Internet Res 2024;26:e60695 [FREE Full text] [doi: 10.2196/60695] [Medline: 39405514]
85. Sachdeva B, Ramjee P, Fulari G, Murali K, Jain M. Learnings from a large-scale deployment of an LLM-powered expert-in-the-loop healthcare chatbot. Computer Science 2024 [FREE Full text] [doi: 10.48550/arXiv.2409.10354]
86. Rathnayaka P, Mills N, Burnett D, De Silva D, Alahakoon D, Gray R. A mental health chatbot with cognitive skills for personalised behavioural activation and remote health monitoring. Sensors (Basel) 2022;22(10):3653 [FREE Full text] [doi: 10.3390/s22103653] [Medline: 35632061]
87. Hamad O, Hamdi A, Shaban K. ASEM: enhancing empathy in chatbot through attention-based sentiment and emotion modeling. Computer Science 2024 [FREE Full text] [doi: 10.48550/arXiv.2402.16194]
88. Chen P, Li Y, Zhang X, Feng X, Sun X. The acceptability and effectiveness of artificial intelligence-based chatbot for hypertensive patients in community: protocol for a mixed-methods study. BMC Public Health 2024;24(1):2266 [FREE Full text] [doi: 10.1186/s12889-024-19667-4] [Medline: 39169305]
89. Kim H, Park S, Park J, Son Y, Kim S, Yim Y, et al. National trends in type 2 diabetes mellitus stratified by central adiposity using waist-to-height ratio in South Korea, 2005-2022. Sci Rep 2024;14(1):24273 [FREE Full text] [doi: 10.1038/s41598-024-75002-2] [Medline: 39414865]

Abbreviations

AI: artificial intelligence
CGM: continuous glucose monitoring
EHR: electronic health record
RAG: retrieval-augmented generation
RQ: research question

Edited by A Mavragani; submitted 12.02.25; peer-reviewed by N Misra, E Dal, T Ogunleye, C Br; comments to author 12.03.25; revised version received 20.06.25; accepted 29.09.25; published 12.11.25.

Please cite as:
Jeon S, Lee S, Kim EH, Eun J, Lee K, Lim H, Lee J
Generative AI Chatbot for Diabetes Management: Formative 2-Part Qualitative Study Using DTalksBot Involving Patients and Clinicians
JMIR Form Res 2025;9:e72553
URL: https://formative.jmir.org/2025/1/e72553
doi:10.2196/72553
PMID:

©Soyun Jeon, Seolhee Lee, Esther Hehsun Kim, Jinsu Eun, Kwangwon Lee, Hajin Lim, Joonhwan Lee. Originally published in JMIR Formative Research (https://formative.jmir.org), 12.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Medical-Economic and Ecological Impact of Anesthesia Teleconsultation: Retrospective Observational Study

Fabrice Ferre[1], MD, PhD; Philippine Furelau[1], MD; François Labaste[1], MD, PhD; Nadège Costa[1], MD; Fanny Vardon[1], MD, PhD; Antoine Piau[1], MD, PhD; Charlotte Martin[1], MD; Vincent Minville[1], MD, PhD

Centre Hospitalier Universitaire de Toulouse, Toulouse, France

Corresponding Author:
Vincent Minville, MD, PhD
Centre Hospitalier Universitaire de Toulouse
1 Ave Jean Poulhes
Toulouse, 31400
France
Phone: 33 5 61 32 27 91
Fax: 33 5 61 32 31 40
Email: minville.v@chu-toulouse.fr

Abstract

Background: Telemedicine, particularly teleconsultation, has emerged as a viable alternative to in-person consultation, especially following the COVID-19 pandemic. Preanesthetic consultations are mandatory before surgery to assess perioperative risk. However, little data exists regarding the combined economic and ecological impacts of replacing in-person consultation with teleconsultation in this context.

Objective: The primary aim was to evaluate the financial and environmental benefits of teleconsultation for preanesthetic consultation. Secondary objectives included assessing patient satisfaction and perioperative safety.

Methods: This retrospective, single-center observational study included patients scheduled for orthopedic surgery between September 2020 and October 2020 at Toulouse University Hospital. Eligible patients completed a preconsultation questionnaire via the MyAnesth digital agent. Patients were allocated to teleconsultation or in-person consultation groups based on predefined criteria. Postoperative data on demographics, transportation, consultation modality, time off work, and patient satisfaction were collected. Economic analysis included travel costs, income loss, and health insurance reimbursements. Ecological analysis quantified greenhouse gas (GHG) emissions based on transportation mode and digital infrastructure use. Statistical comparisons between the teleconsultation and in-person consultation groups used appropriate parametric and nonparametric tests, with significance set at $P \leq .05$.

Results: A total of 401 patients were analyzed (teleconsultations: n=331, 82.5%; in-person consultations: n=70, 17.5%). Teleconsultations reduced the average travel distance by 46,000 km, corresponding to 9.7 tons of carbon dioxide equivalent saved. Mean cost savings per patient were €122 (SD €125; 1 US $=€1.17), with total savings of €42,840 for patients and the national health care system. Teleconsultations also significantly reduced time spent on travel and administrative processes (mean 22, SD 9 minutes vs mean 130, SD 16 minutes for in-person consultations; $P<.001$). No significant differences in postoperative complication rates were observed between groups (teleconsultations: 11/331, 3.3%; in-person consultations: 5/70, 7.1%; $P=.24$). Patient satisfaction scores were high and similar in both groups (median 9, IQR 8-10, of a possible 10), with most patients preferring teleconsultations or expressing no preference for consultation modality. Digital teleconsultation infrastructure contributed minimally to GHG emissions (2.3 kg of carbon dioxide equivalent for 331 teleconsultations), representing a 99% reduction compared to travel-based in-person consultations.

Conclusions: Teleconsultations for preanesthetic assessment demonstrated significant economic and ecological advantages without compromising clinical safety or patient satisfaction. Patients reported high levels of satisfaction and minimal attachment to in-person consultations and appreciated the convenience of remote access. This model reduces unnecessary travel, limits health care–related GHG emissions, and generates considerable cost savings for both patients and public health systems. These findings support broader integration of teleconsultations into routine anesthetic care, particularly for low-risk outpatient surgical candidates. Expanding teleconsultation eligibility criteria could enhance health care system efficiency and contribute to sustainable medical practice.

XSL•FO
RenderX

(JMIR Form Res 2025;9:e70259) doi:10.2196/70259

KEYWORDS

teleconsultation; telemedicine; anesthesia; economic; ecological; carbon impact; carbon dioxide equivalent; CO2eq; greenhouse gas

Introduction

Background

Telemedicine is the remote diagnosis, treatment, and monitoring of patients using telecommunications technology. In 1997, the World Health Organization officially recognized telemedicine for the first time, defining it as "activities, services, and health systems conducted remotely using information and communication technologies to meet global health promotion, care, epidemic control, management, and health-related research needs [1]." This global endorsement of telemedicine and the development of suitable communication tools marked the beginning of a new era of "modern" telemedicine. In this context, the development of teleconsultation facilitated health care access in underserved areas, reduced health care avoidance [2], shortened wait times [3], and improved patient follow-up [4]. A sudden acceleration in the use of telemedicine occurred in March 2020, driven by the COVID-19 pandemic [5,6]. Since then, teleconsultation has experienced a significant rise, becoming widespread across numerous specialties and leading to an exponential increase in its use, even in pediatric and obstetric populations [7,8]. The preanesthetic consultation (PAC) is a mandatory step before any scheduled surgery. Its purpose is to assess the patient's health status, determine the need for additional examinations or specialized consultations, and evaluate the risk of complications during the perioperative period. Beyond reducing interpersonal contact, as required during a pandemic, telemedicine decreases the patient travel costs and associated expenses generated by in-person consultations for both patients and health care systems [6]. As such, teleconsultations could offer significant medical-economic advantages [9,10]. In parallel with digital progress, the health care sector is beginning to participate in the ongoing ecological revolution. The "forced" replacement of in-person consultations with teleconsultations during the COVID-19 pandemic highlighted the feasibility of maintaining care quality while reducing unnecessary travel [6,11]. To our knowledge, no data deal with the real economic and ecological effect of teleconsultations. Several studies have evaluated the benefits of teleconsultation in various medical specialties, demonstrating improved accessibility, patient satisfaction, and potential cost savings. However, these studies have rarely quantified the combined economic and environmental impacts of replacing in-person consultations with teleconsultations. This study aimed to fill this gap by evaluating both the financial and environmental consequences of implementing teleconsultations for PACs.

Objectives

Thus, the primary objective of this study was to evaluate the economic and ecological benefits generated by conducting preanesthesia teleconsultations. Secondary objectives included assessing patient satisfaction and safety in preanesthetic teleconsultations.

Methods

Experimental Design

This was a retrospective, observational, single-center study conducted in the orthopedic surgery anesthesia department at Toulouse University Hospital.

Ethical Considerations

This research is considered an experiment in educational sciences aiming to evaluate the participative and pedagogical quality of a new digital tool implemented in current practice.

This study was exempt from formal institutional review board approval according to French regulations as it involved retrospective analysis of anonymized data collected during routine care, with no intervention or identifiable personal data. As such, it did not fall within the scope of research involving human participants as defined by the French Public Health Code (Articles L1121-1 and R1121-1). Therefore, no ethics committee review was required. This study did not present any risk for the participants and did not modify the usual care pathway or the time required for patient management.

Study Population

All patients scheduled for a PAC for planned orthopedic surgery between September 2020 and October 2020 who responded to the digital conversational agent MyAnesth to prepare for their PAC were eligible. The digital conversational agent MyAnesth was developed in collaboration with a company that creates secure health companions (BOTdesign) [12]. Medical data obtained using MyAnesth allowed for the calculation of a numerical score (MyRISK score) for each patient, stratifying them into low (green), intermediate (orange), or high (red) perioperative risk [13]. Exclusion criteria included the conduct of a relocated PAC (ie, conducted at another hospital center), the presence of a legal guardianship or protection regime (guardianship, trusteeship, or judicial protection), patients who did not speak French, and patients who opposed participation in the study.

PAC Modalities

The modality for the anesthesia consultation (in-person consultation or teleconsultation) was decided based on predefined criteria. Eligibility criteria for teleconsultations included patient consent, a surgical intervention considered minor or intermediate (eg, knee arthroscopy, hand or foot surgery, or removal of osteosynthesis material), good comprehension of French, and sufficient cognitive and technical capabilities for teleconsultation (adequate equipment). Thus, 2 preanesthetic trajectories were identified, defining the

XSL•FO
RenderX

teleconsultation and in-person consultation groups. As per our standard of care, the allocated time for the PAC was 20 minutes.

Teleconsultations were conducted via the TéléO platform (NEHS DIGITAL) and created by the National Mutual of Hospital Workers. TéléO has been used at Toulouse University Hospital since the onset of the COVID-19 pandemic (March 2020). The platform features secure connectivity with internet speed tests and audio and video quality verification. The application is described as "full web," operating solely with an internet connection through any browser. TéléO operates on Web Real-Time Communication, a technology that allows applications and websites to exchange voice calls and video chats and conduct instant file sharing without requiring the user to install plug-ins or other third-party software. TéléO is accessible from any device (smartphone, tablet, desktop, or laptop computer).

Postoperative Data Collection

Patients were recontacted via phone postoperatively by one of the anesthesia team investigators to collect the following study-related data:

- Demographic characteristics, including age, sex, type of surgery, and American Society of Anesthesiologists score.
- PAC modality (in-person consultation or teleconsultation), including mode of transportation used to attend the in-person consultation (car, medical transport, public transportation, bicycle, or scooter), distance between the home and the hospital, presence of a companion, and whether the patient or companion needed to take a full or half day off from work. For this purpose, patients who had undergone teleconsultations were asked to answer these questions hypothetically as if they had attended an in-person consultation ("If you had attended in person...").
- Patient satisfaction with the PAC, collected using a simple numerical scale ranging from 0 ("extremely dissatisfied") to 10 ("extremely satisfied"). Patients were also asked about their preference for a future PAC modality (teleconsultation, in-person consultation, or no preference).

All data were anonymized.

Economic Data Collection

Concerning Patients and Health Insurance

Transportation Costs

Transportation costs by car were estimated based on the standard mileage allowance for vehicle use (order of February 15, 2021; [14]) using the formula distance × 0.541 (average for a vehicle with horsepower of 60 to 150). This rate includes vehicle power, fuel consumption, depreciation, maintenance costs, tire wear, and insurance premiums.

Transportation costs for medical vehicles were calculated based on the applicable rates for the Occitania region [15]. These include a base fee (averaged at €12.70; a conversion rate of 1 US $=€1.17 applies), a patient pickup fee (€15.58), and a per-kilometer rate (€1.02). The patient's copay was €4.00 for a round trip. Short-distance fees (under 18 km) and additional fees for ambulance transport were not considered to avoid overestimating results. Health insurance was assumed to reimburse 65%.

Salary Costs

Time off work for patients or their companions to attend the PAC was recorded and considered as vacation or compensatory time off. Sick leave and employer or health insurance coverage were not considered due to the 1-day waiting period since 2018.

The loss of income for patients was calculated based on the following:

- The French median gross hourly wage in 2018, €15.20 per hour (National Institute of Statistics and Economic Studies 2021), including base salary and overtime or bonuses.
- The average French working hours (National Institute of Statistics and Economic Studies 2019 survey). The average workday is 7.9 hours for full-time employees and 8.4 hours for nonemployees, with a national distribution of 87.9% employees and 12.1% self-employed.

For retirees or unemployed patients, the loss of leisure time was evaluated at 33% of the net salary.

Connection Costs

For teleconsultations, internet or device costs were not included if most patients already had internet access for personal use in 2020.

Travel and Consultation Time

Travel times and distances between patients' homes and the hospital were calculated using Google Maps based on the postal code provided during administrative registration. Cycling speed was estimated at 15 km/h. Public transportation times were not included due to excessive variability. The average time spent at the hospital from administrative check-in was calculated from a sample of 30 in-person consultation patients. For the teleconsultation group, transportation costs and times were extrapolated from patient responses regarding hypothetical in-person consultation attendance.

Concerning the Hospital

Software Investments

The TéléO application was funded by the Regional Health Agency and made available free of charge throughout the region.

Anesthesia Consultation

Under activity-based pricing, the price of a teleconsultation was set at €23 (specialized consultation, hospital billing guide 2020). The average duration of the PAC was calculated from a sample of 60 patients (30 for teleconsultations and 30 for in-person consultations).

Ecological Data Collection

Greenhouse gas (GHG) emissions were estimated based on (1) activity data (energy consumption in kilowatt-hours and distances traveled in kilometers) and (2) carbon emission factors (from the Environment and Energy Management Agency (Agence De l'Environnement et de la Maîtrise de l'Énergie) carbon database, representing the quantity of GHGs emitted by

XSL•FO
RenderX

the activity). Results were expressed in tons of carbon dioxide equivalent (CO_2eq).

Transportation Emissions

For passenger vehicles, a medium-sized car with various types of engines (excluding electric and hybrid) was considered, with an emission factor of 0.218 kg CO_2eq per kilometer. Medical transport was considered to have the same emission factor. Public transport emissions were calculated assuming metro or tramway use, with an average factor of 0.0024 kg CO_2eq per kilometer. Bus use was not considered to avoid overestimating results. Our decision to exclude the bus was based on the following rationale: overlap with other modes. In the study area, bus routes often run parallel to tram and metro lines. This study primarily focused on high-capacity, fixed-route modes (metro and tram) due to their greater relevance for structural transport planning. Buses, being more flexible and variable, would have required a different modeling approach. Flights were considered short haul, with an emission factor of 0.414 kg CO_2eq per kilometer.

The carbon dioxide emission factors used in this study were taken from the carbon database of the Agence De l'Environnement et de la Maîtrise de l'Énergie [16], which provides official French emission data for various transport modes and energy sources. Cost assumptions were based on national sources, including the Centre for Studies and Expertise on Risks, the Environment, Mobility, and Urban Planning (2020) for unit transport costs [17].

Teleconsultation Emissions

Energy consumption from IT equipment using fiber and digital subscriber lines was considered. A 5-minute connection to TéléO and a 20-minute teleconsultation session were assumed. Equipment used for teleconsultations included a fully equipped desktop computer (monitor, speakers, webcam, printer, and internet) consuming 200 W (Wh) per hour of use, equivalent to 67 Wh for 20 minutes. Patient equipment energy consumption was only considered when using a computer (assumed to be a desktop) to avoid overestimation. Energy consumption from smartphones or tablets was not included as their use primarily affects manufacturing emissions rather than use.

Statistical Analysis

Qualitative data were expressed as percentages. Normality of data was tested using the Shapiro-Wilk test. Quantitative data were expressed as medians and IQRs or means and SDs. Qualitative variables were compared using the Fisher exact test or chi-square test, and quantitative variables were compared using the Mann-Whitney *U* test. Statistical analysis was conducted using the MedCalc software (version 12.6.1; MedCalc Software Ltd) and the online statistics platform pvalue.io (Medistica). A P value of ≤.05 was considered statistically significant.

Results

Patient Characteristics

Of the 423 patients included, 401 (94.8%) completed the survey, enabling the analysis of study-related end points. The distribution of patients between the teleconsultation and in-person consultation groups is presented in Table 1. The mean age was 41.4 (SD 16.2) years. Most surgical interventions were considered minor (292/401, 72.8%) and scheduled as outpatient procedures (304/401, 75.8%). Patients seen via teleconsultation were younger, professionally active, and scheduled for minor outpatient surgeries. The primary connection device used was a computer (245.331, 73.6%), followed by a smartphone (53/331, 16%) and a tablet (27/331, 8.2%). A total of 17.5% (70/401) of all patients had previously experienced teleconsultations.

XSL•FO
RenderX

Table 1. Patient characteristics (N=401).a

	Total	Teleconsultations (n=331)	In-person consultations (n=70)	*P* value
Sex, n (%)				.35
Male	238 (59.4)	193 (58.3)	45 (64.3)	
Female	163 (40.6)	138 (41.7)	25 (35.7)	
Age (y), n (%)				
≤25	67 (16.7)	70 (21.1)	7 (10)	
26-45	162 (40.4)	132 (39.9)	29 (41.4)	
46-65	129 (32.2)	109 (32.9)	21 (30)	
>65	43 (10.7)	20 (6)	13 (18.6)	
Age (y), mean (SD)	41.4 (16.2)	40.4 (15.6)	48.0 (17.3)	*<.001*
Age (y), median (IQR)	40 (28-55)	38 (27.0-54.0)	46.5 (33.5-62.0)	
ASA [b] score, n (%)				*<.001*
1	289 (72.1)	250 (75.5)	36 (51.4)	
2	93 (23.2)	69 (20.8)	24 (34.3)	
3	19 (4.7)	9 (2.7)	10 (14.3)	
4	0 (0)	0 (0)	0 (0)	
MyRISK score, n (%)				.23
Green	106 (26.4)	93 (28.1)	13 (18.6)	
Orange	153 (38.2)	125 (37.8)	28 (40)	
Red	142 (35.4)	113 (34.1)	29 (41.4)	
Occupation, n (%)				*<.001*
Student	60 (15)	58 (17.5)	2 (2.9)	
Employed	284 (70.8)	237 (71.6)	47 (67.1)	
Retired	41 (10.2)	24 (7.3)	17 (24.3)	
Unemployed	16 (4)	12 (3.6)	4 (5.7)	
Risk associated with surgery, n (%)				*.02*
Minor	292 (72.8)	250 (75.5)	42 (60)	
Intermediate	92 (22.9)	69 (20.8)	23 (32.9)	
Major	17 (4.2)	12 (3.6)	5 (7.1)	
Type of hospitalization, n (%)				*<.01*
Ambulatory	304 (75.8)	261 (78.9)	43 (61.4)	
Hospitalization	96 (23.9)	69 (20.8)	27 (38.6)	
N/A[c]	1 (0.2)	1 (0.3)	0 (0)	
Previous teleconsultation, n (%)				.94
Yes	70 (17.5)	58 (17.5)	12 (17.1)	
No	331 (82.5)	273 (82.5)	58 (82.9)	
Distance from home to the hospital (km), mean (SD)	66 (96)	70 (101)	50 (70)	.24

[a]*P* values indicate comparisons between the teleconsultation and in-person consultation groups.

[b]ASA: American Society of Anesthesiologists.

[c]N/A: not available

XSL•FO
RenderX

User Satisfaction

Patient satisfaction surveys revealed a median score of 9 (IQR 8-10) and 9 (IQR 8-9) for the teleconsultation and in-person consultation groups, respectively (*P*=.64 for group comparison). Only 13.6% (45/331) of patients in the teleconsultation group and 18.6% (13/70) in the in-person consultation group expressed a preference for having their next PAC in person.

Perioperative Complications

In total, 4.1% (16/391) of the patients experienced complications within the first 6 postoperative months, without difference between groups (*P*=.24): 3.3% (11/331) of the patients had a PAC via teleconsultation, and 7.1% (5/70) of the patients were seen in in-person consultations. The complications were predominantly infectious (10/391, 2.6%) and hemorrhagic (5/391, 1.3%). No patients required intensive care or died.

Transportation Evaluation

The most frequently used mode of transportation for in-person consultations was a car (292/401, 72.8%). The mean distance between the patients' residences and the university hospital was 66 (SD 96) km, with no statistically significant difference (*P*=.46) between groups. Nearly half (191/401, 47.6%) of the patients lived within 25 km, whereas 21.9% (88/401) resided more than 100 km from the hospital (Table 2).

Table 2. Distance and mode of transportation by type of preanesthetic consultation (hypothetical for teleconsultations; N=401).a

	Total, n (%)	Teleconsultations (n=331), n (%)	In-person consultations (n=70), n (%)	*P* value
Distance (km)				
<25	191 (47.6)	159 (48)	32 (45.7)	.32
25-50	58 (14.5)	41 (12.4)	17 (24.3)	.37
51-100	64 (16)	52 (15.7)	12 (17.1)	.56
101-200	65 (16.2)	57 (17.2)	8 (11.4)	.62
201-300	13 (3.2)	13 (3.9)	0 (0)	
≥300	10 (2.5)	9 (2.7)	1 (1.4)	
Mode of transportation				*<.001*
Car	292 (72.8)	252 (76.1)	40 (57.1)	
Ambulance	33 (8.2)	19 (5.7)	14 (20)	
Public transport	63 (15.7)	48 (14.5)	15 (21.4)	
Bicycle or scooter	11 (2.7)	11 (3.3)	0 (0)	
Plane	2 (0.5)	1 (0.3)	1 (1.4)	

[a]Mode of transportation used to attend the in-person consultations (car, medical transport, public transportation, bicycle, or scooter), distance between the home and the hospital, presence of a companion, and whether the patient or companion needed to take a full or half day off from work. For this purpose, patients who had undergone teleconsultations were asked to answer these questions hypothetically as if they had attended an in-person consultation ("If you had attended in person...").

When comparing the type of transportation used by in-person consultation patients with the hypothetical mode of transport that teleconsultation patients would have used for in-person consultations, a statistically significant difference was observed (*P*<.001). Compared to in-person consultation patients, teleconsultation patients were more likely to use their personal car and less likely to use an ambulance (Table 2).

There was no significant difference in distance traveled based on the type of transport (not shown). Overall, teleconsultations reduced the total distance traveled by 44,266 km.

Expenditure Evaluation

For Patients

Financial Savings

One patient traveling by air was excluded from the financial analyses due to the variability of airfare costs. The cost of a round trip was estimated per patient at €85 (SD €100) by car (252/331, 75.5%), €82 (SD €61) by ambulance (19/331, 5.7%), €3.3 (SD €0.70) by public transport (48/331, 14.5%), and €0 by bicycle or scooter (11/331, 3.3%). By avoiding travel, teleconsultations saved patients a total of €23,240 in travel costs (excluding parking). At our university hospital, patients benefit from a fixed parking fee of €1.80 per visit. Including parking, teleconsultations resulted in savings of €87 (SD €102) per patient using a car, totaling €23,694 in overall transportation savings.

More than a third of teleconsultation patients would have had to take time off work to attend an in-person consultation. In addition, 42% (139/331) would have been accompanied by a relative, who, in nearly half (165/331,%) of the cases, would also have needed to take time off work. There was no difference between the groups regarding the presence of an accompanying person or the need to take time off work for either the patient or their companion. There was no statistically significant

XSL•FO
RenderX

association between the presence of an accompanying person and patient age (odds ratio 0.989, 95% CI 0.975-1.00; *P*=.15).

The theoretical income loss was estimated at €122 per workday. Overall, teleconsultations resulted in total savings of €16,408 in income loss for 330 of the patients, including €6627 saved by the accompanying persons. For students, retirees, or unemployed patients seen in teleconsultations (94/331, 28.4%), the financial loss was estimated at €5 per hour, with an average saving of €9.60 (SD €8.40) per patient, totaling €905.

The overall savings for the teleconsultation group combining theoretical travel costs and income loss amounted to €40,102 for 330 patients (Table 3), corresponding to an average of €122 (SD €125) per patient.

Table 3. Total savings for a patient seen in teleconsultation according to the type of transport.a A conversion rate of 1 US $=€1.17 applies.

Mode of transportation	Total, n	Cumulative expenditure (€), mean (SD)	Cumulative expenditure total=€40,102
Car	252	142 (132)	35,815
Ambulance	19	117 (74)	2230
Public transport	48	39 (41)	1926
Bicycle or scooter	11	15 (26)	131

[a]Transportation costs by car were estimated based on the standard mileage allowance for vehicle use (order of February 15, 2021; Service-Public.fr) using the formula distance × 0.541 (average for a vehicle with horsepower of 60 to 150). This rate includes vehicle power, fuel consumption, depreciation, maintenance costs, tire wear, and insurance premiums. Transportation costs for medical vehicles were calculated based on the applicable rates for the Occitania region [15]. These include a base fee (averaged at €12.70), a patient pickup fee (€15.58), and a per-kilometer rate (€1.02). The patient's copay was €4.00 for a round trip. Short-distance fees (under 18 km) and additional fees for ambulance transport were not considered to avoid overestimating results. Health insurance was assumed to reimburse 65%.

Time Savings

The average transport time saved by teleconsultation patients was 123 (SD 100) minutes. In the subset of patients who would have traveled by car, over half (152/252, 60.3%) would have spent at least 2 hours traveling. The average round trip would have taken 125 (SD 100) minutes. In contrast, the mean transport time for in-person consultation patients was 85 (SD 75) minutes (all transport modes combined).

For in-person consultations, the patients' journey included administrative registration; waiting time; the consultation with the anesthesiologist; and a potential discussion with a nurse for additional tests, electrocardiograms, or supplementary planning. The total time dedicated to the PCA (ie, transportation+administrative processes+PCA) for both the in-person consultation and teleconsultation groups was 130 (SD 16) and 22 (SD 9) minutes, respectively (*P*<.001). The average transportation time was 85 (SD 19) minutes in the in-person consultation group. The average time spent at the hospital for in-person consultation patients was 45 (SD 13) minutes. The average time spent for the PCA was 17 (SD 3) minutes for teleconsultations and 20 (SD 3) minutes for in-person consultations (*P*<.001 for group comparison), whereas the administrative time was longer for in-person consultations (mean 25, SD 4 minutes) compared to teleconsultations (mean 5, SD 3 minutes; *P*<.001).

For the National Health Insurance

Medical transportation services (ambulance) would have been used by 5.7% (19/331) of the patients in the teleconsultation group. The mean distance traveled was 89 (SD 83) km. In the teleconsultation group, the transportation cost per patient for a round trip was €226 (SD €174), with a reimbursement rate of 65% by the national health insurance, amounting to €144 (SD €113) per patient. Teleconsultations allowed the national health insurance to save €2737 for this 5.7% (19/331) of the patients. The total savings (patients+national health insurance) achieved in the teleconsultation group amounted to €42,840, corresponding to overall savings of €130 (SD €134) per patient.

For the Hospital

Specialized Software

The funding of TéléO by the Occitania Regional Health Agency did not result in any additional costs related to the use of this specialized software by the hospital.

Carbon Footprint Assessment

Carbon Footprint Related to Transportation

By reducing travel, teleconsultations enabled a reduction in GHG emissions by 9688 kg CO_2eq (Table 4).

XSL•FO
RenderX

Table 4. Carbon impact by mode of transport in the teleconsultation group.[a]

Mode of transport	Distance round trip saved (km)	CO_2eq[b] reduction (kg)
Car	39,792	8675
Ambulance	3386	738
Public transport	1088	2.6
Plane	1928	272
Bicycle	20	0
Total	46,214	9688

[a]For passenger vehicles, a medium-sized car with various types of engines (excluding electric and hybrid) was considered, with an emission factor of 0.218 kg of carbon dioxide equivalent per kilometer. Medical transport was considered to have the same emission factor. Public transport emissions were calculated assuming metro or tramway use, with an average factor of 0.0024 kg of carbon dioxide equivalent per kilometer. Bus use was not considered to avoid overestimating the results. Flights were considered short haul, with an emission factor of 0.414 kg of carbon dioxide equivalent per kilometer.

[b]CO_2eq: carbon dioxide equivalent.

Carbon Footprint Related to Teleconsultations

During an in-person PCA, the anesthesiologist systematically uses computer equipment to access the patient's electronic medical records (ORBIS; Dedalus). The GHG emissions from medical computer equipment were calculated to be 0.004 kg CO_2eq per consultation, corresponding to 0.096 kg CO_2eq per consultation day. For teleconsultations, the digital connection between the patient and physician generated GHG emissions, detailed in Table 5. The total carbon footprint associated with digital connections was 2.30 kg CO_2eq for the 331 teleconsultations conducted (Table 5). Owing to the remote consultation model for the 82.5% (331/401) of patients analyzed who underwent teleconsultations, the anesthesia department was able to reduce its total carbon emissions by 9686 tons of CO_2eq, equating to a reduction of 29.3 kg CO_2eq per patient.

Table 5. Energy consumption and carbon impact by mode of connection in the teleconsultation group.[a]

Mode of connection	Energy consumption (Wh)		CO_2eq[b] impact (kg)
	Patient	Anesthesiologist	
PC (n=244)	16,348	16,348	1.96
Smartphone (n=53)	3551	__[c]	0.21
Tablet (n=28)	1876	—	0.11
NA[d] (n=6)	402	—	0.02
Total	22,177	16,348	2.30

[a]Energy consumption from IT equipment using fiber and digital subscriber lines was considered. A 5-minute connection to TéléO and a 20-minute teleconsultation session were assumed. Equipment used for teleconsultations included a fully equipped desktop computer (monitor, speakers, webcam, printer, and internet) consuming 200 W per hour of use, equivalent to 67 Wh for 20 minutes. Patient equipment energy consumption was only considered when using a computer (assumed to be a desktop) to avoid overestimation. Energy consumption from smartphones or tablets was not included as their use primarily affects manufacturing emissions rather than use.

[b]CO_2eq: carbon dioxide equivalent.

[c]Not available.

[d]NA: not aplicable.

The main results of the economic and ecological impact evaluation of anesthesia teleconsultations are summarized in Table 6.

XSL•FO
RenderX

Table 6. Summary of results in the teleconsultation group. A conversion rate of 1 US $=€1.17 applies.

	Reduction per patient	Total reduction
Economic impact		
Transport savings (€), mean (SD)		
For the patient	72 (94)	23,695 (27,827)
For the health insurance	144 (113)[a]	2737 (3214)
Salary savings (€), mean (SD)		
All patients	30 (40)	9781(11,487)
Students, retirees, or unemployed patients	9.6 (8.4)	905 (1063)
Accompanying persons	20 (41)	6627 (7783)
Total savings (€), mean (SD)	130 (134)	42,840 (50,311)
Time savings (min)		
Transport, mean (SD)	123 (100)	34,082
Administrative registration	20[b]	6620
Consultation	3[c]	993
Total	146	48,326[d]
Round-trip distance savings (km), mean (SD)	140 (201)	46,214
Ecological impact		
Transport-related CO_2eq[e] savings (kg)	29.3	9688

[a]Per patient supported by health insurance.

[b]Average administrative time for in-person consultations.

[c]Difference compared to the average consultation time for in-person consultations.

[d]Equivalent to 805 hours.

[e]CO_2eq: carbon dioxide equivalent.

Discussion

Principal Findings

Through this study, we highlighted the medical, economic, and ecological benefits of conducting preanesthetic teleconsultations. Compared to in-person consultations, teleconsultations led to an average financial savings of €118 per patient, totaling over €42,000 in savings for 82.5% (331/401) of the patients. These economic benefits were primarily due to the reduction in travel-related expenses. Furthermore, approximately 10 tons of CO_2eq emissions were avoided. Opting for home-based teleconsultations without paramedical facilitators could have limited the precision of clinical examinations compared to in-person consultations. However, most anesthetic risk assessments rely on scores obtained through patient interviews (Lee, metabolic equivalent of task, STOP-BANG, HEMSTOP, or APFEL scores). No correlation was found between the consultation modality and perioperative complications. When evaluating airways and intubation conditions, dental health, mouth opening, and the lip test were easily assessed via webcam, whereas the Mallampati score proved more challenging. Nevertheless, no significant differences in predicting difficult intubation were found between teleconsultations and in-person consultations for maxillofacial surgeries, although the positive predictive value for difficult intubation remained low in both cases [3]. Adjustments to airway management strategies following preanesthetic visits were consistent across both modalities.

In this study, patients expressed high satisfaction with teleconsultations, citing ease of use and intuitiveness. There was minimal “attachment” to in-person consultations. A study showed initial patient apprehension toward teleconsultations, with only half of the patients expressing willingness due to concerns about not meeting the physician, lack of equipment, or limited technical knowledge [18]. However, post-teleconsultation satisfaction is now well documented, with most patients reporting high satisfaction and a sense of security [3,19]. However, teleconsultations were not inferior to in-person consultations for preoperative patient evaluation and may be an interesting economic and ecological alternative [9]. During the COVID-19 pandemic, public health campaigns promoting confinement, the expansion of telecommunication technologies, and teleconsultation reimbursements likely influenced patient satisfaction. Technical issues (eg, poor network connectivity and lack of cameras or microphones) were the main limitations [20]. Improvements in technology and digital literacy are expected to enhance patient adherence, especially in fields such as orthopedic surgery where mobility challenges are significant [20].

XSL•FO
RenderX

Nearly 5000 patients are seen annually in PACs for planned orthopedic surgeries at Toulouse University Hospital. Given the hospital's expertise, many patients must travel significant distances. Teleconsultations reduce travel time and distance, which is particularly beneficial in regions with limited mobility options.

For instance, a study found an average reduction of 447 km and 245 minutes per patient, saving over 8.6 million km and nearly 9 years of travel time over 17 years [21]. Another one noted an average saving of 102 km and 67 minutes per patient in the pediatric department in Kentucky, United States [22]. While these figures are not directly transferable to France, comparable results were observed in Portugal (47 km on average) [23]. In this study, teleconsultations reduced patient travel time 6-fold, aligning with the findings of previous studies [24].

Reduced travel also led to significant financial savings for patients. Similar findings have been reported in studies evaluating teleconsultation benefits. As an example, a minimum saving of US $32 per patient was noted in Arkansas, and average savings of US $150 per patient were observed in California [25]. This study found an average saving of €118 (SD €122) per patient [25,26]. The development of the child psychiatry teleconsultation program at the University of Kansas Medical Center has saved an average of US $86 per patient [27]. Over time, these savings accumulate; Dullet et al [21] estimated US $2.9 billion in savings over 17.5 years. In addition, one-third of patients in this study reported needing time off work for in-person consultations, aligning with the findings of Yen et al [28] that this applies to nearly half of patients. Patients living alone or geographically isolated would be the most likely to miss a day of work in the event of in-person consultations [25]. This loss of earnings is not negligible for the patient or their companion. Bynum et al [25] found a saving of US $75 to US $150 for 74% of the families of patients seen in teleconsultations, with reduced absenteeism at work. A Norwegian study suggested an average saving of €131 per patient by limiting absence from work [24], whereas our study estimated this saving at €122 per day. This study also highlights the potential for significant savings in health care reimbursements through teleconsultations. In this study, patients undergoing in-person consultations, often with a higher number of comorbidities, used sanitary vehicles more frequently, leading to higher transport costs [29]. Teleconsultations would help mitigate these expenses. Teleconsultations reduced the average consultation time without compromising patient satisfaction. This time saving could allow anesthesiologists to focus on complex cases or conduct additional consultations daily. Although not assessed in this study, previous research indicates no significant impact of teleconsultations on surgery cancellation rates even for major surgeries [30-32]. Patients living further from hospitals are more likely to cancel in-person appointments.

The health care sector contributes 3% to 8% of GHG emissions in high-income countries. To address these new environmental challenges, our health care services must undergo profound transformations. Among the strategies considered to reduce their ecological impact, hospitals are focusing on eHealth. Telemedicine is viewed as a promising technology, especially due to its potential to reduce transportation and air pollution. In 2021, patient transport was the second largest source of carbon emissions in health care following medication purchases. Digital health and teleconsultation are directly aligned with the ecological shift initiated by our specialty. Most studies on the environmental impact of teleconsultation only consider the reduction in transportation, which is the main source of carbon emissions. Dorrian et al [33], for example, highlighted a reduction of 123 kg CO_2eq per patient for the 42 patients followed via teleconsultations for head and neck cancer. Masino et al [34] concluded that conducting 840 teleconsultations over 6 months in Ontario, Canada, led to a reduction in GHG emissions of 185 tons of CO_2eq solely due to fewer travel requirements. The authors also remind us that carbon dioxide emissions are not the only GHG pollutants, and the emission of 360 kg of other airborne pollutants was also avoided. Over the 17 years of their program, Dullet et al [21] estimate a reduction in carbon emissions equivalent to the electricity consumption of 271 US households of 4 people. Similar results were found in studies from California, Scotland, and Wales [26,35,36]. In comparison, our study found a reduction of 9.7 tons of CO_2eq owing to the implementation of teleconsultations, which represents an average saving of 29.3 kg CO_2eq per patient. We also found that using public transport would have resulted in a CO_2eq emission of 2.6 kg. Although our results strongly favor teleconsultations, the ecological impact of an in-person consultation could be significantly reduced through widespread use of public transport. Obviously, these conclusions are only applicable to urban populations with easy access to public transport networks. Nevertheless, these figures are significant in our study, where 47.6% (191/401) of the patients lived within 25 km of the hospital. However, the environmental trade-offs of the massive deployment of teleconsultations need to be evaluated in terms of a benefit-risk balance. Indeed, the IT infrastructure necessary for their implementation is a major source of carbon emissions. In 2021, IT represented 3% to 5% of the carbon footprint of an average university hospital. The emissions exclusively linked to the internal IT operations of public hospitals were estimated at over 190,000 tons of CO_2eq annually, which is equivalent to 1 million 750 km round-trip flights for 1 person. The energy consumption of the IT sector is linked to 2 main areas: manufacturing (45%)—where the scope for action is limited except for rational purchasing—and use (55%). Use includes the operation of equipment (20%), data transmission via the internet (16%)—which is becoming more efficient and less energy consuming—and data storage in data centers (19%). Currently, only a minority of circulating health data are digitized. Their storage, which is expected to increase in both number and volume, and the associated carbon impact must prompt us to consider controlled IT consumption and digital sobriety. However, the carbon emissions linked to digital technologies for a 15-minute teleconsultation were minimal compared to the emissions avoided from travel. In fact, our study found an emission of 2.3 kg CO_2eq for the 331 teleconsultations, representing a reduction of over 99% in carbon emissions compared to a traditional in-person consultation. Similar data were found during the COVID-19 pandemic [34,37]. These results are particularly noteworthy as the carbon impact of the IT sector remains relatively low in our country

XSL•FO
RenderX

due to its largely decarbonized electricity production, with 92% of energy coming from nuclear, hydroelectric, solar, or wind sources.

Our study has several limitations. First, it was a single-center study. Even if our data are numerous, they can only apply to similar hospitals in similar areas. However, our results are consistent with the current literature on this topic. Second, due to lack of data, some simplifications and exclusions were made (such as type of transportation and underestimation of medical transport). In addition, for the estimation of ecological impact, only GHGs were considered following a simplified calculation approach rather than implementing a comprehensive environmental impact assessment method and life cycle assessment, among others. However, even underestimating some of the data, the results are in favor of teleconsultations. Finally, we did not consider electric or hybrid vehicles. Indeed, including electric vehicles would have required a more detailed energy system analysis, which is beyond the scope of this study.

Conclusions

This retrospective study of 331 patients who underwent teleconsultations demonstrated a savings of over 46,000 km in travel distance, along with a reduction of nearly 10 tons of CO_2eq emissions. In addition, significant savings in time spent on transportation and nonmedical hospital time were observed due to the use of teleconsultations. We were able to demonstrate a considerable direct financial benefit for the patient, as well as for the national health care system, through the reimbursement of medical transport costs. Moreover, our work highlighted the lack of attachment to in-person consultations by patients, as well as a high level of patient satisfaction with teleconsultations without any increase in postoperative complications. The medical-economic and ecological benefits of teleconsultations encourage us to refine and broaden the eligibility criteria for patients to use this mode of PAC.

Acknowledgments

The authors thank the medical and paramedical teams of the orthopedic operating room at Toulouse University Hospital.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

References

1. Oh H, Rizo C, Enkin M, Jadad A. What is eHealth (3): a systematic review of published definitions. J Med Internet Res 2005 Mar 24;7(1):e1 [FREE Full text] [doi: 10.2196/jmir.7.1.e1] [Medline: 15829471]
2. Nelson R. Telemedicine and telehealth: the potential to improve rural access to care. Am J Nurs 2017 Jun;117(6):17-18. [doi: 10.1097/01.NAJ.0000520244.60138.1c] [Medline: 28541980]
3. Applegate RL, Gildea B, Patchin R, Rook JL, Wolford B, Nyirady J, et al. Telemedicine pre-anesthesia evaluation: a randomized pilot trial. Telemed J E Health 2013 Mar;19(3):211-216. [doi: 10.1089/tmj.2012.0132] [Medline: 23384334]
4. Donelan K, Barreto EA, Sossong S, Michael C, Estrada J, Cohen A, et al. Patient and clinician experiences with telehealth for patient follow-up care. Am J Manag Care 2019 Jan;25(1):40-44 [FREE Full text] [Medline: 30667610]
5. Morau E, Blanc A, Boisson C, Sawyers T, Lefrant J, Cuvillon P. Telemedicine for preanesthesia consultations during the first COVID-19 lockdown. Telemed J E Health 2023 Apr 01;29(4):621-624. [doi: 10.1089/tmj.2022.0143] [Medline: 35877785]
6. Grüßer L, Bayram B, Ziemann S, Wallqvist J, Wienhold J, Rossaint R, et al. Teleconsultation for preoperative anesthesia evaluation: identifying environmental potentials by life cycle assessment. Telemed J E Health 2024 Jun 01;30(7):e2050-e2058. [doi: 10.1089/tmj.2023.0700] [Medline: 38656124]
7. Kern D, Bourdaud N, Jaber H, Ezzitouni M, Alacoque X, Larcher C, et al. Teleconsultation in pediatric anesthesia: first assessment of feasibility, quality, and satisfaction in France. Paediatr Anaesth 2023 Jul 03;33(7):552-561. [doi: 10.1111/pan.14663] [Medline: 36929280]
8. Sassi K, Tardif E, Lasry L, Manenc L, Guerby P, Minville V. Assessing feasibility and satisfaction in third-trimester pre-anaesthetic teleconsultations: a prospective observational study. Indian J Anaesth 2024 Feb;68(2):183-188 [FREE Full text] [doi: 10.4103/ija.ija_721_23] [Medline: 38435660]
9. Morau E, Chevallier T, Serrand C, Perin M, Gricourt Y, Cuvillon P. Teleconsultation compared with face-to-face consultation in the context of pre-anesthesia evaluation: TELANESTH, a randomized controlled single-blind non-inferiority study. J Clin Anesth 2024 Feb;92:111318. [doi: 10.1016/j.jclinane.2023.111318] [Medline: 37944402]
10. Quemeneur C, Le Saché F, Ayrault C, Griscenco E, Raft J, N'Guyen P, et al. Anesthesia teleconsultation: quality, medicoeconomic interest, and patient satisfaction of preoperative assessment-a prospective observational study. Telemed J E Health 2024 Feb 01;30(2):457-463. [doi: 10.1089/tmj.2023.0054] [Medline: 37523220]

XSL•FO
RenderX

11. Gillman-Wells CC, Sankar TK, Vadodaria S. COVID-19 reducing the risks: telemedicine is the new norm for surgical consultations and communications. Aesthetic Plast Surg 2021 Feb 03;45(1):343-348 [FREE Full text] [doi: 10.1007/s00266-020-01907-8] [Medline: 32885319]
12. Ferré F, Boeschlin N, Bastiani B, Castel A, Ferrier A, Bosch L, et al. Improving provision of preanesthetic information through use of the digital conversational agent "MyAnesth": prospective observational trial. J Med Internet Res 2020 Dec 04;22(12):e20455 [FREE Full text] [doi: 10.2196/20455] [Medline: 33275108]
13. Ferré F, Laurent R, Furelau P, Doumard E, Ferrier A, Bosch L, et al. Perioperative risk assessment of patients using the MyRISK digital score completed before the preanesthetic consultation: prospective observational study. JMIR Perioper Med 2023 Jan 16;6:e39044 [FREE Full text] [doi: 10.2196/39044] [Medline: 36645704]
14. Welcome to the official information and administrative procedures site. Government of France. URL: https://www.service-public.gouv.fr/ [accessed 2025-05-29]
15. l'Assurance Maladie. URL: https://www.ameli.fr/ [accessed 2025-10-08]
16. Carbon database. Agence De l'Environnement et de la Maîtrise de l'Énergie. URL: https://base-empreinte.ademe.fr/ [accessed 2025-10-08]
17. Evaluation socio-économique de projets de transports collectifs - Retour sur le rendez-vous Mobilités du 16 septembre 2021. Government of France. URL: https://www.cerema.fr/fr/actualites/nouveau-guide-recommandations-evaluation-socio-economique-TCSP [accessed 2025-05-29]
18. Anaïs Roche Interne 4e année du DES d'anesthésie-réanimation, Thilly N, Boileau S, Bouaziz H. Téléconsultation d'anesthésie au domicile : une enquête d'acceptabilité. Can J Anaesth 2018 May 20;65(5):597-599. [doi: 10.1007/s12630-017-1043-5] [Medline: 29264746]
19. Roberts S, Spain B, Hicks C, London J, Tay S. Telemedicine in the Northern territory: an assessment of patient perceptions in the preoperative anaesthetic clinic. Aust J Rural Health 2015 Jun 23;23(3):136-141. [doi: 10.1111/ajr.12140] [Medline: 25615954]
20. Baker J, Stanley A. Telemedicine technology: a review of services, equipment, and other aspects. Curr Allergy Asthma Rep 2018 Sep 26;18(11):60. [doi: 10.1007/s11882-018-0814-6] [Medline: 30259201]
21. Dullet NW, Geraghty EM, Kaufman T, Kissee JL, King J, Dharmar M, et al. Impact of a university-based outpatient telemedicine program on time savings, travel costs, and environmental pollutants. Value Health 2017 Apr;20(4):542-546 [FREE Full text] [doi: 10.1016/j.jval.2017.01.014] [Medline: 28407995]
22. Soares NS, Johnson AO, Patidar N. Geomapping telehealth access to developmental-behavioral pediatrics. Telemed J E Health 2013 Aug;19(8):585-590. [doi: 10.1089/tmj.2012.0226] [Medline: 23756240]
23. Oliveira TC, Barlow J, Gonçalves L, Bayer S. Teleconsultations reduce greenhouse gas emissions. J Health Serv Res Policy 2013 Oct 14;18(4):209-214. [doi: 10.1177/1355819613492717] [Medline: 23945677]
24. Buvik A, Bergmo TS, Bugge E, Smaabrekke A, Wilsgaard T, Olsen JA. Cost-effectiveness of telemedicine in remote orthopedic consultations: randomized controlled trial. J Med Internet Res 2019 Feb 19;21(2):e11330 [FREE Full text] [doi: 10.2196/11330] [Medline: 30777845]
25. Bynum AB, Irwin CA, Cranford CO, Denny GS. The impact of telemedicine on patients' cost savings: some preliminary findings. Telemed J E Health 2003 Dec;9(4):361-367. [doi: 10.1089/153056203772744680] [Medline: 14980093]
26. Yellowlees PM, Chorba K, Burke Parish M, Wynn-Jones H, Nafiz N. Telemedicine can make healthcare greener. Telemed J E Health 2010 Mar;16(2):229-232. [doi: 10.1089/tmj.2009.0105] [Medline: 20156125]
27. Spaulding R, Belz N, DeLurgio S, Williams AR. Cost savings of telemedicine utilization for child psychiatry in a rural Kansas community. Telemed J E Health 2010 Oct;16(8):867-871. [doi: 10.1089/tmj.2010.0054] [Medline: 20925567]
28. Yen C, Tsai M, Macario A. Preoperative evaluation clinics. Curr Opin Anaesthesiol 2010 Apr;23(2):167-172. [doi: 10.1097/ACO.0b013e328336f4b9] [Medline: 20124896]
29. Whetten J, Montoya J, Yonas H. ACCESS to better health and clear skies: telemedicine and greenhouse gas reduction. Telemed J E Health 2019 Oct 01;25(10):960-965. [doi: 10.1089/tmj.2018.0172] [Medline: 30359184]
30. Tam A, Leung A, O'Callaghan C, Fagermo N. Role of telehealth in perioperative medicine for regional and rural patients in Queensland. Intern Med J 2017 Aug 06;47(8):933-937. [doi: 10.1111/imj.13484] [Medline: 28485821]
31. Mullen-Fortino M, Rising KL, Duckworth J, Gwynn V, Sites FD, Hollander JE. Presurgical assessment using telemedicine technology: impact on efficiency, effectiveness, and patient experience of care. Telemed J E Health 2019 Feb;25(2):137-142. [doi: 10.1089/tmj.2017.0133] [Medline: 30048210]
32. Seidel JE, Beck CA, Pocobelli G, Lemaire JB, Bugar JM, Quan H, et al. Location of residence associated with the likelihood of patient visit to the preoperative assessment clinic. BMC Health Serv Res 2006 Feb 22;6(1):13 [FREE Full text] [doi: 10.1186/1472-6963-6-13] [Medline: 16504058]
33. Dorrian C, Ferguson J, Ah-See K, Barr C, Lalla K, van der Pol M, et al. Head and neck cancer assessment by flexible endoscopy and telemedicine. J Telemed Telecare 2009 Apr 13;15(3):118-121. [doi: 10.1258/jtt.2009.003004] [Medline: 19364891]
34. Masino C, Rubinstein E, Lem L, Purdy B, Rossos PG. The impact of telemedicine on greenhouse gas emissions at an academic health science center in Canada. Telemed J E Health 2010 Nov;16(9):973-976. [doi: 10.1089/tmj.2010.0057] [Medline: 20958198]

XSL•FO
RenderX

35. Wootton R, Tait A, Croft A. Environmental aspects of health care in the Grampian NHS region and the place of telehealth. J Telemed Telecare 2010 May 28;16(4):215-220 [FREE Full text] [doi: 10.1258/jtt.2010.004015] [Medline: 20511579]
36. Lewis D, Tranter G, Axford AT. Use of videoconferencing in Wales to reduce carbon dioxide emissions, travel costs and time. J Telemed Telecare 2009 Apr 13;15(3):137-138. [doi: 10.1258/jtt.2009.003010] [Medline: 19364897]
37. Filfilan A, Anract J, Chartier-Kastler E, Parra J, Vaessen C, de La Taille A, et al. Positive environmental impact of remote teleconsultation in urology during the COVID-19 pandemic in a highly populated area. Prog Urol 2021 Dec;31(16):1133-1138 [FREE Full text] [doi: 10.1016/j.purol.2021.08.036] [Medline: 34454847]

Abbreviations

CO2eq: carbon dioxide equivalent
GHG: greenhouse gas
PAC: preanesthetic consultation

Edited by A Mavragani; submitted 18.12.24; peer-reviewed by B Bayram, A Debbarma; comments to author 28.03.25; revised version received 19.05.25; accepted 05.07.25; published 13.11.25.

Please cite as:
Ferre F, Furelau P, Labaste F, Costa N, Vardon F, Piau A, Martin C, Minville V
Medical-Economic and Ecological Impact of Anesthesia Teleconsultation: Retrospective Observational Study
JMIR Form Res 2025;9:e70259
URL: https://formative.jmir.org/2025/1/e70259
doi:10.2196/70259
PMID:

©Fabrice Ferre, Philippine Furelau, François Labaste, Nadège Costa, Fanny Vardon, Antoine Piau, Charlotte Martin, Vincent Minville. Originally published in JMIR Formative Research (https://formative.jmir.org), 13.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Development of a Virtual Home-Based Cycling Intervention for Individuals With Chronic Obstructive Pulmonary Disease: Qualitative Study

Kristina Krogh Christensen[1], PT, MSc; Henrik Bøggild[2], MD, PhD; Ulla Møller Weinreich[1,3], Prof Dr, MD, PhD; Anna Lei Stoustrup[1,3], PT, MSc

[1]Department of Respiratory Diseases, Aalborg University Hospital, Aalborg, Denmark
[2]Department of Health Science and Technology, Public Health and Epidemiology, Aalborg University, Aalborg, Denmark
[3]Department of Clinical Medicine, The Faculty of Medicine, Aalborg University, Aalborg, null, Denmark

Corresponding Author:
Kristina Krogh Christensen, PT, MSc
Department of Respiratory Diseases, Aalborg University Hospital
Hobrovej 18-22
Aalborg, 9000
Denmark
Phone: 45 42462129
Email: Kristinakroghchr@hotmail.com

Abstract

Background: Chronic obstructive pulmonary disease (COPD) is a leading cause of mortality, and exercise has been shown to reduce both. Health conditions, environmental factors, and logistical challenges are often barriers for participation in pulmonary rehabilitation (PR). Given the barriers many individuals with COPD face when attending health care centers for PR, virtual home-based cycling exercise could be an option.

Objective: This study aimed to explore the development of a home-based cycling exercise intervention for individuals with COPD, focusing on aspects such as bicycle selection, app functionality, and pilot testing. Furthermore, it aimed to explore participants' and nonparticipants' attitudes toward the intervention.

Methods: Using a phenomenological-hermeneutic approach, data were gathered from 15 semistructured interviews, including test pilots, participants, and nonparticipants. A thematic analysis was used to analyze the data.

Results: Thematic analysis identified 8 key themes: bicycle selection, individual guidance needs, geographical and video quality, online connectivity, comfort and accessibility of home-based cycling, flexibility, energy levels, and practical limitations. Findings highlighted a preference for pedal bicycles with adjustable intensity, the importance of flexibility in scheduling, and the autonomy provided by a home-based setup. While participants appreciated the virtual journey on videos, barriers such as lack of energy, stress, and limited space were reported by nonparticipants.

Conclusions: Recommendations include enhancing app features and addressing individual needs to improve adherence. The study underscores the potential of tailored home-based exercise interventions in overcoming traditional PR challenges.

International Registered Report Identifier (IRRID): RR2-10.1136/bmjresp-2024-002573

(JMIR Form Res 2025;9:e71234) doi:10.2196/71234

KEYWORDS

chronic obstructive pulmonary disease; COPD; development; home-based exercise intervention; virtual cycling exercise; pulmonary rehabilitation

XSL•FO
RenderX

Introduction

Chronic Obstructive Pulmonary Disease

Chronic obstructive pulmonary disease (COPD) is the most common lung disease worldwide, affecting over 292 million people, and ranking as the third leading cause of death globally [1]. Admissions due to exacerbations of COPD are frequent, and readmissions are concerning, as a large group of individuals experience recurrent hospitalizations [2].

Despite the known benefits of pulmonary rehabilitation (PR) [3-5], many individuals with COPD experience barriers to participation, such as shortness of breath, comorbidities, weather conditions, logistical difficulties related to transportation and parking, and other physical health issues, significantly limiting the individual's ability to engage in physical activity [6,7]. To address motivational aspects beyond logistical barriers, self-determination theory (SDT) offers a valuable framework. SDT emphasizes autonomy, competence, and relatedness—three psychological needs that can guide the design of more engaging and sustainable home-based exercise interventions.

To optimize PR programs, it is important to adjust duration and intensity to each individual's exercise tolerance. This individualized approach ensures that the individual's specific needs and limitations are addressed, enhancing their ability to participate in physical activity [8,9]. Given these physical limitations and logistical difficulties, there is a need for a more adapted and practical solution to improving participation in PR among individuals with COPD.

Home Exercise and Cycling Exercise

Given the barriers many individuals with COPD face when attending health care centers for PR, home exercise could be an option. Exercise activities can safely be carried out at home, and among those, cycling can be offered as an alternative. Cycling has been identified as an effective way to improve physical function and is widely accessible [10]. In Denmark, cycling is one of the most preferred forms of transportation and is deeply ingrained in Danish culture [11], with 70% of Danes aged over 16 years owning a bicycle [12]. Cycling accounts for 21% of all trips under 10 kilometers and 15% of all transportation in the country [11].

4Mvideo App

For individuals with COPD who are unable to cycle outdoors due to health or environmental barriers, virtual solutions can provide a comparable experience from home. The 4Mvideo app is a software system that allows users to experience outdoor cycling virtually by using a cycle at home. The individuals can cycle along a recorded route featuring a cyclist, thus gaining a feeling of real cycling without the challenges of outdoor exercise. For this project, local cyclists will document their journey from Aalborg (Denmark) to Paris (France) with several hours of videos. This content is accessible through the 4Mvideo app, enabling users to virtually cycle from Aalborg to Paris from their homes (Figure 1).

Figure 1. A screen capture of the video on the iPad while cycling. The icon with 62 o/m shows the number of revolutions per minute. The cycling icon shows the performed distance and time. The video icon shows the distance and time of the video. With permission from 4Mvideo.

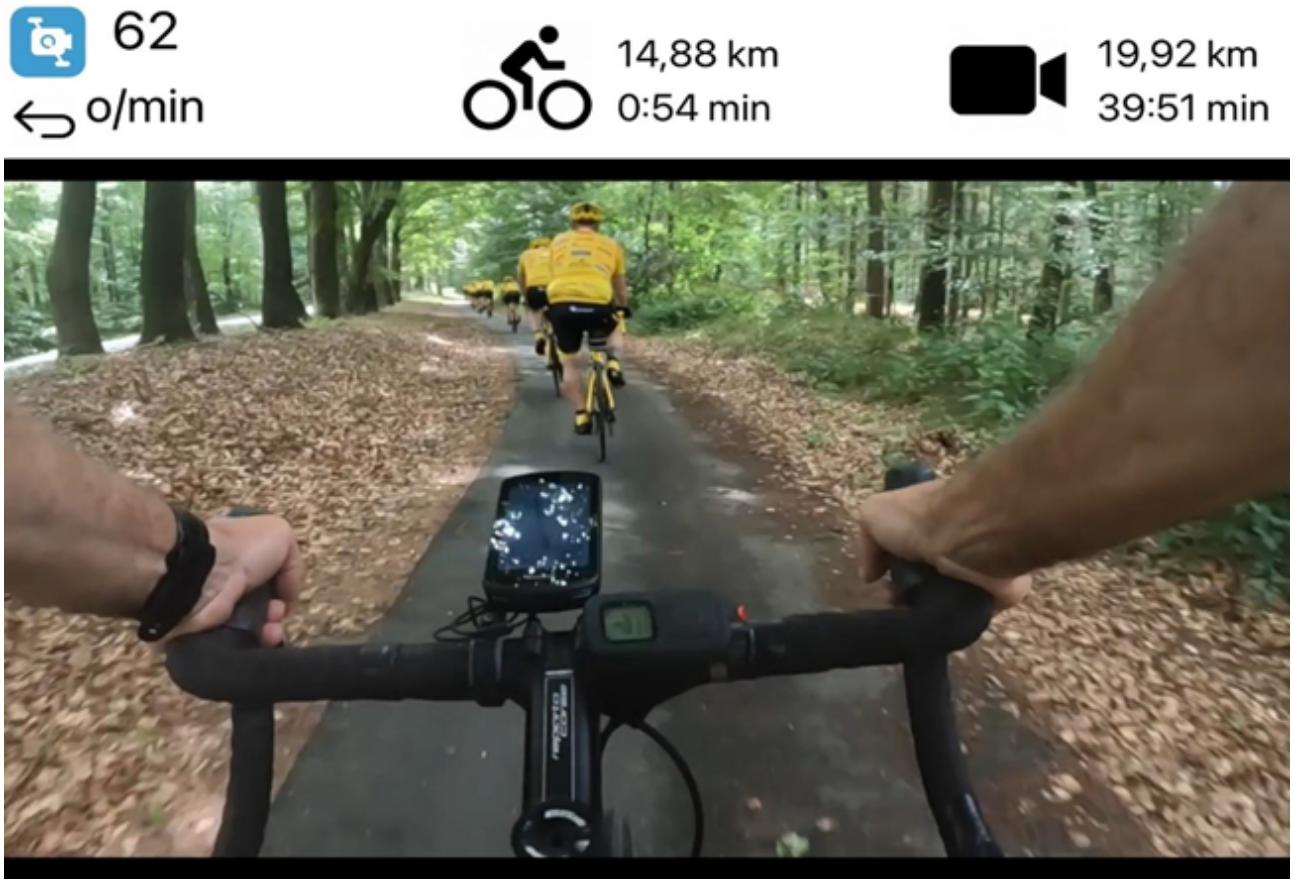

XSL•FO
RenderX

Cyclists participated in teams, and the app had a chat feature to simulate the experience of cycling together as a team. The collective distance covered by the team each day is tracked and shown for the participants, allowing the group to virtually travel from Aalborg to Paris over a 7-week period, regardless of individual contributions. Resistance on the bicycle and the app's responsiveness to pedal rotations are customized according to each participant's exercise tolerance. A map feature helps the individuals track their progress and increases motivation by allowing the participants to visualize their journey (Figure 2). While the app addresses logistical and environmental barriers, it is equally important to consider the motivational aspects that encourage sustained participation in home-based exercise programs. SDT provides a framework for considering these motivational drivers.

Figure 2. Map showing the team's progress on the virtual tour to Paris. With permission from 4Mvideo.

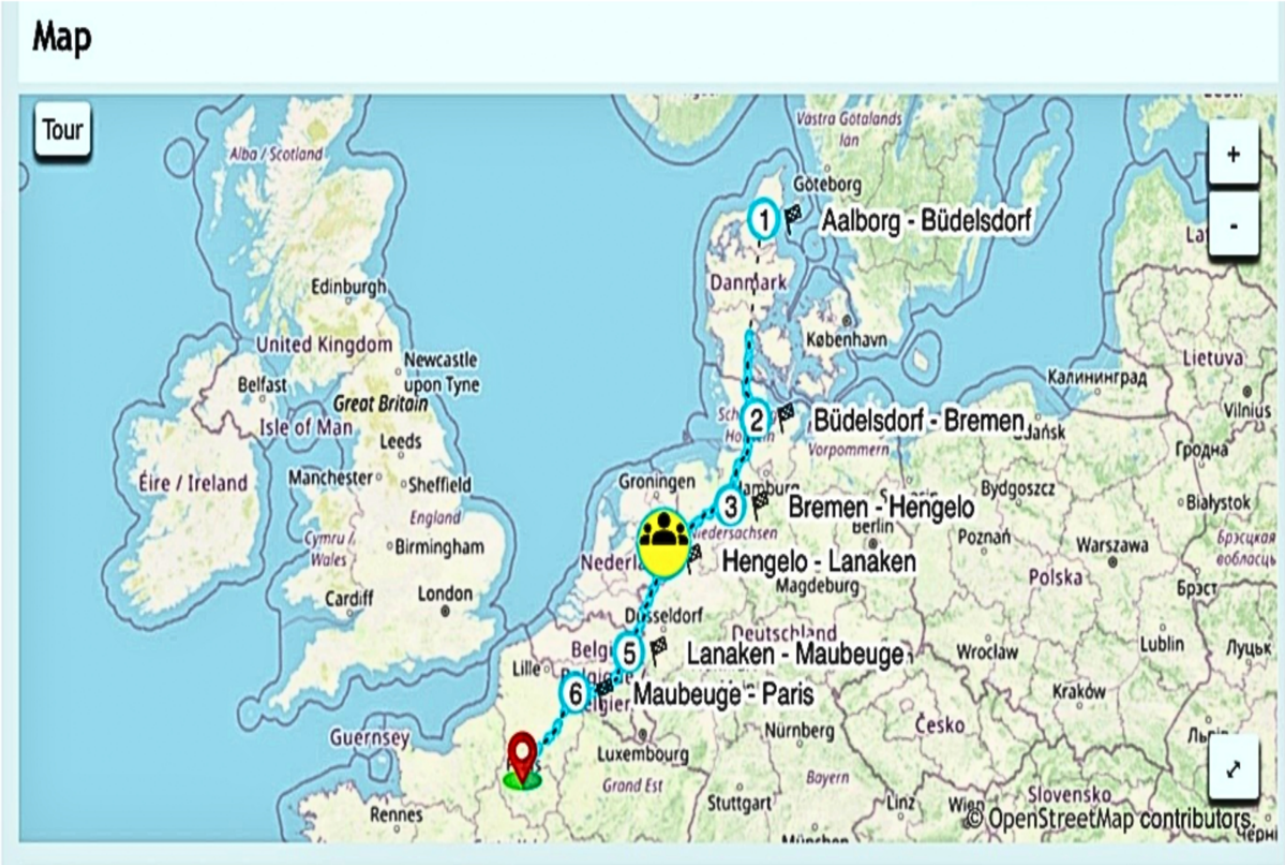

Self-Determination Theory

SDT offers a framework to understand motivational drivers [13]. SDT suggests that motivation is fostered by 3 basic psychological needs: *autonomy* (the sense of control over one's action), *competence* (the ability to master tasks), and *relatedness* (feeling connected to others) [13]. When these needs are fulfilled, the individuals are more likely to experience intrinsic motivation, leading to greater engagement in sustained behaviors, such as exercise.

To analyze the results of a home-based cycling exercise intervention for individuals with COPD, SDT provides a valuable framework for understanding motivation and behavior change in home-based cycling exercise intervention for individuals with COPD. The study aims to explore the development of a home-based cycling exercise intervention for individuals with COPD, focusing on aspects such as bicycle selection, app functionality, and pilot testing. Furthermore, it aims to explore participants' and nonparticipants' attitudes toward the intervention.

Methods

Study Design

This qualitative study was based on a phenomenological-hermeneutic approach, which aimed to explore participants' attitudes and experiences to inform the development of an intervention. This approach explored both the participants' lived experiences and the interpretation of these experiences [14]. It was conducted through individual semistructured telephone interviews. An inductive thematic analysis was used to analyze the data. This study was reported in line with the COREQ (Consolidated Criteria for Reporting Qualitative Research) guideline [15]. This qualitative study was a part of a noninferiority randomized controlled trial conducted at Aalborg University Hospital for the COPDtoParis project [16].

Informants

Inclusion criteria were defined as persons living with COPD with experiences relevant to the development of the intervention. Recruitment was based on a purposive sampling, selecting

XSL•FO
RenderX

informants who could provide diverse insights for the development of the intervention. The number of participants was determined based on information power, with recruitment concluding when the informants provided sufficient content to elucidate the research question [17]. The number of participants was evaluated based on whether each new informant contributed significant new information. Recruitment was concluded once thematic saturation had been reached and no substantial new themes or variations in depth were present in the data to address the study aim. Once sufficient variation and depth in the informants' responses had been reached to answer the research question, the data volume was considered adequate.

Groups selected to inform the development process were (1) test pilots, who were known from the research unit, had COPD, and were assessed by a doctor to be capable of handling home cycling; (2) participants in COPDtoParis; and (3) nonparticipants, individuals who were eligible but declined to participate in COPDtoParis. Additional characteristics of the informants were presented in Table 1.

Table 1. Informant characteristics.

Informant identification	Group	Age (years)	Sex
TP[a] 1	Test pilot	61	Female
TP 2	Test pilot	69	Female
PP[b] 3	Participated	71	Male
PP 4	Participated	75	Female
PP 5	Participated	72	Female
PP 6	Participated	76	Female
PP 7	Participated	78	Female
NP[c] 8	Participated	71	Male
NP 9	Nonparticipants	79	Female
NP 10	Nonparticipants	66	Female
NP 11	Nonparticipants	65	Female
NP 12	Nonparticipants	66	Male
NP 13	Nonparticipants	72	Female
NP 14	Nonparticipants	77	Female
NP 15	Nonparticipants	72	Female

[a]TP: test pilot.

[b]PP: participant.

[c]NP: nonparticipant.

Data Collection

The interview guide included open-ended questions, which allowed specific topics to be addressed while also leaving space to explore new subjects brought up by the informants. Individual semistructured telephone interviews were conducted with the test pilots, participants, and nonparticipants. Nonparticipants were asked “Why did you refrain from participating in the project?” and were encouraged to elaborate where relevant.

The interview guide was based on the SDT framework and covered topics related to the practical implications of applying a home-based cycle exercise for individuals with COPD (see Table 2).

XSL•FO
RenderX

Table 2. Interview guide.

Theme	Examples of content
Introduction	• Details provided to participants about the interview process and its purpose.
Autonomy	• "What motivated you to participate in the cycling project?" • "What motivates you to exercise on the bicycle?" • "What do you think about home exercise versus going out?"
Competence	• "How is it to use the bicycle? Is it easy or difficult to use?" • "What is your overall experience with the bicycle?"
Relatedness	• "How do you experience the social aspect through the tablet?" • "How do you feel you have been supported in your home exercising?" • "What additional support do you believe you need to succeed with your home exercise bicycle in the future?"
Brings the interview to a close	• "What do you feel was missing or could have been improved?"
Additional	• Opportunity for participants to elaborate if they felt any important topics had been overlooked.

All interviews were conducted by a researcher who had an existing relationship with the participants through their involvement in the COPDtoParis study. The interview guide was pilot tested during an initial interview, which resulted in only minimal adjustments. Therefore, the pilot interview was included as part of the data. The interviews were audio-recorded and transcribed. They were conducted from August 2024 to September 2024.

Data Analysis

The analysis followed Braun and Clarke's 6 phases of thematic analysis [18]. Transcripts were read naively to ensure familiarity. Initial coding was generated inductively and then reviewed collaboratively. Themes were developed through discussions among the research team [18]. Saturation was considered achieved when no new themes emerged from 3 consecutive interviews. Reflexivity was ensured through regular team discussions, with explicit attention to the lead researcher's role as a known figure to participants. While no formal triangulation was conducted, interpretations were validated collaboratively.

Ethical Considerations

The study was conducted in accordance with the ethical standards outlined in the Helsinki Declaration [19] and followed the data protection standards set by the Danish Data Protection Agency [20]. In Denmark, qualitative interview studies that do not include biological material do not require approval from a national ethics committee. Before participation, all informants were fully informed about the study's aim and provided oral consent. All data were anonymized and securely stored in accordance with the General Data Protection Regulation [20]. Participants did not receive financial compensation.

Results

The results reflected the perspectives of 15 informants, providing insights into their experiences with the home-based cycling exercise intervention. These experiences, derived from semistructured interviews, were analyzed into 7 themes that highlighted practical and emotional dimensions of the intervention, focusing on patterns and variances. Informants' quotations were included to illustrate the results.

Choosing the Bicycle: Stability Versus Portability

The selection and suitability of the bicycle was guided by the target group's physical limitations. High-performance bicycles were deemed unsuitable due to issues with balance and mobility, leading to the choice of pedal bicycles. Three models were tested by test pilots, and feedback resulted in selecting a stable, compact model. On smooth floors, the bicycle scooted forward while pedaling, but placing it on a thin yoga mat prevented this. It remained stable on carpeted surfaces.

The following quote emphasized the importance of the bicycle's practical design, including its physical size and space requirements:

> *I think the bicycle is practical because it doesn't take up as much space as a regular bicycle.* [Participant 8]

While participants appreciated the practicality and stability of the chosen bicycle, some found its weight problematic. For instance, one participant noted difficulty moving it:

> *I have a cleaning lady and I need to move it when she comes. I can't lift the bicycle myself because it's so heavy. Therefore, I would rather use my own pedal bicycle, which I can lift.* [Participant 6]

Overall, the design was deemed suitable for the participants:

> *I think it's wonderful because I can move it around as it suits me.* [Participant 5]
>
> *I think it's super. It has been easy to figure out how to use the bike.* [Participant 4]

Preferences varied significantly, and the feedback underscored the delicate balance between the bicycle's weight, which ensured stability, and the challenge it posed for some participants due to its heaviness. This highlighted the diversity in individual needs, with some participants valuing the added stability, while others found the weight to be a barrier due to their health

XSL•FO
RenderX

limitations. These variations demonstrated that no single bicycle could fully accommodate the needs of all participants.

Geographical Awareness and Video Quality

The videos in the cycling app received mixed feedback. Some informants appreciated the landscape views, feeling motivated by familiar landmarks, while others preferred alternative distractions like television. The video content's quality and relevance were highlighted as factors influencing engagement. Issues such as impaired hearing or visual barriers also impacted the usability of the videos.

> *I can't hear their talk because of my bad hearing. Therefore, I don't really use the videos but cycle while I watch TV.* [Test pilot 2]

The lack of identifiable landmarks or city signs led to a sense of disorientation, making it difficult for some participants to stay motivated.

> *I like the parts near home. When they cycle south of my hometown, I don't know where I am, and that's less motivating.* [Test pilot 1]

Due to feedback from the test pilots, a feature providing an overview linked to a map was added to the app. In this section of the app, participants can view which routes are associated with specific geographical locations. Furthermore, this feature offers a clear overview of which routes have already been completed and which still need to be cycled.

A participant appreciated the videos featuring beautiful landscapes, expressing that they felt like a part of the cycling group in the videos:

> *I enjoy biking along with the videos, it's cozy, beautiful landscape and I feel like I'm part of the trip. Other times, I watch TV while doing it. It varies.* [Participant 6]

The participants enjoyed watching the landscape while cycling; however, the participants felt the filming quality could be improved:

> *I think it's fun to sit and look at the landscape when I'm cycling. However, I sometimes think there are too many cyclists in front of the one filming, and it blocks my view of the landscape.* [Participant 4]

The virtual landscape was perceived as valuable for enhancing motivation among the participants. While some participants expressed a wish for more visually appealing content, their focus on the journey and the landscape, rather than the specific goal of reaching Paris, implied that the journey itself was more engaging than the destination.

The videos were considered motivating by some participants, while others actively chose other activities, such as watching television while cycling. This suggested that the videos were a supplementary feature appreciated by some participants. It was not the presence of the videos but rather the content that was highlighted as important for engagement.

Connecting Online: Knowing the Teammates

During the test pilot phase, the test pilots were asked if an online chat room could enhance their experience of cycling in terms of getting to know their teammates and improving motivation and adherence. The test pilots expressed a positive attitude toward the notion; therefore, the chat function was developed for the app.

All participants were asked if they wanted access to the chat function, which was intended to increase a sense of community and support. There were significant differences in what participants thought about the chat function. The commonality was that none of them had used the feature. Some participants expressed that they could not see the chat button because it was too small. Others emphasized that it was not a relevant function for them, as they did not want to chat with strangers:

> *To be honest, I don't think anyone will use it. Because you don't know the people and you don't know who they are.* [Participant 4]

Others stated that it might be interesting to try:

> *I haven't tried the chat function yet, maybe I will at some point. I haven't decided on that yet.* [Participant 3]
>
> *I might use the chat function at some point. I just need to start using the bicycle a little more. Anyway, I'm positive to try it out.* [Participant 7]

Based on the quotes, the chat function had not had the desired impact at the time. None of the participants reported using the chat feature, highlighting a disconnect between its intended purpose and participants' needs. Some participants mentioned that they would have been more willing to use the chat if they had the chance to meet the other participants first, as they would no longer be strangers.

Need for Individual Guidance

The participants showed variation in their capacity to manage intensity. Some required a lower intensity level with assistance, while others were able to self-manage effectively. One test pilot needed explicit guidance regarding the intensity of the exercise, as the test pilot reported difficulties in self-regulating. This led to negative physical consequences, such as muscle soreness. The muscle soreness caused concern for the test pilot, as illustrated by the following statement:

> *Today, my legs are very sore after cycling. I didn't feel the strain on my legs while I was cycling, but the day after I was so sore and could hardly walk.* [Test pilot 2]

This highlighted the importance of providing clear guidelines to prevent overtraining, particularly for participants with poor health conditions like COPD. It was also important to inform participants that some leg discomfort was normal and harmless.

Participants valued the ability to adjust the intensity level, as their physical capabilities varied daily due to energy levels and health conditions. This was expressed in a statement:

XSL•FO
RenderX

> *Some days I have the energy for a higher intensity, while other days I have less. I adjust it myself as it suits my energy level and mood.* [Participant 6]

This statement illustrated the necessity for individual adaptation to the intensity level that accommodated the participants' energy levels and moods. Explicit guidance was essential to help prevent overexertion and associated discomfort while allowing for individual adjustment based on daily variations in energy and capability. When providing tailored guidance on adjusting the intensity, no issues with intensity adjustment were reported by other participants.

Flexibility as an Important Factor

Another important theme among the participants was the flexibility provided by the home-based exercise intervention. They valued the ability to cycle at different times throughout the day based on personal preference, which gave them a sense of control over their exercise routines. One participant expressed this point:

> *If I feel like it at 10 AM, then it's frustrating if I must wait until 2 PM to cycle. Now I can go and sit down no matter what time of day. I decide when I have the air and energy for it.* [Participant 7]

The flexibility allowed participants to exercise when they felt physically capable, thereby enhancing adherence to exercise. Participants described significant variability in their exercise capacities, reflecting the need for flexible scheduling and personalized approaches:

> *One day it's hopeless, the second day it's okay and the third day it's great.* [Participant 3]

This flexibility was crucial for maintaining motivation and adherence, particularly for those dealing with poor health conditions. Autonomy in choosing exercise timing and adaptability to health conditions were key factors that supported participating in the intervention.

Barriers Related to Physical and Mental Strain

Energy levels and prioritization were significant factors influencing participation in the intervention. Nonparticipants described a range of barriers that were closely connected to their physical condition and mental state. For the participants, the most important challenge was severe symptoms such as exhaustion and shortness of breath, which made physical activity unmanageable. These symptoms often left them with little energy for exercise, as illustrated below:

> *I feel best when I'm lying down or sitting. If I'm physically active, I feel very unwell. I don't have the energy or strength to exercise...I'm constantly exhausted and could just sleep all the time.* [Nonparticipant 9]

Other nonparticipants chose not to prioritize exercise. For some, their priorities were everyday obligations like work, or they were already feeling sufficiently active through other activities. This was expressed in the following quote:

> *I feel that I'm already very active, so there's no need for more. There isn't much room for additional activity.* [Nonparticipant 13]

One nonparticipant had negative experiences with pedal bicycles and found cycling generally unattractive:

> *Cycling has never been for me, and I have had several of those pedal bicycle at home, and it has not been with any good results.* [Nonparticipant 15]

These accounts illustrate how nonparticipation was not only due to physical limitations but also reflected individual priorities and prior experiences with exercise, shaping attitudes toward exercise.

Beyond physical symptoms and practical considerations, nonparticipants described how stress and mental overload influenced their decisions. They expressed feeling overwhelmed by their health situation and the amount of health care offers they had to manage. This experience of overload created a mental barrier to participation, as described by a nonparticipant:

> *I have felt distressed ever since I was discharged from the hospital. I can't fit anymore into my home, everything is cluttered. I feel like I can't manage anymore.* [Nonparticipant 14]

The strain of balancing physical limitations, practical constraints, and mental distress made participation challenging for some. This ongoing stress since being discharged from the hospital has impacted their ability to manage additional responsibilities or activities, including the cycling intervention. These narratives from nonparticipants highlight how psychological and emotional strain enhanced practical barriers, making nonparticipation a coping strategy rather than a simple disengagement.

The Home Environment as Both Facilitator and Barrier

Participants and nonparticipants described the home environment as central to their experience of the cycling intervention. For participants, exercising at home offered comfort, accessibility, and flexibility. The ability to place the pedal bicycle in familiar spaces, such as the living room or dining area, meant they could exercise while engaging in everyday activities like watching television or using a computer. One test pilot highlighted the privacy and psychological ease of exercising at home:

> *When it comes to exercising, I prefer to do it alone. That's when I get the most out of it. I feel best exercising by myself and in familiar surroundings.* [Test pilot 1]

A sense of privacy and autonomy enhanced motivation and made exercise feel more sustainable for participants. Similarly, other participants emphasized that exercise at home ensured regular engagement, since going outside or to a fitness center would have been too demanding. For these participants, home-based exercise was not only more appealing but also allowed them to adapt routines to their fluctuating condition:

> *The fact that the exercise takes place at home ensures that I actually engage in exercise. If I had to go to a fitness center, I would never attend. By the time I*

XSL•FO
RenderX

> *arrived, I would already be exhausted.* [Participant 5]

In contrast, lack of space was a repeated barrier for nonparticipants. They described their homes as overcrowded with furniture and assistive devices such as walkers and oxygen flasks. Living in smaller apartments, combined with many appointments and care persons coming and going, led nonparticipants to express that they lacked space for more items or engagements. One nonparticipant stated:

> *I live in a 60 square meter apartment, and I already have an oxygen machine and a lot of flowers, so I simply don't have room for it. I know I could just put it aside, but I honestly don't have the energy to keep anything else here. I have just started receiving home care for cleaning, and I must move everything of the floor. I simply can't manage it anymore.* [Nonparticipant 10]

The nonparticipants' struggles with limited space highlighted the importance of considering the practical realities of living environments when introducing home-based interventions. For individuals already managing a high degree of clutter and aids, even small additional equipment felt overwhelming. These physical constraints often interacted with feelings of fatigue and stress, making participation in exercise interventions less feasible.

In contrast, participants found the bicycle compact and movable, which is why they did not report concerns about space and practical limitations. The differing experiences between participants and nonparticipants suggested that perceived or actual lack of space was not only a physical constraint but also a psychological barrier. For nonparticipants, the anticipation of managing additional equipment might have amplified feelings of being overwhelmed, even if the equipment itself was designed to be compact. This contrast shows how the home environment could either enable or obstruct exercise, depending on individual circumstances, thereby underscoring the contextual nature of participation.

Discussion

Summary of Key Findings

The development of a home-based cycling exercise for individuals with COPD aimed to overcome challenges associated with center-based exercise, which often presented logistical barriers, such as transportation and mobility issues. These barriers were particularly relevant for individuals with advanced COPD, who frequently experienced reduced mobility or required supplemental oxygen. Home-based exercise provided the advantage of enabling participation in familiar surroundings, aligning with previous findings that highlighted the benefits of remaining at home while participating in digital health interventions to promote comfort and autonomy [21].

The Journey Was More Engaging Than the Destination

The development and design of virtual maps, incorporating clear markers or familiar landmarks, aimed to enhance participants' motivation. However, the study results revealed that the journey itself was often more engaging for participants than achieving specific goals. This underlined the importance of improving the visual quality of videos by optimizing perspectives and reducing distractions. Beautiful surroundings and engaging visuals motivated participants, suggesting that future interventions should address the preferences of both those motivated by visual content and those who were goal oriented. In practice, this could involve offering a library of cycling routes with different themes (eg, familiar local areas, coastal landscapes, or urban routes), allowing participants to select according to personal preferences. Furthermore, technical improvements such as higher resolution filming and minimizing visual obstructions (eg, cyclists blocking the view) may increase immersion. To ensure scalability, such features could be integrated into existing digital health platforms, with physiotherapists or municipal health services providing initial guidance on setup and personalization. These concrete adjustments may help transform generic suggestions into actionable design elements that strengthen engagement and long-term adherence.

The Bicycles' Design

The participants expressed satisfaction with the bicycle's design and placement, but some encountered challenges due to the weight and portability of the bicycle. For individuals with COPD, who often experience reduced physical strength and mobility, the ability to easily relocate the bicycle within their home is important. It might be necessary in the future to offer different types of bicycles, a lightweight and a heavyweight, or provide guidance on creating a permanent and accessible space for the bicycle. This would enhance autonomy, a key element in SDT, by ensuring that participants feel in control of their exercise environment.

Autonomy and Competence—The Perspective of SDT

From the perspective of SDT, *autonomy* and *competence* were identified as particularly important motivational factors. *Autonomy* was facilitated by the flexibility to exercise when participants felt physically capable. Participants noted that the availability of the equipment at home was crucial for maintaining motivation. *Competence* was fostered through adjustable cycling intensity, which enabled participants to experience progress and mastery. Similarly, participants felt a sense of competence when they were able to plan their own exercise sessions. These results were consistent with prior research, which reported increased adherence to self-managed exercise programs among older individuals who felt competent [22].

Although autonomy and competence emerged as central motivational drivers in this study, the role of relatedness appeared less pronounced. Participants showed limited interest in using the chat function and emphasized privacy rather than social connectedness during exercise. This finding contrasts with the literature in digital health, where social support is often identified as a key facilitator of engagement and adherence to physical activity programs [23,24]. An explanation for this discrepancy may be the specific context of individuals living with COPD, many of whom experience fluctuating health, fatigue, and a strong need to conserve energy. For these individuals, the opportunity to exercise privately and on their

XSL•FO
RenderX

own terms may outweigh the potential benefits of social interaction. Another explanation could be the design of the intervention itself. Participants did not know each other prior to the study, and the virtual chat function may not have provided a sufficiently meaningful basis for social connectedness.

These findings indicate that relatedness should not be dismissed as unimportant, but rather reconsidered in terms of how it can be meaningfully facilitated for this specific population. Future interventions could explore hybrid approaches, such as introductory group meetings or small-group video sessions, to establish a sense of familiarity and trust before transitioning to digital platforms. At the same time, it is important to acknowledge that some individuals may prefer solitude and autonomy, highlighting the need for flexible designs that can accommodate varying preferences for social engagement. By situating our findings within the broader literature, it becomes evident that relatedness remains an important motivational factor, but its relevance and manifestation may depend strongly on the population, health status, and context in which the intervention is delivered.

Barriers From the Nonparticipants

Although exercise is widely known to be beneficial, this intervention was not suitable for everyone. Nonparticipants identified barriers such as lack of energy, stress, mental strain, and practical issues like space constraints. One nonparticipant's negative experiences with pedal bicycles highlighted the importance of accommodating personal preferences in exercise design.

Individual histories and past experiences also played a role in engagement. For example, a participant who expressed enthusiasm for the idea of cycling to Paris may have had prior experiences with group cycling or goal-oriented activities, which positively influenced motivation. In contrast, those less focused on collaborative or goal-driven aspects may not have shared the same background or attachment.

Nonparticipants express feeling overwhelmed by the multitude of health care offers and the overall demands of their condition. This indicated that the mental strain of managing health and daily life responsibilities often overshadowed the potential benefits of engaging in the home-based cycling exercise intervention. Living in smaller or cluttered spaces expressed frustration with the lack of physical room to accommodate the equipment.

To address these issues, interventions should consider strategies to reduce mental overload, such as simplifying health care offers and minimizing the introduction of additional equipment or activities into already crowded and stressful environments. Practical solutions, such as compact or foldable exercise equipment, could also mitigate space-related challenges.

The responses from nonparticipants indicate that this intervention could be most suited for a certain group of individuals with COPD: those who are significantly impacted by their condition, largely homebound, but not dependent on personal care or disability aids. It did not appeal as much to those still capable of and motivated by engaging in activities and exercise outside their home.

Our findings are consistent with earlier work showing that individuals with COPD often prefer home-based exercise because it offers autonomy, privacy, and flexibility in order to be engaging [25,26]. The barriers and motivators we found in our study were further reflected in a systematic review by Robinson et al [27]. They identified key subthemes affecting engagement, including self-efficacy, feedback, support, peer interaction, and creating routines. Likewise, they found common barriers such as exhaustion, space constraints, and emotional strain, which were also reflected in a study [27,28].

Further studies are needed to identify exercise options that could be motivating and acceptable for the nonparticipants, who declined from participating due to resource limitations. While well-established exercise programs already exist for the more active individuals, in line with the points made by Wouters et al [9], there is a need to develop and investigate tailored exercise offers for less mobile patients. COPDtoParis could provide a relevant low-barrier exercise alternative for homebound patients with COPD motivated by home-based options. Tailoring the intervention further to this group could help fill a gap in existing offers, addressing both physical limitations and practical constraints.

Strengths and Limitations

Strengths

A strength of this study was the inclusion of both participants and nonparticipants, which allowed for a more comprehensive understanding of the barriers and facilitators to engagement in the home-based cycling exercise intervention. By capturing insights from both groups, the study not only highlighted the experiences of those who engaged in the intervention but also the reasons for nonparticipation, offering a more nuanced understanding of the intervention's reach and potential areas for improvement. Another strength was the theoretical foundation of SDT for both the interviews and the analysis. This theoretical aspect ensured that the results were not only descriptive but also linked to a broader psychological understanding. The thematic analysis by Braun and Clarke's [18] method strengthened the study's rigor. By following a structured, step-by-step approach to identify, analyze, and report themes, the analysis remained transparent and systematic. One researcher conducted the analytical process, ensuring consistency in coding and theme development, while the other researchers validated these results through discussions. Furthermore, the analysis was presented transparently and systematically, with quotes representing nearly all the informants, which enhanced the trustworthiness of the analysis and results.

Limitations

As the interviewer was familiar with the participants, reflexivity was emphasized throughout the process. An audit trail documenting coding decisions was maintained, and coding structures were discussed within the research team to ensure transparency and minimize bias.

Although the inclusion of nonparticipants is a strength of this study, the depth of their accounts was more limited compared to participants' interviews. Contact with nonparticipants was

XSL•FO
RenderX

less extensive, and data were collected through telephone interviews. This restricted the degree of thematic interpretation.

While SDT provided a useful motivational lens, relying solely on this framework may have limited our analysis. Other models, such as the COM-B (Capability, Opportunity, and Motivation for Behavior Change) model, could have provided a more comprehensive behavioral systems approach, particularly in relation to environmental and capability-related barriers. Similarly, the UTAUT (Unified Theory of Acceptance and Use of Technology) framework might have offered insights into participants' adoption of the app and its technological usability. Future research could benefit from combining SDT with complementary frameworks to capture both motivational and structural determinants of engagement.

Conclusions

This study highlights the value of home-based cycling for individuals with COPD, particularly the importance of autonomy, flexibility, and tailored guidance. For clinicians, the findings suggest offering clear instructions on intensity adjustment and considering patients' daily symptom fluctuations. For developers, practical improvements include customizable training schedules and enhanced video quality with diverse route options. Barriers such as limited space and health care overload underline the need to assess home environments and current treatment burden before enrollment, which may inform both exclusion criteria and supportive services. Future research should evaluate the intervention in randomized controlled trials, cost-effectiveness, and tailoring based on SDT profiles.

Conflicts of Interest

None declared.

References

1. Smoking is the leading cause of chronic obstructive pulmonary disease. World Health Organization. 2023 Nov 15. URL: https://www.who.int/news/item/15-11-2023-smoking-is-the-leading-cause-of-chronic-obstructive-pulmonary-disease [accessed 2025-10-16]
2. Jacobsen PA, Kragholm KH, Torp-Pedersen C, Janssen DJ, Spruit M, Weinreich UM. Employment status, readmission and mortality after acute exacerbation of COPD. Int J Chron Obstruct Pulmon Dis 2021;16:2257-2265 [FREE Full text] [doi: 10.2147/COPD.S319840] [Medline: 34385815]
3. Physical activity and COPD. American Lung Association. URL: https://www.lung.org/lung-health-diseases/lung-disease-lookup/copd/living-with-copd/physical-activity [accessed 2025-10-16]
4. McCarthy B, Casey D, Devane D, Murphy K, Murphy E, Lacasse Y. Pulmonary rehabilitation for chronic obstructive pulmonary disease. Cochrane Database Syst Rev 2015;2015(2):CD003793 [FREE Full text] [doi: 10.1002/14651858.CD003793.pub3] [Medline: 25705944]
5. Kjærgaard JL, Juhl CB, Lange P, Wilcke JT. Early pulmonary rehabilitation after acute exacerbation of COPD: a randomised controlled trial. ERJ Open Res 2020;6(1):00173 [FREE Full text] [doi: 10.1183/23120541.00173-2019] [Medline: 32083113]
6. Thorpe O, Kumar S, Johnston K. Barriers to and enablers of physical activity in patients with COPD following a hospital admission: a qualitative study. Int J Chron Obstruct Pulmon Dis 2014;9:115-128 [FREE Full text] [doi: 10.2147/COPD.S54457] [Medline: 24489465]
7. Oates GR, Niranjan SJ, Ott C, Scarinci IC, Schumann C, Parekh T, et al. Adherence to pulmonary rehabilitation in COPD: a qualitative exploration of patient perspectives on barriers and facilitators. J Cardiopulm Rehabil Prev 2019;39(5):344-349 [FREE Full text] [doi: 10.1097/HCR.0000000000000436] [Medline: 31348127]
8. Gloeckl R, Zwick R, Fürlinger U, Jarosch I, Schneeberger T, Leitl D, et al. Prescribing and adjusting exercise training in chronic respiratory diseases – expert-based practical recommendations. Pulmonology 2023;29(4):306-314 [FREE Full text] [doi: 10.1016/j.pulmoe.2022.09.004] [Medline: 36272962]
9. Wouters EF, Wouters BB, Augustin IM, Houben-Wilke S, Vanfleteren LE, Franssen FM. Personalised pulmonary rehabilitation in COPD. Eur Respir Rev 2018;27(147):170125 [FREE Full text] [doi: 10.1183/16000617.0125-2017] [Medline: 29592864]
10. Cycling – health benefits. Better Health Channel, Department of Health, State Government of Victoria, Australia. URL: https://www.betterhealth.vic.gov.au/health/healthyliving/cycling-health-benefits [accessed 2025-10-16]
11. A nation of cyclists. Ministry of Foreign Affairs of Denmark. URL: https://denmark.dk/people-and-culture/biking [accessed 2025-10-16]
12. Facts and figures on cycling in Denmark. VisitDenmark. URL: https://www.visitdenmark.com/press/latest-news/facts-and-figures-cycling-denmark [accessed 2025-10-16]
13. Ryan RM, Deci EL. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. Am Psychol 2000;55(1):68-78. [doi: 10.1037//0003-066x.55.1.68] [Medline: 11392867]
14. van Manen M. Phenomenology of practice. Phenomenol Pract 2007;1(1):11-30. [doi: 10.29173/pandpr19803]
15. Tong A, Sainsbury P, Craig J. Consolidated Criteria for Reporting Qualitative Research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]

XSL•FO
RenderX

16. Stoustrup AL, Thomsen LP, Andreasen J, Palsson TS, Weinreich UM. Cluster randomised controlled trial on the effects of long-term home-based exercise for patients with chronic obstructive pulmonary disease with recent exacerbation: research protocol of the COPDtoParis Project. BMJ Open Respir Res 2025 Jan 19;12(1):e002573 [FREE Full text] [doi: 10.1136/bmjresp-2024-002573] [Medline: 39832888]
17. Malterud K, Siersma VD, Guassora AD. Sample size in qualitative interview studies: guided by information power. Qual Health Res 2016;26(13):1753-1760. [doi: 10.1177/1049732315617444] [Medline: 26613970]
18. Braun V, Clarke V. Thematic Analysis: A Practical Guide. 1st ed. London: Sage Publications; 2022.
19. WMA Declaration of Helsinki—ethical principles for medical research involving human subjects. World Medical Association. 2024 Oct 19. URL: https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/ [accessed 2025-10-16]
20. GDPR: General Data Protection Regulation [Databeskyttelsesforordningen (Lovgivningen)]. URL: https://gdpr.dk/databeskyttelsesforordningen/ [accessed 2025-10-16]
21. Ienca M, Schneble C, Kressig RW, Wangmo T. Digital health interventions for healthy ageing: a qualitative user evaluation and ethical assessment. BMC Geriatr 2021;21(1):412 [FREE Full text] [doi: 10.1186/s12877-021-02338-z] [Medline: 34215209]
22. Mansson L, Lundin-Olsson L, Skelton DA, Janols R, Lindgren H, Rosendahl E, et al. Older adults' preferences for, adherence to and experiences of two self-management falls prevention home exercise programmes: a comparison between a digital programme and a paper booklet. BMC Geriatr 2020;20(1):209 [FREE Full text] [doi: 10.1186/s12877-020-01592-x] [Medline: 32539711]
23. Smith GSE, Moyle W, Burton NW. The relationship between social support for physical activity and physical activity across nine years in adults aged 60-65 years at baseline. Int J Environ Res Public Health 2023;20(5):4531 [FREE Full text] [doi: 10.3390/ijerph20054531] [Medline: 36901538]
24. Lindsay Smith G, Banting L, Eime R, O'Sullivan G, van Uffelen JGZ. The association between social support and physical activity in older adults: a systematic review. Int J Behav Nutr Phys Act 2017;14(1):56 [FREE Full text] [doi: 10.1186/s12966-017-0509-8] [Medline: 28449673]
25. Lahham A, McDonald CF, Mahal A, Lee AL, Hill CJ, Burge AT, et al. Home-based pulmonary rehabilitation for people with COPD: a qualitative study reporting the patient perspective. Chron Respir Dis 2018;15(2):123-130 [FREE Full text] [doi: 10.1177/1479972317729050] [Medline: 28868892]
26. Santos CD, Rodrigues F, Caneiras C, Bárbara C. From inception to implementation: strategies for setting up pulmonary telerehabilitation. Front Rehabil Sci 2022;3:830115 [FREE Full text] [doi: 10.3389/fresc.2022.830115] [Medline: 36188951]
27. Robinson H, Williams V, Curtis F, Bridle C, Jones AW. Facilitators and barriers to physical activity following pulmonary rehabilitation in COPD: a systematic review of qualitative studies. NPJ Prim Care Respir Med 2018;28(1):19 [FREE Full text] [doi: 10.1038/s41533-018-0085-7] [Medline: 29867117]
28. Russell S, Ogunbayo OJ, Newham JJ, Heslop-Marshall K, Netts P, Hanratty B, et al. Qualitative systematic review of barriers and facilitators to self-management of chronic obstructive pulmonary disease: views of patients and healthcare professionals. NPJ Prim Care Respir Med 2018;28(1):2 [FREE Full text] [doi: 10.1038/s41533-017-0069-z] [Medline: 29343739]

Abbreviations

COM-B: Capability, Opportunity, and Motivation for Behavior Change
COPD: chronic obstructive pulmonary disease
COREQ: Consolidated Criteria for Reporting Qualitative Research
PR: pulmonary rehabilitation
SDT: self-determination theory
UTAUT: Unified Theory of Acceptance and Use of Technology

Edited by A Mavragani, J Sarvestan; submitted 13.01.25; peer-reviewed by S Mandal; comments to author 21.07.25; revised version received 05.09.25; accepted 19.09.25; published 14.11.25.

Please cite as:
Christensen KK, Bøggild H, Weinreich UM, Stoustrup AL
Development of a Virtual Home-Based Cycling Intervention for Individuals With Chronic Obstructive Pulmonary Disease: Qualitative Study
JMIR Form Res 2025;9:e71234
URL: https://formative.jmir.org/2025/1/e71234
doi:10.2196/71234
PMID:

XSL•FO
RenderX

©Kristina Krogh Christensen, Henrik Bøggild, Ulla Møller Weinreich, Anna Lei Stoustrup. Originally published in JMIR Formative Research (https://formative.jmir.org), 14.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Development, Challenges, and Evolution of the Log2Lose Intervention for Weight Management: Randomized Controlled Digital Health Trial

Ryan Jeffrey Shaw[1*], RN, PhD; Hailey Miller[2*], RN, PhD; Cherie Barnes[1*], RD, MS; Clemontina Davenport[3*], PhD; Sarah Morton-Oswald[4*], MS; Sarah Jackson[5*], MPH; Michelle Bean[5*], RD, MPH; Jeff Cohen[6*], BS; Chance Griffin[6], BS; Jane Pendergast[4*], PhD; Maren Olsen[4,7*], PhD; Jennifer Gierisch[5,7*], PhD; Corrine Voils[8,9*], PhD

[1]School of Nursing, Duke University, Durham, NC, United States
[2]School of Nursing, Johns Hopkins University, Baltimore, MD, United States
[3]Department of Biostatistics and Data Science, Wake Forest University, Winston-Salem, NC, United States
[4]Department of Biostatistics & Bioinformatics, Duke University School of Medicine, Duke University, Durham, NC, United States
[5]Department of Population Health Sciences, Duke University School of Medicine, Duke University, Durham, NC, United States
[6]Purple Workshops, Chicago, IL, United States
[7]Duke University, Durham VA Health Care System, Durham, NC, United States
[8]Division of Epidemiology, Department of Internal Medicine, University of Utah, Salt Lake City, UT, United States
[9]William S Middleton Memorial Veterans Hospital, Madison, WI, United States
[*]these authors contributed equally

Corresponding Author:
Ryan Jeffrey Shaw, RN, PhD
School of Nursing
Duke University
307 Trent Dr
DUMC 3322
Durham, NC, 27710
United States
Phone: 1 9196849434
Email: ryan.shaw@duke.edu

Abstract

Background: Long-term adherence to weight loss behaviors is challenging, as most individuals who achieve significant weight loss regain 1-2 kg per year. Financial incentives can reinforce weight-loss initiation and maintenance behaviors, but optimal strategies remain unclear.

Objective: This paper describes the design, technical architecture, and operational workflow of Log2Lose, a 5-year, multisite randomized controlled trial testing different financial incentive strategies to promote weight loss and maintenance. We detail the platform's integration with external devices, automated data collection, and adaptations to maintain intervention fidelity in the context of evolving technology and regulatory requirements.

Methods: The Log2Lose platform collects daily weight and dietary data from cellular scales and fitness tracking applications, calculates weekly incentive eligibility, and sends automated feedback and motivational text messages. We summarize the technical adaptations, message delivery performance, data completeness, and the balance between automation and manual support required to ensure data integrity.

Results: By the end of the study, 706 participants recorded 181,285 weights and 114,144 daily calorie entries. The platform sent 126,283 text messages and calculated 35,187 incentive payments, with 99.4% (34,976/35,187) processed automatically. Adaptations addressed device integration changes, application programming interface discontinuations, and new text messaging regulations. Despite automation, ongoing technical support was essential for resolving delivery errors, device issues, and data anomalies.

Conclusions: Log2Lose demonstrated that large-scale, fully remote weight loss interventions can be implemented using consumer technology paired with a robust, adaptable platform. Success depended on flexible software design, continuous monitoring, and

XSL•FO
RenderX

responsive technical support to navigate regulatory and technological changes. Log2Lose offers a practical model for processing remotely collected longitudinal dietary and weight data, providing valuable guidance for researchers, health care providers, and employers developing similar digital health interventions.

Trial Registration: ClinicalTrials.gov NCT04770909; https://clinicaltrials.gov/study/NCT04770909

(JMIR Form Res 2025;9:e70842) doi:10.2196/70842

KEYWORDS

weight loss; weight maintenance; diet; mobile phone; remote monitoring

Introduction

Over 40% of US adults have obesity [1], increasing the risks for diabetes, hypertension, and higher medical costs [2]. Although numerous weight loss programs achieve an average weight loss of at least 5%, long-term adherence remains a challenge, with most individuals regaining 1-2 kg per year [3].

Financial incentives can reinforce weight-loss initiation and maintenance behaviors, but optimal strategies remain unclear. Although evidence suggests that incentivizing both weight loss and dietary self-monitoring may be effective, it is unknown whether long-term outcomes are best supported by reinforcing the outcome of interim weight loss, the behavior of dietary self-monitoring, or both [4]. In addition, most prior studies distributed rewards at the end of the program rather than as weight loss occurs [5-10]. Incentive timing may influence outcomes, as continuous reinforcement can accelerate behavior acquisition, whereas intermittent reinforcement can sustain it [11]. Little is known about how to structure and implement incentives in remote interventions to maintain adherence to weight loss behaviors over time [12,13].

To address these gaps, we launched Log2Lose, a 5-year, multisite randomized controlled clinical trial in September 2020, to test whether reinforcing interim weight loss, dietary logging, or both improves the proportion of people who achieve clinically significant weight loss over 78 weeks (18 months) relative to a control group with no incentives.

Originally designed to include in-person group meetings, telephone calls, incentives, text messaging, and in-person data collection, the trial pivoted to fully remote delivery during the COVID-19 pandemic. Since launch, the technology landscape has evolved, requiring adaptations to account for changes in ownership, business agreements, and consumer-protection policies among technology providers. In March 2023, the US Federal Communications Commission implemented its first consumer protection rules targeting text messaging [14], requiring mobile providers to block certain messages deemed likely to be illegal or harmful. Concurrently, mobile service providers deployed software to prevent spam- and fraud-prevention software, increasing the risk that legitimate intervention messages would be flagged as spam or “fail to send” if classified as potentially unwanted or high-risk (eg, high-risk financial services, debt collection/forgiveness, “get-rich-quick” schemes) [15].

The primary objective of this study is to describe the Log2Lose intervention’s technical architecture, integration with external devices, and operational workflows, including automated and manual procedures for data monitoring and incentive processing. We detail adaptations made to maintain intervention fidelity amid regulatory and technological changes—which can be expected in a 5-year trial—and we present operational outcomes, including message delivery rates, device data completeness, and weight measurement patterns, along with lessons learned for designing scalable, technology-supported weight-loss interventions.

Methods

Trial Design

This study’s primary objective was to assess the efficacy of financial incentive strategies for achieving person-level loss of ≥5% of baseline weight. The trial spanned three 6-month phases: phase I, incentivized initiation with 13 biweekly group sessions; phase II, incentivized maintenance with 3 monthly group sessions plus 5 monthly individual calls; and phase III, a 26-week, nonincentivized maintenance program with 3 bimonthly calls. Group sessions and calls were led by a registered dietitian.

Participants

Participants with obesity were recruited in Madison, WI, and Durham, NC, USA. The inclusion and exclusion criteria are detailed in ClinicalTrials.gov NCT04770909 and published [16]. Related to the Log2Lose technology, participants met requirements if they were able to speak and read English; able to download and use the Fitbit and MyFitnessPal smartphone apps daily; owned a smartphone with a data and texting plan; had reliable internet access; had the ability to participate in video calls via a smartphone, tablet, or computer; and as described below, later in the trial agreed to create a Gmail account after Google acquired Fitbit.

Interventions

Participants were randomized to one of the 4 arms: incentives for both dietary self-monitoring and weight loss, with partial incentives if only 1 criterion was met (arm 1, combined); incentives for dietary self-monitoring only (arm 2, diet only); incentives for weight loss only (arm 3, weight only); or no financial incentives (arm 4, control) [16].

All participants received a BodyTrace cellular scale and were encouraged to weigh themselves regularly: twice weekly in phase I (weight loss) and once weekly in phase II (maintenance). Participants were instructed to keep the scale on a stable surface and restrict use by others. Participants were asked to track daily food and beverage by using a smartphone app and at 6-month intervals, to wear an activity tracker for 7 days.

XSL•FO
RenderX

All participants received automated motivational text messages twice weekly for 78 weeks (phases I-III), reinforcing educational themes and behavioral skills from group sessions and counseling calls. In phases I and II, participants in incentivized arms received weekly text messages indicating whether they had earned an incentive. Messages were sent between approximately 8 AM and 11 AM local time, with no adjustment for travel across time zones. Due to the COVID-19 pandemic, remote weights replaced in-person measurements as the primary outcome. Prior research supports the validity of remote monitoring compared with in-person weighing, with the added convenience of avoiding travel [17].

Control-arm participants received the same devices and platform access but no financial incentives. All participants received the same frequency and timing of motivational text messages, which were standardized prompts rather than tailored feedback. No additional behavioral counseling was provided via the platform.

Outcomes

This study reports the outcomes for the total number of text messages received by day of the week and message type, the delivery success rate, the total number of weights transmitted, the average number of daily weigh-ins per participant, and the total number of calorie entries received. It summarizes the number of incentive payments processed automatically by the software compared to those processed manually. Finally, we report common technical and troubleshooting examples.

Sample Size and Randomization

Participants were randomized equally to one of the 4 arms by using a computer-generated block randomization sequence accessible only by the study statisticians. Participants were blinded to the incentive criterion of other arms. Sample size and power calculations were conducted using the methodology for multi-arm tests of proportions in PASS 2019. Further details are available in the published protocol paper [16].

Ethical Considerations

Study conduct was overseen by the University of Wisconsin-Madison institutional review board (approval 2020-1693) and a Data Safety Monitoring Board that met at least annually throughout the study period for ethics review and approval. All participants provided informed consent and could opt out of the trial at any time. People completed a web-based prescreening survey, followed by telephone screening and then final in-person eligibility screening and informed consent. Data were stored in secure password-protected databases only accessible to institutional review board–approved study staff. Compensation was provided for assessment completion at weeks 26 (US $40), 52 (US $25), and 78 (US $35).

Platform Description

The Log2Lose platform was developed to automate data collection, incentive processing, and monitoring procedures in support of the randomized trial. The system integrated weight and dietary data from consumer devices and applications, applied prespecified algorithms to determine incentive eligibility, and delivered feedback to participants via text messages. Research staff had access to a dashboard that allowed monitoring of data completeness, device connectivity, and message delivery. These features were designed to maintain intervention fidelity and to provide safeguards against data anomalies or technical failures. Importantly, the platform was developed to be adaptable, enabling adjustments to accommodate vendor-driven changes such as the discontinuation of the MyFitnessPal–Fitbit application programming interface (API) and shifting the US Federal Communications Commission text messaging regulations, which might otherwise have compromised study integrity.

The platform applied prescheduled weekly algorithms to determine incentive eligibility (Figure 1). It operated on a Heroku platform-as-a-service cloud server, utilizing a Ruby on Rails application linked to a PostgreSQL database. Data from BodyTrace scales (weights), Fitbit activity trackers (steps per day, minutes active), and Fitbit and MyFitnessPal dietary apps (calories logged per day) were securely imported into the platform using a web API. The BodyTrace scale measured weight and had an embedded cellular chip using the global system for the mobile communications network, which transmitted weight data to the BodyTrace servers. Each scale's international mobile equipment identity number was recorded at enrollment, linked to the participant ID, and used to retrieve data via API.

During enrollment, participants created a Fitbit account and received an authorization link from the platform. Upon authorization, the platform linked the Fitbit account to the participant ID, enabling import of daily calories, steps, and activity minutes via API. MyFitnessPal did not offer an API to pull data into the platform. Therefore, participants also linked their Fitbit account to the MyFitnessPal app so that calorie data from MyFitnessPal would transfer over to Fitbit.

Participant information (eg, enrollment, surveys) from the Research Electronic Data Capture (REDCap) system, a secure online platform for data capture and storage, was integrated into Log2Lose via API for analysis and study management [18,19]. This integration also included automated notifications for participants and staff.

XSL•FO
RenderX

Figure 1. Overview of the study software infrastructure, which securely integrates data from participant devices (cellular scale, activity tracker, diet app) and incentive payments with study software using a web API. The software operated on a Heroku platform-as-a-service cloud server, utilizing a Ruby on Rails application linked to a PostgreSQL database. Data are transferred via secure protocols to a central database, enabling automated messaging, incentive processing, and synchronization with REDCap for staff oversight. API: application programming interface; MFA: multi-factor authentication; REDCap: Research Electronic Data Capture.

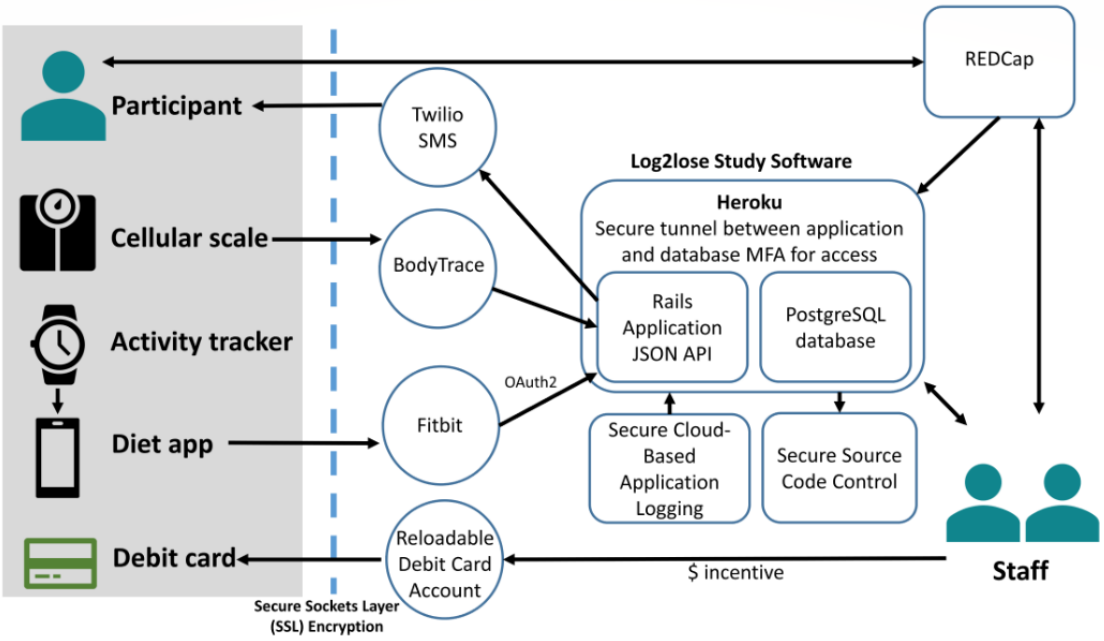

Log2Lose Platform

User Interface

The platform used a messaging API to display study metrics on interactive dashboards (Multimedia Appendices 1 and 2), including participant data, message counts, delivery rates, and text message errors. Additional dashboard features included calorie counts, weights, incentive earnings, participant demographics, Fitbit connection status, and scale battery levels, with all data downloadable. The platform (1) provided interactive dashboards for research staff; (2) generated comprehensive summaries of calorie entries, weigh-ins, incentives, and text message statistics, including detailed views of failed messages; and (3) displayed participant information by study site and cohort, Fitbit connection status, scale signal, and battery status. These features supported operational and safety decisions such as verifying anomalous weights, resolving data integration issues, adjudicating incentive eligibility in edge cases (eg, device outages), and triggering participant outreach when adherence abruptly declined. Decisions were made by trained research coordinators in accordance with the study protocol, with escalation to the study clinician for safety-related concerns. Many platform features were added iteratively in response to emerging operational needs. Figures S1A and S1B in Multimedia Appendix 1 provide detailed views of the dashboards.

The platform allowed filtering of participant data by cohort or site and provided individual profiles displaying study details, timestamps, carrier information, and recent Fitbit and BodyTrace data. Links to full historical records are available from each profile (Multimedia Appendix 2).

Incentive Algorithms, Processing, and Incentive Text Messages

Each week, the platform determined whether participants qualified for incentives for dietary self-monitoring, weight loss, or both, based on data from the participant's cellular scale and fitness application, their randomized study arm, and their position in the study timeline (Multimedia Appendix 1).

This payment algorithm was embedded within the Log2Lose software (hosted on GitHub with secure cloud-based application logging), which generated weekly incentive reports identifying eligible participants. Payments were processed in near real time by a research team member and loaded onto study-provided debit cards.

Participants were notified of their weekly incentive status via Twilio text messages sent on Mondays; if Monday was a holiday, messages were sent the following day. The delivery rate was calculated as the number of messages accepted by the downstream service (carrier for SMS text messages; provider API for push/email) divided by the number attempted by Log2Lose. The success rate was calculated as the proportion of delivered messages that received a positive delivery receipt from the downstream service. Message opens were not tracked (see Limitations).

Data Quality and Monitoring

Automated Procedures

Because errant weights were expected from study-provided scales, the platform automatically screened and flagged measurements showing implausible deviations (>3%) from the most recent accepted weight. Such deviations may have resulted from another person using the scale or user error, such as not calibrating the scale prior to use. Participants were also asked

XSL•FO
RenderX

to weight twice in close succession if the first weight seemed unreasonable. For incentive calculations, the measurements were labeled as rejected, and they were not considered in determining whether an incentive was earned. The system also monitored syncing issues with retrieving dietary data.

Manual Procedures

The statistical team conducted regular checks to detect patterns of repeated rejected weights and notified research coordinators for manual review. Several situations may prompt such review. For example, if a long time had passed since a participant last used the scale, their weight may fall outside the 3% threshold from their previous accepted measurement. In these cases, the participant was contacted, their weight verified, and the last accepted weight was manually reset. Another common scenario occurred when another household member used the scale, generating a weight outside the threshold. Additionally, participants may contact a study coordinator if they believed they incorrectly did not receive an incentive. The coordinator then reviewed the data to determine whether a correction was needed—usually for an unusual situation. If any changes were made, the coordinator received approval from the principal investigators. Such corrections were rare.

Statistical Analyses

We summarized the total number of messages received by day of the week and by type of message (incentive or motivational), as well as the delivery success rate. We also summarized the total number of weights transmitted, the average number of daily weigh-ins per participant, and the total number of calorie entries received. We described the number of incentive payments processed automatically by the software versus those processed manually. Descriptive statistics were used to summarize study participation, message distribution, weight recordings, calorie entries, and incentive payments. Frequencies and percentages were calculated for categorical outcomes, including message type, delivery success, and incentive processing mode, while means, standard deviations, and ranges were used for continuous outcomes such as daily message counts, number of weight recordings per participant, and calorie entries. Temporal patterns in logging behavior were described narratively in relation to technical disruptions, and rare events such as delivery failures and manual incentive processing were reported descriptively given their low frequency.

Results

A total of 706 participants were randomized and provided a baseline weight—348 and 358 across Wisconsin and North Carolina, respectively. Final data collection was completed in April 2025.

Text Messaging

By the end of this study, 126,283 messages were sent, with a daily maximum of 1414. Message distribution was balanced between sites: 62,276 to 348 participants in Wisconsin and 64,007 to 357 participants in North Carolina. One participant withdrew before receiving the first message; thus, 705 participants received messages.

Incentive messages were sent primarily on Mondays (n=26,111) and to a lesser extent on Tuesdays (n=300). Motivational messages were sent on Tuesdays (n=50,352) and Fridays (n=49,319). An additional 201 messages were sent manually on a Wednesday to report a technical error.

Overall, >99% of the messages were delivered successfully. Delivery failures typically resulted from international travel, prolonged service loss, or extended periods with phones turned off. The most common failure type involved messages flagged as spam by mobile carriers despite Twilio's trusted-source agreements. Possible triggers included absence of sender identification, use of dollar signs or shorthand characters, excessive exclamation marks, and repetitive content, although carriers did not disclose specific criteria.

Weights

By study end, 706 participants recorded a total of 181,285 weight measurements from the cellular scales. Of these, 94.9% (172,096/181,285) were accepted. The highest number of weights transmitted in a single day was 349. On weighing days, participants recorded an average of 1.24 (SD 0.316) measurements, reflecting multiple consecutive weigh-ins.

Calorie Data

There were 114,144 daily calorie entries from 694 participants, with a single day maximum of 256. Logging declined sharply around the 90-day mark, partly due to MyFitnessPal–Fitbit API issues in December 2021. Initially, Fitbit data were retrieved via MyFitnessPal for cohort 1, but persistent API problems prompted a switch to Fitbit-only logging. This improved reliability by removing third-party syncing, though participants reported lower satisfaction with Fitbit's interface. Google's acquisition of Fitbit required all users to have Google accounts by the end of 2025. In July 2024, Fitbit consolidated its dashboard into the mobile app, ending web access and affecting our final cohort.

Incentives

The platform automatically calculated and queued 35,187 incentives during phases I and II. Only 222 incentives (0.6%) across both sites required manual processing, mainly due to technical or behavioral complexities (Figure 2; Table 1).

XSL•FO
RenderX

Figure 2. Incentive algorithms. Participants with obesity were randomized to one of four incentive algorithms, which were applied to data collected from a diet app, a weight scale, both, or neither across two study phases. Phase I (weeks 1–26) focused on weight loss, providing rewards for diet logging, weight loss, or both. Phase II (weeks 27–52) emphasized weight maintenance within a defined threshold while continuing diet logging. Algorithm rules specified the required logging frequency, calorie minimums, and weigh-in schedules. The four study arms were: combined incentives, diet-only, weight-only, and no-incentive control. API: application programming interface; REDCap: Research Electronic Data Capture.

Weight Loss Initiation (Phase I Weeks 1-26)	Weight Maintenance (Phase II Weeks 27-52)
Arm 1: Diet Logging + Weight-Based Incentive How to earn weekly incentives: • Log your diet and lose weight = 100% incentive (50% if only one achieved) • Log your diet only = 50% incentive • Lose weight only = 50% incentive For diet logging incentives: • Log food at least 5 days/week, with at least 1 weekend day • Log at least 1000 calories/day (female at birth) or 1200 calories/day (male at birth) • Food logs are checked nightly before midnight (48-hour look-back, except Mondays = 24 hours) For weight loss incentives: • Weigh in at least 2 times/week • Last weight of the week must be lower than the first weight of the week • First daily weight = lowest reading within a 5-minute window • Week = Monday midnight to Sunday 11 PM	**Arm 1: Diet Logging + Weight-Maintenance-Based Incentive** How to earn incentives: • Log your diet and keep weekly weight ≤ your maintenance threshold = 100% incentive • Log your diet only = 50% incentive • Keep weight ≤ threshold only = 50% incentive • Threshold weight = 6-month weight + 3 lbs (may be adjusted by dietitian) For diet logging incentives: • Log food at least 3 days/week, with at least 1 weekend day • Log at least 1000 calories/day (female at birth) or 1200 calories/day (male at birth) • Food logs checked nightly before midnight (48-hour look-back, Mondays = 24 hours) For weight maintenance incentives • Platform executes an API import of threshold weights from REDCap daily. • Current weekly weight must be ≤ threshold • Weigh in at least 1 time/week • If >1 weight collected during week, current weekly weight = average of first accepted weight each day. • First daily weight = lowest reading within a 5-minute window • Week = Monday midnight to Sunday 11 PM
Arm 2: Diet Logging Incentive • 100% incentive for meeting the diet logging rules above	**Arm 2: Diet Logging Incentive** • 100% incentive for meeting the diet logging rules above
Arm 3: Weight-Based Incentive • 100% incentive for meeting the weight loss rules above	**Arm 3: Weight Maintenance-Based Incentive** • 100% incentive for meeting the weight maintenance rules above
Arm 4: No Incentive	**Arm 4: No Incentive**

XSL•FO
RenderX

Table 1. Platform and technical troubleshooting examples.

Issue type	Example	Potential problems	Solutions
Text message delivery	Monitoring for delivery failures	• Spam flagging • Participant opted out • Phone off/out of service • Landline number given • Daylight savings errors	• Automate resends • Investigate delivery failures • Confirm participant preferences
Weight data	Missing or inaccurate data	• Incorrect scale ID entry • Scale out of carrier range • Low battery • Wrong scale used • Scale not tared correctly • Weighing fully clothed • Scale not on a level surface • Household member stepping on scale	• Verify scale ID • Verify participant stepping on scale • Support scale functionality and placement • Ensure battery power
Dietary data	Logged food does not appear in the platform	• Outdated device operating system • Old Fitbit app version • Multiple Fitbit accounts • Enterprise Google email on device • Missed authorizations • Technical bugs with Fitbit app • Accidental revocation of authorization • Data transmission issues	• Update device and Fitbit app • Ensure single account use • Check authorization and data transmission • Manually sync foods • Reauthorize and send new code
Dietary application API[a]	MyFitnessPal dietary data not retrievable	• API connection cut by MyFitnessPal	• Shift to Fitbit dietary app
Incentives	Participant reports nonreceipt	• Scale misuse • Weight discrepancies • Scale shared with others • Scale out of carrier range • Weighing outside incentive period • Time zone issues • Insufficient weigh-ins • Multiple weigh-ins on the same day • Late or insufficient food logging	• Educate on scale use • Verify and reset weights • Confirm food logging
Activity tracker data	Data not received	• Fitbit not syncing to phone due to firmware issues or Bluetooth signal issues (Bluetooth off, devices too far away to sync, interference from other sources) • Technical bugs with Fitbit app • Accidental revocation of authorization • Data transmission issues	• Manual sync • Update firmware • Restart Fitbit • Restart Bluetooth • Increase Bluetooth signal (device proximity/remove interference from other Bluetooth/Wi-Fi signal sources) • Reauthorize and send new code

[a]API: application programming interface.

Discussion

Overview

Novel strategies are needed to support both the initiation and maintenance of weight loss. Because weight management requires frequent, repeated behaviors, technologies that minimize barriers to tracking may be particularly effective. Mobile technologies allow for continuous monitoring of progress, and with appropriate technical support (Table 1), Log2Lose demonstrates that near real-time, large-scale data collection, and feedback are feasible across time zones and diverse settings. Our experience suggests that automation is scalable when paired with pragmatic, responsive support.

Our findings contribute to the broader literature on incentive-based weight management interventions. Prior studies have shown that financial incentives can support initial weight loss but often face challenges in sustaining long-term adherence [5-7]. More recent work such as Ladapo et al [4] have demonstrated the potential of both outcome-based and behavior-based incentives; yet, questions remain regarding how best to deliver and sustain these approaches in real-world contexts. Log2Lose extends this body of work by demonstrating that incentives can be operationalized at scale through a fully remote, multisite trial supported by consumer technologies. Unlike earlier studies that relied on in-person contact or static incentive structures, our trial highlights the importance of adaptability in the face of shifting technology, regulation, and participant needs. In this way, Log2Lose not only reinforces

XSL•FO
RenderX

prior evidence that incentives can be effective but also advances the field by offering a model for integrating automation, flexibility, and responsive technical support into long-term digital health interventions.

Evolution of Technology

The COVID-19 pandemic necessitated a swift pivot to fully remote delivery. Group weight loss sessions were moved to videoconference, nearly all technical support was provided remotely, and primary outcome data were collected with cellular scales instead of during in-person visits. Preparation during our pilot phase positioned us to navigate an increasingly complex technology and regulatory environment during and after the pandemic [11,20].

To avoid reliance on a single application, the platform was designed to be app-agnostic. This flexibility proved essential when the MyFitnessPal–Fitbit API link ended mid-trial. We migrated dietary logging entirely to Fitbit and routed integrations directly through Log2Lose rather than third-party-to-third-party connections. Although this migration required substantial engineering effort, it improved data integrity and reduced opaque synchronization issues. Later vendor changes such as Fitbit's consolidation after Google's acquisition necessitated further protocol revisions and staffing adjustments.

The Federal Communications Commission's introduction of new text messaging regulations added another layer of complexity. Maintaining message delivery required continual problem identification and resolution, often resembling a game of whack-a-mole, especially for issues that threatened intervention integrity. Some messages, particularly those with repetitive content or dollar amounts, were flagged and blocked by telecom carriers. Weekly incentive messages were not personalized with participant-specific information such as names or pronouns, which may have increased the likelihood of being flagged. Although our platform attempts to resend failed messages automatically, evolving security protocols necessitated manual weekly surveillance. The research team reviewed failures by carrier and content, and small adjustments such as removing numerals often restored successful delivery in subsequent weeks. In today's rapidly evolving landscape of technology, regulations, and proprietary ownership, adjustments are to be expected in multiyear studies. Building flexibility into planning is therefore essential.

Automation and Technical Support

Automation processed approximately 99% (126082/126283) of the data and incentives on a fixed weekly schedule, reducing errors and preserving fidelity at scale. Automated systems deliver interventions consistently and without bias, but skilled technical support remains indispensable. Our experience underscores that adequate staffing and technical expertise are critical to the success of large-scale digital health programs.

The platform's monitoring tools enabled staff to address many technical problems proactively without contacting participants. However, when automated processes failed, whether for email alerts, class reminders, surveys, text messages, or weight data collection, immediate intervention was required. With over 700 participants, resolving issues demanded substantial time and resources. For instance, a scale issue we identified was that placement near a bathroom wall with baseboard molding would create a fulcrum effect, causing inaccurate weight measurements. Daylight savings time caused problems with text message delivery that we did not plan for. In each case, staff investigated root causes through messaging API error codes and implemented targeted solutions to prevent recurrence.

Limitations

Although more than 85% of US adults own a smartphone [21], generalizability is limited to those with smartphones and data plans and the digital literacy to use the devices, apps, and videoconferencing technology. Message delivery receipts were recorded, but message opens were not tracked because text messaging lacks reliable open tracking functionality. Email open rates were not implemented to protect participant privacy. Data inaccuracies could occur when other household members used the scale, participants underreported dietary intake, or technology literacy was limited. Although such inaccuracies increase variability in efficacy estimates, they are inherent to real-world settings, generalizable to similar interventions, and likely to bias results toward reduced effect sizes rather than inflate estimated effects.

Conclusions

Log2Lose demonstrated the feasibility of a scalable, digitally enabled weight management program. Success depended on adaptable software built on consumer technology, coupled with continuous monitoring and responsive support. By leveraging widely available consumer technologies, the platform adapted to regulatory and technological changes while maintaining intervention fidelity. Although automation allows for broad reach and efficiency, dedicated staffing for technical support is essential. This program model can inform health care providers, payers, and employers implementing obesity interventions, supporting a healthier workforce and potentially reducing health care costs.

Acknowledgments

We thank the participants in the Log2Lose trial. This study was supported entirely by grants from the US National Institutes of Health, National Heart, Lung and Blood Institute (5U24HL150227, 1UG3HL150558, 5UH3HL150558). CV was also supported by a Research Career Scientist award from the Department of Veterans Affairs Health Services Research & Development service (RCS 14-443).

XSL•FO
RenderX

Data Availability

Following completion of the Log2Lose trial, a dataset will be publicly available in the Biological Specimen and Data Repository Information Coordinating Center (BioLINCC) [22].

Authors' Contributions

Conceptualization: RJS (lead), CV (equal), HM (supporting), CB (supporting)
Data curation: CD (lead), MO (equal), SMO (equal)
Formal analysis: CD (lead), MO (supporting), SMO (supporting)
Funding acquisition: CV (lead), RJS (equal), CG (supporting), JP (supporting)
Investigation: SJ (supporting), MB (supporting), CB (supporting)
Methodology: CV (lead), RJS (equal), JP (supporting)
Project administration: RJS (lead), CV (equal), EF (supporting)
Resources: CV (lead), JG (supporting), SJ (supporting), CD (supporting), MO (supporting), SMO (supporting)
Software: JC (lead), CG (supporting)
Supervision: CV (lead), RJS (equal), JG (equal), MO (equal)
Validation: CD (lead), MO (supporting), SMO (supporting)
Visualization: RJS (lead), CD (supporting), CB (supporting)
Writing – original draft: RJS (lead), HM (supporting), CB (supporting), CD (supporting), SMO (supporting), SJ (supporting), MB (supporting), JC (supporting), CG (supporting), JP (supporting), MO (supporting), JG (supporting), CV (supporting)
Writing – review and editing: RJS (lead), HM (supporting), CB (supporting), CD (supporting), SMO (supporting), SJ (supporting), MB (supporting), JP (supporting), MO (supporting), JG (supporting), CV (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Log2Lose platform population-level dashboards showing (A) messages transmitted in a day with delivery rate and errors; (B) participant information by study site and cohort, connection status, and scale signal and battery; (C) calorie entries, weights, incentives awarded, and messages transmitted; and (D) downloadable reports for study management and analyses.
[PNG File , 93 KB - formative_v9i1e70842_app1.png]

Multimedia Appendix 2
Log2Lose platform individual participant dashboard example showing calories tracked, weights received, and incentive and text message history.
[PNG File , 141 KB - formative_v9i1e70842_app2.png]

References

1. Fryar CD, Carroll M, Afful J. Prevalence of overweight, obesity, and severe obesity among adults aged 20 and over: United States, 1960–1962 through 2017–2018. NCHS. URL: https://www.cdc.gov/nchs/data/hestat/obesity-adult-17-18/overweight-obesity-adults-H.pdf [accessed 2025-10-06]
2. Dai H, Alsalhe TA, Chalghaf N, Riccò M, Bragazzi NL, Wu J. The global burden of disease attributable to high body mass index in 195 countries and territories, 1990-2017: an analysis of the Global Burden of Disease Study. PLoS Med 2020 Jul;17(7):e1003198 [FREE Full text] [doi: 10.1371/journal.pmed.1003198] [Medline: 32722671]
3. Dombrowski SU, Knittle K, Avenell A, Araújo-Soares V, Sniehotta FF. Long term maintenance of weight loss with non-surgical interventions in obese adults: systematic review and meta-analyses of randomised controlled trials. BMJ 2014 May 14;348:g2646 [FREE Full text] [doi: 10.1136/bmj.g2646] [Medline: 25134100]
4. Ladapo JA, Orstad SL, Wali S, Wylie-Rosett J, Tseng C, Chung UYR, et al. Effectiveness of goal-directed and outcome-based financial incentives for weight loss in primary care patients with obesity living in socioeconomically disadvantaged neighborhoods: a randomized clinical trial. JAMA Intern Med 2023 Jan 01;183(1):61-69 [FREE Full text] [doi: 10.1001/jamainternmed.2022.5618] [Medline: 36469353]
5. Leahey TM, Subak LL, Fava J, Schembri M, Thomas G, Xu X, et al. Benefits of adding small financial incentives or optional group meetings to a web-based statewide obesity initiative. Obesity (Silver Spring) 2015 Jan;23(1):70-76 [FREE Full text] [doi: 10.1002/oby.20937] [Medline: 25384463]
6. Volpp KG, John LK, Troxel AB, Norton L, Fassbender J, Loewenstein G. Financial incentive-based approaches for weight loss: a randomized trial. JAMA 2008 Dec 10;300(22):2631-2637 [FREE Full text] [doi: 10.1001/jama.2008.804] [Medline: 19066383]

XSL•FO
RenderX

7. Kullgren JT, Troxel AB, Loewenstein G, Asch DA, Norton LA, Wesby L, et al. Individual- versus group-based financial incentives for weight loss: a randomized, controlled trial. Ann Intern Med 2013 Apr 02;158(7):505-514 [FREE Full text] [doi: 10.7326/0003-4819-158-7-201304020-00002] [Medline: 23546562]
8. Almeida FA, You W, Harden SM, Blackman KC, Davy BM, Glasgow RE, et al. Effectiveness of a worksite-based weight loss randomized controlled trial: the worksite study. Obesity (Silver Spring) 2015 Apr;23(4):737-745 [FREE Full text] [doi: 10.1002/oby.20899] [Medline: 25678325]
9. Shaw PA, Yancy WS, Wesby L, Ulrich V, Troxel AB, Huffman D, et al. The design and conduct of Keep It Off: an online randomized trial of financial incentives for weight-loss maintenance. Clin Trials 2017 Mar;14(1):29-36 [FREE Full text] [doi: 10.1177/1740774516669679] [Medline: 27646508]
10. Yawn B, Kurland M, Madison S, Bertram S, Littlefield D, Wechsler M, et al. Automated patient and medication payment method for clinical trials. OAJCT 2013 Jan:23. [doi: 10.2147/oajct.s38489]
11. Voils C, Pendergast J, Hale S, Gierisch JM, Strawbridge EM, Levine E, et al. A randomized feasibility pilot trial of a financial incentives intervention for dietary self-monitoring and weight loss in adults with obesity. Transl Behav Med 2021 Apr 26;11(4):954-969 [FREE Full text] [doi: 10.1093/tbm/ibaa102] [Medline: 33245118]
12. Foster GD, Wyatt HR, Hill JO, Makris AP, Rosenbaum DL, Brill C, et al. Weight and metabolic outcomes after 2 years on a low-carbohydrate versus low-fat diet: a randomized trial. Ann Intern Med 2010 Aug 03;153(3):147-157 [FREE Full text] [doi: 10.7326/0003-4819-153-3-201008030-00005] [Medline: 20679559]
13. Kurti AN, Davis D, Redner R, Jarvis B, Zvorsky I, Keith DR, et al. A review of the literature on remote monitoring technology in incentive-based interventions for health-related behavior change. Transl Issues Psychol Sci 2016 Jun;2(2):128-152 [FREE Full text] [doi: 10.1037/tps0000067] [Medline: 27777964]
14. FCC adopts its first rules focused on scam texting. Federal Communications Commission. URL: https://www.fcc.gov/document/fcc-adopts-its-first-rules-focused-scam-texting-0 [accessed 2025-10-02]
15. Forbidden message categories in the US and Canada (short code, toll-free, and long code). Twilio. URL: https://tinyurl.com/5hf7f334 [accessed 2025-10-02]
16. Voils CI, Levine E, Gierisch JM, Pendergast J, Hale SL, McVay MA, et al. Study protocol for Log2Lose: a feasibility randomized controlled trial to evaluate financial incentives for dietary self-monitoring and interim weight loss in adults with obesity. Contemp Clin Trials 2018 Mar;65:116-122 [FREE Full text] [doi: 10.1016/j.cct.2017.12.007] [Medline: 29289702]
17. Gavin KL, Almeida EJ, Voils CI, Crane MM, Shaw R, Yancy WS, et al. Comparison of weight captured via electronic health record and cellular scales to the gold-standard clinical method. Obes Sci Pract 2023 Aug;9(4):337-345 [FREE Full text] [doi: 10.1002/osp4.656] [Medline: 37546286]
18. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381 [FREE Full text] [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
19. Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O'Neal L, REDCap Consortium. The REDCap consortium: building an international community of software platform partners. J Biomed Inform 2019 Jul;95:103208 [FREE Full text] [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]
20. Shaw R, Levine E, Streicher M, Strawbridge E, Gierisch J, Pendergast J, et al. Log2Lose: development and lessons learned from a mobile technology weight loss intervention. JMIR Mhealth Uhealth 2019 Mar 13;7(2):e11972 [FREE Full text] [doi: 10.2196/11972] [Medline: 30758297]
21. Internet/broadband fact sheet. Pew Research Center. URL: https://www.pewresearch.org/internet/fact-sheet/internet-broadband/ [accessed 2025-10-02]
22. Biologic specimen and data repository information coordinating center. NIH. URL: https://biolincc.nhlbi.nih.gov/home/ [accessed 2025-10-06]

Abbreviations

API: application programming interface
REDCap: Research Electronic Data Capture

XSL•FO
RenderX

Edited by A Stone, A Mavragani; submitted 03.01.25; peer-reviewed by PH Lin, C Wang, W Demark-Wahnefried; comments to author 12.08.25; revised version received 25.08.25; accepted 26.09.25; published 17.11.25.

Please cite as:
Shaw RJ, Miller H, Barnes C, Davenport C, Morton-Oswald S, Jackson S, Bean M, Cohen J, Griffin C, Pendergast J, Olsen M, Gierisch J, Voils C
Development, Challenges, and Evolution of the Log2Lose Intervention for Weight Management: Randomized Controlled Digital Health Trial
JMIR Form Res 2025;9:e70842
URL: https://formative.jmir.org/2025/1/e70842
doi:10.2196/70842
PMID:

©Ryan Jeffrey Shaw, Hailey Miller, Cherie Barnes, Clemontina Davenport, Sarah Morton-Oswald, Sarah Jackson, Michelle Bean, Jeff Cohen, Chance Griffin, Jane Pendergast, Maren Olsen, Jennifer Gierisch, Corrine Voils. Originally published in JMIR Formative Research (https://formative.jmir.org), 17.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

User-Centered Formative Evaluation of Cognitive Rehabilitation Software: Cognitive Walkthrough and System Usability Scale Study

Seojin Hong[1], MS; Hyun Choi[1], PhD; Hyosun Kweon[1], PhD

National Rehabilitation Center, Seoul, Gangbuk-gu, Republic of Korea

Corresponding Author:
Hyosun Kweon, PhD
National Rehabilitation Center
58 Samgaksan-ro
Seoul, Gangbuk-gu, 01022
Republic of Korea
Phone: 1 02 901 1966
Email: hskweon@korea.kr

Abstract

Background: The usability of software as a medical device plays a critical role in ensuring patient safety, clinical effectiveness, and treatment adherence. Particularly in tablet-based cognitive rehabilitation software, poor usability can increase the risk of user errors, cognitive overload, and low engagement, which may diminish therapeutic outcomes. Despite the growing integration of cognitive rehabilitation software in clinical settings, few studies have systematically evaluated its usability using structured human factor methodologies.

Objective: This study aimed to conduct a formative evaluation of a tablet-based cognitive rehabilitation software by identifying potential user errors, interface design issues, and opportunities for improving the user experience and user interface (UI).

Methods: Following the International Electrotechnical Commission (IEC) 62366-1 framework for usability engineering, a formative evaluation was conducted using a combination of cognitive walkthrough (CW) and the System Usability Scale (SUS). CW was used to identify interaction breakdowns and potential use-related problems through scenario-based expert evaluation. The SUS was administered to quantitatively assess perceived usability. Five occupational therapists with experience in cognitive rehabilitation participated in the formative evaluation.

Results: A total of 14 usability issues were identified across 6 task scenarios. Common problems included difficulty locating buttons, insufficient feedback during task completion, and inconsistent UI behavior. Key improvement areas included keypad layout, visibility of task progress, and accessibility of help content. The mean SUS score was 73.5 (SD 11.4), indicating an acceptable usability level (grade of B–). The lowest scores were related to system complexity and interface integration, suggesting a need to optimize UI flow and content hierarchy.

Conclusions: The combined use of CW and the SUS enabled the identification of both objective and perceived usability challenges in the cognitive rehabilitation software. Findings emphasize the importance of user-centered design and formative evaluation in the early stages of software development. Recommendations such as improving information visibility, reducing unnecessary steps, and enhancing system feedback may contribute to safer and more effective digital cognitive interventions. However, given the small and homogeneous participant sample and the controlled test environment, these findings should be interpreted as preliminary. This study highlights the practical application of the IEC 62366-1 framework in evaluating software as a medical device and demonstrates its utility in guiding iterative user experience and UI improvements in cognitive rehabilitation contexts.

(JMIR Form Res 2025;9:e75805) doi:10.2196/75805

KEYWORDS

usability evaluation; formative evaluation; IEC 62366-1; cognitive walkthrough; System Usability Scale; cognitive rehabilitation software; human factors

XSL•FO
RenderX

Introduction

With advancements in digital technology, the adoption of computer-based cognitive training is expanding in the fields of medicine and rehabilitation [1,2]. Specifically, tablet- and computer-based cognitive rehabilitation software is being used as an effective rehabilitation tool due to its high accessibility and intuitive touch screen interface [3,4]. The primary aim of cognitive rehabilitation software is to enhance patients' cognitive functions through iterative training. However, simplistic program designs alone do not guarantee clinical efficacy. For the effective utility of such software, evaluating the user interface (UI) and user experience (UX) design and usability—factors that shape the experiences of its intended users, such as occupational therapists—is crucial. This usability evaluation is strongly linked to treatment adherence and clinical efficacy [5].

Usability evaluation of medical devices and digital rehabilitation software is critical for reducing therapists' workload, increasing treatment engagement, and ensuring the clinical reliability of the software [6]. In the context of cognitive rehabilitation software, poor usability may impose unnecessary cognitive demands on occupational therapists—such as navigating convoluted menus or locating key functions—which can slow down therapy sessions and distract from patient engagement. For instance, Levac and Miller [7] reported that even experienced therapists encountered usability barriers during a tablet-based rehabilitation tool study, identifying numerous navigational failures that disrupted session flow and required repeated assistance during task completion. Moreover, in analogous rehabilitation contexts—such as robotic exoskeleton use—studies have demonstrated that higher usability and lower cognitive workload are associated with greater therapeutic benefit and user efficiency [8]. These findings suggest that usability deficits in cognitive rehabilitation tools can directly degrade work efficiency and potentially undermine clinical and patient outcomes in real-world settings [9].

In particular, for software functioning as a medical device, it is critical to ensure that users can interact with the system correctly and efficiently in real-world clinical contexts. To address this, the International Electrotechnical Commission (IEC) 62366-1 standard outlines a structured usability engineering process for medical devices, emphasizing the prevention of use-related errors and the enhancement of user-system interaction safety [10]. This standard advocates for the incorporation of formative evaluation during early design stages to identify and rectify potential design flaws before final implementation. Formative evaluations, as defined by the standard, aim to detect use errors and interface issues that could compromise task performance, especially in cognitively demanding contexts. In pediatric cognitive rehabilitation, where users' cognitive abilities and task comprehension may vary widely, adopting such structured usability engineering is essential to ensure both therapeutic effectiveness and user safety. Despite this, few empirical studies have applied the IEC 62366-1 framework specifically to the evaluation of cognitive rehabilitation software, indicating a significant gap in the empirical evidence supporting standard-compliant usability practices in this domain.

Among formative evaluation methods, the cognitive walkthrough (CW) is an effective evaluation method for analyzing design errors that may occur during system navigation, and it also serves as a common methodology for evaluating the intuitiveness of UX and UI design during the inception of medical device development [11]. The System Usability Scale (SUS) is a quantitative tool used to evaluate the overall usability of medical software and is also useful for objectively measuring UX [12].

Previous studies have primarily focused on evaluating the clinical efficacy and treatment outcomes of computerized cognitive rehabilitation software [13,14], with few studies assessing usability to improve UI and UX design. The intuitiveness and ease of the UI and UX in medical device software directly affect clinical applicability and treatment adherence, whereas poorly designed interfaces increase user errors, reducing treatment effectiveness.

Therefore, this study aimed to evaluate the usability of a tablet-based cognitive rehabilitation software by identifying use-related problems and UI and UX design issues. Using CW and the SUS, we assessed the experiences of occupational therapists, the primary users of this software, and proposed design improvement strategies aimed at enhancing the intuitiveness, accessibility, and overall clinical applicability of cognitive rehabilitation software.

Methods

Study Design

This study was conducted as a formative evaluation of a prototype tablet-based cognitive rehabilitation software for children with cognitive impairment. The CW and the SUS were used with occupational therapists to identify use-related problems and gather UX insights to inform subsequent interface improvements.

Cognitive Rehabilitation Software

The cognitive rehabilitation software used in this study was developed for use in occupational therapy to enhance cognitive function, learning, and development among patients with cognitive impairments. This product comprises 12 subprograms, including 7 structured programs arranged by difficulty level (puzzles, finding hidden objects, finding animals, finding the same pictures, finding rules, memorizing, and guessing the sound) and 5 nonstructured programs (daily life play, drawing along, balloon play, farm play, and guessing objects). In addition, a cognitive screening evaluation, result checking, use guidelines, and instructions are available. Specifically, the software assesses patients' cognitive status through an initial screening evaluation and subsequently recommends individualized cognitive therapy programs based on the obtained results. Occupational therapists design and monitor personalized rehabilitation treatment sessions, and analytical outcomes can be reviewed upon completion of each session.

The main UI consisted of a home dashboard displaying program categories, patient management tools, and progress-tracking features. Each rehabilitation module featured consistent navigation patterns, including start and end buttons, difficulty

XSL•FO
RenderX

selection, and progress indicators. Due to intellectual property and confidentiality restrictions requested by the manufacturer, the detailed UI design and proprietary elements of the software could not be disclosed. However, the evaluation was conducted using the actual commercial prototype provided by the manufacturer.

The recommended software specifications include an Android 13–based operating system equipped with either a Samsung Exynos 1380 or Qualcomm Snapdragon 865 processor. In this study, the Samsung Galaxy Tab S6 Lite was used.

Participants

The formative evaluation was conducted in accordance with the IEC 62366-1 standard, which emphasizes formative evaluation during the early stages of medical device development. This approach typically involves a limited number of representative users or domain experts to iteratively identify potential use-related problems and interface design issues. According to the IEC 62366-1 and previous usability research [11,15], a sample of 3 to 8 participants is generally sufficient to detect most usability problems during formative testing. Accordingly, 5 occupational therapists affiliated with the Korea National Rehabilitation Center (NRC)—the primary users of cognitive rehabilitation software—participated in this evaluation. While appropriate for formative evaluation, this narrow sample from a single institution excludes other relevant user groups such as patients and caregivers, thereby limiting the representativeness of the findings. Nonetheless, their participation enabled in-depth identification of task flow difficulties, interface-related usability issues, and UX improvement needs. Expert-based formative evaluation at this stage is particularly important in the context of medical software as it supports early detection of design flaws that could lead to clinical use errors, thereby contributing to risk mitigation and overall safety.

The formative evaluation participants (N=5) were all licensed occupational therapists with previous experience using medical devices for cognitive rehabilitation. Inclusion criteria required sufficient proficiency in both Korean and English to ensure comprehension of the software's terminology and interface instructions. Individuals without previous experience in cognitive rehabilitation therapy were excluded. All participants voluntarily provided written informed consent after receiving a detailed explanation of the study's objectives and procedures and their rights as participants.

The evaluators included 3 usability researchers from the NRC: 1 facilitator and 2 observers. The facilitator was responsible for the overall execution of the evaluation, including orientation, task instruction delivery, informed consent collection, and administration of pre-evaluation surveys and postevaluation interviews. The observers were tasked with monitoring and documenting each participant's task performance and behavioral responses throughout the formative evaluations.

Formative Evaluation Protocol

This study conducted a formative evaluation following the UI validation procedures outlined in the IEC 62366-1 standard, which emphasizes the early identification and mitigation of use-related risks in medical device software. The evaluation was carried out at the NRC using a combination of CW and the SUS.

CW is a scenario-based expert method that examines user interaction flows to detect potential usability issues within the UI. It is particularly well suited for evaluating systems with high cognitive demands, such as pediatric cognitive rehabilitation software. The SUS, a standardized 10-item questionnaire, was used to quantitatively assess users' subjective perceptions of usability. The integration of CW and the SUS enabled a comprehensive evaluation of both objective interaction challenges and subjective UX during early-stage development.

Each evaluation session lasted approximately 50 minutes per participant (Table 1). The facilitator provided participants with an overview of the study objectives and procedures, anonymized product information, and usability testing process. General demographic characteristics and previous experience with cognitive rehabilitation devices were collected, after which participants performed tasks based on predefined software scenarios. Throughout the evaluation, an observer recorded task performance and behavioral responses. Upon completion of the tasks, participants completed the SUS questionnaire to assess the overall usability of the software (Figure 1).

Figure 1. Performance of task scenarios by evaluation participants.

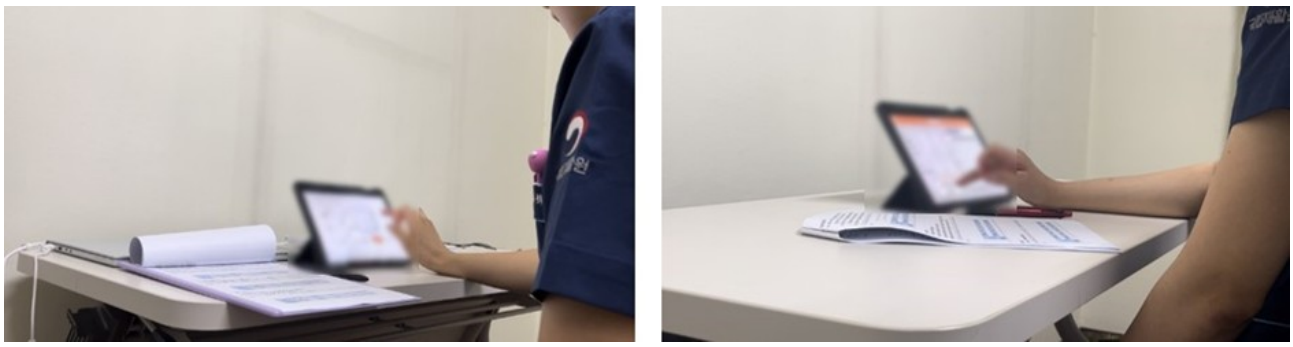

XSL•FO
RenderX

Table 1. Formative evaluation procedures conducted for 50 minutes under the guidance of a facilitator.

Stage	Details	Time (min)
Orientation	• The facilitator introduces the concept of formative evaluation to participants who are unfamiliar with it. • The facilitator explains the target product, purpose, overview, methodology, and other relevant aspects of the formative evaluation to the participant.	5
Guidance on consent	• The facilitator provides the participant with detailed instructions regarding recording and transcription procedures. • The facilitator ensures that the participant fully comprehends the evaluation content and voluntarily provides written consent to take part in the formative evaluation.	5
Cognitive walkthrough	• The facilitator conducts a cognitive walkthrough following a predefined script and product use instructions. • The participant operates the medical device under the facilitator's guidance, verbalizes their thoughts during task performance, and provides feedback on their experiences. • The observer documents and analyzes the participant's responses during task performance.	30
Administration of the SUS[a]	• After completing the cognitive walkthrough, the facilitator administers the SUS to the participant to assess the usability of the user interface.	10

[a]SUS: System Usability Scale.

Before conducting the formative evaluation, researchers reviewed the software's intended use, functional specifications, and workflow documentation included in the usability engineering file provided by the manufacturer. This documentation described the core business processes of the software—such as patient registration, cognitive assessment, training program delivery, and result analysis. The task scenarios were explicitly mapped to these clinical workflows to ensure that the evaluation reflected real-world therapist operations and covered the complete use process. All task scenarios were developed based on the scenario construction procedures outlined in the IEC 62366-1 standard and the usability engineering file. Although no formal pilot test was conducted with intended users, the sequence of actions and task objectives was validated through internal simulation by usability researchers to ensure logical consistency, alignment with the intended use, and appropriateness of user-interface interaction flows. This approach, consistent with formative evaluation standards, was considered sufficient for verifying scenario validity during the early design stage.

The formative evaluation environment simulated a real-world occupational therapy room. To maintain environmental consistency, the test room was premeasured and adjusted to specific parameters: illumination at 316 lx, ambient noise at 28 dBA, room temperature at 22.8 °C, and relative humidity at 67%.

CW Methodology

CW is a commonly used quantitative evaluation method to assess the cognitive load of users during system navigation and task performance [11,16]. To evaluate whether the participants could effectively navigate the software, the facilitator provided the task scenarios without previous training. The evaluation participants navigated the software according to the task scenarios provided and performed the major tasks. The observer and recorder documented UI and UX issues and user errors encountered during the evaluation process.

CW was used to analyze major usability issues that might arise as users navigated the software. For this, the evaluation involved a total of 8 major tasks divided into 17 subtasks, including starting the software, performing a cognitive screening evaluation, setting patient accounts, and selecting and running rehabilitation content. The eight evaluation tasks were (1) checking the power, (2) registering patient information, (3) patient evaluation, (4) checking the user manual, (5) running the program 1, (6) running the program 2, (7) analyzing the results, and (8) exiting the program (Table 2).

XSL•FO
RenderX

Table 2. Task scenarios.

Task	Subtask
Checking the power (task 1)	• The therapist turns the power on in the tablet screen and runs the cognitive rehabilitation software (subtask 1.1).
Patient information registration (task 2)	• The therapist generates patient information (subtask 2.1). • The therapist selects name for the patient (subtask 2.2).
Patient evaluation (task 3)	• The therapist performs a cognitive development screening evaluation on the patient (subtask 3.1). • The therapist ends the cognitive development screening evaluation (subtask 3.2).
Checking the user manual (task 4)	• The therapist checks the guidelines and explanations for the software programs on finding hidden objects and memory (subtask 4.1).
Running the program 1 (task 5)	• The therapist sets the settings for the program on finding hidden objects and applies the program on the patient (subtask 5.1). • The therapist adjusts the difficulty level of the program on finding hidden objects and applies the program on the patient (subtask 5.2). • The therapist provides hints to the patient (select the button within the program; subtask 5.3). • The therapist ends the program on finding hidden objects and returns to the main screen (subtask 5.4).
Running the program 2 (task 6)	• The therapist sets the settings for the memory program and applies the program on the patient (subtask 6.1). • The therapist ends the memory program and returns to the main screen (subtask 6.2).
Analysis of results (task 7)	• The therapist checks the user information (subtask 7.1). • Check the results of “Evaluation participation date” (subtask 7.2). • The therapist checks the specific results for each date (programs applied, levels performed, number of items, correct answer rate, response time, and total performance time; subtask 7.3). • The therapist checks the comprehensive analysis (performance, memory, concentration, hand-eye coordination, language skill, and visual perception; subtask 7.4).
Exit (task 8)	• Exit the cognitive rehabilitation software (subtask 8.1).

The observer monitored the evaluation participants as they performed the subtasks, documenting and analyzing the following: (1) whether the user knew what they needed to do for each task (response options: yes or no), (2) whether the user performed and completed the task properly (response options: yes or no), and (3) task completion status (response options: yes or no) [17,18]. In addition, the reactions and verbal opinions of the participants were recorded.

SUS Questionnaire

The SUS is a quantitative evaluation tool used to assess how users perceive system usability [12]. The SUS comprises 10 items rated on a Likert scale (1=“strongly disagree”; 5=“strongly agree”; Table 3). The total score is converted to a range from 0 to 100 points to evaluate the overall system usability. Moreover, the SUS scores are assigned a letter grade from F (0-60 points) to A (91-100 points) according to a standardized grading system [19,20].

Table 3. System Usability Scale questionnaire item breakdown.

Item category	Statement
Utility	“I think that I would like to use this system frequently.”
Complexity	“I found the system unnecessarily complex.”
Simplicity	“I thought the system was easy to use.”
Professionalism (technician support)	“I think that I would need the support of a technical person to be able to use this system.”
Integration	“I found the various functions in the system were well integrated.”
Unity	“I thought there was too much inconsistency in this system.”
Learnability	“I would imagine that most people would learn to use this system very quickly.”
Convenience	“I found the system very cumbersome to use.”
Satisfaction	“I felt very confident using the system.”
Professionalism (previous learning)	“I needed to learn a lot of things before I could get going with this system.”

XSL•FO
RenderX

Data Collection and Analysis

CW was used to analyze major usability issues that might occur during software navigation. User reactions and opinions, comprising qualitative data collected through CW, were transcribed, managed, and coded using Microsoft Excel based on the observer and recorder records. SUS evaluation was used to quantitatively assess the overall software UX. The mean and SD of the SUS scores were calculated and interpreted as letter grades (A to F) according to the grading system [19].

Ethical Considerations

This study was approved by the institutional review board of the NRC (NRC-2024-04-020). All participants provided voluntary consent to take part after receiving a comprehensive explanation of the study objectives and methods. No compensation was provided for participating in the study. For the protection of personal information and the confidentiality of the participants, all data were anonymized.

Results

Evaluation of Participant Information

The evaluation participants comprised 5 occupational therapists with 3 years and 2 months to 15 years and 4 months of experience. Furthermore, the therapists had 2 to 12 years of experience using similar medical devices or computerized cognitive programs, such as ComCog (Neofect), RAPAEL (Neofect), and RehaCom (HASOMED GmbH; Table 4).

Table 4. General characteristics of the participants.

ID	Sex	Experience (y)	Experience with similar medical devices			
			Used them before	Model name	Experience of use (y)	Frequency of use
P1	Female	≥6	Yes	RAPAEL	2	2-3 times per mo
P2	Female	≥3	Yes	ComCog and RAPAEL	3	1-2 times per wk
P3	Male	12	Yes	ComCog and RehaCom	12	2-3 times per d
P4	Female	≥4	Yes	ComCog and RAPAEL	4	3 times per wk
P5	Male	≥15	Yes	ComCog, RAPAEL, and RehaCom	10	1 time per wk

CW Results

During the CW, which involved observing task performance by the participants based on the task scenarios, frequent typing errors were noted during patient registration (task 2) and patient information generation (subtask 2.1) due to the small size of the on-screen keypad. Participants also experienced inconveniences when checking certain information, as the keypad obscured a section of the patient information input. Furthermore, they experienced difficulty finding the delete button and delays when looking for the button to change to English to input information in English.

To perform the cognitive development screening evaluation (subtask 3.1) within the patient evaluation task (task 3), some participants appeared confused because the evaluation began without prior instruction or explanation. In particular, although the puzzle task required a tapping gesture, several participants attempted to use a dragging gesture, likely because of their previous experience with similar interfaces. This reflected a mismatch between the user's mental model and the intended interaction design. To complete the cognitive development screening evaluation (subtask 3.2), participants were required to press the "Next Item" button as a separate "End" button was not provided. However, evaluation participants mistakenly pressed the "Back" button to end the evaluation, which led to the loss of evaluation results. This issue made it impossible to check the cognitive development results, impacting the running of the program.

To check the user manual (task 4), most users appeared not to recognize the "What is cognitive rehabilitation software?" icon as the button for the software user manual.

To adjust the difficulty (subtask 5.2) within the task of running the program (task 5), the selection of the program stage appeared on the program screen, whereas the selection of the number of items appeared via a modal pop-up. Users made errors by pressing the "Run" button without recognizing the stage selection screen when selecting the stage and number of items. Regarding giving hints (subtask 5.3), 60% (3/5) of the evaluation participants did not recognize the question mark icon as the hint button and instead provided verbal hints, resulting in errors.

To check the results of the "Evaluation participation date" (subtask 7.2) within the result analysis task (task 7), 40% (2/5) of the participants did not end the cognitive development screening evaluation as indicated in the user manual. Consequently, the evaluation results were not saved, and thus, the list of programs performed and performance results could not be confirmed. Moreover, 40% (2/5) of the participants selected the detailed program results (individual training session outcomes) instead of the screening evaluation performance results (summary results from the initial cognitive development screening). Regarding checking the specific results for each date (subtask 7.3), 40% (2/5) of the participants did not complete the cognitive development screening evaluation. Consequently, specific results for each date could not be found and confirmed. Moreover, 40% (2/5) of the participants could not find the icon for specific results by date, and therefore, they were unable to perform this subtask. To check the performance level of the patient for the cognitive domain (subtask 7.4), the correct icon

XSL•FO
RenderX

was selected, but because the cognitive development screening evaluation results were not saved, the cognitive ability graph was not provided and could not be confirmed.

SUS Results

The mean SUS score was 73.5 (SD 7.2) points, equivalent to a letter grade of B– (good or acceptable; Figure 2). Although usability was evaluated to be generally good, some areas for improvement were identified. Specifically, the scores were relatively low for utility (question 1), complexity (question 2), integration (question 5), unity (question 6), and satisfaction (question 9), indicating that the functional needs of the users were not sufficiently met and the method of use was not intuitive (Table 5).

Figure 2. System Usability Scale (SUS) results.

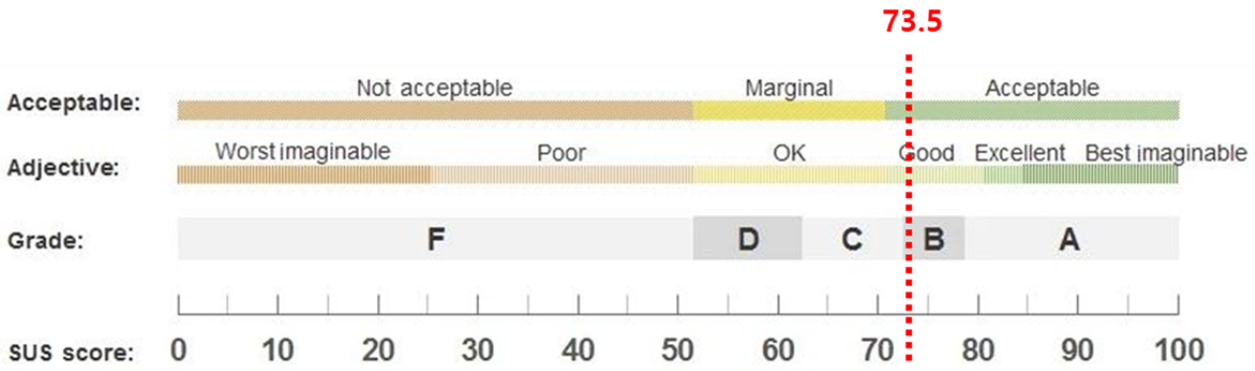

Table 5. System Usability Scale evaluation results (N=5).

Item	Score (1-5), mean (SD; range)	Scaled score (0-100), mean (SD)
Total	3.02 (0.27; N/A[a])	73.50 (11.40)
Utility (question 1)	3.80 (0.45; 3-4)	70.00 (11.18)
Complexity (question 2)	2.40 (0.55; 2-3)	65.00 (13.69)
Simplicity (question 3)	4.00 (0.00; 4-4)	75.00 (0.00)
Professionalism (technician support; question 4)	2.00 (0.71; 1-3)	75.00 (17.68)
Integration (question 5)	3.80 (0.84; 3-5)	70.00 (20.92)
Unity (question 6)	2.40 (0.89; 1-3)	65.00 (22.36)
Learnability (question 7)	4.40 (0.55; 4-5)	85.00 (13.69)
Convenience (question 8)	2.00 (1.00; 1-3)	75.00 (25.00)
Satisfaction (question 9)	3.80 (0.84; 3-5)	70.00 (20.92)
Professionalism (previous learning; question 10)	1.60 (0.55; 1-2)	85.00 (13.69)

[a]N/A: not applicable.

Discussion

Principal Findings

In this study, a formative evaluation was conducted of the usability of a tablet-based cognitive rehabilitation software to identify areas for improvement in the UI and UX. The combined use of CW and the SUS enabled the identification of both interaction-based usability problems and perceived UX among occupational therapists—the primary users of this software.

Key usability issues identified in this study, such as suboptimal keypad design, lack of instructional prompts, nonintuitive task termination mechanisms, and inconsistent iconography, reflect common design failures in medical software interfaces [21,22]. These issues were not merely aesthetic or functional flaws but directly influenced users' ability to complete tasks efficiently and accurately. For example, the absence of clear visual cues for task completion or the ambiguous labeling of interactive elements (eg, the hint button) disrupted workflow continuity, increased cognitive burden, and raised the risk of data loss—factors that undermine both clinical reliability and therapeutic engagement. These findings align with those of previous studies emphasizing that poorly designed interfaces in medical device software can degrade user efficiency and negatively affect patient outcomes [23,24].

From a human factor perspective, many of the observed errors can be interpreted through the lens of cognitive load theory [25]. Inconsistent UI layouts and insufficient feedback mechanisms can impose extraneous cognitive load, diverting attention from therapeutic interaction toward interface management. In pediatric contexts, this can be particularly detrimental as delays

XSL•FO
RenderX

or confusion in therapist-patient interaction may reduce children's motivation and task adherence [24].

The software achieved an SUS score corresponding to a good or acceptable usability grade (B−); however, several dimensions—such as complexity, functional integration, and intuitiveness—scored below the average. These findings suggest a potential misalignment between the system's interface design and the clinical workflows of end users. Consistent with previous research, a high overall SUS score does not necessarily indicate the absence of critical usability issues, as deficiencies in specific subdimensions may still reflect underlying problems related to workflow integration and task efficiency [26,27]. In contrast, higher ratings for learnability and professionalism indicate that, once users became familiar with the interface, they were able to operate the system with relative ease. This discrepancy underscores the importance of incorporating iterative usability testing and applying user-centered design principles throughout the software development life cycle to ensure both functional adequacy and clinical relevance [28,29].

On the basis of previous literature and these findings, several actionable design recommendations are proposed [27,30]: (1) redesign the size and placement of input interfaces to minimize use errors, (2) introduce guided workflows and clearly indicated start and termination controls to enhance procedural clarity and data integrity, (3) improve the visibility and semantic clarity of educational elements such as user manuals and functional icons, and (4) redesign interaction components—such as difficulty adjustment and hint functions—to follow a consistent and intuitive visual language that aligns with user expectations.

These recommendations highlight that formative evaluation should not focus solely on user satisfaction or interface aesthetics but should be systematically integrated with the software's clinical business processes, including patient registration, cognitive assessment, therapy execution, and result review. From a clinical standpoint, improving task flow and feedback mechanisms within cognitive rehabilitation software can directly enhance therapeutic efficiency, reduce therapists' cognitive workload, and increase patient engagement and adherence. Moreover, by aligning formative evaluation with the principles of the IEC 62366-1 standard and the International Organization for Standardization risk management standard 14971, this study demonstrates that usability engineering can serve as a practical framework for manufacturers seeking to design safer and more effective rehabilitation devices. Collectively, these findings emphasize the importance of integrating usability testing not as an isolated activity but as an iterative, evidence-based component of the medical software development process.

Despite its contributions, this study has several limitations. The evaluation involved a small and homogeneous sample of 5 occupational therapists from a single institution, which is appropriate for formative evaluation under the IEC 62366-1 standard but limits generalizability to broader clinical contexts or end users such as patients. Additionally, testing was conducted in a controlled simulation environment rather than during real-world therapy sessions, which may reduce ecological validity. Moreover, task scenarios were internally reviewed but not pilot-tested with end users, which could limit their realism and appropriateness. Finally, this study examined only therapists' perspectives, excluding other user groups (eg, patients) who also interact with cognitive rehabilitation software.

Future studies should expand usability evaluations to include diverse user groups and real-world clinical settings, enabling longitudinal verification of software improvements and their impact on therapeutic outcomes. Furthermore, integrating objective usability measures—such as automated cognitive load assessments or eye tracking—could provide deeper insights into user interaction patterns and interface demands.

Overall, this study contributes empirical evidence supporting structured, user-centered design approaches for medical device software, particularly in cognitively demanding rehabilitation contexts. The findings provide practical guidance for developers and usability engineers seeking to optimize UI and UX design in alignment with clinical workflows and user requirements, thereby improving both safety and clinical effectiveness in digital cognitive rehabilitation interventions.

Conclusions

This study conducted a formative evaluation of a tablet-based cognitive rehabilitation software applying CW and the SUS within the IEC 62366-1 framework. The results revealed critical usability challenges related to interface clarity, consistency, and information visibility—factors that directly affect user efficiency and the clinical reliability of software as a medical device.

In response to the identified issues, several targeted UI and UX improvement strategies were proposed, including resizing and repositioning the keypad, clarifying task initiation and termination processes, enhancing the visibility and labeling of instructional elements, and unifying the design of interaction features such as the hint and difficulty adjustment functions. These refinements are expected to reduce cognitive workload and support error-free operation.

Beyond individual design improvements, this study reinforces the importance of applying structured, standard-compliant usability engineering practices to medical and rehabilitation software. The integration of CW and the SUS proved effective in capturing both interaction-level issues and perceived usability gaps, underscoring the value of iterative, user-centered design processes during the early stages of development.

Future research should expand usability evaluations to real-world clinical environments and include direct feedback from end users. The adoption of objective usability measures such as eye tracking or use log analysis may further deepen insight into interaction behavior and inform next-generation interface design for digital therapeutics.

XSL•FO
RenderX

Acknowledgments

This study was supported by the Korea National Rehabilitation Center Rehabilitation Research & Development Support Program funded by the Ministry of Health and Welfare (NRCRSP-25TB01). Generative artificial intelligence tools (eg, ChatGPT or other large language models) were not used in the generation of this manuscript. All content was authored and reviewed by the listed authors.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: SH, HC, HK
Methodology: SH
Data collection: SH
Formal analysis: SH
Writing – original draft: SH
Writing – review & editing: SH
Supervision: HC, HK

Conflicts of Interest

None declared.

References

1. Maggio MG, De Bartolo D, Calabrò RS, Ciancarelli I, Cerasa A, Tonin P, et al. Computer-assisted cognitive rehabilitation in neurological patients: state-of-art and future perspectives. Front Neurol 2023 Sep 29;14:1255319. [doi: 10.3389/fneur.2023.1255319] [Medline: 37854065]
2. Szturm T, Parmar ST, Mehta K, Shetty DR, Kanitkar A, Eskicioglu R, et al. Game-based dual-task exercise program for children with cerebral palsy: blending balance, visuomotor and cognitive training: feasibility randomized control trial. Sensors (Basel) 2022 Jan 19;22(3):761 [FREE Full text] [doi: 10.3390/s22030761] [Medline: 35161508]
3. Ko EJ, Sung IY, Yuk JS, Jang DH, Yun G. A tablet computer-based cognitive training program for young children with cognitive impairment: a randomized controlled trial. Medicine (Baltimore) 2020 Mar;99(12):e19549 [FREE Full text] [doi: 10.1097/MD.0000000000019549] [Medline: 32195962]
4. Shin SH, Ko MH, Kim YH. Effect of computer-assisted cognitive rehabilitation program for patients with brain injury. J Korean Acad Rehabil Med 2002;26(1):1-8 [FREE Full text]
5. Hong HT, Song SI. Effectiveness of computerized cognitive rehabilitation therapy on cognitive function of children with disabilities: a systematic review. J Ment Health Res Intellect Disabil 2023 Jul 26;17(2):158-172 [FREE Full text] [doi: 10.1080/19315864.2023.2240739]
6. Pearce LM, Howell M, Yamato TP, Bacha JM, Pompeu JE, Howard K, et al. What attributes of digital devices are important to clinicians in rehabilitation? A cross-cultural best-worst scaling study. Int J Med Inform 2024 Nov;191:105589 [FREE Full text] [doi: 10.1016/j.ijmedinf.2024.105589] [Medline: 39137666]
7. Levac DE, Miller PA. Integrating virtual reality video games into practice: clinicians' experiences. Physiother Theory Pract 2013 Oct;29(7):504-512. [doi: 10.3109/09593985.2012.762078] [Medline: 23362843]
8. La Bara LM, Meloni L, Giusino D, Pietrantoni L. Assessment methods of usability and cognitive workload of rehabilitative exoskeletons: a systematic review. Appl Sci 2021 Aug 02;11(15):7146. [doi: 10.3390/app11157146]
9. Zayim N, Yildiz H, Yüce YK. Subjective cognitive load evaluation of a mobile personal health record application. Stud Health Technol Inform 2023 Oct 20;309:262-266. [doi: 10.3233/SHTI230793] [Medline: 37869854]
10. Martjan RS, Weimar SN, Terzidis O. A business model framework for software as a medical device startups in the European Union: mixed methods study. J Med Internet Res 2025 May 23;27:e67328 [FREE Full text] [doi: 10.2196/67328] [Medline: 40408133]
11. Jaspers MW. A comparison of usability methods for testing interactive health technologies: methodological aspects and empirical evidence. Int J Med Inform 2009 May;78(5):340-353. [doi: 10.1016/j.ijmedinf.2008.10.002] [Medline: 19046928]
12. Brooke J. SUS - a quick and dirty usability scale. Usability Evaluation in Industry. URL: https://rickvanderzwet.nl/trac/personal/export/104/liacs/hci/docs/SUS-questionaire.pdf [accessed 2025-02-18]
13. Park SW. Comparison before and after the application of the computerized cognitive rehabilitation program (CoTras-C) for children with cerebral palsy. J Korean Soc Integr Med 2021 Aug 31;9(3):9-18 [FREE Full text] [doi: 10.15268/KSIM.2021.9.3.09]

XSL•FO
RenderX

14. Lee HS, Hyeon JI. The effect of tablet PC cognitive rehabilitation training games on the improvement of game performance and cognitive ability of non-disabled and intellectual disability children and analysis of user effectiveness. Korean Assoc Intellect Disabil 2020 Sep 30;22(3):1-27. [doi: 10.35361/kjid.22.3.1]
15. Neilsen J. Usability Engineering. San Francisco, CA: Morgan Kaufmann; 1993.
16. Liu Y, Osvalder AL, Dahlman S. Exploring user background settings in cognitive walkthrough evaluation of medical prototype interfaces: a case study. Int J Ind Ergon 2005 Apr;35(4):379-390. [doi: 10.1016/j.ergon.2004.10.004]
17. Polson PG, Lewis C, Rieman J, Wharton C. Cognitive walkthroughs: a method for theory-based evaluation of user interfaces. Int J Man Mach Stud 1992 May;36(5):741-773. [doi: 10.1016/0020-7373(92)90039-N]
18. Wharton C, Rieman J, Lewis C, Polson P. The cognitive walkthrough method: a practitioner's guide. In: Nielsen J, Mack RL, editors. Usability Inspection Methods. New York, NY: John Wiley & Sons; Jun 03, 1994:105-140.
19. Bangor B, Kortum P, Miller J. Determining what individual SUS scores mean: adding an adjective rating scale. J Usability Stud 2009;4(3):114-123. [doi: 10.5555/2835587.2835589]
20. Sauro J. A Practical Guide to the System Usability Scale: Background, Benchmarks & Best Practices. Denver, CO: Measuring Usability LLC; 2011.
21. Valdés BA, Hilderman CG, Hung CT, Shirzad N, Van der Loos HF. Usability testing of gaming and social media applications for stroke and cerebral palsy upper limb rehabilitation. In: Proceedings of the 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. 2014 Presented at: IEEE '14; August 26-30, 2014; Chicago, IL p. 3602-3605. [doi: 10.1109/EMBC.2014.6944402]
22. Harte R, Glynn L, Rodríguez-Molinero A, Baker PM, Scharf T, Quinlan LR, et al. A human-centered design methodology to enhance the usability, human factors, and user experience of connected health systems: a three-phase methodology. JMIR Hum Factors 2017 Mar 16;4(1):e8 [FREE Full text] [doi: 10.2196/humanfactors.5443] [Medline: 28302594]
23. Irazoki E, Sánchez-Gómez MC, Contreras-Somoza LM, Toribio-Guzmán JM, Martín-Cilleros MV, Verdugo-Castro S, et al. A qualitative study of the cognitive rehabilitation program GRADIOR for people with cognitive impairment: outcomes of the focus group methodology. J Clin Med 2021 Feb 19;10(4):859 [FREE Full text] [doi: 10.3390/jcm10040859] [Medline: 33669716]
24. Ranzani R, Eicher L, Viggiano F, Engelbrecht B, Held JP, Lambercy O, et al. Towards a platform for robot-assisted minimally-supervised therapy of hand function: design and pilot usability evaluation. Front Bioeng Biotechnol 2021 Apr 15;9:652380 [FREE Full text] [doi: 10.3389/fbioe.2021.652380] [Medline: 33937218]
25. Zahabi M, Kaber DB, Swangnetr M. Usability and safety in electronic medical records interface design: a review of recent literature and guideline formulation. Hum Factors 2015 Aug;57(5):805-834. [doi: 10.1177/0018720815576827] [Medline: 25850118]
26. Cubukcu HC, Canbaz H. Usability assessments in clinical decision support systems. North Clin Istanb 2025 Jun 19;12(3):372-377. [doi: 10.14744/nci.2024.22316] [Medline: 40843327]
27. Olakotan O, Samuriwo R, Ismaila H, Atiku S. Usability challenges in electronic health records: impact on documentation burden and clinical workflow: a scoping review. J Eval Clin Pract 2025 Jun 29;31(4):e70189. [doi: 10.1111/jep.70189] [Medline: 40581989]
28. Blandford A, Furniss D, Vincent C. Patient safety and interactive medical devices: realigning work as imagined and work as done. Clin Risk 2014 Sep;20(5):107-110 [FREE Full text] [doi: 10.1177/1356262214556550] [Medline: 25866466]
29. Norman D. The Design of Everyday Things : Revised and Expanded Edition. New York, NY: Basic Books; 2013.
30. Ehrler F, Haller G, Sarrey E, Walesa M, Wipfli R, Lovis C. Assessing the usability of six data entry mobile interfaces for caregivers: a randomized trial. JMIR Hum Factors 2015 Dec 15;2(2):e15 [FREE Full text] [doi: 10.2196/humanfactors.4093] [Medline: 27025648]

Abbreviations

CW: cognitive walkthrough
IEC: International Electrotechnical Commission
NRC: National Rehabilitation Center
SUS: System Usability Scale
UI: user interface
UX: user experience

XSL•FO
RenderX

Edited by A Mavragani, J Sarvestan; submitted 10.04.25; peer-reviewed by A Gandhi, Q Xu; comments to author 23.06.25; revised version received 31.10.25; accepted 31.10.25; published 18.11.25.

Please cite as:
Hong S, Choi H, Kweon H
User-Centered Formative Evaluation of Cognitive Rehabilitation Software: Cognitive Walkthrough and System Usability Scale Study
JMIR Form Res 2025;9:e75805
URL: https://formative.jmir.org/2025/1/e75805
doi:10.2196/75805
PMID:41252193

©Seojin Hong, Hyun Choi, Hyosun Kweon. Originally published in JMIR Formative Research (https://formative.jmir.org), 18.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Three Body-Worn Accelerometers in the French NutriNet-Santé Cohort: Feasibility and Acceptability Study

Abdouramane Soumaré[1], MSc; Léopold K Fezeu[1], PhD; Jérôme Bouchan[1], PhD; Fabienne Delestre[2], Msc; Alice Bellicha[1], PhD; Greet Cardon[3], PhD; Alan Donnelly[4], PhD; Antje Hebestreit[5], PhD; Mathilde Touvier[1], PhD; Jean-Michel Oppert[1,2*], PhD, MD; Jérémy Vanhelst[1*], PhD

[1]Université Sorbonne Paris Nord and Université Paris Cité, INSERM, INRAE, CNAM, Centre for Research in Epidemiology and StatisticS (CRESS), Nutritional Epidemiology Research Team (EREN), F-93017, Bobigny, Paris, France

[2]Department of Nutrition, Human Nutrition Research Center Ile-de-France (CRNH IdF), Pitié-Salpêtrière Hospital (AP-HP), Sorbonne University, Paris, null, France

[3]Department of Movement and Sports Sciences, Ghent University, Ghent, Belgium

[4]Health Research Institute, University of Limerick, Limerick, Ireland

[5]Leibniz Institute for Prevention Research and Epidemiology-BIPS, Bremen, Germany

[*]these authors contributed equally

Corresponding Author:
Jérémy Vanhelst, PhD
Université Sorbonne Paris Nord and Université Paris Cité, INSERM, INRAE, CNAM, Centre for Research in Epidemiology and StatisticS (CRESS), Nutritional Epidemiology Research Team (EREN), F-93017
74 rue Marcel Cachin
Bobigny, Paris, 93000
France
Phone: 33 01 48 38 89 35
Email: jeremy.vanhelst@eren.smbh.univ-paris13.fr

Abstract

Background: Accurate assessment of physical activity (PA) in large population-based cohorts remains a major methodological challenge. Self-reported questionnaires, although commonly used due to low cost and simplicity, are prone to recall and social desirability biases, causing misclassification and weakened associations with health outcomes. Body-worn accelerometers provide more objective and reliable measurements, but their acceptability and feasibility in large-scale epidemiological studies must be carefully evaluated to ensure compliance, data quality, and scalability.

Objective: The primary objective was to assess the acceptability of using 3 body-worn accelerometer devices (Fitbit, ActivPAL, and ActiGraph) among healthy middle-aged adults participating in the NutriNet-Santé cohort. The secondary objective was to assess the feasibility of these devices in terms of wear-time compliance under free-living conditions.

Methods: This is an ancillary study of the European WEALTH (WEarable sensor Assessment of physicaL and eaTing beHaviors) project that was conducted between 2023 and 2024 in a subsample of participants of the NutriNet-Santé surveillance in France. This sample included 126 healthy participants (62 men), with a mean age of 46.3 (SD 11.3) years. Participants wore simultaneously 3 body-worn accelerometer devices (Fitbit [wrist], ActivPAL [thigh], and ActiGraph [waist]) for 7 consecutive days. After the wear period, participants completed a specific 22-item web-based questionnaire, regarding their acceptability of using each device. This questionnaire was based on the Technology Acceptance Model, which identifies perceived usefulness and ease of use as key determinants of technology acceptance. Items were rated on a 5-point Likert scale (1=strongly disagree to 5=strongly agree). Feasibility was assessed based on the accelerometer wear time data reported in a log diary by participants. A valid day was defined as ≥600 minutes per day of wear time, and a valid week as at least 4 of such days. Acceptability scores were compared between devices using ANOVA, and feasibility outcomes were compared using Kruskal-Wallis tests.

Results: The acceptability assessment based on the questionnaire revealed significant differences among the 3 devices ($P<.001$). The Fitbit achieved the highest acceptability score (mean 80.5/100, SD 8.13) across most criteria such as comfort, ease of use, and social acceptability, while the ActiGraph received the lowest score (mean 71.7, SD 8.68), mainly due to challenges with stability and interference during PA. In terms of feasibility, the 3 accelerometers demonstrated high compliance, with the ActivPAL recording the highest daily wear time, followed by the Fitbit and the ActiGraph ($P<.001$).

XSL•FO
RenderX

Conclusions: Results from our study showed that the Fitbit watch appears as the most accepted device for measuring PA in free-living conditions in the NutriNet-Santé study. The large-scale use of such a device must now be evaluated in terms of logistics, cost, and data privacy.

(JMIR Form Res 2025;9:e76167) doi:10.2196/76167

KEYWORDS

accelerometer; acceptability; feasibility; cohort study

Introduction

Physical activity (PA) is defined as any bodily movement produced by skeletal muscles that results in an energy expenditure higher than the resting state [1]. Robust and consistent evidence supports that PA is a key determinant of health throughout the lifetime [2]. PA is known to improve health and reduce the risk of a wide variety of chronic diseases such as cardiometabolic and respiratory diseases, obesity, and some cancers [3]. Conversely, physical inactivity, defined as not meeting public health PA guidelines, is recognized as the fourth leading risk factor for mortality worldwide [4].

In large epidemiological studies, measuring PA accurately and reliably remains a challenge. Questionnaires are the most commonly used tool in epidemiological studies because of their limited cost and burden (for short questionnaires) and the fact that they are relatively easy to administer on a large scale (especially in web cohorts) [5]. However, self-reported PA methodologies have significant limitations, including recall and social desirability bias, which can lead to meaningful errors in estimating the duration and intensity of PA or to attenuated associations with health end points [6,7]. By comparison, body-worn accelerometers devices measure body segment accelerations (such as wrist, arm, waist, thigh, or ankle) and quantify PA in terms of gravity units or activity counts accumulated over time [8,9]. As such, wearable sensors provide data with a high degree of transparency and harmonization, allowing for comparability across different studies, populations, and countries. High-resolution accelerometry data allow for complex analysis, such as development of PA patterns and performing compositional 24-hour activity analyses in order to associate with health status, to monitor trends and to refine PA guidelines. Hence, accelerometry is considered a more accurate and reliable method for assessing PA in free-living conditions [10,11].

Many accelerometers are available in the health research domain. The most widely used are the research-grade ActiGraph (now branded as Ametris) and the ActivPAL, both devices are found to be reliable and valid for measuring PA and sedentary behavior [12-14]. Recently, a new generation of devices such as the consumer-grade Fitbit has emerged, showing a high popularity among the population due to their user-friendly handling, lower cost, and ability to track multiple metrics in real time [14-16]. Investigators must consider multiple factors to increase participant compliance in wearing the device. While validity and reliability are major factors, the acceptability and feasibility of accelerometer use are also key considerations in determining their potential use in large-scale cohort studies. By definition, acceptability refers to participants' willingness to wear and use the device [17]. This can be influenced by several parameters such as comfort, ease of use, perceived usefulness, and social acceptability [18]. It is crucial to understand the participants' preferences, barriers, and facilitators of using body-worn sensors for ensuring long and seamless wear time of devices. Feasibility evaluates the practicality of using wearable devices in free-living conditions. It encompasses several dimensions, including the cost of the devices, the difficulty of ensuring proper mounting, and the logistical requirements, such as meeting participants or sending instructions to explain how to use the device. While these factors are important, the main criterion considered is compliance, specifically the time recorded when the participant effectively wore the device. Indeed, this aspect is essential for obtaining valid and reliable data over long periods, which is necessary for observational studies.

Currently, there is no study assessing the acceptability of wearable devices in middle-aged adults (aged 40-60 years) without chronic diseases. Most studies on the acceptability of PA measurement devices have been conducted using a mixed methods approach and on specific populations such as the older adults [10,19], people with chronic diseases [20], or children and adolescents [21,22]. Middle-aged adults represent an important target group for PA promotion. Studies have shown that increasing PA versus maintaining PA during middle age was associated with comparable lower mortality risk [23]. Moreover, middle-aged adults account for a large proportion of participants in longitudinal cohort studies, making their adherence to monitoring protocols essential for ensuring valid and scalable accelerometer-based surveillance. They thus represent a key target population for future large-scale accelerometer-based analyses. Therefore, the aim of this study was to assess the acceptability and feasibility of using 3 body-worn accelerometer devices in healthy middle-aged adults from the French NutriNet-Santé cohort.

Methods

Study Design

This study is an ancillary part of the European WEALTH (WEarable sensor Assessment of physicaL and eaTing beHaviors) project, which generally aims to develop standardized methods for classifying daily behaviors, including PA, from accelerometry and other sensor devices using advanced methods [24]. This study used data collected in France among participants of the NutriNet-Santé study. NutriNet-Santé is a web-based cohort launched in 2009 with the aim to investigate the associations between nutrition and health. As previously described elsewhere, participants older than18 years with access to the web have been continuously recruited since 2009 among

XSL•FO
RenderX

the general French population [25]. Individuals reporting chronic diseases, pregnant women, and participants who withdrew their consent before or during data collection did not fulfill eligibility criteria.

Eligibility criteria were (1) a registered participant of the NutriNet-Santé cohort, (2) aged 18 years and older, (3) living in and around Paris (Ile-de-France), (4) be familiar with smartphone technologies, (5) able to understand and communicate in French, and (6) able to attend the laboratory visit for collecting labeled data and to complete data collection during the following 9 days (no holidays, hospital visits, shift work, etc). Among the eligible population, 1350 NutriNet-Santé participants were randomly selected after stratification by sex and age and invited to participate in the study, resulting in a final sample of 156 participants.

Ethical Considerations

This ancillary study is classified as research involving human person (RIPH2), interventional, except for health products, with minimal risk and constraints in accordance with Article L 1121-1 of the French Public Health Code. The WEALTH study and NutriNet-Santé study were approved by the French ethic committee, conducted in accordance with the Declaration of Helsinki, and approved by the institutional review board of the French Institute for Health and Medical Research (IRB INSERM #0000388FWA00005831) and by the National Commission on Informatics and Liberty (CNIL #908,450 and #909,216). All participants were fully informed about the study's aims and procedures. Written informed consent was obtained from all participants prior to inclusion. All data collected were pseudonymized, and no participant could be identified in either the analyses or the reporting. Participants did not receive financial compensation. However, they were provided with detailed feedback on their PA behavior and emotional status as a form of nonmonetary incentive. No identifiable images of participants are presented in this manuscript or in supplementary materials.

Procedures

During the WEALTH study, all French participants were invited to the Pitié Salpêtrière Hospital (APHP, Paris) for an inclusion visit (2 hours) (day 1). During this session, investigators explained the appropriate use of the wearable devices to each participant. Data collection followed a standardized 9-day free-living protocol, consisting of 2 lead-in days (day 1: laboratory visit; day 0: familiarization day), followed by 7 full monitoring days (day 1 to day 7), covering both weekdays and weekend days. Our protocol did not specifically require a valid weekend day, as the primary aim was not to assess detailed PA patterns. Since the Fitbit is widely used in consumer settings, and ActiGraph and ActivPAL are established research-grade devices frequently used in large-scale studies such as National Health and Nutrition Examination Survey, we decided to use these 3 devices in our study. The Fitbit watch (Charge5) was worn on the nondominant wrist, the ActiGraph, now branded as Ametris (wGT3X-BT) on the right waist with an elastic belt, and the ActivPAL (ActivPAL l 3 micro, PAL Technologies Ltd) on the right thigh after being waterproofed in a nitrile sleeve. Participants also wore a LifeQ-enabled smartwatch (Skagen Designs Generation 6) on dominant wrist. However, data from this device were not included in the present analysis. Participants were instructed to wear the ActivPAL continuously, including during sleep, and to remove it only for water-based activities (eg, swimming). In contrast, the ActiGraph was worn only during waking hours and was removed at night, as well as for water-based activities and contact sports. The Fitbit was worn continuously. Prior to the measurement period, participants received an introduction, by investigators, to the wearable devices in a face-to-face session and they could ask any questions. In addition, each participant also obtained written handling manuals to assist with the use and mounting of wearable devices. Finally, participants had an access to a study center hotline for support throughout the wear period. In addition, participants were instructed to keep a log diary collecting when and why the devices (Fitbit, ActiGraph, and ActivPAL) were removed to assess wear and nonwear times. They were also informed about the specific purpose of each wearable device and the rationale to use these different devices simultaneously. The devices were collected after 7-day monitoring period, and data were transferred to a computer for analysis.

Measurements

Participants' Characteristics

Anthropometric measurements were taken during the inclusion visit at Pitié Salpétrière hospital. Weight and height were measured in light clothes, without shoes, using an electronic scale and a wall-mounted scale, respectively. The BMI was calculated as weight (kg) divided by height² (m^2). Nutritional status was assessed according to the World Health Organization definitions: underweight (<18.5 kg/m²), normal weight (≥18.5 kg/m² and <25 kg/m²), overweight (≥25.0 kg/m² and <30 kg/m²), and obesity (≥30.0 kg/m²).

Acceptability Questionnaire

After the 7-day wear period, participants completed a specific web-based questionnaire regarding their acceptance of using the 3 wearable devices. The 22-item acceptance questionnaire was adapted from previous studies [19,20] and developed within the framework of the Technology Acceptance Model (TAM) [26]. According to TAM, technology adoption is driven by perceived usefulness, perceived ease of use, attitude toward use, and behavioral intention to use. These constructs were operationalized in our questionnaire through specific dimensions: (1) comfort, constraints, and interferences in daily activities (perceived ease of use); (2*i*) changes in PA habits, technical aspects, and perceived usefulness; (3) social acceptability (attitude toward using); and (4) satisfaction and long-term use intention (behavioral intention to use). Participants rated each item on a 5-point Likert scale, indicating the extent to which they agreed with each statement, from 1 (strongly disagree) to 5 (strongly agree). In addition, participants completed a ranking of the devices on their preferences for wearing them in future studies and also provided open-ended feedback on their experience with PA measurement technologies. The questionnaire is available on the web (in French) [27].

XSL•FO
RenderX

Feasibility

Feasibility was assessed through participants' compliance with wearing the body-worn accelerometer devices, using data from the log diary. Participants were asked to record the time of all device removals and replacements during the consecutive 7-day period. Wear time for the 3 devices was calculated for each day by summing the durations noted in the log diary. Established accelerometer wear time criteria were used to calculate the number of valid days (defined as a day with wear time ≥600 minutes) for each device, as well as the number of valid weeks (defined as a week with at least 4 valid days) [12]. These 2 thresholds are widely used in accelerometer-based studies among adults [12]. Indeed, in a systematic review with practical considerations, the authors recommended a minimum of 600 minutes per day (10 hours) and at least 4 valid days per week to ensure that the accelerometer captures a representative portion of daily activity and provides reliable estimates of habitual PA patterns under free-living conditions [12].

Statistical Analysis

For the description of participants' characteristics, means and standard deviations were calculated for continuous variables, and number and percentage for categorical variables. For the acceptance questionnaire assessing the use of the activity tracker, the average score for each item was calculated. Total acceptability scores for each device were obtained by summing the scores from individual items and were then compared using an ANOVA test with a significance level of $P<.05$. After checking for normality, the daily wear time of the 3 body-worn accelerometer devices (in minutes per day) was described using medians and the 25th and 75th percentiles. The median wear time for these 3 devices over the 7-day period was calculated and compared using a Kruskal-Wallis test, with a significance level of $P<.05$. All analyses were performed in SAS software (version 9.4; SAS Institute).

Results

Characteristics of Participants

Among the 156 participants, 150 had valid data on wear time for all 3 wearable devices. Of these, 130 completed the acceptance questionnaire, with 126 providing complete data for analysis. The characteristics of the included participants are described in Table 1. The study sample consisted of 126 adults (62 men, 64 women), with a mean average age of 46.3 (SD 11.3) years. In addition, 52.4% (66/126) of participants reported having prior experience with body-worn accelerometer devices.

Table 1. Characteristic of participants (N=126).

Characteristics	Participants, n (%)	Mean (SD)
Gender		
Men	62 (49.2)	N/A[a]
Women	64 (50.8)	N/A
Age (years)	N/A	46.3 (11.3)
Height (cm)	N/A	170 (10.1)
Weight (kg)	N/A	69.7 (12.5)
BMI (kg/m²)	N/A	23.9 (3.19)
BMI categories		
Underweight	1 (0.79)	N/A
Normal	85 (67.5)	N/A
Overweight	35 (27.8)	N/A
Obesity	5 (4)	N/A
Experience using PA[b] measurement technologies		
Yes	66 (52.4)	N/A
No	60 (47.6)	N/A

[a]N/A: not applicable.

[b]PA: physical activity.

Acceptability

Evaluation of the 3 devices using the acceptability questionnaire revealed a significant overall difference in scores ($P<.001$). Pairwise comparisons also showed differences, with the Fitbit obtaining the highest score (mean 80.5, SD 8.13 points) and the ActiGraph (mean 71.7, SD 8.68) the lowest (Table 2). However, there was variability in scores according to the different concepts measured in the acceptance questionnaire (Table 3). The Fitbit watch was perceived as the most comfortable to wear overall, followed by the ActivPAL and then the ActiGraph, despite the ActiGraph obtaining the lowest scores for items related to skin conditions (items 2 and 3). For these same items, the ActiGraph scored higher than the other devices, indicating that it caused fewer skin irritations.

XSL•FO
RenderX

Regarding the items related to constraints and interference with daily activities, all 3 devices obtained satisfying scores (above 3.6/5), although scores were slightly lower for interference with sleep (item 6). The ActiGraph had the lowest scores on these criteria (items 5-8). Changes in PA habits (item 9) were most pronounced for the Fitbit watch, which obtained the lowest score (3.71/5). All 3 devices received relatively high scores (4.32/5) for ease of use (item 10). However, the ActivPAL and the ActiGraph had lower scores (2.70 and 2.65, respectively) than the Fitbit (4.06/5) for understanding the device-wearing instructions (item 11).

Regarding social acceptability (items 14-17), the Fitbit watch obtained higher scores than the other devices, while the ActiGraph had the lowest scores, particularly for discretion (2.79) and esthetics (2.32). All 3 devices scored well on items related to overall satisfaction and intention to use in the long term in the NutriNet Santé study (items 18 and 19), with the Fitbit watch and the ActivPAL showing the highest scores.

Table 2. Average acceptability score obtained by participants for each accelerometer.

	Mean (SD) score	Minimum-Maximum	*P* value[a]
Fitbit	80.5 (8.13)	59-95	N/A[b]
ActivPAL	77.8 (8.87)	50-95	.015
ActiGraph	71.7 (8.68)	45-89	<.001

[a]Reference: Fitbit. Minimum score: 19; maximum score: 95. Global *P* value for ANOVA: *P*<.001.

[b]N/A: not applicable.

XSL•FO
RenderX

Table 3. Results (answer averages) of the acceptance questionnaire assessing the use of wearable accelerometersa.

Item	Fitbit	ActivPAL	ActiGraph	Average
Comfort level				
1. In general, I found "the device" comfortable to wear.	4.11	3.89	3.10	3.7
2. "The device" irritated my skin (itching, redness, unpleasant sensation, etc).[b]	4.01	4.06	4.67	4.25
3. "The device" left traces on my skin (Marks, visible lines).[b]	3.95	3.81	4.66	4.14
4. After a few hours, I was still aware that I was wearing "the device."[b]	3.75	4.00	2.98	3.58
Constraints and interferences in daily activities				
5. I found it easy to work while wearing the "device."	4.70	4.77	4.42	4.63
6. I found it hard to sleep while wearing the "device."[b]	3.72	3.99	3.24	3.65
7. I found it easy to carry out my daily activities (leisure, housework) while wearing "the device."	4.75	4.68	4.27	4.57
8. I found it easy to put on my usual clothes while wearing "the device."	4.83	4.41	3.56	4.27
Changes in PA[c] habits				
9. Wearing the "device" made me move more than usual.[b]	3.71	4.12	4.13	3.99
Ease of use				
10. Overall, I found "the device" easy to use.	4.37	4.47	4.12	4.32
11. The instructions and guidelines provided by the researcher on the use of the "device" were difficult to understand.[b]	4.06	2.70	2.65	3.14
Technical reports				
12. Overall, I was satisfied with the autonomy of the "device."	4.33	4.69	4.36	4.46
13. I found the charging time of the "device" too long.[b]	4.10	4.53	4.52	4.38
Social acceptability				
14. I found the "device" esthetically good.	3.81	2.72	2.32	2.95
15. I found the size of the "device" to be appropriate and discreet.	4.11	3.79	2.79	3.56
16. I found it embarrassing to wear "the device" in public.[b]	4.5	3.93	3.40	3.94
17. I have had some negative comments from other people when I was wearing "the device." [b]	4.84	4.78	4.63	4.75
Satisfaction and long-term use intention				
18. Overall, I'm disappointed with my experience with the "device."[b]	4.29	4.05	3.73	4.02
19. If I were asked to wear the "'device"' for a week within the NutriNet Santé study, I would definitely accept.	4.52	4.44	4.19	4.38
Average by device, mean (SD)	4.23 (0.43)	3.78 (0.46)	4.10 (0.47)	N/A[d]

[a]Scoring: 1=Strongly disagree, 2=disagree, 3=Neither agree nor disagree, 4=Agree, and 5=Strongly agree.

[b]Items were negatively framed and therefore reverse-coded prior to analysis.

[c]PA: physical activity.

[d]N/A: not applicable.

Table 4 summarizes the positive and prohibitive negative points most frequently cited by participants for each accelerometer. For the Fitbit watch, which had the highest acceptability score, the most frequently mentioned positive aspect was its diverse functionalities (sleep tracking, physical behavior, and other parameters). In contrast, for the ActiGraph, the device with the lowest score, the most recurrent noted negative point was its instability during wear, particularly during sports activities.

XSL•FO
RenderX

Table 4. Most recurrent positive and negative statements for each device (N=126).

	Most recurrent statement	Participants, n (%)
Fitbit		
Positive statement	Tracks activity, sleep, and other parameters (heart rate, number of steps, and notifications)	47 (37.6)
Negative statement	Skin affections (irritation, itching, marks, allergies, plastic, discomfort...)	37 (29.6)
ActivPAL		
Positive statement	Discreet and comfortable (easy to forget, lightweight, invisible...)	78 (62.4)
Negative statement	Skin affections (irritation, redness, itching, and marks after removal)	20 (16)
ActiGraph		
Positive statement	Easy to use (easy to remove, replace, attach to belt, and adjustable)	33 (26.4)
Negative statement	Unstable (difficult to hold in place, elastic waistband slips, and moves during certain sports activities)	48 (38.4)

Feasibility

The description and comparison of the wear time for each device (medians and 25th and 75th centiles) over the 7-day wear period are shown in Table 5. A significant difference was found among the devices. Comparative analysis showed that the ActivPAL had the greatest wear time, with a significant difference compared with both ActiGraph (P<.001) and Fitbit (P<.001). Pairwise comparisons also showed differences between ActivPAL and Fitbit (P<.001). The Fitbit also exhibited a higher wear time than the ActiGraph (P<.001). Overall, the ActivPAL had a total of 1049 valid days (person × day), the Fitbit accumulated 1046 valid days (person × day), and the ActiGraph had 1006 valid days (person × day). All participants validated the week of PA evaluation, that is, achieving at least 4 valid days with a wear time of ≥600 minutes per day.

Table 5. Wear time (minutes per day) for the 3 accelerometers.

	Median wear time (25th; 75th centile)			*P* value
	ActiGraph	ActivPAL	Fitbit	
Day 1	871 (795; 941)	1440 (1440; 1440)	1440 (1390; 1440)	
Day 2	930 (885; 983)	1440 (1440; 1440)	1440 (1440; 1440)	
Day 3	950 (887; 1050)	1440 (1440; 1440)	1440 (1415; 1440)	
Day 4	945 (855; 1340)	1440 (1440; 1440)	1440 (1435; 1440)	
Day 5	975 (899; 1335)	1440 (1440; 1440)	1440 (1440; 1440)	
Day 6	964 (895; 1057)	1440 (1440; 1440)	1440 (1440; 1440)	
Day 7	890 (808; 954)	1440 (1440; 1440)	1440 (1440; 1440)	
Over the 7 days	930 (855; 1020)[a,b]	1440 (1440; 1440)[b,c]	1440 (1421; 1440)[a,c]	<.001

[a]Pairwise post hoc comparisons (ActivPAL).

[b]Pairwise post hoc comparisons (Fitbit).

[c]Pairwise post hoc comparisons (ActiGraph).

Discussion

Principal Findings

There is a pressing need to improve methods for assessing PA in large cohort studies. An important challenge is to identify the most suitable device by finding the optimal compromise between acceptability and feasibility for the participants. The aim of our study was to assess the feasibility and acceptability of using body-worn accelerometer devices (ActiGraph, ActivPAL, and the Fitbit watch) in a subsample of the NutriNet-Santé cohort for potential future implementation. In terms of acceptability and feasibility, our study suggests that the Fitbit would be the most suitable device for measuring PA in free-living conditions in the framework of a large cohort such as the NutriNet-Santé study.

The 3 devices were perceived differently on the different dimensions of acceptability. The Fitbit watch and ActivPAL were rated the most comfortable overall, despite receiving lower scores for skin-related issues, whereas the ActiGraph was evaluated more positively in terms of skin conditions. The Fitbit watch was worn on the wrist, a familiar placement for such devices, which could explain why participants perceived it as the most comfortable. However, like the ActivPAL, it requires direct skin contact, which can cause irritation or marks, a concern not shared by the ActiGraph, which can be worn over clothing. The most frequently reported negative aspect by

XSL•FO
RenderX

participants regarding the Fitbit watch is that it causes irritations and marks. These conditions may be due to the plastic or to specific materials of the wristband.

In terms of interference with daily activities, the high scores for all 3 devices indicate that they cause minimal disruption to participants' daily movement and life. However, participants rated the ActiGraph lower score due to interference with wearing regular clothes, possibly due to the need for a fastening belt, which may not be suitable for all types of clothing. The impact on sleep may reflect the comfort level for the ActiGraph and the ActivPAL devices, as well as the notifications and light signals emitted by the Fitbit watch, which may disturb rest. Comfort level and interference with daily activities are key aspects that may influence participants' attitudes toward the devices ("Attitude Toward Using") and impact their intention for future use, as explained by the TAM.

Comparison With Previous Work

Although notifications were sent only to participants' phones as part of this protocol, the Fitbit watch received a valuable score for change in PA habits, which actually reflects its potential to influence behavior through real-time tracking and feedback that may encourage PA changes. Indeed, these features correspond to the most frequently cited positive aspect by participants about the Fitbit watch, reflecting its Perceived Usefulness, 1 of the 2 main determinants of acceptability according in the TAM [26]. Similar findings have been observed in a study performed in adolescents, where devices with interactive features positively impacted PA behavior [28]. However, there is no evidence supporting that this reactivity persists over time. A study on the reactivity of a pedometer with interactive feedback features worn for 2 weeks demonstrated that this reactivity lasts only about a week, and in the absence of intervention, habitual behaviors return during the second week, representing a more accurate estimate of habitual PA in adults [29]. Therefore, we recommend measuring daily PA levels after 1 week of wearing such a device, using the first week as a priming phase. A second option is to consider a blinding procedure when feasible.

All 3 devices were perceived positively in terms of ease of use, although participants found the instructions for the ActivPAL and the ActiGraph more complicated to understand. This suggests that the clarity of the instructions could be improved to enhance Perceived Ease of Use, another key determinant of acceptability in the TAM [26]. In terms of social acceptability, the Fitbit watch exhibited the highest score, while the ActiGraph and the ActivPAL were noted less favorably, with the lowest scores for the ActiGraph, particularly in terms of esthetics, discretion, and wearing in public. The differences observed between the ActiGraph and the ActivPAL lie in the fact that the ActivPAL tends to remain invisible under clothing during the winter season, unlike the ActiGraph, which is hard to miss if worn over clothing. These negatively judged esthetic aspects may lead to a negative attitude of participants toward the devices "Attitude Toward Using," which would unfavorably affect the intention for long-term use, according to the TAM [26].

Concerning feasibility, the results of our study showed that the ActivPAL accelerometer had the highest compliance level, which is not surprising given that it is attached with waterproofed and adhesive dressing to the skin of the thigh. This was followed by the Fitbit watch and the ActiGraph. A review also highlighted that the ActivPAL demonstrated longer wear times than the ActiGraph, potentially due to differences in wear protocols and attachment methods [30]. Indeed, the lower ranking of the ActiGraph can be explained by its placement around the waist and its removal for certain daily activities, such as contact sports, high-impact activities, and water-based activities (swimming, showering, bathing, etc) due to its nonwater resistance. Participants were also asked to remove this device during sleep periods (from bedtime to waking time), which could lead to a risk of forgetting to put it back on, thus affecting the data quality [31]. The Fitbit watch has a high compliance level compared with the ActiGraph, likely due to its wrist placement and water resistance. Indeed, a study has shown that compliance is higher for accelerometers worn on the wrist than on the waist [10]. However, the Fitbit watch needs to be removed at least once for recharging during a 1-week assessment of daily PA. This might explain the lower compliance of this device compared with the ActivPAL, which has a battery life of more than 7 days. Wearing an accelerometer on the wrist and ensuring that it is waterproof enhance continuous use both day and night, improving overall compliance [31,32]. In 2011, the National Health and Nutrition Examination Survey study (United States) using the ActiGraph accelerometer modified the data collection protocol, transitioning from waist-worn to wrist-worn to improve participant compliance and add sleep data [32]. This change aligns with the increasing importance of recent studies investigating 24-hour movement (sedentary lifestyle, PA, and sleep) and their combined effects on health [33-35]. Several studies suggest that health outcomes are influenced by the balance among these 3 behaviors [33-37]. These findings highlight the importance of technical characteristics such as battery life, waterproofing, and placement for ensuring compliance in epidemiological studies. However, to reproduce such a protocol in a large cohort remains a challenge. Indeed, despite the higher compliance found for the ActivPAL, this device requires in-person visits for proper placement, as it is attached to the skin using adhesive dressing, and its removal or replacement is more complex compared with other devices. These logistical and technical constraints contrast with the ActiGraph and the Fitbit, which can be easily mailed to participants with instructions and self-applied, making it more practical for large-scale studies. Nonetheless, the observed compliance level for the 3 accelerometers is excellent, as they have only valid weeks (4 days with wear time ≥600 minutes per day), which is required to obtain valid measurements of PA in free-living conditions, as previously also reported in a review between ActiGraph and ActivPAL [30].

In summary, the results from our study indicate that the Fitbit watch is the most accepted body-worn accelerometer device by participants compared with the ActiGraph and the ActivPAL. Its lower cost (€159.95 [US $174.20]) compared with the other 2 devices (>€292 [US $318]) also makes it a favorable option for use in large-scale population studies. Thus, this tool represents the best compromise between acceptability and feasibility and may be a suitable alternative for a large web

XSL•FO
RenderX

cohort, such as the NutriNet-Santé study. In addition, previous models of the Fitbit watch have shown acceptable validity for assessing moderate to vigorous intensity PA (r=0.73 compared with ActiGraph GT3X+) [38], sedentary behavior (intraclass correlation coefficient=0.95% compared with ActivPAL) [39], total daily energy expenditure (r=0.84 compared with doubly labeled water in free-living conditions) [40], and sleep (sensitivity=0.96; specificity=0.61 compared with polysomnography) [41].

Limitations

Our study has several limitations that should be acknowledged. First, wearing 4 devices simultaneously could have caused discomfort and make it difficult for participants to assess each sensor individually. Second, the study was conducted between September and February. Seasonal variations, particularly during summer, may affect participants' comfort and the discretion of devices, such as the ActivPAL, which is typically worn under clothing. Third, although the ActiGraph is now widely worn on the wrist in research studies, we were unable to assess this placement, as participants were already wearing 2 wrist-based accelerometers (Fitbit and LifeQ), which may have influenced perceptions and comparability. Indeed, we could consider that the position of the device, rather than the device itself, may be a critical factor influencing our results [42]. Fourth, since participation in the NutriNet-Santé cohort is voluntary, individuals tend to be healthier and have higher education levels than the general French population. In addition, participants likely adhered better to the study methodology because they are active cohort involvement, which may lead to higher acceptance. Therefore, our results cannot be extrapolated to all populations. Fifth, the apparent low recruitment rate (11.6%) should not be interpreted as a recruitment difficulty but rather as a direct consequence of the original study design. In the WEALTH project, it had been decided to include only 150 participants per country according to the initial sample size calculation. Therefore, in our ancillary study, that target was met and it was not possible to include more participants. Other factors should also be considered including the validity and reliability of data collected from these different wearable devices, that could affect the findings. Finally, we acknowledge that our study did not assess the quality, integrity, or stability of the data collected by the 3 devices. Future studies should specifically address these aspects to complement our findings.

Conclusions

Our study demonstrates that body-worn accelerometers are both feasible and acceptable among middle-aged adults in a large French web-based cohort. The Fitbit watch appears the most accepted device to measure PA in free-living conditions in participants. The Fitbit is also commercially available, which could facilitate broader participation by allowing individuals to contribute data from their personal devices. This accessibility could have a positive impact on the scalability of PA measurements in epidemiological studies. In addition, this type of device may be suitable to measure PA over several weeks or months at different times throughout the year, providing valuable insights into seasonal variations in PA. However, it is crucial to address data privacy concerns to protect participants' personal information and ensure the accuracy of the PA data collected and analyzed by researchers.

Acknowledgments

The authors are grateful for the participation of the volunteers in France in this examination. The authors also thank Selim Aloui (manager), Thi Hong Van Duong, Régis Gatibelza, Jagatjit Mohinder and Aladi Timera (computer scientists); Julien Allegre, Nathalie Arnault, Laurent Bourhis, and Nicolas Dechamp (data managers and statisticians); and Maria Gomes (participants' support) for their technical contribution to the NutriNet-Santé study, for their dedication and engagement to collect and manage the data used for this study, and for ensuring continuing communication with the cohort participants. This work reflects only the authors' views, and the funders are not responsible for any use that might be made of the information it contains. Researchers were independent from funders. The research question developed in this paper corresponds to a strong concern of the participants involved in the NutriNet-Santé cohort and of the public in general. The results of this study will be disseminated to the NutriNet-Santé participants through the cohort website and public seminars. The WEALTH Project is funded by the Joint Programming Initiative Healthy Diet for a Healthy Life, a research and innovation initiative of EU member states and associated countries (grant 727565 under STAMIFY [Standardized measurement, monitoring and/or biomarkers to study food intake, physical activity, and health]). The funding agencies supporting this work are (in alphabetical order of participating countries): Belgium: Research Foundation—Flanders (Fonds Wetenschappelijk Onderzoek—Vlaanderen); Czechia: Ministry of Education, Youth and Sports (Ministerstvo školství, mládeže a tělovýchovy); France: French National Research Agency (Agence Nationale de la Recherche); Germany: Federal Ministry of Education and Research (BMBF); and Ireland: Health Research Board. In addition, the NutriNet-Santé cohort study was supported by the following public institutions: Ministère de la Santé, Santé Publique France, Institut National de la Santé et de la Recherche Médicale, Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement, Conservatoire National des Arts et Métiers, and Université Sorbonne Paris Nord. Study investigators are independent from the funders. Funders had no role in the study design, the collection, analysis, interpretation of data, the writing of the manuscript, or the decision to submit the paper for publication.

Data Availability

The datasets used and/or analyzed during this study are available from the corresponding author on reasonable request. Researchers from public institutions can submit a collaboration request including their institution and a brief description of the project to

XSL•FO
RenderX

collaboration@etude-nutrinet-sante. All requests will be reviewed by the steering committee of the NutriNet-Santé study. A financial contribution may be requested. If the collaboration is accepted, a data access agreement will be necessary and appropriate authorizations from the competent administrative authorities may be needed. In accordance with existing regulations, no personal data will be accessible.

Authors' Contributions

AS, LKF, JB, JMO, and JV contributed to the study conception and design. AS, LKF, JB, FD, and AB developed the methodology and collected data. AS and LKF performed statistical analysis. AB, GC, AD, MT, and AH contributed to the interpretation of results and provided critical revisions. JV, LKF, and JMO supervised the study. AS drafted the first version of the manuscript, and all authors contributed to reviewing and editing the final version. JMO and JV contributed equally to this work. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Caspersen CJ, Powell KE, Christenson GM. Physical activity, exercise, and physical fitness: definitions and distinctions for health-related research. Public Health Rep 1985;100(2):126-131 [FREE Full text] [Medline: 3920711]
2. Varela AR, Pratt M, Harris J, Lecy J, Salvo D, Brownson RC, et al. Mapping the historical development of physical activity and health research: a structured literature review and citation network analysis. Prev Med 2018;111:466-472. [doi: 10.1016/j.ypmed.2017.10.020] [Medline: 29709233]
3. Pedersen BK, Saltin B. Exercise as medicine - evidence for prescribing exercise as therapy in 26 different chronic diseases. Scand J Med Sci Sports 2015;25 Suppl 3:1-72. [doi: 10.1111/sms.12581] [Medline: 26606383]
4. Bull FC, Al-Ansari SS, Biddle S, Borodulin K, Buman MP, Cardon G, et al. World Health Organization 2020 guidelines on physical activity and sedentary behaviour. Br J Sports Med 2020;54(24):1451-1462 [FREE Full text] [doi: 10.1136/bjsports-2020-102955] [Medline: 33239350]
5. Strath SJ, Kaminsky LA, Ainsworth BE, Ekelund U, Freedson PS, Gary RA, American Heart Association Physical Activity Committee of the Council on LifestyleCardiometabolic HealthCardiovascular, Exercise, Cardiac RehabilitationPrevention Committee of the Council on Clinical Cardiology,Council. Guide to the assessment of physical activity: clinical and research applications: a scientific statement from the American Heart Association. Circulation 2013;128(20):2259-2279. [doi: 10.1161/01.cir.0000435708.67487.da] [Medline: 24126387]
6. Shephard RJ. Limits to the measurement of habitual physical activity by questionnaires. Br J Sports Med 2003;37(3):197-206; discussion 206 [FREE Full text] [doi: 10.1136/bjsm.37.3.197] [Medline: 12782543]
7. Prince SA, Adamo KB, Hamel M, Hardt J, Connor Gorber S, Tremblay M. A comparison of direct versus self-report measures for assessing physical activity in adults: a systematic review. Int J Behav Nutr Phys Act 2008;5:56 [FREE Full text] [doi: 10.1186/1479-5868-5-56] [Medline: 18990237]
8. Vanhelst J. Physical activity assessment using accelerometry. Rev Epidemiol Sante Publique 2019;67(2):126-134. [doi: 10.1016/j.respe.2018.10.010] [Medline: 30658869]
9. Chen KY, Bassett DR. The technology of accelerometry-based activity monitors: current and future. Med Sci Sports Exerc 2005;37(11 Suppl):S490-S500. [doi: 10.1249/01.mss.0000185571.49104.82] [Medline: 16294112]
10. Liu F, Wanigatunga A, Schrack J. Assessment of physical activity in adults using wrist accelerometers. Epidemiol Rev 2022;43(1):65-93 [FREE Full text] [doi: 10.1093/epirev/mxab004] [Medline: 34215874]
11. Skender S, Ose J, Chang-Claude J, Paskow M, Brühmann B, Siegel EM, et al. Accelerometry and physical activity questionnaires—a systematic review. BMC Public Health 2016;16:515 [FREE Full text] [doi: 10.1186/s12889-016-3172-0] [Medline: 27306667]
12. Migueles JH, Cadenas-Sanchez C, Ekelund U, Delisle Nyström C, Mora-Gonzalez J, Löf M, et al. Accelerometer data collection and processing criteria to assess physical activity and other outcomes: a systematic review and practical considerations. Sports Med 2017;47(9):1821-1845 [FREE Full text] [doi: 10.1007/s40279-017-0716-0] [Medline: 28303543]
13. Blackwood J, Suzuki R, Webster N, Karczewski H, Ziccardi T, Shah S. Use of activPAL to measure physical activity in community-dwelling older adults: a systematic review. Arch Rehabil Res Clin Transl 2022;4(2):100190 [FREE Full text] [doi: 10.1016/j.arrct.2022.100190] [Medline: 35756981]
14. Phillips SM, Summerbell C, Hobbs M, Hesketh KR, Saxena S, Muir C, et al. A systematic review of the validity, reliability, and feasibility of measurement tools used to assess the physical activity and sedentary behaviour of pre-school aged children. Int J Behav Nutr Phys Act 2021;18(1):141 [FREE Full text] [doi: 10.1186/s12966-021-01132-9] [Medline: 34732219]
15. Irwin C, Gary R. Systematic review of fitbit charge 2 validation studies for exercise tracking. Transl J Am Coll Sports Med 2022;7(4):1-7 [FREE Full text] [doi: 10.1249/tjx.0000000000000215] [Medline: 36711436]

XSL•FO
RenderX

16. Feehan LM, Geldman J, Sayre EC, Park C, Ezzat AM, Yoo JY, et al. Accuracy of Fitbit devices: systematic review and narrative syntheses of quantitative data. JMIR Mhealth Uhealth 2018;6(8):e10527 [FREE Full text] [doi: 10.2196/10527] [Medline: 30093371]
17. Dillon A, Morris M. User acceptance of information technology: theories and models. Annu Rev Inf Sci Technol 1996;31:3-32.
18. McCallum C, Rooksby J, Gray CM. Evaluating the impact of physical activity apps and wearables: interdisciplinary review. JMIR Mhealth Uhealth 2018;6(3):e58 [FREE Full text] [doi: 10.2196/mhealth.9054] [Medline: 29572200]
19. Zhang Z, Giordani B, Margulis A, Chen W. Efficacy and acceptability of using wearable activity trackers in older adults living in retirement communities: a mixed method study. BMC Geriatr 2022;22(1):231 [FREE Full text] [doi: 10.1186/s12877-022-02931-w] [Medline: 35313819]
20. Mercer K, Giangregorio L, Schneider E, Chilana P, Li M, Grindrod K. Acceptance of commercially available wearable activity trackers among adults aged over 50 and with chronic illness: a mixed-methods evaluation. JMIR Mhealth Uhealth 2016;4(1):e7 [FREE Full text] [doi: 10.2196/mhealth.4225] [Medline: 26818775]
21. Creaser AV, Hall J, Costa S, Bingham DD, Clemes SA. Exploring families' acceptance of wearable activity trackers: a mixed-methods study. Int J Environ Res Public Health 2022;19(6):3472 [FREE Full text] [doi: 10.3390/ijerph19063472] [Medline: 35329166]
22. Clark ELM, Gulley LD, Hilkin AM, Rockette-Wagner B, Leach HJ, Lucas-Thompson RG, et al. Feasibility and acceptability of accelerometer measurement of physical activity in pregnant adolescents. Int J Environ Res Public Health 2021;18(5):2216 [FREE Full text] [doi: 10.3390/ijerph18052216] [Medline: 33668149]
23. Saint-Maurice PF, Coughlan D, Kelly SP, Keadle SK, Cook MB, Carlson SA, et al. Association of leisure-time physical activity across the adult life course with all-cause and cause-specific mortality. JAMA Netw Open 2019;2(3):e190355 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.0355] [Medline: 30848809]
24. Hayes G, Buck C, Cardon G, Cimler R, Elavsky S, Fezeu K L, et al. Standardized methods for evaluating physical and eating behaviours: the wearable sensor assessment of physical and eating behaviours “WEALTH” project. JMIR Res. Protoc 2024.
25. Hercberg S, Castetbon K, Czernichow S, Malon A, Mejean C, Kesse E, et al. The Nutrinet-Santé Study: a web-based prospective study on the relationship between nutrition and health and determinants of dietary patterns and nutritional status. BMC Public Health 2010;10:242 [FREE Full text] [doi: 10.1186/1471-2458-10-242] [Medline: 20459807]
26. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989;13(3):319.
27. Disponible sur. NutriNet-Santé Questionnaire [Internet]. URL: https://info.etude-nutrinet-sante.fr/upload/siteinfo/protected/Quest_Faisabilite_Acceptabilite.pdf [accessed 2025-10-23]
28. Keevil V, Cooper A, Wijndaele K, Luben R, Wareham N, Brage S, et al. Objective sedentary time, moderate-to-vigorous physical activity, and physical capability in a British cohort. Med Sci Sports Exerc 2016;48(3):421-429 [FREE Full text] [doi: 10.1249/MSS.0000000000000785] [Medline: 26501232]
29. Clemes S, Deans N. Presence and duration of reactivity to pedometers in adults. Med Sci Sports Exerc 2012;44(6):1097-1101 [FREE Full text] [doi: 10.1249/MSS.0b013e318242a377] [Medline: 22595985]
30. Edwardson CL, Winkler EA, Bodicoat DH, Yates T, Davies MJ, Dunstan DW, et al. Considerations when using the activPAL monitor in field-based research with adult populations. J Sport Health Sci 2017;6(2):162-178 [FREE Full text] [doi: 10.1016/j.jshs.2016.02.002] [Medline: 30356601]
31. van Hees VT, Renström F, Wright A, Gradmark A, Catt M, Chen KY, et al. Estimation of daily energy expenditure in pregnant and non-pregnant women using a wrist-worn tri-axial accelerometer. PLoS One 2011;6(7):e22922 [FREE Full text] [doi: 10.1371/journal.pone.0022922] [Medline: 21829556]
32. Troiano RP, McClain JJ, Brychta RJ, Chen KY. Evolution of accelerometer methods for physical activity research. Br J Sports Med 2014;48(13):1019-1023 [FREE Full text] [doi: 10.1136/bjsports-2014-093546] [Medline: 24782483]
33. López-Gil JF, Tapia-Serrano MA, Sevil-Serrano J, Sánchez-Miguel PA, García-Hermoso A. Are 24-hour movement recommendations associated with obesity-related indicators in the young population? A meta-analysis. Obesity (Silver Spring) 2023;31(11):2727-2739. [doi: 10.1002/oby.23848] [Medline: 37726964]
34. Riquelme R, Rezende LFM, Marques A, Drenowatz C, Ferrari G. Association between 24-h movement guidelines and cardiometabolic health in Chilean adults. Sci Rep 2022;12(1):5805 [FREE Full text] [doi: 10.1038/s41598-022-09729-1] [Medline: 35388103]
35. Lin L, Liang W, Wang R, Rhodes RE, Liu H. Association of 24-hour movement guideline adherence, mental health and quality of life in young adults: the role of e-Health literacy. Front Public Health 2024;12:1344718 [FREE Full text] [doi: 10.3389/fpubh.2024.1344718] [Medline: 38841676]
36. Janssen I, Clarke AE, Carson V, Chaput J, Giangregorio LM, Kho ME, et al. A systematic review of compositional data analysis studies examining associations between sleep, sedentary behaviour, and physical activity with health outcomes in adults. Appl Physiol Nutr Metab 2020;45(10 (Suppl 2)):S248-S257 [FREE Full text] [doi: 10.1139/apnm-2020-0160] [Medline: 33054342]

XSL•FO
RenderX

37. Rollo S, Antsygina O, Tremblay MS. The whole day matters: understanding 24-hour movement guideline adherence and relationships with health indicators across the lifespan. J Sport Health Sci 2020;9(6):493-510 [FREE Full text] [doi: 10.1016/j.jshs.2020.07.004] [Medline: 32711156]
38. Reid RE, Insogna JA, Carver TE, Comptour AM, Bewski NA, Sciortino C, et al. Validity and reliability of Fitbit activity monitors compared to ActiGraph GT3X+ with female adults in a free-living environment. J Sci Med Sport 2017;20(6):578-582. [doi: 10.1016/j.jsams.2016.10.015] [Medline: 27887786]
39. Carpenter C, Yang C, West D. A comparison of sedentary behavior as measured by the Fitbit and ActivPAL in college students. Int J Environ Res Public Health 2021;18(8):3914 [FREE Full text] [doi: 10.3390/ijerph18083914] [Medline: 33917879]
40. Murakami H, Kawakami R, Nakae S, Nakata Y, Ishikawa-Takata K, Tanaka S, et al. Accuracy of wearable devices for estimating total energy expenditure: comparison with metabolic chamber and doubly labeled water method. JAMA Intern Med 2016;176(5):702-703. [doi: 10.1001/jamainternmed.2016.0152] [Medline: 26999758]
41. de Zambotti M, Goldstone A, Claudatos S, Colrain IM, Baker FC. A validation study of Fitbit Charge 2™ compared with polysomnography in adults. Chronobiol Int 2018;35(4):465-476. [doi: 10.1080/07420528.2017.1413578] [Medline: 29235907]
42. Germini F, Noronha N, Borg Debono V, Abraham Philip B, Pete D, Navarro T, et al. Accuracy and acceptability of wrist-wearable activity-tracking devices: systematic review of the literature. J Med Internet Res 2022;24(1):e30791 [FREE Full text] [doi: 10.2196/30791] [Medline: 35060915]

Abbreviations

PA: physical activity
TAM: Technology Acceptance Model
WEALTH: WEarable sensor Assessment of physicaL and eaTing beHaviors

Edited by A Stone, A Mavragani; submitted 18.04.25; peer-reviewed by O Abasilim, N Peterson, M Liu; comments to author 13.08.25; revised version received 17.10.25; accepted 17.10.25; published 20.11.25.

Please cite as:
Soumaré A, Fezeu LK, Bouchan J, Delestre F, Bellicha A, Cardon G, Donnelly A, Hebestreit A, Touvier M, Oppert JM, Vanhelst J
Three Body-Worn Accelerometers in the French NutriNet-Santé Cohort: Feasibility and Acceptability Study
JMIR Form Res 2025;9:e76167
URL: https://formative.jmir.org/2025/1/e76167
doi:10.2196/76167
PMID:

©Abdouramane Soumaré, Léopold K Fezeu, Jérôme Bouchan, Fabienne Delestre, Alice Bellicha, Greet Cardon, Alan Donnelly, Antje Hebestreit, Mathilde Touvier, Jean-Michel Oppert, Jérémy Vanhelst. Originally published in JMIR Formative Research (https://formative.jmir.org), 20.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Using Human-Centered Design and Development to Create a Digital Sick Day Medication Guidance Application for People With Diabetes, Cardiovascular Disease, or Chronic Kidney Disease: Mixed Methods Study

Julie Babione[1*], MSc; Sarah Gil[1*], BSc; Chidera Okemeziem[1], BSc; Taylor Palechuk[2], MSc; Kaitlyn E Watson[2], BPharm, PhD; Sandra Robertshaw[1]; Nancy Verdin[1], BScOT; Kerry McBrien[3,4], BASc, MPH, MD; David J T Campbell[1,3,5], MD, PhD; Ross Tsuyuki[2,6], MSc, PharmD, ISHF; Neesh Pannu[7], MSc, MD; Matthew James[1,3], MD, PhD; Maoliosa Donald[1,3], BScPT, MSc, PhD

[1]Department of Medicine, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada
[2]EPICORE Centre, Faculty of Medicine and Dentistry, University of Alberta, Edmonton, AB, Canada
[3]Department of Community Health Sciences, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada
[4]Department of Family Medicine, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada
[5]Department of Cardiac Sciences, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada
[6]Department of Pharmacology, Faculty of Medicine and Dentistry, University of Alberta, Edmonton, AB, Canada
[7]Department of Medicine, Faculty of Medicine and Dentistry, University of Alberta, Edmonton, Canada
[*]these authors contributed equally

Corresponding Author:
Maoliosa Donald, BScPT, MSc, PhD
Department of Medicine
Cumming School of Medicine
University of Calgary
3280 Hospital Dr NW
Calgary, AB, T2N 5A1
Canada
Phone: 1 (403) 210 6481
Email: donaldm@ucalgary.ca

Abstract

Background: Diabetes, cardiovascular disease, and chronic kidney disease are associated with high morbidity and costs of care. Medications can reduce long-term complications but may contribute to complications such as hypoglycemia and acute kidney injury during acute illnesses. Sick day medication guidance (SDMG) could help prevent these adverse events, but evidence for effective strategies to deliver this guidance is lacking.

Objective: We iteratively designed and developed a digital prototype user interface (UI) to deliver SDMG for patient self-management. The application, called "*P*reventing medication complications during *AcU*te illness through *S*ymptom *E*valuation and sick day guidance" (PAUSE), delivers personalized knowledge and self-management guidance directly to patients to enhance medication self-management during acute illness, with the goal of reducing preventable emergency visits and hospitalizations and improving patient outcomes during acute illness.

Methods: Using a human-centered design (HCD) approach, we conducted iterative heuristic evaluation and usability testing paired with prototype revisions. Heuristic evaluation involved our team members evaluating the prototype's UI against established criteria. We also conducted formative usability testing with 6 patients (including a patient-caregiver dyad) to provide subjective lived experience perspectives. We analyzed data deductively and pragmatically to rapidly inform subsequent iterations.

Results: We identified 21 and 44 design issues through heuristics evaluation and usability testing, respectively. The development team iteratively revised the PAUSE UI prototype between evaluations, with the final design providing key user flows and integrated supports and reminders for acting on severe acute illness situations that recommend pausing certain medications.

XSL•FO
RenderX

Conclusions: Using an iterative HCD approach, we designed and developed a digital health application to deliver SDMG for patient self-management. We addressed feasible technical and workflow barriers using iterative heuristic evaluations and usability testing resulting in a refined SDMG self-management prototype app for patients taking medications commonly used to treat diabetes, cardiovascular disease, and chronic kidney disease. Further research is needed to test the effectiveness of the current PAUSE app in helping people with these chronic conditions self-manage their medications during acute illness and evaluate the feasibility of integrating the app into community-based chronic disease care.

(JMIR Form Res 2025;9:e77240) doi:10.2196/77240

KEYWORDS

sick day medication guidance; digital health; human-centred design; usability; diabetes; cardiovascular disease; chronic kidney disease

Introduction

Diabetes, cardiovascular disease, and chronic kidney disease (CKD) are associated with high morbidity and health care costs. Effective medications are available to reduce the incidence of long-term complications; however, these same medications are associated with risks of hypoglycemia and acute kidney injury during acute illnesses, which may lead to preventable emergency room visits and hospitalizations [1-3]. These drug-related adverse events are also associated with poor outcomes including extended hospital stays and death [4]. Several international organizations recommend pausing specific medications (ie, those referred to in the SADMANS acronym: sulfonylureas, angiotensin-converting enzyme inhibitors, diuretics, metformin, angiotensin receptor blockers, nonsteroidal anti-inflammatory drugs, and sodium-glucose cotransporter 2 inhibitors) during acute illnesses [5-10]. While some pharmacists and physicians may inform patients of these recommendations, guidance delivery to patients is inconsistent, and many patients remain unaware of it [11,12]. Furthermore, despite the inclusion of sick day medication in clinical practice guidelines [13,14], evidence of effective strategies to deliver it is lacking [5,15,16].

Our previous work explored the barriers and enablers involved in providing sick day medication guidance (SDMG) to patients and found that health care providers (HCPs) often lack SDMG knowledge or confidence in providing SDMG, or do not always remember to provide SDMG [12]. On the other side, patients are often unconcerned with sick day complications or confuse sick day symptoms with chronic condition symptoms or expected side effects of their medications. SDMG guidance was also identified as something that needs to be provided when a sick day event occurs, rather than when medications are first dispensed, because patients are unlikely to remember something they were told months or years ago. Further complicating the issue, patients display a range of behaviors in medication adherence, for example, fear of medication reinitiation or stopping a medication at all [12].

If thoughtfully and rigorously designed, digital tools can offer a new and timely approach to addressing the sick day self-management needs of patients on medications qualifying for SDMG. Digital tools for SDMG should be extremely easy to use, support various levels of patient capacity, and provide the right information at the right moment and in an easily understood manner. These tools can also be integrated into existing electronic medical record (EMR) or pharmacy systems for personalized patient experiences that capitalize on the existing trustworthiness of those systems and thus addressing some of the barriers identified in the preceding paragraphs. Digital SDMG tools might also benefit from integration with wearable technologies, where smartwatch heart rate monitors, continuous glucose monitors, smart scales, and other devices could in future provide key metrics to seamlessly and passively detect a sick day for patients on certain medications and provide timely medication support [17]. To the best of our knowledge, no digital tools have been created to date that support patients with SDMG.

To further address the complex needs of patients needing SDMG as we carefully designed this digital solution, we looked to human-centered design (HCD) for guidance. HCD is a design, development, and evaluation approach that puts the user in the center of all design decisions and ensures that information is provided as clearly and concisely as possible [18-20]. In the digital space, it includes a resource-effective approach that ensures that a user interface (UI) prototype (ie, a rudimentary, interactive representation of a product's design and functionality allowing stakeholders and users to test and validate ideas before the final product is developed) is designed with representative end user involvement and evaluated using heuristic evaluation (ie, examining a UI against established heuristic principles to identify potential problems) and usability testing (ie, observing real users interacting with the interface to identify actual usability issues) with representative end users, and revised before full product development begins [21-23].

The objective of this work was to iteratively design and refine a UI prototype to deliver SDMG for patient self-management using HCD approaches. Here, we describe the "*P*reventing medication complications during *AcU*te illness through *S*ymptom *E*valuation and sick day guidance" (PAUSE) mobile app prototype, developed as part of an overarching research program with the goal of reducing preventable emergency visits and hospitalizations, and improve outcomes during acute illness for people living with diabetes, cardiovascular disease, and CKD.

Methods

Human-Centered Design

A HCD approach informed our prior work addressing the goals of "Understand Current State," "Define," and "Ideate and Design" and have been reported elsewhere [5,11,24]. Here, we

XSL•FO
RenderX

focus on the prototype's iterative UI design and development process, specifically reporting on the heuristic evaluation and formative usability testing within "Evaluate and Refine" goal (Figure 1). These methods aimed to gather qualitative insight into the usefulness, usability, and user satisfaction of the PAUSE prototype, and provide recommendations for improving its usability, accessibility, and design.

Figure 1. Human-centered design approach overview with methods used in this project.

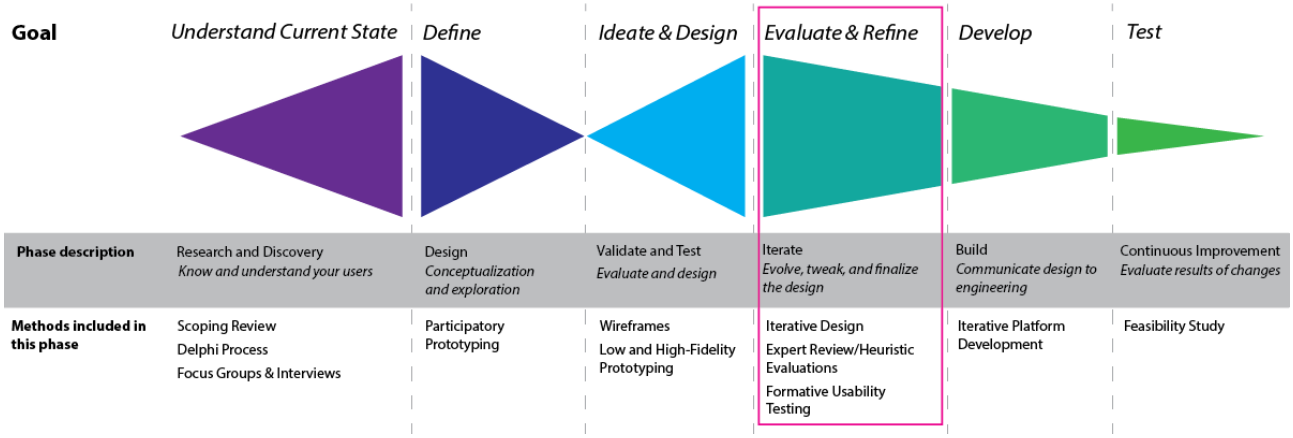

Heuristic Evaluation

We conducted a heuristic evaluation (also known as an expert inspection or evaluation) using an early (ie, rudimentary click-through version of the UI) prototype. The evaluation consisted of applying a set of standardized design principles (ie, heuristics), considering a severity rating of the issue, and recommending potential solutions [25,26]. These evaluations typically reveal most usability-based design problems at a relatively low-resource cost [27]. We used a standardized heuristic evaluation checklist and severity rating guide [25,26,28,29].

The heuristic checklist comprises a set of "common" usability issues inherent in many digital and physical designs [25,30] and gathered by usability engineering experts to facilitate first-pass problems before the design is put in front of representative users as part of later usability testing after making revisions in response to the heuristic evaluations. This prevents common issues from stalling users prematurely during usability testing and allows them to focus more on content and navigation-based issues that developers and designers might otherwise overlook (eg, due to the mere exposure effect—focusing on specific tasks, leading to a decrease in attention to other app elements) [31-33].

We used the severity rating scale shown in Table 1. The severity ratings help evaluators identify which issues are more or less likely to impact elements of usability (as captured in heuristic evaluations) and allow for prioritization of those elements most critical to product success [29].

We followed the pragmatic Nielsen and Norman Group's guidance for conducting heuristic evaluations for this evaluation [28]. Five research team members (SG, JB, TP, DO, and SR) with usability and design expertise or lived-experience expertise (SR) individually and independently examined the prototype interface on various laptops, desktops, and mobile configurations to conduct this heuristic evaluation.

Table 1. Severity rating scale.

Rating	Meaning
1	Violates heuristic; will lead to errors
2	Violates heuristic; might lead to errors
3	Partial adherence to heuristic; will require training to learn
4	Partial adherence to heuristic; will require practice to learn
5	Complete adherence to heuristic; no problems anticipated

Heuristic Evaluation Data Analysis

We followed the Nielsen and Norman Group's guidance for conducting heuristic evaluations for data analysis [28]. Once independent evaluations were completed, the team met to discuss findings item by item, demonstrating issues on screen not seen by others, or where there was discrepancy in severity ratings. Two evaluators (SG and JB) then consolidated the results into a final list of the issues and group discussion notes were consolidated with group-consensus derived severity ratings. We then organized the consolidated list from highest priority to lowest priority (according to the research and design team) with proposed actionable solutions (where possible) and shared with the development team to revise the prototype UI.

Usability Testing

We conducted formative usability testing using semistructured one-on-one interviews to identify the most likely barriers or challenges to using the PAUSE app as experienced by target end users [27,34]. We identified a sample of target end users

XSL•FO
RenderX

who are living with a chronic illness and taking SDMG-qualifying medications and their caregivers through clinical staff champions who provided prospective participants with study invitation letters. Participants then completed the web-based consent forms and a demographic survey and provided contact information for further scheduling. Participants were selected to obtain maximum variation in demographic characteristics (eg, age, gender, location, and eligible conditions). The target sample size for usability testing was set at 6-8 participants in accordance with qualitative or formative usability testing guidelines, where approximately 80% of the issues can be identified in a cost and resource-effective way [34-36]. Testing sessions began with a brief presession interview aimed at collecting basic descriptive demographic information about the study participants and as a way of easing them into the interview process. This was followed by a series of predetermined representative and likely tasks that users would undertake with the prototype app (Multimedia Appendix 1), including onboarding, recording an illness and answering key questions, and responding to reminders to check in. Sessions concluded with a postsession interview with each participant, allowing them to provide subjective perspectives of their experience. Sessions were conducted in participant's homes on their mobile phones, or when technical challenges were encountered on our study laptop where we had the mobile phone mock-up screens loaded with touch screen interaction simulations. Data were collected in the form of audio recordings and field notes.

Usability Testing Data Analysis

Audio recordings of sessions were transcribed and added to field notes that were together analyzed using a deductive content analysis approach [37,38]. This deductive analysis approach was selected to match the practical needs of the collaborating software development team, allowing for rapid synthesis and communication of results for further iteration. We chose the predetermined categories: SDMG content, layout and presentation, interactivity, content (other), functionality and user experience, navigation and workflow, visual elements, medical terminology, and target audience. These categories informed different elements of changes needed for development to move forward. NVivo qualitative analysis software (version 14; Lumivero) was used to facilitate data management and analysis. To include our patient partners as coanalysts and accommodate the dual purpose of usability testing data (ie, academic publication and practical use in rapid turnaround development feedback), coding was completed somewhat unconventionally, although with rigor. A single coder (JB) with access to NVivo coded the transcripts according to smallest units of meaning and initially organized them into the predetermined categories as listed previously. Within these categories, issues were inductively identified (ie, if issues were repeatedly coded within a transcript or between transcripts, they were grouped together). In parallel, to help patient partners and others who did not have access to NVivo identify issues using easier to understand language, we created an Excel sheet with similar headings, which the other research team members used to organize quotes they found in the transcripts to capture issues: (1) ease of understanding (purpose of the app, context the app is used in, and perceptions of app rationale), (2) on-screen content and workflows (anything to do with the UI and content of symptoms listed), (3) performance and efficiency (functionality, How complex the app is to use? Is the app cumbersome? Is the app easy to navigate?), (4) SDMG and adoption (Will they use the advice given? Is it credible? Would they use the app for managing their health?), and (5) parking lot (anything deemed important, interesting, or relevant but unsure how to categorize).

The entire analysis team then met and discussed findings and decided on prioritization and suggested solutions, with JB live cross-checking to confirm analysis alignment between NVivo and Excel coding. The report to the development team was then written based on both analyses with the consensus priorities and recommended solutions, with this manuscript based on that full report.

Ethical Considerations

Ethical approval for this study was obtained from the University of Calgary Conjoint Health Research Ethics Board (REB20-1989). Informed consent was obtained at all stages involving human subjects and the ability for participants to opt out were provided prior to deidentification and analysis. Any protocol modifications were submitted to the ethics review committee and enacted only after approved. Privacy and confidentiality were maintained throughout the study by all research team members (eg, interview transcripts removed any identifying information and any identifying information used for session booking was kept separated from study materials at all times). Only deidentified data were used and data were accessible only to the research team. Usability testing participants were compensated with a CAD $25 (approximately US $18.50) gift card. No conflicts of interest were declared.

Results

Heuristic Evaluation

A total of 21 design issues were documented and are summarized in Multimedia Appendix 2, with some highlights described in Table 2 and the accompanying text description. A full description of all the results is provided in Multimedia Appendix 3.

The analysis of the heuristic evaluation results was practical and focused on real-world application. The categories used in the analysis were decided in advance to facilitate communication with the app developers as these categories would be familiar to them. We provide a summary of the results in Table 2.

XSL•FO
RenderX

Table 2. Heuristic evaluation results showcasing the issues identified in the first iteration prototype UI[a] design.[b]

Issue	Recommendation
Navigation and workflow	
Workflow is incorrect in multiple areas (1): Users experience navigation issues, as some selections do not link to the correct page, selections were preselected, and buttons did not function correctly.	Remove all preselected options, ensure that all pages advance along the correct pathway, verify that the back button functions as users expect, and ensure that users can access the right pages at the right time.
Always end up on the onboarding screen (3): After the onboarding process is completed when users open the app or press the home button, users return to the onboarding page and need to sign in again.	After completing onboarding, users should open the app directly to the home page without needing to sign in each time or view the rotating banner.
No history or archive of advice (1): There is no way for users to revisit past advice, and the information is not saved requiring users to reenter their details. Users wish to share recommendations with their health care providers, 911/811 operators (ie, emergency dispatch), or urgent care doctors without relying solely on memory.	Create an archive section in the app where users can revisit their symptoms and the advice given.
Onboarding process is disjointed and confusing (3): Alignment in lists is off, and there is no consistency across the screens. Specifically, screen 1 of 5 lacks organization for the home-monitoring options.	Organize options into categories, for example, “Screen 1 of 5” should have options grouped according to body areas. Ensure that the app has a consistent look and feel.
Layout and presentation	
Remove resources from the check-in page (2): Users found the option to read resources when checking in distracting. The app must be clear and concise, as users will access it when unwell.	Delete resources from the check-in page and create a dedicated resources page that users can access at their convenience. Overall, to improve clarity and user experience, remove unnecessary information.
App has dead ends and lacks user feedback (2): At the end of the onboarding process, there is no indication that it is complete. When users receive advice from the app, there is no option to exit the screen, no “X” button or “OK” button.	Avoid dead ends; users must be able to exit pages and return to the home page or previous page. User feedback is needed so that users know when they have completed a task. Suggest showing a “Thank You” message, providing a “Home” or “OK” button, or offering an option to close the page.
Important information lost in blocks of text (3): Important information is overlooked when presented in large blocks of text.	Important information or calls to action should be separated from large blocks of text by placing them on their own line, without using brackets. Ensure that text formatting and presentation of information are consistent across all screens.
Visual elements	
Confusion regarding icons in the navigation bar (3): Users are confused about the icons in the bottom navigation bar and are uncertain about their destination within the app.	Enhance the clarity of the icons by using more familiar designs. In addition, consider adding a “tooltip” resource to explain the icons to users.
Red and green color blindness issue (3): The app uses red and green colors to highlight decisions and important areas. People who are red-green color-blind may face challenges.	Update the branding and colors to match the PC health app. Ensure that visuals are color-blind–friendly.
Functionality and user experience	
Rotating banner difficult to read (3): Text on the sign in page continuously scrolls, and users cannot pause to read the information. The banner includes visual indicators (blue circles) that suggest that users can interact with it.	Remove the scrolling banner and add “Next” buttons to allow users to read the information at their own pace. Eliminate circles at the bottom of the page and replace with the “Next” button. Do not show the banner every time a user logs in; instead, place that information in the resource section.
Communication choices during onboarding (3): When selecting notification preferences, users like the option to choose both text and email notifications. Users would prefer to finish setting up their profile before they answered the communication question.	Allow users to choose both text and email or add an option for “text and email.” Move the notification options screen to the end of the onboarding process.
Promote user engagement (2): Language used in the app does not encourage user engagement.	Create messaging that encourages user engagement. Instead of saying, “Contact your healthcare provider if any symptoms arise,” rephrase it to: “Check in the next time you’re feeling worse than usual.” Similarly, change “Day 1 Follow-up” to: “Tap here to check-in.”

XSL•FO
RenderX

Issue	Recommendation
Ensure app is tailored to user (3): During onboarding, users are asked about medications they take and whether they have home monitoring devices. When users enter their symptoms, they are asked the same questions again. Example, user is asked whether they take insulin at onboarding and then given advice that says, "if you take insulin."	After users enter their symptoms, they should not be asked what home devices they use; instead, they should be asked questions about their device measurements. For questions regarding ketones, ensure that the inquiries reflect the type of ketones users measure, or provide an "n/a" option. Users should be directed to the insulin advice page only if they indicate that they take insulin. Also suggest removing the phrase "if you take insulin."
Design of the "My medications" page cluttered and confusing (3): Layout of the "My medication page" is challenging to read due to the amount of information. It is also not clear whether users edit individual medications or the entire page.	Medications should be listed clearly, and the "verified" label should be removed from public view. Each medication listed should include its own edit icon.
Interactivity	
Users can turn all buttons on or off (3): When users respond to questions using "Yes" or "No" buttons, they can click on both responses, even if the question specifies that only 1 answer is allowed.	When a response requires only 1 choice, ensure that only 1 option can be selected.
Medical terminology and SDMG[c] details	
Users are confused about medical advice for symptoms (3): Users feel confused about why symptoms come with varying medical advice. They do not understand the reasons for contacting an HCP[d], what "sick day" medications are, or why they should pause certain medications.	Suggest creating a FAQ[e] resource that explains why users need to contact their HCP, what constitutes a sick day, and why these medications should be paused.
Missing information to provide proper NSAIDs[f] advice (3): During onboarding (screen 4/5), users are asked whether they take NSAIDs but not how frequently they take them.	Add a question asking users how frequently they take NSAIDs (never, rarely, occasionally, and regularly). Provide proper advice for the frequency the user selects (never, rarely, and occasionally do not receive SDMG but regularly does).
Language	
"Educational Resources" as a title adds pressure (3): The phrase "Educational Resources" can be paternalistic and create a sense of top-down pressure. This suggests that the user needs to be educated or they lack knowledge about their medications or illness.	Change the heading to "Resources" to reduce top-down pressure, as this term indicates that users can access additional information as needed.
Consistency	
Design of medication advice page (2): On the "Medications to Stop" page, the display of listed medications resembles selection buttons.	Ensure that the medications are listed in a way that does not suggest that they can be selected. Maintain consistency in how lists and selection options are displayed throughout the app.
Display of the user's name varies across screens (3): The user's name is displayed on some screens and not on others.	Only display the user's name on their profile page, and check-in and follow-up screens.
App design lacks consistency (1): There is a lack of consistency throughout the app, including page designs, button appearances, navigation between pages, and the use of icons.	Review all content for any spelling and grammatical errors, and make sure that design elements, including buttons and selection options, remain consistent throughout the app.

[a]UI: user interface.

[b]The number in brackets behind each issue is the consolidated severity rating.

[c]SDMG: sick day medication guidance.

[d]HCP: health care provider.

[e]FAQ: frequently asked question.

[f]NSAIDs: nonsteroidal anti-inflammatory drugs.

The heuristic evaluation identified most issues in the categories of navigation and workflow (19%), Functionality and user experience (24%) primarily related to UI layout and interaction. Evaluators assessed the app from the perspective of a patient experiencing a relevant sick day, focusing on heuristic elements such as memory recall, consistency, and clarity of language. To support urgent care scenarios, evaluators recommended ensuring that prior advice, such as recommendations to contact emergency or urgent care assistance, could be retrieved and shared with care providers to aid the urgent care process. A notable issue was inconsistent visual design, where similar-looking buttons and visuals made it unclear what was interactive. Given the importance of seamless screen flow in delivering the correct SDMG, evaluators identified several issues for resolution before patient testing. Evaluators recommended additional adjustments to improve usability including removing distracting elements such as additional resources from the assessment workflow to prevent errors and organizing screen content more clearly to avoid confusion.

XSL•FO
RenderX

Evaluators also reviewed language, terminology, and tone to ensure a supportive and user-friendly experience. Suggested changes included (1) removing the term "Educational" from "Educational resources" as this could be perceived as paternalistic, (2) ensuring consistency between the patient's profile-building questions (which may not be needed if this information is imported from an existing EMR or pharmacy platform), and (3) refining the prompt language to encourage users to check in with the app when they are experiencing manageable sick day symptoms.

Usability Testing

A total of 6 participants (5 patients and a patient-caregiver dyad) who spoke and read English completed the usability testing (Table 3). The participants were older than 40 years, educated, and with a 60%/40% split between males and females. An eHEALS (eHealth Literacy Scale) mean score of 30.4 (SD 6.8) indicates a moderate to high level of perceived eHealth literacy, meaning individuals generally feel confident in finding, evaluating, and using web-based health information [39, 40]. A mean System Usability Scale score of 74.2 (SD 9.2) with median score of 73.8 suggests an above average perceived usability of the prototype [41, 42]. The results indicated a highly positive receptivity and anticipated usefulness of the PAUSE app; however, the app's purpose and target audience needed to be refined for optimal effectiveness.

A total of 44 priority-ranked issues were identified to be addressed and resolved prior to software development (Multimedia Appendix 4). A full description of all results is shown in Multimedia Appendix 5. Most issues related to navigation and workflow and interactivity were resolved in the design iteration prior to usability testing, allowing the focus to shift to issues identified by patients and the caregiver with patient dyad with lived experience. Table 4 summarizes the usability testing results, and Multimedia Appendix 3 provides the full results with screen captures to provide further details.

Usability testing identified issues in navigation and workflow, layout and presentation, visual elements, functionality and user experience, interactivity, medical terminology, SDMG content, (general) content, (confusions around) target audience, and in a few cases branding. The research team discussed and prioritized issues and then discussed issues again with the development team to identify appropriate solutions.

We noted a few issues related to navigation and workflow as residual issues identified previously through heuristic evaluations; however, usability testing participants identified new issues as well, specifically around the app's contextual workflow. Questions arose concerning participants' previous experiences with clinician reluctance to discuss medication dosage changes over the phone, prompting the research team to confirm existing assumptions that this capability exists in most local community settings. One participant also noted that some at-home sick day situation management approaches require some preparation and planning, which a checklist included within the app's resources section might facilitate. Another participant asked whether the app might inform their doctor about an ongoing sick day situation. This was deemed potentially complex and might be better for a future iteration.

Within layout and presentation, participants using smaller mobile phones were unable to see the "next" button and there were no suggestions that there was additional content further down (out of sight). A few participants also tried to use the app in landscape mode, which we noted was beyond the capability of the prototype UI; however, this will be something that should be monitored or even addressed (if possible) during app development. On the same topic of screen orientation, the narrow portrait orientation made lengthy text difficult to read. Finally, participants were challenged to identify medications based only on names, asked for pill pictures, and expressed concern about the absence of dosage information.

Several users also mentioned visual elements, with emphasis on accessibility, text size and contrast, and familiar connotations (ie, red signifies alert, danger, or stop) they found confusing in this app. Several participants also found a few inconsistent color uses within the apps (ie, for buttons) to be addressed.

For proposed functionality and user experience, the prototype UI included face and fingerprint recognition as might be included for privacy and security through a modern smartphone. However, users shared challenges and frustrations with or lack of knowledge of these elements. Others found manual medication data entry cumbersome and error prone. A potential solution to this might be to pull medication data from an existing EMR or pharmacy database.

Participants identified a few issues that we classified under interactivity. Form factor preference among participants came up, with smartphones still being somewhat less familiar among older participants than computers. The smaller size of the smartphone also made screen interactions difficult for several users (particularly those with diabetes or age-related vision problems), with checkboxes and "X" (close) buttons difficult to select. Form factor will be something to consider depending on future EMR or pharmacy platform integration plans.

Medical terminology and, most notably, language related to SDMG content came to the forefront during usability testing. For example, patients did not always easily understand content created by medical experts. In some cases, we removed elements such as medication frequency changes for over-the-counter nonsteroidal anti-inflammatory drugs because of their use complexity in real-life situations (eg, daily or occasional use that varies from situation to situation) and the impact of the question in the assessment of symptom severity. Ketones were unfamiliar to those who do not explicitly track them, which should perhaps be shown only to those who do track them. A few users found "Cope with symptoms" confusing, which researchers revised to "Are you able to manage these symptoms by yourself?" The research team collected further contextual feedback that will help position the PAUSE app for pilot testing and implementation. Some users believed that as long as they did not access their health data themselves through secure online web portals that their health data would not be online, leading them to be wary of using digital technologies to access their data. Other complexities might also arise if a patient has medications filled at different pharmacies who might provide different SDMG information.

XSL•FO
RenderX

Table 3. Usability testing demographics.

Characteristics	Values, n (%) or mean (SD)
Patients (n=5)	
Sex, n (%)	
Male	3 (60.0)
Female	2 (40.0)
Gender, n (%)	
Cisgender woman	1 (20.0)
Cisgender man	2 (40.0)
Prefer not to answer	2 (40.0)
Age (years), n (%)	
41-50	1 (20.0)
51-60	1 (20.0)
61-70	1 (20.0)
70	2 (40.0)
Race or ethnicity, n (%)	
White	4 (80.0)
Other	1 (20.0)
Population size, n (%)	
<500,000	3 (60.0)
>500,000	2 (40.0)
Education, n (%)	
Less than university degree	1 (20.0)
University degree or higher	4 (80.0)
Income, n (%)	
<$100,000[a]	2 (40.0)
>$100,000	3 (60.0)
Employment status, n (%)	
Full-time employee or student	2 (40.0)
Retired	3 (60.0)
Marital status, n (%)	
Single	1 (20.0)
Married or common law	4 (80.0)
Chronic condition, n (%)	
Chronic kidney disease	3 (60.0)
Heart failure	1 (20.0)
Diabetes	2 (40.0)
Other	1 (20.0)
eHEALS[b] score, mean (SD)	30.4 (6.8)
SUS[c] total, mean (SD)	5.6 (1.1)
Caregivers (n=1)	
Sex, n (%)	
Female	1 (100.0)

XSL•FO
RenderX

Characteristics	Values, n (%) or mean (SD)
Gender, n (%)	
Cisgender woman	1 (100.0)
Age (years), n (%)	
70	1 (100.0)
Race or ethnicity, n (%)	
White	1 (100.0)
Population size, n (%)	
>500,000	1 (100.0)
Education, n (%)	
University degree or higher	1 (100.0)
Income, n (%)	
>$100,000	1 (100.0)
Employment status, n (%)	
Retired	1 (100.0)
Marital status, n (%)	
Married or common law	1 (100.0)
Chronic condition, n (%)	
Chronic kidney disease	1 (100.0)
Diabetes	1 (100.0)
Other	1 (100.0)
eHEALS score	35 (undef[d])
SUS total	5 (undef)

[a]All dollars are in CAD. Study data were mostly gathered in Summer 2023 and so a currency conversion rate from August 16, 2023 (CAD/1.3518455345177844/0.7397294842244747) is used.

[b]eHEALS: eHealth Literacy Scale.

[c]SUS: System Usability Scale.

[d]Undef: undefined.

XSL•FO
RenderX

Table 4. Usability testing results.

Issue	Recommendation
Navigation and workflow	
Page 4 of 5 (onboarding) does not start at the top of the page: Page 4/5 during onboarding loads with the middle section displayed with no scroll bar.	Ensure that all pages load correctly.
Missing "Next" button leads to unexpected and surprising app behavior: Some pages have "next" buttons and others do not.	Ensure that all pages with steps in a workflow include the "Next" button.
Ensure that notifications do not go to junk mail: Users noted from previous experiences that notification emails sometimes end up in their junk folder, but they may not always think to check there.	After onboarding, a notification should appear prompting users to check their junk folder if they have not received the email.
Fear that a clinician would not discuss information over the phone: Many users lack confidence that their HCP[a] will offer advice over the phone and are unsure about the type of guidance they will receive. The example was specific to insulin adjustments.	If users are advised to call their HCPs in any situation, the app advises them that this type of guidance can be given over the phone. HCPs should also remind patients that HCPs can communicate dosage changes over the phone.
Users want to take proactive steps to prepare for sick days: Users find the self-management advice helpful but can access this advice only during a sick day event. They would like to proactively prepare for a sick day.	Enhance the resource page by emphasizing self-management tips for mild to moderate symptoms. Make archived advice more noticeable.
Users want their HCP aware of sick days and advice provided: Users are concerned that their HCPs will be unaware of a sick day occurrence and what advice was given.	The research team will review this issue, as general practitioners expressed reluctance to interact with notifications from the app and follow up with their patients.
Layout and presentation	
Text alignment and item grouping vary: The alignment and configuration of the conditions and topics during the onboarding process impede users from easily scanning the options.	Organize the conditions either alphabetically or by their relatedness. In addition, ensure that the lists are left-justified and convert the items into checkboxes instead of buttons.
Layout and design present issues on small phones: Users with small phones could not see the "next" button or text below the "fold." Screens have no indication to scroll.	Add a scroll bar to the side of the screen to indicate that scrolling is required. Ensure that key elements are at least partially visible on the "fold."
Ensure responsiveness to changing aspect ratio: When users turned their phone to landscape, potential layout issues arose. The prototype was not able to respond to orientation build, so this was identified as something to monitor during development.	Ensure that the app is responsive to changing device orientation, and the layout is usable and esthetically sound for all orientations.
Narrow portrait screen makes options difficult to read: Some content is difficult to read in a narrow portrait layout, especially when lists are not organized in descending order.	Ensure that text is succinct and left-justify lists with 1 item per row to support scanning.
Include medication doses and photographs in the medication list: Users expressed challenges in identifying medications solely by their names and raised concerns regarding the absence of dosage information.	Add photographs of medications and ensure that the medication list contains details regarding doses and frequency.
Visual elements	
Icon clarity: Icons used within the prototype were not recognizable by several users.	Use icons sparingly, and icons used should be clear and adhere to WCAG[b] 2.1 accessibility guidelines [43]. To promote a better understanding of functionality, use text instead of icons (eg, "edit" and "add").
Ensure sufficient contrast for text colors: Some text was very light in color and difficult to read.	Ensure sufficient contrast between background and text using the WCAG 2.1 accessibility guidelines.
Use colors and shapes consistently: In certain instances, a dark blue circle served as an indicator for a clickable element, while on other pages, it is used purely for decorative purposes.	Maintain consistent color usage throughout the app to prevent any possible confusion. Refrain from applying buttons or link colors to elements that are not interactive.
Use red sparingly and intentionally: Users assume that red button colors indicate an alert, danger, or stop. Users also find that button colors are inconsistent.	Use red sparingly and intentionally and reserve it for error messages. Ensure that button colors are applied consistently and match the context.
Functionality and user experience	

XSL•FO
RenderX

Issue	Recommendation
Face and fingerprint recognition may be insufficient: Users shared their personal experience with apps and facial recognition not working if they wear glasses. Some users were unsure how to use fingerprint recognition.	Ensure multiple ways to log on to the app and consider removing face and fingerprint recognition altogether.
Concerns around the workload of manual medication entry: Users raised concerns about the process of inputting their medications manually, such as errors and the amount of time it would require.	Have the app retrieve the list of medications from relevant pharmacies, NetCare, or an alternative platform.
Interactivity	
App lacks a "Select all" option: When a list contains numerous options, users prefer the ability to "select all" items at once.	Add a "Select all" option to lists where appropriate.
Comfort with a computer versus a smartphone: Certain users prefer using smartphones for calls and texts and are uncomfortable with touch screens. These users would prefer to work on a computer.	If possible, provide a desktop or web version of the app for users who prefer this option.
Lack of checkboxes for selection: Users find it unclear whether they have selected options when only a color change indicates selection.	Make options into a checklist.
"X" button does not always function: The "X" close button did not always work.	Ensure that interactive elements (such as the "X" button) work as expected.
Medical terminology	
Typos in resources (Diabetes Canada handout): Users have identified typos and formatting issues in the Diabetes Canada handout.	Ensure that all content, resources, and handouts are meticulously checked for typos, grammatical errors, and formatting issues, regardless of their source.
Users forget which blood pressure number is systolic and diastolic: Some users were unsure which blood pressure number is systolic (top) and which is diastolic (bottom). In addition, the units of measurement (mm Hg) can be confusing.	Include "top number" or "bottom number" in brackets whenever the tool mentions diastolic or systolic. Ensure that this is consistent throughout the app. Consider removing units to avoid confusion, still to be discussed.
Users are unsure what NSAIDs[c] are: Users are unsure what an NSAID is, even when provided with examples of the generic and brand names.	Remove the term "NSAID" or use statements such as "Do you take any of these over-the-counter medications (nonsteroidal anti-inflammatory drugs)?"
Users are unsure what ketones are: Some users are unfamiliar with ketones, as they do not track this information. This situation can arise during onboarding if users accidentally select an option they do not monitor at home.	Consider implementing a confirmation page during onboarding to verify that the selection choices accurately reflect what users monitor at home.
SDMG[d] content	
Unsure what "cope with symptoms" means: Some users find the question "Can you cope with managing these symptoms on your own?" difficult to interpret. The term "coping" is complex and vague.	Rephrase the question to be more specific: "Are you able to manage these symptoms by yourself?"
The context of "feeling better" and "feeling worse" in relation to living with chronic diseases is unclear: These phrases are subjective and context-sensitive, particularly for individuals living with chronic diseases, who may never feel "better" compared with their predisease "normal."	Ensure language clarity and avoid vague and ambiguous statements that can be misinterpreted.
Users are confused and startled when told to go to urgent care or call 911: Users felt that they would not be able to comfortably explain to the 811/911 operator or the urgent care doctor why they were calling or at urgent care.	At the beginning of the app, inform users that they may receive advice to go to urgent care or call 811/911 based on symptoms. Provide clear and detailed explanations about why users are being asked to do this. Add a button "Information for Healthcare Providers" on the advice page so that users can share information with the emergency doctor.
Type of pill bottle not specified in medication list: Some users felt that it would be beneficial to have the type of pill bottle listed as they have issues opening the standard type.	Include additional information on the type of pill bottle a medication is dispensed in to ensure that all relevant details are accurately captured.
Limited resource topic areas: Most resources are tailored for people with diabetes and not other conditions. This may lead users to question whether they are the intended audience.	Include resources and information for all the conditions the app addresses. Ensure that the app's splash screen communicates the intended audience.

XSL•FO
RenderX

Issue	Recommendation
Limited medical conditions: Users felt that the guidance was based on conditions rather than medications. They were concerned that their symptoms might be related to a condition or topic not covered.	Consider including additional chronic conditions. Clarify that the focus is on medications, not necessarily specific conditions.
Difficult to answer medication frequency question: Users found it challenging to answer the refill frequency question; not all medications are refilled at the same time.	Remove this question.
Are resources for SDMG? Users were unsure whether resources were specific to SDMG, their condition or health information in general.	Label the resources to indicate whether they pertain specifically to SDMG or are general health information. Consider removing all non-SDMG resources.
Context needed for what each symptom entails: Users have reported that some symptom-related questions reflect their usual condition experiences, leading to confusion. For example, people with blood pressure issues often get lightheaded and people with hypertension can have blood pressure fluctuations as part of their condition.	Clarify that these questions are intended to identify symptoms that are sudden in onset or significantly different from the user's normal experience.
Blood pressure measurement questions assume that users have a home device: Users have noted that the questions about blood pressure seem to assume that they can measure their blood pressure at home.	Adjust the question to include, "If you have a blood pressure device at home." In addition, provide an "N/A" option.
Content	
Users misunderstood the term "communication preference": Some users thought that the term "communication preference" was for language preference (eg, French or English) rather than the method (eg, text or email).	Replace "communication preference" with "notifications" in the user profile. List the options as "email," "text," and "email and text." During onboarding, remove the term "communication preference" and replace it with, "Please let us know how you prefer to be notified."
Users need reminders of their previous symptoms: Users have noted that recalling previous symptoms can be challenging, which makes it difficult to answer questions or interact with the app effectively.	Include a summary of the user's last check-in or a history of recent symptoms.
Users forgot the purpose of the app: The app lacks introductory information to explain its purpose and what users can expect. Users mentioned that they would use the app only when unwell, which could lead to forgetting its purpose due to infrequent use.	Provide clear and concise information about the app's purpose, target audience, and expected benefits. Offer optional guided tour screens that demonstrate how to use the app, include a brief video and a handout with this information, and provide detailed information on conditions and medications that fall under SDMGs.
Issue with language and terminology used: Some users were unaware of what caffeine is and did not realize that coffee and other drinks contain it.	Add examples in brackets to ensure that users understand what is being referred to. For instance, "coffee (or tea, coke, or instant flavored coffee) contains caffeine." Provide clear definitions and examples to enhance clarity and understanding.
Target audience	
Users with vision issues had difficulty using the app: Several users found the text small, which could be problematic for users with vision issues, a common concern for the older adults and people living with diabetes.	Increase the text size and provide options for text resizing within the app. Consider implementing a high-contrast mode to further enhance readability. Ensure W3C Content Accessibility Guidelines (WCAG 2.1) are adhered to.
Users have concerns regarding data security and privacy: Users misunderstand the implications of downloading and using an app, believing that avoiding online platforms means their data are not stored. Some users expressed concerns about their privacy and how the research team will handle their data.	Provide clear information about how data are stored and managed, emphasizing that data are already stored digitally even if users avoid engaging with online platforms. Include a section in the app that transparently outlines the privacy policy and data usage terms.
Some users use more than 1 pharmacy: Some users use different pharmacies for various medications and are concerned about whether medications obtained from non-PC pharmacies fall under SDMG and how that would be managed.	Ensure that app resources have sections that clearly explain which medications fall under SDMG, regardless of the pharmacy they are obtained from. Let users know that as of now the app is only available for PC Health app users, but SDMG still applies.
Lack of real-life practical language throughout the app: Users were unclear about the app's purpose, the meaning of "sick day medication guidance," and the phrase "feel worse than my normal."	Revise the app's language to be more accessible to nonmedical users. Avoid medical jargon, use plain language, and provide definitions and use examples.
Branding	

XSL•FO
RenderX

Issue	Recommendation
Pause logo was found to be confusing: Users did not recognize the logo as a pill	This was not an issue for all users, so the logo could be left as is or consider redesigning the logo.

[a]HCP: health care provider.

[b]WCAG: Web Content Accessibility Guidelines.

[c]NSAIDs: nonsteroidal anti-inflammatory drugs.

[d]SDMG: sick day medication guidance.

The PAUSE App UI Design

Overview

Figure 2 shows the overall flow of content within the app and user experience for the most current prototype of the PAUSE app.

Figure 2. PAUSE app flow showing how different patient responses result in pausing or resuming medications or seeking urgent care for serious symptoms. Yellow indicates mild symptoms that can be managed at home by pausing specific SADMANS (sulfonylureas, angiotensin-converting enzyme inhibitors, diuretics, metformin, angiotensin receptor blockers, nonsteroidal anti-inflammatory drugs, and sodium-glucose cotransporter 2 inhibitors) medications. Red indicates situations where urgent care should be sought. Green indicates that the patient has no more risky symptoms and can resume their medications. HCP: health care provider; PAUSE: Preventing medication complications during AcUte illness through Symptom Evaluation and sick day guidance; SDMG: sick day medication guidance.

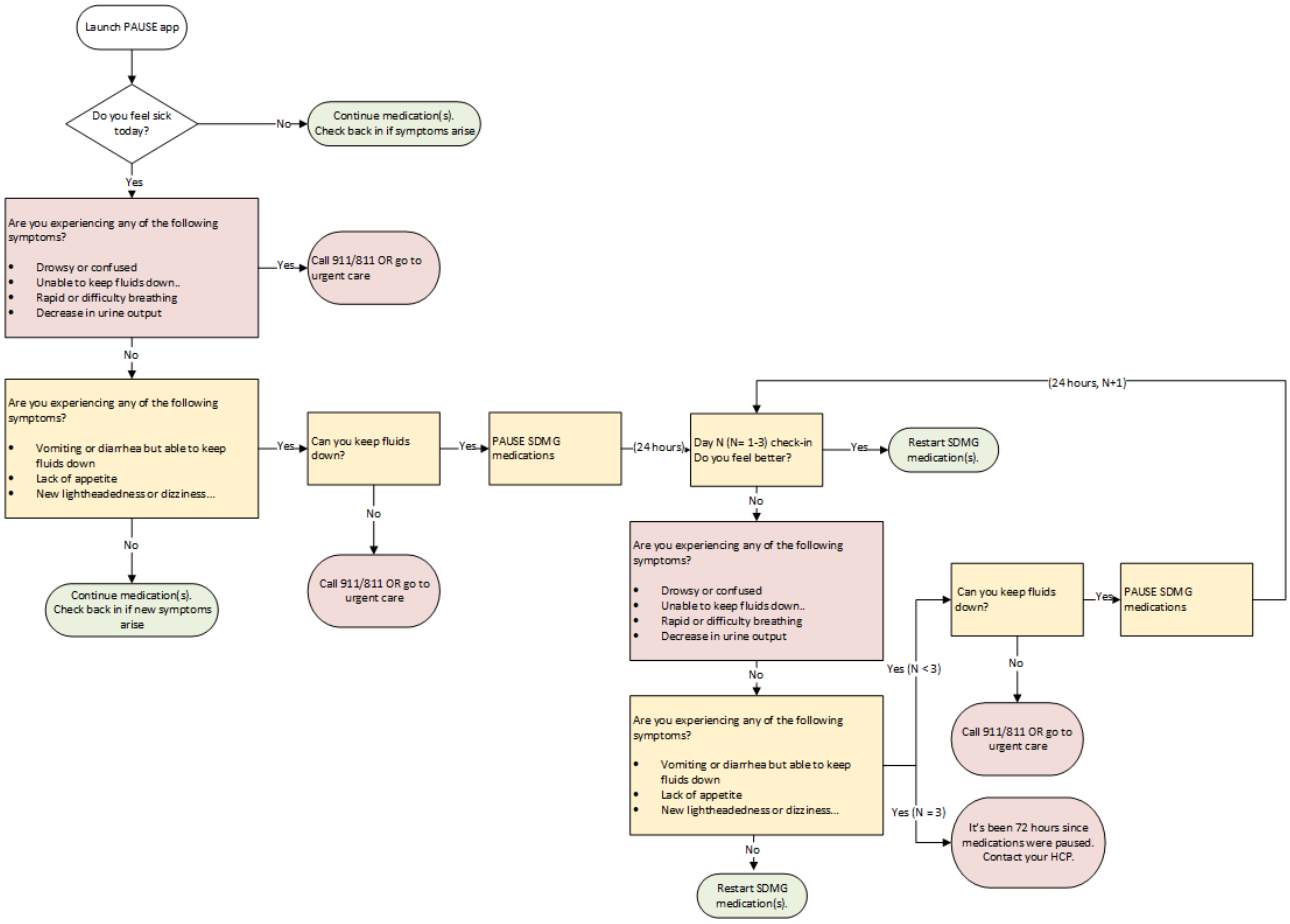

Sick Day Medication Guidance

The PAUSE app UI prototype addresses SDMG in a personalized and time-sensitive manner. The app guides the patient through assessing the severity of their symptoms, appraising their risk of volume depletion, and recommends any temporary medication adjustments to be made. Through push notifications, the app follows up with the patient to reassess symptoms and provide additional recommendations or adjustments. If patient-reported symptoms are severe enough, the app will also recommend that urgent or emergency care be sought. Prompts for resuming medications are provided once qualifying sick day symptoms have resolved. Figure 3 shows key SDMG moments in the PAUSE prototype app UI, including (A) the initial check to see whether the user is feeling unwell, (B) serious and subsequent, and (C) less severe symptoms to be assessed, followed by advice for the user to follow under (D)

XSL•FO
RenderX

mild, (E) moderate, and (F) urgent symptom presentation situations.

The prototype PAUSE app UI assumes that the app will be integrated with an existing patient health record and active medications list; the app will be able to automatically flag qualifying medications that have been prescribed to the patient to be carefully adjusted or temporarily paused to better support the patient during an acute illness.

Figure 3. Screen captures of the prototype Preventing medication complications during AcUte illness through Symptom Evaluation and sick day guidance (PAUSE) app UI at key moments in sick day medication guidance. (A) The app's initial screen shown when launching the app ahead of activating an acute illness tracking and support flow. (B) First screening of symptoms if the user has indicated that they are feeling unwell. Responses to this screen trigger other screens, as shown in the app's flow diagram. (C) Second layer of symptom screening if the patient has indicated that they have any of the previous screen symptoms. (D) Advice provided if the symptoms are deemed mild. (E) Advice to pause specific medications (based on the patient's list of medications) in response to moderate symptoms. (F) Advice to seek urgent care immediately as a result of patient-reported severe symptoms.

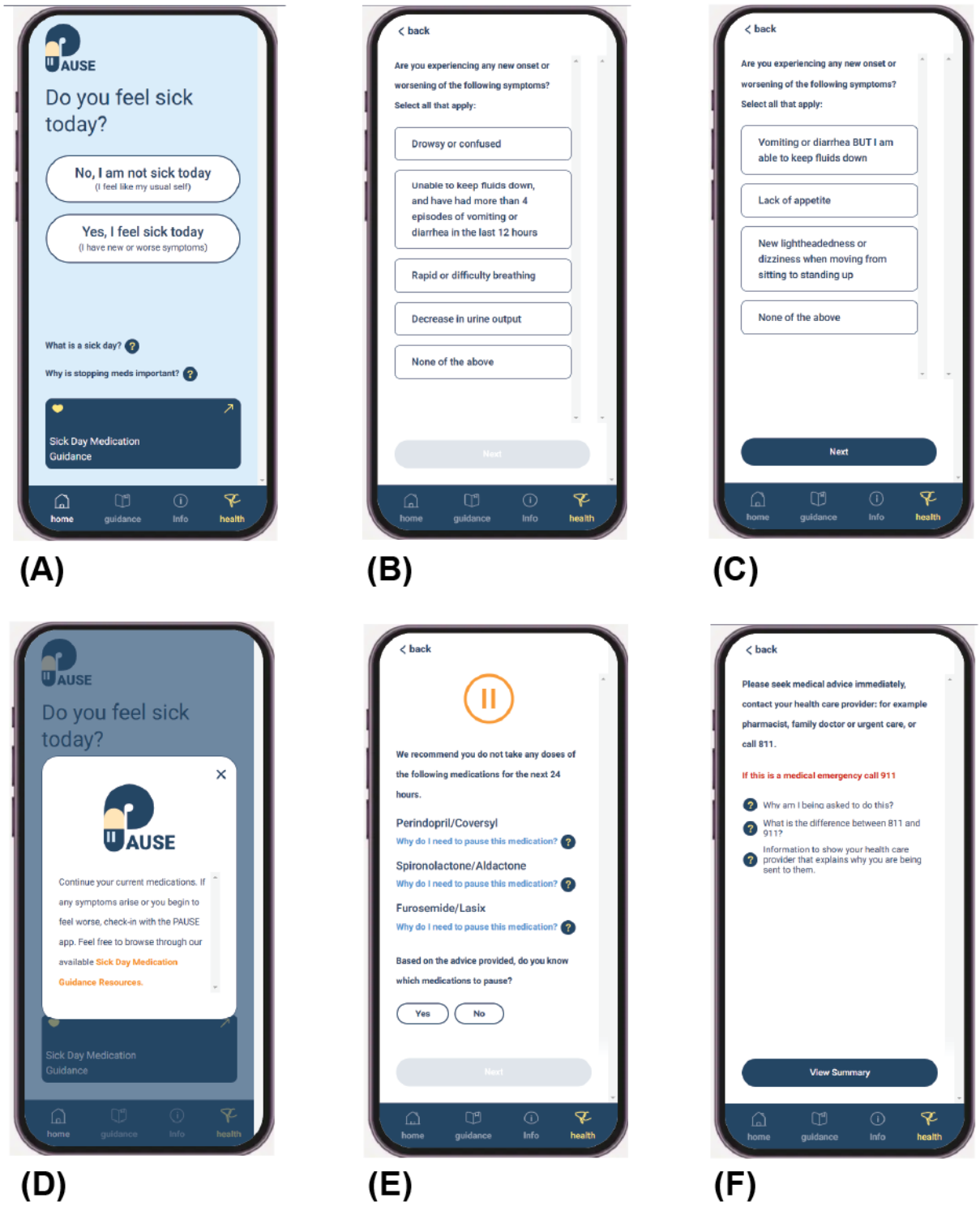

XSL•FO
RenderX

Onboarding

The PAUSE app's user flow designs support the diverse needs of patients (ie, patients with chronic conditions needing SDMG and their caregivers), providing tailored experiences for different groups, including individuals with no diabetes, and those managing diabetes with or without insulin therapy. As referenced in Figure 2, the onboarding user flow gathers information needed to support appropriate and personalized sick day medication adjustments to deliver customized care [44].

Check-in

The check-in user flow is operationalized via push notifications in the days following the patient's initial reporting of symptoms constituting a "sick day" on the app. The prompts check on key symptoms and provide recommendations for how to manage the ongoing acute illness. Figure 2 shows potential outcomes: (1) mild symptoms: continue medication; (2) moderate symptoms: adjust insulin doses and pause specific medications from the SADMANS list, as well as general health advice such as staying hydrated and providing some guidance on hydration; and (3) severe symptoms: seek medical attention urgently.

If the app prompts the patient to seek medical attention, this prompt includes a summary of the information that they are encouraged to share with medical practitioners to support potential symptomatic cognitive impairment, patient knowledge, and awareness, and to promote transparency and trust-building around *why* the app recommends that they urgently seek medical attention.

Other Support Materials

The app also includes additional materials and resources. The home screen's bottom navigation bar is visible throughout the app to help the user be aware of the section. Currently, this section includes the Diabetes Canada SDMG handout to ensure that users would find the materials easily [45]. As the PAUSE prototype is developed into a functional platform, additional materials and guidance can be added to this section.

Discussion

Principal Results

This work showcases the iterative creation and evaluation of an SDMG innovation we previously identified as a key gap in support for SDMG knowledge, awareness, and patient self-management [11]. Through this work, we demonstrate the value of HCD approaches, specifically heuristic evaluation and usability testing, to design a digital SDMG self-management tool that is suited to the specific needs and context of patients requiring SDMG. We designed the resultant PAUSE app UI prototype to support patients to identify and appropriately adjust their medications on sick days.

Comparison With Prior Work

SDMG requires novel solutions for implementation and further evaluation. A study by researchers in Australia found that uptake of SDMG recommendations was poor and that pharmacists recommended laboratory monitoring for potential toxicity but did not provide SDMG despite 71.1% of community-dwelling patient participants using at least one of the SADMANS medications and Kidney Health Australia recommending temporary cessation of these medications during episodes of acute illness for patients with CKD [46]. They concluded that more research is needed to understand barriers to providing SDMG to patients. Another study conducted in The Netherlands found that for both patients with and with no impaired kidney function, most patients continued to take their prescription drugs without adjustment during sick days [47]. This work also noted the challenge created by separating information provided to patients in medical appointments (encouraging consistent adherence) from sick day situations, which happen unexpectedly, and the proper information may not be readily at hand.

SDMG digital tools are an innovation to address sick day needs of patients with diabetes and kidney and cardiovascular conditions who use specific medications that could be harmful during acute illness. Prior research suggests that a digital tool might be an acceptable and feasible strategy to help patients self-manage SADMANS medications during acute illness [48]. Notably, a recent scoping review on sick day management in people with CKD found that a digital intervention would likely be easy to use and an acceptable approach to providing SDMG [16]. Our team previously conducted focus groups and found that both patients and health care practitioner participants supported the concept of a digital tool for SDMG [12]. We have also explored barriers and enablers for both patients and HCPs through focus groups and identified built-in prompts for prescribing and dispensing software as one potential solution [12]. The COVID-19 pandemic also catapulted technology into the role of patient-provider intermediary [17,49]. One review identified strategies for improving kidney disease care using technology, including remote patient monitoring, patient education, and health-related apps to support self-management [50].

This current design and evaluation work builds upon our team's previous work and pairs it with key elements of HCD, namely, heuristic evaluation and usability testing conducted iteratively on successive versions of the PAUSE app UI prototype. A similar SDMG intervention, the Sick Day Protocol developed by the National Health Service Highland in Scotland to prevent acute kidney injury, was iteratively designed, tested, and refined using similar approaches. Patients with sick day medications were given a business card with drugs considered "hazardous" for sick days and instructions to stop those medications for 48 hours while sick, as well as a pamphlet and (as part of a clinical trial) a telecommunications platform to report and monitor sick day events [15]. Usability testing was conducted on the final protocol and card versions and revealed that patients were generally unable to identify which medications should be withheld during a sick day scenario [51]. The lack of patient involvement in earlier phases of the protocol and card design resulted in delayed issue identification, halting project progress after significant resource investment. Informed by this prior research knowledge, we followed a comprehensive HCD approach including heuristic evaluations and usability testing on predevelopment prototypes of the PAUSE app, ensuring a usable design and content, and that patients can act on the app's

XSL•FO
RenderX

recommendations prior to final design and development. The app was conceived and iteratively designed using an HCD approach, considering likely users (ie, the patients), lack of recall of information from a routine physician check-in, and their potentially impaired cognitive states during applicable sick days. In creating the PAUSE app prototype, the end user, the patient, was at the center of all design efforts and the app's interface and information flow designed accordingly.

Limitations

This work is not without limitations. Our heuristics evaluations and usability testing analyses were rapid, with the goal of providing timely feedback to the development team for integration into subsequent prototype versions. As such, the analysis approach was practical and pragmatic oriented, so some nuances related to reported elements and issues may have been missed. As the key value of heuristic evaluations and usability testing is in the practical application to real-world software development projects, nuances are less important, justifying the approach taken. The prototype UI is in English and the participants who completed the usability testing spoke and read English; however, we recognize that future iterations and integrations will need to consider other languages to be universally useful. We had 6 individuals partake in the usability testing; while this sample may be interpreted as small, it is in accordance with formative usability testing guidelines where approximately 80% of the issues are identified, which is intended as a happy balance between effort to solicit feedback and value-add of additional participants [35,36].

The app is being designed to function with minimal training, which supports real-world usability but may limit its usability among populations with lower digital literacy or limited access to mobile technology. This is a key question facing many digital health interventions, where those with access to technology can benefit but those without access are left behind. Reducing barriers to technology access might be one way to address this "digital divide" but is beyond the scope of this current work.

Conclusions

Within an HCD approach, we designed and refined an SDMG self-management prototype app for patients with diabetes, cardiovascular disease, and CKD. Heuristics evaluation and usability testing conducted iteratively between prototype revisions led to a final prototype design with key user flows and integrated supports for acting on severe acute illness situations where pausing SADMANS medications is recommended. The refined PAUSE app is well positioned to better support patients and their caregivers who are on specific medications for diabetes, cardiovascular disease, and CKD to more effectively self-manage at home and in accordance with current guideline recommendations. The HCD approaches used allowed us to address technical and workflow barriers using iterative heuristics evaluation and usability testing of the PAUSE app prototype design; however, we recognize that there remains an impetus for patients to remember to initiate the PAUSE app during an acute illness—which was out of scope of the present evaluation. Further research is needed to implement and evaluate the PAUSE app in community-based care and ensure that patients can accurately follow SDMG using the app, especially in times of acute illness.

Acknowledgments

This work was supported by a Canadian Institutes for Health Research Strategic Patient Oriented Multi-Year Research (grant 433773) and a LevMax grant from Alberta Innovates (no. 222301642). The funder had no role in study design, data collection, analysis, reporting, or the decision to submit for publication. The authors thank the patients and health care providers who provided input on the PAUSE app during the research, design, and development processes. The authors also thank the development team at Bitstrapped for helping them to realize their vision in software or website prototype form.

Conflicts of Interest

RT has received consulting fees from Shoppers Drug Mart and Emergent BioSolutions, is a paid editor in chief of Canadian Pharmacists Journal, and has received investigator-initiated grants from Merck, Pfizer, AstraZeneca, and Sanofi. DJTC has received investigator-initiated research support from Siemens Healthineers. The remaining authors declare that they have no relevant conflicts of interest to disclose.

Multimedia Appendix 1
Usability testing interview guide.
[DOCX File , 35 KB - formative_v9i1e77240_app1.docx]

Multimedia Appendix 2
Preventing medication complications during AcUte illness through Symptom Evaluation and sick day guidance heuristic evaluation results overview, issues within a category shown as a percentage of the total number (n=21). SDMG: Sick day medication guidance.
[PNG File , 391 KB - formative_v9i1e77240_app2.png]

Multimedia Appendix 3
Heuristic evaluation results.

XSL•FO
RenderX

[DOCX File , 399 KB - formative_v9i1e77240_app3.docx]

Multimedia Appendix 4
Usability testing results overview. Each category is shown as a percentage of the total number (n=44) of issues identified. SDMG: sick day medication guidance.
[PNG File , 493 KB - formative_v9i1e77240_app4.png]

Multimedia Appendix 5
Usability testing results.
[DOCX File , 894 KB - formative_v9i1e77240_app5.docx]

References

1. Stirling C, Houston J, Robertson S, Boyle J, Allan A, Norrie J, et al. Diarrhoea, vomiting and ACE inhibitors: an important cause of acute renal failure. J Hum Hypertens 2003;17(6):419-423. [doi: 10.1038/sj.jhh.1001571] [Medline: 12764405]
2. Whiting P, Morden A, Tomlinson LA, Caskey F, Blakeman T, Tomson C, et al. What are the risks and benefits of temporarily discontinuing medications to prevent acute kidney injury? A systematic review and meta-analysis. BMJ Open 2017;7(4):e012674 [FREE Full text] [doi: 10.1136/bmjopen-2016-012674] [Medline: 28389482]
3. Musso G, Saba F, Cassader M, Gambino R. Diabetic ketoacidosis with SGLT2 inhibitors. BMJ 2020;371:m4147 [FREE Full text] [doi: 10.1136/bmj.m4147] [Medline: 33184044]
4. Cardone KE, Bacchus S, Assimon MM, Pai AB, Manley HJ. Medication-related problems in CKD. Adv Chronic Kidney Dis 2010;17(5):404-412. [doi: 10.1053/j.ackd.2010.06.004] [Medline: 20727510]
5. Watson KE, Dhaliwal K, McMurtry E, Donald T, Lamont N, Benterud E, et al. Sick day medication guidance for people with diabetes, kidney disease, or cardiovascular disease: a systematic scoping review. Kidney Med 2022;4(9):100491 [FREE Full text] [doi: 10.1016/j.xkme.2022.100491] [Medline: 36046611]
6. Sick-day guidelines for people with diabetes. HealthLink BC. 2025. URL: https://www.healthlinkbc.ca/illnesses-conditions/diabetes/sick-day-guidelines-people-diabetes [accessed 2025-01-13]
7. Sick day rules: how to manage type 2 diabetes if become unwell with coronavirus (COVID) and what to do with your medication. GPnotebook. URL: https://gpnotebook.com/pages/diabetes-and-endocrinology/sick-day-rules-how-to-manage-type-2-diabetes-if-become-unwell-with-coronavirus-covid-and-what-to-do-with-your-medication [accessed 2025-01-13]
8. Sick day rules—how to manage type 1 diabetes if become unwell with coronavirus (covid). GPnotebook. 2025. URL: https://gpnotebook.com/pages/diabetes-and-endocrinology/sick-day-rules-how-to-manage-type-1-diabetes-if-become-unwell-with-coronavirus-covid [accessed 2025-01-13]
9. American Diabetes Association. Be prepared: sick day management. Diabetes Spectr 2002;15(1):54. [doi: 10.2337/diaspect.15.1.54]
10. Clinical practice guidelines. Diabetes Canada. 2025. URL: https://guidelines.diabetes.ca/cpg/chapter1 [accessed 2025-01-13]
11. Dhaliwal KK, Watson KE, Lamont NC, Drall KM, Donald M, James MT, et al. Managing 'sick days' in patients with chronic conditions: an exploration of patient and healthcare provider experiences. Health Expect 2023;26(4):1746-1756 [FREE Full text] [doi: 10.1111/hex.13789] [Medline: 37291977]
12. Watson KE, Dhaliwal K, Benterud E, Robertshaw S, Verdin N, McMurtry E, et al. Managing medications during "Sick Days" in patients with diabetes, kidney, and cardiovascular conditions: a theory-informed approach to intervention design and implementation. Can J Diabetes 2024;48(4):259-268.e4 [FREE Full text] [doi: 10.1016/j.jcjd.2024.02.003] [Medline: 38395301]
13. NA. KDIGO 2024 clinical practice guideline for the evaluation and management of chronic kidney disease. Kidneys (Ukrainian Journal of Renal Medicine) 2024;13(2):140-171 [FREE Full text] [doi: 10.22141/2307-1257.13.2.2024.456]
14. Clinical practice guidelines. Diabetes Canada. 2025. URL: https://guidelines.diabetes.ca/cpg [accessed 2025-03-31]
15. Fink JC, Maguire RM, Blakeman T, Tomlinson LA, Tomson C, Wagner L, et al. Medication holds in CKD during acute volume-depleting illnesses: a randomized controlled trial of a "Sick-Day" protocol. Kidney Med 2022;4(9):100527 [FREE Full text] [doi: 10.1016/j.xkme.2022.100527] [Medline: 36046613]
16. Duong H, Tesfaye W, Van C, Sud K, Truong M, Krass I, et al. Sick day management in people with chronic kidney disease: a scoping review. J Nephrol 2023;36(5):1293-1306 [FREE Full text] [doi: 10.1007/s40620-022-01497-5] [Medline: 36327038]
17. Stoumpos AI, Kitsios F, Talias MA. Digital transformation in healthcare: technology acceptance and its applications. Int J Environ Res Public Health 2023;20(4):3407 [FREE Full text] [doi: 10.3390/ijerph20043407] [Medline: 36834105]
18. IDEO. Field Guide to Human-Centered Design. DesignKit. New York, NY: IDEO.ORG; 2020. URL: https://www.designkit.org/resources/1.html [accessed 2025-10-15]
19. What is human-centered design—and how does it work in practice? Figma. 2025. URL: https://www.figma.com/resource-library/human-centered-design/ [accessed 2025-10-03]
20. Giacomin J. What is human centred design? Design J 2015;17(4):606-623. [doi: 10.2752/175630614X14056185480186]

XSL•FO
RenderX

21. Maguire M. Methods to support human-centred design. Int J Hum Comput Stud 2001;55(4):587-634. [doi: 10.1006/ijhc.2001.0503]
22. Bazzano AN, Martin J. Designing public health: synergy and discord. Design J 2017;20(6):735-754. [doi: 10.1080/14606925.2017.1372976]
23. Altman M, Huang TT, Breland JY. Design thinking in health care. Prev Chronic Dis 2018;15:E117 [FREE Full text] [doi: 10.5888/pcd15.180128] [Medline: 30264690]
24. Watson KE, Dhaliwal K, Robertshaw S, Verdin N, Benterud E, Lamont N, PAUSE (Preventing Medication Complications During Acute Illness Through Symptom EvaluationSick Day Guidance) Medication Safety Advisory Panel. Consensus recommendations for sick day medication guidance for people with diabetes, kidney, or cardiovascular disease: a modified Delphi process. Am J Kidney Dis 2023;81(5):564-574 [FREE Full text] [doi: 10.1053/j.ajkd.2022.10.012] [Medline: 36470530]
25. Nielsen J. Usability heuristics. In: Usability Engineering. Burlington, MA: Morgan Kaufmann; 1993:115-163.
26. Nielsen J, Molich R. Heuristic evaluation of user interfaces. 1990 Presented at: CHI '90: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems; April 1-5, 1990; Seattle, Washington p. 249-256. [doi: 10.1145/97243.97281]
27. Nielsen J. Usability Engineering. Burlington, MA: Morgan Kaufmann; 1993.
28. Moran K, Gordon K. How to conduct a heuristic evaluation. Nielsen Norman Group. 2023 Jun 25. URL: https://www.nngroup.com/articles/how-to-conduct-a-heuristic-evaluation/ [accessed 2025-03-30]
29. Nielsen J. Severity ratings for usability problems. Nielsen Norman Group. 1994 Nov 1. URL: https://www.nngroup.com/articles/how-to-rate-the-severity-of-usability-problems/ [accessed 2025-03-02]
30. Nielsen J. The theory behind heuristic evaluations. Nielsen Norman Group. 1994 Nov 1. URL: https://www.nngroup.com/articles/how-to-conduct-a-heuristic-evaluation/theory-heuristic-evaluations/ [accessed 2025-03-30]
31. G E. Familiarity feels good: the mere-exposure effect says we like what we know: food, people, and UX. UX Collective. 2020 Oct 5. URL: https://uxdesign.cc/familiarity-feels-good-8826da719933 [accessed 2025-04-01]
32. Zajonc RB. Attitudinal effects of mere exposure. J Pers Soc Psychol 1968;9(2):1-27. [doi: 10.1037/h0025848]
33. Bornstein RF, D'Agostino PR. Stimulus recognition and the mere exposure effect. J Pers Soc Psychol 1992;63(4):545-552. [doi: 10.1037//0022-3514.63.4.545] [Medline: 1447685]
34. Kushniruk AW, Patel VL. Cognitive and usability engineering methods for the evaluation of clinical information systems. J Biomed Inform 2004;37(1):56-76 [FREE Full text] [doi: 10.1016/j.jbi.2004.01.003] [Medline: 15016386]
35. Turner CW, Lewis JR, Nielsen J. Determining Usability Test Sample Size. In: Semantic Scholar. Boca Raton, FL: CRC Press; 2006:3076-3080.
36. Alroobaea R, Mayhew PJ. How many participants are really enough for usability studies? : IEEE; 2014 Presented at: Science and Information Conference; August 27-29, 2014; London, United Kingdom p. 48-56. [doi: 10.1109/sai.2014.6918171]
37. Elo S, Kyngäs H. The qualitative content analysis process. J Adv Nurs 2008;62(1):107-115. [doi: 10.1111/j.1365-2648.2007.04569.x] [Medline: 18352969]
38. Hsieh H, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
39. Norman CD, Skinner HA. eHealth literacy: essential skills for consumer health in a networked world. J Med Internet Res 2006;8(2):e9 [FREE Full text] [doi: 10.2196/jmir.8.2.e9] [Medline: 16867972]
40. Norman CD, Skinner HA. eHEALS: the eHealth literacy scale. J Med Internet Res 2006;8(4):e27 [FREE Full text] [doi: 10.2196/jmir.8.4.e27] [Medline: 17213046]
41. Sauro J. Measuring usability with the System Usability Scale (SUS). MeasuringU. 2011 Feb 11. URL: https://measuringu.com/sus/ [accessed 2025-03-03]
42. Laubheimer P. Beyond the NPS: measuring perceived usability with the SUS, NASA-TLX, and the single ease question after tasks and usability tests. Nielsen Norman Group. 2018 Feb 11. URL: https://www.nngroup.com/articles/measuring-perceived-usability/ [accessed 2025-03-03]
43. Web Content Accessibility Guidelines (WCAG). World Wide Web Consortium. 2025. URL: https://www.w3.org/TR/WCAG21/ [accessed 2025-03-19]
44. Ammenwerth E, Wilk S, Huang Z. Personalization in mHealth: innovative informatics methods to improve patient experience and health outcome. J Biomed Inform 2023;147:104523 [FREE Full text] [doi: 10.1016/j.jbi.2023.104523] [Medline: 37838289]
45. Stay safe when you have diabetes and are sick or at risk of dehydration? Diabetes of Canada. 2025. URL: https://www.diabetes.ca › tools---resources › sick- [accessed 2025-10-03]
46. Truong M, Van C, Sud K, Tesfaye W, Croker N, Seth S, et al. Drug-related problems and recommendations made during home medicines reviews for sick day medication management in Australia. Medicina (Kaunas) 2024;60(5):798 [FREE Full text] [doi: 10.3390/medicina60050798] [Medline: 38792982]
47. Coppes T, Philbert D, van Gelder T, Bouvy ML, Koster ES. Medication management during sick days: no differences between patients with and without impaired renal function. Eur J Clin Invest 2024;54(9):e14231. [doi: 10.1111/eci.14231] [Medline: 38641996]

XSL•FO
RenderX

48. Hall RK, Kazancıoğlu R, Thanachayanont T, Wong G, Sabanayagam D, Battistella M, et al. Drug stewardship in chronic kidney disease to achieve effective and safe medication use. Nat Rev Nephrol 2024;20(6):386-401. [doi: 10.1038/s41581-024-00823-3] [Medline: 38491222]
49. Sakur F, Ward K, Khatri NN, Lau AYS. Self-care behaviors and technology used during COVID-19: systematic review. JMIR Hum Factors 2022;9(2):e35173 [FREE Full text] [doi: 10.2196/35173] [Medline: 35442904]
50. Jain G, Ahmad M, Wallace EL. Technology, telehealth, and nephrology: the time is now. Kidney360 2020;1(8):834-836 [FREE Full text] [doi: 10.34067/KID.0002382020] [Medline: 35372956]
51. Doerfler RM, Diamantidis CJ, Wagner L, Scism BM, Vaughn-Cooke M, Fink WJ, et al. Usability testing of a sick-day protocol in CKD. Clin J Am Soc Nephrol 2019;14(4):583-585 [FREE Full text] [doi: 10.2215/CJN.13221118] [Medline: 30867156]

Abbreviations

CKD: chronic kidney disease
eHEALS: eHealth Literacy Scale
EMR: electronic medical record
HCD: human-centered design
HCP: health care provider
PAUSE: Preventing medication complications during AcUte illness through Symptom Evaluation and sick day guidance
SDMG: sick day medication guidance
UI: user interface

Edited by A Mavragani, J Sarvestan; submitted 12.05.25; peer-reviewed by R Marshall, W Tesfaye; comments to author 18.07.25; revised version received 08.08.25; accepted 01.10.25; published 27.11.25.

Please cite as:
Babione J, Gil S, Okemeziem C, Palechuk T, Watson KE, Robertshaw S, Verdin N, McBrien K, Campbell DJT, Tsuyuki R, Pannu N, James M, Donald M
Using Human-Centered Design and Development to Create a Digital Sick Day Medication Guidance Application for People With Diabetes, Cardiovascular Disease, or Chronic Kidney Disease: Mixed Methods Study
JMIR Form Res 2025;9:e77240
URL: https://formative.jmir.org/2025/1/e77240
doi:10.2196/77240
PMID:

©Julie Babione, Sarah Gil, Chidera Okemeziem, Taylor Palechuk, Kaitlyn E Watson, Sandra Robertshaw, Nancy Verdin, Kerry McBrien, David J T Campbell, Ross Tsuyuki, Neesh Pannu, Matthew James, Maoliosa Donald. Originally published in JMIR Formative Research (https://formative.jmir.org), 27.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

GrantCheck—an AI Solution for Guiding Grant Language to New Policy Requirements: Development Study

Qiming Shi[1], MS; Asil Oztekin[2], PhD; George Matthew[3], MS; Jeffrey Bortle[3], BS; Hayden Jenkins[3], BS; Steven (Koon) Wong[4], MS; Paul Langlois[3], MBA; Anaheed Zaki[3], PhD; Brian Coleman[3], BS; Katherine Luzuriaga[1], MD; Adrian H Zai[4], MPH, MD, PhD

[1]Center for Clinical and Translational Science, UMass Chan Medical School, Worcester, MA, United States

[2]Manning School of Business, UMass Lowell, Lowell, MA, United States

[3]Information Technology, UMass Chan Medical School, Worcester, MA, United States

[4]Department of Population and Quantitative Health Sciences, UMass Chan Medical School, Worcester, MA, United States

Corresponding Author:
Adrian H Zai, MPH, MD, PhD
Department of Population and Quantitative Health Sciences
UMass Chan Medical School
55 N Lake Ave
Worcester, MA, 01655
United States
Phone: 1 617 256 4120
Email: Adrian.Zai@umassmed.edu

Abstract

Background: Academic institutions face increasing challenges in grant writing due to evolving federal and state policies that restrict the use of specific language. Manual review processes are labor-intensive and may delay submissions, highlighting the need for scalable, secure solutions that ensure compliance without compromising scientific integrity.

Objective: This study aimed to develop a secure, artificial intelligence–powered tool that assists researchers in writing grants consistent with evolving state and federal policy requirements.

Methods: GrantCheck (University of Massachusetts Chan Medical School) was built on a private Amazon Web Services virtual private cloud, integrating a rule-based natural language processing engine with large language models accessed via Amazon Bedrock. A hybrid pipeline detects flagged terms and generates alternative phrasing, with validation steps to prevent hallucinations. A secure web-based front end enables document upload and report retrieval. Usability was assessed using the System Usability Scale.

Results: GrantCheck achieved high performance in detecting and recommending alternatives for sensitive terms, with a precision of 1.00, recall of 0.73, and an F_1-score of 0.84—outperforming general-purpose models including GPT-4o (OpenAI; F_1=0.43), Deepseek R1 (High-Flyer; F_1=0.40), Llama 3.1 (Meta AI; F_1=0.27), Gemini 2.5 Flash (Google; F_1=0.58), and even Gemini 2.5 Pro (Google; F_1=0.72). Usability testing among 25 faculty and staff yielded a mean System Usability Scale score of 85.9 (SD 13.4), indicating high user satisfaction and strong workflow integration.

Conclusions: GrantCheck demonstrates the feasibility of deploying institutionally hosted, artificial intelligence–driven systems to support compliant and researcher-friendly grant writing. Beyond administrative efficiency, such systems can indirectly safeguard public health research continuity by minimizing grant delays and funding losses caused by language-related policy changes. By maintaining compliance without suppressing scientific rigor or inclusivity, GrantCheck helps protect the pipeline of research that advances biomedical discovery, health equity, and patient outcomes. This capability is particularly relevant for proposals in sensitive domains—such as social determinants of health, behavioral medicine, and community-based research—that are most vulnerable to evolving policy restrictions. As a proof-of-concept development study, our implementation is tailored to one institution's policy environment and security infrastructure, and findings should be interpreted as preliminary rather than universally generalizable.

(JMIR Form Res 2025;9:e79038) doi:10.2196/79038

XSL•FO
RenderX

KEYWORDS

artificial intelligence; natural language processing; large language model; user-computer interface; grant applications; research compliance; usability evaluation

Introduction

Recent shifts in federal and state policies have imposed restrictions on certain terminology in grant applications, creating new challenges for researchers during proposal development [1-3]. Terms related to concepts such as diversity, gender identity, and climate change [4] have, in some cases, been informally "flagged" or discouraged in funding submissions. This evolving policy environment has introduced uncertainty and increased administrative workload: grant offices must now invest additional time in language screening, and investigators risk delays or rejection if proposals contain politically sensitive terms [5-8]. These pressures highlight an urgent need for tools that enable institutions to ensure compliance with language requirements while preserving the clarity and integrity of scientific communication. These evolving requirements have far-reaching implications for biomedical and public health research. Policy-related grant delays or rejections can interrupt ongoing studies in cancer prevention, infectious disease surveillance, or chronic disease management, leading to gaps in data continuity and slower translation of findings into clinical or population-level benefit. Furthermore, proposals focused on health equity or community engagement are disproportionately affected by language restrictions, risking underrepresentation of vulnerable populations in federally funded research. Addressing these challenges is therefore not merely an administrative priority but a health systems imperative—ensuring compliant communication safeguards the research pipeline that underpins patient care innovation, disease prevention, and equitable health outcomes.

To address this challenge, we pose the following research question: Can a domain-specific artificial intelligence (AI) system automatically detect policy-restricted language in grant documents and assist in revising it to meet new requirements, while maintaining the integrity of the scientific content? Our objective is to develop and evaluate a solution that achieves high accuracy in identifying noncompliant phrases and providing appropriate alternatives. Evaluation metrics include precision and recall in detecting flagged terms, as well as user-centered measures such as the System Usability Scale (SUS) to assess the acceptability and practical utility of the tool's recommendations.

A variety of writing-support technologies have emerged to help researchers navigate increasingly complex requirements. General-purpose AI writing assistants, such as Grammarly (Superhuman Platform Inc) [9] and ChatGPT (OpenAI) [10], can refine grammar and style but are not designed to enforce domain-specific policies. In response, several grant-specific tools have been proposed, and multiple peer-reviewed studies have examined large language models (LLMs) directly in the context of grant writing. Seckel et al [11] provided structured guidelines for ethical and effective LLM use in proposal preparation, highlighting their utility for literature summarization and clarity improvements while cautioning against overreliance and plagiarism risks. Godwin et al [12] reported that a guided generative AI tool for National Institutes of Health grant drafting yielded substantial time savings and enhanced workflow efficiency. Meyer et al [13] observed that LLMs can accelerate early drafting by overcoming writer's block, although human revision is essential to ensure specificity and policy compliance.

In addition, LLMs have been leveraged to streamline documentation and policy adherence. Fukataki et al [14] demonstrated that GPT-4 (OpenAI), when combined with tailored prompts, could accurately extract ethically relevant information from research protocols for Institutional Review Board (IRB) prereview, achieving high reproducibility and alignment with human reviewer assessments. Similarly, Heilmeyer et al [15] showed that domain-tuned open LLMs deployed in a privacy-preserving local environment produced high-quality clinical documentation, with 93% of outputs requiring minimal or no edits before integration into medical records. LLMs have also been shown to enhance general scientific and medical writing. Huang and Tan [16] reported that ChatGPT can improve efficiency and writing quality in scientific review article preparation, particularly for expanding content and refining style.

Collectively, these studies demonstrate the potential for LLMs to support domain-specific writing and compliance-related document preparation. However, to our knowledge, there is no previous study that addresses compliance requirements in grant language while simultaneously delivering an institutionally hosted solution designed to mitigate confidentiality risks.

Building on this body of work, GrantCheck (University of Massachusetts Chan Medical School) introduces an institutionally hosted, hybrid AI system that combines rule-based natural language processing (NLP) with LLM-based rephrasing to support compliant grant writing. Unlike previous tools, GrantCheck explicitly targets policy-sensitive term detection under evolving funding guidelines and provides secure, context-appropriate alternatives within a private cloud environment—addressing both compliance requirements and confidentiality concerns. This line of innovation carries important implications for the broader research and health ecosystem. As grant restrictions increasingly affect studies on population health, behavioral science, and health equity, tools like GrantCheck can help preserve scientific progress in these sensitive domains by ensuring compliant communication rather than topic avoidance. In doing so, GrantCheck not only streamlines institutional compliance but also supports the continuity of critical health research addressing disparities, chronic disease prevention, and emerging public health priorities.

XSL•FO
RenderX

Methods

Overview

This development study was structured as a two-part evaluation: (1) performance testing of GrantCheck using a gold-standard annotated dataset, and (2) usability testing with faculty and research staff. The intervention was delivered via a secure web-based application where participants uploaded grant documents, which were processed by the hybrid NLP–LLM pipeline. Output reports were automatically generated and returned to users through the same interface.

Ethical Considerations

This study was reviewed by the IRB of UMass Chan Medical School (STUDY00002481) and was determined not to constitute research involving human participants under Department of Health and Human Services and Food and Drug Administration regulations. Consequently, formal IRB oversight and informed consent were not required. Usability testing was conducted with faculty and staff members who voluntarily participated in the SUS survey. Participation was anonymous, and no identifiable personal data were collected. All procedures complied with institutional policy, the principles outlined in the Declaration of Helsinki, *Journal of Medical Internet Research*'s ethical requirements for usability studies, and Committee on Publication Ethics guidelines for responsible research conduct. No compensation was provided to participants.

Infrastructure Layer

To support the secure and scalable deployment of GrantCheck, the system was implemented within Amazon Web Services (AWS) Cloud, ensuring strict adherence to institutional data security and privacy standards. This backend infrastructure supports the integration of advanced LLMs through Amazon Bedrock [17] while maintaining a fully secured environment. Microsoft Single Sign-On (SSO) authentication restricts front-end access to authorized users.

GrantCheck's back end runs in a private, encrypted AWS virtual private cloud (VPC) without direct exposure to the public internet. The infrastructure includes Amazon Bedrock for LLM inference (Amazon Nova Pro and Meta Llama 3.1 70B Instruct), Identity and Access Management–controlled AWS Lambda serverless function hosting the processing engine and orchestration logic, handling data in memory without persistent storage. While user uploads are temporarily stored in Amazon S3 to support processing, a lifecycle rule automatically deletes these files after 1 day, ensuring minimal data retention. All data in transit are protected using Transport Layer Security 1.3 encryption, and temporary cryptographic keys are managed via AWS Key Management Service.

Access to the tool is gated through SSO, with audit logging implemented across application and network layers to support traceability. The AWS infrastructure diagram (Figure 1) provides an overview of how the components interact within the secure environment.

Figure 1 depicts the cloud infrastructure supporting the application, including authentication, frontend, backend processing, and LLM orchestration through Cloud, all within a private, encrypted VPC environment.

Figure 1. Cloud infrastructure. API: application programming interface.

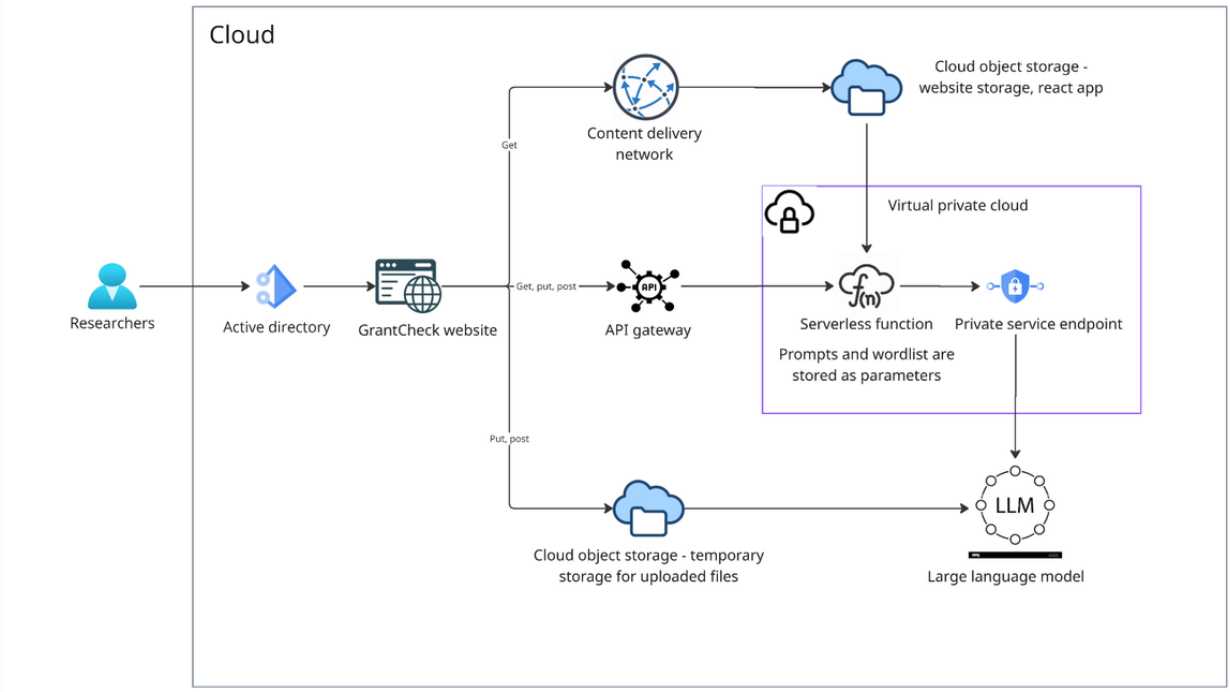

Security Measures

GrantCheck is currently deployed in a preproduction environment within an AWS VPC and was therefore not included in the most recent institutional third-party penetration testing cycle. Upon production release, the system will be incorporated into our institution's routine penetration testing process, which is performed by an external vendor under the

XSL•FO
RenderX

direction of the Information Security Office. The application's configuration adheres to institutional security controls aligned with established standards such as the National Institute of Standards and Technology and the Center for Internet Security benchmarks.

Per-user data isolation is enforced through AWS Amplify's built-in storage authorization, ensuring that files are accessible only to their respective owners. Session token expiration is managed via Amazon Cognito, with refresh tokens expiring after 1 day and both access and ID tokens expiring after 1 hour. Although formal prompt injection resistance testing has not yet been conducted, it is planned as part of the security validation process before production deployment. These measures, combined with the isolated VPC environment and session-based memory clearing, help safeguard user data and maintain compliance with institutional security requirements.

Backend Processing Pipeline

A hybrid pipeline was implemented to improve the precision and reliability of sensitive language detection. It combines rule-based NLP with LLM inference. This layered approach mitigates the limitations of LLM-only methods, especially concerning hallucination, token leakage, and contextual ambiguity.

The initial processing stage uses a custom NLP module built with Python's regular expression (regex) library [18]. We developed a curated dictionary of 238 sensitive terms and phrases for language identification, with the majority derived from Appendix B of the Senate guidelines [19]. This list was expanded through multiple sources and underwent thorough review and approval by the school's faculty council and executive committee, ensuring that it extended beyond terminology alone to reflect emerging areas of policy sensitivity. For example, terms such as "vaccine hesitancy" and "vaccine uptake" were included, given their recent association with grant cancellations [20]. To ensure the dictionary remains up-to-date and accurate, a feedback system is in place where researchers can suggest additions, which undergo thorough review before potential integration into the existing list. Terms are annotated with contextual rules to distinguish benign from problematic uses.

Regex patterns were programmatically generated to ensure consistent and reproducible matching across the sensitive-term dictionary. Patterns were designed with word boundaries to capture only whole words, and case-insensitive matching was applied to avoid missing capitalized variants. For multiword terms, a single fixed whitespace was used to ensure precise detection.

When a document is uploaded, the engine scans its content to identify sensitive terms using regex, extracts the surrounding sentence for context, and generates a structured table listing each term along with its corresponding sentence. The structured table is passed to an LLM via Amazon Bedrock in the second stage to make an alternative term for replacement. Model parameters are tuned with a temperature of 0.7, a top-p setting of 0.9, and a maximum token count of 5000. These settings are version-controlled to ensure consistent inference.

Following LLM inference, the regex engine re-evaluates the proposed alternative text. If any flagged sensitive term is found in the LLM's output, the replacement is suppressed and replaced with a standard message: "No safe alternative found." This safeguard was added to address residual cases of hallucination or subtle term reintroduction. The end-to-end pipeline for this process is illustrated in Figure 2.

Figure 2 illustrates the document flow through the application, beginning with document upload, followed by sensitive term detection, LLM-based suggestion generation, post-processing validation, and finalized output.

Figure 2. End-to-end pipeline for regex-guided sensitive term replacement via large language model. LLM: large language model; UI: user interface.

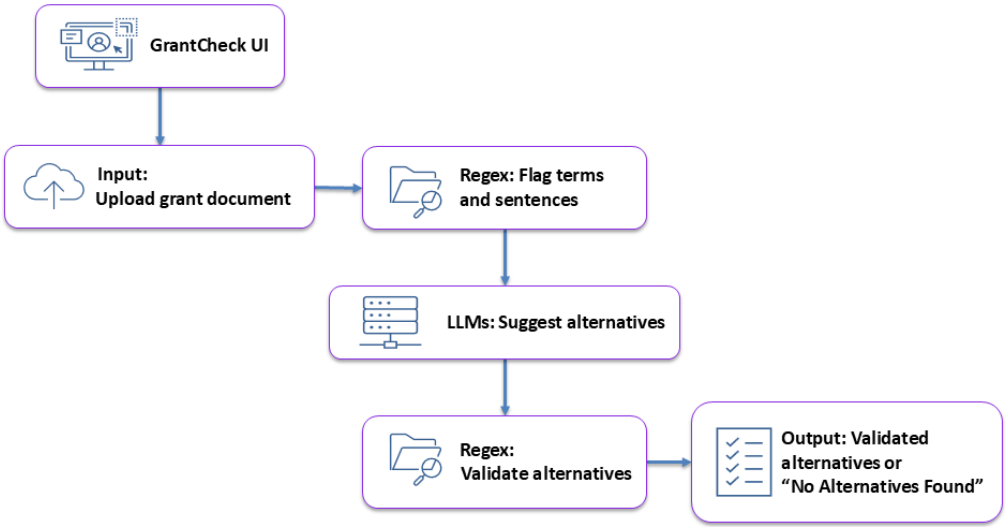

Front End Interface

A lightweight, web-based front end (Figure 3) was developed to provide a secure, user-friendly interface for document upload and report retrieval. This interface was custom-built to comply with institutional security policies. The front end enables users to upload grant documents, which are transmitted securely to the backend processing pipeline. Once the analysis is complete, the user is presented with a report summarizing flagged sensitive

XSL•FO
RenderX

terms and their suggested alternatives. Access to the application is limited to authenticated users through Microsoft SSO. The front end does not store any document content; data are stored in S3 for up to 1 day.

Figure 3. GrantCheck user interface.

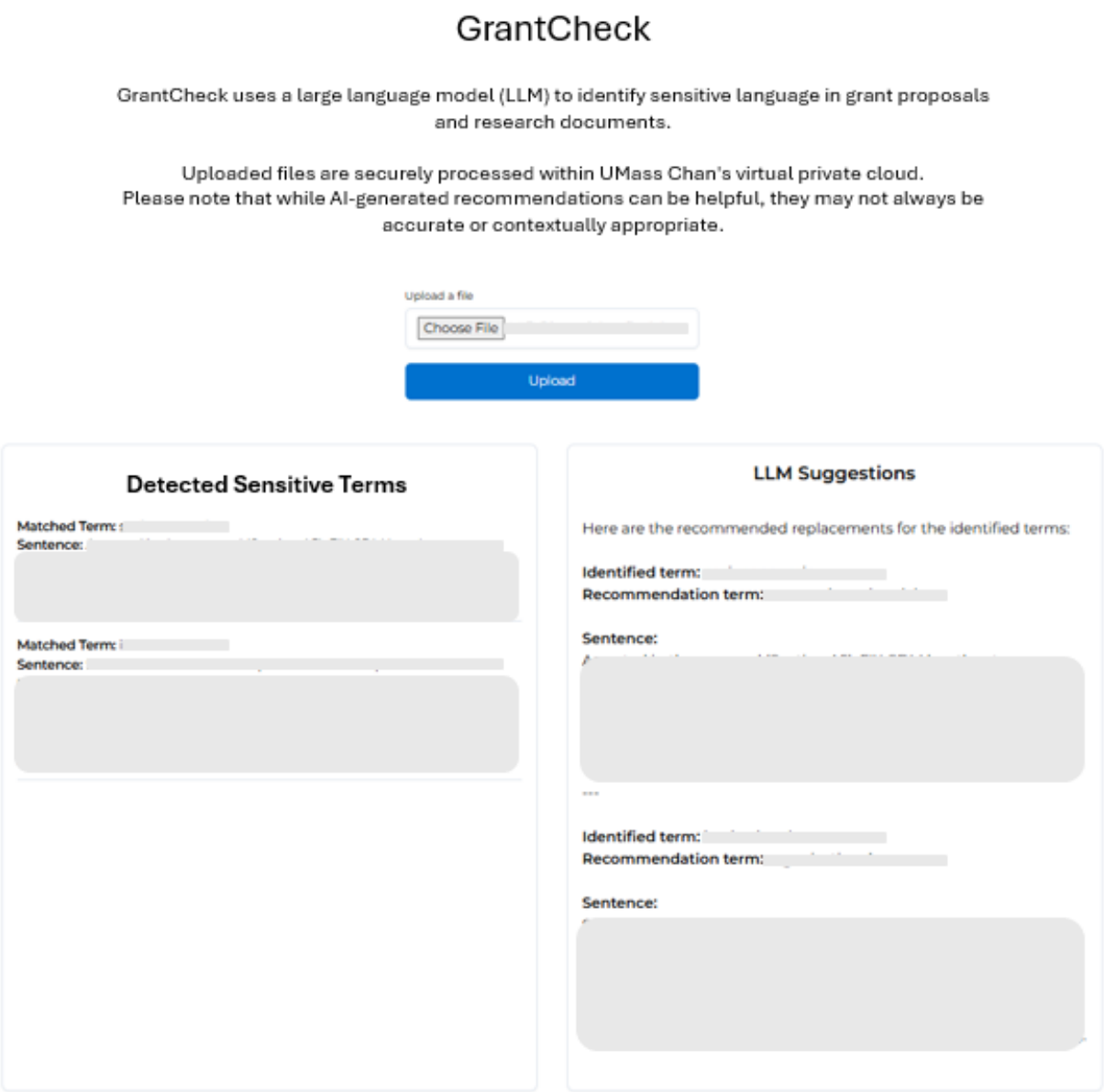

Figure 3 illustrates the front end interface presented to users during the beta testing phase. The interface includes a brief description of GrantCheck's function, a file upload field, and an action button for submission. Results are presented in 2 sections—one listing the detected sensitive terms and another showing suggestions generated by LLM. Documents are processed within the secure VPC, and users are reminded that suggestions are AI-generated and may not always be accurate.

Performance Evaluation

To assess GrantCheck's effectiveness in identifying sensitive language, we conducted a comparative evaluation using a gold standard dataset derived from manually annotated grant proposals. In total, 5 documents were manually annotated by 2 annotators with extensive experience in grant writing and compliance review. Each document was reviewed against the institution's sensitive-term dictionary. True positives, false positives, and false negatives were identified for precision, sensitivity, and specificity calculation. Because identification of terms from the curated sensitive word list is a relatively deterministic task with minimal subjective interpretation, each document was annotated once by a single annotator rather than independently by multiple annotators. We acknowledge that this choice precluded formal assessment of interannotator agreement and represents a limitation that future validation studies should address through double annotation and calculation of interrater reliability metrics such as Cohen κ. The set included both active grant applications under development at our institution and scholarly articles with a strong emphasis on Diversity, Equity, and Inclusion–related principles, selected to test the system's maximum detection capability. Because identifying terms from the curated sensitive word list is a

XSL•FO
RenderX

straightforward task with minimal subjective interpretation, each document was annotated once by a single annotator, rather than having both annotators independently annotate the same document to calculate interannotator agreement. In total, 2 system configurations were tested: version 1, which used a general-purpose LLM with prompts applied to the full document, and version 2 (GrantCheck), which implemented a hybrid pipeline combining rule-based regex detection with targeted LLM inference. For each model, we constructed contingency tables of true positives, false positives, true negatives, and false negatives, and calculated precision, sensitivity (recall), and specificity [21,22]. Furthermore, 95% CIs for these metrics were estimated via 5000 bootstrap resamples [23], and the McNemar exact test [24] was used to assess whether observed differences in error rates between GrantCheck and baseline models were statistically significant ($P<.05$). This framework provided a comprehensive diagnostic evaluation that captured both performance accuracy and error patterns.

Usability Testing

A total of 25 participants were recruited for the SUS survey [25] following statistical guidance from Sauro and Lewis [26], which indicates that, given the typical standard deviation of SUS scores (approximately 12.5), this sample size yields a 95% CI of approximately ±5 points—an accepted precision level for usability benchmarking. This number balances statistical rigor with resource efficiency. Eligible participants were faculty and research staff with recent grant-writing experience, representing the intended end user group for GrantCheck. A convenience sample of 30 faculty and staff with recent grant-writing experience was invited via institutional email, and 25 participated (response rate 83%). All participants completed the SUS in full, with no partial responses. The cohort included 20 principal investigators and 5 scientific writers from diverse departments, including Population & Quantitative Health Sciences, Pulmonary Medicine, the Center for Clinical and Translational Science, Genetics & Cellular Medicine, Cardiovascular Medicine, Molecular, Cell & Cancer Biology, Medicine, Microbiology, Health Systems Science, Emergency Medicine, Pediatrics, Surgery, Family Medicine & Community Health, and Neurobiology.

During the beta testing phase, participants completed the SUS survey (Multimedia Appendix 1), which included 10 structured questions scored on a 5-point Likert scale to evaluate user experience, interface design, and overall satisfaction. An additional open-ended question invited recommendations and suggestions for improvement. Survey responses were collected anonymously, and participation was voluntary. All usability data were reviewed to guide iterative refinements to the user interface and workflow integration. This activity was determined not to constitute human participants research and did not require IRB oversight under institutional policy.

To compute the SUS score, responses to positively worded items (items 1, 3, 5, 7, and 9) were adjusted by subtracting 1 from the raw score. For negatively worded items (items 2, 4, 6, 8, and 10), responses were reverse scored by subtracting the raw score from 5. Each adjusted item score ranged from 0 to 4, and the sum of all 10 adjusted item scores was multiplied by 2.5 to yield a final SUS score ranging from 0 to 100. Higher scores indicate better usability. Scores ≥68 were interpreted as above average. Descriptive statistics (mean, SD, minimum, maximum, and median) were reported for overall SUS results, and open-text feedback was reviewed qualitatively to identify themes related to feasibility and acceptability.

Security and Data Governance

Our Information Security Office reviewed and approved the backend infrastructure design. User uploads are temporarily stored in S3 and automatically deleted after one day through a defined lifecycle rule. Once analysis is completed, the session is terminated, and all memory is cleared within 1 day. To mitigate risks of prompt injection and unintentional data retention, users do not interact with the LLM directly, and the model end points are isolated from internet-facing services. Security audits and threat reviews are conducted quarterly to ensure ongoing compliance. In addition, web application vulnerability scans are performed before each deployment to identify and mitigate potential risks before changes go live.

Plan-Do-Study-Act Cycle

We continuously incorporated user feedback into the development process, using it to refine and enhance the tool through iterative updates.

Results

Across the annotated evaluation set, GrantCheck achieved the strongest overall diagnostic performance (Table 1). It correctly identified 16 of 22 policy-sensitive terms (true positive) while producing no false positives, yielding perfect precision and specificity (both 100%) and a sensitivity of 73% (95% CI 56%-87%). By contrast, ChatGPT-4o (OpenAI), Deepseek R1 (High-Flyer), and Llama 3.1 70B (Meta AI) showed substantially lower sensitivities (41%, 32%, and 18%, respectively) coupled with false positives that reduced specificity. Gemini 2.5 Flash (Google) demonstrated moderate performance (sensitivity 50% and precision 69%), while Gemini 2.5 Pro approached GrantCheck in precision (93%) and specificity (95%) but exhibited lower sensitivity (59%).

Exact McNemar tests confirmed that GrantCheck's improvements over ChatGPT-4o, Deepseek R1, and Llama 3.1 were statistically significant ($P<.05$), while the difference versus Gemini 2.5 Flash was borderline ($P=.09$). The comparison with Gemini 2.5 Pro was not statistically significant ($P=.42$), reflecting its relatively strong performance on this dataset. These findings highlight GrantCheck's ability to maximize compliance safety by avoiding false positives, while retaining higher sensitivity than competing general-purpose LLMs.

XSL•FO
RenderX

Table 1. Comparative diagnostic performance of GrantCheck and baseline models in detecting policy-sensitive terms.

Model	TP[a]	FP[b]	TN[c]	FN[d]	Precision (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	McNemar vs Amandla (*P* value)
GrantCheck (Amandla)	16	0	20	6	1.00 (0.96-1.00)	0.73 (0.56-0.87)	1.00 (0.95-1.00)	__[e]
ChatGPT-4o	9	11	9	13	0.45 (0.26-0.65)	0.41 (0.23-0.59)	0.45 (0.25-0.65)	.04
Deepseek R1	7	6	14	15	0.54 (0.29-0.77)	0.32 (0.14-0.50)	0.70 (0.50-0.85)	.01
Llama 3.1 70B	4	4	16	18	0.50 (0.20-0.80)	0.18 (0.04-0.36)	0.80 (0.61-0.95)	.001
Gemini 2.5 Flash	11	5	15	11	0.69 (0.44-0.89)	0.50 (0.31-0.69)	0.75 (0.56-0.90)	.09
Gemini 2.5 Pro	13	1	19	9	0.93 (0.71-1.00)	0.59 (0.40-0.77)	0.95 (0.82-1.00)	.42

[a]TP: true positive—cases correctly identified as positive.

[b]FP: false positive—cases incorrectly identified as positive.

[c]TN: true negative—cases correctly identified as negative.

[d]FN: false negative—cases incorrectly identified as negative.

[e]Not available.

As illustrated in Table 2, the mean final SUS score across participants was 85.90 (SD 13.44), suggesting above-average usability. The score distribution ranged from 55 to 100, with a median of 92.5 (IQR 80.0-98.0), indicating that most participants found the system highly usable. Individual item means ranged from 3.08 to 3.80 on a scale of 0 to 4. Given that a score above 2 is above the midpoint, this indicates generally positive responses across all items. The highest agreement was observed on the need for minimal technical support and ease of learning, while slightly lower scores were noted for perceived consistency and user confidence. Overall, the item-level averages suggest the application performs well in supporting usability.

Table 2. Usability survey results.

SUS[a] item	Mean (SD)	Median (IQR)
I think that I would like to use this system frequently	3.16 (0.94)	3.00 (2.00-4.00)
I found the system unnecessarily complex	3.52 (1.16)	4.00 (3.00-4.00)
I thought the system was easy to use	3.64 (0.64)	4.00 (3.00-4.00)
I think that I would need the support of a technical person to be able to use this system	3.76 (0.83)	4.00 (3.00-4.00)
I found the various functions in this system were well integrated	3.08 (0.95)	3.00 (2.00-4.00)
I thought there was too much inconsistency in this system	3.08 (1.15)	4.00 (3.00-4.00)
I would imagine that most people would learn to use this system very quickly	3.80 (0.50)	4.00 (3.00-4.00)
I found the system very cumbersome to use	3.68 (0.75)	4.00 (3.00-4.00)
I felt very confident using the system	3.28 (0.89)	4.00 (2.00-4.00)
I needed to learn a lot of things before I could get going with this system	3.36 (1.19)	4.00 (2.00-4.00)
Final SUS score (0-100)	85.90 (13.44)	92.50 (80.00-98.00)

[a]SUS: System Usability Scale.

In addition to the SUS scores, open-ended feedback revealed four recurring themes: (1) export and print functionality, with users requesting the ability to export flagged results into Microsoft Word documents and print reports; (2) clarity and guidance on how to interpret and act upon flagged content, highlighting the need to pair technical recommendations with integrated educational support; (3) accuracy and utility of suggestions, as some replacements such as “women’s health” to “midwifery health” were viewed as unhelpful or overly rigid, whereas substitutions such as replacing “diversity” with “broad representation” and “equity” with “fair access” were considered helpful and appropriate; and (4) highlighting and replacement functionality, with requests for in-document highlighting and inline replacement options rather than a separate spotter report.

Discussion

Design and Features

The development of GrantCheck reflects a practical response to a rapidly evolving policy landscape in academic and research institutions. While this study demonstrates feasibility, it should be understood as a proof-of-concept development study tailored to one institution’s curated dictionary, policy interpretations, and security requirements. As such, the findings provide preliminary evidence of feasibility but are not intended to suggest universal generalizability. As funding agencies and state governments implement new language-related restrictions, institutions face significant challenges in aligning grant narratives with compliance requirements while maintaining the scientific integrity of their proposals. GrantCheck addresses

XSL•FO
RenderX

this emerging need by providing a secure, institutionally compliant solution that assists users in identifying and revising potentially sensitive terminology using a hybrid regex–LLM pipeline.

Initially, we experimented with GPT-4o using few-shot prompting, which yielded promising results but revealed limitations in both overcalling and undercalling sensitive terms. To address these issues, we developed GrantCheck, a hybrid NLP pipeline that integrates rule-based term detection with LLM inference. This approach balances the precision of deterministic methods with the contextual sensitivity of generative AI. A postprocessing validation step further enhances system reliability by mitigating hallucinations and preventing the inadvertent reintroduction of sensitive language. These results should be interpreted with caution, as they are based on a small, annotated dataset and a single-institution user group. Our evaluation shows that GrantCheck consistently achieved perfect precision and specificity, with no false positives across the test set. Our reported performance metrics should be interpreted as preliminary findings based on a small, annotated dataset, designed to demonstrate feasibility rather than to provide definitive benchmarks. This outcome was confirmed through gold-standard annotation and reflects the tool's conservative design, which combines regex-based deterministic detection with a safeguard that suppresses any LLM outputs reintroducing flagged terms. This feature is critical in the grant compliance context, as overcalling benign terms could unnecessarily disrupt proposal preparation, introduce avoidable edits, and risk misinterpretation by reviewers. Its sensitivity of 73% indicates that most, but not all, sensitive terms are detected. By contrast, general-purpose LLMs (eg, ChatGPT-4o, Deepseek R1, and Llama 3.1) showed significantly weaker performance across both sensitivity and specificity, while Gemini 2.5 Pro emerged as the strongest baseline competitor with high precision (93%) and specificity (95%) but lower sensitivity (59%). Reporting 95% CIs and conducting McNemar exact tests confirmed that GrantCheck's improvements over ChatGPT-4o, Deepseek R1, and Llama 3.1 were statistically significant, while differences with Gemini 2.5 Pro were not, underscoring its relative strength. The error analysis provided further insight into system behavior. False negatives in GrantCheck were primarily attributable to information overflow when large numbers of flagged segments exceeded the LLM's optimal processing capacity, causing some terms to be dropped from the suggestion stage. This issue is discussed in more detail later in the manuscript as a limitation of long-context inference. By contrast, baseline models produced false negatives for different reasons, most often missing sensitive terms in less common formulations, and also generated false positives by overflagging benign technical usage. These patterns highlight that GrantCheck's conservative design intentionally prioritizes avoiding false positives—an essential feature in compliance contexts where benign terms should not be mistakenly flagged.

From an infrastructure perspective, the deployment of GrantCheck within a VPC environment demonstrates the feasibility of securely operationalizing LLM-based tools for institutional use. This architecture, which isolates all data processing from public end points and enforces session-based memory clearing, is a model for responsible AI deployment in academic settings. From a cost-efficiency perspective, integrating a regex preprocessing step allows the system to extract and summarize only relevant segments before invoking the LLM. This significantly reduces token consumption by avoiding full-document processing, thereby lowering computational costs without compromising output quality. For example, in a grant containing a high density of sensitive terms, a typical run might involve approximately 3000 input tokens and 500 output tokens. On Amazon Nova Pro pricing, this equates to approximately US $0.036 per request. Even under an upper-bound assumption of 30 users each running the tool once per day, the estimated daily cost would be US $1.08, underscoring that real-world institutional costs are expected to remain very modest relative to the value of preventing grant rejections.

The usability testing results indicate that tools like GrantCheck can be seamlessly integrated into research workflows without burdening end users. Using the SUS enabled standardized assessment, while open-ended feedback informed iterative improvements. Early users helped uncover key limitations—specifically, overcalling and undercalling issues without a regex layer. In response, we implemented a regex-based component to extract identified sensitive terms and their corresponding sentence contexts. These outputs are then passed to the LLM, which generates recommended alternatives. This 2-step approach enhances reliability by narrowing the input scope, mitigating the risk of the LLM content and coherence loss when processing longer text, a known limitation of LLMs. Several users suggested providing a detailed report of the regex-identified terms, including their locations in the source text. To address this, we added a dual-pane display: one window shows each term and its sentence context, while the other presents the LLM's recommended alternatives. One researcher noted uncertainty about whether the term list was exhaustive. By surfacing all regex-identified terms explicitly, the tool builds confidence in its comprehensiveness.

Within a month of its launch, the tool received 128 access requests from researchers across a wide range of departments. This early uptake highlights both the tool's perceived value and the broad relevance across scientific disciplines. The broad user base also suggests strong institutional demand for automated support in navigating shifting grant policies.

Limitations

Several limitations should be noted. GrantCheck was developed and evaluated within a single institution using a locally curated dictionary and infrastructure, so its broader applicability should be interpreted with caution. The findings are best understood as an early proof of concept, with broader impact requiring replication and adaptation at other institutions. In addition, our evaluation relied on a small, annotated dataset (n=5) and a single-institution user group (n=25). While this provided adequate precision for pilot usability assessment, the small sample limits statistical power and generalizability. We also did not conduct interannotator agreement testing, which could have quantified the reliability of manual annotations. Future work will therefore involve larger, multi-institutional datasets

XSL•FO
RenderX

and double annotation with formal reliability statistics (eg, Cohen κ) to provide more robust performance benchmarks. The capabilities of the available LLMs on Amazon Bedrock constrain the system's performance. Furthermore, while the regular expression engine reduces false positives and provides deterministic guardrails, it may miss nuanced or contextually embedded terms. The reliance on a curated dictionary also means updates must be manually maintained to align with policy changes.

GrantCheck is designed to flag all terms included in the institution's approved sensitive-term dictionary, regardless of surrounding context. This approach reflects current policy conditions in which the mere presence of certain terms—whether used in a benign or technical sense—can lead to administrative hold or outright rejection before scientific review. As such, contextual disambiguation was deliberately excluded to minimize the risk of underdetection. While this may result in flagging terms that are not problematic in some settings, it ensures comprehensive identification of language that could trigger compliance concerns under prevailing interpretations.

Another known limitation arises when a document contains a large number of sensitive terms, as demonstrated in our experiments. In such cases, the regex engine successfully identifies all instances and extracts their corresponding sentence contexts; however, when this volume of information is passed to the LLM for alternative language suggestions, the long input may exceed the model's optimal processing capacity. As a result, the LLM may fail to generate recommendations for some flagged terms due to input overload. Our testing showed that, for Amazon Nova Pro, this effect typically emerges when the combined term-context segments exceed approximately 2500 words. This issue is less probable in proposals with few sensitive terms. This 2500-word limit now serves as a benchmark for future optimization strategies, such as batching flagged terms or staging recommendations, to maintain high completion rates even in proposals with a large number of sensitive terms. Furthermore, researchers can iteratively resubmit revised versions of the proposal after addressing some flagged terms, thereby reducing input length and improving LLM response accuracy in subsequent passes. Importantly, even if a term is omitted from the recommendation output, it still appears in the regex findings window, allowing researchers to manually locate and revise it as needed. This ensures that no term is entirely lost in the review process, even if the LLM fails to return a suggestion.

Another limitation of GrantCheck is that its current implementation is tailored to the specific technical and policy environment of our institution. Model choice is constrained by security approvals, which until recently permitted only Amazon Nova and Llama 3.1 within our private VPC; however, at the time of this publication, our institution has obtained approval for access to GPT-4o, which we anticipate will enhance coverage and contextual accuracy in future iterations. Policy-specific term lists are likewise institution-dependent: while our curated dictionary was expanded through faculty council review and local policy interpretations, its foundation is drawn primarily from the Senate report, and broader use would require adapting or rebuilding the list in accordance with that report and the regulatory landscape at other organizations. Security requirements also constrain portability, as our deployment relied on an isolated AWS VPC and institutionally mandated SSO, which may differ across sites. While these constraints limit immediate transferability, the underlying architecture is flexible and can be adapted by other institutions through integration of their own approved models, curated term dictionaries, and security controls. Although GrantCheck was developed within the technical and policy context of a single US academic medical center, the underlying architecture is transferable. Other institutions could adapt the approach by substituting their own sensitive-term lists, approved LLMs, and security controls. Our findings, therefore, demonstrate feasibility but do not imply universal readiness.

A separate limitation involves access to more advanced LLMs. Due to our institution's information security clearance regarding which LLMs we can use safely, the most powerful AI models accessible are Amazon Nova and Llama 3.1. These models provide a lower level of performance when compared with more advanced models such as GPT-4o. We believe that GrantCheck's recommendation quality—both in terms of coverage and the appropriateness of suggested alternatives—will improve with access to more powerful models. To address this limitation and expand functional capabilities, our next stage of development will extend GrantCheck into a multiagent architecture. In this design, 1 subagent will retain the current role of detecting policy-sensitive terms, while a second subagent and its subagents will evaluate whether the grant proposal aligns with administrative and funding agency priorities. This second subagent will leverage tool-calling functions to provide targeted guidance on sections that may benefit from additional emphasis or improvement. A consolidating agent will then synthesize the outputs from these subagents into a unified, prioritized set of recommendations. This roadmap aims to address current limitations by expanding beyond term detection to offer richer, policy-aware feedback, while maintaining the security and compliance features of the existing system.

An additional limitation is that GrantCheck has not yet been formally benchmarked against the Coalition for Health AI [27] assurance framework, which emphasizes transparency, fairness, safety, and continuous monitoring in health-related AI applications. While current safeguards, such as VPC isolation, file deletion, and access controls, provide a strong technical foundation, future work will require mapping GrantCheck's design and evaluation more explicitly to Coalition for Health AI principles to ensure alignment with emerging best practices in health care AI governance. Importantly, UMass Chan Medical School is home to the Massachusetts AI Assurance for Healthcare Laboratory [28], a nationally recognized initiative dedicated to advancing responsible and trustworthy AI in health care. Future iterations of GrantCheck will draw on this institutional expertise to strengthen assurance, benchmarking, and compliance with evolving governance standards. In addition, we plan to expand benchmarking and validation through multisite collaborations facilitated by the Massachusetts AI Assurance for Healthcare Laboratory at UMass Chan Medical School. These collaborations will enable the use of larger, more diverse document sets and external user groups, which are

XSL•FO
RenderX

essential for establishing the external validity and generalizability of GrantCheck across different institutional contexts.

Currently, the tool identifies and suggests replacements for sensitive language, but future updates could also help researchers assess whether their proposals align with funding agency priorities. The tool's underlying architecture is designed to be extensible, enabling future capabilities such as section prioritization, sentiment analysis, and policy-aware rewriting. Once licensing and infrastructure allow, expanding model options beyond Nova and Llama may further improve performance and adaptability.

One important practical note is that the DataFrame must be converted to a string before passing it to an LLM. Feeding a raw DataFrame directly to the LLM can result in loss of structural or contextual information, which is a limitation of how DataFrames are represented in memory. To preserve the data's integrity and readability, it is recommended to use the .to_string() function to transform the data frame to text before submitting it to the model.

Despite these limitations, GrantCheck represents a step forward in operationalizing AI to support research compliance and integrity. Tools like GrantCheck can play a critical role in advancing scientific innovation by enabling researchers to adapt to shifting policies and funding agency priorities while preserving the meaning of their work.

Beyond technical performance, GrantCheck was developed with careful attention to ethical and governance considerations. The system operates solely in an advisory role, supporting researchers in identifying language that may warrant review under evolving administrative guidance. All suggestions are optional and remain under full user discretion, ensuring that authors maintain complete control and academic autonomy. The term dictionary and recommendation framework were developed through transparent institutional review and faculty consultation to uphold neutrality and accountability. The tool does not block, alter, or report content and operates entirely within a secure computing environment. To ensure that the system remains aligned with its advisory purpose, user feedback is actively collected and incorporated into regular updates. Because policies and interpretations evolve over time, GrantCheck is designed to be policy-agnostic and adaptable, allowing institutions to adjust its scope as guidance changes. In this way, the system promotes clarity, consistency, and efficiency in grant writing while safeguarding academic independence and avoiding any form of viewpoint enforcement or censorship.

From a health systems and translational research perspective, the implications of GrantCheck extend well beyond administrative compliance. Language restrictions and shifting policy interpretations can have tangible downstream effects on biomedical research, from delayed project initiation to the loss of funding continuity for critical studies addressing public health challenges. This capability is particularly relevant for proposals in sensitive domains—such as social determinants of health, behavioral medicine, and community-based research—that are most vulnerable to evolving policy restrictions. By proactively identifying and mitigating noncompliant language, GrantCheck helps preserve the operational continuity of research programs across a wide spectrum of domains—including infectious disease surveillance, cancer prevention, precision medicine, behavioral health, and health equity. This is particularly important for proposals that engage underserved or marginalized populations, where delayed or rejected applications can exacerbate existing disparities in research representation and access to innovation.

Furthermore, ensuring compliant grant communication supports institutional stability by protecting revenue streams that maintain clinical research infrastructure, data registries, and translational science programs. In large health systems, even short-term disruptions in funding can slow recruitment, delay protocol activation, and diminish capacity to generate evidence that informs care delivery. GrantCheck thus functions as an upstream enabler of public health resilience—helping academic medical centers maintain continuity in studies that contribute to disease prevention, health system innovation, and evidence-based policymaking.

At a broader societal level, integrating AI-driven compliance tools into research administration can strengthen national readiness for future public health emergencies. During crises, such as pandemics, when timely funding and cross-sector collaboration are essential, reducing administrative friction in the grant process becomes critical for mobilizing rapid scientific response. By maintaining the integrity and inclusivity of research language under evolving policies, GrantCheck helps ensure that vital investigations—especially those addressing social determinants of health and population-level interventions—can proceed without unnecessary delay or topic avoidance.

Finally, the responsible, privacy-preserving design of GrantCheck offers a model for scalable digital infrastructure that reinforces public trust in AI-assisted health research workflows. By aligning institutional compliance with ethical transparency and investigator autonomy, this approach not only reduces administrative burden but also strengthens the long-term sustainability and fairness of the biomedical research enterprise. Together, these features position GrantCheck as an enabling technology that links policy compliance to the broader mission of improving health outcomes, advancing equity, and maintaining scientific innovation.

Conclusions

GrantCheck demonstrates the feasibility of an AI-assisted compliance tool that helps researchers align grant narratives with evolving policy requirements while preserving scientific integrity. Its secure, hybrid regex–LLM architecture and institutional deployment model offer a replicable framework for responsible AI adoption in academic research settings. Future work will expand validation across institutions, integrate stronger assurance benchmarking, and explore multiagent architectures to enhance contextual sensitivity, scalability, and policy adaptability.

XSL•FO
RenderX

Acknowledgments

We thank the researchers who provided thoughtful review and feedback on the tool. We also gratefully acknowledge the AWS team for their technical contributions and valuable suggestions.

ChatGPT was used to improve the clarity, readability, and structure of the manuscript. It was not used for data analysis, interpretation, or result generation. All authors reviewed and verified the accuracy of all artificial intelligence–assisted text. The authors take full responsibility for the content of this publication.

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Data Availability

The deidentified survey data generated and analyzed during this study are available in Multimedia Appendix 2.

Authors' Contributions

Conceptualization: AHZ, QS
Data Curation: AHZ, QS
Formal Analysis: QS
Investigation: QS
Methodology: AHZ, QS, AO, GM, JB, HJ, SKW, PL, AZ, BC, KL
Supervision: AHZ, AO, KL
Writing – original draft: QS
Writing – review & editing: QS, AO, GM, AHZ, KL

Conflicts of Interest

None declared.

Multimedia Appendix 1
System usability survey.
[DOCX File , 319 KB - formative_v9i1e79038_app1.docx]

Multimedia Appendix 2
Survey Data.
[XLSX File (Microsoft Excel File), 16 KB - formative_v9i1e79038_app2.xlsx]

References

1. Ending radical and wasteful government dei programs and preferencing. The White House. 2025. URL: https://www.whitehouse.gov/presidential-actions/2025/01/ending-radical-and-wasteful-government-dei-programs-and-preferencing/ [accessed 2025-10-31]
2. Restoring equality of opportunity and meritocracy. The White House. 2025. URL: https://www.whitehouse.gov/presidential-actions/2025/04/restoring-equality-of-opportunity-and-meritocracy/ [accessed 2025-10-31]
3. Stevens R. Red States Push Forward With DEI Bans in 2025. Duane Morris Government Strategies. 2025. URL: https://statecapitallobbyist.com/dei/red-states-push-forward-with-dei-bans-in-2025/ [accessed 2025-10-31]
4. D.E.I: Division. Extremism. Ideology : How the Biden-Harris NSF Politicized Science. GovInfo. 2024. URL: https://www.govinfo.gov/app/details/GOVPUB-Y4_C73_7-PURL-gpo234941 [accessed 2008-10-31]
5. Miller K, Rabin RC. Ban on D.E.I. language sweeps through the sciences. The New York times. 2025. URL: https://www.nytimes.com/2025/02/09/science/trump-dei-science.html [accessed 2025-10-31]
6. Amendment 109: removing DEIA requirements from ROSES. NASA Science. URL: https://science.nasa.gov/researchers/solicitations/roses-2024/amendment-109-removing-deia-requirements-from-roses-2024/ [accessed 2025-10-31]
7. RESCINDED - Notice of Civil Rights Term and Condition of Award. 2025. URL: https://grants.nih.gov/grants/guide/notice-files/NOT-OD-25-090.html [accessed 2025-10-31]
8. NA. U.S. Department of Education Takes Action to Eliminate DEI. 2025. URL: https://www.ed.gov/about/news/press-release/us-department-of-education-takes-action-eliminate-dei [accessed 2025-10-31]
9. Grammarly. 2025. URL: https://app.grammarly.com/ [accessed 2025-10-31]
10. ChatGPT. URL: https://openai.com/chatgpt/overview/ [accessed 2025-10-31]
11. Seckel E, Stephens BY, Rodriguez F. Ten simple rules to leverage large language models for getting grants. PLoS Comput Biol 2024;20(3):e1011863 [FREE Full text] [doi: 10.1371/journal.pcbi.1011863] [Medline: 38427611]

XSL•FO
RenderX

12. Godwin RC, DeBerry JJ, Wagener BM, Berkowitz DE, Melvin RL. Grant drafting support with guided generative AI software. SoftwareX 2024;27:101784. [doi: 10.1016/j.softx.2024.101784]
13. Meyer JG, Urbanowicz RJ, Martin PCN, O'Connor K, Li R, Peng P, et al. ChatGPT and large language models in academia: opportunities and challenges. BioData Min 2023;16(1):20 [FREE Full text] [doi: 10.1186/s13040-023-00339-9] [Medline: 37443040]
14. Fukataki Y, Hayashi W, Nishimoto N, Ito YM. Developing artificial intelligence tools for institutional review board pre-review: a pilot study on ChatGPT's accuracy and reproducibility. PLOS Digit Health 2025;4(6):e0000695. [doi: 10.1371/journal.pdig.0000695] [Medline: 40587513]
15. Heilmeyer F, Böhringer D, Reinhard T, Arens S, Lyssenko L, Haverkamp C. Viability of open large language models for clinical documentation in german health care: real-world model evaluation study. JMIR Med Inform 2024;12:e59617 [FREE Full text] [doi: 10.2196/59617] [Medline: 39195570]
16. Huang J, Tan M. The role of ChatGPT in scientific communication: writing better scientific review articles. Am J Cancer Res 2023;13(4):1148-1154 [FREE Full text] [Medline: 37168339]
17. Amazon Bedrock. URL: https://aws.amazon.com/bedrock/ [accessed 2025-10-31]
18. re — Regular expression operations. Python documentation. URL: https://docs.python.org/3/library/re.html [accessed 2025-10-30]
19. New Cruz Investigation Reveals How Biden-Harris Diverted Billions from Scientific Research to DEI Activists. U.S. Senate Committee on Commerce, Science, & Transportation. 2024. URL: https://www.commerce.senate.gov/2024/10/new-cruz-investigation-reveals-how-biden-harris-diverted-billions-from-scientific-research-to-dei-activists [accessed 2025-10-31]
20. Stein R. NIH cuts funding for vaccine-hesitancy research. NPR. 2025. URL: https://www.npr.org/2025/03/12/nx-s1-5325863/nih-trump-vaccine-hesitancy-mrna-research [accessed 2025-10-31]
21. Powers DMW. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. Journal of Machine Learning Technologies 2011;2(1):37-63 [FREE Full text]
22. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 2015;10(3):e0118432 [FREE Full text] [doi: 10.1371/journal.pone.0118432] [Medline: 25738806]
23. Efron B, Tibshirani R. An Introduction to the Bootstrap. Philadelphia, PA: Chapman & Hall/CRC; 1994.
24. Fagerland MW, Lydersen S, Laake P. The McNemar test for binary matched-pairs data: mid-p and asymptotic are better than exact conditional. BMC Med Res Methodol 2013;13:91 [FREE Full text] [doi: 10.1186/1471-2288-13-91] [Medline: 23848987]
25. Brooke J. SUS -- a quick and dirty usability scale. In: Usability Evaluation in Industry. Boca Raton, Florida: CRC Press; 1996:189-194.
26. Sauro J. In: Lewis JR, editor. Quantifying the User Experience: Practical Statistics for User Research. United States: Morgan Kaufmann; 2016.
27. Responsible health AI for all. CHAI. URL: https://www.chai.org/ [accessed 2025-10-31]
28. AI assurance lab. UMass Chan Medical School. 2013. URL: https://www.umassmed.edu/ai-assurance-lab/ [accessed 2025-10-31]

Abbreviations

AI: artificial intelligence
AWS: Amazon Web Services
IRB: Institutional Review Board
LLM: large language model
NLP: natural language processing
SSO: single sign-on
SUS: System Usability Scale
VPC: virtual private cloud

XSL•FO
RenderX

Edited by A Mavragani, J Sarvestan; submitted 16.06.25; peer-reviewed by DA Forero, X Zheng; comments to author 06.08.25; revised version received 28.10.25; accepted 29.10.25; published 27.11.25.

Please cite as:
Shi Q, Oztekin A, Matthew G, Bortle J, Jenkins H, Wong S(, Langlois P, Zaki A, Coleman B, Luzuriaga K, Zai AH
GrantCheck—an AI Solution for Guiding Grant Language to New Policy Requirements: Development Study
JMIR Form Res 2025;9:e79038
URL: https://formative.jmir.org/2025/1/e79038
doi:10.2196/79038
PMID:

©Qiming Shi, Asil Oztekin, George Matthew, Jeffrey Bortle, Hayden Jenkins, Steven (Koon) Wong, Paul Langlois, Anaheed Zaki, Brian Coleman, Katherine Luzuriaga, Adrian H Zai. Originally published in JMIR Formative Research (https://formative.jmir.org), 27.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

AI-Powered Thermography for Diabetic Foot Risk Stratification: Multicenter Cross-Sectional Study

Meshari F Alwashmi[1], PhD; Mustafa alghali[1], B.Eng, MSc; Waseem Abu-Ashour[2], MSc, PharmD; Abdullah Mohammed Arabe[3], MSc; Abdullah Al Soheimi[3], CBT, MD; Nasibh Abdulrahman Alharbi[3], MSc; Hani Mohammed Badahdah[3], DPM, MD

[1]Amplifai Health, Riyadh, Saudi Arabia

[2]Memorial University, Health Sciences Centre, St. John's, NL, Canada

[3]Ministry of Health, Riyadh, null, Saudi Arabia

Corresponding Author:
Meshari F Alwashmi, PhD
Amplifai Health
Al Olaya
Riyadh, 12331
Saudi Arabia
Phone: 966 569272937
Email: mfa720@mun.ca

Abstract

Background: Diabetic foot complications are among the most severe and costly outcomes associated with diabetes, with high prevalence particularly in the Middle East and North Africa region. Current screening tools are often limited by subjectivity, invasiveness, or scalability challenges, underscoring the need for innovative approaches.

Objective: This multicenter study aimed to evaluate the performance of an artificial intelligence (AI)–powered thermographic system, Thermal Foot Scan (TFScan), in identifying patients at elevated risk of diabetic foot complications through noninvasive temperature profiling.

Methods: A multicenter cross-sectional analysis of deidentified routine screening data across 4 regions in Saudi Arabia was conducted enrolling 1120 individuals with diabetes. Participants underwent thermal imaging using a smartphone-compatible infrared camera with AI algorithms analyzing angiosomal temperature patterns and asymmetries. Risk was stratified into 4 categories (very low, low, moderate, and high). Associations between TFScan classifications and clinical risk factors, symptoms of neuropathy, and thermal abnormalities were assessed.

Results: While 90.7% (1016/1120) of the participants were classified as very low or low risk, 9.3% (104/1120) were identified as moderate or high risk. This higher-risk group exhibited significantly greater prevalence of key diabetic complications ($P<.001$). Peripheral artery disease was present in 20.2% (21/104) of moderate- and high-risk participants compared to just 0.8% (8/1016) of lower-risk individuals. Cardiovascular disease (60/104, 57.7% vs 313/1016, 30.8%), neuropathy (12/104, 11.5% vs 37/1016, 3.6%), foot deformities (15/104, 14.4% vs 6/1016, 0.6%), and symptoms of loss of protective sensation (53/104, 51% vs 354/1016, 34.8%) were all significantly more frequent in the high-risk subgroup than in the low-risk group, respectively. Thermal imaging further revealed pronounced abnormalities: temperature asymmetries of ≥2.2 °C were observed in 7.1% (79/1120) of the patients overall, with the highest asymmetry and thermal change index scores concentrated in the moderate- and high-risk groups. These individuals also exhibited greater deviations in angiosomal temperature differences—exceeding 2.2 °C in key vascular territories such as the medial plantar and lateral plantar arteries—suggesting both early inflammatory states and critical perfusion deficits.

Conclusions: The TFScan system effectively stratified patients with diabetes into clinically meaningful risk categories, with moderate- and high-risk groups exhibiting a significantly higher burden of vascular, neuropathic, and thermal abnormalities. However, the cross-sectional design, partial reliance on self-report, and low prevalence of advanced complications may limit causal inference. These findings highlight the potential of AI-enhanced thermography to serve as a scalable, objective screening tool for proactive diabetic foot management. Further longitudinal studies are warranted to validate its predictive power and support widespread clinical adoption.

(JMIR Form Res 2025;9:e81289) doi:10.2196/81289

XSL•FO
RenderX

KEYWORDS

artificial intelligence; AI; digital health; thermography; diabetic foot ulcer; risk stratification; screening

Introduction

The Growing Burden of Diabetic Foot Ulcers

Diabetes affects 1 in 10 adults worldwide (537 million), with higher prevalence in the Middle East and North Africa region, affecting 1 in 6 adults (73 million) [1]. Despite advances in medical therapies, the global prevalence of diabetes mellitus is projected to reach 643 million by 2030, with diabetes-related complications also continuing to rise.

One of the most common complications of diabetes is diabetic foot ulcer (DFU). It is estimated that one-third of people with diabetes will develop a DFU during their lifetime [2]. Unfortunately, even after a DFU has been resolved, recurrence is common and is estimated to be 40% within 1 year and 60% within 3 years [3]. Furthermore, diabetes foot care costs are the single largest category of diabetes-related medical costs, accounting for one-third of all diabetes-related costs, and a leading cause of lower-limb amputations [2]. Prevention of these lower-limb complications could lead to a significant positive impact on health outcomes and significant cost savings. Unfortunately, current tools for assessing the risk of developing DFUs have limited scalability in terms of time efficiency and practicality [4].

Limitations of Existing Screening Methods

Traditional tools used to detect diabetic foot complications include the monofilament test, the ankle-brachial index, duplex ultrasonography, and angiography. While widely used, these methods have notable limitations. The monofilament test is designed specifically to assess protective sensations and is not expected to evaluate structural deformities and vascular insufficiencies, with accuracy often dependent on examiner technique [5]. The ankle-brachial index, though noninvasive, is less effective in populations with diabetes due to vessel calcification, which can lead to falsely elevated readings [6,7]. Duplex ultrasonography, while useful for detecting arterial narrowing, is limited in visualizing distal vessels in diabetic limbs [8,9]. Angiography remains the gold standard for vascular imaging but is expensive, time-consuming, and associated with radiation exposure and procedural risks [10]. These limitations highlight the need for a more sensitive, accessible, and objective screening tool.

Human Medical Thermography

Human medical thermography results from decades of research and development in the performance of infrared imaging, standardization of technique, and clinical protocols for thermal imaging [11,12]. It can visualize diseases not readily detected or monitored through other methods. It is a fast, passive, noncontact, and noninvasive imaging method that has been used in numerous peer-reviewed studies [13]. It is currently used worldwide to screen, detect, and monitor diseases. The American Academy of Thermology (AAT) established guidelines for the use of thermography in the evaluation of patients with diabetes. These guidelines provide recommendations for the use of thermal imaging in the detection and monitoring of diabetic neuropathy, including protocols for image acquisition and interpretation [14]. Figure 1 shows the thermal images of the lower extremities of 2 patients with diabetes.

Figure 1. Typical thermal images of the plantar region for healthy, at risk, and amputated patients.

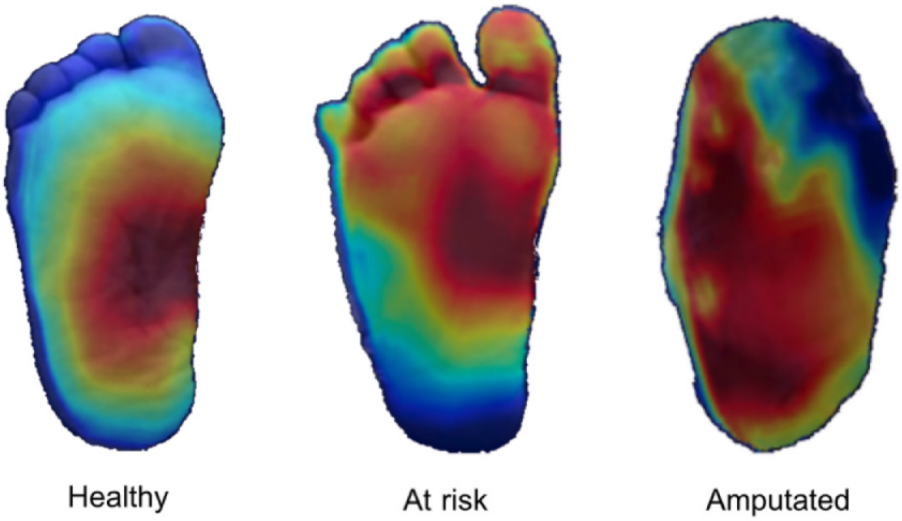

Human medical thermography has many advantages that could encourage widespread adoption. Thermal imaging is relatively inexpensive, compact, and portable; involves no ionizing radiation; and requires little electric power. Recent technological breakthroughs have transformed large and expensive stationary cameras into portable and inexpensive solutions while maintaining quality imaging [15].

A major limitation of the current state of human medical thermography is that even the most skilled human thermographer can only observe, analyze, and successfully interpret a limited number of thermograms. However, computers can process an

XSL•FO
RenderX

image efficiently and extract useful information without tiring. Leveraging artificial intelligence (AI) algorithms, specifically computer vision, can yield findings objectively and minimize interobserver variability. Ongoing progress in software image analysis and reduced reliance on human labor results in faster throughput and centralized processing. This can lead to increased thermographic accuracy and reliability. Nevertheless, computer-aided thermography will require high-level training and experience to ensure quality outcomes.

DFUs and Thermography

Thermography is helpful for the early detection of abnormalities of the foot by analyzing asymmetries and local temperature changes over time. Assessing temperature differences can enable the early detection of ulcers [16,17]. Peripheral vascular disease is a common complication of diabetes that can result in alterations in blood flow that induce changes in skin temperature. These changes in skin temperature may also indicate tissue damage or inflammation resulting from trauma or excessive pressure. The etiology of these traumas is frequently related to moderate repetitive stress that goes unnoticed due to diabetic neuropathy. The application of thermal imaging for the detection of diabetic foot complications is based on the premise that variations in plantar temperature are associated with these types of complications [4,17-23]. Furthermore, there appears to be a positive correlation between BMI and the risk of diabetic foot complications in patients with type 2 diabetes [24,25].

The rapid development of handheld smartphone-based thermal infrared imagers presents a creative solution for predicting DFUs and monitoring existing ones [26]. To address the lack of thermographers, practical computer vision algorithms are needed to automate the process of image acquisition and analysis. These rapidly expanding, low-cost, and widely available resources can help predict one's risk of developing foot ulcers, potentially saving limbs and lives.

AI and its applications are increasingly demonstrating promise in the detection and management of DFUs [27-29]. Diabetes foot syndrome, with its lack of early symptoms and significant impact on patients' quality of life, necessitates the use of AI in timely screening and detection of risk for foot ulcers and possible amputations [28,29].

Studies such as that by Peregrina-Barreto et al [30] have shown the potential of infrared thermography for detecting foot complications in patients with diabetes. Several researchers have demonstrated promising results in the detection of DFUs using machine learning techniques to analyze foot images [31-33]. These studies suggest that AI's application to data derived from thermal plantar images yields promising results, yet several important gaps remain unaddressed. First, most investigations to date have used specialized infrared cameras in controlled environments, but most have not fully capitalized on low-cost smartphone-based thermal imaging for accessible, point-of-care screening. Preliminary work with smartphone infrared cameras has been limited to small feasibility studies rather than integrated AI screening tools. Second, previous AI models have analyzed foot thermograms as whole images or simple regions without segmenting the foot into angiosomal regions, an approach that could pinpoint localized perfusion deficits or inflammation based on vascular anatomy. Finally, clinical scalability remains an open question: existing studies often involve relatively small samples or single-center data, leading to uncertainty about generalizability to larger, real-world populations with diabetes. This emphasizes the need for a validated solution that addresses these limitations.

To our knowledge, no multicenter study has validated a low-cost smartphone-based thermographic imager paired with an AI algorithm across a large population with diabetes. This study addressed this gap by evaluating Thermal Foot Scan (TFScan), a system that leverages AI techniques combined with thermogram images deployed on a low-cost smartphone-based thermal imager and app.

Description of the Technology

We created TFScan, a noninvasive system to identify diabetic foot complications at an early stage. We leveraged off-the-shelf thermal cameras compatible with smartphones or tablets to capture detailed thermal images of participants' feet (Figure 2). We also used AI-based algorithms to perform semantic segmentation and reduce sensor noise in the captured thermograms. The AI models were trained to extract the plantar region as the region of interest and angiosome regions per foot and suppress any background or sensor noise. In addition, the AI was able to detect asymmetric thermal emission of ≥2.2 °C, which can be indicative of pathology in a properly cooled participant [16,34-39].

The system used AI to process images efficiently and extract useful information. It made the findings more objective and minimized interobserver variability. This could lead to faster throughput via centralized cloud-based processing in which samples are anonymized by removing identifiable information from the data, thereby increasing thermographic accuracy and reliability.

The technology analyzes thermal images captured from a specific thermal camera model, the FLIR ONE Edge Pro. We selected the FLIR ONE Edge Pro for its practicality and ease of use in clinical settings. While its resolution is lower than the AAT recommendation of 320 × 240 pixels, several studies have successfully used the FLIR ONE, which has an even lower resolution of 80 × 60 pixels, and concluded that it can effectively capture thermal signals indicative of DFUs [15,37,40,41]. The software identifies temperature variations consistent with inflammation and ulcerative patterns, signaling potential DFUs. The system records, stores, and transmits use events from thermal cameras to a remote storage system.

XSL•FO
RenderX

Figure 2. Thermal imaging device.

Study Objectives

The primary goal of this study was to evaluate the effectiveness of an AI-powered thermographic screening tool in identifying individuals at increased risk of diabetic foot complications through noninvasive, multiregional foot temperature analysis. We also explored the following secondary objectives:

1. To describe the demographic, clinical, and comorbidity profiles of individuals with diabetes enrolled across multiple health care centers in Saudi Arabia
2. To assess the relationship between AI-derived risk classifications and traditional diabetic foot risk factors, including vascular and neuropathic complications
3. To explore temperature-based correlates of AI-predicted risk, including regional foot temperature trends and angiosome-level thermal asymmetry metrics

Methods

Study Design and Population

This study was a multicenter cross-sectional analysis of deidentified routine screening data aligned with the Strengthening the Reporting of Observational Studies in Epidemiology reporting guidelines [42]. This study was conducted across diabetes clinics located in Riyadh, Madinah, Qassim, and Al-Baha in the Kingdom of Saudi Arabia. The dataset included participants who were aged >18 years and had been diagnosed with diabetes (type 1 or type 2). Eligibility for screening was assessed through self-report during the routine clinical workflow, where patients who had a visible foot pathology, such as visible ulcers, infections, or amputations, were excluded. In addition, patients who were unable to stand without assistance were excluded due to the higher risk of falling or injuring themselves. Eligible patients were routinely scanned by the nursing team as part of standard diabetic foot screening procedures. Those who met the inclusion criteria proceeded directly to thermal imaging during their clinic visit.

Data Collection

Overview

Data were collected between February 2025 and April 2025 from 1120 patients with diabetes mellitus (types 1 and 2) across multiple diabetic primary care centers. Patients completed a questionnaire that collected demographic information and self-reported clinical data. In addition, the clinical team at each center conducted AI-augmented thermographic imaging of both feet for each participant.

Self-Reported Questionnaires

Trained health care professionals collected demographic information through a self-reported questionnaire, including anthropometric measurements (height, weight, and BMI), age, gender, lifestyle factors (smoking status and physical activity level), and diabetes-specific information (type and duration of disease). It also included risk factors such as peripheral artery disease (PAD), neuropathy, diabetes-related complications (retinopathy, chronic kidney disease, and renal failure), foot-specific conditions (onychomycosis, gangrene, and Charcot foot), cardiovascular comorbidities, hypertension, and glycated hemoglobin levels from recent laboratory test results. Additionally, health care professionals documented loss of protective sensation (LOPS) symptoms reported by patients, including unsteadiness in walking, pain and tenderness, numbness, and tingling or prickling sensations, using standardized protocols for comprehensive LOPS assessment.

AI-Augmented Thermographic Assessment

Thermal imaging was performed by trained health care professionals who followed a standardized protocol in accordance with the AAT point-of-care guidelines [43]. Patients were acclimatized in a temperature-controlled environment (22-24 °C), after which plantar foot surfaces were captured using a calibrated thermal infrared camera. The captured images were processed through the AI algorithm to extract quantitative thermal parameters, with temperature readings recorded for specific angiosomal regions corresponding to the lateral calcaneal artery, medial calcaneal artery, lateral plantar artery (LPA), and medial plantar artery (MPA), whereas mean foot temperatures were calculated for each foot and temperature asymmetries between corresponding regions of the right and left feet were computed. The AI system analyzed regional thermal distribution patterns for abnormalities, identified hyperthermia and hypothermia in specific angiosomes, quantified temperature asymmetries (with significant asymmetry defined as a ≥2.2 °C difference between corresponding regions), and calculated thermal change index (TCI) scores based on proprietary algorithms.

XSL•FO
RenderX

Data Quality Assurance

To ensure data integrity and reliability, all clinical assessments were conducted by trained health care professionals following standardized protocols [43], whereas thermal imaging was conducted under controlled environmental conditions to minimize external temperature influences. The AI algorithm underwent regular calibration and validation to maintain accuracy, and all data were entered into a secure electronic database with built-in validation checks to identify outliers and inconsistencies. Random audits of 20% of cases were performed to verify data accuracy throughout the collection process. This comprehensive data collection approach enabled thorough evaluation of the AI-augmented thermography system's effectiveness in detecting DFU risk while facilitating analysis of correlations among thermal findings, clinical risk factors, and severity scores.

Ethical Considerations

This study was a secondary analysis of deidentified data obtained during routine clinical procedures and was not part of a prospective research study. All methods were carried out in accordance with relevant guidelines and regulations. The analysis was reviewed and approved by the Ministry of Health Ethics Board, Research Ethics Committee, General Directorate of Research and Studies (H-01-R-009), which waived the requirement for individual informed consent due to the retrospective nature and deidentification of the data. This study adhered to the Declaration of Helsinki and JMIR and Committee on Publication Ethics guidelines on research ethics, privacy, and confidentiality. No additional procedures beyond routine care were performed, and no compensation was provided.

Experimental Equipment and Procedure

The results of infrared thermography can be influenced by various environmental, individual, and technical factors that affect human skin. To obtain accurate results, the thermal images of the participants in the study were obtained in compliance with the protocols and guidelines set by the AAT [14].

The imaging system records and analyzes thermal asymmetry, temperature distribution, and patterns indicative of pathology. Images were classified based on predefined thresholds for asymmetry (≥2.2 °C) and abnormal thermal patterns. Asymmetric thermal emission of ≥2.2 °C can be indicative of pathology in a properly cooled patient [14].

Outcome Measures

The primary outcome was the distribution of TFScan-derived thermal risk categories (very low, low, moderate, and high risk) based on clinical risk factors, foot temperature asymmetry, regional thermal patterns, and angiosomal temperatures. Secondary outcomes included (1) the prevalence of thermal abnormalities (hyperthermia, hypothermia, asymmetry of ≥2.2 °C, or abnormal distribution patterns), (2) associations between TFScan risk categories and clinical risk factors (eg, PAD, neuropathy, and retinopathy), (3) comparison of thermal metrics (eg, TCI score, temperature difference, and mean foot and angiosomal temperatures) by clinical condition and risk group, and (4) identification of thermal phenotypes using unsupervised clustering and heat map analysis. These measures evaluated the clinical relevance of TFScan in stratifying diabetic foot risk and characterizing underlying thermal patterns.

Data Management and Security

All collected data were stored on a password-protected cloud using unique, nonidentifying participant codes. The secure cloud databases are compliant with security standards for access control, data privacy, and encryption requirements. Data were checked for completeness and accuracy, and inconsistencies were resolved through discussion with the research team.

Statistical Analysis Plan

This was an analytical and exploratory study. We summarized cohort demographics, clinical risk factors, and TFScan-derived thermal variables using appropriate descriptive statistics. Means, SDs, medians, and IQRs were calculated for continuous variables (eg, age, BMI, glycated hemoglobin level, and TCI score), and counts and proportions were used for categorical variables (eg, gender, diabetes type, PAD, and neuropathy). We conducted bivariate analyses to explore associations between TFScan risk categories and clinical variables. Chi-square or Fisher exact tests were used for categorical comparisons, and Kruskal-Wallis tests were used to compare continuous thermal metrics between groups (eg, patients with and without PAD). Results were presented in stratified summary tables by TFScan severity group and clinical condition.

Results

Study Population

A total of 1120 individuals with diabetes were enrolled across 4 major regions in Saudi Arabia. The largest proportion of participants was recruited from Riyadh (496/1120, 44.3%), followed by Qassim (305/1120, 27.2%), Madinah (271/1120, 24.2%), and Al-Baha (48/1120, 4.3%), reflecting broad geographic representation. The mean age of the participants was 52.2 (SD 15.85) years, with 73.7% (826/1120) aged >45 years. The cohort was predominantly female (705/1120, 62.9%) and exhibited a mean BMI of 30.62 (SD 6.96) kg/m^2. The average duration of diabetes was 12.87 (SD 8.57) years, with 63.1% (707/1120) of participants living with diabetes for ≥10 years. Type 2 diabetes accounted for 70% (784/1120) of cases.

Thermal Imaging Findings

Thermal imaging revealed several distinct abnormalities. Hyperthermia, defined as elevated localized temperature, was observed in 5.9% (66/1120) of participants, whereas hypothermia was present in 2.1% (23/1120). In addition, 7.1% (79/1120) of individuals demonstrated clinically significant temperature asymmetry, defined as a difference of ≥2.2 °C between corresponding angiosomes on the left and right foot. The mean temperature difference between feet was 0.55 (SD 0.56) °C.

Temperature asymmetries were also examined at the angiosome level. While mean differences were generally under 1 °C, Figure 3 illustrates that a substantial subset of participants exhibited temperature differences exceeding 2 to 3 °C. These outliers may represent individuals at risk of localized inflammation or early tissue compromise not yet detectable through standard clinical

XSL•FO
RenderX

examination. Figure 3 shows the distribution of angiosome-level temperature differences between corresponding anatomical regions of the foot.

Figure 3. Distribution of angiosome-level temperature differences between feet. LCA: lateral calcaneal artery; LPA: lateral plantar artery; MCA: medial calcaneal artery; MPA: medial plantar artery.

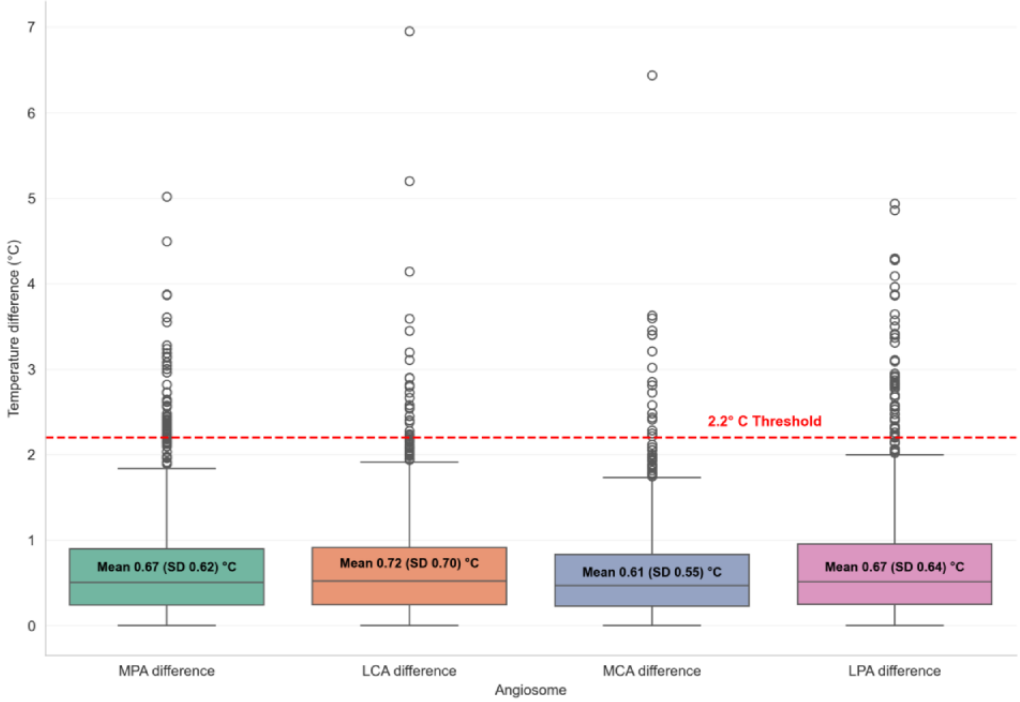

Comorbidities

Several chronic complications commonly associated with diabetes were documented in the study cohort. Cardiovascular disease was the most frequently reported comorbidity, affecting 33.3% (373/1120) of participants. Retinopathy was present in 9.6% (108/1120), PAD was present in 2.6% (29/1120), and chronic kidney disease was present in 2% (22/1120). Renal failure was identified in 0.6% (7/1120) of cases, and Charcot foot was diagnosed in 0.1% (1/1120) of the participants. These comorbidities, particularly those affecting vascular and neurological function, are recognized contributors to diabetic foot complications.

Symptoms of Sensory Loss (LOPS)

Symptoms associated with sensory neuropathy were reported by a substantial proportion of participants. A total of 36.3% (407/1120) experienced at least one symptom consistent with LOPS. The most common symptom was numbness (281/1120, 25.1%), followed by tingling or prickling sensations (185/1120, 16.5%), pain or tenderness (157/1120, 14%), and unsteadiness while walking (82/1120, 7.3%). These sensory abnormalities suggest subclinical neuropathic involvement, which often precedes overt ulceration.

AI-Based Risk Classification

Using the AI-derived classification system, 13.9% (156/1120) of the participants were categorized as very low risk, 76.8% (860/1120) were categorized as low risk, 4.3% (48/1120) were categorized as moderate risk, and 5% (56/1120) were categorized as high risk. When grouped into binary categories, 90.7% (1016/1120) were considered lower risk (very low+low), whereas 9.3% (104/1120) were flagged as higher risk (moderate+high). The detailed demographic and clinical characteristics of the study population are summarized in Table 1.

XSL•FO
RenderX

Table 1. Demographic and clinical characteristics (N=1120).

Characteristic	Values
Study center, n (%)	
Riyadh	496 (44.3)
Madinah	271 (24.2)
Qassim	305 (27.2)
Al-Baha	48 (4.3)
Age (years), mean (SD)	52.2 (15.85)
<45, n (%)	240 (26.2)
45-60, n (%)	446 (39.8)
>60, n (%)	380 (33.9)
Sex, n (%)	
Female	705 (62.9)
Male	415 (37.1)
Anthropometrics, mean (SD)	
BMI (kg/m^2)	30.62 (6.96)
Weight (kg)	78.48 (18.61)
Height (cm)	160.25 (9.1)
Duration of diabetes (years), mean (SD)	12.87 (8.57)
<5, n (%)	215 (19.2)
5-10, n (%)	197 (17.6)
10-20, n (%)	439 (39.1)
≥20, n (%)	269 (24)
Diabetes type, n (%)	
Type 1	336 (30.0)
Type 2	784 (70.0)
Smoker, n (%)	75 (6.7)
Physical activity, n (%)	
Never	345 (30.8)
Rarely	240 (21.4)
Once a week	142 (12.7)
Several times a week	61 (5.4)
Daily	291 (26.0)
Missing	41 (3.7)
Thermal readings (°C), mean (SD)	
Temperature difference (between feet)	0.55 (0.56)
TCI[a]	0.67 (0.48)
Right foot temperature	27.44 (2.67)
Left foot temperature	27.40 (2.64)
Right MPA[b] temperature	28.04 (2.87)
Left MPA temperature	27.95 (2.88)
Right LCA[c] temperature	27.57 (2.53)
Left LCA temperature	27.63 (2.46)

XSL•FO
RenderX

Characteristic	Values
Right MCA[d] temperature	28.32 (2.40)
Left MCA temperature	28.39 (2.38)
Right LPA[e] temperature	27.33 (3.06)
Left LPA temperature	27.27 (2.99)
MPA difference (between right and left angiosomes)	0.67 (0.64)
LCA difference (between right and left angiosomes)	0.67 (0.62)
MCA difference (between right and left angiosomes)	0.61 (0.55)
LPA difference (between right and left angiosomes)	0.72 (0.70)
Thermal abnormalities, n (%)	
Hyperthermia (any angiosome)	66 (5.9)
Hypothermia (any angiosome)	23 (2.1)
Asymmetry of ≥2.2 °C (any foot)	79 (7.1)
Hyperthermia in right foot angiosomes	30 (2.7)
Hyperthermia in left foot angiosomes	36 (3.2)
Hypothermia in right foot angiosomes	12 (1.1)
Hypothermia in left foot angiosomes	11 (1.0)
Clinical risk factors	
HbA_{1c}[f] level (%), mean (SD)	8.23 (2.00)
Cardiovascular disease, n (%)	373 (33.3)
PAD[g], n (%)	29 (2.6)
Neuropathy, n (%)	49 (4.4)
Retinopathy, n (%)	108 (9.6)
CKD[h], n (%)	22 (2.0)
Renal failure, n (%)	7 (0.6)
History of diabetes of ≥10 y, n (%)	707 (63.1)
Charcot foot, n (%)	1 (0.1)
Foot deformity, n (%)	21 (1.9)
LOPS[i] symptoms, n (%)	
Any symptom	407 (36.3)
Numbness	281 (25.1)
Tingling or prickling	185 (16.5)
Pain or tenderness	157 (14.0)
Unsteady walking	82 (7.3)
DFU[j] risk category, n (%)	
Very low	156 (13.9)
Low	860 (76.8)
Moderate	48 (4.3)
High	56 (5.0)
Binary risk classification, n (%)	
Lower risk (very low+low)	1016 (90.7)

XSL•FO
RenderX

Characteristic	Values
Higher risk (moderate+high)	104 (9.3)

[a]TCI: thermal change index.
[b]MPA: medial plantar artery.
[c]LCA: lateral calcaneal artery.
[d]MCA: medial calcaneal artery.
[e]LPA: lateral plantar artery.
[f]HbA_{1c}: glycated hemoglobin.
[g]PAD: peripheral artery disease.
[h]CKD: chronic kidney disease.
[i]LOPS: loss of protective sensation.
[j]DFU: diabetic foot ulcer.

Clinical and Protective Sensation Correlates of AI-Predicted Diabetic Foot Risk

To evaluate the clinical correlates of AI-derived risk categories, we compared the prevalence of comorbidities and LOPS symptoms between participants classified as lower risk (very low risk+low risk) versus those in the higher-risk group (moderate risk+high risk).

Several factors were significantly more prevalent among those flagged as higher risk. PAD was observed in 20.2% (21/104) of higher-risk participants compared to just 0.8% (8/1016) in the lower-risk group (*P*<.001). Similarly, cardiovascular disease affected 57.7% (60/104) of the participants in the higher-risk group, a significant difference from 30.8% (313/1016) in the lower-risk cohort (*P*<.001).

Neuropathy-related indicators were also more frequent among high-risk individuals. Clinically diagnosed neuropathy was nearly 3 times more common (12/104, 11.5% vs 37/1016, 3.6%; *P*<.001), and LOPS symptoms were present in 51% (53/104) of the participants in the higher-risk group versus 34.8% (354/1016) in the lower-risk group (*P*<.001). Moreover, foot deformities were reported by 14.4% (15/104) of the higher-risk participants compared to only 0.6% (6/1016) of the lower-risk participants (*P*<.001). A longer duration of diabetes (≥10 years) was also more common in the high-risk group (77/104, 74% vs 630/1016, 62%; *P*=.02). These findings support the validity of the AI model in identifying individuals with a higher burden of established diabetic foot risk factors even in the absence of active ulceration. The subgroup comparisons are highlighted in Table 2. These differences are further visualized in Figure 4, which shows a consistent pattern of elevated prevalence of multiple comorbid and neuropathic features among individuals classified as moderate or high risk by the AI model.

Table 2. Subgroup comparison by artificial intelligence–classified diabetic foot risk level.

Variable	Moderate and high risk (n=104), n (%)	Very low and low risk (n=1016), n (%)	Prevalence difference (%; 95% CI)	*P* value
PAD[a]	21 (20.2)	8 (0.8)	19.4 (11.7 to 27.1)	<.001
Neuropathy	12 (11.5)	37 (3.6)	7.9 (1.6 to 14.1)	<.001
Retinopathy	9 (8.7)	99 (9.7)	−1.1 (−6.8 to 4.6)	.85
CKD[b]	3 (2.9)	19 (1.9)	1.0 (−2.3 to 4.3)	.73
Diabetes of ≥10 y	77 (74.0)	630 (62.0)	12.0 (3.1 to 21.0)	.02
Charcot foot	1 (1.0)	0 (0.0)	1.0 (−0.9 to 2.8)	.16
Cardiovascular disease	60 (57.7)	313 (30.8)	26.9 (17.0 to 36.8)	<.001
Foot deformity	15 (14.4)	6 (0.6)	14.1 (7.1 to 20.6)	<.001
Any LOPS[c] symptoms	53 (51.0)	354 (34.8)	16.1 (6.1 to 26.2)	.001
Unsteady walking	19 (18.3)	63 (6.2)	12.1 (4.5 to 19.6)	<.001
Pain or tenderness	22 (21.2)	135 (13.3)	7.9 (−0.3 to 16.0)	.04
Numbness	43 (41.3)	238 (23.4)	17.9 (8.1 to 27.7)	<.001
Tingling or prickling	25 (24.0)	160 (15.7)	8.3 (−0.2 to 16.8)	.04

[a]PAD: peripheral artery disease.
[b]CKD: chronic kidney disease.
[c]LOPS: loss of protective sensation.

XSL•FO
RenderX

Figure 4. Prevalence of clinical and protective sensation factors by artificial intelligence–classified risk group. CKD: chronic kidney disease; DM: diabetes mellitus; LOPS: loss of protective sensation; PAD: peripheral artery disease.

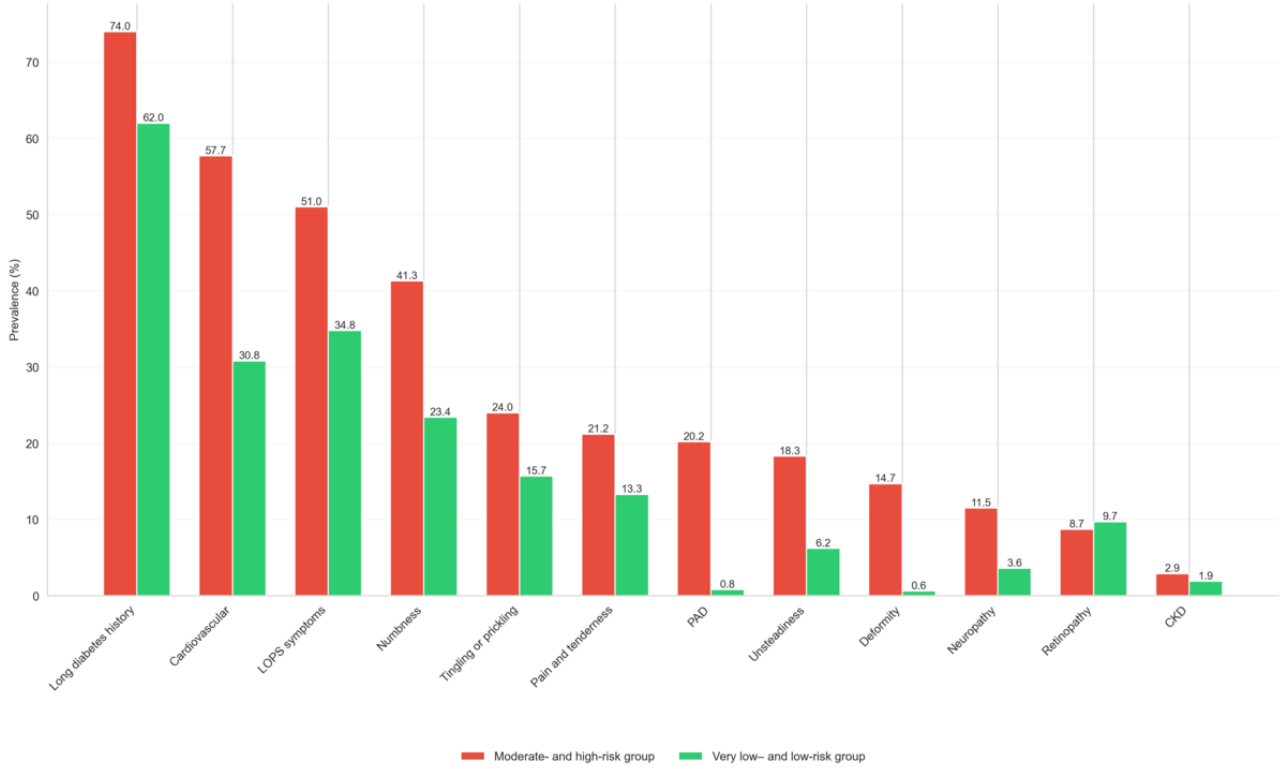

Thermal Pattern Correlates of AI-Predicted Diabetic Foot Risk

To examine how AI-assigned risk categories correspond to objective thermal imaging patterns, we analyzed whole-foot and angiosome-specific temperature metrics across the 4 stratified risk groups.

A clear ascending trend in temperature values was observed from the very low– to moderate-risk categories. For example, mean foot temperatures increased bilaterally from approximately 26.5 °C in the very low–risk group to nearly 28.9 °C in the moderate-risk group. Similarly, temperatures across key angiosomes, including the MPA, lateral calcaneal artery, and medial calcaneal artery, showed rising values across these risk stages. Figure 5 illustrates these dynamic temperature profiles across anatomical regions, reinforcing the value of AI-enhanced thermography in detecting not just risk presence but potentially distinct pathophysiological states within the diabetic foot spectrum.

Figure 5. Mean regional foot temperatures across artificial intelligence–predicted risk categories. LCA: lateral calcaneal artery; MCA: medial calcaneal artery; MFT: mean foot temperature; MPA: medial plantar artery.

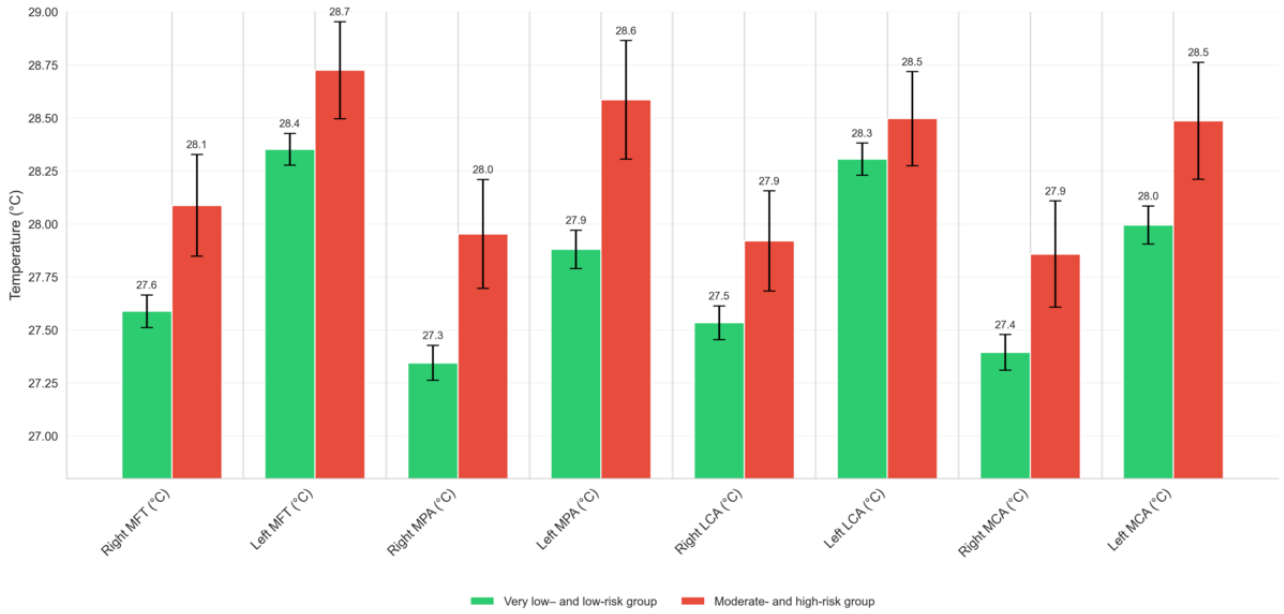

XSL•FO
RenderX

In addition to absolute temperature values, we examined thermal asymmetry and the TCI, 2 key AI-derived metrics. As shown in Table 3, both TCI and temperature differences across corresponding angiosomes (eg, LPA and MPA) increased substantially with increasing risk categories. Notably, the TCI score increased from 0.58 (SD 0.32) in the very low–risk group to 2.05 (SD 0.74) in the high-risk group, indicating significantly greater thermal irregularity. Similarly, LPA and MPA differences exceeded 2 °C in high-risk participants, far above the thresholds typically associated with early inflammatory changes or preulcerative states. These findings are visualized in Figure 6, which highlights the sharp contrast in asymmetry metrics between lower- and higher-risk categories.

All temperature and asymmetry metrics demonstrated statistically significant differences across AI-predicted risk categories as assessed using the Kruskal-Wallis *H* test ($P<.005$ for all comparisons). These findings support the graded relationship between thermal abnormalities and AI-derived diabetic foot risk. A comprehensive summary of both whole-foot and angiosome-specific mean temperatures, as well as thermal asymmetry metrics across all risk categories, is provided in Table 3.

Table 3. Mean foot temperatures and thermal asymmetry metrics by artificial intelligence–predicted diabetic foot risk category.

Variable (°C)	Very low risk, mean (SD)	Low risk, mean (SD)	Moderate risk, mean (SD)	High risk, mean (SD)	*P* value
TCI[a]	0.58 (0.32)	0.58 (0.31)	0.94 (0.55)	2.05 (0.74)	<.001
Temperature difference	0.45 (0.38)	0.46 (0.38)	0.86 (0.65)	2.08 (0.81)	<.001
Right foot temperature	26.53 (2.92)	27.55 (2.61)	28.75 (2.16)	27.09 (2.64)	<.001
Left foot temperature	26.62 (2.86)	27.48 (2.57)	28.93 (2.39)	27.11 (2.53)	<.001
Right MPA[b] temperature	27.11 (3.12)	28.16 (2.79)	29.42 (2.44)	27.69 (2.87)	<.001
Left MPA temperature	27.10 (3.15)	28.02 (2.80)	29.64 (2.55)	27.68 (2.80)	<.001
Right LCA[c] temperature	26.68 (2.80)	27.69 (2.46)	28.76 (1.85)	27.20 (2.61)	<.001
Left LCA temperature	26.92 (2.64)	27.71 (2.40)	28.98 (2.19)	27.32 (2.42)	<.001
Right MCA[d] temperature	27.61 (2.70)	28.43 (2.34)	29.23 (1.83)	27.87 (2.41)	<.001
Left MCA temperature	27.70 (2.66)	28.47 (2.31)	29.63 (1.91)	27.95 (2.40)	<.001
MPA difference (right–left)[e]	0.56 (0.44)	0.58 (0.44)	0.92 (0.74)	2.30 (1.09)	<.001
LCA difference (right–left)	0.63 (0.50)	0.59 (0.46)	0.97 (0.80)	1.81 (1.29)	<.001
MCA difference (right–left)	0.54 (0.44)	0.56 (0.43)	0.77 (0.73)	1.48 (1.22)	<.001
LPA[f] difference (right–left)	0.59 (0.44)	0.60 (0.47)	1.09 (0.83)	2.62 (1.12)	<.001

[a]TCI: thermal change index.
[b]MPA: medial plantar artery.
[c]LCA: lateral calcaneal artery.
[d]MCA: medial calcaneal artery.
[e]Between the right and left angiosomes.
[f]LPA: lateral plantar artery.

XSL•FO
RenderX

Figure 6. Thermal change index (TCI) and regional temperature asymmetries across artificial intelligence–predicted risk categories. LCA: lateral calcaneal artery; LPA: lateral plantar artery; MCA: medial calcaneal artery; MPA: medial plantar artery.

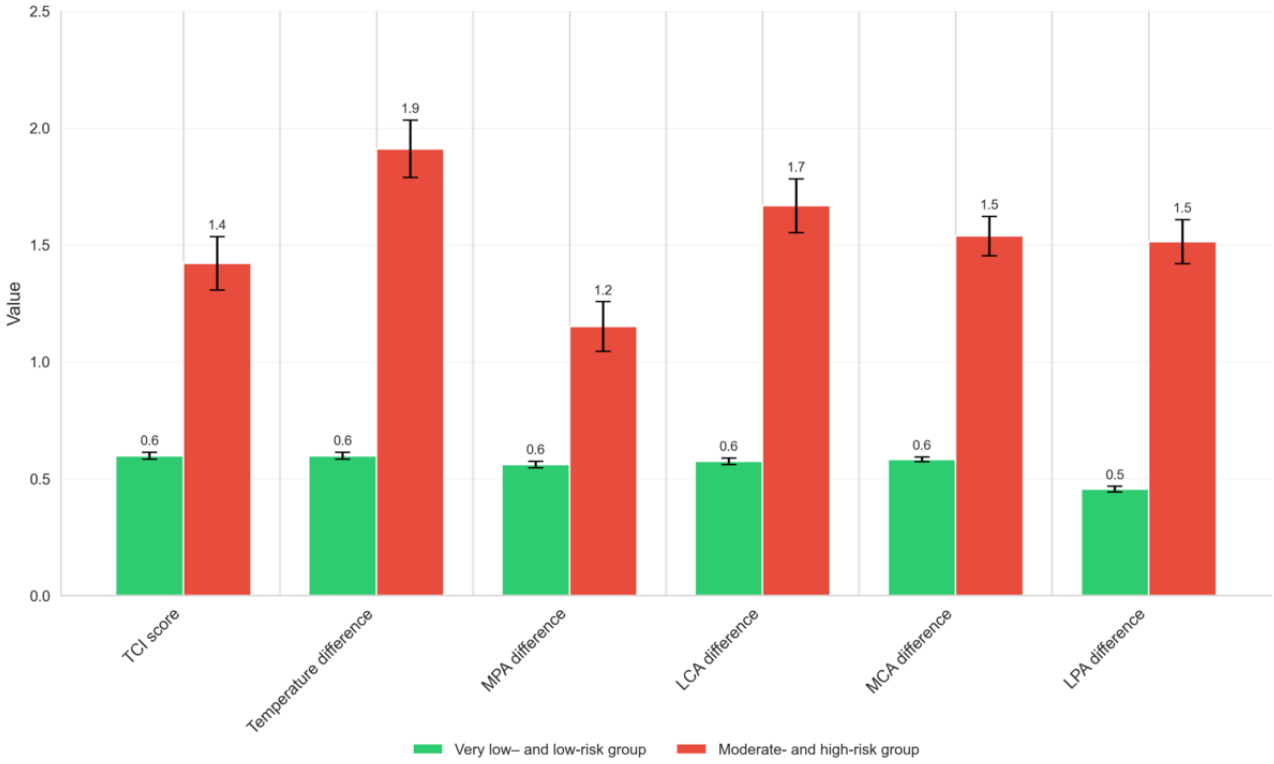

Discussion

Principal Findings

This multicenter, cross-sectional study evaluated the clinical utility of TFScan, an AI-assisted thermographic tool, in detecting thermal abnormalities indicative of early diabetic foot complications. Our results demonstrate that TFScan can stratify patients with diabetes into clinically meaningful risk categories. The distinct thermographic profiles across risk strata and the strong correlation with established clinical risk factors such as PAD, cardiovascular disease, neuropathy, and LOPS support the biological plausibility of TFScan's classifications. Additionally, patients classified as moderate or high risk by the AI algorithm exhibited significantly elevated temperature asymmetry, higher TCI scores, and greater deviations in angiosomal temperatures. These findings suggest that thermal imaging, when integrated with AI algorithms, may offer a scalable and noninvasive approach for early diabetic foot screening in primary care settings.

Comparison With Existing Literature

The significant differences in comorbidity prevalence between risk groups affirm the clinical relevance of TFScan's classifications. PAD and cardiovascular disease were markedly more common in the higher-risk group, consistent with the vascular underpinnings of diabetic foot complications. Neuropathy and LOPS symptoms, including numbness and unsteadiness, were also elevated among high-risk individuals, in line with previous findings that associate sensory loss with foot ulceration risk [4]. However, we extend prior work by evaluating a smartphone-based system across more than 1100 participants.

Mechanistically, hyperthermic patterns likely reflect neuropathic inflammation, whereas marked hypothermia and attenuated angiosomal temperatures may signal ischemic vascular disease. TFScan revealed progressive increases in mean foot and angiosome temperatures from the very low– to moderate-risk groups, followed by a decline in the high-risk category, likely due to underlying hypoperfusion. This nonlinear pattern highlights the algorithm's sensitivity to both hyperthermic inflammation and hypothermic perfusion deficits and, therefore, its ability to detect both neuropathic and vascular complications by integrating both absolute temperatures and asymmetries. The pronounced temperature asymmetry in high-risk participants, particularly in the MPA and LPA angiosomes, aligns with previous reports that asymmetry of ≥2.2 °C is predictive of preulcerative conditions [4,17-23]. Our TCI score results further reinforce thermal irregularity as a viable biomarker.

Clinical Implications

These results suggest that TFScan may serve as a practical adjunct to existing diabetic foot screening methods, especially in resource-limited or high-volume clinical settings, where it could be deployed during routine visits to flag individuals needing vascular assessment or podiatry referral. Its portability, automation, noninvasiveness, and objectivity offer significant advantages over subjective assessments such as the monofilament test, which often suffers from interrater variability and limited sensitivity to early neuropathy or vascular dysfunction. By identifying thermal risk patterns early, TFScan may facilitate timely referral, preventive interventions, and more personalized diabetic foot care. These features suggest strong potential for integration into routine diabetes care, particularly in primary and remote care settings where advanced diagnostics

XSL•FO
RenderX

may not be available. Nevertheless, integration would require training of staff, calibration of devices, and clear referral pathways.

Strengths and Limitations

This study benefited from a large, multicenter design and standardized imaging protocols, enhancing the generalizability of our findings. The use of AI-based image analysis minimized observer bias and enabled quantitative assessment of subtle thermal patterns. Additionally, our inclusion of both clinical and thermographic variables allowed for comprehensive profiling of risk.

Despite these advantages, several limitations must be acknowledged. First, the cross-sectional design precludes causal inference or temporal assessment of ulcer development. Longitudinal follow-up is necessary to determine whether thermal abnormalities predict future foot complications. Second, self-reported data on comorbidities and LOPS symptoms may be affected by recall bias. Moreover, the absence of independent blinding and the relatively low prevalence of advanced complications (eg, PAD and renal failure) in the sample limited statistical power for some comparisons. Finally, while care was taken to standardize image acquisition protocols under controlled environmental conditions, unmeasured confounders (such as prescreening physical activity, ambient room temperature, or skin moisture) may have influenced participants' skin and, therefore, the thermal readings.

These strengths and limitations outline both the promise and the current boundaries of AI-augmented thermographic screening. Future work should build on this foundation through prospective validation studies and technical refinement of the imaging and analysis pipeline.

Future Directions

Future research should evaluate the prognostic utility of TFScan classifications in predicting DFU development and other clinical end points. Longitudinal cohort studies or pragmatic trials integrating TFScan into routine care pathways would be essential to assess real-world effectiveness. Moreover, to address current methodological limitations, subsequent studies should replace self-reported comorbidity and LOPS data with objective measures such as electronic health record linkage and clinician-verified neuropathy or vascular assessments and incorporate independent, blinded outcome assessment to reduce the risk of bias.

Given the low prevalence of advanced complications in this cohort, validation should include enriched and diverse recruitment to ensure adequate event rates, enable subgroup analyses, and strengthen generalizability. Furthermore, efforts should be made to refine the algorithm to incorporate more granular thermal features (eg, localized angiosomal gradients) and explore integration with additional clinical variables or sensor data. Technical refinement should also prioritize robust image acquisition and normalization, for example, the capture of preimaging activity (eg, step count) and skin moisture to allow for covariate adjustment. Finally, implementation science frameworks are needed to study the barriers, facilitators, and cost-effectiveness of TFScan integration at scale.

Conclusions

In this multicenter screening study, AI-assisted thermographic imaging using TFScan effectively stratified patients with diabetes into clinically meaningful risk categories for foot complications. Higher TFScan risk aligns with greater burdens of PAD, cardiovascular disease, and neuropathy and with graded abnormalities in temperature asymmetry, angiosome-level differences, and the TCI, supporting construct validity. However, the cross-sectional design, partial reliance on self-report, and low prevalence of advanced complications may limit causal inference. These findings support the potential role of TFScan as a scalable, noninvasive tool for early identification of patients at risk of diabetic foot complications, meriting further longitudinal validation.

Acknowledgments

The authors would like to express their gratitude to the Ministry of Health and Tamer Mölnlycke Care for their support and guidance throughout this study. They are also deeply appreciative of the staff at the participating diabetic foot centers across 4 health system clusters (King Salman Hospital [Riyadh], King Salman bin Abdulaziz Medical City [Madinah], Diabetic Foot Center [Qassim], and Diabetic Foot Center [Al-Baha]) for their invaluable assistance in the recruitment process. They also want to thank Dr Abdulrahman Alsaqabi, Dr Tarek Kurdi, Dr Fahad Alblaihi, Riad Aqili, and Dr Ahmad Abdulaziz for their tremendous support in facilitating the recruitment and execution of the project. This study was sponsored by Taam and Tamer Mölnlycke Care. The Saudi Ministry of Health supported this study by providing clinical infrastructure and health care personnel to facilitate data collection across the participating diabetic foot centers. Additionally, the authors extend their appreciation to Hussein Ibrahim, Muhammad Alwashmi, Maissa Almagati, Rania Elawad, and Omar Saleh for their contributions to the development of the software used in this research. Moreover, the authors are grateful to all the patients who participated in this study; their willingness to contribute was crucial in advancing the authors' understanding of diabetic foot care. Finally, they acknowledge the use of the Gemini chatbot (Google Workspace with Gemini) for copyediting tasks during the scientific writing process. Specifically, Gemini was used only for proofreading and minor language refinement to improve clarity.

Authors' Contributions

All authors have reviewed and approved the final version of this manuscript.

XSL•FO
RenderX

Conflicts of Interest

MFA and MA are employees and shareholders of Amplifai Health.

References

1. Sun H, Saeedi P, Karuranga S, Pinkepank M, Ogurtsova K, Duncan BB, et al. Erratum to "IDF diabetes atlas: global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045" [Diabetes Res. Clin. Pract. 183 (2022) 109119]. Diabetes Res Clin Pract 2023 Oct;204:110945. [doi: 10.1016/j.diabres.2023.110945] [Medline: 37863776]
2. Armstrong DG, Swerdlow MA, Armstrong AA, Conte MS, Padula WV, Bus SA. Five year mortality and direct costs of care for people with diabetic foot complications are comparable to cancer. J Foot Ankle Res 2020 Mar 24;13(1):16 [FREE Full text] [doi: 10.1186/s13047-020-00383-2] [Medline: 32209136]
3. Armstrong DG, Boulton AJ, Bus SA. Diabetic foot ulcers and their recurrence. N Engl J Med 2017 Jun 15;376(24):2367-2375. [doi: 10.1056/nejmra1615439]
4. Ilo A, Romsi P, Mäkelä J. Infrared thermography and vascular disorders in diabetic feet. J Diabetes Sci Technol 2020 Jan 27;14(1):28-36 [FREE Full text] [doi: 10.1177/1932296819871270] [Medline: 31452395]
5. Wang F, Zhang J, Yu J, Liu S, Zhang R, Ma X, et al. Diagnostic accuracy of monofilament tests for detecting diabetic peripheral neuropathy: a systematic review and meta-analysis. J Diabetes Res 2017;2017:8787261 [FREE Full text] [doi: 10.1155/2017/8787261] [Medline: 29119118]
6. Bhasin N, Scott DJ. Ankle Brachial Pressure Index: identifying cardiovascular risk and improving diagnostic accuracy. J R Soc Med 2007 Jan 01;100(1):4-5 [FREE Full text] [doi: 10.1177/014107680710000103] [Medline: 17197670]
7. Casey SL, Lanting SM, Chuter VH. The ankle brachial index in people with and without diabetes: intra-tester reliability. J Foot Ankle Res 2020 May 12;13(1):21 [FREE Full text] [doi: 10.1186/s13047-020-00389-w] [Medline: 32398142]
8. Eiberg JP, Grønvall Rasmussen JB, Hansen MA, Schroeder TV. Duplex ultrasound scanning of peripheral arterial disease of the lower limb. Eur J Vasc Endovasc Surg 2010 Oct;40(4):507-512 [FREE Full text] [doi: 10.1016/j.ejvs.2010.06.002] [Medline: 20609601]
9. Collins R, Cranny G, Burch J, Aguiar-Ibáñez R, Craig D, Wright K, et al. A systematic review of duplex ultrasound, magnetic resonance angiography and computed tomography angiography for the diagnosis and assessment of symptomatic, lower limb peripheral arterial disease. Health Technol Assess 2007 May;11(20):iii-iv, xi [FREE Full text] [doi: 10.3310/hta11200] [Medline: 17462170]
10. Amin N, Doupis J. Diabetic foot disease: from the evaluation of the "foot at risk" to the novel diabetic ulcer treatment modalities. World J Diabetes 2016 Apr 10;7(7):153-164 [FREE Full text] [doi: 10.4239/wjd.v7.i7.153] [Medline: 27076876]
11. Campbell JS, Mead MN. Human Medical Thermography. Boca Raton, FL: CRC Press; 2022.
12. Serbu G. Infrared imaging of the diabetic foot. In: Proceedings of the 10th Conference on 2009 InfraMation. 2009 Presented at: InfraMation 2009; October 19-22, 2009; Las Vegas, NV URL: https://degroenezuster.nl/wp-content/uploads/2019/11/Thermography-of-Diabetic-foot.pdf
13. Lahiri BB, Bagavathiappan S, Jayakumar T, Philip J. Medical applications of infrared thermography: a review. Infrared Phys Technol 2012 Jul;55(4):221-235 [FREE Full text] [doi: 10.1016/j.infrared.2012.03.007] [Medline: 32288544]
14. Schwartz RG, Brioschi M, O'Young B, Getson P, Bernton T, Brioschi M, et al. The American Academy of Thermology Guidelines for Neuro-Musculoskeletal 2021: infrared medical thermology and sympathetic skin response (SSR) studies. Pan Am J Med Thermol 2022;8:e2021002 [FREE Full text]
15. van Doremalen RF, van Netten JJ, van Baal JG, Vollenbroek-Hutten MM, van der Heijden F. Validation of low-cost smartphone-based thermal camera for diabetic foot assessment. Diabetes Res Clin Pract 2019 Mar;149:132-139. [doi: 10.1016/j.diabres.2019.01.032] [Medline: 30738090]
16. Ilo A. Infrared thermography in vascular disorders - screening and follow-up. University of Oulu. 2020. URL: https://www.oulu.fi/en/theses/infrared-thermography-vascular-disorders-screening-and-follow [accessed 2025-11-13]
17. Nagase T, Sanada H, Takehara K, Oe M, Iizaka S, Ohashi Y, et al. Variations of plantar thermographic patterns in normal controls and non-ulcer diabetic patients: novel classification using angiosome concept. J Plast Reconstr Aesthet Surg 2011 Jul;64(7):860-866. [doi: 10.1016/j.bjps.2010.12.003] [Medline: 21257357]
18. de Deus Passos M, da Rocha AF. Evaluation of infrared thermography with a portable camera as a diagnostic tool for peripheral arterial disease of the lower limbs compared with color Doppler ultrasonography. Arch Med Sci Atheroscler Dis 2022 Aug 8;7(1):e66-e72 [FREE Full text] [doi: 10.5114/amsad/150716] [Medline: 36158072]
19. Zhou Q, Qian Z, Wu J, Liu J, Ren L, Ren L. Early diagnosis of diabetic peripheral neuropathy based on infrared thermal imaging technology. Diabetes Metab Res Rev 2021 Oct 04;37(7):e3429. [doi: 10.1002/dmrr.3429] [Medline: 33307598]
20. Hernandez-Contreras DA, Peregrina-Barreto H, Rangel-Magdaleno JDJ, Renero-Carrillo FJ. Plantar thermogram database for the study of diabetic foot complications. IEEE Access 2019;7:161296-161307. [doi: 10.1109/access.2019.2951356]
21. Sudha BG, Umadevi V, Shivaram JM, Sikkandar MY, Al Amoudi A, Chaluvanarayana HC. Statistical analysis of surface temperature distribution pattern in plantar foot of healthy and diabetic subjects using thermography. In: Proceedings of the 2018 International Conference on Communication and Signal Processing. 2018 Presented at: ICCSP '18; April 3-5, 2018; Chennai, India URL: https://ieeexplore.ieee.org/document/8524310 [doi: 10.1109/iccsp.2018.8524310]

XSL•FO
RenderX

22. Bagavathiappan S, Philip J, Jayakumar T, Raj B, Rao PN, Varalakshmi M, et al. Correlation between plantar foot temperature and diabetic neuropathy: a case study by using an infrared thermal imaging technique. J Diabetes Sci Technol 2010 Nov 01;4(6):1386-1392 [FREE Full text] [doi: 10.1177/193229681000400613] [Medline: 21129334]
23. Papanas N, Papatheodorou K, Papazoglou D, Kotsiou S, Maltezos E. Association between foot temperature and sudomotor dysfunction in type 2 diabetes. J Diabetes Sci Technol 2010 Jul 01;4(4):803-807 [FREE Full text] [doi: 10.1177/193229681000400406] [Medline: 20663441]
24. Renero-Carrillo FJ. The thermoregulation of healthy individuals, overweight–obese, and diabetic from the plantar skin thermogram: a clue to predict the diabetic foot. Diabet Foot Ankle 2017 Aug 16;8(1). [doi: 10.1080/2000625x.2017.1361298]
25. Neves EB, Almeida AJ, Rosa C, Vilaca-Alves J, Reis VM, Mendes R. Anthropometric profile and diabetic foot risk: a cross-sectional study using thermography. Annu Int Conf IEEE Eng Med Biol Soc 2015 Aug;2015:1-3. [doi: 10.1109/EMBC.2015.7445519] [Medline: 28556779]
26. Basatneh R, Najafi B, Armstrong DG. Health sensors, smart home devices, and the internet of medical things: an opportunity for dramatic improvement in care for the lower extremity complications of diabetes. J Diabetes Sci Technol 2018 May 11;12(3):577-586 [FREE Full text] [doi: 10.1177/1932296818768618] [Medline: 29635931]
27. Alwashmi MF, Alghali M, AlMogbel A, Alwabel AA, Alhomod AS, Almaghlouth I, et al. The use of AI-powered thermography to detect early plantar thermal abnormalities in patients with diabetes: cross-sectional observational study. JMIR Diabetes 2025 Jun 13;10:e65209 [FREE Full text] [doi: 10.2196/65209] [Medline: 40512999]
28. Chemello G, Salvatori B, Morettini M, Tura A. Artificial intelligence methodologies applied to technologies for screening, diagnosis and care of the diabetic foot: a narrative review. Biosensors (Basel) 2022 Nov 08;12(11):985 [FREE Full text] [doi: 10.3390/bios12110985] [Medline: 36354494]
29. Kaselimi M, Protopapadakis E, Doulamis A, Doulamis N. A review of non-invasive sensors and artificial intelligence models for diabetic foot monitoring. Front Physiol 2022 Oct 21;13:924546 [FREE Full text] [doi: 10.3389/fphys.2022.924546] [Medline: 36338484]
30. Peregrina-Barreto H, Morales-Hernandez LA, Rangel-Magdaleno JJ, Avina-Cervantes JG, Ramirez-Cortes JM, Morales-Caporal R. Quantitative estimation of temperature variations in plantar angiosomes: a study case for diabetic foot. Comput Math Methods Med 2014;2014:585306 [FREE Full text] [doi: 10.1155/2014/585306] [Medline: 24688595]
31. Cruz-Vega I, Hernandez-Contreras D, Peregrina-Barreto H, Rangel-Magdaleno JD, Ramirez-Cortes JM. Deep learning classification for diabetic foot thermograms. Sensors (Basel) 2020 Mar 22;20(6):1762 [FREE Full text] [doi: 10.3390/s20061762] [Medline: 32235780]
32. Goyal M, Reeves ND, Rajbhandari S, Ahmad N, Wang C, Yap MH. Recognition of ischaemia and infection in diabetic foot ulcers: dataset and techniques. Comput Biol Med 2020 Feb;117:103616 [FREE Full text] [doi: 10.1016/j.compbiomed.2020.103616] [Medline: 32072964]
33. Khandakar A, Chowdhury ME, Ibne Reaz MB, Md Ali SH, Hasan MA, Kiranyaz S, et al. A machine learning model for early detection of diabetic foot using thermogram images. Comput Biol Med 2021 Oct;137:104838 [FREE Full text] [doi: 10.1016/j.compbiomed.2021.104838] [Medline: 34534794]
34. Armstrong DG, Lavery LA. Monitoring neuropathic ulcer healing with infrared dermal thermometry. J Foot Ankle Surg 1996 Jul;35(4):335-372. [doi: 10.1016/s1067-2516(96)80083-4] [Medline: 8872757]
35. Houghton VJ, Bower VM, Chant DC. Is an increase in skin temperature predictive of neuropathic foot ulceration in people with diabetes? A systematic review and meta-analysis. J Foot Ankle Res 2013 Aug 07;6(1):31 [FREE Full text] [doi: 10.1186/1757-1146-6-31] [Medline: 23919736]
36. van Netten JJ, van Baal JG, Liu C, van der Heijden F, Bus SA. Infrared thermal imaging for automated detection of diabetic foot complications. J Diabetes Sci Technol 2013 Sep 01;7(5):1122-1129 [FREE Full text] [doi: 10.1177/193229681300700504] [Medline: 24124937]
37. Vilcahuaman L, Canals R, Canals R, Zequera M, Wilches C, Arista MT, et al. Automatic analysis of plantar foot thermal images in at-risk type II diabetes by using an infrared camera. In: Proceedings of the 2015 World Conference on Medical Physics and Biomedical Engineering. 2015 Presented at: MPBE '15; June 7-12, 2015; Toronto, ON URL: https://link.springer.com/chapter/10.1007/978-3-319-19387-8_55 [doi: 10.1007/978-3-319-19387-8_55]
38. Fraiwan L, AlKhodari M, Ninan J, Mustafa B, Saleh A, Ghazal M. Diabetic foot ulcer mobile detection system using smart phone thermal camera: a feasibility study. Biomed Eng Online 2017 Oct 03;16(1):117 [FREE Full text] [doi: 10.1186/s12938-017-0408-x] [Medline: 28974212]
39. Schaper NC, van Netten JJ, Apelqvist J, Bus SA, Fitridge R, Game F, et al. Practical guidelines on the prevention and management of diabetes-related foot disease (IWGDF 2023 update). Diabetes Metab Res Rev 2024 Mar 27;40(3):e3657 [FREE Full text] [doi: 10.1002/dmrr.3657] [Medline: 37243927]
40. Murphy CA, Laforet K, Da Rosa P, Tabamo F, Woodbury MG. Reliability and predictive validity of inlow's 60-second diabetic foot screen tool. Adv Skin Wound Care 2012 Jun;25(6):261-266. [doi: 10.1097/01.ASW.0000415343.45178.91] [Medline: 22610110]
41. Qin Q, Nakagami G, Ohashi Y, Dai M, Sanada H, Oe M. Development of a self-monitoring tool for diabetic foot prevention using smartphone-based thermography: plantar thermal pattern changes and usability in the home environment. Drug Discov Ther 2022 Sep 17;16(4):169-176 [FREE Full text] [doi: 10.5582/ddt.2022.01050] [Medline: 36002308]

XSL•FO
RenderX

42. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol 2008 Apr;61(4):344-349. [doi: 10.1016/j.jclinepi.2007.11.008] [Medline: 18313558]
43. Guidelines for point of care medical thermography. American Academy of Thermology. URL: https://irp.cdn-website.com/e1c76e4d/files/uploaded/Guidelines-for-Point-of-Care-Medical-Thermography.pdf [accessed 2025-10-09]

Abbreviations

AAT: American Academy of Thermology
AI: artificial intelligence
DFU: diabetic foot ulcer
LOPS: loss of protective sensation
LPA: lateral plantar artery
MPA: medial plantar artery
PAD: peripheral artery disease
TCI: thermal change index
TFScan: Thermal Foot Scan

Edited by J Sarvestan; submitted 25.07.25; peer-reviewed by X Liang; comments to author 11.09.25; revised version received 21.10.25; accepted 29.10.25; published 27.11.25.

Please cite as:
Alwashmi MF, alghali M, Abu-Ashour W, Arabe AM, Al Soheimi A, Alharbi NA, Badahdah HM
AI-Powered Thermography for Diabetic Foot Risk Stratification: Multicenter Cross-Sectional Study
JMIR Form Res 2025;9:e81289
URL: https://formative.jmir.org/2025/1/e81289
doi:10.2196/81289
PMID:

©Meshari F Alwashmi, Mustafa alghali, Waseem Abu-Ashour, Abdullah Mohammed Arabe, Abdullah Al Soheimi, Nasibh Abdulrahman Alharbi, Hani Mohammed Badahdah. Originally published in JMIR Formative Research (https://formative.jmir.org), 27.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A 6-Month Evaluation of the Peer-Ceived Momentary Assessment Method in a Small Sample of Liver Transplant Patients and Their Support Persons: Longitudinal Observational Study

Allan Berrocal[1], BSc, MSc, PhD; Waldo Concepcion[2], MD; Katarzyna Wac[1], BSc, MSc, PhD

[1]Quality of Life Technologies Lab, Center for Informatics, University of Geneva, Geneva, Switzerland

[2]Mohammed Bin Rashid University of Medicine and Health Sciences, Dubai, null, United Arab Emirates

Corresponding Author:
Katarzyna Wac, BSc, MSc, PhD
Quality of Life Technologies Lab
Center for Informatics
University of Geneva
CUI Battelle bat A
Route de Drize 7
Geneva, 1227
Switzerland
Phone: 41 22 379 02 46
Email: katarzyna.wac@unige.ch

Abstract

Background: Patient-reported outcomes, including ecological momentary assessments (EMAs), are acquired from patients via repeated self-reports of their perceived momentary physical and emotional states before and after medical procedures. Patient-reported outcomes are used to measure health outcomes and quality of care. However, certain observable states or behaviors (eg, moods such as fatigue, hope, or medication adherence), or behaviors suggestive of health decline (eg, depression, cognitive decline), are not easily measured via self-reports in certain situations (eg, patients undergoing certain medical procedures, patients with dementia, and others). The peer-ceived momentary assessment (PeerMA) method involves support persons or peers (eg, family members and friends) to report their perception of a patient's subjective physical and emotional states and has been validated in healthy populations.

Objective: We examined the value of the PeerMA method in assessing the disease progression and recovery pathways of patients undergoing liver transplantation. Herein, the PeerMA method is operationalized via the patient's informal caregivers and the patient-based EMA, and wearable-based physical activity datasets from the patients. We report the feasibility results and human factors influencing the acceptance and reliability of the PeerMA method in a small study comprising 8 patients and support persons.

Methods: We conducted a longitudinal observational study of 6 months (autumn 2019 to spring 2020), collecting EMA/self-reports from 8 patients (at the liver transplant clinic at Stanford University Hospital, California) about their perceived levels of hope, sleep, fatigue, depression, and pain in addition to PeerMA-based reports of the same aspects from 7 caregivers. We collected physical activity records from 5 patients using a Fitbit bracelet. Participants completed pre- and poststudy surveys, contributing qualitative data. We implemented the PeerMA method using a smartphone app, making it easy to use by both patients and support persons.

Results: We collected 1142 patient-days and 976 support person–days. On average, each patient received 103 EMAs and responded to 64 (63%) of them, while support persons received 87 PeerMAs and responded to 64 (74%) of them. We report empirical evidence about the methodological feasibility of PeerMA, showing its dual and unique information streams unavailable by EMA alone. We show examples where support person assessments and physical activity data can inform health professionals about the actual state of a patient regarding outcomes such as hope, sleep quality, fatigue, pain, and depression. We discuss human factors influencing the acceptance of the method and make methodological recommendations.

Conclusions: It is possible to leverage data acquired via the PeerMA method and a wearable activity monitor to complement EMA. The PeerMA method incorporates frequent observations from support persons in patients' daily lives, which can be

XSL•FO
RenderX

compared and analyzed next to the patient's self-reports. Such data may help to study and assist patients during disease recovery, which is beneficial for patients recovering from an organ transplant.

(JMIR Form Res 2025;9:e55907) doi:10.2196/55907

KEYWORDS

peer-ceived momentary assessment; PeerMA; ecological momentary assessment; EMA; human state assessment; behavior modeling; human-smartphone interaction; digital health; well-being; mobile phone

Introduction

Background

The ecological momentary assessment (EMA) method was introduced by Stone and Shiffman [1] as a technique to repeatedly collect individual momentary self-assessments for a particular outcome of interest (eg, mood, pain, and happiness), which could be combined with other digital data streams predominantly obtained from smartphone sensors and wearables. An extension of this method, known as peer-ceived momentary assessment (PeerMA), was proposed by Berrocal and Wac [2] and involves peers of the individual (eg, friends or family members) to report peer-based momentary assessments about how they perceive the individual's internal states and external observable behaviors. PeerMA is inspired by the concept of observer and EMA's theoretical and methodological principles. The PeerMA method is a form of EMA where the report is completed by a designated peer of an individual, in the same time observation window as when the individual is prompted to complete his/her own EMA. Peers are defined as close, trusted friends or family members taking the role of observers of the individual completing the EMA. So far, the PeerMA method has been used only in studies involving healthy participants [3]; consequently, the value of PeerMA in patient populations is still an open area of research to which this work contributes.

Related Work

The EMA method has been used in psychology to assess human aspects such as emotional awareness [4], depression [5], happiness [6], or human virtues [7]. In clinical settings [8], EMA was used to assess human states and behavioral disorders such as mood dysregulation, anxiety, substance use, or psychosis [9]. EMA was also used to assess mood and stress changes [10] in patients with chronic conditions such as fatigue, acquired immunodeficiency syndrome, migraine, breast cancer, or kidney disease. Besides health-related uses, the EMA method has been applied to organizational research [11] and in computer science, particularly in the subfields of ubiquitous computing and human-computer interaction [12].

In the literature on health care, the concept of peer is also known as proxy or observer [13] or informant [14]. This concept of peers has been used in multiple areas, including behavior and personality assessment [14-18], chronic stress detection [19], and mental health [20]. In clinical settings, proxies often assess a patient's condition when the patient cannot express himself or herself. For instance, children or patients with dementia [13] or when the patient cannot complete a self-report due to other factors [21]. In clinical care, proxies have been used for rehabilitation and general health [21-24], cancer [25-27], multiple sclerosis, and other chronic diseases [28], among others. Observers can provide assessments about a patient's condition in addition to the patient's self-report. In both clinical and nonclinical scenarios, observers or peers have been documented primarily in studies involving one-time assessments or in longitudinal studies with test and retest periods separated by months or years [29]. In the PeerMA method, peers (observers or support persons) who are usually informal caregivers (eg, friends or family members) contribute more frequently to the patient's care, for example, multiple times per day, which may be important in certain patients. That can be especially important in high-risk populations such as patients undergoing organ transplants. That relationship between the patient and his/her support person could be understood as a form of social support, an important element influencing the patient's quality of life during and after the recovery period [30-33]. Such social support begins with the closest circles, like family and caregivers. Overall, besides complementing someone's self-assessments, peers can also be allies for an individual recovering from an addiction (eg, eating disorders, smoking, and gambling). In this case, peers' assessments could help prevent relapses and their negative consequences.

The work presented in this paper makes a unique contribution by exploring the use of the PeerMA method [2], implying an acquisition of a patient's support persons' assessment to complement the patients' self-reported information obtained via EMA and wearable activity monitors. The combination of these patients' data streams is promising for care personnel to learn more about disease progression and recovery pathways of patients undergoing a liver transplant. This research has foreseeable implications for supporting person-assisted health management programs, ubiquitous technologies, and mobile human-computer interaction.

Study Aims

In this study, we aimed to explore the following questions:

- Aim 1: Is it feasible to leverage the PeerMA method in conjunction with EMA and physical activity and sleep datasets collected from wearables to assess patients' quality of life prior to and after a liver transplant?
- Aim 2: What fundamental human factors drive the patients and support person's acceptance of and sustained engagement with the PeerMA method (as studied by Wulfovich et al [34]).

XSL•FO
RenderX

Methods

Overview

This section describes the experimental design of the longitudinal, observational 6-month study. The protocol was designed in close collaboration with a clinical team.

Study Population

We conducted a longitudinal study of 6 months (autumn of 2019 to spring of 2020) to assess qualitative and quantitative aspects related to the quality of life of adult patients and their caregiving support persons, that is, their family members, neighbors, or friends recruited at the liver transplant clinic of the Stanford University Hospital (Stanford, California). Patients who were in pre- or posttransplant stages would self-report their perceived level of hope [35], quality of sleep [36], fatigue [37], depression [38], and pain [39] using the EMA method. Correspondingly, their support person answered the same questions, leveraging the PeerMA method. Additionally, we collected physical activity and sleep data from those patients (but not from the support persons) using a Fitbit Charge 2 wearable activity monitor.

Recruitment: Patients and Support Persons

During the screening phase of the study (Figure 1), researchers ensured that participants were at least 18 years old and owned a data-enabled smartphone with a suitable OS version. Participants were patients from the transplant clinic at Stanford University Hospital (Stanford), either in the pre or postliver transplant stage, who were recruited during their scheduled visits to the clinic. The patients usually arrive with their support person, who is one of their informal caregivers (eg, family member, neighbor, or friend). As such, the recruitment of each dyad, consisting of one patient and one support person, was facilitated by the medical team at the hospital, where both patients and support persons were available to be informed about the study and to be enrolled if they consented to participate. Overall, 8 patients participated in this study. Some patients were enrolled before receiving the liver transplant (n=3), while others were enrolled posttransplant (n=5). A total of 7 patients had one support person each, and one patient had no support person. It is a coincidence that all the recruited support persons were spouses of the respective patients. Neither patients nor support persons received monetary compensation in this study. Table 1 shows the number of patients and support persons in the study.

Figure 1. Overview of the study design.

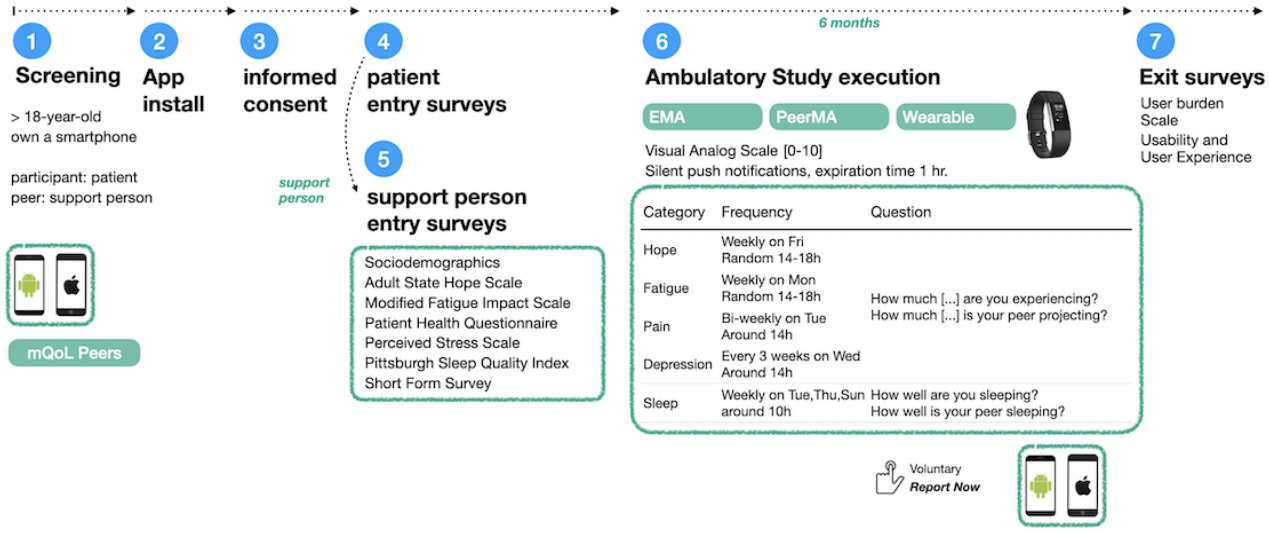

XSL•FO
RenderX

Table 1. Socioeconomic characteristics of the study participants (patients and support persons).

Variables	Patients[a], n (%)	Support persons[a], n (%)
Sex		
Male	7 (88)	1 (14)
Female	1 (13)	6 (86)
Age (years)		
40-49	0 (0)	1 (14)
50-59	1 (13)	1 (14)
60-69	5 (63)	2 (29)
70-79	0 (0)	0 (0)
80 or older	0 (0)	1 (14)
Unknown[b]	2 (25)	2 (29)
Higher education		
High school	1 (13)	1 (14)
Bachelor	1 (13)	1 (14)
Master	2 (25)	2 (29)
Other	2 (25)	1 (14)
Unknown[b]	2 (25)	2 (29)
Employment status		
Employed	1 (13)	2 (29)
Retired	5 (63)	3 (43)
Unknown[b]	2 (25)	2 (29)
Marital status		
Married or domestic partnership	8 (100)	7 (100)

[a]Total number of patients is 8, and the total number of support persons is 7. One patient had no support person.

[b]Did not complete the entry survey.

Proposed Approach

To explore aim 1, we collected quantitative data such as study participants' retention during the study, overall agreement between the EMAs and PeerMAs, and qualitative data for the participant's overall perceived value of the method. The value of the method comes from enabling health care personnel to acquire self-assessments and support person (ie, the observer) assessments paired in time for a more specific assessment of the patient's health outcomes and quality of care. To explore Aim 2, we gathered qualitative data such as study participants' reflections after using the method, perceived difficulty in using the technology, and reliability of the data acquisition technologies (smartphones and wearables) that can influence the quality of the collected data.

Tools

We implemented the PeerMA method by leveraging the mQoL Lab platform [40,41] of the Quality of Life Technologies Lab (University of Geneva). This study used the same toolset from a previous study [3] to collect the EMAs and PeerMAs on the participants' (patients and support persons) own smartphones. Namely, we developed and published a mobile app called mQoL-Peers that was available for Android and iOS, implementing the EMA and PeerMA methods (further explained in Berrocal et al [41]) with configurable EMA/PeerMA questions' content and frequency.

Joining the Study

At the beginning of the study, patients and support persons had a 30-minute face-to-face meeting with a researcher with the following objectives: (1) explain the nature of the study, (2) hand out informed consents, (3) install the mQoL-Peers app and train the participant to use it (concretely, how to complete the entry surveys, and how to address the push notifications for EMA/PeerMA prompts), and (4) answer any questions the participants may have. In particular, we instructed patients to answer the EMA prompts based on how they felt when they completed the EMA. We instructed the support persons to answer the PeerMA prompts based on what they perceived from the patient at the moment when they completed the PeerMA.

Study Surveys and Self-Reports

This section explains the study entry, ambulatory/continual assessments [42], and study exit surveys. The study entry and exit surveys were administered using Qualtrics (Stanford

XSL•FO
RenderX

University), preferred for studies involving clinical patients. The ambulatory EMA and PeerMAs were administered using the mQoL-Peers app.

Study Entry Surveys

The study entry surveys are summarized in Table 2. Patients and support persons completed the entry surveys before beginning the daily EMA/PeerMA. The study entry survey has two parts: (1) socioeconomic status, including gender, age range, education, marital status, and employment status, and (2) standardized questionnaires that provide a quantitative assessment of health-related outcomes for the patients and support persons.

Regarding the standardized patient-reported outcomes–based questionnaires, we used (1) the 6-item Adult State Hope Scale [35] to measure individual's agency and pathways to reach goals in life; (2) the 21-item Modified Fatigue Impact Scale [37], a multiple dimension measure of fatigue including physical, cognitive and psychosocial subscales; (3) the 4-item Patient Health Questionnaire [38] which measures depression and anxiety; (4) the 4-item Perceived Stress Scale [43] to measure self-perceived stress of an individual; (5), the 9-item Pittsburgh Sleep Quality Index [36] to measure quality of sleep in clinical populations, and (6) 20-item Short Form Survey [44].

Table 2. List of study entry surveys with survey construct, survey name, and reference.

Category	Description
Socioeconomic	Patient's profile: gender, age range, education, marital and employment status, support person's profile, and type of relationship with the patient
Hope	6-item Adult State Hope Scale [35]
Fatigue	21-item Modified Fatigue Impact Scale [37]
Depression and anxiety	4-item Patient Health Questionnaire [38]
Stress	4-item Perceived Stress Scale [43]
Sleep quality	12-item Pittsburgh Sleep Quality Index [36]
Medical outcomes	20-item Short Form Survey [44]

Ambulatory Monitoring: EMA/PeerMA Self-Reporting and Wearable Datasets

The ambulatory surveys that continue as repeated measures in the whole study period are summarized in Table 3. We used single-item questions proposed by Rosenzveig et al [39] on a visual analog scale measuring patients' levels of sleep, fatigue, hope, pain, and depression since these aspects reflect relevant physical and psychological states of patients during pretransplant disease progression and posttransplant recovery. With EMA, we asked the following questions: "How well are you sleeping?" and "How much {fatigue, hope, pain, depression} are you experiencing?" Correspondingly, with PeerMA, we asked the support person: "How well is your peer sleeping?" and "How much {fatigue, hope, pain, depression} is your peer projecting?" where "the peer" is the patient. For each PeerMA prompt, support persons were able to indicate how confident they were with each assessment on a visual analog scale from 0 to 1, where 0 means "not confident at all," and 1 means "fully confident." The selected questions can be studied using introquestive methods [45], and they influence a person's well-being (eg, high levels of fatigue or bad sleep quality) [46]. Monitoring and early detection of abnormalities in these measured aspects is expected to provide important information to the clinical team as a basis for additional diagnoses or therapeutic decisions.

To issue the EMAs and PeerMAs, we used time-contingent triggers [47] with different application-triggered surveys at different frequencies for each type of question (Table 3). The frequency was chosen by the research team aiming to minimize user burden by implying the collection of several data points during the week instead of, for example, all data points at once. Patients and support persons received the EMA and PeerMA via silent push notifications (no sound and no vibration). Unattended notifications expired after one hour to prevent questions from accumulating, reducing the burden to the patient or support person. However, the smartphone app additionally allowed patients and support persons to initiate EMAs or PeerMAs voluntarily (ie, self-triggered) at any moment they wished to do so.

During the study duration, we also collected physical activity datasets (steps with each minute of data collected categorized as sedentary, light, moderate, or vigorously active) and sleep data (sleep duration) from patients leveraging a Fitbit Charge 2 wearable provided to them.

XSL•FO
RenderX

Table 3. List of questions used for the ecological momentary assessments (EMA) and peer-ceived momentary assessments (PeerMA).

Category and questions	Frequency
Hope	
• EMA: How much hope are you experiencing? • PeerMA: How much hope is your peer projecting?	• Weekly on Friday only • Random between 2 PM and 6 PM
Sleep	
• EMA: How well are you sleeping? • PeerMA: How well is your peer sleeping?	• Weekly on Tuesday, Thursday, and Sunday • Around 10 AM
Fatigue	
• EMA: How much fatigue are you experiencing? • PeerMA: How much fatigue is your peer projecting?	• Weekly on Monday only • Random between 2 PM and 6 PM
Pain	
• EMA: How much pain are you experiencing? • PeerMA: How much pain is your peer projecting?	• Bi-weekly on Tuesday only • Around 2 PM
Depression	
• EMA: How much depression are you experiencing? • PeerMA: How much depression is your peer projecting?	• Every 3 weeks on Wednesday • Around 2 PM

Exit Survey

At the end of the study, both patients and support persons were invited to complete an exit survey consisting of a subset of 10 questions from the User Burden Scale [48] directly related to the dynamics of this study (level of effort required to use the smartphone app to answer EMAs or PeerMAs). Then, we asked participants to comment on usability aspects of the mobile device app (eg, usability, positive, and negative aspects perceived) and to comment about how patients felt about reflecting on their states during the day. In contrast, support persons answered how they felt about reflecting on their peers' (ie, the patient) states during the day. Additionally, we asked participants to comment about any other physical or psychological behaviors or states related to health or disease that they would like to assess in the future using these paired self-report-based methods. We also asked patients about their perceived value of the wearable activity monitor during the preparation for the transplant or the recovery period. Finally, we asked their opinion about the study design aspects that could be improved.

Ethical Considerations

The Panel on Human Subjects in Medical Research of Stanford University approved the study protocol "Studying the Subjective and Objective Momentary Perception of Quality of Life in Different Contexts of Daily Life" (number 47833, Reg# 351). Both participants and support persons signed an informed consent form explaining the implications of their participation in the study. The informed consent contains a description of the overall process, what data they will contribute, and how we will manage such data in time; the time and effort expected from them; associated risks of participating in the study; compensation scheme (there was no compensation in this study); participants' rights such as withdrawing without any negative consequences; contact information for the participants to approach research personnel other than the front line researchers with whom the participants habitually interact with, among other elements. Additionally, patients signed an authorization form to consent to the use of their health information for research purposes, which is required by the Panel of Human Subjects in Medical Research of Stanford University.

Results

This section presents the study's results in 2 subsections after the study aims.

Aim 1: Feasibility of the Methods in Clinical Settings

This section presents quantitative results related to the feasibility of using the EMA, PeerMA, and physical activity datasets from wearable devices to complement the study of quality-of-life aspects experienced by patients prior to and after receiving an organ transplant. For this feasibility study, we focus on the participation and EMA/PeerMA completion rates, the visual value of the EMA/PeerMA and wearable-based observations, and the similarity scores between the EMA/PeerMA and wearable data streams.

Collected Data Summary

To begin with, Table 1, extracted from the entry survey, shows the socioeconomic characteristics such as gender, age, marital status, education, and employment. Table 4 summarizes the questionnaire/test scores derived for patients and support persons at the beginning of the study, before the collection of daily EMAs and PeerMAs started.

XSL•FO
RenderX

Table 4. Patients' and support persons' scores computed from the study entry surveys (no data available for P4 and SP4).

Instrument/survey name and subscales	Respondents (Px=Patient x, SPx=Support Person x)											
	P1	SP1	P2	SP2	P3	SP3	P5	SP5	P6	SP6	P7	SP7
Adult State Hope Scale [35][a]												
Pathways	24	23	21	24	23	22	21	N/A[b]	23	22	22	17
Agency	24	21	22	22	21	22	16	N/A	22	23	19	18
Total	48	44	43	46	44	44	37	N/A	45	45	41	35
Modified Fatigue Impact Scale [37][c]												
Physical	16	14	22	3	35	1	5	N/A	27	3	22	N/A
Social	19	11	4	3	25	2	8	N/A	4	1	19	N/A
Psychosocial	0	1	4	1	8	0	0	N/A	5	2	4	N/A
Total	35	26	30	7	68	3	13	N/A	36	6	45	N/A
Patient Health Questionnaire [38][d]												
Anxiety	1	0	1	0	1	0	1	N/A	1	2	0	1
Depression	1	0	1	0	2	0	0	N/A	2	0	0	2
Total	2	0	2	0	3	0	1	N/A	3	2	0	3
Perceived Stress Scale [43][e]												
Total	2	0	5	2	7	2	3	N/A	4	4	3	7
Pittsburgh Sleep Quality Index [36][f]												
Subjective sleep quality	0	0	1	0	3	1	1	N/A	1	0	1	N/A
Sleep latency	1	0	3	1	2	1	0	N/A	1	0	0	N/A
Sleep duration	1	0	2	1	1	0	1	N/A	1	0	1	N/A
Habitual sleep efficiency	1	1	3	0	1	1	1	N/A	1	1	0	N/A
Sleep disturbances	1	1	1	1	1	1	1	N/A	1	1	2	N/A
Use of sleep medication	0	0	0	0	0	0	0	N/A	1	0	0	N/A
Daytime dysfunction	1	0	0	0	3	0	0	N/A	1	0	3	N/A
Short Form Survey [44][g]												
Physical functioning	83	100	25	100	33	100	83	N/A	33	100	58	67
Role functioning	0	100	75	100	0	100	100	N/A	0	100	100	50
Social functioning	100	100	20	100	0	100	100	N/A	0	100	80	80
Mental health	88	92	72	84	76	88	72	N/A	60	64	92	56
Health perceptions	72	77	17	95	22	100	62	N/A	42	92	42	25
Pain	20	40	60	0	80	20	20	N/A	60	20	60	80

[a]In this scale, higher scores indicate higher hope levels.

[b]Not applicable.

[c]In this scale, higher scores indicate a greater impact of fatigue on a persons' activities.

[d]In this scale, scores are rated as normal (0-2), mild (3-5), moderate (6-8), and severe (9-12).

[e]In this scale, higher scores are correlated with more stress.

[f]In this scale, scoring is based on a 0 to 3 scale, whereby 3 reflects the negative extreme on the Likert Scale.

[g]In this scale, 0 is the lowest score and 100 is the highest. Higher scores indicate better health except for pain.

In this study, we collected 2118 participant-days, 1142 patient-days (mean 143, SD 63), and 976 support person–days (mean 139, SD 77). Table 5 shows the days each patient and support person contributed to the study. The value of column "Expected number of EMA/PeerMA" is calculated based on the number of weeks the person was in the study and the number of surveys that should have been triggered by that time. Patients received 821 EMA prompts (mean 103, SD 53) and responded

XSL•FO
RenderX

a total of 514 EMAs (mean 64, SD 50) with an average response rate of 63% (SD 26%). Support persons received 612 PeerMA prompts (mean 87, SD 60) and responded to a total of 448 PeerMAs (mean 64, SD 52) with an average response rate of 74% (SD 39%). The number of answered EMA/PeerMAs is divided into application-triggered (prompts presented by the app at the programmed frequency for each question) and self-triggered (those initiated by patients and support persons voluntarily at any moment). Table 5 also shows the response rate, computed as a percentage of the answered EMAs/PeerMAs over the expected number of EMAs/PeerMAs. As Table 5 shows, the response rate can be above 100% when the patients or support persons answer almost all application-triggered surveys, and additionally, they answer many self-triggered surveys, sometimes making up for missed application-triggered surveys and sometimes contributing voluntarily with an additional data point.

Table 5. Summary of the data contributed by each patient and support person in the study.

Patient/ support person ID	Time in the study, days (weeks)	Expected number of EMA[a]/PeerMA[b]	EMA/PeerMA answered		Response rate (%)	Data from wearable	
			Application triggered	Self-triggered		Physical activity (days)	Sleep data (days)
Patient 1	196 (28.0)	164	161	0	98	196	191
Support person 1	196 (28.0)	164	147	0	90	N/A[c]	N/A
Patient 2	169 (24.1)	140	0	69	49	127	95
Support person 2	169 (24.1)	140	0	91	65	N/A	N/A
Patient 3	39 (5.6)	30	1	29	100	39	15
Support person 3	39 (5.6)	30	4	13	57	N/A	N/A
Patient 4	76 (10.9)	59	2	26	47	11	4
Support person 4	76 (10.9)	59	9	61	119	N/A	N/A
Patient 5	95 (13.6)	77	32	1	43	47	18
Support person 5	94 (13.4)	77	27	71	127	N/A	N/A
Patient 6	174 (24.9)	140	17	25	30	0	0
Support person 6	162 (23.1)	135	14	7	16	N/A	N/A
Patient 7	69 (9.9)	53	0	32	60	0	0
Support person 7	12 (1.7)	7	0	4	57	N/A	N/A
Patient 8	195 (27.9)	158	117	2	75	2	2
Support person 8	N/A	N/A	N/A	N/A	N/A	N/A	N/A

[a]EMA: ecological momentary assessment.

[b]PeerMA: peer-ceived momentary assessment.

[c]Not applicable.

As noted in Table 5, not all patients and support persons remained in the study for the expected period of 6 months (approximately 26 weeks). During the exit interview, some of them stated various reasons why they had stopped contributing. For instance, some patients said they felt very well and healthy and even ahead of the recovery process, and thus, they did not perceive much value from answering the EMA/PeerMA prompts. Additionally, some support persons indicated that answering the questions during working hours was challenging, and others indicated that they did not see EMA/PeerMA notifications at all for some time, which could be due to some changes in their smartphone settings.

For each patient and support person, we normalized their EMA/PeerMA responses to 0 to 1 based on the highest and lowest assessment given by each person. Table 6 shows the median, mean, and SD of all assessments for hope, sleep quality, fatigue, depression, and pain. In Table 6, lower values in the “Median” and “Mean” columns represent more desirable states in the case of fatigue, depression, and pain, and less desirable states in the case of hope and sleep quality. Conversely, higher values in the “Median” and “Mean” represent less desirable states for fatigue, depression, and pain, and more desirable states for hope and sleep quality.

Table 7 compares the concordance between the patient’s and the support person’s reports. It shows the number of dyads for which the absolute difference between the mean of each quality of life aspect was under a low threshold of 15% proposed as a minimal importance difference by the Institute for Quality and Efficiency in Health Care in Germany [49].

XSL•FO
RenderX

Table 6. Summary of the ecological momentary assessment (EMA) and peer-ceived momentary assessment (PeerMA) aggregated values obtained from patients and support persons. Each row shows the median, mean, and SD for the corresponding patient or support person calculated from all the assessments issued by that person.

Patient / support person ID	Hope (0-1) (higher is better)		Sleep quality (0-1) (higher is better)		Fatigue (0-1) (lower is better)		Depression (0-1) (lower is better)		Pain (0-1) (lower is better)	
	Median	Mean (SD)	Median	Mean (SD)	Median	Mean (SD)	Median	Mean (SD)	Median	Mean (SD)
Patient 1	0.69	0.64 (0.3)	0.89	0.82 (0.17)	0.08	0.23 (0.29)	0.30	0.31 (0.32)	0.25	0.34 (0.29)
Support person 1	0.81	0.71 (0.29)	0.75	0.72 (0.14)	0.35	0.35 (0.25)	0.14	0.31 (0.38)	0.10	0.21 (0.28)
Patient 2	0.69	0.6 (0.31)	0.85	0.72 (0.28)	0.35	0.43 (0.35)	0.34	0.4 (0.3)	0.27	0.38 (0.34)
Support person 2	0.81	0.64 (0.3)	0.91	0.76 (0.29)	0.12	0.27 (0.26)	0.10	0.28 (0.28)	0.21	0.32 (0.27)
Patient 3	0.69	0.64 (0.32)	0.34	0.44 (0.32)	0.19	0.29 (0.26)	0.18	0.31 (0.28)	0.19	0.3 (0.28)
Support person 3	0.69	0.64 (0.32)	0.46	0.5 (0.37)	0.14	0.22 (0.27)	0.21	0.29 (0.3)	0.16	0.21 (0.28)
Patient 4	0.61	0.59 (0.24)	0.76	0.69 (0.29)	0.16	0.35 (0.36)	0.12	0.19 (0.21)	0.19	0.31 (0.32)
Support person 4	0.75	0.64 (0.24)	0.78	0.69 (0.28)	0.52	0.55 (0.21)	0.51	0.53 (0.26)	0.61	0.59 (0.23)
Patient 5	0.50	0.5 (0.5)	0.88	0.77 (0.29)	0.86	0.62 (0.54)	0.50	0.5 (0.71)	0.00	0.25 (0.5)
Support person 5	0.69	0.66 (0.14)	0.77	0.75 (0.15)	1.00	0.94 (0.2)	1.00	0.87 (0.27)	0.00	0.12 (0.27)
Patient 6	0.41	0.47 (0.31)	0.66	0.6 (0.26)	0.75	0.6 (0.29)	0.44	0.36 (0.24)	0.36	0.39 (0.3)
Support person 6	0.71	0.63 (0.34)	0.43	0.52 (0.37)	0.73	0.61 (0.4)	0.07	0.27 (0.36)	0.74	0.57 (0.41)
Patient 7	0.81	0.77 (0.24)	0.75	0.72 (0.18)	0.70	0.63 (0.29)	0.34	0.33 (0.2)	0.59	0.63 (0.28)
Support person 7	0.29	0.39 (0.49)	0.86	0.68 (0.46)	0.45	0.48 (0.45)	0.36	0.43 (0.49)	0.27	0.39 (0.47)

Table 7. Distribution of the number of dyads with an absolute difference of up to 15% between ecological momentary assessment (EMA) and peer-ceived momentary assessment (PeerMA).

Absolute difference between EMA/PeerMA Means	Number of dyads (from total of 7 dyads)				
	Hope, n (%)	Sleep, n (%)	Fatigue, n (%)	Depression, n (%)	Pain, n (%)
≤5%	3 (43)	4 (57)	1 (14)	2 (29)	0 (0)
≤10%	1 (14)	3 (43)	1 (14)	2 (29)	2 (29)
≤15%	0 (0)	0 (0)	2 (29)	1 (14)	2 (29)
>15%	3 (43)	0 (0)	3 (43)	2 (29)	3 (43)

Visual Exploration of Daily EMA/PeerMA Dynamics

Overview

This section shows selected visualizations of the values reported by patients and support persons via EMA/PeerMA surveys, as well as physical activity datasets collected from a wearable activity monitor by the patient. The intention of the plots is to aid in understanding the magnitude of agreement/disagreement in their EMA/PeerMA self-reported assessments over time.

To use the collected EMA/PeerMAs and wearable-based physical activity data, we performed a few data preparation steps. First, we searched for duplicate EMA/PeerMAs assessments (issued on the same day) and removed incomplete EMA/PeerMAs—if at least one of them was complete. If both duplicate assessments were complete or incomplete, we removed the oldest one (they usually differed only by a few seconds/minutes). Second, we aligned the patient and support person EMA/PeerMAs data streams to a common start and end date. To do this, we used the first day of the patient's data stream as the start (trimming the support person's data stream to match the patient's start date), and for the end date, we picked the day from either data stream that resulted in the largest common set of assessments from both participants. In other words, we used the earliest date among the last patient report and the last support person's report. Third, we imputed missing EMA/PeerMA values using a linear regression model with polynomial features of second degree and the Ridge algorithm with default

XSL•FO
RenderX

parameters in the Scikit-learn data analytics framework. We selected this model based on prior work [50] and the fact that it produces smooth curves with suitable values according to our datasets.

All Outcomes: Visual Comparison

Figure 2 shows a sample chart for the dyad (patient 1, support person 1).

Figure 3 shows the plot of physical activity data for patient 1 collected by the wearable activity monitor for the same period of time.

Figure 2. Sample plots for the combined ecological momentary assessment (EMA)/ peer-ceived momentary assessments (PeerMAs) for patient 1-support person 1. The x-axis represents the time in the study, and the y-axis represents the magnitude of the normalized assessments for hope, sleep quality, fatigue, depression, and pain. The upper plot corresponds to the patient, and the lower plot corresponds to the support person.

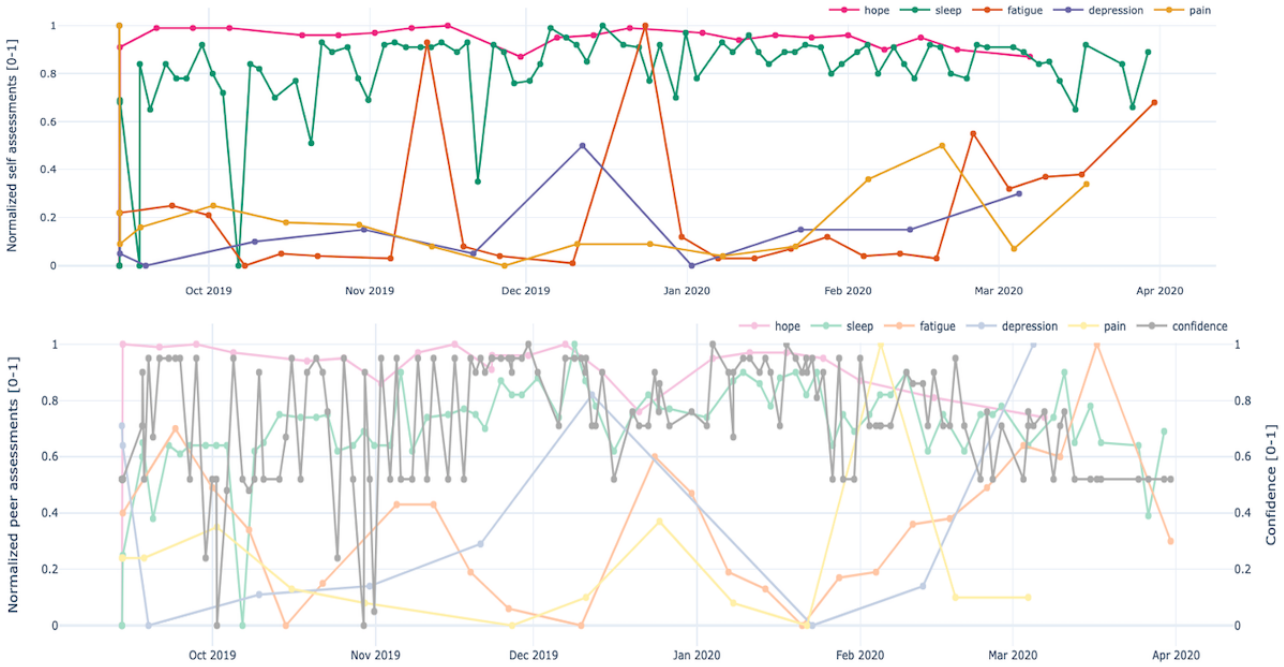

Figure 3. Sample plot for patient 1's physical activity data. The x-axis represents the time in the study, and the y-axis represents the magnitude of the absolute values collected from the wearable. The primary y-axis (left) represents hours for measures such as sleep duration and active time of the patient. The secondary y-axis (right) represents the absolute number of steps measured from the patient.

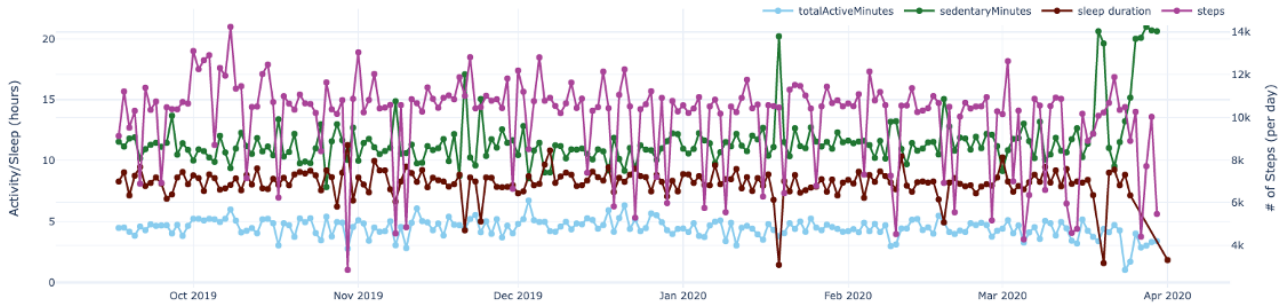

Outcome-by-Outcome Comparison

Figures 4-8 show the individual charts for the dyad (patient 1, support person 1) corresponding to each of the constructs separately for hope, sleep quality, fatigue, depression, and pain. Stronger colors represent the patient's assessments, while lighter colors represent the support person's assessments. The light gray line represents the confidence value estimated by the support person each time they issued a PeerMA.

In Figure 4, we can see a high similarity regarding the perception of hope between the patient and the support person along the timeline. In Figure 5, the patient reported a low value for sleep quality on October 6 and November 21; when we looked at the patient's comments for those days, we learned that one was due to insomnia, and the second one was due to anxiety caused by a meeting the next day. Likewise, in Figure 6, the patient reported a high value for fatigue on November 11 and December 23. High fatigue for a recovering patient could be a sign of alert at times. In this case, the patient's comments reflect that fatigue is coming from usual daily life activities and social activities like entertaining house guests.

Figure 7 shows that on December 11, both the patient and support person reported a value higher than usual for depression. In this case, the patient's comment is, however, rather positive, indicating that everything was going well, hence not aligned with the self-reported level of depression. This might also be a

XSL•FO
RenderX

sign of alert for caregivers; however, as the plot shows, subsequent reports by the patient return to lower values, suggesting that perhaps it was an isolated event that triggered a higher perception of depression. Regarding the perception of pain shown in Figure 8, there are mainly 2 moments when the patient and support person scores differ more: December 23 and February 4. Although there are no comments for these reports, from other patients' comments, we learned that the patient has comorbidities that may affect the perception of physical pain.

Figure 4. Sample plots for individual ecological momentary assessment (EMA)/ peer-ceived momentary assessments (PeerMAs) related to hope. The x-axis represents the time in the study, and the y-axis represents the magnitude of the normalized assessments. The solid spheres represent the acquired values, and the dotted squares represent the regression curves. The light gray diamonds represent the confidence value estimated by the support person each time they issued a PeerMA.

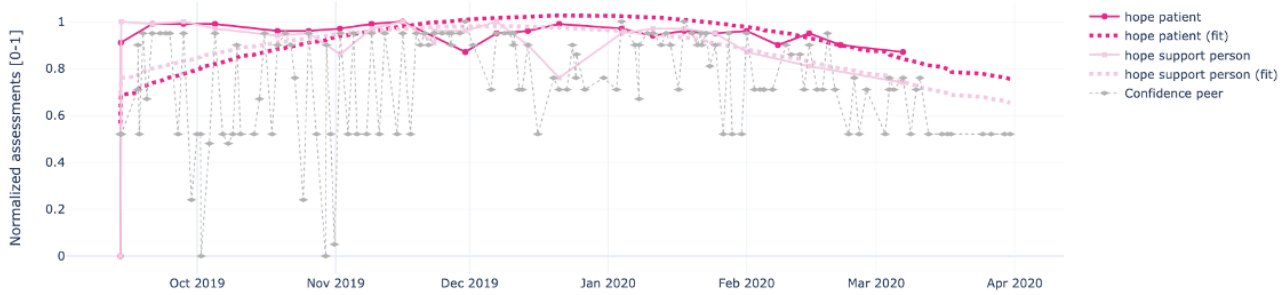

Figure 5. Sample plots for individual ecological momentary assessment (EMA)/ peer-ceived momentary assessments (PeerMAs) related to sleep quality. The x-axis represents the time in the study, and the y-axis represents the magnitude of the normalized assessments. The solid spheres represent the acquired values, and the dotted squares represent the regression lines. The light gray diamonds represent the confidence value estimated by the support person each time they issued a PeerMA.

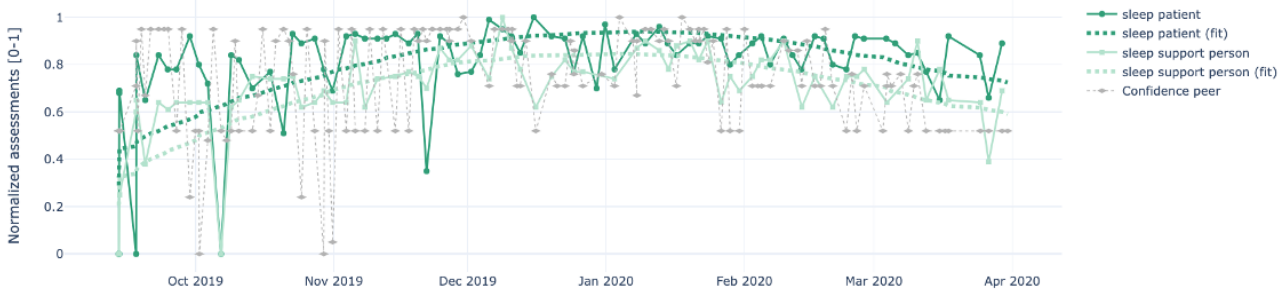

Figure 6. Sample plots for individual ecological momentary assessment (EMA)/ peer-ceived momentary assessments (PeerMAs) related to fatigue. The x-axis represents the time in the study, and the y-axis represents the magnitude of the normalized assessments. The solid spheres represent the acquired values, and the dotted squares represent the regression lines. The light gray diamonds represent the confidence value estimated by the support person each time they issued a PeerMA.

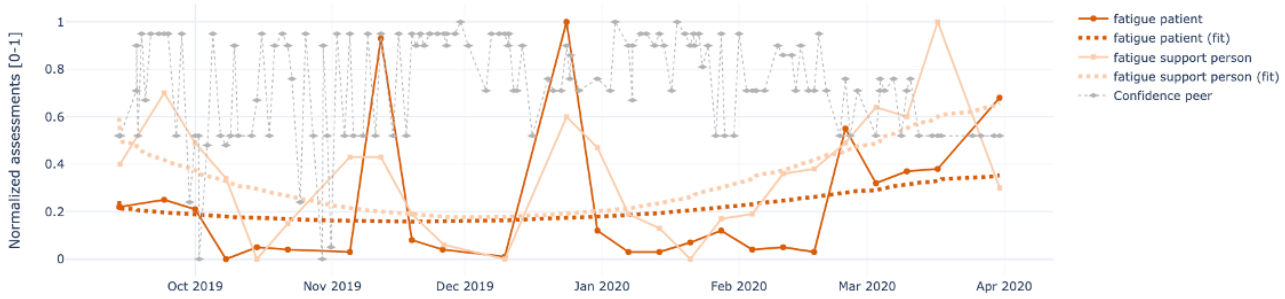

XSL•FO
RenderX

Figure 7. Sample plots for individual ecological momentary assessment (EMA)/ peer-ceived momentary assessments (PeerMAs) related to depression. The x-axis represents the time in the study, and the y-axis represents the magnitude of the normalized assessments. The solid spheres represent the acquired values, and the dotted squares represent the regression lines. The light gray diamonds represent the confidence value estimated by the support person each time they issued a PeerMA.

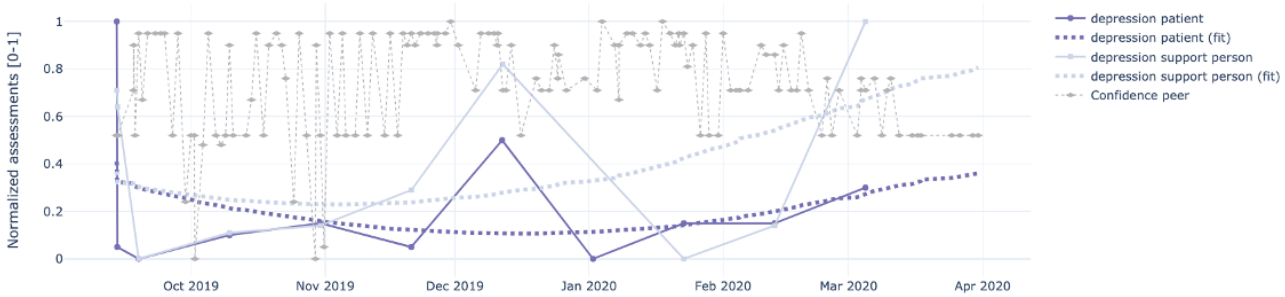

Figure 8. Sample plots for individual ecological momentary assessment (EMA)/ peer-ceived momentary assessments (PeerMAs) related to pain. The x-axis represents the time in the study, and the y-axis represents the magnitude of the normalized assessments. The solid spheres represent the acquired values, and the dotted squares represent the regression lines. The light gray diamonds represent the confidence value estimated by the support person each time they issued a PeerMA.

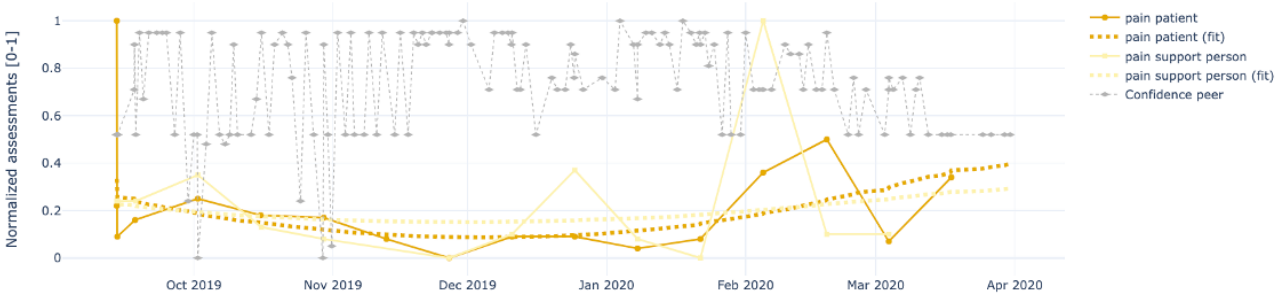

Outcome-by-Outcome and Wearable Outcomes Comparison

Figures 9-13 also show data for the dyad (patient 1, support person 1), focusing on the regressed curves for the EMAs and PeerMAs (hope, sleep quality, fatigue, depression, and pain) along one physical activity marker collected from the patient corresponding to the number of steps walked per day. Table 8 shows the actual correlations between EMA scores and the patient's physical activity, which in this case are moderate to high, both positive and negative.

Figure 9. Sample plots for physical activity and individual ecological momentary assessment (EMA)/ peer-ceived momentary assessments (PeerMAs) related to hope. The x-axis represents the time in the study. The primary y-axis (left) represents regressed curves (pink) for hope by the patient and support person. The secondary y-axis (right) represents the regressed curve (purple) for the number of steps measured by the patient.

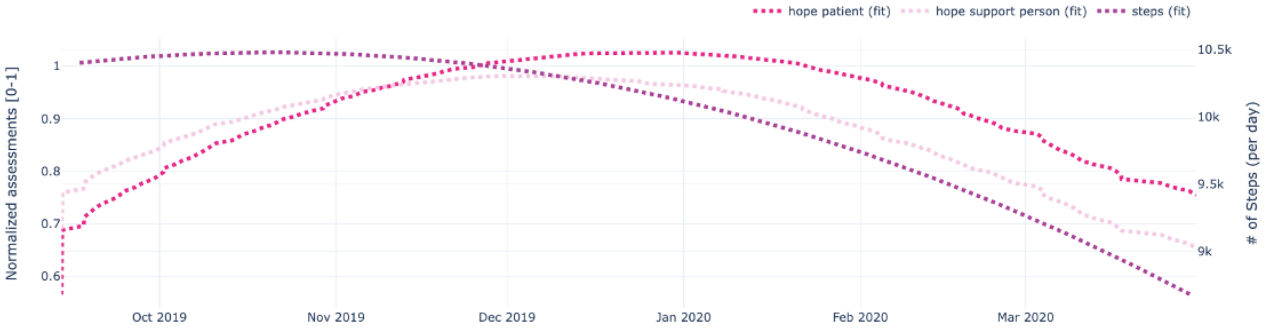

XSL•FO
RenderX

Figure 10. Sample plots for physical activity and individual ecological momentary assessment (EMA)/ peer-ceived momentary assessments (PeerMAs) related to sleep quality. The x-axis represents the time in the study. The primary y-axis (left) represents regressed curves (green) for sleep quality by the patient and support person. The secondary y-axis (right) represents the regressed curve (purple) for the number of steps measured by the patient.

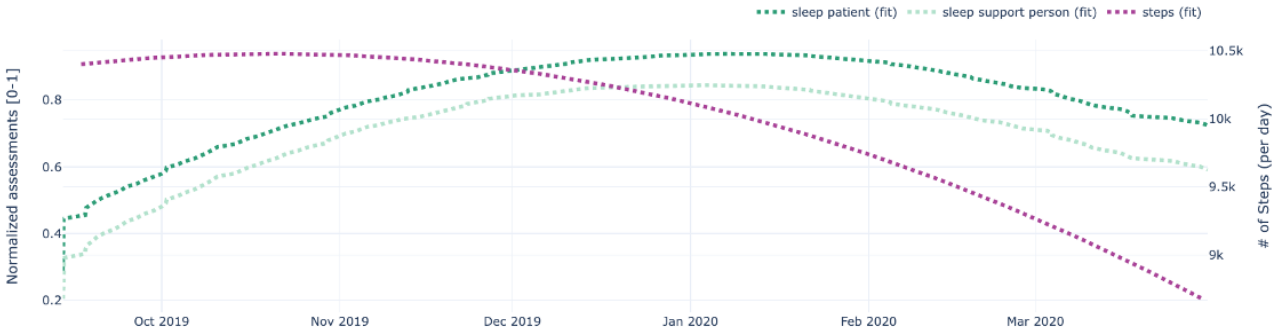

Figure 11. Sample plots for physical activity and individual ecological momentary assessment (EMA)/ peer-ceived momentary assessments (PeerMAs) related to fatigue. The x-axis represents the time in the study. The primary y-axis (left) represents regressed curves (orange) for fatigue by the patient and support person. The secondary y-axis (right) represents the regressed curve (purple) for the number of steps measured by the patient.

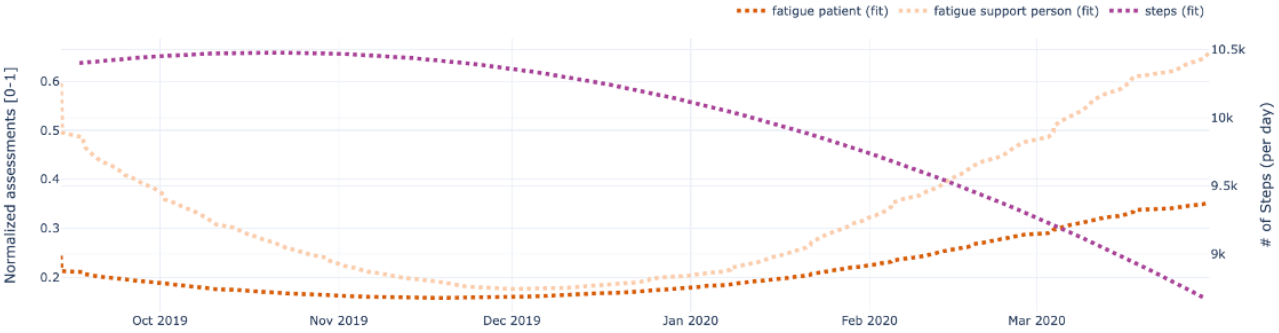

Figure 12. Sample plots for physical activity and individual ecological momentary assessment (EMA)/ peer-ceived momentary assessments (PeerMAs) related to depression. The x-axis represents the time in the study. The primary y-axis (left) represents regressed curves (blue) for depression by the patient and support person. The secondary y-axis (right) represents the regressed curve (purple) for the number of steps measured by the patient.

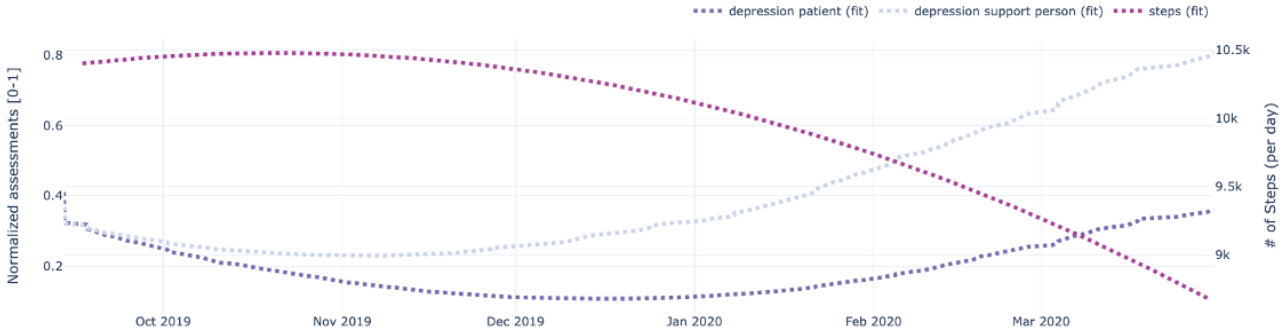

Figure 13. Sample plots for physical activity and individual ecological momentary assessment (EMA)/ peer-ceived momentary assessments (PeerMAs) related to pain. The x-axis represents the time in the study. The primary y-axis (left) represents regressed curves (yellow) for pain by the patient and support person. The secondary y-axis (right) represents the regressed curve (purple) for the number of steps measured by the patient.

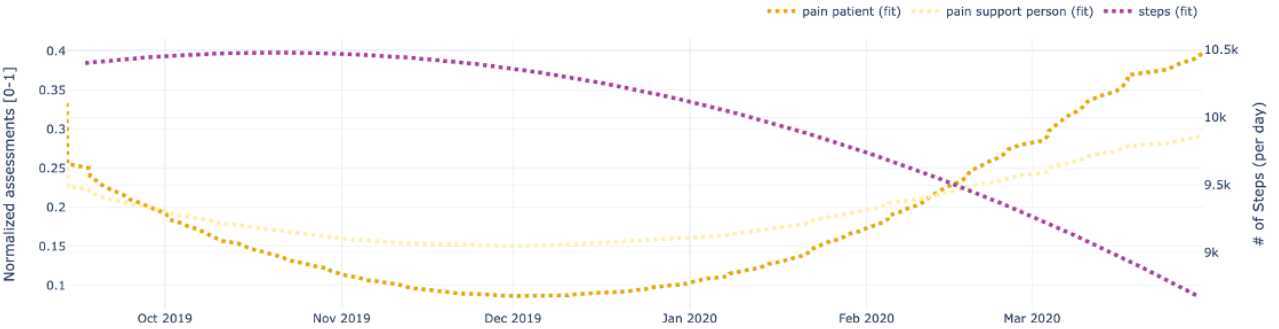

XSL•FO
RenderX

Table 8. Spearman correlations between ecological momentary assessment (EMA) and physical activity (number of steps) collected for each patient in the study. Each row shows the correlation between the patients' self-assessments and the number of steps recorded with the wearable activity monitor, and each column represents a construct of the study.

Dyad	Hope		Sleep		Fatigue		Depression		Pain	
	rs	*P* value	rs	*P* value	rs	*P* value	rs	*P* value	rs	*P* value
1	0.96	<.001	–0.49	<.001	–0.61	<.001	0.98	<.001	–0.56	.04
2	0.97	<.001	0.79	<.001	–0.99	<.001	–0.98	<.001	–0.95	<.001
3	0.18	.40	0.29	.17	–0.80	<.001	0.39	.06	–0.48	.02
4	0.01	.94	1.00	<.001	–0.72	<.001	–0.95	<.001	–0.77	<.001
5	0.91	<.001	0.76	<.001	–0.85	<.001	–0.83	<.001	–0.88	<.001

Similarity Scores

Overview

To assess the value of the scores given by the patient and the support person, we used 2 approaches: similarity of the EMA and PeerMA scores in time, using standard distance-based metrics, and similarity scores based on statistical correlation analysis.

PeerMA as a Model of EMA

We wanted to measure how close or similar the regressed curves between the patient and the support person were for all the health outcomes they assessed during the study. To do this, we took all the data pairs for each dyad (patient, support person) and computed the following similarity metrics: mean absolute error (MAE), mean squared error, root-mean-squared error, and R^2. These metrics are commonly used to assess the fitness of a regression model, in other words, to assess how well a regression model represents the data points from which it was sourced. However, in our case, because we have regression curves representing both the patient and the support person's data points, we consider the curve created with the patients' assessments to be the ground truth and the curve created with the support person's assessments as the approximation model.

Table 9 shows the value for all the similarity scores for each dyad (Dyad X represents the pair [PatientX, Support PersonX]). For example, if we focus on the first segment of Table 9, "Hope," we see that the error (MAE) or mean difference between patient 1's assessment of hope and that of his/her peer, support person 1, was about 4% (0.0412). This value is located under the "Dyad 1" column and can be visually inspected in Figure 4. For dyad 2, that difference was slightly smaller, 3.5% (0.0350), meaning that both patients and support persons in dyads 1 and 2 reported similar values in their assessments. On the contrary, dyad 7 shows a difference of nearly 74% (0.7377), meaning that the curves representing patient 7's and support person 7's assessments of hope are very dissimilar.

XSL•FO
RenderX

Table 9. Summary of ecological momentary assessment (EMA)/peer-ceived momentary assessment (PeerMA) similarity scores: mean absolute error (MAE), mean squared error (MSE), root-mean-squared error (RMSE), and R2.

Aspect and metric	Dyad 1	Dyad 2	Dyad 3	Dyad 4	Dyad 5	Dyad 6	Dyad 7
Hope							
MAE	0.0412	0.0350	0.1589	0.1487	0.2583	0.2970	0.7377
MSE	0.0022	0.0016	0.0542	0.0317	0.0829	0.1185	0.5661
RMSE	0.0469	0.0400	0.2329	0.1780	0.2878	0.3442	0.7524
R^2	0.9346	0.9807	–1.8358	–2.1791	–1.3084	–1.6667	–39.5400
Sleep							
MAE	0.0979	0.0565	0.2474	0.0315	0.1030	0.2321	0.0507
MSE	0.0100	0.0046	0.0988	0.0014	0.0130	0.0740	0.0039
RMSE	0.0998	0.0676	0.3143	0.0371	0.1141	0.2720	0.0628
R^2	–0.6469	0.9288	–15.3476	0.9685	–0.6926	–13.9986	–2.7936
Fatigue							
MAE	0.1202	0.0919	0.0989	0.2211	0.5018	0.2356	0.0992
MSE	0.0199	0.0188	0.0128	0.0601	0.3857	0.0701	0.0163
RMSE	0.1410	0.1369	0.1129	0.2452	0.6210	0.2648	0.1279
R^2	–1.7331	0.8311	–0.3196	–2.0013	–1.1524	–1.2381	–0.1915
Depression							
MAE	0.1475	0.0611	0.1198	0.3004	0.6210	0.2508	0.3307
MSE	0.0280	0.0050	0.0200	0.1011	0.4578	0.0810	0.1187
RMSE	0.1672	0.0704	0.1415	0.3179	0.6766	0.2847	0.3445
R^2	–1.2353	0.9433	–5.6800	–4.1741	–2.1509	–776.5628	–14.0340
Pain							
MAE	0.1443	0.0758	0.1295	0.2076	0.3401	0.1951	0.5581
MSE	0.0367	0.0096	0.0202	0.0510	0.2138	0.0489	0.3352
RMSE	0.1915	0.0978	0.1422	0.2258	0.4624	0.2213	0.5790
R^2	–0.6389	0.9077	–13.4371	–261.7243	–0.3047	–3.8061	–18.1295

EMA/PeerMA Spearman Correlations

We performed a correlation analysis for EMA/PeerMA values for the various states without assuming that EMA and PeerMA measure the same constructs. We applied the Spearman rank correlation method because (1) patient and support person assessments are not independent; they both refer to the state of the patient, and (2) the Shapiro-Wilk, D'Agostino K2, and Anderson-Darling tests indicate that not all patients' and support persons' assessments were normally distributed.

The Spearman rank correlation coefficients for each of the states are presented in Table 10. Each row presents the correlation coefficient rs and *P* value for a patient-support person dyad. For example, the correlation between patient 1's assessment of hope and that of his/her peer, support person 1, is 0.96 (highly positive). Table 11 presents a summary of the values from Table 10, classified into 6 correlation groups. The correlations are predominantly on the positive side; 77% of the correlations are positive (46% high, 20% moderate, and 11% weak), and 23% of the correlations are negative (9% high, 3% moderate, and 11% weak).

XSL•FO
RenderX

Table 10. Spearman correlations between ecological momentary assessment (EMA) and peer-ceived momentary assessment (PeerMA) values collected for each study dyad. Each row shows the correlation between the patients' EMAs and support persons' PeerMAs, and each column represents a construct of the study.

Dyad	Hope		Sleep		Fatigue		Depression		Pain	
	rs	*P* value	rs	*P* value	rs	*P* value	rs	*P* value	rs	*P* value
1	0.96	<.001	0.97	<.001	0.88	<.001	0.78	.01	0.90	<.001
2	0.98	<.001	0.81	<.001	1.00	<.001	1.00	<.001	0.96	<.001
3	0.18	.40	0.29	.17	0.80	<.001	–0.13	.56	0.48	.02
4	0.11	.50	0.92	<.001	0.71	<.001	0.78	<.001	–0.23	.16
5	0.90	<.001	–0.71	<.001	–1.00	<.001	0.80	<.001	–0.81	<.001
6	0.62	<.001	0.15	.52	–0.26	.22	0.54	.01	0.22	.36
7	–0.50	.01	0.07	.72	0.05	.79	–0.23	.25	0.52	.01

Table 11. Summary of ecological momentary assessment (EMA)/ peer-ceived momentary assessment (PeerMA) Spearman correlations from Table 10.

Correlation strength	Hope, n (%)	Sleep, n (%)	Fatigue, n (%)	Depression, n (%)	Pain, n (%)	Total, n (%)
Highly positive[a]	3 (43)	3 (43)	4 (57)	4 (57)	2 (29)	16 (46)
Moderately positive[b]	1 (14)	3 (43)	0 (0)	1 (14)	2 (29)	7 (20)
Weakly positive[c]	2 (29)	0 (0)	1 (14)	0 (0)	1 (14)	4 (11)
Weakly negative[d]	0 (0)	0 (0)	1 (14)	2 (29)	1 (14)	4 (11)
Moderately negative[e]	1 (14)	0 (0)	0 (0)	0 (0)	0 (0)	1 (3)
Highly negative[f]	0 (0)	1 (14)	1 (14)	0 (0)	1 (14)	3 (9)

[a](0.67 to 1.00): Values of the Spearman correlation inside this interval are considered highly positive.

[b](0.34 to 0.66): Values of the Spearman correlation inside this interval are considered moderately positive.

[c](0.00 to 0.33): Values of the Spearman correlation inside this interval are considered weakly positive.

[d](–0.33 to 0.00): Values of the Spearman correlation inside this interval are considered weakly negative.

[e](– 0.66 to –0.34): Values of the Spearman correlation inside this interval are considered moderately negative.

[f](–1.00 to –0.67): Values of the Spearman correlation inside this interval are considered highly negative.

EMA/Physical Activity Spearman Correlations

We performed a correlation analysis between the EMA values for the various states and physical activity (number of steps) measured by a wearable activity monitor. We applied the Spearman rank correlation method because (1) patient and support person assessments are not independent; they both refer to the state of the patient, and (2) the Shapiro-Wilk, D'Agostino K2, and Anderson-Darling tests indicate that not all patients' and support persons' assessments were normally distributed.

The Spearman rank correlation coefficients for each of the states are presented in Table 8, excluding dyads 6 and 7 because we did not collect physical activity records for those patients. Each row presents the correlation coefficient rs and *P* value for a patient-support person dyad. Table 12 presents a summary of the values from Table 8, classified into 6 correlation groups. In this case, the correlations are balanced between positive and negative. We observe that 44% of the correlations are positive (28% high, 4% moderate, and 12% weak), while 56% of the correlations are negative (40% high, 16% moderate, and 0% weak).

XSL•FO
RenderX

Table 12. Summary of ecological momentary assessment (EMA)/steps Spearman correlations from Table 8.

Correlation strength	Hope, n (%)	Sleep, n (%)	Fatigue, n (%)	Depression, n (%)	Pain, n (%)	Total, n (%)
Highly positive[a]	3 (60)	3 (60)	0 (0)	1 (20)	0 (0)	7 (28)
Moderately positive[b]	0 (0)	0 (0)	0 (0)	1 (20)	0 (0)	1 (4)
Weakly positive[c]	2 (40)	1 (20)	0 (0)	0 (0)	0 (0)	3 (12)
Weakly negative[d]	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Moderately negative[e]	0 (0)	1 (20)	1 (20)	0 (0)	2 (40)	4 (16)
Highly negative[f]	0 (0)	0 (0)	4 (80)	3 (60)	3 (60)	10 (40)

[a](0.67 to 1.00): Values of the Spearman correlation inside this interval are considered highly positive.

[b](0.34 to 0.66): Values of the Spearman correlation inside this interval are considered moderately positive.

[c](0.00 to 0.33): Values of the Spearman correlation inside this interval are considered weakly positive.

[d](–0.33 to 0.00): Values of the Spearman correlation inside this interval are considered weakly negative.

[e](– 0.66 to –0.34): Values of the Spearman correlation inside this interval are considered moderately negative.

[f](–1.00 to –0.67): Values of the Spearman correlation inside this interval are considered highly negative.

Aim 2: Human Factors

This section presents human factors driving the acceptance of and the sustained engagement with the methods (as studied by Wulfovich et al [34]). Because the PeerMA method is inspired by 2 related concepts, observers or proxies from clinical research [13], and the existing EMA method [1], it is expected that some of the human factors driving the acceptance of PeerMA are similar to or related to the human factors driving the acceptance of EMA. We report qualitative data such as participants' reflections after using EMA and PeerMA, the perceived difficulty of using the technology (smartphone app), and the reliability of the technologies to preserve the quality of the collected data. Results were drawn from patients' and support persons' comments typed in when they completed an EMA/PeerMA and via the exit survey at the end of the study. Below, we list the relevant human factors identified in this study that impact the acceptance of the PeerMA method.

Self-Awareness and the Value Proposition of the EMA/PeerMA Methods

One important factor determining participants' engagement in the study and motivation to contribute EMA/PeerMA datasets was their prior knowledge and self-awareness about the health and disease-related states we studied (hope, sleep, fatigue, depression, and pain). Patients who are not experiencing negative moods (low hope and depression), or poor physical states (bad sleep, fatigue, or pain) may be less likely to contribute inputs with the EMA or PeerMA methods as valuable sources for better clinical decision-making. These patients may have already communicated their positive state to the clinical team and do not see a value in communicating more. So, the perceived added value of using such a method is a factor that influences participants' acceptance. In general, most patients in the study expressed positive attitudes towards their health during their recovery stage. They exhibited a positive mindset and were committed to following medical indications such as walking and procuring healthy sleep habits. As a result, after a few weeks in the recovery process, many of them indicated that the mobile app was not providing them value. They may have assumed that their current clinical decisions are accurate, as they result in their rather positive state, hence no EMA or PeerMA data is necessary.

The second important factor is to provide feedback or brief information to the participants. In this study, the patients were informed that the EMA or PeerMA served for data collection only, and the mobile app did not display any additional information to the participants or to the clinicians. That was a methodological decision in this preliminary study because one needs to carefully design what and how to show patients and peers this sensitive information about their health. However, that decision reduced the motivation to continue completing EMA and PeerMA for some participants. Unlike our app, the Fitbit app companion that some participants used allowed them to monitor their physical activity and sleep goals, something none of them had experienced before. In fact, all patients who use Fitbit wearables have expressed positive comments about it. The comments from patients included expressions such as "Feel great, loving the push I'm getting from using the Fitbit," "I liked the Fitbit, very informative on my progress," and "It is helpful to know how much exercise I get, and to know about my heart rate during various times of the day."

Individual's Ability, Motivation, and Availability for PeerMA

The acceptance of the PeerMA method is also influenced by how confident individuals feel in using it, based on their smartphone use skills or abilities to perform the assessment task via the EMA/PeerMA method. Another factor is related to time constraints and availability during the day to attend the assessment task via the EMA/PeerMA methods. This is particularly important for patients who may be away from their smartphones during certain hours due to clinical visits, rehabilitation, and rest, among other duties and daily life activities. In this study, we did not impose rules upon the methods or strategies support persons would use to complete their assessments. Unlike patients, the support persons did not express any challenges related to performing the task of providing assessments about someone else.

XSL•FO
RenderX

Because EMA and PeerMA-based studies require that participants invest a bit of time to complete the assessments, it is critical that the instruments involved (eg, smartphone app and wearable devices) are simple and reliable to operate. In our case, participants expressed satisfaction because our study had low complexity, for example, “It was easy to use… though didn’t always see it till later” (patient), “It was a positive experience to share” (support person). Nevertheless, they also criticized the fact that the mobile app becomes boring to use after a few days “the only thing I did not like was every day it was the same questions. Needs more variety” (patient), “Add some variety to the questions each time, more dynamic or based on previous responses such as: yesterday you reported feeling down, how are you feeling today, better, same worse?” (patient).

Usability and Potential Participant Burden of the EMA/PeerMA Methods

Even though the focus of this study was on the PeerMA method and not on the strengths and features of the mQoL-Peers mobile app operationalizing the EMA/PeerMA, participants reported usability aspects of the app. The mQoL-Peers smartphone apps were easy to use, as attested by the participants’ comments. Nevertheless, we received suggestions to improve its iPhone version, which was expected since this version of the app was functionally stable, but its user interface was less mature than the Android version. For instance, among the positive comments about the mobile apps, patients said: “The app was a great vehicle to gather that information, with the least inconvenience to the patient. My complements,” “easy to use and it does not take much time,” “I like the questions, they are simple, direct and give key elements to how I feel about things.” They also made comments about aspects that did not work so well: “Application had bugs and crashed as the OS progressed. User interface was visually unclear, and notifications were inconsistent or misleading” and “the sliding scale was not so responsive, the bar was very small, and it was hard sometimes to move the toggle.”

Additionally, 5 patients and 2 support persons completed a subset of 10 questions from the User Burden Scale [48] at the study exit phase. Table 13 shows the results, suggesting that using the toolset in the study was almost never perceived as a burden. Nevertheless, as attested by 3 respondents, the participant experience could be improved to reduce interaction times.

Table 13. Participants’ evaluation of a subset of questions from the User Burden Scale (N=7) [48].

Question	Never	A little bit of time	Sometimes	Very often	All of the time
App demands high mental effort	7	—[a]	—	—	—
App distracts me from social life	7	—	—	—	—
I worry about the information shared	7	—	—	—	—
I worry about app’s privacy policies	7	—	—	—	—
App is hard to learn	6	1	—	—	—
App has negative effect on social life	6	1	—	—	—
I need assistance to use the app	6	1	—	—	—
I need to remember too much	5	2	—	—	—
I spend too much time in the app	4	1	1	1	—
It takes me too long to do what I want	3	2	1	1	—

[a]Not applicable.

Data Privacy

Careful treatment of sensitive data is an important legal, operational, ethical, and human factor in any human participant study. In adherence to the approved protocol, patients and support persons received a comprehensive set of information about the data privacy policies in four ways: (1) orally, during face-to-face meetings with the researchers at the beginning of the studies; (2) printed, as part of the informed consent forms; (3) displayed upon request from a menu in the mobile app used during the studies; and (4) shared upon request over email during the study. As a result, in this study, neither patients nor support persons made any comments about data privacy considerations, which is considered very positive.

Discussion

Principal Findings

This section discusses the results, providing recommendations for future adopters of the EMA and PeerMA methods in clinical research or clinical practice. We argue that the PeerMA method indeed provides unique information streams that could add value as an input to certain medical decision-making processes. We show examples where support person assessments (ie, PeerMA), as well as physical activity data, can be used to inform health professionals about the actual state of a patient regarding important outcomes such as hope, sleep quality, fatigue, pain, and depression. This section discusses results regarding the in-time similarity and correlations of the patient and the support person’s frequent health assessments conducted in their daily

XSL•FO
RenderX

living. We discuss 4 human factors that influence the acceptance of the PeerMA method.

To begin with, aim 1 of our study, we believe that the time-aligned visual representation of several data streams shown in Figures 2-13 is useful to assist health personnel during the recovery period of patients after a transplant. Likewise, for patients awaiting a transplant, having access to a combination of objective measurements, such as the number of steps per day, with subjective ones reported by the actual patients and their support persons during their daily routines, may allow health professionals to complement their clinical assessment of a patient's disease progression during stages prior to an organ transplant. Our illustrations in this article are only examples of very basic visualizations potentially informative to medical personnel; besides, it is possible to augment such visual representations of data by making them interactive. For example, medical professionals could select specific data points in a chart and study a patient's medical record near that time (eg, accessing recent lab exams, x-ray, or ultrasound imagery, among others) next to these less additional information streams coming from a patient and their support persons. In the literature, when support persons complete a questionnaire with information about a patient, they play the role of proxies [13,21,23,25,26]. According to a study from the International Society for Quality of Life Research, the most common type of use of proxies is for pair studies in which patients and support persons use the same questionnaire (oftentimes with wording change for proxy completion) [29]. Those studies usually collect patients' and support persons' assessments sporadically (once every few weeks or months). Consequently, one benefit of pairing EMA with PeerMA consists of capturing more frequent assessments about a particular aspect of a patient, usually via short questionnaires with one or only a few questions.

Regarding the similarity of the EMA and PeerMA assessments between patients and their support persons, we noticed that the error metrics MAE, mean squared error, and root-mean-squared error are relatively small (around 10%-15%) for dyads 1 and 2, meaning that their regression curves representing the various assessments (hope, sleep, fatigue, depression, and pain) have a similar shape. We found the error metrics to be higher for dyads 3 to 7, with some of them showing high similarity; for instance, dyad 3 showing good similarity for fatigue, depression, and pain, and dyads 4, 5, and 7 showing good similarity for sleep quality. On the other hand, aside from dyad 2, The R^2 metric reflects low-quality models in almost all the cases. It is natural to expect deviations because the support person's assessments are not intended to be as exact as the patient's assessments, but instead, a close approximation built from subjective perceptions issued by the patient's support person. There are many reasons to explain differences in assessments between a patient and his/her support person. One could be a time shift, meaning that some time elapsed from the moment the patient manifested a certain behavior or attitude to the moment the support person perceived such a change and was able to report it in a PeerMA. The type of relationship and the degree of spatio-temporal proximity between a patient and a support person may be an important factor favoring a support persons' ability to better perceive the emotional states of a patient [16-18]. However, we also know that self-reports (including EMA) may be affected by known problems such as socially desirable responding when a person answers questions untruthfully to attain a desirable outcome or consciously misrepresenting his/her actual state [13]. In summary, looking at the similarity between a patient's and a support person's frequent assessments is informative when analyzed carefully, considering the patient's history, personal profile, as well as the type of relationship with his/her support person when possible.

Although we presented correlation results between EMA and PeerMA values, we do not discuss their statistical significance because they derive from a small sample of patients. We mostly wanted to expose the results in detail to motivate the reader and to support the identified findings. Despite that limitation, as shown in Table 11, the majority (77%) of correlations between patients and their support person assessments are positive (46% high, 20% moderate, and 11% weak). It means that for some dyads, the information streams could be valuable for medical personnel studying aspects of disease progression in a patient. For instance, if enough confidence is gained from the support person's assessments about a patient, medical personnel could feed in information from the support person during a given day when a patient is unable to self-report his/her own state. In some cases, the correlations were negative. This is worth studying as well because that could be related to something simple, such as a support person failing to perceive and report the real patient's state. However, it may also be more critical, such as a patient intentionally misrepresenting his/her actual state.

Regarding the identified human factors in this study, we learned that it is particularly important to promote self-awareness and add value to all the participants as they use the tools of EMA and PeerMA. It is worth exploring ways for the mobile app to show selected, meaningful EMA or PeerMA information to all the patients and support persons during the study (a form of compensation known as "in-kind"). We believe doing so will raise engagement and motivation, especially if done in near real-time and linked to the clinical decision-making for that particular patient. The lack of motivation leads to boredom for the patient and the support person. This unintended effect has been studied by Taylor et al [51]; van Berkel et al [52]; Mehrotra et al [53]; and Pejovic and Musolesi [54], who investigate gamification techniques (eg, scoring points, winning prizes, solving puzzles, and others) to increase engagement and participants' satisfaction during mobile subject studies. Making the mobile apps more user-informative and interactive may positively influence the acceptance of EMA/PeerMA-based studies. Some of the patients suggested the addition of more app features to enhance their experience, for example, "add more elements to assess such as more activities, calories burnt and weight," "questions may not be clear—glossary of terms of apps would help, eg, what is pain vs discomfort," "when I get into how well you are sleeping, the scale needs to be thicker, or I should be able to have an option of drop-down, to be expressed differently." However, these must be clearly co-designed with the care team to enhance the patient's self-management skills while assuring the validity of the collected data.

XSL•FO
RenderX

Another important aspect concerns the selection of support persons. Because we recruited the patients during their regular visits to the hospital, their support person was already present for each patient. All the support persons were spouses of the respective patients, which gave us confidence that they were close enough to the patient and were worthy to participate in this exploratory study. However, in general, it should be evident that to use the PeerMA method effectively, the involved support persons should meet specific inclusion criteria at the individual level, as well as criteria regarding his or her relationship with the patient. These criteria may vary based on the nature or main attributes of the phenomena under study. The following are examples of such criteria that are derived from our research. (1) The patient and support person must have a close relationship, that is, in frequent contact with the patient. The overall frequency will depend on the phenomena under study. For instance, to study rapidly oscillating states such as changes in affect, support persons should be in contact with the patient a few times within a day. However, 2 or 3 interactions per week may still be effective for the PeerMA assessment to explore traits such as anxiety. From the reviewed literature [16-18], we know that the social proximity of the relationship is essential (eg, spouse, partner, boyfriend, girlfriend, close friends, and so on) to obtain truthful and meaningful results from the PeerMAs. (2) Patients should trust the support person. During PeerMA-based studies, support persons may start to think about the patient more often than they used to. By doing so and answering PeerMAs, support persons may become more aware of the states being experienced by the patient (or presumably close to the real state). Because these states are private to the patient, PeerMA will be more effective when the patients have high trust in the support person, such that they both feel confident and natural during the study. (3) Support persons should be skilled to perform the task. To begin with, support persons need procedural skills, including the correct use of the smartphone to keep it charged and connected to the internet, answering the PeerMA prompts on time, and providing honest responses. Additionally, support persons need cognitive skills, including the ability to observe and perceive the states of the patient (ie, interpersonal sensitivity) projected in countless forms, such as verbal and nonverbal expressions, behavior changes, and others. (4) Support persons should be motivated to participate. Motivation is not sufficient to perform as a support person, but it is necessary for the success of PeerMA-based studies. When recruiting support persons, researchers should clearly explain the implications of their task, as well as the unique value they bring to these types of studies. However, it must also be clear that playing the role of a support person represents commitment. (5) Support persons should be available to issue assessments. The degree of availability depends on the phenomena under study. For instance, if the PeerMA prompts occur during the day, the support persons should be available to answer the prompts as close to the time they are triggered as possible. Availability is essential to increase the ecological validity and significance of the PeerMA, which increases when they are answered within the same time window as when the EMA is responded to by the patient.

Finally, one may wonder about the optimal number of support persons to invite for each patient. In this study, we experimented only with one support person per patient since, in our sample, it was not possible to recruit 2 or more support persons for each patient. Having more than one support person would favor the collection of more observations, which has the potential to enhance the analyses; however, researchers should define this number based on the support persons' alignment with the inclusion criteria of each study.

Limitations

The first limitation of this work is the small population sample of 8 patients and their support persons. In our study, the support person of each patient always turned out to be the patient's spouse, which is beneficial as we know the support person is close enough to the patient, but it also limits the generalizability of the results in our study. A second important limitation is the fact that part of the study occurred during the early months of 2020, and the patient-support person dynamics may have been altered by the world pandemic due to the COVID-19 virus. Additionally, study participants (both patients and their support persons) were recruited from only one location, the Liver Transplant Clinic at Stanford University Hospital. Consequently, the results may be biased towards this particular group of participants. We also consider as a limitation the absence of ground truth about the state of the patients and support persons when they answer EMAs and PeerMAs about the patients' states, like fatigue [37], depression [38], hope [35], pain [39], and quality of sleep [36] during the study. Although the main goal of the studies was not to assess the reliability or agreement between the self-reported states (via EMA) and the actual state of the patient (ground truth), the lack of an objective measurement of the patient's states affects the analysis of the peer-reported assessments (via PeerMA). Likewise, a related limitation is the lack of information about the physical and mental state of the support person when they issue assessments about the patient. We know that the state of the support persons, including their perceptions and personal beliefs, may bias how they perceive and report particular states occurring to the patient [55]. Another limitation is the lack of baseline or historical EMAs from the patients before they started the protocol, that is, before being "observed" by a support person [56]. The social dynamics between patients and support persons (even if they are family or close acquaintances) may induce a change in the typical behavior of the patient once they know that their support person is assessing them via PeerMA.

Conclusions

In conclusion, we presented the results of a study implying a new, noninvasive data collection method named PeerMA to complement the well-known method of EMA. The motivation came from the limitation that patients are sometimes unable to provide reliable self-assessments of their mental or physical states (eg, due to the complexity of introspection and limitations due to health status, such as pre- or post-operatory, among others). In those cases, support persons could fill in the gaps in assessments and provide their own assessments about the patient based on the information they can perceive from them. In this pioneering study, we explored the PeerMA method in the context of real-life phenomena, collecting qualitative and quantitative data to assess the value of the method and the reason

XSL•FO
RenderX

for its application in clinical care. The study was conducted with patients waiting for or after undergoing a liver transplant. We found empirical evidence about the feasibility of PeerMA as a method and showed that it provides unique information streams that are not available in traditional EMA-based studies. More precisely, we showed examples where support person assessments, as well as wearable-based physical activity data, can be used to inform health professionals about the actual state of a patient regarding important health outcomes such as hope, sleep quality, fatigue, pain, and depression. We identified and discussed 4 human factors that influence the acceptance of the PeerMA method. Overall, we show that it is possible to leverage the data acquired via the PeerMA method as well as wearable activity datasets as complementary to EMA to better study and understand the disease recovery pathways of high-risk populations, such as patients recovering from an organ transplant.

Future Work

Because this is the first study using the PeerMA method in patient populations, we identified several avenues that are worth exploring in future studies to obtain more information about the strengths and weaknesses of the method. The first one implies a type of new interaction where we ask the support person to only "validate" the EMA scores issued by the patient. In this scenario, patients would respond to the EMA (as usual), and at that point, a PeerMA would be sent to the support person. But, instead of asking the support person to complete their independent PeerMA assessment, their task is to react upon it, for instance: (1) they agree/disagree with the EMA; (2) they suggest their own PeerMA assessment; or (3) they opt not to respond. Related to that, we think it is important to validate the implications of using rating scales that were not designed for observer assessments; those rating scales were originally designed for self-assessments [29]. Similarly, there is a need to validate the implications, in both self and observer assessments, of using rating scales for frequently repeated measures when those rating scales were originally designed to be applied just one time, or only very infrequently. It is important to define the frequency of the assessments that researchers use in EMA-based studies, considering the frequency at which the evaluated human states naturally oscillate in the sampled population of patients (eg, hope, fatigue, depression, and pain, among other states).

A second aspect to explore concerns the roles of the participants in the studies. PeerMA-based studies imply the actions of a social group, that is, a patient and one or more support persons. This study used only unidirectional interactions, and both the patient and the support person issued assessments about the patient, which indirectly enabled social support to be provided to the patient by the support person. Future work will study the possible benefits of bidirectional interactions; when the cognitive load of the patient allows it, participants could take both roles of patient and support person simultaneously; that way, each person completes EMAs about their own state and PeerMAs about the other person. This variation may have positive side effects, such as increasing engagement by virtue of the implied mutual support. However, it may also make it difficult to design the study to ensure the validity of the collected data.

Another aspect to explore is the degree of expressiveness embedded in EMA and PeerMA questions. This study used only validated numeric scales, which limits how much one can assess one's own emotions, feelings, or mental or physical state. We would like to explore other assessment alternatives for patients and support persons. The first one is to openly choose the moods, states, or behaviors they wish to report at a given moment, instead of having to assess predefined states over and over. The possible states could be picked from a predefined list to preserve consistency across patients and reduce patient burden. This may have positive effects in cases where individuals feel the monotony of regularly assessing the same fixed states. However, such a patient-driven EMA/PeerMA data collection design may induce bias stemming from collecting only the states that patients and support persons want to report, which may or may not be sufficiently valid or relevant to the study goals or to be used for clinical decision making. Another alternative is to allow participants to describe the current state/behaviors rather than just producing ratings on a scale. This could be beneficial for certain patients to do introspection of their states/behaviors, for example, those who are at risk of relapses from certain addictions or behaviors they wish to change. Instead of typing in text, they could record audio describing their current state. A third alternative is to allow patients to define the states/behaviors they want their support persons to assess. Patients could define (eg, by choosing from a predefined list, or by proposing) the state or construct that they want (and feel comfortable with) their support persons to assess. Possible advantages include the reduction of bias of being observed [56] and emerging knowledge, such as understanding the natural emergence of states that patients allow their support persons to assess (rather than being imposed upon them). Nevertheless, a direct disadvantage is that this type of design may not be directly valuable for clinical researchers who are interested in studying a particular human state that may have clinical, health outcomes, and quality of care implications, or may not be directly valuable for clinical decision-making.

We also would like to study the effects of patients' and support persons' proximity when they participate in EMA/PeerMA studies. The general stance is that being often physically close to the patient adds higher confidence to the support person's reports; however, this may not be guaranteed, as certain human states are hard to detect immediately, even when in front of the individual experiencing that state. In addition, the ability to detect someone's state varies among individuals. As such, not being physically close to the patient or not having had a recent face-to-face encounter does not preclude support persons from reporting with high confidence. Nowadays, individuals can communicate easily via several channels, leveraging a variety of smartphone messaging apps and social networking apps, which may provide enough cues for support persons to confidently infer the state of a given person they are acquainted with. In our study, the EMAs and PeerMAs were triggered based on an interval contingent trigger [57] (ie, using fixed time scheduling), which does not discriminate among physical proximity states. Physical proximity between patients and support persons could be inferred using technologies such as Bluetooth, WiFi signals, and Geo-fences, among other methods to infer a person's location. Identification of digital proximity,

XSL•FO
RenderX

on the other hand, could leverage usage patterns of social network apps and communication exchanges via SMS and phone calls, among others. A less sophisticated but sufficient method could be asking the support person directly, “Is the patient next to you to answer a short survey?” or “Have you recently been in contact with the patient to answer a short survey?” If the answer is no, the support person could still decide if he/she wants to answer the PeerMA survey and provide a confidence level based on his/her knowledge of the target person. Linking EMA and PeerMA surveys to physical or digital proximity may help reduce the burden on patients and support persons and ensure higher data quality.

Additionally, in this study, support persons were able to express their confidence level for each individual PeerMA answer. They may have selected low confidence if they had not been in contact (either physical or digital) with the patient in the recent past. However, our design asked the question of confidence at the end of each PeerMA, which forced the support person to complete the surveys, even in cases where they had low confidence. An alternative approach can be that, after a few days (when the support persons become familiar with the types of questions in the surveys), the question of confidence appears at the beginning of the survey, and then, based on the first response, an algorithmic decision is made to present, or not, the actual PeerMA questions. This may reduce the burden, especially on the support person’s side.

Acknowledgments

The authors acknowledge the Swiss Federal Commission for Scholarships for Foreign Students (2016-19), H2020 WellCo (numbers 769765 and 2018-21), AAL CoME (2014-7-127), GUARDIAN, Swiss NSF project (157003, IZSEZ0_214593), University of Geneva (coQOL@hip2nek), Swiss Universities (P-13 AGE-INT), UCPH Data+ AI@CARE, and the University of Costa Rica. The authors attest that there was no use of generative artificial intelligence technology in the generation of text, figures, or other informational content of this manuscript.

Data Availability

The data sets generated and analyzed during this study are not publicly available because the data usage authorized by study participants via informed consent was restricted to the study researchers. Please refer to the corresponding author for reasonable data requests.

Conflicts of Interest

None declared.

References

1. Stone AA, Shiffman S. Ecological momentary assessment (EMA) in behavioral medicine. Ann Behav Med 1994;16(3):199-202. [doi: 10.1093/abm/16.3.199]
2. Berrocal A, Wac K. Peer-vasive computing: leveraging peers to enhance the accuracy of self-reports in mobile human studies. : ACM Press; 2018 Presented at: Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers - UbiComp'18; 2008 October 8-12; Singapore p. 600-605. [doi: 10.1145/3267305.3267542]
3. Berrocal A, Concepcion W, De Dominicis S, Wac K. Complementing human behavior assessment by leveraging personal ubiquitous devices and social links: an evaluation of the peer-ceived momentary assessment method. JMIR Mhealth Uhealth 2020 Aug 07;8(8):e15947 [FREE Full text] [doi: 10.2196/15947] [Medline: 32763876]
4. Versluis A, Verkuil B, Lane RD, Hagemann D, Thayer JF, Brosschot JF. Ecological momentary assessment of emotional awareness: preliminary evaluation of psychometric properties. Curr Psychol 2018 Nov 24;40(3):1402-1410. [doi: 10.1007/s12144-018-0074-6]
5. Burns MN, Begale M, Duffecy J, Gergle D, Karr CJ, Giangrande E, et al. Harnessing context sensing to develop a mobile intervention for depression. J Med Internet Res 2011 Aug 12;13(3):e55 [FREE Full text] [doi: 10.2196/jmir.1838] [Medline: 21840837]
6. MacKerron G, Mourato S. Happiness is greater in natural environments. Glob Environ Change 2013 Oct;23(5):992-1000. [doi: 10.1016/j.gloenvcha.2013.03.010]
7. Runyan JD, Steinke EG. Virtues, ecological momentary assessment/intervention and smartphone technology. Front Psychol 2015;6:481 [FREE Full text] [doi: 10.3389/fpsyg.2015.00481] [Medline: 25999869]
8. Santangelo P, Ebner-Priemer UW, Trull TJ. Experience sampling methods in clinical psychology. In: The Oxford Handbook of Research Strategies for Clinical Psychology. Oxford: University Press; 2013.
9. Trull TJ, Ebner-Priemer UW. Using experience sampling methods/ecological momentary assessment (ESM/EMA) in clinical assessment and clinical research: introduction to the special section. Psychol Assess 2009 Dec;21(4):457-462 [FREE Full text] [doi: 10.1037/a0017653] [Medline: 19947780]

XSL•FO
RenderX

10. Yang YS, Ryu GW, Choi M. Methodological strategies for ecological momentary assessment to evaluate mood and stress in adult patients using mobile phones: systematic review. JMIR Mhealth Uhealth 2019 Apr 01;7(4):e11215 [FREE Full text] [doi: 10.2196/11215] [Medline: 30932866]
11. Beal DJ. ESM 2.0: state of the art and future potential of experience sampling methods in organizational research. Annu Rev Organ Psychol Organ Behav 2015 Apr 10;2(1):383-407. [doi: 10.1146/annurev-orgpsych-032414-111335]
12. Consolvo S, Walker M. Using the experience sampling method to evaluate ubicomp applications. IEEE Pervasive Comput 2003 Apr;2(2):24-31. [doi: 10.1109/MPRV.2003.1203750]
13. Mayo NE, Figueiredo S, Ahmed S, Bartlett SJ. Montreal accord on patient-reported outcomes (PROs) use series - paper 2: terminology proposed to measure what matters in health. J Clin Epidemiol 2017 Sep;89:119-124. [doi: 10.1016/j.jclinepi.2017.04.013] [Medline: 28433673]
14. Vazire S. Informant reports: a cheap, fast, and easy method for personality assessment. J Res Pers 2006 Oct;40(5):472-481. [doi: 10.1016/j.jrp.2005.03.003]
15. Balsis S, Cooper LD, Oltmanns TF. Are informant reports of personality more internally consistent than self reports of personality? Assessment 2015 Aug;22(4):399-404 [FREE Full text] [doi: 10.1177/1073191114556100] [Medline: 25376588]
16. Gomà-i-Freixanet M. Consensual validity of the EPQ: self-reports and spouse-reports. Eur J Psychol Assess 1997 Sep;13(3):179-185. [doi: 10.1027/1015-5759.13.3.179]
17. Gosling SD, John OP, Craik KH, Robins RW. Do people know how they behave? self-reported act frequencies compared with on-line codings by observers. J Pers Soc Psychol 1998 May;74(5):1337-1349. [doi: 10.1037//0022-3514.74.5.1337] [Medline: 9599447]
18. Watson D, Hubbard B, Wiese D. Self-other agreement in personality and affectivity: the role of acquaintanceship, trait visibility, and assumed similarity. J Pers Soc Psychol 2000 Mar;78(3):546-558. [doi: 10.1037//0022-3514.78.3.546] [Medline: 10743880]
19. Kromm W, Gadinger MC, Schneider S. Peer ratings of chronic stress: can spouses and friends provide reliable and valid assessments of a target person's level of chronic stress? Stress Health 2010 Jan 21;26(4):292-303. [doi: 10.1002/smi.1297]
20. Munson SA, Leary KO, Bhattacharya A, Munson SA, Wobbrock JO, Pratt W. Design opportunities for mental health peer support technologies. 2017 Presented at: CSCW '17: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing; 2017 February 25; Portland, USA. [doi: 10.1145/2998181.2998349]
21. Neumann PJ, Araki SS, Gutterman EM. The use of proxy respondents in studies of older adults: lessons, challenges, and opportunities. J Am Geriatr Soc 2000 Dec;48(12):1646-1654. [doi: 10.1111/j.1532-5415.2000.tb03877.x] [Medline: 11129756]
22. Cusick CP, Brooks CA, Whiteneck GG. The use of proxies in community integration research. Arch Phys Med Rehabil 2001 Aug;82(8):1018-1024. [doi: 10.1053/apmr.2001.25098] [Medline: 11494179]
23. Duncan PW, Lai SM, Tyler D, Perera S, Reker DM, Studenski S. Evaluation of proxy responses to the Stroke Impact Scale. Stroke 2002 Nov;33(11):2593-2599. [doi: 10.1161/01.str.0000034395.06874.3e] [Medline: 12411648]
24. Sneeuw KCA, Aaronson NK, de Haan RJ, Limburg M. Assessing quality of life after stroke. the value and limitations of proxy ratings. Stroke 1997 Aug;28(8):1541-1549. [doi: 10.1161/01.str.28.8.1541] [Medline: 9259746]
25. Sneeuw KCA, Aaronson NK, Osoba D, Muller M, Hsu MA, Yung WK, et al. The use of significant others as proxy raters of the quality of life of patients with brain cancer. Med Care 1997 May;35(5):490-506 [FREE Full text] [doi: 10.1097/00005650-199705000-00006] [Medline: 9140337]
26. Sneeuw KCA, Aaronson NK, Sprangers MA, Detmar SB, Wever LD, Schornagel JH. Comparison of patient and proxy EORTC QLQ-C30 ratings in assessing the quality of life of cancer patients. J Clin Epidemiol 1998 Jul;51(7):617-631. [doi: 10.1016/s0895-4356(98)00040-7] [Medline: 9674669]
27. Sneeuw KCA, Albertsen PC, Aaronson NK. Comparison of patient and spouse assessments of health related quality of life in men with metastatic prostate cancer. J Urol 2001 Feb;165(2):478-482. [doi: 10.1097/00005392-200102000-00029] [Medline: 11176400]
28. O'Kane AA, Park SY, Mentis H, Blandford A, Chen Y. Turning to peers: integrating understanding of the self, the condition, and others' experiences in making sense of complex chronic conditions. Comput Support Coop Work 2016;25(6):477-501 [FREE Full text] [doi: 10.1007/s10606-016-9260-y] [Medline: 32355411]
29. Roydhouse JK, Cohen ML, Eshoj HR, Corsini N, Yucel E, Rutherford C, ISOQOL Proxy Task Forcethe ISOQOL Board of Directors. The use of proxies and proxy-reported measures: a report of the international society for quality of life research (ISOQOL) proxy task force. Qual Life Res 2022 Feb;31(2):317-327. [doi: 10.1007/s11136-021-02937-8] [Medline: 34254262]
30. Liu JB, Pusic AL, Temple LK, Ko CY. Patient-reported outcomes in surgery: listening to patients improves quality of care. Bull Am Coll Surg 2017 Mar;102(3):19-23. [Medline: 28920657]
31. Dunn MA, Rogal SS, Duarte-Rojo A, Lai JC. Physical function, physical activity, and quality of life after liver transplantation. Liver Transpl 2020 May;26(5):702-708 [FREE Full text] [doi: 10.1002/lt.25742] [Medline: 32128971]
32. Golfieri L, Gitto S, Vukotic R, Andreone P, Marra F, Morelli MC, et al. Impact of psychosocial status on liver transplant process. Ann Hepatol 2019;18(6):804-809 [FREE Full text] [doi: 10.1016/j.aohep.2019.06.011] [Medline: 31471202]

XSL•FO
RenderX

33. Schulz KH, Kroencke S. Psychosocial challenges before and after organ transplantation. Transpl Res Risk Manag 2015 Aug:45. [doi: 10.2147/trrm.s53107]
34. Wulfovich S, Fiordelli M, Rivas H, Concepcion W, Wac K. “I must try harder”: design implications for mobile apps and wearables contributing to self-efficacy of patients with chronic conditions. Front Psychol 2019;10:2388 [FREE Full text] [doi: 10.3389/fpsyg.2019.02388] [Medline: 31749733]
35. Snyder CR, Sympson SC, Ybasco FC, Borders TF, Babyak MA, Higgins RL. Development and validation of the State Hope Scale. J Pers Soc Psychol 1996 Feb;70(2):321-335. [doi: 10.1037//0022-3514.70.2.321] [Medline: 8636885]
36. Buysse DJ, Reynolds CF, Monk TH, Berman SR, Kupfer DJ. The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research. Psychiatry Res 1989 May;28(2):193-213. [doi: 10.1016/0165-1781(89)90047-4] [Medline: 2748771]
37. Learmonth YC, Dlugonski D, Pilutti LA, Sandroff BM, Klaren R, Motl RW. Psychometric properties of the Fatigue Severity Scale and the Modified Fatigue Impact Scale. J Neurol Sci 2013 Aug 15;331(1-2):102-107. [doi: 10.1016/j.jns.2013.05.023] [Medline: 23791482]
38. Kroenke K, Spitzer RL, Williams JBW, Löwe B. An ultra-brief screening scale for anxiety and depression: the PHQ-4. Psychosomatics 2009;50(6):613-621. [doi: 10.1176/appi.psy.50.6.613] [Medline: 19996233]
39. Rosenzveig A, Kuspinar A, Daskalopoulou SS, Mayo NE. Toward patient-centered care: a systematic review of how to ask questions that matter to patients. Medicine (Baltimore) 2014;93(22):e120. [doi: 10.1097/md.0000000000000120]
40. De Masi A, Ciman M, Gustarini M, Wac K. MQoL smart Lab: quality of life living lab for interdisciplinary experiments. 2016 Presented at: Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct; 2016 Sep 12; United States. [doi: 10.1145/2968219.2971593]
41. Berrocal A, Manea V, Masi AD, Wac K. mQoL lab: step-by-step creation of a flexible platform to conduct studies using interactive, mobile, wearable and ubiquitous devices. Procedia Comput Sci 2020;175:221-229. [doi: 10.1016/j.procs.2020.07.033]
42. Fahrenberg J, Myrtek M, Pawlik K, Perrez M. Ambulatory assessment - monitoring behavior in daily life settings. Eur J Psychol Assess 2007 Jan;23(4):206-213 [FREE Full text] [doi: 10.1027/1015-5759.23.4.206]
43. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. J Health Soc Behav 1983 Dec;24(4):385. [doi: 10.2307/2136404]
44. Stewart A, Sherbourne CD, Ware JE. Measuring Functioning and Well-Being: The Medical Outcomes Study Approach. United States: Duke University Press; 1992.
45. Uher J. Quantitative data from rating scales: an epistemological and methodological enquiry. Front Psychol 2018;9:2599 [FREE Full text] [doi: 10.3389/fpsyg.2018.02599] [Medline: 30622493]
46. Buysse DJ. Sleep health: can we define it? does it matter? Sleep 2014 Jan 01;37(1):9-17 [FREE Full text] [doi: 10.5665/sleep.3298] [Medline: 24470692]
47. van Berkel N, Ferreira D, Kostakos V. The experience sampling method on mobile devices. ACM Comput Surv 2017 Dec 06;50(6):1-40. [doi: 10.1145/3123988]
48. Suh H, Shahriaree N, Hekler EB, Kientz JA. Developing and validating the user burden scale: a tool for assessing user burden in computing systems. Conf Hum Factors Comput Syst - Proc 2016:3988-3999. [doi: 10.1145/2858036.2858448]
49. Schlichting M, Hennig M, Rudell K, McLeod L, Bennett B, Shaw J, et al. Is IQWiG's 15% threshold universally applicable in assessing the clinical relevance of patient-reported outcomes changes? an ISPOR special interest group report. Value Health 2022 Sep;25(9):1463-1468 [FREE Full text] [doi: 10.1016/j.jval.2022.07.010] [Medline: 36049796]
50. Berrocal A, Wac K. Peers know you: a feasibility study of the predictive value of peer’s observations to estimate human states. Procedia Comput Sci 2020;175:205-213. [doi: 10.1016/j.procs.2020.07.031]
51. Taylor S, Ferguson C, Peng F, Schoeneich M, Picard RW. Use of in-game rewards to motivate daily self-report compliance: randomized controlled trial. J Med Internet Res 2019 Jan 03;21(1):e11683 [FREE Full text] [doi: 10.2196/11683] [Medline: 30609986]
52. van Berkel N, Goncalves J, Hosio S, Kostakos V. Gamification of mobile experience sampling improves data quality and quantity. Proc ACM Interact Mob Wearable Ubiquitous Technol 2017 Sep 11;1(3):1-21. [doi: 10.1145/3130972]
53. Mehrotra A, Pejovic V, Vermeulen J, Hendley R, Musolesi M. My phone and me: understanding people's receptivity to mobile notifications. 2016 Presented at: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems; 2016 May 07; United States p. 1021-1032. [doi: 10.1145/2858036.2858566]
54. Pejovic V, Musolesi M. InterruptMe: designing intelligent prompting mechanisms for pervasive applications. 2014 Presented at: UbiComp '14: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing; 2014 Sep 13; USA p. 897-908. [doi: 10.1145/2632048.2632062]
55. Muraven M. Ego depletion: theory and evidence. Oxf Handb Hum Motiv 2012:1-29. [doi: 10.1093/oxfordhb/9780195399820.013.0007]
56. Levitt SD, List JA. What do laboratory experiments measuring social preferences reveal about the real world? J Econ Perspect 2007;21(2):153-174. [doi: 10.1257/jep.21.2.153]

XSL•FO
RenderX

57. van Berkel N, Goncalves J, Lovén L, Ferreira D, Hosio S, Kostakos V. Effect of experience sampling schedules on response rate and recall accuracy of objective self-reports. Int J Hum Comput Stud 2019 May;125:118-128. [doi: 10.1016/j.ijhcs.2018.12.002]

Abbreviations

EMA: ecological momentary assessment
MAE: mean absolute error
PeerMA: peer-ceived momentary assessment

Edited by A Mavragani; submitted 28.12.23; peer-reviewed by P Robles-Granda, R Pryss, C Nowinski; comments to author 10.10.24; revised version received 31.01.25; accepted 23.04.25; published 27.11.25.

Please cite as:
Berrocal A, Concepcion W, Wac K
A 6-Month Evaluation of the Peer-Ceived Momentary Assessment Method in a Small Sample of Liver Transplant Patients and Their Support Persons: Longitudinal Observational Study
JMIR Form Res 2025;9:e55907
URL: https://formative.jmir.org/2025/1/e55907
doi:10.2196/55907
PMID:

©Allan Berrocal, Waldo Concepcion, Katarzyna Wac. Originally published in JMIR Formative Research (https://formative.jmir.org), 27.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Comparison of ChatGPT and DeepSeek on a Standardized Audiologist Qualification Examination in Chinese: Observational Study

Beier Qi[1], MS; Yan Zheng[1], BS; Yuanyuan Wang[1], BS; Li Xu[2], MD, PhD

[1]Beijing Tongren Hospital, Capital Medical University, Key Laboratory of Otolaryngology - Head and Neck Surgery (Capital Medical University), Ministry of Education, Beijing, China, Beijing, China

[2]Department of Hearing, Speech and Language Sciences, Ohio University, Athens, OH, United States

Corresponding Author:
Li Xu, MD, PhD
Department of Hearing, Speech and Language Sciences
Ohio University
20 E Union St
Athens, OH, 45701
United States
Phone: 1 740 593 0310
Fax: 1 740 593 0287
Email: xul@ohio.edu

Abstract

Background: Generative artificial intelligence (GenAI), exemplified by ChatGPT and DeepSeek, is rapidly advancing and reshaping human-computer interaction with its growing reasoning capabilities and broad applications across fields such as medicine and education.

Objective: This study aimed to evaluate the performance of 2 GenAI models (ie, GPT-4-turbo and DeepSeek-R1) on a standardized audiologist qualification examination in Chinese and to explore their potential applicability in audiology education and clinical training.

Methods: The 2024 Taiwan Audiologist Qualification Examination, comprising 300 multiple-choice questions across 6 subject areas (ie, basic hearing science, behavioral audiology, electrophysiological audiology, principles and practice of hearing devices, health and rehabilitation of the auditory and balance systems, and hearing and speech communication disorders [including professional ethics]), was used to assess the performance of the 2 GenAI models. The complete answering process and reasoning paths of the models were recorded, and performance was analyzed by overall accuracy, subject-specific scores, and question-type scores. Statistical comparisons were performed at the item level using the McNemar test.

Results: ChatGPT and DeepSeek achieved overall accuracies of 80.3% (241/300) and 79.3% (238/300), respectively, which are higher than the passing criterion of the Taiwan Audiologist Qualification Examination (ie, 60% correct answers). The accuracies for the 6 subject areas were 88% (44/50), 70% (35/50), 86% (43/50), 76% (38/50), 82% (41/50), and 80% (40/50) for ChatGPT and 82% (41/50), 72% (36/50), 78% (39/50), 80% (40/50), 80% (40/50), and 84% (41/50) for DeepSeek. No significant differences were found between the two models at the item level (overall P=.79), with a small effect size (accuracy difference=+1%, Cohen h=0.02, odds ratio 0.90, 95% CI 0.53–1.52) and substantial agreement (κ=0.71). ChatGPT scored highest in basic hearing science (88%), whereas DeepSeek performed the best in hearing and speech communication disorders (84%). Both models scored lowest in behavioral audiology (ChatGPT: 70% and DeepSeek: 72%). Question-type analysis revealed that both models performed well on reverse logic questions (ChatGPT: 79/95, 83%; DeepSeek: 80/95, 84%) but performed moderately on complex multiple-choice questions (ChatGPT: 9/17, 53%; DeepSeek: 11/17, 65%). However, both models performed poorly on graph-based questions (ChatGPT: 2/11, 18%; DeepSeek: 4/11, 36%).

Conclusions: Both GenAI models demonstrated strong professional knowledge and stable reasoning ability, meeting the basic requirements of clinical audiologists and suggesting their potential as supportive tools in audiology education. However, the presence of errors underscores the need for cautious use under educator supervision. Future research should explore their performance in open-ended, real-world clinical scenarios to assess practical applicability and limitations.

XSL•FO
RenderX

(JMIR Form Res 2025;9:e79534) doi:10.2196/79534

KEYWORDS

generative artificial intelligence; artificial intelligence; AI; ChatGPT; DeepSeek; audiology; medical education

Introduction

Artificial intelligence (AI), first conceptualized in 1956 [1], has made remarkable progress over the past decade in simulating human cognition and behavior, driving breakthroughs across multiple fields [2]. By constructing brain-inspired computing architectures, AI facilitates automated analysis and decision-making in complex tasks [3]. In recent years, generative AI (GenAI, a subfield of AI that uses generative models to produce text, images, videos, or other forms of data), a leading subfield of AI, has rapidly advanced, driven by algorithmic innovations and the exponential growth of data. Powered by large-scale language models (a language model trained with self-supervised machine learning on a vast amount of text, designed for natural language processing tasks, especially language generation), GenAI has demonstrated broad application potential across medicine, education, scientific research, commerce, and industry [4]. As an evolving form of AI equipped with logical reasoning capabilities, GenAI is fundamentally reshaping the way humans interact with computers.

In the field of education, the leading models of GenAI such as ChatGPT (OpenAI), Gemini Ultra (Google), Meta Llama 3, Anthropic Claude 3, and DeepSeek (DeepSeek) can not only efficiently respond to professional questions and generate a clear problem-solving process but also simulate the thinking process of experts to assist students in academic exploration, problem-solving, and decision-making training [4-6]. GenAI's exceptional generative capabilities offer new avenues for scholarly research, increased learning, classroom improvement, and knowledge sharing, and the interest in how these tools can benefit health education is also growing. Recent studies have shown that GenAI performs well in general medical examinations, such as the United States Medical Licensing Examination, the National Pharmacy Licensing Examination, and the National Nurse Licensing Examination, and has also demonstrated expert-level reasoning ability and diagnostic accuracy in the OTO Chautauqua Otolaryngology specialty examination [7-10]. As the quality and quantity of training data largely determine GenAI's ability to understand and generate text, its outstanding performance in English contexts may not necessarily extend directly to non-English applications. Existing studies have also found that GenAI's accuracy varies across different medical examinations and fields [11,12], with particularly inconsistent performance in examinations conducted in non–English-speaking regions [13,14]. Therefore, it is essential to further explore GenAI's performance in processing non-English information, especially its application potential in Chinese, the most widely spoken first language (12.3%) and the second most spoken language (13.8%) in the world [15].

Audiology is the scientific study of hearing, balance, and related disorders. Audiologists are health care professionals who deliver patient-centered care focused on preventing, identifying, diagnosing, and treating hearing and balance disorders using evidence-based approaches. These conditions can have wide-ranging effects (ie, medical, psychological, physical, social, educational, and occupational). By offering expert and personalized care, audiologists help reduce the impact of these disorders, ultimately enhancing patients' overall well-being and quality of life. In the United States, the doctor of audiology program is typically a 3- or 4-year graduate program. Doctor of audiology students are recommended to take the Praxis Examination in Audiology before their final year of their graduate study. Passing the Praxis Examination is a requirement for the American Speech-Language-Hearing Association Certificate of Clinical Competence in Audiology and state professional licensure in almost all 50 states in the United States [16]. The Educational Testing Service is responsible for administering the Praxis Examination, and the questions in the examination are not published. In Taiwan, an equivalent examination to the Praxis Examination, titled the Taiwan Audiologist Qualification Examination (TAQE), is published online every year [17]. The TAQE is written in traditional Chinese.

Previous studies have begun to explore the feasibility of GenAI in audiology education. Wang et al [18] evaluated the performance of ChatGPT (GPT-4) on the 2023 version of TAQE, showing an overall accuracy of 75%. This study suggests that GenAI possesses a certain level of professional knowledge in audiology and has the potential to support teaching and information acquisition. However, its abilities are still limited by the accuracy and scope of its knowledge, and its feasibility needs further verification.

In this study, the 2024 version of the TAQE was used to compare the performance of GPT-4-turbo and DeepSeek-R1 in audiology professional examinations. We aimed to gain insights into the potential of GenAI models in understanding audiology questions written in Chinese and to assess their applicability in audiology education and clinical practice. At the model level, compared with GPT-4 used by Wang et al [18], GPT-4-turbo demonstrates enhanced inference efficiency and cost-effectiveness, making it more suitable for practical applications. On the other hand, DeepSeek, a Chinese-developed GenAI, is optimized for Chinese language understanding and generation, thereby enhancing its adaptability and effectiveness in Chinese contexts. In addition, it adopts a reinforcement learning framework and incorporates a series of distilled models, which maintain performance while enabling efficient deployment in resource-constrained environments [19].

On the basis of this background, we propose 2 research hypotheses:

- The primary hypothesis is that GenAI models (GPT-4-turbo and DeepSeek-R1) are capable of achieving passing-level performance on the TAQE, thereby reaching the knowledge structure and reasoning ability required of clinical

XSL•FO
RenderX

audiologists and demonstrating their potential applicability in audiology education and clinical training.

- The secondary hypothesis is that compared with the general-purpose model (GPT-4-turbo), DeepSeek-R1 is expected to achieve superior performance in the TAQE, reflecting the advantages of language-specific optimization.

By comparing ChatGPT and DeepSeek, this study not only highlights the technological strengths of both models but also enables a comprehensive evaluation of these 2 leading GenAI systems in the context of professional examinations written in Chinese, thereby shedding light on whether language-specific models (eg, DeepSeek) are superior to general models (eg, ChatGPT).

Methods

Ethical Considerations

This study does not involve human participants and is therefore classified as a non–human subject study. As such, it does not require institutional review board approval or informed consent. All data used in this research were derived from publicly available question banks, and the analysis was conducted exclusively using AI models. The study strictly adheres to data use guidelines and ensures full transparency in model implementation and use.

AI Models

Two GenAI models were used in this study: GPT-4-turbo (released in April 2024) and DeepSeek-R1 (released in May 2024). GPT-4-turbo offers enhanced efficiency, a 128K-token context, and multimodal input [20]. DeepSeek-R1 provides optimized code and mathematical reasoning, a 128K-token context, and open weights for reproducibility. Although newer versions are available, these versions were the latest available GenAI models at the time of this research.

Test Materials

The test questions were from the officially released version of the 2024 TAQE [17]. The TAQE is a professional licensing examination for audiologists and covers six subject areas: (1) basic hearing science, (2) behavioral audiology, (3) electrophysiological audiology, (4) principles and practice of hearing devices, (5) health and rehabilitation of the auditory and balance systems, and (6) hearing and speech communication disorders (including professional ethics), with 50 questions in each subject area, for a total of 300 questions. All questions were in a multiple-choice format. The examination was written in traditional Chinese. All questions and answer sheets were downloaded from the official website [17] as a PDF document.

Procedure

We manually entered questions and options into the dialogue interface and recorded the model's original responses, explanations, and answers. A fixed prompt, "Please answer the following multiple-choice question as a hearing care professional, providing reasoning and explanation," was used to regulate the response style of the models and ensure the outputs were professionally relevant and interpretable (Figure 1). To ensure consistency and minimize potential human bias in the input process, all prompts were predetermined and entered by copy-and-paste rather than manual typing.

XSL•FO
RenderX

Figure 1. Examples of prompts and responses. Left: ChatGPT; right: DeepSeek. This example shows the first question from the basic auditory science examination. All Taiwan Audiologist Qualification Examination questions are multiple choice. The original examination questions are primarily in traditional Chinese and use bilingual (Chinese and English) terminology to express professional concepts.

1106

Please answer the following multiple-choice question as a hearing care professional, providing reasoning and explanation:

1 柯蒂氏器（organ of Corti）之內毛細胞與外毛細胞，其個別之支持細胞為何？
A. Deiters' cell；phalangeal cell B. phalangeal cell；Deiters' cell
C. Hensen's cell；Deiters' cell D. phalangeal cell；Hensen's cell

The correct answer is:
B. phalangeal cell；Deiters' cell
Explanation:
The organ of Corti, located in the cochlea, is responsible for sound transduction. It contains two types of sensory hair cells:
Inner Hair Cells (IHCs) – These are the primary auditory receptors that convert mechanical vibrations into neural signals. They are supported by phalangeal cells (also known as inner phalangeal cells), which provide structural and metabolic support.
Outer Hair Cells (OHCs) – These function in amplifying sound vibrations and improving frequency selectivity. They are supported by Deiters' cells (also known as outer phalangeal cells), which provide mechanical support and form cup-like structures to anchor the OHCs.
Thus, the correct association is:
Inner hair cells → Phalangeal cells
Outer hair cells → Deiters' cells
This matches answer choice B. phalangeal cell；Deiters' cell.

1106

Please answer the following multiple-choice question as a hearing care professional, providing reasoning and explanation:

1 柯蒂氏器（organ of Corti）之內毛細胞與外毛細胞，其個別之支持細胞為何？
A. Deiters' cell；phalangeal cell B. phalangeal cell；Deiters' cell
C. Hensen's cell；Deiters' cell D. phalangeal cell；Hensen's cell
Question 1: What are the supporting cells for the inner hair cells and outer hair cells in the organ of Corti, respectively?
A. Deiters' cell；phalangeal cell
B. phalangeal cell；Deiters' cell
C. Hensen's cell；Deiters' cell
D. phalangeal cell；Hensen's cell

Answer: B. phalangeal cell；Deiters' cell.
Explanation: The organ of Corti contains two types of hair cells: inner hair cells (IHCs) and outer hair cells (OHCs). Each type has specific supporting cells:
Outer hair cells (OHCs) are supported by Deiters' cells, which provide structural support and anchor the OHCs to the basilar membrane.
Inner hair cells (IHCs) are supported by phalangeal cells, which help maintain the structural integrity of the IHCs and their connection to the tectorial membrane.
Thus, the correct pairing is Deiters' cells for outer hair cells and phalangeal cells for inner hair cells, making B the correct answer.

For graph-based questions, both the graph and the text were provided. Although the original test questions were in traditional Chinese, certain scientific terms (eg, the organ of Corti) were expressed in both Chinese and English. Both GenAI models were instructed to respond in English to minimize the impact of language differences. Additionally, the source of the test questions was concealed before testing to prevent the models from drawing on specific training data or external knowledge sources. ChatGPT completed all tests first (February 3-13, 2025), followed by DeepSeek (February 14-24, 2025).

To facilitate detailed error and reasoning analysis, the original questions were classified into 4 categories: graph based, complex, reverse logic, and standard (Table 1).

Table 1. Description of question categories.

Category	Definition in this study	Questions, n (%)	Example
Graph-based questions	Graph in the question or answer	11 (3.7)	
Complex multiple-choice questions	Each option consists of multiple knowledge points	17 (5.7)	
Reverse logic questions	Negative expressions in the questions, such as "wrong," "avoid," "inappropriate," or "not including"	95 (31.7)	
Standard multiple-choice questions	Others	177 (59)	

This study categorized the original questions into 4 categories, namely, diagram-based questions, complex multiple-choice questions, reverse logic questions, and standard multiple-choice questions, for detailed analysis. Table 1 provides definitions, percentages, and representative examples to illustrate the visual and conceptual differences between the different question types.

Statistical Methods

Descriptive statistics were performed overall and by subject areas. Model performance was compared at the item level. The McNemar test was used to assess whether the 2 models differed in accuracy on the same items. Effect sizes were expressed as absolute differences in accuracy, odds ratios, and Cohen *h*. Kappa tests were used to test consistency. Statistical significance was set at $P<.05$. All analyses were performed using R (version 4.2.1).

XSL•FO
RenderX

Results

Both GenAI models achieved test scores exceeding the qualification examination's passing criterion of 60% accuracy, demonstrating overall strong performance. ChatGPT attained an overall score of 80%, with subject-specific scores ranging from 70% to 88%. DeepSeek achieved an overall score of 79%, with subject-specific scores ranging from 72% to 84%, as shown in Figure 2. McNemar test, applied at the item level within each of the 6 subject areas (50 questions each) and in the pooled analysis across all 300 questions, showed no significant differences in accuracy between the 2 models (*P*=.34-.99; overall *P*=.79). The overall effect size was small (accuracy difference=+1%; Cohen *h*=0.02; odds ratio 0.90, 95% CI 0.53–1.52). Cohen κ indicated substantial agreement between the 2 models' responses (κ=0.71, range 0.56-0.78), as presented in Table 2. These results indicate that the 2 models exhibited substantial response consistency despite minor fluctuations, suggesting highly similar overall capabilities in the audiologist qualification examination.

Figure 2. Performance on the 2024 Taiwan Audiologist Qualification Examination by ChatGPT and DeepSeek in the 6 subject areas (ie, basic hearing science, behavioral audiology, electrophysiological audiology, principles and practice of hearing devices, health and rehabilitation of the auditory and balance systems, and hearing and speech communication disorders [including professional ethics]) and the overall scores.

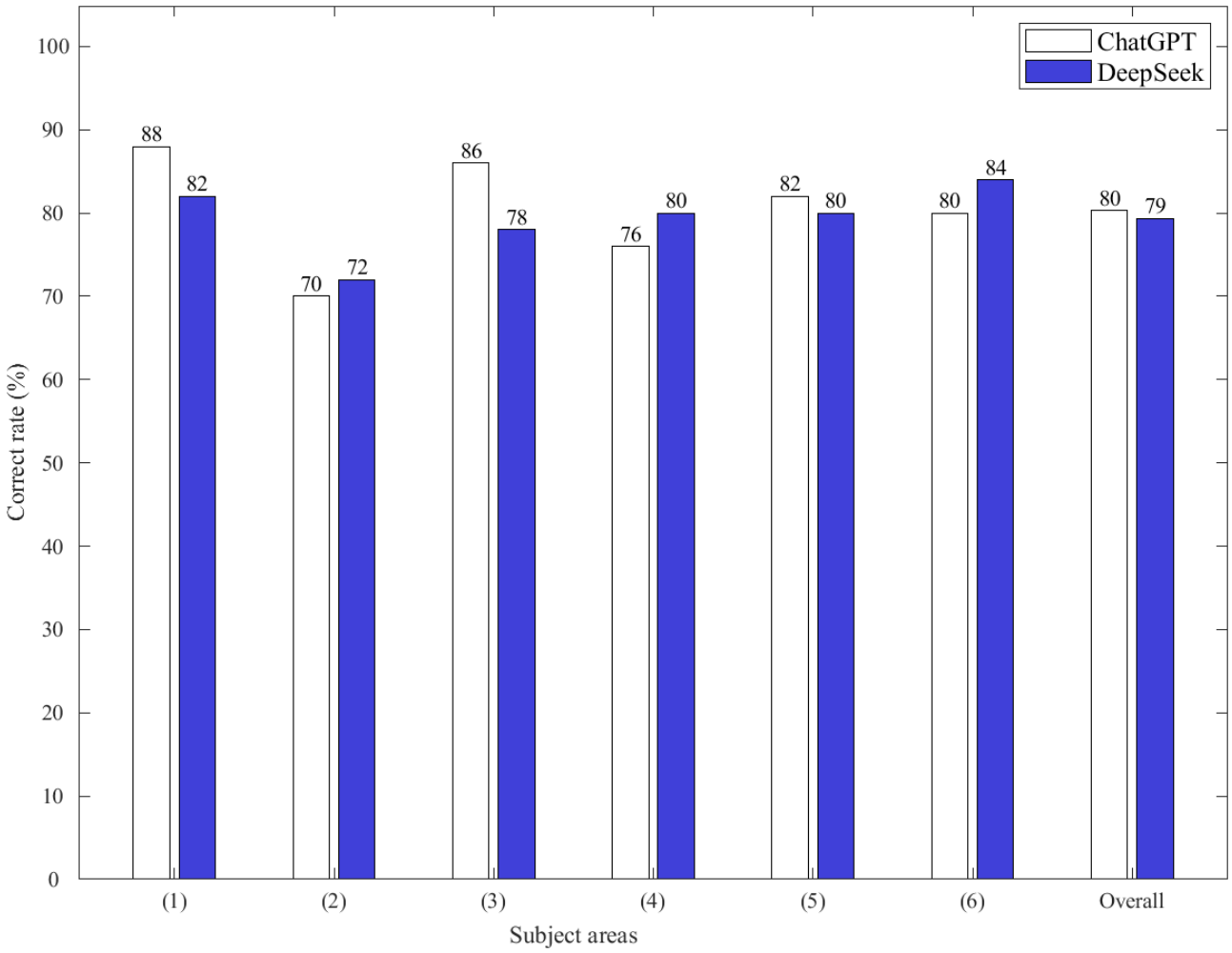

XSL•FO
RenderX

Table 2. Model capability comparison between ChatGPT and DeepSeek.

Subject area	GPT_ACC[a]	DS_ACC[b]	ACC_diff[c]	Cohen *h*	OR[d]	McNemar *P* value	κ
S1	0.88	0.82	+0.06	0.17	0.40 (0.08–2.06)	.45	0.78
S2	0.70	0.72	−0.02	−0.04	1.17 (0.39–3.47)	.99	0.56
S3	0.86	0.78	+0.08	0.21	0.43 (0.11–1.66)	.34	0.67
S4	0.76	0.80	−0.04	−0.10	2.00 (0.37–10.92)	.69	0.76
S5	0.82	0.80	+0.02	0.05	0.80 (0.21–2.98)	.99	0.72
S6	0.80	0.84	−0.04	−0.10	1.50 (0.42–5.32)	.75	0.70
Overall	0.803	0.793	+0.01	0.02	0.90 (0.53–1.52)	.79	0.71

[a]GPT_ACC: accuracy for ChatGPT.

[b]DS_ACC: accuracy for DeepSeek.

[c]ACC_diff: accuracy difference between GPT_ACC and DS_ACC.

[d]OR: odds ratio.

Subject-specific analysis showed that ChatGPT achieved its highest score in basic hearing science (88%), whereas DeepSeek performed best in hearing and speech communication disorders (including professional ethics; 84%). However, both models recorded their lowest scores in behavioral audiology, with ChatGPT at 70% and DeepSeek at 72% (Figure 2).

Question-type analysis revealed that both models underperformed on graph-based questions, with DeepSeek scoring 4 out of 11 (36% correct) and ChatGPT scoring 2 out of 11 (18% correct), both below the passing threshold. For complex multiple-choice questions, both approached the passing level, with DeepSeek scoring 11 out of 17 (65% correct) and ChatGPT scoring 9 out of 17 (53% correct). In reverse logic questions, both models performed well at comparable levels, with DeepSeek scoring 80 out of 95 (84% correct) and ChatGPT scoring 79 out of 95 (83% correct).

Discussion

This study evaluated the performance of 2 GenAI models in a standardized audiologist qualification examination in Chinese. The results showed that (1) both GenAI models passed the qualification examination, indicating that they met the basic requirements of clinical audiologists and possessed the necessary audiological knowledge, and (2) the scores of the GenAI models varied across subjects, with the lowest performance in behavioral audiology. Performance also differed by question type: both models performed poorly on graph-based questions but showed good performance on reverse logic questions and (3) no overall advantage of DeepSeek over ChatGPT was observed, suggesting that language-specific optimization did not translate into a clear superiority in the audiology examination written in Chinese.

This study compared the performance of 2 GenAI models in a real audiologist qualification examination. This examination covers the core areas of audiology, fully reflects the breadth and depth of audiology, and is an important reference for evaluating audiology professional capabilities. The basic requirements for candidates include a degree in audiology or a related field and at least 6 months or 375 hours of clinical practice experience [21]. The passing criterion of the examination has been set at 60% correct, and the pass rate for the 2024 TAQE is 87.61%.

Our test program simulated the process of the candidate's answer without external intervention. The overall accuracy of ChatGPT and DeepSeek were 80% and 79%, respectively. The results show that the 2 GenAI models can stably cope with the standardized (multiple-choice) audiology examination, have good audiology knowledge, and meet the basic standards for clinical audiologists. At the same time, the 2 GenAI models can provide a complete thinking process and clearly show the reasoning chain for solving problems, which makes them potentially useful in audiology education.

In this study, the performance of GenAI on the audiology examination (ie, a non-English examination) was similar to that of other professional examinations in Chinese, including the National Medical Qualification Examination, the National Pharmacist Qualification Examination, and the National Nursing Qualification Examination, with scores ranging from 73% to 84% using GPT-4.0 [10,22-24]. Although these scores are still lower than those of the English version, they are a significant improvement over GPT-3.5, indicating that GenAI is gradually adapting to the non-English examination environment.

In addition, Wang et al [18] evaluated the performance of GPT-4 on the 2023 TAQE, reporting an overall accuracy of 75%. In this study, testing on the 2024 TAQE, GPT-4-turbo (80% correct) and DeepSeek-R1 (79% correct) both outperformed the previous version of ChatGPT. Before concluding that improvements in the GenAI model led to better performance in audiology, we must consider that the difficulty level of the tests may differ between years. We reevaluated GPT-4-turbo on the 2023 TAQE and obtained a score of 76%, which was not significantly different from the score reported by Wang et al [18]. Thus, we do not have enough evidence to support the notion that the performance improvement is likely due to the improved GenAI model itself. To ensure comparability, this study followed the same approach as Wang et al [18], instructing the AI model to generate responses only in English and manually reviewing the accuracy of all outputs. Although no comprehension or accuracy issues attributable to bilingual processing were observed, potential bias arising from

XSL•FO
RenderX

cross-lingual input and output should be carefully considered in future AI evaluation studies.

Both models scored lowest on behavioral audiology, likely due to the complexity of the questions that combine theoretical knowledge with clinical case analysis. For example, the basic knowledge questions are as follows:

> *Question 2.* 聽力測試前進行耳鏡檢查之注意事項，下列何者錯誤？[Which of the following is incorrect regarding precautions for otoscopic examination before hearing tests?]
>
> *Question 4.* 在職業聽力保護計畫中，聽檢儀需多久進行一次聲學校正？[In a hearing conservation program, how often should the audiometer be acoustically calibrated?]

Furthermore, the clinical case–based questions are as follows:

> *Question 18* 針對雙側不對稱感音神經性聽力損失病人，根據語音辨識閾（*SRT*）及舒適聽閾推算遮蔽音設定. [For a patient with bilateral asymmetric sensorineural hearing loss, how should masking levels be set based on the speech recognition threshold (SRT) and the most comfortable listening level?]
>
> *Question 24:* 針對有非器質性聽損嫌疑的患者，從聽力檢查中辨別其表現特徵. [For a patient suspected of nonorganic hearing loss, how can audiological tests be used to identify characteristic patterns?]

Compared to the structured questions relying on memorization or definitions, GenAI models are prone to reasoning biases when dealing with highly interactive and context-dependent questions. Additionally, the behavioral audiology contained the highest number of graph-based questions (5 graphs for 6 questions), which are the question types where both GenAI models performed the worst. These factors might contribute to the poor performance of both GenAI models in this module.

Meanwhile, the poor performance in graph-based questions is consistent with the findings of Wang et al [18] and may be partly due to the low image quality and resolution of the TAQE materials. These results suggest that current GenAI still relies on high-quality original images for effective graphical processing and also highlight limitations in multimodal processing, particularly in image-text integration and visual semantic understanding. However, the results of this study and Wang et al [18] were both lower than the 90.7% accuracy in the United States Medical Licensing Examination test reported by Yang et al [25]. This discrepancy may be attributed to language differences in the test materials [26,27]. In the first 2 studies, the stems of graph-based questions were presented in Chinese, whereas Yang et al [25] used English. Previous studies have shown that language effects are particularly evident in tasks requiring specific cultural background knowledge or relying on the latest literature and data in other languages [28], which may further account for the differences in model performance. In contrast, both GenAI models performed better on the reverse logic questions, demonstrating their strengths in causal reasoning and decision-support tasks.

DeepSeek performed on par with ChatGPT in the audiologist qualification examination written in Chinese (79% vs 80%). As a GenAI model optimized for the Chinese language, DeepSeek-R1 significantly increased its coverage of Chinese in the training corpus and adopted alignment strategies specifically tailored for Chinese during fine-tuning, making it more targeted in Chinese language applications [29].

One would expect that DeepSeek might demonstrate superior performance in professional tasks in Chinese contexts. Our results do not support such a conjecture. It is possible that a vast majority of literature and textbooks on audiology are written in English. The amount of text information on audiology written in Chinese is only a fraction of that in English. Therefore, optimized processing ability in Chinese would not benefit the overall performance of the GenAI in the field of audiology.

This study conducted a preliminary exploration of the application of GenAI in a standardized audiology test written in Chinese. A few limitations were noted. First, all test questions were multiple-choice questions, lacking open-ended questions. In actual clinical practice, audiologists need to not only master discrete knowledge points but also develop integrative reasoning skills to solve complex problems. Therefore, future evaluations should include open-ended questions to assess the ability of GenAI to handle actual clinical scenarios. Second, due to limited access to standardized audiology test materials, this study only examined the TAQE, whose content and format may differ from those used in other countries or regions. In addition, the official TAQE website only reports the annual number of examinees who passed and the overall passing rate, without providing category-specific performance data. These factors may limit the generalizability of the findings in a broader international context and prevent us from further comparing the performance differences between GenAI and human test takers across specific domains. Future studies with access to more granular category–level results would enable a more comprehensive evaluation of similarities and differences between GenAI and human candidates. Third, GenAI is a continuously evolving and updating tool. The versions used in this study may have been updated by the time of publication, so the results may not fully reflect the state-of-the-art performance of the GenAI models.

In conclusion, ChatGPT and DeepSeek demonstrated good professional knowledge and problem-solving ability in the standardized audiologist qualification examination. They also provide a complete reasoning process, enabling teachers and students to quickly grasp key concepts and thought processes, demonstrating their potential as supplementary tools for audiology education. Furthermore, AI's superior ability to process massive data and accumulate knowledge allows it to be updated in real time, thus keeping pace with the evolving new technologies and methods in audiology, highlighting its value as a supportive tool for audiology education. However, GenAI responses still contain a certain error rate, and thus, their use in teaching requires careful judgment and guidance from educators. In addition, the performance of GenAI in open-ended tasks has not yet been fully validated, and its ability to foster humanistic qualities such as empathy and communication remains limited. Future research should prioritize incorporating open-ended tasks based on real-world scenarios to systematically

XSL•FO
RenderX

evaluate the potential and limitations of generative AI in audiology education. Moreover, this study primarily focused on general-purpose models without fine-tuning or domain adaptation. It is reasonable to expect that domain-specific training could further enhance performance, which provides an important direction for future work.

Acknowledgments

This study was supported in part by the China Scholarship Council.

Data Availability

All data generated or analyzed during this study are included in this published paper and Multimedia Appendices 1-3.

Authors' Contributions

LX led the study design, supervised the research process, and was responsible for manuscript revision. BQ assisted in the study design, performed data analysis, and drafted the manuscript. YW and YZ were responsible for data collection.

Conflicts of Interest

None declared.

Multimedia Appendix 1
ChatGPT transcripts.
[PDF File (Adobe PDF File), 1105 KB - formative_v9i1e79534_app1.pdf]

Multimedia Appendix 2
DeepSeek transcripts.
[PDF File (Adobe PDF File), 1204 KB - formative_v9i1e79534_app2.pdf]

Multimedia Appendix 3
Keys to the 2024 Taiwan Audiologist Qualification Examination.
[PDF File (Adobe PDF File), 260 KB - formative_v9i1e79534_app3.pdf]

References

1. Haleem A, Javaid M, Khan I. Current status and applications of artificial intelligence (AI) in medical field: an overview. Curr Med Res Pract 2019 Nov;9(6):231-237 [FREE Full text] [doi: 10.1016/j.cmrp.2019.11.005]
2. Holzinger A, Keiblinger K, Holub P, Zatloukal K, Müller H. AI for life: trends in artificial intelligence for biotechnology. New Biotechnol 2023 May;74:16-24 [FREE Full text] [doi: 10.1016/j.nbt.2023.02.001]
3. Montejo-Ráez A, Jiménez-Zafra SM. Current approaches and applications in natural language processing. Appl Sci 2022 May 11;12(10):4859. [doi: 10.3390/app12104859]
4. Cascella M, Semeraro F, Montomoli J, Bellini V, Piazza O, Bignami E. The breakthrough of large language models release for medical applications: 1-year timeline and perspectives. J Med Syst 2024 Feb 17;48(1):22 [FREE Full text] [doi: 10.1007/s10916-024-02045-3] [Medline: 38366043]
5. Zhang K, Meng X, Yan X, Ji J, Liu J, Xu H, et al. Revolutionizing health care: the transformative impact of large language models in medicine. J Med Internet Res 2025 Jan 07;27:e59069 [FREE Full text] [doi: 10.2196/59069] [Medline: 39773666]
6. Peng L, Luo G, Zhou S, Chen J, Xu Z, Sun J, et al. An in-depth evaluation of federated learning on biomedical natural language processing for information extraction. NPJ Digit Med 2024 May 15;7(1):127 [FREE Full text] [doi: 10.1038/s41746-024-01126-4] [Medline: 38750290]
7. Mihalache A, Huang RS, Popovic MM, Muni RH. ChatGPT-4: an assessment of an upgraded artificial intelligence chatbot in the United States Medical Licensing Examination. Med Teach 2024 Mar;46(3):366-372. [doi: 10.1080/0142159X.2023.2249588] [Medline: 37839017]
8. Jin HK, Lee HE, Kim E. Performance of ChatGPT-3.5 and GPT-4 in national licensing examinations for medicine, pharmacy, dentistry, and nursing: a systematic review and meta-analysis. BMC Med Educ 2024 Sep 16;24(1):1013 [FREE Full text] [doi: 10.1186/s12909-024-05944-8] [Medline: 39285377]
9. Ramchandani R, Guo E, Mostowy M, Kreutz J, Sahlollbey N, Carr MM, et al. Comparison of ChatGPT-4, Copilot, Bard and Gemini Ultra on an otolaryngology question bank. Clin Otolaryngol 2025 Jul;50(4):704-711. [doi: 10.1111/coa.14302] [Medline: 40081900]

XSL•FO
RenderX

10. Wu Z, Gan W, Xue Z, Ni Z, Zheng X, Zhang Y. Performance of ChatGPT on nursing licensure examinations in the United States and China: cross-sectional study. JMIR Med Educ 2024 Oct 03;10:e52746 [FREE Full text] [doi: 10.2196/52746] [Medline: 39363539]
11. Levin G, Horesh N, Brezinov Y, Meyer R. Performance of ChatGPT in medical examinations: a systematic review and a meta-analysis. BJOG 2024 Feb;131(3):378-380 [FREE Full text] [doi: 10.1111/1471-0528.17641] [Medline: 37604703]
12. Alfertshofer M, Hoch CC, Funk PF, Hollmann K, Wollenberg B, Knoedler S, et al. Sailing the seven seas: a multinational comparison of ChatGPT's performance on medical licensing examinations. Ann Biomed Eng 2024 Jun 08;52(6):1542-1545 [FREE Full text] [doi: 10.1007/s10439-023-03338-3] [Medline: 37553555]
13. Zong H, Li J, Wu E, Wu R, Lu J, Shen B. Performance of ChatGPT on Chinese national medical licensing examinations: a five-year examination evaluation study for physicians, pharmacists and nurses. BMC Med Educ 2024 Feb 14;24(1):143 [FREE Full text] [doi: 10.1186/s12909-024-05125-7] [Medline: 38355517]
14. Wang YM, Shen HW, Chen TJ. Performance of ChatGPT on the pharmacist licensing examination in Taiwan. J Chin Med Assoc 2023 Jul 01;86(7):653-658. [doi: 10.1097/JCMA.0000000000000942] [Medline: 37227901]
15. The World Factbook. URL: https://www.cia.gov/the-world-factbook/countries/world/#people-and-society [accessed 2025-11-08]
16. Apply for certification in audiology. American Speech-Language-Hearing Association. URL: https://www.asha.org/certification/audcertification/ [accessed 2025-11-08]
17. Exam question inquiry platform. Ministry of Examination ROC (Taiwan). URL: https://wwwq.moex.gov.tw/exam/wFrmExamQandASearch.aspx [accessed 2024-12-10]
18. Wang S, Mo C, Chen Y, Dai X, Wang H, Shen X. Exploring the performance of ChatGPT-4 in the Taiwan audiologist qualification examination: preliminary observational study highlighting the potential of AI chatbots in hearing care. JMIR Med Educ 2024 Apr 26;10:e55595 [FREE Full text] [doi: 10.2196/55595] [Medline: 38693697]
19. deepseek-ai / DeepSeek-R1. GitHub. URL: https://github.com/deepseek-ai/DeepSeek-R1 [accessed 2025-06-12]
20. Introducing ChatGPT. Open AI. URL: https://openai.com/blog/chatgpt [accessed 2024-10-01]
21. Regulations for examinations of specialized professionals and technical personnel. Ministry of Examination ROC (Taiwan). URL: https://wwwc.moex.gov.tw/main/ExamLaws/wfrmExamLaws.aspx?kind=3&menu_id=320&laws_id=133 [accessed 2025-11-08]
22. Tong W, Guan Y, Chen J, Huang X, Zhong Y, Zhang C, et al. Artificial intelligence in global health equity: an evaluation and discussion on the application of ChatGPT, in the Chinese National Medical Licensing Examination. Front Med (Lausanne) 2023;10:1237432 [FREE Full text] [doi: 10.3389/fmed.2023.1237432] [Medline: 38020160]
23. Wang H, Wu W, Dou Z, He L, Yang L. Performance and exploration of ChatGPT in medical examination, records and education in Chinese: pave the way for medical AI. Int J Med Inform 2023 Sep;177:105173. [doi: 10.1016/j.ijmedinf.2023.105173] [Medline: 37549499]
24. Fang C, Wu Y, Fu W, Ling J, Wang Y, Liu X, et al. How does ChatGPT-4 preform on non-English national medical licensing examination? An evaluation in Chinese language. PLOS Digit Health 2023 Dec 1;2(12):e0000397 [FREE Full text] [doi: 10.1371/journal.pdig.0000397] [Medline: 38039286]
25. Yang Z, Yao Z, Tasmin M, Vashisht P, Jang WS, Wang B, et al. Performance of multimodal GPT-4V on USMLE with image: potential for imaging diagnostic support with explanations. medRxiv. Preprint posted online on October 26, 2023 [FREE Full text] [doi: 10.1101/2023.10.26.23297629]
26. Watari T, Takagi S, Sakaguchi K, Nishizaki Y, Shimizu T, Yamamoto Y, et al. Performance comparison of ChatGPT-4 and Japanese medical residents in the general medicine in-training examination: comparison study. JMIR Med Educ 2023 Dec 06;9:e52202 [FREE Full text] [doi: 10.2196/52202] [Medline: 38055323]
27. Yu P, Fang C, Liu X, Fu W, Ling J, Yan Z, et al. Performance of ChatGPT on the Chinese postgraduate examination for clinical medicine: survey study. JMIR Med Educ 2024 Feb 09;10:e48514 [FREE Full text] [doi: 10.2196/48514] [Medline: 38335017]
28. Seghier ML. ChatGPT: not all languages are equal. Nature 2023 Mar;615(7951):216. [doi: 10.1038/d41586-023-00680-3] [Medline: 36882613]
29. DeepSeek. Wikipedia. URL: https://en.wikipedia.org/wiki/DeepSeek?utm_source=chatgpt.com [accessed 2025-06-18]

Abbreviations

AI: artificial intelligence
GenAI: generative artificial intelligence
TAQE: Taiwan Audiologist Qualification Examination

XSL•FO
RenderX

Edited by A Mavragani; submitted 23.06.25; peer-reviewed by J Hu, OC Balogun; comments to author 06.09.25; revised version received 26.09.25; accepted 20.10.25; published 28.11.25.

Please cite as:
Qi B, Zheng Y, Wang Y, Xu L
Comparison of ChatGPT and DeepSeek on a Standardized Audiologist Qualification Examination in Chinese: Observational Study
JMIR Form Res 2025;9:e79534
URL: https://formative.jmir.org/2025/1/e79534
doi:10.2196/79534
PMID:

©Beier Qi, Yan Zheng, Yuanyuan Wang, Li Xu. Originally published in JMIR Formative Research (https://formative.jmir.org), 28.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Rule-Based Conversational Agent for Mental Health and Well-Being in Young People: Formative Case Series During the Rise of Generative AI

Aimee-Rose Wrightson-Hester[1], PhD; Gee Anderson[1]; Joel Dunstan[2], MIT; Peter M McEvoy[1,3], PhD; Christopher J Sutton[4], PhD; Bronwyn Myers[1,5], PhD; Sarah J Egan[1], PhD; Sara Jane Tai[6], DClinPsy; Melanie Johnston-Hollitt[2], PhD; Wai Chen[1], PhD; Tom Gedeon[7], PhD; Joanna C Moullin[1], PhD; Warren Mansell[1], DClinPsy, DPhil

[1]Curtin enAble Institute, School of Population Health, Faculty of Health Sciences, Curtin University, Perth, Western Australia, Australia

[2]Curtin Institute for Data Science, Curtin University, Perth, Western Australia, Australia

[3]Centre for Clinical Interventions, North Metropolitan Health Service, Nedlands, Australia

[4]Lancashire Clinical Trials Unit,, University of Lancashire, Lancaster, United Kingdom

[5]Substance Use and Tobacco Research Unit, South African Medical Research Council, Tygerberg, Western Cape, South Africa

[6]School of Health Sciences; Division of Psychology and Mental Health, Faculty of Medical and Health Sciences, University of Manchester, Manchester, United Kingdom

[7]School of Electrical Engineering, Computing and Mathematical Sciences, Curtin University, Perth, Western Australia, Australia

Corresponding Author:
Warren Mansell, DClinPsy, DPhil
Curtin enAble Institute, School of Population Health
Faculty of Health Sciences
Curtin University
Kent Street
Perth, Western Australia, 6845
Australia
Phone: 1 82966 2411
Email: warren.mansell@curtin.edu.au

Abstract

Background: There is a shortage of services available to address the growing demand for mental health support in Australia and worldwide. Digital interventions, including conversational agents, can overcome barriers to accessing mental health support. The recent advances in large language models have led to an improvement in the perceived human-like naturalness of chatbot conversations, but there is little research on the experience of chatbots to support mental health. Manage your life online (MYLO) is a rule-based chatbot that was co-designed with young people and uses questions to help users explore their problems. In a case series conducted before the release of ChatGPT (OpenAI), users rated a new smartphone interface for MYLO as acceptable, and there was a large effect size for reduction in problem-related distress.

Objective: This study aimed to evaluate an improved version of MYLO and compare the user experience of MYLO in this case series to the previous version that was completed in November 2022.

Methods: We replicated and extended the previous 2-week case-series, conducted from September to November 2022, by testing 4-week usage of MYLO with a larger sample between October and December 2023. We recruited 24 young people living in Western Australia who self-described as having a lived experience of anxiety or depression. Participants had access to and used MYLO over a 4-week period while completing online weekly surveys that included a range of health and psychological questionnaires. After the 4-week testing phase, participants were invited to provide feedback on their experience of using MYLO through an interview or focus group discussion.

Results: In total, 13 of the 24 participants were retained throughout the study and took part in interviews. On average, participants had around 4 conversations with MYLO. They experienced both benefits and limitations of these conversations. They spoke about their recent experiences with ChatGPT (released in November 2022 after the previous case-series concluded) and other generative artificial intelligence (AI) tools, stating that they had expected MYLO to possess similar functionality, which it did

XSL•FO
RenderX

not. Nonetheless, we found moderate to large effect sizes for improvements in problem-related distress (Cohen *d*=–1.07), anxiety (Cohen *d*=–0.41), and psychiatric impairment (Cohen *d*=–0.60) and some evidence of reliable improvement in clinical outcomes.

Conclusions: These findings have implications for mental health chatbots in the age of ChatGPT and highlight a need for researchers to engage with new technologies to improve user experience, while maintaining the necessary safety and ethical standards that can be a significant challenge for generative AI.

(JMIR Form Res 2025;9:e69841) doi:10.2196/69841

KEYWORDS

mental health; conversational agents; chatbots; acceptability; feasibility; co-design; artificial therapist; youth; child; acceptability; Manage Your Life Online

Introduction

To address the growing need for mental health care [1], researchers and policymakers are investigating the use of digital and artificial intelligence (AI) solutions to improve the scalability of interventions [2]. Digital interventions might be particularly helpful for young people who have grown up with the internet and who experience barriers to accessing appropriate mental health care [3]. Conversational agents or chatbots are one technology that people are using to support their mental health. Mental health conversational agents are specifically designed to support users' mental health by emulating various forms of traditional psychotherapy [4]. Studies have found high user satisfaction with chatbots [5,6], with users enjoying the interactive approach and building a relationship with the chatbots like that of a human therapist or friend [7,8].

Although feedback is generally positive, many studies report negative feedback from users about mental health chatbots, such as repetitive content, not feeling understood [9], inappropriate responses from the bot [7], and technical issues [10]. Before 2022, the majority of mental health chatbots were rule-based systems in that humans explicitly programmed the way that the chatbot recognized the content of the user's text and the rules that decided how the chatbot replied using human-authored statements. Yet, in late 2022, highly naturalistic conversations with chatbots became widely accessible with the release of generative AI agents such as ChatGPT [11]. These systems use large language models to extract layers of regularities within vast databases of text, which leads to human-like responses to users that are not preprogramed or open to the direct oversight of human operators; we return to the issues this brings in the Discussion. Since ChatGPT's release, the use of AI technology has seen substantial growth driven by improvements in natural language processing and the integration of AI technologies into day-to-day personal and business applications [12]. Our paper presents formative data on how the rise in generative AI chatbots may have impacted an existing rule-based chatbot, with recommendations for developers in this fast-paced arena.

There is increasing awareness that digital innovations, such as social media and generative AI chatbots, have the potential for psychological harm [13]. When analyzing risk cases, the harm tends to relate to the "sycophantic" provision of information or advice that confirms a user's pre-existing beliefs [14]. Whilst providing information and advice is often beneficial in the digital domain, this is not universally the case. Indeed, misinformation has been identified as one of the greatest threats of AI [15]. Crucially, the provision of advice and information is not the main ingredient of effective psychological therapy and counselling. Core ingredients of effective therapy include nonjudgmental communication, active listening, and curious inquiry [16]. Yet, the conversational style of most existing chatbots, whether or not they target mental health or use generative AI systems, tends to focus on the provision of advice, interpretation, and information instead of these active therapeutic ingredients. This leaves a gap for innovation for a chatbot that does not provide advice, potential misinformation, or echo-chamber amplification. As such, MYLO is unique because it specializes in asking curious questions to help the user explore a problem in depth and detail, and notice current thoughts and feelings about the problem [17].

Manage your life online (MYLO) is a conversational agent that uses rule-based AI to exchange messages with the user. MYLO aims to emulate a conversation like that of a method of levels therapist [18]. Method of levels therapists help clients explore internal conflicts between their current state and their ideal state [19]. It is these conflicts that lead to psychological distress according to perceptual control theory, which underpins the method of levels therapy [6,18-20]. MYLO responds to user input in real time with curious questions tailored to the user's current problem. These questions aim to help the user sustain their awareness of the conflicts they are experiencing and explore the root cause of their distress [21].

MYLO was initially developed for use on a desktop, and evaluated when used for short sessions in controlled trials within an experimental context [21]. The current MYLO interface (refer to Figure 1) was co-designed with young people as a progressive web application on a smartphone to be accessible and appealing to use [22]. The acceptability of the current interface was evaluated by young people in a case-series in which users tested MYLO for 2 weeks and answered weekly online surveys [22]. Participants rated MYLO as acceptable on the system usability scale (mean 73.57, SD 16.02) and reported that the interface was easy to use and that they liked its functionality in interviews and focus groups. In total, 6 of the 8 participants who attended an interview or focus group said they would recommend MYLO to their friends. Further, reliable change scores [23] were calculated for each participant who completed baseline and posttesting (n=11), and 7 participants showed reliable improvements in their problem-related distress or other clinical outcome (such as depression, anxiety, and psychiatric impairment) scores. Due to the modest sample size

XSL•FO
RenderX

and short testing phase (2 weeks), no formal statistical testing was performed.

Following the previous case series, we made essential changes that we deemed necessary to enhance MYLO's conversational ability and establish safety before moving toward a full-scale trial. We used this case series to assess the acceptability and feasibility of these changes. We updated the MYLO database based on user feedback, alongside guidance from a new youth advisory panel. Specifically, new questions, themes, and terms were added to assist users with a wider range of problems. Importantly, 2 themes to identify risk of suicide or nonsuicidal self-harm were added, which, when triggered, responded to users by reminding them of the local mental health crisis hotline accessible through the MYLO interface. Relatedly, we added a measure of events to assess safety ahead of a full trial. During this development phase, the ability to decrypt user conversations was also added. Analyzing user conversations will allow researchers to identify the helpful and unhelpful elements of MYLO's conversations, building on the work of Gaffney et al [21]. While members of the youth advisory panel had shared sections of their conversations to discuss with the research team, it was unclear whether research participants would consent to share their conversations during a research trial.

In this study, we, therefore, aimed to extend the previous case-series by testing MYLO for a longer period of time and to test the changes made to the methodology described above before conducting a full-scale randomized controlled trial.

We also assessed the feasibility of conversation decryption and our methodology for gaining consent for this procedure. We did not have a target number of conversations or participants. We report the method used and then the number of participants and conversations obtained. We explored the user experience of MYLO using the session impact scale, the system usability scale, and qualitative interviews and focus groups. We also evaluated the effect size with regard to changes in problem-related distress, anxiety, depression, psychiatric impairment, and self-reported ability to engage in goal conflict reorganization after 4 weeks of using MYLO. We compared all of our findings to the previous case series, and through the lens of the widespread rise in use of generative AI chatbots that occurred in the year between the 2 case series. This part of the study is naturalistic, given the timing of the study, rather than making a direct comparison between generative AI and rule-based chatbots.

Figure 1. Screenshots of the MYLO Interface showing, from left to right, the home screen, the start to a conversation, and the drop-down menu. MYLO: Manage your life online.

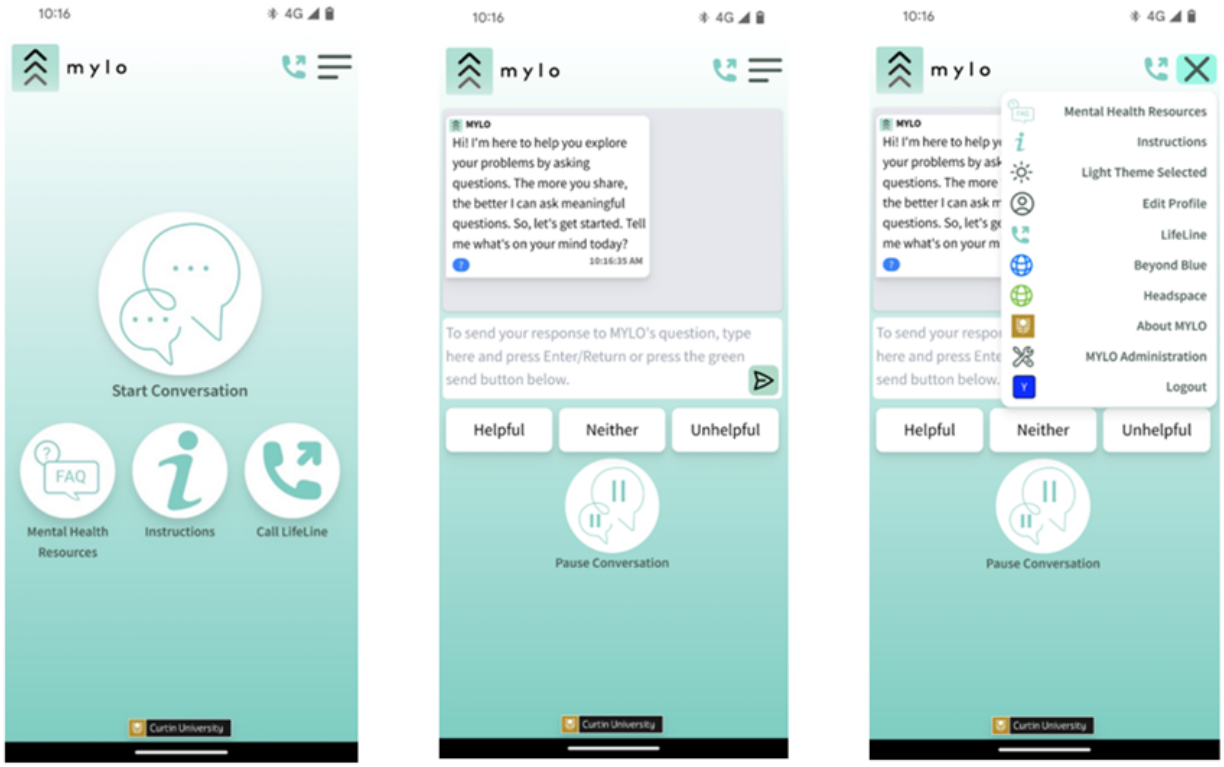

Methods

Design

The case-series protocol was similar to Wrightson-Hester et al [22], which was conducted from September to November 2022 during which participants engaged in the intervention over a period of 2 weeks. In the current study, the intervention duration was extended from 2 to 4 weeks, between October and December 2023. Therefore, a brief method will be described here, but for full details, see Wrightson-Hester et al [22]. One major difference was the procedure used to gain consent for the decryption and analysis of conversations, which will be

XSL•FO
RenderX

described in full in the procedure. To measure any adverse impacts of using MYLO, we also added an adverse events survey post testing.

Ethical Considerations

Approval for the research was obtained from the Curtin University Human Research Ethics Committee (HREC2022-0466). All participants were presented with an information sheet describing the aims and procedures of the study as well as information regarding how their data would be used, and they indicated their consent for each aspect of the study. A subject-generated code was used to link data across phases of the study anonymously. Our recruitment of 16- to 17-year-olds followed advice from ethical guidelines and our consumer panel, and involved either voluntary parental consent or a competency assessment to establish that the participant could be considered a mature minor. Participants were reimbursed for their time completing the online surveys, as well as for the focus group or interview with a digital gift card at a rate of Aus $20 (US $13, A currency exchange rate of Aus $1=US $0.65 is applicable) per hour with a maximum commitment of 5 hours if all surveys and an interview or focus group was completed (total of Aus $100 [US $65]). This report has been prepared so that the identification of individual participants is not possible.

Participants

The study was advertised on social media via Metaverse, X (X Corp, formerly known as Twitter), and LinkedIn (Microsoft) and shared by members of the research group through their existing networks and personal social media pages. The inclusion criteria were (1) being aged 16 to 24 years, (2) currently living in Western Australia, (3) self-reported an experience of anxiety or depression (either current or past), (4) having access to a smartphone and the internet, and (5) being able to read and type in English. The exclusion criteria were (1) scoring greater than 20 on the PHQ-9 (Patient Health Questionnaire-9), the threshold for severe depressive symptoms, or (2) reporting the experience of frequent suicidal thoughts (by scoring 2 or more on the PHQ-9 item 9) [24]. To partly address issues of inclusivity and generalizability, we ensured a minimum of 2 participants of each of the following: male, female, nonbinary gender, and non-Australian living in regional or remote Western Australia. Participants expressed their interest in participating through an online survey via Qualtrics (Qualtrics International Inc) and completed questions to assess their eligibility and demographic information, and gave informed consent to participate along with their contact details.

There was no formal sample size calculation as the purpose of the trial was not to test for effectiveness of MYLO. The initial sample size of 24 was deemed sufficient to assess usability and acceptability of MYLO, and, in particular, we aimed for a larger sample size than the previous case series, to ensure that, after attrition, we had a similar number as the last case series in order to make comparisons.

Materials

Web-Based Survey

Participants completed anonymous online surveys hosted by Qualtrics at baseline, then after each week of testing for a total of 4 weeks, between October and December 2023. A subject-generated identification code [25] was used to link participants' surveys across time points. The self-report questionnaires included in the online survey are presented in Table 1 (this table is adapted from Wrightson-Hester et al [22]).

XSL•FO
RenderX

Table 1. The names and properties of the self-report questionnaires.

Questionnaire	Measures	Scoring
Patient Health Questionnaire-9 [24]	9 items; depression	• 0-4: minimal depression, • 5-9: mild depression, • 10-14: moderate depression, • 15-19: moderately severe depression, and • 20-27: severe depression.
Generalized anxiety disorder assessment-7 [26]	7 items; anxiety	• 0-4: minimal anxiety, • 5-9: mild anxiety, • 10-14: moderate anxiety, and • 15-21: severe anxiety.
General health questionnaire-12 [27]	12 items; psychiatric impairment	• Traditional (acute) scoring method used. Scores range from 0 to 12, and higher scores indicate a greater possibility of psychological distress.
Short Form-6D version 2 [28]	6 items; general health	• Scores range from −0.685 to 1, with 1 indicating perfect health. Australian weights were used for this sample.
Psychological outcome profiles [29]	4 items used for scoring; change in problem-related distress over the course of therapy	• Scores range from 0 to 20. Decreases in scores between pretherapy and posttherapy indicate that a positive change has occurred.
Reorganization of conflict scale [30]	10-item subscale; goal conflict awareness and the proposed mechanism of change in the method of levels therapy	• Each item is scored from 0 (I do not believe this at all) to 100 (I believe this completely). The mean of the 10 items is used as the outcome.
General self-efficacy scale [31]	10 items; self-efficacy	• Scores range from 10 to 40. Higher scores indicate higher perceived general self-efficacy.
Session impact scale[a] [32]	17 items; session (therapeutic) satisfaction	• Each item is scored from 1 (not at all) to 5 (very much). We calculated the mean scores for the unwanted thoughts, relationship impacts, hindering impacts, understanding, and problem-solving subscales. Item 17 measures "other impacts," an optional item that is not used in scoring.
System usability scale[a] [33]	10 items; user experience of digital systems	• Outcome is a percentile ranking from 0 to 100, with scores >68 considered above average.
Adverse event survey[b]	26 items; adverse events	• The items have been adapted from the FOCUS clinical trial [34]. Each item is scored from 0 (Not at all) to 4 (Very much). Participants can also select "Not applicable."

[a]Denotes questionnaires that were only used after the baseline survey.

[b]This survey was included in the posttesting (week 4) survey and also sent to noncompleters separately.

MYLO

Participants accessed MYLO as a progressive web application through a link provided by the research team. MYLO's interface allows users to engage in a text-to-text conversation, access other mental health resources (including a local suicide crisis center hotline), pause and resume conversations, and alter the colors used in MYLO and their own avatars (colored squares with the user's initial). MYLO asks users curious questions in response to free-text provided by the user, for example, if a user states they feel nervous, MYLO might ask, "What do you do when you feel nervous?" The question that MYLO presents is determined by an algorithm that picks the best question based on key terms identified in the user's text and previous ratings of the question and term pairings. The questions aim to sustain the user's awareness of their problems, explore them in new ways, and ultimately, resolve them through shifts in their thinking or actions to resolve them.

Procedure

Participants accessed the Expression of Interest survey via links or QR codes on study adverts. Participants completed questions to access their eligibility (refer to Figure 2), recorded their demographic information (age, gender, cultural background, sexuality, and whether or not they lived in a metropolitan versus regional, rural, or remote location), completed the PHQ-9, and provided their contact details. Eligible participants from the Expression of Interest survey were then contacted by the research team and invited to complete the baseline survey via email or text. Those who completed the baseline survey were then provided with a MYLO account (via Auth0 [Okta]) and sent the link and instructions to download MYLO. Participants then had access to and used MYLO for 4 weeks and completed weekly online surveys between October and December 2023. After the testing phase, participants were invited to attend an online focus group or an individual interview to provide further feedback on their time using MYLO. All focus groups and

XSL•FO
RenderX

interviews were conducted in December 2023. The same topic guide was used to guide the interviews as was used in the previous case-series (Multimedia Appendix 1; see [22]).

At the end of the interview or focus group, participants were informed of the research team's intent to ask for their consent to decrypt and analyze the conversations they had with MYLO during the testing phase. After the interview or focus group, case-series participants were emailed an information sheet and consent form for the decryption of their conversations. Participants were able to ask any questions, and willing participants returned their consent forms via email. Emails with the information sheet and consent form were also sent to all other participants who did not attend an interview or focus group or complete all the surveys when they were sent their gift card for taking part in the study. In the same email, noncompleters were also provided a link to complete the optional adverse events survey.

Figure 2. Participant flowchart through the study from Expression of Interest to posttesting survey.

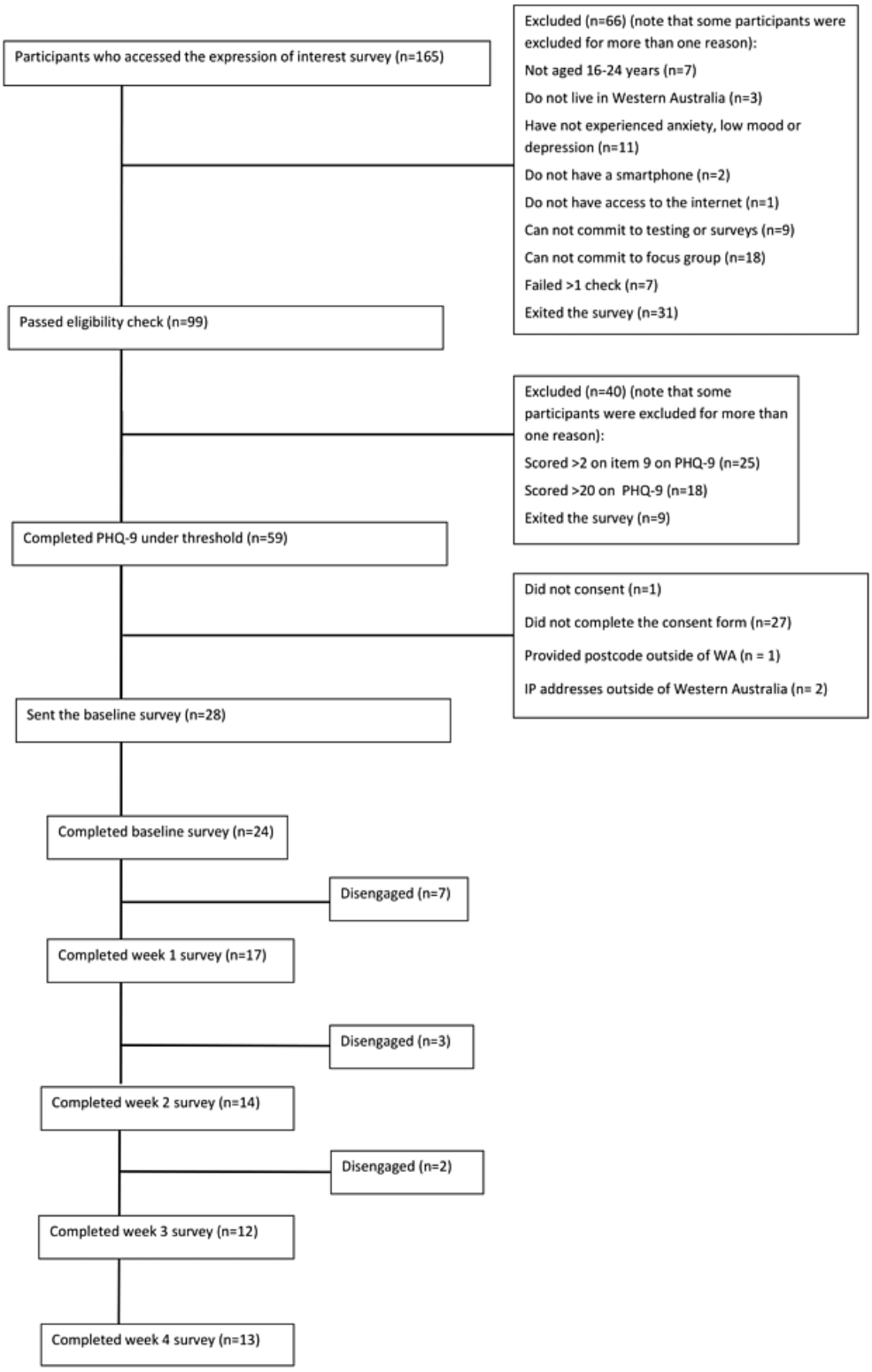

XSL•FO
RenderX

Data Analysis

Analysis was primarily descriptive, using mean (SD), median (IQR [range]), and frequency (%), as appropriate. In addition, (standardized) effect size estimates, with 95% CIs, were computed to assess MYLO's impact on user wellbeing and experience. We also calculated whether any participants achieved a reliable improvement or deterioration [23] during the testing phase on each of the quantitative outcome measures described previously.

To further assess users' experience of MYLO, interviews and focus groups were conducted online via Microsoft Teams, using the same topic guide as Wrightson-Hester et al [22]. Transcription and analysis were conducted by the first author according to Vears and Gillam's [35] inductive content analysis procedure. Coding was iterative, and emerging schemas were discussed and refined at regular research meetings between the first and last author to support confirmability. As the sample size was predefined by the case-series methodology, data saturation did not guide the number of focus groups or interviews conducted. However, the number of participants falls within the range commonly reported for data saturation.

Results

Participants

Figure 2 shows how many participants accessed the Expression of Interest survey and how many were lost at each stage of the survey. This process left 28 eligible participants who were contacted to participate in the study. Of these, 24 completed the baseline survey. The demographics of the participants are summarized in Table 2 and met our diversity targets. Of the 24 participants, 17 completed the week 1 survey, 14 completed week 2, 12 completed week 3, and 13 (54.2%) completed week 4. The 13th respondent in week 4 was unable to complete the week 3 survey. Focus groups and interviews were conducted with 13 participants. In total, 8 (7 completers, 1 noncompleter) participants completed and consented to the analysis of the adverse events survey.

Table 2. Participant demographics at baseline in the total sample and in completers.

Categories	Total sample (n=24), n (%)	Completers (n=13), n (%)
Age (years)		
16-17	12 (50)	9 (69)
18-19	5 (21)	1 (8)
20-21	2 (8)	0 (0)
22-24	5 (21)	3 (23)
Gender		
Woman	13 (54)	8 (62)
Man	7 (29)	3 (23)
Nonbinary or self-described[a]	4 (17)	2 (15)
Sexuality		
Heterosexual	9 (37)	3 (23)
Non-heterosexual	11 (46)	6 (46)
Did not disclose	4 (17)	4 (31)
Culture		
Australian	14 (58)	9 (69)
Australian and other[b]	6 (25)	3 (23)
Other[c]	4 (17)	1 (8)
Metropolitan or regional area		
Metropolitan	19 (79)	10 (77)
Regional or Remote	5 (21)	3 (23)

[a]Three participants described themselves as nonbinary and one as a transgender male.

[b]Australian and other (self-described): Chinese, Dutch and Sri Lankan, English, Italian, Hispanic, and Maori-New Zealander.

[c]Other (self-described): Afrikaans South African, Burmese, and Iranian.

Usage Data

Table 3 presents the MYLO usage data. Participants' usage data were collected in the MYLO app while their survey responses were collected anonymously via an external host (Qualtrics). Therefore, it was not possible to link participant outcomes to their MYLO usage in this phase, and the usage data reported is for all conversations had by all participants who downloaded

XSL•FO
RenderX

MYLO and had at least one conversation during the trial period (n=22).

Table 3. MYLO (Manage Your Life Online) Usage data divided by number of messages, conversations, and users.

	Values
Number of users, n	22
Total conversations, n	90
Conversations per user, mean	4.1
Number of messages from each user per conversation	
Mean (SD)	11.54 (11.10)
Median (range)	8.5 (1-61)
<5 messages	25 (28)
5-10 messages	26 (29)
11-20 messages	29 (32)
≥21 messages	10 (11)
Conversation rating, n (%)	
Helpful	14 (16)
Neither	16 (18)
Unhelpful	37 (41)
Unrated	23 (26)

Conversation Decryption and Consent Acceptability

All 24 participants who completed the baseline survey were sent an information sheet and consent form via email to allow the research team to decrypt and analyze the conversations they had with MYLO during the project. Of the 24 participants, four consented to their conversations being analyzed, one responded saying they did not want their conversations decrypted and analyzed, and 18 participants did not respond to the email. One participant let the research team know they would complete their consent form via email, but then did not contact the team again. From these participants, we were able to collect 14 conversations, which will be analyzed and presented in a future paper.

Adverse Events

In total, 8 participants completed the adverse events survey. All items were scored 0 (Not at all), 1 (Very little), 2 (A little), 3 (Quite a lot), or 4 (Very much). “Not Applicable” responses were not scored or included in calculations. A total of 25 items in the survey describe adverse events, and an average score of each was calculated for each participant was calculated. These scores ranged from 0.04 to 2.48 (mean 0.81, SD 0.78). Only 2 participants achieved a mean of over 1 on the scale. A total of 9 items had mean scores between 1 and 2. Three of these items related to the time and effort required to participate. One item scored 2.38, and that was “Taking part hasn't helped me with my problems.” Item 26 is a positive item and asks participants’ agreement with the statement “My problems have improved to the point whereby I no longer feel I need help,” and this had a mean score of 2. Participants were also able to provide qualitative feedback on the adverse events survey, and 4 participants chose to do so. Three participants left positive comments about MYLO, one of whom chose to explain their responses on the adverse event survey:

> *I had a very easy time figuring out how MYLO worked, but midway had a bout of depression that was unrelated to using MYLO, which made it very difficult for me to find the motivation to use the app.*
>
> *Doing the MYLO trial helped me have a better understanding of mental health and how it can affect you and the people around you. It definitely made me have a larger understanding of how much impact I can have on all the daily activities you do and your thoughts and mindset. I think that MYLO was a very well-created system, and I hope that our community can also enjoy and learn from it.*
>
> *I think the MYLO experience was overall good. I can't wait to see the following phases of MYLO as I can see this AI has the basic structure to become something that is going to be very helpful towards a large population of people.*

One participant left a negative response:

> *During times of mild anxiety/dissociation/depression, etc., I tried to use MYLO to help support me instead of my usual support system. MYLO was very unhelpful and made the problems worse due to its inability to provide any advice, practical help, or even have a proper conversation about the topic of mental health issues. Because of this I found that I needed to use other support systems to not only help with the original problem but counterbalance the lack of support that I felt from MYLO.*

XSL•FO
RenderX

Wellbeing Outcomes

Table 4 presents the mean scores for each self-report questionnaire at each time point as well as the change from baseline to week 4. Table 5 presents how many participants achieved reliable improvement or deterioration from baseline to posttesting for all outcomes. We found a large effect size for problem-related distress and a mean change of 3.77 from baseline to week 4 (Cohen *d*=–1.07). Anxiety scores showed a mean reduction of 2.08 (Cohen *d*=–0.41), indicating a small-to-moderate reduction from baseline to week 4. A medium effect size was found for psychiatric impairment, with the mean score posttesting dropping below the case threshold (ie, clinical threshold for psychiatric impairment) of 3 [30] to 2.46 (Cohen *d*=0.60). Of the 13 participants who completed both the baseline and posttesting (week 4), 11 individuals scored 3 or more at baseline, and 6 scored 3 or more posttesting (refer to Multimedia Appendix 2). The reliable change scores show that, on the measures of problem-related distress, 6 participants achieved a reliable improvement, and none had a reliable deterioration. For anxiety, 2 participants achieved a reliable improvement, and none had a reliable deterioration. For psychiatric impairment, 6 achieved a reliable improvement, and one had a reliable deterioration.

Table 4. Mean scores on clinical outcomes at baseline, week 1, week 2, week 3, and week 4.

	Baseline (n=24), mean (SD)	Week 1 (n=17), mean (SD)	Week 2 (n=14), mean (SD)	Week 3 (n=12), mean (SD)	Week 4 (n=13), mean (SD)	Change[a], mean (SD)	95% CI	*d*[b]
General health	0.56 (0.25)	0.69 (0.19)	0.71 (0.17)	0.65 (0.36)	0.71 (0.22)	0.04 (0.09)	–0.02 to 0.09	0.44
Depression	12.54 (5.45)	10.35 (4.18)	8.71 (4.34)	9.00 (5.31)	9.23 (5.89)	–1.85 (3.96)	–4.24 to 0.54	–0.47
Anxiety	9.58 (4.68)	8.82 (4.67)	8.00 (3.72)	8.58 (4.32)	7.08 (4.03)	–2.08 (5.02)	–5.11 to 0.96	–0.41
Psychiatric impairment	5.63 (2.58)	3.06 (2.73)	2.57 (2.85)	2.92 (2.81)	2.46 (2.96)	–2.31 (3.86)	–4.64 to 0.03	–0.60
Goal conflict reorganization	66.81 (13.12)	63.53 (10.40)	62.21 (13.38)	64.08 (15.20)	64.92 (15.06)	–1.00 (16.50)	–10.97 to 8.97	–0.06
Self-efficacy	28.71 (4.75)	28.53 (4.16)	28.93 (5.37)	28.17 (4.39)	29.46 (4.39)	0.31 (4.13)	–2.19 to 2.80	0.07
Problem-related distress	13.96 (2.71)	10.53 (2.21)	10.14 (2.96)	10.75 (2.99)	9.00 (3.46)	–3.77 (3.52)	–5.89 to –1.65	–1.07

[a]The change column presents the mean change between baseline and week 4; only scores of those who completed the week 4 survey are included (n=13).
[b]Cohen *d* was calculated using Jamovi for each outcome posttesting (end of week 4) relative to baseline [36].

Table 5. Participants' change scores from baseline to posttesting for all target outcomes.

	Cronbach α	Reliable Change Index [33]	Participants' scores												
			1	2	3	4	5	6	7	8	9	10	11	12	13
General health	0.82	0.26	–0.010	0.111	0.053	0.038	–0.126	0.000	0.029	–0.037	–0.005	0.165	0.251	0.014	0.054
Depression	0.86	5.78	–1	–4	–7[a]	–1	3	–5	2	6[b]	–1	–1	–8[a]	–3	–4
Anxiety	0.85	5.72	1	–2	–4	–2	1	–4	4	3	–1	–8[a]	–15[a]	1	–1
Psychiatric impairment	0.78	3.44	–2	–1	–4[a]	3	4[b]	–7[a]	3	0	–5[a]	–5[a]	–7[a]	–3	–6[a]
Goal conflict reorganization	0.81	13.90	–14[b]	17[a]	30[a]	11	–33[b]	–5	–5	5	–12	2	6	5	–20[b]
Self-efficacy	0.86	4.57	–1	1	7[a]	–1	–8[b]	–2	–2	5[a]	1	6[a]	2	0	–4
Problem-related distress	0.72	3.62	–4[a]	0	–12[a]	1	–7[a]	–1	–2	–2	–3	–2	–6[a]	–7[a]	–4[a]

[a]Denotes participant who achieved reliable improvement.
[b]Denotes participant who achieved reliable deterioration.

We found no effect on goal conflict reorganization with a mean change of 1.00 from baseline to week 4 (Cohen *d*=–0.06). The reliable change scores show that 2 participants achieved a reliable improvement in their ability to reorganize goal conflicts, and 3 reliably deteriorated.

Participants showed a small-to-moderate effect size improvement in general health between baseline and posttesting (Cohen *d*=0.44, mean change=0.04). No participants experienced a reliable change in general health. We found a small-to-moderate effect size on participants' depression scores between baseline and posttesting (Cohen *d*=0.47, mean change=1.85). We found no effect on self-efficacy (Cohen *d*=0.07, mean change=–0.31). The reliable change scores show that no participants experienced a reliable change in their general

XSL•FO
RenderX

health, 2 participants achieved reliable improvements in their depression scores, one participant reliably deteriorated, 3 participants achieved reliable improvements in their self-efficacy scores, and one participant reliably deteriorated between baseline and posttesting.

User Experience

In the current and previous case-series [22], participants rated MYLO's usability using the System Usability Scale [33] every week during the testing phase (over 4 weeks and 2 weeks, respectively). In the previous case series, MYLO's System Usability Scale ratings ranged from 37.5-97.5, with a mean rating of 73.57 (SD 16.02). In the current case series, the ratings ranged from 40 – 97.5, with a mean rating of 70.67 (SD 14.59). For both case-series, the mean ratings indicate MYLO is "acceptable" to most users, and scores were above average [37]. This suggests users' perceptions of MYLO's usability did not change between the 2 case series, and the system is performing similarly to the previous case series.

Accessibility of MYLO Interface

These results were echoed in the qualitative data, as many of the participants were generally positive about MYLO "I think it's rad. I think the idea is great" (16-year-old). All the participants described the MYLO interface as easy to use "overall for me, I found it really easy to use" (17-year-old). Many participants liked the accessibility of being able to talk about their problems anywhere "I could just click on the app when I wanted to use it, it would just start right away...it was very convenient" (17-year-old). Please note that quotes throughout have been edited lightly for clarity.

Some participants described times where they would prefer to use MYLO instead of talking to a person, "I found it useful in that respect…instead of feeling like I had to…sit down and talk to someone about it" (23-year-old). These participants were in 2 groups: those who would use MYLO when the people they would usually talk to were not available,

> *I do a lot of thinking in the nighttime…around those times when I'd be sorting problems out. And it's good to have MYLO...especially at those times.* [16-year-old]

And those who were reluctant to talk to people about their problems,

> *it was those points that I sort of reached for MYLO. I don't tend to be someone that usually reaches out and talks to someone about this stuff. I'm more private about it.* [17-year-old]

One participant felt they were able to share more authentically than if they had been talking to a person,

> *Having someone right there in front of you and looking you dead in the eyes, that's really confrontational. For me personally, I'll get nervous...now I don't feel comfortable talking about this. So it [MYLO] definitely felt comfortable. Just to be able to say it as it was.* [23-year-old]

Some participants felt the familiar chat format was helpful to support problem-processing,

> *I prefer that [texting], I wouldn't mind talking on the phone, but I think it's nicer to be able to [text]. When I'm typing, I can process it better in my mind and it helps me come to terms with what I'm upset about.* [17-year-old]

and made them feel comfortable,

> *I like the chat format setting because, I guess it almost feels like you're chatting with a friend in a way, because it's like a text message. So that provides almost a sense of comfort.* [17-year-old]

Satisfaction With Therapy Sessions

We compared participants' satisfaction with MYLO's therapeutic conversations in the current case series to the previous case series. Table 6 presents the mean therapy satisfaction scores measured by the Session Impact Scale [32] across all weeks (excluding baseline, where the scale was not included) for all participants. We found a large difference in (standardized) effect sizes between the previous and current case series on the hindering impact scale, with current case-series participants rating MYLO higher on the subscale (indicating a larger hindering impact) than the previous case-series' participants. We also found a moderate difference in (standardized) effect sizes between the previous and current case series on unwanted thoughts. The higher rating in the current case-series indicates that participants experienced more unwanted thoughts than the previous case-series' participants.

XSL•FO
RenderX

Table 6. Session impact subscale scores for Manage Your Life Online (MYLO) in the current case-series and the previous case-series [22].

Session impact subscale[a]	Current case-series (n[b]=17)	Previous case-series (n=11)	Mean difference (95% CI)[c]	Cohen *d*
Understanding, mean (SD)	2.11 (0.61)	2.36 (0.99)	0.26 (–0.45 to 0.97)	0.33
Problem-solving, mean (SD)	1.91 (0.82)	2.09 (1.04)	0.18 (–0.60 to 0.96)	0.20
Relationship, mean (SD)	2.02 (0.83)	2.22 (0.92)	0.20 (–0.52 to 0.91)	0.23
Hindering, mean (SD)	2.49 (0.87)	1.76 (0.52)	–0.74 (–1.28 to –0.20)	–0.98
Unwanted thoughts, mean (SD)	1.84 (0.70)	1.50 (0.50)	–0.34 (–0.81 to 0.12)	–0.54

[a]Session impact subscale score: 1=not at all, 2=slightly, 3=somewhat, 4=very much, and 5=very much.

[b]Number of responses across all weeks of each case-series.

[c]The mean and SD were calculated by first calculating the mean for each participant in each case-series for each subscale, and the table then displays the mean of these means. The mean difference is the mean from the previous case-series minus the mean from the current case series, and so represents the reduction in mean session impact score from the previous to the current case-series. The SD to perform the standardization for Cohen *d* is the pooled SD across the two case series.

Helpfulness of MYLO Conversations

Some participants found the conversations they had with MYLO useful, and that MYLO achieved its intended goal, “Sometimes it genuinely did make me reflect on my thoughts and stuff like that. Like with those questions” (16-year-old),

> *You know one of the defining characteristics for rumination, it's unproductive and persistent. So, the biggest thing for me about MYLO was that it felt productive because I think it helped me to think about things in a different way.* [23-year-old]

One participant felt at times that MYLO could do a better job of helping the participant than a person,

> *And I almost felt like it's therapeutic…I can just type it [a problem] out and then MYLO would just pick out all the important bits for me. And I don't have to focus. Whereas I feel like sometimes when you… [are] talking to people about it [a problem] …you sometimes [have to] give them like all the context, and they'll get confused in the small details.* [23-year-old]

Difficulties With MYLO Conversations

Despite some positive feedback, the Session Impact Scale scores suggest that participants experienced hindering effects when using MYLO. Participants described 4 main problems: MYLO’s questions were generic, MYLO was repetitive, MYLO had a limited purpose, and MYLO was not “as good” as some freely available generative AI chatbots.

Repetitive

Like the previous case series, the largest reported problem participants had with MYLO was repetition or looping. Although MYLO is programmed not to repeat the same question for at least 20 exchanges, participants felt questions were sometimes too like each other, “It just seems to be quite repetitive” (16-year-old), or that MYLO asked questions about a topic they had already discussed, “Like sometimes I felt like I was going in circles. Like I'd say something. Then it'll give me a question and I'll be like, I just told that” (17-year-old).

Generic

Participants also felt that sometimes MYLO asked generic questions, “The only thing that I sort of felt was on the downside is that sometimes, especially if you're talking about a problem that's quite specific, it would only give general answers” (17-year-old). One participant was conflicted as they sometimes found the similar or general questions useful, but at other times they caused frustration, “[MYLO] asking you to write more, sometimes was a bit frustrating when I didn't have anything to add. Sometimes it did help me to expand on what I was thinking and realize that maybe it's more complicated” (17-year-old).

When either of these issues (ie, repetitive or generic questioning) occur in a conversation they disrupt the flow of conversation and cause the participant to disengage, “It sort of jolted me backwards out of the conversation. I didn't expect [that response], that's not a response I would ever expect from a person that I was talking to” (19-year-old). “If you'd kind of forgotten for a bit, you were reminded that this is just a computer” (23-year-old).

Limited Purpose

Participants also felt that MYLO had a limited purpose and functionality at times. MYLO is programmed to only ask curious questions about a problem, and therefore does not engage in other forms of dialogue. Some participants wanted MYLO to engage in more casual conversation with them and suggested MYLO ask questions about their everyday life as well, “for people who have like a good day…or even just a bit lonely [but] you're still OK…I just want to have a bit of conversation” (17-year-old). This could improve MYLO's scores on the session impact scale, as some participants did not currently feel understood by MYLO: “It doesn't feel like it really knows you” (17-year-old).

Comparisons to Generative AI Chatbots

A novel critique that participants described (6 out of 13), which did not appear in the previous case-series [22], was how MYLO compared to freely available generative AI chatbots (such as ChatGPT). Participants had experience with generative AI chatbots, “ChatGPT started blowing up this year as well. The top 2 examples [for] me personally I would take #1 being Snapchat and second being ChatGPT” (16-year-old). Although

XSL•FO
RenderX

not all participants were as positive about generative AI or digital therapies,

> *Knowing that it's a computer sometimes I don't take it super seriously or don't think that it's gonna be as helpful and that in itself can be a bit of a barrier. I'm not maybe trying or engaging with it as well as I could just because I'm not expecting it to be helpful.* [17-year-old]

Many had expectations that MYLO would look, "I guess something like the interfaces of AI chatbots, like for example like ChatGPT. [Were] what I was kind of expecting…it kind of like subverted my expectations" [16 years old], and behave like other generative AIs (ie, be able to answer user questions and have considerable natural language processing skills). Some were disappointed when these expectations were not met,

> *I've used ChatGPT a bit…that [ChatGPT] kind of framed it for me that I was expecting [a similar experience] …but that's where I got to a bit of a dead end with it [MYLO]. I wanted to ask it things like give me…some time management strategies.* [23-year-old]

In the last case series, most participants would recommend MYLO to their friends, whereas the current participants appear to have a higher expectation for natural language programs, given the inception of ChatGPT and other generative AIs.

> *Definitely could be something I would recommend to other people, but maybe in this stage kind of what he (23-year-old) was saying before when you compare it to things like ChatGPT, which obviously have so much more money and like people behind it, it does kind of fall short.* [17-year-old]

Participants had some specific issues and recommendations informed by their knowledge and experience of generative AI chatbots. For example, ChatGPT has a context window that allows it to "remember" what a user types within each conversation [38]. MYLO, on the other hand, uses the immediate user response to generate its next question rather than all responses within a conversation, aside from rules included to avoid repetition of the same question within a certain number of messages. This is by design, so that MYLO is better able to emulate an method of levels therapist and respond to current statements made by the user. Participants, however, wanted MYLO to recall everything they said, in line with their perception of how ChatGPT operates,

> *I think that would be really, really useful, because when you look at ChatGPT like I've definitely used it and had asked it a question and then asked a question that was related to what it said and the previous question like following on and it's being able to respond and adjust. So I think that sort of capability in MYLO would definitely, really improve the conversation and would also probably really improve the chance of me recommending it.* [19-year-old]

Other participants enjoyed the free text responses they got from generative AI chatbots and the ability to have a general conversation with them, compared to MYLO's preauthored and structured responses.

> *I feel like MYLO doesn't really do that because of the fact that it doesn't have the free speech. As I mentioned before, like for example ChatGPT, …you can actually have a conversation with them, for example, the Snapchat AI.* [16-year-old]

Despite these comments, participants did recognize the dangers of generative AI, "it's a bit more risky, like you have to make sure where the information is drawing from" (17-year-old), and the therapeutic benefit of MYLO's style of conversation,

> *It's good that MYLO was just always asking me questions. In any sort of therapy setting, I guess it's better that the client would do more of the talking and the therapist would do more listening and asking questions.* [23-year-old]

Discussion

Principal Findings

We aimed to build on the acceptability and feasibility analysis within a previous shorter case-series [22] by extending the time that users had access to MYLO, and by increasing the types of data collected; specifically, we added an adverse events survey and the potential for participants to share the conversations they had with MYLO with the research team. The addition of the adverse events survey allowed us to further assess the safety of MYLO as an intervention, and the results suggest that no adverse events were attributed to MYLO, although one individual found its lack of advice made them feel worse and caused them to seek advice from their usual supports. Although MYLO's instructions state that MYLO only asks questions and will not give advice, some users might not read the instructions. To avoid other users experiencing this problem in the future, MYLO's purpose and functionality could be made clearer and unavoidable upon sign-in or when opening a conversation.

Only a small number of participants (4 out of 24) consented to the research team decrypting and analyzing their conversations in this case series. It is possible that participants are uncomfortable sharing their conversations with MYLO due to the sensitive nature of the conversations. It is also possible that the method of informing and collecting consent contributed to this low participation. In future studies, we aim to address both issues by implementing a method for participants to provide consent within the MYLO interface for each conversation they would be willing to have analyzed. This method will increase participants' autonomy over their data and provide an easy and accessible way for participants to provide consent.

To evaluate MYLO's ability to support the well-being of young people, we also assessed the effect size of the change over time in key outcome variables.

Participants' problem-related distress, anxiety, and psychiatric impairment showed substantial effect sizes for improvement after testing MYLO for 4 weeks. These findings were in the same direction as the previous case-series [22].

No change was found in participants' ability to engage in goal conflict reorganization. Inspection of the reliable change scores for all participants showed that 2 participants achieved reliable

XSL•FO
RenderX

improvement, while 3 reliably deteriorated. It is not clear why these results were different from the previous case-series [22], which showed a medium effect size increase in goal conflict reorganization. Goal conflict reorganization is the proposed mechanism of change for method of levels and MYLO, by increasing clients' and users' ability to resolve the conflicts they experience, their problem-related distress should improve, and subsequently their depression and anxiety symptoms should reduce [21]. Several method of levels studies have demonstrated this increase in ROC scores [39-41]. Further research is needed to establish the impact of using MYLO on goal conflict resolution with a fully powered sample. If MYLO provides the opportunity for goal conflict reorganization to be implemented via its interface for problem expression and curious questioning, then we would expect greater evidence of goal conflict awareness in the conversations to mediate its benefits, as found in an earlier study [21]. As it stands, the current study challenges the conceptual credibility of this mechanism, in contrast to earlier studies. Tentatively, it is possible that participants' expectations for near-human conversational ability based on generative AI—which was not available during earlier research on MYLO—reduced their commitment to engaging in depth with MYLO regarding their goal conflicts. We discuss this in more detail later.

We found an effect on depression scores, and the reliable change score of individuals suggests some participants experienced an improvement in their self-efficacy scores. However, more research in a fully powered randomized trial is needed to validate these findings. The null findings related to general health could be due to the items in the Short Form-6D version 2, which focus on 6 domains of health, including mental health. MYLO's focus on mental health and well-being limits its ability to improve the other domains in the Short Form-6D version 2, unless a user is seeking to explore a problem related to one of the other domains.

Regarding the user experience of MYLO, it scored similarly on the session impact scale to the previous case series, except for achieving a higher score on the hindering impacts subscale. As the system usability scale scores were similar, suggesting participants found MYLO similarly acceptable to the participants in the last case-series, we examined the qualitative data to find a possible explanation for this difference. Much of the positive qualitative feedback was similar to the previous case-series and other mental health chatbots [5,6]; participants praised the accessibility of the interface. Participants also described incidents where MYLO had helped them to achieve a new perspective on their problems. Participants also experienced similar difficulties regarding the repetition of questions and instances when MYLO did not understand them. These are difficulties that many chatbots, especially logic or rule-based chatbots, experience [9]. In our study, only 16% of conversations were rated as helpful, which is clearly nonoptimal. However, there appears to be no data from other conversational agents for comparison.

One difference that emerged was comparisons between MYLO and widely available generative AI models such as ChatGPT and Gemini. The widespread availability of these chatbots appears to have altered users' expectations of what interacting with a chatbot is like. Our previous participants had not used these models, as they were not yet available, and while they experienced similar issues and scored MYLO similarly on all other session impact scale subscales (ie, understanding, problem-solving, and relationship), they were much more likely to recommend MYLO in its current form. In contrast, the current participants were less likely to recommend MYLO, despite similar reductions in problem-related distress and other mental health symptoms. Specifically, our users pointed to the natural language processing and generative AI chatbot's ability to adapt and perform many tasks (from writing computer code to answering complex questions on many topics) as superior to MYLO's specific question-answer structure.

Many mental health chatbots lack the functionality of large generative AI models for many reasons (eg, funding), but often by design. For example, as one participant noted, it is often beneficial for the client to do most of the talking during a therapeutic conversation, whereas current generative AI models produce large amounts of text in response to minimal user input. Another drawback from generative AI recognized by our participants was the lack of transparency regarding where information or advice came from, whereas MYLO and other mental health chatbots generate text previously authored by mental health professionals, or generate text constrained by very specific rules created by mental health professionals [42]. Despite this, as technology advances, these interventions might no longer meet users' standards for how naturalistic a digital conversation should be. This could exacerbate the already difficult issue of retention in digital interventions and ultimately limit the effectiveness of MYLO and other mental health chatbots.

One potential explanation for the contrast we found is that rule-based chatbots and generative AI chatbots occupy opposite ends of the "uncanny valley" [43]. The uncanny valley was originally developed to describe a midground between simulations of humans that are clearly experienced as a computer, and those simulations that are almost indistinguishable from humans. When users know they are conversing with a computer, they can nonetheless use its benefits without the concern of sharing their thoughts and feelings with another human. This is thought to be the main reason for what has come to be known as the Eliza effect, identified early on in chatbot development [44], and is likely to have been operating in early versions of MYLO. Yet as computer simulations occupy the midground between a computer and a human, the appeal of the simulation is found to be substantially worse than either extreme and is associated with a feeling of eeriness and detachment from the agent. A recent systematic review indicates that this effect is present in chatbots, with users reverting to either side of the uncanny valley—more simplistic chatbots, or to humans—when their chatbot interaction is near-human [45]. It is possible that the improvements of MYLO, in the context of recent human-like text-based conversations with generative AI, led perceptions of MYLO into the uncanny valley. Indeed, it has been shown that a text-based chatbot programmed to create the uncanny valley effect is experienced as less likable and intelligent than a chatbot programmed to converse exactly like a human [46]. Our findings may therefore reflect the

XSL•FO
RenderX

challenge of the uncanny valley regarding the attempt to improve rule-based chatbots through simply integrating generative AI capable of natural conversation; codesign and user testing will be needed to mitigate the issues.

Future Directions

The concerns regarding generative AI have been expressed more widely by researchers [47-50]. These include the significant cost and environmental impact of maintaining open-access models, which may become prohibitive for widespread applications as costs shift to consumers [51]. In addition, maintaining conversational context is resource-intensive, a limitation noted across systems like ChatGPT and MYLO, though OpenAI's retention of previous discussions illustrates attempts to address this [52]. Generative AI struggles with handling ambiguity, often generating poor responses to unclear inputs, and lacks true reasoning capabilities, relying instead on probabilistic predictions based on training data [52]. Emotional intelligence is another limitation, as these systems, including MYLO, cannot currently detect or adapt to users' emotional states. Furthermore, ethical concerns arise from biases in training datasets, particularly in sensitive areas like mental health [49,50,53]. Unlike commercial Generative AI tools, MYLO's design potentially offers a more controlled and transparent approach to mitigating these risks, and future research could make a direct comparison of benefits and risks between MYLO and a purely generative AI mental health chatbot.

More critically, MYLO's focus exclusively on curious problem exploration and noticing thoughts and feelings has 2 additional advantages. It guards against the issues with advice, misinformation, and sycophancy in generative AI, and supports the intrinsic capacities for users to solve their own problems owing to MYLO's scientific foundations in perceptual control theory [20]. From the perspective of perceptual control theory, longstanding distress emerges from unresolved goal conflicts that require sustained exploration to resolve through a process known as reorganization. Conversely, it is proposed that the advice and information may support the pursuit of the conflicting goals, actually resulting in greater distress [54].

Currently, the logic-based system behind MYLO uses human-selected therapeutic questions about specific themes, and it prioritizes certain questions over others based on the ranking of topics. For example, terms indicating goal conflict, such as "dilemma" or "in two minds," are ranked higher than terms about activities, such as "going to the gym." In addition, questions that are consistently rated by past users as unhelpful are less likely to be selected. These elements will remain, yet they will be supplemented by large language models to address some of the recognized issues. First, we have convened a new, diverse consumer advisory group to make recommendations, give feedback on planned improvements, and carry out user testing. This will proceed in iterative steps to ensure that each phase of improvement enhances the user's experience of MYLO. We are scoping for a natural language processor that is secure and energy efficient to hybridize with MYLO. Its 2 main objectives will be to ensure that themes are more readily and accurately recognized, and to ensure that MYLO selects a response that is appropriate and makes sense to the user. We expect that this hybrid version will improve the user's sense of being understood, maintain their engagement, and ensure that their conversations are long enough to allow for detailed problem exploration, promoting the reorganization of goal conflict as proposed by perceptual control theory [20]. Whilst the exact specifications of the hybridization are unclear at this stage, we plan to retain the existing rule-based selection of prespecified theory-driven questions, to prevent it from confusing the user with a shift in therapeutic principles [17], and to limit the recognized risks of computer-generated text in the arena of mental health [55].

Limitations

We used a case-series design to evaluate MYLO and assess the acceptability of new data collection methods. The lack of a control group, small sample size, and high dropout rates, both in terms of maintaining engagement with MYLO and providing outcome data, suggest that our findings are indicative of a selective group that is not methodologically robust. The findings of this study will be used to inform the design of a randomized controlled trial that compares MYLO users to a waitlist group to provide a robust and valid assessment of MYLO's efficacy. Second, we also experienced high levels of dropout between baseline and the first week of testing when participants were required to download and log in to MYLO, which may have biased the final sample. This step had been added to this case series. It was not necessary in the previous MYLO projects [22] where participants were able to use MYLO without an account through a weblink. The current process requires the participant to inform the lead researcher that they have completed the baseline survey; the lead researcher then creates an account for the user and emails or texts them their login details, as well as instructions on how to download and use MYLO. This process was a necessity for this stage, given the authentication platform being used (Auth0) and MYLO's current stage of development. However, previous consumer panel members have indicated a preference for password-less access, stating that passwords and usernames are a barrier for some young people [22]. We therefore aim to address this issue in the next research and development phase by allowing users to create their own accounts and let them use their saved passwords, biometrics or social media log ins to authenticate their identity; we will also link usage data to individual participants whilst preserving their anonymity so that the potential dose-response relationship between usage and outcome can be analyzed to address this analytic limitation within the current study. Ultimately, the goal is to create a smartphone application version of MYLO that can be placed on all available application stores. Doing so will greatly increase MYLO's accessibility and availability to a broader sample of users.

The number of participants in the qualitative analysis was not guided by data saturation or thematic saturation. We are aware that this contradicts good qualitative practice and may have introduced a selection bias, and so would not be considered generally representative of potential users of MYLO. We also note that our framing of experiences with MYLO after the widespread use of generative-AI chatbots was opportunistic in the absence of similar studies, yet a comparative or crossover design would have been ideal.

XSL•FO
RenderX

Conclusions

This study aimed to extend our previous work on MYLO by providing the application for a longer period of time, 4 weeks, and examining its impact on MYLO users' well-being. We also assessed the feasibility of changes to the trial methodology, that is, adding the option for participants to consent to decryption and analysis of their conversations before conducting a full-scale randomized controlled trial. All changes will be retained moving forward, but we are hoping to integrate consent to decrypt conversations into the application interface to increase the number of participants opting in and to allow participants to opt in for individual conversations. Regarding the quantitative results, the direction of the pre-post effects on problem-related distress, anxiety, and psychiatric impairment was the same as previous research on MYLO. However, the lack of change in participants' goal conflict reorganization and the mixed findings on self-efficacy and general health highlight the need for further investigation with a fully powered randomized controlled trial to better understand the intervention's impact on participants' well-being and mental health, as well as the mechanisms driving these changes.

The study also revealed key insights into the user experience of mental health conversational agents, particularly in comparison to widely available generative AI models like ChatGPT and Gemini. While MYLO scored similarly on user experience and usability metrics, the changing expectations of users in response to advances in AI technologies suggest that integrating more sophisticated natural language processing features might enhance its usability, rapport-building, and therapeutic outcomes. Further, addressing technological barriers, such as streamlining user authentication and developing a smartphone application, will enhance MYLO's accessibility and engagement. Future iterations of rule-based chatbots, incorporating co-designed improvements and hybrid approaches leveraging large language models, hold the potential to better meet users' evolving expectations and improve retention in digital mental health interventions.

Acknowledgments

All authors contributed to the study conception and design. A-RW-H, GA, and JD contributed to data collection. A-RW-H and WM contributed to the analysis and interpretation of the results. A-RW-H prepared the first draft. All authors reviewed the results, provided feedback on drafts, and approved the final version of the manuscript. The authors would like to thank Jason Wright, the youth advisory committees, and the study participants for their contributions to this study. Their expertise, feedback, and time were invaluable in making Manage Your Life Online what it is today.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Topic Guide for the Focus Groups and Interviews.
[DOCX File , 31 KB - formative_v9i1e69841_app1.docx]

Multimedia Appendix 2
Spreadsheet presenting the baseline and 4 week (post-testing) scores on all clinical outcomes for completers.
[XLSX File (Microsoft Excel File), 18 KB - formative_v9i1e69841_app2.xlsx]

References

1. Racine N, McArthur BA, Cooke JE, Eirich R, Zhu J, Madigan S. Global prevalence of depressive and anxiety symptoms in children and adolescents during COVID-19: a meta-analysis. JAMA Pediatr 2021;175(11):1142-1150 [FREE Full text] [doi: 10.1001/jamapediatrics.2021.2482] [Medline: 34369987]
2. Minerva F, Giubilini A. Is AI the Future of mental healthcare? Topoi (Dordr) 2023;42(3):1-9 [FREE Full text] [doi: 10.1007/s11245-023-09932-3] [Medline: 37361723]
3. Lehtimaki S, Martic J, Wahl B, Foster KT, Schwalbe N. Evidence on digital mental health interventions for adolescents and young people: systematic overview. JMIR Ment Health 2021;8(4):e25847 [FREE Full text] [doi: 10.2196/25847] [Medline: 33913817]
4. Koh J, Tng GYQ, Hartanto A. Potential and pitfalls of mobile mental health apps in traditional treatment: an umbrella review. J Pers Med 2022;12(9):1376 [FREE Full text] [doi: 10.3390/jpm12091376] [Medline: 36143161]
5. Boucher EM, Harake NR, Ward HE, Stoeckl SE, Vargas J, Minkel J, et al. Artificially intelligent chatbots in digital mental health interventions: a review. Expert Rev Med Devices 2021;18(sup1):37-49 [FREE Full text] [doi: 10.1080/17434440.2021.2013200] [Medline: 34872429]
6. Gaffney H, Mansell W, Tai S. Agents of change: understanding the therapeutic processes associated with the helpfulness of therapy for mental health problems with relational agent MYLO. Digit Health 2020;6:2055207620911580 [FREE Full text] [doi: 10.1177/2055207620911580] [Medline: 32206331]

XSL•FO
RenderX

7. Gaffney H, Mansell W, Tai S. Conversational agents in the treatment of mental health problems: mixed-method systematic review. JMIR Ment Health 2019;6(10):e14166 [FREE Full text] [doi: 10.2196/14166] [Medline: 31628789]
8. Vaidyam AN, Wisniewski H, Halamka JD, Kashavan MS, Torous JB. Chatbots and conversational agents in mental health: a review of the psychiatric landscape. Can J Psychiatry 2019;64(7):456-464 [FREE Full text] [doi: 10.1177/0706743719828977] [Medline: 30897957]
9. Egan SJ, Johnson C, Wade TD, Carlbring P, Raghav S, Shafran R. A pilot study of the perceptions and acceptability of guidance using artificial intelligence in internet cognitive behaviour therapy for perfectionism in young people. Internet Interv 2024;35:100711 [FREE Full text] [doi: 10.1016/j.invent.2024.100711] [Medline: 38313140]
10. Haque MR, Rubya S. "For an App Supposed to Make Its Users Feel Better, It Sure is a Joke" - An analysis of user reviews of mobile mental health applications. Proc. ACM Hum.-Comput. Interact 2022;6(CSCW2):1-29. [doi: 10.1145/3555146]
11. Ma X, Huo Y. Are users willing to embrace ChatGPT? Exploring the factors on the acceptance of chatbots from the perspective of AIDUA framework. Technology in Society 2023;75:102362. [doi: 10.1016/j.techsoc.2023.102362]
12. Singla A, Sukharevsky A, Yee L, Chui M, Hall B. The state of AI in early 2024: Gen AI adoption spikes and starts to generate value. Quantum Black AI by McKinsey. 2024. URL: https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai#/ [accessed 2025-10-18]
13. Shalaby A. Classification for the digital and cognitive AI hazards: urgent call to establish automated safe standard for protecting young human minds. DESD 2024;2(1):17. [doi: 10.1007/s44265-024-00042-5]
14. Dohnány S, Kurth-Nelson Z, Spens E, Luettgau L, Reid A, Summerfield C, et al. Technological folie\a deux: feedback loops between ai chatbots and mental illness. arXiv:2507.19218 2025:1-19. [doi: 10.48550/arXiv.2507.19218]
15. Williamson SM, Prybutok V. The era of artificial intelligence deception: unraveling the complexities of false realities and emerging threats of misinformation. Information 2024;15(6):299. [doi: 10.3390/info15060299]
16. Wampold BE, Imel ZE. The Great Psychotherapy Debate: the Evidence for What Makes Psychotherapy Work. United Kingdom: Routledge; 2015.
17. Grodniewicz JP, Hohol M. Waiting for a digital therapist: three challenges on the path to psychotherapy delivered by artificial intelligence. Front Psychiatry 2023;14:1190084 [FREE Full text] [doi: 10.3389/fpsyt.2023.1190084] [Medline: 37324824]
18. Carey, TA. The Method of Levels: How to Do Psychotherapy Without Getting in the Way. Hayward, CA: Living Control Systems Publishing; 2006.
19. Mansell W, Goldstein D. Methods of levels therapy. In: The Interdisciplinary Handbook of Perceptual Control Theory. Academic Press: Academic Press; 2020.
20. Powers WT. Behavior: The Control of Perception. New York: Aldine; 1973.
21. Gaffney H, Mansell W, Edwards R, Wright J. Manage your life online (MYLO): a pilot trial of a conversational computer-based intervention for problem solving in a student sample. Behav Cogn Psychother 2014;42(6):731-746. [doi: 10.1017/S135246581300060X] [Medline: 23899405]
22. Wrightson-Hester A, Anderson G, Dunstan J, McEvoy PM, Sutton CJ, Myers B, et al. An artificial therapist (manage your life online) to support the mental health of youth: co-design and case series. JMIR Hum Factors 2023;10:e46849 [FREE Full text] [doi: 10.2196/46849] [Medline: 37477969]
23. Evans C, Margison F, Barkham M. The contribution of reliable and clinically significant change methods to evidence-based mental health. Evid Based Mental Health 1998;1(3):70-72. [doi: 10.1136/ebmh.1.3.70]
24. Kroenke K, Spitzer RL, Williams JBW. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
25. Audette LM, Hammond MS, Rochester NK. Methodological issues with coding participants in anonymous psychological longitudinal studies. Educ Psychol Meas 2020;80(1):163-185 [FREE Full text] [doi: 10.1177/0013164419843576] [Medline: 31933497]
26. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
27. Goldberg D, Williams P. A User's Guide to the General Health Questionnaire. Windsor, UK: NFER-Nelson; 1988.
28. Brazier, Mulhern BJ, Bjorner JB, Gandek B, Rowen D, Alonso J, SF-6Dv2 International Project Group. Developing a new version of the SF-6D health state classification system from the SF-36v2: SF-6Dv2. Med Care 2020;58(6):557-565. [doi: 10.1097/MLR.0000000000001325] [Medline: 32412942]
29. Ashworth M, Shepherd M, Christey J, Matthews V, Wright K, Parmentier H, et al. A client‐generated psychometric instrument: The development of 'PSYCHLOPS'. Couns and Psychother Res 2006;4(2):27-31. [doi: 10.1080/14733140412331383913]
30. Bird T. An Investigation of Transdiagnostic Processes and Interventions in Clinical and Non-clinical Settings (Doctoral Thesis). United Kingdom: University of Manchester; 2013.
31. Schwarzer R, Jerusalem, M. Generalized self-efficacy scale. In: Measures in health psychology: A user's portfolio. Causal and control beliefs. United Kingdom: NFER-NELSON; 1995:35-37.
32. Elliott R, Wexler MM. Measuring the impact of sessions in process—Experiential therapy of depression: the session impacts scale. Journal of Counseling Psychology 1994;41(2):166-174. [doi: 10.1037/0022-0167.41.2.166]

XSL•FO
RenderX

33. Brooke J. SUS: A 'quick and dirty' usability scale. In: Usability Evaluation in Industry. London: Taylor & Francis; 1996:189-194.
34. Morrison AP, Pyle M, Gumley A, Schwannauer M, Turkington D, MacLennan G, FOCUS trial group. Cognitive behavioural therapy in clozapine-resistant schizophrenia (FOCUS): an assessor-blinded, randomised controlled trial. Lancet Psychiatry 2018;5(8):633-643 [FREE Full text] [doi: 10.1016/S2215-0366(18)30184-6] [Medline: 30001930]
35. Vears DF, Gillam L. Inductive content analysis: A guide for beginning qualitative researchers. FoHPE 2022;23(1):111-127. [doi: 10.11157/fohpe.v23i1.544]
36. Navarro DJ, Foxcroft DR. Comparing Two Means: Effect Size. 2025. URL: https://lsj.readthedocs.io/ko/latest/Ch11/Ch11_tTest_07.html [accessed 2025-10-18]
37. Sauro J. 5 Ways to Interpret a SUS Score. Measuring U. 2018. URL: https://measuringu.com/interpret-sus-score/ [accessed 2025-10-18]
38. Pena D. Understanding how Chat-GPT maintains context. Sitepoint. URL: https://www.sitepoint.com/understanding-how-chatgpt-maintains-context/ [accessed 2025-10-18]
39. Gluckman N. An evaluation of online method of levels therapy with young people (Doctoral Thesis). University of East London. 2022. URL: https://repository.uel.ac.uk/item/8v668 [accessed 2025-10-18]
40. Churchman A, Mansell W, Tai S. A process-focused case series of a school-based intervention aimed at giving young people choice and control over their attendance and their goals in therapy. British Journal of Guidance & Counselling 2020;49(4):565-586. [doi: 10.1080/03069885.2020.1815650]
41. Griffiths R, Mansell W, Carey TA, Edge D, Emsley R, Tai SJ. Method of levels therapy for first-episode psychosis: The feasibility randomized controlled next level trial. J Clin Psychol 2019;75(10):1756-1769. [doi: 10.1002/jclp.22820] [Medline: 31240723]
42. Abd-Alrazaq AA, Alajlani M, Alalwan AA, Bewick BM, Gardner P, Househ M. An overview of the features of chatbots in mental health: A scoping review. Int J Med Inform 2019;132:103978 [FREE Full text] [doi: 10.1016/j.ijmedinf.2019.103978] [Medline: 31622850]
43. Mori M. Bukimi no tani [The uncanny valley]. Energy 1970;7:33. [doi: 10.5749/j.ctvtv937f.7]
44. Cristea AI, Sucala M, David D. Can you tell the difference' comparing face-to-face versus computer-based interventions. The Eliza effect in psychotherapy. Journal of Cognitive & Behavioral Psychotherapies 2013;13(2):291-298 [FREE Full text]
45. Greilich A, Bremser K, Wüst K. Consumer response to anthropomorphism of text-based AI chatbots: A systematic literature review and future research directions. Int J Consumer Studies 2025;49(5):e70108. [doi: 10.1111/ijcs.70108]
46. Kishnani D. The Uncanny Valley: An Empirical Study on Human Perceptions of AI-Generated Text and Images. Cambridge, Massachusetts: Massachusetts Institute of Technology; 2025. URL: https://dspace.mit.edu/handle/1721.1/159096 [accessed 2025-10-25]
47. Pandya A, Lodha P, Ganatra A. Is ChatGPT ready to change mental healthcare? Challenges and considerations: a reality-check. Front. Hum. Dyn 2024;5:1289255. [doi: 10.3389/fhumd.2023.1289255]
48. Alanezi F. Assessing the effectiveness of ChatGPT in delivering mental health support: A qualitative study. J Multidiscip Healthc 2024;17:461-471 [FREE Full text] [doi: 10.2147/JMDH.S447368] [Medline: 38314011]
49. Kalam KT, Rahman JM, Islam MR, Dewan SMR. ChatGPT and mental health: friends or foes? Health Sci Rep 2024;7(2):e1912 [FREE Full text] [doi: 10.1002/hsr2.1912] [Medline: 38361805]
50. Raile P. The usefulness of ChatGPT for psychotherapists and patients. Humanit Soc Sci Commun 2024;11:47. [doi: 10.1057/s41599-023-02567-0]
51. Crawford K. Generative AI's environmental costs are soaring - and mostly secret. Nature 2024;626(8000):693. [doi: 10.1038/d41586-024-00478-x] [Medline: 38378831]
52. Open AI. Memory and new controls for ChatGPT. 2024. URL: https://openai.com/index/memory-and-new-controls-for-chatgpt/ [accessed 2025-10-18]
53. Yan L, Greiff S, Teuber Z, Gašević D. Promises and challenges of generative artificial intelligence for human learning. Nat Hum Behav 2024;8(10):1839-1850. [doi: 10.1038/s41562-024-02004-5] [Medline: 39438686]
54. Carey TA. Conflict, as the achilles heel of perceptual control, offers a unifying approach to the formulation of psychological problems. Couns Psych Rev 2008;23(4):5-16. [doi: 10.53841/bpscpr.2008.23.4.5]
55. De Freitas J, Cohen IG. The health risks of generative AI-based wellness apps. Nat Med 2024;30(5):1269-1275. [doi: 10.1038/s41591-024-02943-6] [Medline: 38684859]

Abbreviations

AI: artificial intelligence
MYLO: Manage Your Life Online
PHQ-9: Patient Health Questionnaire-9

XSL•FO
RenderX

Edited by J Sarvestan; submitted 11.12.24; peer-reviewed by A Parks, N Logan; comments to author 02.06.25; revised version received 22.06.25; accepted 22.09.25; published 01.12.25.

Please cite as:
Wrightson-Hester AR, Anderson G, Dunstan J, McEvoy PM, Sutton CJ, Myers B, Egan SJ, Tai SJ, Johnston-Hollitt M, Chen W, Gedeon T, Moullin JC, Mansell W
A Rule-Based Conversational Agent for Mental Health and Well-Being in Young People: Formative Case Series During the Rise of Generative AI
JMIR Form Res 2025;9:e69841
URL: https://formative.jmir.org/2025/1/e69841
doi:10.2196/69841
PMID:

©Aimee-Rose Wrightson-Hester, Gee Anderson, Joel Dunstan, Peter M McEvoy, Christopher J Sutton, Bronwyn Myers, Sarah J Egan, Sara Jane Tai, Melanie Johnston-Hollitt, Wai Chen, Tom Gedeon, Joanna C Moullin, Warren Mansell. Originally published in JMIR Formative Research (https://formative.jmir.org), 01.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

WhatsApp-Based Intervention for Diabetes Prevention and Care in Argentina: Implementation and Process Evaluation

Analia Nejamis[1], MD; Laura Gutierrez[1], MSc; María Victoria López[1], BSc; Alvaro Daniel Ciganda Cha[2], MSc; Omayra Mendoza Quispe[3], MD; Vilma Irazola[1], PhD; Andrea Beratarrechea[4], MSc

[1]Department of Research in Chronic Diseases, Institute for Clinical Effectiveness and Health Policy (IECS), Buenos Aires, Argentina

[2]Institute for Clinical Effectiveness and Health Policy (IECS), Buenos Aires, Argentina

[3]Faculty of Medicine, National University of San Marcos, Lima, Peru

[4]Instituto de Medicina Traslacional e Ingeniería Biomédica (CONICET) – Instituto Universitario Hospital Italiano (IUHI) – Hospital Italiano de Buenos Aires (HIBA), Buenos Aires, null, Argentina

Corresponding Author:
Analia Nejamis, MD
Department of Research in Chronic Diseases
Institute for Clinical Effectiveness and Health Policy (IECS)
Emilio Ravignani 2024
Buenos Aires, C1414CPV
Argentina
Phone: 54 (11) 47778767
Email: anejamis@iecs.org.ar

Abstract

Background: In Argentina, diabetes is a growing public health concern, with a prevalence of 14% in 2024 and projections reaching 15.4% by 2050. In this context, a Diabetes Prevention and Care Program was implemented in low-income areas across 3 provinces. A key component of the program was a WhatsApp (WhatsApp LLC)-based intervention aimed at promoting self-care, encouraging healthy behaviors, and supporting follow-up among people with diabetes, those at risk, and pregnant women.

Objective: This study aimed to describe the implementation and process evaluation of a WhatsApp-based intervention within Argentina's public health system, using the Carroll Implementation Fidelity Framework, focusing on challenges encountered, implementation strategies used, and lessons learned across the 3 target populations.

Methods: The intervention was implemented in 40 primary care centers. The population included adults residing in the catchment areas of the selected primary care centers. Participants included adults with type 2 diabetes, people at moderate or high risk based on the Finnish Diabetes Risk Score, and pregnant women. A set of 192 educational and reminder messages was developed and validated through expert input and community feedback. Messages were tailored to each target population and delivered through WhatsApp via Twilio (Twilio Inc) Business API (application programming interface). We assessed implementation fidelity focusing on adherence to the intervention, participant responsiveness, quality of delivery, and contextual barriers.

Results: A total of 11,029 participants were enrolled in this study, of whom 9983 (90.5%) had a valid mobile phone number registered in the system. Among these, 32.8% (3276/9983) had a diagnosis of type 2 diabetes, 53.3% (5320/9983) were identified as being at moderate or high risk based on the Finnish Diabetes Risk Score questionnaire, and 13.9% (1387/9983) were pregnant women. Overall, 67.3% (n=5749) opted in to receive messages, with the highest acceptance among those with diabetes (n=2169, 74.3%) and the lowest among at-risk people (n=2935, 62.1%). Message adherence was high: 88.7% (n=5004) of participants received at least the minimum number of educational messages expected, and the mean proportion of messages read per participant was 82.2% (SD 29.8). The dropout rate was low (6.1%) but higher among pregnant participants (14.6%). Message delivery issues mostly included problems with WhatsApp on the mobile phones of participants. Technical challenges, including server overload, were addressed during implementation.

Conclusions: The WhatsApp-based intervention was feasible and well-received in public primary care settings in Argentina, particularly among people with diabetes. The experience illustrates how a WhatsApp-based intervention can be leveraged to strengthen service delivery in low-resource contexts, while also highlighting the need for further work on integration with electronic health records, tailoring of content to population needs, and strategies to enhance digital inclusion for underserved populations.

XSL•FO
RenderX

(JMIR Form Res 2025;9:e81098) doi:10.2196/81098

KEYWORDS

diabetes; prediabetes; mHealth; WhatsApp; text-message intervention; process evaluation; Latin America; primary care; implementation science; mobile phone

Introduction

Background

Diabetes mellitus remains one of the leading causes of mortality worldwide and represents a significant public health challenge. In 2024, an estimated 589 million adults (11.1% of the global population aged 20-79 years) were living with diabetes, and this figure is projected to rise to 853 million (13%) by 2050 [1]. Notably, around 95% of this increase is expected to occur in low- and middle-income countries [1]. The global burden is primarily driven by type 2 diabetes (T2D), which accounts for more than 90% of diagnosed cases and is associated with adverse cardiovascular, metabolic, maternal, and neonatal outcomes [1].

In Argentina, the prevalence of diabetes mellitus increased from 5.4% in 2021 to 14% in 2024, with estimates suggesting it may reach 15.4% by 2050 [1]. Combined with high levels of obesity (25.3%) and physical inactivity (44.2%), this trend highlights the urgent need to enhance prevention, early detection, and chronic disease management across the national health system [1,2]. The prevalence of gestational diabetes is estimated at 9.8% [3], further underscoring the national public health concern.

Considering that over 40% of the Argentine population relies exclusively on the public health care system [4], several initiatives have sought to strengthen T2D and gestational diabetes management at the primary care level. Strategies such as health care team training, patient education, scheduled follow-ups, and reminder phone calls have shown improvements in metabolic control and service use, particularly when both health care providers and patients were involved [5,6]. Other multicomponent programs have combined team training with educational outreach visits and digital health strategies, improving access to laboratory tests and ophthalmologic care, blood pressure control, and follow-up visits at 12 months [7]. Despite these efforts, sustaining effective strategies over time remains challenging, especially in resource-constrained primary care settings [8-10].

Mobile phones are now widely used to deliver health information remotely, often through text messaging [11]. This modality provides people with accessible health information and practical guidance to support daily self-management of their health. Given its relatively low cost, potential scalability, and the fact that text messaging does not require advanced technological skills, mobile communication offers a promising way for public health systems seeking to improve care for people living with chronic diseases [12]. These digital tools provide practical and cost-effective alternatives to conventional health promotion strategies, and studies evaluating these interventions generally report high feasibility and acceptability [13].

Argentina has high mobile phone and internet connectivity: 92% of the population uses a mobile phone and 86% uses the internet (Q4 2021), placing the country among the most digitally connected in Latin America [14,15]. WhatsApp is also widely used; 86% of internet users report weekly use in 2021 [16]. This high level of digital connectivity stands in contrast to Argentina's persistent social inequalities, where nearly 40% of the population currently lives below the poverty line [17]. As a result, mobile technology is deeply integrated into daily life, particularly through WhatsApp, the most widely used and culturally accepted messaging platform in the country, offering a feasible channel for delivering health interventions.

WhatsApp-Based Intervention as Part of a Diabetes Prevention and Care Program

Between 2021 and 2024, a Diabetes Prevention and Care Program was implemented within Argentina's public health system, targeting low-income areas of 3 locations in Argentina: Salta, San Juan, and Tandil (Province of Buenos Aires). The program included training for health care teams, as well as screening and care strategies for 3 target populations: people with diabetes, those at high risk of developing the disease, and pregnant women.

A key component of the program was a WhatsApp-based intervention that delivered educational messages aimed at enhancing knowledge, promoting healthy lifestyle changes, supporting medication adherence, and functioning as motivational prompts to enhance self-management behaviors [18].

Using WhatsApp via business APIs (application programming interfaces) allows us to pair a familiar channel with delivery and read status indicators that enable auditable reach, and to implement an explicit, channel-specific opt-in to document and reconfirm consent. This supports a minimally intrusive, scalable approach to public health messaging within the public primary-care system. This study aimed to describe the implementation and process evaluation of this WhatsApp-based intervention within Argentina's public health system, using the Carroll Implementation Fidelity Framework [19], focusing on the challenges encountered, strategies used, and lessons learned across 3 target populations.

Methods

Setting and Population

This study was conducted between 2021 and 2024 in 3 urban settings of Argentina: Salta (capital city of Salta Province, northwest region), San Juan and Gran San Juan (covering 4 of the 5 health zones of San Juan Province, Cuyo region), and Tandil (a midsized city in Buenos Aires Province).

A total of 40 primary care centers (PCCs) were included. Each PCC is responsible for a geographically defined catchment area,

XSL•FO
RenderX

composed mostly of low-income neighborhoods where residents rely primarily on public health services. Eligible PCCs had to be located in socioeconomically disadvantaged areas, report at least 800 monthly outpatient visits, use community health workers (CHWs), and provide free medications through the national and provincial essential medicines supply program. PCCs were selected in coordination with local health authorities to ensure comparability with other centers in each city.

The population included adults residing in the catchment areas of the selected PCCs. Eligibility criteria included adults (aged 18 years or older) with T2D, people at moderate to high risk of developing T2D, and pregnant women. Participants were recruited through two complementary approaches: (1) opportunistic recruitment at PCCs, when individuals attended for routine visits or other health reasons; and (2) community-based recruitment, conducted by trained health workers through door-to-door household visits within the PCC catchment areas and community outreach activities (eg, neighborhood health fairs and community centers). CHWs used standardized screening forms, and received prior training on recruitment protocols, and field activities were supervised by PCC coordinators to ensure consistency across sites.

Data Collection

Trained health care workers administered the validated Finnish Diabetes Risk Score (FINDRISC) questionnaire which was used to screen for people at moderate or high risk of developing T2D [20]. Data were collected using an electronic form developed in REDCap (Research Electronic Data Capture; Vanderbilt University) [21,22]. CHWs and nurses, previously trained to use the system, gathered sociodemographic and clinical information, including age, gender, cardiovascular risk factors, diabetes history, lifestyle behaviors (eg, fruit and vegetable intake and physical activity), and current treatments. Anthropometric measurements, such as height, weight, and abdominal circumference, were also collected to calculate diabetes risk using the FINDRISC tool integrated into the system.

Participants with moderate or high risk of T2D, or with diabetes, or pregnant, were enrolled in this study.

Messaging System

Content Generation and Validation

A total of 192 tailored text messages were developed and validated: 40 for pregnant women, 87 for people with T2D, and 65 for those with moderate or high risk of developing T2D. Messages were categorized into key thematic categories, including general information, lifestyle modification, medication adherence, diabetes care, foot care, insulin use, hypoglycemia prevention, and follow-up reminders.

Content related to behavior change was developed based on principles of the Transtheoretical Model [23], which proposes 5 stages: precontemplation, contemplation, preparation, action, and maintenance. These were grouped into 3 stages, and messages were designed to align with the likely needs of each stage (eg, raising awareness in earlier stages, reinforcing motivation and skills in intermediate stages, and supporting long-term maintenance). A prior validation was conducted and included formative research through cognitive interviews with community members, expert review by local specialists and communicators, and pilot testing using a 7-item questionnaire to evaluate understanding, appeal, and suggestions for improvement. Messages related to diabetes care were adapted from a previous study of our team in Argentina and have shown good acceptability in people with diabetes living in low-resource settings [24-26]. Messages for pregnant women were developed with input from experts and local health authorities to ensure appropriateness and relevance.

A full list of validated messages is available in Multimedia Appendix 1.

Implementation Process

The messaging system followed an automated and organized sequence of steps for collecting participant data, creating messages, and delivering them. As shown in Figure 1, the process started with registering participant information in the REDCap data management platform. Once the forms were completed and all required fields validated, messages were generated using predefined algorithms implemented as an external module in a separate REDCap project. These algorithms considered the specific population subgroup to which the participant belonged, such as people with T2D (with or without insulin dependence), people with a moderate or high risk of developing T2D, or pregnant women classified by gestational week.

An ad hoc monitoring system operated as an intermediary between REDCap and the messaging system (Twilio). It performed automated daily queries of the REDCap database at scheduled intervals to identify pending messages for delivery. When applicable, these messages were transmitted to Twilio, which routed them to WhatsApp for final delivery to participants. Each step in the process was recorded systematically, including the delivery status of each WhatsApp message. Specific reports were generated for monitoring throughout this study to ensure the intervention was operating as intended.

XSL•FO
RenderX

Figure 1. Data flow diagram of the WhatsApp-based messaging system. API: application programming interface; REDCap: Research Electronic Data Capture.

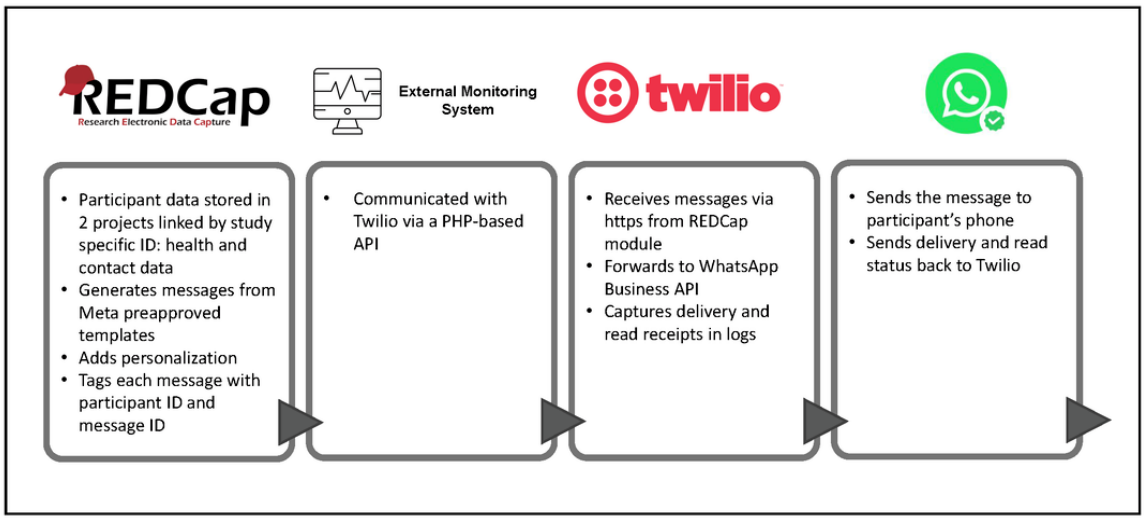

Delivery

Once the REDCap forms were completed and validated, the system automatically created the messages and initiated their delivery. The first message served as an opt-in request, asking participants to confirm their consent to receive messages as part of this study. Only participants who agreed to this initial message were included in the text messaging system. After confirmation, each participant received a welcome message, followed by educational, tailored content based on their population group. Pregnant women received messages according to their gestational week, continuing through the expected duration of their pregnancy. All messages were personalized using the participant's first name. Additionally, regional customization was incorporated into the welcome messages, including province-specific references to health authorities, such as the names of the local diabetes programs or ministries of health.

System Requirements

The following components were required to operate the messaging system: (1) data repository: REDCap was used to store participant information, message templates, and the algorithms for message delivery; (2) Twilio and WhatsApp integration: a verified Meta Business account was required to configure and authorize the WhatsApp Business API via Twilio; (3) cell phone lines: dedicated local numbers were acquired through Twilio to enable automated message delivery via WhatsApp; (4) preapproved message templates: WhatsApp mandated that all message templates be reviewed and approved before deployment; and (5) API-based communication: the control system communicated with Twilio via a PHP-based API, while Twilio interacted with the control system via webhooks.

Data Management and Security

As illustrated in Figure 1, each message was generated in REDCap through an external module that applied Meta-approved templates. Templates included a common header, a personalized greeting with the participant's first name, and the educational content scheduled for that week. Each message was tagged with both a participant ID and a message ID.

Delivery and read receipts were retrieved from Twilio logs and matched to participants using phone number and message text. These logs were consolidated weekly through R (R Project for Statistical Computing; R Foundation) scripts to cover the full study period.

Contact details were stored in a separate REDCap project from health data, and linkage was made only via a study-specific participant ID. Phone numbers were retained solely to enable message delivery.

Access to Twilio logs was restricted to account administrators, who downloaded them to a secure shared folder for statistical analysis. Twilio retained logs of up to 10,000 lines (≈3-6 months of activity), which were then consolidated. Data transfers were encrypted in transit (https). Data were stored on cloud servers without encryption at rest. No additional procedures were implemented to minimize exposure of phone numbers, which were necessary for message delivery.

Participants retained the right to withdraw at any time, either by replying with an opt-out request or by directly contacting this study's team using the information provided in the consent form.

Evaluation Framework

The evaluation was based on Carroll Implementation Fidelity Framework [19], which assesses both the implementation process and the extent to which participants received the intervention as intended (Table 1).

This framework has been widely applied in previous studies [27-29] to evaluate the consistency and quality of intervention delivery. It focuses on key domains such as adherence, quality of delivery, participant responsiveness, and contextual factors.

In our process evaluation, using data collected at various stages of the intervention, adherence was examined by determining whether the planned activities were carried out as intended. This

XSL•FO
RenderX

included verifying that all key messages, including opt-in (ie, initial confirmation of consent to receive weekly messages), welcome, weekly, and final messages, were successfully generated and delivered to participants. Additionally, we analyzed the coverage of the intervention across the 3 participant subgroups: people with T2D, those at moderate or high-risk of developing diabetes, and pregnant participants.

Participant responsiveness was measured by calculating the opt-in acceptance rate. We also tracked the proportion of messages that were read by participants relative to those delivered. Participants' withdrawal from the intervention was also documented.

Contextual factors that may have impacted implementation were also considered, including technical issues and challenges reported by participants.

Table 1. Fidelity dimensions and indicators.

Dimensions	Research questions	Indicators
Adherence	Was the WhatsApp-based intervention implementation as planned?	• Opt-in messages generated: proportion of participants with an opt-in message generated/participants who had a valid cellphone number. • Welcome message generated: proportion with a welcome message generated/participants who accepted the opt-in message. • Final message generated: proportion with a final message generated/participants who completed the intervention.
Adherence	Did the intervention reach all target populations?	• Proportion of participants with a valid cellphone number. • Opt-in messages delivered: proportion receiving the opt-in message/participants with an opt-in message generated. • Welcome messages delivered: proportion receiving the welcome message/participants with a welcome message generated. • Educational messages delivered (minimum expected): proportion receiving at least the minimum number of educational messages expected/participants who accepted the opt-in message. • Educational messages delivered (average): average number of educational messages received per participant. • Final messages delivered: proportion receiving the final message/participants with a final message generated.
Participant responsiveness	What was the level of participant engagement?	• Acceptance rate of opt-in messages: proportion who accepted the opt-in message/participants who received the opt-in message. • Messages read[a]: proportion of delivered messages that were read/messages sent. • Dropout rate: proportion of participants who withdrew from the intervention.
Quality of delivery	Were delivery errors minimized?	• Proportion of message delivery failures.
Contextual factors	What barriers affected implementation?	• Description of technical issues and participant-related challenges.

[a]Indicate message opening.

Data Analysis

Descriptive statistics were used to summarize sociodemographic and clinical characteristics overall and by target group, and to describe implementation fidelity indicators. Group comparisons assessed differences in access (valid cellphone number) and acceptability (opt in). As a sensitivity analysis, multivariable logistic regression models including sex, age, educational attainment, and comorbidity category were fitted for each outcome; odds ratios with 95% CIs are reported in Table S1 in Multimedia Appendix 2 and Table S1 in Multimedia Appendix 3.

Ethical Considerations

This study was reviewed and approved by 3 institutional ethics committees in Argentina: the Institutional Review Board for Research Studies (Consejo Institucional de Revisión de Estudios de Investigación) of the Hospital Privado de Comunidad in Mar del Plata, Province of Buenos Aires (approved date: July 7, 2022); the Research Ethics Committee of the Ministry of Public Health of San Juan (May 12, 2021); and the Provincial Commission for Biomedical Research of the Ministry of Public Health of Salta (March 8, 2021).

Participation in this study was voluntary, and written informed consent was obtained from all participants during the screening process. Consent covered both study data collection and the receipt of digital messages. Participants confirmed their enrollment by explicitly opting in to the messaging system. They retained the right to withdraw at any time, either by replying directly to the system or by contacting this study's team using the information provided in the consent form. Participants did not receive financial or material compensation for their participation in this study, and no identifiable images of individual participants are included in this paper or its supplementary materials.

Confidentiality was strictly maintained, and all procedures complied with Argentina's Personal Data Protection Law (Ley 25.326). The use of WhatsApp Business API required attention to privacy and data protection. Messages were sent using

XSL•FO
RenderX

standardized templates and contained no personal health information. Participant data were stored in REDCap with password-protected access. While messages were transmitted over encrypted https connections, data were stored on secure cloud servers.

Results

Participant Characteristics

A total of 11,029 participants were enrolled in this study, of whom 9983 (90.5%) had a valid mobile phone number registered in the system. Among these, 32.8% (3276/9983) had a diagnosis of T2D, 53.3% (5320/9983) were identified as being at moderate or high risk based on the FINDRISC questionnaire, and 13.9% (1387/9983) were pregnant women (Figure 2).

Table 2 summarizes the main sociodemographic and clinical characteristics of participants enrolled in the Diabetes Prevention and Care Program.

Most participants were women (n=7725, 77.4%), with a mean age of 44.5 (SD 15.6) years. Participants with a valid cellphone number tended to be slightly younger compared to those who could not be reached. Age distribution also varied across target groups, from a mean of 51.9 (SD 13.6) years among people with diabetes to 26.4 (SD 5.9) years among pregnant participants.

More than half of all participants (n=5332, 53.4%) had completed or partially completed secondary education. Among those with a valid cellphone number, a higher proportion had completed secondary education, whereas among participants without a valid phone number, the majority had only completed primary education. In target groups, the proportion with secondary education was higher among the pregnant women group (1008/1387, 72.7%). In contrast, people with diabetes showed lower educational attainment, with 46.3% (1516/3276) having completed only primary education or less.

Comorbidities were common, affecting 76% (7584/9983) of participants overall. Those with a valid cellphone number more frequently reported two or more conditions.

XSL•FO
RenderX

Figure 2. Participants flow diagram. *Among the 5749 participants who consented, 5004 reached the minimum message dose; 745 did not. Dropouts (n=351) occurred independently of reaching the minimum dose.

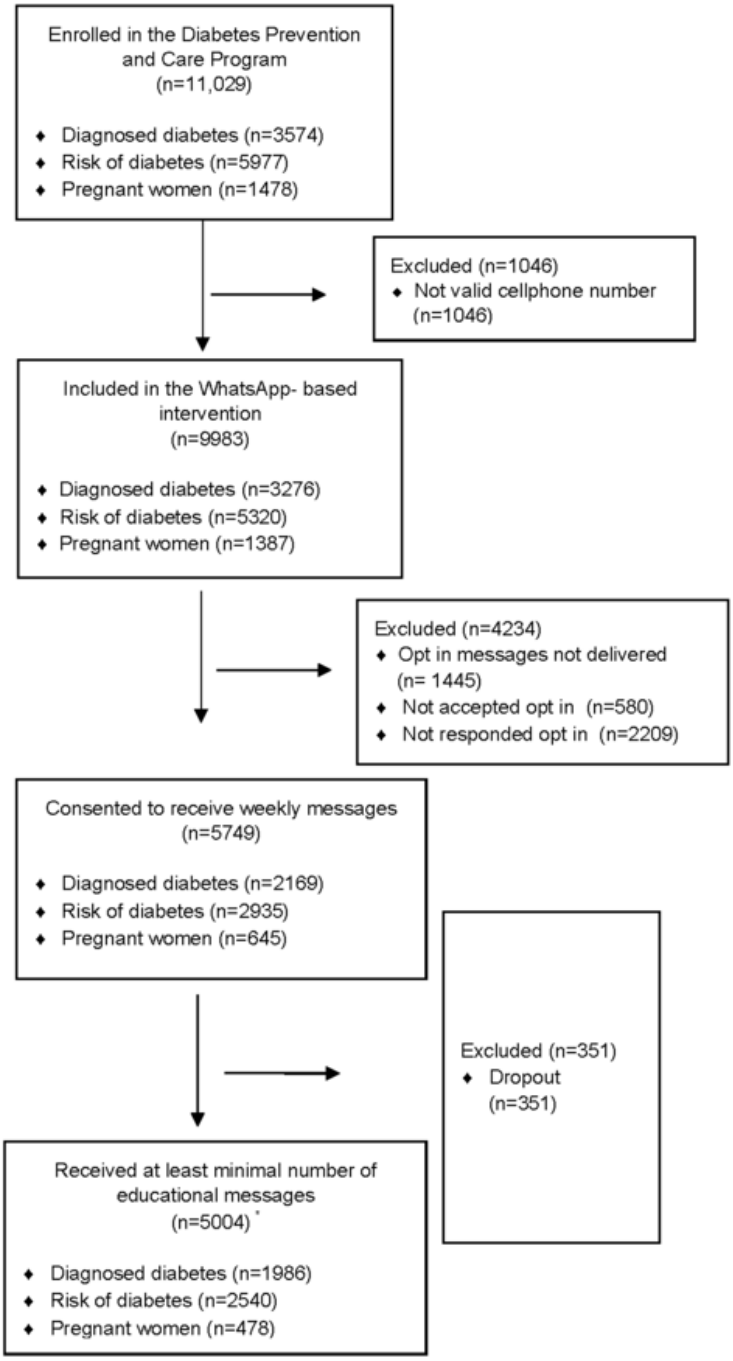

XSL•FO
RenderX

Table 2. Sociodemographic and clinical characteristics of participants enrolled.

Characteristics	Total	Valid cellphone number	
		Yes	No
Participants, N	11,029	9983	1046
Female, n (%)	8476 (76.9)	7725 (77.4)	751 (71.8)
Age (years), mean (SD)	45.0 (15.7)	44.5 (15.6)	49.6(16.9)
Educational level[a], n (%)			
Primary school	4004 (36.3)	3554 (35.6)	450 (43)
Secondary school	5811 (52.7)	5332 (53.4)	479 (45.8)
University or higher	927 (8.4)	858 (8.6)	69 (6.6)
Not available	287 (2.6)	239 (2.4)	48 (4.6)
Comorbidities[b], n (%)			
None	2644 (24)	2395 (24)	249 (23.8)
One condition	3629 (32.9)	3252 (32.6)	377 (36)
Two or more conditions	4743 (43)	4327 (43.3)	416 (39.8)

[a]Includes complete and incomplete levels.

[b]Includes obesity, hypertension, hypercholesterolemia, and cardiovascular disease.

Implementation Evaluation

Adherence to Intervention and Responsiveness

Of the 11,029 enrolled participants, 9983 (90.5%) had a valid cell phone number registered, necessary for intervention delivery. This proportion was highest among pregnant women (n=1387, 93.8%) and lowest among people at risk of diabetes (n=5320, 89%; Table 3).

Opt-in messages were successfully generated for 95.6% (n=9546) of participants with a valid cellphone number, with nearly complete coverage in the diabetes group (n=3253, 99.3%) and the at-risk group (5269/5320, 99%), but lower in the pregnant group (n=1024, 73.8%). Of the generated opt-in messages, 89.4% (n=8538) were successfully delivered, and 91.1% (n=7778) of those delivered were read. Reading rates were slightly higher in the diabetes group (2699/2918, 92.5%) and showed similar levels across groups.

A total of 67.3% (n=5749) of participants accepted the opt-in message, while 25.9% (2209/8538) did not respond, and 6.8% (580/8538) declined to participate. Acceptance was highest among people with diabetes (n=2169, 74.3%), followed by the pregnant group (n=645, 72%) and those at risk of diabetes (n=2935, 62.1%).

Regarding weekly educational content messages, 88.7% (n=5004) of participants received at least the minimum number of messages expected, defined as 20 for people with diabetes and 12 for those at risk. Adherence was highest among the diabetes group (n=1986, 91.6%), followed by the at-risk group (n=2540, 86.5%), and lowest among pregnant participants (n=478, 74.1%), for whom the expected number of messages varied according to gestational week at the time of enrollment, and the mean proportion of messages read per participant was 82.2% (SD 29.8). Final messages were generated for 85.5% (n=1854) of participants in the diabetes group and 90.3% (n=2649) in the at-risk group and delivered in 93.2% (n=4199) of those cases.

The dropout rate among participants who had accepted the messaging intervention was 6.1%, with the lowest rate in the diabetes group (4.4%) and the highest among pregnant participants (14.6%).

XSL•FO
RenderX

Table 3. Indicators of adherence to intervention and responsiveness by target group[a].

Indicators	Total	Diagnosed diabetes	Risk of diabetes	Pregnant women
Participants, n	11,029	3574	5977	1478
With a valid cellphone number, n (%)	9983 (90.5)	3276 (91.7)	5320 (89)	1387 (93.8)
Opt-in message generated, n (%)	9546 (95.6)	3253 (99.3)	5269 (99)	1024 (73.8)
Opt-in message delivered, n (%)	8538 (89.4)	2918 (89.7)	4724 (89.7)	896 (87.5)
Opt-in message read, n (%)	7778 (91.1)	2699 (92.5)	4270 (90.4)	809 (90.3)
Accepted opt-in message[b], n (%)	5749 (67.3)	2169 (74.3)	2935 (62.1)	645 (72)
Welcome message generated, n (%)	5582 (97.1)	2077 (95.8)	2860 (97.4)	645 (100)
Welcome message delivered, n (%)	5318 (95.3)	1988 (95.7)	2724 (95.2)	606 (94)
Welcome message read, n (%)	4991 (93.9)	1887 (94.9)	2547 (93.5)	557 (91.9)
Educational messages				
Received ≥ minimum expected messages[c], n (%)	5004 (88.7)	1986 (91.6)	2540 (86.5)	478 (74.1)
Number of messages delivered per participant				
Mean (SD)	34.2 (20.9)	45 (23.4)	30 (16.5)	19 (9.5)
Median (IQR)	27 (22-44)	40 (27-60)	25 (22-35)	19 (12-27)
Messages read per participant out of delivered (%)				
Mean (SD)	82.2 (29.8)	83.6 (28.9)	81.2 (30.5)	82.3 (29.6)
Median (IQR)	100 (77-100)	100 (82-100)	100 (73-100)	100 (75-100)
Final message generated[d], n (%)	4503 (78.3)	1854 (85.5)	2649 (90.3)	—[e]
Final message delivered, n (%)	4199 (93.2)	1744 (94.1)	2455 (92.7)	—
Final message read, n (%)	3392 (80.8)	1477 (84.7)	1915 (78)	—

[a]Percentages were calculated sequentially, using the frequency in the preceding row as the denominator, unless otherwise specified.

[b]Denominator: opt-in messages delivered.

[c]Participants with diabetes who received at least 20 messages, and those at risk of diabetes who received at least 12 messages. For pregnant women, the number of messages calculated was based on the gestational week at the time of enrollment in this study.

[d]Denominator: opt-in messages accepted.

[e]Not available.

Participant Acceptability

Participant characteristics according to opt-in message acceptance are detailed in Table 4. Additional disaggregated data by target group are provided in Table S1 in Multimedia Appendix 4.

Among the 8538 people who received the opt-in message, 67.3% (n=5749) accepted, while 32.6% (2789/8538) either declined or did not respond. As a complementary analysis to evaluate the impact of initial noncoverage, when considering the total enrolled population as the denominator, the overall acceptance rate was 52.1% (5749/11,029) instead of 67.3% (n=5749) among those with a valid phone number.

A higher proportion of women accepted the intervention (n=4515, 78.5%) compared to those who did not accept (n=2057, 73.8%), while men represented 21.5% (n=1234) of those who accepted and 26.2% (n=732) of those who did not. The mean age was slightly lower among those who agreed to receive the messages (44.0, SD 14.6 years) compared to those who did not (46.7, SD 16.0 years).

Among participants who accepted the intervention, a greater proportion had completed secondary school (n=3123, 54.3%) compared to primary school (n=1936, 33.7%). In the group that did not accept, the proportion with primary education (n=1105, 39.6%) was higher than among those who accepted. The prevalence of self-reported comorbidities was similar across groups, with 78.1% (4489/5749) among those who accepted and 77.8% (2170/2788) among those who did not.

In sensitivity analyses with adjusted models, older age was consistently associated with lower odds of both having a valid cellphone number and accepting opt-in, whereas multiple comorbidities were associated with higher odds. For the opt-in accepted outcome, male sex and lower educational attainment were also associated with reduced odds (Table S1 in Multimedia Appendix 2 and Table S1 in Multimedia Appendix 3).

XSL•FO
RenderX

Table 4. Participant characteristics by opt-in message acceptance.

Characteristics	Opt-in message accepted	
	Yes	No[a]
Participants, N	5749	2789
Female, n (%)	4515 (78.5)	2057 (73.8)
Male, n (%)	1234 (21.5)	732 (26.2)
Age (years), mean (SD)	44.0 (14.6)	46.7 (16.0)
Educational level[b], n (%)		
Primary school	1936 (33.7)	1105 (39.6)
Secondary school	3123 (54.3)	1399 (50.2)
University or higher	571 (9.9)	202 (7.2)
Not available	119 (2.1)	83 (3)
Comorbidities[c], n (%)		
None	1258 (21.9)	615 (22.1)
One condition	1831 (31.8)	980 (35.1)
Two or more conditions	2658 (46.2)	1190 (42.7)

[a]Includes those who declined or did not respond to the opt-in message.

[b]Includes complete and incomplete levels.

[c]Includes obesity, hypertension, hypercholesterolemia, and cardiovascular disease.

Quality of Delivery

The most commonly reported delivery issues included phone numbers not enabled for WhatsApp, users who had not accepted WhatsApp's terms of service, outdated or unsupported versions of the WhatsApp app, and improperly formatted phone numbers.

Contextual Factors

Beyond individual delivery issues, the system faced broader contextual challenges. Some technical difficulties were encountered during the delivery of educational text messages via WhatsApp using the Twilio interface. At some point of this study, during periods of high system load, the server's processing capacity was insufficient to manage the volume of data exchange until the server was upgraded. This occasionally resulted in delays in the registration of message delivery status, particularly when response signals were received before outgoing messages could be properly logged. Additionally, for pregnant women, some opt-in messages could not be generated due to missing or incorrectly coded gestational age data.

Additional limitations included suboptimal code performance and intermittent communication failures between the messaging control system and the REDCap platform. The underlying code was subsequently optimized to handle the increased message volume as more participants were enrolled.

To ensure accurate message tracking, Twilio logs were manually retrieved every week and matched with REDCap data using participant phone numbers.

Discussion

Principal Findings

This study examined the implementation of a WhatsApp-based intervention for people with T2D, moderate or high risk of developing this condition, and pregnant women, in the context of Argentina's public health system, focusing on adherence, engagement, delivery quality, and contextual barriers. The intervention followed a structured delivery process, from opt-in messages to weekly health education content and final closing messages.

There is robust evidence that text messaging interventions can support chronic disease management, such as in T2D [30,31]. However, limited evidence exists on how these programs are implemented in routine primary care settings, particularly in low-resource environments [32]. Our findings contribute to addressing this gap by documenting the novel use of WhatsApp for digital health messaging in the real-world public health sector in Argentina, providing insights on feasibility, potential scalability, and engagement in resource-constrained settings.

While most messages were successfully delivered, a small proportion were affected by technical issues, such as invalid cellphone numbers (approximately 1046/11,029, 10%) or intermittent connectivity problems. Similar challenges have been reported in other mobile health interventions for chronic disease [27,33]. Steinman et al [33] reported that 40% of participants did not receive messages due to frequent changes in personal mobile numbers driven by cost-saving strategies, while Van Olmen et al [27] found that nearly one-third of project phones became unusable because of subscription cancellations, breakage, or loss, despite providing SIM cards and devices.

XSL•FO
RenderX

Both studies also described infrastructural limitations, such as weak network coverage and recurrent software errors [33].

In comparison, the proportion of unreachable participants in our study was lower, suggesting that integrating WhatsApp within existing care structures, with validated contact information and proactive follow-up, may help mitigate some of these barriers. Nonetheless, the percentage of nonreached people highlights possible inequities: those who could not be reached tended to be older, more often male, and with lower educational levels compared to those successfully contacted. Although the differences were modest, these patterns reflect potential digital exclusion linked to socioeconomic and demographic disparities. These trends are consistent with prior evidence showing that older age is negatively associated with adoption of mobile health technologies, while higher educational attainment increases access and use [34]. These patterns could be related to the lower digital literacy, physical or cognitive limitations, and greater concerns about privacy for older adults, whereas people with lower education are more likely to experience restricted access to smartphones, unstable connectivity, or difficulties interpreting health messages. Taking into account the inequities in phone ownership, access, and use [35], future interventions should consider tailored recruitment strategies to ensure that no population is systematically excluded due to digital access limitations, such as providing guidance for phone access and registration, using complementary communication channels for those without WhatsApp, and improving the process of validating contact information.

Another critical component of implementation is participant engagement and acceptability. In this study, we assessed these aspects by examining the read rate of the initial opt-in message, the proportion of participants who agreed to receive weekly messages, the weekly message read rates during the intervention, and dropout rates over time. Engagement levels were high: approximately two-thirds of participants accepted the intervention, over 90% (7778/8538) read the opt-in message, weekly read rates averaged 82.2% (SD 29.8), and only 6.1% of participants dropped out. Our findings are consistent with previous research [24,31,36], showing that text messaging interventions are generally well-received by both participants and health care providers. The engagement observed in this study may be related to the intervention's carefully designed structure, including its integration into routine primary care practice and the proactive efforts of health care teams to validate contact information.

These findings support the potential of WhatsApp as an effective platform to reach and engage broad populations and sustain participation, particularly when integrated into existing care structures. Compared to SMS, WhatsApp might offer advantages such as greater user familiarity, support for multimedia content, and faster message visibility, which could enhance engagement in low-resource settings.

Participant engagement levels differed across the 3 target groups. People with diabetes showed the highest acceptance and lowest dropout rates, likely reflecting a greater perceived need for ongoing support. Those at risk of diabetes had the lowest acceptance, while pregnant women showed higher acceptance of the messages than the at-risk group, but had a higher dropout rate. In this sense, although there is evidence that text messaging interventions can be effective during pregnancy, particularly for promoting healthy behaviors [37-40], the higher dropout rates observed in this group may reflect the competing demands pregnant women often face, such as medical appointments, childcare responsibilities and work, as well as potential preferences for more personalized messaging which may limit their ability to remain engaged with supplementary programs [41,42].

In our study, people identified as being at risk of developing diabetes were the least likely to accept the intervention. This group was, on average, younger than those with diabetes, which may have contributed to a lower perceived awareness, diminished need for support, and reduced motivation to engage with preventive interventions. These findings highlighted the importance of designing targeted strategies that enhance risk awareness and encourage participation in early-stage prevention efforts. Other studies have reported comparable trends, highlighting that engagement levels significantly impact long-term outcomes, particularly in prediabetes self-management programs [43,44].

Sociodemographic factors also influenced acceptability in our study, with men, older participants, and those with lower educational attainment consistently less likely to opt in across all target populations. These patterns underscore that even within targeted populations, digital interventions may reach participants unevenly, potentially reinforcing existing sociodemographic inequities. An analysis of barriers and facilitators of mobile health adoption among people with low socioeconomic position supports these trends [45], identifying multiple factors such as digital literacy, trust, and need for in-person support, that are critical for successful adoption and use. Similarly, a systematic review of mobile health use [34] found that women were more likely to adopt mobile health services than men, indicating that lower engagement among men may reflect differences in health-seeking behavior and health awareness. In light of these findings, tailored strategies that consider gender, age, and educational level, such as proactive outreach, simplified content, and personalized relevant messaging, may enhance relevance and comprehension and could help maximize engagement across all subgroups [35].

Another key aspect in our study was the resolution of technical barriers, particularly those related to the high volume of data processed. One of the main issues encountered was system overload due to the high rate of outgoing messages, which exceeded the initial server's capacity. This led to delays in updating message delivery status, complicating the integration with the REDCap platform and obstructing real-time monitoring. To address these challenges, substantial improvements were implemented to both the server infrastructure and the system's code. The server was upgraded with expanded memory and transitioned to a more robust disk configuration, increasing its capacity to handle large data volumes. In parallel, code optimizations, particularly in the communication protocols with REDCap, streamlined data processing and improved log matching. These enhancements strengthened the reliability of

XSL•FO
RenderX

message delivery and tracking, reduced the need for manual interventions, and improved overall system efficiency.

Additionally, using WhatsApp via the Business API, a familiar and widely accessible platform in Argentina, minimized technical barriers. The API enabled automated logging of delivery and read statuses, and captured explicit, channel-specific opt-ins and opt-outs, providing auditable, high-fidelity process data. Together, these technical improvements contributed to the intervention's feasibility.

Limitations

This real-world implementation study demonstrated that WhatsApp-based text messaging interventions were successfully integrated into public health services; however, several limitations should be considered when interpreting these results.

First, while messages were tailored by participant subgroups (people with T2D, at-risk of diabetes, or pregnant participants), they were not customized at the individual level. This lack of individual tailoring may have limited the intervention's acceptability and adherence, as more personalized content or follow-up could have increased relevance and engagement [11]. However, we implemented a local validation of the message's content at the regional level and included data privacy warnings in some messages, thereby enhancing the contextual relevance and participant trust.

Second, the intervention included general reminder messages encouraging participants to attend medical appointments, but lacked an individualized reminder system. Evidence suggests that individualized reminders, with specific dates and times, can significantly improve attendance in primary care settings [31,46]. Future integration with electronic health records could enable automated, personalized notifications. Third, another limitation of this study is the potential for selection and participation bias. Participants without a valid cellphone number who could not be reached differed modestly in terms of age and educational level compared to those who could be reached. Participation in the intervention also depended on willingness to opt in, which may have led to the overrepresentation of individuals who are more digitally connected, engaged with the health system, or have higher health literacy. As a result, engagement estimates may not be fully generalizable. Future implementations should consider strategies to enhance digital inclusion and monitor differential participation from the outset. Additionally, randomized controlled trials could provide complementary high-quality evidence on intervention effectiveness.

Fourth, despite WhatsApp's widespread use in Argentina, engagement with messages varied across participants. This suggests that high platform penetration alone does not guarantee acceptance or behavioral change. Future adaptations should consider factors such as message content, perceived relevance, and user trust to optimize engagement across different subgroups. Finally, the engagement metrics reported reflect process outcomes rather than behavioral or clinical impact. Changes in self-care practices or glycemic control are beyond the scope of this evaluation and will be analyzed separately alongside other components of the Diabetes Prevention and Care Program.

Conclusions

This study evaluated the implementation of a WhatsApp-based text messaging intervention for diabetes prevention and care in 3 locations within Argentina's public primary health care system. The intervention successfully reached key target groups (people with diabetes, individuals at risk, and pregnant women), demonstrating the feasibility of integrating digital messaging into routine care. However, engagement varied across participants, and several technical and contextual factors likely influenced message adherence and acceptance.

Integrating digital tools into primary care and adapting them to local contexts can strengthen service delivery, but future efforts should focus on improving interoperability with electronic health records, personalizing content by cohort and comorbidity profile, and ensuring equitable access for populations with limited digital connectivity.

Acknowledgments

The authors gratefully acknowledge the support of the collaborating institutions: the Ministry of Public Health of the Province of San Juan, the Ministry of Health of the Province of Salta, and the Ministry of Public Health of Tandil. The authors are also deeply grateful to the health professionals and teams from the primary care centers in San Juan, Salta, and Tandil for their commitment and participation throughout the implementation of the Program. This project was funded by the World Diabetes Foundation (WDF), project ID WDF19-1724. The funder had no involvement in this study's design, data collection, analysis, interpretation, or the writing of this paper.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

AB, VI, and AN conceptualized this study. ADCC and LG curated the data and formally analyzed it. VI and AB acquired the funding. AN and MVL investigated. AN, VI, AB, and LG handled the methodology of this project. VI, AN, and MVL administered the same. ADCC, AB, and LG worked on the software. AN, LG, and OMQ visualized the data. AN, LG, MVL, ADCC, OMQ, VI, and AB wrote the original draft. AN and LG reviewed and edited the writing.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
List of validated text messages.
[DOCX File , 70 KB - formative_v9i1e81098_app1.docx]

Multimedia Appendix 2
Sociodemographic characteristics associated with valid cellphone number in participants.
[DOCX File , 19 KB - formative_v9i1e81098_app2.docx]

Multimedia Appendix 3
Sociodemographic characteristics associated with intervention acceptance (opt-in).
[DOCX File , 19 KB - formative_v9i1e81098_app3.docx]

Multimedia Appendix 4
Participant characteristics by opt-in message acceptance and target group.
[DOCX File , 19 KB - formative_v9i1e81098_app4.docx]

References

1. Diabetes Atlas. 11th edition. International Diabetes Federation. URL: https://diabetesatlas.org/ [accessed 2025-10-11]
2. 4° Encuesta Nacional de Factores de Riesgo: Resultados definitivos. Argentina: Instituto Nacional de Estadística y Censos (INDEC), Secretaría de Gobierno de Salud de la Nación; 2019. URL: https://www.indec.gob.ar/ftp/cuadros/publicaciones/enfr_2018_resultados_definitivos.pdf [accessed 2025-10-10]
3. Gorban de Lapertosa S, Sucani S, Salzberg S, Alvariñas J, Faingold C, Jawerbaum A, DPSG-SAD Group. Prevalence of gestational diabetes mellitus in Argentina according to the Latin American Diabetes Association (ALAD) and International Association of Diabetes and Pregnancy Study Groups (IADPSG) diagnostic criteria and the associated maternal-neonatal complications. Health Care Women Int 2021;42(4-6):636-656. [doi: 10.1080/07399332.2020.1800012] [Medline: 32886577]
4. Censo Nacional de Población, Hogares y Viviendas 2022. Síntesis de resultados. 1a ed. Argentina: Instituto Nacional de Estadística y Censos (INDEC); 2025. URL: https://www.indec.gob.ar/ftp/cuadros/poblacion/sintesis_resultados_censo2022.pdf [accessed 2025-10-10]
5. Prestes M, Gayarre MA, Elgart JF, Gonzalez L, Rucci E, Paganini JM, DIAPREM (DIAbetes Primary Care, Registry, EducationManagement). Improving diabetes care at primary care level with a multistrategic approach: results of the DIAPREM programme. Acta Diabetol 2017 Sep;54(9):853-861. [doi: 10.1007/s00592-017-1016-8] [Medline: 28624898]
6. Gagliardino JJ, Lapertosa S, Pfirter G, Villagra M, Caporale JE, Gonzalez CD, PRODIACOR. Clinical, metabolic and psychological outcomes and treatment costs of a prospective randomized trial based on different educational strategies to improve diabetes care (PRODIACOR). Diabet Med 2013 Sep;30(9):1102-1111. [doi: 10.1111/dme.12230] [Medline: 23668772]
7. Morelli DM, Rubinstein F, Santero M, Gibbons L, Moyano D, Nejamis A, et al. Effectiveness of a diabetes program based on digital health on capacity building and quality of care in type 2 diabetes: a pragmatic quasi-experimental study. BMC Health Serv Res 2023 Jan 31;23(1):101 [FREE Full text] [doi: 10.1186/s12913-023-09082-7] [Medline: 36721213]
8. Kumah E, Kwawuvi DK. Bridging the diabetes care gap: cost-effective impact of self-management education and support. Acad Med 2025;2(2):1-14. [doi: 10.20935/acadmed7742]
9. Davis J, Fischl AH, Beck J, Browning L, Carter A, Condon JE, et al. 2022 National Standards for Diabetes Self-Management Education and Support. Sci Diabetes Self Manag Care 2022 Feb;48(1):44-59. [doi: 10.1177/26350106211072203] [Medline: 35049403]
10. Chatterjee S, Davies MJ, Heller S, Speight J, Snoek FJ, Khunti K. Diabetes structured self-management education programmes: a narrative review and current innovations. Lancet Diabetes Endocrinol 2018 Feb;6(2):130-142. [doi: 10.1016/S2213-8587(17)30239-5] [Medline: 28970034]
11. Tam HL, Wong EML, Cheung K, Chung SF. Effectiveness of text messaging interventions on blood pressure control among patients with hypertension: systematic review of randomized controlled trials. JMIR mHealth uHealth 2021 Sep 22;9(9):e24527 [FREE Full text] [doi: 10.2196/24527] [Medline: 34550078]
12. Déglise C, Suggs LS, Odermatt P. Short message service (SMS) applications for disease prevention in developing countries. J Med Internet Res 2012 Jan 12;14(1):e3 [FREE Full text] [doi: 10.2196/jmir.1823] [Medline: 22262730]
13. Formagini T, Gonzalez CJ, Dias J, Arredondo EM, Hekler E, O'Brien MJ. Advancing type 2 diabetes prevention through text-messaging interventions: a narrative review. Curr Diab Rep 2024 Dec 04;25(1):9. [doi: 10.1007/s11892-024-01568-w] [Medline: 39630283]

XSL•FO
RenderX

14. Acceso y uso de tecnologías de la información y la comunicación. EPH. Cuarto trimestre de 2021. Argentina: Instituto Nacional de Estadística y Censos (INDEC); 2022. URL: https://www.indec.gob.ar/uploads/informesdeprensa/mautic_05_22843D61C141.pdf [accessed 2025-10-10]
15. Tecnologías digitales para un nuevo futuro. Vitacura: Comisión Económica para América Latina y el Caribe (CEPAL) URL: https://edigital.economia.gob.sv/wp-content/uploads/2022/01/Tecnologias-digitales-para-un-nuevo-futuro-CEPAL.pdf [accessed 2025-10-11]
16. Newman N, Fletcher R, Schulz A, Andi S, Robertson C, Nielsen R. The Reuters Institute Digital News Report 2021. England, UK: Reuters Institute for the Study of Journalism; 2021.
17. Incidencia de La Pobreza e Indigencia En 31 Aglomerados Urbanos. Segundo Semestre de 2024. Ministerio de Economía de la República Argentina. Argentina: Instituto Nacional de Estadística y Censos (INDEC); 2024. URL: https://www.indec.gob.ar/uploads/informesdeprensa/eph_pobreza_03_252282AE14D2.pdf [accessed 2025-10-10]
18. Yu J, Fu Y, Du H, Wu Q, Zhang S, Zhao B, et al. The effectiveness of nudge-based interventions on self-monitoring behaviours among patients with cardiometabolic diseases: a systematic review and meta-analysis. Health Psychol Rev 2025 Jul 23:1-28. [doi: 10.1080/17437199.2025.2532017] [Medline: 40701164]
19. Carroll C, Patterson M, Wood S, Booth A, Rick J, Balain S. A conceptual framework for implementation fidelity. Implement Sci 2007 Nov 30;2:40 [FREE Full text] [doi: 10.1186/1748-5908-2-40] [Medline: 18053122]
20. Lindström J, Tuomilehto J. The diabetes risk score: a practical tool to predict type 2 diabetes risk. Diabetes Care 2003 Mar;26(3):725-731. [doi: 10.2337/diacare.26.3.725] [Medline: 12610029]
21. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381 [FREE Full text] [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
22. Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O'Neal L, REDCap Consortium. The REDCap consortium: building an international community of software platform partners. J Biomed Inform 2019 Jul;95:103208 [FREE Full text] [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]
23. Prochaska JO, DiClemente CC. Stages and processes of self-change of smoking: toward an integrative model of change. J Consult Clin Psychol 1983 Jun;51(3):390-395. [doi: 10.1037//0022-006x.51.3.390] [Medline: 6863699]
24. Moyano D, Morelli D, Santero M, Belizan M, Irazola V, Beratarrechea A. Perceptions and acceptability of text messaging for diabetes care in primary care in Argentina: exploratory study. JMIR Diabetes 2019 Mar 18;4(1):e10350 [FREE Full text] [doi: 10.2196/10350] [Medline: 30882362]
25. Diez-Canseco F, Zavala-Loayza JA, Beratarrechea A, Kanter R, Ramirez-Zea M, Rubinstein A, et al. Design and multi-country validation of text messages for an mHealth intervention for primary prevention of progression to hypertension in Latin America. JMIR mHealth uHealth 2015 Feb 18;3(1):e19 [FREE Full text] [doi: 10.2196/mhealth.3874] [Medline: 25693595]
26. Beratarrechea A, Diez-Canseco F, Fernández A, Kanter R, Letona P, Martinez H, et al. Aceptabilidad de una intervención basada en salud móvil para modificar estilos de vida en prehipertensos de Argentina, Guatemala y Perú: un estudio piloto. Rev Peru Med Exp Salud Publica 2015 Jun 19;32(2):221-229. [doi: 10.17843/rpmesp.2015.322.1611]
27. Van Olmen J, Van Pelt M, Malombo B, Ku GM, Kanda D, Heang H, et al. Process evaluation of a mobile health intervention for people with diabetes in low income countries - the implementation of the TEXT4DSM study. J Telemed Telecare 2017 Jan;23(1):96-105. [doi: 10.1177/1357633X15617885] [Medline: 26656894]
28. Muntinga ME, Van Leeuwen KM, Schellevis FG, Nijpels G, Jansen AP. From concept to content: assessing the implementation fidelity of a chronic care model for frail, older people who live at home. BMC Health Serv Res 2015 Jan 22;15:18 [FREE Full text] [doi: 10.1186/s12913-014-0662-6] [Medline: 25608876]
29. van der Laan DM, Langendoen-Gort M, Nijpels G, Boons CCLM, Elders PJM, Hugtenburg JG. Implementation fidelity of an intervention programme to enhance adherence to antihypertensive medication in Dutch community pharmacies. Int J Clin Pharm 2019 Aug;41(4):1031-1046 [FREE Full text] [doi: 10.1007/s11096-019-00845-z] [Medline: 31093942]
30. Beratarrechea A, Diez-Canseco F, Irazola V, Miranda J, Ramirez-Zea M, Rubinstein A. Use of m-Health technology for preventive interventions to tackle cardiometabolic conditions and other non-communicable diseases in Latin America-challenges and opportunities. Prog Cardiovasc Dis 2016;58(6):661-673. [doi: 10.1016/j.pcad.2016.03.003] [Medline: 27041078]
31. Middleton T, Constantino M, McGill M, D'Souza M, Twigg SM, Wu T, et al. An enhanced SMS text message-based support and reminder program for young adults with type 2 diabetes (TEXT2U): randomized controlled trial. J Med Internet Res 2021 Oct 21;23(10):e27263 [FREE Full text] [doi: 10.2196/27263] [Medline: 34524102]
32. Nelson LA, Roddy MK, Bergner EM, Gonzalez J, Gentry C, LeStourgeon LM, et al. Exploring determinants and strategies for implementing self-management support text messaging interventions in safety net clinics. J Clin Transl Sci 2022;6(1):e126 [FREE Full text] [doi: 10.1017/cts.2022.503] [Medline: 36590364]
33. Steinman L, van Pelt M, Hen H, Chhorvann C, Lan CS, Te V, et al. Can mHealth and eHealth improve management of diabetes and hypertension in a hard-to-reach population? -lessons learned from a process evaluation of digital health to support a peer educator model in Cambodia using the RE-AIM framework. mHealth 2020;6:40 [FREE Full text] [doi: 10.21037/mhealth-19-249] [Medline: 33437836]

XSL•FO
RenderX

34. Yang S, Cha MJ, van Kessel R, Warrier G, Thrul J, Lee M, et al. Understanding inequalities in mobile health utilization across phases: systematic review and meta-analysis. J Med Internet Res 2025 Aug 14;27:e71349 [FREE Full text] [doi: 10.2196/71349] [Medline: 40811740]
35. Bommakanti KK, Smith LL, Liu L, Do D, Cuevas-Mota J, Collins K, et al. Requiring smartphone ownership for mHealth interventions: who could be left out? BMC Public Health 2020 Jan 20;20(1):81 [FREE Full text] [doi: 10.1186/s12889-019-7892-9] [Medline: 31959145]
36. Sapre M, Elaiho CR, Brar Prayaga R, Prayaga R, Constable J, Vangeepuram N. The development of a text messaging platform to enhance a youth diabetes prevention program: observational process study. JMIR Form Res 2024 May 29;8:e45561 [FREE Full text] [doi: 10.2196/45561] [Medline: 38809599]
37. Yee L, Taylor S, Young M, Williams M, Niznik C, Simon M. Evaluation of a text messaging intervention to support self-management of diabetes during pregnancy among low-income, minority women: qualitative study. JMIR Diabetes 2020 Aug 10;5(3):e17794 [FREE Full text] [doi: 10.2196/17794] [Medline: 32773367]
38. Saade GR. Pregnancy as a window to future health. Obstet Gynecol 2009 Nov;114(5):958-960. [doi: 10.1097/AOG.0b013e3181bf5588] [Medline: 20168094]
39. Huberty JL, Buman MP, Leiferman JA, Bushar J, Hekler EB, Adams MA. Dose and timing of text messages for increasing physical activity among pregnant women: a randomized controlled trial. Transl Behav Med 2017 Jun;7(2):212-223 [FREE Full text] [doi: 10.1007/s13142-016-0445-1] [Medline: 27800565]
40. Pérez-Lu JE, Bayer A, Iguiñiz-Romero R. Information = equity? How increased access to information can enhance equity and improve health outcomes for pregnant women in Peru. J Public Health (Oxf) 2018 Dec 01;40(suppl_2):ii64-ii73 [FREE Full text] [doi: 10.1093/pubmed/fdy177] [Medline: 30307537]
41. Fry J, Wilkinson SA, Willcox J, Henny M, McGuire L, Guthrie TM, et al. Improving engagement in antenatal health behavior programs-experiences of women who did not attend a healthy lifestyle telephone coaching program. Nutrients 2023;15(8):1860 [FREE Full text] [doi: 10.3390/nu15081860] [Medline: 37111079]
42. Nabovati E, Farzandipour M, Vahedpoor Z, Akbari H, Anvari S, Sharif R, et al. Pregnant women's use and attitude toward mobile phone features for self-management. BMC Med Inform Decis Mak 2023 Apr 26;23(1):77 [FREE Full text] [doi: 10.1186/s12911-023-02172-w] [Medline: 37101302]
43. Stowell M, Dobson R, Garner K, Baig M, Nehren N, Whittaker R. Digital interventions for self-management of prediabetes: a scoping review. PLoS One 2024;19(5):e0303074 [FREE Full text] [doi: 10.1371/journal.pone.0303074] [Medline: 38728296]
44. Arora S, Lam CN, Burner E, Menchine M. Implementation and evaluation of an automated text message-based diabetes prevention program for adults with pre-diabetes. J Diabetes Sci Technol 2024 Sep;18(5):1139-1145. [doi: 10.1177/19322968231162601] [Medline: 36946537]
45. Hengst TM, Lechner L, Dohmen D, Bolman CA. The facilitators and barriers of mHealth adoption and use among people with a low socio-economic position: a scoping review. Digit Health 2023;9:20552076231198702 [FREE Full text] [doi: 10.1177/20552076231198702] [Medline: 37691766]
46. Boksmati N, Butler-Henderson K, Anderson K, Sahama T. The effectiveness of SMS reminders on appointment attendance: a meta-analysis. J Med Syst 2016 Apr;40(4):90. [doi: 10.1007/s10916-016-0452-2] [Medline: 26852337]

Abbreviations

API: application programming interface
CHW: community health worker
FINDRISC: Finnish Diabetes Risk Score
PCC: primary care center
REDCap: Research Electronic Data Capture
T2D: type 2 diabetes

Edited by A Stone, A Mavragani; submitted 22.07.25; peer-reviewed by MS Mekbib, CL Sánchez Bocanegra, C Baxter; comments to author 29.08.25; revised version received 29.09.25; accepted 01.10.25; published 01.12.25.

Please cite as:
Nejamis A, Gutierrez L, López MV, Ciganda Cha AD, Mendoza Quispe O, Irazola V, Beratarrechea A
WhatsApp-Based Intervention for Diabetes Prevention and Care in Argentina: Implementation and Process Evaluation
JMIR Form Res 2025;9:e81098
URL: https://formative.jmir.org/2025/1/e81098
doi:10.2196/81098
PMID:

XSL•FO
RenderX

©Analia Nejamis, Laura Gutierrez, María Victoria López, Alvaro Daniel Ciganda Cha, Omayra Mendoza Quispe, Vilma Irazola, Andrea Beratarrechea. Originally published in JMIR Formative Research (https://formative.jmir.org), 01.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Personalizing a Weight Loss Program Using Cognitive-Behavioral Phenotypes to Improve Engagement and Weight Loss in Adults With Overweight or Obesity: Quasi-Experimental Study

Joanna Szypula[1], PhD; Andreas Jarvstad[1], PhD; Lucy Amelia Jones[2], RD; Katy Tapper[1], PhD

[1]School of Health and Psychological Science, City St George's, University of London, London, United Kingdom
[2]Oviva, London, United Kingdom

Corresponding Author:
Joanna Szypula, PhD
School of Health and Psychological Science
City St George's, University of London
Rhind Building
St John St
London, EC1R0JD
United Kingdom
Phone: 20 70405060
Email: joanna.szypula@city.ac.uk

Abstract

Background: Obesity affects more than one-quarter of adults in the United Kingdom and is a leading cause of preventable disease and health care costs. Digital behavior change programs can provide scalable weight management support, but maintaining engagement is challenging, and engagement is strongly associated with weight loss success. Tailoring interventions to cognitive-behavioral phenotypes, distinct patterns of thinking and behavior, offers one strategy to improve adherence. Although such approaches show promise in controlled settings, evidence from real-world digital programs is limited.

Objective: This study evaluated whether phenotype-tailored weekly advice improved engagement and weight loss in a national digital weight management program. Secondary aims were to assess correlations between advice use and outcomes, explore moderation by socioeconomic status, and capture participants' perceptions of the advice.

Methods: We conducted a quasi-experimental study among UK adults enrolled in a free 12-week program commissioned by the National Health Service. Eligible participants were aged 18-80 years with a BMI greater than 25 kg/m². The phenotype group (n=148; mean age 48 years; 127/148, 86% female; mean BMI 39 kg/m²) completed a 17-item questionnaire, were classified into one of 4 phenotypes, and received weekly tailored advice for 7 weeks. Comparators included a historical cohort enrolled 1 year earlier without phenotype advice (n=241; mean age 44 years; 171/241, 71% female) and nonresponders who did not complete the questionnaire (n=394; mean age 44 years; 299/394, 76% female). Primary outcomes were program engagement (any in-app activity such as meal logging, activity tracking, content reading, or coach messaging) and self-reported weight.

Results: The phenotype group recorded a mean of 257 (SD 232) engagements over 7 weeks, significantly higher than both the historical cohort (mean 159, SD 187; $P<.001$) and nonresponders (mean 135, SD 198; $P<.001$), representing 62%-90% greater activity. All engagement types were significantly elevated ($P<.001$ for all). Mean weight loss was −2.23 kg (SD 7.97) in the phenotype group, compared with −1.60 kg (SD 5.39; $P=.29$) in the historical cohort and −0.69 kg (SD 13.23; $P=.23$) in nonresponders. The number of phenotype-specific advice documents opened correlated with engagement ($r=0.48$; $P<.001$) but not with weight loss ($P=.42$). Socioeconomic status did not moderate outcomes. Posttrial interviews (n=16) provided mixed feedback: many participants described the advice as clear, relevant, and motivating, whereas others considered it too general or poorly matched.

Conclusions: Phenotype-tailored weekly advice was associated with substantially higher engagement in a real-world digital program, although short-term weight differences were not statistically significant. While limited by a nonrandomized design, short follow-up, and reliance on self-reported weight, this study suggests phenotype-based tailoring may be a scalable strategy to strengthen adherence in digital weight loss interventions. Larger randomized trials with longer follow-up are warranted to determine whether increased engagement translates into clinically meaningful weight loss.

XSL•FO
RenderX

(JMIR Form Res 2025;9:e72645) doi:10.2196/72645

KEYWORDS

digital weight management program; health behavior; mobile applications; obesity; patient engagement; real-world evidence; weight loss intervention

Introduction

Global obesity rates have doubled since 1990 [1], with approximately 1 in 4 adults in the United Kingdom now living with the condition [2]. Although early interventions focused on telling individuals to "eat less and move more," research over the past 2 decades has highlighted obesity as a complex, multifactorial issue requiring multicomponent, evidence-based treatment [3].

Many weight loss interventions are generic, offering identical advice to all participants [4,5]. This may partly explain the wide variability in individual responses, even to intensive treatments such as bariatric surgery [6]. Recent efforts have focused on personalizing interventions, often through genetic, anthropometric, or microbiota-based approaches [7-9]. There is growing evidence that personalized nutrition and exercise advice can enhance outcomes [8-10], but comparatively little attention has been paid to tailoring weight loss interventions based on cognitive-behavioral characteristics. Some research has used tools such as the Three-Factor Eating Questionnaire, Dutch Eating Behavior Questionnaire, or 6-Factor Questionnaire to generate behavioral profiles [11-13], but the effects of tailoring advice based on these profiles remain largely untested.

Digital, app-based weight loss interventions are increasingly popular because of their scalability, accessibility, and low cost [14]. They can be as effective as, or more effective than, in-person programs [15-17] and allow for large-scale, real-world data collection to assess adherence and outcomes [18]. Individual differences, such as personality, may influence in-app engagement and outcomes in digital weight management programs [19]. Tailoring interventions has been shown to increase engagement, which may lead to better outcomes [20-22], possibly because personalized advice feels more relevant and persuasive [23-26]. Improving engagement is particularly important in digital settings, where high attrition can limit effectiveness [26].

In-app engagement is commonly used as a proxy for adherence, with more frequent interaction indicating greater engagement [20]. A recent analysis by Lehmann et al [18] found a significant relationship between Oviva (Oviva AG) app engagement and weight loss at 3 months across all countries analyzed but not specifically within the United Kingdom. This nonsignificant relationship in the United Kingdom may be explained by the fact that these data came from more complex patients living with obesity, who often face greater challenges, such as previous unsuccessful weight loss attempts, reliance on prescribed treatments (eg, glucagonlike peptide-1 receptor agonists or total dietary replacement), higher emotional eating, loss of control, significant comorbidities, and psychological barriers [27,28]. These factors could account for the weaker association between app engagement and weight loss in this group. Nevertheless, the broader international findings support the interpretation that app engagement reflects meaningful behavioral activity. App features prompt users to engage in evidence-based techniques such as self-monitoring, reflection, and problem-solving, all known to facilitate weight loss [29]. Therefore, increases in app engagement are likely to reflect active participation in self-management strategies, which may contribute to improved weight outcomes.

For this study, we collaborated with Oviva, a provider of digital weight loss services commissioned by the United Kingdom's National Health Service (NHS). In a previous study, we developed a 17-item questionnaire to assign users to 1 of 4 cognitive-behavioral phenotypes [30]. In this study, we aimed to test whether receiving phenotype-tailored advice would improve app engagement and weight loss in a real-world setting. Using a quasi-experimental design, patients enrolled in the Oviva program were invited by email to complete a "quiz" that provided personalized advice. Those who responded received weekly, tailored emails for 7 weeks. Two comparison groups were included: a historical cohort of patients enrolled a year earlier and a group of nonresponders who received the invitation but did not click the "quiz" link.

As socioeconomic status is known to be associated with obesity prevalence [31-35], but its role in digital program outcomes remains unclear, we additionally explored this in this study. Some evidence suggests that eHealth interventions can support weight loss in low-income adults [36], whereas others report no effect [37-39]. However, many studies may be underpowered to detect interactions or may not recruit representative samples from lower socioeconomic groups [40-44].

We hypothesized that patients receiving phenotype-tailored advice would show a greater number of in-app engagements and would lose more weight than the comparison groups. We also predicted that these effects would be moderated by socioeconomic status (using the Index of Multiple Deprivation [IMD] as a proxy), with less benefit among more deprived participants. Finally, we expected greater app engagement and weight loss among those who accessed more of their phenotype-matched content.

Methods

Ethical Considerations

Ethical approval for this study was granted by the City, University of London Psychology Department Research Ethics Committee (ETH2324-1944; amendments ETH2324-2135 and ETH2324-2449). Participants did not sign a separate consent form because patients enrolling in the Oviva program provide broad consent to be included in trials and for their anonymized data to be used for research, including secondary analyses. Eligible patients received an email inviting them to complete the phenotype questionnaire, with a clear notice that accessing

XSL•FO
RenderX

the survey would enroll them in a 7-week automated email sequence. Participation was voluntary, and recipients could unsubscribe at any time. Historical comparison data were fully anonymized and covered under the original consent. Email addresses were used only to match individuals to anonymized patient user IDs, after which all identifiers were replaced. App engagement data were analyzed in anonymized form. Only authorized internal teams accessed identifiable data for email delivery and reporting, and no personal information was included in the research data set. Participants were not financially compensated but received weekly educational materials as part of the intervention.

Participants

Participants were recruited from Oviva, an NHS-contracted weight loss service provider that is free for patients. They were sent an invitation email if they had started their weight management program with Oviva within the 3-week recruitment period in May 2024. The weight management program used for recruitment is a 12-week digital behavioral and lifestyle intervention offered to NHS-referred patients. It includes support from specialist weight management dietitians and health coaches. Participants gain access to the Oviva app, a class IIa medical device that features lessons on weight management and healthy lifestyle, goal setting with reminders, behavior tracking (eg, food, fluid intake, and physical activity), and communication with their coach to monitor progress. Any activity on the app, such as logging, messaging the coach, or reading a lesson, was recorded as an "engagement." As this was a pragmatic, quasi-experimental study using convenience sampling over a fixed 3-week recruitment period, enrollment was determined by natural sign-up rates.

Program Eligibility

To join the weight management program at Oviva, patients were required to meet the following eligibility criteria: be aged 18-80 years; reside in, or be registered with, a general practitioner practice in England; have a BMI ≥30 kg/m² (or ≥27.5 kg/m² for individuals of Black African, African-Caribbean, or Asian origin); have a diagnosis of diabetes or hypertension; own a smartphone; and be able to speak English. Patients were ineligible if they had severe or moderate frailty as recorded on a frailty register, were pregnant, had been diagnosed with an eating disorder, had a significant unmanaged comorbidity (eg, chronic obstructive pulmonary disease not on medication), or had undergone bariatric surgery in the past 2 years.

Design

This study used a quasi-experimental design with posttest nonequivalent comparison groups. All Oviva weight management patients who started their program within the recruitment window were invited to complete the phenotypes quiz. Those who clicked the link and completed the questionnaire were assigned to the phenotypes group. Two comparison groups were included: (1) patients who enrolled in the same program 12 months before recruitment (historical cohort) and (2) patients who were invited to complete the questionnaire but did not click the link (nonresponders). The historical cohort was randomly selected by the Oviva analytics team, and the researchers had no access to, or influence over, which patients were included. The historical cohort served as a comparable reference group, as no substantial modifications had been made to the app or program structure since their participation. Although minor updates and iterative improvements occurred, these were not expected to meaningfully impact engagement or outcomes.

The dependent variables were weekly app engagements and weight loss over a 7-week period. Any of the following activities counted as an app engagement: tracking an event (eg, meal, fluid, waist circumference, mood), sending a chat message to a coach, opening a lesson page (each page viewed increased the engagement count), or opening a weekly article.

Materials and Measures

Demographics

Participant demographics were provided by Oviva's analytics team, typically sourced from general practitioner practices, with weights self-reported. Socioeconomic status was indicated by IMD decile, a UK measure that estimates poverty based on geographical location and considers factors such as income, employment, crime, and education (higher IMD scores indicate lower deprivation [45]).

Phenotypes Questionnaire

The phenotypes questionnaire, or "quiz," is a 17-item questionnaire developed in our earlier study to match patients to 1 of 4 profiles [30]. Briefly, we identified 5 cognitive and behavioral constructs that (1) could affect weight or weight loss efforts, (2) have an existing validated measure, and (3) are potentially modifiable. The 5 constructs were maladaptive eating, hedonic overeating, reward reactivity, self-regulation, and psychological avoidance (Multimedia Appendix 1 contains the full questionnaire). For further methodological details, please consult the development paper (currently under review) [46].

The quiz was created in Qualtrics (Qualtrics International Inc), with each item rated on a scale from 1 (strongly disagree) to 5 (strongly agree). Participants were matched to a phenotype by calculating a percentage fit for each profile and assigning them to the one with the highest fit. They were shown their profile match along with a short, 3- to 4-bullet point description of their phenotype. Participants could either continue with their matched phenotype or review all profile descriptions and choose the one that resonated most with them. If 2 or more fit scores were equal, participants were shown all profile descriptions and asked to select one.

The Phenotypes

There were 4 phenotypes a participant could be matched with. The profiles were named after colors to keep the assignment emotionally neutral, and each was paired with a spice name to create a distinctive cue for participants, helping them remember their profile name over the following weeks. "Red Chilli" was characterized by high maladaptive and hedonic eating, low self-regulation, and high psychological avoidance. "Yellow Saffron" was based on high hedonic eating and reward reactivity and low maladaptive eating. "Purple Lavender" represented low self-regulation and high psychological avoidance, as well as

XSL•FO
RenderX

low maladaptive and hedonic eating. Lastly, "Green Sage" reflected low maladaptive and hedonic eating, high self-regulation, and low psychological avoidance.

The 4 cognitive-behavioral phenotypes are intended as practical, nonexclusive groupings that highlight dominant behavioral tendencies rather than rigid or mutually exclusive categories. Given the complexity of eating behavior, some overlap in traits and relevance of advice is expected. The profiles aim to guide individuals in identifying and addressing key obstacles to weight loss while acknowledging that multiple factors may coexist.

Phenotype Content

Participants received weekly emails with advice tailored to their phenotype. The first document provided an overview and initial content, followed by weekly strategies addressing specific challenges (eg, overeating). Each document provided evidence-based advice and a short exercise to help participants put this knowledge into practice.

The Red Chilli profile focused on reducing restriction and maladaptive eating using acceptance and commitment therapy, mindfulness, cognitive defusion, and self-compassion. The Yellow Saffron profile addressed hedonic overeating with smaller portions of enjoyable foods, sensory eating techniques, craving management, and environmental architecture for healthier choices. The Purple Lavender profile improved goal setting and adherence by engaging with health goals daily, starting with small behaviors, and using routines, cognitive architecture, and defusion for challenging thoughts. Finally, the Green Sage profile enhanced diet quality and physical activity by adding vegetables, choosing foods with longer oral processing times, applying flexible eating (80% nutritious, 20% less nutritious), and cognitive architecture for activity goals.

Posttrial Feedback

After receiving all content emails, participants were sent a final email requesting feedback through a Qualtrics questionnaire, which included 11 Likert scale questions (1=strongly disagree to 5=strongly agree) assessing the helpfulness, effectiveness, and clarity of the advice. Participants were also given 4 optional open-ended questions to comment on the quiz, useful aspects, and suggestions for improving their profile or the advice.

Procedure

Participants were sent an email inviting them to take part in the phenotypes quiz. Those who completed the quiz were automatically matched to a profile and asked to read through its description, based on which they could decide to accept their match or change it. Once their response was submitted, they received an automatic email reminding them which profile they were matched with and were added to a 7-week long campaign.

Participants received the first email with tailored advice 1 day after they took the quiz. This email contained links to 4 documents (1 for each phenotype), with participants instructed to open the one matching their profile. Six more weekly emails followed, each containing 4 profile-specific documents. Thus, throughout the trial, participants received links to 28 documents but were instructed to only open 7 of them. The final email contained a link to the posttrial feedback questionnaire. The core program and care provided by dietitians and health coaches remained unchanged.

Although the tier 2 program lasts 12 weeks, the phenotype-specific advice was delivered over a 7-week period due to project timelines, which required all data collection and analysis to be completed within a fixed timeframe.

Analysis Plan

The main analyses pertained to the effect of receiving phenotype-related advice on the average number of app engagements and weight loss over a 7-week period. Hierarchical regressions were conducted to compare the average number of app engagements between the phenotype group, and the historical cohort group, and the nonresponder group. We tested whether IMD moderated this relationship by including it as step 2 of the regression and the interaction term as step 3. The same approach was used to analyze the effect of condition on weight loss.

We further carried out equivalent Bayesian analyses using STAN (Stan Development Team) using *stan_glm* in rstanarm version 2.32.1 (Stan Development Team), brms version 2.22.0 [47], R version 4.4.1 (R Foundation for Statistical Computing), and RStudio 2024.12 (Posit Software, PBC). We fit 4 models for each dependent variable (DV): (1) DV ~ 1 (intercept only), (2) DV ~ group, (3) DV ~ group + IMD, and (4) DV ~ group * IMD. The engagement variable was square root–transformed, and the IMD variable was standardized prior to fitting. The weight loss variable contained outliers and a preponderance of zeros. To maintain parity with the standard regression analyses, we did not further process these data in any way. We used weakly informative robust *t*-distribution priors (*df*=3; mean 0, SD 10), and a Gaussian link function. We ran 8 chains, each with 1000 discarded burn-in samples and 5000 retained samples, for a total of 20,000 retained samples per model. We verified model convergence using standard criteria (eg, $\hat{R}$ ~ 1) and visually inspected chains for convergence. For each model, we computed the log marginal likelihood and computed Bayes Factors (BFs) first for target models against the null intercept-only model. For any models which provided BF in favor of the target models over the intercept-only model (ie, BF>3), we additionally computed pairwise BFs to establish which of the surviving models best explain the data.

Change in the average number of engagements on a week-by-week basis was analyzed using a repeated-measures ANOVA.

Qualitative feedback from open-ended survey responses and semistructured interviews was analyzed using a descriptive thematic approach. Responses were reviewed inductively to identify recurring sentiments and contrasting views across participants. Illustrative quotes were selected to reflect the diversity of opinions, and analysis focused on summarizing key perceived benefits, limitations, and personal relevance of the intervention.

The study protocol was uploaded to the Open Science Framework before data analysis, and the analysis plan was preregistered before obtaining engagement and weight loss data

XSL•FO
RenderX

from Oviva [48]. Data were analyzed using SPSS (version 28; IBM Corp).

Results

Participant Characteristics and Study Flow

Figure 1 shows the flow of participants through the study. Participant demographics were reasonably well matched across the 3 groups (Table 1), although participants in the phenotype group were slightly older, more likely to be female, and had a higher average IMD score than those in the comparison groups.

Figure 1. Flow of participants in a quasi-experimental study evaluating the effect of receiving phenotype-matched weight loss advice on app engagement and weight loss among patients actively participating in a UK weight loss program, recruited in May 2024.

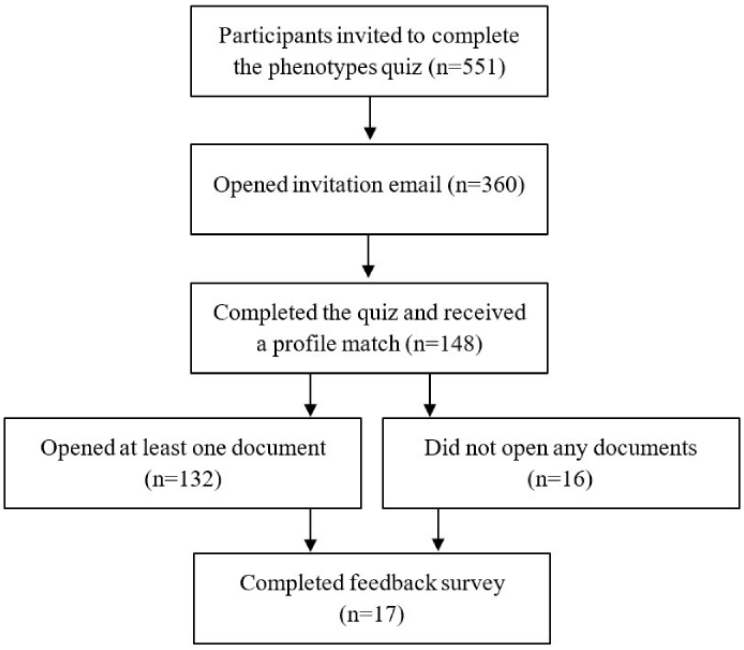

Table 1. Demographic characteristics of participants in the phenotype group and two comparison groups (historical cohort and nonresponders). Scores range from 1-10; lower values indicate higher relative deprivation.

Characteristic	Phenotypes group (n=148)	Historical cohort (n=241)	Nonresponders (n=394)
Age (years), mean (SD)	48 (12)	44 (10)	44 (11)
Female sex, n (%)	127 (86)	171 (71)	394 (76)
Starting weight (kg), mean (SD)	108 (19.7)	111 (24.1)	109 (25.0)
Starting BMI (kg/m^2), mean (SD)	39 (6)	39 (8)	39 (8)
Ethnicity, n (%)			
White	127 (86)	166 (69)	280 (71)
Black	8 (5)	39 (16)	41 (10)
Asian	7 (5)	24 (10)	33 (8)
Mixed	3 (2)	2 (1)	19 (5)
Other	2 (1)	7 (3)	12 (3)
Refused response	0 (0)	2 (1)	6 (2)
Missing data	1 (1)	1 (1)	3 (1)
Index of Multiple Deprivation (IMD)			
Mean (SD)	6 (3)	5 (3)	5 (3)
Median (IQR)	6 (8-3)	3 (7-1)	5 (7-3)

XSL•FO
RenderX

Phenotype Assignment

Figure 2 shows how phenotype-group participants were assigned to their final profile. A total of 18 patients who received an invitation to the quiz clicked on the link but did not complete the questionnaire; therefore, they received the weekly emails but were unaware which links they should click on. Data from these participants were excluded from all analyses.

Figure 2. Flow of participants in the phenotype group from starting the phenotype quiz to receiving a final profile match, which determined the tailored weight loss advice they received during the trial.

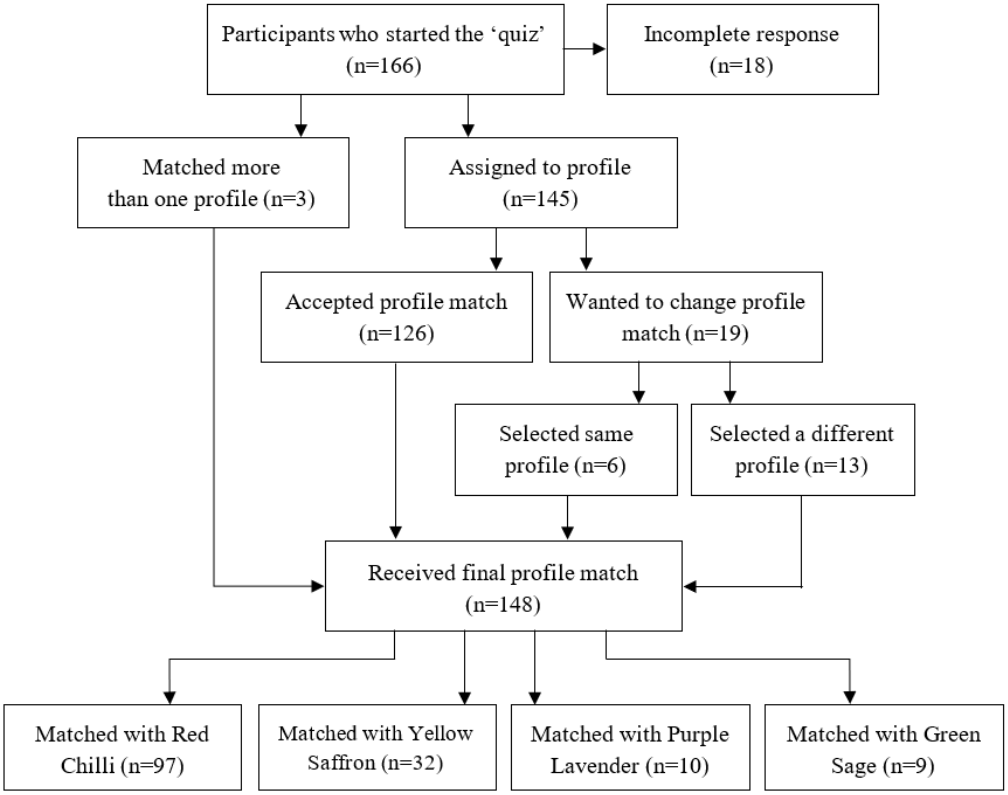

Phenotype Fidelity

Participants were asked to open only documents for the profile they were matched with (eg, those matched with Red Chilli should open Red Chilli documents in weeks 1-7). However, since documents for all profiles were sent each week, it was possible for participants to open advice for multiple profiles. Fidelity was defined as opening at least 1 document tailored to the profile the participant was matched with and not opening more than 1 document from a different profile. The fidelity rate was 73.6% (109/148) throughout the trial. Among the remaining participants, 4.1% (6/148) followed advice for a different profile (opened 3 or more documents for a different profile and no more than 1 document for their original profile), 10.8% (16/148) opened no documents, and 11.5% (17/148) showed a different click pattern.

Main Analyses

Effect of Condition on Total Engagements

The average total number of engagements over a 7-week period was higher in the phenotype group than in the historical cohort (Table 2). A hierarchical linear regression was conducted to compare the average number of engagements between the two groups, and to examine whether this effect was moderated by IMD. Condition was entered at step 1, IMD at step 2, and the interaction term at step 3. The analysis revealed that participants in the phenotype group generated significantly more engagements than those in the historical cohort (b=−93.61, SE B=22.39, β=−0.22; P<.001), but the interaction was not significant (b=−4.44, SE B=7.54, β=−0.12; P=.56). An equivalent Bayesian analysis confirmed that a model including only an intercept term (B=11.91, 95% CI 11.05-12.74) and the main effect of phenotype group (B=1.89, 95% CI 1.05-2.74) was superior to the model that also included a main effect of IMD (BF=144) as well as the interaction model (BF=1429). All the models were superior to the baseline intercept-only model (all BFs>14). Independent t tests (2-tailed) showed that all engagement subtypes (tracking activity, coach messaging, pages of educational modules viewed, and articles read) were significantly higher in the phenotype group than in the historical cohort group, suggesting that the overall increase in engagement was not driven by any single type of app interaction (P<.001 for all; Multimedia Appendix 1 contains all analyses).

XSL•FO
RenderX

Table 2. Mean number of app engagements and mean weight loss in the phenotype group and two comparison groups (historical cohort and nonresponders).

Group	Total app engagement count, mean (SD)	Weight loss (kg), mean (SD)
Phenotypes group (n=148)	257 (232)	2.23 (7.97)
Historical cohort (n=241)	159 (187)	1.60 (5.39)
Nonresponders (n=394)	135 (198)	0.69 (13.23)

Effect of Condition on Weight Loss

Average weight loss over a 7-week period was compared between the phenotype group and the historical cohort (Table 2). Another hierarchical regression was conducted to test the effect of condition on weight loss and whether IMD moderated this relationship. Condition was entered at step 1, IMD at step 2, and the interaction term at step 3. No significant differences in weight loss were found between the two groups (b=0.85, SE B=0.80, β=0.06; *P*=.29), and the interaction was not significant (b=−0.26, SE B=0.27, β=−0.22; *P*=.33). The equivalent Bayesian analysis showed that a baseline intercept model was substantially better than all the main effect and interaction models (all BFs≥567), indicating that neither group, IMD, nor their interaction explained variation in weight loss.

Relationship Between Documents Opened and Primary Outcomes

Among the phenotype cohort there was a significant positive correlation between the number of correct documents opened (ie, documents that matched the phenotype the participant was assigned to) and total engagements (r_{146}=.481; *P*<.001). There was no relationship between the number of correct documents opened and weight loss (r_{119}=−.074; *P*=.42).

Comparison With Nonresponders

An independent-groups *t* test (2-tailed) revealed participants in the phenotypes group generated significantly more engagements over the 7-week period than nonresponders (equal variances not assumed; Levene Test $F_{1,540}$=12.98; *P*<.001; $t_{232.25}$=5.68; *P*<.001; Table 2). Weight loss over the 7-week period was higher among the phenotype group compared with nonresponders, but this difference did not reach statistical significance (t_{445}=−1.21; *P*=.23; Table 2).

BMI and Body Mass Percentage Reduction

On average, participants in the phenotype group reduced their BMI by 0.79 kg/m^2 (SD 2.90), compared with a reduction of 0.56 kg/m^2 (SD 1.94) in the historical cohort over the 7-week period. This translated into a reduction of 1.93% (SD 7.24) in body mass in the phenotype group and a 1.27% (SD 4.14) reduction in body mass in the historical cohort. Nonresponders reduced their BMI by 0.11 kg/m^2 (SD 5.01) and their body mass by 0.09% (SD 13.01).

Exploratory Analyses

A repeated-measures ANOVA was conducted to assess week-by-week change in app engagements and the interactions between the phenotype group, the historical cohort, and the nonresponders. As Mauchly test of sphericity was significant (χ^2_{20}=715.1; *P*<.001; Mauchly W=0.40; Greenhouse-Geisser correction ε=.720), the Greenhouse-Geisser correction was applied. There was a significant effect of week number on app engagement ($F_{4.32,3367.68}$=81.85; *P*<.001), and the interaction with condition was significant ($F_{8.64,3367.68=}$5.57; *P*<.001; Figure 3).

XSL•FO
RenderX

Figure 3. Mean weekly app engagements in the phenotype group (n=148) and the two comparison groups (historical cohort, n=241; nonresponders, n=394). Error bars denote 95% CIs.

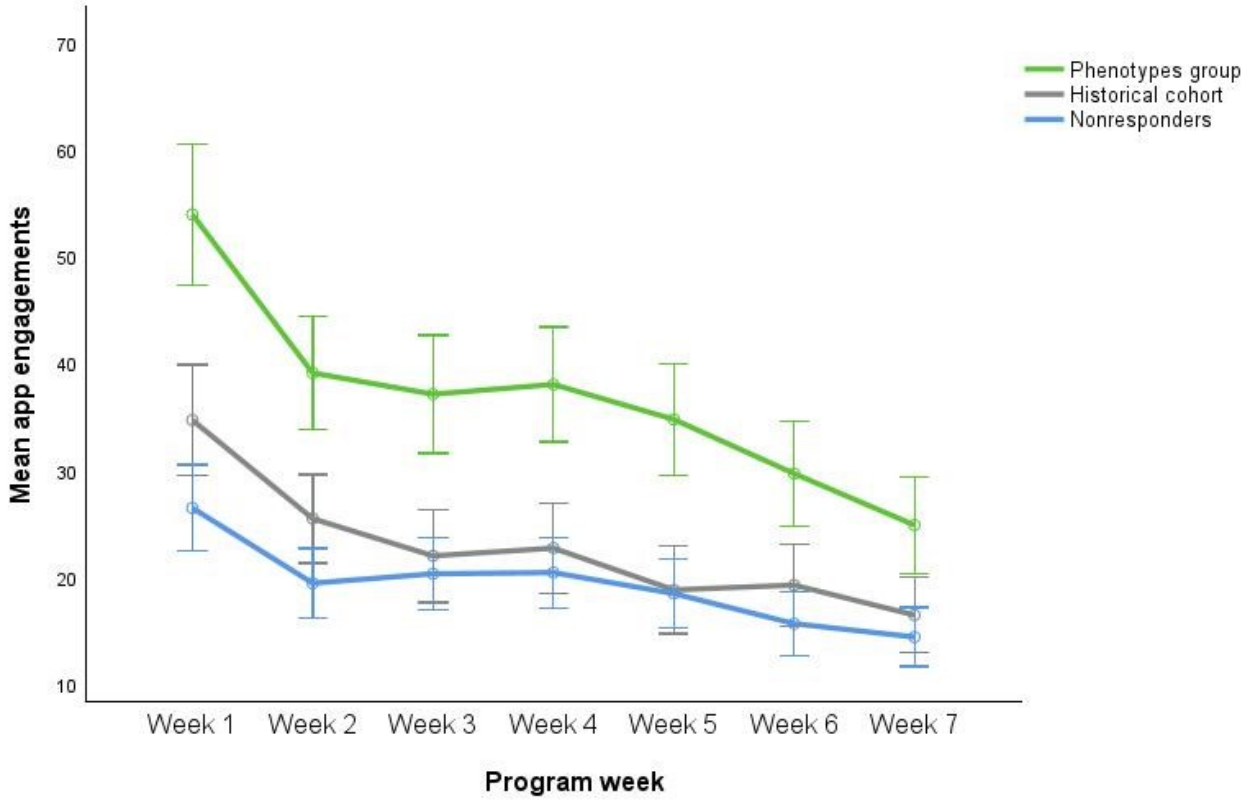

Posttrial Ratings

Posttrial feedback was provided by 17 participants (12.2%); however, 1 participant did not open any documents with phenotype-tailored advice, and their feedback was therefore excluded from the analysis. Overall, feedback on the phenotype-tailored advice was positive (Table 3), as participants felt the advice they received was helpful and easy to understand. Participants also indicated that they were willing to use the advice in the future.

Table 3. Posttrial participant ratings of phenotype-tailored advice (n=16), including perceived helpfulness, clarity, personalization, and intention to use the advice, scored on a 1-5 scale, with higher ratings indicating greater endorsement.

Statement	Rating (SD)
The advice I received was helpful	3.75 (1.13)
The advice helped me make positive changes to my lifestyle	3.63 (1.26)
The advice helped me to lose weight	3.25 (1.24)
The length of the advice was just right (not too short and not too long)	3.50 (1.37)
The advice was clear and easy to understand	4.06 (1.00)
I learned something new by reading the advice in the resources	3.44 (1.31)
The advice I received felt personalized to me	3.13 (1.36)
The advice I received felt relevant to me	3.44 (1.31)
I followed the advice given to me in the resources	3.88 (0.72)
I completed the tasks in the 'Try it out' section in each of the resources	3.69 (1.01)
I will use the advice from the resources in the future	3.81 (0.98)

Qualitative Feedback

Qualitative feedback was collected in two ways: through the posttrial feedback questionnaire and posttrial interviews. The feedback was mixed, with some people feeling very positive about the phenotypes and others feeling they were not as useful. One participant said the advice “was good” (Participant 9), and found it especially helpful to learn about “the reasons behind my eating when not hungry and the food I choose.” Another participant said:

> *all advice was spot on (…) I looked forward to reading them and I will miss them. More would be*

XSL•FO
RenderX

> *great (…) I found this the most helpful even more than the learning modules (…) [understanding] how my mind works in relation to food certainly explained a lot* [Participant 10]

Two more participants wrote, “Excellent advice & profiles” (Participant 11) and “I found the information encouraging to read” (Participant 14). Two participants felt more neutral about their experience, as one said, “it was all useful, except I haven’t lost very much weight” (Participant 13), and another said, “It was ok. To be honest I can’t recall much about it” (Participant 12). Others felt less positively about the phenotypes, saying the “quiz was not really helpful, advice was very vague and generalized (…) no specific tailored help” (Participant 2). Another participant said, “Too much emphasis on how I feel. I hate being constantly asked to think about how I feel. (…) I didn't really fit into any of the categories anyway” (Participant 4). Others said, “it just feels generic and not very personal” (Participant 7) and “[the profile] didn’t reflect me” (Participant 8).

Everyone who completed the posttrial survey was invited to participate in an interview, and 5 participants accepted. Participant 9 felt positively about the quiz and the Red Chilli profile she was matched with. She said the profile “was so accurate that it resonated” and “when I read the profile I was like (…) I know I'm an emotional eater (…) but actually, until I read that, it didn't make sense.” She talked about how the profile:

> *threw up things that you weren't aware of (…) you know what you should be doing. Why can't you do it? And I think that was the big question I really wanted to know the answer to (…) and then finding out that it's probably this and this (…) [that] was the most helpful thing.*

Participant 10, who was also matched with the Red Chilli profile, said, “honestly, every week, I've looked forward to reading them [the advice]. And I'm just sorry they've stopped now.” She talked about how she used the learning from the profile to think:

> *How do I feel? Am I being, you know, not very kind to myself? (…) it put things into perspective for me rather than just saying I've ate this, I've drank that.*

She mentioned how the advice was tailored to her needs, saying:

> *I think I've lost well over a stone (…) I don't have a bad diet, I don't think I just choose fast, convenient food when I'm stressed (…) I've just put more thought into it and the thought that not being cruel to yourself, being actually like, oh, OK, you can do this.*

Moreover, she said:

> *I found the language to be absolutely, you know, spot on for me that it had got me down to a T (…) I'm hoping that it [the advice] will make me stop and think (…) to actually keep the weight off because that's the second challenge, isn't it?*

Participant 6 was matched with the Yellow Saffron profile. She said:

> *quite often, diet or health things, they don't really allow for people who really love food in a foodie way (…) [the quiz] recognized that I was that sort of person (…) the advice was tailored to me (…) which actually I was really pleased with.*

She also said:

> *I don't have to sort of feel defensive about it [being a foodie] (…) for dieting purposes, if you're in a particular profile, and they're not recognizing that in a positive way, so you get defensive and stressed (…) whereas this has recognized it in a positive way, and given me ways of thinking about food, how I like to live, support how I like to eat food (…) but at the same time, it said you could tweak this (…) I've probably been able to take a step back and view how I eat things in a slightly different way, but without feeling guilty about it.*

In an interview with Participant 2, the participant said she did not feel her profile was the right fit and resonated with two profiles when shown all the profile descriptions. When asked for feedback, she was reviewing the core program rather than the phenotype advice and struggled to remember seeing the documents with the advice. Participant 7 opened 4 out of 7 documents with advice for her profile; however, due to a system error, she was not receiving any coaching on the program. Thus, in the interview, she primarily talked about her dissatisfaction with the course in general rather than about the phenotypes.

Discussion

Overview

The aim of this quasi-experimental study was to assess the impact of providing tailored weight loss advice on app engagements and weight loss within a digital weight loss program. Participants were patients enrolled in a weight management program offered by Oviva, a service provider for the NHS. Advice was tailored based on responses to a phenotypes quiz, which matched patients to 1 of 4 cognitive-behavioral profiles. The evidence-based, tailored advice was delivered weekly over email for 7 weeks. App engagement (eg, logging meals, reading lessons) was used as a proxy for program adherence.

Effect on App Engagements

Patients who received phenotype-tailored advice generated significantly more app engagements over a 7-week period compared with a historical cohort of patients enrolled in the same program a year earlier and compared with patients who received an invitation to take the quiz but who did not click on the link (ie, nonresponders). There was a 61.5% increase in app engagements between the phenotypes group and the historical cohort and a 90.5% increase in engagements compared with the nonresponder group. There was a significant positive correlation between the number of documents opened and the number of app engagements generated.

This finding aligns with our hypothesis that receiving tailored advice would boost program adherence. Additionally, the “try it out” section in the documents, which included exercises such

XSL•FO
RenderX

as logging hunger types in the app, likely contributed to increased engagement. This is unlikely to be driven purely by the instructions to use the app, as Oviva regularly encourages users to log activities in the app, and such elevated engagement was not observed in the comparison groups. It could be that the coupling between tailored advice and instructions to log in the app was particularly effective. This is an important finding, as adherence to digital programs, often measured by app engagement, can be challenging. Enhancing adherence to digital interventions is crucial because it directly improves health-related outcomes [49]. Therefore, strategies that increase adherence are highly valuable for the success of eHealth interventions [50-52].

The increased engagement observed in the phenotype group may have been influenced by self-selection bias, given the study's quasi-experimental design. Because participants had to actively complete multiple steps (opening an email, clicking the questionnaire link, and completing it) to receive their phenotype "match," it is possible that only the most motivated individuals chose to participate. This could, in turn, explain the higher engagement rates observed in this group. A randomized controlled trial is required to establish causality.

A randomized controlled trial could be feasibly integrated into digital weight management programs by randomizing eligible patients at the point of onboarding to either receive standard program content or phenotype-tailored advice in addition to the standard content. Randomization and content delivery could be automated within the app infrastructure, and outcomes (eg, app engagement, weight change) could be monitored through existing data pipelines. This design would allow for rigorous testing of the added benefit of phenotype-based personalization within a real-world setting, without changing or disrupting routine care.

Although engagement was higher in the phenotypes group, engagement levels declined week by week across all conditions. This pattern aligns with previous research on weight loss interventions [49-51] and Oviva's internal data. One possible explanation is habituation—as users become familiar with the app's features, the novelty may wear off, leading to reduced interaction. Additionally, some users may have consistent meal routines, logging the same foods frequently in the early weeks but gradually reducing this behavior over time. Given that meal tracking is typically the primary contributor to in-app engagement, a decline in logging frequency could naturally result in lower overall engagement levels.

There was a slight imbalance in IMD scores between the phenotype group and the historical cohort, raising the question of whether relative deprivation status influenced engagement. However, the analysis showed that IMD was not associated with engagement, and adding IMD to the model did not improve its explanatory power beyond the effect of group alone. Furthermore, the interaction term was not informative because of the shared variance between predictor variables. While this may suggest that phenotype-based personalization supported engagement across socioeconomic groups, the study was not powered to detect interaction effects, particularly given the unequal group sizes. Therefore, these null findings should be interpreted with caution.

Effect on Weight Loss

We also predicted that phenotype-tailored weight loss advice would lead to greater reductions in body mass over a 7-week period. While the phenotypes group showed more weight loss (–2.23 kg) compared with the historical cohort (–1.60 kg) and nonresponders (–0.69 kg), these differences were not statistically significant. IMD was not associated with weight loss, and the interaction between group and IMD was also nonsignificant. Bayesian analysis further confirmed that a baseline intercept model was substantially better than models including group, IMD, or their interaction, indicating that none of these variables explained variation in weight loss. This lack of statistically significant differences in weight loss between the groups may be explained by the fact that applying tailored advice takes time both to master and to impact weight loss, and our 7-week trial may have been too short to observe significant effects. A longer-term follow-up with a larger sample may be worthwhile in future studies to more clearly elucidate whether phenotype-tailored advice can increase weight loss.

Posttrial Feedback on the Phenotypes

Only a small proportion of participants (10.8%) provided posttrial feedback. While these results should be interpreted with caution, the phenotypes and tailored advice were generally well received. The 4 most endorsed statements were that the advice was clear and easy to understand, helpful, followed by participants, and likely to be used in the future.

Interviews revealed mixed opinions. Some patients found the phenotypes useful, as they helped them understand the underlying reasons for their maladaptive eating behaviors. Others, however, felt the advice was too generic. In the survey, the statement with the lowest endorsement was that the advice felt personalized. One possible explanation is that the term "personalized" created expectations of a high level of individualization that were ultimately unmet. Additionally, some participants reported that none of the profiles fully resonated with them. This is not unexpected, as the 4 profiles are not exhaustive or mutually exclusive. They were designed to provide general guidance on the cognitive and behavioral patterns most likely to hinder progress, and some individuals may relate to more than one profile or none at all.

A more hyperpersonalized approach, in which individual cognitive and behavioral dimensions are assessed separately and combined into tailored content, may improve both fit and impact. Such an approach is increasingly feasible through artificial intelligence–enabled systems. A recent study demonstrated that health advice can be tailored not only in terms of content but also in delivery, aligning with individuals' natural tendencies and cognitive styles [52]. This dual approach to personalization opens new possibilities for future work.

Recent advances in large language models are enabling digital health companies to deliver increasingly personalized messages, insights, and content [53-55]. Incorporating phenotypes into retrieval-augmented generation systems [56] offers a promising opportunity to enhance both relevance and precision of tailored

XSL•FO
RenderX

interventions [57]. By aligning content with users' cognitive styles and behaviors, such systems may improve engagement and clinical outcomes [57-60]. Although this study did not directly examine artificial intelligence–driven personalization, combining cognitive-behavioral profiling with large language models represents a compelling avenue for future digital health innovation [60].

Another limitation relates to the delivery format. Some participants may have provided feedback on the core Oviva program rather than the phenotype-specific content (posttrial interviews confirmed at least 2 instances), making it unclear whether low personalization ratings reflected dissatisfaction with the advice or broader program frustrations. Delivering tailored advice through a separate medium (by email), while the main program was delivered in-app, may also have reduced coherence. Future studies would benefit from integrating tailored content directly into the digital intervention itself to create a more seamless, unified experience that could further improve engagement and outcomes.

Strengths and Limitations

One strength of this study was testing phenotype-tailored advice on patients enrolled in an NHS-referred weight loss program, making our sample representative of UK patients seeking weight loss treatment and enhancing external validity. Additionally, by using an established program provider, we leveraged a well-developed app and core program, minimizing adherence issues due to app functionality, a common problem in prototype apps used in studies.

However, this study had limitations. The quasi-experimental design limits causal inferences. It is possible that because participants were not randomly assigned, only the most motivated individuals completed the phenotypes quiz, which could explain the higher engagement in this group. Without a randomized controlled trial or baseline motivation measurements, this alternative explanation remains possible.

Another limitation is that tailored content was delivered by email rather than integrated into the app, creating a disjointed experience that may have led participants to confuse the core program with the phenotype content. This separation could also have resulted in some participants missing the email content, leading to a fragmented experience. Finally, the phenotypes were based on data from 4 Oviva weight management programs, but the sample in this study was predominantly matched with the Red Chilli profile. Recruiting from a single program may have contributed to the observed differences in phenotype distribution. Such an uneven distribution might also indicate a need for further refinement of subgroups within profiles to better differentiate patient challenges.

Finally, weight was self-reported, which may have introduced bias (eg, social desirability, systematic underreporting), but self-reporting is standard practice in remotely delivered tier 2 programs and is accepted by the NHS for monitoring patient progress. In the program, participants track weight privately with their coach, with no penalties or incentives for weight change. In the study, phenotype-specific advice focused on cognition and behavior rather than weight outcomes, aligning with the app's broader behavioral framing. Participants were not specifically prompted to report weight or emphasize weight loss, and were unaware it would be assessed for the trial. Importantly, any bias in weight reporting is unlikely to have affected the study's primary aims, as analyses focused on differences across groups or time rather than absolute weight values.

Conclusion

In conclusion, this study provided preliminary evidence that advice tailored to cognitive-behavioral phenotypes may increase program adherence, as indicated by app engagements. Although the quasi-experimental design limits causal conclusions, the initial data trends are promising. Despite limited posttrial feedback, the positive ratings suggest that further exploration of phenotype-tailored advice is warranted. To fully assess its impact on app engagement and weight loss, a randomized controlled trial with a larger sample and longer timeframe is needed.

Acknowledgments

This study was conducted as part of a Knowledge Transfer Partnership (KTP number 10029186) supported by UK Research and Innovation (grant number 13221). The authors attest that no generative artificial intelligence (AI) technology was used in the generation of the text, figures, or any other informational content of this manuscript.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to the potential risk of participant re-identification but are available from the corresponding author on reasonable request.

Authors' Contributions

JS, KT, AJ, and LAJ contributed to the conceptualization and methodology of the study. JS, KT and AJ conducted the formal analysis, with AJ particularly focusing on the Bayesian approach. JS led the execution and coordination of the study, overseeing data collection and the provision of study materials. JS also prepared the original draft of the manuscript, which was subsequently reviewed and edited by KT, LAJ, and AJ.

XSL•FO
RenderX

Conflicts of Interest

LAJ is employed by Oviva UK. The other authors have no conflicts of interest.

Multimedia Appendix 1
All items in the phenotypes questionnaire, and additional app engagement analyses.
[DOC File , 36 KB - formative_v9i1e72645_app1.doc]

Multimedia Appendix 2
TREND checklist.
[DOCX File , 84 KB - formative_v9i1e72645_app2.docx]

References

1. Obesity and overweight. World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight [accessed 2025-01-09]
2. Obesity. National Health Service. URL: https://www.nhs.uk/conditions/obesity/ [accessed 2025-06-09]
3. English S, Vallis M. Moving beyond “eat less, move more” using willpower: reframing obesity as a chronic disease impact of the 2020 Canadian obesity guidelines reframed narrative on perceptions of self and the patient-provider relationship. Clin Obes 2023;13(6):e12615. [doi: 10.1111/cob.12615] [Medline: 37518832]
4. Annesi JJ, Whitaker AC. Psychological factors discriminating between successful and unsuccessful weight loss in a behavioral exercise and nutrition education treatment. Int J Behav Med 2010;17(3):168-175. [doi: 10.1007/s12529-009-9056-2] [Medline: 19653103]
5. Kwasnicka D, Luszczynska A, Hagger MS, Quested E, Pagoto SL, Verboon P, et al. Theory-based digital intervention to promote weight loss and weight loss maintenance (Choosing Health): protocol for a randomised controlled trial. BMJ Open 2020;10(11):e040183 [FREE Full text] [doi: 10.1136/bmjopen-2020-040183] [Medline: 33234638]
6. Dent R, McPherson R, Harper M. Factors affecting weight loss variability in obesity. Metabolism 2020;113:154388. [doi: 10.1016/j.metabol.2020.154388] [Medline: 33035570]
7. Holzapfel C, Dawcynski C, Henze A, Simon MC. Personalized dietary recommendations for weight loss: a scientific perspective from various angles. Ernahrungs Umsch 2021;68(2):26-35. [doi: 10.4455/eu.2021.008]
8. Ordovas JM, Ferguson LR, Tai ES, Mathers JC. Personalised nutrition and health. BMJ 2018;361:bmj.k2173 [FREE Full text] [doi: 10.1136/bmj.k2173] [Medline: 29898881]
9. Livingstone KM, Ramos-Lopez O, Pérusse L, Kato H, Ordovas JM, Martínez JA. Reprint of: precision nutrition: a review of current approaches and future endeavors. Trends Food Sci Technol 2022;130:51-62. [doi: 10.1016/j.tifs.2022.10.010]
10. Vandelanotte C, Trost S, Hodgetts D, Imam T, Rashid M, To QG, et al. Increasing physical activity using an just-in-time adaptive digital assistant supported by machine learning: a novel approach for hyper-personalised mHealth interventions. J Biomed Inform 2023;144:104435 [FREE Full text] [doi: 10.1016/j.jbi.2023.104435] [Medline: 37394024]
11. Romain AJ, Avignon A, Macioce V, Boegner C, Attalin A. Patterns of eating behavior in people with severe obesity seeking weight loss treatment: an exploratory study. Appetite 2022;169:105797. [doi: 10.1016/j.appet.2021.105797] [Medline: 34752827]
12. Pentikäinen S, Arvola A, Karhunen L, Pennanen K. Easy-going, rational, susceptible and struggling eaters: a segmentation study based on eating behaviour tendencies. Appetite 2018;120:212-221. [doi: 10.1016/j.appet.2017.09.001] [Medline: 28888729]
13. Kushner RF, Choi SW, Burns JL. Development of a six-factor questionnaire for use in weight management counseling. Patient Educ Couns 2016;99(12):2018-2025. [doi: 10.1016/j.pec.2016.06.022] [Medline: 27395751]
14. Wren GM, Koutoukidis DA, Scragg J, Preston G, Hennessy M, Estephane D, et al. Optimising a digitally delivered behavioural weight loss programme: study protocol for a factorial cluster randomised controlled trial. Trials 2024;25(1):477 [FREE Full text] [doi: 10.1186/s13063-024-08320-5] [Medline: 39003481]
15. Besson M, Gurviez P, Carins J. Using digital devices to help people lose weight: a systematic review. J Soc Mark 2020;10(3):289-319. [doi: 10.1108/JSOCM-07-2019-0115]
16. Kupila SKE, Joki A, Suojanen L, Pietiläinen KH. The effectiveness of eHealth interventions for weight loss and weight loss maintenance in adults with overweight or obesity: a systematic review of systematic reviews. Curr Obes Rep 2023;12(3):371-394 [FREE Full text] [doi: 10.1007/s13679-023-00515-2] [Medline: 37354334]
17. Beleigoli AM, Andrade AQ, Cançado AG, Paulo MN, Diniz MDFH, Ribeiro AL. Web-based digital health interventions for weight loss and lifestyle habit changes in overweight and obese adults: systematic review and meta-analysis. J Med Internet Res 2019;21(1):e298 [FREE Full text] [doi: 10.2196/jmir.9609] [Medline: 30622090]
18. Lehmann M, Jones L, Schirmann F. App engagement as a predictor of weight loss in blended-care interventions: retrospective observational study using large-scale real-world data. J Med Internet Res 2024;26:e45469 [FREE Full text] [doi: 10.2196/45469] [Medline: 38848556]

XSL•FO
RenderX

19. Hagiwara Y, Shimizu K, Adachi T, Kanai M, Miki T. Impact of personality prototypes on app engagement and weight loss in a mobile app-based lifestyle modification support program for individuals with obesity and non-communicable diseases. Digit Health 2025;11:20552076251347891 [FREE Full text] [doi: 10.1177/20552076251347891] [Medline: 40510189]
20. Jakob R, Harperink S, Rudolf AM, Fleisch E, Haug S, Mair JL, et al. Factors influencing adherence to mHealth apps for prevention or management of noncommunicable diseases: systematic review. J Med Internet Res 2022;24(5):e35371 [FREE Full text] [doi: 10.2196/35371] [Medline: 35612886]
21. Ryan K, Dockray S, Linehan C. A systematic review of tailored eHealth interventions for weight loss. Digit Health 2019;5:2055207619826685 [FREE Full text] [doi: 10.1177/2055207619826685] [Medline: 30783535]
22. Lustria MLA, Noar SM, Cortese J, Van Stee SK, Glueckauf RL, Lee J. A meta-analysis of web-delivered tailored health behavior change interventions. J Health Commun 2013;18(9):1039-1069. [doi: 10.1080/10810730.2013.768727] [Medline: 23750972]
23. Kreuter MW, Strecher VJ, Glassman B. One size does not fit all: the case for tailoring print materials. Ann Behav Med 1999;21(4):276-283. [doi: 10.1007/BF02895958] [Medline: 10721433]
24. Hunot-Alexander C, Croker H, Fildes A, Johnson F, Beeken RJ. Brief “appetitive trait tailored intervention”: development in a sample of adults with overweight and obesity. Behav Change 2021;39(2):106-122. [doi: 10.1017/bec.2021.22]
25. Flores Mateo G, Granado-Font E, Ferré-Grau C, Montaña-Carreras X. Mobile phone apps to promote weight loss and increase physical activity: a systematic review and meta-analysis. J Med Internet Res 2015;17(11):e253 [FREE Full text] [doi: 10.2196/jmir.4836] [Medline: 26554314]
26. Burgess E, Hassmén P, Welvaert M, Pumpa KL. Behavioural treatment strategies improve adherence to lifestyle intervention programmes in adults with obesity: a systematic review and meta-analysis. Clin Obes 2017;7(2):105-114. [doi: 10.1111/cob.12180] [Medline: 28199047]
27. Arhi C, Karagianni C, Howse L, Siddiqui M, Borg C. The effect of participation in tier 3 services on the uptake of bariatric surgery. Obes Surg 2021;31(6):2529-2536 [FREE Full text] [doi: 10.1007/s11695-021-05303-2] [Medline: 33725296]
28. Hazlehurst JM, Logue J, Parretti HM, Abbott S, Brown A, Pournaras DJ, et al. Developing integrated clinical pathways for the management of clinically severe adult obesity: a critique of NHS England Policy. Curr Obes Rep 2020;9(4):530-543 [FREE Full text] [doi: 10.1007/s13679-020-00416-8] [Medline: 33180307]
29. Berry R, Kassavou A, Sutton S. Does self-monitoring diet and physical activity behaviors using digital technology support adults with obesity or overweight to lose weight? A systematic literature review with meta-analysis. Obes Rev 2021;22(10):e13306. [doi: 10.1111/obr.13306] [Medline: 34192411]
30. Szypula J, Jones L, Barron A, Jarvstad A, Tapper K. Personalising weight loss strategies with cognitive-behavioural phenotypes. Appetite 2024;199:107487. [doi: 10.1016/j.appet.2024.107487]
31. Khanolkar AR, Patalay P. Socioeconomic inequalities in co-morbidity of overweight, obesity and mental ill-health from adolescence to mid-adulthood in two national birth cohort studies. Lancet Reg Health Eur 2021;6:100106 [FREE Full text] [doi: 10.1016/j.lanepe.2021.100106] [Medline: 34308407]
32. El-Sayed AM, Scarborough P, Galea S. Unevenly distributed: a systematic review of the health literature about socioeconomic inequalities in adult obesity in the United Kingdom. BMC Public Health 2012;12:18 [FREE Full text] [doi: 10.1186/1471-2458-12-18] [Medline: 22230643]
33. Newton S, Braithwaite D, Akinyemiju TF. Socio-economic status over the life course and obesity: systematic review and meta-analysis. PLoS One 2017;12(5):e0177151 [FREE Full text] [doi: 10.1371/journal.pone.0177151] [Medline: 28510579]
34. Szinay D, Forbes CC, Busse H, DeSmet A, Smit ES, König LM. Is the uptake, engagement, and effectiveness of exclusively mobile interventions for the promotion of weight-related behaviors equal for all? A systematic review. Obes Rev 2023;24(3):e13542. [doi: 10.1111/obr.13542] [Medline: 36625062]
35. Parkhurst A. Fat chance? Exploring the evidence on who becomes obese. 2020health. URL: https://2020health.org/wp-content/uploads/2020/11/2020FATCHANCE-Final.pdf [accessed 2025-10-23]
36. Clark TL, Savin KL, Perez-Ramirez P, Valdez T, Toba G, Gallo LC. eHealth weight loss interventions for adults with low income: a systematic review. Health Psychol 2023;42(6):353-367. [doi: 10.1037/hea0001278] [Medline: 37053431]
37. Richards R, Wren G, Whitman M. The potential of a digital weight management program to support specialist weight management services in the UK National Health Service: retrospective analysis. JMIR Diabetes 2024;9:e52987 [FREE Full text] [doi: 10.2196/52987] [Medline: 38265852]
38. Toon J, Geneva M, Sharpe P, Lavin J, Bennett S, Avery A. Weight loss outcomes achieved by adults accessing an online programme offered as part of Public Health England's 'Better Health' campaign. BMC Public Health 2022;22(1):1456 [FREE Full text] [doi: 10.1186/s12889-022-13847-w] [Medline: 35907834]
39. Birch JM, Jones RA, Mueller J, McDonald MD, Richards R, Kelly MP, et al. A systematic review of inequalities in the uptake of, adherence to, and effectiveness of behavioral weight management interventions in adults. Obes Rev 2022;23(6):e13438 [FREE Full text] [doi: 10.1111/obr.13438] [Medline: 35243743]
40. Coupe N, Cotterill S, Peters S. Enhancing community weight loss groups in a low socioeconomic status area: application of the COM-B model and behaviour change wheel. Health Expect 2022;25(5):2043-2055 [FREE Full text] [doi: 10.1111/hex.13325] [Medline: 34350682]

XSL•FO
RenderX

41. Elfhag K, Rössner S. Who succeeds in maintaining weight loss? A conceptual review of factors associated with weight loss maintenance and weight regain. Obes Rev 2005;6(1):67-85. [doi: 10.1111/j.1467-789X.2005.00170.x] [Medline: 15655039]
42. Lemstra M, Rogers MR. The importance of community consultation and social support in adhering to an obesity reduction program: results from the Healthy Weights Initiative. Patient Prefer Adherence 2015;9:1473-1480 [FREE Full text] [doi: 10.2147/PPA.S91912] [Medline: 26508843]
43. Austin JL, Smith JE, Gianini L, Campos-Melady M. Attitudinal familism predicts weight management adherence in Mexican-American women. J Behav Med 2013;36(3):259-269. [doi: 10.1007/s10865-012-9420-6] [Medline: 22487966]
44. Ahern AL, Boyland EJ, Jebb SA, Cohn SR. Participants' explanatory model of being overweight and their experiences of 2 weight loss interventions. Ann Fam Med 2013;11(3):251-257 [FREE Full text] [doi: 10.1370/afm.1446] [Medline: 23690325]
45. Noble S. English Indices of Deprivation 2019: Research Report. URL: https://assets.publishing.service.gov.uk/media/5d8b364ced915d03709e3cf2/IoD2019_Research_Report.pdf [accessed 2025-10-23]
46. Szypula J, Jarvstad A, Jones L, Tapper K. Developing Cognitive-Behavioural Phenotypes for Weight Management. 2025. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5290327 [accessed 2025-10-23]
47. Bürkner P. brms: An R package for Bayesian multilevel models using Stan. J Stat Softw 2017;80(1):1-28. [doi: 10.18637/jss.v080.i01]
48. Tapper K, Szypula J, Jarvstad A. Testing the Effectiveness of Receiving Phenotype-Tailored Content on Engagement and Weight Loss.: Open Science Framework URL: https://doi.org/10.17605/OSF.IO/UGWXB [accessed 2025-10-27]
49. Acharya SD, Elci OU, Sereika SM, Music E, Styn MA, Turk MW, et al. Adherence to a behavioral weight loss treatment program enhances weight loss and improvements in biomarkers. Patient Prefer Adherence 2009;3:151-160 [FREE Full text] [doi: 10.2147/ppa.s5802] [Medline: 19936157]
50. Lin P, Grambow S, Intille S, Gallis JA, Lazenka T, Bosworth H, et al. The association between engagement and weight loss through personal coaching and cell phone interventions in young adults: randomized controlled trial. JMIR Mhealth Uhealth 2018;6(10):e10471 [FREE Full text] [doi: 10.2196/10471] [Medline: 30341051]
51. Power JM, Phelan S, Hatley K, Brannen A, Muñoz-Christian K, Legato M, et al. Engagement and weight loss in a web and mobile program for low-income postpartum women: fit moms/mamás activas. Health Educ Behav 2019;46(2_suppl):114-123. [doi: 10.1177/1090198119873915] [Medline: 31742447]
52. Briazu R, Bell L, Dodd G, Blackburn S, Massri C, Chang B, et al. The effectiveness of personalised food choice advice tailored to an individual's socio-demographic, cognitive characteristics, and sensory preferences. Appetite 2024;201:107600 [FREE Full text] [doi: 10.1016/j.appet.2024.107600] [Medline: 39002566]
53. Yarkoni T, Westfall J. Choosing prediction over explanation in psychology: lessons from machine learning. Perspect Psychol Sci 2017;12(6):1100-1122 [FREE Full text] [doi: 10.1177/1745691617693393] [Medline: 28841086]
54. Dao D, Teo JYC, Wang W, Nguyen HD. LLM-Powered multimodal AI conversations for diabetes prevention. 2024 Presented at: AIQAM '24: Proceedings of the 1st ACM Workshop on AI-Powered Q&A Systems for Multimedia; June 10, 2024; Phuket p. 1-6. [doi: 10.1145/3643479.3662049]
55. Gencer G, Gencer K. Large language models in healthcare: a bibliometric analysis and examination of research trends. J Multidiscip Healthc 2025;18:223-238 [FREE Full text] [doi: 10.2147/JMDH.S502351] [Medline: 39844924]
56. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H, editors. Advances in Neural Information Processing Systems. Volume 33. United States: Curran Associates, Inc; 2020:9459-9474.
57. Kim DW, Park JS, Sharma K, Velazquez A, Li L, Ostrominski JW, et al. Qualitative evaluation of artificial intelligence-generated weight management diet plans. Front Nutr 2024;11:1374834 [FREE Full text] [doi: 10.3389/fnut.2024.1374834] [Medline: 38577160]
58. Morgan-Bathke M, Raynor HA, Baxter SD, Halliday TM, Lynch A, Malik N, et al. Medical nutrition therapy interventions provided by dietitians for adult overweight and obesity management: an academy of nutrition and dietetics evidence-based practice guideline. J Acad Nutr Diet 2023;123(3):520-545.e10. [doi: 10.1016/j.jand.2022.11.014] [Medline: 36462613]
59. Bush CL, Blumberg JB, El-Sohemy A, Minich DM, Ordovás JM, Reed DG, et al. Toward the definition of personalized nutrition: a proposal by the American Nutrition Association. J Am Coll Nutr 2020;39(1):5-15 [FREE Full text] [doi: 10.1080/07315724.2019.1685332] [Medline: 31855126]
60. Zhang K, Meng X, Yan X, Ji J, Liu J, Xu H, et al. Revolutionizing health care: the transformative impact of large language models in medicine. J Med Internet Res 2025;27:e59069 [FREE Full text] [doi: 10.2196/59069] [Medline: 39773666]

Abbreviations

BF: Bayes Factor
DV: dependent variable
IMD: Index of Multiple Deprivation
NHS: National Health Service

XSL•FO
RenderX

Edited by A Mavragani; submitted 14.02.25; peer-reviewed by N Mungoli, VSK Kancharla, M Gasmi ; comments to author 07.03.25; revised version received 27.06.25; accepted 25.09.25; published 01.12.25.

Please cite as:
Szypula J, Jarvstad A, Jones LA, Tapper K
Personalizing a Weight Loss Program Using Cognitive-Behavioral Phenotypes to Improve Engagement and Weight Loss in Adults With Overweight or Obesity: Quasi-Experimental Study
JMIR Form Res 2025;9:e72645
URL: https://formative.jmir.org/2025/1/e72645
doi:10.2196/72645
PMID:

©Joanna Szypula, Andreas Jarvstad, Lucy Amelia Jones, Katy Tapper. Originally published in JMIR Formative Research (https://formative.jmir.org), 01.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Accuracy and Variability of a Commercial Markerless Motion Capture System Compared to a Pressure Mat for Weight Distribution in Standing: Cross-Sectional Observational Study

Lisa Sheehy[1], PhD; Emma Gal-Dev[1,2]; Heidi Sveistrup[1,3,4,5], PhD; Martin Bilodeau[1,3,5], PhD; Hillel Finestone[1,6,7], MD

[1]Bruyère Health Research Institute, Ottawa, ON, Canada
[2]Interdisciplinary School of Health Sciences, Faculty of Health Sciences, University of Ottawa, Ottawa, ON, Canada
[3]School of Rehabilitation Sciences, Faculty of Health Sciences, University of Ottawa, Ottawa, ON, Canada
[4]Department of Systems and Computer Engineering, Faculty of Engineering and Design, Carleton University, Ottawa, ON, Canada
[5]School of Human Kinetics, Faculty of Health Sciences, University of Ottawa, Ottawa, ON, Canada
[6]Department of Medicine, Division of Physical Medicine and Rehabilitation, Faculty of Medicine, University of Ottawa, Ottawa, ON, Canada
[7]Physical Medicine and Rehabilitation, Bruyère Health, Ottawa, ON, Canada

Corresponding Author:
Lisa Sheehy, PhD
Bruyère Health Research Institute
43 Bruyère St.
Ottawa, ON, K1N5C8
Canada
Phone: 1 6135626262 ext 1593
Email: lsheehy@bruyere.org

Abstract

Background: Commercial markerless motion capture (MMC) systems show promise for use in rehabilitation and have been validated for the assessment of various parameters. However, no prior studies have evaluated MMC systems to detect stance asymmetry.

Objective: The objective of this study was to assess the accuracy and variability of the Jintronix Weight Shift Tool MMC system to estimate the percentage of weight borne on each foot.

Methods: Twelve healthy younger adults, 12 healthy older adults, and 12 people living with stroke were recruited for this cross-sectional study. The percentage of weight borne on each foot was simultaneously recorded by the Weight Shift Tool and a validated pressure mat during 2 series (raising the arm to capture the recording and without arm raise) with left lean, right lean, and equal stance. Agreement between the Weight Shift Tool and the pressure mat was assessed using Bland-Altman analyses.

Results: Bias was greatest for older adults for all stances except for right lean without arm raise. Variability was greatest for people living with stroke for all stances except for left lean with arm raise. On average, the limits of agreement were narrower during equal stance. Although bias between the Weight Shift Tool and the pressure mat was small to moderate (0.0%-11.7%), the limits of agreement were wide (12.8%-33.6% above and below the bias).

Conclusions: The Weight Shift Tool is not clinically acceptable for the estimation of the percentage of weight on each foot due to high variability. Investigation of other MMC systems is required to confirm the validity of MMC for clinical assessment.

(JMIR Form Res 2025;9:e73575) doi:10.2196/73575

KEYWORDS

balance; motion capture; rehabilitation; stroke; stroke rehabilitation; validation study; virtual reality; weight shift

Introduction

The use of markerless motion capture (MMC) for tracking human movement has provided numerous possibilities for exergaming (exercise performed by playing video games that require physical activity), in that it allows for unrestricted movement unhindered by the need to wear a special garment, apply reflective dots, or hold a controller. With a video or infrared camera, software is used to detect the location of the main segments and joints of the body and follow their

XSL•FO
RenderX

movements. Furthermore, software can track and record joint angles and changes in tilt from the vertical. These features may make the use of MMC technology helpful for the clinical assessment of range of motion and posture, as it provides an easier, faster, and cheaper means of assessment in comparison to marker-based motion capture systems [1,2]. In addition, MMC shows promising potential for application in stroke rehabilitation. Exercises performed with MMC technology may provide effective physiotherapy supplementation for rehabilitation after stroke, in a clinical setting or at home [3-7]. The use of MMC technology for telerehabilitation allows feedback to be provided during exercise and for clinicians to track progress from a distance, reducing the need for frequent in-person visits.

There is current interest in the development of accurate and reliable MMC technologies for home rehabilitation and physiotherapy supplementation. Recently, novel MMC systems using commercial infrared camera systems based on deep learning have been developed [8-10]. The most commonly used system in studies evaluating MMC for stroke rehabilitation is the Microsoft Kinect sensor (Microsoft Corp) [4]. The Microsoft Kinect v2 sensor has been tested in healthy individuals and people living with stroke. From testing with healthy individuals, it was deemed reliable for the assessment of gait (r=0.92 for gait velocity and r=0.81 for step length), and valid to assess velocity (no difference between Kinect and Vicon, P=.26) and acceleration (no difference between Kinect and Vicon, P=.42) of one's center of gravity in the anteroposterior direction while performing a sit-to-stand task [11,12]. However, Yeung et al [13] found that the Kinect sensor was more accurate in measuring body sway in the mediolateral direction (mean difference of 2.0 mm between Kinect and Vicon) than in the anteroposterior direction (mean difference of 3.7 mm between Kinect and Vicon). Mobini et al [14] evaluated the Kinect sensor's ability to monitor the progress of upper-body rehabilitation in people living with stroke. The Kinect sensor was able to reliably detect changes in mean hand velocity (R^2=0.96), the logarithm of the median of the hand's dimensionless jerk (R^2=0.94), and the hand's path curvature (R^2=0.97) after 1 month of rehabilitation [14]. The Kinect v2 sensor also tested favorably against manual and digital goniometry for measuring shoulder joint angles in healthy individuals, with differences between readings generally less than 3 degrees except when measuring external rotation [15]. Overall, studies seem to agree that MMC, and specifically the Kinect sensor, shows promise for application as a clinical assessment and rehabilitation tool.

One commercial application of the Kinect v2 sensor for rehabilitation is the Jintronix Rehabilitation System (Jintronix Inc). In addition to using user location and body segment movement for exergaming, clinicians are provided with a measure of stance symmetry through the Weight Shift Tool. Although the use of MMC has been evaluated for many assessments, no studies have been found on its accuracy and variability in measuring weight distribution on each foot. Stance asymmetry is a clinically important measure in physiotherapy. Weight distribution on each foot is an aspect of balance function, which is necessary for independence in activities of daily living and fall prevention [16-18]. In fact, asymmetrical weight-bearing is a common result of stroke [19]. Furthermore, weight-bearing asymmetry can act as an indicator of loss of balance, is positively correlated with postural sway, and may be a useful outcome measure for individuals with lateropulsion [16,20,21]. Therefore, rehabilitation interventions often target improvements in balance and symmetry, both in static and dynamic conditions, as important components of the recovery of patients with stroke, orthopaedic injuries, and gait abnormalities [22-24]. Older adults may also experience declines in balance function as a result of physiological changes from aging [25]. Current assessment methods for weight-bearing asymmetry include the use of two digital bathroom scales, used by standing with one foot on each scale, which has been deemed accurate for patients with stroke (ranging from F_{38}=0.006; P=.94 to F_{38}=0.074; P=.79), although significant variability was noted between trials on the paretic (F_{38}=3.47; P=.004) and nonparetic (F_{38}=4.51; P=.002) sides [26]. An easy, accurate, and reliable measurement of weight distribution in standing could be used to assess and treat stance asymmetries, consequently improving the quality of life in people living with stroke. The validation of commercially available MMC technology, such as the Jintronix Weight Shift Tool, as an acceptable tool for the assessment of weight distribution in standing could be beneficial for clinicians and patient outcomes. Therefore, the purpose of this study was to investigate the ability of the Weight Shift Tool, which incorporates the Kinect v2 sensor, to assess weight distribution in standing by estimating the percentage of weight on each foot.

The objectives of this study were as follows:

1. To assess the accuracy and variability of the Weight Shift Tool, compared to a pressure mat;
2. To determine if raising one's arm (to capture a recording) affected the accuracy or variability of the Weight Shift Tool;
3. To determine if there was a difference in the accuracy or variability of the Weight Shift Tool depending on participant group (younger adults, older adults, people living with stroke).

The hypotheses were as follows:

1. That the accuracy of the Weight Shift Tool is equivalent to a pressure mat for measuring weight distribution in standing and that the variability of the Weight Shift Tool (measured using the Bland-Altman limits of agreement) is less than 5%.
2. That raising one's arm would demonstrate no significant effect on the accuracy or variability of the Weight Shift Tool.
3. That there would be no significant difference in the accuracy or variability depending on participant group.

This preliminary study on the accuracy and variability of the Weight Shift Tool will help inform further research on its validation and clinical utility for assessing weight distribution in standing for different populations.

XSL•FO
RenderX

Methods

Trial Design

This was a cross-sectional observational study which included convenience samples from 3 populations—healthy younger adults, healthy older adults, and people living with stroke.

Participants

Healthy adults (18 years and older) were recruited through posters placed around the community and by word of mouth. Inclusion criteria comprised the ability to stand independently for at least 5 minutes at a time, and being free of any condition that could impair standing balance (ie, vestibular deficits, recent concussion or cervical whiplash injury, lower extremity amputation, second or third trimester of pregnancy, neurological disorders such as multiple sclerosis, severe lower extremity injury, and advanced diabetes). People living with stroke of any type (50 years and older) were recruited from the Bruyère Health inpatient stroke rehabilitation program. Inclusion criteria comprised the ability to stand independently (no gait aid or assistance, although close supervision was acceptable) for at least 5 minutes at a time, and being medically stable, without seizures or any of the previously mentioned conditions that could further impair standing balance. All participants had to have adequate vision to follow an exergame on a TV screen located 2 m away.

As this was a preliminary, observational study, no sample size calculation was performed. Sample size was determined by the availability of resources. People living with stroke were age-matched (within an average of 5 years) to healthy older adults. Because of limited resources, the sample size was not large enough to allow for subgroup analyses. The following demographic information was collected: sex, weight, height, and age, as well as type of stroke and hemiplegic side for people living with stroke.

Equipment

The percentage of weight on each foot assessed by the Jintronix Rehabilitation System Weight Shift Tool was compared to that assessed by the MatScan pressure mat (Tekscan, Inc).

The Weight Shift Tool is an assessment tool included within the Jintronix Rehabilitation System nonimmersive virtual reality platform. The Jintronix Rehabilitation System was originally designed for clinician-led stroke recovery, and includes a variety of customizable games, activities and exercises designed to address standing and sitting balance, gait, and upper extremity movement. The Jintronix Rehabilitation System is based on MMC using the Kinect v2 sensor hardware to interact with the user. The Kinect v2 sensor uses 2 cameras (both sampled at 30 Hz), a red, green, and blue visual camera and an infrared projector combined with an infrared detector that receives returning infrared light reflected from objects [27]. The more distant the object, the longer the “time of flight” between projection and detection, allowing for depth detection. The associated software, Windows Software Development Kit (Microsoft Corp), uses the cameras to detect 20 joints on a skeleton and therefore recognize gestures and motion (Figure 1). The Jintronix Rehabilitation System uses a proprietary algorithm to use the position of the joints to estimate the percent weight on each foot, which is unavailable publicly.

Figure 1. Skeleton of body as seen by the Kinect v2 sensor and Jintronix software.

The Weight Shift Tool measures the center of pressure kinematically to estimate the percentage of weight on each foot. The user stands 1.90-2.20 m from the Kinect v2 sensor within a 0.30 m by 0.30 m area and sees a mirror image of themselves on the screen in real time. The Weight Shift Tool determines the midline of the body (green line in Figure 2), and the degree that this line is off the vertical is translated into an estimation of the percentage of weight on each foot. The percentage can be shown on the screen (Figure 2) or hidden. The percentages are updated in real time as the user’s weight shifts. The user is prompted to take a point-in-time capture of the percentage of weight on each foot by raising their arm vertically. Recordings are stored in the online Jintronix database. To record the percentage of weight on each foot without raising the arm for the purposes of this study, Movavi Video Suite Personal 14 movie editing software (Movavi Software Inc) was used for single point-in-time screen capture of the Jintronix computer screen, triggered by a computer-keyboard hot key.

XSL•FO
RenderX

Figure 2. Screen capture of the Weight Shift Tool showing a participant with the percentage of weight borne on each foot visible on the screen.

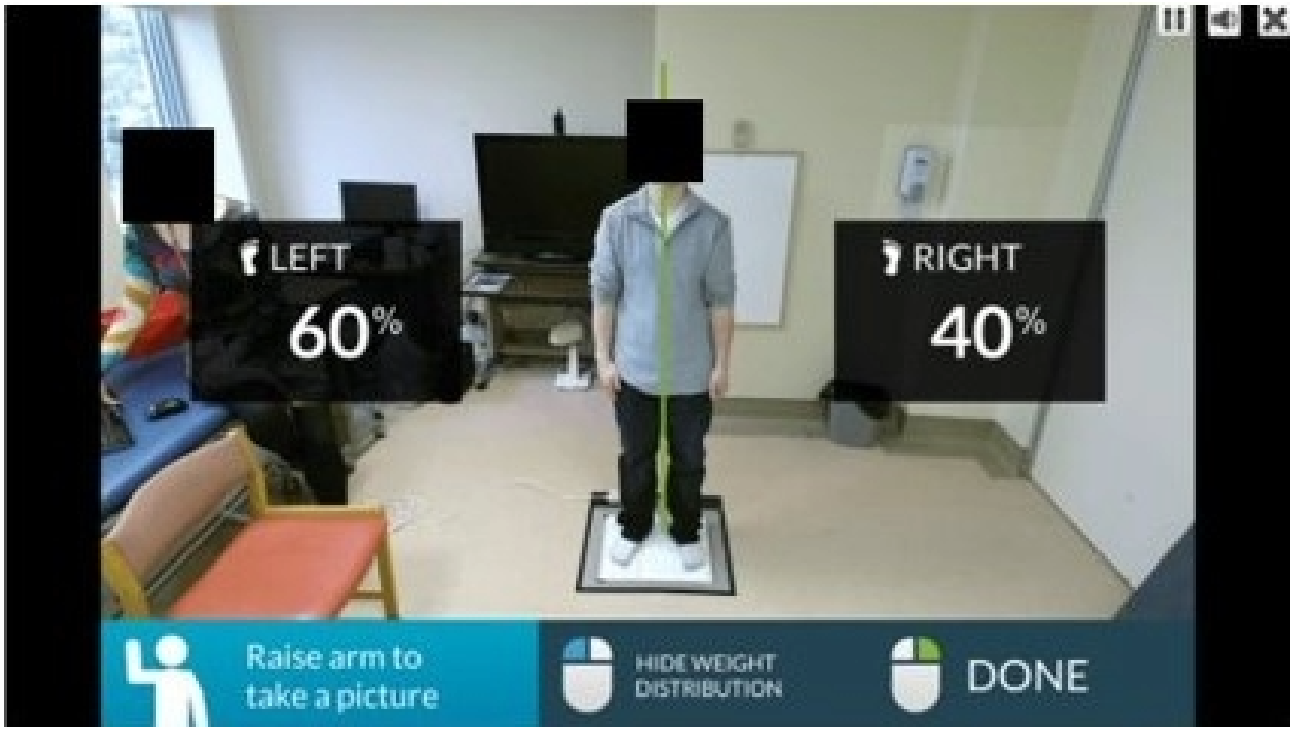

While testing the Weight Shift Tool, participants stood on a MatScan pressure mat, which measures pressure to estimate the percentage of weight on each foot. The MatScan pressure mat consists of a 44 by 52 array of sensels, each measuring 0.84 cm by 0.84 cm. The mat was calibrated to each participant's weight and the pressure under the feet was recorded at one point in time, triggered by a computer keyboard hot key. MatScan Research Software (Tekscan Inc) was used to position a divider between the 2 feet and calculate the percentage of weight under each foot. Although a force plate is regarded as the gold standard for measuring the forces produced in stance, the MatScan pressure mat has been validated against the force plate for healthy, younger adults (r=0.92-0.99 and P<.05 for all center-of-pressure variables during single-leg stance with eyes-open and eyes-closed trials; r=0.85-0.98 and P<.05 for the average percent change for all center-of-pressure variables between eyes-open and eyes-closed trials) and has been deemed reliable for measuring postural stability in older adults with rheumatoid arthritis (intraclass correlation coefficient=0.84 (95% CI 0.63-0.93) to 0.92 (95% CI 0.81-0.97), SEM=1.27-2.35 mm) [28,29]. The pressure mat was chosen for this study because it is more feasible to use in a clinical setting than a force plate.

Procedure

Demographic data were documented for all participants, and people living with stroke were administered the Berg Balance Scale to characterize their balance ability [30]. All participants wore pants and a shirt (ie, no skirts) and stood in sock feet on the pressure mat, in a comfortable stance, with hands by their side and eyes looking at the Jintronix screen (Figure 3). The position of the participant's feet on the pressure mat was drawn on a piece of paper taped to the pressure mat during the initial stance so that the exact foot placement would be reproduced for the following stances. Two series of 9 stances were captured; the order of the series was randomized. In the first series, the Jintronix view of the percentage of weight on each foot was activated but covered on the screen, so that the participants could not see the percentages. Participants held the required position, and a screen capture and pressure mat recording were taken simultaneously. In the second series, the Jintronix view of the percentage of weight on each foot was deactivated, and the pressure mat recording was taken as the participant raised their arm with the shoulder in 90 degrees of abduction and the elbow in 90 degrees of flexion to capture the Weight Shift Tool recording (as seen on the TV screen in Figure 2). Participants could not be blinded to their arm raise condition. In both series, participants were verbally instructed to "stand with equal weight bearing on both feet," "stand with more weight on the left foot," or "stand with more weight on the right foot." People living with stroke completed the tasks to their ability, even if it was not possible for them to put more weight on one foot. Each stance was repeated 3 times in each series, presented in a randomized fashion. All stances were completed by all participants. Participants could take a rest (standing or sitting) during testing, ensuring that their feet were repositioned accurately on the footprints. The purpose of measuring recordings with arm raised (AR) and with no arm raised (NAR) was to determine if the use of the Jintronix software's recommended means of capturing a recording (ie, by raising the arm) would alter measurements of weight distribution by disturbing the participant's natural stance or by exaggerating or diminishing the participant's lean. Therefore, both arm raise conditions were compared.

XSL•FO
RenderX

Figure 3. Experimental set-up with the participant standing on the pressure mat.

Analysis

The pressure mat recordings and Weight Shift Tool recordings from the left foot were analyzed, since the percentage of weight borne on the right foot would always be the recorded percentage of weight borne on the left foot subtracted from 100%, and consistent analysis of the same foot was necessary. An average of the 3 repetitions for each series (AR or NAR), lean type and participant was calculated and used for analysis. Each of the 3 participant groups (younger adults, older adults, and people living with stroke) was analyzed separately. Bland-Altman analyses were performed using Microsoft Excel (Microsoft) to assess agreement between the pressure mat recordings and each method of capturing Weight Shift Tool recordings (raising arm to record, or screen capture without raising arm) [31]. Bland-Altman analysis involves plotting the difference between paired measurements of 2 different methods of clinical assessment (*y*-axis) against the means of the paired measurements (*x*-axis). This allows for an assessment of the agreement of the 2 methods (ie, how interchangeable they may be). The mean of all the differences represents the bias (ie, the average difference between the 2 methods). If the 95% CI for the bias includes zero, then the mean percent weight on the left foot as measured from the Weight Shift Tool is statistically equivalent to the mean percent weight on the left foot as measured from the pressure mat. If the 95% CI for the bias does not include zero, then the Weight Shift Tool will not be considered accurate compared to the pressure mat. The "limits of agreement" of a Bland-Altman plot represent the 95% CI of the distribution of the differences between the 2 methods. If the bias is significantly different from zero, or the limits of agreement are greater than 5% weight on the left foot, then the 2 methods may not be used interchangeably. The acceptable range of variability is informed by a study by Martello et al [32], who established a minimal detectable change (MDC) of 5% for weight-bearing symmetry in people living with stroke. Therefore, limits of agreement exceeding 5% would mean that variability caused by the Weight Shift Tool could be misconstrued as a detectable change of weight distribution in standing. Exceeding this threshold would render the Weight Shift Tool unsuitable for measuring weight distribution in standing, as changes detected by it may not represent true changes in weight-bearing symmetry.

To investigate the impact of participant group, lean, and arm raise on the data, a 3-way mixed methods ANOVA was performed with SPSS (version 28; IBM Corp). The dependent variable was the difference between the percentage of weight measured on the left foot as tested with the Weight Shift Tool and with the pressure mat. The between-subject factor was group. Within-subject factors were the direction of lean and the presence of arm raise. Post hoc analyses were performed using univariate ANOVAs. Statistical significance was defined as $P<.05$.

Data are available from the authors upon reasonable request.

Ethical Considerations

The study was approved by the Research Ethics Boards at Bruyère Health (M16-15-030) and the University of Ottawa (A07-15-02). All participants provided written informed consent. Participants' privacy and confidentiality were maintained

XSL•FO
RenderX

throughout the study. Participants were provided with a CAD $5 (US $3.88) coffee shop gift card to thank them for their time.

Results

Twelve individuals in each of the 3 groups participated. Demographic data are presented in Table 1. The mixed methods ANOVAs analyzed only 11 people living with stroke, due to 1 participant leaning to the incorrect side during every right lean.

Table 1. Demographic data for study participants.

Participant group	Number	Sex	Age (years), mean (SD)	BMI (kg/m^2), mean (SD)	Hemiplegic side	Type of stroke	Berg Balance Scale (out of 56), mean (SD)
Younger adults	12	• 7 male • 5 female	22.9 (3.5)	23.7 (3.2)	N/A	N/A	N/A
Older adults	12	• 2 male • 10 female	67.2 (9.3)	25.3 (3.7)	N/A	N/A	N/A
People living with stroke	12	• 10 male • 2 female	69.8 (13.5)	26.2 (4.8)	• 6 right • 6 left	• 8 ischemic • 2 hemorrhagic • 2 unknown	48.7 (8.4)

Bland-Altman plots were prepared for each participant group (younger adults, older adults, and people living with stroke), separately for NAR (Figure 4) and AR (Figure 5). Overall, for all groups, leans and arm raises, the bias between the Weight Shift Tool and the pressure mat was small to moderate (0.0%-11.7% of body weight on the left foot) (Tables 2-4). However, limits of agreement were generally wide (12.8%-33.6% above and below the bias), indicating high levels of variability. With AR, younger adults had the smallest average bias of all participant groups, in equal (1.9%), left (0.0%), and right (−0.9%) stances. This was not the case with NAR, where no participant group consistently had the lowest average bias across all stances. Biases were smallest for people living with stroke equal stance NAR (0.6%), people living with stroke left lean NAR (0.2%), younger adults left lean AR (0.0%), and younger adults right lean AR (−0.9%). Across all lean conditions and both arm raise conditions, bias was greatest for older adults except for right lean NAR.

Figure 4. Bland-Altman plots for younger adults, older adults, and people living with stroke, showing 3 stance positions (equal stance, left lean, and right lean), without raising an arm to trigger a screenshot (no arm raise). The solid orange line represents the bias between the Weight Shift Tool and the pressure mat. Dashed orange lines represent the limits of agreement for the bias. The orange area represents the 95% CI for the bias. The blue area represents the 95% CIs for the limits of agreement. NAR: no arm raise; PLWS: people living with stroke.

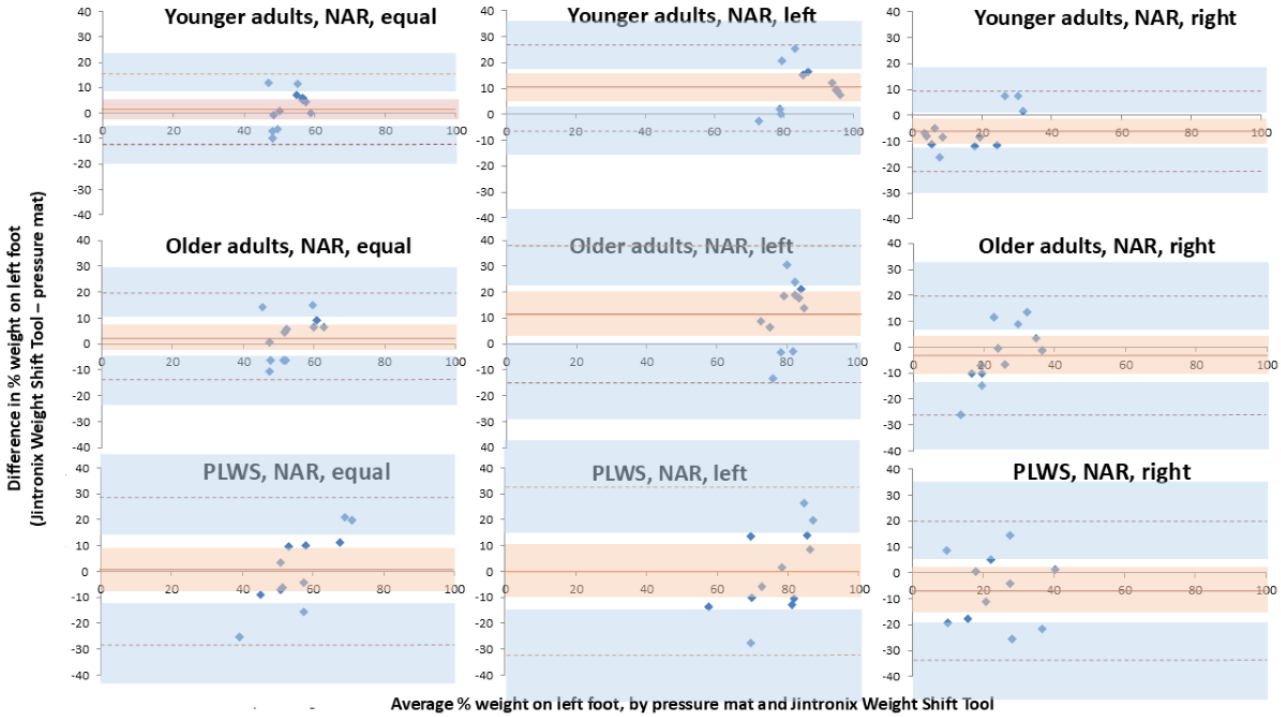

XSL•FO
RenderX

Figure 5. Bland-Altman plots for younger adults, older adults, and people living with stroke, showing 3 stance positions (equal stance, left lean, and right lean), raising an arm to trigger a screenshot (arm raise). The solid orange line represents the offset between the Weight Shift Tool and the pressure mat. Dashed orange lines represent the limits of agreement for the bias. The orange area represents the 95% CI for the bias. The blue area represents the 95% CIs for the limits of agreement. AR: arm raise; PLWS: people living with stroke.

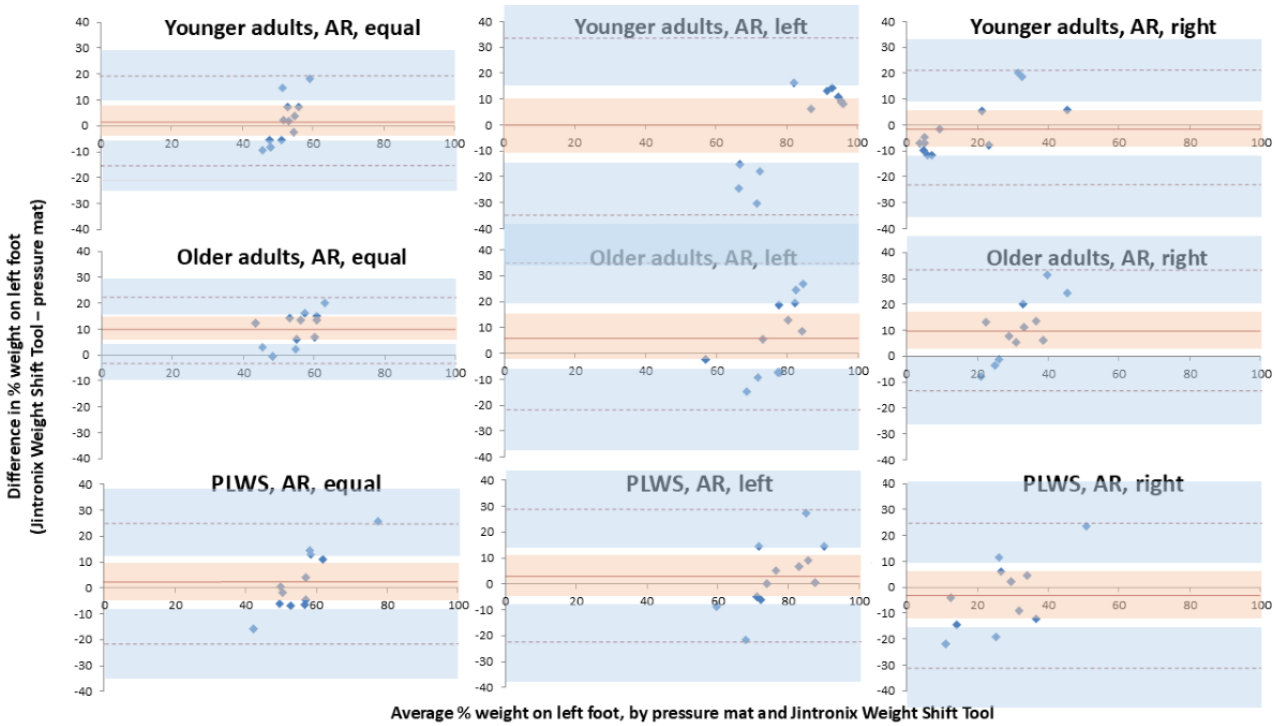

Table 2. Percent weight on the left foot as measured by the Weight Shift Tool and the pressure mat for younger adults. Means, SD, bias, and limits of agreement of 3 trials.

Arm raise condition and stance	Weight on left foot measured by Weight Shift Tool, mean% (SD)	Weight on left foot measured by pressure mat, mean% (SD)	Bias between Weight Shift Tool – pressure mat, % (limits of agreement)
No arm raise			
Equal weight on both feet	53.6 (6.7)	51.6 (4.4)	1.9 (−12.1 to 16.0)
Left lean	92.1 (9.9)	81.8 (8.4)	10.4 (−6.2 to 27.0)
Right lean	12.6 (13.5)	18.5 (8.8)	−5.9 (−21.1 to 9.4)
Arm raise			
Equal weight on both feet	53.2 (7.6)	51.2 (2.9)	2.0 (−15.6 to 19.5)
Left lean	84.3 (19.6)	84.3 (6.2)	0.0 (−33.6 to 33.6)
Right lean	15.8 (18.8)	16.7 (10.6)	−0.9 (−23.2 to 21.3)

Table 3. Percent weight on the left foot as measured by the Weight Shift Tool and the pressure mat for older adults. Means, SD, bias, and limits of agreement of 3 trials.

Arm raise condition and stance	Weight on left foot measured by Weight Shift Tool, mean% (SD)	Weight on left foot measured by pressure mat, mean% (SD)	Bias between Weight Shift Tool – pressure mat, % (limits of agreement)
No arm raise			
Equal weight on both feet	54.5 (8.8)	51.8 (5.6)	2.7 (−14.3 to 19.7)
Left lean	85.9 (9.0)	74.2 (5.8)	11.7 (−14.2 to 37.6)
Right lean	22.8 (12.2)	26.0 (5.2)	−3.2 (−26.4 to 20.0)
Arm raise			
Equal weight on both feet	60.0 (8.5)	49.8 (5.3)	10.3 (−2.6 to 23.1)
Left lean	79.6 (13.7)	73.3 (6.6)	6.3 (−22.4 to 34.9)
Right lean	36.7 (12.4)	26.8 (5.0)	9.9 (−13.1 to 33.0)

XSL•FO
RenderX

Table 4. Percent weight on the left foot as measured by the Weight Shift Tool and the pressure mat for people living with stroke. Means, SD, bias, and limits of agreement of 3 trials.

Arm raise condition and stance	Weight on left foot measured by Weight Shift Tool, mean% (SD)	Weight on left foot measured by pressure mat, mean% (SD)	Bias between Weight Shift Tool – pressure mat, % (limits of agreement)
No arm raise			
Equal weight on both feet	56.3 (16.0)	55.7 (5.6)	0.6 (−27.9 to 29.1)
Left lean	77.1 (15.4)	76.8 (7.9)	0.2 (−32.2 to 32.7)
Right lean	20.2 (11.7)	26.5 (12.3)	−6.3 (−33.2 to 20.6)
Arm raise			
Equal weight on both feet	57.1 (14.0)	54.8 (4.7)	2.2 (−21.2 to 25.6)
Left lean	78.5 (14.3)	75.5 (6.9)	3.0 (−22.7 to 28.7)
Right lean	25.7 (17.1)	28.6 (9.0)	−2.9 (−30.9 to 25.1)

The acceptable threshold for limits of agreement (±5%) was exceeded for all conditions. Across groups, variability was greatest for people living with stroke in all lean and arm raise conditions except for left lean AR, with limits of agreement ranging from 23.4% to 32.5% above and below the bias. On average, limits of agreement were narrower with equal weight on both feet (ranging from 23.4% to 28.5% above and below the bias) than with right lean (ranging from 26.9% to 28.0% above and below the bias) or left lean (ranging from 25.7% to 32.5% above and below the bias). With equal stance, bias was greatest for older adults, for both AR (10.3%) and NAR (2.7%). With equal stance, NAR produced a smaller bias (ranging from 0.6% to 2.7%) than AR (ranging from 2.0% to 10.3%).

There was no 3-way interaction between group, lean and arm raise (P=.12), and no 2-way interaction between group and lean (P=.82). There was, however, a 2-way interaction between group and arm raise (P=.02) and between lean and arm raise (P=.004). While there was no effect of group when there was no arm raise (P=.45), there was an effect when the arm was raised (P=.04), such that there was less bias in younger adults compared to older adults (Figures 4 and 5). Lean direction had an effect on the bias when the arm was not raised (P<.001), such that there was a smaller bias with equal weight-bearing, a tendency to overestimate when leaning to the left, and a tendency to underestimate when leaning to the right (Figure 4). Combining all participant groups and with the arm not raised (Figure 4), the bias during equal weight-bearing versus left lean was similar (P=.10). However, equal weight-bearing versus right lean was significantly different (P=.04), as were left versus right lean (P<.005). There was no effect on the bias from lean directions when the arm was raised (P=.67; Figure 5).

Discussion

Principal Findings

None of the study's hypotheses were confirmed by the results. This study found that, on average, for all stances and groups, the bias between the Weight Shift Tool and the pressure mat was small to moderate and proved not to be statistically significant. However, Bland-Altman analysis showed that variability was high across all groups. In older adults and people living with stroke, the limits of agreement of all stances exceeded acceptable levels of variability (limits of agreement≤±5%). In younger adults, the limits of agreement of all stances were generally narrower and the average bias was the smallest of each participant group for all stances, although still beyond acceptable levels of variability. This suggests that the Weight Shift Tool's proprietary algorithm for estimating weight shift may have been trained using a population of healthy, younger adults and adjusted for accuracy when raising the arm, as originally intended for its use. Variations in posture, body shape, and weight distribution strategies may explain the greater biases of older adults and wider limits of agreement for people living with stroke, as these factors may affect the ability of the Weight Shift Tool to accurately estimate the percentage of weight borne on each foot due to its method of estimating asymmetry. For example, by kinematically monitoring deviation from the vertical of central body segments, the Weight Shift Tool may not interpret a change in weight distribution if a person increases the weight under one foot by bending the ipsilateral knee. Furthermore, for all groups, the limits of agreement were narrower when standing with equal weight on both feet. This shows that the Weight Shift Tool is more variable with drastic asymmetries in weight-bearing, also noted as there were instances in which the Weight Shift Tool estimated that 100% of a participant's weight was on one foot when leaning, although this was not the case. With no arm raise, a statistically significant difference was noted for right lean (tendency to underestimate bias) versus left lean (tendency to overestimate bias), as well as right lean versus equal stance (smaller bias), but not for left lean versus equal stance. This trend was unexpected and could be due to the algorithm of the Weight Shift Tool compensating for differences in symmetry when raising the arm, as this is the designed method for capturing recordings.

Previous studies have compared the Kinect sensor or other MMC to a pressure mat or force plate. Dolatabadi et al [33] validated the Microsoft Kinect sensor for Windows v2 against the GAITRite pressure mat for the analysis of gait parameters (velocity, stance time, step length, and step time) in healthy young adults. Gonzalez et al [34] found that fewer stabilometric parameters of the center of mass measured by the Kinect v2 sensor detected body sway compared to parameters of center of pressure and center of mass measured by a force plate. Dehbandi et al [35] found that the centers of pressure measured by the Kinect v2 sensor and a force plate during a seated balance

XSL•FO
RenderX

task in healthy participants were strongly correlated. As the Kinect v2 sensor was deemed sufficiently accurate, they created an analytical algorithm based on machine learning for the assessment of postural stability. The Vicon MMC system was deemed accurate for estimating ground reaction forces in comparison to a force plate (average root-mean-square difference of 0.75 N/kg across all dimensions and tasks, R^2=0.93-0.99 for peak forces across tasks) [36]. Bland-Altman analysis of the Qualisys MMC system compared to a force plate for evaluating the margin of stability during gait initiation generally showed broad limits of agreement (>20%), which were lower in healthy young adults, greater in healthy older adults, and greatest in patients with Parkinson disease [37]. Therefore, not all MMC systems appear to demonstrate comparable accuracy and variability. Although there is precedent work evaluating the accuracy and variability of the Kinect sensor and other MMC, this is the first study to compare the percentage of weight on each foot as measured by the Kinect sensor and the pressure mat.

The findings of this study are in line with previous studies on the accuracy of the Kinect sensor in detecting movements in the mediolateral direction. Several authors found that the Kinect v2 sensor measured trunk lean and body sway less accurately in the mediolateral direction than in the anteroposterior direction compared to the Vicon marker–based motion tracking system, particularly during double-leg tests [38,39]. Clark [40] found that lateral trunk lean angles measured with the Kinect sensor were inaccurate when tested against the Vicon marker–based motion tracking system without calibration (mean error of 3.2°, SD 2.2°) [40]. Individualized calibration (which was also done in this study) significantly improved the comparison (mean error of 0.8°, SD 0.8°). Thus, the conclusion that the Weight Shift Tool was not accurate in the mediolateral direction is confirmed by other studies evaluating the Kinect sensor. Furthermore, Luizzo et al (2017) [41] evaluated the reliability of the Wii Fit balance board for measuring weight-bearing asymmetry in older adults and people living with stroke. The study determined that the MDC for older adults and people living with stroke was 8.8% and 13.6%, respectively, which were both larger than the MDC established by Martello et al [32]. The authors concluded that the MDC was sufficiently large as to deem the Wii Fit balance board unreliable for measuring weight-bearing asymmetry in older adults and people living with stroke with factory settings [41]. This supports the study's appraisal of the Weight Shift Tool as inaccurate for older adults and people living with stroke, due to limits of agreement exceeding the established MDC of 5%. Clinicians require tools that provide accurate and consistent measurements to inform treatment choices and progressions, and to predict outcomes. Limits of agreement greater than an established MDC, and inconsistency in the bias (especially with older adults) indicate that the Weight Shift Tool is too variable for clinical use, as a patient may not, for example, show a real improvement where one exists. An MMC should not have limits of agreement greater than 5%, and the Weight Shift Tool is far from this threshold [32,41].

Limitations

A limitation of this study is the small sample size, which would make the 3-way ANOVA less likely to determine differences and may increase the limits of agreement in the Bland-Altman analysis. As well, it was not possible to perform subgroup analyses, limiting the generalizability of the results. In addition, the study sample is not representative of the entire stroke population due to the nature of the task. However, individuals who did not meet the inclusion criteria (ie, the ability to stand without assistance or a gait aid) would not use this tool clinically. Therefore, the study sample represents only those for whom the Weight Shift Tool is suitable.

Another limitation is that this study did not take into account which arm was being raised by participants, the dominant leg, or the side of stroke for people living with stroke. The arm that a participant raised during the AR trials could have had an impact on the measurement of weight distribution on each foot. Nevertheless, the arm was not raised overhead, but only to 90 degrees of shoulder abduction with the elbow at approximately 90 degrees, as demonstrated by the Weight Shift Tool during use (Figure 2).

Furthermore, it would have been ideal for leans of varying extents to be tested, rather than solely instructing the participants to "lean to the left (or right)." Additionally, different body shapes, weight distribution strategies, and postures could have been compared to determine if variations would affect the estimation of lean. Finally, there is no minimal clinically important difference data available for comparison.

Future Directions

This preliminary study has identified recommendations for the improvement of the Jintronix Weight Shift Tool, as well as direction for more in-depth evaluations of the accuracy and variability of MMC systems for the measurement of weight distribution in standing.

With respect to the Jintronix Weight Shift Tool, developments in the proprietary software, which analyzes the figure to determine weight distribution, and hardware, which captures the figure for software analysis, could offer improvements. In order to be acceptable for clinical use, the removal of statistical bias and reduction of the limits of agreement to 5% would be required. Recommended improvements include algorithmic calibration adjustments and additional model training (possibly including force plate ground truth for improved computational modeling) to increase accuracy for the estimation of lean from the vertical for varying body shapes, postures, and distribution strategies. Tracking additional landmarks in the trunk and including a measurement of height and BMI for calibration may help as well. Additionally, given that the Weight Shift Tool is intended to be used in the presence of a clinician, removing the need for raising the arm may improve the accuracy of estimates.

With respect to future studies, repeated measures may provide a more accurate average weight distribution, particularly for younger adults and people living with stroke when their arm is used to trigger a recording, as well as for people living with stroke when their arm is not raised, since the bias was small in these instances. Moreover, subgroup analyses could be

XSL•FO
RenderX

performed to determine if individual differences, such as BMI and sex, affect results. A trial design establishing the MDC or minimal clinically important difference for weight-bearing on each foot would allow for a more thorough evaluation of the clinical acceptability of the Weight Shift Tool. Future studies on the ability of the Weight Shift Tool to assess change over time could also better inform the use of the Weight Shift Tool as an outcome measure for research.

While the Weight Shift Tool may be able to track changes in weight distribution in real time (ie, in the context of gaming), its accuracy is poor and it cannot be recommended for clinical use. Currently, other non-MMC tools for measuring weight-bearing asymmetry, such as smart insoles, may also help understand the usefulness of the Weight Shift Tool in comparison to other tools [42].

Conclusions

The objectives of this study were to assess the accuracy and variability of the Weight Shift Tool in comparison to a pressure mat, and to determine if the participant group or raising the arm would affect the accuracy or variability. To do so, the percentage of weight on each foot was measured by both tools in younger adults, older adults, and people living with stroke, when leaning to the left, leaning to the right, or with equal weight-bearing, and with an arm raised or no arm raised. The findings suggest that the bias of the Weight Shift Tool compared to the pressure mat may be accurate for healthy younger adults with an arm raised but the Weight Shift Tool is much too variable for clinical use for all tested groups. Unacceptable levels of bias for older adults and high variability for people living with stroke resulted in poor agreement between the Weight Shift Tool and the validated MatScan pressure mat for these populations. Therefore, these preliminary findings do not support the recommendation of the Weight Shift Tool for clinical use for measuring weight distribution in standing in younger adults, older adults, or people living with stroke. The findings of this study are specific to the Jintronix Rehabilitation System Weight Shift Tool with the Microsoft Kinect v2 sensor. Future versions of this MMC may have enhanced capabilities to measure weight distribution among various populations. Therefore, further research would be required to validate MMC systems for measuring weight shift as a clinical assessment and rehabilitation tool.

Acknowledgments

The authors would like to acknowledge all of our participants, the staff at Bruyère Health's inpatient stroke rehabilitation unit, and research assistant Simon Darbhanga, who assisted with data acquisition and processing.

Funding

This work was supported by the Heart and Stroke Foundation of Canada under Grant-in-aid (G-14-0005830), and by a generous personal donation from Tony and Elizabeth Graham.

Conflicts of Interest

None declared.

References

1. Kanko RM, Laende EK, Davis EM, Selbie WS, Deluzio KJ. Concurrent assessment of gait kinematics using marker-based and markerless motion capture. J Biomech 2021 Oct 11;127:110665. [doi: 10.1016/j.jbiomech.2021.110665] [Medline: 34380101]
2. Milosevic B, Leardini A, Farella E. Kinect and wearable inertial sensors for motor rehabilitation programs at home: state of the art and an experimental comparison. Biomed Eng Online 2020 Apr 23;19(1):25 [FREE Full text] [doi: 10.1186/s12938-020-00762-7] [Medline: 32326957]
3. Hellsten T, Karlsson J, Shamsuzzaman M, Pulkkis G. The Potential of Computer Vision-Based Marker-Less Human Motion Analysis for Rehabilitation. Rehabil Process Outcome 2021;10:11795727211022330 [FREE Full text] [doi: 10.1177/11795727211022330] [Medline: 34987303]
4. Lam WW, Fong KN. The application of markerless motion capture (MMC) technology in rehabilitation programs: a systematic review and meta-analysis. Virtual Reality 2023;27(4):3363-3378. [doi: 10.1007/s10055-022-00696-6]
5. McEwen D, Taillon-Hobson A, Bilodeau M, Sveistrup H, Finestone H. Virtual reality exercise improves mobility after stroke: an inpatient randomized controlled trial. Stroke 2014 Jun;45(6):1853-1855. [doi: 10.1161/STROKEAHA.114.005362] [Medline: 24763929]
6. Sheehy L, Finestone H, Bilodeau M, Sveistrup H. Does virtual reality training performed in sitting challenge sitting balance? A comparison of healthy adults vs. stroke patients. In: Int J Stroke. 2015 Presented at: Canadian Stroke Conference; 17 Sept. 2015; Toronto, ON, Canada p. 81. [doi: 10.1111/ijs.12632]
7. Sheehy L, Taillon-Hobson A, Sveistrup H, Bilodeau M, Yang C, Finestone H. Sitting Balance Exercise Performed Using Virtual Reality Training on a Stroke Rehabilitation Inpatient Service: A Randomized Controlled Study. PM R 2020 Aug;12(8):754-765. [doi: 10.1002/pmrj.12331] [Medline: 31970898]

XSL•FO
RenderX

8. Needham L, Evans M, Wade L, Cosker DP, McGuigan MP, Bilzon JL, et al. The development and evaluation of a fully automated markerless motion capture workflow. J Biomech 2022 Nov;144:111338 [FREE Full text] [doi: 10.1016/j.jbiomech.2022.111338] [Medline: 36252308]
9. Stenum J, Rossi C, Roemmich RT. Two-dimensional video-based analysis of human gait using pose estimation. PLoS Comput Biol 2021 Apr;17(4):e1008935 [FREE Full text] [doi: 10.1371/journal.pcbi.1008935] [Medline: 33891585]
10. Vafadar S, Skalli W, Bonnet-Lebrun A, Khalifé M, Renaudin M, Hamza A, et al. A novel dataset and deep learning-based approach for marker-less motion capture during gait. Gait Posture 2021 May;86:70-76. [doi: 10.1016/j.gaitpost.2021.03.003] [Medline: 33711613]
11. Usami T, Nishida K, Iguchi H, Okumura T, Sakai H, Ida R, et al. Evaluation of lower extremity gait analysis using Kinect V2 tracking system. SICOT J 2022;8:27 [FREE Full text] [doi: 10.1051/sicotj/2022027] [Medline: 35748723]
12. Tanaka R, Ishii Y, Yamasaki T, Kawanishi H. Measurement of the total body center of gravity during sit-to-stand motion using a markerless motion capture system. Med Eng Phys 2019 Apr;66:91-95. [doi: 10.1016/j.medengphy.2018.12.020] [Medline: 30797672]
13. Yeung LF, Cheng KC, Fong CH, Lee WC, Tong KY. Evaluation of the Microsoft Kinect as a clinical assessment tool of body sway. Gait Posture 2014 Sep;40(4):532-538. [doi: 10.1016/j.gaitpost.2014.06.012] [Medline: 25047828]
14. Mobini A, Behzadipour S, Saadat M. Test-retest reliability of Kinect's measurements for the evaluation of upper body recovery of stroke patients. Biomed Eng Online 2015 Aug 04;14:75 [FREE Full text] [doi: 10.1186/s12938-015-0070-0] [Medline: 26238199]
15. Çubukçu B, Yüzgeç U, Zileli R, Zileli A. Reliability and validity analyzes of Kinect V2 based measurement system for shoulder motions. Med Eng Phys 2020 Feb;76:20-31. [doi: 10.1016/j.medengphy.2019.10.017] [Medline: 31882393]
16. Langhorne P, Coupar F, Pollock A. Motor recovery after stroke: a systematic review. Lancet Neurol 2009 Aug;8(8):741-754. [doi: 10.1016/S1474-4422(09)70150-4] [Medline: 19608100]
17. Lee CH, Kim Y, Lee BH. Augmented reality-based postural control training improves gait function in patients with stroke: Randomized controlled trial. Hong Kong Physiotherapy Journal 2014 Dec;32(2):51-57. [doi: 10.1016/j.hkpj.2014.04.002]
18. Marques-Sule E, Arnal-Gómez A, Buitrago-Jiménez G, Suso-Martí L, Cuenca-Martínez F, Espí-López GV. Effectiveness of Nintendo Wii and physical therapy in functionality, balance, and daily activities in chronic stroke patients. J Am Med Dir Assoc 2021 May;22(5):1073-1080. [doi: 10.1016/j.jamda.2021.01.076] [Medline: 33639116]
19. Kamphuis JF, de Kam D, Geurts ACH, Weerdesteyn V. Is weight-bearing asymmetry associated with postural instability after stroke? A systematic review. Stroke Res Treat 2013;2013:692137 [FREE Full text] [doi: 10.1155/2013/692137] [Medline: 23738232]
20. Birnbaum M, Brock K, Clark R, Muir S, Burton E, Hill KD. Standing weight-bearing asymmetry in adults with lateropulsion following stroke. Gait Posture 2021 Oct;90:427-433. [doi: 10.1016/j.gaitpost.2021.09.172] [Medline: 34597984]
21. Marigold DS, Eng JJ. The relationship of asymmetric weight-bearing with postural sway and visual reliance in stroke. Gait Posture 2006 Feb;23(2):249-255 [FREE Full text] [doi: 10.1016/j.gaitpost.2005.03.001] [Medline: 16399522]
22. Pollock A, Baer G, Campbell P, Choo PL, Forster A, Morris J, et al. Physical rehabilitation approaches for the recovery of function and mobility following stroke. Cochrane Database Syst Rev 2014 Apr 22;2014(4):CD001920 [FREE Full text] [doi: 10.1002/14651858.CD001920.pub3] [Medline: 24756870]
23. Farthing JP, Zehr EP. Restoring symmetry: clinical applications of cross-education. Exerc Sport Sci Rev 2014 Apr;42(2):70-75. [doi: 10.1249/JES.0000000000000009] [Medline: 24508737]
24. Viteckova S, Kutilek P, Svoboda Z, Krupicka R, Kauler J, Szabo Z. Gait symmetry measures: A review of current and prospective methods. Biomedical Signal Processing and Control 2018 Apr;42:89-100. [doi: 10.1016/j.bspc.2018.01.013]
25. Daley MJ, Spinks WL. Exercise, mobility and aging. Sports Med 2000 Jan;29(1):1-12. [doi: 10.2165/00007256-200029010-00001] [Medline: 10688279]
26. Bohannon AW, Waldron AM. Weightbearing during comfortable stance in patients with stroke: Accuracy and reliability of measurements. Aust J Physiother 1991;37(1):19-22. [doi: 10.1016/S0004-9514(14)60530-4] [Medline: 25026195]
27. Forbes R, Dayal A. How you become the controller. Xbox.com. 2010 Dec 29. URL: https://web.archive.org/web/20110111115024/https://www.xbox.com/en-US/Live/EngineeringBlog/122910-HowYouBecometheController [accessed 2024-11-23]
28. Goetschius J, Feger MA, Hertel J, Hart JM. Validating center-of-pressure balance measurements using the MatScan® pressure mat. J Sport Rehabil 2018 Jan 01;27(1):1-5. [doi: 10.1123/jsr.2017-0152] [Medline: 28714837]
29. Brenton-Rule A, Mattock J, Carroll M, Dalbeth N, Bassett S, Menz HB, et al. Reliability of the TekScan MatScan® system for the measurement of postural stability in older people with rheumatoid arthritis. J Foot Ankle Res 2012 Aug 13;5:21 [FREE Full text] [doi: 10.1186/1757-1146-5-21] [Medline: 22889288]
30. Berg K, Wood-Dauphinee S, Williams JI. The Balance Scale: reliability assessment with elderly residents and patients with an acute stroke. Scand J Rehabil Med 1995 Mar;27(1):27-36. [Medline: 7792547]
31. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986 Feb 08;1(8476):307-310. [Medline: 2868172]

XSL•FO
RenderX

32. Martello SK, Boumer TC, Almeida JCD, Correa KP, Devetak GF, Faucz R, et al. Reliability and minimal detectable change of between-limb synchronization, weight-bearing symmetry, and amplitude of postural sway in individuals with stroke. Res Biomed Eng 2017 Jun 29;33(2):113-120. [doi: 10.1590/2446-4740.06816]
33. Dolatabadi E, Taati B, Mihailidis A. Concurrent validity of the Microsoft Kinect for Windows v2 for measuring spatiotemporal gait parameters. Med Eng Phys 2016 Sep;38(9):952-958. [doi: 10.1016/j.medengphy.2016.06.015] [Medline: 27387901]
34. Gonzalez DRG, Imbiriba LA, Jandre FC. Comparison of body sway measured by a markerless low-cost motion sensor and by a force plate. Res Biomed Eng 2021 Aug 17;37(3):507-517. [doi: 10.1007/s42600-021-00161-4]
35. Dehbandi B, Barachant A, Smeragliuolo AH, Long JD, Bumanlag SJ, He V, et al. Using data from the Microsoft Kinect 2 to determine postural stability in healthy subjects: A feasibility trial. PLoS One 2017;12(2):e0170890 [FREE Full text] [doi: 10.1371/journal.pone.0170890] [Medline: 28196139]
36. Lichtwark GA, Schuster RW, Kelly LA, Trost SG, Bialkowski A. Markerless motion capture provides accurate predictions of ground reaction forces across a range of movement tasks. J Biomech 2024 Mar;166:112051 [FREE Full text] [doi: 10.1016/j.jbiomech.2024.112051] [Medline: 38503062]
37. Simonet A, Fourcade P, Loete F, Delafontaine A, Yiou E. Evaluation of the margin of stability during gait initiation in young healthy adults, elderly healthy adults and patients with Parkinson's Disease: A comparison of force plate and markerless motion capture systems. Sensors (Basel) 2024 May 23;24(11):3322 [FREE Full text] [doi: 10.3390/s24113322] [Medline: 38894112]
38. Clark RA, Pua Y, Oliveira CC, Bower KJ, Thilarajah S, McGaw R, et al. Reliability and concurrent validity of the Microsoft Xbox One Kinect for assessment of standing balance and postural control. Gait Posture 2015 Jul;42(2):210-213. [doi: 10.1016/j.gaitpost.2015.03.005] [Medline: 26009500]
39. Otte K, Kayser B, Mansow-Model S, Verrel J, Paul F, Brandt AU, et al. Accuracy and reliability of the Kinect version 2 for Ccinical measurement of motor function. PLoS One 2016;11(11):e0166532 [FREE Full text] [doi: 10.1371/journal.pone.0166532] [Medline: 27861541]
40. Clark RA, Pua Y, Bryant AL, Hunt MA. Validity of the Microsoft Kinect for providing lateral trunk lean feedback during gait retraining. Gait Posture 2013 Sep;38(4):1064-1066. [doi: 10.1016/j.gaitpost.2013.03.029] [Medline: 23643880]
41. Liuzzo DM, Peters DM, Middleton A, Lanier W, Chain R, Barksdale B, et al. Measurements of weight bearing asymmetry using the nintendo wii fit balance board are not reliable for older adults and individuals with stroke. J Geriatr Phys Ther 2017;40(1):37-41 [FREE Full text] [doi: 10.1519/JPT.0000000000000065] [Medline: 26288237]
42. Avcı EE, Akgün G, Uygur M, Polat M, Demirbüken İ. Improving partial weight bearing compliance with a smart insole: Validity, reliability, and feasibility study. Foot Ankle Surg 2025 Oct;31(7):645-651. [doi: 10.1016/j.fas.2025.04.007] [Medline: 40280798]

Abbreviations

AR: arm raise
MDC: minimal detectable change
MMC: markerless motion capture
NAR: no arm raise

Edited by J Sarvestan; submitted 07.03.25; peer-reviewed by A Borghese, A Kerr; comments to author 09.06.25; accepted 15.11.25; published 02.12.25.

Please cite as:
Sheehy L, Gal-Dev E, Sveistrup H, Bilodeau M, Finestone H
Accuracy and Variability of a Commercial Markerless Motion Capture System Compared to a Pressure Mat for Weight Distribution in Standing: Cross-Sectional Observational Study
JMIR Form Res 2025;9:e73575
URL: https://formative.jmir.org/2025/1/e73575
doi:10.2196/73575
PMID:

©Lisa Sheehy, Emma Gal-Dev, Heidi Sveistrup, Martin Bilodeau, Hillel Finestone. Originally published in JMIR Formative Research (https://formative.jmir.org), 02.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Depression and Anxiety Outcomes in a Technology-Enabled Psychotherapy Practice: Retrospective Cohort Study

Nicholas R Forand[1*], PhD; Jasmine Nettiksimmons[1*], PhD; Margaret Anton[1*], PhD; Raven Truxson[2*], BS, MPH; Karl Vanderwood[2*], MPH, PhD; Brandn Green[2*], PhD

[1]Two Chairs, San Francisco, CA, United States

[2]JG Research & Evaluation, Bozeman, MT, United States

[*]all authors contributed equally

Corresponding Author:
Nicholas R Forand, PhD
Two Chairs
530 Washington Street
San Francisco, CA, 94111
United States
Phone: 1 415 202 5159
Email: nforand@twochairs.com

Abstract

Background: Mental health conditions account for significant distress, burden, and societal costs. Despite efforts to implement evidence-based practices (EBPs), access to high-quality mental health treatment in general practice remains limited and clinical outcomes suboptimal. Measurement-based care (MBC) is a transtheoretical and transdiagnostic strategy that has the potential, when implemented effectively, to improve the quality of care. Digital tools can also support clinicians by alleviating administrative tasks and providing in-the-moment performance data and clinical decision support.

Objective: This study aimed to examine patient retention in care and depression and anxiety outcomes in a technology-enabled psychotherapy practice where clinicians were supported by a suite of innovations, including an MBC platform and clinical decision support tools.

Methods: This retrospective cohort study examined 2984 adults who initiated mental health treatment with Two Chairs, a hybrid technology–enabled behavioral health provider, from January 1 to June 30, 2024. Rates of reliable change, recovery, remission, deterioration, and the magnitude and trajectory of symptom change in depression and anxiety symptoms were assessed using the Patient Health Questionnaire-9 (PHQ-9) and General Anxiety Disorder-7 (GAD-7).

Results: The participants demonstrated high rates of retention in care (3183/3572, 89.1%) and MBC survey completion (3440/3572, 96.3%). From baseline to the 12th session, patients showed significant symptom improvements in depression and anxiety, achieving high rates of reliable improvement (65.8%, 95% CI 64.0%-67.5%) and recovery (53.2%, 95% CI 51.4%-55.0%), with 71.3% (95% CI 69.7%-72.9%) of patients achieving either outcome. Outcomes continued to improve up to the point of completion of the treatment episode (79.2%, 95% CI 77.5%-80.9%, of patients achieving improvement or recovery). Pre- to posttreatment effect sizes were large (all Cohen d>0.9).

Conclusions: The results of this study show that technology-enabled MBC and clinical decision support systems (CDSSs) may support high-quality patient outcomes in mental health.

(JMIR Form Res 2025;9:e76264) doi:10.2196/76264

KEYWORDS

depression; anxiety; mental health; mental health care; measurement-based care; psychotherapy

XSL•FO
RenderX

Introduction

Background

Mental health conditions are a significant public health concern, contributing substantially to the global burden of disease [1]. The increased prevalence of mental health disorders has a clear relationship with disability, reduced quality of life, and associated increases in avoidable economic costs, including costs incurred by payers while managing the care of individuals with these conditions [2,3]. The economic burden of mental illness is high, affecting both direct health care expenditures and indirect costs connected to loss of productivity in the workforce and increased demand for social supports. An analysis of the change in the economic burden of major depressive disorder between 2010 and 2018 describes a worldwide increase from US $236.6 billion to US $326.2 billion [2]. These broad and significant personal, social, and economic implications underscore the urgency of addressing mental health challenges [4].

Access to Quality Care

Despite a substantial increase in the percentage of US adults receiving mental health services over the past 25 years, access to effective, high-quality mental health care remains problematic [4]. Studies examining early termination of psychotherapy treatment suggest that a significant portion of patients drop out of therapy prematurely [5,6]. One study found that 73.5% of patients in a health system setting dropped out of care prior to their fourth session, suggesting that the majority of patients received an inadequate dose of treatment [6]. High levels of treatment dropout have been attributed to low therapeutic alliance and dissatisfaction with the quality of care [7].

A systematic review of usual care practices highlighted that many mental health interventions also lack a structured, evidence-based approach, leading to inconsistent treatment delivery and variable patient outcomes [8,9]. Outcomes in regular care practices reflect these problems, with rates of improvement being up to 20%-30% lower in these settings than in clinical trials of structured, evidence-based interventions [10-12].

Although there has been a growing recognition of mental illness as a critical health issue, the quality of care provided to individuals with mental disorders remains suboptimal due to limited quality standards, weak therapeutic alliance, and underuse of evidence-based approaches, all of which further contribute to the societal burden of mental illness.

Barriers to Effectively Implementing Evidence-Based Practices

Evidence-based practices (EBPs) are known to improve patient outcomes; however, the implementation of these practices in mental health care can be hindered by cost, complexity, and challenges in sustaining the new clinical model [13,14]. Although standard models of scaling EBPs can be effective, they often require significant financial investment, clinician support, and logistical coordination, limiting their long-term viability [13]. Institutions interested in high-quality care require effective EBPs, as well as efficient and cost-effective means to support clinicians in delivering these practices [15].

Potential of Measurement-Based Care

Measurement-based care (MBC), sometimes referred to as routine outcome monitoring or feedback-informed care, is an EBP that includes the collection, sharing, and use of patient-reported data to collaboratively guide clinical decision-making [16,17]. When implemented effectively, MBC has been demonstrated to improve clinical outcomes [17]. Unlike other evidence-based mental health interventions that are often designed to target a specific clinical diagnosis or concerns in targeted populations, MBC is both transtheoretical and transdiagnostic, suggesting that it is an efficient method for improving outcomes across diverse groups of patients and clinicians. Research indicates that MBC enhances therapy outcomes by enabling clinicians to identify and address care delivery and alliance challenges in real time, leading to reduced dropout [18], improved therapeutic alliance, faster symptom resolution, and a more personalized, effective approach than standard care [17].

Despite these advantages, the implementation of MBC also faces challenges. At the clinician level, the time and administrative burden associated with selecting, delivering, scoring, and tracking scores over time may inhibit uptake [19]. Without support, clinicians may also struggle with generating insights from MBC data or tracking and identifying trends. At the organizational level, leadership alignment, funding, the implementation of new technology, training, and supervision are cited as limitations to adoption [19]. Barriers to the implementation of MBC may be reflected in the rates of clinicians who use these tools. Less than 20% of clinicians indicate using MBC in their practice [20], and even when measures are prescribed by the clinical service, rates of survey completion can be as low as 15% [21].

Technology as a Solution to MBC Barriers

Clinical decision support systems (CDSSs) and other clinical dashboard applications offer technology-enabled solutions to address many of the aforementioned challenges, including the implementation of MBC [16,17]. A CDSS is a form of technology supporting care that can be integrated into health care settings. It can improve patient outcomes and relieve the clinician burden by providing clinicians with evidence-based recommendations at the point of care [20]. CDSSs analyze patient data and suggest appropriate interventions and follow-up, thereby enhancing clinical efficiency and reducing clinical errors [18]. By integrating these systems, clinicians are able to make informed decisions that align with the latest clinical guidelines, leading to improved patient outcomes [20].

Technology can also alleviate administrative tasks, allowing clinicians to focus more on patient care [19]. Additionally, digital tools can facilitate real-time performance feedback, enhancing visibility and accountability across all levels of care [21]. Clinicians can monitor patient progress, symptom changes, and treatment responses [22]. In some instances, patients can also track their own progress, empowering them in their own care [22]. Systematic collection and feedback of data can also

XSL•FO
RenderX

be leveraged for quality improvement efforts, such as monitoring clinician or practice performance and identifying areas for remediation or professional development [22]. Across this broad set of supports, technological innovations in care delivery, patient tracking, and analysis have demonstrated the capacity to increase sustainability of services through enhanced quality and patient engagement, while also decreasing the administrative burden associated with documenting the provision of care. MBC paired with CDSS support may provide a means to improve the efficiency and sustainability of EBP.

Despite the promise of this approach, technology-enabled psychotherapy is not without potential limitations. Design and usability challenges are common in the technology space. When a digital CDSS tool is not perceived by users to be usable, valid, and useful in clinical practice, it can limit adoption or engagement and may even add an additional administrative burden to the workforce [23-25]. CDSSs are more likely to be usable and acceptable if designed with clinician input or "user-centered design" processes. Thoughtfully co-designed CDSS tools can help more effectively realize the benefits of technology-enabled clinical practice.

Aims of This Study

Technology-supported psychotherapy practice, featuring MBC supported by a suite of CDSS tools, may improve the scalability of high-quality care in general psychotherapy practice. Although the literature provides many examples of how technology may support the practice of MBC, many of these studies are confined to randomized controlled trials (RCTs) or other limited implementations or focus primarily on the effectiveness and practice of MBC [26]. This study aimed to describe a broad suite of CDSSs that have been designed and implemented to support clinicians in general practice in the delivery of MBC and other aspects of care. To evaluate the general outcomes and clinical processes associated with practice using this suite of tools, the study reported on the following markers of clinical quality: (1) patient retention in care (ie, reduced dropout); (2) patient and clinician adherence to MBC practices; and (3) clinical outcomes, including depression and anxiety symptom reduction for individuals who enter care with clinically elevated symptoms. The study used a retrospective cohort design and included patients who engaged in weekly outpatient psychotherapy, either virtually or in person. The study's findings will contribute to the limited literature on the real-world use of technology-supported MBC practice at scale.

Methods

Ethical Considerations

The need for formal ethical approval was waived by the Sterling Institutional Review Board (IRB), Atlanta, Georgia. This research was deemed exempt under Department of Health and Human Services (DHHS) Category 4 as retrospective research on an existing and de-identified dataset. The IRB protocol number is 12670-NForand.

Study Design and Study Setting

This was a retrospective cohort analysis focused on adult patients with clinically elevated symptoms of depression or anxiety who enrolled in mental health treatment at Two Chairs in the first half of 2024. Two Chairs is a technology-enabled behavioral health company founded in 2017 that provides psychotherapy to self-paying individuals, commercially insured (health maintenance organizations [HMOs], preferred provider organizations [PPOs]) individuals, and publicly insured (Medicare, Medicaid) individuals. During the data collection period, services were available in California, Washington, and Florida.

Participants

The following details describe general facts about the practice setting. Adult patients (≥18 years old) are enrolled in behavioral health services at Two Chairs and receive outpatient psychotherapy for anxiety, depression, and related conditions. At intake, the patients complete a digital intake assessment and attend an initial clinical interview, after which they are assigned a preliminary diagnosis and subsequently matched with an individual clinician. During onboarding, 98% of all patients are screened for depression and anxiety using the Patient Health Questionnaire-9 (PHQ-9) [27] and General Anxiety Disorder-7 (GAD-7) [28], and 48% and 45% of patients meet the clinical cutoff of ≥10 on these measures, respectively. Of those with a positive screen on either measure, 84% start care within 30 days. Of patients who receive a diagnosis of a depressive disorder, 99% have at least one follow-up PHQ-9 assessment, and of those who receive a diagnosis of anxiety disorder, 99% have at least one follow-up GAD-7 assessment, indicating a high level of follow-up assessment.

The study required patients to have a first treatment session between January 1 and June 30, 2024. Patients were eligible for the study if they had clinical levels of depression or anxiety at their baseline measurement, as evidenced by a PHQ-9 or a GAD-7 score of ≥10. Using symptom cutoffs is common in similar studies of general practice [29] and ensures that patients are likely to have clinically elevated levels of distress and are therefore in need of treatment. Patients reporting active suicidal ideation or otherwise deemed inappropriate for outpatient psychotherapy at intake were referred to health plan or community resources and not included in this study. The cutoff for data included in this study was February 28, 2025 regardless of whether a patient had completed their care by that date. A total of 3572 patients met the initial eligibility criteria.

Patients were excluded from the primary analytic cohort if they terminated from care early (defined as fewer than four sessions), missed a symptom assessment at session 1, or lacked any follow-up symptom assessments between sessions 4 and 12. The cutoff for inclusion of at least four attended sessions was used to ensure the sample had an adequate dose of therapy. The four-session cutoff has been used previously in other large studies of naturalistic care [30,31]. To ensure a comprehensive report of the outcomes for the full cohort, the outcomes of individuals who dropped out prior to four sessions were included as a separate analysis. After these criteria were applied, the primary analytic cohort comprised 2984 (83.5%) patients. Within the primary cohort, inclusion criteria for PHQ-9 outcomes (baseline PHQ-9 score≥10) were met by 2465 (82.6%) patients, and inclusion criteria for GAD-7 outcomes (baseline

XSL•FO
RenderX

GAD-7 score≥10) were met by 2204 (73.9%) patients. A secondary analysis focused on the subset of patients from the primary cohort who had completed their course of care during this time frame (n=2237, 75%; PHQ-9 n=1812, 81%; GAD-7 n=1689, 75.5%).

The Two Chairs Teletherapy Intervention

Patients enrolled at Two Chairs receive weekly 45-55-minute teletherapy or in-person mental health services delivered by licensed clinicians. Approximately 99% of the sessions are conducted via teletherapy. Clinicians use a range of treatment modalities to address patients' needs, including but not limited to cognitive behavioral therapy, acceptance and commitment therapy, time-limited dynamic psychotherapy, and somatic therapy. As part of their practice, clinicians are required to use MBC to evaluate patients' symptom presentation, make care decisions, and track progress in care. Prior to each session, patients complete measures of depression (PHQ-9) and anxiety (GAD-7) symptoms, mental health quality of life (MHQoL), and therapeutic alliance. All symptom data are derived from the regular MBC assessments. For these analyses, only the PHQ-9 and GAD-7 results are reported.

Following standard user-centered design principles, all CDSS tools provided by Two Chairs have been co-designed with user input and tested and revised before launching to the full clinical population. Satisfaction with clinical tools is assessed quarterly and these data are used to drive software and tooling improvements. The product and clinical teams engage in regular review and revision of these tools to ensure they are meeting user needs and adhering to clinical best practices. The technological supports offered by Two Chairs include (1) automated electronic MBC questionnaire delivery to patients; (2) automated scoring and interpretation of patient-reported symptoms over time; (3) clinician alerting for potential safety and risk concerns, such as endorsement of suicidal ideation; and (4) clinical decision support features, including indications of clinically meaningful symptom changes or changes in alliance, standardized recommendations for clinician actions, alerts related to patient-endorsed suicidal ideation, and holistic summaries of change across all measures (see Figures 1-4). Patients also provide a summary of their symptoms, along with highlights of meaningful symptom change, to allow them to reflect on their care progress and feel empowered to engage in care decisions. Clinicians have access to patient goals, diagnoses, and number of sessions on each patients' individual page within the platform. All clinical data are graphically displayed prominently on the home page for each patient on Two Chairs' software platform.

Figure 1. Image of the Two Chairs clinical dashboard for tracking patients.

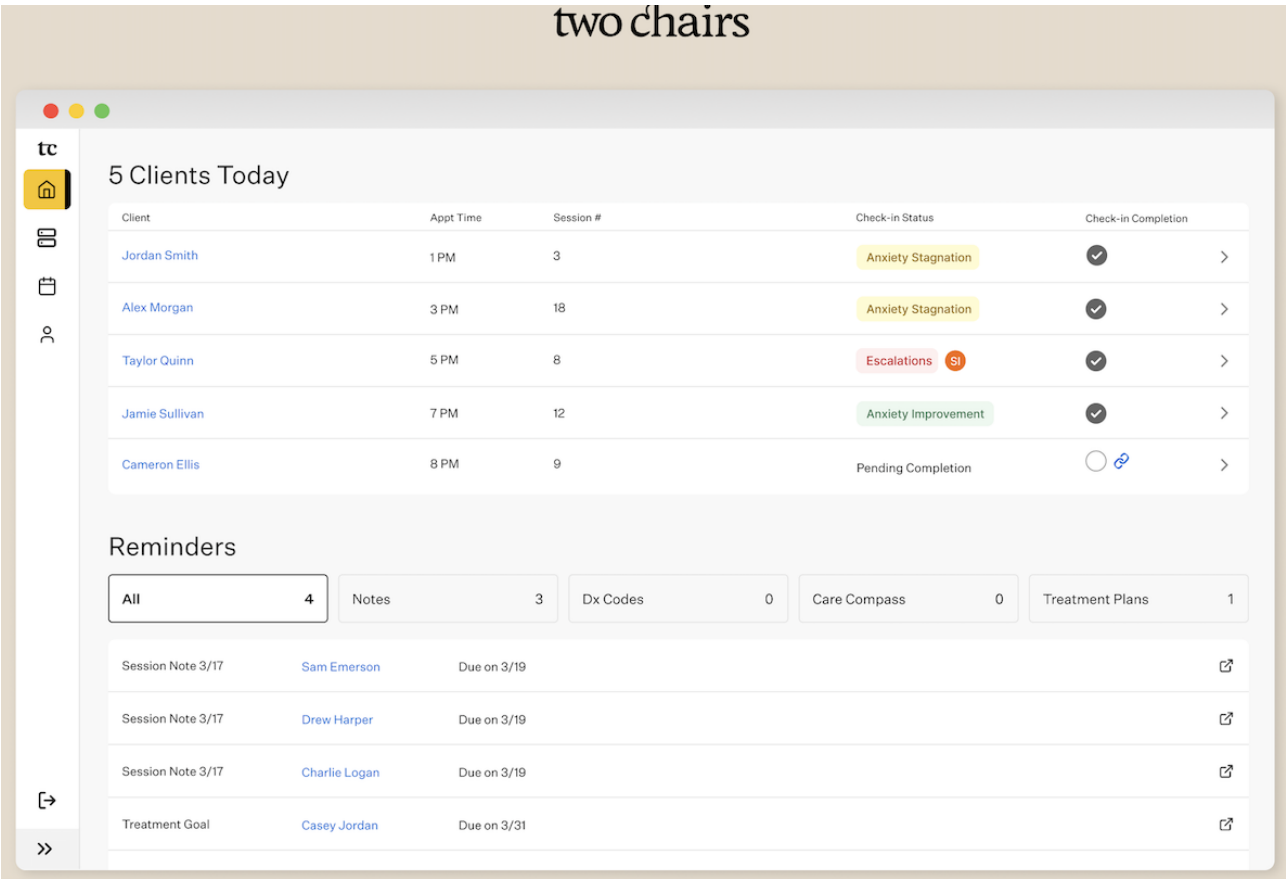

XSL•FO
RenderX

Figure 2. View of GAD-7 scores on the Two Chairs patient page. GAD-7: General Anxiety Disorder-7.

Figure 3. View of PHQ-9 scores on the Two Chairs clinician patient page. PHQ-9: Patient Health Questionnaire-9.

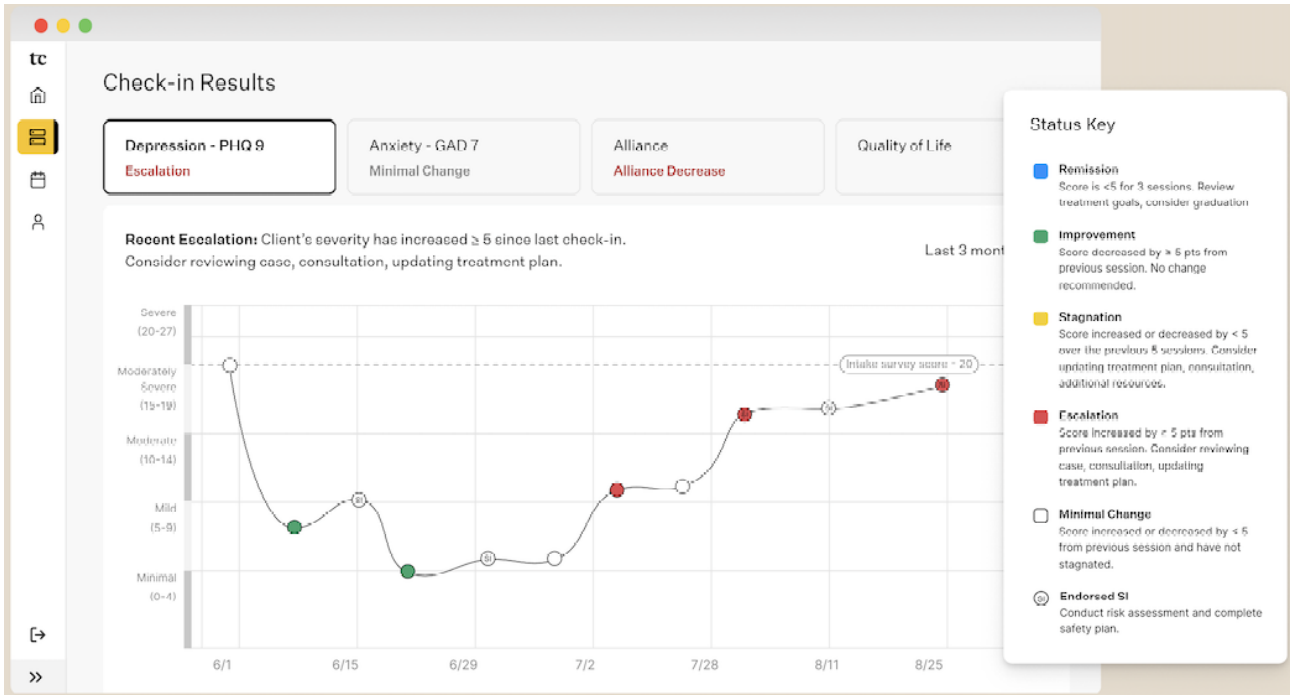

XSL•FO
RenderX

Figure 4. View of patient-facing MBC feedback and Two Chairs clinician-facing tabular view. MBC: measurement-based care.

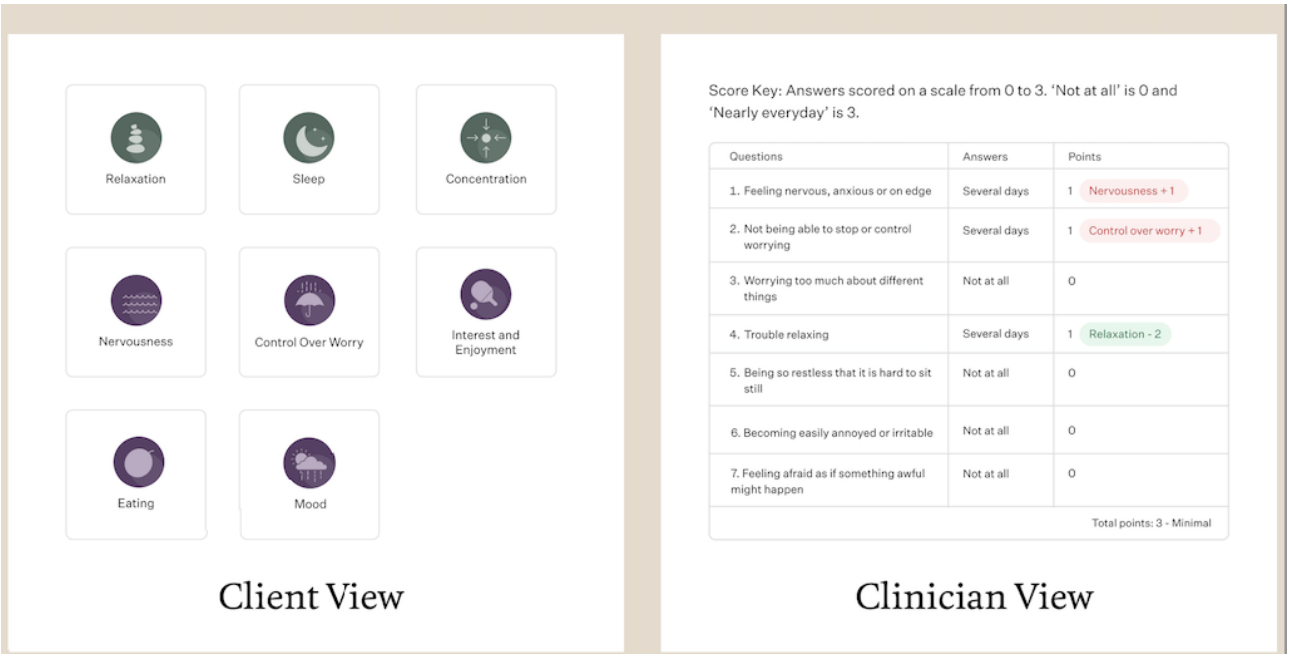

Upon onboarding to Two Chairs, all clinicians receive a full digital training series on the use of MBC in practice, and orientation to Two Chairs tools. Clinicians are trained to integrate patient reported MBC data into care and engage in collaborative decision-making with patients around their treatment progress. The full training package is described in a separate paper [32]. Adherence to MBC practices are tracked by the software system and are part of clinician performance reviews.

Measurement and Variables

Patient Demographics

As part of the onboarding process at Two Chairs, patients provided basic demographic information (eg, age, race/ethnicity, gender, employment status). Data were collected as part of routine care, and patients were not required to complete all questions. As a result, data completeness varied.

Patient Retention

Patient retention was evaluated as the proportion of patients who met the initial study inclusion criteria and were retained until session 4, qualifying for inclusion in the primary analytic cohort.

MBC Completion

MBC completion was assessed as the proportion of observed sessions (up to 12) in which MBC assessments were completed across the entire included population. To assess how well individual clinicians adhered to MBC procedures, we also determined the average completion rates of MBC surveys for patients included within each clinician's panel. The result was a clinician-level assessment of MBC completion for all of their included patients.

Measures

Patients completed the PHQ-9 and GAD-7 at each session throughout care. The PHQ-9 is a 9-item self-report measure of depressive symptoms. Items are rated on a scale ranging from 0 (not at all) to 3 (nearly every day). Total scores range from 0 to 27, and cutoff scores for mild, moderate, moderately severe, and severe depressive symptoms are 5, 10, 15, and 20, respectively. A reliable change on the PHQ-9 is a change of ≥5 in the PHQ-9 score [33]. The GAD-7 is a 7-item self-report measure of anxiety. Items are rated on a scale ranging from 0 (not at all) to 3 (nearly every day). Total scores range from 0 to 21, and the cutoff scores for mild, moderate, and severe anxiety symptoms are 5, 10, and 15, respectively. A reliable change on the GAD-7 is a change of ≥4 in the GAD-7 score [34].

Outcomes

Clinical Outcomes

The number of sessions delivered at Two Chairs varies based on patient presentation; therefore, we used the 12th session of care as a common end point in order to standardize outcomes measurement for the primary analyses. The use of the 12th session as a standard end point in open-ended therapy models is consistent with other studies [11,35]. Baseline clinical scores were derived from the first therapy session. Clinical outcomes were described at two time points: the clinical "check-in" session (defined as session 12 or the last available session between sessions 4 and 11 with a symptom assessment) and termination (defined as the last available symptom measurement for patients who completed their course of care). The latter sample included patients who completed their course of care at any session prior to February 28, 2025. Outcomes were also presented for those patients who dropped out prior to the fourth session, provided those patients met all other inclusion criteria and had at least two observations (n=213).

This study focused on characterizing depression and anxiety symptom improvement using both joint outcomes for the PHQ-9 and GAD-7 and independent outcomes for the PHQ-9 and GAD-7. The goal of the joint measures was to holistically

XSL•FO
RenderX

evaluate change in the commonly comorbid syndromes of depression and anxiety [36], similar to methods used in other settings where these measures are used concurrently [33]. Patients were included in the summaries of joint outcomes if either the PHQ-9 or the GAD-7 baseline score was ≥10. For assessment-specific outcomes (PHQ-9 or GAD-7), inclusion required a baseline score of ≥10 on the given assessment.

Independent Outcomes

Regarding assessment-specific outcomes, *reliable improvement* was defined as an improvement of 5 points on the PHQ-9 or 4 points on GAD-7; *reliable deterioration* was defined as a worsening of 5 points on the PHQ-9 or 4 points on GAD-7; *recovery* was defined as a follow-up score of <10; *remission* was defined as a follow-up score of <5; and *response* was defined as a 50% or greater improvement in the score from baseline.

Joint Outcomes

For joint measures, reliable improvement was defined as reliable improvement on either the PHQ-9 or GAD-7, without reliable deterioration on the other measure. Joint reliable deterioration was defined as reliable deterioration on at least one measure, regardless of a change in the other measure. Recovery was defined as a PHQ-9 and GAD-7 score of <10 at follow-up. Clinically meaningful improvement was defined as meeting either joint recovery or joint reliable improvement criteria. This holistic measure of change across both depression and anxiety encompasses both magnitude of change and symptom severity at the time of measurement.

Statistical Analysis

The study population was characterized using means (SDs) and frequencies with statistical tests to evaluate differences between the primary cohort and the population that was excluded from analysis (*t* tests, chi-squared tests, Fisher exact tests for cell counts<20).

PHQ-9 and GAD-7 scores were measured at baseline, the check-in session, and termination with means (SDs). Change from baseline on the PHQ-9 and GAD-7 was described by the mean difference, 95% CIs, and Cohen d. Summaries of dichotomous outcomes (reliable improvement, recovery, meaningful improvement, remission, and response) were described with frequencies and 95% CIs.

Longitudinal analysis of the PHQ-9 and GAD-7 was performed using linear mixed effects models with intercepts for clinicians and random intercepts and slopes (session numbers) for patients. Fixed effects included age, gender, the session number, and a quadratic term for the session number (see Multimedia Appendix 1). Average effects across sessions were estimated with the most commonly occurring level of age and gender. There were no missing values for age. Missing entries for gender were addressed with an explicit parameter for gender missingness. Missing PHQ-9 or GAD-7 data at the session level were handled within the modeling framework using maximum likelihood estimation.

All analyses were performed using R version 4.4.2 (2024-10-31; R Foundation for Statistical Computing). All data used in the study were extracted from the Two Chairs electronic health record system. Study authors used the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) checklist to ensure complete methodological and scientific content was provided in the manuscript for an observational study. The checklist is included in Multimedia Appendix 2.

Results

Patient Characteristics

Figure 5 shows the CONSORT (Consolidated Standards of Reporting Trials) guidelines–based diagram on study participant inclusion. Of the 3572 patients who met the initial criteria for inclusion, 89.9% (n=3183) were retained for the fourth session and were thus eligible for the primary analytic cohort. A small number of patients were excluded due to missing follow-up MBC scores or because their session 1 baseline was unavailable, leaving 2984 patients (83.5% of the population meeting the original inclusion criteria; Figure 5).

XSL•FO
RenderX

Figure 5. CONSORT (Consolidated Standards of Reporting Trials) diagram on sample size and study participant inclusion. GAD: general anxiety disorder; MBC: measurement-based care; PHQ: Patient Health Questionnaire.

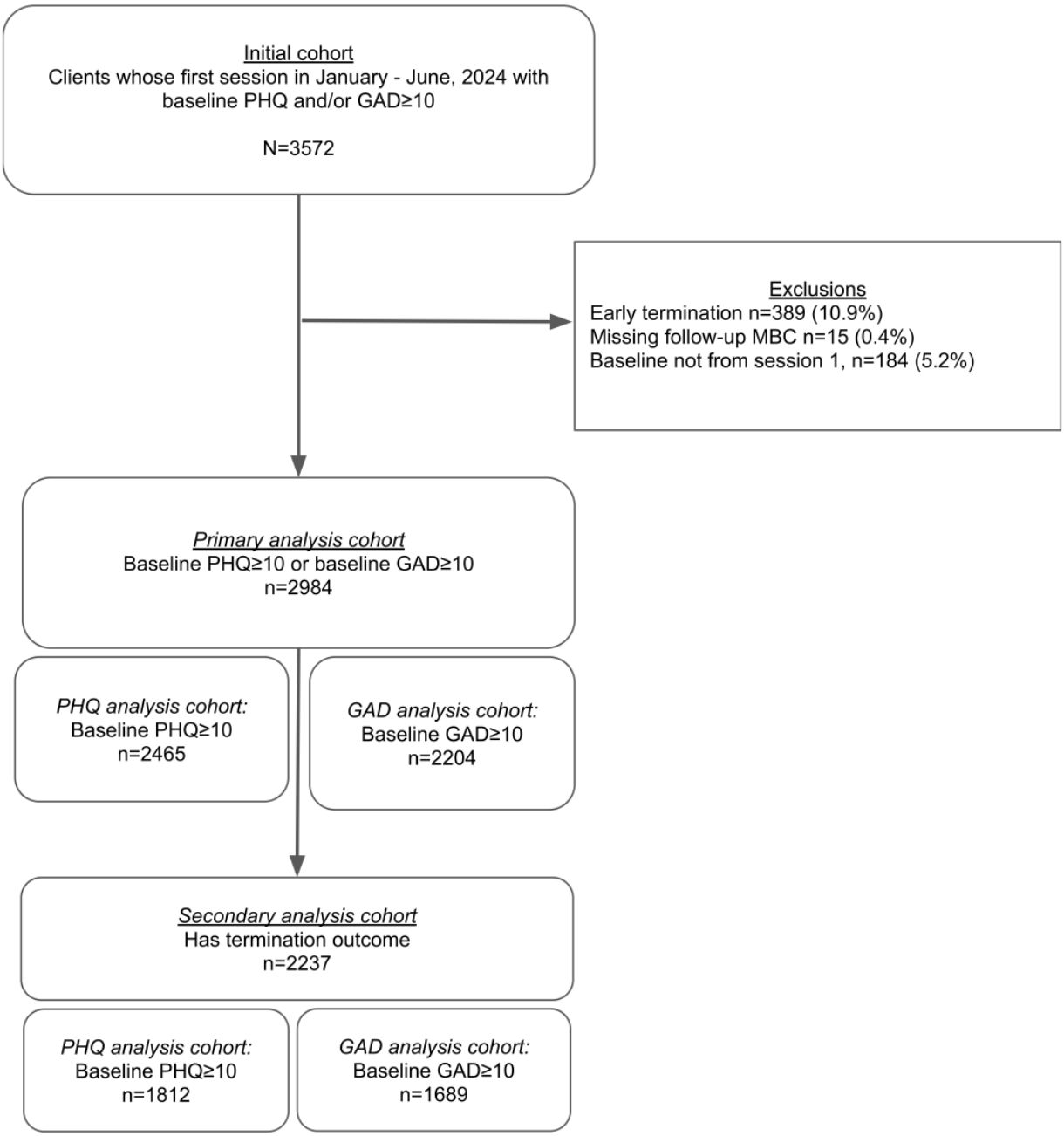

Patient demographics can be found in Table 1. Data for the clinical check-in session were most often collected from session 12 (n=2311, 71.4%) and the average check-in session number of 10.8. A total of 2237 (75%) patients from the primary cohort terminated care during the observation window and therefore had termination outcomes. Of the patients who completed their care, 38.1% (n=852) did so by session 12 and 61.9% (n=1385) did so after session 12. The median termination session was 16 (average 16.9).

XSL•FO
RenderX

Table 1. Demographics and baseline characteristics of patients included and excluded in the primary cohort.

Characteristics	Patients included (n=2984)	Patients excluded (n=588)	*P* value
Age (years), n (%)			
18-29	774 (25.9)	134 (22.8)	.05
30-49	1549 (51.9)	295 (50.2)	.05
50-69	545 (18.3)	127 (21.6)	.05
≥70	116 (3.9)	32 (5.4)	.05
Gender, n (%)			
Woman	1793 (60.1)	361 (61.4)	.006
Man	695 (23.3)	123 (20.9)	.006
Unknown	325 (10.9)	85 (14.5)	.006
Trans/nonbinary/other	171 (5.7)	19 (3.2)	.006
Race and ethnicity, n (%)			
Non-Hispanic White	1297 (43.5)	231 (39.3)	.23
Hispanic	573 (19.2)	118 (20.1)	.40
Unknown race/ethnicity	544 (18.2)	141 (24.0)	.001
Asian	371 (12.4)	59 (10.0)	.20
Black	273 (9.1)	61 (10.4)	.26
Other race/ethnicity	250 (8.4)	47 (8.0)	.99
American Indian/Alaska Native	80 (2.7)	18 (3.1)	.49
Asian Pacific Islander	74 (1.5)	9 (1.5)	.23
Primary payer, n (%)			
HMO[a]	2316 (77.6)	408 (69.4)	<.001
Public (Medicare, Medicaid)	478 (16.0)	123 (20.9)	<.001
Commercial insurance (PPO[b])	142 (4.8)	38 (6.5)	<.001
Self-pay	48 (1.6)	19 (3.2)	<.001
Employment status, n (%)			
Full-time employee	1615 (54.1)	293 (49.8)	.01
Other	774 (25.9)	139 (23.6)	.01
Unknown	315 (10.6)	84 (14.3)	.01
Disabled/unemployed	280 (9.4)	72 (12.2)	.01
Diagnosis category, n (%)			
Anxiety and fear-related disorders	990 (33.2)	178 (30.3)	<.001
Depressive disorders	971 (32.5)	176 (29.9)	<.001
Trauma- and stressor-related disorders	422 (14.1)	61 (10.4)	<.001
Adjustment disorders	289 (9.7)	62 (10.5)	<.001
Other or primary diagnosis unknown	312 (10.5)	111 (18.9)	<.001
Baseline score, mean (SD)			
PHQ-9[c]	13.2 (4.6)	13.3 (5.1)	.70
GAD-7[d]	12 (4.2)	12.6 (4.3)	.001

[a]HMO: health maintenance organization.
[b]PPO: preferred provider organization.
[c]PHQ-9: Patient Health Questionnaire-9.
[d]GAD-7: General Anxiety Disorder-7.

XSL•FO
RenderX

MBC Adherence

Patients in the primary cohort completed the Two Chairs MBC assessments 96.3% (3440/3572) of the time, leading up to the clinical check-in session, which provided near universal weekly data on depression, anxiety, the quality of life, and therapeutic alliance. Among clinicians, the median within-clinician patient adherence to MBC assessments was 96% (IQR 92%-99%). The distribution of within-clinician MBC adherence can be seen in Figure 6.

Figure 6. MBC adherence by clinician. MBC: measurement-based care.

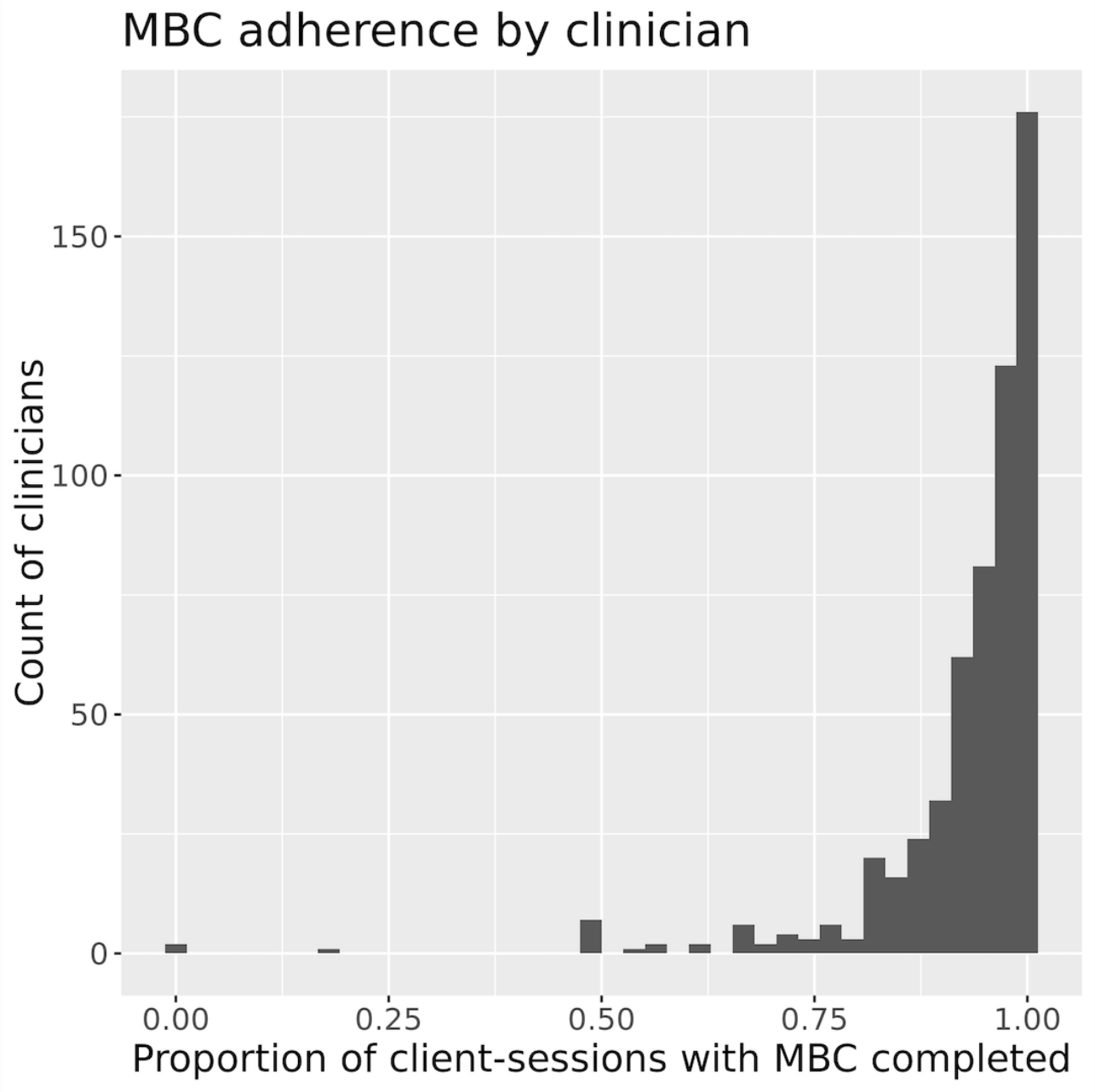

Within the primary analytic cohort, the population was typically between the ages of 30 and 49 years, with slightly more females, was more commonly employed, and more commonly had an HMO as a payer source. Although 18.2% (n=544) had an unknown race or ethnicity, the sample was notable for the racial diversity represented, with just 43.5% (n=1297) self-identifying as non-Hispanic White (Table 1). The excluded population was more likely to have an unknown/unspecified gender, race, and employment status—all features collected from an optional patient profile—and was more likely to have public insurance and primary diagnoses other than anxiety disorders, depressive disorders, and adjustment disorders. Table 2 shows changes in depressive and anxiety symptoms from baseline to the clinical check-in session (for all individuals included in the primary cohort) and changes from baseline to termination (for individuals included in the secondary cohort). At both the check-in assessment and termination, Two Chairs patients showed improvements with large effect sizes across both cohorts and symptom measures.

XSL•FO
RenderX

Table 2. Change in MBC[a] scores from baseline to check-in (primary cohort) and termination (secondary cohort).

Measure	Baseline	Clinical check-in session				Termination session			
	Mean (SD)	Patients (N=2984), n (%)	Mean (SD)	Change (95% CI)	Cohen d	Patients (N=2237), n (%)	Mean (SD)	Change (95% CI)	Cohen d
PHQ-9[b]	14.5 (3.8)	2465 (82.6)	9.5 (5.5)	–5 (–5.3 to –4.8)	0.92	1812 (81.0)	7.9 (5.7)	–6.6 (–6.9 to –6.3)	1.14
PHQ-9 change (%)	__[c]	—	—	–33.9% (–35.4% to –32.5%)	0.94	—	—	–45.0% (–46.8% to –43.3%)	1.2
GAD-7[d]	13.8 (3.1)	2204 (73.9)	8.7 (5.1)	–5.1 (–5.3 to –4.9)	1.03	1689 (75.5)	7.2 (5.2)	–6.6 (–6.8 to –6.3)	1.25
GAD-7 change (%)	—	—	—	-36.6% (-38.1% to –35.1%)	1.04	—	—	-47.3% (-49.0% to –45.5%)	1.31

[a]MBC: measurement-based care.

[b]PHQ-9: Patient Health Questionnaire-9.

[c]Not applicable.

[d]GAD-7: General Anxiety Disorder-7.

From baseline to the most recent check-in session, patients showed, on average, a 5-point (Cohen d=0.92), 34% (Cohen d=0.94) decrease in PHQ-9 scores and a 5.1-point (Cohen d=1.03), 37% (Cohen d=1.04) decrease in GAD-7 scores. From baseline to termination, patients showed, on average, a 6.6-point (Cohen d=1.14), 45% (Cohen d=1.2) decrease in PHQ-9 scores and a 6.6-point (Cohen d=1.25), 47% (Cohen d=1.31) decrease in GAD-7 scores.

Table 3 shows dichotomous outcomes at the clinical check-in and termination sessions and, where available, across assessment-specific and joint measures of achievement. At the clinical check-in session (typically session 12), 71.3% (95% CI 69.7%-72.9%) of patients achieved clinically meaningful improvement, defined as joint recovery in both anxiety and depression or joint reliable improvement, with improvement in either anxiety or depression, without deterioration on the other. Joint reliable deterioration occurred in 7.9% (95% CI 7.0%-8.9%) of patients. Among patients with a termination outcome available, 79.2% (95% CI 77.5%-80.9%) experienced clinically meaningful improvement, with 6% (95% CI 5%-7%) experiencing reliable deterioration. The majority of patients who met the baseline criteria for PHQ-9 outcomes experienced reliable improvement (53%, 95% CI 51.1%-55.0%) and recovery (55.1%, 95% CI 53.2%-57.1%) on PHQ-9 at the clinical check-in session. These numbers improved to 65.3% (95% CI 63.1%-67.5%) and 66.6% (95% CI 64.4%-68.7%), respectively, among patients with available termination sessions. The majority of patients who met the baseline criteria for GAD-7 outcomes experienced reliable improvement (64.4%, 95% CI 62.4%-66.4%) and recovery (61.5%, 95% CI 59.4%-63.5%) on the GAD-7 measures at the clinical check-in session; both of these measures improved to over 70% at the termination session. Across all measures, outcomes at termination in the secondary cohort were higher than outcomes at clinical check-in for the primary cohort. Outcomes for patients with fewer than four sessions are also included in Table 3 (n=213). Although lower than the outcomes for patients who received an adequate dose of care, 55.4% (95% CI 48.7%-62.1%) of patients who dropped out early also experienced clinically meaningful improvement.

XSL•FO
RenderX

Table 3. Patient response, recovery, and remission rates at check-in (primary cohort) and termination (secondary cohort).

Measures	Clinical check-in (primary cohort) Outcome (95% CI), %	Termination (secondary cohort) Outcome (95% CI), %	Early terminators (<4 sessions) Outcome (95% CI), %
Joint measures			
Clinically meaningful improvement[a]	71.3 (69.7-72.9)	79.2 (77.5-80.9)	55.4 (48.7-62.1)
Recovery[b]	53.2 (51.4-55.0)	65.2 (63.2-67.2)	35.7 (29.2-42.2)
Reliable improvement[c]	65.8 (64.0-67.5)	74.8 (73.0-76.6)	47.4 (40.7-54.2)
Remission[d]	16.4 (15.0-17.7)	27.7 (25.9-29.6)	5.6 (2.5-8.8)
Reliable deterioration[e]	7.9 (7.0-8.9)	6.0 (5.0-7.0)	10.3 (6.2-14.4)
PHQ-9[f] measures			
Reliable improvement	53.0 (51.1-55.0)	65.3 (63.1-67.5)	36.6 (29.4-43.8)
Response[g]	36.4 (34.5-38.3)	51.1 (48.8-53.4)	13.7 (8.6-18.9)
Recovery	55.1 (53.2-57.1)	66.6 (64.4-68.7)	35.4 (28.3-42.6)
Remission	19.5 (17.9-21.1)	32.3 (30.2-34.5)	6.9 (3.1-10.6)
Reliable deterioration	3.5 (2.8-4.2)	3.4 (2.6-4.3)	5.1 (1.8-8.4)
GAD-7[h] measures			
Reliable improvement	64.4 (62.4-66.4)	73.6 (71.5-75.7)	38.7 (31.1-46.2)
Response	41.1 (39.0-43.1)	55.8 (53.4-58.1)	14.7 (9.2-20.2)
Recovery	61.5 (59.4-63.5)	71.9 (69.7-74.0)	37.4 (29.9-44.9)
Remission	20.8 (19.1-22.5)	34.0 (31.8-36.3)	4.9 (1.6-8.3)
Reliable deterioration	4.7 (3.8-5.6)	3.9 (3.0-4.8)	3.7 (0.8-6.6)

[a]The patient met the criteria for joint measures for reliable improvement or recovery.

[b]At the clinical end point, both General Anxiety Disorder-7 (GAD-7) and Patient Health Questionnaire-9 (PHQ-9) scores were <10. For individual measures, recovery requires either the PHQ-9 or the GAD-7 score to be <10.

[c]At the clinical end point, the patient showed improvement of at least 5 points on the PHQ-9 or 4 points on GAD-7, without a decline of 5 or 4 points on the other measure, respectively. Reliable improvement requires an improvement of at least 5 points for PHQ-9 measures and at least 4 points for GAD-7 measures.

[d]At the clinical end point, both GAD-7 and PHQ-9 scores were <5. For individual measures, remission requires either the PHQ-9 or the GAD-7 score to be <5.

[e]At the clinical end point, the patient showed an increase of at least 5 points on the PHQ-9 or 4 points on GAD-7. Reliable deterioration requires an increase of at least 5 points for PHQ-9 measures and at least 4 points for GAD-7 measures.

[f]PHQ-9: Patient Health Questionnaire-9. Patients with PHQ-9 scores≥10 at baseline were included.

[g]At the clinical end point, the patient achieved 50% or greater improvement from the baseline score.

[h]GAD-7: General Anxiety Disorder-7. Patients with GAD-7 scores≥10 at baseline were included.

Longitudinal analysis of PHQ-9 and GAD-7 assessments at the session level estimated an approximately 1-point reduction in the assessment score per attended session, which was attenuated over time by approximately 0.05 points per session, accounting for the nonlinear trajectory of change (see Multimedia Appendix 1). Figure 7 shows model-derived average estimates of PHQ-9 and GAD-7 scores across session numbers, along with model parameter-based 95% CIs for each estimated point. The figure demonstrates the swift, early change in symptoms, followed by an attenuated rate of change, with both curves dipping below the clinical threshold (score of 10) near 8 (PHQ-9) and 6 (GAD-7) sessions, respectively.

XSL•FO
RenderX

Figure 7. Model-derived estimates for average PHQ-9 and GAD-7 scores over the treatment duration. GAD-7: General Anxiety Disorder-7; PHQ-9: Patient Health Questionnaire-9.

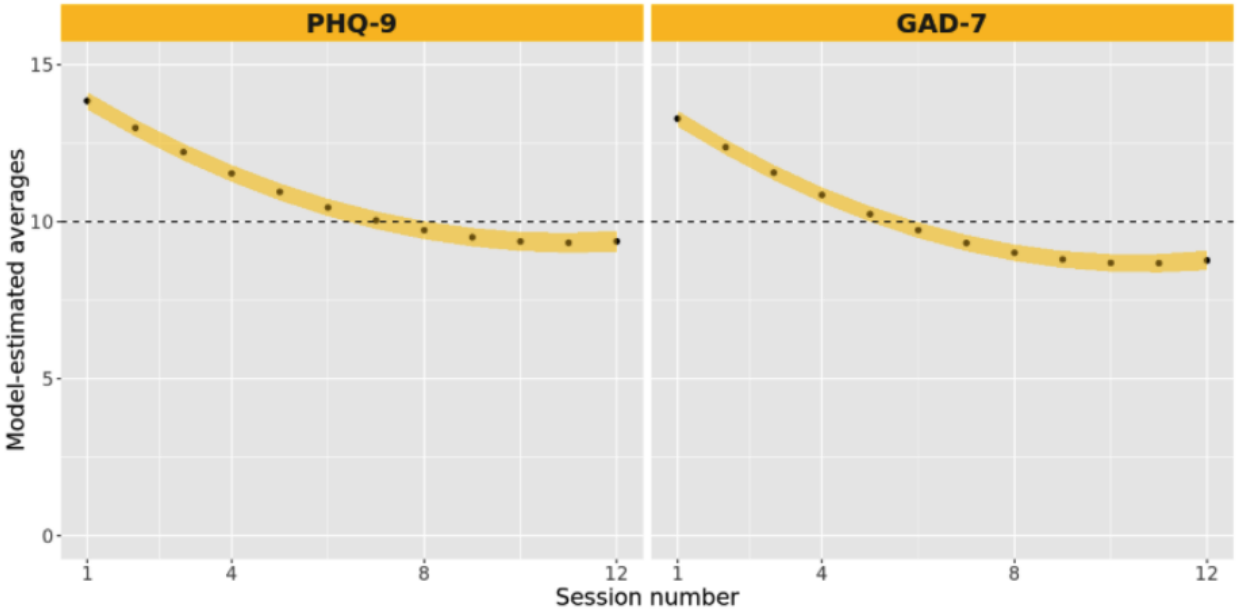

Discussion

Principal Findings

In this study, we examined clinical outcomes among a cohort of 2984 patients who received psychotherapy supported by MBC and a suite of CDSS tools. The study was intended to characterize patient retention and clinical outcomes associated with a large-scale implementation of MBC in a real-world, technology-enabled psychotherapy practice.

The findings indicated that there were clinically meaningful reductions in depression and anxiety symptoms over the course of care among the patients included in the study sample. By the 12th session, 71.3% of patients achieved clinically meaningful improvement in their symptoms, and outcomes improved further when observed at the completion of care (79.2%). On average, improvement in PHQ-9 and GAD-7 scores exceeded the threshold for reliable and clinically significant symptom improvement on these measures [33,34]. We also observed high rates of patient retention in care and over a 95% completion rate for MBC surveys (eg, PHQ-9, GAD-7), indicating high levels of engagement with the care model by clinicians.

Although the study design does not allow us to attribute these outcomes to the suite of CDSS tools or the use of MBC, benchmarks in the literature suggest these outcomes compare favorably to depression and anxiety outcomes achieved in other settings. Meta-analyses of outcomes from RCTs have found depression response rates of 25%-41%, which are lower than the depression response rate of 48% in this study [8,10]. (All Two Chairs outcomes in these comparisons were adjusted to match the inclusion criteria and outcome definitions from the comparator studies. First, we included all individuals who had any follow-up data after baseline in the Two Chairs sample, in contrast to our primary analyses, which included follow-up data only for those who completed at least four sessions. Second, we used a reliable change value of 5 on the GAD-7 for the Blueprint Quality Index comparison to be consistent with its reliable change criteria. Third, we used 12-session outcomes for comparisons to the Blueprint Quality Index comparators, who specified their end point at 12 sessions [8], and outcomes at graduation for all other comparators, who used multiple end-point definitions [8,10,34,37].) When compared to naturalistic studies examining clinical outcomes of clinicians using a range of therapeutic modalities and approaches, response rates (46% in the literature vs 48% in this study) and recovery rates (21% in the literature vs 54% in this study) also compare favorably [11,37]. In anxiety disorders, a systematic review of outcomes for cognitive behavioral therapy showed reliable improvement rates of 45% (vs 62% in this study), and finally, a large naturalistic study showed recovery rates of 31% (vs 60% in this study) [10,38]. These results suggest that the outcomes observed in this study may be superior to outcomes observed in similar large-scale, real-world care settings and comparable to observed outcomes from RCTs of psychotherapy for depression and anxiety.

Rates of dropout and adherence to MBC in this sample also compare favorably to rates observed in the literature. The dropout rates in a standard health care setting were six times higher than observed in our sample (73.5% vs 10.1%) [6]. Completion rates of MBC surveys (PHQ-9 and GAD-7) were higher than 95% across the entire population, with individual clinicians having a median survey adherence rate of 90% within their patient panels. This is in stark contrast to general practice, where only 20% of providers use MBC at all, and 5% of practitioners use it in accordance with an empirically informed schedule (eg, every session) [17].

These findings suggest that in comparison to similar practice settings, patients in this study are more likely to receive an effective dose of care and achieve positive clinical outcomes.

XSL•FO
RenderX

It is possible that the comprehensive use of MBC, supported by the suite of CDSS tools, enables clinicians to identify patients at risk of poor outcomes and adjust treatment in a timely manner, consistent with MBC best practices [17,18].

EBPs are known to be effective for improving overall outcomes in care; however, the financial, organizational, and clinician burden associated with the implementation and sustainment of EBPs likely contribute to their low uptake in clinical practice and limit their potential to improve outcomes [39]. The provision of clinically useful software that facilitates the practice of MBC may reduce some of the barriers to effective adoption and implementation. In this study, clinicians had access to consistent and automated measurement, visual interpretation guidance of results, and actionable insights, in addition to a set of other tools that improve accountability and may increase clinical efficiency and effectiveness. Although these tools cannot change clinical behavior on their own, they can support the practice of MBC and enable clinicians to provide more focused, evidence-based interventions. It is possible that these tools contributed to the high levels of patient retention and clinical outcomes observed in this study.

This study's findings have implications for both value-based payment models as well as standard reimbursement models of care. Psychotherapy is an effective treatment for a broad range of mental health conditions, and health care claim–based analyses suggest that outpatient mental health treatment can save overall health care costs compared to those who do not receive treatment at all [40]. Importantly, evidence suggests that cost savings improve further if individuals are retained in care past one or two sessions, indicating that the dose of treatment must be sufficient to produce clinically meaningful change [40]. Moreover, emerging evidence suggests that decreases in symptoms drive reductions in costs associated with mental health conditions, particularly when compared to those whose symptoms stagnate or worsen over time [41]. Providing patients with the appropriate care, ensuring they remain engaged in care after initiation, and achieving meaningful outcomes can help reduce the overall cost and burden associated with mental health conditions.

Limitations

Although this work offers some suggestive evidence that CDSSs may be associated with higher-quality practice, it is preliminary. The primary limitation of this study is the retrospective design and lack of a comparison group, which limits our ability to examine causal pathways. The study provides benchmarks for comparison to other general practice and structured psychotherapy outcomes; however, these are nonrandomized and unmatched comparators and thus are only illustrative of potential differences and should not be considered rigorous comparisons. Additional research using experimental designs and high-quality comparison samples is needed to isolate the specific impacts of the MBC and CDSS supports offered.

Second, there are numerous potential threats to generalizability outside of the study sample. The CDSSs were examined in one organization, and the study setting incentivizes clinician use of MBC, limiting generalizability to other organizations that may have differences in staffing, policy, culture, patient population, or other factors. The data may also only be generalizable to individuals with clinically elevated levels of depression and anxiety. In addition, to describe the outcomes of patients who received an adequate dose of treatment, the study excluded patients who dropped out prior to the fourth session, which may have inflated the results. Although lower than the results for treatment completers, most of these patients also experienced significant symptom reductions (55%), suggesting this limitation may be minor.

Third, the study design did not allow an examination of other factors that may have influenced the results, such as the clinician skill level, training effectiveness, or patient-level covariates (eg, baseline severity or concurrent treatment during the study period). Future studies should attempt to understand the contributions of these variables on clinical outcomes.

Finally, several authors are employed by the company where the data were generated, and although efforts were made to reduce potential bias by collaborating with external consultants, it is possible conflicts of interest may have influenced the results. A strength of the study is that the extremely high rate of measurement completion ensured that the outcomes of nearly all patients were observable, in contrast to other settings in which many patients may be lost to follow-up, providing a comprehensive view into outcomes in a naturalistic setting.

Conclusion

The results of this retrospective cohort study suggest that an integrated software platform supporting MBC, clinical decision support, and automation of clinically burdensome tasks is associated with high levels of care engagement and clinical outcomes. Although additional research is needed to isolate and clarify the effects of these tools, these findings are suggestive that technology-enabled innovations may help improve the quality of care delivery at scale.

Acknowledgments

We appreciated the review and feedback on this paper from our colleagues at Two Chairs (Maura Church and Marisa Perera) and JG Research & Evaluation (Kristal Jones and Carol Hardy).

Data Availability

The data presented are an accurate and honest account of the research performed and no data were fabricated or manipulated. Given the sensitive nature of the health care data used in the study, requests for raw data will be handled by the contacting author NRF and will be considered on a case-by-case basis.

XSL•FO
RenderX

Authors' Contributions

All authors have made significant contributions to the conception, design, execution, data analysis, or interpretation of the study. All authors have reviewed and approved the final manuscript and agree to be accountable for all aspects of the work. The authors confirm that this manuscript is an original work that has not been previously published and is not under consideration elsewhere. All works and ideas of others are appropriately cited and acknowledged.

Conflicts of Interest

NRF, JN, and MTA are employees of and have equity interest in Two Chairs. RT, KV, and BG are employed by the independent research consultation firm JG Research & Evaluation, contracted by Two Chairs to provide independent consultation and support on the research design, methods, data analyses, and conclusions. Funding for this study was provided by Two Chairs.

Multimedia Appendix 1
Fixed effects output from longitudinal models of change in sessions 1-12 for the PHQ-9 (Patient Health Questionnaire-9) and GAD-7 (General Anxiety Disorder-7).
[DOCX File , 16 KB - formative_v9i1e76264_app1.docx]

Multimedia Appendix 2
STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) checklist.
[PDF File (Adobe PDF File), 157 KB - formative_v9i1e76264_app2.pdf]

References

1. Arias D, Saxena S, Verguet S. Quantifying the global burden of mental disorders and their economic value. EClinicalMedicine 2022 Dec;54:101675-101685 [FREE Full text] [doi: 10.1016/j.eclinm.2022.101675] [Medline: 36193171]
2. Greenberg PE, Fournier A, Sisitsky T, Simes M, Berman R, Koenigsberg SH, et al. The economic burden of adults with major depressive disorder in the United States (2010 and 2018). Pharmacoeconomics 2021 Jun 05;39(6):653-665 [FREE Full text] [doi: 10.1007/s40273-021-01019-4] [Medline: 33950419]
3. Cutler AJ, Keyloun KR, Higa S, Park J, Bonafede M, Gillard P, et al. Annual costs among patients with major depressive disorder and the impact of key clinical events. J Manag Care Spec Pharm 2022 Dec;28(12):1335-1343 [FREE Full text] [doi: 10.18553/jmcp.2022.28.12.1335] [Medline: 36427344]
4. Wang J, Qiu Y, Zhu X. Trends of mental health care utilization among US adults from 1999 to 2018. BMC Psychiatry 2023 Sep 12;23(1):665-676 [FREE Full text] [doi: 10.1186/s12888-023-05156-2] [Medline: 37700243]
5. Swift JK, Callahan JL, Cooper M, Parkin SR. The impact of accommodating client preference in psychotherapy: a meta-analysis. J Clin Psychol 2018 Nov 08;74(11):1924-1937 [FREE Full text] [doi: 10.1002/jclp.22680] [Medline: 30091140]
6. Simon GE, Imel ZE, Ludman EJ, Steinfeld BJ. Is dropout after a first psychotherapy visit always a bad outcome? Psychiatr Serv 2012 Jul;63(7):705-707 [FREE Full text] [doi: 10.1176/appi.ps.201100309] [Medline: 22752034]
7. Barrett MS, Chua W, Crits-Christoph P, Gibbons MB, Casiano D, Thompson D. Early withdrawal from mental health treatment: implications for psychotherapy practice. Psychotherapy (Chic) 2008 Jun 01;45(2):247-267 [FREE Full text] [doi: 10.1037/0033-3204.45.2.247] [Medline: 19838318]
8. Kolovos S, van Tulder MW, Cuijpers P, Prigent A, Chevreul K, Riper H, et al. The effect of treatment as usual on major depressive disorder: a meta-analysis. J Affect Disord 2017 Mar 01;210:72-81 [FREE Full text] [doi: 10.1016/j.jad.2016.12.013] [Medline: 28013125]
9. Sawchuk CN, Craner JR, Berg SL, Smyth K, Mack J, Glader M, et al. Initial outcomes of a real-world multi-site primary care psychotherapy program. Gen Hosp Psychiatry 2018 Sep;54:5-11 [FREE Full text] [doi: 10.1016/j.genhosppsych.2018.06.005] [Medline: 30029160]
10. Cuijpers P, Karyotaki E, Ciharova M, Miguel C, Noma H, Furukawa TA. The effects of psychotherapies for depression on response, remission, reliable change, and deterioration: a meta-analysis. Acta Psychiatr Scand 2021 Sep 04;144(3):288-299 [FREE Full text] [doi: 10.1111/acps.13335] [Medline: 34107050]
11. Freed D. Introducing the Blueprint Quality Index – a new way to benchmark behavioral healthcare quality. Blueprint. 2023 Oct 11. URL: https://www.blueprint.ai/blog/introducing-the-blueprint-quality-index-a-new-way-to-benchmark-behavioral-healthcare-quality [accessed 2025-02-24]
12. McKay CE. Evidence based practices in mental health: advantages, disadvantages, and research considerations. Center for Mental Health Services Research. 2007 Aug. URL: https://www.umassmed.edu/globalassets/center-for-mental-health-services-research/documents/products-publications/issue-briefs/rehab/evidence-based-practices-in-mental-health.pdf [accessed 2025-10-16]
13. Bowser DM, Henry BF, McCollister KE. Cost analysis in implementation studies of evidence-based practices for mental health and substance use disorders: a systematic review. Implement Sci 2021 Mar 12;16(1):26-41 [FREE Full text] [doi: 10.1186/s13012-021-01094-3] [Medline: 33706780]

XSL•FO
RenderX

14. McAleavey AA, de Jong K, Nissen-Lie HA, Boswell JF, Moltu C, Lutz W. Routine outcome monitoring and clinical feedback in psychotherapy: recent advances and future directions. Adm Policy Ment Health 2024 May;51(3):291-305 [FREE Full text] [doi: 10.1007/s10488-024-01351-9] [Medline: 38329643]
15. Edmunds JM, Beidas RS, Kendall PC. Dissemination and implementation of evidence-based practices: training and consultation as implementation strategies. Clin Psychol (New York) 2013 Jun 01;20(2):152-165 [FREE Full text] [doi: 10.1111/cpsp.12031] [Medline: 24072959]
16. Barkham M, De Jong K, Delgadillo J, Lutz W. Routine outcome monitoring (ROM) and feedback: research review and recommendations. Psychother Res 2023 Sep 17;33(7):841-855 [FREE Full text] [doi: 10.1080/10503307.2023.2181114] [Medline: 36931228]
17. Lewis CC, Boyd M, Puspitasari A, Navarro E, Howard J, Kassab H, et al. Implementing measurement-based care in behavioral health: a review. JAMA Psychiatry 2019 Mar 01;76(3):324-335 [FREE Full text] [doi: 10.1001/jamapsychiatry.2018.3329] [Medline: 30566197]
18. Scott K, Lewis CC. Using measurement-based care to enhance any treatment. Cogn Behav Pract 2015 Feb;22(1):49-59 [FREE Full text] [doi: 10.1016/j.cbpra.2014.01.010] [Medline: 27330267]
19. Price MS, Christensen H, Urquhart J, Butte S, Acierno R, Little DM. Patient-facing, semi-automated, solutions to enhance patient participation in measurement-based care practice standards in a tele-mental health specialty clinic for populations at higher risk for self-harm. Int J Med Inform 2023 Sep;177:105155-105159 [FREE Full text] [doi: 10.1016/j.ijmedinf.2023.105155] [Medline: 37467589]
20. Schuckard E, Miller S, Hubble M. Feeback-informed treatment: historical and empirical foundations. In: Prescott D, Maeschalk C, Miller S, editors. Feedback-Informed Treatment in Clinical Practice: Reaching for Excellence. Washington, DC: American Psychological Association; 2017:13-36.
21. Tsai KH, Moskowitz AL, Brown TE, Park AL, Chorpita BF. Interpreting progress feedback to guide clinical decision-making in children's mental health services. Adm Policy Ment Health 2016 Mar 28;43(2):199-206 [FREE Full text] [doi: 10.1007/s10488-015-0630-4] [Medline: 25627140]
22. Dodson P, Haase AM, Jeffreys M, Hales C. Capturing patient experiences of care with digital technology to improve service delivery and quality of care: a scoping review. Digit Health 2024 Oct 22;10:20552076241282900-20552076241282917 [FREE Full text] [doi: 10.1177/20552076241282900] [Medline: 39449855]
23. Sawyer T, Gray M, Umoren R. The global healthcare simulation economy: a scoping review. Cureus 2022 Feb;14(2):e22629-E22637 [FREE Full text] [doi: 10.7759/cureus.22629] [Medline: 35371834]
24. Bayor AA, Li J, Yang IA, Varnfield M. Designing clinical decision support systems (CDSS) – a user centered lens of the design characteristics, challenges, and implications: systematic review. J Med Internet Res 2025 Jun 20;27:e63733-E63752 [FREE Full text] [doi: 10.2196/63733] [Medline: 40540451]
25. Zharima C, Mhlanga S, Abdulla S, Goudge J, Griffiths F. What engagement strategies are useful in facilitating the implementation of electronic health records in health care settings? A rapid review of qualitative evidence synthesis using the normalization process theory. Digit Health 2024 Nov 03;10:20552076241291286-20552076241291304 [FREE Full text] [doi: 10.1177/20552076241291286] [Medline: 39497787]
26. Williams N, Marcus S, Ehrhart M, Sklar M, Esp S, Carandang K, et al. Randomized trial of an organizational implementation strategy to improve measurement-based care fidelity and youth outcomes in community mental health. J Am Acad Child Adolesc Psychiatry 2024 Oct;63(10):991-1004 [FREE Full text] [doi: 10.1016/j.jaac.2023.11.010] [Medline: 38070868]
27. Kroenke K, Spitzer RL, Williams JBW. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
28. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097 [FREE Full text] [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
29. Lungu A, Jun JJ, Azarmanesh O, Leykin Y, Chen CE. Blended care-cognitive behavioral therapy for depression and anxiety in real-world settings: pragmatic retrospective study. J Med Internet Res 2020 Jul 06;22(7):e18723-E18735 [FREE Full text] [doi: 10.2196/18723] [Medline: 32628120]
30. van Dijk DA, Deen ML, van den Boogaard TM, Ruhé HG, Spijker J, Peeters FPML. Prevalence and prediction of dropout during depression treatment in routine outpatient care: an observational study. Eur Arch Psychiatry Clin Neurosci 2023 Aug 17;273(5):1151-1161 [FREE Full text] [doi: 10.1007/s00406-022-01499-1] [Medline: 36253582]
31. Kuyken W, Hayes R, Barrett B, Byng R, Dalgleish T, Kessler D, et al. Effectiveness and cost-effectiveness of mindfulness-based cognitive therapy compared with maintenance antidepressant treatment in the prevention of depressive relapse or recurrence (PREVENT): a randomised controlled trial. Lancet 2015 Jul 04;386(9988):63-73 [FREE Full text] [doi: 10.1016/S0140-6736(14)62222-4] [Medline: 25907157]
32. Forand N, Nettiksimmons J, Brownell A, Anton M, Truxson R, Green B, et al. The impact of measurement based care at scale: examining the effects of implementation on patient outcomes and provider behaviors. Frontiers Health Services 2025:1-16. [doi: 10.2196/preprints.79596]
33. Kroenke K. Enhancing the clinical utility of depression screening. CMAJ 2012 Feb 21;184(3):281-282 [FREE Full text] [doi: 10.1503/cmaj.112004] [Medline: 22231681]

XSL•FO
RenderX

34. Bischoff T, Anderson SR, Heafner J, Tambling R. Establishment of a reliable change index for the GAD-7. Psychol Community Health 2020 Apr 08;8(1):176-187 [FREE Full text] [doi: 10.5964/pch.v8i1.309]
35. Da Costa CP, Padoan CS, Hauck S, Teche SP, Eizirik CL. What is the appropriate time to measure outcome and process factors in psychodynamic psychotherapy? Trends Psychiatry Psychother 2019 May 30;41(2):112-120 [FREE Full text] [doi: 10.1590/2237-6089-2017-0151] [Medline: 31166563]
36. Kalin NH. The critical relationship between anxiety and depression. Am J Psychiatry 2020 May 01;177(5):365-367 [FREE Full text] [doi: 10.1176/appi.ajp.2020.20030305] [Medline: 32354270]
37. Coley R, Boggs J, Beck A, Hartzler AL, Simon G. Defining success in measurement-based care for depression: a comparison of common metrics. Psychiatr Serv 2020 Apr 01;71(4):312-318 [FREE Full text] [doi: 10.1176/appi.ps.201900295] [Medline: 31847739]
38. Loerinc AG, Meuret AE, Twohig MP, Rosenfield D, Bluett EJ, Craske MG. Response rates for CBT for anxiety disorders: need for standardized criteria. Clin Psychol Rev 2015 Dec;42:72-82 [FREE Full text] [doi: 10.1016/j.cpr.2015.08.004] [Medline: 26319194]
39. Boswell J, Schartzman C. What are the pathways and barriers to implementing evidence-based practice and practice-based evidence in psychotherapy? In: Leong FTL, Callahan JL, Zimmerman J, Constantion MJ, Eubanks CF, editors. APA Handbook of Psychotherapy: Evidence-Based Practice, Practice-Based Evidence, and Contextual Participant-Driven Practice. Washington, DC: American Psychological Association; 2024:287-303.
40. Bellon J, Quinlan C, Taylor B, Nemecek D, Borden E, Needs P. Association of outpatient behavioral health treatment with medical and pharmacy costs in the first 27 months following a new behavioral health diagnosis in the US. JAMA Netw Open 2022 Dec 01;5(12):e2244644-e2244654 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.44644] [Medline: 36472875]
41. Popkov AA, Barrett TS, Shergill A, Donohue M, Anderson RJ, Karlin BE. Association between depression symptom severity and total cost of care: findings from a large, 2-year, claims-based, retrospective population health study. J Affect Disord 2025 Jan 01;368:41-47 [FREE Full text] [doi: 10.1016/j.jad.2024.09.056] [Medline: 39271070]

Abbreviations

CDSS: clinical decision support system
EBP: evidence based practice
HMO: health maintenance organization
GAD-7: General Anxiety Disorder-7
MBC: measurement-based care
MHQoL: mental health quality of life
PHQ-9: Patient Health Questionnaire-9
PPO: preferred provider organization
RCT: randomized controlled trial

Edited by J Sarvestan; submitted 19.04.25; peer-reviewed by A Popkov; comments to author 06.08.25; revised version received 04.09.25; accepted 13.10.25; published 02.12.25.

Please cite as:
Forand NR, Nettiksimmons J, Anton M, Truxson R, Vanderwood K, Green B
Depression and Anxiety Outcomes in a Technology-Enabled Psychotherapy Practice: Retrospective Cohort Study
JMIR Form Res 2025;9:e76264
URL: https://formative.jmir.org/2025/1/e76264
doi:10.2196/76264
PMID:

©Nicholas R Forand, Jasmine Nettiksimmons, Margaret Anton, Raven Truxson, Karl Vanderwood, Brandn Green. Originally published in JMIR Formative Research (https://formative.jmir.org), 02.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Feasibility of a Specialized Large Language Model for Postgraduate Medical Examination Preparation: Single-Center Proof-Of-Concept Study

Yun Hao Leong[1,2*], MRCS, MMed; Lathiga Nambiar[1*], MBBS; Victoria Y J Tay[1,2], MMed; Sui An Lie[1,2], MRCP, MMed; Ke Yuhe[1,2], MMed

[1]Division of Anesthesiology and Perioperative Medicine, Singapore General Hospital, Singapore, Singapore

[2]Department of Anesthesiology, Sengkang General Hospital, Singapore, Singapore

[*]these authors contributed equally

Corresponding Author:
Yun Hao Leong, MRCS, MMed
Division of Anesthesiology and Perioperative Medicine
Singapore General Hospital
31 Third Hospital Ave
Singapore, 168753
Singapore
Phone: 65 63213411
Fax: 65 63213411
Email: leong.yun.hao@singhealth.com.sg

Abstract

Background: Large language models (LLMs) are increasingly used in medical education for feedback and grading; yet their role in postgraduate examination preparation remains uncertain due to inconsistent grading, hallucinations, and user acceptance.

Objective: This study evaluates the Personalized Anesthesia Study Support (PASS), a specialized GPT-4 model developed to assist candidates preparing for Singapore's postgraduate specialist anesthesiology examination. We assessed user acceptance, grading interrater reliability, and hallucination detection rates to determine the feasibility of integrating specialized LLMs into high-stakes examination preparation.

Methods: PASS was built on OpenAI's GPT-4 and adapted with domain-specific prompts and references. Twenty-one senior anesthesiology residents completed a mock short answer question examination, which was independently graded by 3 human examiners and 3 PASS iterations. Participants reviewed feedback from both PASS and standard GPT-4 and completed a technology acceptance model (TAM) survey. Grading reliability was evaluated using Cohen and Fleiss κ. Hallucination rates were assessed by participants and examiners.

Results: Of the 21 participants, 17 (81%) completed the TAM survey, generating 136 responses. PASS scored significantly higher than standard GPT-4 in usefulness (mean 4.25, SD 0.50 vs mean 3.44, SD 0.82; $P<.001$), efficiency (mean 4.12, SD 0.61 vs mean 3.41, SD 0.74; $P<.001$), and likelihood of future use (mean 4.13, SD 0.75 vs mean 3.59, SD 0.90; $P<.001$), with no significant difference in ease of use (mean 4.56, SD 0.63 vs mean 4.50, SD 0.61; $P=.35$). Internal grading reliability was moderate for PASS ($\kappa=0.522$) and fair for human examiners ($\kappa=0.275$). Across 316 PASS-generated responses, 67 hallucinations and 189 deviations were labeled. Hallucination labeling rates were comparable between candidates (10/67, 15%) and examiners (57/249, 22.9%; $P=.21$), while examiners labeled significantly more deviations (168/249, 67.5% vs 21/67, 31%; $P<.001$).

Conclusions: PASS demonstrated strong user acceptance and grading reliability, suggesting feasibility in high-stakes examination preparation. Experienced learners could identify major hallucinations at comparable rates to examiners, suggesting potential in self-directed learning but with continued need for caution. Further research should refine grading accuracy and explore multicenter evaluation of specialized LLMs for postgraduate medical education.

(JMIR Form Res 2025;9:e77580) doi:10.2196/77580

KEYWORDS

large language model; artificial intelligence; technology acceptance model; medical education; postgraduate examination

XSL•FO
RenderX

Introduction

The integration of artificial intelligence (AI) into education has accelerated in recent years, particularly with the rapid evolution of large language models (LLMs) such as OpenAI's GPT-4 and Anthropic's Claude. These models employ transformer-based deep learning architectures to process and generate human-like text, enabling dynamic adaptation across educational domains. In medical education, LLMs have been applied for automated question generation, simulated case discussions, feedback provision, and adaptive tutoring [1]. Recent reviews highlight that LLMs can support formative learning through contextual explanations and feedback personalization while also noting persistent issues in transparency, bias, and factual accuracy [2-4].

Despite these promising uses, major challenges remain in applying LLMs to high-stakes assessment. Studies on LLM-based grading report inconsistent results, including conservative grading patterns, variable interrater reliability (IRR) with human assessors [5], and context-dependent scoring variability [6]. Even advanced LLMs such as GPT-4 have shown inconsistent scoring in complex clinical assessments, underscoring the importance of standardized rubrics, transparent prompt reporting, and model calibration [7,8]. Most prior work has evaluated general-purpose LLMs with minimal adaptation, which limits their accuracy in assessing domain-specific medical content. While prompt engineering has been explored to optimize response consistency [9], few studies have explored the feasibility of specialized LLMs tailored to specific educational needs.

Another major concern surrounding LLMs in education is the phenomenon of hallucinations, where models generate responses that appear plausible but are factually incorrect or misleading [10]. Previous studies have reported hallucination rates as high as 27.1% in LLM-generated feedback [11], raising concerns about the potential for misinformation. Recent work in medical education has categorized hallucinations by cognitive level and clinical severity, noting that even subtle deviations can distort learners' conceptual understanding [12]. Given the high-stakes nature of medical training, it is essential to understand not only how frequently these hallucinations occur but also the extent to which educators and students can detect and mitigate their impact.

In parallel, user acceptance strongly influences whether LLM-based systems are adopted into medical curricula. Although several studies report favorable perceptions of LLM-assisted educational tools among undergraduate students [13], there remains a gap in understanding postgraduate learners' perceptions, particularly in specialist examination preparation. Regardless of technical performance, effective integration of LLMs into educational practice depends on user acceptance, which is shaped by perceived usefulness, ease of use, and trustworthiness [14].

To address these issues, our study aims to develop and pilot the Personalized Anesthesia Study Support (PASS), a specialized GPT designed to provide grading and feedback for candidates preparing for Singapore's postgraduate specialist examination in anesthesiology. We present this as a proof-of-concept study to explore the feasibility and potential role of specialized LLMs as supplemental tools in postgraduate examination preparation. Our primary objective was to assess end user acceptance of PASS by using a technology acceptance model (TAM) survey. Secondary objectives included evaluating the grading reliability compared to human examiners and analyzing the hallucination detection trends.

Methods

Ethical Considerations

This study was conducted at Singapore General Hospital under the SingHealth Anesthesiology Residency Program. It was reviewed by the Singapore General Hospital Research Office and was granted exemption from formal institutional review board review, as it involved routine educational activities and the use of anonymized resident data. Participation was voluntary, and no identifiable personal data were collected. Completion of the TAM survey was considered implied consent. All data were deidentified prior to analysis, and no compensation was provided to the participants.

Recruitment and Study Participants

Participants were recruited in a systematic cohort-based approach within the SingHealth Anesthesiology Residency Program at Singapore General Hospital. All anesthesiology residents actively preparing for an upcoming Master of Medicine (MMed) anesthesiology examination were scheduled to complete a routine, in-person mock short answer question (SAQ) assessment and were invited to participate in the study. Eligibility required active enrollment in the residency program with at least 32 months of anesthesia-related training experience, consistent with the criteria for sitting in the MMed examination. No additional exclusion criteria were applied, as the cohort was already homogeneous in the training level and examination eligibility. At the start of the session, a member of the program faculty provided a standardized study briefing, explaining its objectives, procedures, and voluntary nature of participation. Residents were informed that participation would not influence their training evaluation or examination outcomes. Implied consent was obtained following this briefing. A total of 21 residents, representing the full cohort of candidates preparing for that year's MMed examination, consented and participated in the study. The cohort comprised residents in years 2 to 4 of training, with a gender distribution of 14 females and 7 males.

Development Process of PASS

PASS was developed using OpenAI's GPT-4 model and customized via the ChatGPT Custom GPT interface to provide grading and detailed feedback for SAQ-style responses. Two study investigators (LYH and KY) conducted 3 iterative rounds of pilot testing by using simulated essay responses modeled after previous MMed anesthesiology examination questions. Each round involved a systematic review of PASS outputs by both investigators and 2 faculty anesthesiologists to evaluate the clinical accuracy, feedback clarity, and grading consistency. Refinements across iterations focused on clarifying role-based prompts, improving alignment with examiner expectations, and

XSL•FO
RenderX

reducing hallucinated or ambiguous statements in the generated feedback. After each testing round, the prompt was revised to enhance the specificity of the instructions and the standardization of the grading language. The final prompt version, presented in Table S1 in Multimedia Appendix 1, reflected cumulative adjustments derived from these pilot evaluations and represented the configuration used in the study.

The reference materials in PASS included the 2024 MMed anesthesiology syllabus (PDF, 32 pages), examiner feedback reports (n=5, 1-2 pages each), past SAQ papers (2018-2024), and standard textbooks such as Miller's Anesthesia (9th edition) and Yao & Artusio's Anesthesiology: Problem-Oriented Patient Management (9th edition).

Prompt engineering followed the principles outlined by Meskó [15], focusing on specificity and contextualization. Role-playing prompts were used to instruct the LLM to adopt the perspective of an anesthesiologist examiner, ensuring clinically accurate and assessment-calibrated evaluations.

Study Design and Grading Evaluation

Participants completed a closed-book mock examination of 4 SAQs. These were adapted from prior internal mock examinations aligned with recent MMed anesthesiology examination themes and reviewed by another anesthesiology faculty to ensure content validity. Examiner 1 developed a standardized mark scheme and graded all 84 scripts as per standard assessment practices. Each SAQ constituted 1 script, with 21 participants producing 84 total responses.

Additionally, 2 independent human examiners (examiners 2 and 3) also graded all the scripts based on the standardized mark scheme. Examiners 2 and 3 independently evaluated the PASS grading iterations alongside examiner 1. Two study investigators generated 3 separate PASS grading iterations per script in parallel chat windows. PASS was not provided with the official grading rubric and relied solely on its internal knowledge base. SAQ responses were graded on a 0-8 scale and stratified into poor (0-4.5), average (5-5.5), and good (6-8) categories. The stratified gradings were then analyzed for grading IRR, with examiner 1's scores serving as the reference standard when required.

TAM Evaluation

Participants subsequently evaluated both PASS and standard GPT-4.0 feedback by using a customized TAM survey [16]. TAM is grounded in the theory of reasoned action, which posits that an individual's behavioral intention to use a technology is shaped by their attitudes toward the technology and by perceived usefulness. This framework has been widely applied to understand user acceptance of new technologies, including those used in medical education [14,17,18]. Four dimensions were rated on a 5-point Likert scale (1=least agreement, 5=strongest agreement): (1) usefulness for examination preparation, (2) efficiency in study preparation, (3) ease of use, and (4) likelihood of future use. Participation in the survey was voluntary and anonymous. Incomplete surveys were excluded from the analysis.

Hallucination and Deviation Analysis

PASS responses and feedback to participant essays were reviewed independently by study investigators and participants to detect hallucinations and deviations. Hallucinations were defined as "any incorrect information or statements that may cause moderate to major patient harm," while deviations were defined as "variations from usual clinical practice unlikely to result in patient harm." The frequency of hallucinations and deviations detection was compared between groups.

Statistical Analysis

All statistical evaluations were performed in the Excel and Python 3.8 environment. The TAM scores for PASS versus GPT-4 were compared using paired 2-sided *t* tests. Fleiss κ was used to calculate IRR within PASS and human examiner groups, and agreement between individual examiners was evaluated using Cohen κ analysis. Chi-square tests were used to compare hallucination and deviation detection between examiners and candidates. As this was a proof-of-concept feasibility study, no a priori power calculation was performed.

This paper adheres to the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) reporting guideline [19], and a completed STROBE checklist has been included in Multimedia Appendix 2.

Results

TAM Results

Of the 21 participants, 17 (81%) completed the TAM survey, giving a total of 136 survey responses. Four participants did not complete the TAM survey, and no data imputation was performed. User perceptions of PASS were positive. Mean ratings for PASS were as follows: usefulness (4.25, SD 0.50), efficiency (4.12, SD 0.61), ease of use (4.56, SD 0.63), and likelihood of future use (4.13, SD 0.75). Paired 2-sided *t* tests comparing PASS to standard GPT-4 revealed statistically significant differences in 3 of the 4 domains. PASS was rated significantly higher than GPT-4 in usefulness (mean difference=0.81, 95% CI 0.66-0.96; $P<.001$), efficiency (mean difference=0.71, 95% CI 0.53-0.88; $P<.001$), and likelihood of future use (mean difference=0.54, 95% CI 0.39-0.70; $P<.001$). There was no significant difference between PASS and GPT-4 in ease of use (mean difference=0.06, 95% CI –0.07 to 0.18; $P=.35$). A summary of the TAM results can be found in Table 1 and Figure 1.

XSL•FO
RenderX

Table 1. Technology acceptance model survey results comparing PASSa and standard GPT-4 in a single-center observational feasibility study involving postgraduate anesthesiology residents (N=21).

Question	PASS, mean (SD)	GPT, mean (SD)	Mean difference (95% CI)	*P* value
Useful for my examination preparation	4.25 (0.5)	3.441 (0.817)	0.809 (0.66 to 0.96)	<.001
Helps me to study efficiently	4.118 (0.612)	3.412 (0.738)	0.706 (0.53 to 0.88)	<.001
Easy to use	4.559 (0.632)	4.5 (0.611)	0.059 (–0.07 to 0.18)	.35
I will use it in future examinations	4.132 (0.751)	3.588 (0.902)	0.544 (0.39 to 0.70)	<.001

[a]PASS: Personalized Anesthesia Study Support.

Figure 1. Flow diagram summarizing the design and data pathways of this study. A total of 21 MMed anesthesiology residents participated in a mock SAQ examination comprising 4 questions (84 total scripts). Each script was independently graded by 3 human examiners and 3 iterations of the specialized large language model (PASS 1-3). MMed: Master of Medicine; PASS: Personalized Anesthesia Study Support; SAQ: short answer question; TAM: technology acceptance model.

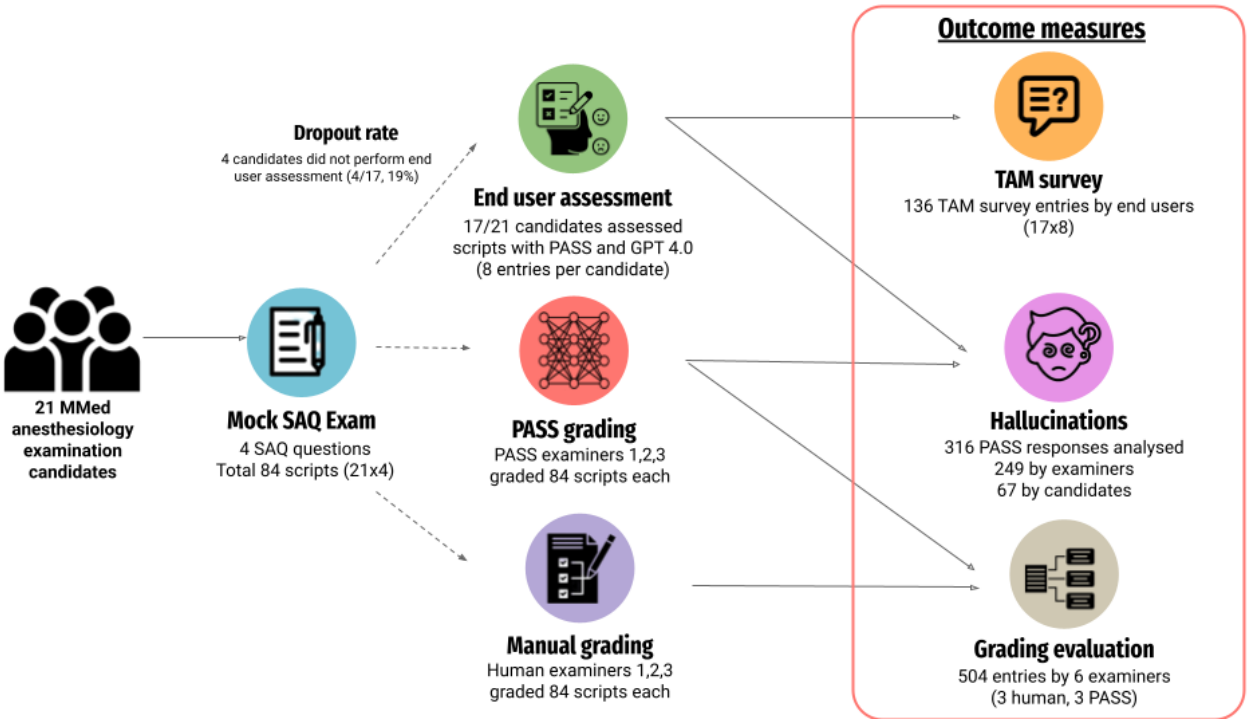

Grading Evaluation

A total of 504 scripts were marked by 6 examiners (3 humans, 3 PASS). PASS examiners demonstrated moderate internal IRR (κ=0.522), whereas human examiners exhibited fair internal IRR (κ=0.275), based on the benchmarks of Landis and Koch [20]. Individually, the agreement of PASS and human examiners with examiner 1 were as follows: PASS 1 (κ=0.237), PASS 2 (κ=0.193), PASS 3 (κ=0.140), examiner 2 (κ=0.149), and examiner 3 (κ=0.470). Among these, examiner 3 demonstrated the highest agreement with examiner 1, while PASS 3 exhibited the lowest. The complete interexaminer agreement matrix can be found in Figure 2. When analyzed as a group (Table 2), PASS examiners approached a moderate level of agreement with the combined group of human examiners (examiners 1, 2, and 3) (κ=0.357, 95% CI 0.156-0.558). Pairwise agreement between PASS and individual human examiners ranged from κ=0.159 (PASS vs examiner 1, slight) to κ=0.429 (PASS vs examiner 3, moderate).

XSL•FO
RenderX

Figure 2. Interexaminer agreement matrix heatmap showing pairwise Cohen κ values differing among human examiners (examiners 1-3) and PASS iterations (PASS 1-3). Each cell represents the level of agreement between 2 graders, with values ranging from 0 (no agreement) to 1 (perfect agreement). Darker red shading indicates higher agreement, lighter blue tones represent moderate agreement, and deep blue tones indicate lower agreement. PASS iterations demonstrated higher internal consistency (κ=0.52-0.57) compared with human examiners (κ=0.15-0.47), suggesting more standardized grading performance across large language model assessments. PASS: Personalized Anesthesia Study Support.

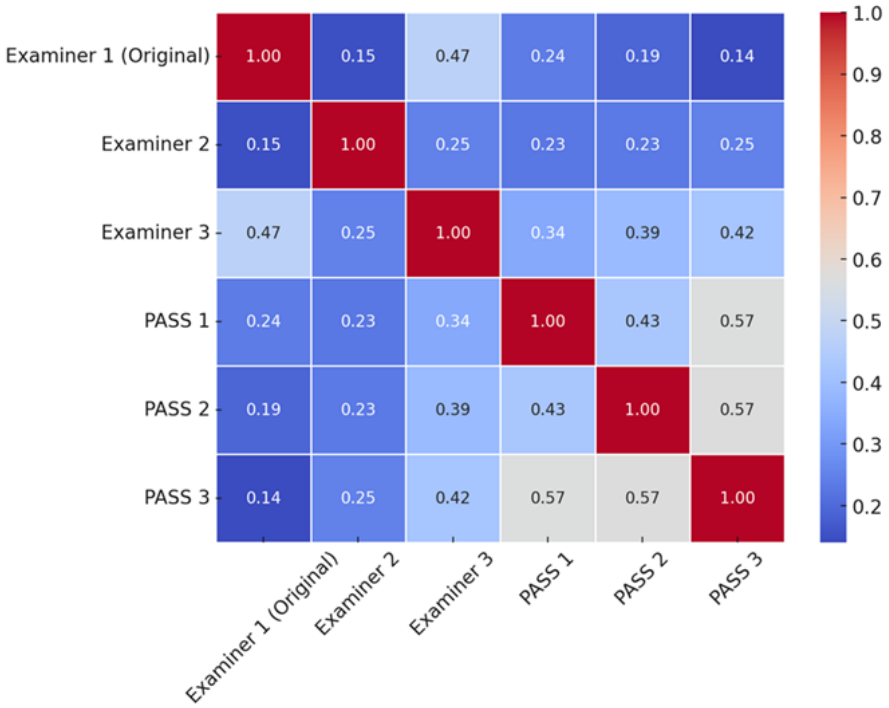

Table 2. Interexaminer agreement between PASSa and human examiners in a single-center observational feasibility study among anesthesiology residents (N=21).

Comparison	κ score (95% CI)
PASS versus combined examiners (1,2,3)	0.357 (0.156 to 0.558)
PASS versus examiner 1	0.159 (–0.053 to 0.372)
PASS versus examiner 2	0.193 (–0.017 to 0.404)
PASS versus examiner 3	0.429 (0.228 to 0.630)

[a]PASS: Personalized Anesthesia Study Support.

Hallucination and Deviation Detection Rates

A total of 316 PASS-generated responses were evaluated for hallucinations and deviations. Examiners conducted 78.8% (249/316) of the evaluations, while candidates accounted for 21.2% (67/316). Across all responses, 67 hallucinations and 189 deviations were labeled. Each hallucination or deviation was labeled by independent reviewers. No statistically significant difference was observed in the hallucination detection rates between candidates and examiners (*P*=.21). Candidates identified 10 hallucinations across 67 responses (15%), while examiners detected 57 hallucinations across 249 responses (22.9%). In contrast, examiners identified significantly more deviations than candidates. Candidates detected 21 deviations (31.3%), whereas examiners identified 168 deviations (67.5%)—a difference that was highly significant (*P*<.001) (Table 3).

Table 3. Comparison of hallucination and deviation labeling by candidates and examiners. Values indicate detection rate (%). Differences were tested using chi-square analysis.

	Candidates (n=67), n (%)	Examiners (n=249), n (%)	*P* value
Hallucinations identified	10 (14.9)	57 (22.9)	.21
Deviations identified	21 (31.3)	168 (67.5)	<.001

XSL•FO
RenderX

Discussion

Primary Outcome: Technology Acceptance and User Perceptions

This proof-of-concept study demonstrates strong end user acceptance of a specialized LLM, supporting the potential role of tailored AI tools in high-stakes medical examination preparation. Younger learners increasingly prefer technology-driven educational tools that provide prompt feedback, personalized learning experiences, and interactive engagement [21,22]. Specialized LLM platforms such as our case example PASS offers an innovative approach to meet these expectations.

Our findings are consistent with prior research showing that LLM-based educational tools are generally well-received across educational contexts [14,23,24]. However, user acceptance specifically for high-stakes examination preparation, where performance directly impacts career progression, remains underexplored [25]. Given that user acceptance is a key factor in the successful adoption of new technologies [26], our results add evidence that specialized LLMs can be successfully incorporated as study aids for postgraduate medical examinations.

Notably, PASS, a specialized version of GPT-4 customized with domain-specific content, received significantly higher ratings in usefulness, efficiency, and likelihood of future use compared to standard GPT-4. This supports the argument that domain-specialized enhancements improve both learner experience and perceived educational value [27].

Grading Reliability

In our study, PASS demonstrated a higher internal grading IRR (κ=0.522) compared to human examiners (κ=0.275), suggesting that LLM-based grading may provide more standardized evaluations. When analyzed as a group, PASS examiners showed statistically significant agreement with human examiners (κ=0.357), indicating that PASS grading trends align with human assessment.

Timely and accurate grading feedback is crucial for effective examination preparation, as it helps candidates refine their responses and align their answering techniques with examiner expectations [28]. However, faculty availability is often limited, restricting or delaying access to grading opportunities. Additionally, human grading is inherently variable, influenced by factors such as individual biases and grading fatigue [29]. Our results suggest that specialized LLMs could offer scalable, consistent supplemental grading and feedback in postgraduate education, potentially alleviating faculty workload and enabling self-directed formative assessment.

However, PASS's grading agreement with humans remained only moderate, indicating that its reliability may not yet be sufficient for formal grading. This finding aligns with previous research highlighting the limitations of LLMs as formal grading tools [30]. Despite these constraints, specialized LLMs remain valuable as self-assessment tools, allowing students to estimate their performance during examination preparation.

PASS was intentionally not provided with the official marking rubric to reflect the realistic conditions of learner-driven use. This design choice likely contributed to moderate agreement levels, as rubric-guided calibration could enhance consistency between AI and human grading. Future studies may investigate whether standardized grading frameworks improve IRR and alignment with human examiners.

Hallucinations and Deviations in AI-Generated Feedback

Hallucinations in LLM generated feedback are a well-documented concern [1]. In medical education, this issue is particularly critical, as even minor misconceptions can lead to significant errors in clinical management, potentially resulting in adverse patient outcomes [31]. While previous studies have discussed potential ways to mitigate the effect of hallucinations [32], there is a notable gap in research comparing the detection rate of hallucinations between tutors and learners.

Encouragingly, our study shows that candidates were able to identify hallucinations at a rate comparable to examiners. The ability of candidates to recognize and filter out major erroneous feedback on their own suggests that specialized LLMs show potential for use under supervision in examination preparation. However, it is important to note that all candidates in our study had at least 32 months of anesthesia experience, which may have contributed to their ability to detect errors effectively. Less experienced learners may not demonstrate the same level of discernment and could require closer supervision when using AI-generated feedback.

Additionally, our results show that candidates were significantly less likely than examiners to detect deviations from standard clinical practice. Failure to recognize clinically relevant, even if minor, deviations from standard practice may inadvertently reinforce incorrect reasoning patterns. If undetected, such deviations could perpetuate unsafe habits or incomplete understanding, which, in a clinical context, might affect patient management. Incorporating structured faculty review or critical appraisal exercises into curricula could help learners identify and mitigate such errors when using AI feedback.

Our findings suggest important considerations for future use of LLMs in education. Educators must assess whether learners possess the critical skills to evaluate AI-generated content before incorporating LLM feedback into curricula. Future research should explore hallucination detection across learner levels and strategies to build critical evaluation skills alongside AI-assisted learning.

Limitations and Future Work

This study has several limitations. First, the small, single-center cohort of 21 participants may limit the generalizability of our findings. All participants were experienced anesthesiology residents from the same training program and were approaching their final examinations. Consequently, the results may not extend to less experienced trainees, learners from other specialties, or different educational settings. Junior trainees may have greater difficulty discerning hallucinations or subtle deviations in AI-generated feedback, underscoring the continued importance of supervised use and faculty guidance. Future

XSL•FO
RenderX

multicenter studies involving a broader range of specialties, experience levels, and training environments will be essential to evaluate the scalability and educational impact of specialized LLMs. Second, no formal power calculation was conducted, given the exploratory, proof-of-concept nature of the study. In addition, intraexaminer reliability was not assessed, which may have influenced the variability in human grading. Finally, this study did not examine the direct effect of using PASS on actual examination performance or learning outcomes. Assessing these educational and behavioral impacts should be a key focus of subsequent research to determine the long-term value and safety of integrating specialized LLMs into postgraduate medical education.

Building on these preliminary findings, future research should focus on multicenter validation of specialized LLMs across different institutions, specialties, and learner experience levels to better define their generalizability. Further work is also needed to evaluate the longitudinal educational impact of LLM-assisted feedback on examination performance, critical thinking, and clinical reasoning. Incorporating standardized grading rubrics or structured prompt frameworks may improve grading reliability and enhance alignment with human examiners. In parallel, research should explore strategies to train learners in AI literacy and critical appraisal, ensuring they can appropriately interpret and verify model outputs. Ultimately, integrating specialized LLMs into postgraduate education should aim to augment, not replace, human judgment, fostering safe, reflective, and supervised use of AI in medical learning environments.

Conclusions

This study provides proof-of-concept evidence supporting the feasibility and user acceptance of specialized LLMs as supplemental study aids for postgraduate medical examination preparation. The specialized LLM (PASS) demonstrated strong user acceptance, outperforming standard GPT-4 in perceived usefulness, study efficiency, and likelihood of future use. PASS also showed moderate grading consistency and agreement with human examiners, indicating potential value in supporting formative, self-directed learning.

While experienced learners were able to recognize major hallucinations at rates comparable to examiners, their lower detection of subtle deviations underscores the continued necessity of human oversight when using AI-generated educational feedback. These findings highlight the potential of specialized LLMs to enhance postgraduate learning under supervision while reinforcing the importance of ongoing faculty involvement, quality control, and iterative model refinement before broader integration into training programs.

Acknowledgments

We would like to extend our gratitude to the SingHealth Anesthesiology Residency Program for their support during the project. During the preparation of this manuscript, the authors employed large language models (ChatGPT) solely to assist with limited aspects of text refinement. These tools were used exclusively for (1) paraphrasing technical methodological descriptions to improve readability while preserving the scientific meaning and (2) optimizing grammatical structure in complex sentences. All large language model–generated content underwent rigorous verification against the original study data and were subject to critical revision by all coauthors to ensure factual accuracy and adherence to the study findings. The final manuscript represents the authors' scholarly work and interpretation.

Data Availability

The datasets generated and analyzed during this study are not publicly available due to its sensitive nature but are available from the corresponding author on reasonable request. Deidentified data supporting the findings of this study may be shared for academic or research purposes upon submission of an appropriate data use agreement.

Funding

The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Authors' Contributions

Conceptualization: YHL, LN, YK
Product development: YHL, YK
Formal analysis: YHL
Investigation: YHL, LN, VYJT, SAL, YK
Methodology: YHL, LN, VYJT, SAL
Supervision: YHL, YK
Writing – original draft: YHL, LN, YK
Writing – review & editing: YHL, LN, VYJT, SAL, YK

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Personalized Anesthesia Study Support prompt.
[DOCX File , 19 KB - formative_v9i1e77580_app1.docx]

Multimedia Appendix 2
STROBE checklist.
[PDF File (Adobe PDF File), 239 KB - formative_v9i1e77580_app2.pdf]

References

1. Abd-Alrazaq A, AlSaad R, Alhuwail D, Ahmed A, Healy PM, Latifi S, et al. Large language models in medical education: opportunities, challenges, and future directions. JMIR Med Educ 2023 Jun 01;9:e48291 [FREE Full text] [doi: 10.2196/48291] [Medline: 37261894]
2. Mohammad B, Supti T, Alzubaidi M, Shah H, Alam T, Shah Z, et al. The pros and cons of using ChatGPT in medical education: a scoping review. Stud Health Technol Inform 2023 Jun 29;305:644-647. [doi: 10.3233/SHTI230580] [Medline: 37387114]
3. Lucas HC, Upperman JS, Robinson JR. A systematic review of large language models and their implications in medical education. Med Educ 2024 Nov;58(11):1276-1285. [doi: 10.1111/medu.15402] [Medline: 38639098]
4. Ye H, Xu J, Huang D, Xie M, Guo J, Yang J, et al. Assessment of large language models' performances and hallucinations for Chinese postgraduate medical entrance examination. Discov Educ 2025 Mar 13;4(1):59 [FREE Full text] [doi: 10.1007/s44217-025-00446-1]
5. Flodén J. Grading exams using large language models: a comparison between human and AI grading of exams in higher education using ChatGPT. British Educational Res J 2024 Sep 16;51(1):201-224. [doi: 10.1002/berj.4069]
6. Quah B, Zheng L, Sng TJH, Yong CW, Islam I. Reliability of ChatGPT in automated essay scoring for dental undergraduate examinations. BMC Med Educ 2024 Sep 03;24(1):962 [FREE Full text] [doi: 10.1186/s12909-024-05881-6] [Medline: 39227811]
7. Law AK, So J, Lui CT, Choi YF, Cheung KH, Kei-Ching Hung K, et al. AI versus human-generated multiple-choice questions for medical education: a cohort study in a high-stakes examination. BMC Med Educ 2025 Feb 08;25(1):208 [FREE Full text] [doi: 10.1186/s12909-025-06796-6] [Medline: 39923067]
8. Emirtekin E. Large language model-powered automated assessment: a systematic review. Applied Sciences 2025 May 20;15(10):5683 [FREE Full text] [doi: 10.3390/app15105683]
9. Fernández AA, López-Torres M, Fernández JJ, Vázquez-García D. ChatGPT as an instructor's assistant for generating and scoring exams. J Chem Educ 2024 Aug 14;101(9):3780-3788. [doi: 10.1021/acs.jchemed.4c00231]
10. Elsayed H. The impact of hallucinated information in large language models on student learning outcomes: a critical examination of misinformation risks in AI-assisted education. Northern Reviews on Algorithmic Research, Theoretical Computation, and Complexity. URL: https://northernreviews.com/index.php/NRATCC/article/view/2024-08-07/7 [accessed 2025-02-28]
11. Jia Q, Cui J, Xi R, Liu C, Rashid P, Li R, et al. On assessing the faithfulness of LLM-generated feedback on student assignments. 2024 Presented at: Proceedings of the 17th International Conference on Educational Data Mining; July 14-17, 2024; Atlanta, GA URL: https://zenodo.org/records/12729868 [doi: 10.5281/zenodo.12729867]
12. Zhou J, Zhang J, Wan R, Cui X, Liu Q, Guo H, et al. Integrating AI into clinical education: evaluating general practice trainees' proficiency in distinguishing AI-generated hallucinations and impacting factors. BMC Med Educ 2025 Mar 19;25(1):406 [FREE Full text] [doi: 10.1186/s12909-025-06916-2] [Medline: 40108629]
13. Mondal H, Karri JKK, Ramasubramanian S, Mondal S, Juhi A, Gupta P. A qualitative survey on perception of medical students on the use of large language models for educational purposes. Adv Physiol Educ 2025 Mar 01;49(1):27-36 [FREE Full text] [doi: 10.1152/advan.00088.2024] [Medline: 39447120]
14. Liu F, Chang X, Zhu Q, Huang Y, Li Y, Wang H. Assessing clinical medicine students' acceptance of large language model: based on technology acceptance model. BMC Med Educ 2024 Nov 03;24(1):1251 [FREE Full text] [doi: 10.1186/s12909-024-06232-1] [Medline: 39490999]
15. Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. J Med Internet Res 2023 Oct 04;25:e50638 [FREE Full text] [doi: 10.2196/50638] [Medline: 37792434]
16. Davis FD, Granić A. The Technology Acceptance Model: 30 Years of TAM. Cham: Springer International Publishing; Mar 04, 2024.
17. Granić A, Marangunić N. Technology acceptance model in educational context: a systematic literature review. Brit J Educational Tech 2019 Jul 09;50(5):2572-2593. [doi: 10.1111/bjet.12864]

XSL•FO
RenderX

18. Lee DY, Lehto MR. User acceptance of YouTube for procedural learning: an extension of the technology acceptance model. Computers & Education 2013 Feb;61:193-208. [doi: 10.1016/j.compedu.2012.10.001]
19. von Elm E, Altman D, Egger M, Pocock S, Gøtzsche PC, Vandenbroucke J, STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Ann Intern Med 2007 Oct 16;147(8):573-577 [FREE Full text] [doi: 10.7326/0003-4819-147-8-200710160-00010] [Medline: 17938396]
20. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159-174. [Medline: 843571]
21. Cruz ROD. Pedagogical practice preferences among generational groups of learners: towards effective twenty-first century higher education. JUTLP 2020 May 02:17(5) [FREE Full text] [doi: 10.53761/1.17.5.6]
22. Shorey S, Chan V, Rajendran P, Ang E. Learning styles, preferences and needs of generation Z healthcare students: scoping review. Nurse Educ Pract 2021 Nov;57:103247. [doi: 10.1016/j.nepr.2021.103247] [Medline: 34768214]
23. Neumann AT, Yin Y, Sowe S, Decker S, Jarke M. An LLM-driven chatbot in higher education for databases and information systems. IEEE Trans Educ 2025 Feb;68(1):103-116. [doi: 10.1109/te.2024.3467912]
24. Bernabei M, Colabianchi S, Falegnami A, Costantino F. Students' use of large language models in engineering education: a case study on technology acceptance, perceptions, efficacy, and detection chances. Computers and Education: Artificial Intelligence 2023;5:100172. [doi: 10.1016/j.caeai.2023.100172]
25. French S, Dickerson A, Mulder RA. A review of the benefits and drawbacks of high-stakes final examinations in higher education. High Educ 2023 Dec 01;88(3):893-918. [doi: 10.1007/s10734-023-01148-z]
26. Davis FD. User acceptance of information technology: system characteristics, user perceptions and behavioral impacts. International Journal of Man-Machine Studies 1993 Mar;38(3):475-487. [doi: 10.1006/imms.1993.1022]
27. Nazir A, Wang Z. A comprehensive survey of ChatGPT: advancements, applications, prospects, and challenges. Meta Radiol 2023 Sep;1(2):2 [FREE Full text] [doi: 10.1016/j.metrad.2023.100022] [Medline: 37901715]
28. Naujoks N, Harder B, Händel M. Testing pays off twice: potentials of practice tests and feedback regarding exam performance and judgment accuracy. Metacognition Learning 2022 Mar 18;17(2):479-498. [doi: 10.1007/s11409-022-09295-x]
29. Faherty A, Counihan T, Kropmans T, Finn Y. Inter-rater reliability in clinical assessments: do examiner pairings influence candidate ratings? BMC Med Educ 2020 May 11;20(1):147 [FREE Full text] [doi: 10.1186/s12909-020-02009-4] [Medline: 32393228]
30. Zhang D, Boey M, Tan YY, Jia AHS. Evaluating large language models for criterion-based grading from agreement to consistency. NPJ Sci Learn 2024 Dec 30;9(1):79 [FREE Full text] [doi: 10.1038/s41539-024-00291-1] [Medline: 39738131]
31. Zhui L, Fenghe L, Xuehu W, Qining F, Wei R. Ethical considerations and fundamental principles of large language models in medical education: viewpoint. J Med Internet Res 2024 Aug 01;26:e60083 [FREE Full text] [doi: 10.2196/60083] [Medline: 38971715]
32. Ji Z, Yu T, Xu Y, Lee N, Ishii E, Fung P. Towards mitigating LLM hallucination via self reflection. In: Findings of the Association for Computational Linguistics. 2023 Dec Presented at: EMNLP 2023; Dec 6-10, 2023; Singapore p. 1827-1843. [doi: 10.18653/v1/2023.findings-emnlp.123]

Abbreviations

AI: artificial intelligence
IRR: interrater reliability
LLM: large language model
MMed: Master of Medicine
PASS: Personalized Anesthesia Study Support
SAQ: short answer question
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
TAM: technology acceptance model

Edited by A Mavragani; submitted 15.05.25; peer-reviewed by M Burns, F Fukuzawa; comments to author 12.10.25; revised version received 07.11.25; accepted 10.11.25; published 03.12.25.

Please cite as:
Leong YH, Nambiar L, Tay VYJ, Lie SA, Yuhe K
Feasibility of a Specialized Large Language Model for Postgraduate Medical Examination Preparation: Single-Center Proof-Of-Concept Study
JMIR Form Res 2025;9:e77580
URL: https://formative.jmir.org/2025/1/e77580
doi:10.2196/77580
PMID:

XSL•FO
RenderX

©Yun Hao Leong, Lathiga Nambiar, Victoria Y J Tay, Sui An Lie, Ke Yuhe. Originally published in JMIR Formative Research (https://formative.jmir.org), 03.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Self-Reported Triggers Evaluation of High-Risk Dietary and Environmental Factors Preceding Migraine Onset by Using a Mobile Tracking App (Migraine Insight): Comparative Analysis Study

Christina Wornom[1*], BS; Brooklyn Brekke-Kumley[1*], BS; Tavsimran Luthra[2*], BS, MD; Lynn J Smith[3*], BS; Jane C Harrington[1*], PhD

[1]Rocky Vista University, Montana College of Osteopathic Medicine, Billings, MT, United States

[2]Department of Medicine, Albert Einstein College of Medicine, Montefiore Medical Center, Bronx, NY, United States

[3]Migraine Insight, Minneapolis, MN, United States

[*]all authors contributed equally

Corresponding Author:
Christina Wornom, BS
Rocky Vista University
Montana College of Osteopathic Medicine
4130 Rocky Vista Way
Billings, MT, 59106
United States
Phone: 1 8034121212
Email: christy.wornom@mt.rvu.edu

Abstract

Background: Migraines are a significant health concern affecting millions of individuals, often requiring habitual tracking of potential triggers to mitigate or predict episodes. Digital health tools such as mobile apps offer a scalable solution for personalized tracking and pattern recognition. Migraine Insight is one such app that facilitates daily logging to quantitatively assess individualized patterns of events preceding migraine onset. However, while various triggers have been reported in migraine literature, there is limited large-scale electronic user-driven data on the frequency and relative impact of specific triggers.

Objective: This study aims to address this gap by analyzing user-reported data from Migraine Insight to identify the most frequently reported triggers and evaluate their potential associations with migraine onset.

Methods: Food-associated migraine triggers were identified by performing a noninterventional, retrospective analysis of self-reported data obtained via the Migraine Insight app. A collaboration was made with the representatives of the Migraine Insight app to extrapolate the data needed for the study. A preliminary keyword list was assessed from a raw data set of 2605 data entry values, extracted from a 30-day period of September 19 to October 18, 2021, to identify high ranking self-reported events, classified by dietary habits, environmental conditions, body physiology, and medications. The variables examined included the following: dietary items, environmental conditions, body physiology, and electronics. Food items were combined into similar groups, considering variable spelling and descriptions of self-reported events. The association of migraine onset after consumption of top 5 dietary products was evaluated to establish the frequency of migraine episodes for all users who reported the items.

Results: Collectively, food (n=353) and beverage (n=252) totaled the highest reported entries, with chocolate, wine, tea, coffee, and cheese as the highest ranked foods for prevalence of reporting across all users and for frequency of migraine onset for users who consumed the items within 48 hours. The 4 highest nonfood entries were altered sleep patterns (n=245), stress or anxiety (n=199), rain or storm conditions (n=192), and bright light or brightness (n=191). Statistical analysis showed that chocolate was the only food trigger significantly associated with migraines (P=.003 vs 50%; P=.04 vs average). Consumption of tea approached significance (P=.051), while consumption of coffee, cheese, and wine were not significant. These findings suggest that chocolate is the most consistent dietary trigger.

Conclusions: High-risk foods, environmental conditions, stress, and lighting with highest prevalence reporting have previously been reported in medical literature, implicating that a migraine tracking app is a valid alternative to paper-based diaries.

XSL•FO
RenderX

(JMIR Form Res 2025;9:e59951) doi:10.2196/59951

KEYWORDS

migraines; headache; environmental triggers; food triggers; food allergies; health tracking application; tracking application; application; app; environmental; migraine; food; diet; dietary; mobile tracking app; comparative analysis; user data; user; tracking

Introduction

Migraines are a chronic neurological condition characterized by episodic attacks that can be debilitating. Identifying potential triggers is a crucial aspect of migraine management, and health care providers often recommend symptom tracking to help patients recognize patterns. Traditionally, migraine diaries have been recorded on paper; however, retrospective reporting following an episode may introduce recall bias, leading to an overemphasis on certain factors [1]. Digital health tools such as mobile apps offer a more systematic and real-time approach to tracking potential triggers.

While migraines have been extensively studied, pinpointing specific, universal triggers remains challenging. Surveys and clinical studies suggest that common natural triggers include dietary factors, stress, weather changes, and fasting [2]. Emerging evidence highlights the role of the gut-brain axis in migraine pathophysiology, with gut microbiota influencing neurological function, stress responses, and migraine susceptibility [3,4]. A significant proportion of migraine sufferers report food-related triggers, with dairy, chocolate, and alcohol being the most frequently cited [5]. Additionally, dietary modifications such as reducing red meat and carbohydrate intake have been associated with a 30% decrease in migraine frequency and severity in certain populations [6]. Caffeine intake is another dietary factor of interest, with studies suggesting a statistically significant association between high caffeine consumption and increased migraine risk [7]. Moreover, immune-mediated responses to food allergens, including IgE-mediated reactions to eggs, milk, seafood, and animal meats, have been implicated in exacerbating migraine episodes, further supporting an inflammatory component in migraine pathogenesis [8].

Despite growing evidence linking diet to migraines, current research lacks definitive conclusions, and the complexity of dietary triggers necessitates further investigation. Additionally, unbiased and large-scale datasets are needed to assess the real-world patterns of migraine onset. Studies analyzing the effects of stress during the COVID-19 pandemic have demonstrated the value of comprehensive, real-time data collection in understanding migraine patterns [9]. Migraine Insight is a mobile app designed to help users track potential triggers, including dietary habits and lifestyle changes, independently of migraine episodes. Using machine-learning algorithms, the app generates individualized reports on high- and low-risk triggers by evaluating combinatorial factors over time.

This study aims to analyze user-reported data from the Migraine Insight app to identify the most frequently reported dietary, environmental, and behavioral triggers and assess their alignment with existing literature. We hypothesized that certain dietary items, particularly chocolate and caffeine-containing beverages, would be significantly associated with migraine episodes, and that app-based data collection would provide results consistent with previous reported triggers in the medical literature. By evaluating trigger patterns within a fixed data set, this research seeks to validate the app's potential utility in migraine management and provide further insights into the multifactorial nature of migraine triggers.

Methods

Study Design

This study uses a noninterventional, retrospective analysis of self-reported data collected via the Migraine Insight mobile app (Multimedia Appendix 1). A collaboration was established with representatives from Migraine Insight to access and extract the relevant dataset. Participants voluntarily subscribed to the app, which offers both free and paid versions with varying features. All users provided informed consent in an "Opt in" manner discussed in the Apps privacy statement of self-reported deidentified data to third parties for research purposes.

An anonymized dataset containing entries from approximately 10,000 US-based users was extracted, covering a 3-year period from November 2018 to November 2021. To ensure data reliability, users who interacted with the app only once were excluded from the analysis. This exclusion was necessary because meaningful pattern recognition—especially temporal associations between triggers and migraine onset—requires multiple data points per participant. Single-use entries do not provide sufficient longitudinal information to evaluate consistent trigger-event relationships.

The dataset included user-reported migraine occurrences and associated potential triggers. Data were analyzed using Microsoft Excel. The variables examined included dietary items, environmental conditions, physiological states, and exposure to electronics. The reported food items were standardized into common categories to address variation in spelling and terminology.

Food-related triggers were classified into high-risk and low-risk groups based on the frequency with which they were reported in association with migraine events. High-risk food items were defined as those with a prevalence of ≥2% in the dataset, while low-risk items were those with <2% prevalence. The 2% threshold was chosen to balance inclusivity with relevance, highlighting commonly reported foods that could meaningfully impact population-level analyses. High-risk foods identified included coffee, tea, chocolate, wine, and cheese. Low-risk foods included meat, food additives, fruits, nuts, carbohydrates, soft beverages, beer, and high-sodium foods. The frequencies of reported high-risk foods and corresponding migraine onsets were further analyzed, with duplicates from the same user excluded to prevent skewed results from repetitive entries.

XSL•FO
RenderX

Ethical Considerations

The investigations of this study were compliant with ethical practices established by the Nuremburg Code. Participants of the Migraine Insight tracking app provided consent for the utilization of data when agreeing to terms of use. To ensure the protection of user privacy, all identifying information was removed from the dataset prior to analysis. None of the data points, contained names, contact information, GPS coordinates, or other personally identifiable information were preserved. The anonymization process was designed to align with best practices in digital health research and data ethics, minimizing the risk of reidentification. Given the retrospective nature of the study and the use of fully anonymized, pre-existing data, the project was reviewed by the institutional review board at St. George's University and was deemed to meet the criteria for exemption from full review. The ethical considerations in this study reflect a commitment to participant confidentiality, data security, and the responsible use of mobile health technologies in clinical research.

Results

This analysis focused on dietary triggers reported by users of the Migraine Insight app during a 30-day sampling window (September-October 2021), using a dataset derived from over 10,000 users. The app enables users to log in migraine episodes and associated lifestyle, environmental, and dietary exposures. From the raw data, entries were cleaned and grouped based on variations in spelling and terminology for consistency. Figures 1 and 2 depict the overall frequency patterns. Within 48 hours of a migraine episode, the most frequently reported categories were dietary and environmental factors. Dietary exposures were classified into high-risk food items (≥2% prevalence), low-risk food items (<2% prevalence), and fasting conditions. Reported frequencies included high-risk foods (400/1786, 22.4%), low-risk foods (94/1918, 4.9%), and fasting (127/1896, 6.7%) (Figure 1).

Among high-risk foods, caffeinated beverages (coffee and tea) were the most commonly reported, followed by chocolate, wine, and cheese. When examining the percentage of exposures that resulted in migraine onset, chocolate led with 61.8% (104/169), followed by tea (75/127, 59.4%), coffee (139/266, 52.2%), cheese (67/144, 46.8%), and wine (55/119, 46.5%). These findings are summarized in Table 1 below, which includes confidence intervals and statistical significance testing.

Figure 1. Self-reported migraine occurrence within 48 hours of consuming high-risk and low-risk food items. This stacked bar graph displays the percentage of self-reported migraines within 48 hours of consuming high-risk and low-risk food items tracked by the Migraine Insight app. Food items were combined into similar groups, considering variable spelling and descriptions of self-reported events. High-risk food items include caffeine, tea, chocolate, wine, liquor, and dairy, while low-risk food items include meat, food additives, fruits, nuts, carbohydrates, soft beverages, beer, and high-sodium–containing foods.

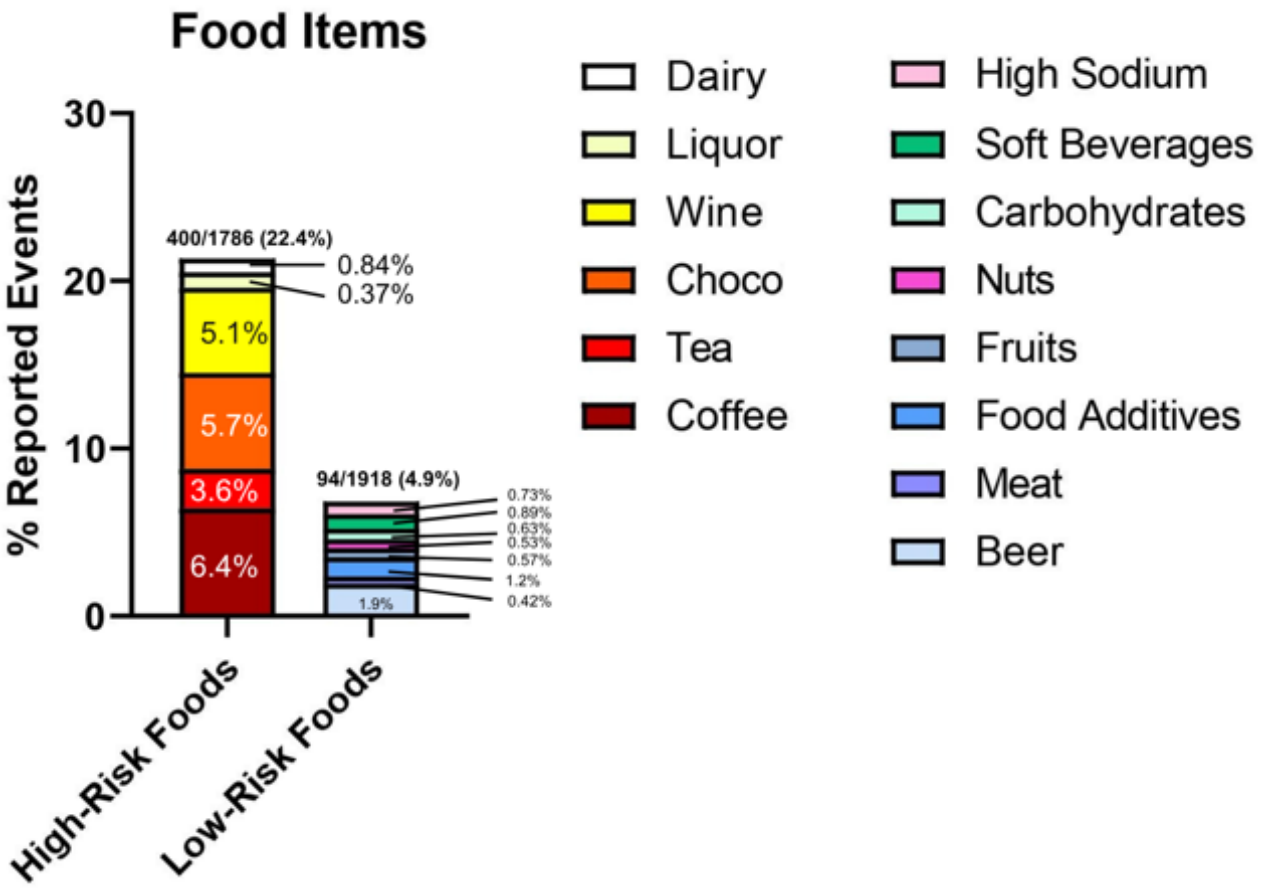

XSL•FO
RenderX

Figure 2. Self-reported migraine occurrence within 48 hours of encountering migraine-associated factors. This stacked bar graph depicts the percentage of self-reported migraines within 48 hours of encountering various migraine-associated factors tracked by the Migraine Insight app. The analysis grouped dietary items into high-risk food items (such as tea, coffee, chocolate, wine, liquor, and dairy), low-risk food items (such as meat, food additives, fruits, nuts, carbohydrates, soft beverages, beer and high-sodium containing foods), and fasting conditions. Environmental factors that were associated with indoor environments such as room lighting were grouped accordingly, whereas factors associated with outdoors, such as weather, were grouped under outside environment. Inside and outside environment-associated factors were grouped under environment. Physiological factors, including pain, fatigue, and stress/anxiety, were grouped into a body physiology category. Medications such as triptans and non-triptans were grouped together, while various electronic devices such as computer and phone screens were grouped under an electronics category. The grouping of dietary items was based on the consideration of variable spelling and descriptions of self-reported events.

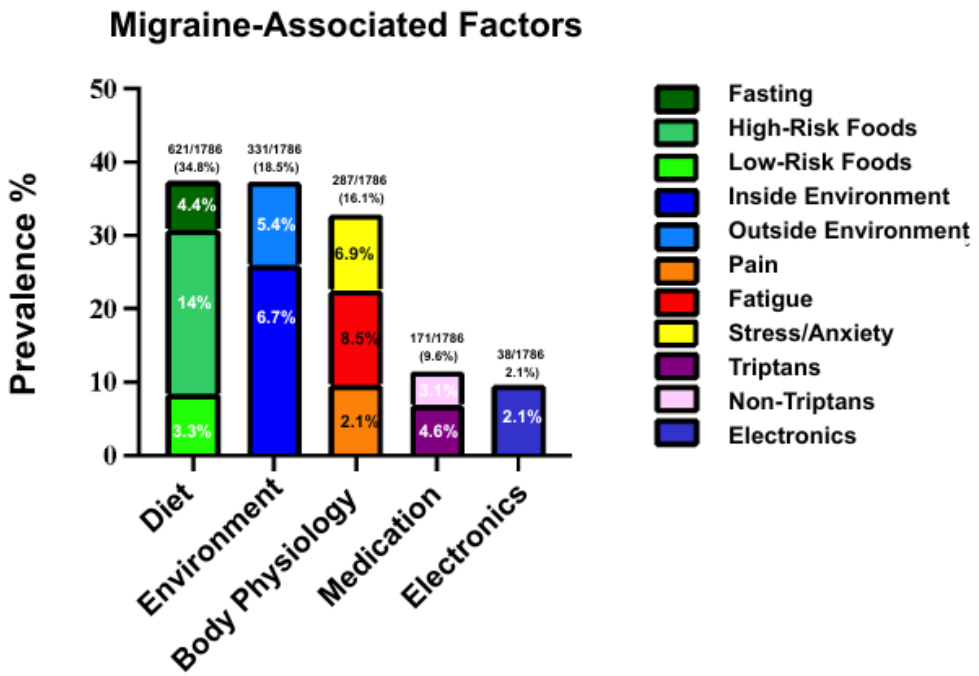

Table 1. Migraine triggers with confidence intervals and statistical significance. This table shows a summary of high-risk dietary triggers (coffee, chocolate, cheese, tea, and wine), showing exposure counts, migraine occurrence rates, 95% CIs, and *P* values. Significance was tested against a 50% threshold and the overall average (53.3%). Chocolate was the only trigger with statistically significant results in both comparisons (*P*<.05).

Trigger	Total exposures, n	Resulted in migraine, n	No migraine, n	Resulted in migraine (%) (95% CI)	*P* (vs 50%)	Sig vs 50%[a]	*P* value vs average	Sig vs avg[b]
Coffee	266	139	127	52.2 (46.3-58.2)	.50	c	.76	c
Chocolate	169	104	65	61.8 (54.0-68.5)	.003	d	.04	d
Cheese	144	67	77	46.8 (38.6-54.7)	.45	c	.11	c
Tea	127	75	52	59.4 (50.4-67.2)	.05	c	.21	c
Wine	119	55	64	46.5 (37.5-55.2)	.46	c	.14	c
Overall	825	440	385	46.5 (37.5-55.2)	.06	c	N/A[e]	N/A

[a]Tests if trigger differs from chance (50%) versus average.

[b]Tests if trigger differs from overall average across all triggers.

[c]Statistically not significant versus 50%.

[d]Statistically significant at *P*<.05.

[e]N/A: not applicable.

Statistical analysis using proportion tests (binomial test vs 50% and vs overall average of 53.3%) revealed that chocolate was the only food trigger that was statistically significant compared to both the 50% threshold (*P*=.003) and the overall average (*P*=.04). Tea approached significance compared to 50% (*P*=.05) but did not meet the conventional *P*<.05 threshold. No other

XSL•FO
RenderX

individual food reached statistical significance, including coffee, cheese, and wine. These findings suggest that while all 5 foods were commonly reported prior to migraines, chocolate stands out as a more consistently associated trigger, both by user reporting (Table 2) and statistical analysis (Table 1).

Table 2. Frequency of migraine occurrence within 48 hours of consuming certain food items.

Food item	Users tracking the trigger, n	Average trigger resulting in a migraine within 48 h, n (%)
Chocolate	169	104 (61.8)
Tea	127	75 (59.4)
Coffee	266	139 (52.2)
Cheese	144	67 (46.8)
Wine	119	55 (46.5)

Table 2 displays the average percentage of times users of the Migraine Insight app reported experiencing a migraine within 48 hours of consuming specific food items, along with the number of users that are tracking that food item. The data were collected through self-reported triggers and tracked by the app. The food items listed were reported as common triggers by many users.

Nondietary triggers were also assessed. Among these, altered sleep patterns (n=245), stress or anxiety (n=199), rain/storm conditions (n=192), and bright light (n=191) were the most frequently reported within 48 hours of migraine onset. These environmental and behavioral factors were frequently reported independently or in conjunction with dietary triggers, though formal multivariate analysis was not conducted in this study.

Physiological triggers, including fatigue (n=245) and pain (n=61), were grouped into a broader “body physiology” category. Medication use, including both triptans (n=131) and non-triptan medications (n=88), and screen exposure (eg, phone, computer use) were also reported. Triptans were captured through various spellings and brand names (eg, Sumatriptan, Rizatriptan), and non-triptan medications included over the counter and prescription agents such as hydroxyzine, Zomig, Toradol, and prochlorperazine.

Discussion

Principal Findings

This study identifies high-risk environmental, dietary, and behavioral factors preceding migraine onset by using data from the Migraine Insight app. The patterns observed align with previous research highlighting certain foods and environmental conditions as potential migraine triggers. High-risk items, including chocolate, tea, coffee, wine, and cheese, emerged as significant contributors, corroborating findings by Martin and Vij [10,11], who identified chocolate, caffeine, and alcohol as common dietary triggers. Similarly, data from Migraine Insight users support research by Nowaczewska et al [12], emphasizing the role of chocolate and aged cheese in migraine risk.

Among dietary triggers, tea and coffee were the second and third most frequently reported triggers (Table 2). There were 127 reports of tea triggering migraines from 45 users, 9 of whom also reported coffee as a trigger. In total, 123 users reported migraines triggered by coffee, with 266 occurrences. Tea consumption led to migraines in 59.4% (75/127) of the users who reported its intake, while coffee triggered migraines in 52.2% (139/266) of the users. One possible explanation for tea’s higher likelihood of inducing migraines is its greater tannin content compared to coffee. Tannins, a type of flavonoid phenol, are found in various dietary sources [13]. A study analyzing tannin levels in both beverages found that roasted coffee beans contain 18 (SD 1.7) $mg \cdot g^{1}$ of tannins, whereas tea contains nearly twice that amount at 37 (SD 2.6) $mg \cdot g^{1}$ [14]. Tannins are also present in high concentrations in chocolate, wine, and dairy—all of which were frequently reported as migraine triggers [15].

Beyond dietary influences, nondietary factors such as weather, light, and stress have been widely recognized as migraine triggers. Stress, in particular, is commonly cited, with a study of over 1200 participants reporting that 80% identified stress as a migraine trigger [16-18]. Similarly, an analysis of 1027 individuals who met the International Headache Society's migraine criteria found that 53.2% reported weather changes as a precipitating factor [19]. Another study identified weather as the second most frequently perceived trigger after stress [18]. Light exposure also emerged as a significant factor, particularly among migraineurs with aura, where 75% identified light as a trigger compared to 35% of those without aura [20].

The use of a mobile tracking app for data collection represents a novel and valuable approach in migraine research. Unlike traditional paper-based diaries, which are prone to recall bias, real-time data logging offers a more accurate and comprehensive view of potential triggers. This method, as demonstrated in our study, aligns with existing literature and enhances the reliability of self-reported data [7,21]. Furthermore, the impact of COVID-19 lockdowns on stress-induced migraine attacks, as studied by Kato et al [9], underscores the value of leveraging unbiased datasets in migraine research.

While this study provides meaningful insights into the dietary, environmental, and behavioral factors preceding migraines, certain limitations must be acknowledged. The reliance on self-reported data introduces the potential for recall bias, as participants may not accurately remember or report their exposures. Additionally, the self-selecting nature of the participant pool may lead to selection bias, limiting the generalizability of the findings. Although mobile apps provide real-time data collection that reduces retrospective inaccuracies,

XSL•FO
RenderX

they also introduce challenges such as inconsistent user engagement and variations in reporting accuracy. Traditional methods such as structured clinical assessments or controlled dietary studies may offer more standardized data but lack the real-world applicability and convenience of mobile tracking. Future research should aim to balance these methodologies by incorporating objective biomarkers or clinician-verified reports alongside mobile app data. Expanding the participant base to encompass more diverse populations and employing more controlled methodologies will further strengthen the validity of these findings. Addressing these gaps will contribute to a deeper understanding of migraine pathophysiology and improved management strategies.

Conclusion

This study highlights the potential of mobile health apps such as Migraine Insight to identify high-risk dietary, environmental, and behavioral factors associated with migraine onset. The findings validate existing literature by confirming that commonly cited triggers—especially chocolate, tea, and coffee—remain prominent among real-world users. Notably, chocolate demonstrated a statistically significant association with migraine episodes, reinforcing its relevance in dietary counseling for migraine prevention.

Beyond individual triggers, this research underscores the utility of app-based, real-time data collection as a reliable alternative to traditional migraine diaries. The ability to capture consistent, personalized data offers new opportunities for understanding complex, multifactorial patterns in migraine occurrence. While limitations related to self-reporting, recall bias, and user engagement remain, the insights gained from large-scale, user-generated datasets provide a strong foundation for further investigation.

Future research should aim to integrate app-based tools with clinical assessments and objective measures such as biomarkers to enhance validity. Expanding participation to include more diverse populations and exploring longitudinal outcomes will further strengthen the evidence base. Ultimately, mobile tracking technologies hold significant promise in advancing personalized migraine management and improving the quality of life for individuals living with these conditions.

Acknowledgments

The original manuscript did not use generative artificial intelligence. During the revision of this work, the authors used OpenAI's ChatGPT in order to assist with language editing and improving the clarity of expression (Multimedia Appendix 2). After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Funding Statement

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Conflicts of Interest

The authors of the study report established partnerships with Migraine Insight, including founder and CEO LJS and academic consultant JCH. The lead and senior authors of the study hold academic roles (faculty and students) at Rocky Vista University and St. George's University and have no nonacademic affiliations.

Multimedia Appendix 1
Analysis of the triggers.
[XLSX File (Microsoft Excel File), 165 KB - formative_v9i1e59951_app1.xlsx]

Multimedia Appendix 2
Document containing original prompts/responses on file from ChatGPT used in the editing of this paper.
[DOCX File , 18 KB - formative_v9i1e59951_app2.docx]

References

1. Nappi G, Jensen R, Nappi RE, Sances G, Torelli P, Olesen J. Diaries and calendars for migraine. A review. Cephalalgia 2006 Aug;26(8):905-916. [doi: 10.1111/j.1468-2982.2006.01155.x] [Medline: 16886925]
2. Marmura MJ. Triggers, protectors, and predictors in episodic migraine. Curr Pain Headache Rep 2018 Oct 05;22(12):81. [doi: 10.1007/s11916-018-0734-0] [Medline: 30291562]
3. Hindiyeh N, Aurora SK. What the gut can teach us about migraine. Curr Pain Headache Rep 2015 Jul;19(7):33. [doi: 10.1007/s11916-015-0501-4] [Medline: 26049770]
4. Mayer EA, Tillisch K, Gupta A. Gut/brain axis and the microbiota. J Clin Invest 2015 Mar 02;125(3):926-938 [FREE Full text] [doi: 10.1172/JCI76304] [Medline: 25689247]

XSL•FO
RenderX

5. Finocchi C, Sivori G. Food as trigger and aggravating factor of migraine. Neurol Sci 2012 May;33 Suppl 1:S77-S80. [doi: 10.1007/s10072-012-1046-5] [Medline: 22644176]
6. Altamura C, Cecchi G, Bravo M, Brunelli N, Laudisio A, Caprio PD, et al. The healthy eating plate advice for migraine prevention: an interventional study. Nutrients 2020 May 28;12(6):1579 [FREE Full text] [doi: 10.3390/nu12061579] [Medline: 32481555]
7. Mostofsky E, Mittleman MA, Buettner C, Li W, Bertisch SM. Prospective cohort study of caffeinated beverage intake as a potential trigger of headaches among migraineurs. Am J Med 2019 Aug;132(8):984-991 [FREE Full text] [doi: 10.1016/j.amjmed.2019.02.015] [Medline: 31402050]
8. Zhao Z, Jin H, Yin Y, Hou Y, Wang J, Tang C, et al. Association of migraine with its comorbidities and food specific immunoglobulin G antibodies and inflammatory cytokines: cross-sectional clinical research. J Pain Res 2021;14:2359-2368 [FREE Full text] [doi: 10.2147/JPR.S316619] [Medline: 34385841]
9. Kato Y, Poh W, Horvath Z, Cadiou F, Shimazu T, Maruki Y. Impact of COVID-19 pandemic on migraine management in the United States: insights from migraine tracking app users. BMC Neurol 2021 Sep 09;21(1):345 [FREE Full text] [doi: 10.1186/s12883-021-02378-3] [Medline: 34503456]
10. Martin VT, Vij B. Diet and headache: part 1. Headache 2016 Oct;56(9):1543-1552. [doi: 10.1111/head.12953] [Medline: 27699780]
11. Martin VT, Vij B. Diet and headache: part 2. Headache 2016 Oct;56(9):1553-1562. [doi: 10.1111/head.12952] [Medline: 27699772]
12. Nowaczewska M, Wiciński M, Kaźmierczak W, Kaźmierczak H. To eat or not to eat: a review of the relationship between chocolate and migraines. Nutrients 2020 Feb 26;12(3):608 [FREE Full text] [doi: 10.3390/nu12030608] [Medline: 32110888]
13. Rudrapal M, Khairnar SJ, Khan J, Dukhyil AB, Ansari MA, Alomary MN, et al. Dietary polyphenols and their role in oxidative stress-induced human diseases: insights into protective effects, antioxidant potentials and mechanism(s) of action. Front Pharmacol 2022;13:806470 [FREE Full text] [doi: 10.3389/fphar.2022.806470] [Medline: 35237163]
14. Savolainen H. Tannin content of tea and coffee. J Appl Toxicol 1992 Jun;12(3):191-192. [doi: 10.1002/jat.2550120307] [Medline: 1629514]
15. Sharma K, Kumar V, Kaur J, Tanwar B, Goyal A, Sharma R, et al. Health effects, sources, utilization and safety of tannins: a critical review. Toxin Reviews 2019 Sep 18;40(4):432-444 [FREE Full text] [doi: 10.1080/15569543.2019.1662813]
16. Stubberud A, Buse DC, Kristoffersen ES, Linde M, Tronvik E. Is there a causal relationship between stress and migraine? Current evidence and implications for management. J Headache Pain 2021 Dec 20;22(1):155 [FREE Full text] [doi: 10.1186/s10194-021-01369-6] [Medline: 34930118]
17. Kelman L. The triggers or precipitants of the acute migraine attack. Cephalalgia 2007 May;27(5):394-402. [doi: 10.1111/j.1468-2982.2007.01303.x] [Medline: 17403039]
18. Pellegrino ABW, Davis-Martin RE, Houle TT, Turner DP, Smitherman TA. Perceived triggers of primary headache disorders: a meta-analysis. Cephalalgia 2018 May;38(6):1188-1198 [FREE Full text] [doi: 10.1177/0333102417727535] [Medline: 28825314]
19. Kesserwani H. Migraine triggers: an overview of the pharmacology, biochemistry, atmospherics, and their effects on neural networks. Cureus 2021 Apr 01;13(4):e14243 [FREE Full text] [doi: 10.7759/cureus.14243] [Medline: 33954064]
20. Denney DE, Lee J, Joshi S. Whether weather matters with migraine. Curr Pain Headache Rep 2024 Apr;28(4):181-187 [FREE Full text] [doi: 10.1007/s11916-024-01216-8] [Medline: 38358443]
21. Spierings EL, Ranke AH, Honkoop PC. Precipitating and aggravating factors of migraine versus tension-type headache. Headache 2001 Jun;41(6):554-558. [doi: 10.1046/j.1526-4610.2001.041006554.x] [Medline: 11437890]

Edited by A Mavragani; submitted 26.04.24; peer-reviewed by MS Khan; comments to author 20.11.24; revised version received 19.10.25; accepted 20.10.25; published 03.12.25.

Please cite as:
Wornom C, Brekke-Kumley B, Luthra T, Smith LJ, Harrington JC
Self-Reported Triggers Evaluation of High-Risk Dietary and Environmental Factors Preceding Migraine Onset by Using a Mobile Tracking App (Migraine Insight): Comparative Analysis Study
JMIR Form Res 2025;9:e59951
URL: https://formative.jmir.org/2025/1/e59951
doi:10.2196/59951
PMID:

©Christina Wornom, Brooklyn Brekke-Kumley, Tavsimran Luthra, Lynn J Smith, Jane C Harrington. Originally published in JMIR Formative Research (https://formative.jmir.org), 03.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use,

XSL•FO
RenderX

distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Mechanism of Online Health Information Seeking Switching to Online Medical Consultation: Cross-Sectional Study

Lijiang Zhang[1,2*], PhD; Jingjing Xia[1*], MS; Hui Chen[1*], MS; Yang Bai[3,4,5], PhD; Jun Wang[4,5], PhD; Liuan Wang[6,7,8], PhD; Wenjie Ren[2], PhD

[1]School of Health Management, Henan Medical University, Xinxiang, Henan, China
[2]Institutes of Health Central Plains, Henan Medical University, Xinxiang, Henan, China
[3]School of Public Administration and Policy, Renmin University of China, Beijing, China
[4]Center for Health Policy Research and Evaluation, Renmin University of China, Beijing, China
[5]Healthy China Institute, Renmin University of China, Beijing, China
[6]School of Management, Beijing Institute of Technology, Beijing, China
[7]Social Science Domain, Beijing Institute of Technology, Zhuhai, China
[8]Qionghai Hospital of Traditional Chinese Medicine, Qionghai, China
[*]these authors contributed equally

Corresponding Author:
Jun Wang, PhD
Center for Health Policy Research and Evaluation
Renmin University of China
59 Zhongguancun Street Hai Dian District
Beijing, 100872
China
Phone: 86 8611290668
Email: hpreruc2022@126.com

Abstract

Background: Internet health care plays a crucial role in addressing the challenge of distributing high-quality medical resources and promoting the optimal allocation of these resources and health equity in China. Online medical consultation (OMC) plays a more significant role than online health information seeking (OHIS). Currently, the proportion of Chinese patients using OMC is low. Therefore, it is essential to enhance patient engagement with OMC and fully leverage the role of internet health care in optimizing the allocation of medical resources.

Objective: This study aims to explore the correlation mechanisms of online medical community users' switching behaviors from OHIS to OMC.

Methods: This study is based on the knowledge-attitude-practice theory, which combines the social support theory and the health belief model to construct a research model of users' willingness to transition from OHIS to OMC. The study adopts a questionnaire survey and structural equation modeling method to conduct an empirical study.

Results: Gaining knowledge about information support has a significant positive impact on perceived susceptibility (β=.339, $P<.001$), perceived severity (β=.348, $P<.001$), and perceived benefits (β=.361, $P<.001$), while having a significant negative impact on perceived barriers (β=–.285, $P<.001$). Gaining knowledge about emotional support positively affects perceived susceptibility (β=.220, $P<.001$) and perceived benefits (β=.149, $P<.01$) but does not significantly influence perceived severity (β=–.006, $P>.05$) or perceived barriers (β=.099, $P>.05$). Perceived susceptibility (β=.123, $P<.05$), perceived severity (β=.174, $P<.001$), and perceived benefits (β=.273, $P<.001$) positively influence patients' transition to online consultation behavior, whereas perceived barriers (β=–.112, $P<.05$) negatively impact this switch. In addition, we found that gaining knowledge about information support not only directly affects patients' behavior in switching to online consultations but also impacts patients' OMCs through perceived susceptibility (14.23%), perceived severity (13.17%), and perceived benefits (25.28%). In contrast, gaining knowledge about emotional support does not directly influence patient behavior transfer; it operates only through perceived susceptibility (46.95%) and perceived benefit (52.90%).

Conclusions: This study integrated the knowledge-attitude-practice framework, social support theory, and health belief model to uncover the internal logic of patients' behavioral transfers within online health communities. It confirmed the mediating role

XSL•FO
RenderX

of the cognitive-emotional dual-drive pathway and health beliefs. The findings provide a scientific basis for the functional design of online health care platforms and for precise health knowledge dissemination strategies.

(JMIR Form Res 2025;9:e78397) doi:10.2196/78397

KEYWORDS

online health information seeking; online medical consultation; switching behavior; knowledge-attitude-practice; social support; health belief model

Introduction

Background

The rapid integration of internet technology into the medical and health sector, along with the strategic demand for the "Healthy China" initiative, has fueled the growth of internet health care in China. Internet health care plays a crucial role in addressing the challenge of distributing high-quality medical resources and promoting the optimal allocation of these resources, as well as health equity. Among these, the online medical community is the most widely used and highly recognized model of internet health care service [1]. Patients' behavior in using online health care communities can be categorized into free online health information seeking (OHIS) and paid online medical consultation (OMC). The former focuses on the search, acquisition, screening, and adoption of health information [2,3]. The latter emphasizes remote consultations with physicians, as well as access to treatment options [4]. Although both can enhance public health awareness and promote access to health services [5], OMC allows quality health care resources to reach grassroots and remote areas more effectively [6], generating an optimal allocation of health care resources that cannot be achieved by simply seeking health information [7]. Meanwhile, OMC plays a more significant role than OHIS in promoting health care system reform and health equity [8]. Currently, the proportion of Chinese patients using OMC is low, and the report of "the 55th Statistical Report on the Development of the Internet in China" shows that as of December 2024, the number of internet health care users in China reached 364 million, but fewer than 10% of them use internet diagnosis and treatment services. Therefore, it is essential to enhance patient engagement with OMC and fully leverage the role of internet health care in optimizing the allocation of medical resources.

The theory of switching behavior posits that users progress through a phased transformation process when engaging with online communities, typically evolving from initial information acquisition to sustained, in-depth participation, and ultimately culminating in proactive engagement behaviors [9]. In online medical communities, patients initially gather health information to assist in their health management either passively (eg, through serendipitous encounters and browsing) or actively (via intentional search and retrieval) [10,11]. However, health information on the internet originates from a variety of sources with differing reliability, which impacts patients' evaluation and screening of health information [12]. Consequently, OHIS does not fully meet patients' health needs, making it more difficult to obtain information with high clinical guidance value [9]. OMC is a type of medical service behavior that depends on the online medical community to facilitate disease diagnosis and treatment plan determination between the doctor and the patient, using either asynchronous or synchronous methods [13]. By using OMC services, doctors can provide patients with accurate disease diagnoses and personalized treatment plans, effectively addressing the inherent drawbacks of OHIS while more efficiently meeting individuals' needs for high-quality online medical services. Consequently, OMC is steadily gaining popularity among patients.

However, OMC presents a challenging decision-making process. First, online medical communities gather a large number of doctors with diverse professional skills, clinical experiences, and qualification backgrounds, and patients often face significant decision-making pressure and cognitive load during the selection process. Second, health care services are typically trust commodities [14] that need to establish sufficient reliability to be selected. Compared with online decision-making behavior in commercial environments, such as online shopping, patients tend to be more cautious when choosing OMC services [15]. OHIS is an important antecedent behavior for patients taking OMC. It has been shown that if an online health care community provides a large number of useful health information resources, patients' knowledge and sense of empowerment increase [16], and patients are more likely to choose OMC services [17]. However, there is a lack of current research on the association between the 2 behaviors, and our literature survey reveals that existing relevant research primarily focuses on 2 aspects. One is the relationship between OHIS and offline health service utilization. Some studies suggest that OHIS promotes health service utilization, while others argue that OHIS behaviors may hinder health service access utilization [18]. The second aspect is the switching behavior in online communities, and the research primarily relies on the theory of "push-pull-anchoring" to examine the formation of transfer willingness, influencing factors, and customer loyalty, among others. The study explores the transition from OHIS to OMC behavior, which is akin to the migration between products with different characteristics within the same medium [19], similar to the shift between free and paid services or the service upgrades of online platform users [20,21]. However, the first type of research overlooks the relationship between OHIS and the use of online health services such as OMC. Moreover, the mechanistic explanation proposed by some studies—that OHIS triggers health anxiety and motivates patients to seek offline consultation—does not effectively explain the relationship between OHIS and OMC. This is because health anxiety, combined with choice overload in the online health care community [22], makes it challenging to encourage OMC behavior. The second type of research primarily focuses on explaining the behavioral transfer mechanism in nonhealth domains. However, the process of

XSL•FO
RenderX

health-related behavioral transfer is more complex, usually involving the precognitive belief change process and the final behavioral change process [23]. Therefore, the findings and theories of online behavioral transfer in nonhealth domains cannot adequately explain the OHIS to the OMC behavioral shift mechanism.

To address the gaps in these researches, this study notes that from a positive driver perspective, health information provided by health care professionals is perceived as authoritative and supportive [24]. In addition, online doctor-patient interactions posted on websites allow patients to perceive emotional support. Moreover, the knowledge related to this health information and interactions can increase patients' involvement and engagement [25]. Meanwhile, the classic theory of health behavior change, knowledge-attitude-practice (KAP), can thoroughly explain the transformation of people's cognition, beliefs, and actions. Therefore, this study proposes constructing a research framework based on the theoretical model of KAP. In addition, grounded in the theory of social support, it will serve as the foundation for the study's theory, incorporating knowledge variables (K), individual subjective belief variables (A) rooted in the health belief model (HBM), and the transition from OHIS to OMC as behavioral variables (P). This study explores whether receiving social support and external knowledge motivates patients to take action online, and whether this relationship is mediated by their health beliefs. This core question aims to offer both theoretical and practical guidance for helping patients engage in OMCs, along with recommendations for the platform's operation and development.

Theoretical Basis

KAP Model

Originating in the field of health behavior research in the mid-20th century, KAP theory is a classic model used to explain how personal knowledge and beliefs influence behavior change. It divides human behavior change into 3 successive processes: acquiring knowledge, generating beliefs, and forming behaviors, which progress layer by layer. Knowledge serves as the foundation for behavior change, while belief acts as the driving force behind behavior change. This theory has been widely used in public health [26], health behavior, health education, and other research fields. KAP theory is suitable for depicting the behavioral change process from OHIS to OMC. When first entering an online medical community, patients often browse disease-related papers and patient comments passively due to limited trust in this community. As individuals gradually gain emotional support, accumulate knowledge, or receive practical help from other patients and health care professionals through these channels (K), they build trust in the effectiveness and timeliness of the online community's medical consultations and diagnoses (A). This trust prompts them to use the OMC service (P). Therefore, this study is grounded in the KAP framework to explore the internal mechanisms of the health belief shift that occurs as patients access health information in the online medical community and how this leads to further online consultation behavior.

Social Support

Social support theory serves as the theoretical foundation for the K (knowledge) part of the KAP framework in this study. Social support refers to people's behaviors such as emotional communication, resource sharing, and information exchange within social relationship networks [27]. This includes information support, emotional support, and instrumental support [28]. In health care research, scholars typically focus more on emotional support (40%) and informational support (31.7%) frequently [29]. Information support encompasses various types of information that individuals receive from others or social networks, aiding them in solving problems, making decisions, and understanding their surroundings. This information can include knowledge about specific matters, general experience, or advice and guidance. Emotional support refers to the psychological comfort and assistance provided to individuals through the expression of care, understanding, sympathy, encouragement, trust, and other emotional factors. This support aims to meet their emotional needs, helping them cope with stress and alleviate negative emotions [27]. The online medical community builds a social relationship network. Patients, on one hand, learn practical knowledge about disease diagnosis and treatment through health care popularization and professional question and answer. On the other hand, they communicate with other patients in the community to exchange their conditions, share experiences, and obtain emotional support. The acquisition of this experiential knowledge significantly enhances patients' trust in the online community [30], which helps drive the adoption and use of upgraded health care services on the platform by patients.

Health Belief Model

HBM is a theoretical framework that explains and predicts individual health behaviors, emphasizing the essential role of cognitive and psychological factors in the decision-making process related to health behaviors. HBM argues that an individual's adoption of a specific health behavior depends on their perceptions and beliefs about health issues. This includes factors such as perceived susceptibility, perceived severity, perceived benefits, perceived barriers, cues to action, and self-efficacy [31]. In simple terms, individuals are more likely to adopt a specific health behavior if they believe they are at risk for a disease, that the disease can lead to severe consequences, and that adopting this health behavior is effective in preventing or reducing the threat of the disease. Furthermore, they should believe they can perform the behavior. In this study, the HBM was used to illustrate the A (belief) component of the KAP model. Given that the cues to action in the HBM typically refer to external triggering factors, this study focuses on an integrated model of individual internal cognition and social support, which is less closely linked to the core of the research framework. Self-efficacy refers to confidence in one's ability to perform a behavior. The actual online consultation platforms are continuously optimizing their interfaces, and the technical threshold is relatively low, which may not be a core influencing factor or can be indirectly reflected through perceived barriers. Therefore, in the HBM, this study includes only 4 constructs: perceived susceptibility, perceived severity, perceived benefits, and perceived barriers. Perceived susceptibility refers to an

XSL•FO
RenderX

individual's subjective assessment of the likelihood of experiencing health issues. Perceived severity is an individual's subjective evaluation of the potential consequences of a health problem, including damage to physical functioning, impact on social functioning, and the economic burden. Perceived benefit is an individual's belief that adopting specific health behaviors will reduce the risk of disease or lessen the consequences of disease. Perceived barriers represent individuals' perceptions of the challenges and obstacles they face in adopting health behaviors.

Research Model and Hypotheses

Supportive Information Seeking and Health Beliefs

Information support in social support theory suggests that individuals who receive information and incorporate it into their existing knowledge systems can influence the formation and transformation of related concepts and beliefs. In an online health community (OHC), patients' access to information about doctors who provide health advice or treatment plans enhances their trust in the community. The more frequent and comprehensive their exposure to such supportive knowledge, the higher their perceived susceptibility [32]. Therefore, this study hypothesized that patients' access to information support positively influences their perceived susceptibility (H1a). Having more health knowledge improves people's perception of health risks and allows for a more accurate assessment of disease severity. Tang et al [21] found that individuals with higher levels of access to social support are more likely to view the consequences of a disease as more serious. Therefore, this paper argues that patients' access to knowledge about information support positively influences their perceived severity (H1b). Moreover, it has been confirmed that seeking information online for support not only conveys knowledge about the disease but also helps patients recognize its benefits [31] and reduces perceived barriers [33], thereby positively influencing online health behaviors. Based on this, the following hypotheses are proposed: patients' acquisition of knowledge about information support positively affects their perceived benefits and negatively impacts their perceived barriers (H1c and H1d).

Emotional support in social support theory suggests that emotional support providers may convey disease-related risk or protective messages. For example, patients in online health care communities often engage in disease risk exchanges, prompting those with access to relevant information to reflect on the likelihood of their illness. Risky narratives, such as "I ignored a symptom in the first place," may also enhance individuals' perceived risk susceptibility [33]. Therefore, this paper proposes the hypothesis that "patients' acquisition of knowledge about emotional support positively affects patients' perceived susceptibility" (H2a). Second, information based on emotional support may trigger in-depth discussions about the consequences of the disease. Patients sharing their experiences with complications in a support group may lead others to take the disease more seriously [34], affecting perceived severity. This study hypothesized that patients' acquisition of knowledge about emotional support positively affects their perceived severity (H2b). Meanwhile, emotional support enhances recognition of the value of health behaviors, and patients are more convinced of the benefits of adopting online health care services after learning about treatment successes through the online health care community. In other words, patients' access to knowledge about emotional support affects their perceived benefits (H2c). Finally, emotional support can help reduce perceived barriers by providing resources such as accompanying patients to medical appointments or recommending an online doctor. It also lessens the patient's fear and anxiety about the disease indirectly [35]. Based on this, this study hypothesized that patients' acquisition of knowledge about emotional support negatively affects their perceived barriers (H2d).

Health Beliefs and the OHIS-OMC Switching Behavior

HBM emphasizes that an individual's subjective judgment of their likelihood of developing a disease (ie, perceived susceptibility) influences their health behavior decisions. When patients believe that they are susceptible to a disease, their worry and uncertainty prompt them to seek medical information and help [36], and their subjective judgment of the potential consequences of the disease (perceived severity)—specifically, the perceived health damage caused by the disease—significantly affects their health behaviors. Simultaneously, individuals' assessment of the benefits they can gain from adopting a health behavior (perceived benefits) also impacts the frequency of their actions [37]. Perceived barriers represent the various obstacles that individuals anticipate when adopting a health behavior; these negative perceptions can inhibit online inquiry behavior [38]. Therefore, based on these researches, this study concludes that patients' perceived susceptibility positively influences their transition from OHIS to OMC (H3a). In addition, patients' perceived severity positively affects their transition from online information acquisition to online consultation (H3b). Moreover, patients' perceived benefits positively impact their shift from OHIS to OMC (H3c), while patients' perceived barriers negatively affect their transition from OHIS to OMC (H3d). Based on these analyses, the theoretical research model developed in this study is illustrated in Figure 1.

XSL•FO
RenderX

Figure 1. The research model.

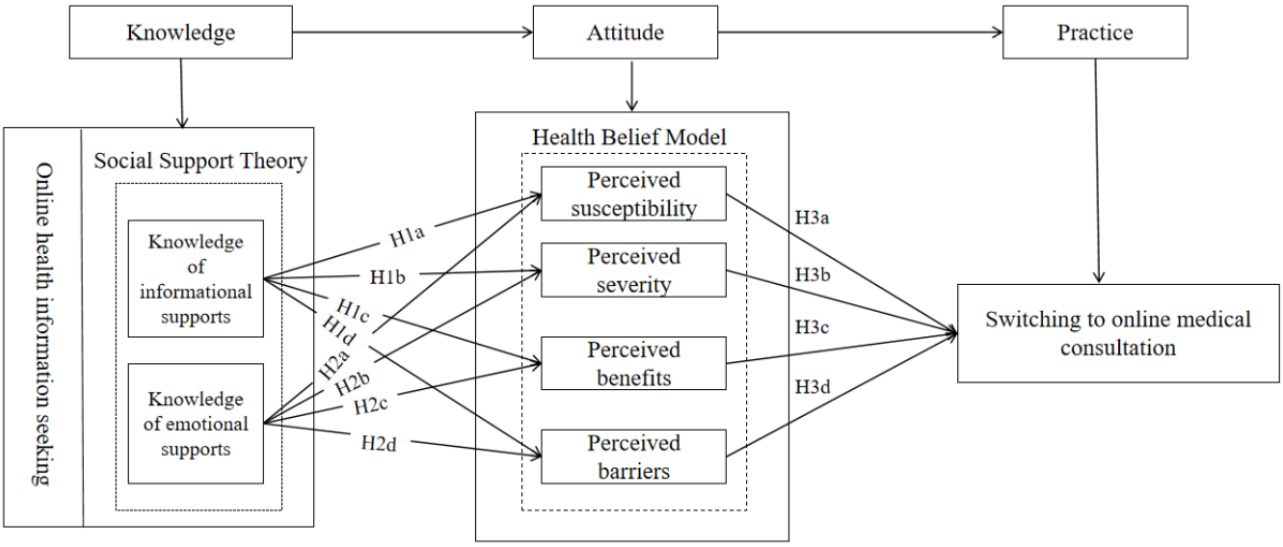

Methods

Study Design

Structural equation modeling (SEM) is a model that analyzes the path relationships among intricate variables, based on factor analysis and linear regression methods. SEM can be used to explore relationships between variables. It incorporates causal relationships among variables and measurement errors into a unified analytical framework, allowing for the simultaneous evaluation of relationships among multiple variables. These variables can be observed data (such as indicators from survey questionnaires) or latent variables (such as cognition, emotion, beliefs, and other factors that are difficult to observe directly). SEM operates at the scale level, meaning it reduces the dimensionality of scales into 1 principal component through factor analysis (transforming multiple variables into a single variable) before conducting path analysis.

Measurement Instrument

Most of the construct items in this study were adapted from validated existing scales. Each item was measured using a 5-point Likert scale, ranging from 1="completely disagree" to 5="completely agree." To ensure the scientific and content validity of the questionnaire, the measurement items were based on established scales from domestic and international studies, and some were revised to suit the online health care application scenario.

The specific measurement items and theoretical sources are shown in Table 1. Among them, seeking informational support and emotional support refer to the scales of Zhu et al [39] and Liang et al [40]. Perceived susceptibility and perceived severity refer to the scales of Noroozi et al [41]. Perceived benefits refer to the scales of Gong et al [42]. Perceived barriers refer to the scales of Huang et al [43] and Yuan and Tang [44]. OMC behavior refers to the scales of Fornell and Larcker [45].

To ensure the scientific accuracy of the questionnaire, we conducted a pilot survey. Fifty individuals who have sought online health information were invited to participate in the pilot survey. We gathered their feedback and suggestions during the completion of the initial questionnaire to further refine the instrument, resulting in the final version of the questionnaire.

XSL•FO
RenderX

Table 1. Constructs and measures.

Constructs and measures	References
OHIS[a] behavior	[39,40]
KIS[b]	
KIS1: While browsing the online health community, I discovered that the platform's doctors assist patients in identifying the causes of their diseases.	
KIS2: While browsing the online health community, I discovered that the platform's doctors provided effective responses and suggestions to patients.	
KIS3: While browsing the online health community, I discovered that the platform's doctors would provide problem-solving information to patients.	
KIS4: While browsing the online health community, I found that the doctors on the platform would provide specific solutions to patients.	
KES[c]	
KES1: While browsing the online health community, I felt the doctor's concern for the patient.	
KES2: While browsing the online health community, I feel that doctors comfort and encourage their patients.	
KES3: While browsing the online health community, I felt that the doctor would confront health problems alongside the patient.	
KES4: While browsing the online health community, I felt that the doctor was attentive to the patient's feelings.	
Health belief model	[41-44]
PSU[d]	
PSU1: In everyday life, people frequently face health issues.	
PSU2: When I maintain an unhealthy lifestyle for an extended period, I feel my chances of getting sick increase.	
PSU3: The chances of developing a disease increase when I am unwell.	
PSU4: I may become ill if I don't seek medical advice promptly for concerning symptoms in my body.	
PSE[e]	
PSE1: I feel more anxious and afraid when I consider the possibility of developing a severe disease.	
PSE2: My daily life would be impacted if I developed a serious illness.	
PSE3: If I don't seek medical advice quickly, I'm worried that the disease may get worse.	
PBE[f]	
PBE1: Online consultations are beneficial to my health.	
PBE2: Online consultations are beneficial for my health management.	
PBE3: Online consultations will help me detect illnesses in advance.	
PBE4: Online consultation will help alleviate my worries about my condition.	
PBA[g]	
PBA1: I am concerned that the private health information submitted to online doctors will be misused.	
PBA2: I worry that the quality of the online consultation service is not guaranteed.	
PBA3: I am concerned about excessive use of online consultation services.	
PBA4: I am concerned about my ability to accurately communicate my consultation needs to community doctors.	
From OHIS to OMC	[43]
OMC[h]	
OMC1: I will consider consulting an online medical community physician if I experience health issues in the future.	
OMC2: In the future, when I encounter health problems, I will consult with online medical community doctors as much as possible.	
OMC3: In the future, when I face health issues, I plan to consult a doctor from an online medical community.	

XSL•FO
RenderX

[a]OHIS: online health information seeking.
[b]KIS: knowledge of informational support.
[c]KES: knowledge of emotional support.
[d]PSU: perceived susceptibility.
[e]PSE: perceived severity.
[f]PBE: perceived benefits.
[g]PBA: perceived barriers.
[h]OMC: online medical consultation.

Ethics Considerations

This study was approved by the ethics committee of Xinxiang Medical University (20250478). Prior to the survey, each participant completed an informed consent form and agreed to the analysis of the obtained data by the research team. The data collection and storage processes in this study strictly adhered to confidentiality principles, ensuring the anonymization of personal information. This study did not involve the use of images or photographs of the participants. Participants were eligible to receive compensation ranging from US $1 to US $2 for their involvement in the study.

Data Collection

Our survey targets the population that has used the internet to search for health information. The target population of this study includes Chinese adults with experience or familiarity with digital health technologies. The eligibility criteria require participants to be aged at least 19 years and confirm at least one of the following: (1) previous use of health-related technologies (such as health apps or smart wearables), (2) functional understanding of the concept of digital health, or (3) indication of willingness to use technology for personal or health management. These criteria ensure that respondents possess the necessary background knowledge to provide meaningful and effective responses to the survey items. We created the questionnaire using the software "Questionnaire Star" [Changsha Ranxing Information Technology Co, Ltd] and distributed it through 2 primary channels. After the preresearch was approved, we officially distributed our questionnaires. The recruited participants would receive a random cash red envelope, with amounts ranging from US $1 to US $2. First, we shared the questionnaire within Chinese communities, such as "QQ groups" and "Douban groups" for users of OHCs. We also recruited participants through OHCs such as "Baidu Tieba" and "Haodaifu Online." A total of 509 questionnaires were collected from March to May 2024. After excluding incomplete and invalid responses according to the eligibility criteria, we obtained 425 valid questionnaires.

Statistical Analysis

In this study, we used SEM to analyze the collected data and evaluate the research model. Following a 2-step procedure, we first used SPSS (version 26.0; IBM Corp) software to examine the measurement model, ensuring its reliability and validity. Tables 2 and 3 report the reliability and validity results. As Table 2 shows, the Cronbach and composite reliability values were greater than the proposed threshold of 0.7 [45], indicating qualified reliability. Convergent validity was assessed by the item loadings and the average variance extracted (AVE) from expected constructs. The results show that AVEs were higher than the recommended value of 0.5 [46], and all indicator loadings exceeded the threshold of 0.7, suggesting satisfactory convergence validity. Table 3 suggested that the square root of AVE values for each construct exceeded all its correlation coefficients with other constructs, indicating promising discriminant validity [47].

We then used AMOS (version 26.0; IBM Corp) software to examine the fitness between the sample data model and the research model. The results are shown in Table 4, indicating that the model's fitness indicators, including the absolute fitness indicator, value-added fitness indicator, and simple fitness indicator, all achieve a reasonable level, demonstrating that the model fits well.

XSL•FO
RenderX

Table 2. Validated factor analysis results.

Constructs	Cronbach α	Convergent validity AVE[a]	Combined reliability
KIS[b]	0.802	0.508	0.804
KES[c]	0.814	0.539	0.822
PSU[d]	0.832	0.568	0.839
PSE[e]	0.809	0.598	0.816
PBE[f]	0.821	0.557	0.832
PBA[g]	0.825	0.556	0.831
OMC[h]	0.813	0.617	0.825

[a]AVE: average variance extracted.
[b]KIS: knowledge of informational support.
[c]KES: knowledge of emotional support.
[d]PSU: perceived susceptibility.
[e]PSE: perceived severity.
[f]PBE: perceived benefits.
[g]PBA: perceived barriers
[h]OMC: online medical consultation.

Table 3. Results of distinction validity test[a].

Constructs	OMC[b]	PBA[c]	PBE[d]	PSE[e]	PSU[f]	KES[g]	KIS[h]
OMC	0.785	N/A[i]	N/A	N/A	N/A	N/A	N/A
PBA	0.206	0.745	N/A	N/A	N/A	N/A	N/A
PBE	0.357	0.132	0.746	N/A	N/A	N/A	N/A
PSE	0.291	0.161	0.265	0.773	N/A	N/A	N/A
PSU	0.272	0.315	0.257	0.342	0.754	N/A	N/A
KES	0.181	0.209	0.288	0.123	0.342	0.734	N/A
KIS	0.316	0.307	0.403	0.311	0.387	0.508	0.713

[a]Values on the diagonal are the square root of average variance extracted, and the lower triangles are the correlation coefficients between the factors.
[b]OMC: online medical consultation.
[c]PBA: perceived barriers.
[d]PBE: perceived benefits.
[e]PSE: perceived severity.
[f]PSU: perceived susceptibility.
[g]KES: knowledge of emotional support.
[h]KIS: knowledge of informational support.
[i]N/A: not applicable.

XSL•FO
RenderX

Table 4. Results of the model fitness test.

Fitness indicators	Judgment criteria	Indicator value	Adaptation effect
Absolute fitness indicator			
CMIN/DF[a]	<3	1.206	Ideal
SEA[b]	<0.08	0.022	Ideal
GFI[c]	>0.9	0.943	Ideal
AGFI[d]	>0.8	0.928	Ideal
Value-added fitness indicator			
CFI[e]	>0.9	0.987	Ideal
IFI[f]	>0.9	0.987	Ideal
TLI[g]	>0.9	0.985	Ideal
Simple fitness indicator			
PGFI[h]	>0.5	0.747	Ideal
PNFI[i]	>0.5	0.794	Ideal

[a]CMIN/DF: chi-square/degrees of freedom.
[b]SEA: standardized error of approximation.
[c]GFI: Goodness-of-Fit Index.
[d]AGFI: Adjusted Goodness-of-Fit Index.
[e]CFI: Comparative Fit Index.
[f]IFI: Incremental Fit Index.
[g]TLI: Tucker-Lewis Index.
[h]PGFI: Parsimony Goodness-of-Fit Index.
[i]PNFI: Parsimony Normed Fit Index.

Results

Characteristics of Respondents

The characteristics of the respondents are illustrated in Table 5. The participants' gender was relatively balanced, with 54.40% (231/425) female and 45.6% (194/425) male. The age range covered individuals from younger than 20 years to older than 50 years. Approximately half (227/425, 53.4%) of the participants held college degrees. In terms of careers, 32.2% (137/425) of the participants were students, while 66.4% (282/425) were used. The respondents' places of residence varied significantly, with 72.7% (309/425) of participants living in urban areas and 27.3% (116/425) residing in rural areas.

XSL•FO
RenderX

Table 5. Characteristics of respondents.

Variable	Participants (N=425)
Sex, n (%)	
Male	194 (45.60)
Female	231 (54.40)
Age (years), n (%)	
≤19	20 (4.70)
20-29	256 (60.20)
30-39	47 (11.10)
40-49	91 (21.40)
50	11 (2.60)
Education level, n (%)	
High school and below	57 (13.40)
Associate degree	98 (23.10)
Bachelor's degree	227 (53.40)
Master's degree and above	43 (10.10)
Employment status, n (%)	
Student	137 (32.20)
Working/freelance	282 (66.40)
Other	6 (1.40)
Place of residence, n (%)	
Rural	116 (27.30)
Urban	309 (72.70)

Structural Model

Path Coefficients

We adopted the AMOS (version 26.0) software to examine the structural model's path coefficients and corresponding significance levels. Table 6 shows the results of the hypothesized path analysis. Among these, seeking knowledge of information support has a significant positive effect on perceived susceptibility (β=.339, *P*<.001), perceived severity (β=.348, *P*<.001), and perceived benefits (β=.361, *P*<.001), while also having a significant negative effect on perceived barriers (β=–.285, *P*<.001). Thus, hypotheses H1a, H1b, H1c, and H1d are valid. Seeking knowledge of emotional support has a significant positive effect on perceived susceptibility (β=.220, *P*<.001) and perceived benefits (β=.149, *P*<.01), confirming hypotheses H2a and H2c. However, it showed no significant effect on perceived severity (β=–.006, *P*>.05) or perceived barriers (β=.099, *P*>.05), meaning hypotheses H2b and H2d were not established. Perceived susceptibility (β=.123, *P*<.05), perceived severity (β=.174, *P*<.001), and perceived benefits (β=.273, *P*<.001) positively affect patients' switching to online consultation behavior, while perceived barriers (β=–.112, *P*<.05) negatively affect switching behavior, validating hypotheses H3a, H3b, H3c, and H3d.

XSL•FO
RenderX

Table 6. Results of hypothesized path analysis.

Hypothesis	Path correlation	Path factor	Results
H1a	KIS[a] → PSU[b]	0.339[c]	Supported
H1b	KIS → PSE[d]	0.348[c]	Supported
H1c	KIS → PBEe	0.361[c]	Supported
H1d	KIS → PBA[f]	–0.285[c]	Supported
H2a	KES[g] → PSU	0.220[c]	Supported
H2b	KES → PSE	–0.006	Not supported
H2c	KES → PBE	0.149[h]	Supported
H2d	KES → PBA	0.099	Not supported
H3a	PSU → OMC[i]	0.123[j]	Supported
H3b	PSE → OMC	0.174[c]	Supported
H3c	PBE → OMC	0.273[c]	Supported
H3d	PBA → OMC	–0.112[i]	Supported

[a]KIS: knowledge of informational support.
[b]PSU: perceived susceptibility.
[c]$P<.001$ (very high level of significance).
[d]PSE: perceived severity.
[e]PBE: perceived benefits.
[f]PBA: perceived barriers.
[g]KES: knowledge of emotional support.
[h]$P<.01$ (high level of significance).
[i]OMC: online medical consultation.
[j]$P<.05$ (a level of significance).

Mediating Effect Test

This study uses model 4 in the Process macro (SPSS [version 26; IBM Corp]) and performs a mediating effect test using 5000 bootstrap resamples to examine the mediating role of health beliefs. As shown in Table 7, both the total effect (β=.339) and the direct effect (β=.135) are significant, as the confidence interval obtained from the bootstrap analysis for the direct effect does not include zero [48]. Among the indirect effects, the confidence intervals of 3 out of 4 effects do not include zero, while the confidence interval of perceived barriers does encompass zero. This indicates that information-supported knowledge acquisition can not only directly affect patients' behavior of switching to online consultation but also influence patients' OMC behavior through perceived susceptibility (14.23%), perceived severity (13.17%), and perceived benefits (25.28%), as these 3 play partial mediating roles in the model path.

Since the path analysis indicates that the acquisition of emotional support knowledge has no significant impact on perceived severity and perceived barriers (H2b and H2d), when conducting the mediating effect test of the HBM related to the acquisition of emotional support knowledge and behavioral transfer, only the acquisition of emotional support knowledge, perceived susceptibility, and perceived benefit are simultaneously included in the regression analysis. The test results are shown in Table 8. The upper and lower limits of the 95% confidence intervals from the Bootstrap for both the total and indirect effects of emotional support knowledge acquisition on patient behavior transfer do not include 0, whereas the upper and lower limits of the 95% confidence intervals from the Bootstrap for the direct effect do contain. This indicates that the acquisition of emotional support knowledge cannot directly influence patient behavior transfer; it can operate only through 2 mediating variables: perceived susceptibility and perceived benefit. Furthermore, perceived susceptibility and perceived benefit entirely mediate the hypothetical path, accounting for 46.95% and 52.90% of the total effect, respectively.

XSL•FO
RenderX

Table 7. Mediating effect of health belief on the relationship between information supports knowledge seeking and switching behavior.

	Effect values	BootLLCI	BootULCI	Relative mediated effect
Total effect	0.339	0.215	0.464	N/A[a]
Direct effect	0.135	0.004	0.266	N/A
Indirect effect				
Perceived susceptibility	0.048	0.008	0.098	14.23%
Perceived severity	0.045	0.003	0.094	13.17%
Perceived benefits	0.086	0.040	0.144	25.28%
Perceived barriers	0.026	–0.010	0.067	N/A

[a]N/A: not applicable.

Table 8. Mediating effect of the health belief between Information supports knowledge seeking and switching behavior.

	Effect values	BootLLCI	BootULCI	Relative mediated effect
Total effect	0.138	0.036	0.239	N/A[a]
Direct effect	0.000	–0.103	0.103	N/A
Indirect effect				
Perceived susceptibility	0.065	0.027	0.114	46.95%
Perceived benefits	0.073	0.036	0.120	52.90%

[a]N/A: not applicable.

Discussion

Principal Findings

Drawing on the KAP theoretical framework, this study combines social support theory and the HBM to examine the mechanism underlying patients' transition from seeking online health information to engaging in OMC within an online medical community. The research found that the OHIS has a positive influence on both informational and emotional support knowledge, thereby promoting OMC. In this process, perceived susceptibility, perceived severity, and perceived benefit within the health beliefs model have a partial mediating effect, while the mediating role of perceived barriers is not significant. This finding supports the underlying research hypothesis proposed in this study, suggesting that patients' online access to positive health information can motivate them to move from OHIS to OMC through shifts in health beliefs.

First, this finding suggests that patients are inclined to acquire positive information and develop health-related knowledge when conducting OHIS. This aligns with a study by Sassenberg and Greving [49]. In their research, a survey of chronic patients indicated that individuals were more likely to obtain positive and supportive information and form favorable views about their health when using the internet for information seeking compared with other information sources.

Second, our study found that patients can positively influence perceived susceptibility and perceived benefit in health beliefs by accessing informational support and emotional knowledge. Both factors play a mediating role in facilitating patients' transition to OMC behavior. On one hand, this suggests that patients can obtain supportive health information through online searches, which helps them observe and evaluate their own health status, thereby increasing their perception of illness [50]; on the other hand, accessing emotionally supportive information heightens their concern for their health status and significantly enhances their perceived susceptibility to disease. Perceived susceptibility, a core variable in health behavioral decision-making, directly affects individuals' demand for specialized medical services. This explains why patients are prompted to seek further specialized online consultations, aligning with the study Li et al [51] on "Perceived susceptibility enhances the use of internet hospitals." Meanwhile, regarding perceived benefits, accessing supportive health information enables patients to better understand, evaluate, and apply professional medical insights. In addition, emotional support strengthens the close relationships among patients and reinforces their recognition of online doctors' professionalism and trust in service quality, all of which enhance patients' perceived benefits. Increased perceived gains foster doctor-patient trust [52], providing a driving force for patients to adopt OMC behaviors. Furthermore, the results did not support H2b and H2d, indicating that the acquisition of emotional support knowledge had no significant effect on perceived severity or perceived barriers. This may be because emotional support primarily offers psychological comfort and empathy rather than concrete cognitive information about disease severity. The emotional exchange provides patients with more emotional relief and psychological support rather than an objective assessment of potential health consequences. On the other hand, perceived barriers are often related to practical factors such as technical difficulties, privacy concerns, and cost issues. Although emotional support can alleviate anxiety, it is less likely to directly alter patients' perceptions of these practical obstacles. Therefore, while emotional support can enhance perceived

XSL•FO
RenderX

susceptibility and perceived benefits, its impact on perceived severity and perceived barriers remains limited.

Meanwhile, this study revealed the differential mediating effect of perceived severity in the HBM. Specifically, patients' access to systematic, information-supportive knowledge of causes, symptoms, and treatment options [53] can enhance their perception and judgment of disease severity. Perceived severity mediates the relationship between patients' access to information-supportive knowledge and their shift to OMC behavior, which is theoretically consistent with the mechanism of facilitating online counseling use in internet hospitals, as noted in a previous study [51]. However, the results of this study also suggest that while knowledge of emotional support can generate positive health beliefs, it does not influence perceived severity, nor change perceived barriers. This contrasts with some literature that suggests that "emotional support may weaken perceived health risks" [54].

Finally, our study also found that while access to information supporting knowledge made it easier for patients to understand physicians' responses and recommendations, reducing perceived barriers, these same barriers prevented patients from shifting to OMC. However, perceived barriers did not mediate the effect, possibly because access to information supporting knowledge had a more substantial influence on perceived susceptibility to severity and benefits, leading patients to prefer those options that could provide immediate diagnosis and treatment through convenient and efficient online consultation methods, counteracting the hindering effect of perceived barriers.

Implications and Limitations

Based on our limited literature review, this study, for the first time, draws on the KAP theoretical framework, integrates the social support theory and the HBM, and adopts an empirical research methodology to explore the mechanisms associated with the transfer of patients' behaviors from online health information acquisition to online consultation. This study also provides practical insights for online health care platforms. From the perspective of information support, online medical communities should encourage medical professionals to engage more in communication and interaction and regularly release authoritative and accurate health information and disease knowledge to enhance patients' understanding of diseases and promote their in-depth utilization of online medical platform functions. From the perspective of emotional support, online medical communities should encourage patients to share their treatment experiences and psychological feelings with one another and also strengthen the emotional bond between doctors and patients. Moreover, they should develop incentive mechanisms to motivate patients to take the initiative in communicating with their doctors, such as launching point reward activities and tutorials for novice medical practitioners, to mitigate the wait-and-see approach of patients in OMCs [55]. Based on the finding that "informational support significantly influences perceived severity, while emotional support does not," we recommend that platforms emphasize different types of content at various stages of facilitating consultation conversion: during the user's disease awareness phase, push authoritative and systematic disease knowledge (informational support); during the user's decision-hesitation phase, showcase warm doctor-patient interaction stories (emotional support) to enhance trust and perceived benefits. In addition, the online medical community should enhance the performance reward and punishment mechanism for doctors, encourage medical professionals to proactively answer patients' inquiries, and demonstrate a profound commitment to humanistic care. This not only establishes trust between doctors and patients but also significantly improves patients' continued adoption and deeper application of online medical services.

This study also has certain limitations, primarily reflected in 3 aspects that can serve as a reference for future research. First, the research data in this paper come from questionnaire surveys, and the subjectivity of users in completing the questionnaires may introduce bias into the research results. The sample primarily consists of young, middle-aged, and urban users, which may overlook the behavioral differences of older or rural groups due to the digital divide. In addition, this study predominantly sampled individuals with at least an undergraduate education; expanding the scope of survey participants in future studies would be beneficial. Finally, this study focused on the transition from acquiring health information to engaging in consultation behavior. Future research could further explore the mechanisms influencing online consultation behavior from alternative perspectives or through mixed methods. Furthermore, the use of cross-sectional data makes it challenging to capture the dynamic process of behavioral change, and the mechanisms could be better understood by incorporating longitudinal designs or mixed methods in future research.

Conclusions

This study integrated the knowledge-attitude-behavior framework, social support theory, and HBM to uncover the internal logic of patients' behavioral transfer within OHCs. It confirmed the mediating role of the cognitive-emotional dual-drive pathway and health beliefs. The findings provide a scientific basis for the functional design of online health care platforms and the precise formulation of health knowledge dissemination strategies. In addition, they highlight the need for future attention to the differences in digital literacy among various groups to promote the deeper development and widespread application of digital health services.

Acknowledgments

The authors would like to thank all the participants in this study.

XSL•FO
RenderX

Funding

This work was jointly supported by grants from the National Natural Science Foundation of China (72061160491 and 72271027) and Beijing Institute of Technology Research Fund Program for Young Scholars (XSQD-202216004).

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

LZ contributed to conceptualization, methodology, writing—original draft, and project administration. JX contributed to data curation, writing—original draft, and visualization. HC participated in investigation, data curation, and writing—original draft. YB and JW conducted the supervision. LW contributed to formal analysis, validation, and software. WR contributed to resources and funding acquisition. LZ, JX, and HC contributed equally to this work as co-first authors.

Conflicts of Interest

None declared.

References

1. Guo X, Guo S, Vogel D, Li Y. Online healthcare community interaction dynamics. J Manag Sci Eng 2016;1(1):58-74. [doi: 10.3724/SP.J.1383.101004]
2. Wang WT, Zhang XP, Luo QF. Quantitative empirical study on the correlation between "digital natives" online health information searching and offline medical treatment behavior. Inf Theory Pract 2021;44(7):86-93. [doi: 10.16353/j.cnki.1000-7490.2021.07.013]
3. Luo XL, Han JT, Fan WG. Health information and health anxiety in the internet era. Inf Stud 2020;40(2):76-86. [doi: 10.31820/pt.29.2.10]
4. Hwang EH, Guo X, Tan Y, Dang Y. Delivering healthcare through teleconsultations: implications for offline healthcare disparity. Inf Syst Res 2022;33(2):515-539. [doi: 10.1287/isre.2021.1055]
5. Luo A, Qin L, Yuan Y, Yang Z, Liu F, Huang P, et al. The effect of online health information seeking on physician-patient relationships: systematic review. J Med Internet Res 2022;24(2):e23354 [FREE Full text] [doi: 10.2196/23354] [Medline: 35142620]
6. Pan X, Zhou X, Yu L, Hou L. Switching from offline to online health consultation in the post-pandemic era: the role of perceived pandemic risk. Front Public Health 2023;11:1121290 [FREE Full text] [doi: 10.3389/fpubh.2023.1121290] [Medline: 37261233]
7. Di Novi C, Kovacic M, Orso CE. Online health information seeking behavior, healthcare access, and health status during exceptional times. J Econ Behav Organ 2024;220:675-690 [FREE Full text] [doi: 10.1016/j.jebo.2024.02.032] [Medline: 38628501]
8. Deng CH, Deng ZH, Fan GR. Can online health platforms reduce health inequity from the perspective of value co-creation. Manage World 2024;40(1):119-140. [doi: 10.19744/j.cnki.11-1235/f.2024.0007]
9. Zhao Y, Liu Z, Song S. Libr Inf 2019;39(2):16-28 [FREE Full text]
10. Panahi S, Watson J, Partridge H. Information encountering on social media and tacit knowledge sharing. J Inf Sci 2015;42(4):539-550. [doi: 10.1177/0165551515598883]
11. Zimmerman MS, Shaw G. Health information seeking behaviour: a concept analysis. Health Info Libr J 2020;37(3):173-191 [FREE Full text] [doi: 10.1111/hir.12287] [Medline: 32052549]
12. Chen D, Zhang R, Liu K, Hou L. Enhancing online patient support through health-care knowledge in online health communities: a descriptive study. Information 2018;9(8):199. [doi: 10.3390/info9080199]
13. Cao BL, Dai WDD. Understanding online doctor-patient communication: perspective on online medical consultation as a communication behavior from the doctor-patient-technology tripartite view. J Univ 2022(11):54-66. [doi: 10.20050/j.cnki.xwdx.2022.11.010]
14. Bloom G, Standing H, Lloyd R. Markets, information asymmetry and health care: towards new social contracts. Soc Sci Med 2008;66(10):2076-2087. [doi: 10.1016/j.socscimed.2008.01.034] [Medline: 18316147]
15. Tang M, Li XG, Wang XY. Research on patients' consultation in online health communities driven by data from a multidimensional trust perspective. Inf Sci 2023;41(5):2-9. [doi: 10.13833/j.issn.1007-7634.2023.05.001]
16. de Looper M, van Weert JCM, Schouten BC, Bolle S, Belgers EHJ, Eddes EH, et al. The influence of online health information seeking before a consultation on anxiety, satisfaction, and information recall, mediated by patient participation: field study. J Med Internet Res 2021;23(7):e23670 [FREE Full text] [doi: 10.2196/23670] [Medline: 34255657]
17. Xu Y, Yang Z, Jiang H, Sun P. Research on patients' willingness to conduct online health consultation from the perspective of web trust model. Front Public Health 2022;10:963522 [FREE Full text] [doi: 10.3389/fpubh.2022.963522] [Medline: 36159287]

XSL•FO
RenderX

18. Li H, Li D, Zhai M, Lin L, Cao Z. Associations among online health information seeking behavior, online health information perception, and health service utilization: cross-sectional study. J Med Internet Res 2025;27:e66683 [FREE Full text] [doi: 10.2196/66683] [Medline: 40085841]
19. Zhou T, Wu X. Humanit Soc Sci Commun 2024;11(1):1083. [doi: 10.1057/s41599-024-03540-1]
20. Zhao YX, Liu ZY, Zhu QH. J China Soc Sci Tech Inf 2020;39(5):534-546. [doi: 10.3772/j.issn.1000-0135.2020.05.006]
21. Tang ZF, Dong XL, Chen SY. Analysis of influencing factors on guardians' willingness and behavior to vaccinate children with non-immunization program vaccines' based on the health belief theory. Chin Rural Health Serv Manag 2023;43(12):907-912. [doi: 10.19955/j.cnki.1005-5916.2023.12.012]
22. Norton J, Raciti MM. Co-creating healthful eating behaviors with very young children: the impact of information overload on primary caregivers. Health Mark Q 2017;34(1):18-34. [doi: 10.1080/07359683.2016.1238664] [Medline: 27782781]
23. Zhang M. Review of the development of the transtheoretical model and related information behavior research. Libr Inf Serv 2022;66(20):141-147. [doi: 10.13266/j.issn.0252-3116.2022.20.015]
24. Wu J, Li SS. Research on patients' willingness to use information services in online health communities. Inf Sci 2017;35(4):119-125. [doi: 10.13833/j.cnki.is.2017.04.022]
25. Zhang X, Han X, Dang Y, Meng F, Guo X, Lin J. User acceptance of mobile health services from users' perspectives: the role of self-efficacy and response-efficacy in technology acceptance. Inform Health Soc Care 2017;42(2):194-206. [doi: 10.1080/17538157.2016.1200053] [Medline: 27564428]
26. Chung D, Wang J, Meng Y. The impact of short-form video and optimistic bias on engagement in oral health prevention: integrating a KAP model. Behav Sci (Basel) 2024;14(10):968 [FREE Full text] [doi: 10.3390/bs14100968] [Medline: 39457840]
27. Zhang N, Li YJ. Empirical study on mobile social media burnout behavior from the perspective of online social support. Inf Exploration 2022;6:71-79. [doi: 10.3969/j.issn.1005-8095.2022.06.009]
28. Wei H, Gao JS, Dai F. The impact of social support in virtual health communities on users' willingness to share knowledge: a moderated chain mediation model. Inf Sci 2021;39(12):146. [doi: 10.13833/j.issn.1007-7634.2021.12.022]
29. Braithwaite DO, Waldron VR, Finn J. Communication of social support in computer-mediated groups for people with disabilities. Health Commun 1999;11(2):123-151 [FREE Full text] [doi: 10.1207/s15327027hc1102_2] [Medline: 16370973]
30. Zhang Q, Zhang R, Lu X, Zhang X. What drives the adoption of online health communities? An empirical study from patient-centric perspective. BMC Health Serv Res 2023;23(1):524 [FREE Full text] [doi: 10.1186/s12913-023-09469-6] [Medline: 37221504]
31. Zhao YC, Zhao M, Song S. Online health information seeking among patients with chronic conditions: integrating the health belief model and social support theory. J Med Internet Res 2022;24(11):e42447 [FREE Full text] [doi: 10.2196/42447] [Medline: 36322124]
32. Peng Y, Yin P, Deng Z, Wang R. Patient-physician interaction and trust in online health community: the role of perceived usefulness of health information and services. Int J Environ Res Public Health 2019;17(1):139 [FREE Full text] [doi: 10.3390/ijerph17010139] [Medline: 31878145]
33. Ma X, Liu Y, Zhang P, Qi R, Meng F. Understanding online health information seeking behavior of older adults: a social cognitive perspective. Front Public Health 2023;11:1147789 [FREE Full text] [doi: 10.3389/fpubh.2023.1147789] [Medline: 36935731]
34. Islam AKMN, Mäntymäki M, Laato S, Turel O. Adverse consequences of emotional support seeking through social network sites in coping with stress from a global pandemic. Int J Inf Manage 2022;62:102431 [FREE Full text] [doi: 10.1016/j.ijinfomgt.2021.102431] [Medline: 34642531]
35. Zhao H. The impact of social support on individual health knowledge creation in online patient communities: the mediating role of attitude. J Knowledge Manage 2024;29(3):937-967. [doi: 10.1108/jkm-02-2024-0135]
36. Jiang L, Liu S, Li H, Xie L, Jiang Y. The role of health beliefs in affecting patients' chronic diabetic complication screening: a path analysis based on the health belief model. J Clin Nurs 2021;30(19-20):2948-2959 [FREE Full text] [doi: 10.1111/jocn.15802] [Medline: 33951248]
37. Jin WY, Zhang Y, Feng CY. Qualitative study on influencing factors of dietary behavior compliance in hemodialysis patients. Chin J Gen Pract 2024;22(3):508-512.
38. Zhang K, Liang J, Hou J, Han B, Zhao Y, Jin T, et al. Effects of health information communication based on social media usage in promoting HPV vaccination intention in China. Health Soc Care Community 2024:9311483 [FREE Full text] [doi: 10.1155/hsc/9311483]
39. Zhu DH, Sun H, Chang YP. Effect of social support on customer satisfaction and citizenship behavior in online brand communities: the moderating role of support source. J Retailing Consum Serv 2016;31:287-293. [doi: 10.1016/j.jretconser.2016.04.013]
40. Liang T, Ho Y, Li Y, Turban E. What drives social commerce: the role of social support and relationship quality. Int J Electron Commerce 2014;16(2):69-90. [doi: 10.2753/jec1086-4415160204]
41. Noroozi A, Jomand T, Tahmasebi R. Determinants of breast self-examination performance among Iranian women: an application of the health belief model. J Cancer Educ 2011;26(2):365-374. [doi: 10.1007/s13187-010-0158-y] [Medline: 20859775]

XSL•FO
RenderX

42. Gong Z, Han Z, Li X, Yu C, Reinhardt JD. Factors influencing the adoption of online health consultation services: the role of subjective norm, trust, perceived benefit, and offline habit. Front Public Health 2019;7:286 [FREE Full text] [doi: 10.3389/fpubh.2019.00286] [Medline: 31637229]
43. Huang H, Zhu J, Chen J, Qin Y, Chen S. Exploring the factors influencing the adoption of online medical services by older adults: a modified UTAUT2 based study. Front Public Health 2025;13:1559701 [FREE Full text] [doi: 10.3389/fpubh.2025.1559701] [Medline: 40247864]
44. Yuan H, Tang N. Research on health information search motivation and perceived barriers of digital immigrants. Inf and Doc Serv 2015:67-72.
45. Fornell C, Larcker DF. Structural equation models with unobservable variables and measurement error: algebra and statistics. J Mark Res 1981;18(3):382. [doi: 10.2307/3150980]
46. Chin W. The partial least squares approach to structural equation modeling. In: Modern Methods for Business Research. New York, NY: Psychology Press; 1998:295-336.
47. Fornell C, Larcker DF. Evaluating structural equation models with unobservable variables and measurement error. J Mark Res 1981;18(1):39-50. [doi: 10.1177/002224378101800104]
48. Chen L, Zhang D, Hou M. The influence of perceived social presence on the willingness to communicate in mobile medical consultations: experimental study. J Med Internet Res 2022;24(5):e31797 [FREE Full text] [doi: 10.2196/31797] [Medline: 35544293]
49. Sassenberg K, Greving H. Internet searching about disease elicits a positive perception of own health when severity of illness is high: a longitudinal questionnaire study. J Med Internet Res 2016;18(3):e56 [FREE Full text] [doi: 10.2196/jmir.5140] [Medline: 26944335]
50. Wang L, Meng YR. A study on the causes of internet health anxiety among migrant workers from the perspective of online information search behavior. J Mod Inf 2024:1-20. [doi: 10.3969/j.issn.1008-0821.2024.12.007]
51. Li D, Hu Y, Pfaff H, Wang L, Deng L, Lu C, et al. Determinants of patients' intention to use the online inquiry services provided by internet hospitals: empirical evidence from China. J Med Internet Res 2020;22(10):e22716 [FREE Full text] [doi: 10.2196/22716] [Medline: 33006941]
52. Segado-Fernández S, Lozano-Estevan MDC, Jiménez-Gómez B, Ruiz-Núñez C, Jiménez Hidalgo PJ, Fernández-Quijano I, et al. Health literacy and critical lecture as key elements to detect and reply to nutrition misinformation on social media: analysis between spanish healthcare professionals. Int J Environ Res Public Health 2022;20(1):23 [FREE Full text] [doi: 10.3390/ijerph20010023] [Medline: 36612345]
53. MacArthur KR. Beyond health beliefs: the role of trust in the HPV vaccine decision-making process among American college students. Health Sociol Rev 2017;26(3):321-338. [doi: 10.1080/14461242.2017.1381035]
54. Rhubart D, Kowalkowski J. Perceived mental health impacts of the COVID-19 pandemic: the roles of social support and social engagement for working age adults in the United States. Prev Med 2022;162:107171 [FREE Full text] [doi: 10.1016/j.ypmed.2022.107171] [Medline: 35878709]
55. Chen D, Zhang R, Feng J, Liu K. Fulfilling information needs of patients in online health communities. Health Info Libr J 2020;37(1):48-59 [FREE Full text] [doi: 10.1111/hir.12253] [Medline: 31090185]

Abbreviations

AVE: average variance extracted
HBM: health belief model
KAP: knowledge-attitude-practice
OHC: online health community
OHIS: online health information seeking
OMC: online medical consultation
SEM: structural equation modeling

Edited by A Stone; submitted 02.06.25; peer-reviewed by BS Paul, G Zhu; comments to author 01.10.25; revised version received 29.10.25; accepted 29.10.25; published 03.12.25.

Please cite as:
Zhang L, Xia J, Chen H, Bai Y, Wang J, Wang L, Ren W
The Mechanism of Online Health Information Seeking Switching to Online Medical Consultation: Cross-Sectional Study
JMIR Form Res 2025;9:e78397
URL: https://formative.jmir.org/2025/1/e78397
doi:10.2196/78397
PMID:

XSL•FO
RenderX

©Lijiang Zhang, Jingjing Xia, Hui Chen, Yang Bai, Jun Wang, Liuan Wang, Wenjie Ren. Originally published in JMIR Formative Research (https://formative.jmir.org), 03.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Mobile App to Enhance Awareness of Vaccination in Adults With Psoriasis and Atopic Dermatitis: Development and Preliminary Evaluation Study

Giulia Gasparini[1,2], MD, PhD; Donatella Panatto[1], PhD; Irene De Kermarek[1], BA; Lucia Valchi[1], MPH; Elvira Massaro[1], MD; Piero Luigi Lai[1], PhD; Martina Burlando[1,2], MD; Giancarlo Icardi[1], MD; Emanuele Cozzani[1,2], MD, PhD

[1]Department of Health Sciences, University of Genoa, Genoa, Italy
[2]Dermatology Unit, Ospedale Policlinico San Martino, Genoa, Italy

Corresponding Author:
Giulia Gasparini, MD, PhD
Department of Health Sciences
University of Genoa
Via Pastore 1
Genoa, 16132
Italy
Phone: 39 010 555 3764
Fax: 39 010 555 6741
Email: giulia.gasparini@unige.it

Abstract

Background: Patients with psoriasis and atopic dermatitis are at increased risk of several vaccine-preventable diseases. Despite this increased susceptibility to infections, vaccination uptake in adults with psoriasis and atopic dermatitis, especially if treated with biologics and other systemic immunomodulators, is insufficient. As mobile health technologies may support behavior change, a mobile app called DermatoVax was developed to raise awareness of immunization in adult patients with psoriasis and atopic dermatitis.

Objective: This paper aims to describe the processes of development of the DermatoVax app and its initial evaluation in terms of technical verification and physicians' quality rating.

Methods: The app was conceived in a user-centered fashion. Its core component was the vaccine checker, which allows the app to produce a sharable list of recommended vaccines, immunization timings, and eventual precautions from a short set of input data. App prototypes were extensively piloted, and feedback from potential end users was obtained to refine the app content. The readability of the textual narratives was measured using the Italian-specific Gulpease index, which ranges from 0 to 100, where 100 indicates the best readability. The quality of the final version was evaluated by 46 medical doctors (n=29, 63% dermatologists and n=17, 37% public health physicians) using a validated Italian user version of the Mobile App Rating Scale (uMARS).

Results: Iterative steps during the app development process allowed us to increase its user-friendliness and comprehensibility. Proper functioning of the checker was confirmed through the correct and complete generation of recommended vaccine lists for 50 mock patients with psoriasis and atopic dermatitis. An overall Gulpease index of 41.0 was observed for the final textual narratives, suggesting acceptable readability properties for patients with a high school diploma. Of a maximum of 5 points, the average uMARS score was 4.22 (SD 0.49). Ratings provided by dermatologists (mean 4.28, SD 0.48) were similar (P=.33) to those provided by public health physicians (mean 4.12, SD 0.51). However, the mean uMARS scores for the quality dimensions of aesthetics (3.88, SD 0.78) and engagement (3.89, SD 0.68) were lower than those for information (4.64, SD 0.42) and functionality (4.47, SD 0.46), suggesting margins for improvement. The app's perceived impact was notably high, with over 80% of physicians agreeing that its use would significantly improve patient awareness (39/46, 85%) and knowledge (41/46, 89%) of vaccination, leading to increased vaccination uptake (37/46, 80%).

Conclusions: DermatoVax is a promising tool to raise awareness of immunization in adult patients with psoriasis and atopic dermatitis. Further assessment of the app, such as its effectiveness in increasing vaccination uptake, is warranted.

(JMIR Form Res 2025;9:e80113) doi:10.2196/80113

XSL•FO
RenderX

KEYWORDS

vaccination; awareness; psoriasis; atopic dermatitis; mobile app; Mobile App Rating Scale

Introduction

Psoriasis and atopic dermatitis (AD) are 2 common chronic skin diseases that affect approximately 2% to 4% of the adult population worldwide [1,2]. Prevalence of adult AD in the overall population is generally lower among men; it decreases with age and shows some important geographic variation, ranging from 2.1% in Japan to 4.9% in the United States [2]. Similarly, the incidence of psoriasis varies from 78.9 per 100,000 person-years in the United States to 230 per 100,000 person-years in Italy [1]. Apart from a significant impact on physical, mental, and social well-being [3], patients with psoriasis and AD are also at increased risk of infections, including several vaccine-preventable diseases (VPDs). For instance, a large US cohort [4] showed that the incidence of herpes zoster (HZ) in adults with psoriasis (11.35 per 1000 person-years) was significantly higher than in adults without psoriasis (7.67 per 1000 person-years), with an adjusted incidence rate ratio of 1.21. Dermatological patients are also at an increased risk of postherpetic neuralgia even if the underlying disease is in remission [5]. Analogously, another US study reported [6] that patients with AD were much more likely (adjusted odds ratio 1.73, 95% CI 1.54-1.95) to develop pneumonia or contract influenza than patients without AD. The increased susceptibility to infections is driven by a complex interplay among disruption of the skin barrier; immune dysregulation; and environmental factors, including the use of immunosuppressants [7]. Indeed, both psoriasis and AD are T cell–mediated inflammatory diseases that can be treated with specific cytokine antagonists or more broad immunosuppressive drugs [8].

Despite the increased risk of VPDs, the available data suggest an insufficient vaccination uptake in adult patients with psoriasis and AD, especially in those of younger age. In a recent US survey of adult patients with AD and psoriasis or psoriatic arthritis treated with biologics and Janus kinase inhibitors [9], coverage rates for the seasonal influenza, HZ, pneumococcal, and COVID-19 vaccines were only 9.5%, 6.8%, 16.7%, and 63.9%, respectively. Analogously low coverage rates for the principal routine vaccines were noted among French patients with psoriasis [10]. While recognizing the individual and societal benefits of vaccines, patients with psoriasis and eczema believe that several barriers (such as lack of access to information about vaccines; lack of time for vaccine appointments; concerns about side effects, including disease flaring; and misinformation) contribute to low vaccination uptake among them [11]. Notably, according to Noe et al [12], influenza vaccine uptake in patients with psoriasis is associated with age, female gender, and some comorbidities but not with systemic treatment, including biologics. Another multinational study [13] showed that, compared with other treatment regimens, patients with psoriasis treated with biologics were much less likely to receive at least one COVID-19 vaccine dose. The latter findings may imply that a substantial proportion of patients with psoriasis and AD on systemic therapy with biologics or immunomodulators and with no other comorbidities are unaware of vaccines that they are recommended to receive or are hesitant to be immunized.

With the potential to reach large numbers of users, mobile health (mHealth) apps offer unique opportunities to deliver self-care interventions and support behavior change in a cost-effective way [14] and are increasingly common in all branches of medicine, including dermatology [15]. Indeed, hundreds of dermatology-specific apps focusing on disease guides, self-surveillance, and self-diagnosis or providing general dermatology reference and educational aid are available at app stores [15]. With regard to vaccines, mobile apps have been shown to raise disease-specific awareness on VPDs [16,17]; improve health care providers communication about vaccination [18]; and, to a lesser extent, increase vaccination uptake [19].

The insufficient vaccination coverage in patients with psoriasis and AD despite their increased risk of VPDs is a clear unmet need in this large population group. Low awareness is likely a major driver of the low vaccination adherence. In proof, three-fourths of German patients with psoriasis and AD reported that they had not been vaccinated against pneumococcal disease because of a lack of recommendation by a physician [20]. Furthermore, to the best of our knowledge, there are no unified Italian recommendations for comprehensive immunization schedules for adult patients with psoriasis and AD. Current guidance is fragmented, exemplified by the limited scope of disease-specific recommendations, such as those for the HZ vaccine [5]. This lack of a centralized resource complicates clinical decision-making for both patients and health care providers. Against this background, we recently developed and launched the DermatoVax mobile app and a companion website that specifically target Italian adults living with psoriasis and AD. The overarching aim of this project was to raise awareness of immunization and, hopefully, increase vaccination uptake among these patients. In this paper, we report the process of development, initial validation, and evaluation of the DermatoVax app.

Methods

Development of the DermatoVax App

The DermatoVax project was conceived as a way to increase knowledge and promote awareness of the importance of vaccination among adult (aged ≥18 years) patients with psoriasis and AD, with the ultimate goal of improving vaccination uptake in this target population group. The process of app development involved several iterative phases and considered our previous experience with the Pneumo Rischio app launched to increase awareness of invasive pneumococcal disease in the general population [16]. Indeed, both projects share several common aspects in their goals and features and were developed in a user-centered fashion.

To decide on the app content and features, a steering committee was first organized; it was composed of 3 dermatologists, 2 public health physicians, 1 specialist in communication, and 2

XSL•FO
RenderX

patient representatives. Following agreement on the scope, content, and features of the future app, an initial on-paper prototype was conceived and reviewed by the committee members. It was unanimously felt that the app should be maximally functional, highly intuitive, and user-friendly.

All textual narratives used on the app were assessed for their readability using the Gulpease index [21], which is specific to the Italian language. It considers the average number of words per sentence and the average number of letters per word. The index ranges from 0 to 100, where 100 indicates the best readability. We chose the target index value of ≥40 as texts with a Gulpease index of <40 are expected to be difficult to read for individuals with a high school diploma. Patient representatives reviewed the app content for comprehensibility, and based on their feedback, some small changes were made. A reconciled paper-based version was shared with a contracted IT company to create an iterative series of wireframes, which were used to test and optimize the overall design, schematic, pathways, and interactions.

An initial app prototype was then developed and piloted in several modalities. First, potential end users (medical residents, nurses, and administrative staff), who were all outside the research team (n=6; 3/6, 50% male and 3/6, 50% female; aged 26-66 years), were asked to navigate the app and perform 5 different technical tasks (to locate the privacy policy, perform a mock session using the vaccine checker, locate information on psoriasis, locate information on influenza vaccination, and identify the nearest vaccination center). Technical effectiveness was measured by means of the task completion rate, which is the proportion of successful attempts to the total number of attempts [22].

Second, the computational algorithm was tested for eventual bugs by performing mock sessions, in which 50 hypothetical patients with psoriasis (n=25, 50%) and AD (n=25, 50%) of different ages, with different comorbidities, and on various treatment regimens completed the checker, which is the core app component (see the Content and Features of the DermatoVax App section), and the output produced was verified for correctness and completeness. Finally, individual usability sessions with 3 patients with psoriasis and 2 patients with AD were conducted. These patients were instructed to navigate through all app components, fill in the checker, download the resulting output report, and assess its clarity. The customer feedback was gathered, and eventual amendments were made. The final version of the app was released on the Apple App Store in May 2023 (Figure 1). A representative set of screenshots of the DermatoVax app is also provided in Multimedia Appendix 1.

Figure 1. Representative screenshots of the DermatoVax app: (A) screenshot of the home page, with project explanation and links to the various parts of the app; (B) screenshot that shows an example of information on the influenza vaccine; and (C) screenshot of the vaccine center finder, where patients can find the closest vaccination center in a 5- or 10-km radius by typing their address.

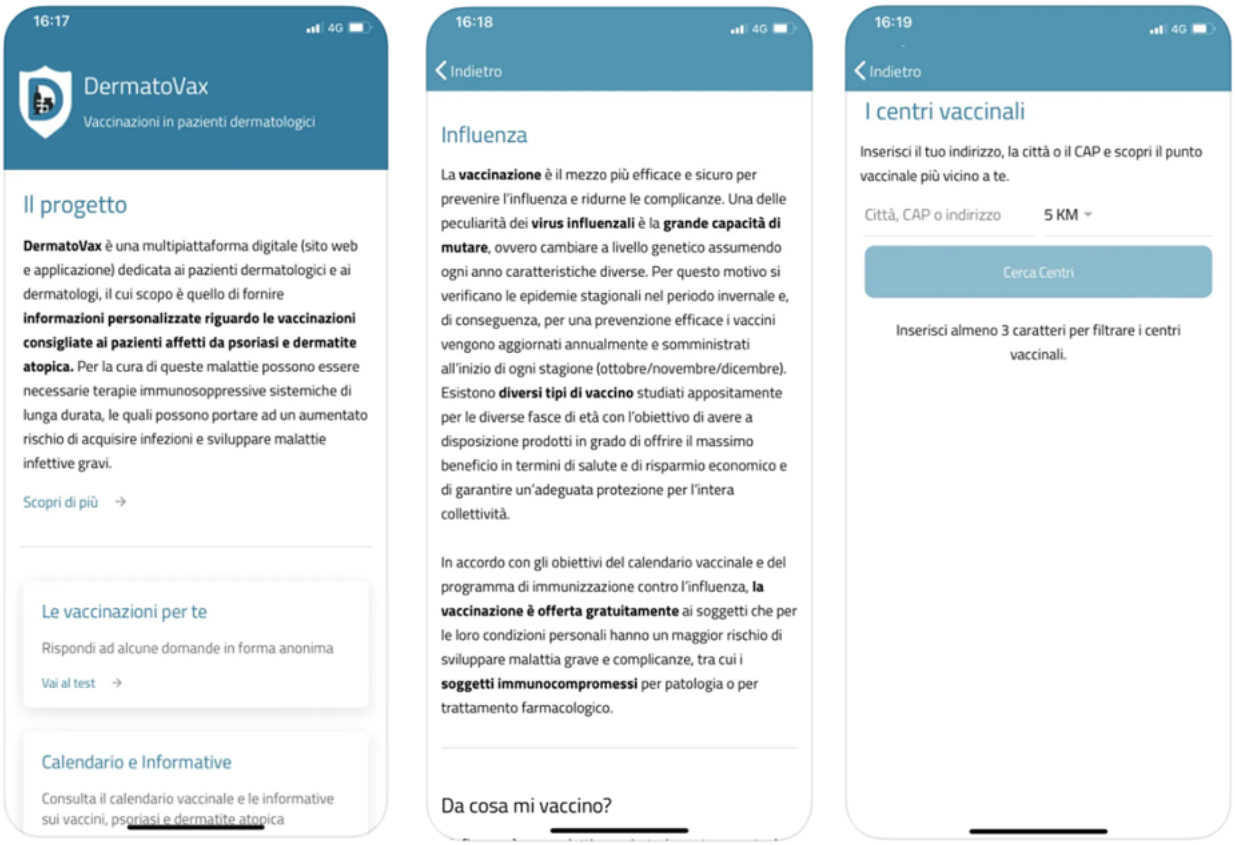

XSL•FO
RenderX

Content and Features of the DermatoVax App

The DermatoVax app consists of 4 main sections. The first section provides a brief description of the project and a periodically updated news feed. The app clearly identifies its owner and developer and the project funder and provides contact details, bibliographic references, terms and conditions, a privacy policy, and links to principal institutional websites.

The second section is purely informative and provides essential information on psoriasis, AD, recommended vaccines, and the associated VPDs. The Italian National Immunization Plan [23] was the primary normative document used to identify the recommended vaccines according to the age group and risk factors. As no specific Italian guidelines for vaccination of patients with psoriasis and AD are available, and given the fact that there is a large overlap between immunosuppressive and immunomodulating drugs for the treatment of psoriasis or AD and rheumatic diseases, we then consulted the recommendations for vaccination of patients with rheumatic diseases issued by the Italian Society of Rheumatology [24]. Indeed, while the Italian National Immunization Plan [23] is intended as a general reference document, the Italian Society of Rheumatology recommendations [24] provide a comprehensive assessment (including grading of the quality of evidence and strength of recommendations) of each inactivated or live attenuated vaccine also in relation to the timing of vaccine administration during immunosuppressive and immunomodulating therapy with the most common drugs. In patients with psoriasis and AD, vaccines may be indicated for 3 main reasons (due to age; the presence of comorbidities such as chronic cardiovascular, respiratory, and other conditions; or the immunosuppressive treatment regimen).

The core part of the app is a vaccine checker, which allows for the creation of a list of recommended vaccines from a short set of input questions. As shown in Table 1, the checker considers patients' nosological entity, eventual hand involvement, age, current or upcoming chronic treatment with 1 of 25 common drugs and their dosing, and presence of comorbidities and other risk factors. Upon informed consent, the checker is filled in, and an output report is produced. The latter summarizes patients' inputs and provides a detailed list of the recommended vaccines. In particular, the following vaccines were considered: diphtheria, tetanus, and acellular pertussis; *Hemophilus influenzae* type B; hepatitis B; hepatitis A; measles, mumps, and rubella; varicella; HZ; quadrivalent meningococcal vaccine; meningococcal B vaccine; conjugated and 23-valent polysaccharide pneumococcal vaccines; seasonal influenza; COVID-19; human papillomavirus; and mpox. Each vaccine-specific recommendation is accompanied by a brief description of vaccination timing in relation to the current or planned treatment (eg, according to the medicine used, live attenuated vaccines should not be administered for a certain period, whereas inactivated vaccines may be generally administered during the immunosuppressive therapy, although their administration is preferable at least 2 weeks before treatment initiation), boosting (eg, a diphtheria, tetanus, and acellular pertussis vaccine booster dose every 10 years), and seasonality (eg, administration of influenza and COVID-19 inactivated vaccines). The list of recommended vaccines may be saved and shared with the patients' dermatologists, general practitioners (GPs), or other health care providers. Patients are warned that the report has an informative character only and are invited to refer to their immunization records and discuss vaccination strategies with their dermatologists or vaccinating physicians. As the algorithm does not consider eventual risks of professional exposure or travel intentions, patients are informed that other vaccines may also be recommended, information on which may be obtained from health care providers.

Finally, the app allows users to identify the nearest vaccination centers and offers links to institutional websites where users can find further relevant information on vaccines, including possible adverse events and their reporting procedures.

XSL•FO
RenderX

Table 1. Items and response options of the vaccine checker used in the DermatoVax mobile app to produce a list of recommended vaccines.

Item	Response options
Disease	"Psoriasis" and "atopic dermatitis"
Hand involvement	"Yes" and "no"
Age	"18-59 years," "60-64 years," and "65 years or more"
Current or upcoming chronic (>2 weeks) therapy	"Yes" and "no"
Medication prescribed (if the previous item was answered with "yes")	"Abrocitinib," "acitretin," "adalimumab," "apremilast," "baricitinib," "bimekizumab," "brodalumab," "certolizumab pegol," "ciclosporin," "dimethyl fumarate," "dupilumab," "etanercept," "golimumab," "guselkumab," "infliximab," "ixekizumab," "methotrexate," "prednisone," "risankizumab," "secukinumab," "tildrakizumab," "topical medications (cremes, ointments, etc.)," "tralokinumab," "upadacitinib," and "ustekinumab"
Ciclosporin daily dose (if the previous item was answered with "ciclosporin")	Dose in milligrams
Methotrexate weekly dose (if the previous item was answered with "methotrexate")	Dose in milligrams
Prednisone daily dose (if the previous item was answered with "prednisone")	Dose in milligrams
Weight (if the previous item was answered with "ciclosporin" or "methotrexate")	Weight in kilograms
Health conditions or other lifestyle characteristics	"Type 1 diabetes," "type 2 diabetes," "respiratory diseases," "cardiovascular diseases," "congenital or acquired immunodeficiency," "HIV infection," "chronic kidney failure/renal disease," "asplenia," "chronic hepatic disease or alcoholism," "man who has sex with men," and "pregnancy"

External Evaluation of App Quality

To evaluate the quality of the DermatoVax app, a convenience sample of board-certified dermatologists and public health physicians were enrolled during 3 local conferences where the project was presented. All physicians were outside the research team and had been unaware of the project. In particular, following a presentation of the app, the conference audience (approximately 80-120 attendees per conference) was asked to download the app (<2 minutes), navigate all its components and functions for at least 10 minutes, and fill in the previously validated Italian user version of the Mobile App Rating Scale (uMARS) [25]. Contrary to the professional version of the Mobile App Rating Scale (MARS) [26,27], the uMARS is simpler and does not require preliminary training [28]. The tool consists of 4 objective quality subscales, namely, engagement, functionality, aesthetics, and information, which comprise a total of 16 items. There are also 2 additional subscales that measure subjective quality (4 items) and perceived impact (6 items) of the evaluated app. Each of the 26 items is rated on a 5-point Likert scale, where a greater score corresponds to a better quality. Scores of individual items are averaged to obtain a mean quality score for each objective subscale, whereas the resulting 4 mean subscale scores are averaged to obtain a uMARS total score. The Italian version of the uMARS has an excellent internal consistency (Cronbach α=0.94) [25]. The questionnaire was administered anonymously using a pencil-and-paper format.

With an assumed SD of the total uMARS score of 0.45 [29], precision of 0.15, and confidence level of 95%, at least 35 physicians were needed to estimate the population mean.

Data Analysis

Total, subscale, and item-specific uMARS scores were summarized as means with SDs and medians with IQRs. Two-tailed independent *t* tests were used to compare uMARS scores between dermatologists and public health physicians. Data analysis was conducted in the R *stats* packages (version 4.1.0; R Foundation for Statistical Computing) and Microsoft Excel.

Ethical Considerations

This study was conducted in accordance with the Declaration of Helsinki of 1964 and its later amendments. The study protocol was approved by the Ethics Committee of Liguria Region, Italy (583/2021; ID 11870). Informed consent was obtained from all participants. All uMARS questionnaires collected were anonymous. No compensation was provided to either patients or health care professionals.

Results

Pilot-Testing of the DermatoVax App

During the piloting of the app prototype, no major changes were deemed necessary. In particular, all 6 testers succeeded in performing the 5 tasks with a task completion rate of 100% (30/30 tasks), and the task completion time was deemed acceptable (range 1-4 min). However, some minor changes regarding navigational components and shortcuts were suggested. The output files with a list of recommended vaccines generated from the input data of 50 mock patients were deemed correct and complete, suggesting a proper functioning of the algorithm.

XSL•FO
RenderX

During the individual usability sessions with 5 real patients, the checker was completed within 6 minutes (range 2-6 min). The feedback obtained from the patients was positive, and no changes were deemed necessary to any app component. In the final app version, the average readability index of the textual narratives was 41.0.

Quality Scores and Perceived Impact of the DermatoVax App

A total of 46 physicians (n=29, 63% dermatologists and n=17, 37% public health physicians) filled in the uMARS tool (Table 2). Their median age was 47 (IQR 38-59) years, and approximately two-thirds (29/46, 63%) were female. The average uMARS score was 4.22/5 (SD 0.49), with a range of 2.91 to 5.00. The mean scores were similar (*P*=.33) between dermatologists (mean 4.28, SD 0.48) and public health physicians (mean 4.12, SD 0.51). An analysis by single objective subscale revealed that the average scores for the information (4.64/5, SD 0.42) and functionality (4.47/5, SD 0.46) dimensions were comparatively high, whereas those for the dimensions of aesthetics (3.88/5, SD 0.78) and engagement (3.89/5, SD 0.68) were substantially lower.

Regarding the subjective quality dimension, the average score of 4 items was 3.59 (SD 0.51). The score was mainly determined by the "Would you pay for this app?" item: only 39% (18/46) of physicians replied affirmatively ("yes" or "definitely yes"), and the average score was only 2.74 (SD 1.44). Conversely, most responders would recommend the DermatoVax app (36/46, 78%) and use it at least 10 times in the following 12 months (31/46, 67%). The average star rating (out of 5 stars) was 3.87 (SD 0.65).

Finally, we investigated physicians' perceived impact of the DermatoVax app on vaccination-related knowledge, attitudes, and practices among patients with psoriasis and AD. As shown in Figure 2, the perceived impact was high: at least four-fifths of physicians agreed or strongly agreed that use of the app would increase patient awareness (39/46, 85%) and knowledge (41/46, 89%) of vaccines and vaccination uptake (37/46, 80%).

Table 2. User version of the Mobile App Rating Scale (uMARS) scores from physicians who evaluated the DermatoVax mobile app by objective subscale and item (N=46).

	Score, mean (SD)	Score, median (IQR)
Engagement subscale		
Entertainment	3.83 (0.77)	4 (3-4)
Interest	4.33 (0.67)	4 (4-5)
Customization	3.50 (1.15)	4 (3-4)
Interactivity	3.52 (0.98)	4 (3-4)
Target group	4.26 (0.74)	4 (4-5)
Total	3.89 (0.68)	4.00 (3.40-4.20)
Functionality subscale		
Performance	4.33 (0.67)	4 (4-5)
Ease of use	4.67 (0.47)	5 (4-5)
Navigation	4.46 (0.66)	5 (4-5)
Gestural design	4.41 (0.62)	4 (4-5)
Total	4.47 (0.46)	4.50 (4.25-4.75)
Aesthetics subscale		
Layout	4.09 (0.81)	4 (4-5)
Graphics	3.78 (0.92)	4 (3-4)
Visual appeal	3.78 (0.76)	4 (3-4)
Total	3.88 (0.78)	4.00 (3.33-4.33)
Information subscale		
Quality of information	4.74 (0.49)	5 (5-5)
Quantity of information	4.52 (0.62)	5 (4-5)
Visual information	4.54 (0.55)	5 (4-5)
Credibility of source	4.80 (0.55)	5 (5-5)
Total	4.64 (0.42)	4.75 (4.38-5.00)
uMARS total score	4.22 (0.49)	4.29 (3.98-4.49)

XSL•FO
RenderX

Figure 2. Physicians' perceived impact of the DermatoVax app on vaccination-related knowledge, attitudes, and practices (N=46).

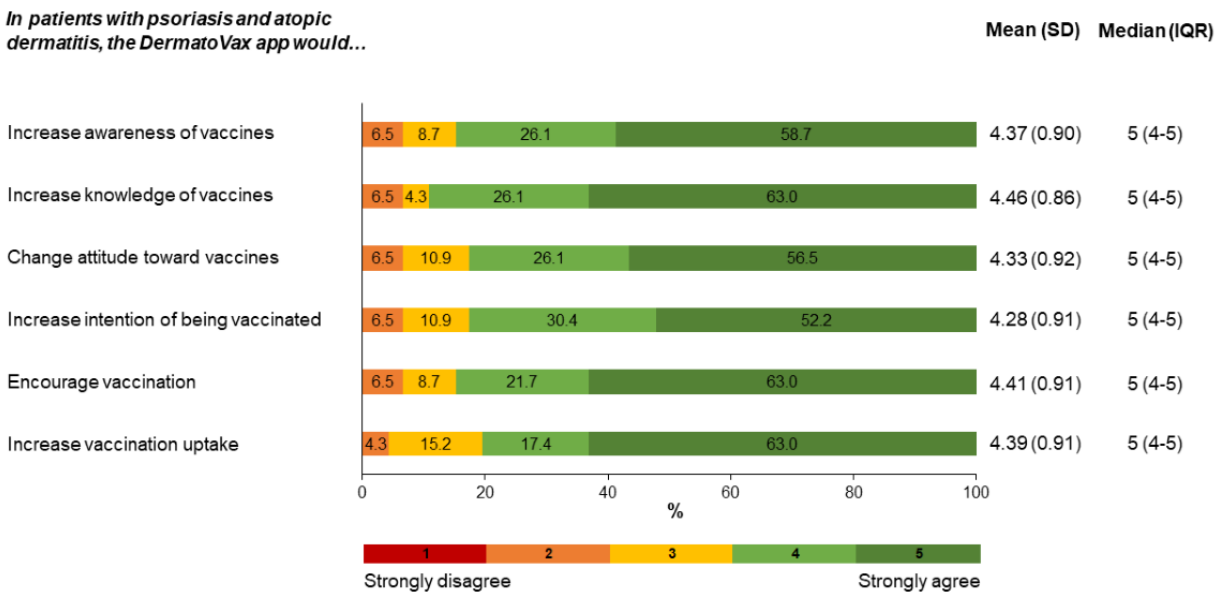

Discussion

Principal Findings

In this paper, we described an initial evaluation of DermatoVax, the first mobile app to increase awareness of vaccines and VPDs that was specifically developed for patients with immune-mediated inflammatory skin diseases. The advent of novel immunomodulatory and biological treatments has revolutionized the treatment of psoriasis and AD. At the same time, these new-generation therapies require particular considerations regarding infections and immunization [30]. Indeed, the recommended vaccination schedule for patients with psoriasis and AD is rather complex and nuanced and depends on age, presence of comorbidities, treatment regimen, and other factors [24,30,31]. This complexity may be burdensome for both patients and health care workers and can lead to reduced vaccination adherence [32]. We believe that mHealth projects such as DermatoVax may empower patients and dermatologists to navigate the complex vaccination landscape. Indeed, most participating dermatologists and public health physicians believed that the DermatoVax app can raise awareness of vaccines (39/46, 85%) and increase vaccination uptake (37/46, 80%) in patients with psoriasis and AD.

The initial validation and individual usability sessions showed that the DermatoVax mobile app was valued by potential end users, who judged it as intuitive, user-friendly, and potentially helpful. Contrary to traditional immunization schedules, which often have numerous and complex footnotes [33], personalized reports on recommended vaccines generated by the app were deemed easily interpretable. Chen et al [34] suggest that human factors can be implemented to improve the usability of adult immunization schedules; the application of some design principles, such as consistency, simplicity, and clarity, enhances schedule efficiency while maintaining a form that fits better with user expectations of the schedule.

The quality of mHealth apps is a multidimensional concept that defines user experience [26]. The MARS [26] and uMARS [28] tools are currently the most widely used scales for evaluating content and the overall and domain-specific quality of mHealth apps. Indeed, the content of a health-related app may not accurately reflect its description and features [26]. An Italian study [35] identified 121 vaccine-related apps, only 32% of which provided some information on vaccines or VPDs. In our study, the mean uMARS scores given by physicians (4.22, SD 0.49) were relatively high. In comparison, the average MARS scores of 21 immunization-related apps available in Chinese app stores is 3.33 (SD 1.07) [36]. A German evaluation of 8 apps focused on psoriasis yielded MARS scores ranging from 3.00 to 4.18 [37]. However, it should be noted that MARS and uMARS scores may be not fully comparable; some studies [37,38] have suggested that mean uMARS scores given by patients may be slightly (0.3 points) lower than average MARS scores given by physicians. This fact underlines the importance of patient involvement in all steps of the app development and validation process.

In evaluating single dimensions of app quality, we found that the mean information (4.64, SD 0.42) and functionality (4.47, SD 0.46) uMARS subscales received the highest scores, whereas the engagement (mean 3.89, SD 0.68) and aesthetics (mean 3.88, SD 0.78) subscales were associated with lower scores. This differential scoring pattern clearly highlights the distinction between the app's utility and its appeal. A similar pattern has been reported in previous studies on the quality of vaccine- [36] and psoriasis-related [37] apps. This finding was expected because the DermatoVax app is intrinsically bound to be scientifically sound and computationally simple and, therefore, highly informative and functional. On the other hand, an

XSL•FO
RenderX

improved functionality and minimalist design may have compromised some engagement attributes and the overall aesthetic perception of the app. In other words, the focus on streamlined utility can clash with consumer expectations for visually rich and interactive apps. This, in turn, could eventually lead to low retention rates (ie, when users delete the app after a few sessions). Future developments of the DermatoVax app will prioritize enhanced engagement and refined aesthetics. This could be achieved through a comprehensive user interface and user experience redesign, the integration of additional gamification elements, and advanced personalization features and feedback elements. In particular, the user interface refinement would ensure a visually polished, modern, and intuitive appearance, whereas the user experience redesign would further streamline the navigation flow, making the latter even less effortful. Implementation of more advanced features that tailor content delivery based on the patient's specific clinical profile (eg, previous vaccines) will further enhance app personalization. Finally, the introduction of dynamic feedback loops would improve long-term user retention.

We then found that only one-third of physicians (18/46, 39%) would purchase the app, which suggests that the willingness to pay for apps such as DermatoVax is likely low. We found a similar proportion when evaluating the Pneumo Rischio mobile app [29], which has several similarities to the DermatoVax app. This finding indicates that, to reach a larger number of end users, such "niche" and prevention-oriented apps should be accessible to everyone at no cost.

To our knowledge, there are no data on vaccination uptake (of any vaccine) among Italian patients with immune-mediated inflammatory skin diseases. Considering the available data from international surveys [9,10], we believe that immunization rates among Italian patients are similarly low. In their pilot work, Feldman et al [39] highlighted numerous unmet preventive health care needs, including vaccination, among patients attending a dermatology clinic and discussed ways of improvement. In particular, one approach could consist of more comprehensive general preventive care provided by dermatologists. However, as noted by the authors, even if dermatologists were specifically trained, this possibility may compromise the time efficiency of dermatological services. Another approach, which seems more promising, consists of a cooperation between dermatologists and primary care physicians: the former may actively identify patients with unmet preventive care needs and refer those patients directly to GPs or other preventive medicine specialists [39]. We believe that mHealth tools such as DermatoVax may facilitate continuity of care, enabling interdisciplinary teamwork that puts patients' need first. Indeed, some research suggests [20] that dermatologists are rarely involved in vaccination counseling. One recent study [40] found that 44% of patients with psoriasis searched the web for additional information as the information provided by their physicians was insufficient. At the same time, 45% wanted their physicians to recommend health-related websites to them. In this regard, dermatologists may recommend their patients free and user-friendly apps to check their vaccination status and eventually receive missing recommended vaccines from their GPs or public health physicians.

This study represents a preliminary evaluation and, as such, is subject to several important limitations. First, external evaluation of the DermatoVax app and its quality was conducted by dermatologists and public health physicians only. Therefore, the most important shortcoming of this study is the lack of formal quality evaluation of the app by patients. Analogously, data on actual app use, and especially its ability to change vaccination status, are yet to be collected and assessed. The app ratings obtained from health care professionals may be systematically different from those provided by patients, who are supposed to be the main app audience. For instance, in evaluating an app, patients could prioritize features such as usability, empowerment, and self-sufficiency, whereas physicians could prioritize clinical aspects such as the use of evidence-based practices and guidelines. More importantly, health-related apps should ideally be trialed for their efficacy or effectiveness. However, only a few mobile apps have undergone formal evaluation of their effectiveness [26]. As underlined by the available systematic reviews [41,42], data on the efficacy and effectiveness of mobile apps and other digital technologies on actual vaccine receipt are limited, heterogeneous, and inconclusive, with some studies showing significant benefits and other trials reporting no effect. To address this limitation, we aim to conduct a larger and more comprehensive evaluation of the DermatoVax app to determine both its quality from the end user perspective and its capacity to modify the immunization knowledge, attitudes, and practices of patients with psoriasis and AD. Second, individual usability sessions were conducted with 5 patients with psoriasis and AD as, according to the "5-user rule," only 5 participants are needed to discover 80% of usability problems [43]. However, we recognize that this number may be small for a comprehensive understanding across diverse patient profiles or for uncovering less frequent but critical issues. Third, owing to the cross-sectional nature of this study without longitudinal follow-up, our findings reflect only a single snapshot in time. Finally, as the app was developed and evaluated solely in Italian, its current utility is restricted to Italian speakers, raising questions about its generalizability across diverse cultural settings and non–Italian-speaking populations such as immigrants.

Importantly, immunization strategies and recommendations, available vaccines, and biological therapies are constantly evolving. This dynamic landscape underscores the critical importance of keeping the app content current. For instance, respiratory syncytial virus vaccines for adults have recently become available [44]. Although these vaccines are not yet part of the Italian vaccination calendar and are not reimbursed, some Italian scientific societies have advocated for the inclusion of respiratory syncytial virus vaccines in the national vaccination schedule [44]. Hence, our commitment to app maintenance involves periodic updates to consistently include these novel advancements.

Conclusions

In conclusion, this preliminary assessment showed that the DermatoVax mobile app is a promising tool to raise awareness of immunization and VPDs in adult patients with psoriasis and AD, whose uptake of the recommended vaccines is insufficient.

XSL•FO
RenderX

Positive feedback of end users and good quality scores of the app from health care providers merit further evaluation, in particular of whether app use may effectively change vaccination-related behaviors. Furthermore, continuously evolving adult immunization schedules [33] and clinical management protocols for immune-mediated inflammatory skin diseases [45] impose the necessity of keeping the app content constantly up-to-date.

Acknowledgments

This project was supported by Pfizer (grant 68984403). The funder had no involvement in the study design; data collection, analysis, or interpretation; or writing of the manuscript. Generative artificial intelligence was not used in any portion of manuscript writing.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: GG (lead), DP (equal), PLL (supporting)
Data curation: LV (lead), EM (supporting)
Formal analysis: LV (lead), EM (supporting)
Funding acquisition: GG
Investigation: GG (lead), LV (equal), MB (supporting), IDK (supporting)
Methodology: DP (lead), GG (equal)
Project administration: GG (lead), DP (supporting), EC (supporting), GI (supporting)
Supervision: EC (equal), GI (equal)
Visualization: PLL
Writing—original draft: GG
Writing—review and editing: DP (lead), EC (supporting), GI (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Representative set of screenshots of the DermatoVax app developed to raise awareness of immunization in adult patients with psoriasis and atopic dermatitis.
[PNG File , 181 KB - formative_v9i1e80113_app1.png]

References

1. Parisi R, Symmons DP, Griffiths CE, Ashcroft DM, Identification Management of Psoriasis and Associated ComorbidiTy (IMPACT) project team. Global epidemiology of psoriasis: a systematic review of incidence and prevalence. J Invest Dermatol 2013 Feb;133(2):377-385 [FREE Full text] [doi: 10.1038/jid.2012.339] [Medline: 23014338]
2. Barbarot S, Auziere S, Gadkari A, Girolomoni G, Puig L, Simpson EL, et al. Epidemiology of atopic dermatitis in adults: results from an international survey. Allergy 2018 Jun 13;73(6):1284-1293. [doi: 10.1111/all.13401] [Medline: 29319189]
3. Gisondi P, Puig L, Richard MA, Paul C, Nijsten T, Taieb C, EADV Burden of Skin Diseases Project Team. Quality of life and stigmatization in people with skin diseases in Europe: a large survey from the 'burden of skin diseases' EADV project. J Eur Acad Dermatol Venereol 2023 Oct 08;37(S7):6-14. [doi: 10.1111/jdv.18917] [Medline: 37806002]
4. Singer D, Thompson-Leduc P, Ma S, Gupta D, Cheng WY, Sendhil SR, et al. Burden of herpes zoster among patients with psoriatic arthritis in the United States. Psoriasis (Auckl) 2024 Jun 22;14:63-78 [FREE Full text] [doi: 10.2147/PTT.S430151] [Medline: 38939905]
5. Ciccarese G, Drago F, Herzum A, Atzori L, Dattola A, Galluzzo M, et al. Anti-Herpes zoster vaccination in patients with dermatologic diseases: a position statement from the Italian SIDeMaST group of sexually transmitted, infectious and tropical diseases. Ital J Dermatol Venerol 2024 Aug;159(4):375-379 [FREE Full text] [doi: 10.23736/S2784-8671.24.07895-2] [Medline: 38780911]
6. Narala S, Hata TR. Adult atopic dermatitis with comorbid atopic disease is associated with increased risk of infections: a population-based cross-sectional study. Dermatol Ther (Heidelb) 2017 Mar 30;7(1):111-121 [FREE Full text] [doi: 10.1007/s13555-017-0172-7] [Medline: 28138890]
7. Chovatiya R, Silverberg JI. Pathophysiology of atopic dermatitis and psoriasis: implications for management in children. Children (Basel) 2019 Oct 04;6(10):108 [FREE Full text] [doi: 10.3390/children6100108] [Medline: 31590274]

XSL•FO
RenderX

8. Guttman-Yassky E, Krueger JG. Atopic dermatitis and psoriasis: two different immune diseases or one spectrum? Curr Opin Immunol 2017 Oct;48:68-73. [doi: 10.1016/j.coi.2017.08.008] [Medline: 28869867]
9. Hren MG, Khattri S. Low rates of vaccination among atopic dermatitis, alopecia areata, psoriasis, and psoriatic arthritis patients on biologics. Arch Dermatol Res 2024 May 25;316(6):285. [doi: 10.1007/s00403-024-03037-6] [Medline: 38796548]
10. Bonhomme A, Fréling E, Reigneau M, Poreaux C, Valois A, Truchetet F, et al. [Vaccination status in psoriasis patients on immunosuppressant therapy (including biologics)]. Ann Dermatol Venereol 2017 Feb;144(2):92-99. [doi: 10.1016/j.annder.2016.09.035] [Medline: 27771122]
11. Noe MH, Archila M, Barbieri JS, Goldman N, Lopez CG, Mostaghimi A, et al. The patient perspective on vaccine uptake in adults with psoriasis and eczema. Arch Dermatol Res 2023 Aug 07;315(6):1583-1592 [FREE Full text] [doi: 10.1007/s00403-023-02546-0] [Medline: 36749390]
12. Noe MH, Shin DB, Hubbard RA, Hennessy S, Gelfand JM. Influenza vaccination rates in adults with psoriasis compared to adults with other chronic diseases. J Invest Dermatol 2019 Feb;139(2):473-475 [FREE Full text] [doi: 10.1016/j.jid.2018.09.012] [Medline: 30315780]
13. Korouri E, Jeong C, Peterson H, Valenzuela F, Romiti R, Didaskalu JA, et al. Global comparison of COVID-19 vaccination rates among psoriasis patients. Life (Basel) 2024 Oct 13;14(10):1297 [FREE Full text] [doi: 10.3390/life14101297] [Medline: 39459597]
14. Schläpfer S, Schneider F, Santhanam P, Eicher M, Kowatsch T, Witt CM, et al. Engagement with a relaxation and mindfulness mobile app among people with cancer: exploratory analysis of use data and self-reports from a randomized controlled trial. JMIR Cancer 2024 May 31;10:e52386 [FREE Full text] [doi: 10.2196/52386] [Medline: 38819907]
15. Alomary SA, Mackenzie E, Khan S, Rao BK. Mobile apps in dermatology: bridging innovation and care. Skin Res Technol 2024 Mar 29;30(3):e13640 [FREE Full text] [doi: 10.1111/srt.13640] [Medline: 38424722]
16. Panatto D, Domnich A, Gasparini R, Bonanni P, Icardi G, Amicizia D, et al. Development and preliminary data on the use of a mobile app specifically designed to increase community awareness of invasive pneumococcal disease and its prevention. Hum Vaccin Immunother 2016 Apr 02;12(4):1080-1084 [FREE Full text] [doi: 10.1080/21645515.2015.1114196] [Medline: 26795065]
17. Ruiz-López T, Sen S, Jakobsen E, Tropé A, Castle PE, Hansen BT, et al. FightHPV: design and evaluation of a mobile game to raise awareness about human papillomavirus and nudge people to take action against cervical cancer. JMIR Serious Games 2019 Apr 08;7(2):e8540 [FREE Full text] [doi: 10.2196/games.8540] [Medline: 30958271]
18. Grabert BK, McRee AL, Henrikson NB, Heisler-MacKinnon J, Blasi PR, Norris CM, et al. Feasibility of using an app-based coaching intervention to improve provider communication about HPV vaccination. Transl Behav Med 2023 Aug 11;13(8):581-588 [FREE Full text] [doi: 10.1093/tbm/ibad002] [Medline: 36999806]
19. Lee WN, Stück D, Konty K, Rivers C, Brown CR, Zbikowski SM, et al. Large-scale influenza vaccination promotion on a mobile app platform: a randomized controlled trial. Vaccine 2020 Apr 16;38(18):3508-3514. [doi: 10.1016/j.vaccine.2019.11.053] [Medline: 31787410]
20. Wellmann P, Kromer C, Siemer R, Klein S, Mohr J, Lippert U, et al. Low pneumococcal vaccination among patients with psoriasis in Germany: results from Vac-Pso. Vaccines (Basel) 2022 Jun 23;10(7):1005 [FREE Full text] [doi: 10.3390/vaccines10071005] [Medline: 35891172]
21. Sentences with high readability. Corpus Italiano. URL: https://www.corpusitaliano.it/en/help/readability.html [accessed 2025-12-02]
22. O'Connor SR, Wilson V, Brown D, Cleland I, McCance TV. Development and optimisation of a mobile app (iMPAKT) for improving person-centred practice in healthcare settings: a multi-methods evaluation study. Digit Health 2024 Oct 29;10:20552076241271788 [FREE Full text] [doi: 10.1177/20552076241271788] [Medline: 39493639]
23. National vaccine prevention plan 2023-2025. Italian Ministry of Health. URL: https://www.trovanorme.salute.gov.it/norme/dettaglioAtto.spring?id=95963&page=newsett [accessed 2025-05-29]
24. Recommendations of the Italian Society of Rheumatology on the vaccination of patients with rheumatic diseases. Italian Society of Rheumatology. URL: https://www.iss.it/-/vaccinazioni-in-pazienti-affetti-da-malattie-reumatologiche [accessed 2025-05-29]
25. Morselli S, Sebastianelli A, Domnich A, Bucchi C, Spatafora P, Liaci A, et al. Translation and validation of the Italian version of the user version of the Mobile Application Rating Scale (uMARS). J Prev Med Hyg 2021 Apr 29;62(1):E243-E248 [FREE Full text] [doi: 10.15167/2421-4248/jpmh2021.62.1.1894] [Medline: 34322643]
26. Stoyanov SR, Hides L, Kavanagh DJ, Zelenko O, Tjondronegoro D, Mani M. Mobile app rating scale: a new tool for assessing the quality of health mobile apps. JMIR Mhealth Uhealth 2015 Mar 11;3(1):e27 [FREE Full text] [doi: 10.2196/mhealth.3422] [Medline: 25760773]
27. Domnich A, Arata L, Amicizia D, Signori A, Patrick B, Stoyanov S, et al. Development and validation of the Italian version of the Mobile Application Rating Scale and its generalisability to apps targeting primary prevention. BMC Med Inform Decis Mak 2016 Jul 07;16(1):83 [FREE Full text] [doi: 10.1186/s12911-016-0323-2] [Medline: 27387434]
28. Stoyanov SR, Hides L, Kavanagh DJ, Wilson H. Development and validation of the user version of the Mobile Application Rating Scale (uMARS). JMIR Mhealth Uhealth 2016 Jun 10;4(2):e72 [FREE Full text] [doi: 10.2196/mhealth.5849] [Medline: 27287964]

XSL•FO
RenderX

29. Panatto D, Domnich A, Gasparini R, Bonanni P, Icardi G, Amicizia D, et al. An eHealth project on invasive pneumococcal disease: comprehensive evaluation of a promotional campaign. J Med Internet Res 2016 Dec 02;18(12):e316 [FREE Full text] [doi: 10.2196/jmir.6205] [Medline: 27913372]
30. Fan R, Cohen JM. Vaccination recommendations for psoriasis and atopic dermatitis patients on biologic therapy: a practical guide. Yale J Biol Med 2022 Jun 30;95(2):249-255 [FREE Full text] [Medline: 35782480]
31. Chat VS, Ellebrecht CT, Kingston P, Gondo G, Bell S, Cordoro KM, et al. Vaccination recommendations for adults receiving biologics and oral therapies for psoriasis and psoriatic arthritis: Delphi consensus from the medical board of the National Psoriasis Foundation. J Am Acad Dermatol 2024 Jun;90(6):1170-1181. [doi: 10.1016/j.jaad.2023.12.070] [Medline: 38331098]
32. Xu J, Cui Y, Huang C, Dong Y, Zhang Y, Fan L, et al. Prevalence and factors associated with pentavalent vaccination: a cross-sectional study in Southern China. Infect Dis Poverty 2023 Sep 15;12(1):84 [FREE Full text] [doi: 10.1186/s40249-023-01134-8] [Medline: 37715293]
33. Kim DK, Hunter P, Advisory Committee on Immunization Practices. Recommended adult immunization schedule, United States, 2019. Ann Intern Med 2019 Feb 05;170(3):182-192. [doi: 10.7326/m18-3600] [Medline: 30716757]
34. Chen DW, Schaeffer LM, Preusse K, Gable TM, Hartzell C, McGlynn S, et al. Improving the U.S. adult immunization schedule by applying usability principles. Proc Hum Factors Ergon Soc Annu Meet 2018 Sep 27;62(1):1316-1320. [doi: 10.1177/1541931218621301]
35. Catozzi D, Gualano MR, Voglino G, Corradi A, Bert F, Siliquini R. Mobile apps for vaccination: an in-depth scan of Apple App Store. Eur J Public Health 2020 Sep 30;30(Suppl. 5):v380-v381 [FREE Full text] [doi: 10.1093/eurpub/ckaa165.1034]
36. Zhang C, Guo X, Zhu R, Hou W, Wang L, Wang F, et al. Mobile apps for vaccination services: content analysis and quality assessment. Online J Public Health Inform 2024 Oct 03;16:e50364 [FREE Full text] [doi: 10.2196/50364] [Medline: 39361418]
37. Lull C, von Ahnen JA, Gross G, Olsavszky V, Knitza J, Leipe J, et al. German mobile apps for patients with psoriasis: systematic search and evaluation. JMIR Mhealth Uhealth 2022 May 26;10(5):e34017 [FREE Full text] [doi: 10.2196/34017] [Medline: 35617014]
38. Lambrecht A, Vuillerme N, Raab C, Simon D, Messner EM, Hagen M, et al. Quality of a supporting mobile app for rheumatic patients: patient-based assessment using the user version of the Mobile Application Scale (uMARS). Front Med (Lausanne) 2021 Jul 22;8:715345 [FREE Full text] [doi: 10.3389/fmed.2021.715345] [Medline: 34368202]
39. Feldman SR, Ravis S, Moran WP, Fleischer Jr AB. Patients seen in a dermatology clinic have unmet preventive health care needs. J Am Acad Dermatol 2001 Apr;44(4):706-709. [doi: 10.1067/mjd.2001.112914] [Medline: 11260555]
40. Erbas ME, Ziehfreund S, Wecker H, Biedermann T, Zink A. Digital media usage behavior and its impact on the physician-patient relationship: cross-sectional study among individuals affected by psoriasis in Germany. J Med Internet Res 2024 Aug 07;26:e57823 [FREE Full text] [doi: 10.2196/57823] [Medline: 39110972]
41. de Cock C, van Velthoven M, Milne-Ives M, Mooney M, Meinert E. Use of apps to promote childhood vaccination: systematic review. JMIR Mhealth Uhealth 2020 May 18;8(5):e17371 [FREE Full text] [doi: 10.2196/17371] [Medline: 32421684]
42. Atkinson KM, Wilson K, Murphy MS, El-Halabi S, Kahale LA, Laflamme LL, et al. Effectiveness of digital technologies at improving vaccine uptake and series completion - a systematic review and meta-analysis of randomized controlled trials. Vaccine 2019 May 21;37(23):3050-3060. [doi: 10.1016/j.vaccine.2019.03.063] [Medline: 31036457]
43. Kanis H. Estimating the number of usability problems. Appl Ergon 2011 Jan;42(2):337-347 [FREE Full text] [doi: 10.1016/j.apergo.2010.08.004] [Medline: 20880514]
44. Manzoni P, Baraldi E, Cetin I, Maggi S, Riccò M, Siliquini R, et al. Prevention of respiratory syncytial virus disease across the lifespan in Italy. Pneumonia (Nathan) 2025 Apr 05;17(1):8 [FREE Full text] [doi: 10.1186/s41479-025-00160-4] [Medline: 40186273]
45. David E, Ungar B, Renert-Yuval Y, Facheris P, Del Duca E, Guttman-Yassky E. The evolving landscape of biologic therapies for atopic dermatitis: present and future perspective. Clin Exp Allergy 2023 Feb 18;53(2):156-172. [doi: 10.1111/cea.14263] [Medline: 36653940]

Abbreviations

AD: atopic dermatitis
GP: general practitioner
HZ: herpes zoster
MARS: Mobile App Rating Scale
mHealth: mobile health
uMARS: user version of the Mobile App Rating Scale
VPD: vaccine-preventable disease

XSL•FO
RenderX

Edited by A Mavragani; submitted 04.07.25; peer-reviewed by G Ciccarese, M Borlu; comments to author 22.07.25; revised version received 01.08.25; accepted 20.10.25; published 09.12.25.

Please cite as:
Gasparini G, Panatto D, De Kermarek I, Valchi L, Massaro E, Lai PL, Burlando M, Icardi G, Cozzani E
A Mobile App to Enhance Awareness of Vaccination in Adults With Psoriasis and Atopic Dermatitis: Development and Preliminary Evaluation Study
JMIR Form Res 2025;9:e80113
URL: https://formative.jmir.org/2025/1/e80113
doi:10.2196/80113
PMID:

©Giulia Gasparini, Donatella Panatto, Irene De Kermarek, Lucia Valchi, Elvira Massaro, Piero Luigi Lai, Martina Burlando, Giancarlo Icardi, Emanuele Cozzani. Originally published in JMIR Formative Research (https://formative.jmir.org), 09.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Development and Implementation of an Internal Quality Control and External Quality Assessment Information System for a Regional Medical Laboratory Center: Pilot Design and Implementation Study

Dayang Chen[1,2*], PhD; Hongmei Mo[1,2*], PhD; Yingying Zhang[3*], BMed; Junlong Qin[1,2], BMed; Lijun Zhang[1,2], BMed; Changxian Jiang[4], BSc; Fuming Li[4], BSc; Xiuming Zhang[1,2,3,5], MRES; Tong Ou[1,2,3], PhD

[1]Department of Medical Laboratory, The Third Affiliated Hospital of Shenzhen University, Shenzhen, China
[2]Department of Medical Laboratory Center, Shenzhen Luohu Hospital Group, Shenzhen, China
[3]Department of Medical College, Shantou University, Shantou, China
[4]Bio-Rad Laboratories (Shanghai) Co, Ltd, Shanghai, China
[5]Department of Pathology, The Hong Kong University - Shenzhen Hospital, Shenzhen, China
[*]these authors contributed equally

Corresponding Author:
Tong Ou, PhD
Department of Medical Laboratory
The Third Affiliated Hospital of Shenzhen University
No 33 Zhongyuan Road, Buji St, Longgang District
Shenzhen, 518000
China
Phone: 86 28715347
Email: ot81061017@163.com

Abstract

Background: The Regional Medical Laboratory Center (RMLC) integrates laboratory departments from hospitals with different levels to optimize resource allocation, enhance testing efficiency, and promote the sharing of professional expertise. Despite these benefits, ensuring consistent quality control (QC) across the RMLC presents a significant challenge due to diverse equipment, procedures, and staff expertise levels. To address these challenges, this study provided an information tool called iLab, developed by Shenzhen Huikang Information Technology Co, Ltd, for managing internal quality control (IQC) and external quality assessment (EQA).

Objective: This study aimed to develop and implement an integrated informatics platform managing IQC and EQA across RMLC.

Methods: We integrated 4 software programs to manage IQC and EQA, laboratory information system (LIS) supplied by Shanghai Tengcheng Medical Tech-Info Co, Ltd; Quality Control Box (QCBOX) and Quality Control EQA Center (QCEC) provided by Bio-Rad Laboratories (Shanghai) Co, Ltd; and the iLab electronic recording system. The platform centralizes QC data processing, with QCBOX and QCEC handling operational workflows, while iLab oversees documentation, analytics, and approval processes. It was successfully deployed in August 2023 across a pilot network comprising 5 district hospitals, 1 central laboratory, and 4 community health stations in Shenzhen's Luohu District.

Results: By leveraging the LIS-QCBOX data flow, the platform streamlined the entire IQC process, which incorporated interlaboratory comparisons and iLab-facilitated postanalysis management. Concurrently, EQA was enhanced through electronic documentation, multirule analysis of acceptable data, regional proficiency testing via QCEC, and measurement uncertainty (MU) estimation. By September 2025, the platform supported 114 users, integrating 133 instruments and 859 active QC items, and generated 2331 monthly IQC analysis reports, 340 EQA multirule analyses, and 289 MU reports. Comparative analysis revealed marked improvements in key metrics from 2023 (preimplementation) to 2024 (postimplementation): the IQC implementation rate rose from 97.79% to 99.87% (P<.001), the report error rate decreased from 0.048% to 0.027% (P<.001), and the intralaboratory turnaround time compliance rate increased from 95.49% to 95.71% (P<.001). The annual EQA unacceptable rate dropped from

XSL•FO
RenderX

0.34% in 2023 to 0% in 2024. The number of ISO (International Organization for Standardization) 15189–accredited test items increased from 203 in 2022 to 206 in 2024.

Conclusions: The iLab system establishes a scalable framework for standardized IQC and EQA management in RMLC, demonstrating significant potential to enhance regional laboratory quality.

(JMIR Form Res 2025;9:e77043) doi:10.2196/77043

KEYWORDS

internal quality control; external quality assessment; quality management; laboratory information system; regional medical laboratory center

Introduction

The Luohu Hospital Group, comprising 5 district hospitals and 43 community health stations, serves as a model for integrated urban health care systems in China. China's National Health Commission endorsed the Luohu model as the template for building a people-centered integrated care model in China [1,2]. Within this framework, the Regional Medical Laboratory Center (RMLC) consolidates laboratory departments from 5 hospitals and 43 community health stations. Meanwhile, the central laboratory handles more complex tests with longer turnaround times, while the branch laboratories manage emergency tests and those requiring shorter turnaround times. This model seeks to optimize intraregional resource sharing, standardize testing quality, unify service standards, and promote mutual recognition of examination outcomes, thereby enhancing the overall efficiency and reliability of health care services.

Despite its many benefits, several challenges persist. Disparities in laboratory quality, differences in staff training, and the geographic spread of primary laboratories make it difficult to ensure the consistency and reliability of test results across the region [3]. Given the complexity and scale of the system, this article proposes a novel information management system designed to address these challenges. By integrating 2 key quality control (QC) strategies—internal quality control (IQC) and external quality assessment (EQA)—the system aims to enhance the consistency and accuracy of laboratory results.

IQC is an essential tool for medical laboratory quality control, continuously monitoring analytical processes to ensure their stability and reliability for patient care [4]. Medical laboratories implement defined IQC procedures, including choosing the control material, setting control limits, analyzing QC charts, detecting random errors, dealing with out-of-control reports, and managing in-house QC data [5]. Standard ISO (International Organization for Standardization) 15189:2022 (section 7.3.7.2) specifies IQC requirements, while the Clinical and Laboratory Standards Institute (CLSI) provides general guidelines for IQC implementation. The day-to-day processes of IQC are often realized through software tools, either standalone applications or those embedded within a laboratory information system (LIS) or laboratory instruments supplied [6]. However, implementing effective IQC within RMLC presents unique challenges. First, RMLC typically includes numerous laboratories with diverse equipment, procedures, and levels of staff expertise. This diversity complicates the standardization of IQC procedures across all facilities. Second, IQC data sharing among these laboratories is hindered by hospital intranet security protocols, creating information silos.

EQA is used to stimulate the harmonization and standardization processes by ensuring laboratory accuracy and monitoring the comparability of a laboratory's testing procedures against external references over time [7]. ISO 15189:2022 (section 7.3.7.3) and CLSI QMS24 outline the requirements essential for the effective management of EQA procedures. Unfortunately, EQA programs can be expensive and time-restricted, making it challenging for basic laboratories to participate [7]. In addition to participating in EQA programs organized by government agencies, laboratories can use interlaboratory comparisons to assess their analytical capabilities and promptly address any potential issues [8-10]. Therefore, there is a need for a regional EQA system to manage the EQA process and carry out routine daily interlaboratory comparisons.

To address these challenges, we developed the iLab informatics platform, an integrated solution for IQC and EQA management within RMLC. By leveraging existing hospital network infrastructure, iLab served as a critical enabler by (1) enforcing standardized IQC and EQA protocols across the network; (2) generating objective evidence through systematic data collection and reporting to ensure compliance with ISO 15189; and (3) establishing automated, traceable workflows that met the documentation mandates of the standard. Deployed successfully in a pilot network of 5 district hospitals, 1 central laboratory, and 4 community health stations, iLab provided the technological infrastructure and standardized processes that were instrumental in achieving ISO 15189 accreditation for 206 testing items. Our findings demonstrate how strategic IT integration can transform quality management in distributed laboratory networks, offering valuable insights for health care systems aiming to standardize testing quality across multiple facilities.

Methods

System Requirements Analysis

The collection and analysis of QC system requirements are a critical stage that shapes the entire project's trajectory. By thoroughly understanding the current issues, inefficiencies, and quality standards, this step defines the specific problems that need to be addressed. After conducting formal and informal interviews with laboratory quality supervisors, laboratory technicians, managers, data managers, and IT experts, the design goal of the QC system was determined. The overall goals are (1) ensuring the system functionality complies with the

XSL•FO
RenderX

management requirements for IQC and EQA as defined by ISO 15189, (2) enabling seamless interconnectivity of QC data, allowing laboratories to access and use data across different institutions, and (3) achieving efficient integration of the QC system with the LIS.

System Architecture and Components

Our IQC and EQA management approach uses 4 software programs LIS, Quality Control Box (QCBOX), Quality Control EQA Center (QCEC), and iLab electronic recording system (hereafter referred to as the iLab system). The LIS, provided by Shanghai Tengcheng Medical Tech-Info Co, Ltd, extracts QC testing data from instruments and interfaces with QCBOX for streamlined management of laboratory QC processes. The QCBOX includes functional modules such as a control dashboard, daily QC monitoring, interlaboratory comparisons, and document management. Meanwhile, the QCEC operates on a browser-server architecture, specifically designed for clinical examination centers and laboratories to efficiently manage EQA programs. The QCBOX and QCEC software, provided by Bio-Rad Laboratories (Shanghai) Co, Ltd. The custom-developed iLab electronic recording system serves as the central platform, integrating data from other components while providing specialized functions for experimental process recording, quality indicator monitoring, and continuous improvement tracking. For further details on the development principles of these software programs, please refer to Multimedia Appendix 1.

Implementation Strategy

Implementation followed a phased approach across the RMLC network, beginning with the central laboratory before extending to 5 district hospitals and 4 community health stations in Luohu District (Figure 1). The deployment process included comprehensive training sessions covering system operations, IQC failure analysis, and QC data management. Technical integration was achieved through application programming interfaces that enable real-time data exchange between the LIS, QCBOX, QCEC, and iLab systems while maintaining strict data security protocols within the hospital network infrastructure.

Figure 1. Overview of quality control process by the information system. EQA: External quality assessment; ERS: Electronic recording system; IQC: Internal quality control;.

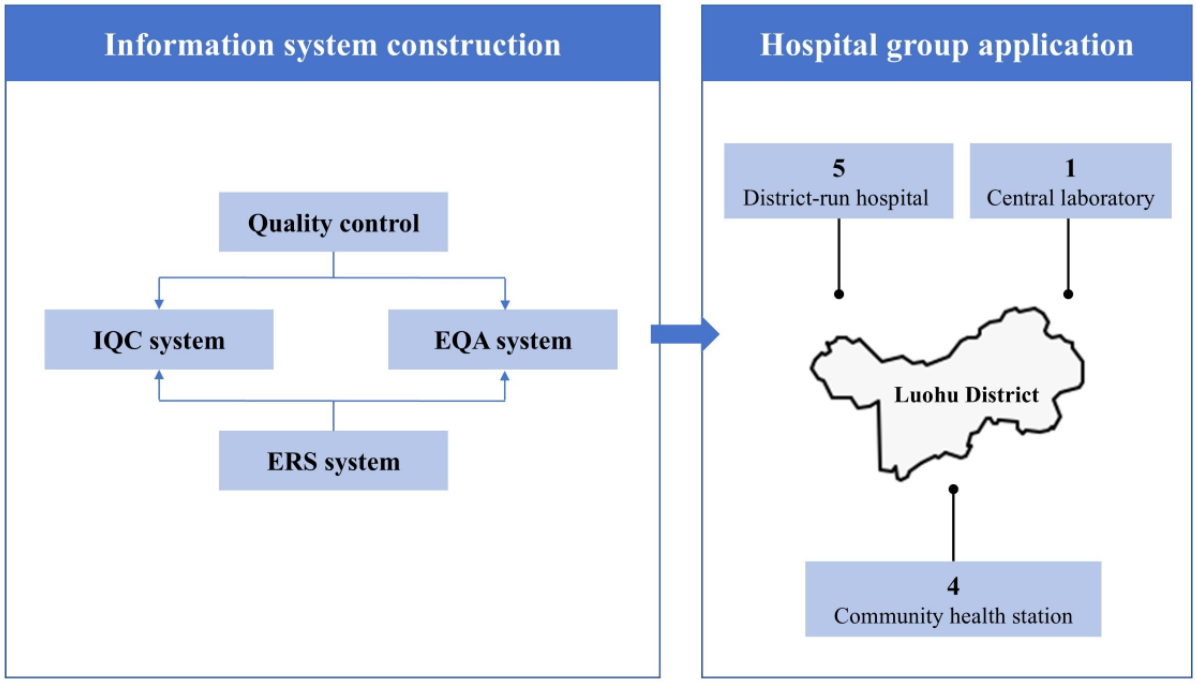

Statistical Analysis

All statistical analyses were performed using R (version 4.0.3; R Core Team). We compared 4 key quality indicators between 2023 (preimplementation) and 2024 (postimplementation): the IQC implementation rate, the annual EQA item unacceptable rate, the report error rate, and the intralaboratory turnaround time compliance rate. The Pearson chi-square test was used to analyze the report error rate and turnaround time compliance rate due to the large sample sizes. Yates continuity correction was applied to the chi-square test for the IQC implementation rate. Fisher exact test was used for the annual EQA unacceptable rate to account for small cell frequencies and zero-event counts. P value<.05 was considered statistically significant.

Ethical Considerations

This study was reviewed and approved by the Institutional Review Board of The Third Affiliated Hospital of Shenzhen University (Approval Number: [2025-LHQRMYY-KYLL-043]). The data were derived from routine quality management activities within the laboratory network and did not involve direct human participants. Consequently, no compensation was provided, and no personally identifiable information is included

XSL•FO
RenderX

in any figures, images, or research materials presented in this study.

Results

Interlaboratory IQC Management Using an Information System

IQC management encompasses 4 critical phases: data generation, data collection, data analysis, and data management. Figure 2 delineates the interconnections among these phases and the software tools instrumental in the execution. LIS serves as the central hub for collecting QC data from instruments across all laboratories and subsequently transmits this data to QCBOX. QCBOX is the software responsible for establishing QC parameters and rules, recording QC results, filing out-of-control reports, etc. The iLab system interfaces with QCBOX to streamline the automation of QC summary report reviews, enhance the monitoring of IQC-related quality indicators, and facilitate the implementation of corrective actions. This integration ensures a seamless workflow that enhances the efficiency and accuracy of the IQC process.

Figure 2. Internal quality control management by the information system. LIS: laboratory information system; QCBOX: Quality Control Box;.

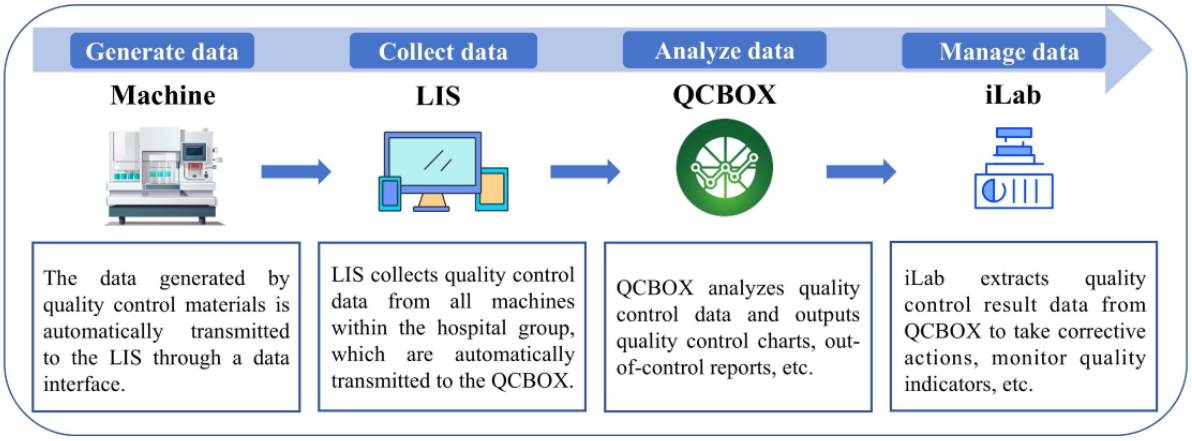

QCBOX automatically archives data following a comprehensive review and finalization of each IQC project. The archived information includes experimental details, QC charts, summary statements, dual review signatures, and other relevant documentation. Furthermore, technicians can access QC data for any time period or from other institutions upon demand. This empowers laboratories to comprehensively assess their overall QC performance and benchmark against external data. The interlaboratory comparison of IQC data module supports the automated retrieval of daily routine IQC data conducted in the laboratory. By comparing the IQC results with those from other regional laboratories using the same methodology and analyzing the same batch of QC materials, the module evaluates the consistency of interlaboratory results using the coefficient of variation ratio (CVR) and standard deviation index (SDI) metrics. This process is elaborated in Case Study 1. The QCBOX dashboard provides a comprehensive overview, displaying metrics such as the IQC implementation rate (percentage of IQC inspection items relative to total inspection items within the same period), the IQC coefficient of variation (CV) failure rate (percentage of inspection items with a CV higher than the requirement relative to the total items with CV requirements for IQC within the same period), and detailed information on out-of-control items for IQC within the region. Leveraging informative graphics and drill-down capabilities, this dashboard enables managers to effectively oversee IQC practices across all regional laboratories.

The iLab system is mainly used for the postanalysis management of IQC data, including four modules: (1) the QC material information record electronically records details such as the name, quantity, batch number, and expiration date; (2) the monthly summary process uses procedural form to extract monthly QC summaries from QCBOX, which undergoes a rigorous multi-stage review and approval; (3) the corrective action process uses a procedural forms to initiate a structured corrective action workflow for the out-of-control test items. The workflow meticulously documents the root cause, the corrective measures implemented, the status of corrective action completion, and subsequent verification outcomes; and (4) quality indicator monitoring, this module continuously monitors the fluctuation of 2 key metrics, IQC implementation rate and IQC CV failure rate. This monitoring is complemented by visual graphics to identify potential risks.

Case Study 1: Interlaboratory Comparisons of Internal Quality Control Data Using QCBOX

The ISO 15189:2022 standard stipulates that interlaboratory comparisons of the results from the examination of identical IQC materials are acceptable alternatives for EQA. Regional testing centers can conduct such interlaboratory comparisons using the same IQC materials at any time through the QCBOX system. This practice holds significant value for promoting the homogenization of test results and the timely detection of potential risks. Our laboratory uses the interlaboratory comparisons feature of QCBOX for routine comparisons. As shown in Multimedia Appendix 2, we compared the consistency of potassium detection results from the central laboratory and Luohu District People's Hospital laboratory using 4 identical instruments of the same brand and model. The laboratory's

XSL•FO
RenderX

results can be compared with the Affiliated Group (a group of laboratories that the Unity Interlaboratory Program groups together to form an ad hoc consensus group), Peer Group (all laboratories using the exact same methodology, instrument, and reagents), and Method Group (all laboratories reporting an analyte using the same methodology code). Relevant terms can be referred to in Bio-Rad's Unity Interlaboratory Program.

Two of the most critical metrics in an interlaboratory program are the CVR and SDI, which are consensus-based metrics for imprecision and bias, respectively [11,12]. In terms of result presentation, the comparison report displays monthly and cumulative statistical data for the laboratory and the consensus groups (Affiliated, Peer, and Method Groups), encompassing parameters such as mean, SD, CV, bias, number of data points, number of laboratories, SDI, and CVR in Table S1 in Multimedia Appendix 2. The distribution of SDI or CVR can be presented through bar charts or pie charts, with automatic assessment of the laboratory testing system's imprecision and bias based on predefined SDI or CVR thresholds [13].

EQA Management Using Information System

EQA serves as a fundamental method for quality control and method validation in RMLC. It ensures that the results obtained from patient investigations are not only accurate and reliable but also comparable across different laboratories. Feedback from EQA results helps identify areas needing attention, such as calibration issues, procedural inconsistencies, or personnel training needs. Laboratories can then implement corrective actions to address these issues. Despite the widespread use of EQA for evaluating the quality of analytical work in medical laboratories, the role of IT in this process often receives limited attention. We therefore developed 2 IT methods to improve the EQA workflow, as depicted in Figure 3. First, we enhanced the conventional EQA processes conducted by the National Center for Clinical Laboratories (NCCL). Second, we used the QCEC software tool to operate EQA within the region, serving as a complementary approach to traditional EQA methods.

Figure 3. Overall workflow of the external quality assessment within the Regional Medical Laboratory Center. EQA: external quality assessment;.

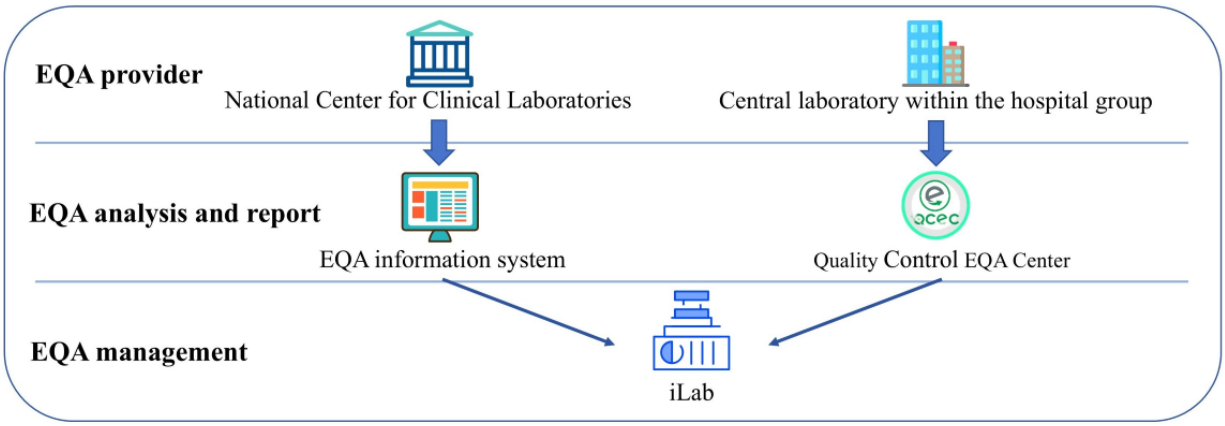

This study primarily used the iLab system to improve the conventional EQA workflow within the laboratory, focusing on elements such as electronic record-keeping, quality indicator monitoring, and detailed data analysis. Several electronic record forms were designed for shared use among regional laboratories. These forms included the EQA plan and implementation record form, interlaboratory comparison record form, interlaboratory quality assessment material receipt and processing record form, and others. Based on data from these records, we established 3 quality indicators interlaboratory, comparison rate for items without an EQA program (the proportion of items that undergo interlaboratory comparison to the total number of items without an EQA program during the same period), the EQA item coverage rate (the proportion of items participating in EQA each year relative to the total number of items with an EQA plan), and the EQA item unacceptable rate (the proportion of items participating in EQA that are deemed unacceptable each year relative to the total number of items that participated in EQA). Data trends were visualized using bar charts and line graphs. For unacceptable EQA items, we used procedural forms to document reasons for nonconformance, corrective actions taken, and verification results. For acceptable EQA items, we applied QC multirules to further analyze the data, identifying potential systematic or random errors, and proposed preventive measures accordingly [14,15]. These processes were detailed in Case Study 2.

The high cost, infrequency, and limited coverage of NCCL-administered EQA programs hindered basic laboratory participation and reduced such programs' effectiveness as quality improvement tools [16]. To address these shortcomings, the central laboratory assumed the role of the NCCL, and QCEC serves as an effective tool for conducting EQA. Using QCEC, the central laboratory assumed the responsibilities of formulating and disseminating the EQA plan, analyzing and evaluating uploaded results and statistics, and providing feedback to division laboratories. Each laboratory was authorized to participate in the central laboratory-issued EQA plan and access the comprehensive result report. The report encompassed laboratory system information relevant to the test data, including methodology, instrumentation, reagents, and calibrators, as well as comparison details such as target values, bias, allowable ranges, evaluation outcomes, and comparative graphical representations.

XSL•FO
RenderX

Case Study 2: Using Quality Control Multirules to Analyze Acceptable EQA Data

Most laboratories focus solely on analyzing unacceptable EQA data to identify root causes and implement corrective actions. However, valuable insights can be gleaned from acceptable EQA data as well. By using appropriate QC rules or rule combinations, laboratories can scrutinize acceptable EQA data to uncover potential systematic or random errors and proactively implement preventive measures (Figure 4A). Different QC rules or combinations of rules have different sensitivities to error, and the QC rules used in this laboratory and their meanings were illustrated in Figure 4B. Without dedicated software, most laboratories face significant challenges in applying QC rules due to the complexities involved in manual calculations. To address this limitation, we have introduced an EQA data multirule scheme within the iLab system.

Figure 4. Application of the quality control multirules module to the analysis of quantitative external quality assessment data. (A) External quality assessment data review process. (B) Quality control multirules description. (C) Quality control multirules analytical process for urea satisfactory external quality assessment data. EQA: external quality assessment; RE: random error; SDI: standard deviation index; TEa: allowable total error;.

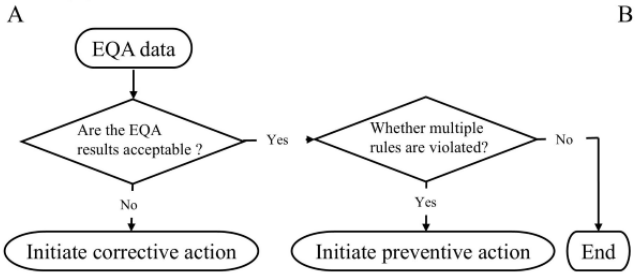

B

Rule name	Explanation	Significance
$\bar{X}_{1.5SDI}$	Mean SDI value of 5 samples exceeding +1.5 or -1.5	SE
R_{4SDI}	Difference between the SDI value of the maximum and minimum results >4.0	RE
$1_{75\%TEa}$	At least one of the 5 samples has a deviation in results of more than 75% of TEa	Uncertain
$5_{\bar{x}}\&1_{75\%TEa}$	5 samples with results on one side of the mean, with 1 result deviating by more than 50% from TEa	SE
$\bar{X}_{1SDI}$	Mean SDI value of 5 samples exceeding +1 or -1	SE
R_{3SDI}	Difference between the SDI value of the maximum and minimum results >3.0	RE

C

Sample number	Measured value	Group mean value	Same group SD	SDI value	|SDI|	Deviation (%)
1	23.7	22.7	1.82	0.55	0.55	4.41
2	13.2	12.41	0.99	0.8	0.8	6.37
3	5.7	5.43	0.43	0.63	0.63	4.97
4	7.6	7.23	0.58	0.64	0.64	5.12
5	15.9	15.51	1.24	0.31	0.31	2.51

Index	Value
Mean |SDI|	0.59
Target TEa	8
75%TEa	6
50%TEa	4

Rules violated	Significance
$1_{75\%TEa}$	Uncertain
$5_{\bar{x}}\&1_{75\%TEa}$	SE

Initiate preventive action

In our laboratory, acceptable EQA results underwent additional multirule analysis incorporating 2 key metrics the SDI, derived from the difference between the measured value and group mean divided by the group SD, and the percentage deviation, calculated as the difference between measured and group mean values divided by the group mean and multiplied by 100%. Laboratory technicians simply import the measured value of the EQA, the group mean value, the same group SD, and the allowable total error (TEa) using the multirule scheme within the iLab system. The system then automatically identifies rule violations.

Figure 4C illustrates the multirule calculation process using a specific case. EQA results for urea from a particular year are presented. Upon importing the relevant data into the iLab system, it is evident that the urea results violated the $1_{75\%TEa}$ and rules . Notably, all test results resided on the same side of the group mean value, suggesting a potential systematic error. Therefore, based on this analysis, preventive measures are necessary to calibrate the instrument and reevaluate the data following retesting of the EQA samples.

An Integrated and Interconnected Information System for Managing IQC and EQA

LIS, QCBOX, QCEC, and the iLab system serve different roles within the laboratory. However, the traditional decentralized management mode often leads to issues such as information silos, resource wastage, and inefficiencies. To address these limitations, we have conducted secondary development to effectively integrate these components into a unified solution, as illustrated in Figure 5. IQC and EQA are distinct yet complementary components ensuring ongoing quality control [17]. The iLab system supports multidimensional in-depth analysis of QC data, generating valuable management information. Particularly, ISO15189:2022 mandates inclusion of measurement uncertainty (MU) calculation in all measurement procedures. We have therefore developed a tool within the iLab system based on IQC and EQA to calculate MU, detailed in Case Study 3 [18,19]. Importantly, personnel in laboratories within each division access the iLab system via authenticated accounts, facilitating interconnected data management.

XSL•FO
RenderX

Figure 5. Schematic illustration of an integrated and interconnected information system for managing internal quality control and external quality assessment. CV: coefficient of variation; EQA: external quality assessment; IQC: internal quality control; LIS: laboratory information system; QC: quality control; QCBOX: Quality Control Box;.

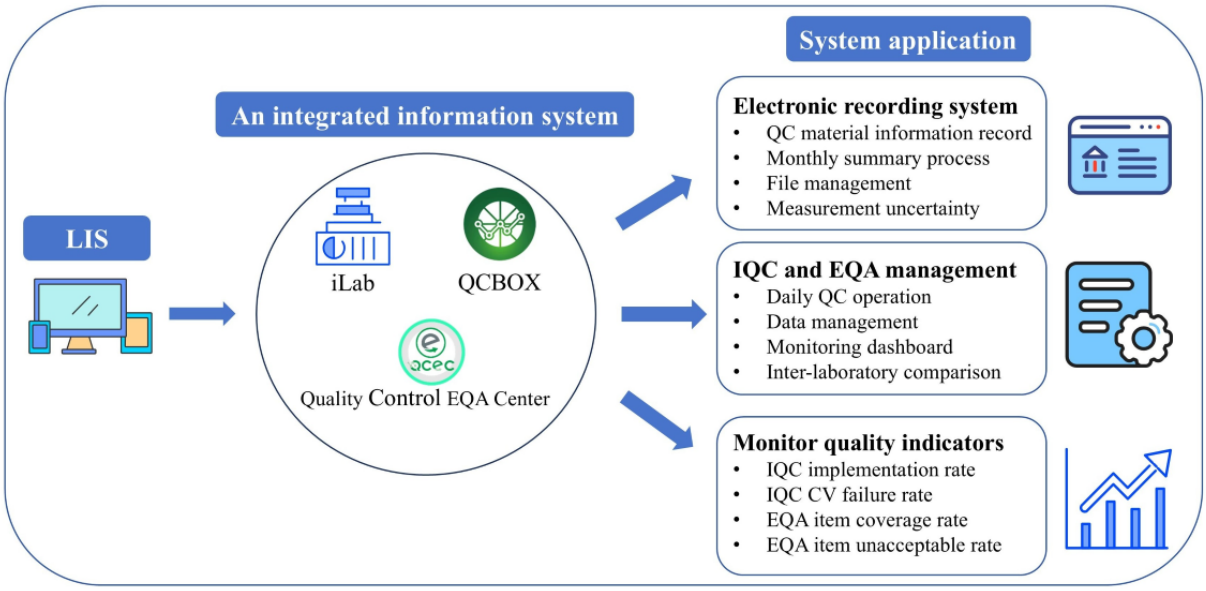

Case Study 3: Using IQC and EQA Data to Calculate MU on the Web Page

The parameter denoted by MU reflects the dispersion associated with a measured value and quantifies the reasonable level of uncertainty inherent in the result. In the context of medical laboratories, MU can be interpreted as the degree of inherent variability or inaccuracy associated with measurements obtained for a specific sample using a particular testing system or protocol. Crucially, a thorough understanding and appropriate application of MU elevate it to a key quality indicator, effectively describing the performance of both the in vitro diagnostics measuring system and the laboratory itself [20]. This empowers clinicians to use medical laboratory measurements more effectively for disease diagnosis and treatment planning.

The complexity of the MU calculation process is considered a significant barrier to its widespread adoption [21]. To efficiently evaluate the uncertainty of test items, our laboratory has developed an automatic MU evaluation function in the iLab system. As shown in Figure 6, this function includes 2 pages, the experimental information entry page (Figure 6A) and the result report page (Figure 6B).

XSL•FO
RenderX

Figure 6. Using internal quality control and external quality assessment data to calculate measurement uncertainty on the webpage. (A) Experimental basic information entry. (B) Measurement uncertainty assessment results report. EQA: external quality assessment; IQC: internal quality control; MU: measurement uncertainty; QCBOX: Quality Control Box; RBP: retinol binding protein;.

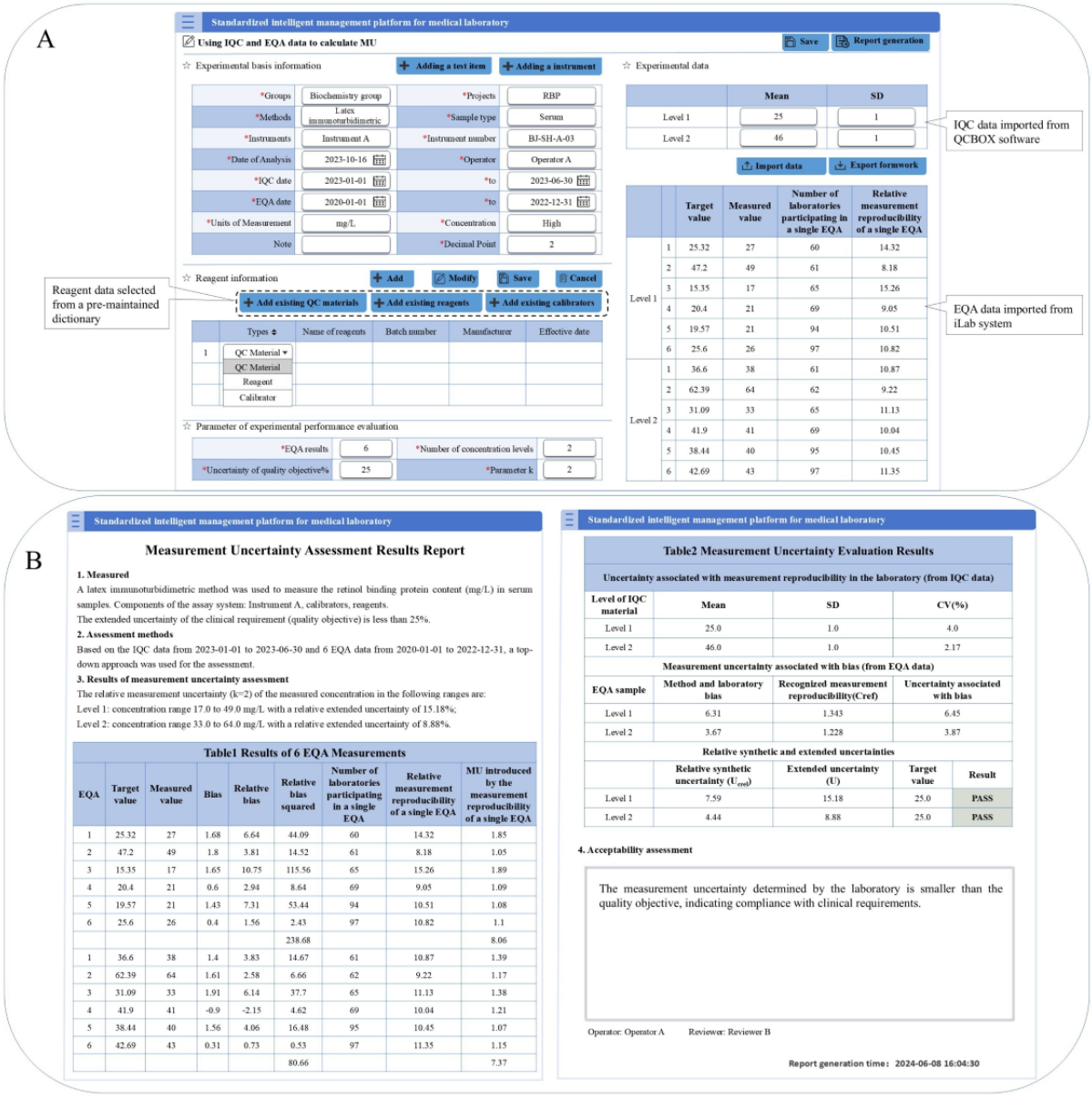

The experimental information entry page comprises several sections:

- Basic laboratory information maintenance: this includes the entry of items to be evaluated, methodologies, instruments, and the statistical period for IQC and EQA data. It is important to note that the types of instruments and test items are maintained in a predefined dictionary, allowing for direct selection from this dictionary during entry.
- Reagent information: this involves the entry of reagent names, catalog numbers, expiration dates, etc. The types of reagents can also be selected from a premaintained dictionary.
- Parameter of experimental performance evaluation: this section includes the number of concentration levels, uncertainty of quality objective, parameter k, and other relevant parameters.
- Experimental data: this section allows for the import or manual entry of IQC and EQA data.

By inputting the appropriate data into the system, an evaluation report can be generated with one click, eliminating the need for manual calculation and significantly reducing the laboratory's workload in MU evaluation. As shown in Figure 6B, the result report page includes a summary of experimental information, EQA statistical results, MU statistical results, and the evaluation conclusion.

Taking retinol binding protein (RBP) as an example, 2 concentrations of IQC data from 2023-01-01 to 2023-06-30 were taken to assess the MU introduced by the reproducibility of laboratory measurements (Figure 6). A total of 6 EQA results from January 1, 2020 to December 31, 2022 were used to assess

XSL•FO
RenderX

the MU associated with the deviation. The mean values of the 2 concentrations of RBP for the IQC were 25 mg/L and 46 mg/L. The final results indicate that the MU at 2 concentration levels meets the quality objective.

In clinical applications, MU serves 2 primary purposes. First, it assists clinicians in decision-making when test results are close to clinical decision thresholds. For example, considering RBP with a reference range of 25-70 mg/L, when the test result is 28 mg/L, the relative expanded uncertainty at Level 1 is 15.18% (Figure 6B), yielding a 95% CI of 23.75-32.25 mg/L. In this situation, pathological factors cannot be directly excluded, and the result should be interpreted in conjunction with the patient's symptoms and medical history. Second, MU is crucial for evaluating serial measurements, enabling clinicians to distinguish true biological trends from random analytical variation. For instance, if a patient's consecutive test results are 71 mg/L and 68 mg/L, with a relative synthetic uncertainty of 4.44% at Level 2 (Figure 6B), the minimum difference threshold at the 95% confidence level is 8.73 mg/L. Since the observed difference of 3 mg/L is below this threshold, clinicians can infer that the variation is attributable to analytical imprecision rather than an actual change in the patient's condition.

Evaluation of System Implementation Effectiveness

The iLab platform achieved rapid, large-scale implementation, resulting in significant measurable gains in operational efficiency and quality performance across the RMLC network. Following its official launch in August 2023, the system integrated 133 QC instruments and 859 active QC items under 114 registered users by September 2025. This deployment was essential in establishing formalized electronic IQC, EQA, and quality management frameworks in basic-level laboratories, where such infrastructure had previously been lacking. The transition from manual to automated processes streamlined workflows and reduced staff workload. This efficiency gain is evidenced by the system's output during this period: generating 2331 monthly IQC analysis reports, 289 MU calculation reports, and 340 qualified EQA multirule analysis reports, identifying 93 rule violations. Automation not only streamlined workflows but also ensured comprehensive digital documentation and full traceability for all quality management actions, further supported by more than 20 training sessions conducted on system operation and IQC result analysis.

Quantitative analysis of the 5 ISO15189-accredited laboratories demonstrated sustained improvements in quality performance metrics following system implementation. The IQC implementation rate significantly increased from 97.79% in 2023 to 99.87% in 2024 ($P<.001$). Operational efficiency and accuracy also showed marked gains. The report error rate declined from 0.048% to 0.027% ($P<.001$), while the intralaboratory turnaround time compliance rate improved from 95.49% to 95.71% ($P<.001$). The annual EQA item unacceptable rate decreased from 0.34% to 0%, though this reduction was not statistically significant. Additionally, the scope of accredited services expanded, with the number of testing items accredited under the ISO 15189 standard increasing from 203 in 2022 to 206 in 2024. These comprehensive results conclusively demonstrate the platform's critical role in achieving centralized standardization, ensuring reliable results, and expanding high-quality services across the regional network.

Discussion

Principal Findings

This study demonstrates that the iLab informatics platform effectively addresses quality heterogeneity across the RMLC by integrating LIS, QCBOX, QCEC, and iLab into a unified management system. The platform establishes standardized workflows for IQC and EQA, enabling centralized data aggregation, automated interlaboratory comparisons, and streamlined quality governance. Key achievements include the implementation of electronic IQC and EQA systems in basic-level laboratories, enhanced EQA efficiency through multirule analysis and in-house assessment capabilities, and simplified measurement uncertainty calculation via an automated template. By ensuring full traceability while reducing administrative burden, the system provides a scalable and transferable model for standardizing quality management across distributed laboratory networks.

Comparison to Prior Work

Despite the recognized benefits of IT in health care collaboration and resource integration, primary health care institutions often lack sufficient IT adoption [22]. The Luohu model aims to bridge this gap by fostering collaboration among health care institutions. Studies have shown that a strong informationization level is a key factor in improving the performance of the county medical community system [23]. IQC and EQA constitute the foundational pillars of QC in every clinical laboratory [4,24,25]. While various IQC implementation methods exist, including IQC modules embedded in LIS, commercial IQC software, and instrument-specific IQC software, these approaches have resulted in significant variability in practices across different laboratories [6,26-28]. For instance, Tanpaiboon et al [29], recently introduced a customized QC program that allows for updates to meet specific and evolving needs. However, their solution is limited to a biochemical genetics laboratory and does not address the complexities of RMLC. Similarly, Giannoli et al [30] used Levey-Jennings control charts with peer-group statistics to detect out-of-control errors when managing multiple instruments measuring the same analyte, a common scenario in RMLC. In contrast, we have integrated a functional template within our QCBOX software to enable interlaboratory comparisons for IQC, automatically selecting participants using the same QC materials and producing evaluation reports inclusive of the SDI and CVR values. This approach reduces both technical and operational barriers, enabling the timely detection of out-of-control instruments.

EQA schemes are a tool for providing participating laboratories with recognition and proof of meeting quality benchmarks [31,32]. However, their value is realized only when feedback drives laboratory improvement. Current research predominantly focuses on analyzing unacceptable EQA data, with limited attention to the mining and management of in-laboratory EQA datasets [33]. In this study, we addressed three key EQA challenges: (1) streamlining traditional EQA workflows via the iLab system to boost efficiency, (2) implementing multirule

XSL•FO
RenderX

analysis of acceptable EQA data for automated detection of systematic or random errors, and (3) enabling in-house EQA organization within the RMLC using QCEC software, allowing flexible, autonomous proficiency testing.

While the routine reporting of MU with patient test results is not recommended by MU guidelines, it should be readily accessible upon request [34]. Recent guidelines and studies have highlighted the significance of MU estimation in medical laboratories, and the usage of IQC and EQA data for MU calculation has gained recognition as a reliable statistical method [19,34,35]. This study presents a user-friendly function template for computing MU using IQC and EQA data. This template simplifies complex calculations for technicians, enabling them to obtain results with minimal effort by simply inputting experimental data. Although the clinical applications described highlight the immense theoretical utility of MU, we acknowledge that its current primary function is to serve laboratory quality governance and compliance. Its routine adoption by frontline clinicians requires significant effort beyond mere technological support and crucially, relies on clinicians to fully understand the clinical value of MU and proactively demand this information from the laboratory.

System Scalability and Generalizability

Our system is designed with generalizability as an important feature, and its fundamental framework is theoretically transferable to a wide range of clinical laboratories, including those in less-resourced or rural settings. However, we acknowledge that successful implementation requires three categories of basic resources that are standard in clinical laboratories and not limited to large urban networks. First, appropriate hardware infrastructure compatible with the software requirements is needed. The system operates on standard computing hardware consistent with the baseline infrastructure of most clinical laboratories, including those in resource-limited settings, without necessitating specialized high-performance computing equipment. Second, sustainable operational costs are necessary, as economic feasibility depends on covering routine expenses such as software licensing and periodic hardware maintenance. These costs are comparable to those of established laboratory workflows and can be easily incorporated into standard operational budgets. Third, personnel competence in protocol implementation is required, as effective deployment relies on laboratory staff possessing basic technical proficiency in operating the software interface and adhering to the standardized protocols described in our manuscript. This can be achieved through targeted training programs that do not demand advanced specialized expertise, ensuring adaptability across varying skill levels. In summary, although our system is inherently generalizable, its practical application presupposes access to these foundational resources.

Limitations

This study has several limitations. First, while the system demonstrated improvements in analytical quality metrics, its downstream effects on clinical decision-making, turnaround time for critical results, or patient outcomes were not directly evaluated. Second, substantial training and workflow adaptation were required for integration, potentially hindering adoption in laboratories with limited technical support. Third, as the remaining community health stations have not yet formally adopted the system, unforeseen issues may emerge during broader deployment that were not identified in the current pilot phase.

Conclusions

This study demonstrates that while RMLC are essential for standardizing laboratory testing, ensuring consistent result quality across diverse laboratories remains a challenge. To address this, we developed and implemented the iLab informatics platform, an integrated solution for managing IQC and EQA. The successful pilot deployment shows that our approach not only standardizes processes and supports continuous quality improvement but also provides a scalable and adaptable model, applicable to other health care networks seeking to enhance laboratory outcome consistency and integrity.

Acknowledgments

We are grateful to the Shenzhen Huikang Information Technology Co, Ltd for their technical support during software development. During the preparation of this work, the authors used generative AI tools for language polishing and refinement. After using these tools, the authors thoroughly reviewed and edited the content as necessary and take full responsibility for all aspects of the published work.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Funding

This work was supported by the Scientific Research Foundation of Shenzhen Health Economics Association (2024204); and the Shenzhen Medical Key Discipline Construction (SZXK054).

Authors' Contributions

XZ and TO contributed to the study conception and design. DC, HM, and YZ developed the methodology. CJ and FL performed data curation and software. DC, HM, YZ, JQ, and LZ performed the analysis and interpretation. DC, YZ, and TO drafted and revised the manuscript. XZ and TO supervised the study. All authors read and approved the final manuscript.

XSL•FO
RenderX

XZ is a co-corresponding author on this paper and can be reached at zxm0760@163.com.

Conflicts of Interest

CJ and FL are employees of Bio-Rad Laboratories (Shanghai) Co, Ltd. Their involvement in this study was strictly limited to providing technical support for data analysis and general guidance on the use of the QCBOX and QCEC systems. They did not participate in study design, data processing, statistical analysis, result interpretation, or manuscript preparation. The other authors declare no competing financial or personal relationships that could influence the work reported herein.

Multimedia Appendix 1
Technical architecture overview of the four software systems (LIS, QCBOX, QCEC, and iLab).
[DOCX File , 13 KB - formative_v9i1e77043_app1.docx]

Multimedia Appendix 2
The interlaboratory comparison results using QCBOX.
[DOCX File , 24 KB - formative_v9i1e77043_app2.docx]

References

1. Liang D, Mei L, Chen Y, Zhou P, Yang X, Huang J. Building a people-centred integrated care model in urban China: a qualitative study of the health reform in Luohu. Int J Integr Care 2020;20(1):9 [FREE Full text] [doi: 10.5334/ijic.4673] [Medline: 32210740]
2. Wang X, Sun X, Gong F, Huang Y, Chen L, Zhang Y, et al. The Luohu model: a template for integrated urban healthcare systems in China. Int J Integr Care 2018;18(4):3 [FREE Full text] [doi: 10.5334/ijic.3955] [Medline: 30483036]
3. Tadeu BTM, Geelhoed D. "This thing of testing our blood is really very important": a qualitative study of primary care laboratory services in Tete Province, Mozambique. Int J Equity Health 2016;15(1):133 [FREE Full text] [doi: 10.1186/s12939-016-0418-5] [Medline: 27558247]
4. Ricós C, Fernandez-Calle P, Perich C, Westgard JO. Internal quality control - past, present and future trends. Adv Lab Med 2022;3(3):243-262 [FREE Full text] [doi: 10.1515/almed-2022-0029] [Medline: 37362142]
5. Giannoli JM, Albarede S, Avellan T, Bouilloux JP, Cartier R, Cohen R, et al. Recommendations for the application and follow-up of quality controls in medical laboratories. Biochem Med (Zagreb) 2021;31(2):020501 [FREE Full text] [doi: 10.11613/BM.2021.020501] [Medline: 33927549]
6. Shang X, Zhao F, Du J, Li R, Xia L, Hou L, et al. Development and clinical evaluation of an online automated quality control system for improving laboratory quality management. Clin Chim Acta 2023;541:117240. [doi: 10.1016/j.cca.2023.117240] [Medline: 36736683]
7. Sciacovelli L, Secchiero S, Padoan A, Plebani M. External quality assessment programs in the context of ISO 15189 accreditation. Clin Chem Lab Med 2018;56(10):1644-1654 [FREE Full text] [doi: 10.1515/cclm-2017-1179] [Medline: 29794248]
8. Preiksaitis J, Pang X, Fox J, Fenton J, Caliendo A, Miller G, American Society of Transplantation Infectious Diseases Community of Practice. Interlaboratory comparison of epstein-barr virus viral load assays. Am J Transplant 2009;9(2):269-279 [FREE Full text] [doi: 10.1111/j.1600-6143.2008.02514.x] [Medline: 19178414]
9. Quartey P, Anokye R, Frimpong M. Inter-laboratory comparability of clinical chemistry testing: a new perspective. Int J Sci Basic Appl Res 2019;48(5):196-207. [doi: 10.34297/AJBSR.2020.08.001235]
10. Payne D, Johansson U, Bloxham D, Couzens S, Carter A, Holtom P, et al. Inter-laboratory validation of a harmonized PNH flow cytometry assay. Cytometry B Clin Cytom 2018;94(5):580-587 [FREE Full text] [doi: 10.1002/cyto.b.21726] [Medline: 30240160]
11. Pant V, Pradhan S, Gautam K. Basics of laboratory statistics. EJIFCC 2023;34(2):90-102 [FREE Full text] [Medline: 37455839]
12. Park H, Cho J, Lee SM, Son J, Kim SR, Yoo S, et al. Development and application of a laboratory-developed quality control program for blood glucose monitoring systems: a single institute experience. Ann Lab Med 2017;37(4):362-364 [FREE Full text] [doi: 10.3343/alm.2017.37.4.362] [Medline: 28445023]
13. Niraula A, Bataju M. External quality assessment practices in clinical biochemistry laboratory: what is the need? Kathmandu Univ Med J 2020;18(69):86-92. [Medline: 33582695]
14. Cembrowski GS, Hackney JR, Carey N. The detection of problem analytes in a single proficiency test challenge in the absence of the health care financing administration rule violations. Arch Pathol Lab Med 1993 Apr;117(4):437-443. [Medline: 8466413]
15. Carey RN, Cembrowski GS, Garber CC, Zaki Z. Performance characteristics of several rules for self-interpretation of proficiency testing data. Arch Pathol Lab Med 2005;129(8):997-1003 [FREE Full text] [doi: 10.5858/2005-129-997-PCOSRF] [Medline: 16048412]

XSL•FO
RenderX

16. Cherie N, Deress T, Berta D, Chane E, Teketelew B, Adane K, et al. Navigating quality assessment hurdles in clinical laboratory services: a comprehensive review in resource-limited settings. Risk Manag Healthc Policy 2024;17:497-504 [FREE Full text] [doi: 10.2147/RMHP.S453020] [Medline: 38476199]
17. Berwouts S, Morris MA, Dequeker E. Approaches to quality management and accreditation in a genetic testing laboratory. Eur J Hum Genet 2010;18 Suppl 1(Suppl 1):S1-19 [FREE Full text] [doi: 10.1038/ejhg.2010.104] [Medline: 20720559]
18. Dannus L, Mulliez A, Berger M, Bourgne C, Veyrat-Masson R. Applicability of the long-term uncertainty in measurement (LTUM) method for analytical performance assessment in clinical cytometry laboratories. Cytometry B Clin Cytom 2022;102(3):254-260 [FREE Full text] [doi: 10.1002/cyto.b.22050] [Medline: 34913579]
19. Padoan A, Antonelli G, Aita A, Sciacovelli L, Plebani M. An approach for estimating measurement uncertainty in medical laboratories using data from long-term quality control and external quality assessment schemes. Clin Chem Lab Med 2017;55(11):1696-1701. [doi: 10.1515/cclm-2016-0896] [Medline: 28245184]
20. Braga F, Panteghini M. The utility of measurement uncertainty in medical laboratories. Clin Chem Lab Med 2020;58(9):1407-1413 [FREE Full text] [doi: 10.1515/cclm-2019-1336] [Medline: 32126011]
21. Martinello F, Snoj N, Skitek M, Jerin A. The top-down approach to measurement uncertainty: which formula should we use in laboratory medicine? Biochem Med (Zagreb) 2020;30(2):020101 [FREE Full text] [doi: 10.11613/BM.2020.020101] [Medline: 32292278]
22. Dong C, Liu J, Mi J. Information-driven integrated healthcare: an analysis of the cooperation strategy of county medical community based on multi-subject simulation. Healthcare (Basel) 2023;11(14):2019 [FREE Full text] [doi: 10.3390/healthcare11142019] [Medline: 37510460]
23. Sun X, Li Y, Liu S, Lou J, Ding Y, Liang H, et al. Enhanced performance of community health service centers during medical reforms in Pudong new district of Shanghai, China: a longitudinal survey. PLoS One 2015;10(5):e0125469 [FREE Full text] [doi: 10.1371/journal.pone.0125469] [Medline: 25950172]
24. Plebani M. Internal quality control and external quality assurance: a great past opens the way to a bright future. Adv Lab Med 2022;3(3):215-220 [FREE Full text] [doi: 10.1515/almed-2022-0075] [Medline: 37362140]
25. Ricós C, Fernández-Calle P, Perich C, Sandberg S. External quality control in laboratory medicine. Progresses and future. Adv Lab Med 2022;3(3):221-242 [FREE Full text] [doi: 10.1515/almed-2022-0058] [Medline: 37362144]
26. Newton DW, Vandepoele N, Yundt-Pacheco JC, Gauthier D, Bachman MA. Management of cytomegalovirus, epstein-barr virus, and HIV viral load quality control data using unity real time. J Clin Microbiol 2022;60(1):e0167521 [FREE Full text] [doi: 10.1128/JCM.01675-21] [Medline: 34669458]
27. Quinton M, Newton DW, Neil B, Mitchell S, Mostafa HH. Quality control data management with unity real-time in molecular virology. J Clin Virol 2024;171:105655 [FREE Full text] [doi: 10.1016/j.jcv.2024.105655] [Medline: 38367294]
28. Kinns H, Pitkin S, Housley D, Freedman DB. Internal quality control: best practice. J Clin Pathol 2013;66(12):1027-1032. [doi: 10.1136/jclinpath-2013-201661] [Medline: 24072731]
29. Tanpaiboon P, Salazar D, Pan M, Xu L, Sharma R. Development of an automated quality control and assay performance management system for biochemical genetics laboratory. Mol Genet Metab Rep 2025;44:101247 [FREE Full text] [doi: 10.1016/j.ymgmr.2025.101247] [Medline: 40893138]
30. Giannoli JM, Bernard M, L'Hirondel J, Heim A, Badrick T. A model for managing quality control for a network of clinical chemistry instruments measuring the same analyte. Clin Chem Lab Med 2024;62(5):853-860. [doi: 10.1515/cclm-2023-0965] [Medline: 37999926]
31. Krleza JL, Celap I, Tanaskovic JV. External quality assessment in Croatia: problems, challenges, and specific circumstances. Biochem Med (Zagreb) 2017;27(1):86-92 [FREE Full text] [doi: 10.11613/BM.2017.011] [Medline: 28392730]
32. Laudus N, Nijs L, Nauwelaers I, Dequeker E. The significance of external quality assessment schemes for molecular testing in clinical laboratories. Cancers (Basel) 2022;14(15):3686 [FREE Full text] [doi: 10.3390/cancers14153686] [Medline: 35954349]
33. Kristensen GBB, Meijer P. Interpretation of EQA results and EQA-based trouble shooting. Biochem Med (Zagreb) 2017;27(1):49-62 [FREE Full text] [doi: 10.11613/BM.2017.007] [Medline: 28392726]
34. ISO/TS 20914:2019 Medical laboratories-Practical guidance for the estimation of measurement uncertainty. International Organization for Standardization. URL: https://ab.iacld.com/UpFiles/ABIacld/Documents/4c58d3f2-5f80-45bc-993a-e5e37998eea7.pdf [accessed 2025-11-22]
35. Hemanth T, Renuka P, Vinodhini V. The use of internal and external quality control data for the calculation of measurement uncertainty of glycated hemoglobin. Biomed Pharmacol J 2023;16(4):2371-2375 [FREE Full text] [doi: 10.13005/bpj/2812]

Abbreviations

CV: Coefficient of Variation
CLSI: Clinical and Laboratory Standards Institute
CVR: Coefficient of Variation Ratio
EQA: external quality assessment
LIS: laboratory information system

XSL•FO
RenderX

IQC: internal quality control
ISO: International Organization for Standardization
MU: measurement uncertainty
NCCL: National Center for Clinical Laboratories
QC: quality control
QCBOX: Quality Control Box
QCEC: Quality Control EQA Center
RBP: retinol binding protein
RMLC: Regional Medical Laboratory Center
SDI: Standard Deviation Index
TEa: allowable total error

Edited by J Sarvestan; submitted 06.05.25; peer-reviewed by C Buchta, VC Mishra; comments to author 14.10.25; revised version received 19.11.25; accepted 20.11.25; published 09.12.25.

Please cite as:
Chen D, Mo H, Zhang Y, Qin J, Zhang L, Jiang C, Li F, Zhang X, Ou T
Development and Implementation of an Internal Quality Control and External Quality Assessment Information System for a Regional Medical Laboratory Center: Pilot Design and Implementation Study
JMIR Form Res 2025;9:e77043
URL: https://formative.jmir.org/2025/1/e77043
doi:10.2196/77043
PMID:

©Dayang Chen, Hongmei Mo, Yingying Zhang, Junlong Qin, Lijun Zhang, Changxian Jiang, Fuming Li, Xiuming Zhang, Tong Ou. Originally published in JMIR Formative Research (https://formative.jmir.org), 09.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Access to Specialized Medical Training in Spain and Determinants of Failure in the National Entrance Examination: Econometric Modeling Study

Montserrat Diaz-Fernandez[1*], PhD; Mar Llorente-Marron[1*], PhD; Victor Asensi[2*], MD, PhD

[1]Department of Quantitative Economics, University of Oviedo, Oviedo, Spain

[2]Department of Medicine, University of Oviedo, Oviedo, Spain

[*]all authors contributed equally

Corresponding Author:
Montserrat Diaz-Fernandez, PhD
Department of Quantitative Economics, University of Oviedo
Avda del Cristo s/n
Oviedo, 33006
Spain
Phone: 34 985103000 ext 3749
Email: mdiaz@uniovi.es

Abstract

Background: The process of accessing specialized medical training in Spain is a complex issue, involving not only the evaluation of medical knowledge acquired throughout university training but also the interaction of factors of a contextual and structural nature, which can influence the results obtained in the entrance examination. In this context, research on the variables that determine performance in this test is of great relevance form not only an academic but also a social and economic point of view. The interaction among factors such as academic performance, gender, nationality, and timing offers a unique opportunity to evaluate the functioning of the educational system at a critical moment in its recent history. Research that has focused specifically on access to specialized medical training has shown mixed results on how these factors impact examination performance.

Objective: This study aimed to approximate the factors that determine failure in the entrance test for specialized medical training in Spain with the aim of better understanding the extent to which differences based on sex, nationality, and the context of the COVID-19 pandemic contribute to explaining such failure.

Methods: We carried out econometric modeling of the final results obtained in the entrance examination to specialized medical training and identified the explanatory factors that determine the results, their relevance, effect, and significance. Econometric modeling provides a rigorous framework for estimating the causal effect of different variables on the final examination score. It helps identify not only which variables have an impact on performance but also to what extent they do so and under what conditions.

Results: Based on the results obtained in the 2019-2021 test calls (7217 eliminated candidates), academic records ($P<.001$) and examination scores ($P<.001$), together with demographic factors including sex ($P=.54$) and nationality ($P<.001$), and calendar year ($P<.001$) were determinants of the behavior observed in the final results. Our results do not indicate whether being male or female favors or decreases the final grade obtained; however, being Spanish constitutes a relevant explanatory factor in our final results. The calendar effect, directly related to the COVID-19 pandemic, allows us to quantify the negative impact on the final results.

Conclusions: This study investigated the impact of factors such as sex, nationality, and the COVID-19 pandemic on access to specialized medical training in Spain. Empirically, not being Spanish acts as an unfavorable fixed characteristic in the baseline econometric model, but it becomes favorable when considering the candidate's academic record. The impact of language is not perceived as a limiting factor; the test exclusively evaluates knowledge of medical content. The negative effects of the COVID-19 pandemic are visualized in the final scores.

(JMIR Form Res 2025;9:e72440) doi:10.2196/72440

XSL•FO
RenderX

KEYWORDS

medical profession; specialized medical training; access; gender; nationality; academic scale; COVID-19 pandemic

Introduction

The world's population is growing, and life expectancy is increasing. This implies a greater consumption of health services, which is essential for guaranteeing social well-being. However, recourses are limited. In this context, adequate planning for training of health professionals to meet the current and future needs of the population is unavoidable. In light of this, health care resource planning studies have been conducted [1,2].

The Spanish health care system has serious problems due to high levels of life expectancy and medical insufficiency. As a result, the increase in life expectancy, population growth, and the number of people in confinement together contribute to the increase in the demand for health services [3]. Spain has experienced a rapid decline in birth rate and population growth, which has caused a significant demographic change that has affected the country's present and future. In 2022, life expectancy in Spain was 3 years longer than the average within the European Union (EU). Currently, 20% of the population of Spain is older than 65 years.

Health spending in Spain has grown in recent years, although it is lower than the EU average. In 2019, Spain dedicated 9.1% of its gross domestic product to health spending, that is, 0.8 percentage points less than the EU average, with a per-capita spending of €2488 (US $ 2885.15), which is 30% lower than the EU average (€3523 [US $ 4085.46]) [4]. The number of doctors per inhabitant in Spain (4.33 doctors per 1000 inhabitants) exceeds the average (3.6 doctors per 1000 inhabitants) in OECD (Organisation for Economic Co-operation and Development) countries. However, doctor shortages and supply rigidities remain key problems. In this debate, the difficulties in reconciling family life in a female-dominant sector play a relevant role, as does the impact of the mass retirement of many professionals and the expatriation of talent to other countries.

The medical profession has experienced significant social change in recent decades, with a greater presence of women at all levels and specialties. This feminization process has occurred in most high-income countries and is expected to continue in the future [5-9]. In Spain, while 2 out of every 3 doctors who retire are men, 2 out of every 3 physicians who join are women. Currently, 7 out of 10 students enrolled in medical schools are women, and a similar metric is observed for resident physicians (6 out of every 10 residents are women) [10]. However, the aging of the health care workforce and the flight of talent to other countries can threaten the quality and viability of Spain's health care system [3]. The average age of active physicians in 2021 was 49.76 years, with a bimodal distribution of those in the age groups of 45-64 years (37.75%) and 35-54 years (38.78%). The perception of professional opportunities influences the mobility of physicians, who may feel demotivated by their current situation or attracted by other options [11]. The outflow of physicians from Spain to other countries in search of better working conditions, salaries, and professional recognition has continued in recent years, despite the COVID-19 pandemic. According to data from the Collegiate Medical Organization, 2504 certificates of suitability for work abroad were issued in 2021, which was similar to the number issued in 2020. Most of the chosen destinations are EU countries where better options are offered, especially with regard to working conditions and compensation. On the other hand, the arrival of foreign doctors in Spain is related to the economic cycle of public health spending, which has suffered cuts during the economic crisis [11,12].

Public health constitutes a fundamental aspect of the development and quality of life of any country. Different indicators can be used for evaluation (patients' rights, waiting lists, prevention programs, free access, etc). The universal right to health in Spain is guaranteed through the National Health System (NHS), which covers 96.5% of health care services. The NHS offers free, quality services to all citizens throughout their lives and contributes to the state of well-being. Spain stands out as one of the countries with the best health systems worldwide.

Technological capacity and highly qualified human capital constitute strengths of the NHS. The implementation and official recognition of the medical specialization system in Spain through the Medical Intern Residency (MIR) program as a means of accessing the specialist title has led to decisive advances in Specialized Health Training (SHT) tests, which serve as a mark of prestige and recognition for NHS professionals [13]. The SHT based on the residency system consists of a programmed, supervised, and paid clinical practice that pursues the progressive assumption of responsibilities throughout the training period. The SHT aims to train health professionals in the knowledge, techniques, and skills of each specialty, guaranteeing comprehensive and efficient care for users. The assignment of candidates to MIR placements by specialty and hospital center is based on an ordered election based on the results of a national examination with multiple-choice questions that are determined based on the principles of merit and ability and the academic scale of the degree in medicine [14,15]. The choice of specialty is therefore based on the principle of vertical equity by allowing the most productive candidates to have priority in the choice of the SHT program using the ranking position as a proxy variable. Certification examinations play a fundamental role in terms of social responsibility, as they provide objective and transparent criteria for assessing candidates' performance [16-18].

Various studies have shown that factors such as sex, nationality, or the COVID-19 pandemic have an impact on the final results of participants who pass the SHT access test in Spain [7,19]. However, not everyone passes the test; approximately 1 in 5 participants are eliminated from the process during the access attempt. However, the literature does not include the profile of eliminated applicants who do not exceed the access threshold. The shortage of human and material resources in some areas of

XSL•FO
RenderX

the country, labeled medical deserts, makes access to quality health care appropriate to the needs of the population difficult. Health planning is a complex process that requires evaluating health needs, establishing objectives and strategies to improve health, and allocating health resources efficiently and equitably. Its purpose is to offer solutions to health problems that affect people through actions to prevent, treat, control, or eliminate diseases. To do this, not only conventional health services but also the social, economic, and material aspects that influence health must be considered [20]. Studying whether there are significant differences among the factors that determine success or failure in access to the SHT can be useful in the context of a worrying shortage of health professionals.

In this study, we aim to examine the factors that determine the performance of candidates in terms of access to the SHT, which is a demanding test that provides access to specialized training in health care in Spain. However, not everyone manages to overcome it, leaving some outside the selection process. What factors influence success or failure? Are there differences based on sex or nationality? Has the COVID-19 pandemic had an influence on the outcomes? Based on the data of applicants who did not achieve the minimum grade between the 2019 and 2021 test calls, we explore potential differences based on sex, nationality, and the context of the COVID-19 pandemic. We intend to determine the explanatory factors that are associated with the final score obtained and offer relevant information for health planning, which must respond to the needs and expectations of citizens in terms of health, as well as the social, economic, and material progress of people. This work can help optimize access to the SHT and mitigate medical desert areas in Spain.

Does the use of traditional scoring metrics for resident recruitment create racial and gender bias [21]? Our objective is to identify the explanatory factors that approximate the final score on the national entrance examination to access the SHT. We intend to analyze quantitatively to what extent the examination, sociodemographic factors, and calendar data explain the results of the MIR examination. To do so, we will use an econometric approach that will allow us to establish a causal relationship between the results obtained and the variables that determine them.

Methods

Data

The statistical information used in this study was obtained from the final list of results of the selective medical examinations, prepared by the Ministry of Health for the annual calls during the 2019-2021 period [22]. The recruitment process was conducted through a national public call, in which applicants voluntarily register and meet basic administrative and academic requirements. The registration forms collect demographic information of the applicants, examination data, and details of the test schedule, allowing systematic tracking of each stage of the process. The analyzed period corresponds to the 3 examinations conducted between 2020 and 2022, with complete information available for all variables across all observations and no missing data. For the analysis, only the number of candidates who did not pass the examination was considered, defined as those scoring below the minimum established threshold (35% of the arithmetic mean of the 10 highest exam scores). In the 2019, 2020, and 2021 calls, the cutoff scores were 157, 156, and 177 points, respectively. In the 2022 call, the cutoff criterion was adjusted to 25% of the arithmetic mean of the 10 highest examination scores, lowering the minimum passing score to 113 points. In total, 7217 observations were analyzed: 2796 from 2019, 2526 from 2020, and 1895 from 2021.

Ethical Considerations

This study did not require review or approval by an institutional ethics committee, as it relies exclusively on publicly available, anonymized, and aggregated secondary data.

In accordance with international research ethics guidelines (such as the Declaration of Helsinki), the recommendations of the Spanish Ministry of Science and Innovation's Research Ethics Committee for studies using public data, and JMIR's editorial policies, studies using such data are considered exempt from formal ethical review.

Procedure

The final grade obtained, that is, the score, is obtained as a weighted average of the test examination (90%) and academic records (10%). This variable approximates the final result in which factors specifically related to the examination—valid answers, validity, academic records, scale, demographic factors, sex, nationality and language, and the calendar factors MIR_2020 and MIR_2021—contribute to the observed behavior of the final result (Table 1). The effect of the calendar year is approximated from the annual data corresponding to the calls made between 2019 and 2021 and the variables MIR_2019, MIR_2020 (MIR_2020=1), and MIR_2021 (MIR_2021=1), with the 2019 call serving as the control category.

XSL•FO
RenderX

Table 1. Selected variables used in the study.

Indicators, variables, and their definitions	Expected sign
Test	The expected sign of this variable is positive
Valid: quantitative variable indicating the total number of correct answers obtained by each candidate	
Scale: quantitative variable obtained by multiplying by 10 the evaluation of each candidate's file and dividing the product by the average of the 10 best examinations of the corresponding exercise	
Demographic factors	A priori; the expected sign of this variable can be + or –
Sex: dichotomous variable that takes the value 0 if the applicant is female, 1 otherwise	
Nationality: dichotomous variable that takes the value 0 if the applicant is of Spanish nationality, 1 otherwise	
Language: dichotomous variable that takes the value 0 if the applicant communicates in Spanish, 1 otherwise	
Calendar	A priori; the expected sign of this variable can be + or –
MIR_2020[a]: dichotomous variable that takes the value 1 if the sample observation corresponds to the 2020 call, otherwise 0	
MIR_2021: dichotomous variable that takes the value 1 if the sample observation corresponds to the call for the year 2021, 0 otherwise	

[a]MIR: Medical Intern Residency.

The specification of the model to be used requires the determination of the variables and functional relationship to be considered:

$$Score_i = f\,(Valid_i, Scale_i, Sex_i, Nationality_i, Language_i, MIR2020_i, MIR2021_i, u_i)$$

where, for the *ith* observation, the variable that includes the final grade obtained in the test, *Score*, denotes the dependent variable, and *u* the random disturbance term. Linear regression models were used and estimated using the ordinary least squares (OLS) method to explain *Score* based on a set of explanatory variables. All explanatory variables were selected based on theoretical relevance and data availability. Additionally, interaction terms were included where supported by prior literature or exploratory analysis. Multicollinearity was assessed using the linear correlation matrix of the explanatory variables, which showed no signs of high multicollinearity. However, when heteroskedasticity is present—meaning that the variance of the error terms is not constant—the SEs produced by OLS may be unreliable, compromising the validity of statistical inference. To address this, we use heteroskedasticity-robust SEs, specifically the Huber-White-Hinkley estimator, which adjusts the variance-covariance matrix to yield consistent SE estimates even under nonconstant error variance. This ensures more accurate hypothesis testing. Moreover, the use of this multivariate approach enables the estimation of each predictor's unique contribution while accounting for the influence of other variables, which is crucial for reducing potential biases due to confounding—particularly in nonexperimental research settings.

To make the sprint and background effects visible, taking into account the gender of the student as well as the impact that the COVID-19 pandemic had on the results for each sex, interaction terms of the sex variable with the valid, scale, MIR_2020, and MIR_2021 variables are included in the model. Similarly, to make the effect of the nationality variable visible, interaction terms are included for the valid, scale, MIR_2020, and MIR_2021 variables.

Results

Table 2 shows the descriptive statistics of the variables considered based on a sample of 7217 observations. For qualitative variables, information is made visible on the number of observations that satisfy the analyzed characteristics and their relative frequency. For the quantitative variables, the mean, median, IQR, minimum value, maximum value, and sample SD are used. Statistical analysis and model estimation were performed using the EViews program (IHS Global Inc) and our predetermined α level was .05.

XSL•FO
RenderX

Table 2. Variable definition and summary statistics (2019-2021; N=7217). This is our own elaboration based on Ministry of Health 2022 data.

Indicators	Values
Demographic factors	
Qualitative variables, n (%)	
Sex	
Male	4155 (57.57)
Female	3062 (42.43)
Nationality	
Spanish	3409 (47.24)
Non-Spanish	3808 (52.76)
Language	
Spanish	6468 (89.62)
Not Spanish	749 (10.37)
Calendar	
MIR[a] call, n (%)	
MIR_2019	2796 (38.74)
MIR_2020	2526 (35.00)
MIR_2021	1895 (26.26)
Test	
Quantitative variables	
Valid	
Mean (SD)	68.05 (12.44)
Median (IQR)	69 (60-77)
Range	0-94
Scale	
Mean (SD)	6.11 (1.13)
Median (IQR)	6.2 (5.00-7.04)
Range	4.37-5.00

[a]MIR: Medical Intern Residency.

During the selected period, the total number of eliminated candidates decreased. Of a total of 7217 eliminated candidates who attempted to earn an SHT place, 38.74% (n=2796) corresponded to the 2019 call. In the following calls, the weight was slightly lower. Between the initial and final calls, the total number of eliminated candidates decreased by 10.48 percentage points. Of the total sample, 57.57% (n=4155) of the eliminated candidates were women and 47.24% (n=3409) were Spanish. The native language of 1 in 10 of the eliminated applicants was not Spanish. The median number of valid responses and scale scores during the examined period was 69 (IQR 60-77) correct answers and 6.2 (IQR 5.00-7.04) points, respectively. The linear correlation matrix of the explanatory variables shows no signs of high multicollinearity, indicating that the independent variables are not excessively correlated with each other and can be reliably included in the regression model (Table 3). It also indicates that during the period analyzed, the direction of the linear association between the qualitative factors sex and nationality and each of the quantitative factors—validity and scale—was inversely proportional ($r_{\text{Sex-Valid}}$=–0.0202; $r_{\text{Sex-Scale}}$=–0.0505; $r_{\text{Nationality-Valid}}$=–0.0766; $r_{\text{Nationality-Scale}}$=–0.0109).

XSL•FO
RenderX

Table 3. Correlation coefficients determined using observations 1 to 7217. The critical value at 5% (2-tailed) was 0.0231 for N=7217.

	Nationality	Valid	Scale	Sex	MIR_2020	MIR_2021	Sex × valid	Sex × MIR_2020	Sex × MIR_2021	Nationality × valid	Nationality × scale	Language
Nationality	1	–0.0766	–0.0109	0.0032	–0.0492	0.0136	–0.0066	–0.0105	–0.0103	0.9652	0.9613	0.2912
Valid	__[a]	1	0.2013	–0.0202	–0.2033	0.3388	0.1397	–0.1369	0.2028	0.1171	–0.0359	–0.0444
Scale	—	—	1	–0.0505	0.0601	0.1388	–0.0146	0.0095	0.0688	0.0271	0.2005	–0.041
Sex	—	—	—	1	–0.001	0.0032	0.9709	0.486	0.4139	0.0013	–0.0069	0.0186
MIR_2020	—	—	—	—	1	–0.4379	–0.0352	0.5686	–0.2607	–0.089	–0.0328	–0.0192
MIR_2021	—	—	—	—	—	1	0.0577	–0.249	0.5955	0.0772	0.0461	–0.0079
Sex × valid	—	—	—	—	—	—	1	0.426	0.478	0.0214	–0.0102	0.0072
Sex × MIR_2020	—	—	—	—	—	—	—	1	–0.1482	–0.0363	–0.0082	–0.0026
Sex ×MIR_2021	—	—	—	—	—	—	—	—	1	0.0271	0.0066	–0.0086
Nationality × valid	—	—	—	—	—	—	—	—	—	1	0.9414	0.2729
Nationality × scale	—	—	—	—	—	—	—	—	—	—	1	0.2672
Language	—	—	—	—	—	—	—	—	—	—	—	1

[a]Not applicable.

Following the approach adopted for the analysis of applicants regarding access to the SHT who passed the MIR examination [19], Table 4 shows the results corresponding to the eliminated applicants. The model is estimated with SEs and robust covariances given the cross-sectional nature of the support. The regression coefficient is globally significant ($F_{9,7201}$=4038.137), and the collinearity indicators do not show severe linear association problems. These results indicate that not all the selected factors constitute relevant determinants that help to explain the observed behavior of the final result obtained.

Table 4. Regression model for the final score (2019-2021; MIR_2019: n=2796; MIR_2020: n=2526; MIR_2021: n=1895; MIR: Medical Intern Residency). This is our own elaboration based on the Ministry of Health 2022 data.

Dependent variable: score	β coefficient	*P* value
Constant	28.7606	<.001
Valid	0.6012	<.001
Scale	0.9972	<.001
Sex	0.3186	.54
Nationality	–0.6178	<.001
MIR_2020	–0.3359	<.001
MIR_2021	–6.2058	<.001
Sex × valid	–0.0049	.48
Sex × MIR_2020	–0.0574	.65
Sex × MIR_2021	0.0699	.65
Adjusted *R*	0.8344	__[a]
F test (*df*)	4038.137 (9,7201)	<.001

[a]Not applicable.

The effects on the final results of the examination component and academic records are determined through the validity and scale variables. The effect on the final test result in both cases is shown to be directly proportional and significant. The effect of the applicant's sex is approximated by the sex variable and its interaction with the variable that approximates the number of correct answers provided on the examination. In both cases, the effect on the final result is not perceived as significant. The nationality of the applicant is approximated by the nationality variable, and its effect is found to be significant. The results obtained show the differential effect on the final result between Spanish (control category) and non-Spanish individuals. The

XSL•FO
RenderX

difference in the total score obtained between a Spanish and non-Spanish eliminated candidate, ceteris paribus, is 0.6178 lower for the latter. The contribution of the calendar year is captured by the variables MIR_2020 and MIR_2021, which collect the observations corresponding to the tests carried out in March 2021 and January 2022, respectively. The result obtained is negative and significant in both instances. The final results of the tests carried out in 2021 and 2022, ceteris paribus, are 0.3116 and 6.2355 points lower than those of the control category, respectively. The effects of the interaction of the calendar factor and the sex factor are collected through the interaction variables sex × MIR_2020 and sex × MIR_2021, whose effect on the final result of the call is not perceived as being significant.

During the period analyzed, 1 in 4 applicants were not Spanish and neither were 45.24% (n=3265) of the total number of eliminated applicants, of which 81% came from South America. Table 5 explores the impact of the origin of the eliminated candidates on the final result. The effect of origin and its interaction with both the examination and academic records are analyzed through the variables nationality, nationality × valid, and nationality × scale, respectively. Three alternative models are estimated to approximate the behavior of the final test result. Model 1 incorporates the variable nationality × scale, model 2 adds the interaction with the examination through the nationality × valid variable, and model 3 adds the nationality × valid variable. Based on the results obtained, the model chosen is the one that offers the greatest predictive capacity and incorporates the smallest number of regressors. Based on the goodness-of-fit test, model 2, which simultaneously incorporates the interaction of nationality with the variables validity and scale, registered the highest value (0.8390). However, model 1 minimizes the values of the Akaike information criterion (27721.32), the Schwart criterion (27797.04) and the Hannan-Quinn criterion (27747.37); it also incorporates a smaller number of explanatory factors. The regression coefficient is globally significant ($F^*_{10,7206}$=3660.368), and the collinearity indicators do not show severe linear association problems.

Table 5. Regression model for the final score (2019-2021; MIR_2019: n=2796; MIR_2020: n=2526; MIR_2021: n=1895; MIR: Medical Intern Residency). This is our own elaboration based on the Ministry of Health 2022 data.

Dependent variable: score	Model 1[a]		Model 2[b]		Model 3[c]	
	β coefficient	*P* value	β coefficient	*P* value	β coefficient	*P* value
Constant	28.7631	<.001	28.7988	<.001	28.7943	<.001
Valid	0.601825	<.001	0.5967	<.001	0.5972	<.001
Scale	0.9892	<.001	0.9952	<.001	0.9953	<.001
Sex	0.2701	.60	–0.0561	.92	0.0261	.95
Nationality	–1.4682	<.001	–6.8459	<.001	–6.7054	<.001
MIR_2020[d]	–0.3425	<.001	–0.3259	<.001	–0.3261	<.001
MIR_2021	–6.2203	<.001	–6.1005	<.001	–6.1010	<.001
Sex × valid	–0.0044	.53	7.0636×10^{-05}	.99	–0.0006	.89
Sex × MIR_2020	–0.0470	.71	–0.0452	.72	–0.0517	.68
Sex × MIR_2021	0.0741	.63	0.0094	.95	0.0165	.91
Nationality × scale	0.1362	.003	0.0214	.64	__[e]	—
Nationality × valid	—	—	0.0791	<.001	0.0791	<.001

[a]Adjusted *R*=0.8342; $F_{10,7200}$=3660.368; *P*<.001; Akaike information criterion=27721.32; Schwarz criterion=27797.04; Hannan-Quinn criterion=27747.37.

[b]Adjusted *R*=0.8390; $F_{11,7199}$=3417.508; *P*<.001; Akaike information criterion=27769.02; Schwarz criterion=27851.62; Hannan-Quinn criterion=27797.44.

[c]Adjusted *R*=0.8388; $F_{10,7200}$=3154.413; *P*<.001; Akaike information criterion=27773.24; Schwarz criterion=27848.96; Hannan-Quinn criterion=27799.30.

[d]MIR: Medical Intern Residency.

[d]Not applicable.

The impact on the final test score of the explanatory factors validity and scale is, in both cases, directly proportional and significant. The sex and sex × valid variables approximate the effect on the final result of the qualitative factor sex and its interaction with the examination component, respectively, none of which are perceived as significant. The effects that origin and academic records have on the final result are approximated through the variables nationality and nationality × scale, respectively. The impact of both variables is significant, although the signs of the results are opposite in nature. The difference in the total score obtained between Spanish and non-Spanish eliminated candidates is, ceteris paribus, 1.4682 points lower for the latter. However, the differential effect corresponding to the scale between the final result of a Spanish and non-Spanish eliminated candidate, ceteris paribus, is 0.1362 points greater for the non-Spanish candidate.

The impact that the calendar year has on the final result is collected by the variables MIR_2020 and MIR_2021, which group together the observations corresponding to the tests carried out in March 2021 and January 2022, respectively. The

XSL•FO
RenderX

result obtained is negative and significant in both references. The final results of the tests carried out in 2021 and 2022 are, ceteris paribus, 0.3425 and 6.2203 points lower than those of the other years, respectively. The effects of the interaction of the calendar factor and the sex factor through the variables sex × MIR_2020 and sex × MIR_2021 did not significantly impact the final result.

Table 6 shows the estimation of the standardized coefficients, which facilitates the ranking of the relative weight of the explanatory variables in the model over the dependent variable. According to this criterion, the dominance of the examination variable in the data distribution is visualized in the final result through the variable that includes the number of correct answers provided on the test. Next, the weights of the variables that approximate the calendar year represent the second and sixth positions in the rankings, MIR_2021 and MIR_2020, respectively, and the academic records obtained through the scale variable rank third. The fourth and fifth positions are occupied by the origin of the applicant and its interaction with the academic record, respectively. The weight of the sex factor completes the distribution.

Table 6. Standardized β coefficients. This is our own elaboration based on the Ministry of Health 2022 data.

Variable	Standardized β coefficient	Order position
Constant	__[a]	—
Valid	0.9639	1
Scale	0.1451	3
Sex	0.0172	8
Nationality	–0.0944	4
MIR_2020[b]	–0.0210	6
MIR_2021	–0.3525	2
Sex × valid	–0.0194	7
Sex × MIR_2020	–0.0018	9
Sex × MIR_2020	0.0024	10
Nationality × scale	0.0544	5

[a]MIR: Medical Intern Residency.

Of a total of 7217 candidates eliminated, 3808 physicians of non-Spanish origin did not pass the SHT access test, with 2639 coming from South America. The native language of 749 of the eliminated test candidates was not Spanish. The consideration of language as a possible limiting factor for taking the test is only reflected in the examination score by determining the total number of correct answers. Table 7 shows the results of the examination score, namely, exam_score, based on the validity and language variables. The regression coefficient is globally significant, and the collinearity indicators do not show linear association problems. The effect of the valid variable on the examination score is shown to be directly proportional and significant. However, the effect of the explanatory factor of language is not perceived as significant.

Table 7. Examination score (2019-2021; MIR_2019: n=2796; MIR_2020: n=2526; MIR_2021: n=1895; MIR: Medical Intern Residency). This is our own elaboration based on the Ministry of Health 2022 data[a].

Dependent variable: examination score	β coefficient	*P* value
Constant	20.9447	<.001
Valid	0.5294	<.001
Language	−0.0494	.75

[a]Adjusted R=0.6603; $F_{2,7214}$=7013.959; P<.001.

Discussion

This study fills an important gap in the literature by focusing on the profile and factors influencing candidates who fail to gain access to SHT in Spain, which is approximately 1 in 5 applicants. Using econometric modeling of data from the 2019-2021 MIR examination calls, we identified the key quantitative and qualitative variables affecting the final scores of those excluded, including examination results, academic records, nationality, gender, and the impact of the COVID-19 pandemic.

Some studies have analyzed the factors that determine the results of SHT entrance examinations in Spain [7,19]. However, there is little information on the profile of applicants who do not gain access and are ultimately excluded from the process, that is approximately 1 in 5 applicants. Knowing the differences between the factors that determine the success/failure of access to the SHT could favor the management of health care in the context of a scarcity of human resources.

XSL•FO
RenderX

Our empirical exercise focused on econometrically modeling the final MIR examination results for participants eliminated from the process during 2019-2021. The econometric approach allows causal analysis to be carried out by identifying and classifying the explanatory factors that help elucidate the final scores and their relevance, effect, and meaning [23]. Although the final results combine examination scores and academic records, they also reflect examinees' interaction with real-world environments, justifying the econometric integration of both quantitative and qualitative factors such as sex, nationality, and examination calendar.

The study's findings reveal significant differences between Spanish and non-Spanish applicants, with a persistent structural disadvantage in the performance of the latter. Among the relevant factors defining the final score, the validity and scale variables—corresponding to the so-called sprint effect and background effect, respectively—stand out [7,24]. An additional valid examination answer or one more point in the academic record, ceteris paribus, increases the final score, though the sprint effort (ie, examination performance) outweighs the background effort (ie, academic record) in standardized terms. This dominance aligns with that reported in studies emphasizing differential risk exposures by gender [25,26], though these variables alone do not clearly indicate success or failure, and their relationship remains coincidental [19].

Regarding sociodemographic factors, sex did not have a significant effect on final results, contrary to some previous studies suggesting male candidates historically achieve higher scores [15,27-29]. Furthermore, the interaction between sex and examination validity did not demonstrate a consistent effect. In contrast, nationality emerged as a significant explanatory factor: being non-Spanish carried a fixed, unfavorable impact on final scores, ranking fourth in standardized importance after academic record. The Ministry of Health requires degree homologation based on MIR examination success for professional practice in Spain [13]. Approximately 1 in 4 candidates during the study period was non-Spanish, with 45.24% (n=3265) of these candidates having been eliminated.

Importantly, the interaction between nationality and academic record was positive and significant, suggesting that strong academic backgrounds can partly compensate for structural disadvantages faced by non-Spanish applicants. This finding may point to systemic barriers such as degree recognition issues or implicit biases rather than linguistic obstacles, as native language was not a significant factor despite 749 eliminated non-Spanish candidates having a different native language [19,30,31].

Another key conclusion relates to the negative impact of the COVID-19 pandemic on final grades. The model's calendar variables show that candidates from the 2020 and 2021 examination calls scored significantly lower than the prepandemic cohort. Pandemic-induced disruptions—including rapid shifts to virtual or hybrid learning and elevated stress—impaired examination preparation [32-36]. Notably, the interaction between sex and examination year was not significant.

From a causal perspective, examination results, calendar year, academic records, and sociodemographic factors are key determinants of the final scores of excluded candidates. While sex remains a determinant for successful candidates regardless of nationality [19], among those excluded, nationality plays a more prominent role. This highlights the importance of internationalization in the medical profession as a health policy factor to balance supply and demand in Spain's health system. The regulated SHT program aims to ensure quality and comprehensive care, and immigration of doctors prepared for the MIR examination may help alleviate short-term workforce deficits and increase labor market flexibility.

In light of these conclusions, it is essential to analyze, evaluate, and prioritize the factors that influence final scores and the allocation of places. This approach can serve as a basis for designing policies that improve equity and efficiency in access to the health care system. The observed disparities underscore the need for targeted support mechanisms for internationally trained candidates. These may include structured mentoring programs, individualized guidance, and orientation strategies aimed at improving their integration and performance. Furthermore, in anticipation of future crises, educational authorities should consider implementing academic recovery measures and psychological support systems to ensure more balanced and inclusive preparation environments.

To address the underlying causes of these disparities, it is also advisable to implement specific support policies—such as guidance programs, tutoring or integration strategies—for candidates trained abroad. Key policy actions could include reviewing and streamlining the recognition process for foreign medical degrees, creating mentorship networks that connect international candidates with experienced professionals in the Spanish health care system, and developing crisis-response protocols that include flexible examination preparation resources and integrated emotional support services for all applicants.

By addressing these structural challenges, the Spanish health care system can take meaningful steps toward a more inclusive and resilient process for accessing specialized medical training.

One of the main limitations of this study lies in the nature of the data used. The dataset provided by the Ministry of Health is fully anonymized, which prevents any longitudinal tracking of individual candidates across different examination sessions or stages of the selection process. As a result, it is not possible to analyze repeated attempts by the same individuals or to evaluate changes in performance over time. This limits the ability to draw conclusions about the evolution of candidates' profiles or the cumulative impact of training and preparation efforts. Despite this constraint, the study offers a robust cross-sectional analysis of a large and representative sample of excluded candidates over 3 MIR calls, contributing valuable insights into a largely understudied population.

A promising future line of research involves conducting independent qualitative studies to complement the quantitative findings. Since the current dataset is anonymized and does not allow for the identification of individual applicants, future studies could use interviews, focus groups, or surveys with a representative sample of excluded candidates. This would

XSL•FO
RenderX

provide deeper insights into their personal experiences, perceived barriers, and contextual challenges such as preparation conditions, psychological factors, or socioeconomic constraints, which may not be fully captured through econometric modeling alone.

Acknowledgments

We certify that no generative artificial intelligence tools were used in the preparation of this manuscript. All parts of the manuscript—including conceptualization, methodology, data analysis, writing, and editing—were carried out solely by the authors listed. The authors take full responsibility for the content of the manuscript and affirm its originality and integrity.

Data Availability

The datasets generated during this study are available from the corresponding author on reasonable request.

Funding

This work has not received specific funding for its development. The only associated funding has been used exclusively to cover the costs arising from the publication process. No funds have been received for the analysis, writing of the manuscript, or any other activity related to the scientific content of the study. We received a Foundation University of Oviedo Professor Víctor Asensi research grant (FI-21-026).

Authors' Contributions

Conceptualization: MDF, MLM, and VA
Data curation: MDF, MLM, and VA
Formal analysis: MDF and MLM
Writing—original draft: MDF, MLM, and VA
Writing—review and editing: MDF, MLM, and VA

Conflicts of Interest

None declared.

References

1. Roberfroid D, Leonard C, Stordeur S. Physician supply forecast: better than peering in a crystal ball? Hum Resour Health 2009 Feb 13;7:10 [FREE Full text] [doi: 10.1186/1478-4491-7-10] [Medline: 19216772]
2. Safarishahrbijari A. Workforce forecasting models: A systematic review. Journal of Forecasting 2018 Aug 17;37(7):739-753. [doi: 10.1002/for.2541]
3. Szabo S, Nove A, Matthews Z, Bajracharya A, Dhillon I, Singh DR, et al. Health workforce demography: a framework to improve understanding of the health workforce and support achievement of the Sustainable Development Goals. Hum Resour Health 2020 Jan 29;18(1):7 [FREE Full text] [doi: 10.1186/s12960-020-0445-6] [Medline: 31996212]
4. España: Perfil Sanitario del país 2021. Organisation for Economic Co-operation and Development. 2021. URL: https://doi.org/10.1787/3d1c2ed2-es [accessed 2025-11-26]
5. Kwon E. “For Passion or for Future Family?” Exploring factors influencing career and family choices of female medical students and residents. Gend. Issues 2016 Jun 25;34(2):186-200. [doi: 10.1007/s12147-016-9168-3]
6. Riska E. Gender and medical careers. Maturitas 2011 Mar;68(3):264-267. [doi: 10.1016/j.maturitas.2010.09.010] [Medline: 21030166]
7. Rodriguez Santana I. Becoming a resident in a high demanded medical specialty: an unequal race? Evidence from the Spanish resident market. Hum Resour Health 2021 Jan 06;19(1):3 [FREE Full text] [doi: 10.1186/s12960-020-00543-2] [Medline: 33407526]
8. Babaria P, Abedin S, Berg D, Nunez-Smith M. "I'm too used to it": a longitudinal qualitative study of third year female medical students' experiences of gendered encounters in medical education. Soc Sci Med 2012 Apr;74(7):1013-1020. [doi: 10.1016/j.socscimed.2011.11.043] [Medline: 22341202]
9. Heru AM. Pink-collar medicine: Women and the future of medicine. Gend. Issues 2005 Dec;22(1):20-34. [doi: 10.1007/s12147-005-0008-0]
10. Estadística de Estudiantes. Ministerio de Ciencia, Innovación y Universidades. URL: https://www.universidades.gob.es/ [accessed 2022-06-29]
11. Estimación de la oferta y demanda de médicos especialistas: España 2018-2030. Universidad de Las Palmas de Gran Canaria. 2019. URL: https://accedacris.ulpgc.es/jspui/handle/10553/106259 [accessed 2025-12-26]
12. Pastor-Bravo M, Nelson S. Migration of Latin American nurses to Spain 2006-2016: a case study. Int Nurs Rev 2019 Jun;66(2):183-190 [FREE Full text] [doi: 10.1111/inr.12511] [Medline: 30989654]

XSL•FO
RenderX

13. Pardo JF. Formación sanitaria especializada: un cambio es inevitable. Educación Médica 2021 May;22:47-48. [doi: 10.1016/j.edumed.2021.05.002]
14. Aranda Briones B. Advancing the delimitation of electronic voting: the saga continues with a new trilogy. Annotated Spanish Parliamentary Case Law. 2024 Dec 30. URL: https://www.boe.es/boe/dias/2010/08/07/ [accessed 2025-11-26]
15. Bonillo A. [Access tests to specialized health training for doctors and other healthcare professionals in Spain: examining the exam and the examined candidates]. Gac Sanit 2012 May;26(3):231-235 [FREE Full text] [doi: 10.1016/j.gaceta.2011.09.020] [Medline: 22197321]
16. Kelleher M, Schumacher DJ, Zhou C, Kwakye D, Santen SA, Warm E, et al. Public board acore reporting undermines holistic review for residency selection. J Gen Intern Med 2025 Jan 04;40(1):17-21. [doi: 10.1007/s11606-024-09133-7] [Medline: 39496852]
17. Muecke T, Rao A, Walker H, Tinnion J, Jesudason D, Bacchi S, et al. Specialists of tomorrow: an umbrella review of evidence supporting criteria used in medical and surgical specialty training selection processes. Discov Educ 2024 Aug 05;3(1):119. [doi: 10.1007/s44217-024-00205-8]
18. Muecke T, Bacchi S, Casson R, Chan WO. Building the workforce of tomorrow: The weighting of rural exposure in standardised curriculum vitae scoring criteria for entrance into Australian specialty training programs. Aust J Rural Health 2024 Aug 08;32(4):827-833. [doi: 10.1111/ajr.13131] [Medline: 38715522]
19. Díaz-Fernández M, Llorente-Marrón M, Cocina-Díaz V, Asensi V. COVID-19 and access to medical professional careers: Does gender matter? Int J Environ Res Public Health 2023 Jul 31;20(15):6477 [FREE Full text] [doi: 10.3390/ijerph20156477] [Medline: 37569018]
20. Lombardi CV, Chidiac NT, Record BC, Laukka JJ. USMLE step 1 and step 2 CK as indicators of resident performance. BMC Med Educ 2023 Jul 31;23(1):543 [FREE Full text] [doi: 10.1186/s12909-023-04530-8] [Medline: 37525136]
21. Lai SH, Suarez-Pierre A, Jaiswal K, Travis C, Steward L, Nehler M, et al. Implementation of a holistic review process of US allopathic medical students eliminates non-comparable metrics and bias in general surgery residency interview invitations. J Surg Educ 2023 Nov;80(11):1536-1543. [doi: 10.1016/j.jsurg.2023.07.008] [Medline: 37507300]
22. Formación Especializada. Ministerio de Sanidad. URL: https://www.sanidad.gob.es/areas/profesionesSanitarias/formacionEspecializada/home.htm [accessed 2025-03-26]
23. Nuñez LF. Econometría de evaluación de impacto. Economía 2011;34(67):81-125 [FREE Full text] [doi: 10.18800/economia.201101.004]
24. Dávila-Quintana CD, López-Valcárcel BG, Barber P, Ortún V. El baremo académico en el acceso a la formación médica especializada en España. FEM (Ed. impresa) 2015 Jun;18(3):219-224. [doi: 10.4321/s2014-98322015000400011]
25. Gneezy U, Niederle M, Rustichini A. Performance in competitive environments: Gender differences. The Quarterly Journal of Economics 2003 Aug 01;118(3):1049-1074. [doi: 10.1162/00335530360698496]
26. Niederle M, Vesterlund L. Explaining the gender gap in math test scores: The role of competition. Journal of Economic Perspectives 2010 May 01;24(2):129-144. [doi: 10.1257/JEP.24.2.129]
27. Dayal A, O'Connor DM, Qadri U, Arora VM. Comparison of male vs female resident milestone evaluations by faculty during emergency medicine residency training. JAMA Intern Med 2017 May 01;177(5):651-657 [FREE Full text] [doi: 10.1001/jamainternmed.2016.9616] [Medline: 28264090]
28. Sobol DL, Berfield KS, Shalhub S, Tatum RP, Yale LA, Perkins JD, et al. Exploring the influence of gender on surgical clerkship grades and test scores: A single institution, multisite comparison. J Surg Educ 2022 Sep;79(5):1132-1139. [doi: 10.1016/j.jsurg.2022.05.008] [Medline: 35660307]
29. Thackeray EW, Halvorsen AJ, Ficalora RD, Engstler GJ, McDonald FS, Oxentenko AS. The effects of gender and age on evaluation of trainees and faculty in gastroenterology. Am J Gastroenterol 2012 Nov;107(11):1610-1614. [doi: 10.1038/ajg.2012.139] [Medline: 23160284]
30. Fenoll-Brunet MR. El concepto de internacionalización en enseñanza superior universitaria y sus marcos de referencia en educación médica. Educación Médica 2016 Jul;17(3):119-127. [doi: 10.1016/j.edumed.2016.07.002]
31. Hidalgo DT, Pérez JAA, Rojas PAD, Aguilera NG. The internationalization of higher education and medical universities as a source of development. EDUMECENTRO 2022;14:i-iv.
32. Valdivieso MA, BURBANO VM, BURBANO AS. Percepción de estudiantes universitarios colombianos sobre el efecto del confinamiento por el coronavirus, y su rendimiento académico. Espacios 2020 Nov 05;41(42):269-281. [doi: 10.48082/espacios-a20v41n42p23]
33. García Prieto FJ. Competencia digital del alumnado universitario y rendimiento académico en tiempos de COVID-19. Pixel-Bit. 2022. URL: https://institucional.us.es/revistas/PixelBit/64/7_91862.pdf [accessed 2025-11-26]
34. Kang L, Li Y, Hu S, Chen M, Yang C, Yang BX, et al. The mental health of medical workers in Wuhan, China dealing with the 2019 novel coronavirus. Lancet Psychiatr 2020 Mar;7(3):e14. [doi: 10.1016/s2215-0366(20)30047-x]
35. Lozano-Vargas A. Impacto de la epidemia del Coronavirus (COVID-19) en la salud mental del personal de salud y en la población general de China. Rev Neuropsiquiatr 2020 Apr 11;83(1):51-56. [doi: 10.20453/rnp.v83i1.3687]
36. Shigemura J, Ursano RJ, Morganstein JC, Kurosawa M, Benedek DM. Public responses to the novel 2019 coronavirus (2019-nCoV) in Japan: Mental health consequences and target populations. Psychiatry Clin Neurosci 2020 Apr;74(4):281-282 [FREE Full text] [doi: 10.1111/pcn.12988] [Medline: 32034840]

XSL•FO
RenderX

Abbreviations

EU: European Union
MIR: Medical Intern Residency
NHS: National Health System
OECD: Organisation for Economic Co-operation and Development
OLS: ordinary least squares
SHT: Specialized Health Training

Edited by A Mavragani, J Sarvestan; submitted 10.02.25; peer-reviewed by CP Fu; comments to author 30.06.25; revised version received 21.10.25; accepted 29.10.25; published 09.12.25.

Please cite as:
Diaz-Fernandez M, Llorente-Marron M, Asensi V
Access to Specialized Medical Training in Spain and Determinants of Failure in the National Entrance Examination: Econometric Modeling Study
JMIR Form Res 2025;9:e72440
URL: https://formative.jmir.org/2025/1/e72440
doi:10.2196/72440
PMID:

©Montserrat Diaz-Fernandez, Mar Llorente-Marron, Victor Asensi. Originally published in JMIR Formative Research (https://formative.jmir.org), 09.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Proactively Delivered Digital Mental Health Support for Health Care Workers: Usability and Acceptability Evaluation

Lauren Southwick[1,2], MPH; Rachel Gonzales[1,2], MPH; Lisa Bellini[3], MD; David A Asch[3], MD, MBA; Nandita Mitra[4], PhD; Mohan Balachandran[2], MA, MS; Courtney Benjamin Wolk[5], PhD; Emily M Becker-Haimes[5], PhD; Rachel Kishton[2,5], MD; Sarah Beck[2], MA; Raina M Merchant[1,2], MD, MSHP; Anisk K Agarwal[1,2], MPH, MS, MD

[1]Department of Emergency Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States

[2]Center for Health Care Transformation and Innovation, Penn Medicine, Philadelphia, PA, United States

[3]Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States

[4]Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States

[5]Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States

Corresponding Author:
Anisk K Agarwal, MPH, MS, MD
Department of Emergency Medicine
Perelman School of Medicine
University of Pennsylvania
3600 Civic Center Boulevard
Philadelphia, PA, 19104
United States
Phone: 1 2674280125
Email: Anish.Agarwal@pennmedicine.upenn.edu

Abstract

Background: Health systems are investing in mental health and well-being support tools and resources for health care workers (HCW). Considering the mental health strain facing HCWs, there is a need to optimize the current mental health delivery model.

Objective: This study aimed to evaluate the usability and acceptability of a proactive digital mental health approach (Cobalt+;Penn Medicine), which included services proactively sent to HCWs via text messaging, including (1) monthly automated text messaging reminders and links to Cobalt, and (2) bimonthly text-message–based measures of depression and anxiety.

Methods: This study used the System Usability Scale (SUS), Net Promoter Score (NPS), and open-ended questions to capture Cobalt+ participants who received proactive digital mental health tools and resources. Descriptive summary statistics were used for SUS and NPS outcome measures, and a chi-square test was used to detect group differences. Open-ended questions were analyzed using a qualitative open coding process by 2 coders. Research team members calculated interrater agreement (Cohen κ above 0.80).

Results: A total of 162 of 642 HCWs randomized to Cobalt+ (25.2%) visited Cobalt due to a proactive text message and completed usability and acceptability measures. The mean age was 38.9 years, most were female (90.7%), 56.8% White, 53.1% married or partnered, and 34.6% engaged in shift work. The mean SUS score was 74.43 (median score 72.5). Participants said they mostly “browsed” the online mental health platform. Cobalt+ received an NPS of 13.7. When asked to elaborate on their experience, 2 categories (eg, positive and negative experiences) with 13 subcategories were identified. Most participants noted the brief process that helped prioritize mental health: “Forget otherwise. Puts in forefront of my mind,” and “Your texts do remind me to take stock of my current feelings.”

Conclusions: A proactive digital mental health approach may help overcome barriers in the uptake of services that are otherwise passively available to HCWs. This study demonstrated that the proactive approach is generally usable, modestly acceptable, and further supplemented by HCW feedback. These findings suggest the approach’s viability and the need for additional research toward improvement and broader implementation.

Trial Registration: ClinicalTrials.gov NCT05028075; https://clinicaltrials.gov/study/NCT05028075

(JMIR Form Res 2025;9:e74086) doi:10.2196/74086

XSL•FO
RenderX

KEYWORDS

well-being; health care workforce; digital intervention; feasibility; COVID-19 pandemic; usability; acceptability

Introduction

Health care workers' (HCWs) well-being is essential to ensure high-quality patient care and the durability of the workforce [1]. The rates of stress and burnout among HCWs were high before the COVID-19 pandemic, and these rates have persisted or increased [2]. The pandemic resulted in acute stressors directly impacting an already strained health care system. HCWs faced unique challenges during the pandemic related to rapid shifts in care, moral injury, and concerns about risks of infection [3-5].

In response to the demands of the pandemic, health systems have invested in mental health and well-being tools, including in-person resources such as peer-to-peer networks and digital mental health programs that offer remote or asynchronous mental health care and resources to their employees [2,6-9]. Digital mental health programs have garnered attention as they offer privacy and confidentiality to overcome stigma and can be scaled to employee bases [10]. These digital platforms offer a variety of resources such as individual health assessments, resilience tools, journaling prompts, progress visualization, and connection to mental health resources [9,11]. Research has begun to explore the utility of digital platform use among diverse US employers. Bondar et al [10] found that among 66 US employers across 40 states from 2018 to 2021 (of which 5% were included health care organizations), 1132 employees participated in an employee-sponsored digital program. They found among their sample of employees with moderate anxiety or depression and who attended at least 1 mental health appointment had improved depression and anxiety symptoms [10]. These results are promising, but a gap exists in exploring how digital platforms were leveraged specifically among HCWs.

Health systems have created their own digital platforms and tools to support HCW well-being and mental health [5,12,13]. The Mount Sinai Health System launched the Wellness Hub app [12] which includes self-administered surveys, resilience tools, journaling prompts and well-being videos, and personalization features. Understanding of the tool use was measured by metrics such as the number of downloads, surveys submitted, and engagement metrics (eg, number of openings). The US Department of Veterans Affairs created the COVID Coach app [9], which included tools and resources to support well-being, track personal growth, and visualize progress over time. Penn Medicine created Cobalt, an open-source online centralized suite of mental health resources designed to support HCWs through a stepped care model [12]. Cobalt offers depression and anxiety screening and connects users with a well-being coach (for mild symptoms), a care manager (for moderate symptoms), or a clinician (for more severe symptoms). In 2020 Cobalt had over 10,000 users, 200,000 page views, 1400 one-on-one appointments, and over 1000 group sessions [12].

There is a need to optimize the current mental health delivery model for HCWs as mental health strains persist in medicine. Tong et al [14] notes a need to shift from focusing solely on symptom alleviation but toward proactive individual health-seeking and health-promoting behaviors [15]. For example, the current and usual care model requires HCWs to personally initiate multiple steps (1) identify a personal need for support; (2) know where to seek support and care; (3) download the app, visit a website, or schedule an appointment; and (4) use the resource or attend the appointment. In this context, the HCW has to seek the resources they need, and there may be several barriers to completing each step. This is particularly relevant, as mental health conditions can compromise insight, motivation, and decision-making, thereby making self-directed engagement in care more challenging.

To address this gap, this study team conducted a randomized controlled trial (RCT) among 1275 HCWs at a large, urban, academic health system that demonstrated reduced depression and anxiety symptoms after 6 months of a proactive digital approach to Cobalt [16]. Despite these encouraging findings, little is known about the HCW experience receiving proactive resources. Although they may overcome barriers to accessing tools, some may view receiving proactive mental health services as off-putting, intrusive, and counterproductive. Research is needed to better understand the HCW experience of workplace mental health programs and proactive approaches.

This analysis aims to evaluate the usability and acceptability of Cobalt+ (Penn Medicine), a proactive and digital mental health approach, and capture the participant. Two key objectives guided the analyses: (1) evaluate usability and acceptability through widely used and validated measures; (2) capture participants' perceptions and experience through open-ended questions. We hypothesize that participants will find this approach both usable and acceptable. As digital mental health programs continue to scale across institutions, it is critical to better understand how participants experienced and engaged with the intervention, as these insights can directly inform the ongoing efforts to support HCW mental health.

Methods

Overview

This is a subgroup analysis of the Cobalt+ RCT, a separately reported RCT [16]. The parent RCT randomized 1275 HCWs to test the effect of proactive digital mental health messaging (Cobalt+). A total of 633 of the HCWs were randomized to usual care, and 642 HCWs were randomized to Cobalt+. Eligible HCWs were employed by a large, urban, academic health system in the Northeast and provided informed consent. Participants were excluded if they were non–English speaking, did not have daily access to a smartphone, or did not participate in at least 4 hours per week of patient care. Recruitment was conducted via email from January to May 2022. Eligible and consenting individuals were enrolled and randomized on a rolling basis, and the final participant completed follow-up in March 2023 [16].

XSL•FO
RenderX

Cobalt+ intervention participants received a suite of services via proactive text messaging, including (1) monthly automated text messaging reminders and links to Cobalt, and (2) bimonthly text-message–based measures of depression and anxiety. Participants included in this analysis are those randomized to Cobalt+ and who visited the Cobalt platform due to the text messages and completed usability and acceptability measures. The trial was approved by the institutional review board at the University of Pennsylvania (IRB protocol #: 848844) and preregistered on ClinicalTrials.gov (identifier: NCT05028075).

Data Collection

The intervention and data collection process was entirely remote. Outcome measures and text messages were administered in English and electronically using a Health Insurance Portability and Accountability Act (HIPAA)–compliant research platform to facilitate remote engagement for clinical trials and care [17].

Outcome Measures and Data Sources

All participants completed baseline demographics before randomization; variables included age, sex, race, ethnicity, marital status, and professional role. They also completed depression (Patient Health Questionnaire-9 [PHQ-9] [18]) and anxiety assessments (Generalized Anxiety Disorder-7 [GAD-7] [19]), both of which are widely used and are validated instruments and have been used in multiple studies and efficiently screen for symptoms of depression and anxiety. All participants completed PHQ-9 and GAD-7 follow-up assessments at months 6 and 9. Usability and acceptability outcome measures for this analysis were collected cross-sectionally at the 6-month survey only. The datasets generated or analyzed during this study are not publicly available due to participant privacy, but a deidentified dataset is available from the corresponding author on reasonable request. Of note, the PHQ-9 and GAD-7 data are available for the public to access through the National Institute of Mental Health Data Archive, which makes data available across studies from a variety of scientific domains.

Usability

Usability was measured using the System Usability Scale (SUS), a 10-item validated survey [20]. The SUS is a widely used measure of perceived usability across digital products and platforms. We selected SUS because it is brief, simple to administer electronically, and extensively used to evaluate digital products. It has been used to evaluate HCW experience of novel, innovative tools [21]. SUS is calculated by scoring the 10 responses on a 5-point Likert scale. Odd-numbered items' scores are subtracted by 1, while even-numbered items are subtracted by subtracting the response from 5. Then the total score is multiplied by 2.5 to yield a final usability score. SUS produces a score ranging from 0 to 100, with higher scores indicating better usability. A SUS score above 68 is typically considered above average. Open-ended questions captured how Cobalt+ participants used the platform. We used the SUS scores and open-ended questions to measure Cobalt's usability among study participants.

Acceptability

Acceptability was measured by the Net Promoter Score (NPS), a common assessment to capture user loyalty and satisfaction [22] NPS was selected for its brevity, interpretability, and widespread use in evaluating new products or services. NPS is often used to evaluate health systems or clinical programs and is a common metric for acceptability partly due to its sensitivity to detractors [21,23]. NPS measures the likelihood of recommending the program to a friend or colleague. It creates a score by subtracting the "detractors" (scores of 0-6) from "promoters" (scores of 9-10). A score of 7 or 8 is considered neutral. For instance, to calculate NPS, you calculate the proportion of promoters minus the proportion of detractors. NPS can range from –100 to 100; scores above 20 indicate "good" customer satisfaction, while scores above 50 represent "strong" customer satisfaction [22]. Open-ended responses assessed general satisfaction and experience with Cobalt+.

Analyses

Descriptive summary statistics were used for usability (SUS) and acceptability (NPS) outcome measures, and a chi-square test was used to detect group differences (Stata SE 17, Stata Corp LLC). All available data were analyzed; no additional procedures for handling missingness were undertaken.

Open-ended questions were analyzed using a qualitative open coding process and a general inductive approach. Two trained coders (MS and LS) jointly read 50 responses and drafted the outline of the codebook together. The 2 coders self-identified as female, and one as White and one as South Asian. Each coder had 10+ years of experience conducting and analyzing qualitative data. Each coder then independently read 50 more responses to adjust the codebook as needed. The coders met in person every 100 responses to resolve discrepancies and finalize coding procedures. If consensus was not met during these sessions, a senior research team member and qualitative researcher (AA) adjudicated. Coding decisions and codebook revisions were documented in an audit trail to enhance transparency and rigor. Following adjudication of any areas of discrepancy, the coders shared an additional 10 responses to review coding for interrater reliability and overall comprehensiveness. Following adjudication of any areas of discrepancy, coders independently coded the remaining responses. The coders used Microsoft Excel (Version 16.91; Microsoft Corp) to calculate interrater agreement (Cohen κ above 0.80). Strategies used to ensure reliability and validity in the qualitative data included a comprehensive audit trail, checks between coders, and biweekly team debriefings.

Ethical Considerations

The trial was approved by the institutional review board at the University of Pennsylvania (IRB protocol #: 848844), preregistered on ClinicalTrials.gov (Identifier: NCT05028075). Written informed consent was obtained from all participants. A safety protocol was approved by the institutional review board and overseen by a 3-member independent data and safety monitoring board. Participants reporting suicidal ideation through surveys or text messages were promptly contacted by the University Employee Assistance Program (EAP) for a safety

XSL•FO
RenderX

assessment, with up to three outreach attempts within 3 days. Automated safety messages with crisis resources were triggered by predetermined self-harm–related keywords. All participants received information about national suicide prevention resources and mental health support at study completion. Participants received compensation for completing the baseline survey ($50), 6-month survey ($100), and 9-month survey ($50) via electronic gift card. The study was conducted in accordance with the Declaration of Helsinki and local regulations. Participant data were deidentified and stored on encrypted servers.

Results

Overview

Of the 642 HCWs randomized to the intervention. Of the 642 Cobalt+ participants, 25.2% (162/642) visited the Cobalt platform due to the proactive text messaging, completed usability and acceptability outcomes, and were included in this analysis. Among the 162 participants who completed the outcomes, the mean age was 38.9 (SD 10.5) years, with nearly half (48.2%) aged 18-35 years. The majority were female (147/162, 90.7%), 56.8% White (92/162), 53.1% (86/162) married or living with a partner and approximately one-third engaged in shift work (56/162, 34.6%). In terms of professional roles, 27.8% were nurses, 12.4% were physicians or advanced practice providers, and 25.9% were in managerial positions. Baseline mental health symptom scores were mild, with mean PHQ-9 and GAD-7 scores of 6.5 (SD 5.3) and 6.4 (SD 4.9), respectively. There were no significant differences in terms of demographics and baseline mental health measures (eg, PHQ-9 and GAD-7) between those who completed the assessments and the overall study sample, except for gender (χ^2_{640}=8.86; P=.00), with a higher proportion of female participants and managers (χ^2_{640}=6.58; P=.04) who completed the outcome measures (Multimedia Appendix 1).

Evaluation Outcomes

Usability

The mean SUS score was 74.43 (median score 72.5). Many respondents strongly agreed or agreed to “I felt very confident using the system,” (125, 75.7%) and strongly agreed or agreed to “I think that I would like to use this system frequently,” (118, 71.5%). Most participants (222, 81%) strongly disagreed or disagreed with “I think that I would need the support of a technical person to be able to use this system,” (134, 81.2%). Despite Figure 1’s supplemental insights, validated usability rests on the total SUS score rather than individual usability domains.

XSL•FO
RenderX

Figure 1. System Usability Scale responses.

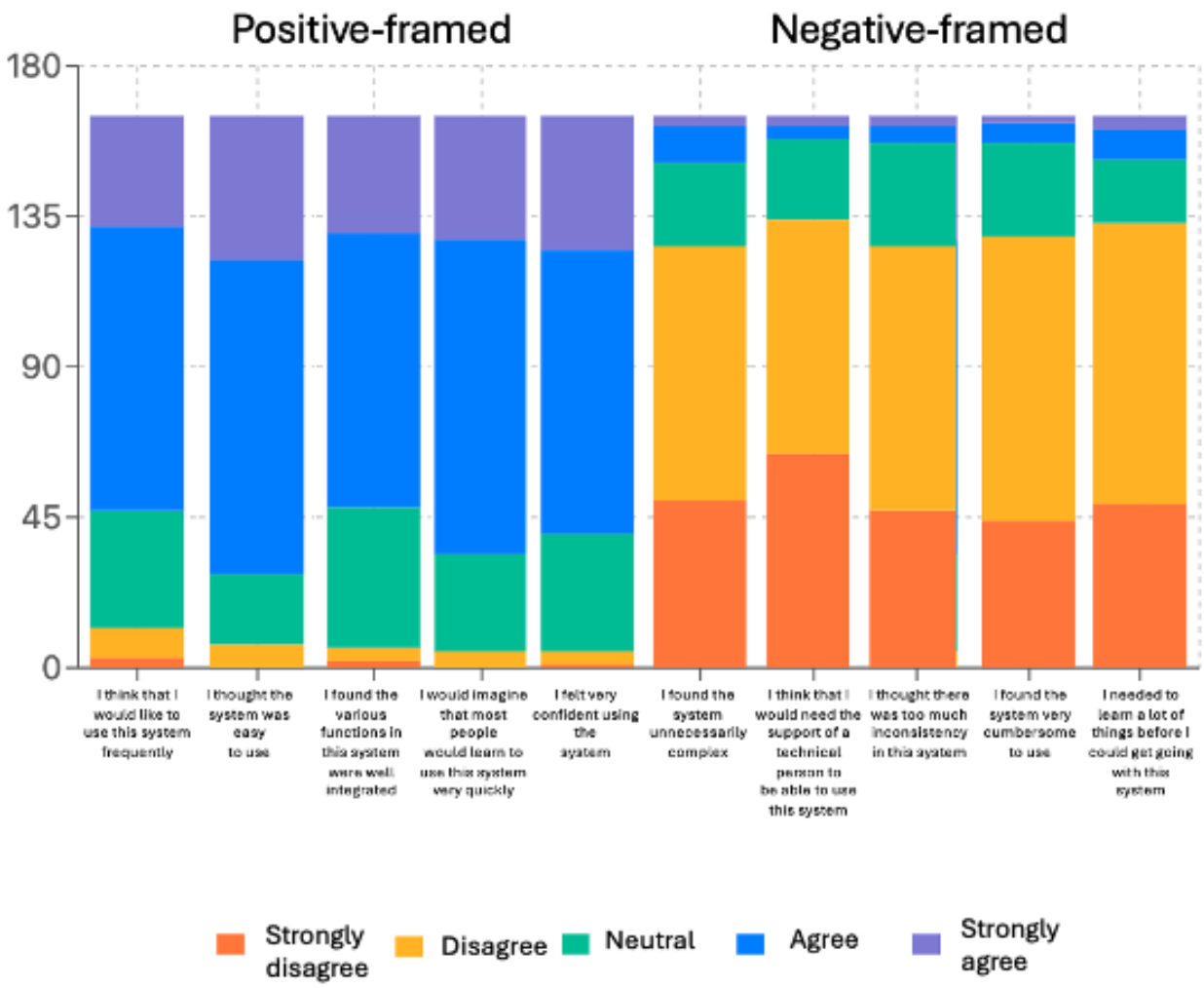

Qualitative Feedback

Participants who visited Cobalt due to a Cobalt+ text message most visited Cobalt's videos, podcasts, and articles (95/162, 58%; Table 1).

Table 1. Cobalt+ resources are the most used.

Resource type	Frequency, n (%)
Videos, podcasts, and articles	95 (58.3)
Group virtual well-being session	5 (3.1)
Individual coaching session	23 (14.1)
Appointment with a mental health professional	30 (18.4)
Other	10 (6.1)

While the open-ended comments suggest a range of uses. A few respondents explicitly stated that they did not use any resources. Some participants passively browsed out of curiosity or preliminary interest, such as "didn't use anything, just browsed," and "seeing what options are available if needed."

Acceptability

Among 162 respondents, the majority of respondents were "promoters" 45.7% (74/162, 45.9%), 22% were "neutral" (36/162, 22.2%), and 32% were "detractors" (52/162, 32.1%);

XSL•FO
RenderX

which equated to an NPS of 13.7. This NPS score and respondent breakdown suggest "good" acceptability.

Qualitative Feedback

Two categories (eg, positive and negative experiences) with 13 subcategories were identified through open-ended questions asking participants to elaborate on their satisfaction and experience receiving the proactive text messages (Table 2).

Table 2. Categories, subcategories, and illustrative quotes.

Categories and subcategories	Frequency, n (%)	Illustrative Quote
Positive experience		
Mental health	36 (16.3)	• It was a very easy way to be reminded to check in on myself. • Your text do remind me to take stock of my current feelings
Easy	35 (15.8)	• This tool is very useful and easy to access • Text messages make it easier in busy times
Positive sentiment	23 (10.4)	• I like that they reach out to me, and offer suggestions • I did not mind receiving the text with updated information and suggestions.
Care	19 (8.6)	• Nice to know there is someone there • It feels like someone understands and is thinking of you • Was not an overbearing amount of messages just reminders that you were there for me.
Helpful	15 (6.8)	• I like that they reach out to me, and offer suggestions • The reminders were helpful for me to take time to care for myself. • Helpful reminders that I can look at on my own terms
Brief	9 (4.1)	• They are concise and to the point. • Brief messages, nice to be checked on :)
Negative experience		
Negative sentiment	18 (8.1)	• Too many texts • I do not find any benefit from the text messages. They just feel like spam. • Did not find it helpful
Not applicable	16 (7.2)	• I do not always see my text in a timely mater • Little time to schedule yet more appointments
Did not use	13 (5.9)	• Did not feel the need • I already see a therapist
Ignored	12 (5.4)	• Most of the time when I received text messages I just look at them real quick and then easily forget about them.
Intrusive	7 (3.2)	• They were helpful reminders, but felt like it was in the way and obstructive to my day • Just intrusive, deal with it right away or it disappears. Rarely able to stop and deal with it. Text messages also do not come to my laptop or iPad and the site is easier to access from there
Overwhelming	3 (1.4)	• I have been so overwhelmed with everything else that I did not really take the time to read them because it seemed cumbersome

The positive experiences included general reflections on how participants felt about Cobalt+ components. Within this, subcategories emerged around messages being brief (19/162, 4%), feeling cared for (22/162, 5%), the importance of mental health support (49/162, 11%), and their preference for text message outreach (52/162, 11%). Most participants noted that the brief engagement helped prioritize their mental health. A participant said, "Your texts remind me to take stock of my current feelings." Several participants appreciated the convenience of receiving messages directly on their phones and valued the freedom to engage with them at their own pace. One participant shared that it "was not an overbearing amount of messages, just reminders that you were there for me," while another shared, "It was nice being reached out to and checked on." Participants highlighted the study's benefits, such as "I did not mind receiving the text with updated information and suggestions." Overall, participants found the messages easy to engage with (98/642, 21%); a participant said, "Text messages

XSL•FO
RenderX

make it easier in busy times," and they were often helpful, especially when delivering reminders.

The negative experiences captured critiques and challenges associated with the proactive approach. Some participants reported ignoring or not using the messages altogether, feeling that they did not align with their needs or were intrusive. A participant shared, "Most of the time when I received text messages, I just look at them real quick and then easily forget about them." Others felt the texts were "too many" or that they were "helpful reminders but [were] in the way and obstructive to my day." Furthermore, a small minority found them to be "cumbersome" because they were "so overwhelmed with everything else that I didn't really take the time to read" (Table 2).

Discussion

Principal Results

The results of the primary RCT demonstrated improved mental health outcomes among those HCWs randomized to Cobalt+, a proactive service sent via text messaging, including (1) monthly automated text messaging reminders and links to Cobalt, and (2) bimonthly text-message-based measures of depression and anxiety [16] This analysis investigates how usable and acceptable the proactive support approach was to HCWs and offers 2 main findings. First, Cobalt+ had "good" usability (SUS score=74.43). Previous literature notes that an SUS score above 68 is above average, while scores between 70 and 84 are considered "good usability" [20]. In addition to the SUS composite score, descriptive statistics for individual SUS items were provided to illustrate response patterns across usability domains. When analyzing open-ended usability feedback, the most used resources were videos, podcasts, and articles. This is an important finding as videos, podcasts, and articles are largely cost-effective and can reach a broad audience simultaneously. Second, the NPS score (n=13.7) was positive and, in a range, considered to reflect moderate to low loyalty, suggesting both promise and opportunity for improvement [22]. Despite the positive NPS score, it is essential to note that our sample had a substantial proportion of detractors (52, 32%) and this proportion should not be overlooked.

Open-ended questions provided additional context about opportunities to improve the proactive approach. Lack of time, not needing therapeutic services, or having their own mental health provider were often expressed as the main reasons for not accessing resources. However, some participants enjoyed the proactive outreach and wrap-around support. This finding speaks to the value of the proactive approach where the support and resources are centered around the user. This is particularly relevant, as many digital platforms require users to seek out resources, and mental health conditions can compromise insight, motivation, and decision-making, thereby making self-directed engagement in care. This analysis highlights the importance of "quick and easy" survey assessments and how these periodic check-ins "allow [participants] the freedom to answer when it is convenient… reminds [them] when it's time to self-reflect" and how the frequent outreach made participants feel "cared for." Yet, a smaller group found the text messaging to be intrusive. Future research could allow individuals to customize their communication type (ie, text message or email) and how often they receive notifications. Yet, this customization does not overlook implementation challenges (eg, integration across systems, resource allocation, and clinical support) [14,24], structural considerations (eg, disparities in access to digital tools), and cultural concerns (eg, stigma and trust in employer-sponsored tools outside of a research context) [25]; however, it is a first step to personalized mental health and well-being outreach, which will lay the foundation for future iterations, refinements, and adaptations to meet the HCWs' mental health and well-being needs [26].

The digital strategy to send resources and assessments to HCWs was overall found to have good usability and was modestly received. This study reveals formative and important preliminary data that remote text message engagement can extend the reach of mental health platforms for HCWs. Proactive outreach to intermittently engage, assess, and connect health care workers allows one to prioritize mental health and well-being and motivate them using existing resources [27]. This approach also maintains privacy and provides a unique strategy at scale, which can be tailored to the needs of an individual over time.

Comparison With Prior Work

This is among the first large mental health–focused RCTs to collect participants' feedback and feasibility metrics among HCWs during the pandemic and beyond. Prior research found a positive impact of an employer-sponsored behavioral health program on improved depression and anxiety symptoms through increased psychotherapy sessions [28,29]. This is a noteworthy difference compared to the Cobalt+ proactive approach and study sample for 2 reasons. First, Maeng et al [28] analyzed individuals whose baseline scores were more than or equal to 10 on depression (PHQ-9) or anxiety (GAD-7) validated measures [28], whereas the Cobalt+ sample had mild baseline PHQ-9 and GAD-7 scores (Multimedia Appendix 1). Given that Cobalt+'s sample exhibited mild baseline symptoms, there was limited potential for measurable improvement in depression and anxiety. Second, despite the Cobalt+'s samples' mild baseline scores, nearly 20% used appointments with a mental health professional (Table 1). This suggests that individuals with subthreshold forms of depression and anxiety also seek out mental health professional support.

Among digital platforms catered to HCWs, our analysis is most similar to the COVID Coach [9] and Wellness Hub [11] apps. The COVID Coach, smartphone-based app, was tested through an RCT among 40 South African physicians. It was noted to have adequate feasibility acceptability as measured by the SUS and client satisfaction questionnaire (CSQ-8). COVID Coach's SUS score was slightly higher than Cobalt+'s with 76.6 (SD 14.6). Our analysis did not use the CSQ-8 since Cobalt+ provided a variety of tools and resources and was not limited to mental health services. Nevertheless, the authors recommend future in-depth qualitative interviews to capture users' experience with the app and patterns of use [9]. In its design and implementation, Cobalt+ also closely parallels the Wellness Hub app [11], sharing similarities in its resource offering and being deployed within a large health system in the Northeast

XSL•FO
RenderX

United States. However, to our knowledge, Wellness Hub has not been extensively evaluated. Golden et al [11], reports back-end metrics such as the number of HCWs downloaded the app, the total number of baseline surveys submitted, the number of users submitting additional surveys postbaseline, and general app engagement metrics (eg, the number of app openings). Similarly, their evaluation is lacking the user experience of the app [11].

Limitations

This analysis has several limitations. First, not all study participants (24.6%) completed the month 6 assessment, and fewer completed the outcome measures (n=162). This limits our generalizability in reflecting the larger participant experience due to the limited responses. Usability and acceptability findings should be interpreted carefully due to this response bias. Second, study participants were recruited from one large, urban health system, and second, the inclusion criteria required participants to have daily access to a smartphone; this criterion may limit generalizability, especially within the context of the known racial and ethnic "digital divide." This divide may be smaller in employed health care workers, and smartphone ownership continues to expand in the United States [30]. Third, a higher proportion of female participants completed the outcome than the overall study sample measures (Multimedia Appendix 1). Interpretation of our findings and external validity may be limited for dissimilar HCW samples. Fourth, study participants were HCWs willing to complete 2 online surveys; therefore, selection and nonresponder bias may be present. Fifth, participants who opted into the study may have experienced greater distress and were likelier to prefer mobile interventions than those who declined participation. Sixth, validated survey measures lack some specification. For instance, the SUS provides a subjective, global measure of usability and unfortunately does not specify specific domains such as navigation, error prevention, and efficiency. Similarly, despite the NPS's widespread use, the acceptability thresholds are derived mainly from commercial rather than health care contexts [22]. Our analysis reported SUS and NPS scores and open-ended feedback separately. Seventh, there remains stigma associated with accessing mental health care, which may also contribute to selection bias and limited generalizability. Rather than integrating mental health records or treatments, we relied on self-reported data from participants and data on appointments made. Self-report measures have limitations due to social desirability and recall errors [31,32]. Further research is needed to continuously seek feedback from those using the program to stay relevant and responsive to ongoing and changing stressors.

Conclusions

Proactive mental health services may help overcome expected barriers to uptake of passively available services. This study demonstrated that Cobalt+, a proactive digital mental health program using proactive text messaging, is modestly usable and acceptable based on two key evaluation metrics, further supplemented by HCW feedback. These findings suggest the program's viability and the need for additional research toward improvement and broader implementation. In this study, we report on the insights gained using simple, scalable text messaging to reach a large group of diverse HCWs. Future studies are needed to test this novel approach across clinical settings and professional roles and in large health systems.

Acknowledgments

Meghana Sharma for her work coding the open-ended responses and Art Pelullo for conducting back-end data extraction and analyses. We would also like to thank Michael Josephs, Neda Khan, and Michael Kopinsky for building the intervention on Way to Health.

Funding

National Institute of Neurological Disorders and Stroke (NINDS), National Institute of Mental Health (NIMH), National Institute on Minority Health and Health Disparities (NIMHD; award number: 1R01MH127686-01). The funder had no role in the study design; collection, analysis, or interpretation of data; writing of the manuscript; or decision to submit it for publication. None of the authors serve on the JMIR Publications Editorial Board (including Editors-in-Chief, Associate Editors, or Editorial Board members). None of the authors have acted as an expert witness in any relevant legal proceedings.

Conflicts of Interest

Penn Medicine Cobalt resources and delivery system was developed by inventors at Penn Medicine and if it were to be successful, the University of Pennsylvania would benefit financially. Co-author DAA has Clinical Trial Intellectual Property and Financial Interest relating to this research. The University requires written notification to team members and that we included the below statement in our informed consent document, "Penn Medicine Cobalt resources and delivery system used as a part of this protocol was developed by inventors at Penn Medicine and if it were to be successful, the University of Pennsylvania would benefit financially." To address this, we restricted DAA's involvement in the actual conduct of this clinical trial, such as subject selection, enrollment, the consent process, data collection, cleaning or adjudication of primary or secondary outcomes, including whether adverse events are related to the trial. DAA participated in study design and as an author on publications resulting from this study and participated in data analysis, interpretation, and reporting of data, but only after the analytic dataset is final ("locked").

Multimedia Appendix 1

XSL•FO
RenderX

Supplementary material.
[DOCX File , 18 KB - formative_v9i1e74086_app1.docx]

References

1. Tawfik DS, Profit J, Morgenthaler TI, Satele DV, Sinsky CA, Dyrbye LN, et al. Physician burnout, well-being, and work unit safety grades in relationship to reported medical errors. Mayo Clin Proc 2018;93(11):1571-1580 [FREE Full text] [doi: 10.1016/j.mayocp.2018.05.014] [Medline: 30001832]
2. Sexton JB, Adair KC. Well-being outcomes of health care workers after a 5-hour continuing education intervention: the WELL-B randomized clinical trial. JAMA Netw Open 2024;7(9):e2434362 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.34362] [Medline: 39298170]
3. Bansal A, Trieu M, Eriksson EM, Zhou F, McVernon J, Brokstad KA, et al. SARS-CoV-2 infection rates and associated risk factors in healthcare workers: systematic review and meta-analysis. Sci Rep 2025;15(1):4705 [FREE Full text] [doi: 10.1038/s41598-025-89472-5] [Medline: 39922967]
4. Fantus S, Cole R, Usset T, Hawkins L. Multi-professional perspectives to reduce moral distress: a qualitative investigation. Nurs Ethics 2024;31(8):1513-1523 [FREE Full text] [doi: 10.1177/09697330241230519] [Medline: 38317421]
5. DePierro J, Katz CL, Marin D, Feder A, Bevilacqua L, Sharma V, et al. Mount sinai's center for stress, resilience and personal growth as a model for responding to the impact of COVID-19 on health care workers. Psychiatry Res 2020;293:113426 [FREE Full text] [doi: 10.1016/j.psychres.2020.113426] [Medline: 32861094]
6. DePierro JM, Chan CC, Mohamed N, Starkweather S, Ripp J, Peccoralo LA. Determinants of staff intent to leave health care during the COVID-19 pandemic. Am J Public Health 2024;114(S2):200-203. [doi: 10.2105/AJPH.2024.307574] [Medline: 38354356]
7. Ameli R, Sinaii N, West CP, Luna MJ, Panahi S, Zoosman M, et al. Effect of a brief mindfulness-based program on stress in health care professionals at a us biomedical research hospital: a randomized clinical trial. JAMA Netw Open 2020;3(8):e2013424 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.13424] [Medline: 32840621]
8. Stone RSB. Code lavender: a tool for staff support. Nursing 2018;48(4):15-17. [doi: 10.1097/01.NURSE.0000531022.93707.08] [Medline: 29561364]
9. Jaworski BK, Taylor K, Ramsey KM, Heinz A, Steinmetz S, Pagano I, et al. Exploring usage of COVID coach, a public mental health app designed for the COVID-19 pandemic: evaluation of analytics data. J Med Internet Res 2021;23(3):e26559 [FREE Full text] [doi: 10.2196/26559] [Medline: 33606656]
10. Bondar J, Babich Morrow C, Gueorguieva R, Brown M, Hawrilenko M, Krystal JH, et al. Clinical and financial outcomes associated with a workplace mental health program before and during the COVID-19 pandemic. JAMA Netw Open 2022;5(6):e2216349 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.16349] [Medline: 35679044]
11. Golden EA, Zweig M, Danieletto M, Landell K, Nadkarni G, Bottinger E, et al. A resilience-building app to support the mental health of health care workers in the COVID-19 era: design process, distribution, and evaluation. JMIR Form Res 2021;5(5):e26590 [FREE Full text] [doi: 10.2196/26590] [Medline: 33872189]
12. Livesey C, Kugler K, Huang JJ, Burton E, Rangnekar A, Vojta G, et al. COBALT: supporting the mental well-being of the health care workforce with technology-facilitated care during Covid-19 and beyond. Healthc (Amst) 2022;10(3):100640 [FREE Full text] [doi: 10.1016/j.hjdsi.2022.100640] [Medline: 35841839]
13. Albott CS, Wozniak JR, McGlinch BP, Wall MH, Gold BS, Vinogradov S. Battle buddies: rapid deployment of a psychological resilience intervention for health care workers during the covid-19 pandemic. Anesth Analg 2020;131(1):43-54 [FREE Full text] [doi: 10.1213/ANE.0000000000004912] [Medline: 32345861]
14. Graham AK, Lattie EG, Powell BJ, Lyon AR, Smith JD, Schueller SM, et al. Implementation strategies for digital mental health interventions in health care settings. Am Psychol 2020;75(8):1080-1092 [FREE Full text] [doi: 10.1037/amp0000686] [Medline: 33252946]
15. Tong ACY, Wong KTY, Chung WWT, Mak WWS. Effectiveness of topic-based chatbots on mental health self-care and mental well-being: randomized controlled trial. J Med Internet Res 2025;27:e70436 [FREE Full text] [doi: 10.2196/70436] [Medline: 40306635]
16. Agarwal AK, Southwick L, Gonzales RE, Bellini LM, Asch DA, Shea JA, et al. Digital engagement strategy and health care worker mental health: a randomized clinical trial. JAMA Netw Open 2024;7(5):e2410994 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.10994] [Medline: 38787562]
17. Asch DA, Volpp KG. On the way to health. LDI Issue Brief 2012;17(9):1-4. [Medline: 22934330]
18. Kroenke K, Spitzer RL, Williams JBW. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
19. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
20. Brooke J. Sus - A Quick and Dirty Usability Scale. Milton Park: Taylor and Francis group; 1996.
21. Duggan MJ, Gervase J, Schoenbaum A, Hanson W, Howell JT, Sheinberg M, et al. Clinician experiences with ambient scribe technology to assist with documentation burden and efficiency. JAMA Netw Open 2025;8(2):e2460637 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.60637] [Medline: 39969880]

XSL•FO
RenderX

22. Lucero KS. Net promoter score (NPS): what does net promoter score offer in the evaluation of continuing medical education? J Eur CME 2022;11(1):2152941 [FREE Full text] [doi: 10.1080/21614083.2022.2152941] [Medline: 36465493]
23. Alismail A, Schaeffer B, Oh A, Hamiduzzaman S, Daher N, Song H, et al. The use of the net promoter score (NPS) in an outpatient allergy and pulmonary clinic: an innovative look into using tablet-based tool vs traditional survey method. Patient Relat Outcome Meas 2020;11:137-142 [FREE Full text] [doi: 10.2147/PROM.S248431] [Medline: 32547281]
24. Berardi C, Antonini M, Jordan Z, Wechtler H, Paolucci F, Hinwood M. Barriers and facilitators to the implementation of digital technologies in mental health systems: a qualitative systematic review to inform a policy framework. BMC Health Serv Res 2024;24(1):243 [FREE Full text] [doi: 10.1186/s12913-023-10536-1] [Medline: 38408938]
25. Sevov B, Huettemann R, Zinner M, Meister S, Fehring L. Employee preference and use of employee mental health programs: mixed methods study. JMIR Hum Factors 2025;12:e65750 [FREE Full text] [doi: 10.2196/65750] [Medline: 40324169]
26. Hornstein S, Zantvoort K, Lueken U, Funk B, Hilbert K. Personalization strategies in digital mental health interventions: a systematic review and conceptual framework for depressive symptoms. Front Digit Health 2023;5:1170002 [FREE Full text] [doi: 10.3389/fdgth.2023.1170002] [Medline: 37283721]
27. Morrison LG, Hargood C, Pejovic V, Geraghty AWA, Lloyd S, Goodman N, et al. The effect of timing and frequency of push notifications on usage of a smartphone-based stress management intervention: an exploratory trial. PLoS One 2017;12(1):e0169162 [FREE Full text] [doi: 10.1371/journal.pone.0169162] [Medline: 28046034]
28. Maeng D, Cornell AE, Nasra GS. Utilization of an employee behavioral health program and its effects on outcomes for depression and anxiety disorders. J Occup Environ Med 2019;61(10):812-817. [doi: 10.1097/JOM.0000000000001678] [Medline: 31425323]
29. Wang PS, Simon GE, Avorn J, Azocar F, Ludman EJ, McCulloch J, et al. Telephone screening, outreach, and care management for depressed workers and impact on clinical and work productivity outcomes: a randomized controlled trial. JAMA 2007;298(12):1401-1411 [FREE Full text] [doi: 10.1001/jama.298.12.1401] [Medline: 17895456]
30. Saeed SA, Masters RM. Disparities in health care and the digital divide. Curr Psychiatry Rep 2021;23(9):61 [FREE Full text] [doi: 10.1007/s11920-021-01274-4] [Medline: 34297202]
31. Paulhus DL. Two-component models of socially desirable responding. J. Pers Soc Psychol 1984;46(3):598-609. [doi: 10.1037/0022-3514.46.3.598]
32. Definition of recall bias - NCI Dictionary of Cancer Terms. NCI. URL: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/recall-bias [accessed 2025-09-09]

Abbreviations

CSQ-8: client satisfaction questionnaire-8
GAD-7: Generalized Anxiety Disorder-7
HCW: health care worker
HIPAA: Health Insurance Portability and Accountability Act
NPS: Net Promoter Score
PHQ-9: Patient Health Questionnaire-9
RCT: randomized controlled trial
SUS: System Usability Scale

Edited by J Sarvestan; submitted 17.03.25; peer-reviewed by H Holmen; comments to author 27.08.25; revised version received 12.09.25; accepted 13.11.25; published 09.12.25.

Please cite as:
Southwick L, Gonzales R, Bellini L, Asch DA, Mitra N, Balachandran M, Wolk CB, Becker-Haimes EM, Kishton R, Beck S, Merchant RM, Agarwal AK
Proactively Delivered Digital Mental Health Support for Health Care Workers: Usability and Acceptability Evaluation
JMIR Form Res 2025;9:e74086
URL: https://formative.jmir.org/2025/1/e74086
doi:10.2196/74086
PMID:

©Lauren Southwick, Rachel Gonzales, Lisa Bellini, David A Asch, Nandita Mitra, Mohan Balachandran, Courtney Benjamin Wolk, Emily M Becker-Haimes, Rachel Kishton, Sarah Beck, Raina M Merchant, Anisk K Agarwal. Originally published in JMIR Formative Research (https://formative.jmir.org), 09.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly

XSL•FO
RenderX

cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Localized Muscular Fatigue in Robotic-Assisted Laparoscopic Surgery: Predictive Modeling Study

Daniel Caballero[1], PhD; Manuel J Pérez-Salazar[1], MSc; Juan A Sánchez-Margallo[1], PhD; Francisco M Sánchez-Margallo[2], PhD

[1]Bioengineering and Health Technologies Unit, Jesús Usón Minimally Invasive Surgery Centre, Cáceres, Cáceres, Spain

[2]Scientific Direction, Jesús Usón Minimally Invasive Surgery Centre, Cáceres, Cáceres, Spain

Corresponding Author:
Juan A Sánchez-Margallo, PhD
Bioengineering and Health Technologies Unit
Jesús Usón Minimally Invasive Surgery Centre
Carretera N-521, km. 41.8
Cáceres, Cáceres, 10071
Spain
Phone: 34 927 18 10 32
Email: jasanchez@ccmijesususon.com

Abstract

Background: Robotic-assisted surgery (RAS) has grown rapidly in recent decades, and several RAS procedures have become the standard. However, the physical and mental demands of minimally invasive surgery (MIS) techniques can lead to ergonomic shortcomings for surgeons. Advances in wearable technology and artificial intelligence favor the development of innovative solutions to analyze and improve ergonomic conditions during surgical practice.

Objective: The main objective is the development and validation of a predictive model of localized muscle fatigue from electromyography (EMG) data during conventional laparoscopic surgery (LAP) and RAS.

Methods: Four different tasks were performed on LAP and RAS: dissection, labyrinth, peg transfer, and suturing. A wireless EMG sensor system was used to record muscle activity. Joint analysis of the spectrum and analysis graphs was used to evaluate the localized muscle fatigue. A dataset was generated for each task as a function of surgeons' expertise level and surgical type. Each dataset was scaled as preprocessing and divided into 2 datasets: 80% for training and 20% for testing. Multiple linear regression (MLR) and multilayer perceptron (MLP) were applied as predictive techniques and validated on all test datasets. R^2 coefficient and root-mean-square error were used to measure the accuracy of the models.

Results: RAS showed less muscle fatigue for novice surgeons compared to LAP practice, although it was higher for expert surgeons. The predictive model achieved satisfactory R^2 and root-mean-square error coefficients for all parameters extracted from the EMG signal, predicting with high accuracy localized muscle fatigue values. The MLR predictive model demonstrated superior performance relative to the MLP model.

Conclusions: Wearable technology and artificial intelligence techniques have been successfully applied for the development and validation of a novel predictive model based on MLR and MLP to predict localized muscle fatigue in MIS.

(JMIR Form Res 2025;9:e68536) doi:10.2196/68536

KEYWORDS

laparoscopic surgery; robotic-assisted surgery; artificial intelligence; predictive techniques; electromyography; localized muscle fatigue

Introduction

Background

Apart from the well-known advantages of conventional laparoscopic surgery (LAP) and robotic-assisted surgery (RAS) for the patient, these techniques have a steep learning curve and are physically and mentally demanding for surgeons, as well as presenting ergonomic challenges. Consequently, this mental and physical burden, combined with potential ergonomic deficiencies in the working environment, can lead to musculoskeletal problems in surgeons [1]. Although ergonomic

XSL•FO
RenderX

conditions are considered to have improved in RAS compared to LAP, mainly for the primary surgeon operating from the console, scientific evidence remains scarce [2]. Therefore, further ergonomic analysis studies are needed to thoroughly investigate the ergonomic conditions of the surgeon during RAS, identify potential deficiencies, and develop possible solutions and recommendations.

Minimally invasive surgery (MIS) and, in particular, RAS have grown rapidly in recent decades, and several robotic-assisted procedures have become standard surgical techniques [3]. The advantages of MIS and RAS for patients are widely known, including reduced tissue trauma, better cosmetic outcomes, and reduced hospital stay [4]. However, these techniques present some limitations for the surgeon, such as possible ergonomic deficiencies during long surgeries, or high levels of stress in certain procedures [5], which can be detrimental to the surgeon's health and have an impact on the quality of surgical procedures and patient care [6].

Surgical robotics offers greater surgical precision, better visual perception by allowing 3D images to be obtained compared to the 2D images of LAP, and better ergonomics for the surgeon, who can operate in a seated position. There are some traditional methods to assess the workload, both mental and physical, of surgeons during their surgical activity, mainly oriented to subjective evaluations such as, for example, the SURG-TLX scales [7]. However, the evolution and miniaturization of sensors have allowed the increasing incorporation of wearable technology in ergonomic and physiological analysis in the surgical setting, which has made it possible to offer objective solutions without interrupting the surgeons' surgical practice. Some of these objective parameters are the surgeon's posture through body movement techniques, force/torque analysis [8], muscle activity through electromyography (EMG) signal analysis [9], and stress level through electrocardiogram examination or electrodermal activity (EDA) signals [10].

The application of artificial intelligence (AI) has grown exponentially in its use and development. These algorithms are based on nontrivial processes to discover potentially useful knowledge initially hidden in the data [11,12]. Among the different AI techniques, there are several algorithms that allow the development of predictive models [13,14]. Thus, the use of these predictive models in the field of ergonomics during surgical practice offers us a wide range of possibilities to predict those risk situations related to the surgeon's health, for example, inadequate posture, the appearance of muscle fatigue, or high levels of stress. Being able to predict and, therefore, reduce these risk situations would potentially improve the health of surgeons and, consequently, the quality of surgical practice.

Related Works

Several studies have been presented focusing on the ergonomic analysis in RAS. Some of these studies use subjective surveys [15], demonstrating that 56.1% of regularly practicing robotic surgeons continue to experience related physical symptoms or discomfort, including neck stiffness, finger and eye fatigue, among the most common.

Other studies analyzed the surgeon's posture at the console during surgical practice using traditional photogrammetry methodologies [16]. Ergonomically risky positioning of the neck and elbow was observed in medical students, and robotic experts showed risky positioning in the knee and hip region. Another study objectively assessed ergonomics in surgical robotics using rapid entire body assessment and the rapid upper limb assessment to quantify ergonomic effectiveness [17]. An Xbox Connect camera was used to record the surgeon's posture at the console. The results indicated a medium risk, with the recommendation that measures need to be taken to improve surgeon ergonomics. However, these photogrammetry-based methods are prone to occlusion problems in such cumbersome environments as operating rooms.

In order to deal with the above limitations, wearable systems such as inertial measurement unit–based systems are used. These systems continuously track the surgeon's movements without interfering with the sterile environment. In a study during the performance of robotic prostatectomies, it was concluded that the console can limit postures, causing static loads that have been associated with musculoskeletal symptoms for the surgeon's neck, torso, and shoulders [18].

Other studies have objectively analyzed the surgeon's muscle loading using surface EMG systems [9]. The results indicated that laparoscopic practice presented more muscle fatigue in the forearm compared to robotic-assisted laparoscopic surgery. Other studies analyzed physiological parameters to examine the surgeon's health, such as heart rate as a possible indicator of stress [16]. The highest level of stress was obtained in surgical residents. In previous work, we used different wearable technologies to record the posture, muscle activity, EDA, and electrocardiographic signal of surgeons during robotic-assisted surgical practice [19]. The results indicated that robot-assisted procedures showed better ergonomic outcomes for the surgeon compared to LAP.

Regarding the use of predictive models in ergonomics in MIS based on AI techniques, in previous studies, we developed predictive models to predict stress from ergonomic (kinematic, body posture, and angular rate) and physiological (EDA, blood pressure, and body temperature) parameters of the surgeon [5] during RAS procedures. The linear models (multiple linear regression [MLR]) proposed in this study were successfully validated, improving the results achieved by the nonlinear models (support vector machine, multilayer perceptron [MLP]), demonstrating the possibility of predicting factors that help us to improve the surgeon's health during robotic surgery.

Other studies analyzed and compared the surgeon's posture between RAS and LAP techniques using motion analysis technology with inertial sensors and developed a predictive model of ergonomically inappropriate postures for the surgeon [20]. MLR and MLP were compared, with the best results obtained when applying MLR as a predictive technique.

In this study, the predictive model developed would allow the evaluation of localized muscle fatigue for RAS using MLR and MLP. To do this, we will use novel wearable technologies in the surgical field in order to study the surgeon's localized muscle fatigue and to develop predictive models based on AI that enable

XSL•FO
RenderX

predicting situations and factors that lead to increased muscle fatigue in the field of MIS, both in LAP surgery and RAS.

This work represents a significant step in the application of cutting-edge technology for the improvement of ergonomic conditions in the surgical field, with the aim of improving the surgeon's health during surgery, contributing to the improvement of surgical performance and, consequently, the quality of patient care.

Objective

The main objective of this study is to develop and validate a predictive model of localized muscle fatigue in surgeons based on EMG sensor data during conventional and robot-assisted laparoscopic surgery, using AI techniques (MLP and MLR) and considering different levels of experience (experienced and novice surgeons). The hypothesis of this study is that it would be possible to predict the localized muscle fatigue in MIS by combining wearable technology and AI techniques, with RAS surgery leading to less muscle fatigue for novice surgeons, who would be less conditioned by the habit of LAP practice.

In addition, this study seeks to take a step forward in the prevention and improvement of ergonomic conditions of surgeons, to seek the improvement of their health and, consequently, the quality of surgery and health care of patients.

Methods

Experimental Design

In this study, 8 different surgeons with different levels of experience in LAP performed 4 different surgical tasks by LAP and RAS, all of them on a task-simulator training model during 25 surgical sessions.

For each surgical session, the EMG signal was recorded by using a wireless system (EMG TRIGNO Avanti System, Delsys, Natick, Massachusetts, USA) on 7 muscle groups bilaterally: Brachioradialis, Brachii Triceps, Erector Spinae, Gastrocnemius Medial, Middle Trapezius, Upper Trapezius, and Vastus Lateralis.

From these data, a dataset (140,198 records) was generated, on which a preprocessing technique (scaling) has been applied, and then, 2 datasets were generated: 80% of the data for the training dataset and the cross-validation of 10-folds (112,158 records), and the remaining 20% of the data for the test dataset (28,040 records). Records from all 25 sessions appeared in both datasets.

In this study, 2 predictive AI approaches were applied to the training dataset to generate the predictive models: MLP and MLR.

Finally, to validate the generated predictive models, they were validated on the 10-fold cross-validation and test dataset. Figure 1 shows the experimental design followed by this study.

Figure 1. Experimental design. EMG: electromyography; LAP: conventional laparoscopic surgery; RAS: robotic-assisted surgery.

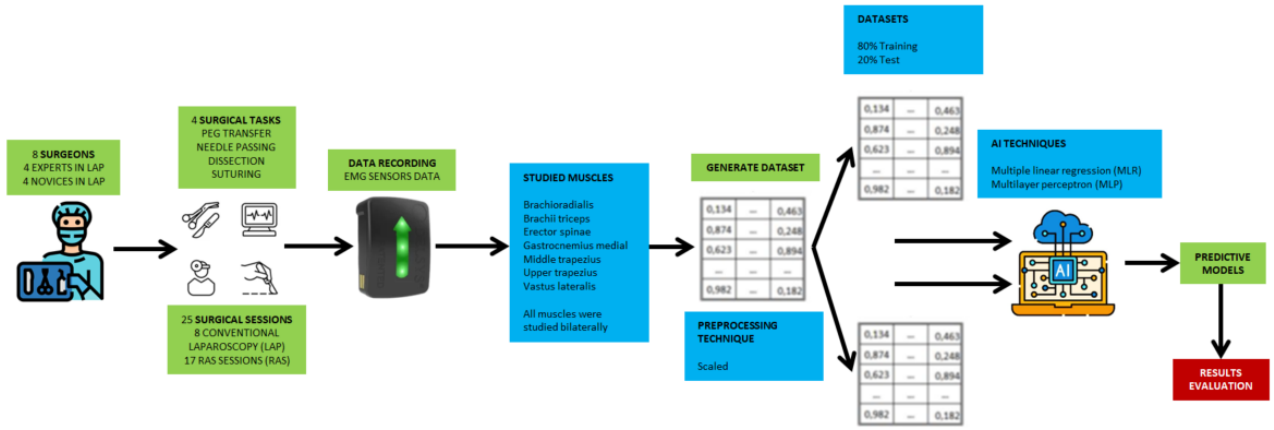

Experimental Setup

Overview

In this study, 4 different surgical activities were performed both by LAP and RAS on a task-simulator training model. For LAP (conventional laparoscopic) practice, the Olympus Visera III System (Shinjuku, Tokyo, Japan) and Karl Storz instruments (Karl Storz SE & Co. KG, Tuttlingen, Germany) were used. For RAS (robotic-assisted laparoscopic surgery), the Versius Surgical System (CMR Surgical, Cambridge, United Kingdom) was used. It is a modular robotic platform with an open console and 3D vision.

In this study, 4 different laparoscopic tasks were performed.

Peg Transfer (Hand-Eye Coordination)

This task consists of moving rubber pieces with the shape of an elongated toroid from one point to another point, changing the piece from one hand to another hand [21]. For that, 2 fenestrated forceps or a Maryland dissector and a fenestrated forceps were used (Figure 2A). It is considered a repetition when it takes the 3 pieces to the 3 goal points. Participants were asked to complete 2 repetitions with a limit of 10 minutes.

XSL•FO
RenderX

Figure 2. Simulator tasks performed in this study: (A) peg transfer (hand-eye coordination), (B) labyrinth (needle passing), (C) dissection in organic tissue, and (D) suturing.

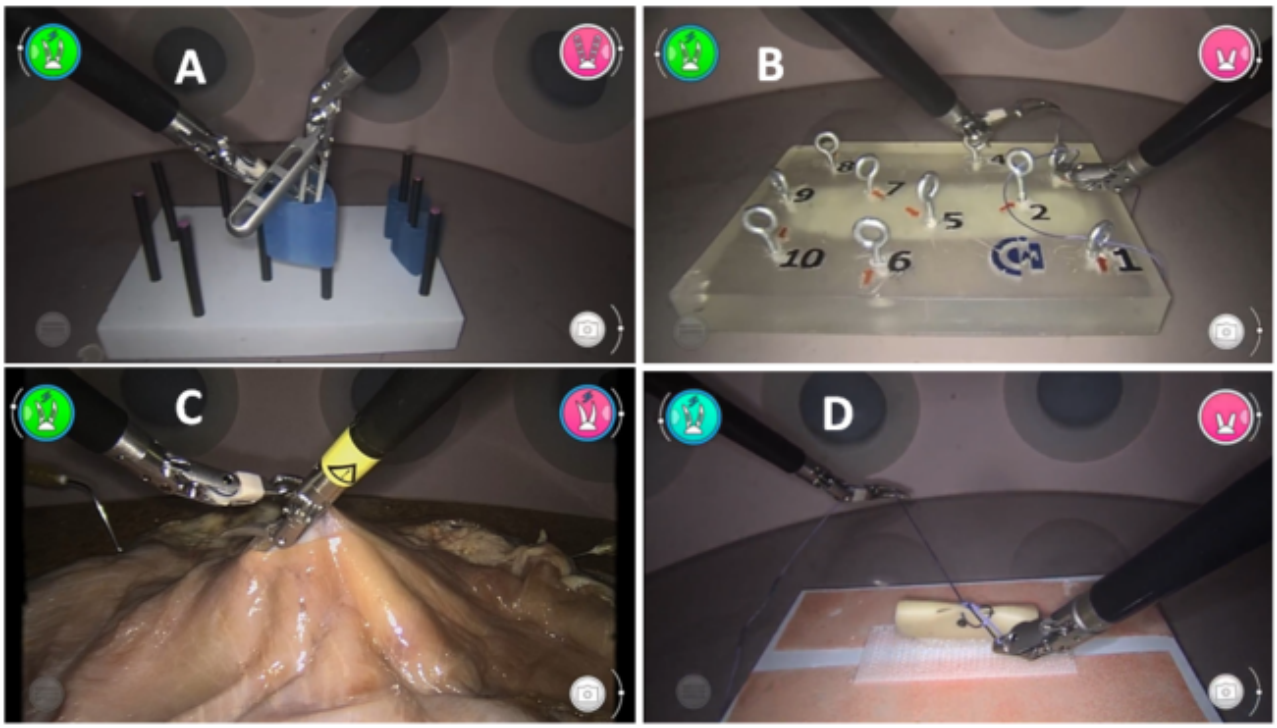

Labyrinth (Needle Passing)

In this task, it is necessary to thread a needle through a circuit with holes, in order and in different directions. This task aims to force uncomfortable postures, especially in the wrists, to evaluate the surgeon's skill and the ease of returning to a correct posture [22]. It is necessary to insert the needle with the dominant hand and remove it from the other side with the nondominant hand. A needle holder was used in the dominant hand and a Maryland dissector in the nondominant hand (Figure 2B). Participants were asked to complete the entire circuit within 10 minutes.

Dissection in Organic Tissue

An ex vivo porcine stomach was used for this task. A 3-cm-long incision was made, and the participant was asked to dissect the serous layer from the muscular layer [23]. A scissor was used in the dominant hand and a Maryland dissector in the nondominant hand (Figure 2C). A time limit of 5 minutes was set for this task.

Suturing

Participants were asked to perform a suture on the previously performed dissection. The suture consisted of 2 simple sutures and one double suture in opposite directions [24] (Figure 2D). A needle holder was used in each hand, with a time limit of 10 minutes.

Ethical Considerations

Written informed consent was taken and signed by all participants for the record and use of data generated in this study. In this informed consent, the participants could select to opt out of the study or to participate.

As this research focused only on recording the data that did not contain identifying information and posed no additional risk to subjects and participants, its activities were deemed exempt from ethics committee review. This study was conducted under the guidelines of the Spanish normative 14/2007, of July 3, on Biomedical research [25], since this study does not involve human biological samples or critical data that allow the identification of the participants, and it is not invasive for the human participants.

Data Recording

The wireless EMG TRIGNO Avanti System from DELSYS was used to record the muscle activity by means of EMG signals (Figure 3). This system has up to 16 sensors with a sampling rate of 1024 Hz. The EMG signal of the following 7 muscle groups was recorded bilaterally: Brachioradialis, Brachii Triceps, Erector Spinae, Gastrocnemius Medial, Middle Trapezius, Upper Trapezius, and Vastus Lateralis. The EMG sensors were placed on each muscle according to the SENIAM guidelines [26]. These guidelines are widely used in scientific literature, providing detailed recommendations for sensor placement in more than 30 muscles. Among the guidelines, it can be found longitudinal and transverse location, sensor orientation, and interelectrode distance, among others. Prior to the placement of each sensor, the skin was cleaned by gently rubbing it with 70% isopropyl alcohol.

Raw EMG signals were initially processed using a band-pass filter of 20-300 Hz. The filtered EMG was then smoothed with a moving window of 125 milliseconds and calculated as root-mean-square in order to remove artifacts and noise. To normalize the results for each subject, the EMG values were presented as a percentage of maximum voluntary contraction (%MVC). %MVC was computed separately for each muscle

XSL•FO
RenderX

group, just before each test, by asking each participant to perform specific contractions against a fixed resistance. The MVC was performed separately for each muscle group just before each test, asking each subject to perform specific contractions against a fixed resistance. These contractions are performed with the aim of measuring the maximum effort of each muscle group and normalizing the EMG signals of each subject.

Figure 3. Surgical setting performing labyrinth task on robotic-assisted surgery.

Data Analysis

Overview

A total of 8 participants performed the 4 laparoscopic tasks using both LAP and RAS techniques. EMG data were gathered from all sessions. A dataset was generated for each task and as a function of the expertise level of the surgeons and the type of surgery, for a total of 16 datasets. The %MVC was computed for each dataset. Muscular fatigue was quantified using the joint analysis of the spectrum and amplitude method [27].

Predictive Model

From all EMG data collected, 16 datasets were generated as a function of the surgical task (peg transfer, needle passing, organic tissue dissection, and suturing), the surgeon's level of experience (novice and experienced), and the type of surgical technique (LAP and RAS), with 140,198 records. After that, the original datasets were transformed by applying a scaling as a preprocessing technique. The scaled preprocessing technique allows describing each %MVC of each muscle in the scale between 0 and 1 [28]. For this purpose, each value is subtracted by the minimum value and then divided by the interval between the maximum and the minimum values.

The 16 preprocessed datasets were divided into 32 datasets: 80% of the data for the calibration and training (112,158 records) and 20% for validation and test (28,040 records) purpose [29]. On all calibration and test datasets, data from each participant were added. On all calibration subsets, 2 AI techniques were applied to generate predictive models: MLR and MLP. The free software WEKA (Waikato Environment for Knowledge Analysis; University of Waikato, New Zealand) [30] was used for implementing the predictive models.

MLR is a predictive technique that allows the creation of future models that can be predicted from current data by trend analysis [31]. In this study, MLR was applied as a predictive linear approach on the datasets. MLR was performed, showing the linear relationship between a dependent variable and several independent variables. This technique reaches a linear regression equation that can be used to predict future values.

XSL•FO
RenderX

The M5 method of attribute selection was applied in this study. This method steps through the attributes, removing the one with the smallest standardized coefficient until no improvement in the estimation of the error is observed. In addition, a ridge value of 1.0×10^{-4} was applied [32].

MLP was applied as a predictive machine learning approach to datasets. MLPs are a type of artificial neural network model that are developed by using principles of neuronal organization [33]. Thus, different numbers of neurons are aggregated into layers. Different layers may perform different transformations on their inputs. Signals travel from the first layer (the input layer) to the last layer (the output layer), possibly after traversing the layers multiple times [34]. The default configuration was used in this study, with the learning rate equal to 0.3, the number of epochs equal to 500, the threshold equal to 20, and 30 nodes in the first hidden layer, 10 nodes in the second hidden layer, and 3 nodes in the third hidden layer.

Finally, to validate the generated predictive models, they were validated in 2 steps. First, a 10-fold cross-validation, in which the calibration dataset was divided into 10 equally subset, was performed. Each time 1 subset was used to validate, the remaining 9 subsets were used to fit the model. The process was repeated sequentially until all subsets were validated. Therefore, all data were used for both purposes, calibration and validation, making it a very robust method for cross-validation [35]. Finally, the test dataset was used for external validation of the generated predictive models.

The R^2 coefficient was used to evaluate the goodness of fit of the prediction and for its validation, according to the rules given by Colton [36]. This considers R^2 from 0 to 0.25 as a poor to null relationship; from 0.25 to 0.50 indicates a weak degree of relationship; from 0.50 to 0.75 designates a moderate to good relationship; and from 0.75 to 1 shows a very good to excellent relationship.

The root-mean-square error (RMSE) was also used to validate the prediction results. The RMSE measures the difference between actual and predicted values. RMSE values less than 0.05 are considered appropriate [37].

Results

Overview

Eight surgeons participated in this study: 4 experts in laparoscopic surgery (> 100 laparoscopic surgeries performed; 1.80, SD 0.08 m tall and 36.5, SD 11.03 years old) and 4 novices in laparoscopic surgery (< 100 laparoscopic surgeries performed; 1.82, SD 0.04 m tall and 28.5, SD 7.59 years old). All participants had no previous experience in robotic-assisted laparoscopic surgery. They performed 8 sessions by means of LAP techniques and 17 sessions using RAS techniques.

The results show the comparison of kinematic data as a function of the surgical technique and the level of expertise of the surgeon, as well as the goodness of the predictive models to predict the localized muscle fatigue for each combination of surgical technique and level of expertise on each muscle group analyzed. These results demonstrate the influence of previous experience in LAP (without previous experience in RAS) on the onset of localized muscle fatigue in robot-assisted LAP practice [38].

Kinematic Analysis

Comparative as a Function of the Surgical Technique

In general, the erector spinae, middle trapezius, and upper trapezius were the muscles that required the most muscle activity (Figure 4). Similarly, in the case of expert surgeons, for the labyrinth, peg transfer, and suturing tasks, they had lower muscle activity in the muscles on the right side, usually the dominant side, (Figure 4), and for the dissection task, they had lower muscle activity in the muscles on the left side (Figure 4). For the novice surgeons, the peg transfer tasks had lower muscle activity in the muscles on the right side (Figure 4), and for the remaining tasks, lower muscle activity in the muscles on the left side (Figure 4).

The analysis of localized muscle fatigue showed greater use of force and greater muscle fatigue for RAS in expert surgeons (Figure 5). In the case of novice surgeons, the results were more balanced (Figure 5). However, during LAP, the use of force and localized muscle fatigue is higher, being remarkable for the suturing task (Figure 5).

XSL•FO
RenderX

Figure 4. Comparison of muscle activity (percentage of maximum voluntary contraction) of (A) experienced and (B) novice surgeons during performance of 4 simulator tasks, from upper to bottom: dissection, labyrinth, peg transfer, and suturing. Using conventional laparoscopic surgery (red) and robotic-assisted (blue) surgery for the following muscle groups bilaterally: Erector spinae (ER_SPIN), Middle trapezius (MID_TRAP), and Upper trapezius (UP_TRAP).

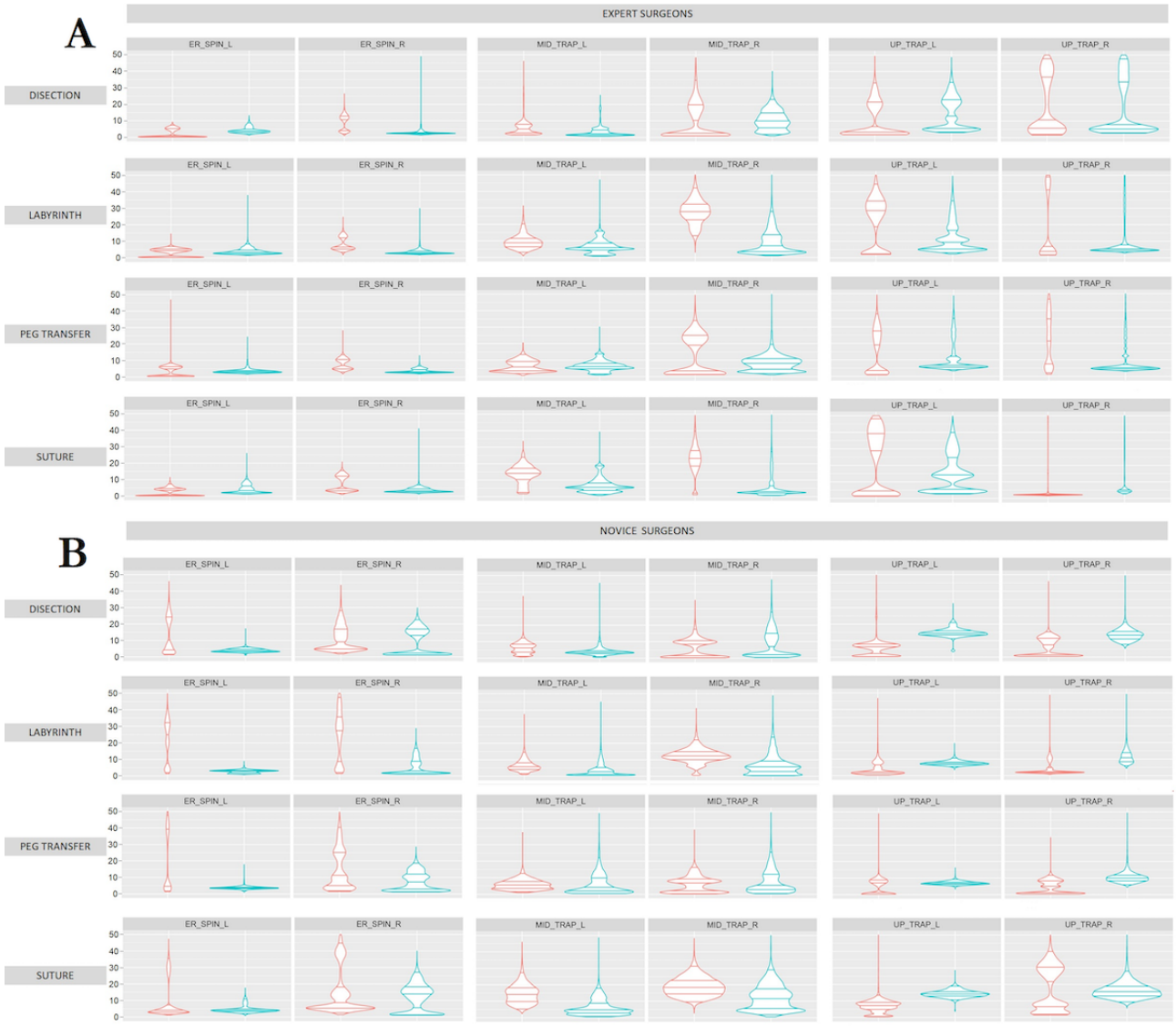

XSL•FO
RenderX

Figure 5. Comparison of fatigue/recovery muscle localized and muscle strength increase/decrease for (A) experienced and (B) novice surgeons on 4 simulator tasks, from upper to bottom: dissection, labyrinth, peg transfer, and suturing. Performing conventional laparoscopic (red) and robotic-assisted (blue) surgeries.

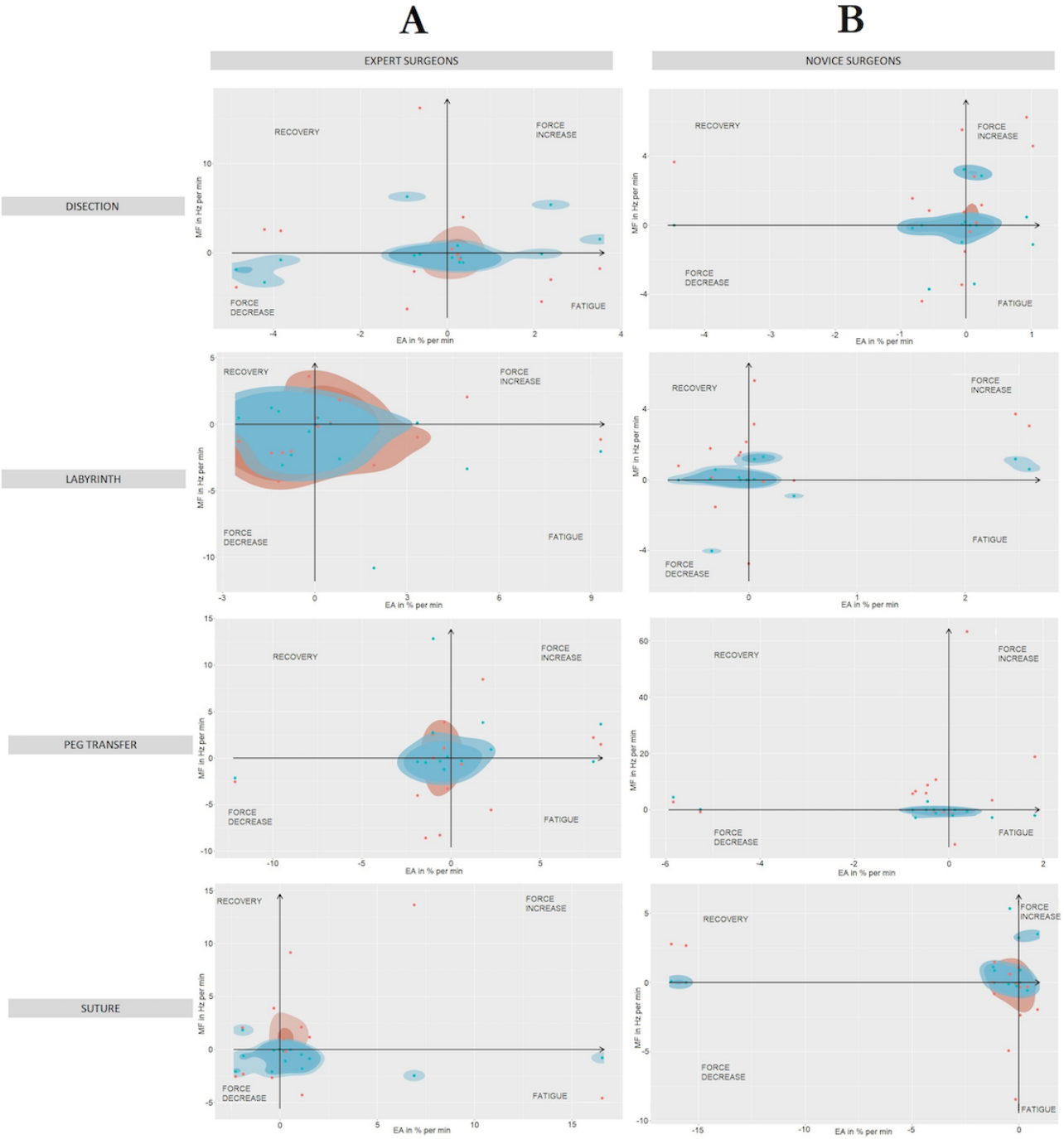

Comparative as a Function of the Level of Expertise

Expert surgeons showed greater muscle activity in the forearm muscles (middle trapezius, upper trapezius, triceps brachii, and brachioradialis) than novice surgeons during LAP practice (Figure 6). However, in the rest of the muscle groups, expert surgeons showed less muscle activity. In the case of RAS, in general, novice surgeons showed less muscle activity than expert surgeons (Figure 6).

In general, expert surgeons showed greater localized muscle fatigue and greater use of force during LAP (Figure 7). On the other hand, novice surgeons showed less muscle fatigue, greater use of force, and muscle recovery (Figure 7). In the case of RAS, in general, expert surgeons showed greater use of force and greater localized muscle fatigue (Figure 7). In the case of novice surgeons, they showed less localized muscle fatigue and greater muscle recovery with less use of force (Figure 7).

XSL•FO
RenderX

Figure 6. Comparison of muscle activity (percentage of maximum voluntary contraction) of (A) conventional laparoscopic and (B) robotic-assisted surgery during performance of 4 simulator tasks, from upper to bottom: dissection, labyrinth, peg transfer, and suturing. Performing experienced (red) and novice (blue) surgeons for the following muscle groups bilaterally: Erector spinae (ER_SPIN), Middle trapezius (MID_TRAP), and Upper trapezius (UP_TRAP).

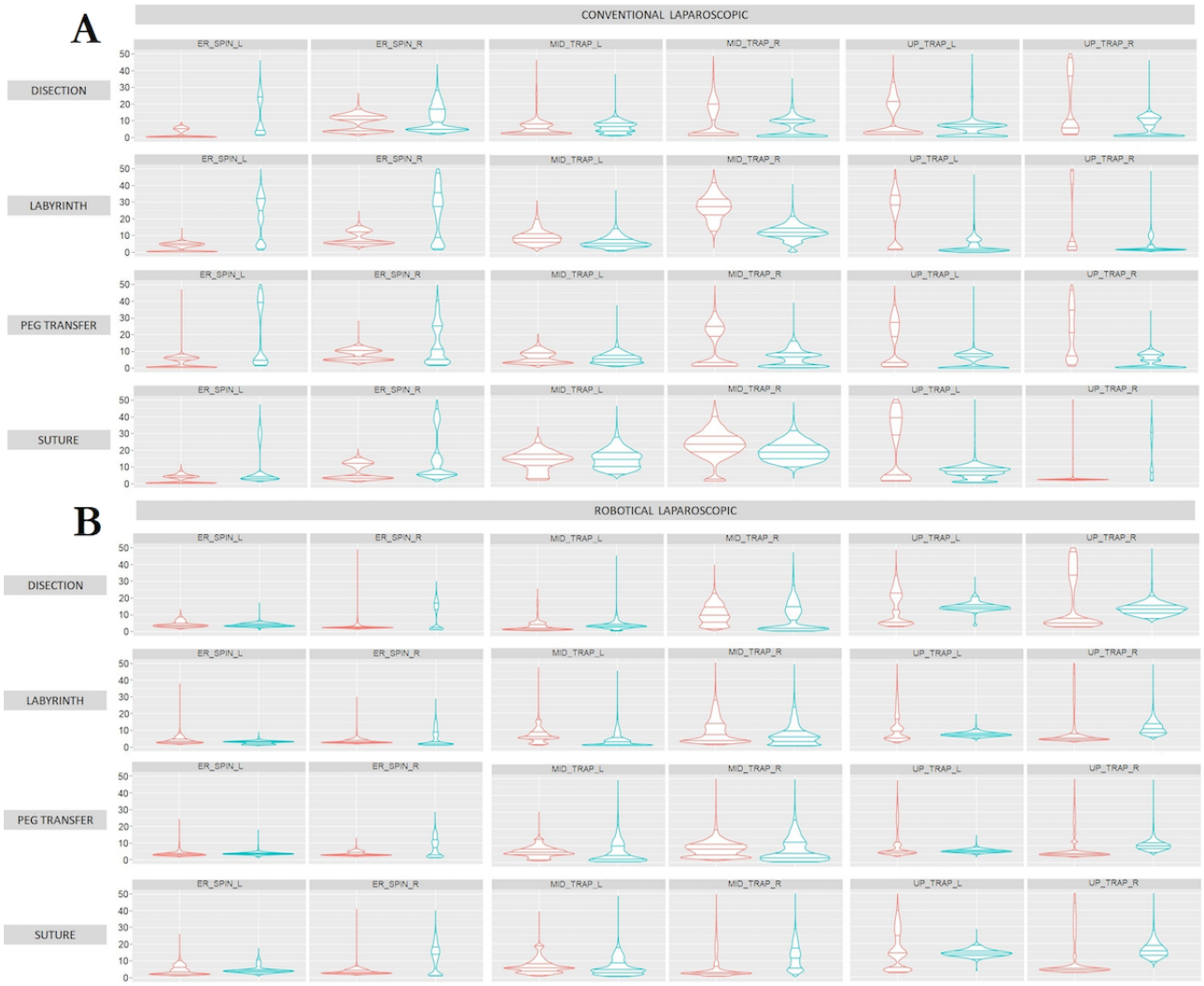

XSL•FO
RenderX

Figure 7. Comparison of fatigue/recovery muscle localized and muscle strength increase/decrease during (A) conventional laparoscopic and (B) robotic-assisted surgeries on 4 simulator tasks, from upper to bottom: dissection, labyrinth, peg transfer, and suturing. Performed by experienced (red) and novice (blue) surgeons.

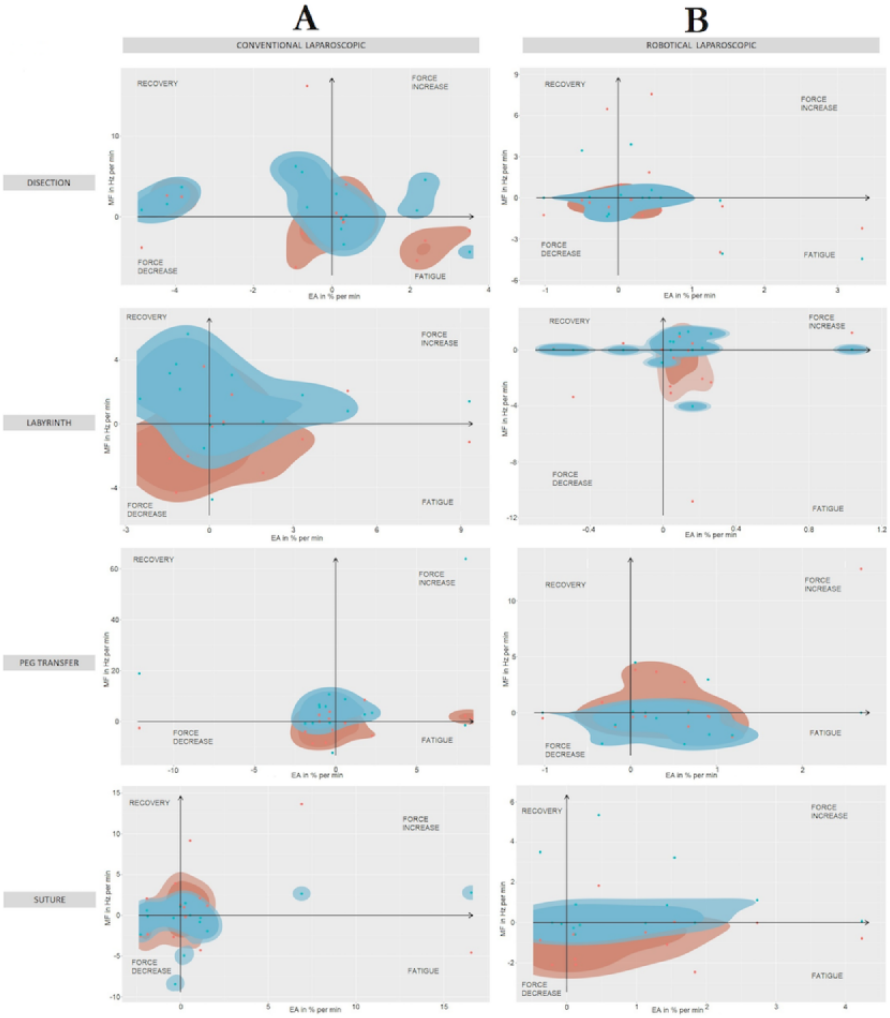

Predictive Models

Training Dataset

The predicted results for the training dataset are shown in Figure 8 for the different tasks. This shows the R^2 coefficient values of the training dataset for the predictive analysis performed applying MLR and MLP, showing the higher of the two in Figure 8, with the dataset preprocessed using the scaling technique. In general, MLR showed slightly higher values than MLP for the R^2 coefficient with a low RMSE error (RMSE < 0.05) for both cases.

In general, the results obtained in the present study are satisfactory according to the standards given by Colton [36]. High to excellent correlations (R^2>0.75), close to 1, and RMSE close to 0 were achieved for all the muscle groups studied. Of note were the R^2 coefficient values during LAP for the right brachioradialis in the dissection task (R^2=0.9963) and the left erector spinae in the pin transfer task (R^2=0.9828) of novice surgeons, and the right erector spinae of expert surgeons in the suturing task (R^2=0.9881). For RAS surgery, right triceps brachii in the dissection task (R^2=0.9953), left vastus lateralis in the dissection (R^2=0.9955), suture (R^2=0.9951), and pin transfer (R^2=0.9949) tasks, and right vastus lateralis in the suture (R^2=0.9933) and labyrinth (R^2=0.9904) tasks of novice surgeons were noteworthy.

XSL•FO
RenderX

Figure 8. Results from the calibration dataset indicate the R2 values applying the best model in each case on four simulator tasks: (A) dissection, (B) labyrinth, (C) peg transfer, and (D) suturing. For the 4 analyzed groups as a function of the surgical type and the surgeons' level of expertise, being novice surgeons and conventional laparoscopic surgery (yellow), novice surgeons and robotic-assisted surgery (red), experienced surgeons and conventional laparoscopic surgery (blue), and experienced surgeons and robotic-assisted surgery (green).

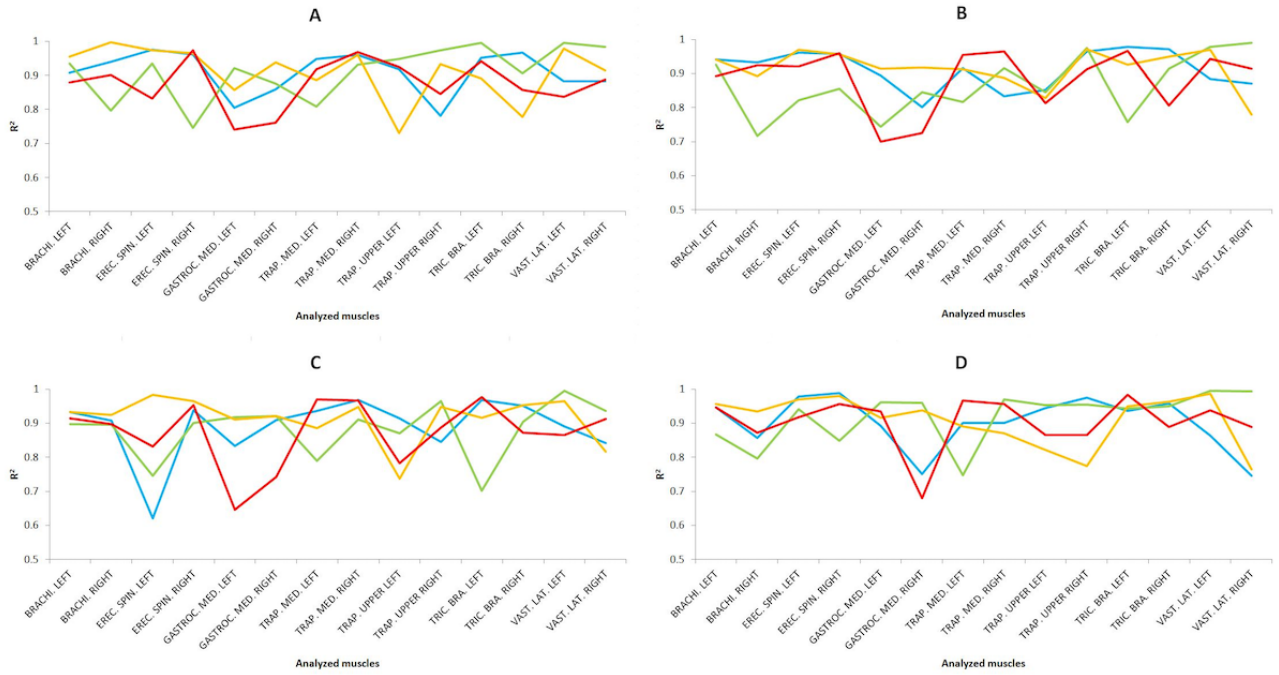

Prediction Results of the Cross-Validation

The results obtained from the cross-validation are shown in Figure 9 for each task. Thus, the R^2 coefficient values of the cross-validation for the predictive analysis performed by applying MLR and MLP are shown, with the higher of the two in Figure 9, with the dataset preprocessed using the scaling technique. Overall, MLR showed slightly higher values than MLP for the R^2 coefficient with a low RMSE error (RMSE<0.05) for both cases.

In general, the results obtained in this study are adequate, according to the standards established by Colton [36]. Of note were the R^2 coefficient values for the LAP technique with regard to the right erector spinae during the suturing task (R^2=0.9606) for novice surgeons. Regarding RAS, of note were the right middle trapezius in the suturing task (R^2=0.9472) and the left vastus lateralis in the dissection (R^2=0.9541) and suturing (R^2=0.9746) tasks for expert surgeons, as well as the left triceps brachii (R^2=0.9408) and the right vastus lateralis (R^2=0.9511) in the suturing task for novice surgeons.

In almost all cases, a good to excellent relationship was achieved (R^2>0.6), highlighting that for the dissection task, expert surgeons with RAS achieved the highest R^2 values for most parameters, for the labyrinth task, novice surgeons with LAP, for the peg transfer task, novice surgeons with LAP, and for the suturing task, expert surgeons with RAS.

XSL•FO
RenderX

Figure 9. Results from cross-validation indicate the R2 values applying the best model in each case on four simulator tasks: (A) dissection, (B) labyrinth, (C) peg transfer, and (D) suturing. For the 4 analyzed groups as a function of the surgical type and the surgeons' level of expertise, being novice surgeons and conventional laparoscopic surgery (yellow), novice surgeons and robotic-assisted surgery (red), experienced surgeons and conventional laparoscopic surgery (blue), and experienced surgeons and robotic-assisted surgery (green).

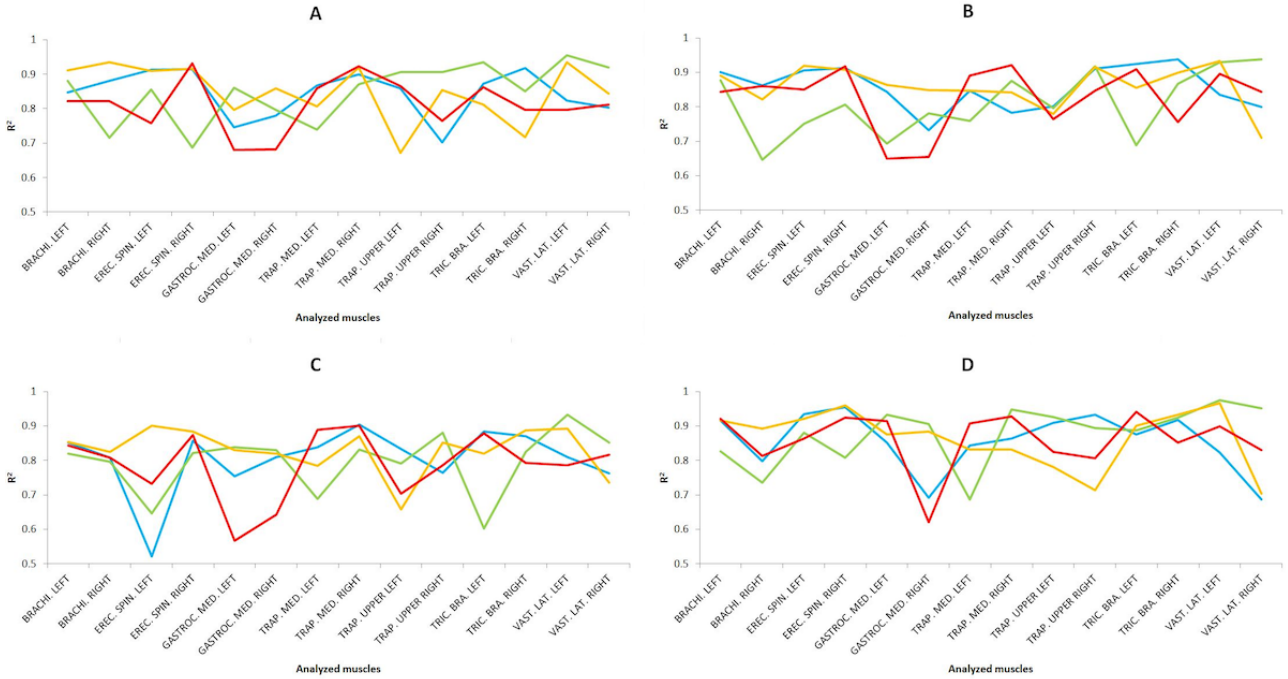

Test Dataset

To validate the predictive models, the predicted results for the test dataset were analyzed and are shown in Figure 10 for each task. As in the previous sections, Figure 10 shows the R^2 coefficient values of the test dataset for the predictive analysis performed by applying MLR and MLP, showing the higher of the two in Figure 10, with the dataset preprocessed using the scaling technique. Overall, MLR showed slightly higher values than MLP for the R^2 coefficient with a low RMSE error (RMSE<0.05) for both cases.

As in the previous cases, in general, the results obtained are adequate according to the standards given by Colton [36]. We should highlight the R^2 coefficient values for LAP of the right erector spinae (R^2=0.8782), the left erector spinae (R^2=0.8681), and the left upper trapezius (R^2=0.8751) in the suturing task for expert surgeons. As for RAS, we will highlight the right medial gastrocnemius (R^2=0.8599), the right middle trapezius (R^2=0.8502), the left vastus lateralis (R^2=0.8751) and the right vastus lateralis (R^2=0.8933) and labyrinth (R^2=0.8604) in the suturing task for expert surgeons, as well as the left middle trapezius (R^2=0.8671), the left triceps brachii (R^2=0.8837) in the suturing task for novice surgeons.

XSL•FO
RenderX

Figure 10. Results from the test dataset indicate the R2 values applying the best model in each case on four simulator tasks: (A) dissection, (B) labyrinth, (C) peg transfer, and (D) suturing. For the 4 analyzed groups as a function of the surgical type and the surgeons' level of expertise, being novice surgeons and conventional laparoscopic surgery (yellow), novice surgeons and robotic-assisted surgery (red), experienced surgeons and conventional laparoscopic surgery (blue), and experienced surgeons and robotic-assisted surgery (green).

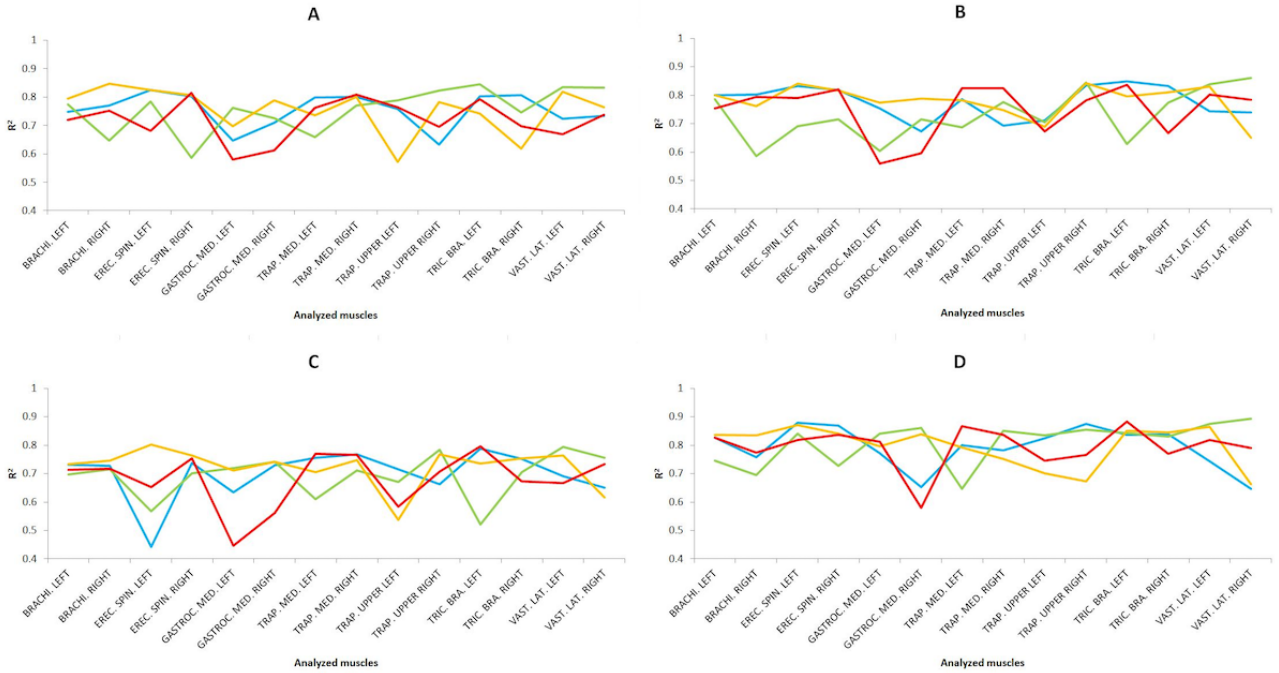

Discussion

Principal Findings

The muscles analyzed in this study have been extensively evaluated in previous studies [19,39] for LAP and RAS, with the main muscles analyzed being the upper trapezius, middle trapezius, and triceps brachii.

Considering the type of surgical technique, the expert surgeons had less muscle activity with the RAS technique for the dissection, peg transfer, and suturing tasks, mainly for the forearm muscles (triceps brachii, middle trapezius, and upper trapezius). However, the labyrinth task had less muscle activity with the LAP technique for the nondominant hand. Conversely, the dominant hand showed less muscle activity in the RAS technique. For the novice surgeons, the middle trapezius bilaterally and the right upper trapezius had less muscle activity in LAP for the dissection, labyrinth, and pin transfer tasks. For the other cases, they had less muscle activity using the RAS technique. In all cases, the muscle activity was lower using the RAS technique during the suturing task. The results obtained during this study agree with those obtained in previous studies [19,40], which demonstrated that from an ergonomic point of view, RAS is more comfortable than LAP [41].

In the same way, muscle activity is closely related to muscle fatigue, and prolonged muscle activity of small, single muscle fibers can cause degenerative muscle changes even at very low levels of muscle activity. For this reason, proper ergonomics is very important for LAP activities [42]. This fact reinforces the need for future surgeons in surgical robotics to receive structured, progressive, and adequate training in ergonomic aspects in both LAP and RAS [43].

Joint analysis of spectrum and analysis graphs has been widely applied in scientific literature for the analysis of muscle fatigue [44]. This method relates muscle activity and the mean frequency of the EMG spectrum, obtaining a schematic representation of the relationship between localized muscle fatigue and force [45]. Previous studies have studied trapezius muscle fatigue during LAP, as the trapezius is frequently vulnerable to localized muscle fatigue [46]. This fact is being corrected with RAS, as the activity of the main dominant muscles in LAP has been reduced [47]. Furthermore, these results showed less localized muscle fatigue and greater use of force in RAS than in LAP in novice surgeons. However, in the case of expert surgeons, the use of force and localized muscle fatigue was lower in LAP. This fact could be related to their previous experience in LAP [43]. However, it would be necessary to take into account other aspects that may affect surgeons' ergonomics, such as the ergonomically optimal configuration of the working environment, their prior training in the field of surgical ergonomics, activities performed prior to the study, or the use of simulators of similar surgical environments, among others.

In general, all muscles bilaterally show statistically significant differences between the LAP technique and RAS. This fact demonstrates the different use of muscles between LAP and RAS for both types of surgeons, experts and novices [39]. Similarly, obtaining better results by applying MLR as an AI technique could be related to the limitations of using MLP, such as its black box nature or the need to use a large amount of data for training.

Considering more specifically the results for the tasks performed by the LAP technique (Figure 6), for the dissection task, the left brachioradialis showed lower muscle activity for novice

XSL•FO
RenderX

surgeons, and the right medial gastrocnemius showed lower muscle activity for expert surgeons. For the peg transfer task, the muscles that did not follow the general trend were the medial gastrocnemius bilaterally, showing less muscle activity for novice surgeons, and the right vastus lateralis, also showing less muscle activity for novice surgeons. In the same way, the results obtained for the suturing task followed the general trend, except for the medial gastrocnemius bilaterally, displaying less muscle activity for novice surgeons, highlighting the similarity for the middle trapezius bilaterally.

In the case of the tasks using RAS (Figure 6), the dissection and labyrinth tasks followed the general trend, with less muscle activity being shown by novice surgeons, except for the right erector spinae, which showed less muscle activity for the expert surgeons. For the peg transfer task, the general trend was followed, with less muscle activity for novice surgeons, except for the erector spinae bilaterally and right middle trapezius, which displayed less muscle activity for expert surgeons. Finally, for the suturing task, the general trend was followed, with less muscle activity for novice surgeons, except for the erector spinae bilaterally and the right middle trapezius and right upper trapezius, which displayed less muscle activity for expert surgeons.

It is noteworthy that, in general, muscle activity on the right side is lower than on the left side, both for experienced and novice surgeons. This fact could be related to the dominant side of surgeons, who require less effort to complete a manual task [48]. Consistent with this fact, muscle activity on the dominant side also appears to increase slightly over time compared to the nondominant side, suggesting that the dominant side may be more susceptible to developing muscle fatigue and possible future musculoskeletal disorders [49]. In the same way, the results obtained for Figure 7 are consistent with those obtained in previous analyses [50].

The results from Figure 8 could be related to the variability of the surgeon's surgical experience and the stress generated during the performance of surgical activities [51]. Moreover, in almost all cases, a high to excellent relationship was achieved (R^2>0.75), highlighting that for the dissection task, expert surgeons with RAS achieved the highest R^2 values for most parameters for the labyrinth task, expert surgeons with LAP, for the peg transfer task, novice surgeons with LAP and for the suturing task, expert surgeons with RAS. These positive predictive results are in agreement with previous studies that also showed high to excellent predictive results performing LAP and RAS [5]. As for the previous results, the predictive results for Figure 9 are in agreement with previous studies [5].

As in previous sections, these satisfactory predictive results for Figure 10 are in agreement with previous studies [5]. Moreover, in almost all cases, a good to excellent relationship was achieved (R^2>0.5), highlighting that for the dissection task, expert surgeons with RAS obtained the highest R^2 values for most parameters, for the labyrinth and pin transfer tasks, novice surgeons with LAP, and for the suturing task, expert surgeons with RAS. In general, the outcomes of this study exhibit a slight superiority over those obtained in preceding studies that used AI models for localized muscle fatigue prediction [52-55].

Among the limitations of the present study is the small number of participating surgeons. Although previous studies have established the relationship between muscle activity and localized muscle fatigue experienced during simulated tasks compared to surgical procedures [19], further research in this area is warranted. This should involve increasing the number of surgeons involved and conducting more intricate surgical procedures that closely resemble real-world scenarios, obtaining more definitive and representative results [56]. On the other hand, the limitations due to the number of EMG sensors available and that can be used simultaneously have prevented us from recording the muscle activity of all the muscle groups of interest. For future work, we intend to include recordings of other muscle groups, such as the deltoid or biceps brachii, which have also been shown to be related to LAP activity. Furthermore, the incorporation of additional wearable and physiological sensors, such as EDA or an electrocardiogram, should be considered in subsequent studies to enhance the representativeness of the results.

For future research, the predictive models derived from this study are intended to be integrated and validated within surgical training activities and in actual surgical practice across diverse specialties, including general surgery, urology, and gynecology, among others. This integration will contribute to the prevention of musculoskeletal issues during surgical practice and, consequently, enhance the quality of surgical outcomes. Furthermore, it would be beneficial for future studies to include participating surgeons with prior experience in RAS and to analyze the impact of experience on the ergonomics of surgeons who perform robotic surgery. Additionally, it would be insightful to investigate the correlation between the onset of localized muscle fatigue and the quality of surgical performance during RAS surgery.

Similarly, the development of real-time feedback systems would allow surgeons to be alerted to situations of high localized muscle fatigue and prevent possible musculoskeletal risks during surgical practice. In this way, predictive models could be integrated into surgical training and actual clinical practice, improving the health of the surgical team and, consequently, the quality of patient care.

Conclusions

In this study, a novel predictive model based on MLR and MLP has been developed to predict the localized muscle fatigue of a set of surgeons' muscle groups during LAP and RAS. This model achieved satisfactory R^2 and RMSE coefficients for all parameters extracted from the EMG signal, predicting with high accuracy the localized muscle fatigue values. The MLR predictive model demonstrated superior performance relative to the MLP model. Therefore, it is a promising starting point for predicting surgeons' ergonomic risks during surgical practice in MIS.

Regarding the analysis of localized muscle fatigue, in general, RAS showed less muscle fatigue for novice surgeons compared to LAP practice, although it was higher for expert surgeons.

XSL•FO
RenderX

Regarding ergonomic aspects, postural habits and motor control are different among surgeons with different surgical experience, which affects muscle activity patterns. The prevention and improvement of ergonomics in surgical practice minimizes localized muscle fatigue and forced postures, improves the health of surgeons, favors surgical performance, and consequently, contributes to improving the quality of patient care.

Ergonomic analysis makes it possible to study individual variations in the different ergonomic parameters according to the surgeons' experience or the surgical technique used, thus allowing for improved training and surgical guidelines. Considering the rapid expansion of RAS in clinical practice, the analysis and improvement of the ergonomic conditions of surgeons and the surgical team are essential to provide optimal working conditions, ensuring the well-being of the surgical team and the quality of patient care. The training of the surgical team in the optimization of ergonomic conditions is fundamental to maximizing the benefits of MIS for both patients and surgeons.

Acknowledgments

The authors would also like to thank the colleagues who collaborated in this study (Carlos Sánchez, David Durán, Javier Salas, Ramón González, Laura Pires, and Javier Vela). Generative artificial intelligence (AI) was not used in the writing of this manuscript.

Funding

This work has been financed by the Ministry of Science and Innovation with funds from the European Union Next Generation EU; from the Recovery, Transformation, and Resilience Plan (PRTR-C17.I1); and the European Regional Development Fund (ERDF) of Extremadura Operational Program 2021-2027. However, the views and opinions expressed are those of the authors.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

DC, MJP-S, JAS-M, and FMS-M contributed to the conceptualization of the study and participated in the formal analysis, investigation, methodology development, and validation. DC, MJP-S, and JAS-M were responsible for data curation, software development, and visualization. Funding acquisition, project administration, resources, and supervision were undertaken by JAS-M and FMS-M. DC, MJP-S, and JAS-M prepared the original draft of the manuscript, and all 4 authors—DC, MJP-S, JAS-M, and FMS-M—contributed to the review and editing of the manuscript. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

None declared.

References

1. Gabrielson AT, Clifton MM, Pavlovich CP, Biles MJ, Huang M, Agnew J, et al. Surgical ergonomics for urologists: a practical guide. Nat Rev Urol 2021;18(3):160-169. [doi: 10.1038/s41585-020-00414-4] [Medline: 33432182]
2. Müller DT, Ahn J, Brunner S, Poggemeier J, Storms C, Reisewitz A, et al. Ergonomics in robot-assisted surgery in comparison to open or conventional laparoendoscopic surgery: a narrative review. Int J Abdom Wall Hernia Surg 2023;6(2):61-66. [doi: 10.4103/ijawhs.ijawhs_52_22]
3. Hurley AM, Kennedy PJ, O'Connor L, Dinan TG, Cryan JF, Boylan G, et al. SOS save our surgeons: stress levels reduced by robotic surgery. Gynecol Surg 2015;12(3):197-206. [doi: 10.1007/s10397-015-0891-7]
4. Turner SR, Mormando J, Park BJ, Huang J. Attitudes of robotic surgery educators and learners: challenges, advantages, tips and tricks of teaching and learning robotic surgery. J Robot Surg 2020;14(3):455-461 [FREE Full text] [doi: 10.1007/s11701-019-01013-1] [Medline: 31463878]
5. Caballero D, Pérez-Salazar MJ, Sánchez-Margallo JA, Sánchez-Margallo FM. Applying artificial intelligence on EDA sensor data to predict stress on minimally invasive robotic-assisted surgery. Int J Comput Assist Radiol Surg 2024;19(10):1953-1963. [doi: 10.1007/s11548-024-03218-8] [Medline: 38955902]
6. Kaplan JR, Lee Z, Eun DD, Reese AC. Complications of minimally invasive surgery and their management. Curr Urol Rep 2016;17(6):47. [doi: 10.1007/s11934-016-0602-6] [Medline: 27075019]
7. Wilson MR, Poolton JM, Malhotra N, Ngo K, Bright E, Masters RSW. Development and validation of a surgical workload measure: the surgery task load index (SURG-TLX). World J Surg 2011;35(9):1961-1969 [FREE Full text] [doi: 10.1007/s00268-011-1141-4] [Medline: 21597890]

XSL•FO
RenderX

8. Nillahoot N, Pillai BM, Sharma B, Wilasrusmee C, Suthakorn J. Interactive 3D force/torque parameter acquisition and correlation identification during primary trocar insertion in laparoscopic abdominal surgery: 5 cases. Sensors (Basel) 2022 Nov 19;22(22):8970 [FREE Full text] [doi: 10.3390/s22228970] [Medline: 36433567]
9. Armijo PR, Huang C, High R, Leon M, Siu K, Oleynikov D. Ergonomics of minimally invasive surgery: an analysis of muscle effort and fatigue in the operating room between laparoscopic and robotic surgery. Surg Endosc 2019;33(7):2323-2331. [doi: 10.1007/s00464-018-6515-3] [Medline: 30341653]
10. Guzmán-García C, Sánchez-González P, Margallo JAS, Snoriguzzi N, Rabazo JC, Margallo FMS, et al. Correlating personal resourcefulness and psychomotor skills: an analysis of stress, visual attention and technical metrics. Sensors (Basel) 2022;22(3):837 [FREE Full text] [doi: 10.3390/s22030837] [Medline: 35161582]
11. Avila-Tomás JF, Mayer-Pujadas M, Quesada-Varela V. [Artificial intelligence and its applications in medicine I: introductory background to AI and robotics]. Aten Primaria 2020;52(10):778-784 [FREE Full text] [doi: 10.1016/j.aprim.2020.04.013] [Medline: 32660768]
12. Janiesch C, Zschech P, Heinrich K. Machine learning and deep learning. Electron Mark 2021;31(3):685-695. [doi: 10.1007/s12525-021-00475-2]
13. Luo X, Kang Y, Duan S, Yan P, Song G, Zhang N, et al. Machine learning-based prediction of acute kidney injury following pediatric cardiac surgery: model development and validation study. J Med Internet Res 2023;25:e41142 [FREE Full text] [doi: 10.2196/41142] [Medline: 36603200]
14. Lee GI, Lee MR, Green I, Allaf M, Marohn MR. Surgeons' physical discomfort and symptoms during robotic surgery: a comprehensive ergonomic survey study. Surg Endosc 2017;31(4):1697-1706. [doi: 10.1007/s00464-016-5160-y] [Medline: 27515836]
15. Kim HJ, Yang JH, Chang D, Lenke LG, Pizones J, Castelein R, et al. Assessing the reproducibility of the structured abstracts generated by chatgpt and bard compared to human-written abstracts in the field of spine surgery: comparative analysis. J Med Internet Res 2024;26:e52001 [FREE Full text] [doi: 10.2196/52001] [Medline: 38924787]
16. Brunner S, Müller D, Krauss DT, Datta RR, Eckhoff JA, Storms C, et al. Cologne ergonomic measurement for robotic surgery (CEMRobSurg) using the Hugo™ RAS system. Surg Endosc 2024;38(10):6128-6138. [doi: 10.1007/s00464-024-11129-7] [Medline: 39187727]
17. Dwyer A, Huckleby J, Kabbani M, Delano A, De Sutter M, Crawford D. Ergonomic assessment of robotic general surgeons: a pilot study. J Robot Surg 2020;14(3):387-392. [doi: 10.1007/s11701-019-00996-1] [Medline: 31302826]
18. Yu D, Dural C, Morrow MMB, Yang L, Collins JW, Hallbeck S, et al. Intraoperative workload in robotic surgery assessed by wearable motion tracking sensors and questionnaires. Surg Endosc 2017;31(2):877-886. [doi: 10.1007/s00464-016-5047-y] [Medline: 27495330]
19. Pérez-Salazar MJ, Caballero D, Sánchez-Margallo JA, Sánchez-Margallo FM. Comparative study of ergonomics in conventional and robotic-assisted laparoscopic surgery. Sensors (Basel) 2024;24(12):3840 [FREE Full text] [doi: 10.3390/s24123840] [Medline: 38931624]
20. Pérez-Salazar MJ, Caballero D, Sánchez-Margallo JA, Sánchez-Margallo FM. Correlation study and predictive modelling of ergonomic parameters in robotic-assisted laparoscopic surgery. Sensors (Basel) 2024;24(23):7721 [FREE Full text] [doi: 10.3390/s24237721] [Medline: 39686259]
21. Hwang M, Ichnowski J, Thananjeyan B, Seita D, Paradis S, Fer D, et al. Automating surgical peg transfer: calibration with deep learning can exceed speed, accuracy, and consistency of humans. IEEE Trans Automat Sci Eng 2023;20(2):909-922. [doi: 10.1109/tase.2022.3171795]
22. Healey J, Picard R. Detecting stress during real-world driving tasks using physiological sensors. IEEE Trans Intell Transport Syst 2005;6(2):156-166. [doi: 10.1109/TITS.2005.848368]
23. Shi W, Liu PX, Zheng M. Cutting procedures with improved visual effects and haptic interaction for surgical simulation systems. Comput Methods Programs Biomed 2020;184:105270. [doi: 10.1016/j.cmpb.2019.105270] [Medline: 31881400]
24. Phan PT, Hoang TT, Thai MT, Low H, Davies J, Lovell NH, et al. Smart surgical sutures using soft artificial muscles. Sci Rep 2021;11(1):22420 [FREE Full text] [doi: 10.1038/s41598-021-01910-2] [Medline: 34789808]
25. NA. [Law 14/2007, 3 July, on biomedical research (BOE, 4 July 2007)]. Rev Derecho Genoma Hum 2007(26):283-325. [Medline: 18201045]
26. Hermens HJ, Freriks B, Disselhorst-Klug C, Rau G. Development of recommendations for SEMG sensors and sensor placement procedures. J Electromyogr Kinesiol 2000;10(5):361-374. [doi: 10.1016/s1050-6411(00)00027-4] [Medline: 11018445]
27. Dufaug A, Barthod C, Goujon L, Marechal L. New joint analysis of electromyography spectrum and amplitude-based methods towards real-time muscular fatigue evaluation during a simulated surgical procedure: a pilot analysis on the statistical significance. Med Eng Phys 2020;79:1-9. [doi: 10.1016/j.medengphy.2020.01.017] [Medline: 32273161]
28. Oka M. Interpreting a standardized and normalized measure of neighborhood socioeconomic status for a better understanding of health differences. Arch Public Health 2021;79(1):226 [FREE Full text] [doi: 10.1186/s13690-021-00750-w] [Medline: 34911564]
29. Dietterich TG. Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput 1998;10(7):1895-1923. [doi: 10.1162/089976698300017197] [Medline: 9744903]

XSL•FO
RenderX

30. Frank E, Hall MA, Witten IH. The weka workbench. In: Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques. Burlington, Massachusetts: Morgan Kaufmann; 2016.
31. Fayyad U, Piatetsky‐Shapiro G, Smyth P. From data mining to knowledge discovery in databases. AI Mag 1996;17(3):37-54. [doi: 10.1609/aimag.v17i3.1230]
32. Caballero D, Caro A, Dahl AB, ErsbØll BK, Amigo JM, Pérez-Palacios T, et al. Comparison of different image analysis algorithms on MRI to predict physico-chemical and sensory attributes of loin. Chemom Intell Lab Syst 2018;180:54-63. [doi: 10.1016/j.chemolab.2018.04.008]
33. Wu X, Kumar V, Ross Quinlan J, Ghosh J, Yang Q, Motoda H, et al. Top 10 algorithms in data mining. Knowl Inf Syst 2007;14(1):1-37. [doi: 10.1007/s10115-007-0114-2]
34. Bishop CM. Pattern Recognition and Machine Learning. New York: Springer; 2006.
35. Grossman R, Seni G, Elder J. Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions. Williston, Vermont, U.S.A: Morgan and Claypool Publishers; 2010.
36. Colton T. Statistics in Medicine. New York: Little Brown and Co; 1974.
37. Hyndman RJ, Koehler AB. Another look at measures of forecast accuracy. Int J Forecast 2006;22(4):679-688. [doi: 10.1016/j.ijforecast.2006.03.001]
38. Porta-Serra M, Plasencia A, Sanz F. La calidad de la información clínica: estadistícamente significativo o clínicamente importante? [The quality of clinical information: statistically significant or clinically important?]. Med Clin (Barcelona) 1988;90:463-468.
39. Jarc AM, Curet MJ. Viewpoint matters: objective performance metrics for surgeon endoscope control during robot-assisted surgery. Surg Endosc 2017;31(3):1192-1202 [FREE Full text] [doi: 10.1007/s00464-016-5090-8] [Medline: 27422247]
40. Pérez-Duarte FJ, Lucas-Hernández M, Matos-Azevedo A, Sánchez-Margallo JA, Díaz-Güemes I, Sánchez-Margallo FM. Objective analysis of surgeons' ergonomy during laparoendoscopic single-site surgery through the use of surface electromyography and a motion capture data glove. Surg Endosc 2014;28(4):1314-1320. [doi: 10.1007/s00464-013-3334-4] [Medline: 24337915]
41. Willuth E, Hardon SF, Lang F, Haney CM, Felinska EA, Kowalewski KF, et al. Robotic-assisted cholecystectomy is superior to laparoscopic cholecystectomy in the initial training for surgical novices in an ex vivo porcine model: a randomized crossover study. Surg Endosc 2022;36(2):1064-1079 [FREE Full text] [doi: 10.1007/s00464-021-08373-6] [Medline: 33638104]
42. Wee IJY, Kuo L, Ngu JC. A systematic review of the true benefit of robotic surgery: ergonomics. Int J Med Robot 2020;16(4):e2113. [doi: 10.1002/rcs.2113] [Medline: 32304167]
43. Sjøgaard G, Lundberg U, Kadefors R. The role of muscle activity and mental load in the development of pain and degenerative processes at the muscle cell level during computer work. Eur J Appl Physiol 2000;83(2-3):99-105. [doi: 10.1007/s004210000285] [Medline: 11104050]
44. Straker L, Mathiassen SE. Increased physical work loads in modern work--a necessity for better health and performance? Ergonomics 2009;52(10):1215-1225. [doi: 10.1080/00140130903039101] [Medline: 19787501]
45. Mastaglia FL. The relationship between muscle pain and fatigue. Neuromuscul Disord 2012;22 Suppl 3:S178-S180. [doi: 10.1016/j.nmd.2012.10.003] [Medline: 23182635]
46. Luttmann A, Jäger M, Laurig W. Electromyographical indication of muscular fatigue in occupational field studies. Int J Ind Ergon 2000;25(6):645-660. [doi: 10.1016/s0169-8141(99)00053-0]
47. Nordander C, Hansson G, Ohlsson K, Arvidsson I, Balogh I, Strömberg U, et al. Exposure-response relationships for work-related neck and shoulder musculoskeletal disorders--analyses of pooled uniform data sets. Appl Ergon 2016;55:70-84. [doi: 10.1016/j.apergo.2016.01.010] [Medline: 26995038]
48. Krzysztofik M, Jarosz J, Matykiewicz P, Wilk M, Bialas M, Zajac A, et al. A comparison of muscle activity of the dominant and non-dominant side of the body during low versus high loaded bench press exercise performed to muscular failure. J Electromyogr Kinesiol 2021;56:102513. [doi: 10.1016/j.jelekin.2020.102513] [Medline: 33422743]
49. Luger T, Bonsch R, Seibt R, Krämer B, Rieger MA, Steinhilber B. Intraoperative active and passive breaks during minimally invasive surgery influence upper extremity physical strain and physical stress response-a controlled, randomized cross-over, laboratory trial. Surg Endosc 2023;37(8):5975-5988 [FREE Full text] [doi: 10.1007/s00464-023-10042-9] [Medline: 37084097]
50. Diederichsen LP, Nørregaard J, Dyhre-Poulsen P, Winther A, Tufekovic G, Bandholm T, et al. The effect of handedness on electromyographic activity of human shoulder muscles during movement. J Electromyogr Kinesiol 2007;17(4):410-419. [doi: 10.1016/j.jelekin.2006.03.004] [Medline: 16839778]
51. Rodrigues Armijo P, Huang C, Carlson T, Oleynikov D, Siu K. Ergonomics analysis for subjective and objective fatigue between laparoscopic and robotic surgical skills practice among surgeons. Surg Innov 2020;27(1):81-87. [doi: 10.1177/1553350619887861] [Medline: 31771411]
52. Keshavarz Panahi A, Cho S. Prediction of muscle fatigue during minimally invasive surgery using recurrence quantification analysis. Minim Invasive Surg 2016;2016:5624630 [FREE Full text] [doi: 10.1155/2016/5624630] [Medline: 27313884]
53. Al-Mulla MR, Sepulveda F, Colley M. A review of non-invasive techniques to detect and predict localised muscle fatigue. Sensors (Basel) 2011;11(4):3545-3594 [FREE Full text] [doi: 10.3390/s110403545] [Medline: 22163810]

XSL•FO
RenderX

54. Lambay A, Liu Y, Morgan PL, Ji Z. Machine learning assisted human fatigue detection, monitoring, and recovery: a review. Digit Eng 2024;1:100004. [doi: 10.1016/j.dte.2024.100004]
55. Wei K, Kimura C, Shimura M, Shimomura Y, Zhao X, Tamura T, et al. Predicting task performance in robot-assisted surgery using physiological stress and subjective workload: a case study with interpretable machine learning. Front Hum Neurosci 2025;19:1611524 [FREE Full text] [doi: 10.3389/fnhum.2025.1611524] [Medline: 40606493]
56. Amirthanayagam A, Zecca M, Barber S, Singh B, Moss EL. Impact of minimally invasive surgery on surgeon health (ISSUE) study: protocol of a single-arm observational study conducted in the live surgery setting. BMJ Open 2023;13(3):e066765 [FREE Full text] [doi: 10.1136/bmjopen-2022-066765] [Medline: 36882245]

Abbreviations

%MVC: percentage of maximum voluntary contraction
AI: artificial intelligence
EDA: electrodermal activity
EMG: electromyography
LAP: conventional laparoscopic surgery
MIS: minimally invasive surgery
MLP: multilayer perceptron
MLR: multiple linear regression
RAS: robotic-assisted surgery
RMSE: root-mean-square error
WEKA: Waikato Environment for Knowledge Analysis

Edited by J Sarvestan; submitted 08.11.24; peer-reviewed by H Ghorbani, X Wang, X Xiao; comments to author 22.07.25; revised version received 29.08.25; accepted 02.10.25; published 10.12.25.

Please cite as:
Caballero D, Pérez-Salazar MJ, Sánchez-Margallo JA, Sánchez-Margallo FM
Localized Muscular Fatigue in Robotic-Assisted Laparoscopic Surgery: Predictive Modeling Study
JMIR Form Res 2025;9:e68536
URL: https://formative.jmir.org/2025/1/e68536
doi:10.2196/68536
PMID:

©Daniel Caballero, Manuel J Pérez-Salazar, Juan A Sánchez-Margallo, Francisco M Sánchez-Margallo. Originally published in JMIR Formative Research (https://formative.jmir.org), 10.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Awareness, Perceptions, Willingness, and Feasibility of mHealth Apps Among People Living With Epilepsy: Cross-Sectional Questionnaire Study

Bushra Batool Zahra[1], PharmD, MPhil; Muhammad Amir Hamza[1], PharmD, MPhil; Rehana Sarwat[1], PharmD, MPhil; Hina Batool Zahra[1], PharmD; Ayesha Azam[1], PharmD, MPhil; Ali Ahmed[2], PharmD, MPhil, PhD

[1]Department of Pharmacy Practice, Faculty of Pharmaceutical Sciences, Riphah International University, Islamabad, Pakistan

[2]Division of Infectious Diseases and Global Public Health, School of Medicine, University of California San Diego, San Diego, CA, United States

Corresponding Author:
Ali Ahmed, PharmD, MPhil, PhD
Division of Infectious Diseases and Global Public Health
School of Medicine
University of California San Diego
9500 Gilman Drive, La Jolla
San Diego, CA, 92093
United States
Phone: 1 6199539767
Email: aliahmed@health.ucsd.edu

Abstract

Background: The rapid expansion of mobile health (mHealth) apps has transformed health care delivery worldwide. Despite their potential to improve epilepsy care, a substantial treatment gap remains, especially in low- and middle-income countries, due to limited resources, stigma, and low adoption of digital technologies. Although mHealth apps can bridge these disparities, their impact depends on acceptance and use by the target population.

Objective: We aimed to assess the awareness, feasibility, willingness, perception, and factors influencing these behaviors for the usage of mHealth apps among people living with epilepsy in Pakistan.

Methods: We conducted a cross-sectional analytical survey between March and July 2024 among people living with epilepsy attending the Pakistan Institute of Medical Sciences (PIMS). Participants completed a validated, self-administered questionnaire with 33 items across 5 domains. We recruited 406 participants through convenience sampling and analyzed the data using SPSS version 23.0 (IBM Corp). Through multivariable linear regression analysis, we explored factors associated with people living with epilepsy willingness to use mHealth apps. Correlation analysis was used to elucidate the association among awareness, perception, feasibility, and willingness.

Results: Among 406 participants, 53.7% (n=218) were male, 64.5% (n=262) were married, and 89.2% (n=362) were identified as Muslim. Although 86.2% (n=350) of participants have heard about mHealth apps for epilepsy management, 78.1% (n=317) expressed negative perceptions of their use. More than half, 69% (n=280), reported concerns about the privacy of their medical information online, and 78.1% (n=317) were not comfortable using mHealth apps on smartphones or tablets. Multivariable linear regression analysis revealed that rural residents (P=.05), those with a college education (P<.001), and participants with a treatment duration of 2-3 years (P<.001) significantly influenced participants' willingness. Correlation analysis showed a weak negative relationship between awareness and feasibility (ρ=–0.124; P=.01) and a weak positive relationship between awareness and willingness (ρ=0.013; P=.07).

Conclusions: To expand mHealth use for epilepsy care in Pakistan, stakeholders must address concerns about digital literacy, data privacy, and trust. Collaborative efforts involving government, technologists, nongovernmental organizations, academia, and health care providers can improve education, enhance data security, and adapt mHealth tools to local needs, ultimately improving treatment access and outcomes for people living with epilepsy.

(JMIR Form Res 2025;9:e80283) doi:10.2196/80283

XSL•FO
RenderX

KEYWORDS

mobile health apps; epilepsy; feasibility; treatment compliance; digital tool; cross-sectional study

Introduction

Epilepsy is one of the chronic neurological diseases that require special medical care due to susceptibility to experiencing epileptic seizures, resulting in neurological, cognitive, psychological, and social implications [1,2]. The global prevalence of active epilepsy, encompassing both idiopathic and secondary forms, is 658 per 100,000 individuals. The Global Burden of Disease Study 2021 Collaborators found that 80% of the bulk of the epilepsy burden was reported in low- and middle-income countries (LMICs) according to World Bank country income classifications [3]. In Pakistan, the prevalence of epilepsy is estimated to be 9.99/1000 according to a population-based survey in southern Pakistan [4], and recent epidemiology data is unavailable [5]. About 80% of people living with epilepsy live in areas with limited resources, encountering significant barriers to health care services [6]. This makes it exceedingly challenging for them to adhere to daily self-management practices for their condition, including taking medications as directed, keeping consistent sleep patterns, engaging in regular physical activity, and effectively managing stress [7].

The evolution of mobile health (mHealth) apps has transformed them from basic information sources into advanced tools that can diagnose, monitor, and manage chronic neurological diseases [8,9]. mHealth apps, which are within the domain of eHealth and are defined as the usage of mobile and wireless communication technology to enhance health care delivery, outcomes, and research [10,11]. mHealth apps can provide substantial support to people living with epilepsy, encompassing medication reminders and alerts, educational resources, personalized feedback, cost management, medication education, and support [12-14].

In recent years, numerous mobile phone apps have been developed and effectively used to enhance the self-management of individuals with diabetes mellitus, hypertension, and various other conditions [15]. Applications for self-management of epilepsy have been implemented in advanced nations like Australia and the United States [16]. Although Pakistan claims over 90% mobile network coverage, this statistic pertains mostly to basic signal availability and does not reflect smartphone accessibility [17]. National Information and Communication Technologies studies indicate smartphone ownership at approximately 50%-60%, with diminished adoption in rural regions [18]. Digital literacy barriers, such as literacy levels, gender gaps, and socioeconomic disparities, strongly influence whether mobile coverage translates into meaningful use of health applications [19]. These factors form critical preconditions for the feasibility of epilepsy-focused mHealth solutions. Moreover, the extent to which patients are willing to use mHealth apps also affects their widespread adoption and successful integration into current health care systems [20,21]. Shreds of evidence [22,23] have reported that the sustainability of mHealth apps depends on feasibility studies. In LMICs, there is insufficient study investigating public perceptions of mHealth apps [24].

Several international studies have been conducted to evaluate the feasibility and acceptability of these mHealth apps. For example, in Uganda, Tumuhimbise and colleagues [25] found the use of mHealth apps feasible among health care workers to increase patient-based care to control tuberculosis. A pilot intervention study in British Columbia demonstrated the feasibility of the mHealth app among patients, which used automated SMS text messaging from a clinic's computer system [26]. An observational research in South Africa found epileptic-specific mHealth apps feasible among children with refractory epilepsy [27]. These studies collectively indicate the potential of mobile technologies to improve self-management. Despite promising international feasibility evidence, there is a lack of empirical evidence from Pakistan on perception, feasibility, and willingness to use epilepsy-focused mHealth apps. Nonetheless, these contexts are indicative of high- or middle-income environments with a stronger digital infrastructure, which restricts their applicability to Pakistan [28-30]. In Pakistan, the disparities in smartphone ownership, digital literacy, and cultural acceptance of health technologies indicate that feasibility has yet to be examined [8,31]. The absence of context-specific evidence highlights the significance of this study.

We contextualize feasibility within broader access and literacy constraints, and mobile coverage alone is not sufficient. Therefore, this study sought to investigate the feasibility, awareness, and perceptions of mHealth apps among people living with epilepsy and to analyze the factors that may affect this population's willingness to adopt mHealth apps for epilepsy management.

Methods

Overview

The study was reported by following the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) checklist, adhering to the principles established by the Equator network [32]. The STROBE checklist has been attached in Multimedia Appendix 1.

Study Design and Study Setting

This hospital-based quantitative research follows a cross-sectional analytical study, which was conducted among people living with epilepsy from March to July 2024 using a self-administered questionnaire.

The study was conducted at the Neurology department of the Pakistan Institute of Medical Sciences (PIMS), Islamabad [33]. The institute is the largest hospital in Islamabad, with a specialized 24-bed care unit for people living with epilepsy, equipped with the latest neurodiagnostic facilities. The institute is a well-known government hospital representing Pakistan due to its centrality. The Outpatient Department (OPD) serves 80-100 people living with epilepsy, providing free-of-cost treatment and diagnostic services [34].

XSL•FO
RenderX

Study Population

The study participants' having sex (male and female), consisted of people living with epilepsy aged 18 years and older who were able to provide informed consent, had a clinical diagnosis, and had been undergoing treatment for more than 6 months. People living with epilepsy with limited verbal communication, aged more than 65 years, experiencing hearing or cognitive impairments, and with a treatment history of less than 3 months were excluded from the study. Furthermore, the study excluded people living with epilepsy who were unable to complete the questionnaire.

Sample Size and Sampling Technique

The online Rao soft calculator [35] was used to determine the sample size, incorporating a 5% margin of error, a 95% CI, and a 50% response rate. The calculation determined that a total sample size of 379 people was required. To secure a sufficient respondent pool, 427 people living with epilepsy were solicited, resulting in 406 being incorporated into the final analysis. A nonprobability sampling method, specifically convenience sampling, was used due to constraints in gathering data from a certain group of people living with epilepsy at specified times and locations.

Questionnaire Development and Data Collection Procedure

The data-gathering questionnaire was developed in English based on a relevant scientific literature review [36-41]. A preliminary evaluation was performed before commencing the data collection process to assess the content validity, face validity, and internal consistency of the questionnaire. Five pharmacy practice professionals, in conjunction with the head of the Neurology department at PIMS, conducted a comprehensive assessment of the questionnaire to ensure face and content validity. The feedback obtained was used to enhance the questionnaire, ensuring its conformity with the study's aims.

A pilot study with 28 participants was done to ascertain the clarity, comprehensibility, and alignment of the questions with the study's aims. The pilot study results exhibited clarity, readability, and applicability. The feedback from the pilot study was integrated, resulting in the development of the final improved questionnaire before full implementation. Cronbach value of the finalized questionnaire was determined to be 0.723. The pilot study was excluded from the final analysis.

The self-administered questionnaire was comprised of 5 sections containing 33 questions. Throughout the development process, these items were allocated into 5 segments, with certain items dependent on the responses to prior items. The initial segment sought to gather data regarding the demographic attributes, educational background, and treatment records of the participants. The second section aimed to collect data regarding the people living with epilepsy awareness of mHealth apps. In the third section, the people living with epilepsy evaluated their perceptions of mHealth apps using a 5-point Likert scale. The fourth section sought to assess people living with epilepsy comfort and practicality with mHealth apps. The final fifth component aimed to collect data on the willingness, level of agreement, or degree of concern regarding the use of mHealth apps for disease management. BBZ also developed the Urdu-translated version of the questionnaire for Urdu-speaking participants by using the World Health Organization's translation guidelines [42]. Pharmacy practice specialists and neurologists meticulously evaluated the translated versions for clarity and contextual appropriateness.

For recruitment, BBZ was assigned a room within the neurology OPD by the department head in PIMS. After people living with epilepsy had completed their routine consultation with the attending neurologist, those meeting the predefined inclusion and exclusion criteria were approached by BBZ. The study objectives were explained, and informed written consent was obtained. The structured questionnaire was then administered. Each questionnaire lasted approximately 20-25 minutes. In total, 427 individuals were invited to participate, of whom 406 provided complete responses and were included in the final analysis. Twenty-one incomplete questionnaires were excluded, resulting in a study response rate of 95.1%.

Statistical Analysis

The questionnaire was checked for completeness, and data were entered and analyzed using SPSS (IBM Corp). Descriptive statistics were applied to estimate the frequencies and percentages. The categorical data were presented as percentages and frequencies, while means and SDs were computed for continuous variables. The Mann-Whitney *U* test and Kruskal-Wallis test, when appropriate, were used for unadjusted comparisons. To examine the relationship between awareness, perception, feasibility, and willingness, the Spearman rank correlation coefficient was used. The independent relationships between willingness scores and clinical, sociodemographic, and attitudinal characteristics were investigated using multivariable linear regression analysis. Age, gender, type of residence, education level, employment status, family income, duration of treatment, daily pill burden, prescribed antiepileptic medication, awareness, perception, and feasibility scores were all considered independent variables. Variables were selected based on their theoretical relevance and significant associations identified through univariate analysis. Model assumptions, such as linearity, residual normality, homoscedasticity, and multicollinearity, were assessed prior to analysis and were determined to be adequate. A 95% CI and a *P* value of .05 were considered statistically significant. Due to the cross-sectional design of the study, these analyses were interpreted as exploratory associations rather than causal inferences.

Ethical Considerations

The study was conducted in adherence with the guidelines outlined in the Declaration of Helsinki [43]. The Research Ethics Committee of Riphah Institute of Pharmaceutical Sciences, Riphah International University, Islamabad, Pakistan (reference number REC/RIPS/30), and Shaheed Zulfiqar Ali Bhutto University, Islamabad, Pakistan (reference number F.1-1/2015/ERB/SZABMU/1251), approved the research study. Participants were informed about the purpose of the research and its benefit to the population, and informed consent forms were obtained for each participant. The data confidentiality was also assured. No financial or nonfinancial compensation was provided to study participants.

XSL•FO
RenderX

Results

Sociodemographic and Clinical Characteristics of Participants

The predominant study population consists of males (218/406, 53.7%), married individuals (262/406, 64.5%), and middle-aged individuals of 3-45 years (152/406, 30%) of age, with a majority adhering to the Muslim faith (362/406, 89.2%). The majority of the people living with epilepsy underwent treatment for 2-3 years (199/406, 49%) and experienced a daily pill burden of 1-5 pills (257/406, 63.3%), and the predominantly prescribed antiepileptic medicine was valproic acid (297/406, 73.2%). The comprehensive demographic and clinical attributes of the study population are delineated in Table 1.

XSL•FO
RenderX

Table 1. Sociodemographic and clinical characteristics of the participants.

Characteristics	Participants, n (%)
Age range (years)	
18-30	122 (30)
31-45	152 (37.4)
46-60	85 (20.9)
≥61	47 (11.6)
Sex	
Male	218 (53.7)
Female	188 (46.3)
Religion	
Muslim	362 (89.2)
Non- Muslim	44 (10.8)
Marital status	
Married	262 (64.5)
Single	121 (29.8)
Divorced	10 (2.5)
Widowed	13 (3.2)
Type of residence	
Rural	139 (34.2)
Urban	267 (65.8)
Educational background	
Uneducated	47 (11.6)
Primary or secondary	65 (16)
College	226 (55.7)
Bachelors or postgraduation	68 (16.7)
Employment status	
Employed	206 (50.7)
Self-employed	76 (18.7)
Retired	12 (3)
Unemployed	61 (15)
Housewife	51 (12.6)
Family income (US $)	
≤177	223 (54.9)
≥177	167 (41.1)
≥354	16 (3.9)
Duration of treatment	
>6 months	125 (30.8)
2-3 years	199 (49)
≥4 years	82 (20.2)
Daily pill burden	
1-5	257 (63.3)
6-10	121 (29.8)
>10	28 (6.9)

XSL•FO
RenderX

Characteristics	Participants, n (%)
Medicine prescribes (antiepileptics)	
Valproic acid	297 (73.2)
Levetiracetam	44 (10.8)
Carbamazepine	30 (7.4)
Clonazepam	19 (4.7)
Lamotrigine	16 (3.9)

Correlation Among Awareness, Perception, Feasibility, and Willingness About mHealth Apps

There is a very weak positive correlation between awareness and perception (ρ=0.012; *P*=.08). This suggests that as the awareness scores increase, perception scores decrease slightly. Awareness shows a negative correlation with feasibility (ρ=–0.124; *P*=.01), which indicates that increased awareness toward mHealth apps is associated with increased feasibility of using mHealth apps. Awareness is weakly and positively correlated with willingness (ρ=0.013; *P*=.07); however, the correlation is not statistically significant. Perception shows a positive but weak correlation with feasibility (ρ=0.189; *P* .001), and this correlation is statistically significant.

Awareness About the mHealth Apps

The majority of the participants, 86.2% (350/406), reported that they have heard about mHealth apps used for epilepsy management. Similarly, 86.2% (350/406) of participants were aware of the potential benefits of mHealth apps. About 72.9% (296/406) of respondents used mHealth apps and found them helpful. In contrast, 57.1% (232/406) of participants have not entered information into mobile apps related to disease and medication, highlighting a deficiency in the use of technology for managing chronic diseases (Table 2).

Table 2. Awareness about the mobile health apps.

Statements	Yes, n (%)	No, n (%)	I don't know, n (%)
Have you ever heard of mobile health apps for epilepsy management?	350 (86.2)	49 (12.1)	7 (1.7)
Are you familiar with specific mobile apps for tracking seizures, medications, or other aspects of epilepsy management?	343 (84.5)	39 (9.6)	24 (5.9)
Have you ever used a mobile health app to monitor your epilepsy symptoms, medication, or overall health?	313 (77.1)	58 (14.3)	35 (8.6)
If you have used any mobile health apps for epilepsy, did you find the app features helpful?	296 (72.9)	92 (22.7)	18 (4.4)
Are you aware of the potential benefits of using mobile health apps for medication adherence?	350 (86.2)	36 (8.9)	20 (4.9)
Have you ever entered information into a mobile appn about your disease, medications, and diet?	133 (32.8)	232 (57.1)	41 (10.1)

Perception, Feasibility, and Willingness About mHealth Apps

Overall, 69.2% (282/406) of participants were not open to using mHealth apps as an additional tool for epilepsy management. More than half of the participants (317/406, 78.1%) perceived that mHealth apps are not a valuable addition to their epilepsy management toolkit. In total, 75.9% (308/406) of the participants reported that mHealth apps can make medical care cheaper and faster. Meanwhile, 77.3% (314/406) reported that mHealth apps can save them from transportation costs, offering them significant financial relief. Around 69% (280/406) of participants had privacy concerns regarding medical information security offered by mHealth apps.

A total of 82.8% (336/406) of respondents indicated challenges in the usability of mHealth apps, and due to this reason, more than half participants (317/406, 78.1%) were not comfortable while using the mHealth apps on smartphones or tablets. Similarly, 79.3% (322/406) of participants reported that they did not find the app feasible or convenient for managing their disease. Overall, 36.7% (149/406) were not interested, and 73.4% (298/406) did not prefer mHealth apps for epilepsy management. Detailed perception, feasibility, and willingness about mHealth apps can be seen in Tables 3-5.

XSL•FO
RenderX

Table 3. Perception response of participants toward mobile health apps.

Statement and response	Strongly disagree, n (%)	Disagree, n (%)	Neutral, n (%)	Agree, n (%)	Strongly agree, n (%)
Are you open to using mobile health apps as an additional tool for managing epilepsy?	281 (69.2)	59 (14.5)	15 (3.7)	28 (6.9)	23 (5.7)
Do you think a mobile health app is a valuable addition to your epilepsy management toolkit?	317 (78.1)	14 (3.4)	35 (8.6)	31 (7.6)	9 (2.2)
Do you think mobile health apps make medical care cheaper and faster?	2 (0.5)	34 (8.4)	308 (75.9)	36 (8.9)	26 (6.4)
Do you think mobile health apps save money and transportation costs?	3 (0.7)	43 (10.6)	24 (5.9)	314 (77.3)	22 (5.4)
Do you think mobile health apps can improve and increase antiepileptic medication adherence?	18 (4.4)	294 (72.4)	44 (10.8)	39 (9.6)	11 (2.7)
Do you think mobile health apps increase communication with healthcare providers?	16 (3.9)	310 (76.4)	42 (10.3)	25 (6.2)	13 (3.2)
Do you think using mobile health apps to access health care offers great security?	25 (6.2)	25 (6.2)	280 (69.0)	68 (16.7)	8 (2)
Do you think it is an important tool that can be used to complement treatment strategies?	16 (3.9)	286 (70.4)	44 (10.8)	50 (12.3)	10 (2.5)

Table 4. Feasibility response of participants toward mobile health apps.

Statement and response	Participants, n (%)
How comfortable are you in using smartphones or tablets to manage your epilepsy?	
Very uncomfortable	2 (0.5)
Uncomfortable	317 (78.1)
Neutral	53 (13.1)
Comfortable	18 (4.4)
Very comfortable	16 (3.9)
Do you think it is easy to handle a mobile health app?	
Strongly disagree	2 (0.5)
Disagree	5 (1.2)
Neutral	336 (82.8)
Agree	62 (15.3)
Strongly agree	1 (0.2)
Do you think it is convenient to use and share medical information on mobile health apps?	
Strongly disagree	2 (0.5)
Disagree	26 (6.4)
Neutral	322 (79.3)
Agree	53 (13.1)
Strongly agree	3 (0.7)
Would you prefer the mobile app to have medication reminders and tracking features?	
Strongly do not prefer	298 (73.4)
Do not prefer	44 (10.8)
Neutral	27 (6.7)
Prefer	17 (4.2)
Strongly prefer	20 (4.9)

XSL•FO
RenderX

Table 5. Willingness response of participants toward mobile health apps.

Statement and response	Participants, n (%)
I feel positive about using mobile health apps to manage my epilepsy condition	
Strongly disagree	11 (2.7)
Disagree	32 (7.9)
Neutral	117 (28.8)
Agree	142 (35)
Strongly agree	104 (25.6)
I am interested in using mobile health apps to monitor medication adherence	
Not interested at all	8 (2)
Not very interested	47 (11.6)
Neutral	114 (28.1)
Somewhat interested	149 (36.7)
Very interested	88 (21.7)
I am worried about my privacy and confidentiality if my health information is online	
Not worried at all	17 (4.2)
Not very worried	40 (9.9)
Neutral	122 (30)
Worried	146 (36)
Very worried	81 (20)
I prefer digital health tools over conventional hospital visits	
Strongly do not prefer	19 (4.7)
Do not prefer	48 (11.8)
Neutral	94 (23.2)
Prefer	149 (36.7)
Strongly prefer	96 (23.6)

Impact of Sociodemographic and Clinical Characteristics on Awareness, Perception, Feasibility, and Willingness About mHealth Apps

The median level of appropriate awareness among people living with epilepsy possessing a college education exceeded that of people living with epilepsy with primary or secondary education or those with bachelor's or postgraduate degrees (*P* .001). The awareness of mHealth apps among people living with epilepsy was higher for the self-employed compared to retirees, the unemployed, employees, and housewives (*P* .001). Statistically significant differences were also found between awareness of mHealth apps and duration of treatment, which was higher in people living with epilepsy whose duration of treatment was ≤1 year as compared to 2-3 years or ≥4 years (*P* .001). The perception score was higher in people living with epilepsy aged 31-45 years as compared with those aged 18-30 years, 46-60 years, and ≥61 years (*P*=.04). Furthermore, the perception score was elevated in people living with epilepsy whose family income was ≥$354 compared to those with a family income of ≥$177 and ≤$177(*P*=.01). People living with epilepsy having a daily 6-10 pill burden have greater perception scores than those with a daily pill burden of 1-5 pills and 10 pills (*P* .001).

The feasibility score for using mHealth apps among people living with epilepsy with bachelor's or postgraduate degrees surpassed that of people living with epilepsy with primary or secondary education, as well as those with college degrees and those who are uneducated (*P* .001). The feasibility score was also greater in people living with epilepsy having a family income of ≤$177 as compared to a family income of ≥$177 and ≥$354 (*P*≥.99). Moreover, people living with epilepsy having a treatment duration of ≥4 years have a higher feasibility score than those with 1 year and 2-3 years (*P* .001). The willingness score was greater among participants aged 31-45 years as compared to those aged 18-30 years, 46-60 years, and ≥61 years (*P*=.03). People living with epilepsy following the Muslim religion has a higher willingness toward mHealth apps as compared to non-Muslims (*P* .001). People living with epilepsy

XSL•FO
RenderX

from urban residential areas exhibit a higher willingness score compared to those from rural locations ($P \geq .99$). Comprehensive findings about the impact of demographic and clinical variables on the awareness, perception, feasibility, and willingness to adopt mHealth apps are provided in Table 6.

XSL•FO
RenderX

Table 6. Comparison of sociodemographic and clinical characteristics with awareness, perception, feasibility, and willingness of participants toward mobile health apps.

Characteristics	Participants, n (%)	Awareness		Perception		Feasibility		Willingness	
		Median (IQR)	*P* value	Median (IQR)	*P* value	Median (IQR)	*P* value	Median (IQR)	*P* value
Age (years)			.04		.002		.002		.03
18-30	122 (30)	5 (3-5)		18 (18-19.25)		9 (9-11)		15 (12-17)	
31-45	152 (37.4)	5 (4-6)		19 (18-23)		9 (9-9)		15 (12-18)	
46-60	85 (20.9)	5 (3-5)		18 (18-20)		9 (9-10)		15 (14-17)	
≥61	47 (11.6)	5 (4-6)		18 (18-20)		9 (9-11)		13 (11-16)	
Sex			.65		<.001		.18		.80
Male	218 (53.7)	5 (4-6)		19 (18-22)		9 (9-10)		1 5 (12-17)	
Female	188 (46.3)	5 (3-5)		18 (18-19.75)		9 (9-10)		15 (12-17)	
Religion			.05		.09		.56		<.001
Muslim	362 (89.2)	5 (4-5)		18 (18-21)		9 (9-10)		15 (12-17)	
Non-Muslim	44 (10.8)	4 (2.25-5)		18 (17.25-19)		9 (9-11)		12.50 (10-15)	
Marital status			.09		.03		.12		.10
Married	262 (64.5)	5 (3-5)		19 (18-21)		9 (9-10)		15 (12-17)	
Single	121 (29.8)	5 (4-5)		18 (18-20)		9 (9-10)		15 (12-17)	
Divorced	10 (2.5)	5 (2.75-5)		18 (17.75-24.50)		10.50 (9-15.25)		13.50 (9.50-14)	
Widowed	13 (3.2)	5 (5-6)		18 (17-18)		9 (9-10)		13 (12-14)	
Type of residence			.06		.39		.44		.006
Rural	139 (34.2)	5 (4-6)		19 (18-20)		9 (9-10)		14 (12-16)	
Urban	267 (65.8)	5 (3-5)		18 (18-20)		9 (9-10)		15 (12-18)	
Educational background			.001		.17		<.001		<.001
Uneducated	47 (11.6)	5 (3-5)		19 (18-21)		9 (9-10)		15 (12-16)	
Primary or Secondary	65 (16)	4 (3-5)		18 (18-20)		10 (9-12.50)		14 (11-16)	
College	226 (55.7)	5 (4-6)		18.50 (18-21)		9 (9-9)		16 (14-19)	
Bachelors or post-graduation	68 (16.7)	5 (3.25-6)		18 (18-19)		9 (9-14.75)		12 (10-15)	
Employment status			<.001		.001		.74		.01
Employed	206 (50.7)	5 (3-5)		18 (18-20)		9 (9-10)		16 (13-18)	
Self-employed	76 (18.7)	5 (5-6)		19 (18-21.75)		9 (9-10)		13.50 (11-16)	
Retired	12 (3)	5 (4-6)		18 (16.25-24.75)		9.50 (9-13)		13.50 (12-15.75)	
Unemployed	61 (15)	4 (3-5)		19 (18-23)		9 (9-11.50)		15 (12-17)	
Housewife	51 (12.6)	4 (3-5)		18 (18-18)		9 (9-9)		15 (13-17)	
Family income (US dollars)			.19		.02		.002		.002
≤$177	223 (54.9)	5 (3-5)		19 (18-22)		9 (9-11)		15 (13-17)	
≥$177	167 (41.1)	5 (3-5)		18 (18-20)		9 (9-9)		14 (12-16)	
≥$354	16 (3.9)	5 (5-6)		18 (18-23.50)		9 (9-9)		16.50 (15.25-19)	
Duration of treatment (years)			.001		.45		<.001		<.001

XSL•FO
RenderX

Characteristics	Participants, n (%)	Awareness		Perception		Feasibility		Willingness	
		Median (IQR)	*P* value	Median (IQR)	*P* value	Median (IQR)	*P* value	Median (IQR)	*P* value
≤1	125 (30.8)	5 (4-6)		18 (18-21)		9 (9-9)		15 (13-17)	
2-3	199 (49)	5 (3-5)		18 (18-20)		9 (9-10)		16 (13-19)	
≥4	82 (20.2)	4.50 (3-5)		18.50 (17-20.25)		10.50 (9-13.25)		12 (10-14)	
Daily pill burden			<.001		<.001		.41		.95
1-5	257 (63.3)	5 (3-5)		18 (18-20)		9 (9-10)		15 (12-17)	
6-10	121 (29.8)	5 (4-6)		19 (18-24.50)		9 (9-10.50)		15 (12-17)	
>10	28 (6.9)	4 (3-5)		19 (17.25-21.75)		9 (9-10.75)		14.50 (12-19)	
Medicine prescribes (antiepileptic)			.10		.22		<.001		<.001
Valproic acid	297 (73.2)	5 (4-6)		18 (18-20)		9 (9-10)		15 (13-18)	
Levetiracetam	44 (10.8)	4 (3-5)		18 (18-22)		9 (9-10)		15 (12-17)	
Carbamazepine	30 (7.4)	4 (3.75-5)		18 (18-19)		9 (9-10)		12 (11-14.25)	
Clonazepam	19 (4.7)	4 (3-5)		19 (16-23)		11 (10-16)		12 (10-14)	
Lamotrigine	16 (3.9)	5 (3.25-6)		18 (18-19)		9 (9-11)		13 (9.25-14.75)	

Factors Associated With Participants' Willingness to Use mHealth Apps

Multivariable linear regression analysis showed that the willingness to use mHealth apps differed among participants with varying educational backgrounds, residences, and family incomes. Rural resident participants have lower willingness than urban residents, with a *P* value of .05, which is borderline statistically significant. Participants having a college educational background have a significantly higher willingness (*P* .001) as compared to a bachelor's or postgraduate degree. There was no statistically significant effect for uneducated or primary or secondary educated individuals. People living with epilepsy whose family income is ≥$177 show significantly lower willingness (*P*=.01). People living with epilepsy having a treatment duration of ≤1 year (*P*=.001) and 2-3 years (*P* .001) show significantly higher willingness as compared to those on treatment for ≥4 years. Other demographic and clinical factors, and awareness, perception, and feasibility scores do not show statistically significant influence. The findings from the multivariable linear regression analysis are summarized in Table 7.

XSL•FO
RenderX

Table 7. Multivariable linear regression analysis of the factors influencing the willingness of participants to use mobile health apps.

Characteristics	Participants, n (%)	Unstandardized coefficients, β (95% CI)	Standardized coefficients (β)	*P* value
Age range (years)				
18-30	122 (30)	0.446 (–0.81 to 1.70)	0.058	.49
31-45	152 (37.4)	–0.361 (–1.62 to 0.89)	–0.049	.57
46-60	85 (20.9)	0.738 (–0.58 to 2.06)	0.085	.28
≥61	47 (11.6)	Reference	—[a]	—
Sex				
Male	218 (53.7)	–0.025 (–0.75 to 0.70)	–0.004	.95
Female	188 (46.3)	Reference	—	—
Religion				
Muslim	362 (89.2)	0.275 (–0.83 to 1.38)	0.024	.63
Non-Muslim	44 (10.8)	Reference	—	—
Marital status				
Married	262 (64.5)	0.521 (–1.61 to 2.65)	0.070	.63
Single	121 (29.8)	0.155 (–2.02 to 2.34)	0.020	.89
Divorced	10 (2.5)	–0.374 (–3.22 to 2.47)	–0.016	.80
Widowed	13 (3.2)	Reference	—	—
Type of residence				
Rural	139 (34.2)	–0.729 (–1.47 to 0.01)	–0.098	.05
Urban	267 (65.8)	Reference	—	—
Educational background				
Uneducated	47 (11.6)	0.889 (–0.38 to 2.16)	0.080	.17
Primary or Secondary	65 (16)	0.530 (–0.69 to 1.75)	0.055	.39
College	226 (55.7)	2.27 (1.23-3.30)	0.319	<.001
Bachelor's or postgraduation	68 (16.7)	Reference	—	—
Employment status				
Employed	206 (50.7)	0.678 (–0.56 to 1.91)	0.096	.28
Self-employed	76 (18.7)	–0.426 (–1.76 to 0.91)	–0.047	.53
Retired	12 (3)	–0.143 (–2.44 to 2.15)	–0.007	.90
Unemployed	61 (15)	–0.008 (–1.42 to 1.41)	–0.001	.99
Housewife	51 (12.6)	Reference	—	—
Family income (US dollars)				
≤177	223 (54.9)	–1.648 (–3.36 to 0.07)	–0.232	.06
≥177	167 (41.1)	–2.186 (–3.92 to –0.44)	–0.304	.01
≥354	16 (3.9)	Reference	—	—
Duration of treatment				
≤1 year	125 (30.8)	1.964 (0.83-3.09)	0.257	.001
2-3 years	199 (49)	2.203 (1.20-3.19)	0.311	<.001
≥4 years	82 (20.2)	Reference	—	—
Daily pill burden				
1-5	257 (63.3)	0.023 (–1.34 to 1.38)	0.003	.97
6-10	121 (29.8)	0.260 (–1.16 to 1.68)	0.034	.72
>10	28 (6.9)	Reference	—	—

XSL•FO
RenderX

Characteristics	Participants, n (%)	Unstandardized coefficients, β (95% CI)	Standardized coefficients (β)	*P* value
Medicine prescribes (antiepileptics)				
Valproic acid	297 (73.2)	1.381 (–0.35 to 3.11)	0.173	.12
Levetiracetam	44 (10.8)	1.778 (–0.14 to 3.70)	0.156	.07
Carbamazepine	30 (7.4)	–0.265 (–2.27 to 1.74)	–0.020	.80
Clonazepam	19 (4.7)	0.616 (–1.54 to 2.77)	0.037	.58
Lamotrigine	16 (3.9)	Reference	—	—
Awareness score				
Continuous	406 (100)	–0.058 (–0.31 to 0.20)	–0.022	.66
Perception score				
Continuous	406 (100)	–0.041 (–0.11 to 0.03)	–0.053	.29
Feasibility score				
Continuous	406 (100)	–0.039 (–0.20 to 0.12)	–0.025	.63

[a]Not applicable.

Discussion

Principal Findings

This study found that while most participants have heard about mHealth apps used for epilepsy management, the majority expressed concerns about privacy and security associated with these apps. More than half of the participants felt uncomfortable while using them due to challenges faced in the usability of mHealth apps. Similarly, the majority of participants reported that mHealth apps were not a valuable addition to their epilepsy management toolkit. In contrast, the study findings indicate that middle-aged urban residents with higher educational attainment and extended durations of epilepsy treatment exhibit a greater self-reported willingness to embrace mHealth apps as a digital health tool. This study highlights that awareness of mHealth apps and trial use do not necessarily translate into active, disease-focused usage.

The study findings indicated that participants with a high level of education were more likely to use mHealth apps for epilepsy management. Our findings are in line with the results of previous studies [44,45]. A considerable number of individuals with lower levels of education face significant health care needs; yet, they frequently struggle to comprehend the information provided on mHealth apps [46]. Ongoing endeavors to guarantee that mHealth apps are user-friendly and that health and treatment information is accessible, comprehensible, and retrievable are essential.

Participants living in rural regions exhibit a reduced willingness to engage with mHealth apps. This result supports earlier findings [47,48] indicating that the urban population exhibited a greater level of agreement regarding the use of mHealth apps. In resource-limited settings, traditional forms of health care provision are used more than advanced health care services [49]. The implementation of mHealth apps is difficult in LMICs due to the perpetuation of digital divides [50]. Therefore, by increasing the accessible treatment facilities and addressing the digital divides, we may assist the rural population affected by epilepsy in benefiting from these apps.

In total, 50% of the participants in the study expressed apprehension about sharing their medical information online, as this heightens their privacy concerns over the usage of mHealth apps that monitor their symptoms and facilitate illness treatment, which is compatible with the study conducted in Germany [51]. The transmission of patient information via digital tools presents legal concerns. Encryption must be adequate to comply with the Health Insurance Portability and Accountability Act (HIPAA) standards, ensuring patient confidentiality and the security of personal data [52]. Awareness about data privacy and the availability of medical information of people living with epilepsy should be provided to people living with epilepsy visiting hospitals for follow-ups, and privacy-preserving edge federated learning should be incorporated in mobile health systems [53].

In the study, we found that people living with epilepsy have a negative perception regarding the use of mHealth apps for epilepsy management. The negative perception surrounding the digitalization of the health care system arises from various cultural, systematic, and technological barriers, including stigma and inadequate digital health literacy [54]. Previous research [55] indicates the disparity between the creation and implementation of mHealth apps and the patients for whom they are designed. We recommend that the government and health care institutions take proactive steps to establish a digital literacy program aimed at educating people living with epilepsy and their caregivers on how to effectively access and navigate the mHealth apps, ultimately leading to improved treatment outcomes.

A recently published systematic review [8] indicates that people living with epilepsy are generally cognizant of mHealth apps, but there is a deficiency in comprehension of the app features and benefits that could facilitate timely disease management. This lack of understanding may impede the usage of these apps and exacerbate the atypical concerns of people living with epilepsy, which is consistent with previous research findings [56-58]. Awareness and practical implementation of mHealth app features for epilepsy management should be initiated in

XSL•FO
RenderX

government hospitals to alleviate the disease burden on the impoverished population affected by epilepsy in Pakistan.

We also found that occupation correlates with the viability of using mHealth apps among people living with epilepsy. This study indicated that self-employed and employed participants were inclined to use mHealth apps. This is attributed to the high educational and socioeconomic status that facilitates their access to mHealth apps [59]. In contrast, the unemployed participants demonstrated a lower feasibility for using the mHealth app. This finding was similar to a multiethnic study [60] in the United States, which indicates that individuals from lower socioeconomic backgrounds had limited access to advanced treatment options.

A significant portion of the participants expressed discomfort and demonstrated a decreased preference for mHealth apps that include medication reminders and tracking functionalities. Previous studies [61,62] have demonstrated the significance of using mHealth apps in the management of chronic diseases. A key difficulty for 50%-60% of chronically ill persons is medication adherence [63]. mHealth apps appear to be a promising strategy for addressing issues related to medication nonadherence and enhancing treatment outcomes [64]. To improve medication adherence and management in people living with epilepsy, it is crucial to offer guidance by health care providers on mHealth apps for those receiving care in hospitals with suboptimal treatment outcomes for epilepsy.

Our findings may further substantiate the necessity for rules and frameworks for the creation and categorization of epilepsy apps within the existing epilepsy app ecosystem for less formally educated individuals [65,66]. Moreover, studies reported [67,68] challenges in conveying referrals to health care facilities due to limited knowledge and obstacles encountered by the uneducated people living with epilepsy. The regular enhancement of digital literacy skills and education for uneducated people living with epilepsy in need is essential to optimize the efficacy of an mHealth intervention.

In this study, people living with epilepsy prefer the mHealth apps to the conventional hospital visits, as the majority of participants belong to rural areas. Frequent hospital visits can incur significant expenses, especially in rural regions, owing to transportation expenditures [69]. A comprehensive systematic literature review [70] substantiates that mHealth apps demonstrate significant potential in assisting individuals and health care systems amid the rising prevalence and financial burdens of chronic diseases globally by providing informational support and an easier way to contact health care providers [71]. Consequently, the incorporation of mHealth apps may assist people living with epilepsy in managing their condition by alleviating financial strains and treatment burden.

Our findings indicate a notable level of awareness regarding mHealth apps in Pakistan when compared to various other LMICs. However, engagement with functions tailored to specific diseases remains limited. This aligns with global observations, as documented in the systematic review by Gotlieb et al [8]. The analysis emphasizes that, although feasibility and acceptability appear encouraging, genuine adoption and sustained use are dependent upon factors such as usability, literacy, and support from the health system. This highlights the necessity for strategies that are specifically tailored to the context of Pakistan. The comparison of the study data with previously published research has been elaborated in Table 8.

Table 8. Comparison of the study data with previously published studies targeting mobile health apps.

Study	Country and setting	Population	Sample Size	Key variables measured	Awareness and use of mHealth apps	Active engagement	Key findings or predictors
Our study	Pakistan, [urban+rural] clinic	Adults with epilepsy on treatment	406	• Awareness • Perception • Feasibility • Willingnes • Treatment • Duration • Education	77.1% reported having used mHealth apps	42.9% input disease or medication data	Higher self-reported willingness among middle-aged, urban, higher-educated participants
Patterson et al [72], 2022	Pakistan and United Kingdom	Patients with epilepsy referred via health workers	59	• Diagnostic accuracy • Referral timing	Not applicable (usage by health workers)	Not applicable	High appropriateness of referrals; timeliness improved
Alzamanan et al [73], 2024	Malaysia	Expert panel & literature review	-	• Domains or features of self-management apps	Not applicable	Framework focus, not direct patient use	Identified key domains; usability rated highly among experts
Gotlieb et al [8], 2025	Multiple countries	Mixed epilepsy populations	20 studies in review	• Effectiveness • Feasibility • Usability	Varied installation and uptake	Mixed engagement with disease-specific functions	Barriers: usability, literacy, infrastructure; predictors include usefulness, interface quality

XSL•FO
RenderX

Limitations

The study presents multiple limitations. First, we used convenience sampling, which results in limited generalizability. Second, data were collected from PIMS, primarily frequented by people living with epilepsy with low educational backgrounds resulting from illiteracy; the participants exhibited a lack of focus on the study's objectives. Third, in the questionnaire, certain questions were addressed on behalf of a people living with epilepsy. As a result, the responses may be influenced by the perspective of the person conducting the survey on behalf of a patient, leading to self-reported data. Fourth, this study is deficient in qualitative insights and factor analytic approaches. Moreover, the assessment of willingness relied on self-reporting instead of direct behavioral measures. While face and content validity were established, no formal psychometric validation (eg, factor analysis) of the feasibility, perception, and willingness scales was performed. Due to the execution of several subgroup comparisons without formal adjustment, there exists a risk of increased type I error; hence, these findings should be interpreted as exploratory and hypothesis-generating rather than confirmatory.

Future Implications

In the future, policymakers, health care providers, and app developers need to work together in co-designing and implementing evidence-based strategies for the use of mHealth apps, while also ensuring the protection of privacy within these apps. Addressing public health concerns necessitates the implementation of digital literacy sessions and training programs. Future research may concentrate on longitudinal and interventional studies, examining the functionality and efficacy of these mHealth apps with larger sample sizes, alongside user observations and perspectives, particularly highlighting the dynamics of user-platform interaction, to evaluate the causal pathways of mHealth usage and the establishment of trustworthiness. Moreover, it is recommended to conduct psychometric validation of the feasibility, perception, and willingness scales using factor analytic methods. Objective behavioral metrics, such as app usage logs, need to be incorporated to validate self-reported willingness. Probability sampling across multiple centers should be used to improve representativeness.

Conclusions

This study demonstrates that urban residents with higher socioeconomic status were more likely to use mHealth apps, although they raised concerns about privacy protection. Stakeholders should implement a robust privacy policy in these apps. Future research may focus on addressing the digital divides, along with cultural, systemic, and technological challenges, as informed by the identified impacting factors. To improve the comfort of people living with epilepsy with mHealth apps, it is crucial to implement training programs and digital literacy sessions for both users and health care providers, focused on enhancing the functionality and effectiveness of mHealth apps.

Acknowledgments

We thank the head of the neurology department, Dr. Mazhar Badshah, for his valuable suggestions in the development of the questionnaire, the administration of PIMS for their assistance, and people living with epilepsy, who cooperated with us during our research work there.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Data Availability

The datasets used and analyzed during the current study are available from the corresponding author upon reasonable request.

Authors' Contributions

AA and BBZ contributed to conceptualization, investigation, and methodology. BBZ performed data collection and writing–original draft. BBZ and MAH handled the data curation and formal analysis. BBZ, AA, and MAH managed software. AA provided supervision. AA, MAH, RS, HBZ, and AA writing–review and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
STROBE checklist.
[PDF File (Adobe PDF File), 158 KB - formative_v9i1e80283_app1.pdf]

References

XSL•FO
RenderX

1. Sezgin E, Noritz G, Lin S, Huang Y. Feasibility of a voice-enabled medical diary app (SpeakHealth) for caregivers of children with special health care needs and health care providers: mixed methods study. JMIR Formative Research 2021;5(5):e25503 [FREE Full text] [doi: 10.2196/25503] [Medline: 33865233]
2. Catalán-Aguilar J, González-Bono E, Cano-López I. Perceived stress in adults with epilepsy: a systematic review. Neuroscience and Biobehavioral Reviews 2025;170:106065 [FREE Full text] [doi: 10.1016/j.neubiorev.2025.106065] [Medline: 39961554]
3. Huang J, Man Y, Shi Z, Fu X, Shi W, Liang X. Global, regional, and national burden of maternal disorders, 1990-2021: a systematic analysis from the global burden of disease study 2021. BMC Public Health 2025;25(1):2576 [FREE Full text] [doi: 10.1186/s12889-025-23814-w] [Medline: 40730990]
4. Aziz H, Ali SM, Frances P, Khan MI, Hasan KZ. Epilepsy in Pakistan: a population-based epidemiologic study. Epilepsia 1994;35(5):950-958. [doi: 10.1111/j.1528-1157.1994.tb02539.x] [Medline: 7925166]
5. Tanveer M, Tahir A, Iqbal M, Aslam F, Ahmed A. Health-related quality of life and medication adherence of people living with epilepsy in Pakistan: a cross-sectional study. Brain Behavior 2023;13(9):e3127 [FREE Full text] [doi: 10.1002/brb3.3127] [Medline: 37515419]
6. Ali A, Clarke DF. Digital measures in epilepsy in low-resourced environments. Expert Rev Pharmacoeconomic Outcomes Research 2024;24(6):705-712. [doi: 10.1080/14737167.2023.2270163] [Medline: 37818647]
7. Moalong K, Espiritu A, Fernandez M, Jamora R. Treatment gaps and challenges in epilepsy care in the Philippines. Epilepsy Behavior 2021;115:107491. [doi: 10.1016/j.yebeh.2020.107491] [Medline: 33323340]
8. Gotlieb E, Marzoughi S, Kwon C, Harmon M, Kimura M, Truesdale A, et al. Clinical effectiveness, feasibility, acceptability, and usability of mobile health applications for epilepsy: a systematic review. Epilepsia 2025;66(5):1349-1373. [doi: 10.1111/epi.18288] [Medline: 39945400]
9. Fatehi V, Salahzadeh Z, Mohammadzadeh Z. Mapping and analyzing the application of digital health for stroke rehabilitation: scientometric analysis. Disability and Rehabilitation: Assistive Technology 2025;20(2):321-330. [doi: 10.1080/17483107.2024.2387101] [Medline: 39140131]
10. Bernard-Willis Y, Oliveira ED, Lakhan SE. An overview of digital health in the transition of pediatric to adult epilepsy care. Journal of Pediatric Epilepsy 2020;09(04):106-113. [doi: 10.1055/s-0040-1716825]
11. Sánchez Rodríguez M, Collado Vázquez S, Martín Casas P, Cano de la Cuerda R. Neurorehabilitation and apps: a systematic review of mobile applications. Neurología (English Edition) 2018;33(5):313-326. [doi: 10.1016/j.nrleng.2015.10.002]
12. Nevin SM, Wakefield CE, Schilstra CE, McGill BC, Bye A, Palmer EE. The information needs of parents of children with early-onset epilepsy: a systematic review. Epilepsy Behavior 2020;112:107382. [doi: 10.1016/j.yebeh.2020.107382] [Medline: 32854014]
13. Choi S, Lim K, Baek H, Yoo S, Cho A, Kim H, et al. Impact of mobile health application on data collection and self-management of epilepsy. Epilepsy Behavior 2021;119:107982. [doi: 10.1016/j.yebeh.2021.107982] [Medline: 33946011]
14. Feldman AG, Beaty BL, Moore SL, Bull S, Wilson K, Atkinson KM, et al. Feasibility, acceptability, and effectiveness of a smartphone app to increase pretransplant vaccine rates: usability study. JMIR Formative Research 2025;9:e68855 [FREE Full text] [doi: 10.2196/68855] [Medline: 40237553]
15. Jiang J, Zhu Q, Zheng Y, Zhu Y, Li Y, Huo Y. Perceptions and acceptance of mhealth in patients with cardiovascular diseases: a cross-sectional study. JMIR Mhealth Uhealth 2019;7(2):e10117 [FREE Full text] [doi: 10.2196/10117] [Medline: 30714942]
16. Liu X, Wang R, Zhou D, Hong Z. Feasibility and acceptability of smartphone applications for seizure self-management in China: questionnaire study among people with epilepsy. Epilepsy Behavior 2016;55:57-61. [doi: 10.1016/j.yebeh.2015.11.024] [Medline: 26745631]
17. Khan UI, Shah S, Qureshi A, Viswanathan S, Merchant AT, Virani SS, et al. Burden of cardiometabolic diseases and depression in a low-income, urban community in Pakistan: a cross-sectional survey. BMC Public Health 2025;25(1):757 [FREE Full text] [doi: 10.1186/s12889-025-21939-6] [Medline: 39994704]
18. Abid M, Chaudhry M. Socioeconomic determinants of digital technology adoption: a case study of south Punjab, Pakistan. Pakistan Journal of Social Sciences (PJSS) 2024;44(2).
19. Kazi AM, Ahsan N, Jamal S, Khan A, Mughis W, Allana R, et al. Characteristics of mobile phone access and usage among caregivers in Pakistan - A mHealth survey of urban and rural population. International Journal of Medical Informatics 2021;156:104600. [doi: 10.1016/j.ijmedinf.2021.104600] [Medline: 34638012]
20. van Velsen L, Flierman I, Tabak M. The formation of patient trust and its transference to online health services: the case of a Dutch online patient portal for rehabilitation care. BMC Medical Informatics and Decision Making 2021;21(1):188 [FREE Full text] [doi: 10.1186/s12911-021-01552-4] [Medline: 34118919]
21. Ncube B, Mars M, Scott RE. Perceptions and attitudes of patients and healthcare workers towards the use of telemedicine in Botswana: an exploratory study. PLoS One 2023;18(2):e0281754 [FREE Full text] [doi: 10.1371/journal.pone.0281754] [Medline: 36795740]
22. van Leersum CM, Bults M, Siebrand E, Olthuis TJJ, Bekhuis REM, Konijnendijk AAJ, et al. Exploring opportunities and challenges for the spread, scale-up, and sustainability of mhealth apps for self-management of patients with type 2 diabetes

XSL•FO
RenderX

mellitus in the Netherlands: citizen science approach. JMIR Diabetes 2024;9:e56917 [FREE Full text] [doi: 10.2196/56917] [Medline: 39689302]
23. Kaboré SS, Ngangue P, Soubeiga D, Barro A, Pilabré AH, Bationo N, et al. Barriers and facilitators for the sustainability of digital health interventions in low and middle-income countries: a systematic review. Frontiers Digit Health 2022;4:1014375 [FREE Full text] [doi: 10.3389/fdgth.2022.1014375] [Medline: 36518563]
24. Gholamzadeh M, Abtahi H, Safdari R, Haghshenas E, Gharabaghi M, Rahimi B. Attitudes, awareness, and willingness of iranians toward using telemedicine in post-COVID era: a cross-sectional study. Health Science Reports 2025;8(1):e70321 [FREE Full text] [doi: 10.1002/hsr2.70321] [Medline: 39777285]
25. Tumuhimbise W, Atwine D, Kaggwa F, Musiimenta A. Acceptability and feasibility of a mobile health application for enhancing public private mix for TB care among healthcare Workers in Southwestern Uganda. BMC Digital Health 2023;1(1):9 [FREE Full text] [doi: 10.1186/s44247-023-00009-0] [Medline: 38014370]
26. van der Kop ML, Memetovic J, Smillie K, Coleman J, Hajek J, Van Borek N, et al. Use of the WelTel mobile health intervention at a tuberculosis clinic in British Columbia: a pilot study. Journal of Mobile Technology in Medicine 2013;2(3):7-14. [doi: 10.7309/jmtm.2.3.3]
27. Davies E, Fieggen K, Wilmshurst J, Anyanwu O, Burman R, Komarzynski S. Demonstrating the feasibility of digital health to support pediatric patients in South Africa. Epilepsia Open 2021;6(4):653-662 [FREE Full text] [doi: 10.1002/epi4.12527] [Medline: 34310860]
28. Bruno E, Simblett S, Lang A, Biondi A, Odoi C, Schulze-Bonhage A, RADAR-CNS Consortium. Wearable technology in epilepsy: the views of patients, caregivers, and healthcare professionals. Epilepsy Behavior 2018;85:141-149. [doi: 10.1016/j.yebeh.2018.05.044] [Medline: 29940377]
29. Peterson CL, Coleman H, Walker C. Needs for aids and equipment for the management of epilepsy in an Australian cohort. Epilepsy Behavior 2020;108:107062. [doi: 10.1016/j.yebeh.2020.107062] [Medline: 32375096]
30. Esmaeili B, Vieluf S, Dworetzky BA, Reinsberger C. The potential of wearable devices and mobile health applications in the evaluation and treatment of epilepsy. Neurologic Clinics 2022;40(4):729-739. [doi: 10.1016/j.ncl.2022.03.005] [Medline: 36270687]
31. Latif S, Rana R, Qadir J, Ali A, Imran MA, Younis MS. Mobile health in the developing world: review of literature and lessons from a case study. IEEE Access 2017;5:11540-11556. [doi: 10.1109/access.2017.2710800]
32. Cuschieri S. The STROBE guidelines. Saudi Journal of Anesthesia 2019;13(Suppl 1):S31-S34 [FREE Full text] [doi: 10.4103/sja.SJA_543_18] [Medline: 30930717]
33. Ahmed A, Saqlain M, Umair MM, Hashmi FK, Saeed H, Amer M, et al. Stigma, social support, illicit drug use, and other predictors of anxiety and depression among HIV/AIDS patients in Pakistan: a cross-sectional study. Frontiers in Public Health 2021;9:745545 [FREE Full text] [doi: 10.3389/fpubh.2021.745545] [Medline: 34660521]
34. Ahmed A, Dujaili JA, Jabeen M, Umair MM, Chuah L, Hashmi FK, et al. Barriers and enablers for adherence to antiretroviral therapy among people living with HIV/AIDS in the Era of COVID-19: a qualitative study from Pakistan. Frontiers in Pharmacology 2021;12:807446 [FREE Full text] [doi: 10.3389/fphar.2021.807446] [Medline: 35153763]
35. Liu H, Li VL, Liu Q, Liu Y, Su C, Wong H, et al. Lac-Phe induces hypophagia by inhibiting AgRP neurons in mice. Nature Metabolism 2025 Oct;7(10):2004-2017. [doi: 10.1038/s42255-025-01377-9] [Medline: 40957996]
36. Houta S, Bader A, Effert J, Esser B, Henze J, Spaic A, et al. Digital health applications in the self-management of epilepsy-A survey on patients' perspective. Epilepsia Open 2023;8(4):1288-1299 [FREE Full text] [doi: 10.1002/epi4.12788] [Medline: 37438936]
37. Belachew EA, Getachew D, Netere AK, Gizachew E, Sendekie AK. Perception, willingness, and practices of telemedicine in patients with chronic diseases: implication of digital health in patients' perspective at a tertiary care hospital in Ethiopia. Frontiers in Public Health 2023;11:1234436 [FREE Full text] [doi: 10.3389/fpubh.2023.1234436] [Medline: 37608985]
38. Hofstede J, de Bie J, van Wijngaarden B, Heijmans M. Knowledge, use and attitude toward eHealth among patients with chronic lung diseases. International Journal of Medical Informatics 2014;83(12):967-974. [doi: 10.1016/j.ijmedinf.2014.08.011] [Medline: 25269992]
39. Tilahun M, Habte N, Mekonnen K, Srahbzu M, Ayelegne D. Nonadherence to antiepileptic medications and its determinants among epileptic patients at the university of Gondar Referral Hospital, Gondar, ethiopia, 2019: an institutional-based cross-sectional study. Neurology Research International 2020;2020:8886828 [FREE Full text] [doi: 10.1155/2020/8886828] [Medline: 33194230]
40. Patel S, Akhtar A, Malins S, Wright N, Rowley E, Young E, et al. The acceptability and usability of digital health interventions for adults with depression, anxiety, and somatoform disorders: qualitative systematic review and meta-synthesis. Journal of Medical Internet Research 2020;22(7):e16228 [FREE Full text] [doi: 10.2196/16228] [Medline: 32628116]
41. Simblett SK, Bruno E, Siddi S, Matcham F, Giuliano L, López JH, RADAR-CNS Consortium. Patient perspectives on the acceptability of mHealth technology for remote measurement and management of epilepsy: a qualitative analysis. Epilepsy Behavior 2019;97:123-129. [doi: 10.1016/j.yebeh.2019.05.035] [Medline: 31247523]
42. Process of translation and adaptation of instruments. Organization WH.: World Health Organization, Geneva; 2009. URL: http://www who int/substance_abuse/research_tools/translation/en/ [accessed 2025-11-07]

XSL•FO
RenderX

43. Ashcroft R. The declaration of Helsinki. In: The Oxford textbook of clinical research ethics. England: Oxford Academic; 2008:141-148.
44. Jembai JVJ, Wong YLC, Bakhtiar NAMA, Lazim SNM, Ling HS, Kuan PX, et al. Mobile health applications: awareness, attitudes, and practices among medical students in Malaysia. BMC Medical Education 2022;22(1):544 [FREE Full text] [doi: 10.1186/s12909-022-03603-4] [Medline: 35836223]
45. Slater K, Taylor R, Collins CE, Hutchesson M. Awareness of cardiovascular disease risk and care received among Australian women with a history of hypertensive disorders of pregnancy: a cross-sectional survey. BMC Pregnancy Childbirth 2025;25(1):15 [FREE Full text] [doi: 10.1186/s12884-024-07018-5] [Medline: 39780057]
46. Giebel GD, Speckemeier C, Abels C, Plescher F, Börchers K, Wasem J, et al. Problems and barriers related to the use of digital health applications: scoping review. Journal of Medical Internet Research 2023;25:e43808 [FREE Full text] [doi: 10.2196/43808] [Medline: 37171838]
47. Patel S, Sreelal B, Kalyani S, Joy S, Pravalika T, Ramya T, et al. Digital health literacy among the ageing population: a comparative cross-sectional study between rural and urban Telangana, India. Medical Journal Armed Forces India 2024;80(3):313-319. [doi: 10.1016/j.mjafi.2023.09.006] [Medline: 38799990]
48. Cao L, Chongsuvivatwong V, McNeil EB. The sociodemographic digital divide in mobile health app use among clients at outpatient departments in inner Mongolia, China: cross-sectional survey study. JMIR Human Factors 2022;9(2):e36962 [FREE Full text] [doi: 10.2196/36962] [Medline: 35587367]
49. Osei E, Mashamba-Thompson TP. Mobile health applications for disease screening and treatment support in low-and middle-income countries: a narrative review. Heliyon 2021;7(3):e06639 [FREE Full text] [doi: 10.1016/j.heliyon.2021.e06639] [Medline: 33869857]
50. Mbunge E, Batani J, Gaobotse G, Muchemwa B. Virtual healthcare services and digital health technologies deployed during coronavirus disease 2019 (COVID-19) pandemic in South Africa: a systematic review. Global Health Journal 2022;6(2):102-113 [FREE Full text] [doi: 10.1016/j.glohj.2022.03.001] [Medline: 35282399]
51. Zöllner JP, Noda AH, McCoy J, Schulz J, Tsalouchidou P, Langenbruch L, et al. Use of health-related apps and telehealth in adults with epilepsy in germany: a multicenter cohort study. Telemedicine and e-Health 2023;29(4):540-550. [doi: 10.1089/tmj.2022.0238] [Medline: 35984859]
52. Patel UK, Malik P, DeMasi M, Lunagariya A, Jani VB. Multidisciplinary approach and outcomes of tele-neurology: a review. Cureus 2019;11(4):e4410 [FREE Full text] [doi: 10.7759/cureus.4410] [Medline: 31205830]
53. Aminifar A, Shokri M, Aminifar A. Privacy-preserving edge federated learning for intelligent mobile-health systems. Future Generation Computer Systems 2024;161:625-637 [FREE Full text] [doi: 10.1016/j.future.2024.07.035]
54. Shegog R, Braverman L, Hixson J. Digital and technological opportunities in epilepsy: toward a digital ecosystem for enhanced epilepsy management. Epilepsy Behavior 2020;102:106663. [doi: 10.1016/j.yebeh.2019.106663] [Medline: 31778878]
55. Madanian S, Nakarada-Kordic I, Reay S, Chetty T. Patients' perspectives on digital health tools. PEC Innovation 2023;2:100171 [FREE Full text] [doi: 10.1016/j.pecinn.2023.100171] [Medline: 37384154]
56. Yoo S, Lim K, Baek H, Jang S, Hwang G, Kim H, et al. Developing a mobile epilepsy management application integrated with an electronic health record for effective seizure management. International Journal of Medical Informatics 2020;134:104051. [doi: 10.1016/j.ijmedinf.2019.104051] [Medline: 31837500]
57. Ernawati I, Yasin NM, Setyopranoto I, Ikawati Z. Effect of mobile health applications on improving self-management knowledge and seizure control in epilepsy patients: a scoping review. Healthcare Informatics Research 2024;30(2):127-139 [FREE Full text] [doi: 10.4258/hir.2024.30.2.127] [Medline: 38755103]
58. Verma A, Pathak P, Mishra A, Upadhya S. Factors linked with perceived stigma amid people with Epilepsy -across sectional study. Epilepsy Research 2024;205:107428. [doi: 10.1016/j.eplepsyres.2024.107428] [Medline: 39116512]
59. Batada I. Impact of Digital Marketing on Consumer Buying Perspectives in the Pakistan Market. J Entrep Manag Innov 2023 Jun 30;5(3):319-345. [doi: 10.52633/jemi.v5i3.305]
60. Patel RJS, Ding J, Marvel FA, Shan R, Plante TB, Blaha MJ, et al. Associations of demographic, socioeconomic, and cognitive characteristics with mobile health access: MESA (Multi-Ethnic Study of Atherosclerosis). Journal of the American Heart Association 2022;11(17):e024885 [FREE Full text] [doi: 10.1161/JAHA.121.024885] [Medline: 36056720]
61. Ridho A, Alfian SD, van Boven JFM, Levita J, Yalcin EA, Le L, et al. Digital health technologies to improve medication adherence and treatment outcomes in patients with tuberculosis: systematic review of randomized controlled trials. Journal of Medical Internet Research 2022;24(2):e33062 [FREE Full text] [doi: 10.2196/33062] [Medline: 35195534]
62. Quisel T, Foschini L, Zbikowski SM, Juusola JL. The association between medication adherence for chronic conditions and digital health activity tracking: retrospective analysis. Journal of Medical Internet Research 2019;21(3):e11486 [FREE Full text] [doi: 10.2196/11486] [Medline: 30892271]
63. Conway CM, Kelechi TJ. Digital health for medication adherence in adult diabetes or hypertension: an integrative review. JMIR Diabetes 2017;2(2):e20 [FREE Full text] [doi: 10.2196/diabetes.8030] [Medline: 30291093]
64. Du S, Zhao Y. Enhancing digital health interventions for medication adherence: considerations for broader applicability and long-term impact. Journal of Medical Internet Research 2025;27:e69204 [FREE Full text] [doi: 10.2196/69204] [Medline: 40085837]

XSL•FO
RenderX

65. Torous J, Andersson G, Bertagnoli A, Christensen H, Cuijpers P, Firth J, et al. Towards a consensus around standards for smartphone apps and digital mental health. World Psychiatry 2019;18(1):97-98 [FREE Full text] [doi: 10.1002/wps.20592] [Medline: 30600619]
66. Huckvale K, Nicholas J, Torous J, Larsen ME. Smartphone apps for the treatment of mental health conditions: status and considerations. Current Opinion in Psychology 2020;36:65-70 [FREE Full text] [doi: 10.1016/j.copsyc.2020.04.008] [Medline: 32553848]
67. Ezeala-Adikaibe BA, Ekochin CF, Ezeme MS, Mbadiwe CN, Okwara CC, Nwobodo MU. Knowledge, attitude, and practice of epilepsy in a rural community in Enugu, Southeast Nigeria. International Journal of Medicine & Health Development 2024;29(2):77-87. [doi: 10.4103/ijmh.ijmh_67_23]
68. Makhado L, Maphula A, Ngomba R, Musekwa O, Makhado T, Nemathaga M, et al. Epilepsy in rural South Africa: patient experiences and healthcare challenges. Epilepsia Open 2024;9(4):1565-1574 [FREE Full text] [doi: 10.1002/epi4.12999] [Medline: 38884148]
69. Phimha S, Sirikarn P, Tiamkao S. Economic burden and determinants among hospitalized patients with epilepsy in Thailand. BMC Health Service Research 2025;25(1):413 [FREE Full text] [doi: 10.1186/s12913-025-12504-3] [Medline: 40114158]
70. Jakob R, Harperink S, Rudolf AM, Fleisch E, Haug S, Mair JL, et al. Factors influencing adherence to mhealth apps for prevention or management of noncommunicable diseases: systematic review. Journal of Medical Internet Research 2022;24(5):e35371 [FREE Full text] [doi: 10.2196/35371] [Medline: 35612886]
71. Polus M, Keikhosrokiani P, Uusimaa J, Komulainen-Ebrahim J, Annunen J, Khan S. The role of digital care pathway for epilepsy on patients' treatment burden: clinicians' perspective. 2024 Presented at: First Nordic Conference, NCDHWS 2024; 2024 May 7–8; Oulu, Finland p. 267-268. [doi: 10.1007/978-3-031-59080-1_19]
72. Patterson V, Larik AB, Soomro MZ, Coates S. Validation of an epilepsy management smartphone application in Pakistan. Acta Epileptol 2022;4:12 [FREE Full text] [doi: 10.1186/s42494-021-00075-9] [Medline: 40477854]
73. Alzamanan MZ, Lim KS, Ismail MA, Ghani NA. Development of an epilepsy self-management mobile health app framework: content validity study results. PLoS One 2024;19(6):e0302844 [FREE Full text] [doi: 10.1371/journal.pone.0302844] [Medline: 38848353]

Abbreviations

HIPAA: Health Insurance Portability and Accountability Act
LMIC: low- and middle-income country
mHealth: mobile health
OPD: Outpatient department
PIMS: Pakistan Institute of Medical Sciences
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by A Stone, A Mavragani; submitted 07.07.25; peer-reviewed by MD Butt, C Ochoa-Zezzatti; comments to author 08.09.25; revised version received 19.10.25; accepted 23.10.25; published 10.12.25.

Please cite as:
Zahra BB, Hamza MA, Sarwat R, Zahra HB, Azam A, Ahmed A
Awareness, Perceptions, Willingness, and Feasibility of mHealth Apps Among People Living With Epilepsy: Cross-Sectional Questionnaire Study
JMIR Form Res 2025;9:e80283
URL: https://formative.jmir.org/2025/1/e80283
doi:10.2196/80283
PMID:

©Bushra Batool Zahra, Muhammad Amir Hamza, Rehana Sarwat, Hina Batool Zahra, Ayesha Azam, Ali Ahmed. Originally published in JMIR Formative Research (https://formative.jmir.org), 10.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

AI-Enabled Personalized Smoking Cessation Intervention With the Aipaca Chatbot: Mixed Methods Feasibility Study

Yunlong Liu[1], MS; Paul Calle[1], PhD; Mariah Vadakekut[1], BS; Daniel Rubin[2], MD; Zsolt Nagykaldi[3], PhD; Mark Doescher[3], MD, MScPH; Lisa Hightow-Weidman[4], MD, MPH; Chongle Pan[1*], PhD; Ruosi Shao[5*], PhD

[1]School of Computer Science, University of Oklahoma, Norman, OK, United States

[2]Department of Community Health and Family Medicine, College of Medicine, University of Florida, Gainesville, FL, United States

[3]Department of Family and Preventive Medicine, College of Medicine, University of Oklahoma Health Sciences Center, Oklahoma City, OK, United States

[4]College of Nursing, Florida State University, Tallahassee, FL, United States

[5]College of Communication and Information, Florida State University, Tallahassee, FL, United States

[*]these authors contributed equally

Corresponding Author:
Ruosi Shao, PhD
College of Communication and Information
Florida State University
296 Champions Way, Building C
Tallahassee, FL, 32306
United States
Phone: 1 850 644 5034
Email: rshao@fsu.edu

Abstract

Background: Tobacco use remains the leading cause of preventable mortality in the United States; yet, evidence-based cessation services remain underused due to staffing constraints, limited access to counseling, and competing clinical priorities. Generative artificial intelligence (GenAI) chatbots may address these barriers by delivering personalized, guideline-aligned counseling through naturalistic dialogue. However, little is known about how GenAI chatbots support smoking cessation at both outcome and communication process levels.

Objective: This feasibility study evaluated the implementation of an evidence-based smoking cessation counseling session delivered by a GenAI-powered chatbot, Aipaca. We examined (1) pre-post changes in cessation preparedness, (2) communication dynamics during counseling sessions, and (3) user perceptions of the chatbot's value, limitations, and design needs.

Methods: We conducted an observational, single-arm, mixed methods study with 29 adult smokers. Participants completed pre-post surveys measuring knowledge of smoking-related health risks and cessation methods, self-efficacy, and readiness to quit. Each engaged in a 30-minute text-based counseling session with Aipaca, powered by GPT-4 and structured using the 5A's framework (Ask, Advise, Assess, Assist, Arrange). Sessions were transcribed for microsequential conversation analysis. Twenty-five participants completed semistructured interviews exploring perceived value, challenges, and design suggestions. Quantitative data were analyzed with paired-samples *t* tests, qualitative data were thematically analyzed, and transcripts were analyzed for interactional practices. The methodological strength of this study lies in its triangulated approach, which combines quantitative measurement of intervention effectiveness, qualitative analysis of user interviews, and conversational analysis of counseling transcripts to generate a comprehensive understanding of both outcomes and underlying mechanisms.

Results: Participants demonstrated significant improvements in all preparedness indicators: knowledge of health risks, knowledge of cessation methods, self-efficacy, and readiness to quit. Conversation analysis identified three recurrent patterns enabling counseling-relevant dynamics: (1) contextual referencing and continuity, (2) formulations with elaboration prompts, and (3) narrative progression toward collaborative planning. Interview themes underscored Aipaca's perceived value as an accessible, nonjudgmental, and motivating resource, capable of delivering personalized and interactive support. Criticisms included limited accountability, reduced cultural resonance, and overly goal-directed style. Participants emphasized design needs such as proactive engagement, gamified progress tracking, empathetic or anthropomorphic personas, and safeguards for accuracy.

XSL•FO
RenderX

Conclusions: This mixed methods feasibility study demonstrates that GenAI can deliver evidence-based smoking cessation counseling with measurable short-term gains in cessation preparedness and process-level communication patterns consistent with motivational interviewing. Users valued Aipaca's accessibility, empathy, and personalization, while also articulating expectations for richer social roles and long-term accountability. Findings highlight both the promise and challenges of integrating GenAI into digital health: pairing adaptive language generation with human-centered design, embedding accuracy safeguards, and ensuring integration into multilevel cessation infrastructures will be essential for future clinical deployment.

(JMIR Form Res 2025;9:e73319) doi:10.2196/73319

KEYWORDS

generative AI; conversational agent; usability; tobacco treatment; health intervention

Introduction

Background

Tobacco use remains the leading cause of preventable mortality in the United States, accounting for more than 480,000 deaths each year, nearly 1 in 5 of all US deaths [1]. Smoking drives premature death through its role in diseases such as lung cancer, chronic obstructive pulmonary disease, and coronary artery disease [2]. Implementing evidence-based tobacco treatment interventions is therefore a continuing public health imperative [3]. While these in-person interventions are effective in both the short and long term, their widespread implementation is often hindered by high demands on staff and clinical resources, limited access to in-person counseling, and competing priorities for clinicians' time [4-6]. These persistent barriers highlight the need for scalable, innovative approaches to complement existing cessation services.

Recent advancements in generative artificial intelligence (GenAI) have created new opportunities for implementing evidence-based health interventions through conversational agents [7-10]. GenAI-powered chatbots are uniquely positioned to deliver highly personalized, interactive, and natural language-based counseling, adapting in real time to users' evolving needs and preferences [11]. This potential for dynamic, human-like dialogue has not only expanded the technical capabilities of chatbots but also raised users' expectations regarding naturalness, engagement, and the fulfillment of complex social roles in digital health.

Despite these advances, the mechanisms by which GenAI-driven chatbots influence health behavior change, particularly through communication processes and user perceptions, remain poorly understood. Most published studies have focused either on quantitative outcomes or on qualitative user interviews but seldom combine these with in-depth analysis of the counseling conversations themselves. Methodological triangulation, combining quantitative measures of intervention outcomes, qualitative interviews of user perceptions, and conversational analysis of the counseling process and communication dynamics between participants and artificial intelligence (AI) chatbots, strengthens research by enabling a more comprehensive understanding of complex phenomena [12]. In the context of digital health, triangulation that encompasses quantitative effectiveness, qualitative interviews, and conversational analysis of transcripts is particularly valuable for elucidating both what works and how it works.

Related Work

The Rise of Conversational Agents and GenAI in Substance Use Intervention

Conversational agents (CAs), such as chatbots and avatar-led systems, have increasingly been adopted to deliver tobacco treatment and broader substance use interventions. A systematic review of 13 studies (N=8236) found that CAs significantly improved cessation outcomes (sample-weighted odds ratio 1.66, 95% CI 1.33-2.07) and were generally well-accepted by users, commonly assessed via self-reported satisfaction, perceived ease of use, comfort interacting with the agent, and willingness to recommend [13]. Chatbot-delivered interventions vary from 4 to 24 weeks, with interaction frequencies ranging from daily to multiple times per week, and median engagement levels around 36 of 42 intervention days. Engagement was high, with over 85% of participants initiating at least one interaction and about 75% reading all messages [13]. Similarly, a meta-analysis of 5 randomized controlled trials (RCTs; N=58,796) revealed that conversational AI interventions significantly increased 6-month abstinence rates compared to control conditions (risk ratio [RR] 1.29, 95% CI 1.13-1.46) [14]. These trials examined CA in various delivery formats, such as chatbots embedded in smoking cessation apps with daily interactive counseling, social media–based CAs embedded in group chats, and internet-based avatars delivering acceptance and commitment therapy (ACT) [15-17]. Control groups across these RCTs included usual care with in-person counseling and pharmacotherapy, smoking-cessation apps without chatbot functionality, or static smoking cessation tips with no interactive support.

Evidence from single-arm and pilot trials extends these findings to specific populations, including young adults and sexual and gender minority users. For example, avatar-led ACT interventions improved self-efficacy and quit intentions among university students through six 25-minute digital sessions, with 51.9% of completers achieving abstinence at postintervention [18]. Similarly, a digital ACT-based program tailored for sexual and gender minority young adults yielded 34% biochemically confirmed abstinence at postintervention, approximately 3 times higher than the 9%-12% rates observed in previous digital interventions [19]. Beyond cessation counseling, CAs have also been applied to screening and intake. Jeanne, a 3D embodied CA designed to assess tobacco and alcohol use disorders, can conduct a 4.4-minute structured interview using validated measures (eg, Cigarette Dependence Scale-5 for tobacco) [20]. Findings indicated high acceptability (mean score=24.8/30), strong concordance with paper-based tools (r>0.89), and

XSL•FO
RenderX

excellent diagnostic validity (eg, area under the curve=0.97 for tobacco), supporting the feasibility of automating early stages of behavioral health interventions.

Taken together, existing evidence supports the initial feasibility and effectiveness of CAs across the tobacco treatment continuum, including screening, initial counseling, quit planning, motivational interviewing (MI), relapse prevention, and follow-up coaching. However, barriers were identified, especially for earlier rule-based or prescripted systems, including a lack of human-likeness, limited response coherence, and constrained conversational depth, which hindered relational communication, reduced engagement, and compromised long-term intervention effectiveness [13]. Moreover, attrition rates remain substantial in CA-based trials, ranging from 14% to 80% across studies, reflecting heterogeneous study durations and definition of "dropout," and underscoring engagement as a persistent challenge [14]. Importantly, this engagement challenge is broadly comparable to other cessation modalities: for example, 50% of participants dropped out during treatment, and 61% were not reached at 7-month follow-up in a large state Quitline cohort (N=49,347) [21].

Chatbot Design and Integration of Evidence-Based Intervention

Chatbot-delivered smoking cessation interventions increasingly combine structured, evidence-based counseling components with human-centered design principles and, in many cases, generative AI capabilities. Systematic reviews indicate that effective systems typically integrate multiple intervention elements through frequent, tailored, multiweek interactions, encompassing screening and intake, initial counseling and psychoeducation, MI, personalized quit-plan development, skill-building and coping strategies, relapse prevention, and referral to human services [13,14]. For example, Florence, the World Health Organization's web-based virtual human, provides 24/7 access to guideline-aligned brief tobacco treatment, assisting users in setting quit dates, developing quit plans, and connecting to local quitlines and cessation resources [22]. Evidence-based MI and cognitive behavioral therapy (CBT) approaches have also been integrated into chatbot systems. A web-based MI-style chatbot delivers 2 structured MI sessions: an assessment session of smoking history and quitting intentions, followed by a personalized normative feedback session designed to elicit reasons for quitting [23]. Roby, another chatbot-based intervention, provides a 5-session program that combines MI and CBT, guiding users through quit date setting, withdrawal and craving management, and relapse prevention [24]. Despite these advances, substantial variability remains in the reporting of conversational design and fidelity, particularly in how chatbots structure sessions and implement features such as content sequencing, turn-taking, prompt logic, and follow-up cadence [25].

Across development efforts, human-centered design and iterative co-design processes have been central to optimizing usability, therapeutic relevance, and long-term engagement of smoking cessation chatbots. The development of QuitBot followed a 4-year, 11-step user-centered design process that included needs assessments, iterative prototyping, diary studies, and a pilot RCT; user feedback led to refinements in persona, tone, session flow, and eventually the addition of an open-ended Q&A function supported by fine-tuned large language models (LLMs) [25]. Similar iterative design principles guided the World Health Organization's Florence chatbot, which was codeveloped with user testing to ensure accessibility, personalization, and integration with local cessation resources [22]. Systematic reviews reveal that users consistently value design elements that enhance empathy, interactivity, and personalization, while expressing dissatisfaction with rigid, scripted conversations or excessive repetition [13,14]. Evidence from human-centered studies further highlights that users prefer supportive personas, clear quit-planning tools, flexible frequency of interaction, and integration of progress tracking features [25].

The natural language processing (NLP) techniques underpinning smoking cessation chatbots have evolved from prescripted and retrieval-based systems to more advanced generative AI models. Early systems, such as Florence, relied on prescripted content to deliver guideline-based brief interventions with high safety and fidelity, but limited conversational depth and flexibility [22]. Retrieval-based designs introduced greater adaptability by using intent classification and predefined response libraries, as seen in TAMI, the tobacco addiction management intelligent agent [26]. TAMI incorporated a human-in-the-loop approach, in which uncovered intents and user inputs were iteratively analyzed and mapped to new response templates, thereby extending the flexibility and coverage of its conversations while maintaining control and consistency. More recent systems leverage generative models to create dynamic, context-aware counseling exchanges. For example, MIBot used a hybrid design that combined scripted prompts with fine-tuned open-source transformer models to generate MI-style reflections [27]. Subsequent evaluations showed that user satisfaction with MIBot decreased after the release of ChatGPT, reflecting users' rising expectations for conversational empathy and coherence. In addition, users with prior ChatGPT experience produced longer responses and reported greater increases in the perceived importance of quitting smoking, suggesting that exposure to generative AI systems is shaping user expectations of therapeutic chatbots in meaningful ways. Beyond LLMs, reinforcement learning has also been applied to optimize persuasive message strategies by modeling user traits and motivational states [28]. To date, most research-oriented cessation chatbots use open-source fine-tuned models (eg, GPT-2/3 derivatives) or custom NLP pipelines, whereas clinical deployment of closed-source commercial models such as OpenAI's GPT-4/5 remains limited, reflecting both the technical opportunities and ongoing concerns regarding safety, bias, and transparency.

Study Objectives

This feasibility study used a mixed methods design to evaluate the implementation of an evidence-based smoking cessation intervention delivered by a GenAI-powered chatbot. We recruited 29 adult smokers to participate in an initial counseling session with the AI chatbot, Aipaca (AI for Patient Care), during which each participant developed a personalized quit plan and set a quit date in accordance with the U.S. Public Health Service guideline [29]. Following this session, participants completed a structured interview to share their experiences, perceptions,

XSL•FO
RenderX

and suggestions for improvement. We further conducted conversational analysis with all counseling transcripts to examine communication dynamics between users and the AI.

The methodological strength of this study lies in its triangulated approach, which combines quantitative measurement of intervention effectiveness, qualitative analysis of user interviews, and conversational analysis of counseling transcripts to generate a comprehensive understanding of both outcomes and underlying mechanisms [12]. Our study addressed three key questions: (1) How effectively can a GenAI chatbot deliver smoking cessation counseling? (2) What communication dynamics arise during user interactions with the chatbot? (3) How do users perceive the chatbot's role in their cessation efforts?

By integrating these perspectives, our work contributes to the growing literature on AI-augmented health interventions and provides practical guidance for the human-centered design and evaluation of GenAI-powered chatbots in evidence-based behavioral counseling.

Methods

Chatbot Development

System Architecture and Implementation

The text-based chatbot, Aipaca, was developed using the GPT-4 API [30], which enables interactive and personalized responses through the generative pretrained transformer model. The system architecture consists of 4 key components: user interface, messaging system, model service, and storage system (Figure 1). The front-end user interface was designed using HTML and JavaScript, accessible via web browsers and smartphones. It incorporated design enhancements such as auto-scrolling during conversations and a pop-up displaying a survey link upon completion of the counseling session. The messaging system manages communication between users and model services, facilitating user authentication through Google sign-in or a 6-digit user ID, implemented with Python and RESTful APIs. It handles request routing and chat history management, with key endpoints for login, menu retrieval, chat initiation, response generation, survey access, and chat history retrieval. JSON is used for data interchange, ensuring efficient and secure communication across the system. Each model service operates independently, using Python Flask to enable a modular and scalable architecture. This setup allows for seamless integration of additional APIs, advanced models, or increased server capacity to support future enhancements. Independent operation across different machines or ports further enhances system flexibility and scalability. The storage system offers 2 methods for information retrieval: users authenticated through Google sign-in can access their complete chat history, while those using a 6-digit ID can retrieve chat history from the past 2 hours, accommodating the typical duration of an initial counseling session (approximately 30 minutes). This setup is optimized for initial counseling interactions and aligns with the cross-sectional nature of our study design, ensuring secure authentication and efficient data management.

Figure 1. The system architecture of the Alpaca chatbot consists of four key components: user interface (web front end used by participants), messaging system (session management, routing, logging, and attention-check handling), model services (LLM for counseling responses), and storage system (database saving deidentified inputs/outputs and loading chat history). Arrows indicate data flow during the counseling sessions. LLM: large language model.

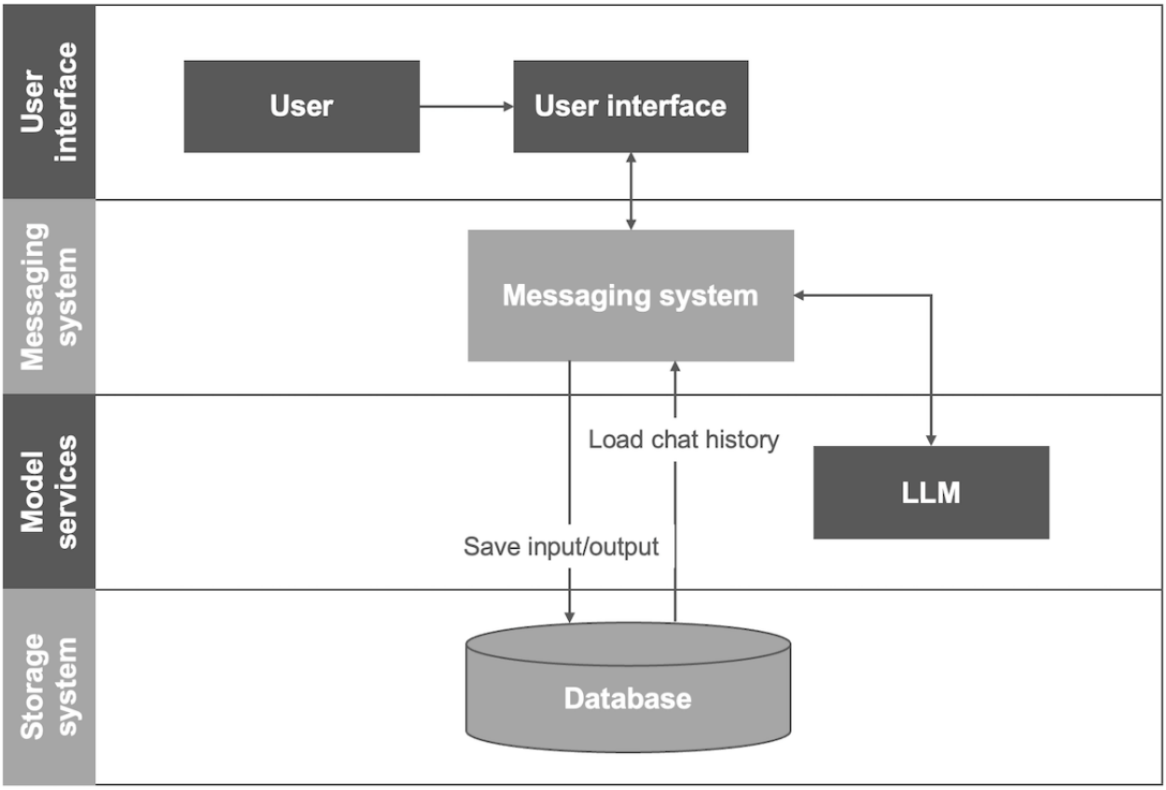

XSL•FO
RenderX

Implementing Evidence-Based Smoking Cessation Counseling Through Prompting

The evidence-based standards of practice developed by the Tobacco Treatment Specialist (TTS) training program were implemented in Aipaca through structured prompting [31]. The prompt outlined Aipaca's role as a TTS and provided a comprehensive framework for conducting initial cessation counseling. This framework emphasized a client-centered, MI approach, encouraging participants to openly discuss their tobacco use, reflect on smoking triggers, identify strategies for behavior change, and develop coping mechanisms to support cessation. The counseling structure adhered to the 5A's framework (Ask, Advise, Assess, Assist, Arrange) as outlined in clinical practice guidelines [4,32], and was operationalized into a five-step process (Figure 2): (1) asking about tobacco use and previous quit attempts, (2) advising on tailored quitting strategies and resources, (3) assessing readiness to quit, (4) assisting in developing a personalized quit plan, and (5) arranging a quit date.

Figure 2. Five-step counseling structure following the 5A framework. Intervention workflow for a guideline-concordant, evidence-based session (approx. 10-15 minutes) delivered to adult US smokers in a single-arm study. Steps include ask (review smoking history, discuss health risks, and assess past quit attempts), advise (manage withdrawal, introduce evidence-based resources, and identify support), assess (evaluate readiness to quit), assist (create a personalized quit plan), and arrange (set a quit date).

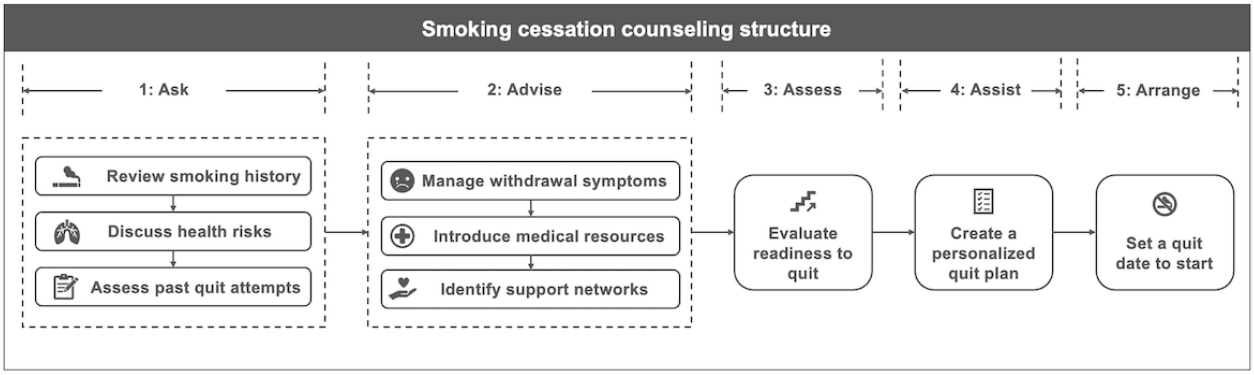

The prompting framework underwent iterative pilot testing and refinement by 3 certified TTS through simulated counseling sessions, incorporating insights from clinical practice and case-based scenarios in tobacco treatment. Consistent with clinical practice guidelines, the initial counseling session was designed to last 10-15 minutes, providing low-to-medium intensity counseling [4]. Prior to study launch, we evaluated Aipaca's performance using a standard TTS certification exam, a requirement for accredited TTS training programs by the Council for Tobacco Treatment Training Programs. Aipaca achieved a score of 81%, exceeding the required passing threshold of 70%.

Study Design

We conducted an observational, single-arm, mixed methods study using methodological triangulation to develop a comprehensive understanding of both outcomes and underlying mechanisms of GenAI-enabled smoking cessation counseling. The study design integrated three complementary components: (1) quantitative pre-post surveys to assess changes in preparedness for cessation, (2) conversation analysis of counseling transcripts to examine communication dynamics during human-chatbot interaction, and (3) qualitative thematic analysis of semistructured interviews to explore user perceptions and identify design implications (Figure 3). This triangulated approach is well suited for evaluating complex communication interventions, as it aligns proximal behavioral outcomes with observed interaction processes and user meaning-making [12].

Figure 3. Mixed methods study design. Participants completed a pre-study survey, a single smoking-cessation counseling session with Aipaca, a post-study survey, and an optional qualitative interview. Quantitative pre-post outcomes (preparedness or readiness, self-efficacy, and knowledge) were triangulated with conversation-analysis of transcripts and thematic analysis of interviews to explain observed changes and user experience. Population: adult US smokers (≥5 cigarettes per day) recruited via MTurk.

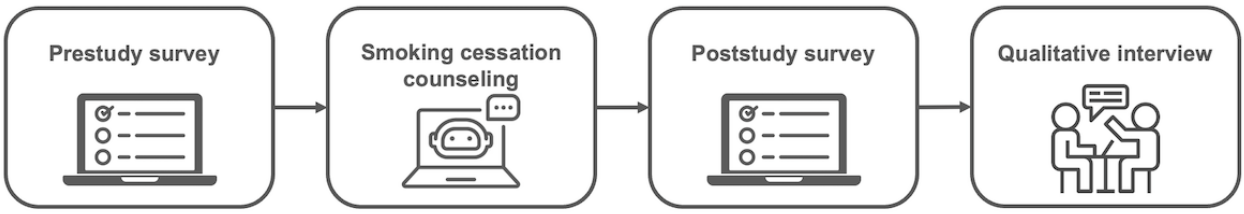

XSL•FO
RenderX

Ethical Considerations

This study was reviewed and approved by the institutional review board (IRB) of the University of Oklahoma (IRB15693). All procedures complied with institutional policies and applicable regulations for research involving human participants. Prior to any study activities, participants reviewed an IRB-approved information sheet and provided electronic informed consent within a secure survey environment. Participation was voluntary, and respondents could withdraw at any time without penalty. To protect privacy, enrollment and follow-up occurred through MTurk (Amazon Mechanical Turk), single participation was enforced via Worker ID with IP/device checks, and all data were stored on access-controlled servers. Monetary compensation was provided through MTurk after verification of task completion. Before participant enrollment, we evaluated Aipaca's performance with a standard TTS certification examination required by the Council for Tobacco Treatment Training Programs–accredited courses; Aipaca scored 81% (70% passing threshold). These preparatory evaluations were used to confirm content alignment with evidence-based counseling standards and to minimize participant risk prior to initiating data collection.

Participants and Recruitment

Participants were recruited via MTurk between May and June 2024. Eligibility criteria included (1) currently smoking ≥5 cigarettes per day, (2) at least 18 years of age, (3) residing in the United States, and (4) having a ≥97% approval rate on human intelligence tasks on MTurk. Twenty-nine participants (16 males, 13 females; mean age 45.5, SD 12.5, range 25-65 years) completed the pre-post surveys and counseling session, of whom 25 completed the semistructured interview. Descriptive smoking history and nicotine dependence characteristics are summarized in Table S1 in Multimedia Appendix 1, with Fagerström Test for Nicotine Dependence (FTND) categories reported to contextualize likely withdrawal severity and inform the therapeutic tailoring recommended in clinical practice.

Recruitment used systematic quality and uniqueness safeguards. We limited enrollment to one submission per Worker ID, implemented IP/device duplication checks, embedded attention-check items in both screening and baseline instruments, and prevented back-button resubmissions. Geo-IP verification confirmed US access. MTurk completion codes were generated only upon successful postsurvey submission and were validated against Worker IDs prior to payment to prevent duplicate participation. Survey data were transmitted over Transport Layer Security and stored on a secure server; the linkage file between Worker ID and survey ID was maintained separately and deleted after payments were issued.

Procedures

Pre-Post Survey Measures

After consent, participants completed a preintervention survey on Qualtrics measuring demographics, tobacco use (years of smoking, cigarettes per day, prior quit attempts, and longest abstinence), nicotine dependence (FTND) [33], and cessation preparedness. Preparedness included (a) perceived sufficiency of knowledge about smoking-related health risks (0-100) and cessation methods (0-100) [34,35], (b) cessation self-efficacy (1-7) [36], and (c) readiness to quit (10-point readiness ladder) [37]. The same preparedness measures were collected again following the counseling session.

Counseling Session With Aipaca

Participants then engaged in a text-based initial counseling session with Aipaca, a GenAI chatbot designed to deliver evidence-informed tobacco treatment (eg, brief advice, assessment, and collaborative planning) and MI microskills (open questions, reflections, affirmations, and summaries). The chatbot was implemented via GPT-4 guided by a system-level prompt specifying counseling goals, turn-taking structure, safety guardrails (eg, escalation for medical danger and avoidance of diagnostic claims), and content boundaries. A programmatic scaffold ensured coverage of core counseling elements (history, motivations, barriers/triggers, quit-aid education, and collaborative plan) while allowing free-text responses.

Semistructured Interviews

Following the postsession survey, participants were invited to a 30-minute semistructured interview exploring (1) perceived value and critiques of the chatbot session, (2) opportunities and challenges for real-world implementation, and (3) design suggestions regarding interaction style, features, and safeguards. Twenty-five interviews were completed, as four participants (participants 6, 10, 25, and 27) opted out after the postintervention survey due to time constraints.

Reporting Standards

We reported this mixed methods feasibility study in accordance with the GRAMMS (Good Reporting of a Mixed Methods Study) guideline [38-40].

Data Sources and Analysis

Quantitative Analysis

We used paired-samples *t* tests to assess pre-post change in each preparedness measure. For interpretability, we report mean differences with 95% CIs and standardized within-person effect sizes (Cohen d_z). Analyses were conducted with 29 participants.

Conversation Analysis of Counseling Transcripts

We analyzed 29 counseling transcripts (1514 total turns; session range=50-112 turns) using a microsequential approach to characterize how counseling interactions unfolded turn by turn (eg, adjacency pairs, question-answer sequences, formulations, and contextual referencing of prior talk) and how these sequences supported personalization and collaborative plan formation. Two authors (YL and RS) met weekly to iteratively develop and refine an interactional codebook that included categories such as topic initiation and uptake, referencing earlier content, agenda-setting, formulation and summarization, and plan negotiation. Discrepancies were resolved through discussion until consensus was achieved.

Thematic Analysis of Interviews

Interview data were transcribed and analyzed using reflexive thematic analysis focused on three a priori domains: (1) perceived values and criticisms, (2) opportunities and challenges

XSL•FO
RenderX

of GenAI chatbot for counseling support, and (3) design feedback. Initial codes were generated by the corresponding author (RS) following established procedures [41], and two additional authors (YL and PC) independently reviewed a subset to refine code definitions and theme boundaries. Disagreements were resolved by consensus. The final thematic structure included 3 overarching themes and 11 subthemes (Table S2 in Multimedia Appendix 1), with exemplar participant quotes selected to illustrate each theme.

Results

Pre-Post Changes in Preparedness for Cessation

Participants demonstrated statistically significant improvements across all 4 cessation preparedness indicators (Figure 4). On the 0-100 scales, perceived knowledge about smoking-related health risks increased from 82.59 (SD 13.35) to 86.21 (SD 8.69), t_{28}=−2.12, P=.043, 95% CI 0.13-7.11, and d_z=0.39. The t tests conducted in this study were 2-tailed. Knowledge of cessation methods increased from 56.86 (SD 24.02) to 76.86 (SD 18.50), t_{28}=−6.77, P<.001, 95% CI 13.95-26.05, and d_z=1.26. On the 1-7 scale, cessation self-efficacy improved from 3.10 (SD 1.32) to 4.38 (SD 1.52), t_{28}=−5.05, P<.001, 95% CI 0.76-1.79, and d_z=0.94. On the 10-point readiness ladder, readiness to quit increased from 5.76 (SD 2.80) to 6.83 (SD 2.52), t_{28}=−4.08, P<.001, 95% CI 0.53-1.61, and d_z=0.76. Of the 29 participants, 19 (65.5%) moved up at least one rung on the readiness ladder, more often shifting from contemplation (eg, “I should quit, but I’m not quite ready”) to preparation (eg, “I am starting to think about how to change my smoking patterns”) or initial action (eg, “I am taking actions to quit”).

Figure 4. Pre-post changes in preparedness for cessation. The left panel displays four boxplots comparing pre- and postintervention scores for knowledge of health risks (0-100), knowledge of cessation methods (0-100), self-efficacy (1-7), and readiness to quit (0-10). The right panel illustrates the readiness to quit ladder, with shaded circles representing preintervention data and blue circles representing postintervention data. Within-person pre-post comparisons were conducted (2-sided α=.05).

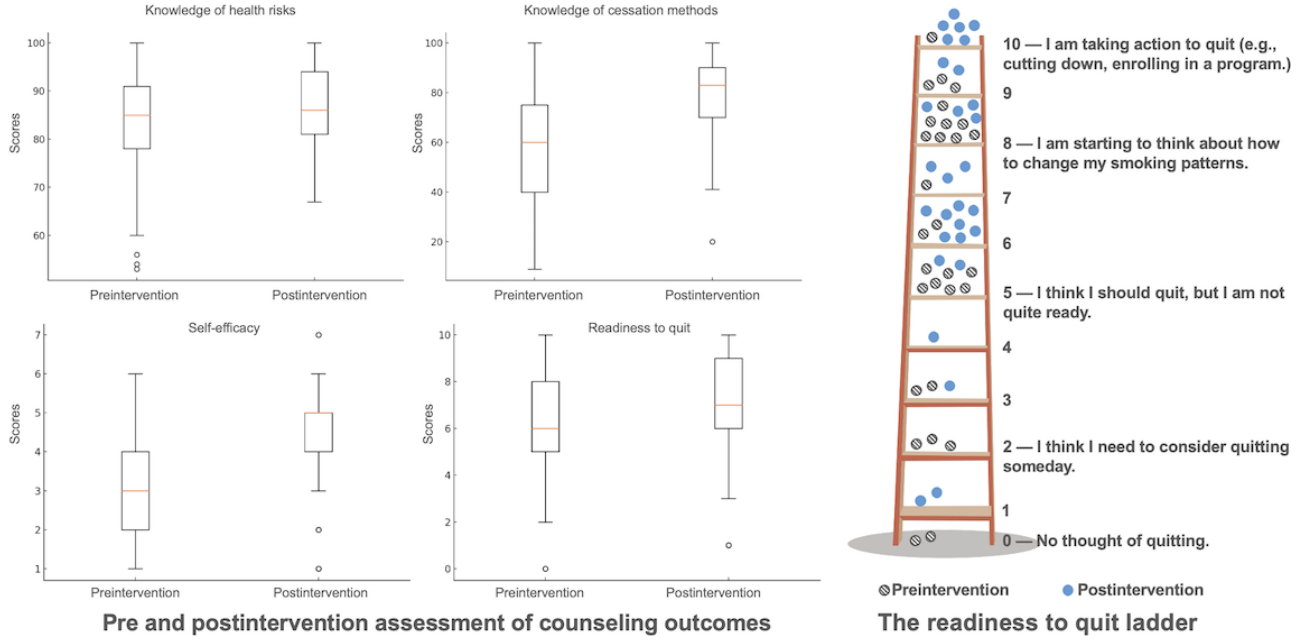

Communication Dynamics During Counseling Sessions

The microsequential analysis of 29 transcripts offered insights into the interactional dynamics of human-chatbot counseling sessions, illuminating not only *what* transpired but also *how* the counseling work was co-constructed turn by turn [42]. Three recurrent patterns were identified: contextual referencing and continuity, formulations with elaboration prompts, and narrative progression toward collaborative planning.

Contextual Referencing and Continuity

Both participants and the chatbot regularly referenced prior talk (eg, earlier-stated reasons for quitting, identified triggers, or previous quit attempts) to maintain topical continuity and conversational coherence. For instance, when Participant 13 revisited a prior cardiac hospitalization as a motivation for quitting, the chatbot reformulated and extended this into a concrete health benefit:

Participant 13: Like I said before, the main reason I need to quit is my heart. I really want to stay healthy and live a long life.

Aipaca: That is a compelling reason to quit. You know, quitting can really make a huge difference for your heart. It cuts the risk of heart disease and can help your body recover over time. You might even notice better stamina as you get further along in the process.

Formulations, Clarifications, and Elaboration Prompts

The chatbot frequently restated participants’ prior input and followed with clarifying or elaborative questions (eg, “You mentioned stress as an important trigger, can you tell me more about which situations set it off?”), encouraging deeper disclosure around triggers and coping contexts [43]. For example, Participant 23 mentioned stress-related cravings, and the chatbot prompted further elaboration:

XSL•FO
RenderX

> *Participant 23: I tend to get cravings when I'm stressed.*
>
> *Aipaca: I see, stress is a big one for a lot of people. Can you tell me a bit more about what kinds of stress tend to trigger those cravings? Is it work, family, or maybe certain situations that really set it off?*

Narrative Progression and Collaborative Planning

Sessions often followed a trajectory from orientation and history-taking into reflective exploration of motivations and barriers, culminating in collaborative plan building (eg, selecting a quit date, discussing quit aids, and drafting coping strategies). Early turns frequently used inclusive pronouns (eg, "we" and "let's") to facilitate affiliation and joint problem-solving, as in "Let's dig into it" and "Now we know better about what triggers you to smoke" [43]. As therapeutic rapport developed, the chatbot transitioned to more in-depth questions about previous quit attempts, strategies used, and the struggles encountered. The progression encouraged participants to move from brief factual responses to more personal disclosure of feelings, motivations, and barriers to quitting. For example, Participant 21 initially gave brief responses about smoking history (eg, "35 years") but later shared detailed reflections on difficulties with cravings, stress, and sleep disruption during a past quit attempt:

> *The longest I managed to stay smoke-free was about a month. I was really trying, like sticking to it every day, but it was hard. The cravings were really intense, especially when I was stressed. I was feeling pretty irritable and had trouble sleeping too. But eventually, when things got tough, like work stress and some family stuff, I just caved and went back to smoking. It felt like the only way to deal with everything at the time.*

This progressive nature of the counseling discourse allowed the chatbot to adapt to participants' needs and preferences, fostering more personalized and supportive interactions. In later sequences, Aipaca pivoted toward readiness assessment and co-construction of tailored quit plans by linking motivations to actionable strategies:

> *Aipaca: You've mentioned family as a big reason for quitting. How do you think quitting will impact your relationship with them?*
>
> *Participant 3: I think it'll be good. I don't want my kids to see me smoke, and I don't want to be out of breath when I'm playing with them.*
>
> *Aipaca: That sounds like a strong motivator. How do you feel about involving your family in your quit plan? Maybe they could support you when cravings hit or during stressful moments.*

Taken together, these sequential practices, including contextual referencing, formulation with elaboration, and a progressive trajectory from orientation to collaborative planning, facilitated contingent, personalized counseling within sessions. The patterns suggest that the chatbot's turn design consistently leveraged participants' prior contributions to sustain coherence and scaffold the movement from information gathering toward actionable next steps aligned with their motivations and barriers.

User Perceptions of Smoking Cessation Counseling With GenAI Chatbot

Most participants (n=18) indicated an intention to continue using Aipaca as a resource to support smoking cessation. Commonly cited strengths included accessibility, nonjudgmental nature, perceived empathy, and personalized responses. Several participants described Aipaca as a useful complement rather than a stand-alone solution:

> *A good starting point...but I would use it in conjunction with sessions with a specialist.* [Participant 9]
>
> *I would trust a doctor's advice over a chatbot. But if I wanted some quick information, I would use it.* [Participant 15]

In contrast, 5 participants expressed reluctance to engage further with Aipaca, citing an impersonal tone, lack of accountability, or a communication style they felt was overly goal-directed. Notably, 8 participants reported no dislikes, underscoring the system's overall acceptability.

Perceived Value of Aipaca in Quit-Smoking

Reliable and Accessible Support Resource

Participants recognized Aipaca as a valuable source of accessible, always available, and high-quality cessation support. They emphasized the clarity, relevance, and simplicity of the information provided, describing responses as "very informative and detailed" [Participant 2], "fast and accurate" [Participant 24], "straightforward and clear...relevant to my initial inputs" [Participant 1], and "done so very well, wasn't overly complex and seemed helpful" [Participant 17]. Several participants highlighted its potential as a comprehensive informational resource: "Aipaca gives me all the relevant answers I need without searching" [Participant 24].

Availability in moments of temptation was especially valued. Participants noted the benefit of having real-time support when cravings struck:

> *You can message all day when you have a craving, and it provides you motivation not to smoke.* [Participant 12]

Others emphasized its reliability as a stand-in for social support when traditional resources were unavailable:

> *Just having something there for support would be a great guide to quitting. A lot of times there may not be anyone around to talk to. But with Aipaca, you wouldn't have to worry about that.* [Participant 29]

For participants with limited support networks, Aipaca was viewed as a valuable alternative: "valuable in cases where a person does not have a social support system in any other way" [Participant 3]. Some even preferred it to traditional supports, citing its less personal nature as advantageous:

> *As an introvert, I personally prefer the chatbot since it's less personal than talking to a doctor.* [Participant 26]

XSL•FO
RenderX

Nonjudgmental and Motivating Interaction

Participants valued Aipaca's humanlike and supportive qualities, describing the interaction as personal, encouraging, and motivating. As Participant 1 stated:

> *I would use Aipaca for the personality of the responses that it gave...as it resonated easily with me and would influence me to continue to use it in my quitting phase.*

Participants frequently described the chatbot as "encouraging and supportive" [Participant 20], with several noting that it boosted their confidence and self-efficacy: "something that can constantly boost your mentality and help you feel important enough to quit" [Participant 12].

Several participants reported experiencing a sense of sincerity and care during interactions.

> *I felt heard and encouraged. I felt I could ask anything. The chatbot really had my best interest in mind.* [Participant 23]

Others echoed this:

> *It felt like I was talking with a real person, one who was concerned about me and wanted to help.* [Participant 28]

Importantly, Aipaca also created a safe, nonjudgmental space for discussing struggles. As Participant 5 noted:

> *I feel more comfortable chatting to it than talking to someone in my family who has never smoked and doesn't understand how addicting it is.*

Contingent, Personalized, and Interactive Communication

Participants perceived Aipaca as capable of understanding their inputs and generating relevant, contextually appropriate responses, laying the foundation for meaningful contingent communication.

> *I felt like Aipaca really understood what I was saying. There was no confusion or having to word my questions differently like when I have used other chatbots in the past...I felt heard and understood.* [Participant 11]

Similarly, Participant 14 emphasized its ability to maintain coherent, long-form dialogue:

> *It definitely understood what I said every single input. There was no confusing interaction, which even now is hard to do with a chat in a conversation that long.*

Participants also highlighted its corrective capacity, as Participant 19 observed:

> *There were also a few times that I had misspelled a word, and it gave me a response that was related to the correct word. For example, I put "wait" and meant "weight," but it gave me an answer for "weight."*

The chatbot's ability to reference prior exchanges enhanced personalization and coherence:

> *It recalled the things I could do to replace the moments when I'm triggered, such as playing piano. When it seemed to recall and respond properly to what I was saying and was detailed in its response, it made me feel like I could really be helped.* [Participant 7]

Participants emphasized that Aipaca "seemed to listen to me...It didn't seem like it just gave canned responses" [Participant 9], and several remarked on the suitability of its recommendations: "it answered questions that I had, and made recommendations suited to me specifically" [Participant 17].

Finally, participants valued the chatbot's interactive qualities and conversational pacing.

> *It was very conversational; it understood my context and went "back and forth" very well.* [Participant 8]

Others appreciated features that mimicked human typing: "responses came back gradually instead of as one big chunk of text. It gave the illusion that the chatbot was typing in real-time" [Participant 3]. These qualities reinforced the perception of natural, 2-way dialogue.

User Criticisms of Aipaca for Quit-Smoking

While participants generally emphasized Aipaca's strengths, criticisms varied, reflecting individual differences in counseling needs and preferences. Notably, 8 participants reported no dislikes. Some participants felt Aipaca's focus on smoking cessation left little room for personal connection or small talk, making the interaction feel "cold and clinical" [Participant 3] or "like a helicopter mom" [Participant 14]. Others found its strict goal orientation tiring:

> *The chatbot returned everything back to the topic at hand, which was tiring.* [Participant 28]

Yet preferences diverged; while some described the counseling style as pushy, others wanted more firmness:

> *It could be a little more compelling and stricter. I need a kick in the ass.* [Participant 26]

In addition, several participants noted that the impersonal nature of a chatbot reduced accountability:

> *I would be able to "cheat" and not be accountable...knowing it is only a computer.* [Participant 7]

Others pointed to limited cultural resonance:

> *I doubt it ever saw Pulp Fiction and saw how cool John Travolta looked smoking.* [Participant 19]

Aipaca's response style also drew mixed reactions. Quick replies sometimes seemed insincere:

> *It answered within seconds...not seeming to have time to actually read what I wrote.* [Participant 7]

Conversely, a few found the delay between messages frustrating:

> *The seconds of response time felt very long...I thought I lost my internet.* [Participant 26]

XSL•FO
RenderX

Opportunities and Challenges of GenAI Chatbots for Smoking Cessation

Enhance Understanding and Aid Decision-Making

Participants consistently highlighted Aipaca's potential to expand knowledge about smoking's effects and support informed decision-making. They valued access to accurate, comprehensive, and up-to-date information, including actionable guidance on cessation methods and medications:

> *It could help someone become knowledgeable on the effects of smoking and help them make decisions.* [Participant 2]
>
> *It could bring opportunities to provide the latest info on how to best quit smoking.* [Participant 16]

Some appreciated its ability to answer questions difficult to research independently:

> *I could ask Aipaca questions that I don't know how to google.* [Participant 8]

However, concerns were raised about accuracy and reliability:

> *My only concern is it giving false or inaccurate information.* [Participant 2]
>
> *The chatbot has to be very accurate and not make mistakes.* [Participant 4]

Eliciting and Sustaining Motivation

Participants valued Aipaca's capacity to provide continuous, on-demand encouragement and reminders to reinforce motivation:

> *It would be helpful in that it is always available on demand.* [Participant 9]
>
> *It could help remind me why I need to quit and offer support.* [Participant 5]

Several emphasized the importance of proactive engagement, such as follow-ups when inactive:

> *I would like it to check in with me if I haven't used it in a while.* [Participant 17]

At the same time, barriers to sustained use were noted, including fading commitment, forgetting to engage during cravings, and technical issues:

> *I could see potential challenges such as users abandoning their goal completely and forgetting about the chatbot.* [Participant 1]
>
> *What if Aipaca was down for maintenance and I couldn't get advice?* [Participant 11]

Long-Term Support and Companionship

Many participants viewed Aipaca as a potential companion and accountability partner:

> *I view it as a "friend" to stick with me even if I fail.* [Participant 26]
>
> *If used on a consistent basis, it would become my partner, and I wouldn't feel alone.* [Participant 28]

Others saw broader applications for behavior change, such as dieting, or even societal benefits through reduced smoking prevalence:

> *I expect Aipaca to positively affect society and significantly lower the number of smokers.* [Participant 20]

Yet limitations of nonhuman support were evident. Some doubted its capacity for empathy: "It cannot truly understand what this process might be like psychologically" [Participant 9], while others questioned its accountability: "With a chatbot, there's zero social pressure or consequences if I don't follow through" [Participant 14].

Design Feedback and Suggestions From Participants

Participants recommended design improvements to optimize Aipaca for long-term cessation support, emphasizing sustained engagement and stronger interpersonal connections.

Personalized, Gamified, and Proactive Engagement

Many participants suggested personalized reminders, adaptive progress tracking, and milestone recognition to reinforce motivation:

> *I would like it to keep track of my quitting, and to message me if I haven't spoken to it lately.* [Participant 17]
>
> *A progress bar...with periodic reminders and check-ins.* [Participant 28]

Others wanted milestones tied to health or financial benefits, such as heart recovery or money saved [Participants 9 and 17]. Gamification was proposed to boost engagement, including points, badges, or rewards:

> *I could earn points for staying abstinent...making it fun would help me stay committed.* [Participant 13]
>
> *Points I could redeem for items...would be more motivating.* [Participant 15]

Participants also highlighted the value of proactive accountability, with some preferring motivational nudges and others requesting firmer reminders:

> *Keep me on task...chide me if I mess up.* [Participant 28]
>
> *It is objective, so I can be accountable without feeling guilty.* [Participant 7]

Anthropomorphic Design and Empathy

Participants emphasized the need for more humanlike and empathetic interactions. Suggested features included voice-based communication, visual aids, and avatars to make conversations feel natural and engaging:

> *Voice-to-text will make the experience much more convenient.* [Participant 1]
>
> *Charts or supportive images...would make it feel less like a wall of text.* [Participant 3]

Customizable personalities and tones were also requested:

> *Different people need different tones.* [Participant 2]

XSL•FO
RenderX

> *I want to choose a personality that matches mine.* [Participant 23]

Above all, participants stressed empathy and relatability as critical for emotional support:

> *Aipaca should reassure the user that their feelings are normal and encourage them to stay strong.* [Participant 20]
>
> *The most important feature is empathy...like someone in a support group who has also quit smoking.* [Participant 3]

Several suggested giving Aipaca the identity of a former smoker to strengthen credibility and connection:

> *It would be beneficial for the chatbot to have an identity as another smoker, sharing inspirational stories on how it quit successfully.* [Participant 20]

Discussion

Principal Findings

This study evaluated the feasibility of delivering evidence-based smoking cessation counseling through a GenAI chatbot using a triangulated mixed methods design. By integrating quantitative pre-post outcomes, qualitative interviews, and conversation analysis, we provide a multilayered account of how generative AI may support both the effectiveness and the relational dynamics of digital cessation counseling. In addition, we examined the communicative processes through which intervention effectiveness was achieved and situated user perceptions within broader theoretical debates in human-computer interaction and digital health.

Quantitative Outcomes: Feasibility in the Context of Evolving Standards

Pre-post analyses revealed significant improvements in participants' preparedness to quit smoking across knowledge, self-efficacy, and readiness indicators. These proximal effects extend meta-analytic findings that CAs can improve cessation outcomes by demonstrating that GenAI can achieve measurable motivational gains even in the initial counseling session [13,14]. While prior trials often tested scripted or retrieval-based systems with modest effect sizes, our findings suggest that generative models may accelerate short-term change by sustaining more naturalistic engagement. At the same time, our results must be interpreted cautiously. Considering that high attrition and methodological heterogeneity have hampered generalizability in prior CA studies [14], single-session effects risk inflating positive impressions due to novelty, and sustained engagement remains an open question. Moreover, as participants explicitly compared Aipaca to commercial LLMs such as ChatGPT, future trials must account for a shifting sociotechnical baseline: users now expect counseling systems to match the coherence and empathy of widely available generative models. This creates a higher threshold for both effectiveness and acceptability than in earlier CA research.

Qualitative Insights: User Perceptions, Empathy, and Expectations

Interviews highlighted Aipaca's perceived value as an accessible, nonjudgmental, and motivating resource, echoing earlier findings that empathy and personalization drive acceptability in digital cessation tools [18,44]. Importantly, participants did not report uncanny valley effects or any psychological reactance, which are commonly documented when chatbots simulate but fall short of human empathy [45,46]. Instead, they embraced Aipaca's empathetic tone and requested more anthropomorphic features, such as distinct personalities, cultural resonance, and long-term relational continuity. This divergence suggests that uncanny valley discomfort may emerge primarily when a chatbot's language capabilities misalign with its social role. By contrast, Aipaca's generative fluency supported alignment between its assigned role (counselor) and user expectations, mitigating reactance. Yet critiques about accountability, cultural understanding, and rigid goal orientation highlight persistent gaps. Participants expressed concern that an impersonal chatbot could not provide social pressure to maintain abstinence or reflect insider knowledge of smoking culture. These limitations underscore Rogers' client-centered counseling principle that therapeutic alliance depends not only on empathy but also on authenticity and shared understanding [47]. Users' suggestions for gamified progress tracking, proactive reminders, and empathetic personas extend prior human-centered design research (eg, QuitBot and Florence), reinforcing the need to embed co-design processes in future development [22,25].

Conversational Analysis: Contingency, Sequential Structure, and Therapeutic Work

The microsequential analysis of transcripts illuminated how generative AI supports counseling processes at the turn-by-turn level. Three recurrent practices—contextual referencing, formulation-plus-elaboration, and progression toward collaborative planning—demonstrated contingency, the property of conversations in which each utterance builds on prior talk [48]. This interdependence is critical for trust, perceived competence, and social connection [49,50]. By maintaining topical continuity, referencing earlier reasons for quitting, and scaffolding from orientation to collaborative plan-building, Aipaca enacted the very sequencing associated with MI and client-centered therapy. This extends the literature in two ways. First, it provides empirical evidence of process-level mechanisms rarely described in chatbot research. Prior reviews have highlighted relational communication as a gap in CA design [13], but few have shown how GenAI achieves such dynamics in practice. Second, it underscores the limitations of rigid scripting: while Aipaca could sustain coherent therapeutic narratives, its strict redirection of off-topic talk sometimes strained rapport. Future systems may need adaptive topic-switching, reinforcement learning, or sentiment-aware discourse modeling to balance counseling focus with the social nuances that participants valued.

Implications for Practice and Policy

Generative AI chatbots could augment established services such as quitlines by providing continuous, low-intensity support

XSL•FO
RenderX

between counselor calls. Embedding tools such as Aipaca into multilevel cessation infrastructures could increase reach and continuity, particularly for those with limited access to in-person treatment. However, evaluations of ChatGPT and related models show inconsistent adherence to cessation guidelines and occasional misinformation [51]. Clinical deployment, therefore, requires transparent validation protocols, integration of guideline-concordant knowledge bases, and escalation pathways to human specialists. Frameworks such as FUTURE-AI emphasize fairness, robustness, explainability, and traceability as essential for trustworthy health AI. Although chatbots may reduce access barriers for populations with limited social support, mobility, or geographic access to cessation services [18], their design must incorporate cultural adaptation, language inclusivity, and sensitivity to avoid reproducing inequities. Training on diverse linguistic and cultural datasets, together with participatory co-design involving underrepresented populations, will be essential to ensuring fair access and cultural relevance. Such inclusive design practices can strengthen trust, usability, and engagement across high-need populations and ensure that AI-augmented cessation tools contribute to equitable improvements in public health outcomes.

Limitations and Future Work

This study has several limitations that inform directions for future research. First, its cross-sectional design restricts our ability to assess long-term cessation outcomes. While participants showed immediate improvements in preparedness indicators, it remains unclear whether these gains can be sustained over time. Future research should use longitudinal designs to evaluate Aipaca's capacity to support maintenance, relapse prevention, and multisession counseling. Second, the relatively small sample size (n=29) limits statistical power and generalizability. Future studies with larger, more diverse samples would increase the ability to detect meaningful subgroup differences, such as baseline readiness to engage in a quit attempt or variability in digital literacy. Third, the absence of a control group precludes comparison with standard evidence-based approaches to tobacco cessation counseling. Future research should include randomized or noninferiority designs to evaluate Aipaca's effectiveness relative to human-delivered counseling and existing digital interventions. Fourth, the within-subject, pre-post design precludes direct comparison with alternative chatbot architectures. Incorporating between-subject comparisons with rule-based or retrieval-based systems would clarify whether generative AI's adaptive, personalized responses confer a distinctive advantage, both in counseling effectiveness and communication dynamics. Such comparisons could also illuminate how different system designs align with specific counseling goals. Fifth, participants varied in their stage of change, from contemplation to preparation and action. This heterogeneity may have influenced the perceived relevance of the single-session intervention, which emphasized readiness assessment, health education, and quit planning. Future studies should tailor chatbot content to the readiness stage, delivering more targeted, stage-appropriate support. Finally, the sample was predominantly White, limiting insight into the sociocultural factors shaping cessation experiences in minoritized groups. Smoking behavior and treatment engagement are deeply influenced by cultural norms, identity, and social context. Future research should prioritize the recruitment of more diverse populations to ensure that generative AI chatbots can be culturally responsive and equitably address the needs of high-need and underrepresented groups.

Conclusions

This study reported findings from a mixed methods feasibility evaluation of Aipaca, a GenAI chatbot designed to deliver smoking cessation counseling. By integrating quantitative pre-post surveys, qualitative interviews, and conversation analysis, we demonstrated that GenAI can facilitate measurable improvements in cessation preparedness, while also enacting counseling-relevant communication dynamics such as contextual referencing, collaborative planning, and narrative progression. Our findings highlight both the promise and the challenges of deploying GenAI in behavioral health. Participants valued Aipaca's accessibility, personalization, and empathy, yet also articulated expectations for richer social roles, cultural resonance, and long-term accountability. Design feedback emphasized the importance of personalized and proactive engagement, gamified progress tracking, and anthropomorphic features to sustain motivation and strengthen interpersonal connections. Viewed through a communication contingency lens, this study illustrates how GenAI reshapes the dynamics of human-chatbot interaction and redefines users' expectations of digital health tools. More broadly, it offers practical guidance for implementing evidence-based interventions with GenAI chatbots: pair adaptive language generation with human-centered design, embed safeguards for accuracy and trust, and integrate systems into existing cessation infrastructures to ensure clinical relevance and equity.

Acknowledgments

Generative AI tools (ChatGPT, OpenAI) were used solely for grammar checking and language refinement during manuscript preparation. No AI-generated text, data, or references were incorporated without author verification, and all substantive content was conceived, written, and approved by the authors.

Funding

This work was supported by the AIM-AHEAD Coordinating Center, funded by the National Institutes of Health (NIH). Research reported in this publication was supported by the Office of the Director, NIH Common Fund, under award number 1OT2OD032581-02. The content is solely the responsibility of the authors and does not necessarily represent the official views of AIM-AHEAD or the NIH.

XSL•FO
RenderX

This work was also supported by the Accelerating Research Translation (ART) Program, National Science Foundation, under grant 2331409. The funder had no role in study design, data collection and analysis, the decision to publish, and the preparation of the manuscript.

Data Availability

Due to institutional review board (IRB) restrictions (IRB15693), raw counseling transcripts cannot be publicly shared because they may contain health-related information and potentially identifiable details. Deidentified excerpts are included in the manuscript. Deidentified summary data supporting the findings (anonymized pre-post survey data with codebook and aggregate counts for qualitative themes) and analysis scripts are available from the corresponding author upon reasonable request for noncommercial research, subject to an IRB-consistent data use agreement. The full survey instruments and interview protocol are provided in Multimedia Appendices 2 and 3, and a partial system prompt is provided in Multimedia Appendix 4. Materials subject to third-party licensing (eg, Tobacco Treatment Specialist curriculum content embedded in prompts) cannot be redistributed.

Authors' Contributions

CP and RS are co–senior authors and contributed equally to this manuscript.
Conceptualization: YL, DR, LH-W, CP, RS
Data curation: YL, PC, MV
Investigation: YL, PC, MV
Methodology: YL, LH-W, CP, RS
Resources: DR, LH-W, PC
Writing – original draft: RS
Writing – review & editing: all authors

Conflicts of Interest

None declared.

Multimedia Appendix 1
Participants smoking history and nicotine dependence characteristics and the thematic analysis.
[DOCX File , 39 KB - formative_v9i1e73319_app1.docx]

Multimedia Appendix 2
Full survey instruments.
[DOCX File , 63 KB - formative_v9i1e73319_app2.docx]

Multimedia Appendix 3
Semistructured interview guide.
[DOCX File , 27 KB - formative_v9i1e73319_app3.docx]

Multimedia Appendix 4
System prompt.
[DOCX File , 26 KB - formative_v9i1e73319_app4.docx]

Multimedia Appendix 5
GRAMMS checklist.
[DOCX File , 27 KB - formative_v9i1e73319_app5.docx]

References

1. Health effects of smoking. American Lung Association. 2024 Nov 20. URL: https://www.lung.org/quit-smoking/smoking-facts/health-effects/smoking [accessed 2025-11-24]
2. How tobacco smoke causes disease – what it means to you. U.S. Centers for Disease Control and Prevention. 2011 Feb 4. URL: https://stacks.cdc.gov/view/cdc/153198 [accessed 2025-11-24]
3. Pipe AL, Evans W, Papadakis S. Smoking cessation: health system challenges and opportunities. Tob Control 2022 Mar;31(2):340-347. [doi: 10.1136/tobaccocontrol-2021-056575] [Medline: 35241609]
4. Fiore M. Treating Tobacco Use and Dependence: 2008 Update – Clinical Practice Guideline. Rockville, MD: U.S. Department of Health and Human Services; May 2008.

XSL•FO
RenderX

5. Levy DE, Regan S, Perez GK, Muzikansky A, Friedman ER, Rabin J, et al. Cost-effectiveness of implementing smoking cessation interventions for patients with cancer. JAMA Netw Open 2022;5(6):e2216362 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.16362] [Medline: 35679043]
6. Sharpe T, Alsahlanee A, Ward KD, Doyle F. Systematic review of clinician-reported barriers to provision of smoking cessation interventions in hospital inpatient settings. J Smok Cessat 2018;13(4):233-243. [doi: 10.1017/jsc.2017.25]
7. Aggarwal A, Tam CC, Wu D, Li X, Qiao S. Artificial intelligence-based chatbots for promoting health behavioral changes: systematic review. J Med Internet Res 2023;25:e40789 [FREE Full text] [doi: 10.2196/40789] [Medline: 36826990]
8. Laranjo L, Dunn AG, Tong H, Kocaballi AB, Chen J, Bashir R, et al. Conversational agents in healthcare: a systematic review. J Am Med Inform Assoc 2018;25(9):1248-1258 [FREE Full text] [doi: 10.1093/jamia/ocy072] [Medline: 30010941]
9. Schachner T, Keller R, V Wangenheim F. Artificial intelligence-based conversational agents for chronic conditions: systematic literature review. J Med Internet Res 2020;22(9):e20701 [FREE Full text] [doi: 10.2196/20701] [Medline: 32924957]
10. Zhang J, Oh YJ, Lange P, Yu Z, Fukuoka Y. Artificial intelligence chatbot behavior change model for designing artificial intelligence chatbots to promote physical activity and a healthy diet: viewpoint. J Med Internet Res 2020;22(9):e22845 [FREE Full text] [doi: 10.2196/22845] [Medline: 32996892]
11. Yim D, Khuntia J, Parameswaran V, Meyers A. Preliminary evidence of the use of generative AI in health care clinical services: systematic narrative review. JMIR Med Inform 2024;12:e52073 [FREE Full text] [doi: 10.2196/52073] [Medline: 38506918]
12. Carter N, Bryant-Lukosius D, DiCenso A, Blythe J, Neville AJ. The use of triangulation in qualitative research. Oncol Nurs Forum 2014;41(5):545-547. [doi: 10.1188/14.ONF.545-547] [Medline: 25158659]
13. He L, Balaji D, Wiers RW, Antheunis ML, Krahmer E. Effectiveness and acceptability of conversational agents for smoking cessation: a systematic review and meta-analysis. Nicotine Tob Res 2023;25(7):1241-1250 [FREE Full text] [doi: 10.1093/ntr/ntac281] [Medline: 36507916]
14. Bendotti H, Lawler S, Chan GCK, Gartner C, Ireland D, Marshall HM. Conversational artificial intelligence interventions to support smoking cessation: a systematic review and meta-analysis. Digit Health 2023;9:20552076231211634 [FREE Full text] [doi: 10.1177/20552076231211634] [Medline: 37928336]
15. Olano-Espinosa E, Avila-Tomas JF, Minue-Lorenzo C, Matilla-Pardo B, Serrano Serrano ME, Martinez-Suberviola FJ, Dejal@ Group. Effectiveness of a conversational chatbot (Dejal@bot) for the adult population to quit smoking: pragmatic, multicenter, controlled, randomized clinical trial in primary care. JMIR Mhealth Uhealth 2022;10(6):e34273 [FREE Full text] [doi: 10.2196/34273] [Medline: 35759328]
16. Masaki K, Tateno H, Nomura A, Muto T, Suzuki S, Satake K, et al. A randomized controlled trial of a smoking cessation smartphone application with a carbon monoxide checker. NPJ Digit Med 2020;3:35 [FREE Full text] [doi: 10.1038/s41746-020-0243-5] [Medline: 32195370]
17. Wang H, Zhang Q, Ip M, Fai Lau JT. Social media–based conversational agents for health management and interventions. Computer 2018;51(8):26-33. [doi: 10.1109/mc.2018.3191249]
18. Karekla M, Savvides SN. Smoking cessation avatar-led acceptance and commitment therapy digital intervention: feasibility and acceptability in young adults. Transl Behav Med 2021;11(1):198-205. [doi: 10.1093/tbm/ibz128] [Medline: 31628476]
19. Heffner JL, Watson NL, Serfozo E, Kelly MM, Reilly ED, Kim D, et al. An avatar-led digital smoking cessation program for sexual and gender minority young adults: intervention development and results of a single-arm pilot trial. JMIR Form Res 2021;5(7):e30241 [FREE Full text] [doi: 10.2196/30241] [Medline: 34328430]
20. Auriacombe M, Moriceau S, Serre F, Denis C, Micoulaud-Franchi J, de Sevin E, et al. Development and validation of a virtual agent to screen tobacco and alcohol use disorders. Drug Alcohol Depend 2018;193:1-6. [doi: 10.1016/j.drugalcdep.2018.08.025] [Medline: 30321739]
21. Nair US, Rabe B, Brady BR, Bell ML. Predictors of client retention in a state-based tobacco quitline. J Smok Cessat 2020;15(2):67-75. [doi: 10.1017/jsc.2020.10]
22. Loveys K, Lloyd E, Sagar M, Broadbent E. Development of a virtual human for supporting tobacco cessation during the COVID-19 pandemic. J Med Internet Res 2023;25:e42310 [FREE Full text] [doi: 10.2196/42310] [Medline: 38051571]
23. He L, Basar E, Wiers RW, Antheunis ML, Krahmer E. Can chatbots help to motivate smoking cessation? A study on the effectiveness of motivational interviewing on engagement and therapeutic alliance. BMC Public Health 2022;22(1):726 [FREE Full text] [doi: 10.1186/s12889-022-13115-x] [Medline: 35413887]
24. He L, Basar E, Wiers RW, Antheunis ML, Krahmer E. Chatting your way to quitting: a longitudinal exploration of smokers' interaction with a cessation chatbot. Internet Interv 2025;39:100806 [FREE Full text] [doi: 10.1016/j.invent.2025.100806] [Medline: 39996147]
25. Bricker JB, Sullivan B, Mull K, Santiago-Torres M, Lavista Ferres JM. Conversational chatbot for cigarette smoking cessation: results from the 11-step user-centered design development process and randomized controlled trial. JMIR Mhealth Uhealth 2024;12:e57318 [FREE Full text] [doi: 10.2196/57318] [Medline: 38913882]
26. Saiyed A, Layton J, Borsari B, Cheng J, Kanzaveli T, Tsvetovat M, et al. Technology-assisted motivational interviewing: developing a scalable framework for promoting engagement with tobacco cessation using NLP and machine learning. Proc Comput Sci 2022;206:121-131. [doi: 10.1016/j.procs.2022.09.091]

XSL•FO
RenderX

27. Brown A, Kumar AT, Melamed O, Ahmed I, Wang YH, Deza A, et al. A motivational interviewing chatbot with generative reflections for increasing readiness to quit smoking: iterative development study. JMIR Ment Health 2023;10:e49132 [FREE Full text] [doi: 10.2196/49132] [Medline: 37847539]
28. Albers N, Neerincx MA, Brinkman W. Addressing people's current and future states in a reinforcement learning algorithm for persuading to quit smoking and to be physically active. PLoS One 2022;17(12):e0277295 [FREE Full text] [doi: 10.1371/journal.pone.0277295] [Medline: 36454782]
29. The Clinical Practice Guideline Treating Tobacco Use and Dependence 2008 Update Panel, Liaisons, and Staff. A clinical practice guideline for treating tobacco use and dependence: 2008 update – a U.S. Public Health Service report. Am J Prev Med 2008;35(2):158-176 [FREE Full text] [doi: 10.1016/j.amepre.2008.04.009] [Medline: 18617085]
30. Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, et al. GPT-4 technical report. arXiv Preprint posted online March 15, 2023. [doi: 10.48550/arXiv.2303.08774]
31. Council for Tobacco Treatment Training Programs. URL: https://www.ctttp.org/ [accessed 2025-11-06]
32. Lawson PJ, Flocke SA, Casucci B. Development of an instrument to document the 5A's for smoking cessation. Am J Prev Med 2009;37(3):248-254 [FREE Full text] [doi: 10.1016/j.amepre.2009.04.027] [Medline: 19666161]
33. Heatherton TF, Kozlowski LT, Frecker RC, Fagerström KO. The Fagerström test for nicotine dependence: a revision of the Fagerström tolerance questionnaire. Br J Addict 1991;86(9):1119-1127. [doi: 10.1111/j.1360-0443.1991.tb01879.x] [Medline: 1932883]
34. Griffin RJ, Dunwoody S, Neuwirth K. Proposed model of the relationship of risk information seeking and processing to the development of preventive behaviors. Environ Res 1999;80(2):S230-S245. [doi: 10.1006/enrs.1998.3940] [Medline: 10092438]
35. Noh G, Lee SY, Choi J. Exploring factors influencing smokers' information seeking for smoking cessation. J Health Commun 2016;21(8):845-854. [doi: 10.1080/10810730.2016.1177140] [Medline: 27367187]
36. Shiffman S, Balabanis MH, Paty JA, Engberg J, Gwaltney CJ, Liu KS, et al. Dynamic effects of self-efficacy on smoking lapse and relapse. Health Psychol 2000;19(4):315-323. [doi: 10.1037//0278-6133.19.4.315] [Medline: 10907649]
37. Layoun N, Hallit S, Waked M, Aoun Bacha Z, Godin I, Dramaix M, et al. Predictors of readiness to quit stages and intention to quit cigarette smoking in 2 and 6 months in Lebanon. J Res Health Sci 2017;17(2):e00379 [FREE Full text] [Medline: 28676591]
38. O'Cathain A, Murphy E, Nicholl J. The quality of mixed methods studies in health services research. J Health Serv Res Policy 2008;13(2):92-98. [doi: 10.1258/jhsrp.2007.007074] [Medline: 18416914]
39. Cameron R. Lessons from the field: applying the Good Reporting of a Mixed Methods Study (GRAMMS) framework. Electron J Bus Res Methods 2013;11(2):55-56 [FREE Full text]
40. Willinge GJA, Spierings JF, Romijnders KAGJ, Mathijssen EGE, Twigt BA, Goslings JC, et al. Feasibility of a web-based and mobile-supported follow-up treatment pathway for adult patients with orthopedic trauma in the Netherlands: concurrent mixed methods study. JMIR Form Res 2024;8:e57579 [FREE Full text] [doi: 10.2196/57579] [Medline: 39622674]
41. Harry B, Sturges KM, Klingner JK. Mapping the process: an exemplar of process and challenge in grounded theory analysis. Educ Res 2005;34(2):3-13. [doi: 10.3102/0013189x034002003]
42. Sutherland O, Strong T. Therapeutic collaboration: a conversation analysis of constructionist therapy. J Fam Ther 2011;33(3):256-278. [doi: 10.1111/j.1467-6427.2010.00500.x]
43. Wickman SA, Campbell C. An analysis of how Carl Rogers enacted client-centered conversation with Gloria. J Couns Dev 2011;81(2):178-184. [doi: 10.1002/j.1556-6678.2003.tb00239.x]
44. Prochaska JJ, Vogel EA, Chieng A, Kendra M, Baiocchi M, Pajarito S, et al. A therapeutic relational agent for reducing problematic substance use (Woebot): development and usability study. J Med Internet Res 2021;23(3):e24850 [FREE Full text] [doi: 10.2196/24850] [Medline: 33755028]
45. Liu B, Sundar SS. Should machines express sympathy and empathy? Experiments with a health advice chatbot. Cyberpsychol Behav Soc Netw 2018;21(10):625-636. [doi: 10.1089/cyber.2018.0110] [Medline: 30334655]
46. Shao R. An empathetic AI for mental health intervention: conceptualizing and examining artificial empathy. In: EmpathiCH '23: Proceedings of the 2nd Empathy-Centric Design Workshop. New York: Association for Computing Machinery; 2023 Presented at: EmpathiCH '23: Proceedings of the 2nd Empathy-Centric Design Workshop; April 23, 2023; Hamburg, Germany p. 1-6. [doi: 10.1145/3588967.3588971]
47. Rogers CR. The attitude and orientation of the counselor in client-centered therapy. J Consult Psychol 1949;13(2):82-94. [doi: 10.1037/h0059730] [Medline: 18118261]
48. Rafaeli S. From new media to communication. SAGE Annu Rev Commun Res 1988;16:110-134.
49. Burgoon JK, Bonito JA, Bengtsson B, Ramirez A, Dunbar NE, Miczo N. Testing the interactivity model: communication processes, partner assessments, and the quality of collaborative work. J Manag Inf Syst 2015;16(3):33-56. [doi: 10.1080/07421222.1999.11518255]
50. Kellermann K, Sleight C. Coherence: a meaningful adhesive for discourse. Annal Int Commun Assoc 2016;12(1):95-129. [doi: 10.1080/23808985.1989.11678714]

XSL•FO
RenderX

51. Abroms LC, Yousefi A, Wysota CN, Wu T, Broniatowski DA. Assessing the adherence of ChatGPT chatbots to public health guidelines for smoking cessation: content analysis. J Med Internet Res 2025;27:e66896 [FREE Full text] [doi: 10.2196/66896] [Medline: 39883917]

Abbreviations

ACT: acceptance and commitment therapy
AI: artificial intelligence
CA: conversational agent
CBT: cognitive behavioral therapy
FTND: Fagerström Test for Nicotine Dependence
GenAI: generative artificial intelligence
GRAMMS: Good Reporting of a Mixed Methods Study
IRB: institutional review board
LLM: large language model
MI: motivational interviewing
NLP: natural language processing
RCT: randomized controlled trial
RR: risk ratio
TTS: Tobacco Treatment Specialist

Edited by A Mavragani, J Sarvestan; submitted 01.03.25; peer-reviewed by N Albers, N Baharom, MK Milic; comments to author 12.05.25; revised version received 16.10.25; accepted 20.10.25; published 11.12.25.

Please cite as:
Liu Y, Calle P, Vadakekut M, Rubin D, Nagykaldi Z, Doescher M, Hightow-Weidman L, Pan C, Shao R
AI-Enabled Personalized Smoking Cessation Intervention With the Aipaca Chatbot: Mixed Methods Feasibility Study
JMIR Form Res 2025;9:e73319
URL: https://formative.jmir.org/2025/1/e73319
doi:10.2196/73319
PMID:

©Yunlong Liu, Paul Calle, Mariah Vadakekut, Daniel Rubin, Zsolt Nagykaldi, Mark Doescher, Lisa Hightow-Weidman, Chongle Pan, Ruosi Shao. Originally published in JMIR Formative Research (https://formative.jmir.org), 11.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Role of Practitioner- and User-Set Goals in Engagement and Psychological Distress Among Kooth Digital Health Users: Retrospective Analysis

Jacqlyn Yourell[1], PhD; Jennifer Huberty[1], PhD; Terry Hanley[2], PhD; Louisa Salhi[2,3], PhD

[1]Fit Minded Inc., Phoenix, AZ, United States
[2]University of Manchester, Manchester, United Kingdom
[3]Kooth Digital Health, London, United Kingdom

Corresponding Author:
Jacqlyn Yourell, PhD
Fit Minded Inc.
2901 E Greenway Road PO Box 30271
Phoenix, AZ, 85046
United States
Phone: 1 (602) 935 6986
Email: jackie@fit-minded.com

Abstract

Background: Youth and young adult mental health concerns are rising globally, with digital mental health platforms offering a promising solution for accessible support. Among the various features these platforms provide, goal setting and achievement have been shown to positively influence behavior change and mental health outcomes. However, there is limited understanding of how user-set goals compare to those set collaboratively with a practitioner regarding their impact on user engagement and mental health outcomes in digital mental health platforms.

Objective: The purpose of this study was to examine the relationship between various goal-related variables (eg, the number of goals created and progress in user-set and practitioner-set goals) and user engagement as well as mental health (ie, psychological distress) on a free digital mental health platform. A secondary exploratory aim was to assess how different user-presenting issues were associated with platform engagement.

Methods: We leveraged secondary data from a free, web-based mental health platform for youth aged 10 to 25 years in the United Kingdom that offers goal-setting features, emotional journaling, peer support, asynchronous chat with practitioners, and various self-guided well-being activities. Data included in the analyses were from youth and young adults (mean age 15.84 years, SD 2.88; 522/691, 75.5% female) who engaged with the goal-setting feature and completed both pre- and postengagement psychological distress measures between January 2020 and December 2023. We examined the relationship between user-set goals and practitioner-set goals on user engagement and psychological distress via linear regressions. The impact of different user-presenting issues on engagement was also explored via linear regression.

Results: The number of practitioner-set goals created was positively associated with platform engagement (β=.16; P<.001), whereas the number of self-set goals and goal progress, whether self or practitioner set, were not. Progress on practitioner-set goals was significantly associated with reduced psychological distress (β=–.27; P<.001), while progress on self-set goals showed no significant association (P=.16). Physical health–related and school-related presenting issues were the strongest predictors of increased platform engagement (β=.21; P<.001 and β=.17; P<.001, respectively).

Conclusions: These findings underscore the importance of collaborative goal setting in improving mental health outcomes for youth and young adults on digital mental health platforms. By highlighting the role of guided support and goal progression, this study enhances our understanding of how digital mental health platforms can better support young people's mental health and well-being. This paper also highlights how digital mental health platforms can serve as a valuable resource for addressing a wide range of mental health needs.

(JMIR Form Res 2025;9:e70818) doi:10.2196/70818

XSL•FO
RenderX

KEYWORDS

youth; young adults; mental health; psychological distress; digital mental health; goal setting; practitioner

Introduction

Youth mental health concerns are continuing to grow globally, with 1 in 7 adolescents aged 10 to 19 years experiencing a mental disorder. Suicide is the third leading cause of death among those aged 15 to 29 years [1]. In the United Kingdom, more than 1 in 5 children and young adults (aged 8-25 y) experience a mental health disorder [2]. The growing need for accessible mental health support has fueled the popularity of digital mental health platforms, which offer scalable, real-time support that can be accessed from almost anywhere, anytime [3,4]. As the demand and use of digital mental health platforms among youth rise, understanding the factors that drive engagement and positive mental health outcomes is imperative.

User engagement with digital mental health platforms, even those that are free, can be significantly low. For example, a review of more than 93 free digital mental health apps with at least 10,000 downloads revealed that more than 90% of the users discontinued use within 30 days of installing the app [5]. However, this does not account for "episodic engagement," in which users engage with the app as needed rather than daily [6]. The review also excluded apps with therapist-based components to focus on unguided interventions. More research is needed to identify the factors influencing user engagement in digital mental health platforms with guided components, particularly among adolescents and young adults, and over longer periods (>30 d).

Interestingly, the review found that self-guided apps did not show improvements for adolescent and young adult mental health [5]. However, some have reported that therapist-based or guided interventions may lead to better engagement and mental health outcomes, particularly for those experiencing greater distress, as they provide users with structured support and accountability [7,8]. Research suggests that professional guidance in digital mental health interventions is more effective at improving mental health outcomes in adults (aged >18 y) compared to unguided interventions [9]. Further research is needed to identify aspects of self-guided and practitioner-guided strategies that influence mental health outcomes among adolescents and young adults in digital mental health platforms.

Goal setting is an effective strategy for behavior change. A systematic review and meta-analysis of 155 studies found that interventions that involve goal setting can improve a range of behaviors, with the strongest effects observed in younger populations (eg, children and young adults) [10]. The review identified the need for more research into long-term effects (eg, beyond 12 months), specific health domains (eg, mental health), and optimal methods for supporting goal setting. A systematic literature review of 26 studies found that users of digital mental health technologies valued features that supported them in setting and tracking their progress with goals [11], and evidence suggests a link between goal achievement and increased user engagement. For example, a study with more than 4000 users aged 11 to 25 years on a digital mental health app showed that more than half (55.6%) of the users had meaningful improvement in the goals they set collaboratively with a practitioner [12]. Those who had meaningful improvement in their goals spent more time with the app than those who did not, suggesting goal achievement may be associated with higher engagement. Notably, the study only evaluated goals set collaboratively with a practitioner and did not examine relationships with user-set goals. Another study examining goal setting in adult users aged 18 years and older found that those with a practitioner-recorded presenting issue made the most progress in their goals, and goals were more frequently achieved when set with a practitioner than when set by users themselves [13]. Further research is needed to explore how both user-set and practitioner-set goals influence user engagement and psychological distress among adolescents and young adults.

The purpose of this paper was to examine the relationship between various goal-related variables on user engagement and mental health (ie, psychological distress) in a free digital mental health platform designed for adolescents and young adults. A secondary, exploratory aim was to assess how different user-presenting issues were associated with engagement. We hypothesized the following:

1. The number of goals created will be positively associated with increased engagement.
2. Progress in both user-set and practitioner-set goals will be positively associated with engagement.
3. Progress in user-set and practitioner-set goals will be associated with decreases in psychological distress.

Methods

Digital Mental Health Platform

Kooth is a free, innovative web-based mental health platform designed for children and young people aged 10 to 25 years in the United Kingdom. Children and young people hear about Kooth predominantly through schools, health care professionals, friends and family, community events, social media, or online search engines. For more than 2 decades, Kooth has provided the UK National Health Service with an anonymous mental health support platform that is accessible without a formal referral, provided Kooth is commissioned in their local area. Kooth provides a wide array of resources and support options, including therapeutic content, peer-created therapeutic articles, online community forums, asynchronous and synchronous messaging with practitioners, goal setting, emotional journaling, and a variety of self-guided well-being activities.

Participants

Participants in this study were youth who engaged in Kooth and provided consent for their data to be used for research purposes during Kooth registration. Users could opt in or out for their platform data to be used for research purposes at any point by messaging the Kooth team inbox on the platform. All data were collected by Kooth as part of standard care practices. Data were included in analyses for this paper only for users who had data

XSL•FO
RenderX

for goal-related variables (because not all users engaged with this feature on the autonomous platform) and completed a psychological distress measure before and after engaging with the goal feature. Only anonymized data that fit these criteria were shared with the researchers.

Measures

Goals

Goal Creation

Users can set their own goals through the platform using the goal feature, independently of any practitioner interactions. A brief introduction to goals and how to monitor their goal is provided. These goals are moderated by the Kooth moderation team to check any risk related to safeguarding concerns. If a goal is not suitable or a risk is flagged, a practitioner will contact the user through the platform to discuss and safeguard, if needed, based on safeguarding processes in Kooth. Practitioners can also set goals collaboratively with users during chats. For analysis, goal creation was operationalized as the number of goals created, separated into user-set goals and practitioner-set goals. Both were treated as continuous count variables.

Goal Progress

The goal-based outcome tool is a widely used measure for evaluating progress toward goals in mental health interventions for children and young people [14]. This tool supports collaborative goal setting by involving the young person and, optionally, their practitioner. Goals can be established either by the young person, the practitioner, or both and are rated at regular intervals on an 11-point scale from 0 (no progress) to 10 (goal achieved). The goal-based outcome tool helps track progress by comparing initial goal settings with subsequent ratings, providing a clear measure of advancement over time, as all goals start from 0 by default. Goals are regularly reviewed and discussed with the practitioner for users who choose to attend chats on the platform. A movement of 3 points is considered a meaningful change, with higher scores reflecting more significant progress toward goal achievement [12].

For our analysis, we computed goal progress as the difference between the most recent follow-up score and the baseline score of 0 on the primary goal. This change score, which ranges from 0 to 10, reflects the amount of progress made; higher scores indicate greater improvement toward the goal.

Engagement

Platform engagement was operationalized as the number of logins recorded for each user between registration (start date) and last activity (end date). This was treated as a continuous variable in regression analyses. To provide additional descriptive context, we also reported the span of activity in days (number of days between registration and last active date), though this variable was not used in regression models.

Psychological Distress

Psychological distress was assessed via either the Young Person's Clinical Outcomes in Routine Evaluation (YP-CORE; for users aged ≤16 years [15]) or Clinical Outcomes in Routine Evaluation-10 (CORE-10; for users aged ≥17 years [16]) at 6 time points. The final YP-CORE and CORE scores used in analyses represented the last registered score a user completed on the platform, which did not always correspond to their final time point of platform use. Both measures are 10 items that ask participants to rate their experiences of the past week on a 4-point Likert scale, ranging from "not at all" to "all or most of the time." Example items include "I have felt despairing or hopeless" and "I have felt tense, anxious, or nervous." Total scores range from 0 to 40, with higher scores indicating greater levels of psychological distress. These scales were used at registration to provide a baseline score of psychological distress. In chats, practitioners also used this scale to track and monitor psychological distress. Completion of these measures was always optional for users.

Presenting Issues

Presenting issues were psychological, social, or behavioral concerns disclosed by users during interactions on the platform, identified through one-to-one chats, moderation of journals or goal entries on the platform, or moderation of community engagement activities. Presenting issues were not mutually exclusive, meaning users could have multiple presenting issues. These concerns were systematically categorized based on their nature or impact. In this paper, categories included emotional difficulties; home and relationships; mental health; risk, abuse, or safeguarding risk; physical health; school-related issues; suicidality, self-harm, and maladaptive coping; and other presenting issues. The grouping of these presenting issues into these categories was validated by Kooth clinicians and researchers who provided the data extract. For analysis, each category was coded as a binary variable (0=not present and 1=present), and users could be represented in multiple categories simultaneously.

Demographic Data

Demographic data were routinely collected upon registration to the Kooth platform using predefined drop-down boxes where users selected their responses. Categories included age (measured in years), gender identity (female, male, gender fluid, agender, nonbinary, other self-disclosed, and prefer not to say), and ethnicity (White British, African, Asian, Black, mixed backgrounds, Bangladeshi, Caribbean, Chinese, Indian, Irish, Pakistani, any other White background, and not stated). Racial and ethnic categories were grouped as follows: White British, any other White background, African, not stated, and a combined "remaining ethnicities" category (including Asian, Black, mixed backgrounds, Bangladeshi, Caribbean, Chinese, Indian, Irish, Pakistani, White and Asian, White and Black African, and White and Black Caribbean, collectively representing 12.74% (88/691) of the sample, with each subroup individually accounting for fewer than 2% of participants.

Statistical Analysis

Descriptive statistics were used to examine participant demographics (age, gender identity, and ethnicity), engagement, psychological distress, goals set by both the user and the practitioner (average number of goals set and goal progress), and presenting issues. Pearson correlations were used to assess relationships between these key variables.

XSL•FO
RenderX

Paired samples 2-tailed *t* tests were conducted to assess changes in psychological distress. Normality of difference scores was assessed using Shapiro-Wilk (W=0.932; *P*<.001) and Kolmogorov-Smirnov (D=0.173; *P*<.001) tests. Although both were significant, skewness (–0.683) and kurtosis (1.440) indicated only mild deviations, and Q-Q plots supported approximate normality. Given the large sample size (N=691), use of the paired *t* test was deemed robust to these mild deviations. These tests compared distress levels at registration and the user's last distress score to determine whether significant changes occurred. Analyses controlled for age, ethnicity, and gender identity.

All regression models included relevant control variables (eg, age, gender identity, and ethnicity) to adjust for potential confounding factors. For predicting engagement, linear regression was used to examine how goal-related variables (number of goals created and goal progress) predicted user engagement. We ran three models, namely (1) user-set goal variables only, (2) practitioner-set goal variables only, and (3) both user-set and practitioner-set goal variables together, to explore their unique and combined contributions to engagement. An additional, separate linear regression was conducted to assess how user-presenting issues were related to user engagement. Presenting issues were analyzed in a separate regression to assess their relationship with platform engagement to maintain conceptual clarity and avoid model overfitting.

For predicting psychological distress, linear regressions were used to determine how goal-related variables (number of goals created and goal progress by the user and practitioner) predicted decreases in psychological distress. Specifically, we first conducted a regression model, including both user-set and practitioner-set goals variables (number of goals created and progress) as simultaneous predictors, to assess their unique contributions while accounting for shared variance. We then conducted separate regression models for user-set and practitioner-set goals, each including both the number of goals created and goal progress, to examine their individual associations with changes in distress scores.

Ethical Considerations

This study is a retrospective analysis of existing Kooth platform data. At registration, users choose whether their anonymized data may be used for research, and they can change this preference at any time. Only data from users who opted in were included. All data were fully anonymized prior to analysis, and no compensation was provided, as the data reflect naturally occurring platform use. The project design was reviewed and approved by the university research ethics committee at the University of Manchester, United Kingdom (2024-20916-36703).

Results

Overview

The study sample for this paper included 691 Kooth users in the United Kingdom who registered on the platform between January 3, 2020, and December 14, 2023. Notably, a larger number of users engaged with the platform over this time frame, but this analysis focused solely on those who provided consent to participate in the research and met the criteria for this study (ie, setting a goal and having psychological distress scores recorded both before and after setting a goal; a high proportion of users did not have a second psychological distress score). Participants logged in an average of 262.5 times (SD 520.84; range 1-7179 logins). The average interval between participants' registration date and last recorded activity was 98.15 (SD 122.25; median 57; IQR 11-132) days, with a range of 1 to 1134 days, reflecting varied use over time by young people as needed. The average number of user-set goals in the entire sample was very low (<1 goal), largely due to a lack of data for user-set goals, indicating that users were choosing not to engage with the goals feature independently. However, among users who set at least 1 goal themselves, users set an average of 3 goals. For goals set collaboratively with a practitioner in the entire sample, an average of 9 goals were set. Among users whose practitioners set at least 1 goal with them, an average of 13 goals were set. The lack of data for user-set goals also impacted the sample size for goal progress, with user-set goal progress (n=116) having a lower sample size compared to practitioner-set goal progress (N=691).

Participants had a mean age of 15.84 (SD 2.88) years at registration, with ages ranging from 10 to 25 years. Among the 691 participants, the gender identity distribution was predominantly female (n=522, 75.54%), with smaller percentages self-identifying as male (n=112, 16.21%), gender fluid (n=22, 3.18%), agender (n=17, 2.46%), nonbinary (n=8, 1.16%), and other identities (self-disclosed: n=2, 0.28%; preferred not to say: n=8, 1.16%). Participants were predominantly White British (n=544, 78.72%) followed by those from another White background (n=26, 3.76%), African (n=17, 2.46%), not stated (n=16, 2.32%), and each of the remaining ethnicities were all less than 2%, collectively representing 12.74% (88/691) of the population (Bangladeshi, Caribbean, Chinese, Indian, Irish, Pakistani, White and Asian, White and Black African, White and Black Caribbean, and any other Asian, Black, Mixed, or another ethnic group background). Table 1 presents all participant descriptive statistics.

XSL•FO
RenderX

Table 1. Descriptive statistics (N=691).

Demographics	Values
Gender identity, n (%)	
Female	522 (75.54)
Male	112 (16.21)
Gender fluid	22 (3.18)
Agender	17 (2.46)
Prefer not to say	8 (1.16)
Nonbinary	8 (1.16)
Other self-disclosed	2 (0.28)
Race and ethnicity, n (%)	
White British	544 (78.72)
Any other White background	26 (3.76)
African	17 (2.46)
Not stated	16 (2.32)
Remaining ethnicities[a]	88 (12.74)
Age at registration (y), mean (SD)	15.84 (2.88)
Goal variables, mean (SD)	
Number of user-set goals	<1 (1.88)
User-set goal progress (0-10)	5.69 (4.04)
Number of practitioner-set goals	9.04 (10.47)
Practitioner-set goal progress (0-10)	4.68 (3.40)
Platform engagement[b] (logins), mean (SD)	262.5 (520.84)
Psychological distress (0-40), mean (SD)	
Preengagement	27.15 (7.10)
Postengagement	26.16 (8.27)
Difference	–2.47 (6.93)
Presenting issues[c] (yes or no), n (%)	
Suicidality, self-harm, or maladaptive coping	333 (48.19)
Emotion issue	330 (47.75)
Home or relationship issues	345 (49.92)
Mental health issue	391 (56.58)
Other issue	117 (16.93)
Risk, abuse, or safeguarding risk issue	194 (28.07)
Physical issue	116 (16.78)
School issue	257 (37.19)

[a]"Remaining ethnicities" included any other Asian, Black, mixed backgrounds, Bangladeshi, Caribbean, Chinese, Indian, Irish, Pakistani, White and Asian, White and Black African, and White and Black Caribbean.

[b]Data for the engagement variable were from the platform between January 3, 2020, and December 14, 2023, and did not reflect overall platform engagement for all users.

[c]Presenting issues were not mutually exclusive; users may have presented with any combination of issues or none at all.

Engagement Predictors

Linear regression analysis revealed that goal progress, whether for self-set (P=.56) or practitioner-set goals (P=.30), and the number of self-set goals (P=.18) were not associated with engagement. However, the number of practitioner-set goals was positively associated with engagement (β=.16; P<.001). Four types of presenting issues were significantly positively

XSL•FO
RenderX

associated with increased engagement (Table 2), including physical issues (β=.21; *P*<.001); school issues (β=.17; *P*<.001); risk, abuse, or safeguarding issues (β=.14; *P*=.002); and other issues (β=.10; *P*=.01).

Table 2. Linear regression results of the user presenting issues and the Kooth platform engagement.

Independent variable	B[a] (SE; 95% CI)	β[b]	*P* value
Suicidality, self-harm, or maladaptive coping	48.64 (51.75; –52.98 to 150.26)	.04	.35
Emotion issue	58.26 (60.30; –90.16 to 176.67)	.06	.33
Home or relationship issues	16.49 (60.95; –103.20 to 136.28)	.02	.79
Mental health issue	–102.95 (63.77; –228.17 to 22.28)	–.09	.11
Other issue	94.76 (56.71; –16.60 to 206.12)	.07	.09
Risk, abuse, or safeguarding risk issue	163.38 (52.93; 59.45 to 267.31)	.14	.002
Physical issue	294.99 (56.65; 183.74 to 406.24)	.21	<.001
School issue	182.00 (52.96; 78.02 to 285.986)	.17	<.001
Controls			
Age	2.41 (6.19; –9.73 to 14.57)	.01	.70
Ethnicity[c]	71.64 (47.78; –22.18 to 165.47)	.05	.13
Agender[d]	–9.68 (118.71; –242.78 to 223.42)	–.03	.94
Gender fluid[d]	–118.24 (107.48; –329.28 to 92.81)	–.04	.27
Nonbinary[d]	–108.14 (173.16; –448.16 to 231.87)	–.02	.53
Other self-disclosed[d]	–236.11 (342.89; –909.41 to 437.20)	–.02	.49
Male[d]	–100.95 (51.16; –201.41 to –0.49)	–.07	.049

[a]Unstandardized regression coefficient. SEs and CIs apply to B only.

[b]Standardized coefficient.

[c]Reference group: not British.

[d]Reference group: female.

Psychological Distress

Among youth aged 10 to 16 years, 87.44% (390/446) scored above the established clinical cutoff for significant psychological distress on the YP-CORE. For young adults aged 17 years and older, 60% (147/245) experienced severe distress on the CORE-10, 20% (49/245) experienced moderate to severe distress, and 13.06% (32/245) experienced moderate distress, with the remainder experiencing mild, low, or healthy distress levels.

Changes in Psychological Distress Over Time

The paired samples *t* test showed a significant decrease in psychological distress from time point 1 to time point 2 across all users (t_{683}=–9.32; *P*<.001). The mean difference was –2.47 (SD 6.93) points, indicating a small to medium effect size (Cohen *d*=–0.356). When examining differences by age (10-16 y and ≥17 y), both groups showed significant decreases in psychological distress. For those aged 16 years and younger (t_{445}=–6.97; *P*<.001), the mean difference was –2.35 and Cohen *d* was –0.330. For those aged 17 years and older (t_{244}=–6.36; *P*<.001), the mean difference was –2.67 and Cohen *d* was –0.41. The effect size was slightly higher for the older group.

Relationship Between Goal-Related Variables and Psychological Distress

Linear regression results indicated that greater progress in goals set with a practitioner was significantly associated with a decrease in psychological distress (β=–.27; t_{563}=–6.360; *P*<.001) after controlling for engagement. Specifically, for each 1-unit increase in goal progress, psychological distress decreased by 0.59 units (β =–.59, 95% CI –0.77 to –0.41).

Linear regression results indicated that greater progress in goals set by the user themselves was not significantly associated with changes in psychological distress (β=–.14; t_{103}=–1.405; *P*=.16), after controlling for engagement. Results from linear regression analyses are presented in Table 3.

XSL•FO
RenderX

Table 3. Linear regression results of practitioner-set goal progress and the user's change in psychological distress.

Independent variable	B[a] (SE; 95% CI)	β[b]	*P* value
Practitioner-set goal progress	–.579 (0.09; –0.77 to –0.40)	–.27	<.001
Controls			
Age	0.06 (0.10; –0.14 to 0.25)	.02	.58
Ethnicity[c]	1.74 (0.79; 0.19 to 3.28)	.09	.03
Agender[d]	1.24 (1.89; –2.46 to 4.95)	.03	.51
Gender fluid[d]	–2.13 (1.67; –5.45 to 1.19)	–.05	.21
Nonbinary[d]	4.66 (2.95; –1.14 to 10.46)	.06	.12
Other self-disclosed[d]	–0.82 (5.10; –10.85 to 9.20)	–.01	.87
Male[d]	–1.65 (0.91; –3.44 to 0.14)	–.07	.07
Engagement (d)	0.001 (0.001; 0.00 to 0.002)	.09	.03

[a]Unstandardized regression coefficient. SEs and CIs apply to B only.

[b]Standardized coefficient.

[c]Reference group: not British.

[d]Reference group: female.

Discussion

Principal Findings

The purpose of this paper was to examine the relationship among various goal-related variables, presenting issues, engagement, and mental health (ie, psychological distress) in a free digital mental health platform designed for adolescents and young adults. This platform provides an observational perspective on how youth choose to autonomously engage with the platform and goal setting, both by setting goals themselves and collaboratively with a practitioner. Our primary aim was to understand how goal setting and achievement are related to platform engagement and changes in psychological distress among adolescents and young adults using a digital mental health platform. We also explored how different users' presenting issues were related to platform engagement. Our findings revealed that the number of practitioner-set goals was associated with platform engagement, whereas other goal-related variables were not associated with platform engagement. In addition, presenting issues related to physical health; school; and risk, abuse, or safeguarding were associated with platform engagement. Progress in collaboratively set goals by practitioners was linked to a decrease in psychological distress.

Within this sample, the number of practitioner-set (but not self-set) goals created was positively associated with user engagement, while goal progress was not. This suggests that practitioner involvement in collaboratively setting goals with users may play an important role in fostering engagement, aligning with broader evidence from digital mental health research showing that human contact is a key factor in sustaining platform engagement [17]. It is possible that the relationship operates in the other direction—that more engaged users are also more likely to connect with practitioners and set goals collaboratively. In this context, practitioner-set goals may reflect engagement rather than drive it, whereas self-set goals may lack the accountability and support needed to sustain platform use. Other potential influences on engagement, such as peer support features or users' personal motivations, may influence engagement but were not captured in this study [18]. Notably, this dataset was restricted to focus specifically on goal-related variables from users who had 2 measurement points before and after goal engagement relating to psychological distress. This limited the sample and therefore does not capture all factors that may influence user engagement.

Interestingly, users presenting with physical health issues, school-related issues, and risk, abuse, or safeguarding issues demonstrated the strongest associations with engagement in the platform. This suggests that these specific presenting issues might be more salient or urgent for users, prompting users to engage more actively with Kooth. Given the platform's primary advertisements through schools and communities, it may have resonated more with youth experiencing challenges related to school and physical health, particularly because the sample was collected during and shortly after the COVID-19 pandemic, when many youth faced ongoing school-related difficulties [19,20]. The nonsignificant relationship between engagement in Kooth and mental health and home or relationship issues suggests that these specific presenting concerns may not be the sole factors driving how young people engage with the platform. It is possible that users view Kooth as a broader support resource, using it to address a wide range of life challenges beyond mental health alone. This aligns with how Kooth is marketed in the United Kingdom, which emphasizes its role as an early intervention and prevention tool for higher-need populations, offering solution-focused support and encouraging users to seek additional mental health resources (eg, neurodivergence assessments or social care) [21,22]. Our findings are also consistent with previous research showing that users of the digital mental health platform, including youth and young adults, present with a wide range of issues (eg, self-worth, work-related challenges, and physical health concerns) [13,20].

XSL•FO
RenderX

In addition, one study found that users set goals in various areas, including self-help, self-care, emotional exploration, career aspirations, and self-help skills for life, demonstrating the platform's focus on a wide range of personal development beyond mental health [13]. However, it is also possible that the lack of association between mental health and home or relationship issues with engagement in the platform reflects alternative coping strategies, greater reliance on offline support, or different help-seeking preferences among users presenting with these issues. Future research is needed to more fully understand the specific drivers of engagement and the diverse ways young people use Kooth.

Users' psychological distress significantly decreased over time, indicating that the platform had a positive impact on mental health outcomes. However, the relationship between goal progress and decreases in psychological distress reduction was not the same across user-set goals and collaboratively set goals between users and a practitioner. Specifically, progress in practitioner-set goals was associated with a significant decrease in psychological distress, while progress in user-set goals did not show the same relationship (ie, nonsignificant). This finding suggests that goals set with trained practitioners may be more effective in reducing psychological distress than those set independently by users. Practitioner-set goals are likely to be more targeted, structured, and grounded in professional therapeutic approaches, which could explain their greater impact. These results align with existing research that highlights the importance of professional guidance in goal setting within mental health interventions [7,9]. However, the lack of a significant association between user-set goal progress and decreases in psychological distress might be attributable to users setting fewer goals on their own than with a practitioner, as the sample size for user-set goal progress (n=116) was substantially smaller than that for practitioner-set goal progress (N=691).

Limitations

This study had several limitations. First, the small sample size for some variables restricted the ability to stratify results by demographic factors, such as age, gender identity, or ethnicity, which could have provided more detailed insights into different user groups. Second, the sample was predominantly female, which may limit generalizability. Third, the time frame was limited to the period when goals were set, which may not fully capture the long-term impact of engagement on the entire platform. Fourth, more granular use metrics (eg, feature use, session length, and frequency patterns) were not available, which limits the ability to capture the full range of user engagement behaviors. Fifth, presenting issues were determined by practitioners based on text-based chats rather than self-reports from users, which may introduce bias in how issues are categorized and understood. Finally, as the user's choice is highly important to Kooth, users were not required to complete the psychological distress measures. Collecting these data outside of the routinely collected data would provide a more complete dataset, although leveraging real-world data is a strength, as it provides a more accurate understanding of the true effectiveness of digital platforms.

Future Research

Future research with larger sample sizes should further investigate the impact of both self-set and practitioner-set goals on psychological distress to better understand how these different types of goals influence mental health outcomes. Larger samples would provide more robust data to examine potential differences in goal-setting mechanisms and their effectiveness. Future research should aim to include a more balanced gender identity distribution to better understand potential differences in outcomes across gender identities. In addition, this research could help identify whether self-set goals have a distinct or complementary effect compared to practitioner-set goals, ultimately informing more tailored and effective strategies for improving psychological distress. Future research could include a qualitative study to explore what practitioners are doing to set goals and whether this is a mechanism that improves psychological distress progress, particularly in scaling digital platforms. In addition, understanding how users can be guided to set their own goals outside of practitioner engagement would be valuable for developing tools that facilitate greater user autonomy. Considering users did not engage with the goals feature as much on their own compared to setting goals with a practitioner, one suggestion is to revise the platform's language to guide users through the independent goal-setting process, making it more integrated into the platform journey to create a more guided experience similar to that of practitioner-set goals. Future studies should consider additional demographic factors (eg, socioeconomic status) or other individual characteristics to explore how these factors may influence the relationship between goal progress and psychological distress. In addition, examining these relationships in other populations beyond the United Kingdom would improve generalizability across young people. These recommendations are consistent with a recent scoping review where the need for more research to better define, implement, and measure goal-oriented practices in youth mental health was highlighted [23].

Conclusions

This study examined the relationship among presenting issues, goal setting, and progression (both user set and those set collaboratively with a practitioner) and young people's engagement with a free, anonymous digital mental health platform, Kooth. We also examined the impact of goal setting and progress on users' psychological distress. We leveraged a unique dataset that provided real-world evidence from a nonstructured digital platform where users had the autonomy to engage with different aspects of the platform based on their preferences. Broadly, we found that progress in goals set collaboratively with a practitioner predicted a decrease in psychological distress. For users' presenting issues, school-related and physical health concerns were the strongest predictors of platform engagement. The Kooth platform may have the potential to serve as a resource for a wide range of young people's mental health needs, providing vital early intervention and prevention support for navigating diverse challenges, including those beyond mental health.

Importantly, this study highlights the value of setting and progressing toward goals, particularly those collaboratively

XSL•FO
RenderX

established with practitioners. Goals cocreated with practitioners were associated with decreases in psychological distress while self-set goals were not, emphasizing the role of guided support in making meaningful changes in mental health. Fostering collaboration between young users and practitioners may optimize mental health improvements on digital mental health platforms. This paper adds to the existing literature by deepening our understanding of how goal setting and achievement can be optimized in digital mental health platforms to better support adolescents and young adults facing a variety of issues, while also providing insights into how engagement differs based on presenting issues.

Data Availability

The datasets presented in this study are not readily available because they are proprietary and subject to confidentiality restrictions. Requests to access the datasets should be directed to Kooth Digital Health (research@kooth.com).

Conflicts of Interest

JH is the chief executive officer of Fit Minded and consults with Kooth as a fractional chief science officer. JH was hired by Kooth to contribute to scientific strategy but does not receive incentives for the outcomes of the research. JY is an employee of JH and is independent of Kooth leadership. JH and JY's roles are to ensure the quality of science regarding Kooth. The authors do not have stock in Kooth and receive no financial incentives from the sales of Kooth. The authors do not receive any financial benefit from the outcomes of this study. TH is contracted by Kooth Digital Health. LS is employed by Kooth Digital Health and receives honorarium and shares in Kooth annually; however, these are not contingent on study performance or scientific outputs. LS is also an honorary senior research fellow at the University of Manchester.

References

1. Mental health of adolescents. World Health Organization. 2025 Sep 25. URL: https://www.who.int/news-room/fact-sheets/detail/adolescent-mental-health [accessed 2025-11-20]
2. One in five children and young people had a probable mental disorder in 2023. National Health Service England. 2023 Nov 21. URL: https://www.england.nhs.uk/2023/11/one-in-five-children-and-young-people-had-a-probable-mental-disorder-in-2023 [accessed 2024-12-20]
3. Lehtimaki S, Martic J, Wahl B, Foster KT, Schwalbe N. Evidence on digital mental health interventions for adolescents and young people: systematic overview. JMIR Ment Health 2021 Apr 29;8(4):e25847 [FREE Full text] [doi: 10.2196/25847] [Medline: 33913817]
4. Pretorius C, Chambers D, Coyle D. Young people's online help-seeking and mental health difficulties: systematic narrative review. J Med Internet Res 2019 Nov 19;21(11):e13873 [FREE Full text] [doi: 10.2196/13873] [Medline: 31742562]
5. Baumel A, Muench F, Edan S, Kane JM. Objective user engagement with mental health apps: systematic search and panel-based usage analysis. J Med Internet Res 2019 Sep 25;21(9):e14567 [FREE Full text] [doi: 10.2196/14567] [Medline: 31573916]
6. Boucher EM, Raiker JS. Engagement and retention in digital mental health interventions: a narrative review. BMC Digit Health 2024 Aug 08;2:52. [doi: 10.1186/s44247-024-00105-9]
7. Jacob J, Stankovic M, Spuerck I, Shokraneh F. Goal setting with young people for anxiety and depression: what works for whom in therapeutic relationships? A literature review and insight analysis. BMC Psychol 2022 Jul 13;10(1):171 [FREE Full text] [doi: 10.1186/s40359-022-00879-5] [Medline: 35831897]
8. Leech T, Dorstyn D, Taylor A, Li W. Mental health apps for adolescents and young adults: a systematic review of randomised controlled trials. Child Youth Serv Rev 2021 Aug;127:106073. [doi: 10.1016/j.childyouth.2021.106073]
9. Baumeister H, Reichler L, Munzinger M, Lin J. The impact of guidance on internet-based mental health interventions — a systematic review. Internet Interv 2014 Oct;1(4):205-215. [doi: 10.1016/j.invent.2014.08.003]
10. Epton T, Currie S, Armitage CJ. Unique effects of setting goals on behavior change: systematic review and meta-analysis. J Consult Clin Psychol 2017 Dec;85(12):1182-1198. [doi: 10.1037/ccp0000260] [Medline: 29189034]
11. Catania J, Beaver S, Kamath RS, Worthington E, Lu M, Gandhi H, et al. Evaluation of digital mental health technologies in the United States: systematic literature review and framework synthesis. JMIR Ment Health 2024 Aug 30;11:e57401 [FREE Full text] [doi: 10.2196/57401] [Medline: 39213023]
12. Jacob J, Costa da Silva L, Sefi A, Edbrooke‐Childs J. Online counselling and goal achievement: exploring meaningful change and the types of goals progressed by young people. Couns Psychother Res 2021;21(3):502-513. [doi: 10.1002/capr.12363]
13. Banwell E, Salhi L, Hanley T, Facey‐Campbell N. The use of goal-based outcome measures in digital therapy with adults: what goals are set, and are they achieved? Couns Psychother Res 2022 Sep 24;23(3):770-780. [doi: 10.1002/capr.12584]
14. Law D, Jacob J. Goals and Goal Based Outcomes (GBOs): Some Useful Information. London, UK: CAMHS Press; 2013.
15. Twigg E, Cooper M, Evans C, Freire E, Mellor-Clark J, McInnes B, et al. Acceptability, reliability, referential distributions and sensitivity to change in the Young Person's Clinical Outcomes in Routine Evaluation (YP-CORE) outcome measure:

XSL•FO
RenderX

replication and refinement. Child Adolesc Ment Health 2016 May 03;21(2):115-123. [doi: 10.1111/camh.12128] [Medline: 32680371]
16. Barkham M, Bewick B, Mullin T, Gilbody S, Connell J, Cahill J, et al. The CORE - 10: a short measure of psychological distress for routine use in the psychological therapies. Couns Psychother Res 2013 Feb 26;13(1):3-13. [doi: 10.1080/14733145.2012.729069]
17. Werntz A, Amado S, Jasman M, Ervin A, Rhodes JE. Providing human support for the use of digital mental health interventions: systematic meta-review. J Med Internet Res 2023 Feb 06;25:e42864 [FREE Full text] [doi: 10.2196/42864] [Medline: 36745497]
18. Banwell E, Hanley T, De Ossorno Garcia S, Mindel C, Kayll T, Sefi A. The helpfulness of web-based mental health and well-being forums for providing peer support for young people: cross-sectional exploration. JMIR Form Res 2022 Sep 09;6(9):e36432 [FREE Full text] [doi: 10.2196/36432] [Medline: 36083629]
19. Chaabane S, Doraiswamy S, Chaabna K, Mamtani R, Cheema S. The impact of COVID-19 school closure on child and adolescent health: a rapid systematic review. Children (Basel) 2021 May 19;8(5):415 [FREE Full text] [doi: 10.3390/children8050415] [Medline: 34069468]
20. Knipe D, de Ossorno Garcia S, Salhi L, Afzal N, Sammut S, Mainstone-Cotton L, et al. Digital mental health service engagement changes during Covid-19 in children and young people across the UK: presenting concerns, service activity, and access by gender, ethnicity, and deprivation. PLoS One 2025;20(2):e0316468 [FREE Full text] [doi: 10.1371/journal.pone.0316468] [Medline: 39946352]
21. Gledhill J. In response to the recent NHS Providers Report: early intervention and prevention remains the best approach to addressing the mental health crisis. Kooth Digital Health. URL: https://gov.kooth.com/kooth-uk/kooths-response-to-the-nhs-providers-report [accessed 2024-12-23]
22. Newhouse K. The urgent need for preventative mental health support for young people. Kooth Digital Health. URL: https://gov.kooth.com/kooth-uk/preventative-mental-health-support-for-young-people [accessed 2024-12-23]
23. Jacob J, Wozney L, Oddli HW, Duncan C, Chorney J, Emberly D, et al. Goal-oriented practices in youth mental health and wellbeing settings: a scoping review and thematic analysis of empirical evidence. Psychol Psychother 2025 Jun 13;98(2):431-477. [doi: 10.1111/papt.12564] [Medline: 39670986]

Abbreviations

CORE: Clinical Outcomes in Routine Evaluation
YP-CORE: Young Person's Clinical Outcomes in Routine Evaluation

Edited by D Gan; submitted 02.01.25; peer-reviewed by MI Bhatti, Z Zhang, S Lord; comments to author 02.06.25; accepted 16.10.25; published 15.12.25.

Please cite as:
Yourell J, Huberty J, Hanley T, Salhi L
The Role of Practitioner- and User-Set Goals in Engagement and Psychological Distress Among Kooth Digital Health Users: Retrospective Analysis
JMIR Form Res 2025;9:e70818
URL: https://formative.jmir.org/2025/1/e70818
doi:10.2196/70818
PMID:

©Jacqlyn Yourell, Jennifer Huberty, Terry Hanley, Louisa Salhi. Originally published in JMIR Formative Research (https://formative.jmir.org), 15.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Impact of a Digital Peer-Supported App on Daily Steps and Lifestyle Changes Among Individuals With Prediabetes and Early-Stage Type 2 Diabetes: Prospective, Nonrandomized Controlled Trial

Shota Yoshihara[1,2], MSc, OTR; Takeo Shibata[3], PhD; Mizuki Kosone[3], MSc; Megumi Shibuya[2], MSc; Kayoko Takahashi[1,4], OTR, ScD

[1]Department of Rehabilitation Sciences, Graduate School of Medical Sciences, Kitasato University, Kanagawa, Japan
[2]Department of Business Development, A10 Lab Inc, Tokyo, Japan
[3]Department of Health Management, School of Health Studies, Tokai University, Kanagawa, Japan
[4]Department of Occupational Therapy, School of Allied Health Science, Kitasato University, Kanagawa, Japan

Corresponding Author:
Shota Yoshihara, MSc, OTR
Department of Rehabilitation Sciences
Graduate School of Medical Sciences
Kitasato University
1-15-1 Kitasato, Sagamihara
Kanagawa, 252-0373
Japan
Phone: 81 042 778 9305
Email: yoshihara.shota@st.kitasato-u.ac.jp

Abstract

Background: Physical activity is a simple, low-risk intervention that could be integrated into daily life to improve glycemic control in individuals with prediabetes and early-stage type 2 diabetes mellitus (T2DM). However, maintaining physical activity remains challenging, even when its benefits are well understood. Although digital peer support has the potential to promote and maintain physical activity, its effectiveness has not yet been sufficiently established.

Objective: This study examined the impact of a digital peer-supported app on daily step goal achievement and average daily step counts among individuals with prediabetes and early-stage T2DM.

Methods: This 3-month, prospective, nonrandomized controlled trial recruited participants aged 40-79 years with prediabetes or early-stage T2DM. The participants were divided into a digital peer-supported app group and a control group. The digital peer-supported app group tracked their daily steps, shared their progress with small peer groups, and received real-time feedback and support within the app. The control group tracked their steps individually using a pedometer. The primary outcome was the achievement rate of daily step goals. Secondary outcomes included the average daily step count, BMI, glycosylated hemoglobin A_{1c} level, blood pressure, and self-reported lifestyle behaviors.

Results: A total of 32 participants (digital peer-supported app group: n=18 and control group: n=14) completed the study. The digital peer-supported app group reported a significantly higher median daily step goal achievement rate (57.2%, IQR 32.2%-90% vs 26.7%, IQR 10%-64.4%; *P*=.04) and daily step count (6854, IQR 4846-10388 steps vs 3946, IQR 3176-6832 steps; *P*=.03) compared to the control group. No significant differences were observed in glycosylated hemoglobin A_{1c} levels, blood pressure, BMI, or lifestyle behaviors.

Conclusions: Our findings inform research in this field by suggesting that a digital peer-supported app may support daily step goal achievement and increase step counts among individuals with prediabetes and early-stage T2DM over the 3-month study period. The digital peer-supported app facilitated real-time feedback, peer approval, and continuous engagement to support participation in light physical activity.

Trial Registration: UMIN-CTR UMIN000039466; https://center6.umin.ac.jp/cgi-open-bin/ctr_e/ctr_view.cgi?recptno=R000044999

XSL•FO
RenderX

(JMIR Form Res 2025;9:e75953) doi:10.2196/75953

KEYWORDS

apps; digital peer support app; eHealth; mHealth; mobile health; mobile phone; peer support; type 2 diabetes mellitus; prediabetes

Introduction

Type 2 diabetes mellitus (T2DM) is a global public health issue that increases health care costs, strains health care systems, and reduces life expectancy and quality of life [1]. Unhealthy lifestyle behaviors, such as excessive consumption of fat, sugar, and alcohol, as well as insufficient physical activity, contribute significantly to the development of T2DM [2].

Previous studies have shown that physical activity is an important nonpharmacological intervention for the management of patients with T2DM [3-5]. Among the various forms of physical activity, walking is a beneficial form of physical activity that can be easily integrated into daily commuting for individuals with T2DM, including busy workers and older adults, with minimal risk of adverse effects [6]. A meta-analysis of 20 randomized controlled trials (RCTs) reported that walking significantly decreased glycosylated hemoglobin A_{1c} (HbA_{1c}) by 0.50% (95% CI –0.78% to –0.21%) among patients with T2DM [7]. Similarly, another meta-analysis including 10 RCTs showed that individuals with prediabetes showed a decrease in HbA_{1c} levels after engaging in aerobic training compared to the control group, with a weighted mean difference of 0.34% (95% CI –0.45 to –0.23) [8].

Individuals with prediabetes and T2DM often struggle to enhance and maintain their physical activity, including walking [9]. However, several previous studies have reported that peer support can be an effective form of external assistance in addressing this challenge [10-12]. Nevertheless, traditional lifestyle counseling in clinical settings, such as peer support, requires significant human and time resources and is often provided through face-to-face interactions [13].

To address the limitations of conventional peer support, previous studies have examined digital peer support, which involves group-based apps that enable users to encourage each other toward healthy behaviors [14-16]. Given the evidence from face-to-face peer support studies showing increased physical activity, including walking, digital peer support may also be a promising approach for promoting physical activity, as measured by step counts. Moreover, amid growing interest in diverse digital tracking technologies, including continuous glucose monitoring [17], smartphone-derived step counts offer low-burden, automatically captured, dynamic data that are scalable across populations and may complement contemporary diabetes care.

Therefore, in this study, we aimed to examine the effect of a digital peer support app on the achievement rate of daily step goals and average daily step count, recorded by smartphones, among individuals with prediabetes and early-stage T2DM compared to nondigital peer support. If implemented effectively, the digital peer-supported app can facilitate step count increases in a convenient, cost-effective, and remote manner.

Methods

Study Settings

This study was a 3-month, prospective, non-RCT conducted between October and December 2019 as part of the Kanagawa "ME-BYO Living Lab" program [18]. The program partners with local municipalities, health-conscious companies (in this study, A10 Lab Inc, the developer and operator of the digital peer-supported app), and academic institutions that conducted independent evaluations of program effectiveness, providing a real-world evaluation platform for assessing the functionality and effectiveness of health-related interventions and their public health impact. Within this framework, A10 Lab Inc submitted the proposal and supplied the intervention app; program-operating funds supported this study. The protocol was reviewed by the Ethics Review Committee of Healthcare Systems Co, Ltd, a private company ethics committee retained under the program. Local municipalities assisted with participant recruitment at the request of A10 Lab Inc.

Participants

Participants were recruited from individuals living or working in Kanagawa, Japan, through a municipality-led program aimed at increasing step counts for diabetes management. Participant recruitment was conducted through the dissemination of study invitations via posters, leaflets, and municipal public relations newsletters. Interested individuals were required to complete a preliminary registration via email.

Eligible participants were adults aged 40-79 years with prediabetes and early-stage T2DM, defined as HbA_{1c} levels between 5.6% and 7.0%. Participants were required to own a mobile phone. Individuals with severe diabetes-related symptoms or disabilities, those diagnosed with conditions that could interfere with the study, and those deemed unsuitable by the research team were excluded. The HbA_{1c} range of 5.6%-7.0% was selected to align with the municipal referral program, identifying prediabetes (5.6%-6.4%) and early-stage T2DM (≥6.5%), and to reflect the real-world behavioral focus of the intervention on increasing step counts via peer support.

Participants self-selected into either the digital peer-supported app group or the control group (with no digital peer-supported app). Although initially planned, randomization was not undertaken as the study was embedded in a municipality-led public program that prioritized participant autonomy. Eligibility was determined from recent health checkups or routine clinical testing results, and participants continued their usual clinical care throughout the trial. As the intervention comprised a minimal-risk behavioral program (step-count promotion with peer support) that did not alter medical management, no pre-enrollment medical examination or physician clearance was required, and no formal physician-verification workflow was implemented. Four participants who owned only feature phones

XSL•FO
RenderX

(nonsmartphones) were placed in the control group and were provided with pedometers because they were unable to use the app. In addition, 3 participants who owned smartphones but lacked the technical skills to use the app were assigned to the control group for feasibility reasons.

Participants continued usual care throughout the trial. Medication use and related modifications, including those concerning glucose-lowering agents (eg, metformin, GLP-1 receptor agonists, SGLT2 inhibitors, and insulin), antihypertensives, or lipid-lowering therapies, were not systematically recorded, and no restrictions were imposed on routine prescribing.

Program

All participants attended a face-to-face program introduction and set personal daily step goals for prediabetes and early-stage T2DM prevention and management. They were then divided into a digital peer-supported app group and a control group. The digital peer-supported app group received a tutorial during their initial login and used the app for 3 months. They could track their daily steps and work toward their goals as part of a team of up to 5 members, posting their step count and a photo once a day to share their progress and encourage mutual support. In contrast, the control group tracked their steps using a smartphone pedometer app, but worked individually without team participation, posting, or progress sharing.

Participants completed questionnaires and health checks (eg, sex, age, height, weight, HbA_{1c} level, and blood pressure) at the beginning and end of the program to assess changes in lifestyle habits and health behaviors. All participants submitted their 24-hour step counts via smartphones using a dedicated QR code at the end of the study.

Intervention of Digital Peer-Supported Apps

In this study, we used Minchalle (A10 Lab Inc), a commercially available digital peer-supported app for both Android and iPhone [14-16]. Originally developed to help users build desirable habits such as exercising, dieting, and learning English, the app has been downloaded over 1.5 million times. For this study, the app was customized to promote daily step achievement through anonymous group chats with up to 5 participants (Figure 1). Participants in the digital peer-supported app group posted their step counts, photos, and comments on the group chat once a day.

Figure 1. Examples of app screens. (A) Select a group, (B) post photos, view step counts, and comment within the group, posting a picture taken that day and commenting on the day's events; (C) view the contents of the posts in the group, (D) respond to the posts by group members, and (E) check the records.

The main functions of the app were as follows: (1) posting step counts, photos, and comments about the day (Figures 1A-C); (2) receiving approval from group members (Figure 1D); (3) setting step-count goals at the group level; (4) receiving feedback on the team's total daily step count; and (5) automatic group removal if a participant remained inactive for 15 days. Step counts were recorded using a smartphone, either manually or automatically, via Google Fit or Apple Health, and the app reported the daily step count at the time of posting. Participants could post comments or photos multiple times a day and interact with other members; however, this was not mandatory. The app was available to participants at no cost (Multimedia Appendix 1).

Measurements

Participants' demographic characteristics were assessed using a questionnaire that included sex (male or female), age group (40-49, 50-59, 60-69, or 70-79 years), number of cohabiting family members (continuous), occupational status (full-time or regular employee, self-employed, part-time or temporary worker, homemaker, or unemployed), and diabetes care support (yes or no). In addition, participants reported their average daily step count over the past month (<1000, 1000-1999, 2000-2999, 3000-3999, 4000-4999, 5000-5999, 6000-6999, 7000-7999, 8000-8999, 9000-9999, ≥10000, or unknown). They were also asked to specify their daily step goals (steps/day) to assess individual physical activity targets (continued).

XSL•FO
RenderX

Our primary endpoint was the achievement rate of average daily step goals over the 3-month period. The achievement rate was calculated as the number of days in which the participants exceeded their individually set step goals, recorded via smartphones and pedometers, divided by the total number of study days. The secondary endpoint was the average daily step count during the study period. Additional secondary outcomes included BMI, HbA_{1c} levels, blood pressure, and lifestyle questionnaire responses. BMI was calculated by dividing weight (kg) by height (m^2). Participants completed a web-based questionnaire and had their weight, blood pressure, and HbA_{1c} levels measured at clinics or other facilities at baseline and after 3 months.

The same assessment schedule was applied for the lifestyle questionnaire, which evaluated the health status, lifestyle behaviors, and physical activity. Participants rated their overall health status (good, fairly good, not very good, or poor). Balanced meal consumption (staple food, main dish, and side dishes at least twice per day) was reported as (almost every day, 4-5 days per week, 2-3 days per week, or rarely). Healthy eating habits were assessed based on the frequency with which participants consciously selected healthy foods and practiced mindful eating (always, often, rarely, or never). Engagement in moderate-intensity physical activity (≥30 minutes/session) was categorized based on frequency as 2 or more times per week, once per week, 2-3 times per month, once per month, or rarely. Confidence in maintaining regular physical activity was rated on a 4-point scale (very confident, confident, not very confident, and not confident at all). Participants also reported their average daily sleep duration over the past month (<5, 5-6, 6-7, 7-8, 8-9, or ≥9 hours). Alcohol consumption frequency was recorded as daily, 3-4 times per week, 1-2 times per week, 1-2 times per month, or rarely or never. Finally, perceived stress levels were assessed on a 4-point scale as very high, high, low, or very low.

Statistical Analysis

The Pearson chi-square test was used to confirm the categorical variable homogeneity between the 2 groups at baseline, and the Mann-Whitney *U* test was used for continuous variables.

The primary outcome was analyzed using the Mann-Whitney *U* test to compare daily step achievement rates between the digital peer-supported and control groups. Secondary outcomes included the daily average step count, analyzed using the Mann-Whitney *U* test. Changes in BMI, HbA_{1c} levels, and blood pressure were assessed using within-group *P* values calculated using the Wilcoxon signed-rank test and between-group *P* values calculated using the Mann-Whitney *U* test. Questionnaire-based analyses were conducted to examine lifestyle behavior–related changes. The Stuart-Maxwell test was used to evaluate within-group changes in categorical variable proportions over 3 months, and between-group comparisons were again performed using the Mann-Whitney *U* test.

To test the robustness of results, we performed a linear mixed-effects model with fixed effects for groups, days, and their interactions, and a random intercept for each participant to examine group-level trends in daily average step counts over time. Moreover, separate mixed-effects models were fitted within each group to assess within-group trends. Days with 0 recorded steps were treated as missing, assuming nonwear or device error, and were imputed using each participant's average daily step count derived from their nonmissing data.

All analyses were performed using the Stata statistical software (version 18.0; StataCorp LLC) and IBM SPSS Statistics (version 24.0; IBM Corp). Two-tailed *P* values <.05 were considered statistically significant.

Ethical Considerations

All participants provided written informed consent before their enrollment. This study adhered to the ethical principles outlined in the Declaration of Helsinki and was approved by the independent Ethics Review Committee of Healthcare Systems Co, Ltd, a private company ethics committee (approval no 1911). All collected data were anonymized before analysis to ensure participant confidentiality and privacy.

All participants who completed the study received a JPY 5000 (approximately US $46.26) gift card as compensation, irrespective of whether they submitted end-of-study self-reported outcomes. The 3-month endline survey invitation and compensation distribution were issued concurrently. The trial was registered with the University Hospital Medical Information Network (UMIN) Clinical Trials Registry (UMIN000039466).

Results

Thirty-eight participants were initially enrolled in this study. In the digital peer-supported app group, 1 participant did not respond at the time of follow-up, 2 participants were excluded because of inaccurate step count data, and 2 participants were excluded because they had no time to walk due to shift work. One participant in the control group who did not respond at the time of the follow-up was excluded. The final statistical analysis included 18 participants in the digital peer support app group and 14 in the control group. Figure 2 shows a flowchart of the study participants.

XSL•FO
RenderX

Figure 2. Flowchart of the study participants.

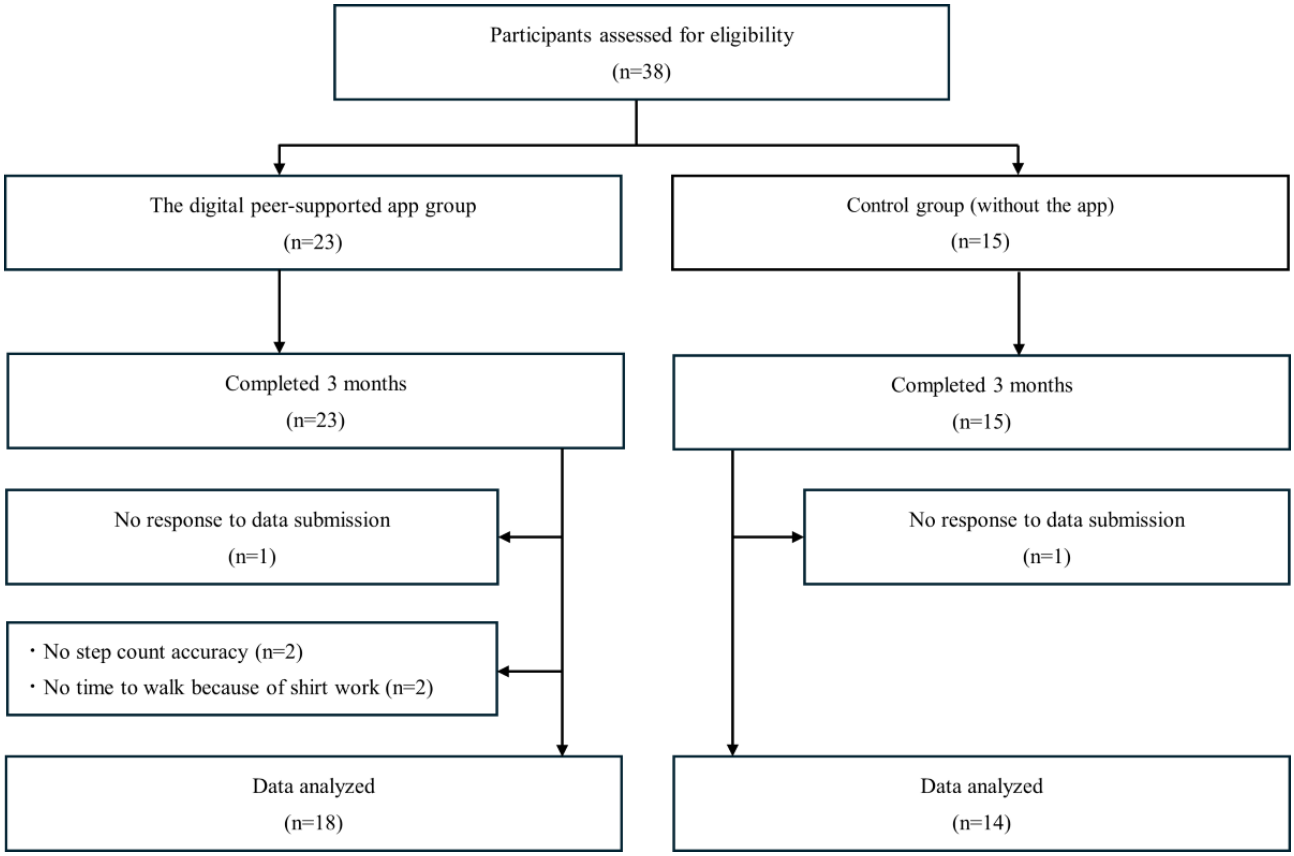

Table 1 shows the baseline characteristics of the participants in the digital peer-supported app (n=18) and control (n=14) groups. Age, sex, household composition, and employment status were similar between the 2 groups. No notable differences were observed in the prevalence of diabetes care support. Moreover, physical measurements (BMI, HbA_{1c} level, and blood pressure) were comparable between the groups, and there were no significant differences between the groups in terms of average daily step counts and daily step goals. The proportion with prediabetes (HbA_{1c} 5.6%-6.4 %) was 15/18 (83.3%) and 11/14 (78.6%) in the digital peer-supported app and control groups, respectively, with no statistically significant differences. Therefore, baseline characteristics were considered well-balanced between the 2 groups.

XSL•FO
RenderX

Table 1. Baseline participant characteristics.

Variable	Digital peer-supported app group (n=18)	Control group (n=14)	*P* value
Age (years), n (%)			.98
40-49	2 (11.1)	2 (14.3)	
50-59	7 (38.9)	5 (35.7)	
60-69	7 (38.9)	5 (35.7)	
70-79	2 (11.1)	2 (14.3)	
Sex, n (%)			.75
Male	8 (44.4)	7 (50)	
Female	10 (55.6)	7 (50)	
Number of household members, n (%)			.33
Living alone	2 (11.1)	2 (14.3)	
2 people	10 (55.6)	5 (35.7)	
3 people	2 (11.1)	4 (28.6)	
4 people	4 (22.2)	1 (7.1)	
5 people	0 (0.0)	1 (7.1)	
7 people	0 (0.0)	1 (7.1)	
Employment status, n (%)			.33
Full-time employee	9 (50)	5 (35.7)	
Self-employed	1 (5.6)	2 (14.3)	
Part-time worker	1 (5.6)	3 (21.4)	
Homemaker	0 (0.0)	1 (7.1)	
Unemployed	7 (38.9)	3 (21.4)	
Support for diabetes care, n (%)	13 (72.2)	7 (50)	.20
BMI (kg/m²), median (IQR)	24 (21-26)	23 (22-26)	.88
HbA_{1c}[a] (%), median (IQR)	6.0 (5.6-6.4)	6.0 (5.7-6.4)	.74
Prediabetes (HbA_{1c} 5.6%-6.4 %), n (%)	15 (83.3)	11 (78.6)	.73[b]
Early-stage T2DM[c] (HbA_{1c} 6.5%-7%), n (%)	3 (16.7)	3 (21.4)	—[d]
Systolic blood pressure (mm Hg), median (IQR)	120 (110-128)	126 (121-134)	.22
Diastolic blood pressure (mm Hg), median (IQR)	77 (70-81)	76 (72-79)	.97
Average daily step count during the past month, n (%)			.26
<1000	1 (5.6)	0 (0.0)	
1000-<2000	2 (11.1)	5 (35.7)	
2000-<3000	1 (5.6)	3 (21.4)	
3000 -<4000	5 (27.8)	0 (0.0)	
4000-<5000	3 (16.7)	2 (14.3)	
5000-<6000	0 (0.0)	1 (7.1)	
6000-<7000	1 (5.6)	1 (7.1)	
7000-<8000	1 (5.6)	1 (7.1)	
8000-<9000	1 (5.6)	0 (0.0)	
9000-<10000	3 (16.7)	1 (7.1)	

XSL•FO
RenderX

Variable	Digital peer-supported app group (n=18)	Control group (n=14)	*P* value
≥10000	1 (5.6)	0 (0.0)	
Unknown	2 (11.1)	5 (35.7)	
Daily step goal (steps/day), median (IQR)	7000 (5000-8000)	6000 (5000-7000)	.64

[a]HbA_{1c}: glycated hemoglobin A_{1c}.

[b]The *P* value of .73 is based on an analysis that included both participants with prediabetes and those with early-stage type 2 diabetes mellitus. Therefore, the nonsignificant group difference (*P*=.73) applies to the combined group of prediabetes and early-stage type 2 diabetes mellitus, not to each condition separately.

[c]T2DM: type 2 diabetes mellitus.

[d]Not applicable.

Figure 3 shows a comparison of the daily step achievement rates between the digital peer-supported and control groups. The median daily step achievement rate was significantly higher in the digital peer-supported app group (57.2%, IQR 32.2%-90%) than in the control group (26.7%, IQR 10%-64.4%; *P*=.04).

Figure 4 compares the average daily step counts between the digital peer-supported and control groups. The median daily step count was significantly higher in the digital peer-supported app group (6854, IQR 4846-10388 steps) than in the control (3946, IQR 3176-6832 steps; *P*=.03). A linear mixed-effects model was constructed to further evaluate step count trajectories over time (Figure 5). The group-by-time interaction was statistically significant (*P* for interaction=.03), indicating a differential trend in the step counts between groups. In the digital peer-supported app group, no significant linear trend was observed in the daily step counts across the study period (*P* for trend=.26). In contrast, the control group showed a significantly decreasing trend over time (*P* for trend<.001).

Figure 3. Comparison of daily step achievement rates between the digital peer-supported app and control groups. The median daily step achievement rate was significantly higher in the digital peer-supported app group (57.2%, IQR 32.2%-90%) than in the control group (26.7%, IQR 10%-64.4%), as examined using the Mann-Whitney U test (*P*=.04).

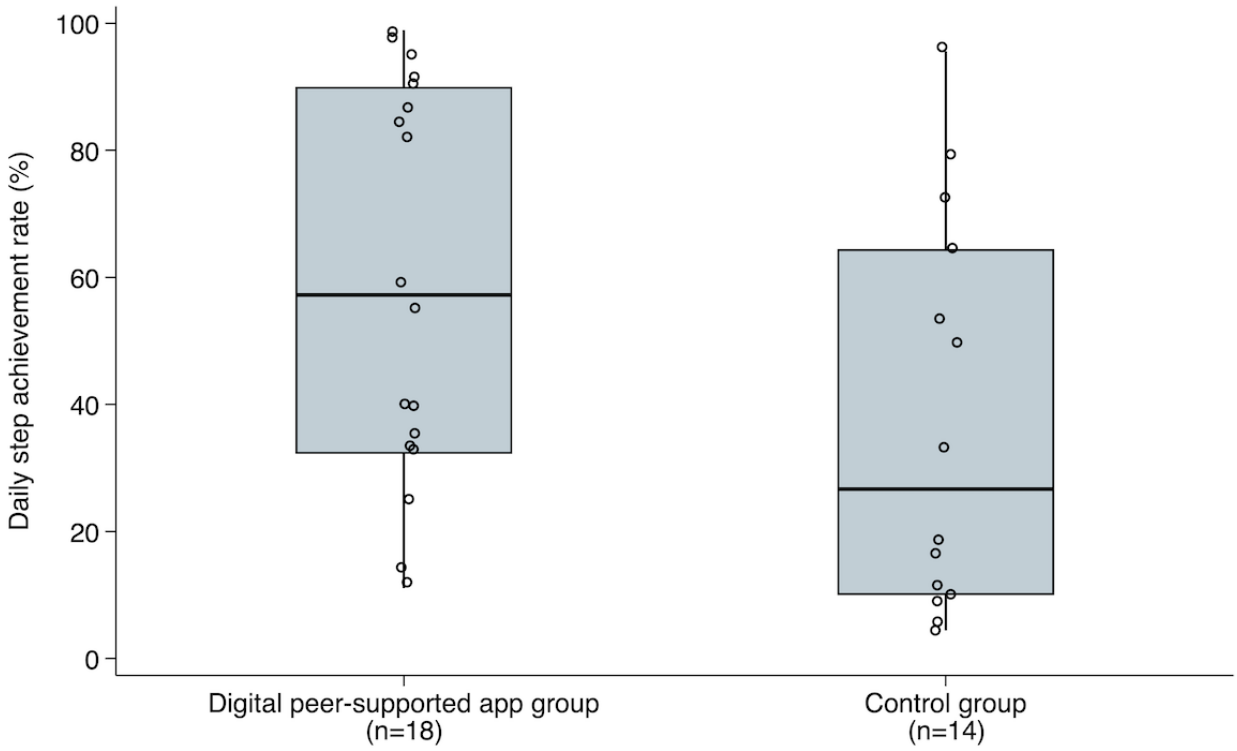

XSL•FO
RenderX

Figure 4. Comparison of average daily step counts between the digital peer-supported app and control groups. The median of the mean daily step count was significantly higher in the digital peer-supported app group (6854, IQR 4846-10388 steps) than in the control group (3946, IQR 3176-6832 steps), as determined by the Mann-Whitney U test (*P*=.03).

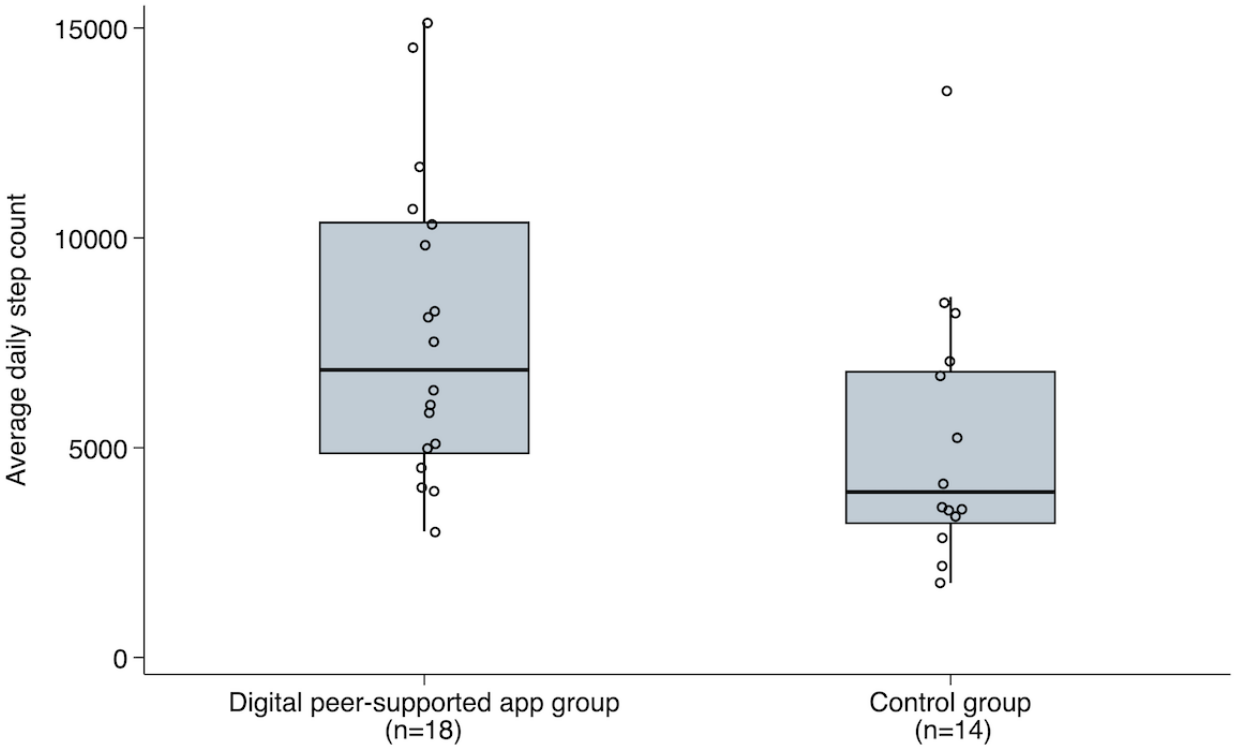

Figure 5. Group-level trends in average daily step counts over time. Each line represents the mean daily step count per day within each group.

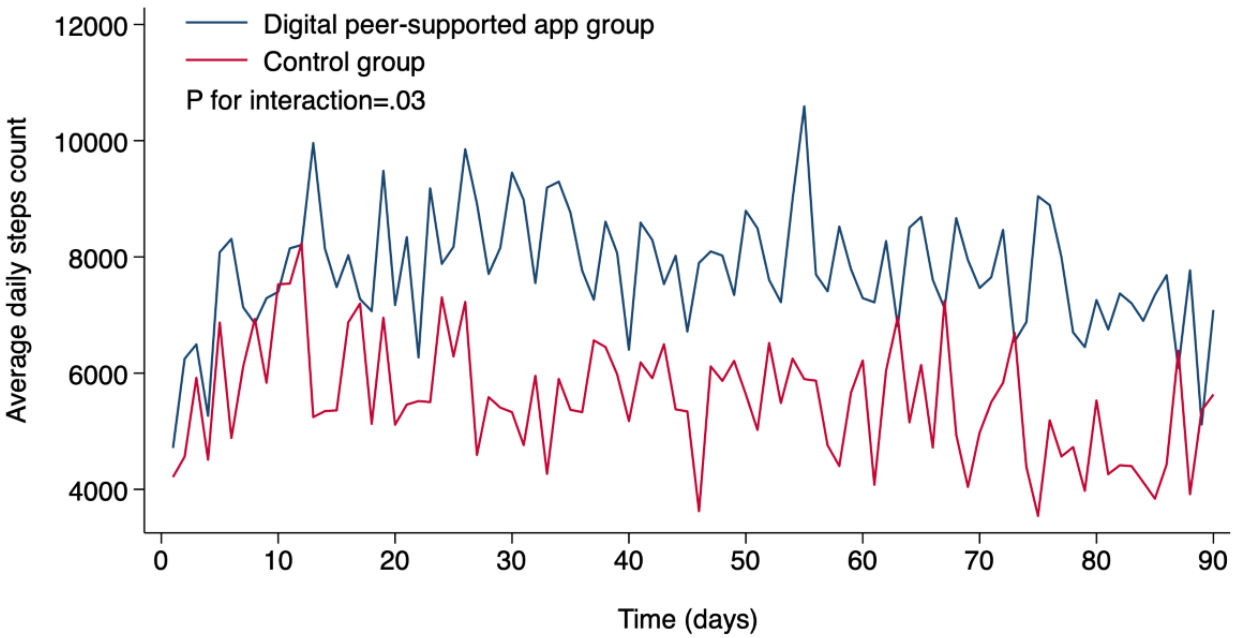

Table 2 summarizes changes in secondary health indicators and lifestyle habits in the digital peer-supported app and control groups. Although no significant between-group differences were observed for HbA_{1c} levels, blood pressure, or lifestyle habits, pre-post changes in BMI showed a trend toward improvement in the digital peer-supported app group (*P*=.07). The digital peer-supported app group showed minor improvements in perceived health status, daily staple food consumption, and exercise frequency, whereas the control group experienced slight declines in perceived health and exercise frequency. Confidence

XSL•FO
RenderX

in exercising increased in the digital peer-supported app group, but slightly decreased in the control group. Sleep duration of ≥6 hours increased slightly in the digital peer-supported app group but remained unchanged in the control group. Daily alcohol consumption decreased slightly in the digital peer-supported app group, but increased in the control group.

Table 2. Changes in anthropometric, metabolic, and lifestyle-related outcomes between the digital peer-supported app group and control group.

Variable	Digital peer-supported app group			Control group			$P_{\text{Between-group}}$
	Baseline	Follow-up	$P_{\text{Within-group}}$	Baseline	Follow-up	$P_{\text{Within-group}}$	
BMI (kg/m²)[a], median (IQR)	24.4 (22.0-26.1)	24.1 (21.7-25.9)	.07	23.4 (21.8-27.0)	23.0 (21.8-26.7)	.12	.92
HbA_{1c}[b], median (IQR)	6.1 (5.7-6.4)	6.0 (5.6-6.3)	.81	6.2 (5.7-6.6)	6.2 (5.8-6.4)	.71	.73
Systolic blood pressure (mm Hg)[c], median (IQR)	120 (110-128)	127 (120-134)	.45	126 (121-134)	124 (113-136)	.35	.32
Diastolic blood pressure (mm Hg)[c], median (IQR)	77 (68-81)	75.5 (68-84)	.75	76 (72-79)	75 (73-78)	.86	.80
Perceived good overall health status, n (%)[d]	9 (50)	10 (55.6)	.65	6 (42.9)	4 (28.6)	.22	.10
Every day of balanced meal consumption, n (%)[d]	9 (50)	10 (55.6)	.65	5 (35.7)	5 (35.7)	.22	.24
Always careful about food and healthy foods, n (%)[d]	3 (16.7)	5 (27.8)	.29	3 (21.4)	2 (14.3)	.11	.95
Physical activity at least twice a week, n (%)[d]	8 (44.4)	12 (66.7)	.33	6 (42.9)	5 (35.7)	.57	.42
Very confident in exercising, n (%)[d]	3 (16.7)	6 (33.3)	.34	4 (28.6)	3 (21.4)	.51	.17
Average sleep duration ≥6 hours, n (%)[d]	8 (44.4)	10 (55.6)	.26	6 (42.9)	6 (42.9)	.95	.22
Drink alcohol daily, n (%)[d]	5 (27.8)	3 (16.7)	.41	2 (14.3)	3 (21.4)	.67	.36
Perceived as very stressful in daily life, n (%)[d]	3 (16.7)	3 (16.7)	.51	1 (7.1)	2 (14.3)	.39	.25

[a]BMI (digital peer-supported app group: n=16 and control group: n=11).

[b]Glycosylated hemoglobin A_{1c} levels (digital peer-supported app group: n=12 and control group: n=10).

[c]Systolic and diastolic blood pressure (digital peer-supported app group: n=14 and control group: n=11).

[d]Questionnaire data for health status (digital peer-supported app group: n=18 and control group: n=14).

Discussion

Principal Findings

This study showed that users of a digital peer-supported app had significantly higher daily step goal achievements and average daily step counts compared to controls who did not receive digital peer support among individuals with prediabetes and early-stage T2DM, providing real-world evidence of the efficacy of this application. Given the limited research on individuals with prediabetes and early-stage T2DM, our findings expand the available evidence, suggesting that a digital peer-supported app may promote light physical activity, such as walking, and support its maintenance.

Our findings in individuals with prediabetes and early-stage T2DM are consistent with those of previous studies on patients with T2DM, which reported that eHealth and mobile health interventions can sustain and increase physical activity [19-24]. For example, a previous study reported that an interactive smartphone game aimed at increasing daily steps in people with T2DM resulted in a significant increase in activity levels compared to a control group that received standard lifestyle counseling [22]. In addition, another study showed that compared with the control group, the intervention group increased their daily step counts by using an app designed to increase physical activity [23].

Several features of digital peer-supported apps may explain why participants in the digital peer-supported app group were more likely to achieve their daily step goals and maintain a higher

XSL•FO
RenderX

average step count than those in the control group without the app. First, real-time feedback and peer approval (similar to receiving a "Like" on Facebook) likely played an important role. Participants could visually monitor their progress, share daily activities, and receive feedback and approval through chats from team members, which may have maintained motivation by facilitating comparison, mutual support [25,26], and a sense of accomplishment while encouraging walking behavior [27]. In addition, notifications and reminders likely helped maintain participants' motivation to engage in physical activity [25,27].

In the digital peer-supported app group, pre-post analysis showed no significant changes in HbA_{1c} or blood pressure; however, a trend toward BMI reduction was observed. A meta-analysis suggested that walking significantly decreases HbA_{1c} levels, BMI, and diastolic blood pressure in patients with T2DM [7]. The BMI reduction observed in our study may be due to the sustained increase in the daily step count over the 3 months. Finally, the absence of a significant HbA_{1c} alteration was consistent with the short study duration (3 months) and the early stage of T2DM in our cohort. As HbA_{1c} is a lagging, composite index averaging glycemia over 8-12 weeks, a single value has limited sensitivity to early metabolic dysfunction [28], thereby potentially masking subtle activity-related improvements. Future studies should incorporate more sensitive, high-frequency glycemic endpoints (eg, continuous glucose monitoring–derived time-in-range and glycemic variability) [17,29] to better detect the glycemic effects of app-induced physical activity increase.

Future studies are required to determine whether short-term or sustained use of a digital peer-supported app is more effective in supporting lasting behavioral changes and health. A digital peer-supported app may act as a short-term catalyst that initiates behavioral change or serves as a long-term support system that reinforces and sustains it. In this study, the digital peer-supported app may have triggered behavioral changes during the intervention period, potentially facilitating the development of long-lasting habits. However, these habits may not persist without continuous digital peer support.

Limitations

This study examined the potential role of digital peer-supported apps in increasing and maintaining daily step counts. However, some challenges remain unaddressed. First, methodological limitations, including a small sample size, low statistical power, and potential confounding factors such as academic background, digital literacy, and motivation, cannot be ruled out. In addition, group enrollment was nonrandomized, and participation was voluntary. A formal sample size calculation was not conducted because of a lack of previous studies on the effectiveness of digital peer support on health outcomes. An RCT with a larger sample size and a parallel control group is ongoing to address these limitations (UMIN000046892) and to improve the validity and generalizability of the results. Second, the primary outcome (individual step count goals) was determined by the participants. Accordingly, the appropriateness and validity of these goals were not objectively verified. Third, we did not collect data on medication use or adjustments. Therefore, glucose-lowering or antihypertensive therapy initiation or titration during the follow-up might have contributed to the changes observed in HbA_{1c} and blood pressure, along with other unmeasured factors. Fourth, the daily step counts recorded by smartphone applications might have been underestimated due to variations in phone placement on the body or intermittent activity tracking. Finally, generalizability remains limited. Although most control participants used a smartphone pedometer app (ie, 4 feature-phone users received stand-alone pedometers), the intervention required a smartphone and a higher level of digital literacy to engage with the social features of the app.

Conclusions

This study provides real-world evidence demonstrating that a digital peer-supported app may improve daily step goal achievement and average daily step counts in individuals with prediabetes and early-stage T2DM. The digital peer-supported app offers real-time feedback, peer approval, and continuous engagement to support sustained participation in light physical activities, such as walking. Given the limited number of studies on this population, our findings add to the evidence on the potential of digital peer support to promote and sustain light physical activity. Further studies are required to evaluate whether short-term app use is sufficient to initiate lasting behavioral changes or if long-term engagement is necessary to sustain health benefits.

Acknowledgments

This study received financial support from A10 Lab Inc, the developer and operator of the habit-formation app, through grant funding provided by the Ministry of Economy, Trade, and Industry (METI) under the FY2020 Healthcare Service Social Implementation Project Subsidy Program. Portions of this manuscript were edited for grammar and clarity using ChatGPT (OpenAI) and Grammarly (Grammarly Inc). The final content was reviewed and approved by all authors, who take full responsibility for its accuracy and integrity. The authors would also like to thank Editage for English language editing.

Funding

This study was supported by A10 Lab Inc.

XSL•FO
RenderX

Authors' Contributions

TS, MS, and MK designed the study. MS collected the data. SY and KT analyzed and interpreted the data. SY drafted the first version of the manuscript.

Conflicts of Interest

MS was an employee of A10 Lab, Inc at the time of this study. SY was an employee of A10 Lab Inc when the manuscript was written, but was not employed by the company during the study period.

Multimedia Appendix 1
Examples of user interfaces of digital peer support applications.
[PDF File (Adobe PDF File), 7812 KB - formative_v9i1e75953_app1.pdf]

References

1. Association American Diabetes. Updates to the. Diabetes Care 2018 Sep;41(9):2045-2047. [doi: 10.2337/dc18-su09] [Medline: 30135199]
2. Zheng Y, Ley SH, Hu FB. Global aetiology and epidemiology of type 2 diabetes mellitus and its complications. Nat Rev Endocrinol 2018;14(2):88-98. [doi: 10.1038/nrendo.2017.151] [Medline: 29219149]
3. Duclos M, Virally ML, Dejager S. Exercise in the management of type 2 diabetes mellitus: what are the benefits and how does it work? Phys Sportsmed 2011;39(2):98-106. [doi: 10.3810/psm.2011.05.1899] [Medline: 21673489]
4. Colberg SR, Sigal RJ, Yardley JE, Riddell MC, Dunstan DW, Dempsey PC, et al. Physical activity/exercise and diabetes: a position statement of the American Diabetes Association. Diabetes Care 2016;39(11):2065-2079 [FREE Full text] [doi: 10.2337/dc16-1728] [Medline: 27926890]
5. Asano RY, Sales MM, Browne RAV, Moraes JFVN, Coelho Júnior HJ, Moraes MR, et al. Acute effects of physical exercise in type 2 diabetes: a review. World J Diabetes 2014;5(5):659-665 [FREE Full text] [doi: 10.4239/wjd.v5.i5.659] [Medline: 25317243]
6. Morris JN, Hardman AE. Walking to health. Sports Med 1997;23(5):306-332. [doi: 10.2165/00007256-199723050-00004] [Medline: 9181668]
7. Qiu S, Cai X, Schumann U, Velders M, Sun Z, Steinacker JM. Impact of walking on glycemic control and other cardiovascular risk factors in type 2 diabetes: a meta-analysis. PLoS One 2014;9(10):e109767 [FREE Full text] [doi: 10.1371/journal.pone.0109767] [Medline: 25329391]
8. Wang Y, Li H, Yang D, Wang M, Han Y, Wang H. Effects of aerobic exercises in prediabetes patients: a systematic review and meta-analysis. Front Endocrinol (Lausanne) 2023;14:1227489 [FREE Full text] [doi: 10.3389/fendo.2023.1227489] [Medline: 37522127]
9. Ried-Larsen M, MacDonald CS, Johansen MY, Hansen KB, Christensen R, Almdal TP, et al. Why prescribe exercise as therapy in type 2 diabetes? We have a pill for that!. Diabetes Metab Res Rev 2018;34(5):e2999. [doi: 10.1002/dmrr.2999] [Medline: 29488311]
10. Wang Y, Pei L, Wang M, Sun S, Wang S, Zhang Q. Effect of peer support on adults with diabetes-related peripheral neuropathy. Patient Educ Couns 2022;105(4):828-834. [doi: 10.1016/j.pec.2021.07.047] [Medline: 34373175]
11. Sazlina SG, Browning CJ, Yasin S. Effectiveness of personalized feedback alone or combined with peer support to improve physical activity in sedentary older Malays with type 2 diabetes: a randomized controlled trial. Front Public Health 2015;3:178 [FREE Full text] [doi: 10.3389/fpubh.2015.00178] [Medline: 26217658]
12. Liu S, Bi A, Fu D, Fu H, Luo W, Ma X, et al. Effectiveness of using group visit model to support diabetes patient self-management in rural communities of Shanghai: a randomized controlled trial. BMC Public Health 2012;12:1043 [FREE Full text] [doi: 10.1186/1471-2458-12-1043] [Medline: 23198694]
13. Ricci-Cabello I, Ruiz-Pérez I, Rojas-García A, Pastor G, Rodríguez-Barranco M, Gonçalves DC. Characteristics and effectiveness of diabetes self-management educational programs targeted to racial/ethnic minority groups: a systematic review, meta-analysis and meta-regression. BMC Endocr Disord 2014;14:60 [FREE Full text] [doi: 10.1186/1472-6823-14-60] [Medline: 25037577]
14. Tabira K, Oguma Y, Yoshihara S, Shibuya M, Nakamura M, Doihara N, et al. Digital peer-supported app intervention to promote physical activity among community-dwelling older adults: nonrandomized controlled trial. JMIR Aging 2024;7:e56184 [FREE Full text] [doi: 10.2196/56184] [Medline: 38814686]
15. Okamoto M, Saito Y, Nakamura S, Nagasawa M, Shibuya M, Nagasaka G, et al. Smartphone-based digital peer support for a walking intervention among public officers in Kanagawa prefecture: single-arm pre- and postintervention evaluation. JMIR Form Res 2024;8:e53759 [FREE Full text] [doi: 10.2196/53759] [Medline: 39316793]
16. Takebayashi M, Namba M, Koyama T, Kaneda Y, Kawaguchi H, Uemura C, et al. Impact on step count by commitment-based health application. PLoS One 2024;19(8):e0305765 [FREE Full text] [doi: 10.1371/journal.pone.0305765] [Medline: 39133726]

XSL•FO
RenderX

17. Liu X, Zhang J. Continuous glucose monitoring: a transformative approach to the detection of prediabetes. J Multidiscip Healthc 2024;17:5513-5519 [FREE Full text] [doi: 10.2147/JMDH.S493128] [Medline: 39600717]
18. Kanagawa ME-BYO Living Lab. Kanagawa Prefectural Government. URL: https://www.pref.kanagawa.jp/docs/bs5/cnt/f536534/index.html [accessed 2025-03-11]
19. Bonn SE, Hummel M, Peveri G, Eke H, Alexandrou C, Bellocco R, et al. Effectiveness of a smartphone app to promote physical activity among persons with type 2 diabetes: randomized controlled trial. Interact J Med Res 2024;13:e53054 [FREE Full text] [doi: 10.2196/53054] [Medline: 38512333]
20. Thorsen IK, Yang Y, Valentiner LS, Glümer C, Karstoft K, Brønd JC, et al. The effects of a lifestyle intervention supported by the interwalk smartphone app on increasing physical activity among persons with type 2 diabetes: parallel-group, randomized trial. JMIR Mhealth Uhealth 2022;10(9):e30602 [FREE Full text] [doi: 10.2196/30602] [Medline: 36170002]
21. Poppe L, De Bourdeaudhuij I, Verloigne M, Shadid S, Van Cauwenberg J, Compernolle S, et al. Efficacy of a self-regulation-based electronic and mobile health intervention targeting an active lifestyle in adults having type 2 diabetes and in adults aged 50 years or older: two randomized controlled trials. J Med Internet Res 2019;21(8):e13363 [FREE Full text] [doi: 10.2196/13363] [Medline: 31376274]
22. Höchsmann C, Müller O, Ambühl M, Klenk C, Königstein K, Infanger D, et al. Novel smartphone game improves physical activity behavior in type 2 diabetes. Am J Prev Med 2019;57(1):41-50 [FREE Full text] [doi: 10.1016/j.amepre.2019.02.017] [Medline: 31128953]
23. Glynn LG, Hayes PS, Casey M, Glynn F, Alvarez-Iglesias A, Newell J, et al. Effectiveness of a smartphone application to promote physical activity in primary care: the SMART MOVE randomised controlled trial. Br J Gen Pract 2014;64(624):e384-e391 [FREE Full text] [doi: 10.3399/bjgp14X680461] [Medline: 24982490]
24. Rossen J, Larsson K, Hagströmer M, Yngve A, Brismar K, Ainsworth B, et al. Effects of a three-armed randomised controlled trial using self-monitoring of daily steps with and without counselling in prediabetes and type 2 diabetes-the Sophia Step Study. Int J Behav Nutr Phys Act 2021;18(1):121 [FREE Full text] [doi: 10.1186/s12966-021-01193-w] [Medline: 34496859]
25. Emberson MA, Lalande A, Wang D, McDonough DJ, Liu W, Gao Z. Effectiveness of smartphone-based physical activity interventions on individuals' health outcomes: a systematic review. Biomed Res Int 2021;2021:6296896 [FREE Full text] [doi: 10.1155/2021/6296896] [Medline: 34409104]
26. Likhitweerawong N, Boonchooduang N, Kittisakmontri K, Chonchaiya W, Louthrenoo O. Effectiveness of mobile application on changing weight, healthy eating habits, and quality of life in children and adolescents with obesity: a randomized controlled trial. BMC Pediatr 2021;21(1):499 [FREE Full text] [doi: 10.1186/s12887-021-02980-x] [Medline: 34758761]
27. Schoeppe S, Alley S, Van Lippevelde W, Bray NA, Williams SL, Duncan MJ, et al. Efficacy of interventions that use apps to improve diet, physical activity and sedentary behaviour: a systematic review. Int J Behav Nutr Phys Act 2016;13(1):127 [FREE Full text] [doi: 10.1186/s12966-016-0454-y] [Medline: 27927218]
28. Zhang J, Zhang Z, Zhang K, Ge X, Sun R, Zhai X. Early detection of type 2 diabetes risk: limitations of current diagnostic criteria. Front Endocrinol (Lausanne) 2023;14:1260623 [FREE Full text] [doi: 10.3389/fendo.2023.1260623] [Medline: 38027114]
29. Battelino T, Alexander CM, Amiel SA, Arreaza-Rubin G, Beck RW, Bergenstal RM, et al. Continuous glucose monitoring and metrics for clinical trials: an international consensus statement. Lancet Diabetes Endocrinol 2023;11(1):42-57. [doi: 10.1016/S2213-8587(22)00319-9] [Medline: 36493795]

Abbreviations

HbA1c: glycosylated hemoglobin A1c
RCT: randomized controlled trial
T2DM: type 2 diabetes mellitus
UMIN: University Hospital Medical Information Network

Edited by A Stone; submitted 18.04.25; peer-reviewed by C Baxter, J Zhang; comments to author 07.11.25; revised version received 19.11.25; accepted 24.11.25; published 15.12.25.

Please cite as:
Yoshihara S, Shibata T, Kosone M, Shibuya M, Takahashi K
The Impact of a Digital Peer-Supported App on Daily Steps and Lifestyle Changes Among Individuals With Prediabetes and Early-Stage Type 2 Diabetes: Prospective, Nonrandomized Controlled Trial
JMIR Form Res 2025;9:e75953
URL: https://formative.jmir.org/2025/1/e75953
doi:10.2196/75953
PMID:

XSL•FO
RenderX

©Shota Yoshihara, Takeo Shibata, Mizuki Kosone, Megumi Shibuya, Kayoko Takahashi. Originally published in JMIR Formative Research (https://formative.jmir.org), 15.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

ROOM to Grow, a Mobile Well-Being Intervention for University Students: Overview of the Design Process and Outcomes

Tajda Laure[1], PhD; Camila Villegas Meija[1], PhD; Danielle Remmerswaal[1], PhD; Djameela Dulloo[2], MSc; Ruth Van der Hallen[1], PhD; Birgit Mayer[1], PhD; Marianne Littel[1], PhD; Bram Dierckx[3], MD, PhD; Jeroen S Legerstee[3,4,5], PhD; Rutger C M E Engels[1], PhD; Marilisa Boffo[1], PhD

[1]Department of Psychology, Education, and Child Studies, Erasmus School of Social and Behavioural Sciences, Erasmus University Rotterdam, Rotterdam, The Netherlands

[2]Erasmus University Rotterdam, Rotterdam, The Netherlands

[3]Department of Child and Adolescent Psychiatry/Psychology, Erasmus MC - Sophia Children's Hospital, Rotterdam, South Holland, The Netherlands

[4]Academic Center for Child and Adolescent Psychiatry and Specialized Youth Care, Levvel, Amsterdam, North Holland, The Netherlands

[5]University of Amsterdam, Research Institute of Child Development and Education, Amsterdam, The Netherlands

Corresponding Author:
Tajda Laure, PhD
Department of Psychology, Education, and Child Studies
Erasmus School of Social and Behavioural Sciences
Erasmus University Rotterdam
Burgemeester Oudlaan 50
Rotterdam, 3062 PA
The Netherlands
Phone: 31 638436367
Email: laure@essb.eur.nl

Abstract

Background: University students are facing a multitude of challenges and an increase in mental health issues that affect their academic performance and overall well-being. In response, Erasmus University Rotterdam launched the Student Wellbeing Programme in 2019, offering comprehensive, tailored support through a stepped-care framework to enhance student success and well-being. One of the tools developed for students is ROOM to Grow, an anonymous and accessible preventative mental health app.

Objective: This paper describes the process and outcomes of ROOM's design and development, guided by the Centre for eHealth Research (CeHRes) road map and privacy-by-design principles, and highlights the lessons learned throughout this process.

Methods: This paper describes the first 4 phases of the CeHRes Road map: contextual inquiry, value specification, design, and operationalization. It outlines the population (ie, stakeholders), methods (ie, literature reviews, expert groups, cognitive walkthroughs, interviews, experimental designs), and outcomes of each phase.

Results: The most common mental health struggles among our target population were stress, anxiety symptoms, perfectionistic tendencies, and loneliness. Students often recognized these issues only once they became overwhelming. Regarding digital tools, students seek credible content specific to their experiences, as well as adaptable and intuitive systems; they are mindful of data privacy and aim to reduce their screen time. ROOM was developed to address the diverse needs and preferences of university students through a transdiagnostic approach to mental health. It targets emotion regulation (ER) skills and self-awareness, which underlie the mental health challenges experienced by our target users. ROOM comprises 26 brief exercises (ie, micro-interventions) that support the development and use of adaptive ER. The exercises incorporate techniques from various therapeutic approaches (ie, self-compassion, positive psychology, mindfulness, cognitive behavioral therapy, and acceptance and commitment therapy) to accommodate students' diverse content preferences. To help students recognize their struggles earlier, ROOM includes a mood tracker and a self-assessment module with questionnaires that evaluate both traits (eg, perfectionistic tendencies) and states (eg, stress levels), providing personalized feedback. ROOM further implements an intelligent recommender system that connects users to relevant content, enhancing the tool's personalization and responsiveness to users' needs. As students aim to minimize screen time, ROOM's goal is not prolonged app use but the application of ER skills in real life, supported by features that facilitate

XSL•FO
RenderX

skill transfer into everyday settings. Finally, ROOM was developed within a "privacy-by-design" framework to address students' privacy concerns, implementing strict privacy and security regulatory standards.

Conclusions: Compromises were necessary to balance user needs, resource constraints, and privacy-by-design and security standards, often limiting ROOM's interactivity. Other challenges included simplifying complex psychological concepts into brief formats, fostering interdisciplinary collaboration, balancing academic rigor with industry production pace, and operating with fixed resources while maintaining an iterative process. This paper may serve as a primer for designing transdiagnostic, adaptive mental health interventions for youth, blending therapeutic approaches and promoting skill transfer.

(JMIR Form Res 2025;9:e63325) doi:10.2196/63325

KEYWORDS

mobile intervention; mHealth; micro-intervention; mental health; emotion regulation, self-awareness; transdiagnostic; skill transfer; student well-being; university students

Introduction

Background

In the last decade, universities have been compelled to evolve beyond their traditional roles. Historically centered on imparting knowledge and academic skills, universities are now taking on greater responsibilities to support the holistic development of their students [1]. This shift is driven by growing recognition of the multifaceted challenges university students face today. There is greater emphasis on the development of soft and personal skills that are crucial for participation in the modern workforce and integration into today's society. This expanded educational mission reflects the need to move beyond merely study success—often gauged by grade point average, dropout rates, and study career progression—to student success, which encompasses personal (eg, development of a value system to inform future decisions, self-efficacy) and professional (eg, effective communication, problem-solving) growth in addition to academic achievements [2,3].

The necessity for this shift has partly arisen from increasing reports of mental health problems among university students, coupled with a noticeable shortage of available resources [4,5]. Seminal reports indicate that one-third of university students globally experience mental health problems such as depression, anxiety, and substance abuse, yet only 1 in 5 receive adequate help [4,5]. Mental health problems not only affect the individuals themselves but also have a detrimental impact on society. In particular, they are directly linked to poorer academic performance and higher dropout rates [6,7], negatively influencing individual well-being [8] as well as career and financial prospects [9]. On a broader scale, when students falter, society loses out on individuals equipped with essential skills vital for economic growth and innovation, posing challenges in meeting the demand for highly skilled workers required for sustainable development and maintaining competitiveness in the global market [10].

What Does It Mean to Be a Student Nowadays?

The current generation of students faces pressures from global concerns such as climate change [11], COVID-19, housing crises [12], and geopolitical tensions, alongside academic pressures [13], adding layers of complexity to their university experience and overall development. Most university students fall within a developmental phase known as emerging adulthood [14], during which individuals transition from adolescence to full adulthood. Emerging adulthood typically spans from the late teens to the mid- to late twenties and is characterized by identity exploration, instability, self-focus, a sense of in-betweenness, and a feeling of possibilities [14,15]. This period often involves frequent relocations and significant changes in social and personal identities. For university students, the typical milestones of this life stage—such as moving away from family and gaining greater independence—are further compounded by the added pressures of academic demands and the potential accumulation of student debt [16]. Although the transition to university presents an opportunity for positive personal development, the inherent instability of this period, combined with academic and other pressures, may create conditions that contribute to the development of mental health issues. A shift toward a more holistic approach to student success within academia aligns well with the priorities of Generation Z (1997-2012), who comprise the majority of today's university students. Generation Z tends to be less focused on traditional career paths and more attuned to the broader implications of their roles in society [3]. Considering the present-day context, universities need to expand their focus beyond academic success and career progression to also foster self-exploration and personal growth among students [3]. By doing so, they can cultivate well-rounded individuals who are better prepared to navigate both the workforce and life's challenges.

Erasmus University Rotterdam: Student Wellbeing Programme

In 2019, Erasmus University Rotterdam (EUR), located in Rotterdam, The Netherlands, launched the Student Wellbeing Programme [17], a university-wide initiative designed to foster student success by supporting their well-being and personal development, grounded in a whole-system perspective. This multifaceted, research-driven programme considers contextual, organizational, and individual factors that are crucial for student success. It systematically addresses student well-being through tailored support services that account for the diverse needs and characteristics of its student body, while also facilitating systemic changes to the university's role in students' lives. The programme integrates services, events, initiatives, and policies within the university's educational and professional services infrastructure (ie, a so-called chain of care). Such a comprehensive, grassroots approach can better reach students

XSL•FO
RenderX

and provide them with effective, holistic support [18]. Central to this initiative is a stepped-care framework [19,20]—a structured system of support services and interventions tailored to the severity and nature of students' needs. For example, the programme offers universal prevention and self-help measures, skill-building workshops, mental health awareness events, and spaces designed to facilitate social connection. One of the interventions at this level is *ROOM*, a mobile app designed to provide tools that foster mental health and self-growth, empowering students to manage their well-being proactively. For those requiring more intensive support, the stepped-care model extends to increasingly targeted and specialized services, such as academic and well-being coaching and access to university psychologist services.

The Case for Building Yet Another Mobile App

eHealth solutions provide accessible, scalable, anonymous, and adaptable support, making them a promising avenue for enhancing the delivery of mental health services in both preventive and therapeutic settings [21]. Literature suggests that online interventions can effectively improve emotional health and reduce symptoms of anxiety and depression among university students [22-25]. While today's students are more open to discussing mental health compared with previous generations, many hesitate to seek help due to stigma, time and financial limitations, and a preference for self-managing their mental health challenges [24,26]. At the same time, this generation engages with digital tools such as mobile phones more than any other [27,28]. Providing students with a mobile mental health tool (ie, ROOM) represents a viable approach to addressing some of the barriers to seeking psychological help (as suggested by Andersson and Titov [29]; Ebert et al [30]). Moreover, a tool implemented within the university ecosystem serves as an accessible, low-threshold entry point to the stepped-care model established by the EUR Student Wellbeing Programme.

Given the increasing popularity of digital health solutions, the current landscape of mobile mental health offers a wide range of off-the-shelf solutions for users. These apps hold considerable promise, as they can be cost-effective and provide timely support and privacy to their users [31]. Despite their popularity, however, these tools vary in quality and frequently exhibit several limitations [31,32]. For instance, while some apps incorporate evidence-based therapeutic techniques such as cognitive behavioral therapy (CBT) or mindfulness practices, only a very small fraction have been empirically tested and validated [33]. Of the approximately 15,000 commercial mental health apps available on the market, only 3%-4% are evidence based [34]. Research on mental health apps in the university context further indicates that only a few commercially available apps meet the specific needs of university students [35-37]. University students prefer digital tools that are easy to use, integrate credible and high-quality educational content, allow customization (such as notification settings), offer personalization (ie, responsiveness to their needs), are cost-free or include free trial periods, and implement robust security measures to protect their data privacy [35,36,38]. However, these features are often lacking in currently available apps. More rigorous research and development are needed to create effective online interventions for this target population, with an emphasis on user experience (UX) and engagement to ensure sustainable implementation within universities—an aspect often overlooked [23-25].

Recognizing (1) the potential of digital tools and the inherent familiarity of university students with mobile devices, (2) students' behavioral patterns and preferences regarding help-seeking and mental health management, and (3) the current state of existing digital mental health solutions, we sought to leverage mobile technology to develop a service that is accessible, private, evidence based, and scientifically validated. This service is aligned with students' needs and designed to integrate within the broader university support ecosystem, helping students cultivate the skills necessary to better manage their mental health.

Our Approach to Mobile Intervention Development

To develop this mobile intervention, we formed a multidisciplinary team that brought together diverse knowledge bases and areas of expertise. The team included clinical and health psychologists, behavioral researchers, visual designers, game and interaction designers, behavior change intervention designers, human-computer interaction specialists, information technology developers, research methodologists, machine learning engineers, and data scientists. Master's students enrolled in clinical, health, and positive psychology programs also joined the team to conduct research and assist in developing intervention materials.

In designing our mobile intervention, we adhered to the Centre for eHealth Research (CeHRes) road map, a comprehensive framework for developing eHealth solutions [39,40]. This approach is grounded in human-centered design principles, persuasive systems design, business modeling, and empirical evidence [39,41]. It departs from traditional linear approaches to designing digital interventions and products (eg, waterfall development methods) and instead proposes a progressive, cyclical design process involving iterative assessment and design cycles across 5 development phases, aligning more closely with Agile development methods: (1) contextual inquiry (ie, assessing the problem, context, barriers and opportunities, and stakeholder roles), (2) value specification (ie, establishing guidelines to align the development process with stakeholder needs), (3) design (ie, developing and refining prototypes through iterative testing to ensure relevance, acceptability, and usability among stakeholders), (4) operationalization (ie, finalizing and implementing the eHealth technology), and (5) summative evaluation (ie, evaluating the impact and uptake of the eHealth technology; see Multimedia Appendix 1 for more details [40]). Within this framework, users and other stakeholders are actively involved as co-designers of the end solution, which evolves from a conceptual prototype to an increasingly functional and complex system. Decisions are research driven, evaluated, and cyclically improved until the product is finalized.

Objectives

The goal of this paper is 2-fold: to present the developed solution and to share our journey and lessons learned from creating a mobile mental health app for university students. We aim to

XSL•FO
RenderX

provide intervention developers working with this target population with actionable insights to create relevant and engaging mental health tools, as well as to serve as a primer on developing adaptive digital interventions.

The focus of this paper is on the first 4 phases of the CeHRes road map: contextual inquiry, value specification, design, and operationalization (as depicted in Figure 1), excluding the summative evaluation phase, which is reported elsewhere [42]. Although the intervention development process was iterative, we present the information linearly, aligned with the road map phases, for clarity. In the following sections, we outline the methods and main outcomes for the first 3 phases. The paper concludes with a description of the developed product, ROOM, and its integration within the EUR ecosystem, including an overview of our implementation strategy (ie, the operationalization phase). Finally, the discussion synthesizes the main insights and lessons learned from the intervention development process.

Figure 1. Centre for eHealth Research (CeHRes) road map approach to ROOM's (the intervention) development.

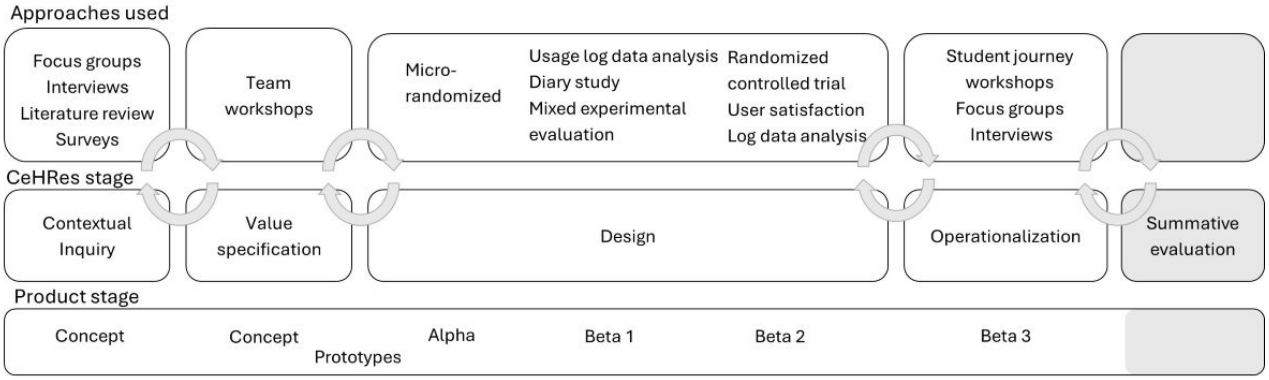

Contextual Inquiry

Methods

The goal of this phase was to develop an understanding of university students, their context, and their challenges; to identify other relevant stakeholders; and to determine their needs to ensure seamless integration of the final product within the broader system in which the tool will be used. In this phase, we addressed the following research questions:

- What personal or academic challenges do students at EUR experience? How and where do they seek support, and what could be improved in terms of available support at the university?
- What is the prevalence of mental health problems among EUR students?
- How do students relate to technology and well-being, and how do they spend their time?
- What aspects of digital media and technology do students value most?
- Which mental health approaches would be most helpful for students in managing their mental health during their time at university?

Participants in the contextual inquiry phase included EUR students (aged 18-30); staff members who were in direct contact with students and aware of their struggles (eg, university psychologists, study advisors, student counselors); and researchers and mental health professionals working with youth and experienced in mental health approaches.

To address the research questions, we used a mixed-methods approach that combined qualitative and quantitative techniques, including focus groups, semistructured interviews, large-scale surveys, and comprehensive literature and desk research. Data analysis methods were tailored to each methodology and research question, employing techniques such as descriptive statistics, affinity mapping, and summary reporting. Key outcomes were identified through triangulation of insights across all research activities.

Results

Overview

A total of 3827 students participated in the contextual inquiry phase (female: n=2564, 67%; male: n=1231, 32.17%; nonbinary or preferred not to disclose: n=21, 0.55%; and gender not disclosed: n=11, 0.29%), along with 17 university staff in student-facing roles and 8 experts in mental health and technology (see Table 1). Further research details and results from the contextual inquiry phase are available in the Open Science Framework (OSF) environment [43].

XSL•FO
RenderX

Table 1. Overview of research goals, evaluation methods, participant information, and analysis.

Research question and method	Participant group	Number	Key details	Analysis
Struggles and support				
7 focus groups	Students	33	21 (64%) female and 12 (36%) male	Descriptive statistics and the participants' input were summarized and documented in subsequent reports.
3 focus groups	University staff	17	13 study advisors, 3 student psychologists, and 1 student counselor	Descriptive statistics
Mental and physical healthnology, well-being, activities				
Large-scale survey [44]	Students	3769	2536 female (67.29%), 1212 male (32.16%), and 21 nonbinary (0.56%)	Descriptive statistics
Technology, well-being, and activities				
Semistructured interviews	Students	11	No demographic details provided	Affinity mapping
Semistructured interviews	Students	9	4 (44%) female, 5 (56%) male	Affinity mapping
A focus group	Students	5	3 (60%) female, 2 (40%) male	The participants' input was summarized and documented in subsequent reports.
Therapeutic approaches				
2 focus groups	Experts	8	1 psychiatrist, 1 health care psychologist, 2 clinical psychology researchers/mindfulness practitioners, 1 clinical psychology researcher/therapist, and 2 mental health researchers	The participants' input was summarized and documented in subsequent reports.

Outcomes of the Contextual Inquiry

Psychological and Situational Challenges Among University Students

Results from the focus groups with students (n=33) and university staff in student-facing roles (n=17), as well as from the large-scale survey [44] (n=3769), indicated that students face significant internal and external challenges. Psychologically, they commonly struggle with anxiety, depression, stress, loneliness, and the pressure to perform and achieve perfection, often compounded by inadequate study skills. Typically, students recognize these issues only once they become overwhelming, suggesting limited mental health literacy and self-awareness. Many also report poor service accessibility, such as long waiting times, insufficient information, and confusion about available resources. At the same time, situational difficulties—such as financial constraints and high academic expectations—further contribute to their stress.

University Students' Perspectives on Technology, Digital Media, and Well-Being

Research exploring how students relate to technology and well-being, their leisure activities, and what they value in digital media (including interviews with 20 students and a focus group with 5) yielded 3 key insights: students' experimental preferences, engagement dilemmas, and implementation considerations.

Key Experiential Preferences

Students expressed distinct preferences regarding digital tools. First, they emphasized the need for credible, trustworthy content that resonates with their personal experiences. Second, they placed strong importance on privacy and data security, expecting digital tools to be intuitive, adaptable to their needs, require minimal user input, and use relatable, nonpatronizing language. They expressed a clear dislike for social comparisons (ie, comparing their progress or achievements with those of their peers) and valued their time, preferring activities that yield clear, quick, and tangible results. Finally, this group often seeks to reduce screen time due to screen-related stress, demonstrating a preference for minimizing online interactions [45].

The Engagement Dilemma

We found that users were saturated with online content and actively attempted to limit their phone usage [43]. However, digital behavior change interventions (DBCIs) require sufficient user engagement to be effective [46]. This highlighted conflicting requirements that needed to be addressed: *How can a DBCI engage users without relying on sustained app usage?* Engagement was therefore redefined beyond mere app interaction, focusing instead on the active application of skills acquired from the app in real-world contexts, guided by the concept of *learning transfer* [42,43].

Implementation Considerations

Findings from the Contextual Inquiry phase revealed that students needs and receptiveness to well-being support or technology vary across individuals, phases of the academic year, and years of study. Key periods—such as the start of the academic year, examination periods, and seasonal changes—shape these fluctuations, with notable differences observed between students in different years of study and

XSL•FO
RenderX

between national and international students. While some students proactively manage their stress, others are more reactive, seeking support primarily during moments of crisis. Students prefer light, low-effort interactions on mobile devices, reserving longer tasks (eg, journaling) for computers, and are most likely to engage with phones in the mornings, evenings, and during commutes. They often hesitate to seek formal support, typically turning first to friends or family. Additionally, some students hold negative perceptions of the university, viewing it as a culture of overwork with limited care. These varying attitudes and levels of receptiveness may directly influence the app's acceptability and uptake, highlighting the importance of accounting for these factors when designing the app's content and communication strategy.

Identifying a Suitable Mental Health Approach to Support University Students: Transdiagnostic Approach

First, a pool of potential intervention topics was created, including stress management, emotion regulation (ER), social skills, goal setting, and self-awareness. Next, by triangulating data gathered from the literature review and expert focus groups (n=8), candidates for therapeutic approaches were identified. These included positive psychology [47], mindfulness [48], acceptance and commitment therapy (ACT) [49], self-compassion [50], and CBT [51]. The intervention approaches were tiered from less to more intensive based on problem severity—starting with *positive psychology* for simple mood improvements, progressing to *self-compassion and ACT* for mild to moderate mental health problems, and escalating to *CBT* for more severe issues (see Villegas et al [43]). Initially, the idea was to target specific problems such as anxiety and depressive symptoms using distinct therapeutic approaches. However, to accommodate such a heterogeneous target population, a more parsimonious approach was needed to make the app more flexible and approachable. Therefore, we diverged from traditional problem-specific methods and adopted a transdiagnostic approach, targeting common factors underlying the various mental health challenges experienced by university students. Emotion (dys)regulation was identified as a suitable target, as it emerges as a key factor across a wide spectrum of mental health problems [52]. ER is a complex process involving the initiation, inhibition, or modification of one's emotional state or behavior in specific contexts [53]. It encompasses various strategies, including choosing or altering the situations one engages in, redirecting attention, reappraising situations, and adjusting one's emotional or behavioral responses accordingly. Fostering ER skills can lead to improved well-being, interpersonal relationships, and resilience [18].

Value Specification

Methods

During the value specification phase, our objective was to identify aspects of the tool that would add value for end users and translate these into specific system requirements. Using the value specification approach outlined in the CeHRes road map [54], we identified recurring themes from the contextual inquiry results and extracted items that would provide value to key stakeholders (ie, students). These were then converted into actionable requirements covering the following: *content* (ie, what the intervention should provide); *UX* (ie, interaction with the content, intervention look and feel); *technical aspects* (ie, app functionalities and back-end system); and *implementation and organizational aspects* (ie, integration of the tool within the university ecosystem and compliance with operational, legal, privacy, and other regulatory frameworks) [54]. These requirements were prioritized based on their relevance to users and other stakeholders (eg, university regulations), as well as financial and technical limitations. A final list of requirements was then created to guide subsequent development. The identified values and requirements were continuously refined in the following phase as we designed, tested, and learned more about our users and available resources.

Results

The value specification process resulted in themes that were translated into 36 values, which were then operationalized into 175 requirements. A full overview of these requirements is available in the OSF [43]. To maintain conciseness, Table 2 presents an excerpt of the most important themes, subthemes, and related values and requirements that informed ROOM's creation and its implementation within the broader university ecosystem.

Based on the value and requirements specification, the tool was designed to focus on 2 key areas: improving *self-awareness*—to help students recognize when and why they are struggling (and, when necessary, prompt them to seek support)—and *enhancing ER*. The emphasis was on concise content and a design that is responsive to users' needs while offering the freedom to engage through different modalities. In terms of implementation, the focus was on maintaining strong security and privacy standards, creating a tool that is visually and functionally distinct from EUR branding and educational platforms to build student trust and separate academic matters from personal mental health, and developing a tailored communication strategy that leverages student ambassadors and aligns with the student journey to ensure accessibility throughout the academic year.

XSL•FO
RenderX

Table 2. Summary of key themes, subthemes, and values from the contextual inquiry, mapped to content, UX[a], and technical and organizational requirements informing intervention design.

Themes	Subtheme	Value to the user and stakeholders	Requirements
Common mental health struggles among EUR[b] students	• Anxiety/rumination/worry, comparison pressure, perfectionism, and fear of failure	• Getting techniques to deal with negative thoughts and feelings that might be maladaptive	• Content: The app contains exercises that train adaptive emotional regulation skills.
Lack of self-awareness	• Students do not know when to get help or what is normal stress, anxiety, and rumination	• Having the space and tools to be more self-aware	• Content: The app allows students to engage in assessments to understand themselves better, including their traits, states, and psychological mechanisms.
Time sensitivity	• Getting something in return for invested time, fast, "time is precious," reducing phone time.	• Learning in easy, quick spurts of time	• UX: Interactions and exercises should be short or be segmented into smaller chunks.
Personalized content	• Like reduced and curated content, dislike generic nonapplicable content.	• Getting content that feels relevant and personal.	• UX: The app provides tailored recommendations of activities relevant to the user's intentions.
Tone and content preference	• Undesirable—"cringe" or over-the-top content; dislike condescending tones; like well-researched, evidence-based information; need for validation and compassion; sensitive to pressure/judgment; empathetic vibes; calming music/audio guides when stressed.	• Down-to-earth yet science-based advice to deal with emotions. • Feeling cared for and not pressured.	• UX: Content is candid and to the point. The voice is not rigid or too silly. • Content: Scientific sources are included in the app. • Content and UX: Content acknowledges societal and academic pressures and provides feedback that feels compassionate and validating. • Content: Content emphasizes the humanness of the user (and avoids overtly quantifying the user).
Relationship with screen time	• Dislike the addictiveness of social media; dislike notifications; technostress—tired/saturated.	• Having the option to follow exercises without needing to access the app.	• Content and UX: The app contains content and interactions that support users in practicing, applying, and remembering skills offscreen [42,43].
Visual preferences	• Aesthetics are important; "Good Vibes"	• Learning from a tool that is aesthetically pleasing and calming.	• UX: The aesthetic is "laid back"—resembles chill-hop and other study/relax YouTube channels familiar to the users.
Privacy considerations	• Mistrust technology that feels "creepy" or "knows too much"; international and university standards for privacy and security.	• A system that prioritizes privacy by design.	• Technical: User log data are stored for research or system monitoring purposes only and is done so anonymously. Full usage and users' data are stored locally on the user's phone (ie, use of a privacy-by-design development approach). • Organizational: The system must adhere to and pass review for relevant privacy and security standards.
Peer endorsement versus university skepticism	• Students value and trust advice from other students; the university is seen as a source of stress and is distrusted	• Introducing the tool by someone students trust • Separation of the well-being tool from the university branding • Separation of mental health and personal development from academic performance and studying	• Organizational: Student ambassadors promote the tool • UX: The tool has a distinct visual style that is separate from the university branding • Organizational: The tool is disjointed from the university's branding and educational support environments.

XSL•FO
RenderX

Themes	Subtheme	Value to the user and stakeholders	Requirements
Attitudes to well-being	• Interest in personal development and preventive approaches to well-being; attention to well-being only in reaction to distress	• Being aware of the tool's availability and being able to easily access it when the student's interest/need arises • The communication about the tool is tailored to the receptivity of different student groups • Universal and indicated prevention	• Content and organizational: The tool and related marketing strategy integrate positive mental health framing and avoid use of language implying pathology (eg, instead of "distorted thoughts," the term "unhelpful thoughts" is used). • Content: The tool provides resources for seeking specialized help when necessary, directing students to the support available through the university's stepped care model. • Organizational: The tool is introduced to students as they are starting university. • Organizational: Students are referred to the tool while waiting for an appointment with university psychologists.

[a]UX: user experience.

[b]EUR: Erasmus University Rotterdam.

Design

Methods

Following the value specification requirements, we developed, evaluated, and iterated multiple versions of ROOM at different fidelity levels. The development process comprised 4 main stages: (1) early content and concept prototypes; (2) high-fidelity clickable prototypes; and 2 fully functional app versions—(3) alpha and (4) beta (ie, fully coded mobile apps; see Figure 2).

In the initial stages of prototype development, the focus was primarily on content. During this phase, formative evaluations were conducted by mental health experts to assess the appropriateness, adequacy, and accuracy of the materials. Concurrently, university students reviewed the content for clarity and acceptability. Observational pilot studies explored whether engagement with the intervention content produced changes in targeted skills or moods, aiming to assess improvements in cognitive defusion, cognitive restructuring, self-compassion, mindfulness, and emotional states.

Clickable prototypes were used to assess the acceptability and appeal of the app's concept and content among students. These evaluations also examined the app's visual and interaction design to ensure alignment with user expectations and to enhance the overall UX.

Figure 2. Excerpts from ROOM's versions from concept to final product: the raw content, an early prototype of the content, an early prototype of interactions, and the latest version (beta I).

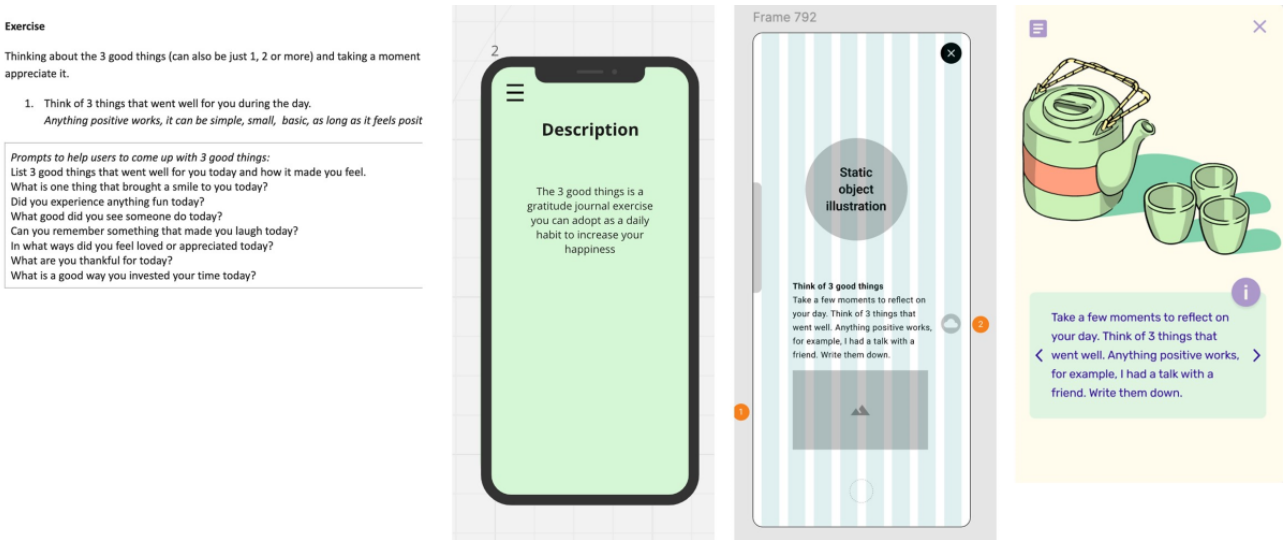

The alpha version was developed to experimentally evaluate the effects of the exercises on target outcomes—specifically, users' emotional states immediately after completing an exercise (proximal effects)—and to explore more distal effects on distress and ER skills over time [55,56]. The beta version was a more comprehensive iteration of the app, incorporating all key components of the final architecture (ie, exercises, self-assessment module, recommender system). It was designed to test features supporting the application of learned techniques in users' daily lives (ie, learning transfer elements) [43].

Evaluations were conducted with varying levels of rigor—ranging from rapid iteration tests to longer experimental studies. The methods employed at this stage included both qualitative (ie, semistructured interviews, focus groups, design sessions, diary studies) and quantitative approaches (ie,

XSL•FO
RenderX

micro-randomized trials and randomized controlled trial designs [A/B testing]).

Results

Overview

In the design phase, a total of 692 EUR students participated, of whom 506 (73.1%) were female, 168 (24.3%) male, and 18 (2.6%) nonbinary or preferred not to disclose. This phase also involved 30 experts, including psychiatrists, psychologists, researcher-practitioners, and digital health specialists. An overview of the research is provided in Table 3. Detailed research methods and results from the design phase are available in the OSF environment [43].

Table 3. Overview of key research phases in the design stage: goals, methods, participant information, and analysis.

Fidelity stage, research goal, and method		Participants	Data analysis
Early concept (low-fidelity)			
Expert evaluation of intervention content			
	4 expert focus groups	1 psychiatrist, 1 health care psychologist, 3 researcher-practitioners, and 2 digital mental health researchers	Input was summarized and documented in subsequent reports
	Semistructured interviews	4 psychologists	Thematic analysis
Evaluating the efficacy of content on key targets			
	6 quasi-experimental pre-post studies (mixed methods)	192 EUR[a] students: 139 (72.4%) female, 48 (25.0%) male, and 5 (2.6%) nonbinary	Paired *t* test or Wilcoxon signed rank test
Qualitative evaluation of content acceptability and clarity			
	3 think-aloud sessions and 6 semistructured interviews	85 EUR students: 55 (65%) female, 26 (31%) male, and 4 (5%) nonbinary	Thematic analysis
	Informal user experience interviews	Informal check-in (sample not detailed)	Input was summarized and documented in subsequent reports
General user engagement			
Explore onscreen and offscreen user engagement			
	Semistructured interviews	30 in total: 10 mental health professionals, 9 digital experts, and 11 DBCI[b] users	Grounded theory analysis
High-fidelity prototype			
Review acceptability of content, design, and interactions	Think-aloud sessions + user experience interviews (clickable prototype)	15 EUR students: 9 (60%) female, 4 (27%) male, and 2 (13%) nonbinary	Input was summarized and documented in subsequent reports
Assess likelihood of usage and skill transfer [43]	5 co-design workshops (high-fidelity prototype)	10 EUR students: 9 (90%) female and 1 (10%) male	Input was summarized and documented in subsequent reports
Alpha			
Evaluate effects on emotional states (micro-randomized trial)	Micro-randomized trial + semistructured interviews	146 EUR students: 119 (81.5%) female, 24 (16.4%) male, and 3 (2.1%) nonbinary; a subsample of 18 in interviews: 11 female and 7 male	Weighted and centered least-squares analysis, paired *t* test, and Wilcoxon signed rank test (see Laure et al [56])
Beta I			
Test effectiveness and acceptability of transfer elements [43]	A/B testing on ROOM versions	231 EUR students: 168 (72.7%) female, 60 (26.0%) male, and 3 (1.3%) no disclosure	Mixed-effects linear models and Wilcoxon signed rank test
Evaluate experience with learning transfer [43]	Diary study + follow-up interviews	13 EUR students: 7 (54%) female, 5 (38%) male, and 1 (8%) nonbinary	Thematic analysis
Evaluate ROOM app usage patterns	Usage log data analysis (A/B + diary participants)	62 A/B study participants + 13 diary study participants	Log data analysis

[a]EUR: Erasmus University Rotterdam.

[b]DBCI: digital behavior change intervention.

XSL•FO
RenderX

Summary of Key Outcomes of the Formative Evaluations

Low-Fidelity Prototype Phase (Early Design and Concept Testing)

Expert focus groups and interviews (n=11) indicated that the intervention content was accurate, appropriate for the targeted problems, and sufficiently faithful to the original therapeutic approaches. Experts recommended incorporating compassionate debriefing messages, including disclaimers that the app does not replace professional therapy, tailoring examples to university life, addressing privacy concerns, and using relatable, unpolished examples to avoid reinforcing perfectionism. They also emphasized the importance of expanding the research to include more male participants, as most early users were female.

Quasi-experimental pre-post studies (n=192) using low-fidelity prototypes (delivered via Qualtrics, Miro boards, or slides) indicated that the exercises improved proximal outcomes, including cognitive defusion, thought believability and discomfort, emotional states, and self-compassion. However, effects on trait-level characteristics were less consistent: one study reported improvements in cognitive defusion after 5 days, whereas another did not, and mindfulness levels remained unchanged [43]. Qualitative research with students (n=85) indicated that participants understood and were able to carry out the presented techniques. Students appreciated the short length of the exercises and the psychoeducational explanations, which helped make abstract concepts more concrete. They also emphasized the need for repetition and longer-term practice to sustain benefits, rather than relying solely on brief interventions. Challenges included emotional distress triggered by certain exercises—particularly those targeting unhelpful thinking patterns—highlighting the need for postexercise debriefs, occasional confusion over complex metaphors, and feelings of being overwhelmed by the volume of text.

High-Fidelity Clickable Prototype Phase (Interactive App Design and User Testing)

Evaluation of the high-fidelity clickable prototype using think-aloud methods and interviews (n=15), which simulated the app's visual design, navigation, animations, and interactions, indicated that users found the app's aesthetics appealing and the therapeutic content engaging. However, they experienced difficulties with nonintuitive navigation, unclear icons, hidden background information, and audio or animation timing that occasionally made interactions uncomfortable. Users frequently overlooked key elements, such as the "info" button, which disrupted the intended flow of exercises. These findings highlighted the need for clearer navigation, more intuitive interaction design, and enhanced visual cues to guide users smoothly through the exercises.

Co-design workshops (n=10) provided detailed feedback on features related to learning transfer [43]. Action-planning tools (eg, if-then plans) and body-based cues (eg, hand gestures) were perceived as practical and directly applicable to real-life situations [43]. By contrast, social-sharing features (eg, inviting friends to join exercises) were generally rejected, as participants preferred to engage with the app individually and privately. Users favored short, playful exercises that could be easily integrated into daily routines, supported by habit-formation tools such as reminders or personalized challenges, while avoiding complex or cognitively demanding tasks. They further emphasized that more complex or emotionally challenging techniques required repeated, guided practice rather than immediate off-screen transfer [43].

Alpha Prototype Phase (Structured Testing With Micro-Randomized Trial)

The alpha version, which included 20 ER exercises, was tested in a micro-randomized trial (n=146) over 21 days. Participants completed exercises twice daily, in the morning and evening, and were subsequently interviewed about their UX (n=18) [55,56]. The trial demonstrated high adherence to intervention completion, with an average completion rate of 65.98% (4,046/6,132), likely supported by monetary incentives for completing daily assessments. Results indicated immediate positive effects on emotional states following the exercises, along with gradual improvements in ER and stress symptoms over the intervention period; depressive and anxiety symptoms remained stable. Qualitative interviews revealed that, while some participants appreciated the structured delivery and felt motivated by the imposed schedule, others found it restrictive and preferred greater autonomy. Exercises lasting longer than 7 minutes were most frequently left incomplete, and those requiring participants to speak aloud or change their environment received mixed feedback. Breathing exercises were the most liked, whereas cognitive defusion exercises were the least liked. Qualitative findings also identified notifications and context-sensitive recommendations as among the most commonly expressed user needs [55,56].

Beta Prototype Phase (A/B Testing and Transfer Effectiveness)

The beta I version expanded the app to include 24 exercises and learning transfer elements [43], along with a self-assessment module and a recommendation system based on data from the micro-randomized trial (alpha prototype). A/B testing (n=231) compared app versions with and without transfer elements, focusing on their effects on ER skills and the application of learned skills off-screen [43]. Overall, app usage was low, with an average of only 2.5 completed exercises and brief durations of active engagement. However, the data suggested that the version incorporating transfer elements facilitated earlier application of skills in real-life contexts [43].

A diary study (n=13) evaluating transfer elements [43] showed that participants applied skills learned in the app using transfer elements, such as action plans and environmental cues (eg, stickers) [43]. Participants often simplified the techniques, favoring short, easy practices with immediate effects—such as breathing or grounding exercises—over more complex ones. Perceived positive changes included learning new coping strategies, increased self-compassion, and momentary boosts in positive affect. Barriers to transferring skills to real-life contexts included cognitive demands, forgetting action plans or the purpose of environmental cues (eg, stickers), and feeling overwhelmed by lengthy or complex techniques [43].

User log data indicated that active users (n=75, who participated in the A/B and diary studies) completed an average of 6

XSL•FO
RenderX

exercises and saved the objects earned by completing exercises in the virtual room. Exercises were generally rated positively, with an average likeability score of 73.7/100 and perceived helpfulness of 68.6/100. While 66 users accessed the self-assessment module, only 39 completed a survey. Perfectionism was the most frequently selected topic. Longer questionnaires were often abandoned. Users reported that the app required frequent log-ins, which they found frustrating, and that the overall structure was not user-friendly. Specifically, many struggled to explore content beyond what was highlighted by the recommender system. Additionally, the app lacked features to track progress or revisit past exercises and action plans. The app's key feature of facilitating the application of learned strategies outside the app (ie, learning transfer) [43] was not presented in a way that felt actionable or salient to users.

Outcome of the Design Phase: ROOM to Grow

Overview of the ROOM App and Its Core Features

Based on the results of the design phase, we refined the app to create a functional early version, referred to as a minimum viable product. This section summarizes the key features of the ROOM app, including intervention targets, content, engagement strategies, and UX. Detailed descriptions of the components, rationale, underlying theories, and the user journey are available in the OSF environment [43].

Intervention Content and Targets

ROOM is a mobile-based, transdiagnostic mental health intervention targeting ER through short (3-10 minutes) behavior change exercises grounded in mindfulness, self-compassion, positive psychology, ACT, and CBT. Each approach addresses different stages of the ER process; for example, cognitive restructuring helps users reframe triggering situations, while self-compassion supports the modulation of emotional responses. The variety of approaches caters to the diverse needs and preferences of university students, allowing them to explore and identify strategies that work best for them. Beyond ER, ROOM is designed to enhance self-awareness. Recognizing that students often delay addressing mental health issues, the app includes a brief monitoring feature that evaluates 7 emotional states (see OSF [43]) and longer questionnaires assessing both transient psychological states (eg, depressive symptoms) and trait-level characteristics (eg, perfectionism). The app identifies risk and protective factors and educates users on topics such as perfectionism, depression, and self-efficacy, while providing personalized feedback.

Engagement Strategy

First, students perceived mobile phone use as a source of stress and sought to minimize screen time. Second, expert feedback emphasized the importance of designing content that supports the application of learned skills beyond the app [43]. Third, although users desired autonomy, our studies showed that complete freedom in app usage tended to reduce engagement. To meet the needs of this target group, ROOM is designed to promote real-life skill application while limiting prolonged screen time [43]. To achieve its intended impact, users are encouraged to engage with the app for at least 21 days—based on findings from the alpha version study showing improvements in ER skills and stress [55]—exploring various content, discovering what works best for them, building well-being habits, and mastering the skills learned in the exercises. In line with this, ROOM includes (1) short, actionable therapeutic content for immediate use; (2) a gamified collectible system enabling quick access to favorite exercises from the home screen; (3) features that support the application of learned techniques in real-world contexts (learning transfer) [42,43]; and (4) a 21-day challenge that prompts users to monitor their mood daily and try 1 exercise per day.

User Experience

The goal was to design an intuitive, aesthetically pleasing, user-friendly, and time-efficient interface responsive to users' needs. We prioritized credible, evidence-based content presented in short, focused segments to accommodate limited attention spans. The visual design drew inspiration from Chill Hop, a style popular with the target audience. To avoid overwhelming users, additional details are nested under "insights" and "info" buttons, allowing users to control the amount of content they access.

Each exercise and questionnaire explains its relevance and includes practical, student-focused examples. An intelligent recommendation system addresses users' need for responsiveness and personalization, connecting them to content that aligns with their individual needs and preferences. The recommender system learns and adapts based on user input while adhering to a privacy-by-design principle. It uses a federated machine learning model, which trains directly on users' mobile devices, keeping data private by sharing only anonymized model updates with a cloud-based model aggregator (see Laure et al [42] and Villegas Mejia et al [43] for more details). To address students' privacy concerns, no invasive tracking (eg, geo-location) is employed.

Lastly, to ensure user safety and provide a positive end-to-end experience, the app incorporates several safety measures. For example, rules embedded in the recommender system prevent the suggestion of certain exercise types based on reported moods. The app also provides reassurance after completing exercises or questionnaires through debriefing information. Additionally, it includes up-to-date information on relevant resources, such as university psychologists or general practitioners, for users who may require further support.

Operationalization

The planning for the integration and dissemination of ROOM within the EUR context was intertwined throughout the development process, particularly during the design phase, when highly refined prototypes were tested before reaching the minimum viable product. Below, we describe ROOM's implementation within the EUR ecosystem and the strategy for reaching the end users (ie, students).

Technology—Context Implementation

First, the app is integrated into the broader EUR Student Wellbeing Programme's chain of care, leveraging its visibility and awareness among most EUR students. Second, considerable effort was devoted to building strong relationships with the

XSL•FO
RenderX

university's faculties, student support services, and information technology services, which supported both the research and the development of the tool. As a result, they have been aware of the initiative and committed to supporting ROOM's implementation at EUR. Third, ROOM was developed in alignment with both internal university and international standards and regulations. In particular, it adheres to the privacy-by-design framework [57], ensuring that data protection is integrated into the app architecture from its inception. Data privacy was also highlighted by students as an important consideration (see Table 3). Compliance with data and technology regulations (ie, European Union General Data Protection Regulation [EU GDPR] [58], EU Artificial Intelligence [AI] Act [59]), and security frameworks—including robust technical decisions, penetration testing, and International Organization for Standardization certifications—was a significant part of the development, ensuring that ROOM can be sustained beyond the research setting and avoid the "valley of death" phenomenon [60]. ROOM is fully hosted on the university's cloud infrastructure, and student authentication protocols were tested and integrated into the app, allowing EUR students secure and free access to ROOM.

Communication and Marketing Strategy for ROOM's User Adoption

Resources were allocated for implementation, and an implementation officer and a communication officer were hired to coordinate the rollout of ROOM and its communication campaign at EUR. User research indicated the need for diverse outreach strategies: some students engage with well-being proactively, while others seek support primarily during challenging periods. Furthermore, regarding messaging and slogans, some students respond positively to personal development themes (eg, "Be the best you"), while others prefer more compassionate messaging. Student associations, sports clubs, and fraternities may be more receptive to messages incorporating humor and references to popular internet culture. Peer-to-peer recommendations are also perceived as highly relevant by students (see Table 3). Early adopters of ROOM play a key role in spreading information and creating a bottom-up movement to engage other students. Those less interested in well-being and personal development are likely to engage later, once ROOM gains popularity.

Therefore, ROOM's implementation and communication strategy includes promotion through multiple channels, such as social media, the university intranet, the Student Wellbeing Programme's communication channels, student ambassadors, and interactive installations displaying ROOM's signature objects (eg, the Lava lamp) alongside QR codes directing students to the ROOM App Store page for more information and downloads. These initiatives are being rolled out through targeted campaigns across the academic year, ensuring that students are reached when they are most receptive—whether seeking support (eg, during periods of stress) or focusing on self-development (see Table 2). This approach is designed to meet students where they are, aligning with their diverse needs and readiness to engage with the tool.

Ethical Consideration

All research activities conducted across the different phases of the road map adhered to the Netherlands Code of Conduct for Research Integrity and received approval from the Ethics Review Board of the Department of Psychology, Education, and Child Studies at EUR (records number 20-066a2, ETH2122-0677, and TH2324-0086). The final randomized controlled trial evaluating ROOM's effectiveness was registered at ClinicalTrials.gov (NCT05576883). All participants in our studies provided informed consent after reviewing information about the specific round of data collection they joined. Privacy measures included data anonymization and secure storage. Participants were compensated with vouchers of varying amounts, generally reflecting the time invested in each study (eg, €5 [US $6.7] for 30 minutes). For smaller studies, participants were entered into a lottery as a form of compensation.

Discussion

Principal Findings

The objective of this paper was 2-fold: to present the developed solution, ROOM, and to share insights from its development process. Specifically, we detailed how the 4 phases of the CeHRes road map [39,40]—contextual inquiry, value specification, design, and operationalization—shaped this mobile intervention. The design process was iterative, employing a range of qualitative and quantitative research methods and engaging key stakeholders, including university students, mental health experts, study advisers, and university psychologists, throughout the development phases. The outcome of this process is ROOM, which incorporates the following signature features: it adopts a transdiagnostic approach to mental health, addressing the underlying mechanisms of a range of mental health issues common among university students by targeting ER and self-awareness. The intervention functions as a micro-intervention, promoting active engagement with its content both inside and outside the app. It also features a recommender system that adapts content to the user based on their preferences (eg, managing unhelpful thoughts or practicing relaxation) and momentary needs (ie, current emotional states). ROOM adheres to high standards of privacy and security, ensuring transparency in data management. It also gives users autonomy over their engagement with the app and deliberately avoids engagement patterns commonly seen on social media, promoting genuine and meaningful user interaction.

In the following sections, we discuss the resulting intervention, outline the development process, highlight the challenges encountered during the app's design, and provide practical recommendations for developers creating adaptive mental health interventions for youth.

The Developed Solution

ROOM employs a transdiagnostic approach to mental health, providing flexible, personalized support tailored to each student's unique needs. This approach has gained popularity for its potential to prevent common mental health issues by targeting underlying vulnerability factors rather than focusing

XSL•FO
RenderX

on disorder-specific symptoms [61]. Previous studies have shown that applying a transdiagnostic approach in digital interventions can successfully reduce anxiety and depressive symptoms—2 of the most common concerns among university students [62]. In ROOM, the transdiagnostic factor targeted is ER [63]. The use of adaptive ER strategies—such as recognizing, expressing, and appropriately regulating emotions—is associated with enhanced well-being [64] and improved academic performance [65]. Our studies indicate that engaging with ROOM reduces distress and enhances ER skills [55]. Focusing on ER among university students is a promising approach for maintaining and promoting their well-being.

Another key target of the ROOM intervention, closely linked to ER, is self-awareness. Awareness of one's emotional state is essential for selecting appropriate ER strategies [66]. ROOM fosters self-awareness by encouraging users to monitor their emotional states and gain insights about themselves through questionnaires assessing various psychological states (eg, distress symptoms) and traits (eg, perfectionism) that influence well-being. Because of the psychoeducational content embedded in the self-assessment modules (as well as in the exercises), we expect ROOM to enhance users' mental health literacy. Mental health literacy plays a key role in fostering help-seeking behaviors, encouraging users to recommend help to others, and reducing mental health stigma [67]. Thus, ROOM may help decrease stigma and promote help-seeking among students who require additional support. While this was not the focus of the current studies, it will be evaluated in future research.

This target population has a strong preference for systems that offer personalization [35,36] and aims to reduce screen time. Therefore, rather than protocolized, longer, and intensive mental health interventions such as internet-based CBT [68] or mindfulness-based stress reduction [69], ROOM incorporates brief (3-10 minutes) behavior change exercises that can be conceptualized as micro-interventions—that is, targeted, short interventions integrated into a person's daily routine, causing minimal disruption [70]. These exercises provide practical, real-time strategies that help students regulate their emotions and enhance self-awareness in the moment. Research on micro-interventions suggests that this type of approach is a promising way to improve moods [71,72], promote value-based behaviors [73], and encourage behavioral activation [74]. Consistent with this, our studies indicate that engaging with ROOM has successfully improved momentary moods [55], fostered growth in ER skills, and reduced stress levels. However, the long-term effects of this intervention remain to be evaluated.

ROOM also incorporates an intelligent recommender system that delivers meaningful, context-sensitive suggestions based on users' history, preferences, and current emotional states, ensuring relevant support in the moment. Consequently, ROOM users do not follow a predetermined regimen; their engagement with the content is personal and varies for each student, enabling genuine personalization. Recommender systems are playing an increasingly important role in mental health, serving multiple functions such as reducing choice overload, strengthening the digital therapeutic alliance (ie, the relationship between the user and the digital tool [75]), and enhancing user engagement [76].

A distinctive feature of ROOM is its engagement strategy. The app is not designed for long-term retention but aims to teach users relevant ER strategies and facilitate the application of these skills in real-life settings [43]. The app places less emphasis on traditional engagement metrics—such as time spent online (ie, micro-engagement)—and instead prioritizes meaningful interactions. It scaffolds engagement with the content and supports the associated behavior change process beyond the screen (ie, macro-engagement [45,46]). This design choice is grounded in our research, which indicates that digital screens often cause stress for students, prompting them to minimize screen time. This finding aligns with recent studies showing that users are increasingly saturated with online tools and experience digital stress [77,78]. Moreover, applying learned skills is an aspect that Generation Z considers crucial in their learning process [3,79]. ROOM incorporates learning transfer elements—such as implementation intentions [80], associative cues, and tactile guidance [81]—to support real-world application of skills, aligning with the values of this generation. Our studies have shown that the transfer elements may accelerate the practice of learned techniques in users' daily lives [43], suggesting that engagement modalities extending interaction with the content beyond the app and into real-life activities represent a promising avenue for enhancing learning and behavior change outcomes.

Long-term, sustained usage is not a goal of ROOM. However, users need to engage with the app for a defined period for it to achieve its intended impact. To support effective engagement, ROOM incorporates persuasive design elements, including gamification mechanics and a recommender system—both recognized as key drivers of user engagement [82]. One gamification element in ROOM is the collectibles mechanic [83], in which users collect objects linked to exercises and use them to decorate and personalize their virtual room. This feature provides opportunities for self-expression (ie, personalizing their room [84]) and highlighting exercises they enjoyed, while also offering quick access to revisit those exercises, addressing users' need for time efficiency. Collectible game mechanics are a popular gamification element [83], and formative evaluations indicated that users greatly enjoyed this feature. The second gamification element is the 21-day challenge, which encourages users to engage with the app consistently for 3 weeks to experience its benefits. Although this element has not yet been formally assessed, previous research suggests that challenges can effectively motivate users to continue engaging with a system or application [85].

Lastly, user autonomy and privacy emerged as central themes in our research, consistent with prior studies on this population [35,37]. ROOM supports autonomy by allowing users to engage flexibly at their preferred pace and in their chosen manner, while also incorporating features that facilitate learning transfer [43] and maximize positive outcomes. The app is fully transparent about how user data are managed and provides clear insights into the workings of its recommender system. Many commercial apps lack robust privacy standards and may sell users' mental health data to third parties [86]. By contrast, ROOM stands out by effectively balancing persuasive strategies with user autonomy while upholding strict data privacy standards.

XSL•FO
RenderX

The Development Process

ROOM is the result of a 4-year development process guided by the CeHRes road map [40]. Throughout this period, the target population and other relevant stakeholders were continuously involved to ensure the app aligned with users' needs and preferences. We employed a range of research approaches, from UX studies to experimental research, to design and refine the app's content and functionalities. Using both quantitative and qualitative methods deepened our understanding of the student population, their needs and preferences, how they engage with interventions independently (ie, in their private time), and their perceptions of these interventions. By identifying problems early and iteratively optimizing the design, ROOM's content and functionality became better aligned with the target users, enhancing both its UX and effectiveness. Moreover, collaboration with experts from diverse disciplines—each contributing unique perspectives, expertise, and methodologies—strengthened problem-solving, fostered innovation, and supported the development of comprehensive solutions throughout the project [87]. Their involvement was crucial for evaluating content for fidelity to the intervention approaches, addressing issues identified in user research, and designing user-friendly, engaging app elements that support behavior change within the available resources.

Strengths and Limitations

Currently, evaluation of the intervention's effectiveness and uptake is ongoing [42], and ROOM is being implemented university-wide, achieving over 3200 downloads within the first few weeks. While results from the final effectiveness study [42] are not yet available, our research already provides preliminary support for our approach to digital intervention development. Particularly, ROOM is generally well received by students, functions as intended, and is safe to use, as indicated by improved or stable levels of distress symptoms [55]. The main strength of our approach lies in its iterative, agile development, the inclusion of a variety of stakeholders, and, most importantly, the consistent consideration of information gathered from our target population (ie, university students).

A notable limitation of our development process is the overrepresentation of female participants in these studies. This is unsurprising, as female students often report experiencing more mental health difficulties [4]. Moving forward, we plan to actively involve more male and other gender participants in the development process (eg, when adding content or modifying the visual design) to enhance the inclusivity of ROOM and ensure it incorporates perspectives and information that may currently be underrepresented.

Challenges, Compromises, and Recommendations

Developing ROOM required navigating multiple challenges, including balancing student needs with constraints such as privacy concerns and interdisciplinary collaboration, all of which influenced the app's final design and functionality. Below, we outline key challenges and compromises made, along with practical recommendations.

Depth of Information Versus Preference for Brevity

Adapting mental health interventions for Generation Z, who prefer short and engaging content, led us to develop brief 3-10-minute behavior change exercises. The challenge was maintaining the depth and quality of psychological content within this condensed format. Through iterative testing and expert input, we achieved a balance between brevity and preserving essential content. For interventions of this type, it is recommended to use a streamlined flow that presents only the necessary information to complete the exercise, while offering optional "learn more" buttons for users who want additional details or support. Developers should continuously test content with users and consult experts (ie, mental health professionals) to ensure fidelity to therapeutic principles and maintain exercise quality.

Content Diversity Versus Interactivity

Balancing content diversity with interactivity posed another significant challenge. While our goal was to offer a wide range of content to meet the varied needs of the target population, prioritizing content variety within the iterative development process limited our ability to incorporate highly interactive elements, such as videos or interactive gameplay, due to resource constraints. Currently, the app primarily relies on written content, quizzes, and audio guides. Given that this target population often learns through visual media (eg, videos) [3] and is accustomed to high levels of interactivity in digital tools, incorporating varied media formats—especially videos—is strongly recommended when developing digital solutions for youth. If such formats are not feasible, presenting content in brief segments, such as 2 sentences per page, can serve as an effective alternative.

Privacy by Design Versus System Interactivity and Personalization

Designing for privacy while maintaining system interactivity and personalization posed another significant challenge. Unlike many commercial apps that prioritize UX over privacy [88], ROOM adhered to the "privacy-by-design" principle, reflecting both the sensitivity of mental health data and Generation Z's strong privacy concerns [35]. We opted to store and process user data locally on their phones rather than in the cloud, ensuring compliance with privacy regulations such as the EU GDPR law [58]. This approach, however, limited the level of personalization that could otherwise be offered. For example, we forwent passive data collection, such as GPS tracking, which could have enabled more tailored recommendations and the design of just-in-time adaptive interventions [89]. However, the use of federated machine learning [90] enabled us to provide a personalized experience while complying with strict privacy-by-design principles. It remains essential to continue protecting user privacy and promoting ethical practices by exploring innovative, privacy-preserving technologies. Future research should investigate the balance between personalization and privacy, examining user attitudes and the effects of these trade-offs.

XSL•FO
RenderX

Fixed Resources Versus Iterative Process

Another challenge was operating within fixed time, budget, and scope constraints while committing to an iterative development process, in our case following the CeHRes road map, and using agile project management [91]. Agile project management offers various strengths, such as an iterative and flexible approach to product development, but it also presents challenges, including limited process visibility, difficulties with long-term planning, and resource allocation issues [92]. In our development process, we struggled with resource allocation and had to continuously adjust the project scope [91]. Balancing technical requirements for a sustainable product, UX, ethical guidelines, and scientific rigor made the prioritization of values and features paramount. We implemented regular review checkpoints to assess progress and remained adaptable to evolving requirements within the existing resource constraints. We recommend that developers carefully select an appropriate project management approach and conduct regular budget reviews after each iteration (eg, monthly). These reviews should include refining priorities and ensuring that remaining resources are focused on high-priority items, helping to keep development on track within budget limits.

Interdisciplinary Teams and Collaboration

Working with interdisciplinary teams and across sectors (ie, academia and industry) required developing a shared language and vision—both a challenge and an opportunity [93]. While the diverse expertise brought by team members from different fields enriched the project, it required time and effort to synchronize our understanding and approaches. For example, academic-industry collaborations are crucial for building viable products that can survive "in the real world," but they face challenges due to differing processes and pace [94,95]. Academia operates under rigorous, protocol-driven environments that prioritize deliberative, gradual progress, often with changing requirements, whereas industry settings demand speed, innovation, and rapid implementation [95]. To bridge these differences, it is paramount to allocate time at the start of the project to select an appropriate collaboration model, align processes, and ensure that neither the integrity nor the innovativeness of the project is compromised.

Conclusions

This paper presents the multifaceted process behind the development of a mobile transdiagnostic well-being intervention for university students. Several key insights emerged that may inform the development of future digital mental health tools for young people. First, our findings suggest that for a heterogeneous population such as university students, a transdiagnostic approach to mental health can effectively address their varied needs. Second, tech-savvy Generation Z often seeks to reduce screen time due to technostress, highlighting the importance of designing digital mental health tools that support both in-app and off-screen engagement strategies. Third, system responsiveness and content curation are highly important to users, highlighting the need for adaptive interventions and the integration of tools such as AI-based recommender systems to effectively meet their needs. Fourth, prioritizing security and privacy in digital mental health tools is essential for building and maintaining users' trust. This project demonstrates that, even when creating a "data-hungry" system, privacy-friendly practices are achievable when integrated from the outset. Lastly, developing such a tool requires user-centered approaches and a strong interdisciplinary team to address the design, functionality, and content needs of stakeholders. Although the full evaluation of ROOM's impact is still pending, initial feedback indicates that it is well-received, feasible, and safe for student use.

Acknowledgments

We are grateful to everyone involved in this project. We thank the entire EUR Student Wellbeing Programme team, and especially Rianne van der Zanden, our program coordinator, for her unwavering encouragement and expert support in managing and embedding ROOM within the programme and the broader EUR ecosystem. A special thanks goes to the IJsfontein team, especially Ronald Benschop, Stijn Raaijmakers, Rose-Anne Dotinga, Eddy Klappe, Vera Koot, Kate Romadina, Kelly van Evert, and Salko Joost Kattenberg, for their dedication and efforts in bringing ROOM to reality. Thanks also to Thoughtworks, and particularly Oege Dijk, for developing the recommender system. We are grateful to all the students and colleagues involved in the development of the tool: Aydan Yusifova, Iman Shadid, Ximena M. Zúñiga, Bhavana Madiraju, Bas Oppenheim, Lara Ritter, Karina Riffelsberger, Dimitra Papadaki, Bianca Rozs, Eren Sadi, Andia Taliadorou, Gabriëlle Lijten, Tonka Milošević, Marissa Duess, Mina Awad, Matylda Kępa, Iliana Liakea, Andrea Reyna Zorrilla, Momin Qazi, Sara Serrao, Vid Verčko, and Aljaž Laure. We extend special thanks to Eveline Crone, Donna Spruijt-Metz, and Susan Murphy for their expert advice and support throughout the entire project development.

Funding

This project is funded by internal EUR funding, the Nationaal Programma Onderwijs (NPO) from the Dutch Ministry of Education, Culture, and Science, and a grant by the Erasmus Trustfonds (GWA.100939).

Data Availability

The data supporting the findings of the design process will be made available on the OSF [43] following an embargo period after publication, to allow for the commercialization of the research findings. Data from qualitative interviews, however, will not be shared to protect participants' privacy and confidentiality.

XSL•FO
RenderX

Conflicts of Interest

TL, CVM, MB, DR, DD, and RCMEE declare a conflict of interest, as they were directly involved in the development of the ROOM app. The remaining authors declare no conflicts of interest.

Multimedia Appendix 1
Overview of Centre for eHealth Research (CeHRes) stages, its objectives, and related activities.
[DOCX File , 29 KB - formative_v9i1e63325_app1.docx]

References

1. Newton J, Dooris M, Wills J. Healthy universities: an example of a whole-system health-promoting setting. Glob Health Promot 2016 Mar 19;23(1 Suppl):57-65. [doi: 10.1177/1757975915601037] [Medline: 27199018]
2. Enwefa S, Enwefa R. Redefining the future of higher education in the 21st century: educating and preparing for today and tomorrow. International Journal of Innovative Business Strategies (IJIBS) 2016 Dec 01;2(2):95-101. [doi: 10.20533/ijibs.2046.3626.2016.0015]
3. Seemiller C, Grace M. Generation Z: educating and engaging the next generation of students. About Campus: Enriching the Student Learning Experience 2017 Jul 01;22(3):21-26. [doi: 10.1002/abc.21293]
4. Auerbach RP, Mortier P, Bruffaerts R, Alonso J, Benjet C, Cuijpers P, WHO WMH-ICS Collaborators. WHO World Mental Health Surveys International College Student Project: prevalence and distribution of mental disorders. J Abnorm Psychol 2018 Oct;127(7):623-638 [FREE Full text] [doi: 10.1037/abn0000362] [Medline: 30211576]
5. Auerbach RP, Alonso J, Axinn WG, Cuijpers P, Ebert DD, Green JG, et al. Mental disorders among college students in the World Health Organization World Mental Health Surveys. Psychol Med 2016 Oct;46(14):2955-2970 [FREE Full text] [doi: 10.1017/S0033291716001665] [Medline: 27484622]
6. Bruffaerts R, Mortier P, Kiekens G, Auerbach RP, Cuijpers P, Demyttenaere K, et al. Mental health problems in college freshmen: prevalence and academic functioning. J Affect Disord 2018 Jan 01;225:97-103 [FREE Full text] [doi: 10.1016/j.jad.2017.07.044] [Medline: 28802728]
7. Hubble S, Bolton P. Support for students with mental health issues in higher education in England. Commons Library Briefing. London, UK: House of Commons Library, UK Parliament; 2019 Oct 14. URL: https://dera.ioe.ac.uk/id/eprint/34234/ [accessed 2024-03-17]
8. Jenkins PE, Ducker I, Gooding R, James M, Rutter-Eley E. Anxiety and depression in a sample of UK college students: a study of prevalence, comorbidity, and quality of life. J Am Coll Health 2021 Jan 29;69(8):813-819. [doi: 10.1080/07448481.2019.1709474] [Medline: 31995452]
9. Veldman K, Reijneveld SA, Verhulst FC, Ortiz JA, Bültmann U. A life course perspective on mental health problems, employment, and work outcomes. Scand J Work Environ Health 2017 Jul 01;43(4):316-325 [FREE Full text] [doi: 10.5271/sjweh.3651] [Medline: 28570743]
10. Behr A, Giese M, Teguim Kamdjou HD, Theune K. Dropping out of university: a literature review. Review of Education 2020 Mar 23;8(2):614-652. [doi: 10.1002/rev3.3202]
11. Cianconi P, Betrò S, Janiri L. The impact of climate change on mental health: a systematic descriptive review. Front Psychiatry 2020;11:74-11 [FREE Full text] [doi: 10.3389/fpsyt.2020.00074] [Medline: 32210846]
12. Hooft VHL, Elfferich A, Marchal B, Molenaar T, van ZS. Landelijke monitor studentenhuisvesting 2023. Het kenniscentrum voor studentenhuisvesting. The Netherlands: Het kenniscentrum voor studentenhuisvesting; 2023 Oct 07. URL: https://www.kences.nl/publicaties/landelijke-monitor-studentenhuisvesting-2023/ [accessed 2024-12-06]
13. Sheldon E, Simmonds-Buckley M, Bone C, Mascarenhas T, Chan N, Wincott M, et al. Prevalence and risk factors for mental health problems in university undergraduate students: a systematic review with meta-analysis. J Affect Disord 2021 May 15;287:282-292. [doi: 10.1016/j.jad.2021.03.054] [Medline: 33812241]
14. Arnett JJ. Emerging adulthood: what is it, and what is it good for? Child Dev Perspectives 2007 Dec 07;1(2):68-73. [doi: 10.1111/j.1750-8606.2007.00016.x]
15. Arnett JJ. Emerging adulthood: a theory of development from the late teens through the twenties. American Psychologist 2000;55(5):469-480. [doi: 10.1037/0003-066x.55.5.469]
16. How many people have a student debt? Statistics Netherlands. The Netherlands: Centraal Bureau voor de Statistiek; 2023. URL: https://longreads.cbs.nl/the-netherlands-in-numbers-2023/how-many-people-have-a-student-debt/ [accessed 2024-06-12]
17. The Student Wellbeing Platform. Erasmus Universiteit Rotterdam. URL: https://www.eur.nl/en/education/study-rotterdam/student-wellbeing [accessed 2024-10-06]
18. Troy A, Mauss I. Resilience in the face of stress: Emotion regulation as a protective factor. In: Southwick SM, Litz BT, Charney D, Friedman MJ, editors. Resilience and Mental Health: Challenges Across the Lifespan. Cambridge, MA: Cambridge University Press; 2011:30-45.
19. Ho FY, Yeung W, Ng TH, Chan CS. The efficacy and cost-effectiveness of stepped care prevention and treatment for depressive and/or anxiety disorders: a systematic review and meta-analysis. Scientific Reports 2016 Dec 05;6:29281 [FREE Full text] [doi: 10.1038/srep29281] [Medline: 27377429]

XSL•FO
RenderX

20. Stepped care for post-secondary campuses: a promising model to improve access to mental health care on campus. The Centre for Innovation in Campus Mental Health. Ontario, Canada: The Centre for Innovation in Campus Mental Health; 2019 Jul 25. URL: https://campusmentalhealth.ca/toolkits/stepped-care/ [accessed 2024-06-12]
21. Short CE, DeSmet A, Woods C, Williams SL, Maher C, Middelweerd A, et al. Measuring engagement in eHealth and mHealth behavior change interventions: viewpoint of methodologies. J Med Internet Res 2018 Nov 16;20(11):e292 [FREE Full text] [doi: 10.2196/jmir.9397] [Medline: 30446482]
22. Ferrari M, Allan S, Arnold C, Eleftheriadis D, Alvarez-Jimenez M, Gumley A, et al. Digital interventions for psychological well-being in university students: systematic review and meta-analysis. J Med Internet Res 2022 Sep 28;24(9):e39686 [FREE Full text] [doi: 10.2196/39686] [Medline: 36169988]
23. Lattie EG, Adkins EC, Winquist N, Stiles-Shields C, Wafford QE, Graham AK. Digital mental health interventions for depression, anxiety, and enhancement of psychological well-being among college students: systematic review. J Med Internet Res 2019 Jul 22;21(7):e12869 [FREE Full text] [doi: 10.2196/12869] [Medline: 31333198]
24. Harrer M, Adam SH, Baumeister H, Cuijpers P, Karyotaki E, Auerbach RP, et al. Internet interventions for mental health in university students: a systematic review and meta-analysis. Int J Methods Psychiatr Res 2019 Jun 26;28(2):e1759 [FREE Full text] [doi: 10.1002/mpr.1759] [Medline: 30585363]
25. Davies EB, Morriss R, Glazebrook C. Computer-delivered and web-based interventions to improve depression, anxiety, and psychological well-being of university students: a systematic review and meta-analysis. J Med Internet Res 2014 May 16;16(5):e130 [FREE Full text] [doi: 10.2196/jmir.3142] [Medline: 24836465]
26. Theurel A, Witt A. Identifying barriers to mental health help-seeking in French university students during the Covid-19 pandemic. Creative Education 2022;13(02):437-449. [doi: 10.4236/ce.2022.132025]
27. Niaz Ahmed. Generation Z's smartphone and social media usage: a survey. Journalism and Mass Communication 2019 Mar 28;9(3):437-449. [doi: 10.17265/2160-6579/2019.03.001]
28. Sidoti O, Gelles-Watnick R, Faverio M, Atske S, Radde K, Park E. Mobile fact sheet. Pew Research Center. Washington, DC: Pew Research Center; 2024 Jan 31. URL: https://www.pewresearch.org/internet/fact-sheet/mobile/ [accessed 2024-06-12]
29. Andersson G, Titov N. Advantages and limitations of internet-based interventions for common mental disorders. World Psychiatry 2014 Feb;13(1):4-11 [FREE Full text] [doi: 10.1002/wps.20083] [Medline: 24497236]
30. Ebert DD, Cuijpers P, Muñoz RF, Baumeister H. Prevention of mental health disorders using internet- and mobile-based interventions: a narrative review and recommendations for future research. Front Psychiatry 2017;8:116 [FREE Full text] [doi: 10.3389/fpsyt.2017.00116] [Medline: 28848454]
31. Koh J, Tng GYQ, Hartanto A. Potential and pitfalls of mobile mental health apps in traditional treatment: an umbrella review. J Pers Med 2022 Aug 25;12(9):1376. [doi: 10.3390/jpm12091376] [Medline: 36143161]
32. Bakker D, Kazantzis N, Rickwood D, Rickard N. Mental health smartphone apps: review and evidence-based recommendations for future developments. JMIR Ment Health 2016 Mar 01;3(1):e7 [FREE Full text] [doi: 10.2196/mental.4984] [Medline: 26932350]
33. Marshall JM, Dunstan DA, Bartik W. Apps with maps-anxiety and depression mobile apps with evidence-based frameworks: systematic search of major app stores. JMIR Ment Health 2020 Jun 24;7(6):e16525. [doi: 10.2196/16525] [Medline: 32579127]
34. Eis S, Solà-Morales O, Duarte-Díaz A, Vidal-Alaball J, Perestelo-Pérez L, Robles N, et al. Mobile applications in mood disorders and mental health: systematic search in Apple App Store and Google Play Store and review of the literature. Int J Environ Res Public Health 2022 Feb 15;19(4):2186 [FREE Full text] [doi: 10.3390/ijerph19042186] [Medline: 35206373]
35. Melcher J, Camacho E, Lagan S, Torous J. College student engagement with mental health apps: analysis of barriers to sustained use. J Am Coll Health 2022 Oct 13;70(6):1819-1825. [doi: 10.1080/07448481.2020.1825225] [Medline: 33048626]
36. Becker TD, Torous JB. Recent developments in digital mental health interventions for college and university students. Curr Treat Options Psych 2019 Jun 14;6(3):210-220. [doi: 10.1007/s40501-019-00178-8]
37. Torous J, Nicholas J, Larsen ME, Firth J, Christensen H. Clinical review of user engagement with mental health smartphone apps: evidence, theory and improvements. Evid Based Ment Health 2018 Aug 05;21(3):116-119 [FREE Full text] [doi: 10.1136/eb-2018-102891] [Medline: 29871870]
38. Melcher J, Hays R, Torous J. Digital phenotyping for mental health of college students: a clinical review. Evid Based Ment Health 2020 Nov;23(4):161-166 [FREE Full text] [doi: 10.1136/ebmental-2020-300180] [Medline: 32998937]
39. van Gemert Pijnen LJEWC, Nijland N, van LM, Ossebaard HC, Kelders SM, Eysenbach G, et al. A holistic framework to improve the uptake and impact of eHealth technologies. J Med Internet Res 2011 Dec;13(4):e111 [FREE Full text] [doi: 10.2196/jmir.1672] [Medline: 22155738]
40. Kip H, Beerlage-de Jong N, Kelders S, van Gemert-Pijnen L. The CeHRes roadmap. In: eHealth Research, Theory and Development: A Multidisciplinary Approach. London, UK: Routledge; 2024:107-126.
41. Kip H, Keizer J, da Silva MC, Beerlage-de Jong N, Köhle N, Kelders SM. Methods for human-centered eHealth development: narrative scoping review. J Med Internet Res 2022 Jan 27;24(1):e31858 [FREE Full text] [doi: 10.2196/31858] [Medline: 35084359]

XSL•FO
RenderX

42. Laure T, Boffo M, Engels RC, Remmerswaal D. Effectiveness and uptake of a transdiagnostic emotion regulation mobile intervention among university students: protocol for a randomized controlled trial. Internet Interv 2024 Sep;37:100750 [FREE Full text] [doi: 10.1016/j.invent.2024.100750] [Medline: 38827123]
43. Villegas MC, Laure T, Remmerswaal D, Dulloo D, Boffo M. The ROOM app: transdiagnostic mobile intervention for emotion regulation. Conceptualization, design and evaluation. Open Science Framework. 2024 Jan 31. URL: https://osf.io/e9utb/overview?view_only=24f0f05a74074ad89786785cb6960719 [accessed 2024-01-08]
44. Student Wellbeing Monitor General Report on First Assessment Wave The EUR Student Wellbeing Team. Erasmus University Rotterdam. 2020. URL: https://www.eur.nl/en/media/2021-06-studentmonitorgeneralreport2020 [accessed 2022-09-18]
45. Villegas Mejía C, Remmerswaal D, Engels RC, Ludden GD, Boffo M. Macro-engagement in mHealth: exploring user engagement beyond the screen. Digit Health 2024 Jan 23;10:20552076231225591-20552076231225518 [FREE Full text] [doi: 10.1177/20552076231225591] [Medline: 38269371]
46. Yardley L, Spring BJ, Riper H, Morrison LG, Crane DH, Curtis K, et al. Understanding and promoting effective engagement with digital behavior change interventions. Am J Prev Med 2016 Nov;51(5):833-842. [doi: 10.1016/j.amepre.2016.06.015] [Medline: 27745683]
47. Seligman M. Positive psychology, positive prevention, positive therapy. In: Snyder CR, Lopez SJ, editors. Handbook of Positive Psychology. Oxford, England/New York, USA: Oxford University Press; 2002:3-12.
48. Creswell JD. Mindfulness Interventions. Annu Rev Psychol 2017 Jan 03;68(1):491-516. [doi: 10.1146/annurev-psych-042716-051139] [Medline: 27687118]
49. Hayes S, Strosahl K, Wilson K. Acceptance and Commitment Therapy: The Process and Practice of Mindful Change. New York, NY: The Guilford Press; 2012.
50. Neff KD, Hsieh Y, Dejitterat K. Self-compassion, achievement goals, and coping with academic failure. Self and Identity 2005 Jul;4(3):263-287. [doi: 10.1080/13576500444000317]
51. Beck J. Cognitive Behavior Therapy: Basics and Beyond (2nd Edition). New York, USA: The Guildford Press; 2011.
52. Sloan E, Hall K, Moulding R, Bryce S, Mildred H, Staiger PK. Emotion regulation as a transdiagnostic treatment construct across anxiety, depression, substance, eating and borderline personality disorders: a systematic review. Clin Psychol Rev 2017 Nov;57:141-163. [doi: 10.1016/j.cpr.2017.09.002] [Medline: 28941927]
53. Gross J. Emotion regulation: conceptual foundations and empirical foundations. In: Gross J, editor. Handbook of Emotion Regulation. New York, NY: The Guildford Press; 2014:3-24.
54. Burns C. Human-centred design. In: Kip H, Beerlage-de Jong N, van Gemert-Pijnen L, Sanderman R, Kelders S, editors. eHealth Research, Theory and Development: A Multidisciplinary Approach. London, UK: Routledge; 2018:207-227.
55. Laure T, Remmerswaal D, Konigorski S, Engels RCME, Boffo M. Optimization of a mobile transdiagnostic emotion regulation intervention for university students: a micro-randomized trial. Stress Health 2025 Feb 01;41(1):e3507-e3514. [doi: 10.1002/smi.3507] [Medline: 39707816]
56. Laure T, Engels RCME, Remmerswaal D, Spruijt-Metz D, Konigorski S, Boffo M. Optimization of a transdiagnostic mobile emotion regulation intervention for university students: protocol for a microrandomized trial. JMIR Res Protoc 2023 Oct 27;12:e46603 [FREE Full text] [doi: 10.2196/46603] [Medline: 37889525]
57. Cavoukian A. Understanding how to implement privacy by design, one step at a time. IEEE Consumer Electron Mag 2020 Mar 1;9(2):78-82. [doi: 10.1109/mce.2019.2953739]
58. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). EUR-Lex. 2017 Apr 27. URL: http://data.europa.eu/eli/reg/2016/679/oj; [accessed 2024-03-20]
59. European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying down Harmonised Rules on Artificial Intelligence and Amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). EUR-Lex. URL: https://data.europa.eu/eli/reg/2024/1689/oj [accessed 2024-07-22]
60. Sandberg B, Aarikka-Stenroos L. What makes it so difficult? A systematic review on barriers to radical innovation. Industrial Marketing Management 2014 Nov;43(8):1293-1305. [doi: 10.1016/j.indmarman.2014.08.003]
61. Dalgleish T, Black M, Johnston D, Bevan A. Transdiagnostic approaches to mental health problems: current status and future directions. J Consult Clin Psychol 2020 Mar;88(3):179-195 [FREE Full text] [doi: 10.1037/ccp0000482] [Medline: 32068421]
62. Schaeuffele C, Meine LE, Schulz A, Weber MC, Moser A, Paersch C, et al. A systematic review and meta-analysis of transdiagnostic cognitive behavioural therapies for emotional disorders. Nat Hum Behav 2024 Mar 16;8(3):493-509 [FREE Full text] [doi: 10.1038/s41562-023-01787-3] [Medline: 38228727]
63. Berking M, Lukas CA. The affect regulation training (ART): a transdiagnostic approach to the prevention and treatment of mental disorders. Current Opinion in Psychology 2015 Jun;3:64-69. [doi: 10.1016/j.copsyc.2015.02.002]

XSL•FO
RenderX

64. Daros AR, Haefner SA, Asadi S, Kazi S, Rodak T, Quilty LC. A meta-analysis of emotional regulation outcomes in psychological interventions for youth with depression and anxiety. Nat Hum Behav 2021 Oct 20;5(10):1443-1457 [FREE Full text] [doi: 10.1038/s41562-021-01191-9] [Medline: 34545236]
65. Nadeem A, Umer F, Anwar MJ. Emotion regulation as predictor of academic performance in university students. JPAP 2023 Mar 04;4(1):20-33. [doi: 10.52053/jpap.v4i1.157]
66. Im S, Kahler J. Evaluating the empirical evidence for three transdiagnostic mechanisms in anxiety and mood disorders. J Gen Psychol 2022 Oct 08;149(2):232-257. [doi: 10.1080/00221309.2020.1828252] [Medline: 33030123]
67. Koutra K, Pantelaiou V, Mavroeides G. Breaking barriers: unraveling the connection between mental health literacy, attitudes towards mental illness, and self-stigma of psychological help-seeking in university students. Psychology International 2024 May 02;6(2):590-602. [doi: 10.3390/psycholint6020035]
68. Oliveira C, Pacheco M, Borges J, Meira L, Santos A. Internet-delivered cognitive behavioral therapy for anxiety among university students: a systematic review and meta-analysis. Internet Interv 2023 Mar;31:100609 [FREE Full text] [doi: 10.1016/j.invent.2023.100609] [Medline: 36873307]
69. Baumgartner Ph DN, Schneider Ph DR. A randomized controlled trial of mindfulness-based stress reduction on academic resilience and performance in college students. J Am Coll Health 2023 Aug 16;71(6):1916-1925. [doi: 10.1080/07448481.2021.1950728] [Medline: 34398703]
70. Baumel A, Fleming T, Schueller SM. Digital micro interventions for behavioral and mental health gains: core components and conceptualization of digital micro intervention care. J Med Internet Res 2020 Oct 29;22(10):e20631 [FREE Full text] [doi: 10.2196/20631] [Medline: 33118946]
71. Elefant AB, Contreras O, Muñoz RF, Bunge EL, Leykin Y. Microinterventions produce immediate but not lasting benefits in mood and distress. Internet Interv 2017 Dec;10:17-22 [FREE Full text] [doi: 10.1016/j.invent.2017.08.004] [Medline: 29270366]
72. Meinlschmidt G, Lee J, Stalujanis E, Belardi A, Oh M, Jung EK, et al. Smartphone-based psychotherapeutic micro-interventions to improve mood in a real-world setting. Front Psychol 2016 Jul 28;7:1112-1111 [FREE Full text] [doi: 10.3389/fpsyg.2016.01112] [Medline: 27516747]
73. Thomas EBK, Sagorac Gruichich T, Maronge JM, Hoel S, Victory A, Stowe ZN, et al. Mobile acceptance and commitment therapy with distressed first-generation college students: microrandomized trial. JMIR Ment Health 2023 May 15;10:e43065 [FREE Full text] [doi: 10.2196/43065] [Medline: 37184896]
74. Bunge EL, Walker H, Hunt G, Leykin Y, Kim E. The impact of a behavioral activation online micro intervention on mood and activity level. J Technol Behav Sci 2023 Apr 12;8(2):196-204. [doi: 10.1007/s41347-023-00314-5]
75. D'Alfonso S, Lederman R, Bucci S, Berry K. The digital therapeutic alliance and human-computer interaction. JMIR Ment Health 2020 Dec 29;7(12):e21895 [FREE Full text] [doi: 10.2196/21895] [Medline: 33372897]
76. Valentine L, D'Alfonso S, Lederman R. Recommender systems for mental health apps: advantages and ethical challenges. AI Soc 2022 Jan 17:1-12. [doi: 10.1007/s00146-021-01322-w] [Medline: 35068708]
77. Smith AC, Fowler LA, Graham AK, Jaworski BK, Firebaugh M, Monterubio GE, et al. Digital overload among college students: implications for mental health app use. Social Sciences 2021 Jul 21;10(8):279. [doi: 10.3390/socsci10080279]
78. Steele RG, Hall JA, Christofferson JL. Conceptualizing digital stress in adolescents and young adults: toward the development of an empirically based model. Clin Child Fam Psychol Rev 2020 Mar 7;23(1):15-26. [doi: 10.1007/s10567-019-00300-5] [Medline: 31392451]
79. Seemiller C, Grace M. Generation Z Goes to College. San Francisco, CA: Jossey-Bass; 2016.
80. Gollwitzer P, Sheeran P. Implementation intentions and goal achievement: a meta-analysis of effects and processes. Advances in Experimental Social Psychology 2006;38:69-119. [doi: 10.1016/s0065-2601(06)38002-1]
81. Bieleke M, Keller L, Gollwitzer PM. If-then planning. European Review of Social Psychology 2020 Dec 10;32(1):88-122. [doi: 10.1080/10463283.2020.1808936]
82. Looyestyn J, Kernot J, Boshoff K, Ryan J, Edney S, Maher C. Does gamification increase engagement with online programs? A systematic review. PLoS One 2017;12(3):e0173403 [FREE Full text] [doi: 10.1371/journal.pone.0173403] [Medline: 28362821]
83. Toups DP, Crenshaw N, Wehbe R, Tondello G, Nacke L. The collecting itself feels good: towards collection interfaces for digital game objects. In: Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play. New York, NY: Association for Computing Machinery (ACM); 2016 Presented at: Annual Symposium on Computer-Human Interaction in Play; Oct 16-19, 2016; Austin, TX, USA p. 276-290.
84. Nguyen A. Why do students engage in gamification? An exploratory study using means-end chains. Journal of Theoretical and Applied Information Technology 2021;99(23):5783-5793 [FREE Full text]
85. Zichermann G, Cunningham C. Game mechanics: designing for engagement (I, II). In: Treseler M, editor. Gamification by Design: Implementing Game Mechanics in Web and Mobile Apps. Sebastopol, CA: O'Reilly Media, Inc; 2011:35-67.
86. Top mental health and prayer apps fail spectacularly at privacy, security. Mozilla Foundation. 2022. URL: https://foundation.mozilla.org/en/blog/top-mental-health-and-prayer-apps-fail-spectacularly-at-privacy-security/ [accessed 2024-10-03]

XSL•FO
RenderX

87. Blandford A, Gibbs J, Newhouse N, Perski O, Singh A, Murray E. Seven lessons for interdisciplinary research on interactive digital health interventions. Digit Health 2018;4:2055207618770325 [FREE Full text] [doi: 10.1177/2055207618770325] [Medline: 29942629]
88. Parker L, Halter V, Karliychuk T, Grundy Q. How private is your mental health app data? An empirical study of mental health app privacy policies and practices. Int J Law Psychiatry 2019 May;64:198-204. [doi: 10.1016/j.ijlp.2019.04.002] [Medline: 31122630]
89. Nahum-Shani I, Smith SN, Spring BJ, Collins LM, Witkiewitz K, Tewari A, et al. Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. Ann Behav Med 2016 Sep 23:446-462. [doi: 10.1007/s12160-016-9830-8] [Medline: 27663578]
90. Kaissis GA, Makowski MR, Rückert D, Braren RF. Secure, privacy-preserving and federated machine learning in medical imaging. Nat Mach Intell 2020 Jun 08;2(6):305-311. [doi: 10.1038/s42256-020-0186-1]
91. Vierlboeck M, Gövert K, Trauer J, Lindemann U. Budgeting for Agile product development. 2019 Jul 26 Presented at: 22nd International Conference on Engineering Design; August 4-8, 2019; Delft, The Netherlands p. 2169-2178. [doi: 10.1017/dsi.2019.223]
92. Gustavsson T. Benefits of Agile project management in a non-software development context: a literature review. Riga, Latvia: University of Latvia/Professional Association of Project Managers (Latvia); 2016 Presented at: Project Management Development – Practice and Perspectives: Fifth International Scientific Conference on Project Management in the Baltic Countries; April 14-15, 2016; Riga, Latvia. [doi: 10.22364/pmdpp.2016]
93. Specht A, Crowston K. Interdisciplinary collaboration from diverse science teams can produce significant outcomes. PLoS ONE 2022 Nov 29;17(11):e0278043. [doi: 10.1371/journal.pone.0278043]
94. Wohlin C, Aurum A, Angelis L, Phillips L, Dittrich Y, Gorschek T, et al. The success factors powering industry-academia collaboration. IEEE Softw 2012 Mar;29(2):67-73. [doi: 10.1109/ms.2011.92]
95. O'Dwyer M, Filieri R, O'Malley L. Establishing successful university–industry collaborations: barriers and enablers deconstructed. J Technol Transf 2022 Mar 30;48(3):900-931. [doi: 10.1007/s10961-022-09932-2]

Abbreviations

ACT: acceptance and commitment therapy
AI: artificial intelligence
CBT: cognitive behavioral therapy
CeHRes: Centre for eHealth Research
DBCI: digital behavior change intervention
ER: emotion regulation
EU: European Union
EUR: Erasmus University Rotterdam
GDPR: General Data Protection Regulation
OSF: Open Science Framework
UX: user experience

Edited by A Mavragani; submitted 17.06.24; peer-reviewed by C Peuters; comments to author 30.08.24; revised version received 06.10.24; accepted 05.06.25; published 17.12.25.

Please cite as:
Laure T, Villegas Meija C, Remmerswaal D, Dulloo D, Van der Hallen R, Mayer B, Littel M, Dierckx B, Legerstee JS, Engels RCME, Boffo M
ROOM to Grow, a Mobile Well-Being Intervention for University Students: Overview of the Design Process and Outcomes
JMIR Form Res 2025;9:e63325
URL: https://formative.jmir.org/2025/1/e63325
doi:10.2196/63325
PMID:

©Tajda Laure, Camila Villegas Meija, Danielle Remmerswaal, Djameela Dulloo, Ruth Van der Hallen, Birgit Mayer, Marianne Littel, Bram Dierckx, Jeroen S Legerstee, Rutger C M E Engels, Marilisa Boffo. Originally published in JMIR Formative Research (https://formative.jmir.org), 17.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete

XSL•FO
RenderX

bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Performance of DeepSeek-R1, ChatGPT (GPT-o3-mini), and Gemini 2.0 Flash on German Medical Multiple-Choice Questions: Comparative Evaluation

Annika Meyer[1], Dr med; Yassin Karay[2], Dr rer med; Andrea U Steinbicker[1], Prof Dr Med; Thomas Streichert[3], Prof Dr Med; Remco Overbeek[1], Dr med

[1]Department of Anesthesiology and Operative Intensive Care, Faculty of Medicine and University Hospital, University Hospital Cologne, Cologne, Germany

[2]Dean's Office for Student Affairs, Faculty of Medicine, University Hospital Cologne, Cologne, Germany

[3]Institute for Clinical Chemistry, Faculty of Medicine and University Hospital, University Hospital Cologne, Cologne, Germany

Corresponding Author:
Annika Meyer, Dr med
Department of Anesthesiology and Operative Intensive Care
Faculty of Medicine and University Hospital
University Hospital Cologne
Kerpener Str. 62
Cologne, 50937
Germany
Email: annika.meyer1@uk-koeln.de

Abstract

Background: Despite the transformative potential of artificial intelligence (AI)–based chatbots in medicine, their implementation is hindered by data privacy and security concerns. DeepSeek offers a conceivable solution through its capability for local offline operations. However, as of 2025, it remains unclear whether DeepSeek can achieve an accuracy comparable to that of conventional, cloud-based AI chatbots.

Objective: This study aims to evaluate whether DeepSeek, an AI-based chatbot capable of offline operation, achieves answer accuracy on medical multiple-choice questions (MCQs) comparable to that of leading chatbots (ie, ChatGPT and Gemini) on German medical MCQs, thereby assessing its potential as a privacy-preserving alternative for clinical use.

Methods: A total of 200 interdisciplinary MCQs from the German Progress Test Medicine were administered to ChatGPT (GPT-o3-mini), DeepSeek (DeepSeek-R1), and Gemini (Gemini 2.0 Flash). Accuracy was defined as the proportion of correctly solved questions. Overall differences among the 3 models were tested with the Cochran Q test, while pairwise comparisons were conducted using the McNemar test. Subgroup analyses were performed by medical domain (Fisher exact test) and question length (Wilcoxon rank-sum test). An a priori power analysis indicated a minimum sample size of 195 questions.

Results: All 3 chatbots surpassed the conventional passing threshold of 60%, with accuracies of 96% (192/200) for DeepSeek, 94% (188/200) for Gemini, and 92.5% (185/200) for ChatGPT. The overall difference among models was not statistically significant (P=.10) nor were pairwise comparisons. However, incorrect responses were significantly associated with longer question length for DeepSeek (P=.049) and ChatGPT (P=.04) but not for Gemini. No significant differences in performance were observed across clinical versus preclinical domains or medical specialties (all P>.05).

Conclusions: Overall, DeepSeek demonstrates outstanding performance on German medical MCQs comparable to the widely used chatbots ChatGPT and Gemini. Similar to ChatGPT, DeepSeek's performance declined with increasing question length, highlighting verbosity as a persistent challenge for large language models. While DeepSeek's offline capability and lower operational costs are advantageous, its safe and reliable application in clinical contexts requires further investigation.

(JMIR Form Res 2025;9:e77357) doi:10.2196/77357

KEYWORDS

ChatGPT; Gemini; DeepSeek; large language model; chatbots; artificial intelligence

XSL•FO
RenderX

Introduction

In recent years, artificial intelligence (AI) has experienced a remarkable surge in development and adoption [1]. Although applications in medicine are not entirely new, with substantial investments in health care–related AI already initiated nearly a decade ago, the capabilities of AI have advanced considerably [1,2]. These developments offer opportunities for research, automation of routine tasks, and diagnostic support while simultaneously raising persistent ethical, governance, and regulatory challenges [1,3].

A major catalyst for the rapid uptake of AI has been the publication of AI-based chatbots on the World Wide Web, which have substantially lowered barriers related to the usability and accessibility of sophisticated AI systems [4]. Their applications in medicine and research are increasingly under investigation. For example, ChatGPT (OpenAI), introduced in 2022, demonstrates strong performance in medical state examinations, can generate medical reports and radiology documentation, and supports medical programming [5-7]. In addition, these chatbots exhibit considerable linguistic fluency and empathy in response to patients' inquiries, although answer quality can vary across medical specialties [8,9].

Despite their promising capabilities, relevant concerns regarding data privacy and security impede the clinical adoption of online AI-based chatbots in hospitals [4]. One proposed solution is the recently introduced offline-capable chatbot, DeepSeek (Hangzhou DeepSeek Artificial Intelligence Basic Technology Research Co, Ltd) [10]. Notably, initial analyses already suggest that DeepSeek excels in various benchmarks, such as mathematics-based assessments, surpassing previous chatbot iterations [11,12]. Developed by a small team of computer scientists in 2025 [13,14], DeepSeek gained immediate prominence due to its efficient computational design that requires fewer graphical processing units. Operational costs are thereby reduced, and user fees and the carbon footprint are lowered [12-15]. Moreover, its partially open-source nature might further encourage ongoing innovation of AI [10,13,14].

However, empirical evidence on the accuracy of answers on medical multiple-choice questions (MCQs) compared with established chatbots remains scarce, underscoring the need for systematic evaluation [10,12]. Therefore, this study aimed to systematically evaluate DeepSeek's performance in applying medical knowledge by comparing it with ChatGPT and Gemini (Google DeepMind) on 200 MCQs from the German Progress Test Medicine (PTM).

Methods

Berlin Progress Test

The PTM used in this study is a knowledge test designed at the Charité in Berlin for students of human medicine. It is taken by approximately 12,000 students from 19 universities in Germany, Austria, and Switzerland through their medical studies. The PTM consists of 200 interdisciplinary MCQs at the graduate level and is designed to provide students with objective feedback on their personal growth in knowledge over the course of their studies. The questions cover a broad spectrum of domains, including internal medicine, surgery, pediatrics, obstetrics and gynecology, psychiatry, anesthesiology, radiology, laboratory medicine, and the basic sciences (eg, anatomy, physiology, and biochemistry), ensuring balanced representation of both clinical and preclinical content [16]. By covering both clinical and preclinical domains, the test provides a balanced and integrated assessment of medical knowledge, which was the rationale for its selection in this study. Moreover, the test results and, in particular, the knowledge gained per semester have proven to be suitable criteria for predicting academic success with regard to the German state examinations [17]. For this study, we used the 51st PTM, published in October 2024, which had a mean discrimination index of 0.45 and a Cronbach α of 0.98 [16]. Each question was categorized by a physician according to the subject area and study phase (eg, clinical phase and preclinical phase).

Data Collection

To address our research question, we evaluated the performance of 3 chatbots, ChatGPT (GPT-o3-mini), DeepSeek (DeepSeek-R1), and Gemini (Gemini 2.0 Flash), using 200 MCQs from the PTM 51 published in October 2024 between February 21, 2025, and March 4, 2025. These chatbots were selected because they represent 3 leading approaches to large language models in medicine: ChatGPT as the widely used benchmark, Gemini as a major proprietary competitor, and DeepSeek as a novel offline-capable alternative with potential privacy advantages. The specific versions were chosen because they were the most recent publicly accessible releases and free of cost at the time of data collection, reflecting the default user-facing performance available in February 2025.

All questions were used with formal permission, presented in German, and included without modification or exclusion. The number of the included 200 questions was determined based on a sample size calculation for the McNemar test, assuming a Cohen d of 0.28 and a statistical power of 80%, yielding a required sample size of 195. In accordance with the findings of Alfertshofer et al [18], the word count of each question was subsequently determined.

Each unaltered question was entered into the publicly accessible default web-based interface of each chatbot without supplementary prompting or user-directed modifications. Browsing and integrated tools were left enabled, reflecting the standard user-facing functionality of each system. To avoid memory effects or response contamination, each question was submitted in a separate, newly initiated chat session. On occasions where the chatbot failed to generate a response initially, the query was reinitiated.

In addition to the quantitative analyses, 2 physicians independently conducted an exploratory qualitative review of discrepancies between the officially defined correct answers and the responses generated by the chatbots. Potential reasons for discrepancies were derived through independent review and subsequent discussion, without the application of a predefined coding framework, aimed at providing illustrative examples of chatbot limitations.

XSL•FO
RenderX

Statistical Analysis

Statistical analyses were performed using "R" (version 2025.09.1+401: R Foundation for Statistical Computing) [19]. Sample size calculations were performed with the *pwr* package [20]. Data wrangling and analysis were carried out using *rio* [21], *tidyverse* [22], *gtsummary* [23], *rstatix* [24], *labelled* [25], and *fastDummies* [26], while data visualization was accomplished using the *tidyverse* [22], *RColorBrewer* [27], and *cowplot* packages [28]. ChatGPT was used to facilitate statistical programming, while all AI-generated content was critically reviewed by a human. Categorical variables were summarized by absolute and relative frequencies, whereas continuous variables were described using medians and IQRs. Normality was assessed using the Shapiro-Wilk test. Differences in chatbot performance were evaluated using the McNemar test for paired categorical data with Bonferroni correction and the Fisher exact test for unpaired categorical data, while the Wilcoxon rank-sum test was applied for continuous data. A binomial test was applied to test the performance of the chatbots against the threshold of 60%. A *P* value of <.05 was considered statistically significant (Multimedia Appendices 1 and 2).

Ethical Considerations

As this study was limited to medical state examination questions and publicly available results, no research involving human participants was conducted. In accordance with the guidelines of the ethics committee of the University of Cologne, ethics approval was therefore not required [29].

Results

Question Characteristics

A total of 200 original medical progress test questions were used to evaluate the 3 chatbots, with an average length of 55 (IQR 40-74) words. Most (177/200, 88.5%) questions focused on clinical knowledge, with nearly one-quarter (47/200, 23.5%) specifically addressing internal medicine (Table 1).

Table 1. Comparison of DeepSeek, Gemini, and ChatGPT in the Progress Test Medicine.

Characteristic	All questions (N=200)	DeepSeek				Gemini				ChatGPT			
		True (n=192)	False (n=8)	*P* value	Q value	True (n=188)	False (n=12)	*P* value	Q value	True (n=185)	False (n=15)	*P* value	Q value
Word count, median (IQR)	55 (40, 74)	54 (39, 73)	70 (60, 83)	.049[a]	0.098	54 (39, 74)	65 (56, 74)	.12[a]	0.175	53 (39, 73)	64 (55, 87)	.04[a]	0.082
Phase, n (%)				.60[b]	0.72			.15[b]	1.75			>.99[b]	>.99
Clinical phase	177 (89)	169 (88)	8 (100)			168 (89)	9 (75)			163 (88)	14 (93)		
Preclinical phase	23 (12)	23 (12)	0 (0)			20 (11)	3 (25)			22 (12)	1 (6.7)		
Specialty, n (%)				>.99[b]	>.99			.49[b]	.498			.15[b]	0.22
Internal medicine	47 (24)	45 (23)	2 (25)			46 (24)	1 (8.3)			45 (24)	2 (13)		
Surgery	24 (12)	23 (12)	1 (13)			23 (12)	1 (8.3)			20 (11)	4 (27)		
Others	129 (65)	124 (65)	5 (63)			119 (63)	10 (83)			120 (65)	9 (60)		

[a]Wilcoxon rank-sum test.

[b]Fisher exact test.

Accuracy of the Chatbots

All chatbot models significantly exceeded the predefined performance threshold of 60% (*P*<.001 for all comparisons). Accuracy was 96% (95% CI 92.9%-100%) for DeepSeek, 94% (95% CI 90.5%-100%) for Gemini, and 92.5% (95% CI 88.7%-100%) for ChatGPT. Accuracy differences among the 3 chatbots were small and not statistically significant, and there was no difference in pairwise comparison after Bonferroni adjustment (Multimedia Appendix 3; Figures 1A and 1B). DeepSeek correctly answered MCQs with a median word count of 54 (IQR 39-73), whereas incorrectly answered questions had a significantly higher median word count of 70 (IQR 60-83; *P*=.049). A similar pattern was also observed for ChatGPT (*P*=.04) but not for Gemini (Figure 1C). No accuracy variations were found across medical specialties or clinical or preclinical categorization (Figures 1A and 1B; Table 1).

XSL•FO
RenderX

Figure 1. Comparison of ChatGPT, Gemini, and DeepSeek in answering German-language medical multiple-choice questions. (A) Shows the performance categorized by clinical and preclinical phases. (B) Shows the performance on internal medicine questions. The dashed blue line indicates the conventional passing threshold of 60% in the German Progress Test Medicine. (C) Shows a box plot illustrating the relationship between question word count and accuracy.

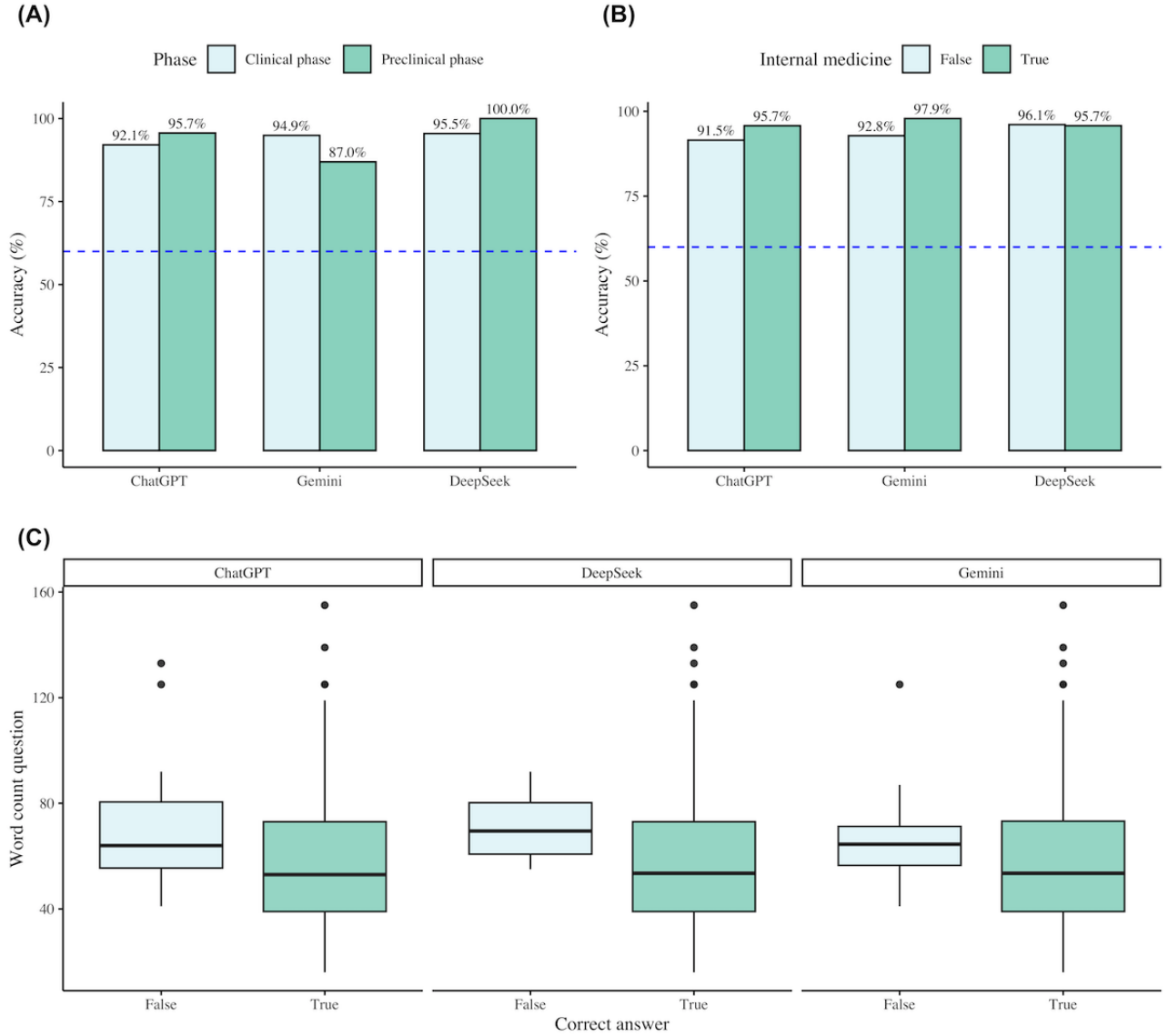

Exploratory Qualitative Assessment of Chatbot Answers

Despite their high accuracy, all 3 chatbots occasionally produced highly plausible but incorrect explanations. In the context of large language models, such errors are commonly referred to as *hallucinations*, meaning the confident generation of factually incorrect or fabricated information that is not supported by the input or external knowledge [30]. For example, in one case, Gemini fabricated an incorrect response label for a multiple-choice answer. In another case, both DeepSeek and ChatGPT proposed a diagnosis that was not among the available options. The clinical scenario described recurrent morning stiffness with enlargement of the distal interphalangeal joints and proximal interphalangeal joints but without redness or swelling. Despite the absence of *arthrosis* from the listed answer choices, both DeepSeek and ChatGPT chose it as the most likely cause. In contrast, the correct answer provided by the questionnaire's designer was rheumatoid arthritis.

Moreover, some discrepancies among the chatbots appeared to mirror inconsistencies in the medical literature itself [31,32]. For example, in explaining metamizole-induced agranulocytosis, ChatGPT emphasized antibody-mediated granulocyte destruction, whereas DeepSeek and Gemini attributed the condition to direct toxic effects on the bone marrow.

Discussion

Principal Findings

In this evaluation of 200 German-language medical MCQs spanning both preclinical and clinical domains, DeepSeek, ChatGPT, and Gemini achieved very high overall accuracies (93%-96%), indicating that an offline - deployable model, such as DeepSeek, can match leading chatbots in core medical reasoning tasks [10]. Moreover, the mean difficulty index of the PTM 51 was 0.35, corresponding to an average student accuracy of 35% on the same questions [16]. Thus, all 3 chatbots substantially outperformed the comparison group of medical

XSL•FO
RenderX

students. Notably, performance for both DeepSeek and ChatGPT declined with increasing question length, underscoring that verbose prompts remain a persistent challenge and echoing previous observations that longer MCQs amplify opportunities for error [18]. Sporadic yet plausible *hallucinations* and *out - of - options* answers—well - documented safety concerns in large language models—were observed across all 3 chatbots [8].

High Performance of AI-Based Chatbots in MCQs

Compared to earlier ChatGPT versions on the PTM, our results illustrate rapid progress in the previous years [33,34]. Similarly, studies reporting improvements from GPT-3.5 (58%) to GPT-4 (81%) in medical state examinations mirror our findings [5,35]. Moreover, Alfertshofer et al [36] analyzed 1200 medical licensing MCQs and identified question length and language as key determinants of accuracy, mirroring our finding that verbosity negatively impacts performance. Within German-language contexts, Friederichs et al [34] found that ChatGPT answered roughly two-thirds of PTM items correctly and outperformed early-year medical students, consistent with our result that all 3 chatbots surpassed conventional pass thresholds on the PTM 51 item pool. Notably, the accuracies reported in our study exceed those of earlier PTM estimates, underscoring the rapid capability gains since 2023.

Our results also fit into a broader international landscape of benchmarking studies that have consistently reported high but context-dependent performance of generative AI in medical MCQs. DeepSeek-R1, for example, achieved accuracies of 97% on English and Chinese licensing items [1]; 92% to 95% on the Chinese National Medical Licensing Examination, with significant and stable advantages over ChatGPT [2,6]; and more than 90% in oncology [10] and microbiology [11]. In contrast, head-to-head comparisons on the United States Medical Licensing Examination found DeepSeek slightly inferior to ChatGPT-o1 (92% vs 95%) [3], while in ophthalmology board-style examinations, ChatGPT o1 Pro (83.4%) clearly outperformed DeepSeek-R1 (72.5%) [4]. Conversely, in pediatric board preparation questions, DeepSeek-R1 reached 98% accuracy, markedly surpassing ChatGPT-4 (82.7%) [5]. These findings underline that apparent global accuracy masks substantial domain-specific variability, where some models excel in pediatrics or oncology, while others dominate in ophthalmology. Importantly, several studies found minimal overlap in the specific questions missed by different models, suggesting complementary rather than uniform knowledge gaps [5].

Language-specific effects further add nuance to these comparisons. In bilingual ophthalmology MCQs, DeepSeek performed better in Chinese (86.2%) than in English (80.8%), while Gemini and OpenAI models showed weaker robustness across languages [8]. Similar findings in multiyear Chinese National Medical Licensing Examination evaluations confirm DeepSeek's consistent advantage in Chinese [2,6], supporting the view that training corpus composition can strongly shape performance across languages. This resonates with our observation that even in German-language testing, performance is influenced not only by specialty but also by linguistic and structural features of the questions.

Taken together, previous literature and our findings converge on 3 themes. First, the overall performance of state-of-the-art chatbots on medical MCQs is now consistently at or above medical student thresholds, often exceeding 90%. Second, accuracy varies by language, with models such as DeepSeek particularly advantaged in Chinese settings. Third, accuracy differs by domain, with ophthalmology and pediatrics illustrating opposite outcomes across models. Finally, our demonstration that verbosity predicts chatbot errors echoes the hypothesis by Alfertshofer et al [18] that longer questions amplify opportunities for error [35].

Ophthalmology-specific studies likewise found DeepSeek on par with, or superior to, contemporaneous versions of ChatGPT and Gemini [37,38], although 1 study found it lagged behind ChatGPT on pediatric MCQs, suggesting specialty-dependent effects or influences of question format [39]. Indeed, we observed word count to be a critical determinant of chatbot accuracy on medical MCQs, supporting the hypothesis by Alfertshofer et al [18] that longer questions create more opportunities for error.

Implications for Safety and Transparency

Consistent with the literature, hallucinations remained a relevant obstacle for the chatbots in addressing medical MCQs [14,40]. Thus, DeepSeek's *think-aloud* feature, where it exposes intermediate reasoning steps, may help end users detect hallucinations, overgeneralization, and dataset biases, provided these outputs are critically reviewed [12,14,40,41]. In addition, this feature might facilitate the scrutiny of DeepSeek's novel or out-of-scope reasoning.

However, transparency does not inherently mitigate the risk of generating unsafe content. After all, the literature suggests that DeepSeek produces unsafe responses 10 times more often than ChatGPT [42]. While DeepSeek is released under a permissive license with publicly available model weights, sometimes described as partially open source [10,13,14], the very notion of openness in generative AI remains contested. Many so-called open models are only *open-weight*, sharing parameters but withholding training and fine-tuning data, a practice termed *open-washing* [43]. This highlights that openness is best understood as a graded and multidimensional property rather than a binary state.

Furthermore, DeepSeek's operation within government - regulated frameworks [14] highlights the tension between transparency and regulatory compliance. Thus, despite the potential to save time, costs, and personnel resources in clinical and research-related decision-making processes in health care [44], the hope that AI-based chatbots can serve as reliable decision-support tools must be critically questioned at this point in time.

Future Directions

While DeepSeek, ChatGPT, and Gemini all excel on German-language MCQs, persistent issues, such as hallucinations, bias, and unsafe outputs mentioned in the

XSL•FO
RenderX

literature, underscore the need for further refinement of these chatbots. Thus, future research should assess chatbot performance on tasks that demand free-text generation and complex reasoning without predefined answer options in real-world simulated environments. Evaluations should also extend across diverse languages, specialties, and examination formats. In addition, studies need to refine prompt engineering approaches to reduce verbosity-related errors and systematically monitor safety, bias, and regulatory compliance over time while accounting for the threat to reproducibility posed by continuous model updates.

Limitations

Nevertheless, our reliance on medical MCQs may overestimate real-world clinical utility of such AI-based chatbots because such formats cannot capture the nuance of actual patient-clinician interactions [8]. Furthermore, presenting every item in German limits generalizability to other languages, health care settings, and assessment formats. In the literature on ophthalmology MCQs, for instance, DeepSeek's accuracy rose from 81% in English to 86% in Chinese, whereas the accuracies of ChatGPT and Gemini fell from 72% to 75% down to 68% to 71%, a pattern attributed to the higher proportion of Chinese tokens in DeepSeek's training data [38]. This finding challenges the common assumption that multilingual AI models are inherently biased toward the English language [35,45] and highlights how training‐corpus composition can drive language‐specific performance [38]. To address concerns regarding reproducibility, we conducted a stability spot-check on a random subset of 20 questions in September 2025. As the original versions tested between February 21, 2025, and March 4, 2025, were no longer available, we used the most recent publicly accessible default web-based versions at that time (ChatGPT5, DeepSeek [version 3.2], and Gemini 2.5 Flash, all accessed on September 30, 2025). A total of 19 (95%) questions were answered correctly by all models, while 1 (5%) error initially made by ChatGPT (GPT-o3-mini) was now reproduced by DeepSeek (version 3.2). This demonstrates that ongoing model updates can change outcomes, underscoring that our findings are specific to particular model versions and access dates.

Conclusions

In conclusion, DeepSeek matches Gemini and ChatGPT in accuracy on German-language medical MCQs while offering novel insights and a transparent *thinking-aloud* glimpse into its reasoning. Yet, recurring hallucinations and documented biases make expert oversight and critical appraisal indispensable. Furthermore, the literature points to potential safety and regulatory concerns that could outweigh DeepSeek's offline-deployment advantages, such as lower environmental footprint and operational costs. Ultimately, targeted research is needed to delineate DeepSeek's failure modes, rigorously validate its safety and impartiality, and establish best-practice strategies.

Acknowledgments

The authors would like to thank Ari Soleman for proofreading this manuscript.

DeepL and ChatGPT (OpenAI) were used for proofreading, providing translation support, and assisting with the linguistic aspects of this paper. In addition, ChatGPT aided in the statistical programming, with all artificial intelligence–generated output being critically reviewed by the authors.

Funding

The German Research Foundation provided funding for the article processing fee of this work.

Data Availability

The datasets analyzed during this study are not publicly available due to copyright restrictions but are available from the corresponding author on reasonable request.

Authors' Contributions

AM, TS, and YK designed the study. AM and RO collected the data and conducted the exploratory qualitative analysis. AM performed the data analysis and interpretation and drafted the manuscript. TS, RO, YK, and AUS critically reviewed and revised the manuscript.

Conflicts of Interest

AM received speaker support at both the Congress of the German Society for Clinical Chemistry and Laboratory Medicine 2024 and Medica 2024, as well as the Digital Laboratory research award from the German Society for Clinical Chemistry and Laboratory Medicine for her previous work on artificial intelligence–based chatbots. In addition, for another research manuscript, TS and AM were supported by OpenAI's Researcher Access Program and Application Programming Interface. AM and TS received funding from the German Research Foundation for the article processing charges of previous work and will also receive German Research Foundation support for the article processing charges of this manuscript.

XSL•FO
RenderX

Multimedia Appendix 1
Cleaned data for data analysis.
[XLSX File (Microsoft Excel File), 16 KB - formative_v9i1e77357_app1.xlsx]

Multimedia Appendix 2
Code for data analysis.
[PDF File (Adobe PDF File), 222 KB - formative_v9i1e77357_app2.pdf]

Multimedia Appendix 3
Accuracy of Gemini, ChatGPT, and DeepSeek in the German Progress Test Medicine.
[DOCX File , 15 KB - formative_v9i1e77357_app3.docx]

References

1. Abulibdeh R, Celi LA, Sejdić E. The illusion of safety: a report to the FDA on AI healthcare product approvals. PLOS Digit Health 2025 Jun 5;4(6):e0000866. [doi: 10.1371/journal.pdig.0000866] [Medline: 40471897]
2. Amisha, Malik P, Pathania M, Rathaur VK. Overview of artificial intelligence in medicine. J Family Med Prim Care 2019 Jul;8(7):2328-2331 [FREE Full text] [doi: 10.4103/jfmpc.jfmpc_440_19] [Medline: 31463251]
3. Beam AL, Drazen JM, Kohane IS, Leong TY, Manrai AK, Rubin EJ. Artificial intelligence in medicine. N Engl J Med 2023 Mar 30;388(13):1220-1221. [doi: 10.1056/NEJMe2206291] [Medline: 36988598]
4. Thirunavukarasu AJ, Ting DS, Elangovan K, Gutierrez L, Tan TF, Ting DS. Large language models in medicine. Nat Med 2023 Aug 17;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
5. Meyer A, Riese J, Streichert T. Comparison of the performance of GPT-3.5 and GPT-4 with that of medical students on the written German medical licensing examination: observational study. JMIR Med Educ 2024 Feb 08;10:e50965 [FREE Full text] [doi: 10.2196/50965] [Medline: 38329802]
6. Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. J Med Internet Res 2023 Jun 28;25:e48568 [FREE Full text] [doi: 10.2196/48568] [Medline: 37379067]
7. Meyer A, Ruthard J, Streichert T. Dear ChatGPT – can you teach me how to program an app for laboratory medicine? J Lab Med 2024 May;48(5):197-201 [FREE Full text] [doi: 10.1515/labmed-2024-0034]
8. Meyer A, Soleman A, Riese J, Streichert T. Comparison of ChatGPT, Gemini, and Le Chat with physician interpretations of medical laboratory questions from an online health forum. Clin Chem Lab Med 2024 Nov 26;62(12):2425-2434 [FREE Full text] [doi: 10.1515/cclm-2024-0246] [Medline: 38804035]
9. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023 Jun 01;183(6):589-596 [FREE Full text] [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
10. Temsah A, Alhasan K, Altamimi I, Jamal A, Al-Eyadhy A, Malki KH, et al. DeepSeek in healthcare: revealing opportunities and steering challenges of a new open-source artificial intelligence frontier. Cureus 2025 Feb;17(2):e79221. [doi: 10.7759/cureus.79221] [Medline: 39974299]
11. DeepSeek-AI. DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning. ArXiv Preprint posted online on January 22, 2025 [FREE Full text] [doi: 10.48550/arXiv.2501.12948]
12. Peng Y, Malin BA, Rousseau JF, Wang Y, Xu Z, Xu X, et al. From GPT to DeepSeek: significant gaps remain in realizing AI in healthcare. J Biomed Inform 2025 Mar;163:104791 [FREE Full text] [doi: 10.1016/j.jbi.2025.104791] [Medline: 39938624]
13. Poo M. Reflections on DeepSeek's breakthrough. Natl Sci Rev 2025 Mar;12(3):nwaf044. [doi: 10.1093/nsr/nwaf044] [Medline: 40041025]
14. Kayaalp ME, Prill R, Sezgin EA, Cong T, Królikowska A, Hirschmann MT. DeepSeek versus ChatGPT: multimodal artificial intelligence revolutionizing scientific discovery. From language editing to autonomous content generation-redefining innovation in research and practice. Knee Surg Sports Traumatol Arthrosc 2025 May;33(5):1553-1556. [doi: 10.1002/ksa.12628] [Medline: 39936363]
15. Gibney E. China's cheap, open AI model DeepSeek thrills scientists. Nature 2025 Feb;638(8049):13-14. [doi: 10.1038/d41586-025-00229-6] [Medline: 39849139]
16. Progress Test Medizin. URL: https://ptm-dashboard.charite.de/ [accessed 2025-11-18]
17. Karay Y, Schauber SK. A validity argument for progress testing: examining the relation between growth trajectories obtained by progress tests and national licensing examinations using a latent growth curve approach. Med Teach 2018 Nov;40(11):1123-1129. [doi: 10.1080/0142159X.2018.1472370] [Medline: 29950124]
18. Alfertshofer M, Knoedler S, Hoch CC, Cotofana S, Panayi AC, Kauke-Navarro M, et al. Analyzing question characteristics influencing ChatGPT's performance in 3000 USMLE®-style questions. Med Sci Educ 2024 Sep 28;35(1):257-267. [doi: 10.1007/s40670-024-02176-9]

XSL•FO
RenderX

19. R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing. 2022. URL: https://www.r-project.org/ [accessed 2025-11-18]
20. Champely S, Ekstrom C, Dalgaard P, Gill J, Weibelzahl S, Anandkumar A, et al. pwr: basic functions for power analysis. The Comprehensive R Archive Network. 2020. URL: https://cran.r-project.org/web/packages/pwr/pwr.pdf [accessed 2025-11-18]
21. Chan C, Leeper TJ, Becker J, Schoch D. rio: a Swiss-army knife for data I/O. The Comprehensive R Archive Network. URL: https://cran.r-project.org/web/packages/rio/readme/README.html [accessed 2025-11-18]
22. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, et al. Welcome to the Tidyverse. J Open Source Softw 2019;4(43):1686 [FREE Full text] [doi: 10.21105/JOSS.01686]
23. Sjoberg DD, Larmarange J, Curry M, Lavery J, Whiting K, Zabor EC. gtsummary: presentation-ready data summary and analytic result tables. Daniel D. Sjoberg. URL: https://www.danieldsjoberg.com/gtsummary//reference/gtsummary-package.html [accessed 2025-11-18]
24. Kassambara A. rstatix: pipe-friendly framework for basic statistical tests. The Comprehensive R Archive Network. 2025. URL: https://cloud.r-project.org/web/packages/rstatix/index.html [accessed 2025-11-18]
25. Larmarange J, Ludecke D, Wickham H, Bojanowski M, Briatte F. labelled: manipulating labelled data. The Comprehensive R Archive Network. URL: https://cran.r-project.org/web/packages/labelled/index.html [accessed 2025-11-18]
26. Kaplan J. fastDummies: fast creation of dummy (binary) columns and rows from categorical variables. The Comprehensive R Archive Network. 2025 Jul 22. URL: https://cran.r-project.org/web/packages/fastDummies/fastDummies.pdf [accessed 2025-11-18]
27. Neuwirth E. RColorBrewer: ColorBrewer palettes. The Comprehensive R Archive Network. 2022. URL: https://cran.r-project.org/web/packages/RColorBrewer/index.html [accessed 2025-11-18]
28. Wilke CO. cowplot: streamlined plot theme and plot annotations for ggplot2. The Comprehensive R Archive Network. 2025 Jul 7. URL: https://cran.r-project.org/web/packages/cowplot/index.html [accessed 2025-11-18]
29. Köln U. Ethikkommission der medizinischen fakultät der Universität zu Köln. Universität zu Köln. URL: https://medfak.uni-koeln.de/forschung-transfer/translation-i-klinische-forschung/ethikkommission [accessed 2025-09-29]
30. Farquhar S, Kossen J, Kuhn L, Gal Y. Detecting hallucinations in large language models using semantic entropy. Nature 2024 Jun;630(8017):625-630 [FREE Full text] [doi: 10.1038/s41586-024-07421-0] [Medline: 38898292]
31. Tomidis Chatzimanouil MK, Goppelt I, Zeissig Y, Sachs UJ, Laass MW. Metamizole-induced agranulocytosis (MIA): a mini review. Mol Cell Pediatr 2023 Aug 17;10(1):6 [FREE Full text] [doi: 10.1186/s40348-023-00160-8] [Medline: 37589909]
32. Rudin D, Roos NJ, Duthaler U, Krähenbühl S. Toxicity of metamizole on differentiating HL60 cells and human neutrophil granulocytes. Toxicology 2019 Oct 01;426:152254. [doi: 10.1016/j.tox.2019.152254] [Medline: 31356851]
33. Rodrigues Alessi M, Gomes HA, Lopes de Castro M, Terumy Okamoto C. Performance of ChatGPT in solving questions from the progress test (Brazilian National Medical Exam): a potential artificial intelligence tool in medical practice. Cureus 2024 Jul;16(7):e64924. [doi: 10.7759/cureus.64924] [Medline: 39156244]
34. Friederichs H, Friederichs WJ, März M. ChatGPT in medical school: how successful is AI in progress testing? Med Educ Online 2023 Dec;28(1):2220920 [FREE Full text] [doi: 10.1080/10872981.2023.2220920] [Medline: 37307503]
35. Liu M, Okuhara T, Chang X, Shirabe R, Nishiie Y, Okada H, et al. Performance of ChatGPT across different versions in medical licensing examinations worldwide: systematic review and meta-analysis. J Med Internet Res 2024 Jul 25;26:e60807 [FREE Full text] [doi: 10.2196/60807] [Medline: 39052324]
36. Alfertshofer M, Hoch CC, Funk PF, Hollmann K, Wollenberg B, Knoedler S, et al. Sailing the seven seas: a multinational comparison of ChatGPT's performance on medical licensing examinations. Ann Biomed Eng 2024 Jun;52(6):1542-1545 [FREE Full text] [doi: 10.1007/s10439-023-03338-3] [Medline: 37553555]
37. Mikhail D, Farah A, Milad J, Nassrallah W, Mihalache A, Milad D, et al. Performance of DeepSeek-R1 in ophthalmology: an evaluation of clinical decision-making and cost-effectiveness. Br J Ophthalmol 2025 Aug 20;109(9):976-981. [doi: 10.1136/bjo-2025-327360] [Medline: 40701781]
38. Xu P, Wu Y, Jin K, Chen X, He M, Shi D. DeepSeek-R1 outperforms Gemini 2.0 Pro, OpenAI o1, and o3-mini in bilingual complex ophthalmology reasoning. Adv Ophthalmol Pract Res 2025 Aug;5(3):189-195 [FREE Full text] [doi: 10.1016/j.aopr.2025.05.001] [Medline: 40678192]
39. Mondillo G, Colosimo S, Perrotta A, Frattolillo V, Masino M. Comparative evaluation of advanced AI reasoning models in pediatric clinical decision support: ChatGPT O1 vs. DeepSeek-R1. medRxiv Preprint posted online on January 28, 2025 [FREE Full text] [doi: 10.1101/2025.01.27.25321169]
40. Smith J. Daily briefing: the pros and cons of DeepSeek. Online ahead of print. Nature 2025 Jan 30. [doi: 10.1038/d41586-025-00330-w] [Medline: 39890911]
41. Meyer A, Wetsch WA, Steinbicker AU, Streichert T. Through ChatGPT's eyes: the large language model's stereotypes and what they reveal about healthcare. J Med Syst 2025 Feb 05;49(1):20. [doi: 10.1007/s10916-025-02159-2] [Medline: 39907718]
42. Arrieta A, Ugarte M, Valle P, Parejo JA, Segura S. o3-mini vs DeepSeek-R1: which one is safer? ArXiv Preprint posted online on January 30, 2025 [FREE Full text] [doi: 10.48550/arXiv.2501.18438]

XSL•FO
RenderX

43. Liesenfeld A, Dingemanse M. Rethinking open source generative AI: open-washing and the EU AI Act. 2024 Presented at: FAccT '24: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency; June 3-6; Rio de Janeiro, Brazil URL: https://dl.acm.org/doi/10.1145/3630106.3659005 [doi: 10.1145/3630106.3659005]
44. Laymouna M, Ma Y, Lessard D, Schuster T, Engler K, Lebouché B. Roles, users, benefits, and limitations of chatbots in health care: rapid review. J Med Internet Res 2024 Jul 23;26:e56930 [FREE Full text] [doi: 10.2196/56930] [Medline: 39042446]
45. Gabriel N, Bhatia A. Lost in translation: large language models in non-English content analysis. ArXiv Preprint posted online on June 12, 2023 [FREE Full text] [doi: 10.48550/arXiv.2306.07377]

Abbreviations

AI: artificial intelligence
MCQ: multiple-choice question
PTM: Progress Test Medicine

Edited by A Stone; submitted 12.05.25; peer-reviewed by L Mosca, P Sharma; comments to author 26.09.25; revised version received 16.10.25; accepted 16.10.25; published 18.12.25.

Please cite as:
Meyer A, Karay Y, Steinbicker AU, Streichert T, Overbeek R
Performance of DeepSeek-R1, ChatGPT (GPT-o3-mini), and Gemini 2.0 Flash on German Medical Multiple-Choice Questions: Comparative Evaluation
JMIR Form Res 2025;9:e77357
URL: https://formative.jmir.org/2025/1/e77357
doi:10.2196/77357
PMID:

©Annika Meyer, Yassin Karay, Andrea U Steinbicker, Thomas Streichert, Remco Overbeek. Originally published in JMIR Formative Research (https://formative.jmir.org), 18.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Identification of Design Requirements for a Software Application for Use by Clinicians That Collects Acute Stroke Treatment Data During Clinical Workflow: Pilot Study

Adam Forward[1], BEng, MASc; Gizem Koca[1], BSc, MSc; Aymane Sahli[1], PhD; Noreen Kamal[1], BSc, MASc, PhD

Department of Industrial Engineering, Faculty of Engineering, Dalhousie University, Halifax, NS, Canada

Corresponding Author:
Noreen Kamal, BSc, MASc, PhD
Department of Industrial Engineering
Faculty of Engineering
Dalhousie University
PO Box 15000
Halifax, NS
Canada
Phone: 1 902 494 3293
Email: Noreen.Kamal@dal.ca

Abstract

Background: Clinical registries are critical for monitoring processes of care in diseases and driving quality improvements. However, many smaller hospitals lack the required resources to collect the necessary data to contribute to registries.

Objective: This study aims to design and evaluate a data collection tool for acute stroke treatment that streamlines the collection of process data and provides tools to aid clinician users while not interfering with clinical workflow. The evaluation will identify key design requirements that facilitate prospective data collection and add value for clinicians.

Methods: We developed a prototype tool for testing using Figma Pro for use on an iPad. Clinicians were recruited through convenience sampling to test the prototype's use in a small-scale simulated clinical field experiment, during which participant were asked to think aloud and then complete a series of tasks to mimic a mock stroke treatment while inputting the required data into the prototype. Follow-up semistructured interviews were conducted to gain feedback on how the prototype integrated into the workflow and on the aspects of the prototype they felt helped and hindered their use of it. Qualitative data analysis combined review of the experiment recordings to identify the most frequent errors made during the scenario and deductive thematic analysis from the follow-up interviews to determine user needs for the following prototype iteration. The insights from the feedback identified design requirements that were implemented in the iterated design and documented to provide a reference for future product designers.

Results: Three participants were recruited from 2 hospitals between April 18 and June 6, 2024, for the simulated field experiment. The scenario took 10-12 minutes, with 1.2-3.7 minutes spent using the prototype, depending on whether optional features such as the NIHSS (National Institute of Health Stroke Scale) calculator were used. The simple and condensed layout and features such as NIHSS calculators, benchmark metric timers, and the final pop-up summary received the most positive feedback from each participant. Issues identified included small target sizes causing higher error rates, lack of color in important features reducing their visibility, and grouping of mandatory and optional information field layouts leading to a disjointed flow. The key design requirements include prioritizing simple dynamic layouts, sufficient target sizes to prevent errors, useful features with clear visual cues, and prompt data feedback to facilitate seamless integration.

Conclusions: A prospective data collection tool for clinicians to use during stroke treatment can add value for clinicians and, with further testing, can be integrated into workflow. The design requirements identified through this study can provide a basis for streamlining the collection of accurate data while increasing the value of the tool for users and should be considered by future product designers to add value to their software and improve user experience.

(JMIR Form Res 2025;9:e64800) doi:10.2196/64800

KEYWORDS

data collection; medical technology; simulated field test; software; user interface design

XSL•FO
RenderX

Introduction

Clinical registries are databases that hospitals use to monitor disease and delivery of care to better understand outcomes and drive improvements. They aggregate data containing patient demographics, medical history, hospital processes, and outcomes of care to measure the quality of care and guide improvements. The use of clinical registries has been shown to improve the quality of care, especially in chronic diseases such as cancer and diabetes [1], but they are also beneficial in acute diseases such as stroke [2].

Multiple stroke registries have demonstrated success in improving treatment processes. One of the largest clinical registries used in stroke treatment is Get With the Guidelines–Stroke in the United States, which combines registry data collection and feedback with hospital toolkits, stakeholder meetings, and collaborative workshops to guide improvements in enrolled hospitals and has significantly increased the rates of patients receiving treatment while ensuring that adverse events remain low [3]. Other large-scale stroke registries that have been associated with improved quality of care include the Registry of Stroke Care Quality in Europe [4,5] and the Australian Stroke Clinical Registry [6]. Finally, in Canada, the current national clinical registry for stroke is OPTIMISE (Optimising Patient Treatment in Major Ischemic Stroke With Endovascular Thrombectomy) [7,8], which records the care path of patients that are treated with endovascular thrombectomy.

For a stroke registry to successfully highlight the processes of care, it must have accurate data so stakeholders can make informed decisions on how to improve their processes. Data collection for stroke is typically conducted retrospectively through chart reviews; however, prospective data collection for stroke has been practiced in various centers [9]. Both retrospective and prospective data collection methods are resource intensive and require staff dedicated to either abstracting or collecting the data, respectively. Because of the resource-intensive nature of data gathering and the time sensitivity necessary to treat a stroke, it is difficult for smaller hospitals with fewer clinicians to contribute to clinical registries.

In Canada, the stroke centers currently contributing data to OPTIMISE are comprehensive stroke centers (CSCs) [7,8]. These centers are frequently larger academic teaching hospitals with the resources available to collect stroke treatment data. However, primary stroke centers (PSCs), which are capable of treating patients with thrombolysis only, are typically small community and rural hospitals. These hospitals make up most stroke treatment centers in Canada; yet, most are unable to contribute data to OPTIMISE under the current system data collection for stroke registries. Many PSCs do not have the resources required to retrospectively abstract stroke treatment data. Additionally, prospective data collection at a small hospital could risk taking time from patient care during the treatment process to input data without dedicated staff.

Novel methods to streamline data collection for stroke treatment using software applications has been studied [10,11], and while there has been success in reducing treatment times, multiple issues have arisen related to data completion and low clinician user satisfaction due to the added work to existing processes [11,12]. The ability to collect accurate and complete data while minimizing the added workload and improving user satisfaction is crucial for prospective registries to succeed in sustainably driving quality improvement initiatives.

There are known barriers to the adoption of data collection technologies in health care settings, including cost, technical training, time, usefulness, and personnel availability [13,14]. By reducing these barriers, adoption of novel clinical technology can increase through improvements in user experience and accessibility for data collection. The use of software to aid in prospectively collecting data during acute stroke treatment can facilitate contributing data to clinical registries. However, there is currently an unmet need to understand the design of a tool that can aid in the prospective collection of data during the acute stroke treatment process.

There is a knowledge gap in understanding the design requirements of a software interface, such as the information layout, interaction styles for inputting information, desired features, and feedback mechanisms that facilitate seamless prospective data collection for clinician users in the acute stroke treatment process and minimize barriers to clinician use. While existing mobile software has been tested and barriers to technology adoption have been studied, there has been no analysis of the specific features and barriers that affect clinician users' ability and satisfaction in integrating a prospective data collection software into the stroke treatment workflow.

This study proposes a prototype data collection tool that is designed, validated, and evaluated by clinician users to specify the key design requirements for collecting data during the workflow of the acute stroke treatment process. The aim of the prototype is to achieve 2 main goals: first, to allow seamless and accurate collection of prospective stroke treatment data, and second, to provide features within the prototype that offer value to clinician users. If these 2 conditions are met, it is anticipated that clinicians will be more open to adopting the technology and integrating it into their stroke treatment operations.

Methods

Overview

This research is an exploratory study anchored in the design thinking process. The process includes 5 stages: empathize, define, ideate, prototype, and test. The empathize and define stages of design thinking were completed in a previous study through the completion of semistructured interviews [15]. This study focuses on the ideation, prototyping, and testing of the data collection prototype.

The first iteration of the prototype used a combination of established usability rules, existing literature on health care software, results from previously conducted semistructured interviews [15] analyzing differences in the processes of stroke centers across Canada, and expert opinions from members of the Canadian Stroke Consortium, a group of stroke professionals who agreed on the variables required for the stroke registry.

XSL•FO
RenderX

The prototype was designed to incorporate data capture for all key data metrics agreed on by the Canadian Stroke Consortium, and it was then tested by clinician users who provided feedback to identify key design requirements to improve the tool's usability and utility for a second iteration. These requirements were then integrated into the prototype for future testing.

Ideation

Key considerations when ideating this prototype included the information to be collected in the software, required features, and the layout and organization of information. The prototype was first developed as low-fidelity pen-and-paper designs and then upgraded to a medium-fidelity prototype using Figma.

Preidentified Requirements

The information for the prototype corresponds with established performance benchmarks in stroke treatment [16,17]. The required information was determined by members of the Canadian Stroke Consortium. Two variable lists were created based on the stroke center classification (PSCs and CSCs) and were used to design the information architecture and determine the structure of the prototype. The complete variable lists can be found in Multimedia Appendix 1.

The previous study identified features that clinicians stated would facilitate use of the prototype and aid in seamless integration with workflow. Five main aids were discussed as adding value for clinicians in stroke treatment: a calculator to determine the National Institute of Health Stroke Scale (NIHSS), inclusion-exclusion criteria to determine a patient's eligibility for treatment, a summary of the information collected, a "now" button to collect process times, and 2 on-screen timers recording the time since patient arrival and benchmark times.

Information Architecture

An information architecture (IA) was developed to define the layout and organization of the prototype and is provided in Multimedia Appendix 2 (Figure S1). The main wireframe contains the information fields on a single page; thus, as opposed to the standard use for navigation, this IA highlights the content and organization of the main page. The IA was organized into 3 sections: patient and stroke information, hospital process, and treatment process. Two sub-IA's were included to show the structure of the NIHSS calculator and the inclusion-exclusion criteria for treatment. The sub-IA's (Figures S2 and S3 in Multimedia Appendix 2) highlight the order of the information given in the featured tools and the terminology for each information field.

Prototyping

The foundation of the design was anchored in Nielsen's [18] usability heuristics. A stroke treatment data collection software naturally requires speed and learnability, so some heuristics were prioritized in the initial design. For example, 3 of Nielsen's heuristics prioritized in the design included "match between the system and the real world," "flexibility and efficiency of use," and "aesthetic and minimalist design." These heuristics were most critical in minimizing the user's cognitive load and ensuring they would not be overwhelmed while inputting data into the software.

Pen and Paper Designs

The low-fidelity prototype was created using the pen-and-paper software called Concepts (TopHatch Inc). The design considered the data fields, features, required number of screens, and navigation. The pen-and-paper designs were developed by AF and reviewed by NK.

Medium Fidelity Prototype

The final pen-and-paper design was upgraded to a medium-fidelity prototype using Figma Pro (Figma Inc; 2016). The medium-fidelity prototype was built for use on an iPad to test how the interactivities and modality integrated into the clinical workflow for users. The prototype contained working links and conditional variables that updated in real time to act as a properly functioning application.

Testing

Recruitment

The research study was conducted at Dalhousie University in Nova Scotia, Canada. The geographic location limited the clinician sample pool, as the province has a small population with only 1 CSC and 2 stroke-trained neurologists. Additionally, the research team had to ensure that each participant completed the experiment under the same conditions and with the same equipment, so participation was required to be in person. A participation invitation was sent to any prospective participant who had a direct role in the acute stroke treatment process and was within driving distance of the research laboratory where the study took place. Two hospitals—a CSC and a thrombolytic-capable center (which was often bypassed due to its proximity to the CSC)—were within a reasonable driving distance of the research laboratory. There were no additional screenings for eligibility, as the sampling pool was already limited by geographic location, participant expertise, and the required time to participate. Because of these factors, convenience sampling was used to recruit clinician participants. Participants were invited by email to participate in the user testing session.

Simulated Scenario

The user testing environment was prepared to ensure that the prototype was evaluated in a scenario that best mimicked the stroke treatment process with respect to the space and equipment available while minimizing any human risk. The simulated environment was prepared to ensure that participants used the prototype while moving around, carrying objects, and speaking to people but did not replicate other aspects of an emergency department, such as background noise or navigating corridors. A single room was set up with arrows to guide the participant to each area. The room was divided into areas related to hospital departments involved in acute stroke treatment. The scenario required a set of preprepared materials, including a mock patient chart, a prearrival form, and a mock thrombolytic syringe. Two volunteers participated as actors and took different roles throughout the scenario, including the roles of paramedic, emergency nurse, computed tomography (CT) technologist, and interventional radiologist. The actors were provided one of two unique scripts, depending on whether the participant was from a PSC or a CSC, to communicate information to the participant

XSL•FO
RenderX

in a manner similar to how they would during a real stroke. Participants were required to complete all mandatory data fields during the scenario.

The simulated scenario was recorded using a Meeting Owl camera (Owl Labs; 2014), allowing for a 360° view of the room. The scenario and follow-up interviews were audio- and video-recorded. Automatic transcription software was used to record the participants' feedback. Participants were asked to "think aloud" as they completed the simulated scenario, with reminders given throughout as necessary.

Qualitative Feedback

Follow-up one-on-one interviews were conducted with the participants to obtain their feedback on the prototype. An interview guide was used to ensure consistency in topics discussed with participants, and to minimize bias by ensuring each question was asked impartially. The interview guide is provided in Multimedia Appendix 3. The questions concerned understanding the participants' thoughts on the layout, features, viability of integrating the prototype into workflow, and possible improvements as they completed the scenario. The feedback from both the simulated scenario and follow-up interviews was transcribed on the researcher's laptop, manually reviewed for accuracy, and then coded using NVivo (QSR International Pty Ltd).

Data Analysis of Recordings

Two recordings were analyzed to evaluate the prototype: a room recording of the simulated scenario and a screen recording of the iPad. They were reviewed concurrently to evaluate the interactions made, time spent using the prototype, errors made, and total time completing the simulated scenario. The time spent using the prototype was determined by timestamping each required task's start and end times, noting the participant's first and last interactions with the prototype for each task, and then adding all task times. The number of errors made was identified through a combination of the screen recording, using a Figma feature that highlights the screen when a misinput is made, and manual review of the simulated clinical scenario.

Deductive thematic analysis was conducted for the interviews, referencing Nielsen's heuristics for the codebook. Each participant transcript was analyzed line by line, with codes assigned to participant quotes representing an idea from one of Nielsen's heuristics. Each code was labeled as a "positive" or "negative" application of Nielsen's heuristics, where positive uses were to be kept for following prototype iterations and negative uses were to be changed for the following iteration. Participant quotes that best captured the group consensus on each aspect of the prototype were selected, then grouped by the primary investigator and discussed with the research team to determine the underlying heuristic the design features promoted. Because this was a pilot study and the themes were predetermined, no interrater reliability was conducted. Additionally, this study was not intended to be an exhaustive study reaching thematic saturation. However, the primary investigator ensured that feedback reported in the Results was discussed by at least 2 participants to identify overlap in participant feedback. The features and design requirements were analyzed and categorized to determine the considerations, features, and design choices for future iterations.

Ethical Considerations

Ethical approval was obtained from the Dalhousie University Research Ethics Board for prospective research involving human subjects. The study protocol, with REB file number 2023-6753, was approved by the Dalhousie Ethics Board. All participants provided recorded verbal informed consent before participating in the study, in accordance with the requirements of the Dalhousie Ethics Board. Participants were informed that participation was voluntary and that they could withdraw from the study at any time; that recordings of the experiment would be taken for analysis; that follow-up interviews would be transcribed and deidentified by coding to protect their privacy; and that confidentiality would be preserved when reporting results. Participants were also informed that no compensation would be provided for study participation.

Results

First Iteration Design

The first iteration was designed to be simple, intuitive, and efficient. Five types of input elements were included for data entry: manual input (ie, typing), buttons, steppers (ie, buttons that increase or decrease an initial value by one using a "+" or "-" button), time pickers (ie, a slider that uses a finger swipe to change an initial time value), and drop-down menus. These input elements were chosen based on 3 main factors: anticipated time to enter the information, the amount of screen space used, and the variable data types (integers, alphanumeric, list, etc). Default values were agreed upon for 3 fields to reduce error potential and interaction time: all date fields would default to "today's" date; arrival method would default to "Ambulance," as that is the primary method of arrival for stroke patients; and the "CTA/CTP occurred immediately after CT" would default to "yes."

The presentation of information was defined in this stage. To reduce the cognitive load for users, expandable information fields were designed based on user inputs. This included the "treatment information" section not expanding until the user chose the treatment method, and if CT angiography (CTA) or CT perfusion (CTP) did not occur immediately after the CT scan, the form would expand to require inputting the CTA or CTP time.

A final summary was included to provide feedback to the user after the data were collected. Two methods to summarize the information were used: a sidebar summary that condensed the information of the main page for easier readability and a final pop-up summary at the end providing the main information documented and benchmark times. Both methods were included to determine participants' preferences during user testing. Figure 1 shows the layout and features of the first iteration.

XSL•FO
RenderX

Figure 1. First prototype iteration screenshots, including: (A) main data entry page with sample information filled, (B) National Institute of Health Stroke Scale (NIHSS) calculator feature, (C) Inclusion-exclusion criteria feature, and (D) final summary pop-up.

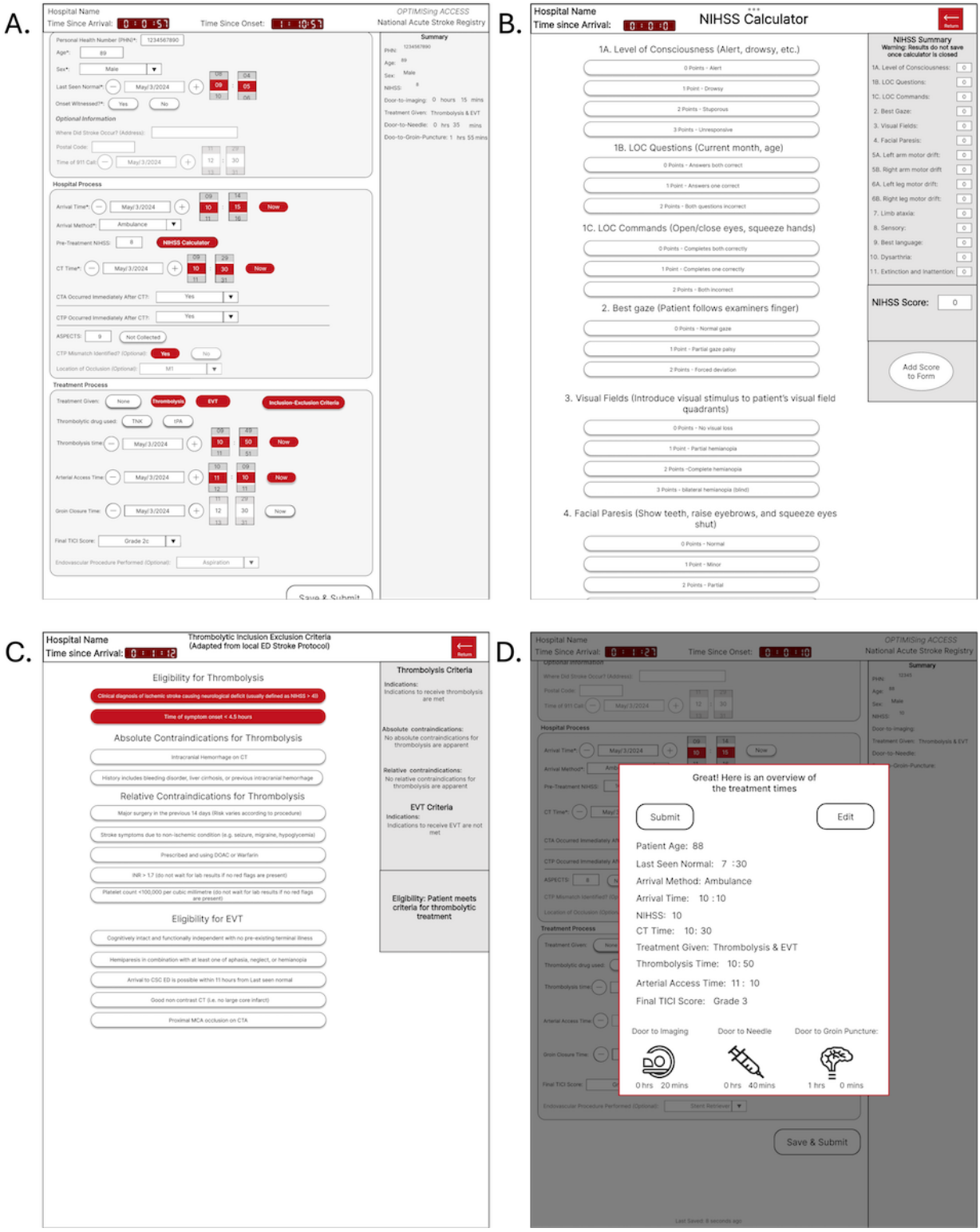

User Testing

Simulated Scenario

Recruitment occurred between April 18, 2024, and June 6, 2024. During that time, 3 clinicians were recruited to complete the simulated clinical scenario from 2 different hospitals—one CSC and one PSC. The participants consisted of an emergency physician, a stroke-trained nurse, and a neurologist. Each simulated scenario completed by the participants took 10-12 minutes. The total time using the prototype ranged from 1 minute and 12 seconds to 3 minutes and 43 seconds.

The longest task time was observed for participants who chose to use the NIHSS calculator, which was completed by 2 of the 3 participants (taking 1 minute and 50 seconds and 2 minutes and 8 seconds of the prototype use time). The cumulative number of tasks completed by the participants for each input

XSL•FO
RenderX

element and comparisons in the time taken to complete tasks with each related input element are provided in Table 1. A 2-sample *t* test (2-tailed) was completed comparing related interaction styles (ie, data fields that could be filled in multiple ways, such as by manual typing or selecting a premade list). No interaction style had a statistically significant difference in task times, except for the comparison of buttons to drop-down menus (*P*=.008).

Table 1. Comparison in the cumulative number of tasks per input element completed by the participants and comparisons of the mean, SD, 95% CI and *P* value comparing for each related interaction style.

Comparison group and interaction style	Frequency, n	Task time (seconds), mean (SD)	95% CI	*P* value
Group 1				.25
Manual input	12	4.17 (2.91)	2.31-6.02	
Time picker	16	5.50 (2.92)	3.92-7.08	
Group 2				.008
Button	11	1.46 (0.66)	1.63-1.86	
Drop-down menu	8	4.50 (2.27)	2.60-6.40	
Group 3				.30
Time picker	16	5.50 (2.92)	3.92-7.08	
Drop-down menu	8	4.50 (2.27)	2.60-6.37	
Group 4				.60
Button	11	1.46 (0.66)	1.63-1.86	
Stepper	6	1.67 (0.82)	0.81-2.52	

The number of errors made by each participant was reviewed and categorized. Six different types of errors were identified. “Missed target” errors occurred when a participant attempted to select or interact with a data entry field but missed the target box. Missed target errors for each interaction style (manual input, time picking, dropdown menu, and buttons) were recorded separately to determine which interaction styles were most prone to errors. When combined, missed targets made up 80% (32/40) of the identified errors. Typing errors were exclusive to manual input fields and occurred when participants activated the manual input correctly but entered incorrect information into the main form. Finally, “feature missed” errors occurred when participants did not recognize a feature they had to use, such as the “save and submit” button, or did not use a feature during the field study that they stated in the follow-up interviews they would have used if they recognized it. An average of 13 errors were made per participant. Table 2 provides the number of errors of each type made by each participant.

Table 2. Number of errors made by each participant (P1-P3) by error type, categorized by interaction style: manual information fields (typing), dropdown menus, time pickers (sliders), and buttons. Missed targets are counted individually.

Error type	Error frequency, n			
	P1	P2	P3	Total
Missed target (info field)	3	7	6	16
Missed target (dropdown)	1	3	5	9
Feature missed	1	3	2	6
Missed target (time)	3	1	0	4
Missed target (button)	2	1	0	3
Typing error	1	1	0	2

Qualitative Feedback

Follow-up interviews conducted with participants took between 60 and 75 minutes. The interviews reviewed each aspect of the prototype with participants to gather their feedback. Participants’ responses were coded using Nielsen’s heuristics as a framework to evaluate how they felt each aspect of the prototype met the heuristic criteria.

The most consistently positive responses about the prototype among all 3 participants regarded the minimalist design, which presented information only as needed; the usefulness of the provided features, including the NIHSS calculator, inclusion-exclusion criteria, and “now” buttons; and the value of the final pop-up summary. Participants felt that the organization of information made inputting the information intuitive and simple.

> *I think that it captures really good information. It is very clear, it is a one-pager, succinct, I like the collapsible boxes, it is very organized. It is neat and it is in an order that makes sense.* [Participant 3]

XSL•FO
RenderX

The timers, NIHSS calculator, and inclusion-exclusion criteria features included in the design were well received by all 3 participants. Tools such as the NIHSS calculator were already being used in participants' workflows, and they discussed using their phones to check the time throughout the process, so having the main tools available in one place was beneficial to participants.

> *The timers, the NIHSS calculator, the inclusion exclusion criteria, those are the big things that I liked.* [Participant 2]

When discussing the potential integration of the prototype into clinical workflow, the main facilitating factor identified was the pop-up summary. Two participants noted that if the final summary could be emailed to colleagues, printed, or linked to a patient's chart, they would be more likely to use the prototype.

> *That is a beautiful summary to take a quick screenshot of and be like, "hey, I have everything I need for that patient in terms of their ASP (acute stroke protocol). So, I love the summary.* [Participant 1]

The most negatively received aspects of the prototype discussed by all 3 participants were the small target sizes causing errors, small text sizes reducing readability, and the lack of visual cues in the design. The error potential in the design was the most frequently recognized issue discussed by participants, as it led to longer use times and increased frustration.

> *I think just everything is pretty small and you have to be very accurate as to where you push, and then if you are just a bit off, it won't accept that you are doing something.* [Participant 3]

A similar issue discussed concerned the readability of the text. Two participants found that if they wanted to review their information, they would need to focus on the iPad screen rather than the next task in the scenario. This concern was also applicable to the timer feature, as it was not immediately apparent how much time had passed when clinicians glanced at the iPad, reducing the usefulness of the feedback the software attempted to provide. Additionally, for the timers, participants stated that an added indication of how much time had passed would be useful.

> *I would make [the text] a bit bigger, a bit easier to see on the go because I find I was really concentrating on trying to read, as opposed to walking.* [Participant 1]
>
> *I would make the timers bigger. The time since arrival, time since onset counters, and maybe even some indication of how close we are to the benchmarks.* [Participant 2]

Another issue regarding the lack of recognizable cues concerned the navigational features in the prototype being easy to miss. Two participants used the NIHSS calculator feature, but both hesitated to press the "add score to form" button due to its lack of color and its placement on the side of the screen. Additionally, the inclusion-exclusion criteria feature was missed by one participant because its location in the "treatment process" section did not align with their mental workflow, in which they began determining eligibility for treatment as soon as the patient arrived.

> *(When referring to the NIHSS calculator) It is probably a bit more intuitive to have a button on the bottom, and then maybe a different color so that it is a bit more obvious.* [Participant 3]
>
> *[I did the] Inclusion and exclusion criteria, in my head, I think of it earlier on in the medical history taking, so I would have that closer to the top.* [Participant 2]

Two other issues with the layout discussed were the placement of optional information and the efficiency of the drop-down menus. All 3 participants found the first set of optional information fields interfered with the flow of how information was presented. They discussed that, while the sections made sense, the last seen normal information naturally felt as though it should be followed by the arrival time.

> *To me, intuitively, the information with last seen normal, was the onset witnessed, I gather once the patient is in hospital. So, I think of those pieces together. So, the fact that they are divided is not the most intuitive way that I think about it.* [Participant 3]

The other concern with the layout involved the drop-down menus. Two participants found that the dropdown menus were unnecessary for data fields that had only a few options. They felt that buttons displaying all options or checkboxes would be easier to use and more efficient.

> *CTA and CTP would be easier if it was just a checkbox, like did CTA occur? Yes, instead of needing to select a drop-down, it would just be a little bit faster.* [Participant 2]

All 3 participants discussed their concerns with integrating the first prototype iteration into their workflow. Four points were discussed as requirements for integrating the software into a hospital: the ability to reliably save data, return to previously incomplete records if retrospective data collection is needed, add comments for outlier records, and reaccess the data. Participants discussed the inclusion of an autosave feature to ensure that collected data would not be lost if an extenuating circumstance arose.

> *I think one of the hopes would be that this would have some interval save function so that if say, you pass it off to somebody else, and they just set it down, and the iPad in the interval has died, that the information does not immediately disappear.* [Participant 3]

The ability to return to incomplete records was discussed as an important element. Although all 3 participants agreed that most information could be gathered in real time, concerns arose that if certain data elements—either those captured by existing electronic software (such as CT time) or those with difficult-to-identify starting points (such as groin puncture time)—were to be captured in real time, it might lead to discrepancies in the data.

XSL•FO
RenderX

> *Even if you are standing in the room, the interventionalist is not really announcing when they punctured [the groin] and when they have closed, and so that would have to come from the interventionalist sheet. And, although the 'now' buttons are good, I think that it would be difficult for those to be recorded in real time.* [Participant 3]

Two participants discussed the expansion of the prototype into workflow and how its value could be improved if a free-text section were added to comment on cases. Participants explained that patient factors can influence the time to treat a patient and that a way to comment on how these factors influenced the treatment times could be beneficial when the data are reviewed.

> *A quick comment to say, 'the transfer got delayed by a snow blizzard' or 'treatment was a little weird', a little comment box to just say those things that are not captured here that come up would help.* [Participant 1]

The final topic discussed regarding integration of the software was the ability to reaccess the data. Participants found the idea of real-time feedback of the process through the pop-up summary beneficial, but they also wanted the ability to return to the data even after the information was completed.

> *A question I would have is where can I get this data back? That would be a very important question, because the biggest complaint I have is we put all this work in to get data and then we are not allowed to access it.* [Participant 1]

Second Iteration Design

A second iteration of the prototype was designed accounting for the main usability issues identified in the simulated field experiment. The revised design, including the layout and modified features, are shown in Figure 2. The main changes for the second iteration of the prototype aimed to reduce the error potential for selecting data fields and improve the visibility of the system status with visual cues. About 80% of the errors made occurred due to missing an interaction target, indicating the need to increase the size of the targets. However, the information presented still had to be condensed to a single page, and there was no available space to increase the target sizes, since the sidebar summary used approximately 25% of the screen space. Therefore, the sidebar summary was removed to provide more screen space for the main page. The additional space was used to reorganize the layout of the main page by moving the optional information fields to the right-hand side. This allowed the heights of the manual input fields to be increased from 30 pixels (approximately 0.8 cm) to 45 pixels (approximately 1.2 cm) to reduce potential error rate and increase user satisfaction, while keeping the content of the prototype condensed to one page.

To improve visual cues, color, size and location were considered the primary additions to increase the visibility of important features. The "inclusion-exclusion criteria" button was moved to the "hospital process" section beside the "NIHSS calculator" button to correspond with the users' mental model. In the NIHSS calculator, the "add score to form" button was replaced with a "Save NIHSS score" button at the bottom of the screen and colored red to further increase visibility. For the timers, 3 colors were used to visually signify to the user how close they were to meeting the treatment benchmark times, and 2 timers were added to correspond with the 3 benchmark metrics. Green signified the user is well within the benchmark, orange indicates the user was nearing the benchmark, and red indicated the user had missed the benchmark.

Finally, the dropdown menus for the "CTA/CTP occurred immediately after imaging" fields were replaced with buttons. The remaining dropdown menus were not replaced either due to the lack of screen space or the number of options within the dropdown menu.

XSL•FO
RenderX

Figure 2. Second prototype iteration screenshots including: (A) main page with relocated information, timers with colors, and new buttons replacing dropdown menus; (B) National Institute of Health Stroke Scale (NIHSS) calculator with larger selection buttons and a "Save NIHSS" button located at the bottom; (C) new "additional information" slider displaying the NIHSS scores and inclusion exclusion criteria; and (D) modified final summary including the date of arrival and location of occlusion.

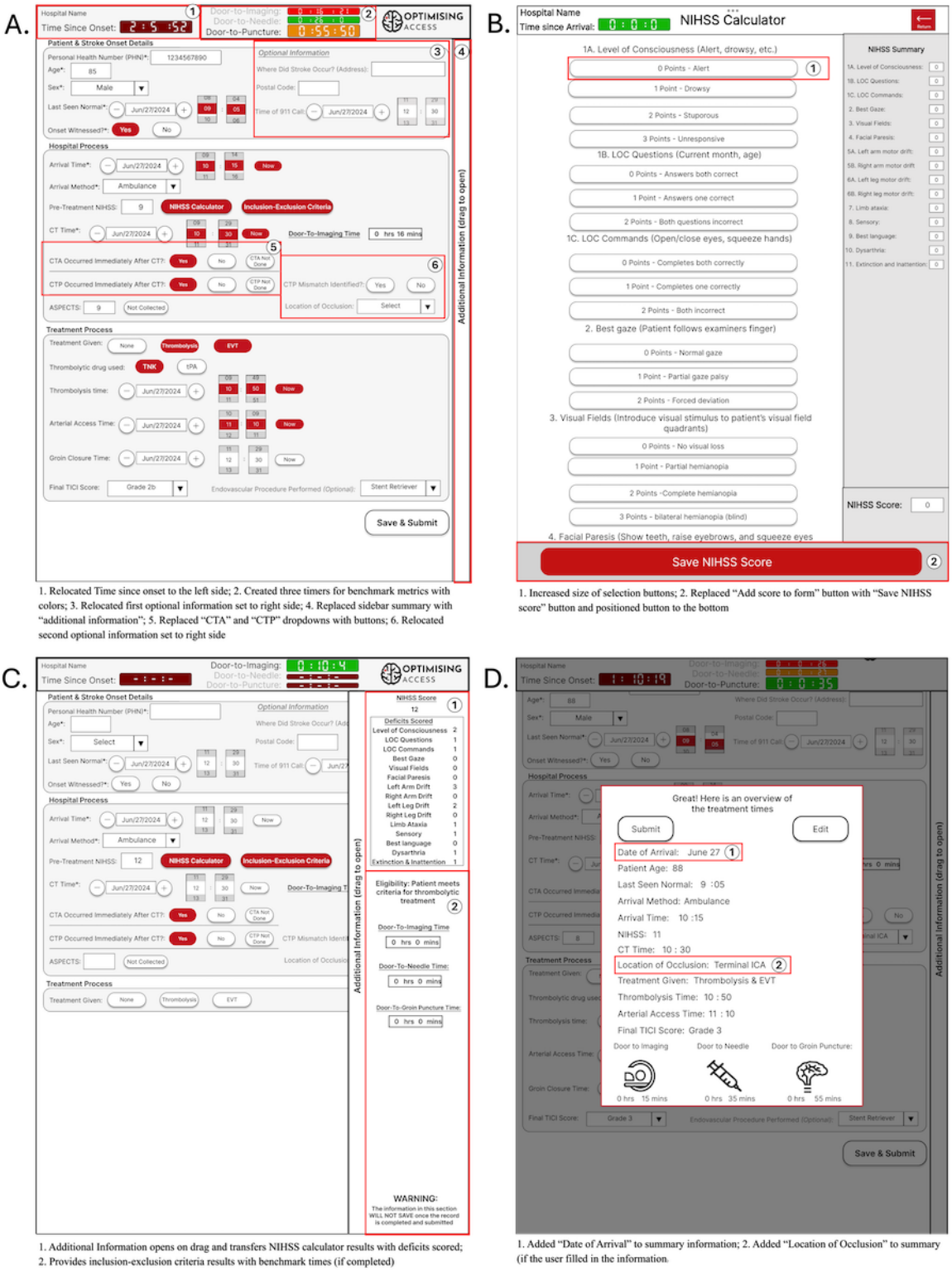

Design Requirements

The results of the user testing identified key requirements for the acute stroke treatment data collection software to effectively integrate into clinical workflow. Textbox 1 provides a summary of the design requirements gathered from the user tests and follow-up interviews, as well as their categorization using Nielsen's heuristics as a framework, highlighting how they add value to the design.

XSL•FO
RenderX

Textbox 1. Design requirements identified and categorized by Nielsen's heuristics.

Visibility of system status

- Final pop-up summary of information entered, and times collected throughout the process that can be saved and shared
- Timers counting arrival and onset times with visual cues corresponding to treatment benchmark times

Match system to the real world

- NIHSS (National Institute of Health Stroke Scale) calculator and inclusion-exclusion criteria features available in "Hospital process" section
- Separate mandatory and optional data fields to create a logical information sequence

User control and freedom

- Provide a way to search and review previously completed records
- A free-text comment system to note unique cases

Consistency and standards

- Position navigation buttons at the top or bottom of the screen
- Provide an autosave feature (and a manual save option)

Error prevention

- Target sizes must be greater than 50 pixels in width and 30 pixels in height
- Add color to navigation buttons for recognizability
- Provide default values based on date, expected method of arrival, and hospitals' scanning capabilities

Recognition rather than recall

- Transfer results of NIHSS calculator with deficits scored and inclusion-exclusion criteria features to main page

Flexibility and efficiency of use

- Font sizes greater than 14 point for quick scanning
- Use buttons in place of dropdown menus (where space allows)
- "Now" buttons to timestamp processes (with time pickers to allow adjustability if the time was missed)

Aesthetic and minimalist design

- Keep information condensed to one page
- Have information fields expand and collapse based on user inputs

Help users recognize, diagnose, and recover from errors

- Allow an incomplete record to be saved, highlighting the missing information fields

Discussion

Principal Findings

This qualitative study identified and categorized design requirements for a prototype software aimed at collecting data during clinical workflow through an exploratory small-scale user test. The results highlight some of the unique needs of clinicians recruited to collect stroke treatment data prospectively in a software application. Usability testing guidelines recommend 5 users to identify 80% of usability issues and conducting multiple small-scale tests to achieve the best iterative results [19]. Although only 3 users were able to test the first prototype iteration due to the limited participant pool for in-person participation with a naturally busy clinician demographic, the results from the simulated scenario and follow-up interviews provided rich data in a more accurate context in which the final software is anticipated to be used compared with a stationary usability test, and overlapping insights concerning the usability of the prototype were identified. Additionally, by setting up the experiment environment so that participants tested the prototype while moving and talking to multiple people while inputting data, the feedback and requirements gathered reflected the needs of clinicians in a scenario more accurate to the real-world context in which the software is ultimately intended to be used compared with a remotely conducted usability test. In this discussion, we compare the requirements gathered from the participants and features implemented in the prototype in relation to previously identified techniques to improve data quality and stroke treatment processes. Additionally, we discuss potential ideas to iterate the

XSL•FO
RenderX

final version of the software to incorporate design requirements that could realistically scale for real-world adoption.

Facilitators for Improved Data Quality

The value of the prototype depended on the ability to efficiently capture information. An obvious consideration was the speed of data entry by minimizing time spent using the prototype. The design aimed to balance a condensed, simple layout with all pertinent information on one page of the software; however, this increased the likelihood of errors due to small target sizes that must be tapped with the user's finger. Other unique insights that were critical when considering the usefulness of the software in the context of stroke data collection included the importance of clear visual cues, adaptable design, immediate feedback, and the clear information grouping. Previous studies have emphasized the importance of adaptability in data collection tools, allowing users to bypass information fields that are not relevant to them [20].

A feature discussed that was not implemented in the second prototype iteration was the inclusion of free-text comments. While viewed as potentially beneficial for data analysis, some concerns arise due to the time required to input free-text comments during workflow and the impact this could have on data quality. A previous study identified excessive free-text entry as a contributor to poor data quality in a stroke registry [21]. Additionally, much of the information found in the patient chart is not necessary for registry data and hinders the standardized format. If the final version of the software were to be used by multiple centers, it is important to ensure that the information captured is standardized, as it could potentially be shared to compare metrics between different hospitals. However, the inclusion of free-text fields could be combined with built-in iPad features such as a microphone to allow hands-free notetaking, and free-text comments could be a private field viewable only by specific clinicians for security purposes if deemed valuable and necessary to integrate with clinical documentation.

Usability Issues

Despite aspects of the prototype being positively received, there were important issues concerning the prototype's usability that impacted the participants' perspective on integrating the first iteration into workflow. Primary issues included the size of interaction targets affecting task times, use of screen space for unnecessary features, lack of visual cues for aspects such as navigation and system feedback, unintuitive sequence of information groups causing hesitance during use, and less efficient interaction styles that required excessive inputs and increased task times. Several of the identified issues could be considered by future designers of mobile medical software beyond stroke treatment, who should consider the context in which clinicians will need to input data to identify specific needs for these usability aspects.

The biggest usability issue was the size of interaction targets. The small size of targets for manual information fields, time picking, and drop-down menus caused participants to require more time to complete tasks by reselecting targets and increased frustration during use, which reduced participant buy-in. Additionally, if implemented on a larger scale, a higher error potential could reduce overall data quality. The small target sizes were largely due to the aim of fitting the prototype's content to a single page; however, the "sidebar summary" was removed, as participants did not view it as adding value due to its redundancy in presenting easily accessible on-screen information. After removing the sidebar summary, the information was reorganized to allow enough space to increase the target sizes while still containing the content on a single page.

Another usability issue that particularly affects the usefulness of a prototype for prospective stroke data collection is the visual cues. Due to the time sensitivity of stroke treatment, it is critical that a clinician does not have to thoroughly inspect software to understand the information they need to input, how to access a feature they want to use, or where to find information they want to review. The main issue identified with the prototype's first iteration was that the visual cues were not clear enough to immediately signify important information to users. This included not knowing how to navigate or activate features due to lack of color in buttons, not knowing how close they were to meeting different stroke treatment benchmarks using the timers, and being unsure if a data field was mandatory or optional.

A consideration for the final software (and future testing) is training to understand how to interpret the visual cues, as they are meant to indicate how close someone is to meeting benchmark treatment times, but not as a measure of clinician performance. While the cues could help promote a sense of urgency, they are not intended to make the user feel as if they have failed when the timer turns red. Therefore, it is important to train clinician users to understand that the visual cues are purely a signifier of time and not an indirect judgment of their performance.

Finally, the information grouping and interaction styles contributed to making the first iteration less efficient. The information was presented linearly, with all data fields positioned on the left-hand side, but the order of information and lack of distinction between mandatory and optional data fields interrupted the participants' thought process, as they assumed the next data field would be the next mandatory data point in the treatment process. Additionally, the drop-down menus (which also had higher error rates) were less efficient to use than buttons. While drop-down menus are useful in some cases, converting them to buttons allows for faster interactions and minimizes use time without compromising data quality. Design guidelines do not specify exact criteria for using drop-down menus compared with buttons; however, drop-down menus are often considered clunky if not used to conserve screen space [22].

Features Integrated Into Stroke Treatment Processes

The NIHSS calculator, inclusion-exclusion criteria, benchmark timers, and pop-up summary at the end of the scenario were seen as value-added features rather than distractions. The use of these tools has been shown to improve treatment times [15,23], indicating that additional tools may be added in future prototype iterations if they provide value to the users. However,

XSL•FO
RenderX

to accommodate the variety of potential hospitals and clinician experience levels (with both technology and stroke treatment), it was important to keep features optional to allow for the most universal appeal. Additionally, any future added features must be tested and verified with clinicians, as the concise and condensed layout was considered a critical design requirement by participants. For the PSC version of the software, a potential value-added aid could include a method of rapid telestroke consultation, as effective use and minimum delay time in telestroke consultations have been associated with improved patient outcomes [24-28]. This could be achieved by having the final software link contacts of on-call physicians to use FaceTime or perform a phone assessment, streamlining the process for completing telestroke assessments in rural and remote centers.

This study aimed to understand whether it is realistic for clinicians to prospectively collect stroke data during clinical workflow without interfering with the treatment process. While there were reservations concerning the potential to distract users, the opportunities associated with integrating the software into practice were not ignored. The pop-up summary was seen as an effective incentive for using the software during clinical workflow if it could be shared with the stroke team to provide feedback and saved to hospital systems to potentially reduce documentation load. Extensive research has shown that prompt data feedback is an effective strategy for improving treatment times [21,23,29-31]. Furthermore, a previous study indicated that documentation is one of the most time-consuming tasks for nursing staff [32], so the ability to either link the pop-up summary to a hospitals electronic medical record (EMR) or to print the summary as a sticker to attach to the patient chart would be a significant facilitator for adopting data collection software [13,33]. The final version of the software would ideally connect with the hospitals EMR, but an alternative option could be to program the final software to link with Apple AirPrint, allowing the iPad to connect with a hospital printer. This method would be easily accessible and beneficial for hospitals that have not yet converted to EMRs.

Limitations and Future Directions

This study has several limitations. First, the prototype was designed only for the iPad, so there was no phone version to compare modality preferences. In a future test, two versions of the prototype could be designed for use on an iPad and an iPhone, both incorporating the design requirements (within their limitations), to determine how each modality impacts user experience.

The second limitation was that the sample size of 3 participants means that the findings may not be generalizable, and the results should be interpreted with caution. Additionally, it meant there was not a sufficiently large amount of task time data to find statistically significant differences between most interaction styles. With a larger sample size, more requirements are expected to be gathered, validated, and compared, so the results of this study should be taken as exploratory results that can provide direction for future software considerations but should not be seen as definitive. Further testing with additional clinicians of different demographics, experience, and geographic location can determine additional findings.

A third limitation was that the convenience sampling method can lead to selection bias, as the interested participants may not be indicative of the greater clinical population. Convenience sampling was chosen because the natural pool of potential participants was quite small, as they had to be clinicians within the study's small, low-population province who were directly involved in stroke treatment, within driving distance of the research laboratory, and available to participate during the one-and-a-half-month timeframe allotted for user testing. However, this method was chosen to complete the simulated clinical field test, which allowed for recordings of both the prototype screen and the participants' movement for more accurate analysis of when participants would use the prototype during the simulated scenario. Additionally, all requirements identified had to be independently identified by at least 2 (but more often all 3) participants to be considered. Future tests could recruit more participants by having a longer timeframe for recruitment, having the research team travel directly to hospitals for testing sessions, or providing incentives to recruit a larger sample for usability tests. Additionally, a larger sample can be gathered through online surveys with clinicians to review the future iteration's tools and features.

The fourth and final limitation is that the simulated clinical scenario was conducted in a single room with individual participants with no background noise. This environment is not similar to a real emergency department, and a more realistic simulated scenario with added stimuli could result in identifying additional design requirements. To further develop the prototype, a larger in-person field test with multiple hospitals in different regions, taking place at the participants' hospital of practice, will help identify further design requirements in the user's environment. A high-fidelity prototype should be created using the data gathered, and the following prototype can then be tested in a real emergency department, either as a more advanced pilot test or as a clinical trial.

Conclusions

In this study, we conducted a small-scale exploratory test to identify key design requirements for prospective data collection software to integrate into the clinical workflow of the acute stroke treatment process. User testing in a simulated clinical scenario was conducted to provide participants with a more accurate sense of how the prototype would integrate into workflow, so that feedback would be more accurately reflect the needs of clinicians in a stroke treatment scenario. Deductive thematic analysis highlighted the main factors of the prototype that both positively and negatively affected clinicians' ability to capture data and their overall experience using the prototype. The first prototype iteration met clinician participants' requirements of providing prompt data feedback and a simple layout. However, issues concerning high error potential, lack of visual cues to highlight system status, and information grouping hindered the ability to collect required information seamlessly, reduced the value of some features, and negatively affected the user experience. The user testing and analysis resulted in a list of design requirements and how they improve

XSL•FO
RenderX

the prototype's usability, which was then implemented in a second prototype iteration. This study identified preliminary key needs for adopting prospective data collection software and the requirements to improve the software's value to users without hindering their workflow. The requirements identified and heuristics most applicable to the software's features could serve as a basis for future developers when designing similar software.

Conflicts of Interest

None declared.

Multimedia Appendix 1
List of variables.
[PDF File (Adobe PDF File), 452 KB - formative_v9i1e64800_app1.pdf]

Multimedia Appendix 2
Information architecture.
[DOCX File , 416 KB - formative_v9i1e64800_app2.docx]

Multimedia Appendix 3
Interview guide.
[PDF File (Adobe PDF File), 126 KB - formative_v9i1e64800_app3.pdf]

References

1. Hoque DME, Kumari V, Hoque M, Ruseckaite R, Romero L, Evans SM. Impact of clinical registries on quality of patient care and clinical outcomes: a systematic review. PLoS One 2017;12(9):e0183667 [FREE Full text] [doi: 10.1371/journal.pone.0183667] [Medline: 28886607]
2. Hills NK, Johnston SC. Duration of hospital participation in a nationwide stroke registry is associated with improved quality of care. BMC Neurol 2006;6:20 [FREE Full text] [doi: 10.1186/1471-2377-6-20] [Medline: 16740167]
3. Schwamm LH, Fonarow GC, Reeves MJ, Pan W, Frankel MR, Smith EE, et al. Get with the guidelines-stroke is associated with sustained improvement in care for patients hospitalized with acute stroke or transient ischemic attack. Circulation 2009;119(1):107-115. [doi: 10.1161/CIRCULATIONAHA.108.783688] [Medline: 19075103]
4. Fasugba O, Sedani R, Mikulik R, Dale S, Vařecha M, Coughlan K, et al. How registry data are used to inform activities for stroke care quality improvement across 55 countries: a cross-sectional survey of registry of stroke care quality (RES-Q) hospitals. Eur J Neurol 2024;31(1):e16024 [FREE Full text] [doi: 10.1111/ene.16024] [Medline: 37540834]
5. Palaiodimou L, Kargiotis O, Katsanos AH, Kiamili A, Bakola E, Komnos A, et al. Quality metrics in the management of acute stroke in Greece during the first 5 years of registry of stroke care quality (RES-Q) implementation. Eur Stroke J 2023;8(1 Suppl):5-15 [FREE Full text] [doi: 10.1177/23969873221103474] [Medline: 36793743]
6. AuSCR Annual Report. AuSCR. 2023. URL: https://auscr.com.au/about/annual-reports/ [accessed 2025-10-30]
7. Stotts G, Poppe A, Hill MD, Kelly M, Swartz R, Shankar J. OPTIMISE: a national endovascular treatment quality assurance program. Int J Stroke 2019;14(3_suppl):3-52 [FREE Full text] [doi: 10.1177/1747493019872147]
8. Trivedi R, Alcock S, Trivedi A, Ghrooda E, McEachern J, Shankar J. Suboptimal outcome with endovascular thrombectomy: will acute stroke unit help? Can J Neurol Sci 2021;48(4):567-569. [doi: 10.1017/cjn.2020.244] [Medline: 33161913]
9. Yoon SS, George MG, Myers S, Lux LJ, Wilson D, Heinrich J, et al. Analysis of data-collection methods for an acute stroke care registry. Am J Prev Med 2006;31(6 Suppl 2):S196-S201. [doi: 10.1016/j.amepre.2006.08.010] [Medline: 17178303]
10. Andrew BY, Stack CM, Yang JP, Dodds JA. mStroke: "Mobile Stroke"-improving acute stroke care with smartphone technology. J Stroke Cerebrovasc Dis 2017;26(7):1449-1456. [doi: 10.1016/j.jstrokecerebrovasdis.2017.03.016] [Medline: 28434773]
11. Noone ML, Moideen F, Krishna RB, Pradeep Kumar V, Karadan U, Chellenton J, et al. Mobile app based strategy improves door-to-needle time in the treatment of acute ischemic stroke. J Stroke Cerebrovasc Dis 2020;29(12):105319. [doi: 10.1016/j.jstrokecerebrovasdis.2020.105319] [Medline: 32992177]
12. Gutiérrez-Zúñiga R, Uclés J, Sánchez-Manso JC, Fuentes B, de Celis E, Ruiz-Ares G, et al. The use of a smartphone application to improve stroke code workflow metrics: a pilot study in a comprehensive stroke centre. Digit Health 2022;8:20552076221137252 [FREE Full text] [doi: 10.1177/20552076221137252] [Medline: 36406153]
13. Garrett P, Brown CA, Hart-Hester S, Hamadain E, Dixon C, Pierce W, et al. Identifying barriers to the adoption of new technology in rural hospitals: a case report. Perspect Health Inf Manag 2006;3:9 [FREE Full text] [Medline: 18066367]

XSL•FO
RenderX

14. Boonstra A, Broekhuis M. Barriers to the acceptance of electronic medical records by physicians from systematic review to taxonomy and interventions. BMC Health Serv Res 2010;10:231 [FREE Full text] [doi: 10.1186/1472-6963-10-231] [Medline: 20691097]
15. Forward A, Sahli A, Evans R, Kamal N. Exploring differences in stroke treatment between urban and rural hospitals: a thematic analysis of practices in Canada. Can J Neurol Sci 2025:1-9. [doi: 10.1017/cjn.2024.369] [Medline: 39773573]
16. Salbach NM, Mountain A, Lindsay MP, Blacquiere D, McGuff R, Foley N, Canadian Stroke Best Practice Recommendations Advisory Committee, in collaboration with the Canadian Stroke Consortiumthe Canadian Partnership for Stroke Recovery. Canadian stroke best practice recommendations: virtual stroke rehabilitation interim consensus statement 2022. Am J Phys Med Rehabil 2022;101(11):1076-1082. [doi: 10.1097/PHM.0000000000002062] [Medline: 35767008]
17. Heran M, Lindsay P, Gubitz G, Yu A, Ganesh A, Lund R, et al. Canadian stroke best practice recommendations: acute stroke management, 7 edition practice guidelines update, 2022. Can J Neurol Sci 2024;51(1):1-31. [doi: 10.1017/cjn.2022.344] [Medline: 36529857]
18. Nielsen J. Jakob Nielsen's 10 usability heuristics. Nielsen Norman Group Internet. 2012. URL: https://media.nngroup.com/media/articles/attachments/Heuristic_Summary1-compressed.pdf [accessed 2024-02-20]
19. Nielsen J. How many test users in a usability study. Nielsen Norman Group Internet. 2012. URL: https://www.nngroup.com/articles/how-many-test-users/ [accessed 2020-03-10]
20. Lynch E, Laver K, Levy T, Schultz T. 'The way that we are collecting and using data has evolved' evaluating the Australian National Stroke Audit programme to inform strategic direction. BMJ Open Qual 2023;12(1):e002136 [FREE Full text] [doi: 10.1136/bmjoq-2022-002136] [Medline: 36693674]
21. Lanzola G, Parimbelli E, Micieli G, Cavallini A, Quaglini S. Data quality and completeness in a web stroke registry as the basis for data and process mining. J Healthc Eng 2014;5(2):163-184 [FREE Full text] [doi: 10.1260/2040-2295.5.2.163] [Medline: 24918182]
22. Angie L. Dropdowns: design guidelines. Nielsen Norman Group. 2017. URL: https://www.nngroup.com/articles/drop-down-menus/ [accessed 2025-10-30]
23. Kamal N, Smith E, Jeerakathil T, Hill M. Thrombolysis: improving door-to-needle times for ischemic stroke treatment - a narrative review. Int J Stroke 2018;13(3):268-276. [doi: 10.1177/1747493017743060] [Medline: 29140185]
24. Jagolino-Cole AL, Bozorgui S, Ankrom CM, Vahidy F, Bambhroliya AB, Randhawa J, et al. Variability and delay in telestroke physician alert among spokes in a telestroke network: a need for metric benchmarks. J Stroke Cerebrovasc Dis 2019;28(11):104332. [doi: 10.1016/j.jstrokecerebrovasdis.2019.104332] [Medline: 31439524]
25. French B, Boddepalli R, Govindarajan R. Acute ischemic stroke: current status and future directions. Mo Med 2016;113(6):480-486 [FREE Full text] [Medline: 30228538]
26. Nguyen-Huynh MN, Klingman JG, Avins AL, Rao VA, Eaton A, Bhopale S, KPNC Stroke FORCE Team. Novel telestroke program improves thrombolysis for acute stroke across 21 hospitals of an integrated healthcare system. Stroke 2018;49(1):133-139 [FREE Full text] [doi: 10.1161/STROKEAHA.117.018413] [Medline: 29247142]
27. Zerna C, Jeerakathil T, Hill MD. Telehealth for remote stroke management. Can J Cardiol 2018;34(7):889-896. [doi: 10.1016/j.cjca.2017.12.025] [Medline: 29459240]
28. Kennedy BA, Stout PJ. Telestroke process at a community hospital: a quality improvement project. J Emerg Nurs 2023;49(4):546-552. [doi: 10.1016/j.jen.2022.12.008] [Medline: 36710095]
29. Ruff IM, Ali SF, Goldstein JN, Lev M, Copen WA, McIntyre J, et al. Improving door-to-needle times: a single center validation of the target stroke hypothesis. Stroke 2014;45(2):504-508. [doi: 10.1161/STROKEAHA.113.004073] [Medline: 24399372]
30. Xian Y, Xu H, Lytle B, Blevins J, Peterson ED, Hernandez AF, et al. Use of strategies to improve door-to-needle times with tissue-type plasminogen activator in acute ischemic stroke in clinical practice: findings from target: stroke. Circ Cardiovasc Qual Outcomes 2017;10(1). [doi: 10.1161/CIRCOUTCOMES.116.003227] [Medline: 28096207]
31. Xian Y, Smith EE, Zhao X, Peterson ED, Olson DM, Hernandez AF, et al. Strategies used by hospitals to improve speed of tissue-type plasminogen activator treatment in acute ischemic stroke. Stroke 2014;45(5):1387-1395. [doi: 10.1161/STROKEAHA.113.003898] [Medline: 24713527]
32. Philipsen N, Carruthers W, Chi G, Ensey D, Shmorhun A, Valdez R. A mixed-methods assessment of time spent documenting by nurses using an electronic medical records system. 2014 Presented at: Systems and Information Engineering Design Symposium (SIEDS); April 25, 2014; Charlottesville p. 118-123. [doi: 10.1109/sieds.2014.6829925]
33. Bowens FM, Frye PA, Jones WA. Health information technology: integration of clinical workflow into meaningful use of electronic health records. Perspect Health Inf Manag 2010;7(Fall):1d [FREE Full text] [Medline: 21063545]

Abbreviations

CSC: comprehensive stroke center
CT: computed tomography
CTA: CT angiography
CTP: CT perfusion

XSL•FO
RenderX

EMR: electronic medical record
IA: information architecture
NIHSS: National Institute of Health Stroke Scale
OPTIMISE: Optimising Patient Treatment in Major Ischemic Stroke With Endovascular Thrombectomy
PSC: primary stroke center

Edited by J Sarvestan; submitted 26.07.24; peer-reviewed by P Watson, K Shaw; comments to author 29.04.25; revised version received 10.06.25; accepted 26.08.25; published 19.12.25.

Please cite as:
Forward A, Koca G, Sahli A, Kamal N
Identification of Design Requirements for a Software Application for Use by Clinicians That Collects Acute Stroke Treatment Data During Clinical Workflow: Pilot Study
JMIR Form Res 2025;9:e64800
URL: https://formative.jmir.org/2025/1/e64800
doi:10.2196/64800
PMID:

©Adam Forward, Gizem Koca, Aymane Sahli, Noreen Kamal. Originally published in JMIR Formative Research (https://formative.jmir.org), 19.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Usability of a Tablet-Based Cognitive Assessment Administered by Medical Assistants in General Practice: Implementation Study

Philipp Schaper[1], Dr rer nat; Alexander Hanke[2], MSc; Stephan Jonas[3], Prof Dr; Leon Nissen[3], MSc; Lara Marie Reimer[3], Dr rer nat; Florian Schweizer[3,4], MSc; Michael Wagner[2], Prof Dr; Kristin Rolke[5], MSc; Carolin Rosendahl[5], MSc; Judith Tillmann[5], PhD; Klaus Weckbecker[5], Prof Dr Med; Jochen René Thyrian[6], Prof Dr

[1]Psychological Aging Research, Department of Psychology, University of Siegen, Siegen, Germany
[2]Department of Old Age Psychiatry and Cognitive Disorders, University Hospital Bonn, Bonn, North Rhine-Westphalia, Germany
[3]Institute for Digital Medicine, University Hospital Bonn, Bonn, North Rhine-Westphalia, Germany
[4]TUM School of Computation, Information and Technology, Technical University of Munich, Garching bei München, Bavaria, Germany
[5]General Practice I and Interprofessional Care, Institute of General Practice and Primary Care, Witten/Herdecke University, Witten, Germany
[6]German Center for Neurodegenerative Diseases, Greifswald, Germany

Corresponding Author:
Philipp Schaper, Dr rer nat
Psychological Aging Research
Department of Psychology
University of Siegen
Adolf-Reichwein-Straße 2a
Siegen, 57076
Germany
Phone: 49 271 740 5541
Email: philipp.schaper@uni-siegen.de

Abstract

Background: Digital short cognitive tests administered by medical assistants (MAs) in general practitioners' (GPs) practices have great potential for the timely identification of patients with dementia, because they can lead to targeted specialist referrals or to immediate reassurance of patients regarding their perceived concerns. However, integration of this testing approach into clinical practice requires good usability for the test itself, especially for cognitively impaired older adults.

Objective: In this implementation study, the digital version of the Montreal Cognitive Assessment (MoCA) Duo was conducted by MAs in general practice. We tested if the interaction with the test is associated with usability problems for the patients and aimed to find additional relevant constructs that should be considered for the potential implementation of such digital tests into clinical practice. We focused the study on subjective success, usability, and workload as well as their association with the result of the cognitive test to assess whether the MoCA Duo can be implemented into general practice.

Methods: In total, 10 GPs took part in the study. Within their practices, 299 GP patients (aged 51-97 years) with cognitive concerns completed the MoCA Duo administered by MAs. Subsequently, patients and MAs completed digital questionnaires regarding the interaction with the app. Usability was measured using the adapted System Usability Scale, and perceived workload using the National Aeronautics and Space Administration Task Load Index. For the perceived workload, we included an assessment of the patient by the MA. Results of the MoCA Duo were supplied to the GPs for their consultation with the patient.

Results: The results indicated good usability for the MoCA Duo. Self-assessment by the patients indicated that 64% (191/299) could perform in the test to the best of their ability, affected by their MoCA score. We found significant higher usability ratings by patients with better MoCA scores as well as by younger patients. Furthermore, the perceived workload showed overall medium workload. We found significant correlations between the subjective perceived workload of the patients and the assessment by MAs. Self-assessments as well as assessments by the MAs were significantly influenced by the MoCA scores and the age of the participants.

Conclusions: The results indicate good usability of the digital MoCA within the sample, supporting the idea that the resulting scores are adequate to assess cognitive status without dependence on technological affinity. Furthermore, the results highlight the relevance of heterogenous samples for comparable evaluation studies, based on the significant effect of cognitive status and age on usability and workload.

XSL•FO
RenderX

(JMIR Form Res 2025;9:e76010) doi:10.2196/76010

KEYWORDS

computerized assessment; dementia; digital assessment; digital platform; mild cognitive impairment; MoCA; Montreal Cognitive Assessment

Introduction

Based on the predictions by the Global Burden of Disease 2019 Dementia Forecasting Collaborators, more than 57.4 million people globally were living with dementia in 2019, with an expected increase to 152.8 million in 2050 [1,2]. Even though most of the predicted future cases will be in developing countries, an increase in cases is also expected for industrialized countries. In the case of Germany, in addition to the existing 1.8 million cases, an additional 360,000 to 440,000 new cases are expected [2]. The high number of cases in itself is a challenge for public health. However, due to the progressive nature of dementia, timely diagnosis severely affects intervention and potential treatment options, which currently only slow progression of the disease. Therefore, timely diagnosis mediates the personal, medical, social, and economic impact of the disease [3]. For the best results, it is crucial that testing for early signs is accessible within standard health care [4].

The German guideline on dementia already recommends the administration of cognitive short tests such as the Montreal Cognitive Assessment (MoCA) or the Mini Mental State Examination (MMSE) if self-reported memory problems are indicated by the patient or in response to the general practitioner's (GP) request [5]. This is due to the gatekeeping role of GPs in German health care, where they not only treat patients but also function as coordinators for further treatment by specialists, including neurologists in the case of dementia symptoms. However, within general practice, patients with manifested dementia are often not recognized as such [6,7] due to no or insufficient testing [8], resulting in delayed diagnosis and belated management of the disease. The rate of affected patients without an adequate diagnosis could be as high as 50% [9]. Due to the late diagnosis, patients are often referred to specialists once the symptoms are severe and advanced diagnostics are no longer indicated. In contrast, patients benefit most if they receive a timely diagnosis [10].

There are multiple reasons for the lack of timely diagnosis, including the high workload in GP practices, lack of expertise with mild cognitive impairment (MCI) tests, and low monetary compensation for the current diagnostic methods [11,12]. These problems have already resulted in many GPs withdrawing from dementia diagnostics altogether [13]. Even though such structural problems cannot be changed rapidly, novel assessment approaches involving supervised or unsupervised digital cognitive testing could benefit patients as well as GPs [14,15].

The inclusion of digital cognitive testing into standard processes within GP practices allows for a more approachable and standardized testing [16,17]. In turn, GPs have an additional indicator or a starting point for dementia diagnostics, which helps to recognize initially mild symptoms [10]. This would lead to more targeted referrals to specialists because patients are identified early enough to benefit from their timely diagnosis in terms of treatment or at least disease management, while healthy patients and severe cases of dementia receive no superfluous extended diagnostics. It has been shown that using IT-supported assessments is perceived as beneficial by professional users [18] and in the context of GP practice [19].

However, the relatively sudden increase in apps within the German health care system [20], including several options for dementia screening [21,22] or care, results in apps not being used at all, different preferences between practitioners and patients, and missing standardization of the process [23]. Even though their potential benefit is recognized in theory, the integration of apps within primary care is still lacking [24]. Therefore, it is necessary not only to evaluate the validity of a cognitive screening app within an experimental study but also whether the app and the resulting procedure are suited for integration into general practice.

This study focuses on the quantitative usability evaluation of the iCreate study ("Digitally supported case finding to improve the diagnosis and care of patients with dementia in primary care"), a study that aims to assess the feasibility, validity, usability, and clinical impact of tablet-based cognitive testing by medical assistants (MAs) in GP practices. This implementation study compares the patient's interaction with the digital cognitive test against established usability criteria, as well as the perceived workload for the patients with that reported by the MAs. The results should allow assessment of whether the test can be integrated into general practice and highlight potential problems as well as avenues for further research.

Methods

Overview

This study addresses the usability aspects of the tablet-based MoCA Duo within the iCreate study, a prospective feasibility study of tablet-based MoCA testing administered by MAs in GP practices. Producing meaningful data in digital dementia diagnostics requires that the testing process as well as the results are not influenced by usability problems [25].

Study Procedure

The study took place from June 1, 2023, to March 3, 2025, and was conducted by 4 university institutes and 1 research institution in North Rhine-Westphalia, Germany. Potential GPs in the city of Bonn and the Rhine-Sieg district in western Germany were contacted by the study team and offered participation in the project until the criterion of integrating 10 GPs, as set by the overarching iCreate study, was met.

Each GP practice was visited at least twice for presentation of the study and for a 1-hour training of the MAs to administer the tablet-based MoCA Duo. Patients were informed by posters

XSL•FO
RenderX

and flyers in the GPs' waiting room about optional cognitive testing in case of concerns regarding cognition and either requested this test themselves or followed the suggestion of the GPs or MAs to examine a suspected impairment. Integration into the study was therefore not randomized but rather based on selection by the participants themselves or the GP practice team.

The MoCA Duo is a 1:1 app replication of the paper-and-pencil MoCA [26], which is used in clinical practice, and good equivalence of both MoCA versions has been established [27]. The app and all subsequent questionnaires were completed on an iPad Air (11-inch, Apple Inc). Study participation including consent forms, the MoCA Duo, and all questionnaires, took approximately 25 minutes. The GP practices were compensated for the additional time and effort to conduct the test (15€ [US $17.36] for the MAs, 25€ [US $28.94] for the practice).

The MoCA Duo creates a result summary, which was supplied to the GP, who suggested an adequate next step, ranging from no further procedures to a retest in the GPs office in a year, or additional testing and diagnostic workup in the memory clinic of the University Clinic Bonn. Results of qualitative interviews with a smaller set of GPs, Mas, and patients have recently been published [28]. The interviews indicated positive reception of the digital test from all 3 groups, and GPs and MAs indicated interest in long-term implementation into standard practice. Results on diagnostic accuracy (prediction of MCI or dementia diagnoses in subsequent memory clinic assessment) will be reported elsewhere.

Questionnaires

First the MoCA Duo was administered by the MAs to assess the cognitive status of the patient. Subsequently, the patients used the same tablet to answer questionnaires on usability, workload, and secondary measures regarding the MoCA Duo. Finally, the MA completed an assessment on the perceived workload of the patient. There were no additional questionnaires for the MA. Participants were free not to take part in the study; however, there were no aborts during testing, and therefore only complete data for all measures could be used.

Cognitive Status

To assess cognitive impairment, we used a commercially available digital adaptation of the MoCA [29,30], which directly imitates the paper version of the test [27], with instructions for the administering MA as well as for the patient. Two cutoffs were used based on large German-speaking normative and clinical samples [31]. A score of 23 or lower indicates high risk for MCI, a score of 27 or more indicates low risk for MCI, and scores in between require further examination, subsequently referred to as "within threshold."

Usability

The System Usability Scale (SUS) is an established measure used to assess the user's subjective perception of an interface [32] and has been previously validated [33]. For this study, we used the already published simplified version of the questionnaire, especially designed for cognitively impaired and older adults based on the expected sample population [34]. This scale uses similar but slightly reworded statements, which refer to "confusion" caused by the system rather than to "inconsistency." As in the original version, the items are presented as statements to be rated on a Likert scale from 1 to 5, ranging from "strongly disagree" to "strongly agree." The German translation of the original SUS by Gao et al [35] was used but had to be slightly adapted by the authors to reflect the adaptations by Holden [34]. The resulting score, between 0 and 100, is subsequently categorized into different categories (Bangor et al [32], as applied by Souza et al [36]): below 20.5 (worst imaginable), 21 to 38.5 (poor), 39 to 52.5 (average), 53 to 73.5 (good), 74 to 85.5 (excellent), and 86 to 100 (best imaginable).

Workload

The National Aeronautics and Space Administration Task Load Index (NASA TLX) is an established measure to assess the workload of a task or interaction in regard to mental, physical, and temporal demand as well as performance, effort, and frustration [37]. The scale can be used regarding one specific task, which necessarily interrupts the interaction, or for an entire process. The instrument has been used in diverse contexts, often for work-related tasks, but also for workloads in older patients in interaction with health care instruments [38,39]. In addition to the commonly used self-assessment for the scale, we also asked the MAs to rate the respective patients on the same scales regarding their interaction with the MoCA Duo. In contrast to the SUS, the resulting scores of the NASA TLX do not result in predefined categories; however, differences in the subscales can pinpoint the source of the workload. Furthermore, the scale is sensitive to small differences in workload when comparing different applications [39], therefore creating a meaningful baseline for subsequent research.

Secondary Variables

We added additional questions for more demographic information, such as age, relationship status, and whether living on one's own, but also addressed the self-assessed performance of the patients within the MoCA Duo, as well as their daily use of technology. For assessment of their own performance, participants were asked if they could show their full ability within the test. If not, participants could choose from a range of potential reasons or could indicate one in free text.

To assess daily use of technology, participants were asked how often they used a smartphone or tablet in a day, with the options of "never," "rarely," "regularly," and "often." We also assessed if they had prior experience with cognitive tests. As an additional indicator for the usability of the digital MoCA, we asked the participants to compare the interaction with the test relative to their interaction with their own smartphone or tablet, with the options of "easier," "similar," "more difficult," and "I don't use a similar device."

Data Analysis

To test the plausibility of our results, we plan an ANOVA for the scores in the MoCA using the age groups as a factor, expecting a decline in the scores indicating increasing risk with higher age. Subsequently, we plan an analysis of the usability assessment based on the categories of the MoCA to test if

XSL•FO
RenderX

usability is affected by cognitive status. For the workload, we compare self-assessment and assessment by the MAs using correlations. The different subscales of the self-assessment will be analyzed using multivariate analysis of variance (MANOVA), using the age groups in one analysis and the categories of the MoCA in a separate analysis as factors. These analyses will also be repeated for the assessment by the MAs.

Because low affinity with technology might be associated with worse results in a digital MoCA [40], a chi-square test on the daily smartphone use of the patients and the 3 categories for the MoCA scores is conducted.

All analyses were conducted using R software (version 4.0.0; R Foundation for Statistical Computing).

Ethical Considerations

There were no invasive procedures involved in the study. The instruments were developed and administered as efficiently as possible to keep time demands low. The Ethics Committee of the Medical Faculty of the University of Bonn, Germany (No 258/23-EP), granted approval for the whole iCreate feasibility study. All participants gave written informed consent based on a study description and explanation of data use in German. Study data were clearly indicated as deidentified. The GP practices were compensated for the additional time and effort to conduct the test (€15 [US $17.36] for the MAs, €25 [US $28.94] for the practice); the participants themselves received no additional compensation for study participation.

Results

Sample

Medical Practices

Overall, a total of 10 GPs in the city of Bonn and the Rhine-Sieg district in western Germany took part in the study. These practices tested between 4 and 70 patients during the study period.

Patients

The sample consisted of 299 patients who decided to partake and agreed to the use of the MoCA Duo. The participants were dominantly female (193/299, 65%), had a mean age of 73.23 years (SD 9.82; range 51-97 years), mostly lived with someone else (206/299, 69%), and were mainly in a relationship (204/299, 68%). Based on participants' recollection, a total of 13% (40/299) had prior experience with cognitive tests, 6% (17/299) were uncertain, 80% (238/299) reported not having prior experience, and 1% (4/299) chose not to answer the question.

Cognitive Status

The mean score in the MoCA for the sample was 23.21 (SD 5.22). According to the thresholds by Thomann et al [31], our sample contained 33% (100/299) patients with low risk for MCI, 43% (130/299) patients with high risk for MCI, and 23% (69/299) within the threshold. The MoCA scores based on equally spaced age groups are presented in Table 1. A 1-way ANOVA showed a significant decline in MoCA scores with increase in age (F_2=28.84; P<.001).

XSL•FO
RenderX

Table 1. Mean values and SDs for usability and workload based on age groups.

Scale and subscale	Age groups (years), mean (SD)			
	50 to 60 (n=29)	61 to 70 (n=87)	71 to 80 (n=96)	81 and older (n=87)
MoCA[a] scores	26.21 (3.37)	25.97 (3.43)	22.73 (4.91)	19.99 (5.58)
SUS[b]	76.03 (18.52)	77.56 (16.97)	66.20 (20.77)	57.76 (20.75)
NASA TLX>[c] self-assessment				
Mental workload	6.55 (5.42)	6.70 (5.15)	8.10 (5.40)	8.70 (5.81)
Physical workload	3.21 (3.88)	2.89 (3.64)	4.26 (4.72)	5.14 (5.74)
Temporal workload	4.45 (4.28)	4.09 (4.75)	4.54 (4.36)	5.75 (5.41)
Performance	13.07 (5.69)	13.29 (5.31)	11.09 (5.34)	8.94 (4.62)
Effort	8.72 (5.01)	7.78 (4.91)	8.55 (4.91)	8.68 (5.62)
Frustration	6.62 (5.31)	6.24 (5.62)	7.35 (5.91)	7.36 (6.01)
NASA TLX external assessment				
Mental workload	5.76 (5.69)	5.91 (5.29)	8.64 (5.98)	11.78 (6.05)
Physical workload	2.79 (3.82)	3.43 (4.80)	4.71 (4.73)	5.83 (6.12)
Temporal workload	4.66 (5.07)	5.17 (5.04)	5.59 (4.77)	6.17 (5.54)
Performance	13.24 (6.15)	13.80 (6.05)	11.33 (5.82)	8.00 (5.50)
Effort	6.10 (5.17)	7.21 (5.80)	9.33 (5.68)	11.71 (5.92)
Frustration	5.21 (5.21)	6.16 (5.81)	8.45 (6.26)	8.49 (6.30)

[a]MoCA: Montreal Cognitive Assessment.

[b]SUS: System Usability Scale.

[c]NASA TLX: National Aeronautics and Space Administration Task Load Index.

Usability

The mean usability rating on the SUS was relatively high (mean 68.00, SD 21.02) within the possible range from 0 to 100. Based on the benchmarks [32] added to the original SUS scale [41], the digital test would be considered to show good usability, falling in the range of 53 to 73.5. The overall distribution of results is shown in Multimedia Appendix 1.

Table 2 presents the SUS scores relative to the categories for the MoCA. A 1-way ANOVA resulted in a significant effect of the MoCA categories on the SUS score (F_2=62.79; P<.001), with lower scores in the threshold category and even lower scores in the high-risk category. The SUS scores relative to equally spaced age groups in Table 1 show a comparable pattern, with a 1-way ANOVA indicating a significant main effect (F_3=16.85; P<.001), with lower scores for older patients.

XSL•FO
RenderX

Table 2. Mean values and SDs for usability and workload relative to the 3 Montreal Cognitive Assessment (MoCA) categories.

Scale and subscale	Low risk for MCI[a] (n=100), mean (SD)	Within threshold (n=69), mean (SD)	High risk for MCI (n=130), mean (SD)
SUS[b]	81.18 (16.52)	72.54 (16.74)	55.46 (18.95)
NASA TLX>[c] self-assessment			
Mental workload	6.14 (4.93)	5.83 (4.28)	9.94 (5.72)
Physical workload	2.54 (3.24)	2.93 (3.84)	5.72 (5.59)
Temporal workload	3.56 (3.93)	3.86 (4.21)	6.14 (5.38)
Performance	14.82 (5.03)	11.39 (4.97)	8.54 (4.29)
Effort	6.78 (4.75)	6.87 (4.15)	10.42 (5.19)
Frustration	4.88 (5.01)	6.41 (5.44)	8.85 (5.96)
NASA TLX external assessment			
Mental workload	4.20 (4.85)	7.97 (5.17)	12.04 (5.54)
Physical workload	2.50 (3.76)	3.42 (4.43)	6.55 (5.78)
Temporal workload	3.43 (3.86)	5.54 (4.51)	7.18 (5.64)
Performance	16.65 (4.72)	12.42 (4.54)	6.52 (4.04)
Effort	5.06 (4.92)	8.58 (4.74)	12.47 (5.45)
Frustration	3.84 (4.66)	7.07 (5.27)	10.50 (6.05)

[a]MCI: mild cognitive impairment.

[b]SUS: System Usability Scale.

[c]NASA TLX: National Aeronautics and Space Administration Task Load Index.

Workload

Overall mean ratings for the self-assessment of the patients and the assessment of the patients by the MAs are shown in Table 3. The self-assessment on all 6 subscales showed significant positive correlations with the external assessment in a range from 0.44 to 0.57. The lowest scores occurred for physical and temporal demand, and the highest for self-assessed performance.

Tables 1 and 2 present self-assessment and external assessment relative to the categories of the MoCA (Table 2) and relative to the equally spaced age groups (Table 1). A MANOVA for all self-assessments showed an overall significant main effect of the MoCA categories on the subscales ($F_{12,584}$=9.93; P<.001), with all scores rising from patients with low risk for MCI to patients with high risk for MCI, with the exception of the performance subscale, which showed an opposing trend. A MANOVA for all external assessments also showed a significant effect ($F_{12,584}$=20.18; P<.001) and the same pattern in the descriptive values.

A MANOVA on all self-assessments regarding the age groups showed an overall significant main effect of the age groups on the subscales ($F_{18,876}$=2.53; P<.001), with all scores rising from younger to older patients, except for the performance subscale, which showed an opposing trend. A MANOVA for all external assessments also showed a significant effect ($F_{18,876}$=4.38; P<.001) and the same pattern in the descriptive values.

Table 3. Mean values and SDs of self-assessment and external-assessment of National Aeronautics and Space Administration Task Load Index (NASA TLX) subscales including correlations of the assessments.

NASA TLX[a] subscale[b]	Self-assessment, mean (SD)	External assessment, mean (SD)	Correlation, *r*	*P* value
Mental workload	7.72 (5.50)	8.48 (6.24)	0.44	<.001
Physical workload	4.01 (4.76)	4.47 (5.21)	0.55	<.001
Temporal workload	4.75 (4.82)	5.55 (5.11)	0.54	<.001
Performance	11.3 (5.43)	11.3 (6.25)	0.57	<.001
Effort	8.38 (5.13)	9.09 (6.04)	0.44	<.001
Frustration	6.96 (5.79)	7.48 (6.15)	0.57	<.001

[a]NASA TLX: National Aeronautics and Space Administration Task Load Index.

[b]Values for the scores range from 1 to 20.

XSL•FO
RenderX

Secondary Variables

We found that 64% (191/299) of participants felt unrestricted in displaying their ability within the test. Of those who could not perform to the best of their ability, 41% (44/108) reported agitation as the main reason, 27% (29/108) thought the unusual tasks hindered them, and 17% (18/108) felt uncomfortable with being in a test situation. Additionally, 34% (37/108) reported other reasons in free text, which included anxiety, the language barrier, and physical ailments. The remaining categories were rarely selected and included 8% (9/108) problems understanding specific tasks and 1% (1/108) problems with the interaction.

Within the categories of the MoCA, patients within the group with low risk for MCI could show their full potential in 84% (84/100) of cases, for the within-threshold group this value dropped to 59% (41/69) and was lowest in the high-risk-for-MCI group with 51% (66/130).

Concerning affinity to technology, the sample was heterogenous, with 24% (72/299) reporting using their smartphone or tablet often in one day, 32% (97/299) using it regularly, 21% (63/299) using it rarely, and 22% (67/299) never using it. Table 4 presents the number of patients within the categories regarding cognitive status relative to their technology affinity. A subsequent chi-square test on technology affinity and the 3 categories for the MoCA scores indicated a significant association between both measures (N=299; χ^2_6=66.3; P<.001). Figure 1 shows the percentage contribution of deviation to the statistic, highlighting two main patterns (1) for patients with low technology affinity, the count in the low-risk group is very low, while it is high in the high-risk group (top row, columns 1 and 3); and (2) for patients with high technology affinity, the count is high in the low-risk group while it is low in the high-risk group (bottom row, columns 1 and 3).

Table 4. Cross table on technology affinity and categories on the Montreal Cognitive Assessment (MoCA) score.

Technology affinity	Low risk for MCI[a], n (%)	Within threshold, n (%)	High risk for MCI, n (%)
Never	4 (9)	11 (16)	52 (40)
Rarely	15 (15)	15 (22)	33 (25)
Regularly	39 (39)	25 (36)	33 (25)
Often	42 (42)	18 (26)	12 (9)

[a]MCI: mild cognitive impairment.

Figure 1. Percentage contribution of affinity with technology and Montreal Cognitive Assessment (MoCA) categories. Affinity to technology is indicated by smartphone or tablet use on a daily bases, with “never” indicating low affinity to technology and “often” indicating high affinity to technology. MCI: mild cognitive impairment.

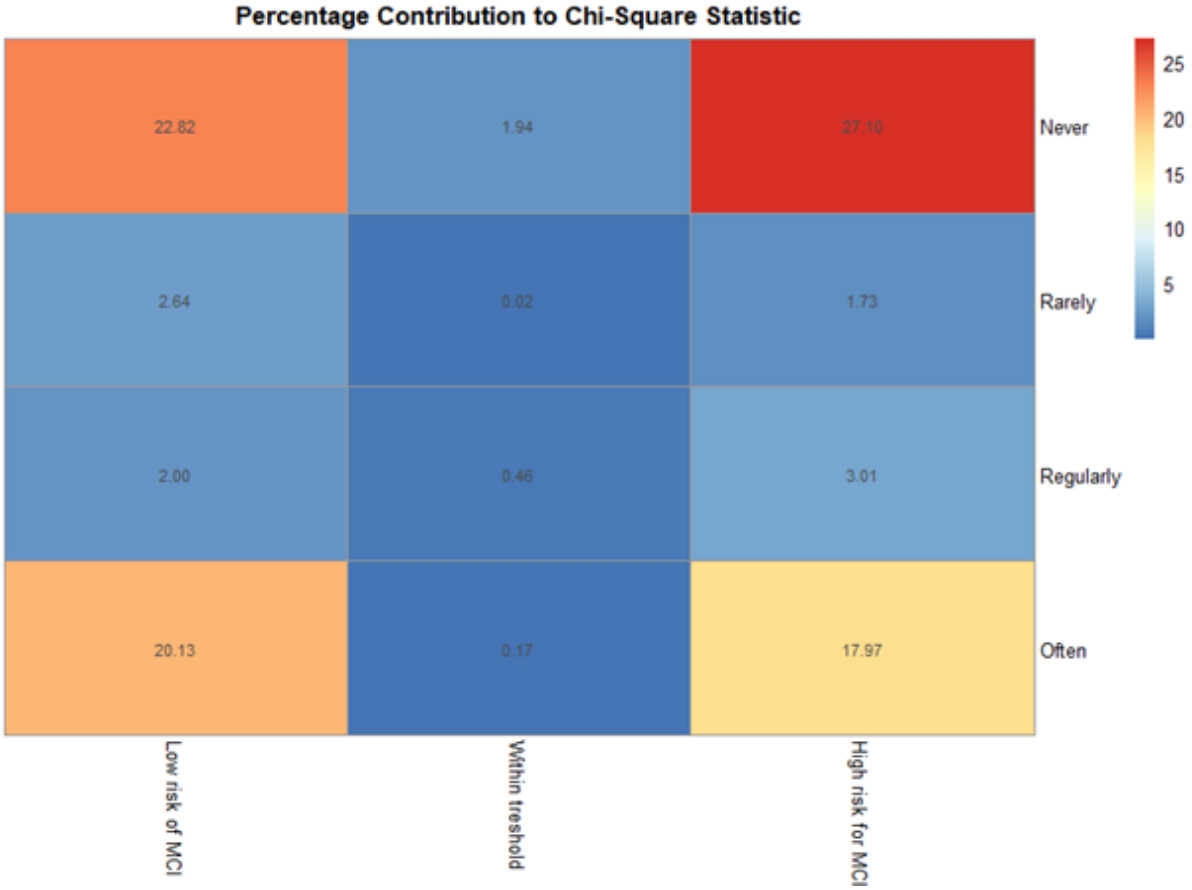

XSL•FO
RenderX

Regarding the comparison of the interaction with the app relative to their own device, a total of 31% (94/299) of patients considered the digital MoCA simpler, 33% (99/299) considered it of similar difficulty, 8% (23/299) found it more difficult, and 28% (83/299) reported not having a smartphone or tablet for comparison.

Discussion

Overview

In this evaluation study on the usability of the integration of a digital MoCA test to be conducted by MAs within GP practices, we found overall positive results. Our analyses show the good usability of the digitalized MoCA and support its feasible integration into the routine care setting in GP practices. As outlined in the review on mobile technology for cognitive assessment by Koo and Vizer [42], usability studies have to include patients who are cognitively impaired in addition to healthy controls. This study presents (to our knowledge) the first usability evaluation on a heterogenous sample including patients who are cognitively impaired for the MoCA Duo [27].

Findings

Within a heterogenous sample in terms of MoCA scores, that is, potential cognitive impairment, as well as different affinity to technology, we found that most patients could show their full potential (191/299, 64%) within the cognitive test while indicating good usability based on the SUS scores. Our analyses indicated significantly lower scores regarding usability for patients with higher risk for MCI, as well as for older patients. Both workload assessments were also significantly influenced by both factors. This further highlights the relevance of including patients who are cognitively impaired and older in evaluation studies.

The instances in which patients reported not being able to perform to the best of their ability were (at least in the self-assessment) not based on the usability or the fact that the test was digital, but rather on factors such as anxiety, which would also be relevant in the paper-based version. However, a study addressing this potential link found no indication of a detrimental effect of anxiety [43]. However, it is also noteworthy that within the group of patients with high risk for MCI the self-report of being able to perform to the best of their ability was lowest (66/130, 51%) compared to both other groups.

The assessment of the workload within the test showed a plausible pattern of results. We found high but not excessive mental workload within the scales for mental demand and effort and low physical demand. The self-assessment of performance was descriptively high, which appears understandable given how easy it is within the tasks of the MoCA to assess one's own performance. Because there are no time restrictions in the MoCA tasks, the resulting perceived time pressure was low. Critically, the assessment of the workload by the MA was positively correlated with the self-assessment of the patients. Besides the additional confidence in the results of the patients, this supports the role of MAs within GP practices, because their assessment, even though it might be informal, could serve as an additional indicator for subsequent treatment by the practitioner.

Considering the decline in SUS scores, the lower self-report of ability to perform to the best of one's own ability, and the increase in workload with increasing risk for MCI and age, it must be considered if the usability problems indicate problems with the app or rather the result of the cognitive decline. Given the present results, we suggest that the usability for the MoCA Duo is high for people not significantly cognitively challenged by the test, but with insufficient cognitive resources for the tasks, the usability is perceived as lower, indicated by the increased workload.

The diagnostic value of the paper-based MoCA has been demonstrated across different languages [44], and comparisons between digital and paper-based versions show high consistency [27], except in instances where physical interaction is necessary [40]. This highlights the relevance of taking experience with technology into account but also demonstrates the relevance of good usability in digital applications, especially for patients without experience with the respective technology.

We found a significant relationship between the MoCA assessment and the daily use of a smartphone or tablet as an indicator of technological affinity, the main differences in observed and expected values were especially prominent in the edges of the 3 × 4 table in Figure 1, with higher-than-expected values for patients with low risk for MCI often using their smartphone, as well as for patients with high risk for MCI never using such a device. Conversely, fewer-than-expected observations were made for patients with low risk for MCI never using a smartphone and patients with high risk for MCI using it often. Even though unfamiliarity with the technology might have a negative impact on the MoCA score [40], this would only account for the increase of patients who never use a smartphone with low MoCA scores. Furthermore, technology use has been previously shown as a predictor for MoCA scores [45]. Therefore, we suggest that the pattern of results might be better explained by patients with healthy cognition being more able and interested in learning and subsequently using smartphones, fitting the significant decline in MoCA scores with age. However, further study regarding this link is still necessary.

Limitations

Some limitations must be considered for this study. Even though the high number of participants is encouraging, there is likely to be a selection bias regarding which patients took part in the study due to subjective cognitive problems or based on the offer in GP practices. Additionally, self-selection to not participate in the study was not recorded and might have influenced the final sample and the subsequent results. This aspect must be addressed in future work in terms of an evaluation study assessing the applicability of the approach in standard health care.

Based on the large and diverse sample in this study, we do not expect issues regarding usability in future work. Other potential barriers such as proficiency in German, deficits with eyesight, or motor problems (ie, tremors) would have to be considered when administering the test. Due to the setting of the study, the validity of the MoCA scores was not tested against other measures, but based on the validation study by Klil-Drori et al

XSL•FO
RenderX

[30] and the plausible patterns of our secondary measures, we do not expect severe deviations from a paper-based MoCA.

Future Work

Future work should address the reported effects of technological affinity and the MoCA scores, especially regarding the digital assessment of the test. Finally, an evaluation of the integration of the cognitive test by MAs and the subsequent integration of the results into GP practices into standard health care should be conducted. Ideally, a standardized training module would be used to standardize the application of the test by MAs.

Overall, this implementation study further supports the positive outlook of integrating digital cognitive testing by MAs into GP practices as outlined by the results of other published work of the iCreate project, such as the results of qualitative interviews with practitioners, MAs, and patients. The ease of data collection across several GP practices also highlights the benefit of digital testing; this is further highlighted by the automatic analysis of the results, which allowed discussion of the performance immediately following the test. If adopted in standard health care, the test result should also be available to the specialist in case of further diagnostic testing, removing the need to repeat tests.

Acknowledgments

The authors thank all participating general practitioners (GPs), medical assistants (MAs), and especially patients for taking part in the study.

Funding

This study was funded by the Federal Ministry of Health (BMG), Germany (grant ZMII2-2523FEP40).

Data Availability

The datasets generated or analyzed during this study are not publicly available, as this has not been specifically allowed in the written informed consent of the participants, but they are available from the corresponding author on reasonable request.

Authors' Contributions

PS developed the quantitative evaluations in iCreate and wrote the majority of the manuscript. LN, LMR, and FS were responsible for the digital aspects in iCreate, such as tablet setup and survey programming. AH is the iCreate project coordinator, trained the MAs, presented the project to and corresponded with GPs, and provided further neuropsychological testing of patients in the memory clinic. KR and CR conducted the qualitative analysis focused on 1:1 interviews with patients, Mas, and GPs. JT conducted background research, contributed essentially to the development of the qualitative study and analysis, and wrote the ethics application. KW recruited the practices and provided practical medical expertise. MW, SJ, and JRT developed the iCreate project and accompanied it throughout the course. MW is the project leader of iCreate. JRT devised the basic idea of the quantitative evaluations and contributed to the manuscript. All authors gave feedback on previous versions of the manuscript, read and approved the final manuscript, and are accountable for all aspects of the work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Distribution of System Usability Scale (SUS) scores.
[PNG File , 15 KB - formative_v9i1e76010_app1.png]

References

1. GBD 2019 Dementia Forecasting Collaborators. Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the Global Burden of Disease Study 2019. Lancet Public Health 2022;7(2):e105-e125 [FREE Full text] [doi: 10.1016/S2468-2667(21)00249-8] [Medline: 34998485]
2. Blotenberg I, Hoffmann W, Thyrian JR. Dementia in Germany: epidemiology and prevention potential. Dtsch Arztebl Int 2023;120(27-28):470-476 [FREE Full text] [doi: 10.3238/arztebl.m2023.0100] [Medline: 37226316]
3. Jessen F, Bohr L, Kruse C, Dodel R. Die S3-Leitlinien Demenzen [The German S3 guidelines on dementia]. Nervenarzt 2023;94(7):609-613. [doi: 10.1007/s00115-023-01492-6] [Medline: 37233772]
4. Dubois B, Padovani A, Scheltens P, Rossi A, Dell'Agnello G. Timely diagnosis for Alzheimer's disease: a literature review on benefits and challenges. J Alzheimers Dis 2016;49(3):617-631 [FREE Full text] [doi: 10.3233/JAD-150692] [Medline: 26484931]

XSL•FO
RenderX

5. S3-Leitlinie Demenzen, Version 4.0. 2025. URL: https://www.dgppn.de/_Resources/Persistent/1f641e4edaf5c5d5a5114ee69146ba459a7da6b3/S3-Leitlinie+Demenzen_Langversion_2023_11_28_Final+%28003%29.pdf [accessed 2025-11-21]
6. Eichler T, Thyrian JR, Hertel J, Köhler L, Wucherer D, Dreier A, et al. Rates of formal diagnosis in people screened positive for dementia in primary care: results of the DelpHi-Trial. J Alzheimers Dis 2014;42(2):451-458 [FREE Full text] [doi: 10.3233/JAD-140354] [Medline: 24898640]
7. Eichler T, Thyrian JR, Hertel J, Michalowsky B, Wucherer D, Dreier A, et al. Rates of formal diagnosis of dementia in primary care: the effect of screening. Alzheimers Dement (Amst) 2015;1(1):87-93 [FREE Full text] [doi: 10.1016/j.dadm.2014.11.007] [Medline: 27239495]
8. Wangler J, Fellgiebel A, Jansky M. Das Praxispersonal im Kontext der hausärztlichen Demenzerkennung ? ein ungehobenes Potenzial? [The practice staff in primary care dementia recognition-is there an untapped potential?]. Z Gerontol Geriatr 2019;52(7):661-666. [doi: 10.1007/s00391-018-01484-1] [Medline: 30478791]
9. Early Detection and Diagnosis of Alzheimer's. Chicago: Alzheimer's Association; 2020.
10. Santacruz KS, Swagerty D. Early diagnosis of dementia. Am Fam Physician 2001 Feb 15;63(4):703-13, 717-718 [FREE Full text] [Medline: 11237085]
11. Wangler J, Jansky M. What requirements do primary care physicians have with regard to dementia diagnostics and dementia care? - a survey study among general practitioners in Germany 2022/2023. Arch Public Health 2023;81(1):155 [FREE Full text] [doi: 10.1186/s13690-023-01174-4] [Medline: 37620850]
12. Zimmermann GW. Untersuchungen bei Demenz richtig abrechnen. 2021. URL: https://www.medical-tribune.de/praxis-und-wirtschaft/artikel/untersuchungen-bei-demenz-richtig-abrechnen [accessed 2025-11-21]
13. Wangler J, Fellgiebel A, Jansky M. Hausärztliche Demenzdiagnostik ? Einstellungen, Vorgehensweisen und Herausforderungen von Hausärzten in Rheinland-Pfalz [Dementia diagnosis in general practitioner care - attitudes, procedures and challenges from the perspective of general practitioners in rhineland-palatinate]. Dtsch Med Wochenschr 2018;143(19):e165-e171. [doi: 10.1055/a-0651-1978] [Medline: 30231278]
14. Lancaster C, Koychev I, Blane J, Chinner A, Wolters L, Hinds C. Evaluating the feasibility of frequent cognitive assessment using the mezurio smartphone app: observational and interview study in adults with elevated dementia risk. JMIR Mhealth Uhealth 2020;8(4):e16142 [FREE Full text] [doi: 10.2196/16142] [Medline: 32238339]
15. Thompson LI, Harrington KD, Roque N, Strenger J, Correia S, Jones RN, et al. A highly feasible, reliable, and fully remote protocol for mobile app-based cognitive assessment in cognitively healthy older adults. Alzheimers Dement (Amst) 2022;14(1):e12283 [FREE Full text] [doi: 10.1002/dad2.12283] [Medline: 35415201]
16. Sabbagh MN, Boada M, Borson S, Chilukuri M, Dubois B, Ingram J, et al. Early detection of mild cognitive impairment (MCI) in primary care. J Prev Alzheimers Dis 2020;7(3):165-170. [doi: 10.14283/jpad.2020.21] [Medline: 32463069]
17. Düzel E, Thyrian JR. Mobile, alltagsnahe digitale Technologien für die Prävention der Alzheimer-Demenz: kognitive Gesundheit und kognitive Sicherheit [Mobile everyday-life digital technologies for the prevention of Alzheimer's dementia: Cognitive health and cognitive safety]. Nervenarzt 2023;94(5):400-407 [FREE Full text] [doi: 10.1007/s00115-023-01478-4] [Medline: 37115257]
18. Eichler T, Thyrian JR, Fredrich D, Köhler L, Wucherer D, Michalowsky B, et al. The benefits of implementing a computerized intervention-management-system (IMS) on delivering integrated dementia care in the primary care setting. Int Psychogeriatr 2014;26(8):1377-1385 [FREE Full text] [doi: 10.1017/S1041610214000830] [Medline: 24811145]
19. Blotenberg I, Boekholt M, Lieberknecht N, Säring P, Thyrian JR. Acceptance of unsupervised app-based cognitive assessment in outpatient care: an implementation study. JMIR Form Res 2025;9:e62706 [FREE Full text] [doi: 10.2196/62706] [Medline: 39946414]
20. Wangler J, Jansky M. The use of health apps in primary care-results from a survey amongst general practitioners in Germany. Wien Med Wochenschr 2021;171(7-8):148-156 [FREE Full text] [doi: 10.1007/s10354-021-00814-0] [Medline: 33570692]
21. Thabtah F, Peebles D, Retzler J, Hathurusingha C. Dementia medical screening using mobile applications: a systematic review with a new mapping model. J Biomed Inform 2020;111:103573 [FREE Full text] [doi: 10.1016/j.jbi.2020.103573] [Medline: 32961306]
22. Düzel E, Thyrian JR, Berron D. Innovation in der Diagnostik ? mobile Technologien [Innovation in diagnostics-mobile technologies]. Nervenarzt 2019;90(9):914-920. [doi: 10.1007/s00115-019-0773-8] [Medline: 31420690]
23. Armstrong S. Which app should I use? Patients and doctors are making increasing use of health apps, but there is little guidance about how well they work. Br J Sports Med 2017;51(16):1237-1239. [doi: 10.1136/bjsports-2016-g8027rep] [Medline: 28760947]
24. Dahlhausen F, Zinner M, Bieske L, Ehlers JP, Boehme P, Fehring L. Physicians' attitudes toward prescribable mHealth apps and implications for adoption in Germany: mixed methods study. JMIR Mhealth Uhealth 2021;9(11):e33012 [FREE Full text] [doi: 10.2196/33012] [Medline: 34817385]
25. Fillit HM, Simon ES, Doniger GM, Cummings JL. Practicality of a computerized system for cognitive assessment in the elderly. Alzheimers Dement 2008;4(1):14-21. [doi: 10.1016/j.jalz.2007.09.008] [Medline: 18631946]
26. MoCA Duo: MoCA Test Inc. 2024. URL: https://mocacognition.com/ [accessed 2025-11-21]

XSL•FO
RenderX

27. Berg J, Durant J, Léger GC, Cummings JL, Nasreddine Z, Miller JB. Comparing the electronic and standard versions of the montreal cognitive assessment in an outpatient memory disorders clinic: a validation study. J Alzheimers Dis 2018;62(1):93-97 [FREE Full text] [doi: 10.3233/JAD-170896] [Medline: 29439349]
28. Rolke K, Rosendahl C, Weckbecker K, Hanke A, Wagner M, Nissen L, et al. How do patients, medical assistants and physicians accept and experience tablet-based cognitive testing by medical assistants in general practice? A qualitative study. BMC Prim Care 2025;26(1):174. [doi: 10.1186/s12875-025-02823-z] [Medline: 40382582]
29. Nasreddine ZS, Phillips NA, Bédirian V, Charbonneau S, Whitehead V, Collin I, et al. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. J Am Geriatr Soc 2005;53(4):695-699. [doi: 10.1111/j.1532-5415.2005.53221.x] [Medline: 15817019]
30. Klil-Drori S, Bodenstein KC, Sun S, Kojok L, Gruber J, Ghantous Y, et al. Montreal Cognitive Assessment (MoCA) XpressO: validation of a digital self-administered cognitive prescreening tool. J Am Geriatr Soc 2024;72(8):2516-2522. [doi: 10.1111/jgs.18902] [Medline: 38558263]
31. Thomann AE, Berres M, Goettel N, Steiner LA, Monsch AU. Enhanced diagnostic accuracy for neurocognitive disorders: a revised cut-off approach for the Montreal Cognitive Assessment. Alzheimers Res Ther 2020;12(1):39 [FREE Full text] [doi: 10.1186/s13195-020-00603-8] [Medline: 32264975]
32. Bangor A, Kortum PT, Miller JA. The System Usability Scale (SUS): an empirical evaluation. Int J Hum-Comput Interact 2008;24(6):574-594. [doi: 10.1080/10447310802205776]
33. Peres SC, Pham T, Phillips R. Validation of the System Usability Scale (SUS). Proc Hum Factors Ergon Soc Annu Meet 2013;57(1):192-196. [doi: 10.1177/1541931213571043]
34. Holden RJ. A simplified System Usability Scale (SUS) for cognitively impaired and older adults. Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care 2020;9(1):180-182. [doi: 10.1177/2327857920091021]
35. Gao M, Kortum P, Oswald FL. Multi-language toolkit for the System Usability Scale. Int J Hum-Comput Interact 2020;36(20):1883-1901. [doi: 10.1080/10447318.2020.1801173]
36. de Souza GDPM, De Abreu JN, Sena RDO, Silva ADS, Silva JCC, Bitar WM, et al. Concordance, reproducibility, and usability of a Brazilian version of the Montreal Cognitive Assessment (MoCA) questionnaire in electronic format (appsheet) to screen cognitive impairment in older women. Arch Gerontol Geriatr Plus 2024;1(4):100088. [doi: 10.1016/j.aggp.2024.100088]
37. Hart SG. Nasa-Task Load Index (NASA-TLX); 20 years later. Proc Hum Factors Ergon Soc Annu Meet 2006;50(9):904-908.
38. Patel T, Ivo J, Faisal S, McDougall A, Carducci J, Pritchard S, et al. A prospective study of usability and workload of electronic medication adherence products by older adults, caregivers, and health care providers. J Med Internet Res 2020;22(6):e18073 [FREE Full text] [doi: 10.2196/18073] [Medline: 32348292]
39. Cornet VP, Daley CN, Srinivas P, Holden RJ. User-centered evaluations with older adults: testing the usability of a mobile health system for heart failure self-management. Proc Hum Factors Ergon Soc Annu Meet 2017;61(1):6-10 [FREE Full text] [doi: 10.1177/1541931213601497] [Medline: 30930610]
40. Wallace S, Donoso Brown EV, Simpson R, D'Acunto K, Kranjec A, Rodgers M, et al. A comparison of electronic and paper versions of the Montreal Cognitive Assessment. Alzheimer Dis Assoc Disord 2019;33(3):272-278. [doi: 10.1097/WAD.0000000000000333] [Medline: 31335458]
41. Brooke J. SUS- a quick and dirty usability scale. In: Jordan PW, Thomas B, Weerdmeester BA, McClelland L, editors. Usability Evaluation in Industry. Florida: CRC Press; 1996:189-194.
42. Koo BM, Vizer LM. Mobile technology for cognitive assessment of older adults: a scoping review. Innov Aging 2019;3(1):igy038 [FREE Full text] [doi: 10.1093/geroni/igy038] [Medline: 30619948]
43. Patel R, Nair A. Does anxiety affect performance on attention task (digit span forward) on the MOCA test? A clinical correlation study. Innov Aging 2020;4(Suppl 1):878.
44. Malek-Ahmadi M, Nikkhahmanesh N. Meta-analysis of montreal cognitive assessment diagnostic accuracy in amnestic mild cognitive impairment. Front Psychol 2024;15:1369766 [FREE Full text] [doi: 10.3389/fpsyg.2024.1369766] [Medline: 38414877]
45. Scanlon L, O'Shea E, O'Caoimh R, Timmons S. Usability and validity of a battery of computerised cognitive screening tests for detecting cognitive impairment. Gerontology 2016;62(2):247-252. [doi: 10.1159/000433432] [Medline: 26113397]

Abbreviations

GP: general practitioner
MA: medical assistant
MANOVA: multivariate analysis of variance
MCI: mild cognitive impairment
MMSE: Mini-Mental-State Examination
MoCA: Montreal Cognitive Assessment
NASA TLX: National Aeronautics and Space Administration Task Load Index

XSL•FO
RenderX

SUS: System Usability Scale

Edited by A Stone; submitted 15.04.25; peer-reviewed by M Bayen, P Nohl-Deryk; comments to author 26.09.25; accepted 21.10.25; published 22.12.25.

Please cite as:
Schaper P, Hanke A, Jonas S, Nissen L, Reimer LM, Schweizer F, Wagner M, Rolke K, Rosendahl C, Tillmann J, Weckbecker K, Thyrian JR
Usability of a Tablet-Based Cognitive Assessment Administered by Medical Assistants in General Practice: Implementation Study
JMIR Form Res 2025;9:e76010
URL: https://formative.jmir.org/2025/1/e76010
doi:10.2196/76010
PMID:

©Philipp Schaper, Alexander Hanke, Stephan Jonas, Leon Nissen, Lara Marie Reimer, Florian Schweizer, Michael Wagner, Kristin Rolke, Carolin Rosendahl, Judith Tillmann, Klaus Weckbecker, Jochen René Thyrian. Originally published in JMIR Formative Research (https://formative.jmir.org), 22.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Providing Mobile Patient Access to Their Electronic Secondary Care Patient Record in Adults With Cystic Fibrosis: Results of a Prospective, Parallel, Randomized Open-Pilot Quantitative Study

Helen K Chadwick[1,2], BSc, MSc, PhD; Akhil Sawant[2], MBBS, MSc, MRCP; Helen White[3], MSc, PGCHE, SRD, PhD; Lindsey Gillgrass[2], RGN; Giulia Spoletini[2], MD, PhD; Ian J Clifton[2], MD; Christine Etherington[2], MBChB; Daniel G Peckham[1,2], MBBS, DM

[1]Respiratory Medicine, Leeds Institute of Medical Research at St James's, University of Leeds, Leeds, West Yorkshire, United Kingdom
[2]Adult Cystic Fibrosis Unit, St James's University Hospital, Leeds Teaching Hospitals NHS Trust, Leeds, West Yorkshire, United Kingdom
[3]School of Clinical & Applied Science, Leeds Beckett University, Leeds, West Yorkshire, United Kingdom

Corresponding Author:
Helen K Chadwick, BSc, MSc, PhD
Respiratory Medicine
Leeds Institute of Medical Research at St James's
University of Leeds
Clinical Sciences Building
Leeds, West Yorkshire, LS2 9JT
United Kingdom
Phone: 44 113206 9106
Email: h.k.chadwick@leeds.ac.uk

Abstract

Background: The Leeds regional adult and pediatric cystic fibrosis (CF) services introduced a modified primary care electronic health care record (EHR) in 2007. This resulted in a dramatic improvement in efficiency while providing the benefits of primary care developments, including full Patient Access to their records.

Objective: This study aims to evaluate the feasibility, benefits, usability, and acceptability to patients of providing secure access to linked secondary care in CF.

Methods: A prospective, parallel, randomized, open, pilot study with an intervention (EHR access; I) and a control group (no EHR access; C). People with CF were recruited on a consecutive basis, from outpatient clinics or as inpatients on the regional Leeds adult CF unit. At baseline and 6 months, paper-based self-report questionnaires were completed by participants to assess having access to EHR on psychological impact, patient satisfaction, quality of life (QoL), patient and physician relationships, and patterns and rates of adherence to treatment. Perceptions and engagement with Patient Access and computer literacy were also assessed. Once completed, participants were randomized into either the intervention or control group (1:1), with those in the intervention group given instructions about how to gain access and the functions of Patient Access by the research team.

Results: A total of 91 people with CF completed the 6-month study (intervention n=45; median age 27.5, IQR 12.0 years; 22 male participants; control group n=46; median age 27.0, IQR 15.0 years; 29 male participants). Median number of logins was 9 (range 1-205). There was no effect of Patient Access on levels of anxiety (Generalized Anxiety Disorder-7; I=3.0, C=5.0), all symptom QoL scales and seven QoL domains (Cystic Fibrosis Questionnaire-Revised; respiratory I=63.89, C=55.56; weight I=100.00, C=66.67; digestion I=88.89, C=88.89; physical I=60.42, C=50.00; vitality I=54.17, C=41.67; emotional I=86.67, C=66.67; role I=75.0, C=75.0; body image I=77.78, C=66.67; eating disturbances I=88.89, C=100.0; treatment burden I=55.56, C=55.56), levels of depression (Patient Health Questionnaire-9; I=3.0, C=7.0), confidence in managing health care (Patient Activation Measure-13; I=66.67, C=60.63), level of trust in health care professionals (patient and provider perceived efficacy in patient-physician interaction; I=49.0, C=47.0), and computer literacy. Patient Access scored 86% for satisfaction, 82% for ease of use, and 80% for usefulness (Perceived Health Web Site Usability Questionnaire). Of those who had EHR access, 41 of 42 (98%) participants agreed that access to EHR should continue.

Conclusions: This pilot study suggests that providing access to EHR in adults with CF does not appear to have a negative effect (increase levels of anxiety or decrease QoL), and uptake by patients has been very positive. Prospective studies are needed to

XSL•FO
RenderX

investigate the long-term effect on objective health outcomes and how we can improve the functionality of such apps from the patient perspective.

Trial Registration: ClinicalTrials.gov NCT06122025; https://clinicaltrials.gov/study/NCT06122025

(JMIR Form Res 2025;9:e69747) doi:10.2196/69747

KEYWORDS

Patient Access; anxiety; quality of life; questionnaires; electronic health care records

Introduction

Over the past decade, a revolution in technology has fundamentally changed the way people live and how they access information. This new digital age has resulted in an exponential growth in social media and web-based access to services such as banking and shopping. The COVID-19 pandemic resulted in access to digital health solutions being implemented and adopted at unprecedented rates [1]. While digital technology can transform the way patients engage in health care, misinformation, misinterpretation, and information overload have the potential to negatively impact physical and mental health [1-4].

The National Health Service and other national health care services have recognized the importance of a more joined-up working with an integrated approach to health care services. This has the potential to improve safety and efficiency, clinical standards, and disease prevention, and empowers users to take control of their own health [3,5-7].

The vision for primary care was that all citizens should have digital access to their general practitioner records by 2015 [8]. This vision has been successful, and by 2017, 95% of general practitioner practices were able to offer digital access to detailed primary care records, including test results. Since 2020, all health care users are now able to access the full record, including note annotations, largely due to the COVID-19 pandemic and the significant improvement in remote health services [1,9]. This new transparency in record sharing has met some resistance from health care professionals due to concerns that accessing unfiltered medical information may lead to increased workload, patient anxiety, and potential litigation [3,10-12]. Equally, some patients have expressed anxiety over the potential for breaches in security and confidentiality [4].

In contrast to primary care, full digital access in secondary care is lagging [13], limited by the accessibility of platforms and the fragmentation and lack of digitized records.

In Leeds, the regional adult and pediatric cystic fibrosis (CF) services introduced a modified primary care electronic health care record (EHR) in 2007 and went paperless. This resulted in a dramatic improvement in efficiency while providing the benefits of primary care developments, including full Patient Access to their records [14]. However, the multidisciplinary team had concerns that providing patients with full access to their records might increase anxiety and affect the way the team recorded information. Therefore, prior to opening the portal to all patients, we performed a pilot study to evaluate the feasibility, benefits, and acceptability to patients of providing secure access to linked secondary care in CF. We also wanted to explore technological usability and the impact of the shared records on communication and patient satisfaction. It was hypothesized that providing Patient Access would have an effect on levels of anxiety and quality of life (QoL) scores in the intervention group.

Methods

Design

A prospective, parallel, randomized, open, pilot study was conducted between April 2018 and May 2019 (ClinicalTrials.gov Identifier: NCT06122025).

People with CF were recruited on a consecutive basis, either from routine CF outpatient clinics (Seacroft Hospital, Leeds, UK) or as an inpatient on the regional Leeds adult CF unit (St James's University Hospital, Leeds, UK). Eligible people were identified using the Leeds adult CF unit EHRs [14]. Inclusion criteria consisted of diagnosis of CF (confirmed by the presence of a CF phenotype with either two CF-causing mutations or a single mutation with two positive sweat chloride [>60 mmol/L]), aged 17 years or older, able to give written informed consent, and presence of an EHR system at the regional Leeds adult CF unit. Patients were excluded if they were taking part in a clinical trial.

Once consented, individuals were allocated to either the intervention or control group (1:1 ratio) by the research team using a randomized sampling technique. Random sequenced numbers were generated with the smallest value of 1 and the largest of 100, and even numbers were assigned to the intervention.

Measures

Overview

A baseline paper-based self-report questionnaire was used to collect demographic data, including age and gender.

Participants completed paper-based self-report questionnaires at week 0, and again at week 26±1, to evaluate the psychological impact, and effect on patient satisfaction, QoL, patient and physician relationships, and patterns and rates of adherence to treatment of having access to EHR in secondary care. Perceptions and engagement with Patient Access, and computer literacy were also assessed.

Primary Outcome Measures

Level of anxiety

The Generalized Anxiety Disorder-7 (GAD-7) is a validated questionnaire, further validated for use in CF, to assess the level

XSL•FO
RenderX

of anxiety through seven items covering the preceding 2 weeks [15,16].

QoL

The Cystic Fibrosis Questionnaire-Revised (CFQ-R) was used to assess QoL [17,18].

Secondary Outcome Measures

Severity of Depression

The Patient Health Questionnaire-9 (PHQ-9) is a validated questionnaire to measure the severity of depression [19], further validated for use in CF [16].

Knowledge, Skills, and Confidence in Managing Health

The Patient Activation Measure-13 (PAM-13) is a validated questionnaire to assess the knowledge, skills, and confidence a person with CF has in managing their health care [20].

Level of Trust and Interaction With Health Care Professionals

The patient and provider perceived efficacy in patient-physician interaction (PEPPI) is a validated questionnaire to assess the participants' level of trust and interaction with their health care professionals [21].

Computer Literacy

This was assessed using a questionnaire modified from the "My diabetes, my way" survey to assess types, frequency, and experience of computer or internet use.

Perception of and Intention to Engage With Patient Access

A questionnaire was devised to assess patients' thoughts about having access to their medical records in terms of reasons for accessing the records, expected effect of having access to them, and barriers to acceptance. This questionnaire was based on a modified OpenNotes presurvey [22] and the Physician and Patient Attitudes toward Technology in Medicine survey [23,24].

Perception of and Engagement With Patient Access (Intervention Group Only)

This questionnaire contained items from the baseline questionnaire modified to reflect having had access to their records for 6 months.

Satisfaction, Ease of Use, and Usefulness of Patient Access

The Perceived Health Web Site Usability Questionnaire (PHWSUQ; intervention group only) [25] was used to assess satisfaction, ease of use, and usefulness of the intervention.

Use Patterns

For the intervention group, data relating to the frequency of logins to their records was collected from the system audit trail.

A detailed description of the questionnaires used is available in Multimedia Appendix 1.

Procedures

Eligible adults with CF were approached face-to-face at either routine CF outpatient clinics (Seacroft Hospital, Leeds, UK) or as an inpatient on the regional Leeds adult CF unit (St James's University Hospital, Leeds, UK). If they expressed interest, they were provided with a participant information sheet by a member of the research team (who was also part of their direct care team), the study was explained to them, and they had the opportunity to ask questions and have them answered satisfactorily. Upon enrollment, participants completed the paper consent form and baseline paper-based self-report questionnaires in the following order: GAD-7, PHQ-9, PAM-13, CFQ-R, PEPPI, computer literacy, and "perceptions of and intention to engage with Patient Access." Participants were then randomized into either the intervention or control group. Those in the intervention group were instructed in how to gain access, and the functions of Patient Access were explained to them by a member of the research team (who was also part of their direct care team); access to current problems and medication, test requests, letters, consultations, allergies, and immunizations. Information about Patient Access security and privacy was also explained. No prompts were used in the study; those in the intervention used Patient Access at their discretion during the 6-month study period. Help and support were provided only if the participant requested them. After 26 (+/–1) weeks, participants completed the follow-up paper-based questionnaires (GAD-7, PHQ-9, PAM-13, CFQ-R, PEPPI, and computer literacy). Those in the intervention group also completed the end of intervention "Perception of Patient Access and their engagement" questionnaire and PHWSUQ on paper. Following successful completion of the questionnaires, participants in the control group were granted access to their EHR if they expressed interest.

Ethical Considerations

The study was approved by London—Camberwell St Giles Research Ethics Service (17/LO/19; 26/11/2017). Informed consent was obtained prior to any study-specific assessments. A participant information sheet and consent form were provided to potential participants, the content of the document was fully explained to them, and they were given enough time to consider the information. Participants had the opportunity to ask any questions about the study, and answers were provided to their satisfaction. A signed and dated copy of the informed consent form was provided to participants. All participants were provided with a unique study identifier after informed consent so that study data were anonymous. No compensation was provided for study participation.

Statistical Analysis

An a priori sample size calculation was performed using the following formula, assuming a 2-sided 5% significance level, a standardized effect size of 0.6, 80% power, and a dropout rate of 10%. The calculation indicated that 100 participants would be required, allocated to group on a 1 intervention:1 control, resulting in 50 individuals being recruited to the intervention (given mobile access to their own secondary care EHR) and 50 people to the control group (no access to their secondary care EHR).

Analysis was conducted separately for participants with data at both baseline and follow-up. Descriptive statistics were produced for computer literacy. Data were expressed as a number or median (IQR) as appropriate. For proportions, a Fisher exact test was used, and for measures, a Mann-Whitney

XSL•FO
RenderX

U test was used for between-subjects. No correction for multiplicity was applied. *P*<.05 was used as a guide to significance, and all computations used SPSS (v26.0 or higher; IBM Corp).

Results

Participants' Characteristics at Baseline

A total of 350 people met the inclusion criteria. In total, 102 patients were recruited, 51 in each group (Figure 1). An extra individual was recruited to each group to account for two people who dropped out shortly after consent and randomization, respectively. Baseline characteristics are shown in Table 1.

Figure 1. CONSORT flow diagram for patient enrollment, allocation, follow-up, and analysis. CONSORT: Consolidated Standards of Reporting Trials; EHR: electronic health care record.

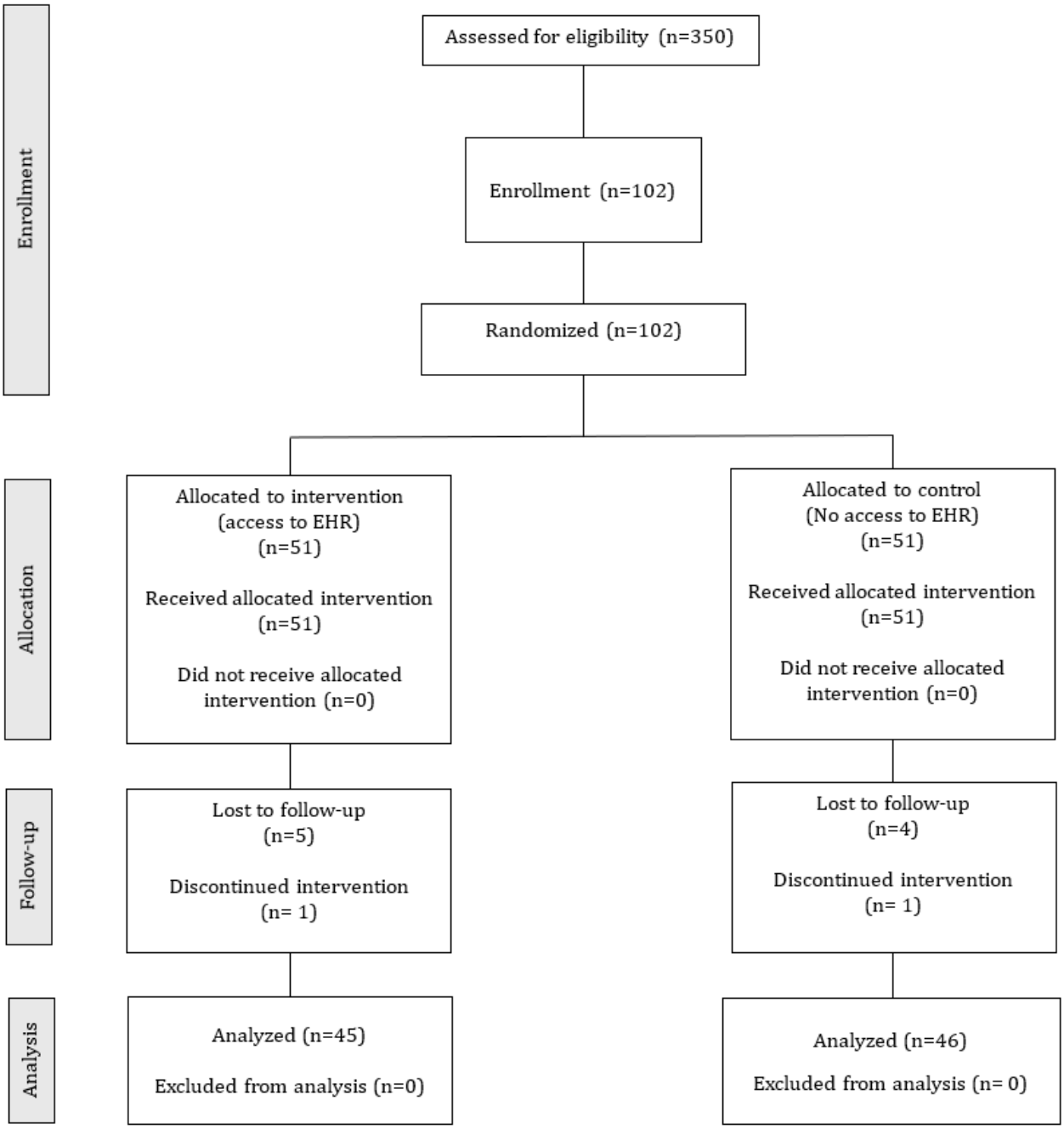

XSL•FO
RenderX

Table 1. Questionnaire scores for levels of anxiety (GAD-7[a]), QoL[b] (CFQ-R[c]), depression (PHQ-9[d]), knowledge, skills, and confidence in managing health (PAM-13[e]), and level of trust and interaction with health care professionals (PEPPI[f]), at baseline and follow-up (6 months) for the intervention and control groups.

	Baseline, mean (SD)			Follow-up, mean (SD)		
	Intervention group	Control group	*P* value	Intervention group	Control group	*P* value
Anxiety level (GAD-7)	4.0 (5.0)	3.0 (7.0)	≥.99	3.0 (5.0)	5.0 (5.0)	≥.99
QoL physical functioning (CFQ-R)	62.50 (51.04)	45.83 (62.50)	.37	60.42 (65.63)	50.00 (62.50)	.48
QoL vitality (CFQ-R)	50.0 (35.42)	41.67 (25.0)	.76	54.17 (41.67)	41.67 (41.67)	.35
QoL emotional functioning (CFQ-R)	73.33 (26.67)	66.67 (93.33)	.17	86.67 (38.33)	66.67 (43.33)	.22
QoL social functioning (CFQ-R)	72.22 (22.22)	61.11 (33.33)	<.001	66.67 (27.78)	61.11 (36.11)	.65
QoL role functioning (CFQ-R)	75.00 (33.33)	66.67 (33.33)	.15	75.00 (41.67)	75.00 (45.83)	.38
QoL body image (CFQ-R)	66.67 (55.56)	66.67 (44.44)	.29	77.78 (33.33)	66.67 (44.44)	.25
QoL eating disturbances (CFQ-R)	100 (22.22)	88.89 (22.22)	.30	88.89 (22.22)	100.0 (33.33)	.53
QoL treatment burden (CFQ-R)	55.56 (25.00)	55.56 (33.33)	.81	55.56 (33.33)	55.56 (38.89)	.33
QoL respiratory symptoms (CFQ-R)	55.56 (29.17)	61.11 (27.78)	.74	63.89 (33.33)	55.56 (38.39)	.90
QoL health perceptions (CFQ-R)	61.11 (33.33)	44.44 (44.44)	.04	55.56 (44.44)	55.56 (38.89)	.79
QoL weight (CFQ-R)	83.33 (66.67)	66.67 (66.67)	.14	100.0 (58.33)	66.67 (66.67)	.20
QoL digestion (CFQ-R)	88.89 (22.22)	88.89 (33.33)	.68	88.89 (3.33)	88.89 (33.33)	.34
Depression level (PHQ-9)	4.0 (6.0)	6.0 (8.0)	.17	3.0 (7.8)	7.0 (10.5)	.05
PAM-13	65.47 (17.53)	65.47 (19.20)	.24	66.65 (17.11)	60.63 (19.36)	.05
PEPPI	46.00 (10.00)	44.00 (10.00)	.26	49.00 (9.0)	47.00 (8.0)	.38

[a]GAD-7: Generalized Anxiety Disorder-7.

[b]QoL: quality of life.

[c]CFQ-R: Cystic Fibrosis Questionnaire-Revised.

[d]PHQ-9: Patient Health Questionnaire-9.

[e]PAM-13: Patient Activation Measure-13.

[f]PEPPI: patient and provider perceived efficacy in patient-physician interaction.

Groups were similar for age (intervention: 27.5, IQR 14.0 years vs control 27, IQR 15.0 years; *P*=.96) and gender distribution (intervention 27 male participants, control 32 male participants; *P*=.32).

At baseline, there was no difference between groups for levels of anxiety (GAD-7), the three-symptom scales (respiratory, weight, and digestion), and seven of the nine QoL (CFQ-R) domains (physical, vitality, emotional, role, body image, eating disturbances, and treatment burden), levels of depression (PHQ-9), knowledge, skills, and confidence a person with CF has in managing their health care (PAM-13), or their level of trust in and interactions with their health care professionals (PEPPI; Table 1). Those in the intervention group had higher levels of health perceptions and social functioning (*P*=.04 and *P*<.001, respectively). One participant in the intervention group did not complete the baseline computer literacy questionnaire. Computer literacy was similar across groups (Table 2).

XSL•FO
RenderX

Table 2. Assessment of computer skills (literacy) for the intervention and control group at baseline and follow-up (6 months).

	Baseline, n (%)		Follow-up, n (%)	
	Intervention	Control	Intervention	Control
How would you describe your computer skills?[a]				
Poor	0 (0)	2 (4)	0 (0)	3 (6)
Moderate	2 (4)	2 (4)	3 (7)	3 (6)
Average	14 (28)	11 (22)	10 (22)	9 (20)
Good	13 (26)	13 (25)	15 (33)	15 (33)
Very good	21 (42)	23 (45)	17 (38)	16 (35)
Which of the following devices do you own?[a]				
Desktop computer	13 (26)	10 (20)	14 (31)	13 (28)
Laptop computer	23 (46)	23 (45)	33 (73)	30 (65)
“Internet-ready” smartphone	46 (92)	49 (96)	40 (89)	43 (93)
Computer tablet (eg, iPad)	35 (70)	32 (63)	28 (62)	25 (54)
How often do you use these devices?[a]				
Everyday	48 (96)	47 (92)	44 (98)	45 (98)
Several times a week	0 (0)	3 (6)	1 (2)	0 (0)
Several times a month	2 (4)	0 (0)	0 (0)	1 (2)
Less than monthly	0 (0)	1 (2)	0 (0)	0 (0)

[a]One intervention participant did not complete the question at baseline.

At baseline, 48 of 50 (96%) participants assigned to the intervention and 46 of 50 (92%) controls agreed that Patient Access was a good idea (Table 3). In terms of access, 45 of 50 (90%) patients in the intervention and 42 of 50 (84%) controls thought that patients have a right to see all consultations (rather than health care professionals only sharing what they think is appropriate) and the majority of each group thought that having access would improve the patient-health care relationship (n=47 intervention, n=43 control). Reasons for wanting access to their records were because they were curious (n=29 intervention, n=33 control), they wanted to see test results (n=40 intervention, n=36 control), they wanted to know about their health (n=44 intervention, n=40 control), and they had a right to see what was in their record (n=23 intervention, n=30 control). Over half of the individuals gave no particular reason for not wanting to access their record (n=28 intervention, n=27 control), whereas the most popular specific reasons were worrying about the privacy of information (n=6 intervention, n=8 control) and because they were afraid about finding something negative about their health which they were not aware about (n=6 intervention, n=8 control).

XSL•FO
RenderX

Table 3. Perceptions of and intention to engage with Patient Access (based on a modified OpenNotes presurvey and the Physician and Patient Attitudes toward Technology in Medicine survey) for the intervention and control group at baseline; reasons for accessing the records, expected effect of having access, and barriers to acceptance[a].

	Intervention, n (%)	Control, n (%)	*P* value
In general, making EHRs[b] available to people with CF[c] (ie, having Patient Access) is a good idea?			.74
Disagree	0 (0)	0 (0)	
Somewhat disagree	0 (0)	0 (0)	
Somewhat agree	2 (4)	3 (6)	
Agree	48 (96)	46 (92)	
Don't know	0 (0)	1 (2)	
How much access should patients have to consultations?			.67
A patient has a right to see them all	45 (90)	42 (84)	
Health care professionals should only share what they think is appropriate	5 (10)	8 (16)	
Why might you like to be able to read your EHRs (tick or cross all that apply)?			
I am curious	29 (58)	33 (66)	
I have the right to see what's in my medical record	23 (46)	30 (60)	
I want to see the diagnosis or test results	40 (80)	36 (72)	
I want to check the notes for accuracy	11 (22)	10 (20)	
I want to know about my health	44 (88)	40 (80)	
Other	9 (18)	1 (2)	
Why might you NOT like to be able to access your EHRs (tick/cross all that apply)?			
I don't think it would be useful	0 (0)	0 (0)	
I do not use the internet very much	0 (0)	0 (0)	
I worry about the privacy of information	6 (12)	8 (16)	
I do not need to see test results	0 (0)	1 (2)	
I do not want to think about my health any more than I have to	1 (2)	8 (16)	
I am afraid I will find out something bad about my health that I didn't know	6 (12)	8 (16)	
I do not need to see what health care professionals wrote about my visit	2 (4)	0 (0)	
No particular reason	28 (56)	27 (54)	
Other	2 (4)	4 (8)	
Do you think you might show or discuss your EHRs with other people?			.13
Yes	38 (76)	27 (54)	
No	4 (8)	10 (20)	
Don't know	8 (16)	13 (26)	
With whom do you think you might share your EHRs (tick/cross all that apply)?			
Partner	40 (80)	28 (56)	
Family friend or relative	36 (72)	32 (64)	
Friend	11 (22)	8 (16)	
Doctor (outside of cystic fibrosis)	20 (40)	19 (38)	
Nurse or Health care professional (outside of cystic fibrosis)	16 (32)	11 (22)	
Someone else	5 (10)	2 (4)	
Having access to my EHR would improve the patient-health care professional relationship?			.54
Agree	47 (94)	43 (86)	
Disagree	3 (6)	6 (12)	

XSL•FO
RenderX

	Intervention, n (%)	Control, n (%)	*P* value
Don’t know	0 (0)	1 (2)	
I would better understand my health and medical conditions			.48
Disagree	1 (2)	1 (2)	
Somewhat disagree	2 (4)	3 (6)	
Somewhat agree	15 (30)	14 (28)	
Agree	32 (64)	28 (56)	
Don’t know	0 (0)	4 (8)	
I would better remember the plan for my care			.69
Disagree	1 (2)	2 (4)	
Somewhat disagree	0 (0)	2 (4)	
Somewhat agree	10 (20)	11 (22)	
Agree	38 (76)	33 (66)	
Don’t know	1 (2)	2 (4)	
I would feel more in control of my health care			.28
Disagree	1 (2)	4 (8)	
Somewhat disagree	1 (2)	4 (8)	
Somewhat agree	11 (22)	9 (18)	
Agree	37 (74)	31 (62)	
Don’t know	0 (0)	2 (4)	
I would worry more			.70
Disagree	25 (50)	22 (44)	
Somewhat disagree	14 (28)	13 (26)	
Somewhat agree	5 (10)	8 (16)	
Agree	0 (0)	2 (4)	
Don’t know	6 (12)	5 (10)	
I would be concerned about my privacy			.60
Disagree	30 (60)	25 (50)	
Somewhat disagree	10 (20)	12 (24)	
Somewhat agree	10 (20)	10 (20)	
Agree	0 (0)	1 (2)	
Don’t know	0 (0)	2 (4)	
I would be concerned about security of my record			.53
Disagree	28 (56)	22 (44)	
Somewhat disagree	9 (18)	12 (24)	
Somewhat agree	13 (26)	13 (26)	
Agree	0 (0)	1 (2)	
Don’t know	0 (0)	2 (4)	
The information would be more confusing than helpful			.75
Disagree	25 (50)	21 (42)	
Somewhat disagree	16 (32)	15 (30)	
Somewhat agree	5 (10)	5 (10)	
Agree	1 (2)	1 (2)	

XSL•FO
RenderX

	Intervention, n (%)	Control, n (%)	*P* value
Don't know	3 (6)	8 (16)	
It could make my doctor's job more difficult			.60
Disagree	17 (34)	21 (42)	
Somewhat disagree	18 (36)	14 (28)	
Somewhat agree	9 (18)	5 (10)	
Agree	0 (0)	1 (2)	
Don't know	6 (12)	9 (18)	
Learning how to use and access to my EHR will be easy for me			.99
Strongly disagree	2 (4)	2 (4)	
Disagree	1 (2)	1 (2)	
Neither agree nor disagree	4 (8)	5 (10)	
Agree	19 (38)	18 (36)	
Strongly agree	24 (48)	23 (46)	
I have the resources necessary to use and have access to my EHR (eg, the internet)			.16
Strongly disagree	3 (6)	1 (2)	
Disagree	0 (0)	0 (0)	
Neither agree nor disagree	0 (0)	4 (8)	
Agree	10 (20)	8 (16)	
Strongly agree	37 (74)	36 (72)	
I plan to use and access to my EHR in the next 6 months			.02
Strongly disagree	2 (4)	2 (4)	
Disagree	0 (0)	2 (4)	
Neither agree nor disagree	1 (2)	8 (16)	
Agree	17 (34)	7 (14)	
Strongly agree	30 (60)	30 (60)	

[a]One intervention participant and one control participant did not complete the question at baseline.

[b]EHR: electronic health care record.

[c]CF: cystic fibrosis.

Over 50% (38/50 intervention; 27/50 control) of patients in each group thought that they might show or discuss their records with other people, such as their partner (n=40/50 intervention, n=28/50 control) or family friend or relative (n=36/50 intervention, n=32/50 control).

Prior to being randomized, the majority agreed (reporting either somewhat agreed or agreed) that they plan to use and access their records for the intervention period (n=47 intervention, n=37 control), that having access would help them understand their health and medical conditions (n=47 intervention, n=42 control), they would remember their care plan better (n=48 intervention, n=44 control), feel more control of their health care (n=48 intervention, n=40 control), and that learning how to use Patient Access would be easy for them (n=43 intervention, n=41 control) because they had the resources (n=47 intervention, n=44 control). The majority also disagreed (reporting either somewhat disagreed or disagreed) that as a result of having access, they would worry more (n=39 intervention, n=35 control), be concerned about privacy (n=40 intervention, n=37 control), or security (n=37 intervention, n=34 control), information would be more confusing than helpful (n=41 intervention, n=36 control), and it would make the doctor's job more difficult (n=35 intervention, n=35 control).

Follow-Up

In total, 91 patients completed the 6-month study. No differences in age or gender were observed across the two groups, considering only those who had completed the study (intervention n=45; median age 27.5, IQR 12.0 years; 22 male participants; control group n=46, median age 27.0, IQR 15.0 years; 29 male participants; *P*=.17).

There was no evidence of an effect of Patient Access on levels of anxiety (GAD-7), the three symptom QoL scales (respiratory *P*=.90; weight *P*=.20; and digestion *P*=.39), seven of the nine QoL (CFQ-R) domains (physical *P*=.48; vitality *P*=.35; emotional *P*=.22; role *P*=.38; body image *P*=.25; eating

XSL•FO
RenderX

disturbances P=.53; and treatment burden P=.33), levels of depression (PHQ-9), confidence in managing health care (PAM-13), their level of trust in health care professionals (PEPPI), or computer literacy (Tables 1 and 2). The QoL domain "social functioning" decreased by six points in the intervention but remained stable in the control group. The health perceptions domain (CFQ-R) decreased in the intervention by six points but increased in the control group by nine points at follow-up.

In the intervention group, 41 of the 42 (98%) participants agreed that access was still a good idea and wanted to continue having access, respectively (Table 4). At baseline, patients planned to use and access their records over the 6-month intervention period (n=35), and did so because they were curious (n=25), wanted to see test results (n=26) and know about their health (n=18), they have a right to see what is in their record (n=16), remember what happened at a clinic visit (n=12), be sure they understood what the health professional said (n=10), and know what the professional was thinking (n=6). Over the 6-month intervention, eight people did not log into Patient Access at all. Reasons for not doing so were that they had forgotten they had access (n=3), they did not need to see test results (n=2), no particular reason (n=3), and due to worry about the privacy of information (n=1).

XSL•FO
RenderX

Table 4. Perceptions of and engagement with Patient Access by the intervention group at follow-up (having had access for 6 months).

	Value, n (%)
In general, making EHRs[a] available to people with CF[b] (ie, having Patient Access) is a good idea? (3 people did not answer)	
Disagree	0 (0)
Somewhat disagree	0 (0)
Somewhat agree	2 (5)
Agree	39 (93)
Don't know	1 (2)
Did you log in and access your EHR at any point during the past 6 months? (4 people did not answer)	
No	8 (20)
Yes	33 (80)
Why might you like to be able to read your EHRs (tick/cross all that apply)?	
I was curious	25 (56)
I have the right to see what's in my medical record	16 (36)
I wanted to see the diagnosis/ test results	26 (58)
I wanted to know what my health professional was thinking	6 (13)
I wanted to check my record for accuracy	5 (11)
I wanted to be sure I understood what the health professional said	10 (22)
I wanted to remember what happened during the visit	12 (27)
I wanted to know about my health	18 (40)
No particular reason	0 (0)
Other	3 (7)
Why did you not access your EHRs (tick/cross all that apply)?	
I didn't think it would be useful	0 (0)
I do not use the internet very much	0 (0)
I thought reading the record would make me nervous or anxious	0 (0)
I do not want to think about my health more than I have to	0 (0)
I am afraid I will find out something bad about my health that I didn't know	0 (0)
I do not need to see what health care professionals wrote about my visit	0 (0)
I forgot I could access my health care records online	3 (7)
I worry about the privacy of information	1 (2)
I do not need to see test results	2 (4)
No particular reason	3 (7)
Other	3 (7)
How easy was it to understand your health care records? (5 people not answer)	
Very difficult	0 (0)
Somewhat difficult	3 (8)
Someway easy	12 (30)
Very easy	20 (50)
Don't know	5 (12)
Did you ever contact the CF unit about something in your health care record? (6 people did not answer)	
Yes	5 (13)
No, I did not feel any need to	32 (84)
I considered it, but decided not to	1 (3)

XSL•FO
RenderX

	Value, n (%)
Don’t know/ don’t remember	0 (0)
(If applicable) Why did you decide to contact your health care professional about something in your health care records (38 people did not answer)	
I wanted an explanation, for example, of a test result	4 (57)
I wanted something removed from my record	0 (0)
I wanted to report something I thought was an error in my record	1 (14)
I wanted to discuss something I disagreed with	0 (0)
Another reason	2 (29)
Were you satisfied with the health care professional’s response to your request? (38 people did not answer)	
Yes	7 (100)
Somewhat	0 (0)
No	0 (0)
Did you show, discuss, or share your health care record with other people? (6 people did not answer)	
Yes	27 (69)
No	11 (28)
Don’t know/don’t remember	1 (3)
With whom do you show/ discuss/share your health care records? (6 people did not answer)	
My partner	21 (54)
A family friend or relative	23 (59)
A friend	4 (10)
A doctor (outside of cystic fibrosis care)	3 (8)
A nurse or health care professional (outside of cystic fibrosis care)	2 (5)
Someone else	2 (5)
I would like to continue having access to my EHR (5 people did not answer)	
Yes	39 (98)
No	1 (2)
I understand my health and medical conditions better? (5 people did not answer)	
Disagree	0 (0)
Somewhat disagree	2 (5)
Somewhat agree	8 (20)
Agree	26 (65)
Don’t know	4 (10)
I remember to plan for my care better (2 people did not answer)	
Disagree	1 (2)
Somewhat disagree	3 (7)
Somewhat agree	11 (26)
Agree	21 (49)
Don’t know	4 (9)
Don’t know	3 (7)
I feel more in control of my health care (5 people did not answer)	
Disagree	0 (0)
Somewhat disagree	6 (15)
Somewhat agree	7 (18)

XSL•FO
RenderX

	Value, n (%)
Agree	26 (65)
Don’t know	1 (2)
I am concerned about my privacy (5 people did not answer)	
Disagree	26 (65)
Somewhat disagree	11 (28)
Somewhat agree	2 (5)
Agree	0 (0)
Don’t know	1 (2)
I am concerned about the security of my record (5 people did not answer)	
Disagree	26 (65)
Somewhat disagree	9 (23)
Somewhat agree	4 (10)
Agree	0 (0)
Don’t know	1 (2)
I felt offended (5 people did not answer)	
Disagree	34 (85)
Somewhat disagree	3 (8)
Somewhat agree	2 (5)
Agree	0 (0)
Don’t know	1 (2)
Did reading your EHRs, specifically the consultations, change the way you feel about health care professionals? (5 people did not answer)	
Yes, felt much worse	0 (0)
Yes, felt much better	9 (23)
No	25 (62)
Don’t know	6 (15)
(At the start of the study) I planned to use and access to my EHR in the next 6 months (5 people did not answer)	
Strongly disagree	0 (0)
Disagree	0 (0)
Neither agree nor disagree	5 (13)
Agree	6 (15)
Strongly agree	29 (72)

[a]EHR: electronic health care record.

[b]CF: cystic fibrosis.

Patient Access scored 86% for satisfaction, 82% for ease of use, and 80% for usefulness. The median number of logins over the study period was 9 (range 1-205). For those who did use Patient Access, patients agreed that they understood their CF better (n=34), felt more in control of their health (n=33), that information was easy to understand (n=32), and they could plan for their health condition better (n=32). There were no privacy or security concerns reported by 37 and 35 participants, respectively. Reading the consultations did not impact the way 25 participants felt about health care professionals; 9 participants felt much better, and 6 participants had not considered whether their feelings had changed. Seven patients contacted the CF Unit to discuss something in their record, with four people wanting an explanation of a test result, one patient wanted to report something they thought was an error, and two patients cited another reason not provided on the questionnaire. All were satisfied with the health care professional’s response. Two-thirds of patients (n=27) had shown, discussed, or shared their record with other people, with the majority reporting sharing it with their partner (n=21) or family member (n=23). Five participants had shown it to health care professionals outside of CF.

XSL•FO
RenderX

Discussion

Principal Findings

It was hypothesized that providing Patient Access would have an effect on levels of anxiety and QoL scores in the intervention group. A total of 91 adults with CF completed the 6-month study, and the median number of logins for those in the intervention group over the study period was 9 (range 1-205). There was no effect of Patient Access in the intervention group on levels of anxiety, all symptom QoL scales (respiratory, weight, and digestion), and seven QoL domains (physical, vitality, emotional, role, body image, eating disturbances, and treatment burden).

In contrary to the literature [26], levels of anxiety were not increased in our population and may reflect the chronic nature of the disease and the regular and close communication with a familiar multidisciplinary team. The slight decrease in anxiety in the intervention group after 6 months supports that Patient Access may decrease anxiety among patients [27].

At follow-up, the QoL domain "social functioning" had decreased by six points in the intervention but remained stable in the control group. To date, no other literature has reported a decrease in social functioning in response to Patient Access. This QoL domain measures how CF affects a person's ability to participate in social activities, maintain relationships, and engage in social interactions. Viewing their health care record may have made the patient reflect on their health status and decreased their participation in activities, impacting social interactions and relationships. It is unknown if any participants were experiencing a pulmonary exacerbation, associated with a decline in lung function, hospitalizations, and reduced QoL [28], at the time of follow-up, which may have contributed to the decrease in social functioning. However, as the respiratory domain increased at follow-up for the intervention group, this reason is unlikely.

Interestingly, the QoL health perceptions domain score had decreased in the intervention by six points but increased in the control group by nine points at follow-up. The decrease in the intervention group might reflect the improved understanding of their CF and ability to review their health status. In contrast, the control group might have been optimistic about their health, potentially due to an improvement in physical functioning, and adopting a coping style that has been shown to be an independent predictor of survival in CF [29].

Although all patients had relatively high activation levels (knowledge, skills, and confidence in managing their own health and care) at baseline, and were similar between groups, the results from the PAM-13 questionnaire suggest that having access to their EHR may help sustain and reinforce this. At follow-up, those in the control group had significantly lower activation levels than the intervention group. This supports the literature showing that Patient Access can empower individuals in managing their own health [5,30,31].

Research suggests that uptake of Patient Access is affected by privacy and security concerns [6,32]. In our cohort, this was true for one participant in the intervention group who never accessed their record for this reason. Nevertheless, the opportunity to access their records was positively received by the majority of the patients who completed the study. In addition, the percentage of people who had privacy and security concerns, respectively, decreased over the 6 months (10/50, 20% of participants and 13/50, 26% of participants vs 2/40, 5% of participants and 4/40, 10% of participants).

On average, participants in the intervention group accessed their records more than once a month, which suggests that information contained in their EHR may need to be accessed outside of their clinic visit, and the convenience reduces the need for the patient to contact the CF unit. Increased workload is cited in the literature as a concern by health care professionals as a result of patients having access [3]. Seven patients contacted the CF Unit during the intervention in relation to their records. Although the method of communication (telephone, text, email, and face-to-face) or impact on the health care professional's workload is unknown, it is positive that only a relatively small percentage of patients made contact, and all patients were satisfied with the professional's response.

The level of trust in, and interactions with, their health care professionals and levels of self-efficacy (confidence) in self-care and self-management did not change for the majority of those with access. For those who had considered the impact of reading consultations, 22% said they felt much better toward health care professionals afterwards, therefore enhancing patient-provider communication. However, it is unknown if health care professionals altered their consultation writing style, knowing patients would now be able to view their records [33,34].

Over the next few years, there is likely to be a significant change in the sharing of medical information between patients and their health care providers. This transference of key information will be bidirectional and include clinical and physiological data collected on individual mobile devices within the home and work environment. In the United Kingdom, secure access to personal health records is now well established in primary care. However, access to secondary care records remains limited, due in part to the underinvestment in hospital IT infrastructure and the presence of multiple legacy systems. A key driving force for the expansion of digital access is the need to deliver health information transparency, improve health care, quality of records, and empower patients to comanage their own health [35-37]. Changes in digital information exchange of personal medical information have the potential to have positive effects, such as greater patient empowerment, ownership, and improved knowledge of the underlying conditions [30,32,38,39]. This was reflected in this study, with patients perceiving that digital access aided them to better understand their condition, plan for their care better, and feel more in control of their health, while providing medical information that was relatively easy to understand and proved more helpful than confusing.

The use of digital access to EHR is likely to vary between individuals, with uptake being influenced by accessibility, digital literacy, health condition, and age of the user. Previous studies have reported improved patient engagement following the introduction of digital access to EHR, although the design and function of the portal can be perceived as nonpatient-centric

XSL•FO
RenderX

and inadequate by users [40]. In this study, patients gave a high score for satisfaction, with ease of use and usefulness slightly lower, suggesting that improvement in usability and functions may be needed. All but one patient in the intervention group agreed that Patient Access was a good idea and wished to continue with access after the 6 months. Since the end of the intervention, access to EHR has been granted to all adult patients with CF at the Leeds CF unit who wish to have this resource.

Strengths and Limitations

In contrast to primary care, secondary care is in the early stages of offering full digital access. This is the first study to provide secure access to secondary care records in CF, which has incorporated the feedback about which aspects of their EHR people with CF wish to access and priorities for development [41].

Participants were mostly White British, which is reflective of CF being a Caucasian disease [42], and in young-middle adulthood. These factors may have influenced the positive uptake of the app, as older adults are less likely to use digital services [43,44]. The findings from the PHWSUQ suggest that ease of use and usefulness could be improved and highlight the importance of having an intuitive, simple interface [45]. However, as Patient Access is a UK-based mobile app, updating the app functionality is outside the remit of the research team.

Conclusions

Our results suggest that increased patient information sharing through Patient Access to EHR is beneficial and desirable to patients, and should be implemented in other disease areas in secondary care where possible. There was no evidence of an effect of Patient Access on levels of anxiety and overall QoL, but improved subjective understanding, engagement, and control in the management of their CF. Prospective studies are needed to investigate the long-term effect of such interventions on objective health outcomes and how we can improve the functionality of such apps from the patient perspective.

Acknowledgments

The authors would like to thank all the participants involved for their time and cooperation. Previous research sought patient feedback regarding which aspects of their electronic health care record people with cystic fibrosis (CF) wish to access, and their priorities for development [41]. This was incorporated into the design of this study. People with CF were also involved in the design of this study and were asked to assess the burden of questionnaire completion.

Funding

This research was supported by a UK Cystic Fibrosis Trust Clinical Excellence and Innovation Award. Award or grant number is not applicable.

Data Availability

The datasets analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

HKC was involved in the conception, design, acquisition, data analysis, interpretation, drafting, and critical revision of the work and approved the final manuscript. AS, GS, IJC, and CE were involved in the design, acquisition, and critical revision of the work and approved the final manuscript. HW was involved in the conception, design, and critical revision of the work and approved the final manuscript. LG was involved in the acquisition and approved the final manuscript. DGP was involved in conception, design, data interpretation, critical revision, and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed description of questionnaires.
[DOCX File , 23 KB - formative_v9i1e69747_app1.docx]

Multimedia Appendix 2
CONSORT eHEALTH checklist (V 1. 6. 1).
[DOCX File , 38 KB - formative_v9i1e69747_app2.docx]

References

1. Hägglund M, McMillan B, Whittaker R, Blease C. Patient empowerment through online access to health records. BMJ 2022;378:e071531 [FREE Full text] [doi: 10.1136/bmj-2022-071531] [Medline: 36175012]

XSL•FO
RenderX

2. Blease CR, Delbanco T, Torous J, Ponten M, DesRoches CM, Hagglund M, et al. Sharing clinical notes, and placebo and nocebo effects: can documentation affect patient health? J Health Psychol 2022;27(1):135-146. [doi: 10.1177/1359105320948588] [Medline: 32772861]
3. Mold F, de Lusignan S, Sheikh A, Majeed A, Wyatt JC, Quinn T, et al. Patients' online access to their electronic health records and linked online services: a systematic review in primary care. Br J Gen Pract 2015;65(632):e141-e151 [FREE Full text] [doi: 10.3399/bjgp15X683941] [Medline: 25733435]
4. Fernández L, Fossa A, Dong Z, Delbanco T, Elmore J, Fitzgerald P, et al. Words matter: what do patients find judgmental or offensive in outpatient notes? J Gen Intern Med 2021;36(9):2571-2578 [FREE Full text] [doi: 10.1007/s11606-020-06432-7] [Medline: 33528782]
5. Delbanco T, Walker J, Bell SK, Darer JD, Elmore JG, Farag N, et al. Inviting patients to read their doctors' notes: a quasi-experimental study and a look ahead. Ann Intern Med 2012;157(7):461-470 [FREE Full text] [doi: 10.7326/0003-4819-157-7-201210020-00002] [Medline: 23027317]
6. Ross SE, Moore LA, Earnest MA, Wittevrongel L, Lin C. Providing a web-based online medical record with electronic communication capabilities to patients with congestive heart failure: randomized trial. J Med Internet Res 2004;6(2):e12 [FREE Full text] [doi: 10.2196/jmir.6.2.e12] [Medline: 15249261]
7. Bronson DL, O'Meara K. The impact of shared medical records on smoking awareness and behavior in ambulatory care. J Gen Intern Med 1986;1(1):34-37. [doi: 10.1007/BF02596322] [Medline: 3772566]
8. McMillan B, Eastham R, Brown B, Fitton R, Dickinson D. Primary care patient records in the United Kingdom: past, present, and future research priorities. J Med Internet Res 2018;20(12):e11293 [FREE Full text] [doi: 10.2196/11293] [Medline: 30567695]
9. Louch G, Albutt A, Smyth K, O'Hara JK. What do primary care staff think about patients accessing electronic health records? A focus group study. BMC Health Serv Res 2022;22(1):581. [doi: 10.1186/s12913-022-07954-y] [Medline: 35488233]
10. Blease CR, Kharko A, Dong Z, Jones RB, Davidge G, Hagglund M, et al. Experiences and opinions of general practitioners with patient online record access: an online survey in England. BMJ Open 2024;14(1):e078158 [FREE Full text] [doi: 10.1136/bmjopen-2023-078158] [Medline: 38302414]
11. Coulter A, Richards T, Giles C, Walker S. Patients need access to their medical records—now. BMJ 2024;386:q1385. [doi: 10.1136/bmj.q1385] [Medline: 38991694]
12. Turner A, Morris R, McDonagh L, Hamilton F, Blake S, Farr M, et al. Unintended consequences of patient online access to health records: a qualitative study in UK primary care. Br J Gen Pract 2023;73(726):e67-e74 [FREE Full text] [doi: 10.3399/BJGP.2021.0720] [Medline: 36316163]
13. Clarke A, Watt I, Sheard L, Wright J, Adamson J. Implementing electronic records in NHS secondary care organizations in England: policy and progress since 1998. Br Med Bull 2017;121(1):95-106. [doi: 10.1093/bmb/ldw055] [Medline: 28043952]
14. Peckham D, Etherington C, White H, Mehta A, Shaw N, Morton A, et al. The development and deployment of integrated electronic care records in a regional adult and paediatric cystic fibrosis unit. J Cyst Fibros 2014;13(6):681-686 [FREE Full text] [doi: 10.1016/j.jcf.2014.03.002] [Medline: 24670246]
15. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
16. Quon BS, Bentham WD, Unutzer J, Chan Y, Goss CH, Aitken ML. Prevalence of symptoms of depression and anxiety in adults with cystic fibrosis based on the PHQ-9 and GAD-7 screening questionnaires. Psychosomatics 2015;56(4):345-353. [doi: 10.1016/j.psym.2014.05.017] [Medline: 25556569]
17. Quittner AL, Buu A, Messer MA, Modi AC, Watrous M. Development and validation of the cystic fibrosis questionnaire in the United States: a health-related quality-of-life measure for cystic fibrosis. Chest 2005;128(4):2347-2354. [doi: 10.1378/chest.128.4.2347] [Medline: 16236893]
18. Gee L, Abbott J, Conway SP, Etherington C, Webb AK. Development of a disease specific health related quality of life measure for adults and adolescents with cystic fibrosis. Thorax 2000;55(11):946-954 [FREE Full text] [doi: 10.1136/thorax.55.11.946] [Medline: 11050265]
19. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
20. Hibbard J, Gilburt H. Supporting people to manage their health: an introduction to patient activation. King's Fund. 2014. URL: https://assets.kingsfund.org.uk/f/256914/x/d5fbab2178/supporting_people_manage_their_health_2014.pdf [accessed 2025-09-20]
21. Maly RC, Frank JC, Marshall GN, DiMatteo MR, Reuben DB. Perceived efficacy in patient-physician interactions (PEPPI): validation of an instrument in older persons. J Am Geriatr Soc 1998;46(7):889-894. [doi: 10.1111/j.1532-5415.1998.tb02725.x] [Medline: 9670878]
22. Vodicka E, Mejilla R, Leveille SG, Ralston JD, Darer JD, Delbanco T, et al. Online access to doctors' notes: patient concerns about privacy. J Med Internet Res 2013;15(9):e208 [FREE Full text] [doi: 10.2196/jmir.2670] [Medline: 24072335]
23. Miller G. Physician and patient attitudes toward technology in medicine. Medscape 2016:1-23.

XSL•FO
RenderX

24. Tavares J, Oliveira T. Electronic health record patient portal adoption by health care consumers: an acceptance model and survey. J Med Internet Res 2016;18(3):e49 [FREE Full text] [doi: 10.2196/jmir.5069] [Medline: 26935646]
25. Nahm E, Resnick B, Mills ME. Development and pilot-testing of the perceived health Web Site usability questionnaire (PHWSUQ) for older adults. Stud Health Technol Inform 2006;122:38-43. [Medline: 17102214]
26. Ross SE, Lin C. The effects of promoting patient access to medical records: a review. J Am Med Inform Assoc 2003;10(2):129-138 [FREE Full text] [doi: 10.1197/jamia.m1147] [Medline: 12595402]
27. Robinson S, Reed M, Quevillon T, Hirvi E. Patient perceptions and interactions with their web portal-based laboratory results. BMJ Health Care Inform 2019;26(1) [FREE Full text] [doi: 10.1136/bmjhci-2019-000012] [Medline: 31039117]
28. Stanford GE, Dave K, Simmonds NJ. Pulmonary exacerbations in adults with cystic fibrosis: a grown-up issue in a changing cystic fibrosis landscape. Chest 2021;159(1):93-102 [FREE Full text] [doi: 10.1016/j.chest.2020.09.084] [Medline: 32966813]
29. Abbott J, Hurley MA, Chadwick H, Peckham D. Ways of coping and survival in cystic fibrosis: a 20-year longitudinal study. J Cyst Fibros 2023;22(1):112-118 [FREE Full text] [doi: 10.1016/j.jcf.2022.04.011] [Medline: 35461783]
30. Rexhepi H, Åhlfeldt RM, Cajander, Huvila I. Cancer patients' attitudes and experiences of online access to their electronic medical records: a qualitative study. Health Informatics J 2018;24(2):115-124 [FREE Full text] [doi: 10.1177/1460458216658778] [Medline: 27440056]
31. Mák G, Smith Fowler H, Leaver C, Hagens S, Zelmer J. The effects of web-based patient access to laboratory results in British Columbia: a patient survey on comprehension and anxiety. J Med Internet Res 2015;17(8):e191 [FREE Full text] [doi: 10.2196/jmir.4350] [Medline: 26242801]
32. Mold F, de Lusignan S, Sheikh A, Majeed A, Wyatt JC, Quinn T, et al. Patients' online access to their electronic health records and linked online services: a systematic review in primary care. Br J Gen Pract 2015;65(632):e141-e151 [FREE Full text] [doi: 10.3399/bjgp15X683941] [Medline: 25733435]
33. Alpert JM, Morris BB, Thomson MD, Matin K, Sabo RT, Brown RF. Patient access to clinical notes in oncology: a mixed method analysis of oncologists' attitudes and linguistic characteristics towards notes. Patient Educ Couns 2019;102(10):1917-1924 [FREE Full text] [doi: 10.1016/j.pec.2019.05.008] [Medline: 31109771]
34. Turner A, Morris R, McDonagh L, Hamilton F, Blake S, Farr M, et al. Unintended consequences of patient online access to health records: a qualitative study in UK primary care. Br J Gen Pract 2023;73(726):e67-e74 [FREE Full text] [doi: 10.3399/BJGP.2021.0720] [Medline: 36316163]
35. Bell SK, Delbanco T, Elmore JG, Fitzgerald PS, Fossa A, Harcourt K, et al. Frequency and types of patient-reported errors in electronic health record ambulatory care notes. JAMA Netw Open 2020;3(6):e205867 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.5867] [Medline: 32515797]
36. Johansen MA, Kummervold PE, Sørensen T, Zanaboni P. Health professionals' experience with patients accessing their electronic health records: results from an online survey. Stud Health Technol Inform 2019;264:504-508. [doi: 10.3233/SHTI190273] [Medline: 31437974]
37. Vermeir P, Degroote S, Vandijck D, Van Tiggelen H, Peleman R, Verhaeghe R, et al. The patient perspective on the effects of medical record accessibility: a systematic review. Acta Clin Belg 2017;72(3):186-194. [doi: 10.1080/17843286.2016.1275375] [Medline: 28056665]
38. Zanaboni P, Kummervold PE, Sørensen T, Johansen MA. Patient use and experience with online access to electronic health records in Norway: results from an online survey. J Med Internet Res 2020;22(2):e16144 [FREE Full text] [doi: 10.2196/16144] [Medline: 32031538]
39. Giardina TD, Menon S, Parrish DE, Sittig DF, Singh H. Patient access to medical records and healthcare outcomes: a systematic review. J Am Med Inform Assoc 2014;21(4):737-741 [FREE Full text] [doi: 10.1136/amiajnl-2013-002239] [Medline: 24154835]
40. Dendere R, Slade C, Burton-Jones A, Sullivan C, Staib A, Janda M. Patient portals facilitating engagement with inpatient electronic medical records: a systematic review. J Med Internet Res 2019;21(4):e12779 [FREE Full text] [doi: 10.2196/12779] [Medline: 30973347]
41. White H, Gillgrass L, Wood A, Peckham DG. Requirements and access needs of patients with chronic disease to their hospital electronic health record: results of a cross-sectional questionnaire survey. BMJ Open 2016;6(10):e012257 [FREE Full text] [doi: 10.1136/bmjopen-2016-012257] [Medline: 27742623]
42. UK cystic fibrosis registry 2023 annual data report. Cystic Fibrosis Trust. 2024. URL: https://www.cysticfibrosis.org.uk/sites/default/files/2024-11/CFT_2023_Annual_Data_Report_Oct2024%201.pdf [accessed 2024-09-20]
43. Fisher R, Fraser C. Who gets in? What does the 2020 GP patient survey tell us about access to general practice? The Health Foundation. 2022. URL: https://www.health.org.uk/news-and-comment/charts-and-infographics/who-gets-in [accessed 2025-06-01]
44. Wilson J, Heinsch M, Betts D, Booth D, Kay-Lambkin F. Barriers and facilitators to the use of e-health by older adults: a scoping review. BMC Public Health 2021;21(1):1556 [FREE Full text] [doi: 10.1186/s12889-021-11623-w] [Medline: 34399716]
45. Zoorob D, Hasbini Y, Chen K, Wangia-Anderson V, Moussa H, Miller B, et al. Ageism in healthcare technology: the older patients' aspirations for improved online accessibility. JAMIA Open 2022;5(3):ooac061 [FREE Full text] [doi: 10.1093/jamiaopen/ooac061] [Medline: 35855421]

XSL•FO
RenderX

Abbreviations

CF: cystic fibrosis
CFQ-R: Cystic Fibrosis Questionnaire-Revised
CONSORT: Consolidated Standards of Reporting Trials
EHR: electronic health care record
GAD-7: Generalized Anxiety Disorder-7
PAM-13: Patient Activation Measure-13
PEPPI: patient and provider perceived efficacy in patient-physician interaction
PHQ-9: Patient Health Questionnaire-9
PHWSUQ: Perceived Health Web Site Usability Questionnaire
QoL: quality of life

Edited by A Mavragani; submitted 06.12.24; peer-reviewed by S Georgopoulou, Y Zhao; comments to author 07.04.25; revised version received 03.06.25; accepted 05.08.25; published 25.12.25.

Please cite as:
Chadwick HK, Sawant A, White H, Gillgrass L, Spoletini G, Clifton IJ, Etherington C, Peckham DG
Providing Mobile Patient Access to Their Electronic Secondary Care Patient Record in Adults With Cystic Fibrosis: Results of a Prospective, Parallel, Randomized Open-Pilot Quantitative Study
JMIR Form Res 2025;9:e69747
URL: https://formative.jmir.org/2025/1/e69747
doi:10.2196/69747
PMID:

©Helen K Chadwick, Akhil Sawant, Helen White, Lindsey Gillgrass, Giulia Spoletini, Ian J Clifton, Christine Etherington, Daniel G Peckham. Originally published in JMIR Formative Research (https://formative.jmir.org), 25.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Mobile App (MyPeer) Co-Designed With Immigrant Adolescents for Better Sexual and Reproductive Health: Usability Study

Salima Meherali[1*], PhD; Amyna Ismail Rehmani[1*], MN; Mariam Ahmad[1*], MPH; Piper Scott-Fiddler[2], MS; Paula Pinzón-Hernández[2], PhD; Zeba Khan[2], MSc; Sarah Flicker[3], PhD; Philomina Okeke-Ihejirika[4], PhD; Bukola Salami[5], PhD; Eleni Stroulia[6], PhD; Ashley Vandermorris[7], PhD; Josephine Pui-Hing Wong[8], PhD; Wendy V Norman[9], MD, MHSc.; Shannon D Scott[1], PhD; Sarah Munro[10], PhD

[1]Faculty of Nursing, College of Health Sciences, University of Alberta, Edmonton, AB, Canada
[2]Department of Obstetrics and Gynecology, University of British Columbia, Vancouver, BC, Canada
[3]Faculty of Environmental and Urban Change, York University, Toronto, ON, Canada
[4]Department of Women's and Gender Studies, Faculty of Arts, University of Alberta, Edmonton, AB, Canada
[5]Department of Community Health Science, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada
[6]Faculty of Science, University of Alberta, Edmonton, AB, Canada
[7]Department of Paediatrics, University of Toronto, Toronto, ON, Canada
[8]Daphne Cockwell School of Nursing, Toronto Metropolitan University, Toronto, ON, Canada
[9]Department of Family Practice, Faculty of Medicine, University of British Columbia, Vancouver, BC, Canada
[10]Department of Health Systems and Population Health, School of Public Health, University of Washington, Seattle, WA, United States
[*]these authors contributed equally

Corresponding Author:
Salima Meherali, PhD
Faculty of Nursing
College of Health Sciences
University of Alberta
11405 87 Ave NW
Edmonton, AB, T6G 2V2
Canada
Phone: 1 (780) 248 1401
Email: meherali@ualberta.ca

Abstract

Background: Adolescents require comprehensive sexual and reproductive health (SRH) education to successfully transition from puberty into adulthood. However, they often experience barriers and challenges while trying to promote their SRH or access SRH services. Such challenges are amplified among youth from migrant backgrounds, who may further be constrained by societal stigmas and cultural taboos regarding SRH. Mobile health interventions have the potential to provide culturally relevant, accessible, and evidence-based SRH educational resources; however, few SRH mobile apps in Canada are co-designed with immigrant youth or meaningfully integrate their voices and lived experiences.

Objective: We aimed to co-design a culturally relevant and evidence-based mobile app with immigrant adolescents to provide accurate SRH resources. In this paper, we present the findings of the usability testing of our SRH mobile app—MyPeer.

Methods: Throughout our study, we used a community-based participatory research approach and implemented the principles of human-centered design to co-design our mobile app. For our usability study, we recruited immigrant adolescents and interest holders working with the target population. Adolescents participated in moderated focus group discussions (FGDs) and interest holders evaluated the app's quality using the standardized Mobile App Rating Scale (rating components on a scale of 1-5). All FGDs were audio-recorded and later analyzed to implement changes in the app. Mobile App Rating Scale (MARS) scores and responses were analyzed descriptively to evaluate the app's engagement, functionality, aesthetics, quality of information, and subjective app quality.

Results: Overall, 25 adolescents and 17 interest holders participated in this usability study. We analyzed the findings from the FGDs and categorized them into four categories: (1) navigation and interface, (2) SRH information quality and learning, (3)

XSL•FO
RenderX

technical performance, and (4) accessibility and multimedia usability. Adolescents found the app visually appealing and the interface easy to navigate. They appreciated interactive features, such as quizzes, and the presentation of information through various media (eg, videos and infographics). However, they also identified technical issues, such as map glitches and navigation inconsistencies, and requested deeper content on SRH topics. The data from the MARS checklist completed by interest holders were analyzed descriptively. The app received the highest scores in the domains of functionality, with mean scores of 4.3 (performance and navigation); engagement, with mean scores of 3.7 (interest); and aesthetics, with mean scores of 4.1 (graphics) and 3.9 (visual appeal). The lowest rated items were customization, with a mean score of 2.5, and interactivity, with a mean score of 3.1.

Conclusions: Our app—MyPeer—has promising usability and appeal for adolescents looking for SRH information. Incorporating feedback from youth and content experts helped identify both technical refinements and content requirements. Our findings support the app's potential as a scalable, youth-centered SRH digital tool and emphasize the value of participatory design in youth digital interventions.

(JMIR Form Res 2025;9:e81115) doi:10.2196/81115

KEYWORDS

adolescent health; mHealth; sexual and reproductive health; co-design; usability testing

Introduction

Adolescence (ages 10-19 years) is a critical period of development during which individuals develop their identity and lay solid foundations for subsequent phases of life [1]. These foundations include a healthy and comprehensive understanding of their sexual and reproductive health (SRH), including puberty, sexual maturity, relationships, and much more [1]. However, research in the field of adolescent SRH suggests that adolescents do not always have the opportunity or access to the best possible SRH education during this crucial phase of their life [2,3]. In many situations, such access is even more restricted for adolescents from migrant backgrounds due to the stigma surrounding SRH, parental unwillingness to discuss SRH, and inability to seek help due to fear and shame [4,5]. A combination of these factors can result in a lack of medically accurate information and increase the risks of unhealthy behaviors.

With the onset of digital technology, health education has become a promising channel to fill information gaps and potentially change behavior [6,7]. According to Statistics Canada [8], 99.2% of people aged 15-24 years use the internet, and 96% own a smartphone, making digital platforms highly impactful for health education. Mobile health (mHealth) apps provide private, youth-friendly platforms that are accessible on demand, offering trusted information and linking youth to services [9,10]. This is particularly beneficial for designing SRH digital interventions as innovative solutions to address sensitive SRH issues [11]. Research globally suggests that such interventions have been successful in reaching, engaging, and educating adolescents and young adults; in some cases, the rate of preventive testing, delayed sexual activity, and SRH knowledge has been improved [11-13]. Several apps have been developed to promote adolescent SRH; however, there is limited information on their quality and accuracy, and few are specifically designed to meet the needs of adolescents from migrant backgrounds in Canada.

To better understand the range of mobile apps available for adolescent SRH in North America and to support the potential development of a new app tailored for adolescents from migrant backgrounds in Canada, we conducted an environmental scan in 2022 [14]. This scan assessed the quality and usability of these apps and identified 15 relevant apps. Our findings revealed that very few apps offered comprehensive, reliable, and evidence-based information, and most lacked features for customization [14]. Additionally, insights from this scan, combined with extensive research involving adolescents that highlight gaps in cultural adaptation and user engagement strategies, underscore the critical need for evidence-informed, youth-centered design in SRH app development [15]. Therefore, we aimed to co-design an accessible and culturally sensitive knowledge translation resource (ie, mobile app) to provide immigrant adolescents with evidence-based resources to access SRH information. Our paper aims to present the findings of a comprehensive usability study of our mobile app, incorporating feedback from adolescents and SRH experts.

Methods

Study Design

We conducted a multisite study (2023-2024) across 3 Canadian provinces (ie, Alberta, British Columbia, and Ontario), with large immigrant populations. We undertook this study with the purpose of engaging immigrant adolescents in developing, implementing, and evaluating an mHealth app that will deliver evidence-based SRH information to this population. We used a systematic approach to develop a resource that reflects the perceptions and experiences of the target populations as well as to meet their unique needs. Using the principles of human-centered design and community-based participatory research [16], we conducted this study in 4 stages (Figure 1). The protocol for the study and results of the individual stages are published elsewhere [3,15]. This study reports on the findings from stage 4—usability testing.

XSL•FO
RenderX

Figure 1. Stages of the co-design project. AAG: adolescent advisory group.

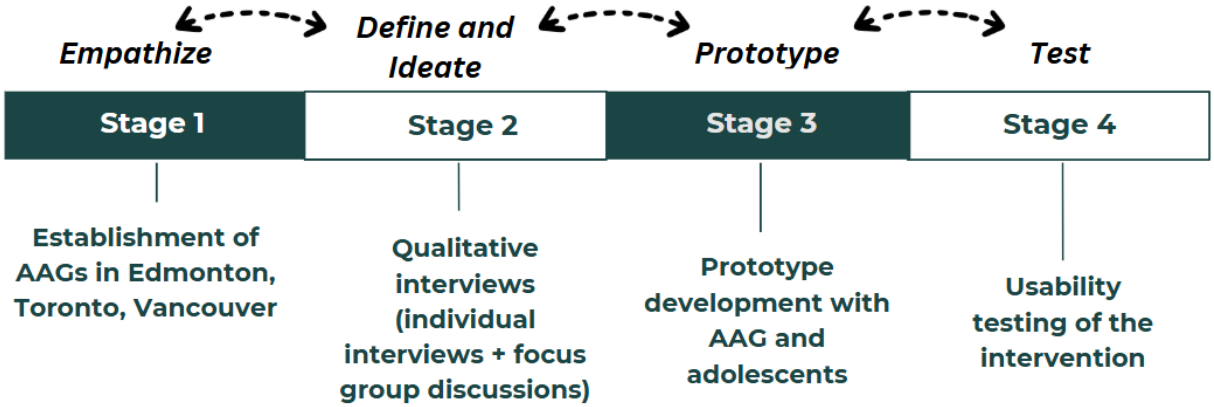

Project Stages

Stage 1: Empathize

To establish adolescent advisory groups (AAGs), we recruited 26 adolescents (aged 12-19 years) as equal partners in research activities, such as data collection, data analysis, resource development, concurrent feedback, and knowledge mobilization.

Stage 2: Define and Ideate

We conducted individual qualitative interviews with immigrant adolescents (n=58) to explore their SRH information needs, access to services, and barriers to promoting their SRH.

Stage 3: Prototype

After assessing the SRH information needs of immigrant adolescents and gaps in their knowledge, we developed a prototype for a mobile app specifically catered to their needs.

Stage 4: Usability Testing

The multidisciplinary team involved in prototype development included the primary researchers (principal investigator, research coordinator, and youth research assistants), software and app developers from the Department of Computing Science at the University of Alberta, AAG members, and interest holders as content experts. At this stage, we aimed to evaluate whether the designed intervention (prototype) was usable by its targeted end users and to identify errors and issues to make it more user-friendly [10]. This paper reports on the process and outcomes of the usability testing of the MyPeer app prototype.

Ethical Considerations

We received ethics approval for this project from the University of Alberta Research Ethics Board (August 26, 2021; Pro00113664), the Toronto Metropolitan University Research Ethics Board (April 27, 2022; REB 2022-069), and the University of British Columbia (May 31, 2022; #H22-00395). All participants provided informed consent to participate in the study and were given a gift card (CAD $20 [US $14.5] value) for their time and contribution. Participants were also encouraged to sign a confidentiality agreement to protect the identities of other participants and the experiences shared within discussions.

Prototype: MyPeer

Our app prototype, MyPeer, is designed to provide adolescents with evidence-based, culturally relevant resources for SRH. This app was co-designed with adolescents from migrant backgrounds, incorporating their unique needs and cultural perspectives. The app provides SRH educational content on 6 SRH priorities (ie, puberty, menstruation, sexually transmitted infections, sexual assault, contraception, and healthy relationships), which were identified and informed through qualitative interviews (stage 2) and AAG design thinking sessions (stage 1). The content was developed by systematically reviewing current literature, evidence-based SRH frameworks (World Health Organization), and national health agency guidelines (Public Health Agency of Canada). We collaborated with SRH experts (eg, clinicians, sexual health consultants, and public health nurses) to validate the content and to ensure its accuracy, and we recommended enhancements to promote its youth friendliness. These experts worked with racialized and migrant populations on issues pertaining to SRH and had practical knowledge about their current challenges. We offered the content through a variety of multimedia formats, such as videos, infographics, text-based information, and animations. The content was developed with an aim to address common cultural stigmas, misconceptions, and knowledge gaps, while promoting positive SRH attitudes and conversations within families and communities. In addition to providing comprehensive SRH content, the app linked adolescents with nearby SRH resources and services (eg, sexual assault centers and birth control clinics), educated them on the latest developments in the field of SRH (via blogs), and promoted engagement through quizzes. Our AAG members contributed to the app's design and layout, offering suggestions to improve its features and overall appeal. The prototype was made available for testing on both iOS and Android devices.

Participants and Recruitment

For the usability testing of our prototype, we deemed it beneficial not only to test the app with its intended target audience (ie, adolescents) but also with relevant interest holders (eg, youth advocacy organizations, SRH experts, and health care professionals) working with immigrant adolescents. This approach provided us with a more nuanced understanding of the app's functionality, acceptability, and relevance within the

XSL•FO
RenderX

broader SRH ecosystem. Moreover, involving interest holders in the testing gave us an overview of how digital interventions may be perceived among youth, considering their consistent involvement with this demographic.

We recruited participants from Edmonton, Toronto, and Vancouver between April and December 2024. Our AAG members assisted with recruiting adolescent participants for the app usability testing. We used a multifaceted approach to promote recruitment activities, which included sharing posters via social media, distributing recruitment materials within community networks, email listservs, and youth organizations, and following up with adolescents who participated in our qualitative interviews (phase 2). We used our study's criteria for inclusion of participants who were either first-generation (born outside of Canada and immigrated to Canada) or second-generation (born in Canada with at least 1 foreign-born parent) immigrants. Additionally, our criteria required that adolescents should be residing in Canada during the time of the study and have access to a mobile device with internet (Android or iOS) to use the mobile app and participate in research activities. We ensured that we had an equal number of participants testing the app on both iOS and Android platforms. We provided all potential participants with an overview of the study, its objectives, the tasks associated with testing, and the time commitment required for each phase of testing. Participants were required to review and sign a consent form and complete a demographic form, which was offered securely through Google Forms.

To recruit interest holders, we contacted relevant interest holders through our network who work with adolescents and have expertise in youth SRH, such as through reproductive health clinics and youth advocacy organizations. We included a diverse range of interest holders, including frontline professionals working with youth (eg, physicians, nurses, and educators) and experts involved in youth-focused initiatives and advocacy efforts (eg, Action Canada).

Testing Procedures and Measures

We used a multimethod approach to evaluate the usability of the mobile app prototype, incorporating exploratory testing through open-ended focus group discussions (FGDs) and quantitative assessment testing surveys.

Testing Measures With Adolescents

We invited adolescents to participate in FGDs (5-6 members each) via Zoom, which lasted for 60-90 minutes. The FGDs moderated by the research team members (SM, MA, and AIR) included open-ended discussions to capture user perceptions, challenges, app performance, and recommendations (see Multimedia Appendix 1). We provided participants with case-based scenarios that involved using various features of the app. They shared feedback on their experience and engaged in discussions about the most effective features and areas for improvement. All the FDGs were audio-recorded for later analysis. After the FGDs, we invited participants to complete a survey rating the app's attractiveness, user-friendliness, navigation, and usability on a Likert scale, which was offered via Google Forms. It also collected descriptive responses from the participants on their thoughts about the app interface, design, graphics, and suggestions for improvement.

Testing Measures With Interest Holders

With the diverse expertise of our interest holder group, we leveraged their insights to thoroughly evaluate the app's quality using the Mobile App Rating Scale (MARS) checklist [17]. MARS is a validated tool that provides a multidimensional, reliable, and flexible app quality assessment criterion, which researchers and academics can effectively use to gather data on the app's engagement, functionality, aesthetics, quality of information, and subjective app quality [17,18]. We contacted individual interest holders and gave them an overview of the app and the MARS survey. We asked them to review the app content and its features in detail and then complete the survey on their own time. Any concerns or questions regarding the sections of the survey were addressed by the team.

Data Analysis

Quantitative Analysis

Quantitative data from the MARS checklist were analyzed descriptively, and the mean scores were calculated for each of the 5 MARS domains: engagement, functionality, aesthetics, information quality, and subjective quality. Higher mean scores indicated better app quality and perceived usability. Moreover, the demographic characteristics of adolescents were analyzed through descriptive statistics.

Qualitative Analysis

We analyzed the qualitative data from the FGDs iteratively, which helped us to modify our questions and expand our understanding of each subsequent FGD. The research team independently reviewed audio transcripts to familiarize themselves with the data and to identify recurrent ideas related to content relevance, app features, navigation, and usability through inductive coding. Categories were iteratively refined and organized into overarching themes. Any discrepancies in the final categorization process were resolved through team discussion to ensure analytic consistency.

Results

Participant Characteristics

We recruited 25 adolescents and 17 interest holders for the usability testing. We ensured that our sample included individuals from diverse backgrounds to gather a variety of perceptions and experiences that would help inform the continuous improvement of the app. We collected sociodemographic information from adolescents only, which is presented in Table 1. Of the 25 adolescents, 23 completed the post-FGD survey evaluating the app's features. The group of interest holders included 5 public health nurses and clinicians, 4 health promotion facilitators (eg, Birth Control Centre), 3 SRH educators, and 5 individuals from youth advocacy organizations (eg, Action Canada and Canadian Advisory of Women Immigrants).

XSL•FO
RenderX

Table 1. Demographic characteristics of adolescent participants (N=25).

Characteristics	Values, n (%)
Sex	
Female	19 (76)
Male	6 (24)
Age (years)	
14-15	1 (4)
16-17	3 (12)
18-20	21 (84)
Place of birth	
Northern America	13 (52)
Sub-Saharan Africa	2 (8)
Southern Asia	7 (28)
Western Asia	2 (8)
Eastern Europe	1 (4)
Ethnicity	
Asian (South Asian, Southeast Asian, East Asian, and Persian)	17 (68)
Black	6 (24)
White	2 (8)
Current level of education	
Grade 10	1 (4)
Grade 12	4 (16)
Postsecondary school	20 (80)
Do you access SRH[a] services?	
Yes	11 (44)
No	14 (56)

[a]SRH: sexual and reproductive health.

Key Findings

Adolescent Perspectives

Adolescents provided detailed feedback on the app's design, functionality, and user experience. The findings from each FGD were analyzed iteratively to inform the questioning for the next discussion and make necessary changes in the app (eg, technical glitches). We present below the findings from adolescents' feedback in 4 categories.

Navigation and Interface

At this stage, it was insightful for us to learn how adolescents perceived the initial design of the app and adjusted to ensure user-friendliness. Many participants appreciated the app's design and commented on its simplicity and color theme (Figure 2): "The fonts, colors and graphics are probably my favorite thing about the app" [Participant 10].

XSL•FO
RenderX

Figure 2. Comparison of the current (A) and previous (B) interfaces of the MyPeer app, highlighting simplified navigation and enhanced visual appeal.

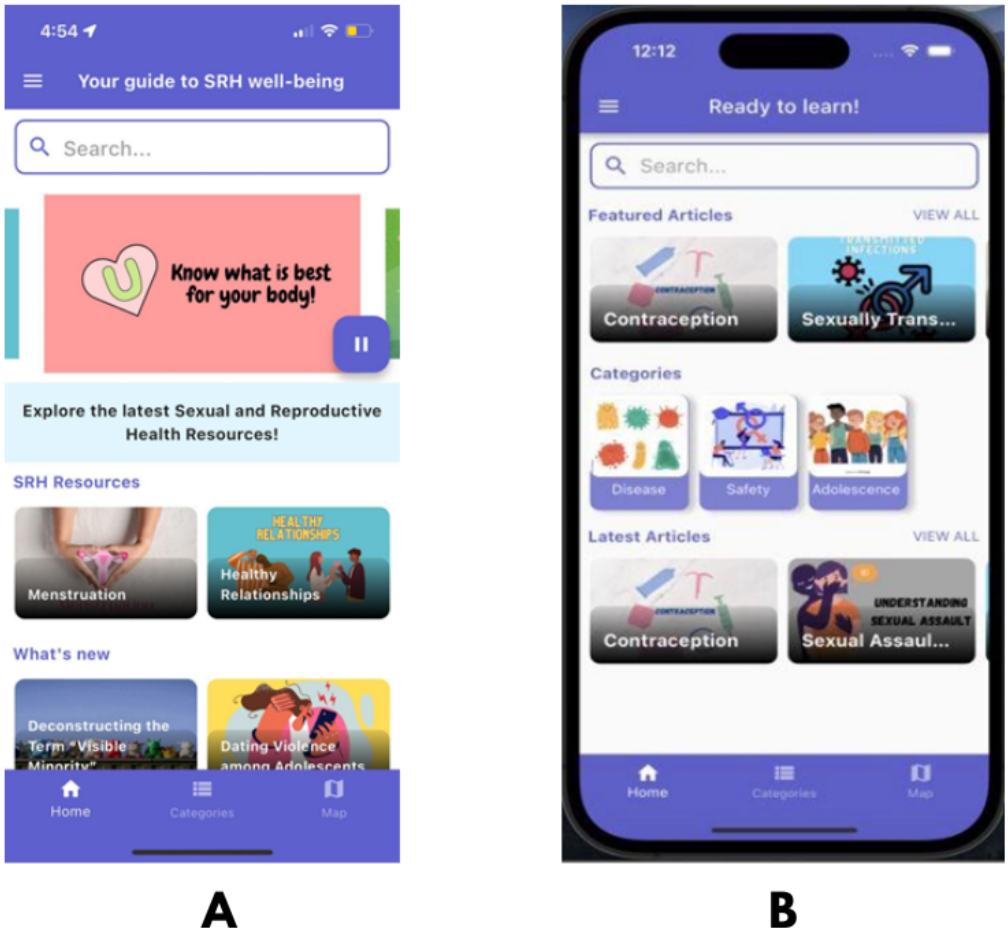

They liked the display and arrangement of information:

> *The topics were well arranged, especially the horizontal icons, hence easy to identify. The different bright colors were eye-catching, and also the animations were relatable.* [Participant 2]

However, several others reported difficulties navigating the app as they found the layout unclear and the content unorganized. Participants emphasized that the home page appeared cluttered and overwhelming, making it difficult to understand how to access desired content. One participant commented that they found the layout confusing:

> *Having the "categories" button at the bottom confused me into thinking there were more options to choose from, but we had already been shown what the app has to offer on the first page.* [Participant 13]

Adolescents expressed that it took a lot of exploring to understand the format of the content.

Participants also highlighted how the inconsistent use of fonts, image sizes, and information presentation hindered their ability to intuitively explore the app:

> *The categories aren't set up well...the organization can be a lot better. Some tabs look different even though they serve the same purpose.* [Participant 8]

Another adolescent expressed similar thoughts: "I couldn't tell where one section ended and another began...the dropdowns made it worse" [Participant 14].

Along with their perceptions, adolescents provided feedback on how certain features could be improved that would make the app more appealing. Participants suggested repositioning the navigation elements (eg, sidebar and top panel) and making the sliding images on the home page clickable: "The sidebar order felt random...maybe sort it based on what's most useful first" [Participant 3]. Another participant expressed:

> *It would've been simpler if I liked the sliding images, but assumed they could be clicked on. Not knowing what I could or couldn't click was a little difficult to rummage through as some options were clickable like the categories and quizzes, but the top sliding images were not.* [Participant 15]

Based on their feedback, we reorganized the navigation elements to present the content more clearly and added the feature to make the sliding images clickable. We also included content

XSL•FO
RenderX

within each picture to provide a better understanding of their SRH rights and how they can take care of their bodies.

SRH Information Quality and Learning

We inquired about the perceptions of adolescents regarding the SRH information and content presented through different categories and whether the information felt youth-friendly and easy to understand. We emphasized that the selection of content was informed by findings from adolescent interviews and advisory group discussions, and encouraged participants to suggest any additional topics they felt should be included.

Many participants appreciated the SRH topics covered and indicated that these topics were essential for SRH education during adolescence. Adolescents indicated that the information presented throughout the app, including FAQs, quizzes, articles, and videos, was accurate and appropriate. They indicated that the language used was simple and avoided unnecessary jargon, but could be simpler and more engaging. One participant mentioned:

> *I really like the information that the app provides and the use of color in every section. I would like to see more consistency with the fonts and also making it more reader-friendly.* [Participant 11]

Some participants inquired about including more expansive explorations of SRH concepts with embedded definitions in quizzes or links to further learning resources. Participants expressed a desire for more in-depth information on specific topics, such as contraception, and recommended including content that addresses the cultural dimensions of these subjects. One participant mentioned:

> *Some of the content was too intensive and some was lacking information. This mainly applied to the contraceptive section where it goes through every possible form of contraceptive, but didn't really go in depth on how to use them. If I were a new adolescent and reading through it I would be confused by the description of how to use it and how to acquire such things.* [Participant 20]

Another participant expressed:

> *I wish there were more diverse stories or even real experiences shared to reflect on. I was glad that there were lots of information available for things like menstruation and contraceptives—which are important! But I was slightly disappointed to see less information about sexual assault or dating violence.* [Participant 9]

Based on this feedback, we expanded information on topics such as contraception and provided links to additional resources at the bottom of each topic. We also included blogs by our team on topics such as dating violence to raise awareness on these issues and support informed decision-making among users.

Some participants also expressed a desire to see features, such as chatbots or helper chats and discussion forums, embedded within the app, where users can ask questions and get accurate answers or participate in discussions among their peers: “I would like to see like a topic forum where I can search and get like an expert opinion” [Participant 13]. While the idea was appreciated, concerns were raised about the potential harm, such as negative conversations (eg, bullying) and the spread of misinformation through these forums.

Participants shared their feedback regarding the app’s quiz feature and whether this would be a good motivator for app users to continue to improve their knowledge. Many adolescents found the quizzes interactive and engaging and expressed that they challenged them to learn more: “I liked the quizzes because it sort of challenged me to try to retain as much information as possible and made it a more interactive app” [Participant 18].

Participants also reflected on how the quiz section can be improved to ensure it is impactful. They suggested that definitions for difficult terms or acronyms can be presented as a pop-up window or at the bottom of the content.

One adolescent mentioned:

> *I would want the quiz section to improve because when you get the right answer, it's very short. I have done quizzes in different apps before, and the long explanatory answer was what I liked the most about the other apps.* [Participant 17]

Technical Performance

Throughout our discussions, we found that participants experienced recurrent technical difficulties, particularly those related to the map feature and quiz navigation. Some adolescents commented that their map feature did not automatically detect their location, and they were unable to locate services nearby: “I liked the map feature even though it wasn't working perfectly for me” [Participant 22]. Some adolescents also suggested including a pop-up explaining the purpose of the map and a brief overview of the content within it.

As adolescents tested various features of the app, we were able to identify several bugs and performance issues that allowed us to highlight areas requiring further improvements (Figure 3). For example, one participant noted that after completing a quiz and returning to the home page, they observed more categories than were originally displayed. This feedback enabled us to identify an underlying coding issue, which was subsequently addressed by our app developer. Similarly, other participants pointed out glitches with the app interface, such as the inability to close the keyboard after performing a search or rapid transitions between quiz questions that left little time to process: “The quizzes still skip questions, which prevents me from answering all the questions” [Participant 24]. Another participant mentioned, “My keyboard popped up randomly when I opened the map or search, and then I had to tap like five times to make it go away” [Participant 16].

XSL•FO
RenderX

Figure 3. Reported technical glitches during the usability testing of the MyPeer app, including slow loading times and inconsistent navigation.

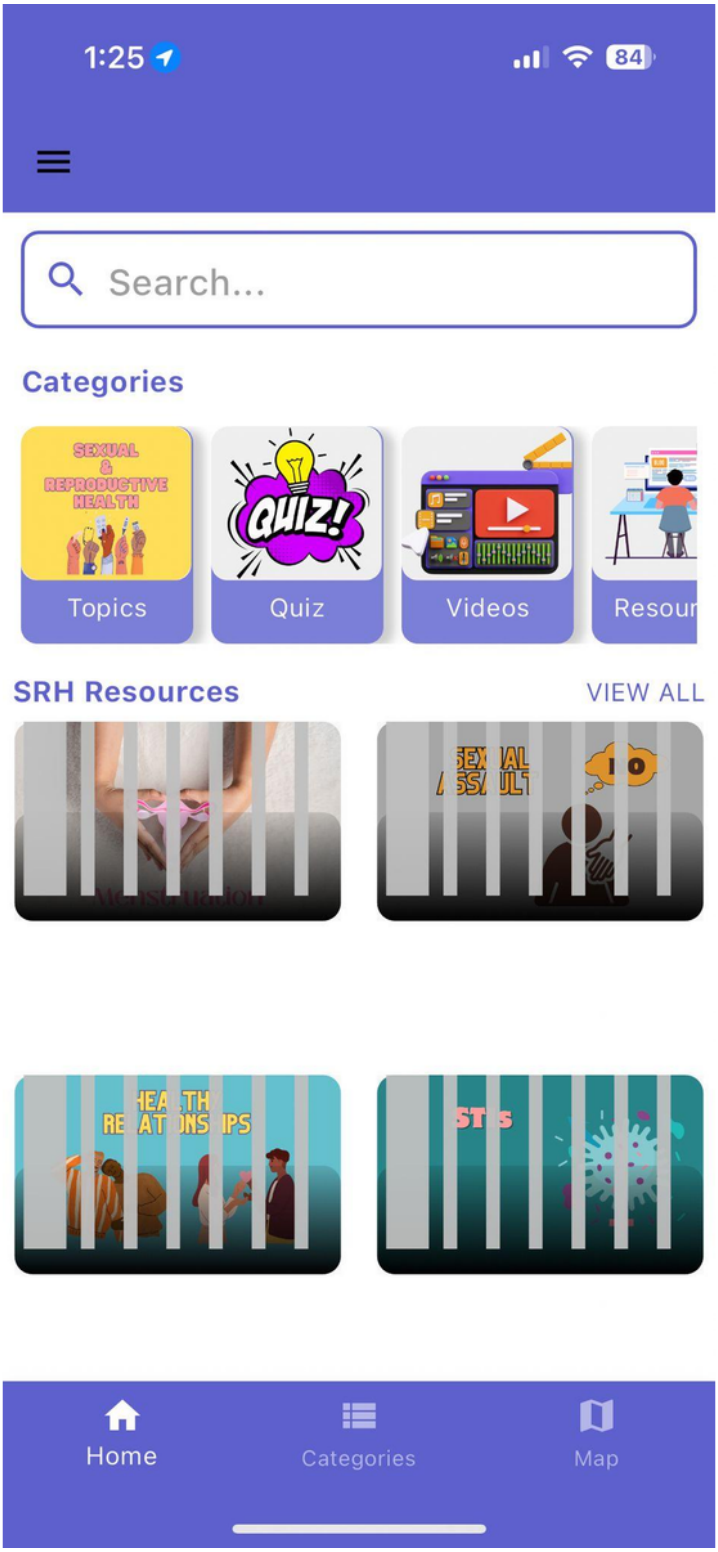

We tried to address the challenges regarding the technical performance of the app as soon as we received feedback. This helped us to rectify those errors and test whether the issues were resolved in the subsequent FGD. As we progressed to a new FGD, we also identified some new technical issues that hadn't been present initially and might have been generated as a result of updates made to the firmware. For example, one participant mentioned that the videos embedded within each category were auto-playing as they opened that category. With the expertise of our app developer, we were able to fix all bugs and restore the app to its optimal performance, which were then not reported in the subsequent FGD.

Accessibility and Multimedia Usability

We inquired about adolescents' views on the information presented through various means (eg, infographics, videos, and pictures) by using case-based scenarios during FGDs. By participating in those activities, adolescents could indicate which

XSL•FO
RenderX

resources proved effective and which could be improved. Adolescents emphasized the need for improved accessibility features, such as text-to-speech options, transcripts for videos, and more intuitive media control (eg, full-screen buttons). As one participant mentioned:

> *I liked the videos, as it makes the content more accessible for those who are more adept at listening. It would be helpful to have a transcript or subtitle to make this information accessible. For the infographics, I would have appreciated being able to click and enlarge the photos.* [Participant 5]

Many participants commented on the design and limited Zoom functionality of the infographics, citing difficulties reading the font. Aligning with other feedback from participants regarding the consistency of fonts across different categories, we made major improvements in the design of the infographics to ensure their readability and consistency. We also made improvements to the color scheme of the infographics to make it more appealing.

Participants had contrasting views on the way the information within categories was presented. Some adolescents liked how information was categorized and presented under drop-down menus, which made it look less cluttered. However, others did not find manually minimizing the dropdowns appealing. One participant indicated alternatives for drop-down style content:

> *The drop-down menu for the infographics page makes it hard to access and read the information. It would be helpful if the infographic went to a different page.* [Participant 17]

While we appreciated the feedback, this feature was not supported by our current app coding framework and was not deemed possible to implement at that stage.

Along with the content within categories, we also inquired about the accessibility of other features within the app. Participants commented that they appreciated the maps feature as it distinguished the app from other information that is available online. They also liked including blogs within the content as they can provide more recent and up-to-date information on SRH topics. Initially, the rotating images banner on our home screen did not have a pause feature. After receiving the following feedback, we integrated a pause and play button for the rotating banner (Figure 4):

XSL•FO
RenderX

Figure 4. Landing page of the MyPeer app providing adolescents access to educational content, blogs, and navigation to additional resources and services.

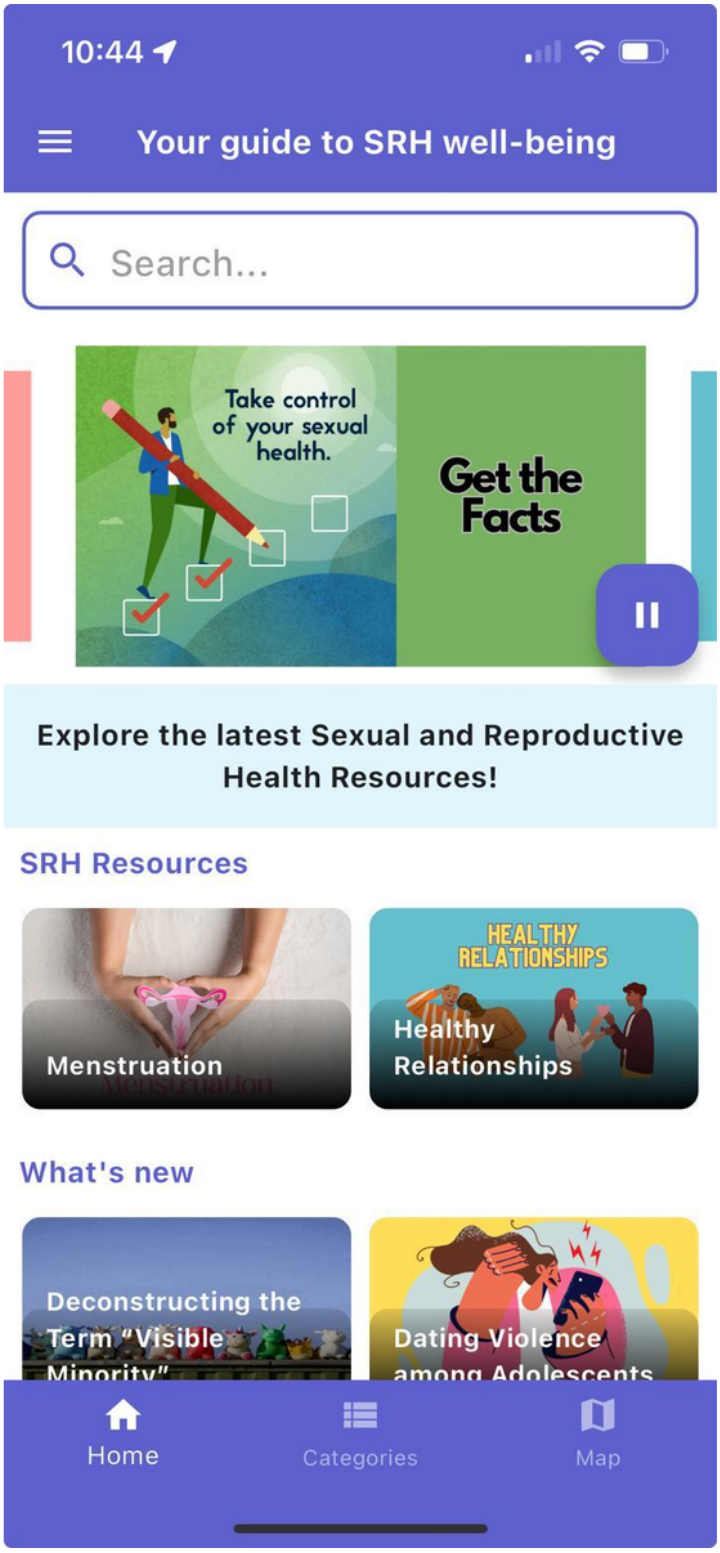

> *After a while, it was a bit disorienting to watch the banners keep rotating. It would be helpful to have the ability to pause or click on a banner.* [Participant 9]

A couple of participants mentioned that it would be beneficial to include a tour-style guide for users when they first download the app to avoid confusion for new users. We implemented this feedback and included a walk-through of different features of the app, which was well received by consecutive FGDs (Figure 5). Some participants also pointed out that not all adolescents would like to register and provide their information; such individuals, in their view, should be able to access the content without registering. We considered this feedback and added the option to access content without registering.

XSL•FO
RenderX

Figure 5. Guided tour of the MyPeer mobile app outlining key features and disclaimers (privacy and security).

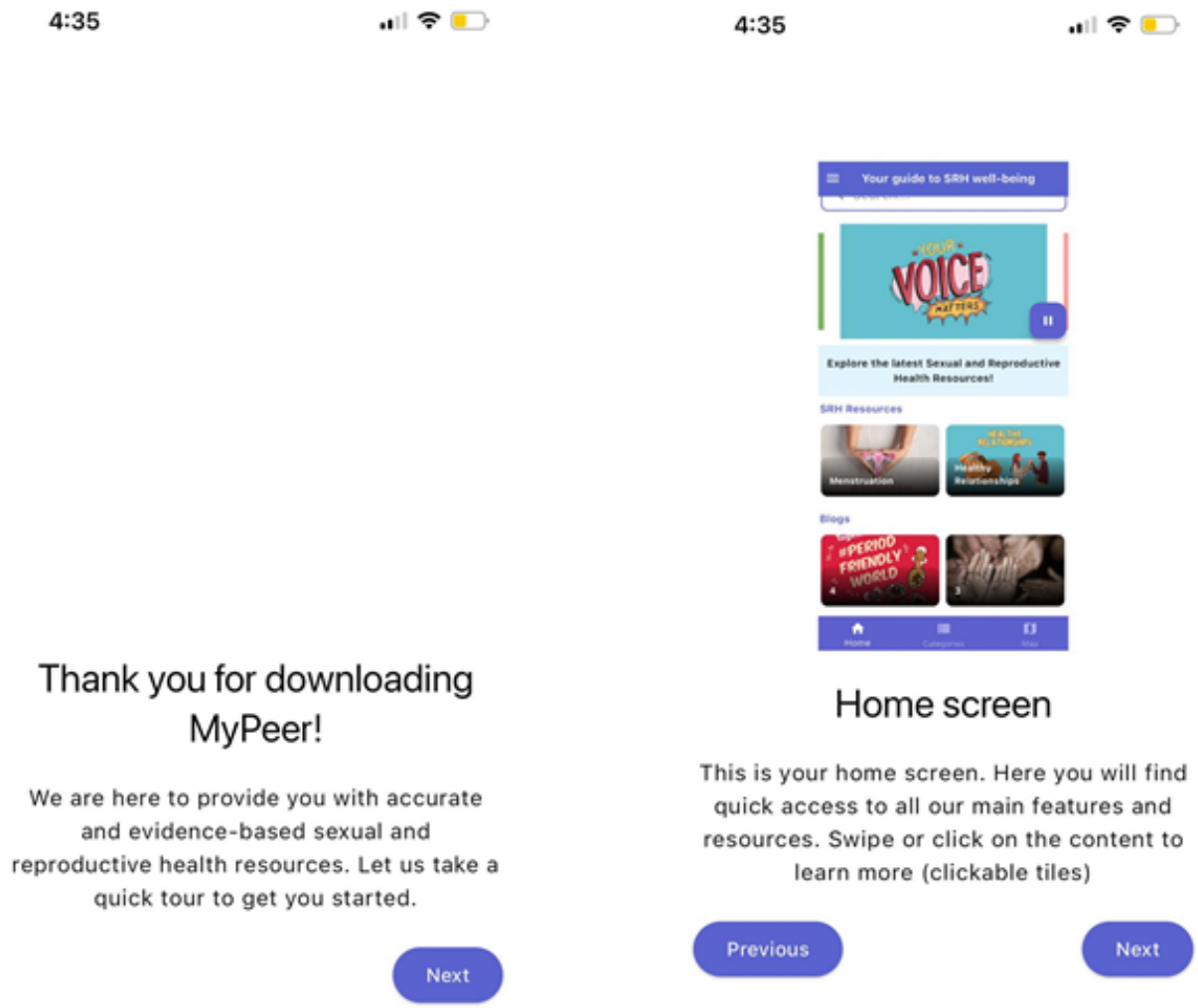

Stakeholder Assessment Using the MARS Checklist

A total of 17 interest holders, including sexual health educators, researchers, academics, and health professionals, completed the MARS checklist. We conducted this stage of this study after making most of the changes identified through adolescents' feedback. According to the checklist, their feedback was assessed across four domains: (1) engagement, (2) functionality, (3) aesthetics, and (4) information quality.

Engagement

Most interest holders indicated that the app was moderately entertaining, with 29% (5/17) indicating that it was highly entertaining (Table 2). The interactive quizzes, videos, and articles were appreciated for their potential to maintain user engagement and promote learning. Among the participants, 59% (10/17) of interest holders agreed that the app was moderately interesting and would keep users engaged for some time. Some interest holders indicated that the lack of more gamification strategies might limit long-term retention.

XSL•FO
RenderX

Table 2. Mean scores for the engagement domain of the Mobile App Rating Scale checklist.

Item	Mean score
Entertainment	
Is the app fun and entertaining to use?	3.6
Interest	
Is the app interesting to use?	3.7
Customization	
Does it provide or retain necessary settings and preferences?	2.5
Interactivity	
Does it allow user input and feedback?	3.1
Target group	
Is the content appropriate for adolescents?	3.9

Functionality

This domain received high ratings, with 53% (9/17) of participants reporting that the app's performance was perfect, while 29% (5/17) reported minor problems, and 18% (3/17) of interest holders reported some technical problems. These issues were primarily due to known bugs and technical glitches, which were identified during the adolescent FGDs as well. Despite these issues, the app was considered functional in its core delivery and was ranked highly for its ease of navigation (Table 3).

Table 3. Mean scores for the functionality domain of the Mobile App Rating Scale checklist.

Item	Mean score
Performance	
How accurately or fast do the app features work?	4.3
Ease of use	
How easy is it to learn to use the app?	4.2
Navigation	
Is moving between screens logical or accurate or uninterrupted?	4.3
Gestural design	
Are interactions consistent and intuitive?	4.2

Aesthetics

The majority of interest holders rated the app's graphics and aesthetics as high quality, with 41% (7/17) describing the visual appeal as "very attractive" and an additional 24% (4/17) indicating a high level of appeal. Interest holders noted some inconsistencies in the graphics and visual design, and a few also reported issues with icon and content sizing on smaller screens (Table 4).

Table 4. Mean scores for the aesthetics domain of the Mobile App Rating Scale checklist.

Item	Mean score
Layout	
Are the arrangement and size of icons and content appropriate?	3.8
Graphics	
How high is the quality of graphics?	4.1
Visual appeal	
How good does the app look?	3.9

Information

The majority of the interest holders found the SRH content to be credible, evidence-based, and appropriate for the target audience, with 59% (10/17) of interest holders indicating that the quality of information was highly relevant, coherent, and correct. Moreover, 47% (8/17) of interest holders reported that the information covered was comprehensive and concise and included links to further information and resources (Table 5). However, they suggested increasing comprehensiveness by

XSL•FO
RenderX

providing more elaborate explanations, especially in areas where adolescents may seek additional clarification.

Table 5. Mean scores for the information domain of the Mobile App Rating Scale checklist.

Item	Mean score
Accuracy of app description	
Does the app contain what is described?	4.3
Quality of information	
Is the app content correct, well-written, and relevant?	4.4
Quantity of information	
Is the extent of coverage within scope?	4.2
Visual information	
Is the visual explanation clear and logical?	4.3
Credibility	
Does the app come from a legitimate source?	3.8

Discussion

Principal Findings

Our study evaluated the usability and perceived effectiveness of our mobile app "MyPeer," co-designed with immigrant youth to disseminate SRH educational content and raise awareness about existing services. We found that our app was well-received among adolescents and interest holders alike, with feedback and suggestions to improve its user-friendliness, content range, technical issues, and interactive features. By integrating a participatory approach at all stages, our process and findings provide essential insights into developing digital health interventions with youth.

Our content and app development were grounded and centered around adolescent perspectives and experiences, with adolescents involved at all stages through a participatory approach [17], allowing them to take ownership of the process and co-design an intervention that is tailored to their needs. A systematic review of SRH apps recommends that, for SRH apps to be impactful, they should not only be of high-quality design and have relevant information but also be tailored to their target group and well-tested with real users [19]. Through FGDs and individualized testing, we provided participants with the opportunity to critically analyze the app (eg, case-based scenarios) along with using the various features to their best potential. We made efforts to integrate their feedback into the design and content of the app, considering the diverse range of thoughts and perceptions among participants. Some features and suggestions were deemed not feasible to implement (eg, helper chat and discussion forum); however, we tried to find alternative solutions to address those unique ideas. We also asked participants in each FGD to express their thoughts on different features through various modalities (eg, whiteboard activities and surveys). Using various feedback mechanisms enabled us to receive feedback based on each participant's strengths and weaknesses while ensuring that no participant was left unheard. Martin et al [20] suggest that offering alternative ways to participants with diverse capabilities and strengths enhances the chance for effective technological development by acquiring valuable data from a population with high intra- and intervariability.

Participants initially found the app a bit too simple but easy to use. However, with multiple rounds of feedback and changes, they reported the app's design to be attractive and visually appealing. The variety of content delivery approaches (eg, videos and infographics) improved the user's motivation to engage with the app. Studies on youth-centered digital interventions have echoed similar findings, emphasizing the importance of multiple formats of information delivery to cater to youth at different cognitive levels and learning styles [10,12]. The quiz feature was appreciated for its interactive nature, with participants indicating enhancing the questions and adding more detailed explanations. Gamification within mobile apps has been seen to improve motivation, comprehension, and long-term retention of health-related information in various studies [11,20]. Overall, participants believed that the app was an effective resource to improve SRH education and knowledge among youth. To enhance the user experience, participants recommended addressing visual inconsistencies and technical glitches (eg, map calibration errors and unintended category duplications) within the app, which highlighted the importance of robust quality assurance during app development. Similar challenges have been identified in other usability studies with adolescent-centered apps, emphasizing the value of iterative testing with end users throughout the development process and considerations for sustainability of mobile technology [21,22].

A recurring theme within our findings was the desire among adolescents for more comprehensive and contextually relevant SRH information, with specific suggestions for including detailed content on topics such as contraception, aspects of sexual relationships, and cultural aspects of SRH. This is consistent with findings in similar studies where authors argue that the effectiveness of SRH interventions relies not only on the presentation of biological facts but also on the social, emotional, and cultural aspects of sexuality, which is meaningful to youth [9,12,19]. Another important consideration with youth-centered digital interventions is the need for user privacy and data security [23,24]. Many users expressed making the

XSL•FO
RenderX

content accessible for individuals not willing to register, and to include statements on how their data will be used and stored, which will enhance user trust and the legitimacy of the app. This may also be connected to social stigma and negative judgment that youth experience while accessing sexual health resources and care.

Stakeholder Feedback and Engagement

The feedback that we gathered from interest holders using the MARS checklist gave us valuable external insights and validation of the app's quality, design, functionality, and usability. Including interest holders and field experts in app testing has been recommended to obtain targeted data on the intended population, gain multidisciplinary perspectives, and assess the provider's perceptions on the perceived impact of such interventions [24]. Most interest holders rated the app's visual appeal and engagement to be high, along with strong agreement on the accurate and evidence-based nature of the app's content. Their observation reinforces the value of co-designing digital health interventions with youth and SRH experts, which is suggested by the World Health Organization for developing youth-friendly interventions [10]. However, some similar concerns—as expressed by adolescents—were raised regarding the inconsistencies within the graphics, app design, and content presentation on different screen sizes. This highlights the need for responsive design approaches in assuring that the app runs smoothly across a variety of devices and platforms [25,26].

Limitations

Although our study implemented co-design and participatory approaches at each step, we acknowledge a range of limitations. First, the sample size may not be representative of the entire adolescent population from migrant backgrounds in Canada. It is possible that certain cultural or ethnic groups' views may not have been expressed. Moreover, the limited number of interest holders may not have captured the full range of institutional perspectives. Second, although the data and feedback we received from the participants were rich, longer-term data on behavior change or knowledge retention were not collected, which could have provided insights into the app's effectiveness in increasing SRH knowledge. Moreover, we initially aimed to offer the MARS checklist to adolescents as well, so we can have a better understanding and quantitative data on the app's usability. However, very few adolescents completed the checklist, and the data were not sufficient for analysis. We, therefore, incorporated the various domains of the checklist throughout our qualitative discussions. Third, although our instructions requested that participants use the app and review all features through various approaches for a period of time, we are unable to determine if each participant gave the same thought and attention to each aspect of the app. To mitigate this risk, we applied measures to encourage participants to use the app routinely through daily email reminders, a detailed checklist, and timely feedback. We ensured that each participant in the FGD had an opportunity to use the app and a chance to express their concerns.

Future Implications

Our findings suggest that integrating comprehensive SRH education and culturally sensitive content can enhance the educational value of SRH apps for immigrant adolescents. Addressing interface issues and technical glitches is important to maintain user engagement and motivation to continue using the app. Moreover, it is essential that adolescents know how their data is being used and protected, considering the sensitive nature of app content and the need for privacy. To evaluate the app's effectiveness in improving SRH knowledge, future research is needed to assess the impact of the app on knowledge, attitudes, and behaviors through longitudinal designs or randomized controlled trials. Our future research will focus on conducting an effectiveness trial to evaluate the app's impact on the knowledge, attitudes, and behaviors of immigrant adolescents toward their SRH. We will also work to continuously update and evolve our content based on the latest evidence-based resources and guidelines, and add further content identified through the usability study, which is inclusive and appropriate to diverse populations. For this purpose, we will form further partnerships with youth and community-based organizations to evaluate how we can benefit from our respective work with adolescents collaboratively and engage in knowledge-sharing initiatives.

Conclusions

Our findings demonstrate that a co-designed, youth-friendly mobile app (MyPeer) can be impactful in addressing stigma, misconceptions, and barriers to access SRH information for adolescents from migrant backgrounds. Through a co-design and participatory approach, we gathered feedback from adolescents and interest holders on the design, functionality, and usability of our mobile app. Our app rated highly in its aesthetics, navigation, and design. The content was also appreciated for its accuracy and was deemed evidence-based by SRH experts. We received feedback and suggestions to enhance the breadth and depth of SRH content, address technical problems, and resolve inconsistencies related to graphics and design. Our findings suggest that immigrant adolescents need clear, evidence-based information offered through a variety of mediums. Beyond its acceptability and perceived effectiveness, our process provides valuable insights into incorporating youth voices within digital health interventions, indicating the importance of participatory design in promoting relevance, usability, and impact of such interventions. Our findings have broader implications for researchers, practitioners, and policymakers exploring the development of inclusive, evidence-based, youth-centered digital interventions. Health educational apps developed for youth to support their health and well-being should be cognizant of their perspectives and their preference to receive educational information.

XSL•FO
RenderX

Acknowledgments

We would like to acknowledge Nilanjan Ghatak for the development of the mobile app.

Funding

This work was supported by the Women and Children's Health Research Institute (WCHRI) and the Canadian Institute of Health Research (CIHR; grant #RES0053962).

Data Availability

The datasets used or analyzed during the study are available from the corresponding author upon reasonable request.

Authors' Contributions

S Meherali conceived the study. MA, AIR, and S Meherali conducted the interviews. MA, AIR, and S Meherali participated in data analysis. S Meherali and AIR prepared the initial draft (writing). S Meherali, MA, AIR, PSF, PP-H, ZK, SF, BS, PO-I, AV, ES, JPHW, SDS, WVN, and S Munro edited and reviewed the manuscript draft and commented on all drafts of the manuscript (writing).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Focus group facilitation guide.
[DOCX File , 17 KB - formative_v9i1e81115_app1.docx]

References

1. Adolescent health. World Health Organization. 2019 Nov 26. URL: https://www.who.int/health-topics/adolescent-health#tab=tab_1 [accessed 2025-12-03]
2. Langat EC, Mohiddin A, Kidere F, Omar A, Akuno J, Naanyu V, et al. Challenges and opportunities for improving access to adolescent and youth sexual and reproductive health services and information in the coastal counties of Kenya: a qualitative study. BMC Public Health 2024;24(1):484. [doi: 10.1186/s12889-024-17999-9] [Medline: 38365655]
3. Meherali S, Rehmani AI, Ahmad M, Kauser S, Scott Fiddler P, Pinzon-Hernandez P, et al. Between cultures and traditions: a qualitative investigation of sexual and reproductive health experiences of immigrant adolescents in Canada. Cult Health Sex 2024;27(12):1533-1549 [FREE Full text] [doi: 10.1080/13691058.2024.2436677] [Medline: 39648708]
4. Corley AG, Sprockett A, Montagu D, Chakraborty NM. Exploring and monitoring privacy, confidentiality, and provider bias in sexual and reproductive health service provision to young people: a narrative review. Int J Environ Res Public Health 2022;19(11):6576 [FREE Full text] [doi: 10.3390/ijerph19116576] [Medline: 35682160]
5. Sidamo N, Kerbo A, Gidebo K, Wado YD. Exploring barriers to accessing adolescents sexual and reproductive health services in South Ethiopia Regional State: a phenomenological study using Levesque's framework. Adolesc Health Med Ther 2024;15:45-61 [FREE Full text] [doi: 10.2147/AHMT.S455517] [Medline: 38562442]
6. Brayboy LM, McCoy K, Thamotharan S, Zhu E, Gil G, Houck C. The use of technology in the sexual health education especially among minority adolescent girls in the United States. Curr Opin Obstet Gynecol 2018;30(5):305-309 [FREE Full text] [doi: 10.1097/GCO.0000000000000485] [Medline: 30153129]
7. Dennison L, Morrison L, Conway G, Yardley L. Opportunities and challenges for smartphone applications in supporting health behavior change: qualitative study. J Med Internet Res 2013;15(4):e86 [FREE Full text] [doi: 10.2196/jmir.2583] [Medline: 23598614]
8. A data snapshot of Canadian youth and technology. Statistics Canada. 2024 Aug 12. URL: https://www.statcan.gc.ca/o1/en/plus/6782-data-snapshot-canadian-youth-and-technology [accessed 2025-12-03]
9. Ippoliti NB, L'Engle K. Meet us on the phone: mobile phone programs for adolescent sexual and reproductive health in low-to-middle income countries. Reprod Health 2017;14:11 [FREE Full text] [doi: 10.1186/s12978-016-0276-z] [Medline: 28095855]
10. Youth-centered digital health interventions: a framework for planning, developing and implementing solutions with and for young people. World Health Organization. 2021 Jul 19. URL: https://www.who.int/publications/i/item/9789240011717 [accessed 2025-12-03]
11. Tebb K. Leveraging serious video games to transform HIV prevention and care for adolescents and young adults: the case for playtest!. J Adolesc Health 2024;74(2):220-222. [doi: 10.1016/j.jadohealth.2023.11.002] [Medline: 38237976]
12. Guse K, Levine D, Martins S, Lira A, Gaarde J, Westmorland W, et al. Interventions using new digital media to improve adolescent sexual health: a systematic review. J Adolesc Health 2012;51(6):535-543. [doi: 10.1016/j.jadohealth.2012.03.014] [Medline: 23174462]

XSL•FO
RenderX

13. Ibitoye M, Lappen H, Freeman R, Jordan AE, Aronson ID. Technology-based interventions to increase point-of-care HIV testing and linkage to care among youth in the US: a systematic review. AIDS Behav 2021;25(6):1829-1838 [FREE Full text] [doi: 10.1007/s10461-020-03112-9] [Medline: 33231846]
14. Benoit JRA, Louie-Poon S, Kauser S, Meherali S. Promoting adolescent sexual and reproductive health in North America using free mobile apps: environmental scan. JMIR Pediatr Parent 2022;5(4):e33826 [FREE Full text] [doi: 10.2196/33826] [Medline: 36194450]
15. Meherali S, Munro S, Puinean G, Salami B, Wong JP, Vandermorris A, et al. Co-designing a sexual health app with immigrant adolescents: protocol for a qualitative community-based participatory action research study. JMIR Res Protoc 2023;12:e45389 [FREE Full text] [doi: 10.2196/45389] [Medline: 36947124]
16. Chen E, Leos C, Kowitt SD, Moracco KE. Enhancing community-based participatory research through human-centered design strategies. Health Promot Pract 2019;21(1):37-48. [doi: 10.1177/1524839919850557] [Medline: 31131633]
17. Stoyanov SR, Hides L, Kavanagh DJ, Zelenko O, Tjondronegoro D, Mani M. Mobile app rating scale: a new tool for assessing the quality of health mobile apps. JMIR Mhealth Uhealth 2015;3(1):e27 [FREE Full text] [doi: 10.2196/mhealth.3422] [Medline: 25760773]
18. Terhorst Y, Philippi P, Sander LB, Schultchen D, Paganini S, Bardus M, et al. Validation of the Mobile Application Rating Scale (MARS). PLoS One 2020;15(11):e0241480 [FREE Full text] [doi: 10.1371/journal.pone.0241480] [Medline: 33137123]
19. Muehlmann M, Tomczyk S. Mobile apps for sexual and reproductive health education: a systematic review and quality assessment. Curr Sex Health Rep 2023;15(2):77-99. [doi: 10.1007/s11930-023-00359-w]
20. Martin A, Caon M, Adorni F, Andreoni G, Ascolese A, Atkinson S, et al. A mobile phone intervention to improve obesity-related health behaviors of adolescents across Europe: iterative co-design and feasibility study. JMIR Mhealth Uhealth 2020;8(3):e14118 [FREE Full text] [doi: 10.2196/14118] [Medline: 32130179]
21. Fedele DA, Cushing CC, Fritz A, Amaro CM, Ortega A. Mobile health interventions for improving health outcomes in youth: a meta-analysis. JAMA Pediatr 2017;171(5):461-469 [FREE Full text] [doi: 10.1001/jamapediatrics.2017.0042] [Medline: 28319239]
22. Shikako K, Mogo ERI, Grand-Maison V, Simpson R, Pritchard-Wiart L, Majnemer A, Jooay App Research Group. Designing user-centered mobile health initiatives to promote healthy behaviors for children with disabilities: development and usability study. JMIR Form Res 2021;5(9):e23877 [FREE Full text] [doi: 10.2196/23877] [Medline: 34528886]
23. Pretorius C, Chambers D, Coyle D. Young people's online help-seeking and mental health difficulties: systematic narrative review. J Med Internet Res 2019;21(11):e13873 [FREE Full text] [doi: 10.2196/13873] [Medline: 31742562]
24. Seretlo RJ, Smuts H, Mokgatle MM. Development of an mHealth app by experts for queer individuals' sexual-reproductive health care services and needs: nominal group technique study. JMIR Form Res 2024;8:e59963 [FREE Full text] [doi: 10.2196/59963] [Medline: 39167434]
25. Durgekar SR, Rahman SA, Naik SR, Kanchan SS, Srinivasan G. A review paper on design and experience of mobile applications. EAI Endorsed Scal Inf Syst 2024 Jan 29 [FREE Full text] [doi: 10.4108/eetsis.4959]
26. Tomlinson M, Rotheram-Borus MJ, Swartz L, Tsai AC. Scaling up mHealth: where is the evidence? PLoS Med 2013;10(2):e1001382 [FREE Full text] [doi: 10.1371/journal.pmed.1001382] [Medline: 23424286]

Abbreviations

AAG: adolescent advisory group
FGD: focus group discussion
MARS: Mobile App Rating Scale
mHealth: mobile health
SRH: sexual and reproductive health

Edited by A Stone; submitted 22.07.25; peer-reviewed by N Levitz, AA Villa-Rueda; comments to author 22.09.25; revised version received 20.11.25; accepted 24.11.25; published 25.12.25.

Please cite as:
Meherali S, Rehmani AI, Ahmad M, Scott-Fiddler P, Pinzón-Hernández P, Khan Z, Flicker S, Okeke-Ihejirika P, Salami B, Stroulia E, Vandermorris A, Pui-Hing Wong J, Norman WV, Scott SD, Munro S
A Mobile App (MyPeer) Co-Designed With Immigrant Adolescents for Better Sexual and Reproductive Health: Usability Study
JMIR Form Res 2025;9:e81115
URL: https://formative.jmir.org/2025/1/e81115
doi:10.2196/81115
PMID:41447463

XSL•FO
RenderX

©Salima Meherali, Amyna Ismail Rehmani, Mariam Ahmad, Piper Scott-Fiddler, Paula Pinzón-Hernández, Zeba Khan, Sarah Flicker, Philomina Okeke-Ihejirika, Bukola Salami, Eleni Stroulia, Ashley Vandermorris, Josephine Pui-Hing Wong, Wendy V Norman, Shannon D Scott, Sarah Munro. Originally published in JMIR Formative Research (https://formative.jmir.org), 25.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

A Blended Educational Program to Promote Dialogue on Patient Safety Between Patient and Family Advisory Councils and Health Care Organizations: Codevelopment Study

Yannick Blum, MSc; Larissa Brust, MSc; Matthias Weigl, Prof Dr; Qëndresa Rramani Dervishi, PhD

Institute for Patient Safety, Medical Faculty, University Hospital Bonn, Venusberg-Campus 1, Bonn, Germany

Corresponding Author:
Yannick Blum, MSc
Institute for Patient Safety, Medical Faculty, University Hospital Bonn, Venusberg-Campus 1, Bonn, Germany

Abstract

Background: Reducing patient harm and improving patient safety is a central objective in global health care. Effective communication and meaningful patient engagement are considered essential strategies to achieve this goal. However, implementation of structured and strategic patient engagement at the organizational level remains limited, particularly in the context of patient safety. Patient and family advisory councils (PFACs) offer a promising model to enhance organizational-level patient engagement, yet guidance on implementation and targeted training for PFAC members is scarce.

Objective: This study aimed to codevelop an evidence-informed, blended educational program designed to strengthen PFAC members' competencies in patient safety and communication, and to foster strategic collaboration between PFACs and health care organizations.

Methods: The intervention was systematically developed using a logic model framework that structures the development process from available and required resources to the ultimate objectives and impacts. The primary target group included PFAC members, such as patients, relatives, and advocates, as well as health care representatives in leadership, quality management, or coordination roles. The program's content and structure were informed by a nationwide needs and requirements analysis among PFAC members, conducted using a mixed methods Delphi approach, and by a rapid scoping review on existing educational resources and evidence on PFAC engagement in patient safety.

Results: Our *Partners for Patient Safety* blended educational program consisted of 2 modular components: a self-paced e-learning module and a subsequent on-site workshop module. Content addressed three core topics: (1) fundamentals of patient safety, (2) engagement of PFACs, and (3) communication and collaborative goal setting. The e-learning module provided theoretical knowledge using diverse didactic formats, such as interactive tasks, videos, and downloadable materials, and included applied examples using established decision-making and goal-setting frameworks. The workshop module built on the e-learning module and facilitated local implementation through collaborative exercises focused on stakeholder perspectives, communication barriers, and joint goal development. Both modules were aligned with defined learning objectives and combined passive and active learning strategies to promote engagement and practical application.

Conclusions: The *Partners for Patient Safety* program seeks to develop PFAC members' competencies, promote collaboration in patient safety, and foster a culture of safety and partnership within health care organizations. By combining theoretical knowledge with practical, collaborative learning, the program addresses key barriers to effective PFAC engagement at the organizational level. Its modular design allows flexible implementation and has the potential to strengthen cooperation between PFACs and health care representatives, ultimately improving patient safety outcomes. Further evaluation of the program's implementation and effectiveness is needed.

Trial Registration: German Clinical Trials Registry DRKS00034733; https://drks.de/search/en/trial/DRKS00034733

(JMIR Form Res 2025;9:e79286) doi:10.2196/79286

KEYWORDS

patient safety; communication; collaboration; patient engagement; educational program

Introduction

Reducing patient harm and improving patient safety is a central goal in global health care, as emphasized by the World Health Organization [1]. Effective communication within health care organizations and improved patient engagement are essential strategies for achieving this objective [1-3]. Nonetheless, respective approaches to foster effective patient engagement for strategic promotion of patient safety on the level of health care organizations are rarely reported [4]. Thus, effective and

XSL•FO
RenderX

evidence-based strategies for successfully implementing patient engagement at the organizational level are greatly needed.

Patient engagement refers to involving patients and their representatives in a patient's health care journey, potentially taking place at all levels of care; it can occur at all levels of care and range from consultation to partnership [5]. Despite widespread advocacy in research and policy for increased meaningful patient engagement across all levels of health care [6,7], in reality, its implementation remains limited, especially in the context of patient safety [8-10]. These shortcomings are particularly pronounced at the organizational level, where strategic engagement of patients and their representatives faces several barriers [10]. One promising approach to increase patient engagement at the organizational level is the establishment and integration of patient and family advisory councils (PFACs) in health care organizations [11,12]. PFACs, or "patient advisory boards," are often constituents of patients and their families, as well as members of other patient representative groups (eg, self-help groups), who might also be engaged in existent processes such as quality improvement and patient safety [11].

Despite a few general recommendations on implementation [13,14], practice guidelines and standards regarding implementation of patient engagement at the organizational level are rare [11,15]. While available research suggests that engagement of advisors in safety and quality improvement initiatives yields positive outcomes [16,17], our previous rapid scoping review found that hands-on recommendations or practice standards for advisors assuming this role are mostly lacking (reported elsewhere). This gap persists despite long-standing calls from patient representatives for more training and education on patient safety topics [18]. Moreover, a previous survey conducted in the German health care setting revealed that levels of patient engagement and patient advocacy are perceived as low, partly attributed to a shortage of respective educational programs for patients and their representatives [10]. Addressing these persistent needs requires targeted interventions to enhance engagement and skills of patient representatives such as those within PFACs.

To this end, we codeveloped a blended educational program on relevant topics on patient safety and communication for PFACs and associated health care representatives, drawing upon needs of PFAC members and with continuous support of patient representatives. This intervention program will strengthen patient safety and communication skills of PFAC members and increase patient engagement at the organizational level of health care organizations, thereby promoting patient safety.

Our main objective was to describe the process of codevelopment, the framework, and content of the *Partners for Patient Safety* (P4PS) program for PFACs and associated health care representatives. This paper presents the underlying evidence and provides a detailed overview of the structure, methods, and components of the intervention.

Methods

Overview

This section covers (1) the systematic development process; (2) the evidence base of previous extensive work, specifically a needs and requirements analysis and a rapid scoping review (reported elsewhere in detail) [19]; and (3) the setting and target group of the P4PS program.

We used the Template for Intervention Description and Replication (TIDieR; Checklist 1) and the Guidance for the Reporting of Intervention Development (GUIDED; Checklist 2); both frameworks ensure comprehensive reporting of the intervention [20] and enhance its reproducibility [21,22].

Ethical Considerations

The study was preregistered in the German Clinical Trials Registry (DRKS00034733) and approved by the ethics committee of the medical faculty at the University of Bonn (2024 - 236-BO). Data collection and processing were carried out in accordance with the Declaration of Helsinki, the European Union's General Data Protection Regulation, and the German Federal Data Protection Act. The privacy and confidentiality of all participants were strictly maintained; all identifying information was removed or anonymized before data analysis. All participants signed an informed consent form and received a monetary compensation of €120 (US $1.16) for their participation.

Intervention Development

Intervention Framework

For systematic development of the intervention, we used the logic model [23] as a structured framework and well-established tool in public health and quality improvement research, facilitating the systematic planning, implementation, and evaluation of interventions by mapping the relationships between inputs, activities, outputs, objectives, results, and goals [24,25]. Specifically, the model guided the derivation of necessary inputs to promote exchange and collaboration between the PFAC and the health care organization through the development of skill acquisition in terms of patient safety and communication among learners. The details of the model were developed iteratively by 2 members of the research team based on problem identification and objectives of the intervention. Figure 1 shows the general process and procedure for developing the P4PS program. This methodological approach ensured a goal-oriented intervention design based on key assumptions and evidence. An overview of the columns of the logic model relevant to the process of intervention development is provided in Table 1. The complete model can be found in Multimedia Appendix 1.

XSL•FO
RenderX

Figure 1. Process of developing the Partners for Patient Safety program. PE: patient engagement; PFAC: patient and family advisory council; PS: patient safety.

XSL•FO
RenderX

Table . Logic model for Partners for Patient Safety program development. The table outlines the levels, descriptions, and key assumptions of the logic model. The vertical logic works backward from inputs to goals to establish a conceptual framework and operational definitions for each level, while the horizontal logic describes the requirements needed to achieve each level. The columns objective verifiable indicators and means of verification of the model are intended for evaluation purposes and are therefore not included here.

Level	Description	Key assumptions
Goal	• Improving PS[a] and quality of care at an organizational level by engaging patients, their families, and representatives in health care	• Increased engagement of PFACs[b] in health care in Germany improves PS • Engagement in PS is hindered by inadequate knowledge of PFAC members regarding PS and communication
Results	• Increased PFAC engagement in decisions on quality, care processes, and PS issues • Sustainable networks and partnerships between PFACs and HCO[c]	• PFACs and HCRs[d] are interested in promoting PFAC engagement in PS topics • PFACs and HCO know how PS-relevant topics can be evaluated
Objectives	• Increased engagement of PFACs through improved quality and quantity of communication between participating PFACs and HCR on PS topics	• PFACs need better PS and dialogue skills for gaining a competent voice in health care • Communication and partnership-based dialogue between PFACs and HCR must be expanded

Outputs	• Participating PFAC and HCR
	  • Understand the essential principles of PS and recognize events relevant to PS
	  • Understand the models and strategies of effective communication presented
	  • Understand examples of good PFAC engagement in care processes
	  • Understand the perspectives of various stakeholders and explain their activities and functions
	  • Identify barriers and misunderstandings in communication between PFAC and HCO
	  • Develop new possibilities for future activities and collaboration
		• Participating PFAC and HCO
		  • Are interested in promoting PFAC engagement in PS processes
		  • Are interested in discussing and reflecting on cases of (near) harm that have occurred in patient care
		  • Speak openly about communication barriers and PS risks
		• Additionally to the above
		  • HCOs provide opportunities for closer collaboration
		  • Resources are made available to ensure improved collaboration

Activities	• Intervention development based on needs and requirements as well as international literature
	• PFAC and HCR
	  • Are trained in PS content, examples of PFAC engagement, and communication skills
	  • Are presented with new communication activities to foster dialogue
	  • Are given the opportunity to systematically apply the options presented to their own situation
	  • Receive support to promote common goals and future collaboration
		• PFAC and HCR take part in surveys and interventions
		• PFACs share specific needs and requirements on PS and dialogue topics
		• Interventions developed are effective in promoting PS and communication skills
		• Interventions developed are effective in promoting collaboration

Inputs	• Interview guide for needs and requirements analysis
	• Online questionnaire to prioritize content for intervention
	• Literature research on comparable, already existing programs
	• Other infrastructure and resources, for example,
	  • Funding
	  • Personnel expertise of study team
	  • Support with and resources for the development of educational programs
		• Responsible people provide
		  • Resources to develop the intervention
		  • Materials for the participants
		• Generally existing
		  • Adequate funding
		  • Availability of PFAC and HCR
		  • Willingness and consent of PFACs and HCR to participate
		  • No ethical concerns about the project

XSL•FO
RenderX

[a]PS: patient safety.

[b]PFAC: patient and family advisory council.

[c]HCO: health care organization.

[d]HCR: health care representative.

Target Population

The target group of the P4PS program comprised members of PFACs affiliated with health care organizations. This included representatives of patient-oriented interests (eg, patients, relatives, patient advocates, and representatives of other patient-related organizations) as well as professionals of the health care organization who were responsible for leadership, general management, or administration of the respective PFAC (eg, coordinators and moderators). In line with the intervention's objective of promoting collaboration between PFACs and health care organizations, program development was additionally geared toward the participation of senior professionals and leaders from collaborating health care organizations, such as senior physicians and head nurses, quality and risk managers, or members of hospital management and board of directors.

Evidence to Inform Development of the Intervention

Two previous main sources informed the development of the P4PS program in terms of topics and formats: a needs and requirement analysis (a mixed methods study with a Delphi approach) and a literature review on existing educational resources for PFACs; both are reported elsewhere in detail [19]. However, a brief summary of the key findings from both studies is provided in Multimedia Appendix 2 [19].

The mixed methods study with a Delphi approach previously identified PFACs' needs regarding competencies in patient safety and communication as well as PFAC engagement [19]. This approach ensured a high degree of participation in informing and developing this intervention. According to the continuum of patient engagement outlined by Carman et al [5], it actually achieved "active involvement" of PFAC members. According to the stages of participation described by Wright [26], the degree of participation in this study can be classified as medium, meaning participants had partial decision-making authority. A total of 19 participants from 6 nationwide PFACs took part in the two rounds of surveys, which consisted of (1) interviews and (2) questionnaires. The responses were evaluated descriptively and consensus-oriented (criterion: 85% agreement). The mixed methods analysis was performed sequentially and convergently. Among other findings, the results highlighted the need to address the following: (1) fundamentals of patient safety, error occurrence and error prevention, and knowledge on quality and risk management; (2) support in understanding and defining PFAC roles, responsibilities, and tasks; and (3) precise and clear communication of information, complaints, and problems. In terms of formats of intervention, participants desired a modularized program with theoretical and practical components, delivered online or on-site depending on the topics and organization, with opportunities for partial implementation in their respective settings [19].

To further inform the development of the P4PS program, we previously used peer-reviewed and relevant gray literature from various databases and search engines to conduct a parallel literature search in the form of a rapid scoping review of existing evidence and educational resources on patient safety for PFACs in health care organizations (reported elsewhere in detail). Although the number of the identified sources was limited, 13 articles and sources featuring various formats, content, and topics could be identified [16,27-38]. This information was extracted, and, together with the results of the needs and requirement analysis, used to inform the content and format of the intervention. In addition, other sources, such as previous projects and teaching materials, as well as those from various organizations, were used to generate intervention content [1,39-43].

While the general objectives of the intervention were defined on the basis of the logic model (refer to the Intervention Framework section), the specific learning objectives for the P4PS program on the core topics of patient safety, PFAC engagement, and communication (Textbox 1) were defined based on the taxonomy proposed by Bloom et al [44] and its later revision by Krathwohl [45].

XSL•FO
RenderX

Textbox 1. Learning objectives of the Partners for Patient Safety program for each topic covered.

- Patient safety (PS)
  - Understand the fundamentals of PS and quality and risk management as well as recognize their relevance in health care.
  - Analyze examples of errors in health care and complex factors influencing PS and understand different measures to reduce errors (eg, standardization of processes and reporting systems).
  - Apply legal principles in the context of PS in health care.
- Patient and family advisory council (PFAC) engagement
  - Analyze the different levels of engagement of patients and their representatives and the characteristics of these levels.
  - Understand examples of the engagement of PFACs and reflect on their potential impact on the quality and safety of care.
  - Analyze their own understanding of their role, and that of the PFAC as a whole, and reflect on internal and external responsibilities.
  - Adopt and analyze different perspectives (PFAC, health care representatives, and administration) to develop a deeper understanding of the challenges and needs of the other groups.
  - Identify and evaluate at least 3 specific misunderstandings or barriers to collaboration.
  - Analyze possible activities and tasks to engage the PFAC in quality and safety of care in the health care organization.
  - Using methods learned, jointly formulate realistic and measurable goals for collaboration between the PFAC and the health care organization.
- Communication
  - Understand the importance of open and appropriate communication for collaboration in health care contexts.
  - Understand basic communication models and apply them specifically in different contexts.
  - Apply supportive models for decision-making and goal setting in collaborative settings.

Intervention Development Group and Supportive Consultations

A total of 5 patient safety researchers, consisting of psychologists, health scientists, and a physician, designed and developed the P4PS program. The research team received support from local institutional services (center for human resources development, university computer center, and faculty members), who provided technical and pedagogical advice regarding the digital realization, organization, and structure of the intervention. Finally, the P4PS program was informally reviewed by an external patient representative, whose user-oriented feedback was incorporated iteratively.

Results

In this section, we primarily describe the P4PS program in detail and complement this with descriptive information on program use (module completion and use duration) and participant feedback.

Content and Format of the Intervention

On the basis of key educational needs identified in preceding needs and requirements analysis and the literature review, the P4PS program addressed three core topics: (1) fundamentals of patient safety, (2) engagement of PFACs, and (3) communication and collaborative goal setting.

The P4PS program comprised two major modules: (1) an e-learning module, followed by (2) an on-site workshop module. This structure allows for a combination of passive and active learning approaches and enables participants to receive information as well as apply it in real-world examples. Figure 2 shows the contents, methods, and didactics of the 2 modules of the P4PS program. Details on the e-learning module and workshop module are presented in the following sections.

XSL•FO
RenderX

Figure 2. Modules of the Partners for Patient Safety program. PS: patient safety; PFAC: patient and family advisory council.

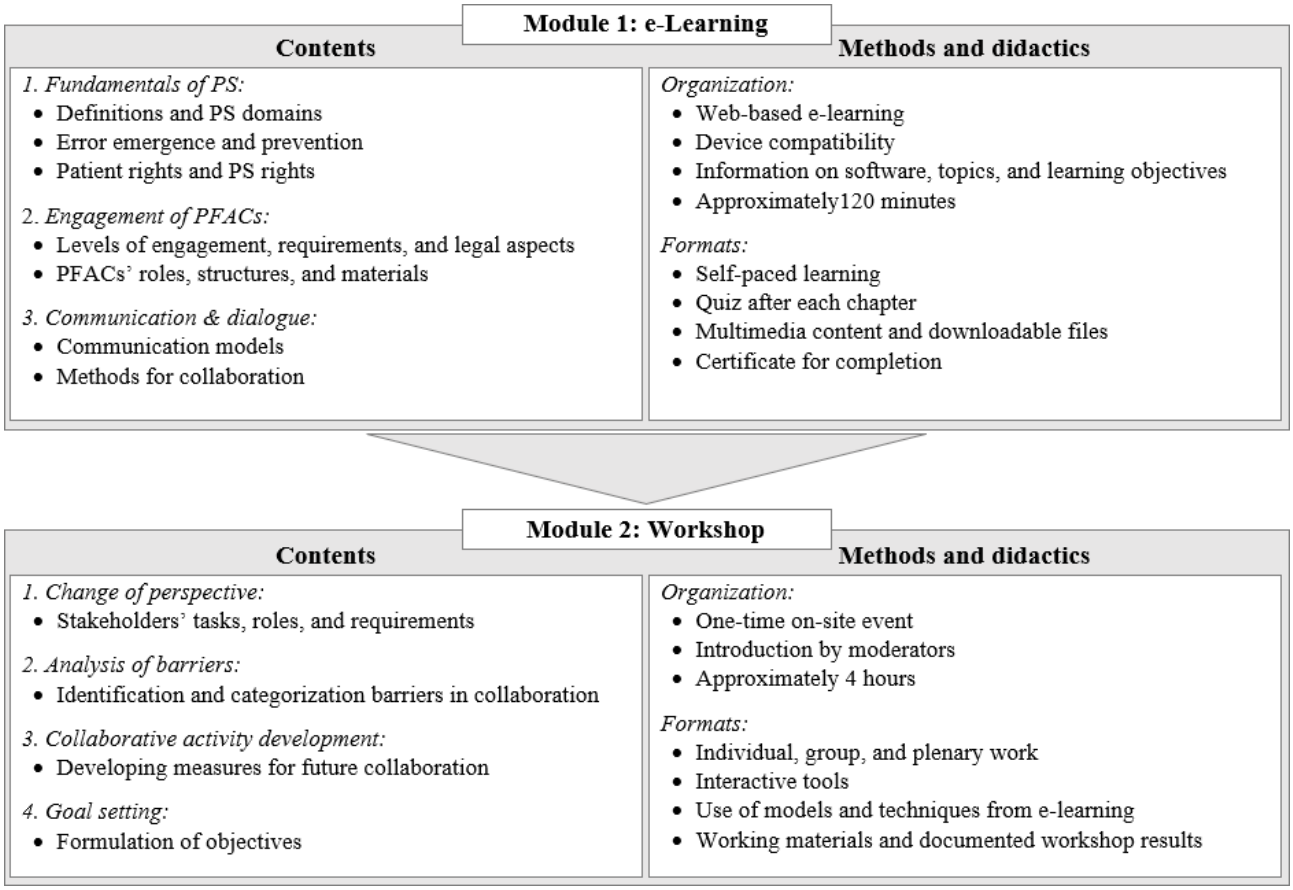

e-Learning (Module 1)

Topics and Key Content

The e-learning module was designed based on the objective and core topics developed for the P4PS program (refer to the Content and Format of the Intervention section) with three chapters: (1) fundamentals of patient safety, (2) engagement of PFACs, and (3) communication and dialogue.

The first chapter covered definitions of patient safety and adverse events, key patient safety domains, the significance of the topic (eg, adverse event rates and mismanaged conditions), and the sociotechnical complexities that affect patient safety and contribute to errors. Learners were informed about patients' rights (eg, World Health Organization Patient Safety Rights Charter) and methods for identifying and reducing errors (eg, Swiss cheese model, standardization, hygiene measures, and critical incident reporting).

The second chapter captured engagement of PFACs and provided basic information on patient engagement (ie, levels of engagement, legal requirements, and practical examples of positive influence of PFAC engagement). Learners were presented with videos and various downloadable content covering internal role understanding, examples of high-functioning PFACs organization and structure, and other materials (eg, meeting minutes, PFAC charter, and orientation manual for new members).

The third chapter focused on communication and dialogue for patient safety and included the Speak Up initiative [46] and communication models that fostered respectful dialogue (eg, 4-side model, nonviolent communication, and sandwich method [47-49]). To improve collaboration between PFACs and health care facilities, 2 frameworks were introduced: the facts, options, risks and benefits, decision, execution, and check (FORDEC) decision-making method [50] and the specific, measurable, attainable, realistic, and time-bound (SMART) goal-setting framework [51]. The FORDEC method was selected due to its potential for team decision support and preplanning of risks and failures in complex situations [52,53]. The SMART method is an established approach that has been successfully applied in various disciplines, including medical education [54-57].

The e-learning module concluded with a practice example that demonstrated how a hospital addressed a patient safety issue related to an error, linking it to the course content. More specifically, in this example, a fictitious PFAC collaborated with hospital management to reduce readmission rates. To this end, they set common goals and design strategies to achieve these goals using the SMART and FORDEC methods, which were introduced in earlier chapters of the e-learning module. The developed strategies included informing patients about the Speak Up initiative and providing them with a logbook to record medication changes, questions, and contact information after discharge. This logbook is made available to learners as a downloadable document.

The e-learning module provided various didactic elements, allowing learners to choose suitable options for obtaining information. Options included audio texts, labeled graphics with extra content, and short videos. Learners also engaged with

XSL•FO
RenderX

diagrams, graphs, matching exercises, and scenario-based tasks. The intensity of the content delivered increased over the course of the e-learning module, such that basic information on patient safety was followed by in-depth information on strategies, activities, and tasks to promote the engagement of PFACs in patient safety in health care organizations. Learners received downloadable files with information, checklists, and templates for application to their PFAC. Each topic concluded with a short quiz intended to reinforce learning.

We used Rise 360 software (Articulate Global, LLC) to create the e-learning module. Embedded, self-created videos were generated using Microsoft PowerPoint and edited in the DaVinci Resolve editing program (Blackmagic Design Pty Ltd). Audio tracks were generated using the ElevenLabs' artificial intelligence Text to Speech software. The final e-learning module [58] has been made available on the home page of the Institute for Patient Safety.

Didactic Approach, Design, and Procedure

Prior to the introduction of the first chapter, learners received instructions on software management, an overview of topics, learning objectives (Textbox 1), and information on taking breaks.

The e-learning module was designed to be flexible and completed by each learner independently at their own pace. It was accessible via a link and was to be completed before the workshop module (module 2) to aid knowledge retention. It worked on any device (PC, smartphone, tablet, or laptop) with a stable internet connection. A larger device was recommended for better visibility. This module had a duration of approximately 120 minutes.

Workshop (Module 2)

Topics and Key Content

The activities carried out in the workshop module were based on completion of the previous e-learning module (module 1). The workshop module addressed the individual situation of each PFAC and health care organization as well as the respective circumstances. While the e-learning module provided a theoretical foundation, the workshop module focused on translation and contextualization to locally promote mutual communication and collaborative goal development.

On the basis of the formulated learning objectives (Textbox 1), 4 different activities were performed by the participants during the workshop module:

1. A change of perspective, aiming to strengthen learners' understanding of the complexity of tasks and requirements of the different stakeholders and their individual functions in the care process
2. An identification and categorization of existing barriers and misunderstandings in the previous collaboration between the PFAC and the health care organization
3. The identification or refinement of activities to promote future collaboration
4. The collaborative formulation of concrete objectives for these activities

Didactic Approach, Design, and Procedure

In the beginning, the moderators introduced the topic with a short presentation outlining the learning objectives (Textbox 1). Learners were then informed about the course of the workshop module and presented reminders on the preceding e-learning module (eg, communication techniques).

All tasks were performed using different methodological approaches (eg, individual assignments, small or whole group tasks, discussions, and short presentations) and different media (eg, interactive queries, working papers, handouts, and flipcharts). Learners were encouraged by the moderators to apply the communication techniques and models presented in the e-learning module (refer to the e-Learning section). Collaborative development and statement or refinements of objectives were carried out by the learners using the FORDEC and SMART methods, learned in the e-learning module (refer to the e-Learning section). Learners receive additional handouts and worksheets for this purpose. A tabular overview of the workshop activities, methodology, and scheduling can be found in Multimedia Appendix 3.

The results of the workshop modules were documented and shared with participants in a summary. The workshop modules were organized as one-time and on-site events that lasted approximately 4 hours.

Program Use and Participant Feedback

A total of 36 participants (PFAC members and health care representatives) engaged with the P4PS program. Of these 36 participants, 33 (92%) completed module 1 (ie, e-learning module) and 26 (72%) took part in module 2 (ie, workshop module). Completion rates across the 8 chapters of the e-learning module ranged from 94% (31/33) to 100% (36/36). On average, participants spent 140.5 (SD 65.3) minutes completing the e-learning module. Regarding feasibility and preliminary effectiveness, the program was generally described as usable, user-friendly, relevant, and effective for building competencies in patient safety and communication. Challenges included technical skills, lay language, and the degree of interactivity, which did not, however, hinder completion. Details of the formal evaluation, including outcomes on feasibility and preliminary effectiveness, are reported in a separate publication.

Discussion

Principal Findings

Responding to the call for increased engagement of patients and their representatives in patient safety at the organizational level and associated challenges of inadequate programs to promote collaboration, we developed the P4PS program on patient safety and communication for members of PFACs and health care representatives. To the best of our knowledge, this is one of the first codeveloped training initiatives specifically designed to build PFAC capacity in patient safety–related care issues. Moreover, this P4PS program can be a novel step toward institutionalizing collaborative structures. Evaluating its feasibility and effectiveness is crucial to understand how such innovations can be sustainably implemented and scaled; however, to provide more detailed reporting, this study focuses

XSL•FO
RenderX

on the systematic development of the intervention itself, while the evaluation is reported in detail elsewhere.

In the course of the program's development, we sought to establish conceptual and methodological rigor by drawing on established frameworks such as the logic model and the taxonomy proposed by Bloom et al [44], as well as by engaging PFAC members and stakeholders. The use of the logic model [59] is expected to support targeted planning based on existing resources and to facilitate the program's implementation, replication, and translation into practice. In turn, the taxonomy proposed by Bloom [60] supports the gradual development of cognitive skills and continuous learning progress during the P4PS program. Additionally, the engagement of PFAC members for intervention development ensured a needs-based intervention and suitable content for knowledge transfer. Through our comprehensive reporting, which is often lacking in similar initiatives [20,61], we enhance the potential for replication in other health care settings and adaptation to different contexts.

Concerning its future implementation, our intervention development addressed several challenges that often limit sustainable success of such patient engagement measures in practice. For instance, to ensure its relevance to the national context (ie, Germany), we screened the international literature [37,38,41,42] and finalized key content in consultation with a patient representative. Furthermore, we incorporated findings from a previous needs and requirements analysis [19], a scoping review, and an earlier German intervention study on skills development into the intervention [39]. Despite the thorough planning stage, we acknowledge that this approach requires significant time and resource investment, which can be a barrier for feasibility across health care organizations [62]. As involvement of stakeholders and recruiting participants can be challenging, health care organizations should proactively communicate the relevance of the program in dialogue with PFACs. Visibility and attractiveness can be increased by monitoring participation data and linking it to improvements in patient safety initiatives [63].

Another factor affecting implementation is the background of PFAC members, which can vary widely across organizations. Studies show that a lack of motivation, feedback, or knowledge [64,65] as well as technical and time-related hurdles inhibit engagement [66]. Educational programs integrated into the onboarding process can help establish a shared knowledge base. Support personnel from the health care sector may facilitate access to resources [27], while incentives such as certificates or perks for long-term project participation may strengthen willingness to engage. Optional e-learning modules that allow for tailored content selection may address the varying needs among stakeholders of different PFACs [67]. However, feasibility must be maintained to ensure accessibility and focus. Consideration of such context-specific factors will also be relevant in evaluating the program's transferability.

Future evaluation of the P4PS program plays a decisive role in further development and improvement. As one of the first initiatives of its kind in Germany, the P4PS program builds on specific needs and fulfills the specific purpose of training PFACs, which requires a detailed description and has led us to address these needs pragmatically. Future program testing for feasibility and effectiveness using a mixed methods approach will allow identification of strengths and weaknesses.

It can be assumed that the P4PS program will only reach its full potential if favorable policy, structural, and procedural changes are made in the health care system at the same time [68-70]. Future research should pursue measures to implement such programs, especially depending on different underlying organizational contexts [71]. A comparable intervention, if properly implemented, has the potential to initiate appropriate changes to strengthen the voice of patients and their advocates in promoting safe care systems.

Conclusions

This study provides a comprehensive account of the systematic development of the P4PS program for PFACs and health care representatives. Building on our previous efforts to identify PFACs' needs and relevant evidence, the program has the potential to be used as a viable measure to promote PFACs' engagement in fostering patient safety in health care organizations. As the landscape of health care services evolves quickly, continued commitment to empowering patients and patient advocates through needs-oriented and pragmatic measures is essential in fostering a culture of safety and mutual collaboration.

Acknowledgments

The authors thank their colleagues at the Institute for Patient Safety for providing teaching materials that enriched the development of their e-learning module. They also thank the Center for Human Resources Development and the University Computer Center at the University Hospital Bonn for their guidance in software selection and program development. Special thanks go to Loes Broekmate for her support and thorough review of the program. This work was supported by the German Federal Ministry of Health (ZMII2-2524PAT001) and the Open Access Publication Fund of the University of Bonn. During the preparation of this work, the authors used ChatGPT-4 to improve the readability and language of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the final version of the publication.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request. The intervention developed in the course of the study (e-learning module and workshop module) is not based on any programming syntax. The intervention can be accessed at the University Hospital Bonn website [58].

XSL•FO
RenderX

Authors' Contributions

Conceptualization: YB, LB, MW, QRD
Data curation: YB, LB, QRD
Funding acquisition: MW
Investigation: YB, LB, QRD
Methodology: YB, LB, MW, QRD
Project administration: YB, LB, QRD
Supervision: MW, QRD
Validation: YB, LB, QRD
Visualization: YB
Writing – original draft: YB
Writing – review & editing: YB, LB, MW, QRD

Conflicts of Interest

None declared.

Multimedia Appendix 1
Complete logic model for the intervention description and potential evaluation.
[PDF File, 149 KB - formative_v9i1e79286_app1.pdf]

Multimedia Appendix 2
Summary of key findings from the main sources that informed intervention development.
[PDF File, 215 KB - formative_v9i1e79286_app2.pdf]

Multimedia Appendix 3
Workshop activities, methodology, and schedule.
[PDF File, 169 KB - formative_v9i1e79286_app3.pdf]

Checklist 1
TIDieR checklist.
[PDF File, 307 KB - formative_v9i1e79286_app4.pdf]

Checklist 2
GUIDED checklist.
[PDF File, 137 KB - formative_v9i1e79286_app5.pdf]

References

1. Global Patient Safety Action Plan 2021-2030: Towards Eliminating Avoidable Harm in Health Care: World Health Organization; 2021.
2. Burgener AM. Enhancing communication to improve patient safety and to increase patient satisfaction. Health Care Manag (Frederick) 2017;36(3):238-243. [doi: 10.1097/HCM.0000000000000165] [Medline: 28657914]
3. Park M, Giap TT. Patient and family engagement as a potential approach for improving patient safety: a systematic review. J Adv Nurs 2020 Jan;76(1):62-80. [doi: 10.1111/jan.14227] [Medline: 31588602]
4. Marzban S, Najafi M, Agolli A, Ashrafi E. Impact of patient engagement on healthcare quality: a scoping review. J Patient Exp 2022 Sep 16;9:23743735221125439. [doi: 10.1177/23743735221125439] [Medline: 36134145]
5. Carman KL, Dardess P, Maurer M, et al. Patient and family engagement: a framework for understanding the elements and developing interventions and policies. Health Aff (Millwood) 2013 Feb;32(2):223-231. [doi: 10.1377/hlthaff.2012.1133] [Medline: 23381514]
6. Interim report - based on the first survey of patient safety in WHO member states. : World Health Organization; 2023 URL: https://www.who.int/publications/m/item/interim-report---based-on-the-first-survey-of-patient-safety-in-who-member-states [accessed 2025-10-31]
7. O'Hara JK, Canfield C. The future of engaging patients and families for patient safety. Lancet 2024 Mar 2;403(10429):791-793. [doi: 10.1016/S0140-6736(23)01908-6] [Medline: 37722399]
8. Sharma AE, Grumbach K. Engaging patients in primary care practice transformation: theory, evidence and practice. Fam Pract 2017 Jun 1;34(3):262-267. [doi: 10.1093/fampra/cmw128] [Medline: 28034916]

XSL•FO
RenderX

9. Wieczorek CC, Nowak P, Frampton SB, Pelikan JM. Strengthening patient and family engagement in healthcare - The New Haven Recommendations. Patient Educ Couns 2018 Aug;101(8):1508-1513. [doi: 10.1016/j.pec.2018.04.003] [Medline: 29661703]
10. Koch A, Brust L, Weigl M. Barriers and facilitators of patient engagement activities to improve patient safety in healthcare organizations: a Delphi-based expert survey. Risk Manag Healthc Policy 2024 Dec 19;17:3217-3226. [doi: 10.2147/RMHP.S489522] [Medline: 39717361]
11. Oldfield BJ, Harrison MA, Genao I, et al. Patient, family, and community advisory councils in health care and research: a systematic review. J Gen Intern Med 2019 Jul;34(7):1292-1303. [doi: 10.1007/s11606-018-4565-9] [Medline: 30051331]
12. Sharma A, Angel L, Bui Q. Patient advisory councils: giving patients a seat at the table. Fam Pract Manag 2015;22(4):22-27. [Medline: 26176505]
13. Anderson A, Getz KA. Insights and best practices for planning and implementing patient advisory boards. Ther Innov Regul Sci 2018 Jul;52(4):469-473. [doi: 10.1177/2168479017720475] [Medline: 29714530]
14. Patient and family advisory councils blueprint: a start-up map and strategy guide. American Hospital Association. 2022. URL: https://www.aha.org/system/files/media/file/2022/01/alliance-pfac-blueprint-2022.pdf [accessed 2025-10-31]
15. Liang L, Cako A, Urquhart R, et al. Patient engagement in hospital health service planning and improvement: a scoping review. BMJ Open 2018 Jan 30;8(1):e018263. [doi: 10.1136/bmjopen-2017-018263] [Medline: 29382676]
16. Lavallee DC, Blakeney EA, Yu Y, et al. Engaging patients and families to transform heart failure care. J Interprof Care 2020;34(6):835-838. [doi: 10.1080/13561820.2019.1696286] [Medline: 31829762]
17. Sharma AE, Knox M, Mleczko VL, Olayiwola JN. The impact of patient advisors on healthcare outcomes: a systematic review. BMC Health Serv Res 2017 Oct 23;17(1):693. [doi: 10.1186/s12913-017-2630-4] [Medline: 29058625]
18. Grenz-Farenholtz B, Langner D, Hohmann E, Verheyen F, Kranich C. Patientenvertretung auf der ebene der bundesländer – ergebnisse einer quantitativen befragung zu den bildungsbedürfnissen von patientenvertretern und patientenvertreterinnen [Article in German]. Gesundheitswesen 2014;76(11):735-741. [doi: 10.1055/s-0034-1366985]
19. Brust L, Blum Y, Weigl M. Promoting patient safety through patient engagement at the organisational level: a Delphi-based needs assessment among patient and family advisory councils. Health Expect 2025 Jun;28(3):e70319. [doi: 10.1111/hex.70319] [Medline: 40492414]
20. Duncan E, O'Cathain A, Rousseau N, et al. Guidance for reporting intervention development studies in health research (GUIDED): an evidence-based consensus study. BMJ Open 2020 Apr 8;10(4):e033516. [doi: 10.1136/bmjopen-2019-033516] [Medline: 32273313]
21. Hoffmann TC, Glasziou PP, Boutron I, et al. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. BMJ 2014 Mar 7;348:g1687. [doi: 10.1136/bmj.g1687] [Medline: 24609605]
22. Hoffmann TC, Walker MF. 'TIDieR-ing up' the reporting of interventions in stroke research: the importance of knowing what is in the 'black box'. Int J Stroke 2015 Jul;10(5):657-658. [doi: 10.1111/ijs.12524] [Medline: 26094670]
23. Using logic models to bring together planning, evaluation & action: logic model development guide. University of Arizona Prevention Research Center. 2000. URL: https://azprc.arizona.edu/sites/default/files/CHWtoolkit/PDFs/LOGICMOD/LOGICMOD.PDF [accessed 2025-10-31]
24. Goeschel CA, Weiss WM, Pronovost PJ. Using a logic model to design and evaluate quality and patient safety improvement programs. Int J Qual Health Care 2012 Aug;24(4):330-337. [doi: 10.1093/intqhc/mzs029] [Medline: 22745358]
25. The logframe handbook: a logical framework approach to project cycle management (English). : World Bank Group; 2005 URL: https://documents.worldbank.org/en/publication/documents-reports/documentdetail/783001468134383368/the-logframe-handbook-a-logical-framework-approach-to-project-cycle-management [accessed 2025-10-31]
26. Bethmann A, Hilgenböcker E, Wright M. Partizipative qualitätsentwicklung in der prävention und gesundheitsförderung [Book in German]. In: Tiemann M, Mohokum M, editors. Prävention und Gesundheitsförderung: Springer; 2021.
27. Clavel N, Pomey MP, Ghadiri DP. Partnering with patients in quality improvement: towards renewed practices for healthcare organization managers? BMC Health Serv Res 2019 Nov 8;19(1):815. [doi: 10.1186/s12913-019-4618-8] [Medline: 31703678]
28. Wahlberg KJ, Burnett M, Muthukrishnan P, et al. Partnering with patients in a quality improvement curriculum for internal medicine residents. J Patient Exp 2021 Mar 21;8:2374373521999604. [doi: 10.1177/2374373521999604] [Medline: 34179411]
29. Tsang VW, Chew SY, Junker AK. Facilitators and barriers to the training and maintenance of young persons' advisory groups (YPAGs). Int J Pediatr Adolesc Med 2020 Dec;7(4):166-173. [doi: 10.1016/j.ijpam.2019.10.002] [Medline: 33319014]
30. Martinez W, Browning D, Varrin P, Sarnoff Lee B, Bell SK. Increasing patient-clinician concordance about medical error disclosure through the patient TIPS model. J Patient Saf 2019 Dec;15(4):305-307. [doi: 10.1097/PTS.0000000000000284] [Medline: 28492422]
31. Haesebaert J, Samson I, Lee-Gosselin H, et al. "They heard our voice!" Patient engagement councils in community-based primary care practices: a participatory action research pilot study. Res Involv Engagem 2020 Sep 21;6:54. [doi: 10.1186/s40900-020-00232-3] [Medline: 32974050]
32. Hamilton health sciences homepage. Hamilton Health Sciences. URL: https://mylearning.hhsc.ca/course/index.php?categoryid=11 [accessed 2024-08-01]

XSL•FO
RenderX

33. Patient family advisor handbook. Royal Victoria Regional Health Centre. 2022. URL: https://www.rvh.on.ca/wp-content/uploads/2022/11/2022-Patient-and-Family-Advisor-Handbook-June-2022-new.pdf [accessed 2025-10-31]
34. Course offerings. Hamilton Health Sciences Centre for People Development. URL: https://centreforpeopledevelopment.ca/course-offerings/ [accessed 2024-08-01]
35. Engaging patients in patient safety – a Canadian guide. Healthcare Excellence Canada. 2017. URL: https://www.healthcareexcellence.ca/en/resources/engaging-patients-in-patient-safety-a-canadian-guide/ [accessed 2025-10-31]
36. Patient family advisory councils. : Colorado Hospital Association; 2019 URL: https://cha.com/wp-content/uploads/2019/04/CHA.158-PFAC-Toolkit_2019_final.pdf [accessed 2025-10-31]
37. Including the patient voice: a guide to engaging the public in programs and policy development. Betsy Lehman Center for Patient Safety. 2024. URL: https://betsylehmancenterma.gov/assets/uploads/general/IncludingThePatientVoice-Revised2024.pdf [accessed 2025-10-31]
38. Patient and family advisor orientation manual. Agency for Healthcare Research and Quality. 2013. URL: https://www.ahrq.gov/sites/default/files/wysiwyg/professionals/systems/hospital/engagingfamilies/strategy1/Strat1_Tool_9_AdvisorTrain_508.pdf [accessed 2014-06-15]
39. Körner M, Dinius J, Ernstmann N, et al. Effectiveness and feasibility of an interprofessional training program to improve patient safety-a cluster-randomized controlled pilot study. Front Psychol 2023 Nov 7;14:1186303. [doi: 10.3389/fpsyg.2023.1186303] [Medline: 38022945]
40. Patient Safety Rights Charter: World Health Organization; 2024.
41. PFACQS® 1.0 toolkit. The Healthcare and Patient Partnership Institute. URL: https://www.h2pi.org/pfac-tools [accessed 2025-01-13]
42. Strategically advancing patient and family advisory councils in New York State hospitals. : New York Health Foundation; 2018 URL: https://nyhealthfoundation.org/resource/strategically-advancing-patient-and-family-advisory-councils-in-new-york-state-hospitals/ [accessed 2025-10-31]
43. Wright J, Lawton R, O'Hara JK, et al. Improving patient safety through the involvement of patients: development and evaluation of novel interventions to engage patients in preventing patient safety incidents and protecting them against unintended harm. Programme Grants Appl Res 2016;4(15):1-296. [doi: 10.3310/pgfar04150]
44. Bloom B, Engelhart MD, Furst EJ, et al. Taxonomy of Educational Objectives, Handbook 1: Cognitive Domain: Addison-Wesley Longman Ltd; 1956.
45. Krathwohl DR. A revision of Bloom's Taxonomy: an overview. Theory Pract 2002;41(4):212-218. [doi: 10.1207/s15430421tip4104_2]
46. Nacioglu A. As a critical behavior to improve quality and patient safety in health care: speaking up!. Saf Health 2016;2:10. [doi: 10.1186/s40886-016-0021-x]
47. von Thun FS. Miteinander Reden 1: Störungen und Klärungen Allgemeine Psychologie der Kommunikation [Book in German]: Rowohlt E-Book; 2013.
48. Rosenberg MB, Griffin S, Carrington KL. Nonviolent communication: a language of compassion. In: Transforming Terror: University of California Press; 2019:304-311. [doi: 10.1525/9780520949454-111]
49. Snyder G, Daniels AC. Oops!: 13 Management Practices That Waste Time and Money (and What to Do Instead): Performance Management Publications; 2009.
50. Hoermann HJ. FOR-DEC: a prescriptive model FOR aeronautical decision making. Presented at: Proceedings of the WEAAP Conference; Mar 28-31, 1994 URL: https://elib.dlr.de/27044/ [accessed 2025-10-31]
51. Drucker PF. People and Performance: The Best of Peter Drucker on Management: Harvard Business School Press; 1977.
52. Merkle W, editor. Risk Management in Medicine: Springer-Verlag; 2015.
53. Fuzier R, Izard P, Petiot E, Jaulin F. Safety in healthcare: from the flight deck to the operating room. Turk J Anaesthesiol Reanim 2023 Aug 18;51(4):290-296. [doi: 10.4274/TJAR.2023.231264] [Medline: 37587655]
54. Chang A, Chou CL, Teherani A, Hauer KE. Clinical skills-related learning goals of senior medical students after performance feedback. Med Educ 2011 Sep;45(9):878-885. [doi: 10.1111/j.1365-2923.2011.04015.x] [Medline: 21848715]
55. Grant H, Dweck CS. Clarifying achievement goals and their impact. J Pers Soc Psychol 2003 Sep;85(3):541-553. [doi: 10.1037/0022-3514.85.3.541] [Medline: 14498789]
56. Webb TP, Merkley TR, Wade TJ, Simpson D, Yudkowsky R, Harris I. Assessing competency in practice-based learning: a foundation for milestones in learning portfolio entries. J Surg Educ 2014;71(4):472-479. [doi: 10.1016/j.jsurg.2014.01.019] [Medline: 24776864]
57. Li ST, Tancredi DJ, Co JP, West DC. Factors associated with successful self-directed learning using individualized learning plans during pediatric residency. Acad Pediatr 2010;10(2):124-130. [doi: 10.1016/j.acap.2009.12.007] [Medline: 20206911]
58. Partners for Patient Safety program, e-Learning (Module 1) [Web page in German]. Institute for Patient Safety, University Hospital Bonn. URL: https://www.ukbonn.de/tools/patientensicherheit-und-kommunikation-fur-patientenbeirate/ [accessed 2025-11-04]
59. Mccawley P. Logic model for program planning and evaluation. : University of Idaho Extension; 2000 URL: https://verso.uidaho.edu/esploro/outputs/report/Logic-Model-for-Program-Planning-and/996822658601851 [accessed 2025-10-31]

XSL•FO
RenderX

60. Adams NE. Bloom's taxonomy of cognitive learning objectives. J Med Libr Assoc 2015 Jul;103(3):152-153. [doi: 10.3163/1536-5050.103.3.010] [Medline: 26213509]
61. Möhler R, Bartoszek G, Köpke S, Meyer G. Proposed criteria for reporting the development and evaluation of complex interventions in healthcare (CReDECI): guideline development. Int J Nurs Stud 2012 Jan;49(1):40-46. [doi: 10.1016/j.ijnurstu.2011.08.003] [Medline: 21924424]
62. Reis T, Faria I, Serra H, Xavier M. Barriers and facilitators to implementing a continuing medical education intervention in a primary health care setting. BMC Health Serv Res 2022 May 13;22(1):638. [doi: 10.1186/s12913-022-08019-w] [Medline: 35562695]
63. Lewis B. Success of patient and family advisory councils: the importance of metrics. J Patient Exp 2023;10:23743735231167972. [doi: 10.1177/23743735231167972] [Medline: 37064819]
64. Chegini Z, Arab-Zozani M, Shariful Islam SM, Tobiano G, Abbasgholizadeh Rahimi S. Barriers and facilitators to patient engagement in patient safety from patients and healthcare professionals' perspectives: a systematic review and meta-synthesis. Nurs Forum 2021 Oct;56(4):938-949. [doi: 10.1111/nuf.12635] [Medline: 34339525]
65. Zdanski de Souza AD, Hoffmeister LV, Souto De Moura GM. Facilitators and barriers of patient involvement in hospital services: integrative review. Texto Contexto Enfermagem 2022;31(1). [doi: 10.1590/1980-265x-tce-2020-0395en]
66. Whitelaw S, Pellegrini DM, Mamas MA, Cowie M, Van Spall HG. Barriers and facilitators of the uptake of digital health technology in cardiovascular care: a systematic scoping review. Eur Heart J Digit Health 2021 Feb 4;2(1):62-74. [doi: 10.1093/ehjdh/ztab005] [Medline: 34048508]
67. Nielssen I, Santana M, Pokharel S, et al. Operationalizing the principles of patient engagement through a Patient Advisory Council: lessons and recommendations. Health Expect 2024 Feb;27(1):e13909. [doi: 10.1111/hex.13909] [Medline: 37942678]
68. Bombard Y, Baker GR, Orlando E, et al. Engaging patients to improve quality of care: a systematic review. Implement Sci 2018 Jul 26;13(1):98. [doi: 10.1186/s13012-018-0784-z]
69. Miller T, Reihlen M. Assessing the impact of patient-involvement healthcare strategies on patients, providers, and the healthcare system: a systematic review. Patient Educ Couns 2023 May;110:107652. [doi: 10.1016/j.pec.2023.107652] [Medline: 36804578]
70. Mesko B, deBronkart D, Dhunnoo P, Arvai N, Katonai G, Riggare S. The evolution of patient empowerment and its impact on health care's future. J Med Internet Res 2025 May 1;27:e60562. [doi: 10.2196/60562] [Medline: 40311140]
71. Ansmann L, Nöst S, Körner M, et al. Navigating the future of Organisational Health Services Research in Germany and beyond: a position paper. Gesundheitswesen 2024 Sep;86(S 04):S259-S266. [doi: 10.1055/a-2308-7384] [Medline: 39038484]

Abbreviations

FORDEC: facts, options, risks and benefits, decision, execution, and check
GUIDED: Guidance for the Reporting of Intervention Development
P4PS: Partners for Patient Safety
PFAC: patient and family advisory council
SMART: specific, measurable, attainable, realistic, and time-bound
TIDieR: Template for Intervention Description and Replication

Edited by A Stone; submitted 19.06.25; peer-reviewed by C Recsky, SA Scielzo; revised version received 30.09.25; accepted 30.09.25; published 24.11.25.

Please cite as:
Blum Y, Brust L, Weigl M, Rramani Dervishi Q
A Blended Educational Program to Promote Dialogue on Patient Safety Between Patient and Family Advisory Councils and Health Care Organizations: Codevelopment Study
JMIR Form Res 2025;9:e79286
URL: https://formative.jmir.org/2025/1/e79286
doi:10.2196/79286

© Yannick Blum, Larissa Brust, Matthias Weigl, Qëndresa Rramani Dervishi. Originally published in JMIR Formative Research (https://formative.jmir.org), 24.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Feasibility of Physical Therapy Evaluation Symptom Provocation Tests in Older Adults With Mild Traumatic Brain Injury: Mixed Methods Study

Carrie A Barrett[1*], PT, DPT, PhD; Mark G Goetting[2*], MD; Alessander Danna-dos-Santos[1*], PT, PhD

[1]Department of Physical Therapy, College of Health and Human Services, Western Michigan University, 1903 W. Michigan Ave., Kalamazoo, MI, United States

[2]Department of Pediatric and Adolescent Medicine, Western Michigan University Homer Stryker School of Medicine, Kalamazoo, MI, United States

[*]all authors contributed equally

Corresponding Author:
Carrie A Barrett, PT, DPT, PhD
Department of Physical Therapy, College of Health and Human Services, Western Michigan University, 1903 W. Michigan Ave., Kalamazoo, MI, United States

Abstract

Background: Early standard assessment protocols have decreased costs and better identified treatment strategies in individuals with mild traumatic brain injury (mTBI). Clinical practice guidelines contained strong recommendations for the use of provocation tests in evaluating younger adults (<65 y of age). Currently, the recommended protocols and literature regarding tolerance to self-reported and physical exertion outcome testing in older adults (>65 y of age) with mTBI have been limited. To start bridging some of these shortcomings and aid the development of practice guidelines for older adults, this study explored the feasibility of physical therapy evaluation protocols for individuals aged >65 years.

Objective: The study aimed to (1) assess the feasibility and tolerance of using evidence-based physical therapy evaluation outcome tests, and (2) apply the International Classification of Health, Disability, and Function (ICF) domains to an emergent thematic analysis of research protocols involving older adults with and without mTBI.

Methods: The feasibility study was a mixed methods design that included 13 community participants (>65 y) with and without mTBI. Investigators documented completion of the health form, participant-reported outcomes including the Patient-Specific Function Scale (PSFS) and Post Concussion Symptom Scale, and physical performance measures using the Motion Sensitivity Quotient (MSQ) and a submaximal recumbent stepper test, modified from the Buffalo Concussion Bike Test. Emergent contextual themes were identified within the study protocol, testing space, and participant responses.

Results: The sample included 13 participants (aged 65 - 91 y; 7 females, 6 males), 4 with mTBI and 9 without mTBI. All completed the health form, PSFS, and Post Concussion Symptom Scale, with moderate verbal cues required in 15% (PSFS) of the cases. The Motion Sensitivity Quotient and Buffalo Concussion Bike Test-Modified were completed to the participants' maximum safe effort. No participants experienced adverse mTBI symptoms. The categories for the theme of study protocol were time, communication, and logistics coordination. The categories for the theme of study environment were assessment, documentation, and safety. Unanticipated positive emotions were evoked with comments such as feedback, including participation was "fun" and "I didn't know I could do all this."

Conclusions: Older adults were able to tolerate the collection of participant-reported and physical outcome measures that allowed for individual maximal safe efforts without mTBI symptom provocation. The contextual themes using the ICF model of study protocol and environment were categorized and coded for future research consideration. Positive emotion participant responses were captured at the end of data collection. Clinical Relevance: This study applied standard physical therapy assessment protocols, which were safely tolerated by older adults with mTBI. The study used ICF domains from a research procedure perspective for protocol and environment considerations. The findings included participants' positive emoted responses, which may inform future large-sample trials in the evaluation of older adults with mTBI.

(JMIR Form Res 2025;9:e76799) doi:10.2196/76799

KEYWORDS

older adults; geriatric; mild traumatic brain injury; concussion; physical therapy evaluation; symptom provocation tests

XSL•FO
RenderX

Introduction

Researchers reported over 4.3 million individuals with traumatic brain injury (TBI) annually in the United States [1]. Of all TBI severities, experts considered 90% to be "mild" traumatic brain injury (mTBI) [2]. However, it is important to note that despite the "mild" designation, individuals with mTBI developed cognitive, physical, psychological, and social dysfunctions [3,4]. Of importance, persisting symptoms posed a higher risk to older adult populations (aged >65 y) [3-6]. Investigators have reported that 34% (all ages) have mTBI symptoms after 6 months, whereas over 44% of older adult populations continued to have symptoms beyond 6 months [3,6]. Researchers described ongoing mTBI symptoms to include headache, irritability or emotional lability, sleep disruption, drowsiness or fogginess, dizziness, and forgetfulness [7].

The aging population of older adults with mTBI has negatively contributed to the US health care burden [4,8,9]. Researchers reported the overall health care cost of those with mTBI was US $46 million (2017) [8]. Pavlov et al [9] found that health care costs were 2 times higher in older adults than in individuals in younger populations among those with mTBI. Older adults with mTBI reported having difficulty in self-care, mobility, and quality of life [4]. Due to the complex clinical scenario resulting from mTBI, researchers reported a pressing need for the development of clinical assessment tools that effectively assess the extent of symptoms and use intervention strategies [10,11].

Research has supported that early standard assessment or evaluation protocols may help decrease costs and better identify treatment strategies for individuals with mTBI [11,12]. Clinical practice guidelines for physical therapists contained strong recommendations for the use of test measures in those who have persistent symptoms after mTBI in younger populations [7,13]. The recommended tests included patient positional testing and aerobic testing, which mildly provoked mTBI symptoms [13]. These tests provided the necessary information to initiate rehabilitative treatment [13,14]. Unfortunately, the provocation testing in mTBI research failed to include findings on older adult populations [11,13]. The purpose of this study was to explore the tolerance to the recommended testing in a small group of older adults with and without mTBI. The study aims were (1) to determine the feasibility and tolerance in older adults in implementing the recommended outcome tests for mTBI, and (2) to describe an emergent thematic analysis during the testing of older adults with and without mTBI within the domains of the International Classification of Health, Disability, and Function (ICF) model. This study has detailed the methodology and participant responses to provide preliminary considerations in older adult research. Further development of physical therapy evaluation guidelines for these tests will guide clinicians in safe, evidence-informed assessment of the aging population with mTBI [7,11].

Methods

Recruitment

The feasibility study was a mixed methods design conducted at Western Michigan University (WMU) Laboratory for Advancements in Rehabilitation Sciences. The participants were recruited from local clinics, hospital emergency departments, senior living facilities, and senior centers. A total of 15 community members were interested in the study. Of the members not enrolled (n=2), 1 expressed severe physical limitations prohibiting travel to the study site, and 1 declined due to the data collection schedule. The investigators established the sample size based on the timeline of funding and successfully enrolled 13 eligible older adults (June 2022- June 2023).

The inclusion criteria included individuals aged >65 years, who had the cognitive means to answer health-related questionnaires and the physical ability to perform testing. The general cognitive and physical abilities were screened via verbal communication over the telephone during the study explanation. In this screening, the primary investigator (PI) explained the study protocol and asked participants if they had concerns about memory testing or exercising on a stationary stepper. The PI (CAB) assessed for verbal inconsistencies in language, memory, and processing throughout all interactions with participants. In addition, the PI (CAB) screened the cardiovascular system, the cervical spine and neck, and mobility status of each participant to determine medical clearance and safety before data collection. The exclusion criteria excluded ages younger than 65 years; screening abnormal for neck screening (ligamentous instability or vertebral artery insufficiency) [13]; or not having the physical ability to perform the aerobic testing [15].

The PI (CAB) collecting data was state-licensed, certified in advanced older adult cognitive screening and cardiopulmonary resuscitation (basic life support), and worked as a board-certified specialist in neurologic physical therapy for 17 years in direct access, providing services to patients with neurologic conditions.

Ethical Considerations

The WMU institutional review board (IRB) reviewed and approved this study under the expedited review with a waiver of written informed consent (application IRB-2022 - 205). The participants were assigned study numbers, and all data were collected on hard copy forms. The data were entered into the computer as deidentified data. The raw data were analyzed under the study numbers. The coinvestigators (AD-d-S, MGG, Lizzie Butkis, Ashley Helmer, and Mike Peres) analyzed all deidentified data. All participants verbalized and signed the written IRB-approved consent before enrollment. The study was grant-funded by the WMU Faculty Research and Creative Award (grant 2020). Through this funding, the participants received US $30 gift card incentive to participate in the study.

Study Design

Research supported that older adults with mTBI could complete participant-reported and health status forms [3,5]. Previous studies included individuals with mTBI experiencing symptom provocation when they were exposed to visual, cognitive, and movement stimulation [16,17]. Investigators (CAB, Ashley Helmer, Mike Peres, and Lizzie Butkis) created an a priori observational table to document the completion of each participant-reported form. There was limited research on a similar tool. The senior coinvestigators (AD-d-S and MGG)

XSL•FO
RenderX

peer-reviewed the observation table. After the collaboration, all investigators unanimously agreed to mark the completeness of each of the forms to be marked as "completed without or minimal assistance," "completed with moderate assistance," and "unable to complete: [reason]." Participants were categorized as "completed without or with minimal assistance" when they independently marked all form prompts without or with less than 50% cues provided by the investigator (verbal, visual, and physical) to complete missed form prompts. Participants were marked as "completed with moderate assistance" when they needed cues more than 50% and up to 100% of the time in filling out each form. Participants were considered "unable to complete: [reason]" when they were not able to complete the form, even with visual, verbal, or physical cues from the PI, and the reason reported. The "unable to complete" was anticipated to be used for participants who became overstimulated from significant mTBI symptoms due to cognitive or physical processing demands.

Statistical Analysis

Feasibility

The first focus of this study was to explore the participants' tolerance and ability to complete physical therapy evaluation testing in older adults with and without mTBI. The PI (CAB) was to monitor and record the participants' responses, tolerance, and vitals during the participant-reported and physical measurements. The participant-reported measures included a general health form, the Post Concussion Symptom Scale (PCSS) [18], and the Patient Specific Functional Scale (PSFS) [13,19,20]. The PI (CAB) collected data regarding participant abilities (including field notes on impairments of mobility, communication, and hearing) and responses to filling out the forms. The PI (CAB) recorded the physical measures of the Motion Sensitivity Quotient (MSQ) [21], and the Buffalo Concussion Bike Test-Modified (BCBT-M) [14,15,22] to determine mTBI sign or symptom provocation and safe participation. A recumbent stepper with adjustable arm handles, leg pedals, and a supportive seat was used as a modification to the Buffalo Bike Concussion Test to accommodate a variety of abilities in older adult participants for aerobic submaximal physical exertion testing [23,24]. Previous research has supported the validity and reliability of the test measures and parameters collected (Table 1). The PI (CAB) recorded observations on the participant's ability to safely complete each health medical form, PCSS, PSFS, MSQ, and BCBT-M; therefore, we used descriptive statistics for analysis.

XSL•FO
RenderX

Table . Data collection measurement summary feasibility study in older adults with mild traumatic brain injury.

Measure	ICF[a] domains	How measured	Variable type	Psychometrics of geriatric or mTBI[b] outcomes
Health form and screening	Body structure function - physical screen	Participant-reported and investigator	Binary (yes or no)	Clinical practice guidelines [13]
PCSS[c]	Body structure function - symptom limitation	Participant-reported	Numeric (0 - 132)	Responsive rate (effect size and standardized response mean>1.3) [18]
PSFS[d]	Activity participation	Participant-reported	Numeric (0 - 10)	Reliability and validity (92% completed, β=.477, *P*=.02) [19,20]
Completion of participant-reported forms (PCSS, PSFS)	Tolerance to cognitive exertion	Investigator	Categorical (completed without or minimal assist, completed with moderate assist, not completed [reason])	__[e]
MSQ[f]	Body structure function, activity - screen of vestibular, cervical spine, motion sensitivity, benign paroxysmal positional vertigo, and balance	Investigator	Numeric (0 - 20)	Reliability (ICC[g] 1.00 - 0.95) and validity (*r*=0.64, *P*<.001, Cronbach α=0.95) [21]
BCBT-M[h]	Body structure function - exertional tolerance or aerobic capacity	Investigator	Numeric (blood pressure, heart rate, relative perceived exertion, and symptom provocation intensity)	Reliability (95% CI <0.01-0.03, *P*=.03) and validity (*z*=−2.5, IQR 16-20; *P*=.013) [15,22,25]
Completion of physical performance (MSQ, BCBT-M)	Tolerance to movement or physical exertion	Investigator	Binary (yes or no)	—

[a]ICF: International Classification of Function, Disability, and Health.

[b]mTBI: mild traumatic brain injury.

[c]PCSS: Post Concussion Symptom Scale.

[d]PSFS: Patient Specific Functional Scale.

[e]Not applicable.

[f]MSQ: Motion Sensitivity Quotient.

[g]ICC: intraclass correlation coefficient.

[h]BCBT-M: Buffalo Concussion Bike Test-Modified.

Study Themes

The secondary focus of this study was to describe emergent themes [26], or concepts that investigators recorded during the study that could impact the data collection and findings in older adults with and without mTBI. In patient-care settings, physical therapists commonly use the World Health Organization–developed ICF model to capture a comprehensive understanding of the patient [27]. Conversely, researchers have not commonly used the ICF model for research procedures or setup. We determined that contextual domains of the ICF model (environment and personal), that commonly have been integrated into a clinical physical therapy appointment, were not captured in the data collection of this study. Therefore, we created a thematic analysis early in the study to explore similar contextual factors from a research perspective [26]. We defined the contextual factors as the interaction of the environment and study protocol among older adults with and without mTBI. After all data were collected and deidentified, we (CAB, Ashley Helmer, Mike Peres, and Lizzie Butkis) reviewed the data to derive categories and codes. We met 1-2 times a week via videoconference call or electronically, to develop common observations among all participants, and all investigators met to determine final agreement on terminology.

The 2 major themes were identified as the study protocol and study environment. We defined the study protocol as an outlined plan for conducting the research that included considerations of methodological processes, participant contextual details (participant ability and interaction during the task), and factors related to recording data that may be specific to older adult research. The study protocol categories included (1) time, or a unit of measure needed to complete a task or action; (2) communication, or an informational exchange that included an understanding; and (3) logistics, or coordination of person, place, and task. The study environment was defined as the physical space and the investigator's interaction with older adult participants before, during, and after the data collection. The study environment theme included (1) assessment, or the action of making a judgment of the participant's ability; (2) documentation, or accurate recording of information as evidence; and (3) safety, or mitigation of risk of harm or injury

XSL•FO
RenderX

(Table 2). The PI (CAB) documented the observations of the study protocol and test environment, and the coinvestigators (MGG, AD-d-S, Ashley Helmer, Mike Peres, and Lizzie Butkis) iteratively derived emergent themes, categories, codes, and examples based on deidentified written responses. The PI (CAB) performed member checking after each recorded participant's comment by repeating the comment back to the participant to ensure accuracy. The code of "unsolicited participant feedback post data collection" emerged as a single content area that the senior investigators agreed to further explore as important (MGG, AD-d-S, Ashley Helmer, Mike Peres, and Lizzie Butkis). The PI (CAB) recorded these unanticipated responses from participants as field notes. The PI (CAB) captured the quotes and repeated them back, or member checked, to ensure accuracy. The quotes were later analyzed (MGG, AD-d-S, Ashley Helmer, Mike Peres, and Lizzie Butkis) for contextual relevance. In previous studies, researchers described contextual relevance as a nuanced understanding, tailored to a specific population within a given event or circumstance [28].

XSL•FO
RenderX

Table . Emergent thematic analysis of data from a feasibility study in older adults with mild traumatic brain injury: study protocol and study environment (n=13).

Themes (categories)	Codes	Examples (subject number) (P[a] and or I[b]) impact
• Study protocol (time, communication, and logistics)	Participant contact, study explanation, and scheduling calls	• Telephone communication, "I'm having a hard time hearing you" (1,3) (P and I) • "I have another call and will be putting you on hold/call you back" (4,6,7) (P and I) • Participant leaving a message and their perception of not calling back is a returned call 2 d after the message (4,6,13) (P) • "Can you repeat what I'll be doing?" (4,13) (P and I) • "I'll need a ride". "because I can't drive" (1,3,13,10) (P)
• Study protocol (time, communication, and logistics)	Delivery of study instructions to participants for the day of the study	• "Can you mail them to me? I don't use e-mail" (5,6,13) (I)
• Study protocol (time, communication, and logistics)	After testing, participant recovery	• Assess for independent mobility, or cool down required for more than 10 min after testing (3,5,9,13) (I)
• Study protocol (time) • Study environment (documentation, assessment, and safety)	Physical assist vehicle to or from the data collection site	• Due to way-finding and participant requests, PI[c] met participants in the parking lot and walked them to the test site (1-12) (I) • Due to slight abnormal gait after testing, the participant was assisted to their vehicle (3,5,7,12,13) (I) • Poor straight path walking to and from the test room observed (5,7,12) (I) • After testing observed to have postexercise coughing upon leaving the test space (9,13) (P and I)
• Study environment (assessment) • Study protocol (time and logistics)	Individual testing, arriving with partner	• Couples signed up or needed to ride together (3,4,6,8,11,12) (P and I)
• Study protocol (time and communication)	Review of consent before data collection	• Time needs to adjust PI speech for response (speak louder, lower, or bias to the left or right ear), slow speech pace for the participant to understand, and provide verbal space for participants with a delayed response (9,10,12) (I)
• Study protocol (time and communication) • Study environment (documentation, assessment, and safety)	Data collection	• Social discussion by participants during (1-13) (I) • Repeat instructions, 1-word cues, and time for the participant to move and process requested task (9,10,12) (I) • Assessment of participant tolerance, safety (1-13) (I) • Skill in providing safety in the use of gait belt, assistive device, when to stop testing in RPE[d] and BP[e] or HR[f] critical ranges (1-13) (I)

XSL•FO
RenderX

Themes (categories)	Codes	Examples (subject number) (P[a] and or I[b]) impact
• Study environment (documentation, assessment, and safety)	Equipment use (stationary recumbent stepper, exam table, gait belt, BP stethoscope, cuff, and HR monitor)	• Aerobic equipment modification for UE[g] or LE[h] usage and position modifications (ability, joint safety, comfort) (1-13) (P and I) • Determine variant size BP cuff (10,13) (I) • Use of gait belt when observed compensations, slight veer, or reach for furniture in balance or mobility (5,7,10,12) (I) • Document safety considerations, modifications of equipment (1-13) (I) • Document responses to exertion while observing and assisting the participant (1-13) (I)
• Study protocol (time and logistics) • Study environment (assessment and safety)	Lab space (standard chair without wheels, fan, water, bathroom nearby)	• “I can’t sit in a roller chair - I’ll never get up” (3,12) (I) • Observed difficulty getting up from a chair (3,5,12) (I) • “It’s hot”, observed to need to cool off-fan (9,13) (I) • Request to use the bathroom during data collection (5,8,11) (I) • Request of bathroom close to the data collection room (2,5,8,9,12,13) (P) • Request of water for medication or thirst during data collection (1,4,9,12,13) (I)
• Study protocol (time and communication) • Study environment (documentation and assessment)	Unsolicited participant feedback post data collection	• Language included in the study was “fun” (1,6,7,11) (P) • “I didn’t know what to expect, and I did o.k.” (5,9) (P) • “I didn’t know I could do all this” (2) (P) • Language that requested participation in “the next study”. (5,6,7,12) (P) • “Is that it? I thought it would take longer!” (3) (P) • “I really liked this” (2) (P) • “I’m happy to help out” (7) (P) • Assess to walk unassisted, “I’m fine to walk to my car by myself, I feel really good” (13) (P)

[a]P: participant impact to consider during the study.

[b]I: investigator impact to consider for study set-up.

[c]PI: primary investigator.

[d]RPE: relative perceived exertion (0=best, 10=worst).

[e]BP: blood pressure.

[f]HR: heart rate.

[g]UE: upper extremity.

[h]LE: lower extremity.

Results

Feasibility of Completion

The total sample was 13 (n=9 without mTBI, n=4 with mTBI), including participants aged 65 - 91 years (n=7 female). The participants reported their race to be White (n=10) or chose not to fill out race (n=3). The participants fully completed the health form without or with minimal assistance (N=13, 100%). Most participants fully completed the PSFS (n=11, 85%) with minimal to no cues. The participants who completed the PSFS with moderate verbal cues (n=2, 15%; n=1 without mTBI, n=1 with mTBI) required verbal cues to define “activities” and redirection due to their extensive details of chosen activities. All the participants completed the PCSS (N=13, 100%) with minimal to no cues. The participants completed the MSQ and BCBT-M to their maximum safe effort. In the MSQ, 1 participant with mTBI tested positive for benign paroxysmal positional vertigo (BPPV) and needed 50% assistance to complete testing. A different participant without mTBI, in completing the MSQ, required 75% - 90% assistance for the standing and turning component due to a previous diagnosis of bilateral lower

XSL•FO
RenderX

extremity neuropathy. Most of the participants (n=10, 77%) were not able to fully complete the 10 minutes of the BCBT-M due to contraindicated cardiovascular pulmonary responses or reported joint discomfort. Of importance, no participants experienced significant adverse mTBI signs or symptoms from any of the testing [15].

Emergent Themes

The participants' and the investigators' perspectives were observed within the identified themes of study protocol and environment. For example, the investigators described the theme of the study protocol of time, as the time needed for the participant to communicate, transfer, walk, and write; and the time allowed by the investigator to allow space to complete the tasks (Table 2). Similarly, the investigators defined the theme for the study environment to include categories of assessment, documentation, and safety from each of the participants and the investigator's perspectives. For example, documentation influenced the participant's ability to complete participant-reported forms (denoted by P, participant impact) and the investigator's ability to record data while concurrently monitoring the participant (designated by I, investigator consideration). The codes under only the study protocol theme included "patient contact, explaining the study, scheduling calls, and study site visit" which related to telephone usage; "delivery of patient test day instructions" which revealed a need to communicate through general postal mailing as email was not an option; "explaining the consent slowly allowing space to ask questions," which increased the timeframe of data collection; and "after testing participant recovery," which required time, skilled assessment, and directions to guide the participant to safely cool down. The code under only the study environment theme was "equipment use." Various participant and investigator interactions and perspectives contributed to the overlapping of the categories, codes, and examples in both the study protocol and environment themes (Table 2). The codes that included both themes were physical assistance and testing, data collection, laboratory space, and participant feedback.

Contextual Relevance

The contextual relevance findings were further analyzed due to the unanticipated positive emotions evoked by the participants [28]. In the code "unsolicited participant feedback," participants stated comments such as, "that was fun," and "I didn't know I could do all this." In the analysis of the participants' quotes, we conceptualized the participants' perceptions of their roles and how they identified in society through the framework, the Role Theory of Active Aging (Figure 1) [29]. According to researchers who described the Role Theory of Active Aging, as individuals age, their perceived societal roles change, as does a potential loss of identity and a lack of understanding of their new roles [29,30]. The three main contributing categories that intersect within the Role Theory of Active Aging included (1) value or positive social role recognition, (2) environment or the space that allows for role fulfillment, and (3) behavior or interaction based on role expectation. Respectively, under these categories, this study applied the Role Theory of Active Aging concepts of (1) self-worth or a term for one's internal sense of being worthy of or good enough for a task, person, or situation; (2) belonging or one's affinity of inclusion, acceptance, or identity; and (3) participation or the action of taking part in an action or event (Figure 1). The concepts all overlapped, thereby contributing to active aging [29]. We adapted this framework to active aging through older adult participant research involvement.

Figure 1. Role theory of active aging in a feasibility study with corresponding older adult participant feedback (the Role Theory concept was adapted from the works of Ding and Ran [29] and Anglin et al [30]).

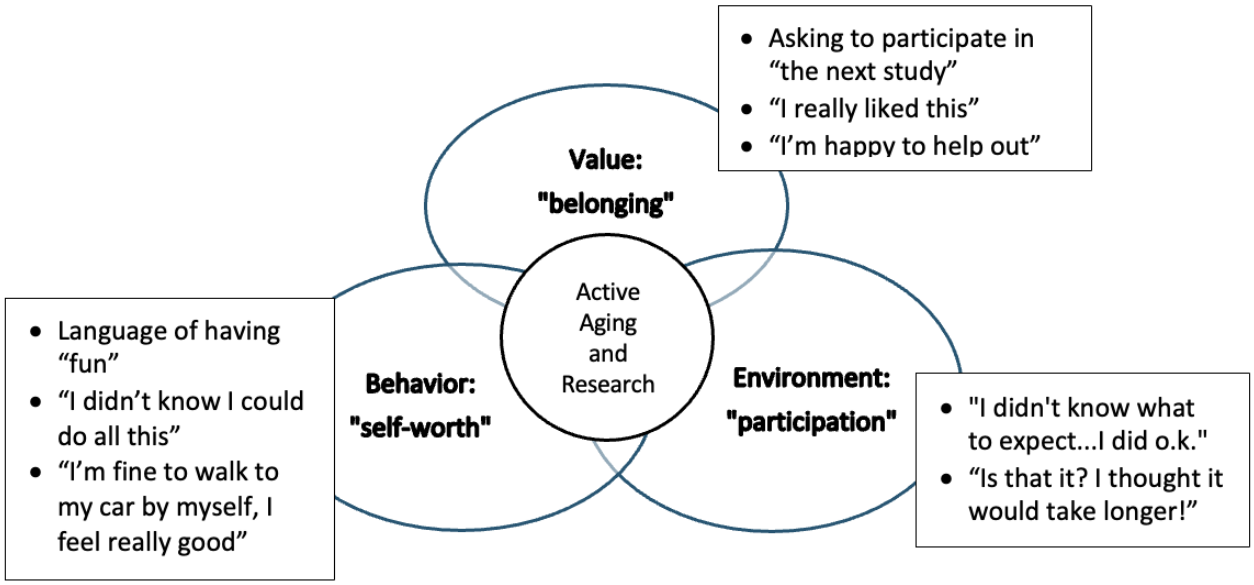

XSL•FO
RenderX

Discussion

Overview

Previous research has supported standard evaluation and assessment protocols for individuals aged <65 years with mTBI, including medical history interviews, self-reported outcomes, motion provocation screening, and submaximal aerobic testing [11,13,25]. The research supporting mTBI management in older adults has been limited [11]. This study provided early findings that the use of standard evaluation measures was safe in participants aged >65 years. This feasibility study included two aims: (1) to determine the feasibility of older adults' tolerance to the mTBI practice guidelines with a focus on participant-reported forms and symptom provocation tests, and (2) to describe an emergent thematic analysis contextualized within the ICF model relevant to research protocols and data collection in aging populations.

Principal Results

Tolerance to Provocation Testing

The first aim was to determine the feasibility of older adults tolerating participant-reported and physical outcomes of mTBI testing. These preliminary findings supported that older adults (aged 65 - 91 y) were able to complete the participant-reported outcomes of the medical health form (N=13, 100%), the PCSS (N=13, 100%), and the PSFS (n=11, 85%) without or with minimal assistance (Table 3). A small subset of participants required assistance to complete the PSFS (n=2, 15%). Assistance was related to defining the activity and providing redirection to remain on task (n=1 without mTBI, n=1 with mTBI). None of the participants reported post-mTBI symptoms from the task of completing these forms. These findings of completing the PSFS were consistent with previous findings in the use of the PSFS in older adult populations and individuals with acquired brain injuries [19,20]. In addition, these findings were also consistent with previous research in symptom-reported measures on individuals with complete and incomplete recovery after mTBI [31] and activity limitation subjective forms in individuals with acquired brain injury [19]. This study supported that older adults were able to complete standard forms that provided information about mTBI history, symptom intensity, and functional impact. The information from the forms has been crucial in guiding clinicians throughout the examination and development of early management strategies [13]. These preliminary findings additionally supported that older adults were able to complete the physical measures of MSQ (n=12, 92%) and the BCBT-M (n=13, 100%) without significant adverse effects. Quatman-Yates et al [13] reported that more research was needed to determine the outcome utility of motion-provoking symptom testing after mTBI. Researchers have supported that the MSQ outcome was reliable and valid in screening for symptom provocation related to the vestibular system, cervical spine, BPPV, motion sensitivity, and balance, which differentiated impairments and guided treatments [13,21]. This study's preliminary findings supported that the older adults with and without mTBI tolerated the MSQ testing and further identified 1 participant who was positive for BPPV. Our early findings in the BCBT-M were consistent with the research on the Buffalo Concussion Treadmill Test and the Buffalo Concussion Bike Test in the use of submaximal aerobic testing in individuals with mTBI; however, we modified the aerobic equipment to a recumbent stationary stepper to allow for the flexibility of appropriate fitting in this older adult population [23,24]. Importantly, all the participants tolerated the stationary recumbent stepper testing (n=13), completing the BCBT-M to their safest optimal level without provocation of mTBI symptoms. We anticipated the participants with mTBI to have exertional symptom provocation, and the results did not indicate this, possibly due to the small sample size. Most of the participants (n=10, 77%) discontinued the test before the standard 10-minute mark due to suboptimal cardiovascular responses or musculoskeletal discomfort during movement. These findings suggested the standard submaximal exertion test protocols designed for populations aged <65 years may not be the same for individuals >65 years with or without mTBI. Participants reported cardiovascular-altering medication usage during data collection. We are in the process of comparing participant medication use, early test termination due to poor cardiovascular responses, and participant perception of exertion to inform future research. Previous studies have described the need to develop BCBT-M protocols for older adults with and without mTBI, and although not generalizable, this study informed tolerance to BCBT-M in older adults [13,15]. The preliminary data in this study also indicated aerobic limitations attributable to musculoskeletal and deconditioning and did not indicate adverse or unsafe mTBI symptom responses. Similar exertional testing in individuals younger than 65 years with mTBI has been used to establish performance levels and to define aerobic exercise parameters that support brain health and accelerate healing after mTBI [13,15]. The development of age-appropriate parameters in older adults is essential to ensure safe and effective rehabilitation following mTBI.

XSL•FO
RenderX

Table . Results of feasibility in testing older adults with mild traumatic brain injury within 2-hour sessions (n=13).

Outcome	Participant-reported outcome measures				Physical mobility and exertion measures		
	Provoked mTBI[a] signs or symptoms, n (%)	Complete without or minimal assist, n (%)	Complete with assist, n (%)	Unable to complete, n (%)	Safely completed, no provoked mTBI signs or symptoms, n (%)	Completed safely, ended test due to other physical reason, n (%)	Unable to complete, n (%)
Health form	0 (0)	13 (100)	0 (0)	0 (0)	—[b]	—	—
PCSS[c]	0 (0)	13 (100)	0 (0)	0 (0)	—	—	—
PSFS[d]	0 (0)	11 (85)	2 (15)	0 (0)	—	—	—
MSQ[e]	—	—	—	—	12 (95)	1 (8)	0 (0)
BCBS-M[f]	—	—	—	—	13 (100)	10 (77)	0 (0)

[a]mTBI: mild traumatic brain injury.

[b]Not applicable.

[c]PCSS: Post Concussion Symptom Scale.

[d]PSFS: Patient Specific Functional Scale.

[e]MSQ: Motion Sensitivity Quotient.

[f]BCBT-M: Buffalo Concussion Bike Test-Modified.

Emergent Thematic Analysis

The second aim was an emergent thematic analysis focused on research procedures, participant interaction, physical space, and data collection with consideration to older adults with and without mTBI. We relied on the ICF contextual domains to support the research-related themes of the study protocol and study environment. The theme of the study protocol included categories of time, communication, and logistics. The theme of the study environment included categories of documentation, assessment, and safety. Some of the categories had various perspectives to consider, equally important in considerations derived from both the participant (P) and investigator (I) perspectives (Table 2). Examples of the codes under only the study protocol referred to communication modifications for individual participants and investigators skilled in participant safety assessment. Examples of the codes under the study environment related to skilled investigators who modified equipment to optimize size and fit and mitigate safety risks. These findings were important considerations for conducting future research. For example, investigators may need to consider the physical test environment, equipment adjustability, communication options, and monitored participant assessment (licensed health care providers as investigators) for safe and optimal research study procedures and data collection. Hume et al [17] reported that older adults with mTBI required specific recovery considerations and highlighted limitations in research involving this population [17]. The limitations included recruiting individuals aged >65 years and accounting for age-related psychological and biological differences. Elias et al [11] described that older adults have unique psychological, social, and biological characteristics that have not been developed in the mTBI research to effectively translate into postinjury mTBI management. Although the preliminary findings of this study required further development, the results provided insight into safe and accurate research considerations for older adults with and without mTBI.

This study supported unanticipated contextual relevance within the thematic analysis from the code of "unsolicited participant feedback." Because the feedback was unexpected, we had no established formal data collection methodology before testing. Surprisingly, the participants (n=10) organically stated positive comments after data collection (Table 2). We interpreted the participants' responses as novel reflections of their societal role within the framework of the Role Theory in Active Aging [29,30]. The participant feedback findings (Table 2) suggested that participating in research may have meaning to the participants (Figure 1). The participants emoted responses that reflected self-worth, a sense of belonging to a societal role, and successful participation in a social role [29]. These early findings needed further development to support the benefits of older adults participating in research and how this may impact their perceived role in active aging. By identifying and validating positive psychological and functional outcomes associated with research involvement, future studies can help promote broader inclusion of older adults in scientific inquiry. This, in turn, may enhance generalizability and empower older adults to view themselves as active contributors to knowledge generation and societal advancement, supporting healthier, more engaged aging trajectories [32].

Limitations

This study was a feasibility study in evaluating older adults with and without mTBI and therefore had inherent limitations. The sample size was small, the enrollment was a sample of convenience, and the shortened recruitment timeline was based on resource availability. Although we recruited from demographically different areas and worked with senior services across local facilities, the participants were not representative of the US population. We have discussed more rigorous

XSL•FO
RenderX

recruitment and enrollment strategies for improved representation in future research.

Another limitation was having 1 investigator collect data. Early in the study, the investigators established that to ensure participant safety, data collection must be performed by a licensed health care provider with experience in the assessment of older adults. The PI (CAB) addressed mitigating bias by member checking documented participant feedback and deidentifying all data before analysis. Future studies would include training specific to older adults with mTBI and extend data collection to multiple sites.

Diffusion in the study design was also a concern, as some of the participants arrived at the testing site in pairs. The PI (CAB) separated the participants during testing to maintain data collection integrity; however, this was verbalized by the participants as challenging. The participants who had more comorbidities relied heavily on their partners for accuracy in reporting timelines. The PI (CAB) encouraged participants to independently answer to their best ability. Future studies may involve acquiring electronic medical records to triangulate the accuracy of health form reporting. Despite this, the participants were able to complete the health forms without or with minimal cues (N=13, 100%), and this simulated individual participant reporting in a clinic appointment.

Conclusion

The evaluation of individuals for baseline assessment and initiation of treatment has been identified as crucial in reducing recovery times following mTBI [13-15]. The evaluation of older adults with mTBI, clinical guideline recommendations, and testing tolerance has been limited in the research [11,13]. This feasibility study was critical for further development of methodology and implementation of future large sample studies using provocation testing in older adults with mTBI. This study contributed to a deeper understanding of key contextual variables, including temporal constraints, safety protocols, communication strategies, and environmental settings, that shape research methodologies for evaluating older adults. The study also highlighted the potential benefits of research participation for older adults, particularly with societal expectations surrounding active aging. Further development and validation of these findings could guide physical therapists in delivering safe, evidence-informed care to the aging population with mTBI. Continued research is essential to establish standardized testing protocols that are tailored to the unique needs and vulnerabilities of older adults with mTBI [13].

Acknowledgments

We thank the College of Health and Human Services and the Doctor of Physical Therapy Program faculty and staff for your consistent support of this work. We extend a special thank you to Dr. Ashley Helmer, Dr. Lizzie Butkis, and Dr. Mike Peres for their early work in data analysis. We thank the participants as we appreciate their time and dedication to advancing research.

This study was funded for a 1 year period starting in June 2022, supported by Western Michigan University, under the Faculty Research and Creative Activities Award (#2020) for US $9800.

Data Availability

Deidentified data sets are not available currently due to ongoing research analysis.

Conflicts of Interest

None declared.

Checklist 1
CONSORT 2010 checklist.
[PDF File, 248 KB - formative_v9i1e76799_app1.pdf]

Checklist 1
Mixed Methods Reporting checklist.
[PDF File, 238 KB - formative_v9i1e76799_app2.pdf]

References

1. Loftin MC, Zynda AJ, Pollard-McGrandy A, Eke R, Covassin T, Wallace J. Racial differences in concussion diagnosis and mechanism of injury among adults presenting to emergency departments across the United States. Brain Inj 2023 Dec 6;37(12-14):1326-1333. [doi: 10.1080/02699052.2023.2248581] [Medline: 37607067]
2. Maas AIR, Menon DK, Manley GT, et al. Traumatic brain injury: progress and challenges in prevention, clinical care, and research. Lancet Neurol 2022 Nov;21(11):1004-1060. [doi: 10.1016/S1474-4422(22)00309-X] [Medline: 36183712]
3. Abdulle AE, de Koning ME, van der Horn HJ, et al. Early predictors for long-term functional outcome after mild traumatic brain injury in frail elderly patients. J Head Trauma Rehabil 2018;33(6):E59-E67. [doi: 10.1097/HTR.0000000000000368] [Medline: 29385011]

XSL•FO
RenderX

4. Karr JE, Iverson GL, Berghem K, Kotilainen AK, Terry DP, Luoto TM. Complicated mild traumatic brain injury in older adults: post-concussion symptoms and functional outcome at one week post injury. Brain Inj 2020;34(1):26-33. [doi: 10.1080/02699052.2019.1669825] [Medline: 31550173]
5. De Koning ME, Scheenen ME, Van Der Horn HJ, Spikman JM, Van Der Naalt J. From “miserable minority” to the “fortunate few”: the other end of the mild traumatic brain injury spectrum. Brain Inj 2018;32(5):540-543. [doi: 10.1080/02699052.2018.1431844] [Medline: 29388851]
6. Voormolen DC, Haagsma JA, Polinder S, et al. Post-concussion symptoms in complicated vs. uncomplicated mild traumatic brain injury patients at three and six months post-injury: results from the CENTER-TBI study. J Clin Med 2019 Nov 8;8(11):1921. [doi: 10.3390/jcm8111921] [Medline: 31717436]
7. Silverberg ND, Iaccarino MA, Panenka WJ, et al. Management of concussion and mild traumatic brain injury: a synthesis of practice guidelines. Arch Phys Med Rehabil 2020 Feb;101(2):382-393. [doi: 10.1016/j.apmr.2019.10.179] [Medline: 31654620]
8. Reid LD, Fingar KR. Inpatient stays and emergency department visits involving traumatic brain injury, 2017. In: Healthcare Cost and Utilization Project (HCUP) Statistical Briefs [Internet]: Agency for Healthcare Research and Quality (US). URL: https://www.ncbi.nlm.nih.gov/books/NBK556732/ [accessed 2025-10-20] [Medline: 32379407]
9. Pavlov V, Thompson-Leduc P, Zimmer L, et al. Mild traumatic brain injury in the United States: demographics, brain imaging procedures, health-care utilization and costs. Brain Inj 2019;33(9):1151-1157. [doi: 10.1080/02699052.2019.1629022] [Medline: 31241427]
10. Haarbauer-Krupa J, Pugh MJ, Prager EM, Harmon N, Wolfe J, Yaffe K. Epidemiology of chronic effects of traumatic brain injury. J Neurotrauma 2021 Dec;38(23):3235-3247. [doi: 10.1089/neu.2021.0062] [Medline: 33947273]
11. Elias J, Sutherland E, Kennedy E. Concussion management in older people: a scoping review. J Head Trauma Rehabil 2024;39(4):293-303. [doi: 10.1097/HTR.0000000000000933] [Medline: 38453625]
12. Marshall S, Bayley M, McCullagh S, et al. Updated clinical practice guidelines for concussion/mild traumatic brain injury and persistent symptoms. Brain Inj 2015;29(6):688-700. [doi: 10.3109/02699052.2015.1004755] [Medline: 25871303]
13. Quatman-Yates CC, Hunter-Giordano A, Shimamura KK, et al. Physical therapy evaluation and treatment after concussion/mild traumatic brain injury. Journal of Orthopaedic & Sports Physical Therapy 2020 Apr;50(4):CPG1-CPG73. [doi: 10.2519/jospt.2020.0301]
14. Leddy JJ, Haider MN, Ellis MJ, et al. Early subthreshold aerobic exercise for sport-related concussion: a randomized clinical trial. JAMA Pediatr 2019 Apr 1;173(4):319-325. [doi: 10.1001/jamapediatrics.2018.4397] [Medline: 30715132]
15. Haider MN, Johnson SL, Mannix R, et al. The Buffalo Concussion Bike Test for concussion assessment in adolescents. Sports Health 2019;11(6):492-497. [doi: 10.1177/1941738119870189] [Medline: 31486715]
16. Tyler CW, Likova LT. Brain trauma impacts retinal processing: photoreceptor pathway interactions in traumatic light sensitivity. Doc Ophthalmol 2022 Jun;144(3):179-190. [doi: 10.1007/s10633-022-09871-1] [Medline: 35445376]
17. Hume C, Mitra B, Wright B, Kinsella GJ. Cognitive performance in older people after mild traumatic brain injury: trauma effects and other risk factors. J Int Neuropsychol Soc 2023 Aug;29(7):651-661 [FREE Full text] [doi: 10.1017/S1355617722000674] [Medline: 36102332]
18. Langevin P, Frémont P, Fait P, Roy JS. Responsiveness of the post-concussion symptom scale to monitor clinical recovery after concussion or mild traumatic brain injury. Orthop J Sports Med 2022 Oct;10(10):23259671221127049. [doi: 10.1177/23259671221127049] [Medline: 36250029]
19. Evensen J, Soberg HL, Sveen U, Hestad KA, Bronken BA. The applicability of the Patient-Specific Functional Scale (PSFS) in rehabilitation for patients with acquired brain injury (ABI) - a cohort study. J Multidiscip Healthc 2020;13:1121-1132. [doi: 10.2147/JMDH.S259151] [Medline: 33116558]
20. Heldmann P, Hummel S, Bauknecht L, Bauer JM, Werner C. Construct validity, test-retest reliability, sensitivity to change, and feasibility of the patient-specific functional scale in acutely hospitalized older patients with and without cognitive impairment. J Geriatr Phys Ther 2022;45(3):134-144. [doi: 10.1519/JPT.0000000000000303] [Medline: 33734156]
21. Heusel-Gillig L, Santucci V, Hall CD. Development and validation of the modified motion sensitivity test. Otol Neurotol 2022 Sep 1;43(8):944-949. [doi: 10.1097/MAO.0000000000003641] [Medline: 35970158]
22. Guadagni V, Drogos LL, Tyndall AV, et al. Aerobic exercise improves cognition and cerebrovascular regulation in older adults. Neurology (ECronicon) 2020 May 26;94(21):e2245-e2257. [doi: 10.1212/WNL.0000000000009478] [Medline: 32404355]
23. Fiems CL, Bales M, Ganley S, Michel D, Rice K, Silverman K. Reliability of a recumbent stepper submaximal exercise test after traumatic brain injury. Physiother Theory Pract 2022 Nov;38(13):2896-2904. [doi: 10.1080/09593985.2021.1934923] [Medline: 34112064]
24. Shirazi SY, Huang HJ. Older adults use fewer muscles to overcome perturbations during a seated locomotor task. J Neurophysiol 2024 Jun 1;131(6):1250-1259. [doi: 10.1152/jn.00263.2023] [Medline: 38717333]
25. Graham RF, van Rassel CR, Burma JS, et al. Concurrent validity of a stationary cycling test and the Buffalo Concussion Treadmill Test in adults with concussion. J Athl Train 2021 Dec 1;56(12):1292-1299. [doi: 10.4085/1062-6050-0003.21] [Medline: 34911073]

XSL•FO
RenderX

26. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
27. Leonardi M, Lee H, Kostanjsek N, et al. 20 years of ICF-International Classification of Functioning, Disability and Health: uses and applications around the world. Int J Environ Res Public Health 2022 Sep 8;19(18):11321. [doi: 10.3390/ijerph191811321] [Medline: 36141593]
28. Lim WM. What is qualitative research? An overview and guidelines. Australasian Marketing Journal 2025 May;33(2):199-229. [doi: 10.1177/14413582241264619]
29. Ding X, Ran M. Research on the application of role theory in active aging education service system design. In: HCI International 2021 - Late Breaking Papers: Cognition, Inclusion, Learning, and Culture: Springer; 2021:205-222. [doi: 10.1007/978-3-030-90328-2_13]
30. Anglin AH, Kincaid PA, Short JC, Allen DG. Role theory perspectives: past, present, and future applications of role theories in management research. J Manage 2022 Jul;48(6):1469-1502. [doi: 10.1177/01492063221081442]
31. Bittencourt M, Balart-Sánchez SA, Maurits NM, van der Naalt J. Self-reported complaints as prognostic markers for outcome after mild traumatic brain injury in elderly: a machine learning approach. Front Neurol 2021;12:751539. [doi: 10.3389/fneur.2021.751539] [Medline: 34925214]
32. Dogra S, Dunstan DW, Sugiyama T, Stathi A, Gardiner PA, Owen N. Active aging and public health: evidence, implications, and opportunities. Annu Rev Public Health 2022 Apr 5;43(1):439-459. [doi: 10.1146/annurev-publhealth-052620-091107] [Medline: 34910580]

Abbreviations

BCBT-M: Buffalo Concussion Bike Test-Modified
BPPV: benign paroxysmal positional vertigo
ICF: International Classification of Function, Disability, and Health
IRB: institutional review board
MSQ: Motion Sensitivity Quotient
mTBI: mild traumatic brain injury
PCSS: Post Concussion Symptom Scale
PI: primary investigator
PSFS: Patient-Specific Function Scale
TBI: traumatic brain injury
WMU: Western Michigan University

Edited by A Mavragani; submitted 30.04.25; peer-reviewed by D Lawson, P Geels; revised version received 11.09.25; accepted 12.09.25; published 23.10.25.

Please cite as:
Barrett CA, Goetting MG, Danna-dos-Santos A
Feasibility of Physical Therapy Evaluation Symptom Provocation Tests in Older Adults With Mild Traumatic Brain Injury: Mixed Methods Study
JMIR Form Res 2025;9:e76799
URL: https://formative.jmir.org/2025/1/e76799
doi:10.2196/76799

© Carrie A Barrett, Mark G Goetting, Alessander Danna-dos-Santos. Originally published in JMIR Formative Research (https://formative.jmir.org), 23.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Predictors for Acceptance of Sexual Aggression Myths Among People Using Cyberporn: Cross-Sectional Study

Farah Ben Brahim[1,2*], MSci; Germano Vera Cruz[3*], Prof Dr; Robert Courtois[2], Prof Dr Med; Yasser Khazaal[1], Prof Dr Med

[1]Addiction Medicine, Faculty of Biology and Medicine, University of Lausanne, Rue du Bugnon 21, Lausanne, Switzerland

[2]QualiPsy UR 1901, Faculty of Arts et Human Sciences, University of Tours, Tours, France

[3]Department of Psychology, UR 7273 CRP-CPO, UFR Human, Social Sciences and Philosophy, University of Picardie Jules Verne, Amiens, France

[*]these authors contributed equally

Corresponding Author:
Farah Ben Brahim, MSci
Addiction Medicine, Faculty of Biology and Medicine, University of Lausanne, Rue du Bugnon 21, Lausanne, Switzerland

Abstract

Background: Although the acceptance of sexual aggression myths (ASAM) has been intensively studied over the past 30 years, there is still a lack of conclusive information on its relationships with a given set of cognitive and behavioral variables.

Objective: This study aimed to examine to what extent sexual coercion, compulsive cyberporn use (CCU), arousing cyberporn scenes, moral incongruence, impulsivity, and sexual self-esteem can predict the ASAM—assessed with the Acceptance of Modern Myths About Sexual Aggression (AMMSA) score—among people who use cyberporn.

Methods: Overall, 1557 English speakers who used cyberporn at least once during the previous 6 months (mean age 33.3, SD 10.9 years; men: n=1000, 64.2%; women: n=557, 35.8%) were included in this study after completing an online questionnaire. The questionnaire included measures of the AMMSA, sexual coercion experiences including perpetration and victimization, CCU, arousing cyberporn scenes, moral incongruence, impulsivity, and sexual self-esteem. Data analytics included descriptive statistics, Pearson correlation, analysis of variance, and multivariate linear regression.

Results: Male participants had significantly higher scores than their female counterparts regarding the following variables: AMMSA (t_{1551}=12.46, P<.001), sexual violence perpetration (t_{1551}=3.10, P<.001), CCU (t_{1551}=8.05, P=.003), aroused by "groups with several females" porn scenes (t_{1551}=8.71, P=.002), and age (t_{1551}=7.11, P<.001). Female participants had significantly greater mean scores than their male counterparts regarding the following variables: sexual violence victimization (t_{1551}=−11, P<.001), aroused by "submission" pornographic scenes (t_{1551}=−5, P<.001), aroused by "soft" pornographic scenes (t_{1551}=−2.40, P=.03), and aroused by "groups with several males" pornographic scenes (t_{1551}=−5.93, P<.001). Among the male participants, AMMSA scores were significantly predicted by scores of sexual coercion (both perpetration and victimization; β=.042, P<.001 and β=.023, P=.03, respectively), CCU (β=.339, P<.001), arousing cyberporn scenes displaying humiliation and groups of males (β=.136, P=.003), sexual self-esteem (β=.164, P<.001), and moral incongruence (β=.131, P<.001). Among the female participants, AMMSA scores were significantly predicted by the scores of sexual coercion perpetration (β=.060, P<.001), CCU (β=.287, P<.001), arousing cyberporn scenes with several females (β=−.111, P=.02), sexual self-esteem (β=.188, P<.001), and age (β=.025, P<.001). Age emerged as a distinct factor in women, with older participants more inclined to accept sexual aggression myths.

Conclusions: This study identified key psychological and behavioral correlates of the ASAM across genders. CCU, sexual coercion perpetration, and sexual self-esteem predicted the ASAM. Impulsivity and moral incongruence were salient among men, and age was salient among women. Gender differences emerged in arousal of pornographic content.

(JMIR Form Res 2025;9:e75485) doi:10.2196/75485

KEYWORDS

compulsive cyberporn; cyberporn; multivariate regression; sexual aggression myths; sexual coercion

Introduction

Sexual aggression myths [1], including rape myths [2], are biased, stereotyped, or false beliefs about sexual aggression, survivors [3], and sexual offenders that create an adverse atmosphere for survivors, including justifications for coercion, victim blaming, and gendered power dynamics [4,5]. Those beliefs affect how people interact with survivors, especially whether they believe them or blame them for their abuse [1,6]. Such beliefs are strong barriers to the dissemination of asking for sexual consent [7] and to the prevention of all kinds of sexual

XSL•FO
RenderX

aggression [8]. Among others, deeply rooted sex, gender, and social stereotypes are associated with the acceptance of sexual aggression myths (ASAM) [2,4,5,9], especially among men [10]. A questionnaire related to the Acceptance of Modern Myths About Sexual Aggression (AMMSA) [1] conceptualizes such myths about sexual aggression, including subtle and still "socially acceptable" beliefs that downplay or justify sexual violence, not limited to rape in the legal sense. These myths are part of broader cultural dynamics and media narratives, which warrant further exploration.

In addition to distorting the reality of sexual violence, these myths reinforce negative cultural stereotypes [3] that uphold male sexual dominance and minimize or excuse acts of aggression. Payne et al's [3] approach draws attention to the common social beliefs that support the persistence and propagation of these myths, emphasizing that the ASAM is part of a broader cultural narrative. Such narratives are partly vehicled by media, including some porn content [11]. Given these societal impacts, researchers have explored how media, particularly pornography, are associated with the ASAM. In this context, the relationship between sexual aggression myths and porn use has been particularly examined.

For instance, an earlier meta-analysis concluded that pornography is associated with the ASAM in experimental research but not in nonexperimental studies [12]. A more recent meta-analysis by Hedrick [13] showed that porn, especially porn involving violent content, is positively associated with the ASAM. Borgogna et al [14] indicated, however, that hostile sexism positively predicts the ASAM regardless of gender or porn use. Possibly, the conflicting results arise from cultural differences, differences in porn content, and violence-related porn materials classifications [15], as well as possible individual differences, including variations in the way people are using porn or cyberporn (internet porn use) [16-19] and in compulsive cyberporn use (CCU) [16,17]; a behavioral marker of problematic engagement in porn—rather than frequency alone [20,21]—may better account for these inconsistencies. Of note, CCU is characterized by an increase in pornography use, accompanied by a loss of control and associated negative consequences [22-26]. However, studies directly examining the link between CCU and ASAM are still limited (eg, [16]). This highlights the need to examine how individual differences intersect with both CCU and the ASAM.

Individual differences among variables previously linked to CCU—such as pornography-related moral incongruence [27], impulsivity [28,29], and sexual self-esteem [16]—may help explain some possible relationship with the ASAM. One such factor, moral incongruence, refers to the psychological distress that arises when pornography use conflicts with an individual's internalized moral values [27,30]. To reduce this tension, some may adopt rationalizations or belief systems—such as the ASAM—that serve to legitimize or minimize the moral implications of their behavior. As such, porn-related moral incongruence may not only be a correlate of CCU but also a psychological mechanism contributing to the endorsement of the ASAM. Impulsivity, a neuropsychological trait marked by a tendency for spontaneous actions without careful consideration, has been positively associated with the ASAM [31], suggesting that individuals with higher impulsivity may be more inclined to rationalize or justify sexual violence due to limited self-control and reflection on their actions. Similarly, sexual self-esteem, which has shown a modest yet positive association with sexual coercion [16,32], refers to the individual's perception of their own sexual value, confidence, and competence in sexual contexts [33]. It has been associated with sexual satisfaction, relational dynamics, and vulnerability to coercion-related beliefs and behaviors [34,35], making it relevant in the study of the ASAM. The link between sexual self-esteem and sexual coercion may be particularly relevant for people with cyberporn use, as pornography could shape or reinforce perceptions of sexual worth and attitudes toward sexual behaviors [6].

In addition to these individual factors, past personal experiences may interfere with beliefs about sexual aggression. Thus, personal experiences related to sexual coercion [36-38] may shape possible associations between CCU and the ASAM, given the documented links between prior sexual coercion experiences and the ASAM [39].

Another factor that may help to explain these associations is the type of pornographic content that arouses users. This neglected dimension that may contribute to such an association between porn use and the ASAM is the arousal that people may have in relation to specific porn contents [40,41] (ie, differences may exist between people reporting arousal from humiliation scenes and those reporting arousal from soft ones). Such porn scripts may influence, or reciprocally interact, with people's beliefs about gender roles and sexual conduct, especially when they portray themes of control, dominance, or coercion [42-44]. By normalizing aggressive sexual relations and minimizing the value of consent, these scripts have the potential to reinforce the ASAM [42-44].

Bringing these threads together, this study aimed to explore the relationships between the above-discussed variables and the ASAM. Specifically, it aimed at exploring, by gender group (male vs female), the relationships between a set of variables related to the participants' sexual coercion experiences, CCU, arousing cyberporn scenes, moral incongruence, impulsivity, sexual self-esteem, and ASAM—assessed with the AMMSA score—among people who use cyberporn.

Methods

Design of the Study

In order to achieve the aim, the study assessed a relatively large sample of participants (when compared with previous studies, eg, [2,36-38]) who used cyberporn at least once during the previous 6 months.

No precise hypotheses were formulated before data collection, as the investigation was exploratory and data-driven. Exploratory research permits adaptation and exploration of new pathways or new data analysis methodologies as the study progresses, in contrast to structured research with preconceived assumptions. This approach encourages the use of different analytical methods for extracting information from data without being constrained by particular ideas or by following the path

XSL•FO
RenderX

carved by previous research [45]. This has the benefit of opening up novel opportunities for investigation and avoiding bias or errors from earlier research or established theories [45].

Participants

Participants were recruited through Prolific, a crowdsourcing platform widely used for academic research [46,47]. The study was conducted entirely online using the SphinxOnline platform. Eligible participants were English-speaking adults (aged 18 y or older) who reported having viewed pornography at least once during the past 6 months. After accessing the survey link, participants were presented with a detailed information sheet and eligibility questions. Only those who met the inclusion criteria were granted access to the full questionnaire.

Materials

Criterion Variables

Acceptance of Sexual Aggression Myths

Gerger et al's [1] 30-item AMMSA was shortened [48] to measure the ASAM (Multimedia Appendix 1; the measures are available for consultation if needed). This 11-item (7-point scale from "I totally disagree" to "I totally agree") assesses participants' tolerance for the ASAM and female sexual aggression (eg, "When a woman starts a relationship with a man, she should be aware that the man will assert his right to have sex"; "Many women tend to exaggerate the problem of male violence"). Each participant has an overall rating. Higher scores indicate greater acceptance of myths. The Cronbach α scale was 0.94 in this study.

This questionnaire has been validated and implemented in numerous countries; it is, in fact, one of the most frequently employed measures in this field [1,48,49].

Predictor Variables

Demographic Data

Sex, age, sexual orientation, and relationship status were the demographic details provided by the participants.

Violent and Coercive Sexuality

The Sexual Experience Survey (SES) assessed sexual coercion experiences of perpetration and victimization in youths over the age of 14 years [50,51]. Victimization (eg, touching, kissing, rape) was determined by the 11-item victimization form. Perpetration (eg, touching, kissing, rape) was determined by the 11-item perpetration form. We determined the total perpetration and victimization scores of each participant for this research. The subscales' α values were 0.91 and 0.93, respectively.

Compulsive Cyberporn Use

We used the 8-item short form of the Compulsive Internet Use Scale (CIUS) to measure CCU [52,53]. Previous research [54,55] adapted the CIUS for cybersex, while this study adapted the CIUS for cyberporn, following previous methods [16,17], with the word "Internet" signifying pornographic websites. This unidimensional short form is internally consistent [52]. A 5-point scale (from "Never" to "Very often") tested CCU, with higher values indicating higher CCU. In this study, this scale's α was 0.90.

Arousing Cyberporn Scenes

We assessed 5 cyberporn scenes by asking about the arousal level toward these scenes (submission, humiliation, soft porn, groups with several males, and groups with several females), using the following instruction: "The following pornographic scenes are:". We used a 4-point scale ranging from "Very arousing" to "Not arousing at all."

Moral Incongruence

We relied on Grubbs et al's [27] item "I believe that pornography use is morally wrong" to assess pornography's negative moral perception with a 7-point scale ("Strongly disagree" to "Strongly agree").

Impulsivity

It was assessed using the Short Urgency, Premeditation, Perseverance, Sensation Seeking, Positive Urgency (UPPS-P) Impulsive Behavior Scale [56,57]. Specifically, the two features most frequently associated with addictive behaviors, positive and negative urgency (4-point response scale from "Agree strongly" to "Disagree strongly") [58-60], were assessed using only 8 of the measure's 20 items; these 8 items are related to two dimensions: positive urgency (measures the tendency to act impulsively due to positive affect; eg, "When I'm really excited, I tend not to think about the consequences of my actions") and negative urgency (measures the tendency to act impulsively due to negative affect; eg, "When I am upset, I often act without thinking"). Two scores were awarded to each participant. Before calculating the two scores, the scale items were reverse-coded. Impulsivity increased with higher scores. The subscales' α values were 0.79 and 0.82 for positive and negative urgency, respectively.

Sexual Self-Esteem

Based on the self-esteem single-item scale [61], we rated the sexual self-esteem of the participants with the following derived item: "I have high sexual self-esteem" on a 4-point scale from "Not very true of me" to "Very true of me."

Ethical Considerations

The study relied on an anonymous SphinxOnline survey. The Research Ethics Committee of Tours-Poitiers in France (2020-04-05) approved the study. All procedures were conducted in accordance with the Declaration of Helsinki and relevant institutional guidelines. Participants provided informed consent online before participation, after being fully informed about the study's aims, procedures, anonymity, and their right to withdraw at any time without consequences. No personally identifying data were collected. Data were collected, stored, and analyzed anonymously. Participants were recruited through Prolific, a platform widely used for ethical participant recruitment in academic research [46,47]. Inclusion criteria included being at least 18 years old and having watched pornographic material at least once in the past 6 months. The survey took approximately 15 minutes to complete. The participants received time compensation for their study participation via the Prolific platform (approximately £10

XSL•FO
RenderX

[approximately US $14] per hour of participation). This study was not registered on a public trial registry; however, the protocol was approved in advance by the ethics committee, and no deviations from the original procedures occurred. No adverse events were reported during data collection.

Reporting

This study follows the recommendations of the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines for reporting observational research [62]. Although STROBE was initially developed for epidemiological studies, it is broadly applicable to cross-sectional observational designs in psychology and social sciences. We ensured transparency and completeness in reporting, in line with the EQUATOR Network recommendations [62].

Statistical Analysis

It must be noted that the data had 0.2%-0.7% of missing values in the sociodemographic variables.

First, we conducted descriptive analyses on the study's predictor and criterion variables. These analyses were conducted separately on male versus female participants.

Second, we built a correlation matrix including all the study variables.

Third, we tested the normality of the dependent variable (AMMSA score) using the Shapiro-Wilk test for normality. The test was significant (W=0.878, df=1557; P<.001), indicating that the distribution of this variable was significantly different from a normal distribution [63]. In addition, we conducted multicollinearity tests on the independent variables. The results indicated no significant multicollinearity.

Fourth, and given that the assumption of normality had not been met, we built two robust multivariate linear regression models (one per gender group). These linear regressions used the above predictor variables (except for sexual orientation and relationship status) to predict AMMSA scores, the study's criterion variable. Robust linear regression can be used to improve the accuracy of predictions when the residuals are nonnormal and when the data contain outliers [64].

Finally, we conducted an ANOVA to test the effect of sociodemographic variables (sexual orientation and relationship status) on the AMMSA score.

All statistical analyses were conducted using the software SPSS Statistics version 29 (IBM Corp), except for the regression analyses, which were conducted using the R package "robustbase" [65]. These analyses were conducted with different statistical software that have different ways of dealing with missing values (eg, SPSS allows analysis considering only the participants without any missing values in each variable; on the contrary, the algorithm used for regression requires missing values imputation using the nearest neighbor approach).

Results

Descriptive Statistics and Inferential Comparison

Overall, 1590 English-speaking individuals fully completed the online questionnaire. However, 1557 participants were included in this study, among them 64.2% (n=1000) were men, and 35.8% (n=557) were women. The age of participants ranged from 18 to 75 (mean 33.3, SD 10.9) years. During the study period, 497 (31.9%) participants were single; 663 (42.6%) were in a relationship, not married; 389 (25%) were in a relationship, married; and 8 (0.5%) were widows. The participants' sexual orientation distribution was as follows: heterosexual 1228 (78.9%), homosexual 98 (6.3%), bisexual 200 (12.8%), and other 31 (2%). It must be noted that, given this study's objective, outlined above, 33 participants were excluded; among them, 27 were nonbinary, and 6 did not indicate their gender. Participants' countries of residence can be found in Multimedia Appendix 2.

Table 1 presents the descriptive statistics of all variables by gender (male vs female participants) and the respective mean comparison.

XSL•FO
RenderX

Table . Descriptive statistics (by gender: male n=998 vs female n=555) of the continuous and the ordinal variables included in this study design to predict the acceptance of sexual aggression myths in a multicountry sample.

Variable categories/variables	Scale/range	Mean male	SD male	Mean female	SD female	*t* test (*df*)	*P* value
Criterion variable							
AMMSA[a] acceptance of sexual aggression myths (total score)	1 - 7	3.08	1.32	2.25	1.20	12.46 (1551)	<.001[b]
Predictor variables							
Violent and coercive sexuality (2 variables)							
SES-P[c] (total score)	0 - 77	1.63	6.68	.81	3.78	3.10 (1551)	<.001[b]
SES-V[d] (total score)	0 - 105	1.37	4.50	9.09	16.19	−11 (1551)	<.001[b]
CCU[e] (1 variable)							
CIUS[f] adapted to cyberporn use (total score)	1 - 5	2.58	0.88	2.18	0.97	8.05 (1551)	.003[b]
Arousing cyberporn scenes (5 variables)							
Aroused by “submission” pornographic scenes	1 - 4	2.51	1.05	2.77	0.98	−5 (1551)	<.001[b]
Aroused by “humiliation” pornographic scenes	1 - 4	1.74	1.01	1.66	1.02	1.39 (1551)	.83
Aroused by “soft” pornographic scenes	1 - 4	2.62	0.91	2.75	0.99	−2.40 (1551)	.02[g]
Aroused by “groups with several males” pornographic scenes	1 - 4	1.83	1.05	2.18	1.13	−5.93 (1551)	<.001[b]
Aroused by “groups with several females” porn scenes	1 - 4	2.83	1.04	2.34	1.08	8.71 (1551)	.002[b]
Impulsivity (2 variables)							
UPPS-P[h] negative urgency impulsivity	1 - 4	2.63	0.65	2.40	0.65	6.61 (1551)	.89
UPPS-P positive urgency impulsivity	1 - 4	2.55	0.58	2.42	0.60	4.33 (1551)	.18
Sexual self-esteem (1 variable)							
Sexual self-esteem	1 - 4	2.35	0.87	2.30	0.94	1.09 (1551)	.22
Moral incongruence (1 variable)							

XSL•FO
RenderX

Variable categories/variables	Scale/range	Mean male	SD male	Mean female	SD female	*t* test (*df*)	*P* value
Pornography negative moral perception	1 - 7	2.21	1.62	2.30	1.57	−1.14 (1551)	.75
Sociodemographics (1 variable)							
Age (years)	18 - 75	34.67	11.45	30.87	9.26	7.11 (1551)	<.001[b]

[a]AMMSA: Acceptance of Modern Myths About Sexual Aggression.

[b]Significant at <.01.

[c]SES-P: Sexual Experience Survey–Perpetration.

[d]SES-V: Sexual Experience Survey–Victimization.

[e]CCU: compulsive cyberporn use.

[f]CIUS: Compulsive Internet Use Scale.

[g]Significant at <.05.

[h]UPPS-P: Urgency, Premeditation, Perseverance, Sensation Seeking, Positive Urgency.

As can be seen in this table, male participants had significantly higher scores than their female counterparts regarding the following variables: ASAM (AMMSA total score), SES–Perpetration (total score), CCU (CIUS total score), aroused by “groups with several females” porn scenes, and age. Female participants had significantly greater mean scores than their male counterparts on the following variables: SES–Victimization (total score), aroused by “submission” pornographic scenes, aroused by “soft” pornographic scenes, and aroused by “groups with several males” pornographic scenes.

All Variables Correlation Matrix

Table 2 shows the correlation coefficients (by gender) between the variables included in the study, two-by-two combinations. As shown in these two tables, the strongest correlations are ($r \geq 0.20$) between AMMSA and SES–Perpetration (+), AMMSA and CCU (+), SES–Perpetration and SES–Victimization (+), humiliation porn scenes and submission porn scenes (+), soft porn scenes and humiliation porn scenes (−), porn scenes with group of males and humiliation (+), UPPS-P negative and positive urgency and CCU (+), and age and submission porn scenes (−) with a far more important association among women than men.

XSL•FO
RenderX

Table . Correlation (by gender: male n=998 vs female n=555) between the variables included in this study design to predict the acceptance of sexual aggression myths in a multicountry sample.

Variables by gender	AMMSA[a] (total score)	SES-P[b] (total score)	SES-V[c] (total score)	CIUS[d] adapted to cyberporn use (total score)	Aroused by "submission" pornographic scenes	Aroused by "humiliation" pornographic scenes	Aroused by "soft" pornographic scenes	Aroused by "groups with several males" pornographic scenes	Aroused by "groups with several females" porn scenes	UPPS-P[e] negative urgency impulsivity	UPPS-P positive urgency impulsivity	Sexual self-esteem	Pornography negative moral perception	Age
AMMSA (total score), *r*														
Male participants	—[f]													
Female participants	—													
SES-P (total score), *r*														
Male participants	0.235[g]	—												
Female participants	0.217[g]	—												
SES-V (total score), *r*														
Male participants	0.076[h]	0.556[g]	—											
Female participants	0.017	0.248[g]	—											
CIUS adapted to cyberporn use (total score), *r*														
Male participants	0.271[g]	0.093[g]	0.083[g]	—										
Female participants	0.247[g]	0.127[g]	0.081	—										
Aroused by "submission" pornographic scenes, *r*														
Male participants	0.124[g]	0.032	0.050	0.174[g]	—									
Female participants	0.052	0.033	0.069	0.123[g]	—									
Aroused by "humiliation" pornographic scenes, *r*														
Male participants	0.144[g]	0.088[g]	0.054	0.121[g]	0.420[g]	—								

XSL•FO
RenderX

Variables by gender	AMM-SA[a] (total score)	SES-P[b] (total score)	SES-V[c] (total score)	CIUS[d] adapted to cyberporn use (total score)	Aroused by "submission" pornographic scenes	Aroused by "humiliation" pornographic scenes	Aroused by "soft" pornographic scenes	Aroused by "groups with several males" pornographic scenes	Aroused by "groups with several females" porn scenes	UPPS-P[e] negative urgency impulsivity	UPPS-P positive urgency impulsivity	Sexual self-esteem	Pornography negative moral perception	Age
Female participants	0.053	0.073	0.105[h]	0.073	0.341[g]	—								
Aroused by "soft" pornographic scenes, *r*														
Male participants	−0.012	0.013	−0.032	0.090[g]	−0.058	−0.215[g]	—							
Female participants	−0.013	0.088[h]	−0.024	0.094[h]	−0.009	−0.168[g]	—							
Aroused by "groups with several males" pornographic scenes, *r*														
Male participants	−0.008	0.045	0.092[g]	0.078[h]	0.164[g]	0.366[g]	−0.145[g]	—						
Female participants	0.029	0.025	0.009	0.122[g]	0.254[g]	−0.168[g]	−0.199[g]	—						
Aroused by "groups with several females" porn scenes, *r*														
Male participants	0.030	0.036	−0.038	0.047	0.006	−0.061	0.092[g]	−0.113[g]	—					
Female participants	−0.056	0.042	0.017	0.067	0.152[g]	0.188[g]	−0.086[h]	0.267[g]	—					
UPPS-P negative urgency impulsivity, *r*														
Male participants	−0.144[g]	−0.152[g]	−0.068[h]	−0.223[g]	−0.085[g]	−0.079[h]	0.016	−0.028	−0.015	—				
Female Participants	−0.104[h]	−0.066	−0.015	−0.228[g]	−0.100[h]	−0.144[g]	0.065	−0.065	−0.122[g]	—				
UPPS-P positive urgency impulsivity, *r*														
Male participants	−0.120[g]	−0.116[g]	−0.041	−0.233[g]	−0.129[g]	−0.078[h]	−0.030	−0.020	−0.064[h]	0.606[g]	—			
Female participants	−0.078	−0.144[g]	−0.120[h]	−0.276[g]	−0.102[h]	−0.072	0.002	−0.053	−0.177[g]	0.565[g]	—			
Sexual self-esteem, *r*														
Male participants	0.086[g]	0.081[h]	0.031	−0.090[g]	0.005	−0.040	−0.013	−0.030	0.030	0.214[g]	0.108[g]	—		

XSL•FO
RenderX

Variables by gender	AMMSA[a] (total score)	SES-P[b] (total score)	SES-V[c] (total score)	CIUS[d] adapted to cyberporn use (total score)	Aroused by "submission" pornographic scenes	Aroused by "humiliation" pornographic scenes	Aroused by "soft" pornographic scenes	Aroused by "groups with several males" pornographic scenes	Aroused by "groups with several females" porn scenes	UPPS-P[e] negative urgency impulsivity	UPPS-P positive urgency impulsivity	Sexual self-esteem	Pornography negative moral perception	Age
Female participants	0.164[g]	0.064	0.044	0.005	−0.024	−0.034	0.065	−0.013	0.013	0.096[h]	0.005	—		
Pornography negative moral perception, *r*														
Male participants	0.202[g]	0.097[g]	0.044	0.120[g]	−0.010	−0.010	0.074[h]	−0.047	0.019	−0.006	−0.054	−0.004	—	
Female participants	0.082	0.084[h]	−0.030	0.088[h]	−0.027	−0.044	0.148[g]	−0.171	−0.097[h]	0.008	−0.023	−0.030	—	
Age, *r*														
Male participants	−0.037	0.051	−0.006	−0.232[g]	−0.076[h]	−0.076[h]	−0.062	0.005	−0.025	0.084[g]	0.128[g]	0.079[h]	−0.075[h]	—
Female participants	0.152	0.015[g]	0.071	−0.118[g]	−0.201[g]	−0.067	−0.105[h]	0.043	0.004	0.090[h]	0.195[g]	0.124[g]	−0.123[g]	—

[a]AMMSA: Acceptance of Modern Myths About Sexual Aggression.

[b]SES-P: Sexual Experience Survey–Perpetration.

[c]SES-V: Sexual Experience Survey–Victimization.

[d]CIUS: Compulsive Internet Use Scale.

[e]UPPS-P: Urgency, Premeditation, Perseverance, Sensation Seeking, Positive Urgency.

[f]Not applicable.

[g]$P<.001$.

[h]$P<.05$.

Associations Statistics

Table 3 (regression model 1, R^2 [explained variance]=0.42, male participants; regression model 2, R^2=0.41, female participants) displays the relationships between the independent variables and the dependent variable, resulting from the multivariate regression analyses.

XSL•FO
RenderX

Table . Multivariate associations (by gender: male n=1000 vs female n=557) between the predictor variables and the criterion variable (Acceptance of Modern Myths About Sexual Aggression total score) included in this study design to predict the acceptance of sexual aggression myths in a multicountry sample.

Variables by gender	β (95% CI)	Error	*P* value[a]	VIF[b]
Violent and coercive sexuality (2 variables)				
SES-P[c] (total score)				
Male participants	.042 (0.028–0.056)	0.007	<.001	1.53
Female participants	.060 (0.034–0.086)	0.013	<.001	1.11
SES-V[d] (total score)				
Male participants	−.023 (−0.043 to −0.003)	0.010	.03	1.47
Female participants	−.006 (−0.012 to 0.000)	0.003	.07	1.11
CCU[e] (1 variable)				
CIUS[f] adapted to cyberporn use (total score)				
Male participants	.339 (0.246–0.431)	0.047	<.001	1.17
Female participants	.287 (0.185–0.389)	0.052	<.001	1.15
Arousing cyberporn scenes (5 variables)				
Aroused by “submission” pornographic scenes				
Male participants	.059 (−0.022 to 0.139)	0.041	.15	1.28
Female participants	.080 (−0.024 to 0.184)	0.053	.13	1.22
Aroused by “humiliation” pornographic scenes				
Male participants	.136 (0.047–0.224)	0.045	.003	1.42
Female participants	.040 (−0.063 to 0.143)	0.053	.45	1.30
Aroused by “soft” pornographic scenes				
Male participants	−.044 (−0.129 to 0.042)	0.044	.32	1.08
Female participants	−.065 (−0.163 to 0.033)	0.050	.19	1.11
Aroused by “groups with several males” pornographic scenes				
Male participants	−.087 (−0.164 to −0.010)	0.039	.03	1.18
Female participants	−.017 (−0.110 to 0.077)	0.047	.73	1.31
Aroused by “groups with several females” porn scenes				
Male participants	.004 (−0.069–0.077)	0.037	.92	1.03
Female participants	−.111 (−0.202 to −0.020)	0.046	.02	1.12
Impulsivity (2 variables)				
UPPS-P[g] negative urgency impulsivity				
Male participants	.180 (−0.328 to −0.032)	0.075	.02	1.68
Female participants	.123 (−0.298 to 0.051)	0.089	.17	1.54
UPPS-P positive urgency impulsivity				
Male participants	−.031 (−0.133 to 0.196)	0.084	.71	1.63
Female participants	.007 (−0.201 to 0.187)	0.099	.94	1.64
Sexual self-esteem (1 variable)				
Sexual self-esteem				
Male participants	.164 (0.076–0.253)	0.045	<.001	1.07
Female participants	.188 (0.088–0.288)	0.051	<.001	1.04
Moral incongruence (1 variable)				

XSL•FO
RenderX

Variables by gender	β (95% CI)	Error	*P* value[a]	VIF[b]
Pornography negative moral perception				
Male participants	.131 (0.084–0.178)	0.024	<.001	1.03
Female participants	.054 (−0.007–0.114)	0.031	.08	1.07
Sociodemographics (1 variable)				
Age				
Male participants	.003 (−0.003 to 0.010)	0.004	.33	1.11
Female participants	.025 (0.014–0.035)	0.005	<.001	1.15

[a]*P* value (significant at <.05).

[b]VIF: variance inflation factors; it is a test of multicollinearity (a value of 1 indicates that there is no multicollinearity; VIFs between 1 and 5 suggest that there is a moderate multicollinearity, but it is not severe enough to warrant corrective measures; VIFs higher than 5 represent critical levels of multicollinearity, indicating that there is a need for corrective measures).

[c]SES-P: Sexual Experience Survey–Perpetration.

[d]SES-V: Sexual Experience Survey–Victimization.

[e]CCU: compulsive cyberporn use.

[f]CIUS: Compulsive Internet Use Scale.

[g]UPPS-P: Urgency, Premeditation, Perseverance, Sensation Seeking, Positive Urgency.

The results from model 1 suggest that male participants were more likely to endorse the ASAM if they had higher scores of sexual coercion perpetration, CCU, arousal with pornographic scenes displaying humiliation, negative urgency impulsivity, sexual self-esteem, and moral incongruence. They were less likely to endorse the ASAM if they had higher scores of sexual coercion victimization and arousal toward pornographic scenes displaying groups of several males.

The findings from model 2 suggest that female participants were more likely to endorse the ASAM if they had higher scores of sexual coercion perpetration, CCU, sexual self-esteem, and older age. They were less likely to endorse the ASAM if they had higher scores of arousal toward pornographic scenes displaying groups of several females.

In addition, the ANOVA results show that among male participants, there was no statistical difference between the AMMSA mean score of single (n=338) and in-relationship (n=650) participants. On the contrary, the AMMSA mean scores of heterosexual (n=841), homosexual (n=61), and bisexual (n=92) participants differ statistically (3.14, SD 1.31; 2.84, SD 1.42; 2.75, SD 1.29, respectively; $F_{3,991}$=3.81, *P*=.02, η^2p=0.008). In particular, the AMMSA mean score of heterosexual participants is statistically higher than the AMMSA mean score of bisexual participants (t_{931}=2.44, *P*=.04). There was no difference between the AMMSA mean score of heterosexual versus homosexual and homosexual versus bisexual participants (t_{900}=1.56, *P*=.26; t_{151}=1.21, *P*=.45, respectively).

Among female participants, there was no statistical difference between the AMMSA mean score of single (n=143) and in-couple (n=391) participants. On the contrary, the AMMSA mean scores of heterosexual (n=387), homosexual (n=37), and bisexual (n=112) participants differ statistically (2.47, SD 0.23; 2.02, SD 1.21; 1.74, SD 0.89, respectively; $F_{3,533}$=12.48, *P*<.001, η^2p=0.045. In particular, the AMMSA mean score of heterosexual participants is statistically higher than the AMMSA mean score of bisexual participants (t_{497}=4.92, *P*<.001). There was no difference between the AMMSA mean score of heterosexual versus homosexual and homosexual versus bisexual participants (t_{422}=1.56, *P*=.26; t_{147}=1.21, *P*=.45, respectively).

Discussion

The two robust multivariate linear regression models highlighted the complex interplay of factors that contribute to the endorsement of the ASAM among male and female participants. The acceptance of these myths was assessed in this study with the AMMSA scale [1], which includes modern and socially embedded myths that justify or downplay a wide range of aggressive sexual behaviors—not limited to rape in its legal definition.

Principal Results and Comparison With Prior Work

Relationships Between Predictors and the Outcome: Male Participants

In the male cohort, as delineated by regression model 1, a multifaceted relationship was observed. Male participants who have perpetrated sexual coercion or have higher CCU scores displayed a greater propensity to endorse sexual aggression myths. This association is further compounded by heightened arousal to specific pornographic content featuring humiliation, indicative of a disturbing association between sexual aggression myths endorsement and the consumption of such specific pornographic content. Interestingly, arousal to scenes involving multiple males correlates with a reduced tendency to endorse sexual aggression myths. The association between arousal to scenes involving multiple males and the AMMSA scores is, however, weak. As such, any interpretations regarding their impact on broader understandings of sexual relationships should be considered tentative and approached with caution. Moreover, it is important to consider the role of the participants' sexual orientation in this association. Individuals with different sexual preferences might interpret and respond to such content in varied

XSL•FO
RenderX

ways, which could further contribute to the complex and nuanced understanding observed in this study.

The significant relationship between the male participants' negative urgency impulsivity, sexual self-esteem (both positively correlated with sexual aggression myths), and their AMMSA scores found in this study underlines the cognitive and emotional dimensions that fuel these beliefs. These two dimensions suggest the existence of a dual mechanism—emotional dysregulation (via urgency) and cognitive self-enhancement (via sexual self-esteem)—through which individuals may be more susceptible to endorsing harmful myths surrounding sexual aggression. In particular, negative urgency may contribute to justifying coercive behaviors as a way to manage internal tension or frustration, aligning with myths that minimize the responsibility of the aggressor, while sexual self-esteem may support rationalizations of aggression when inflated or contingent on dominance.

The findings from this study also highlighted moral incongruence as an important factor predicting male participants' AMMSA scores, hinting at an internal dissonance that may perpetuate harmful stereotypes. This particular finding is consistent with our initial hypothesis that moral incongruence would be positively associated with the ASAM, particularly given its established role in compulsive pornography use [66]. Furthermore, it highlights the role of internal psychological conflict—particularly when one's behaviors (eg, compulsive pornography use) are at odds with personal moral standards. One may hypothesize that this dissonance may lead to defensive cognitive strategies (ie, rationalization of morally conflicting behaviors [67]), such as endorsing rape myths or minimization of aggression, to alleviate guilt or justify behavior [68].

Relationships Between Predictors and the Outcome: Female Participants

Conversely, regression model 2 reveals that among female participants, patterns emerged that mirror some aspects of their male counterparts. For instance, female participants with higher scores of perpetration of sexual coercion and those with higher CCU scores are more inclined to validate sexual aggression myths. Notably, the results of this study showed that arousal to pornographic scenarios involving multiple females inversely predicts sexual aggression myths endorsement (negative relationship). Additionally, sexual self-esteem once again appears as a positive predictor—female participants with higher scores of sexual self-esteem are more likely to endorse sexual aggression myths. Age emerges as a factor unique among women, with older female participants being more likely to endorse these myths, potentially reflecting generational shifts in attitudes toward sexual violence. These findings collectively highlight the need for gender-sensitive interventions that address the roots of the ASAM, emphasizing the multifaceted and gender-specific factors that underpin this social issue.

Relationships Between Predictors and the Outcome: Male and Female Participants

When comparing male and female participants, common features explained the endorsement of sexual aggression myths, including higher CCU, sexual self-esteem, and perpetration of sexual coercion. All these factors are significantly and positively related to the participants' AMMSA scores for both sexes. A unique finding in the female cohort was the influence of age, with older women showing a higher likelihood of endorsing sexual aggression myths. The phenomenon was not observed in the male group, highlighting the possible asymmetric evolution of sexual aggression myths across gender. For both groups, arousal from porn scenes involving several people of the same sex was associated with a lower level of the ASAM. More refinement in assessing these results may be useful to understand this apparent paradox. For male participants only, arousing humiliation scenes were associated with higher endorsement of sexual aggression myths. In addition, our findings, like in Worthen's [69] study, revealed that people with heterosexual orientation have statistically higher AMMSA scores compared to the participants with a bisexual orientation. Similarly, complementary results [70] found that people identifying themselves as queer exhibit lower ASAM levels compared to people with homosexual orientation. These findings suggest that sexual minorities are generally less accepting of sexual aggression myths.

Relationships Between All the Variables

One of the notable correlations was between the sexual perpetration scores and the sexual victimization scores among men and women. This association was stronger in men. Research suggests that there can be an overlap between individuals who perpetrate sexual violence and those who are victimized, indicating that these experiences may be interconnected cyclically. People may be both perpetrators and victims of sexual violence, especially in intimate partner violence contexts. Psychological intimate partner violence seems to have an important perpetration-victim overlap [71]. In a technology-facilitated sexual violence context, victimization and perpetration are also linked, especially among adolescent males. Victimization may lead to perpetration as retaliation or learned behavior [72].

Age and arousal to submission pornographic scenes were negatively associated among both men and women. This association was particularly weak in men. This suggests that submissive porn scenarios may be more arousing for younger women compared to older women. According to the results of this study, as women age, they appear to lose interest in submission porn, which may reflect a shift in their sexual preferences. Consequently, questions could be raised about how their interest in submissive real-life sexual behavior evolves in relation to their porn use.

Theoretical Implications

Our findings indicate that both individual and societal factors significantly predict the ASAM. The associations observed between CCU, sexual self-esteem, and the ASAM across genders suggest that psychological traits and exposure to pornography-related activities are significantly related to the endorsement of sexual aggression myths. These results extend Payne et al's [3] conception of sexual aggression myths, which emphasized their cultural foundations, by highlighting the importance of individual-level characteristics—such as

XSL•FO
RenderX

impulsivity and moral incongruence—which were found to be associated with the ASAM differently according to gender.

The results illustrate the complex nature of sexual content consumption and sexual identity, therefore extending ASAM theory. According to Worthen [69], people from sexual minorities critique sexual aggression myths more due to discrimination. This study's findings on pornography use complement and extend this understanding. The multiple associations between a variety of pornographic content and the ASAM (eg, the negative association with same-sex group scenes vs the positive association with humiliation content) suggest that individual traits and content type may moderate the relationship between sexual content exposure and sexual aggression myths. If true, this questions fundamental theoretical paradigms that see porn use as either consistently negative or positive and offers an elaborate theoretical model that can include viewer traits and content-specific effects.

The positive correlation between sexual self-esteem and the ASAM presents a complex and somewhat paradoxical finding. While self-esteem is generally considered protective, higher levels of sexual self-esteem, when combined with certain sexual behaviors (such as compulsive cyberporn consumption), may instead reinforce harmful stereotypes and myths surrounding sexual violence. This may indicate that the ASAM is not merely a product of static cultural beliefs but is dynamically shaped by personal experiences, psychological traits, and behaviors, with distinct gender-specific pathways.

To fully understand the mechanisms surrounding sexual aggression myth endorsement, future theoretical frameworks should account for both cultural and dispositional dimensions. By addressing social norms and psychological factors—such as the consumption of certain types of pornography, sexual behaviors, and individual differences—interventions aimed at reducing the ASAM can become more targeted and effective in mitigating sexual aggression–supportive attitudes.

Takeaways for Possible Interventions

Based on the findings of this study, it appears necessary to address the ASAM and investigate viable solutions targeting possible associations with the porn industry and compulsive porn use. The porn industry, like other entertainment industries, may expose individuals and society to some potentially harmful stereotypes. Given the cross-sectional nature of most studies, including this study, the direction of some associations remains unclear. It is equally plausible that individuals who already endorse sexual aggression myths are more likely to be aroused by specific pornographic content (eg, humiliation or submission), rather than pornography use leading to greater ASAM. This highlights the importance of considering potential bidirectional mechanisms when exploring these associations. Indeed, while causality cannot be established, it is, however, important to find ways to remind viewers about sexual consent when promoting scenes involving such content and to promote the distinction between fantasy and reality in sexual encounters. Possible regulation mechanisms could be considered [73] for violent content but also concerning addictive porn use.

At a population level, interventions aiming to increase pornography literacy [74] and inform about the relational and affective aspects of sexual relations [75] and how to check for sexual consent can be developed and assessed, including in school-based programs [76,77], taking into consideration possible age, gender, and cultural-related issues. The association found with sexual self-esteem suggests the importance of reinforcing such a dimension within the ability to respect oneself and others' sexual and affective needs. The promotion of sexual self-esteem (if needed) must be done with precautions, given its complex relationship with the ASAM, as shown in our findings. For instance, the promotion of sexual self-esteem can be done while emphasizing healthy sexual practices and attitudes, rather than reinforcing negative stereotypes. Finally, interventions designed to address moral incongruence might be of great importance, as they can help people resolve the cognitive dissonance associated with it and help people, especially men, reach harmony between their fantasies, porn use, and sexual behavior outside of the ASAM influence.

Limitations and Future Studies

The study's limitations include its cross-sectional design and a small number of people endorsing nonbinary gender. It is also limited by some single-item measurements (eg, sexual self-esteem and moral incongruence). Although we employed measurements used in prior research, we are aware of the difficulties of generalizing these results and the need for more elaborate measures in future studies.

The strengths of the data-driven approach used in this study may have been limited by the absence of certain dimensions that should be included in future research. For instance, a more refined assessment of sexual self-esteem, along with other potentially important dimensions, such as sexual function and relational interactions, could enhance the comprehensiveness of future studies.

The study introduced the assessment of different kinds of porn-arousing scenes, considering relational aspects of the contents. We cannot exclude that some wordings, such as the group scenes, may include different kinds of relational interactions. Additionally, the arousal to pornographic content was measured through participants' self-reported responses to scene type labels (eg, humiliation, submission), without detailed descriptions or audiovisual material. This may limit the ecological validity of the measure. In addition, all participants responded to the same list of scenes, regardless of their sexual orientation, which may have affected the relevance of some items. Future research could benefit from more tailored assessments based on individual preferences.

Furthermore, given the cross-sectional design of this study, causal inferences cannot be drawn regarding the relationship between arousal to specific pornographic content and the ASAM. It remains unclear whether exposure to or preference for certain types of pornographic scenes increases the ASAM or whether individuals with higher ASAM are more likely to find such content arousing. Future longitudinal studies are needed to disentangle these directional effects and better understand the dynamics at play.

XSL•FO
RenderX

As far as we are aware, this study is the first (or one of very few) that has examined the relationships between "preferred" cyberporn scenes, sexual self-esteem, moral incongruence, and the ASAM. However, this study tested only the main effects; as such, all predictor variables were set at the same level. Future research based on this study could explore some predictor variables' mediating and moderating effects. For instance, it would be interesting to examine if sexual coercion experiences diminish sexual self-esteem and, thus, increase the ASAM, and if the arousal to humiliation/submission cyberporn scenes is associated with potential problematic cyberporn use and the ASAM, and so forth.

Conclusions

The regression models utilized in this study yielded important insights into the psychological and behavioral correlates of ASAM across genders. CCU, the perpetration of sexual coercion, and sexual self-esteem were found to be significant predictors of the ASAM among both male and female participants. Among men, impulsivity and moral incongruence emerged as particularly salient predictors, whereas age was a significant factor exclusively among women. Furthermore, marked gender differences were observed in the degree of arousal elicited by specific pornographic content and in the association between participants' subjective appreciation of such content and their levels of ASAM endorsement.

The findings from this study can be used to develop interventions in the following directions: (1) gender-sensitive educational interventions focused on critical consumption of pornographic content, particularly addressing how arousal and appreciation may unconsciously reinforce acceptance of sexual aggression myths; (2) targeted interventions for reducing CCU, such as app-based self-monitoring tools and cognitive-behavioral strategies aimed at improving online sexual self-regulation; (3) given the role of moral incongruence and impulsivity in male participants, incorporate modules on ethical decision-making, impulse control, and empathy training in school- and community-based programs; (4) tailor sexual health and consent education for women by addressing developmental and age-related factors that may influence their beliefs around sexual aggression myths; (5) integrate findings into rehabilitation programs for individuals who have engaged in sexually coercive behavior, focusing on restructuring harmful beliefs, enhancing sexual self-esteem in healthy ways, and reducing impulsivity; and (6) launch broader societal campaigns to debunk sexual aggression myths and reduce their normalization, with particular attention to gender-specific messaging and outreach.

Acknowledgments

The authors wish to acknowledge the study participants for their contribution.

Generative artificial intelligence (ChatGPT; OpenAI) was used solely to assist with grammar refinement and language polishing. The scientific content, data analysis, interpretation, and conclusions were developed entirely by the authors.

Data Availability

The data and questionnaires applied in this research can be obtained upon request from the corresponding author.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Measures.
[PDF File, 2378 KB - formative_v9i1e75485_app1.pdf]

Multimedia Appendix 2
Participants by country of residency.
[DOCX File, 16 KB - formative_v9i1e75485_app2.docx]

References

1. Gerger H, Kley H, Bohner G, Siebler F. The acceptance of modern myths about sexual aggression scale: development and validation in German and English. Aggress Behav 2007;33(5):422-440. [doi: 10.1002/ab.20195] [Medline: 17683102]
2. Burt MR. Cultural myths and supports for rape. J Pers Soc Psychol 1980;38(2):217-230. [doi: 10.1037/0022-3514.38.2.217]
3. Payne DL, Lonsway KA, Fitzgerald LF. Rape myth acceptance: exploration of its structure and its measurement using the Illinois Rape Myth Acceptance Scale. J Res Pers 1999 Mar;33(1):27-68. [doi: 10.1006/jrpe.1998.2238]
4. Abrams D, Viki GT, Masser B, Bohner G. Perceptions of stranger and acquaintance rape: the role of benevolent and hostile sexism in victim blame and rape proclivity. J Pers Soc Psychol 2003 Jan;84(1):111-125. [doi: 10.1037/0022-3514.84.1.111] [Medline: 12518974]
5. Suarez E, Gadalla TM. Stop blaming the victim: a meta-analysis on rape myths. J Interpers Violence 2010 Nov;25(11):2010-2035. [doi: 10.1177/0886260509354503] [Medline: 20065313]

XSL•FO
RenderX

6. Vera Cruz G. Human Sexuality: Discourse, Theories, Studies and Practice: Lambert Academic Publishing; 2015.
7. Fansher AK, Musamali T, Self M. Fear and consent: an exploratory study of fear of false accusations of sexual assault and consent-seeking practices. J Sch Violence 2023 Jan 2;22(1):75-88. [doi: 10.1080/15388220.2022.2138410]
8. Silver N, Hovick SR. A schema of denial: the influence of rape myth acceptance on beliefs, attitudes, and processing of affirmative consent campaign messages. J Health Commun 2018;23(6):505-513. [doi: 10.1080/10810730.2018.1473532] [Medline: 29843565]
9. Barnett MD, Hale TM, Sligar KB. Masculinity, femininity, sexual dysfunctional beliefs, and rape myth acceptance among heterosexual college men and women. Sex Cult 2017 Sep;21(3):741-753. [doi: 10.1007/s12119-017-9420-3]
10. Fansher AK, Zedaker SB. The relationship between rape myth acceptance and sexual behaviors. J Interpers Violence 2022 Jan;37(1-2):NP903-NP924. [doi: 10.1177/0886260520916831] [Medline: 32401140]
11. Bridges AJ, Wosnitzer R, Scharrer E, Sun C, Liberman R. Aggression and sexual behavior in best-selling pornography videos: a content analysis update. Violence Against Women 2010 Oct;16(10):1065-1085. [doi: 10.1177/1077801210382866] [Medline: 20980228]
12. Allen M, Emmers T, Gebhardt L, Giery MA. Exposure to pornography and acceptance of rape myths. J Commun 1995 Mar 1;45(1):5-26. [doi: 10.1111/j.1460-2466.1995.tb00711.x]
13. Hedrick A. A meta-analysis of media consumption and rape myth acceptance. J Health Commun 2021 Sep 2;26(9):645-656. [doi: 10.1080/10810730.2021.1986609] [Medline: 34641769]
14. Borgogna NC, Lathan EC, McDermott RC. She asked for it: hardcore porn, sexism, and rape myth acceptance. Violence Against Women 2022 Feb;28(2):510-531. [doi: 10.1177/10778012211037378] [Medline: 34855559]
15. Mestre-Bach G, Villena-Moya A, Chiclana-Actis C. Pornography use and violence: a systematic review of the last 20 years. Trauma Violence Abuse 2024 Apr;25(2):1088-1112. [doi: 10.1177/15248380231173619] [Medline: 37309642]
16. Ben Brahim F, Courtois R, Vera Cruz G, Khazaal Y. Predictors of compulsive cyberporn use: a machine learning analysis. Addict Behav Rep 2024 Jun;19:100542. [doi: 10.1016/j.abrep.2024.100542] [Medline: 38560011]
17. Ben Brahim F, Vera Cruz G, Courtois R, May J, Khazaal Y. Strength of Pornography Craving Experience (PCE-S): psychometric properties of a new measure based on the elaborated intrusion theory of desire. Addict Behav 2024 Jan;148:107858. [doi: 10.1016/j.addbeh.2023.107858] [Medline: 37774528]
18. Li L, Feng R, Yang L, et al. Unraveling the impact of cyberporn motivations on mental health: insights from Chinese college students. BMC Psychol 2025 May 26;13(1):562. [doi: 10.1186/s40359-025-02901-y] [Medline: 40420131]
19. Sallie SN, Ritou VJE, Bowden-Jones H, Voon V. Assessing online gaming and pornography consumption patterns during COVID-19 isolation using an online survey: highlighting distinct avenues of problematic internet behavior. Addict Behav 2021 Dec;123:107044. [doi: 10.1016/j.addbeh.2021.107044] [Medline: 34311186]
20. Camilleri C, Perry JT, Sammut S. Compulsive internet pornography use and mental health: a cross-sectional study in a sample of university students in the United States. Front Psychol 2020;11:613244. [doi: 10.3389/fpsyg.2020.613244] [Medline: 33510691]
21. Weinstein AM, Zolek R, Babkin A, Cohen K, Lejoyeux M. Factors predicting cybersex use and difficulties in forming intimate relationships among male and female users of cybersex. Front Psychiatry 2015;6:54. [doi: 10.3389/fpsyt.2015.00054] [Medline: 25941496]
22. Allen A, Kannis-Dymand L, Katsikitis M. Problematic internet pornography use: the role of craving, desire thinking, and metacognition. Addict Behav 2017 Jul;70:65-71. [doi: 10.1016/j.addbeh.2017.02.001] [Medline: 28214738]
23. Antons S, Brand M. Diagnostic and classification considerations related to compulsive sexual behavior disorder and problematic pornography use. Curr Addict Rep 2021 Sep;8(3):452-457. [doi: 10.1007/s40429-021-00383-7]
24. Chen L, Jiang X. The assessment of problematic internet pornography use: a comparison of three scales with mixed methods. Int J Environ Res Public Health 2020 Jan 12;17(2):488. [doi: 10.3390/ijerph17020488] [Medline: 31940928]
25. Cooper A, Scherer CR, Boies SC, Gordon BL. Sexuality on the Internet: from sexual exploration to pathological expression. Professional Psychol Res Pract 1999;30(2):154-164. [doi: 10.1037//0735-7028.30.2.154]
26. de Alarcón R, de la Iglesia JI, Casado NM, Montejo AL. Online porn addiction: what we know and what we don't—a systematic review. J Clin Med 2019 Jan 15;8(1):91. [doi: 10.3390/jcm8010091] [Medline: 30650522]
27. Grubbs JB, Kraus SW, Perry SL. Self-reported addiction to pornography in a nationally representative sample: the roles of use habits, religiousness, and moral incongruence. J Behav Addict 2019 Mar 1;8(1):88-93. [doi: 10.1556/2006.7.2018.134] [Medline: 30632378]
28. Beyens I, Vandenbosch L, Eggermont S. Early adolescent boys' exposure to internet pornography: relationships to pubertal timing, sensation seeking, and academic performance. J Early Adolesc 2015;35(8):1045-1068. [doi: 10.1177/0272431614548069]
29. Bőthe B, Tóth-Király I, Potenza MN, Griffiths MD, Orosz G, Demetrovics Z. Revisiting the role of impulsivity and compulsivity in problematic sexual behaviors. J Sex Res 2019 Feb;56(2):166-179. [doi: 10.1080/00224499.2018.1480744] [Medline: 29913087]
30. Grubbs JB, Exline JJ, Pargament KI, Hook JN, Carlisle RD. Transgression as addiction: religiosity and moral disapproval as predictors of perceived addiction to pornography. Arch Sex Behav 2015 Jan;44(1):125-136. [doi: 10.1007/s10508-013-0257-z] [Medline: 24519108]

XSL•FO
RenderX

31. Garza AD, Franklin CA. The effect of rape myth endorsement on police response to sexual assault survivors. Violence Against Women 2021 Mar;27(3-4):552-573. [doi: 10.1177/1077801220911460] [Medline: 32241227]
32. Sex, Mind, and Emotion: Innovation in Psychological Theory and Practice, 1st edition: Taylor and Francis; 2018.
33. Doyle Zeanah P, Schwarz JC. Reliability and validity of the sexual self-esteem inventory for women. Assessment 1996 Mar;3(1):1-15. [doi: 10.1177/107319119600300101]
34. Andersen BL, Cyranowski JM. Women's sexual self-schema. J Pers Soc Psychol 1994;67(6):1079-1100. [doi: 10.1037//0022-3514.67.6.1079]
35. Impett EA, Schooler D, Tolman DL. To be seen and not heard: femininity ideology and adolescent girls' sexual health. Arch Sex Behav 2006 Apr;35(2):131-144. [doi: 10.1007/s10508-005-9016-0] [Medline: 16752117]
36. Bernstein E, Kanefsky R, Cook M, Newins AR. Acceptance of rape myths and psychological symptoms: the indirect effect of self-blame. J Am Coll Health 2024;72(6):1641-1645. [doi: 10.1080/07448481.2022.2086005] [Medline: 35728073]
37. Brewer G, Forrest-Redfern A. Attachment anxiety, rape myth acceptance, and sexual compliance. J Interpers Violence 2022 Apr;37(7-8):NP4626-NP4639. [doi: 10.1177/0886260520948526] [Medline: 32783495]
38. Trottier D, Benbouriche M, Bonneville V. A meta-analysis on the association between rape myth acceptance and sexual coercion perpetration. J Sex Res 2021;58(3):375-382. [doi: 10.1080/00224499.2019.1704677] [Medline: 31865775]
39. Zidenberg AM, Sparks B, Harkins L, Lidstone SK. Tipping the scales: effects of gender, rape myth acceptance, and anti-fat attitudes on judgments of sexual coercion scenarios. J Interpers Violence 2021 Oct;36(19-20):NP10178-NP10204. [doi: 10.1177/0886260519872978] [Medline: 31475616]
40. Marshall EA, Miller HA. The role of sexual scripts in the relationship between pornography use and sexual coercion. J Interpers Violence 2023 Apr;38(7-8):5519-5541. [doi: 10.1177/08862605221123291] [Medline: 36181240]
41. Tomaszewska P, Krahé B. Attitudes towards sexual coercion by Polish high school students: links with risky sexual scripts, pornography use, and religiosity. J Sex Aggress 2016 Sep;22(3):291-307. [doi: 10.1080/13552600.2016.1195892]
42. Vera Cruz G, Humeau A. Sexual intercourse amongst Mozambican adults: reversal theory based inventory of motives. Sex Cult 2019 Jun;23(2):425-443. [doi: 10.1007/s12119-018-9569-4]
43. Vera Cruz G. Men's sexual sadism towards women in Mozambique: influence of pornography? Curr Psychol 2020 Jan 20;39(2):694-704. [doi: 10.1007/s12144-018-9794-x]
44. Vera Cruz G, Sheridan T. The normalization of violence during sex among young Mozambicans reportedly under the influence of pornography. Sex Cult 2022 Feb;26(1):397-417. [doi: 10.1007/s12119-021-09898-7]
45. Swedberg R, Mahoney J. Exploratory research. In: Elman C, Gerring J, editors. The Production of Knowledge: Enhancing Progress in Social Science: Cambridge University Press; 2020:17-41. [doi: 10.1017/9781108762519.002]
46. Palan S, Schitter C. Prolific.ac—a subject pool for online experiments. J Behav Exp Finance 2018 Mar;17:22-27. [doi: 10.1016/j.jbef.2017.12.004]
47. Peer E, Rothschild D, Gordon A, Evernden Z, Damer E. Data quality of platforms and panels for online behavioral research. Behav Res 2022;54(4):1643-1662. [doi: 10.3758/s13428-021-01694-3]
48. Helmke S, Kobusch PR, Rees J, et al. Beliefs about the Strauss-Kahn case in France and Germany: political orientation and sexual aggression myths as local versus global predictors. Int J Confl Violence 2004;38(1):186-186. [doi: 10.4119/UNIBI/ijcv.375]
49. Horvath MAH, Brown J, editors. Rape: Challenging Contemporary Thinking: 10 Years On, 2nd edition: Routledge; 2023. [doi: 10.4324/9781003163800]
50. Koss MP, Abbey A, Campbell R, et al. Revising the SES: a collaborative process to improve assessment of sexual aggression and victimizationn. Psychol Women Q 2007 Dec;31(4):357-370. [doi: 10.1111/j.1471-6402.2007.00385.x]
51. Testa M, VanZile-Tamsen C, Livingston JA, Koss MP. Assessing women's experiences of sexual aggression using the sexual experiences survey: evidence for validity and implications for research. Psychol Women Q 2004 Sep;28(3):256-265. [doi: 10.1111/j.1471-6402.2004.00143.x]
52. Gmel G, Khazaal Y, Studer J, Baggio S, Marmet S. Development of a short form of the compulsive internet use scale in Switzerland. Int J Methods Psych Res 2019 Mar;28(1):e1765. [doi: 10.1002/mpr.1765]
53. Meerkerk GJ, Van Den Eijnden R, Vermulst AA, Garretsen HFL. The Compulsive Internet Use Scale (CIUS): some psychometric properties. Cyberpsychol Behav 2009 Feb;12(1):1-6. [doi: 10.1089/cpb.2008.0181] [Medline: 19072079]
54. Brahim FB, Rothen S, Bianchi-Demicheli F, Courtois R, Khazaal Y. Contribution of sexual desire and motives to the compulsive use of cybersex. J Behav Addict 2019 Sep 1;8(3):442-450. [doi: 10.1556/2006.8.2019.47] [Medline: 31505968]
55. Varfi N, Rothen S, Jasiowka K, Lepers T, Bianchi-Demicheli F, Khazaal Y. Sexual desire, mood, attachment style, impulsivity, and self-esteem as predictive factors for addictive cybersex. JMIR Ment Health 2019 Jan 21;6(1):e9978. [doi: 10.2196/mental.9978] [Medline: 30664470]
56. Billieux J, Rochat L, Ceschi G, et al. Validation of a short French version of the UPPS-P Impulsive Behavior Scale. Compr Psychiatry 2012 Jul;53(5):609-615. [doi: 10.1016/j.comppsych.2011.09.001] [Medline: 22036009]
57. Lynam DR. Development of a short form of the UPPS-P Impulsive Behavior Scale (unpublished technical report). 2013.
58. Billieux J, Heeren A, Rochat L, et al. Positive and negative urgency as a single coherent construct: evidence from a large-scale network analysis in clinical and non-clinical samples. J Pers 2021 Dec;89(6):1252-1262. [doi: 10.1111/jopy.12655] [Medline: 34114654]

XSL•FO
RenderX

59. Fouyssac M, Peña-Oliver Y, Puaud M, et al. Negative urgency exacerbates relapse to cocaine seeking after abstinence. Biol Psychiatry 2022 Jun 15;91(12):1051-1060. [doi: 10.1016/j.biopsych.2021.10.009] [Medline: 34922736]
60. Quintero MJ, Navas JF, Perales JC. The associative underpinnings of negative urgency and its role in problematic gambling behavior. Addict Behav 2020 Dec;111:106533. [doi: 10.1016/j.addbeh.2020.106533] [Medline: 32771795]
61. Robins RW, Hendin HM, Trzesniewski KH. Measuring global self-esteem: construct validation of a single-item measure and the Rosenberg self-esteem scale. Pers Soc Psychol Bull 2001 Feb;27(2):151-161. [doi: 10.1177/0146167201272002]
62. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. PLoS Med 2007 Oct 16;4(10):e296. [doi: 10.1371/journal.pmed.0040296] [Medline: 17941714]
63. Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). Biometrika 1965 Dec 1;52(3-4):591-611. [doi: 10.1093/biomet/52.3-4.591]
64. Andersen R. Modern Methods for Robust Regression: Sage Publications; 2008. [doi: 10.4135/9781412985109]
65. Todorov V, Filzmoser P. An object-oriented framework for robust multivariate analysis. J Stat Soft 2009;32(3). [doi: 10.18637/jss.v032.i03]
66. Lewczuk K, Glica A, Nowakowska I, Gola M, Grubbs JB. Evaluating pornography problems due to moral incongruence model. J Sex Med 2020 Feb;17(2):300-311. [doi: 10.1016/j.jsxm.2019.11.259] [Medline: 31818724]
67. Lerner MJ. The belief in a just world. In: The Belief in a Just World: A Fundamental Delusion: Springer; 1980:9-30. [doi: 10.1007/978-1-4899-0448-5_2]
68. Grubbs JB, Perry SL, Wilt JA, Reid RC. Pornography problems due to moral incongruence: an integrative model with a systematic review and meta-analysis. Arch Sex Behav 2019 Feb;48(2):397-415. [doi: 10.1007/s10508-018-1248-x] [Medline: 30076491]
69. Worthen MGF. Rape myth acceptance among lesbian, gay, bisexual, and mostly heterosexual college students. J Interpers Violence 2021 Jan;36(1-2):NP232-NP262. [doi: 10.1177/0886260517733282] [Medline: 29294932]
70. Schulze C, Koon-Magnin S. Gender, sexual orientation, and rape myth acceptance: preliminary findings from a sample of primarily LGBQ-identified survey respondents. Violence Vict 2017 Feb 1;32(1):159-180. [doi: 10.1891/0886-6708.VV-D-15-00017] [Medline: 28234204]
71. Clemens V, Fegert JM, Kavemann B, et al. Epidemiology of intimate partner violence perpetration and victimisation in a representative sample. Epidemiol Psychiatr Sci 2023 Apr 19;32:e25. [doi: 10.1017/S2045796023000069] [Medline: 37073832]
72. Gámez-Guadix M, Sorrel MA, Martínez-Bacaicoa J. Technology-facilitated sexual violence perpetration and victimization among adolescents: a network analysis. Sex Res Soc Policy 2023 Sep;20(3):1000-1012. [doi: 10.1007/s13178-022-00775-y]
73. Vera-Gray F, McGlynn C, Kureshi I, Butterby K. Sexual violence as a sexual script in mainstream online pornography. Br J Criminol 2021 Sep 2;61(5):1243-1260. [doi: 10.1093/bjc/azab035]
74. Rothman EF, Daley N, Alder J. A pornography literacy program for adolescents. Am J Public Health 2020 Feb;110(2):154-156. [doi: 10.2105/AJPH.2019.305468] [Medline: 31855489]
75. Astle S, McAllister P, Emanuels S, Rogers J, Toews M, Yazedjian A. College students' suggestions for improving sex education in schools beyond 'blah blah blah condoms and STDs'. Sex Educ 2021 Jan 2;21(1):91-105. [doi: 10.1080/14681811.2020.1749044]
76. Bradley AR, Yeater EA, O'Donohue W. An evaluation of a mixed-gender sexual assault prevention program. J Prim Prev 2009 Nov;30(6):697-715. [doi: 10.1007/s10935-009-0198-4] [Medline: 19937385]
77. Salazar LF, Vivolo-Kantor A, Hardin J, Berkowitz A. A web-based sexual violence bystander intervention for male college students: randomized controlled trial. J Med Internet Res 2014 Sep 5;16(9):e203. [doi: 10.2196/jmir.3426] [Medline: 25198417]

Abbreviations

AMMSA: Acceptance of Modern Myths about Sexual Aggression
ASAM: Acceptance of Sexual Aggression Myths
CCU: compulsive cyberporn use
CIUS: Compulsive Internet Use Scale
SES-P: Sexual Experience Survey–Perpetration
SES-V: Sexual Experience Survey–Victimization
UPPS-P: Urgency, Premeditation, Perseverance, Sensation Seeking, Positive Urgency

XSL•FO
RenderX

Edited by A Mavragani; submitted 04.04.25; peer-reviewed by D Henry, JC Navarro; revised version received 24.07.25; accepted 24.07.25; published 16.10.25.

Please cite as:
Ben Brahim F, Vera Cruz G, Courtois R, Khazaal Y
Predictors for Acceptance of Sexual Aggression Myths Among People Using Cyberporn: Cross-Sectional Study
JMIR Form Res 2025;9:e75485
URL: https://formative.jmir.org/2025/1/e75485
doi:10.2196/75485

© Farah Ben Brahim, Germano Vera Cruz, Robert Courtois, Yasser Khazaal. Originally published in JMIR Formative Research (https://formative.jmir.org), 16.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Associations Between Online Casual Sexual Behavior and HIV-Related Risk Behaviors Among Men Who Have Sex With Men in Southeast China: Cross-Sectional Study

Lin Chen[1], MD; Zhongrong Yang[2], MD; Wanjun Chen[1], MD; Tingting Jiang[1], MD; Xiaohong Pan[1], MD

[1]Department of HIV/AIDS and STDs control and prevention, Zhejiang Provincial Center for Disease Control and Prevention, 3399 Binsheng Road, Binjiang District, Hangzhou, Zhejiang Province, China

[2]Department of HIV/TB control and prevention, Huzhou Center for Disease Control and Prevention, Huzhou, Zhejiang Province, China

Corresponding Author:
Xiaohong Pan, MD
Department of HIV/AIDS and STDs control and prevention, Zhejiang Provincial Center for Disease Control and Prevention, 3399 Binsheng Road, Binjiang District, Hangzhou, Zhejiang Province, China

Abstract

Background: With the growing popularity and convenience of the internet, an increasing number of men who have sex with men (MSM) are seeking casual sexual partners online. However, the effect of online casual sexual behavior on other HIV-related risk behaviors remains unclear.

Objective: This study aims to explore the characteristics of internet-based casual sexual behavior and its relationship with HIV-related risk behaviors among MSM.

Methods: This cross-sectional study was conducted between June and December 2018 in 4 cities in Zhejiang Province, China. Peer-driven sampling was used for recruitment. Announcements were disseminated by 4 community-based organizations and 10 voluntary counseling and testing clinics online and offline. After informed consent, participants completed an electronic questionnaire covering demographic characteristics, casual sexual behaviors, HIV-related risk behaviors, and HIV prevention. SPSS (version 19.0; IBM Corp) was used to conduct chi-square tests, univariate and multivariate logistic regression analyses using a backward stepwise method based on the likelihood ratio test, and Poisson regression with robust variance to identify associations between finding casual sexual partners online and other risk behaviors. *P* values of <.05 were considered statistically significant.

Results: In the past 6 months, 40.2% (302/751) of participants reported finding casual sexual partners online; 18.9% (142/751) reported finding casual sexual partners offline; 7.6% (57/751) reported having sexual intercourse with MSM without condoms after drinking alcohol; and 6.9% (52/751) reported condomless sex after using stimulants. Among those who found partners online, 62.5% (188/301) did so more than once per month and 39.5% (113/286) had more than one online sexual partner. In total, 39.3% (114/290) had sex with online partners at home and 10.1% (30/297) sought partners in other cities. Compared with participants who engaged in receptive anal intercourse (or both roles), those who engaged only in insertive intercourse reported a higher proportion of finding partners online more than once per month (72.7% vs 57.4%, *P*=.01), having more than 2 online sexual partners (52.1% vs 33.3%, *P*=.002), and conducting inconsistent condom use with online sexual partners (40.0% vs 25.8%, *P*=.01). Regression analysis showed that, compared with MSM who did not find partners online, those who did were more likely to report finding casual sexual partners offline (adjusted odds ratio [aOR] 9.398; 95% CI 5.956 - 14.829), having sex without condoms after drinking alcohol (adjusted prevalence ratio [aPR] 1.788; 95% CI 1.062 - 3.011), and having sexual intercourse without condoms after using stimulants (aPR 2.064; 95% CI 1.178 - 3.617).

Conclusions: Internet-based casual sexual behavior is increasingly common among MSM. Finding partners online was associated with offline partner-seeking and condomless use after alcohol or stimulant use. Future HIV prevention efforts should emphasize behavioral interventions tailored to MSM who use dating apps.

(JMIR Form Res 2025;9:e71690) doi:10.2196/71690

KEYWORDS

human immunodeficiency virus; men who have sex with men; sexual behavior on the internet; cross-sectional study; stimulants use

XSL•FO
RenderX

Introduction

In recent years, men who have sex with men (MSM) have been one of the populations most seriously affected by HIV globally [1-3]. In China, the HIV-positive rate was the highest among MSM, compared with other high-risk populations (injecting drug users and commercial sex workers) in China [4,5]. In Zhejiang Province, nearly 40% of HIV-positive individuals acquired HIV through male-to-male sexual behavior. MSM have become a high-risk group for HIV infection, largely due to factors such as unprotected sex, multiple sexual partners, casual sex, and substance abuse [6,7]. These behaviors are closely tied to broader developments in social science and technology—particularly the rise of the internet, which has facilitated widespread online sexual behavior.

In recent years, smartphone apps such as Grindr (Grindr LLC), Jack'd (Lucid Dreams LLC), Blued (BlueCity Holdings Ltd), Manhunt (Online Buddies, Inc), Scruff (Perry Street Software, Inc), and Black Gay Chat (BGCLive, LLC) have become increasingly popular among MSM [8]. In China, Blued is the most widely used social app among MSM, with more than 30 million registered users nationwide. It is estimated that there are 409,000 MSM in Zhejiang Province and 8.28 million MSM in mainland China, accounting for roughly 1.732% (8,288,536/478,552,887) of all adult men aged 18-64 years in China [9]. Data show that 71.8% (18/252) of MSM report engaging in online sexual behavior, with this figure reaching 60% in China [10,11].

The internet has brought convenience and greater privacy to sexual activity among MSM, particularly for individuals with higher levels of education. However, the increased frequency of sexual encounters and the unknown HIV status of online partners have contributed to the ongoing spread of HIV and other sexually transmitted infections (STIs) within the MSM community [12]. A cohort study found that seeking sexual partners online was an independent risk factor for HIV infection [9]. This trend calls for urgent public health attention.

Regarding online casual sex, recent research has focused on the prevalence, characteristics, risk factors, and potential interventions related to internet-based sexual behaviors [13-15]. One study in India, for instance, compared current internet-related sexual behaviors between MSM and male-to-female transgender individuals [13]. However, the broader impact of online sex behavior on other HIV-related risk behaviors has often been overlooked. A systematic narrative review indicated that the existing literature does not clearly establish a link between online partner-seeking and condom use or STI status [16]. This gap highlights the need for more in-depth research in the field.

This study hypothesized that casual sexual behavior online influences HIV-related risk behaviors. Based on the literature, the following behaviors were selected as key outcomes potentially linked to HIV infection: offline casual sex, inconsistent condom use with MSM after alcohol consumption, and inconsistent condom use with MSM after stimulant use [17-19]. Compared with those who find partners online, individuals seeking partners offline report more sexual partners, a higher frequency of sexual activity, and lower rates of condom use [17]. In addition, alcohol and stimulant use may further reduce condom use and increase the likelihood of HIV infection [18,19]. One study found that 39.5% (230/582) of men reported using alcohol before sex in the past 3 months [18], and 40.7% (483/1186) reported stimulant use among MSM [20]. These 3 behaviors are closely associated with HIV transmission among MSM.

While a few studies have explored the relationship between online sex behavior and condom use, there remains a lack of research on condom use in the context of alcohol and stimulant use [21]. A cross-sectional study was conducted in 2018 in Zhejiang Province to gain a deeper understanding of the characteristics of online casual sex and its impact on other high-risk sexual behaviors. The findings aim to provide empirical evidence for future policy and strategy development.

Methods

Study Participants

This cross-sectional study was conducted between June and December 2018 among MSM in 4 cities—Hangzhou, Ningbo, Wenzhou, and Shaoxing in Zhejiang Province, China. The inclusion criteria were as follows: individuals who (1) had anal or oral intercourse with a male within the past 6 months, (2) were aged 18 years or older, (3) resided locally for more than 6 months, and (4) consented to participate in the study.

Study Design and Data Collection

Participants were recruited through 4 local community-based organizations (CBOs) and 10 voluntary counseling and testing (VCT) clinics in Zhejiang Province, via venues for gay men and networks frequented by gay men. More than half of all MSM in the province resided in the cities of Hangzhou, Ningbo, Wenzhou, and Shaoxing. A peer-driven sampling method was used. First, CBOs posted recruitment advertisements in locations such as bathrooms, bars, and parks, as well as on online platforms, including social apps (Blued, WeChat [Tencent], and QQ [Tencent]) and social media. In addition, VCT clinics promoted the study in their physical offices and on their WeChat official accounts. Second, MSM already enrolled through CBOs or VCT clinics were encouraged to refer their peers.

To screen for eligibility, the participants answered 4 preliminary questions. After receiving training from either a CBO member or a physician, the participants scanned a QR code to access an electronic informed consent form. Those who selected "Agree" proceeded to complete the questionnaire.

The electronic questionnaire was developed using Wenjuanxing (Changsha Ranxing Information Technology Co, Ltd), a secure online survey platform that requires a username and password for access and data management. After its initial design, the questionnaire was pilot-tested with 5 respondents to refine its structure. Skip logic and mandatory responses were implemented to improve data quality. The final questionnaire included 39 items.

Participants received a gift valued at 30 RMB (approximately US $5) upon completion. To verify against duplicate entries,

XSL•FO
RenderX

participants were asked to authorize their mobile phone numbers—no duplicates were identified. Five MSM were involved in designing the survey protocol and reviewing the questionnaire. Respondents were also able to review and revise their answers using a "Back" button.

The required sample size was calculated based on an estimated 40% - 60% rate of online sexual partners seeking online among MSM. Using PASS software version 11.0 (NCSS, LLC), the minimum sample size was determined to be 266 participants, assuming α=.1 and β=.1.

Definition of Key Variables

Regarding demographic characteristics, participants reported their age, educational background, registered permanent residence, income, and sexual role. Age was calculated from birth dates provided in the questionnaire. Background was categorized into illiterate, primary school, junior middle school, high school, college, and postgraduate education. For logistic regression analysis, educational background was dichotomized as "high school or below" and "college or above." Participants who reported having their registered in the local county or district were classified as "native." Previous research has shown that individuals with a receptive sexual role are at higher risk of HIV infection than are individuals with an insertive role. In this study, sexual role was coded as "receptive or both" or "insertive only" for use in regression analyses and in characterizing individuals who seek partners online.

Characteristics of HIV-Related Risky Behavior

Participants were asked, "Have you ever dated and had casual sexual intercourse with men you met on the internet, such as via Blued, WeChat, a chat room, or other platforms?" Those who answered "Yes" were defined as having "found casual sexual partners online." Participants were also asked, "Have you ever had casual sex intercourse with men you met in a bar, park, bathing pool, or other place in the past 6 months?" Those who answered "Yes" were categorized as having "found sexual partners offline."

To assess condom use after alcohol consumption, participants were asked: "Have you ever engaged in sexual behavior without a condom after drinking in the last 6 months?" Those who answered "Yes" were defined as having engaged in the relevant risk behavior.

For stimulant use, participants were asked: "Have you ever engaged in sexual behavior without a condom after using stimulants, such as 'rush poppers' or 'zero capsules,' in the last 6 months?" Again, participants answering "Yes" were categorized as having engaged in this risk behavior. The main components of "rush poppers" and "zero capsules" are various nitrite esters and dimethyltryptamine (5-MeO-DiPT), respectively. Both stimulants are popular among MSM and enhance sexual stimulation.

Characteristics of Finding Sexual Partners Online

The variables used to describe the characteristics of finding casual partners online were the frequency of finding partners online (once a month or more often), number of online sexual partners (one or more), place where casual sex occurred (home, karaoke lounge, club, or hotel), city where the encounter took place (local city or other city), age of the online sexual partner (0 - 30 or >30 years), and condom use with online sexual partners (every time, sometimes, or never).

In the chi-square test, the place of sexual intercourse was recorded as "hotel or karaoke lounge or club" or "home." Condom use was recorded into "consistent condom use" (reported use every time) and "inconsistent condom use" (reported use sometimes or never). All variables refer to behaviors within the previous 6 months.

The selection of variables related to finding sexual partners online was based on their importance in the literature and frequency of appearance in previous research reports.

Other Confounders

Participants were asked the following question regarding HIV intervention: "Have you ever seen any HIV–related information or propaganda on a dating app, such as Blued, WeChat, Aloha, etc.?" Responding "Yes" was defined as "acquired HIV information from a dating app." For logistic regression, responses were dichotomized as "No" and "Yes."

Another variable was assessed through the question: "Have you seen a sexually explicit video on the internet in the last 6 months?" Response options included: "3 - 7 times per week," "1 - 2 times per week," "2 times per month," "1 time per month," "less than 1 time per month," and "Never." For logistic regression analysis, this variable was also dichotomized as "No" and "Yes."

Data Analysis

Data were analyzed using SPSS software (version 19.0; IBM Corp). Descriptive analyses were conducted to summarize demographic characteristics and risky sexual behaviors among all MSM participants. The Pearson chi-square tests were used to examine differences in characteristics of web-based casual sexual behavior between groups with different educational backgrounds and sexual roles. Univariate and multivariate logistic regression analyses using a backward stepwise method based on the likelihood ratio test were performed to assess the relationship between finding casual sexual partners online and finding casual sexual partners offline. Poisson models with robust variance were performed to evaluate the relationship between finding casual sexual partners online and sexual intercourse with MSM without condoms after drinking alcohol, as well as sexual intercourse with MSM without condoms after using stimulants. All multivariate models were adjusted for 4 confounders: education level (0=high school or below, 1=college or above), sexual role (0=receptive or both, 1=insertive only), acquisition of HIV information from a dating app (0=no, 1=yes), and viewing sexually explicit videos on the internet (0=no, 1=yes). Missing data were excluded from the analysis but are reported in the tables. A P value of <.05 was considered statistically significant, and all hypothesis tests were 2-sided.

Ethical Considerations

All procedures in this study were approved by the Ethics Committee of the Zhejiang Provincial Center for Disease Control and Prevention (approval number: 2018 - 033).

XSL•FO
RenderX

Electronic informed consent was obtained from all participants. Before completing the questionnaire, respondents were informed—either in person by CBO staff or by phone—about the purpose, estimated duration, survey content, and data confidentiality. The informed consent form specified that deidentified data could be used for secondary analysis without the need for reconsent. To ensure confidentiality, phone numbers were anonymized and permanently deleted from the dataset after verification. Each respondent will receive a gift worth US $4, and CBOs will get a reward of US $7 for completing the questionnaire survey of each respondent.

Results

In total, 812 individuals met the inclusion criteria. Of these, 19 did not complete the questionnaire, resulting in 793 (97.7%) valid initial responses being included. After excluding 42 participants with missing key information, the final sample included 751 participants. The recruitment procedure is shown in Figure 1.

Figure 1. Recruitment process flow chart for participants. CBO: community-based organization; VCT: voluntary counseling and testing.

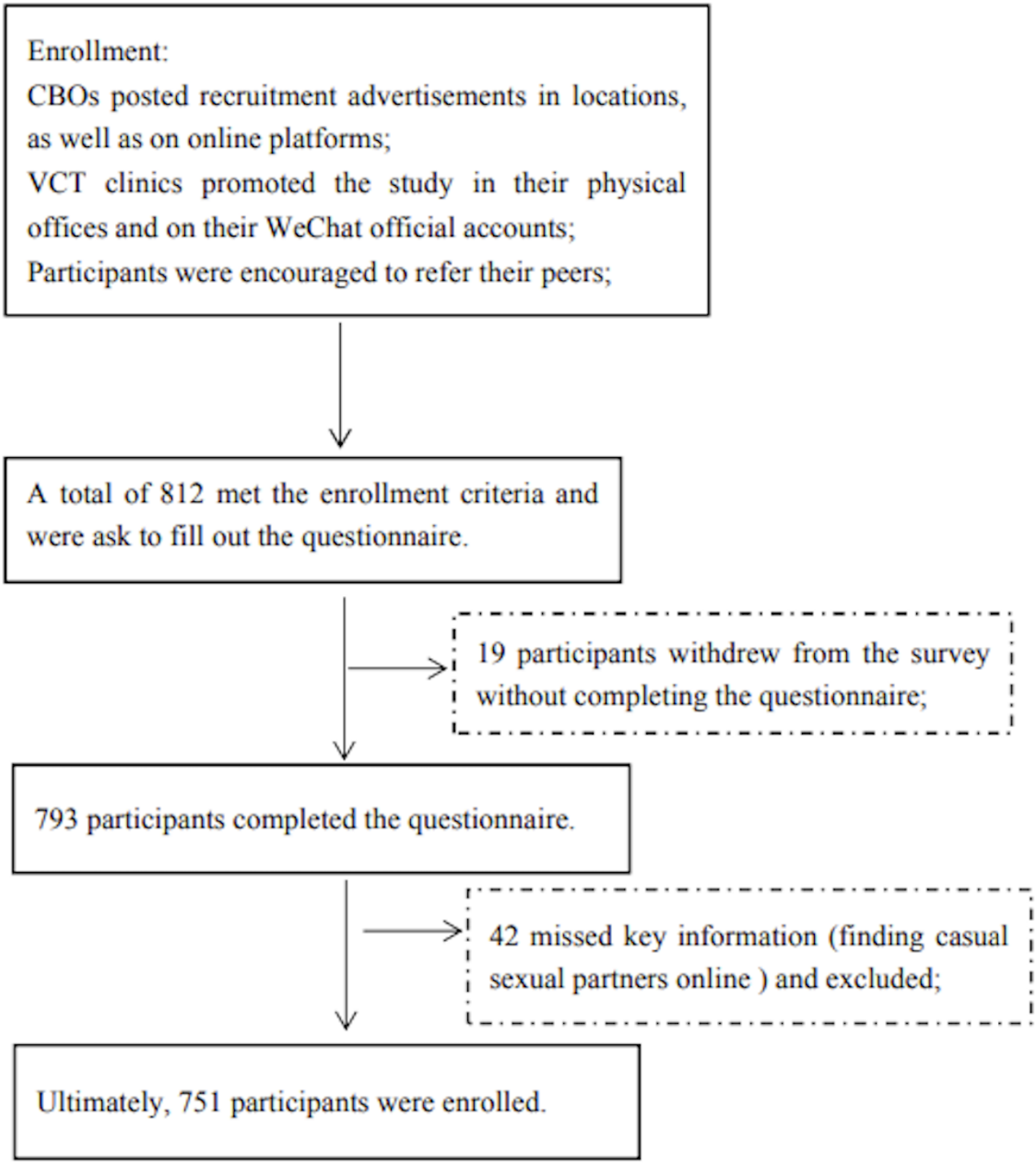

XSL•FO
RenderX

Among the 751 participants, 24.1% (181/751) were aged 35 years or older, 61.9% (465/751) had a college education or higher, and 63.9% (480/751) had ever engaged in a receptive sexual role. In the past 6 months, 67.0% (503/751) reported having acquired HIV information from dating apps and 78.7% (591/751) had viewed sexually explicit videos on the internet (Table 1).

Table . Sociodemographic characteristics, HIV-related risk behaviors, intervention exposure, and associations with other HIV-related risk behaviors. Data are presented as n (%).

Variables	All MSM[a] (N=751)	Found casual sexual partners offline (n=142)	Chi-square (*df*)	*P* value	Sexual intercourse with MSM without condoms after drinking alcohol (n=57)	Chi-square (*df*)	*P* value	Sexual intercourse with MSM without condoms after using stimulants (n=52)	Chi-square (*df*)	*P* value
Age (years)			5.133 (2)	.08		0.094 (2)	.95		0.159 (2)	.92
18 - 24	241 (32.1)	45 (18.7)			18 (7.5)			18 (7.5)		
25 - 34	329 (43.8)	53 (16.1)			26 (7.9)			22 (6.7)		
≥35	181 (24.1)	44 (24.3)			13 (7.2)			12 (6.6)		
Education level			11.831 (1)	.001		8.530 (1)	.003		0.003 (1)	.09
High school or below	286 (38.1)	72 (25.2)			32 (11.2)			20 (7.0)		
College or above	465 (61.9)	70 (15.1)			25 (5.4)			32 (6.9)		
Sexual role			4.359 (1)	.03		0.169 (1)	.68		0.004 (1)	.95
Receptive or both	480(63.9)	80(16.7)			35(7.3)			33(6.9)		
Insertive only	271(36.1)	62(22.9)			22(8.1)			19(7.0)		
Acquired HIV information from a dating app			8.708 (1)	.003		0.407 (1)	.52		0.133 (1)	.72
No	248 (33.0)	32 (12.9)			21 (8.5)			16 (6.5)		
Yes	503 (67.0)	110 (21.9)			36 (7.2)			36 (7.2)		
Watched sexually explicit videos on the internet			0.003 (1)	.95		1.119 (1)	.29		3.194 (1)	.07
No	160 (21.3)	30 (18.8)			9 (5.6)			6 (3.8)		
Yes	591 (78.7)	112 (19.0)			48 (8.1)			46 (7.8)		
Found casual sexual partners online			112.861 (1)	<.001		5.1531 (1)	.02		8.826 (1)	.003
No	449 (59.8)	29 (6.5)			26 (5.8)			21 (4.7)		
Yes	302 (40.2)	113 (37.4)			31 (10.3)			31 (10.3)		

[a]MSM: men who have sex with men.

Within the past 6 months, 40.2% (302/751) of the participants reported having found casual sexual partners online; 18.9% (142/751) had found casual partners at bars, bathrooms, or parks; 7.6% (57/751) reported having sexual intercourse with MSM without condoms after drinking alcohol; and 6.9% (52/751) reported having sex without condoms after using stimulants (Table 1).

The chi-square test results showed statistically significant differences in the prevalence of 3 HIV-related risk behaviors between those who found partners online and those who did not ($P<.05$): finding casual sexual partners offline (37.4%, 113/302 vs 6.5%, 29/449; $P<.001$), sexual intercourse with MSM without a condom after drinking alcohol (10.3%, 31/302, vs 5.8%, 26/449; $P=.02$), and sexual intercourse without a condom after using stimulants (10.3%, 31/302 vs 4.7%, 21/449; $P=.003$; Table 1). Among MSM who reported finding casual sexual partners online, 62.5% (188/301) met partners more than once per month; 39.5% (113/286) had more than one online sexual partner; 60.7% (114/290) had sex with partners at a hotel, karaoke lounge, or club; 89.9% (267/297) found casual partners in the local city; 72.2% (213/298) preferred partners younger than 30 years; and 69.4% (206/297) reported using condoms

XSL•FO
RenderX

every time with online sexual partners. The most common partner selection preferences were body shape and age (Table 2).

Table . Frequency, condom use, and locations of online sexual encounters among men who have sex with men who sought sexual partners online in the past 6 months.

Variables	N (%)	Sexual role		Chi-square (*df*)	*P* value
		Receptive anal or both, n (%)	Inserted anal only, n (%)		
Frequency of finding partners online (time per month)				6.634 (1)	.01
1	113 (37.5)	86 (42.6)	27 (27.3)		
>1	188 (62.5)	116 (57.4)	72 (72.7)		
Missing	1				
Number of online sexual partners				9.327 (1)	.002
2	173 (60.5)	128 (66.7)	45 (47.9)		
>2	113 (39.5)	64 (33.3)	49 (52.1)		
Missing	16				
Place where casual sexual intercourse occurred				0.361 (1)	.55
Hotel or karaoke lounge or club	176 (60.7)	116 (59.5)	60 (63.2)		
Home	114 (39.3)	79 (40.5)	35 (36.8)		
Missing	12				
City in which casual sexual intercourse occurred				0.167 (1)	.68
Local city	267 (89.9)	177 (89.4)	90 (90.9)		
Other city	30 (10.1)	21 (10.6)	9 (9.1)		
Missing	5				
Age of online sexual partners				0.230 (1)	.63
0 - 30 years	213 (71.5)	144 (72.4)	69 (69.7)		
>30 years	85 (28.5)	55 (27.6)	30 (30.3)		
Missing	4				
Condom use with online sexual partners				6.662 (1)	.01
Consistent	206 (69.4)	147 (74.2)	59 (59.6)		
Inconsistent	91 (30.6)	51 (25.8)	40 (40.0)		
Missing	5				

Compared with participants who engaged in receptive anal intercourse (or both roles), those who engaged only in insertive intercourse reported a higher proportion of finding partners online more than once per month (72.7%, 72/99 vs 57.4%, 116/202; *P*=.01), having more than 2 online sexual partners (52.1%, 49/94 vs 33.3%, 64/192; *P*=.002), and inconsistent condom use with online sexual partners (40.0%, 40/99 vs 25.8%, 51/198; *P*=.01; Table 2).

In the univariate and multivariate logistic regression analyses, the relationships between finding casual sexual partners online and other risky sexual behaviors were evaluated, after controlling for education (0=High School and under, 1=College and above), sex role (0=Receptive or both, 1= Insertive only), acquisition of HIV information from a dating app (0=No, 1=Yes), and watching obscene videos on the internet (0=No, 1=Yes). Compared to MSM who did not find partners online, those who did reported a higher likelihood of finding casual sexual partners offline (adjusted odds ratio [aOR] 9.398; 95% CI 5.956 - 14.829), having sexual intercourse without condoms after drinking alcohol (adjusted prevalence ratios [aPR] 1.788; 95% CI 1.062 - 3.011), and having sexual intercourse without condoms after using stimulants (aPR 2.064; 95% CI 1.178 - 3.617; Table 3).

XSL•FO
RenderX

Table . Univariate and multivariate logistic regression analyses of associations between finding casual sexual partners online and 3 HIV-related risk behaviors in the past 6 months among men who have sex with men.

Variables	Found casual sexual partners offline		Sexual intercourse with MSM[a] without condoms after drinking alcohol		Sexual intercourse with MSM without condoms after using stimulants	
	Crude OR[b] (95% CI)	Adjusted OR (95% CI)	Crude PR[c] (95% CI)	Adjusted PR (95% CI)	Crude PR (95% CI)	Adjusted PR (95% CI)
Education	0.527 (0.364 - 0.762)	0.460 (0.304 - 0.697)	0.481 (0.285 - 0.811)	0.477 (0.283 - 0.805)	0.986 (0.564 - 1.724)	__[d]
Sex role	1.483 (1.023 - 2.150)	1.723 (1.132 - 2.623)	1.113 (0.653 - 1.898)	—	1.018 (0.579 - 1.790)	—
Acquired HIV information from dating app	1.889 (1.232 - 2.896)	1.829 (1.148 - 2.918)	0.845 (0.493 - 1.448)	—	1.112 (0.617 - 2.003)	—
Watched sexually explicit videos on the internet	1.013 (0.648 - 1.584)	—	1.444 (0.708 - 2.943)	—	2.079 (0.888 - 4.868)	—
Found casual sexual partners online	8.659 (5.562 - 13.480)	9.398 (5.956 - 14.829)	1.773 (1.053 - 2.985)	1.788 (1.062 - 3.011)	2.202 (1.265 - 3.832)	2.064 (1.178 - 3.617)

[a]MSM: men who have sex with men.

[b]OR: odds ratio.

[c]PR: prevalence ratio.

[d]Not available.

Discussion

Principal Findings

With the growing popularity of the internet, online casual sexual behaviors have become increasingly common. In this study, the rate of finding partners online for casual sexual behavior was nearly 40%, more than twice the rate of offline sex encounters. This may be attributed to the convenience and anonymity of the internet, which can facilitate the occurrence of casual sexual activity. Notably, the rate of online casual sex in this study was higher than that in Mexico but lower than that in the United States and India [13,21,22]. These differences may be linked to the earlier and broader adoption of internet technologies, as well as social culture.

When examining behaviors associated with increased risk of STI or HIV transmission, MSM with online sexual partners were found in other studies to report a greater number of partners, and higher proportions reported exchanging sex for money or drugs and having sex with partners who themselves had multiple partners [21]. This may represent one of the biggest challenges to HIV prevention in China over the next decade and deserves heightened attention.

Existing research exploring the complex dynamics of online casual sex among MSM remains limited, with only a few studies available [23]. This study contributes by outlining the demographic and behavioral characteristics of MSM and their online sexual partners. Notably, not all encounters took place in hotels; approximately half occurred at home, possibly because of cost-saving considerations. This shift complicates HIV prevention efforts because traditional strategies—such as distributing condoms and educational materials in hotels—may not be as effective in private domestic settings [24]. Data also indicate that a small proportion of MSM engage in online sexual activity while traveling to other regions. This includes individuals with stable partners who travel for work, as well as single individuals traveling for leisure. The increasing availability of mobile technology has linked mobility with heightened sexual risk behavior [25]. These findings raise questions about the accessibility and timeliness of intervention services, such as postexposure prophylaxis and HIV testing, for individuals in transit or away from home.

We also found that approximately one-third of the participants reported having online sexual partners aged 30 years or older. Compared with other populations, MSM are younger when they become infected with HIV [26]. The older a sexual partner is, the higher the probability that they are living with HIV. Seeking older sexual partners may increase HIV risk, particularly among those new to MSM social networks or those engaging in receptive anal intercourse—groups that warrant focused attention in HIV prevention strategies.

Regarding condom use, nearly one-third of participants reported inconsistent condom use with online sexual partners. This rate was especially high among those who reported only insertive anal intercourse and those with a high school education or below. Though the rate of inconsistent condom use in this study was lower than that reported in some other countries [27], it remains a concern. One study found that MSM were less likely to engage in condomless anal sex with partners met online compared with those met offline, possibly because they were more likely to ask about HIV status and testing history beforehand [28]. Nonetheless, the rate of condomless sex continues to represent a key driver of HIV transmission in this population and should not be overlooked.

XSL•FO
RenderX

Many researchers have found that MSM who meet their partners online are more likely to engage in risky sexual behaviors, such as receptive anal sexual partners and reporting condomless receptive anal sex, past 6 months of illicit drug use in the previous 6 months, and having sex with anonymous partners whose HIV status is unknown [17,29]. One study found that HIV-positive MSM had significantly more partners whom they had met online than did HIV-negative MSM [13]. In this study, after adjusting for confounders, we found persistent associations between finding casual sexual partners online and 3 HIV-related risk behaviors: finding casual sexual partners offline, engaging in sexual activity without condoms after drinking alcohol, and engaging in sexual activity without condoms after using stimulants. These findings support existing evidence that online sexual behavior is linked to increased HIV risk among MSM.

The effects of drinking alcohol and stimulant use on HIV infection risk have been well-documented [18,19]. Rush poppers have become one of the most popular stimulants among MSM in China, used for both pain relief and sexual pleasure. Reports suggest that 20% - 30% of MSM in China use rush poppers [11,30]. Users were more likely to have multiple sexual partners, casual partners, or STIs [31]. A 2018 US study found that 59.9% of MSM reported alcohol or illicit drug use before or during sex [10]. Both an increased number of insertive and receptive partners and substance use during sex were associated with higher odds of seeking partners online [10].

This study implies that online sexual behavior may play a driving role in the occurrence of other risk behaviors. For example, individuals who engage in sex with unfamiliar online partners—particularly receptive partners—may be more likely to use sexual stimulants, leading to impaired judgment or reduced control over condom use. In the case of alcohol, some MSM may intentionally meet for drinks or meals before sex to reduce awkwardness or build comfort, which may contribute to condomless sex. These patterns suggest that HIV-related risk behaviors tend to cluster among MSM—especially those at highest risk of infection—who should be identified as a priority population for receiving high-quality, targeted interventions [32]. Cohort studies are needed to further clarify these relationships. Although it is not feasible to prevent online partner-seeking behavior, risk-reduction strategies and services can be strengthened and tailored to reduce harms when such behaviors occur.

This study has some limitations. First, the cross-sectional design limits our ability to determine causal relationships between online partners seeking and other risk behaviors. Second, the peer-driven sampling method may lead to sampling bias and thus fail to represent the real MSM population. To reduce such bias, we made efforts to disseminate recruitment information through various channels to decrease the homogeneity of the respondents. In addition, we identified repeated participants using their mobile phone numbers. Third, data were collected via self-administered online questionnaires, which may have introduced information bias. However, we implemented strict quality control procedures, such as phone follow-ups and questionnaire checks, to help minimize this bias. Future multicenter cohort studies with larger sample sizes are needed to confirm and expand on these findings.

Conclusion

Finding sexual partners online is common among MSM in Zhejiang Province. Most MSM seek partners online more than once per month, and the majority of their partners are younger than 30 years of age. Hotels, lounges, and clubs are the main locations for sexual encounters. MSM with a high school education or below were more likely to report condomless sexual behavior. Those who engaged in insertive anal intercourse only reported higher frequencies of online partner-seeking and a greater number of online partners. Among MSM, there were clear associations between finding sexual partners online and other HIV-related risk behaviors, including finding partners offline, engaging in condomless sex after drinking alcohol, and engaging in condomless sex after using stimulants. These findings underscore the need for targeted, integrated prevention strategies focused on MSM populations who use online platforms to seek sexual partners.

Acknowledgments

The authors thank all participants for their time and effort in contributing to the survey. We also thank the staff of the CBOs and VCT clinics for their assistance with participant enrollment, data collection, and verification. The CBOs involved were Coastal, Love, Glowworm-Light, and Blue-Sky public welfare groups. This research was supported by the Zhejiang Science and Technology Plan for Disease Prevention and Control (Project No.: 2025JK154) and the Zhejiang Key Laboratory of Vaccine, Infectious Disease Prevention and Control. Funding supported the study design and implementation, data collection, and revision of the manuscript.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

XHP contributed to the conceptualization, funding acquisition, and manuscript review. WJC, TTJ, and LC were responsible for data collection, data management, and data analysis. LC and ZRY drafted the manuscript. All authors reviewed and approved the final manuscript.

XSL•FO
RenderX

Conflicts of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Kloek M, Bulstra CA, van Noord L, Al-Hassany L, Cowan FM, Hontelez JAC. HIV prevalence among men who have sex with men, transgender women and cisgender male sex workers in sub-Saharan Africa: a systematic review and meta-analysis. J Int AIDS Soc 2022 Nov;25(11):e26022. [doi: 10.1002/jia2.26022] [Medline: 36419343]
2. Mann LM, Le Guillou A, Goodreau SM, et al. Correlations between community-level HIV preexposure prophylaxis coverage and individual-level sexual behaviors among United States MSM. AIDS 2022 Nov 15;36(14):2015-2023. [doi: 10.1097/QAD.0000000000003343] [Medline: 35876641]
3. Kazemian P, Ding DD, Scott JA, et al. The cost-effectiveness of a resilience-based psychosocial intervention for HIV prevention among MSM in India. AIDS 2022 Jul 15;36(9):1223-1232. [doi: 10.1097/QAD.0000000000003231] [Medline: 35471644]
4. Wu Z, Chen J, Scott SR, McGoogan JM. History of the HIV epidemic in China. Curr HIV/AIDS Rep 2019 Dec;16(6):458-466. [doi: 10.1007/s11904-019-00471-4] [Medline: 31773405]
5. Yuan D, Liu Y, Zhou Y, et al. Men who have sex with men is the high-risk drug resistance population: a meta-analysis of HIV-1 drug resistance profiles and trends in China. J Clin Pharm Ther 2022 Nov;47(11):1729-1737. [doi: 10.1111/jcpt.13772] [Medline: 36117263]
6. Balaji AB, Bowles KE, Le BC, Paz-Bailey G, Oster AM, NHBS Study Group. High HIV incidence and prevalence and associated factors among young MSM, 2008. AIDS 2013 Jan 14;27(2):269-278. [doi: 10.1097/QAD.0b013e32835ad489] [Medline: 23079807]
7. Gorbach PM, Javanbakht M, Shover CL, Bolan RK, Ragsdale A, Shoptaw S. Associations between cannabis use, sexual behavior, and sexually transmitted infections/human immunodeficiency virus in a cohort of young men who have sex with men. Sex Transm Dis 2019 Feb;46(2):105-111. [doi: 10.1097/OLQ.0000000000000919] [Medline: 30640212]
8. Chan PA, Towey C, Poceta J, et al. Online hookup sites for meeting sexual partners among men who have sex with men in Rhode Island, 2013: a call for public health action. Public Health Rep 2016;131(2):264-271. [doi: 10.1177/003335491613100210] [Medline: 26957661]
9. Hu MG, Xu CD, Wang JF. Spatiotemporal analysis of men who have sex with men in Mainland China: social app capture-recapture method. JMIR Mhealth Uhealth 2020 Jan 24;8(1):e14800. [doi: 10.2196/14800] [Medline: 32012086]
10. Pravosud V, Ballard AM, Holloway IW, Young AM. Online partner seeking and sexual behaviors among men who have sex with men from small and midsized towns: cross-sectional study. JMIR Form Res 2022 Jun 10;6(6):e35056. [doi: 10.2196/35056] [Medline: 35687395]
11. Ge R, Chen L, Chen W, et al. Comparison of human immunodeficiency virus-related risky sexual behaviors between men who have sex with men only and men who have sex with men and women: a cross-sectional study in Eastern China. Prev Med Rep 2024 Dec;48:102923. [doi: 10.1016/j.pmedr.2024.102923] [Medline: 39555239]
12. Zhao T, Cao Z, Jiang Z, et al. Providing HIV testing in men who have sex with men through a gay social networking app in China: a qualitative interview study with multisectoral service providers. Front Public Health 2022;10:1056720. [doi: 10.3389/fpubh.2022.1056720] [Medline: 36466506]
13. Acharya S, Karanjkar V, Chougule S, et al. Online partner seeking and high-risk behaviour in men who have sex with men and male-to-female transgendered people in Mumbai, India: implications for prevention strategies. PLoS One 2023;18(4):e0284602. [doi: 10.1371/journal.pone.0284602] [Medline: 37115735]
14. Wu D, Tang W, Lu H, et al. Leading by example: web-based sexual health influencers among men who have sex with men have higher HIV and syphilis testing rates in China. J Med Internet Res 2019 Jan 21;21(1):e10171. [doi: 10.2196/10171] [Medline: 30664490]
15. Bolding G, Davis M, Hart G, Sherr L, Elford J. Where young MSM meet their first sexual partner: the role of the Internet. AIDS Behav 2007 Jul;11(4):522-526. [doi: 10.1007/s10461-007-9224-9] [Medline: 17347876]
16. Tsai JY, Sussman S, Pickering TA, Rohrbach LA. Is online partner-seeking associated with increased risk of condomless sex and sexually transmitted infections among individuals who engage in heterosexual sex? A systematic narrative review. Arch Sex Behav 2019 Feb;48(2):533-555. [doi: 10.1007/s10508-018-1235-2] [Medline: 30155796]
17. Chen L, Chen W, Jiang T, Ni Z, Ma Q, Pan X. The characteristics and risk factors of web-based sexual behaviors among men who have sex with men in Eastern China: cross-sectional study. JMIR Public Health Surveill 2021 Sep 2;7(9):e25360. [doi: 10.2196/25360] [Medline: 34473066]
18. Zhao P, J, Xu W, Wang C. Sexual uses of drug and alcohol among men who have sex with men in China: implications for HIV prevention. BMC Infect Dis 2022;22(1):895. [doi: 10.1186/s12879–022–07880–8] [Medline: 36447161]
19. Simons JS, Maisto SA, Beckius BZ, Kim K. How cannabis and alcohol use influence sexual behavior among men who havesex with men (MSM): a secondary data analysis using ecological momentary assessment. J Stud Alcohol Drugs 2025:1938-4114. [doi: 10.15288/jsad.24–00172]

XSL•FO
RenderX

20. Chen W, Chen L, He L, Chai C. Urban–rural disparity in risky sexual behavior, HIV knowledge, and healthy practices among men who have sex with men: a cross-sectional study in Southeast China. PLoS ONE 2024;19(11):e0312006. [doi: 10.1371/journal.pone.0312006]
21. Kachur RE, Copen CE, Strona FV, Bernstein KT, Furness BW, Hogben M. Factors associated with online sex partners among gay, bisexual and other men who have sex with men: results from a national survey. Sex Transm Dis 2021 Aug 1;48(8):542-546. [doi: 10.1097/OLQ.0000000000001376] [Medline: 34110750]
22. Espinosa da Silva C, Smith LR, Patterson TL, et al. Stigma and web-based sex seeking among men who have sex with men and transgender women in Tijuana, Mexico: cross-sectional study. JMIR Public Health Surveill 2020 Jan 30;6(1):e14803. [doi: 10.2196/14803] [Medline: 32031963]
23. Choi SK, Bauermeister J. A latent profile analysis of online dating patterns among single young men who have sex with men. AIDS Behav 2022 Apr;26(4):1279-1288. [doi: 10.1007/s10461-021-03485-5] [Medline: 34609630]
24. Passaro RC, Castañeda-Huaripata A, Gonzales-Saavedra W, et al. Contextualizing condoms: a cross-sectional study mapping intersections of locations of sexual contact, partner type, and substance use as contexts for sexual risk behavior among MSM in Peru. BMC Infect Dis 2019 Nov 11;19(1):958. [doi: 10.1186/s12879-019-4517-y] [Medline: 31711433]
25. Eshraghian EA, Ferdos SN, Mehta SR. The impact of human mobility on regional and global efforts to control HIV transmission. Viruses 2020 Jan 6;12(1):67. [doi: 10.3390/v12010067] [Medline: 31935811]
26. Zhang L, Shen Z, Yang Y, et al. Status of HIV infection among men who have sex with men in Harbin, 2009–2018. Zhonghua Liu Xing Bing Xue Za Zhi 2021;42(3):538-543. [doi: 10.3760/cma.j.cn112338–20200328–00462] [Medline: 34814426]
27. Nemoto T, Xie H, Iwamoto M, Sakata MH. HIV risk behaviors in relation to psychosocial factors and internet usage among Asian and Pacific Islander (API) men who had sex with men (MSM) in California. AIDS Educ Prev 2020 Apr;32(2):117-136. [doi: 10.1521/aeap.2020.32.2.117] [Medline: 32539477]
28. Wray TB, Monti PM. Characteristics of sex wvents, partners, and motivations and their associations with HIV–risk behavior in a daily diary study of high–risk men who have sex with men (MSM). AIDS Behav 2020;24(6):1851-1864. [doi: 10.1007/s10461–019–02760–w] [Medline: 31832855]
29. Pravosud V, Ballard AM, Holloway IW, Young AM. Latent class analysis of online platforms for partner-seeking and sexual behaviors among men who have sex with men from Central Kentucky. AIDS Behav 2024 Mar;28(3):1015-1028. [doi: 10.1007/s10461-023-04210-0] [Medline: 37824036]
30. Zhu Z, Yan H, Wu S, et al. Trends in HIV prevalence and risk behaviours among men who have sex with men from 2013 to 2017 in Nanjing, China: a consecutive cross-sectional survey. BMJ Open 2019 Jan 30;9(1):e021955. [doi: 10.1136/bmjopen-2018-021955] [Medline: 30705234]
31. Ma Q, Jiang T, Chen W, et al. Awareness of sexual partner's HIV status among men who have sex with men in China: cross-sectional survey study. JMIR Public Health Surveill 2025 Jan 13;11:e66139. [doi: 10.2196/66139] [Medline: 39810334]
32. Mwaniki SW, Kaberia PM, Mugo PM, Palanee-Phillips T. HIV prevalence and associated risk factors among young tertiary student men who have sex with men (MSM) in Nairobi, Kenya: a respondent-driven sampling survey. AIDS Res Ther 2023 Feb 6;20(1):7. [doi: 10.1186/s12981-023-00502-6] [Medline: 36747178]

Abbreviations

aOR: adjusted odds ratio
aPR: adjusted prevalence ratio
CBO: community-based organization
MSM: men who have sex with men
STI: sexually transmitted infection
VCT: voluntary counseling and testing

Edited by A Mavragani; submitted 24.01.25; peer-reviewed by JM Guerras, Z Yang; revised version received 19.08.25; accepted 09.09.25; published 17.10.25.

Please cite as:
Chen L, Yang Z, Chen W, Jiang T, Pan X
Associations Between Online Casual Sexual Behavior and HIV-Related Risk Behaviors Among Men Who Have Sex With Men in Southeast China: Cross-Sectional Study
JMIR Form Res 2025;9:e71690
URL: https://formative.jmir.org/2025/1/e71690
doi:10.2196/71690

XSL•FO
RenderX

© Lin Chen, Zhongrong Yang, Wanjun Chen, Tingting Jiang, Xiaohong Pan. Originally published in JMIR Formative Research (https://formative.jmir.org), 17.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Experiences of Health Research Data Sharing Among Researchers in Sub-Saharan Africa: Cross-Sectional Study

Felix Sukums[1], PhD; Bernard Ngowi[2,3], PhD; Rebecca Chaula[4], MSc; Kim L Weiszhar[5], MSc; Akili Kalinga[6], PhD; Olena Ivanova[7,8], PhD; Clarissa Prazeres da Costa[7,8], PhD; Achim Hoerauf[9], PhD; Sundeep Sahay[10], PhD; Emilia Virginia Noormahomed[11,12], PhD; Linda B Debrah[13], PhD; Chummy Sikasunge[14], PhD; Helena Ngowi[15], PhD; Andrea S Winkler[5,16,17], PhD

[1]Digital Health Unit, School of Public Health and Social Sciences, Muhimbili University of Health and Allied Sciences, 9 United Nations Road, Upanga West, Dar es Salaam, United Republic of Tanzania

[10]Department of Informatics, University of Oslo, Oslo, Norway

[11]Faculty of Medicine, Eduardo Mondlane University (UEM), Maputo, Mozambique

[12]Mozambique Institute for Health Education and Research (MIHER), Maputo, Mozambique

[13]School of Medical Sciences, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana

[14]School of Veterinary Medicine, University of Zambia, Lusaka, Zambia

[15]Department of Veterinary Medicine and Public Health, Sokoine University of Agriculture, Morogoro, United Republic of Tanzania

[16]Department of Community Medicine and Global Health, Institute of Health and Society, University of Oslo, Oslo, Norway

[17]Department of Global Health and Social Medicine, Harvard Medical School, Boston, MA, United States

[2]Mbeya College of Health and Allied Sciences, University of Dar es Salaam, Mbeya, United Republic of Tanzania

[3]Muhimbili Medical Research Centre, National Institute for Medical Research, Dar es Salaam, United Republic of Tanzania

[4]Directorate of ICT, Muhimbili University of Health and Allied Sciences, Dar es Salaam, United Republic of Tanzania

[5]Department of Neurology, TUM University Hospital, and Center for Global Health, School of Medicine and Health, Technical University of Munich (TUM), Munich, Germany

[6]National Institute for Medical Research, Dar es Salaam, United Republic of Tanzania

[7]Institute of Infectious Diseases and Tropical Medicine, LMU University Hospital, LMU Munich, Munich, Germany

[8]German Centre for Infection Research (DZIF), Partner Site Munich, Munich, Germany

[9]Institute of Medical Microbiology, Immunology and Parasitology (IMMIP), University Hospital Bonn, Bonn, Germany

Corresponding Author:
Felix Sukums, PhD
Digital Health Unit, School of Public Health and Social Sciences, Muhimbili University of Health and Allied Sciences, 9 United Nations Road, Upanga West, Dar es Salaam, United Republic of Tanzania

Abstract

Background: Digital platforms play a vital role in improving the availability and access to health research outputs, enhancing the engagement of policy makers and practitioners in the research processes. Despite their potential, it needs to be explored how digital platforms are used to manage and share health research datasets and publications, and to translate research findings among health networks or institutions in sub-Saharan Africa (SSA).

Objective: This study aimed to assess the practices of health research data management, including sharing among researchers and their support staff within 3 large research networks for health innovations in SSA.

Methods: A cross-sectional mixed methods survey was conducted across 3 research networks in SSA, showing experiences of sharing research data using digital platforms among researchers of 3 large research and innovation networks in SSA and affiliated institutions in the Global North. A total of 160 respondents completed a self-administered web-based questionnaire, and following data cleansing, the survey data were analyzed using both descriptive and inferential statistics.

Results: Most respondents (91/160, 56.9%) used electronic data collection tools to collect research data. Almost half (79/160, 49.4%) of the respondents have a digital research data management platform. More than half of the respondents shared their research datasets (102/160, 63.8%), and 61.3% (98/160) shared research findings with the research community through different channels. Furthermore, most respondents shared their research datasets and research outputs through institutional data repositories (42/160, 26.1%), scientific conferences (123/160, 76.9%), and journal articles (110/160, 68.8%). This study found that parameters such as sex, professional category (health professional, information and communication technology professional, and data managers), and the role (researcher or student) influence health research data sharing within the community. The results show

XSL•FO
RenderX

that the roles of the individual have the strongest association with the sharing of research datasets, followed by years of experience in research, then sex, and profession. Females were less likely to share their research datasets than males. Data managers and information and communication technology professionals exchanged datasets less frequently in the professional group, and the researcher's role was statistically significant in sharing research datasets.

Conclusions: This study demonstrates that most researchers share research datasets and outputs through various channels. It was further found that digital platforms were essential in managing and sharing research datasets and publications since more than half (85/160, 53.1%) of the respondents have and use digital platforms. In addition, the study identified factors that influenced researchers' practices of sharing research datasets and publications. Furthermore, key gaps limit the sharing of these research datasets, including inadequate infrastructure, insufficient African dataset sharing platforms, a lack of institutional policy, and limited skills to use available platforms.

(JMIR Form Res 2025;9:e69411) doi:10.2196/69411

KEYWORDS

health research data; research data management; research data sharing; health research network; ICT; digital research data repositories; sub-saharan Africa; information and communication technology

Introduction

The World Health Organization (WHO) defined knowledge translation as the exchange and application of knowledge by relevant stakeholders to accelerate the benefit of global and local innovation in strengthening health systems and improving people's health [1]. Research findings need timely translation into practice and policy [2,3]. Translation of research findings may lead to evidence-based practice guidelines, adoption of technologies, or health care delivery models with proven value, and a change in practice and behaviors [3,4]. Thus, there is a need to establish substantial programs of health research data sharing for improvement in health care delivery and policy formulation [4].

Data sharing and its significance connect to research findings. Findable, accessible, interoperable, and reusable (FAIR) principles are the foundation for effective data sharing [5]. FAIR guiding principles for scientific data management support a vision for data sharing by becoming more findable, accessible, interoperable, and reusable, whereby these principles were intended to apply to many kinds of digital assets or content types [6,7].

Several systematic reviews have assessed barriers and facilitators to the use of research evidence in health decision-making [8-11], pointing out the need for timely access to good quality, relevant, and applicable research evidence, collaborations with decision makers, and building relationships and skills. Interventions to boost the translation of research findings into policy and practice include formats such as evidence briefs, issue briefs, policy briefs, press releases, web portals, consultations, deliberative dialogs, and workshops [12]. Literature shows that one of the best practices for research data sharing within the scientific field is using digital data repositories [13-15].

Data sharing platforms are systems that enable the collection, storage, and reuse of research datasets across all disciplines. Despite global progress in the development of digital platforms for research data sharing, sub-Saharan Africa (SSA) remains significantly underrepresented in the open science ecosystem. Existing platforms, such as CKAN, Zenodo, Figshare, Dryad, and Dataverse, have played a pivotal role in facilitating open access to research data by enabling data publication, discoverability, and reuse through standardized metadata, persistent identifiers, and open licensing models [16,17]. Collaborative environments, like GitHub and Kaggle, also support data sharing alongside reproducible analytics and community engagement. Furthermore, initiatives such as OpenAIRE and the European Open Science Cloud illustrate how regional efforts can institutionalize open science practices [18]. However, in the context of SSA, participation in and benefit from such platforms remain limited due to infrastructural barriers, uneven internet access, limited institutional support for data management, and capacity constraints in data stewardship [8,19]. While platforms like the African Open Science Platform and the DataFirst repository at the University of Cape Town signal regional progress, they remain relatively few and often underresourced [20,21]. This imbalance restricts the visibility, accessibility, and global integration of African research outputs, exacerbating existing inequalities in knowledge production and dissemination. Therefore, addressing these disparities through investment in regional platforms, policy alignment, and capacity building is essential to ensure equitable participation in the global open science movement.

Despite global advances in open science, SSA continues to face significant barriers to research data sharing. High-income countries benefit from strong digital infrastructure, FAIR data policies, and established repositories that promote open access and reuse [5,22]. In contrast, the African region struggles with poor connectivity, limited funding, inadequate data stewardship skills, and fragmented policies [19,23,24]. Concerns around data ownership, language diversity, and lack of awareness further hinder participation [8,25,26]. While platforms like European Open Science Cloud, Zenodo, and Dryad support global data sharing [27,28], similar efforts in Africa, such as African Open Science Platform and DataFirst, remain underresourced. Bridging this gap requires investment in infrastructure, harmonized policies, and capacity building tailored to local contexts.

Even though data sharing in science is not yet practiced routinely [29], there has been an increasing emphasis on the governance and storage of research data in recent times. Accordingly, research stakeholders and funders more often recognize the

XSL•FO
RenderX

value of scientific data [29,30]. Open Science, including Open Data, is focused mainly on promoting the sharing of scientific research work among stakeholders [15,29-32]. It is being advocated to ensure open access to research data for further study and for informed decision-making among stakeholders, including policy makers, practitioners, librarians, data managers, and research funders. Furthermore, the WHO emphasizes that the global scientific community promotes the culture of sharing datasets even before publication to validate the findings and data reuse, as well as early translation of research results for health practices and policy [1]. Consequently, new guidelines entitled "Sharing and reuse of health-related data for research purposes: WHO policy and implementation guidance" were published in 2022 [1]. The European Commission has also been actively promoting making research data more accessible for years. A prominent and successful example of this endeavor is the European COVID-19 data platform launched in 2020 to combat the pandemic worldwide [33].

Data sharing is being advocated to be in electronic formats for easier management and accessibility [34]. Digital platforms or data repositories help create online spaces within which the producers and users of information can share resources and learn [8,26]. Research data shared among stakeholders includes various types and formats. These consist of quantitative and qualitative datasets, such as survey responses, clinical data, patient records, and demographic surveillance, along with real-time data from human participants, health system data, epidemiological data, and human biological specimens. Shared data may also comprise images, technical reports, financial documents, models, and digital health data from wearables and eHealth apps [7,17,25,26,35].

Despite the rich potential of platforms to enable the sharing of research findings between researchers, policy makers, and practitioners, their value has not been effectively realized, particularly in the context of health research in low- and middle-income countries (LMICs). Many issues surrounding digital scientific data sharing, such as legal, ethical, cultural, and intellectual rights, remain unknown. New problems and questions are emerging as the old ones are increasingly addressed, but technical challenges still hinder open data sharing on a broader scale.

Leading publishers, such as Springer Nature, Elsevier, Taylor and Francis, and Wiley, have introduced and started encouraging authors to provide quality research datasets through the publisher's digital platforms or other publicly available platforms, for example, a general-purpose data repository for easier accessibility of the data to a broad audience [34,36,37]. In 2016, Wiley surveyed 4600 authors from 112 countries to get insights into research data sharing. The survey indicated that 69% (n=3174) of the authors shared data. Barriers to data sharing included intellectual property or confidentiality issues, ethical concerns, fear of data misinterpretation or misuse, and fear of scooping [37,38]. Only 92 respondents (authors) were from SSA, and many countries, such as Tanzania and Ghana, were not represented.

Overall, as transparent data sharing has become increasingly a requirement for good scientific practice at different levels, little is known about how digital platforms have recently been used to manage and share health research data and publications within SSA.

Health research consortia, such as CYSTINET-Africa, TAKeOFF, and TB Sequel, are spearheading efforts to co-develop a best-practice model for the early digital translation of disease-related research data across 5 health innovation networks in SSA. Simultaneously, leading academic publishers—including Springer Nature, Elsevier, Taylor & Francis, and Wiley—have begun requiring authors to submit quality-assured research datasets via their own digital platforms or through publicly accessible repositories to enhance data discoverability and reuse [11-13]. A 2016 Wiley survey involving 4600 authors from 112 countries revealed that while 69% (3174/4600) of respondents shared data, significant barriers remained, such as intellectual property concerns, ethical restrictions, fears of misinterpretation, and risks of being scooped [13]. Notably, only 92 of those respondents were from SSA, with key countries such as Tanzania and Ghana unrepresented. This highlights a pressing need to assess and address regional gaps. Accordingly, this study aims to examine existing practices and future needs regarding the timely sharing of research data related to infectious diseases, neglected tropical diseases, noncommunicable diseases, and One Health across subnational and national levels within the 5 targeted research networks. The findings will inform the development of long-term, scalable digital data-sharing prototypes that are aligned with the needs of researchers, policy makers, and stakeholders. These platforms are intended to enhance accessibility to research outputs and facilitate the integration of evidence into policy and practice, while ensuring alignment with both user requirements and reporting mandates at national and subnational levels.

This study aimed to assess the practices of health research data management, including sharing among researchers and their support staff within 3 large Research Networks for Health Innovations in SSA funded by the German Federal Ministry of Education and Research (BMBF). We specifically assessed the types of data shared, the methods of sharing, the timing of sharing, and the recipients of the data. In addition, the study assessed how researchers disseminated research data for reuse by other researchers and how they shared research outputs to facilitate translation for informing policy practice.

Methods

Study Setting

The respondents of this study were researchers, including postgraduate students, and research support staff of 3 BMBF-funded Research Networks for Health Innovations in SSA, namely CYSTINET-Africa, TAKeOFF, and TB Sequel [39]. The partner institutions are located in different SSA countries (Figure 1). Participating institutions from Germany were also included in the study. Partner institutions conducted research in various areas, including infectious diseases, neglected tropical diseases, noncommunicable disease, One Health, and digital health.

XSL•FO
RenderX

Figure 1. Geographical locations of the supported health networks [39]. TB: tuberculosis.

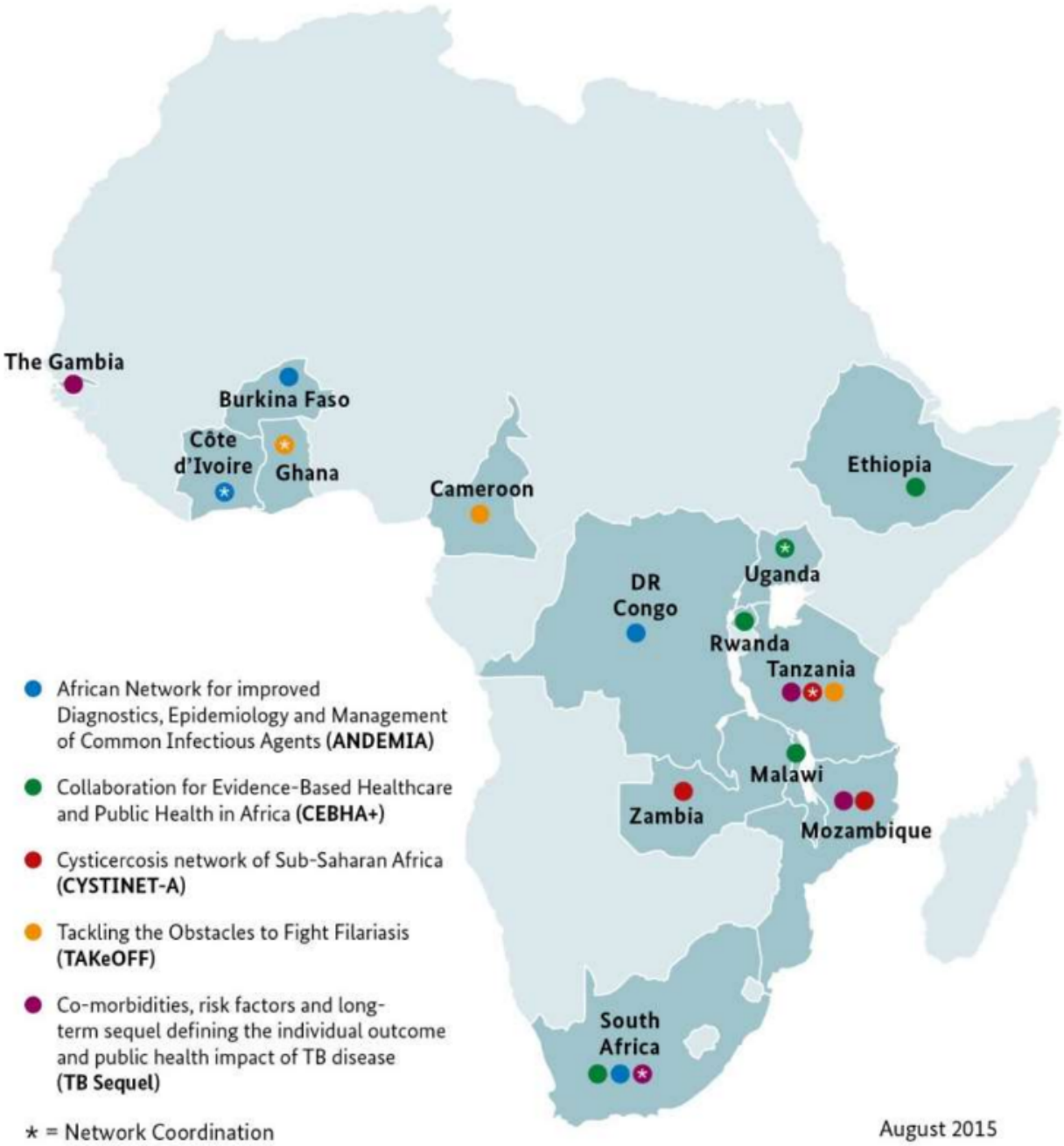

Study Design

This study used an online cross-sectional design to assess the experiences of researchers and support staff with research data management and sharing at subnational and national levels and to determine factors hindering the use of digital platforms for this purpose within SSA's 3 health networks.

Ethical Considerations

Ethical clearance was sought from the National Health Research Ethics Review Committee of the National Institute for Medical Research (NIMR) in Tanzania (ethical clearance certificate NIMR/HQ/R.8a/Vol.IX/3836). Verbal permission to implement this research was obtained from the coordinators of each research network. Respondents to the self-administered internet-based questionnaire through Google Forms agreed to participate in this study through informed consent, which was attached as the first section of the questionnaire. Respondents were asked to fill in the questionnaire after reading and accepting the contents of the consent form. No personally identifiable information about respondents was collected. Research data were kept secure and confidential by 2 researchers (FS and BN) and 1 data manager at NIMR in Tanzania. There was no compensation provided to participants.

Questionnaire Development

The questionnaire used in this study was developed through a structured and iterative process designed to capture relevant practices, challenges, and needs related to research data sharing across health research networks in SSA. The development started with a review of existing literature and validated instruments on data sharing practices, open science frameworks,

XSL•FO
RenderX

and digital health governance. Key domains were identified and aligned with the study objectives, covering topics such as research data management, research data management platforms, dissemination of research data, general needs and requirements for data sharing platforms, and stakeholder engagement. Experts from the participating research networks—CYSTINET-Africa, TAKeOFF, and TB Sequel—were consulted to ensure contextual relevance and content validity. A draft version of the questionnaire was then pretested with a small sample of 15 researchers from the target networks to assess clarity, comprehensiveness, and length. Based on feedback from the pilot, minor revisions were made to improve wording, flow, and response options. The final instrument included both closed- and open-ended questions and was administered in English through an online platform to facilitate broad participation. The complete questionnaire is provided in Multimedia Appendix 1 for reference and transparency.

Data Collection Tools and Techniques

A self-administered web-based questionnaire was developed using Google Forms and shared with study participants for completion. Data were collected using a nonprobability sampling method, in which participants were directly contacted and referred to the research network team by sharing the link. Each of the network coordinators provided the contact information from the network. The duration of data collection was initially planned for 2 months (September and October 2021) but was later extended to December 2021 due to delays in responses. A reminder message was sent every Friday of the week. The questionnaire was pretested among 15 researchers outside the 3 research networks to determine the ease of completing it and to ensure that the questions were clearly understood.

The experience gained from the pretest was used to improve the tools. Web-based questionnaire administration enables access to large and diverse populations and is an inexpensive, quick, and convenient approach to collecting data [40-42]. Furthermore, the questionnaire had mandatory fields for quality checks to ensure that all relevant fields were filled in before the data were submitted to a database.

The questionnaire explored practices of sharing raw research datasets and outputs to facilitate the early translation of research into policy and practice within the 3 networks. The questionnaire also examined the existence and use of digital platforms within the 3 health research networks to improve the early translation of research findings into practice and policy. The questionnaire collected researchers' sociodemographic data, explored the existence of data or information systems and data sharing practices within the research projects, and assessed experience using research data management digital platforms and disseminating research data, including anonymized raw research datasets and outputs.

Data Analysis

The quantitative data were transferred from the Google Forms server onto a Microsoft Excel file to verify data completeness and consistency and liaise with respective network coordinators. After extracting data from Google Drive, it was stored in a secure computer with a password, ready for cleaning and analysis. Researchers assessed and verified the data for quality before compiling it for analysis. Frequency tables were prepared to assist in data cleaning. The software developed by IBM, SPSS version 23.0, was used for univariate and bivariate analyses. Differences between categories were measured using chi-square or Fisher exact tests, and binary logistic regression was used to determine the factors that influence the sharing of research datasets.

After data collection, several variables were transformed to enhance analytical clarity and improve the robustness of inferential statistics. These transformations included recoding research network membership, categorizing age groups according to WHO standards, and standardizing variables such as sex, education level, professional role, and years of experience.

We explored factors (variables) that influence the sharing of research datasets and publications through different channels among the respondents. The binary logistic regression (crude odds ratio [cOR] and adjusted odds ratio [OR]) was used in this study since the dependent variable was dichotomous, involving 2 choices: whether the respondents shared their research datasets or outputs. In this study, the independent variables were all categorical.

Measure the Importance and Associations of the Categorical Variables

Cramer V measures associations and importance for the categorical variables. The statistics range from 0 to 1, where 0 indicates no association, and 1 indicates a perfect association. A larger value for Cramer V suggests a stronger association between variables. It is derived from the chi-square test and provides a value that shows the strength of associations between 2 variables. In our case, we measure the importance of the independent variables, which include sex, profession, roles, and years of experience [43].

Results

Overview

Table 1 illustrates the sociodemographic characteristics of the respondents. The majority (98/160, 61.3%) were from 3 health innovation research networks of SSA, of which 31.3% (50/160) were from CYSTNET-Africa, and 63.1% (101/160) worked at universities. Slightly more than half were aged 40 years or younger (83/160, 52.2%), the vast majority resided in East African countries (130/160, 81.3%), and over half (108/160, 67.5%) had 1 to 10 years of experience in research-related work. Half of the respondents were researchers (81/160, 50.6%), and the highest education level was almost equal between Master's and PhD levels (64/160, 40.1% and 77/160, 48.1%, respectively).

XSL•FO
RenderX

Table . Sociodemographic characteristics of the respondents (n=160).

Sociodemographic characteristics	Frequency, n (%)
Research network[a]	
The 3 networks	98 (61.3)
Others	62 (38.8)
Institution category	
University	101 (63.1)
Research Institution	35 (21.9)
Others	24 (15)
Age group (y)	
21 - 30	17 (10.7)
31 - 40	66 (41.5)
41 - 50	56 (35.2)
51 and above	20 (12.6)
Sex	
Female	58 (36.3)
Male	101 (63.1)
Preferred not to mention	1 (0.6)
Highest level of education	
PhD degree	64 (40)
Master's degree	77 (48.1)
Bachelor's degree	19 (11.9)
Research experience (years)	
1 - 5	71 (44.4)
6 - 10	37 (23.1)
11 and above	52 (32.5)
Professional category	
Health researchers	114 (71.3)
ICT[b] officers and data managers	27 (16.9)
Others	19 (11.9)
Roles	
Researchers	81 (50.6)
ICT and data managers	30 (18.8)
Postgraduate (PhD and Masters)	35 (21.9)
Others	14 (8.8)
Types of research data collected in the research platforms[a]	
Human health data	139 (86.9)
Animal health data	40 (25)
Environmental health data	27 (16.9)
Other-not specified data	12 (7.5)
Sources of research data[a]	
Community-based sources	115 (71.9)
Health facilities	108 (67.5)

XSL•FO
RenderX

Sociodemographic characteristics	Frequency, n (%)
Other sources	48 (30)
Means of collecting primary and secondary research data[a]	
Using paper-based forms	51.9 (31.9)
Using electronic data collection tools	91 (56.9)

[a]Multiple options applied.
[b]ICT: information and communication technology.

The study indicated that different types of research data are collected from various sources, including, but not limited to, communities, health facilities, villages, and subnational and national levels, covering a wide range of health domains. Both paper and electronic data collection and management tools were used. Most respondents (91/160, 56.9%) used electronic data collection tools to collect research data.

Frequency of Sharing Research Outputs among Respondents

In total, 53.6% (112/209) indicated that they share or disseminate their research findings for early translation of research findings as per the project plan. In comparison, 41.6% (87/209) indicated that they would share their research output when the opportunity arises. Also, in 76.9% (123/160) and 68.8% (110/160) of the cases, participants responded to sharing their research outputs for early translations of research findings through scientific conferences and journal articles, respectively, as illustrated in Table 2, on which different means of sharing research datasets were reported. Furthermore, 26.1% (61/234) and 12% (28/234) of the respondents used institutional data repositories and cloud file sharing as the channel to share research datasets, as illustrated in Table 3.

Table . Proportion of respondents and their different channels of sharing research datasets.

Channels	Frequency, n (%)
Institutional data repository	61 (26.1)
Funder data repository	19 (8.1)
Discipline-specific data repository	13 (5.6)
Journal data repository	28 (12)
As supplementary material in a journal	25 (10.7)
External repository or general-purpose data repository (eg, Figshare, Zenodo)	7 (3)
Cloud file sharing (eg, Dropbox, Google Drive, and Blog or website)	27 (11.5)
I do not share research data beyond my fellow researchers	54 (23.1)

Table . Proportion of respondents and their different channels of sharing research outputs.

Means	Frequency, n (%)
Project website	44 (7.2)
Social media channels	45 (7.3)
Newsletters	39 (6.4)
Scientific conferences	123 (20)
Journal articles and other publications	110 (17.9)
Policy briefs	50 (8.1)
Feedback meetings with policy makers	49 (8)
Feedback meetings with health professionals	60 (9.8)
Feedback meetings with community members	51 (8.3)
Reports submitted to the ministry, region, and district officials	43 (7)

Furthermore, access to the existing digital research platforms to a wider audience was also reported as one of the limitations. About one-third (53/169, 31.4%) of the respondents indicated they could share their research datasets mostly with their fellow researchers within their research institution or research networks (32/169, 18.9%), as presented in Table 4.

XSL•FO
RenderX

Table . Stakeholders who had access to research datasets through different platforms (n=160).

Sharing raw research datasets	Frequency, n (%)
Other researchers within their (own) research institution	53 (31.4)
Other researchers within the (own) research network	32 (18.9)
Other researchers in the other health innovation research networks	8 (4.7)
Other researchers outside their institutions and the health innovation research networks	13 (7.7)
Health professionals or practitioners in general	17 (10.1)
Policy makers at the subnational level	10 (5.9)
Policy makers at the national level	14 (8.3)
Policy makers at the international level	5 (3)
Own research funders	17 (10.1)

Researchers were asked if their research project had a data transfer or sharing agreement or policy with other partners. The results show that 100% (N=160) of the ANDEMIA and TAKeOFF have sharing agreements with the partners of their research network, followed by CYSTINET-Africa (13/15, 86.7%), TB Sequel (7/9, 77.8%), and CEBHA+ (2/4, 50%). Also, research institutes (19/23, 82.6%), European countries (N=9, 100%), and South Africa (N=4, 100%) only share their data with partners within their research network.

More females share their data through their research partners than males. Graduates with experience between 1 and 5 years (17/19, 89.5%) share their research project through their partners in the research network.

We analyze the availability of data-sharing agreements among the research projects (Table 5). There was cross-sharing between networks and third parties through data-sharing agreements. However, there is no statistically significant difference in the availability of data-sharing agreements among the 3 groups, as indicated in Table 5.

Table . Availability of data sharing agreements among the research projects.

Independent variables	Availability of data sharing agreement with (n=61)			Fisher exact test
	Partners of their research network, n (%)	Partners of any of the health innovation research networks, n (%)	Third-party, n (%)	
Research network				0.11
3 networks	33 (84.6)	7 (17.9)	8 (20.5)	
Others	14 (63.6)	8 (36.4)	4 (18.2)	
Institution category				0.35
University	24 (66.7)	6 (16.7)	6 (16.7)	
Research Institute	19 (65.5)	5 (17.2)	5 (17.2)	
Others	4 (44.4)	4 (44.4)	1 (11.1)	

Practices in Sharing Research Datasets and Publications Among Respondents

Table 6 shows respondents' experiences in sharing research datasets and research outputs and the availability of digital research data management platforms. Almost half (79/160, 49.4%) of the respondents reported that they had digital research data management platforms, which include Afyadata, KoboCollect, DHIS2, DSpace, REDCap, and Open Data Kit, 63.8% (102/160) were sharing research datasets, and 46.3% (74/160) were sharing research publications via the channels. The presence of digital research data management platforms showed a statistically significant difference between the institution categories (P=.013), in which 51.9% (41/79) of the respondents from universities reported using digital research data management platforms. Furthermore, statistically significant differences in sharing research datasets and publications were found in respondents' sex, professional category, years of research experience, and roles.

XSL•FO
RenderX

Table . Respondents' characteristics in the context of management and sharing of research datasets.

Independent variables	Shared research datasets (n=102)		Shared research findings (n=98)		Had digital research data management platforms (n=79)	
	n (%)	*P* value	n (%)	*P* value	n (%)	*P* value
Research network		.40		.51		.27
3 networks	60 (58.8)		62 (63.3)		45 (57)	
Others	42 (41.2)		36 (36.7)		34 (43)	
Institution category		.89		.07		.01
University	63 (61.8)		63 (64.3)		41 (51.9)	
Research Institution	23 (22.5)		25 (25.5)		23 (29.1)	
Others	16 (15.7)		10 (10.2)		15 (19)	
Age group (y)		.10		.11		.44
21 - 30	9 (8.8)		10 (10.3)		11 (14.1)	
31 - 40	38 (37.3)		36 (37.1)		35 (44.9)	
41 - 50	43 (42.2)		34 (35.1)		23 (29.5)	
51 and older	12 (11.8)		17 (17.5)		9 (11.5)	
Sex		.01		.25		.95
Female	30 (29.4)		32 (33)		29 (36.7)	
Male	72 (70.6)		65 (67)		50 (63.3)	
Highest level of education		.10		.15		.08
PhD	47 (46.1)		45 (45.9)		32 (40.5)	
Masters	45 (44.1)		43 (43.9)		34 (43)	
Graduate	10 (9.8)		10 (10.2)		13 (16.5)	
Professional category		.04		.79		.38
Health professionals	69 (67.6)		69 (70.4)		53 (67.1)	
ICT[a] and data managers	16 (15.7)		16 (16.3)		16 (20.3)	
Others	17 (16.7)		13 (13.3)		10 (12.7)	
Roles		.001		.20		.44
Researchers	61 (59.8)		54 (55.1)		43 (54.4)	
ICT and data managers	14 (13.7)		19 (19.4)		17 (21.5)	
Postgraduates(PhD and Masters)	16 (15.7)		16 (16.3)		13 (16.5)	
Others	11 (10.8)		9 (9.2)		6 (7.6)	
Research experience (y)		.03		.54		.31
1 - 5	38 (37.3)		41 (41.8)		31 (39.2)	
6 - 10	29 (28.4)		22 (22.4)		22 (27.8)	
11 and above	35 (34.3)		35 (35.7)		26 (32.9)	

[a]ICT: information and communication technology.

XSL•FO
RenderX

By using the Cramer V test, it shows that sex had a Cramer V coefficient of 0.20, indicating a strong association between sex and the datasets sharing and publications; profession had a coefficient of 0.20 and also had a strong association between the profession and the datasets sharing and publications; roles had a Cramer coefficient of 0.30, indicating a very strong association between roles and the datasets sharing and publications compared with sex and profession. Years of experience had a Cramer coefficient of 0.21, suggesting a strong association between years of experience and sharing research datasets and publications. So, based on Cramer V coefficient, the most important variables in association with the outcome variable are roles (0.30), years of experience in research (0.21), sex (0.20), and profession (0.20).

Factors Influencing Research Data Sharing via Digital Research Channels or Data Management Platforms

Table 7 illustrates the logistic regression analysis to determine factors influencing whether respondents shared their research datasets. As presented in Table 7, sex, professional category, and roles significantly influence dataset sharing among respondents.

XSL•FO
RenderX

Table . Logistic regression analysis on factors influencing the sharing of research datasets.

Variables	cOR[b] (95% CI)	*P* value	aOR[c] (95%)	*P* value
Research network				
3 networks	0.752 (0.385 - 1.469)	.40	1.063 (0.464 - 2.439)	.89
Others	Reference	Reference	Reference	Reference
Institution category				
University	0.829 (0.324 - 2.121)	.70	0.656 (0.198 - 2.174)	.49
Research Institute	0.958 (0.319 - 2.876)	.94	1.123 (0.293 - 4.296)	.87
Others	Reference	Reference	Reference	Reference
Age group (y)				
21 - 30	0.750 (0.203 - 2.770)	.67	1.970 (0.287 - 13.544)	.49
31 - 40	0.905 (0.327 - 2.507)	.85	1.486 (0.328 - 6.727)	.61
41 - 50	2.205 (0.742 - 6.550)	.16	2.921 (0.746 - 11.429)	.12
51 and older	Reference	Reference	Reference	Reference
Sex				
Female	0.432 (0.220 - 0.845)	.01	0.385 (0.160 - 0.927)	.03
Male	Reference	Reference	Reference	Reference
Education				
PhD	2.488 (0.864 - 7.165)	.09	3.904 (0.826 - 18.459)	.09
Masters	1.266 (0.462 - 3.469)	.65	2.658 (0.710 - 9.949)	.15
Graduate	Reference	Reference	Reference	Reference
Professional category				
Health professionals	0.180 (0.040 - 0.819)	.03	0.108 (0.019 - 0.614)	.01
ICT[a] and data managers	0.171 (0.033 - 0.895)	.04	0.244 (0.038 - 1.578)	.14
Others	Reference	Reference	Reference	Reference
Roles				
Researchers	0.832 (0.211 - 3.283)	.79	1.011 (0.178 - 5.743)	.99
ICT and data managers	0.239 (0.055 - 1.032)	.06	0.165 (0.026 - 1.043)	.06
Students	0.230 (0.054 - 0.969)	.05	0.239 (0.042 - 1.361)	.11
Others	Reference	Reference	Reference	Reference
Research experience (y)				
1 - 5	0.559 (0.266 - 1.177)	.13	0.814 (0.245 - 2.704)	.74
6 - 10	1.761 (0.665 - 4.663)	.26	1.748 (0.441 - 6.923)	.43
11 and more	Reference	Reference	Reference	Reference

[a]ICT: information and communication technology.

[b]cOR: crude odds ratio.

[c]aOR: adjusted odds ratio.

Before and after adjusting for additional parameters, as illustrated in Table 7, sex has a statistically significant impact on the sharing of research datasets. Specifically, females are 56.8% less likely to share their datasets than males (cOR 0.432, CI 0.220 - 0.845; *P*=.01). This difference persists even after adjusting for OR of 0.385 (CI 0.160 - 0.927; *P*=.03).

Health professionals are significantly less likely than others to share their research datasets, even after controlling for other variables, according to a study on the professional category that is statistically significant before and after adjustment for OR 0.180 (CI 0.019 - 0.614; *P*=.01).

There is statistical significance in students' probabilities of sharing their study datasets compared with other respondents (OR 0.23, CI 0.054 - 0.969; *P*=.05).

The OR increased after controlling for other factors, leading to a 60.3% higher likelihood of sharing the datasets in 3 networks

XSL•FO
RenderX

than others (OR 1.603, CI 0.464 - 2.439; *P*=.89). However, the *P* value indicates no significant difference in both cOR and adjusted OR.

The institutional category (university, research institutions, and others) shows a statistically insignificant association in sharing research datasets. After adjustment, the values (OR 0.656, CI 0.198 - 2.174; *P*=.49) indicate a 34.4% lower likelihood of dataset sharing.

Discussion

Principal Findings

Our results show that different types of data are mainly collected using electronic tools. Most data collected for research platforms are human health data from health facilities and the community. The study investigated the experiences of researchers sharing research data using digital platforms in 3 large research and innovation networks of SSA and affiliated institutions. It also explored the sharing of research datasets, the availability of data sharing agreements among the research projects, and the practices in sharing research datasets and publications among respondents under SSA's 3 health innovation research networks.

Respondents used various digital platforms, such as KoboCollect and REDCap, for data management. The presence of such platforms was more common in universities (*P*=.01), which may have more advanced research infrastructures than other institutions. The effectiveness of these tools in managing and sharing large datasets has been widely documented, especially in global health research.

The results show that most of the researchers from the networks have agreements or policies for sharing research datasets with their partners. A study conducted by Waithira et al [21], on data management and policy, emphasizes that each institution should have a policy and agreement to ensure ethical and smooth data sharing and promote the sharing of datasets among researchers. From the independent variables, the role of the researcher was more important in sharing their datasets, followed by years of experience in research. However, it did not show statistical significance then, followed by the profession of the researchers, and finally, sex.

Practice in Data Sharing

The results show that different people share research datasets and outputs in different ways. Researchers share data at different stages of the research process—both during and after research completion—through various channels, such as institutional repositories, journal publications, and scientific conferences.

Of the respondents, approximately half (112/209, 53.6%) communicated research findings in accordance with project plans, and a comparable percentage (87/209, 41.6%) shared their research outputs when possibilities presented themselves. The main means of disseminating research outputs were conferences (123/160, 76.9%) and journal articles (110/160, 68.8%), with 26.1% (61/234, due to multiple responses) and 33.8% (61/180) of respondents using institutional repositories and cloud platforms, respectively. Unlike in our study, a study by Khan and colleagues [7] found that data sharing and reuse were more common among experienced researchers. Their findings indicated that researchers with previous experience in reusing data were more likely to share their own data than those without such experience. On the contrary, Tenopir et al [44] showed differences among age groups, with younger respondents feeling more positively about data sharing and reuse yet providing less of their data than older respondents.

Also, most researchers use institutional repositories to share their datasets [7]. About one-third (53/169, 31.4%) of respondents said they had access to digital research platforms within their institutions, indicating a lack of widespread availability.

Ethical and responsible data sharing requires appropriate agreements, particularly for cross-border data flows, to ensure compliance with national and international data protection laws [25,26]. Our study assessed the availability of data-sharing agreements across research projects. While data were shared between networks and third parties through such agreements, some projects lacked formal agreements. However, no statistically significant difference was found in the availability of agreements and data-sharing practices across the 3 networks.

Data-sharing behaviors are influenced by various major factors, as indicated by the logistic regression study. One important effect was gender: women were less likely than men to share their datasets (OR 0.38, *P*=.015). The author believes that women may be encountering previous challenges in data sharing that could discourage them. This was also observed in another study where there was a significant interaction between gender and age, whereby there was a gender difference in confidence in privacy protection [44]. A professional group was also significant, with data managers and information and communication technology (ICT) exchanging datasets less frequently (OR 0.12, *P*=.012). Compared with other jobs, the researcher's role was associated with a substantial decrease in student dataset sharing (OR 0.16, *P*=.034). Interestingly, years of research expertise had no significant impact on data-sharing behaviors. The study conducted by Piwowar [45] also indicates that few researchers are still sharing their research datasets. They only share their research datasets if they have experience in doing so. Also, the study shows that researchers in health subjects were least likely to share their datasets.

Associated Factors Influencing Data Sharing

Data sharing has become an imperative in every sector. However, most organizations still maintain data silos, exercising unnecessary restrictions on data access and discouraging data sharing [46]. In times of COVID-19, disease surveillance and timely sharing of large digital datasets have taken center stage. For instance, as data sharing was deemed crucial for tackling the pandemic, the WHO and its partners established a global coalition on data sharing. The global coalition aimed to facilitate and promote effective, ethical, and equitable data sharing across geographies and disciplines for various purposes [47]. Furthermore, health research data sharing is being promoted to strengthen academic research, inform health policy and practice, and strengthen the integrity of the clinical trial system. The American politician and senior United States senator from Massachusetts, Elizabeth Warren, said

XSL•FO
RenderX

> *When researchers have access to complete data, they can answer new questions, explore different lines of analysis, and more efficiently conduct large-scale analyses across trials.* [Elizabeth Warren 48]

The Senator also emphasized the need to publicly share data that produced null, inconclusive, or negative results. Data sharing has the potential to strengthen the practice of medical research.

The findings of this study contribute to the body of knowledge about health research data sharing in SSA. The results present experiences of health research data sharing from 15 African countries. It is interesting to note that slightly more than half of the respondents were using digital tools to collect research data from varied sources, including community and health facilities. Only about half (79/160, 49.4%) of the respondents indicated that they had digital research data management platforms, which means they lack essential foundations for enabling proper research data management and sharing.

Access to Digital Research Platforms

Limited access to digital platforms for sharing research datasets was reported. The most common sharing was among colleagues within the same research institutions (53/169, 31.4%), while access to wider audiences, including policy makers at national and international levels, was significantly lower (19/169, 17.2% and 5/169, 3%, respectively). These results underscore the need for improved digital infrastructure and policies to enhance data accessibility. Studies from other regions highlight similar barriers to the widespread use of digital platforms for data dissemination.

Factors Influencing Electronic Data Sharing

Several factors were identified as influencing data-sharing practices. Researchers with less experience (1 - 5 y) were less likely to share datasets than those with more experience. This aligns with previous findings that indicate early-career researchers may face more barriers in data sharing, such as a lack of resources or networks. In addition, universities showed a higher likelihood of using digital research data management platforms (41/79, 51.9%) compared with research institutions, highlighting the disparity in digital capacity between different types of institutions.

The need to promote digital platforms for simplified research data management and sharing is evident, as about one-third of the respondents still use manual or paper-based tools. This may be due to limited knowledge and skills in using digital platforms, as well as having no or unreliable internet connections for uploading data [10,26,38,49]. The fear of data loss on digital platforms could also account for this. Furthermore, only about two-thirds were sharing research datasets. These findings correspond with the global trends in research data sharing, indicating that only 60% (n=2760) of more than 4600 Wiley authors from 112 countries participated in a survey conducted in 2016 and shared their research data and outputs [37].

Furthermore, in our study, some researchers did not share health research datasets (disaggregated anonymized or pseudonymized or analyzed data). Similarly, findings indicated that only 4% (n=119) of 2972 of the Directory of Open Access Repositories contain research datasets, while the rest host publications and related materials [50]. This could be due to researchers' reluctance to share their research data with the public.

A study in Jordan reported that slightly more than half of the researchers (93/175, 53.1%) were unwilling to share their data with the public [51]. Another study reported that 80% (407/507) of researchers were willing to share their research data [52], while in another study, 54% (133/250) of the respondents used research repositories for data sharing [53]. In this study, most respondents indicated that they shared their research outputs through journal articles and scientific conferences, which may not be easily accessible to targeted users (eg, health practitioners and policy makers) for early translation to inform practice and policy. Interestingly, about one-third share their outputs through feedback meetings with the stakeholders.

Studies report diverse research data sharing experiences in LMICs [54]. Numerous challenges have been pointed out to hinder effective and efficient research data sharing and reuse in all economies [37,54,55]. Some of the challenges reported included limited privacy issues or protection, lack of regulations, heterogeneity and fragmentation of data sources, data storage limitations, incompatible data formats, political concerns, insufficient data sharing skills, and costs associated with data preparation and sharing [37,51,54].

Several factors influence stakeholders' practices in sharing research data with different beneficiaries. These factors include the availability of robust digital platforms for data management and sharing, digital skills for data deposition and adaptation, incentives and motivations, organization support for data sharing, and legal and policy frameworks supporting data sharing. Furthermore, the willingness of researchers to contribute to ensuring the timely sharing of their research data [34,36,37].

There are, however, some challenges or barriers to the early sharing of research data or outputs for translation of research outputs. For example, reluctance to publish or share data was reported despite efforts for open data or open access, for example, journals emphasize data availability requirements when submitting their articles. Studies also reported challenges affecting the sharing of research data, including ethics and confidentiality concerns, fear of data misuse by people who had no or little contributions, lack of incentives and motivations, limited knowledge and skills for data management, and sharing through different platforms as well as limited financial support to data management and sharing. Furthermore, there are no clear guidelines and plans on health research data management, sharing, publications, authorships, and peer review of research datasets [44,53,54,56].

The findings reveal the need to develop research data governance skills, including acquiring suitable digital platforms for research data management and sharing, harmonizing data sharing requirements, and selecting and using repositories or platforms for data sharing. It is also imperative for research projects to develop data management plans that are in line with relevant national data governance or data protection and sharing legal frameworks, including data access and sharing agreements. Capacity building in health research data governance is important among researchers, research administrators, and

XSL•FO
RenderX

members of ethics review committees. The capacity building mentioned here relates to training in data governance, covering aspects of data stewardship, protection, and sharing, as well as awareness of and compliance with existing data protection and privacy laws, including cross-border data sharing, as well as institutional policies and guidelines for data management and sharing [22,26].

Maximizing the Benefits of and Minimizing the Risks of Sharing Health Research Data

Understanding the risks hindering health data sharing or access, Tiffin and colleagues [57] developed a digital health data governance framework in LMICs. The framework addresses 4 key domains to maximize the benefits of using health data while minimizing potential risks: ethical oversight, informed consent processes, data protection through data access controls, sustainability of fair data use, and application of relevant legislation [57].

Different studies have explored the benefits of sharing research data. For instance, data sharing allows or supports the following: replication and validation of research findings and reuse of the research data to answer new research questions, thus, cost savings [53]. Previous studies explored the attitudes and experiences of researchers on open access in LMICs, whereby there were variations in awareness and access to institutional repositories [52,58]. Furthermore, there are also inequalities between data-sharing practices among researchers in low-resource and high-resource settings [37,54].

Our study revealed positive attitudes toward openly sharing research data, including datasets. Similarly, this study reported practices in research data sharing but focused on health research data management and sharing health research datasets and publications through digital platforms. Another study focused on institutional repositories and indicated that they are mainly used to deposit and share research publications such as journal articles, reports, and dissertations [58].

Limitations and Strengths

The findings of this study are self-reported; hence, they are likely to be overestimated for respondents who want to exaggerate their practices, thus expecting to be praised, or may underestimate their capacity, thus expecting extra support. We used web-based questionnaires to gather information from the respondents. This could have limited the opportunity for respondents to clarify any unclear questions or terms. For instance, there could be confusion between sharing research outputs such as disaggregated anonymized or pseudonymized datasets, analyzed data outputs, and other publications such as journal articles, theses, or reports.

The questionnaire pretesting stage ensured the survey questions were clear and understandable to the respondents. However, the possibility of misunderstandings was not completely eliminated. The web-based administration of questionnaires enables access to large and different populations and is an inexpensive, quick, and convenient approach to collecting data [40-42]. However, this approach has the disadvantage of having relatively lower response rates than face-to-face administration [40]. Nevertheless, in this study, the response rate was higher (75.9%) than the mean response rate (40.5%) for web-based data collection approaches reviewed by Blumenberg and Barros in 2018 [40].

We furthermore limited our exploratory study to quantitative responses on the use of digital platforms for sharing research data and the experience and practices of sharing the data. However, further insights are needed into what kinds of platforms are required, what it takes to share research data, and the facilitators and barriers to using appropriate digital research tools, ideally with a mixed method approach integrating quantitative and qualitative research data.

The study is confined to assessing the existence of research data sharing agreements and the influences on data sharing practices. It did not explore the contents or types of agreements or how they ought to be structured. Future studies should concentrate on the contents and types of agreements, particularly regarding the transfer and sharing of digital data.

However, this study has documented the extent of the use of digital platforms for research data management and sharing, and practices of sharing research datasets and publications in the health sector across SSA. The experiences gathered in this study may contribute to facilitating the adoption of sustainable and scalable digital platforms that have great potential to connect researchers, policy, and decision makers, and other relevant stakeholders to enable access to research outputs for facilitating the translation of research into policy and practice in SSA. Furthermore, the platforms should conform to users' experiences and requirements and comply with national and subnational authorities' project and research reporting requirements.

Conclusion and Recommendations

The study presents an increased understanding of the existence and usage of digital research data platforms and stakeholders' characteristics in the context of research data management and sharing. The study contributes further knowledge to using digital platforms to facilitate the early translation of research data and outputs to inform practice and policy, and reuse for research in the special context of SSA.

Most respondents used electronic data collection tools for research. Nearly half indicated they have a digital research data management platform and shared their research datasets and findings with the research community through various channels. Furthermore, many respondents shared their datasets and outputs via institutional repositories, scientific conferences, and journal articles. The study identified factors such as sex, professional category (health professionals, ICT specialists, and data managers), and role (researcher or student) influencing health research data sharing.

The results showed that an individual's role had the strongest association with dataset sharing, followed by research experience, sex, and profession. In addition, logistic regression revealed that females were less likely to share their research datasets than males. Data managers and ICT professionals shared datasets less frequently within their professional group, while the researcher's role was statistically significant in promoting data sharing.

XSL•FO
RenderX

Researchers and data management teams should be equipped with proper research data management platforms or data repositories to facilitate the sharing of research outputs. A research data governance and protection framework should be established and enforced to create an enabling environment for centralized research data repositories, curation, sharing, and reuse.

Acknowledgments

We are grateful to various individuals who have provided valuable support during the design and conduct of this research project and data management. We thank the management and staff of ANDEMIA, CEBHA+, CYSTINET-Africa, TAKeOFF, and TB Sequel for supporting data collection preparation and coordination in their respective networks. We are also grateful to Juda Lyamai for his dedication to supporting data collection and analysis.

This project was funded by the German Federal Ministry of Education and Research (BMBF) under the CYSTINET-Africa (01KA1618) project. It was conducted within the sub-project on digital data sharing.

Authors' Contributions

FS designed the study and drafted the manuscript. RC, and FS conducted the data analysis. BN, AH, SS, LBD, CS, and ASW contributed to the design of the study. All authors reviewed the manuscript and contributed substantially to its development.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Questionnaire.
[DOCX File, 25 KB - formative_v9i1e69411_app1.docx]

References

1. Sharing and reuse of health-related data for research purposes: WHO policy and implementation guidance. World Health Organization. 2022 Jan 17. URL: https://www.who.int/publications/i/item/9789240044968 [accessed 2023-04-12]
2. Edwards A, Zweigenthal V, Olivier J. Evidence map of knowledge translation strategies, outcomes, facilitators and barriers in African health systems. Health Res Policy Sys 2019 Dec;17(1):16. [doi: 10.1186/s12961-019-0419-0]
3. Grimshaw JM, Eccles MP, Lavis JN, Hill SJ, Squires JE. Knowledge translation of research findings. Implementation Sci 2012 Dec;7(1):50. [doi: 10.1186/1748-5908-7-50]
4. Wensing M, Grol R. Knowledge translation in health: how implementation science could contribute more. BMC Med 2019 Dec;17(1):1-6. [doi: 10.1186/s12916-019-1322-9]
5. Wilkinson MD, Dumontier M, Aalbersberg I, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016 Mar 15;3(1):160018. [doi: 10.1038/sdata.2016.18]
6. Barker M, Chue Hong NP, Katz DS, et al. Introducing the FAIR Principles for research software. Sci Data 2022 Oct 14;9(1):622. [doi: 10.1038/s41597-022-01710-x]
7. Khan N, Thelwall M, Kousha K. Data sharing and reuse practices: disciplinary differences and improvements needed. OIR 2023 Oct 13;47(6):1036-1064. [doi: 10.1108/OIR-08-2021-0423]
8. Anane-Sarpong E, Wangmo T, Tanner M. Ethical principles for promoting health research data sharing with sub-Saharan Africa. Dev World Bioeth 2020 Jun;20(2):86-95. [doi: 10.1111/dewb.12233] [Medline: 31115148]
9. Anane-Sarpong E, Wangmo T, Ward CL, Sankoh O, Tanner M, Elger BS. “You cannot collect data using your own resources and put It on open access”: perspectives from Africa about public health data-sharing. Dev World Bioeth 2018 Dec;18(4):394-405. [doi: 10.1111/dewb.12159] [Medline: 28745008]
10. Oliver K, Innvar S, Lorenc T, Woodman J, Thomas J. A systematic review of barriers to and facilitators of the use of evidence by policymakers. BMC Health Serv Res 2014 Dec;14(1). [doi: 10.1186/1472-6963-14-2]
11. Orton L, Lloyd-Williams F, Taylor-Robinson D, O’Flaherty M, Capewell S. The use of research evidence in public health decision making processes: systematic review. PLoS ONE 2011;6(7):e21704. [doi: 10.1371/journal.pone.0021704] [Medline: 21818262]
12. Gagliardi AR, Kothari A, Graham ID. Research agenda for integrated knowledge translation (IKT) in healthcare: what we know and do not yet know. J Epidemiol Community Health 2017 Feb;71(2):105-106. [doi: 10.1136/jech-2016-207743] [Medline: 27647137]
13. Michener WK. Ecological data sharing. Ecol Inform 2015 Sep;29:33-44. [doi: 10.1016/j.ecoinf.2015.06.010]
14. Conrad LY, Delahunty R, Ding W. The promise and the future of research data sharing. Learn Publ 2022 Jan;35(1):4-6. [doi: 10.1002/leap.1432]

XSL•FO
RenderX

15. Elger K, Biskaborn BK, Pampel H, Lantuit H. Open research data, data portals and data publication–an introduction to the data curation landscape. Polarforschung 2016;85(2):119-133.
16. d'Aquin M, Kirstein F, Oliveira D, Schimmler S, Urbanek S. FAIREST: A framework for assessing research repositories. Data Intelligence 2023 Mar 8;5(1):202-241. [doi: 10.1162/dint_a_00159]
17. Amorim RC, Castro JA, Rocha da Silva J, Ribeiro C. A comparison of research data management platforms: architecture, flexible metadata and interoperability. Univ Access Inf Soc 2017 Nov;16(4):851-862. [doi: 10.1007/s10209-016-0475-y]
18. Runeson P, Olsson T, Linåker J. Open Data Ecosystems — An empirical investigation into an emerging industry collaboration concept. Journal of Systems and Software 2021 Dec;182:111088. [doi: 10.1016/j.jss.2021.111088]
19. Mansour E, Srinivas K, Hose K. Federated Data Science to Break Down Silos [Vision]. SIGMOD Rec 2022 Jan 31;50(4):16-22. [doi: 10.1145/3516431.3516435]
20. Okafor IA, Mbagwu SI, Chia T, Hasim Z, Udokanma EE, Chandran K. Institutionalizing Open Science in Africa: limitations and prospects. Front Res Metr Anal 2022;7:855198. [doi: 10.3389/frma.2022.855198] [Medline: 35494419]
21. Waithira N, Mutinda B, Cheah PY. Data management and sharing policy: the first step towards promoting data sharing. BMC Med 2019 Dec;17(1):80. [doi: 10.1186/s12916-019-1315-8]
22. Cengiz N, Kabanda SM, Moodley K. Cross-border data sharing through the lens of research ethics committee members in sub-Saharan Africa. PLoS ONE 2024;19(5):e0303828. [doi: 10.1371/journal.pone.0303828]
23. Mboera LEG, Rumisha SF, Magesa SM, Kitua AY. Utilisation of health management information system in disease surveillance in Tanzania. Tanzania J Hlth Res 2001;3(2):15-17. [doi: 10.4314/thrb.v3i2.14213]
24. Sukums F, Simba D, Kumalija C, Asiimwe S, Pothepragada S, Githendu P. Avoiding pitfalls: Key insights and lessons learnt from customizing and rolling out a national web - based system in Tanzania. E J Info Sys Dev Countries 2021 Nov;87(6):12193. [doi: 10.1002/isd2.12193]
25. Obiora OL, Shead DA, Olivier B. Data sharing considerations and practice among health researchers in Africa: a scoping review. Digit Health 2024;10:20552076241290955. [doi: 10.1177/20552076241290955] [Medline: 39493630]
26. Kabanda SM, Cengiz N, Rajaratnam K, et al. Data sharing and data governance in sub-Saharan Africa: Perspectives from researchers and scientists engaged in data-intensive research. S Afr J Sci 2023;119(5-6):15129. [doi: 10.17159/sajs.2023/15129] [Medline: 39483790]
27. Garev KV. Design principles of a domestic platform for scientific dataset exchange. Softw Syst 2025;38(1):122-133. [doi: 10.15827/0236-235X.149.122-133]
28. Habermann T. Connecting repositories to the Global Research Community: a re-curation process. J EScience Librariansh 2023;12(3). [doi: 10.7191/jeslib.739]
29. Fecher B, Friesike S, Hebing M. What drives academic data sharing? PLoS ONE 2015;10(2):e0118053. [doi: 10.1371/journal.pone.0118053] [Medline: 25714752]
30. MacFarlane A. The importance of effective data sharing and reuse to funders and others supporting research. Learn Publ 2022 Jan;35(1):71-74. [doi: 10.1002/leap.1443]
31. Hampson G, DeSart M, Steinhauer J, Gadd E, Hinchliffe LJ, Vandegrift M, et al. Open science roadmap: recommendations to UNESCO. open scholarsh initiat proc. 2020.
32. Vicente-Saez R, Martinez-Fuentes C. Open Science now: a systematic literature review for an integrated definition. J Bus Res 2018 Jul;88:428-436. [doi: 10.1016/j.jbusres.2017.12.043]
33. COVID-19 data: accelerating research through data sharing. European Commission. 2020. URL: https://www.covid19dataportal.org/ [accessed 2023-04-12]
34. Lipton V. Open scientific DATA - why choosing and reusing the RIGHT DATA matters. In: InTechOpen 2020. [doi: 10.5772/intechopen.87201]
35. Manu T, Asjola V, Gowda M, Shashikumara A, Chaudhary P, Muduli PK. Analysis of research data repositories in india. Presented at: International Conference on Knowledge Organisation In Academic Libraries (I-Koal 2018); Nov 26-27, 2018; Conference Hall (DST), University of Hyderabad.
36. Johnson R, Watkinson A, Mabe M. The STM report: an overview of scientific and scholarly publishing. 2018.
37. Global data sharing trends. Wiley. 2014. URL: https://tinyurl.com/d534vd6f [accessed 2023-04-12]
38. Laine H. Afraid of scooping – case study on researcher strategies against fear of scooping in the context of open science. Data Sci J 2017 Jun 15;16:29-29. [doi: 10.5334/dsj-2017-029]
39. Research networks for health innovations in Sub-Saharan Africa 2016. BMBF.: Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH; 2016. URL: https://www.gesundheitsforschung-bmbf.de/en/research-networks-for-health-innovations-in-sub-saharan-africa-8690.php [accessed 2025-10-01]
40. Blumenberg C, Barros AJD. Response rate differences between web and alternative data collection methods for public health research: a systematic review of the literature. Int J Public Health 2018 Jul;63(6):765-773. [doi: 10.1007/s00038-018-1108-4]
41. Granello DH, Wheaton JE. Online data collection: strategies for research. Jour of Counseling & Develop 2004 Oct;82(4):387-393 [FREE Full text] [doi: 10.1002/j.1556-6678.2004.tb00325.x]
42. Hunter L. Challenging the reported disadvantages of e-questionnaires and addressing methodological issues of online data collection. Nurse Res 2012 Sep 21;20(1):11-20. [doi: 10.7748/nr2012.09.20.1.11.c9303]

XSL•FO
RenderX

43. Widyawati C, Paskarini I. Correlation between age, work period, knowledge, and attitude's workers with behavior on the use of PPE. Jurnal Berkala Kesehatan 2021;7(2):134 [FREE Full text] [doi: 10.20527/jbk.v7i2.10891]
44. Tenopir C, Dalton ED, Allard S, et al. Changes in data sharing and data reuse practices and perceptions among scientists worldwide. PLoS ONE 2015;10(8):e0134826. [doi: 10.1371/journal.pone.0134826] [Medline: 26308551]
45. Piwowar HA. Who shares? Who doesn't? Factors associated with openly archiving raw research data. PLOS ONE 2011;6(7):e18657. [doi: 10.1371/journal.pone.0018657] [Medline: 21765886]
46. Goasduff L. Data sharing is a business necessity to accelerate digital business. : Gartner; 2021.
47. Fegan G, Cheah PY. Solutions to COVID-19 data sharing. The Lancet Digital Health 2021 Jan;3(1):e6. [doi: 10.1016/S2589-7500(20)30273-9]
48. Warren E. Strengthening research through data sharing. N Engl J Med 2016 Aug 4;375(5):401-403. [doi: 10.1056/NEJMp1607282] [Medline: 27518656]
49. Chinakidzwa M, Phiri M, Mashingaidze M. Research data sharing and re-use practices, perceptions and behaviours: evidence from Zimbabwe. JAE 2022 Apr 12;3(1):75-101 [FREE Full text] [doi: 10.31920/2633-2930/2022/v3n1a4]
50. Onyancha OB. Open research data in Sub-Saharan Africa: a bibliometric study using the data citation index. Pub Res Q 2016 Sep;32(3):227-246. [doi: 10.1007/s12109-016-9463-6]
51. Al-Ebbini L, Khabour OF, Alzoubi KH, Alkaraki AK. Biomedical data sharing among researchers: a study from Jordan. J Multidiscip Healthc 2020;13:1669-1676. [doi: 10.2147/JMDH.S284294] [Medline: 33262602]
52. Nobes A, Harris S. Open access in low- and middle-income countries: attitudes and experiences of researchers. EOR 2023 Dec 12;1(3):17. [doi: 10.1108/EOR-03-2023-0006]
53. Kratz JE, Strasser C. Researcher perspectives on publication and peer review of data. PLoS ONE 2015;10(2):e0117619. [doi: 10.1371/journal.pone.0117619] [Medline: 25706992]
54. Tenopir C, Rice NM, Allard S, et al. Data sharing, management, use, and reuse: practices and perceptions of scientists worldwide. PLoS ONE 2020;15(3):e0229003. [doi: 10.1371/journal.pone.0229003] [Medline: 32160189]
55. Aljunid SM, Srithamrongsawat S, Chen W, et al. Health-care data collecting, sharing, and using in Thailand, China mainland, South Korea, Taiwan, Japan, and Malaysia. Value Health 2012;15(1 Suppl):S132-S138. [doi: 10.1016/j.jval.2011.11.004] [Medline: 22265060]
56. Alsheikh-Ali AA, Qureshi W, Al-Mallah MH, Ioannidis JPA. Public availability of published research data in high-impact journals. PLoS ONE 2011;6(9):e24357. [doi: 10.1371/journal.pone.0024357] [Medline: 21915316]
57. Tiffin N, George A, LeFevre AE. How to use relevant data for maximal benefit with minimal risk: digital health data governance to protect vulnerable populations in low-income and middle-income countries. BMJ Glob Health 2019 Apr;4(2):e001395. [doi: 10.1136/bmjgh-2019-001395]
58. Nunda IM, Elia EF. Institutional repositories adoption and use in selected Tanzanian higher learning institutions. IJEDICT 2019;15(1).

Abbreviations

BMBF: Federal Ministry of Education and Research
cOR: crude odds ratio
FAIR: findable, accessible, interoperable, and reusable
ICT: information and communication technology
LMICs: low- and middle-income countries
NIMR: National Institute for Medical Research
OR: odds ratio
SSA: sub-Saharan Africa
WHO: World Health Organization

Edited by A Mavragani; submitted 28.11.24; peer-reviewed by G Humphreys, SM Kabanda; revised version received 15.05.25; accepted 16.05.25; published 23.10.25.

Please cite as:
Sukums F, Ngowi B, Chaula R, Weiszhar KL, Kalinga A, Ivanova O, Prazeres da Costa C, Hoerauf A, Sahay S, Noormahomed EV, Debrah LB, Sikasunge C, Ngowi H, Winkler AS
Experiences of Health Research Data Sharing Among Researchers in Sub-Saharan Africa: Cross-Sectional Study
JMIR Form Res 2025;9:e69411
URL: https://formative.jmir.org/2025/1/e69411
doi:10.2196/69411

XSL•FO
RenderX

© Felix Sukums, Bernard Ngowi, Rebecca Chaula, Kim L Weiszhar, Akili Kalinga, Olena Ivanova, Clarissa Prazeres da Costa, Achim Hoerauf, Sundeep Sahay, Emilia Virginia Noormahomed, Linda B Debrah, Chummy Sikasunge, Helena Ngowi, Andrea S Winkler. Originally published in JMIR Formative Research (https://formative.jmir.org), 23.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Factors Influencing the Use of Online Symptom Checkers in the United Kingdom: Cross-Sectional Study

Austen El-Osta[1], MSc, MPA, PhD; Eva Riboli-Sasco[1], MA; Mahmoud Al Ammouri[1], MD, MPH; Sami Altalib[1], MD, MPH; Ana Luisa Neves[2], MD, MPH, MSc; Azeem Majeed[2], MD, PhD; Benedict Hayhoe[2], MD

[1]Self-Care Academic Research Unit (SCARU), Imperial College London, School of Public Health, 90 Wood Lane, London, United Kingdom
[2]Department of Public Health and Primary Care, Imperial College London, London, United Kingdom

Corresponding Author:
Austen El-Osta, MSc, MPA, PhD
Self-Care Academic Research Unit (SCARU), Imperial College London, School of Public Health, 90 Wood Lane, London, United Kingdom

Abstract

Background: The National Health Service (NHS) faces increasing strain. Concurrently, demand for health information, consumer empowerment, and health awareness continues to grow. These trends, coupled with the ubiquity of smartphones and internet access, are positioning online symptom checkers (OSCs) as promising tools for preliminary diagnosis and triage. While there is increasing data on the demographics, motivations, and perspectives of current and potential users of OSCs globally, no study has yet quantified or ranked the various factors associated with the use of OSCs in the United Kingdom.

Objective: This study aimed to assess key trends and user perceptions on the usability and effectiveness of OSC in the United Kingdom. We also sought to identify concerns related to the privacy, security, and accuracy of OSCs and to quantify the weight of these various factors on the use of OSCs.

Methods: A cross-sectional survey of UK adults was conducted using an electronic questionnaire. A convenience sample was recruited between February and March 2024 through web-based platforms and personal networks. The survey included questions on awareness, use, perceptions, and concerns regarding OSCs, as well as respondents' demographics. Responses were pseudo-anonymized and analyzed using univariable and multivariable logistic regression models to assess relationships between demographic factors; perceived usability, reliability, and risks; and OSC use.

Results: The survey collected responses from 634 participants. The majority (543/634, 85.7%) had used OSCs, primarily the NHS 111 service (498/634, 78.6%). Younger age (<46 years old), being female (adjusted odds ratio [aOR] 1.79, 95% CI 1.05 - 3.06), and having children (aOR 3.19, 95% CI 1.56 - 6.51) were associated with higher odds of using OSCs. Key motivations for using OSCs included understanding symptoms (501/634, 79.0%) and determining the need for medical care (491/634, 77.4%). Key concerns negatively impacting use related to privacy (aOR 0.58, 95% CI 0.35-0.97) and fear of replacing traditional, face-to-face consultations (aOR 0.47, 95% CI 0.26 - 0.87). The most important factor found to affect the decision to use OSCs was the perceived ease of use (aOR 8.17, 95% CI 4.25-15.71), followed by the perceived helpfulness in decision-making (aOR 2.96, 95% CI 1.62 - 5.42), and respondents' trust in their diagnostic accuracy (aOR 2.24, 95% CI 1.32 - 3.79).

Conclusions: OSCs are widely used in the United Kingdom, particularly the NHS 111 service, driven primarily by ease of use and perceived helpfulness in decision support. However, privacy and security concerns, as well as fears of OSCs replacing traditional consultations, pose significant barriers. Addressing these concerns is crucial for enhancing user trust and maximizing the benefits of OSCs in supporting self-care and improving health care efficiency.

(JMIR Form Res 2025;9:e65314) doi:10.2196/65314

KEYWORDS

digital health; online symptom checker; self-care; symptom checker; United Kingdom; user experience; cross sectional; survey; questionnaire; opinions; perceptions; usability; triage

Introduction

Overview

The global shortage of health care workers has exacerbated the challenges faced by health systems worldwide. The World Health Organization projects a deficit of 10 million health care workers by 2030. While low- and lower-middle-income countries will be the most affected, all countries will face significant challenges [1]. In the United Kingdom, the National Health Service (NHS) is already under considerable strain [2], which significantly impacts the quality of care and health outcomes for patients [3].

XSL•FO
RenderX

At the same time, there is a growing demand for health information and increasing consumer empowerment [4]. In this context, the reliance on digital health tools has surged, particularly decision support tools including online symptom checkers (OSCs) [5,6]. The ubiquitous access to the internet supports this trend, with 96% of households in the United Kingdom having internet access in 2020 [7], which accelerated following the advent of the COVID-19 pandemic, given the need for health services to avoid face-to-face contact and preserve urgent care capacity [8].

Symptom checkers are available as websites or applications and can generate a prioritized list of potential diagnoses based on the entered symptoms and suggest suitable actions, such as self-care, consulting a general practitioner (GP), or seeking urgent medical care [9]. By providing preliminary diagnostic guidance and triage recommendations, these tools could potentially alleviate some of the burdens on health care systems with the potential to reduce unnecessary health care visits by providing timely medical advice and empowering individuals to make informed decisions about their health [10-12]. Despite this, the diagnostic and triage accuracy varies greatly among OSCs, thus raising concerns and calling for caution [13].

Evidence regarding the sociodemographic profile of OSC users shows that they tend to be young [6,14,15], women [6,14-16], and more educated [6,14,16], whereas confidence in OSC and intentions to discuss recommendations with providers varied by ethnicity [17]. A Chinese study that investigated the use of self-diagnosis health chatbots found a higher prevalence of male users [18], highlighting likely variations across countries. Having a chronic health condition has also been associated with greater use in one study conducted among US users of the Isabel Symptom Checker [16]. A 2022 systematic review on user experiences of symptom checkers identified 8 relevant aspects of user experience that have been explored in the literature, including motivation, trust, acceptability, satisfaction, accuracy, usability, safety/security, and functionality [19]. A recent review identified convenience, supplementing offline medical visits, gaming, and social influence as key motivations for using OSCs, whereas other studies highlighted the quest for a better understanding of the causes of symptoms and support for deciding whether to seek care [16]. Individual factors like internet access, health technology literacy, and awareness of OSC availability also play significant roles in determining user engagement with OSCs [5,12,20], with one UK survey showing higher mean health technology literacy across users of an OSC [21]. Collectively, these findings highlight the importance of enhancing the accuracy, accessibility, usability, and trustworthiness of OSCs to improve user experience and integration into health care services.

While there is a growing body of evidence regarding OSC use, no study has yet quantified or ranked the various factors associated with the use of OSCs in the community setting.

Objective

This study aimed to identify, characterize, and quantify the factors associated with the use of OSCs among community-dwelling adults in the United Kingdom. Specifically, we sought to (1) assess the demographic characteristics of OSC users, (2) evaluate user perceptions of the usability and effectiveness of OSCs, (3) identify concerns related to the privacy, security, and accuracy of OSCs, and (4) quantify the weight of these various factors on the adoption and use of OSCs.

Methods

Ethical Considerations

The Imperial College Research Ethics Committee granted ethical clearance for the study (ICREC# 20IC5974). All experimental protocols were approved by the Imperial College London Research Ethics Committee. All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. Informed consent was obtained from all participants. Consent for publication is not applicable.

All data collected were pseudo-anonymized at the point of submission by assigning unique identifiers, with no directly identifiable personal information (eg, name, address, and IP address) retained in the final dataset. The survey platform (Qualtrics XM) was configured to prevent the collection of IP addresses and location metadata. All data were stored securely on password-protected servers at Imperial College London, accessible only to the study team. Participants were informed that their responses would be confidential, that their participation was voluntary, and that they could withdraw at any time prior to submitting their completed responses. No incentives were offered for participation in the survey unless recruited via the Prolific Academic platform, where participants received a small financial compensation in line with the platform's fair pay guidelines. This compensation rate was preapproved by the platform and considered appropriate for a brief (under 10-minute) survey. No identifying data were linked to compensation claims.

Study Design

This cross-sectional study aimed to explore factors influencing the use of OSCs among community-dwelling adults in the United Kingdom. The study used a quantitative methodology using an electronic survey tool (eSurvey).

Data Collection

This was an open eSurvey, accessible to anyone with the survey link, and required less than 10 minutes to complete. The eSurvey was developed and tested to ensure clarity, usability, and technical functionality before fielding. The link to the eSurvey was active on the Imperial College Qualtrics platform between February 23 and March 25, 2024. Study information was disseminated, including the participant information sheet and link to the survey. Participants were recruited through convenience sampling. The researcher's personal and professional networks were mobilized to respond and further disseminate the eSurvey among potentially eligible participants. Most participants were recruited via Prolific Academic's panel [22]. The participant information sheet included information regarding the study's aims, the protection of participants' personal data, their right to withdraw from the study at any time,

XSL•FO
RenderX

which data were stored, where and for how long, who the investigator was, the purpose of the study, and the survey length.

Participants were informed that this was a voluntary survey. Informed consent was obtained from all participants. Data collected were stored on a secure database at Imperial College London and were only accessible to the research team. All responses were pseudo-anonymized to ensure confidentiality.

The selection of factors that may affect the use of OSCs to be included in the survey was guided by a review of existing literature on the topic. Demographic factors, including age and gender, as well as perceptions such as perceived usability, effectiveness, reliability, accuracy, safety, and privacy, have been identified as influential factors from prior studies [6,16,19,20,23,24].

Before publication, the eSurvey was tested, piloted, and revised internally by the study team. In its final version, the eSurvey comprised a total of 25 questions displayed across 4 screens and gathered data regarding respondents' awareness, use, and perspectives regarding OSCs, as well as basic demographic information (Textbox 1).

Textbox 1. Independent and dependent variables.

Independent variables

- Age
- Gender
- Ethnicity
- Highest level of education
- Having children (less than 16 years of age)
- Disability
- Long-term health conditions (LTCs)
- I find symptom checkers easy to use
- Help me make better choices when seeking medical care
- Help me improve my health literacy and support my self-care journey
- Useful when I have limited access to a health care professional (eg, rural setting and out of hours)
- I trust their suggested diagnosis to be accurate
- I trust their triage recommendation (when and where to seek appropriate care) to be accurate
- My general practitioner would encourage me to use symptom checkers
- My family and friends would encourage me to use symptom checkers
- I find using a symptom checker reassuring, and it makes me feel less anxious about my health
- Not yet safe enough to rely solely on them, and may put my health at risk
- I am concerned that using the symptom checker may put my privacy and health information at risk
- I think symptom checkers may increase inequalities between patients
- I am worried about symptom checkers replacing face-to-face or phone consultations

Dependent variable

- Using the online symptom checker

The questionnaire underwent preliminary testing with a small group of potential respondents to refine the items based on feedback on clarity and relevance. This included 4 departmental colleagues (2 primary care physicians, a Master of Public Health student, and a researcher), and 8 members of a lay reference group who are diverse in age, gender, and ethnicity. This pilot testing helped in adjusting the survey to better address the study objectives. The study eSurvey was reviewed during beta testing to improve usability and wording of questions, although no attempt was made to validate the tool, given the nature and scope of this study. The final version of the survey instrument is available as Multimedia Appendix 1.

Data Analysis

Only questionnaires fully completed were included in the analysis. Duplicate entries from the same IP address within a 24-hour period were also eliminated before analysis. Participant characteristics and responses were summarized using total (n) and relative (%) frequencies. For inferential analyses, "strongly agree" and "somewhat agree" were categorized into "agree," and "strongly disagree" and "somewhat disagree" were categorized into "disagree," to create binary variables for logistic regression analysis.

In the statistical analysis, we used a series of logistic regression models to examine the associations between various factors and

XSL•FO
RenderX

the use of OSC. We conducted both unadjusted and adjusted analyses for four distinct categories of predictors: (1) demographic factors, (2) usability and effectiveness, (3) reliability and accuracy, and (4) risks and concerns. For each category, we first performed univariable logistic regression to assess the unadjusted relationship between the predictors and OSC. Subsequently, we constructed multivariable logistic regression models to account for potential confounding factors. In the adjusted models, we included the following covariates: age, gender, ethnicity, highest level of education, parenting status (having children), disability, and long-term health conditions. For the demographic factors category, we examined the individual associations of each demographic variable with OSC use in the univariable model. In the multivariable model for this category, we included all demographic variables simultaneously to assess their independent effects while controlling for one another. For the other 3 categories (usability and effectiveness, reliability and accuracy, and risks and concerns), the adjusted models included the respective predictors of interest along with the demographic covariates. This approach allowed us to evaluate the relationships between each category of predictors and OSC use, both in isolation and while accounting for potential confounding effects of demographic characteristics.

Results were deemed statistically significant at a P value <.05. The odds ratios for these relationships were quantified to understand the influence of each factor on the use of OSCs and compare them. All analyses were performed using STATA (version 17; StataCorp LP). The CHERRIES (Checklist for Reporting Results of Internet E-Surveys) was used to guide reporting [25] (Checklist 1).

Results

Participant Characteristics

A total of 641 individuals took part in the survey, with complete responses obtained from 634 respondents. A full description of participants according to age, gender, ethnicity, and educational background is provided in Table 1. The largest proportion of respondents was between 26 and 35 years old (207/634, 32.7%) and female (295/634, 46.5%). The majority identified as White (534/634, 84.2%) and had a college or university degree (453/634, 71.5%). Nearly a third (200/634, 31.5%) had children under 16 years, 15.1% (n=96) reported having a disability, and 22.9% (n=145) had one or more long-term health conditions.

XSL•FO
RenderX

Table . Respondent characteristics (n=634).

	Values, n (%)
Age (years)	
18 - 25	108 (17.0)
26 - 35	207 (32.7)
36 - 45	156 (24.6)
46 - 55	89 (14.0)
56 - 65	55 (8.7)
Older than 65	19 (3.0)
Sex	
Male	265 (41.8)
Female	295 (46.5)
Other	65 (10.3)
Prefer not to say	9 (1.4)
Ethnicity	
Asian or Asian British	45 (7.1)
Black, African Caribbean, or Black British	25 (3.9)
White	534 (84.2)
Other ethnic group	25 (3.9)
Prefer not to say	5 (0.8)
Highest level of education	
Primary school or secondary school (up to 16 years)	19 (3.0)
Higher or secondary or further education (A-levels, BTEC, etc)	158 (24.9)
College or university degree	453 (71.5)
Prefer not to say	4 (0.6)
Having children (less than 16 y of age)	
Yes	200 (31.5)
No	434 (68.5)
Disability	
Yes	96 (15.1)
No	509 (80.3)
Prefer not to say	29 (4.6)
Long-term health conditions	
Yes	145 (22.9)
No	453 (71.4)
Prefer not to say	36 (5.7)

Main Survey Findings

Use of OSC

The majority (543/634, 85.7%) of participants had used an OSC, while 14.4% (n=91) reported never having used one. The reasons for nonuse of OSC reported by the largest proportion (41/91, 45.1%) were preference for consulting a health care professional directly rather than using OSC, followed by having never heard of them (36/91, 39.6%); 24.2% (n=22) of respondents did not trust them. However, two-thirds of the nonusers (59/91, 64.8%) expressed a likelihood of using a symptom checker in the future (Table S1 in Multimedia Appendix 2).

The most used OSC was NHS 111, with 78.6% (n=498) of respondents indicating that they had used this tool, followed by Healthline (142/634, 22%). Participants predominantly used symptom checkers before seeking medical advice (533/634, 94.5%), primarily to better understand symptoms (501/634, 79.0%) and to determine the need for medical care (491/634,

XSL•FO
RenderX

77.4%). Most respondents (429/634, 77.9%) indicated that symptom checkers offered recommendations for action or triage. Among those who received such recommendations, a significant proportion (337/429, 79.7%) were directed to seek a consultation with a health care professional. Regarding adherence, most participants (59.8% [n=253]) reported following the recommendations most of the time, 27.7% (n=117) stated they always adhered to the suggestions provided by the symptom checker, while a smaller proportion (52/429, 12.3%) admitted to rarely doing so.

Association Between Demographics and Use of OSC

An increase in age was associated with a decrease in OSC use. Specifically, individuals aged 46 - 55, 56 - 65 and >65 years showed significantly decreased odds of using OSC (adjusted odds ratio [aOR] 0.29, 95% CI 0.11 - 0.72), (aOR 0.27, 95% CI 0.10 - 0.71), and (aOR 0.22, 95% CI 0.06 - 0.78) respectively, compared with the younger [18-25] age group. Similarly, females exhibited higher odds of using OSC compared with males (aOR 1.79, 95% CI 1.05 - 3.06). Having children younger than 16 years of age also showed significantly higher odds of using OSC (aOR 3.19, 95% CI 1.56 - 6.51) compared with those who do not have children younger than 16 years of age. In contrast, neither ethnicity nor educational background exhibited any statistically significant association with the use of OSC. Similarly, disability and long-term health conditions did not contribute to the outcome of using OSC (Table 2).

XSL•FO
RenderX

Table . The association between demographics and using an online symptom checker (OSC), adjusted for age, gender, ethnicity, highest level of education, having children, disability, and long-term health conditions.

Variable	Using the OSC			
	Unadjusted		Adjusted	
	OR[a] (95% CI)	*P* value	OR (95% CI)	*P* value
Age (years)				
18 - 25	__[b]	—	—	—
26 - 35	1.14 (0.50 - 2.59)	.75	1.07 (0.45 - 2.54)	.89
36 - 45	0.69 (0.31 - 1.55)	.37	0.52 (0.21 - 1.29)	.16
46 - 55	0.35 (0.16 - 0.80)	.01	0.29 (0.11 - 0.72)	.008
56 - 65	0.23 (0.10 - 0.54)	.001	0.27 (0.10 - 0.71)	.008
Older than 65	0.18 (0.06 - 0.55)	.003	0.22 (0.06 - 0.78)	.02
Sex				
Male	—	—	—	—
Female	1.98 (1.22 - 3.21)	.006	1.79 (1.05 - 3.06)	.03
Other	1.28 (0.61 - 2.68)	.52	0.81 (0.32 - 2.01)	.65
Ethnicity				
White	—	—	—	—
Black, African Caribbean, or Black British	1.97 (0.45 - 8.51)	.365	1.16 (0.25 - 5.49)	.85
Asian or Asian British	1.75 (0.61 - 5.03)	.297	1.41 (0.47 - 4.25)	.54
Other ethnic group	0.90 (0.30 - 2.69)	.847	0.51 (0.16 - 1.63)	.25
Highest level of education				
Primary school or secondary school (up to 16 years)	—	—	—	—
Higher or secondary or further education (A-levels, BTEC, etc)	0.87 (0.24 - 3.19)	.83	0.76 (0.19 - 3.08)	.70
College or university degree	1.23 (0.35 - 4.34)	.75	1.00 (0.26 - 3.88)	.99
Having children (less than 16 y of age)				
No	—	—	—	—
Yes	2.62 (1.46 - 4.68)	.001	3.19 (1.56 - 6.51)	.001
Disability				
No	—	—	—	—
Yes	1.21 (0.63 - 2.32)	.57	1.17 (0.50 - 2.77)	.72
Long-term health conditions (LTCs)				
No	—	—	—	—
Yes	1.53 (0.85 - 2.77)	.16	1.91 (0.94 - 3.88)	.07

[a]OR: odds ratio.

[b]Reference.

Associations Between OSC Use and Users' Perceptions of OSCs

The main survey findings are shown in Table S1 in Multimedia Appendix 2. The segment below highlights key associations between perceived usability and effectiveness, reliability and accuracy, and risks and concerns with using OSCs.

Perceived Usability, Usefulness, and Use of OSCs

Most of the participants found OSCs easy to use (566/634, 89.3%), believed they could help with medical decisions (545/634, 86.0%), and support their health literacy and self-care (537/634, 84.7%). A great proportion (581/634, 94.6%) also agreed that OSCs are useful tools in scenarios with limited

XSL•FO
RenderX

access to health care professionals, such as rural settings or out-of-hours situations.

Participants who found the symptom checkers easy to use were more likely to use them compared with those who did not (aOR 8.17, 95% CI 4.25 - 15.71). Similarly, individuals who found these tools helpful in making better medical care choices were more likely to use them than those who did not (aOR 2.96, 95% CI 1.62 - 5.42). Moreover, those who agreed that the symptom checkers improved health literacy and supported self-care showed a heightened likelihood of use (aOR 2.36, 95% CI 1.30 - 4.28). Participants who perceived the OSC as useful in scenarios with limited access to health care professionals were twice as likely to use them compared with those who disagreed (aOR 2.15, 95% CI 1.01 - 4.59) (Table 3).

Table . The association between usability and effectiveness and the use of an online symptom checker (OSC), adjusted for age, gender, ethnicity, highest level of education, having children, disability, and long-term health conditions.

	Using the OSC			
	Unadjusted		Adjusted	
	OR[a] (95% CI)	*P* value	OR (95% CI)	*P* value
I find symptom checkers easy to use				
Disagree	—[b]	—	—	—
Agree	8.93 (5.16 - 15.46)	<.001	8.17 (4.25 - 15.71)	<.001
Help me make better choices when seeking medical care				
Disagree	—	—	—	—
Agree	3.05 (1.80 - 5.15)	<.001	2.96 (1.62 - 5.42)	<.001
Help me improve my health literacy and support my self-care journey				
Disagree	—	—	—	—
Agree	2.66 (1.58 - 4.47)	<.001	2.36 (1.30 - 4.28)	.005
Useful when I have limited access to a health care professional (eg, rural setting, out of hours)				
Disagree	—	—	—	—
Agree	2.35 (1.22 - 4.53)	.011	2.15 (1.01 - 4.59)	.05

[a]OR: odds ratio.

[b]Reference.

Perceived Reliability, Accuracy, and Use of OSCs

Just over half of the respondents expressed confidence and trust in OSCs' information (363/634, 57.3%). The perceived accuracy of OSCs' diagnosis was slightly higher (399/634, 63.0%), while 79.2% (n=502) of respondents said they trusted the triage provided by the tools. A total of 69.2% (n=438) of respondents found using a symptom checker reassuring and made them feel less anxious about their health.

Participants who trusted the suggested diagnosis to be accurate were more likely to use the symptom checkers compared with those who disagreed (aOR 2.24, 95% CI 1.32 - 3.79). Similarly, individuals who trusted the triage recommendation provided by these tools showed a heightened likelihood of use (aOR 2.33, 95% CI 1.33 - 4.06). Participants who found using symptom checkers reassuring and anxiety-reducing were significantly more likely to use them (aOR 3.85, 95% CI 2.28 - 6.50). Additionally, the encouragement from family and friends to use symptom checkers significantly influenced their use (aOR 2.05, 95% CI 1.24 - 3.41). However, encouragement from GPs was not significantly associated with the use of symptom checkers (aOR 1.34, 95% CI 0.79 - 2.24) (Table 4).

XSL•FO
RenderX

Table . The association between reliability and accuracy and the use of an online symptom checker (OSC), adjusted for age, gender, ethnicity, highest level of education, having children, disability, and long-term health condition.

	Using the OSC			
	Unadjusted		Adjusted	
	OR[a] (95% CI)	*P* value	OR (95% CI)	*P* value
I trust their suggested diagnosis to be accurate				
Disagree	__[b]	—	—	—
Agree	2.14 (1.36 - 3.34)	.001	2.24 (1.32 - 3.79)	.003
I trust their triage recommendation (when and where to seek appropriate care) to be accurate				
Disagree	—	—	—	—
Agree	2.55 (1.58 - 4.12)	<.001	2.33 (1.33 - 4.06)	.003
My general practitioner (GP) would encourage me to use symptom checkers				
Disagree	—	—	—	—
Agree	1.58 (0.99 - 2.52)	.054	1.34 (0.79 - 2.24)	.28
My family and friends would encourage me to use symptom checkers				
Disagree	—	—	—	—
Agree	2.08 (1.33 - 3.26)	.001	2.05 (1.24 - 3.41)	.005
I find using a symptom checker reassuring, and it makes me feel less anxious about my health				
Disagree	—	—	—	—
Agree	3.35 (2.12 - 5.27)	<.001	3.85 (2.28 - 6.50)	<.001

[a]OR: odds ratio.
[b]Reference.

Perceived Risks and Concerns and Use of OSCs

Reported concerns included safety (481/634, 75.9%), privacy (258/634, 41.7%), and exacerbating inequalities (257/634, 41.6%). More than half of the respondents (417/634, 65.9%) worried about replacing traditional consultations, and more than a quarter (170/634, 26.9%) would not feel confident discussing the outcomes of their symptom checker consultation with their GP. Participants who agreed that symptom checkers are not yet safe enough to rely solely on them and may put their health at risk did not show a significant association with use compared with those who disagreed (aOR 0.59, 95% CI 0.30 - 1.17). However, concerns regarding privacy and health information security were significantly inversely associated with use, with individuals expressing such worries being less likely to use symptom checkers (aOR 0.58, 95% CI 0.35 - 0.97). Similarly, those who believed that symptom checkers might increase inequalities between patients were less likely to use them (aOR 0.47, 95% CI 0.28 - 0.79). Concern about OSCs replacing face-to-face or phone consultations was significantly associated with decreased use (aOR 0.47, 95% CI 0.26 - 0.87) (Table 5).

Figure 1 highlights the relative weight of each factor on the use of OSCs. The main factor that significantly increased the probability of using OSCs was the tools' ease of use (aOR 8.17, 95% CI 4.25 - 15.71). This was followed by feeling reassured by using the tool and having children. Users of these decision support tools also usually thought of OSCs as helpful in influencing and improving their health care choices, their health literacy, and self-care capacity. Encouragement from friends and family, limited health care access, being female, and trust in the triage and diagnostic accuracy of OSCs were also important factors.

Demographic factors associated with decreased odds of using OSCs included male gender and older age. Concerns regarding privacy and data security, as well as the risk of increased inequalities and loss of face-to-face consultations due to OSCs, were also identified as reducing the likelihood of using these tools (Figure 1).

XSL•FO
RenderX

Table . The association between risks and concerns and the use of an online symptom checker (OSC), adjusted for age, gender, ethnicity, highest level of education, having children, disability, and long-term health conditions.

	Using the OSC			
	Unadjusted		Adjusted	
	OR[a] (95% CI)	*P* value	OR (95% CI)	*P* value
Not yet safe enough to rely solely on them, and may put my health at risk				
Disagree	—[b]	—	—	—
Agree	0.69 (0.39 - 1.21)	.19	0.59 (0.30 - 1.17)	.13
I am concerned that using the symptom checker may put my privacy and health information at risk				
Disagree	—	—	—	—
Agree	0.46 (0.29 - 0.71)	.001	0.58 (0.35 - 0.97)	.04
I think symptom checkers may increase inequalities between patients				
Disagree	—	—	—	—
Agree	0.59 (0.38 - 0.92)	.02	0.47 (0.28 - 0.79)	.005
I am worried about symptom checkers replacing face-to-face or phone consultations				
Disagree	—	—	—	—
Agree	0.50 (0.29 - 0.84)	.010	0.47 (0.26 - 0.87)	.015

[a]OR: odds ratio.

[b]Reference.

Figure 1. Factors that significantly influence the use of an online symptom checker. The size of the bubbles is proportional to the odds ratios of each factor's influence, where green depicts factors that increase the likelihood, and red indicates factors that decrease the likelihood of using online symptom checkers.

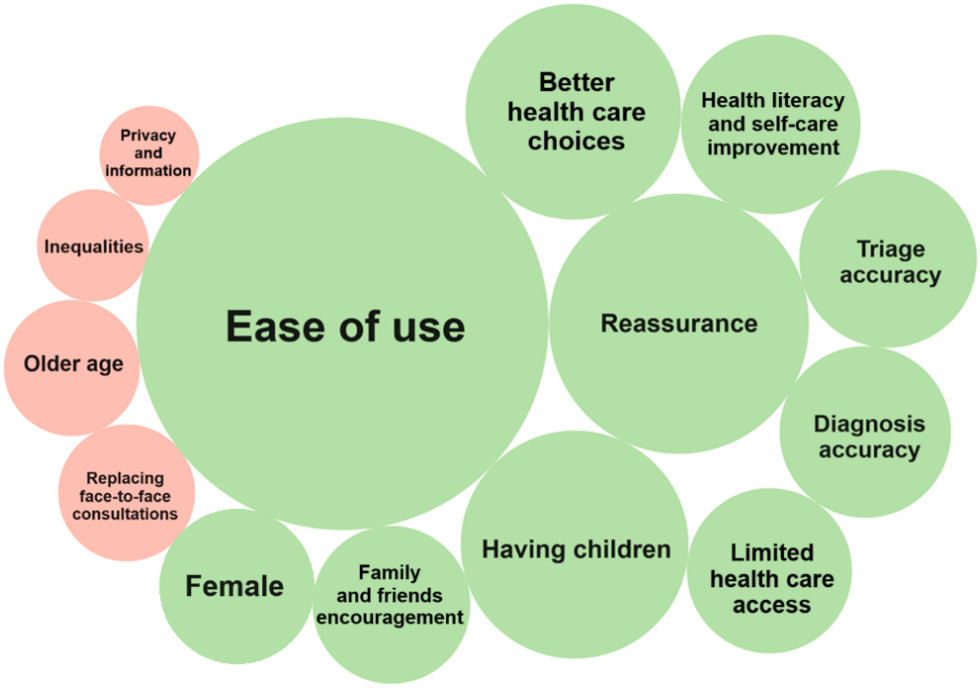

Discussion

Summary of Principal Findings

This study investigated the factors that influence the use of OSC through a cross-sectional survey of community-dwelling adults in the United Kingdom. Most participants (545/634, 86%) had used a symptom checker at some point, with the NHS 111 platform being the most widely used (501/634, 79%), followed by Healthline (142/634, 22%). NHS 111 is a free-to-use service provided by the UK's NHS that offers both online and telephone triage for urgent but nonemergency health concerns. This platform guides users through a series of questions about their symptoms, providing advice on the next steps to take, which may include self-care, contacting a GP, or seeking emergency services. The service aims to reduce unnecessary visits to

XSL•FO
RenderX

emergency departments while ensuring patients receive appropriate care [26]. Healthline, on the other hand, is a US-based health information website that provides comprehensive, evidence-based health content covering various topics, conditions, and wellness information [27]. These tools and other OSCs named by respondents to our survey were predominantly used before seeking medical advice (602/634, 95%), primarily to better understand symptoms and to determine the need for care.

We found that while most OSC nonusers expressed a strong likelihood of future use, respondents had varying concerns regarding the accuracy of information, safety, privacy, and the potential increase of inequalities. Older individuals (aged between 46 - 55, 56 - 65, and older than 65 years) showed significantly decreased odds of using OSC compared with the 18 - 25 years age group, independently of other variables. However, females exhibited higher odds of using these tools compared with males, and individuals with children younger than 16 years of age were more likely to use OSC compared with those who did not have children. Further, although trust in the accuracy of diagnoses and triage recommendations, as well as encouragement from family and friends, positively influenced use, concerns regarding privacy, health information security, inequalities, and the potential displacement of traditional consultations were significantly associated with decreased use (Figure 1). No significant associations were found between the use of OSCs and encouragement from GPs to use these tools, concerns about the safety of relying solely on symptom checkers, or potential health risks. These findings highlight the complex nature of the many factors that could influence the use of OSCs in the contemporary setting.

Study Strengths and Limitations

A key strength of this study is that it examined a wide range of variables, including demographics, motivations, perceived effectiveness, reliability, and concerns regarding OSC and associations with their use, providing valuable insights into the complexities surrounding the wide-scale adoption and diffusion of these tools in the contemporary setting. While supporting existing evidence on this topic, we identified additional factors associated with the use of OSCs, including the perceived family and friends support, as well as having children. By using regression models, we were also able to quantify the associations between the use of OSCs and a variety of factors, including both demographic factors and respondents' perceptions regarding the usability, safety, accuracy, and concerns associated with OSCs. Crucially, we identified significant predictors of use while controlling for potential confounders and enhancing the internal validity. The study's relatively large and diverse sample also closely reflects the ethnic distribution of the UK population in terms of broad categories [28] as well as the prevalence of long-term conditions and disabilities.

The principal limitation of this study is its cross-sectional nature, as it cannot establish causality and temporal relationships between the factors examined. The other main limitation is concerned with the sampling method. While convenience sampling allowed for efficient data collection and provided initial insights into the perceptions of OSC users, it potentially limits the generalizability of our findings. Future studies should consider using probability sampling techniques to ensure a more representative sample of the general population, thereby enhancing the external validity of the results. Another key limitation is that the study used a single-item measure to assess the usability of OSC. Recognizing the complexity of usability as a construct, we acknowledge that this simplified approach may not capture all dimensions of usability effectively, but we deemed it sufficient for the purpose of keeping the number of questions in the survey to a minimum, especially as we sought to measure responses to other factors. Future research would benefit from incorporating multidimensional, validated usability scales to provide a more detailed understanding of user interactions with digital health tools.

As the survey was only accessible via web, it is likely that potentially eligible participants with limited access to the internet or those less confident with digital technology may have been excluded and their views absent or underreported. In addition, we used a convenience sample and could not measure response rates. The voluntary nature of the survey meant that respondents with an interest in or previous experience with OSCs might have been more likely to take part, which could explain the high prevalence of people who had previously used an OSC (543/634, 85.6%). While we adjusted for several demographic factors in our analyses, the possibility of unmeasured confounding cannot be ruled out, which may influence the observed associations. Our primary analysis also did not extensively explore interactions between covariates, potentially overlooking complex relationships between factors associated with OSC use. We acknowledge also that, as this study relied on self-reported data, this may be subject to recall and social desirability biases, and for this reason, participants' responses regarding their use patterns, preferences, and adherence to recommendations may therefore not fully reflect their actual behaviors. Because these limitations are common in survey studies, longitudinal studies are indicated to follow up users over a longer time horizon to better understand how their interactions with OSCs evolve and how these tools may influence their health behaviors, lifestyle choices, health care use, and health outcomes.

Comparison With Existing Literature

The findings of this study are in line with prior research, including primary studies and reviews reporting on the sociodemographics of OSC users who tend to be young [6,12,14,15], women [6,12,14-16], and with higher education levels [6,12,14,16], and higher digital literacy scores [21]. Although having a chronic health condition or a disability was associated with greater use in a study by Meyer et al [16], we did not find this association in this study's sample. Regarding the motivations for using OSCs, a better understanding of the causes of symptoms has also been found to be the primary motivation among US users of the Isabel Symptom Checker [16], followed by support for deciding whether to seek care.

The finding that users of OSCs tend to find these tools easy to use and helpful was corroborated by Meyer et al [16] and Pairon et al [6]. The strong correlation identified in this study highlights the importance of user-friendly interfaces in promoting the

XSL•FO
RenderX

adoption of OSCs. The review by Pairon et al [6] also emphasized that users value OSCs for their ability to support health-related decisions, especially in determining whether to seek medical care. The results of this study reinforce this, as individuals who perceived OSCs as helpful in making better medical care choices were nearly 3 times more likely to use them.

Compliance with OSC recommendations has also been a point of focus in the literature. Previous studies reported varying levels of adherence to OSC advice, with compliance rates ranging from 57% to 67% [6], although in our study, we report a significantly higher compliance rate, with 87.9% (557/634) of participants following OSC recommendations most or all of the time. This higher rate of adherence may reflect an increasing reliance on digital health tools, particularly in the context of the COVID-19 pandemic, which has accelerated the adoption of telemedicine and web-based health resources. Further research is needed to continue monitoring these trends and to explore the long-term impact of OSC use on health care outcomes.

Finally, issues relating to perceived accessibility, accuracy, security, and privacy of OSCs were also identified by Aboueid et al [20] in their qualitative study exploring young adults' perspectives on the use of OSCs. Most of their respondents thought of OSCs as more useful for self-triage than self-diagnosis, which reflects the fact that only 62.9% (399/634) of the respondents in this study trusted the diagnosis provided by the OSC, compared with 78.9% (500/634) for the triage suggestion.

Implications for Research

Although this study identified demographic disparities in the use of OSCs, further research is warranted to understand the underlying reasons for these disparities. Research focusing on the sociocultural factors, digital literacy, and health care-seeking behaviors among different demographic groups could provide valuable insights into addressing disparities and promoting equitable access to OSC. In addition, this study's findings highlight the importance of usability, effectiveness, and trust in driving the adoption and use of OSCs. Future research could investigate the specific features and functionalities of these tools, such as user interface design and decision support algorithms that contribute to their perceived usability, effectiveness, and adherence to the recommendations.

Our study highlighted concerns regarding the privacy and health information security of OSC, which could impact their acceptance and use, necessitating the development of robust frameworks, regulatory standards, and guidelines for OSC platforms to ensure transparency, accuracy, and user privacy. Additionally, studies investigating the potential implications of OSC on health care inequalities and the doctor-patient relationship are essential for informing policy and practice, whereas research exploring effective strategies for educating users about the capabilities and limitations of these tools, as well as enhancing communication and collaboration between users and health care providers, could help build trust and confidence in OSC.

Future research should focus on understanding the sociocultural factors influencing OSC use and developing strategies to address privacy and security concerns. Additionally, efforts to improve the usability and reliability of OSCs, alongside targeted interventions to promote equitable access, are essential for integrating these tools effectively into the health care system. By addressing these issues, OSCs can play a key role in supporting self-care and improving health care accessibility and efficiency in the United Kingdom.

Conclusions

This study provides insights into the factors influencing the use of OSCs in the United Kingdom, highlighting both their increasing and widespread adoption and some of the concerns associated with these digital health tools that will likely be crucial to inform the self-care journey of people in different settings and from different walks of life. While most of the respondents to our survey of British adults used OSCs, particularly the NHS 111 service, primarily for understanding symptoms and determining the need for medical care, younger individuals, females, and those with children are more likely to use OSCs overall. Ease of use, perceived helpfulness in medical decision-making, and trust in the accuracy of diagnoses and triage recommendations are key factors driving OSC use, but these are coupled to concerns about privacy, health information security, and the potential for OSCs to exacerbate health care inequalities, posing significant barriers to their adoption. The fear of OSCs replacing traditional consultations with health care professionals remains common among users, and these concerns must be addressed to enhance user trust and maximize the benefits of OSCs in health care delivery.

Acknowledgments

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. AE-O, ALN, AM, and BH are supported by the National Institute for Health and Care Research (NIHR) Applied Research Collaboration Northwest London. The views expressed are those of the authors and not necessarily those of the National Health Service, the NIHR, or the Department of Health and Social Care. The authors confirm that no use of generative artificial intelligence was made at any stage of this work, either in conducting the research or in the drafting of the manuscript.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

XSL•FO
RenderX

Authors' Contributions

All authors provided substantial contributions to the conception (AE-O, ER-S, AM, and BH), design (AE-O, ER-S, MAA, SA, and BH), acquisition (AE-O, ER-S, MAA, and SA), and interpretation (AE-O, ER-S, MAA, SA, ALN, BH, and AM) of study data and approved the final version of the paper. AE-O took the lead in planning the study with support from co-authors. MAA, SA, and ER-S carried out the data analysis with support from AE-O. AE-O is the guarantor.

Conflicts of Interest

BH worked for an online symptom checker (Healthily) between November 2019 and May 2021. The authors declare they have no other competing interests.

Multimedia Appendix 1
Survey export (raw xls).
[XLSX File, 17 KB - formative_v9i1e65314_app1.xlsx]

Multimedia Appendix 2
Main survey findings.
[PDF File, 156 KB - formative_v9i1e65314_app2.pdf]

Checklist 1
CHERRIES checklist.
[PDF File, 87 KB - formative_v9i1e65314_app3.pdf]

References

1. WHO health workforce. World Health Organization. 2024. URL: https://www.who.int/health-topics/health-workforce#tab=tab_1 [accessed 2025-08-25]
2. Cooksley T, Clarke S, Dean J, et al. NHS crisis: rebuilding the NHS needs urgent action. BMJ 2023:1. [doi: 10.1136/bmj.p1]
3. Jones S, Moulton C, Swift S, et al. Association between delays to patient admission from the emergency department and all-cause 30-day mortality. Emerg Med J 2022 Mar;39(3):168-173. [doi: 10.1136/emermed-2021-211572] [Medline: 35042695]
4. Lober WB, Flowers JL. Consumer empowerment in health care amid the internet and social media. Semin Oncol Nurs 2011 Aug;27(3):169-182. [doi: 10.1016/j.soncn.2011.04.002] [Medline: 21783008]
5. Chambers D, Cantrell AJ, Johnson M, et al. Digital and online symptom checkers and health assessment/triage services for urgent health problems: systematic review. BMJ Open 2019 Aug;9(8):e027743. [doi: 10.1136/bmjopen-2018-027743] [Medline: 31375610]
6. Pairon A, Philips H, Verhoeven V. A scoping review on the use and usefulness of online symptom checkers and triage systems: how to proceed? Front Med 2022;9:1040926. [doi: 10.3389/fmed.2022.1040926] [Medline: 36687416]
7. Internet access – households and individuals, Great Britain: 2020 2021. Office for National Statistics (ONS). URL: https://tinyurl.com/bdcwd9cb [accessed 2025-08-25]
8. Espinoza J, Crown K, Kulkarni O. A guide to chatbots for COVID-19 screening at pediatric health care facilities. JMIR Public Health Surveill 2020 Apr 30;6(2):e18808. [doi: 10.2196/18808] [Medline: 32325425]
9. Semigran HL, Linder JA, Gidengil C, Mehrotra A. Evaluation of symptom checkers for self diagnosis and triage: audit study. BMJ 2015 Jul 8;351:h3480. [doi: 10.1136/bmj.h3480] [Medline: 26157077]
10. El-Osta A, Webber I, Alaa A, et al. What is the suitability of clinical vignettes in benchmarking the performance of online symptom checkers? an audit study. BMJ Open 2022 Apr 27;12(4):e053566. [doi: 10.1136/bmjopen-2021-053566] [Medline: 35477872]
11. Riboli-Sasco E, El-Osta A, Alaa A, et al. Triage and diagnostic accuracy of online symptom checkers: systematic review. J Med Internet Res 2023;25:e43803. [doi: 10.2196/43803]
12. Arellano Carmona K, Chittamuru D, Kravitz RL, Ramondt S, Ramírez AS. Health information seeking from an intelligent web-based symptom checker: cross-sectional questionnaire study. J Med Internet Res 2022;24(8):e36322. [doi: 10.2196/36322]
13. Wallace W, Chan C, Chidambaram S, et al. The diagnostic and triage accuracy of digital and online symptom checker tools: a systematic review. npj Digit Med 2022;5(1):118. [doi: 10.1038/s41746-022-00667-w]
14. Kopka M, Scatturin L, Napierala H, et al. Characteristics of users and nonusers of symptom checkers in Germany: cross-sectional survey study. J Med Internet Res 2023;25:e46231. [doi: 10.2196/46231]
15. Morse KE, Ostberg NP, Jones VG, Chan AS. Use characteristics and triage acuity of a digital symptom checker in a large integrated health system: population-based descriptive study. J Med Internet Res 2020;22(11):e20549. [doi: 10.2196/20549]

XSL•FO
RenderX

16. Meyer AND, Giardina TD, Spitzmueller C, Shahid U, Scott TMT, Singh H. Patient perspectives on the usefulness of an artificial intelligence–assisted symptom checker: cross-sectional survey study. J Med Internet Res 2020 Jan 30;22(1):e14679. [doi: 10.2196/14679] [Medline: 32012052]
17. Painter A, Hayhoe B, Riboli-Sasco E, El-Osta A. Online symptom checkers: recommendations for a vignette-based clinical evaluation standard. J Med Internet Res 2022 Oct 26;24(10):e37408. [doi: 10.2196/37408] [Medline: 36287594]
18. Fan X, Chao D, Zhang Z, Wang D, Li X, Tian F. Utilization of self-diagnosis health chatbots in real-world settings: case study. J Med Internet Res 2021 Jan 6;23(1):e19928. [doi: 10.2196/19928] [Medline: 33404508]
19. You Y, Ma R, Gui X. User experience of symptom checkers: a systematic review. AMIA Annu Symp Proc 2022:1198-1207. [Medline: 37128443]
20. Aboueid S, Meyer S, Wallace JR, Mahajan S, Chaurasia A. Young adults' perspectives on the use of symptom checkers for self-triage and self-diagnosis: qualitative study. JMIR Public Health Surveill 2021;7(1):e22637. [doi: 10.2196/22637]
21. Turnbull J, Prichard J, MacLellan J, Pope C. eHealth literacy and the use of NHS 111 online urgent care service in England: cross-sectional survey. J Med Internet Res 2024;26:e50376. [doi: 10.2196/50376]
22. Prolific. URL: https://www.prolific.com [accessed 2025-08-25]
23. Miller S, Gilbert S, Virani V, Wicks P. Patients' utilization and perception of an artificial intelligence–based symptom assessment and advice technology in a british primary care waiting room: exploratory pilot study. JMIR Hum Factors 2020;7(3):e19713. [doi: 10.2196/19713]
24. Powley L, McIlroy G, Simons G, Raza K. Are online symptoms checkers useful for patients with inflammatory arthritis? BMC Musculoskelet Disord 2016 Dec;17(1):1-6. [doi: 10.1186/s12891-016-1189-2]
25. Eysenbach G. Improving the quality of web surveys: the checklist for reporting results of internet e-surveys (CHERRIES). J Med Internet Res 2004;6(3):e34. [doi: 10.2196/jmir.6.3.e34]
26. NHS 111. NHS England. URL: https://www.england.nhs.uk/urgent-emergency-care/nhs-111 [accessed 2025-08-25]
27. Healthline. URL: https://www.healthline.com/about/about-us [accessed 2025-08-25]
28. Religion, England and Wales: census 2021. Office for National Statistics. URL: https://www.ons.gov.uk/peoplepopulationandcommunity/culturalidentity/religion/bulletins/religionenglandandwales/census2021 [accessed 2024-04-04]

Abbreviations

aOR: adjusted odds ratio
CHERRIES: Checklist for Reporting Results of Internet E-Surveys
GP: general practitioner
NHS: National Health Service
OSC: online symptom checker

Edited by A Mavragani; submitted 12.08.24; peer-reviewed by S Kujala, Z Narita; revised version received 01.04.25; accepted 09.07.25; published 15.09.25.

Please cite as:
El-Osta A, Riboli-Sasco E, Al Ammouri M, Altalib S, Luisa Neves A, Majeed A, Hayhoe B
Factors Influencing the Use of Online Symptom Checkers in the United Kingdom: Cross-Sectional Study
JMIR Form Res 2025;9:e65314
URL: https://formative.jmir.org/2025/1/e65314
doi:10.2196/65314

© Austen El-Osta, Eva Riboli-Sasco, Mahmoud Al Ammouri, Sami Altalib, Ana Luisa Neves, Azeem Majeed, Benedict Hayhoe. Originally published in JMIR Formative Research (https://formative.jmir.org), 15.9.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Interaction of Long Working Hours and Sense of Coherence on Objective Total Sleep Time: Cross-Sectional Study From the SLEPT Study

Kei Muroi[1], MD, PhD; Emi Morita[1,2], PhD; Sumire Matsumoto[1], PhD; Asuka Ishihara[1], PhD; Sumi Hasegawa[1], MSc; Mami Ishitsuka[3,4], MSN; Daisuke Hori[5], MD, PhD; Shorato Doki[5], MD, PhD; Tsukasa Takahashi[5], MD, PhD; Shin-ichiro Sasahara[5], MD, PhD; Takashi Kanbayashi[1,6], MD, PhD; Masashi Yanagisawa[1,7], MD, PhD; Makoto Satoh[1], MD, PhD; Ichiyo Matsuzaki[1,5], MD, PhD

[1]International Institute for Integrative Sleep Medicine (WPI-IIIS), Tsukuba Institute for Advanced Research (TIAR), University of Tsukuba, Tsukuba, Japan

[2]Forestry and Forest Products Research Institute, Forest Research and Management Organization, Tsukuba, Japan

[3]Graduate School of Comprehensive Human Sciences, University of Tsukuba, Tsukuba, Japan

[4]School of Nursing at Narita, International University of Health and Welfare, Narita, Japan

[5]Institute of Medicine, University of Tsukuba, 1-1-1 Tennodai, Ibaraki, Tsukuba, Japan

[6]Ibaraki Prefectural Medical Center of Psychiatry, Kasama, Ibaraki, Japan

[7]Department of Molecular Genetics, University of Texas Southwestern Medical Center, Dallas, TX, United States

Corresponding Author:
Shin-ichiro Sasahara, MD, PhD
Institute of Medicine, University of Tsukuba, 1-1-1 Tennodai, Ibaraki, Tsukuba, Japan

Abstract

Background: Long working hours are a significant risk factor for reduced sleep duration among workers. Sense of coherence (SOC), a dispositional orientation that enhances resilience to daily stressors, may serve as a protective factor for sleep duration under work-related stress. However, previous studies examining SOC and sleep duration have relied on subjective measures, which may be subject to recall bias. The interaction between long working hours and SOC on objective sleep duration has not been previously investigated.

Objective: This study aimed to investigate the interaction between SOC and long working hours on objectively measured total sleep time (TST) among Japanese workers. We hypothesized that individuals with higher SOC would demonstrate less susceptibility to sleep reduction associated with long working hours compared to those with lower SOC.

Methods: A cross-sectional survey was conducted from 2016 to 2017 as part of the Sleep Epidemiology Project at the University of Tsukuba (SLEPT) study among workers in Japan. The 13-item Sense of Coherence scale (SOC-13) was administered to assess participants' SOC levels, and weekly working hours were self-reported. TST was objectively measured using actigraphy devices worn continuously for 1 week. Long working hours were defined as ≥50 hours per week. Multiple regression analysis was performed with TST as the dependent variable, including long working hours, SOC-13, and their interaction term as independent variables. Simple slope analysis was conducted to examine the interaction effect at different SOC levels of ±1 SD from the mean.

Results: A total of 540 workers were included in the final analysis. The study population had a mean age of 43.2 years, with 41.1% female participants. Mean TST was 322.0 (SD 58.0) minutes, and mean SOC score was 58.3 (SD 11.9). Long working hours were reported by 304 (56.3%) participants. Multiple regression analysis revealed a significant main effect of long working hours on reduced TST (β=−.115, *P*=.023), with workers in the long hours group sleeping 13.5 minutes less per night. Importantly, a significant interaction between long working hours and SOC was observed (β=.147, *P*=.026), indicating that SOC moderated the relationship between long working hours and sleep duration. Simple slope analysis demonstrated that at low SOC levels (−1 SD), long working hours were significantly associated with reduced TST (β=−24.7, *P*=.0015).

Conclusions: Workers with lower SOC experienced significantly greater sleep reduction when working long hours, while those with higher SOC maintained relatively stable sleep duration despite extended work schedules. These findings suggest that interventions aimed at enhancing SOC may be effective in protecting workers' sleep health, particularly for those unable to reduce their working hours.

(JMIR Form Res 2025;9:e73784) doi:10.2196/73784

XSL•FO
RenderX

KEYWORDS

actigraphy; long working hours; sleep duration; salutogenesis; worker; employee

Introduction

Sleep duration is an important biomarker in human sleep, and short sleep duration is associated with increased cardiovascular disease, diabetes, mental disorders, and all-cause mortality [1-3]. In addition, short sleep duration is a risk for occupational accidents, incidents, and presenteeism among workers [4,5]. Japan has the lowest average sleep duration among high-income countries [6], especially among the working-age population. Ikeda et al [7] report that permanent daytime workers in Japan sleep about 360 minutes on average. It is important to identify factors that protect sleep duration in this population.

In workers, long working hours are a risk factor for insufficient sleep [8,9]. International organizations and countries have established standards for maximum working hours. For example, International Labor Organization conventions stipulate a maximum of 48 hours of work per week [10]. In Japan, the Labor Standards Law stipulates that an employer shall not make an employee work more than 8 hours per day or 40 hours per week [11]. An agreement should be concluded between an employer and an employee and submitted to the Labor Standards Inspection Office when the employer wants to extend working hours or have an employee work on holidays. As a general rule, the upper limit of overtime work is 45 hours per month or 360 hours per year. When there are special temporary circumstances, the upper limit can be up to 720 hours per year or less than 100 hours per month or up to an average of 80 hours per month over multiple months [11]. Assuming 45 hours of overtime per month, this translates into an approximate 50-hour workweek, which is roughly in line with the conventions set forth by the International Labor Organization. Against this background, long working hours are often defined as around 50 hours per week, and a meta-analysis by Wong et al [9] reported that working hours in excess of around 50 hours per week are associated with risk of cardiovascular and cerebrovascular diseases, depression, anxiety disorder, and sleep disorder in workers. They also suggest that the mechanism of short sleep duration due to long working hours is related to high work stress and the sympathetic nervous system [12,13], in addition to difficulty in securing sleep. Research on the effect of increased sleep on reduced work hours is limited, although a study by Schiller et al [14] reported that a 25% reduction in weekly work hours resulted in an increase of 23 minutes of sleep. On the other hand, it has been noted that sleep duration may not increase [15]; therefore, methods other than the reduction of working hours are important to maintain workers' sleep duration.

Salutogenesis is an approach to human health that examines the factors that promote and maintain physical and mental health [16]. Antonovsky defined sense of coherence (SOC) as the ability to maintain health in the face of stressful events and situations [17,18]. SOC is defined as comprehensibility, manageability, and meaningfulness [17,18]. People with high SOC have a higher ability to cope with stress and maintain physical and mental health in the face of stress, and associations between SOC and depression and all-cause mortality have been reported [19,20]. Furthermore, SOC has been reported to protect against stress in workers and even against serious stresses such as post-traumatic stress disorder [21,22]. Long working hours are a stressful situation, and workers' SOC may be associated with sleep duration.

To our knowledge, only 2 studies have previously examined the association between SOC and sleep duration: in a study of software office workers, Morita et al [23] reported that workers with high SOC maintained more than 6 hours of sleep per day as compared with workers with low SOC; in a study of civil servants, Kósa et al [24] also reported a U-shaped association between SOC and sleep duration, with a positive association between SOC and sleep duration of up to 8 hours per day. However, the interviews in these studies focused only on subjective sleep duration, which may be subject to recall bias [25]. Since subjective sleep duration is reported to be longer than actual sleep duration, it is more important to measure objective sleep duration. The gold standard for objective sleep duration is polysomnography (PSG) [26], but it is not suitable for measuring large populations. Recently, actigraphy, with its built-in accelerometer, has been used in epidemiologic studies to measure objective sleep duration, with a reported agreement rate of approximately 90% with PSG [27,28]. Furthermore, previous studies have not taken into account the factor of working hours in the relationship between sleep duration and SOC, and it would be useful to examine the interaction between long working hours and SOC, as SOC can help maintain health under stress. It is important to examine interactions in epidemiology to consider effective interventions to improve outcomes with limited resources [29]. By examining interactions, we can gain clues about the reasons and mechanisms whereby exposure factors affect outcomes [30]. While the interaction between SOC and long working hours in workers' psychological stress has been studied [21], the interaction between SOC and long working hours with sleep duration has not been reported. Examining the interaction between long working hours and SOC in objective sleep can provide important insights into the protection of workers' sleep.

Hence, we examined the interaction between long working hours and SOC by using actigraphy to measure sleep duration in workers.

Methods

Study Design and Recruitment

Study Framework and Timeline

Data for this study were obtained from the Sleep Epidemiology Project at University of Tsukuba (SLEPT) study, a comprehensive sleep epidemiology investigation conducted jointly by the University of Tsukuba and the International Institute for Integrative Sleep Medicine, University of Tsukuba. The SLEPT study uses a cross-sectional design with prospective data collection conducted from 2016 to 2017. Data were collected through multiple complementary methods:

XSL•FO
RenderX

self-administered questionnaires, participant sleep diaries maintained for 7 consecutive days, and objective actigraphy measurements.

Study Settings and Participant Recruitment

The study was conducted across 4 diverse workplace settings in Japan: a national university hospital in Ibaraki Prefecture, a national research institute, a corporate research institute, and a medical corporation in Tokyo. To maximize participation across diverse worker populations, a multi-modal recruitment approach was implemented during the recruitment period from 2016 to 2017. The recruitment strategy encompassed multiple complementary approaches. The participants were recruited by use of flyers, posters, workplace group emails, and workplace web-based bulletin boards. Some participants were referred by research staff or other study participants.

During the systematic recruitment process conducted across all 4 sites, potential participants received comprehensive study information including objectives, time commitments, data collection procedures, and the voluntary nature of participation. Interested individuals were scheduled for enrollment sessions where informed consent was obtained and study materials were distributed.

Participation Incentives and Feedback

Although no monetary compensation was provided, several non-monetary incentives were offered to encourage participation and provide value to workers. All participants who completed the study protocol received personalized sleep reports offering insights into their individual sleep patterns and characteristics. Additionally, basic sleep hygiene information and educational materials were provided to enhance participants' understanding of healthy sleep practices.

Measurements

Basic Attributes

Age, sex, height, weight, drinking habits, exercise habits, and smoking status were surveyed. Survey items related to exercise habits were "Yes" (="I exercise at least 30 min, twice a week, for 1 y") or "No" (="other") [31]; those related to drinking habits were "Yes" (="more than once a week") or "No" (="less than once a week"); and those related to smoking habits were "never smoker," "former smoker," or "current smoker." They were also asked about any diseases the participant was currently being treated for, and any medications the participant was currently taking were also surveyed.

Occupational Factors

The participants were asked about their working hours (hours) for the last week, their daily 1-way commuting time (minutes), and their employment status as defined by the Ministry of Health, Labor and Welfare (full-time regular employees or regular staff, full-time employees with short hours, part-time employees, contract employees, managers or staff, and others). They were also asked about whether they were currently engaged in shift work.

The Brief Scales for Job Stress (BSJS) [32] was used for job stress; the BSJS is a 20-item questionnaire based on a job demands-control-support model similar to the Job Content Questionnaire [33]. All participants were asked, "Please select the statement about your current work environment that most closely matches your feelings." Responses were rated on a 4-point scale (1="disagree" to 4="agree") and mean scores (range: 1.00 - 4.00) were calculated for 6 subscales: "workload," "mental workload," "interpersonal relationships," "job control," "reward from work," and "support from colleagues and superiors." The reliability and validity of this scale have been reported [34-36]. Workload, mental workload, and interpersonal relationship were defined as stress factors, and job control, reward from work, and support from colleagues and superiors as stress-buffering factors. The Cronbach α coefficient for the BSJS in this study was 0.766.

Objective Total Sleep Time

The MTN-220 (ACOS Co Ltd.) was used to record total sleep time (TST). Participants wore the MTN-220 on their trunk by clipping it to their trousers or pants, 24 hours a day for 1 week, except when bathing; the MTN-220 records activity using a built-in 3-axis acceleration sensor [37]. In a laboratory validation study, the agreement with PSG data ranged from 84.7% to 85.4% [37]. Data were extracted from the MTN-220 device by use of SleepSignAct2 software (Kissei Comtec Co. Ltd.). Sleep and wake detection from MTN-220 data followed the previously reported algorithm of SleepSignAct2 [37]. A sleep diary was used to measure subjective sleep patterns. Participants were required to estimate their sleep duration each morning and record the results in a sleep diary. After data acquisition, the data from the sleep measurements and the sleep diary were checked for consistency. Objective sleep duration on actigraphy was calculated after determining the wake times and bedtimes in the participants' sleep diaries, and the TST (minutes) was the average of all nights of data collected by actigraphy.

Sense of Coherence

The 13-item version of the Sense of Coherence scale (SOC-13) was used to assess SOC [18]. A shortened version of the original 29-item SOC scale, the SOC-13 scale was developed by Antonovsky [18,38]. The SOC-13 includes items categorized into 3 subscales: Comprehensibility (consisting of 5 items: 2, 6, 8, 9, and 11), Manageability (consisting of 4 items: 3, 5, 10, and 13), and Meaningfulness (consisting of 4 items: 1, 4, 7, and 12). The final score for each participant is the sum of the scores of the items, including reversed scores for questions 1, 2, 3, 7, and 10 (where a score of 7 is considered as 1, 6 as 2, 5 as 3, 4 as 4, 3 as 5, 2 as 6, and 1 as 7). Scores range from 13.0 to 91.0 points, with higher total scores indicating a stronger SOC. The Japanese version of the SOC-13 questionnaire [39], which has been translated and validated in this language, was used in our study. Members of the Society for Theory and Research on Salutogenesis are authorized to use this questionnaire for academic research [40]. The authors, KM, SD, and SS, are members of the Society for Theory and Research on Salutogenesis. The Cronbach α coefficient for the SOC-13 in this study was 0.856.

XSL•FO
RenderX

Statistical Analysis

A total of 785 workers participated in the study. Written informed consent was obtained from all the participants. In total, 4 participants subsequently withdrew their consent, and their data were deleted. An additional 36 participants were excluded because they did not meet the eligibility criteria. This was due to inconsistencies between the sleep measurement data and the sleep diaries. Of the remaining 745 participants, 205 participants having sleep disorders such as sleep apnea, having a history of psychiatric disorders, taking sleeping pills or psychotropic drugs, working short hours or part-time, or missing data were excluded, leaving a total of 540 participants for analysis.

Long working hours were defined as more than 50 hours of work per week. BMI was calculated from height and weight and divided into 3 groups (underweight:<18.5, normal weight: 18.5 - 24.9, overweight:≥25), age was divided into 5 groups (20 - 29, 30 - 39, 40 - 49, 50 - 59, and 60 or older), and commuting time was doubled to daily 1-way commuting time. Multiple regression analysis was conducted with TST as the dependent variable and with long working hours, SOC-13, and the interaction term between the 2 variables as the independent variables. The SOC-13 was centralized to avoid multicollinearity due to the correlation between the main effect term and the interaction term. Age, sex, BMI, drinking habits, exercise habits, smoking status, BSJS, presence of shift work, and commuting time were entered as confounders in the multiple regression model. When an interaction was observed, we performed simple slope analysis and plotted the results based on the studies by Aiken and West [41] and Cohen et al [42] and tested the simple slopes of the regression lines corresponding to combinations of low (one standard deviation below the mean: −1SD) and high (one SD above the mean:+1 SD) levels of the SOC. Furthermore, we made the plot of the 2-way interaction between long working hours and SOC. The statistical software used was R (version 4.3.1; R Foundation for Statistical Computing), the significance level was 0.05 on both sides, and the pequod package [43] was used for single slope analysis.

Ethical Considerations

Ethics Review and Approvals

This study was conducted in accordance with the ethical standards of the Declaration of Helsinki and was approved by the Medical Ethics Committee of the University of Tsukuba (approval number: 1065 - 10). The research protocol underwent comprehensive review to ensure compliance with national research ethics guidelines and institutional policies for human subjects research. All procedures involving human participants were conducted in accordance with the ethical standards of the institutional or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

Informed Consent

Written informed consent was obtained from all participants prior to their enrollment in the study. Participants were provided with comprehensive information about the study objectives, procedures, potential risks and benefits, data collection methods, and their rights as research participants. The consent process included detailed explanations of the actigraphy measurement procedures, questionnaire requirements, and sleep diary maintenance. Participants were explicitly informed of their right to withdraw from the study at any time without penalty or impact on their employment or health care. Four participants subsequently exercised their right to withdraw consent, and their data were immediately deleted from all study databases and analyses in accordance with their wishes and ethical requirements.

Privacy and Confidentiality Protection

All study data were deidentified and anonymized to protect participant privacy and confidentiality. Personal identifiers were separated from research data and stored securely with access restricted to authorized research personnel only. Study databases were password-protected and maintained on secure servers with appropriate data encryption protocols. Participants were assigned unique study identification numbers to ensure anonymity throughout the data collection, analysis, and reporting processes. No individual participants can be identified from any published results, tables, figures, or supplementary materials. Data-sharing procedures adhere to institutional policies and ethical requirements for protection of participant privacy.

Compensation and Incentives

No monetary compensation was provided to study participants. However, as an incentive for participation and to provide value to participants, individual sleep measurement results and basic sleep health information were provided as feedback to all participants who completed the study protocols. This feedback was presented in a summary format to help participants better understand their sleep patterns while maintaining the scientific integrity of the study. The feedback process was designed to be educational rather than diagnostic, with appropriate disclaimers about the research nature of the measurements.

Participant Identification and Images

This manuscript and all supplementary materials contain no images, photographs, or other materials that could potentially identify individual study participants. All data presentations use aggregated statistical summaries, and no individual-level data points are displayed in ways that could compromise participant anonymity. The research team ensured that all tables, figures, and data visualizations maintain participant confidentiality and comply with privacy protection requirements.

Results

Table 1 shows the descriptive characteristics of the participants. The mean overall sleep duration was 322 (SD 58.0) minutes, and the mean SOC was 58.3 (SD 11.9). The long hours group (≥50 hours/week) showed a shorter mean TST (<50 hours/week: mean 332 SD 56.1 minutes; ≥50 hours/week: mean 315, SD 58.5 minutes; P<.001); no difference was found between the 2 groups for SOC (<50 hours/week: mean 58.0, SD 11.8; ≥50 hours/week: mean 58.5, SD 12.0; P=.62).

XSL•FO
RenderX

Table . Demographic and occupational characteristics of Japanese workers by weekly working hours in a cross-sectional analysis from the SLEPT[a] study (N=540). Cross-sectional analysis from the SLEPT study conducted in 4 workplaces in Ibaraki and Tokyo prefectures, Japan (2016-2017). Participants were full-time daytime workers aged 20-60 years. Working hours were categorized as <50 hours/week (n=236, 43.7%) and ≥50 hours/week (n=304, 56.3%).

	Working hours		*P* value
	<50 hours/week (n=236)	≥50 hours/week (n=304)	
Sex, n (%)			
Male	85 (36.2)	180 (59.2)	<.001[b]
Female	150 (63.8)	124 (40.8)	
Age (years), n (%)			
20 - 29	35 (14.8)	54 (17.8)	.12[b]
30 - 39	70 (29.7)	82 (27)	
40 - 49	68 (28.8)	73 (24)	
50 - 59	46 (19.5)	82 (27)	
60 and older	17 (7.2)	13 (4.3)	
BMI, n (%)			
Underweight	23 (9.7)	18 (5.9)	.042[b]
Normal weight	166 (70.3)	242 (79.6)	
Overweight	47 (19.9)	44 (14.5)	
Drinking habit, n (%)			
No	81 (34.3)	69 (22.7)	.004
Yes	155 (65.7)	235 (77.3)	
Exercise habit, n (%)			
No	188 (79.7)	244 (80.3)	.91
Yes	48 (20.3)	60 (19.7)	
Smoking status, n (%)			
Never smoker	180 (76.3)	229 (75.3)	.94
Current smoker	20 (8.5)	25 (8.2)	
Former smoker	36 (15.3)	50 (16.4)	
Shift work, n (%)			
No	219 (92.8)	262 (86.2)	.018
Yes	17 (7.2)	42 (13.8)	
BSJS[c], mean (SD)			
Workload	1.83 (0.630)	2.60 (0.740)	<.001[d]
Mental workload	1.95 (0.700)	2.46 (0.710)	<.001[d]
Interpersonal relationship	1.75 (0.670)	2.02 (0.78)	<.001[d]
Reward from work	2.81 (0.820)	3.01 (0.77)	.004[d]
Support from colleagues and superiors	3.00 (0.650)	2.90 (0.630)	.072[d]
Job control	2.83 (0.720)	2.88 (0.770)	.500[d]
Commuting time	73.4 (75.5)	62.9 (50.2)	.053[d]
SOC-13[e]	58.0 (11.8)	58.5 (12.0)	.62[d]
TST[f]	332 (56.1)	315 (58.5)	<.001[d]

XSL•FO
RenderX

[a]SLEPT: Sleep Epidemiology Project at the University of Tsukuba.

[b]Chi-square test.

[c]BSJS, Brief Scales for Job Stress.

[d]Unpaired *t* test.

[e]SOC-13: 13-item Sense of Coherence scale.

[f]TST: total sleep time.

Table 2 shows the results of the multiple regression analysis with TST as the dependent variable and long hours worked, SOC-13, and the interaction of the 2 variables as the independent variables. A main effect of long working hours was found (β=−.115, *P*=.023), as well as an interaction between long working hours and SOC (β=.147, *P*=.026), while SOC alone was not associated with TST (β=−.0537, *P*=.45).

Table 3 shows the results for the covariates in the multiple regression analysis. Female sex (β=0.155, *P*=.0020), having an overweight BMI (β=−0.0954, *P*=.030), and working a shift (β=−0.0914, *P*=.044) were associated with TST. The adjusted *R*-squared value on multiple regression analysis was 0.0560.

Table 4 shows the simple slope analysis for moderating effects at (−1 SD) low and (+1 SD) high levels of SOC between TST and long working hours. Only low levels of SOC were significant. At low levels of SOC, long working hours significantly and negatively predicted TST (β=−24.7, t_{516}=−3.19, *P*=.0015). However, at high levels of SOC, long working hours did not significantly predict TST (β=−2.21, t_{516}=−0.29, *P*=.78).

Figure 1 shows the plot of the 2-way interaction between long working hours and SOC on TST.

Table . Multiple regression analysis results examining the interaction between long working hours and sense of coherence on objectively measured total sleep time among Japanese workers (N=540). Results from the cross-sectional SLEPT[a] study of Japanese workers across 4 workplaces (2016-2017). Dependent variable: 7-day actigraphy-measured total sleep time. The model is adjusted for demographic, lifestyle, and occupational factors.

	B[b]	SE	*t* test (*df*)	β[c]	*P* value
Working hours (Ref[d] <50 hours/week)					
≥50 hours/week	−13.5	5.89	−2.29 (516)	−.115	.023
SOC-13[e]	−0.261	0.344	−0.760 (516)	−.0537	.45
Working hours≥50 hours/week×SOC-13	0.943	0.423	2.23 (516)	.147	.026

[a]SLEPT: Sleep Epidemiology Project at the University of Tsukuba.

[b]B: nonstandard coefficient.

[c]β: standard coefficient.

[d]Ref: reference.

[e]SOC-13: 13-item Sense of Coherence scale.

XSL•FO
RenderX

Table . Covariate effects in multiple regression analysis predicting total sleep time among Japanese workers with long working hours interaction model (N=540). Effects of demographic, lifestyle, and occupational covariates in the working hours×SOC[a] interaction model from the SLEPT[b] study (Japan, 2016-2017).

	B[c]	SE	*t* test (*df*)	β[d]	*P* value
Female sex (ref: male)	18.0	5.80	3.11 (516)	.155	.0020
Age (years; ref: 20 - 29 years)					
30 - 39	−2.43	7.81	−0.311 (516)	−.0188	.76
40 - 49	−15.3	8.10	−1.89 (516)	−.116	.059
50 - 59	−10.6	8.78	−1.21 (516)	−.0777	.23
60 and older	−19.8	13.0	−1.52 (516)	−.0783	.13
BMI (ref[e]: normal weight)					
Underweight	−0.157	9.47	−0.0165 (516)	−.000720	.99
Overweight	−14.8	6.81	−2.17 (516)	−.0954	.030
Having a drinking habit (ref: not having one)	2.87	5.74	0.500 (516)	.0221	.62
Smoking status (ref: never smoker)					
Current smoker	−0.961	9.24	−0.104 (516)	−.00458	.92
Former smoker	2.18	7.12	0.306 (516)	.0138	.76
Having an exercise habit (ref: not having one)	−0.998	6.34	−0.158 (516)	−.00688	.88
Doing shift work (ref: not doing shift work)	−17.0	8.41	−2.02 (516)	−.0914	.044
BSJS[f] workload	−3.39	4.29	−0.790 (516)	−.0465	.43
BSJS mental workload	0.774	4.52	0.171 (516)	.0100	.86
BSJS interpersonal relationship	5.60	4.16	1.34 (516)	.0716	.18
BSJS reward from Work	−1.09	3.83	−0.285 (516)	−.0150	.78
BSJS job control	4.28	3.90	1.10 (516)	.0548	.27
BSJS support from colleagues and superiors	2.81	4.71	0.596 (516)	.0310	.55
Commuting time	−0.0323	0.0400	−0.808 (516)	−.0348	.42

[a]SOC: sense of coherence.

[b]SLEPT: Sleep Epidemiology Project at the University of Tsukuba.

[c]B: nonstandard coefficient.

[d]β: standard coefficient.

[e]Ref: reference.

[f]BSJS: Brief Scales for Job Stress.

XSL•FO
RenderX

Table . Simple slope analysis examining the effect of long working hours on total sleep time at different levels of sense of coherence among Japanese workers (N=540). Analysis following significant working hours×SOC[a] interaction in the SLEPT[b] study (Japan, 2016-2017).

	β[c]	SE	*t* test (*df*)	*P* value
Low SOC-13[d] (−1 SD)	−24.7	7.76	−3.19 (516)	.0015
High SOC-13 (+1 SD)	−2.21	7.76	−0.29 (516)	.78

[a]SOC: sense of coherence.

[b]SLEPT: Sleep Epidemiology Project at the University of Tsukuba.

[c]β: simple slope.

[d]SOC-13: 13-item Sense of Coherence scale.

Figure 1. Cross-sectional analysis from the Sleep Epidemiology Project at University of Tsukuba (SLEPT) study conducted across 4 workplaces in Ibaraki and Tokyo prefectures, Japan (2016 - 2017). The plot illustrates the significant 2-way interaction (β=0.147, *P*=.026) between working hours (categorized as <50 vs ≥50 hours/week) and sense of coherence on total sleep time measured by 7-day waist-worn actigraphy. Low sense of coherence (SOC) represents scores 1 SD below the mean (–1 SD; triangle markers, approximately 46.4 points), while high SOC represents scores 1 SD above the mean (+1 SD; circle markers, approximately 70.2 points). Workers with low SOC showed a significant 24.7-minute reduction in sleep duration when working ≥50 hours/week (*P*=.0015), whereas workers with high SOC showed no significant sleep reduction (2.21 minutes, *P*=.78). The interaction demonstrates that a sense of coherence protects against sleep loss associated with long working hours. The y-axis shows total sleep time in minutes; the x-axis shows working hour categories. SOC-13: sense of coherence measured by 13-item scale; TST: total sleep time.

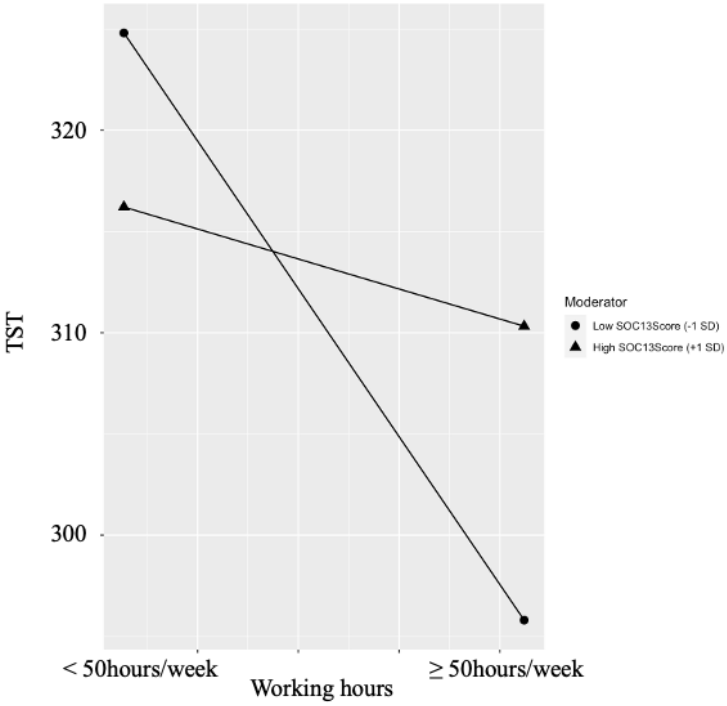

Discussion

Principal Findings

To our knowledge, this is the first report examining the association of SOC with objective sleep duration. The TST of the participants was 322 minutes, which was less than the national average. Previously, Takahashi et al [44] measured the sleep duration of 55 Japanese workers by use of actigraphy and reported an average TST of 384 minutes. By comparison, the sleep duration of the present study population was shorter; on the other hand, the mean SOC of the present study population was 58.5, which was comparable to the national survey sample [39]. The reliabilities of the BSJS and SOC-13 in this study were confirmed, as the Cronbach α coefficients exceeded 0.70 [45].

Multiple regression analysis showed that high BMI and shift work were associated with shorter sleep duration, consistent with the results of previous studies [46-48]. Our results showed that women tended to sleep more than men. In a previous study of sex differences in sleep duration among adult non–shift workers in 47 countries, men tended to have shorter sleep durations measured by wearable devices than those of women throughout their lifetime, consistent with our results [49]. On the other hand, sleep duration measured by actigraphy in Japanese was shorter for women than for men, which was speculated to be because women in Japan have more domestic roles [50]. The reason for the difference between our results and those of this previous study may be the method of recruiting participants in the present study. Since women are more health conscious and have more health behaviors than men [51,52], it

XSL•FO
RenderX

is possible that the study attracted women who were trying to get more sleep as a result of their health behaviors.

A decrease in TST was observed for long working hours (≥50 hours/week) alone, while an interaction between long working hours and SOC was associated with an increase in TST. Single slope analysis showed that both long hours worked groups decreased TST, but that lower SOC was associated with a significantly greater decrease in TST. The interaction results in this study suggest that the effect of long working stress varies with SOC. Recent studies suggest that sleep disturbances are not directly caused by stress but rather by stress coping and reactivity [53,54] and that those with higher SOC are more likely to develop coping behaviors and use more appropriate stress coping strategies, such as seeking assistance, during stress load and seeking help [18]. Since high SOC plays an important role in coping with stress, it is possible that workers' SOC may be protective in sleep. Furthermore, although studies are limited to health care workers, biological mechanisms have been reported to associate stress with SOC and the dorsolateral prefrontal cortex (DLPFC) [55,56], suggesting that high SOC is associated with a high-functioning DLPFC. With respect to stress-coping mechanisms, the prefrontal cortex plays an important role in translating stress experiences into adaptive behaviors by integrating cognitive and emotional behaviors and promoting neuroendocrine and autonomic nervous system flexibility in response to stress [57,58]. Those with higher SOC have associated higher functioning of the DLPFC and may be better adapted to work stress in long working hours. Further investigation of the biological mechanisms in stress, SOC, and sleep duration is warranted.

Implications

An approach that focuses on SOC to protect workers' duration is warranted [14,15], as it has been suggested that reducing working hours alone may not be sufficient to maintain sleep duration. SOC has been suggested to grow in adulthood, and to improve SOC, it is important to increase generalized resistance resources (GRRs), or resources that can be used in coping with stress [59,60]. Within the framework of SOC-GRRs, increasing resources in the workplace is important and can be used as an occupational health measure. Useful GRRs in the workplace include good interpersonal relationships, employee decision-making authority, and opportunities to use skills and knowledge [60]. For occupational health professionals and workplace managers, focusing on GRRs in the workplace and increasing GRRs could help increase workers' sleep duration in addition to reducing working hours.

Limitations

This study was a cross-sectional study, so causal relationships are unknown. In addition, there is a generalizability issue because the study population was white-collar workers. Furthermore, there is a possibility of selection bias, which is a healthy worker effect. The multiple regression analysis resulted in an adjusted R^2 of 0.0560, which is not a good fit for the model. It is possible that unmeasured confounders such as marital status and economic status [61,62] were not adjusted for in the present study. In addition, it has been reported that 31% to 55% of sleep duration is related to genetic factors [63,64], and future analyses should take genetic factors into account to construct a model to predict sleep duration. Furthermore, since a study of healthy college students aged 18 to 40 years reported that waist-worn actigraphy devices underestimated TST in comparison with arm-worn devices [65], it is possible that the present waist-worn actigraphy device (ACOS MTN-220) also underestimated TST in the same way. Therefore, electroencephalography-based sleep tracking devices similar to PSG would be suitable for more accurate evaluation of objective sleep measurement. Finally, since the working hours surveyed in this study are a subjective assessment, an objective assessment such as reporting using a timecard or smartphone applications would be more accurate [66].

Conclusions

The interaction between long working hours (≥50 hours/week) and SOC for workers indicated that lower SOC could lead to shorter sleep duration. The results suggest that an SOC-focused approach is important to maintain workers' sleep duration. Future studies should include large-scale follow-up and investigation of the biological mechanisms.

Acknowledgments

We wish to thank the participants of this study. We express our gratitude to Medirom, Inc. and the University of Tsukuba Health Center for their support and cooperation. We are grateful to Noriko Nakamura, Emi Sakurai, and Noriko Fujiwara of the University of Tsukuba International Institute for Integrative Sleep Medicine (WPI-IIIS) and Hiroyuki Koda of the Health and Safety Office for their valuable guidance and suggestions. We also wish to thank Shigeki Teradaira, Takumi Fukuda, Kohei Kawahara, and Hirokazu Tachikawa of University of Tsukuba for their help in data collection and data management. We express our gratitude to all the volunteers who donated through crowdfunding and the members of IIIS, University of Tsukuba, for their help in implementing crowdfunding. We also wish to thank Flaminia Miyamasu, Medical English Communications Center, Institute of Medicine, University of Tsukuba, for grammatical revision. This study was supported by JSPS Kakenhi grants (grant numbers JP16H03245 and JP21H02246). This work was also supported by Japan Society for the Promotion of Science (JSPS) KAKENHI Grant-in-Aid for Scientific Research (22K07571 and 24K10674 to TK). This work was also funded by volunteers through crowdfunding, READYFOR [67]. This research was funded by the World Premier International Research Center Initiative (WPI) by the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan, and Japan Agency for Medical Research and Development under Grant Number JP21zf0127005. No artificial intelligence tools were used in manuscript preparation.

XSL•FO
RenderX

Data Availability

The datasets generated or analyzed during this study are not publicly available due to ethical restrictions and privacy regulations to protect participant confidentiality, but anonymized data may be available from the corresponding author upon reasonable request and with appropriate ethical approval.

Authors' Contributions

Conceptualization: KM (lead), SS (supporting)
Data curation: EM (lead), KM (supporting)
Formal analysis: KM (lead)
Funding acquisition: MY (lead), MS (equal), EM (equal), TK (equal)
Investigation: EM (lead)
Methodology: KM (lead), EM (supporting)
Project administration: SS (lead), IM (equal)
Resources: EM (lead)
Software: KM (lead)
Supervision: SS (lead), EM (supporting)
Validation: KM (lead), SS (supporting)
Visualization: KM (lead)
Writing – original draft: KM (lead)
Writing – review & editing: KM (lead), SS (supporting), EM (supporting), SM (supporting), AI (supporting), SH (supporting), MI (supporting), DH (supporting), SD (supporting), TT (supporting), TK (supporting), MY (supporting), MS (supporting), IM (supporting)

Conflicts of Interest

None declared.

References

1. Svensson T, Saito E, Svensson AK, et al. Association of sleep duration with all- and major-cause mortality among adults in Japan, China, Singapore, and Korea. JAMA Netw Open 2021 Sep 1;4(9):e2122837. [doi: 10.1001/jamanetworkopen.2021.22837] [Medline: 34477853]
2. Grandner MA, Seixas A, Shetty S, Shenoy S. Sleep duration and diabetes risk: population trends and potential mechanisms. Curr Diab Rep 2016 Nov;16(11):106. [doi: 10.1007/s11892-016-0805-8] [Medline: 27664039]
3. Yu B, Fu Y, Dong S, Reinhardt JD, Jia P, Yang S. Identifying potential action points for improving sleep and mental health among employees: a network analysis. Sleep Med 2024 Jan;113:76-83. [doi: 10.1016/j.sleep.2023.11.020]
4. Alhainen M, Härmä M, Pentti J, et al. Sleep duration and sleep difficulties as predictors of occupational injuries: a cohort study. Occup Environ Med 2022 Apr;79(4):224-232. [doi: 10.1136/oemed-2021-107516] [Medline: 34650000]
5. Tsuchida M, Monma T, Ozawa S, Kikuchi A, Takeda F. Relationships between lifestyle habits and presenteeism among Japanese employees. J Public Health (Berl) 2025 Jul;33(7):1597-1604. [doi: 10.1007/s10389-023-02136-4]
6. OECD. Gender data portal. 2021. URL: https://www.oecd.org/gender/data/OECD_1564_TUSupdatePortal.xlsx [accessed 2024-02-07]
7. Ikeda H, Kubo T, Sasaki T, et al. Cross-sectional Internet-based survey of Japanese permanent daytime workers' sleep and daily rest periods. J Occup Health 2018 May 25;60(3):229-235. [doi: 10.1539/joh.17-0165-OA] [Medline: 29311437]
8. Qiu D, Li Y, Li R, et al. Long working hours, work-related stressors and sleep disturbances among Chinese government employees: a large population-based follow-up study. Sleep Med 2022 Aug;96:79-86. [doi: 10.1016/j.sleep.2022.05.005] [Medline: 35613538]
9. Wong K, Chan AHS, Ngan SC. The effect of long working hours and overtime on occupational health: a meta-analysis of evidence from 1998 to 2018. Int J Environ Res Public Health 2019 Jun 13;16(12):13-19. [doi: 10.3390/ijerph16122102] [Medline: 31200573]
10. C001 - hours of work (industry) convention, 1919 (no 1). International Labour Organization. 1919. URL: https://www.ilo.org/dyn/normlex/en/f?p=NORMLEXPUB:12100:0::NO::P12100_ILO_CODE:C001 [accessed 2025-08-14]
11. Ministry of Health Labour and Welfare. Laws relating to labour standards in japan. 2021. URL: https://www.mhlw.go.jp/new-info/kobetu/roudou/gyousei/foreigner/technical_intern/dl/english_01.pdf
12. Jarczok MN, Jarczok M, Mauss D, et al. Autonomic nervous system activity and workplace stressors--a systematic review. Neurosci Biobehav Rev 2013 Sep;37(8):1810-1823. [doi: 10.1016/j.neubiorev.2013.07.004] [Medline: 23891906]
13. Jarczok MN, Jarczok M, Thayer JF. Work stress and autonomic nervous system activity BT -. In: Theorell T, Theorell T, editors. Handbook of Socioeconomic Determinants of Occupational Health: From Macro-Level to Micro-Level Evidence: Springer International Publishing; 2020:625-656. [doi: 10.1007/978-3-030-31438-5_27]

XSL•FO
RenderX

14. Schiller H, Lekander M, Rajaleid K, et al. The impact of reduced worktime on sleep and perceived stress—a group randomized intervention study using diary data. Scand J Work Environ Health 2017 Mar 1;43(2):109-116. [doi: 10.5271/sjweh.3610] [Medline: 27942734]
15. Imai T, Kuwahara K, Kabe I, et al. Impact of a reduction in overtime hours on sleep duration among Japanese employees. Environ Occup Health Pract 2020;2(1):n. [doi: 10.1539/eohp.2020-0012-CM]
16. Mittelmark MB, Bauer GF. Salutogenesis as a theory, as an orientation and as the sense of coherence BT. In: Mittelmark MB, Bauer GF, Vaandrager L, editors. The Handbook of Salutogenesis 2022. [doi: 10.1007/978-3-030-79515-3_3]
17. Antonovsky A. Health, Stress and Coping: Jossey-Bass; 1979.
18. Antonovsky A. Unraveling the Mystery of Health: How People Manage Stress and Stay Well: Jossey-Bass; 1987.
19. Haukkala A, Konttinen H, Lehto E, Uutela A, Kawachi I, Laatikainen T. Sense of coherence, depressive symptoms, cardiovascular diseases, and all-cause mortality. Psychosom Med 2013 May;75(4):429-435. [doi: 10.1097/PSY.0b013e31828c3fa4] [Medline: 23592751]
20. Piiroinen I, Tuomainen TP, Tolmunen T, et al. Sense of coherence and mortality: a systematic review and meta-analysis. Psychosom Med 2020;82(6):561-567. [doi: 10.1097/PSY.0000000000000812] [Medline: 32398414]
21. Ohta M, Higuchi Y, Yamato H, Kumashiro M, Sugimura H. Sense of coherence modifies the effect of overtime work on mental health. J Occup Health 2015;57(3):297-301. [doi: 10.1539/joh.14-0164-BR] [Medline: 25739337]
22. Schäfer SK, Becker N, King L, Horsch A, Michael T. The relationship between sense of coherence and post-traumatic stress: a meta-analysis. Eur J Psychotraumatol 2019;10(1):1562839. [doi: 10.1080/20008198.2018.1562839] [Medline: 30693079]
23. Morita Y, Ohta M, Inoue T, et al. Sense of coherence is significantly associated with both metabolic syndrome and lifestyle in Japanese computer software office workers. Int J Occup Med Environ Health 2014 Dec;27(6):967-979. [doi: 10.2478/s13382-014-0322-4] [Medline: 25556337]
24. Kósa K, Vincze S, Veres-Balajti I, Bába É. The pendulum swings both ways: evidence for u-shaped association between sleep duration and mental health outcomes. Int J Environ Res Public Health 2023 Apr 26;20(9):5650. [doi: 10.3390/ijerph20095650] [Medline: 37174169]
25. Jackson CL, Patel SR, Jackson WB II, Lutsey PL, Redline S. Agreement between self-reported and objectively measured sleep duration among White, Black, Hispanic, and Chinese adults in the United States: multi-ethnic study of atherosclerosis. Sleep 2018 Jun 1;41(6):zsy057. [doi: 10.1093/sleep/zsy057]
26. Kushida CA, Littner MR, Morgenthaler T, et al. Practice parameters for the indications for polysomnography and related procedures: an update for 2005. Sleep 2005 Apr;28(4):499-521. [doi: 10.1093/sleep/28.4.499] [Medline: 16171294]
27. Cole RJ, Kripke DF, Gruen W, Mullaney DJ, Gillin JC. Automatic sleep/wake identification from wrist activity. Sleep 1992 Oct;15(5):461-469. [doi: 10.1093/sleep/15.5.461] [Medline: 1455130]
28. Ancoli-Israel S, Cole R, Alessi C, Chambers M, Moorcroft W, Pollak CP. The role of actigraphy in the study of sleep and circadian rhythms. Sleep 2003 May 1;26(3):342-392. [doi: 10.1093/sleep/26.3.342] [Medline: 12749557]
29. Greenland S. Interactions in epidemiology: relevance, identification, and estimation. Epidemiology (Sunnyvale) 2009 Jan;20(1):14-17 [FREE Full text] [doi: 10.1097/EDE.0b013e318193e7b5] [Medline: 19234397]
30. VanderWeele TJ. The interaction continuum. Epidemiology (Sunnyvale) 2019 Sep;30(5):648-658 [FREE Full text] [doi: 10.1097/EDE.0000000000001054] [Medline: 31205287]
31. National health and nutrition survey. : Ministry of Health Labour and Welfare; 2024. URL: https://www.mhlw.go.jp/bunya/kenkou/kenkou_eiyou_chousa.html [accessed 2024-02-06]
32. Nishikido N, Kageyama T, Kobayashi T, Haratani T. Assessment of job stress using a brief questionnaire: its relations to depression among male workers of an information processing company. Japanese J Occup Ment Heal 2000;8(2):73-82 [FREE Full text]
33. Karasek R, Brisson C, Kawakami N, Houtman I, Bongers P, Amick B. The Job Content Questionnaire (JCQ): an instrument for internationally comparative assessments of psychosocial job characteristics. J Occup Health Psychol 1998;3(4):322-355. [doi: 10.1037/1076-8998.3.4.322]
34. Muroi K, Ishitsuka M, Hori D, et al. Relationships between childhood bullying/domestic violence experience and insomnia among employees in Japan. F1000Res 2023;12:103. [doi: 10.12688/f1000research.129340.1] [Medline: 37384304]
35. Hori D, Arai Y, Morita E, et al. Morning preference is associated with subjective happiness among Japanese female workers: a moderation analysis by sleep characteristics from the SLEPT study. Chronobiol Int 2022 May;39(5):690-703. [doi: 10.1080/07420528.2022.2028801] [Medline: 35078389]
36. Hori D, Sasahara S, Oi Y, et al. Relationships between insomnia, long working hours, and long commuting time among public school teachers in Japan: a nationwide cross-sectional diary study. Sleep Med 2020 Nov;75:62-72. [doi: 10.1016/j.sleep.2019.09.017] [Medline: 32853920]
37. Nakazaki K, Kitamura S, Motomura Y, et al. Validity of an algorithm for determining sleep/wake states using a new actigraph. J Physiol Anthropol 2014 Oct 4;33(1):31. [doi: 10.1186/1880-6805-33-31] [Medline: 25282356]
38. Eriksson M, Lindström B. Validity of Antonovsky's sense of coherence scale: a systematic review. J Epidemiol Community Health 2005 Jun;59(6):460-466. [doi: 10.1136/jech.2003.018085] [Medline: 15911640]

XSL•FO
RenderX

39. Togari T, Yamazaki Y, Nakayama K, Yokoyama Y, Yonekura Y, Takeuchi T. Nationally representative score of the Japanese language version of the 13-item 7-point sense of coherence scale. Nihon Koshu Eisei Zasshi 2015;62(5):232-237. [doi: 10.11236/jph.62.5_232] [Medline: 26118706]
40. Society for theory and research on salutogenesis. 2024. URL: https://stars-society.org/ [accessed 2024-02-06]
41. Aiken LS, West SG. Multiple Regression: Testing and Interpreting Interactions: Sage Publications, Inc; 1991.
42. Cohen J, Cohen P, West SG, Aiken LS. Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences, 3rd edition: Lawrence Erlbaum Associates Publishers; 2003.
43. Mirisola A, pequod SL. Moderated Regression Package. 2022. URL: https://cran.r-project.org/web/packages/pequod/pequod.pdf [accessed 2025-08-14]
44. Takahashi M, Tsutsumi A, Kurioka S, et al. Occupational and socioeconomic differences in actigraphically measured sleep. J Sleep Res 2014 Aug;23(4):458-462. [doi: 10.1111/jsr.12136] [Medline: 24628714]
45. Taber KS. The use of Cronbach's alpha when developing and reporting research instruments in science education. Res Sci Educ 2018 Dec;48(6):1273-1296. [doi: 10.1007/s11165-016-9602-2]
46. Moraes W, Poyares D, Zalcman I, et al. Association between body mass index and sleep duration assessed by objective methods in a representative sample of the adult population. Sleep Med 2013 Apr;14(4):312-318. [doi: 10.1016/j.sleep.2012.11.010] [Medline: 23391395]
47. Boersma GJ, Mijnster T, Vantyghem P, Kerkhof GA, Lancel M. Shift work is associated with extensively disordered sleep, especially when working nights. Front Psychiatry 2023;14:1233640. [doi: 10.3389/fpsyt.2023.1233640] [Medline: 38161719]
48. Cheng WJ, Cheng Y. Night shift and rotating shift in association with sleep problems, burnout and minor mental disorder in male and female employees. Occup Environ Med 2017 Jul;74(7):483-488. [doi: 10.1136/oemed-2016-103898] [Medline: 27810939]
49. Jonasdottir SS, Minor K, Lehmann S. Gender differences in nighttime sleep patterns and variability across the adult lifespan: a global-scale wearables study. Sleep 2021 Feb 12;44(2):zsaa169. [doi: 10.1093/sleep/zsaa169] [Medline: 32886772]
50. Li L, Nakamura T, Hayano J, Yamamoto Y. Age and gender differences in objective sleep properties using large-scale body acceleration data in a Japanese population. Sci Rep 2021;11(1):9970. [doi: 10.1038/s41598-021-89341-x]
51. Ek S. Gender differences in health information behaviour: a Finnish population-based survey. Health Promot Int 2015 Sep;30(3):736-745. [doi: 10.1093/heapro/dat063] [Medline: 23985248]
52. Otsuka T, Konta T, Sho R, et al. Factors associated with health intentions and behaviour among health checkup participants in Japan. Sci Rep 2021 Oct 5;11(1):19761. [doi: 10.1038/s41598-021-99303-y] [Medline: 34611263]
53. Kalmbach DA, Anderson JR, Drake CL. The impact of stress on sleep: pathogenic sleep reactivity as a vulnerability to insomnia and circadian disorders. J Sleep Res 2018 Dec;27(6). [doi: 10.1111/jsr.12710]
54. Otsuka Y, Itani O, Matsumoto Y, Kaneita Y. Associations between coping strategies and insomnia: a longitudinal study of Japanese workers. Sleep 2022 Feb 14;45(2):zsab244. [doi: 10.1093/sleep/zsab244] [Medline: 34585730]
55. Tei S, Fujino J. Sense of coherence, spontaneous brain activity, and burnout severity. Eur Psychiatr 2021 Apr;64(S1):S737-S737. [doi: 10.1192/j.eurpsy.2021.1953]
56. Kanda K, Tei S, Takahashi H, Fujino J. Neural basis underlying the sense of coherence in medical professionals revealed by the fractional amplitude of low-frequency fluctuations. PLoS ONE 2023;18(6):e0288042. [doi: 10.1371/journal.pone.0288042] [Medline: 37390054]
57. Arnsten AFT. Stress signalling pathways that impair prefrontal cortex structure and function. Nat Rev Neurosci 2009 Jun;10(6):410-422. [doi: 10.1038/nrn2648] [Medline: 19455173]
58. Cerqueira JJ, Almeida OFX, Sousa N. The stressed prefrontal cortex. Left? Right!. Brain Behav Immun 2008 Jul;22(5):630-638. [doi: 10.1016/j.bbi.2008.01.005] [Medline: 18281193]
59. Langeland E, Vaandrager L, Nilsen ABV, Schraner M, Meier Magistretti C. Effectiveness of interventions to enhance the sense of coherence in the life course BT. In: Mittelmark MB, Bauer GF, Vaandrager L, editors. The Handbook of Salutogenesis: Springer International Publishing; 2022. [doi: 10.1007/978-3-030-79515-3_20] [Medline: 36122015]
60. Vaandrager L, Koelen M. Salutogenesis in the workplace: building general resistance resources and sense of coherence BT - salutogenic organizations and change: the concepts behind organizational health intervention research. In: Bauer GF, Jenny GJ, editors. Salutogenic Organizations and Change: Springer Netherlands; 2013:77-89. [doi: 10.1007/978-94-007-6470-5_5]
61. Grandner MA, Patel NP, Gehrman PR, et al. Who gets the best sleep? Ethnic and socioeconomic factors related to sleep complaints. Sleep Med 2010 May;11(5):470-478. [doi: 10.1016/j.sleep.2009.10.006] [Medline: 20388566]
62. Matsumoto Y, Uchimura N, Ishitake T. The relationship between marital status and multifactorial sleep in Japanese day workers. Sleep Biol Rhythms 2022 Apr;20(2):211-217. [doi: 10.1007/s41105-021-00357-2] [Medline: 38469263]
63. de Castro JM. The influence of heredity on self-reported sleep patterns in free-living humans. Physiol Behav 2002 Aug;76(4-5):479-486. [doi: 10.1016/s0031-9384(02)00699-6] [Medline: 12126983]
64. Watson NF, Buchwald D, Vitiello MV, Noonan C, Goldberg J. A twin study of sleep duration and body mass index. J Clin Sleep Med 2010 Feb 15;06(01):11-17. [doi: 10.5664/jcsm.27704]
65. Takeshima M, Echizenya M, Inomata Y, Shimizu K, Shimizu T. Comparison of sleep estimation using wrist actigraphy and waist actigraphy in healthy young adults. Sleep Biol Rhythms 2014 Jan;12(1):62-68. [doi: 10.1111/sbr.12048]

XSL•FO
RenderX

66. Chiang TW, Chen SY, Pan YC, Lin YH. Automatic work-hours recorder for medical staff (staff hours): mobile app development. JMIR Mhealth Uhealth 2020 Feb 25;8(2):e16063. [doi: 10.2196/16063] [Medline: 32130165]
67. Why do people sleep? what is optimal sleep? approaching the "mysteries of sleep" with the latest science. READYFOR. URL: https://readyfor.jp/projects/wpi-iiis [accessed 2025-08-14]

Abbreviations

BSJS: Brief Scales for Job Stress
DLPFC: dorsolateral prefrontal cortex
GRR: generalized resistance resource
PSG: polysomnography
SLEPT: Sleep Epidemiology Project at University of Tsukuba
SOC: sense of coherence
SOC-13: 13-item Sense of Coherence scale
TST: total sleep time

Edited by A Mavragani; submitted 19.03.25; peer-reviewed by K Otani, T Sasaki; revised version received 18.06.25; accepted 27.06.25; published 03.09.25.

Please cite as:
Muroi K, Morita E, Matsumoto S, Ishihara A, Hasegawa S, Ishitsuka M, Hori D, Doki S, Takahashi T, Sasahara SI, Kanbayashi T, Yanagisawa M, Satoh M, Matsuzaki I
Interaction of Long Working Hours and Sense of Coherence on Objective Total Sleep Time: Cross-Sectional Study From the SLEPT Study
JMIR Form Res 2025;9:e73784
URL: https://formative.jmir.org/2025/1/e73784
doi:10.2196/73784

©Kei Muroi, Emi Morita, Sumire Matsumoto, Asuka Ishihara, Sumi Hasegawa, Mami Ishitsuka, Daisuke Hori, Shorato Doki, Tsukasa Takahashi, Shin-ichiro Sasahara, Takashi Kanbayashi, Masashi Yanagisawa, Makoto Satoh, Ichiyo Matsuzaki. Originally published in JMIR Formative Research (https://formative.jmir.org), 3.9.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Environmental Triggers of Specific Subtypes of Agitation in People With Dementia: Observational Study

Hannah Davidoff[1,2], MSc; Alex Van Kraaij[3], MSc; Erika Lutin[2], PhD; Laura Van den Bulcke[4,5], MD; Mathieu Vandenbulcke[4,5], MD, PhD; Nick Van Helleputte[2], PhD; Maarten De Vos[1], PhD; Chris Van Hoof[2,3], PhD; Maarten Van Den Bossche[4,5], MD, PhD

[1]ESAT, Department of Electrical Engineering, KU Leuven, Leuven, Belgium
[2]imec, Heverlee, Belgium
[3]OnePlanet Research Center, Wageningen, The Netherlands
[4]Geriatric Psychiatry, University Psychiatric Center KU Leuven, Herestraat 49, Leuven, Belgium
[5]Neuropsychiatry, Department of Neurosciences, Leuven Brain Institute, KU Leuven, Leuven, Belgium

Corresponding Author:
Maarten Van Den Bossche, MD, PhD
Geriatric Psychiatry, University Psychiatric Center KU Leuven, Herestraat 49, Leuven, Belgium

Abstract

Background: Among the most critical behavioral and psychological symptoms of dementia, agitation can lead to decreased quality of life of people with dementia and their caregivers. Monitoring triggers of agitation and its subtypes could enable early detection or prediction of agitated moments, which could be used to guide preventive or mitigating interventions. However, at this point in time, limited research exists on quantifying environmental triggers of agitation or its subtypes.

Objective: In this paper, we aim to quantify the relationships between specific environmental factors and agitation as well as specific agitation subtypes, such as motor and verbal agitation.

Methods: Using a cross-sectional repeated measures design, 37 people with dementia, admitted to a specialized neuropsychiatric ward for patients with dementia and severe behavioral and psychological problems, were each included for 1 week. During this period, the Pittsburgh Agitation Scale was filled in by the nurses on the ward following an experience sampling methodology to assess a patient's agitation level on a momentary basis. Continuous environmental data (light, sound, and temperature) were collected from fixed sensors mounted on the ward. Generalized linear mixed models were used to quantify relationships between environmental variables and outcome variables (agitation, motor agitation, and verbal agitation). These models accounted for the hierarchical nature of our dataset as well as confounding factors, such as time of day and the room-level location of the patient. The time window for analysis was selected through a comparison of β coefficient estimates across various window lengths. Models were built up sequentially, per outcome variable, using selected features per environmental modality.

Results: We found that different environmental factors captured in the window of 33 to 12 minutes before the agitation moment were most informative for different subtypes of agitation: mean light level (β=−0.61, 95% CI −1.12 to −0.10; P=.02) for motor agitation and SD of sound level (β=0.68, 95% CI 0.34 - 1.02; P<.001) for verbal agitation. Contextual factors such as time of day (β range=0.51 - 0.94; P<.05 to <.001) and room-level location (β range=0.85 - 1.08; P<.01 to <.001) were also significant predictors of agitation.

Conclusions: Integrating the key differences between predictors of verbal and motor agitation, respectively, the higher SD in sound level and the lower mean light level, in a model predicting the occurrence of subtype-specific agitation, could substantially improve model performance. Overall, these findings can aid in the development of predictive models for agitation based on environmental data and enable subsequent just-in-time interventions, improving the quality of life for both patients and caregivers.

(JMIR Form Res 2025;9:e60274) doi:10.2196/60274

KEYWORDS

dementia; agitation; multimodal sensing; digital health; medical informatics; observational study; quality of life; environmental factors; neuropsychiatric; mental illness; Alzheimer disease; experience sampling; linear mixed models; health informatics; clinical informatics

XSL•FO
RenderX

Introduction

Alongside the cognitive symptoms of dementia, behavioral and psychological symptoms (BPSDs) significantly impact the quality of life of patients and their caregivers [1-4]. Agitation, one of the most critical BPSDs, has been associated with increased caregiver burden [5], earlier institutionalization [6], and more rapid disease progression [7]. Behaviors that fall under the International Psychogeriatric Association's definition of agitation include excessive motor activity, verbal aggression, or physical aggression [8]. The International Psychogeriatric Association further defines that for these behaviors to be considered agitation, they must be consistently or recurringly present for at least 2 weeks, have a large impact on interpersonal, social relationships or activities of daily living, and not be completely explainable by a comorbid condition. Several theoretical frameworks exist to explain the etiology of agitation in people with dementia. The progressively lowered stress threshold (PLST) model posits that behavioral symptoms, like agitation, arise when internal or external demands of the environment exceed a person's stress threshold [9]. Additionally, this model hypothesizes that typical threshold levels are higher in the morning and lower throughout the day [10]. The PLST model is supported by Lawton and Simon's [11] earlier environmental docility hypothesis, which states that people become more susceptible to environmental factors as their "level of competence" decreases.

Parallel to these frameworks, the Unmet Needs model [12] and the need-driven dementia-compromised behavior (NDB) model [13,14] focus on the needs of people with dementia and how they drive behaviors such as agitation. Both models stress the inability of a person with dementia to adequately communicate their needs. Unmet needs range from "loneliness or need for social interaction" and "boredom or sensory deprivation" to discomfort and pain [12]. One of the key factors in the Unmet Needs model is the need for adequate levels of stimulation [12] to mitigate boredom or social isolation in particular. Cohen-Mansfield et al [12] also state that different unmet needs relate to different behaviors. Specifically, they hypothesize that discomfort and pain would be associated with verbal agitation, while boredom or sensory deprivation would be associated with physically nonaggressive agitation [12]. Similarly, the NDB model postulates that different behaviors represent different unmet needs. For example, wandering could reflect the inability of a person with dementia to find their way or their need to avoid overstimulating or overcrowded locations. Vocalizations, on the other hand, could be an attempt to communicate a need. Algase et al [13] state that several background factors can affect NDBs: neurological factors, cognitive factors, and health status. In addition to these background factors, proximal factors such as emotions, physiological needs, functional status, and the social and physical environment can also impact NDBs [13].

These theoretical models form the inspiration for our research into what aspects of the environment could be triggers for the onset of agitation and how these triggers could differ by agitation type. Existing research has investigated univariate relationships between aspects of the environment and agitation. For example, Tartarini et al [15] have researched the relationship between temperature and agitation, where they found that the variation of temperature from "ideal" (22.6 °C) increased the frequency of agitation. Figueiro et al [16] have found that their bright light intervention decreased the occurrence of agitation. Janus et al [17] and Joosse [18] have looked at the impact of sound on agitation, determining that different aspects of sound (such as accumulation or variation) can increase agitation. However, these typically focus on an overall effect over a set period and not the effect of an environmental factor on a momentary level. In addition to unimodal research, there have been several studies looking into multimodal relationships between several environmental factors and agitation [19-22]. These studies point toward the relevance of analyzing environmental factors closer in time to the onset of agitation. A case study (n=1) by Au-Yeung et al [20] looked at environmental variables (ambient light, sound, humidity, etc) and other data from passive sensors using time windows of 8 hours aligned with the shifts of the nurses. There, the humidity and sound parameters were associated with agitation, although the humidity parameter was influenced by showers during visits from the included patients' wives (which led to decreased levels of agitation). Homdee [21] used a sequence of nine 6-minute windows of environmental time series data forming a 54-minute input to a predictive model. HekmatiAthar et al [22] used windows of 30 minutes of environmental time series data to predict subsequent agitation. Both prediction modeling studies focused primarily on which machine learning framework performed best for this predictive problem and did not go into detail regarding feature importance. These analyses and modeling approaches also did not further refine models by agitation type, focusing only on agitation as a whole. Additionally, despite the clinical knowledge of increased agitation in the late afternoon or early evening [23-25], "sundowning," these models do not take into account contextual factors like time of day. Here, we use an exploratory statistical approach to analyze important environmental factors not only on the presence or absence of agitation but also specifically on the presence or absence of motor and verbal agitation. This paper expands on current literature both through this extra level of detail on agitation and its subtypes as well as through the multimodal approach, providing a more holistic view on a patient's environment.

The focus of this paper is to further understand which factors in the environment, measured by our study protocol [26], are associated with which type of agitation. With this, we aim to steer feature selection for agitation prediction models, eventually enabling just-in-time interventions as well as supporting the development of nonpharmacological alternatives (eg, environmental adjustment) to reduce agitation.

Methods

Patient Demographics and Study Design

The patients included in the study (N=37) were recruited from the population of a neuropsychiatric hospital ward specialized in high intensive care for people with dementia and accompanying severe BPSD with typically 4- to 8-week stays. The study was observational; no specific treatment interventions

XSL•FO
RenderX

were introduced for research purposes, and routine care of patients continued throughout the study. The minimum inclusion criteria for the study were a diagnosis of dementia and the presence of agitation. Patients recruited from June 2021 to August 2023 formed a convenience sample. All patients admitted who met the minimum inclusion criteria or their representative were asked to participate. The clinical demographics of the included patients can be found in Table S1 in Multimedia Appendix 1. The ward unit layout consists of centralized communal living spaces with patient rooms and therapy rooms connected via hallways. More details can be found as additional information in Davidoff et al [26].

The study design from which the data used in these analyses originated was a cross-sectional repeated measures design, where patients were included for 1 week each. Sensor data, from both fixed and wearable sensors, were collected continuously throughout the week, and agitation labels were acquired by the nurses on the ward through the use of experience sampling [27] with a study-specific app, which prompted the nurse 9 times a day, semirandomly distributed. When prompted, nurses were asked to visit the patient and report the agitation level from that observation. On top of the 9 prompts, nurses had the option to self-initiate an ecological momentary assessment (EMA) to report an occurring agitation event and fill out the Pittsburgh Agitation Scale (PAS) after calming down the patient. More details related to the study protocol and rationale can be found in Davidoff et al [26].

Ethical Considerations

The data collection protocol was approved by the ethics committee research of the University Hospitals Leuven (ID S62882). Since patients were experiencing dementia, written informed consent was mainly obtained through the patient's representative. However, information on the study was explained as much as possible to the patient and on a level corresponding to the patient's current cognitive capacities. Patient data were treated confidentially and pseudonymized. Participants did not receive financial compensation. The decision whether or not to participate in the study did not have any influence on the care or treatment patients received during their admission.

Description of Data and Data Processing

The ground truth of the patient's agitation state used in this study was the binarized score derived from the PAS [28], where the presence of each of the agitation subtypes (motor, verbal, resistance to care, and aggression) is scored on a scale of 0 to 4, and the presence or absence of each subtype is subsequently translated into 1 and 0, respectively. For this analysis, we focused only on verbal and motor agitation subtypes and not on the resistance to care and aggression subtypes, as these were rarely recorded as present in our survey dataset. This scale was a part of the survey notified for completion up to 9 times per day by the study-specific app.

Given the focus on the prediction and not the detection of agitation in this paper, a buffer (the spacing between the survey timestamp and the end of the window of environmental data analyzed) of 12 minutes was used. This was done to avoid detection of verbal agitation, one of the 2 most observed subtypes (as defined by the PAS) of agitation in our dataset, as verbal agitation could influence the measurement of sound on the ward. This buffer ensures that the input features are used to predict the subsequent onset of agitation and not detect agitation. Although we cannot completely exclude that verbal agitation had been ongoing beyond the buffer length, the buffer of 12 minutes is in line with buffers used in the literature [21]. Moreover, we have visually analyzed the effect of changing the buffer length and found that, as expected, the highest correlation between sound and verbal agitation is at the time of the EMA.

The length of the time windows explored in this paper ranged from 3 to 30 minutes at intervals of 3 minutes. This range is in line with time windows used in studies exploring the relationship between environmental parameters and agitation [21,22].

Using the survey timestamps, the environmental data were extracted for each of the 3 modalities focused on in this paper: sound (leqDBA), light (lux), and temperature (degrees Celsius). This was done by first determining the room the patient was in during the relevant time window. When an environmental sensor is present in that room, the time series segment includes the data from that sensor. If there was no environmental sensor, the data are left as not applicable, and the location is recorded as "off map" (eg, the patient location or environmental data for their location is unknown). Four rooms on the ward have 2 sensor enclosures present because of the above-average room area. When patient presence in any of those 4 rooms is detected, the respective sensors are averaged. If the patient visits multiple rooms in the relevant time window, the environmental data from each of the relevant sensors are appended. A moving average over 5 seconds is applied to each of the environmental data modality time series to smooth the effect of room transitions in the data. A visual representation of this data processing flow can be found in Figure 1.

XSL•FO
RenderX

Figure 1. Data processing flow. (Not to scale) In this figure, 3 different standardized time series (one for each modality: temperature, sound, and light) are shown. The color of each time series denotes the location the data were extracted from. The red vertical line with a marker triangle indicates the timestamp of the survey. The spacing between this vertical line and the start of the shaded area is called the "buffer" and is used to avoid detection instead of the prediction of verbal agitation. It is set to 12 minutes. The small, shaded area is one example window. All the data for that window are used for feature calculation. For all other windows, the right side (end point) of the shaded area stays the same (−12 minutes from the survey timestamp), and the length of the time window is expanded to include more minutes of data before the survey timestamp. The final time window encompasses all the shaded areas. dBA: A-weighted decibel.

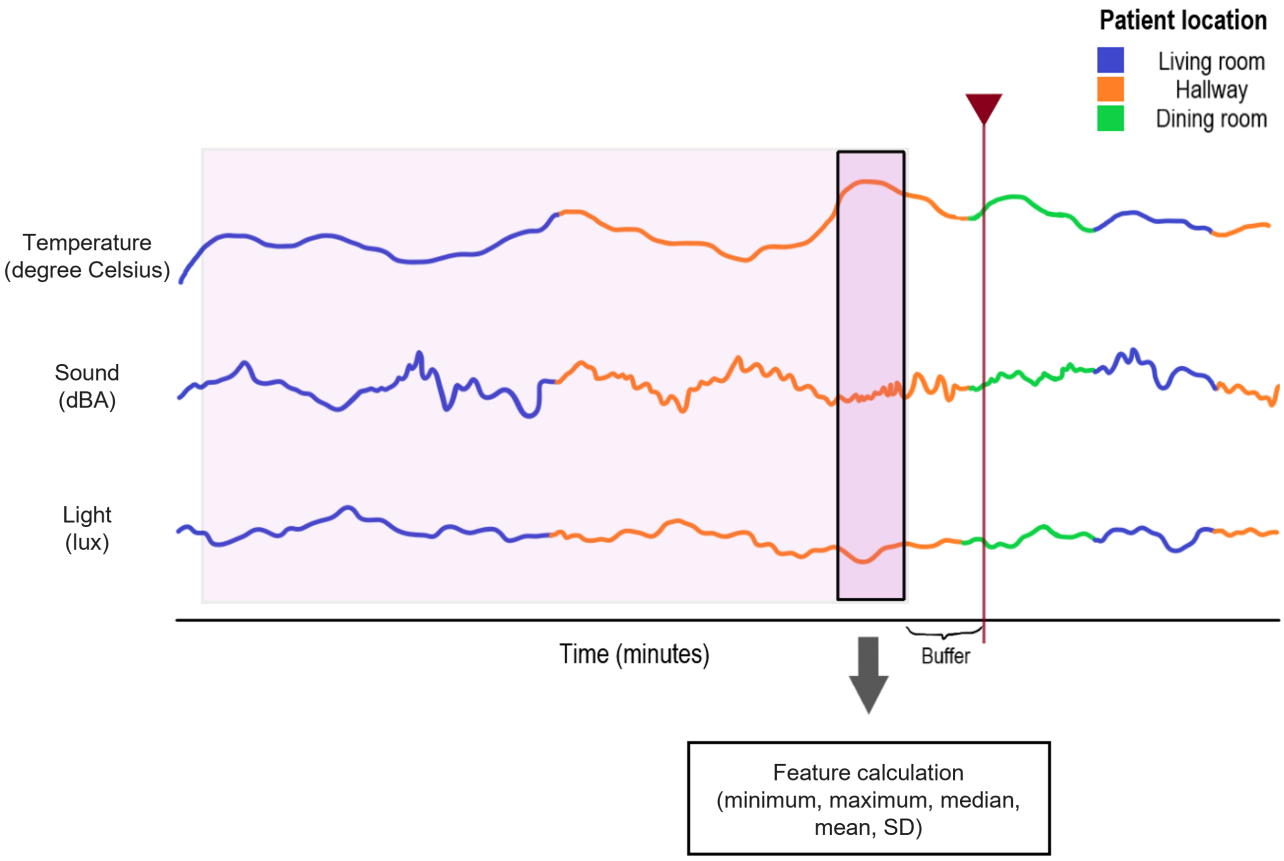

Per the relevant time window, the room in which the most time was spent was recorded along with the exact amount of time spent in that room. Additionally, the number of data points present in the respective modality table was recorded per window. From all data points present per window, the following descriptive features were calculated: minimum, maximum, mean, median, and SD. Windows where any of the following quality conditions were not met were discarded from subsequent analysis: (1) the percentage of data points for which the patient location was unknown must be less than 50%, (2) there must be at least 50% of the expected number of data points present for that window length, (3) the majority room for that window must be a location with environmental sensors (eg, majority room is not "off map"), and (4) the survey was recorded within the time frame of 8 AM to 8 PM, as outlined in the protocol.

The first 3 criteria can be influenced by times when (1) the patient was not wearing the tracking wearable or (2) the patient is tracked to a location that does not have a direct environmental sensor present.

The features calculated per modality were standardized within the majority room, meaning that feature values are compared to all feature values for that specific location on the ward. A feature value of 1 would mean that the value of the feature is 1 SD higher than the mean value of that feature in that room. This makes the interpretation of the environmental sensor data generalizable across rooms and wards.

Exploratory Modeling for Feature Selection

Given the hierarchical nature of this dataset (repeated measures within patients) and a binarized outcome variable, generalized linear mixed models ("glmer" models in the *lme4* package in R; R Foundation for Statistical Computing) were used for modeling with the family set to "binomial." These types of models allow for random effects, taking into account group-level (in this case, patient-level) differences. The models developed in this step served as an initial exploration of informative features.

To explore the relationship between environmental variables and the different types of agitation, 3 different models (1 per outcome variable) were built per explanatory feature. The models follow the structure below.

Outcome variable[1]~explanatory variable[2]+time_group[3]+(1 | p_id)[4]

XSL•FO
RenderX

The outcome variable has 3 options: all agitation, motor agitation, and verbal agitation.

The explanatory variable has 15 values for iteration: the 5 standardized features (minimum, maximum, mean, median, and SD) for each of the 3 modalities (light, sound, and temperature).

The "time_group" stays constant throughout all models. This is the time of day split into 4 groups by the hour in which the survey was completed: (1) 8 - 12 (exclusive), (2) 12 - 16 (exclusive), (3) 16 - 20 (exclusive), and (4) 20 - 8 (exclusive).

p_id is the random intercept per patient. It is included to account for the nested structure of this dataset.

Per model built, we extracted the β coefficients quantifying the relationship between the explanatory variable [2] and the outcome variable [1]. These β coefficient estimates, analogous to "effect size," can be interpreted as the log-odds change in the outcome variable with a 1 unit increase in the explanatory variable. This can be exponentiated to get the odds ratio. Since standardized features were used, a 1-unit increase is a 1 SD increase in the explanatory variable.

In addition to the feature-level analysis, different time window lengths ranging from 3 to 30 minutes were explored. An optimal window length was set for future modeling steps based on the window with the maximum amount of information (magnitude and significance of β coefficient estimates) across all features and modalities. The most informative features, in this window, were then selected per modality and outcome variable by the largest (absolute) β coefficient. These features were used as the environmental variable input in subsequent modeling steps.

Modeling Agitation Types

Building on the selection of informative features in the initial exploratory modeling step, these models aimed to further quantify the relationships between environmental variables and agitation. This was done through a forward model selection process to build an optimal model per agitation type. Differences in informative features seen per outcome variable (ie, agitation type), both in magnitude and sign, were the reason for building an optimal model per outcome and comparing results. In this step, the selected features for light and sound were used as the basis for outcome-specific models. In addition to the time of day included in the exploratory modeling step, these models also included a majority room group to add additional context. The majority room group is a contextual variable that indicates the room where the majority of the time window was spent (see the Description of Data and Data Processing section) was grouped into 1 of 5 groups: "living room," "dining room," "hallway," "patient room," and "other." This variable was included in the model to account for any confounding effects the patient's location may have on the outcome. The temperature was dropped from the modeling analysis due to a lack of significant features for any outcome variable.

For each of the outcome variables, the model was built starting from a model with only an intercept by the patient (model 0). The full model build-up flow along with each model structure can be found in Figure 2. First, the contextual variables are added one by one (models 1 and 2). Subsequently, both sound and light are added separately and then combined (models 3 and 4 and model 5). Finally, the interaction between either time or majority room (maj_room) was tested for both the sound-specific and light-specific models (sound: models 6 and 7 and light: models 8 and 9). The added value of including each variable versus the added complexity was tested with an ANOVA test using the *anova* function in R. The variable was kept in the model if the ANOVA test showed a significant improvement or if there was sufficient domain knowledge to support the inclusion of confounding variables (time or majority room). No Bonferroni correction or Benjamini-Hochberg correction has been used, as this analysis is purely explorative and not driven by a main hypothesis.

Figure 2. GLMER model build-up flow, which was used for each of the agitation types as outcome variables. The model improvement was checked at each step with ANOVA in R. This flow was used to find the most contributing environmental features.

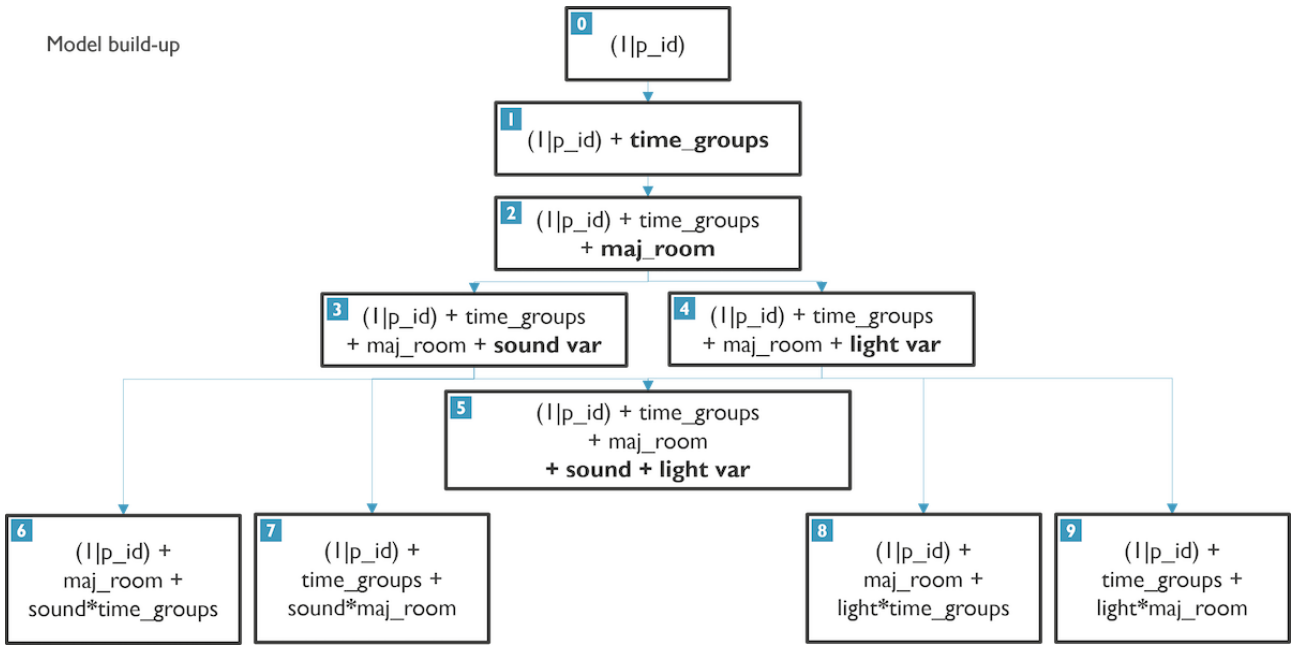

XSL•FO
RenderX

The coefficients were extracted for all model components along with model parameters for each of the models shown in Figure 2.

Results

Data Availability, Demographics, and Descriptive Statistics

Data on clinical demographics of the included patients (n=29) can be found in Table S1 of Multimedia Appendix 1 along with reasons and counts of patients not included in the analyses. Further details on the number of surveys present after each filtering step can be found in Table S2 in Multimedia Appendix 1. The filtering step that removes most surveys from the analysis is the “off map” filter.

Of the 694 surveys used in the mixed models, 503 are nonagitated. Of the remaining 191 recorded agitated surveys, 5 contain agitation that was neither motor nor verbal, 34 had only verbal agitation, 86 had only motor agitation, and 66 had coexisting motor and verbal agitation. The co-occurrence of motor and verbal agitation in this dataset is important in the subsequent interpretation of the results, specifically for verbal agitation where there are fewer “pure” verbal agitation surveys (n=34) than those co-occurring with motor agitation (n=66). For motor agitation, interpretation is still influenced by this co-occurrence but to a lesser extent due to the larger number of “pure” motor agitation surveys (n=86). Figure 3 shows the distribution of surveys and occurrences of agitation across the majority room group. The differing number of surveys recorded as well as the differing proportion of agitated surveys per room seen in Figure 3 aid in the interpretation of the room-level impact on subsequent model performance. A more detailed figure showing the proportion of agitation per room, split by type, can be found in Figure S1 in Multimedia Appendix 1.

XSL•FO
RenderX

Figure 3. Occurrence and counts of surveys (dark gray), split for with annotated agitation (light gray) and without agitation (yellow) per room grouping.

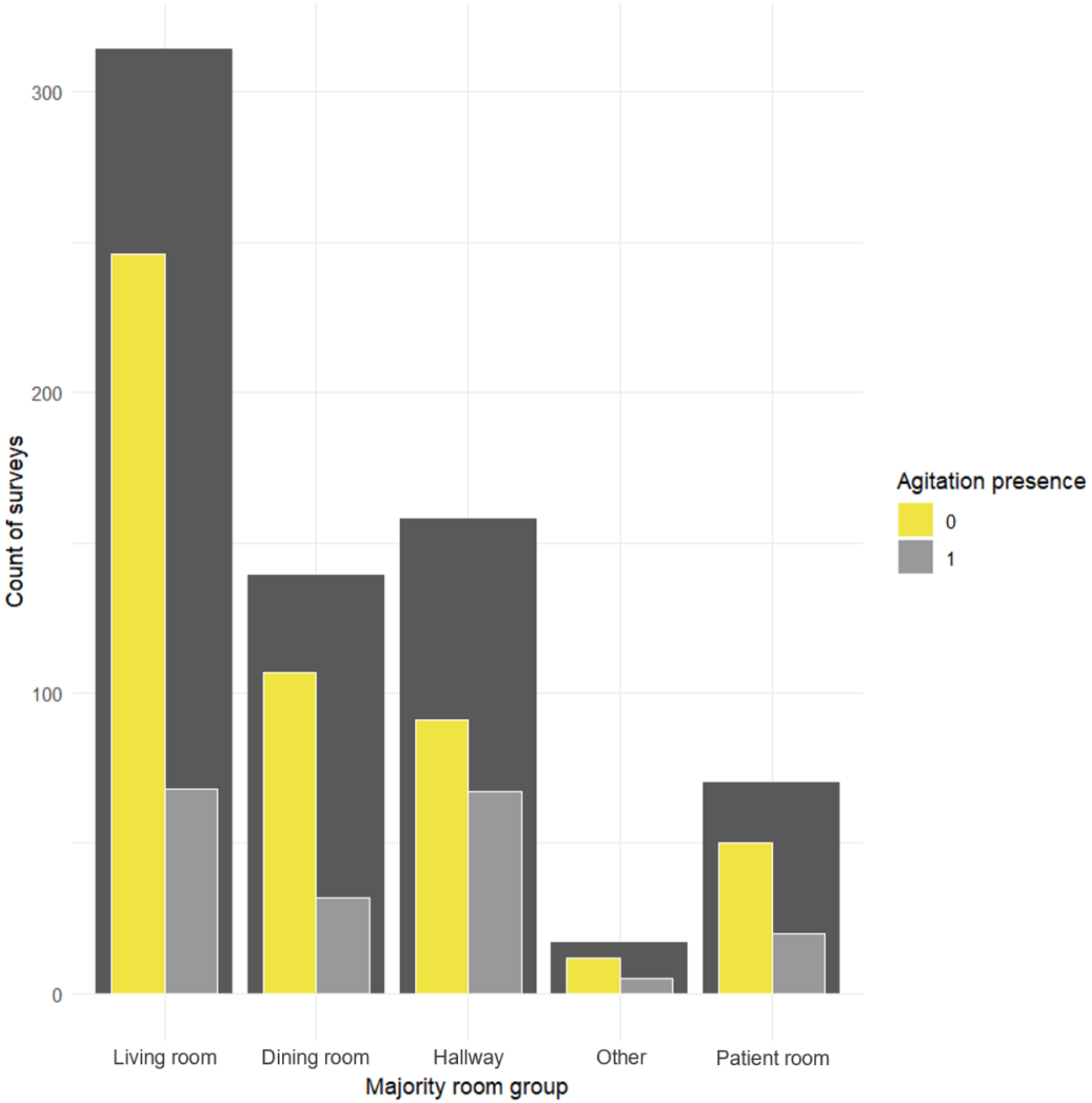

Exploring Optimal Time Window Length and Feature Selection

Time Window

The 21-minute window from 33 to 12 minutes prior to an EMA contained the majority of the maximum β coefficients, indicating that this window contained the most information for the prediction of agitation across modalities. For more details on the exploration of time window length, please see Figures S2 and S3 in Multimedia Appendix 1.

Feature Selection

For the selected 21-minute window, the β coefficients for each of the 5 features (minimum, maximum, median, mean, and SD) from each modality and for each outcome variable are summarized in Figure 4. To make agitation type–specific models, we extracted the most informative feature per modality for each outcome variable (agitation, motor agitation, or verbal agitation). The selected features are indicated with red triangles in Figure 4. Across each outcome variable, the SD of sound was selected for being the largest significant β coefficient (agitation: β=0.452; $P<.001$, motor: β=0.274; $P=.04$, and verbal: β=0.653; $P<.001$). The mean light level was only selected and significant for motor agitation (β=−0.449; $P=.04$). Other light parameters selected for having the largest β coefficients were SD (β=−0.196) and minimum (β=0.239) for agitation and verbal agitation, respectively. There were no features significantly associated with any of the outcome variables from the temperature modality. Median, mean, and minimum temperature had the largest β coefficients for agitation, motor agitation, and verbal agitation, respectively. Temperature was dropped from all subsequent modeling because of the lack of significant

XSL•FO
RenderX

relationships between predictors and outcome variables at the preselected buffer and window lengths.

Figure 4. β Coefficients for all features of each environmental modality (light, sound, and temperature) per outcome variable (overall agitation, motor agitation, and verbal agitation). The data used as input for the features are derived from the period 12 to 33 minutes before the ecological momentary assessment. The features that were selected for further model build-up per modality and outcome variable are marked with a red triangle. The significance level is indicated with the size and filling of the dots.

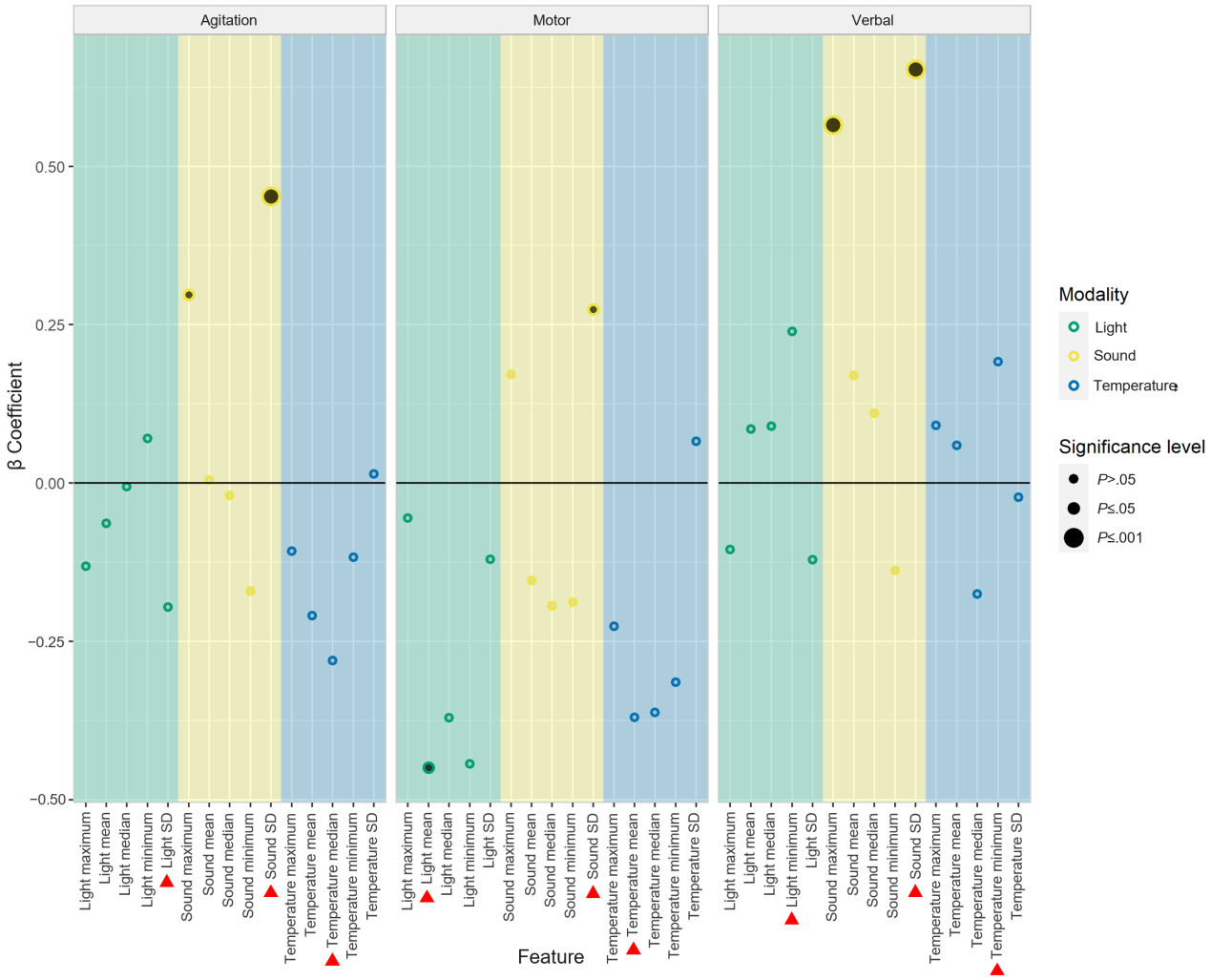

Outcome-Specific Models

Model Optimization

The selected light and sound features were used to build up agitation type–specific models, following the build-up flow shown in Figure 2. For each variable or interaction added to the model, an ANOVA test was conducted to determine if the information added resulted in a significant improvement in the model fit. The results from these model fit tests can be found in Table 1.

XSL•FO
RenderX

Table . ANOVA model fit testing[a].

Model[b]	Variable added	Agitation, *P* value	Motor, *P* value	Verbal, *P* value
0	Patient intercept only	N/A[c]	N/A	N/A
1	+ time_groups	.001	.008	.12
2	+ maj_room	.005	<.001	.02
3	+ sound parameter	.001	.08	<.001
4	+ light parameter (to 2)	.03	.004	.16
5	+ both light and sound (to 2)	<.001	.003	<.001
5	+ both light and sound (to 3)	.06	.003	.23
5	+ both light and sound (to 4)	.002	.06	<.001
6	+ sound*time	.26	.20	.16
7	+ sound*room	.74	.34	.93
8	+ light*time	.24	.47	.19
9	+ light*room	.24	.94	.17

[a]The *P* values here indicate the significance of the ANOVA test, which examines the added value of including each variable versus the added complexity of the addition. An insignificant result indicates that the added complexity of a model with an extra parameter is not worth the amount of extra variance explained by adding that parameter to the model.

[b]Model numbers correspond to the numbering in Figure 2.

[c]N/A: not applicable.

Each of the numbers of the models listed in column 1 of Table 1 corresponds to the model structures found in Figure 2. Including the “time_groups” variable in model 1 resulted in a significantly improved model fit for both agitation and motor agitation but not for verbal agitation. This indicates that the “time_groups” variable contributes important information for the prediction of agitation and motor agitation. This result, in combination with domain knowledge supporting the inclusion of time as a factor in agitation occurrence (eg, the PLST framework), led to the inclusion of “time_groups” in all subsequent models. In addition to time, including a second contextual variable “majority_room” also resulted in improved fits (Table 1). The inclusion of the selected environmental parameters had varying effects on model fits depending on the type of agitation used as the outcome variable. The addition of light led to significant improvements in models for agitation and its subtype motor agitation but not for verbal agitation. Conversely, the addition of sound led to significant improvements for agitation and its subtype verbal agitation but not for motor agitation. As seen in Table 1, including environmental parameters led to significant improvements in model fits compared to models where they were not included. However, including both environmental parameters in a single model was only informative when the most informative modality for that subtype was not yet included in the model. This is seen in the results when comparing model 5 to models 2 - 4 per outcome variable; adding light when light was not yet included in a model for motor agitation resulted in a significant improvement in model fit. A similar trend was followed for the addition of sound for agitation or verbal agitation.

Model Parameter Estimates

The coefficient estimates for each of the predictor variables included in models 0 - 5 can be found in Multimedia Appendix 1. The time group variables, for levels 2 and 3 (the hour groups 12 - 16 exclusive and 16 - 20 exclusive), were significant predictors of the occurrence of agitation and motor agitation. The β coefficients for time group 3 were higher (β=0.83 - 0.94) than the coefficients for time group 2 (β=0.51 - 0.61) across all agitation and motor agitation models. This indicates an increase in the log odds of agitation and motor agitation as the day progresses. The time group variables were not significant predictors of verbal agitation. The “hallway” level of the majority room group variable was a significant predictor for all types of agitation across all models (β=0.85 - 1.08). The β coefficients for the “hallway” room group were consistently higher for motor agitation than for verbal agitation and agitation. The reference level used for the calculation of these estimates for the majority room group variable was the “living room.” This means that the effects of other locations are relative to the outcomes and variables at the living room locations.

Regarding the inclusion of environmental variables (models 3 - 5), the SD of sound was a significant predictor of agitation and verbal agitation (β=0.42 and β=0.68, respectively). The β coefficient estimate for verbal agitation was higher than that of agitation. The mean light level was a significant predictor for motor agitation only with a coefficient estimate of −0.61 for the light-only model (model 4) and an estimate of −0.60 for the light and sound model (model 5). These negative coefficient estimates indicate that an increase in the mean light level would lead to a decrease in the log odds of the outcome variable (motor agitation).

XSL•FO
RenderX

Discussion

Principal Findings and Comparison to Prior Work

The results presented in this paper explore the relationship between environmental factors, such as light, sound, and temperature, on agitation and 2 of its subtypes: verbal and motor agitation. Exploratory models were used to identify the most informative features per modality and outcome variable, based on the magnitude and significance of their relationships to agitation, motor, and verbal agitation. With these features identified, agitation and agitation type–specific models were built up, considering additional contextual variables like the room-level location of the patient, to quantify further the relationship between each selected environmental variable and the relevant outcome variable.

With the time window length set at 21 minutes, the most informative features were selected using the results summarized in Figure 4. The selected features were then used in agitation type–specific models built up from a model containing only a random intercept per patient (model 0). The most informative features were not consistent across all modalities and outcome variables, already an initial indication of possible type-specific triggers of agitation. This indication is reinforced throughout the forward selection process, where the most informative modality also differed depending on the outcome variable analyzed.

For motor agitation, light (mean) was consistently the predictor with the largest β coefficient estimate. This is present in Multimedia Appendix 1 (β mean light=−0.61; *P*=.02) as well as in the model build-up results summarized in Table 1. After the inclusion of both contextual variables, time of day and majority room, the addition of light led to an improved model fit for motor agitation (*P*=.004), where the addition of the SD of sound did not. This further solidified the specific importance of light for this type of agitation. Current literature relates indoor light levels to spatial or temporal disorientation [29,30], where lower light levels lead to increased levels of disorientation, which can manifest as wandering. Additionally, light and sundowning are thought to be related through the disturbed sleep-wake cycle often seen in people with dementia [16,23]. While studies on the impact of lighting in BPSDs have shown mixed results, several have shown an association between lighting parameters and specific behavioral symptoms. Figueiro et al [16] reported that their lighting intervention (consisting of tailored "bluish-white" lighting) significantly decreased agitation scores on the Cohen-Mansfield Agitation Inventory and resulted in an increase in circadian entrainment. The increase in the β coefficient estimates when comparing time group 2 and time group 3 (2: β=0.51 - 0.61 and 3: β=0.83 - 0.94) could also be explained by a relationship between motor agitation and sundowning. This contrasts with the insignificant time predictors seen in the models for verbal agitation. It is possible that observed sundowning is primarily driven by the observation of motor agitation. Algase et al [31] reported that wanderers wandered less when light levels were lowered. At first glance, this contradicts the result found here; however, this could be explained by a difference in baseline light level [31]. Their study is based on the NDB model, which identifies internal or "background" and proximal factors for behaviors like wandering and agitation. However, in this model, and in other theoretical frameworks focusing on agitation triggers, there are often 2 sides to environmental triggers: overstimulation and understimulation or boredom [12,32]. In this case, our result showing that low mean light levels increase the occurrence of motor agitation, and the result of Algase et al [31] showing that lowered light levels decreased the occurrence of wandering (a subtype of motor agitation), can coexist. This is especially the case when comparing results with different baselines: the light at one ward can be baseline overstimulating, therefore leading to decreases in wandering when stimulation is lowered, while another ward—like the one here (median light levels per room <100 lux; median 22, IQR 7-64 lux)—can be baseline understimulating (eg, low light levels) and lead to decreases in wandering or motor agitation when stimulation is increased. Incorporating additional information on light parameters (eg, circadian stimulus of light, correlated color temperature, and red, green, and blue) on the ward could expand this analysis, given the relationship between the specific aspects of light and the regulation of the circadian system [16,33]. It is also possible that differences between our results and those from Algase et al [31] originate from subtle differences in wandering and motor agitation. Despite the disagreement on the influence of light on agitation, the significance of the "hallway" room location as a predictor variable for agitation (β=0.85 - 1.08 depending on the model structure) is in agreement with the location, where the highest proportion of wandering (along with in the dining room) is seen in Algase et al [31].

For verbal agitation, sound was the most informative modality. This was seen both in the model build-up ANOVA results (Table 1) and in the predictor estimates from the models in Table S3 in Multimedia Appendix 1. The SD of sound was chosen as the most informative sound feature across all outcome variables, with the largest β estimate (β=0.653) for verbal agitation, followed by agitation (β=0.452), and 0.274 for motor agitation. This result is in line with Algase et al [31], where wandering was reported to occur more frequently with higher SDs of sound. However, wandering is considered a subtype of motor agitation [28]. Therefore, we would expect that the SD of sound is also a significant predictor of motor agitation. For our results, this is not the case, although the coefficient estimate does have the same sign, indicating a similar trend (increased SD of sound, increased motor agitation). The difference in magnitudes of the β coefficient estimates between SD of sound and verbal agitation compared to agitation is likely driven by the "agitation" outcome variable, including all types of agitation, and not specifically comparing the presence or absence of verbal agitation. Sound, in general, being a significant predictor for agitation, is in line with prior research, showing that increased sound levels lead to increased agitation [17,18]. Following the NDB model, this could be influenced by the number of people present in a particular room (eg, possible crowding) increasing sound parameters [13]. The relationship between verbal agitation and sound parameters could also be due to overstimulation, as hypothesized in the Unmet Needs model [12].

XSL•FO
RenderX

Temperature was left out of model optimization, as it was not significantly related to any outcome variables. Existing research has shown that the variation in temperature from ideal is a predictor of agitation. It is therefore possible that this variation, which could be in either direction from ideal (too hot or too cold), is not adequately captured in the features used for these analyses. Future work could use different features to quantify, for example, the amount of time that the temperature differed from the mean by at least 1 or 2 SDs. This feature is more closely aligned with the significant predictor reported in Tartarini et al [15] than the SD of temperature used here.

These results, correlating environmental sound parameters in time windows before an agitated survey is recorded, build upon existing research of predictive agitation models by making the distinction between different agitation types and possible environmental triggers. This is in line with the Cohen-Mansfield "Unmet Needs" model [12], where different unmet needs (possibly triggered by environmental parameters) are said to be related to different behaviors, like the expression of agitation of different types. Key differences between predictors for motor agitation and verbal agitation have been highlighted, and taking these differences into account could substantially improve the performance of predictive models. These agitation type–specific results have the potential to drive explainable machine learning models by keeping clinically interpretable knowledge in the model development process. Nevertheless, if an agitation model is being developed without the integration of type-specific information, sound appears to be a stronger predictor due to improved model fits (Table 1) compared to the addition of light as well as larger coefficient estimates across models as seen in Table S3 in Multimedia Appendix 1. It is also important to keep important contextual information about the agitation episode in the model as well, as indicated by the significant time of day and majority room parameters. This is in line with the work by Stojchevska et al [34], where adding context to stress modeling improved model performance.

Strengths and Limitations

A larger sample size, allowing for a confirmatory analysis instead of an exploratory one, could further strengthen the results reported here. It is possible that β coefficients not reported as significant would be significant when enough surveys were included to capture the full range of each feature more confidently. Additionally, the dataset size here could explain disagreements with existing literature. Other research has investigated how specific aspects of sound were correlated with agitation. Including frequency-level features of sound could allow for a more in-depth analysis of the relationship between sound and different agitation types.

Future Directions

Future work can combine the environmental triggers identified here with the quantified physiological response seen when a patient is agitated to form a more holistic model of agitation without the need for survey-based labels. Furthermore, due to the differences seen between agitation subtypes, future work comparing the efficacy of interventions for specific types of agitation may be more informative than focusing on all types of agitation.

Conclusions

The results presented here provide fundamental information supporting the understanding of environmental triggers of agitation. Using various theoretical frameworks to guide this analysis, we have quantified significant relationships between several environmental factors and specific agitation types as well as agitation as a whole. We found that motor agitation and verbal agitation differ in the environmental data modality that provides the most information.

Motor agitation was shown to be most related to light levels on the ward, where low light levels increased the levels of motor agitation shown. In contrast, increased verbal agitation was related to higher SDs of sound levels. These results are in line with existing theoretical frameworks of the etiology of agitation, particularly they focus on balancing stimulation levels on the ward. In addition, we have identified the relevant time window for (just-in-time) prediction of agitation using these factors. Future agitation prediction models can incorporate this additional domain knowledge of agitation type–specific triggers and related contextual parameters to improve model performance. Furthermore, nonpharmacological interventions aimed at the patient's direct environment can be tailored depending on the type of agitation most often expressed. Models like these, supporting autonomous identification and prediction of agitation, can help alleviate the burden on the caregiver by reducing the impact of agitation on care, therefore improving both patient and caregiver quality of life.

Acknowledgments

The authors thank the nurses and other clinical staff of Cog K for their support of data collection. This study was funded by the province of Gelderland (to AVK and CVH); the KU Leuven (C2M/23/053 to MVDB and MDV); the "Funds Malou Malou, Perano, Georgette Paulus, JMJS Breugelmans and Gabrielle, François and Christian De Mesmaeker," managed by the King Baudouin Foundation of Belgium (2021-J1990130-222081 to MVDB); the KU Leuven and the University of Melbourne (KU Leuven Global PhD partnerships to MVDB); and the "Klinische onderzoeks- en opleidingsraad (KOOR)" of the University Hospitals Leuven (to MVDB). The funders played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript.

XSL•FO
RenderX

Data Availability

The datasets generated or analyzed during this study are not publicly available due to their containing information that could compromise the privacy of research participants but are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Further study and analysis details.
[DOCX File, 611 KB - formative_v9i1e60274_app1.docx]

References

1. Beeri MS, Werner P, Davidson M, Noy S. The cost of behavioral and psychological symptoms of dementia (BPSD) in community dwelling Alzheimer's disease patients. Int J Geriatr Psychiatry 2002 May;17(5):403-408. [doi: 10.1002/gps.490] [Medline: 11994927]
2. Hongisto K, Hallikainen I, Selander T, et al. Quality of Life in relation to neuropsychiatric symptoms in Alzheimer's disease: 5-year prospective ALSOVA cohort study. Int J Geriatr Psychiatry 2018 Jan;33(1):47-57. [doi: 10.1002/gps.4666] [Medline: 28067961]
3. Karttunen K, Karppi P, Hiltunen A, et al. Neuropsychiatric symptoms and quality of life in patients with very mild and mild Alzheimer's disease. Int J Geriatr Psychiatry 2011 May;26(5):473-482. [doi: 10.1002/gps.2550] [Medline: 21445998]
4. Chen P, Guarino PD, Dysken MW, et al. Neuropsychiatric symptoms and caregiver burden in individuals with Alzheimer's disease: the TEAM-AD VA cooperative study. J Geriatr Psychiatry Neurol 2018 Jul;31(4):177-185. [doi: 10.1177/0891988718783897] [Medline: 29966477]
5. Patrick KS, Gunstad J, Martin JT, Chapman KR, Drost J, Spitznagel MB. Specific agitation behaviours in dementia differentially contribute to aspects of caregiver burden. Psychogeriatrics 2022 Sep;22(5):688-698. [doi: 10.1111/psyg.12871] [Medline: 35853570]
6. Okura T, Plassman BL, Steffens DC, Llewellyn DJ, Potter GG, Langa KM. Neuropsychiatric symptoms and the risk of institutionalization and death: the aging, demographics, and memory study. J Am Geriatr Soc 2011 Mar;59(3):473-481. [doi: 10.1111/j.1532-5415.2011.03314.x] [Medline: 21391937]
7. Peters ME, Schwartz S, Han D, et al. Neuropsychiatric symptoms as predictors of progression to severe Alzheimer's dementia and death: the Cache County Dementia Progression Study. Am J Psychiatry 2015 May;172(5):460-465. [doi: 10.1176/appi.ajp.2014.14040480] [Medline: 25585033]
8. Cummings J, Mintzer J, Brodaty H, et al. Agitation in cognitive disorders: International Psychogeriatric Association provisional consensus clinical and research definition. Int Psychogeriatr 2015 Jan;27(1):7-17. [doi: 10.1017/S1041610214001963] [Medline: 25311499]
9. Richards KC, Beck CK. Progressively lowered stress threshold model: understanding behavioral symptoms of dementia. J Am Geriatr Soc 2004 Oct;52(10):1774-1775. [doi: 10.1111/j.1532-5415.2004.52477.x] [Medline: 15450062]
10. Smith M, Hall GR, Gerdner L, Buckwalter KC. Application of the progressively lowered stress threshold model across the continuum of care. Nurs Clin North Am 2006 Mar;41(1):57-81. [doi: 10.1016/j.cnur.2005.09.006] [Medline: 16492454]
11. Lawton MP, Simon B. The ecology of social relationships in housing for the elderly. Gerontologist 1968;8(2):108-115. [doi: 10.1093/geront/8.2.108] [Medline: 5657480]
12. Cohen-Mansfield J, Dakheel-Ali M, Marx MS, Thein K, Regier NG. Which unmet needs contribute to behavior problems in persons with advanced dementia? Psychiatry Res 2015 Jul 30;228(1):59-64. [doi: 10.1016/j.psychres.2015.03.043] [Medline: 25933478]
13. Algase DL, Beck C, Kolanowski A, et al. Need-driven dementia-compromised behavior: an alternative view of disruptive behavior. Am J Alzheimers Dis (Columbia) 1996 Nov;11(6):10-19. [doi: 10.1177/153331759601100603]
14. Whall AL, Kolanowski AM. Special Section—Behavioral symptoms of dementia: their measurement and intervention. Editorial: The need-driven dementia-compromised behavior model—a framework for understanding the behavioral symptoms of dementia. Aging Ment Health 2004 Mar;8(2):106-108. [doi: 10.1080/13607860410001649590]
15. Tartarini F, Cooper P, Fleming R, Batterham M. Indoor air temperature and agitation of nursing home residents with dementia. Am J Alzheimers Dis Other Demen 2017 Aug;32(5):272-281. [doi: 10.1177/1533317517704898] [Medline: 28429641]
16. Figueiro MG, Plitnick BA, Lok A, et al. Tailored lighting intervention improves measures of sleep, depression, and agitation in persons with Alzheimer's disease and related dementia living in long-term care facilities. Clin Interv Aging 2014;9:1527-1537. [doi: 10.2147/CIA.S68557] [Medline: 25246779]

XSL•FO
RenderX

17. Janus SIM, Kosters J, van den Bosch KA, Andringa TC, Zuidema SU, Luijendijk HJ. Sounds in nursing homes and their effect on health in dementia: a systematic review. Int Psychogeriatr 2021 Jun;33(6):627-644. [doi: 10.1017/S1041610220000952] [Medline: 32594932]
18. Joosse LL. Do sound levels and space contribute to agitation in nursing home residents with dementia? Res Gerontol Nurs 2012 Jul;5(3):174-184. [doi: 10.3928/19404921-20120605-02] [Medline: 22716652]
19. Au‐Yeung WM, Miller LM, Lindauer A, et al. High dimensional longitudinal assessment of ambient home environments and their association with behaviors of people with cognitive decline. Alzheimer's & Dementia 2023 Dec;19(S11):e082052. [doi: 10.1002/alz.082052]
20. Au-Yeung WTM, Miller L, Beattie Z, et al. Sensing a problem: proof of concept for characterizing and predicting agitation. Alzheimers Dement (N Y) 2020;6(1):e12079. [doi: 10.1002/trc2.12079] [Medline: 32864417]
21. Homdee N. Prediction of dementia-related agitation using multivariate ambient environmental time-series data. arXiv. Preprint posted online on Feb 17, 2020. [doi: 10.48550/arXiv.2002.07237]
22. HekmatiAthar S, Goins H, Samuel R, Byfield G, Anwar M. Data-driven forecasting of agitation for persons with dementia: a deep learning-based approach. SN Comput Sci 2021;2(4):326. [doi: 10.1007/s42979-021-00708-3] [Medline: 34109317]
23. Bliwise DL, Carroll JS, Lee KA, Nekich JC, Dement WC. Sleep and "sundowning" in nursing home patients with dementia. Psychiatry Res 1993 Sep;48(3):277-292. [doi: 10.1016/0165-1781(93)90078-u] [Medline: 8272449]
24. Guu T, Aarsland D, ffytche D. Light, sleep‐wake rhythm, and behavioural and psychological symptoms of dementia in care home patients: revisiting the sundowning syndrome. Int J Geriat Psychiatry 2022 May;37(5) [FREE Full text] [doi: 10.1002/gps.5712]
25. Silva MWB, Sousa-Muñoz RL, Frade HC, Fernandes PA, Magalhães ADO. Sundown syndrome and symptoms of anxiety and depression in hospitalized elderly. Dement Neuropsychol 2017 Jun;11(2):154-161. [doi: 10.1590/1980-57642016dn11-020008]
26. Davidoff H, Van den Bulcke L, Vandenbulcke M, et al. Toward quantification of agitation in people with dementia using multimodal sensing. Innov Aging 2022;6(7):igac064. [doi: 10.1093/geroni/igac064] [Medline: 36600807]
27. Myin-Germeys I, Kasanova Z, Vaessen T, et al. Experience sampling methodology in mental health research: new insights and technical developments. World Psychiatry 2018 Jun;17(2):123-132. [doi: 10.1002/wps.20513] [Medline: 29856567]
28. Rosen J, Burgio L, Kollar M, et al. The Pittsburgh Agitation Scale: a user‐friendly instrument for rating agitation in dementia patients. Am J Geriatr Psychiatry 1994;2(1):52-59. [doi: 10.1097/00019442-199400210-00008] [Medline: 28531073]
29. Bautrant T, Grino M, Peloso C, et al. Impact of environmental modifications to enhance day-night orientation on behavior of nursing home residents with dementia. J Am Med Dir Assoc 2019 Mar;20(3):377-381. [doi: 10.1016/j.jamda.2018.09.015] [Medline: 30455049]
30. Goudriaan I, van Boekel LC, Verbiest MEA, van Hoof J, Luijkx KG. Dementia enlightened?! A systematic literature review of the influence of indoor environmental light on the health of older persons with dementia in long-term care facilities. Clin Interv Aging 2021;16:909-937. [doi: 10.2147/CIA.S297865] [Medline: 34079240]
31. Algase DL, Beattie ERA, Antonakos C, Beel-Bates CA, Yao L. Wandering and the physical environment. Am J Alzheimers Dis Other Demen 2010 Jun;25(4):340-346. [doi: 10.1177/1533317510365342] [Medline: 20378834]
32. Kales HC, Gitlin LN, Lyketsos CG, Detroit Expert Panel on Assessment and Management of Neuropsychiatric Symptoms of Dementia. Management of neuropsychiatric symptoms of dementia in clinical settings: recommendations from a multidisciplinary expert panel. J Am Geriatr Soc 2014 Apr;62(4):762-769. [doi: 10.1111/jgs.12730] [Medline: 24635665]
33. Blume C, Garbazza C, Spitschan M. Effects of light on human circadian rhythms, sleep and mood. Somnologie (Berl) 2019 Sep;23(3):147-156. [doi: 10.1007/s11818-019-00215-x] [Medline: 31534436]
34. Stojchevska M, Steenwinckel B, Van Der Donckt J, et al. Assessing the added value of context during stress detection from wearable data. BMC Med Inform Decis Mak 2022 Oct 15;22(1):268. [doi: 10.1186/s12911-022-02010-5] [Medline: 36243691]

Abbreviations

BPSD: behavioral and psychological symptom
EMA: ecological momentary assessment
NDB: need-driven dementia-compromised behavior
PAS: Pittsburgh Agitation Scale
PLST: progressively lowered stress threshold

XSL•FO
RenderX

Edited by A Mavragani; submitted 06.05.24; peer-reviewed by A Hidki, K Sarma; revised version received 30.04.25; accepted 06.05.25; published 27.08.25.

Please cite as:
Davidoff H, Van Kraaij A, Lutin E, Van den Bulcke L, Vandenbulcke M, Van Helleputte N, De Vos M, Van Hoof C, Van Den Bossche M
Environmental Triggers of Specific Subtypes of Agitation in People With Dementia: Observational Study
JMIR Form Res 2025;9:e60274
URL: https://formative.jmir.org/2025/1/e60274
doi:10.2196/60274

© Hannah Davidoff, Alex Van Kraaij, Erika Lutin, Laura Van den Bulcke, Mathieu Vandenbulcke, Nick Van Helleputte, Maarten De Vos, Chris Van Hoof, Maarten Van Den Bossche. Originally published in JMIR Formative Research (https://formative.jmir.org), 27.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Exploring Co-production as an Implementation Strategy for Trauma-Informed Care in a Youth-Focused HIV Clinic in Memphis, Tennessee: Mixed Methods Research

Leslie Lauren Brown[1], MSSW, PhD; Megan L Wilkins[2], PhD; Robert Kenneth Douglas McLean[3], PhD; Jamie Lynn Stewart[1], MSSW; Nehali D Patel[2], MD; Almariana Jessica Acuña[1], MA; Carolyn Marie Audet[4], MPH, PhD; Jessica M Sales[5], PhD; April C Pettit[6], MD, MPH; Latrice C Pichon[7], MPH, PhD

[1]Family and Community Medicine, Psychiatry and Behavioral Sciences, School of Medicine, Meharry Medical College, 1005 Dr. D.B. Todd Jr. Boulevard, Nashville, TN, United States

[2]Department of Infectious Diseases, St. Jude Children's Research Hospital, Memphis, TN, United States

[3]The Centre for Practice-Changing Research, Ottawa Hospital Research Institute, Ottawa, ON, Canada

[4]Department of Health Policy and the Institute for Global Health, Vanderbilt University Medical Center, Nashville, TN, United States

[5]Rollins School of Public Health, Emory University, Atlanta, GA, United States

[6]Infectious Disease Division, School of Medicine, Vanderbilt University Medical Center, Nashville, TN, United States

[7]School of Public Health, University of Memphis, Memphis, TN, United States

Corresponding Author:
Leslie Lauren Brown, MSSW, PhD
Family and Community Medicine, Psychiatry and Behavioral Sciences, School of Medicine, Meharry Medical College, 1005 Dr. D.B. Todd Jr. Boulevard, Nashville, TN, United States

Abstract

Background: Memphis, Tennessee is second in the nation for HIV incidence, with one in three diagnoses among youth. Psychological trauma disproportionately impacts youth with HIV, compared with HIV-negative counterparts, requiring community-led and trauma-informed solutions to address mental wellness among youth with HIV. However, a dearth of research concentrates on trauma-informed care (TIC) for this population, with little exploration among youth-centered HIV care settings or into strategies for mobilizing communities to develop solutions. Research co-production, an approach in which research beneficiaries engage in research as cooperative partners, aligns with the TIC focus on collaborative decision-making and could be an effective strategy for facilitating collaborative TIC adoption, but formative research is needed to explore this potential.

Objective: We sought to explore TIC implementation determinants and contextual factors that might influence research co-production as a strategy for implementation, including appetite for evidence-based approaches, support for co-production, and resources for capacity building.

Methods: We applied an exploratory sequential mixed methods design to identify potential barriers and facilitators to TIC implementation in a youth-focused clinic and contextual factors relative to co-production. All clinic personnel were purposively invited to complete semistructured interviews. Thematic analysis, via four cycles of coding, was applied using the Consolidated Framework for Implementation Research 2.0 to qualitative data. Subsequently, a steering committee of clinic personnel was invited to complete surveys, applying the Research Quality Plus for Co-Production framework to explore co-production factors. A deliberative dialog approach was applied to analyze these findings and synthesize them with Consolidated Framework for Implementation Research.

Results: A total of 20 personnel completed interviews, and 9 completed surveys. Potential facilitators included perceived clinic cohesiveness, equity focus, and prioritization or compatibility of TIC. Potential barriers included perceived disconnect between the clinic and larger hospital (in which youth with HIV were seen as stigmatized in other areas of the hospital), sustainability concerns related to a perceived lack of championing by leaders, insufficient mental health protocols, a lack of formal patient feedback procedures, and a lack of protected time for personnel activity engagement. Survey responses suggested that the clinic is likely supportive of evidence-based approaches (mean 3.6, SD 0.70) and collaborative research (mean 3.1, SD 0.31) and empowers personnel to participate (mean 3.1, SD 0.22). Conducive to co-production, the environment was seen as learning-centered, where evidence and standardized or validated approaches are prioritized, and there is an openness for innovation, with a focus on health disparities and quality improvement. Potential barriers included change-resistant staff, role silos, and underutilization of staff skills, coupled with a lack of formal research training and time constraints.

XSL•FO
RenderX

Conclusions: Findings suggested that TIC implementation is likely to be embraced in the clinic, with co-production perceived as useful and fitting. However, greater effort is needed to integrate patient experiences and test co-production as a TIC implementation strategy.

(JMIR Form Res 2025;9:e66426) doi:10.2196/66426

KEYWORDS

HIV; trauma-informed care; co-production; youth; research quality

Introduction

Memphis, Tennessee, has seen a worsening HIV epidemic and is now second in the nation for incidence [1,2]. Youth aged 13 to 24 years in the area account for one in three people newly diagnosed with HIV [3], representing a 40% increase in HIV incidence between 2018 and 2023 [4]. Psychological trauma is a modifiable but understudied driver of poor health outcomes among youth with HIV, having been connected with care attrition, poor medication adherence, and unsuppressed viremia [5-9]. As youth with HIV in Tennessee experience disproportionately higher rates of posttraumatic stress disorder than HIV-negative counterparts (41% vs 2%, respectively) [510,11], local plans to end the HIV epidemic in Memphis and other priority jurisdictions identify implementation of trauma-informed care (TIC) as critical for improving HIV outcomes [12,13]. Yet, strategies for implementing TIC effectively in youth-focused HIV care clinics remain understudied. Meanwhile, HIV governing bodies urge that the HIV epidemic will not be halted without community-led solutions to attenuate health disparities [2,14].

A key feature of TIC is the belief that factors related to culture, history, and gender influence health outcomes, and intentional efforts are needed to acknowledge and minimize power differentials as part of decision-making processes [15]. As an evidence-based approach for preparing personnel to realize, recognize, and respond to trauma to resist retraumatization [16], TIC is associated with multilevel improvements. At the personnel level, it has been connected to improvements in organizational culture and professional quality of life [17], and at the patient level, it has the potential to improve health outcomes among youth [18] and adults with HIV [19]. In Tennessee, a trauma intervention was associated with a 15-times increase in retention in HIV care and a 5-times increase in viral suppression [20]. Although TIC is a burgeoning approach to care, strategies for trauma-informed research appear less frequently [21]. Trauma-informed research focuses attention toward fostering safe research activities, so that all involved in research recognize potential signs of trauma-related distress and protocols are in place to minimize distress [21]. For example, TIC principles call for efforts to promote collaborative decision-making and mutuality while considering historical contexts [2]. Despite the potential for TIC to improve health outcomes, replicable implementation of TIC presents as challenging, as it relies on flexible principles that vary by setting as to how they can or should be applied [22].

Research co-production is an umbrella term for many approaches to creating knowledge where persons who will benefit from the research (ie, knowledge users and beneficiaries)—such as people with HIV or those with experience providing care in HIV care organizations—are engaged in the research process as cooperative partners [23]. Collaborative research efforts, in which research knowledge users or beneficiaries are engaged as equal decision makers, are therefore well-positioned to ensure solutions are rooted in lived experience, experiential understandings of preferences for care, and viable research strategies. Research conducted without such feedback is more likely to be wasteful, less effective [24], and carries potential to propagate harm by reenacting traumatic power dynamics [25], including the objectification, minimization, and dismissal of voices, perspectives, and priorities of people with little or no political or financial power [22] or those considered to lack “scientific expertise.” [23]

Co-production is highly compatible with TIC, as both emphasize the importance of collaboration, the empowerment and elevation of the voices of persons with lived experience, and attention to minimizing power differentials [15]. However, to our knowledge, no frameworks exist to guide collaborative TIC implementation via research co-production. This knowledge and practice gap poses challenges for developing community-led solutions in HIV care organizations, as scant literature presents findings relative to the usefulness or potential for co-production as a strategy for implementing TIC. Overall, limited literature explores the utility of collaborative research as part of TIC implementation [15]. Specifically, gaps remain relative to how to replicate TIC, indicating a need to develop and test reproducible strategies and indications for how they might be culturally adapted across settings. Further, little information exists about how HIV care personnel perceive the potential usefulness of adopting a co-production approach.

There is an urgent need for community-led approaches to attenuate trauma effects in adolescent HIV clinics in the Southern United States. However, formative research is needed to first understand how personnel working in these care spaces view TIC and co-production. Thus, the primary objective of this study was to conduct exploratory research to guide the preliminary development of a collaborative TIC implementation plan in a youth-focused HIV clinic. We sought to explore personnel-perceived barriers and facilitators to TIC implementation in the setting, as well as perception of current research practices that might influence adoption of a co-production strategy. Specifically, we sought to understand how personnel perceived the research milieu relative to knowledge use environment, evidence as a priority, and the overall application of collaborative research. By answering these research questions, we aim to apply findings to develop a replicable approach for guiding collaborative TIC

XSL•FO
RenderX

implementation with and for people with lived experience who have been most affected by trauma.

Methods

Study Setting

This study focuses primarily on the experiences of personnel from an adolescent and young adult-focused HIV clinic in the priority jurisdiction of Memphis, Tennessee. This clinic, serving approximately 200 patients, is housed within a larger tertiary pediatric hospital, with a primary focus on hematology and oncology. The clinic is the primary provider for HIV care for youth aged 14‑24 years in a tristate region. Patients are primarily racial minorities, with about 70% represented by cisgender men who have sex with men and the remaining cisgender women. Patients of this clinic have access to institutional transportation, meals, research activities, school advocacy, mental health services, and case management via social workers, a psychologist, a psychiatrist, and a chaplain, sexually transmitted infection testing and treatment, medical care from advanced practice providers supervised by physicians, and an on-site laboratory and pharmacy.

Ethical Considerations

Ethics approval was granted by the primary institution from which data were collected to Meharry Medical College (21-07-1105), with review also conducted by the University of Memphis (PRO-FY2022-201). All participants completed an informed consent process prior to participating in the study; these were administered verbally, with written consent provided. All identifying information was kept confidential, with only the study principal investigator and two coders having identifiable information. All identifiers were removed, including connections of any quotes with participant demographic information that could be discernible. All personnel participating in interviews received a US $50 gift card as compensation, and all steering committee members received a US $50 gift card as compensation for attending the 1.5-hour meeting.

Study Approach

Overall, this research is a product of a pragmatic approach in which research is focused on practical understandings of concrete, real-world issues [26]. To advance a pragmatic approach, we applied a constructivist stance to collecting data through multiple methods and from multiple sources, interpreting their meaning, and reporting key findings and implications. A constructivist stance emphasizes the importance of individuals constructing their own understanding and interpretation of the world through personal experience and reflection and lends itself well to the objectives of research co-production, where ongoing, collaborative input shapes the research design and execution [26-29].

We used a multiple-methods design to complement the constructivist stance, recognizing that this combination would best facilitate the integration of varied perspectives within our team [26]. Mixed methods designs support the generation of stronger inferences when compared with a single medium of data collection by enabling the capturing of a more holistic and richer illustration of complex phenomena [30]. Moreover, this design is recommended for research that is community-engaged [31]; it aligns with the iterative nature of co-production and is well-suited for TIC as a paradigm, indicating the need to include voices from multiple platforms in a system and through various mediums to enhance inclusivity. An exploratory sequential mixed methods design was applied for data collection: first, collecting qualitative data through in-depth interviews; second, administering a brief quantitative and qualitative survey instrument; and then, synthesizing these findings. This approach facilitated the triangulation of multiple lines of evidence, multiple sources of insight, and an integrated analysis and discussion.

The current formative research presents one part of an ongoing effort to collaboratively implement TIC in the youth-focused HIV clinic. Previous findings have informed the current stage of research, and current findings will be applied to direct and tailor future TIC implementation specific to the setting. The first phase of the overall study was conducted during the first year of the project (2021) and has been published elsewhere [32]. This phase marked the inception of the project and was initiated by the focal clinic's request for technical assistance with TIC implementation from the study team. During that phase, the study team primarily focused on soliciting or integrating local feedback on grant objectives and research questions via an established local community workgroup, Connect 2 Protect, and with internal and external local champions. These efforts were followed by information-sharing and rapport-building activities, including the study team presenting on TIC to clinic staff, as well as the institution as a whole, via a hospital-wide Grand Rounds didactic and the construction of a process map of clinic services. The process map was generated through discussion groups with personnel and a patient peer representative group; those findings served to construct a joint understanding of standard care and available resources, as well as limitations and the need to develop protocols relative to TIC. Findings from the process map were used to update a pre-existing interview guide that had also been developed through community workgroup discussions for TIC in HIV care spaces [33,34].

Current Data Collection

The next phase of activities for this study occurred in the spring of 2022 and primarily centered on engaging clinic personnel to purposively participate in one-on-one key informant interviews to assess barriers and facilitators to TIC implementation. Inclusion criteria included all patient-facing clinic personnel. Once identified as clinic personnel by the site principal investigator, eligible individuals were invited via email to participate. All individuals who volunteered to participate were invited for an interview and asked to choose if they preferred a web-based (Zoom; Zoom Video Communications Inc) or an in-person platform. The 30-minute to 1-hour interviews were conducted by two master's-level public health students operating under the supervision of a PhD-level investigator (LCP), with the interviewers having no previous relationship with personnel of the clinic. Both interviewers received training on the topics of HIV and TIC prior to conducting the interviews. Interviews were then audio recorded and transcribed verbatim.

XSL•FO
RenderX

During the final phase of this study (summer 2023), a steering committee of approximately 18 clinic personnel was convened to begin designing a plan for collaborative TIC implementation. The steering committee composition was built primarily from clinic personnel who volunteered to participate and was comprised of providers, researchers, and administrators and additionally included the project principal investigator and several topic experts specializing in TIC and co-production, with the intent to invite patients onto the committee after discussions within the group of best practices to promote patient emotional safety. The steering committee met three times and was then invited to participate in a web-based survey to elicit a contextual understanding of baseline levels of the co-production environment using the Research Quality Plus for Co-Production (RQ+4 Co-Pro) framework [35]. Participants received US $50 gift cards for attending steering committee meetings. Multimedia Appendix 1 illustrates how each phase of the research thus far has iteratively built on a collaborative approach for TIC implementation.

Statistical Data Analysis

This paper focuses on findings from a mixed methods designed study (Multimedia Appendix 1), which collectively sought to (1) explore implementation determinants and (2) conduct a dynamic evaluation. First, primary data from interviews were analyzed using the Consolidated Framework for Implementation Research (CFIR) 2.0 [36] to deductively code qualitative data, allowing CFIR constructs and domains to emerge de novo in our analysis. Four cycles of coding were conducted by researchers who were external to the clinic environment, with all three being female, with a mixed racial composition, including AJA and JLS for all cycles, and LLB was involved in the first and final cycle. First-cycle coding began with the three coders coding the first transcript together, followed by the two coders (JLS and AJA) independently coding each transcript by reading transcripts as PDFs and documenting notes into Excel (Microsoft Corp) by CFIR domains. Spreadsheets were organized first by CFIR domains, with columns listed for a summary of each domain, themes presenting as barriers with sample quotes, and themes presenting as facilitators with sample quotes. CFIR subdomains and constructs were then listed within each CFIR domain, with the same information for each subdomain and construct. Upon review of these results, all coders felt data saturation had been reached and no further interviews were required. Second-cycle coding occurred with the two reviewers verbally discussing the independently-developed results to begin developing a codebook based on presenting CFIR themes and revising contextually-defined definitions and summaries of each construct. Third-cycle coding entailed reaching consensus on which quotes belonged in which CFIR domains and constructs (eg, differentiating similar themes by grouping together ideas into one theme most befitting). A draft of these findings was then developed as a narrative Word (Microsoft Corp) document, with researcher LLB reviewing themes and discussing discrepancies with JLS and AJA to reach consensus. Results from Phase II were then reviewed with the steering committee to discuss the meaningfulness of findings and a general plan for coproducing the implementation of TIC in light of the findings.

Results from interviews were used to adapt the pre-existing framework, RQ+4 Co-Pro, which uses both quantitative and qualitative survey responses [37].

Adaptations focused on ensuring the language was appropriate to the current setting and study, relative to TIC in the HIV clinic, with personnel who were generally unfamiliar with the concept of co-production and included insights from the first phase findings [32]. Three authors (LLB, RKDM, and JLS) worked together to adapt the tool. The RQ+4 Co-Pro framework was designed to explore the local context for co-production. The RQ+4 Co-Pro contextual factors are: (1) Knowledge Use Environment, measuring appetite for research-generated knowledge and the ability to use evidence-based approaches to inform practices and policies (restrictive vs empowering); (2) Research Environment, measuring the extent to which an environment is supportive of co-production as an effective and legitimate approach to research (restrictive vs empowering); and (3) Capacities for co-production, measuring the extent to which the environment is devoting time and resources to build the capacity of researchers and users or beneficiaries to engage in co-production (no focus vs strong focus). Each of the three contextual factors includes a 4-point rubric to enable a quantitative rating and an open-ended short response to allow examples and qualitative descriptions.

To analyze RQ+4 Co-Pro results, we applied content analysis, using the RQ+4 Co-Pro framework to organize responses, with themes reviewed between three coders (LLB, JLS, and RKDM). Descriptive statistics were used to analyze quantitative response data for each of the three RQ+4 Co-Pro rubrics. Following the analysis of interviews using CFIR, a final cycle of coding was conducted by LLB and RKDM to synthesize findings from CFIR with those of RQ+4 Co-Pro. To synthesize data, we followed a deliberative dialog approach [38] to collectively interpret how findings from each phase of the research contributed to an overall understanding of the potential utility of using a collaborative approach as a trauma-informed research strategy in the clinic. Deliberative dialog is an approach that allows multiple parties to consider relevant information from multiple points of view [38]. Our process involved two team members (LLB and RKDM) undertaking independent analysis, followed by shared reflection to identify any points of divergence or similarity. These points were explored between the two team members, and consensus was then achieved through review by the full research team. Because the CFIR instrument is not intended for use at the patient level, patient interviews required a separate analytical framework for interpretation and will be presented in a forthcoming study.

Results

Sample

Interviews were conducted (March-August 2022) with 20 (57%) clinic personnel, of which 18 (90%) identified as female, one male (1%), and one (1%) nonbinary; 8 (40%) participants identified as Black, 8 (40%) as White, 2 (10%) as Asian, and 1

XSL•FO
RenderX

(1%) as Native American/Hawaiian; median age was 48 (IQR 42-54) years; median time in the position was 14 IQR (1-20) years. Positions included mental health practitioners (eg, social workers, chaplains, and psychologists), physicians, advanced practice providers (eg, physician assistants and nurse practitioners), nurses, clinical researchers, patient representatives, and administrators. RQ+4 Co-Pro surveys were conducted with a total of 9 (50%) steering committee members in July 2023.

Interview Results

Four CFIR domains emerged as potential implementation determinants, with various CFIR constructs presenting as relevant for each domain: (1) Outer Setting, (2) Inner Setting, (3) Individuals, and (4) Implementation Processes. Findings are organized by domain and constructs below with exemplary quotes. Definitions and other quotes are provided in Multimedia Appendix 2.

Outer Setting

The Critical incidents construct indicated that personnel overall felt the COVID-19 pandemic presented opportunities for strengthening engagement between personnel via new web-based group meetings during a difficult time (eg, hospital-wide groups convened to discuss national and local reports of police brutality). However, personnel also discussed the perception that COVID-19 presented challenges for patient engagement, coupled with a greater need for patient support, with an indication that such challenges potentially moved them further from a TIC paradigm. Additional relevant incidents included the clinic's move to a new building and a new electronic medical record, with both identified as slowing down efforts to work toward TIC implementation. The move to the new building presented both opportunities for the HIV clinic to be more integrated with the overall hospital (ie, potentially destigmatizing HIV), as well as the concern that patients may be exposed to personnel, patients, or families with less experience interacting with persons with HIV or identifying as Lesbian, Gay, Bisexual, Transgender, Queer, Intersex (LGBTQI).

> *I think the pandemic has definitely impacted us and caused more trauma. So, I think we're definitely all very intimate with it. I think everybody's experienced COVID trauma, whether you've had a personal experience or just the way that the world is. So, I think it maybe might open us up to talk […] But I also think there's a barrier with time still.* [Transcript 9]

Local attitudes indicated there was a perceived disconnect between the larger institution and the HIV clinic, with observations that patients with HIV do not receive the same services or amenities as patients in other hospital clinics. One person explained that this was perceived to contribute to stigma not only among patients but also among personnel of the HIV clinic.

> *I don't feel a larger embrace of [the] hospital for our patient population. We tend to be left out of events, there have been times we haven't even heard of events or giveaways. It's like we have been excluded for a long time.* [Transcript 1]

Themes in Partnerships and Connections indicated that personnel felt there was hospital-wide support for multidisciplinary team meetings for providers to discuss difficult cases, and the option for research teams to engage a community advisory board (CAB) for feedback on research. While patients' needs were seen as addressed through advanced referral networks for external mental health services, more community networks may be needed to help patients with basic needs (eg, housing).

> *So that [CAB] feedback is very, very important, and I think that's one of the things that we rely on when we're looking at how we're going to change things or if we are thinking of implementing new things.* [Transcript 10]

Themes in Policies and Regulations indicated that the hospital had several policies and initiatives aligned with the HIV clinic's human equity focus, including employee resource groups with an LGBTQI focus, family advisory groups, and a diversity, equity, and inclusion (DEI) curriculum for personnel. Moreover, it was elucidated that personnel felt the best way to implement sustainable change is by anchoring efforts into hospital-wide policies. Prioritization of confidentiality and privacy was highly valued throughout the institution, with attention to policies that promote patient choice and autonomy regarding information-sharing. While there were institutional mechanisms discussed for obtaining patient feedback and experiences via an annual patient survey and CABs or advisory groups, these mechanisms were seen as needing to be revamped, with attention to reducing power differentials within group settings to increase transparency and dissemination of patient survey findings.

> *Once you make it a structural change, then everybody's on board and responsible for it, and so those tend to be the changes that are most impactful and last.* [Transcript 1]

Inner Setting

The Physical Infrastructure care space was perceived as generally safe and inviting for patients (eg, gender-neutral bathrooms, privacy, closed doors, affirming posters, and well-lit rooms with televisions, games, blankets, and snacks); however, some expressed that the space could be more adolescent-friendly, and the new space (ie, integrated within the hospital) may have reduced privacy and less exam rooms.

> *Now we are in a different clinic, and it's a much newer building. However, you're kind of wide […] out and in the open; you are mingled in with the rest of individuals in the hospital.* [Transcript 17]

The Information, Technology, and Work Infrastructure was perceived as well-developed. Responsibilities and chain of command were seen as clearly demarcated, with various types of mental health assessments conducted by personnel from different disciplines, and comprehensive care provided by an on-site pharmacy (eg, adherence counseling). Conversely, a need was raised to make screenings more comprehensive and deduplicated to reduce repeat suicide assessments and patient

XSL•FO
RenderX

fatigue, and some felt that there was a lack of understanding of duties and activities among personnel of different positions.

> *I feel like some additional questions could be asked specifically about sexual traumas and things of that nature. Because I don't recall asking, we ask about domestic violence, but not anything specifically about sexual.* [Transcript 19]

Communication and Related Connectedness indicated that personnel perceived clinic practices generally as promoting inclusivity and safety among patients, and a strong sense of team integration allowed patients to be well supported. While most staff reported feeling prepared to handle patient trauma due to on-the-job experience, there was an expressed need for patients to be provided more options regarding the order of clinic services received (eg, blood draws at the start of appointments vs later in appointments), and a need for more frequent staff meetings to improve communication about patients' needs, with specific attention to meeting the mental health demands stemming from trauma.

> *We have such […] great communication with the multidisciplinary team, that we're all on board of what is going on with patient[s]. What have they shared.* [Transcript 5]

Human-equity centeredness themes emerged relative to how the clinic fosters respect for patient cultural differences as seen with patient-affirming policies (eg, use of preferred names and patient-identified pronouns). Further, clinic personnel expressed having a high level of care and compassion, with an accentuated respect for providing care equitably and working to destigmatize HIV hospital-wide. Despite this, there was a stated need to train personnel on covert HIV stigma and unconscious bias and to expand the personnel understanding of racial trauma.

> *We always let them know that they can be who they want to be regardless of how they were born or how they identify. We will respect whatever decisions they make […] or […] for the person that they are.* [Transcript 19]

The clinic environment was perceived to be recipient-, deliverer-, and learning-centered with shared values, beliefs, and norms around caring, supporting, and addressing the needs and welfare of recipients and deliverers, as well as the prioritization of psychological safety, continual improvement, and the use of data to inform practice. Personnel were reportedly attuned to patient experiences of trauma and HIV-related stigma, as well as a perceived stigma against providers working in the field of HIV. Generally, personnel felt that work-life balance was promoted in the setting. On the one hand, personnel were seen as empowering patients, as patients are encouraged to communicate their needs and concerns and see themselves as equal players in health care decisions. On the other hand, some providers were seen as having limited flexibility for patients, prioritizing their own work schedule over the patients' when booking appointments. More peer-to-peer support opportunities for patients were an identified need to enhance patient care, and personnel identified that a clinic-level patient needs assessment should be conducted to elicit patient clinic experiences related to safety and culture. Finally, the clinic was seen as a space in which there is a positive learning climate where new practices are relatively easily adopted (eg, past studies evaluating the impact of staff being trained in motivational interviewing and staff wearing pronoun buttons). Greater efforts are reportedly needed in the clinic to reduce burnout and provide more opportunities for staff promotion or appreciation, but some identified more tenured staff appear resistant to innovations.

> *This will just go back to the inability for people to be adaptable to change. So being stuck in those cultural norms that they've already learned, they've already practiced and not being adaptable to learning new things.* [Transcript 12]

A strong tension for change was observed. Personnel stated that TIC implementation was a priority (eg, expressed need for comprehensive and routine patient trauma assessments) but identified time constraints and barriers to providing more robust mental health services (eg, high levels of staff turnover or burnout, regular exposure to patient trauma, need for inclusive policies hospital-wide to reduce patient stigma, or likelihood for retraumatizing patients).

> *… I think primarily it's in staff morale, staff turnover, and then seeing compassion fatigue just in individual providers in stress and overwhelm and maybe even a lack of confidence in their ability to do the work that we know they know how to do.* [Transcript 3]

TIC was stated by all personnel interviewed as highly compatible with the desired clinic culture, the need for training staff, and the need for expanding patient and provider mental health care. Staff reported being mindful of patient stigma and responsive to how trauma effects influence patient behaviors, medication adherence, and treatment outcomes. There was an expressed need to educate patients and nonclinical personnel on how trauma affects health and about trauma response. Finally, there was a stated need to regularly integrate patient feedback and preferences to improve clinic services.

> *I think it [TIC] fits [at our clinic]. I think it's important because if we don't come from that lens, I think we're not providing optimal services. So yeah, I think it's extremely important ... Because the personnel in the [HIV] Clinic are very embracing of our patients and understand they've been exposed to a variety of circumstances.* [Transcript 2]

Personnel stated that TIC implementation was seen as a priority by all clinic personnel, but had been temporarily paused during the relocation and adoption of the new electronic medical record system. Resources identified to further prioritize TIC implementation were time, space, money, tools, and staffing (ie, limited exam rooms for private discussions and overworked providers). Assessments were identified as needing to be reconsidered, including adverse childhood experiences, which were seen as not comprehensively capturing some social determinants of trauma (eg, racial trauma and housing instability), general psychosocial assessments, and suicide risk assessments. While direct service providers reported having access to trainings (eg, motivational interviewing, cultural humility, and implicit bias or diversity), there was a perceived need to ensure all staff, including front desk and administrative,

XSL•FO
RenderX

have access to the same trainings. Enhancements to training, such as more videos and role-plays to help prepare for interactions with patients and trauma, were suggested. Notably, personnel who mentioned tangible barriers (eg, training needs and learning curve) were generally longer-term employees.

> *If all staff [get] involved and [receive] more training, we could pick up on triggers better and be able to get the patient help or share what we saw or perceived. And maybe work more closely with the team that way, especially with the social worker or the psychologist.* [Transcript 5]

Individuals

Project role themes emerged relative to high-level leaders and innovation deliverers. While providers unanimously supported TIC implementation and universal patient trauma assessments, and clinic leaders expressed unequivocal support for TIC, clinic and hospital leadership were perceived as potential barriers if they do not more greatly prioritize staff wellness.

> *I think our biggest hurdle would be leadership. I think if something is not in line with their vision on their time, then it won't be a primary focus.* [Transcript 17]

The project characteristics subdomain included individuals' needs, capabilities, and motivations. Needs emerged relative to staff stress or burnout or turnover and the impact of providing trauma stewardship for patients. Facilitators, conversely, included that personnel generally felt supported by a nurturing team environment and (eg, staff-peer support program and clinician-focused presentations) wellness initiatives (eg, institution-wide mindfulness hour).

> *Listening to people's traumas so often can lead an employee to be overwhelmed. It can lead an employee to being in a space of hopelessness because there's so much trauma and not enough response to that trauma...* [Transcript 16]

Most reported feeling adequately prepared to handle patient trauma due to experience and team support, but in-person trainings (vs web-based or recorded trainings) were seen as a means for improving skills for nonclinical staff to assist with HIV disclosure and effectively responding to patient trauma. Finally, personnel expressed a commitment to providing optimal care and addressing their own personal trauma, privilege, and biases, but efforts to adopt innovations were sometimes stymied by feeling overworked and immersed in a culture where patients are prioritized over personnel well-being.

> *Just high levels of stress ... We're not in a great state right now in terms of emotional well-being of staff ... We need protected time specific to providing self-care. It needs to be a priority and it needs to be modeled as a priority by our staff, by our managers.* [Transcript 5]

Implementation Processes

Overall, the success of TIC implementation was perceived to rest on how well it could impact professional quality-of-life outcomes (ie, burnout) and clinic improvements in trauma screening and assessment protocols so that culturally responsive care is enhanced and more services are provided by people who mirror patients in race or ethnicity, age, gender identity, and lived experience.

> *I feel like that can actually be a hindrance if you don't have people who are relatable or look like the patient. How does that impact them to give you that information and make them feel safe?* [Transcript 17]

Interview and Survey Results Synthesized

Numeric results, summaries of qualitative responses, and sample quotes are provided below from the RQ+4 Co-Pro analysis for each of the framework's three contextual factors. The tables presented inMultimedia Appendices 3-5 present the synthesized results with the two frameworks, CFIR and RQ+4 Co-Pro, integrated together. Specifically, this table shows how CFIR findings reflected levels of evidence, support, and capacity for the collective co-production approach to research. We present a summary of the synthesis findings below, as well as in Multimedia Appendices 3-5.

The contextual factor Knowledge Use Environment (eg, ability to use evidence-based approaches) was on a 4-point rubric (4=Empowering, 3=Supportive, 2=Unsupportive, and 1=Restrictive). The mean response was 3.6 (SD 0.70), indicating an environment highly conducive to using research results in decision-making. Qualitative responses indicated themes that appeared relative to current knowledge or research use, evidence as a priority, and efforts to adopt data-driven approaches (Multimedia Appendix 4). The hospital was perceived to be highly supportive of biomedical research by virtue of being a "research" hospital, with a culture for encouraging personnel to produce research and seek out innovations. However, one participant mentioned an example in which the HIV clinic faced a barrier to conducting research that other clinics in the hospital may not; this can affect the extent to which the clinic can engage in research. Finally, committee members reported approaching the process of research from a place of autodidactic curiosity and openness.

> *I have learned through seasoned employees that surveys and research have contributed to programs that exist today like DEI - Diversity Equity and Inclusion.*

CFIR findings both affirmed these results and provided deeper context to demonstrate some potential challenges and facilitators that might present as teams consider research co-production approaches in the setting. Potential facilitators emerging from interviews focused on the clinic being seen as a learning-centered space, where new practices are relatively easily adopted, with personnel consistently referencing the use of standardized and validated instruments, evidence-based protocols, and various forms of hospital- and clinic-level efforts to improve outcomes via data-driven methods. While the RQ+4 Co-Pro items did not elicit barriers to the prioritization of evidence-based practices, CFIR themes indicated that while there are areas in which patient care services could be improved, some more tenured staff were perceived as resistant to innovations, and the perceived disconnect between the larger

XSL•FO
RenderX

institution and HIV clinic could constrain collaborative research production.

The contextual factor Research Environment (eg, level for which co-production is seen as effective and legitimate) was on a 4-point rubric (4=Empowering, 3=Supportive, 2=Unsupportive, and 1=Restrictive). The mean response was 3.1 (SD 0.31), indicating a supportive environment for research being cogenerated by personnel and patients. Qualitative responses indicated themes of examples of collaborative research in the setting, perceived value of collaborative approach, and the extent to which collaborative research approaches are used as the status quo (Multimedia Appendix 4). Findings suggested that research is done through a team approach. Members reported having collectively participated in many quality improvement projects to improve patient outcomes, in which all are treated as equal members of research teams; there are some leaders who encourage personnel to participate in conducting research, with the hopes of using results to reduce health disparities. The general belief expressed was that the collaborative approach aligned with the clinic and institutional culture. Other members felt that personnel had competing priorities, which slowed down research processes. Finally, one committee member indicated that they had no knowledge of co-production occurring in the environment.

> *We all are able to propose potential changes in our practice and back it up with evidence.*

Some of the facilitators identified during interviews appeared to support findings from RQ+4 Co-Pro. Specifically, personnel indicated a shared value for seeking and integrating patient feedback to improve care, and evaluation approaches were discussed as embedded in clinic policies, procedures, and practices in which end users are perceived to contribute equally to research teams and actively contribute to change efforts. Conversely, potential barriers centered around perceived silos between positions and some staff feeling their skills are underused relative to research activities. Overall, personnel indicated novel approaches should be anchored into standard operating procedures to be sustainable.

Finally, the contextual factor Capacities for Co-production (eg, the extent to which research beneficiaries are trained on and have resources for co-production) was also assessed on a 4-point rubric (4=Strong focus, 3=Significant focus, 2=Minimal focus, and 1=No focus). The mean response was 3.1 (SD 0.22), indicating there was significant capacity for research co-production. Qualitative responses indicated themes included training and capacity-building, resources, and support (Multimedia Appendix 5). Responses suggested a range of perceptions—that the clinic was more engaged in co-production than the hospital overall, to the inclusion of an example project in which the hospital had applied similar methods to develop the institution's DEI initiative. Some members warned there would be a need for more time and resources to make co-production a standard approach, including patient care coverage. Other members (including some in nonresearch roles) reported they were learning by doing, and "a clear plan is not in place but in progress."

Interviews indicated that many aspects of the environment are currently facilitative of coproduced TIC, including team integration, camaraderie, and the stated prioritization of TIC, underscoring the potential collaborative participation in germane research activities. Potential barriers included a perceived lack of formal training in research activities for patients and personnel, coupled with the perception that there are fewer resources currently available for the HIV clinic, which could limit personnel capacity for focusing on efforts outside of clinical care.

Discussion

Principal Findings

Clinic culture presented as highly aligned with TIC values, with personnel expressing unanimous support for its adoption in the clinic. Although leaders expressed unequivocal support, some staff felt this support needed to be demonstrated through a greater focus on personnel wellness. Moreover, the potential for co-production also appeared strong but will require formal training and resource allocation to bring to action. We draw comparisons with other studies below for contextualization. Overall, findings from this study have clinical relevance for future implementation in the clinic but also progress knowledge in several meaningful ways. Generally, results contribute to the small but growing body of research on TIC with youth with HIV, as well as the potential for co-production as a useful strategy for TIC implementation.

Our findings reflect much of what has been observed across various settings, with studies consistently finding strong personnel support for TIC implementation and excitement for expanding trauma assessment and treatment protocols; this has been observed in other national pediatric care settings [39], HIV care organizations, and general health care spaces [15,33,40,41]. The current clinic's equity-focused culture (ie, compatible with TIC) is consistent with a similar study from a Southern HIV clinic, with both indicating that despite a positive and compatible culture, TIC implementation will likely require capacity-building (eg, trainings and procedural integration into workflow) [24]. Findings from other youth-focused outpatient counseling centers echo this need, adding that trainings must be supplemented with attention to system-wide efforts (eg, interorganizational networks, staffing, and fidelity monitoring), and highlighting the shared need to advance systematic service user feedback and mobilize change-resistant personnel [40]. Concern for sustainability also appears to be shared across settings [15,40,41], with a need for interagency collaborations, support from leadership, and policy changes identified as key indicators of sustainability across youth-focused care settings [40]. Finally, our findings align with those of a large study across various pediatric mental health settings, indicating that internal champions are critical for supporting implementation of expanded trauma assessment and treatment protocols; however, this study additionally found that external change agents were also critical [39]. While the role of external change agents was not a theme in this study, this study is nested within a larger implementation project that relies on external change agents.

XSL•FO
RenderX

Thus, clinic personnel could benefit from hearing how external change agents may act as facilitators.

Several barriers to TIC appear to be unique to this study, when compared with a systematic review of TIC across health care settings, in HIV clinics, and youth-focused organizations. Some of these included the perception that patients receive differing care or are stigmatized outside of the clinic, the lack of peer-to-peer support or platforms for soliciting or integrating patient preferences for care systematically, and the presence of repetitive suicide assessments, which could contribute to the likelihood of retraumatization. While we observed that more tenured staff were more likely to discuss tangible resource barriers, it was not discernible if these individuals had a more realistic view of challenges related to implementing innovations or if these perceived barriers could have been signs of perceived change resistance.

Regardless, most perceived barriers are consistent with other TIC literature, including among youth and in HIV care spaces. Some of these shared needs include enhancements in workforce development [17,40,42], staff wellness initiatives [17,40], trauma screening, assessment, referrals, and intervention [43], and the urgency for antiracism training for personnel [44]. McGeown et al [15] similarly found a need to direct attention to power dynamics as part of collective decision-making processes, and that more research skills, time, and resources are needed to implement TIC. Other collaborative trauma-informed research indicates that there are inherent power differentials in participatory research; academicians should stay cognizant to avoid having their own perspectives become the default and should solicit all members' perspectives through formal feedback loops [45].

The general cohesiveness among clinic staff aligns with the TIC principle of Collaboration and differs from studies in which TIC has been scaled across macrolevel settings (eg, entire cities) [17,34]. However, cohesiveness may wane across larger spaces, as seen with the perceived disconnect between the HIV clinic and the wider institution. Methods for integrating initiatives are a focal point for future TIC research in this setting to ensure policy changes are most likely to be sustained. The hospital-wide DEI initiative presents as a prime area for initiating multilevel efforts to be able to provide all patient-facing staff with culturally relevant TIC trainings.

To our knowledge, findings mark the first depiction of co-production as an implementation strategy for trauma-informed HIV care. While others have suggested co-production may be applied to TIC implementation [15,46,47], no literature showcases this implementation strategy with TIC in the United States or with LGBTQI-specific populations, men, or persons with HIV [15,40]. Results suggest many aspects of the environment would be facilitative of collaborative TIC (eg, cohesive or welcoming culture, trialability, human-equity focus, positive learning environment, and prioritization of TIC). Identified needs for consideration in future interventions include enhanced engagement among hospital leaders to ensure all personnel have access to resources for and protected time to engage in TIC and research activities through policy changes for effort distribution and mandating training for all patient-facing staff. Clinic leadership could also support the development of more peer-to-peer patient support opportunities by collaborating with psychosocial providers and increasing attention to change-resistant staff during adoption through individual discussions.

The existence of patient- and community-advisory boards demonstrates there is a path for engaging end users for feedback, but personnel perceive that the activities of these committees could be improved by including more current patients. Efforts to engage active patients in current CABs are recommended. Findings indicated that the environment likely would be facilitative of research co-production; however, the inclusion of personnel outside of the HIV clinic (eg, transportation representatives and security guards) in initiatives may present challenges (ie, training nonclinic personnel in a trauma-informed approach). Engagement with leadership or management of personnel outside of the HIV clinic is recommended to create buy-in among direct reports and support intervention participation for all patient-facing staff in the hospital environment.

Responses suggested that the hospital and clinic environments, in which staff are empowered to participate in research, spur personnel to seek out innovations. Specifically, the team evaluation approach appears facilitative of co-production. However, competing priorities for nonresearch personnel present challenges to the collaborative approach, with an identified need to develop a clear path and capacities for co-production to succeed. Generally, results suggested that committee members supported the approach for TIC research, with an emphasized need to ensure time and resources are reserved for personnel engagement in activities.

Limitations and Future Research

The current findings have several notable limitations. Application of purposive sampling may have introduced bias, as personnel most supportive of the study may have more readily volunteered to participate. Hence, findings may not be representative of all clinic personnel and cannot be generalized to represent personnel from HIV clinics outside of the current sample. Moreover, while data saturation was reached, the sample size for the RQ+4 Co-Pro component was small, precluding stratified results by demographics. Several of the RQ+4 Co-Pro items did not reap strong results, indicating full saturation may not have been reached. Patient-level data have not been presented here but will be presented separately in a follow-up report. Given that some of the RQ+4 Co-Pro items were not fully assessed, follow-up quantitative surveys should be conducted with a greater sample size of clinic and hospital personnel, which will serve to further validate qualitative results. While this study presents preliminary findings of our process to begin synthesizing a co-production framework (RQ+4 Co-Pro) with preimplementation stage research on TIC implementation (using CFIR), future research is needed to build on these methodological ambitions to further develop an integrated model. Such an effort will require a larger case study across multiple sites to extrapolate contextual factors enabling and constraining TIC research co-production and apply the

XSL•FO
RenderX

Classical Test Theory (including principal components analysis) to test and refine psychometric properties of such a model.

Conclusions

We conducted a mixed methods formative study to assess potential implementation determinants of coproduced TIC in a pediatric, youth, and young adult–focused HIV clinic in Memphis, Tennessee. Findings suggested that clinic personnel unanimously support TIC and there is an urgency for its implementation, but clinic and hospital leadership were perceived as potential barriers without more explicitly prioritizing staff wellness. Additionally, some staff shared concerns about change-resistant attitudes among some personnel, stating there would be a need for more peer-to-peer support opportunities for patients and greater efforts to anchor adoption into the larger hospital procedures, and strong support from multidisciplinary clinic leaders is perceived to be a vital factor in implementation. The clinic environment was also seen as an ideal space to conduct research, with a generally empowering milieu for co-production, but also, there was a stated need for a clear path to be formalized so that personnel have protected time for engagement and personnel wellness. Our results lead us to believe that TIC is likely to be an embraced evidence-based approach in the clinic, and co-production will likely be seen as a useful and fitting research strategy for TIC adoption among personnel of the sampled HIV clinic. However, follow-up research is needed to focus on integrating patient experiences and further test co-production (eg, for acceptability, appropriateness, and feasibility) as an implementation strategy for understanding and leveraging local contexts during TIC implementation. We aim to use results from this study to develop a site-tailored TIC implementation plan via research co-production so that personnel from the clinic remain equal contributors throughout the research process. Overall, both TIC and research co-production remain underused approaches in HIV care organizations, yet present as critical means for advancing national goals to end the HIV epidemic via collaborative, community-led solutions most likely to reduce health disparities. Findings from this study present potential implications for the multiple intersecting arenas of implementation science, adolescent health, and trauma research, all of which could be advanced from more greatly integrating either TIC or co-production into HIV care.

Acknowledgments

The authors would like to thank all the personnel and steering committee members from the participating institution; LLB acknowledges salary support from K01MH131471-01A1; Funding was provided by the Tennessee Center for AIDS Research (P30AI110527). No generative artificial intelligence was used to develop any portion of this manuscript.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

LLB, MW, ACP, and LCP conceived of the qualitative portion of the study and sought and received funding to support it. LLB and RKDM conceived the additional survey and mixed method approach. LLB wrote the first draft. LLB, MW, RKDM, JLS, NP, AJA, CMA, JMS, ACP, and LCP revised the manuscript. All authors approved of the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Co-production engagement in all phases of research.
[DOCX File, 15 KB - formative_v9i1e66426_app1.docx]

Multimedia Appendix 2
Personnel-perceived barriers and facilitators to implementation of trauma-informed care in an adolescent-focused pediatric HIV clinic in Memphis, Tennessee
[DOCX File, 25 KB - formative_v9i1e66426_app2.docx]

Multimedia Appendix 3
Knowledge use environment (eg, application of research).
[DOCX File, 19 KB - formative_v9i1e66426_app3.docx]

Multimedia Appendix 4
Research environment (eg, value and actual use of collaborative approach).
[DOCX File, 18 KB - formative_v9i1e66426_app4.docx]

XSL•FO
RenderX

Multimedia Appendix 5
Capacities for co-production (support for collaborative research on topic or in general).
[DOCX File, 18 KB - formative_v9i1e66426_app5.docx]

References

1. HIV in the Southern United States. Centers for Disease Control and Prevention. 2016. URL: https://stacks.cdc.gov/view/cdc/92279 [accessed 2025-07-30]
2. Harnessing the power of community engagement and innovation to end the HIV epidemic. 2023 ryan white HIV/AIDS program highlights. Rockville, Maryland. : U.S. Department of Health and Human Services; 2023.
3. HIV surveillance report 2020. Centers for Disease Control and Prevention. 2022. URL: https://www.cdc.gov/nchhstp/director-letters/2020-hiv-surveillance-report.html
4. Shelby county emphasizing HIV and syphilis testing amid 'alarming' numbers. Action News 5. 2024. URL: https://www.actionnews5.com/2024/06/02/shelby-county-emphasizing-hiv-syphilis-testing-amid-alarming-numbers/ [accessed 2025-07-30]
5. Brown LL, McLean RDK, Osman A, et al. Trauma-resilient HIV care and capacity for community-led research [Abstract]. Presented at: The 16th AIDSImpact Conference; May 26-28, 2025; Morocco.
6. Ingerski LM, Means B, Wang F, et al. Preventing medication nonadherence of youth (13-24 years) with HIV initiating antiretroviral therapy. J Adolesc Health 2021 Oct;69(4):644-652. [doi: 10.1016/j.jadohealth.2021.04.006] [Medline: 34059425]
7. Tarantino N, Brown LK, Whiteley L, et al. Correlates of missed clinic visits among youth living with HIV. AIDS Care 2018 Aug;30(8):982-989. [doi: 10.1080/09540121.2018.1437252] [Medline: 29455553]
8. Friedman MR, Stall R, Silvestre AJ, et al. Effects of syndemics on HIV viral load and medication adherence in the multicentre AIDS cohort study. AIDS 2015 Jun 1;29(9):1087-1096. [doi: 10.1097/QAD.0000000000000657] [Medline: 25870981]
9. Sales JM, Swartzendruber A, Phillips AL. Trauma-informed HIV prevention and treatment. Curr HIV/AIDS Rep 2016 Dec;13(6):374-382. [doi: 10.1007/s11904-016-0337-5] [Medline: 27704251]
10. Vanderbilt child health poll (2020-2023). Vanderbilt University Medical Center. 2023. URL: https://pediatrics.vumc.org/child-health-poll [accessed 2025-07-30]
11. Henderson SC, Wu F, Spears M. Child and adolescent mental health in Tennessee. : The Sycamore Institute; 2023 URL: https://sycamoretn.org/wp-content/uploads/2024/09/2023.09.08-final-child-and-adolescent-mental-health-in-tennessee.pdf [accessed 2025-08-08]
12. End HIV 901: ending the HIV epidemic Shelby county. : End HIV 901; 2020 URL: https://endhiv901.org/wp-content/uploads/2020/12/End-HIV-901-Plan-Presentation-Final-for-web-12-30.pdf [accessed 2025-08-08]
13. HRSA HIV/AIDS bureau division of metropolitan HIV/AIDS programs. HRSA-22-018 (H89) Ryan White HIV/AIDS Program Part A HIV Emergency Relief Grant Program. : US DHHS; 2021.
14. Ayala G, Sprague L, van der Merwe LLA, et al. Peer- and community-led responses to HIV: a scoping review. PLoS ONE 2021;16(12):e0260555. [doi: 10.1371/journal.pone.0260555] [Medline: 34852001]
15. McGeown H, Potter L, Stone T, et al. Trauma-informed co-production: collaborating and combining expertise to improve access to primary care with women with complex needs. Health Expect 2023 Oct;26(5):1895-1914. [doi: 10.1111/hex.13795] [Medline: 37430474]
16. Center for Substance Abuse Treatment (US). Trauma-Informed Care in Behavioral Health Services: Rockville (MD): Substance Abuse and Mental Health Services Administration (US); 2014. URL: https://www.ncbi.nlm.nih.gov/books/NBK207201/ [accessed 2025-08-08]
17. Damian AJ, Gallo J, Leaf P, Mendelson T. Organizational and provider level factors in implementation of trauma-informed care after a city-wide training: an explanatory mixed methods assessment. BMC Health Serv Res 2017 Nov 21;17(1):750. [doi: 10.1186/s12913-017-2695-0] [Medline: 29157241]
18. Bendall S, Eastwood O, Cox G, et al. A systematic review and synthesis of trauma-informed care within outpatient and counseling health settings for young people. Child Maltreat 2021 Aug;26(3):313-324. [doi: 10.1177/1077559520927468] [Medline: 32536207]
19. Sales JM, Piper K, Riddick C, Getachew B, Colasanti J, Kalokhe A. Low provider and staff self-care in a large safety-net HIV clinic in the Southern United States: implications for the adoption of trauma-informed care. SAGE Open Med 2019 Jan;7:2050312119871417. [doi: 10.1177/2050312119871417]
20. Brown LL, Perkins JM, Shepherd BE, et al. Piloting safety and stabilization: a multi-component trauma intervention to improve HIV viral suppression, retention in care, and post-traumatic stress disorder in a Southern United States HIV service organization. AIDS Behav 2024 Jan;28(1):174-185. [doi: 10.1007/s10461-023-04174-1] [Medline: 37751108]
21. Allison SM, Parker KL, Senn TE. Incorporating a trauma-informed perspective in HIV-related research with transgender and gender diverse individuals. J Int AIDS Soc 2022 Oct;25 Suppl 5(Suppl 5):e25976. [doi: 10.1002/jia2.25976] [Medline: 36225137]
22. Goldhammer H, Marc LG, Chavis NS, et al. Interventions for addressing trauma among people with HIV: a narrative review. AIDS Care 2022 Apr;34(4):505-514. [doi: 10.1080/09540121.2021.1984382] [Medline: 34612097]

XSL•FO
RenderX

23. McLean RKD, Carden F, Graham ID, et al. Evaluating research co-production: protocol for the Research Quality Plus for Co-Production (RQ+ 4 Co-Pro) framework. Implement Sci Commun 2022 Mar 14;3(1):28. [doi: 10.1186/s43058-022-00265-7] [Medline: 35287758]
24. Chambers R. Inclusive rigour for complexity. Journal of Development Effectiveness 2015 Jul 3;7(3):327-335. [doi: 10.1080/19439342.2015.1068356]
25. Pichon LC, Teti M, Brown LL. Triggers or prompts? When methods resurface unsafe memories and the value of trauma-informed photovoice research practices. Int J Qual Methods 2022 Apr;21:1-7. [doi: 10.1177/16094069221113979]
26. Kelly LM, Cordeiro M. Three principles of pragmatism for research on organizational processes. Methodological Innovations 2020 May;13(2). [doi: 10.1177/2059799120937242]
27. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005 Nov;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
28. Schwandt T. Constructivist, interpretivist approaches to human inquiry. In: Handbook of Qualitative Research: Thousand Oaks: Sage; 1994.
29. McLean RKD, Graham ID, Carden F. Evaluating research coproduction. In: Graham ID, Rycroft-Malone J, Kothari A, Mccutcheon C, editors. Research Coproduction in Healthcare: Hoboken: Wiley; 2022.
30. Wasti SP, Simkhada P, van Teijlingen ER, Sathian B, Banerjee I. The growing importance of mixed-methods research in health. Nepal J Epidemiol 2022 Mar;12(1):1175-1178. [doi: 10.3126/nje.v12i1.43633] [Medline: 35528457]
31. Ochieng BMN, Meetoo D. Using mixed methods when researching communities. Nurse Res 2015 Sep;23(1):16-19. [doi: 10.7748/nr.23.1.16.e1323] [Medline: 26365071]
32. Brown LL, Wilkins ML, Pichon LC, et al. Process mapping to inform implementation of trauma-informed care for youth aged 14-24 with HIV in the Southern United States. Res Sq. 2023 Aug 14 p. rs.3.rs-3234952. [doi: 10.21203/rs.3.rs-3234952/v1] [Medline: 37645914]
33. Piper KN, Brown LL, Tamler I, Kalokhe AS, Sales JM. Application of the Consolidated Framework for Implementation Research to facilitate delivery of trauma-informed HIV care. Ethn Dis 2021;31(1):109-118. [doi: 10.18865/ed.31.1.109] [Medline: 33519161]
34. Brown L, Perkins J, Acuña J, et al. Community-based participatory research to guide adoption of culturally responsive trauma-informed HIV care throughout Nashville, Tennessee. Research Square. Preprint posted online on Dec 15, 2023. [doi: 10.21203/rs.3.rs-3739954/v1] [Medline: 38168447]
35. McLean RKD, Carden F, Aiken AB, et al. Evaluating the quality of research co-production: Research Quality Plus for Co-Production (RQ + 4 Co-Pro). Health Res Policy Sys 2023;21(1). [doi: 10.1186/s12961-023-00990-y]
36. Damschroder LJ, Reardon CM, Widerquist MAO, Lowery J. The updated Consolidated Framework for Implementation Research based on user feedback. Implement Sci 2022;17(1). [doi: 10.1186/s13012-022-01245-0]
37. McLean R, Carden F, Aiken A, et al. The research quality plus for co-production (RQ+ 4 co-pro) assessment instrument. : Integrated Knowledge Translation Research Network. Ottawa, Canada; 2024 URL: https://idl-bnc-idrc.dspacedirect.org/server/api/core/bitstreams/5cc7cef8-ced5-4781-add5-80dfead504bb/content [accessed 2025-08-14]
38. Culyer AJ, Lomas J. Deliberative processes and evidence-informed decision making in healthcare: do they work and how might we know? Evid Policy 2006;2(3):357-371. [doi: 10.1332/174426406778023658]
39. Purbeck CA, Briggs EC, Tunno AM, Richardson LM, Pynoos RS, Fairbank JA. Trauma-informed measurement-based care for children: implementation in diverse treatment settings. Psychol Serv 2020 Aug;17(3):311-322. [doi: 10.1037/ser0000383] [Medline: 31403809]
40. Huo Y, Couzner L, Windsor T, Laver K, Dissanayaka NN, Cations M. Barriers and enablers for the implementation of trauma-informed care in healthcare settings: a systematic review. Implement Sci Commun 2023 May 5;4(1):49. [doi: 10.1186/s43058-023-00428-0] [Medline: 37147695]
41. As K, Adam E, Livingston M III, Root C, Sales J. Support for trauma-informed care implementation among Ryan White HIV clinics in the Southeastern United States. AIDS Behav 2023 Mar;27(3):939-947. [doi: 10.1007/s10461-022-03830-2]
42. Piper KN, Schiff J, Riddick C, et al. Barriers and facilitators to implementation of trauma support services at a large HIV treatment center in the Southern United States. AIDS Care 2021 Mar;33(3):316-325. [doi: 10.1080/09540121.2020.1764473] [Medline: 32423225]
43. Ali S, Stanton M, McCormick K, Reif S. The landscape of trauma informed care in community-based HIV service organizations in the United States south. AIDS Care 2023 Feb;35(2):244-248. [doi: 10.1080/09540121.2022.2079599] [Medline: 35637568]
44. Piper KN, Elder A, Renfro T, Iwan A, Ramirez M, Woods-Jaeger B. The importance of anti-racism in trauma-informed family engagement. Adm Policy Ment Health 2022 Jan;49(1):125-138. [doi: 10.1007/s10488-021-01147-1] [Medline: 34195916]
45. Jumarali SN, Nnawulezi N, Royson S, Lippy C, Rivera AN, Toopet T. Participatory research engagement of vulnerable populations: employing survivor-centered, trauma-informed approaches. Journal of Participatory Research Methods 2021;2(2). [doi: 10.35844/001c.24414]

XSL•FO
RenderX

46. Edelman NL. Trauma and resilience informed research principles and practice: a framework to improve the inclusion and experience of disadvantaged populations in health and social care research. J Health Serv Res Policy 2023 Jan;28(1):66-75. [doi: 10.1177/13558196221124740] [Medline: 36326603]
47. Stokes Y, Cloutier P, Aggarwal D, et al. Youth, caregiver and healthcare professional perspectives on planning the implementation of a trauma-informed care programme: a qualitative study. J Adv Nurs 2024 Mar 7:38450840. [doi: 10.1111/jan.16095] [Medline: 38450840]

Abbreviations

CAB: community advisory board
CFIR: Consolidated Framework for Implementation Research
DEI: diversity, equity, and inclusion
LGBTQI: Lesbian, Gay, Bisexual, Transgender, Queer, Intersex
RQ+4 Co-Pro: Research Quality Plus for Co-Production
TIC: trauma-informed care

Edited by A Mavragani; submitted 12.09.24; peer-reviewed by R Nooripour; revised version received 23.05.25; accepted 25.05.25; published 21.08.25.

Please cite as:
Brown LL, Wilkins ML, McLean RKD, Stewart JL, Patel ND, Acuña AJ, Audet CM, Sales JM, Pettit AC, Pichon LC
Exploring Co-production as an Implementation Strategy for Trauma-Informed Care in a Youth-Focused HIV Clinic in Memphis, Tennessee: Mixed Methods Research
JMIR Form Res 2025;9:e66426
URL: https://formative.jmir.org/2025/1/e66426
doi:10.2196/66426

© Leslie Lauren Brown, Megan Wilkins, Robert Kenneth Douglas McLean, Jamie Lynn Stewart, Nehali Patel, Almariana Jessica Acuña, Carolyn Marie Audet, Jessica M Sales, April C Pettit, Latrice C Pichon. Originally published in JMIR Formative Research (https://formative.jmir.org), 21.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Food Handling Practices Among Food Businesses in Jigjiga, Eastern Ethiopia, During the COVID-19 Pandemic: Cross-Sectional Study

Alinoor Mohamed Farah[1], MPH; Tahir Yousuf Nour[1], MPH; Abdurahman Kedir Roble[2], MSc Nursing; Muse Obsiye[1], MPH; Mowlid Akil Aden[1], MPH; Abdulahi Siraj Abdi[2], MSc Nursing; Fentabil Getnet[1], PhD

[1]School of Public, College of Medicine and Health Sciences, Jigjiga University, CoMHS Building, 2nd Floor, Jigjiga, Ethiopia

[2]School of Nursing and Midwifery, College of Medicine and Health Sciences, Jigjiga University, Jigjiga, Ethiopia

Corresponding Author:
Alinoor Mohamed Farah, MPH
School of Public, College of Medicine and Health Sciences, Jigjiga University, CoMHS Building, 2nd Floor, Jigjiga, Ethiopia

Abstract

Background: The COVID-19 pandemic has posed significant challenges to food safety practices globally, profoundly affecting the knowledge, attitudes, and practices of both food handlers and consumers.

Objective: This study aimed to investigate food safety knowledge and practices of food handlers in the context of COVID-19.

Methods: A cross-sectional study was conducted in Jigjiga during the pandemic. A total of 384 food handlers were surveyed using a structured questionnaire and an observational checklist. The questionnaire assessed knowledge of COVID-19 symptoms, transmission, and prevention measures, and the checklist evaluated food safety practices and the implementation of COVID-19 prevention measures in food businesses. Categorical variables (eg, sufficient vs insufficient COVID-19 knowledge and good vs poor food-safety practice) were summarized as frequencies and percentages. Pearson chi - square test was used to assess differences in these binary outcomes across demographic and other categorical subgroups (eg, sex, age category, education level, and source of COVID-19 information). A *P* value <.05 was considered statistically significant.

Results: A total of 384 food handlers were approached, and all responded (response rate=100%). The majority of participants (276/384, 71.9%) had received food hygiene training, and the main source of COVID-19 information was government news media (170/384, 44.3%). The majority of respondents (264/384, 68.8%) correctly identified the key COVID-19 symptoms, and 52.1% (200/384) accurately understood that respiratory droplets from coughs or sneezes drive transmission. However, less than 50% of participants consistently practiced preventive measures such as avoiding handshaking, frequently sanitizing food contact surfaces, and reminding customers to follow physical distancing. Participants who obtained information from government sites and the media had sufficient knowledge compared to other participants (*P*=.07). Females (*P*=.03), younger adults (*P*=.03), married individuals (*P*=.04), those with secondary education (*P*=.014), and those who had received previous food safety training (*P*=.004) demonstrated better food handling practices than their counterparts. Furthermore, 61.4% (236/384) of food businesses had handwashing facilities at the entrance, 70.2% (270/384) implemented crowd control measures, and 56.1% (215/384) used floor markings to facilitate physical distancing. However, only 57.7% (221/384) of food establishments routinely cleaned and disinfected their work surfaces and touch points.

Conclusions: These findings highlight the need for targeted education and training interventions to improve food handlers' knowledge and practices, particularly during the COVID-19 pandemic.

(JMIR Form Res 2025;9:e64214) doi:10.2196/64214

KEYWORDS

COVID-19; knowledge; practice; Ethiopia; food safety; food; food supply; good hygiene practices; proper hygiene; hygiene; cross-sectional study; data collection; employee; food handler; food safety training; food handling

Introduction

The COVID-19 pandemic significantly impacted food safety practices and consumer behavior. Although there is no evidence of SARS-CoV-2 transmission through food [1,2], the virus can survive on surfaces, including food packaging [3]. Consequently, food handlers and consumers have adopted enhanced hygiene measures, such as increased handwashing, produce cleaning, and food thermometer use [4,5]. The food industry has implemented additional practices to protect workers, including social distancing, face coverings, and employee screening [6]. However, pandemic-related disruptions may increase food safety risks [6]. To mitigate these risks, experts recommend adhering

XSL•FO
RenderX

to established food safety protocols, proper sanitization, and maintaining good hygiene practices throughout the food supply chain [7]. Postpandemic, it is expected that there will be a greater understanding of the importance of food hygiene and creative adaptations in food service [2].

Food handlers in businesses demonstrated improved knowledge and attitudes toward food safety during the pandemic [8]. Although home-based online food businesses have expanded, food handlers have shown low knowledge and improper practices in some areas [9]. Age, income, education, and gender influence food safety practices [4]. Overall, the pandemic increased awareness of food safety but also led to some misconceptions and improper practices, highlighting the need for targeted educational programs [5,10].

Due to numerous factors, food safety is significantly more complicated in developing nations. Poverty is one of the leading causes of unsafe food consumption due to a lack of access to sufficient food and clean water, poor government structure, the persistence of infectious diseases in the community, inconvenient environmental conditions to ensure food safety, and poor food handling and sanitation practices [11,12]. Ethiopia is not exceptional because of the prevalence of poor food handling and sanitation practices, inadequate food safety laws, weak regulatory systems, lack of financial resources to invest in food safety, and lack of training and education for food handlers [12].

In this context, safeguarding the food supply chain during the COVID-19 pandemic requires not only industry-level interventions but also strong frontline food safety practices at every point of contact. In Ethiopia, where food handling often occurs under resource constraints and regulatory oversight may be limited, understanding how food handlers adapt and where gaps remain is critically important. Although several studies have documented shifts in knowledge and attitudes among food handlers, few have examined these dynamics in resource-constrained settings, where unique socioeconomic and infrastructural challenges may amplify both COVID-19–related and other food safety risks. Therefore, this study aimed to investigate the food safety knowledge and practices of food handlers in Jigjiga Town. The findings will inform targeted, context-appropriate training and policy measures to reinforce safe food handling practices both during and after the COVID-19 era.

Methods

Study Design, Setting, and Period

A cross-sectional study was conducted from August 12 to 23, 2020, to evaluate the food safety knowledge and practices of food handlers and assess the sanitary conditions of groceries in Jigjiga town. Jigjiga is the capital town of the Somali Regional State located in the eastern part of Ethiopia, and the city has an elevation of 1609 meters above sea level. Based on the 2007 Census conducted by the Central Statistical Agency of Ethiopia, Jijiga had a total population of 203,588, of whom 109,138 were men and 94,450 were women. The town is divided into 30 kebeles (the smallest administrative units), of which 20 are urban and 10 are rural ones.

Sample Size Determination

The sample size was determined using Epi Info version 3.5.1 software (Center for Disease Control and Prevention, Atlanta, 2004) by single population proportion formula with the assumptions of 95% confidence level and 5% precision. A sample size of 384 was estimated with the conservative assumption of a 50% prevalence of good food safety practices, with 95% certainty.

Sampling Technique

A master list of 519 food-vending outlets, including shops, grocery stores, supermarkets, and minimarkets, was obtained from the Jigjiga City Council Office. To capture unregistered businesses, we conducted a door-to-door census in each kebele, which identified an additional 201 outlets and increased the total sampling frame to 720 establishments. A proportional sample size was determined for each kebele, and groceries were randomly selected from each kebele using a random-number generator. A total of 384 grocery stores and shops were randomly selected for the study. The study targeted both food retail establishments (grocery stores, supermarkets, minimarkets, and food shops) and the individual food handlers working within them.

Inclusion and Exclusion Criteria

Food vending outlets, specifically shops, grocery stores, supermarkets, and mini-markets, regardless of whether they were listed or registered by the Jigjiga City Council office, were included.

Data Collection and Quality Control

In total, 4 sanitarians were recruited as data collectors and supervisors for the study. All team members participated in an orientation covering the study objectives, questionnaire components, and data quality management. We used a structured questionnaire that was pretested on 10% (38/384) of a similar study population, and expert feedback was incorporated into the final tool.

The questionnaire was drafted in English and translated into Somali using forward–backward translation to ensure semantic equivalence of the questions. It included items on common COVID-19 symptoms and food handling knowledge, adapted from the World Health Organization (WHO) and Food and Agriculture Organization (FAO) COVID-19 and food safety protocols [13]. Data collectors interviewed food handlers in stores, shops, and groceries and observed their on-the-job practices using a standardized checklist to assess prevention measures.

We evaluated the internal consistency of the tools using Cronbach α (α=0.702), indicating acceptable reliability. For observational data, we conducted interrater reliability exercises: data collectors applied the checklist side by side, discrepancies were discussed, and retraining was provided until the standardized criteria were consistently applied. The final questionnaire was refined to improve clarity, minimize bias, and align with the study objectives.

XSL•FO
RenderX

We took multiple precautions to ensure that the participants did not feel compelled to provide socially desirable responses. First, the questionnaire was carefully worded and structured so that the questions were neutral and nonleading. Before data collection began, the participants were assured that their answers would remain anonymous, used only in aggregate form, and would have no impact on their employment or licensing status. This reassurance helped reduce the fear of negative consequences for honest reporting. The interviewers received dedicated training in neutral probing techniques and were instructed to avoid any verbal or nonverbal cues that might suggest approval or disapproval of particular answers. Finally, the participants were reminded that their honest feedback was valuable and that they could pause or take brief breaks at any time, helping ensure that the respondents remained attentive and invested in providing thoughtful answers.

Methods of Data Analysis

Data were coded and entered into Epi Info version 3.5.1 software and exported into STATA version 14 for analysis. Knowledge and practice questions were scored as 1 or 0 for correct and incorrect responses, respectively. Mean scores were calculated and used as cut-off points to dichotomize the outcome variables. Scores below the mean score were considered as indicating inadequate knowledge or practice, whereas scores equal to or above the mean score were considered as indicating adequate knowledge and favorable adequate practice. Knowledge questions were scored as 1 or 0 for correct and incorrect responses, respectively. The total knowledge score ranged from 0 (no correct answers) to 6 (all correct answers); a mean score of ≤3.5 indicated poor knowledge, and a mean score of >3.5 indicated good knowledge. There were 14 practice questions scored on a 3-point scale from 1 (“always”) to 3 (“never”). Total practice scores ranged from 14 to 42; a mean score of >18.8 (answering “never” or “occasionally”) indicated poor practices, and a score of ≤18.8 (answering “always”) indicated good practices. Thus, the lower the practice scores, the higher the probability of good practices, and vice versa.

Summary statistics, such as frequencies and proportions, were computed as appropriate. Categorical variables (eg, sufficient vs insufficient COVID-19 knowledge and good vs poor food-safety practice) were summarized as frequencies and percentages. Pearson’s chi - square test was used to assess differences in these binary outcomes across demographic and other categorical subgroups (eg, sex, age category, education level, and source of COVID-19 information). *P* value <.05 was considered statistically significant.

Operation Definition

Food handling refers to one or more operations of food production, manufacture, offering or displaying for sale, storage, preservation, wrapping, transportation, delivery, importation, exportation, or the licensing or approval for any of such activities.

Food handling knowledge refers to a food handler’s accurate understanding of COVID-19–related health and safety information as it applies to one or more operations of food production, manufacture, sale, storage, preservation, wrapping, transportation, delivery, importation, exportation, or licensing or approval of such activities.

Food handling practices refer to the consistency with which food handlers perform COVID-19–related safety measures, such as personal protective equipment use, hand hygiene, respiratory etiquette, surface sanitation, and physical distancing, across all stages of food production, sale, storage, transport, or regulatory activities.

COVID-19 prevention measures refer to the extent to which a food retail establishment implements structural and managerial interventions designed to reduce viral transmission.

Food businesses or establishments refer to any commercial establishment engaged in the retail sale of edible products, including but not limited to supermarkets, grocery stores, and convenience shops.

Ethical Considerations

This study was conducted in accordance with the guidelines of the Declaration of Helsinki, and all procedures involving research participants were approved by the Jigjiga University institutional review board (RERC/039/2020). Written informed consent was obtained from all participants. The participants were not compensated for participating in the study. The study participants’ data were confidential and protected.

Results

Sociodemographic Characteristics

A total of 384 food handlers were approached, and all of them responded (response rate=100%). As shown in Table 1, of the total participants, 50.8% were male, with a mean age of 32.1 (SD 4.8) years. Out of 384, the majority of participants were under 40 years of age (n=369, 96.1%) and 108 (28.1%) respondents had college and university-level education, and 296 (77.1%) were married. The majority of the participants received food hygiene training (n=276, 71.9%), and 108 (28.1%) were working at grocery stores. The main source of information about COVID-19 among participants was government news media (n=170, 44.3%).

XSL•FO
RenderX

Table . Sociodemographic characteristics of food handlers in food businesses in Jigjiga, Eastern Ethiopia (2020).

Variable	Value
Sex, n (%)	
Male	195 (50.8)
Female	189 (49.2)
Age (years), mean (SD)	32.1 (4.8)
18 - 39	369 (96.1)
>40	15 (3.9)
Marital status, n (%)	
Married	296 (77.1)
Single	57 (14.8)
Divorced	24 (6.3)
Widowed	7 (1.8)
Education status, n (%)	
No education	98 (25.5)
Elementary	82 (21.4)
Secondary	96 (25.0)
College and university	108 (28.1)
Type of food business, n (%)	
Wholesale shops	104 (27.1)
Retail shops	95 (24.7)
Grocery store	108 (28.1)
Minimarket/supermarket	77 (20.1)
License status, n (%)	
Licensed	249 (64.8)
Not licensed	135 (35.2)
Food hygiene training, n (%)	
Yes	276 (71.9)
No	108 (28.1)
Source of information on COVID-19, n (%)	
Government news media	170 (44.3)
Social media	143 (37.2)
Other	71 (18.5)

Knowledge

As shown in Table 2, the majority of respondents (264/384, 68.8%) correctly identified the key COVID-19 symptoms. The majority of the respondents understood the dynamics of COVID-19 infectiousness: 63.8% (245/384) of respondents were aware of the possibility of infection before the onset of symptoms; 52.1% (200/384) of the participants responded true on droplets as a major transmission route. More than half of the respondents (215/384, 56.0%) understood the need for self-isolation, and 54.4% (209/384) were aware of the standard precautions that need to be followed when dealing with customers.

XSL•FO
RenderX

Table . Knowledge of food handlers in food businesses about COVID-19 transmission, prevention, and food safety practices in Jigjiga, Eastern Ethiopia (2020).

Questions	Response (n=384)		
Knowledge	True, n (%)	False, n (%)	I do not know, n (%)
The common symptoms of COVID-19 are fever, fatigue, and dry cough.	264 (68.8)	93 (24.2)	27 (7.0)
COVID-19 can be transmitted through direct contact of respiratory droplets when infected persons cough or sneeze.	200 (52.1)	156 (40.6)	28 (7.3)
The disease can be transmitted from asymptomatic patients.	245 (63.8)	110 (28.7)	29 (7.5)
The disease can be spread by touching a surface or object that has the virus on it and then touching their own mouth, nose, or possibly their eyes.	214 (55.7)	135 (35.2)	35 (9.1)
A person with mild symptoms of COVID-19 must remain at home until resolution of clinical symptoms.	215 (56.0)	126 (32.8)	43 (11.2)
Standard precautions such as wearing gloves and masks should be followed by food handlers when dealing with customers.	209 (54.4)	111 (28.9)	64 (16.7)

Practice

Table 3 explains the practical measures in place by the food handlers to protect themselves and their customers; 52.9% (203/384) of the respondents reported always wearing masks and gloves when dealing with the customers, and up to 51.3% (197/384) washed their hands when they were soiled. Other measures included maintaining social distance and avoiding shaking hands, hugging, or kissing. Unfortunately, less than 50% (192/384) of the participants had avoided shaking hands, touching their nose and mouth, frequently sanitized all food contact surfaces, and made regular announcements to remind their customers to follow physical distancing.

XSL•FO
RenderX

Table . Food safety and hygiene practices of food handlers in food businesses in Jigjiga, Eastern Ethiopia (2020).

Questions	Response (n=384)		
Practice	Always, n (%)	Occasionally, n (%)	Never, n (%)
I always wear masks and gloves when dealing with customers.	203 (52.9)	136 (35.4)	45 (11.7)
I wash my hands with water and soap before putting on the gloves.	176 (45.8)	191 (49.8)	17 (4.4)
I use sanitizer between customers (before and after handling food/packing food).	202 (52.6)	104 (27.1)	78 (20.3)
I wash my hands whenever they are soiled.	197 (51.3)	104 (27.1)	83 (21.6)
I wash my hands after blowing my nose or covering a sneeze.	188 (49.0)	113 (29.4)	83 (21.6)
When I'm sneezing or coughing, I cover my nose or mouth with a tissue/clean cloth or into my elbow.	146 (38.0)	122 (31.8)	116 (30.2)
I keep at least one-meter distance from my work colleagues and the customers.	217 (56.5)	96 (25.0)	71 (18.5)
I avoid shaking hands, hugging, or kissing with colleagues/customers.	184 (47.9)	127 (33.1)	73 (19.0)
I avoid touching my mouth, nose, and eyes with my hand if I have not washed them with water and soap.	183 (47.7)	123 (32.0)	78 (20.3)
I frequently wash and sanitize all food contact surfaces.	178 (46.4)	128 (33.3)	78 (20.3)
I make regular announcements to remind customers to follow physical distancing advice and clean their hands regularly.	142 (37.0)	121 (31.5)	121 (31.5)

COVID-19 Prevention Measures

As shown in Table 4, 61.4% (235/383) of food businesses had handwashing facilities at the entrance. Furthermore, 70.2% (269/384) of food businesses were able to control the crowd. More than half (215/383, 56.1%) of the food businesses used floor markings inside the retail store to facilitate compliance with physical distancing. Moreover, 74.7% (286/383) and 54.1% (207/383) of the food businesses had barriers at tills and counters introduced as an additional level of protection for food handlers and signs placed at entry points to request customers not to enter the shop if they were unwell or had COVID-19 symptoms, respectively. In addition, 57.7% (221/383) of food businesses frequently cleaned or disinfected work surfaces and touch points.

XSL•FO
RenderX

Table . COVID-19 prevention measures practiced by food handlers in food businesses in Jigjiga, Eastern Ethiopia (2020).

COVID-19 prevention measures	Response, (n=383)	
	Yes, n (%)	No, n (%)
Handwashing facility observed at the establishment.	235 (61.4)	148 (38.6)
Queue/crowd control managed.	269 (70.2)	114 (29.8)
Floor markings used inside the retail store to facilitate compliance with the physical distancing, particularly in the most crowded areas, such as serving counters and tills.	215 (56.1)	168 (43.9)
Barriers at tills and counters introduced as an additional level of protection for food handler.	286 (74.7)	97 (25.3)
Signs placed at entry points to request customers not to enter the shop if they are unwell or have COVID-19 symptoms.	207 (54.1)	176 (45.9)
Frequent cleaning/disinfection of work surfaces and touch points.	221 (57.7)	162 (42.3)

Comparison of COVID-19 Knowledge and Practice Between Different Groups

As shown in Table 5, less than half (191/384, 49.7%) of the respondents had sufficient knowledge. Of the total participants, more male participants had sufficient knowledge than female participants (n=100, 52.4% vs n=91, 47.6%), and the difference was not statistically significant (*P*=.54). More participants who obtained information from government sites and the media had sufficient knowledge than other participants (*P*=.07). Female (*P*=.03), younger (*P*=.06), and married respondents (*P*=.04) practiced more than their counterparts. Furthermore, respondents with secondary education (*P*=.01) and those who received food hygiene training (*P*=.004) had better practice than their counterparts.

XSL•FO
RenderX

Table . Comparison of COVID-19 knowledge and food safety practices among food handlers across sociodemographic groups in Jigjiga, Eastern Ethiopia (2020).

Variable	Sufficient knowledge (n=384)		Good practice (n=384)	
	Frequency, n (%)	*P* value[a]	Frequency, n (%)	*P* value[a]
Overall	191 (49.7)		233 (60.7)	
Sex		.54		.03[b]
Male	100 (52.4)		108 (46.4)	
Female	91 (47.6)		125 (53.6)	
Age (years)		.78		.006[b]
18 - 39	183 (95.8)		229 (98.3)	
>40	8 (4.2)		4 (1.7)	
Marital status		.52		.04[b]
Married	145 (75.9)		189 (81.1)	
Single	32 (16.7)		32 (13.7)	
Divorced	12 (6.4)		9 (3.9)	
Widowed	2 (1.0)		3 (1.3)	
Education status		.35		.01[b]
No education	51 (26.7)		60 (25.7)	
Elementary	37 (19.4)		38 (16.3)	
Secondary	43 (22.5)		67 (28.8)	
College and university	60 (31.4)		68 (29.2)	
Type of food business		.56		.35
Wholesale	52 (27.2)		61(26.2)	
Retail	47 (24.6)		53(22.8)	
Grocery	49 (25.7)		73(31.3)	
Minimarket/supermarket	43 (22.5)		46(19.7)	
Food hygiene training		.77		.004[b]
Yes	136/191 (71.2)		180 (77.3)	
No	55/191 (28.8)		53 (22.7)	
License status		.38		.37
Licensed	128 (67.0)		147 (63.1)	
Not licensed	63 (33.0)		86 (36.9)	
Source of information on COVID-19		.07		.52
Government site and media	94 (49.2)		98 (42.1)	
Social media	69 (36.1)		89 (38.2)	
Other	28 (14.7)		46 (19.7)	

[a]Pearson chi-squared test to compare knowledge, attitudes, and practices between different groups.

[b]Statistically significant at $P<.05$

XSL•FO
RenderX

Discussion

Principal Findings

This study examined food safety knowledge and practices among 384 food handlers in Jigjiga Town during the COVID-19 outbreak. More than half of the participants were not adequately informed about COVID-19. Higher levels of COVID-19 knowledge were observed among men, older individuals, and those with a college or university education. Studies conducted in different countries, including India, Jordan, Malaysia, and Bangladesh, have shown that food handlers generally possess fair to excellent knowledge, positive attitudes, and good practices regarding food safety during the pandemic [8,11,12,14]. However, some studies have identified areas for improvement, particularly in cross-contamination prevention and safe food handling practices [9].

The food safety practices of the female participants were better than those of the male participants, which is consistent with the findings of other studies [15,16]. This could be explained by the fact that females are more involved in-home food preparation, multitask more, and pay more attention to COVID-19. Similarly, younger respondents and married respondents practiced more than their counterparts. Our findings disagree with those of other studies that have shown insufficient practice among younger adults [17].

Participants with secondary education and those trained in food safety had better practice than their counterparts, which is in line with other similar studies conducted in Iran and Malaysia [14,16]. Attending food safety courses was also found to be a significant factor influencing food safety knowledge and practices among Malaysian food truck vendors [14]. A brief educational intervention study on restaurant food handlers' knowledge of food allergies conducted in Brighton in 2014 revealed improvements in absolute knowledge and practice change [18]. Food safety and hygiene interventions influenced the food safety knowledge, behavior, and habits of employees in the food and beverage departments of hotels and restaurants in Turkey's tourism industry [19]. A systematic review conducted in Canada found that routine intervention education of food service premises is effective in reducing the risk of foodborne illnesses and improving food handler knowledge and practices. Community-based education programs have been shown to increase public awareness of food safety [20].

Participants who obtained information from government websites and traditional media had higher levels of knowledge. In our study, most respondents relied on government media for information updates. Although government agencies, healthcare professionals, and scientists are generally trusted sources of food safety information, social media is considered less reliable [21]. However, the primary information source can significantly influence the link between COVID-19 knowledge and protective behaviors; interpersonal communication and social media often elicit stronger behavior change than official channels [22].

The food service industry was among the first frontline employment sectors to experience the effects of the COVID-19 pandemic. To date, no study has indicated that COVID-19 spreads through food. In addition, there is no evidence that viruses that infect the respiratory tract can be transmitted through food or food packaging [7]. SARS-CoV and Middle East Respiratory Syndrome Coronavirus do not appear to be transmitted through food consumption [7].

According to the FAO and World Food Program, touching food packages or containers contaminated with SARS-CoV-2 could transmit the virus to the mouth, nose, or eyes [7]. However, because the virus does not survive on these surfaces, this is not considered the primary route of disease transmission. A previous study found that food products can be used to spread respiratory viruses, such as SARS-CoV-1 and influenza [23]. Another study found that the risk of Ebola infection in individual humans in the United States was negligible to low as a result of contaminated cocoa beans, palm oil, or cashews imported from South Africa [24].

More work is needed to prevent SARS-CoV-2 transmission from the respiratory tract to food packaging surfaces or through food consumption. The Food and Drug Administration has proposed guidelines for consumers to follow when shopping for, handling, and preparing food. To avoid SARS-CoV-2 transmission and comply with the food safety system's minimum requirements, food handlers, including food establishment employees and consumers, should follow good sanitation and hygiene practice guidelines [25].

Ensuring smooth food movement through the supply chain requires vigilance and cooperation from all involved. It is essential to maintain consumer confidence in both food safety and availability by providing a secure environment at every stage, especially in retail shops and canteens, where both consumers and retailers must uphold rigorous hygiene and safety standards [7]. Retailers can help by providing sanitary facilities, such as wipes, disinfectants, and sanitizers, as well as displaying sanitary practices with visual aids. Physical separation can be maintained by marking the floors as a reference for the minimum distance. To avoid contact at cashier counters, plexiglass can be installed, and food tasting for promotional campaigns should be avoided during the pandemic [7].

Consumers should ensure that family members in vulnerable groups (immune-compromised, elderly, children, and COVID-19 patients) remain at home [26]. Furthermore, wearing masks, gloves, using hand sanitizers and wipes before handling food carts, avoiding reusable shopping bags, and practicing acceptable respiratory etiquette should be prioritized. If reusable bags are deemed regionally acceptable, they should be disinfected immediately after use [27].

This study had some limitations. First, its cross-sectional design precludes any causal inferences regarding the relationship between food handlers' COVID-19 knowledge and their practices. Because the data were collected at a single point in time, we cannot determine whether greater knowledge drives better practices or vice versa, nor can we track how these measures evolve alongside changing public health guidelines. Second, relying on self-reported behaviors carries the risk of social desirability bias, as respondents may overstate their adherence to mask-wearing, handwashing, and other preventive measures. Although we complemented interviews with direct

XSL•FO
RenderX

observation, the mere presence of data collectors may have temporarily improved performance (the Hawthorne effect), meaning that observed practices might not fully reflect routine behavior. Finally, by recruiting food stores and groceries, our sample may not represent larger or more varied food service settings, such as restaurants and industrial processors, where knowledge and practice patterns could differ substantially.

Conclusion

In conclusion, less than half (191/384, 49.7%) of food handlers had sufficient knowledge about COVID-19. Our findings also showed that food handlers with previous food safety training practiced more than their counterparts did. As a result, researchers, food safety communicators, the media, and all other related sectors should work to educate food handlers to advance their health and food safety training. Food handler training courses should be prioritized by public health officials, especially during the COVID-19 pandemic.

Acknowledgments

We would like to extend our special thanks to Jigjiga University for funding research on COVID-19, which may be used in planning health programs for emerging viral diseases. All authors declared that they had insufficient funding to support open access publication of this manuscript, including from affiliated organizations or institutions, funding agencies, or other organizations. JMIR Publications provided article processing fee (APF) support for the publication of this article. This research was funded by Jigjiga University.

Data Availability

The dataset for this study is available from the corresponding author on reasonable request.

Authors' Contributions

AMF conceived the study, prepared the proposal, analyzed the data, interpreted the findings, and wrote the manuscript. TYN, MO, AKR, MAA, ASA, and FG were involved in developing the study proposal, data analysis, and reviewing the manuscript.

Conflicts of Interest

None declared.

References

1. Desai AN, Aronoff DM. Food safety and COVID-19. JAMA 2020 May 19;323(19):1982. [doi: 10.1001/jama.2020.5877] [Medline: 32271867]
2. Maragoni-Santos C, Serrano Pinheiro de Souza T, Matheus JRV, et al. COVID-19 pandemic sheds light on the importance of food safety practices: risks, global recommendations, and perspectives. Crit Rev Food Sci Nutr 2022;62(20):5569-5581. [doi: 10.1080/10408398.2021.1887078] [Medline: 33591233]
3. Makroo HA, Majid D, Siddiqi MA, Greiner R, Dar BN. COVID-19 pandemic and its implications on food systems. Preprints. Preprint posted online on Aug 13, 2020. [doi: 10.20944/preprints202008.0321.v1]
4. Thomas MS, Feng Y. Food handling practices in the era of covid-19: a mixed-method longitudinal needs assessment of consumers in the United States. J Food Prot 2021 Jul 1;84(7):1176-1187. [doi: 10.4315/JFP-21-006] [Medline: 33666666]
5. Goswami RG. Knowledge on safe handling of food during covid-19 pandemic: a questionnaire based survey. Int J Res Rev 2020;7(5):103-109 [FREE Full text]
6. Trmčić A, Demmings E, Kniel K, Wiedmann M, Alcaine S. Food safety and employee health implications of COVID-19: a review. J Food Prot 2021 Nov 1;84(11):1973-1989. [doi: 10.4315/JFP-21-201] [Medline: 34265068]
7. Olaimat AN, Shahbaz HM, Fatima N, Munir S, Holley RA. Food safety during and after the era of COVID-19 pandemic. Front Microbiol 2020;11:1854. [doi: 10.3389/fmicb.2020.01854] [Medline: 32849446]
8. Omar S. Hygienic food practice among food workers in Jordanian food factories during pandemic crisis: COVID-19: comparative study. Research Square. Preprint posted online on May 29, 2020. [doi: 10.21203/rs.3.rs-29784/v1]
9. Osaili TM, Al-Nabulsi AA, Al-Jaberi TM. Food safety knowledge, attitudes, and practices among female food handlers engaged in home-based online food businesses in Jordan during the COVID-19 pandemic. Heliyon 2022 Aug;8(8):e10427. [doi: 10.1016/j.heliyon.2022.e10427] [Medline: 36065214]
10. Limon MR. Food safety practices of food handlers at home engaged in online food businesses during COVID-19 pandemic in the Philippines. Curr Res Food Sci 2021;4:63-73. [doi: 10.1016/j.crfs.2021.01.001] [Medline: 33665620]
11. Kannu MJS, Evangeline MJV, Karthick K. Impact of Covid19: knowledge, attitude and practice (KAP) in food industry, India. IJTSRD 2020;4(4):1666-1669 [FREE Full text]
12. Noor N, Zaman LK, Yaacob N, Hassan MS, Khadzil S. The awareness of food safety knowledge, attitudes and practices among food handlers during COVID-19. Food Science and Technology 2023;11(2):125-132. [doi: 10.13189/fst.2023.110203]

XSL•FO
RenderX

13. COVID-19 and food safety: guidance for food businesses. : World Health Organization and Food and Agriculture Organization of the United Nations; 2020.
14. Wan Nawawi WNF, Ramoo V, Chong MC, Zaini NH, Chui PL, Abdul Mulud Z. The food safety knowledge, attitude and practice of Malaysian food truck vendors during the COVID-19 pandemic. Healthcare (Basel) 2022;10(6):998. [doi: 10.3390/healthcare10060998]
15. Haapala I, Probart C. Food safety knowledge, perceptions, and behaviors among middle school students. J Nutr Educ Behav 2004;36(2):71-76. [doi: 10.1016/s1499-4046(06)60136-x] [Medline: 15068755]
16. Mohammadi‐Nasrabadi F, Salmani Y, Esfarjani F. A quasi‐experimental study on the effect of health and food safety training intervention on restaurant food handlers during the COVID‐19 pandemic. Food Sci Nutr 2021 Jul;9(7):3655-3663. [doi: 10.1002/fsn3.2326]
17. sanlier N. The knowledge and practice of food safety by young and adult consumers. Food Control 2009 Jun;20(6):538-542. [doi: 10.1016/j.foodcont.2008.08.006]
18. Bailey S, Billmeier Kindratt T, Smith H, Reading D. Food allergy training event for restaurant staff; a pilot evaluation. Clin Transl Allergy 2014;4(1):26. [doi: 10.1186/2045-7022-4-26] [Medline: 25225607]
19. Sanli˙er N, Sormaz Ü, Güneş E. The effect of food safety education on food safety knowledge, attitudes, behaviors of individuals who work in food and beverage departments in Turkey. International Journal of Gastronomy and Food Science 2020 Dec;22:100259. [doi: 10.1016/j.ijgfs.2020.100259]
20. Campbell ME, Gardner CE, Dwyer JJ, Isaacs SM, Krueger PD, Ying JY. Effectiveness of public health interventions in food safety: a systematic review. Can J Public Health 1998;89(3):197-202. [doi: 10.1007/BF03404474] [Medline: 9654807]
21. Thomas MS, Feng Y. Consumer risk perception and trusted sources of food safety information during the COVID-19 pandemic. Food Control 2021 Dec;130:108279. [doi: 10.1016/j.foodcont.2021.108279] [Medline: 36568483]
22. Kim S, Capasso A, Cook SH, et al. Impact of COVID-19-related knowledge on protective behaviors: The moderating role of primary sources of information. PLoS ONE 2021;16(11):e0260643. [doi: 10.1371/journal.pone.0260643]
23. Klein G. Spread of viruses through the food chain. Dtsch Tierarztl Wochenschr 2004 Aug;111(8):312-314. [Medline: 15469057]
24. Borges do Nascimento IJ, Cacic N, Abdulazeem HM, et al. Novel coronavirus infection (COVID-19) in humans: a scoping review and meta-analysis. J Clin Med 2020 Mar 30;9(4):941. [doi: 10.3390/jcm9040941] [Medline: 32235486]
25. Shopping for food during the COVID-19 pandemic-information for consumers. : US Food and Drug Administration; 2020.
26. Running essential errands: grocery shopping, take-out, banking, getting gas, and doctor visits: May 11, 2020. Centers for Disease Control and Prevention. URL: https://stacks.cdc.gov/view/cdc/88371/ [accessed 2025-08-18]
27. Bialek S, Gierke R, Hughes M, et al. Coronavirus disease 2019 in children — United States, February 12–April 2, 2020. MMWR Morb Mortal Wkly Rep 2020;69(14):422-426. [doi: 10.15585/mmwr.mm6914e4]

abbreviations

FAO: Food and Agriculture Organization
WHO: World Health Organization

Edited by A Mavragani; submitted 11.07.24; peer-reviewed by AN Ali, A Ahmad, Z Batu; revised version received 02.06.25; accepted 03.06.25; published 20.08.25.

Please cite as:
Farah AM, Nour TY, Roble AK, Obsiye M, Aden MA, Abdi AS, Getnet F
Food Handling Practices Among Food Businesses in Jigjiga, Eastern Ethiopia, During the COVID-19 Pandemic: Cross-Sectional Study
JMIR Form Res 2025;9:e64214
URL: https://formative.jmir.org/2025/1/e64214
doi:10.2196/64214

© Alinoor Mohamed Farah, Tahir Yousuf Nour, Abdurahman Kedir Roble, Muse Obsiye, Mowlid Akil Aden, Abdulahi Siraj Abdi, Fentabil Getnet. Originally published in JMIR Formative Research (https://formative.jmir.org), 20.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

A Novel Approach to Care Redesign Collaboration Between Emergency and Specialty Departments: Qualitative Experience Report

Sonia Rose Harris[1*], MPH, MSW; Christian Rose[2*], MD; Samantha MR Kling[1], RD, PhD; Brett A Cohen[2,3], MD; Olga Goldberg[4], MD; Hannah Walton[5], MS; Darlene Veruttipong[1], MPH; Mohamed Alhadha[6], MBA; Sheneé K Laurence[6], MPH, BSN, RN; Shashank Ravi[2], MD, MBA; Carl A Gold[4], MD, MS; Jonathan G Shaw[1], MD, MS; Laurice Yang[4*], MD, MHA; Cati G Brown-Johnson[1], PhD

[1]Evaluation Sciences Unit, Division of Primary Care and Population Health, Department of Medicine, School of Medicine, Stanford University, 3180 Porter Drive, Palo Alto, CA, United States

[2]Department of Emergency Medicine, School of Medicine, Stanford University, Stanford, CA, United States

[3]Department of Emergency Medicine, Vanderbilt University Medical Center, Vanderbilt University, Nashville, TN, United States

[4]Department of Neurology and Neurological Sciences, School of Medicine, Stanford University, Stanford, CA, United States

[5]School of Medicine, Stanford University, Stanford, CA, United States

[6]Stanford Health Care, Stanford, CA, United States

[*]these authors contributed equally

Corresponding Author:
Samantha MR Kling, RD, PhD
Evaluation Sciences Unit, Division of Primary Care and Population Health, Department of Medicine, School of Medicine, Stanford University, 3180 Porter Drive, Palo Alto, CA, United States

Abstract

Background: Given the rising demand for emergency department (ED) services and coupled with the scarcity of specialty care availability, there is an urgency to design a system for appropriate, effective, and timely ED-to-specialty ambulatory referrals. Efficient care transitions are important to patient outcomes and experience and require cross-specialty cooperation, as care transitions affect practices and resources of individuals, departments, and institutions.

Objective: Here, our objective was to (1) describe a collaboration between Stanford's Emergency Medicine and Neurology and Neurological Sciences departments aimed at designing and implementing an optimized discharge process and transition of care from ED to ambulatory neurology for follow-up care and (2) the resulting intervention from the collaboration.

Methods: We describe the process for barrier identification, tools used to foster partnership and intervention ideation, and the resulting intervention. Our experience and findings are integrated into a 4-component framework for future interdepartmental collaborations: (1) cross-specialty team meetings, (2) preimplementation interviews, (3) a design thinking focus group session, and (4) small group meetings. Qualitative data included observational notes and document review from biweekly cross-specialty and small group meetings, preimplementation interviews, and design-thinking focus groups.

Results: Our process included 10 cross-specialty team meetings with 8 physicians and operational representatives, 18 individual preimplementation interviews, 10 focus group participants, and 9 small group meetings. The process components fostered collaboration and teamwork among the multidisciplinary team; supported early identification of barriers and facilitators, including divergent understanding of project goals; and developed creative ideas that contributed to intervention development. The collaboration resulted in a 4-pronged multimodal intervention. Two elements focused on modifying clinical practice to better triage clinically appropriate ED referrals to ambulatory neurology: (1) optimizing management of conditions in the ED to reduce preventable referrals, and (2) increasing deferral to Primary Care clinicians to direct appropriate specialty follow-up care. Two additional structural elements sought to directly improve appropriate referral timeliness by (3) streamlining insurance authorization processes and (4) increasing neurology appointment availability.

Conclusions: This cross-specialty collaboration resulted in a multimodal intervention that called for both structural and practice changes, which were novel and supported by the results of comparable interventions. Future applications of this framework can validate its utility among different collaborative groups in new settings.

(JMIR Form Res 2025;9:e22028) doi:10.2196/22028

XSL•FO
RenderX

KEYWORDS

emergency service; referral and consultation; interdisciplinary communication; neurology; ambulatory care; hospital

Introduction

Emergency departments (EDs) are critical interfaces between acute care and the broader health care system, requiring coordination with multiple specialties to provide comprehensive patient care. ED visits among adults have been increasing, with the United States reporting 155 million ED visits in 2022, equating to 47 visits per 100 people [1]. This sustained growth in ED usage, coupled with persistent shortages in specialty care, emphasizes the importance of effective ED-to-specialty referral processes for continuous, comprehensive care [2-4].

Shortages in neurological specialty care are pronounced, a significant threat as characterized by the American Academy of Neurology [5,6]. Nationally, challenges in neurology include longer wait times for appointments compared with other specialty clinics, difficulties filling positions, and establishing care for Medicaid patients [5]. Neurology clinics have competing needs where they must balance referrals from the hospital, ED, and community physicians with their existing case load [7].

Traditionally, health care resource allocation has followed a "more is better" paradigm, assuming that increased availability translates to improved access [8,9]. However, this approach fails to account for real-world limitations such as space and availability of specialty clinicians [8-10]. Given finite resources and growing demand, health care systems must shift toward optimization of resources through strategic, evidence-based interventions that address both clinical practices and structural barriers [11,12].

Increasingly recognized as a key to comprehensive health care, interdisciplinary collaboration is associated with increased patient safety, lower hospitalization rates, and reduced rates of complications and medical errors [13,14]. However, traditional approaches to developing such collaborative interventions are often expert-driven and may not adequately address end user needs or contextual complexities.

Design thinking, a human-centered approach to problem-solving, has emerged as a promising methodology in health care. Evidence suggests that design thinking results in more usable, acceptable, and effective interventions compared to traditional approaches [15]. This approach emphasizes stakeholder engagement, iterative refinement, and empathy-driven design, making it particularly well-suited for complex health care challenges that require coordination across multiple disciplines and care settings.

Using this foundation, we describe the collaboration between the Stanford Department of ED and the Stanford Department of Neurology and Neurological Sciences (hereinafter "Neurology") that aimed to design and implement an optimized discharge process and transition of care. Here, we describe the resulting intervention and our novel framework that integrates design thinking principles with cross-specialty collaboration.

Methods

Setting

Institutional Setting

The ED and Neurology departments collaborated on a quality improvement (QI) project from November 2022 to July 2023 at Stanford Medicine, a quaternary academic medical center that serves the Bay Area of Northern California, United States. Within our medical center, both departments are overseen by the same operational service line that provides member departments with strategic operations support and shared infrastructure. The QI initiatives described here were under the Stanford Improvement Capability Development Program [16].

The ED is staffed by 90 physicians (60 clinical full-time equivalents) and 60 trainees (15 per residency class) and conducts 95,000 ED encounters yearly. Neurology provides inpatient and ambulatory services. Ambulatory care includes 9 satellite clinics with 66 physicians, 7 advanced practice providers, 9 nurses, and 60 trainees (residents and fellows) and delivers 48,000 ambulatory encounters yearly. The inpatient service includes 3 consult services (Neurohospitalist, Stroke, and Neurocritical Care) and 4 admitting services (Neurohospitalist, Stroke, Neurocritical Care, and Epilepsy Monitoring Unit). Stanford is a Comprehensive Stroke Center and a Level I Trauma Center.

Project Origin

To improve care transitions between emergency services and ambulatory specialty care, the ED first partnered with the Department of Orthopedic Surgery. In June 2022, this partnership implemented a workflow that triaged ED referrals to orthopedic surgery as "routine," "urgent," or "immediate," with specified follow-up timelines; results will be reported elsewhere. Following this collaboration, the ED and Neurology departments partnered on a similar QI initiative, as the ED places the second most referrals to ambulatory Neurology. Per month, the ED sends approximately 72 referrals to Neurology, and approximately 40 are scheduled with an appointment. The most common referrals are for headache or migraine (230 per year) and dizziness or vertigo (58 per year).

The Stanford Medicine Evaluation Sciences Unit (ESU) joined this team to provide enhanced evaluation support. The ESU is a group of clinicians and nonclinicians, doctorate- and master's-trained faculty, and staff with methodological expertise in evaluation and health services research [16].

Clinical and evaluation members who participated in this collaborative project will be referred to as the cross-specialty team.

Data Collection

Overview

The ESU collected qualitative data throughout the partnership to assess participant perspectives, document collaborative

XSL•FO
RenderX

processes, and provide rapid feedback. This included defining and reflecting on goals, intervention, and implementation. Qualitative data included observational notes and document review from biweekly cross-specialty team meetings (Process Component 1), preimplementation interviews [17] (Process Component 2), design-thinking [18] focus group (Process Component 3), and observational notes and document review from ad hoc individual or small group meetings (Process Component 4).

Meeting Observations—Cross-Specialty Team Meetings and Small Group Meeting Observations (Process Components 1 and 4)

The cross-specialty team held biweekly meetings on Zoom (Zoom Video Communications) from November 2022 to June 2023. Meetings aimed to develop a shared goal, an intervention, and an implementation approach centered around transitioning patients from ED to ambulatory neurology for follow-up care. Meeting attendees could include physicians, QI consultants, new patient coordinators (NPCs), service line managers, and ESU. Small group meetings focused on gathering in-depth information on department-specific issues; these mostly occurred between Neurology and ESU.

The ESU provided documentation, rapid feedback, and technical assistance during meetings. Two ESU team members (SRH and CB-J) took detailed observational notes to capture discussion and decision-making around goals, intervention, and implementation approaches. Rapid ethnographic principles were followed for observation notes [19]. The ESU collected meeting documents and products, for example, PowerPoints, spreadsheets, PDFs, and screenshots, to capture the development of the intervention.

In addition, the ESU (SRH, CB-J, and JGS) provided rapid, data-informed feedback to the team following methods outlined in the Rapid Implementation Framework (RIF) [20]. Finally, the ESU answered questions and contributed insights related to their expertise, such as primary care (JGS), local QI projects on care coordination (JGS, SMRK, and CB-J), and implementation science (JGS, SMRK, and CB-J) [21].

Preimplementation Interviews (Process Component 2)

Semistructured interviews (Table 1) were informed by the premortem approach [17], RIF [20], and the Stanford Lightning Report [22]. Semistructured interview guides (Multimedia Appendix 1) were designed to capture barriers and facilitators to implementation and hypothetical reasons for project failure. The study used a purposive sampling strategy, including the cross-specialty team and key stakeholders identified through snowball sampling. Cross-specialty team members identified additional stakeholders (ie, information technology staff and chief residents) during their interviews. Prospective participants received an email invitation to participate from the ESU. Two ESU team members (SRH and HW) conducted interviews from January to April 2023, recording via Zoom with verbatim transcription (Rev Inc).

Table . Roles of preimplementation interview participants and team members in a cross-specialty collaboration.

Role	Number of preimplementation interview participants	Number of cross-specialty team members[a]
Physicians	10	5
Operations and technology solutions	5	3
Scheduling	3	0
Total	18	8

[a]Cross-specialty team denotes participants who attended cross-specialty team meetings (Process Component 1)

Design Thinking Focus Group (Process Component 3)

In March 2023, ESU facilitated a 1.5-hour design thinking focus group discussion via Zoom. The session had four goals: (1) better understand potential intervention users, (2) challenge initial assumptions, (3) redefine problems to be user-focused, and (4) create innovative solutions to prototype and test. A total of 10 people from the cross-specialty collaboration team attended, representing Neurology, ED, Primary Care, ESU, Operations, and Quality Consulting. The session included a brief introduction to design thinking and three activities: empathy mapping [23], the worst possible idea [24], and brainstorming [25]. Participants completed an anonymous feedback survey (Multimedia Appendix 2).

Data Analysis

Rapid analysis methods [20,22] were applied to interview data to synthesize and provide prompt, actionable feedback to the cross-specialty team and capture iterations of the intervention. Two coauthors (SRH and HW) used a priori codes in a matrix analysis [26] in Excel (Microsoft Corp). Early results were presented to the cross-specialty team during a biweekly meeting to validate findings and in small group sessions to generate feedback.

Data from observational notes and document review were used to triangulate findings reported in the preimplementation interviews. Observational notes and interview transcripts were revisited as needed to validate or query findings. For the design thinking focus group, data were analyzed into three broad categories: what worked, what did not work, and ideas for future sessions.

Ethical Considerations

This project received a nonresearch determination by Stanford University's institutional review board (IRB #68792). Interview participants gave verbal consent before each recorded interview. Participants were told their privacy and confidentiality would be maintained. Interview participants were not compensated. As this is a report on the early phases of a QI project focused

XSL•FO
RenderX

on cross-specialty collaboration, we did not involve patients or the public in the research design, conduct, reporting, or dissemination. However, we sought to include diverse health care employees to understand workflow perspectives. Participants did not receive any form of compensation for participation. The data supporting this study's findings are available on reasonable request.

Results

Novel Framework to Increase Cross-Specialty Collaboration

This cross-specialty collaboration with enhanced evaluation support identified core activities and milestones for intervention development and implementation. Major process components of our proposed Framework to Increase Cross-Specialty Collaboration (Figure 1) include (1) cross-specialty team meetings, (2) preimplementation interviews, (3) design thinking focus group session, and (4) small group meetings.

XSL•FO
RenderX

Figure 1. Four-component framework that was used to support cross-specialty collaboration to design and implement an optimized discharge process and transition of care from emergency department to ambulatory Neurology for follow-up care.

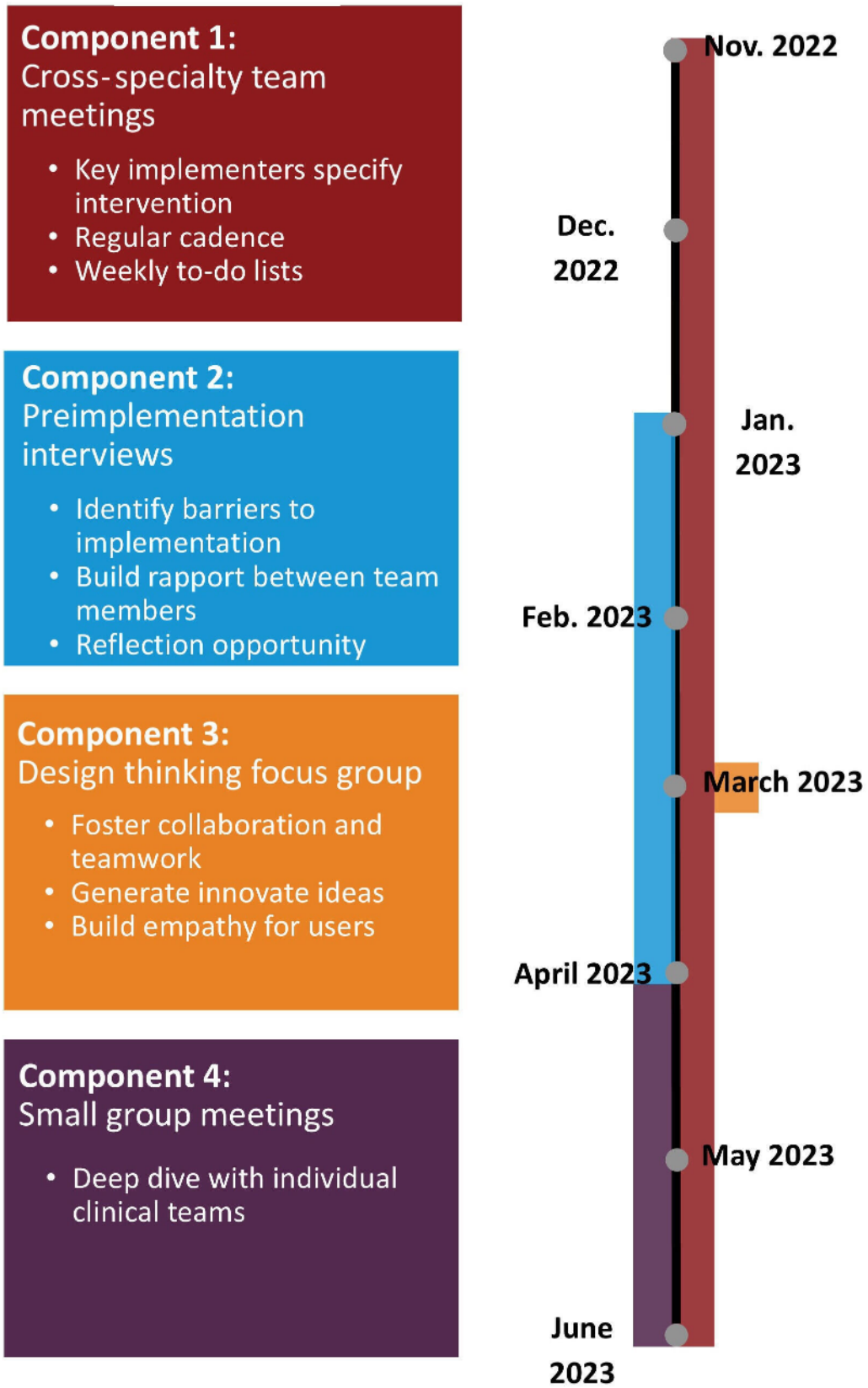

Engaging in these process components fostered collaboration and teamwork among the multidisciplinary team, supported early identification of barriers and facilitators, including divergent understanding of project goals and identification of creative ideas that contributed to intervention development.

The process resulted in a multimodal intervention that was informed by qualitative data (as described in Table 2), that addressed diverse needs and perspectives, and aimed to enhance quality through greater attention toward timeliness and appropriateness of ED-to-Neurology referrals. The intervention

XSL•FO
RenderX

included two modifications to clinical practice and two structural changes. Element 1, enhancing clinical practices through procedures or treatments, optimized clinical management of conditions in the ED to reduce ED-to-Neurology referrals. For example, increasing the ED's capacity to perform occipital nerve blocks to eliminate the need for referral. Element 2 embedded a referral considerations document within the electronic medical record, guiding physicians toward primary care or neurology follow-up when appropriate. Element 3 addressed timeliness by streamlining insurance authorization by having financial counselors call patients who could not be seen by Neurology. Element 4 addressed access by adding three Neurology appointment slots per week by the headache specialist and a potential weekend or night clinic.

Table . The multimodal intervention that resulted from a cross-specialty collaboration that aimed to design and implement an optimized discharge process and transition of care from an ED[a] to ambulatory neurology for follow-up[b].

Element of intervention	Description of element	Specific theme or learning from the collaboration framework (process component) that informed the element
Increasing ED clinical practice (practice change)	• Investigated procedures or treatments (eg, occipital nerve block) that could be administered in the ED, in lieu of an urgent referral • Procedures would not require prior insurance authorization in the ED • ED physicians received specific training from Neurology	• Creative brainstorming during the design thinking focus group led to the idea for optimizing the management of conditions in the ED through increasing clinical practice • Collaborative meeting time used to strategize to implement into practice • Conversations to address divergent understanding of project goals led to recognition that timeliness and appropriateness are intertwined, and both needed to be addressed
Use all referral services (practice change)	• Focus on headache or migraine and dizziness • Increase the percentage of complex cases to neurology • Coordinate care with Primary Care for appropriate cases • The framework defines appropriate referrals to a primary care clinician (PCP[c]) versus specialty referral to Neurology • Short two-sentence commentary embedded into the Epic electronic health record (Epic Systems Corporation) referral order, as a point of order entry education • Detailed diagnosis-specific consideration document linked at the point of order entry	• The decision to focus attention on two diagnoses (headache and dizziness) emerged after rapid implementation feedback from interviews • Primary Care perspective integrated after conversations in cross-specialty collaboration meetings • Tensions related to appropriateness versus timeliness led to a need for a referral consideration document • Two-sentence commentary and detailed description reviewed during cross-specialty collaboration meetings
Streamline insurance authorization (structural change)	• NPCs[d] spend approximately 1 - 4 hours per week calling patients who cannot be scheduled due to insurance • Causes frustration for both NPCs and patients • New model to change this responsibility from NPCs to finance counselors	• Emerged in response to problem statements for the scheduler identified during the design thinking session
Increase access in neurology (structural change)	• Stanford headache specialist (3 total) each open 3 urgent slots per patients from the ED with an urgent referral • As of July 2023, awaiting approval for a weekend or night clinic and APP[e] transition clinic	• In response to the goal of increasing the timeliness of referrals for urgent cases • In response to failure points highlighted in the preimplementation interviews, increasing access occurred as a final step after addressing insurance coverage, and appropriateness of referral

[a]ED: emergency department.

[b]Intervention elements were mapped to the components of the collaboration framework that most informed the element.

[c]PCP: primary care physician.

[d]NPC: new patient coordinator.

[e]APP: advanced practice provider.

XSL•FO
RenderX

Cross-Specialty Team Meetings (Process Component 1)

The team met bimonthly from November 2022 to May 2023 (10 meetings; range 1 - 2 h; Figure 1). Team members (5 - 15 per meeting) included clinical, scheduling, QI, operations, and ESU. Members of the cross-specialty team attributed meeting success to having regular cadence, a primary ED leader, team-specific action items, and attendance by at least one physician from both ED and Neurology.

The initial meetings were attended by operations, schedulers, and physicians from Neurology and ED. The administrative director for quality and operations for the service line provided oversight and served as liaison between the cross-specialty team and operations. Initial aims included identifying all applicable contributors and establishing a specific, measurable goal for the collaboration. In these meetings, ED discussed prior experience with Orthopedics.

The ESU used the cross-specialty team meetings in the middle and late stages of the collaboration to provide rapid feedback [20] informed by the preimplementation interviews, design thinking focus group, and small group meetings (Process Components 2 - 4). Two of the most prominent factors of the RIF reports from February to March 2023 were (1) hypothetical barriers to implementation, that is, limited neurology appointments; and (2) tensions between departments caused by divergent understanding of project goals. To address concerns regarding limited neurology appointments, the project team decided to focus only on the two most common reasons for referral—headaches and dizziness. The team also invited a Primary Care representative to these meetings, recognizing Primary Care's role in the intervention. From March to May 2023, cross-specialty team meetings (Table 2) focused on developing and launching the intervention to enhance Primary Care usage.

Component 2: Preimplementation Interviews

Overview

Interviews with participants (Table 1) asked about elements necessary for success and potential threats that might thwart success ("premortem" [17] questioning). Interviews identified a need for trust, divergent understanding of project goals, and potential failure points, all of which contributed to goal refinement and intervention development.

Need for Trust

Interviewees recognized that the process to refer and coordinate follow-up care post-ED discharge involves many team members and workflows housed in multiple departments; here, this included ED, Neurology, and Primary Care. Proposing to change another team's processes concerned interviewees and required them to trust that team members were working collaboratively. One participant reflected:

> *We are asking another service to change their processes and we can only suggest so much.* [Participant 10, ED]

Divergent Understanding of Project Goals

Interviews revealed that team members had different views on the intended project goal, contributing to an "us and them" (Participant 9, Neurology) dynamic. Interviewed participants described the collaboration's goal as either to increase the timeliness of Neurology follow-up care or the appropriateness of ED referrals to Neurology.

Although both timeliness and appropriateness are factors influencing patient access to care, approaches differed depending on the driving aim. The ED prioritized timeliness by focusing on different levels of referral priorities (ie, routine, urgent, and immediate) tied to defined timelines similar to their experience with Orthopedic Surgery. In contrast, Neurology prioritized appropriateness, proposing to define the clinical characteristics of patients needing a Neurology referral versus those needing a primary care referral, for example.

Through our collaborative approach, including regular meetings, design thinking sessions, and evaluation team rapid feedback, the participants came to recognize the intertwined nature and necessity to address both timeliness and appropriateness.

Potential Failure Points

Overview

Preimplementation interviews also asked participants to anonymously name potential failure points, allowing team members to explore sensitive barriers. Potential failure points included the need for widespread education for frontline clinicians, balancing capacity, and insurance coverage.

ED Education

Interviewees recognized that any intervention would require widespread ED education. One participant flagged a concern about scale:

> *[when] you're teaching, you can teach one or five people, but this is a whole department.* [Participant 3, Neurology]

This would be further complicated by resident rotations and the need for continuously available, asynchronous education.

Limited Capacity to Increase Access

Interviewees, especially those from Neurology, were concerned that current capacity limited intervention options and that the project itself could exacerbate capacity issues. Interviewees reported that their clinics had substantial wait times for appointments and overburdened schedules as urgent post-ED or hospital follow-ups were being scheduled outside of clinic hours. Balancing capacity issues with other concerns was of clear concern.

> *If we expect that patients will follow up in a certain time period, but at the end of the day the capacity of providers and clinics doesn't match that expectation, then it may fail.* [Participant 10, ED]

Insurance Coverage

Interviewees identified that health insurance coverage may complicate the referral process and patient expectations for follow-up care. Referring a patient to appropriate follow-up

XSL•FO
RenderX

care was a top priority to interviewees, but depending on the patients' insurance, the process of referral was not straightforward.

> *Coming up with a plan that's the same no matter someone's insurance status is part of our mission ... If I send a referral to someone who is under no medical, legal responsibility to have to follow up with a patient, they might not, not because they disagree with the management plan of the ED.* [Participant 1, ED]

Design Thinking Focus Group (Process Component 3)

Overview

The design thinking focus group session (March 30, 2023), which was based on published best practices [18], engaged team members in 3 activities: empathy mapping, worst possible idea, and brainstorming (Figure 2).

XSL•FO
RenderX

Figure 2. Results of empathy mapping and worst possible ideas from the design thinking focus group that was a component of the framework to support cross-specialty collaboration to optimize the discharge process and transition of care from the emergency department (ED) to ambulatory neurology for follow-up care.

XSL•FO
RenderX

Empathy Mapping

Participating team members jointly identified 3 personas (hypothetical end users) affected by the problem and the proposed interventions: ED physician, an ED patient with a migraine who has tried three different medications, and an NPCor scheduler in Neurology. The team was split into two arbitrary groups. Per the empathy mapping [23] protocol, each individual team member was given a few minutes to note what each persona might say, think, do, and feel on their team's web-based whiteboards (JamBoard; Google Inc). For example, team members noted that an NPC may say [to a patient]: "Sorry, we just don't have appointments for a couple months," which might make the NPC think that they need "more clarity on what is urgent." The NPC might "call a neurologist to try to find a slot," which makes the NPC feel "overwhelmed and frustrated."

Teams discussed their whiteboards, summarized findings, and then summarized the goal of each persona as a large group, including (1) patients want to be seen by a clinician and feel better, and (2) Neurology NPCs want to schedule a patient without unnecessary burden and frustration. Team members noted the lack of control and high level of uncertainty for all roles engaged, including patients, physicians, and staff, in the existing referral system.

Worst Possible Idea

Small groups created a list of the least effective ideas to address problems identified through "empathy mapping." One of the "worst possible ideas" was to make clinic staff and clinicians work 18-hour shifts to accommodate all referrals, overburdening clinical staff and creating burnout. This led participants to discuss the importance of not significantly increasing the workload for any staff.

Brainstorming

The session ended with a brainstorming [25,27] discussion wherein participants were encouraged to generate all ideas, big or small, without hesitation or consideration of practical limitations. Ideas included (1) all ED patients to go to primary care first to determine need for a Neurology consult, (2) a Neurology "doctor of the day" who is available to see urgent Neurology patients in ambulatory settings, and (3) a Neurology express care model.

Design Thinking Feedback Survey

Survey participants (n=8) reported that the design thinking session fostered collaboration and teamwork, generated innovative ideas, and built empathy for users. Participants reported they felt high engagement, were interested in the design thinking process, and were challenged by the activities. A few participants suggested that design thinking exercises would have been helpful earlier in the collaboration because of the emphasis on generating innovative solutions.

Problem statements identified during empathy mapping and ideas generated during brainstorming contributed to the final multimodal intervention, for example, streamlined insurance authorization to address NPC needs.

Component 4: Small Group Meetings

The ESU met with individual or small groups, usually Neurology, ad hoc to explore emerging work thought to be peripherally related to the primary objective of the cross-specialty team meetings (Figure 1). These small group meetings were used to document intervention development progress in detail and then prioritize what to discuss in the larger implementation meetings. Small group meetings were essential for capturing the two structural changes: (1) streamlined insurance authorization processes and (2) increasing Neurology appointment availability (Table 2).

Discussion

Principal Findings

Given rising demand for ED services [28], coupled with a scarcity of specialty care availability [5], the urgency for effective and prompt ED-to-specialty referrals has never been more pronounced. To address this need, we outlined the process used by a cross-specialty collaboration at a single institution and a subsequent multimodal intervention aimed at enhancing quality through appropriateness and timeliness of ED-to-Neurology referrals. The intervention included two elements focused on modifying clinical practice: (1) optimizing management of the most common conditions like headache and dizziness in the ED to reduce preventable referrals, and (2) enhancing the use of Primary Care services to direct specialty follow-up care as needed. Two additional structural elements sought to directly improve referral timeliness by (3) streamlining insurance authorization processes and (4) increasing Neurology appointment availability.

The cross-specialty initiative was enabled through an iterative collaboration process that included regular cross-specialty team meetings, preimplementation interviews, a design thinking focus group session, and small group meetings. Each component provided key insights that shaped the development of the intervention elements and helped build rapport between specialties. Our experience informed the proposed Framework to Increase Cross-Specialty Collaboration for future QI endeavors requiring interdepartmental partnership for future collaboration between ED and specialties.

Significance of Intervention

Our intervention shares commonalities with other successful initiatives aimed at improving transitions from the ED to ambulatory Neurology for follow-up, while also incorporating its own innovative but broadly applicable elements. For example, the emphasis on streamlining referrals and insurance authorization resembles process improvements made in clinics like UCLA Fast Neuro Clinic [29] and Florida Access Clinic [30]. However, our intervention's focus on optimizing ED management and leveraging Primary Care expertise before referring to Neurology represents a more nuanced approach to ensuring appropriate referrals.

While some access initiatives have focused narrowly on increasing the number of Neurology clinicians or clinic availability [29-31], our intervention considered clinical appropriateness to promote more judicious referrals, as well as

XSL•FO
RenderX

supporting ED clinicians to work at the top of their license. This balancing of timeliness and appropriateness through clinical practice and structural changes is a unique contribution based on currently available literature. There were concerns raised by the Neurology team about educating the entire ED on new workflows, but these activities are relatively straightforward for the ED, given existing processes for QI and education at faculty and resident monthly meetings. It was during these monthly meetings that inclusion and exclusion criteria to Neurology referrals versus Primary Care were identified.

Value of Collaboration and Design Thinking

The proposed Framework to Increase Cross-Specialty Collaboration fostered integration of diverse insights from ED, Neurology, operations, and Primary Care. Rather than moving forward with an intervention based on team assumptions about referral issues and prior experiences, we gathered input directly from participants through techniques such as preimplementation interviews and design thinking. The Framework allowed the team to identify and address four key hurdles to intervention design: (1) “us versus them” mentality; (2) divergent understanding of project goals (appropriateness vs timeliness); (3) narrowing scope to top Neurology referrals (ie, headache and dizziness); and (4) considering longitudinal health services, like Primary Care, for the patient when appropriate.

Existing evidence on the impact of design thinking in health care demonstrates that it improves intervention usability, acceptability, and outcomes compared to traditional expert-driven methods [15]. Indeed, design thinking is increasingly recognized as a promising human-centered approach for complex problem-solving in health care, moving away from authoritarian dynamics toward cocreative environments [15]. It emphasizes empathy, equitable engagement, and continuous refinement based on end user input [32].

Our use of design thinking in the proposed framework helped elicit openness to nontraditional interventions for improving access. By gathering perspectives on lived experiences of physicians and patients, the intervention resolved underlying problems, instead of creating just symptomatic fixes. This human-centered foundation was key in developing a contextualized, multimodal intervention acceptable to both specialties and aligned with end user needs. Collaborative approaches informed by design-thinking principles have also been shown to result in effective and usable intervention elsewhere [15].

Limitations

This project has three major limitations related to generalizability, effectiveness, and representation. First, this work represents a single academic institution, which allowed us to leverage existing infrastructure but limited generalizability. Furthermore, the ensuing intervention focused on only the two most common reasons for Neurology referrals, headache and dizziness. Second, the implementation, effectiveness, and sustainability of the intervention remain untested. Third, although we included a range of perspectives, we lack frontline nursing and patient perspectives.

Overall, our process showcases the promise of partnership and design thinking to tailor localized improvements in complex systems like ED to specialty care transitions.

Conclusions

Interventions to optimize ED to specialty service transitions are vital given the increasing demand for ED services and growing shortage of specialty care availability. Our four-component framework enabled trust, rapport, and early identification of barriers and facilitators to implementation while increasing cross-specialty collaboration and was comprised of: (1) cross-specialty team meetings; (2) preimplementation (“premortem”) interviews; (3) a design thinking focus group; and (4) small group meetings. Development of future access interventions can further refine the Framework to Increase Cross-Specialty Collaboration and look to leverage key elements of our intervention: modifications to clinical practice (eg, including occipital block in ED clinician treatment repertoire) and structural changes (eg, streamlining insurance authorization processes).

Ultimately, this collaboration resulted in a multimodal intervention that called for both structural and practice changes, which were novel and supported by the results of comparable interventions. Future applications of this framework can validate its utility among different collaborative groups in new settings.

Acknowledgments

The authors acknowledge Ian Brown of Stanford School of Medicine and Stanford Health Care and Marcy Winget, director of the Evaluation Sciences Unit. This project was supported by Stanford Health Care as part of the Improvement Capability Development Program. Generative artificial intelligence (AI) was used to edit small sections of the manuscript (eg, single sentences) for grammar or clarity; all AI-generated edits were reviewed and further edited by authors.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
The semistructured guide for preimplementation interviews that were designed to capture barriers and facilitators to implementation and hypothetical reasons for failure of the project to optimize the discharge process and transition of care from ED to ambulatory neurology for follow-up care.
[PDF File, 66 KB - formative_v9i1e22028_app1.pdf]

Multimedia Appendix 2
An anonymous survey administered to participants at the design thinking focus group. The aim of the survey was to evaluate the experience of the focus group participants.
[PDF File, 59 KB - formative_v9i1e22028_app2.pdf]

References

1. Cairns C, Ashman JJ, Kang K. Emergency department visit rates by selected characteristics: United States, 2022. NCHS Data Brief 2024 Aug(503):1-8. [doi: 10.15620/cdc/159284] [Medline: 39302262]
2. Cairns C, Ashman JJ, King JM. Emergency department visit rates by selected characteristics: United States, 2020. NCHS Data Brief 2022 Nov(452):1-8. [Medline: 36453979]
3. Jones L, McClean J, Avitzur O. US health care system in 2035: the neurology future forecasting series. Ovid. URL: https://oce.ovid.com/article/00006114-202203080-00005?relatedarticle=y [accessed 2023-10-02]
4. Zhao C, Lee K, Do D. Neurology consults in emergency departments. Neur Clin Pract 2020 Apr;10(2):149-155. [doi: 10.1212/CPJ.0000000000000712]
5. Dall TM, Storm MV, Chakrabarti R, et al. Supply and demand analysis of the current and future US neurology workforce. Neurology (ECronicon) 2013 Jul 30;81(5):470-478. [doi: 10.1212/WNL.0b013e318294b1cf] [Medline: 23596071]
6. Majersik JJ, Ahmed A, Chen IHA, et al. A shortage of neurologists—we must act now: a report from the AAN 2019 transforming leaders program. Neurology (ECronicon) 2021 Jun 14;96(24):1122-1134. [doi: 10.1212/WNL.0000000000012111] [Medline: 33931527]
7. Chapman FA, Pope AE, Sorensen D, Knight RS, Al-Shahi Salman R. Acute neurological problems: frequency, consultation patterns and the uses of a rapid access neurology clinic. J R Coll Physicians Edinb 2009 Dec;39(4):296-300. [doi: 10.4997/JRCPE.2009.402] [Medline: 21152464]
8. Coombs NC, Campbell DG, Caringi J. A qualitative study of rural healthcare providers' views of social, cultural, and programmatic barriers to healthcare access. BMC Health Serv Res 2022 Apr 2;22(1):438. [doi: 10.1186/s12913-022-07829-2] [Medline: 35366860]
9. Pugh J, Penney LS, Noël PH, et al. Evidence based processes to prevent readmissions: more is better, a ten-site observational study. BMC Health Serv Res 2021 Mar 1;21(1):189. [doi: 10.1186/s12913-021-06193-x] [Medline: 33648491]
10. Anderson GF, Chalkidou K. Spending on medical care: more is better? JAMA 2008 May 28;299(20):2444-2445. [doi: 10.1001/jama.299.20.2444] [Medline: 18505956]
11. Harris DCH, Davies SJ, Finkelstein FO, et al. Increasing access to integrated ESKD care as part of universal health coverage. Kidney Int 2019 Apr;95(4S):S1-S33. [doi: 10.1016/j.kint.2018.12.005] [Medline: 30904051]
12. Patrick J, Puterman ML. Reducing wait times through operations research: optimizing the use of surge capacity. hcpol 2008;3(3):75-88. [doi: 10.12927/hcpol.2008.19575]
13. Warren JL, Warren JS. The case for understanding interdisciplinary relationships in health care. Ochsner J 2023;23(2):94-97. [doi: 10.31486/toj.22.0111] [Medline: 37323516]
14. Bendowska A, Baum E. The significance of cooperation in interdisciplinary health care teams as perceived by Polish medical students. Int J Environ Res Public Health 2023 Jan 5;20(2):954. [doi: 10.3390/ijerph20020954] [Medline: 36673710]
15. Altman M, Huang TTK, Breland JY. Design thinking in health care. Prev Chronic Dis 2018 Sep 27;15:E117. [doi: 10.5888/pcd15.180128] [Medline: 30264690]
16. Vilendrer S, Saliba-Gustafsson EA, Asch SM, et al. Evaluating clinician-led quality improvement initiatives: a system-wide embedded research partnership at Stanford Medicine. Learn Health Syst 2022 Oct;6(4):e10335. [doi: 10.1002/lrh2.10335] [Medline: 36263267]
17. Gilmartin H, Lawrence E, Leonard C, et al. Brainwriting premortem: a novel focus group method to engage stakeholders and identify preimplementation barriers. J Nurs Care Qual 2019;34(2):94-100. [doi: 10.1097/NCQ.0000000000000360] [Medline: 30148746]
18. Foster MK. Design thinking: a creative approach to problem solving. Manage Teach Rev 2021 Jun;6(2):123-140. [doi: 10.1177/2379298119871468]
19. Mulhall A. In the field: notes on observation in qualitative research. J Adv Nurs 2003 Feb;41(3):306-313. [doi: 10.1046/j.1365-2648.2003.02514.x] [Medline: 12581118]
20. Finley EP, Chrystal JG, Gable AR, et al. The Rapid Implementation Feedback (RIF) report: real-time synthesis of qualitative data for proactive implementation planning and tailoring. Implement Sci Commun 2024 Jun 21;5(1):69. [doi: 10.1186/s43058-024-00605-9] [Medline: 38907331]

XSL•FO
RenderX

21. Kling SMR, Saliba-Gustafsson EA, Winget M, et al. Teledermatology to facilitate patient care transitions from inpatient to outpatient dermatology: mixed methods evaluation. J Med Internet Res 2022 Aug 3;24(8):e38792. [doi: 10.2196/38792] [Medline: 35921146]
22. Brown-Johnson C, Safaeinili N, Zionts D, et al. The Stanford Lightning Report Method: a comparison of rapid qualitative synthesis results across four implementation evaluations. Learn Health Syst 2020 Apr;4(2):e10210. [doi: 10.1002/lrh2.10210] [Medline: 32313836]
23. Peng F. Design thinking: from empathy to evaluation. In: Herath D, St-Onge D, editors. Foundations of Robotics: A Multidisciplinary Approach with Python and ROS: Springer Nature, Vol. 2022:63-81. [doi: 10.1007/978-981-19-1983-1_3]
24. What is worst possible idea? Interaction Design Foundation. URL: https://www.interaction-design.org/literature/topics/worst-possible-idea [accessed 2023-10-02]
25. Ingram C, Langhans T, Perrotta C. Teaching design thinking as a tool to address complex public health challenges in public health students: a case study. BMC Med Educ 2022 Apr 12;22(1):270. [doi: 10.1186/s12909-022-03334-6] [Medline: 35413916]
26. Hamilton AB, Finley EP. Qualitative methods in implementation research: an introduction. Psychiatry Res 2019 Oct;280:112516. [doi: 10.1016/j.psychres.2019.112516] [Medline: 31437661]
27. What is brainstorming? 10 effective techniques you can use. nteraction Design Foundation. URL: https://www.interaction-design.org/literature/topics/brainstorming [accessed 2023-10-02]
28. Trzeciak S, Rivers EP. Emergency department overcrowding in the United States: an emerging threat to patient safety and public health. Emerg Med J 2003 Sep;20(5):402-405. [doi: 10.1136/emj.20.5.402] [Medline: 12954674]
29. Roy S, Keselman I, Nuwer M, Reider-Demer M. Fast neuro: a care model to expedite access to neurology clinic. Neurol Clin Pract 2022 Apr;12(2):125-130. [doi: 10.1212/CPJ.0000000000001152] [Medline: 35747888]
30. Yarbrough AJ, Johnson L, Vats A, Jaffee MS, Busl KM. Neurology access clinic: a model to improve access to neurologic care in an academic medical center. Neurol Clin Pract 2022 Jun;12(3):203-210. [doi: 10.1212/CPJ.0000000000001158] [Medline: 35747536]
31. Hill CE, Thomas B, Amendt A, Lawler M, Kalfin M. A multidisciplinary care model to increase access to neurologic ambulatory care. J Am Assoc Nurse Pract 2021;33(3):254-259. [doi: 10.1097/JXX.0000000000000435]
32. Kim SH, Myers CG, Allen L. Health care providers can use design thinking to improve patient experiences. : Harvard Business Review; 2017.

Abbreviations

ED: emergency department
ESU: Evaluation Sciences Unit
NPC: new patient coordinator
QI: quality improvement
RIF: Rapid Improvement Framework

Edited by A Mavragani; submitted 15.01.25; peer-reviewed by CJ Engstrom, K McCauley; revised version received 23.06.25; accepted 02.07.25; published 26.08.25.

Please cite as:
Harris SR, Rose C, Kling SMR, Cohen BA, Goldberg O, Walton H, Veruttipong D, Alhadha M, Laurence SK, Ravi S, Gold CA, Shaw JG, Yang L, Brown-Johnson CG
A Novel Approach to Care Redesign Collaboration Between Emergency and Specialty Departments: Qualitative Experience Report
JMIR Form Res 2025;9:e22028
URL: https://formative.jmir.org/2025/1/e22028
doi:10.2196/22028

© Sonia Rose Harris, Christian Rose, Samantha MR Kling, Brett A Cohen, Olga Goldberg, Hannah Walton, Darlene Veruttipong, Mohamed Alhadha, Sheneé K Laurence, Shashank Ravi, Carl A Gold, Jonathan G Shaw, Laurice Yang, Cati Brown-Johnson. Originally published in JMIR Formative Research (https://formative.jmir.org), 26.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Variability in the Diagnosis and Treatment of Interstitial Cystitis/Bladder Pain Syndrome: Internet Survey

Eleanora Melkonian[1]; A Leigh Garrett[2], PhD; Erika Kline[2], MS, PhD; Pressley Smith[3], MS; Madelyn Wiesenhahn[3], BA; Jeanette Petit[3], BS; Alison Swierczynski[3], BS; Carley Zhou[3], BS; Stuart B Bauer[4,5], MD; Rosalyn Adam[4,5], PhD; Kamil E Barbour[6], MPH, MS, PhD; Sonja I Ziniel[7], MA, PhD; Catherine A Brownstein[3,8], MPH, PhD

[1]APV-DTW LP, Bonita Springs, FL, United States
[2]Inspire, Arlington, VA, United States
[3]Division of Genetics and Genomics, Boston Children's Hospital, 300 Blackfan Street, Boston, MA, United States
[4]Department of Urology, Boston Children's Hospital, Boston, MA, United States
[5]Department of Surgery, Harvard Medical School, Boston, MA, United States
[6]Division of Population Health, Centers for Disease Control and Prevention, Atlanta, GA, United States
[7]Department of Pediatrics, University of Colorado School of Medicine, Aurora, CO, United States
[8]Department of Pediatrics, Harvard Medical School, Boston, MA, United States

Corresponding Author:
Catherine A Brownstein, MPH, PhD
Division of Genetics and Genomics, Boston Children's Hospital, 300 Blackfan Street, Boston, MA, United States

Abstract

Background: Interstitial cystitis/bladder pain syndrome (IC/BPS) is a complex, chronic condition affecting the urinary bladder. Symptoms commonly associated with IC/BPS include painful urination, pain during intercourse, a persistent or recurrent sensation of bladder discomfort or pressure that often worsens as the bladder fills and eases after urination, urgency, frequent urination with little warning, nighttime urination disrupting sleep, and burning or other unusual urinary sensations. These symptoms can profoundly impact emotional and mental health, hinder participation in daily activities, disrupt social interactions, and strain personal relationships.

Objective: This study aimed to compare the experiences of different races and ethnicities with IC/BPS regarding symptoms, diagnosis, treatment status, and treatment methods. We hypothesized that there would be differences in racial and ethnic minority groups.

Methods: A cross-sectional web-based survey was administered between June and August 2022 through the Interstitial Cystitis Association and the Inspire web-based health community. Eligible adults resided in the United States, self-reported IC/BPS symptoms, and completed the survey in English. The instrument gathered demographic information, details regarding age at symptom onset, formal diagnosis status, and treatment use. Validated symptom and problem indices (the O'Leary-Sant Interstitial Cystitis Symptom Index and Problem Index) captured symptom severity and quality-of-life impact. Comparative analyses, including Fisher exact and median tests, were conducted across racial or ethnic groups (minority or multiple-race vs White and Hispanic vs non-Hispanic), and multivariable logistic regression assessed predictors of race or ethnicity on IC/BPS diagnosis status and treatment outcomes.

Results: In total, 1631 individuals completed the survey. Racial or ethnic minority or multiple-race respondents constituted 11.6% (n=189) of the sample. Although IC/BPS symptom severity (Interstitial Cystitis Symptom Index or Interstitial Cystitis Problem Index scores) did not significantly differ by race or ethnicity, minority or multiple-race respondents were 50% less likely to have a formal medical diagnosis of IC/BPS than White respondents (adjusted odds ratio 0.50, 95% CI 0.30 - 0.83). Overall, 86.7% (n=1408) of participants reported having received a formal IC/BPS diagnosis, and the single strongest determinant of receiving any form of treatment was having a formal diagnosis (odds ratio 29.67, 95% CI 18.32 - 48.05). Over 25% (n=385) of all respondents reported using narcotic or opioid medications, indicating the challenging nature of IC/BPS symptom management.

Conclusions: Minority or multiple-race participants were significantly less likely to have ever been diagnosed with IC/BPS by a health care professional, and those who were not diagnosed with IC/BPS were less likely to have used self-care behavioral and nonpharmacological treatments for their symptoms. Streamlining the diagnostic process and public health awareness campaigns outlining treatment options may help individuals manage IC/BPS symptoms.

(JMIR Form Res 2025;9:e70813) doi:10.2196/70813

XSL•FO
RenderX

KEYWORDS

interstitial cystitis; painful bladder syndrome; social media; social networking; pain; treatment

Introduction

Interstitial cystitis/bladder pain syndrome (IC/BPS) is a multifactorial, chronic urinary bladder syndrome. Common symptoms of IC/BPS include dysuria, dyspareunia, urgency, a persistent or recurrent sense of bladder pain or pressure that is often exacerbated as the bladder fills and is temporarily relieved after voiding, frequent urination or with little warning, nocturia that disrupts sleep, and associated burning or other abnormal urinary sensations [1]. The nature of these symptoms can significantly impair emotional well-being and psychological health [2,3], participation in daily activities [4], social functioning [5], and personal relationships [6].

Because the key symptoms of interstitial cystitis overlap with many other, more common, diagnoses, it is difficult to diagnose IC/BPS [7]. The condition may remain a diagnosis of exclusion even after careful patient history, examination, and laboratory measures [8]. IC/BPS can have a nonspecific presentation that overlaps with other urological and gynecological disorders, such as urinary tract infections, endometriosis, and vulvodynia [9]. Male population with IC/BPS are often misdiagnosed as having overactive bladder or chronic prostatitis [10-12].

Treating IC/BPS requires a multimodal approach, as no single therapy is universally effective [13]. Treatments typically involve lifestyle modifications, dietary changes, stress management, and pelvic floor physical therapy, which can help some patients but may not provide sufficient relief for others. Medications such as oral pentosan polysulfate sodium [14], antihistamines, tricyclic antidepressants, and bladder instillations with agents like lidocaine or heparin are also used, though their effectiveness varies widely and can come with side effects [15]. Some patients benefit from neuromodulation techniques like sacral nerve stimulation, while others require more invasive options such as botulinum neurotoxin injections or bladder hydrodistention [15]. While certain treatments may provide symptom relief, none consistently cure the condition. The variability in treatment effectiveness underscores the need for personalized approaches and further research into more targeted therapies.

Racial and ethnic disparities may exist within the realm of the IC/BPS research. While minority populations encompass just under 40% of the total US population [16], their representation in studies [17], including those focused on IC/BPS [18], has historically been inadequate. Emerging evidence suggests a potential variation in the natural history of IC/BPS across racial and ethnic groups. For instance, Black individuals with IC/BPS report a greater degree of urinary urgency compared to their White counterparts and are more likely to be diagnosed at a younger age [19,20].

Given the evidence of racial and ethnic differences in experiences with IC/BPS [19,20], there is a need for additional information on demographic profiles, clinical characteristics, diagnostic experiences, and patterns of treatment among minority or multiple-race individuals with IC/BPS. Therefore, we sought to comprehensively assess a large web-based population of people diagnosed with IC/BPS or "suspected to have IC/BPS based on their symptoms" for their spectrum of experiences, including demographics, symptomatology, treatments received, and quality of life.

Methods

Study Design and Population

We administered a cross-sectional web-based survey, focusing on multiple IC/BPS-related domains, between June and August 2022. The survey was created and administered in collaboration with the Interstitial Cystitis Association (ICA) and Inspire. ICA is a leading national nonprofit organization dedicated to advocacy, education, and support for those affected by IC/BPS. Inspire is the world's largest web-based patient health community with over 45,000 self-identified individuals with IC/BPS.

To be eligible for study inclusion, respondents had to be members of the Inspire IC/BPS support group and discussion community or the ICA, be 18 years and older of age, be capable of responding to a web-based survey in English, be residing in the United States, and have self-reported symptoms potentially attributable to IC/BPS, regardless of having received a formal medical diagnosis of the condition. To encourage minority or multiple-race representation, recruitment materials clearly stated that minority or multiple-race individuals were invited to participate. Inspire members were invited by a message sent to their Inspire inbox. ICA members were invited by an email sent to their mailing list. Three reminder emails were sent, each 2 weeks apart.

Ethical Considerations

The study protocol and survey content were approved by the Institutional Review Board of Boston Children's Hospital (IRB 04-11-160). Individuals clicking through the informational page and completing the survey were considered to be consenting to the study. The participants received a US $25 e-gift card as compensation for completing the full survey. To prevent fraud, provided emails were compared to those on the mailing and membership lists, and data from unknown individuals were not included. In the case of duplicates, only the first completed survey was included in the analysis. Deidentified survey responses, with all personally identifiable information removed, are available for secondary data analysis.

Survey Details

The survey data collected included self-reported data across key areas: (1) demographics including current age, gender identity, race and ethnicity (measured separately), highest level of education completed, current employment status, health insurance coverage status and type over the past 12 months, and presence of a direct family member with an IC/BPS diagnosis; (2) age at IC/BPS symptom onset and whether the condition had been formally diagnosed (options were "diagnosed with IC/BPS by a health care professional" or "suspected

XSL•FO
RenderX

diagnosis"; the primary outcome); (3) treatment status and an inventory of the specific treatments ever used, categorized into "self-treatment" methods not requiring a prescription or clinical administration (eg, dietary modifications, over-the-counter products, and supplements) versus methods requiring a prescription or an in-office procedural treatment under medical supervision; and (4) current symptomatology using the validated O'Leary-Sant Interstitial Cystitis Symptom Index (ICSI), which grades symptom severity, as well as the corresponding Interstitial Cystitis Problem Index (ICPI), which measures the extent to which IC/BPS negatively impacts daily functional status and overall quality of life [21].

Regarding race and ethnicity, respondents were able to indicate with which of the following categories they identified: Asian, Black or African American, Hispanic or Latinx, Native American, Native Hawaiian or Pacific Islander, White, or other. Those indicating more than 1 racial category or identification with a racial category other than "White" were classified as minority or multiple-race for analysis purposes, whereas Hispanic or non-Hispanic ethnicity was recorded as a separate demographic category.

Analysis

Descriptive statistics were used to summarize the overall demographic and clinical characteristics of the survey sample and to compare the distributions of these variables across racial or ethnic minority or multiple-race versus White respondents, as well as Hispanic versus non-Hispanic respondents, using the Fisher exact test or nonparametric median tests, as appropriate. The mean ICSI and ICPI scores were also compared between the racial or ethnic and Hispanic groups using 2-tailed *t* tests with unequal variance. Treatment rates (both overall and categorized by self-treatment vs prescription or in-office treatment) were compared across racial or ethnic and Hispanic groups using the Fisher exact test. Finally, we conducted a series of multivariable logistic regression analyses to identify factors independently associated with several key outcome variables, including having received an IC/BPS diagnosis from a health care professional and having ever received any treatment for IC/BPS symptoms, having ever used self-treatment methods, having ever required prescription or in-office treatment, and being currently treated at the time of the survey. The covariates included in these regression models were age, gender, race or ethnicity, Hispanic ethnicity, education level, employment status, health insurance, diagnosis status, family history of IC/BPS, age of symptom onset, and ICPI and ICSI scores. All statistical analyses were performed using Stata (version 16.1; StataCorp) software. The Benjamini-Hochberg adjustment was used to correct for multiple testing, with a P value <.02 considered statistically significant.

Results

Sample Characteristics

A total of 23,121 individuals were invited to the survey. Of those, the email open rate was 19.5% (4509 emails). Of the opened emails, 1631 (36%) completed the survey and formed the survey cohort. The majority of the cohort (n=1434, 87.9%) were assigned female at birth. Within the cohort, 7.2% (n=118) reported being Hispanic persons, and 11.6% (n=189) reported being members of racial minority groups: 2.3% (n=37) Asian only, 6% (n=98) Black or African American only, 0.6% (n=9) Native American or Alaskan Native only, 0.4% (n=6) Native Hawaiian or Pacific Islander only, 0.5% (n=8) other only, and 1.9% (n=31) identifying as multiple races (Table 1).

XSL•FO
RenderX

Table . Demographic characteristics of respondents by race and ethnicity[a].

Characteristics	Overall	Race			Ethnicity		
		Minority or multiple-race	White	*P* value[b]	Hispanic	Non-Hispanic	*P* value[b]
Female (assigned gender at birth), n (%)	1434 (87.9)	132 (69.8)	1296 (90.3)	<.001	89 (75.4)	1345 (88.9)	<.001
Gender identity (n=1630), n (%)				<.001			<.001
Male	190 (11.7)	55 (29.1)	135 (9.4)		28 (23.7)	162 (10.7)	
Female	1420 (87.1)	129 (68.3)	1285 (89.6)		87 (73.7)	1333 (88.2)	
Nonbinary	11 (0.7)	1 (0.5)	10 (0.7)		1 (0.9)	10 (0.7)	
Transgender	8 (0.5)	4 (2.1)	4 (0.3)		1 (0.9)	7 (0.5)	
None of these describe me	1 (0.1)	0 (0)	1 (0.1)		1 (0.9)	0 (0)	
Age (years), median (IQR)	56 (42 - 67)	42 (34 - 53)	58 (44 - 68)	<.001	43 (35 - 55)	57 (43 - 67)	<.001
Hispanic ethnicity, n (%)	118 (7.2)	47 (24.9)	65 (4.5)	<.001	118 (100)	0 (0)	<.001
Race (multiple selection possible), n (%)				<.001			
Asian only	37 (2.3)	37 (19.6)	0 (0)		0 (0)	37 (2.5)	.18
Black only	98 (6)	98 (51.9)	0 (0)		37 (33)	61 (4)	<.001
Native American or Alaskan Native only	9 (0.6)	9 (4.8)	0 (0)		3 (2.7)	6 (0.4)	.02
Native Hawaiian or Pacific Islander only	6 (0.4)	6 (3.2)	0 (0)		0 (0)	6 (0.4)	≥.99
White only	1436 (88.4)	0 (0)	1436 (100)		65 (58)	1371 (90.6)	<.001
Other only	8 (0.5)	8 (4.2)	0 (0)		5 (4.5)	3 (0.2)	<.001
Multiple-race categories	31 (1.9)	31 (16.4)	0 (0)		2 (1.8)	29 (1.9)	≥.99
Minority or multiple-race (n=1625)	189 (11.6)	189 (100)	0 (0)		47 (42)	142 (9.4)	<.001
Education (n=1628), n (%)				<.001			<.001
Some high school or less	25 (1.5)	15 (7.9)	10 (0.7)		10 (8.6)	15 (1)	
High school graduate or GED[c]	175 (10.8)	27 (14.3)	148 (10.3)		19 (16.2)	156 (10.3)	
Vocational training	81 (5)	21 (11.1)	60 (4.2)		11 (9.4)	70 (4.6)	
Associate or technical degree	240 (14.7)	26 (13.8)	210 (14.7)		19 (16.2)	221 (14.6)	
Some college	239 (14.7)	28 (14.8)	211 (14.7)		12 (10.2)	227 (15)	
Bachelor's degree	472 (29)	44 (23.3)	428 (29.9)		27 (23.1)	445 (29.5)	
Master's degree	296 (18.2)	19 (10.1)	276 (19.3)		15 (12.8)	281 (18.6)	
Professional degree	41 (2.5)	7 (3.7)	33 (2.3)		2 (1.7)	39 (2.6)	

XSL•FO
RenderX

Characteristics	Overall	Race			Ethnicity		
		Minority or multiple-race	White	*P* value[b]	Hispanic	Non-Hispanic	*P* value[b]
Postgraduate degree	59 (3.6)	2 (1.1)	57 (4)		2 (1.7)	57 (3.8)	
Employment status (n=1578), n (%)				<.001			<.001
Full-time	508 (32.2)	75 (40.1)	431 (31.1)		51 (44)	457 (31.3)	
Part-time	218 (13.8)	33 (17.7)	183 (13.2)		20 (17.2)	198 (13.5)	
Furloughed or temporary layoff	11 (0.7)	3 (1.6)	8 (0.6)		3 (2.6)	8 (0.6)	
Maternity or medical leave	17 (1.1)	5 (2.7)	12 (0.9)		3 (2.6)	14 (1)	
Homemaker	111 (7)	14 (7.5)	97 (7)		7 (6)	104 (7.1)	
Student	21 (1.3)	3 (1.6)	18 (1.3)		1 (0.9)	20 (1.4)	
Unemployed on disability	145 (9.2)	22 (11.8)	122 (8.8)		8 (6.9)	137 (9.4)	
Unemployed	84 (5.3)	13 (7)	70 (5.1)		12 (10.3)	72 (4.9)	
Retired	463 (29.3)	19 (10.2)	444 (32.1)		11 (9.5)	452 (30.9)	
Health insurance for past 12 months (n=1498), n (%)	1431 (95.5)	148 (89.2)	1278 (96.3)	<.001	85 (87.6)	1346 (96.1)	.001
Health insurance (n=1543), n (%)				<.001			<.001
Through employer	679 (44)	71 (41.8)	605 (44.2)		49 (49.5)	630 (43.6)	
Purchase directly from insurance	165 (10.7)	41 (24.1)	124 (9.1)		18 (18.2)	147 (10.2)	
Medicare	467 (30.3)	24 (14.1)	443 (32.4)		12 (12.1)	455 (31.5)	
Medicaid	(124 (8)	27 (15.9)	97 (7.1)		14 (14.1)	110 (7.6)	
Tricare, VA, Indian Health Service, other	108 (7)	7 (4.1)	99 (7.2)		6 (6.1)	102 (7.1)	
Diagnosis of IC/BPS[d] by health care professional, n (%)	1408 (86.7)	143 (75.7)	1265 (88.1)	<.001	93 (78.8)	1321 (87.3)	.02
Age at symptom onset (n=1594) (years), n (%)				<.001			<.001
<30	547 (34.3)	63 (33.7)	482 (34.4)		43 (37.1)	504 (34.1)	
30 - 39	352 (22.1)	57 (30.5)	292 (20.8)		37 (31.9)	315 (21.3)	
40 - 49	279 (17.5)	40 (21.4)	239 (17.1)		24 (20.7)	255 (17.3)	
50 - 59	231 (14.5)	22 (11.8)	208 (14.9)		9 (7.8)	222 (15)	
60+	185 (11.6)	5 (2.7)	180 (12.9)		3 (2.6)	182 (12.3)	
Direct relative with IC/BPS diagnosis[e], n (%)	160 (9.8)	17 (9)	143 (10)	.80	11 (9.3)	149 (9.9)	≥.99
ICSI[f]				.08			.27
Values, n (%)	1411 (100)	175 (12.4)	1236 (87.6)		108 (7.7)	1303 (92.3)	
Mean (SD)	11.5 (4.12)	11 (3.57)	11.5 (4.19)		11.1 (3.78)	11.5 (4.14)	
ICPI[g]				.37			.55

XSL•FO
RenderX

Characteristics	Overall	Race			Ethnicity		
		Minority or multiple-race	White	*P* value[b]	Hispanic	Non-Hispanic	*P* value[b]
Values, n (%)	1551 (100)	184 (11.9)	1367 (88.1)		114 (7.4)	1437 (92.6)	
Mean (SD)	10.7 (3.58)	10.9 (3.24)	10.7 (3.62)		10.9 (3.28)	10.7 (3.60)	

[a]This table presents the demographic characteristics of the study population (n=1631), stratified by race and ethnicity. Key variables include gender identity, age, educational attainment, employment status, health insurance coverage, and diagnosis-related factors. Comparisons between racial and ethnic groups were conducted using Fisher exact tests for categorical variables and median tests for the continuous variable, age. Significant differences (*P* values) highlight variations across racial and ethnic groups, providing insights into disparities in demographics, socioeconomic factors, and health care access among respondents.

[b]Fisher exact tests were used for categorical variables, while a median test was used for the continuous variable, age.

[c]GED: general equivalency diploma.

[d]IC/BPS: interstitial cystitis/bladder pain syndrome.

[e]A direct relative was defined as a blood relative including parents, siblings, or children.

[f]ICSI: Interstitial Cystitis Symptom Index.

[g]ICPI: Interstitial Cystitis Problem Index.

Compared to White respondents, those classified as minority or multiple-race tended to be younger (median 42, IQR 34 - 53 years vs median 58, IQR 44 - 68 years; *P*<.001) and had lower levels of educational attainment (*P*<.001). Hispanic respondents reported an earlier age of initial symptom onset than their non-Hispanic counterparts (69% (n=80) vs 55.4% (n=819) before age 40 years; *P*<.001). No statistically significant differences were observed in the mean ICSI urinary symptom scores or ICPI quality-of-life impact scores across the racial or ethnic and Hispanic groups, suggesting a similar level of overall symptom burden and condition severity between the populations (Table 1).

Over 85% (n=1408, 86.7%) of respondents were diagnosed with IC/BPS by a health care professional; the remaining 13.3% (n=223) reported having a “suspected diagnosis” of IC/BPS.

Respondents having a “suspected diagnosis” of IC/BPS reported most of the classic symptoms associated with IC/BPS but to a lesser degree versus those diagnosed by a health care provider. Significantly more individuals diagnosed with IC/BPS self-reported currently having any IC/BPS symptoms (98.1%, n=1381) based on the ICPI than those with a suspected diagnosis (92.5%, n=206); *P*<.001): frequent urination during the day, getting up at night to urinate, needing to urinate with little warning, and burning, pain, discomfort, or pressure in the bladder. The mean and median ICSI were also significantly lower in those suspected of having IC/BPS than in those diagnosed: 10.31 (SD 3.52) versus 11.59 (SD 4.17) and 10 (IQR 8 - 13) versus 11 (IQR 9 - 15), respectively; both *P*<.001. The same was observed for the mean and median ICPI: 9.84 (SD 3.46) versus 10.85 (SD 3.58) and 10 (IQR 7 - 13) versus 11 (IQR 8 - 14), respectively; both *P*<.001.

Treatment Use

When queried about treatments received for IC/BPS symptoms, over 90% of respondents reported having used some form of therapy, with a median of 6 (IQR 4-10) treatment types tried by each participant (Table 2). However, treatment rates were significantly lower among racial or ethnic minority or multiple-race respondents than among White counterparts (n=151, 79.9% vs n=1317, 92.6%; *P*<.001) and among Hispanic versus non-Hispanic respondents (n=91, 77.8% vs n=1383, 92.1%; *P*<.001). Minority or multiple-race participants reported using fewer cumulative treatment types than their White and non-Hispanic counterparts: medians of 4 (IQR 3 - 7) versus 7 (IQR 4 - 10; minority or multiple-race vs White: *P*<.001) and 5 (IQR 3 - 8) versus 6 (IQR 4 - 10) types (Hispanic vs non-Hispanic: *P*=.003).

XSL•FO
RenderX

Table . Treatment use among respondents by treatment type, race, and ethnicity[a].

Treatment	Overall	By race (n=1612)			By ethnicity (n=1618)		
		Minority or multiple-race (n=151)	White (n=1423)	*P* value[b]	Hispanic (n=117)	Non-Hispanic (n=1501)	*P* value[b]
Any (n=1618), n (%)	1474 (91.1)	151 (79.9)	1317 (92.6)	<.001	91 (77.8)	1383 (92.1)	<.001
Number of treatment types ever used (n=1474), median (IQR)	6 (4-10)	4 (3-7)	7 (4-10)	<.001	5 (3-8)	6 (4-10)	.003
Treatment type categories (n=1474), n (%)							
Self-treatment	1313 (89.4)	127 (84.1)	1186 (90.1)	.04[c]	75 (82.4)	1243 (89.9)	.03[c]
Prescription or in-office treatment	1195 (90.9)	140 (92.7)	1195 (90.7)	.55	83 (91.2)	1257 (90.9)	≥.99
Treatment types (n=1474), n (%)[d]							
IC[e] diet, modified or elimination diet	1079 (73.2)	79 (52.3)	995 (75.6)	<.001	54 (59.3)	1025 (74.1)	.003
Over-the-counter products	780 (52.9)	38 (25.2)	737 (56)	<.001	38 (41.8)	742 (53.7)	.03[c]
Anti-inflammatory medications	739 (50.1)	68 (45)	669 (50.8)	.20	43 (47.3)	696 (50.3)	.59
Antidepressants	731 (49.6)	44 (29.1)	683 (51.9)	<.001	37 (40.7)	694 (50.2)	.08
Bladder instillations	696 (47.2)	50 (33.1)	642 (48.8)	<.001	40 (44)	656 (47.4)	.59
Antihistamines	647 (43.9)	47 (31.1)	597 (45.3)	.001	27 (29.7)	620 (44.8)	.005
Pelvic floor PT[f]	618 (41.9)	45 (29.8)	569 (43.2)	.002	33 (36.3)	585 (42.3)	.28
Bladder hydrodistention	526 (35.7)	36 (23.8)	487 (37)	.001	23 (25.3)	503 (36.4)	.03[c]
Pentosan polysulfate sodium	521 (35.4)	35 (23.2)	483 (36.7)	.001	25 (27.5)	496 (35.9)	.11
Nonnarcotic pain medication	500 (33.9)	26 (17.2)	472 (35.8)	<.001	18 (19.8)	482 (34.9)	.003
Supplements	464 (31.5)	24 (15.9)	438 (33.3)	<.001	24 (26.4)	440 (31.8)	.30
Narcotics or opioids	385 (26.1)	33 (21.9)	350 (26.6)	.24	16 (17.6)	369 (26.7)	.06
Topical medications	365 (24.8)	38 (25.2)	327 (24.8)	.92	17 (18.7)	348 (25.2)	.21
Histamine blockers	364 (24.7)	28 (18.5)	336 (25.5)	.07	15 (16.5)	349 (25.2)	.06
Antispasmodics	318 (21.6)	28 (18.5)	289 (21.9)	.40	20 (22)	298 (21.6)	.90
Acupuncture	279 (18.9)	26 (17.2)	253 (19.2)	.66	15 (16.5)	264 (19.1)	.68
Cannabinoids	266 (18.1)	24 (15.9)	242 (18.4)	.50	8 (8.8)	258 (18.7)	.02
Biofeedback	145 (9.8)	15 (9.9)	130 (9.9)	≥.99	7 (7.7)	138 (10)	.59

XSL•FO
RenderX

Treatment	Overall	By race (n=1612)			By ethnicity (n=1618)		
		Minority or multiple-race (n=151)	White (n=1423)	*P* value[b]	Hispanic (n=117)	Non-Hispanic (n=1501)	*P* value[b]
Neurostimulation	140 (9.5)	14 (9.3)	126 (9.6)	≥.99	8 (8.8)	132 (9.5)	≥.99
Botulinum neurotoxin	139 (9.4)	17 (11.3)	122 (9.3)	.46	6 (6.6)	133 (9.6)	.49
Percutaneous tibial nerve stimulation	132 (9)	14 (9.3)	118 (9)	.88	8 (8.8)	124 (9)	≥.99
Laser	117 (7.9)	12 (8)	105 (8)	≥.99	9 (9.9)	108 (7.8)	.43
Surgery	102 (6.9)	10 (6.6)	92 (7)	≥.99	5 (5.5)	97 (7)	.83
Triamcinolone	98 (6.7)	13 (8.6)	85 (6.5)	.30	9 (9.9)	89 (6.4)	.19
Immunosuppressants	52 (3.5)	13 (8.6)	39 (3)	.002	3 (3.3)	49 (3.5)	≥.99
None of these	6 (0.4)	1 (0.7)	5 (0.4)	.48	1 (1.1)	5 (0.4)	.32
Other	270 (18.3)	12 (8)	254 (19.3)	<.001	18 (19.8)	252 (18.2)	.68
Currently treated (n=1618), n (%)	1015 (62.7)	112 (59.3)	897 (63)	.34	60 (51.3)	955 (63.6)	.01

[a]This table summarizes the treatment use patterns among study participants (n=1618), stratified by race and ethnicity. The table presents the proportion of respondents who reported ever using any treatment, the median number of treatment types used, and the distribution of specific treatment modalities, including self-treatment, prescription or in-office treatments, and various pharmacologic and nonpharmacologic interventions. Treatment types are listed in descending order of overall use. Comparisons between racial and ethnic groups were conducted using Fisher exact tests for categorical variables and median tests for continuous variables. Statistically significant *P* values indicate disparities in treatment access and use across demographic groups. The Benjamini-Hochberg adjustment was applied to account for multiple comparisons. These findings highlight potential differences in treatment access and preferences among diverse patient populations.

[b]Fisher exact tests were used for categorical variables, while a median test was used for the continuous variable, age.

[c]Treatment types are ordered from most common to least common among the overall sample population.

[d]This *P* value is no longer statistically significant due to the new corrected significance level (*P*<.02) after the Benjamini-Hochberg adjustment for multiple testing.

[e]IC: interstitial cystitis.

[f]PT: physical therapy.

Of all respondents, 89.4% (n=1313) indicated using behavioral and nonpharmacological treatments, such as dietary modifications and over-the-counter products, which were self-implemented. Likewise, 90.9% (n=1195) reported using treatments requiring clinical supervision such as prescription medications, instillations, pelvic floor physical therapy, interventional procedures, and other provider-administered therapies.

Significant differences in treatment rates for minority or multiple-race versus White respondents were observed for the use of dietary modifications such as elimination or restrictive IC/BPS diets (n=79, 52.3% vs n=995, 75.6%; *P*<.001), over-the-counter products (n=38, 25.2% vs n=737, 56%; *P*<.001), antidepressants (n=44, 29.1% vs n=683, 51.9%; *P*<.001), bladder instillations (n=50, 33.1% vs n=642, 48.8%; *P*<.001), antihistamines (n=47, 31.1% vs n=597, 45.3%; *P*=.001), pelvic floor physical therapy (n=45, 29.8% vs n=569, 43.2%; *P*=.002), bladder hydrodistention (n=36, 23.8% vs n=487, 37%; *P*=.001), pentosan polysulfate sodium (n=35, 23.2% vs n=483, 36.7%; *P*=.001), nonnarcotic pain medications (n=26, 17.2% vs n=472, 35.8%; *P*<.001), and nutraceuticals and other supplements (n=24, 15.9% vs n=438, 33.3%; *P*<.001). Minority or multiple-race respondents reported significantly higher rates of immunosuppressant use than White respondents (n=13, 8.6% vs n=39, 3%; *P*=.002).

Significant differences in treatment rates for Hispanic versus non-Hispanic respondents were observed for the use of dietary modifications, such as elimination or restrictive IC/BPS diets (n=54, 59.3% vs n=1025, 74.1%; *P*<=.003), antihistamines (n=27, 29.7% vs n=620, 44.8%; *P*=.005), and nonnarcotic pain medications (n=18, 19.8% vs n=482, 34.9%; *P*=.003). There was also a significant difference in those being currently treated, among Hispanic and non-Hispanic respondents (n=60, 51.3% vs n=955, 63.6%; *P*=.01).

No significant difference was identified in the use of narcotics or opioids, with 26.1% (n=385) of surveyed individuals reporting using them, and no significant difference by race (minority or multiple-race: 21.9% (n=33) vs White: 26.6% (n=350); *P*=.24) or ethnicity (Hispanic: 17.6% (n= 16) vs non-Hispanic: 26.7% (n=369); *P*=.06)

Multivariable Analyses

After controlling for age, gender, socioeconomic status, and other relevant factors, the odds of having been diagnosed with

XSL•FO
RenderX

IC/BPS by a health care professional were 50% lower among racial or ethnic minority or multiple-race respondents compared to their White counterparts (odds ratio [OR] 0.50, 95% CI 0.30 - 0.83; Table 3). A similar magnitude difference in diagnostic likelihood was observed between Hispanic and non-Hispanic respondents in adjusted analyses but did not reach statistical significance (OR 0.56, 95% CI 0.31 - 1.03). Other factors associated with higher odds of having received a diagnosis included assigned female at birth (OR 2.87, 95% CI 1.82 - 4.55 vs male) and an earlier age of symptom onset. Those experiencing initial symptoms at an older age reported having a lower likelihood of diagnosis compared to those with onset by age 29 years (ORs ranging from 0.01 for onset at age 60+ years to 0.24 for onset at age 30 - 39 years).

XSL•FO
RenderX

Table . Multivariable logistic regression models predicting interstitial cystitis/bladder pain syndrome (IC/BPS) diagnosis and treatment use[a].

	Diagnosed with IC/BPS[b]		Ever been treated in any way[c]		Ever self-treated[c]		Ever received prescription or in-office treatment[c]		Currently treated[d]	
	OR[e] (95% CI)	*P* value	OR (95% CI)	*P* value	OR (95% CI)	*P* value	OR (95% CI)	*P* value	OR (95% CI)	*P* value
Hispanic (comparison group: non-Hispanic)	0.56 (0.31 - 1.03)	.06	0.64 (0.30 - 1.33)	.23	0.54 (0.29 - 1.01)	.05	0.88 (0.53 - 2.12)	.75	0.75 (0.46 - 1.25)	.27
Minority or multiple-race (comparison group: White)	0.50 (0.30 - 0.83)	.008	0.93 (0.51 - 1.71)	.82	0.61 (0.36 - 1.03)	.06	1.06 (0.53 - 2.12)	.86	1.12 (0.74 - 1.69)	.60
Female (gender assigned at birth; comparison group: male)	2.87 (1.82 - 4.55)	.001	2.24 (1.31 - 3.83)	.003	1.88 (1.14 - 3.10)	.01	1.06 (0.65 - 2.16)	.85	1.35 (0.91 - 2.00)	.14
Age (years)	1.08 (1.05 - 1.11)	<.001	1.00 (0.99 - 1.02)	.68	0.98 (0.97 - 1.00)	.01	0.88 (0.38 - 2.00)	.11	0.98 (0.97 - 1.00)	.007
Education (comparison group: high school graduate or GED[f] or less)										
Vocational training, associate or technical degree, some college	1.08 (0.57 - 2.05)	.82	1.24 (0.66 - 2.36)	.51	1.21 (0.71 - 2.04)	.48	0.95 (0.51 - 1.76)	.86	1.04 (0.69 - 1.57)	.86
Bachelor's degree	0.95 (0.49 - 1.87)	.89	1.26 (0.65 - 2.46)	.50	1.69 (0.96 - 2.99)	.07	1.41 (0.73 - 2.74)	.31	1.51 (0.98 - 2.33)	.06
Master's degree or higher	0.74 (0.37 - 1.48)	.40	3.34 (1.51 - 7.39)	.003	2.07 (1.13 - 3.78)	.02	1.46 (0.74 - 2.89)	.28	1.42 (0.91 - 2.23)	.13
Current employment status (comparison group: full-time employed)										
Part-time employed	—[g]	—	—	—	—	—	—	—	0.65 (0.44 - 0.97)	.03[h]
Retired	—	—	—	—	—	—	—	—	0.93 (0.62 - 1.39)	.71
Unemployed	—	—	—	—	—	—	—	—	0.54 (0.31 - 0.94)	.03[h]
Medical leave or disability	—	—	—	—	—	—	—	—	1.14 (0.71 - 1.82)	.59
Homemaker or student	—	—	—	—	—	—	—	—	0.43 (0.27 - 0.69)	<.001

XSL•FO
RenderX

	Diagnosed with IC/BPS[b]		Ever been treated in any way[c]		Ever self-treated[c]		Ever received prescription or in-office treatment[c]		Currently treated[d]	
	OR[e] (95% CI)	*P* value	OR (95% CI)	*P* value	OR (95% CI)	*P* value	OR (95% CI)	*P* value	OR (95% CI)	*P* value
Continuous health insurance during past 12 months	—	—	—	—	—	—	—	—	1.08 (0.60 - 1.94)	.79
Diagnosed with IC/BPS by health care provider	—	—	29.67 (18.32 - 48.05)	<.001	2.36 (1.10 - 5.06)	.03[f]	3.99 (1.95 - 8.16)	<.001	5.55 (3.41 - 9.04)	<.001
Direct relative with IC/BPS diagnosis	1.80 (0.84 - 3.89)	.13	—	—	—	—	—	—	—	—
Age at symptom onset (years) (comparison group:<30 years)										
30 - 39	0.24 (0.12 - 0.46)	<.001	—	—	—	—	—	—	—	—
40 - 49	0.09 (0.04 - 0.20)	<.001	—	—	—	—	—	—	—	—
50 - 59	0.09 (0.03 - 0.26)	<.001	—	—	—	—	—	—	—	—
60+	0.01 (0.004 - 0.05)	<.001	—	—	—	—	—	—	—	—
ICPI[b,i]	—	—	—	—	—	—	—	—	1.02 (0.97 - 1.08)	.48

XSL•FO
RenderX

	Diagnosed with IC/BPS[b]		Ever been treated in any way[c]		Ever self-treated[c]		Ever received prescription or in-office treatment[c]		Currently treated[d]	
	OR[e] (95% CI)	*P* value	OR (95% CI)	*P* value	OR (95% CI)	*P* value	OR (95% CI)	*P* value	OR (95% CI)	*P* value
ICSI[b,j]	—	—	—	—	—	—	—	—	1.01 (0.96 - 1.06)	.79

[a]This table presents the results of multivariable logistic regression models examining the associations between demographic factors and the likelihood of being diagnosed with IC/BPS as well as different forms of treatment use. Odds ratios (ORs) with 95% CIs are reported for key predictors, including race, ethnicity, gender assigned at birth, age, education, employment status, insurance coverage, family history of IC/BPS, and age at symptom onset. The table includes five regression models: (1) diagnosis of IC/BPS by a health care provider, (2) ever having received any treatment, (3) self-treatment, (4) prescription or in-office treatment, and (5) current treatment. Significant findings indicate disparities in IC/BPS diagnosis and treatment use across racial and ethnic groups as well as differences based on socioeconomic and health care access factors. Notably, being of minority or multiple-race status was associated with lower odds of receiving an IC/BPS diagnosis compared to White respondents. Female gender and older age were significant predictors of diagnosis, while higher educational attainment was associated with increased odds of treatment receipt. Additional models adjusted for symptom severity (ICPI and ICSI indices) and current life circumstances, such as employment and insurance status, in predicting current treatment use. The Benjamini-Hochberg adjustment for multiple testing was applied, and adjusted *P* values are reported where applicable. These findings provide critical insights into potential disparities in IC/BPS diagnosis and treatment access, emphasizing the need for further investigation into barriers to care for underrepresented populations.

[b]The logistic regression predicting diagnosis with IC/BPS by a health care provider includes, besides ethnicity and race, other demographic characteristics such as gender assigned at birth, current age in years, as well as educational status. We hypothesized that having an immediate family member with an IC/BPS diagnosis and age at symptom onset influenced the likelihood of being diagnosed. Measures that reflect the current status of IC/BPS symptoms, such as the ICPI and the ICSI, were not included in this model as predictive variables because the diagnosis of IC/BPS occurred longer than 12 months ago. This applies also to variables describing the current employment status of the respondent as well as having had continuous health insurance in the past 12 months.

[c]The logistic regression models predicting if respondents ever received treatment in the past, if they self-treated in the past, and if they received prescription or in-office treatments in the past included the same demographic predictor variables as the logistic regression model predicting diagnosis: ethnicity, race, gender assigned at birth, current age in years, and education. In addition, we hypothesized that having a diagnosis by a health care provider would influence the receipt of treatment and therefore added this as a predictive variable to the model.

[d]The logistic model predicting if respondents receive current treatment should also be influenced by current symptom measures as well as current life circumstances. In addition to the demographic predictor variables from the previous models, such as ethnicity, race, gender assigned at birth, current age in years, and education, we added current employment status and having had continuous health insurance in the past 12 months. Besides having been diagnosed with IC/BPS by a health care provider, we added 2 current symptom measures, ICPI and ICSI, since worse symptoms should increase the likelihood of current treatment.

[e]OR: odds ratio.

[f]GED: general equivalency diploma.

[g]Not available.

[h]This *P* value is no longer statistically significant due to the new corrected significance level ($P<.02$) after the Benjamini-Hochberg adjustment for multiple testing.

[i]ICPI: Interstitial Cystitis Problem Index.

[j]ICSI: Interstitial Cystitis Symptom Index.

Having received a formal medical diagnosis of IC/BPS emerged as the strongest predictor of whether a participant had ever received any form of treatment for IC/BPS symptoms (OR 29.67, 95% CI 18.32 - 48.05). Race and ethnicity were not independent predictors of treatment status in these models. However, those assigned female gender at birth had significantly higher odds than male gender of having ever received treatment (OR 2.24, 95% CI 1.31 - 3.83) as well as having used self-treatment methods (OR 1.88, 95% CI 1.14 - 3.10).

Self-reporting as a homemaker or student was associated with a significantly lower likelihood of currently receiving treatment compared to respondents employed full-time (OR 0.43, 95% CI 0.27 - 0.69). Higher levels of educational attainment correlated with a greater likelihood of receiving treatment: compared to those with a high school education or less, those with a master's degree or higher had 2.07 times higher odds of ever using self-treatment methods (95% CI 1.13 - 3.78) and 3.34 times higher odds of having received any form of treatment (95% CI 1.51 - 7.39).

Discussion

Principal Findings

This large survey included individuals with diagnosed or suspected IC/BPS and documented their experiences across diagnostic and therapeutic landscapes. A central finding was that individuals who had IC/BPS symptoms and who identified as a racial or ethnic minority were substantially less likely to report having received a formal medical IC/BPS diagnosis from a health care provider than non-Hispanic White respondents, even after accounting for multiple potential confounding factors such as age, gender, age at symptom onset, educational

XSL•FO
RenderX

attainment, and family history. This disparity in the rates of receiving a diagnosis appeared to drive other inequities observed in access to treatments requiring physician prescription or administration. Specifically, once the diagnostic status was controlled for in the multivariable analyses, racial and ethnic minorities were no longer statistically significantly less likely to receive any form of treatment for IC/BPS symptoms.

The survey data suggest that a major barrier to equal receipt of care is not inherently tied to race or ethnicity per se but rather stems from the reduced rates of diagnosis among minority patient populations. The definitive reasons underlying these diagnostic disparities could not be directly ascertained from this study but could possibly be due to individuals with suspected diagnoses reporting somewhat less severe symptomology as we observed, which may result in a lack of referral to a specialist or clinician who is familiar with the diagnosis of IC/BPS or other difficulties navigating the US health care system. Diagnostic disparities may also be due to a multitude of other factors, such as unconscious biases influencing medical decision-making [22-24], racism [25], poor communication [26], or differing abilities in advocating for health care needs [27]. Provider disbelief and pain dismissal have been reported in the female population diagnosed with IC/BPS [28], and people from racial and ethnic minority groups are less likely to have their pain taken seriously or treated [29,30], suggesting that this may also play a role. Regardless of the cause, this diagnostic gap represents the importance of intervention to achieve more equitable access to care across all patient demographic groups.

Compounding the issue of reduced diagnosis rates among minority or multiple-race individuals, this study also revealed gaps in the use of behavioral and nonpharmacologic treatments for IC/BPS. Compared to White and non-Hispanic individuals, minority or multiple-race and Hispanic participants reported lower rates of using several forms of self-treatment that could be implemented in an accessible, low-risk manner, such as dietary modifications, over-the-counter analgesics, anti-inflammatory agents, antihistamines, and nutritional supplements. Similarly, minority or multiple-race participants reported lower use rates of some prescription medications, such as antidepressants and pentosan polysulfate, as well as reduced exposure to commonly used interventional therapies, such as bladder instillations, hydrodistention, and pelvic floor physical therapy. Differences in the use of interventional therapies may be due to insurance coverage for these interventions; insurance type varied significantly between minority or multiple-race and White survey respondents and Hispanic and non-Hispanic respondents.

However, these discrepancies across a range of therapeutic options, many of which do not necessarily require formal IC/BPS diagnosis, highlight the need to improve awareness of and access to evidence-based treatment guidelines [1,8]. A recent study reported that the female population with IC/BPS has a strong interest in guided programs that teach self-care practices [31]. Campaigns could focus on easy-to-adopt therapies that have the potential to mitigate symptom burden prior to pursuing more invasive approaches [8].

The finding that over 25% (n=385) of the individuals who responded to the survey are using narcotic medications was striking and indicative of the persistent pain of this condition. Many patients remain symptomatic or undertreated (as evidenced by the ICSI/ICPI scores), despite multimodal therapy. Additionally, though there were meaningful differences in treatment use, there was no difference in ICSI/ICPI scores by race or ethnicity, possibly reflecting that treatments are not that effective overall.

Limitations

Limitations of this study include the potential for biases inherent to the recruitment of participants via web-based surveys—such as recall and self-selection bias, which could limit the generalizability of the findings to the broader IC/BPS patient population. There is also the possibility that participants may have subjectively interpreted and responded to questions in a manner that introduced inconsistencies compared to objective clinical assessments. The survey did not ask any questions that can be a proxy for poor access to care or providers who may diagnose interstitial cystitis, such as having a primary care provider or getting routine screening. Additionally, despite targeted efforts to oversample racial and ethnic minority or multiple-race groups, the overall survey population was still underrepresented in these patient populations compared to the US demographic composition, which may have affected the power to detect differences between groups. Due to the small sample size, the ethnic minority or multiple-race groups were categorized together for analyses, which may have missed differences that would have been revealed by a more nuanced approach. The minority or multiple-race individuals were also younger than the White respondents, which may have impacted the results. While the age difference is consistent with the finding that Black and Hispanic individuals are more likely to receive a diagnosis of IC/BPS at a younger age [20], it may also be due to older minority or multiple-race individuals being less likely to use digital health information or access the internet [32].

Finally, because the survey was administered to all individuals self-identifying as having IC/BPS in a web-based community regardless of diagnosis status, it is possible that there was some degree of heterogeneity or imprecision in the extent to which undiagnosed respondents fully met the established clinical criteria for IC/BPS. However, IC/BPS is commonly misdiagnosed [7,20], indicating that restricting the sample to diagnosed individuals would not eliminate heterogeneity and imprecision. In addition, the “suspected diagnosis” respondents were more likely to be minority or multiple-race, so eliminating “suspected diagnosis” from the analyses would greatly reduce the diversity of individuals included in the study. It is important to note that the surveyed individuals with “suspected diagnoses” were still symptomatic, despite having reported milder symptoms. For example, the mean ICSI score of 10.3 (SD 3.52) for those with “suspected diagnosis” means that these individuals reported strong need to urinate about half the time, needing to urinate more than once per 2 hours, getting up 2 or more times a night, or experiencing pain or burning in their bladder fairly often.

XSL•FO
RenderX

Conclusions and Future Directions

In this study of over 1600 respondents across diverse racial and ethnic backgrounds, a key disparity identified was the reduced likelihood of individuals who identified as racial or ethnic minority or multiple-race to have received a formal medical diagnosis of their condition compared to White individuals, even after controlling for multiple potentially confounding sociodemographic and clinical factors. This diagnostic gap appears to be a driver of downstream disparities in access to guideline treatments and therapies requiring clinical supervision or prescription. The findings indicate a need to focus on quality improvement efforts to facilitate accurate and timely IC/BPS diagnosis across all patient populations, regardless of race, ethnicity, or demographic characteristics. Initiatives to raise awareness of guideline-recommended and self-treatment approaches may reduce the identified disparities in the use of appropriate evidence-based therapies. Additional investigations into understanding the interplay between demographic factors and diagnosis are warranted.

Acknowledgments

The authors are grateful to the participants and their families. The authors also thank their colleagues at the Boston Children's Hospital (in particular Dr Richard Lee and Dr Alan Beggs). The authors thank Dr Kurt Greenlund for valuable advice and mentorship. This publication was supported by the Centers for Disease Control and Prevention (CDC) of the US Department of Health and Human Services (HHS) as part of a financial assistance award totaling US $2,400,000 with 100% funding from the CDC/HHS (U01DP006634-01). The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the CDC.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

The design and implementation of the research were carried out by CAB, EM, and ALG. PS, MW, JP, EK, AS, CZ, KEB, and SIZ contributed to the analysis of the results. SBB, RA, and KEB were responsible for drafting the manuscript, while CAB provided overall supervision. All authors reviewed the final manuscript.

Conflicts of Interest

None declared.

References

1. Lim Y, Leslie SW, O'Rourke S. Interstitial Cystitis/Bladder Pain Syndrome: StatPearls Publishing; 2025.
2. Crawford A, Tripp DA, Nickel JC, et al. Depression and helplessness impact interstitial cystitis/bladder pain syndrome pain over time. Can Urol Assoc J 2019 Oct;13(10):328-333. [doi: 10.5489/cuaj.5703] [Medline: 31364973]
3. McKernan LC, Walsh CG, Reynolds WS, Crofford LJ, Dmochowski RR, Williams DA. Psychosocial co-morbidities in interstitial cystitis/bladder pain syndrome (IC/BPS): a systematic review. Neurourol Urodyn 2018 Mar;37(3):926-941. [doi: 10.1002/nau.23421] [Medline: 28990698]
4. Cunningham AR, Gu L, Dubinskaya A, et al. Quality-of-life impact of interstitial cystitis and other pelvic pain syndromes. Front Pain Res (Lausanne) 2023;4(1149783):1149783. [doi: 10.3389/fpain.2023.1149783] [Medline: 37305204]
5. Kirkham A, Swainston K. Women's experiences of interstitial cystitis/painful bladder syndrome. West J Nurs Res 2022 Feb;44(2):125-132. [doi: 10.1177/0193945921990730] [Medline: 33514302]
6. Tonyali S, Yilmaz M. Sexual dysfunction in interstitial cystitis. Curr Urol 2017 Nov;11(1):1-3. [doi: 10.1159/000447186] [Medline: 29463969]
7. Hanno PM, Erickson D, Moldwin R, Faraday MM, American Urological Association. Diagnosis and treatment of interstitial cystitis/bladder pain syndrome: AUA guideline amendment. J Urol 2015 May;193(5):1545-1553. [doi: 10.1016/j.juro.2015.01.086] [Medline: 25623737]
8. Clemens JQ, Erickson DR, Varela NP, Lai HH. Diagnosis and treatment of interstitial cystitis/bladder pain syndrome. J Urol 2022 Jul;208(1):34-42. [doi: 10.1097/JU.0000000000002756]
9. Dell JR, Mokrzycki ML, Jayne CJ. Differentiating interstitial cystitis from similar conditions commonly seen in gynecologic practice. Eur J Obstet Gynecol Reprod Biol 2009 Jun;144(2):105-109. [doi: 10.1016/j.ejogrb.2009.02.050]
10. Forrest JB, Schmidt S. Interstitial cystitis, chronic nonbacterial prostatitis and chronic pelvic pain syndrome in men: a common and frequently identical clinical entity. J Urol 2004 Dec;172(6 Pt 2):2561-2562. [doi: 10.1097/01.ju.0000145102.40773.43] [Medline: 15538208]
11. Clemens JQ, Clauw DJ, Kreder K, et al. Comparison of baseline urological symptoms in men and women in the MAPP research cohort. J Urol 2015 May;193(5):1554-1558. [doi: 10.1016/j.juro.2014.11.016] [Medline: 25463989]

XSL•FO
RenderX

12. Arora HC, Shoskes DA. The enigma of men with interstitial cystitis/bladder pain syndrome. Transl Androl Urol 2015 Dec;4(6):668-676. [doi: 10.3978/j.issn.2223-4683.2015.10.01] [Medline: 26813678]
13. Beltran E. Female pelvic conditions: interstitial cystitis/bladder pain syndrome. FP Essent 2024 Dec;547(33-9):33-39. [Medline: 39692795]
14. Grigoryan B, Kasyan G, Pivazyan L, Pushkar D. Pentosan polysulfate in patients with bladder pain syndrome/interstitial cystitis with Hunner's lesions or glomerulations: systematic review and meta-analysis. Ther Adv Urol 2022;14(17562872221102809):17562872221102809. [doi: 10.1177/17562872221102809] [Medline: 35677571]
15. Cacciatore L, Territo A, Minore A, Testa A, Mantica G, Esperto F. Bladder pain syndrome (BPS): a comprehensive review of treatment strategies and management approaches. Res Rep Urol 2024;16:273-282. [doi: 10.2147/RRU.S387749] [Medline: 39479186]
16. Improved race and ethnicity measures reveal U.S. population is much more multiracial. United States Census Bureau. 2021 Jul 1. URL: https://www.census.gov/library/stories/2021/08/improved-race-ethnicity-measures-reveal-united-states-population-much-more-multiracial.html [accessed 2025-05-17]
17. National Academies of Sciences, Engineering, and Medicine. In: Bibbins-Domingo K, Helman A, editors. Improving Representation in Clinical Trials and Research: Building Research Equity for Women and Underrepresented Groups: The National Academies Press; 2022.
18. Sant GR. Interstitial cystitis in minority women. J Assoc Acad Minor Phys 1993;4(3):89-92. [Medline: 8369621]
19. Jarman A, Janes JL, Shorter B, et al. Food sensitivities in a diverse nationwide cohort of veterans with interstitial cystitis/bladder pain syndrome. J Urol 2023 Jan;209(1):216-224. [doi: 10.1097/JU.0000000000002938] [Medline: 36001744]
20. Dallas KB, Bresee C, De Hoedt A, et al. Demographic differences and disparities in the misdiagnosis of interstitial cystitis/bladder pain syndrome in a national cohort of VA patients. Urology 2022 May;163(22-8):22-28. [doi: 10.1016/j.urology.2021.07.019] [Medline: 34348123]
21. O'Leary MP, Sant GR, Fowler FJ, Whitmore KE, Spolarich-Kroll J. The interstitial cystitis symptom index and problem index. Urology 1997 May;49(5A Suppl):58-63. [doi: 10.1016/s0090-4295(99)80333-1] [Medline: 9146003]
22. Shah HS, Bohlen J. Implicit Bias: StatPearls Publishing; 2024.
23. Hall WJ, Chapman MV, Lee KM, et al. Implicit racial/ethnic bias among health care professionals and its influence on health care outcomes: a systematic review. Am J Public Health 2015 Dec;105(12):e60-e76. [doi: 10.2105/AJPH.2015.302903] [Medline: 26469668]
24. Gopal DP, Chetty U, O'Donnell P, Gajria C, Blackadder-Weinstein J. Implicit bias in healthcare: clinical practice, research and decision making. Future Healthc J 2021 Mar;8(1):40-48. [doi: 10.7861/fhj.2020-0233] [Medline: 33791459]
25. Hamed S, Bradby H, Ahlberg BM, Thapar-Björkert S. Racism in healthcare: a scoping review. BMC Public Health 2022 May 16;22(1):988. [doi: 10.1186/s12889-022-13122-y] [Medline: 35578322]
26. Cooper LA, Roter DL, Carson KA, et al. The associations of clinicians' implicit attitudes about race with medical visit communication and patient ratings of interpersonal care. Am J Public Health 2012 May;102(5):979-987. [doi: 10.2105/AJPH.2011.300558] [Medline: 22420787]
27. Wiltshire J, Cronin K, Sarto GE, Brown R. Self-advocacy during the medical encounter: use of health information and racial/ethnic differences. Med Care 2006 Feb;44(2):100-109. [doi: 10.1097/01.mlr.0000196975.52557.b7] [Medline: 16434908]
28. Brown VL, James A, Hunleth J, et al. Believing women: a qualitative exploration of provider disbelief and pain dismissal among women with interstitial cystitis/bladder pain syndrome from the MAPP research network. Int Urogynecol J 2024 Jan;35(1):139-148. [doi: 10.1007/s00192-023-05677-0] [Medline: 37991567]
29. Hoffman KM, Trawalter S, Axt JR, Oliver MN. Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites. Proc Natl Acad Sci U S A 2016 Apr 19;113(16):4296-4301. [doi: 10.1073/pnas.1516047113] [Medline: 27044069]
30. Baker TA, Booker SQ, Janevic MR. A progressive agenda toward equity in pain care. Health Psychol Behav Med 2023;11(1):2266221. [doi: 10.1080/21642850.2023.2266221] [Medline: 37818413]
31. Hassani D, Flick L, Sangha H, Brown LA, Andy U, Arya L. How do women with interstitial cystitis/bladder pain syndrome make treatment choices? Int Urogynecol J 2022 Mar;33(3):583-593. [doi: 10.1007/s00192-021-04994-6] [Medline: 34599672]
32. Tappen RM, Cooley ME, Luckmann R, Panday S. Digital health information disparities in older adults: a mixed methods study. J Racial Ethn Health Disparities 2022 Feb;9(1):82-92. [doi: 10.1007/s40615-020-00931-3] [Medline: 33415705]

Abbreviations

IC/BPS: interstitial cystitis/bladder pain syndrome
ICA: Interstitial Cystitis Association
ICPI: Interstitial Cystitis Problem Index
ICSI: Interstitial Cystitis Symptom Index
OR: odds ratio

XSL•FO
RenderX

Edited by A Mavragani; submitted 02.01.25; peer-reviewed by B Grigoryan, S Hajebrahimi; revised version received 13.03.25; accepted 08.04.25; published 05.08.25.

Please cite as:
Melkonian E, Garrett AL, Kline E, Smith P, Wiesenhahn M, Petit J, Swierczynski A, Zhou C, Bauer SB, Adam R, Barbour KE, Ziniel SI, Brownstein CA
Variability in the Diagnosis and Treatment of Interstitial Cystitis/Bladder Pain Syndrome: Internet Survey
JMIR Form Res 2025;9:e70813
URL: https://formative.jmir.org/2025/1/e70813
doi:10.2196/70813

© Eleanora Melkonian, A Leigh Garrett, Erika Kline, Pressley Smith, Madelyn Wiesenhahn, Jeanette Petit, Alison Swierczynski, Carley Zhou, Stuart B Bauer, Rosalyn Adam, Kamil E Barbour, Sonja I Ziniel, Catherine Brownstein. Originally published in JMIR Formative Research (https://formative.jmir.org), 5.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Continuous Rural High School Educational Outreach and Lasting Impact on Health Care Career Attitudes: Qualitative Pilot Study

Rebecca Bolen[1*], BA; Meagan Flesch[2*], BS; Jerrica Dennis[2], BA; Logan Shouse[2], BS; Mark Payton[2], PhD; David Ross[3], DO

[1]Department of Tracks and Special Programs, College of Osteopathic Medicine, Rocky Vista University, 8401 S Chambers Rd, Englewood, CO, United States

[2]Department of Biomedical Science, College of Osteopathic Medicine, Rocky Vista University, Parker, CO, United States

[3]Department of Tracks and Special Programs, College of Osteopathic Medicine, Rocky Vista University, Parker, CO, United States

[*]these authors contributed equally

Corresponding Author:
Rebecca Bolen, BA
Department of Tracks and Special Programs, College of Osteopathic Medicine, Rocky Vista University, 8401 S Chambers Rd, Englewood, CO, United States

Abstract

Background: Rural communities face persistent challenges in recruiting and retaining health care professionals. Research has shown that individuals from rural backgrounds are more likely to return to practice in these areas, yet most existing pipeline programs focus on undergraduate and medical students rather than high school students. Early exposure to health care careers is essential, as many students have already selected their career paths by the time they enter college. A previous study conducted in 2020 analyzed the effects of a single educational workshop at a rural high school in New Hampshire. The results suggested that students had a better understanding of the health care field.

Objective: This study evaluates the impact of an ongoing educational outreach program designed to introduce rural high school students to diverse health care professions.

Methods: This study was conducted at West Grand High School, a rural high school in Kremmling, Colorado, between September and December 2023. The intervention consisted of 4 monthly sessions, each focusing on a different medical specialty—primary care, sports medicine, dermatology, and neurology. These sessions, led by second-year medical students, provided an overview of common conditions, treatment approaches, and various health care roles involved in patient care. Participants completed pre- and postsession surveys assessing their interest in health care careers, perceived barriers, and likelihood of returning to their rural hometown to practice. A follow-up survey was conducted 4 months after the final session to assess long-term impact.

Results: While individual session surveys showed no significant changes, overall interest in and likelihood of pursuing a health care career increased significantly over the course of the presentation series (P=.03 and P=.04, respectively). However, there was no significant change in students' likelihood of returning to their rural hometowns to practice or their perceived access to resources for a health care career. Financial constraints (43/66, 65%) were identified as the most significant barrier, followed by lack of exposure (19/66, 29%) and support (17/66, 26%), while interest and education were least likely perceived as obstacles.

Conclusions: This study highlights the importance of early and sustained outreach efforts in rural communities to increase awareness of diverse health care career pathways. While short-term educational interventions can positively influence career interest, long-term mentorship and structured support systems are essential for fostering a sustained commitment to rural health care careers. Future initiatives should integrate financial counseling, ongoing mentorship, and collaborations with existing rural pipeline programs to enhance the effectiveness of such interventions.

(JMIR Form Res 2025;9:e70505) doi:10.2196/70505

KEYWORDS

rural health; health disparities; career pathways; high school students; education; healthcare workforce

Introduction

Rural and remote populations frequently contend with a lack of health care providers, reduced access to specialty services, and health outcomes that lag behind those in urban centers [1]. A Merritt Hawkins White Paper in 2018 reported that 20% of the US population lives in rural areas, yet only 9% of US physicians practice in these communities [2]. Limited health care infrastructure, long travel distances for care, and professional isolation all contribute to the difficulty in recruiting

XSL•FO
RenderX

and retaining qualified health care professionals in rural areas [2].

Research consistently shows that health care professionals who have lived in or originated from rural areas are significantly more likely to return to these communities to practice [3-5]. As a result, early exposure to health care careers for rural students is a critical strategy to bolster the rural health workforce. Without targeted interventions, students in rural communities may lack awareness of health care career pathways or perceive such careers as unattainable, perpetuating the cycle of inadequate health care services and an insufficient number of health care providers. Structural and socioeconomic barriers, such as fewer academic role models and geographic isolation, can further limit students' access to the information and mentorship necessary to pursue health-related careers [6]. However, by implementing continuous educational outreach initiatives, we can bridge the gap between rural high school students and health care–related professions such as physicians, nurses, various technicians, and related health care field careers.

Early engagement is especially important because many students make key career decisions during high school [7,8]. Research indicates that structured exposure to health care professions before college can significantly enhance students' interest, motivation, and self-efficacy in pursuing those careers [8,9]. These programs not only provide critical information about health professions and educational pathways but also offer students the opportunity to connect with real-world role models, demystify professional stereotypes, and envision themselves as future health care providers. For example, outreach programs that include hands-on experiences, simulations, facility tours, and personal stories from health professionals have been shown to increase students' interest in health careers, particularly in rural and underserved settings [10-14].

Many existing "pipeline" programs support undergraduate and medical students in pursuing careers in rural health care, but there is a deficit in programs that target rural high school students [15,16]. These students are at a formative stage where exposure can have a lasting impact on their academic and career trajectories. Without outreach during this period, we risk missing students who may never pursue postsecondary education due to a lack of early encouragement or awareness of viable career options [7,17].

The benefits of inspiring rural high school students to pursue careers in health care are twofold. First, it addresses the shortage of health care professionals in rural areas. By encouraging students from these communities to pursue health care careers, we increase the likelihood that they will return to their hometowns to practice, improving access to health care for the local population [18]. Furthermore, the early introduction of students to a diverse range of career paths within health care, beyond traditional roles such as doctors and nurses, can help students align their interests, strengths, and passions with a particular career path, allowing them to make informed decisions about their future careers [17].

Previous studies have demonstrated the effectiveness of educational outreach initiatives. For example, a 2020 project in New Hampshire provided rural high school students with a day of health care workshops and mentorship opportunities, fostering connections between students and health care professionals [19]. Another such initiative is the Canadian Healthcare Travelling Roadshow, a short-term yet highly interactive program that visits rural communities and delivers hands-on sessions designed to inspire youth toward health careers. Following participation, students reported increased interest in health professions and a stronger sense of connection with health care providers [13]. Similarly, a growing body of literature supports the value of high school–level outreach in increasing health career consideration, particularly in underserved populations [7-10]. However, many existing programs lack continuity and structured follow-up, limiting their long-term impact. Outreach programs that are sustained over time, rather than limited to a single event, may have a greater impact by building mentorship relationships, reinforcing career aspirations, and providing ongoing academic support.

This study aims to analyze how more continuous educational outreach initiatives affect rural high school students' attitudes toward a career in health care. Through multiple sessions over a longer period than previously outlined initiatives, it is anticipated that ongoing mentorship and education will inspire a lasting interest in health care professions. By empowering students and equipping them with knowledge, skills, and exposure, we can cultivate a new generation of health care professionals who are motivated to serve their communities.

Methods

Recruitment

This study was conducted from September to December 2023 at West Grand High School (WGHS) in Kremmling, Colorado. WGHS was selected due to its rural location. Kremmling, a town of less than 1500 residents in the Rocky Mountains, is at least a 2-hour drive from any larger populated area and 3 hours from a metropolitan district. WGHS ranks 245 of 334 Colorado high schools and 13,242 of 17,655 nationally [20]. Minority enrollment is 36%, and 44% of students are economically disadvantaged [20]. The school's total enrollment for grades 9 - 12 is 112 students [20]. Given its rural setting, many students may have limited exposure to health care professionals or career pathways in health care–related fields. Understanding this context is essential in evaluating the impact of the intervention.

Participants were recruited based on their interest in the program, with all grade levels invited. Each session included approximately 8 - 10 students (~9% of the student body), all of whom were female, ranging from freshmen to seniors. The same students attended each session throughout the study.

Program Design and Data Collection

The outreach program was structured as a health specialty educational series designed to introduce students to a variety of health care careers and training pathways. Data collection occurred over 4 sessions, each spaced approximately 1 month apart. Each session included a 20-minute presentation focused on a specific specialty: primary care, sports medicine, dermatology, or neurology. These topics were selected not only based on presenter interest but also for their ability to

XSL•FO
RenderX

incorporate a diverse range of health care careers, ensuring that students could explore multiple professional roles beyond physician pathways. In addition, faculty insights were gathered to align the topics with students' expressed interests. The sessions were specifically designed for a high school level of education, using accessible language, interactive activities, and relatable role models to foster engagement. The program built upon existing college counseling efforts at the school, integrating health care–specific content to complement and expand the resources already available to students.

Presentations were delivered by 4 second-year medical students from Rocky Vista University and were designed to highlight common conditions within each specialty while showcasing the interdisciplinary nature of health care. Each session emphasized the roles of various professionals involved in patient care, including physicians (MD or DO), physician assistants, nurse practitioners, nurses, athletic trainers, aestheticians, psychologists, technicians, and other allied health professionals. Discussions covered the educational and training requirements for each profession to provide students with a broader understanding of health care career pathways.

To enhance engagement, each session incorporated real-world clinical case presentations to demonstrate how different health care professionals collaborate in patient care. These case-based discussions encouraged critical thinking and allowed students to visualize the diverse roles within the health care system.

Before and after each presentation, participants completed anonymous voluntary surveys assessing their interest in health care. All surveys consisted of the following five questions: (Q1) How likely are you to stay in or return to your hometown to pursue your career?; (Q2) How interested are you in a career in health care? (Q3) What is the likelihood that you will go into a career in health care?; (Q4) What barriers affect the likelihood that you will go into a career in health care?; and (Q5) Do you feel that you have the resources (support, money, knowledge, etc) to pursue a career in health care?

Responses to questions 1, 2, 3, and 5 were based on a Likert scale from 1 to 5, corresponding to the answer choices of strongly disagree, disagree, neither agree nor disagree, agree, and strongly agree. Q4 allowed for participants to select any of the following that apply: money, educational background, lack of social support or family support, knowledge or exposure to health care, interest in health care.

In addition, the postpresentation surveys included five open-ended questions: (1) What did you like about this workshop?; (2) What did you dislike about this workshop?; (3) What questions or comments do you still have?; (4) Does anyone in your family or close friends work in the health care field? If so, who and what do they do?; and (5) Do you have any experience working or volunteering in the health care field? If so, what is your experience?

Statistical Analysis

A 2-factor ANOVA was run for each of the 4 Likert scale questions with month, presurvey, and postsurvey as the factors of interest. Pre-post analyses were performed for each month and longitudinally for the entire presentation series. *P* values, means, SEs, and sample sizes were calculated for each comparison. If a comparison between months given pre-post data was significant ($P<.05$), pairwise comparisons were made.

For Q4, contingency tables were constructed to examine potential relationships between students' likelihood of entering the health care field and each listed barrier. Contingency tables allow for the evaluation of interrelationships between categorical variables. A χ^2 test was then applied to determine statistical significance.

Ethical Considerations

This survey study was approved by the Rocky Vista University Institutional Review Board for participation by high school–aged students (approval #2023-150). Assent forms, which outlined the purpose and voluntary nature of the study, were completed by students, and parental consent was obtained prior to the presentation series. The original informed consent permitted secondary data analysis without the need for additional consent. All survey responses were collected anonymously, with no identifying information recorded. Students were informed of their right to withdraw from the study at any time without forfeiting the opportunity to attend the presentations. No monetary or material compensation was provided for participation.

Results

Analysis of pre- and postpresentation survey responses showed no significant changes within each individual session. However, responses to Q2 and Q3 (interest in a career in health care and the likelihood of going into a career in health care) indicated a significant increase over the course of the presentation series ($P=.03$ and $P=.04$, respectively), as seen in Figure 1. There was no significant change in the likelihood of the students staying or returning to their hometown to pursue a career (Q1) or in their feelings about having the resources to pursue a career in health care (Q5; Figure 1).

XSL•FO
RenderX

Figure 1. Representation of student responses over the 4 months of presentations: September (n=10), October (n=7), November (n=8), December (n=8). A self-reported Likert scale was used, with 5 being "strongly agree" and 1 being "strongly disagree."

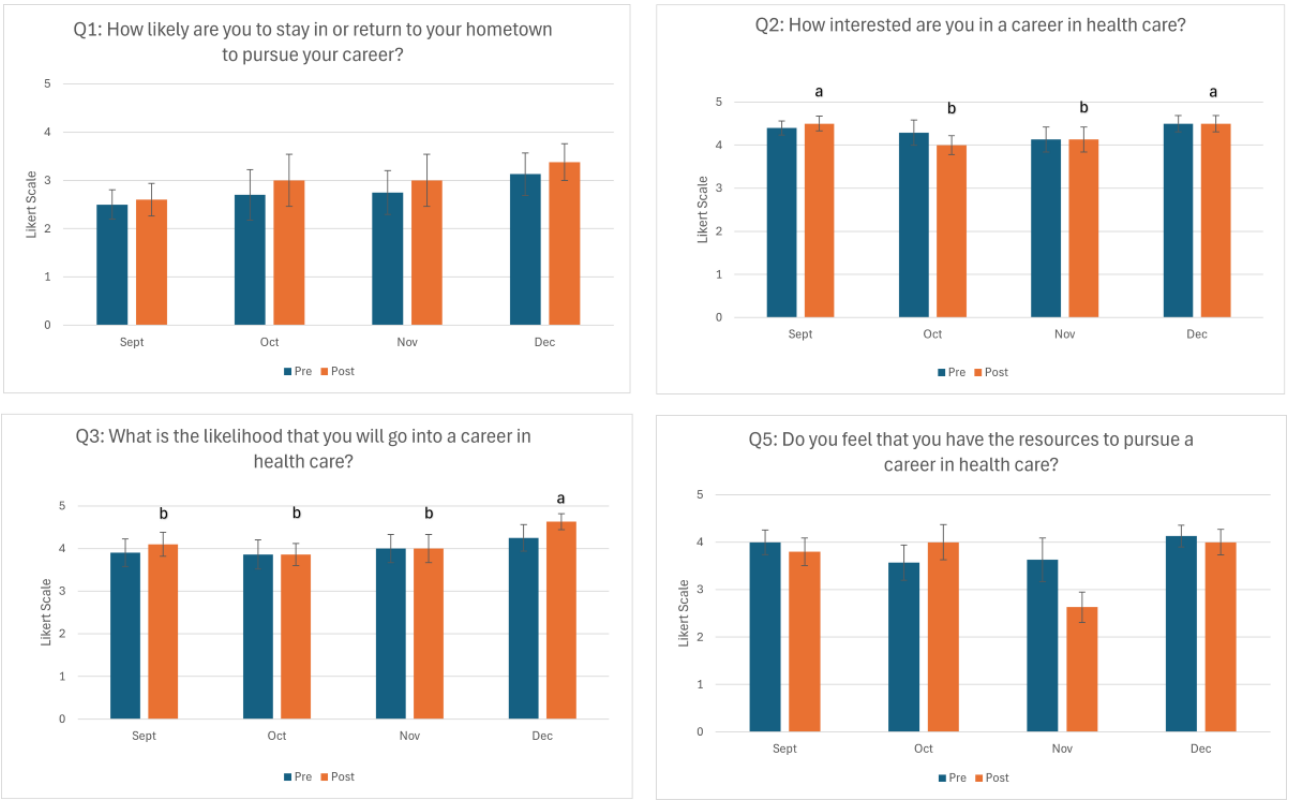

Analysis of the barriers to pursuing a health care career (Q4) in relation to Q3 revealed significant findings. In order of likelihood deemed as a barrier was money > exposure > support > interest, with education and interest being significantly less likely to be considered barriers compared to the first three (Table 1). However, statistical significance was only present as a negative correlation regarding interest and education, suggesting that students did not perceive these as barriers to pursuing a health care career.

Table . Representation of student-reported barriers to pursuing a career in health care over the whole 4-month period. In order of most perceived as a barrier to least: money > exposure > support > education > interest.

Q4: What barriers affect the likelihood that you will go into a career in health care?	Responses, (n/N, %)
Money	43/66, 65%
Educational background	10/66, 15%
Exposure	19/66, 29%
Lack of support	17/66, 26%
Interest	3/66, 5%

Several key themes emerged from student responses to the open-ended questions. Students expressed enthusiasm for learning about health care careers beyond becoming a physician. Many were unaware of the wide range of available roles, highlighting the need to further expand the series to explore additional career paths. Students also provided feedback on areas for improvement. Several mentioned a desire for more interaction with the lecture series team, as the separate presentation format limited engagement with all presenters. In addition, students requested a more detailed exploration of each health care career, which may necessitate extending the series if high school schedules allow. About half of the students reported having a family member or acquaintance in health care, raising questions about how to better engage those without prior exposure. Most students also reported little to no health care–related work or volunteer experience, with many citing personal patient experiences as their primary exposure to the field.

Discussion

Principal Findings

Previous studies have primarily focused on short-term outreach programs that rely on isolated lectures or single-day simulations [13,19]. While these models showed to spark initial interest, their limited duration may lack the continuity needed to sustain long-term impact. In contrast, this program emphasized consistent exposure and ongoing engagement through a series

XSL•FO
RenderX

of sessions and hands-on experiences. This repeated contact not only reinforces student interest but also fosters mentorship and deeper career exploration, potentially leading to more lasting shifts in students' aspirations toward health care professions.

The data demonstrate that the educational outreach program produced a statistically significant increase in interest in and likelihood of pursuing health care careers among rural high school students. Specifically, interest in pursuing health care careers (Q2) significantly increased from October to December, but not from September to December. This lack of significance over this study period may be attributed to fluctuations in sample size between September and December. We did observe a linear increase in interest from October to December, suggesting that consistent exposure to the intervention may contribute to sustained interest. As previously discussed, research shows that early and sustained exposure to health care significantly increases student interest, especially when introduced before college [8]. The consistent engagement from October to December suggests that continued exposure during this stage of education can serve as an effective early intervention, inspiring students to consider careers in health care. This approach may help address the rising demand for physicians in rural areas, as students from these communities are more likely to return and practice in their hometowns [3-5].

The likelihood of pursuing health care careers (Q3) increased over the entirety of the program. These findings support the potential benefit of implementing similar health care educational series in other rural high schools. Furthermore, the results showed no significant difference in responses from the beginning to the end of each individual presentation. This suggests that cumulative exposure to various health care specialties and career pathways, rather than interest in any single specialty, increased overall interest in the health care field. This impact may, in part, be attributed to the progressive development of mentorship and a growing sense of connection fostered throughout the educational series. This evolving rapport may have contributed to making a career in health care feel more accessible and attainable. Openly sharing experiences can help humanize the profession. Future research could explore whether specific specialties or health care roles have a greater impact on student engagement and how continued exposure shapes career aspirations over a more longitudinal time.

It is important to note that an increased interest in health care careers does not necessarily correspond to an increased interest in rural health care careers. While a rural background is a known factor influencing the choice to practice in rural settings, it is not a guarantee [3]. Many rural students pursue higher education with the intent of leaving rural areas. Literature has demonstrated that mentorship and ongoing exposure to rural practice settings are essential to maintaining interest in returning to rural health care [7]. Future iterations of this program could incorporate structured mentorship with local rural health care professionals and opportunities for hands-on exposure to rural health care settings, reinforcing the connection between career interest and rural practice.

The lack of statistically significant change in responses to Q1 (likelihood to stay or return to hometown) and Q5 (perception of having resources to pursue a health care career) is noteworthy. Since students' likelihood to stay or return to their hometown remained unchanged, this suggests that their career interest was independent of geographic location. Future interventions could address this by incorporating discussions on the unique opportunities and challenges of rural health care practice, including scope of practice, job demand, and lifestyle considerations. Increasing awareness of rural-specific health care career pathways, financial incentives for rural providers, and community impact could potentially enhance students' interest in remaining in their hometowns and addressing the shortage of rural health care professionals.

Similarly, the lack of change in students' perceptions of having the necessary resources to pursue a health care career (Q5) suggests that the presentations did not alleviate concerns about financial and educational barriers. Many students expressed concerns about the cost of health care education, particularly medical school. Expanding the curriculum to include sessions on scholarships and loan repayment programs could help demystify the financial aspects of health care training and provide students with concrete strategies to navigate educational barriers.

Regarding Q4 (barriers to pursuing a health care career), analysis revealed a negative correlation between perceived barriers related to interest or education and the likelihood of entering health care. This suggests that students did not view these factors as significant obstacles, which may reflect a growing confidence in their ability to pursue a health care career following the program.

To assess the long-term impact of this intervention, the same questionnaire was administered to students 4 months after the final presentation. The analysis compared responses from the initial survey in September to those from the follow-up questionnaire in April. The results indicated no significant change in students' responses over this period. This can be interpreted in several ways. One plausible explanation is a Type II error due to small sample size, which limits statistical power and may have resulted in a false negative. Expanding the program to a larger student population across multiple schools could provide a more robust evaluation of its effectiveness. Another possible explanation is the need for sustained exposure and reinforcement. Literature suggests that continued mentorship and health care exposure are necessary to solidify long-term interest in health care careers, particularly in rural settings [7]. Without additional interventions, students may not receive sufficient reinforcement of their interest in health care. To overcome this phenomenon, the program will need to be expanded over a longer duration or provide additional opportunities for students to learn outside of the presentation hours. The Canadian Healthcare Travelling Roadshow demonstrated that hands-on exposure to diverse health care fields can significantly boost student interest in medical careers [13]. Drawing on this model, this program added a full-day experience at the medical school of the presenters, offering interactive stations such as ultrasound, intubation, cardiopulmonary resuscitation (CPR), and phlebotomy. This immersive component allowed students to envision themselves in health care roles through direct engagement. While not

XSL•FO
RenderX

formally studied, student feedback indicated that the experience greatly enhanced their enthusiasm for pursuing careers in health care.

The findings regarding the open-ended questions highlight important areas for refinement and expansion in future lecture series. Many students found brief discussions on financing a medical education particularly valuable, as they were previously unaware of loan options, scholarships, and repayment strategies. The strong interest in financial education suggests that a dedicated lecture on funding options could play a crucial role in reducing perceived barriers to a health care career. Similarly, the enthusiasm for diverse health care career paths reinforces the importance of broadening the scope of future presentations to ensure students are aware of all available opportunities. Engaging students with no prior connection to health care remains a challenge, but targeted outreach efforts, such as early advertising or school assembly presentations, may help attract a more diverse audience. In addition, students expressed a desire for greater interaction with presenters, as the separate presentation format limited engagement. Personal storytelling has been shown to enhance student interest in health care careers, particularly within rural and underserved communities [13]. To address this, future series will incorporate an introductory question-and-answer session, allowing presenters to share their diverse journeys and foster a more interactive learning environment.

A key component of the program was providing students with mentorship opportunities. While mentorship was not a formalized aspect of the study, students were given the emails of all medical student presenters and were encouraged to reach out with questions. If a medical student could not answer a question, they connected students with a relevant health care professional who could provide guidance. Follow-up plans are also in place to sustain and expand this program. The initiative has been passed down to the next class of medical students to ensure continuity and further develop the program across additional rural high schools in Colorado. In addition, connecting students to existing mentorship programs through organizations such as the Colorado Area Health Education Center could provide more structured career development opportunities. This could be integrated into the series by incorporating a lecture, as discussed in Q4, that informs students about additional volunteer, shadowing, and health care exposure opportunities. Future research could evaluate the long-term impact of such mentorship connections on students' health care career trajectories.

Limitations include a small, fluctuating sample size and data collection from only one rural high school. Expanding this program to additional rural high schools would allow for a broader evaluation of its impact across different communities and a more diverse student population. Future research could also explore how different program components (eg, mentorship, financial education, and hands-on exposure) contribute to student engagement and career decision-making.

Although this study includes 4 sessions over 4 months, it represents the first phase of a continuous outreach model intended to be repeated annually and expanded to additional rural high schools. This sustained approach aims to create a lasting impact by reinforcing health care career interest over time through ongoing exposure, mentorship, and community engagement. Overall, the findings suggest that educational outreach programs like this health care specialty series can provide valuable career insight, mentorship, and resources to rural high school students. By refining and expanding such initiatives, we can continue to foster student interest in health care careers, address barriers to entry, and ultimately strengthen the rural health care workforce.

Conclusion

This study underscores the value of early and sustained outreach in encouraging rural high school students to explore health care careers. By providing repeated exposure, practical skill-building, and relatable role models, the program fostered increased interest in the field. While short-term interventions may spark initial curiosity, long-term engagement, through mentorship, structured support, and clear educational pathways, is key to sustaining student commitment, particularly to rural health care. Although some measures, such as intent to return to rural communities or perceived access to resources, showed limited change, these findings point to clear opportunities for program enhancement. With continued development, this initiative can serve as a scalable model to guide rural students toward health professions and address long-standing workforce gaps in underserved areas.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

References

1. Haakenstad A, Irvine CMS, Knight M, et al. Measuring the availability of human resources for health and its relationship to universal health coverage for 204 countries and territories from 1990 to 2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet 2022 Jun 4;399(10341):2129-2154. [doi: 10.1016/S0140-6736(22)00532-3] [Medline: 35617980]
2. Merritt Hawkins. 2018 review of physician and advanced practitioner recruiting incentives. AMN Healthcare. 2018. URL: https://www.merritthawkins.com/Rural-PhysicianRecruiting-Challenges-and-Solutions [accessed 2025-07-02]

XSL•FO
RenderX

3. Ogden J, Preston S, Partanen RL, Ostini R, Coxeter P. Recruiting and retaining general practitioners in rural practice: systematic review and meta-analysis of rural pipeline effects. Med J Aust 2020 Sep;213(5):228-236. [doi: 10.5694/mja2.50697] [Medline: 32696519]
4. Henry JA, Edwards BJ, Crotty B. Why do medical graduates choose rural careers? Rural Remote Health 2009;9(1):1083. [doi: 10.22605/rrh1083] [Medline: 19257797]
5. Walters L, Seal A, McGirr J, Stewart R, DeWitt D, Playford D. Effect of medical student preference on rural clinical school experience and rural career intentions. Rural Remote Health 2016;16(4):3698. [doi: 10.22605/rrh3698] [Medline: 27854470]
6. Fisher KA, Fraser JD. Rural health career pathways: research themes in recruitment and retention. Aust Health Rev 2010 Aug;34(3):292-296. [doi: 10.1071/AH09751] [Medline: 20797360]
7. Warshaw R. Priming the medical school pipeline: schools reach out to teens in minority and underserved communities. Association of American Medical Colleges. 2016. URL: https://www.aamc.org/news/priming-medical-school-pipeline-schools-reach-out-teens-minority-and-underserved-communities [accessed 2025-06-11]
8. Lauver LS, Swan BA, West MM, et al. Kids into health careers: a rural initiative. J Rural Health 2011;27(1):114-121. [doi: 10.1111/j.1748-0361.2010.00316.x] [Medline: 21204978]
9. Anibas J. Expanding the pipeline of rural students interested in medicine. WMJ 2022 Jul;121(2):71. [Medline: 35857678]
10. Moro C, Phelps C. Encouraging study in health sciences: informing school students through interprofessional healthcare simulations. Simul Healthc 2024 Jun 1;19(3):144-150. [doi: 10.1097/SIH.0000000000000732] [Medline: 37255339]
11. Horslen LC, Kotova S, Hankins V, et al. CHEST Watch: a high school outreach program. Semin Thorac Cardiovasc Surg 2022;34(3):1134-1139. [doi: 10.1053/j.semtcvs.2021.07.013] [Medline: 34284071]
12. Yeh VJH, Mukhtar F, Rajesh A, et al. Inspiring young learners toward a career in healthcare using simulation. Sim Healthcare 2020;15(6):404-408. [doi: 10.1097/SIH.0000000000000432]
13. Maurice S, Mytting K, Gentles JQ, et al. The Healthcare Travelling Roadshow: a qualitative study of a rural community engagement initiative in Canada. Rural Remote Health 2019 Sep;19(3):2019. [doi: 10.22605/RRH5238] [Medline: 31500435]
14. MacAskill W, Fallon AB, Cotter N, Purea P. Aspire2Health and COVID-19: the impact of the pandemic on outcomes from an outreach program to increase high school student interest in rural health careers. Rural Remote Health 2023 Aug;23(3):8007. [doi: 10.22605/RRH8007] [Medline: 37633313]
15. Longenecker RL, Andrilla CHA, Jopson AD, et al. Pipelines to pathways: medical school commitment to producing a rural workforce. J Rural Health 2021 Sep;37(4):723-733. [doi: 10.1111/jrh.12542] [Medline: 33244824]
16. Reeves NS, Cheek C, Hays R, et al. Increasing interest of students from under-represented groups in medicine-a systematised review. Aust J Rural Health 2020 Jun;28(3):236-244. [doi: 10.1111/ajr.12609] [Medline: 32462758]
17. Kumar K, Jones D, Naden K, Roberts C. Rural and remote young people's health career decision making within a health workforce development program: a qualitative exploration. Rural Remote Health 2015;15(4):3303. [doi: 10.22605/rrh3303] [Medline: 26556399]
18. Feldman K, Woloschuk W, Gowans M, et al. The difference between medical students interested in rural family medicine versus urban family or specialty medicine. Can J Rural Med 2008;13(2):73-79. [Medline: 18405464]
19. Dunn MC, Saks NS. Inspiring rural high school students to consider careers in medicine and the health professions. TRE 2020;41(2):64-69. [doi: 10.35608/ruraled.v41i2.867]
20. West Grand High School. US News. 2025. URL: https://www.usnews.com/education/best-high-schools/colorado/districts/west-grand-school-district-no-1/west-grand-high-school-4243 [accessed 2025-02-09]

Abbreviations

WGHS : West Grand High School

Edited by A Mavragani; submitted 23.12.24; peer-reviewed by J Birney, M Deutchman; revised version received 05.05.25; accepted 11.05.25; published 14.07.25.

Please cite as:
Bolen R, Flesch M, Dennis J, Shouse L, Payton M, Ross D
Continuous Rural High School Educational Outreach and Lasting Impact on Health Care Career Attitudes: Qualitative Pilot Study
JMIR Form Res 2025;9:e70505
URL: https://formative.jmir.org/2025/1/e70505
doi:10.2196/70505

© Rebecca Bolen, Meagan Flesch, Jerrica Dennis, Logan Shouse, Mark Payton, David Ross. Originally published in JMIR Formative Research (https://formative.jmir.org), 14.7.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution,

XSL•FO
RenderX

and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Mediating Effects of Psychological Independence and Social Support on the Association Between Family Strength and Depression in Young Korean Adults: Cross-Sectional Study

Sunyoung Kim[1], MSN; Suin Park[2], PhD; Hyunlye Kim[3], PhD; Dabok Noh[4], PhD

[1]College of Nursing, Korea University, Seoul, Republic of Korea

[2]Gyeongju Mental Health Welfare Center, Gyeongju, Republic of Korea

[3]Department of Nursing, College of Medicine, Chosun University, Gwangju, Republic of Korea

[4]College of Nursing, Eulji University, 553 Sanseong-daero, Sujeong-gu, Seongnam-si, Gyeonggi-do, Republic of Korea

Corresponding Author:
Dabok Noh, PhD
College of Nursing, Eulji University, 553 Sanseong-daero, Sujeong-gu, Seongnam-si, Gyeonggi-do, Republic of Korea

Abstract

Background: Although family strength is potentially associated with a reduced risk of depression, little is known about the underlying pathways and mediating factors.

Objective: This study aimed to investigate the mediating effects of psychological independence and social support on the relationship between family strength and depression in young adults.

Methods: A cross-sectional web-based survey was conducted among 1,000 young Korean adults aged 19 to 24 years. We used a web-based survey agency to recruit participants using an independent panel and quota sampling, with stratification based on gender and age. The participants completed self-reported questionnaires that assessed family strength, psychological independence, social support, and depression. To examine the mediating effects of psychological independence and social support on the relationship between family strength and depression, we performed path analysis with AMOS 26 software (IBM Corp) using maximum standard likelihood estimation.

Results: The path analysis revealed that gender (female) had a direct positive effect on depression (β=.09, *P*=.004) and an indirect negative effect on depression through social support (β=−.03, *P*=.001). Although there were no significant direct effects of living status (with parents) on depression, it had a significant and positive indirect effect through psychological independence (β=.03, *P*=.001). Family strength had a significant and negative direct effect on depression (β=−0.19, *P*=.001) and significant indirect and negative effects through psychological independence and social support (β=−0.17, *P*=.001). Therefore, the overall effect of family strength on depression was significantly negative (β=−0.37, *P*=.001). Psychological independence influenced depression both directly (β=−0.16, *P*=.001) and indirectly through social support (β=−0.12, *P*=.001), and social support influenced depression directly (β=−0.21, *P*=.001). The overall model explained 23% of the total variance in depression.

Conclusions: The findings highlight that gender, living with parents, family strength, psychological independence, and social support in reduce depression among young adults. Additionally, the mediating effects of psychological independence and social support on the relationship between family strength and depression were significant in this population. Therefore, strategies to increase psychological independence and social support could reduce the risk of depression in young adults who have low family strength.

(JMIR Form Res 2025;9:e71485) doi:10.2196/71485

KEYWORDS

depression; family strength; psychological independence; social support; young adults

Introduction

The World Mental Health Report claims that of all mental disorders, depression has the strongest burden [1]. A national US survey found that three-quarters of all lifetime cases of depression started before people were 24 years old [2]. In South Korea, depression in young adults has become increasingly prevalent in recent years. In March 2022, Korea's National Mental Health Survey found a depression rate of 18.5% in people in their 20s [3]. Early adulthood or "emerging adulthood," which was recently recognized as a developmental psychology life stage, is a crucial early stage for preventive interventions to treat depression [4].

XSL•FO
RenderX

Family strength promotes positive growth; therefore, children raised in strong families are more likely to be healthy and develop strong family units as adults [5]. Sittner et al [6] described family strength as "commitment, appreciation and affection, positive communication, enjoyable time together, a sense of spiritual well-being, and the ability to cope with stress and crisis." Other studies have also found that family strength can reduce the likelihood of depression in family members [7-9].

Young adulthood is defined as being between the ages of 19 and 24 years [10]. This is generally a period of transition from childhood dependence and adolescence to assuming adult roles and responsibilities [11]. While receiving continued support from their parents, young adults are also seeking to establish their identities and prepare for independence [12]. Psychological independence in young adults is defined as autonomous thinking and behavior and concurrently maintaining a positive, secure relationship with their parents based on attachment [13]. Higher psychological independence is associated with lower rates of depression in young adults [14].

Social support can be an important protective factor against mental health problems in emerging adulthood, which is a transitional period in which people go through substantial changes in their social roles and responsibilities [4]. Previous studies have revealed that higher perceived social support is associated with lower levels of depression in young adults [15,16].

Psychological independence could also play a mediating role between family strength and depression. Previous studies have found that higher perceived family strength is associated with greater psychological independence [17], and dysfunctional family lives and conflicted dependence on parents are associated with higher levels of suicidal ideation [18]. Psychological independence was found to mediate the relationship between family strength and depression [19].

Social support may also mediate the relationship between family strength and depression. For example, a previous study found that family strength mediated through social support had a significant effect on happiness in young adults [20], and another study found that social support mediated the relationship between family strength and psychological well-being in children [21].

The theoretical framework of this study is based on the Social Support Theory Model, which emphasizes the role of social support in enhancing an individual's ability to cope with stress and positively influencing their health [22]. A previous study found that individuals with high levels of psychological independence are more actively engaged in social relationships and are more likely to receive support from others, suggesting that psychological independence is a significant factor in promoting social support [23]. Another study emphasized that social support increases psychological stability and reduces the perception of stress, thereby lowering the risk of depression [24]. Psychological independence facilitates social support, which in turn reduces the risk of depression, ultimately supporting the mental health of young adults through this mechanism.

These study results suggest that psychological independence and social support could mediate the relationship between family strength and depression; however, little is known regarding these structural relationships. In this study we conducted a path analysis of the relationships among family strength, psychological independence, social support, and depression in young adults. Figure 1 shows our hypothesized model, in which family strength influences depression, with this relationship being mediated by psychological independence and social support.

XSL•FO
RenderX

Figure 1. Hypothetical study model: mediating effects of psychological independence and social support on the relationship between family strength and depression among young Koreans adults aged 19 - 24 years. Arrows indicate the hypothesized causal relationships between variables.

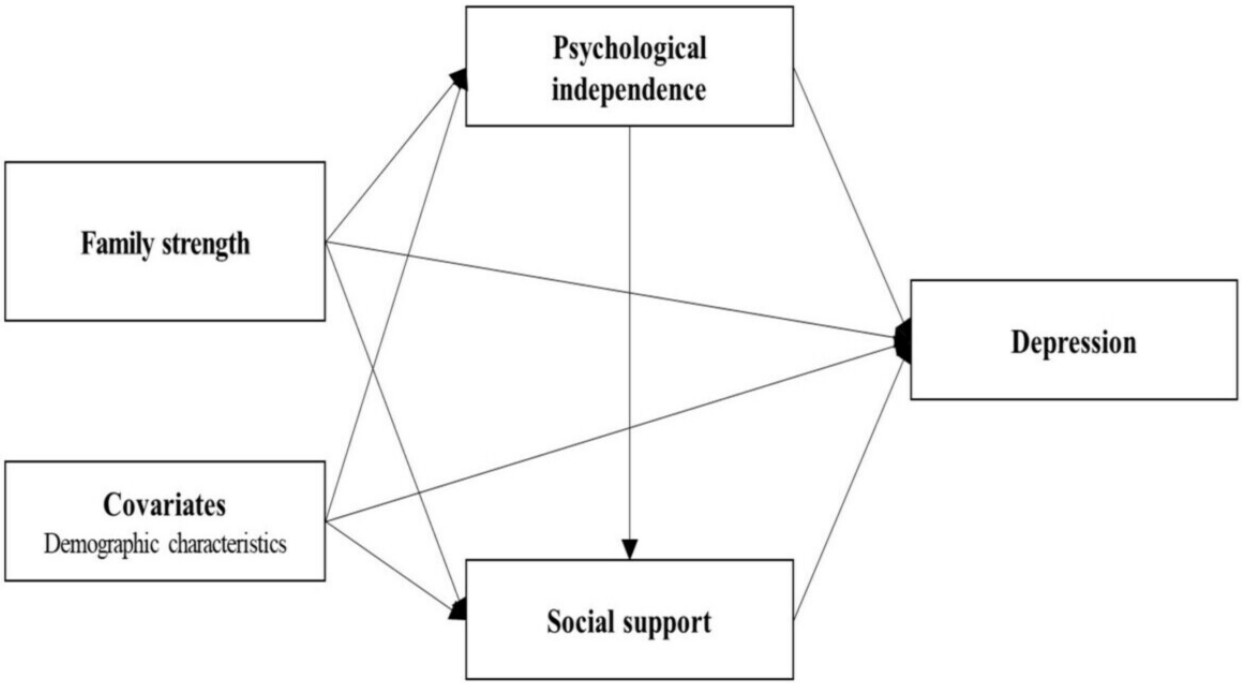

Methods

Study Design and Participants

A cross-sectional web-based survey was conducted among 1,000 young Korean adults aged 19 - 24 years. This study adhered to the STROBE (Strengthening the Reporting of Observational Studies) guidelines, ensuring rigorous compliance with the reporting standards for cross-sectional studies. Using the G-power version 3.1.9.4 program [25], the required sample size was calculated based on a significance level of 5%, statistical power of 85%, an effect size of 0.02 (small), and 11 predictors. The minimum required sample size was determined to be 945. Considering a 6% dropout rate, the final sample size was set at 1,000.

Measures

Family Strengths

We used the Korea Family Strengths Scale II [26] to assess family strength within the Korean cultural context. This scale comprises 5 dimensions: family resilience; valuing each other and acceptance; qualitative bonding; economic stability and cooperation; and caring about the community, within which 22 self-reported items are scored on a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). An example questionnaire item is “My family trusts each other.” Scores range from 22 to 110, with higher scores indicating greater family strength. Cronbach α coefficient was 0.94 in a prior study [26] and 0.97 in this study.

Psychological Independence

Psychological independence was evaluated using the Psychological Independence Scale [12], which comprises 3 factors: supportive relationship with parents; voluntary decision-making; and self-reliability. The “supportive relationship with parents” factor has 6 items related to feeling a sense of being accepted as they are, having their decisions respected, and having their opinions valued when solving family issues [12]. The “voluntary decision-making” factor has 6 items related to autonomously acting based on one’s own thoughts or approaches rather than parental opinions in decision-making [12]. The “self-reliance” factor has 6 items related to having confidence in one’s career path, life goals, and potential, and using this confidence to establish personal goals and plans [12]. The scale has 18 self-report items that are scored on a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree), with 6 negatively worded items being reverse-coded. An example questionnaire item is “I have my own life goals.” Scores range from 18 to 90, with higher scores indicating greater psychological independence. Cronbach α coefficient was 0.83 in a prior study [12] and 0.83 in this study.

Social Support

The Korean version of the Social Provision Scale [27], which was translated into Korean by Yoo and Lee [28], was used to assess perceived social support. The scale has 6 factors—guidance, reassurance of worth, social integration, attachment, opportunity for nurturance, and reliable alliance—to assess relationships with friends, family, coworkers, community

XSL•FO
RenderX

members, and others. The 24 items are rated on a 4-point Likert scale ranging from 1 (strongly disagree) to 4 (strongly agree). An example questionnaire item is "There are people I can depend on to help me if I really need it." Of these 24 items, the 12 negatively worded items are reverse-coded. Scores range from 24 to 96, with higher scores indicating greater perceived social support. Cronbach α coefficient was 0.92 in a prior study [27] and 0.93 in this study.

Depression

Depression was assessed using the Korean version of the Center for Epidemiologic Studies Depression Scale-Revised (K-CESD-R) [29], which is the Korean translation of CESD-R [30]. The K-CESD-R has 20 self-report items scored on a 5-point Likert scale ranging from 0 (less than 1 day in the last week) to 4 (nearly every day in the last 2 weeks). An example questionnaire item is "I could not shake off the blues." Scores range from 0 to 80, with higher scores indicating a higher level of depression. The Cronbach α coefficient for the K-CESD-R was 0.98 in a prior study [29] and 0.95 in this study.

General Characteristics

We also collected participant characteristics, including gender, age, living status, educational level, religion, job, household economic status, and residential area.

Data Collection

A web-based survey agency distributed research descriptions and survey links via email to an independent panel of young Korean adults aged 19‑24 years. They collected data using quota sampling with gender and age stratification, ensuring an even distribution of participants to minimize potential sampling bias. Participants reviewed the survey information, provided informed consent, and accessed the web-based survey via the provided URL. The data were collected from the 1000 participants between October 24, 2022, and October 31, 2022.

Data Analyses

The data were analyzed using SPSS (version 26.0) and AMOS (version 26.0) software (IBM Corp). Descriptive analyses were generated for demographic characteristics, family strength, psychological independence, social support, and depression. Independent 2-tailed *t* tests, ANOVA with the Scheffe post hoc test, and Pearson correlations were used to analyze family strength, psychological independence, social support, and depression based on demographic characteristics. To examine the mediating effects of psychological independence and social support on the relationship between family strength and depression, we performed path analysis with AMOS 26 software using maximum standard likelihood estimation.

Ethical Considerations

The study protocol was reviewed by the institutional review board of Eulji University (approval number: EU22-71, date of approval: October 17, 2022). Participants were provided with study information on the first screen of the web-based questionnaire. After reading the information, they provided electronic consent by clicking the consent button. To prevent duplicate submissions, a system was implemented to detect and block multiple responses; if a repeat attempt was made, a notification was displayed and access was restricted. If a participant withdrew from the survey—for example, by closing the website—their data were automatically discarded. Upon survey completion, all data were processed automatically to ensure anonymity and were coded using nonidentifiable markers. Nonmonetary incentives emphasized the importance of participants' contributions in advancing mental health research and intervention development.

Results

Main Variables Based on Demographic Characteristics

Table 1 shows the data for demographic characteristics, family strength, psychological independence, social support, and depression. Out of 1000 participants in the study, 520 (52%) were male, the mean age was 21.68 (SD 1.68) years, and more than two-thirds lived with their parents (n=676, 67.6%). Most of the 1,000 participants had a college education or higher (n=844, 84.4%), had no religion (n=720, 72.0%), and had no job (n=826, 82.6%). The self-reported household economic status of 1,000 participants was middle 434 (43.4%), low 284 (28.4%), and high 282 (28.2%), with nearly half living in big cities 489 (48.9%), followed by small- or medium-sized cities 404 (40.4%), and rural areas 107 (10.7%).

Compared to females, the males reported higher family strength (t_{998}=5.03, P<.001) and lower social support ($t_{997.45}$=−2.16, P=.031) and depression (t_{998}=−3.69, P<.001). Participants not living with their parents showed higher psychological independence than those living with their parents (t_{998}=2.69, P=.007). Compared to high school graduates, participants with a college education or higher had higher levels of family strength (t_{998}=−4.80, P<.001), psychological independence (t_{998}=−2.72, P=.007), and social support (t_{998}=−2.86, P=.004), and lower levels of depression ($t_{190.21}$=3.46, P=.001). Religious participants had higher family strength than those who were nonreligious (t_{998}=−3.77, P<.001). Participants with low self-reported household economic status had significantly lower family strength than those with middle or high economic status ($F_{2, 997}$ =77.55, P<.001). Participants with low self-reported household economic status had the lowest levels of psychological independence ($F_{2, 997}$=23.84, P<.001) and social support ($F_{2, 997}$ =22.23, P<.001), and the highest level of depression ($F_{2, 997}$ =16.51, P<.001), followed by those with middle or high household economic status. Participants living in big cities had higher family strength than those living in small- or medium-sized cities and rural areas ($F_{2, 997}$ =3.42, P=.033). Participants living in rural areas had lower levels of psychological independence than those living in big cities and small- or medium-sized cities ($F_{2, 997}$=4.69, P=.009).

XSL•FO
RenderX

Table . Family strength, psychological independence, social support, and depression according to demographic characteristics (N=1000).

Variables		Values	Family strengths			Psychological independence			Social support			Depression		
			Mean (SD)	Statistics	*P* value	Mean (SD)	Statistics	*P* value	Mean (SD)	Statistics	*P* value	Mean (SD)	Statistics	*P* value
Gender				$t_{998}=5.03$	<.001		$t_{998}=1.66$	.10		$t_{974.5}=2.16$	.03		$t_{998}=-3.69$	<.001
	Male	n=520 (52.0%)	81.46 (17.36)			65.08 (9.49)			71.46 (12.38)			12.55 (15.43)		
	Female	n=480 (48.0%)	76.00 (16.98)			64.05 (10.34)			73.07 (11.16)			16.10 (14.98)		
Age (years)		Mean 21.68 (SD 1.68)	—[a]	$r_{998}=-0.05$	.11	—	$r_{998}=-0.03$	.40	—	$r_{998}=-0.04$	.23	—	$r_{998}=-0.06$	.07
Living status				$t_{998}=-0.30$	.76		$t_{998}=2.69$	.01		$t_{998}=1.46$	.15		$t_{998}=1.32$	.19
	Without parents	n=324 (32.4%)	78.60 (17.69)			65.80 (9.48)			73.02 (11.40)			15.18 (15.35)		
	With parents	n=676 (67.6%)	78.95 (17.25)			64.00 (10.07)			71.86 (12.02)			13.81 (15.29)		
Education level				$t_{998}=-4.80$	<.001		$t_{998}=-2.72$	.01		$t_{998}=-2.86$	.004		$t_{992.1}=-3.46$	.001
	High school	n=156 (15.6%)	72.78 (18.71)			62.61 (9.61)			69.76 (12.92)			18.88 (18.79)		
	College or higher	n=844 (84.4%)	79.96 (16.90)			64.95 (9.93)			72.69 (11.57)			13.40 (14.43)		
Religious views				$t_{998}=-3.77$	<.001		$t_{998}=-1.74$	.08		$t_{998}=-1.67$	.10		$t_{998}=-0.03$	.97
	No	n=720 (72.0%)	77.56 (17.56)			64.25 (10.04)			71.84 (11.96)			14.25 (15.02)		
	Yes	n=280 (28.0%)	82.14 (16.49)			65.46 (9.55)			73.24 (11.44)			14.28 (16.06)		
Having a job				$t_{998}=1.00$	.32		$t_{998}=0.87$	.39		$t_{998}=0.03$	.97		$t_{998}=0.96$	.34
	No	n=826 (82.6%)	79.09 (17.42)			64.71 (10.00)			72.24 (12.03)			14.47 (15.30)		
	Yes	n=174 (17.4%)	77.64 (17.20)			63.99 (9.49)			72.21 (10.86)			13.25 (15.36)		
Household economic status				$F_{2,997}=7.55$	<.001, a<b,c[b]		$F_{2,997}=23.84$	<.001, a<b<c		$F_{2,997}=22.23$	<.001, a<b<c		$F_{2,997}=16.51$	<.001, c<b<a
	Low	n=284 (28.4%)	68.89 (17.11)			61.46 (10.55)			68.65 (12.81)			19.08 (17.77)		
	Middle	n=434 (43.4%)	81.68 (15.60)			65.08 (9.40)			72.76 (11.03)			12.46 (13.45)		
	High	n=282 (28.2%)	84.49 (16.12)			66.97 (9.22)			75.03 (11.12)			12.16 (14.29)		
Residential area				$F_{2,997}=3.42$	.03, e,f<d[c]		$F_{2,997}=4.69$	.01, f<d,e		$F_{2,997}=3.00$	.05		$F_{2,997}=1.15$	.32
	Big city	n=489 (48.9%)	80.19 (17.11)			65.09 (9.80)			72.75 (11.86)			13.52 (14.14)		
	Small- or medium-sized city	n=404 (40.4%)	77.95 (17.18)			64.69 (9.99)			72.29 (11.78)			14.85 (16.34)		

XSL•FO
RenderX

Variables	Values	Family strengths	Psychological independence	Social support	Depression
Rural area	n=107 (10.7%)	76.03 (18.96)	61.88 (9.78)	69.66 (11.67)	15.38 (16.41)

[a]Not applicable.

[b] a: Low; b: Middle; c: High

[c] d:Big city; e: Small- or medium- sized city; f: Rural area

Levels of Main Variables

Table 2 shows the results for family strength, psychological independence, social support, and depression. The respective mean family strength, psychological independence, social support, and depression scores were 78.84 (SD 17.38), 64.59 (SD 9.91), 72.23 (SD 11.83), and 14.26 (SD 15.31).

Table . Family strength, psychological independence, social support, and depression (N=1000).

Characteristics	Mean (SD)	Observed range	Possible range
Family strengths	78.84 (17.38)	22 - 110	22 - 110
Psychological independence	64.59 (9.91)	30 - 90	18 - 90
Social support	72.23 (11.83)	26 - 96	24 - 96
Depression	14.26 (15.31)	0 - 80	0 - 80

Path Analysis on the Determinants of Depression

Our previous bivariate analyses found that gender, living status, educational level, household economic status, and residential area were significantly associated with psychological independence, social support (mediating variables), and depression (dependent variable). Therefore, these demographic variables—gender, living status, educational level, household economic status, and residential area—were included as covariates in the initial model of this study. However, of the demographic variables, household economic status was excluded in the final model because it was closely related to educational level, and to better enhance the model fit, the residential area was excluded due to the lack of meaningful results. Figure 2 shows the final path diagram for the modified model.

The final model had a good fit to the data (minimum discrepancy function/degrees of freedom=3.000, goodness of fit index=0.998, comparative fit index=0.998, Tucker Lewis index=0.990, standardized root-mean-squared residual=0.015, and root-mean-square error of approximation=0.028).

The direct, indirect, and total effects of the final model are shown in Table 3. Gender (female) had a direct positive effect on depression (β=.09, *P*=.004), and an indirect negative effect on depression through social support (β=−0.03, *P*=.001). Although there were no significant direct effects of living status (with parents) on depression, it had a significant and positive indirect effect through psychological independence (β=.03, *P*=.001).

Family strength had a significant and negative direct effect on depression (β=−0.19, *P*=.001), and significant indirect negative effects through psychological independence and social support (β=−0.17, *P*=.001). Therefore, the overall effect of family strength on depression was significantly negative (β=−0.37, *P*=.001).

Our analysis revealed that there were significant psychological independence and social support mediation effects. Psychological independence influenced depression both directly (β=−0.16, *P*=.001) and indirectly through social support (β=−0.12, *P*=.001), and social support influenced depression directly (β=−0.21, *P*=.001). The overall model explained 23% of the total variance in depression.

XSL•FO
RenderX

Figure 2. Final path diagram for depression in young adults: The direct and indirect effects of family strengths, social support, psychological independence, and demographic factors (gender and living status) on depression. Path coefficients are presented as standardized beta values with significant levels (*P<.05, **P<.01, *** P<.001).

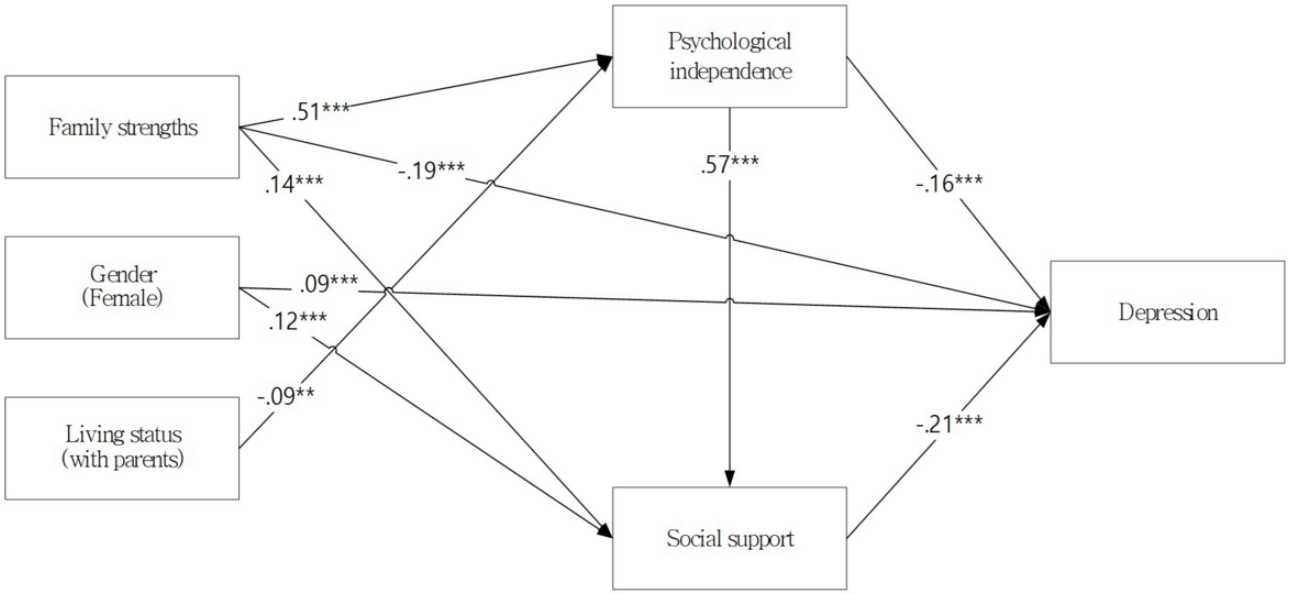

Table . Direct, indirect, and total effects of the final modified model.

Path	Direct	Indirect	Total
	β (P value)	β (P value)	β (P value)
Living status (with parents) → psychological independence (SMC[a]=.263)	−0.09 (.001)	—[b]	−0.09 (.001)
Family strength → psychological independence (SMC=.263)	0.51 (.001)	—	0.51 (.001)
Gender (female) → social support (SMC=.430)	0.12 (.001)	—	0.12 (.001)
Living status (with parents) → social support (SMC=.430)	—	−0.05 (.001)	−0.05 (.001)
Family strength → social support (SMC=.430)	0.14 (.001)	0.29 (.001)	0.43 (.001)
Psychological independence → social support (SMC=.430)	0.57 (.001)	—	0.57 (.001)
Gender (female) → depression (SMC=.230)	0.09 (.004)	–0.03 (.001)	0.07 (.04)
Living status (with parents) → depression (SMC=.230)	—	0.03 (.001)	0.03 (.001)
Family strength → depression (SMC=.230)	−0.19 (.001)	−0.17 (.001)	−0.37 (.001)
Psychological independence → depression (SMC=.230)	−0.16 (.001)	−0.12 (.001)	−0.28 (.001)
Social support → depression (SMC=.230)	−0.21 (.001)	—	−0.21 (.001)

[a]SMC: squared multiple correlations for structure equations.

[b]Not applicable.

XSL•FO
RenderX

Discussion

Principal Findings

This study examined the mediating effects of psychological independence and social support on the relationship between family strengths and depression among young adults. The findings indicate that both factors serve as significant mediators.

The mean depression score (14.26, SD 15.31) in our young adult study sample was slightly lower than that of other studies using the K-CESD-R. Yi et al [31] reported a mean depression score of 15.55 (SD 11.87) in Korean female college students, and Kim and Park [32] reported a mean score of 17.31 (SD 8.29) in Korean college students. As a cutoff score of 13 indicates a risk of clinical depression [29], most studies have found the average depression level in Korean young adults to be relatively high. This could be a result of job insecurity [33], economic instability, or the lack of quality, stable jobs, and employment opportunities [34] for young Korean adults. In addition, the relatively large SD of 15.31 observed in this study highlights significant variability in depression levels among participants, implying the existence of a subgroup with particularly high levels of depression. These findings emphasize the need for more detailed analyses of depression levels in future research to better understand and address this issue.

Our results confirm previous findings regarding significant gender differences in depression levels. Kim et al [35] and Salk et al [36] reported higher levels of depression in females in globally representative samples. Our study found that these gender differences are more pronounced in Korean society, likely due to the unique cultural norms present in Confucian-oriented societies. In Korea, Confucian traditions emphasize harmony, social order, and hierarchy, which restrict individual emotional expression and particularly lead men to suppress their emotions to a greater degree than women [37], thereby contributing to a lower reported depression level among men. In Confucian cultural contexts, young adult women experience elevated levels of depression, primarily attributable to the dual burden of fulfilling traditional familial roles and conforming to heightened societal expectations [38]. This dual burden likely explains the higher depression levels observed in the female participants of our study.

Our results also found that the perceived social support score for females was significantly higher than that of males. A qualitative analysis reported that the types of preferred social support also differed by gender in young adults [39]. They found that males tended to prioritize social network support, which helped them forget and control their emotional distress by engaging in activities, such as having fun, whereas females prioritized social support, which helped them talk about their problems and analyze their emotional distress. The social support measurement scale we used assessed guidance, reassurance of worth, social integration, attachment, opportunity for nurturance, and reliable alliance in relationships with friends, family, coworkers, community members, and others [28]. Therefore, as the measurement tool's characteristics may have aligned more closely with women's preferred social support types, there may have been higher perceived social support levels for women than for men. However, further research may be needed to determine the gender differences in social support.

Our research results also revealed that young adults living with their parents had significantly lower levels of psychological independence than those not living with their parents. Our findings align with those of prior research conducted with Japanese university students, suggesting that the negative impact of cohabiting with parents on psychological independence may be a shared phenomenon within the Confucian-influenced East Asian cultural sphere [40]. This may be because of the current socioeconomic conditions in Korea, as factors such as increased job insecurity and rising housing costs are preventing many young adults from becoming independent from their parents [41]. Older generations in Korea also tend to have family values that consider it natural to continue looking after their children [41]. Consequently, many young adults who continue to live with their parents are also dependent on their parents, which could impact their psychological independence.

Our results revealed a direct negative effect of family strength on depression in young adults. Family strength comprises 5 factors: family resilience, valuing each other and acceptance, qualitative bonding, economic stability and cooperation, and caring about the community [26]. Consistent with our findings, a study targeting college students reported a negative direct effect of family strength on depression [19]. Other studies targeting adolescents have also reported that higher perceived family resilience was associated with lower depression [42,43]. Rahman et al [44] also reported that higher family cohesion, that is, the bond and interconnectedness between family members, resulted in lower depressive symptoms in young adults. Because the results suggested that young adults with weak family strength are possibly more vulnerable to depression, psychiatric and mental health nurses should assess family strength when dealing with depression and develop preventive interventions for young people who have poor family strength.

The path analysis revealed that family strength influenced psychological independence, which, in turn, influenced depression. This result was also similar to previous studies on college students [19], and a previous study on Korean high school students that found that the lower the perception of family strength, the lower the psychological independence [17]. An earlier study on Korean college students [14] also reported that the lower the psychological independence level, the higher the level of depression. Jeon [45] examined the relationship mechanism between psychological independence and depression, finding that the lower the psychological independence from the parents, the higher the attachment anxiety (fear of rejection or abandonment) and the attachment avoidance (persistent avoidance or discomfort with intimacy), and the higher the depression levels. Therefore, to prevent depression, the psychological independence of young people from families that have low family strength should be assessed and relevant interventions provided.

"Psychological independence" is defined as a state in which young adults exhibit autonomy in their thoughts and behaviors while maintaining supportive relationships with their parents based on stable attachment [12]. However, parenting methods

XSL•FO
RenderX

such as helicopter parenting, which is characterized by overprotective and controlling behavior toward children [46], may impair the development of independence and autonomy in young adults [47]. In their study on college students, Kim and Park [48] found that the higher the perceived level of helicopter parenting, the lower the level of assertiveness, and the higher the level of depression. Vigdal and Brønnick's [46] systematic review also concluded that helicopter parenting is associated with depression in young adults. Therefore, it is possible that rather than overprotective and controlling parenting, providing psychological independence through supportive parental relationships could reduce the possibility of depression in young adults.

This study found that social support played a mediating role between family strength and depression, which again suggests that interventions that seek to improve social support could reduce the prevalence of depression in young adults with weak family strength. Consistent with this finding, Cano et al [49] reported that interpersonal resources, such as family cohesion and social support, are associated with depression in young adults, and Mecha et al [50] found that perceived social support from family, friends, and significant others is associated with depression in young adults. Young adults with weak family strengths, such as those in out-of-home care, often struggle to receive social support from their families. Therefore, interventions are needed that promote social support through relationships with friends and significant others.

Limitations

This study has several limitations. First, it was limited to Korean young adults aged 19 to 24 years recruited through a web-based panel; therefore, the sample may not have fully represented the young emerging adult population in Korea. Future studies should recruit individuals from various age groups and geographic regions to enhance generalizability. Second, due to the cross-sectional nature of this study, causal inferences could not be made. The directionality of the relationships between family strength, psychological independence, social support, and depression variables, therefore, could not be determined. Because economic and social conditions change over time, longitudinal studies are needed to track changes in depression levels and related factors. Third, the use of self-reported measures could lead to social desirability and recall bias, which could have resulted in overreporting or underreporting. To mitigate this, participants were provided with clear instructions prior to commencing the web-based survey via mobile devices, ensuring they had sufficient time to respond thoughtfully. Finally, the analysis of the factors anticipated to influence depression in this study yielded an explanatory power of 23%. Given the potential existence of additional variables impacting depression among young adults that were not included in this investigation, future research should explore other factors that may contribute to depression to provide a more comprehensive understanding.

Practical Implications

Our findings suggest that strategies to enhance psychological independence and social support can significantly reduce depression among young adults with weak family strengths, underscoring the need for tailored preventive interventions. Mental health professionals should assess family strengths and depression levels to identify high-risk groups and provide targeted support. Psychological independence can be fostered through coaching-based programs involving influential peers or professionals outside the family [40], while social connections can be strengthened through group activities and mentoring initiatives [51]. Although the government provides professional psychological counseling services, these are not specifically designed for young adults, and limited awareness leads to underutilization. Therefore, it is essential to implement mental health support programs in universities and workplaces and improve accessibility through comprehensive promotion efforts.

Conclusions

Although the average depression scores in this study sample were slightly lower than those in previous Korean studies, they still indicated a high risk of depression, with young adult females having higher depression levels than young adult males. Young adults living with their parents were found to have lower psychological independence than those not living with their parents. The main finding of this study was the indirect effect of family strength on depression mediated through psychological independence and social support. This result suggests that strategies targeting these mediating factors could reduce the risk of depression in young adults who have weaker family strength. Mental health professionals should assess young adult family strengths, psychological independence, and social support to provide more focused preventive interventions for depression.

Acknowledgments

This work was supported by the National Research Foundation of Korea grant funded by the Korean government (Ministry of Science and Information and Communications Technology; NRF-2022R1F1A1063208).

Data Availability

The datasets used and/or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

All authors listed meet the authorship criteria according to the latest guidelines of the International Committee of Medical Journal Editors and agree with the manuscript. Conceptualization was done by SP and DN; SP, HK, and DN are responsible for

XSL•FO
RenderX

methodology; validation was done by SP and DN; HK and DN contributed to investigation; analysis was done by SP and DN; SK, SP, HK, and DN involved in writing—original draft preparation;
SK and DN involved in writing—review and editing; project administration was done by DN; funding acquisition was arranged by DN. All four authors have critically read the text and contributed with inputs and revisions, and all have also read and approved the final version of manuscript.

Conflicts of Interest

None declared.

References

1. World Health Organization. World Mental Health Report: Transforming Mental Health for All. : Geneva: World Health Organization; 2022. URL: http://hdl.handle.net/10713/20295 [accessed 2025-01-12]
2. Kessler RC, Berglund P, Demler O, Jin R, Merikangas KR, Walters EE. Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. Arch Gen Psychiatry 2005 Jun;62(6):593-602. [doi: 10.1001/archpsyc.62.6.593] [Medline: 15939837]
3. Ministry of Health and Welfare. Announcement of the results of the COVID-19 national mental health survey for the second quarter of 2022. Ministry of Health and Welfare website. URL: https://www.mohw.go.kr/board.es?mid=a10503000000&bid=0027&tag=&act=view&list_no=372545&cg_code=.arnet [accessed 2025-01-12]
4. Arnett JJ, Žukauskienė R, Sugimura K. The new life stage of emerging adulthood at ages 18-29 years: implications for mental health. Lancet Psychiatry 2014 Dec;1(7):569-576. [doi: 10.1016/S2215-0366(14)00080-7] [Medline: 26361316]
5. Defrain J. NF02-493 creating a strong family: looking at life from a family strengths perspective. Lincoln, NE: University of Nebraska-Lincoln Extension. 2002. URL: https://digitalcommons.unl.edu/extensionhist/345 [accessed 2025-01-12]
6. Sittner BJ, Hudson DB, DeFrain J. Using the concept of family strengths to enhance nursing care. MCN Am J Matern Child Nurs 2007;32(6):353-357. [doi: 10.1097/01.NMC.0000298130.16914.47] [Medline: 17968217]
7. Bae KE, Kim EH. Relationship of family strengths, family function, ego-identity and depression in adolescence. Child Health Nurs Res 2007;13(3):357-366.
8. Guerrero-Muñoz D, Salazar D, Constain V, Perez A, Pineda-Cañar CA, García-Perdomo HA. Association between family functionality and depression: a systematic review and meta-analysis. Korean J Fam Med 2021 Mar;42(2):172-180. [doi: 10.4082/kjfm.19.0166] [Medline: 32521579]
9. Wang E, Zhang J, Peng S, Zeng B. The association between family function and adolescents' depressive symptoms in China: a longitudinal cross-lagged analysis. Front Psychiatry 2021;12:744976. [doi: 10.3389/fpsyt.2021.744976] [Medline: 34975563]
10. United States National Library of Medicine. Young adult - mesh. National Center for Biotechnology Information. 2024. URL: https://www-ncbi-nlm-nih-gov [accessed 2025-01-20]
11. Arnett JJ. Emerging adulthood. A theory of development from the late teens through the twenties. Am Psychol 2000 May;55(5):469-480. [doi: 10.1037//0003-066x.55.5.469] [Medline: 10842426]
12. Moon JH, Cheon SM. Development and validation of psychological independence scale for university students. Korean J Couns 2020;21(3):253-276. [doi: 10.15703/kjc.21.3.202006.253]
13. Lin HH, Chew PYG, Wilkinson RB. Young adults' attachment orientations and psychological health across cultures: the moderating role of individualism and collectivism. J Relat Res 2017;8:e17. [doi: 10.1017/jrr.2017.17]
14. Jeon HO. Relationship of psychological separation, depression and antidepressive coping behaviors in university students. J Korea Acad Ind Coop Soc 2012 Jun 30;13(6):2655-2663. [doi: 10.5762/KAIS.2012.13.6.2655]
15. Ioannou M, Kassianos AP, Symeou M. Coping with depressive symptoms in young adults: perceived social support protects against depressive symptoms only under moderate levels of stress. Front Psychol 2018;9:2780. [doi: 10.3389/fpsyg.2018.02780] [Medline: 30692958]
16. Scardera S, Perret LC, Ouellet-Morin I, et al. Association of social support during adolescence with depression, anxiety, and suicidal ideation in young adults. JAMA Netw Open 2020 Dec 1;3(12):e2027491. [doi: 10.1001/jamanetworkopen.2020.27491] [Medline: 33275154]
17. Ryu N, Yoo S, Song H, Choi J, Chung A. Relationship between perceived family strength and psychological separation of high school students. Korean J Youth Stud 2011;18(10):65-87.
18. Gouveia-Pereira M, Abreu S, Martins C. How do families of adolescents with suicidal ideation behave? Psicol Reflex Crit 2014;27(1):171-178. [doi: 10.1590/S0102-79722014000100019]
19. Bu-Ja K, Youn-Shil C. The effect of family health on the depression of university students: mediating psychological independence and self-efficacy. Korean J Hum Dev 2016 Sep 30;23(3):71-90. [doi: 10.15284/kjhd.2016.23.3.71]
20. Analysis of the structural relationship among family strength, social support, hope and happiness of university students. Korea J Youth Counsel 2017 Nov;25(2):93-122. [doi: 10.35151/kyci.2017.25.2.005]
21. Wi J, Cho G. An analysis of structural relations among children's perceived family strength, self-resilience, social support, stress-coping behavior, and psychological well-being. J Korean Assn Learn Cent Curric Instr 2014;14(8):237-264.

XSL•FO
RenderX

22. Cobb S. Social support as a moderator of life stresss. Psychosom Med 1976 Sep;38(5):300-314. [doi: 10.1097/00006842-197609000-00003]
23. Lakey B, Cohen S. Social support theory and measurement. In: Cohen S, Underwood LG, Gottlieb BH, editors. Social Support Measurement and Intervention: A Guide for Health and Social Scientists: New York, NY: Oxford University Press; 2000:29-52. [doi: 10.1093/med:psych/9780195126709.003.0002]
24. Thoits PA. Mechanisms linking social ties and support to physical and mental health. J Health Soc Behav 2011 Jun;52(2):145-161. [doi: 10.1177/0022146510395592] [Medline: 21673143]
25. Faul F, Erdfelder E, Buchner A, Lang AG. Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. Behav Res Methods 2009 Nov;41(4):1149-1160. [doi: 10.3758/BRM.41.4.1149] [Medline: 19897823]
26. Yoo Y, Lee I, Kim S, Choi H. Development of Korea Family Strengths Scale (KFSS-II). J Korean Home Manage Assoc 2013 Aug 30;31(4):113-129. [doi: 10.7466/JKHMA.2013.31.4.113]
27. Cutrona CE, Russell DW. The provisions of social relationships and adaptation to stress. Adv Pers Relatsh 1987;1(1):37-67.
28. Yoo YR, Lee JY. Adult attachment and help-seeking intention: the mediating roles of self-concealment, perceived social support, and psychological distress. Korean J Couns Psychother 2006;18(2):441-460.
29. Lee S, Oh ST, Ryu SY, et al. Validation of the Korean version of Center for Epidemiologic Studies Depression Scale-Revised (K-CESD-R). Korean J Psychosom Med 2016;24(1):83-93. [doi: 10.0000/kjpm.2016.24.1.83]
30. Eaton WW, Smith C, Ybarra M, Muntaner C, Tien A. Center for Epidemiologic Studies Depression Scale: Review and Revision (CESD and CESD-R). In: Maruish ME, editor. Use of Psychological Testing for Treatment Planning and Outcomes Assessment, 3rd edition 2004:363-377. [doi: 10.4324/9781410610614]
31. Yi SJ, Kim M, Park I. Investigating influencing factors on premenstrual syndrome (PMS) among female college students. BMC Womens Health 2023 Nov 10;23(1):592. [doi: 10.1186/s12905-023-02752-y] [Medline: 37950208]
32. Kim MH, Park SH. Effects of sleep duration on stress and depression in college students. J Korean Soc Integr Med 2021;9(4):261-270. [doi: 10.15268/ksim.2021.9.4.261]
33. The impact of employment instability on mental health: the case of South Korea. Health Soc Welfare Rev 2018 Sep;38(3):129-160. [doi: 10.15709/hswr.2018.38.3.129]
34. Song DW, Lee K. A cultural analysis of the varying modes of survival and the particular structures of feeling among young adults in contemporary South Korea in an era of fierce competition and widespread social uncertainty. Korean J Commun Inform 2017 Aug 31;84:28-98. [doi: 10.46407/kjci.2017.08.84.28]
35. Kim GE, Jo MW, Shin YW. Increased prevalence of depression in South Korea from 2002 to 2013. Sci Rep 2020 Oct 12;10(1):16979. [doi: 10.1038/s41598-020-74119-4] [Medline: 33046758]
36. Salk RH, Hyde JS, Abramson LY. Gender differences in depression in representative national samples: Meta-analyses of diagnoses and symptoms. Psychol Bull 2017 Aug;143(8):783-822. [doi: 10.1037/bul0000102] [Medline: 28447828]
37. Moon YS, Yoon Y, Hwang DS, Pak TK. Navigating the discussion of mental illness with Korean Americans. J Psychiatr Pract 2025 Jan 1;31(1):8-12. [doi: 10.1097/PRA.0000000000000830] [Medline: 39836675]
38. Park SY, Bernstein KS. Depression and Korean American immigrants. Arch Psychiatr Nurs 2008 Feb;22(1):12-19. [doi: 10.1016/j.apnu.2007.06.011]
39. Martínez-Hernáez A, Carceller-Maicas N, DiGiacomo SM, Ariste S. Social support and gender differences in coping with depression among emerging adults: a mixed-methods study. Child Adolesc Psychiatry Ment Health 2016;10(1):2. [doi: 10.1186/s13034-015-0088-x] [Medline: 26744601]
40. Oda M, Yoshioka SI. Factors influencing psychological independence in adolescents and their relationship to coaching-based support from significant others. Yonago Acta Med 2021 Feb;64(1):34-45. [doi: 10.33160/yam.2021.02.007] [Medline: 33642902]
41. Kang YJ. Relationship quality between unmarried adult children and their coresident parents: focus on intergenerational exchanges and family values. Korean J Clin Lab Sci 2016 Aug 31;27(3):387-403. [doi: 10.7856/kjcls.2016.27.3.387]
42. Smith OA, Nelson JA, Adelson MJ. Interparental and parent–child conflict predicting adolescent depressive symptoms. J Child Fam Stud 2019 Jul;28(7):1965-1976. [doi: 10.1007/s10826-019-01424-6]
43. Ramos G, Ponting C, Bocanegra E, et al. Discrimination and internalizing symptoms in rural latinx adolescents: the protective role of family resilience. J Clin Child Adolesc Psychol 2022;51(6):997-1010. [doi: 10.1080/15374416.2021.1923018] [Medline: 34038290]
44. Rahman A, Sanchez M, Bursac Z, et al. Depressive symptoms, family cohesion, and acculturation gap conflicts among Latinx emerging adults living in the United States. Fam Relat 2023 Jul;72(3):932-947. [doi: 10.1111/fare.12722]
45. Jeon HK. The effect of maternal psychological control on the depression of university students: the mediation effects of the need for approval, psychological separation and adult attachment. Stud Korean Youth 2016 Aug;27(3):97-131. [doi: 10.14816/sky.2016.27.3.97]
46. Vigdal JS, Brønnick KK. A systematic review of “helicopter parenting” and its relationship with anxiety and depression. Front Psychol 2022;13:872981. [doi: 10.3389/fpsyg.2022.872981] [Medline: 35693486]
47. Brown MT, Jensen M, Hussong AM. Parent-emerging adult text interactions and emerging adult perceived parental support of autonomy. J Soc Pers Relat 2024 Feb;41(2):499-520. [doi: 10.1177/02654075231167347]

XSL•FO
RenderX

48. Kim GY, Park JH. Perceived helicopter parenting and depression among Korean university students: the mediating effect of assertiveness. Korean J Child Stud 2019 Aug;40(4):165-177. [doi: 10.5723/kjcs.2019.40.4.165]
49. Cano M, Castro FG, De La Rosa M, et al. Depressive symptoms and resilience among hispanic emerging adults: examining the moderating effects of mindfulness, distress tolerance, emotion regulation, family cohesion, and social support. Behav Med 2020;46(3-4):245-257. [doi: 10.1080/08964289.2020.1712646] [Medline: 31935162]
50. Mecha P, Martin-Romero N, Sanchez-Lopez A. Associations between social support dimensions and resilience factors and pathways of influence in depression and anxiety rates in young adults. Span J Psychol 2023 May 2;26:e11. [doi: 10.1017/SJP.2023.11] [Medline: 37129012]
51. Anderson AJ, DuBois DL. Are adults influenced by the experience of mentoring youth? A scoping review. J Community Psychol 2023 Apr;51(3):1032-1059. [doi: 10.1002/jcop.22954] [Medline: 36322942]

Abbreviations

K-CESD-R: Korean version of the Center for Epidemiologic Studies Depression Scale-Revised
STROBE: Strengthening the Reporting of Observational Studies

Edited by A Mavragani; submitted 22.01.25; peer-reviewed by HJ Park, IA Aziz; revised version received 14.04.25; accepted 24.04.25; published 27.06.25.

Please cite as:
Kim S, Park S, Kim H, Noh D
Mediating Effects of Psychological Independence and Social Support on the Association Between Family Strength and Depression in Young Korean Adults: Cross-Sectional Study
JMIR Form Res 2025;9:e71485
URL: https://formative.jmir.org/2025/1/e71485
doi:10.2196/71485

© Sunyoung Kim, Suin Park, Hyunlye Kim, Dabok Noh. Originally published in JMIR Formative Research (https://formative.jmir.org), 27.6.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Determinant Factors of Stress in Caregivers of Patients With Schizophrenia: Cross-Sectional Study

Isymiarni Syarif[1]; Hasnawati Amqam[1], DrPH; Saidah Syamsuddin[2], Dr med; Veni Hadju[3], PhD; Syamsiar Russeng[4], DrPH; Yusran Amir[5], MPH

[1]Department of Public Health, Faculty of Public Health, Hasanuddin University, Jl. Perintis Kemerdekaan Km. 10 Tamalanrea, Makassar, Indonesia
[2]Department of Psychiatry, Faculty of Medicine, Hasanuddin University, Makassar, Indonesia
[3]Department of Nutrition, Faculty of Public Health, Hasanuddin University, Makassar, Indonesia
[4]Department of Occupational Safety and Health, Faculty of Public Health, Hasanuddin University, Makassar, Indonesia
[5]Department of Health Administration and Policy, Faculty of Public Health, Hasanuddin University, Makassar, Indonesia

Corresponding Author:
Hasnawati Amqam, DrPH
Department of Public Health, Faculty of Public Health, Hasanuddin University, Jl. Perintis Kemerdekaan Km. 10 Tamalanrea, Makassar, Indonesia

Abstract

Background: Caregivers of individuals with schizophrenia face ongoing psychological and emotional burdens due to the chronic and relapsing nature of the disorder and the complexity of caregiving. Prolonged exposure to caregiving stress characterized by emotional exhaustion, role overload, and lack of social support has been consistently associated with poor mental health outcomes among caregivers, including depression and anxiety.

Objective: This study aimed to assess stress levels among caregivers of patients with schizophrenia and identify the key determinants of caregiver stress.

Methods: This study used a cross-sectional survey that was conducted between June and August 2024 at the Labakkang District Health Center, South Sulawesi, Indonesia. A total of 110 female caregivers participated in the study. Data were collected using validated questionnaires to measure stress levels and related factors. Statistical analyses included chi-square tests to identify associations and partial least squares structural equation modeling to examine the strength and direction of relationships between variables.

Results: This study included 110 female caregivers of individuals with schizophrenia. The majority were early older people (48/110, 44%), had a basic level of education (elementary to junior high school; 45/110, 46%), were unemployed (83/110, 75%), and had been providing care for more than 10 years (42/110, 38%). A total of 58 of 110 (53%) caregivers experienced mild levels of stress, while 63 of 110 (57%) caregivers reported a moderate caregiving burden. Additionally, 64 of 110 (58%) caregivers reported challenges related to patient treatment nonadherence, and 58 of 110 (53%) caregivers experienced low levels of social stigma. Most caregivers (69/110, 63%) adopted adaptive coping strategies; however, more than half reported low levels of knowledge (59/110, 54%) and limited access to health information (73/110, 66%). The chi-square analysis identified several statistically significant associations with stress: age (P=.03), education (P<.001), caregiving burden (P<.001), knowledge (P<.001), coping strategies (P<.001), treatment nonadherence (P=.004), and perceived stigma (P=.003). Further, partial least squares structural equation modeling analysis showed that caregiving burden (r=0.672), stigma (r=0.921), and limited knowledge (r=0.909) were positively correlated with stress. In contrast, social support was strongly negatively associated with stress (r=−0.872), indicating its protective role.

Conclusions: These findings underscore the critical need for targeted interventions that enhance social support networks, reduce stigma, and strengthen caregivers' coping capacities. Strengthening these dimensions is essential to mitigating the psychological toll of caregiving and sustaining caregivers' functional well-being. Evidence increasingly supports that empowering caregivers through structured support systems and educational initiatives can substantially alleviate stress-related burdens and improve care continuity for individuals with schizophrenia.

(JMIR Form Res 2025;9:e70648) doi:10.2196/70648

KEYWORDS

caregiver; stress caregiver; schizophrenia; determinant factors; cross-sectional study

XSL•FO
RenderX

Introduction

The role of caregivers in supporting individuals with mental disorders, such as schizophrenia, is crucial, particularly in assisting patients with medication adherence, providing emotional support, and helping with daily living activities. However, this often results in significant physical, emotional, and social burdens [1]. These challenges are exacerbated when caregivers are confronted with unpredictable psychotic symptoms, limited knowledge about their condition, and social stigma from the surrounding community [2,3].

Several studies have indicated that caregivers of individuals with schizophrenia experience moderate to high levels of stress, which directly impacts their psychological well-being, manifesting in symptoms such as anxiety, depression, and chronic fatigue [4,5]. This emotional burden not only affects caregivers' quality of life but also diminishes the effectiveness of patient care, potentially leading to higher relapse rates, increased hospitalization, and additional strain on mental health systems [6,7].

In Indonesia, the situation is even more complex owing to limited access to mental health information, pervasive societal stigma, and a lack of structured social support systems [8,9]. Many caregivers struggle to fully understand a patient's condition and often feel inadequately equipped or unsupported in providing optimal care. Moreover, there are few community-based or primary care interventions specifically designed to help caregivers manage their psychosocial burdens. This highlights the urgent need for evidence-based approaches that are practically applicable in local contexts and primary health care settings.

The psychosocial stressors faced by caregivers are highly complex and interrelated, encompassing factors such as low mental health literacy, prolonged caregiving duration, caregiver's psychological state, subjective burden, patient nonadherence to treatment, and insufficient family or social support [10,11]. Therefore, a comprehensive psychosocial approach is required. Previous research has demonstrated that strong social support, improved mental health literacy, and adaptive coping skills are protective factors against caregiver stress [12].

Nonetheless, most existing studies have addressed emotional and social burdens in isolation without considering the dynamic interplay among multiple psychosocial factors [13]. Furthermore, the application of advanced statistical techniques, such as partial least squares structural equation modeling (PLS-SEM), to study caregiver stress remains limited, especially in evolving countries like Indonesia. This method holds significant promise for mapping complex latent relationships among variables, such as stress, stigma, knowledge, subjective burden, social support, and trust [14,15]. Such studies are essential for advancing the development of robust evidence-based interventions [16].

Therefore, this study aimed to holistically identify and analyze the psychosocial factors influencing stress among caregivers of individuals with schizophrenia in Indonesia using a cross-sectional design and PLS-SEM approach. These findings are expected to offer meaningful empirical contributions to the field of community mental health and serve as a foundation for developing more contextual, humane, and effective interventions to enhance caregiver well-being and patient quality of life [17].

Methods

Design, Setting, and Participants

This study used a cross-sectional survey design to assess stress levels and identify the key determinant factors among caregivers of individuals with schizophrenia. A cross-sectional web-based survey was conducted with adults (aged 18 y and older) living with a mental health disorder. The self-reported survey was administered on the web-based survey platform Qualtrics. The survey is described according to the CHERRIES (Checklist for Reporting Results of Internet E-Surveys) [18]. The study was conducted within the jurisdiction of 3 community health centers in Labakkang Subdistrict, Pangkep Regency, South Sulawesi, Indonesia. At the time of the survey, there were 138 registered caregivers of individuals diagnosed with schizophrenia, of whom 110 were female. These caregivers were distributed across the 3 participating community health centers.

The participants were selected through purposive sampling. The sample size was determined using the Roscoe formula [19], which states that for multivariate analysis, the sample size should be at least 10 times the number of variables (R=n×<10), where n is the number of variables. Given that the study involved 11 variables, a minimum of 110 participants was deemed necessary.

Purposive sampling was used to recruit caregivers who met the following inclusion criteria: (1) registered as caregivers for a person diagnosed with schizophrenia at one of the selected health centers, (2) aged 18 years or older, and (3) actively involved in daily care activities. Caregivers who were unable to participate because of illness, language barriers, or unwillingness to provide consent were excluded. The selected sample represented diverse demographic and caregiving backgrounds, contributing to the generalizability of the findings to the target population.

Data collection was conducted over three months: 28 caregiver responses were collected in June, 55 in July, and 27 in August. These data were used to assess caregivers' stress levels and identify contributing personal, situational, and environmental factors.

Data Collection

The survey was conducted between June and August 2024, through direct home visits to caregivers of individuals with schizophrenia. To maximize response rates and ensure data quality, trained public health enumerators were engaged in the survey in collaboration with mental health officers from participating health centers. The fieldwork was coordinated by local government authorities and the Pangkep District Health Office. Community health volunteers (cadres) also supported field activities by facilitating access to and encouraging community participation.

XSL•FO
RenderX

Data were collected using a structured questionnaire that included questions regarding emotional burden, care demands, social support, and caregiver stress levels. Stress levels were evaluated using the National Alliance for Caregiving Questionnaire in collaboration with the American Association of Retired Person [20]. Stigma was measured using the Internalized Stigma of Mental Illness scale [21]. Caregivers' coping mechanisms were assessed using the Coping Orientation to Problems Experienced Inventory and McMaster Family Assessment Device [22,23], and social support was evaluated using the Multidimensional Perceived Social Support questionnaire [24]. The burden of caregiving was quantified using Zarit's Caregiver Burden [25], and caregiver knowledge was measured through the Knowledge Assessment Schizophrenia Test [26,27]. Additionally, medication noncompliance among patients with schizophrenia was assessed using the Medication Compliance Report Scale [28]. Each instrument was rigorously validated at the Bungoro Health Center, Pangkep Regency, to ensure high reliability and validity (Table 1).

Table . Validity and reliability of assessment tools determinant factors of stress level on caregivers of patients with schizophrenia in Pangkep Regency, Indonesia, 2024.

Assessment tools	Reliability
National Alliance for Caregiving Questionnaire in collaboration with AARP[a]	0.982[b]
Internalized Stigma of Mental Illness scale	0.9[c]
The COPE[d] Inventory and the McMaster Family Assessment Device	0.989[c]
Multidimensional Perceived Social Support	0.952[b]
Zarit's Caregiver Burden	0.989[c]
Knowledge Assessment Schizophrenia Test	0.966[c]
Medication Compliance Report Scale	0.949[c]

[a]AARP: American Association of Retired Persons.
[b]$P<.01$.
[c]$P<.001$.
[d]COPE: Coping Orientation to Problems Experienced.

Data Statistical Analysis

Manual thematic coding was conducted to categorize data into predefined themes, such as stress levels and determinant factors, including patient characteristics, caregiver burden, stigma, knowledge, coping strategies, access to health information, medication nonadherence, social support, and caregiving duration. The coding framework was informed by the existing literature and refined during the initial data familiarization. The coded data were then tabulated using Excel (Microsoft Corp) and further analyzed using SPSS (version 29; IBM Corp) for both descriptive and inferential statistics.

To examine the relationship between exogenous and endogenous variables, the data were analyzed using PLS-SEM with SmartPLS (version 4; SmartPLS GmbH) [29-32]. The evaluation process consists of two stages: the measurement model and the structural model [32,33]. The PLS-SEM was selected for this study for several reasons. First, it is well-suited for advancing existing theories [34]. Second, PLS-SEM is highly effective for analyzing complex models, especially in exploratory research. Third, it allows for the analysis of the entire model as a cohesive entity, rather than isolating individual components [35]. Finally, PLS-SEM enables the simultaneous evaluation of both structural and measurement models, ensuring more robust and precise results [36]. The exogenous variables included personal, situational, and social environmental determinants, while the endogenous variables included stress levels. The proposed model is shown in the supplement in Multimedia Appendix 1 and Figures 1 and 2. This study quantitatively analyzed the determinants of caregiver stress in schizophrenia care, focusing on caregiver coping, burden of care, stigma, medication nonadherence, and social support. Chi-square tests were used to examine the interrelationships among the measurable variables.

XSL•FO
RenderX

Figure 1. PLS model specification diagram (first estimation) determinant factors of stress level on caregiver of patients with schizophrenia in Pangkep Regency, Indonesia, 2024. PLS: partial least squares.

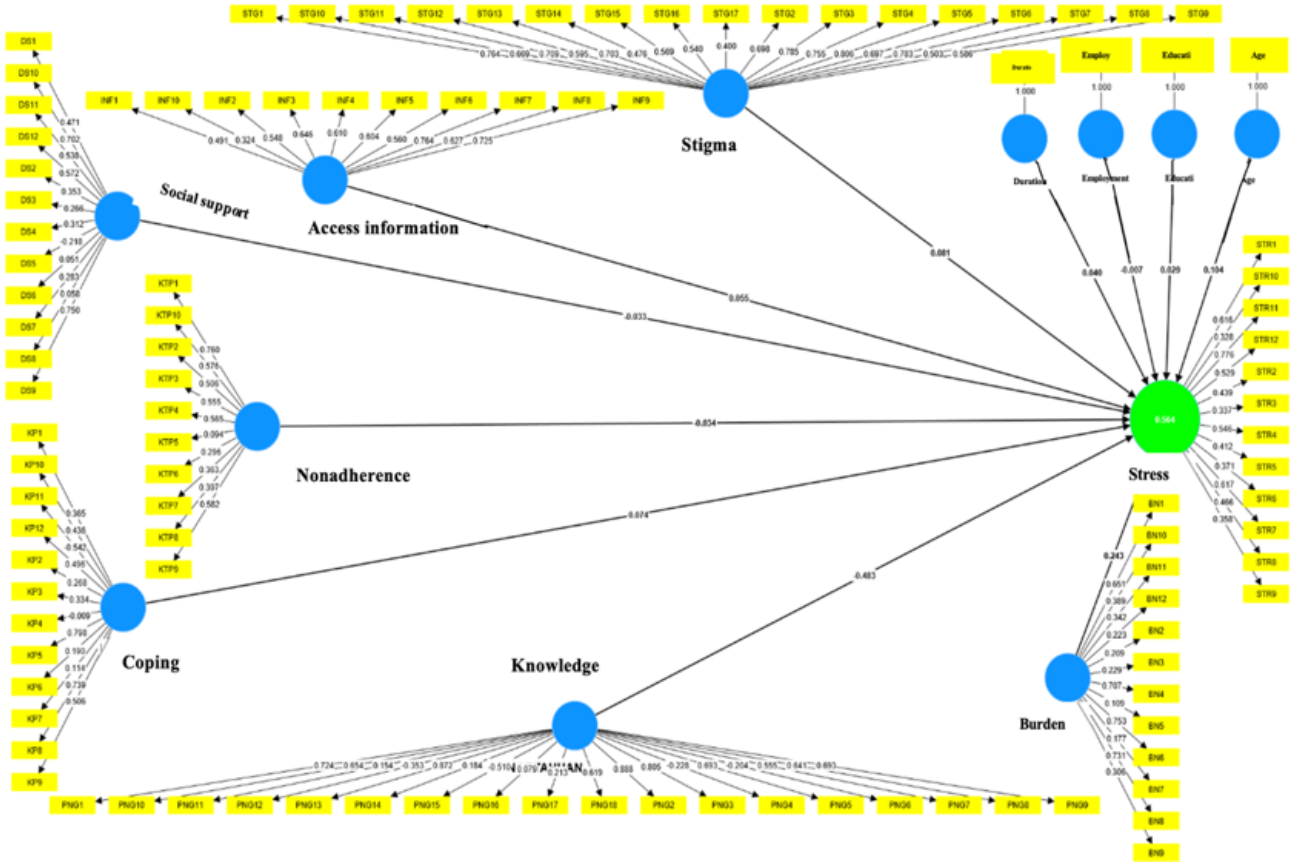

Figure 2. Outer model diagram of PLS respecification (second estimation) determinant factors of stress level on caregiver of patients with schizophrenia in Pangkep Regency, Indonesia, 2024. PLS: partial least squares.

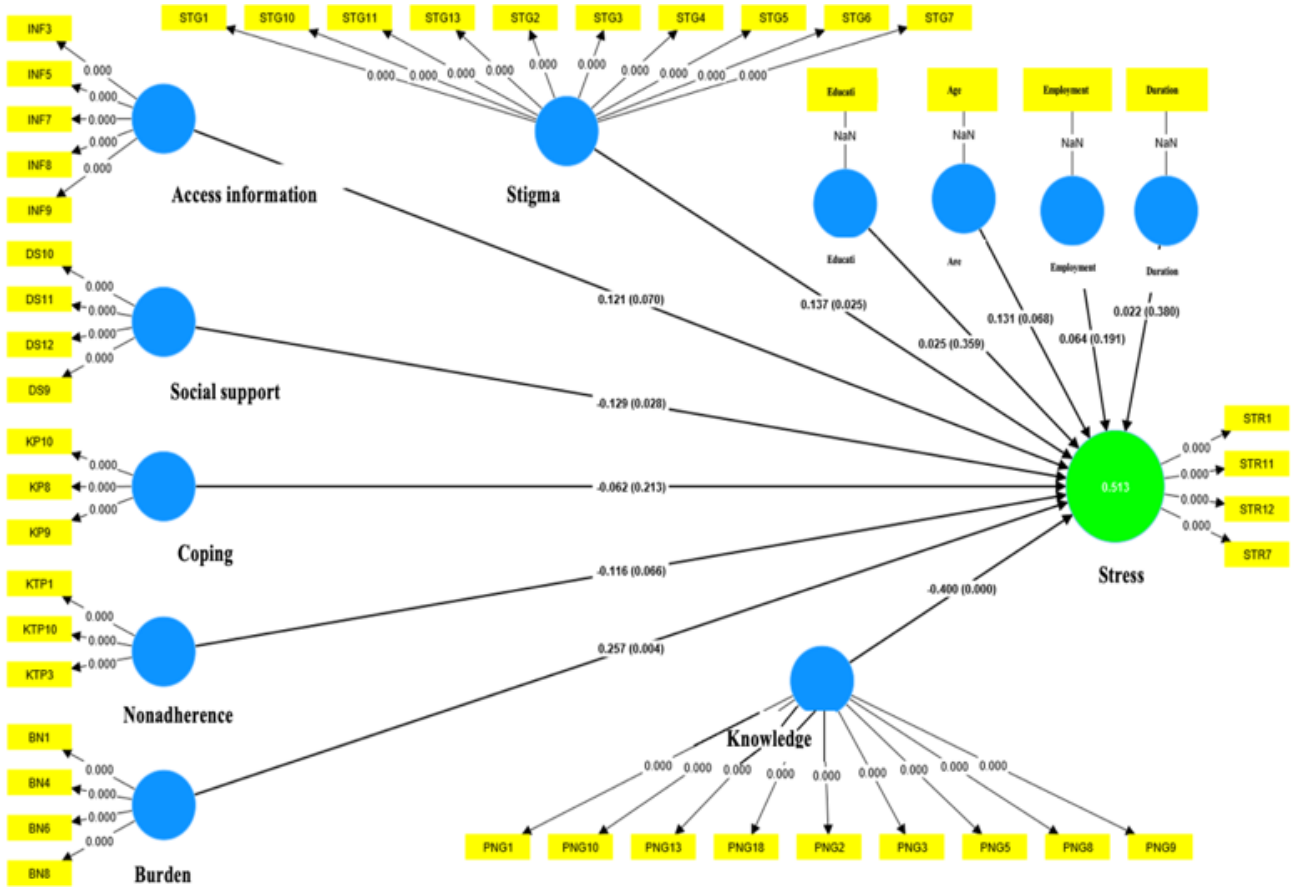

XSL•FO
RenderX

Ethical Considerations

Ethical approval was obtained from the Faculty of Public Health Ethics Committee, Hasanuddin University (25240930270). Written informed consent was obtained from all participants before data collection. The participants were assured of their right to withdraw at any time, and strict confidentiality was maintained by anonymizing and securely storing all data. Nonmonetary incentives were provided to acknowledge the participants' time and contribution, as the study involved minimal risk.

Results

Stress Levels of Caregivers

Caregivers of individuals with schizophrenia often experience elevated stress levels owing to the complex emotional, physical, and social demands associated with caregiving. These stressors can substantially affect psychological well-being and social functioning. As shown in Table 2, a majority of caregivers (58/110, 53%) reported experiencing mild stress, while 6 out of 110 (5%) caregivers reported severe stress. These findings suggest that caregiving for individuals with schizophrenia imposes considerable psychological pressure, even though extreme stress levels are not prevalent among most respondents. According to Lazarus and Folkman's stress and coping theory, stress arises when individuals perceive that their environmental demands exceed their available coping resources. In this context, although most caregivers did not exhibit signs of severe stress, the predominance of mild stress nonetheless indicated a persistent psychological strain embedded in the caregiving experience.

Table . Distribution of stress levels of caregivers of patients with schizophrenia in Pangkep Regency, Indonesia.

Stress levels	Value, n (%)
Normal	46 (42)
Mild stress	58 (53)
Severe stress	6 (5)
Total	110 (100)

From a practical perspective, these findings underscore the need for structured support systems at the community and primary health care levels. Early intervention to manage mild stress is essential to prevent escalation to severe stress, particularly considering the long-term burden of caring for individuals with schizophrenia. Primary care centers and mental health services can play an active role by providing education, stress management training, and access to caregiver support networks. This may include the promotion of mobile health interventions as digital platforms for ongoing monitoring and caregiver education.

Personal Determinant Factors

Understanding the personal determinant factors that influence caregivers is crucial for addressing their stress and challenges. The analysis categorizes these determinants into 4 key aspects: age, education, employment, and care duration. Each of these factors plays a significant role in shaping the caregiving experience and the associated psychological burden.

The majority of caregivers (48/110, 44%) were early older individuals (Table 3). Of the 110 caregivers, 37 (34%) were late adults. Age influences both physical and psychological capacities to manage caregiving-related stress [37]. Early older caregivers may benefit from greater life experiences, but may also begin to face declining energy levels. In practice, this highlights the need for tailored attention to the physical capacity and emotional support of older caregivers to help them sustain their caregiving roles without excessive fatigue.

XSL•FO
RenderX

Table . Distribution of personal determinant factors of patients with schizophrenia in Pangkep Regency, Indonesia, 2024.

Personal determinant factors	Value, n (%)
Age	
Early adult	25 (22)
Late adult	37 (34)
Early older individuals	48 (44)
Education	
Without school	23 (21)
Elementary school	50 (45)
Senior high school	33 (30)
Higher school	4 (4)
Employment	
Unemployed	83 (75)
Employed	27 (25)
Duration of care (years)	
Less than 1 to 4	31 (28)
5 to 10	37 (34)
Over 10	42 (38)

The highest proportion of caregivers had only completed elementary school (50/110, 45%), whereas 23 of 110 (21%) caregivers had no formal education (Table 3). Lower educational attainment may affect their ability to access health care information, cope with caregiving stress, and seek professional assistance when required. Individuals with lower educational backgrounds often face difficulties in accessing, processing, and understanding essential information for caring for people with mental illnesses, which may exacerbate caregivers' emotional burdens [38]. From a practical standpoint, simple, visual-based educational tools, particularly those delivered through technology-based interventions, may help reduce caregiver stress by providing easy-to-understand information and enhancing caregivers' capacity to care for individuals with schizophrenia.

A striking 83 out of 110 (75%) caregivers were unemployed, whereas only 25% (27/110) were employed (Table 3). This high rate of unemployment among caregivers may intensify stress levels, as the inability to work or maintain stable employment often leads to feelings of being trapped in the caregiving role without adequate financial or social support [39]. In practical terms, economic assistance or skill training programs that enable caregivers to pursue flexible employment opportunities could help alleviate both financial and emotional burdens. The inability to balance work and caregiving responsibilities may also contribute to burnout and reduce overall quality of life.

The duration of caregiving varied: 42 out of 110 (38%) caregivers provided care for over 10 years, followed by 37 out of 110 (34%) caregivers for 5 to 10 years (Table 3). The duration of caregiving may affect caregivers' stress levels. According to the stress theory, the longer an individual serves as a caregiver, the greater the potential for emotional and physical exhaustion, commonly referred to as burnout [38,40]. In practical terms, interventions that provide continuous support, such as coping training and psychosocial assistance, are crucial for reducing long-term caregiver stress and helping caregivers manage the challenges associated with prolonged caregiving.

Situational Determinant Factors

Caregiving for individuals with schizophrenia presents significant challenges, often leading to psychological, emotional, and physical burdens for caregivers. Situational determinant factors play a crucial role in understanding how external conditions influence caregivers' well-being and ability to provide adequate care. These factors include caregiver burden, coping mechanisms, and knowledge of schizophrenia, all of which have direct implications for caregivers' quality of care and overall mental health.

Based on Table 4, it can be concluded that the burden experienced by caregivers of patients with schizophrenia was significant. The majority (63/110, 57%) of caregivers reported experiencing moderate burden, while 33 out of 110 (30%) caregivers reported a heavy burden reflecting the ongoing pressure resulting from complex caregiving demands. These findings are consistent with the Transactional Model of Stress and Coping, which posits that stress arises when demands exceed an individual's available resources [41]. In practical terms, this highlights the need for interventions that not only focus on stress management but also strengthen social support, education, and technological tools such as mobile health to help caregivers cope with challenges adaptively.

XSL•FO
RenderX

Table . Distribution of situational determinant factors of patients with schizophrenia in Pangkep Regency, Indonesia, 2024.

Situational determinant factors	Value, n (%)
Caregiver burden	
Minimal burden	14 (13)
Moderate burden	63 (57)
Heavy burden	33 (30)
Coping caregiver	
Maladaptive coping	41 (37)
Adaptive coping	69 (63)
Knowledge	
Low knowledge	59 (54)
Adequate knowledge	51 (46)
Medication nonadherence	
High nonadherence	46 (42)
Low nonadherence	64 (58)

Table 4 shows the coping mechanisms, indicating that 69 out of 110 (63%) caregivers applied adaptive coping, while 41 out of 110 (37%) caregivers still used maladaptive coping. Based on the Transactional Model of Stress and Coping, these findings suggest that the majority of caregivers are able to adapt to stress constructively, but a significant proportion still rely on maladaptive coping, highlighting the need for targeted interventions [40,41]. Interventions such as coping skills training and social support are effective in enhancing adaptive coping and reducing caregiver burden.

Table 4 shows the level of knowledge about schizophrenia. The data indicated that 59 of 110 (54%) caregivers had low knowledge, while 46% (51/110) possessed adequate knowledge. Low knowledge may influence caregivers' perceptions of risk, benefits of actions, and self-efficacy in caring for individuals with schizophrenia [42]. Practically, these findings highlight the importance of structured educational interventions, such as digital psychoeducation or community-based training, to enhance caregivers' mental health literacy. Nurfurqoni et al [8] demonstrated that digital psychoeducational interventions significantly improve caregivers' knowledge and skills, which in turn, positively impact burden reduction and coping improvement.

Table 4 shows that another factor influencing caregiver well-being is their level of medication nonadherence. The data indicated that 46 of 110 (42%) caregivers had high medication nonadherence. The finding that 42% (46/110) of caregivers reported high medication nonadherence highlights a critical risk factor for schizophrenia management. According to the Health Belief Model, nonadherence may result from perceived barriers or low perceived benefits of consistent medication use [42]. Practically, this underscores the need for targeted interventions that address caregiver knowledge, beliefs, and routines, such as mobile-based reminders, psychoeducation, and caregiver-inclusive treatment planning, to improve adherence and prevent relapse.

Social Environment Determinant Factors

The social environment plays a crucial role in determining stress levels among caregivers of patients with schizophrenia, where low social support, societal stigma, and limited access to health information contribute to an increased psychological burden.

Figure 3A shows that the stigma related to schizophrenia is a significant challenge for caregivers. The data revealed 2 categories of stigma: low stigma, with 58 of 110 (53%) caregivers and high stigma, with 13 of 110 (12%) caregivers. Internalized stigma can negatively affect caregivers' mental health, self-efficacy, and willingness to seek support [43]. Practically, these findings highlight the need for community-based interventions such as peer support groups and antistigma education to reduce negative perceptions and promote caregiver resilience, particularly for those still experiencing high stigma.

XSL•FO
RenderX

Figure 3. Distribution of social environment determinant factors—(A) stigma; (B) social support; and (C) access to health information—of patients with schizophrenia in Pangkep Regency, Indonesia 2024.

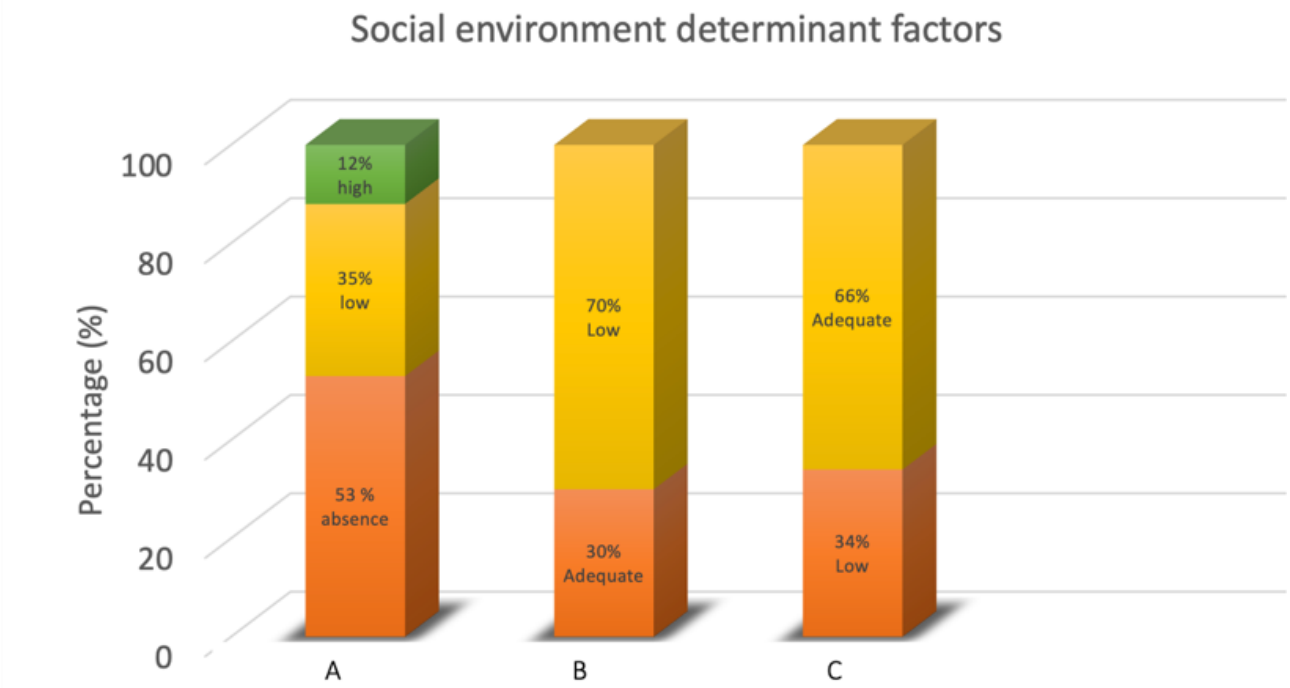

Figure 3B shows that social support is a crucial factor in reducing stress and enhancing caregivers' well-being. The data indicated that 30 of 110 (33%) caregivers experienced low social support, while 70 of 110 (77%) caregivers received adequate support. Social support functions as a resource that helps individuals cope with stress, serving as both an emotional and practical buffer for caregivers caring for individuals with schizophrenia [44]. Adequate social support can reduce burnout and improve caregivers' quality of life, which, in turn, enhances the quality of patient care. Conversely, lack of social support increases vulnerability to stress, consistent with Social Support Theory, which suggests that caregivers who feel isolated are more susceptible to emotional stress and difficulties in caring for patients [45]. These findings highlight the importance of community-based interventions to strengthen social support networks and improve access to mental health services to reduce caregiver burden and enhance the quality of care provided to patients.

Figure 3C shows that access to health information played a crucial role in determining the effectiveness of caregivers' support for patients. The data revealed that 73 of 110 (66%) caregivers had limited access to health information, while only 37 of 110 (34%) caregivers had sufficient access. Limited access to health information can hinder decision-making and increase stress, consistent with Health Information Access Theory [46]. Practically, such limitations can reduce the quality of care provided to patients with schizophrenia. Therefore, interventions such as digital education and health literacy programs are essential to strengthen caregivers' capacities.

Significant Determinant Factors of Caregiver Stress

Caregiving for individuals with schizophrenia presents significant challenges that often lead to elevated stress levels. Understanding the key determinants of caregiver stress is essential for developing targeted interventions to enhance well-being. The results indicated that certain factors were significantly related to caregiver stress ($P<.05$).

The chi-square analysis identified several factors that were significantly associated with caregiver stress. Regarding personal determinants, the findings revealed that age was significantly associated with caregiver stress ($P=.03$), with younger caregivers tending to experience higher stress levels than older caregivers(Table 5). The educational level also demonstrated a significant relationship with stress levels ($P<.001$), with caregivers with lower educational attainment experiencing higher stress than those with secondary or higher education levels. In contrast, employment status did not show a statistically significant relationship with caregiver stress ($P=.05$), although employment may provide positive distractions and emotional support. Similarly, caregiving duration did not exhibit a significant association with caregiver stress ($P=.07$), suggesting that caregiving experience beyond 10 years does not necessarily alleviate or exacerbate stress.

XSL•FO
RenderX

Table . Determinant factors of stress level on caregivers of patients with schizophrenia in Pangkep Regency, Indonesia, cross-sectional study, 2024.

Personal determinant factors	Stress level			Total, n (%)	*P* value[a]
	Normal, n (%)	Mild stress, n (%)	Severe stress, n (%)		
Age					.03[a]
Early adult	8 (7.3)	15 (13.7)	2 (1.8)	25 (22.7)	
Late adult	16 (14.5)	19 (17.3)	2 (1.8)	37 (33.6)	
Early older individual	22 (20)	24 (21.8)	2 (1.8)	48 (43.7)	
Education					<.001
Without school	12 (10.9)	11 (10)	0 (0)	23 (20.9)	
Elementary school	28 (25.5)	10 (9.1)	12 (10.9)	50 (45.5)	
Senior high school	6 (5.5)	26 (23.7)	1 (0.9)	33 (30)	
Higher school	0 (0)	4 (3.6)	0 (0)	4 (3.6)	
Employment					.05
Employed	16 (14.5)	10 (9.1)	1 (0.9)	27 (24.5)	
Unemployed	30 (27.3)	48 (43.6)	5 (4.5)	83 (75.5)	
Duration of caregiving (years)					.07
<1-4	18 (16.4)	13 (11.7)	0 (0)	31 (28.2)	
5-10	15 (13.6)	21 (19.1)	1 (0.9)	37 (33.6)	
10+	13 (11.8)	24 (16.4)	5 (4.5)	42 (38.2)	
Burden of caregiver					<.001
Minimal	13 (11.8)	1 (0.9)	0 (0)	14 (12.7)	
Moderate burden	32 (29.1)	25 (24.5)	4 (3.6)	63 (57.3)	
High burden	1 (0.9)	30 (27.2)	2 (1.8)	33 (30)	
Coping of caregiver					<.001
Adaptive	10 (9.1)	54 (49.1)	6 (5.5)	69 (62.7)	
Maladaptive	36 (32.7)	4 (3.6)	1 (0.9)	41 (37.3)	
Knowledge					<.001
Adequate	45 (40.9)	6 (5.4)	0 (0)	51 (46.7)	
Low	1 (0.9)	36 (32.7)	6 (5.5)	59 (53.6)	
Medication noncompliance					.004[a]
High	11 (10)	34 (31)	1 (0.9)	46 (41.8)	
Low	35 (31.8)	24 (21.9)	5 (4.5)	64 (58.2)	
Stigma					.003[a]
Absence	12 (10.9)	26 (23.7)	1 (0.9)	39 (35.5)	
Low	34 (30.9)	29 (26.3)	5 (4.5)	70 (64.7)	
High	0 (0)	3 (2.7)	0 (0)	3 (2.7)	

XSL•FO
RenderX

Personal determinant factors	Stress level			Total, n (%)	*P* value[a]
	Normal, n (%)	Mild stress, n (%)	Severe stress, n (%)		
Social support of caregiver					.26
Adequate	34 (30.9)	38 (34.5)	5 (4.5)	77 (70)	
Low	12 (10.9)	20 (18.1)	1 (0.9)	33 (30)	
Access to health information					.25
Adequate	13 (11.8)	23 (20.9)	1 (0.9)	37 (33.6)	
Low	33 (30)	35 (31.9)	5 (4.5)	73 (66.4)	

[a]Chi-square test (significance set at *P*<.05).

Regarding situational determinants**,** the results indicated that caregiving burden was significantly associated with caregiver stress (*P*<.001), where a higher burden was correlated with increased stress (Table 5) Caregivers' coping also showed a significant relationship with stress (*P*<.001), with caregivers using maladaptive coping strategies experiencing higher stress levels than those using adaptive coping mechanisms. Furthermore, medication noncompliance by the care recipient was significantly related to caregiver stress (*P*=.004).

Regarding social environmental determinants, the findings revealed that social stigma was significantly associated with caregiver stress (*P*=.003), with caregivers experiencing high stigma being more vulnerable to moderate to severe stress than those experiencing low stigma (Table 5). Social support did not show a significant relationship with caregiver stress (*P*=.26), although caregivers with limited social support reported higher stress levels than those with sufficient support (Table 5). Similarly, access to health information was not significantly associated with stress levels (*P*=.25), although caregivers with inadequate access to health-related information reported increased anxiety and uncertainty, which may have indirectly contributed to stress.

These findings underscore that the factors significantly associated with caregiver stress include age, education level, caregiving burden, coping strategies, medication compliance, caregiver knowledge, and social stigma (Table 5). In contrast, employment status, caregiving duration, social support, and access to health information were not significantly associated with caregiver stress (Table 5). These results highlight the importance of targeted interventions to reduce caregiver stress by enhancing knowledge, promoting adaptive coping strategies, and addressing social stigmas.

Measurement Model Analysis Using PLS-SEM

The PLS-SEM is an exploratory study with structural model development. When the evaluation results of the research model in the second-order factor do not sufficiently meet the levels of validity, reliability, and convergent validity in measuring the relationship between variables and their measurement dimensions, the PLS model needs to be considered. In the respective PLS model, a direct estimation was conducted between information access, social support, stigma, burden, noncompliance, coping, stress, knowledge, duration of care, occupation, education, and age in relation to increased stress. Thus, the model was revised from a second-order factor to a first-order factor.

Evaluation of the Respecified PLS Outer Model

Evaluation of the inner model or structural model is related to testing the hypothesis of the influence between previously hypothesized research variables. The first structural model evaluation consists of checking the collinearity between variables with the Inner variance inflated factor measure where the inner variance inflated factor value <5 means there is no multicollinearity. The simplified PLS model is expressed as follows. The PLS model is an approach with first-order factor measurement, in which each dimension or aspect directly influences the stress variable. The first estimation result indicates that some indicators are not valid for measuring their respective dimensions and require adjustments by adding or removing certain indicators. In terms of reliability, most variables had a composite reliability value above 0.60, except for coping, which was less reliable. Furthermore, convergent validity has not yet been achieved as the average variance extracted (AVE) value is below 0.50, necessitating further model refinement.

The initial estimation results of the PLS model specification (Figure 1) show that several indicators are invalid for measuring the dimensions of each variable. These indicators exhibited modest correlations when the same variable dimensions were measured. To obtain a successful PLS model, it is necessary to continually add and delete indicators from the model. The first estimation result in the PLS respecification model shows that there are several indicators that are not valid for measuring each dimension of the variable measurement. The indicators were not strongly correlated when measuring the same variable dimensions. Therefore, it is necessary to repeatedly enter and remove indicators from the model to obtain a valid PLS respecification model.

According to Table 6, the PLS respecification model estimation results show that, according to the composite reliability measure, the two variables have low reliability. For the coping dimension, a Cronbach α value of 0.622 was considered trustworthy; however, a composite reliability score of 0.591<0.60 is regarded as unreliable. These findings highlight the need to eliminate erroneous indicators from the PLS model in order to enhance it from the initial estimation. The AVE value of the PLS respecification model was less than 0.50, making the convergent validity results with the AVE unacceptable for the first

XSL•FO
RenderX

estimation. Consequently, either eliminating indicators with low outer loadings or selecting indicators with a high degree of correlation are required to improve the model.

Table . Reliability level of respecification (first estimation) determinant factors of stress level on caregiver of patients with schizophrenia in Pangkep Regency, Indonesia, cross-sectional study, 2024.

Dimension	Cronbach α[a]	Composite reliability (ρ_a)[b]	Composite reliability (ρ_c)[c]	Note
Access to health information	0.842	0.829	0.845	Reliable
Social support caregiver	0.622	0.677	0.639	Reliable
Stigma	0.918	0.934	0.927	Reliable
Burden caregiver	0.709	0.721	0.712	Reliable
Medication nonadherence	0.677	0.685	0.746	Reliable
Coping caregiver	0.622	0.564	0.591	Nonreliable
Knowledge	0.685	0.916	0.775	Reliable
Stress	0.715	0.757	0.789	Reliable

[a]α<0.6; nonreliable.

[b]ρ_a (composite reliability)>0.6.

[c]A measure of internal consistency in a variable in the structural equation modeling (SEM) measurement model.

In the second estimation (Figure 2), the results showed an improvement in validity, with all indicators having outer loadings greater than 0.60 (Table 7). The strongest correlating indicators were identified for each dimension: DS9 for social support, INF3 for information access, STG5 and STG7 for stigma, and STR11 for stress. Additionally, personal determinant factors, such as education, length of care, employment, and age, also had valid outer loadings. Thus, the refined PLS model demonstrated stronger and more reliable results when the variables studied were measured.

Table . Reliability level of second estimation on determinant factors of stress level on caregiver of patients with schizophrenia in Pangkep Regency, Indonesia, cross-sectional study, 2024.

Dimension	Cronbach α[a]	Composite reliability (ρ_a)[b]	Composite reliability (ρ_c)[c]	AVE[d,e]
Access to health information	0.561	0.705	0.807	0.680
Social support caregiver	0.806	0.872	0.887	0.727
Burden caregiver	0.659	0.672	0.815	0.597
Stigma	0.916	0.921	0.930	0.548
Medication nonadherence	0.775	0.820	0.897	0.813
Coping caregiver	0.737	0.790	0.849	0.655
Knowledge	0.897	0.909	0.918	0.588
Stress	0.710	0.728	0.821	0.537

[a]α<0.6; nonreliable.

[b]ρ_a (composite reliability)>0.6.

[c]ρ_c (composite reliability)>0.6.

[d]AVE: average variance extracted.

[e]AVE>0.6.

Figure 2 indicates that all indications are true when the outer loadings are greater than 0.60. DS1, DS2, and DS3 are three reliable indicators that measure social support; their outer loadings ranged from 0.696 to 0.938. Access to the information dimension is measured by two reliable indicators, INF3 and INF6, whose outer loadings range from 0.709 to 0.926. A total of 11 indicators, STG1, STG2, STG3, STG4, STG5, STG6, STG7, STG9, STG10, STG11, and STG13, are valid for assessing the stigma dimension, and their outer loadings range from 0.662 to 0.846. BN1, BN6, and BN8 were reliable indicators of the burden dimension. The outer loading ranged from 0.678 to 0.837. Three reliable indicators that measured the coping dimension were KP5, KP8, and KP9; their outer loadings ranged from 0.732 to 0.907. Among these indicators, KP8 exhibited the highest outer loading. Medication noncompliance can be measured using two reliable indicators, KTP3 and KTP10, which have outer loadings ranging from 0.872 to 0.930. Indicator KTP3 exhibited the highest outer load. A total of 8 reliable indicators with outer loadings ranging from 0.648 to 0.900 were used to measure knowledge dimension: PNG1, PNG2, PNG3, PNG5, PNG8, PNG9, PNG10, and PNG13. The PNG2 and PNG13 indicators exhibited the highest

XSL•FO
RenderX

external loading. A total of 4 reliable indicators, STR1, STR7, STR11, and STR12, with outer loadings between 0.655 and 0.855, were used to measure the stress dimension, and have identified STR11 as having the largest outer loading.

According to Table 7, the PLS respecification model estimation results show that, according to the composite reliability measure, the two variables have low reliability. For the coping dimension, a Cronbach α value of 0.622 was considered trustworthy; however, a composite reliability score of 0.591<0.60 is regarded as unreliable. These findings highlight the need to eliminate erroneous indicators from the PLS model in order to enhance it from the initial estimation. The AVE value of the PLS respecification model was less than 0.50, making the convergent validity results with the AVE unacceptable for the first estimation. Consequently, either eliminating indicators with low outer loadings or selecting indicators with a high degree of correlation is required to improve the model.

The second estimation indicated that all measurement dimensions were dependable, with Cronbach α and composite reliability values over 0.60 (reliable), indicating a satisfactory reliability level for the second estimation in the PLS model. However, information access has a reliability value of 0.807>0.60 (reliable) according to the composite reliability measure, but 0.561<0.60 (not reliable) according to Cronbach α. Furthermore, the AVE value, which indicates convergent validity, was valid (>0.50).

To evaluate the predictive power of a PLS-SEM model, it is important to ensure that the model can not only explain the relationships in the sample but also accurately predict new data. In the PLS prediction analysis, a comparison between the root mean square error and mean absolute error using a linear regression model was performed to assess the predictive performance of the model. This approach helps identify potential overfitting and ensures the predictive validity of the model.

According to Table 8 in the PLS model, all indicators of the endogenous variable stress (STR1, STR7, STR11, and STR12) exhibited lower root mean square error and mean absolute error values than those of the linear regression model, as shown in Table 8. The test results show that the PLS model proposed in this study has good predictive power. Consequently, the strong predictive power of the PLS model was validated. This study's PLS model was deemed acceptable based on an examination of the entire model using R-square, Q-square, standardized root mean squared residual, and PLS Predict. Knowledge, burden, social support, and stigma are the three aspects of this model that impact caregiver stress. Although greater load and noncompliance with medication might increase stress, increased understanding can help alleviate stress. This model suggests that caregiver stress is influenced by four factors: knowledge, burden, social support, and stigma. Increased knowledge can help reduce stress, while a greater burden and noncompliance with treatment can increase stress. With this understanding, interventions aimed at increasing caregiver knowledge and reducing caregiver burden by teaching effective coping techniques, such as relaxation, EFT, spiritual therapy, meditation, and yoga, can help caregivers manage stress.

Table . PLS[a] predict factors of stress level on caregivers of patients with schizophrenia in Pangkep Regency, Indonesia, cross-sectional study, 2024.

	Q^2predict	PLS-SEM_RMSE[b,c]	PLS-SEM_MAE[d]	LM_RMSE	LM_MAE
STR1	0.254	0.871	0.640	1.150	0.900
STR7	0.083	0.844	0.597	1.107	0.823
STR11	0.276	0.779	0.579	0.904	0.732
STR12	0.119	1.041	0.812	1.431	1.103

[a]PLS: partial least squares.

[b]SEM: structural equation modeling.

[c]RMSE: root mean square error.

[d]MAE: mean absolute error.

Discussion

Principal Results

This study found that 53% (58/110) of caregivers of individuals with schizophrenia experienced mild stress, whereas 5% (6/110) reported severe stress, indicating that a subset of caregivers faced significant challenges. The PLS-SEM analysis demonstrated that stress was significantly associated with caregiver burden, social support, stigma, and knowledge. The predictive model showed strong relevance, supporting the use of data-driven approaches for identifying the key psychosocial determinants. These findings highlight the importance of targeted interventions to mitigate caregiver stress and improve the quality of care.

These results support the transactional model of stress proposed by Lazarus and Folkman, emphasizing the interplay between personal and environmental factors in shaping stress responses. From a practical perspective, the findings underscore the need to strengthen caregiver capacity through mental health literacy, stigma reduction, and technology-assisted education and counseling. Interventions such as mobile apps, caregiver training, and support programs may help to reduce caregiver burden and improve outcomes for both caregivers and patients.

Policy makers should consider implementing structured and scalable strategies to support caregivers' well-being and reduce the risk of relapse in individuals with schizophrenia. The findings of this study provide a robust foundation for crafting targeted interventions that address both the psychological needs of caregivers and the quality of care provided to patients. Such

XSL•FO
RenderX

interventions should include mental health literacy programs, caregiver training initiatives, and antistigma campaigns. By enhancing caregivers' mental well-being and equipping them with the necessary skills and resources, such strategies can sustainably reduce caregiver burden and improve the overall quality of care for individuals with schizophrenia, ultimately contributing to a reduced risk of relapse.

Limitations

This study had several limitations that should be considered. First, the cross-sectional design limits the ability to draw causal conclusions between variables. As the data were collected at only one point in time, the dynamics of caregiver stress and its impact on long-term well-being cannot be fully explored. Therefore, longitudinal studies are recommended to observe changes in caregivers' conditions over time. Additionally, the sample was limited to female caregivers from a specific geographic region, which may have affected the generalizability of the findings. Future research should involve male caregivers and explore cultural differences in their stress responses.

Comparison With Prior Work

These findings align with the current literature indicating that caregivers often face emotional strain, high caregiving burdens, and limited social support and digital health literacy, especially in resource-limited settings [47-49].

Level Stress Caregivers

The majority of caregivers (58/110, 53%) reported experiencing mild stress, whereas 5% (6/110) reported severe stress. These findings support previous literature suggesting that caregiver stress exists on a spectrum and is influenced by factors such as caregiving duration, social support, and coping capacity [50,51]. The predominance of mild stress suggests the potential for resilience, particularly within collective cultures such as Indonesia, where family roles are highly valued [52].

Caregivers with mild stress experience ongoing psychological pressure related to their caregiving responsibilities. Caregivers typically have strong social support and functional coping mechanisms [46,53]. In contrast, caregivers who experience severe stress are more vulnerable to emotional exhaustion, social isolation, and stigma associated with mental illness. These factors align with the Transactional Model of Stress and Coping [40], which posits that stress arises when caregiving demands exceed an individual's coping ability.

These findings underscore the importance of social buffering, whereby social support mitigates the negative effects of stress. Family and community support has been shown to significantly reduce caregiver stress [46,54,55]. From a practical standpoint, these findings highlight the need for routine screening for caregiver stress as part of a holistic and preventive health care approach. Digital interventions, such as digital psychoeducation and caregiver support apps, have proven effective in improving coping skills and reducing stress [40,46,53-55]. By understanding the distribution and determinants of caregiver stress, health care providers can design more comprehensive, sustainable, community-based, and technology-driven intervention strategies tailored to local social and cultural contexts. Such strategies can enhance caregivers' quality of life and help prevent relapses in individuals with schizophrenia.

Significant Determinant Factors of Caregiver Stress

Results revealed that several personal and psychosocial variables were significantly associated with caregiver stress. Demographic characteristics, such as age (particularly those aged 45 - 55 y) and lower levels of education, were significantly related to increased stress. These findings indicate the importance of tailoring interventions based on age-related needs and cognitive capacity. Older caregivers with limited formal education may face challenges in navigating digital platforms or comprehending complex medical information, which hinders their ability to effectively manage caregiving responsibilities. Previous studies have shown that web-based psychoeducational interventions can bridge this gap. For instance, Voineskos et al [56] found that tailored digital modules improved caregiver understanding and stress management, particularly among low-educated groups. Similarly, Mueser et al [57] emphasized that user-friendly interfaces and multimedia content significantly enhance engagement among digitally inexperienced caregivers. Beyond education and age, caregiver burden was strongly associated with higher stress levels ($P<.001$), consistent with prior findings that caregiving over extended periods, lack of respite, and insufficient social support contribute to both physical and psychological fatigue [58].

Coping strategies also showed a significant association ($P<.001$), with maladaptive strategies such as avoidance and denial being linked to elevated stress levels, while adaptive coping approaches such as peer support were associated with better psychological outcomes [59,60]. Moreover, caregiver knowledge played a pivotal role ($P<.001$). Caregivers with higher levels of illness literacy were better equipped to manage care challenges, align expectations, and support adherence to treatment. As explained by Al-Awad [61], when caregivers understand the rationale behind treatment protocols, medication adherence improves, although stigma and limited accessibility can obstruct compliance. Notably, medication nonadherence ($P=.004$) was significantly associated with stress, reinforcing the need for education-based strategies. These findings align with digital health research [62], which showed that mobile coping programs enhanced caregivers' health literacy and emotional regulation, thereby reducing stress. Another critical factor was stigma ($P=.003$). As framed by Goffman's Social Stigma Theory, mental illness-related stigma can lead to social exclusion and emotional strain in caregivers. Shimazu [63] supported this finding, noting that caregivers of people with schizophrenia often internalize stigma, exacerbating emotional distress.

In contrast, social support ($P=.26$) and access to health information ($P=.25$) were not statistically significant in this study, although their relevance was supported by existing theory. According to the Social Support Theory, emotional and instrumental support buffers individuals from stress. Fonseka and Woo [64] emphasized that even when caregivers receive support, those who perceive it as insufficient remain at psychological risk [65]. Similarly, while information access did not show a statistically significant relationship with stress, its

XSL•FO
RenderX

theoretical importance remained. As proposed by the Health Information Access Theory, equitable access to health information enhances health literacy, decision-making, and caregiving confidence. Castillo et al [46] and Romm et al [65] found that disparities in access could hinder caregivers' coping mechanisms and increase anxiety. Chen et al [66] also found that uncertainty and poor access to accurate information exacerbated emotional burdens among caregivers [67].

These findings reinforce the notion that caregiver stress is shaped by a complex interplay of personal, social, and informational variables. Interventions should be designed holistically, addressing not only empirically significant variables such as education, burden, and stigma but also theoretically important elements such as social support and information equity to foster comprehensive caregiver resilience and reduce psychological strain. Thus, these findings not only align with but also extend existing theoretical frameworks by offering practical implications for designing targeted, evidence-based interventions that address both individual and systemic determinants of caregiver stress.

Structural equation modeling using PLS-SEM confirmed that caregiving burden, knowledge, social support, and stigma were significant predictors of caregiver stress. The model demonstrated good internal consistency (Cronbach $\alpha>0.70$), adequate convergent validity (AVE>0.50), and strong outer loading (>0.60). The PLS-predict analysis further supported the model's predictive relevance, outperforming benchmark linear regression models. These findings underscore the need for culturally contextualized, targeted interventions, particularly digital health solutions that address psychosocial stressors and improve caregiver literacy and resilience in mental health care.

Conclusions

This study revealed that the majority of caregivers of individuals with schizophrenia experienced mild stress; however, a substantial subset also encountered moderate to severe stress. Key psychosocial determinants included caregiving burden, stigma, insufficient knowledge, and limited social support. These findings provide important empirical evidence that emphasizes the urgency of developing targeted support strategies for caregivers.

Given the growing role of digital health in improving mental health outcomes, future research should focus on the design and evaluation of telehealth-based interventions. Specifically, randomized controlled trials and longitudinal studies are required to assess the long-term effectiveness of community-based psychoeducational programs delivered through mobile or web platforms. These interventions should be tailored to caregivers' digital and health literacy levels, particularly in resource-constrained settings, to ensure accessibility and sustained engagement.

Acknowledgments

We extend our gratitude and acknowledge the support and encouragement provided by the Directorate of Research, Technology, and Community Service; the Directorate General of Higher Education, Research, and Technology; and the Ministry of Education, Culture, Research, and Technology for this research endeavor provided under the Postgraduate Research Dissertation Grant with contract 02035/UN4.22.2/PT.01.03/2024. No portion of this manuscript was generated using generative artificial intelligence. All the content was written and edited by the authors.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

IS developed the initial concept and design of the study, including formulating the research objectives, scope, and key research questions that guided the entire investigation. HA was responsible for collecting, organizing, and processing the data used in this study, including validation, data cleaning, and ensuring the data were properly structured for subsequent analysis. SS conducted the formal analysis of the curated data, applying appropriate statistical and quantitative methods to interpret the data and extract meaningful results that support the study's conclusions. HA developed and established the methodological framework for the research, including the experimental design, selection of variables, data collection instruments, and analytical procedures used throughout the study. SR managed the overall administration of the research project, including coordinating the research team, scheduling, and ensuring that all phases of the study were executed according to the planned timeline. YA created visual representations of the data and findings, including the design of graphs, tables, diagrams, and other visual aids that facilitated the clear presentation of results. IS prepared the initial draft of the manuscript, including writing the introduction, methods, results, and preliminary discussion sections of the study. IS, HA, SS, VH, SR, and YA collectively contributed to reviewing, revising, and editing the manuscript to ensure accuracy, clarity, completeness, and compliance with publication standards. All authors reviewed the final manuscript.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
Data analysis model of partial least squares structural equation modeling (PLS-SEM).
[PDF File, 1196 KB - formative_v9i1e70648_app1.pdf]

Checklist 1
Checklist for Reporting Results of Internet E-Surveys (CHERRIES)
[DOCX File, 17 KB - formative_v9i1e70648_app2.docx]

References

1. Sołtys A, Bidzan M, Tyburski E. The moderating effects of personal resources on caregiver burden in carers of Alzheimer's patients. Front Psychiatry 2021;12:772050. [doi: 10.3389/fpsyt.2021.772050] [Medline: 34916974]
2. Peng MM, Xing J, Tang X, Wu Q, Wei D, Ran MS. Disease-related risk factors for caregiver burden among family caregivers of persons with schizophrenia: a systematic review and meta-analysis. Int J Environ Res Public Health 2022 Feb 7;19(3):1862. [doi: 10.3390/ijerph19031862] [Medline: 35162886]
3. Lestari R, Yusuf A, Setyawan FEB, Ahsan A, Hargono R. A societal adaptation model as a novel approach toward the recovery of people with schizophrenia. Healthcare Low-Resour Sett 2023;11(s1). [doi: 10.4081/hls.2023.11208]
4. Wang Y, Tian S, Yi J, Cai C, Zhu Y, Shi C. The impact of family burden on the psychological distress of informal caregivers for older adults living with dementia: the role of internal and external coping resources. Geriatr Nurs 2024;60:92-98. [doi: 10.1016/j.gerinurse.2024.08.017] [Medline: 39236371]
5. Pertiwi M. The influence of caregiver burden on the quality of life of family caregivers of schizophrenia patients with coping strategies as a moderator variable [Thesis in Indonesian]. : Universitas Airlangga; 2020.
6. Sustrami D, Suhardiningsih AS, Setyowati A, Fauk NK, Arifin H. Family caregiver burden in schizophrenia: a structural equation model of caregiver, patient, environmental, and family function factors. J Kep Padjadjaran 2024;12(3):261-269. [doi: 10.24198/jkp.v12i3.2599]
7. Jung M, Jeong Y, Park BJ. Adaptation of families of adult patients with brain tumor: partial least squares structural equation modeling. PLoS ONE 2023;18(5):e0285677. [doi: 10.1371/journal.pone.0285677] [Medline: 37167209]
8. Nurfurqoni FA, Hastuti D, Sunarti E. The influence of family function, social support, and infants' health on mothers' postnatal depression. Psikohumaniora 2024;9(1):12-28. [doi: 10.21580/pjpp.v9i1.22877]
9. Shih CY, Pai HC. Factors affecting the relationship between stress and anxiety in critically ill patients: a PLS-SEM approach. Clin Nurs Res 2023;32(4):356-367. [doi: 10.1177/10547738211062346]
10. Nguyen VHA, Tran TNA, Vu TT, et al. The interplay of psychological distress, stigma, and social support in determining quality of life among Vietnamese people living with HIV. Discover Ment Health 2025;5(1):22. [doi: 10.1007/s44192-025-00171-z]
11. Wang YX, Cai C, Zhu YX, et al. Family burden and psychological distress among Chinese caregivers of elderly people with dementia: a moderated mediation model. BMC Nurs 2024 Oct 8;23(1):723. [doi: 10.1186/s12912-024-02382-1] [Medline: 39379874]
12. Khan A, Zeb I, Zhang Y, Fazal S, Ding J. Relationship between psychological capital and mental health at higher education: role of perceived social support as a mediator. Heliyon 2024 Apr;10(8):e29472. [doi: 10.1016/j.heliyon.2024.e29472]
13. Zhang Z, Li Z, Fan Y, et al. PLS-SEM modeling in healthcare research: a review of recent studies. J Affect Disord 2024;329:1220. [doi: 10.1016/j.jad.2024.01.067]
14. François J, Audrain-Pontevia AF, Boudhraâ S. Impact of mental health apps on trust in health providers and treatment compliance: cross-sectional PLS-SEM study. J Med Internet Res 2024;26(1):e48182. [doi: 10.2196/48182]
15. Ma Y, Liang C, Gu D, Zhao S, Yang X. How social media use at work affects improvement of older people's willingness to delay retirement: a PLS-SEM approach. J Med Internet Res 2021;23(2):e18264. [doi: 10.2196/18264]
16. Zhang D, Lim J, Zhou L, Dahl AA. Breaking the data value-privacy paradox in mobile mental health systems through user-centered privacy protection: a web-based survey study. JMIR Ment Health 2021 Dec 24;8(12):e31633. [doi: 10.2196/31633] [Medline: 34951604]
17. Siregar R, Kritpracha C, Chinnawong T. Quality of life among Indonesian family caregivers caring for dependent older persons: a cross-sectional study using PLS-SEM. Belitung Nurs J 2025;11(1):33-44. [doi: 10.33546/bnj.2025.11770268]
18. Eysenbach G. Improving the quality of web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34. [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
19. Sugiyono. Metode Penelitian Kuantitatif, Kualitatif, dan R&D [Report in Indonesian]. : CV.Alfabeta; 2019.
20. National Alliance for Caregiving. Caregiving in the US 2009. AARP. 2009. URL: https://www.aarp.org/pri/topics/ltss/family-caregiving/caregiving_09/ [accessed 2025-06-15]
21. Boyd JE, Adler EP, Otilingam PG, Peters T. Internalized Stigma of Mental Illness (ISMI) scale: a multinational review. Compr Psychiatry 2014 Jan;55(1):221-231. [doi: 10.1016/j.comppsych.2013.06.005] [Medline: 24060237]
22. Carver CS, Scheier MF, Weintraub JK. COPE inventory. : APA PsycTests; 1989. [doi: 10.1037/t10027-000]

XSL•FO
RenderX

23. Cordingley L, Wearden A, Appleby L, Fisher L. The Family Response Questionnaire: a new scale to assess the responses of family members to people with chronic fatigue syndrome. J Psychosom Res 2000. [doi: 10.1016/S0022-3999(01)00193-3]
24. Zimet GD, Dahlem NW, Zimet SG, Farley GK. The Multidimensional Scale of Perceived Social Support. J Pers Assess 1988 Mar;52(1):30-41. [doi: 10.1207/s15327752jpa5201_2]
25. Zarit SH, Reever KE, Bach-Peterson J. Relatives of the impaired elderly: correlates of feelings of burden. Gerontologist 1980 Dec;20(6):649-655. [doi: 10.1093/geront/20.6.649] [Medline: 7203086]
26. Reine G, Lancon C, Simeoni MC, Duplan S, Auquier P. [Caregiver burden in relatives of persons with schizophrenia: an overview of measure instruments]. Encephale 2003;29(2):137-147. [Medline: 14567165]
27. Compton MT, Quintero L, Esterberg ML. Assessing knowledge of schizophrenia: development and psychometric properties of a brief, multiple-choice knowledge test for use across various samples. Psychiatry Res 2007 May 30;151(1-2):87-95. [doi: 10.1016/j.psychres.2006.05.019] [Medline: 17379321]
28. Thompson K, Kulkarni J, Sergejew AA. Reliability and validity of a new Medication Adherence Rating Scale (MARS) for the psychoses. Schizophr Res 2000 May 5;42(3):241-247. [doi: 10.1016/s0920-9964(99)00130-9] [Medline: 10785582]
29. Hair JF, Hult G, Tomas M, et al. Partial Least Squares Structural Equation Modeling (PLS-SEM) Using R: A Workbook: Springer Nature; 2021. URL: https://library.oapen.org/handle/20.500.12657/51463 [accessed 2025-06-20]
30. Mouzaek E, Marzouqi A, Alaali N, Salloum S, Aburayya A, Suson O. An empirical investigation of the impact of service quality dimensions on guests satisfaction: a case study of Dubai hotels. J Contemp Issues Bus Gov Apr 2021;27(3):1187. [doi: 10.47750/cibg.2021.27.03.160]
31. Salloum SA, Mhamdi C, Kurdi BA, Shaalan K. Factors affecting the adoption and meaningful use of social media: a structural equation modeling approach. Int J Inf Technol Lang Stud 2018;2(3):96-109 [FREE Full text]
32. Hair J, Hollingsworth CL, Randolph AB, Chong AYL. An updated and expanded assessment of PLS-SEM in information systems research. Ind Manag Data Syst 2017 Apr 10;117(3):442-458. [doi: 10.1108/IMDS-04-2016-0130]
33. Salloum SA, Al-Emran M, Habes M, Alghizzawi M, Ghani MA, Shaalan K. What impacts the acceptance of E-learning through social media? An empirical study. Recent Adv Technol Accept Model Theor ;2021:419-431. [doi: 10.1007/978-3-030-64987-6_24]
34. MacKenzie SB, Podsakoff PM, Jarvis CB. The problem of measurement model misspecification in behavioral and organizational research and some recommended solutions. J Appl Psychol 2005;90:710-730. [doi: 10.1037/0021-9010.90.4.710] [Medline: 16060788]
35. Goodhue DL, Lewis W, Thompson R. Does PLS have advantages for small sample size or non-normal data? MIS Q 2012;36(3):981. [doi: 10.2307/41703490]
36. Barclay D, Higgins C, Thompson R. The partial least squares (PLS) approach to casual modeling: personal computer adoption and use as an illustration. Technol Stud 1995 [FREE Full text]
37. Jansen JE, Gleeson J, Cotton S. Towards a better understanding of caregiver distress in early psychosis: a systematic review of the psychological factors involved. Clin Psychol Rev 2015 Feb;35:56-66. [doi: 10.1016/j.cpr.2014.12.002] [Medline: 25531423]
38. Aysun B, Nesrin İ. Predictors of caregiver burden in caregivers of bedridden patients. J Nurs Res 2019 Jun;27(3). [doi: 10.1097/jnr.0000000000000297]
39. Lee MMS, Yeoh EK, Wong ELY. Employers' perspectives of caregiver-friendly workplace policies for caregiver-employees caring for older adults in Hong Kong: thematic analysis. JMIR Aging 2025 Mar 31;8:e68061. [doi: 10.2196/68061] [Medline: 40163861]
40. Soni A, Kumar R, Das A. Experiences of the family caregivers of persons with schizophrenia in North Indian region: a qualitative inquiry. Arch Psychiatr Nurs 2024 Aug;51:17-24. [doi: 10.1016/j.apnu.2024.05.005] [Medline: 39034075]
41. Lazarus RS, Folkman S. Stress, Appraisal, and Coping: Springer publishing company; 1984.
42. Pearlin LI, Mullan JT, Semple SJ, Skaff MM. Caregiving and the stress process: an overview of concepts and their measures. Gerontologist 1990 Oct;30(5):583-594. [doi: 10.1093/geront/30.5.583] [Medline: 2276631]
43. Rosenstock IM, Strecher VJ, Becker MH. Social learning theory and the Health Belief Model. Health Educ Q 1988;15(2):175-183. [doi: 10.1177/109019818801500203] [Medline: 3378902]
44. Penner F, Sharp C, Marais L, Shohet C, Givon D, Boivin M. Community-based caregiver and family interventions to support the mental health of orphans and vulnerable children: review and future directions. New Dir Child Adolesc Dev 2020 May;2020(171):77-105. [doi: 10.1002/cad.20352] [Medline: 32618434]
45. Stanley S, Balakrishnan S. Family caregiving in schizophrenia: do stress, social support and resilience influence life satisfaction? A quantitative study from India. Soc Work Ment Health 2023 Jan 2;21(1):67-85. [doi: 10.1080/15332985.2022.2070051]
46. Castillo EG, Ijadi-Maghsoodi R, Shadravan S, et al. Community interventions to promote mental health and social equity. Curr Psychiatry Rep 2019 Mar 29;21(5):35. [doi: 10.1007/s11920-019-1017-0] [Medline: 30927093]
47. Zhou H, Ma F, Zhang Y, et al. Perceptions of family caregivers of patients with schizophrenia towards antipsychotics associated side-effects in china:a qualitative study. Patient Prefer Adherence 2022;16:2171-2179. [doi: 10.2147/PPA.S372487] [Medline: 36003799]

XSL•FO
RenderX

48. Glecia A, Li H. Mental health and wellbeing in family caregivers of patients with schizophrenia disorder: a literature review. Curr Psychol 2024 Mar;43(12):10914-10941. [doi: 10.1007/s12144-023-05220-w]
49. Chan KKS, Yip CCH, Tsui JKC. Self-compassion mediates the impact of family support on clinical and personal recovery among people with mental illness. Mindfulness (N Y) 2023 Mar;14(3):720-731. [doi: 10.1007/s12671-023-02088-6]
50. Lök N, Bademli K. The effect of a mindful self-compassion intervention on burden, express emotion and mental well-being in family caregivers of patients with schizophrenia: a randomized controlled trial. Community Ment Health J 2024 Jul;60(5):997-1005. [doi: 10.1007/s10597-024-01253-y] [Medline: 38393493]
51. Waller S, Reupert A, Ward B, McCormick F, Kidd S. Family-focused recovery: perspectives from individuals with a mental illness. Int J Ment Health Nurs 2019 Feb;28(1):247-255. [doi: 10.1111/inm.12528] [Medline: 30142231]
52. Caqueo-Urízar A, Urzúa A, Miranda-Castillo C. Quality of life and burden in caregivers of patients with schizophrenia: a literature review. Health Qual Life Outcomes 2021;19(1):76. [doi: 10.1186/s12955-021-01718-1]
53. Xiao LD, Wang J, He GP, Bellis A, Verbeeck J, Kyriazopoulos H. Family caregiver challenges in dementia care in Australia and China: a critical perspective. BMC Geriatr 2020;20(1):153. [doi: 10.1186/s12877-020-01534-9]
54. Liu D, Li X, Zhang Y. A web-based support program for caregivers of people with schizophrenia: development and usability evaluation. JMIR Ment Health 2020;7(9):e18795. [doi: 10.2196/18795]
55. Naslund JA, Aschbrenner KA, Araya R, et al. Digital technology for treating and preventing mental disorders in low-income and middle-income countries: a narrative review of the literature. JMIR Ment Health 2020;7(1):e11875. [doi: 10.2196/11875]
56. Voineskos D, Foussias G, Correll CU. Psychosocial interventions for the treatment of schizophrenia in the era of digital psychiatry: a narrative review. JMIR Ment Health 2022;9(1):e29683. [doi: 10.2196/29683]
57. Mueser KT, Achtyes ED, Gogate J, Mancevski B, Kim E, Starr HL. Telehealth-based psychoeducation for caregivers: the family intervention in recent-onset schizophrenia treatment study. JMIR Ment Health 2022 Apr 15;9(4):e32492. [doi: 10.2196/32492] [Medline: 35436231]
58. Buck B, Wingerson M, Whiting E, Snyder J, Monroe-DeVita M, Ben-Zeev D. User-centered development of bolster, an mhealth intervention for early psychosis caregivers: needs assessment, prototyping, and field trial. JMIR Ment Health 2023 Nov 30;10:e50522. [doi: 10.2196/50522] [Medline: 38032692]
59. Sharaev MG, Malashenkova IK, Maslennikova AV, et al. Diagnosis of schizophrenia based on the data of various modalities: biomarkers and machine learning techniques (review). Sovrem Tekhnologii Med 2022;14(5):53-75. [doi: 10.17691/stm2022.14.5.06]
60. Kochhar SS, Mishra AK, Chadda RK, Sood M, Bhargava R. Changes in psychosocial variables among caregivers of patients with schizophrenia: a short-term follow-up study. Cureus 2024 Mar;16(3):e55887. [doi: 10.7759/cureus.55887] [Medline: 38463408]
61. Al-Awad F. Perceived burden and quality of life in caregivers of patients with schizophrenia in Saudi Arabia's eastern province: a cross-sectional study. Clin Pract Epidemiol Ment Health 2024;20:e17450179314013. [doi: 10.2174/0117450179314013240417105321] [Medline: 39355198]
62. Sari A, Duman Z. Family resilience experiences of individuals with schizophrenia-family caregivers dyads: qualitative study. J Psychiatr Ment Health Nurs 2025 Feb;32(1):125-137. [doi: 10.1111/jpm.13091] [Medline: 39105323]
63. Shimazu A, Totsuka Y, Nagasawa S. Digital intervention to improve family caregivers' coping skills: a randomized controlled trial. JMIR Ment Health 2021;8(6):e23648. [doi: 10.2196/63743]
64. Fonseka LN, Woo BKP. Wearables in schizophrenia: update on current and future clinical applications. JMIR Mhealth Uhealth 2022 Apr 7;10(4):e35600. [doi: 10.2196/35600] [Medline: 35389361]
65. Romm KL, Nilsen L, Gjermundsen K, et al. Remote care for caregivers of people with psychosis: mixed methods pilot study. JMIR Ment Health 2020;7(7):e19497. [doi: 10.2196/19497]
66. Chen CR, Lee YW, Chen TT, et al. The effects of mindful exercise on cognition in patients with schizophrenia: a systematic review and meta-analysis of randomized controlled trials. Schizophr Res 2024 Sep;271:237-245. [doi: 10.1016/j.schres.2024.07.042] [Medline: 39059247]
67. Rybakowski J. A half-century of participant observation in psychiatry. Part I. Schizophrenia. Psychiatr Pol 2020 Jun 30;54(3):405-419. [doi: 10.12740/PP/119379] [Medline: 33038877]

Abbreviations

AVE: average variance extracted
CHERRIES: Checklist for Reporting Results of Internet E-Surveys
PLS: partial least squares
SEM: structural equation modeling

XSL•FO
RenderX

Edited by A Mavragani; submitted 28.12.24; peer-reviewed by J Ramon-Cardona, ML Duc; revised version received 12.05.25; accepted 12.05.25; published 03.07.25.

Please cite as:
Syarif I, Amqam H, Syamsuddin S, Hadju V, Russeng S, Amir Y
Determinant Factors of Stress in Caregivers of Patients With Schizophrenia: Cross-Sectional Study
JMIR Form Res 2025;9:e70648
URL: https://formative.jmir.org/2025/1/e70648
doi:10.2196/70648

© Isymiarni Syarif, Hasnawati Amqam, Saidah Syamsuddin, Veni Hadju, Syamsiar Russeng, Yusran Amir. Originally published in JMIR Formative Research (https://formative.jmir.org), 3.7.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Association of Self-Rated Health in Pregnancy With Maternal Childhood Experiences, Socioeconomic Status, Parity, and Choice of Antenatal Care Providers: Cross-Sectional Study

Bjarne Austad, PhD; Gunnhild Åberge Vie, PhD; Mari Hegnes Hansen, MD; Hanna Sandbakken Mørkved, MD; Linn Okkenhaug Getz, PhD; Bente Prytz Mjølstad, PhD

General Practice Research Unit, Department of Public Health and Nursing, Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology, P.O. Box 8905, Trondheim, Norway

Corresponding Author:
Bjarne Austad, PhD
General Practice Research Unit, Department of Public Health and Nursing, Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology, P.O. Box 8905, Trondheim, Norway

Abstract

Background: During pregnancy, self-rated health (SRH) and self-rated mental health (SRMH) are key indicators of health status and predictors of future health care needs. The relationship between pregnant women's health perceptions and their choice of antenatal care providers, midwives, or general practitioners (GPs) is not known. Factors like childhood experiences and socioeconomic status are important determinants of health throughout life. Understanding these health determinants can help health care providers better address the diverse needs of pregnant women.

Objective: This study aims to assess how SRH and SRMH during pregnancy are associated with maternal childhood experiences, socioeconomic status, parity, and antenatal care provided by midwives or GPs.

Methods: An anonymous, web-based cross-sectional survey was conducted from January to March 2022 among pregnant women in Norway, distributed via Facebook and Instagram. The survey included questions on SRH, SRMH, socioeconomic status, childhood perceptions, and antenatal program participation. Pearson's chi-squared test and logistic regression models were used to explore associations and estimate odds ratios for good SRH and SRMH.

Results: Among 1402 participants, 94.7% (1328/1402) reported good or very good health before pregnancy, dropping to 67.8% (950/1402) during pregnancy ($P<.001$). Reporting your childhood as good was associated with better SRH compared with those who reported average or difficult childhood (70.2% [755/1076] vs 64% [114/178] vs 53.2% [74/139]; $P<.001$). This corresponds to 48% lower odds of good SRH for those reporting a difficult childhood compared to those reporting a good childhood (OR 0.52, 95% CI 0.36 - 0.76). Financial security and higher education were associated with better SRH (both $P<.001$). First-time mothers reported better SRH than those with previous births (73.9% [533/722] vs 61.4% [417/680]; $P<.001$). For SRMH, 89.9% (1260/1402) reported good or very good SRMH before pregnancy, decreasing to 73.1% (1024/1401) during pregnancy ($P<.001$). Women who reported a good childhood, financial security, higher education, and first-time mothers reported better SRMH during pregnancy ($P<.001$ for all). Nearly all women participated in the antenatal program, regardless of their subjective health, and most expressed satisfaction. Among participants, 55.6% (753/1354) received shared antenatal care, 38.6% (520/1354) were seen only by midwives, and 6% (81/1354) only by GPs. The proportion of women receiving antenatal care solely from a midwife decreased with declining SRH, from 42.6% (78/183) among those with very good SRH to 27.3% (15/55) among those with poor SRH.

Conclusions: A difficult maternal childhood, low socioeconomic status, and having given birth before were associated with poorer SRH and SRMH during pregnancy. Both midwives and GPs played vital roles in providing antenatal care, though few women received antenatal care exclusively from GPs. The likelihood of physician involvement in care increased slightly with worsening health.

(JMIR Form Res 2025;9:e68811) doi:10.2196/68811

KEYWORDS

pregnancy; antenatal care; self-rated health; self-rated mental health; primary care

XSL•FO
RenderX

Introduction

Throughout pregnancy, women undergo significant physical and psychological changes. Overall health can decline due to common conditions such as nausea, fatigue, and pelvic pain, as well as more serious complications like anemia, gestational diabetes, or preeclampsia. Ensuring maternal health and well-being is a multifaceted challenge that places high demands on health care services and antenatal care [1-3].

Self-rated health (SRH) is a reliable indicator of an individual's overall health status [4,5] and serves as a strong predictor of future health care needs and mortality [6,7]. In Norway, 9% of individuals assess their health as poor or very poor, consistent with averages observed in other countries from the Organisation for Economic Co-operation and Development (OECD) [8]. Among various factors influencing SRH in adulthood, adverse childhood experiences play a significant role [9-11]. Recent research indicates that maternal adverse childhood experiences can also impact health during pregnancy and contribute to preterm birth [12-14]. In addition, adverse childhood experiences are associated with socioeconomic status [15], which in turn is strongly associated with quality of life during pregnancy [16,17]. A Swedish study found an association between poor SRH during pregnancy and adverse outcomes such as premature birth and small-for-gestational-age newborns [18]. Furthermore, a Brazilian study indicate that primiparous women reported better health compared to multiparous, while another Swedish study noted that women's physical and emotional SRH is negatively affected by pregnancy but positively influenced by childbirth [19,20]. These findings underscore the importance of assessing pregnant women's health perceptions to effectively tailor antenatal care.

Poor SRH in pregnancy is associated with elevated rates of depression, anxiety, and stress [21,22]. Prenatal mental health is particularly crucial as it can significantly affect both the mother's and the baby's outcomes [23,24]. Mental health issues are prevalent in pregnancy, with an American study estimating that approximately 12% of women experience a major depressive episode during pregnancy [25,26]. In Norway, over 9% of women report experiencing a postpartum depression [27]. While some studies suggest that mental health may remain stable or even improve during pregnancy [16], women with low social support are at an increased risk of developing mental health challenges [28]. Therefore, health care professionals are encouraged to address psychosocial stressors, enhance social support, and evaluate self-rated mental health (SRMH) within antenatal care [24,29].

Antenatal care serves a preventive health service designed to ensure safe pregnancies and childbirth while effectively identifying and managing health risks [30]. In Norway, a comprehensive free antenatal program is available, which includes 9 routine check-ups and early ultrasound examinations during weeks 11 - 13 as well as standard ultrasounds at weeks 17 - 19 (Textbox 1). National guidelines recommend addressing mental health, although SRH is not specifically mentioned [31]. Norway has one of the lowest infant mortality rates globally and a low rate of maternal mortality, indicating a healthy population and high-quality antenatal care [32-34].

The Regular GP Scheme in Norway ensures that all residents have access to a personal general practitioner (GP), who is expected to provide antenatal care. In addition, midwives in municipalities offer antenatal services. A strong emphasis is placed on the preferences of pregnant individuals, allowing them to choose between midwife- or GP-led check-ups, or a combination of both [31]. However, the limited availability of GPs and midwives may restrict these choices. Notably, pregnant women in Norway have historically attended more antenatal visits than the national guidelines recommend (12 visits instead of the advised 8 - 9) [35,36], and over 60% report taking sick leave during pregnancy [37]. Unlike physicians, midwives in Norway are not authorized to certify sick leaves but can issue "maternity allowance" if a mother's work poses risks to the fetus [38].

Textbox 1. Characteristics of current Norwegian antenatal care as described in the national guidelines.

- All pregnant women living in Norway are entitled to free antenatal care.
- The routine program is standardized for all women, regardless of previous childbirth experience, and consists of 9 checkups.
- The first visit is recommended between weeks 6-8, as early as possible, to address lifestyle factors.
- The routine antenatal care program includes ultrasound scans during week 11-13 (gradually implemented since 2022) and between weeks 17-19 (established in the early 1990s).
- Antenatal care is delivered in primary care settings by midwives and regular general practitioners (GPs). Most women in Norway have a designated regular GP and can choose to receive antenatal consultations from their GP, a midwife, or both.
- For high-risk or complicated pregnancies, referrals to specialized health care (gynecologists) are made, although some follow-up care still occurs in primary care settings.

Despite the importance of antenatal care as a preventive health service, details on antenatal care across different health care providers are not readily identified in Norwegian national registries. Updated data on pregnant women's perceived health and how this impacts their choice of antenatal care providers, midwives, or GPs can aid in the organization of services. Identifying social risk factors associated with poor perceived health in pregnancy can help target preventive efforts, thus mitigating a potential intergenerational impact.

The aim of this study was to examine how SRH and SRMH in pregnancy are associated with (1) maternal childhood experiences, current socioeconomic status, and parity, and (2) antenatal care provided by midwives or GPs.

XSL•FO
RenderX

Methods

Study Design and Setting

We conducted a web-based cross-sectional survey among Norwegian pregnant women aged 18 years and older between January and March 2022. The survey was distributed on 6 Facebook (Meta Platforms) groups for pregnant women, organized by the due date, and 5 Instagram (Meta Platforms) accounts focused on pregnancy and childbirth. These groups and accounts had a combined total of 109,000 followers although many women were likely active on multiple platforms. For comparison, 51,292 women gave birth in Norway in 2022.

To promote the survey, we used an introduction video and postings. After publishing the survey on a new platform, all responses were received within 4 days. Data collection was closed after 2 - 4 weeks for the different groups as we received no further responses. Of the 1402 women who answered the survey; 868 responses came from Facebook and 534 from Instagram. All participants consented to participate by answering the questionnaire, which included an information letter about the study. The study was anonymous, and the women were asked to respond only once. We followed the Checklist for Reporting Results of Internet E-Surveys (CHERRIES), a quality checklist for internet-based surveys [39].

Questionnaire

In the absence of validated questionnaires that fully met the study's broader objectives, we developed a new one. We incorporated validated questions from previous surveys, including questions from the Cambridge Worry Scale and questions related to demographic variables, SRH and childhood experiences from the HUNT study [40,41]. The question "worries about money problems" in the Cambridge Worry Scale seemed somewhat unfamiliar within the context of the Norwegian welfare system. We chose to ask about capacity to manage an unexpected expense as this is easy to understand and has previously been used in Norwegian surveys [42]. In addition, the capacity to manage unexpected expenses of varying amounts has been defined as financial security; we chose NOK 10,000 (approximately US $1000) [43,44]. A total of 10 pregnant women were invited to pilot the survey for assessing completion time and question clarity, of whom 6 responded, suggesting minor linguistic adjustments. The questionnaire was administered by using a data collection tool (Nettskjema.no developed by the University of Oslo), which is user-friendly and compatible with various digital devices. The questionnaire contained up to 57 questions depending on the respondent's answers, covering five different topics: (1) demographic variables, including sick leave and financial security; (2) assessment of SRH and SRMH, both during the current pregnancy and before pregnancy; (3) antenatal care, including visits to midwives, GPs, gynecologists, ultrasounds, and satisfaction with the controls; (4) sources where the women seek advice and information during pregnancy; (5) thoughts about the time after birth and the parental role, including assessment of own childhood. In total, 12 questions were open-ended with free text boxes, while the rest were closed-ended items. The complete questionnaire is translated into English and attached in Multimedia Appendix 1. This paper examines the parts related to health and attendance in the antenatal care program. Other parts of the project are published separately [45].

Study Variables

The questions concerning SRH were "Before you became pregnant, how would you characterize your own health?" and "How is your health now?" For SRMH, we asked, "Before you became pregnant, how were you feeling emotionally/mentally?" and "How are you feeling emotionally/mentally now during the pregnancy?" The answer options were very good, good, not so good, or poor in line with other studies [18,19], and were dichotomized for the logistic regression [46]. Participants were asked if they had been or were still on sick leave (used in cases of maternal illness) or maternity allowance (used when mother's work could be harmful to the fetus). The questions regarding the socioeconomic status included the level of completed education and finances. We did not inquire about income. Instead, participants were asked if they could manage an unforeseen expense of NOK 10,000 (approximately 1000 USD), with response options of yes, uncertain, or no. We merged uncertain and no and defined that as "financial insecurity," and yes as "financial security." Furthermore, they were asked a single-item question to describe their childhood as either very good, good, average, difficult, or very difficult. We merged very good and good into good and difficult and very difficult into difficult. Antenatal visits were defined as the regular controls according to the Norwegian antenatal program. The women were asked if they had attended all of them, most of them, not attended, or not relevant (not been to the first control yet) with possible free-text answers for reasons for not attending. In addition, they were asked about the profession of the health care provider (GP, midwife, gynecologist, and others).

Statistical Analysis

Data were analyzed using SPSS Statistics. The Pearson chi-squared test was used to examine associations between SRH and SRMH, and childhood experiences, financial security, parity, and attendance to antenatal care providers (GP or midwife). We further estimated the odds ratio (OR) of good SRH and SRMH using logistic regression models, and present crude associations as well as associations adjusted for age, education and trimester. While χ^2 tests allowed comparison of all four levels of SRH and SRMH, logistic regression allowed adjustment for potential confounders after dichotomizing perceived health. Most respondents answered all questions; data on trimester was missing for 2 participants, financial security for 2 and childhood experiences for 9 participants. All participants reported SRH and all but 1 SRMH. We performed complete case analyses and excluded those with missing responses from each analysis. Free text responses were analyzed for content and categorized.

Ethical Considerations

A remit assessment was submitted to the Regional Committee for Medical Research Ethics in Central Norway (REK), which determined that formal ethics approval was not required under national regulations (reference 267956). The study was also

XSL•FO
RenderX

approved by the Norwegian Centre for Research Data (NSD; reference 134424), which confirmed that no direct or indirect personal identifiers were processed. All procedures involving human participants were conducted in accordance with relevant national and institutional guidelines and regulations. An information letter, empathetically worded and based on a template provided by NSD, was distributed to all participants. Informed consent was obtained from all participants prior to data collection. No compensation was provided for participation.

Results

Table 1 shows the characteristics of the study population of 1402 participants. In Multimedia Appendix 2, we have compared these with the national birth statistics for Norway in the year of the survey, 2022.

XSL•FO
RenderX

Table . Demographic characteristics of the study population of pregnant women in Norway (N=1393 - 1402).

Characteristic	n (%)
Age (years)	
<25	119 (8.5)
25 - 37	1226 (87.4)
>37	57 (4.1)
Parity	
First-time mothers	722 (51.5)
Previous birth	680 (48.5)
Current trimester	
First trimester	99 (7.1)
Second trimester	776 (55.4)
Third trimester	525 (37.5)
Education	
Lower secondary school	28 (2)
Upper secondary school	315 (22.5)
College or university (less than 4 years)	495 (35.3)
College or university (4 years or more)	564 (40.2)
Employment status	
Employed	1210 (86.3)
Student or apprentice	123 (8.8)
Unemployed (looking for a job)	12 (0.9)
On welfare benefits, unable to work	44 (3.1)
Housewife with housework or caring responsibilities	13 (0.9)
Sick leave during pregnancy	
Yes	820 (58.6)
No	580 (41.4)
Maternity allowance[a]	
Yes	197 (14.2)
No	1195 (85.8)
Ability to handle unforeseeable expenses of NOK 10,000 (US $1000)	
Yes	1166 (83.3)
No or do not know	234 (16.7)
Childhood	
Good or very good	1076 (76.7)
Average	178 (12.7)
Difficult or very difficult	139 (9.9)
Have you attended all your antenatal care visits?	
Yes, all	1323 (94.4)
Yes, most of them	35 (2.5)
No	3 (0.2)
Not relevant or have not been to the first control yet	40 (2.9)

XSL•FO
RenderX

Characteristic	n (%)
Who has performed your antenatal care visits thus far?[b]	
Only GP[c]	81 (6)
Only midwife	520 (38.4)
GP and midwife	753 (55.6)
Women in third trimester: Who has performed your antenatal care visits thus far?	
Only GP	15 (2.9)
Only midwife	240 (45.8)
GP and midwife	269 (51.3)
Neither midwife nor GP	1 (0.2)
Have you had an antenatal visit by a gynecologist in addition to GP or midwife?	
Yes	177 (12.6)
No	1225 (87.4)
Overall, how satisfied are you with your antenatal care visits?	
Large or very large extent	1079 (77.1)
Some or small extent	273 (19.5)
Not at all	9 (0.6)
Not relevant or have not been to the first control yet	39 (2.8)
Have you attended an early ultrasound (before week 17)?	
Yes	1247 (89)
No	154 (11)
Have you attended routine ultrasound (weeks 17 - 19)?[d]	
Yes	521 (99.4)
No	3 (0.6)

[a]Maternity allowance is an economic compensation to healthy pregnant women who are unable to work because their work may be harmful for the fetus.

[b]Participants with no controls by GP or midwife thus far in pregnancy were excluded (n=48).

[c]GP: general practitioner.

[d]Based on women in 3rd trimester.

Overall Self-Rated Health Before and During Pregnancy

Before pregnancy, 94.7% (1328/1402) of participants reported their health as either very good (35.2%, 494/1402)) or good (59.5%, 834/1402). During pregnancy, this combined percentage dropped to 67.8% (950/1402) (very good 13.7% (192/1402) and good 54.1% (758/1402). Conversely, the proportion of 5.3% (74/1402) reporting their health as either not so good (4.9%, 68/1402) or poor (0.4%, 6/1402) before pregnancy increased to 32.2% (452/1402) (not so good 28.2% (396/1402) and poor 4% (56/1402) during pregnancy. This indicates a significant decline in SRH (χ^2_9=564.7; P<.001).

In Figure 1, we present the distribution of overall SRH during pregnancy, categorized by childhood experiences, socioeconomic factors (education and financial status), and parity. In Multimedia Appendix 3, we present crude and adjusted OR for good versus poor SRH. Participants who reported a good childhood had better SRH compared to those with average or difficult childhoods (70.2% [755/1076] vs 64% [114/178] vs 53.2% [74/139]; χ^2_6=34.7; P<.001). This corresponds to 52% lower odds of good SRH for those with a difficult childhood (OR 0.48, 95% CI 0.34 - 0.69), and the difference was robust to adjustment for age, education, and trimester (adjusted OR 0.52, 95% CI 0.36 - 0.76; see Multimedia Appendix 3).

XSL•FO
RenderX

Women with financial security reported better SRH than those with financial insecurity (70.2% [819/1166] vs 55.6% [130/234]; *P*<.001), corresponding to 36% lower odds of good SRH for women with financial insecurity in the adjusted model (OR 0.64, 95% CI 0.47 - 0.87). Higher education was also associated with better SRH (χ^2_9=49.3; *P*<.001), with 76% lower odds for good SRH for those with only lower secondary school compared to 4 years or more of college or university (adjusted OR 0.33, 95% CI 0.16 - 0.72). First-time mothers reported better SRH than mothers with previous births (73.9% [533/722] vs 61.4% [417/680]; *P*<.001), with 47% lower odds of good SRH for women with previous births (adjusted OR 0.53, 95% CI 0.41 - 0.67).

XSL•FO
RenderX

Figure 1. Overall self-rated health during pregnancy distributed by reported childhood, completed education, financial security, and parity. Financial security: the ability to handle unforeseeable expenses of NOK 10,000 (US $1000); n=1400 - 1402.

XSL•FO
RenderX

Self-Rated Mental Health Before and During Pregnancy

Before pregnancy, 89.9% (1260/1402) of participants reported their SRMH as very good (33.8%, 474/1402) or good (56.1%, 786/1402). During pregnancy, this combined percentage dropped to 73.1% (1024/1401; very good 17.1% [239/1401] and good 56% [785/1401]). Conversely, while 10.2% (142/1402) reporting their SRMH as either not so good (9.3%, 130/1402) or poor (0.9%, 12/1402) before pregnancy, this increased to 26.9% (377/1401; not so good 23.1%, 323/1401) and poor (3.9%, 54/1401) during pregnancy. This indicates a significant decline in SRMH (χ^2_9=365; *P*<.001).

Figure 2 shows the distribution of SRMH during pregnancy by childhood experiences, socioeconomic factors (education and financial status), and parity. Multimedia Appendix 3 shows crude and adjusted ORs for good SRMH. Participants who described their childhood as good reported better SRMH compared to those with average or difficult childhoods (76.7% [825/1075] vs 65.7% [117/178] vs 53.3% [65/139]; χ^2_6=49.7; *P*<.001). This corresponds to 65% lower odds of good SRMH for those with a difficult childhood (OR 0.35, 95% CI 0.24 - 0.56), and the difference was robust to adjustment for age, education, and trimester (adjusted OR 0.37, 95% CI 0.25 - 0.54; see Multimedia Appendix 4).

The proportion of women reporting good or very good SRMH increased with longer education, except for the few (n=28) with only lower secondary school (78% [440/564] vs 73.1% [361/494] vs 64.4% [203/315] vs 71.4% [20/28]; χ^2_9=30.2; *P*<.001). The odds of good SRMH were 45% lower for those with upper secondary school compared to those with more than 4 years of college or university (adjusted OR 0.55, 95% CI 0.40 - 0.76). Women with financial security reported better SRMH than those with financial insecurity (75.9% [884/1402] vs 59.8% [140/234]; *P*<.001), corresponding to 46% lower odds for good SRMH for those with financial insecurity (adjusted OR 0.54, 95% CI 0.39 - 0.74). First-time mothers reported better SRMH than mothers with previous births (77.4% [558/721] vs 68.5% [466/680]; *P*<.001), with 43% lower odds for good SRMH for women with previous births (adjusted OR 0.57, 95% CI 0.45 - 0.73).

XSL•FO
RenderX

Figure 2. Self-rated mental health during pregnancy distributed by reported childhood, completed education, financial security, and parity. Financial security: the ability to handle unforeseeable expenses of NOK 10,000 (US $1000); n=1400 - 1402.

XSL•FO
RenderX

Sick Leave and Maternity Allowance

In total, 58.6% (820/1400) reported that they were or had been on sick leave so far during pregnancy (see Table 1), cumulating to 69.8% (366/525) among women in the third trimester (not shown in the table). Sick leave was strongly associated with SRH: 29.8% (57/191) of those reporting very good health, 53% (401/757) with good health, 80.6% (319/396) with not so good health, and 76.8% (43/56) with poor health were or had been on sick leave (*P*<.001). A similar pattern was observed for SRMH: 44.4% (106/239) of those with very good SRMH, 56.4% (442/784) with good SRMH, 70.5% (227/322) with not so good SRMH, and 83.3% (45/54) with poor SRMH were or had been on sick leave (*P*<.001). The proportion receiving maternity allowance remained stable at approximately 14% (197/1392; see Table 1) and was not affected by SRH (*P*=.99) or SRMH (*P*=.68).

Attendance in the Antenatal Program and Satisfaction

Since nearly all attended the antenatal program (see Table 1), there was no numerical basis for analyzing whether the attenders had any different SRH than the nonattenders. 24 participants (1.7%) responded to an open-ended question about reasons for not attending all antenatal checkups thus far. The reasons for missing appointments were postponing due to illness (either their own or the health care personnel's, such as COVID-19; n=9), being early in pregnancy and not yet needing a check-up (n=4), being advised by their health care provider to wait or experiencing a wait time for the first appointment (n=4), delaying due to vacation or other reasons (n=3), forgetting the appointment (n=2), or not needing a check-up because they received treatment at the hospital (n=2).

In total, 55.6% (753/1354) received shared care from both their GP and midwife, 38.4% (520/1354) had only been to a midwife and 6% (81/1354) to only their GP. The 13 % (177/1357) that had seen a gynecologist during pregnancy were mostly also seen by either a GP or midwife, and only 0.2% (3/1357) reported having solely been to a gynecologist (not shown in the table). We did a separately analyze for women in the 3rd trimester as they probably have been to most antenatal visits and found an even lower proportion women only seeing their GP (2.9%, 15/525) and only 0.2% (1/525) responded to neither have seen a GP nor midwife.

On a 5-point Likert scale, most participants were satisfied to a large or very large extent with the antenatal care visits (79.3%, 1079/1361). While the satisfaction levels were slightly lower for those who received antenatal visits solely from their GPs compared to those who had seemed midwife only or shared care, this difference was not statistically significant (70.4% [57/81] vs 81.7% [424/519] vs 78.9% [593/752]; *P*=.05). In addition, there were no notable differences in the proportion of participants reporting being satisfied to a little extent or not at all (3.7% [3/81] vs 3.5% [18/519] vs 3.6% [27/752]).

Self-Rated Health and Attendance

We analyzed the distribution of antenatal care visits based on SRH (see Figure 3) and SRMH (see Figure 4). The proportion of women receiving antenatal care solely from a midwife decreased as SRH declined, from 42.6% (78/183) among those with very good SRH to 27.3% (15/55) among those with poor SRH. The proportion receiving combined care increased correspondingly with poorer SRH, while the proportion receiving antenatal care solely from a GP remained low. These differences were not statistically significant across all 3 categories of attendance (χ^2_6=9.1, *P*=.17). However, we additionally compared visits to GPs, either solely or as shared care, to visits solely to midwives, with results indicating a statistically significant difference (χ^2_3=8.1; *P*=.04). We also examined antenatal visits by gynecologists, whether solely or in combination with primary health care, and found an increase from 13% among women with very good SRH to 23.2% with poor SRH (*P*<.001; not shown in Figure 3).

The proportion of women receiving antenatal care solely from a midwife remained stable for those with very good (39.3%, 90/129), good (38.8%, 292/412), and not so good (39%, 124/175) SRMH, but dropped for those reporting poor SRMH (24.5%, 13/53). However, this difference was not statistically significant, whether assessed across all 3 categories of attendance (χ^2_6=6.2; *P*=.40) or comparing any visits to GPs to visits solely to midwives (χ^2_3=5.2; *P*=.16). For gynecologists, there were no statistically significant differences by different SRMHs (*P*=.14; not shown in Figure 4).

XSL•FO
RenderX

Figure 3. Self-rated health during pregnancy for participants having antenatal consultation with only a general practitioner (GP; n=81), only a midwife (n=520) or combined GP and midwife (n=753). Participants with no controls by GP or midwife thus far in pregnancy were excluded (n=48).

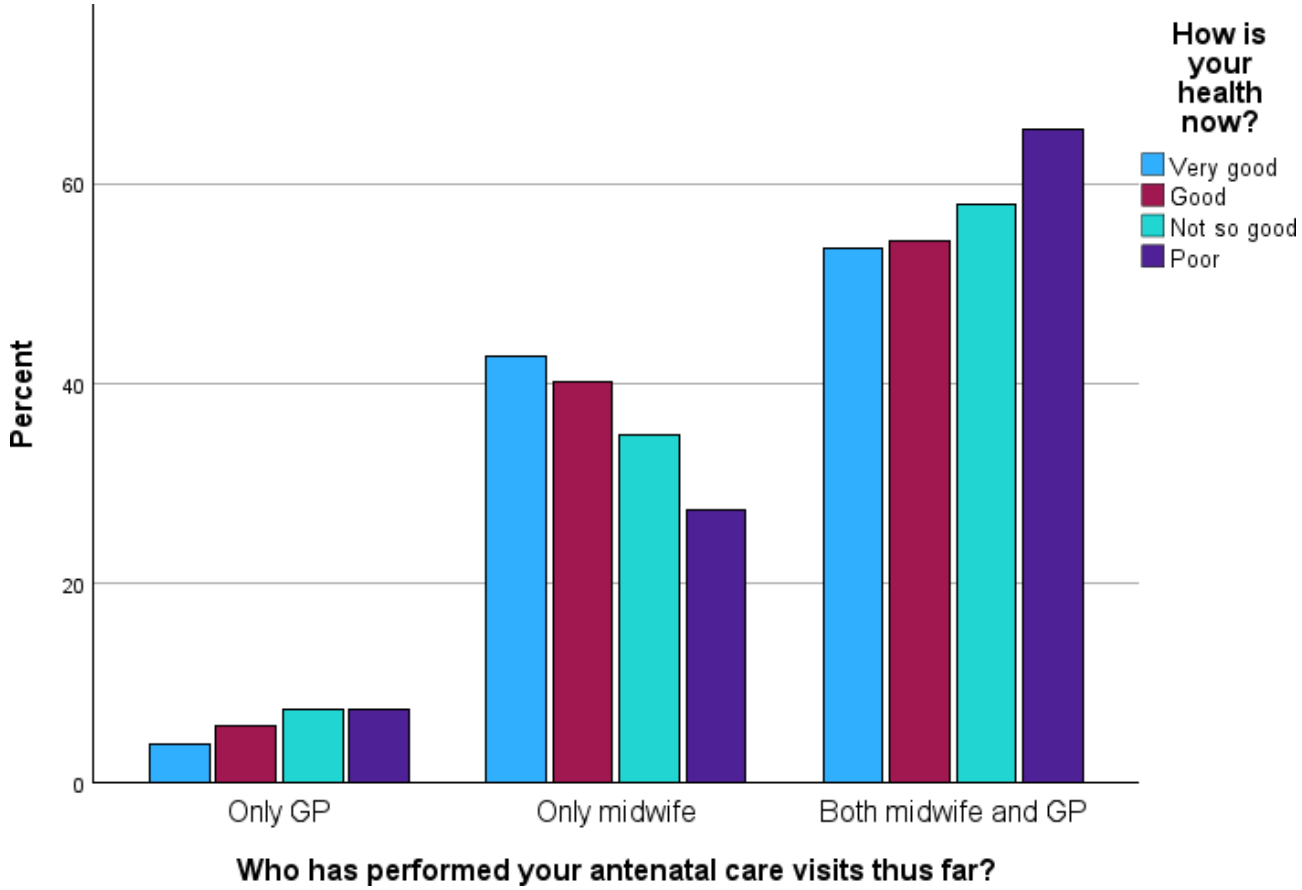

XSL•FO
RenderX

Figure 4. Self-rated mental health during pregnancy for participants having antenatal consultation with only a general practitioner (GP; n=81), only a midwife (n=520) or combined GP and midwife (n=753). Participants with no controls by GP or midwife thus far in pregnancy were excluded (n=48).

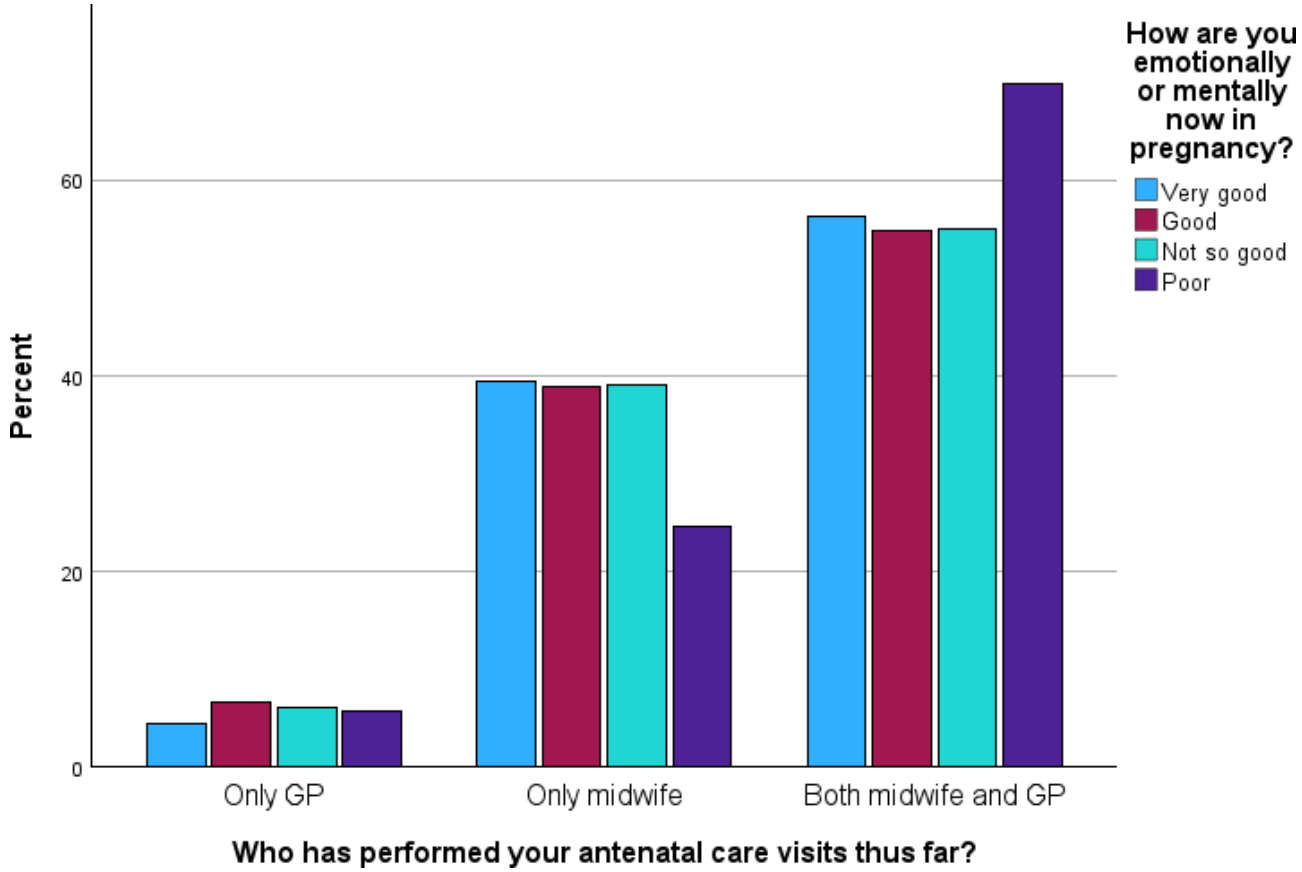

Discussion

Principal Findings

We found that both SRH and SRMH were reported to be worse during pregnancy compared to before pregnancy. Women who had experienced difficult childhoods, had low socioeconomic status, or those who have given birth before reported poorer SRH and SRMH, even after adjusting for age, trimester, and education. Nearly all pregnant women participated in the antenatal program regardless of their health status. Both midwives and GPs played crucial roles in providing antenatal care, shared care being the most common type of care. Few women received antenatal visits exclusively from a GP. Notably, the proportion of women receiving antenatal care from a physician (either GP or gynecologist) increased slightly with declining SRH.

Findings in Relation to Comparable Studies

Self-Rated Health Before and During Pregnancy

While it is well established that health can temporarily deteriorate during pregnancy [47,48], the extent of worsening health reported by the women in our study is striking. The proportion of women indicating their SRH as not so good or poor, raised from 5.3% (74/1402) to 32.2% (452/1402), and their SRMH from 10.2% (142/1402) to 26.9% (377/1402). Direct comparisons with other studies are complicated by variations in exact phrasing and participant selection, in addition to sociocultural differences across countries [49]. A Swedish register study noted an increase in poor SRH from approximately 3% before pregnancy to 9% during pregnancy [50], while an American study found that 16% of women indicated fair or poor SRH during pregnancy [5]. In contrast, another Swedish study reported similar findings to ours, with the proportion of women rating their SRH as poor increasing from 20% to 37% between mid- and late pregnancy [19]. In the Norwegian context, our survey question—“Before you became pregnant, how would you characterize your own health?”—yielded a proportion of poor SRH that aligns with or is slightly lower than national averages regardless of pregnancy status [8]. We collected data shortly after the lifting of COVID-19 restrictions in Norway, and it remains unclear whether some participants’ reporting of their health before or during pregnancy were still negatively impacted by the pandemic [51,52].

Adverse childhood experiences are associated with poor health in adult life [9-11,53] and epigenetic aging later in life, including during pregnancy [12]. Furthermore, these experiences are associated with an increased risk of preterm birth and hypertensive disorders in pregnancy [14,54]. A recent study highlighted links between such experiences and risk factors among pregnant women, including smoking, depression, and being exposed to physical abuse, all of which can adversely impact maternal, pregnancy and infant health [55]. A Danish study found that adverse childhood experiences was associated with antenatal depression and anxiety, however, the association was not as strong as that of recent stressful life events [56]. Our findings, which indicate 52% lower odds of good SRH and 65% lower odds of good SRMH if they perceived their childhood as

XSL•FO
RenderX

difficult, align well with these studies. Although we did not explore various categories of adverse childhood experiences, the single question regarding women's assessment of their childhood is validated and represents an empirically supported method for evaluating such experiences [57]. This underscores the importance of addressing maternal childhood experiences during antenatal visits in relation to SRH and SRMH [18].

Improved socioeconomic status is well-documented to correlate with better health outcomes, influencing both pregnancy and the likelihood of complications, possibly linked to increased stress [56,58,59]. A global analysis emphasized how social determinants can affect maternal health in a complex interplay between superdeterminants, individual factors and health care systems [3]. Our findings indicate that higher education and financial security is associated with better SRH and SRMH. This is particularly noteworthy given that Norway has one of the world's most comprehensive welfare systems and robust birth rights, including up to 12 months of paid parental leave and free antenatal care. In another study using the same dataset, we found that 11.8% (165/1402) of the participants were greatly concerned about their financial situation, and nearly 3 of 4 worried to some extent [45]. As a comparison, in a study of pregnant and postpartum women in the United States, 59% reported health care unaffordability and 54% general financial stress [60].

Sick Leave and Maternity Allowance

International comparisons of sick leave prevalence are challenging due to differences in welfare schemes and labor markets; however, Norway reports some of the highest rates in Europe [37]. Previous studies indicate that between 60% and 75% of pregnant women in Norway report sick leave during pregnancy [37,61]. This aligns with our findings, where nearly 60% (820/1,402) reported at least one instance of sick leave, increasing to 70% (366/525) among women in the third trimester. The reasons for frequent sick leave are complex [62]. Although an appropriate level of sick leave is difficult to establish, the observed increase in sick leave among those with poorer SRH or SMRH aligns well with the intention of its use. In addition, the finding that maternity allowance [38]—granted when a mother's work may be harmful to the fetus—was not influenced by the women's health further supports the rationale behind its use.

Antenatal Care Provision by Midwife, GP, or Both?

While other studies have found several barriers to antenatal care, our findings indicate that nearly all pregnant women participated in the antenatal care program, irrespective of their health status [63]. Shared care was the most common type of antenatal care (55.6%), compared to only seeing a midwife (38.4%) or a GP (6%). According to the Norwegian Health Atlas, based on registry data, midwives handled 57% of antenatal contacts in primary care, while GPs or out-of-hours accounted for 43%. However, the proportion receiving shared care was not accounted for [36]. A Norwegian non–peer-reviewed survey among pregnant women and parents of young children reported that two-thirds of women received shared antenatal care between midwives and GPs [64].

Most women were satisfied with the antenatal care. Our study found lower involvement of GPs compared to other studies, which was unexpected. Several factors could contribute to the higher attendance of pregnant women at midwifery services. A possible explanation could be the limited availability of GPs due to high workloads and a growing number of unfilled positions. By September 2022, approximately 4% of Norwegians were without a GP [65]. While midwifery services in municipalities have been significantly strengthened, with a nearly 70% increase from 2015 to 2020 [66], the number of GPs grew by less than 8% over the same period [67]. Another reason for the higher attendance at midwives' services could be the longer consultations and their specialized training and experience in antenatal care. This discrepancy is worth debating, as midwives are also a limited resource at a time when Norway, like many other countries, needs to focus on prioritization and sustainability in health services to secure good antenatal care for all, including those who need it the most [68].

A key goal of antenatal care is to identify individual vulnerability factors that indicate an increased risk of complications during pregnancy, childbirth, or the postpartum period [69-71]. Midwives can provide valuable continuity of care throughout pregnancy, which has been shown to be beneficial [72,73]. A GP working within a continuity-of-care care model is ideally positioned to assess vulnerability, drawing on their comprehensive knowledge of both the woman's psychosocial circumstances and prior medical history [74,75]. Furthermore, since newborns are typically registered with the same GP as their mother, both mother and child are likely to benefit from the regular GP's unique life course perspective [76]. A registry study found that an increase in GP density in Norway significantly improved perinatal outcomes, including reduced fetal deaths and higher birth weights [77]. Although our study was not designed to assess complications, our finding that poor SRH was linked to more frequent contact with physicians (GPs or gynecologists) supports this argument. We do not know the reason for the more frequent contact, but the Norwegian antenatal guideline recommends seeking a physician for certain conditions, such as serious health issues, contagious diseases, and serious symptoms like bleeding and strong pain [31].

While pregnant women in Norway can chose between midwife- or GP-led checkups, the antenatal program in Denmark is designed for shared care. It includes three GP assessments (at gestational weeks 6 - 10, 25, and 32) combined with five to seven appointments with a midwife [78]. The fact that most participants in our study received shared antenatal care from both GPs and midwives indicates that Norwegian women might prefer this kind of care [64]. However, the coordination of antenatal visits between midwives and GPs is often left to the woman herself, and this could be a reason for the established habits of attending more antenatal checkups than formally recommended in national guidelines [36]. Therefore, there are compelling arguments for formalizing shared care and securing GP involvement in pregnancy care in Norway, also from a sustainability perspective. Further research is needed for this topic, especially concerning pregnant women with poorer health. The key challenge will be to establish effective communication

XSL•FO
RenderX

among all involved professionals as interdisciplinary collaboration has been found to be complex [74,79]. Although electronic messages between health personnel is possible in Norway, continuity of information flow in pregnancy care is currently secured by the woman carrying a physical "Antenatal Card" [80]. An ongoing national digitalization project is, however, expected to facillitate a seamless electronic communication in the future.

Methodological Strengths and Limitations

We consider the face validity of the questionnaire to be strong. This is supported by a successful pilot test with minimal feedback, and very few missing data, suggesting that participants found the study both relevant and well-aligned with its objectives. Although the authors developed the questionnaire, it contained several validated questions [41]. The questionnaire was anonymous, and participants had the opportunity to answer multiple times, as the data collection tool (Nettskjema.no) did not track IP addresses. Members of overlapping social media groups would receive an invitation for each group they belonged to. Although we considered this before the study began, we concluded that multiple responses from the same participant were unlikely, among other reasons because the survey was quite comprehensive. Furthermore, receiving all answers within few days lowers the risk of multiple responses. However, we cannot completely rule it out. Missing data was negligible and would not have substantially changed results, even if the missing data were not at random. Therefore, more advanced methods like multiple imputation were not considered. We collected data shortly after COVID-19 restrictions were lifted in Norway, and its possible that participants' health and their recall of their health before pregnancy was still adversely affected by the pandemic [51,52].

Although we received 1402 responses, which is substantial, it represents only a portion of the 51,292 women who gave birth in Norway in 2022. Therefore, while we recognize the significance of our findings, we must remain cautious about the weight we assign to the results. Odds ratios are greater than risk ratios when the outcome is frequent and should not be read as the relative risk. The sociodemographic characteristics of the study population (age, parity, and marital status) were comparable to the general population of pregnant women in Norway, except for a lower proportion of immigrants. Findings therefore might not be generalizable to immigrants. Self-selection to participation might also depend on the included topics, such as worries and strategies for information seeking. The broad age categories increase the chances of residual confounding by age. Beyond these sociodemographic factors, it is challenging to evaluate the representativity of SRH and SRMH among our participants. Since SRH and SRMH were assessed retrospectively, recall bias could be a concern. However, it may have only been a few months since the participants were not pregnant. In addition, we expect recall of such broad concepts to be reasonably accurate, and our findings of their SRH align closely with average SRH reported in national statistics [8].

We used single-item measures for SRH to describe participants' health. Although this subjective measure might seem simplistic and some nuances in variations in health may be lost, it is a comprehensive indicator that has been shown to correlate well with morbidity and mortality [4,81]. Furthermore, single-item SRMH correlates moderately with mental health scales and appears to be a robust measure of population mental health [82]. However, SRMH is not precisely defined, and assessment using the word "emotionally" in addition to "mentally" may have lowered the threshold for reporting poor SRMH, but overall, these single-item measures remain valuable tools for understanding the health status of our participants. Still, to get more comprehensive insights, future studies may consider incorporating multi-item scales.

Conclusion

Both SRH and SRMH were reported to be worse during pregnancy compared to before. Factors such as a difficult maternal childhood, low socioeconomic status, and previous childbirth were associated with poorer health during pregnancy. Both midwives and GPs played crucial roles in delivering antenatal care, shared care being the most common, however few women received care exclusively from GPs. The likelihood of physician involvement in care increased slightly with declining health. Recognizing these factors is essential for providing appropriate support and tailoring effective antenatal care.

Acknowledgments

The authors want to thank Ingvill Engvik Øvsthus, clinical psychologist and PhD candidate, for valuable advice and assistance with the recruitment of respondents via various social media platforms.

MHH and HSM collaborated on this study as medical students at NTNU until their graduation in June 2023. BPM and BA are contract GPs, specialists in general practice, and associate professors of general practice. GAV is an MD and associate professor of health services research. LOG is an MD, a professor of behavioral sciences in medicine, and the head of the General Practice Research Unit at NTNU.

Data Availability

The datasets used and analyzed during this study are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Authors' Contributions

BA, BPM, LOG, MHH, and HSM developed the questionnaire with inputs from Ingvill Øvsthus. MHH and HSM conducted the study as part of their MD student thesis, supervised by BPM, BA, and LOG. BA wrote the first draft of the study, and discussed the paper several times with BPM and GAV, who contributed in the writing and review of the paper. GAV contributed to the statistics. All authors gave feedback and have approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The questionnaire, translated into English.
[DOCX File, 27 KB - formative_v9i1e68811_app1.docx]

Multimedia Appendix 2
Demographic characteristics of pregnant women in Norway compared to this study.
[DOCX File, 17 KB - formative_v9i1e68811_app2.docx]

Multimedia Appendix 3
Odds ratios with 95% CIs for good self-rated health according to perception of childhood, socioeconomic situation, and parity.
[DOCX File, 17 KB - formative_v9i1e68811_app3.docx]

Multimedia Appendix 4
Odds ratios with 95% CIs for good self-rated mental health according to perception of childhood, current socioeconomic situation and parity.
[DOCX File, 17 KB - formative_v9i1e68811_app4.docx]

References

1. Grand-Guillaume-Perrenoud JA, Origlia P, Cignacco E. Barriers and facilitators of maternal healthcare utilisation in the perinatal period among women with social disadvantage: A theory-guided systematic review. Midwifery 2022 Feb;105:103237. [doi: 10.1016/j.midw.2021.103237] [Medline: 34999509]
2. Dube A, Mwandira K, Akter K, et al. Evaluating theory of change to improve the functioning of the network for improving quality of care for maternal, newborn and child health. PLOS Glob Public Health 2024;4(8):e0003532. [doi: 10.1371/journal.pgph.0003532] [Medline: 39088520]
3. Souza JP, Day LT, Rezende-Gomes AC, et al. A global analysis of the determinants of maternal health and transitions in maternal mortality. Lancet Glob Health 2024 Feb;12(2):e306-e316. [doi: 10.1016/S2214-109X(23)00468-0] [Medline: 38070536]
4. Jylhä M. What is self-rated health and why does it predict mortality? Towards a unified conceptual model. Soc Sci Med 2009 Aug;69(3):307-316. [doi: 10.1016/j.socscimed.2009.05.013] [Medline: 19520474]
5. Christian LM, Iams J, Porter K, Leblebicioglu B. Self-rated health among pregnant women: associations with objective health indicators, psychological functioning, and serum inflammatory markers. Ann Behav Med 2013 Dec;46(3):295-309. [doi: 10.1007/s12160-013-9521-7] [Medline: 23765366]
6. Palladino R, Tayu Lee J, Ashworth M, Triassi M, Millett C. Associations between multimorbidity, healthcare utilisation and health status: evidence from 16 European countries. Age Ageing 2016 May;45(3):431-435. [doi: 10.1093/ageing/afw044] [Medline: 27013499]
7. Schnittker J, Bacak V. The increasing predictive validity of self-rated health. PLoS One 2014;9(1):e84933. [doi: 10.1371/journal.pone.0084933] [Medline: 24465452]
8. Health at a glance 2023: OECD indicators. OECD. 2025. URL: https://www.oecd.org/en/publications/health-at-a-glance-2023_7a7afb35-en.html [accessed 2025-05-17]
9. Felitti VJ, Anda RF, Nordenberg D, et al. Relationship of childhood abuse and household dysfunction to many of the leading causes of death in adults. The Adverse Childhood Experiences (ACE) Study. Am J Prev Med 1998 May;14(4):245-258. [doi: 10.1016/s0749-3797(98)00017-8] [Medline: 9635069]
10. Hughes K, Bellis MA, Hardcastle KA, et al. The effect of multiple adverse childhood experiences on health: a systematic review and meta-analysis. Lancet Public Health 2017 Aug;2(8):e356-e366. [doi: 10.1016/S2468-2667(17)30118-4] [Medline: 29253477]
11. Russ SA, Larson K, Tullis E, Halfon N. A lifecourse approach to health development: implications for the maternal and child health research agenda. Matern Child Health J 2014 Feb;18(2):497-510. [doi: 10.1007/s10995-013-1284-z] [Medline: 23955383]

XSL•FO
RenderX

12. Dye CK, Alschuler DM, Wu H, et al. Maternal adverse childhood experiences and biological aging during pregnancy and in newborns. JAMA Netw Open 2024 Aug 1;7(8):e2427063. [doi: 10.1001/jamanetworkopen.2024.27063] [Medline: 39120899]
13. Racine N, Devereaux C, Cooke JE, Eirich R, Zhu J, Madigan S. Adverse childhood experiences and maternal anxiety and depression: a meta-analysis. BMC Psychiatry 2021 Jan 11;21(1):28. [doi: 10.1186/s12888-020-03017-w] [Medline: 33430822]
14. Hardcastle K, Ford K, Bellis MA. Maternal adverse childhood experiences and their association with preterm birth: secondary analysis of data from universal health visiting. BMC Pregnancy Childbirth 2022 Feb 16;22(1):129. [doi: 10.1186/s12884-022-04454-z] [Medline: 35172776]
15. Harter CL, Harter JFR. The link between adverse childhood experiences and financial security in adulthood. J Fam Econ Issues 2022;43(4):832-842. [doi: 10.1007/s10834-021-09796-y] [Medline: 34522076]
16. Lagadec N, Steinecker M, Kapassi A, et al. Factors influencing the quality of life of pregnant women: a systematic review. BMC Pregnancy Childbirth 2018 Nov 23;18(1):455. [doi: 10.1186/s12884-018-2087-4] [Medline: 30470200]
17. Moor I, Spallek J, Richter M. Explaining socioeconomic inequalities in self-rated health: a systematic review of the relative contribution of material, psychosocial and behavioural factors. J Epidemiol Community Health 2017 Jun;71(6):565-575. [doi: 10.1136/jech-2016-207589] [Medline: 27682963]
18. Viirman F, Hesselman S, Wikström AK, et al. Self-rated health before pregnancy and adverse birth outcomes in Sweden: A population-based register study. Birth 2021 Dec;48(4):541-549. [doi: 10.1111/birt.12567] [Medline: 34184771]
19. Schytt E, Hildingsson I. Physical and emotional self-rated health among Swedish women and men during pregnancy and the first year of parenthood. Sex Reprod Healthc 2011 Apr;2(2):57-64. [doi: 10.1016/j.srhc.2010.12.003] [Medline: 21439522]
20. Ferreira MS, da Silva ZP, de Almeida MF, Alencar GP. Is parenthood associated with self-rated health among women in Brazil? PLoS One 2023;18(10):e0293262. [doi: 10.1371/journal.pone.0293262] [Medline: 37903132]
21. Lee E, Song J. The effect of physical and mental health and health behavior on the self-rated health of pregnant women. Healthcare (Basel) 2021 Aug 28;9(9):1117. [doi: 10.3390/healthcare9091117] [Medline: 34574890]
22. Juarez Padilla J, Singleton CR, Pedersen CA, Lara-Cinisomo S. Associations between self-rated health and perinatal depressive and anxiety symptoms among latina women. Int J Environ Res Public Health 2022 Sep 22;19(19):11978. [doi: 10.3390/ijerph191911978] [Medline: 36231278]
23. Federenko IS, Wadhwa PD. Women's mental health during pregnancy influences fetal and infant developmental and health outcomes. CNS Spectr 2004 Mar;9(3):198-206. [doi: 10.1017/s1092852900008993] [Medline: 14999160]
24. Ghahremani T, Magann EF, Phillips A, Ray-Griffith SL, Coker JL, Stowe ZN. Women's mental health services and pregnancy: a review. Obstet Gynecol Surv 2022 Feb;77(2):122-129. [doi: 10.1097/OGX.0000000000000994] [Medline: 35201363]
25. Karalexi MA, Eberhard-Gran M, Valdimarsdóttir UA, Karlsson H, Munk-Olsen T, Skalkidou A. Perinatal mental health: how nordic data sources have contributed to existing evidence and future avenues to explore. Nord J Psychiatry 2022 Aug;76(6):423-432. [doi: 10.1080/08039488.2021.1998616] [Medline: 35057712]
26. Le Strat Y, Dubertret C, Le Foll B. Prevalence and correlates of major depressive episode in pregnant and postpartum women in the United States. J Affect Disord 2011 Dec;135(1-3):128-138. [doi: 10.1016/j.jad.2011.07.004] [Medline: 21802737]
27. Shakeel N, Sletner L, Falk RS, et al. Prevalence of postpartum depressive symptoms in a multiethnic population and the role of ethnicity and integration. J Affect Disord 2018 Dec 1;241:49-58. [doi: 10.1016/j.jad.2018.07.056] [Medline: 30096592]
28. Bedaso A, Adams J, Peng W, Sibbritt D. The relationship between social support and mental health problems during pregnancy: a systematic review and meta-analysis. Reprod Health 2021 Jul 28;18(1):162. [doi: 10.1186/s12978-021-01209-5] [Medline: 34321040]
29. Ibanez G, Blondel B, Prunet C, Kaminski M, Saurel-Cubizolles MJ. Prevalence and characteristics of women reporting poor mental health during pregnancy: findings from the 2010 French National Perinatal Survey. Rev Epidemiol Sante Publique 2015 Apr;63(2):85-95. [doi: 10.1016/j.respe.2015.02.023] [Medline: 25841615]
30. WHO recommendations on antenatal care for a positive pregnancy experience. World Health Organization. 2016. URL: https://www.who.int/publications/i/item/9789241549912 [accessed 2025-05-17]
31. Nasjonal faglig retningslinje: svangerskapsomsorgen [national clinical practice guideline: the antenatal care] [Article in Norwegian]. The Norwegian Direcorate of Health. 2019. URL: https://www.helsedirektoratet.no/retningslinjer/svangerskapsomsorgen [accessed 2025-05-17]
32. Health at a Glance: Europe 2024: State of Health in the EU Cycle: OECD/European Commission; 2024.
33. Infant mortality rate, for both sexes combined (infant deaths per 1,000 live births). United Nations Data. 2022. URL: http://data.un.org/Data.aspx?q=infant+mortality+rate&d=PopDiv&f=variableID%3a77 [accessed 2025-05-17]
34. Diguisto C, Saucedo M, Kallianidis A, et al. Maternal mortality in eight European countries with enhanced surveillance systems: descriptive population based study. BMJ 2022 Nov 16;379:e070621. [doi: 10.1136/bmj-2022-070621] [Medline: 36384872]

XSL•FO
RenderX

35. Backe B. Antenatal care in Norway--many unnecessary check ups. Tidsskr Nor Laegeforen 2002 Aug 30;122(20):1989-1992. [Medline: 12555444]
36. Byhring H, Shu J, Mathisen S, et al. Obstetrics healthcare atlas. Helseatlas. 2019. URL: https://apps.skde.no/helseatlas/en/v2/fodsel [accessed 2025-05-17]
37. Truong BT, Lupattelli A, Kristensen P, Nordeng H. Sick leave and medication use in pregnancy: a European web-based study. BMJ Open 2017 Aug 3;7(8):e014934. [doi: 10.1136/bmjopen-2016-014934] [Medline: 28775180]
38. Who can get maternity allowance? [Article in Norwegian]. NAV (Norwegian Labour and Welfare Administration). 2023. URL: https://familie.nav.no/om-svangerskapspenger [accessed 2025-05-17]
39. Eysenbach G. Improving the quality of web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34. [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
40. Tomasdottir MO, Sigurdsson JA, Petursson H, et al. Self reported childhood difficulties, adult multimorbidity and allostatic load. a cross-sectional analysis of the Norwegian HUNT study. PLoS ONE 2015;10(6):e0130591. [doi: 10.1371/journal.pone.0130591] [Medline: 26086816]
41. The HUNT study - a longitudinal population health study in norway [Article in Norwegian]. Spørreskjema HUNT [Questionnaires in The HUNT studies]. 2025. URL: https://www.ntnu.no/hunt/sp-rreskjema [accessed 2025-05-17]
42. Lyshol H, Grøtvedt L, Fagerhaug TN, Feuerherm AJ, Jakhelln G, Sen A. A study of socio-economic inequalities in self-reported oral and general health in South-East Norway. Sci Rep 2022 Aug 12;12(1):13721. [doi: 10.1038/s41598-022-18055-5] [Medline: 35962044]
43. Deevy M, Liu Streeter J, Hasler A, Lusardi A. Financial resilience in America global financial literacy excellence center (GFLEC). : The George Washington University School of Business; 2021.
44. Revold MK, Holmøy A. Levekårsundersøkelsen EU-SILC 2015 Tema: Boforhold, Utsatthet Og Uro for Lovbrudd [Living Conditions Survey EU-SILC 2015 Theme: Housing Conditions, Vulnerability, and Fear of Crime]: Statistics Norway; 2015. URL: https://www.nsd.no/data/individ/publikasjoner/NSD2279-Dokumentasjonsnotat.pdf [accessed 2025-05-17]
45. Hansen MH, Mørkved HS, Austad B, Vie G, Getz LO, Mjølstad BP. Worries and information seeking during pregnancy: a cross-sectional study among 1402 expectant Norwegian women active on social media platforms. Scand J Prim Health Care 2025 Feb 15;2025:1-12. [doi: 10.1080/02813432.2025.2461036] [Medline: 39953944]
46. Manor O, Matthews S, Power C. Dichotomous or categorical response? Analysing self-rated health and lifetime social class. Int J Epidemiol 2000 Feb;29(1):149-157. [doi: 10.1093/ije/29.1.149] [Medline: 10750617]
47. Haas JS, Jackson RA, Fuentes-Afflick E, et al. Changes in the health status of women during and after pregnancy. J Gen Intern Med 2005 Jan;20(1):45-51. [doi: 10.1111/j.1525-1497.2004.40097.x] [Medline: 15693927]
48. Khalaf A, Johansson M, Ghani RMA, Ny P. Self-rated health in Swedish pregnant women: a comprehensive population register study. Br J Midwifery 2022 Jun 2;30(6):306-315. [doi: 10.12968/bjom.2022.30.6.306]
49. Self-rated health, health at a glance 2023. OECD. 2023. URL: https://www.oecd.org/en/publications/health-at-a-glance-2023_7a7afb35-en/full-report/self-rated-health_ff2c161e.html [accessed 2025-05-17]
50. Henriksson P, Söderström E, Blomberg M, et al. Self-rated health in migrant and non-migrant women before, during and after pregnancy: a population-based study of 0.5 million pregnancies from the Swedish pregnancy register. J Clin Med 2020 Jun 6;9(6):1764. [doi: 10.3390/jcm9061764] [Medline: 32517225]
51. Eberhard-Gran M, Engelsen LY, Al-Zirqi I, Vangen S. Depressive symptoms and experiences of birthing mothers during COVID-19 pandemic. Tidsskr Nor Laegeforen 2022 Feb 15;142(3). [doi: 10.4045/tidsskr.21.0450] [Medline: 35170925]
52. The Norwegian Government. Timeline: news from Norwegian ministries about the coronavirus disease COVID-19. 2025. URL: https://www.regjeringen.no/en/topics/health-and-care/public-health/timeline-for-news-from-norwegian-ministries-about-the-coronavirus-disease-covid-19/id2692402 [accessed 2025-05-17]
53. Jahn A, Rysgaard TK, Andersen JH, Winding TN. Adverse childhood experiences and future self-rated health: a prospective cohort study. BMC Public Health 2021 May 12;21(1):895. [doi: 10.1186/s12889-021-10941-3] [Medline: 33975578]
54. Miller ES, Fleming O, Ekpe EE, Grobman WA, Heard-Garris N. Association between adverse childhood experiences and adverse pregnancy outcomes. Obstet Gynecol 2021 Nov 1;138(5):770-776. [doi: 10.1097/AOG.0000000000004570] [Medline: 34619717]
55. Swedo EA, D'Angelo DV, Fasula AM, Clayton HB, Ports KA. Associations of adverse childhood experiences with pregnancy and infant health. Am J Prev Med 2023 Apr;64(4):512-524. [doi: 10.1016/j.amepre.2022.10.017] [Medline: 36697281]
56. Zilmer BM, Overbeck G, Siersma V, Wilson P. Recent life events and adverse childhood experiences in predicting antenatal depression and anxiety: cross-sectional study. Fam Pract 2025 Jan 17;42(1):cmaf001. [doi: 10.1093/fampra/cmaf001] [Medline: 39820312]
57. Vederhus JK, Timko C, Haugland SH. Adverse childhood experiences and impact on quality of life in adulthood: development and validation of a short difficult childhood questionnaire in a large population-based health survey. Qual Life Res 2021 Jun;30(6):1769-1778. [doi: 10.1007/s11136-021-02761-0] [Medline: 33534031]
58. Bilsteen JF, Andresen JB, Mortensen LH, Hansen AV, Andersen AMN. Educational disparities in perinatal health in Denmark in the first decade of the 21st century: a register-based cohort study. BMJ Open 2018 Nov 8;8(11):e023531. [doi: 10.1136/bmjopen-2018-023531] [Medline: 30413512]

XSL•FO
RenderX

59. Kim MK, Lee SM, Bae SH, et al. Socioeconomic status can affect pregnancy outcomes and complications, even with a universal healthcare system. Int J Equity Health 2018 Jan 5;17(1):2. [doi: 10.1186/s12939-017-0715-7] [Medline: 29304810]
60. Taylor K, Compton S, Kolenic GE, et al. Financial hardship among pregnant and postpartum women in the United States, 2013 to 2018. JAMA Netw Open 2021 Oct 1;4(10):e2132103. [doi: 10.1001/jamanetworkopen.2021.32103] [Medline: 34714338]
61. Dørheim SK, Bjorvatn B, Eberhard-Gran M. Sick leave during pregnancy: a longitudinal study of rates and risk factors in a Norwegian population. BJOG 2013 Apr;120(5):521-530. [doi: 10.1111/1471-0528.12035] [Medline: 23130975]
62. Pedersen P, Labriola M, Nielsen CV, Maimburg RD, Nohr EA, Momsen AM. Systematic review of interventions targeting sickness absence among pregnant women in healthcare settings and workplaces. BMJ Open 2018 Oct 25;8(10):e024032. [doi: 10.1136/bmjopen-2018-024032] [Medline: 30366917]
63. Escañuela Sánchez T, Linehan L, O'Donoghue K, Byrne M, Meaney S. Facilitators and barriers to seeking and engaging with antenatal care in high-income countries: A meta-synthesis of qualitative research. Health Soc Care Community 2022 Nov;30(6):e3810-e3828. [doi: 10.1111/hsc.14072] [Medline: 36240064]
64. Gravide og foreldre ønsker bedre kvalitet fra helsetjenesten [pregnant women and parents want better quality from the health service] [Article in Norwegian]. Helseoversikt. 2022. URL: https://www.helseoversikt.no/nyheter/2022/6/14/gravide-og-foreldre-savner-bedre-kvalitet-fra-helsetjenesten [accessed 2025-05-17]
65. One out of 24 is without a GP [Article in Norwegian]. The Norwegian Medical Association. 2022. URL: https://www.legeforeningen.no/foreningsledd/yf/allmennlegeforeningen/krisen-i-fastlegeordningen/lister-uten-fast-lege/ [accessed 2025-05-17]
66. Jordmødre i kommunene [midwives in the municipalities] [Article in Norwegian]. Norwegian Directorate of Health. 2021. URL: https://www.helsedirektoratet.no/rapporter/tilgang-pa-og-behov-for-jordmodre/jordmodre-i-helsetjenestene/jordmodre-i-kommunene [accessed 2025-05-17]
67. Allmennlegetjenesten [regular GP service] [Article in Norwegian]. Statistisk sentralbyrå [Statistics Norway]. 2023. URL: https://www.ssb.no/statbank/table/12005/tableViewLayout1 [accessed 2025-05-17]
68. Helsepersonellkommisjonen [health personnell comission]. tid for handling: personellet i en bærekraftig helse- og omsorgstjeneste [time for action: the personnell in a sustainable healthcare service] [article in norwegian]. : Government.no; 2023 URL: https://www.regjeringen.no/no/dokumenter/nou-2023-4/id2961552/?ch=2
69. Brygger Venø L, Jarbøl DE, Pedersen LB, Søndergaard J, Ertmann RK. General practitioners' perceived indicators of vulnerability in pregnancy- A qualitative interview study. BMC Fam Pract 2021 Jun 26;22(1):135. [doi: 10.1186/s12875-021-01439-3] [Medline: 34174822]
70. Ertmann RK, Lyngsøe BK, Nicolaisdottir DR, Kragstrup J, Siersma V. Mental vulnerability before and depressive symptoms during pregnancy and postpartum: a prospective population-based cohort study from general practice. Nord J Psychiatry 2022 May;76(4):243-249. [doi: 10.1080/08039488.2021.1953583] [Medline: 34355638]
71. Biaggi A, Conroy S, Pawlby S, Pariante CM. Identifying the women at risk of antenatal anxiety and depression: A systematic review. J Affect Disord 2016 Feb;191:62-77. [doi: 10.1016/j.jad.2015.11.014] [Medline: 26650969]
72. Cibralic S, Pickup W, Diaz AM, et al. The impact of midwifery continuity of care on maternal mental health: A narrative systematic review. Midwifery 2023 Jan;116:103546. [doi: 10.1016/j.midw.2022.103546] [Medline: 36375410]
73. Perriman N, Davis DL, Ferguson S. What women value in the midwifery continuity of care model: A systematic review with meta-synthesis. Midwifery 2018 Jul;62:220-229. [doi: 10.1016/j.midw.2018.04.011] [Medline: 29723790]
74. Brygger Venø L, Pedersen LB, Søndergaard J, Ertmann RK, Jarbøl DE. Challenges in the cross-sectoral collaboration on vulnerable pregnant women: a qualitative study among Danish general practitioners. BMC Prim Care 2022 Jul 26;23(1):187. [doi: 10.1186/s12875-022-01773-0] [Medline: 35883047]
75. Sandvik H, Hetlevik Ø, Blinkenberg J, Hunskaar S. Continuity in general practice as predictor of mortality, acute hospitalisation, and use of out-of-hours care: a registry-based observational study in Norway. Br J Gen Pract 2022 Feb;72(715):e84-e90. [doi: 10.3399/BJGP.2021.0340] [Medline: 34607797]
76. Halfon N, Larson K, Lu M, Tullis E, Russ S. Lifecourse health development: past, present and future. Matern Child Health J 2014 Feb;18(2):344-365. [doi: 10.1007/s10995-013-1346-2] [Medline: 23975451]
77. Kinge JM, Grytten J. The impact of primary care physician density on perinatal health: Evidence from a natural experiment. Health Econ 2021 Dec;30(12):2974-2994. [doi: 10.1002/hec.4426] [Medline: 34498332]
78. Anbefalinger for svangreomsorgen [recommendations for antenatal care] [Article in Norwegian]. Sundhedsstyrelsen. 2021. URL: https://sst.dk/-/media/Udgivelser/2021/Anbefalinger-svangreomsorgen/Svangreomsorg-2022-ny.ashx [accessed 2025-05-17]
79. Espejord S, Auberg SH, Kvitno TK, Furskog-Risa C, Lukasse M. Norwegian community midwives' experience of interdisciplinary collaboration in care of pregnant women with vulnerabilities. Sex Reprod Healthc 2024 Mar;39:100951. [doi: 10.1016/j.srhc.2024.100951] [Medline: 38306732]
80. Antental health card. Helsenorge. 2023. URL: https://www.helsenorge.no/en/pregnancy-and-maternity-care-in-norway/antenatal-health-card/ [accessed 2025-05-17]
81. Bowling A. Just one question: If one question works, why ask several? J Epidemiol Community Health 2005 May;59(5):342-345. [doi: 10.1136/jech.2004.021204] [Medline: 15831678]

XSL•FO
RenderX

82. McAlpine DD, McCreedy E, Alang S. The meaning and predictive value of self-rated mental health among persons with a mental health problem. J Health Soc Behav 2018 Jun;59(2):200-214. [doi: 10.1177/0022146518755485] [Medline: 29406825]

Abbreviations

GP: general practitioner
OR: odds ratio
SRH: self-rated health
SRMH: self-rated mental health

Edited by A Mavragani; submitted 20.11.24; peer-reviewed by R Appiah-Kubi, S Satthapisit; revised version received 22.03.25; accepted 08.04.25; published 03.06.25.

Please cite as:
Austad B, Vie GÅ, Hansen MH, Mørkved HS, Getz LO, Mjølstad BP
Association of Self-Rated Health in Pregnancy With Maternal Childhood Experiences, Socioeconomic Status, Parity, and Choice of Antenatal Care Providers: Cross-Sectional Study
JMIR Form Res 2025;9:e68811
URL: https://formative.jmir.org/2025/1/e68811
doi:10.2196/68811

© Bjarne Austad, Gunnhild Åberge Vie, Mari Hegnes Hansen, Hanna Sandbakken Mørkved, Linn Okkenhaug Getz, Bente Prytz Mjølstad. Originally published in JMIR Formative Research (https://formative.jmir.org), 3.6.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Assessing the Causal Association Between COVID-19 and Graves Disease: Mendelian Randomization Study

Hui Nian[1*], MM; Yu Bai[2*], MM; Hua Yu[3], PhD

[1]Department of Thoracic Surgery, Shanghai Xuhui Central Hospital, Shanghai, China

[2]Department of Intensive Care Unit, Shanghai Xuhui Central Hospital, Shanghai, China

[3]Department of Plastic and Reconstructive Surgery, Shanghai Fourth People's Hospital Affiliated to Tongji University School of Medicine, No. 1279 Sanmen Road, Hongkou District, Shanghai, China

[*]these authors contributed equally

Corresponding Author:
Hua Yu, PhD
Department of Plastic and Reconstructive Surgery, Shanghai Fourth People's Hospital Affiliated to Tongji University School of Medicine, No. 1279 Sanmen Road, Hongkou District, Shanghai, China

Abstract

Background: Graves disease (GD) is an autoimmune thyroid disorder characterized by hyperthyroidism and autoantibodies. The COVID-19 pandemic has raised questions about its potential relationship with autoimmune diseases like GD.

Objective: This study aims to investigate the causal association between COVID-19 and GD through Mendelian randomization (MR) analysis and assess the impact of COVID-19 on GD.

Methods: We conducted an MR study using extensive genome-wide association study data for GD and COVID-19 susceptibility and its severity. We used stringent single nucleotide polymorphism selection criteria and various MR methodologies, including inverse-variance weighting, MR-Egger, and weighted median analyses, to assess causal relationships. We also conducted tests for directional pleiotropy and heterogeneity, as well as sensitivity analyses.

Results: The MR analysis, based on the largest available dataset to date, did not provide evidence supporting a causal relationship between COVID-19 susceptibility (odds ratio [OR] 0.989, 95% CI 0.405 - 2.851; *P*=.93), COVID-19 hospitalization (OR 0.974, 95% CI 0.852 - 1.113; *P*=.70), COVID-19 severity (OR 0.979, 95% CI 0.890 - 1.077; *P*=.66), and GD. Tests for directional pleiotropy and heterogeneity, as well as sensitivity analyses, supported these findings.

Conclusions: This comprehensive MR study does not provide sufficient evidence to support a causal relationship between COVID-19 and the onset or exacerbation of GD. These results contribute to a better understanding of the potential association between COVID-19 and autoimmune diseases, alleviating concerns about a surge in autoimmune thyroid diseases due to the pandemic. Further research is warranted to explore this complex relationship thoroughly.

(JMIR Form Res 2025;9:e66003) doi:10.2196/66003

KEYWORDS

Graves disease; COVID-19; Mendelian randomization; causal relationship; autoimmune disease

Introduction

Graves disease (GD) is an autoimmune thyroid disorder characterized by hyperthyroidism and the presence of autoantibodies. It is influenced by a combination of genetic, environmental, and immunological factors. Viral infections have long been implicated as potential triggers of autoimmune diseases, including autoimmune thyroid disorders [1-3]. Notably, the global COVID-19 pandemic has intensified concerns regarding its potential role in the development or exacerbation of autoimmune diseases such as GD [1,2].

Previous studies have explored the associations between viral infections and autoimmune thyroid diseases, with evidence suggesting that infections such as Epstein-Barr virus and hepatitis C virus may contribute to the onset or progression of these disorders [4,5]. More recently, research has focused on the potential immunological impact of SARS-CoV-2, the virus responsible for COVID-19. It has been proposed that SARS-CoV-2 can disrupt immune homeostasis, potentially leading to autoimmune reactions, including thyroid dysfunction [6]. While observational studies have reported an increased incidence of thyroid abnormalities in COVID-19 patients, these findings remain inconclusive in establishing a direct causal link between COVID-19 and autoimmune thyroid diseases [7-9].

The impact of COVID-19 on thyroid function has been further highlighted in recent systematic reviews and meta-analyses, which have suggested that thyroid abnormalities may be a common sequela of SARS-CoV-2 infection [10]. Additionally,

XSL•FO
RenderX

genetic predisposition plays a crucial role in determining susceptibility to autoimmune thyroid diseases, which underscores the need for a robust methodological approach to examine causality.

To address these uncertainties, this study uses Mendelian randomization (MR) analysis to investigate the causal relationship between COVID-19 and GD. MR leverages genetic variants as instrumental variables to approximate the effects of a randomized controlled trial, thereby minimizing confounding biases. This approach is based on 3 key assumptions: (1) genetic variants are not associated with confounding factors, (2) genetic variants influence an intermediate exposure variable, and (3) genetic variants affect the outcome solely through their impact on the exposure variable [11].

By elucidating the potential causal association between COVID-19 and GD, this research aims to contribute to a deeper understanding of the impact of SARS-CoV-2 on autoimmune thyroid diseases. These findings could inform future studies and clinical strategies for managing autoimmune thyroid disorders in the context of COVID-19.

Methods

Study Design

This study used a 2-sample MR design, relying exclusively on publicly accessible data for all analyses. We meticulously adhered to the recommendations presented in the Strengthening the Reporting of Observational Studies in Epidemiology-Mendelian Randomization (STROBE-MR) guidelines, ensuring a comprehensive and rigorous reporting approach.

Ethical Considerations

This study is a secondary analysis of publicly available genome-wide association study (GWAS) summary statistics obtained from the Integrative Epidemiology Unit OpenGWAS database [12]. As no new individual-level data were collected and no direct human participation was involved, ethical approval from an institutional review board was not required. The original GWASs adhered to their respective ethical guidelines and received necessary approvals. Since the study exclusively used deidentified, publicly available data, informed consent was not applicable; however, the original studies obtained informed consent from participants at the time of data collection and explicitly permitted secondary data analysis without additional consent. All data used in this study were fully anonymized, ensuring participant confidentiality, and no personally identifiable information was accessed, stored, or analyzed. In accordance with data protection regulations and ethical standards, strict measures were taken to maintain privacy and confidentiality. As no human participants were directly involved, no compensation was provided. Additionally, this study does not include any identifiable images or figures.

Data Sources for Exposures

We procured summary-level data for COVID-19 susceptibility (n=1,683,768), COVID-19 hospitalization (n=1,882,773), and COVID-19 severity (n=1,388,342) from the Integrative Epidemiology Unit OpenGWAS database [12]. This database offers extensive GWAS summary statistics encompassing a diverse array of traits and outcomes.

The data on COVID-19 susceptibility included 38,984 individuals, comprising individuals diagnosed with COVID-19 through laboratory confirmation of SARS-CoV-2 infection, electronic health records using *International Classification of Diseases* (*ICD*) codes or physician notes, or self-reported cases. For COVID-19 hospitalization, 9986 individuals were included, while COVID-19 severity encompassed 5101 individuals categorized as being in an extremely critical condition, which included individuals who experienced mortality or required advanced respiratory support [13-15].

Detailed information regarding the genetic variants associated with these COVID-19-related outcomes is presented in 3/7/2025 [16], offering a comprehensive overview of the exposure variables used in our MR analysis.

Outcome Data

The outcome data for GD were extracted from publicly available summary statistics, involving a total of 2176 GD cases and 210,277 controls (GWAS ID: bbj-a-123). The cases represented individuals hospitalized for GD, and the control group consisted of individuals from cohorts without identified cases. The genetic instruments used for GD in this MR study were derived from these data, facilitating a comprehensive evaluation of the causal relationship between GD and pertinent genetic variations.

Instrument Construction

For constructing genetic instruments for each exposure, we used a rigorous approach. We identified single nucleotide polymorphisms (SNPs) displaying significant associations with the respective exposure at a genome-wide significance threshold ($P<5\times10^{-}$), which, while less stringent than the conventional threshold ($P<5\times10^{-}$), was chosen to balance instrument strength and the number of available SNPs. These SNPs underwent linkage disequilibrium clumping to remove highly correlated SNPs ($r^2\geq0.001$). Our reference population was drawn from the European cohort within the 1000 Genomes Project, which may limit the generalizability of our findings to non-European populations. To ensure consistency in effect sizes and address issues related to palindromic SNPs, we harmonized SNP-exposure and SNP-GD associations. In cases where specific SNPs were missing in the GD GWAS dataset, we used proxy SNPs ($r^2>0.8$). To mitigate potential bias from weak instruments, we performed *F* statistic calculations and conducted multiple sensitivity analyses, including leave-one-out analysis and MR-Egger regression. The processes of SNP extraction from GWAS summary-level data, linkage disequilibrium clumping, and harmonization were executed using the TwoSample MR package in R version 4.3.0 (R Core Team) [17,18]. The strength of the instrumental variables was evaluated using the F statistic, computed as follows: $F = (R^2/K)/[(1-R^2)(N-K-1)]$, where R^2 denotes the variance explained by the instruments, K is the number of instruments, and N is the sample size [19].

XSL•FO
RenderX

Statistical Analyses

Using a 2-sample MR approach, we estimated the causal effects of COVID-19 susceptibility and its severity on GD. The primary MR analysis used the inverse-variance weighting (IVW) method, assuming that the genetic instruments collectively satisfy the core MR assumptions. In addition to the primary MR analysis, we conducted supplementary sensitivity analyses, including MR-Egger, weighted median, and MR-pleiotropy residual sum and outlier (MR-PRESSO) methods. These methods enhance robustness against potential pleiotropic effects by relaxing certain MR assumptions. To assess horizontal pleiotropy, we used the intercept term from MR-Egger regression, where an intercept *P* value exceeding .05 indicated an absence of pleiotropy. Furthermore, we applied MR-PRESSO as a supplementary method to detect and exclude outlier genetic instruments with pleiotropic effects, subsequently estimating causal effects. All tests were 2-sided and executed using the TwoSampleMR and MR-PRESSO packages in R version 4.3.0. A significance threshold was set at *P*<.05. Since our analysis exclusively used publicly available data, no ethical approval from a local committee was sought.

Results

In summary, this MR analysis encompassed GWAS data for 3 COVID-19 statuses, including COVID-19 susceptibility and its severity, in relation to GD (detailed information provided in Table 1).

Regarding COVID-19 susceptibility, the IVW MR analysis indicated that GD did not significantly increase the risk (odds ratio [OR] 0.989, 95% CI 0.405 - 2.851; *P*=.93; Figures 1A-3A and Table 2). No heterogeneity among SNPs was observed (Cochran Q *P*=.39). The MR-Egger intercept was close to zero (intercept=−0.0051; *P*=.85), implying the absence of directional pleiotropic effects (Table 3). MR-PRESSO did not detect any outlier SNPs (*P*=.47). The results of the leave-one-out sensitivity analysis supported the observed causal association (Figure 4A).

The results of the IVW MR analysis indicated that there was no significant association between COVID-19 hospitalization and the risk of GD (OR 0.974, 95% CI 0.852 - 1.113; *P*=.70; Figures 1B-3B and Table 2). Heterogeneity was not observed among the various studies (*P*=.27). Sensitivity analysis using the MR-PRESSO method also failed to yield statistically significant causal estimates (*P*=.33). Furthermore, the MR-Egger intercept suggested an absence of horizontal pleiotropy (intercept=−0.0039; *P*=.89). The MR-PRESSO method did not identify any outlier SNPs (*P*=.33; Table 3). Leave-one-out sensitivity analysis did not reveal any individual SNPs that significantly influenced the causal estimates (Figure 4B).

In the primary MR analysis using the IVW method, there was no significant association observed between COVID-19 severity and the risk of GD. The OR for this relationship was 0.979 (95% CI 0.890 - 1.077; *P*=.66; Figures 1C-3C and Table 2). The results were consistent when using the weighted median method (OR 0.975, 95% CI 0.852 - 1.116; *P*=.71) and conducting the MR-PRESSO sensitivity analysis (*P*=.20). Furthermore, the Cochran Q test revealed no evidence of heterogeneity (*P*=.20), and the MR-Egger intercept indicated the absence of horizontal pleiotropy (intercept=0.0352; *P*=.37). No potential outliers were identified using the MR-PRESSO method (Table 3). Leave-one-out sensitivity analysis did not reveal any individual SNPs that significantly influenced the causal estimates (Figure 4C). The symmetry in the funnel plot further supported the reliability of the MR analysis (Figure 3).

Table . Data sources for the analysis.

Phenotype and consortium	Source of genetic variations	
	Cases	Controls
Graves disease	2176 individuals hospitalized for Graves disease	210,277 individuals from cohorts without identified cases
COVID-19 susceptibility and its severity		
Susceptibility	38,984 individuals who were diagnosed with COVID-19 using laboratory confirmation of SARS-CoV-2 infection, electronic health records (which included the use of *ICD*[a] codes or physician notes), or self-reported cases	1,644,784 individuals who were registered within the same groups but were excluded from the cases under investigation
Hospitalization	9986 individuals who were admitted to hospitals as a result of contracting COVID-19	1,877,672 individuals who were enrolled in the study cohorts but did not meet the criteria for classification as cases, indicating that they did not require hospitalization due to COVID-19
Very severe	5101 individuals categorized as being in an extremely critical condition, which encompasses those who experienced mortality or required advanced respiratory support, including interventions such as continuous positive airway pressure (CPAP), bilevel positive airway pressure (BiPAP), endotracheal intubation, or high-flow nasal cannula therapy	1,383,241 individuals who were included in the study cohorts but did not meet the criteria for classification as cases

[a]*ICD*: *International Classification of Diseases*.

XSL•FO
RenderX

Figure 1. Forest plot of the results of Mendelian randomization (MR) analysis. (A) COVID-19 susceptibility, (B) COVID-19 hospitalization, and (C) COVID-19 severity.

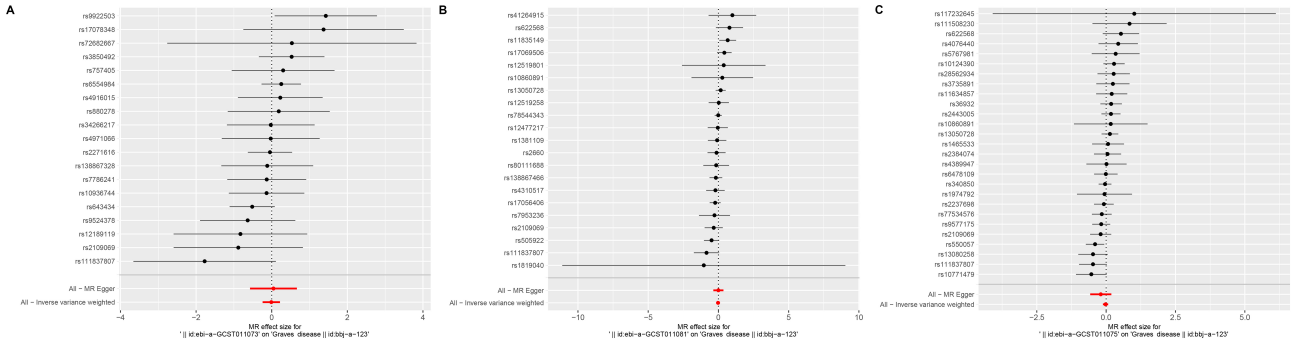

Figure 2. Scatter plot of the results of Mendelian randomization (MR) analysis. (A) COVID-19 susceptibility, (B) COVID-19 hospitalization, and (C) COVID-19 severity. SNP: single nucleotide polymorphism.

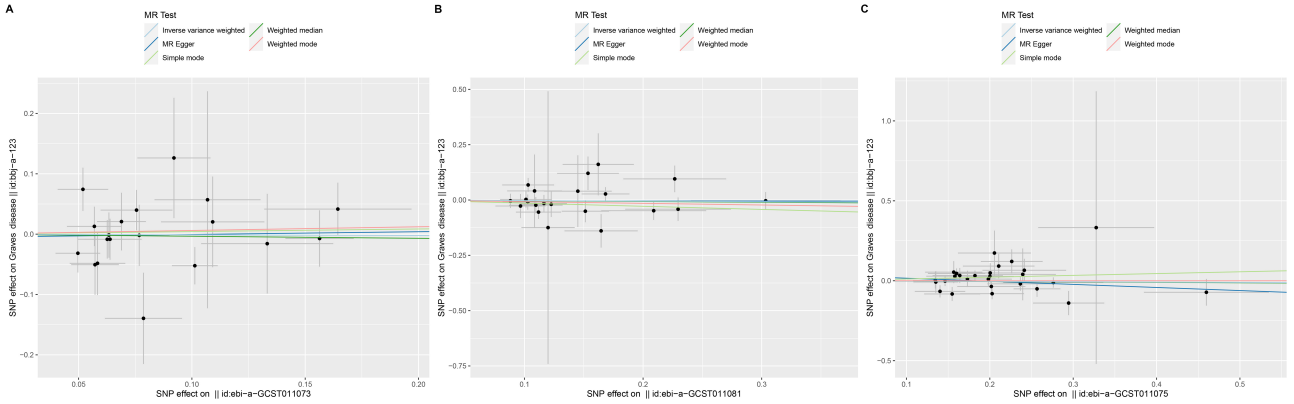

Figure 3. Funnel plot of the results of Mendelian randomization (MR) analysis. (A) COVID-19 susceptibility, (B) COVID-19 hospitalization, and (C) COVID-19 severity.

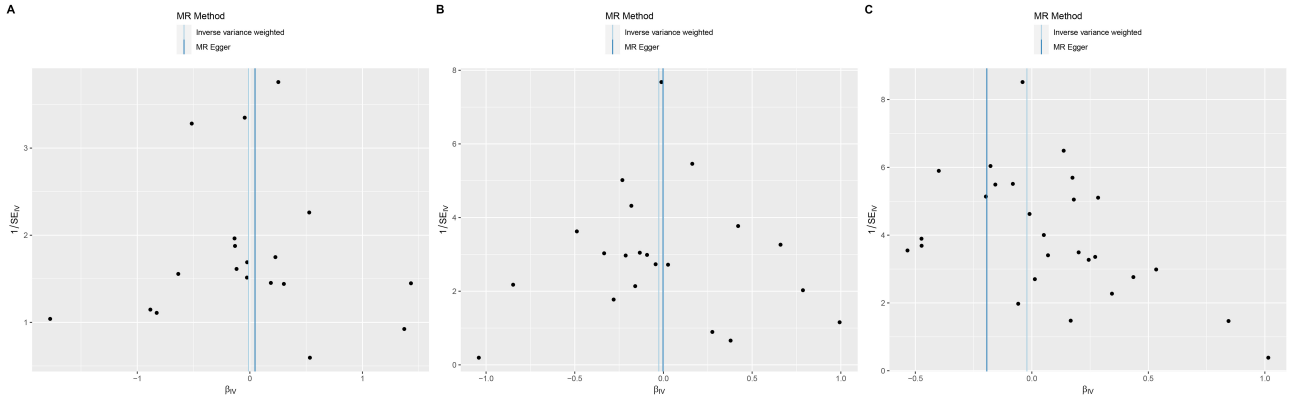

XSL•FO
RenderX

Table . Mendelian randomization estimates of the association between Graves disease and COVID-19.

Outcome	COVID-19 susceptibility			COVID-19 hospitalization			COVID-19 very severe		
	SNPs[a], n	OR[b] (95% CI)	*P* value	SNPs, n	OR (95% CI)	*P* value	SNPs, n	OR (95% CI)	*P* value
Graves disease									
IVW[c]	19	0.989 (0.405 - 2.851)	.93	21	0.974 (0.852 - 1.113)	.70	27	0.979 (0.890 - 1.077)	.66
Weighted median	19	0.966 (0.685 - 1.362)	.84	21	0.967 (0.797 - 1.175)	.74	27	0.975 (0.852 - 1.116)	.71
MR-Egger[d]	19	1.047 (0.786 - 1.245)	.89	21	0.997 (0.696 - 1.431)	.99	27	0.824 (0.563 - 1.207)	.33
Simple mode	19	1.043 (0.632 - 1.723)	.87	21	0.868 (0.655 - 1.151)	.34	27	1.119 (0.871 - 1.436)	.39

[a]SNP: single nucleotide polymorphism.

[b]OR: odds ratio.

[c]IVW: inverse-variance weighting.

[d]MR-Egger: Mendelian randomization-Egger regression.

Table . Results of heterogeneity and horizontal pleiotropy analyses.

Outcome	Exposure	*P* value (heterogeneity)[a]	MR-Egger[b] intercept	*P* value (pleiotropy)[c]	*P* value (global test)[d]
Graves disease	COVID-19 susceptibility	.39	−0.0051	.85	.47
Graves disease	COVID-19 hospitalization	.27	−0.0039	.89	.33
Graves disease	COVID-19 very severe	.20	0.0352	.37	.20

[a]The *P* value of Cochran Q value in the heterogeneity test.

[b]MR-Egger: Mendelian randomization-Egger.

[c]The *P* value for the intercept in the MR-Egger regression.

[d]The *P* value for the global test in the Mendelian randomization-pleiotropy residual sum and outlier test.

Figure 4. Leave-one-out analysis plots of the results of Mendelian randomization (MR) analysis. (A) COVID-19 susceptibility, (B) COVID-19 hospitalization, and (C) COVID-19 severity.

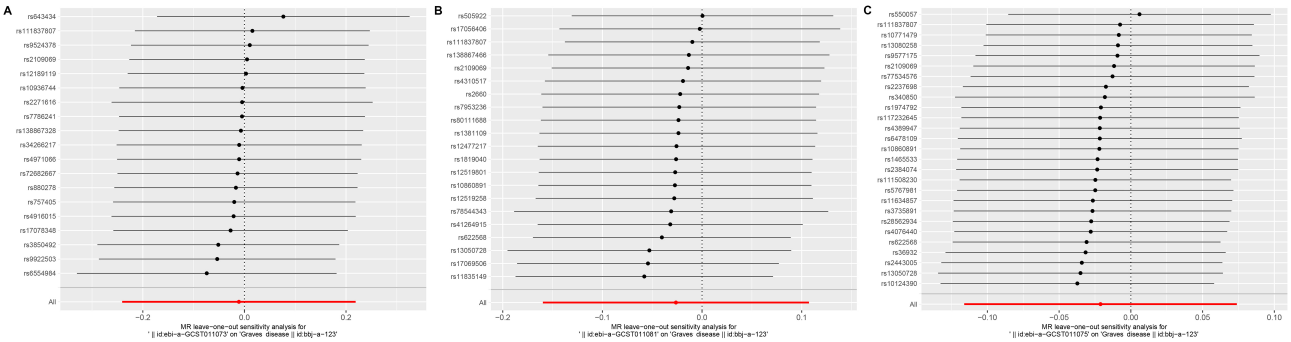

Discussion

Principal Findings

In this MR study, we extensively explored the potential causal relationships between COVID-19 susceptibility, COVID-19 hospitalization, and COVID-19 severity in relation to GD. Using a variety of MR methods, we aimed to provide a comprehensive understanding of these relationships. Our study results underscore the absence of a causal link between COVID-19 susceptibility and GD. Recent research continues to unveil the intricate interplay between autoimmune diseases and viral infections. For example, a study conducted by Widhani et al [20] delved into the potential associations between viral infections, including COVID-19, and autoimmune thyroid diseases, further corroborating our findings. Additionally, the work of Smith and Hegedüs [21] has provided in-depth insights into the pathogenesis of GD, emphasizing its multifactorial nature. These recent studies lend robust support to our research. Furthermore, investigations into the relationship between COVID-19 and autoimmune diseases have expanded in recent

XSL•FO
RenderX

years. A systematic review by Tutal et al [22] examined the impact of autoimmune thyroid diseases on the severity of COVID-19, enriching our understanding of the bidirectional influences between these health conditions. To provide a broader clinical context, we have integrated evidence on how COVID-19 susceptibility and outcomes differ across other autoimmune diseases, which highlights the need for more nuanced investigations beyond GD alone [23].

Similarly, we found that COVID-19 hospitalization does not significantly increase the risk of GD. The consistency of this result across various MR methods suggests that COVID-19 hospitalization is not a causal risk factor for GD. This conclusion is pivotal for understanding the etiology of GD and its relationship with other health conditions. Recent research has provided further insights into the connection between COVID-19 and autoimmune thyroid diseases. For example, a study by Fallahi et al [24], conducted within the past 5 years, investigated the impact of COVID-19 hospitalization on the development of autoimmune thyroid diseases, including GD. The study used a large-scale patient database and used statistical methods to evaluate the risk of developing autoimmune thyroid diseases after COVID-19 hospitalization. Their findings, which align with our MR results, indicated that COVID-19 hospitalization was not significantly associated with an increased risk of GD or other autoimmune thyroid diseases. This further supports the critical importance of recognizing that COVID-19 hospitalization itself does not have a direct causal role in the development of GD. Furthermore, we examined the relationship between COVID-19 severity and the risk of developing GD. Our study revealed that COVID-19 severity does not contribute to the onset of GD. This finding provides important insights into the genetic and environmental factors influencing GD's pathogenesis. A recent study by Bostan and colleagues [25] explored the genetic mechanisms underlying the development of GD in individuals with severe COVID-19. Their research highlighted the role of genetic susceptibility in the development of GD among those with severe COVID-19, adding to our understanding of the complex interaction between COVID-19 and autoimmune thyroid diseases.

Moreover, recent research has extensively explored the associations between COVID-19 and thyroid-related conditions. Studies have indicated that COVID-19 infection can influence thyroid function, potentially leading to hyperthyroidism or thyrotoxicosis in certain individuals, further emphasizing the potential link between COVID-19 and thyroid diseases. However, the specific mechanisms remain unclear and warrant further investigation. Recent studies have provided more in-depth insights into the complex interplay between COVID-19 and thyroid function. For instance, a study by Smith et al [21] identified COVID-19-related thyroid dysfunction, including subacute thyroiditis and hyperthyroidism, reporting cases of patients experiencing thyroid-related symptoms during or following COVID-19 infection. Similarly, Damara et al [10] conducted a systematic review and meta-analysis, revealing a significant association between autoimmune thyroid diseases and the severity of COVID-19. These studies underscore the importance of further understanding the relationship between COVID-19 and thyroid dysfunction. Additionally, Muller et al [26] examined the underlying mechanisms of COVID-19-induced thyroid dysfunction, particularly focusing on the immune response's role in thyroid autoimmunity. Their findings suggest that the cytokine storm triggered by COVID-19 may play a crucial role in the development of thyroid diseases. Liu et al [27] investigated the impact of SARS-CoV-2 infection on the expression of thyroid-related genes in thyroid tissue, shedding light on the potential molecular pathways involved. To broaden the context, we have also incorporated recent research on COVID-19 vaccine immunogenicity in patients with autoimmune diseases, providing additional understanding of how immune responses to COVID-19 vaccination may differ in individuals with pre-existing autoimmune conditions [28].

Furthermore, recent investigations have explored the complex interplay between GD and other thyroid-related conditions in the context of COVID-19. For example, Nakano et al [29] found that individuals with GD may be at an increased risk of contracting COVID-19 and could face more severe outcomes postinfection, including complications such as thyrotoxicosis. This study highlights the bidirectional relationship between GD and COVID-19, emphasizing the importance of comprehensive research to better understand the complex interactions between thyroid-related conditions and COVID-19. In a study by Sethi et al [30], researchers observed that COVID-19 infection was associated with thyroid abnormalities, with some patients developing thyrotoxicosis. This suggests that COVID-19 may influence thyroid function, raising concerns about potential implications for individuals with GD. Additionally, Croce et al [31] examined the bidirectional effects of GD and COVID-19, proposing mechanisms through which GD could increase susceptibility to COVID-19 and worsen clinical outcomes, such as immune system dysregulation and cytokine imbalances. However, both COVID-19 and GD can significantly impact physical and mental health, and thus, we have integrated insights from recent research on physical activity and mental health interventions, which may provide potential strategies for improving long-term patient outcomes [32]. Collectively, these studies underscore the urgent need for further research into the multifaceted interactions between COVID-19 and thyroid-related diseases, including GD. This will not only deepen our understanding of the pathogenesis of these conditions but also offer critical insights for health care practitioners, aiding in the development of more informed and effective clinical strategies.

Limitations

It is important to note that despite the strengths of the MR approach, our study has several limitations. MR methods rely on certain assumptions, including the absence of horizontal pleiotropy. Although we used various strategies to address potential pleiotropic effects, we cannot entirely exclude the possibility of unmeasured confounders influencing our results. Furthermore, our analysis primarily relied on data from the European population, which may limit the generalizability of our findings to other ethnic groups. Specifically, the wide CIs for COVID-19 susceptibility suggest that the study may be underpowered to detect modest effects. Additionally, the definition of COVID-19 cases included self-reported cases, which could introduce reporting bias. The combination of laboratory-confirmed, electronic health record–derived, and

XSL•FO
RenderX

self-reported cases may also affect the reliability of exposure classification. However, sensitivity analyses on subsets of laboratory-confirmed and electronic health record–derived cases support the robustness of our findings. Similarly, the definition of GD cases was restricted to hospitalized patients, which could result in selection bias, underrepresenting mild or moderate cases. While hospitalization-based diagnoses enhance diagnostic accuracy, future research should incorporate outpatient and primary care datasets to capture the full spectrum of disease severity. Lastly, this study did not fully explore the potential bidirectional relationships or temporal dynamics between COVID-19 and GD. Future studies should address these aspects using bidirectional MR and longitudinal cohort studies to further clarify the interactions between these conditions.

Conclusions

In summary, our MR analysis did not identify significant causal relationships between COVID-19 susceptibility, COVID-19 hospitalization, or COVID-19 severity and the risk of developing GD. These findings highlight the complex interactions between viral infections, autoimmune diseases, and genetic factors. However, they also emphasize the need for further investigation into the bidirectional relationships and temporal dynamics between COVID-19 and thyroid-related conditions. Expanding this research will deepen our understanding of the pathogenesis of both diseases, contributing valuable insights for clinical decision-making and management strategies.

Acknowledgments

The authors express their sincere gratitude to the Integrative Epidemiology Unit OpenGWAS Project [12] for granting access to the genome-wide association study (GWAS) summary-level data used in this research. This project is supported by the Talent Boosting Program (SY-XKZT-2020-3007) and the Discipline Boosting Program (SY-XKZT-2020-1021) of Shanghai Fourth People's Hospital.

Data Availability

The analysis in this research relied on publicly available datasets, which can be accessed through the Integrative Epidemiology Unit OpenGWAS Project [12].

Authors' Contributions

HN analyzed the data and wrote the manuscript. YB designed the study. HY collected the data and corrected the paper. All authors have read and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Chen K, Gao Y, Li J. New-onset and relapsed Graves' disease following COVID-19 vaccination: a comprehensive review of reported cases. Eur J Med Res 2023 Jul 13;28(1):232. [doi: 10.1186/s40001-023-01210-7] [Medline: 37443067]
2. Murugan AK, Alzahrani AS. SARS-CoV-2 plays a pivotal role in inducing hyperthyroidism of Graves' disease. Endocrine 2021 Aug;73(2):243-254. [doi: 10.1007/s12020-021-02770-6] [Medline: 34106438]
3. Almutairi H, Alqadi FS, Alsulaim RK, et al. Unveiling promising modalities and enhancing patient outcomes in Graves' disease treatment: a systematic review and meta-analysis. Cureus 2024 May;16(5):e60829. [doi: 10.7759/cureus.60829] [Medline: 38910658]
4. Msheik AN, Al Mokdad Z, Hamed F, et al. Epstein-Barr virus flare: a multiple sclerosis attack. Surg Neurol Int 2024;15:355. [doi: 10.25259/SNI_457_2024] [Medline: 39524581]
5. Cyna W, Wojciechowska A, Szybiak-Skora W, Lacka K. The impact of environmental factors on the development of autoimmune thyroiditis-review. Biomedicines 2024 Aug 7;12(8):1788. [doi: 10.3390/biomedicines12081788] [Medline: 39200253]
6. Mobasheri L, Nasirpour MH, Masoumi E, Azarnaminy AF, Jafari M, Esmaeili SA. SARS-CoV-2 triggering autoimmune diseases. Cytokine 2022 Jun;154:155873. [doi: 10.1016/j.cyto.2022.155873] [Medline: 35461172]
7. Naguib R. Potential relationships between COVID-19 and the thyroid gland: an update. J Int Med Res 2022 Feb;50(2). [doi: 10.1177/03000605221082898] [Medline: 35226548]
8. Darvishi M, Nazer MR, Shahali H, Nouri M. Association of thyroid dysfunction and COVID-19: a systematic review and meta-analysis. Front Endocrinol (Lausanne) 2022;13:947594. [doi: 10.3389/fendo.2022.947594] [Medline: 36387848]
9. Rossetti CL, Cazarin J, Hecht F, et al. COVID-19 and thyroid function: what do we know so far? Front Endocrinol (Lausanne) 2022;13:1041676. [doi: 10.3389/fendo.2022.1041676] [Medline: 36601011]
10. Damara FA, Muchamad GR, Ikhsani R, Syafiyah AH, Bashari MH. Thyroid disease and hypothyroidism are associated with poor COVID-19 outcomes: a systematic review, meta-analysis, and meta-regression. Diabetes Metab Syndr 2021;15(6):102312. [doi: 10.1016/j.dsx.2021.102312] [Medline: 34731819]

XSL•FO
RenderX

11. Sanderson E, Glymour MM, Holmes MV, et al. Mendelian randomization. Nat Rev Methods Primers 2022 Feb 10;2:6. [doi: 10.1038/s43586-021-00092-5] [Medline: 37325194]
12. GWAS summary data. MRC Integrative Epidemiology Unit. URL: https://gwas.mrcieu.ac.uk/ [accessed 2025-03-28]
13. COVID-19 Host Genetics Initiative. The COVID-19 Host Genetics Initiative, a global initiative to elucidate the role of host genetic factors in susceptibility and severity of the SARS-CoV-2 virus pandemic. Eur J Hum Genet 2020 Jun;28(6):715-718. [doi: 10.1038/s41431-020-0636-6] [Medline: 32404885]
14. Zhang Z, Fang T, Chen L, Qiu B, Lv Y. Thyroid function and COVID-19 susceptibility and its severity: a two-sample Mendelian randomization study. Endocrinology 2022 Oct 11;163(11):bqac139. [doi: 10.1210/endocr/bqac139] [Medline: 36048436]
15. Köhler A, Chen B, Gemignani F, et al. Genome-wide association study on differentiated thyroid cancer. The Journal of Clinical Endocrinology & Metabolism 2013 Oct;98(10):E1674-E1681. [doi: 10.1210/jc.2013-1941]
16. IEU OpenGWAS project. URL: https://gwas.mrcieu.ac.uk/ [accessed 2025-04-07]
17. Hu X, Zhao J, Lin Z, et al. Mendelian randomization for causal inference accounting for pleiotropy and sample structure using genome-wide summary statistics. Proc Natl Acad Sci U S A 2022 Jul 12;119(28):e2106858119. [doi: 10.1073/pnas.2106858119] [Medline: 35787050]
18. Zhu X, Li X, Xu R, Wang T. An iterative approach to detect pleiotropy and perform Mendelian Randomization analysis using GWAS summary statistics. Bioinformatics 2021 Jun 16;37(10):1390-1400. [doi: 10.1093/bioinformatics/btaa985] [Medline: 33226062]
19. Zhao SS, Holmes MV, Zheng J, Sanderson E, Carter AR. The impact of education inequality on rheumatoid arthritis risk is mediated by smoking and body mass index: Mendelian randomization study. Rheumatology (Oxford) 2022 May 5;61(5):2167-2175. [doi: 10.1093/rheumatology/keab654] [Medline: 34436562]
20. Widhani A, Koesnoe S, Maria S, et al. Factors related to severity, hospitalization, and mortality of COVID-19 infection among patients with autoimmune diseases. Trop Med Infect Dis 2023 Apr 18;8(4):227. [doi: 10.3390/tropicalmed8040227] [Medline: 37104352]
21. Smith TJ, Hegedüs L. Graves' disease. N Engl J Med 2016 Oct 20;375(16):1552-1565. [doi: 10.1056/NEJMra1510030] [Medline: 27797318]
22. Tutal E, Ozaras R, Leblebicioglu H. Systematic review of COVID-19 and autoimmune thyroiditis. Travel Med Infect Dis 2022;47:102314. [doi: 10.1016/j.tmaid.2022.102314] [Medline: 35307540]
23. Shin JI, Kim SE, Lee MH, et al. COVID-19 susceptibility and clinical outcomes in autoimmune inflammatory rheumatic diseases (AIRDs): a systematic review and meta-analysis. Eur Rev Med Pharmacol Sci 2022 May;26(10):3760-3770. [doi: 10.26355/eurrev_202205_28873] [Medline: 35647859]
24. Fallahi P, Ferrari SM, Elia G, et al. Thyroid autoimmunity and SARS-CoV-2 infection: report of a large Italian series. Autoimmun Rev 2022 Nov;21(11):103183. [doi: 10.1016/j.autrev.2022.103183] [Medline: 36007802]
25. Bostan H, Ucan B, Kizilgul M, et al. Relapsed and newly diagnosed Graves' disease due to immunization against COVID-19: a case series and review of the literature. J Autoimmun 2022 Apr;128:102809. [doi: 10.1016/j.jaut.2022.102809] [Medline: 35220164]
26. Muller I, Cannavaro D, Dazzi D, et al. SARS-CoV-2-related atypical thyroiditis. Lancet Diabetes Endocrinol 2020 Sep;8(9):739-741. [doi: 10.1016/S2213-8587(20)30266-7] [Medline: 32738929]
27. Liu H, Xin J, Cai S, Jiang X. Mendelian randomization analysis provides causality of smoking on the expression of ACE2, a putative SARS-CoV-2 receptor. Elife 2021 Jul 6;10:e64188. [doi: 10.7554/eLife.64188] [Medline: 34227468]
28. Cho K, Park S, Kim EY, et al. Immunogenicity of COVID-19 vaccines in patients with diverse health conditions: a comprehensive systematic review. J Med Virol 2022 Sep;94(9):4144-4155. [doi: 10.1002/jmv.27828] [Medline: 35567325]
29. Nakano Y, Takeshima K, Furukawa Y, Morita S, Sakata M, Matsuoka TA. Concomitant exacerbation of Graves orbitopathy and double-seronegative Myasthenia Gravis after SARS-CoV-2 Infection. JCEM Case Rep 2025 Feb;3(2):luaf019. [doi: 10.1210/jcemcr/luaf019] [Medline: 39882352]
30. Sethi Y, Uniyal N, Maheshwari S, Sinha R, Goel A. Thyroid function abnormalities in the acute phase of COVID-19: a cross-sectional hospital-based study from North India. Cureus 2022 May;14(5):e24942. [doi: 10.7759/cureus.24942] [Medline: 35706733]
31. Croce L, Gangemi D, Ancona G, et al. The cytokine storm and thyroid hormone changes in COVID-19. J Endocrinol Invest 2021 May;44(5):891-904. [doi: 10.1007/s40618-021-01506-7] [Medline: 33559848]
32. Rahmati M, Lee S, Yon DK, et al. Physical activity and prevention of mental health complications: an umbrella review. Neurosci Biobehav Rev 2024 May;160:105641. [doi: 10.1016/j.neubiorev.2024.105641] [Medline: 38527637]

Abbreviations

GD: Graves disease
GWAS: genome-wide association study
ICD: *International Classification of Diseases*
IVW: inverse-variance weighting

XSL•FO
RenderX

MR: Mendelian randomization
MR-PRESSO: Mendelian randomizaion-pleiotropy residual sum and outlier
OR: odds ratio
SNP: single nucleotide polymorphism
STROBE-MR: Strengthening the Reporting of Observational Studies in Epidemiology-Mendelian Randomization

Edited by A Mavragani; submitted 01.09.24; peer-reviewed by M Chen, Q Du, SW Lee; revised version received 12.02.25; accepted 20.02.25; published 08.04.25.

Please cite as:
Nian H, Bai Y, Yu H
Assessing the Causal Association Between COVID-19 and Graves Disease: Mendelian Randomization Study
JMIR Form Res 2025;9:e66003
URL: https://formative.jmir.org/2025/1/e66003
doi:10.2196/66003

© Hui Nian, Yu Bai, Hua Yu. Originally published in JMIR Formative Research (https://formative.jmir.org), 8.4.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

COVID-19 Perceptions Among Communities Living on Ground Crossings of Somali Region of Ethiopia: Community Cross-Sectional Survey Study

Alinoor Mohamed Farah[1], MPH; Abdifatah Abdulahi[2], MD; Abdulahi Hussein[2], MD; Ahmed Abdikadir Hussein[2], MD; Abdi Osman[2], MD; Mohamed Mohamud[1], MPH; Hasan Mowlid[1], MPH; Girum Hailu[3], MD; Fathia Alwan[3], MBA; Ermiyas Abebe Bizuneh[4], MSCE; Ahmed Mohammed Ibrahim[1], MPH; Elyas Abdulahi[4], PhD

[1]Faculty of Public Health, Institute of Health, Jigjiga University, Kebele 16, Jigjiga, Ethiopia

[2]Faculty of Medical Sciences, Institute of Health, Jigjiga University, Jigjiga, Ethiopia

[3]Health and Development Division, Intergovernmental Authority on Development, Addis Ababa, Ethiopia

[4]Research and Community Service Directorate, Jigjiga University, Jigjiga, Ethiopia

Corresponding Author:
Alinoor Mohamed Farah, MPH
Faculty of Public Health, Institute of Health, Jigjiga University, Kebele 16, Jigjiga, Ethiopia

Abstract

Background: The COVID-19 pandemic has profoundly affected the movement of people across borders in Eastern and Southern Africa. The implementation of border closures and restrictive measures has disrupted the region's economic and social dynamics. In areas where national authorities lack full control over official and unofficial land crossings, enforcing public health protocols to mitigate health risks may prove challenging.

Objective: This study aimed to assess perceived factors that influence the spread and control of COVID-19 among Somali communities living on and near ground crossings in Tog Wajaale, Somali region, Ethiopia.

Methods: A community-based cross-sectional study was conducted using a multistage sampling technique. Beliefs and perceptions of the virus's spread and control were partially adapted from the World Health Organization (WHO) resources, exploring four main perception themes: (1) perceived facilitators for the spread of the virus, (2) perceived inhibitors, (3) risk labeling, and (4) sociodemographic variables. A sample size of 634 was determined using the single proportion formula. Standardized mean scores (0 - 100) and SDs categorized perception themes, with mean differences by sociodemographic variables analyzed using ANOVA and *t* tests. Statistical significance was established with a 95% CI and a *P* value below .05. The data were analyzed using STATA version 14.1.

Results: Factors influencing COVID-19 spread and control include behavioral nonadherence and enabling environments. A total of 81.9% (439/536) did not comply with social distancing, and 92.2% (493/536) faced constraints preventing them from staying home and enabling environments. Misconceptions were prevalent, including beliefs about hot weather (358/536, 66.8%), traditional medicine (36/536, 6.7%), and religiosity (425/536, 79.3%). False assurances also contributed, such as feeling safe due to geographic distance from hot spots (76/536, 14.2%) and perceiving the virus as low-risk or exaggerated (162/536, 30.2%). Only 25.2% (135/536) followed standard precautions and 29.9% (160/536) were vaccinated. Employment, region, income, sex, education, and information sources significantly influenced behavioral nonadherence, myth prevalence, and false assurances.

Conclusions: The findings highlight the need for substantial risk communication and community engagement. Only 46.6% (250/536) of individuals adhered to precautionary measures, there was a high perception of nonadherence, and essential COVID-19 resources were lacking. Additionally, numerous misconceptions and false reassurances were noted. Understanding cross-border community behavior is crucial for developing effective, contextually appropriate strategies to mitigate COVID-19 risk in these regions.

(JMIR Form Res 2025;9:e66751) doi:10.2196/66751

KEYWORDS

COVID-19; perceptions; border-crossings; SARS-CoV-2; coronavirus; respiratory; infectious; pandemic; Somali; Ethiopia; Africa; community-based; World Health Organization; WHO; cross-sectional study; multistage sampling technique; transmission; cross-border; community engagement

XSL•FO
RenderX

Introduction

Cross-border movement has been identified as an important factor in COVID-19 transmission in Eastern and Southern Africa, and most countries in the region have restricted entry to reduce virus importation. However, the region's economy and social fabric are dependent on the cross-border movement of goods and people, and it was anticipated that border closures would have the same, if not greater, impact than COVID-19 [1,2]. As a result, countries have relaxed COVID-19 containment measures and lifted border closures, increasing the risk of the virus spreading internationally.

Under normal circumstances, a diverse range of people crosses national borders in the Eastern and Southern Africa region to participate in vital activities. Truck and boda-boda (motorcycle taxi) drivers, small- and large-scale traders, tourists, migrant workers, fisherfolk and pastoralists, refugees, returnees, border community members, and people traveling to seek health care and other social services are among those affected. These people interact with the communities they pass through on their way to and from work. Border officials, security officers, and support staff interact with travelers daily, increasing the risk of community spread [2-6].

Since the International Health Regulations 2005 came into effect in 2007, there has been growing recognition that, unlike airports and ports, ground crossings frequently represent informal passages between 2 countries with no physical structure, barriers, or borders. Furthermore, ground crossing is important for the international spread of diseases. Travelers and those who live and work on or near borders are particularly vulnerable to this threat [7].

Communities that live on and near ground crossings vary in size and density. Cross-border movement is a daily necessity for many people living in these communities for work, trade, family visits, schooling, health care services, religious activities, entertainment, and other reasons. However, health measures to control public health risks may be difficult to implement in places where national authorities are unable to fully monitor formal and informal ground crossings [7].

Given that borders have already been reopened, it is critical to understand cross-border transmission dynamics, movement patterns, and the behavior of cross-border actors to develop effective, contextually appropriate risk communication and community engagement (RCCE) strategies to mitigate COVID-19 risk. Thus, this study aimed to assess the perceived factors that influence the spread and control of COVID-19 among Somali communities living on and near ground crossings.

Methods

Study Design, Setting, and Period

A community-based cross-sectional study was conducted from December 25 to 31, 2022, in Tog Wajaale, Somali region, Ethiopia. Tog Wajaale, also known as Wajaale, is a city in the Somali region near the border with Somaliland. Tog Wajaale is the main border crossing for goods entering and leaving Somaliland, primarily from Berbera, the country's main port. Tog Wajaale is located on the border between Ethiopia and Somaliland. It is 70 km east of Jigjiga, the capital of the Somali region.

Populations

All adult populations older than 18 years were included in the source population. The study participants were all adult populations older than 18 years who had lived in the Tog Wajaale town for at least 6 months.

Eligibility Criteria

The study included individuals aged 18 years and older who had resided in the study area for a minimum of 6 months. Those who were experiencing mental illness or severe health issues during the study period were excluded from participation.

Sample Size Determination

The sample size was obtained using the single proportions formula with the specified parameters of 50% (P=.50, as there is no comparable study in Ethiopia), 95% CI ($Z_1 - \alpha/2$) 1.96%, and 5% margin of error (d). Adding a 10% nonresponse rate and a design effect of 1.5 resulted in a required sample size of 634.

Sampling Technique

A multistage sampling technique is used to select participants for the study. In the first stage, 1 city council in the region, based on the border with the neighboring country and demographic diversity, is selected. In the second stage, 5 kebeles (subdistricts) are randomly selected, and the total sample size is allocated by probability proportional to size (PPS) to each kebele based on the number of registered households. In the third stage, using the list of households as the sampling frame, a systematic random sampling technique is used to select the study participants from the kebeles.

Data Collection and Quality Control

Data were collected using a pretested structured questionnaire. Ten BSc nurses fluent in the local language collected data through face-to-face interviews at the household level. Beliefs and perceptions of the virus's spread and control were partially adapted from the World Health Organization (WHO) resources. Specifically, some of the questions were adapted from the COVID-19 rapid quantitative assessment tool [8,9]. Three main perception themes were explored: perceived facilitators for the spread of the virus (6 items), perceived inhibitors (7 items), and risk labeling (8 items), as well as sociodemographic variables, including access to communication resources. Prior to participation, the household head or any individual over 18 years of age was fully informed of the study details and asked for consent. Questionnaires were reviewed daily for completeness, and respondents were contacted to see if any information was missing. The investigators regularly supervised the data collection process. To ensure data quality, ongoing supervision and follow-ups were conducted throughout the collection period.

Data Analysis

KoBoToolbox, an open-source data collection platform, was used to collect data. KoBo-collected responses were initially

XSL•FO
RenderX

encoded in a Microsoft Excel database and then exported to STATA version 14.1 for analysis. The frequency tables were used to present the respondents' background variables as well as specific belief items. Standardized mean scores (0 - 100) and SDs were used to describe the lists of factor categories based on the perception themes. To compare mean differences by sociodemographic variables, 1-way ANOVA and *t* tests were used. A factor loading score of 0.4 was used as a cutoff value to retain items in each category [10]. Exploratory factor analysis (EFA) categorized perceptions into 4 themes: facilitators of virus transmission (6 factors), inhibitors (7 factors), risk identification (8 factors), and sociodemographic factors, including communication resource access. The sample's adequacy for EFA was confirmed by the Kaiser-Mayer-Olkin measure (KMO>50%). Statistical significance was determined using a 95% CI and a *P* value below .05.

Operation Definition and Variables

The *perceived factors* influencing the spread and control of COVID-19 include exacerbating factors, inhibiting factors, and risk labeling.

Perceived COVID-19 Exacerbating Factors

Behavioral Nonadherence

Noncompliance with public health measures designed to control the spread of COVID-19 encompasses various behaviors exhibited by individuals or groups. These actions diverge from the recommended guidelines and may include disregarding social distancing protocols, neglecting to use face masks, refusing COVID-19 vaccination, or displaying discriminatory attitudes toward others in relation to the pandemic.

Lack of Enabling Environment

The absence of essential conditions that support the implementation of public health measures to prevent COVID-19 transmission. For example, economic and social constraints that force people to leave their homes for work or social issues; overcrowded living and working conditions that may not allow physical distancing; lack of access to personal protective equipment such as face masks; and insufficient availability of clean water, soap, hand rub alcohol, or sanitizers.

Perceived COVID-19 Inhibiting Factors

False Assurance

Misconceptions or preconceptions that diminish the perceived threat of COVID-19 frequently result in a relaxed compliance with preventive protocols. Examples include the belief that one's location is far removed from epidemic centers or the underestimation of the seriousness of the illness.

Myths

Misinformed or culturally rooted beliefs that influence public perception of COVID-19 and its risks often undermine scientific evidence. Examples include beliefs that hot weather kills the virus, reliance on traditional medicine as a cure, or attributing protection to religious practices alone.

Standard Precautions

The actions taken by individuals to minimize the risk of COVID-19 transmission were consistent with the recommended public health guidelines. These include wearing masks, maintaining hand hygiene, avoiding crowded areas, and vaccinating against COVID-19.

Risk Labeling

Risk labeling involves perceiving a specific sect of the community to be at a high risk of contracting COVID-19, especially older adults and people with underlying medical conditions.

Ethical Considerations

This study was conducted in accordance with the guidelines of the Declaration of Helsinki, and all procedures involving research study participants were approved by the Jigjiga University institutional review board (RERC/052/2022). Written informed consent was obtained from all the participants. Participants were not compensated for participating in the study. The study participants' data are confidential and protected.

Results

Sociodemographic Characteristics of Participants

The survey questionnaire was completed by 536 participants, resulting in a response rate of 84.5% (536/634). Table 1 summarizes the demographic information of the survey respondents. The majority of respondents (439/536, 81.9%) were younger than 40 years and male (313/536, 58.4%). Most of the respondents (153/536, 28.5%) had a high school education, followed by those who could read and write but with no formal education (114/536, 21.3%). The majority of participants in the study were married (335/536, 62.5%), ran their own businesses (261/536, 48.7%), were Muslim (501/536, 93.5%), and were of Somali origin (423/536, 78.9%). Furthermore, the majority of participants (287/536, 53.5%) used cellular data as their primary internet source, had an income greater than 5000 ETB (Ethiopian Birr; US $93) (324/536, 60.4%), and their primary source of information about COVID-19 was television (230/536, 42.9%).

XSL•FO
RenderX

Table . Sociodemographic characteristics of the respondents (N=536).

Characteristics		Values, n (%)
Age (years)		
	<40	439 (81.9)
	≥40	97 (18.1)
Sex		
	Male	313 (58.4)
	Female	223 (41.6)
Education		
	Cannot read and write	10 (1.9)
	Read and write	114 (21.3)
	Primary (grade 1 - 8)	78 (14.6)
	High school (grade 9 - 12)	153 (28.5)
	Diploma or level I-IV	104 (19.4)
	First degree (BSc or BA) and above	77 (14.4)
Marital status		
	Single (unmarried)	174 (32.5)
	Married	335 (62.5)
	Divorced	13 (2.4)
	Widowed	4 (0.7)
	Engaged (in relationship)	10 (1.9)
Employment		
	Private or business	261 (48.7)
	Daily laborer	75 (14)
	Government employee	80 (14.9)
	Student	34 (6.3)
	No job	86 (16)
Religion		
	Muslim	501 (93.5)
	Orthodox	27 (5)
	Protestant	8 (1.5)
Place of origin		
	Somali region	423 (78.9)
	Other regions	113 (21.1)
Internet connection mostly used to get information about COVID-19		
	Cellular data (telle)	287 (53.5)
	Wireless Wi-Fi	46 (8.6)
	Do not use internet	203 (37.9)
Monthly income		
	<5000 ETB[a] (US $93)	212 (39.6)
	≥5000 ETB (US $93)	324 (60.4)
Main source of information about COVID-19		
	Television	230 (42.9)
	Radio	16 (3)

XSL•FO
RenderX

Characteristics		Values, n (%)
	Friends and neighbors	58 (10.8)
	Social media (Facebook, Twitter, YouTube, etc)	196 (36.6)
	Health workers	36 (6.7)
Ever used the call center for information about COVID-19		
	Yes	22 (4.1)
	No	514 (95.9)

[a]ETB: Ethiopian Birr.

Perceived COVID-19 Exacerbating and Inhibiting Factors

Table 2 shows the prevalence of specific factors in the exacerbating and inhibiting categories. The prevalence of specific factors associated with behavioral nonadherence ranged from 81.9% (439/536) to 32.5% (174/536). The fact that people are still not adhering to physical or social distance accounted for a greater proportion of nonadherence. The magnitude of the absence of enablers that would support behavioral adherence ranged from 92.2% (494/536) to 5.4% (29/536). The fact that people could not stay home for economic and social reasons and living or working in overcrowded conditions contributed significantly to the lack of enablers. Environmental factors were perceived as more deterrent than behavioral nonadherence, indicating a strong tendency for externalizing factors that could exacerbate the spread of COVID-19 in the community.

XSL•FO
RenderX

Table . Perceived COVID-19 exacerbating and inhibiting factors (N=536).

Perceived factors			Descriptive statistics, n (%)
Perceived COVID-19 exacerbating factors			
	Behavioral nonadherence		
		People fear stigma and bias related to COVID-19	205 (38.3)
		People still do not respect social distance, ie, use crowded transportation means, still hug and shake each other's hands while greeting	439 (81.9)
		People refused to take COVID-19 vaccine	174 (32.5)
	Lack of enabling environment		
		People do not stay at home for economic and social reasons	494 (92.2)
		People still live and work in a very crowded condition	403 (75.2)
		People do not have access to face masks, water, and sanitizers	29 (5.4)
Perceived COVID-19 inhibiting factors			
	Myths		
		The weather we live-in is too hot for coronavirus to survive	358 (66.8)
		We believe we have traditional medicine against COVID-19	36 (6.7)
		We are religious enough to control COVID-19	425 (79.3)
	False assurances		
		We live far away from COVID-19's hot spot areas	76 (14.2)
		COVID-19 is not risky or exaggerated	162 (30.2)
	Engaged in precaution		
		Engaged in standard precaution measures of COVID-19	135 (25.2)
		Vaccinated against COVID-19	160 (29.9)

The most common myths were perceived as protectiveness from hot weather and religiosity, accounting for 66.8% (358/536) and 79.3% (425/536), respectively. Beliefs that the specific localities where respondents are currently living are far away from coronavirus hot spot areas and that COVID-19 is not dangerous or exaggerated contributed to 14.2% (76/536) and 30.2% (162/536) of the false assurances, respectively. On the other hand, the prevalence of people believing that the spread of COVID-19 would be controlled because of their active participation in standard precautionary measures and the COVID-19 vaccine were 25.2% (135/536) and 29.9% (160/536), respectively.

Perceived Categories of COVID-19 Exacerbating and Inhibiting Factors

Table 3 presents the categories and lists of perceived COVID-19 exacerbations and inhibitors, along with their respective prevalence. EFA identified 2 major categories of perceived exacerbating and inhibiting COVID-19 factors in the target population. The first category of factors was classified as behavioral nonadherence, indicating noncompliance with recommended precautionary measures. Individuals engaging in handshaking, using crowded transportation, refusing COVID-19 vaccination, and exhibiting a factor loading score of 0.714 collectively contribute to behavioral nonadherence.

The lack of enabling environmental conditions, which are supposed to support the adaptation of precautionary measures, was the second category of exacerbating facilitating factors. With factor loading scores of 0.71, the lack of enablers was made up of economic reasons that challenged the stay-at-home principle, overcrowded living or working conditions, and a lack of personal protective equipment such as face masks and sanitizers. Behavioral nonadherence and lack of enabler factors

XSL•FO
RenderX

explained 28.4% (152/536) of the variance in perceived virus facilitators.

Table . Perceived categories of COVID-19 exacerbating and inhibiting factors. Kaiser-Mayer-Olkin measure of sampling adequacy (KMO=53%) and degrees of freedom (*df*=10).

Perceived factors about COVID-19	Principal components and factor loading scores		Descriptive statistics (N=536)	
	Exacerbating factors	Inhibiting factors	n (%)	95% CI
Behavioral nonadherence[a]	0.71	—[b]	335 (62.5)	58.3-66.5
Lack of enabling environment[c]	0.71	—	494 (92.2)	89.6-94.2
Myths[d]	—	0.70	425 (79.3)	75.6-82.5
False assurances[e]	—	0.32	216 (40.3)	36.2-44.5
Engaged in precautions[f]	—	0.71	250 (46.6)	42.4-50.9

[a]Behavioral nonadherence: People fear stigma and bias related to COVID-19, people still use crowded transportation means, people with flu-like symptoms are not well screened for COVID-19, people refuse to take the COVID-19 vaccine, and people still hug and shake each other's hands while greeting.

[b]Not applicable.

[c]Lack of enabling environment: People do not stay at home for economic and social reasons, people still live and work in a very crowded condition, people do not have personal protective equipment like face masks, and people do not have access to water or hand rub alcohol or sanitizers.

[d]Myths: The weather we live in is too hot for coronavirus to survive, we believe we have traditional medicine against COVID-19, and we are religious enough to control COVID-19.

[e]False assurances: We live far away from COVID-19's rampant areas and COVID-19 is not risky or exaggerated.

[f]Engaged in precautions: Engaged in standard precaution measures for COVID-19 and vaccinated against COVID-19.

The table also includes categories and lists of perceived COVID-19 inhibitors as well as their respective prevalence. EFA produced 3 major categories of perceived factors that inhibited COVID-19. Two of the three categories were incorrectly identified (myths and false assurances), whereas one was correctly identified as an inhibitor. The myths category included factors that are thought to inhibit the virus but have not been scientifically proven. In this case, the myths included perceived religiosity (perceiving oneself as an effective religious man or woman in dealing with challenges), people's perceived confidence in owning effective traditional medicines that were not clinically confirmed, and living in hot weather, with a factor loading score of 0.70.

The second category of perceived inhibitors, which were frequently associated with false assurances that people were immune to COVID-19. With factor loading scores of 0.32, the category consisted of 2 main beliefs: we live far away from COVID-19 hot spot areas and there have been no locally reported COVID-19 cases so far. The beliefs appeared to be false assurances in that people perceive themselves to be living outside a risk zone, giving the impression of invulnerability.

With a factor score loading of 0.71, the third, promotable category of perceived inhibitors concerned people who had taken standard precautions and were vaccinated against COVID-19. Factors related to the 3 categories explained 20.1% (108/532) of the variance in perceived inhibitors, indicating the presence of several other unreported myths and unhealthy beliefs that require further investigation.

Risk Labeling

The COVID-19 risk labels and groups are presented in Table 4. People with underlying illnesses (118/536, 22%) and older adults (351/536, 65.5%) were perceived by the community to be at a high risk of COVID-19.

Table . Perceived COVID-19 risk groups and labels (N=536).

Perceived high-risk groups	Descriptive statistics, n (%)
Children (aged 0 - 9 years)	38 (7.1)
Adolescents (aged 10-15 years)	3 (0.6)
Youths (aged 16 - 29 years)	2 (0.4)
Adults (aged 30 - 50 years)	11 (2.1)
Older persons (aged 51 years or more)	351 (65.5)
Pregnant women	9 (1.7)
Health workers	4 (0.7)
People with underlying chronic illness (eg, diabetes, hypertension, and cancer)	118 (22)

XSL•FO
RenderX

Distribution of Perceived Exacerbating Factors

Table 5 illustrates the relationships between sociodemographic characteristics, behavioral nonadherence, and lack of an enabling environment in relation to COVID-19. The mean percentages of behavioral nonadherence and lack of an enabling environment as facilitators of virus spread were both high, at 62.5% (335/536) and 92.2% (494/536), respectively.

XSL•FO
RenderX

Table . Distribution of perceived exacerbating factors among the study participants.

Variables	Behavioral nonadherence[a]				Lack of enabling environment[b]			
	% Mean (SE)	*t* test (*df*=534)	*F* test (*df*=535)	*P* value	% Mean (SE)	*t* test (*df*=534)	*F* test (*df*=535)	*P* value
Overall	62.5 (2.1)				92.2 (1.2)			
Sex		–1.20		.12		2.78		.003[c]
Male	60.4 (2.8)				94.9 (1.2)			
Female	65.5 (3.2)				88.3 (2.2)			
Age (years)		–0.39		.35		0.51		.31
<40	62.1 (2.3)				92.4 (1.5)			
≥40	64.4 (5.2)				90.8 (3.1)			
Education			1.37	.23			3.55	.003[c]
Cannot read and write	50.0 (15.3)				80.0 (8.4)			
Read and write	63.2 (4.5)				86.8 (2.5)			
Primary (grade 1 - 8)	55.1 (5.5)				87.2 (3.0)			
High school (grade 9 - 12)	58.8 (3.9)				92.8 (2.1)			
Diploma or level I-IV	70.2 (4.7)				97.1 (2.6)			
First degree (BSc or BA) and above	67.5 (5.5)				98.7 (3.0)			
Marital status			0.24	.92			4.12	.003[c]
Single (unmarried)	63.2 (3.7)				94.8 (2.0)			
Married	62.1 (2.7)				90.7 (1.5)			
Divorced	53.8 (13.5)				92.3 (7.5)			
Widowed	75.0 (24.3)				75.0 (13.4)			
Engaged	70.0 (15.5)				__[d]			
Employment			3.33	.01[c]			0.75	.56
Private or business	64.4 (2.9)				90.8 (1.7)			
Daily laborer	48.0 (5.5)				90.7 (3.1)			
Government employee	60.0 (5.4)				96.3 (3.1)			
Student	82.4 (8.2)				94.1 (4.6)			
No job	64.0 (5.2)				93.0 (2.9)			
Religion			0.34	.71			0.35	.71
Muslim	62.9 (2.2)				92.0 (1.4)			
Orthodox	59.3 (9.3)				92.6 (6.2)			
Protestant	50.0 (17.2)				—			
Place of origin		1.89		.03[c]		–1.52		.06
Somali region	64.5 (2.3)				91.3 (1.4)			

XSL•FO
RenderX

Variables	Behavioral nonadherence[a]				Lack of enabling environment[b]			
	% Mean (SE)	*t* test (*df*=534)	*F* test (*df*=535)	*P* value	% Mean (SE)	*t* test (*df*=534)	*F* test (*df*=535)	*P* value
Other regions	54.9 (4.7)				95.6 (1.9)			
Internet connection mostly used to get information about COVID-19			1.03	.36			7.68	<.001[c]
Cellular data (telle)	65.2 (2.9)				94.8 (1.6)			
Wireless Wi-Fi	56.5 (7.1)				—			
Do not use internet	60.1 (3.4)				86.7 (1.8)			
Monthly income		–2.84		.002[c]		0.2		.42
<5000 ETB[e] (US $93)	55.2 (3.4)				92.4 (1.8)			
≥5000 ETB (US $93)	67.6 (2.6)				91.9 (1.5)			
Source of information			1.31	.25			6.84	<.001
Official websites	50.0 (34.1)				—			
Television	60.1 (3.2)				92.1 (1.7)			
Radio	50.0 (12.1)				93.8 (6.5)			
Friends or neighbors	55.2 (6.3)				75.8 (3.4)			
Social media (Facebook, Twitter, YouTube, etc)	68.9 (3.5)				97.9 (1.9)			
Health workers	61.1 (8.1)				86.1 (4.4)			

[a]Behavioral nonadherence: People fear stigma and bias related to COVID-19, people still use crowded transportation means, people with flu-like symptoms are not well screened for COVID-19, people refuse to take the COVID-19 vaccine, and people still hug and shake each other's hands while greeting.

[b]Lack of enabling environment: People do not stay at home for economic and social reasons, people still live and work in a very crowded condition, people do not have personal protective equipment like face masks, and people do not have access to water or hand rub alcohol or sanitizers.

[c]Statistically significant at *P*<.05 (two-tailed).

[d]Not applicable.

[e]ETB: Ethiopian Birr.

In terms of behavioral nonadherence, respondents' employment, region, and income showed significant mean percentage differences. Students had the highest mean percentage of behavioral nonadherence (442/536, 82.4%; *P*=.01) when it came to employment. Respondents of Somali origin and those with an income of more than 5000 ETB (US $93) had the highest mean percentages of behavioral nonadherence (346/536, 64.6%; *P*=.03) and (362/536, 67.5%; *P*=.003), respectively. There was no difference in behavioral nonadherence by age, gender, education, marital status, religion, primary internet source, or primary information source for COVID-19.

Similarly, the lack of an enabling environment that was perceived as a facilitator of the spread of the virus was more prevalent among male respondents (509/536, 95%; *P*=.003), respondents whose education status was first degree and above (529/536, 98.7%; *P*=.003), respondents who were single and not married (508/536, 94.8%; *P*=.001), those who used cellular data as their primary source of the internet (508/536, 94.8%; *P*=.001), and those who used social media as their primary source of information (525/536, 97.9%; *P*<.001). There was no difference in behavioral nonadherence by age, employment, religion, place of origin, and monthly income.

XSL•FO
RenderX

Distribution of Perceived Inhibiting Factors

The factors deemed to hinder the transmission and control of the virus exhibited considerable sociodemographic disparities (Table 6). The highest perceived inhibitor to control the spread of COVID-19 was seen to be myth (425/536, 79.3%), followed by preventive measures (250/536, 46.6%) and false guarantees (214/536, 39.9%).

XSL•FO
RenderX

Table . Distribution of perceived inhibiting factors among the study participants.

Variables	Myth[a]				False assurance[b]				Engaged in precaution[c]			
	% Mean (SE)	*t* test (*df*=534)	*F* test (*df*=535)	*P* value	% Mean (SE)	*t* test (*df*=534)	*F* test (*df*=535)	*P* value	% Mean (SE)	*t* test (*df*=534)	*F* test (*df*=535)	*P* value
Overall	79.3 (1.8)				40.3 (2.1)				46.6 (2.2)			
Sex		2.79		.003[d]		0.33		.37		–0.17		.43
Male	83.4 (2.1)				40.9 (2.8)				46.3 (2.8)			
Female	73.4 (3.0)				39.4 (3.3)				47.1 (3.4)			
Age (years)		–0.58		.28		–0.70		.76		0.73		.23
<40	78.8 (1.9)				39.6 (2.3)				47.0 (2.3)			
≥40	81.6 (4.2)				43.7 (5.3)				44.8 (5.4)			
Education			7.37	<.001[d]			2.12	.06			3.72	.003[d]
Cannot read and write	90.0 (12.5)				50.0 (15.4)				60.0 (15.6)			
Read and write	79.8 (3.7)				46.5 (4.6)				40.4 (4.6)			
Primary (grade 1 - 8)	60.3 (4.5)				30.8 (5.5)				47.4 (5.6)			
High school (grade 9 - 12)	74.5 (3.2)				46.4 (3.9)				36.6 (4.0)			
Diploma or level I-IV	88.5 (3.9)				37.5 (4.8)				58.7 (4.8)			
First degree (BSc or BA) and above	93.5 (4.5)				31.2 (5.6)				57.1 (5.6)			
Marital status			1.42	.23			0.8	.52			0.94	.44
Single (unmarried)	83.3 (3.1)				36.8 (3.7)				47.7 (3.8)			
Married	77.6 (2.2)				41.8 (2.7)				45.7 (2.7)			
Divorced	76.9 (11.2)				38.4 (13.6)				61.5 (13.9)			
Widowed	__[e]				75.0 (24.6)				75.0 (25.0)			
Engaged (in relationship)	60.0 (12.8)				40.0 (15.5)				30.0 (15.8)			
Employment			2.51	.041[d]			4.81	<.001[d]			0.47	.76

XSL•FO
RenderX

Variables	Myth[a]				False assurance[b]				Engaged in precaution[c]			
	% Mean (SE)	*t* test (*df*=534)	*F* test (*df*=535)	*P* value	% Mean (SE)	*t* test (*df*=534)	*F* test (*df*=535)	*P* value	% Mean (SE)	*t* test (*df*=534)	*F* test (*df*=535)	*P* value
Private or business	82.0 (2.5)				49.0 (3.0)				45.2 (3.1)			
Daily laborer	66.7 (4.7)				30.7 (5.6)				48.0 (5.8)			
Government employee	83.8 (4.5)				38.8 (5.4)				43.8 (5.6)			
Student	82.4 (6.9)				32.4 (8.3)				55.9 (8.6)			
No job	76.7 (4.3)				26.7 (5.2)				48.8 (5.4)			
Religion			0.21	.81			0.95	.39			0.17	.85
Muslim	79.0 (1.8)				39.7 (2.2)				46.9 (2.2)			
Orthodox	81.5 (7.8)				44.4 (9.4)				44.4 (9.6)			
Protestant	87.5 (14.4)				62.5 (17.4)				37.5 (17.7)			
Place of origin			0.68	.25		–1.82		.03[d]		0.99		.16
Somali region	79.9 (2.0)				38.3 (2.4)				47.8 (2.4)			
Other regions[f]	77.5 (4.0)				47.8 (4.7)				42.5 (4.7)			
Internet connection mostly used to get information about COVID-19			0.46	.63			5.37	.005[d]			4.1	.02[d]
Cellular data (telle)	78.7 (2.9)				36.9 (2.9)				49.8 (2.9)			
Wireless Wi-Fi	84.8 (7.2)				26.1 (7.2)				58.7 (7.3)			
Do not use internet	78.8 (3.4)				48.3 (3.4)				39.4 (3.5)			
Monthly income		–0.02		.49		3.74		<.001[d]		–4.29		<.001[d]
<5000 ETB[g] (US $93)	79.2 (2.8)				50.0 (3.4)				35.4 (3.3)			
≥5000 ETB (US $93)	79.3 (2.3)				33.9 (2.6)				54.0 (2.8)			

XSL•FO
RenderX

Variables	Myth[a]				False assurance[b]				Engaged in precaution[c]			
	% Mean (SE)	*t* test (*df*=534)	*F* test (*df*=535)	*P* value	% Mean (SE)	*t* test (*df*=534)	*F* test (*df*=535)	*P* value	% Mean (SE)	*t* test (*df*=534)	*F* test (*df*=535)	*P* value
Source of information			2.32	.042[d]			3.18	.008[d]			2.75	.02[d]
Official website	—				50.0 (34.4)				50.0 (35.0)			
Television	77.2 (2.7)				36.8 (3.2)				45.6 (3.3)			
Radio	56.3 (10.1)				31.3 (12.1)				18.8 (12.4)			
Friends and neighbors	74.1 (5.3)				62.1 (6.4)				32.8 (6.5)			
Social media[h]	85.2 (2.9)				36.7 (3.5)				53.6 (3.5)			
Health workers	77.8 (6.7)				50.0 (8.1)				50.0 (8.3)			

[a]Myths: The weather we live in is too hot for coronavirus to survive, we believe we have traditional medicine against COVID-19, and we are religious enough to control COVID-19.

[b]False assurances: We live far away from COVID-19's rampant areas, and COVID-19 is not risky or exaggerated.

[c]Engaged in precautions: Engaged in standard precautions for COVID-19 and vaccinated against COVID-19.

[d]Statistically significant at *P*<.05 (two-tailed).

[e]Not applicable.

[f]Other regions include Addis Ababa, Afar, Amhara, Oromia, southern nations and nationalities, and Tigray.

[g]ETB: Ethiopian Birr.

[h]Facebook, Twitter, YouTube, etc.

Male respondents (447/536, 83.4%; *P*=.003), those with a first degree or higher (501/536, 93.5%; *P*<.001), government employees (449/536, 83.8%; *P*=.041), and those who used social media as their primary source of information (457/536, 85.3%; *P*=.042) had a significantly higher prevalence of myths. There was no difference in myth prevalence by age, marital status, religion, place of origin, primary internet source, or monthly income.

False assurances were also more common among those who owned or ran a private business (263/536, 49.1%; *P*=.001), non-Somalis (256/536, 47.8%; *P*=.03), noninternet users (259/536, 48.3%; *P*=.005), those with monthly income less than 5000 ETB (US $93; 268/536, 50%; *P*=.001), and those who get COVID-19 information from friends and neighbors (333/536, 62.1%; *P*<.001). There was no significant variation in the prevalence of false assurance by age, sex, education, marital status, religion, employment, or primary information source for COVID-19.

Likewise, participants who were unable to read and write, those who used Wi-Fi as their primary source of internet, those who earned more than 5000 ETB (US $93), and those who used social media as their primary source of COVID-19 information were more likely to take precautionary measures than their counterparts. There was no significant variation in the prevalence of precautionary measures according to age, sex, marital status, religion, and employment.

Discussion

Principal Findings

This study produced relevant findings regarding community perceptions of the factors that facilitate and inhibit the spread of COVID-19, risk labeling, and information needs in a border town in the Somali region. The perceived factors were classified into the following major categories: behavioral adherence, a lack of enabling environmental conditions, and risk labeling.

Factors perceived to exacerbate virus spread were grouped into 2 thematic categories: behavioral nonadherence (335/536, 62.5%) and a lack of enabling environmental conditions (494/536, 92.2%). The study revealed a concerning lack of awareness and adherence with 80% (429/536) of respondents noting that individuals continue to disregard social distancing. Additionally, 38.1% (204/536) of those surveyed indicated that people fear the stigma associated with the virus, while 32.5% (174/536) expressed unwillingness to receive the COVID-19 vaccine. These results highlight a significant gap in the understanding of and compliance with preventive measures.

The use of crowded transportation modes was not only due to a lack of alternatives but also associated with behavioral

XSL•FO
RenderX

nonadherence. Fears, threat appraisals, and worries have consistently been linked to increased adherence in COVID-19 protective behaviors [11,12]. An Iranian study found that increased fear predicted less adherence [13]. This study shows that clear communication of risk could increase participation in protective behaviors, especially among disengaged groups.

Conversely, people could not stay at home for economic and social reasons (494/536, 92.2%); they did not have access to face masks, water, or sanitizers (29/536, 5.4%); and they still lived and worked in crowded conditions (403/536, 75.2%). According to behavioral and communication theories, people's perceived lack of resources has a negative impact on their actual practices [14]. Despite this, the high prevalence of perceived facilitators indicates 2 major and urgent issues. First, it indicates a strong effort to reduce behavioral nonadherence and a lack of enablers that facilitate viral spread. Second, even a higher perceived lack of enabling conditions appears concerning because it may lead people to externalize their ability to control the virus while ignoring their own efforts.

Poverty and livelihood issues are major challenges in COVID-19 prevention in low-income countries. This is because COVID-19 has impacted people's livelihoods and income opportunities, resulting in food insecurity [15]. Most people in Tog Wajaale are low-income earners who work in the informal sector and rely on daily wages to feed their families and cover their related expenses. As most of them struggle to earn a living, COVID-19 has added to their difficulties. As a result, they are confronted with the harsh reality of contracting the virus while attempting to survive. The observation results show that despite the lockdown, daily laborers in the construction sector, informal businesses, and street vendors continued their daily activities [16].

Factors perceived as inhibitors of viral spread were divided into 3 categories: false assurances (216/536, 40.3%), myths (425/536, 79.3%), and standard precautions (250/536, 46.6%). False assurances were perceptions of invulnerability, characterized by people believing that they were living outside the COVID-19 risk zone. The 2 main false assurances in this study were the perception of COVID-19 as less risky or exaggerated and residence far away from the COVID-19 rampant areas. Myths in this study included religious effectiveness (425/536, 79.3%), living in hot weather (358/536, 66.8%), and traditional medicine (36/536, 6.7%). The WHO myth busters list the majority of the misconceptions presented in this study, indicating that they were universally held during the pandemic [17]. Evidence suggests that myths or misperceptions, such as denial of presence and misperceived causes, transmission modes, and prevention methods, can prevent and control efforts during pandemics of HIV, Zika virus, yellow fever, and Ebola unless traced and addressed [18-21].

Culture and religion are important components of Somalis' social and moral fabrics. As a result, cultural and religious misunderstandings as well as misinformation about the virus are among the factors contributing to complacency when it comes to prevention measures. According to a study conducted in northern Ethiopia, people believe that COVID-19 is a "punishment from God" or "Allah" for people's immoral acts, and that the power of prayer can prevent the pandemic's severe effects [16].

The magnitude of correctly perceived factor engagement in standard precautionary measures (134/536, 25%) and vaccination against COVID-19 (160/536, 29.9%) was too low, necessitating hard work to promote this perception, until a larger segment of the community embraced an accurate reason for virus protection.

Respondents' employment, region, and income all showed significant mean percentage differences in behavioral nonadherence. When it came to employment, students had the highest mean percentage of behavioral nonadherence (442/536, 82.5%). Respondents of Somali origin and those with an income of more than 5000 ETB (US $93) had the highest mean percentages of behavioral nonadherence (64.6%, 346/536 and 67.5%, 362/536, respectively). Students were most likely to be of younger age, and being younger was found to be a predicator of nonadherence to preventive behavior [22]. Regular communication regarding SARS-CoV-2 emphasizes that COVID-19 mostly threatens the older population. This could send a false message of safety to the younger generations. Younger individuals' lower adherence may be related to their perceived lower vulnerability and propensity to make impulsive decisions, take greater risks to gain emotional and social stimulation (such as attending social gatherings), and fail to recognize the long-term effects of their actions [23].

Similarly, almost 60.1% (322/536) of the communities surveyed are reliant on cross-border commerce and daily labor, and 20% (107/536) of them live on less than US $100 per month. Thus, they are driven to leave their homes, even during the COVID-19 epidemic, to sustain their families. Within the Somali population, males typically serve as primary income earners and feel obligated to venture outside their homes, even during the COVID-19 pandemic, to earn money to sustain their families. This observation aligns with a study conducted in Somalia, which also noted that Somali men tend to disregard COVID-19 preventive protocols [24].

Obtaining COVID-19 information from official sources (radio, television, and government announcements) increased the likelihood of having a high adherence score, whereas obtaining it from social media decreased the likelihood of having a high adherence score. This highlights the role of misinformation in determining adherence to scientifically proven COVID-19 prevention measures [25]. In addition, a previous study showed that social media users tend to be more extroverted, suggesting that these people may find it more challenging to stay at home and observe COVID-19 preventive measures than introverted non–social media users [26].

Limitations

First, the cross-sectional design does not provide conclusive conclusions about causality, and social desirability bias may skew the self-reported data. Second, only one border crossing town in parts of the Somali region was surveyed, and the results of this study may not reflect the perceptions of other towns in the Somali region and Ethiopia. A similar study could be

XSL•FO
RenderX

extended to other border-crossing towns in the Somali region and Ethiopia.

Recommendations

There is a need for local activities that encourage participation in standard precautions by addressing locally perceived hurdles and fostering a sense of shared responsibility and community ownership during the fight against COVID-19. Consequently, RCCE efforts in Tog Wajaale should (1) consider sociodemographic variations in myths and false assurances, (2) investigate more beliefs that could facilitate or inhibit the spread of the virus, and (3) design local initiatives that enhance community ownership of tasks associated with controlling the virus, thus supporting and advocating engagement in standard precautionary measures. That is, using community and religious leaders to spread correct information, model precautionary behaviors, and organize COVID-19–related activities.

Conclusions

The results suggest that extensive efforts in RCCE are necessary. This conclusion is supported by several factors. First, only 44.6% (239/536) of individuals correctly engaged in precautionary measures. Second, there was a high perceived prevalence of behavioral nonadherence and a shortage of essential resources in the COVID-19 response. Lastly, numerous misconceptions and false reassurances were identified among the population. Thus, understanding the behavior of cross-border communities is crucial for developing effective, contextually appropriate RCCE strategies to mitigate COVID-19 risk in border communities.

Acknowledgments

The authors wish to thank the district health office and health facility staff for their support and facilitation of the data collection. This study was funded by the Intergovernmental Authority on Development (IGAD). The views expressed herein are solely those of the authors and do not reflect the views of IGAD or any other stakeholders. All authors declared that they had insufficient funding to support open access publication of this manuscript, including from affiliated organizations or institutions, funding agencies, or other organizations. JMIR Publications provided article processing fee (APF) support for the publication of this article.

Data Availability

The dataset for this study is available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: AMF, AA, AH, EA
Formal analysis: AMF
Writing – original draft: AMF
Writing – review & editing: AMF, AA, AH, AAH, AO, MM, HM, GH, FA, EAB, AMI, EA

Conflicts of Interest

None declared.

References

1. Jones L, Schmidt-Sane M. Covid-19 RCCE strategies for cross-border movement in Eastern and Southern Africa. : Social Science in Humanitarian Action Platform; 2020 URL: https://www.socialscienceinaction.org/resources/covid-19-rcce-strategies-for-cross-border-movement-in-eastern-and-southern-africa/ [accessed 2025-02-20]
2. Bouet A, Laborde Debucquet D. COVID-19 border policies create problems for African trade and economic pain for communities. In: COVID-19 and Global Food Security: International Food Policy Research Institute; 2020:69-72.
3. Storer E, Pearson G. Cross-border dynamics and healthcare in West Nile, Uganda. UNICEF, IDS, and Anthrologica. 2019 URL: https://opendocs.ids.ac.uk/articles/online_resource/Cross-Border_Dynamics_and_Healthcare_in_West_Nile_Uganda/26433706?file=48083425 [accessed 2025-02-20]
4. Baluku M, Akello G, Parker M, Grant C. How the ‘disease of the radio’ is the next burden impacting the lives of those on the Uganda-DRC border. : Social Science in Humanitarian Action Platform; 2020 Apr 8 URL: https://www.socialscienceinaction.org/blogs-and-news/covid-19-uganda-disease-radio-next-burden-impacting-lives-uganda-drc-border/ [accessed 2025-02-20]
5. Nshimbi CC. The human side of regions: informal cross-border traders in the Zambia–Malawi–Mozambique growth triangle and prospects for integrating Southern Africa. J Borderl Stud 2020 Jan 1;35(1):75-97. [doi: 10.1080/08865655.2017.1390689]
6. Rutherford B, Addison L. Zimbabwean farm workers in Northern South Africa. Rev Afr Polit Econ 2007 Jan;34(114):619-635. [doi: 10.1080/03056240701819491]
7. Controlling the spread of COVID-19 at ground crossings: interim guidance, 20 May 2020. : World Health Organization; 2020 URL: https://iris.who.int/handle/10665/332165 [accessed 2025-02-20]

XSL•FO
RenderX

8. Risk communication and community engagement (RCCE) action plan guidance COVID-19 preparedness and response. : World Health Organization; 2020 Mar 16 URL: https://www.who.int/publications/i/item/risk-communication-and-community-engagement-(rcce)-action-plan-guidance [accessed 2025-02-20]
9. COVID-19: vulnerable and high risk groups 2020. World Health Organization. URL: https://www.who.int/westernpacific/emergencies/covid-19/information/high-risk-groups [accessed 2025-02-20]
10. Katchova A. Principal component analysis. Econometrics Academy. 2017. URL: https://sites.google.com/site/econometricsacademy/econometrics-models/principal-component-analysis [accessed 2025-02-20]
11. Brouard S, Vasilopoulos P, Becher M. Sociodemographic and psychological correlates of compliance with the COVID-19 public health measures in France. Can J Pol Sci 2020 Dec;53(2):253-258. [doi: 10.1017/S0008423920000335]
12. Barber SJ, Kim H. COVID-19 worries and behavior changes in older and younger men and women. J Gerontol B 2021 Jan 18;76(2):e17-e23. [doi: 10.1093/geronb/gbaa068]
13. Ezati Rad R, Mohseni S, Kamalzadeh Takhti H, et al. Application of the protection motivation theory for predicting COVID-19 preventive behaviors in Hormozgan, Iran: a cross-sectional study. BMC Public Health 2021 Mar 8;21(1):466. [doi: 10.1186/s12889-021-10500-w] [Medline: 33685426]
14. Glanz K, Rimer BK, Viswanath K. Health Behavior and Health Education: Theory, Research, and Practice: John Wiley & Sons; 2008.
15. Kassegn A, Endris E. Review on socio-economic impacts of 'triple threats' of COVID-19, desert locusts, and floods in East Africa: evidence from Ethiopia. Cogent Social Sciences 2021 Jan 1;7(1):1885122. [doi: 10.1080/23311886.2021.1885122]
16. Nigussie H. The coronavirus intervention in Ethiopia and the challenges for implementation. Front Commun 2021;6:562512. [doi: 10.3389/fcomm.2021.562512]
17. Coronavirus disease (COVID-19) advice for the public: mythbusters. World Health Organization. 2020. URL: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/myth-busters [accessed 2025-02-20]
18. Tenkorang EY. Myths and misconceptions about HIV transmission in Ghana: what are the drivers? Cult Health Sex 2013 Mar;15(3):296-310. [doi: 10.1080/13691058.2012.752107]
19. Carey JM, Chi V, Flynn DJ, Nyhan B, Zeitzoff T. The effects of corrective information about disease epidemics and outbreaks: evidence from Zika and yellow fever in Brazil. Sci Adv 2020 Jan;6(5):eaaw7449. [doi: 10.1126/sciadv.aaw7449] [Medline: 32064329]
20. Buli BG, Mayigane LN, Oketta JF, et al. Misconceptions about Ebola seriously affect the prevention efforts: KAP related to Ebola prevention and treatment in Kouroussa Prefecture, Guinea. Pan Afr Med J 2015;22(Suppl 1):11. [doi: 10.11694/pamj.supp.2015.22.1.6269] [Medline: 26740839]
21. Odlum M, Yoon S. Health information needs and health seeking behavior during the 2014-2016 Ebola outbreak: a Twitter content analysis. PLoS Curr 2018 Mar 23;10. [doi: 10.1371/currents.outbreaks.fa814fb2bec36e29b718ab6af66124fa] [Medline: 29707416]
22. Urbán R, Király O, Demetrovics Z. Who complies with coronavirus disease 2019 precautions and who does not? Curr Opin Psychiatry 2021 Jul 1;34(4):363-368. [doi: 10.1097/YCO.0000000000000723] [Medline: 34001700]
23. Casey BJ, Getz S, Galvan A. The adolescent brain. Dev Rev 2008;28(1):62-77. [doi: 10.1016/j.dr.2007.08.003] [Medline: 18688292]
24. Ahmed MAM, Siewe Fodjo JN, Gele AA, et al. COVID-19 in Somalia: adherence to preventive measures and evolution of the disease burden. Pathogens 2020 Sep 6;9(9):735. [doi: 10.3390/pathogens9090735] [Medline: 32899931]
25. Ahinkorah BO, Ameyaw EK, Hagan JE, Seidu AA, Schack T. Rising above misinformation or fake news in Africa: another strategy to control COVID-19 spread. Front Commun 2020;5:45. [doi: 10.3389/fcomm.2020.00045]
26. Correa T, Hinsley AW, de Zúñiga HG. Who interacts on the web?: The intersection of users' personality and social media use. Comput Human Behav 2010:247-253. [doi: 10.1016/j.chb.2009.09.003]

Abbreviations

EFA: exploratory factor analysis
ETB: Ethiopian Birr
KMO: Kaiser-Mayer-Olkin measure
PPS: probability proportional to size
RCCE: risk communication and community engagement
WHO: World Health Organization

XSL•FO
RenderX

Edited by A Mavragani; submitted 22.09.24; peer-reviewed by R Ibrahim; revised version received 10.12.24; accepted 11.12.24; published 24.04.25.

Please cite as:
Farah AM, Abdulahi A, Hussein A, Hussein AA, Osman A, Mohamud M, Mowlid H, Hailu G, Alwan F, Bizuneh EA, Ibrahim AM, Abdulahi E
COVID-19 Perceptions Among Communities Living on Ground Crossings of Somali Region of Ethiopia: Community Cross-Sectional Survey Study
JMIR Form Res 2025;9:e66751
URL: https://formative.jmir.org/2025/1/e66751
doi:10.2196/66751

© Alinoor Mohamed Farah, Abdifatah Abdulahi, Abdulahi Hussein, Ahmed Abdikadir Hussein, Abdi Osman, Mohamed Mohamud, Hasan Mowlid, Girum Hailu, Fathia Alwan, Ermiyas Abebe Bizuneh, Ahmed Mohammed Ibrahim, Elyas Abdulahi. Originally published in JMIR Formative Research (https://formative.jmir.org), 24.4.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Synthetic Data-Driven Approaches for Chinese Medical Abstract Sentence Classification: Computational Study

Jiajia Li[1,2,3*], PhD; Zikai Wang[1,2,4*], PhD; Longxuan Yu[5], MSc; Hui Liu[6], PhD; Haitao Song[1,2], PhD

[1]Shanghai Artificial Intelligence Research Institute Co., Ltd, Shanghai, China
[2]Xiangfu Laboratory, Jiaxing, China
[3]School of Chemistry and Chemical Engineering, Shanghai Jiao Tong University, Shanghai, China
[4]Inner Mongolia Academy of Science and Technology, Hohhot, China
[5]University of California San Diego, San Diego, CA, United States
[6]Shanghai Civil Aviation College, Shanghai, China
[*]these authors contributed equally

Corresponding Author:
Haitao Song, PhD
Shanghai Artificial Intelligence Research Institute Co., Ltd, Shanghai, China

Abstract

Background: Medical abstract sentence classification is crucial for enhancing medical database searches, literature reviews, and generating new abstracts. However, Chinese medical abstract classification research is hindered by a lack of suitable datasets. Given the vastness of Chinese medical literature and the unique value of traditional Chinese medicine, precise classification of these abstracts is vital for advancing global medical research.

Objective: This study aims to address the data scarcity issue by generating a large volume of labeled Chinese abstract sentences without manual annotation, thereby creating new training datasets. Additionally, we seek to develop more accurate text classification algorithms to improve the precision of Chinese medical abstract classification.

Methods: We developed 3 training datasets (dataset #1, dataset #2, and dataset #3) and a test dataset to evaluate our model. Dataset #1 contains 15,000 abstract sentences translated from the PubMed dataset into Chinese. Datasets #2 and #3, each with 15,000 sentences, were generated using GPT-3.5 from 40,000 Chinese medical abstracts in the CSL database. Dataset #2 used titles and keywords for pseudolabeling, while dataset #3 aligned abstracts with category labels. The test dataset includes 87,000 sentences from 20,000 abstracts. We used SBERT embeddings for deeper semantic analysis and evaluated our model using clustering (SBERT-DocSCAN) and supervised methods (SBERT-MEC). Extensive ablation studies and feature analyses were conducted to validate the model's effectiveness and robustness.

Results: Our experiments involved training both clustering and supervised models on the 3 datasets, followed by comprehensive evaluation using the test dataset. The outcomes demonstrated that our models outperformed the baseline metrics. Specifically, when trained on dataset #1, the SBERT-DocSCAN model registered an impressive accuracy and F_1-score of 89.85% on the test dataset. Concurrently, the SBERT-MEC algorithm exhibited comparable performance with an accuracy of 89.38% and an identical F_1-score. Training on dataset #2 yielded similarly positive results for the SBERT-DocSCAN model, achieving an accuracy and F_1-score of 89.83%, while the SBERT-MEC algorithm recorded an accuracy of 86.73% and an F_1-score of 86.51%. Notably, training with dataset #3 allowed the SBERT-DocSCAN model to attain the best with an accuracy and F_1-score of 91.30%, whereas the SBERT-MEC algorithm also showed robust performance, obtaining an accuracy of 90.39% and an F_1-score of 90.35%. Ablation analysis highlighted the critical role of integrated features and methodologies in improving classification efficiency.

Conclusions: Our approach addresses the challenge of limited datasets for Chinese medical abstract classification by generating novel datasets. The deployment of SBERT-DocSCAN and SBERT-MEC models significantly enhances the precision of classifying Chinese medical abstracts, even when using synthetic datasets with pseudolabels.

(JMIR Form Res 2025;9:e54803) doi:10.2196/54803

KEYWORDS

medical abstract sentence classification; large language models; synthetic datasets; deep learning; Chinese medical; dataset; traditional Chinese medicine; global medical research; algorithm; robustness; efficiency; accuracy

XSL•FO
RenderX

Introduction

Background

In the realms of machine learning and artificial intelligence, the significance of data cannot be overstated, yet accessing authentic real-world datasets often presents substantial hurdles, including elevated costs, extended timeframes, and privacy issues. To navigate these obstacles, there is a growing pivot toward the utilization of synthetic datasets. While synthetic datasets have predominantly been associated with computer vision applications, the landscape is changing for natural language processing (NLP) due to advancements in text generation capabilities, largely attributed to the multihead self-attention mechanism integral to the transformer family of models. This trend began with the transformer itself [1] and has since given rise to models such as Bidirectional Encoder Representations from Transformers (BERT) [2], OpenAI GPT [3], Transformer-XL [4], OpenAI GPT-2 [5], and Grover [6]. In recent years, large language models have transformed the field of NLP, demonstrating exceptional performance in a wide range of tasks. These models generate increasingly coherent text, prompting researchers to explore the potential of synthetic datasets. Particularly noteworthy, Zellers et al [6] claimed that their Grover model for conditional text generation outperforms human-generated text in both style and content, especially in the “fake news” and “propaganda” categories, as evaluated by human raters. This highlights the promise of synthetic datasets to improve NLP performance and the continued need to develop text generation techniques.

Prior Work

Sentence-Level Text Classification

Text classification at the sentence level has been explored using various deep learning models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs). Kim [7] first proposed a single-layer CNN with pretrained word embeddings for text classification, achieving excellent results. Yang et al [8] developed a 2-level attention mechanism using gate recurrent units for document classification, while Conneau et al [9] introduced a CNN-nonstatic model with character-level CNNs and an average pooling layer. Recently, pretrained models, such as BERT [2], RoBERTa [10], XLNet [11], and ALBERT [12], have also been used for sentence-level text classification tasks. These transformer-based models generate sentence representations and can be combined with various classifiers to achieve state-of-the-art performance. Additionally, hybrid approaches using word-level and character-level CNNs initialized with ELMo [13] or BERT embeddings have been explored to improve the robustness and performance of sentence-level text classification models. Overall, pretrained models have significantly advanced the state-of-the-art in sentence-level text classification, and further research in this area is expected to yield even more sophisticated models.

Unsupervised Text Clustering

Unsupervised text clustering is an important task in NLP that groups similar text documents without relying on labeled examples. While traditional methods such as hierarchical agglomerative clustering [14], k-means clustering [15], nonnegative matrix factorization [16], and latent dirichlet allocation [17] have been widely used for this task, recent advances in pretrained embeddings have led to new and competitive methods. These include paragraph vector [18] and USE+KMeans [19], which have shown promising results in various text clustering benchmarks. More recently, BERT-based methods, such as BERT-EMD [20], have been proposed for unsupervised text clustering. Additionally, SBERT [21] and DocSCAN [22] are the most recent methods that generate high-quality text embeddings and use graph-based clustering for unsupervised document clustering. Together, these developments highlight the continued importance of unsupervised text clustering in NLP research, with recent methods based on pretrained embeddings, and SBERT in particular, showing promising results.

Synthetic Data

Synthetic data, generated using various techniques such as generative adversarial network (GAN)-based text generation and language model-based data augmentation, has become a popular way to expand small datasets in NLP. These synthetic datasets aim to improve model performance and generalization by providing additional examples for training. For instance, Xu et al [23] used a GAN-based method to generate synthetic data for image captioning, which showed promising results. Similarly, Wang and Wan [24] expanded GAN-based text generation to create synthetic datasets for sentiment analysis, which performed similarly to real data. Additionally, language model–based data augmentation, such as using GPT-2 for auto-completion, has been effective in generating synthetic data for NLP tasks [5]. Zhang et al [25] improved Chinese text classification using a language model–based data augmentation technique, while Zhou et al [26] showed that their language model–based data augmentation method improved low-resource language modeling. However, the use of synthetic data in NLP also has its limitations and potential drawbacks, which require further research. It is important to explore the advantages and disadvantages of synthetic datasets, as well as the different techniques used to generate them, to fully understand their impact on model performance and generalization.

The Goal of This Study

This paper aims to explore the creation and use of synthetic datasets to address the lack of real-world datasets for Chinese medical abstract classification, as there are currently no open-source datasets with sentence-level classification labels for Chinese medical abstracts, as shown in Figure 1. In response to this challenge, we have harnessed the capabilities of GPT-3.5 to generate distinct datasets for the classification of Chinese medical abstracts. Moreover, we designed and trained the clustering and supervised models for this task. Through this innovative approach, this study not only showcases the immense potential of synthetic datasets to bridge the gaps inherent in real-world datasets but also illuminates their profound impact on enhancing the performance of NLP tasks.

XSL•FO
RenderX

Figure 1. Goal of this study. It has been observed that Chinese medical abstract classification faces a significant number of limitations due to the lack of corresponding labeled datasets. Achieving accurate sentence-level classification in these abstracts will be instrumental in contributing Chinese medical information to the global medical field.

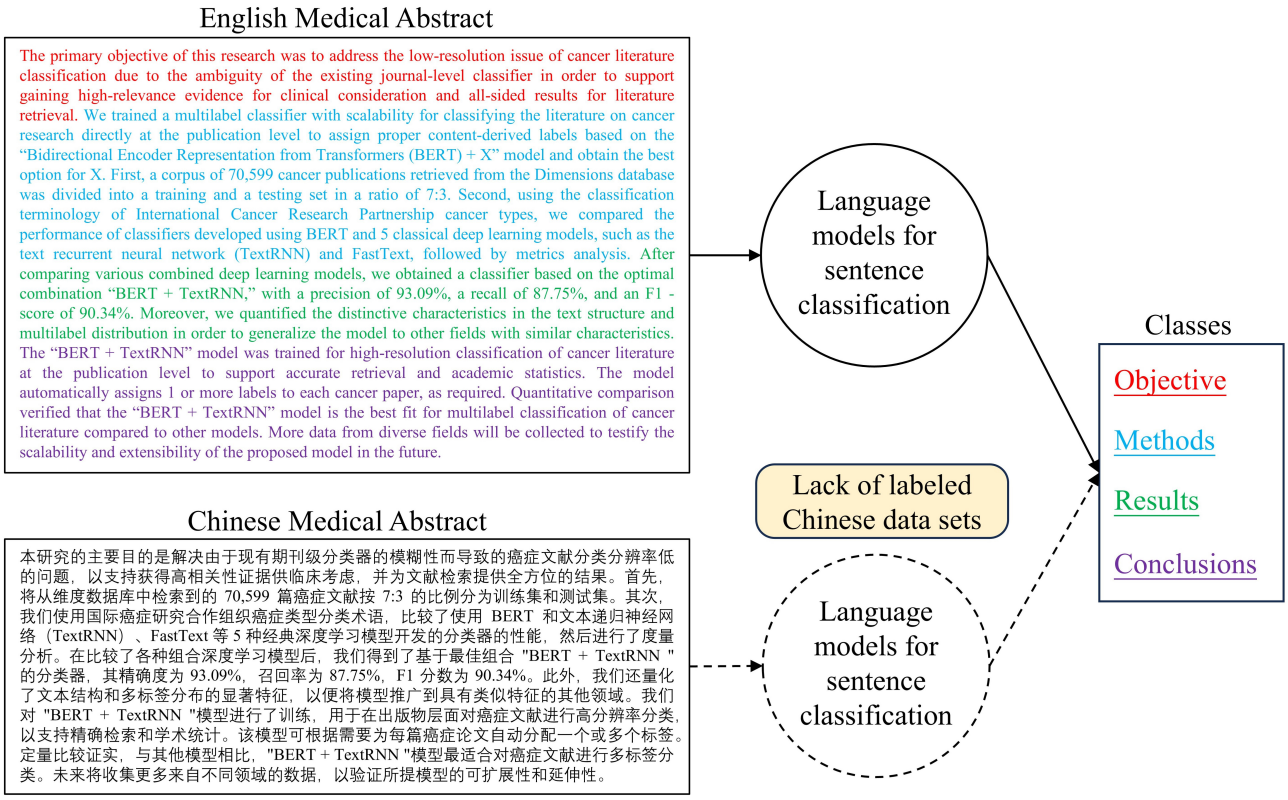

Methods

Overall Framework

The methodology used in this paper is structured into three distinct sections, systematically addressing the creation and application of synthetic datasets for Chinese medical abstract classification:

1. Synthetic dataset generation: the process begins by translating the PubMed dataset [27] into Chinese and generating 2 distinct synthetic abstract datasets based on provided titles, keywords, disciplines, and categories by GPT-3.5.
2. Unsupervised clustering: this step involves fine-tuning the SBERT model with the synthetic datasets and then using the DocSCAN algorithm to cluster the generated sentence embeddings in an unsupervised manner, termed SBERT-DocSCAN (Figure 2).
3. Supervised classification: a new supervised method, SBERT-MEC, is proposed to classify the synthetic data set, enhancing the ability to accurately categorize synthetic medical abstracts (Figure 2).

XSL•FO
RenderX

Figure 2. Framework of the proposed methods for abstract classification. The unsupervised clustering method (SBERT-DocSCAN) leverages SBERT embeddings and DocSCAN to group sentences into abstract sections based on clustering. The supervised learning method (SBERT-MEC) uses a multiencoder cascade (MEC) to enhance classification accuracy by extracting and integrating features for abstract sentence classification.

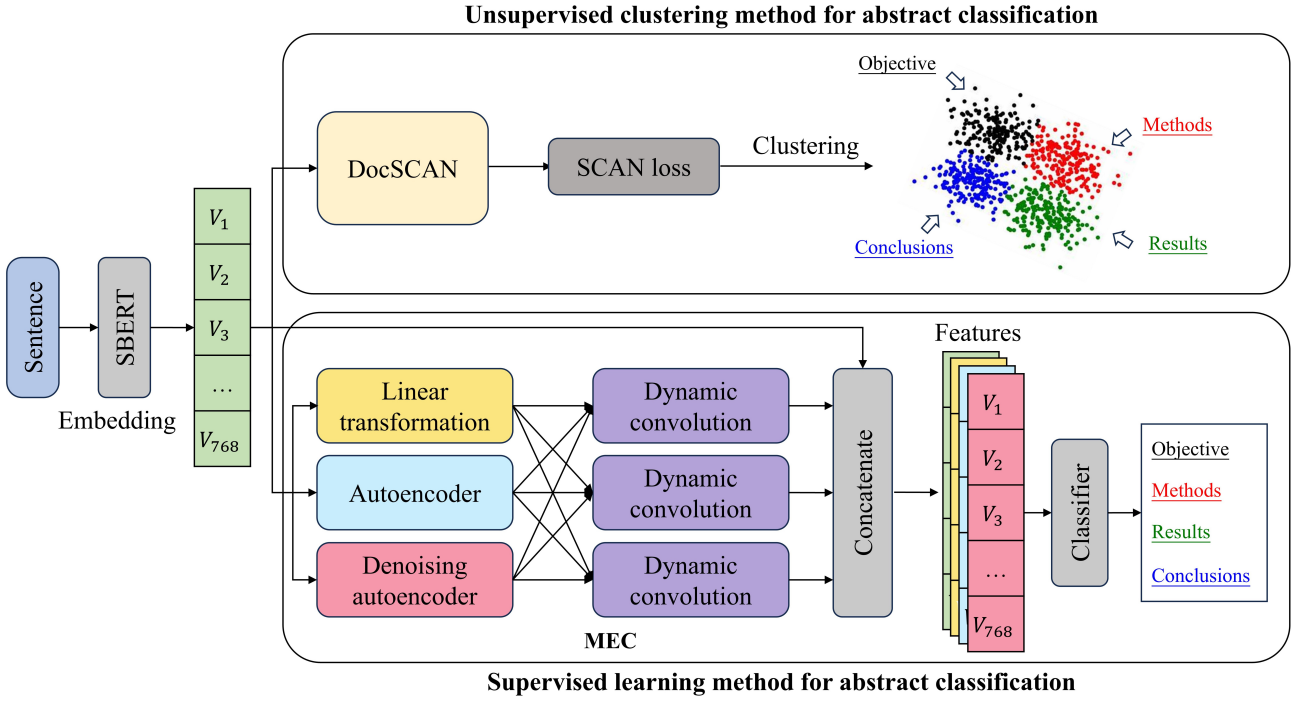

Synthetic Dataset Generation

As there is currently no Chinese dataset for sentence-level classification in the medical abstract field, in this step, we used OpenAI's text generation model, GPT-3.5, to generate 3 types of small synthetic datasets, which only contain around 15,000 sentences each yet still perform well on classification tasks. The first one is the translated PubMed dataset, which is translated into Chinese from the PubMed 200k RCT dataset by DeepL [28], and we choose 15,000 sentences with clear labels as dataset #1. The second dataset is created by using GPT-3.5 to generate abstracts based on title, keywords, discipline, and category from the CSL dataset [29]. The third dataset is also generated using GPT-3.5, but instead, it is paraphrased with rewritten abstracts assigned pseudolabels as dissimilar to the original text as possible.

Generating a diverse training dataset using large language models is a challenge. Even when the generation temperature is set to a high value, these models still tend to generate highly repetitive datasets that lack the diversity required for effective language model training. To address this issue, we selected the CSL corpus as the base corpus, which contains over 40,000 Chinese medical abstracts. Although these abstracts are not annotated at the sentence level, 20,000 of them have clear structure division into the 4 parts mentioned above. We extracted these 20,000 abstracts and manually labeled them as the test dataset. For the remaining 20,000 abstracts without clear structure division, we removed some low-quality data and the corpus of abstracts, then extracted a portion of them as the input corpus for GPT-3.5, which includes corresponding titles, keywords, subjects, and categories. These 4 types of data were inputted into GPT-3.5 to generate abstracts in 4 parts: purpose, method, results, and conclusion. These abstracts were then cleaned to produce a dataset of 15,000 sentences with pseudolabels (dataset #2). Similarly, the structure-less abstracts were inputted into GPT-3.5, which rewrote them to generate clear structure abstracts in 4 parts, resulting in a dataset of 15,000 sentences with pseudolabels (dataset #3). The construction process of the above datasets is shown in Figure 3.

XSL•FO
RenderX

Figure 3. Construction process of the synthetic datasets used for sentence-level classification in the medical abstract field. Dataset #1 is derived by translating the PubMed 200k RCT dataset into Chinese using DeepL and selecting 15,000 sentences with clear labels. Dataset #2 is generated by GPT-3.5, using titles, keywords, disciplines, and categories from the CSL dataset to produce structured abstracts divided into 4 parts (objective, methods, results, and conclusions). Dataset #3 is created by paraphrasing structure-less abstracts using GPT-3.5, generating diverse and rewritten abstracts with pseudolabels. The test dataset comprises 20,000 manually labeled abstracts from the CSL corpus with clear structural divisions. For each of the models in the paper (eg, SBERT), we train them on the training sets separately to produce corresponding ones (eg, SBERT #1, SBERT #2, and SBERT #3).

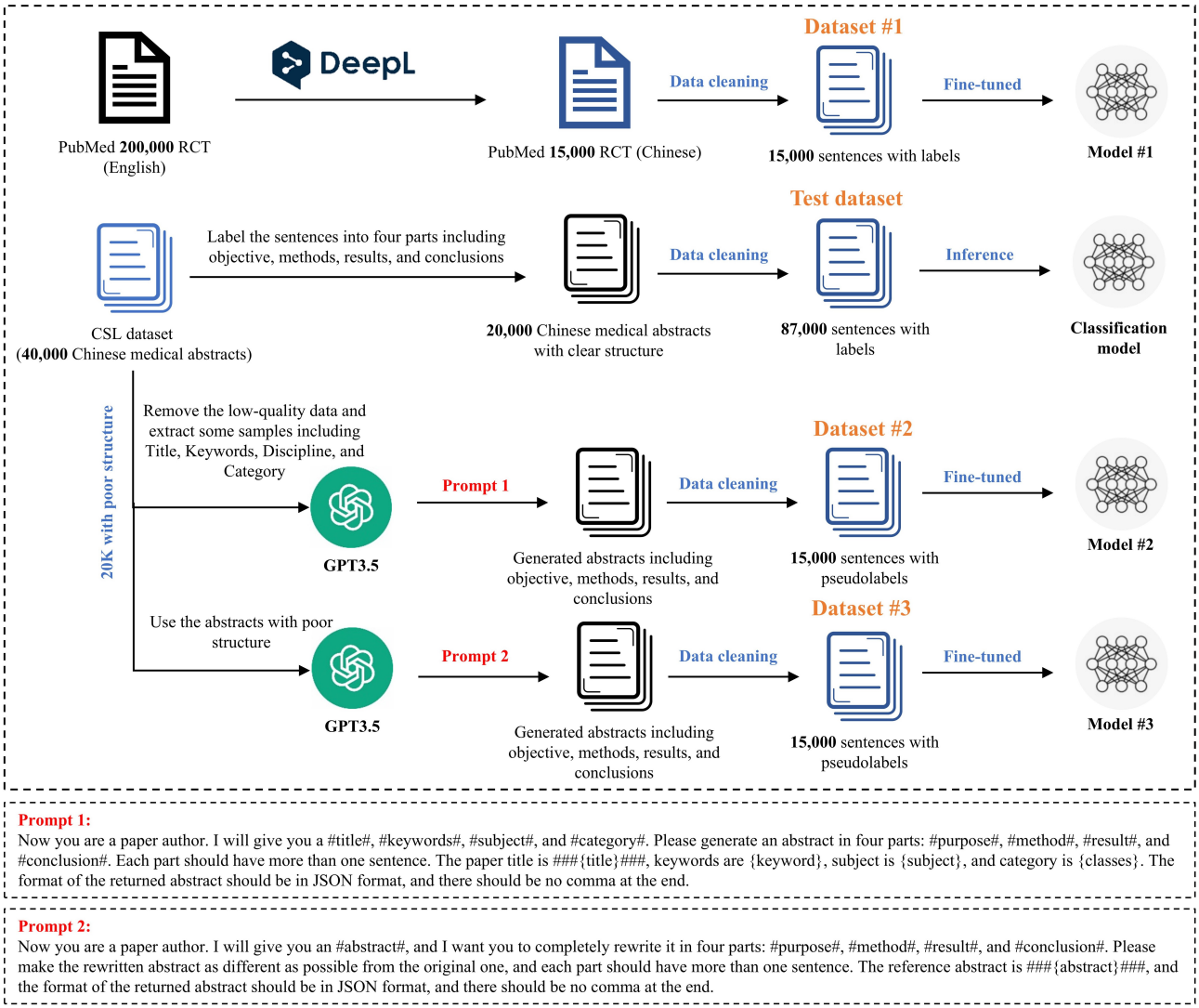

Unsupervised Clustering

In the second part, we fine-tuned the sentence transformer and then used the DocSCAN algorithm to cluster the synthetic datasets. We chose the sbert-chinese-general-v2 model, which is a model pretrained on the SimCLUE dataset [30], as the base model due to its outstanding performance on embedding Chinese sentences. We fine-tuned the pretrained model on the 3 datasets into 3 models, which is quite crucial as it leads to the overall distribution more inclined toward the functional aspects of each sentence rather than the literal content. We then used these fine-tuned models to embed the input data, followed by clustering by the DocSCAN algorithm which is an algorithm relying on the intuition that a datapoint and its nearest neighbors in representation space often share the same class label, and outperform others on unsupervised classification tasks. Specifically, in this part, we generated embeddings for each sentence in the test set using the 3 fine-tuned sentence transformers in the first part and then used the DocSCAN algorithm to cluster these embeddings for unsupervised abstract classification, which uses the SCAN loss LS to fine-tune the model f(x), defined as:

$$LS=-1|D|\sum x\in D\sum k\in Nx\log(f(x)\cdot f(k))+\lambda\sum i\in Cpi\log(pi)$$

which can be broken down into a consistency loss and an auxiliary loss. The consistency loss aims to ensure that the same label is assigned to a data point and its neighbors. Our model f(·) computes a label for a given data point x from the dataset *D*, as well as for each data point *k* in the set of mined neighbors from *x* in Nx. To achieve this, we calculate the dot product between the output distribution (normalized by softmax(·) of datapoint x and its neighbor k. The auxiliary loss is used to achieve regularization via entropy, scaled by a weight λ. This loss encourages the model to distribute probability mass across all clusters C, where pi denotes the assigned probability of cluster i in C by the model. Without this term, there could be a shortcut where all examples collapse into a single cluster. The entropy term ensures that the distribution of class labels resulting from applying DocSCAN tends to be roughly uniform.

XSL•FO
RenderX

Supervised Classification

In the third part, we develop a supervised learning method, SBERT-MEC, which leverages the proposed multiencoder cascade (MEC) module for feature extraction, designed to enhance the classification of synthetic datasets through refined weak supervision. Figure 2 shows the framework of SBERT-MEC and the modeling process. The first step of the SBERT-MEC is to input the sentence embeddings generated by SBERT into the MEC module, which includes 3 encoders: linear transformation, autoencoder model, and denoising autoencoder model. Formally, given a 768-dimensional sentence embedding V, the linear transformation V transforms the sentence embedding V into W by adding a small random number n:

$$W=LV=V1+n,V2+n,\ldots,V768+n$$

The autoencoder model can be represented as follows:

$$W=AE\theta(V)$$

where AE is a fully connected network that maps V into a duplicate embedding W and θ is the parameters of AE. The autoencoder is trained by duplicate pairs from the Quora dataset [31]. The denoising autoencoder (DA), in another way, adds Gaussian noise to the sentence embedding and forms a noisy embedding VN, and the training process forces output W to be as close to the original embedding V as possible:

$$VN=V+\ ,\ \sim N0,\sigma 2I$$

$$W = DA\theta(VN)$$

To further enhance performance, 3 embeddings are input into the dynamic convolutional layer for mutual supervision. Specifically, for each embedding, the embeddings generated by the other 2 encoders are concatenated together and used as inputs to supervise this embedding. Finally, 3 output embeddings of the dynamic convolutional layers are concatenated with the original embedding and then put into a classifier to predict test data. This process is repeated for the 3 models fine-tuned in the first part respectively.

To summarize, first, we create synthetic datasets by different methods and leverage them to fine-tune sentence transformers. These datasets are small in size while effective in training classification tasks. Second, fine-tuned models are used on an unsupervised classification algorithm DocSCAN to analyze the performance of these datasets. Further, fine-tuned models are used on a weakly supervised algorithm SBERT-MEC proposed by us, which performs data augmentation on embeddings. This process demonstrates the performances of small synthetic datasets again, as well as the effectiveness of our algorithm.

Ethical Considerations

This study did not involve human participants, animal subjects, or personally identifiable data. All data used in the research were either synthetic or publicly available, ensuring compliance with ethical guidelines. As such, no ethical approval or informed consent was required. The study adheres to principles of academic integrity and research transparency.

Results

Dataset

For empirical evaluation, we crafted 3 datasets: 2 synthetic ones generated using GPT-3.5, and a pre-existing dataset. Each synthetic dataset comprises approximately 15,000 sentences pertinent to medical abstracts. These datasets underwent a meticulous generation process, including cleaning and pseudolabel assignment for sentences within each abstract section. Detailed methodology for creating the synthetic datasets is outlined in the section on Synthetic Dataset Generation. Table 1 illustrates the distribution of sentences across various sections within each dataset.

Table . Distribution of development datasets. Dataset #1 (translated PubMed dataset with labeled sentences), dataset #2 (GPT-3.5-generated abstracts with pseudolabels), and dataset #3 (paraphrased GPT-3.5-generated abstracts). Each dataset contains 15,000 sentences, divided into 4 sections: objective, methods, results, and conclusion.

Dataset	Classes of sentences				Total
	Objective	Methods	Results	Conclusion	
Dataset #1	3750	3750	3750	3750	15,000
Dataset #2	3750	3750	3750	3750	15,000
Dataset #3	3537	4140	3826	3497	15,000

Baselines

We selected multiple algorithms of different kinds as baselines to demonstrate the performance of our algorithms. To begin with, we used classical non-BERT models. TextCNN is built upon the CNN paradigm, applying multiple convolutional filters and pooling operations to capture local features of varying lengths within the text. TextRNN [32] leverages RNNs to model the sequential information in the text, capturing dependencies between words. TextRNN-Att [33] extends TextRNN by incorporating attention mechanisms, enabling the model to focus on essential words. TextRCNN [34] combines convolutional and RNNs, simultaneously considering word order and contextual information. DPCNN [35] uses multiple layers of convolution and pooling operations to capture hierarchical text features. FastText [36] is a simple and efficient text classification method based on the bag-of-words model, representing text at the word level.

Additionally, we used BERT-based methods. BERT is a transformer-based pretrained language model, capable of learning rich contextual representations. BERTCNN [37] and BERT-RNN [8] integrate CNNs and RNNs, respectively, on top of BERT, combining the contextual representations from BERT with the ability to extract local features. BERT-RCNN

XSL•FO
RenderX

[38] combines BERT with both convolutional and RNNs, simultaneously considering word order and contextual information. BERT-DPCNN [39] incorporates a deep pyramid CNN on top of BERT, combining the contextual representations from BERT with multilevel feature extraction capabilities. ERNIE [40] is another transformer-based pretrained language model, built upon BERT with further improvements and optimizations, with better performance in Chinese contexts.

In our ablation study, we conducted an experiment using the k-means algorithm to cluster SBERT embeddings for comparative analysis with DocSCAN. Additionally, we examined the impact of omitting the dynamic convolutional layer in the MEC module. Subsequent experiments involving BERT serve as an extension of this analysis, further exploring the effects of removing 3 encoders within MEC.

Experiment Settings

In this study, we used the PyTorch framework and Python for all experiments. Data processing and model training were performed on an NVIDIA GeForce RTX 3090 (24G) GPU and an Intel(R) Core(TM) i9-10900K CPU @3.70GHz. Experimental parameters are detailed in Table 2. To mitigate overfitting, a dropout rate of 0.5 was used, randomly deactivating 50% of neuron connections during training. We conducted training over 30 epochs using either the Adam or BertAdam optimization algorithms, with a learning rate set at 5e-5. Training batches comprised 64 samples each, and the RNN featured a hidden size of 256. The CNN component used filter sizes of (2, 3, 4), with a total of 256 filters. Input sequence length was capped at 64 tokens, and the feature dimension was maintained at 768. Pretrained models used in the experiments included BERT (bert-base-chinese), ERNIE (ernie,-3.0-base-zh), and SBERT (sbert-chinese-general-v2), which were critical to the study's execution and result analysis.

Table . Settings for training models.

Parameter	Setting
Dropout	0.5
Epoch number	30
Optimizer	Adam/BertAdam
Learning rate	e−5
Batch size	64
RNN[a] hidden size	[2]
CNN[b] filter number	256
Max length	64
Feature dimension	768
BERT[c]	Bert-base-chinese
ERNIE	Ernie-3.0-base-zh
SBERT	Sbert-chinese-general-v2

[a]RNN: recurrent neural network.

[b]CNN: convolutional neural network.

[c]BERT: Bidirectional Encoder Representations from Transformers.

Evaluation Metrics

The experiments used accuracy and F_1-score to evaluate the model performance. The formulas for each index are as follows:

Accuracy= TP+TNTP+TN+FP+FN

where TP (true positive) corresponds to the number of instances correctly predicted as positive; TN (true negative) corresponds to the number of instances correctly predicted as negative; FP (false positive) corresponds to the number of instances incorrectly predicted as positive; FN (false negative) corresponds to the number of instances incorrectly predicted as negative.

Precision= TPTP+FP

Recall= TPTP+FN

F1-score= 2×Precision×RecallPrecision+Recall

In the above formulations, the Precision measures the proportion of correctly predicted positive instances out of the total instances predicted as positive and Recall measures the proportion of correctly predicted positive instances out of the total actual positive instances. F1-score is the harmonic mean of precision and recall, providing a single value that considers both metrics. This technique is highly effective for imbalanced datasets, where disparities in class sizes can skew model performance. It helps ensure fair representation of all classes, enhancing model reliability.

Model Performance

Table 3 shows the performance of various algorithms across different training datasets when evaluated on the test dataset. Notably, when trained on dataset #1, the SBERT-DocSCAN algorithm emerges as the leading performer, securing an

XSL•FO
RenderX

accuracy and F_1-score of 0.8985 on the test dataset. This standout performance highlights the algorithm's capability to classify medical domain data with high precision. Additionally, the SBERT-MEC algorithm also displays comparable performance on the same dataset, with an accuracy and F_1-score of 0.8938, making it the second most effective algorithm in our evaluation. For Data se t#2, the SBERT-DocSCAN and SBERT-MEC algorithms again demonstrate superior performance. SBERT-DocSCAN leads with exceptional accuracy and F_1-score of 0.8983, reinforcing its effectiveness in managing generated data. Meanwhile, SBERT-MEC remains a strong contender with an accuracy and F_1-score of 0.8673, marking it as the second most proficient algorithm for this dataset. Furthermore, considering dataset #3, SBERT-DocSCAN and SBERT-MEC algorithms consistently demonstrate outstanding performance. SBERT-DocSCAN achieves the highest accuracy and F_1-score of 0.9130, affirming its effectiveness in handling paraphrased data. Similarly, the SBERT-MEC algorithm achieves the second-highest accuracy and F_1-score of 0.9039 and 0.9035, respectively, highlighting its competence in dealing with paraphrased data. Additionally, Table 4 illustrates the confusion matrices of the SBERT-DocSCAN model on the test dataset using different training datasets.

In summary, the comprehensive analysis of the table elucidates that the SBERT-DocSCAN and SBERT-MEC algorithms consistently outperform other algorithms across multiple datasets. The superior performance of these algorithms in terms of accuracy and F1 scores underscores their significance and efficacy in the domain of text classification.

Table . Comparative performance of various models on the test dataset[a].

Method	Training on dataset #1		Training on dataset #2		Training on dataset #3	
	Accuracy	F_1-score	Accuracy	F_1-score	Accuracy	F_1-score
TextCNN	0.6122	0.6074	0.7843	0.7795	0.6789	0.6755
TextRNN	0.5045	0.5413	0.7315	0.7321	0.7295	0.7239
TextRNN-Att	0.6439	0.6421	0.7114	0.6955	0.7521	0.7521
TextRCNN	0.6562	0.6543	0.7642	0.7643	0.7439	0.7392
DPCNN	0.6953	0.6967	0.6947	0.6988	0.6758	0.6759
FastText	0.6562	0.6543	0.7208	0.7101	0.7391	0.7371
BERT[b]	0.7842	0.7857	0.8426	0.8400	0.7348	0.7083
BERT-CNN	0.7338	0.7340	0.8242	0.8213	0.7181	0.7136
BERT-RNN	0.8548	0.8546	0.8536	0.8532	0.7955	0.7882
BERT-RCNN	0.8245	0.8253	0.8313	0.8252	0.8364	0.8332
BERT-DPCNN	0.7839	0.7850	0.8186	0.8186	0.8301	0.8304
ERNIE	0.8801	0.8808	0.8681	0.8675	0.8895	0.8882
SBERT-Kmeans	0.8875	0.8875	0.8709	0.8709	0.8788	0.8788
SBERT-DocSCAN	0.8985	0.8985	0.8983	0.8983	0.9130	0.9130
SBERT-MEC	0.8938	0.8938	0.8673	0.8651	0.9039	0.9035

[a]It includes traditional models, BERT-based improved models, and the model proposed in this paper. These models were trained on 3 distinct datasets and subsequently evaluated on the test dataset in terms of accuracy and F_1-score.

[b]BERT: Bidirectional Encoder Representations from Transformers.

XSL•FO
RenderX

Table . Confusion matrices for the SBERT-DocSCAN model on the test dataset[a].

Label or cluster	0	1	2	3	Total	Recall
Model training on dataset #1						
Objective	20,302	1304	25	490	22,051	0.9175
Methods	1533	19,825	448	137	21,943	0.9035
Results	179	901	18,676	2075	21,831	0.8555
Conclusions	649	280	861	19,914	21,704	0.9175
Total	22,593	22,310	20,010	22,616	78,647[b]	0.8985
Precision	0.8595	0.8886	0.9333	0.8805	0.8985	__[c]
Model training on dataset #2						
Objective	20,250	1247	24	530	22,051	0.9183
Methods	705	20,496	397	345	21,943	0.934
Results	210	626	18,005	2990	21,831	0.8247
Conclusions	102	141	1585	19,876	21,704	0.8983
Total	21,267	22,510	20,011	23,741	78,647[b]	0.8983
Precision	0.9522	0.9105	0.8998	0.8372	0.8983	—
Model training on dataset #3						
Objective	20,401	251	129	1447	22,051	0.9178
Methods	505	19,673	1325	29	21,943	0.9137
Results	288	1655	19,864	593	21,831	0.8868
Conclusions	749	252	386	19,982	21,704	0.9131
Total	21,943	21,831	21,704	22,051	78,647[b]	0.913
Precision	0.9297	0.9012	0.9152	0.9062	0.913	—

[a]This table displays the confusion matrices for the SBERT-DocSCAN model, which was trained separately on 3 distinct training datasets. The matrices detail the accuracy and recall for each category when evaluated on the test dataset.

[b]Out of 87,529.

[c]Not applicable.

Ablation Study

To assess the impact of various submodules, we conducted ablation studies. We began by training 2 clustering models with identical structures (SBERT-DocSCAN). The key distinction between them was that 1 model was fine-tuned on our datasets, while the other was not. In addition, we trained 3 supervised models on different datasets and evaluated their performance on the test dataset. Notably, 1 of these models was the SBERT-MEC, which lacked the proposed dynamic convolution (DC) module. The results, presented in Table 5, clearly demonstrate that the SBERT-DocSCAN method, when fine-tuned with SBERT, outperforms others in terms of efficiency. Furthermore, within the supervised learning category, the SBERT-MEC model equipped with the DC module surpassed those lacking this module, underlining the value of the DC module in enhancing model performance.

XSL•FO
RenderX

Table . Performance of ablation model on test dataset[a].

Method	Training on dataset #1		Training on dataset #2		Training on dataset #3	
	Accuracy	F_1-score	Accuracy	F_1-score	Accuracy	F_1-score
SBERT-DocSCAN (without finetune)	0.3891	0.3895	0.3891	0.3895	0.3891	0.3895
SBERT-DocSCAN (with finetune)	0.8985	0.8985	0.8983	0.8983	0.9130	0.9130
BERT[b]	0.8330	0.8324	0.8424	0.8421	0.8406	0.8397
SBERT-MEC[c] (without DC)[d]	0.8714	0.8717	0.8390	0.8332	0.8736	0.8735
SBERT-MEC (with DC)	0.8938	0.8938	0.8673	0.8651	0.9039	0.9035

[a]This table compares the performance of 2 SBERT-DocSCAN clustering models—one fine-tuned on our datasets and the other not—and 3 supervised models, including an SBERT-MEC model without the dynamic convolution module. All models were evaluated on the test dataset.

[b]BERT: Bidirectional Encoder Representations from Transformers.

[c]MEC: multiencoder cascade.

[d]DC: dynamic convolution.

Discussion

Principal Findings

This study addresses the critical lack of real-world datasets for Chinese medical abstract classification by leveraging GPT-3.5 to generate synthetic datasets and developing models tailored for this task. Our findings confirm that synthetic datasets, when carefully designed, can match or surpass the performance of manually labeled datasets in sentence-level classification tasks. The SBERT-DocSCAN and SBERT-MEC models developed in this study demonstrate the potential of clustering and supervised approaches to effectively classify Chinese medical abstracts, illustrating the profound impact of synthetic datasets on enhancing NLP tasks.

Classical non-BERT models, such as TextCNN [7], TextRNN [32], and TextRCNN [34], exhibit strong performance in capturing local or sequential features. However, these models lack the deep contextual understanding provided by transformer-based models. While FastText [36] offers a lightweight and efficient alternative, its bag-of-words approach limits its ability to capture complex semantic relationships. These limitations underscore the advantages of leveraging pretrained language models in tasks that demand rich contextual understanding.

BERT-based models significantly improve performance by providing deep contextual representations. The success of models like BERT-CNN [37] and BERT-RCNN [38] aligns with prior studies, which highlight the effectiveness of combining BERT embeddings with convolutional and recurrent structures for enhanced local and sequential feature extraction. However, our proposed SBERT-DocSCAN and SBERT-MEC models outperform these baselines, indicating the added value of advanced clustering methods and DC layers. Specifically, DocSCAN [22], with its graph-based clustering approach, demonstrates superior clustering quality compared to k-means.

The ablation study further emphasizes the contributions of individual components in our models. The removal of the dynamic convolutional layer in the MEC module resulted in a notable decline in performance, highlighting its role in refining language representations. Similarly, the omission of the 3 encoders within MEC led to a significant reduction in accuracy, underscoring the importance of multiencoder architecture in capturing diverse linguistic features. These findings align with the broader NLP literature [39], which emphasizes the benefits of combining multiple feature extraction techniques for enhanced model performance.

Our exploration of ERNie [40], a BERT-based model optimized for Chinese contexts, further validates the importance of leveraging models tailored to specific languages and domains. While ERNie offers improvements over standard BERT in Chinese text classification, it does not surpass the performance of our SBERT-based approaches. This suggests that task-specific architectural innovations, such as the integration of clustering methods and DC modules, can provide greater benefits than domain-specific pretraining alone.

Overall, these comparisons and ablation studies highlight the robustness and versatility of our proposed methods, providing valuable insights into the design of models for sentence-level text classification tasks. By demonstrating the effectiveness of integrating advanced clustering methods, DC layers, and multiencoder architectures, this study contributes to the growing body of research focused on optimizing transformer-based models for real-world applications.

Limitations

While promising, our proposed methods have 2 main limitations. First, the quality of synthetic data heavily relies on the GPT-3.5 model, with performance contingent on effective prompt design. Refining prompt engineering strategies will be essential for future improvements. Second, the 2-stage approach used in both models results in an increased parameter size, raising concerns about computational resource efficiency. Further optimization

XSL•FO
RenderX

is necessary to address these scalability challenges, especially in real-world applications where resources may be limited.

Conclusions

This study demonstrates the significant potential of synthetic datasets, generated using GPT-3.5, in addressing the scarcity of labeled datasets for Chinese medical abstract classification. Our findings reveal that compact synthetic datasets can achieve performance comparable to, and in some cases surpass, that of manually labeled datasets. The proposed SBERT-DocSCAN and SBERT-MEC models further highlight the benefits of combining advanced clustering techniques, multiencoder architectures, and DC modules, showcasing their ability to enhance sentence-level classification tasks in specialized domains. These contributions provide valuable insights for leveraging generative artificial intelligence in NLP applications.

Beyond this specific task, this work underscores the transformative potential of synthetic datasets in reducing reliance on costly manual labeling, enabling broader adoption of NLP technologies in resource-limited fields. Future research can expand upon these findings by exploring more sophisticated data generation strategies, optimizing model architectures for efficiency, and fostering interdisciplinary collaborations to develop tailored solutions for complex, real-world challenges. By bridging gaps in data availability, this study provides a foundation for advancing NLP capabilities in medical and other specialized domains.

Acknowledgments

This research was supported by Shanghai Science and Technology Commission Research Project 24YF2731300, 23692107300, and 202121511101200, Postdoctoral Fellowship Program of CPSF under grant GZC20231604, the Xiangfu Lab Youth Program XF052024B0100, and Shanghai Municipal Commission of Economy and Information Technology (SHEITC) Project 2024-GZL-RGZN-01020.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

References

1. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Presented at: 31st International Conference on Neural Information Processing Systems (NeurlPS); Dec 4-9, 2017; Long Beach, CA, United States. [doi: 10.5555/3295222.3295349]
2. Devlin J, Chang MW, Lee K, Toutanova K. B. Pre-training of deep bidirectional transformers for language understanding. Presented at: 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; Jun 2-7, 2019; Minneapolis, MN, United States. [doi: 10.18653/v1/N19-1423]
3. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. OpenAI. 2020. URL: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf [accessed 2020-09-25]
4. Dai Z, Yang Z, Yang Y, Carbonell J, Le Q, Salakhutdinov R. Transformer-XL: attentive language models beyond a fixed-length context. Presented at: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; Jul 28 to Aug 2, 2019; Florence, Italy. [doi: 10.18653/v1/P19-1285]
5. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. Open AI. URL: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf [accessed 2025-03-03]
6. Zellers R, Holtzman A, Rashkin H, Bisk Y, Farhadi A, Roesner F, et al. Defending against neural fake news. Presented at: 33rd International Conference on Neural Information Processing Systems (NeurlPS); Dec 8-14, 2019; Vancouver, BC, Canada. [doi: 10.5555/3454287.3455099]
7. Kim Y. Convolutional neural networks for sentence classification. Presented at: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); Oct 25-29, 2014; Doha, Qatar. [doi: 10.3115/v1/D14-1181]
8. Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E. Hierarchical attention networks for document classification. Presented at: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics; Jun 12-17, 2016; San Diego, CA, United States. [doi: 10.18653/v1/N16-1174]
9. Conneau A, Kiela D, Schwenk H, Barrault L, Bordes A. Supervised learning of universal sentence representations from natural language inference data. Presented at: ; Sep 9-11, 2017; Copenhagen, Denmark. [doi: 10.18653/v1/D17-1070]
10. Yinhan L, Myle O, Naman G, Jingfei D, Mandar J, Danqi C, et al. RoBERTa: a robustly optimized BERT pretraining approach. arXiv. Preprint posted online on Jul 26, 2019. [doi: 10.48550/arXiv.1907.11692]

XSL•FO
RenderX

11. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov RR, Le QV. Xlnet: generalized autoregressive pretraining for language understanding. Presented at: 33rd International Conference on Neural Information Processing Systems (NeurlPS); Dec 8-14, 2019; Vancouver, BC, Canada. [doi: 10.5555/3454287.3454804]
12. Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R. Albert: a lite BERT for self-supervised learning of language representations. arXiv. Preprint posted online on Sep 26, 2019. [doi: 10.48550/arXiv.1909.11942]
13. Peters M, Neumann M, Iyyer M, et al. Deep contextualized word representations. Presented at: 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; Jun 1-6, 2018; New Orleans, LA, United States. [doi: 10.18653/v1/N18-1202]
14. Ward Jr JH. Hierarchical grouping to optimize an objective function. J Am Stat Assoc 1963 Mar;58(301):236-244. [doi: 10.1080/01621459.1963.10500845]
15. MacQueen J. Some methods for classification and analysis of multivariate observations. Presented at: 5th Berkeley Symposium on Mathematical Statistics and Probability; Dec 27, 1965 to Jan 7, 1966; Oakland, CA, USA.
16. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nat New Biol 1999 Oct 21;401(6755):788-791. [doi: 10.1038/44565] [Medline: 10548103]
17. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. J Mach Learn Res 2003 Mar 1;3:993-1022. [doi: 10.5555/944919.944937]
18. Le Q, Mikolov T. Distributed representations of sentences and documents. Presented at: 31st International Conference on International Conference on Machine Learning (ICML); Jun 21-25, 2014; Beijing, China. [doi: 10.5555/3044805.3045025]
19. Cer D, Yang Y, Kong SY, et al. Universal sentence encoder for english. Presented at: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations; Oct 31 to Nov 4, 2018; Brussels, Belgium. [doi: 10.18653/v1/D18-2029]
20. Li J, Liu X, Zhao H, Xu R, Yang M, Jin Y. BERT-EMD: many-to-many layer mapping for BERT compression with earth mover's distance. Presented at: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP); Nov 16-20, 2020. [doi: 10.18653/v1/2020.emnlp-main.242]
21. Reimers N, Gurevych I. Sentence-bert: sentence embeddings using siamese bert-networks. Presented at: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); Nov 3-7, 2019; Hong Kong, China. [doi: 10.18653/v1/D19-1410]
22. Stammbach D, Ash E. DocSCAN: unsupervised text classification via learning from neighbors. Presented at: Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022); Sep 12-15, 2022; Potsdam, Germany URL: https://aclanthology.org/2022.konvens-1.4 [accessed 2025-03-03]
23. Xu T, Zhang P, Huang Q, et al. AttnGAN: fine-grained text to image generation with attentional generative adversarial networks. Presented at: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); Jun 18-23, 2018; Salt Lake City, UT, United States. [doi: 10.1109/CVPR.2018.00143]
24. Wang K, Wan X. SentiGAN: generating sentimental texts via mixture adversarial networks. Presented at: 27th International Joint Conference on Artificial Intelligence (IJCAI); Jul 13-19, 2018; Stockholm, Sweden. [doi: 10.5555/3304222.3304388]
25. Zhang X, Zhao J, LeCun Y. Character-level convolutional networks for text classification. Presented at: 28th International Conference on Neural Information Processing Systems (NeurlPS); Dec 8-13, 2014; Montreal, QC, Canada. [doi: 10.5555/2969239.2969312]
26. Zhou R, Li X, He R, Bing L, Cambria E, Si L, et al. MELM: data augmentation with masked entity language modeling for low-resource NER. Presented at: The 60th Annual Meeting of the Association for Computational Linguistics; May 22-27, 2022; Dublin, Ireland. [doi: 10.18653/v1/2022.acl-long.160]
27. PubMed. URL: https://pubmed.ncbi.nlm.nih.gov [accessed 2023-10-20]
28. DeepL. URL: https://www.deepl.com/translator [accessed 2023-10-20]
29. Yudong L, Yuqing Z, Zhe Z, Linlin S, Weijie L, Weiquan M, et al. CSL: a large-scale chinese scientific literature dataset. Presented at: 29th International Conference on Computational Linguistics; Oct 12-17, 2022; Gyeongju, Republic of Korea.
30. CLUEbenchmark/SimCLUE. GitHub. URL: https://github.com/CLUEbenchmark/SimCLUE [accessed 2024-01-08]
31. Aghaebrahimian A. Quora question answer dataset. Presented at: 20th International Conference on Text, Speech, and Dialogue; Aug 27-31, 2017; Prague, Czech Republic. [doi: 10.1007/978-3-319-64206-2_8]
32. Pengfei L, Xipeng Q, Xuanjing H. Recurrent neural network for text classification with multi-task learning. Presented at: 25th International Joint Conference on Artificial Intelligence (IJCAI); Jul 9-15, 2016; New York City, NY, United States. [doi: 10.5555/3060832.3061023]
33. Zhou P, Shi W, Tian J, et al. Attention-based bidirectional long short-term memory networks for relation classification. Presented at: The 54th Annual Meeting of the Association for Computational Linguistics; Aug 7-12, 2016; Berlin, Germany. [doi: 10.18653/v1/P16-2034]
34. Wang R, Li Z, Cao J, Chen T, Wang L. Convolutional recurrent neural networks for text classification. Presented at: 2019 International Joint Conference on Neural Networks (IJCNN); Jul 14-19, 2019; Budapest, Hungary. [doi: 10.1109/IJCNN.2019.8852406]
35. Johnson R, Zhang T. Deep pyramid convolutional neural networks for text categorization. Presented at: 55th Annual Meeting of the Association for Computational Linguistics; Jul 30 to Aug 4, 2017; Vancouver, Canada. [doi: 10.18653/v1/P17-1052]

XSL•FO
RenderX

36. Joulin A, Grave E, Bojanowski P, Mikolov T. Bag of tricks for efficient text classification. Presented at: 15th Conference of the European Chapter of the Association for Computational Linguistics; Apr 3-7, 2017; Valencia, Spain. [doi: 10.18653/v1/E17-2068]
37. Kaur K, Kaur P. BERT-CNN: improving BERT for requirements classification using CNN. Procedia Comput Sci 2023;218:2604-2611. [doi: 10.1016/j.procs.2023.01.234]
38. Kaur K, Kaur P. BERT-RCNN: an automatic classification of app reviews using transfer learning based RCNN deep model. Research Square. Preprint posted online on Jan 24, 2023. [doi: 10.21203/rs.3.rs-2503700/v1]
39. Yanjun L, Haijun Z, Weimin P, Rujia F. Microblog rumor detection based on bert-DPCNN. In: Artificial Intelligence in China 2021:524-530. [doi: 10.1007/978-981-15-8599-9_60]
40. Zhang Z, Han X, Liu Z, Jiang X, Sun M, Liu Q. ERNIE: enhanced language representation with informative entities. Presented at: 57th Annual Meeting of the Association for Computational Linguistics; Jul 28 to Aug 2, 2019; Florence, Italy. [doi: 10.18653/v1/P19-1139]

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
CNN: convolutional neural network
DC: dynamic convolution
GAN: generative adversarial network
MEC: multiencoder cascade
NLP: natural language processing
RNN: recurrent neural network

Edited by A Mavragani; submitted 14.04.24; peer-reviewed by Y Wang; revised version received 28.11.24; accepted 28.11.24; published 19.03.25.

Please cite as:
Li J, Wang Z, Yu L, Liu H, Song H
Synthetic Data-Driven Approaches for Chinese Medical Abstract Sentence Classification: Computational Study
JMIR Form Res 2025;9:e54803
URL: https://formative.jmir.org/2025/1/e54803
doi:10.2196/54803

© Zikai Wang, Longxuan Yu, Hui Liu, Haitao Song. Originally published in JMIR Formative Research (https://formative.jmir.org), 19.3.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Identifying High-Priority Ethical Challenges for Precision Emergency Medicine: Nominal Group Study

Christian Rose[1], MD; Emily Shearer[2], MD; Isabela Woller[3], BSc; Ashley Foster[4], MD; Nicholas Ashenburg[1], MD; Ireh Kim[1]; Jennifer Newberry[1], MS, MD, JD

[1]Department of Emergency Medicine, Stanford University School of Medicine, 500 Pasteur Dr, Stanford, CA, United States

[2]Department of Emergency Medicine, Alpert School of Medicine, Brown University, Providence, RI, United States

[3]Loyola University Chicago, Chicago, IL, United States

[4]Department of Emergency Medicine, University of California, San Francisco, San Francisco, CA, United States

Corresponding Author:
Christian Rose, MD
Department of Emergency Medicine, Stanford University School of Medicine, 500 Pasteur Dr, Stanford, CA, United States

Abstract

Background: Precision medicine promises to revolutionize health care by providing the right care to the right patient at the right time. However, the emergency department's unique mandate to treat "anyone, anywhere, anytime" creates critical tensions with precision medicine's requirements for comprehensive patient data and computational analysis. As emergency departments serve as health care's safety net and provide a growing proportion of acute care in America, identifying and addressing the ethical challenges of implementing precision medicine in this setting is crucial to prevent exacerbation of existing health care disparities. The rapid advancement of precision medicine technologies makes it imperative to understand these challenges before widespread implementation in emergency care settings.

Objective: This study aimed to identify high priority ethical concerns facing the implementation of precision medicine in the emergency department.

Methods: We conducted a qualitative study using a modified nominal group technique (NGT) with emergency physicians who had previous knowledge of precision medicine concepts. The NGT process consisted of four phases: (1) silent generation of ideas, (2) round-robin sharing of ideas, (3) structured discussion and clarification, and (4) thematic grouping of priorities. Participants represented diverse practice settings (county hospital, community hospital, academic center, and integrated managed care consortium) and subspecialties (education, ethics, pediatrics, diversity, equity, inclusion, and informatics) across various career stages from residents to late-career physicians.

Results: A total of 12 emergency physicians identified 82 initial challenges during individual ideation, which were consolidated to 48 unique challenges after removing duplicates and combining related items. The average participant contributed 6.8 (SD 2.9) challenges. These challenges were organized into a framework with 3 themes: values, privacy, and justice. The framework identified the need to address these themes across 3 time points of the precision medicine process: acquisition of data, actualization in the care setting, and the after effects of its use. This systematic organization revealed interrelated concerns spanning from data collection and bias to implementation challenges and long-term consequences for health care equity.

Conclusions: Our study developed a novel framework that maps critical ethical challenges across 3 domains (values, privacy, and justice) and 3 temporal stages of precision medicine implementation. This framework identifies high-priority areas for future research and policy development, particularly around data representation, privacy protection, and equitable access. Successfully addressing these challenges is essential to realize precision medicine's potential while preserving emergency medicine's core mission as health care's safety net.

(JMIR Form Res 2025;9:e68371) doi:10.2196/68371

KEYWORDS

precision medicine; emergency medicine; ethical considerations; nominal group study; consensus framework

Introduction

Historically, medical treatments have been tailored to the average patient. However, no individual is ever the average patient—everyone has their own risk factors, needs, and ideal outcomes. Precision medicine, as defined by the White House's nationwide Precision Medicine Initiative, aims to address this problem by moving from a one-size-fits-all model toward one of providing the "right care to the right person at the right time" through the use of vast amounts of health data, including

XSL•FO
RenderX

genomic information, electronic health records, environmental and lifestyle data, and medical imaging data, among others [1-5]. However, even the most advanced, targeted therapies are limited by biases that already exist in the US health care system [6,7].

Variation in care, even when "evidence based," can result in worsened gaps between the haves and the have-nots, disproportionately burdening already marginalized communities [8]. For example, patients with limited resources or inadequate insurance coverage may be unable to access these evidence-based interventions, while well-insured patients at major medical centers can, creating a 2-tiered system where advances in precision medicine may primarily benefit those with better health care access. For emergency physicians, who serve in the safety net for American health care, precision medicine's "right care, right patient, right time" must be balanced with our "anyone, anywhere, anytime" mantra and the practical and ethical challenges that this setting presents.

Early precision medicine initiatives have focused on the use of genetics to better understand or treat diseases. Much has been made of determining the right medication for the right person (eg, warfarin dosing for those with a CYP [cytochrome P450] mutation) or when to start cancer screening based on genetic profiles (eg, colorectal screening for CHD1 [chromodomain helicase DNA binding protein 1]) [9]. It is clear, however, that genetics alone cannot explain much of the variation in outcomes across the population [2,10]. The effect of behavioral factors like diet and exercise are well-established, and studies have begun to illuminate the effect of social and environmental determinants of health like access to health care resources or housing, which may be responsible for over half of health outcomes [11,12]. As such, there has been a broadening approach to precision medicine that may also include home monitoring, social media and social networks, as well as patient-generated data like diet, daily steps, exercise, or heart rate variability [13,14].

While useful when available, not all patients can afford these tools and services. Furthermore, "precise" treatments rely on the accuracy of this data, which must be built into datasets and shared. Few comprehensive datasets that incorporate all of these data exist, and even fewer have been carefully evaluated for missing or inaccurate data that could bias results [15]. Current datasets suffer from a lack of representation from some populations while simultaneously overrepresenting others. Healthy patients or those without access to care rarely present to hospitals, while those who are chronically ill are seen more often. Furthermore, a skewed representation occurs because research data often comes primarily from large academic centers rather than community hospitals or rural health care settings [16]. Data that were collected where patients are provided differential care might inadvertently support antiquated practices [17-21] due to the association of outcomes with variables like the ability to pay for a higher level of care. Though precision approaches aim to limit quality gaps and provide more equitable care, decisions that are made without the full picture of their context might inadvertently exacerbate systemic biases, for example, when risk assessment scores built from data of principally white cohorts over- or underestimate the risk of disease in non-White populations [7].

The emergency department setting offers additional unique hurdles for the growth of precision medicine. Little patient-generated data are readily available at the point of care. Highly important features like code status are haphazardly available among the sickest patients. Nevertheless, emergency care providers strive to provide better care, and to actualize the benefits of a learning health care system by integrating evidence and experience into continuous improvement and innovation. How can health care's safety net, responsible for a growing and significant portion of acute care in America [22,23], begin to approach the problems that are inherent to the implementation of precision medicine so that it might offer real-time, equitable care for patients with the widest spectrum of disease in medicine and a microcosm of the health care system?

Thus, the burgeoning expansion of precision medicine underscores the need to identify challenges facing its nascent deployment and ethical use in the emergency setting such that its optimal benefits may be realized. Our main objective was to identify critical ethical challenges in implementing precision medicine in the emergency setting using a consensus technique. This approach would also help identify themes and gaps in current research. This manuscript presents the resulting consensus framework.

Methods

Study Design and Setting

We used a modified nominal group technique (NGT), a consensus methodology widely used in emergency medicine, to identify high priority ethical challenges relevant to the implementation of precision medicine in the emergency medicine setting and arrive at a consensus framework [24-31]. This study followed the COREQ (Consolidated Criteria for Reporting Qualitative Research) reporting framework for qualitative studies [32].

NGT is a 4-step process in which stakeholders generate and then prioritize questions or research ideas (Figure 1) [33]. NGT was well-suited for our research question as our aim was to generate a comprehensive set of ethical challenges which could then be organized and ordered. This is in contrast to other frequently used consensus techniques such as the Delphi technique, which are designed instead to reach a consensus opinion on one particular question [24]. NGT has previously been used successfully to identify research priorities in the field of emergency medicine [25,27,34-37].

XSL•FO
RenderX

Figure 1. Flow chart of the nominal group technique (NGT) process.

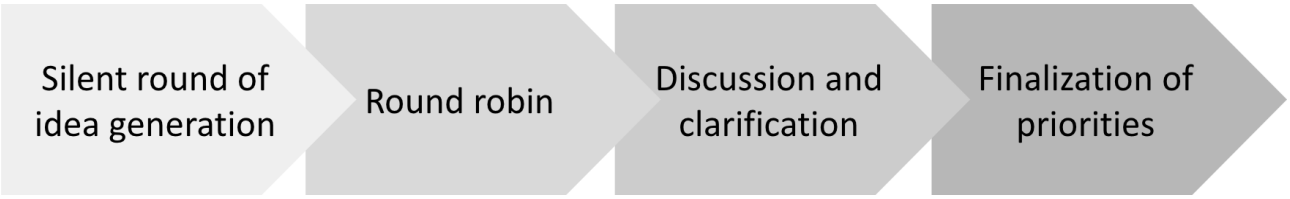

Selection and Participants

A NGT session group size is recommended at greater than 6 participants, with the ideal between 9 and 12 [34]. As precision emergency medicine is an emerging field, we sought participants who had familiarity with the clinical practice of emergency medicine, including an understanding of the types of data that are collected throughout a patient's journey and the ways in which data is used for clinical decision-making. Participants were recruited through email to the Society of Academic Medicine Ethics Committee listserve; the email invitation also asked recipients for referral and sharing to others who may be interested. A quota sample of the first 14 respondents to the email were registered as potential participants in the NGT group to meet the ideal group size and allow for scheduling conflicts. In total, 12 emergency medicine physicians who reported working knowledge of precision medicine topics, as evidenced by publication records and participation in precision medicine conferences or similar activities, joined our NGT session. Participants represented a spectrum of subspecialties (education, ethics, pediatrics, diversity, equity, inclusion, and informatics) across various stages of training (resident, early-, mid-, and late-career) and practice settings (county hospital, community hospital, academic center, and integrated managed care consortium) (Specialty, Level of Training, and Practice Setting included in Table S1 in Multimedia Appendix 1) to provide a range of emergency medicine practice perspectives.

Participants took part in the NGT session during the American College of Emergency Physicians annual conference in Boston in October 2021. The process was moderated by the lead researcher of the group (CR), and notes were taken by 2 separate researchers (ES and MG).

NGT Process

Phase 1: Silent Idea Generation.

The first step of NGT is a round of silent idea generation. In this phase, after being given an initial prompt, participants silently write down what they believe to be high priority challenges. It is imperative that this phase is silent, as this allows each member of the group to write their initial thoughts without influence from other members. Our prompt asked participants to identify high priority ethical concerns for the implementation of precision medicine initiatives in the emergency medicine setting. They had 10 minutes to complete this phase.

Phase 2: Sharing Ideas.

After individual idea generation, participants then shared their ideas with the group in a round-robin (going around the table one person at a time) format. NGT has the advantage of not allowing any one particular voice to outweigh others, by asking each participant to share their complete list of ideas in turn during the round robin phase, rather than ad hoc discussion, thus minimizing potential bias [38,39]. This phase lasted 10 minutes.

Phase 3: Discussion and Clarification

After the round-robin format, participants undertook discussion and clarification of topics. In this phase, participants were encouraged to ask other members for clarification of topics generated as needed. Cross-cutting themes were identified, and duplicate topics consolidated. Over 30 minutes, key topics were then placed on moveable cards for prioritization and further thematic grouping.

Phase 4: Finalization of Priorities

Finally, participants finalized the topics they believed represented the highest priority areas to address. In traditional NGT, this is the ranking phase, where members are allocated votes to generate a ranked list of priorities. Overwhelmingly, however, participants felt that the topics they raised were interrelated and of commensurate importance over several themes such that their categorization—rather than prioritization—would provide greater impactful focus for interventions. Consequently, participants grouped challenges through consensus opinion over a 40-minute period.

Ethical Considerations

This study was approved by the Stanford University Institutional Review Board (IRB #62687); participants gave written informed consent before taking part in this study. All data were collected anonymously. Participants received a US $125 gift card in compensation for their time.

Results

During phases 1 and 2 of the NGT process, 82 challenges were generated (complete list in Table S2 in Multimedia Appendix 1) with the average participant contributing 6.8 (SD 2.9) challenges. After refinement and separation of cross-cutting themes and consolidation of duplicate topics in phase 3, a list of 48 unique challenges remained.

In phase 4, participants grouped these challenges into common realms based on the nature of the ethical issue and where in the precision medicine process the issue was likely to occur. The participants identified 3 key time points: the acquisition of data, the actualization of precision medicine in practice, and the aftereffects of the process (Textboxes 1-3). The result of this mapping was a 3 by 3 matrix consisting of ethical challenges grouped into the domains of Values (eg, patient values and priorities as well as how precision medicine values those individuals), Privacy (eg, patient expectations as well as literal data security), and Justice (eg, overcoming and addressing

XSL•FO
RenderX

systemic biases) as related to their temporal relationship to the precision medicine process (Figure 2).

Textbox 1. Challenges identified by participants related to how the unique values of patients, providers, and the health care system will be incorporated into precision medicine.

Values

- **Acquisition**
- How are decisions made about which disease conditions to study and treat?
- How accurate is data in the medical record at representing the whole of an individual?
- How much does lifestyle matter and is it captured?
- How will social determinants of health be evaluated?
- Should individuals worry about the impact of nonmodifiable risk factors like family history or genetic predisposition?
- What if individuals identify differently from measures used in precision medicine?
- Can we mitigate the mistrust of people toward engaging with the system?
- **Actualization**
- How do we address patient choice and preference in the face of precision recommendations?
- What do we do if health care providers' interpretations and values conflict with those of the patient?
- How might this affect the patient-physician relationship?
- Will patients be able to opt out?
- Does the delivery of precision medicine change in the emergency scenario?
- How do we ensure provider comfort with using precision medicine?
- How do health care providers stay up to date with rapid changes in precision medicine?
- **Aftereffects**
- How do we maintain humility as evidence for decision-making changes over time?
- What if precision recommendations turn out to be wrong?
- Might there be unintended psychological consequences like stress after decisions are made?
- How do we ensure patient choice and preference is communicated longitudinally between medical teams?
- What prevents a capitalist model from researching only the most common or lucrative conditions?

XSL•FO
RenderX

Textbox 2. Challenges identified by participants regarding privacy in the precision medicine process.

Privacy

- **Acquisition**
- Who will be accountable for the collected information's storage, maintenance, and safety?
- How will protected information be handled?
- How do we protect minors while ensuring appropriate data sharing with their caregivers?
- How do we include underrepresented or excluded groups in a way that avoids perception of surveillance?
- **Actualization**
- How will patients interface with their personal medical data?
- Will patients lose the power to make private decisions about their own health when it may have a downstream effect on others?
- How will changing evidence or incidental findings like new genetic risk factors be communicated with patients to ensure follow-up?
- **Aftereffects**
- Can anonymized data be reidentified?
- How do we ensure safe data sharing can occur among differing institutions or electronic medical record systems?
- What are the legal ramifications of an infinitely variable standard of care?

Textbox 3. Challenges identified by participants about how to ensure just care in the precision medicine model and process.

Justice

- **Acquisition**
- How do we incorporate all people into precision research?
- Can we address disparities like access to care that make it hard for patients to participate?
- Will those who are historically excluded continue to be so, while some populations will see over- and undersampling?
- What do we do with historical data that may be biased?
- Who will fund the research and data acquisition?
- How are precision medicine priorities determined (ie, institutional level or government level)?
- Could precision research resources be better spent addressing direct care access?
- **Actualization**
- Will flawed data inputs result in flawed data outputs and recommendations?
- How do we build a relationship of trust with a patient in the emergency department?
- How do we prevent nonmodifiable predispositions from leading to biased care (ie, genetic determinism)?
- How do we allow patient choice and values in the face of recommendations or limited resources?
- Can variations in care still be equitable?
- Will people get disproportionate care based on affordability?
- How can we balance precision medicine safety with accessibility?
- **Aftereffects**
- How will we maintain monitoring and vigilance of the accuracy of these systems?
- How do we ensure systems are learning and adapting over time?
- Can we ensure precision medicine is equitable across all patient populations?
- Might the choice of who and what to study and treat have unintended consequences?
- Might this further biased or racist ideologies and marginalize patients?

XSL•FO
RenderX

Figure 2. Priority matrix of ethical concerns for precision medicine.

	Values	Privacy	Justice
	(Complete list in Textbox 1)	(Complete list in Textbox 2)	(Complete list in Textbox 3)
Acquisition	Who decides what to include in datasets?	How will data be collected and maintained?	Will some populations be over- or undersampled?
Actualization	Do patients get to choose what treatment is right for them?	Will patients lose the power to make private decisions?	Will precision medicine be available to everyone?
Aftereffect	Who determines if we made the right choice?	Can data be reidentified later?	Can we ensure equitable distribution of resources?

Discussion

Principal Findings

In this study, we have identified critical ethical challenges facing the implementation of precision medicine in the unique practice environment of the emergency department. The growing use of vast amounts of genomic, clinical, and lifestyle data to guide individualized therapies holds immense potential to improve emergency care. However, the "anyone, anywhere, anytime" mandate of emergency medicine presents distinct hurdles for the deployment of precision approaches that rely on the timely availability of complete patient information in a fast-paced and high-acuity setting. Using a modified nominal group technique, we engaged a diverse panel of emergency physicians to develop a consensus framework mapping key ethical challenges along the continuum of precision medicine implementation, from data acquisition to use in clinical practice to monitoring of downstream consequences. Crucially, this framework highlights the need to proactively identify and mitigate potential pitfalls to best realize the promise of precision medicine in the ED while promoting equitable access and benefit for the diverse patient populations we serve. Our findings provide a framework to guide research, practice innovation, and policy aimed at the ethical translation of precision medicine to the front lines of emergency care.

Shared Ethical Challenges: Values, Privacy, and Justice

A recurring theme in our NGT discussion was "whose values are we accounting for?" [40]. Participants acknowledged the sometimes conflicting needs of the individual versus those of the system. Providers may recommend medications that prolong survival, but side effects may be undesirable or cost prohibitive for some patients [41]. The definition of high-quality care envisioned by precision medicine may not align with the individual goals of patients. There are multiple stakeholders driving the discourse of precision medicine—health care providers, health care insurers, regulators, and industry—the least empowered of which is the patient [40]. There were many questions raised regarding challenges to patient autonomy and the ambiguity of who is the ultimate decision maker in precision medicine and how that might change over time, as the needs of the many may be at odds with the needs of a few.

XSL•FO
RenderX

This tension maybe at odds with the ethical principle of autonomy [42]. The definition of respect for autonomy requires that patients are able to make their own decisions while incorporating their own values. This potentially extends to the use of their data and specimens, but the limits of patient autonomy remain under debate [43-45]. For example, the appropriate approach to consent for enrollment in biobanks that start in emergency departments remains under debate [46,47].

Ensuring that patient autonomy is maintained can be particularly challenging in the emergency care context. In the emergency department, providers make decisions with patients in a time-constrained setting, particularly compared with those historical clinical environments of precision medicine, such as oncology, which may occur over weeks. Not only does this limit time for discussion, it may also create a sense of external pressure to make a decision [48]. In such time-sensitive situations, patient personal information and preferences may be disclosed and decisions required under duress, thus also implicating privacy [49].

Privacy includes not only the protection of one's identity and personal information, but also the ability to make choices without them being disclosed to others [50]. Another challenge to the patient's privacy is that the practice of precision medicine in the emergency department will require massive amounts of data transiting among and across institutions and devices. Ensuring the safety and security of this protected information is a major hurdle for precision medicine initiatives, and the abovementioned time-sensitive nature of the emergency department encounter makes it necessary to share this data in real time, which is unique from other medical specialties [51-54]. In addition, when patients know that the information they provide to the system will be used to make recommendations for themselves or others, it may have downstream effects. For example, alcohol use and history for the allotment of liver transplant is one area that may leave patients in a moral quandary, unsure whether to be open with their behavior but fearing the effects of those disclosures [55,56]. Patient perceptions of the trustworthiness of health care providers and institutions will impact the success of any scientific or research endeavor, particularly among historically disadvantaged populations [57].

It is no surprise that our consensus framework includes justice as a key area for research in precision medicine ethics as it grows in the emergency department and in the prehospital setting [58-60]. Emergency physicians carry the distinct duty under federal law to provide care to anyone who presents for emergency care [61]. This duty is widely recognized to rest on a fundamental moral responsibility to provide care "promptly and expertly, without prejudice or partiality" [49]. At first it may seem that precision medicine would help to limit personal biases which might affect care, using data as opposed to gestalt or local practice variation. But structural biases may encode results and outcomes on patients that might inadvertently limit the benefits of the precision medicine model. When they are present in the initial development of the model, such as lack of representation and personal biases, affects the validity of results used to make targeted recommendations [62].

Justice, or more specifically with respect to medical ethics, distributive justice, emphasizes that resources are to be distributed fairly, equitably, and appropriately. This can be heavily hindered if providers rely on precision models to make recommendations without a critical appraisal of how well the results fit the patient in front of them. Even if done well, the outputs of precision medicine may not be available to all [63]. Precision medicine which is ultimately only accessible to a few but built on the data of many, would be an affront to the growing field of social emergency medicine [59]. In oncology, an area with a much longer history of precision medicine, 78% of patients and providers reported concerns regarding the accessibility of precision medicine [64].

Continuum of Precision Emergency Medicine: Data Availability, Actualization, and Aftereffects

As noted by participants, challenges in implementing precision medicine in the emergency medicine context occur anywhere from when data is collected (or not collected) to the downstream effects of using precision medicine interventions. In emergency medicine, providers stand prepared to address a range of conditions at any time. We often must rely on incomplete patient histories (eg, due to the severity of illness or mental status changes frequently encountered in emergency medicine) and limited access to previous medical records (eg, our patients may be uninsured, underinsured, and without access to routine primary care providers) [65]. Precision medicine approaches will need to be deployable in a timely manner and maintain their precision with only the limited patient-specific information available to emergency physicians.

Furthermore, even when data are available, the actualization of precision medicine in the care environment may serve only to perpetuate systemic flaws in the health care system if that data is inaccurate, biased, or exclusionary from the start [7,66]. A 12-year review of the National Hospital Ambulatory Medical Care Survey (NHAMCS) illustrates this risk: Black patients were 10% less likely to receive immediate or emergent Emergency Severity Index (ESI) scores and to be admitted, but 1.26 times more likely to die in the emergency department or the hospital than White patients [65]. Using such historically biased data to make precision medicine recommendations about resource allocation could amplify these disparities. While precision medicine aims to deliver individually tailored care, it still requires categorizing patients into discrete groups by shared characteristics—a process that carries its own implications and potential for bias [4,67].

Additional logistic concerns for the actualization of precision medicine in emergency departments, include costs, legal issues, and patient education. Efforts made to collate data from emergency care settings to be used in precision medicine research may result in products or treatment pathways unaffordable to these patients [68-70]. This may be mitigated through insurers [71], but the emergency department is a safety net, and as such this approach may not benefit the most vulnerable patients while those who are insured may find that their plans have little incentive to offer this option [72]. Research has shown that the cost of precision medicine is a concern shared by patients and providers alike [63,64].

XSL•FO
RenderX

Even when payment is not an issue, a data- and machine-supported approach may not be conducive to truly informed consent given the limitations of adequately educating emergency department patients regarding how precision medicine and its associated research works [63,73]. Beyond the challenge of consenting in the emergency context, providers must also consider issues regarding disclosing incidental information and the potential changes associated with the use of data over time which interconnected to other growing datasets. Stewardship of these data, defined as "the responsible management of something entrusted to one's care" [74], will be critical to maintaining trust of patients and communities as this science evolves [57].

Discussion in this NGT process highlighted the need to actively maintain a holistic view of the patient without being overly reductive to these categories. Currently, only a limited amount of data is collected with respect to health measures or lab studies, and even less is recorded for social determinants of health. Future work should focus on systematic collection of environmental and social factors in emergency care settings, including neighborhood-level data on environmental exposures, community resources, and barriers to care access. However, this data collection must be done thoughtfully to avoid reinforcing biases—for example, by ensuring community engagement in data collection methods, using validated measurement tools that have been tested across diverse populations, and analyzing social factors as modifiable system-level variables rather than immutable patient characteristics. In emergency medicine specifically, this could include developing standardized screening tools for social needs that are feasible in acute care settings, creating automated systems to gather community-level data that do not burden individual patient encounters, and establishing partnerships with community organizations to better understand local health contexts. These approaches could help create more comprehensive precision medicine models while maintaining emergency medicine's core mission of providing equitable care to all patients.

Limitations

The use of a NGT and participant diversity are key strengths of this study. However, several limitations warrant discussion. While NGT aims to limit the biases associated with other consensus models, expertise bias is still a potential limitation, as all participants were aware of and knowledgeable on this topic beforehand. While this expertise strengthens the confidence in the working knowledge of the topic, it can also lead to anchoring. We took several steps to limit its impact. NGT, by way of its structured format, is meant to limit the influence of any one dominant individual in discussion [39]. In addition, we included a diverse spectrum of emergency physicians from residents to attending physicians with varying years of experience, as well as emergency physicians with a variety of subspeciality training.

Self-selection bias should also be considered, as participants volunteered in response to an email invitation to the Society of Academic Medicine Ethics Committee listserve. This recruitment method may have attracted participants with stronger interest in ethics and precision medicine, though their engagement likely enhanced the depth of insights generated during the NGT session.

Finally, though the number of participants was within ideal range for an NGT [24], the relatively small group size remains a limitation. Future research could benefit from additional strategies that would incorporate a higher number of emergency physicians and experts from outside of the field which might offer perspective to our work.

Conclusions

Realizing the promise of precision medicine's "right person, right place, right time" must be balanced by the context of the "anyone, anywhere, anytime" of emergency medicine. Our work has illuminated a range of concerns for the development of precision emergency medicine, and provides a focused framework for the most pressing ethical considerations along the continuum of data acquisition to implementation. This framework may be used to direct research and innovation toward addressing these challenges in the emergency medicine setting.

Acknowledgments

This study was made possible thanks to the financial support provided through a generous seed grant from Stanford University's McCoy Family Center for Ethics in Society - Ethics, Society, and Technology Hub Seed Grant.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Authors' Contributions

CR, JN, ES, and IK conceived of, designed, and completed the research. CR and JN obtained research funding. AF and NA performed data analysis. CR and JN drafted the manuscript. IK and IW contributed to the manuscript development and all authors contributed substantially to manuscript editing. CR takes responsibility for the work as a whole. No generative AI was used in any portion of the manuscript writing.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
Additional tables.
[DOCX File, 20 KB - formative_v9i1e68371_app1.docx]

References

1. Salari P, Larijani B. Ethical issues surrounding personalized medicine: a literature review. Acta Med Iran 2017 Mar;55(3):209-217. [Medline: 28282721]
2. Liu X, Luo X, Jiang C, Zhao H. Difficulties and challenges in the development of precision medicine. Clin Genet 2019 May;95(5):569-574. [doi: 10.1111/cge.13511] [Medline: 30653655]
3. Sitapati A, Kim H, Berkovich B, et al. Integrated precision medicine: the role of electronic health records in delivering personalized treatment. Wiley Interdiscip Rev Syst Biol Med 2017 May;9(3):e1378. [doi: 10.1002/wsbm.1378] [Medline: 28207198]
4. Seymour CW, Gomez H, Chang CCH, et al. Precision medicine for all? Challenges and opportunities for a precision medicine approach to critical illness. Crit Care 2017 Oct 18;21(1):257. [doi: 10.1186/s13054-017-1836-5] [Medline: 29047353]
5. Mensah GA, Jaquish C, Srinivas P, et al. Emerging concepts in precision medicine and cardiovascular diseases in racial and ethnic minority populations. Circ Res 2019 Jun 21;125(1):7-13. [doi: 10.1161/CIRCRESAHA.119.314970] [Medline: 31219738]
6. Lee SJ, Fullerton SM, Saperstein A, Shim JK. Ethics of inclusion: Cultivate trust in precision medicine. Science 2019 Jun 7;364(6444):941-942. [doi: 10.1126/science.aaw8299] [Medline: 31171685]
7. Geneviève LD, Martani A, Shaw D, Elger BS, Wangmo T. Structural racism in precision medicine: leaving no one behind. BMC Med Ethics 2020 Feb 19;21(1):17. [doi: 10.1186/s12910-020-0457-8] [Medline: 32075640]
8. Stein JN, Charlot M, Cykert S. Building toward antiracist cancer research and practice: The case of precision medicine. JCO Oncol Pract 2021 May;17(5):273-277. [doi: 10.1200/OP.20.01070] [Medline: 33974820]
9. Goetz LH, Schork NJ. Personalized medicine: motivation, challenges, and progress. Fertil Steril 2018 Jun;109(6):952-963. [doi: 10.1016/j.fertnstert.2018.05.006] [Medline: 29935653]
10. Elemento O. The future of precision medicine: towards a more predictive personalized medicine. Emerg Top Life Sci 2020 Sep 8;4(2):175-177. [doi: 10.1042/ETLS20190197] [Medline: 32856697]
11. Hill-Briggs F, Adler NE, Berkowitz SA, et al. Social determinants of health and diabetes: a scientific review. Diabetes Care 2020 Nov 2;44(1):258-279. [doi: 10.2337/dci20-0053] [Medline: 33139407]
12. Marmot M, Allen JJ. Social determinants of health equity. Am J Public Health 2014 Sep;104 Suppl 4(Suppl 4):S517-S519. [doi: 10.2105/AJPH.2014.302200] [Medline: 25100411]
13. Kalra D. The importance of real-world data to precision medicine. Per Med 2019 Mar;16(2):79-82. [doi: 10.2217/pme-2018-0120] [Medline: 30724116]
14. Hollister B, Bonham VL. Should electronic health record-derived social and behavioral data be used in precision medicine research? AMA J Ethics 2018 Sep 1;20(9):E873-E880. [doi: 10.1001/amajethics.2018.873] [Medline: 30242819]
15. Sukumar SR, Natarajan R, Ferrell RK. Quality of big data in health care. Int J Health Care Qual Assur 2015;28(6):621-634. [doi: 10.1108/IJHCQA-07-2014-0080] [Medline: 26156435]
16. DesHarnais S. Current uses of large data sets to assess the quality of providers. Construction of risk-adjusted indexes of hospital performance. Int J Technol Assess Health Care 1990;6(2):229-238. [doi: 10.1017/s0266462300000751] [Medline: 2118489]
17. Williams WW, Hogan JW, Ingelfinger JR. Time to eliminate health care disparities in the estimation of kidney function. N Engl J Med 2021 Nov 4;385(19):1804-1806. [doi: 10.1056/NEJMe2114918]
18. Wiens J, Saria S, Sendak M, et al. Do no harm: a roadmap for responsible machine learning for health care. Nat Med 2019 Sep;25(9):1337-1340. [doi: 10.1038/s41591-019-0548-6]
19. Char DS, Shah NH, Magnus D. Implementing machine learning in health care - addressing ethical challenges. N Engl J Med 2018 Mar 15;378(11):981-983. [doi: 10.1056/NEJMp1714229] [Medline: 29539284]
20. Owens K, Walker A. Those designing healthcare algorithms must become actively anti-racist. Nat Med 2020 Sep;26(9):1327-1328. [doi: 10.1038/s41591-020-1020-3]
21. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. Ann Intern Med 2018 Dec 18;169(12):866-872. [doi: 10.7326/M18-1990] [Medline: 30508424]
22. Marcozzi D, Carr B, Liferidge A, Baehr N, Browne B. Trends in the contribution of emergency departments to the provision of hospital-associated health care in the USA. Int J Health Serv 2018 Apr;48(2):267-288. [doi: 10.1177/0020731417734498] [Medline: 29039720]
23. Hsia RY, Sabbagh SH, Guo J, Nuckton TJ, Niedzwiecki MJ. Trends in the utilisation of emergency departments in California, 2005–2015: a retrospective analysis. BMJ Open 2018 Jul;8(7):e021392. [doi: 10.1136/bmjopen-2017-021392]
24. McMillan SS, King M, Tully MP. How to use the nominal group and Delphi techniques. Int J Clin Pharm 2016 Jun;38(3):655-662. [doi: 10.1007/s11096-016-0257-x] [Medline: 26846316]

XSL•FO
RenderX

25. Ranney ML, Fletcher J, Alter H, et al. A consensus-driven agenda for emergency medicine firearm injury prevention research. Ann Emerg Med 2017 Feb;69(2):227-240. [doi: 10.1016/j.annemergmed.2016.08.454] [Medline: 27998625]
26. Elder E, Johnston AN, Byrne JH, Wallis M, Crilly J. Core components of A staff wellness strategy in emergency departments: A clinician-informed nominal group study. Emerg Med Australas 2021 Feb;33(1):25-33. [doi: 10.1111/1742-6723.13561] [Medline: 32592326]
27. Nik Hisamuddin R, Tuan Hairulnizam TK. Developing key performance indicators for emergency department of teaching hospitals: a mixed fuzzy Delphi and nominal group technique approach. Malays J Med Sci 2022 Apr;29(2):114-125. [doi: 10.21315/mjms2022.29.2.11] [Medline: 35528822]
28. Peltzer-Jones J, Nordstrom K, Currier G, Berlin JS, Singh C, Schneider S. A research agenda for assessment and management of psychosis in emergency department patients. West J Emerg Med 2019 Mar;20(2):403-408. [doi: 10.5811/westjem.2019.1.39263] [Medline: 30881564]
29. Members of the SAEM Consensus Conference Emergency Medical Services Subcommittee, Adelgais KM, Hansen M, et al. Establishing the key outcomes for pediatric emergency medical services research. Acad Emerg Med 2018 Dec;25(12):1345-1354. [doi: 10.1111/acem.13637] [Medline: 30312993]
30. Marynowski-Traczyk D, Wallis M, Broadbent M, et al. Optimising emergency department and acute care for people experiencing mental health problems: a nominal group study. Aust Health Rev 2022 Oct;46(5):519-528. [doi: 10.1071/AH21092] [Medline: 34793296]
31. Safdar B, Greenberg MR. Organization, execution and evaluation of the 2014 Academic Emergency Medicine consensus conference on Gender-Specific Research in Emergency Care - an executive summary. Acad Emerg Med 2014 Dec;21(12):1307-1317. [doi: 10.1111/acem.12530] [Medline: 25420469]
32. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
33. Agency for Healthcare Research and Quality. Nominal group technique. URL: https://digital.ahrq.gov/health-it-tools-and-resources/evaluation-resources/workflow-assessment-health-it-toolkit/all-workflow-tools/nominal-group-technique [accessed 2024-12-18]
34. Harvey N, Holmes CA. Nominal group technique: an effective method for obtaining group consensus. Int J Nurs Pract 2012 Apr;18(2):188-194. [doi: 10.1111/j.1440-172X.2012.02017.x] [Medline: 22435983]
35. Kottmann A, Krüger AJ, Sunde GA, et al. Establishing quality indicators for pre-hospital advanced airway management: a modified nominal group technique consensus process. Br J Anaesth 2022 Feb;128(2):e143-e150. [doi: 10.1016/j.bja.2021.08.031] [Medline: 34674835]
36. Sunde GA, Kottmann A, Heltne JK, et al. Standardised data reporting from pre-hospital advanced airway management - a nominal group technique update of the Utstein-style airway template. Scand J Trauma Resusc Emerg Med 2018 Jun 4;26(1):46. [doi: 10.1186/s13049-018-0509-y] [Medline: 29866144]
37. Nelson V, Dubov A, Morton K, Fraenkel L. Using nominal group technique among resident physicians to identify key attributes of a burnout prevention program. PLoS One 2022;17(3):e0264921. [doi: 10.1371/journal.pone.0264921] [Medline: 35303009]
38. Gallagher M, Hares T, Spencer J, Bradshaw C, Webb I. The nominal group technique: a research tool for general practice? Fam Pract 1993 Mar;10(1):76-81. [doi: 10.1093/fampra/10.1.76] [Medline: 8477899]
39. Humphrey-Murto S, Varpio L, Gonsalves C, Wood TJ. Using consensus group methods such as Delphi and Nominal Group in medical education research. Med Teach 2017 Jan;39(1):14-19. [doi: 10.1080/0142159X.2017.1245856] [Medline: 27841062]
40. Chan S, Erikainen S. What's in a name? The politics of "Precision Medicine". Am J Bioeth 2018 Apr;18(4):50-52. [doi: 10.1080/15265161.2018.1431324] [Medline: 29621451]
41. Kelly TD, Hawk KF, Samuels EA, Strayer RJ, Hoppe JA. Improving uptake of emergency department-initiated buprenorphine: barriers and solutions. West J Emerg Med 2022 Jul 11;23(4):461-467. [doi: 10.5811/westjem.2022.2.52978] [Medline: 35980414]
42. Sedig L. What's the role of autonomy in patient- and family-centered care when patients and family members don't agree? AMA J Ethics 2016 Jan 1;18(1):12-17. [doi: 10.1001/journalofethics.2016.18.1.ecas2-1601] [Medline: 26854631]
43. Entwistle VA, Carter SM, Cribb A, McCaffery K. Supporting patient autonomy: the importance of clinician-patient relationships. J Gen Intern Med 2010 Jul;25(7):741-745. [doi: 10.1007/s11606-010-1292-2] [Medline: 20213206]
44. Lewis J, Holm S. Organoid biobanking, autonomy and the limits of consent. Bioethics 2022 Sep;36(7):742-756. [doi: 10.1111/bioe.13047] [Medline: 35531912]
45. Caulfield T, Murdoch B. Genes, cells, and biobanks: Yes, there's still a consent problem. PLoS Biol 2017 Jul;15(7):e2002654. [doi: 10.1371/journal.pbio.2002654] [Medline: 28742850]
46. Ashcroft JW, Macpherson CC. The complex ethical landscape of biobanking. Lancet Public Health 2019 Jun;4(6):e274-e275. [doi: 10.1016/S2468-2667(19)30081-7] [Medline: 31155216]
47. Saben JL, Shelton SK, Hopkinson AJ, et al. The Emergency Medicine Specimen Bank: An innovative approach to biobanking In acute care. Acad Emerg Med 2019 Jun;26(6):639-647. [doi: 10.1111/acem.13620] [Medline: 30239069]

XSL•FO
RenderX

48. Naess AC, Foerde R, Steen PA. Patient autonomy in emergency medicine. Med Health Care Philos 2001;4(1):71-77. [doi: 10.1023/a:1009982710625] [Medline: 11315422]
49. American College of Emergency Physicians. Code of ethics for emergency physicians. Ann Emerg Med 2008 Nov;52(5):581-590. [doi: 10.1016/j.annemergmed.2008.07.041] [Medline: 18970998]
50. Geiderman JM, Moskop JC, Derse AR. Privacy and confidentiality in emergency medicine: obligations and challenges. Emerg Med Clin North Am 2006 Aug;24(3):633-656. [doi: 10.1016/j.emc.2006.05.005] [Medline: 16877134]
51. Lee SSJ, Cho MK, Kraft SA, et al. “I don’t want to be Henrietta Lacks”: diverse patient perspectives on donating biospecimens for precision medicine research. Genet Med 2019 Jan;21(1):107-113. [doi: 10.1038/s41436-018-0032-6]
52. Ratcliff CL, Kaphingst KA, Jensen JD. When personal feels invasive: foreseeing challenges in precision medicine communication. J Health Commun 2018;23(2):144-152. [doi: 10.1080/10810730.2017.1417514] [Medline: 29279048]
53. Stiles D, Appelbaum PS. Cases in precision medicine: concerns about privacy and discrimination after genomic sequencing. Ann Intern Med 2019 May 21;170(10):717. [doi: 10.7326/M18-2666]
54. Sankar PL, Parker LS. The Precision Medicine Initiative’s All of Us Research Program: an agenda for research on its ethical, legal, and social issues. Genet Med 2017 Jul;19(7):743-750. [doi: 10.1038/gim.2016.183] [Medline: 27929525]
55. Moss AH, Siegler M. Should alcoholics compete equally for liver transplantation? JAMA 1991 Mar 13;265(10):1295-1298. [Medline: 1995977]
56. Sedki M, Ahmed A, Goel A. Ethical and allocation issues in liver transplant candidates with alcohol related liver disease. Transl Gastroenterol Hepatol 2022;7:26. [doi: 10.21037/tgh-2020-13] [Medline: 35892052]
57. Kraft SA, Cho MK, Gillespie K, et al. Beyond consent: building trusting relationships with diverse populations in precision medicine research. Am J Bioeth 2018 Apr;18(4):3-20. [doi: 10.1080/15265161.2018.1431322] [Medline: 29621457]
58. Zink BJ. Social justice, egalitarianism, and the history of emergency medicine. Virtual Mentor 2010 Jun 1;12(6):492-494. [doi: 10.1001/virtualmentor.2010.12.6.mhst1-1006] [Medline: 23158453]
59. Anderson ES, Lippert S, Newberry J, Bernstein E, Alter HJ, Wang NE. Addressing social determinants of health from the emergency department through social emergency medicine. West J Emerg Med 2016 Jul;17(4):487-489. [doi: 10.5811/westjem.2016.5.30240] [Medline: 27429706]
60. Ben-Assuli O. Electronic health records, adoption, quality of care, legal and privacy issues and their implementation in emergency departments. Health Policy 2015 Mar;119(3):287-297. [doi: 10.1016/j.healthpol.2014.11.014] [Medline: 25483873]
61. Brown HL, Brown TB. EMTALA: The evolution of emergency care in the United States. J Emerg Nurs 2019 Jul;45(4):411-414. [doi: 10.1016/j.jen.2019.02.002] [Medline: 30902349]
62. Boulware LE, Purnell TS, Mohottige D. Systemic kidney transplant inequities for black individuals: examining the contribution of racialized kidney function estimating equations. JAMA Netw Open 2021 Jan 4;4(1):e2034630. [doi: 10.1001/jamanetworkopen.2020.34630] [Medline: 33443577]
63. Erdmann A, Rehmann-Sutter C, Bozzaro C. Patients’ and professionals’ views related to ethical issues in precision medicine: a mixed research synthesis. BMC Med Ethics 2021 Aug 31;22(1):116. [doi: 10.1186/s12910-021-00682-8] [Medline: 34465328]
64. Miller AM, Garfield S, Woodman RC. Patient and Provider Readiness for Personalized Medicine. Personalized Medicine in Oncology. URL: https://www.personalizedmedonc.com/articles/2816:patient-and-provider-readiness-for-personalized-medicine [accessed 2025-01-29]
65. Zhang X, Carabello M, Hill T, Bell SA, Stephenson R, Mahajan P. Trends of racial/ethnic differences in emergency department care outcomes among adults in the United States from 2005 to 2016. Front Med (Lausanne) 2020;7:300. [doi: 10.3389/fmed.2020.00300] [Medline: 32671081]
66. Mersha TB, Abebe T. Self-reported race/ethnicity in the age of genomic research: its potential impact on understanding health disparities. Hum Genomics 2015 Jan 7;9(1):1. [doi: 10.1186/s40246-014-0023-x] [Medline: 25563503]
67. Manrai AK, Patel CJ, Ioannidis JPA. In the era of precision medicine and big data, who is normal? JAMA 2018 May 15;319(19):1981-1982. [doi: 10.1001/jama.2018.2009] [Medline: 29710130]
68. O’Sullivan BP, Orenstein DM, Milla CE. Pricing for orphan drugs: will the market bear what society cannot? JAMA 2013 Oct 2;310(13):1343-1344. [doi: 10.1001/jama.2013.278129] [Medline: 24084916]
69. Ferkol T, Quinton P. Precision medicine: at what price? Am J Respir Crit Care Med 2015 Sep 15;192(6):658-659. [doi: 10.1164/rccm.201507-1428ED] [Medline: 26207804]
70. Ramaswami R, Bayer R, Galea S. Precision medicine from a public health perspective. Annu Rev Public Health 2018 Apr 1;39(153–68):153-168. [doi: 10.1146/annurev-publhealth-040617-014158] [Medline: 29166244]
71. Geruso M, Jena AB, Layton TJ. Will personalized medicine mean higher costs for consumers? Harvard Business Review. URL: https://hbr.org/2018/03/will-personalized-medicine-mean-higher-costs-for-consumers [accessed 2025-01-17]
72. Horgan D, Jansen M, Leyens L, et al. An index of barriers for the implementation of personalised medicine and pharmacogenomics in Europe. Public Health Genomics 2014;17(5-6):287-298. [doi: 10.1159/000368034] [Medline: 25401385]

XSL•FO
RenderX

73. Fiore RN, Goodman KW. Precision medicine ethics: selected issues and developments in next-generation sequencing, clinical oncology, and ethics. Curr Opin Oncol 2016 Jan;28(1):83-87. [doi: 10.1097/CCO.0000000000000247] [Medline: 26569425]
74. Fullerton SM, Anderson NR, Guzauskas G, Freeman D, Fryer-Edwards K. Meeting the governance challenges of next-generation biorepository research. Sci Transl Med 2010 Jan 20;2(15):20371468. [doi: 10.1126/scitranslmed.3000361] [Medline: 20371468]

Abbreviations

CHD1: chromodomain helicase DNA binding protein 1
COREQ: Consolidated Criteria for Reporting Qualitative Research
CYP: cytochrome P450
ESI: Emergency Severity Index
NGT: nominal group technique
NHAMCS: National Hospital Ambulatory Medical Care Survey

Edited by A Mavragani; submitted 04.11.24; peer-reviewed by M Siermann, U Khan; revised version received 18.12.24; accepted 18.12.24; published 06.02.25.

Please cite as:
Rose C, Shearer E, Woller I, Foster A, Ashenburg N, Kim I, Newberry J
Identifying High-Priority Ethical Challenges for Precision Emergency Medicine: Nominal Group Study
JMIR Form Res 2025;9:e68371
URL: https://formative.jmir.org/2025/1/e68371
doi:10.2196/68371

© Christian Rose, Emily Shearer, Isabela Woller, Ashley Foster, Nicholas Ashenburg, Ireh Kim, Jennifer Newberry. Originally published in JMIR Formative Research (https://formative.jmir.org), 6.2.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Impact of Psychological Resilience on the Fear of Pain and Activity Recovery in Postsurgical Patients: Observational Cohort Study

Yang Luo[1*], MN; Sisi Li[2*]; Lijuan Feng[1]; Junyi Zheng[1]; Chunfen Peng[3]; Lihong Bao[1]

[1]Department of Thoracic Surgery, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, 1277 Jiefang road, Jianghan District, Wuhan, China

[2]Department of Colorectal and Anal Surgery, Zhongnan Hospital of Wuhan University, Wuhan, China

[3]Cancer Center, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China

[*]these authors contributed equally

Corresponding Author:
Lihong Bao
Department of Thoracic Surgery, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, 1277 Jiefang road, Jianghan District, Wuhan, China

Abstract

Background: The fear of pain (FOP) and fear avoidance belief (FAB) play a crucial role in the occurrence and development of chronic pain. However, the dynamics of these factors in postsurgical pain appear to differ, with the FOP often learned from others rather than directly caused by painful experiences. Psychological resilience refers to an individual's capacity to adapt effectively to adversity, challenges, and threats, and may play a significant role in overcoming the FOP and avoidance behavior.

Objective: The aim of this study was to investigate the role of psychological resilience in overcoming the FOP and avoidance behavior among surgical patients undergoing lung surgery.

Methods: Participants were recruited at the Wuhan Union Hospital. Psychological resilience was measured using the Connor-Davidson Resilience Scale. The FOP was assessed using the simplified Chinese version of the Fear of Pain-9 items. The FAB was measured using the Physical Activity subscale of the Fear-Avoidance Beliefs Questionnaire. Activity recovery was assessed through questions related to social activities and household responsibilities. The adaptive least absolute shrinkage and selection operator (Lasso) regression analysis under nested cross-validation was used to identify key factors affecting postoperative FOP and activity recovery.

Results: A total of 144 participants were included in the final analysis. The results showed that preoperative FOP (coefficient=8.620) and FAB (coefficient=8.560) were mainly positively correlated with postoperative FOP, while psychological resilience (coefficient=−5.822) and age (coefficient=−2.853) were negatively correlated with it. These average R^2 of these models was 73% (SD 6%). Psychological resilience was the most important factor in predicting activity recovery, and these models obtained an average accuracy of 0.820 (SD 0.024) and an average area under the curve of 0.926 (SD 0.044).

Conclusions: Psychological resilience was negatively associated with the postoperative FOP and positively with activity recovery in patients who underwent lung surgery. Patients with higher resilience are more likely to cope effectively with postsurgical pain and recover activities sooner. These findings highlight the importance of assessing and potentially enhancing psychological resilience in the perioperative period to improve postoperative outcomes.

(JMIR Form Res 2025;9:e63556) doi:10.2196/63556

KEYWORDS

psychological resilience; fear of pain; postsurgical recovery; fear avoidance belief; pain management; activity recovery

Introduction

Psychological factors such as the fear of pain (FOP) and fear avoidance belief (FAB) are thought to play an important role in the occurrence and development of chronic pain [1-3]. Based on the fear-avoidance model, among patients with chronic pain, painful experiences from activities will lead to the FOP, and then cause patients to refuse activities until disability, which in turn can lower the pain threshold and thus more easily lead to painful experiences after activities, forming a vicious cycle of fear reinforcement [4,5].

However, this seems to be different in the case of postsurgical pain. Our previous research [6] has shown that the FOP is related to postsurgical pain and can lead to avoidance behavior. A significant portion of this FOP is learned from others' teaching or observational learning, while the FOP being directly caused by painful experiences is relatively rare. This seems to indicate that some people refuse to engage in certain activities such as

XSL•FO
RenderX

coughing and ambulation from the beginning based on their cognition, rather than pain experience, as they have heard others tell them that such activities are dangerous.

A good pain experience after activities can help break the cycle of the FOP. The fear-avoidance model does not explain the dynamics underlying the FOP and functional recovery [4], such as how some postoperative patients can still continue to perform rehabilitation activities despite pain. It seems that there is an internal drive that prompts patients to overcome their fears and produce positive recovery behaviors.

Psychological resilience, also known as mental toughness [7], refers to an individual's capacity to adapt effectively to adversity, challenges, and threats [8]. Psychological resilience has a positive impact in patients undergoing knee surgery and those with pain catastrophizing [9], as well as its predictive role in dental phobia [10]. However, its relationship with postsurgical pain and the FOP remains understudied.

Individuals with low psychological resilience tend to avoid stressful situations [11]. In this study, we aimed to understand the role of psychological resilience in overcoming the FOP and avoidance behavior among surgical patients. To this end, we followed up a sample from a previous study [12] to observe the impact of psychological resilience on the FOP and behavior, aiming to assist in postsurgical pain management.

Methods

Participants

This study was a single-center, cohort study of patients who were hospitalized to undergo thoracic surgery at the Wuhan Union Hospital, a tertiary hospital located in Wuhan, China, from May 2022 to January 2023. Participants were recruited based on the following inclusion criteria: (1) they were at least 18 years old; (2) they had undergone lung surgery; and (3) they did not have a history of neurological or psychiatric disorders. The exclusion criteria included individuals who experienced critical conditions during the study period or those who did not complete follow-up assessments. In the context where the sample size exceeds 10 events per variable [13], and considering a potential 10% dropout rate, we would terminate the further enrollment of participants.

Ethical Considerations

This study was a follow-up study based on earlier research [12] (Registration number: ChiCTR2200056651), and has been approved by the institutional ethics board of Wuhan Union Hospital of Tongji Medical College, Huazhong University of Science and Technology (No. 20220026‐2). All the participants provided written and oral informed consent. The original informed consent form, which was approved by the same ethics committee, explicitly allows for secondary analysis without the need for additional consent (No. 20200351). Participants were provided with free counseling services during the follow-up period as a form of compensation. All study data have been anonymized or deidentified to protect the privacy and confidentiality of the participants.

Fear of Pain

Before surgery and 6 months after surgery, we used the simplified Chinese version of the Fear of Pain-9 items (FOP-9) [14] to gather data on the FOP. The FOP-9 is a 5-point ordinal Likert scale where total scores range from 9 to 45, and a higher score signifies a greater level of FOP.

Fear Avoidance Belief

On the third day after surgery, the Physical Activity subscale of the Fear-Avoidance Beliefs Questionnaire (FABQ-PA) was used to measure the FAB. The FABQ-PA is a self-reported questionnaire comprising the first 5 questions of the original Fear-Avoidance Beliefs Questionnaire [15]. A higher score on this subscale signifies a greater tendency towards the FAB.

Psychological Resilience

Psychological resilience was collected 1 month after surgery by using the Connor-Davidson Resilience Scale (CD-RISC) [16]. The CD-RISC comprises 25 items, each rated on a 5-point scale (0‐4), with higher scores reflecting greater psychological resilience.

Activity Recovery

Activity recovery was defined as the recovery of social activities and housework after discharge. Activity recovery was assessed at 6 months after surgery through questions such as, "Are your current social activities, like work and community activities, similar to what they were before the surgery?" and "Have your current family responsibilities returned to what they were before the surgery?" The answers were categorized as being lower or similar/higher than they were before surgery. Both the answers of "similar/higher" meant that the activity recovery was positive.

Statistics

We used the adaptive least absolute shrinkage and selection operator (Lasso) with nested cross-validation to conduct variable selection, coefficient estimation, and performance prediction. All the statistics were completed using Python 3.12 (Python Software Foundation) and the scikit-learn library (developed by a community of open-source contributors, initially started by David Cournapeau). The process was as follows [17]: (1) Feature preprocessing: We conducted standardization on numeric features and one-hot encoding on categorical features. In order to avoid the problem of multicollinearity, we set the first one as the baseline and dropped it for each categorical variable, such as "Male" in sex, "Primary School" in education, "Presence" in chronic pain, and "Wedge Resection" in the surgery method. (2) Model training: We estimated the initial coefficients using Ridge regression with 5-fold cross-validation, and then conducted model training with the adaptive Lasso with 5-fold cross-validation after weighing the coefficients. (3) Model evaluation: In the 5-fold outer cross-validation, we selected the optimal model and calculated the average value of the model fitting parameters. (4) Variable estimation: We reported the regularization parameters and variable coefficients of the optimal model.

XSL•FO
RenderX

Results

Characteristics

A total of 150 patients met the study's inclusion criteria. However, 6 (4%) patients were excluded as they were lost to follow-up. A total of 144 participants were included in the final analysis, and the baseline characteristics are listed in Table 1. The study flow is shown in Figure 1.

Table . Baseline patient characteristics (n=144).

Characteristics	Value
Age (years), mean (SD)	52.40 (13.37)
Sex, n (%)	
Male	73 (50.7)
Female	71 (49.3)
Education, n (%)	
Primary school	20 (13.9)
Middle school	30 (20.8)
High school	43 (29.9)
University	51 (35.4)
Chronic pain, n (%)	
Absence	137 (95.1)
Presence	7 (4.9)
Preoperative FOP[a], median (IQR)	16 (13-23)
Surgery method, n (%)	
Wedge resection	24 (16.7)
Segmental resection	64 (44.4)
Lobectomy	56 (38.9)
Pain intensity, median (IQR)	3 (1-4)
Fear avoidance belief, median (IQR)	11 (8-15)
Psychological resilience, mean (SD)	49.54 (12.99)
Activity recovery, n (%)	
Lower	38 (26.4)
Similar/Higher	106 (73.6)
Postoperative FOP, median (IQR)	20 (14-28)

[a]FOP: fear of pain.

XSL•FO
RenderX

Figure 1. Flowchart of the research process and data collection. FOP: fear of pain; FAB: fear avoidance belief.

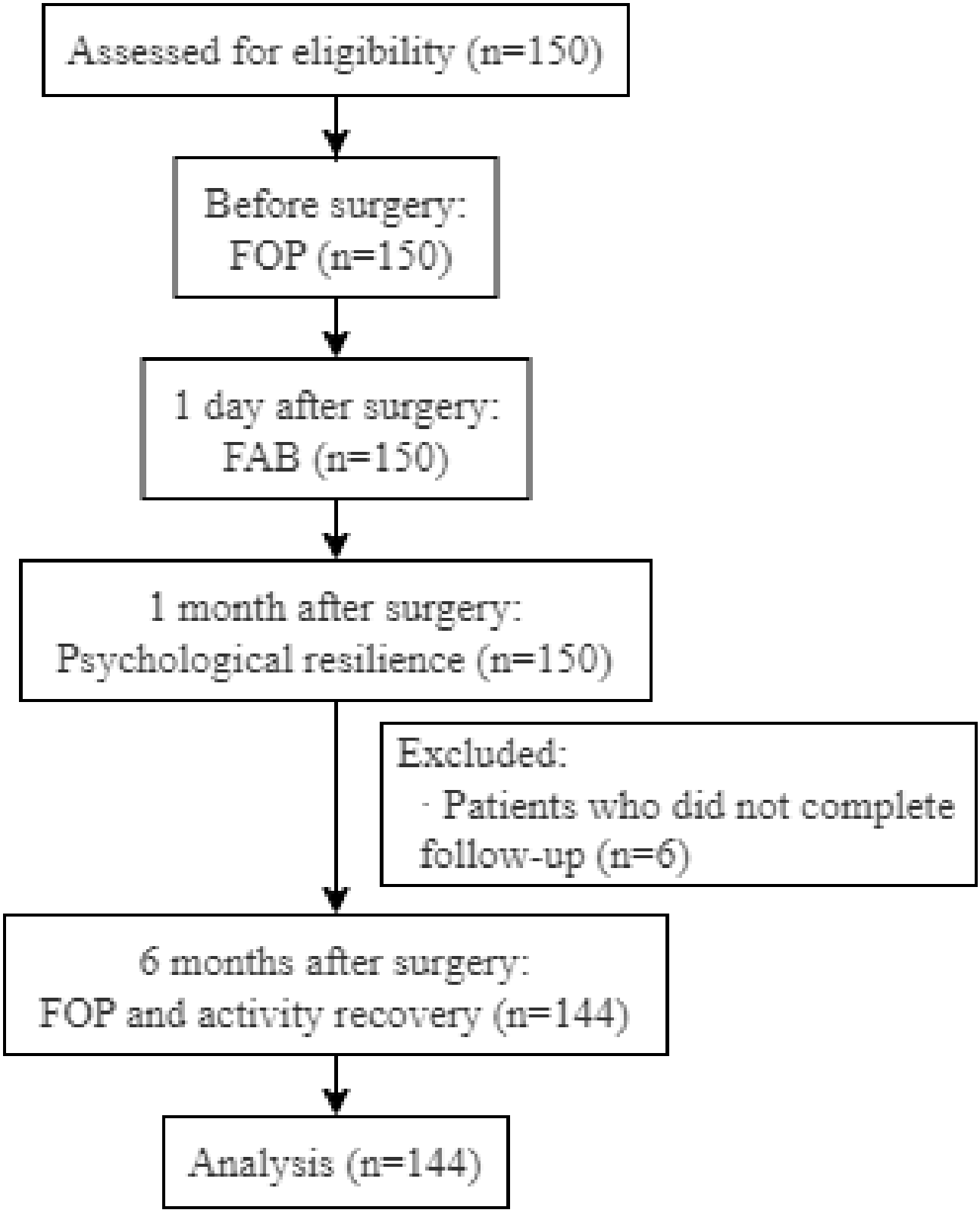

Postoperative FOP as the Dependent Variable

The statistical analysis showed that the mean squared error was 14.273 (SD 3.254), indicating a reasonable prediction error range with room for improvement. The mean R^2 value was 0.731 (SD 0.060), suggesting that the models could account for approximately 73.1% (SD 6%) of the variance in postoperative FOP and had a relatively good fit. The optimal alpha of the optimal model was 0.081. The coefficient of the optimal model revealed that the preoperative FOP, FAB, and pain intensity had significant positive effects on the postoperative FOP with coefficients of 8.620, 8.560, and 0.417, respectively. In contrast, psychological resilience and age had significant negative impacts, with coefficients of −5.822 and −2.853, respectively. Among categorical variables, patients with a university education had lower postoperative FOP compared to those with

XSL•FO
RenderX

primary school education, while female patients had a higher postoperative FOP than male patients did. The surgical method and chronic pain had no significant effect on the postoperative FOP (Table 2).

Table . Optimal model of adaptive least absolute shrinkage and selection operator (Lasso) regression for patients who underwent lung surgery in predicting the postoperative fear of pain (FOP; n=144).

Independent variables[a]	Coefficient
Age	–2.853
Sex	
Male	0.000
Female	0.778
Education	
Primary school	0.000
Middle school	−0.000
High school	0.000
University	−0.190
Chronic pain	
Absent	0.000
Present	0.000
Preoperative FOP	8.620
Surgery method	
Wedge resection	0.000
Segmental resection	-0.005
Lobectomy	0.000
Pain intensity	0.417
Fear avoidance belief	8.560
Psychological resilience	−5.822

[a]Lambda is 0.081 for all variables.

Activity Recovery as the Dependent Variable

The model performed excellently after analysis, with an average accuracy of 0.820 (SD 0.024) and an average area under the curve of 0.926 (SD 0.044), indicating that the model had a good predictive ability and stability for activity recovery. Psychological resilience, FAB, and pain intensity were significant factors that affected the activity recovery, with coefficients of 1.185, −0.966, and −0.170, respectively. Both university education and segmental resection significantly positively affected activity recovery compared to baseline, with coefficients of 0.283 and 0.042, respectively. Age, sex, chronic pain, and preoperative FOP had no significant effect on activity recovery (Table 3).

XSL•FO
RenderX

Table . Optimal model of adaptive least absolute shrinkage and selection operator (Lasso) logistic regression for patients who underwent lung surgery in predicting activity recovery (n=144).

Independent variables[a]	Coefficient
Age	0.000
Sex	
Male	0.000
Female	0.000
Education	
Primary school	0.000
Middle school	0.000
High school	0.000
University	0.283
Chronic pain	
Absent	0.000
Present	0.000
Preoperative FOP[b]	0.000
Surgery method	
Wedge resection	0.000
Segmental resection	0.042
Lobectomy	0.000
Pain intensity	−0.170
Fear avoidance belief	−0.966
Psychological resilience	1.185

[a]Lambda is 2.783 for all variables.

[b]FOP: fear of pain.

Discussion

Study Findings

This study delved into the role of psychological resilience in perioperative FOP and avoidance behavior among surgical patients, particularly those undergoing lung surgery. The findings contribute novel insights to the existing literature, suggesting that psychological resilience may serve as a crucial factor in breaking the vicious cycle of FOP and facilitating recovery.

Impact of Psychological Resilience on Postoperative FOP

Our study employed adaptive Lasso regression under nested cross-validation to analyze the factors influencing postoperative FOP, a common phenomenon experienced by patients following surgery [18]. The developed model demonstrated robust performance, with a high proportion of variance explained in postoperative FOP. Among the factors, psychological resilience emerged as the most significant inhibitory factor influencing postoperative FOP. Preconceived biases regarding postoperative activities and pain can lead to the development of the FOP [19,20]. We observed that some patients strictly adhered to the doctor's orders despite potential pain, while others rejected and avoided them. This disparity can be attributed to the varying levels of psychological resilience among individuals. Resilient individuals are more likely to accept and face pain, actively adjust their pain concepts, and gradually expose themselves to pain stimuli [21,22]. For instance, they may perceive postoperative pain as a normal part of the recovery process and actively communicate with medical staff to learn pain management strategies, gradually adapting to and overcoming the FOP. Conversely, individuals with lower psychological resilience may avoid stressful situations, including postoperative activities, leading to a vicious cycle of pain, fear, and avoidance [9,10,23].

Impact of Psychological Resilience on Postoperative Activity Recovery

The adaptive Lasso logistic regression exhibited higher accuracy and area under the curve values with a lower standard deviation, indicating that the model has high accuracy and stability in predicting postoperative activity recovery. The fact that the coefficient of psychological resilience is the largest revealed that psychological resilience had a stronger influence on postoperative activity recovery than other factors. This suggests that psychological resilience primarily enables individuals to initiate activities that may cause pain despite the presence of FOP, serving as a positive factor in overcoming fear. Patients

XSL•FO
RenderX

with higher psychological resilience tend to have better ability to resume their activities postoperatively [24,25], and our research has also confirmed this. They may be more actively involved in social interactions and resume daily housework, thereby promoting physical function recovery and improving psychological well-being, resulting in a virtuous cycle. This may be because patients with high psychological resilience are better able to cope with negative emotions, have more reasonable cognitive expectations of the rehabilitation process, and can better overcome the fear of activities, thus promoting the recovery of postoperative activities [26,27].

Study Limitations

This study has several limitations. First, psychological resilience was measured postoperatively, which may introduce bias due to the impact of surgery. Second, although both models demonstrated certain predictive abilities, there is room for improvement in the R^2 values. Third, the generalizability of the findings is limited, as this study specifically focused on patients undergoing lung surgery, and the results may not be applicable to all surgical patients. Therefore, conducting multicenter studies with a bigger sample size would further validate the robustness of the results.

Conclusion

This study revealed the significant role of psychological resilience in postsurgical pain management, particularly in reducing the FOP and promoting recovery of activities. Individuals with higher psychological resilience are more likely to effectively cope with postsurgical pain and reduce avoidance behaviors. These findings underscore the importance of assessing and potentially enhancing psychological resilience during the perioperative period, especially for those with higher levels of FOP. By fostering resilience and promoting correct pain cognition, clinical practitioners can empower patients to overcome the FOP, engage in rehabilitation activities, and achieve better postoperative outcomes.

Acknowledgments

This study was supported by the general program of Hubei Provincial Natural Science Foundation of China (No. 2021CFB588).

Data Availability

The datasets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

References

1. Sugano R, Ikegami K, Ando H, et al. The relationship between fear-avoidance beliefs in employees with chronic musculoskeletal pain and work productivity: a longitudinal study. J UOEH 2020;42(1):13-26. [doi: 10.7888/juoeh.42.13] [Medline: 32213739]
2. Meier ML, Stämpfli P, Humphreys BK, Vrana A, Seifritz E, Schweinhardt P. The impact of pain-related fear on neural pathways of pain modulation in chronic low back pain. Pain Rep 2017 May;2(3):e601. [doi: 10.1097/PR9.0000000000000601] [Medline: 29392216]
3. Luo H, Cai Z, Huang Y, et al. Study on pain catastrophizing from 2010 to 2020: a bibliometric analysis via CiteSpace. Front Psychol 2021;12:759347. [doi: 10.3389/fpsyg.2021.759347] [Medline: 34975649]
4. Crombez G, Eccleston C, Van Damme S, Vlaeyen JWS, Karoly P. Fear-avoidance model of chronic pain: the next generation. Clin J Pain 2012 Jul;28(6):475-483. [doi: 10.1097/AJP.0b013e3182385392] [Medline: 22673479]
5. Varangot-Reille C, Pezzulo G, Thacker M. The fear-avoidance model as an embodied prediction of threat. Cogn Affect Behav Neurosci 2024 Oct;24(5):781-792. [doi: 10.3758/s13415-024-01199-4] [Medline: 38890209]
6. Luo Y, He J, Bao L, Meng H, Hu C, Chen Q. Fear of pain as a predictor for postoperative pain intensity among the patients undergoing thoracoscopic surgery. Pain Res Manag 2022;2022:2201501. [doi: 10.1155/2022/2201501] [Medline: 35757293]
7. Soundara Pandian PR, Balaji Kumar V, Kannan M, Gurusamy G, Lakshmi B. Impact of mental toughness on athlete's performance and interventions to improve. J Basic Clin Physiol Pharmacol 2023 Jul 1;34(4):409-418. [doi: 10.1515/jbcpp-2022-0129] [Medline: 35792085]
8. Bögemann SA, Puhlmann LMC, Wackerhagen C, et al. Psychological resilience factors and their association with weekly stressor reactivity during the COVID-19 outbreak in Europe: prospective longitudinal study. JMIR Ment Health 2023 Oct 17;10:e46518. [doi: 10.2196/46518] [Medline: 37847551]
9. Stamatis A, Morgan GB, Spinou A, Tsigaridis KG. Mental toughness and osteoarthritis: postsurgery improvement in knee pain/functionality in older adults. Rehabil Psychol 2023 May;68(2):212-219. [doi: 10.1037/rep0000499] [Medline: 37126026]
10. Tian H, Ding X, Lin C, Qu P, Fan X. Study of dental phobia, psychological resilience, and related factors in children aged 8-9 years in Weifang city. Hua Xi Kou Qiang Yi Xue Za Zhi 2023 Feb 1;41(1):73-79. [doi: 10.7518/hxkq.2023.01.010] [Medline: 38596944]

XSL•FO
RenderX

11. Babić R, Babić M, Rastović P, et al. Resilience in health and illness. Psychiatr Danub 2020 Sep;32(Suppl 2):226-232. [Medline: 32970640]
12. Bao L, Peng C, He J, Sun C, Feng L, Luo Y. The relationship between fear avoidance belief and threat learning in postoperative patients after lung surgery: an observational study. Psychol Res Behav Manag 2023;16:3259-3267. [doi: 10.2147/PRBM.S420724] [Medline: 37605755]
13. Tabachnick BG, Fidell LS. Using Multivariate Statistics: Pearson Education; 2013. URL: https://books.google.com.sg/books?id=ucj1ygAACAAJ [accessed 2023-12-11]
14. Luo Y, Li S, He J, et al. Translation and validation of fear of pain-9 items into simplified Chinese version for Mainland China. J Pain Res 2021;14:35-40. [doi: 10.2147/JPR.S275227] [Medline: 33469354]
15. Waddell G, Newton M, Henderson I, Somerville D, Main CJ. A fear-avoidance beliefs questionnaire (FABQ) and the role of fear-avoidance beliefs in chronic low back pain and disability. Pain 1993 Feb;52(2):157-168. [doi: 10.1016/0304-3959(93)90127-B] [Medline: 8455963]
16. Connor KM, Davidson JRT. Development of a new resilience scale: the Connor-Davidson Resilience Scale (CD-RISC). Depress Anxiety 2003;18(2):76-82. [doi: 10.1002/da.10113] [Medline: 12964174]
17. Ballout N, Etievant L, Viallon V. On the use of cross-validation for the calibration of the adaptive lasso. Biom J 2023 Jun;65(5):e2200047. [doi: 10.1002/bimj.202200047] [Medline: 36960476]
18. Ustunel F, Erden S. Evaluation of fear of pain among surgical patients in the preoperative period. J Perianesth Nurs 2022 Apr;37(2):188-193. [doi: 10.1016/j.jopan.2021.02.003] [Medline: 34922831]
19. Pak R, Mahmoud Alilou M, Bakhshipour Roudsari A, Yousefpour F. Experiential avoidance as a factor in generalized psychological vulnerability: in the relationship between chronic pain and pain anxiety with pain disability. Pain Manag Nurs 2024 Jun;25(3):e256-e264. [doi: 10.1016/j.pmn.2024.01.007] [Medline: 38418316]
20. Shojaie D, Hoffman AS, Amaku R, et al. Decision making when cancer becomes chronic: needs assessment for a web-based medullary thyroid carcinoma patient decision aid. JMIR Form Res 2021 Jul 16;5(7):e27484. [doi: 10.2196/27484] [Medline: 34269691]
21. Dursun A, Kaplan Y, Altunbaş T, Bahtiyar M. The mediating effect of experiential avoidance on the relationship between psychological resilience and psychological needs in the COVID-19 pandemic. Curr Psychol 2022 May 19;43(14):1-11. [doi: 10.1007/s12144-022-03085-z] [Medline: 35607644]
22. Seçer İ, Ulaş S, Karaman-Özlü Z. The effect of the fear of COVID-19 on healthcare professionals' psychological adjustment skills: mediating role of experiential avoidance and psychological resilience. Front Psychol 2020;11:561536. [doi: 10.3389/fpsyg.2020.561536] [Medline: 33192830]
23. Kartal M, Kapikiran G, Karakas N. The effect of emergency nurses' psychological resilience on their thanatophobic behaviors: a cross-sectional study. Omega (Westport) 2024 Aug;89(3):1273-1285. [doi: 10.1177/00302228221128156] [Medline: 36117453]
24. The REGAIN (Regional versus General Anesthesia for Promoting Independence after Hip Fracture) Investigators. Preoperative psychological resilience and recovery after hip fracture: secondary analysis of the regain randomized trial. J American Geriatrics Society 2023 Dec;71(12):3792-3801. [doi: 10.1111/jgs.18552]
25. Berthold DP, Bormann M. Editorial commentary: psychological resilience promotes positive outcomes after surgery. Arthroscopy: The Journal of Arthroscopic & Related Surgery 2024 Dec;40(12):2895-2896. [doi: 10.1016/j.arthro.2024.04.004]
26. Al Ta'ani Z, Al Ta'ani O, Gabr A, et al. From fear to resilience: a scoping review of psychological components in anterior cruciate ligament rehabilitation. J Sport Rehabil 2024 Nov 1;33(8):591-618. [doi: 10.1123/jsr.2023-0420] [Medline: 39299682]
27. Broche-Perez Y, Jimenez-Morales RM, Vázquez-Gómez LA, Bauer J, Fernández-Fleites Z. Fear of relapse and quality of life in multiple sclerosis: the mediating role of psychological resilience. Mult Scler Relat Disord 2023 Nov;79:105026. [doi: 10.1016/j.msard.2023.105026] [Medline: 37839366]

Abbreviations

CD-RISC: Connor-Davidson Resilience Scale
FAB: fear avoidance belief
FABQ-PA: Physical Activity subscale of the Fear-Avoidance Beliefs Questionnaire
FOP: fear of pain
FOP-9: Fear of Pain-9 items
Lasso: least absolute shrinkage and selection operator

XSL•FO
RenderX

Edited by A Mavragani; submitted 23.06.24; peer-reviewed by KE Cain, P Humburg; revised version received 18.12.24; accepted 19.12.24; published 07.02.25.

Please cite as:
Luo Y, Li S, Feng L, Zheng J, Peng C, Bao L
Impact of Psychological Resilience on the Fear of Pain and Activity Recovery in Postsurgical Patients: Observational Cohort Study
JMIR Form Res 2025;9:e63556
URL: https://formative.jmir.org/2025/1/e63556
doi:10.2196/63556

© Yang Luo, Sisi Li, Lijuan Feng, Junyi Zheng, Chunfen Peng, Lihong Bao. Originally published in JMIR Formative Research (https://formative.jmir.org), 7.2.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Impact of Preventive Intravenous Amiodarone on Reperfusion Ventricular Fibrillation in Patients With Left Ventricular Hypertrophy Undergoing Open-Heart Surgery: Randomized Controlled Clinical Trial

Chan-Juan Gong*, MD; Xiao-Kai Zhou*, MD; Zhen-Feng Zhang, MD; Yin Fang, MD

Department of Anesthesiology and Perioperative Medicine, The First Affiliated Hospital of Nanjing Medical University, No. 300 Guangzhou Road, Nanjing, China

*these authors contributed equally

Corresponding Author:
Yin Fang, MD
Department of Anesthesiology and Perioperative Medicine, The First Affiliated Hospital of Nanjing Medical University, No. 300 Guangzhou Road, Nanjing, China

Abstract

Background: Ventricular fibrillation (VF) is a vicious arrhythmia usually generated after removal of the aortic cross-clamp (ACC) in patients undergoing open-heart surgery, which could damage cardiomyocytes, especially in patients with left ventricular hypertrophy (LVH). Amiodarone has the prominent properties of converting VF and restoring sinus rhythm. However, few studies concentrated on the effect of amiodarone before ACC release on reducing VF in patients with LVH.

Objective: The study was designed to explore the effectiveness of prophylactic intravenous amiodarone in reducing VF after the release of the ACC in patients with LVH.

Methods: A total of 54 patients with LVH scheduled for open-heart surgery were enrolled and randomly divided (1:1) into 2 groups—group A (amiodarone group) and group P (placebo-controlled group). Thirty minutes before removal of the ACC, the trial drugs were administered intravenously. In group A, 150 mg of amiodarone was pumped in 15 minutes. In group P, the same volume of normal saline was pumped in 15 minutes. The primary outcome was the incidence of VF 10 minutes after removal of the ACC.

Results: The incidence of VF was lower in group A than in group P (30% vs 70%, P=.003). The duration of VF, the number of defibrillations, and the defibrillation energy were also lower in group A than in group P (P<.001, P=.002, and P=.002, respectively). After the end of cardiopulmonary bypass, the heart rate and mean arterial pressure were lower in group A, and the mean pulmonary arterial pressure and the dose of vasoactive drugs were higher than those in group P (P<.001, P<.001, P=.04, and P=.02, respectively). However, there were no significant differences in the use of vasoactive-inotropic agents and hemodynamic status between the 2 groups before the end of surgery.

Conclusions: In patients with LVH who undergo open-heart surgery, amiodarone can be safely used to reduce the incidence of VF, the duration of VF, the frequency of defibrillation, and the energy of defibrillation after ACC removal.

Trial Registration: Chinese Clinical Trial Registry ChiCTR2000035057; https://www.chictr.org.cn/showprojEN.html?proj=57145

(JMIR Form Res 2025;9:e64586) doi:10.2196/64586

KEYWORDS

amiodarone; left ventricular hypertrophy; reperfusion ventricular fibrillation; open-heart surgery; randomized controlled trial; RCT; clinical trial; ventricular fibrillation; vicious arrhythmia; aortic cross-clamp; surgery; effectiveness; defibrillation

Introduction

Ventricular fibrillation (VF) is commonly seen after release of the aortic cross-clamp (ACC) during open-heart surgery, which can result in reperfusion VF (rVF) when reperfusion of the myocardium begins. The incidence of rVF after the ACC release has been reported to range from 10% to 80% [1]. It has been proven in the animal model that persistent VF could cause more myocardial fibrosis [2]. The use of internal electrical defibrillation is a successful approach to terminate VF. However, multiple shocks and prolonged VF duration may exacerbate myocardial damage and diminish the patient's prognosis [3]. Besides, repeated defibrillations would increase the incidence of postoperative bradycardia in cardiac surgeries [4]. Therefore, it is necessary to reduce the rVF after removal of the ACC during cardiac surgery.

XSL•FO
RenderX

In patients with left ventricular hypertrophy (LVH), the unique myocardial structure hinders the complete perfusion of fluid during cardiopulmonary bypass (CPB) surgery to reach the subendocardial cardiomyocytes. This limitation diminishes the efficacy of myocardial protection in open-heart surgery [5]. Following the release of the ACC, there is a notable rise in the occurrence of rVF, leading to an escalation in both the energy and frequency of cardiac shocks within the hypertrophic myocardium [6]. Thus, it is also of vital clinical significance to decrease the occurrence of rVF in patients with LVH.

Amiodarone, known for its classification as a potent Class III antiarrhythmic drug with diverse electrophysiological properties, has been extensively used in clinical practice for the treatment of atrial and ventricular arrhythmias [7,8]. Intravenous amiodarone has been confirmed to protect the myocardium from ischemia-reperfusion injury and prevent reperfusion-induced VF [9]. It also showed a good hemodynamic tolerance when administered intravenously [10]. Previous studies have demonstrated the efficacy of amiodarone in reducing the occurrence of VF after the removal of the ACC [11,12]. However, there is only 1 clinical trial investigating the role of amiodarone in reducing rVF in patients with LVH [13]. Therefore, we designed the study to investigate the effects of prophylactic intravenous amiodarone on rVF after the release of the ACC in patients with LVH undergoing open-heart surgery.

Methods

Settings and Patients

This was a prospective, double-blind, randomized controlled trial conducted in the First Affiliated Hospital of Nanjing Medical University, Nanjing, China. The patients undergoing open-heart surgery were screened by preoperative transthoracic echocardiography (TTE). When meeting the eligible criteria, the patients were recruited in the study. A total of 54 patients with LVH undergoing open-heart surgery were enrolled and participated in this study from August 2022 to November 2023. This manuscript adheres to all applicable CONSORT (Consolidated Standards of Reporting Trials) guidelines (Checklists 1 and 2). The protocol of this study has already been published [14]. The eligibility criteria are hence described.

The inclusion criteria included (1) patients undergoing open-heart surgery, (2) aged between 18 and 75 years, (3) diagnosed with LVH evaluated by TTE preoperatively, and (4) classified as American Society of Anesthesiology (ASA) Physical Status I-III. LVH was characterized by a left ventricular mass index>115 g/m^2 for men and >95 g/m^2 for women [15].

The exclusion criteria included preoperative administration of Class I or III antiarrhythmic medications; QT interval of 500 milliseconds or higher or corrected QT interval of 480 milliseconds or higher; left ventricular ejection fraction below 40%; and presence of thyroid dysfunction, electrolyte imbalances, liver or renal dysfunction, pulmonary interstitial fibrosis, or known allergy to amiodarone. Patients undergoing emergency surgery, redo surgery, or concomitant procedures, including coronary artery bypass grafting (CABG), were also excluded.

Ethical Considerations

This study was approved by the Ethical Committee of the First Affiliated Hospital of Nanjing Medical University (2020-SR-175) and registered in the Chinese Clinical Trial Registry (ChiCTR2000035057). All participants provided verbal informed consent before being enrolled in the study. To secure and protect all participant information, all data were recorded by the specific ID number instead of the participant's name throughout the study. Only select research team members could access these data. The process was monitored by the Ethical Committee of the First Affiliated Hospital of Nanjing Medical University.

Randomization and Blinding

After anesthetic induction, the enrolled participants were assigned with a random permuted block method in a 1:1 ratio to 2 groups, namely, group A (treated with amiodarone) and group P (placebo control). An unblinded nurse anesthetist prepared the drugs according to the allocation information in the randomization envelope. The drugs were uniformly packaged, numbered, and administered by anesthesiologists. All involved parties, including the anesthesiologists, surgeons, and participants, were unaware of the specific treatment regimen, ensuring a blinded study design. However, nonblinded nurse anesthetists were not involved in evaluating treatment interventions, analyzing data, or interpreting results in the study.

Treatment Regimen of Amiodarone

The trial drugs were administered 30 minutes before ACC release through the central venous line. In group A, 150 mg of amiodarone was diluted to 20 ml of normal saline and pumped in 15 minutes. In group P, 20 ml of normal saline was pumped in 15 minutes.

Anesthesia Protocols

All patients received standard general anesthesia. In the operating room, patients were routinely monitored, including 5-lead electrocardiograph, peripheral blood oxygen saturation, invasive arterial blood pressure, bispectral index, and end-tidal CO_2. Anesthesia induction was carried out as follows: 0.05 mg/kg midazolam, 0.3 mg/kg etomidate, 0.15 mg/kg cis-atracurium, and 4 - 6 µg/kg fentanyl. Approximately 5 minutes after induction, patients were intubated and ventilated to maintain end-tidal CO_2 at 35 - 45 mm Hg. Then, ultrasound-guided central vein catheterization was performed through the right internal jugular or subclavian vein. The Swan-Gans catheter was floated to the pulmonary artery through the right internal jugular vein. The patients continued to receive anesthesia by inhalation of sevoflurane 1% - 2% and continuous pump of propofol 1 - 2 mg/kg/h, dexmedetomidine 0.2 - 0.7 µg/kg/min, and cis-atracurium 0.05 - 0.1 mg/kg/h. The total dose of fentanyl was 30 - 50 µg/kg. The bispectral index was maintained between 40 and 60. Inotropic agents, vasopressors, vasodilators, or fluid administration were applied to achieve the following hemodynamic status: a mean arterial pressure

XSL•FO
RenderX

(MAP) of 50 - 80 mm Hg, cardiac index (CI) ≥2.2 L/min/m^2, and urine output ≥1.0 mL/kg/h.

Surgical and CPB Procedures

The surgical procedures were performed by median sternotomy with heparinization and standard CPB. Anticoagulation was started with 400 U/kg of heparin and maintained according to an activated clotting time of >480 seconds. The CPB was initiated using a membrane oxygenator and the circuit was primed with 1500 ml of Ringer acetate solution and 200 ml of 50% human serum albumin. The patients were perfused with a nonpulsatile roller pump at a flow rate of 2.0 - 2.4 L/min/m^2 and with MAP of 50 - 80 mm Hg. Moderate systemic hypothermia was maintained at 30 - 32°C during CPB. Myocardial protection was achieved by intermittent antegrade or retrograde infusions of histidine-tryptophane-ketoglutarate cardioplegia. After weaning from CPB, the reversal of heparin was achieved with protamine.

Management of VF

Thirty seconds before ACC release, 1 mg/kg of esmolol was administered to all patients. When VF occurred after ACC release, internal electrical defibrillation with 20 J was performed in the first attempt. If not successful, the second defibrillation with 20 J was attempted after reusing 0.5 - 1 mg/kg of esmolol. If VF persisted, the third 30 J defibrillation was performed after giving 1 mg/kg of lidocaine. If termination of VF was not achieved, the fourth defibrillation with 40 J was attempted after the administration of 50 mg of amiodarone administration. If VF still existed, the administration of 100 mg of amiodarone and defibrillation with 50 J was repeated until VF terminated.

Outcome Measures

The primary outcome was the incidence of VF 10 minutes after ACC release. The secondary outcomes included the duration of VF, the number and total energy of defibrillations needed to terminate VF, and the use of other antiarrhythmic agents and epicardial temporary pacemakers after the release of the ACC. Furthermore, MAP, heart rate (HR), mean pulmonary arterial pressure (MPAP), central venous pressure (CVP), and CI were measured 15 minutes after anesthetic induction (T1), 15 minutes after weaning from CPB (T2), and 15 minutes before the end of surgery (T3). In addition, the vasoactive-inotropic score (VIS) and the total dose of vasoactive and inotropic agents (TDA) were also recorded at T2 and T3 (VIS = dopamine [μg/kg/min] + dobutamine [μg/kg/min] + 10 × milrinone [μg/kg/min] + 100 × epinephrine [μg/kg/min] + 100 × norepinephrine [μg/kg/min] + 1000 × vasopressin [U/kg/min]; TDA = dopamine [mg] + dobutamine [mg] + 10 × milrinone [mg] + 100 × epinephrine [mg] + 100 × norepinephrine [mg] + 1000 × vasopressin [U]).

Sample Size Estimation

The sample size was calculated based on the primary outcome using PASS 15.0 (NCSS, LLC). According to published studies, the incidence of VF after ACC release ranged from 10% (2/20) to 80% (38/47) [1]. Based on our preliminary data, the incidence of VF 10 minutes after ACC release was 70% (7/10) in the placebo-controlled group, while the VF rate was 30% (3/10) in the amiodarone group. To reach statistical significance with a power of 80% and an α value of .05 (2-sided), 21 patients were needed in each group. Assuming a 20% dropout or missing data rate, we recruited 27 patients in each group for a total of 54 patients.

Statistical Analysis

All data were analyzed using SPSS 22.0 (IBM Corp) and GraphPad Prism 8.0 (GraphPad Software Inc). All statistical tests were considered significant if $P<.05$. Only 2-sided tests were used. The normality assumption was assessed with the Kolmogorov-Smirnov test for all tests. Continuous variables were presented as mean (SD) or median (IQR), depending on the distribution. Categorical variables were presented as frequency (%). Continuous variables were compared using Student *t* tests if normally distributed or Mann-Whitney *U* tests if not normally distributed. The proportions were compared using Pearson chi-square or continuity correction tests. Data for repeated measures were compared using repeated measures ANOVA.

Results

Baseline Characteristics

In total, 77 patients with LVH who underwent open-heart surgery were selected for this study. Among them, 54 patients (27 in group A and 27 in group P) completed this trial and 23 patients were excluded, which were shown in the CONSORT flow diagram (Figure 1). The demographic and perioperative data of the participants were comparable between groups (Table 1).

XSL•FO
RenderX

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) flow diagram of patient recruitment, randomization, and withdrawal. A total of 27 patients in each group were eventually included in the final analysis. ASA: American Society of Anesthesiologists; CABG: coronary artery bypass grafting.

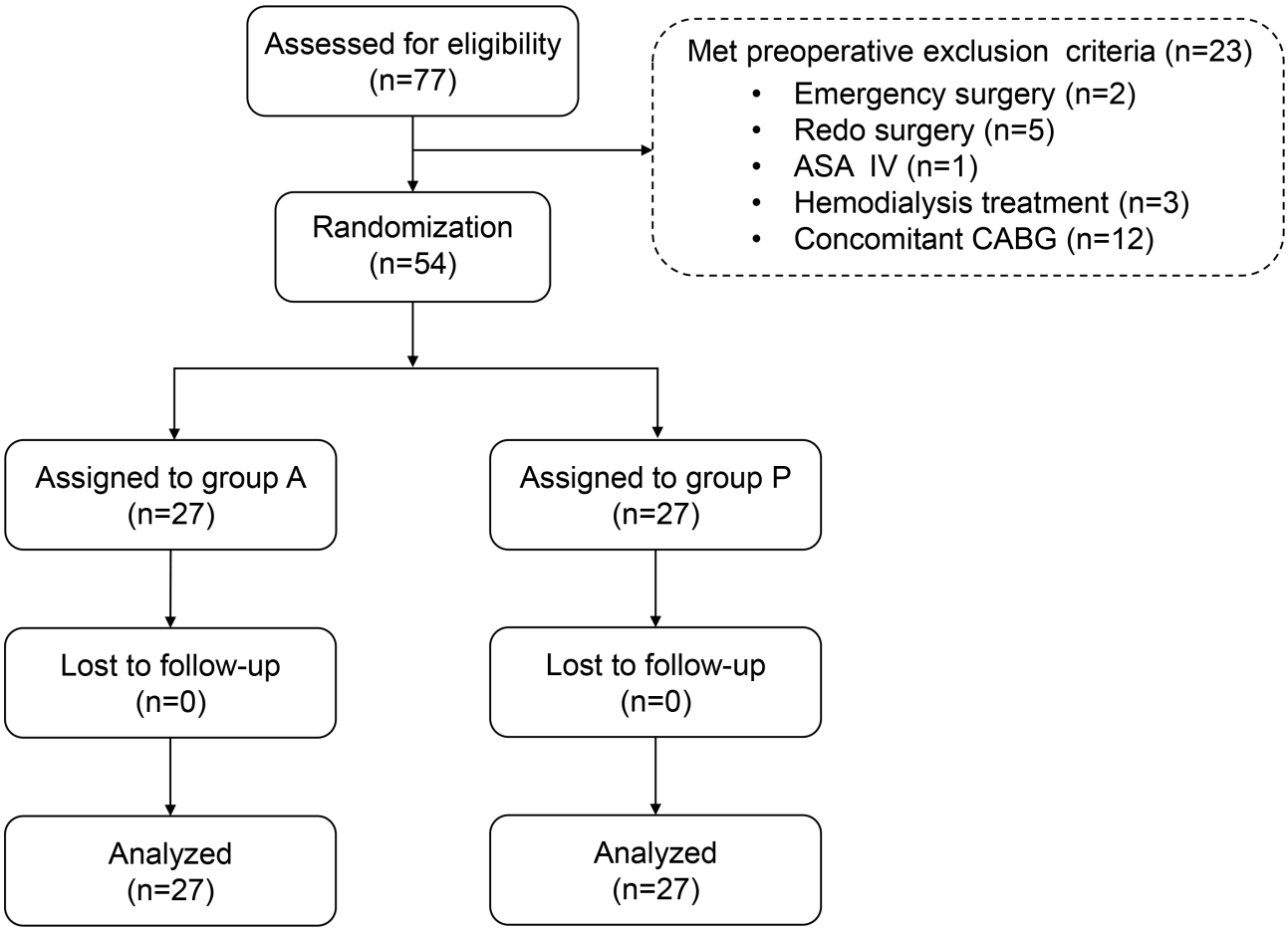

XSL•FO
RenderX

Table . Demographic characteristics and perioperative data.

Perioperative information		Group A[a] (N=27)	Group P[b] (N=27)	*P* value
Preoperative				
	Age (y), mean (SD)	64.44 (9.94)	65.59 (6.30)	.62
	Male, n	19	14	.16
	BMI (kg/m^2), mean (SD)	22.74 (2.42)	23.50 (1.92)	.21
	Hypertension, n	10	6	.23
	Diabetes mellitus, n	4	7	.50
	Hyperlipidemia. n	3	2	.99
	Smoking, n	6	4	.73
	NYHA[c] class, mean (SD)	2.63 (0.49)	2.41 (0.50)	.11
	LVMI[d] (g/m^2), mean (SD)	134.40 (11.74)	132.44 (15.05)	.60
	LVEF[e] (%), mean (SD)	57.11 (3.94)	58.78 (5.51)	.21
	Calcium blockers, n	6	8	.54
	β-blockers, n	20	16	.19
	Diuretics, n	7	6	.75
	ACEIs[f], ARBs[g], n	10	11	.78
Intraoperative				
	Procedures, n			.20
	AVR[h]	18	23	
	Morrow	9	4	
	Surgical time (min), mean (SD)	358.44 (35.67)	375.04 (35.55)	.09
	CPB[i] time (min), mean (SD)	154.89 (30.98)	168.70 (33.48)	.12
	ACC[j] time (min), mean (SD)	135.33 (28.91)	143.81 (29.53)	.29
	Time for CPB weaning (min), mean (SD)	28.04 (10.81)	30.59 (8.69)	.42
	Blood loss (mL), mean (SD)	528.89 (213.87)	496.74 (184.61)	.56
	Urine output (mL), mean (SD)	1350.00 (527.36)	1417.04 (754.10)	.71
	Total fluid volume (mL), mean (SD)	3727.04 (423.72)	3562.59 (448.78)	.17
	Total cardioplegia (mL), mean (SD)	1385.19 (317.39)	1412.96 (374.06)	.77
	Abnormal electrolyte or acid-base metabolism, n	0	0	.99
Postoperative, mean (SD)				
	Ventilation time (h)	10.26 (4.69)	11.44 (4.93)	.37
	ICU[k] stay (d)	4.96 (3.20)	5.93 (3.76)	.32
	Hospital stays (d)	11.85 (4.02)	13.74 (5.41)	.15

[a]Group A: amiodarone group.
[b]Group P: placebo control group.
[c]NYHA: New York Heart Association.
[d]LVMI: left ventricular mass index.
[e]LVEF: left ventricular ejection fraction.
[f]ACEI: angiotensin converting enzyme inhibitor.
[g]ARB: angiotensin receptor blocker.
[h]AVR: aortic valve replacement.
[i]CPB: cardiopulmonary bypass.

XSL•FO
RenderX

[j]ACC: aortic cross-clamp.
[k]ICU: intensive care unit.

Intraoperative Outcomes

The incidence of VF after the release of the ACC was significantly lower in group A than in group P (8/27, 30% vs 19/27, 70%; *P*=.003) (Table 2). Furthermore, the number of internal shocks and the total energy of defibrillations were lower in group A than in group P (*P*=.002). Furthermore, a shorter duration of VF was also observed in group A (*P*<.001). Regarding the restoration of cardiac rhythm after ACC release, there were no significant differences in the recovery time of cardiac contraction or the proportion of epicardial pacing between the groups. However, the duration of epicardial pacemaker use in group A was higher than those in group P (*P*=.03). When dealing with VF after the release of the ACC, there were no significant differences in the consumption of antiarrhythmic drugs between the groups (Table 3). At the T2 time point, group A had a higher VIS than group P (*P*=.02). However, no differences in VIS were observed between the groups at the time of T3. During the whole intraoperative period, there were no significant differences in TDA between the groups.

Table . Incidence of ventricular fibrillation (VF) and restoration of cardiac rhythm after aortic cross-clamp release.

Changes of cardiac rhythm	Group A[a] (N=27)	Group P[b] (N=27)	*P* value
VF, n (%)	8 (30)	19 (70)	.003
Defibrillations, n			.002
0	19	8	
1	6	9	
≥2	2	10	
Total energy (J), median (IQR)	0 (0 - 20)	20 (0 - 40)	.002
VF duration (min), median (IQR)	0 (0 - 1)	1.5 (0 - 3.5)	<.001
Time for cardiac contraction recovery (min), mean (SD)	2.63 (1.15)	2.61 (1.60)	.96
Epicardial pacing, n	14	8	.10
Duration of epicardial pacing (min), median (IQR)	10 (0 - 41)	0 (0 - 9)	.03

[a]Group A: amiodarone group.
[b]Group P: placebo control group.

Table . Consumption of antiarrhythmic and vasoactive-inotropic agents.

Intravenous agents	Group A[a] (N=27)	Group P[b] (N=27)	*P* value
Esmolol (mg), median (IQR)	70 (65 - 75)	130 (70 - 150)	.20
Lidocaine (mg), median (IQR)	0 (0 - 0)	0 (0 - 0)	.36
VIS[c] (T2[d]), mean (SD)	15.32 (3.66)	12.83 (4.19)	.02
VIS (T3[e]), mean (SD)	10.61 (3.99)	10.39 (4.10)	.84
TDA[f], mean (SD)	32.20 (8.98)	27.54 (8.64)	.06

[a]Group A: amiodarone group.
[b]Group P: placebo control group.
[c]VIS: vasoactive-inotropic score.
[d]T2: fifteen minutes after weaning from cardiopulmonary bypass.
[e]T3: fifteen minutes before the end of surgery;
[f]TDA: total dose of vasoactive and inotropic agents.

Hemodynamic Status During the Intraoperative Period

The hemodynamic status at various intraoperative times is shown in Figure 2. During the intraoperative period, repeated measures ANOVA indicated significant differences in MAP between the 2 groups (*F1*=5.161, *P*=.03). The global MAP was lower in group A than in group P (Figure 2). Compared with group P, MAP in group A was significantly lower at the time point T2 (mean 61.19 SD 4.16 vs mean 69.67 SD 6.23; *P*<.001). However, no significant differences were found in MAP at the time points T1 and T3 between the groups. During surgery, repeated measures ANOVA did not show significant differences in HR, MPAP, CVP, and CI between the groups (*F1,1,1,1*=3.098, 0.747, 0.271, and 0.001; and *P*=.08, .39, .61, and .97, respectively) (Figure 2). Compared with group P, HR

XSL•FO
RenderX

(mean 86.44 SD 8.07 vs mean 94.63 SD 6.90; *P*<.001) was lower and MPAP (mean 20.15 SD 4.63 vs mean 17.56 SD 4.44; *P*=.04) was higher in group A at the time point T2. However, no significant differences were found in CVP and CI at the T2 time point. Furthermore, there were no significant differences in HR, MPAP, CVP, and CI at the time points T1 and T3 between the groups.

Figure 2. Changes in hemodynamic status during the intraoperative period. The global mean arterial pressure curve was lower in group A than in group P by repeated measures ANOVA. There were no significant differences in the global curves of heart rate, mean pulmonary arterial pressure, central venous pressure, and cardiac index between the 2 groups by repeated measures ANOVA. At T2 point, mean arterial pressure and heart rate were significantly higher in Group P while mean pulmonary arterial pressure was significantly higher in Group A. Group A: amiodarone group; Group P: placebo control group. CI: cardiac index; CVP: central venous pressure; MAP: mean arterial pressure; MPAP: pulmonary arterial pressure; T1: fifteen minutes after anesthetic induction; T2: fifteen minutes after weaning from cardiopulmonary bypass; T3: fifteen minutes before the end of surgery. **P*<.05 was found between the two groups.

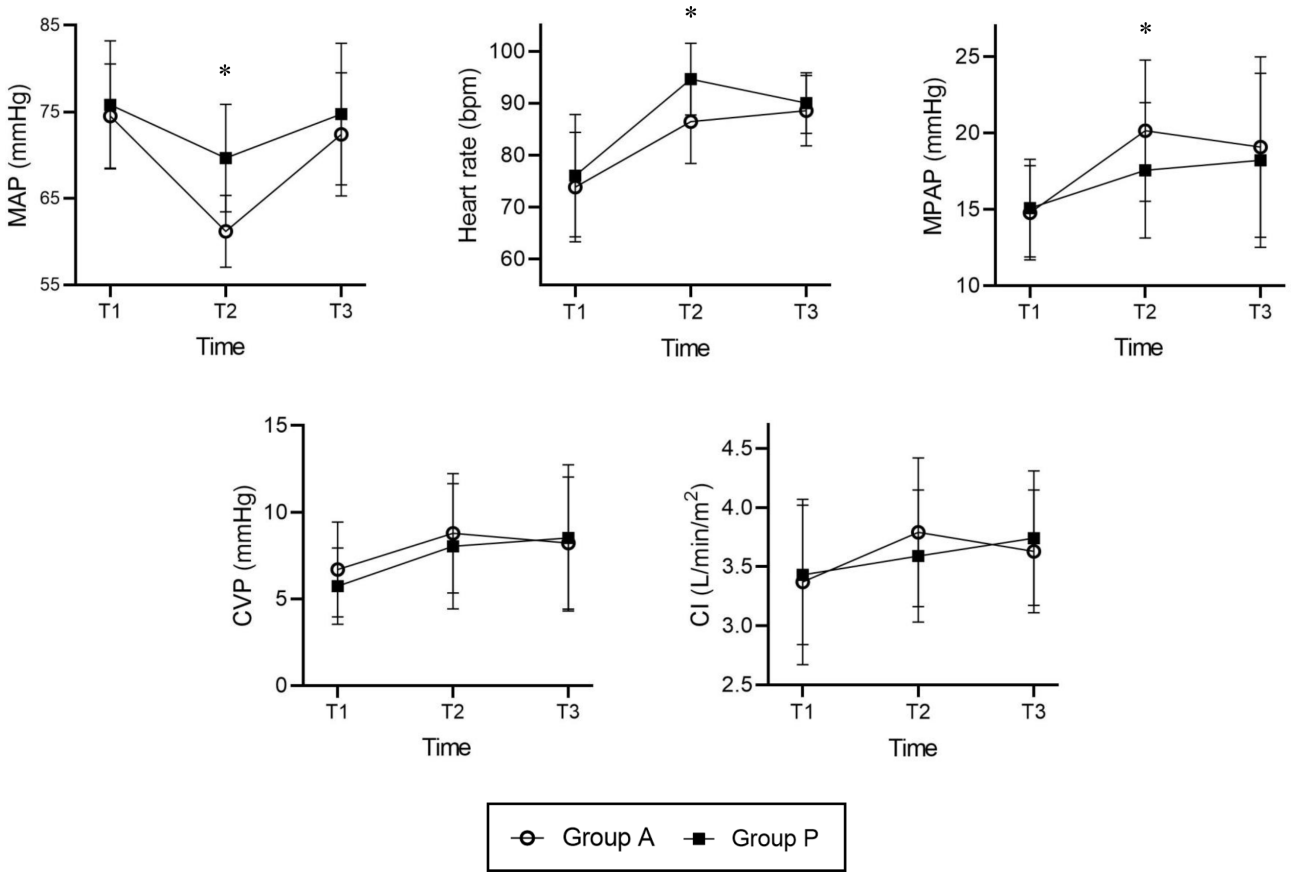

Discussion

Principal Findings

In this randomized trial, we primarily found that prophylactic intravenous use of amiodarone effectively reduced the incidence of VF in patients with LVH after ACC removal during open-heart surgery, while also decreasing the duration of VF, the frequency of defibrillation, and the energy required for defibrillation.

VF is a malignant arrhythmia that occurs after the release of the ACC in patients undergoing open-heart surgery. There are various factors that can contribute to VF after the removal of the ACC, which include surgical trauma, inadequate myocardial protection, insufficient intracardiac deairing or left ventricular venting, abnormal electrolyte and acid-base metabolism, delayed rewarming prior to ACC release, and myocardial reperfusion injury after ACC release. Among these factors, myocardial reperfusion injury is often considered the primary cause [16]. The prolonged duration of VF following ACC release can have several adverse effects on cardiomyocytes. Research findings indicated that VF led to alterations in the distribution of coronary artery blood supply, consequently causing subendocardial myocardial ischemia and subsequent cardiac dysfunction [17]. As the duration of VF extends, the resistance of subendocardial vessels progressively rises, exacerbating the subendocardial ischemia in cardiomyocytes. Furthermore, myocardial edema was observed to develop with the prolonged duration of VF [17]. During VF, there is an elevation in ventricular wall tension and venous pressure, necessitating a higher perfusion pressure to restore normal blood supply during CPB. The erratic and frequent myocardial contractions, along with elevated pressure in the cardiac cavity during VF, may lead to increased oxygen consumption by the myocardium. The described factors indeed result in an imbalance between oxygen supply and demand in the myocardium, leading to severe impairment of cardiomyocyte function.

In individuals with LVH, the cardiomyocytes enlarge in volume, yet the associated coronary arteries fail to grow in tandem, leading to a lack of subendocardial capillaries. Elevated end-diastolic pressure and enhanced systolic cardiac function

XSL•FO
RenderX

contribute to an increase in myocardial oxygen consumption. These histological alterations result in an imbalance between myocardial oxygen supply and demand, causing subendocardial myocardial ischemia and reduced myocardial tolerance to ischemia-reperfusion injury [18]. Patients with LVH have been reported to be more likely to suffer from VF after removal of the ACC [5,6,13]. Therefore, this situation has created a need for prophylactic treatment.

Amiodarone, as a broad-spectrum antiarrhythmic agent, has several electrophysiological mechanisms to treat VF. The main antiarrhythmic effect of amiodarone occurs through the inhibition of IKr and IKs channels, leading to a homogeneous prolongation of myocardial repolarization [7]. In addition, it can reduce HR and slow atrioventricular nodal conduction by blocking calcium channels and β-receptors [19]. Furthermore, amiodarone extends refractoriness and decreases intracardiac conduction velocity by inhibiting the sodium channel [20]. During the period of ischemia-reperfusion, amiodarone has the ability to enhance the metabolic efficiency of cardiomyocytes and decrease the dispersion of repolarization across the cell membrane, which is a significant contributor to the development of ventricular arrhythmias [21].

Several clinical trials have investigated whether amiodarone could lower the occurrence of VF after ACC removal. Samantaray et al [11] reported that, compared with the placebo group, administering 150 mg of amiodarone 3 minutes before aortic opening during CABG surgery reduced the occurrence of VF after ACC release. In addition, Yilmaz et al [12] revealed that both 300 mg of amiodarone and 1.5 mg/kg of lidocaine reduced the occurrence of VF after ACC release compared with the placebo group. However, there were no significant differences in the reduction of VF between the amiodarone and lidocaine groups. On the contrary, in the clinical trial by Ayoub et al [22], it was found that the administration of amiodarone 2 minutes before aortic opening did not result in a reduction in the occurrence of VF after the release of the ACC during CABG surgery. On the other hand, lidocaine was effective in reducing the occurrence of VF in this study [22]. In addition, Mauermann et al [23] reported that both 300 mg of amiodarone and 1.5 mg/kg of lidocaine did not lead to a reduction in the occurrence of VF. Despite some conflicting results in the studies mentioned, there is a general indication that amiodarone has a therapeutic effect on VF after ACC removal.

In a documented case, a 66-year-old female patient underwent aortic valve replacement for severe aortic stenosis. The preoperative TTE screening indicated a thickness of up to 18 mm in the ventricular septum and left ventricular posterior wall. Following the release of the ACC and administration of lidocaine, refractory VF ensued, with multiple cardiac shocks showing no effectiveness. Finally, a dose of 150 mg of amiodarone was delivered through the aortic root, ultimately leading to the termination of VF with a single defibrillation [24]. Suzuki et al [5] also reported a similar case of using amiodarone to terminate rVF during cardiac surgery. In a clinical trial by Mita et al [13], patients with LVH undergoing aortic valve replacement were randomly divided into 2 groups. One group received a 150 mg bolus of amiodarone followed by a maintenance dose of 30 mg/h, while the other group received a 1 mg/kg bolus of lidocaine followed by a maintenance dose of 1 mg/kg/h after anesthetic induction. The occurrence of VF was markedly reduced in the amiodarone group compared with the lidocaine group (7/34, 20.6% vs 17/34, 50%; P=.02) [9]. Meanwhile, amiodarone has the potential to decrease the levels of interleukin-6 and tumor necrosis factor following the surgery.

In this study, our results showed that prophylactic intravenous amiodarone could lower the occurrence of VF in patients with LVH after the release of the ACC during open-heart surgery. Furthermore, among patients with VF, amiodarone reduced the number and total energy of internal cardiac shocks. In addition, the duration of VF could also be reduced. These results further proved the conclusions of the abovementioned studies. With the use of amiodarone, the recovery of cardiac contractions was not affected and there was no increase in the use of epicardial pacemakers. However, the duration of epicardial pacemakers was longer among patients with amiodarone, which may be correlated with the effect of amiodarone on slowing HR. In terms of hemodynamic effects, after weaning from CPB, HR and MAP decreased and MPAP increased, while the use of vasoactive and inotropic drugs was also higher in patients treated with amiodarone, which could be related to the pharmacological effect of amiodarone. However, CI and CVP were not affected and the hemodynamic status at the end of surgery was not significantly different between the groups. Regarding the consumption of vasoactive and inotropic drugs, although the dose of the drugs was higher among patients with amiodarone after weaning from CPB, the dose and total doses of vasoactive and inotropic agents at the end of surgery were not significantly different among the 2 groups.

Limitations

Our study presents several limitations. First, this was a single-center randomized controlled clinical trial with a relatively small sample size, which limited the power of the statistical analysis. Second, lidocaine was not compared as a control group in the study, because it may be more effective than a placebo but less effective than amiodarone to reduce the incidence of VF according to previous studies [12,13], which could increase the sample size and the cost of the study. Third, this study only included patients with ASA I-III. Therefore, these findings cannot be generalized to patients with ASA IV, especially those with severe left ventricular dysfunction. Therefore, subsequent studies should further address the abovementioned issues.

Conclusions

Preventive intravenous administration of amiodarone can reduce the incidence of VF in patients with LVH after ACC removal during open-heart surgery and decrease the duration of VF, defibrillation frequency, and defibrillation energy. It is safe and has little impact on intraoperative hemodynamic management. The findings may provide some insight for anesthesiologists in dealing with similar problems in clinical practice.

XSL•FO
RenderX

Acknowledgments

This research was supported by the Clinical Capacity Improvement Project of The First Affiliated Hospital of Nanjing Medical University in 2020 (NO. JSPH-MC-2020-9).

Data Availability

The datasets generated and analyzed during this study are not publicly available due to ethical considerations but are available from the corresponding author on reasonable request.

Authors' Contributions

YF and Z-FZ designed the research study. Z-FZ, XKZ, C-JG, and YF performed the research. C-JG and YF analyzed the data. Z-FZ and X-KZ wrote the manuscript. All authors participated in the writing (review and editing) of the manuscript and read and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Checklist 1
SPIRIT (Standard Protocol Items: Recommendations for Interventional Trials) checklist.
[PDF File, 140 KB - formative_v9i1e64586_app1.pdf]

Checklist 2
CONSORT-EHEALTH checklist (V 1.6.1).
[PDF File, 1226 KB - formative_v9i1e64586_app2.pdf]

References

1. He LM, Zhang A, Xiong B. Effectiveness of amiodarone in preventing the occurrence of reperfusion ventricular fibrillation after the release of aortic cross-clamp in open-heart surgery patients: a meta-analysis. Front Cardiovasc Med 2022;9:821938. [doi: 10.3389/fcvm.2022.821938] [Medline: 35187132]
2. Stommel AM, Högler S, Mueller M, et al. A ventricular fibrillation cardiac arrest model with extracorporeal cardiopulmonary resuscitation in rats: 8 minutes arrest time leads to increased myocardial damage but does not increase neuronal damage compared to 6 minutes. Front Vet Sci 2023;10:1276588. [doi: 10.3389/fvets.2023.1276588] [Medline: 38026669]
3. Yamaguchi H, Weil M, Tang W, Kamohara T, Jin X, Bisera J. Myocardial dysfunction after electrical defibrillation. Resuscitation 2002 Sep;54(3):289-296. [doi: 10.1016/s0300-9572(02)00149-1] [Medline: 12204463]
4. Udzik J, Sienkiewicz S, Biskupski A, Szylińska A, Kowalska Z, Biskupski P. Cardiac complications following cardiac surgery procedures. J Clin Med 2020 Oct 18;9(10):3347. [doi: 10.3390/jcm9103347] [Medline: 33081028]
5. Suzuki S, Iwasaki T, Morimatsu H, et al. Successful use of intravenous amiodarone for refractory ventricular fibrillation just after releasing aortic cross-clamp. Masui 2010 Oct;59(10):1266-1270. [Medline: 20960899]
6. Kodaka M, Mori T, Ichikawa J, Ando K, Komori M. Refractory ventricular arrhythmias during aortic valve replacement and cardiac artery bypass requiring 16 attempts of electrical cardioversion: a case report. JA Clin Rep 2020 Aug 11;6(1):60. [doi: 10.1186/s40981-020-00369-w] [Medline: 32783130]
7. Mujović N, Dobrev D, Marinković M, Russo V, Potpara TS. The role of amiodarone in contemporary management of complex cardiac arrhythmias. Pharmacol Res 2020 Jan;151:104521. [doi: 10.1016/j.phrs.2019.104521] [Medline: 31756386]
8. orodoc V, Indrei L, Dobroghii C, et al. Amiodarone therapy: updated practical insights. J Clin Med 2024 Oct 12;13(20):6094. [doi: 10.3390/jcm13206094] [Medline: 39458044]
9. Sattler SM, Lubberding AF, Skibsbye L, et al. Amiodarone treatment in the early phase of acute myocardial infarction protects against ventricular fibrillation in a porcine model. J Cardiovasc Transl Res 2019 Aug;12(4):321-330. [doi: 10.1007/s12265-018-9861-6] [Medline: 30617762]
10. Khataminia M, Najmeddin F, Najafi A, et al. Effect of heart rate control with amiodarone infusion on hemodynamic and clinical outcomes in septic shock patients with tachycardia: a prospective, single-arm clinical study. J Pharm Health Care Sci 2021 Oct 11;7(1):37. [doi: 10.1186/s40780-021-00219-6] [Medline: 34629112]
11. Samantaray A, Chandra A, Panigrahi S. Amiodarone for the prevention of reperfusion ventricular fibrillation. J Cardiothorac Vasc Anesth 2010 Apr;24(2):239-243. [doi: 10.1053/j.jvca.2009.07.007] [Medline: 19800815]
12. Yilmaz M, Aydin U, Arslan ZI, et al. The effect of lidocaine and amiodarone on prevention of ventricular fibrillation in patients undergoing coronary artery bypass grafting. Heart Surg Forum 2014 Oct 1;17(5):E245-E249. [doi: 10.1532/HSF98.2014402] [Medline: 25367235]

XSL•FO
RenderX

13. Mita N, Kagaya S, Miyoshi S, Kuroda M. Prophylactic effect of amiodarone infusion on reperfusion ventricular fibrillation after release of aortic cross-clamp in patients with left ventricular hypertrophy undergoing aortic valve replacement: a randomized controlled trial. J Cardiothorac Vasc Anesth 2019 May;33(5):1205-1213. [doi: 10.1053/j.jvca.2018.10.005] [Medline: 30416026]
14. Zhou X, Zhang Z, Gong C, Fang Y. The effect of prophylactic intravenous amiodarone administration on reperfusion ventricular fibrillation in patients with left ventricular hypertrophy undergoing cardiopulmonary bypass surgery: protocol for a randomized double-blind clinical trial. JMIR Res Protoc 2023 Jan 27;12:e40115. [doi: 10.2196/40115] [Medline: 36705966]
15. Lang RM, Badano LP, Mor-Avi V, et al. Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging. J Am Soc Echocardiogr 2015 Jan;28(1):1-39. [doi: 10.1016/j.echo.2014.10.003] [Medline: 25559473]
16. Zheng Y, Gu Q, Chen HW, et al. Efficacy of amiodarone and lidocaine for preventing ventricular fibrillation after aortic cross-clamp release in open heart surgery: a meta-analysis of randomized controlled trials. J Zhejiang Univ Sci B 2017;18(12):1113-1122. [doi: 10.1631/jzus.B1700229] [Medline: 29204991]
17. Hottenrott C, Maloney JV Jr, Buckberg G. Studies of the effects of ventricular fibrillation on the adequacy of regional myocardial flow. 3. Mechanisms of ischemia. J Thorac Cardiovasc Surg 1974 Oct;68(4):634-645. [Medline: 4278292]
18. Zhu YC, Zhu YZ, Spitznagel H, Gohlke P, Unger T. Substrate metabolism, hormone interaction, and angiotensin-converting enzyme inhibitors in left ventricular hypertrophy. Diabetes 1996 Jan;45 Suppl 1:S59-S65. [doi: 10.2337/diab.45.1.s59] [Medline: 8529802]
19. Siddoway LA. Amiodarone: guidelines for use and monitoring. Am Fam Physician 2003 Dec 1;68(11):2189-2196. [Medline: 14677664]
20. Pannone L, D'Angelo G, Gulletta S, et al. Amiodarone in ventricular arrhythmias: still a valuable resource? Rev Cardiovasc Med 2021 Dec 22;22(4):1383-1392. [doi: 10.31083/j.rcm2204143] [Medline: 34957778]
21. Moreau D, Clauw F, Martine L, Grynberg A, Rochette L, Demaison L. Effects of amiodarone on cardiac function and mitochondrial oxidative phosphorylation during ischemia and reperfusion. Mol Cell Biochem 1999 Apr;194(1-2):291-300. [doi: 10.1023/a:1006935323491] [Medline: 10391152]
22. Ayoub CM, Sfeir PM, Bou-Khalil P, et al. Prophylactic amiodarone versus lidocaine for prevention of reperfusion ventricular fibrillation after release of aortic cross-clamp. Eur J Anaesthesiol 2009 Dec;26(12):1056-1060. [doi: 10.1097/EJA.0b013e32832f0dfb] [Medline: 19809326]
23. Mauermann WJ, Pulido JN, Barbara DW, et al. Amiodarone versus lidocaine and placebo for the prevention of ventricular fibrillation after aortic crossclamping: a randomized, double-blind, placebo-controlled trial. J Thorac Cardiovasc Surg 2012 Nov;144(5):1229-1234. [doi: 10.1016/j.jtcvs.2012.06.039] [Medline: 22770549]
24. Morita Y, Mizuno J, Yoshimura T, Morita S. Efficacy of amiodarone on refractory ventricular fibrillation resistant to lidocaine and cardioversion during weaning from cardiopulmonary bypass in aortic valve replacement for severe aortic stenosis with left ventricular hypertrophy. J Anesth 2010 Oct;24(5):761-764. [doi: 10.1007/s00540-010-0991-5] [Medline: 20665054]

Abbreviations

ACC: aortic cross-clamp
ASA: American Society of Anesthesiology
CABG: coronary artery bypass grafting
CI: cardiac index
CONSORT: Consolidated Standards of Reporting Trials
CPB: cardiopulmonary bypass
CVP: central venous pressure
HR: heart rate
LVH: left ventricular hypertrophy
MAP: mean arterial pressure
MPAP: mean pulmonary arterial pressure
$P_{ET}CO_2$: end-tidal CO_2
rVF: reperfusion ventricular fibrillation
TDA: total dose of vasoactive and inotropic agents
TTE: transthoracic echocardiography
VF: ventricular fibrillation
VIS: vasoactive-inotropic score

XSL•FO
RenderX

Edited by A Mavragani; submitted 21.07.24; peer-reviewed by R Agarwal, Z Yu; revised version received 21.11.24; accepted 26.11.24; published 13.01.25.

Please cite as:
Gong CJ, Zhou XK, Zhang ZF, Fang Y
Impact of Preventive Intravenous Amiodarone on Reperfusion Ventricular Fibrillation in Patients With Left Ventricular Hypertrophy Undergoing Open-Heart Surgery: Randomized Controlled Clinical Trial
JMIR Form Res 2025;9:e64586
URL: https://formative.jmir.org/2025/1/e64586
doi:10.2196/64586

© Chan-Juan Gong, Xiao-Kai Zhou, Zhen-Feng Zhang, Yin Fang. Originally published in JMIR Formative Research (https://formative.jmir.org), 13.1.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Identification of Behavioral, Clinical, and Psychological Antecedents of Acute Stimulant Poisoning: Development and Implementation of a Mixed Methods Psychological Autopsy Study

Marley Antolin Muñiz[1*], BS; Vanessa M McMahan[1*], MS, PhD; Xochitl Luna Marti[1], MPH; Sarah Brennan[1], BA; Sophia Tavasieff[1]; Luke N Rodda[2,3], PhD; James Knoll[4], MD; Phillip O Coffin[1,3], MIA, MD

[1]Center on Substance Use and Health, San Francisco Department of Public Health, San Francisco, CA, United States

[2]Forensic Laboratory Division, Office of the Chief Medical Examiner, San Francisco, CA, United States

[3]Department of Medicine, University of California, San Francisco, San Francisco, CA, United States

[4]Department of Psychiatry, SUNY Upstate Medical University, Syracuse, NY, United States

[*]these authors contributed equally

Corresponding Author:
Phillip O Coffin, MIA, MD
Center on Substance Use and Health, San Francisco Department of Public Health, San Francisco, CA, United States

Abstract

Background: Despite increasing fatal stimulant poisoning in the United States, little is understood about the mechanism of death. The psychological autopsy (PA) has long been used to distinguish the manner of death in equivocal cases, including opioid overdose, but has not been used to explicitly explore stimulant mortality.

Objective: We aimed to develop and implement a large PA study to identify antecedents of fatal stimulant poisoning, seeking to maximize data gathering and ethical interactions during the collateral interviews.

Methods: We ascertained death records from the California Electronic Death Reporting System (CA-EDRS) and the San Francisco Office of the County Medical Examiner (OCME) from June 2022 through December 2023. We selected deaths determined to be due to acute poisoning from cocaine or methamphetamine, which occurred 3 - 12 months prior and were not attributed to suicide or homicide. We identified 31 stimulant-fentanyl and 70 stimulant-no-opioid decedents. We sought 2 informants for each decedent, who were able to describe the decedent across their life course. Informants were at least 18 years of age, communicated with the decedent within the year before death, and were aware that the decedent had been using substances during that year. Upon completion of at least one informant interview conducted by staff with bachelor's or master's degrees, we collected OCME, medical record, and substance use disorder treatment data for the decedent. Planned analyses include least absolute shrinkage and selection operator regressions of quantitative data and thematic analyses of qualitative data.

Results: We identified and interviewed at least one informant (N=141) for each decedent (N=101). Based on feedback during recruitment, we adapted language to improve rapport, including changing the term "accidental death" to "premature death," offering condolences, and providing content warnings. As expected, family members were able to provide more data about the decedent's childhood and adolescence, and nonfamily informants provided more data regarding events proximal to death. We found that the interviews were stressful for both the interviewee and interviewer, especially when participants thought the study was intrusive or experienced significant grief during the interviews.

Conclusions: In developing and implementing PA research on fatal stimulant poisoning, we noted the importance of recruitment language regarding cause of death and condolences with collateral informants. Compassion and respect were critical to facilitate the interview process and maintain an ethical framework. We discuss several barriers to success and lessons learned while conducting PA interviews, as well as recommendations for future PA studies.

(JMIR Form Res 2025;9:e64873) doi:10.2196/64873

KEYWORDS

psychological autopsy; acute stimulant poisoning; overdose; cocaine; methamphetamine; fentanyl

Introduction

Deaths attributed to acute stimulant poisoning have been increasing in the United States [1]. From 2015 to 2022, poisoning deaths in the United States attributed to cocaine or methamphetamine increased 4- and 6-fold, respectively [2]. This increase in mortality, largely driven by deaths also involving fentanyl, a potent synthetic opioid, has been referred

XSL•FO
RenderX

to as the "fourth wave" of the overdose crisis [3-6]. While the increase in deaths involving stimulants and fentanyl may be primarily due to the toxic effects of fentanyl, the presence of stimulants likely reflects rising polysubstance and, in some cases, unintentional fentanyl use [4,5,7].

Fatal opioid overdose has a well-defined mechanism—respiratory depression eventually leading to cardiac arrest and death—and decades of research exploring the nature of and antecedents to death. However, this framework and associated research are absent for fatal stimulant poisoning, leading investigators and public health providers to assume stimulant poisoning deaths are similar to opioid overdose deaths and a result of an "overdose." This has led to efforts to identify prevention and treatment strategies, like bystander interventions, reversal agents, and strategies to prevent an acute overdose, akin to strategies used for opioid overdose. However, deaths attributed to stimulant poisoning, which are attributed to cardiac or cerebrovascular events far more often than for opioid overdose deaths, may actually be due to chronic disease rather than a single event of substance use [8,9]. Thus, prevention efforts focused on bystander interventions may not be effective to prevent stimulant poisoning deaths, but instead longer-term interventions, such as screening, prevention, and treatment of chronic diseases, may be appropriate [10].

In the Leveraging Psychological Autopsies to Accelerate Research into Stimulant Overdose Mortality study, we aimed to identify the clinical, behavioral, and psychological antecedents of fatal stimulant poisoning. We selected the psychological autopsy (PA) [11,12], which has been used for suicide [12-15] and opioid overdose research [16,17], because this exploratory approach identified concepts decades ago that still drive opioid overdose prevention today [18]. We hypothesized that fatal stimulant poisoning likely has individual, social, and environmental contributors, and thus employed a socioecological conceptual framework [19]. To identify multilevel antecedents of fatal stimulant poisoning, we conducted PAs involving detailed interviews with close contacts of decedents (collateral informants), analyzed alongside medical examiner (case narrative, autopsy, and toxicology reports) and medical record data [20]. We describe the adaptation of PAs to study fatal stimulant poisoning and lessons learned.

Methods

Ethical Considerations

We obtained approval for this study from the University of California San Francisco, Human Research Protection Program (#21 - 35305). To use informant information in death records from the California Electronic Death Reporting System (CA-EDRS), we also received approval from the State of California Health and Human Services Agency Committee for the Protection of Human Subjects (#2022 - 091).

Decedent Case Selection

We reviewed records from the CA-EDRS and the Office of the County Medical Examiner (OCME) of San Francisco for eligible cases from June 2022 through December 2023. We selected closed cases of deaths determined to be due to acute poisoning from cocaine or methamphetamine. Suicides and homicides were excluded because the mechanism and antecedents to death would likely differ from accidental, natural, or undetermined deaths [21]. We sought 30 stimulant-fentanyl decedents (15 involving cocaine and 15 involving methamphetamine) and 70 stimulant-no-opioid decedents (35 involving cocaine and 35 involving methamphetamine). We suspected that stimulant-fentanyl deaths would be similar to opioid deaths, which are well understood. However, little is known about stimulant deaths not involving opioids, thus we targeted a larger sample of that group [9]. Decedents without a known name or informant were excluded. Interviews were conducted 3 - 12 months after decedent death to minimize both acute bereavement and forgetfulness among informants [22,23]. Decedents were considered "enrolled" in the study when at least one informant interview was completed, at which time OCME, medical record, and substance use treatment data were collected.

Informant Selection

We sought 2 informants for each decedent, who ideally could describe the decedent across the life course (eg, childhood experiences and behaviors proximal to death). Informants had to be at least 18 years of age, have communicated with the decedent within the year before death, be aware that the decedent was using substances during that year, and be able to provide informed consent in English. We identified potential informants using CA-EDRS and OCME data, medical record review, flyers, and word of mouth. We conducted outreach at harm reduction programs and public housing settings where decedents had resided. If a mailing address was available for a potential informant, we mailed a recruitment letter that included condolences, the purpose of the study, details of participation, and the option to opt out of additional contact from the study team (Multimedia Appendix 1). We then waited 1 week before reaching out by phone if a phone number was available. If only a phone number was available, we reached out immediately with up to 3 calls, typically separated by at least 1 week.

Interview Procedures

Bachelor's and master's-level staff conducted the interviews after receiving 2 hours of training by a psychiatrist experienced in PAs. All informants provided informed consent through review of an information sheet. Semistructured, mixed methods interviews lasting 1 - 2 hours were then conducted over telephone, Zoom, or in person. The interview included 14 thematic sections involving sociodemographic characteristics, neighborhood and housing, adverse childhood experiences, emotional and psychological health, substance use, and social environment (Multimedia Appendix 1). Each section began with quantitative questions, including validated scales, followed by qualitative questions to add nuance and to contextualize quantitative data. Quantitative interview responses were entered into University of California San Francisco's REDCap (Research Electronic Data Capture) [24,25] in real time by the interviewer. Interviews were audio-recorded and transcribed verbatim via the Rev platform. Informants were compensated US $50 and offered an additional US $10 for any referral that resulted in a completed informant interview. After the interview, informants were offered bereavement resources and additional information

XSL•FO
RenderX

about the cause of death as appropriate. Because some informants had superior knowledge about different aspects of a decedent's life, the interviewer scored the reliability of the responses to each of the survey sections as highly unreliable, somewhat unreliable, neutral, somewhat reliable, or highly reliable.

Additional Data Sources

Upon completion of at least one informant interview, we collected additional OCME, medical record, and substance use disorder treatment data. We waited until the informant interview was completed to ensure we only collected these data for enrolled decedents. The data abstraction form was developed based on prior research the team had conducted [26], the Centers for Disease Control and Prevention (CDC) State Unintentional Drug Overdose Reporting System (SUDORS) [27], and a preliminary review of 15 decedents. OCME data included the case narrative, toxicology, and autopsy reports. Medical record data included regularity of medical care (eg, established primary care, urgent care only, emergency or hospital care only); recent emergency or inpatient care; latest electrocardiogram and echocardiogram results; latest troponin, brain natriuretic peptide, and urine drug screen results; any evidence of opioid, cocaine, or methamphetamine use; all problem list diagnoses; and all active medications at the time of death.

Analytic Plan

We plan to conduct descriptive statistics of quantitative data, with a focus on a priori variables of interest, including cardiac health, social isolation, fentanyl exposure, suicidality, and resilience. We will then conduct least absolute shrinkage and selection operator (LASSO) regression analyses [28] to identify variables related to fatal stimulant poisoning and analyze results by fentanyl involvement and by cocaine compared to methamphetamine as the causal stimulant. We will also conduct targeted, hypothesis-driven regression analyses by fentanyl involvement exploring cardiac health and opioid use history. Qualitative data will be analyzed by study team members using ATLAS.ti 23.2.2 with an a priori codebook refined during the consensus-driven coding process. Qualitative analyses will focus on salient and recurrent themes and describe theme co-occurrence.

Results

We adapted the PA approach to conduct a large study of the antecedents of fatal stimulant poisoning (Figure 1). We identified and interviewed at least one informant for each decedent. During this period, we reviewed 177 potential stimulant-fentanyl cases prior to enrolling 31 decedents (18% of potential decedents; an extra decedent was erroneously enrolled) and 187 potential stimulant-no-opioid cases prior to enrolling 70 decedents (37% of potential decedents). For 19 (5%) cases, there were no informants available. We contacted 725 potential informants for a total of 364 decedents; 410 (57%) did not complete an interview or follow-up with the study team, 112 (15%) were not interested in participating, 62 (9%) were ineligible, and 141 (19%) completed an interview (Table 1).

Recruitment phone calls varied in length, usually lasting 15 minutes, or as long as 40 minutes when a potential informant was in emotional distress. Calls that were longer typically seemed welcome by potential participants, during which they discussed the lives of the decedents and their own feelings of grief. In some calls, potential participants also discussed frustration with the unexpected contact.

Most potential informants were family members (n=409, 56%), including 133 (18%) parents, 80 (11%) children, and 196 (27%) siblings or other family. We also contacted friends (n=79, 11%), spouses or partners (n=43, 6%), and service providers (n=49, 7%). Many decedents had lived in supportive housing settings (eg, single residence occupancy hotels); thus, 77 (11%) potential informants were building staff. Informants who completed interviews had a similar distribution (Table 2).

The majority of stimulant-involved deaths in San Francisco were also attributed to fentanyl [29], thus we rapidly enrolled stimulant-fentanyl decedents. We then shifted our focus to enrolling stimulant-no-opioid decedents, at which point 2 potential informants reached out to us with concerns about the study procedures; both were concerned with the unexpected outreach, the lack of condolences offered in our recruitment materials, and the phrase "accidental death" used to describe the purpose of the study. In dialogue with these potential informants, it became clear that they considered "accidental death" to mean "overdose," and they did not feel that their loved one had died from an "overdose."

We subsequently revised the recruitment letter and our telephone script to begin with content warnings about the sensitivity of the topic we would be discussing, refer to "premature death," instead of "accidental death," offer condolences, and allow for more check-ins. We also made the introduction of our recruitment materials intentionally vague and did not discuss the study's focus of fatal stimulant poisoning until the individual expressed comfort with talking with us, we suggested they speak with us from a private location, and we confirmed their eligibility (Multimedia Appendix 1). These changes were made to promote comfort and engagement among potential informants during recruitment and to limit disclosure of the involvement of stimulants in death to individuals without prior knowledge of the decedent's substance use. After making these changes, we completed all remaining interviews with only one potential informant voicing concern about the purpose of the study. We speculate that the potential informant had reviewed our center's website, which was included on the letterhead of the recruitment letter, providing them with discomforting details about the study.

In addition to the modified recruitment language, we made several other modifications to the study during implementation. Communicating with potential informants proved challenging via telephone calls only, so we modified the study to allow for text messaging with informants. We only corresponded via text message when contact had already been established with the potential informant. After recruitment commenced, it became clear that it would be challenging to interview 2 informants for each decedent, so we modified the study to allow for only one informant if we could not reach a second informant. We also allowed for more than 2 informants in the event one informant

XSL•FO
RenderX

who was interviewed was not knowledgeable about the decedent's life. We modified our inclusion criteria to allow for the enrollment of decedents who died from both methamphetamine and cocaine toxicity, since we did not think we would lose scientific information by their inclusion, and, depending on the number enrolled, we might be able to identify unique characteristics of this group. Finally, during the study we added additional recruitment strategies, including snowball sampling (ie, compensating a participant who referred another participant an additional US $10), contacting participants after the interview who had consented to being contacted for recruitment purposes, and posting flyers in settings where we could reach potential informants (eg, syringe services programs, public housing, etc).

Interviews lasted 1-2 hours. Interviewers coded the reliability of informants for each section of questions in the interview. In general, family members had superior information about childhood experiences and less reliable information about recent substance use. Staff experienced stress in conducting the interviews, particularly when potential informants felt that recruitment efforts were intrusive or when enrolled informants experienced significant grief during interviews. Staff were not specifically trained in psychology or bereavement counseling and at times were overwhelmed by the grief among informants. To alleviate some of the emotional burden on the study team, limits were set around interview capacity, with interviewers performing no more than 1 interview daily and a maximum of 3 interviews per week. Weekly team calls included conversations about recent informant interactions and upcoming capacity for recruitment calls and interviews. Staff could also access biweekly debrief sessions led by a clinical psychologist on staff, addressing topics such as the varying signs of grief and how those signs may show up in outreach and interviews. Interviewers and recruiters were trained in motivational interviewing techniques to provide staff with tools like reflective listening to express empathy during outreach and interviews.

XSL•FO
RenderX

Figure 1. Procedural flow for a mixed methods psychological autopsy study of decedents of fatal stimulant poisoning in San Francisco, CA, from June 2022 to December 2023. CDPH: California Department of Public Health; OCME: Office of the County Medical Examiner.

XSL•FO
RenderX

Table . The number of potential informants for eligible decedents and the recruitment status of potential informants contacted for a mixed methods psychological autopsy study of decedents of acute stimulant poisoning in San Francisco, CA, from June 2022 to December 2023.

	n	%
Eligible decedents (N=364)[a]		
No informants	19	5
At least one potential informant, not enrolled	244	67
At least one potential informant, enrolled	101	28
Potential informants contacted (N=725)		
Ineligible[b]	62	9
Refused participation	112	15
Contacted but no interview completed or did not follow-up[c]	410	57
Completed interview	141	19

[a]Closed cases due to cocaine or methamphetamine acute poisoning with and without fentanyl involvement. Deaths involving other opioids were excluded. Once 31 fentanyl-involved decedents were enrolled, eligibility was restricted to stimulant poisoning without any opioid involvement.

[b]Potential informants were ineligible if they were <18 years old, had not been in contact with the decedent in the year prior to death, did not know the decedent used substances in the year before death, were not able to provide informed consent in English, or the decedent had not died 3 - 12 months before recruitment.

[c]Includes potential informants who were sent a letter or called but never responded, as well as those who were successfully contacted but never followed up to complete an interview. Tracking database did not allow for further disaggregation.

Table . The relationships to decedents of potential informants who were contacted and interviewed for a mixed methods psychological autopsy study of decedents of fatal stimulant poisoning in San Francisco, CA, from June 2022 to December 2023.

Relationship of informant to decedent	Contacted (N=725)		Interviewed (N=141)	
	n	%	n	%
Parent	133	18	18	13
Spouse or partner	43	6	8	6
Child	80	11	17	12
Friend	79	11	30	21
Sibling or other family member	196	27	44	31
Service provider (eg, clinician, case manager)	49	7	11	8
Building staff	77	11	4	3
Other	68	9	9	6

Discussion

We successfully adopted the PA methodology for a large investigation of antecedents to fatal stimulant poisoning. We enrolled at least one informant for 101 decedents. We ascertained OCME and medical record data for all enrolled decedents. We identified challenges including recruiting informants, adapting our recruitment process to address informant grief, and ensuring staff were prepared to manage emotionally challenging interactions.

Enrolling informants was a substantial challenge of the project. We sought 2 informants per decedent; however, a high proportion of our decedents were markedly socially isolated. Nineteen of the decedents we initially identified were either unidentifiable (ie, “John Doe”) or had no contacts listed in their death or medical records. Among those who had any contacts, most had very few and often only family members, limiting our ability to obtain detailed information related to substance use. Few decedents had multiple contacts listed who knew the decedent proximal to death (eg, current friends, partners). We attempted to overcome these challenges by asking informants for referrals of others who knew the decedent and may be interested in participating in an interview, conducting direct outreach to public housing, posting flyers with our contact information, contacting health care providers for recommendations, and using personal connections research staff had with service programs in the community. We also altered the protocol to allow for a single informant when 2 could not be identified; as social isolation was a frequent antecedent to

XSL•FO
RenderX

death, we felt this adjustment would help to reduce bias by including more socially isolated decedents. Nonetheless, a significant proportion of decedents could not be enrolled due to an inability to reach any eligible informants.

Because stimulant-no-opioid cases were far less common than stimulant-fentanyl cases, we devoted substantially more effort to locating informants and ended up enrolling a higher proportion of potential stimulant-no-opioid decedents (37%) compared to potential stimulant-fentanyl decedents (18%). Many potential informants were ineligible due to our requirement that the informant be aware of the decedent's substance use, as well as speak sufficient English to provide informed consent. The dearth of informants for many decedents was compounded by ethical limitations in outreach efforts. Our human subjects committee required a letter be sent to potential informants with a mailing address but allowed direct phone contact for those without a mailing address.

Some informants experienced significant emotional distress during interviews and outreach. Potential informants received "cold calls" from study staff, rather than a warm handoff from a known provider, and did not directly consent to initial contact from our research team. This element of surprise and the sensitive nature of the call left some potential informants feeling the outreach was intrusive. However, for others who spoke at length during the recruitment calls, it seemed like a welcome opportunity to talk about the decedent. While a warm handoff from known clinical providers or the medical examiner's office may be preferred for initial outreach, this may not be feasible for larger PA studies or non-research "death review" projects attempting to track and improve our understanding of drug poisoning deaths.

During the study, we adapted recruitment materials and processes to respond to concerns from potential informants, including more sympathetic language, frequent check-ins, and a content warning. We only noted the role of stimulants in the decedent's death after confirming that the informant knew about the decedent's substance use, and we replaced the phrase "accidental death" with "premature death." These later changes were consistent with preliminary findings from early reviews of transcripts, as informants often did not perceive stimulant poisoning deaths to be "overdose" deaths but instead natural or medical deaths. Most potential informants reacted well to the improved script and recruitment approaches. As in other PA studies [30,31], many informants had a positive experience participating in the study and expressed gratitude to the study team for conducting the research, showing interest in the decedent's life, and providing an opportunity for them to speak about the decedent.

During interviews, we checked in frequently, offered breaks, asked informants how the experience was for them, and if they would like any additional resources. Interviewers also offered bereavement resources consisting of a list of local and national bereavement meetings or websites. The content of recruitment calls and informant interviews was often emotionally draining for staff as well, which was expected given the topics addressed with close contacts of decedents [11]. Additional strategies to address grief among informants may be beneficial, such as postinterview check-ins with informants to offer additional resources [30]. Future PA studies may also benefit from employing study staff with a background in psychiatric work, experience with bereavement, experience with death investigations and next-of-kin notifications, or experience working with populations who use substances [16].

Despite these challenges, we were able to effectively use PAs to conduct a large study of fatal stimulant poisoning using data from multiple independent sources. The marked social isolation of people who use stimulants and the sensitive nature of interviews were significant challenges, some of which were distinct from studies of opioid overdose. The PA requires that informants voluntarily discuss their memories and insights about an extremely distressing event [31]. Thus, future developments in PA research should focus on informant and interviewer well-being. This could include adapting recruitment language to the specific population being studied, including the use of terms such as "premature death" rather than any specific mechanism of death; offering condolences, content warnings, and check-ins as needed; seeking out staff with experience in psychology, bereavement, and substance use; incorporating postinterview check-ins with informants; and using warm handoffs of informants when feasible.

Acknowledgments

We thank the Leveraging Psychological Autopsies to Accelerate Research into Stimulant Overdose Mortality study research team, the San Francisco Office of the County Medical Examiner (OCME) staff, and study participants. This study was supported by Centers for Disease Control and Prevention (CDC) grant R01CE003364.

Data Availability

The datasets generated during and/or analyzed during this study are not publicly available due to the possibility of individual identification but are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1

XSL•FO
RenderX

No image included. Copies of outreach letter, REDCap (Research Electronic Data Capture) survey, and telephone recruitment script.
[DOCX File, 121 KB - formative_v9i1e64873_app1.docx]

References

1. Stimulant overdose. Centers for Disease Control and Prevention. 2023. URL: https://www.cdc.gov/drugoverdose/deaths/stimulant-overdose.html [accessed 2025-01-03]
2. Drug overdose death rates. National Institute on Drug Abuse. 2023. URL: https://nida.nih.gov/research-topics/trends-statistics/overdose-death-rates [accessed 2025-01-03]
3. Jalal H, Buchanich JM, Roberts MS, Balmert LC, Zhang K, Burke DS. Changing dynamics of the drug overdose epidemic in the United States from 1979 through 2016. Science 2018 Sep 21;361(6408):eaau1184. [doi: 10.1126/science.aau1184] [Medline: 30237320]
4. Friedman J, Shover CL. Charting the fourth wave: geographic, temporal, race/ethnicity and demographic trends in polysubstance fentanyl overdose deaths in the United States, 2010-2021. Addiction 2023 Dec;118(12):2477-2485. [doi: 10.1111/add.16318] [Medline: 37705148]
5. Ciccarone D. The rise of illicit fentanyls, stimulants and the fourth wave of the opioid overdose crisis. Curr Opin Psychiatry 2021 Jul 1;34(4):344-350. [doi: 10.1097/YCO.0000000000000717] [Medline: 33965972]
6. Rawson RA, Erath TG, Clark HW. The fourth wave of the overdose crisis: examining the prominent role of psychomotor stimulants with and without fentanyl. Prev Med 2023 Nov;176:107625. [doi: 10.1016/j.ypmed.2023.107625] [Medline: 37468073]
7. Coffin PO, McMahan VM, Murphy C. Evidence of pre-mortem opioid use among fentanyl overdose decedents in a safety net healthcare system. J Urban Health 2022 Oct;99(5):865-872. [doi: 10.1007/s11524-022-00663-1] [Medline: 35879486]
8. Darke S, Kaye S, Duflou J. Comparative cardiac pathology among deaths due to cocaine toxicity, opioid toxicity and non-drug-related causes. Addiction 2006 Dec;101(12):1771-1777. [doi: 10.1111/j.1360-0443.2006.01601.x] [Medline: 17156176]
9. Turner C, Chandrakumar D, Rowe C, Santos GM, Riley ED, Coffin PO. Cross-sectional cause of death comparisons for stimulant and opioid mortality in San Francisco, 2005-2015. Drug Alcohol Depend 2018 Apr 1;185:305-312. [doi: 10.1016/j.drugalcdep.2017.12.030] [Medline: 29486419]
10. Coffin PO, Suen LW. Methamphetamine toxicities and clinical management. NEJM Evid 2023 Dec;2(12):EVIDra2300160. [doi: 10.1056/EVIDra2300160] [Medline: 38320504]
11. Conner KR, Beautrais AL, Brent DA, Conwell Y, Phillips MR, Schneider B. The next generation of psychological autopsy studies. Suicide Life Threat Behav 2011 Dec;41(6):594-613 [FREE Full text] [doi: 10.1111/j.1943-278X.2011.00057.x]
12. Favril L, Yu R, Uyar A, Sharpe M, Fazel S. Risk factors for suicide in adults: systematic review and meta-analysis of psychological autopsy studies. Evid Based Ment Health 2022 Nov;25(4):148-155. [doi: 10.1136/ebmental-2022-300549] [Medline: 36162975]
13. Conner KR, Pinquart M, Duberstein PR. Meta-analysis of depression and substance use and impairment among intravenous drug users (IDUs). Addiction 2008 Apr;103(4):524-534. [doi: 10.1111/j.1360-0443.2007.02118.x] [Medline: 18261192]
14. Knoll JL 4th. The psychological autopsy, part I: applications and methods. J Psychiatr Pract 2008 Nov;14(6):393-397. [doi: 10.1097/01.pra.0000341894.35877.1b] [Medline: 19057241]
15. Cavanagh JTO, Carson AJ, Sharpe M, Lawrie SM. Psychological autopsy studies of suicide: a systematic review. Psychol Med 2003 Apr;33(3):395-405. [doi: 10.1017/s0033291702006943] [Medline: 12701661]
16. Oliver P, Horspool M, Rowse G, Parry M, Keen J, Mathers N. A psychological autopsy study of non-deliberate fatal opiate-related overdose. CORE. URL: https://core.ac.uk/download/pdf/34714006.pdf [accessed 2025-01-03]
17. Hakansson A, Gerle E. Proposing a psychological autopsy procedure for postmortem examination of accidental fatal overdose cases. J Alcohol Drug Depend 2018;6(2):1000308. [doi: 10.4172/2329-6488.1000308]
18. Manning FJ, Ingraham LH, DeRouin EM, Vaughn MS, Kukura FC, St Michel GR. Drug “overdoses” among U.S. soldiers in Europe, 1978-1979. II. Autopsies following deaths and near-deaths. Int J Addict 1983 Feb;18(2):153-166. [doi: 10.3109/10826088309027349] [Medline: 6862734]
19. Bircher J, Hahn EG. Will the Meikirch model, a new framework for health, induce a paradigm shift in healthcare? Cureus 2017 Mar 6;9(3):e1081. [doi: 10.7759/cureus.1081] [Medline: 28405531]
20. INSERM Collective Expertise Centre. Suicide: psychological autopsy, a research tool for prevention. In: INSERM Collective Expert Reports: Institut national de la santé et de la recherche médicale (INSERM); 2005. [Medline: 21348161]
21. Athey AJ, Beale EE, Overholser JC, Stockmeier CA, Bagge CL. Acute stressors and clinical characteristics differentiate death by suicide, accident, or natural causes among illicit and prescription opiate users. Drug Alcohol Depend 2020 Mar 1;208:107847. [doi: 10.1016/j.drugalcdep.2020.107847] [Medline: 31951908]
22. Hawton K, Appleby L, Platt S, et al. The psychological autopsy approach to studying suicide: a review of methodological issues. J Affect Disord 1998 Sep;50(2-3):269-276. [doi: 10.1016/s0165-0327(98)00033-0] [Medline: 9858086]
23. Appleby L, Cooper J, Amos T, Faragher B. Psychological autopsy study of suicides by people aged under 35. Br J Psychiatry 1999 Aug;175:168-174. [doi: 10.1192/bjp.175.2.168] [Medline: 10627801]

XSL•FO
RenderX

24. Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: building an international community of software platform partners. J Biomed Inform 2019 Jul;95(103208):103208. [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]
25. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381. [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
26. Hurstak E, Rowe C, Turner C, et al. Using medical examiner case narratives to improve opioid overdose surveillance. Int J Drug Policy 2018 Apr;54(35-42):35-42. [doi: 10.1016/j.drugpo.2017.12.017] [Medline: 29353022]
27. SUDORS dashboard: fatal drug overdose data. Centers for Disease Control and Prevention. 2024. URL: https://www.cdc.gov/overdose-prevention/data-research/facts-stats/sudors-dashboard-fatal-overdose-data.html?CDC_AAref_Val=https://www.cdc.gov/drugoverdose/fatal/dashboard/index.html [accessed 2025-01-03]
28. Tibshirani R. Regression shrinkage and selection via the LASSO. J R Stat Soc Series B Stat Methodol 1996 Jan 1;58(1):267-288. [doi: 10.1111/j.2517-6161.1996.tb02080.x]
29. Anderson N, McMahan V, Coffin PO. Substance use trends in San Francisco through 2022. : Center on Substance Use and Health, Department of Public Health, City and County of San Francisco; 2023 Oct 17 URL: https://www.csuhsf.org/_files/ugd/91710f_ea77b3b62d81455c8143aa2f97b2d5d5.pdf [accessed 2025-01-03]
30. Mérelle S, Van Bergen D, Looijmans M, et al. A multi-method psychological autopsy study on youth suicides in the Netherlands in 2017: feasibility, main outcomes, and recommendations. PLoS One 2020;15(8):e0238031. [doi: 10.1371/journal.pone.0238031] [Medline: 32853213]
31. Johal G, Appleby L, Turnbull P. Is there still a place for psychological autopsy in suicide research? A literature review of methodological limitations and recommendations for future development. Int Rev Psychiatry 2024 Jul 3;36(4-5):494-502. [doi: 10.1080/09540261.2024.2378075]

Abbreviations

CA-EDRS: California Electronic Death Reporting System
CDC: Centers for Disease Control and Prevention
OCME: Office of the County Medical Examiner
PA: psychological autopsy
REDCap: Research Electronic Data Capture
SUDORS: State Unintentional Drug Overdose Reporting System

Edited by A Mavragani; submitted 31.07.24; peer-reviewed by B Ray; revised version received 08.11.24; accepted 13.11.24; published 13.01.25.

Please cite as:
Antolin Muñiz M, McMahan VM, Luna Marti X, Brennan S, Tavasieff S, Rodda LN, Knoll J, Coffin PO
Identification of Behavioral, Clinical, and Psychological Antecedents of Acute Stimulant Poisoning: Development and Implementation of a Mixed Methods Psychological Autopsy Study
JMIR Form Res 2025;9:e64873
URL: https://formative.jmir.org/2025/1/e64873
doi:10.2196/64873

© Marley Antolin Muñiz, Vanessa M McMahan, Xochitl Luna Marti, Sarah Brennan, Sophia Tavasieff, Luke N Rodda, James Knoll, Phillip O Coffin. Originally published in JMIR Formative Research (https://formative.jmir.org), 13.1.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Resilience Informatics in Public Health: Qualitative Analysis of Conference Proceedings

Maiya G Block Ngaybe[1], MPH; Lidia Azurdia Sierra[1], MPH; Andrew McNair[2], BA; Myla Gonzalez[1], BSc; Mona Arora[3], MScPH, PhD; Kacey Ernst[4], MPH, PhD; Enrique Noriega-Atala[5], MSc, PhD; M Sriram Iyengar[6], MSc, MS, PhD

[1]Health Promotion Sciences, University of Arizona, Mel and Enid Zuckerman College of Public Health, Tucson, AZ, United States

[2]Department of Linguistics, University of Arizona, College of Social & Behavioral Sciences, Tucson, AZ, United States

[3]Community, Environment & Policy Department, University of Arizona, Mel and Enid Zuckerman College of Public Health, Tucson, AZ, United States

[4]Department Chair & Professor, Epidemiology and Biostatistics, University of Arizona, Mel and Enid Zuckerman College of Public Health, Tucson, AZ, United States

[5]University of Arizona, Tucson, AZ, United States

[6]College of Medicine - Phoenix, University of Arizona, Phoenix, AZ, United States

Corresponding Author:
M Sriram Iyengar, MSc, MS, PhD
College of Medicine - Phoenix
University of Arizona
475 N 5th St
Phoenix, AZ, 85004
United States
Phone: 1 281 793 4733
Email: msiyengar@arizona.edu

Abstract

Background: In recent years, public health has confronted 2 formidable challenges: the devastating COVID-19 pandemic and the enduring threat of climate change. The convergence of these crises underscores the urgent need for resilient solutions. Resilience informatics (RI), an emerging discipline at the intersection of informatics and public health, leverages real-time data integration from health systems, environmental monitoring, and technological tools to develop adaptive responses to multifaceted crises. It offers promising avenues for mitigating and adapting to these challenges by proactively identifying vulnerabilities and fostering adaptive capacity in public health systems. Addressing critical questions regarding target audiences, privacy concerns, and scalability is paramount to fostering resilience in the face of evolving health threats.

Objective: The University of Arizona held a workshop, titled Resilience Informatics in Public Health, in November 2023 to serve as a pivotal forum for advancing these discussions and catalyzing collaborative efforts within the field. This paper aims to present a qualitative thematic analysis of the findings from this workshop.

Methods: A purposive sampling strategy was used to invite 40 experts by email from diverse fields, including public health, medicine, weather services, informatics, environmental science, and resilience, to participate in the workshop. The event featured presentations from key experts, followed by group discussions facilitated by experts. The attendees engaged in collaborative reflection and discussion on predetermined questions. Discussions were systematically recorded by University of Arizona students, and qualitative analysis was conducted. A detailed thematic analysis was performed using an inductive approach, supported by MAXQDA software to manage and organize data. Two independent researchers coded the transcripts; discrepancies in coding were resolved through consensus, ensuring a rigorous synthesis of the findings.

Results: The workshop hosted 27 experts at the University of Arizona, 21 (78%) of whom were from public health–related fields. Of these 27 experts, 8 (30%) were from the field of resilience. In addition, participants from governmental agencies, American Indian groups, weather services, and a mobile health organization attended. Qualitative analysis identified major themes, including the potential of RI tools, threats to resilience (eg, health care access, infrastructure, and climate change), challenges with RI tools (eg, usability, funding, and real-time response), and standards for RI tools (eg, technological, logistical, and sociological). The attendees emphasized the importance of equitable access, community engagement, and iterative development in RI projects.

XSL•FO
RenderX

Conclusions: The RI workshop emphasized the necessity for accessible, user-friendly tools bridging technical knowledge and community needs. The workshop's conclusions provide a road map for future public health resilience, highlighting the need for scalable, culturally sensitive, community-driven interventions. Future directions include focused discussions to yield concrete outputs such as implementation guidelines and tool designs, reshaping public health strategies in the face of emerging threats.

(JMIR Form Res 2025;9:e63217) doi:10.2196/63217

KEYWORDS

resilience; public health; informatics; mobile phone; artificial intelligence; AI

Introduction

Background

Over the past few years, 2 major threats to public health have emerged. The recent COVID-19 pandemic caused an estimated 1 million deaths in the United States [1], and the long-term effects of COVID-19 are still being identified. Meanwhile, climate change is the long-term shift in predicted temperatures and weather behaviors for a given region [2]. While COVID-19 emerged within just a few months, making it an acute stressor, climate change's effects have developed gradually, making it a chronic stressor [3]. In the long term, the public health effects of climate change could be even more serious as global warming, infectious disease, air pollution, and other consequences increase in severity [4-7]. The effects of climate change threaten the physical and mental health of the public via respiratory disease, infection, stress, malnutrition, heat exhaustion, and a variety of other negative consequences [4-7].

In the aftermath of the COVID-19 pandemic, public health is grappling with the growth of misinformation [8,9], persistent health inequities [10-12], a deepening mental health crisis [13,14], rising malnutrition rates [15,16], and a myriad of other critical challenges. There is now great pressure on the field of public health to prevent a future pandemic and control public response. In navigating a post–COVID-19 world, public health is now faced with the challenges of improving communication across disciplines [17,18], preventing the spread of misinformation and disinformation [19,20], responding quickly in the face of another public health emergency [21], addressing mental health issues [13,14], and tackling the nutritional crisis within the United States [15,16]. On a global scale, the health impacts of political tensions and climate change, particularly in relation to refugee displacement, add to these challenges [22,23].

To successfully cope with these inevitable threats to public health, people and communities need to develop resilience, defined as the ability of people, households, communities, countries, and systems to mitigate, adapt to, and recover from shocks and stresses in a manner that reduces chronic vulnerability and facilitates inclusive growth [24]. It is recognized that resilience is a necessary precursor to achieving the United Nations' Sustainable Development Goals, including "No poverty," "Zero hunger," and "Climate action" [25].

What Is Resilience Informatics?

Informatics has been defined as "the science of information, the practice of information processing, and the engineering of information systems" [26], although specific definitions for the field and its subfields, such as health and medical informatics, still continue to be debated [27]. Informatics tools include data science, artificial intelligence (AI), mobile health, and augmented and virtual reality. Resilience informatics (RI) is an emerging discipline concerned with harnessing informatics to materially improve and promote the ability of people, communities, and organizations to effectively cope with natural and man-made stressors [3]; for example, there is evidence in the literature that informatics tools such as medical informatics systems or informatics tools such as telemedicine and enhanced health information systems may be able to strengthen the resilience of health care facilities [28].

While modern informatics technologies can play a major role in helping people and communities develop resilience, much formative research needs to be conducted to understand the contours of this emerging discipline. The following fundamental research and development questions need to be addressed to lay a solid foundation for RI:

- For whom should we develop RI tools?
- What are the important resilience challenges to be tackled using informatics? What are the priorities?
- How can we provide privacy and security, protect sensitive data, and account for social determinants of health?
- How can we safeguard the rights of Indigenous populations and help bridge the digital divide?
- If, as seems necessary, developing resilience entails behavior change in people and communities, how can we develop tailored and personalized tools to help support behavior change?

In addition, RI systems need to be scalable and extensible while not consuming excess energy or environmental resources that could exacerbate climate change. To answer these questions, input is needed from frontline workers across diverse fields, areas of expertise, and levels of implementation in public health and resilience areas.

The Workshop's Purpose

On November 20, 2023, the University of Arizona hosted a workshop, titled Resilience Informatics in Public Health, with experts from a variety of relevant fields, including resilience, public health, climate, informatics, policy, and technology. These experts were primarily based in Arizona, but a few purposefully selected experts from outside of the state with relevant areas of expertise were also invited to participate and present relevant information to help set the groundwork for the workshop discussions. The purpose of the gathering was to shape the future directions and applications of the emerging discipline of RI.

XSL•FO
RenderX

A secondary purpose of the workshop was to provide networking opportunities to build connections across this field and inspire new projects in this up-and-coming area of research and development of RI tools and systems. Attendees had the opportunity to meet and network with other experts in the field whom they may not have had the chance to meet in other professional settings.

Methods

Recruitment

A group of 40 experts were purposefully selected and invited via email to attend the workshop. The workshop participants were from the fields of public health, medicine, weather services, informatics, environmental science, and resilience. The experts were chosen using a convenience and snowball sampling method based on whom the RI workshop leadership had developed relationships with and who seemed to be best suited to represent the desired sectors and enrich the conversations at the workshop. The institutions represented included Weill Cornell Medicine, the University of Colorado, Northern Arizona University, the Arizona Center for Rural Health, the Mel and Enid Zuckerman College of Public Health, the Arizona Institute for Resilience, and the College of Medicine-Phoenix.

Workshop and Data Collection

During the workshop, participants listened to presentations from a few key experts to spark discussion and ideas. In between the talks, 6 groups of 3 to 7 people each were formed by workshop facilitators such that the groups contained as much diversity as possible in terms of the members' backgrounds (field of work and discipline). Subsequently, with the help of the facilitators, the attendees reflected and collaboratively discussed answers to the questions presented in Textbox 1.

During the discussions, participants were asked to write down their ideas independently on sticky notes and compile them on flipcharts to ensure that no one participant dominated the discussion. Data saturation was not discussed because this was a 1-time event, and additional data collection was not possible after the workshop. The discussions around the key questions (Textbox 1) were systematically audio recorded, and notes were taken by students from the University of Arizona who were part of the organizing team. Workshop attendees were made aware of the audio recording and note-taking and were informed that the findings of the workshop were to be compiled into a proceedings report.

Textbox 1. Questions for the Resilience Informatics in Public Health workshop discussion aided by facilitators.

Session 1 (approximately 30 min)

- What does resilience mean to me?
- To what do we need to be resilient?

Session 2 (approximately 1 h)

- What informatics tools are you using?
- What are the barriers to technological solutions? How do we address them?
- What tools do we need?
- How do we ensure equity across these solutions?

Qualitative Analysis

Three University of Arizona students conducted a rapid analysis of the notes taken during the workshop to produce a report for dissemination to the greater community for awareness. Although transcripts were not sent to participants for comment, the report was shared with participants via email to solicit feedback to increase the validity of the analysis as a form of participant checking. Workshop participants were also asked to participate in the in-depth thematic analysis process. A group of 4 students and the principal investigator worked with 1 expert attendee who volunteered to help develop the codebook. The team used a primarily inductive approach in the data analysis with a phenomenological theoretical approach [29]. The analysis team members iteratively constructed their list of themes based on the transcripts and notes using standard identification techniques [30]. The themes were compiled together using a Google Jamboard and grouped to determine the final list of codes and subcodes (Figure 1).

The 4 student researchers then coded the transcripts independently. In the case that the notetakers forgot to record the discussions, notes were used in place of transcripts. The coding team also coded the notes that were taken on sticky notes by the groups during discussions to ensure that no ideas were missed. The team then worked together to finalize the first draft of the codebook with definitions and examples (Figure 1; Multimedia Appendix 1). The codebook was entered into MAXQDA (VERBI Software GmbH), and the team began coding. The first transcript was coded by all team members, and discrepancies were resolved through discussion and consultation with senior members of the team. The remaining transcripts and notes documents were coded independently by 2 coders each. Conflicts were again resolved through discussion. The codebook was adjusted through the coding process because themes may have come up inductively. After coding all documents, the coders reviewed all documents once more to ensure that nothing was missed.

The team then wrote a narrative analysis of the findings, summarizing major points that came up to help answer the

XSL•FO
RenderX

workshop's questions. The analysis write-up process was aided in part by ChatGPT (OpenAI) for certain sections in the synthesis of the summary of the information gathered from the text; these syntheses were subsequently reviewed manually to ensure accuracy, validity, and appropriate language use in the narrative. The research team used the COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist to guide the reporting of the workshop findings and qualitative analysis [31].

Figure 1. Brainstorming document for the development process of the codebook for the qualitative analysis of transcripts and notes from the Resilience Informatics in Public Health workshop.

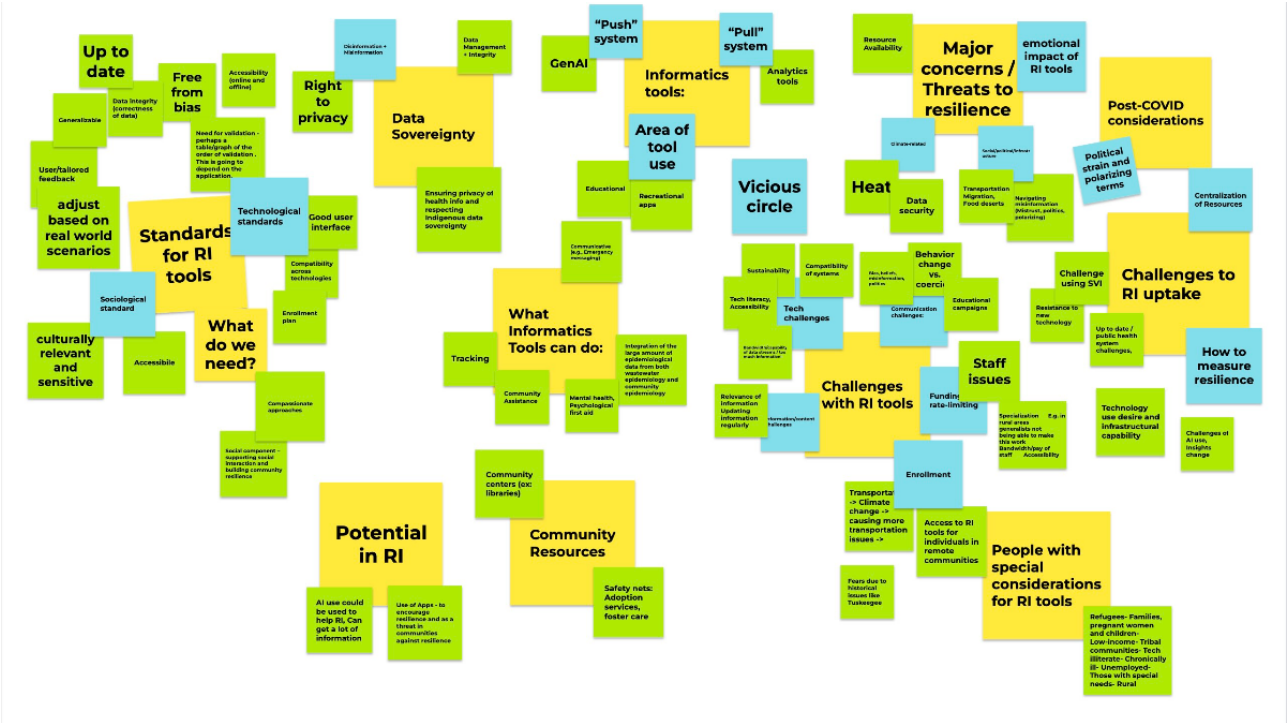

Facilitator Reflexivity

The facilitators of the workshop were 3 female professors and researchers with PhD degrees and experience in the fields of epidemiology, environmental health, climate, and resilience. One of the facilitators is a widely recognized infectious disease epidemiologist with practical experience in public health practice at a health department level and a focus on the environmental determinants of vector-borne disease transmission and control. Another facilitator is an assistant professor at the Mel and Enid Zuckerman College of Public Health with a focus on building public health capacity to address "complex" public health challenges through developing decision-support tools, enhancing science communication and translation, and integrating a health and equity lens into adaptation planning. At the time of the workshop, she was leading the Arizona Department of Health Services–Centers for Disease Control and Prevention COVID Health Disparities Initiative at the Arizona Center for Rural Health aimed at mobilizing partnerships to advance health equity and address social determinants of health related to COVID-19 health disparities among underserved populations considered high risk. Finally, the last facilitator is an experienced research professor with the Lyda Hill Institute for Human Resilience, faculty at the University of Colorado Denver and the Mel and Enid Zuckerman College of Public Health, a guest researcher with the Centers for Disease Control and Prevention, and an affiliate scientist at the National Center for Atmospheric Research in Boulder, Colorado. She focuses on the intersection of weather, climate, and health, with an emphasis on vector-borne diseases.

Workshop participants were primarily invited because they had a relationship with at least 1 of the facilitators in their line of work related to public health and resilience. The participants were familiar with the researchers and their public health interventions, which have benefited the state of Arizona, especially in relation to the COVID-19 pandemic.

Ethical Considerations

The RI workshop team reviewed the criteria for determining human subject research and concluded that the study protocol did not meet the criteria, classifying it as nonhuman subject research. The study team took all appropriate precautions to maintain data confidentiality. Participants were made aware that information collected from the conversations would be compiled in a report and manuscript for publication and dissemination of the workshop insights. Transcripts and quotes were deidentified to further protect participant identities as a precautionary measure. No compensation was offered, although travel, food, and accommodation reimbursements were provided for those who requested assistance to attend the workshop.

Results

Workshop Attendees

Of the 40 experts invited, 27 (68%) participated in the workshop at the Health Sciences Innovations Building at the University

XSL•FO
RenderX

of Arizona. Of these 27 experts, 21 (78%) were from public health–related fields (n=5, 24% with a concentration in epidemiology; n=5, 24% with a concentration in community outreach; and n=3, 14% with a concentration in medicine). Of the 27 experts, 8 (30%) worked in the field of resilience, and 7 (26%) had a concentration in a field of technology such as data science (n=4, 57%). Finally, of the 27 experts, 3 (11%) had extensive experience in emergency management or preparedness, 1 (4%) in environmental science, and 2 (7%) in meteorology.

The organizing team present included the 3 facilitators, 1 principal investigator, and 8 students from relevant fields who helped with logistics, note-taking, and audio recording of conversations.

The largest number of participants were from academia, mostly affiliated with the University of Arizona (19/27, 70%), followed by Arizona State University (3/27, 11%), and specially invited attendees from academic institutions outside of Arizona (University of Colorado: 1/27, 4%; and Weill Cornell Medicine: 1/27, 4%). Of the 27 attendees, 2 (7%) came from the Arizona Department of Health Services, while 3 (11%) were from county-level health departments (Pima: n=1, 33%; Coconino: n=1, 33%; and Maricopa: n=1, 33%); furthermore, 3 (11%) participants came from organizations representing American Indian groups (Salt River Pima-Maricopa Indian Community: n=1, 4%; and Arizona Advisory Council on Indian Health Care: n=2, 7%), 2 (7%) were from the National Oceanic and Atmospheric Administration and National Weather Service, 1 (4%) was the chief executive officer and founder of a mobile health organization (Wehealth), and 1 (4%) was from a statewide community health worker organization (Arizona Community Health Workers Association).

Codes

The codes that were identified during the qualitative analysis included the potential in RI, threats to resilience, people with special considerations, challenges with RI, standards for RI tools (technological, sociological, and logistical), and data sovereignty, among others (Multimedia Appendix 1; Figure 2). The major findings associated with the codes are included in the following subsections.

Figure 2. Frequency of codes in the transcript and notes documents included in qualitative analysis from MAXQDA.

Code System	Grace_Notes for Resili...	Rhett_RI Recording	Maiya_RI Recording	Sage_RI Recording	Andrew_Notes for Re...	Myla_RI Recording	Lidia_RI Recording	TOTAL
Standards for RI Tool	1%		9%	4%	0%		2%	4%
Technological Standard	7%			4%	19%	1%		3%
Logistical Standard	8%	0%	5%	1%	19%			2%
Iterative process	2%			3%	1%			1%
Sociological Standard	13%		1%	24%	38%			13%
Tailoring	12%		7%	7%	2%			5%
What Do We Need?	3%	5%			2%	10%	5%	2%
Communication	12%				1%			0%
long-term solutions	7%			9%	9%			5%
Challenges with RI Tools				1%	0%		15%	2%
Funding-related issues	2%		7%		11%			2%
Accessibility	10%	1%		1%	14%	8%	2%	2%
Communication	3%	1%	11%	4%	3%		5%	4%
Information			11%		1%			2%
Centralization of Resources	5%					12%		2%
Measuring Resilience	2%	3%		4%			43%	6%
Communication between disciplines	3%			9%	0%			5%
Clinical Translation				1%				0%
Mixed methods approach				6%				3%
Human centered design	1%			0%	3%			0%
succession planning				2%				1%
Behavioral economics				3%				1%
Generational differences				11%	2%			6%
Infrastructure					2%			0%
Elasticity / Scale	2%		2%					0%
Real-time response	2%		11%	7%	1%			5%
Data Sovereignty	3%		3%	0%	8%		1%	1%
Disinformation + Misinformation			1%	2%			1%	1%

Potential in RI

RI tools were mentioned in many of the conversations. Specific examples included the 4 tools highlighted during the expert presentations at the event—Pilas, AZCOVIDTXT, AZX, and Wehealth—as well as ≥200 essential tools for public health data monitoring (Textbox 2).

RI tools were also used for educational purposes such as community education, health care worker training and simulation tools. Such tools were also mentioned as having a multitude of different purposes and functions (Textbox 3).

XSL•FO
RenderX

Textbox 2. List of key existing resilience informatics tool examples mentioned by participants in the workshop.

- Organization assessments
- Quality assurance tools
- Heat maps
- Telemapping tools
- Mental health screening
- The 988 public helpline for mental health

Textbox 3. Purposes of key existing resilience informatics tools.

- Creating clinical guidelines with community perspective and provider perspective
- Recording membership sign-ups
- Listing contact information for outreach in emergencies
- Taking raw data and putting it into a usable format
- Digesting complex information, summarizing relevant information, and then pushing it to the relevant audience
- Recording how many people have enrolled in a course
- Evaluating a course or intervention
- Measuring the intensity of reactions from the public
- Surveying the prevalence of infectious diseases or chronic illnesses
- Acting as a decision tool for gauging and assessing data
- Projecting the impact of an intervention

Finally, emerging technologies such as generative AI hold the potential to enhance the impact of RI, as highlighted by a participant:

> *So this is actually a really hot area of research for generative AI. And you have to look at it a little bit differently than what you see on the surface. But what generative AI is, you can feed it a corpus of knowledge. And it builds a giant tree out of that knowledge. And so when you ask a specific question, it'll try to navigate that part of the tree and find those things...I mean, there's reasons to be skeptical because basically the ChatGPT systems that we know have gotten their information from all over the internet. There's some really bad information. But there are other ways where, for example, Maricopa County could put all of your policies, plans, procedures and every...Build your own corpus of knowledge and use the same generative AI algorithms so that you could ask a question like, "How do I answer this question?" And so then it's taking your internal information, sort of trying to like capture your brain and everyone else's brain and all of the documents and you can produce stuff.* [Participant from Arizona State University Knowledge Enterprise]

In summary, RI tools have the potential to span a wide range of functions. While many RI tools are already available, there remain numerous possible areas for further expansion.

Threats to Resilience

Resilience faces challenges across various health sectors, which are often intersecting and complex. Protecting the resilience of communities and ensuring their health and well-being begins with understanding these threats and how individual communities are impacted.

> *Health care access was one of the threats to resilience mentioned at the workshop. The cost of health care alone makes it inaccessible to many groups of people or, at the very least, a burden to individuals who are barely able to afford health care. However, threats to health care access extend beyond cost because the geographic location of health services also plays a critical role. Our participants also mentioned that within health care services, workforce strain can pose a threat to public health. Health practitioners' mental health was also mentioned as having a direct impact on the quality of treatment that patients receive. Overworked health care professionals introduce the potential for practitioner mistakes or lead to a shortage of health care professionals in the community, as noted by a participant:*
>
> *But then you take into account that 30% of the public health workforce left the profession during COVID. And one of the biggest challenges in health departments right now is retention, not just retention of people, retention of historical knowledge and community understanding. I mean, of this idea. And I'm like, "Yes." But it really sucks when you have a*

XSL•FO
RenderX

> *room full of people who don't really know, understand and haven't worked, been on the ground during the pandemic, haven't done these things, don't know the community.* [Participant from Arizona Advisory Council on Indian Health Care]

Inadequate infrastructure was mentioned by our participants as a large threat with respect to the centralization of resources and transportation access, as previously discussed with health care access. However, the threat of infrastructure extends beyond these areas, encompassing the housing crisis, internet access, and utilities. The centralization of resources and transportation services is necessary to protect the health of communities, as discussed before, in terms of the ability to access health services. However, beyond health care, a lack of transportation services and decentralized resources continue to threaten resilience because transportation is crucial for individuals in food deserts to be able to reach grocery stores and for those in areas of job shortages to reach work. Housing accessibility and affordability were also mentioned during the workshop because housing prices have increased in the past few years in Arizona.

In the case of water, communities rely on having access to safe, running water for consumption and bathing. This important resource, especially in the desert environment of Arizona, has not always been adequately protected during droughts and instances of contaminated water supply, as mentioned by a participant:

> *So, what has happened in a lot of the orchards, like the nut orchards, have moved from California to the Willcox area [in Arizona] to pump the water out because they don't have any restrictions. And so basically, they're sucking it dry. So, the concern is, there won't be any water in the future, 50 years from now, in that area.* [Participant in environmental science field at the University of Arizona]

Internet access was mentioned as being of great importance for protecting health as well because much health-related information and many health services are available on the web. Being able to view this health information or find health services through web searches is important for individuals to be able to protect their health.

Large-scale events such as the ongoing climate crisis were mentioned during the workshop as one of the greatest threats to resilience. Climate change has negatively impacted water resources, raised temperatures, and increased the prevalence of infectious diseases. Furthermore, modern infrastructure is not equipped to withstand the challenges that extreme weather events pose, leaving communities more at risk of being devastated by such an event. One participant stated as follows:

> *How, for me, is its emergency response, but it's also how to prevent and mitigate issues with climate change. And we're getting into it with the hottest heat on record in the city of Phoenix. I'm trying to get more involved with getting areas that don't have cooling centers, doesn't have things like that, like the tribal nations...Outlying areas of, even in the Phoenix area, doesn't have cooling centers...So, we see a big phase for central Phoenix and getting in some rural areas, but when you start going outside the Phoenix area or into tribal nations, there's nothing to help them recover or even get supplies. So, to me, that's a big issue. I live in Queen Creek, so even looking at my own town, they have a station. Yeah, it's the fire station. You drop off the water, but you have no cooling stations for the homeless or individuals or even the reservations.* [Participant from the Arizona Coalition for Healthcare Emergency Response]

The climate crisis has become even more difficult to address due to its politicization.

Political strain was mentioned at the workshop as another threat to the resilience of communities in working against the polarization of health concerns. Participants mentioned that during the COVID-19 pandemic, the communication of accurate information proved difficult because the public was inundated with mass amounts of misinformation and disinformation. One participant stated as follows:

> *Politics can modify how we approach public health disasters.* [Participant from a county health department]

Communication emerged as an individual threat to the resilience of communities because approaches to addressing the COVID-19 pandemic were disjointed and limited as a result, according to our workshop participants. There were disparities in information between clinicians and public health practitioners who lacked the means to effectively engage with each other. The inability to properly communicate exacerbated the issue of trust in information and the prevalence of misinformation. Creating adaptable communication systems can pose difficulty in the field of public health, as mentioned by a workshop participant:

> *I think uncertainty too. There's a lot of the things we're talking about happening in the future. I think especially when we talk about extreme heat and things like that, or even other things like how AI is going to come into play. There's just a lot of uncertainty of what's going to happen. I think that really heavily influences the way people respond to information. How do you communicate something that, things that are changing, things that are uncertain, and how do you make decisions in an uncertain environment, the burden that puts on a household person trying to plan financially or somebody just mentally the load, or even a city planner trying to figure out what's the best decision?* [Participant from Arizona State University Knowledge Enterprise]

Issues related to access to *resources* also emerged as a potential threat to resilience for the communities our participants serve. A lack of financial resources limits what the community can afford to support its health, as noted by a participant:

> *And it's for our community organizations so that they can leverage that piece to apply funding, to ensure there's more programs within the services, within the communities they serve. Because we can't do*

XSL•FO
RenderX

> *everything. We don't have the capacity.* [Participant from a county health department]

Other resource-related issues mentioned by participants included food supply and air.

Several *social issues* were mentioned during the workshop that directly work as threats to resilience, with public health professionals facing concerns surrounding food insecurity, the housing crisis, transportation, migration, racism, gentrification, war and politics, cultural sensitivity, the cost of living, employment security, and the education system. One participant stated as follows:

> *Yeah. I think distrust is more just the fact that if we need to be resilient, we need to trust each other and we don't anymore. And so that's a thing we need to be aware of. How do we build trust in communities? And the disinformation is more about are we just getting the right information or not? That's why I meant trust between communities.* [Participant who is a software development professional and chief executive officer of a mobile health organization]

Mental health was a concern raised by several participants; for instance, aging populations were mentioned as being vulnerable to the empty nest syndrome and isolation issues. In addition, substance misuse was directly related to mental health (eg, the growing prevalence of drugs such as fentanyl). Concerns surrounding the lack of empathy and social capacity were discussed in depth in 1 group at the workshop. The lack of empathy in American culture was mentioned as possibly stemming from the colonial mindset focused on individualism. By contrast, social capacity was identified as an issue due to the ever-growing role of mobile phones and social media today, especially among younger generations. Mindful interventions were mentioned as having the potential to address these concerns and help ensure that populations remain socially capable. One participant stated as follows:

> *It's compassion fatigue or something. You get so much information because you're connected to the internet, you just get bombarded every day.* [Participant from a county health department]

People With Special Considerations

A recurrent theme that emerged during the workshop discussions was that certain communities may face greater challenges to resilience than others, as mentioned by a participant:

> *No one would choose not to have running water.* [Participant from a county health department]

It was apparent in many of the discussions that some communities lack basic resources such as consistent access to water. Indigenous communities, asylum seekers, the aging and older adult populations, people with disabilities, non–English-speaking communities, and other groups considered marginalized often face unique challenges to resilience that merit special considerations when considering the use of RI tools.

Asylum seekers and refugees were mentioned as having challenges related to privacy and access to resilience tools because they are vulnerable to exploitation and often face barriers to legal recourse. The aging and older adult populations, people with disabilities, non–English-speaking communities, and other groups considered marginalized were also mentioned as needing special consideration when designing RI tools (eg, considering translation, accessibility, and other tailoring approaches to fit their needs).

Challenges With RI Tools

Some workshop participants mentioned that RI tools face challenges of usability that must be understood and addressed to maximize the positive impact of the tools in improving community resilience while also preventing possible adverse effects.

Accessibility is the first of these challenges. RI tools can be limited in impact based on the size of the audience they can reach and effectively assist. Tools should be tailored to the community being served and designed with end users in mind. Elasticity and scalability were mentioned as issues that should be at the forefront of RI tool design considerations so that the tools may serve the greatest number of communities, which can vary in size, geography, and so on. Accessibility concerns further include issues of resource centralization, language barriers, technology literacy, and related considerations. One participant remarked as follows:

> *So, one thing too, I would say the integration of social determinants of health in city planning. So general plans as well as emergency response plans. One of the things we learned during COVID is a lot of our emergency response plans really weren't built to address supporting diverse populations, multilanguage speakers, all of these other things. And so being able to go in and make those the norm now.* [Participant from a county health department]

Real-time response was mentioned as being especially desirable in RI tools. Between collecting and analyzing data and publishing the results, the process can be all too slow to be able to properly respond to fast-moving issues that may threaten resilience, such as a pandemic. RI tools must be able to quickly respond in times of crisis to be able to limit the amount of people who are affected by the crisis, as highlighted by a participant:

> *Informatics can assist me in finding what that trigger point is, but it can't assist me in actual response. It's too late. Once it's already happening, now I'm in the door. I'm working with a problem and actually implementing things. Informatics takes a backseat; they don't care anymore. But I want informatics to inform that response. That would be very valuable to tell me, "Hey, you're done. You can stop now. You can turn it off." That would be nice. But I don't know how informatics can assist in the response phase or the recovery phase.* [Participant from a county health department]

Funding was mentioned as being a particular challenge for existing public health and resilience-building initiatives. RI tools and projects must be able to secure funding, such as through grants, to establish and sustain the tool. An example

XSL•FO
RenderX

given by a participant concerned the difficulty of fitting grant expectations. If an RI project does not fit expectations, adjustments will need to be made to maintain funding, and the project's efficiency may suffer. Funding is necessary for developing RI tools, but it is also a potential threat if the tool does appeal to funders. A participant stated as follows:

> *Make recommendations to other departments and divisions within the health department to say, "Hey, we need to allocate resources here. Here's what we're hearing." And to your point, we're listening to people's individual stories. We're also looking at data as a whole. And the reality is, we've tried to push a lot of these dashboards and a lot of these things to the general public...They're not really interested in a lot of that unless it's very easy to use and speaks directly to them. So again, my job is that, "Okay, that's not working. And how can I frame it in a way that's really going to focus on that individual so they can make an important decision themselves?" So a lot of what we've done is really just change the way we present data.* [Participant from a county health department]

RI tools are also tasked with the challenge of measuring resilience, especially across the various fields that contribute to resilience research; for instance, resilience must be defined in a standard way, and units of measurement may need to be standardized according to the definition. A group of participants had the following discussion:

> *Speaker 6: Do you guys think that's a problem, the way that we measure their resilience, and solve problems in this case? So the thing that we should think about how we are measuring this resilience about other things right here, I don't know if we have the right tools right now to do that. The resilience measurement tools.*
>
> *Speaker 9: And then the definition of resilience, which alters person to person, agency to agency.*
>
> *Speaker 6: Yeah, that's true.*
>
> *Speaker 9: Race, ethnicity, there's so many implications, things that affect that. So I always look at it as you have to define what the meaning is first in order to measure it. Otherwise, its positive outcomes, just how you go about it. What do you define as success in the situation? Again, that's always going to alter and shift.*
>
> *Speaker 8: Or how do you weigh different? If you can agree that resilience is sort of a combination of metrics that we already account for, okay, so we're counting. We're trying to measure things like food security and vulnerability, heat and different mental health metrics. But then you try to combine that into some sort of measurement of resilience. It's like, "Well, how do I weigh in some sort of statistical way, the weight of my mental health versus my heat resilience, which are all..."*

Later, in the same group, a participant stated as follows:

> *We spent all this time trying to measure the wrong things, like risk. Risk is great. Let's take probability times consequence and its risk. And then we've got everything solved and we're measuring the wrong things. So at least now that we can talk about resiliency and all of the constructs that come with it, you package it up and you've got to talk about social cohesion and adaptivity and vulnerability, and then you sprinkle in some of that probability and consequence stuff. But that's so reductive that you miss the point of what resilience is, which it's an intended positive outcome about how we can work together and have shared goals that are different than individual goals. But you have to have all of that. So, I think it's a really positive thing for results is hard to measure, and you have to roll around in the mud with a pig for a while in order to put some thoughts together. Yeah, it is refreshing.* [Participant from Arizona State University Knowledge Enterprise]

Technological Challenges With RI Tools

Several challenges that are more technological were also mentioned by many participants; for instance, infrastructure issues such as broadband access are problematic for the effectiveness of RI interventions, which often rely heavily on the presence, functioning, and quality of technological tools. At times, government intervention may aid in increasing the necessary infrastructure; for example, some government programs could help to provide free Wi-Fi or expand hotspots. Rural communities were mentioned as being especially vulnerable to such issues, where there may be higher turnover in technology. In this case, low-cost smartphones may be needed to support these communities, or perhaps a community-operated mobile phone provider could be enlisted to help increase mobile phone coverage and access.

Communication

RI communication faces a spectrum of barriers and facilitators, prominently highlighted by the necessity for 2-way communication. This involves ensuring a cohesive message and fostering a shared understanding among stakeholders. However, intergenerational disparities present challenges because differing communication preferences and styles may hinder effective dialogue. Establishing rapport amid these differences can prove arduous, as illustrated by the struggle to navigate technological nuances and preferences, such as the preference for traditional email over newer communication methods.

Moreover, the landscape of informatics communication is marked by silos, where various disciplines converge, including developers, computer scientists, behavioral economists, public health practitioners, and social scientists. This interdisciplinary nature can be both a boon and a bane. While it offers diverse perspectives, it also necessitates concerted efforts to bridge disciplinary divides. One expert at the workshop asked the following question:

> *What about in nonemergent situations? Nonemergent public health situations? How can that gap the bridge? And I'm talking about between the academics*

XSL•FO
RenderX

> *in between the field of public health and then between the clinical side? That's my question.* [Participant from Arizona Advisory Council on Indian Health Care]

In addition, the gap between the clinical and public health realms underscores the need for enhanced collaboration and communication strategies to align goals and priorities effectively. Overcoming these barriers requires deliberate efforts to foster a culture of inclusivity and open communication across sectors and disciplines.

Components of an RI Project

Several different essential components of a successful RI tool or project were mentioned during many of the conversations (Textbox 4).

Textbox 4. Key considerations and essential components when designing a resilience informatics (RI) intervention.

- Systems and infrastructure to process data (broadband access)
- Data integration hub and intelligence subsystem for information sharing
- Anomaly finder (could be an artificial intelligence [AI] function with data search)
- AI to digest complex information, summarize relevant information, and then push it to the relevant audience
- Centralized hotline or integration to allow individuals to have a resource to call
- Bidirectional flow (receive community input)
- Equity assessment (part of before, during, and after evaluations)
- Human elements and user tests
- Coalition oversight to have a big-picture view and report back
- Communication component to hit all different networks

Standards for RI Tools

Overview

A major theme that emerged from the RI workshop discussions was the need for standards or guidelines to help guide the design and implementation of RI tools in the future. Through an analysis of participant conversations, a list of standards for RI tools was developed based on the characteristics presented as ideal for RI tools. While it is recognized that this list may not be comprehensive and that some of the standards may apply to some but not all tools, it may serve as a starting point for future iterations of RI tool standard development. The standards have been grouped into 3 categories based on the nature of the characteristics in question: technological, logistical, and sociological.

Technological Standards for RI Tools

Technological standards were mentioned in many of the discussions. These standards may be helpful for software developers, statisticians, and other technical professionals involved in RI projects.

RI Tools Should Be Up to Date

Similar to other forms of technology, RI tools must be regularly maintained and updated to ensure continued accessibility and usefulness, according to our participants. RI practitioners must also put into place standards for data used by RI tools. This may entail connecting multiple resources to get a broader scope of data, training AI tools by using data from diverse populations to remove bias, establishing standards to ensure the accuracy of data, and using validated tools and objective measurements when implementing RI.

Data Must Be Accessible

Another concern regarding RI tools mentioned in the expert discussions at the workshop was accessibility: data should be designed to be accessible (following accessibility guidelines) while still being protected. Tiered access to data (eg, nonpublic facing or anonymized reports to preserve the privacy and security of data) should be implemented. There should also be regulations in place related to the sharing of private corporate data, such as the number of COVID-19 test kits sold.

RI Tools Should Be Interoperable

Furthermore, RI tools need to be optimized for all major platforms to ensure interoperability across different technologies. This could involve creating versions of RI tools for Apple and Microsoft operating systems, among others.

Data Sovereignty and Privacy Should Be Prioritized

Finally, protecting data sovereignty and privacy must be of utmost importance for RI tool developers and users. This especially applies to the use of RI tools with populations who may be particularly susceptible to abuse or exploitation by bad actors. Data-sharing agreements can be beneficial tools for parties collaborating on resilience-related public health projects with RI tools; however, these agreements must explicitly be written to protect the sovereignty, ownership, and privacy of the data collected and shared.

Logistical Standards for RI Tools

Another type of RI tool standard that emerged from the discussions among workshop attendees focused on logistical characteristics. These are more related to the planning, organization, design, and practical aspects of implementing or rolling out the tool, rather than the technological or sociological aspects of the tool or RI project.

XSL•FO
RenderX

RI Tools Require Thoughtful Design

It was mentioned that when designing RI tools or interventions, it is important to use a systematic approach, following step-by-step procedures. These procedures should include data collection and performance evaluation to ensure continued monitoring and evaluation of the project, ensuring that objectives are met and that the necessary information is available to demonstrate the effectiveness of the intervention. Examples of disciplines or approaches from which RI tools could draw best practices include human-centered design, mixed methods research, and clinical translational science. The acceptability of the desired behavior that the RI tool may be promoting should always be considered as well in the design and implementation of the project. The design could additionally include an implementation model, considering infrastructure, service, providers, users involved, and so on.

The Process of RI Tool Development Should Be Iterative

The workshop discussions highlighted the need for continual community testing and evaluation. One participant stated as follows:

> *We're always reinventing.* [Moderator from the University of Colorado]

RI Tools Need to Have Real-Time Responses

RI tools should work in real time and be able to pivot and adapt quickly to fit the needs of end users. Rapid and clear information should be conveyed in a simple report afterward as well.

RI Tools Need to Have Multiple-Layer Solutions

RI tools should function at a variety of levels to meet the needs of the communities they serve; for example, an ideal RI tool should consider all channels that the end users may use: SMS text messaging, apps, social media, and so on.

RI Tools Require Funding

While perhaps evident, RI tools need funding to be most effective, and this was emphasized many times at the workshop. In the planning and design stages, it is important to consider where these funds will be coming from to make sure that the tool is most effective and sustainable: grants, sales, and so on. Multiple funding streams may be best to consider, should it be possible. In addition, funding is important to consider when budgeting. One participant stated as follows:

> *How can I give our community partners something that they can turn around and use as leverage for money, resources, whatever?* [Participant from a county health department]

Sociological Standards for RI Tools

The following sociological standards for RI tools were identified at the workshop.

Focus on Community-Identified Needs First

Our participants mentioned that RI tools should prioritize populations with the greatest needs within the communities being served. To do so, stakeholder engagement in critical conversations about design and intervention implementation is crucial.

Build Trust: "Partnerships Are Key!"

Upfront community engagement is key to RI tool success. Community leaders and partners, particularly gatekeepers to the communities, need to be engaged in the RI tool from the outset. Community members should be additionally involved in decision-making surrounding resource allocation if this is relevant to the RI tool or project. Successful community engagements are characterized by humility; in other words, our experts mentioned that it is important to be upfront about what you do or do not know. Along the same lines, community values must be recognized early on to ensure successful community engagement, such as transparency, humility, and so on. To identify these values and develop community partnerships, we must listen to groups (eg, tribal communities), understand and learn from others, and recognize that all people have important information, no matter their education level or background. Cultural humility and responsibility as well as community-based participatory research methods may provide important lessons from which RI tool developers and implementers can gain a better understanding of successful community engagement and cultural value identification.

Integrate Social Determinants of Health

The RI workshop participants mentioned how important it is to integrate social determinants of health, such as income, education level, and health care access, into the design, implementation, and evaluation of the RI tool. These factors are crucial when considering public health outcomes because they can influence and potentially confound the tool's impacts.

Long-Term Solutions and Sustainability Should Be the Goal

The RI tool must have a long-term vision and sustainability as the goal to most effectively build resilience in communities. One participant stated as follows:

> *Systemic change takes time, and it builds, and you've got to take things and use things to your advantage.* [Participant from the Arizona Department of Health Services]

Along these lines, the RI project will need a way to follow up with participants over time to continually evaluate and assess the long-term impacts of the RI intervention.

Tailor the Intervention

Tailoring was a strong theme in many of the conversations that experts had at the workshop. Each RI tool must be tailored to the specific population of interest. One participant stated as follows:

> *So you know what's dangerous about that idea is that it's measuring urban capability against rural ability, because you're talking to the one of two people in the entire county that is thinking about public health weather. There is no one else.* [Participant from a county health department]

RI tool designers must additionally contextualize the problem not only by population but also by the specific event as well.

Participants also mentioned that intergenerational solutions (eg, physical community spaces) are important when developing RI tool interventions. Language must always be considered—not

XSL•FO
RenderX

only in its conventional sense but also in terms of vernacular expressions and subgroup dialects—to increase understanding when using RI tools to communicate with different communities. Cultural representation must also be considered when tailoring interventions, considering differences in emergency response language and acronyms and literacy levels. One participant warned against reliance on AI translation, although it seems that AI tools are evolving at a rapid rate and may improve translation capabilities over time. In the meantime, it may be crucial for most communities to either rely on manual translation with the help of local community members or, at the very least, double-check AI-generated translations through manual review. Multimedia and multimodal strategies were suggested as being important to address language barriers or varying literacy levels. Differences in technology accessibility must also be acknowledged and taken into consideration in RI tool design and implementation. To be most inclusive, low-tech or even no-tech alternatives may need to be integrated into the design. One participant remarked as follows:

> *I always say, "It's the generation of someone who prefers a phone call than a text."* [Participant from the cooperative extension office, affiliated with the University of Arizona]

Finally, it was emphasized that it is important for all RI tools to be inclusive for people with disabilities.

Equity Must Be Prioritized

During the workshop, participants were asked to answer a question about how equity would be prioritized in the RI tools; for instance, it is important to collect data to evaluate equity within the RI intervention, such as documenting equity awareness and feedback. Another way to prioritize equity in the RI project could be by involving people from the community in project roles and paying them equitably. Assessing policies and tools (eg, AI) for discriminatory impacts may be an equitable goal or activity for an RI project. Finally, it was recognized in a conversation that the Culturally and Linguistically Appropriate Services standards are an existing guideline that may be useful for RI tool developers to reference and use to increase equitable outcomes.

Issues Related to Data Use

Two main issues emerged in the discussions regarding data use (Textbox 5).

Textbox 5. Issues regarding data use and illustrative quotes.

Data are not always presented in an appealing or accessible way for practitioners

"The informatics can inform our response, but we might be just focusing so heavily on the data that the practitioners out in the field start going, yeah, it's a wildfire, what do you want me to do about it? It's going to burn. It's going to burn that long." [Participant from a county health department]

The right information must be there for somebody to grab it

"I used to develop a lot of software...and we used to do pharmacy applications and our user group would say, 'We need these 20 data fields.' We go, 'Yeah, that makes sense. Okay, we'll put the 20 data fields in.' Three years later we go back, two of them are used. Now did that mean the other 18 weren't important or shouldn't have been used? It probably meant that there was this really subset that needed those data fields to make better decisions. But if you looked at it from utilization of the data field, you would say, 'Well, just get rid of them, we really don't need them.'" [Participant from a county health department]

Indigenous Data Sovereignty

Indigenous data sovereignty was a key theme that emerged from the guest presentation from the Salt River Pima-Maricopa Indian Community and the discussions among the workshop attendees. Indigenous data sovereignty can be defined as "the right of Indigenous Peoples to govern how data from or about them is collected, accessed, used, stored, and disposed of" [32]. Culturally centered sovereignty requires a conscious effort by public health officials, researchers, policy makers, and RI practitioners alike. Explicit protections and acknowledgments of Indigenous ownership of data and the rights surrounding these data must be put into place before RI tools are deployed in Indigenous communities. This involves an acknowledgment of cultural practices, the political sovereignty of Indigenous nations, collective knowledge and shared wisdom, and the relationship between Indigenous peoples and the land.

Martinez outlined 4 principles that must be considered for Indigenous governance known as the "CARE principles": collective benefit, authority to control, responsibility, and ethics. Following these principles is a minimum standard for public health work in Indigenous communities, particularly with respect to sensitive health data collected about Indigenous peoples [33].

Discussion

Principal Findings

In summary, the RI workshop experts confirmed the need for tools that translate technical knowledge into something useful for the community. These tools must be low cost or free, accessible (easy to use), available on any platform (especially mobile phones), and simple in design; not require special domain knowledge; and be connected, bidirectional, responsive in real time, and respectful of data sovereignty. Technological and multimodal alternatives tailored for different groups with varying accessibility needs are crucial for the success of RI tools. The end goals of these RI tools are sharing resources, effecting behavior change, and advancing health equity.

Comparison to Existing Literature

Some of the findings from our workshop resonate with those in similar studies in the literature; for instance, our attendees confirmed that informatics tools have the potential to help build resilience by improving the capacity for households to respond to and make effective decisions during disasters, as well as through improving social capital [34]. Another finding that resonates with existing literature is that there is a need for a

XSL•FO
RenderX

structured guide for the development of informatics tools [35]. Other studies have discussed logistical issues such as data quality and difficulties related to collecting data being obstacles to the success of health informatics projects in addressing pandemics [36]. In addition, similar studies on the COVID-19 pandemic have noted that human factors, ethics, data privacy, and the diversity of participants are vital to consider in the implementation of medical informatics tools [37]. Another study emphasized the iterative nature of informatics tools, necessitating continuous evaluation and reformulation for success [38]. This paper presents a few initial ideas that could be used to begin developing a checklist or framework of standards for future RI tools. Finally, it was demonstrated in this summary from our workshop that RI, and health informatics in general, encompasses a very wide range of functions and has the potential to help make health care services more resilient in the face of climate change impacts, similar to findings from other studies on health informatics [39].

It is worth noting that our workshop did not include much explicit discussion about big data and its potential, which would be an important area to explore in the future, especially considering the potential it has shown in building resilience in other sectors [40,41]. Another important issue that did not come up, despite some conversation about the positive impacts that informatics may have on reducing workforce strain, was the concern that informatics may impact jobs in health care and resilience-adjacent sectors if not introduced with care [42]. Other concerns include the fact that while these tools may be emerging, the workforce to implement them may not be readily available [43]. This topic has been explored in the literature and has been a concern with many other major technological breakthroughs and should be explored further to preemptively avoid any major impacts in restructuring the workforce to ensure that new informatics tools do not have large negative impacts [42].

Strengths and Limitations

This workshop included a wide range of participants from different fields, with each being an expert in their respective area, thus lending credibility and wide-reaching relevance to the findings. A limitation of this workshop is that the findings may be limited primarily to Arizona because most participants were from the state. Nonetheless, it was recognized that many of the discussions were relevant to contexts outside of Arizona as well. In addition, because this was not a qualitative study with a systematic approach to data collection, the findings may not be comprehensive, and future research should be conducted to ensure that a saturation of ideas is reached to gain a fuller understanding of the reach and potential of this field.

Future Directions

Future research questions to explore that emerged from the workshop discussions included the following:

- How do you physically change the environment to protect against infectious disease?
- How do we leverage community knowledge? What is the compensation for this?
- How do we teach empathy?
- What standards are necessary for RI tools to be most effective at addressing public health and climate change–related issues?

It is important for future research in this area to adopt mixed methods and be interdisciplinary, human centered, and community based. The research should ensure that equity considerations are at the forefront.

A follow-up workshop was held in June 2024 in Tucson, Arizona, to present the findings and work toward developing solutions to resilience and public health problems together with the same experts from this initial workshop. For future directions in the field of RI developments, it would be helpful to have more focused discussions with clear outputs such as implementation guidelines or designs of RI tools to put into practice. It is important to continue to build relationships with communities, stay engaged and curious about different cultures, and learn from others.

Conclusions

The RI workshop described in this paper has laid a critical foundation for advancing the development and application of RI tools, especially in the context of public health and climate change. By emphasizing the need for accessible, user-friendly, and equitable technologies that can effectively bridge technical knowledge and community needs, this workshop has underscored the importance of multidisciplinary and community-centered approaches. The discussions highlighted key areas for future research, including physically transforming environments to mitigate disease risk, leveraging community knowledge, and developing empathetic and culturally sensitive tools. With a follow-up workshop planned, the next steps involve moving from discussion to actionable outcomes, focusing on the creation of implementation guidelines and tool designs that prioritize health equity and resilience. In conclusion, continued collaboration between experts, communities, and interdisciplinary teams is essential to realizing the full potential of RI in addressing public health challenges in an era of rapid global change.

Data Availability

The datasets generated during and analyzed during this study are not publicly available to protect the privacy of our participants but are available from the corresponding author on reasonable request.

Acknowledgments

The authors recognize and thank all partners who helped them with developing, and by participating in, this workshop, including Antonio Hernandez, Charles Gerba, Charles Kmet, Darien Fuller, Jose F Florez-Arango, Kerri Rittschof, Patrick Wightman,

XSL•FO
RenderX

Sameer Halai, Sheryl Nelson, Summer Hassan, Sybil Peters, Theresa Cullen, Tim Lant, Yevheniia Varyvoda, Enrique Noriega, Anna Hartman, Ken Komatsu, and Jahnavi Shriram, as well as the others who participated in this workshop. Special thanks to the organizing team, including Rhett Hill, Grace Bohan, Sage Vu, Mary Hayden, Zackry Guido, and Corrie Hannah. The authors additionally acknowledge the use of a generative AI tool (ChatGPT) [44] to aid in creating parts of the qualitative analysis in the Results section. All generative AI output was reviewed and edited by a human to ensure accuracy and validity before the final manuscript was submitted.

This research was funded by a grant from the University of Arizona Technology Research Initiative Fund or Water, Environmental, and Energy Solutions. The authors gratefully acknowledge support from this fund and also from Dr Sharon Collinge, Zack Guido, and the Arizona Institute for Resilience.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Codebook with definitions and examples used for qualitative coding and analysis.
[DOCX File , 18 KB - formative_v9i1e63217_app1.docx]

References

1. Covid-19 map. Johns Hopkins University & Medicine. 2024. URL: https://coronavirus.jhu.edu/map.html [accessed 2024-12-09]
2. Sustainable Development Goals: climate change. United Nations. 2023. URL: https://www.un.org/sustainabledevelopment/climate-change/ [accessed 2023-06-12]
3. Iyengar MS, Merchant N, Ernst K, Rains S, Arora M, Ngaybe MG, et al. Resilience informatics for public health. Stud Health Technol Inform 2024 Jan 25;310:1276-1280. [doi: 10.3233/SHTI231170] [Medline: 38270020]
4. Frumkin H, Hess J, Luber G, Malilay J, McGeehin M. Climate change: the public health response. Am J Public Health 2008 Mar;98(3):435-445. [doi: 10.2105/ajph.2007.119362]
5. Huang C, Vaneckova P, Wang X, Fitzgerald G, Guo Y, Tong S. Constraints and barriers to public health adaptation to climate change: a review of the literature. Am J Prev Med 2011 Feb;40(2):183-190. [doi: 10.1016/j.amepre.2010.10.025] [Medline: 21238867]
6. Gould S, Rudolph L. Challenges and opportunities for advancing work on climate change and public health. Int J Environ Res Public Health 2015 Dec 09;12(12):15649-15672 [FREE Full text] [doi: 10.3390/ijerph121215010] [Medline: 26690194]
7. Haines A, Kovats RS, Campbell-Lendrum D, Corvalan C. Climate change and human health: impacts, vulnerability and public health. Public Health 2006 Jul;120(7):585-596. [doi: 10.1016/j.puhe.2006.01.002] [Medline: 16542689]
8. Varma R, Verma Y, Vijayvargiya P, Churi PP. A systematic survey on deep learning and machine learning approaches of fake news detection in the pre- and post-COVID-19 pandemic. Int J Intell Comput Cybernet 2021 Aug 16;14(4):617-646. [doi: 10.1108/ijicc-04-2021-0069]
9. Mirza SA, Sheikh AA, Barbera M, Ijaz Z, Javaid MA, Shekhar R, et al. COVID-19 and the endocrine system: a review of the current information and misinformation. Infect Dis Rep 2022 Mar 11;14(2):184-197 [FREE Full text] [doi: 10.3390/idr14020023] [Medline: 35314653]
10. Berger Z, Altiery DE Jesus V, Assoumou SA, Greenhalgh T. Long COVID and health inequities: the role of primary care. Milbank Q 2021 Jun 30;99(2):519-541 [FREE Full text] [doi: 10.1111/1468-0009.12505] [Medline: 33783907]
11. Benatar SR. Health in a post-COVID-19 world. S Afr J Sci 2022 Nov 30;118(11/12). [doi: 10.17159/sajs.2022/14995]
12. de Miranda WD, Silveira F, Santos FP, Magalhães Junior HM, Paes-Sousa R. Challenges, consequences, and possible paths for confronting post-COVID-19 health inequalities and vulnerabilities. Saúde Debate 2022 Dec;46(spe8):141-155. [doi: 10.1590/0103-11042022e811]
13. Shanbehzadeh S, Tavahomi M, Zanjari N, Ebrahimi-Takamjani I, Amiri-Arimi S. Physical and mental health complications post-COVID-19: scoping review. J Psychosom Res 2021 Aug;147:110525 [FREE Full text] [doi: 10.1016/j.jpsychores.2021.110525] [Medline: 34051516]
14. Ren FF, Guo RJ. Public mental health in post-COVID-19 era. Psychiatr Danub 2020 Aug 12;32(2):251-255 [FREE Full text] [doi: 10.24869/psyd.2020.251] [Medline: 32796794]
15. Tosato M, Calvani R, Ciciarello F, Galluzzo V, Martone AM, Zazzara MB, et al. Malnutrition in COVID-19 survivors: prevalence and risk factors. Aging Clin Exp Res 2023 Oct;35(10):2257-2265 [FREE Full text] [doi: 10.1007/s40520-023-02526-4] [Medline: 37665556]
16. Grund S, Bauer JM. Malnutrition and sarcopenia in COVID-19 survivors. Clin Geriatr Med 2022 Aug;38(3):559-564 [FREE Full text] [doi: 10.1016/j.cger.2022.04.001] [Medline: 35868673]
17. Lewis M, Holland K, Govender E. The interdisciplinary communication dynamics of the COVID-19 pandemic. In: Lewis M, Govender E, Holland K, editors. Communicating COVID-19. Cham, Switzerland: Palgrave Macmillan; 2021.

XSL•FO
RenderX

18. Cox A, Iedema R, Li S, Sabbe M, Salaets H, Dauby N. Adding barriers to barriers during the COVID-19 pandemic: a call for interdisciplinary research on communication with migrant patients in the emergency department. Eur J Emerg Med 2021 Jan 01;28(1):5-6. [doi: 10.1097/MEJ.0000000000000777] [Medline: 33369953]
19. Pool J, Fatehi F, Akhlaghpour S. Infodemic, misinformation and disinformation in pandemics: scientific landscape and the road ahead for public health informatics research. Stud Health Technol Inform 2021 May 27;281:764-768. [doi: 10.3233/SHTI210278] [Medline: 34042681]
20. Bin Naeem S, Kamel Boulos MN. COVID-19 misinformation online and health literacy: a brief overview. Int J Environ Res Public Health 2021 Jul 30;18(15):8091 [FREE Full text] [doi: 10.3390/ijerph18158091] [Medline: 34360384]
21. Salvador-Carulla L, Rosenberg S, Mendoza J, Tabatabaei-Jafari H, Pandemic-Mental Health International Network. Rapid response to crisis: health system lessons from the active period of COVID-19. Health Policy Technol 2020 Dec;9(4):578-586 [FREE Full text] [doi: 10.1016/j.hlpt.2020.08.011] [Medline: 32874862]
22. Abbas M, Aloudat T, Bartolomei J, Carballo M, Durieux-Paillard S, Gabus L, et al. Migrant and refugee populations: a public health and policy perspective on a continuing global crisis. Antimicrob Resist Infect Control 2018 Sep 20;7(1):113 [FREE Full text] [doi: 10.1186/s13756-018-0403-4] [Medline: 30250735]
23. Ford-Paz RE, Santiago CD, Coyne CA, Rivera C, Guo S, Rusch D, et al. You're not alone: a public health response to immigrant/refugee distress in the current sociopolitical context. Psychol Serv 2020;17(S1):128-138. [doi: 10.1037/ser0000381] [Medline: 31464470]
24. Branham L. Overcoming recurring crisis through resilience: an analysis of USAID's definition of resilience. Clark University. 2019 May. URL: https://commons.clarku.edu/idce_masters_papers/236/ [accessed 2024-12-09]
25. Carlsen L, Bruggemann R. The 17 United Nations' sustainable development goals: a status by 2020. Int J Sustain Dev World Ecol 2021 Jul 30;29(3):219-229. [doi: 10.1080/13504509.2021.1948456]
26. Prabhu A. Informatics. In: Schintler LA, McNeely CL, editors. Encyclopedia of Big Data. Cham, Switzerland: Springer International Publishing; 2020.
27. Bath PA. Health informatics: current issues and challenges. J Inf Sci 2008 Jun 13;34(4):501-518. [doi: 10.1177/0165551508092267]
28. Mosadeghrad AM, Afshari M, Isfahani P, Ezzati F, Abbasi M, Farahani SA, et al. Strategies to strengthen the resilience of primary health care in the COVID-19 pandemic: a scoping review. BMC Health Serv Res 2024 Jul 25;24(1):841 [FREE Full text] [doi: 10.1186/s12913-024-11278-4] [Medline: 39054502]
29. Moran D. Introduction to Phenomenology. Milton Park, UK: Routledge; 2002.
30. Ryan GW, Bernard HR. Techniques to identify themes. Field Methods 2003 Feb 01;15(1):85-109. [doi: 10.1177/1525822x02239569]
31. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec 16;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
32. Garba I, Sterling R, Plevel R, Carson W, Cordova-Marks FM, Cummins J, et al. Indigenous peoples and research: self-determination in research governance. Front Res Metr Anal 2023 Nov 15;8:1272318 [FREE Full text] [doi: 10.3389/frma.2023.1272318] [Medline: 38033627]
33. Jennings L, Anderson T, Martinez A, Sterling R, Chavez DD, Garba I, et al. Applying the 'CARE Principles for Indigenous Data Governance' to ecology and biodiversity research. Nat Ecol Evol 2023 Oct 09;7(10):1547-1551. [doi: 10.1038/s41559-023-02161-2] [Medline: 37558804]
34. Virapongse A, Duerr R, Metcalf EC. Knowledge mobilization for community resilience: perspectives from data, informatics, and information science. Sustain Sci 2018 Aug 1;14(4):1161-1171. [doi: 10.1007/s11625-018-0612-z]
35. Sijm-Eeken ME, Arkenaar W, Jaspers MW, Peute LW. Medical informatics and climate change: a framework for modeling green healthcare solutions. J Am Med Inform Assoc 2022 Nov 14;29(12):2083-2088 [FREE Full text] [doi: 10.1093/jamia/ocac182] [Medline: 36219786]
36. Fernandez-Luque L, Kushniruk AW, Georgiou A, Basu A, Petersen C, Ronquillo C, et al. Evidence-based health informatics as the foundation for the COVID-19 response: a joint call for action. Methods Inf Med 2020 Dec 11;59(6):183-192 [FREE Full text] [doi: 10.1055/s-0041-1726414] [Medline: 33975375]
37. Basit MA, Lehmann CU, Medford RJ. Managing pandemics with health informatics: successes and challenges. Yearb Med Inform 2021 Aug 21;30(1):17-25 [FREE Full text] [doi: 10.1055/s-0041-1726478] [Medline: 33882594]
38. Smith MW, Ash JS, Sittig DF, Singh H. Resilient practices in maintaining safety of health information technologies. J Cogn Eng Decis Mak 2014 Sep 06;8(3):265-282 [FREE Full text] [doi: 10.1177/1555343414534242] [Medline: 25866492]
39. Gray SW, Ottesen RA, Currey M, Cristea M, Nikowitz J, Shehayeb S, et al. Leveraging an informatics approach to identify an unmet clinical need for BRCA1/2 testing among patients with ovarian cancer. JCO Clin Cancer Inform 2022 Sep;6. [doi: 10.1200/cci.22.00034]
40. Harwell LC, Vivian DN, McLaughlin MD, Hafner SF. Scientific data management in the age of big data: an approach supporting a resilience index development effort. Front Environ Sci 2019 Jun 04;7(Article 72):1-13 [FREE Full text] [doi: 10.3389/fenvs.2019.00072] [Medline: 33123540]

XSL•FO
RenderX

41. Shah SA, Seker DZ, Rathore MM, Hameed S, Ben Yahia S, Draheim D. Towards disaster resilient smart cities: can internet of things and big data analytics be the game changers? IEEE Access 2019;7:91885-91903. [doi: 10.1109/access.2019.2928233]
42. Smith SE, Drake LE, Harris JG, Watson K, Pohlner PG. Clinical informatics: a workforce priority for 21st century healthcare. Aust Health Rev 2011;35(2):130-135. [doi: 10.1071/ah10935]
43. McFarlane TD, Dixon BE, Grannis SJ, Gibson PJ. Public health informatics in local and state health agencies: an update from the public health workforce interests and needs survey. J Public Health Manag Pract 2019;25 Suppl 2, Public Health Workforce Interests and Needs Survey 2017(2 Suppl):S67-S77 [FREE Full text] [doi: 10.1097/PHH.0000000000000918] [Medline: 30720619]
44. ChatGPT homepage. ChatGPT. URL: https://chatgpt.com/ [accessed 2024-12-09]

Abbreviations

AI: artificial intelligence
COREQ: Consolidated Criteria for Reporting Qualitative Research
RI: resilience informatics

Edited by A Mavragani; submitted 13.06.24; peer-reviewed by M Khairulbahri; comments to author 03.09.24; revised version received 05.09.24; accepted 06.11.24; published 16.01.25.

Please cite as:
Block Ngaybe MG, Azurdia Sierra L, McNair A, Gonzalez M, Arora M, Ernst K, Noriega-Atala E, Iyengar MS
Resilience Informatics in Public Health: Qualitative Analysis of Conference Proceedings
JMIR Form Res 2025;9:e63217
URL: https://formative.jmir.org/2025/1/e63217
doi:10.2196/63217
PMID:

©Maiya G Block Ngaybe, Lidia Azurdia Sierra, Andrew McNair, Myla Gonzalez, Mona Arora, Kacey Ernst, Enrique Noriega-Atala, M Sriram Iyengar. Originally published in JMIR Formative Research (https://formative.jmir.org), 16.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Assessment, Decision, Adaptation, Production, Topical Experts-Integration, Training, and Testing (ADAPT-ITT) Framework to Tailor Evidence-Based Posttraumatic Stress Disorder Treatment for People With HIV to Enhance Engagement and Adherence: Qualitative Results from a Feasibility Randomized Controlled Trial

Cristina M Lopez[1], PhD; Angela D Moreland[1], PhD; Stephanie Amaya[1], PhD; Erin Bisca[1], MA, LPC; Christin Mujica[2], MA; Tayler Wilson[1], MA; Nathaniel Baker[3], MS; Lauren Richey[4], MD; Allison Ross Eckard[5], MD; Patricia A Resick[6], PhD; Steven A Safren[7], PhD; Carla Kmett Danielson[1], PhD

[1]Department of Psychiatry & Behavioral Sciences, Medical University of South Carolina, Charleston, SC, United States
[2]Department of Psychological Science, University of Arkansas, Fayetteville, AR, United States
[3]Department of Public Health Sciences, Medical University of South Carolina, Charleston, SC, United States
[4]Department of Medicine, Infectious Disease Section, Louisiana State University Health Sciences Center, New Orleans, LA, United States
[5]Department of Infectious Diseases, Medical University of South Carolina, Charleston, SC, United States
[6]Department of Psychiatry and Behavioral Sciences, Duke University Medical Center, Durham, AL, United States
[7]Department of Psychology, University of Miami, Coral Gables, FL, United States

Corresponding Author:
Cristina M Lopez, PhD
Department of Psychiatry & Behavioral Sciences
Medical University of South Carolina
67 President St
Charleston, SC, 29425-5712
United States
Phone: 1 843 792 2945
Email: lopezcm@musc.edu

Abstract

Background: Individuals with co-occurring posttraumatic stress disorder (PTSD) and HIV are at high-risk for negative HIV-related outcomes, including low adherence to antiretroviral therapy, faster disease progression, more hospitalizations, and almost twice the rate of death. Despite high rates of PTSD in persons with HIV (PWH) and poor HIV-related health outcomes associated with PTSD, an effective evidence-based treatment for PTSD symptoms in PWH does not exist.

Objective: This study aimed to describe the adaptation and theater testing of an evidence-based intervention designed for people with co-occurring PTSD and HIV.

Methods: The Assessment, Decision, Adaptation, Production, Topical experts-integration, Training, and Testing (ADAPT-ITT) framework guided the formative process used to modify an evidence-based PTSD treatment (cognitive processing therapy; CPT) to meet the unique needs of PWH experiencing PTSD. With the integration of Life-Steps for Medication Adherence (Life-Steps), the adapted protocol (CPT-Life-Steps for integration of adherence; CPT-L) targeted HIV-related stigma and HIV medication adherence within a trauma-informed framework. Theater testing was completed with 7 participants to evaluate acceptability of CPT-L for PWH. The qualitative data (N=54 recordings) used to evaluate and adapt CPT-L emerged from individual interviews conducted with participants after each therapy session as well as exit interviews conducted at posttreatment data collection.

Results: After challenging stigma-related appraisals, participants expressed feeling less constrained by maladaptive thoughts. These shifts translated to increased self-efficacy with both HIV-related care and mental health.

Conclusions: These results indicate that trauma-informed work with PWH should consider the impact of HIV on trauma-related stuck points, intersecting identities (including living with HIV), and challenging internalized stigma. Findings provide evidence

XSL•FO
RenderX

that CPT-L is acceptable and effective in addressing internalized HIV stigma that impacts PTSD symptom maintenance and HIV treatment engagement.

Trial Registration: ClinicalTrials.gov; NCT05275842; https://clinicaltrials.gov/study/NCT05275842?id=NCT05275842&rank=1

International Registered Report Identifier (IRRID): RR2-10.1016/j.conctc.2023.101150

(JMIR Form Res 2025;9:e64258) doi:10.2196/64258

KEYWORDS

PTSD; HIV; adherence; minoritized populations; adaptation; evidence-based; treatment; engagement; posttraumatic stress disorder; stress; trauma; antiretroviral therapy; therapy; symptoms; acceptability; self-efficacy

Introduction

Background

By 2019, an estimated 1,189,700 people in the United States were living with HIV [1]. Although people with HIV (PWH) face many unique psychosocial risk factors that impact health outcomes, trauma exposure and posttraumatic stress disorder (PTSD) have been consistently identified among the most prominent stressors for this population. The prevalence of PTSD among PWH, 30%-74% [2,3] far exceeds that of the general population, 3.9% [4]. Trauma history endorsed for *DSM-5-TR* (*Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, Text Revision*) criterion A, diagnostic information across the lifespan can include a wide range of adverse childhood experiences, interpersonal violence, survivors of natural disasters, combat, immigration-related trauma, HIV-related trauma if identified as traumatic by patient, and other events endorsed as inducing fear of serious harm to self or a loved one. Data suggest that traumatic events most often reported by PWH include interpersonal violence (physical, sexual, or psychological in nature), witnessing domestic violence or physical abuse as children, and seeing someone be seriously injured or killed [5]. This high comorbidity has important treatment implications as patients with co-occurring PTSD and chronic medical conditions report forgetting (41%) or skipping (24%) medications at significantly higher rates than patients without PTSD (29% and 13%, respectively) [6]. In addition, cognitive impairment and increased avoidant behavior experienced as a function of PTSD may contribute to lower levels of antiretroviral therapy (ART) adherence for PWH with a comorbid PTSD diagnosis [7]. For instance, PWH reported avoiding their ART because it was a reminder of their status [7] and of a traumatic life event [8], and they reported avoidance of clinic care visits due to fear of prejudice and discrimination from health care providers [9].

Notably, advances in ART have led to significant improvements in disease-related morbidity, mortality, and health outcomes for PWH. With consistent use, ART markedly slows disease progression and improves quality of life by suppressing viral load, which can potentially be reduced to undetectable levels. However, failure to reach near perfect (≥95%) adherence rates may be insufficient and detrimental to achieving viral suppression. More specifically, suboptimal ART adherence is strongly linked to viral proliferation [10,11], medication resistance [10,12], disease progression [13], and even death [14]. Unfortunately, the necessary 95% to 100% ART adherence rates are not easily achieved considering the multifaceted barriers that PWH must overcome to reach and maintain the recommended benchmark.

The stigmas attached to HIV, mental health, and minoritized identities are one such barrier. Stigmas can impact PWH through three primary mechanisms: (1) enacted stigma experienced in the form of actual rejection or discrimination by people and society; (2) anticipated stigma, which is the expectancy of discrimination and marginalization; and (3) internalized stigma or negative self-perceptions related to having the disease [15]. In addition, intersectional stigma [16,17], has been shown to differentially impact PWH, such that individuals with additional minoritized identities (eg, women, individuals that identify as sexual and gender minorities, and minoritized racial groups) report higher rates of stigmatization [18], and medication nonadherence [19], compared with those with majority-based identities. Furthermore, stigmatization has also been associated with multiple social determinants of health, such as low education level, poverty, lack of access to health insurance, incarceration, and non-US country of origin [18]. In turn, stigmatization can manifest in significant health disparities including poorer health outcomes, elevated PTSD symptoms due to reinforcement of avoidant coping behaviors, and delayed enrollment in HIV care until significantly ill [18-22].

Theoretical models jointly attribute the observed health disparities to the presence of “minority stress” and the absence of “social safety” among PWH, especially those who experience intersectional stigma. Specifically, the minority stress model [23] proposes that heightened vulnerability to health problems results from the cumulative burden of psychological stress experienced in response to social stigmatization (ie, minority stress). In effect, chronic exposure to stigma may lead individuals to be more susceptible to the development of posttraumatic distortions about the world (eg, enacted stigma) and oneself (eg, internalized stigma) following a traumatic event [15,23]. In addition, stigma may reinforce avoidant coping behaviors that are commonly associated with PTSD (eg, lack of trust in new relationships like with a new provider; not taking medication because it reminds them of a negative thought or feeling, etc). In other words, minority stress can impact mental health by reinforcing maladaptive beliefs associated with PTSD and emotional disorders after traumatic or difficult experiences [24]. These findings emphasize the importance of addressing stigma within evidence-based psychosocial interventions as a means of reducing distress and paving the way for improvements in HIV medication adherence and health outcomes. Expanding on the minority stress model, social safety theory posits that, in the absence of social safety (ie, reliable social connection and

XSL•FO
RenderX

protection), stigmatized PWH routinely navigate the world in a state of maintained vigilance that adversely impacts immune system response [25]. Threats in the social environment are thought to trigger physiological stress responses, including chronic inflammation, that can impair immunological function [25]. Thus, social safety theory suggests that targeting maladaptive beliefs (eg, internalized stigma) and interpersonal patterns (eg, avoidance) may amplify social safety, and improve health outcomes by extension, for trauma exposed PWH [25].

Current Study

Due to the distinct connection between HIV and PTSD, it is important that evidence-based treatments be tailored to address both trauma symptoms and barriers to HIV care for PWH. Although effective interventions to address HIV medication adherence and PTSD are widely available, there are no established integrated approaches that address the dual issues of cooccurring HIV with PTSD and related health disparities. In addition, no research to date has examined the critical question of whether HIV outcomes can be improved among the large number of PWH with comorbid PTSD by treating PTSD symptoms to maximize benefits of health promotion interventions. To address this tremendous void in the field, we propose the integration and evaluation of 2 empirically supported interventions that have been successfully used to improve PTSD- and HIV-related outcomes in this population, namely cognitive processing therapy (CPT) [26], and Life-Steps for Medication Adherence (Life-Steps) [27].

Using the Assessment, Decision, Adaptation, Production, Topical experts-integration, Training, and Testing (ADAPT-ITT) model, a framework for adapting HIV-related evidence-based interventions [28], CPT and Life-Steps were integrated to encompass HIV-related stigma and ART adherence within the scope of PTSD treatment. Upon adapting the intervention for PWH and PTSD, the resulting model (CPT-L) was theater-tested with treatment-seeking members of the intended population. More specifically, maladaptive beliefs influenced by exposure to trauma can interfere with important correlates of health, including a participant's motivation to remain ART adherent [29], access social support [30], and establish trusting relationships with medical care providers [31]. To tackle these psychosocial factors, CPT-L includes cognitive restructuring components introduced weekly through modified worksheets and handouts reflecting topics that are more relevant to the realities of living with co-occurring HIV and PTSD. Acknowledging that these health conditions are experienced within a broader personal and cultural context, CPT-L also placed emphasis on engaging in discussion and exploration of the participants' intersecting identities (eg, HIV status, racial, ethnic, sexual, and gender identity discrimination) [32]. This study presents qualitative findings of feedback solicited to assess reactions to the new intervention and elicit critiques that will guide future efforts to refine and tailor CPT-L.

Methods

Treatment Adaptation

Aligned with the ADAPT-ITT model, the CPT-L adaptation process will comprise 8 phases: assessment, decision, adaptation, production, topical expert, integration, training, and testing; and it has primarily been used in the context of HIV [33]. Findings in the current manuscript reflect the first 4 completed phases of the ADAPT-ITT protocol. During the initial step (assessment) of the ADAPT-ITT framework, the team consulted with the Community Advisory Board (CAB), comprising of constituents with lived experience (with PTSD or HIV), Ryan White clinic personnel (eg, providers, case managers, and staff), peer advocates, and community-based mental health providers. CAB members and topical experts were provided with a review of the extant literature to assess the scope of the problem and whether it merited development of new evidence-based treatments to address comorbid PTSD and HIV. Consideration was given to studies on mental health disparities in PWH as well as conflicting evidence about the efficacy of existing treatment modalities for HIV and co-occurring mood disorders. Evidence suggests that although existing psychosocial interventions can be effective in treating depression among PWH, symptom improvement does not translate consistently to improvements in HIV-specific markers (eg, ART adherence and viral load) [34]. Therefore, the team heeded CAB feedback and determined that adaptation was necessary to attend to the unique and unmet needs of this population.

In the second step (decision), the CAB and team's selection of CPT among the available evidence-based PTSD interventions was driven by its flexibility to integrate HIV-specific content and CPT's established training model for successful dissemination [35,36]. CPT, is an evidence-based treatment shown to be effective for reducing posttraumatic distress, including PTSD and depressive symptoms, among civilian and military adult populations with complex histories of exposure to traumatic events [37,38]. The protocol is designed for 12 sessions, but the length of treatment can be adapted based upon the patient's needs [39]. In terms of structure, after centering the first half of treatment on dysfunctional cognitions about the worst traumatic event, the focus shifts to maladaptive beliefs that highlight the following 5 areas commonly impacted by trauma: safety, trust, power and control, esteem, and intimacy. CPT has been shown to reduce trauma-related self-blame and guilt [40,41] and may be particularly beneficial for PWH because of the focus on "stuck points." Stuck points are negative, inaccurate, and unhelpful beliefs about yourself, others, or the world that occur as a result of trauma. Maintaining these stuck points prevents the individual from recovering from PTSD. The goal of CPT is to assist the patient in identifying stuck points related to their traumatic event, and then challenge these stuck points or negative thinking patterns through cognitive restructuring [42]. The process of cognitive restructuring is used to identify alternative, more accurate, and more helpful thoughts through Socratic questioning (ie, open-ended questions from the therapist that stimulates the patient to question current beliefs as flawed). In addition, the broad scope covered under the 5 themes allows for the flexibility to address internalized stigma related to HIV and marginalized identities. Indeed, although it was not originally developed as an HIV intervention, CPT has shown potential in promoting HIV risk-reduction [43].

Adaptation focused on modifying CPT to bolster its effectiveness with HIV medication adherence and related health

XSL•FO
RenderX

outcomes. The CAB's decision to incorporate Life-Steps for medication adherence [27], into CPT was made based on studies showing its efficacy and successful integration with various mental health treatment modalities [27,41]. Life-Steps is a one-session intervention using cognitive behavioral strategies, problem-solving, and motivational interviewing to promote adherence. The modified intervention (CPT-L) used the single-session adjunct to establish baseline HIV and ART psychoeducation and problem-solving skills. CPT-L later leveraged this foundational knowledge by incorporating continuous reminders of the benefits of remaining ART adherent and restructuring maladaptive thoughts that posed barriers to HIV care. The integration of the Life-Steps component also addressed health literacy and concerns about ART through tailored explanations about the need for consistent adherence, psychoeducation on the co-occurrence of PTSD and HIV, and rehearsal of adaptive health behaviors (eg, recognizing barriers to adherence and using problem-solving strategies to confront them). In the context of the larger study aims, CAB members chose Life-Steps as the control condition to ensure that participants assigned to the study arm would receive benefit as well.

The third step (adaptation) was completed through a larger pilot trial. The protocol was theater-tested with members of the target population, namely PWH and co-occurring PTSD. Aligned with the ADAPT-ITT model, reactions and feedback about the intervention were elicited from participants and stakeholders and include the qualitative findings presented here. Currently, CPT-L has a drafted manual (stage 4 of the ADAPT-ITT model) and plans to continue the adaptation process with topical expert feedback are underway. Refer to Table 1 for CPT-L protocol modifications based on ADAPT-ITT steps 1-4 that have been completed. Next stages of the project include an iterative process of refining the manual by eliciting comments and suggestions from our CAB, clinical team, peer advocates and other key constituents (step 5–Topical expert feedback), integration of that feedback in the CPT-L manual (step 6–Integration), and training of personnel to deliver the adapted intervention (step 7).

Table 1. Components and modifications for a tailored, evidence-based PTSD treatment for people with HIV (cognitive processing therapy- Lifesteps): adaptations informed by a qualitative study using the empirically supported ADAPT-ITT framework.

Sessions	Intervention component	CPT-L[a] refinements
1	Life-Steps; begin overview of PTSD[b] and CPT[c]	Psychoeducation about connection between HIV status, stigma, and trauma; how held identities impact trauma and HIV experiences; assign additional impact statement assigned if trauma and receiving news about HIV status are separate events
2-3	Finding maladaptive HIV beliefs	Identify beliefs related to internalized stigma through HIV-related impact statement; use HIV-related examples in context of ABC worksheets
4-5	Processing the index event	Distinguish trauma event processing from receiving news about HIV status; processing stigma from learning of HIV status; use HIV-related examples in context of challenging questions worksheets
6-7	Learning to self-challenge	Use HIV-related examples in context of challenging beliefs worksheets
8-10	Trauma themes (safety, trust, and power-control)	Regain "control" of health by challenging stigma-related stuck points; adaptation of Trust Star Worksheet to include status disclosure
11-12, 12+	Esteem, intimacy, and facing the future	Challenging stigma-related non-adaptive cognitions, duty to warn review for clinician
Delivery adaptations		
Format change	Variable length	Could use additional "stressor sessions" to discuss stigma-related stuck points[39] in CPT[c] variable length format
All	Adherence self-monitoring	In addition to risk assessment in CPT[c], CPT-L[a] includes ART[d] adherence monitoring
All	Trauma, stigma, and HIV connection	Continually developing the connection between HIV status and trauma; discuss impacts of intersectionality on HIV and trauma experience, as appropriate
All	Practice assignments	Therapist to encourage challenging both trauma-related and HIV-related stuck points between sessions (if separate events)

[a]CPT-L: CPT-Life-Steps for integration of adherence.

[b]PTSD: post-traumatic stress disorder.

[c]CPT: cognitive processing therapy.

[d]ART: antiretroviral therapy.

Often, when adaptations of evidence-based treatments are attempted, the process generally ends there. However, a critical element of the ADAPT-ITT model includes conducting a test of the full intervention (step 8). While the involvement of stakeholders across multiple phases of the adaptation processes helps ensure that the adaptation resonates with the community and reflects their needs, the last and final phase of the ADAPT-ITT model determines that the adapted intervention

XSL•FO
RenderX

has met the goal of creating positive and incremental impact for a specific population. Thus, subsequent steps of the adaptation process for CPT-L include a larger trial to evaluate the final product in a rigorous and systematic way. The submitted project proposal for the large-scale testing of CPT-L also includes an aim focused on implementation science and CPT-L intervention scale up, an important aspect to the team to remain aligned with the feedback from the CAB members.

Participants and Recruitment

Participants were recruited from the Infectious Disease Department's Ryan White Clinic at the academic medical center. A full description of the study procedures is reported elsewhere [44]. All protocols were approved by the institutional review board at the affiliated university. The study's program coordinator recruited participants during their regularly scheduled HIV care clinic appointments and used a "cold call" list to contact individuals with HIV outside of these appointments. The study's purpose, methods, and compensation were discussed with participants, and if they were interested in participating, the initial contact assessment was completed. This initial assessment included completing a shortened version of the PTSD checklist for *DSM-V* (*Diagnostic and Statistical Manual of Mental Disorders* [Fifth Edition]; PCL-5), as well as the Montreal Cognitive Assessment (MoCA). If participants scored at least 3 out of 5 on the shortened PCL-5, indicating possible PTSD, and scored within the mild or moderate range on the MoCA assessment (ie, no indication of severe cognitive impairments), then the consent process was completed. Consented participants then completed the clinician administered PTSD scale for DSM-5 to determine if they met criteria for PTSD. Participants who met criteria for PTSD were able to continue through the study. Additional eligibility criteria included (1) age 18 years or older enrolled in a Ryan White clinic; (2) able to speak, read, and write English; (3) meet at least one of the following HIV care criteria: (a) diagnosed with HIV in last 3 months; (b) detectable viral load in last 12 months; (c) failed to show up for or missed 1 or more HIV care appointments in the past 12 months; (d) last HIV care visit was more than 6 months ago; and (e) self-reporting less than 90% ART adherence in the past 4 weeks.

CPT-L was completed individually with a licensed professional counselor and was offered virtually and in person, depending on the participant's location and preference. All participants of the theater testing phase were provided with the standard CPT workbook, as well as modified worksheets specific to CPT-L (eg, Stuck Point log with examples of HIV related stuck points, modified ABC, Challenging Question, Challenging Beliefs worksheets with examples using HIV related stuck points, and a Trust Star geared toward aspects of HIV status disclosure).

As part of a larger pilot trial using theater testing for the first stage, 7 patients completed end of session interviews at the end of each CPT therapy session and exit interviews at the end of protocol (54 total session recordings; number of sessions per patient ranged from 1 to 12; M=7). Race and ethnicity included African American (n=5) and non-Hispanic White (n=2). Participants' ages ranged from 23-64 years. There were 4 males and 3 females, and participants identified as gay (n=2), bisexual (n=2), or heterosexual (n=3). Index traumas included life threatening illness or injury (n=1), physical assault (n=1), transportation accident (n=1), sexual assault or rape (n=2), and other unwanted sexual experiences (n=2).

Procedures

Patients participated in end of the session adaptation questions following each session and intervention exit interviews that explored strengths and weaknesses of CPT-L, including feedback on adaptations of CPT to people with HIV and suggestions for improvement in the adaptation before the large scale randomized controlled trial. The end of session adaptation form consisted of 2 questions that asked for participants to provide a rating from 1 to 10 (1=not at all and 10=very much): "How much did the materials covered in today's session address aspects of people living with HIV?" and "How much did the handouts reflect your experiences living with HIV and PTSD?" Participants were also asked to provide open-ended feedback on these questions and additional information regarding how useful the materials were and how they will use the skills taught in therapy in their everyday lives. In addition, exit interviews were conducted at the end of treatment. Exit interviews consisted of 6 primary questions and 13 follow-on questions (Table 2) that took place in 2022, in English, over videoconference and lasted approximately 5-15 minutes. Participants were compensated US $25 for completion of each exit interview.

XSL•FO
RenderX

Table 2. Exit interview questions completed by patients with HIV and comorbid PTSD at the end of a qualitative study exploring acceptability and relevance of a tailored, evidence-based PTSD treatment for people with HIV (cognitive processing therapy-Lifesteps).

Primary question	Follow-up question
Tell me about the mental health treatment or counseling you have received before	• What did you like? • What did you dislike? • What was missing? What could have been more helpful? • How did you feel about the overall experience?
Tell me your thoughts about the 12-session trauma and adherence-based intervention that you just completed over the past 6-8 weeks	• Was the length of treatment problematic? • Was it beneficial to discuss the topic of your trauma or stressful life experience? • Was the intervention beneficial to discuss and process thoughts about your HIV status? • What reservations do you have about recommending this treatment to be available clinically? • What changes, if any, do you think would need to be made to trauma treatment to make it most helpful, relevant and appropriate for PWHa? • Do you think that the same trauma treatment that is used with other groups would be equally helpful for PWH[a]? Why or Why not?
Did you feel that the CPT-L[b] intervention specifically addressed HIV-related topics?	• How do you feel that your experience of the trauma is different or unique than a survivor who is not living with HIV?
What coping behaviors for reducing your traumatic stress symptoms did you learn from participating in CPT-L[b]?	• Are these coping strategies different than the ones you used before CPT-Lb? How so?
What additional insights did you gain about the identities you hold?	• Did the examples and discussion included in the intervention resonate with the different identities you hold?
What concerns do you have if a friend were asked to participate in this study?	__[c]

[a]PWH: persons with HIV.

[b]CPT-L: CPT-Life-Steps for integration of adherence.

[c]Not applicable.

Data Analysis

All qualitative recordings (end of session adaptation questions and exit interviews) were transcribed verbatim. Following transcription, content from the qualitative recordings were analyzed using a systematic thematic analysis within a grounded theory approach [45,46]. A “ground-up” approach involves an iterative process of establishing codes, grouping the codes into key concepts, and organizing the concepts into broader themes. NVivo (version 12) software was used to organize data and conduct the qualitative analyses. This software package facilitates coding, storage, retrieval, and analysis of large amounts of transcribed data. The codebook was created in a collaborative, inductive, and iterative process. All qualitative recordings were coded by a primary coder (second author) and 25% of recordings were double coded for interrater reliability. Both coders have expertise in qualitative coding of mental health research; neither coder had any interaction nor relationships with participants. Meetings between coders were held to reach a consensus on inconsistent codes for each transcript.

The qualitative researchers were careful to ensure data saturation in the current study given that only 7 participants were included. In qualitative research, there are 5 components that influence data saturation which include predetermined codes and themes, sample size, relevancy of research subjects, number of research tools, and length of data collection sessions [47]. When analyzing the responses from the 7 participants, the qualitative experts determined that data saturation was reached based upon the themes directly mapping onto the questions asked in the interview, previous research showing that saturation can be reached in approximately 10 interviews, that all participants in this study were patients receiving CPT-L and thus, were the most relevant to answer the questions about treatment, and the length of data collection certainly increased the chance of early saturation in this study given that 54 interviews were collected over the data collection period.

Ethical Considerations

The study was approved by the Medical University of South Carolina Institutional Review Board for Human Research (Pro00106801). Before engaging in the study and collecting data, each participant was required to give written informed consent, either in person or virtually using the Medical University of South Carolina’s institutional review board approved REDCap (Research Electronic Data Capture) e-Consent. Within the consent process, research staff discussed with potential participants that the results of the research study, including interviews, may be published and information would not be identifiable. The procedures also clarified that all information obtained from participants would be confidentially maintained. Research staff received training from the principal investigator on the importance of confidentiality and had weekly

XSL•FO
RenderX

supervision meetings with the principal investigator ensuring strict compliance with the Human Subjects Protection plan. Additional measures for protecting privacy and confidentiality also included assigning each participant a unique project identification number. The codes that linked the name of the participant and the study ID number were kept confidential by the principal investigator in a secured, locked cabinet. The secure servers that housed the data files collected from the web applications (REDCap) during were located at MUSC in Secure Sockets Layer 128-bit encrypted servers behind firewalls. Participant confidentiality was masked in all data files by the use of project identification numbers rather than personal information.

After all the relevant information was reviewed verbally with trained research staff, participants signed the document (either virtually or in person) and the signed consent document was stored electronically (from e-consent or scanned in) in the REDCap database, which provides password protected, secure, web-based collection and management of research. Informed consent was conducted by members of the study team that received training through the University of Miami CITI training in both the Conduct of Human Subject Research and Good Clinical Practice, and were institutional review board–approved study personnel. Study data collected, including exit interviews, was conducted either in-person in the privacy of a clinic room, or remotely through telephone. Participants were informed when the interview recording began and ended and were instructed not to use the names of individuals or reveal any personally identifiable information. These audio files were deidentified within 24 hours and uploaded into the REDCap electronic study database to later be transcribed. Any files on a portable recording device were deleted following this upload, and only the project therapist, fidelity raters, and qualitative coders heard the content of the audio recordings. Participants were not compensated for interviews conducted at the end of each CPT-L session, but were compensated US $25 for end of study exit interviews.

Results

Descriptive Characteristics

Regarding the materials addressing aspects of people living with HIV, participants rated the materials very highly (mean 9.86, SD 0.93) and indicated that the handouts greatly reflected the experiences of living with HIV and PTSD (mean 9.71, SD 0.61).

Qualitative Results

Four overarching themes, each with their own subthemes, emerged from the semistructured interviews: (1) overall strengths of CPT-L, (2) overall suggested improvements of CPT-L, (3) helpful handouts/activities/materials, and (4) ways that patient would like to incorporate skills learned from CPT-L.

Theme 1: Overall Strengths of CPT-L

Most patients discussed overall strengths of CPT-L, with the main subthemes emerged including that patients generally felt materials were relatable and relevant to HIV, information was relevant to HIV, learned tools unique and specific to trauma related to HIV, discussed topics never addressed before, reported general improvements in well-being and hope, and learned a different (less negative) view of HIV status following CPT-L engagement. In discussing that they generally felt materials were relatable and relevant to HIV, one patient reported, “It’s making perfect sense and I think it’s really helpful,” and another patient noted, “[CPT-L] is very sensible. It’s pragmatic, it’s logical. I wish I had this a long time ago.” Several patients discussed that the model was very relevant to HIV, making statements such as, “everything was covered. HIV, on what living with HIV is,” and “Oh yes, those would be concerns that people living with HIV would have.”

All patients discussed at some point in the exit interviews that they learned tools to address relevant issues. For example, 1 patient stated,

> *I felt increasingly as I participated that it was giving me some tools to live more – well, to be more whole and not feel as constrained by like stuck points in a way.*

Another patient discussed,

> *I wish I’d had these tools a very long time ago ‘cause I could have probably worked through some things more successfully.*

Relatedly, several patients reported that they discussed topics in CPT-L that they have never discussed before, including

> *It’s honestly the first time since that sexual abuse that I even tried to tackle this,” and “PTSD as far as a self-recognized condition is kind of new to me. So there’s still kind of a process of exploration about that.*

In describing that they now have a different view of their HIV status following engagement in CPT-L, 1 patient stated,

> *I don’t see HIV as a threat to me as much as I once did.*

Another patient described,

> *Knowing that it’s not my fault. That you know, I’m HIV and still can be successful.*

Finally, most of the patients discussed that they felt general improvements in well-being and hope following CPT-L. One patient described,

> *I honestly feel a little more hopeful. I feel better about my decision about how to deal with this feeling of hopelessness. And I think that I’m finding that I put more reasonable weight on events than I used to and some of it is coming from this where I’m just not looking. I feel like I’ve got a prism if you will look at things and not necessarily have them so heavily weighted one way or the other.*

Other patients discussed improvements in well-being and hope by stating,

> *I feel a little stronger about being able to take care of myself. Not just as far was physiologically with procedures but with my state of mind*

and

XSL•FO
RenderX

> *I think [CPT-L] is increasing, you know, kind of my recognition of my place in life. And that I can improve things. You know? And you know, not feel so crestfallen about my place in the world.*

Theme 2: Suggested Improvements in CPT-L

About half of the patients discussed suggested improvements in CPT-L, with the main subthemes including adding more HIV-relevant examples, to individualize the activities, and to clarify some of the handouts. Specifically, some patients who suggested adding more HIV relevant examples stated, "Is there a way to make the trust worksheets more reflective of HIV," "There were like only 3 choices and they weren't really talking about HIV," and "The handouts might be better with advocate stories from people with HIV." One patient who thought the activities could be more individualized mentioned, "I think that for me though because of my generation so to speak and some of the things I've seen while I've been HIV positive as well as had my own sexual orientation that those things might be, – they don't touch on me so much." Another patient stated, "I just think that some people will respond well to this method and other's less so." Finally, some patients discussed ways to clarify some of the handouts with statements such as, "Some of them were not self-explanatory altogether," and "As far as like the terminology that maybe there's something to you know actually define it."

Theme 3: Helpful Handouts, Activities, and Materials

All 7 of the patients gave feedback on their perceived helpful handouts, activities, and materials during completion of CPT-L, with the main subthemes including handouts, challenging stuck points, and exposure activities. All patients made positive comments about the handouts, including general comments about the handouts such as

> *The handouts made me reflect on the things that I asked me to do and things that I need to let go of,*
>
> *The handouts reflected someone living with both HIV and PTSD. They reflected a lot,*

and

> *I think that they could equally be applied to a lot of other difficult circumstances people have. And I think with certain individuals that it might be more helpful for them to have something that's tangible.*

Several patients also made positive comments about specific handouts, including

> *I actually like this trust star so to speak. I think that it could actually be pretty useful just as an exercise about trust generally rather than specifically.*

Almost all of the patients discussed that the challenging stuck points activities were helpful. Specifically, 1 patient noted,

> *To reflect on, you know, an issue and look at it from a lot of different facets in a lot of different ways,*

and another mentioned,

> *It works. I mean it's asking me directly to address the [stuck points] and to put things down and look at them and go through a systematic process. You don't like activating them but it's working.*

In addition, most patients reported that the exposure techniques were very helpful in decreasing symptoms. One patient described,

> *Just talking about it. I mean, you let me trigger some things that I didn't even think I was gonna think about, yeah?*

Theme 4: Incorporating Skills Learned in CPT-L

Many of the patients described ways that they plan to incorporate skills learned in CPT-L moving forward, including setting different goals, engaging in different activities, practicing challenging thoughts and stuck points, being more aware of feelings, and improving self-acceptance and self-esteem. Regarding goals, many patients described that they plan to set different goals for themselves. One patient reported,

> *I wanna set different goals for myself and maybe think about how I'm doing things a little bit differently and as far as you know, making a more dive into this state.*

Another patient stated some personal goals,

> *Not being in my PJs and taking a shower before session. And combing my hair.*

In addition, a patient discussed a goal of consistently taking medication by stating,

> *I need to make sure I'm on my meds for the next three months to get shots and I hate shots, so this has helped me.*

Relatedly, several patients described that they plan to engage in different activities following completion of CPT-L, including

> *I can start planning something. I will start with people from church" and "If I have to do at least one thing just for myself, not because I did anything to deserve it, or whatever right? I can do that.*

Most patients described that they plan to practice changing thoughts and challenging stuck points following completion of CPT-L. Specifically, 1 patient reported,

> *I think that there's a great likelihood that I'm probably going to filter what I think and how I react,*

while another patient stated,

> *Reconstruct my perception or think about alternatives.*

A few patients reported that they plan to be more aware of their feelings, as 1 patient stated,

> *Assessing feelings, and probably being more aware of them. And their impact on people around me.*

Finally, several of the patients described that they would like to improve their self-acceptance and self-esteem following the completion of CPT-L. Specifically, 1 patient made a comment about self-acceptance and self-esteem including,

> *The way I think about what happened with my mother and trying to use it as a lens to help me have a better sense of self instead of looking for some reason why I'm the way I am. Just more acceptance of the way I*

XSL•FO
RenderX

> *am. Because I mean, I don't feel like I'm altogether horrible.*

Other patients described,

> *Try to think more positive about myself,*

and

> *I can already see this has given me a construct with some new things that float into my mind. That are kind of self-judgments that I can, you know, have a better way of dealing with.*

Discussion

Recent years have seen a growing awareness of the need for trauma-informed, HIV treatment approaches as converging evidence suggests a reciprocal association between HIV and exposure to trauma. Having an HIV diagnosis is a significant predictor of physical and sexual victimization [48-49], which in turn has been associated with more engagement in HIV-risk behaviors, poorer physical functioning, increased symptomology, and higher rates of mortality among PWH [12,50-54]. Unfortunately, PWH with co-occurring PTSD must overcome several structural, psychosocial, and behavioral challenges with potentially devastating health consequences. These mutually reinforcing risk factors call for an integrated treatment approach [55]. Indeed, integrated interventions for cooccurring disorders have been found to increase client engagement with treatment and improve patient outcomes [56]. To address this unmet need, the current study introduced a modified treatment protocol incorporating gold standard, evidence-based practices used with trauma exposed populations CPT and PWH (Life-Steps) [27,57]. CPT was selected from similarly effective treatments for PTSD for its mechanism of change (ie, cognitive restructuring), which allows for the flexibility to simultaneously target maladaptive appraisals related to trauma, intersectional stigma, and HIV using a skills-based, modular format.

The resulting framework (CPT-L) leverages treatment components from each protocol to target factors known to contribute to PTSD and ART nonadherence. First, structural barriers, such as low access to (and understanding of) health education, can considerably impede adherence given the complexity of HIV medication regimens [58]; components of CPT-L addressing these documented barriers were well-received by participants in the current study. The integration of Life-steps with CPT provided explanations about consistent adherence and psychoeducation about the co-occurrence of PTSD and HIV. Generally, participants found that the information adapted from both protocols was well integrated and accurately reflected their lived experience. Findings suggest that demystifying co-occurring HIV and PTSD allowed participants to gain a sense of mastery over their own medical care, which then translated to actionable goals to increase and maintain higher levels of ART adherence. The observed patterns were consistent with studies supporting the effectiveness of education-based interventions on treatment adherence for PWH [59,60] but extended existing interventions by incorporating information on the impacts of PTSD. Although some participants lamented not having access to the information sooner, they also noted that it has contributed to a newfound self-awareness and sense of hope. Within the context of the roles of stigma and attitudinal barriers in delayed mental health services, the high patient acceptability of CPT-L are promising and warrant further testing to help improve access to trauma-focused care, both in the United States and abroad [61].

Another known factor contributing to PTSD and ART nonadherence that is addressed by CPT-L are maladaptive beliefs in a multicultural context. In addition to "traditional" (ie, more trauma-focused) maladaptive beliefs that are identified for cognitive restructuring in CPT, increased acceptability and relevance of CPT-L by PWH may be due to CPT-L's emphasis on participants' intersecting identities (eg, HIV status, racial, ethnic, sexual, and gender identity discrimination) to acknowledge and validate broader personal and cultural contexts that can affect health behaviors and management. This multicultural orientation opened a conversation about experiences of stigma (enacted and internalized) they may have experienced. Participants were then guided in processing intersectional stigma through a trauma-informed lens (ie, themes of safety, trust, power and control, esteem, and intimacy). For example, aligned with evidence of the association between disclosure of HIV status and increased social support [43], the "Trust Star Worksheet" was adapted to reflect disclosure as an important type of trust. The Trust Star is an activity completed within CPT to assist the patient in identifying aspects and levels of trust, and how this impacts the way the individual may trust themselves and others. While many individuals with PTSD view trust as an all or none concept, this activity assists the patient in seeing trust as multi-faceted. This in turn assists the patient in challenging extreme or exaggerated stuck points related to trust. Based on results from feedback on the qualitative study, the new conceptualization of the Trust Star in CPT-L included the participant's trust in their own ability to navigate the process of disclosure of their status using their newly acquired tools (captured in the theme "different (less negative) view of HIV status". Upon completing the protocol, participants recognized that much of the shame they experienced about their HIV status, victimization experiences, and intersecting identities was founded on flawed beliefs about the nature of the disease, the world, themselves, and others (captured in themes "practicing challenging thoughts and stuck points" and "improving self-acceptance and self-esteem"). Using thought challenging skills, participants developed more adaptive worldviews that fostered greater self-acceptance (especially regarding sexual orientation and gender identity), trust in others, and trust in themselves (captured in themes "practicing challenging thoughts and stuck points"). Consequently, they were better able to engage in self-care (eg, hygiene and medication adherence) and were more motivated to connect with their social support networks. From the perspective of a social safety theoretical framework, in removing inappropriate shame and internalized stigma that hinders social bonds, feedback from participants highlights CPT-L's potential as an effective tool to decrease some of the deleterious health consequences that traumatized PWH experience as a function of social threat and disconnection [62].

XSL•FO
RenderX

Finally, PTSD-associated avoidance may underlie ART nonadherence such that the medications themselves may cue trauma memories and HIV-related stigma that deter consistent use. In other words, regardless of the reason (eg, forgetting vs choosing), it is possible that medication noncompliance is functionally an avoidant behavior used to escape distressing internal experiences for people struggling with co-occurring HIV and PTSD. In response, CPT-L features exposure as a core intervention component. Specifically, rather than the standard trauma narrative, participants are assigned 2 reflection statements recounting their index trauma and their experience with the news of their HIV status (unless HIV status is the index trauma). For many participants, the therapy setting was their first experience talking, or even deliberately thinking, about their index trauma and HIV status. They generally observed that actively approaching these topics allowed them to reorganize their thoughts and reframe their experiences. To their surprise, participants discovered that their broadened perspectives cultivated esteem and intimacy as they reflected on their relationships to themselves and others. These findings were consistent with evidence suggesting that disclosing trauma histories within the context of research or health care settings poses minimal risk of worsening symptoms, and in some cases can have secondary therapeutic benefit [63]. This data may help address misinformation and provider fears about "retraumatizing" PWH when screening for PTSD. In addition, participants were able to reconceptualize the cause and meaning of their index trauma and HIV status. Coupled with their fostered sense of trust in their own abilities, the shift away from self-blame and internalized stigma helped them regain a sense of control of their health and greater confidence in their ability to manage their own care.

While the intervention was largely well-received, several themes reflecting the need for modifications also emerged within the data. Participants primarily focused on improving handouts and worksheets stating that although the material was helpful, it was not intuitive, and the terminology was sometimes unclear. Several participants specified that they would benefit from additional examples pointedly touching on issues at the intersection of HIV and PTSD. Participants maintained that they would like to see self-acceptance and self-esteem for trauma-exposed PWH modeled through treatment materials (eg, including participant "success stories" and examples of adaptive alternatives to common stuck points).

Notwithstanding these limitations, CPT-L appears to be acceptable and relevant to PWH experiencing traumatic stress and merits further investigation as an integrated intervention for co-occurring PTSD and HIV. After challenging stigma-related stuck points, participants expressed feeling less constrained by maladaptive thoughts. These shifts translated to increased self-efficacy with both HIV-related care and mental health. Although limited in size and scope, the current study yielded preliminary support for the tolerability and effectiveness of the intervention. Findings highlight the importance of tailoring mental health treatment approaches to address trauma-related stuck points, intersecting identities (including living with HIV), and challenging internalized stigma. Such adaptations support goals for health outcomes (eg, ART adherence) by reducing the intensity of trauma-related distress and patterns of avoidance that interfere with HIV-treatment engagement. Future studies can expand on these efforts to complete the ADAPT-ITT process, including refining and testing CPT-L on a wider scale. Based on these qualitative results, integration and evaluation of strategies that foster social safety in mental health interventions for PWH warrants further attention.

Acknowledgments

This manuscript was supported by a grant (R34MH125706) from the National Institute of Mental Health (NIMH), National Institutes of Health (NIH) and partly supported by a grant (T32MH18869) from the NIMH, NIH and grant (K24DA39783; National Institute on Drug Abuse). Views expressed in this article do not necessarily reflect those of the funding agencies acknowledged. We acknowledge and thank our CAB members for their input and guidance across multiple aspects of our research program. No artificial intelligence (AI) tools were used in any portion of the manuscript.

Data Availability

The data generated for this study are not publicly available due to the sensitive nature of qualitative data and collection procedures. Recordings and transcripts of interviews are not publicly available given the confidentiality agreements with participants that were reviewed upon study enrollment.

Conflicts of Interest

PR receives royalties from Guilford Publications and conducts workshops on cognitive processing therapy. SS receives royalties from Oxford University Press, Guildford Publications, and Springer Humana Press as an author of books related to cognitive behavioral therapy, health psychology, and LGBT (lesbian, gay, bisexual, and transgender)-affirmative psychotherapy. The authors declare that they have no other known competing financial interest.

References

XSL•FO
RenderX

1. HIV surveillance report, 2019.: Centers for Disease Control and Prevention URL: https://www.cdc.gov/hiv/pdf/library/reports/surveillance/cdc-hiv-surveillance-report-2018-updated-vol-32.pdf [accessed 2024-11-25]
2. Safren SA, Gershuny BS, Hendriksen E. Symptoms of posttraumatic stress and death anxiety in persons with HIV and medication adherence difficulties. AIDS Patient Care STDS 2003;17(12):657-664 [FREE Full text] [doi: 10.1089/108729103771928717] [Medline: 14746659]
3. McLean CP, Fitzgerald H. Treating posttraumatic stress symptoms among people living with HIV: a critical review of intervention trials. Curr Psychiatry Rep 2016;18(9):83 [FREE Full text] [doi: 10.1007/s11920-016-0724-z] [Medline: 27439305]
4. Sherr L, Nagra N, Kulubya G, Catalan J, Clucas C, Harding R. HIV infection associated post-traumatic stress disorder and post-traumatic growth--a systematic review. Psychol Health Med 2011;16(5):612-629 [FREE Full text] [doi: 10.1080/13548506.2011.579991] [Medline: 21793667]
5. Pantalone DW, Rood BA, Morris BW, Simoni JM. A systematic review of the frequency and correlates of partner abuse in HIV-infected women and men who partner with men. J Assoc Nurses AIDS Care 2014;25(1 Suppl):S15-S35 [FREE Full text] [doi: 10.1016/j.jana.2013.04.003] [Medline: 24070646]
6. Kronish IM, Edmondson D, Li Y, Cohen BE. Post-traumatic stress disorder and medication adherence: results from the mind your heart study. J Psychiatr Res 2012;46(12):1595-1599 [FREE Full text] [doi: 10.1016/j.jpsychires.2012.06.011] [Medline: 22809686]
7. Hou J, Fu J, Meng S, Jiang T, Guo C, Wu H, et al. Posttraumatic stress disorder and nonadherence to treatment in people living with HIV: a systematic review and meta-analysis. Front Psychiatry 2020;11:834 [FREE Full text] [doi: 10.3389/fpsyt.2020.00834] [Medline: 32973579]
8. Katz IT, Ryu AE, Onuegbu AG. Impact of HIV-related stigma on treatment adherenceystematic review and meta-synthesis. Journal of the International AIDS Society 2013;16(3):18640 [FREE Full text] [doi: 10.7448/ias.16.3.18640]
9. Global HIV & AIDS statistics—fact sheet.: UNAIDS; 2019. URL: https://www.unaids.org/sites/default/files/media_asset/UNAIDS_FactSheet_en.pdf [accessed 2024-11-25]
10. Maggiolo F, Airoldi M, Kleinloog HD, Callegaro A, Ravasio V, Arici C, et al. Effect of adherence to HAART on virologic outcome and on the selection of resistance-conferring mutations in NNRTI- or PI-treated patients. HIV Clin Trials 2007;8(5):282-292 [FREE Full text] [doi: 10.1310/hct0805-282] [Medline: 17956829]
11. Paterson DL, Swindells S, Mohr J, Brester M, Vergis EN, Squier C, et al. Adherence to protease inhibitor therapy and outcomes in patients with HIV infection. Ann Intern Med 2000;133(1):21-30 [FREE Full text] [doi: 10.7326/0003-4819-133-1-200007040-00004] [Medline: 10877736]
12. Harrigan PR, Hogg RS, Dong WWY, Yip B, Wynhoven B, Woodward J, et al. Predictors of HIV drug-resistance mutations in a large antiretroviral-naive cohort initiating triple antiretroviral therapy. J Infect Dis 2005;191(3):339-347 [FREE Full text] [doi: 10.1086/427192] [Medline: 15633092]
13. Blake Helms C, Turan JM, Atkins G, Kempf MC, Clay OJ, Raper JL, et al. Interpersonal mechanisms contributing to the association between HIV-related internalized stigma and medication adherence. AIDS Behav 2017;21(1):238-247 [FREE Full text] [doi: 10.1007/s10461-016-1320-2] [Medline: 26864692]
14. Lima VD, Harrigan RA, Bangsberg DR, Hogg RS, Gross R, Yip B, et al. The combined effect of modern highly active antiretroviral therapy regimens and adherence on mortality over time. J Acquir Immune Defic Syndr 2009;50(5):529-536 [FREE Full text] [doi: 10.1097/QAI.0b013e31819675e9] [Medline: 19223785]
15. Earnshaw VA, Chaudoir SR. From conceptualizing to measuring HIV stigma: a review of HIV stigma mechanism measures. AIDS Behav 2009;13(6):1160-1177 [FREE Full text] [doi: 10.1007/s10461-009-9593-3] [Medline: 19636699]
16. Bogart LM, Wagner GJ, Galvan FH, Klein DJ. Longitudinal relationships between antiretroviral treatment adherence discrimination due to HIV-serostatus, race,sexual orientation among African?American men with HIV. Annals of Behavioral Medicine 2010:184-190 [FREE Full text] [doi: 10.1007/s12160-010-9200-x]
17. Turan JM, Elafros MA, Logie CH, Banik S, Turan B, Crockett KB, et al. Challenges and opportunities in examining and addressing intersectional stigma and health. BMC Med 2019;17(1):7 [FREE Full text] [doi: 10.1186/s12916-018-1246-9] [Medline: 30764816]
18. Baugher AR, Beer L, Fagan JL, Mattson CL, Freedman M, Skarbinski J, et al. Prevalence of internalized HIV-related stigma among HIV-infected adults in care, United States, 2011-2013. AIDS Behav 2017;21(9):2600-2608 [FREE Full text] [doi: 10.1007/s10461-017-1712-y] [Medline: 28213821]
19. Sevelius JM, Patouhas E, Keatley JG, Johnson MO. Barriers and facilitators to engagement and retention in care among transgender women living with human immunodeficiency virus. Ann Behav Med 2014;47(1):5-16 [FREE Full text] [doi: 10.1007/s12160-013-9565-8] [Medline: 24317955]
20. Bunn JY, Solomon SE, Miller C, Forehand R. Measurement of stigma in people with HIV: a reexamination of the HIV Stigma Scale. AIDS Educ Prev 2007;19(3):198-208 [FREE Full text] [doi: 10.1521/aeap.2007.19.3.198] [Medline: 17563274]
21. Gesesew HA, Tesfay Gebremedhin A, Demissie TD, Kerie MW, Sudhakar M, Mwanri L. Significant association between perceived HIV related stigma and late presentation for HIV/AIDS care in low and middle-income countries: A systematic review and meta-analysis. PLoS One 2017;12(3):e0173928 [FREE Full text] [doi: 10.1371/journal.pone.0173928] [Medline: 28358828]

XSL•FO
RenderX

22. Gonzalez A, Locicero B, Mahaffey B, Fleming C, Harris J, Vujanovic AA. Internalized HIV stigma and mindfulness: associations with PTSD symptom severity in trauma-exposed adults with HIV/AIDS. Behav Modif 2016;40(1-2):144-163 [FREE Full text] [doi: 10.1177/0145445515615354] [Medline: 26584609]
23. Meyer IH. Prejudice, social stress, and mental health in lesbian, gay, and bisexual populations: conceptual issues and research evidence. Psychol Bull 2003;129(5):674-697 [FREE Full text] [doi: 10.1037/0033-2909.129.5.674] [Medline: 12956539]
24. Spinhoven P, Drost J, van Hemert B, Penninx BW. Common rather than unique aspects of repetitive negative thinking are related to depressive and anxiety disorders and symptoms. J Anxiety Disord 2015;33:45-52 [FREE Full text] [doi: 10.1016/j.janxdis.2015.05.001] [Medline: 26004746]
25. Diamond LM, Alley J. Rethinking minority stress: a social safety perspective on the health effects of stigma in sexually-diverse and gender-diverse populations. Neurosci Biobehav Rev 2022;138:104720 [FREE Full text] [doi: 10.1016/j.neubiorev.2022.104720] [Medline: 35662651]
26. Resick PA, Nishith P, Weaver TL, Astin MC, Feuer CA. A comparison of cognitive-processing therapy with prolonged exposure and a waiting condition for the treatment of chronic posttraumatic stress disorder in female rape victims. J Consult Clin Psychol 2002;70(4):867-879 [FREE Full text] [doi: 10.1037//0022-006x.70.4.867] [Medline: 12182270]
27. Safren SA, Otto M, Worth JL, Salomon E, Johnson W, Mayer K, et al. Two strategies to increase adherence to HIV antiretroviral medication: life-steps and medication monitoring. Behav Res Ther 2001;39(10):1151-1162 [FREE Full text] [doi: 10.1016/s0005-7967(00)00091-7] [Medline: 11579986]
28. Wingood GM, DiClemente RJ. The ADAPT-ITT model: a novel method of adapting evidence-based HIV Interventions. J Acquir Immune Defic Syndr 2008;47(1):S40-S46. [doi: 10.1097/QAI.0b013e3181605df1] [Medline: 18301133]
29. Rueda S, Park-Wyllie LY, Bayoumi A, Tynan AM, Antoniou TA, Rourke SB, et al. Patient support and education for promoting adherence to highly active antiretroviral therapy for HIV/AIDS. Cochrane Database Syst Rev 2006;2006(3):CD001442 [FREE Full text] [doi: 10.1002/14651858.CD001442.pub2] [Medline: 16855968]
30. Koenen KC, Ratanatharathorn A, Ng L. Posttraumatic stress disorder in the world mental health surveys. Psychological Medicine 2017;47(13):2260-2274 [FREE Full text] [doi: 10.1017/9781316336168.011]
31. Kalichman S, Kalichman MO, Cherry C. Medication beliefs and structural barriers to treatment adherence among people living with HIV infection. Psychol Health 2016;31(4):383-395 [FREE Full text] [doi: 10.1080/08870446.2015.1111371] [Medline: 26534788]
32. Breet E, Kagee A, Seedat S. HIV-related stigma and symptoms of post-traumatic stress disorder and depression in HIV-infected individuals: does social support play a mediating or moderating role? AIDS Care 2014;26(8):947-951 [FREE Full text] [doi: 10.1080/09540121.2014.901486] [Medline: 24666226]
33. Wingood GM, DiClemente RJ, Villamizar K, Er DL, DeVarona M, Taveras J, et al. Efficacy of a health educator–delivered HIV prevention intervention for latina women: a randomized controlled trial. Am J Public Health 2011;101(12):2245-2252 [FREE Full text] [doi: 10.2105/ajph.2011.300340]
34. Olatunji BO, Mimiaga MJ, O'Cleirigh C, Safren SA. Review of treatment studies of depression in HIV. Top HIV Med 2006;14(3):112-124. [Medline: 16946456]
35. Karlin BE, Cross G. From the laboratory to the therapy room: national dissemination and implementation of evidence-based psychotherapies in the U.S. department of veterans affairs health care system. Am Psychol 2014;69(1):19-33. [doi: 10.1037/a0033888] [Medline: 24001035]
36. Monson CM, Langkaas TF, Schnurr PP. Dissemination and implementation of evidence-based practices for posttraumatic stress disorder in post-9/11 veterans. Clinical Psychology: Science and Practice 2020;27(3):2329.
37. Chard KM. An evaluation of cognitive processing therapy for the treatment of posttraumatic stress disorder related to childhood sexual abuse. Journal of Consulting and Clinical Psychology 2005;73(5):965-971 [FREE Full text] [doi: 10.1037/0022-006x.73.5.965]
38. Monson CM, Fredman SJ, Adair KC. Cognitive-behavioral conjoint therapy for posttraumatic stress disorder: application to operation enduring and iraqi freedom veterans. J Clin Psychol 2008;64(8):958-971. [doi: 10.1002/jclp.20511] [Medline: 18613094]
39. Galovski TE, Blain LM, Mott JM, Elwood L, Houle T. Manualized therapy for PTSD: flexing the structure of cognitive processing therapy. J Consult Clin Psychol 2012;80(6):968-981 [FREE Full text] [doi: 10.1037/a0030600] [Medline: 23106761]
40. Nishith P, Nixon RDV, Resick PA. Resolution of trauma-related guilt following treatment of PTSD in female rape victims: a result of cognitive processing therapy targeting comorbid depression? J Affect Disord 2005;86(2-3):259-265 [FREE Full text] [doi: 10.1016/j.jad.2005.02.013] [Medline: 15935245]
41. Sobel AA, Resick PA, Rabalais AE. The effect of cognitive processing therapy on cognitions: impact statement coding. J Trauma Stress 2009;22(3):205-211 [FREE Full text] [doi: 10.1002/jts.20408] [Medline: 19479979]
42. Resick PA, Monson CM, Chard KM. Cognitive Processing Therapy for PTSD: A Comprehensive Manual. New York: Guilford Press; 2017.

XSL•FO
RenderX

43. Pearson CR, Kaysen D, Huh D, Bedard-Gilligan M. Randomized control trial of culturally adapted cognitive processing therapy for PTSD substance misuse and HIV Sexual Risk Behavior for Native American Women. AIDS Behav 2019;23(3):695-706 [FREE Full text] [doi: 10.1007/s10461-018-02382-8] [Medline: 30607757]
44. Lopez CM, Baker NL, Moreland AD, Bisca E, Wilson T, Slick N, et al. Development and feasibility testing of an integrated PTSD and adherence intervention cognitive processing therapy-life steps (CPT-L) to improve HIV outcomes: trial protocol. Contemp Clin Trials Commun 2023;33:101150 [FREE Full text] [doi: 10.1016/j.conctc.2023.101150] [Medline: 37273831]
45. Corbin JM, Strauss A. Grounded theory research: procedures, canons, and evaluative criteria. Qualitative Sociology 1990:3-21 [FREE Full text] [doi: 10.1007/bf00988593]
46. Naeem M, Ozuem W, Howell K, Ranfagni S. A step-by-step process of thematic analysis to develop a conceptual model in qualitative research. International Journal of Qualitative Methods 2023;22(11) [FREE Full text] [doi: 10.1177/16094069231205789]
47. Mwita KM. Factors influencing data saturation in qualitative studies. IJRBS 2022;11(4):414-420. [doi: 10.20525/ijrbs.v11i4.1776]
48. Global and regional estimates of violence against women: prevalence and health effects of intimate partner violence and non-partner sexual violence.: World Health Organization; 2013. URL: https://www.who.int/publications/i/item/9789241564625 [accessed 2024-11-24]
49. Anderson JC, Campbell JC, Glass NE, Decker MR, Perrin N, Farley J. Impact of intimate partner violence on clinic attendance, viral suppression and CD4 cell count of women living with HIV in an urban clinic setting. AIDS Care 2018 Apr;30(4):399-408 [FREE Full text] [doi: 10.1080/09540121.2018.1428725] [Medline: 29397777]
50. Brief DJ, Bollinger AR, Vielhauer MJ, Berger-Greenstein JA, Morgan EE, Brady SM, et al. Understanding the interface of HIV, trauma, post-traumatic stress disorder, and substance use and its implications for health outcomes. AIDS Care 2004;16(1):S97-120 [FREE Full text] [doi: 10.1080/09540120412301315259] [Medline: 15736824]
51. Gore-Felton C, Koopman C. Traumatic experiences: harbinger of risk behavior among HIV-positive adults. Journal of Trauma & Dissociation 2002;3(4):121-135 [FREE Full text] [doi: 10.1300/j229v03n04_07]
52. LeGrand S, Reif S, Sullivan K, Murray K, Barlow ML, Whetten K. A review of recent literature on trauma among individuals living with HIV. Curr HIV/AIDS Rep 2015;12(4):397-405 [FREE Full text] [doi: 10.1007/s11904-015-0288-2] [Medline: 26419376]
53. Leserman J. Role of depression, stress, and trauma in HIV disease progression. Psychosom Med 2008;70(5):539-545. [doi: 10.1097/PSY.0b013e3181777a5f] [Medline: 18519880]
54. Mugavero MJ, Pence BW, Whetten K, Leserman J, Swartz MJ, Stangl D, et al. Predictors of AIDS-related morbidity and mortality in a southern U.S. cohort. AIDS Patient Care STDs 2007;21(9):681-690 [FREE Full text] [doi: 10.1089/apc.2006.0167] [Medline: 17919095]
55. Remien RH, Stirratt MJ, Nguyen N, Robbins RN, Pala AN, Mellins CA. Mental health and HIV/AIDS: the need for an integrated response. AIDS 2019;33(9):1411-1420 [FREE Full text] [doi: 10.1097/QAD.0000000000002227] [Medline: 30950883]
56. Hilton N, McKee S, Ham E, Green M, Wright L. Co-occurring mental illness and substance use disorders in canadian forensic inpatients: underdiagnosis and implications for treatment planning. International Journal of Forensic Mental Health 2018;17(2):145-153 [FREE Full text] [doi: 10.1080/14999013.2018.1451416]
57. Resick PA, Schnicke MK. Cognitive processing therapy for sexual assault victims. J Consult Clin Psychol 1992;60(5):748-756 [FREE Full text] [doi: 10.1037//0022-006x.60.5.748] [Medline: 1401390]
58. Kalichman SC, Pope H, White D, Cherry C, Amaral CM, Swetzes C, et al. Association between health literacy and HIV treatment adherence: further evidence from objectively measured medication adherence. J Int Assoc Physicians AIDS Care (Chic) 2008;7(6):317-323 [FREE Full text] [doi: 10.1177/1545109708328130] [Medline: 19056866]
59. Balfour L, Kowal J, Silverman A, Tasca GA, Angel JB, Macpherson PA, et al. A randomized controlled psycho-education intervention trial: Improving psychological readiness for successful HIV medication adherence and reducing depression before initiating HAART. AIDS Care 2006;18(7):830-838 [FREE Full text] [doi: 10.1080/09540120500466820] [Medline: 16971295]
60. Altice FL, Mostashari F, Friedland GH. Trust and the acceptance of and adherence to antiretroviral therapy. J Acquir Immune Defic Syndr 2001;28(1):47-58 [FREE Full text] [doi: 10.1097/00042560-200109010-00008] [Medline: 11579277]
61. Hays PA. Addressing cultural complexities in practice: A framework for clinicians and counselors. American Psychological Association 2009;40(4):354-360. [doi: 10.1037/a0016250]
62. Machtinger EL, Lavin SM, Hilliard S, Jones R, Haberer JE, Capito K, et al. An expressive therapy group disclosure intervention for women living with HIV improves social support, self-efficacy, and the safety and quality of relationships: a qualitative analysis. J Assoc Nurses AIDS Care 2015;26(2):187-198 [FREE Full text] [doi: 10.1016/j.jana.2014.05.001] [Medline: 25027284]
63. Jaffe AE, DiLillo D, Hoffman L, Haikalis M, Dykstra RE. Does it hurt to ask? A meta-analysis of participant reactions to trauma research. Clin Psychol Rev 2015;40:40-56 [FREE Full text] [doi: 10.1016/j.cpr.2015.05.004] [Medline: 26051308]

XSL•FO
RenderX

Abbreviations

ADAPT-ITT: Assessment, Decision, Adaptation, Production, Topical experts-integration, Training, and Testing
ART: antiretroviral therapy
CAB: Community Advisory Board
CPT: cognitive processing therapy
CPT-L: CPT-Life-Steps for integration of adherence
DSM-V: Diagnostic and Statistical Manual of Mental Disorders (Fifth Edition)
DSM-5-TR: Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, Text Revision
Life-Steps: Life-Steps for Medication Adherence
MoCA: Montreal Cognitive Assessment
PTSD: posttraumatic stress disorder
PWH: persons with HIV
REDCap: Research Electronic Data Capture

Edited by A Mavragani; submitted 12.07.24; peer-reviewed by MKS Smith, JI Campbell; comments to author 02.09.24; revised version received 23.09.24; accepted 27.09.24; published 16.01.25.

Please cite as:
Lopez CM, Moreland AD, Amaya S, Bisca E, Mujica C, Wilson T, Baker N, Richey L, Eckard AR, Resick PA, Safren SA, Danielson CK
Assessment, Decision, Adaptation, Production, Topical Experts-Integration, Training, and Testing (ADAPT-ITT) Framework to Tailor Evidence-Based Posttraumatic Stress Disorder Treatment for People With HIV to Enhance Engagement and Adherence: Qualitative Results from a Feasibility Randomized Controlled Trial
JMIR Form Res 2025;9:e64258
URL: https://formative.jmir.org/2025/1/e64258
doi:10.2196/64258
PMID:

©Cristina M Lopez, Angela D Moreland, Stephanie Amaya, Erin Bisca, Christin Mujica, Tayler Wilson, Nathaniel Baker, Lauren Richey, Allison Ross Eckard, Patricia A Resick, Steven A Safren, Carla Kmett Danielson. Originally published in JMIR Formative Research (https://formative.jmir.org), 16.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Designing the First Pregnancy Guaranteed Income Program in the United States: Qualitative Needs Assessment and Human-Centered Design to Develop the Abundant Birth Project

Deborah Karasek[1,2], MPH, PhD; Jazzmin C Williams[3], MD; Michaela A Taylor[4], MPH; Monica M De La Cruz[5], MPH; Stephanie Arteaga[5], MPH; Sabra Bell[6]; Esperanza Castillo[3], MS; Maile A Chand[4,6], BA; Anjeanette Coats[5]; Erin M Hubbard[5], MPH; Latriece Love-Goodlett[4], BA; Breezy Powell[4,6]; Solaire Spellen[7], MPH; Zea Malawa[3,4], MPH, MD; Anu Manchikanti Gomez[5], MSc, PhD

[1]School of Public Health, Oregon Health & Science University and Portland State University, Portland, OR, United States
[2]Department of Obstetrics, Gynecology, and Reproductive Sciences, University of California, San Francisco, San Francisco, United States
[3]University of California, San Francisco, San Francisco, CA, United States
[4]Expecting Justice, San Francisco, CA, United States
[5]Sexual Health and Reproductive Equity Program, School of Social Welfare, University of California, Berkeley, Berkeley, CA, United States
[6]Department of Human Ecology, Univeristy of California, Davis, Davis, CA, United States
[7]California Preterm Birth Initiative, University of California, San Francisco, San Francisco, CA, United States

Corresponding Author:
Anu Manchikanti Gomez, MSc, PhD
Sexual Health and Reproductive Equity Program
School of Social Welfare
University of California, Berkeley
110 Haviland Hall, MC 7400
Berkeley, CA, 94720-7400
United States
Phone: 1 510 6420722
Email: anugomez@berkeley.edu

Abstract

Background: Racial inequities in pregnancy outcomes persist despite investments in clinical, educational, and behavioral interventions, indicating that a new approach is needed to address the root causes of health disparities. Guaranteed income during pregnancy has the potential to narrow racial health inequities for birthing people and infants by alleviating financial stress.

Objective: We describe community-driven formative research to design the first pregnancy-guaranteed income program in the United States—the Abundant Birth Project (ABP). Informed by birth equity and social determinants of health perspectives, ABP targets upstream structural factors to improve racial disparities in maternal and infant health.

Methods: The research team included community researchers, community members with lived experience as Black or Pacific Islander pregnant, and parenting people in the San Francisco Bay Area. The team conducted needs assessment interviews and facilitated focus groups with participants using human-centered design methods. Needs assessment participants later served as co-designers of the ABP program and research, sharing their experiences with financial hardships and government benefits programs and providing recommendations on key program elements, including fund disbursement, eligibility, and amount.

Results: Housing affordability and the high cost of living in San Francisco emerged as significant sources of stress in pregnancy. Participants reported prohibitively low income eligibility thresholds and burdensome enrollment processes as challenges or barriers to existing social services. These insights guided the design of prototypes of ABP's program components, which were used in a design sprint to determine the final components. Based on this design process, the ABP program offered US $1000/month for 12 months to pregnant Black and Pacific Islander people, selected through a lottery called an abundance drawing.

Conclusions: The formative design process maximized community input and shared decision-making to co-design a guaranteed income program for Black and Pacific Islander women and people. Our upstream approach and community research model can inform the development of public health and social service programs.

XSL•FO
RenderX

(JMIR Form Res 2025;9:e60829) doi:10.2196/60829

KEYWORDS

maternal and child health; economics; public health; qualitative research methods; programs (evaluation and funding); community-centered; pregnancy; first pregnancy; behavioral interventions; racial health; financial stress; Abundant Birth Project; infant health; infant; Black

Introduction

Background

Pregnancy is a critical period of development when stress can have deleterious effects on both the pregnant person and their baby, leading to adverse birth and long-term health outcomes [1,2] that disproportionately affect Black, Indigenous, and Native Hawaiian or Pacific Islander people [3,4]. In the United States, Black infants are twice as likely to be born preterm or with low birth weight compared to White infants [3]. These inequities extend to maternal health, with Black women being 3.3 times more likely to die during or after pregnancy than White women [5]. While rates of adverse outcomes are often obscured by aggregation of data among Asian and Pacific Islander subgroups [6], Native Hawaiian and other Pacific Islander people experience higher rates of low birth weight, preterm birth, and macrosomia [3,7,8]. Even in San Francisco, a wealthy, well-resourced city, the shrinking Black and Pacific Islander populations experience significant inequities in birth outcomes [9]. Black and Pacific Islander pregnant women and people experienced the highest rates of preterm birth, at 12.7% and 8.9%, respectively, in 2014-2016, the last years with enough births to Pacific Islander people to reliably report [10]. Research increasingly indicates that racism—structural, institutional, and interpersonal—is a key underlying driver of these inequities [11,12]. It is unsurprising, therefore, that clinical, educational, and behavioral interventions have failed to meaningfully narrow racialized birth disparities [3,13-15].

Efforts to reduce inequities must address exposure to interpersonal, institutional, and structural racism as the root cause of disproportionate risk [16-19]. Racial inequity in income and wealth is one manifestation of structural racism, with important public health ramifications [16,20]. The racial wealth gap has resulted from a long history of exclusion of Black families from economic opportunity, through a combination of redlining, exclusionary zoning, subprime loans, and “urban renewal” programs, among other tactics [21]. This leaves Black families and other families of color more vulnerable to economic stress during pregnancy. Financial stress may be particularly acute during pregnancy, a period of income volatility when families frequently experience changes in housing, employment, health care, and childcare costs [22]. Indeed, in a nationally representative study of pregnant and postpartum women, over half reported general financial stress [23]. Yet, efforts to directly alleviate poverty and financial stress to address birth inequities are limited. In San Francisco, the median annual income in 2020 for White households was US $146,043, while it was just US $38,862 for Black households and US $90,917 for Pacific Islander households [24], making it the city with the third largest income disparity in the United States [25]. The Insight Center Family Needs calculator estimated the minimum income required to live in San Francisco in 2021 without public or private assistance as US $60,232 for a single adult, US $120,519 for a single adult with 1 preschooler, and US $167,432 for a household with 2 adults, 1 infant, and 1 preschooler [26]. To meet the needs of a family of three, Black and Pacific Islander households would have needed to earn 4 and 2 times their current average incomes, respectively.

Guaranteed income describes the disbursement of unconditional cash to individuals or families in a community without restrictions, allowing flexibility to close gaps where needs are not being met by public benefits programs [27]. Guaranteed income programs offer the potential to address poverty and stress caused by economic racism and the racial wealth gap. Philanthropic funders, cities, and states, including California, are increasingly investing in guaranteed income as a poverty reduction and equity-generating strategy [28]. Emerging findings from the growing research on United States guaranteed income pilots have found that monthly income supplements improve the sense of trust in institutions, dignity, and autonomy and increase savings, without changing spending patterns [29,30]. Magnolia Mothers Trust, a program providing US $1000 per month for 12 months to Black mothers living in federally subsidized housing in Jackson, Mississippi, allowed mothers to save for emergencies, buy food, pay childcare expenses, and prioritize their health [30].

At the time of the Abundant Birth Project’s (ABP) inception, there were no guaranteed income programs in the United States that addressed the critical period of pregnancy. However, evidence from other contexts supported a guaranteed income in pregnancy as a health intervention. A quasi-experimental study among 14,591 pregnant women in Manitoba, Canada, found that an approximately 10% increase in income from a pregnancy income benefit resulted in reductions of low birthweight by 21% and preterm birth by 18% [31,32]. Since the launch of ABP, there has been a proliferation of guaranteed income programs across California, including other programs that target pregnancy [28,33,34].

The Abundant Birth Project

ABP is a program of Expecting Justice, a Black-led, cross-sector collaborative focused on birth equity in San Francisco that includes mothers, public health professionals, health and social systems partners, and governmental leaders. Expecting Justice uses a collective impact model, which is a process to bring together diverse stakeholders to create a common agenda, establish shared measurement, and foster mutually reinforcing activities for stakeholders [35]. As part of its work to advance birth equity, Expecting Justice is guided by community voice and emphasizes upstream solutions that address the social determinants of health. As defined by the National Birth Equity Collaborative, birth equity is the assurance of optimal birth

XSL•FO
RenderX

conditions for all people, with an ongoing effort to address racial and social inequities [36]. The social determinants of health result from the larger systems at play, such as economic policies, racism, and political systems that shape the conditions of daily life and health [37,38]. These frameworks informed the focus on this pregnancy-guaranteed income program as an equity-generating strategy to directly address income insecurity as a social determinant of health by interrupting the pathway from economic exclusion and stress to adverse birth outcomes. Expecting Justice targeted this intervention toward communities most likely to experience preterm birth to help narrow disparities in preterm birth rates. Focusing on the most marginalized aligns with the birth equity framework, emphasizing that targeted support, rather than universal availability, is essential for addressing birth inequities [36].

Expecting Justice entered a community-academic partnership with the University of California, Berkeley, and the University of California, San Francisco, to design and evaluate ABP. We developed the ABP program and evaluation by actively, meaningfully, and intentionally centering the voices and experiences of the communities most affected by the adverse birth outcomes the program seeks to improve. ABP launched in June 2021, with the goal of enrolling 150 Black and Pacific Islander pregnant women and people (including gender-diverse individuals with the capacity for pregnancy) over 18 months.

This paper describes the formative research process, under the co-leadership of community members, used to design ABP.. Our primary objectives were to (1) determine the financial needs of Black and Pacific Islander women and people during pregnancy, and (2) collaboratively design a pregnancy-guaranteed income pilot program with Black and Pacific Islander mothers. We hope that the process and results from the formative research to design ABP will be useful to other researchers, practitioners, and community groups developing interventions to address racial health disparities, including guaranteed income pilot programs.

Methods

Community Research Model

ABP used shared decision-making and community-centered practices in all phases of the program development and evaluation, by purposely centering the voices of community members with lived expertise [39]. As a community-based participatory research project, ABP recognized that this relationship is dynamic and requires continual effort to rebalance traditional power structures [40]. To support these efforts, 4 community researchers (CRs) served as key members of the community-academic partnership. The CRs are Black and Pacific Islander community members who have lived experience of pregnancy and parenting in the San Francisco Bay Area. They come from a variety of professional backgrounds and have an interest in learning or experience using research methods to advance birth equity. We adopted a shared power structure and decision-making process that included following Expecting Justice's "expert opinion protocol" for all final decisions. This protocol enabled ABP stakeholders who identified as Black and/or Pacific Islander mothers to have protected time to share their insights with the larger group without contradiction before voting. We also used weighted voting in decision-making to prioritize the perspectives of those who identified as Black and/or Pacific Islander mothers.

The process of designing the ABP program and evaluation involved (1) conducting interviews as part of the needs assessment, (2) a Rapid Assessment Process (RAP) for analysis, (3) a prototype development design sprint, and (4) a usability testing design sprint (Figure 1). The CRs contributed to the full process, bringing expertise from their lived experience with pregnancy and parenting in the Black and Pacific Islander communities: developing research questions, formulating survey and interview instruments, recruiting participants, conducting interviews, facilitating design sprints, interpreting results, and making key program decisions. CRs received training in data collection and analysis supported by the university-based staff, which coupled with the team's shared power structure, helped to dismantle the hierarchy often present in research institutions that keeps community members out of decision-making processes [41,42]. Expecting Justice also led an ABP working group with birth equity stakeholders, including Black and Pacific Islander mothers, community organization representatives, social service providers, maternal health leaders and researchers, and representatives from health plans to collaborate and provide strategic resources to support program development. The ABP working group met monthly to inform key program design and implementation decisions before, during, and following the formative phase.

XSL•FO
RenderX

Figure 1. The design process for the Abundant Birth Project involved conducting needs assessment interviews, an RAP to analyze findings, a prototype development design sprint, and a usability design sprint to arrive at the final design. ABP: Abundant Birth Project.

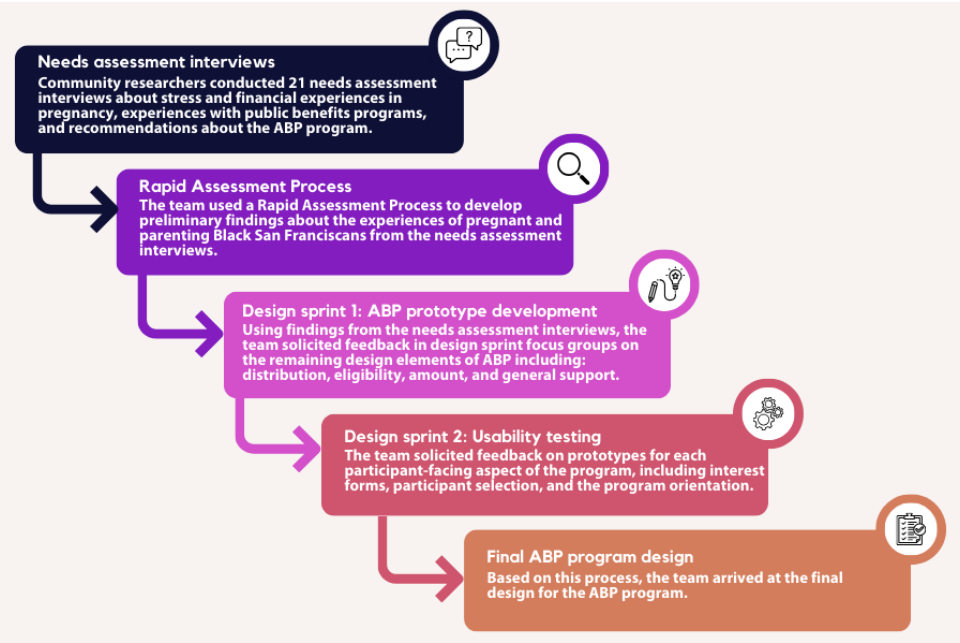

The shared decision-making structure with the ABP program and research teams, CRs, and ABP working group led to tangible improvements in research and program design. For example, the CRs voiced concerns that research studies rarely compensate participants in a way that truly values their time. Under their leadership, the team increased the amount of participant remuneration for the interviews and design sprints from US $50 to US $75 and US $100 to US $200, respectively.

Needs Assessment Interviews

To inform ABP's design, we first conducted interviews to understand the experiences Black and Pacific Islander women and people face during pregnancy in San Francisco. Inclusion criteria were aged 18 years or older, identified as Black and/or Pacific Islander, spoke English, and either currently pregnant or had been pregnant while residing in San Francisco in the past 5 years. Participants were recruited from social media and local community organizations. The CRs closely collaborated on the development of the interview guide to ensure that language and questions would capture appropriate topics in a comfortable environment for participants. We conducted interviews from April to June 2020. The CRs conducted the interviews using a web-based videoconference platform or by telephone. Interviews were recorded and professionally transcribed.

The interview guide addressed stress and financial experiences during pregnancy and experiences with and perception of public benefits programs. We also asked participants for recommendations on the ABP program design, including how much money would alleviate financial stress during pregnancy, how to best distribute the income supplement, and program eligibility criteria and restrictions.

Before conducting interviews, the CRs underwent training in qualitative research methods led by Dr Brittany Chambers, a Black scholar, mother, and community engagement expert. The training included an introduction to asking research questions, qualitative data collection, and interviewing skills. The entire research team went through the training together, building trust and rapport throughout the process. The CRs described the importance of being trained by a researcher from their own communities. Dr Chambers used art-based activities that allowed the CRs to celebrate their stories and the whole team to tap into their imagination and creativity and engage with one another outside of hierarchies. CRs received compensation for their time and expertise in the training.

About RAP

To identify key topics, we used RAP, an intensive, team-based, iterative analytic approach that uses data triangulation to develop preliminary findings from the perspective of an insider—pregnant and parenting Black and Pacific Islander San Franciscans [43]. RAP is well-suited for qualitative research informing the design of interventions, as it allows for rapid and timely data reduction and ongoing analysis during data collection. We brought together the team, including CRs, program staff, and student research assistants for shared qualitative analysis training. CRs, other research team members, and program staff used summary templates to create a memo for each interview. Members of the research team reviewed memos for consistency. The first step of data reduction involved matching domains from the summary template to interview questions. The domains included concerns about stress and pregnancy, employment, experiences with social services, desired programs and services, feelings about the income supplement, recommended supplement amount, eligibility restrictions, disbursement, recruitment recommendations, and additional financial services. Finally, we used an individual-level data matrix to synthesize key data across participants and identify themes.

XSL•FO
RenderX

Human-Centered Design Sprints

Overview

Using design thinking methodology, we conducted 2 focus groups, called human-centered design sprints, to develop and refine the facets of the guaranteed income program following themes identified from the need assessment interviews. Originating in consumer technology, design thinking is a method of developing innovative solutions and products with leadership and input from the end user [44]. A design sprint refers to a rapid, human-centered design activity, in which individuals are brought together to discuss and iterate on program design components [45]. Design sprints prioritize qualitative feedback from stakeholders and target end users, and encourage creative thinking [45]. Human-centered design methodology is especially well-suited for developing urgently needed interventions to reduce health inequities and has previously been used to gather community input on solutions for perinatal disparities [46]. We partnered with design thinking firms that helped determine the structure of the design sprints and trained the ABP program staff and CRs on how to facilitate these sessions.

Due to the COVID-19 pandemic, we conducted the design sprints using videoconferencing software and recorded the sessions. We provided tablets with internet connectivity to all design sprint participants who lacked internet access to participate virtually. For continuity, we invited participants from the needs assessment interviews to participate in the design sprints. Co-designing with the same participants also ensured that their voices and perspectives were reflected in the program prototypes. Each design sprint session was 4 hours.

Design Sprint 1: ABP Prototype Development

Guided by findings from the needs assessment interviews, our team determined 4 remaining design questions to be answered through the first design sprint.

1. Distribution: How can we best deliver money in a way that is both inclusive and convenient?
2. Eligibility: How can we best serve the Black and Pacific Islander communities in San Francisco while allowing for flexibility in eligibility criteria?
3. Amount: How can we balance financially supporting pregnancy needs with the desire to support as many people as possible?
4. General support: How can we support pregnant people, not just financially, but holistically, during their pregnancy?

Before the design sprint, we provided participants with a welcome packet to introduce ABP and the purpose of the session. The welcome packet included a journey map to contextualize the human-centered design sprints for participants (Figure 2). The journey map was developed using common themes from the needs assessment interviews to guide participants through a story of a Black or Pacific Islander pregnant woman in San Francisco, focusing on how racism contributes to financial and housing challenges, stress, and stigma.

Figure 2. The journey map was developed using themes from needs assessment interviews conducted from April to June 2020 for the Abundant Birth Project. It described how racism contributes to financial and housing challenges, stress, and stigma for Black and Pacific Islander pregnant women in San Francisco. SF: San Francisco; PI: Pacific Islander; WIC: Women, Infant and Children.

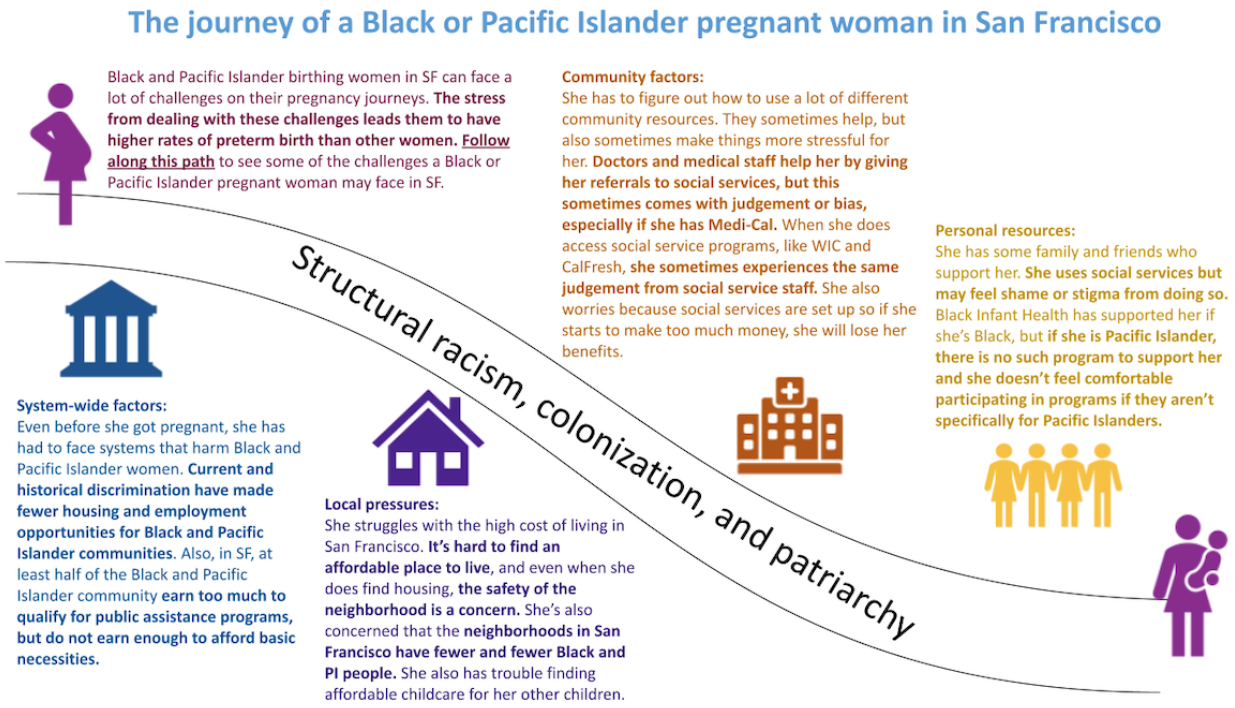

Following the first design sprint, we synthesized feedback from participants with findings from the needs assessment interviews to develop ABP program prototypes. Specific recommendations from participants in interviews and the first design sprint informed the prototypes for key components of ABP, including distribution method, amount, and eligibility criteria. The themes surrounding previous experiences with social services informed

XSL•FO
RenderX

components of the ABP prototype that focused on support that should be offered in addition to the supplement itself.

Design Sprint 2: Usability Testing

Using the learnings from the first design sprint, our team developed prototypes for each participant-facing aspect of the program, including the interest form, eligibility interview, participant selection, program enrollment, orientation, program experience, and research engagement. Design thinking is iterative; therefore, the second design sprint focused on usability testing with the ABP prototypes. Participants gave feedback on the prototypes, which helped ensure that the final program design aligned with community values and perspectives. CRs facilitated breakout groups that focused on each aspect of the program experience. For research engagement, participants were asked about what concerns they would have about engaging in research and what would make them excited to opt into the research. We made final decisions regarding the prototypes using Expecting Justice's "expert opinion" protocol.

Ethical Considerations

The institutional review board of the University of California, San Francisco, approved this study's protocol (IRB #19-27608, reference #267614). Participants signed consent forms for interviews and design sprints after reviewing them with study personnel. The research team removed identifiers (eg, names) from interview transcripts to deidentify data. Participants received a US $75 gift card incentive for taking part in the needs assessment interviews and a US $200 gift card incentive for participation in each design sprint.

Results

Needs Assessment Interviews

Participant Demographics

We conducted 21 in-depth interviews from April to June 2020. Seventeen participants identified as Black or African American, 5 as Pacific Islander, 2 as Latinx, and 1 as Native American or Indigenous (Table 1). All 21 participants identified as women. Eleven participants had experienced homelessness in the past, and 15 were at risk of food insecurity [47].

Table 1. Selected demographics for Abundant Birth Project formative research participants in San Francisco, California. Formative research needs assessment interviews and design sprints occurred from April to December 2020 (N=21).

Characteristic	Value, n (%)
Racial identity[a]	
Black or African American	17 (81)
Latinx	2 (10)
Native American or Indigenous	1 (5)
Pacific Islander	5 (24)
Pregnancy status	
Pregnant at the time of interview	6 (29)
Pregnant in the last 5 years	15 (71)
Income	
100% federal poverty level	7 (33)
100%-200% federal poverty level	6 (29)
>200% federal poverty level	6 (29)
Do not know	2 (9)

[a]Racial identity categories are not mutually exclusive.

Experiences of Financial Stress in Pregnancy

All participants discussed feeling stressed during their pregnancies; sources included negotiating employment leave, the health of their baby and themselves, and most frequently, housing and the high cost of living in San Francisco. Some participants reported being housing insecure or homeless during their pregnancies, underscoring the precarity of meeting basic needs in San Francisco. One participant described how the cost of living in the Bay Area was a primary driver of her stress.

> *Especially living in San Francisco Bay Area, everyone is so uptight and on edge because it's so expensive to live here, and we're always stressed-out about working or money and just trying to make it, so with all of that already going on top of bringing another life into the world, things can get really, really stressful.* [Currently pregnant, identified as Black and Latinx]

Another participant discussed her distressing personal experience with homelessness and the tradeoffs she had to consider to keep her child safe.

> *I was couch surfing. I was homeless. I became homeless March of 2019, so I was couch surfing waiting for a shelter to come about or some type of services, so I've been just couch surfing, sleeping on*

XSL•FO
RenderX

> *the…well, I will sleep at the bus stop, but my son would be with one of my friends during the night, so that he would [be] able to sleep comfortably.* [Last pregnant in 2020, identified as Pacific Islander]

Participants discussed the need for assistance programs that could include middle-income families. One participant underscored this by saying:

> *The hard part is being in this random middle class just because of my husband's one job, even though nobody's taking into account all of our expenses and things. And so, it's still kind of like we're low-income, but it doesn't read low-income on paper.* [Last pregnant in 2018, identified as Black, Latinx, and Native American or Indigenous]

Other participants who considered themselves middle-income echoed this "catch-22" situation, in which they struggled financially because they did not make enough to meet the costs of living in the San Francisco Bay Area, but their income surpassed the federal poverty line used to calculate public benefit eligibility, rendering them ineligible for assistance. Describing this impossible situation, one participant said:

> *Just because someone has a job doesn't mean that they shouldn't not qualify and get this extra help, because they still have stress. They still are worried about something financially, you know…what I'm saying is like most programs if you make less than $35,000 a year. Well, that's poor, poor, poor in San Francisco. But what about like the person who makes $32 an hour, you know, that can't make ends meet or is stressing and struggling. Like there's never any kind of help or support for them. They can't get the Medi-Cal. They can't get the WIC. No welfare. Children's Council. None of that.* [Most recently pregnant in 2019, identified as Black]

Experiences With Government Benefits Programs

Most participants (18/21) had experiences with government benefits programs, including CalFresh (California's Supplemental Nutrition Assistance Program), CalWorks (California's Temporary Assistance to Needy Families program, commonly known as "welfare"), Medi-Cal (California's Medicaid program), the Special Supplemental Nutrition Program from Women, Infant and Children, and Black Infant Health, a statewide program to improve the health of Black mothers and babies. Participants found most government benefits programs financially helpful and were grateful for the aid. One participant said:

> *CalFresh helped a lot. It was a lifesaver, actually. We were just so broke at that time and paying San Francisco rent and everything, so that definitely helped us get by until I got a more full-time job, and things changed after that, but it definitely got us through.* [Currently pregnant, identified as Pacific Islander]

Despite appreciating the financial assistance, participants found the application processes for many benefits programs burdensome and income eligibility requirements too restrictive, given the cost of living in San Francisco. Additionally, participants reported feeling stigmatized by the benefits program staff or when using certain resources, such as a CalFresh Electronic Benefit Transfer card at a grocery store or Medi-Cal at their health care provider. Describing previous stigmatizing experiences while using Electronic Benefit Transfer at the grocery store, one participant said:

> *I feel like you're looked at differently, you're really treated differently. Like, I would have people, like, cashiers, kind of put me to the side like, 'Oh, figure out your check stuff, and then let me get these people that are paying.' Like, I've had people say that to me, and so that part of it is just like—and then, you know, [the cashiers] getting up on a loudspeaker—just it just, all that stuff, nobody's discreet about anything…* [Most recently pregnant in 2016, identified as Black]

Recommendations for a Pregnancy Guaranteed Income Program

Overall, participants enthusiastically supported the idea of an income supplement for Black and Pacific Islander people during pregnancy, feeling that the flexible resources would significantly reduce stress during pregnancy.

> *That would be so helpful and also show compassion towards the women [and] what they go through because people don't know really what we go through unless you had the baby...that money would definitely be helpful towards whether or not you pay for diapers or food, you know what I'm saying. Like sometimes you can't do both, but you need both, right?* [Most recently pregnant in 2018, identified as Pacific Islander]

We asked participants how much money per month they believed would reduce financial stress during pregnancy. Three participants felt "anything would help," 4 recommended less than US $500, 11 recommended an amount between US $700 and US $1000, and 4 believed more than US $1000 would be necessary due to the high cost of living in San Francisco. Participants recommended electronic fund distribution for speed and because many bills are paid on the web. They felt that cash may be spent too quickly, makes tracking spending difficult, and cannot be replaced if lost. Some also expressed concerns that cash could not be used for remote or web-based purchases, which was essential during the data collection period at the beginning of the COVID-19 pandemic. Many mentioned the need for an option for individuals without a bank account. When asked about potential other services that should be offered alongside ABP, most felt it would be valuable to offer optional support with finances (eg, credit building or repair, savings, banking, and taxes), childcare, and housing.

Some participants expressed concerns about how recipients would spend the income supplement. This may reflect false and racist narratives about how Black mothers use public assistance programs [48]. Often speaking in generalities, participants discussed concerns that income supplement recipients may not use the money in ways society deems "right." Participants also

XSL•FO
RenderX

reported concerns that if hypothetical recipients misused the funds, it may result in the cessation of the program.

> *Of course, we have those folks who would abuse it and maybe not really using it, well, I guess if it's no strings attached, it doesn't matter what they use it for, but I guess I would just be afraid of anything happening like that so where it would be abused to a point that we'd lose it altogether, you know what I mean, so that would be a concern.* [Most recently pregnant in 2018, identified as Pacific Islander]

Several participants discussed eligibility requirements, such as an income cap or debt-to-income ratio. These criteria were proposed to ensure that the limited resources are directed to the target population.

Design Sprint 1: ABP Prototype Development

Overview

Fifteen interviewees from the needs assessment participated in the first design sprint, which occurred in July 2020. The CRs facilitated breakout sessions to discuss the concept of an income supplement and address the 4 key program design questions.

Amount

Participants believed that a monthly supplement should be at least US $1000 to provide a noticeable benefit during pregnancy. As participants also wanted the program to accommodate as many people as possible, they preferred a supplement to reach 150 participants over a larger supplement reaching fewer people.

Disbursement

Some participants recommended multiple payment types for the disbursement of funds. As this was not feasible with the disbursement organizations, participants came to the consensus that an ABP-specific debit card (as opposed to cash, check, or direct deposit) would be the most acceptable way to distribute the supplement. Participants discussed what would make the debit card the most accessible, such as not requiring a bank account or social security number. They also expressed they wanted the ABP debit card to have a beautiful design that reflected their communities, so recipients would not feel stigmatized or judged as they often did in other benefits programs (Figure 3).

Figure 3. The Abundant Birth Project debit card included imagery that reflected participants' desire for a beautiful design to prevent feelings of stigma or judgment often associated with using other benefits programs.

Eligibility

Participants felt the program should have additional eligibility criteria aside from identifying as Black and/or Pacific Islander, such as gestational age in pregnancy and income. For income, participants expressed a desire for the program to serve individuals who do not qualify for government benefits but still struggle to meet their needs in San Francisco. Design sprint participants believed that ABP participants should be randomly selected to give everyone interested an equal chance of enrolling in the program. Importantly, participants cautioned against using the term "lottery" to describe this process, given that their communities had a history of negative experiences with lottery-based programs, such as the local public housing selection process.

Other Support

Participants believed that the program should prioritize offering participants opt-in case management support. While they felt that engaging with case management should not be required with an unconditional program, they acknowledged that people have trouble accessing existing resources and navigating the benefits system.

Design Sprint 2: Usability Testing

Eight participants from the first design sprint returned in December 2020 to test and provide feedback on all components of the ABP prototypes.

We elicited feedback on example recruitment documents. Participants preferred interest and referral forms that could be completed in 5 minutes or less. They suggested that the ABP website contain a portal to allow applicants to check their status from interest form submission to enrollment. To create a welcoming atmosphere, participants felt it was important for the website to have imagery representing Black and Pacific Islander people. This was especially important for the Pacific Islander participants because few programs tailored outreach specifically to this community. With this in mind, Expecting Justice collaborated with Black and Pacific Islander artists to design outreach materials such as flyers and social media posts. Expecting Justice consulted the San Francisco Pacific Islander Maternal Advisory Board about the best language to use to refer to the Pacific Islander population. They learned that the community felt most comfortable using "Pacific Islander," "Pasifika," or "Pasefika"—as opposed to "PI"—and consequently used these terms in their outreach materials. Participants also recommended that individuals be given a small welcome gift once selected. Participants felt that a list of local

XSL•FO
RenderX

resources was insufficient in terms of offering support, and instead favored having program staff directly link participants to services (ie, a warm handoff).

During this design sprint, participants expressed concerns about people “misusing” the ABP funds. While we originally wanted to create the lowest documentation burden for applicants, participants in the second design sprint expressed a desire for stricter documentation requirements to ensure eligibility. These preferences mirrored those expressed in the needs assessment interviews and the first design sprint. In addition, participants discussed what would happen in the case of a pregnancy loss.

Participants also weighed in on their willingness to participate in research to evaluate the impact of ABP. They noted the factors that would make them more likely to participate in the research, including transparency in how the results would be used and knowing that the data collection would be conducted by Black and Pacific Islander community members.

Final ABP Program Design

Overview

Through steps 1-4 of the formative community-centered research, we established the final program design of the ABP (Textbox 1). As a guaranteed income program, the income supplements were unconditional, which meant there were no requirements to receive or restrictions on the use of the funds. Recipients could use the funds in a manner that best supported their unique needs during pregnancy and postpartum.

Textbox 1. Final San Francisco Abundant Birth Project pilot elements, including the amount, duration, and disbursement of the income supplements, program eligibility criteria and assessment, participant selection, and additional support that resulted from the design process.

Program feature and selected design

Amount and duration

- US $1000 per month for 12 months

Disbursement

- Direct debit card that could be used for cash or banking

Eligibility

- Identify as Black and/or Pacific Islander
- Living in San Francisco or recently displaced
- <27 weeks pregnant
- Annual household income of <US $100,000

Assessment of eligibility criteria

- Self-reported income
- Proof of residence
- Provider verification of pregnancy

Selection

- Random selection through the bimonthly abundance drawing
- Participants were re-entered into the drawing until they reached the third trimester

Additional support: abundance coaching

- Opt-in abundance coaching to support housing needs and connections to additional resources.

Eligibility and Verification

Eligibility criteria listed in Textbox 1 were assessed using a web-based interest form and confirmed in an intake interview that occurred over Zoom (Zoom Video Communications). Applicants were required to provide verification of pregnancy and documentation of residence in San Francisco to ensure the program served its prioritized population. The allowable documents to confirm residence included a California-issued ID, school documents, employer pay stubs, or utility bill statements. Gestational weeks of pregnancy were verified by a medical provider. The documentation requirements reflected the intention to strike a balance between honoring community voices and not creating overly burdensome eligibility requirements. Household income was self-reported and, to make eligibility assessment simpler for program staff, not adjusted by household size. The maximum income requirements were purposely selected to allow low- and middle-income families to participate, given the high cost of living in San Francisco and that many families may not qualify for other governmental benefits but struggle to meet basic needs. On the interest form, applicants were able to self-identify as Black and/or Pacific Islander, the communities most impacted by preterm birth in San Francisco.

XSL•FO
RenderX

Eligibility Changes Related to Pregnancy Outcomes

The design processes helped inform what would happen if a participant's eligibility changed over the course of participation, particularly as it related to pregnancy outcomes. Using the expert opinion protocol, the ABP working group decided that in the event of a miscarriage, pregnancy termination, adoption, or death of the infant, ABP participants would continue to receive the income supplement for a 3-month grace period.

Implementing Randomized Selection

Bimonthly randomized drawings, called the "abundance drawing," ensured that all eligible individuals had an equal chance of being selected for ABP within each drawing. If applicants were not selected into ABP, they remained in the eligibility pool until they were selected or entered into the third trimester of pregnancy and therefore became ineligible.

Additional Support: Abundance Coaching

ABP offered optional case management services called "abundance coaching." In the needs assessment interviews, participants expressed a desire for more Black and Pacific Islander representation among staff from the local organizations that provided them with support. They also expressed previous stigmatizing experiences with government benefits programs. These perspectives informed the decision that all abundance coaches identify as Black and/or Pacific Islander to ensure racial and cultural concordance. Once enrolled, participants accessed optional support from abundance coaches who provided resource linkage and support with financial and other life goals. This included referrals to culturally specific community resources and providers. Unlike traditional case management which is often deficit-focused and paternalistic by directing and managing goals for participants, the abundance coaches provided individualized care and support based on the unique goals ABP participants set for themselves.

Evaluation

Using foundational strategies of equitable evaluation [49], we developed evaluation principles to abide by (1) leading with justice, (2) recognizing harm, (3) respecting autonomy, and (4) honoring community voice. We developed these principles based on ongoing work with the CRs and the responses from the interview and design sprint participants. The CRs, Expecting Justice collaborators, and research participants reported that too often Black and Pacific Islander people are excluded from decision-making in research and that there is a lack of transparency surrounding how data are used. The evaluation was opt-in for all ABP participants, based on the belief that the supplement should be truly unconditional, including research participation. The formative research also informed the decision that only CRs and other research team members who identified as Black and/or Pacific Islander would conduct surveys and interviews with participants.

Discussion

Principal Findings

Through a community-centered design process, we designed a novel guaranteed income program to center the needs of pregnant Black and Pacific Islander women and people in San Francisco. To our knowledge, no previous studies have qualitatively explored the financial, social services, and pregnancy and parenting experiences of these communities in San Francisco. Our needs assessment interviews revealed significant financial and life stressors experienced by Black and Pacific Islander people during and after pregnancy. Housing insecurity and previous poor experiences with social service programs were underscored by multiple participants. These findings are consistent with a report documenting that increases in San Francisco housing prices led to new concentrations of poverty especially for Black families from 2000 to 2015 [50]. Our findings also align with previous studies that report disproportionate unaffordable housing costs among people of color [51,52]. Additionally, the high cost of living in San Francisco left participants far above the federal poverty line ineligible for social services yet struggling to make ends meet. While all participants were enthusiastic about the potential of a guaranteed income program, they expressed concerns surrounding the eligibility and additional support needed to navigate the benefits system.

These experiences served as the foundation of the ABP prototypes and final program design. This formative process allowed for the establishment of eligibility and verification, supplement amount, disbursement options, and optional case management supports that were directly aligned with community members' expertise and preferences.

Strengths and Limitations

There are several strengths to this community-centered model of formative research and program development. Guaranteed income is a novel public health intervention, and the ABP program contributed to a swell of guaranteed income pilots in California and across the country [53-56]. Integrating community voice and expertise throughout the design process tailored ABP to the needs of the pregnant women and people it aimed to serve. We note that formative research with the CR model and shared decision-making requires more time and resources in the design phase, and institutions and academic hierarchies often do not allow space for this community-engaged work. Nonetheless, we strongly believe that community-centered formative research and program design ensures authentic engagement and leads to more relevant, responsive, and just programs and interventions. Our process can provide a roadmap for other programs to guide community co-design of an intervention.

Our results should also be considered in light of some limitations. Due to the COVID-19 pandemic, the needs assessment interviews and design sprints were implemented through videoconferencing only, and this mode of data collection may have affected the way participants engaged in the sessions. Participants' experiences of pregnancy and parenting were also likely changed by the pandemic. Additionally, eligibility for the interviews and design sprints included the ability to speak and read English, so these findings may not capture the experiences of non–English-speaking Black and Pacific Islander pregnant people in San Francisco, particularly immigrants.

XSL•FO
RenderX

Conclusions

The model of formative research to design ABP may be useful to other programs centering on community-driven, equity-oriented approaches. Interventions intended to address health equity should include experts with lived experience in the design process. Too often, well-intentioned programs do not meet the needs of recipients or center equity because of a lack of community participation in all elements of design.

The inclusion of CRs as integral members of the ABP staff was key to maximizing community involvement and shared decision-making. Ongoing staff and ABP working group meetings had designated spaces for CRs and community experts to advise ABP program implementation and evaluation design. Using the expert opinion protocol allowed for any ongoing issue to be decided with a consensus that prioritizes lived experience in developing solutions.

ABP aims to directly address the historical disinvestment of Black and Pacific Islander communities in San Francisco and positively impact birth outcomes in these communities through the provision of an unconditional income supplement. The COVID-19 pandemic has disproportionately impacted Black families, further underscoring the vital need for interventions that address structural deficiencies [57]. ABP applied a community-centered approach to design an intervention focused on the communities most impacted by maternal and birth disparities to reduce the effects of the racial wealth gap and economic insecurity among Black women. This novel approach and design process has great promise to reduce inequities in birth outcomes and center health and racial justice in antipoverty programs.

Acknowledgments

DK, JCW, MAT, MMDLC, SB, EC, MAC, AC, LL-G, BP, SS, ZM, and AMG contributed to the conceptualization, methodology, and implementation of the research. DK, JCW, MAT, MMDLC, SB, EC, MAC, AC, BP, SS, and AMG contributed to the analysis. DK and JCW contributed to the first draft of this paper. All authors contributed to the review and editing of subsequent drafts and reviewed the final paper. This research was supported by the March of Dimes Clinical and Social Science Challenge Grant, the California Preterm Birth Initiative funded by Marc and Lynne Benioff, and the UCSF-Kaiser Department of Research Building Interdisciplinary Research Careers in Women's Health Program (BIRCWH) funded by the National Institute of Child Health and Human Development (NICHD) and the Office of Research on Women's Health (ORWH; K12 HD052163. The contents of this paper are solely the responsibility of the authors and do not necessarily represent the official view of the funders.

Data Availability

Data are not available due to their sensitive nature and the risk of deductive disclosure.

Conflicts of Interest

None declared.

References

1. Almeida J, Bécares L, Erbetta K, Bettegowda VR, Ahluwalia IB. Racial/ethnic inequities in low birth weight and preterm birth: the role of multiple forms of stress. Matern Child Health J 2018;22(8):1154-1163 [FREE Full text] [doi: 10.1007/s10995-018-2500-7] [Medline: 29442278]
2. Behrman RE, Butler AS, editors. Institute of Medicine Committee on Understanding Premature Birth and Assuring Healthy Outcomes. US: National Academies Press; 2007.
3. Martin JA, Hamilton BE, Osterman MJK, Driscoll AK. Births: final data for 2020. Natl Vital Stat Rep 2021;70(2):1-51 [FREE Full text] [doi: 10.15620/cdc:100472] [Medline: 35157571]
4. Matthews TJ, MacDorman MF, Thoma ME. Infant mortality statistics from the 2013 period linked birth/infant death data set. Natl Vital Stat Rep 2015;64(9):1-30 [FREE Full text] [Medline: 26270610]
5. Petersen EE. MMWR. morbidity and mortality weekly report; vital signs: pregnancy-related deaths, United States, 2011-2015, and strategies for prevention, 13 States, 2013-2017. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/mmwr/volumes/68/wr/mm6818e1.htm [accessed 2019-09-02]
6. Sarnquist CC, Grieb EM, Maldonado YA. How racial and ethnic groupings may mask disparities: the importance of separating Pacific Islanders from Asians in prenatal care data. Matern Child Health J 2010;14(4):635-641. [doi: 10.1007/s10995-009-0494-x] [Medline: 19582560]
7. Rao AK, Daniels K, El-Sayed YY, Moshesh MK, Caughey AB. Perinatal outcomes among Asian American and Pacific Islander women. Am J Obstet Gynecol 2006;195(3):834-838. [doi: 10.1016/j.ajog.2006.06.079] [Medline: 16949421]
8. Singh GK, Yu SM. Pregnancy outcomes among Asian Americans. Asian Am Pac Isl J Health 1993;1(1):63-78. [Medline: 11567243]
9. Maternal, child, and adolescent health; SF preterm birth 5-year dashboard by social determinants 2018. San Francisco Department of Public Health. URL: https://www.sf.gov/departments/department-public-health/maternal-child-and-adolescent-health [accessed 2018-04-11]

XSL•FO
RenderX

10. California Department of Public Health. Preterm birth. Maternal, child, and adolescent health division. California Department of Public Health. URL: https://www.cdph.ca.gov/Programs/CFH/DMCAH/surveillance [accessed 2024-09-05]
11. Bailey ZD, Feldman JM, Bassett MT. How structural racism works - racist policies as a root cause of U.S. racial health inequities. N Engl J Med 2021;384(8):768-773. [doi: 10.1056/NEJMms2025396] [Medline: 33326717]
12. Jones CP. Levels of racism: a theoretic framework and a gardener's tale. Am J Public Health 2000;90(8):1212-1215. [doi: 10.2105/ajph.90.8.1212] [Medline: 10936998]
13. Lu MC, Halfon N. Racial and ethnic disparities in birth outcomes: a life-course perspective. Matern Child Health J 2003;7(1):13-30. [doi: 10.1023/a:1022537516969] [Medline: 12710797]
14. Hughes D, Simpson L. The role of social change in preventing low birth weight. Future Child 1995;5(1):87-102. [Medline: 7633870]
15. Blackwell SC, Gyamfi-Bannerman C, Biggio JR, Chauhan SP, Hughes BL, Louis JM, et al. 17-OHPC to prevent recurrent preterm birth in singleton gestations (PROLONG Study): a multicenter, international, randomized double-blind trial. Am J Perinatol 2020;37(2):127-136 [FREE Full text] [doi: 10.1055/s-0039-3400227] [Medline: 31652479]
16. Bailey ZD, Krieger N, Agénor M, Graves J, Linos N, Bassett MT. Structural racism and health inequities in the USA: evidence and interventions. Lancet 2017;389(10077):1453-1463. [doi: 10.1016/S0140-6736(17)30569-X] [Medline: 28402827]
17. Braveman P, Gottlieb L. The social determinants of health: it's time to consider the causes of the causes. Public Health Rep 2014;129(Suppl 2):19-31 [FREE Full text] [doi: 10.1177/00333549141291S206] [Medline: 24385661]
18. Ford CL, Airhihenbuwa CO. Critical race theory, race equity, and public health: toward antiracism praxis. Am J Public Health 2010;100(S1):S30-S35. [doi: 10.2105/ajph.2009.171058]
19. Malawa Z, Gaarde J, Spellen S. Racism as a root cause approach: a new framework. Pediatrics 2021;147(1):e2020015602. [doi: 10.1542/peds.2020-015602] [Medline: 33386339]
20. Wallace ME, Mendola P, Liu D, Grantz KL. Joint effects of structural racism and income inequality on small-for-gestational-age birth. Am J Public Health 2015;105(8):1681-1688. [doi: 10.2105/AJPH.2015.302613] [Medline: 26066964]
21. Hillier AE. Redlining and the home owners' Loan corporation. J Urban History 2003;29(4):394-420. [doi: 10.1177/0096144203029004002]
22. Stanczyk AB. The dynamics of household economic circumstances around a birth. Equitable Growth. 2016. URL: https://equitablegrowth.org/working-papers/income-volatility-around-birth/ [accessed 2019-11-18]
23. Taylor K, Compton S, Kolenic GE, Scott J, Becker N, Dalton VK, et al. Financial hardship among pregnant and postpartum women in the United States, 2013 to 2018. JAMA Netw Open 2021;4(10):e2132103 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.32103] [Medline: 34714338]
24. American Community Survey 2016-2020 5-year data release. US Census Bureau. URL: https://www.census.gov/newsroom/press-kits/2021/acs-5-year.html [accessed 2023-03-07]
25. Kendall M. Income inequality in the Bay Area is among nation's highest. The Mercury News. 2018. URL: https://www.mercurynews.com/2018/02/15/income-inequality-in-the-bay-area-is-among-nations-highest/ [accessed 2018-02-15]
26. Family needs calculator. Insight Center. URL: https://insightcced.org/family-needs-calculator/ [accessed 2022-05-25]
27. Guaranteed income: a primer for funders. Asset Funders Network, Center for High Impact Philanthropy, Economic Security Project, Springboard to Opportunities. URL: https://economicsecurityproject.org/wp-content/uploads/2022/10/220516-Funder-Primer.pdf [accessed 2022-05-16]
28. Guaranteed basic income projects. California Department of Social Services. URL: https://www.cdss.ca.gov/inforesources/guaranteed-basic-income-projects [accessed 2022-03-04]
29. West S, Castro BA, Samra S, Coltrera E, Addo M, Carlson M, et al. Preliminary analysis: SEED?s first year. Stockton Economic Empowerment Demonstration 2023;11(4):25 [FREE Full text]
30. Magnolia Mother's Trust evaluation report. Springboard to Opportunity. 2020. URL: https://springboardto.org/wp-content/uploads/2024/09/STO_MMT-2nd-Cohort-2020-Evaluation-Summary-1.pdf [accessed 2022-03-04]
31. Brownell MD, Nickel NC, Chartier M, Enns JE, Chateau D, Sarkar J, et al. An unconditional prenatal income supplement reduces population inequities in birth outcomes. Health Aff (Millwood) 2018;37(3):447-455. [doi: 10.1377/hlthaff.2017.1290] [Medline: 29505359]
32. Brownell MD, Chartier MJ, Nickel NC, Chateau D, Martens PJ, Sarkar J, et al. Unconditional prenatal income supplement and birth outcomes. Pediatrics 2016;137(6):e20152992. [doi: 10.1542/peds.2015-2992] [Medline: 27244846]
33. Kuang J. More than 12,000 Californians are getting cash from guaranteed income experiments. CalMatters. URL: http://calmatters.org/california-divide/2023/02/universal-basic-income/ [accessed 2024-08-27]
34. La L. Guaranteed basic income programs proliferate across California. CalMatters. URL: http://calmatters.org/newsletter/california-guaranteed-income-programs/ [accessed 2024-08-27]
35. Kania J, Kramer M. Collective impact. China: FSG Beijing; 2011.
36. National Birth Equity Collaborative (NBEC) - birth equity for all black birthing people. National Birth Equity Collaborative. URL: https://birthequity.org/ [accessed 2023-07-11]

XSL•FO
RenderX

37. Social determinants of health. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/about/priorities/why-is-addressing-sdoh-important.html?CDC_AAref_Val=https://www.cdc.gov/about/sdoh/index.html [accessed 2023-07-12]
38. Crear-Perry J, Correa-de-Araujo R, Lewis Johnson T, McLemore MR, Neilson E, Wallace M. Social and structural determinants of health inequities in maternal health. J Womens Health (Larchmt) 2021;30(2):230-235 [FREE Full text] [doi: 10.1089/jwh.2020.8882] [Medline: 33181043]
39. Ortiz K, Nash J, Shea L, Oetzel J, Garoutte J, Sanchez-Youngman S, et al. Partnerships, processes, and outcomes: a health equity-focused scoping meta-review of community-engaged scholarship. Annu Rev Public Health 2020;41(1):177-199 [FREE Full text] [doi: 10.1146/annurev-publhealth-040119-094220] [Medline: 31922931]
40. Boland SE, Homdayjanakul KJ, Reed ND, Wesner CA, Basha E, Kaufman CE, et al. Addressing the cycle of inaction: a DrPH student perspective on the decolonization of public health. HPHR. 2021. URL: https://bcphr.org/35-article-boland/ [accessed 2023-06-20]
41. Getrich CM, Sussman AL, Campbell-Voytal K, Tsoh JY, Williams RL, Brown AE, et al. Cultivating a cycle of trust with diverse communities in practice-based research: a report from PRIME Net. Ann Fam Med 2013;11(6):550-558 [FREE Full text] [doi: 10.1370/afm.1543] [Medline: 24218379]
42. Isler MR, Corbie-Smith G. Practical steps to community engaged research: from inputs to outcomes. J Law Med Ethics 2021;40(4):904-914. [doi: 10.1111/j.1748-720x.2012.00719.x]
43. Hamilton AB. Qualitative methods in rapid turn-around health services research. U.S. Department of Veterans Affairs. 2013. URL: https://www.hsrd.research.va.gov/for_researchers/cyber_seminars/archives/780-notes.pdf [accessed 2021-06-23]
44. Brown T. Design thinking. Harv Bus Rev 2008;86(6):84-92, 141. [Medline: 18605031]
45. Jake-Schoffman DE, McVay MA. Using the design sprint process to enhance and accelerate behavioral medicine progress: a case study and guidance. Transl Behav Med 2021;11(5):1099-1106. [doi: 10.1093/tbm/ibaa100] [Medline: 33057685]
46. Nijagal MA, Patel D, Lyles C, Liao J, Chehab L, Williams S, et al. Using human centered design to identify opportunities for reducing inequities in perinatal care. BMC Health Serv Res 2021;21(1):714 [FREE Full text] [doi: 10.1186/s12913-021-06609-8] [Medline: 34284758]
47. Hager ER, Quigg AM, Black MM, Coleman SM, Heeren T, Rose-Jacobs R, et al. Development and validity of a 2-item screen to identify families at risk for food insecurity. Pediatrics 2010;126(1):e26-e32. [doi: 10.1542/peds.2009-3146] [Medline: 20595453]
48. Floyd I, Pavetti L, Meyer L, Safawi A, Schott L, Bellew E, et al. TANF policies reflect racist legacy of cash assistance. Center on budget and policy priorities. URL: https://www.cbpp.org/research/family-income-support/tanf-policies-reflect-racist-legacy-of-cash-assistance [accessed 2022-06-24]
49. Equitable evaluation framework framing paper. San Rafael, CA: equitable evaluation initiative. Center for Equitable Evaluation, Institute for Foundation and Donor Learning, Luminaire Group, Dorothy A Johnson Center for Philanthropy. 2017. URL: https://sdbjrfoundation.issuelab.org/resource/equitable-evaluation-framework-framing-paper.html [accessed 2017-06-01]
50. Verma P, Rinzler D, Zuk M. Rising housing costs and re-segregation in San Francisco. UC Berkeley's Urban Displacement Project And California Housing Partnership. URL: https://www.urbandisplacement.org/wp-content/uploads/2021/08/sf_final.pdf [accessed 2024-12-25]
51. Kimberlin S. California?s Housing Affordability Crisis Hits Renters and Households With the Lowest Incomes the Hardest. Sacramento, CA: California Budget & Policy Center; 2019.
52. Levin M. Black Californians' housing crisis, by the numbers. CalMatters. URL: http://calmatters.org/housing/2020/06/black-californians-housing-crisis-by-the-numbers/ [accessed 2024-10-15]
53. Philly Joy Bank. The Philadelphia City Fund. URL: https://philacityfund.org/programs/philly-joy-bank/ [accessed 2004-11-09]
54. Our work. The Bridge Project. URL: https://bridgeproject.org/our-work/ [accessed 2024-11-09]
55. The nest (guaranteed income). Hummingbird Indigenous Family Services. URL: https://www.hummingbird-ifs.org/programs/nest/the-nest [accessed 2024-11-09]
56. About. Black Economic Equity Movement (BEEM). URL: https://beemproject.org/about/ [accessed 2024-11-09]
57. Lopez MH, Rainie L, Budiman A. Financial and health impacts of COVID-19 vary widely by race and ethnicity. Pew Research Center. URL: https://www.pewresearch.org/fact-tank/2020/05/05/financial-and-health-impacts-of-covid-19-vary-widely-by-race-and-ethnicity/ [accessed 2020-05-05]

Abbreviations

ABP: Abundant Birth Project
CR: community researcher
RAP: Rapid Assessment Process

XSL•FO
RenderX

Edited by A Mavragani; submitted 05.06.24; peer-reviewed by KEC Massengale, EC Rhodes, X Yu; comments to author 13.08.24; revised version received 10.11.24; accepted 06.12.24; published 27.01.25.

Please cite as:
Karasek D, Williams JC, Taylor MA, De La Cruz MM, Arteaga S, Bell S, Castillo E, Chand MA, Coats A, Hubbard EM, Love-Goodlett L, Powell B, Spellen S, Malawa Z, Gomez AM
Designing the First Pregnancy Guaranteed Income Program in the United States: Qualitative Needs Assessment and Human-Centered Design to Develop the Abundant Birth Project
JMIR Form Res 2025;9:e60829
URL: https://formative.jmir.org/2025/1/e60829
doi:10.2196/60829
PMID:

©Deborah Karasek, Jazzmin C Williams, Michaela A Taylor, Monica M De La Cruz, Stephanie Arteaga, Sabra Bell, Esperanza Castillo, Maile A Chand, Anjeanette Coats, Erin M Hubbard, Latriece Love-Goodlett, Breezy Powell, Solaire Spellen, Zea Malawa, Anu Manchikanti Gomez. Originally published in JMIR Formative Research (https://formative.jmir.org), 27.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Tobacco Use, Risk Perceptions, and Characteristics of Adults Who Used a Heated Tobacco Product (IQOS) in the United States: Cross-Sectional Survey Study

Hui G Cheng[1], PhD; Brendan Noggle[1], MPH; Andrea R Vansickel[1], PhD; Edward G Largo[1], PhD; Pierpaolo Magnani[2]

[1]Altria Client Services LLC, Richmond, VA, United States

[2]PMI R&D, Philip Morris Products S.A., Neuchâtel, Switzerland

Corresponding Author:
Andrea R Vansickel, PhD
Altria Client Services LLC
601 East Jackson Street
Richmond, VA, 23219
United States
Phone: 1 8595765695
Email: Andrea.R.Vansickel@altria.com

Abstract

Background: The Tobacco Heating System (THS; commercialized as IQOS) is a smoke-free heated tobacco product introduced in the United States in 2019 and authorized by the US Food and Drug Administration as a modified risk tobacco product (MRTP) in 2020. THS consists of a holder and specially designed tobacco sticks that are heated instead of burned to produce a nicotine-containing aerosol. THS was available in Atlanta, Georgia; Richmond, Virginia; Charlotte, North Carolina; the Northern Virginia region; and South Carolina before its market removal in November 2021.

Objective: This study aims to describe selected sociodemographic characteristics and self-reported health history of adults who used IQOS (AUIs), their tobacco use patterns (eg, tobacco use history, exclusive and dual-use, and switching from cigarette smoking), their risk perceptions of the product, and their understanding of MRTP messages.

Methods: The IQOS Cross-Sectional Postmarket Adult Consumer Study was a study of AUIs aged 21 years or older who were recruited from a consumer database via direct postal mail and emails. Participants completed the online survey between September and November 2021.

Results: The survey was completed by 645 current and 43 former AUIs who had used at least 100 tobacco sticks (considered established THS use) before the assessment. Of the 688 participants, 424 (61.6%) were male, 502 (73.0%) were non–Hispanic White, and the mean age was 45 years. The vast majority (680/688, 98.8%) of AUIs had ever smoked combusted cigarettes before first trying THS and 628 (91.3%) had smoked cigarettes in the 30 days before first using THS. At the time of assessment, 161 (23.4%) reported using e-cigarettes (vs 229, 33.3%, before THS use), 92 (13.4%) reported smoking cigars (vs 114, 16.6%, before THS use), and 338 (49.1%) were still smoking after an average of 1 year of THS use. Among those currently using THS who were still smoking (n=298), 249 (83.6%) smoked fewer cigarettes compared with before first trying THS; 362 of 688 (52.6%) AUIs reported having no physical health conditions evaluated in this study and almost three-quarters reported having no mental health conditions. Among all AUIs, over 563 (81.8%) had never used a cessation treatment or had not used it in the past 12 months, and 555 (80.7%) AUIs demonstrated a correct understanding of the MRTP message and AUIs perceived THS as having a lower risk than cigarettes (43.8 vs 64.4 on a 100-point composite score scale).

Conclusions: This study provides evidence that THS can help adult smokers in the United States completely switch away from cigarettes or reduce smoking.

(JMIR Form Res 2025;9:e57398) doi:10.2196/57398

KEYWORDS

United States Food and Drug Administration; FDA; IQOS; Tobacco Heating System; THS; heated tobacco products; modified risk tobacco products; MRTP; tobacco; nicotine; smokers; tobacco harm reduction; cross-sectional surveys; cigarettes

XSL•FO
RenderX

Introduction

Cigarette smoking is a leading contributor to mortality and morbidity worldwide [1]. In the United States, smoking contributes to an estimated 500,000 premature deaths and an expenditure exceeding US $170 billion in annual medical costs to treat smoking-related lung and cardiovascular illnesses [2]. Although quitting is the best way to mitigate health risks from smoking, it remains challenging for many smokers. Despite 70% of smokers wanting to quit smoking and more than half attempting each year [3], the majority of those who attempt to quit relapse within 12 months [4-7]. Individuals who smoke make many attempts before successfully quitting, rendering the quitting journey a protracted relapsing process [7,8]. Given that health risks from cigarette smoking are mainly attributed to harmful constituents generated through combustion, smokers who are unable or unwilling to quit may benefit from switching to smoke-free products such as heated tobacco products including the Tobacco Heating System (THS; commercialized as *IQOS* by Philip Morris Products S.A.) [9-12].

THS is a smoke-free tobacco product that heats specially designed tobacco sticks (commercialized as *HeatSticks* by Philip Morris Products S.A.) to temperatures below the level of combustion and produces an aerosol with substantially reduced (ie, 90%-95% lower, on average) levels of harmful and potentially harmful constituents compared with cigarette smoke while delivering similar levels of nicotine [13-15]. The product debuted in 2014 in Japan and was available in more than 80 countries as of September 2024.

Evidence collected to date supports the role of THS in tobacco harm reduction. The product was rapidly adopted by individuals who smoked in Japan [16], and recent data suggest that the majority of individuals who used THS have switched to exclusive use and stopped smoking [17]. Moreover, individuals who used heated tobacco products exclusively, as opposed to smoking, had more favorable health-related biomarker profiles compared with smokers [18]. In a prospective study of 38 Italian smokers with chronic obstructive pulmonary disease who started to use THS, 58% stopped smoking within 3 years, and those who were still smoking significantly reduced their cigarette consumption (mean cigarette consumption dropped from 21 per day at baseline to 3.7 per day at the last follow-up). Most importantly, consistent improvements were observed in respiratory symptoms, exercise tolerance, quality of life, and rate of disease exacerbations, whereas no such change was observed among individuals who smoked cigarettes during the same follow-up period [19]. These results suggest that switching from cigarette smoking to THS may be associated with long-term health benefits.

The US Food and Drug Administration (FDA) authorized the marketing of THS in the United States in April 2019. Subsequently, it granted a modified risk tobacco product (MRTP) order in July 2020 authorizing the use of the following reduced exposure claim:

"AVAILABLE EVIDENCE TO DATE:

- The IQOS system heats tobacco but does not burn it.
- This significantly reduces the production of harmful and potentially harmful chemicals.
- Scientific studies have shown that switching completely from conventional cigarettes to the IQOS system significantly reduces your body's exposure to harmful or potentially harmful chemicals."

THS was first introduced to the United States in Atlanta (Georgia) in October 2019. Marketing expanded to Richmond (Virginia) in January 2020, Charlotte (North Carolina) in July 2020, and northern Virginia and South Carolina as of May 2021 before it was removed from all US markets in November 2021 as a result of a Cease-and-Desist Order (CDO) issued by the United States International Trade Commission (ITC). (The ITC issued a CDO prohibiting the importation, marketing, sale, and distribution of *IQOS* devices and *HeatSticks* on September 29, 2021. To comply with the CDO, Philip Morris USA stopped marketing and selling the products as of November 28, 2021, in all channels.) Three types of tobacco sticks were sold: amber/regular (nonmenthol), green/smooth menthol (menthol), and blue/fresh menthol (menthol). At the end of August 2021, the tobacco sticks were available in 4842 third-party retail outlets, and the THS devices were available in 1235 third-party retail outlets.

This study aimed to (1) characterize adults who used *IQOS* (AUIs) and their history and patterns of tobacco use, as well as tobacco cessation treatment use; (2) describe risk perceptions of THS and the understanding of the MRTP claim; and (3) describe complete switching from cigarette smoking to THS, transitions to/back to cigarette smoking, and relevant quitting behaviors among AUIs. We used data from the 2021 *IQOS* Cross-sectional Postmarket Adult Consumer Study to provide the first evidence about AUIs, how AUIs used the product, and the potential role of such products in tobacco harm reduction in real-world settings in the United States.

Methods

Study Design, Study Population, and Recruitment

This was a cross-sectional study. The study population consisted of adults, aged 21 years or older, who had used at least 100 tobacco sticks with THS before the assessment. Individuals were excluded if (1) they were unable to read, speak, or understand English; (2) they or their first-degree relative were current or former employees of the tobacco industry, the Central Research Organization, or any market research firm; or (3) they or their first-degree relative were currently involved in litigation against any tobacco industry entity.

In this study, we recruited AUIs from a consumer database, which covered the majority of individuals who purchased an *IQOS* device. The *IQOS* Consumer Database is a database of registered *IQOS* consumers in the United States. These consumers entered the database by voluntarily registering their devices either at the point of sale or later by themselves. Based on the number of devices registered and sold, we estimated that approximately 70% of *IQOS* consumers were included in the *IQOS* Consumer Database. We sent invitations by both direct mail and email (if available) to all individuals in the database

XSL•FO
RenderX

who agreed to be contacted for research purposes: 1 initial invitation and 1 reminder were sent to all potential participants. Figure 1 depicts the recruitment process and final sample. Data were collected between September 15 and November 14, 2021. In brief, of the 19,258 individuals invited, 2022 (10.50%) completed the screener; of the 698 eligible individuals who consented to participate, 688 (98.6%) completed the survey.

Figure 1. Flowchart depicting sample recruitment for the Cross-Sectional Postmarket Adult Consumer Study (2021). Participants sampled from the IQOS consumer database were primarily residing in Georgia, Virginia, North Carolina, or South Carolina, United States.

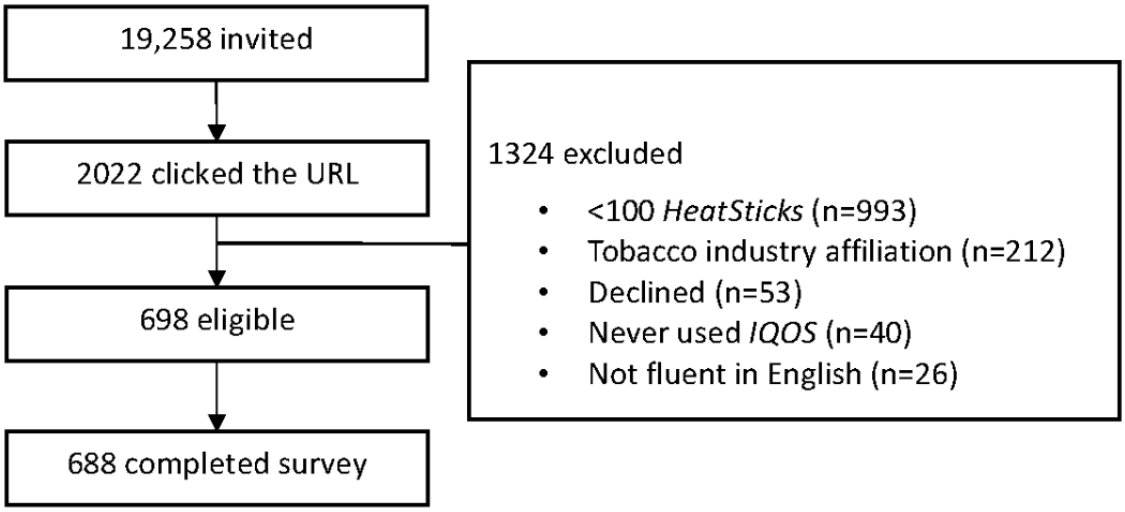

Ethical Considerations

The study was reviewed (study protocol, questionnaire, and informed consent statement) and approved by the Sterling Institutional Review Board (IRB ID number: 9102-HCheng). This study was conducted in compliance with the study protocol and, where applicable, in accordance with the Insights Association Code of Standards and Ethics for Market Research and Data Analytics [20] and the International Chamber of Commerce/European Society for Opinion and Marketing Research's (ESOMAR) International Code on Market, Opinion and Social Research [21]. All participants read the informed consent statement and voluntarily consented to participate in the study via electronic agreement. All data generated through this study were considered highly confidential by study staff and participants were identified only via unique participation identification numbers. Upon completion of the questionnaire, participants were provided instructions to obtain a US $40 electronic gift card.

Assessment and Definitions

All eligible participants, regardless of the mode of recruitment, self-administered the questionnaire online. The assessment consisted of modules of questions on the history of use of various tobacco products (including THS, cigarettes, cigars, pipe, hookah, e-cigarettes, smokeless tobacco, and oral tobacco-derived nicotine products). Additional modules included risk perceptions and understanding of claim-related information, demographic characteristics, and diagnoses of selected physical and mental conditions.

In this study, we defined ever use as a "yes" response to the question "Have you EVER used/smoked...EVEN ONE TIME?" Current use was defined as "every day" or "some days" responses to the question "Do you now use/smoke...every day, some days, or not at all?" Former use was defined as a "yes" response to the ever-use question and a "not at all" response to the current-use question. Established cigarette smoking and THS use were defined as having smoked at least 100 cigarettes or used at least 100 tobacco sticks in a lifetime. Menthol/nonmenthol preference was assessed using the question "Which *one* type of *HeatSticks* are you *CURRENTLY* using *most often*?" Responses of "green menthol" (previously "smooth menthol") and "blue menthol" (previously "fresh menthol") were classified as preferring menthol; responses of "amber" (previously "regular") were classified as preferring nonmenthol. We identified menthol/nonmenthol cigarette preference via questions about the type of cigarettes smoked most often.

In this study, we asked time-centered questions to establish a temporal sequence of smoking and THS use to assess changes in smoking behaviors after first trying THS. For example, we classified current smokers before first trying THS as individuals who answered "every day" or "some days" to the question "During the 30 days before you first tried *IQOS*, did you smoke cigarettes every day, some days, or not at all?". Using this approach, we defined individuals who had *completely switched* as current established AUIs, who (1) had smoked at least 100 cigarettes in their lifetime, (2) were smoking every day or some days in the 30 days before first trying THS, (3) were not smoking at the time of assessment, and (4) had smoked their last cigarette after trying THS (via the question "Was the last time you smoked cigarettes...?" with response options "Before trying *IQOS* for the first time" and "After trying *IQOS* for the first time").

Perceived health risks were assessed using the validated 18-item Perceived Health Risk scale, which has demonstrated good psychometric properties [22]. Two batteries of questions were asked for cigarette smoking and THS use, respectively. Each item was rated on a 5-point Likert-type scale ranging from 0 (no risk) to 4 (very high risk). An overall score was calculated by summing all items in the scale (possible range 0-72), with higher scores representing greater perceived health risk [22].

XSL•FO
RenderX

The understanding of the MRTP message was assessed using 2 questions shown in Textbox 1.

Textbox 1. Questions to assess understanding of the modified risk tobacco product message.

- 1. Based on what you know or believe, please complete the following: Smokers who switch completely from cigarettes to IQOS:
  - Have more exposure to harmful or potentially harmful chemicals
  - Have the same exposure to harmful or potentially harmful chemicals
  - Have less exposure to harmful or potentially harmful chemicals (correct answer)
  - Have no exposure to harmful or potentially harmful chemicals
  - Don't know

IF "HAVE LESS EXPOSURE..." SELECTED

- 2. Based on what you know or believe, what do smokers need to do in order to reduce their body's exposure to harmful or potentially harmful chemicals?
  - Stop smoking cigarettes completely and only use IQOS (correct answer)
  - Smoke fewer cigarettes and also use IQOS
  - Keep smoking the same amount of cigarettes and also use IQOS
  - Don't know

Demographic characteristics included sex (male or female), age, race/ethnicity, level of education, household income, employment status, and marital status. A lifetime history of physical conditions was assessed using the question "Have you ever been told by a doctor, nurse, or other health professional that you had...?" Options included cardiovascular conditions, respiratory conditions, cancers, and diabetes (see Table 2 for details). A lifetime history of mental conditions was assessed using 2 questions: "A mental illness or disorder refers to a wide range of mental health conditions or disorders that affect your mood, thinking, or behavior. Has a doctor, nurse, or other health professional ever told you that you had a mental health condition such as depression, stress, or problems with emotions?" and "Are you now taking medicine or receiving treatment from a doctor, nurse, or other health professional for a mental health condition or emotional problem?" All information was based on self-report.

Analysis

Descriptive statistics were used to summarize outcome variables. For categorical variables, frequencies and proportions (95% CIs) were generated. For numeric variables (eg, age), means (95% CI) and medians (IQRs) were calculated. Estimates (eg, proportions and means) were presented for current and ever AUIs, respectively. Demographic characteristics and THS use behaviors were further stratified for AUIs who preferred menthol and nonmenthol tobacco sticks. Estimates are not shown for former established AUIs due to imprecision as a result of the small sample size (n=43). Missing values were kept as missing except for missing values from logic skips in which case, values were assigned based on the logic skip. For example, if a participant indicated that they had never used an ENDS (electronic nicotine delivery systems) product, then the current ENDS use question would be skipped and a "no" response value assigned. Data analyses were conducted using SAS version 9.4 (SAS Institute Inc.). The protocol for this study was published previously [23]. Data are available for interested researchers. Please contact the corresponding author for details.

Results

Characteristics of Study Participants

Of the 688 established AUIs, 273 (39.7%) resided in Georgia, 179 (26.0%) resided in North Carolina, 177 (25.7%) resided in Virginia, and 45 (6.5%) resided in South Carolina. Nominal numbers (n<5) of participants resided in other states. Of the total 688 participants, 424 (61.6%) were males, and the mean age was 45 years, with the majority between 35 and 54 years of age. On average, current established AUIs who preferred menthol tobacco sticks were younger than those who preferred nonmenthol tobacco sticks (mean age 43.1 years, 95% CI 42.0-44.3 vs 47.3 years, 95% CI 46.1-48.6). Of all established AUIs who had ever smoked cigarettes (n=680), 279 (41.0%; 95% CI 36.9%-44.3%) smoked menthol cigarettes most often. Individuals who were non–Hispanic White accounted for the majority of established AUIs (502/688, 73%). Differences in tobacco stick varieties were observed across race/ethnicity groups; greater proportions of non–Hispanic Whites preferred nonmenthol tobacco sticks, whereas greater proportions of non–Hispanic Black and non–Hispanic Asians preferred menthol varieties. Approximately two-thirds of established AUIs had an annual family income greater than US $60,000, 556 (80.8%) had at least some college education, and 547 (79.5%) were in the workforce. Table 1 provides more detailed information about the demographic characteristics of the sample.

XSL•FO
RenderX

Table 1. Demographic characteristics of established AUIs[a,b].

Characteristics	Ever-established AUIs (n=688)	Current established AUIs (n=645)	Current established AUIs who prefer menthol tobacco sticks (n=337)	Current established AUIs who prefer nonmenthol tobacco sticks (n=8)
Sex, n (%); 95% CI				
Male	424 (61.6); 58 to 65.3	394 (61.1); 57.3 to 64.9	205 (61.1); 55.9 to 66.3	188 (61.0); 55.6 to 66.5
Age (years)				
Mean (95% CI)	45.0 (44.1 to 45.9)	45.1 (44.3 to 46)	43.1 (42 to 44.3)	47.3 (46.1 to 48.6)
Median (IQR)	43.5 (36-54)	44 (36-54)	41 (35-50)	47 (39 to 55.5)
21-24, n (%); 95% CI	14 (2.0); 1.0 to 3.1	13 (2.0); 0.9 to 3.1	6 (1.8); 0.7 to 3.8	7 (2.3); 0.9 to 4.6
25-34, n (%); 95% CI	121 (17.6); 14.7 to 20.4	109 (16.9); 14 to 19.8	76 (22.6); 18.1 to 27	33 (10.7); 7.3 to 14.2
35-44, n (%); 95% CI	222 (32.3); 28.8 to 35.8	209 (32.4); 28.8 to 36	121 (35.9); 30.8 to 41	88 (28.6); 23.5 to 33.6
45-54, n (%); 95% CI	178 (25.9); 22.6 to 29.1	172 (26.7); 23.3 to 30.1	81 (24.0) 19.5 to 28.6	91 (29.6); 24.5 to 34.6
55-64, n (%); 95% CI	112 (16.3); 13.5 to 19	103 (16); 13.1 to 18.8	37 (10.9); 7.6 to 14.3	66 (21.4); 16.9 to 26
65+, n (%); 95% CI	41 (6.0); 4.2 to 7.7	39 (6.1); 4.2 to 7.9	16 (4.8); 2.5 to 7	23 (7.5); 4.5 to 10.4
Race/ethnicity, n (%); 95% CI				
Hispanic/Latino	40 (5.8); 4.1 to 7.6	39 (6.1); 4.2 to 7.9	20 (5.9); 3.4 to 8.5	19 (6.2); 3.5 to 8.9
Non–Hispanic White	502 (73.0); 70.0 to 76.3	463 (71.8); 68.3 to 75.3	212 (62.9); 57.8 to 68.1	251 (81.5); 77.2 to 85.8
Non–Hispanic Black	38 (5.5); 3.8 to 7.2	37 (5.7); 3.9 to 7.5	28 (8.3); 5.4 to 11.3	9 (2.9); 1.3 to 5.5
Non–Hispanic Asian	91 (13.2); 10.7 to 15.8	90 (13.9); 11.3 to 16.6	68 (20.2); 15.9 to 24.5	22 (7.1); 4.3 to 10.0
Non–Hispanic Native Hawaiian or other Pacific Islander	0 (0); <0.1 to 0.5	0 (0); <0.1 to 0.6	0 (0); <0.1 to 1.1	0 (0); <0.1 to 1.2
Non–Hispanic American Indian or Alaska Native	8 (1.2); 0.5 to 2.3	8 (1.2); 0.5 to 2.4	6 (1.8); 0.7 to 3.8	2 (0.7); 0.1 to 2.3
Non–Hispanic Other	9 (1.3); 0.6 to 2.5	8 (1.2); 0.5 to 2.4	3 (0.9); 0.2 to 2.6	5 (1.6); 0.5 to 3.8
Household income (US $), n (%); 95% CI				
<60,000	231 (33.6); 30.1 to 37.1	212 (32.9); 29.2 to 36.5	120 (35.6); 30.5 to 40.7	92 (29.9).8 to 35
60,000-74,999	67 (9.7); 7.5 to 12.0	63 (9.8); 7.5 to 12.1	35 (10.4); 7.1 to 13.6	28 (9.1); 5.9 to 12.3
75,000-99,999	124 (18); 15.2 to 20.9	118 (18.3); 15.3 to 21.3	59 (17.5); 13.5 to 21.6	59 (19.2); 14.8 to 23.6
100,000-149,999	125 (18.2); 15.3 to 21.1	117 (18.1); 15.2 to 21.1	65 (19.3); 15.1 to 23.5	52 (16.9); 12.7 to 21.1
≥150,000	95 (13.8); 11.2 to 16.4	90 (14); 11.3 to 16.6	35 (10.4); 7.1 to 13.6	55 (17.9); 13.6 to 22.1
Prefer not to answer	46 (6.7); 4.8 to 8.6	45 (7.0); 5.0 to 8.9	23 (6.8); 4.1 to 9.5	22 (7.1); 4.3 to 10.0
Education, n (%); 95% CI				
High school or less	132 (19.2); 16.2 to 22.1	125 (19.4); 16.3 to 22.4	62 (18.4); 14.3 to 22.5	63 (20.5); 16.0 to 25.0

XSL•FO
RenderX

Characteristics	Ever-established AUIs (n=688)	Current established AUIs (n=645)	Current established AUIs who prefer menthol tobacco sticks (n=337)	Current established AUIs who prefer nonmenthol tobacco sticks (n=8)
Some college	196 (28.5); 25.1 to 31.9	175 (27.1); 23.7 to 30.6	89 (26.4); 21.7 to 31.1	86 (27.9); 22.9 to 32.9
Associate's degree	85 (12.4); 9.9 to 14.8	82 (12.7); 10.1 to 15.3	51 (15.1); 11.3 to 19.0	31 (10.1); 6.7 to 13.4
Bachelor's degree	179 (26.0); 22.7 to 29.3	171 (26.5); 23.1 to 29.9	93 (27.6); 22.8 to 32.4	78 (25.3); 20.5 to 30.2
Master's degree	76 (11.1); 8.7 to 13.4	73 (11.3); 8.9 to 13.8	32 (9.5); 6.4 to 12.6	41 (13.3); 9.5 to 17.1
Professional school degree	4 (0.6); 0.2 to 1.5	4 (0.6); 0.2 to 1.6	1 (0.3); <0.1 to 1.6	3 (1); 0.2 to 2.8
Doctorate	9 (1.3); 0.6 to 2.5	9 (1.4); 0.6 to 2.6	4 (1.2); 0.3 to 3	5 (1.6); 0.5 to 3.8
Other	7 (1); 0.4 to 2.1	6 (0.9); 0.3 to 2.0	5 (1.5); 0.5 to 3.4	1 (0.3); <0.1 to 1.8
Employment status, n (%); 95% CI				
Employed for wages	439 (63.8); 60.2 to 67.4	412 (63.9); 60.2 to 67.6	223 (66.2); 61.1 to 71.2	189 (61.4); 55.9 to 66.8
Self-employed	108 (15.7); 13.0 to 18.4	104 (16.1); 13.3 to 19.0	47 (14.0); 10.3 to 17.7	57 (18.5); 14.2 to 22.8
Not employed	141 (20.5); 17.5 to 23.5	129 (20.0); 16.9 to 23.1	67 (19.9); 15.6 to 24.1	62 (20.1); 15.7 to 24.6
Marital status, n (%); 95% CI				
Married	370 (53.8); 50.1 to 57.5	355 (55.0); 51.2 to 58.9	179 (53.1); 47.8 to 58.4	176 (57.1); 51.6 to 62.7
Widowed	8 (1.2); 0.5 to 2.3	8 (1.2); 0.5 to 2.4	4 (1.2); 0.3 to 3.0	4 (1.3); 0.4 to 3.3
Divorced	100 (14.5); 11.9 to 17.2	93 (14.4); 11.7 to 17.1	42 (12.5); 8.9 to 16.0	51 (16.6); 12.4 to 20.7
Separated	15 (2.2); 1.1 to 3.3	12 (1.9); 0.8 to 2.9	8 (2.4); 1.0 to 4.6	4 (1.3); 0.4 to 3.3
Never married	116 (16.9); 14.1 to 19.7	109 (16.9); 14.0 to 19.8	68 (20.2); 15.9 to 24.5	41 (13.3); 9.5 to 17.1
Living with partner	60 (8.7); 6.6 to 10.8	50 (7.8); 5.7 to 9.8	27 (8.0); 5.1 to 10.9	23 (7.5); 4.5 to 10.4
Do not wish to answer	19 (2.8); 1.5 to 4.0	18 (2.8); 1.5 to 4.1	9 (2.7); 1.2 to 5.0	9 (2.9); 1.3 to 5.5

[a]AUI: adult who used *IQOS*.

[b]Participants sampled from the *IQOS* consumer database were primarily residing in Georgia, Virginia, North Carolina, or South Carolina.

Just over half of AUIs reported they had never received any physical diagnoses included in this study (362/688, 52.6%; 95% CI 48.9%-56.4%; see Table 2 for more details). The most common conditions reported were hypertension (165/688, 24.0%; 95% CI 20.8%-27.2%) and high cholesterol (148/688, 21.5%; 95% CI 18.4%-24.6%). Approximately a quarter of participants reported they had ever had a mental health condition (181/688, 26.3%; 95% CI 23.0%-29.6%).

XSL•FO
RenderX

Table 2. Lifetime diagnosis of selected physical and mental health conditions among established AUIs[a,b].

Lifetime diagnosis	Ever-established AUIs (n=688), n (%); 95% CI	Current established AUIs (n=645), n (%); 95% CI
Physical condition		
A heart attack, also called myocardial infarction	10 (2.8); 1.5 to 4.0	17 (2.6); 1.4 to 3.9
Angina, also called angina pectoris (chest pain or discomfort)	12 (1.7); 0.8 to 2.7	11 (1.7); 0.7 to 2.7
Congestive heart failure	5 (0.7); 0.2 to 1.7	5 (0.8); 0.3 to 1.8
Coronary heart disease	7 (1.0); 0.4 to 2.1	7 (1.1); 0.4 to 2.2
High blood pressure (hypertension)	165 (24.0); 20.8 to 27.2	154 (23.9); 20.6 to 27.2
High cholesterol (hyperlipidemia)	148 (21.5); 18.4 to 24.6	139 (21.6); 18.4 to 24.7
Stroke	9 (1.3); 0.6 to 2.5	9 (1.4); 0.6 to 2.6
Any other heart condition or heart disease	7 (1.0); 0.4 to 2.1	7 (1.1); 0.4 to 2.2
Chronic obstructive pulmonary disease	28 (4.1); 2.6 to 5.6	28 (4.3); 2.8 to 5.9
Chronic bronchitis	25 (3.6); 2.2 to 5.0	23 (3.6); 2.13 to 5
Emphysema	5 (0.7); 0.2 to 1.7	4 (0.6); 0.2 to 1.6
Asthma	66 (9.6); 7.4 to 11.8	63 (9.8); 7.5 to 12.1
Apnea (pauses in breathing during sleep)	51 (7.4); 5.5 to 9.4	46 (7.1); 5.2 to 9.1
Any other respiratory or lung condition	2 (0.3); <0.1 to 1.1	2 (0.3); <0.1 to 1.1
Cancer	23 (3.3); 2.0 to 4.7	20 (3.1); 1.8 to 4.4
Diabetes	46 (6.7); 4.8 to 8.6	43 (6.7); 4.7 to 8.6
None of the above	362 (52.6); 48.9 to 56.4	342 (53.0); 49.2 to 56.9
Mental health condition		
Yes	181 (26.3); 23.0 to 29.6	166 (25.7); 22.4 to 29.1
No	477 (69.3); 65.9 to 72.8	450 (69.8); 66.2 to 73.3
Do not know	30 (4.4); 2.8 to 5.9	29 (4.5); 2.9 to 6.1
Taking medication or receiving treatment for a mental health condition or emotional problem		
Yes	121 (17.7); 14.9 to 20.6	113 (17.5); 14.6 to 20.5
No	548 (79.7); 76.6 to 82.7	516 (80.0); 76.9 to 83.1
Do not know	18 (2.6); 1.4 to 3.8	16 (2.5); 1.3 to 3.7

[a]AUI: adult who used *IQOS*.

[b]Participants sampled from the *IQOS* consumer database were primarily residing in Georgia, Virginia, North Carolina, or South Carolina.

Tobacco Use History and Cessation Treatment Use History

As shown in Tables 3 and 4, 680 of the 688 (98.8%) established AUIs had ever smoked cigarettes and 628 (91.3%) were currently smoking cigarettes before they tried THS for the first time. By contrast, at the time of assessment, only 338 (49.1%) were smoking. At the time of the assessment, 144 of 645 (22.3%) current established AUIs were also using e-cigarettes and 86 (13.3%) were also using cigars. Use of other tobacco products was relatively rare (≤5%, 33 or fewer/645). Current established smokers (n=326) had an average of 24.0 (95% CI 22.6-25.4) years of smoking history; former established smokers (ie, AUIs who had stopped smoking; n=329) had an average of 20.9 (95% CI 19.6-22.3) years of smoking history.

When asked about cigarette smoking during the 30 days before first trying THS, 519 (75.4%) participants indicated they were smoking every day, 109 (15.8%) were smoking some days but not every day, and 52 (7.6%) indicated they were not smoking cigarettes (including 16 participants who had not smoked cigarettes for at least 12 months, ie, former established smokers). Only a small proportion (n=16, 2.3%) of ever-established THS users were former established smokers before first trying THS. Of note, 11 of these 16 were using at least one other type of tobacco product during the 30 days before trying THS.

XSL•FO
RenderX

The vast majority of ever-established AUIs who were former established smokers at the time of the assessment became former smokers after first trying THS. Of the 329 former established smokers at the time of the assessment, 289 (87.8%) were smoking cigarettes during the 30 days before first trying THS or smoked after first trying THS, which suggested that they became former smokers after first trying THS. Of the 30 participants who did not smoke during the 30 days before first trying THS and indicated that the last time they smoked was before trying THS, 14 indicated they had not smoked for less than 1 month before trying THS, and 16 had not smoked for at least 12 months before first trying THS.

Approximately one-third of established AUIs were using e-cigarettes and about 114 (16.6%) were using cigars; use of other tobacco products was less common (ie, ≤44/688, <7%) before first trying THS.

With respect to cessation treatment, 330 out of 645 (51.2%; 95% CI 47.3%-55.0%) current established AUIs had never used any tobacco cessation treatment, 202 (31.3%; 95% CI 27.7%-34.9%) had used it more than 12 months ago, 39 (6.0%; 95% CI 4.2%-7.9%) used it within the past 12 months but more than 30 days ago, and 27 (4.2%; 95% CI 2.6%-5.7%) used it within the past 30 days at the time of the assessment.

Table 3. Types of tobacco products used by ever-established AUIs[a] (n=688) before first trying the THS[b] and at the time of assessment[c].

Tobacco products	Tobacco use before first trying THS			Current tobacco use at the time of assessment
	Ever tried, n (%); 95% CI	Ever-established use, n (%); 95% CI	Current use, n (%); 95% CI	Current use, n (%); 95% CI
Cigarettes	680 (98.8); 97.7-99.5	632 (91.9); 89.8-93.9	628 (91.3); 89.2-93.4	338 (49.1); 45.4-52.9
Cigars	428 (62.2); 58.6-65.8	125 (18.2); 15.3-21.1	114 (16.6); 13.8-19.4	92 (13.4); 10.8-15.9
Pipe filled with tobacco	151 (22.0); 18.9-25.0	32 (4.7); 3.1-6.2	18 (2.6); 1.4-3.8	17 (2.5); 1.3-3.6
Hookah	225 (32.7); 29.2-36.2	29 (4.2); 2.7-5.7	44 (6.4); 4.6-8.2	34 (4.9); 3.3-6.6
E-vapor products	493 (71.7); 68.3-75.0	298 (43.3); 39.6-47.0	229 (33.3); 29.8-36.8	161 (23.4); 20.2-26.6
Smokeless tobacco	174 (25.3); 22.0-28.5	96 (14.0); 11.4-16.5	38 (5.5); 3.8-7.2	28 (4.1); 2.6-5.6
Oral tobacco-derived nicotine products	143 (20.8); 17.8-23.8	36 (5.2); 3.6-6.9	41 (6.0); 4.2-7.7	27 (3.9); 2.5-5.4
Any tobacco product other than THS	685 (99.6); 98.7-99.9	659 (95.8); 94.3-97.3	658 (95.6); 94.1-97.2	461 (67); 63.5-70.5

[a]AUI: adult who used *IQOS*.

[b]THS: Tobacco Heating System.

[c]Participants sampled from the *IQOS* consumer database were primarily residing in Georgia, Virginia, North Carolina, or South Carolina.

XSL•FO
RenderX

Table 4. Types of tobacco products used by current established AUIs[a] (n=645) before first trying the THS[b] and at the time of assessment[c].

Tobacco products	Tobacco use before first trying THS			Current tobacco use at the time of assessment
	Ever tried (n=645), n (%); 95% CI	Ever-established use (n=645), n (%); 95% CI	Current use (n=645), n (%); 95% CI	Current use (n=645), n (%); 95% CI
Cigarettes	638 (98.9); 97.8 to 99.6	591 (91.6); 89.5 to 93.8	586 (90.9); 88.6 to 93.1	316 (49.0); 45.1 to 52.9
Cigars	396 (61.4); 57.6 to 65.2	118 (18.3); 15.3 to 21.3	106 (16.4); 13.6 to 19.3	86 (13.3); 10.7 to 16.0
Pipe filled with tobacco	138 (21.4); 18.2 to 24.6	30 (4.7); 3.0 to 6.3	17 (2.6); 1.4 to 3.9	17 (2.6); 1.4 to 3.9
Hookah	209 (32.4); 28.8 to 36.0	27 (4.2); 2.6 to 5.7	41 (6.4); 4.5 to 8.2	33 (5.1); 3.4 to 6.8
E-vapor products	462 (71.6); 68.2 to 75.1	275 (42.6); 38.8 to 46.5	211 (32.7); 29.1 to 36.3	144 (22.3); 19.1 to 25.5
Smokeless tobacco	157 (24.3); 21 to 27.7	87 (13.5); 10.9 to 16.1	36 (5.6); 3.8 to 7.4	26 (4.0); 2.5 to 5.6
Oral tobacco-derived nicotine products	128 (19.8); 16.8 to 22.9	32 (5.0); 3.3 to 6.6	35 (5.4); 3.7 to 7.2	23 (3.6); 2.1 to 5.0
Any tobacco product other than THS	643 (99.7); 98.9 to >99.9	617 (95.7); 94.1 to 97.2	616 (95.5); 93.9 to 97.1	428 (66.4); 62.7 to 70.0

[a]AUI: adult who used *IQOS*.

[b]THS: Tobacco Heating System.

[c]Participants sampled from the *IQOS* consumer database were primarily residing in Georgia, Virginia, North Carolina, or South Carolina.

THS Use

Overall Use Behaviors

On average, AUIs had used THS for 1.2 years (mean 1.2, 95% CI 1.1-1.2 for all established AUIs; mean 1.2, 95% CI 1.1-1.3 for current AUIs). Among current AUIs (n=645), one-third were using THS only (n=217, 33.6%; 95% CI 30.0%-37.3%), 280 (43.4%; 95% CI 39.6%-47.2%) were using it and 1 other tobacco product (including 187, 29.0%, who were smoking cigarettes and 93, 14.4%, who were not smoking cigarettes), and 148 (23.0%; 95% CI 19.7%-26.2%) were using it and at least two other tobacco products (including 129, 20.0%, who were smoking cigarettes and 19, 2.9%, who were not smoking cigarettes). Of current AUIs, 443 (68.7%) used THS every day during the 30 days before the assessment. A greater proportion of current established THS users who had used the device for more than a year used THS daily compared with those who had used THS for a year or less (218/300, 72.7%, vs 225/345, 65.2%, *P*=.04). Current AUIs used a median of 15 tobacco sticks per day on days used (Table 5).

XSL•FO
RenderX

Table 5. Patterns of Tobacco Heating System use in the past 30 days at the time of the assessment[a].

Patterns	Current established AUIs[b] (n=645)	Current established AUIs who prefer menthol tobacco sticks (n=337)	Current established AUIs who prefer nonmenthol tobacco sticks (n=308)
Number of days used, n (%); 95% CI			
0-9	36 (5.6); 3.8-7.4	14 (4.2); 2.0-6.3	22 (7.1); 4.3-10.0
10-19	66 (10.2); 7.9-12.6	31 (9.2); 6.1-12.3	35 (11.4); 7.8-14.9
20-29	100 (15.5); 12.7-18.3	52 (15.4); 11.6-19.3	48 (15.6); 11.5-19.6
30	443 (68.7); 65.1-72.3	240 (71.2); 66.4-76.1	203 (65.9); 60.6-71.2
Number of days used, mean (95% CI)	25.9 (25.3-26.5)	26.3 (25.6-27.1)	25.4 (24.5-26.2)
Number of days used, median (IQR)	30 (25-30)	30 (26-30)	30 (22-30)
Number of tobacco sticks used per day on days used, n (%); 95% CI			
<1	7 (1.1); 0.4-2.2	5 (1.5); 0.5-3.4	2 (0.7); 0.1-2.3
1	4 (0.6); 0.2-1.6	2 (0.6); 0.1-2.1	2 (0.7); 0.1-2.3
2	9 (1.4); 0.6-2.6	5 (1.5); 0.5-3.4	4 (1.3); 0.4-3.3
3	17 (2.6); 1.4-3.9	9 (2.7); 1.2-5	8 (2.6); 1.1-5.1
4	10 (1.55); 0.6-2.5	4 (1.2); 0.3-3	6 (2); 0.7-4.2
5-9	104 (16.1); 13.3-19.0	55 (16.3); 12.4-20.3	49 (15.9); 11.8-20
10-14	146 (22.6); 19.4-25.9	79 (23.4); 18.9-28	67 (21.8); 17.2-26.4
15-19	94 (14.6); 11.9-17.3	49 (14.5); 10.8-18.3	45 (14.6); 10.7-18.6
20+	250 (38.8); 35.0-42.5	126 (37.4); 32.2-42.6	124 (40.3); 34.8-45.7
Number of tobacco sticks per day on days used, median (IQR)	15 (10-20)	15 (10-20)	15 (10-20)
Number of tobacco sticks per day, median (IQR)	13 (7-20)	12.5 (8-20)	14 (7-20)

[a]Participants sampled from the *IQOS* consumer database were primarily residing in Georgia, Virginia, North Carolina, or South Carolina.

[b]AUI: adult who used *IQOS*.

Menthol and Nonmenthol Tobacco Stick Preferences

The majority of current AUIs had tried all 3 varieties of tobacco sticks (461/645, 71.5%; 437/645, 67.8%; and 359/645, 55.7%, for amber; green menthol; and blue menthol, respectively) by the time of the assessment. Amber was the most common variety tried first (344/645, 53.3%; 202/645, 31.3%; and 99/645, 15.3%, for green menthol and blue menthol, respectively); at the time of the assessment, nearly half indicated amber tobacco sticks as the variety currently used most often (308/645, 47.8%; 191/645, 29.6%; and 146/645, 22.6%, for green menthol and blue menthol, respectively).

When stratified by menthol and nonmenthol cigarette preference, we observed a clear preference for menthol tobacco sticks among participants who smoked menthol cigarettes most often. For example, among participants who formerly smoked menthol cigarettes most often (n=145), 135 (93.1%) used either green menthol or blue menthol tobacco sticks most often. Among participants who formerly smoked nonmenthol cigarettes most often (n=195), 144 (73.8%) used amber tobacco sticks most often. Similar patterns were observed among participants who currently smoked cigarettes: Among the 127 participants who currently smoked menthol cigarettes most often, 121 (95.3%) used menthol tobacco sticks most often; among the 204 participants who currently smoked nonmenthol cigarettes most often, 161 (78.9%) used nonmenthol tobacco sticks most often. There was little difference in the frequency and amount of tobacco stick use between those who preferred menthol and nonmenthol tobacco sticks (Table 5).

Perceptions of THS

THS was perceived as having lower health risk as compared with cigarette smoking: the mean composite score of the Perceived Health Risk scale was 43.8 (95% CI 42.6-45.1) for THS and 64.4 (95% CI 63.0-65.7) for cigarette smoking.

With respect to understanding the MRTP message, among the 688 ever-established AUIs, 555 (80.7%; 95% CI 77.7%-83.6%) correctly identified “less exposure” to harmful or potentially harmful chemicals when switching completely from cigarettes to THS. Only 33 (4.8%; 95% CI 3.2%-6.4%) perceived “no exposure.” Other participants answered “the same exposure” (59/688, 8.6%; 95% CI 6.5%-10.7%), “more exposure” (6/688, 0.9%, 95% CI 0.3%-1.8%), or they “do not know” (35/688, 5.1%, 95% CI 3.5%-6.7%). Among those who correctly identified “less exposure” (n=555), 471 (84.9%; 95% CI 81.9%-87.9%) understood that “smokers must stop smoking

XSL•FO
RenderX

completely and only use *IQOS*" to reduce their exposure to harmful or potentially harmful chemicals. Only 6 (1.1%; 95% CI 0.4%-2.3%) selected "keep smoking the same amount of cigarettes and also use *IQOS*."

Switching to THS and Transitions to/Back to Cigarette Smoking

Complete Switching

Among all current established AUIs (n=645), 191 (29.6%; 95% CI 26.1%-33.1%) met our predefined "complete switching" from cigarette smoking criteria, that is, participants who (1) had smoked at least 100 cigarettes in their lifetime, (2) were currently smoking during the 30 days before first trying THS, (3) were not smoking at the time of assessment, and (4) smoked their last cigarettes after trying THS. To note, this estimate is best interpreted as the proportion of current AUIs who had switched from cigarettes to THS from a cross-sectional assessment. It should not be interpreted as a "switching rate" as it does not provide any information about the "speed" of transition. Moreover, 350 of 688 (50.9%) AUIs were not smoking at the time of assessment (ie, 338/688, 49.1%, AUIs were also smoking; see Tables 3 and 4). The majority of individuals who were not smoking but did not meet the "complete switching" criteria were those who indicated that the last time they smoked cigarettes was within 30 days *before* first trying THS. (To meet the "complete switching" criteria, participants needed to indicate the last time they smoked was *after* first trying THS.) It is possible that these individuals immediately switched and did not smoke a cigarette once they started using THS. Therefore, we consider our estimate of "complete switching" to be conservative because it does not include those who stopped smoking immediately upon first trying the product.

When stratified by daily and nondaily smoking during the 30 days before first trying THS, we found similar proportions of AUIs who completely switched from cigarette smoking to THS between daily (ie, every day) and nondaily (ie, someday) smokers (158/519, 30.4%, and 33/109, 30.3%, respectively). Among all AUIs, we did not observe any differences in proportions who met the "complete switching" criteria between those who used menthol tobacco sticks and those who used nonmenthol tobacco sticks most often (102/358, 28.5%, and 89/330, 27.0%, respectively). Participants who completely switched from cigarettes to THS had used THS for a longer duration (mean duration 15.6 months) compared with those who did not meet the complete switching definition (mean duration 13.6 months; estimated difference 2.0, 95% CI 0.4-3.7, P=.01).

Among AUIs who were also smoking at the time of assessment (ie, dual-use; n=298), 249 (83.6%; 95% CI 79.4%-87.8%) indicated that they smoked fewer cigarettes per day at the time of assessment; 7 (2.3%; 95% CI 1.0%-4.8%) indicated they smoked more and 42 (14.1%; 95% CI 10.1%-18.0%) indicated they smoked the same number of cigarettes per day at the time of assessment compared with the 30 days before trying THS.

Transition to or Back to Cigarettes

In this study, we observed no relapse of cigarette smoking (defined as the current use of cigarettes and having smoked at least 100 cigarettes and having had not smoked cigarettes for <12 months before first trying THS); we observed 1 individual who reinitiated cigarette smoking (defined as the current use of cigarettes and having smoked at least 100 cigarettes and having had not smoked cigarettes for at least 12 months before first trying THS).

Discussion

Principal Findings

To our knowledge, this study is the first to describe the characteristics of AUIs and comprehensively assess use behaviors and risk perceptions relevant to THS use in a real-world setting among established AUIs in the United States. We found that AUIs tended to be male, non–Hispanic White or non–Hispanic Asian, and middle-aged (age 35-54 years), with at least some college education and a household income over US $60,000. More than half of AUIs reported no physical health conditions and the majority (548/688, 79.7%) reported not receiving treatment for any mental health conditions. Our results showed that the vast majority of established AUIs were existing smokers with an average of over 20 years of smoking history; 570 of 645 (88.4%) had either never used or had tried but stopped using tobacco cessation treatment. After an average of 1.2 years of THS use, almost half of AUIs were not smoking at the time of the assessment, and more than 80% (156/187, 83.4%) of participants who were still smoking indicated they had reduced their cigarette consumption. In addition, almost one-third met our strict definition for "complete switching" from cigarettes to THS and none reported relapse back to smoking. Finally, 555 of 688 (80.7%) participants perceived THS use as less harmful compared with cigarette smoking; few perceived it as risk-free and the vast majority understood the THS-modified risk claim. These findings provide firsthand empirical evidence in a real-world setting on THS consumers and their use behaviors during a relatively early stage of adoption in the United States.

Interpretation of Findings

These findings are in line with evidence from other countries in that the vast majority of AUIs smoked cigarettes when they first tried THS, and few had never smoked a cigarette [24,25]. Our findings suggested that THS can facilitate adults who smoked cigarettes—especially those who did not use cessation treatment or did not successfully quit smoking after trying cessation treatment—to switch to THS.

Previous studies have found that young adults were more likely to use e-vapor products compared with older adults [26,27]. By contrast, we found that those who used THS tended to be middle-aged or older adults with a relatively long smoking history. In addition, the majority of current established AUIs had never tried a cessation treatment and another 202 of 645 (31.3%) tried but discontinued cessation treatment more than 12 months ago. Therefore, THS may be particularly beneficial to those who are older, have a long history of smoking, and did not use or were not successful in using tobacco cessation treatment. Along with the finding of minimal uptake among individuals who had never smoked cigarettes, our results support

XSL•FO
RenderX

the FDA's conclusion that marketing this product is appropriate for the protection of public health.

With respect to product use, the majority of AUIs used it daily and used more than 15 tobacco sticks per day on days used. This level of consumption was generally in line with cigarette consumption among smokers in national surveys [28-30], and daily use of THS has been associated with switching [31,32]. Taken together, these results suggest the replacement of cigarettes with THS use.

The correct perception of risk can play an integral role in switching behaviors. In line with the large body of literature showing that individuals who perceive smoke-free tobacco products as less harmful compared with cigarettes are more likely to switch to such products from cigarette smoking [33,34], a previous study documented that individuals who used THS exclusively (rather than smoking) were more likely to perceive it as less harmful than cigarettes, compared with those who used THS and smoked cigarettes (ie, dual-use) [31]. Given this context, it is encouraging to observe that over 80% (555/688, 80.7%) of participants in this study correctly identified the reduced exposure to harmful or potentially harmful chemicals when switching completely from cigarettes to THS, and few participants incorrectly identified "no exposure."

With respect to tobacco stick varieties, we found an interesting pattern suggesting that menthol tobacco sticks were used and preferred not only by individuals who preferred menthol cigarettes, but also by a sizable portion of those who preferred nonmenthol cigarettes. That is, although amber (or original) was the most commonly ever used (456/645, 70.7%) and first tried (342/645, 53.0%) variety, the 2 menthol varieties accounted for the majority of varieties used most often at the time of assessment. This aligns with the observation that approximately 1 in 5 individuals who preferred nonmenthol cigarettes preferred menthol tobacco sticks; by contrast, only 1 in 20 individuals who preferred menthol cigarettes preferred amber (or original). Although it is not clear why menthol tobacco sticks were more commonly preferred than amber, such findings support the potential role of menthol varieties in continued use. In this study, we found that ever and current use of other tobacco products was more common among participants of this study (with e-vapor products being the most common) compared with that among smokers based on estimates from national surveys [35]. This may suggest that individuals open to trying and using other tobacco products were more likely to try and use THS. When compared with the time before trying THS, we observed a ~30% (229/688 vs 161/688) reduction in e-cigarette use. It is possible that some participants were using e-vapor products to help stop cigarette smoking but were not successful; after initiating THS use, they stopped using e-vapor products. Future studies that assess reasons to stop using other tobacco products will shed new light on this topic.

Limitations

Our findings should be interpreted with the following limitations in mind. First, the study is based on self-reported information. Although we are not aware of any validation studies on self-reported use behaviors, the larger literature supports the reasonable validity of self-reported behaviors in tobacco research [36]. In addition, many measures relied on recalled information, so recall bias cannot be ruled out. We tried to minimize recall bias by (1) asking about prominent behaviors generally easy to recall, such as whether they smoked cigarettes during the 30 days before first trying THS, and (2) limiting the length of the recall period (eg, the past 30 days). Second, it is estimated that approximately 70% of individuals who purchased THS registered themselves in the database. Moreover, of all individuals (N=19,258) invited to complete the eligibility assessment, only 2022 (10.50%) did so. It is unknown whether individuals who registered in the database or responded to the invitation are representative of all AUIs in the United States. Also, due to limited distribution, it is not clear whether these results would apply to all adults who smoke in the United States. Nonetheless, we provided the first estimates of perceptions and behaviors related to THS use from a sample large enough to produce meaningful results. Third, the study was conducted among AUIs, so we cannot provide information about the prevalence of THS use in the general US population. Fourth, no definitive evidence can be drawn for causal relationships from this cross-sectional study. Perhaps future longitudinal studies can explore changes in behavior over time among individuals who adopt THS. Finally, it is noteworthy that during the recruitment period, a notification was sent to consumers on October 13, 2021, informing them that the product would become unavailable after November 28, 2021. In consideration of the potential impact of this communication on consumer behaviors, we also conducted analyses only using data collected before that date (n=463). Results from this subsample were generally in line with the results for the full sample.

Conclusions

Our results provide supportive evidence for the potential of THS to help individuals switch from smoking by comprehensively assessing use behaviors and risk perceptions in a real-world setting among established AUIs in the United States.

Acknowledgments

The study was funded internally and did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The authors thank Dr. Annie Heremans for study design input and reviewing early drafts of the manuscript.

Availability of Data and Materials

The datasets used or analyzed during this study are available from the corresponding author upon reasonable request.

XSL•FO
RenderX

Authors' Contributions

All authors designed the study. HGC, BN, and ARV acquired the data. HGC prepared the first draft. All authors reviewed and revised the manuscript. All authors have approved the final version of the manuscript.

Conflicts of Interest

HCG is a former and BN, ARV, and EGL are current employees of Altria Client Services LLC, a service company to Altria and its tobacco operating companies. PM is an employee of Philip Morris International, a tobacco manufacturer. Both companies were involved in the review and approval of the manuscript for publication.

References

1. GBD 2019 Risk Factors Collaborators. Global burden of 87 risk factors in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet 2020 Oct 17;396(10258):1223-1249 [FREE Full text] [doi: 10.1016/S0140-6736(20)30752-2] [Medline: 33069327]
2. Smoking cessation: a report of the surgeon general- executive summary. US Department of Health and Human Services. Atlanta Georgia: US Department of Health and Human Services; 2020. URL: https://www.hhs.gov/sites/default/files/2020-cessation-sgr-executive-summary.pdf [accessed 2024-11-27]
3. Babb S, Malarcher A, Schauer G, Asman K, Jamal A. Quitting smoking among adults - United States, 2000-2015. MMWR Morb Mortal Wkly Rep 2017 Jan 06;65(52):1457-1464 [FREE Full text] [doi: 10.15585/mmwr.mm6552a1] [Medline: 28056007]
4. García-Rodríguez O, Secades-Villa R, Flórez-Salamanca L, Okuda M, Liu S, Blanco C. Probability and predictors of relapse to smoking: results of the National Epidemiologic Survey on Alcohol and Related Conditions (NESARC). Drug Alcohol Depend 2013 Oct 01;132(3):479-485 [FREE Full text] [doi: 10.1016/j.drugalcdep.2013.03.008] [Medline: 23570817]
5. Fiore M, Jaén C, Baker T. Treating Tobacco Use and Dependence: 2008 Update Clinical Practice Guideline. Rockville, MD: U.S. Department of Health and Human Services. Public Health Service; 2008.
6. Centers for Disease ControlPrevention (CDC). Quitting smoking among adults--United States, 2001-2010. MMWR Morb Mortal Wkly Rep 2011 Nov 11;60(44):1513-1519 [FREE Full text] [Medline: 22071589]
7. Borland R, Partos TR, Yong H, Cummings KM, Hyland A. How much unsuccessful quitting activity is going on among adult smokers? Data from the International Tobacco Control Four Country cohort survey. Addiction 2012 Mar;107(3):673-682 [FREE Full text] [doi: 10.1111/j.1360-0443.2011.03685.x] [Medline: 21992709]
8. Chaiton M, Diemert L, Cohen JE, Bondy SJ, Selby P, Philipneri A, et al. Estimating the number of quit attempts it takes to quit smoking successfully in a longitudinal cohort of smokers. BMJ Open 2016 Jun 09;6(6):e011045 [FREE Full text] [doi: 10.1136/bmjopen-2016-011045] [Medline: 27288378]
9. U.S. Department of Health And Human Services (U.S. DHHS). How Tobacco Smoke Causes Disease: The Biology and Behavioral Basis for Smoking-Attributable Disease: A Report of the Surgeon General. Atlanta, GA: U.S. Department of Health And Human Services, Public Health Service, Office of the Surgeon General; 2010.
10. McNeill A. Harm reduction. BMJ 2004 Apr 10;328(7444):885-887 [FREE Full text] [doi: 10.1136/bmj.328.7444.885] [Medline: 15073074]
11. Abrams DB, Glasser AM, Pearson JL, Villanti AC, Collins LK, Niaura RS. Harm minimization and tobacco control: reframing societal views of nicotine use to rapidly save lives. Annu Rev Public Health 2018 Apr 01;39:193-213 [FREE Full text] [doi: 10.1146/annurev-publhealth-040617-013849] [Medline: 29323611]
12. Gottlieb S, Zeller M. A nicotine-focused framework for public health. N Engl J Med 2017 Sep 21;377(12):1111-1114. [doi: 10.1056/NEJMp1707409] [Medline: 28813211]
13. Simonavicius E, McNeill A, Shahab L, Brose LS. Heat-not-burn tobacco products: a systematic literature review. Tob Control 2019 Sep 04;28(5):582-594 [FREE Full text] [doi: 10.1136/tobaccocontrol-2018-054419] [Medline: 30181382]
14. Lüdicke F, Baker G, Magnette J, Picavet P, Weitkunat R. Reduced exposure to harmful and potentially harmful smoke constituents with the Tobacco Heating System 2.1. Nicotine Tob Res 2017 Feb;19(2):168-175 [FREE Full text] [doi: 10.1093/ntr/ntw164] [Medline: 27613951]
15. Haziza C, de La Bourdonnaye G, Merlet S, Benzimra M, Ancerewicz J, Donelli A, et al. Assessment of the reduction in levels of exposure to harmful and potentially harmful constituents in Japanese subjects using a novel tobacco heating system compared with conventional cigarettes and smoking abstinence: a randomized controlled study in confinement. Regul Toxicol Pharmacol 2016 Nov;81:489-499 [FREE Full text] [doi: 10.1016/j.yrtph.2016.09.014] [Medline: 27693654]
16. Tabuchi T, Gallus S, Shinozaki T, Nakaya T, Kunugita N, Colwell B. Heat-not-burn tobacco product use in Japan: its prevalence, predictors and perceived symptoms from exposure to secondhand heat-not-burn tobacco aerosol. Tob Control 2018 Jul 16;27(e1):e25-e33 [FREE Full text] [doi: 10.1136/tobaccocontrol-2017-053947] [Medline: 29248896]
17. Afolalu EF, Langer P, Fischer K, Roulet S, Magnani P. Prevalence and patterns of tobacco and/or nicotine product use in Japan (2017) after the launch of a heated tobacco product (IQOS(R)): a cross-sectional study. F1000Res 2021;10:504 [FREE Full text] [doi: 10.12688/f1000research.52407.2] [Medline: 35528952]

XSL•FO
RenderX

18. Sakaguchi C, Nagata Y, Kikuchi A, Takeshige Y, Minami N. Differences in levels of biomarkers of potential harm among users of a heat-not-burn tobacco product, cigarette smokers, and never-smokers in Japan: a post-marketing observational study. Nicotine Tob Res 2021 Jun 08;23(7):1143-1152 [FREE Full text] [doi: 10.1093/ntr/ntab014] [Medline: 33502518]
19. Polosa R, Morjaria JB, Prosperini U, Busà B, Pennisi A, Gussoni G, et al. Health outcomes in COPD smokers using heated tobacco products: a 3-year follow-up. Intern Emerg Med 2021 Apr 23;16(3):687-696 [FREE Full text] [doi: 10.1007/s11739-021-02674-3] [Medline: 33754228]
20. Insights Association Code of Standards and Ethics for Market Research and Data Analytics. URL: https://www.insightsassociation.org/Portals/INSIGHTS/Insights%20Association%20Code%20of%20Standards_November%202023_v2.pdf [accessed 2024-12-06]
21. ICC/ESOMAR International Code on Market, Opinion and Social Research and Data Analytics. URL: https://esomar.org/uploads/attachments/ckqtawvjq00uukdtrhst5sk9u-iccesomar-international-code-english.pdf [accessed 2024-12-06]
22. Cano S, Chrea C, Salzberger T, Alfieri T, Emilien G, Mainy N, et al. Development and validation of a new instrument to measure perceived risks associated with the use of tobacco and nicotine-containing products. Health Qual Life Outcomes 2018 Sep 21;16(1):192 [FREE Full text] [doi: 10.1186/s12955-018-0997-5] [Medline: 30241527]
23. Noggle B, Cheng H, Iyer S, Vansickel A. IQOS® cross-sectional and cohort US study documentation. Qeios 2023:e1 [FREE Full text] [doi: 10.32388/vkd77j]
24. Sutanto E, Miller C, Smith DM, O'Connor RJ, Quah ACK, Cummings KM, et al. Prevalence, use behaviors, and preferences among users of heated tobacco products: findings from the 2018 ITC Japan Survey. Int J Environ Res Public Health 2019 Nov 21;16(23):4630 [FREE Full text] [doi: 10.3390/ijerph16234630] [Medline: 31766410]
25. Gallus S, Lugo A, Liu X, Borroni E, Clancy L, Gorini G, TackSHS Project Investigators. Use and awareness of heated tobacco products in Europe. J Epidemiol 2022 Mar 05;32(3):139-144 [FREE Full text] [doi: 10.2188/jea.JE20200248] [Medline: 33456019]
26. Levy DT, Yuan Z, Li Y. The prevalence and characteristics of e-cigarette users in the U.S. Int J Environ Res Public Health 2017 Oct 11;14(10):1200 [FREE Full text] [doi: 10.3390/ijerph14101200] [Medline: 29019917]
27. Bandi P, Cahn Z, Goding Sauer A, Douglas CE, Drope J, Jemal A, et al. Trends in e-cigarette use by age group and combustible cigarette smoking histories, U.S. adults, 2014-2018. Am J Prev Med 2021 Feb;60(2):151-158. [doi: 10.1016/j.amepre.2020.07.026] [Medline: 33032869]
28. Cornelius ME, Loretan CG, Wang TW, Jamal A, Homa DM. Tobacco product use among adults - United States, 2020. MMWR Morb Mortal Wkly Rep 2022 Mar 18;71(11):397-405 [FREE Full text] [doi: 10.15585/mmwr.mm7111a1] [Medline: 35298455]
29. American Lung Association. Tobacco trends brief. American Lung Association. 2022. URL: https://www.lung.org/research/trends-in-lung-disease/tobacco-trends-brief/overall-smoking-trends [accessed 2024-11-27]
30. Substance Abuse and Mental Health Services Administration (SAMHSA). Results from the 2020 National Survey on Drug Use and Health: detailed tables. SAMHSA. Rockville, MD: SAMHSA; 2020. URL: https://www.samhsa.gov/data/sites/default/files/reports/rpt35323/NSDUHDetailedTabs2020v25/NSDUHDetailedTabs2020v25/NSDUHDetTabsSect2pe2020.htm [accessed 2024-11-27]
31. Sutanto E, Miller C, Smith DM, Borland R, Hyland A, Cummings KM, et al. Concurrent daily and non-daily use of heated tobacco products with combustible cigarettes: findings from the 2018 ITC Japan Survey. Int J Environ Res Public Health 2020 Mar 22;17(6):2098 [FREE Full text] [doi: 10.3390/ijerph17062098] [Medline: 32235702]
32. Baig SA, Giovenco DP. Behavioral heterogeneity among cigarette and e-cigarette dual-users and associations with future tobacco use: findings from the Population Assessment of Tobacco and Health Study. Addict Behav 2020 May;104:106263 [FREE Full text] [doi: 10.1016/j.addbeh.2019.106263] [Medline: 32028096]
33. Persoskie A, O'Brien EK, Poonai K. Perceived relative harm of using e-cigarettes predicts future product switching among US adult cigarette and e-cigarette dual users. Addiction 2019 Dec;114(12):2197-2205. [doi: 10.1111/add.14730] [Medline: 31278802]
34. Brose LS, Brown J, Hitchman SC, McNeill A. Perceived relative harm of electronic cigarettes over time and impact on subsequent use. A survey with 1-year and 2-year follow-ups. Drug Alcohol Depend 2015 Dec 01;157:106-111 [FREE Full text] [doi: 10.1016/j.drugalcdep.2015.10.014] [Medline: 26507173]
35. Li L, Borland R, Cummings KM, Gravely S, Quah ACK, Fong GT, et al. Patterns of non-cigarette tobacco and nicotine use among current cigarette smokers and recent quitters: findings from the 2020 ITC Four Country Smoking and Vaping Survey. Nicotine Tob Res 2021 Aug 18;23(9):1611-1616 [FREE Full text] [doi: 10.1093/ntr/ntab040] [Medline: 33693833]
36. Patrick DL, Cheadle A, Thompson DC, Diehr P, Koepsell T, Kinne S. The validity of self-reported smoking: a review and meta-analysis. Am J Public Health 1994 Jul;84(7):1086-1093. [doi: 10.2105/ajph.84.7.1086] [Medline: 8017530]

Abbreviations

AUI: adult who used IQOS
CASRO: Council of American Survey Research Organization
CDO: Cease-and-Desist Order

XSL•FO
RenderX

ENDS: electronic nicotine delivery system
ESOMAR: European Society for Opinion and Marketing Research
FDA: Food and Drug Administration
ITC: United States International Trade Commission
MRTP: modified risk tobacco product

Edited by A Mavragani; submitted 15.02.24; peer-reviewed by A Hassan, K Cummings; comments to author 16.07.24; revised version received 13.09.24; accepted 11.10.24; published 07.02.25.

Please cite as:
Cheng HG, Noggle B, Vansickel AR, Largo EG, Magnani P
Tobacco Use, Risk Perceptions, and Characteristics of Adults Who Used a Heated Tobacco Product (IQOS) in the United States: Cross-Sectional Survey Study
JMIR Form Res 2025;9:e57398
URL: https://formative.jmir.org/2025/1/e57398
doi:10.2196/57398
PMID:

©Hui G Cheng, Brendan Noggle, Andrea R Vansickel, Edward G Largo, Pierpaolo Magnani. Originally published in JMIR Formative Research (https://formative.jmir.org), 07.02.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Predictors of the Intention to Stop Using Smart Devices at Bedtime Among University Students in Saudi Arabia: Cross-Sectional Survey

Manal Almalki[1], PhD

Public Health Department, College of Nursing and Health Sciences, Jazan University, Jazan, Saudi Arabia

Corresponding Author:
Manal Almalki, PhD
Public Health Department
College of Nursing and Health Sciences
Jazan University
1st Fl
Al Maarefah Rd
Jazan, 45142
Saudi Arabia
Phone: 966 173290000 ext 5548
Email: almalki@jazanu.edu.sa

Abstract

Background: The widespread use of smart devices, particularly among university students, has raised concerns about their impact on sleep quality. Bedtime usage of smart devices is associated with sleep disruptions and poor sleep quality.

Objective: This study aimed to explore the behavioral and perceptual factors influencing university students' intention to stop using smart devices at bedtime in Saudi Arabia.

Methods: A cross-sectional survey was conducted in June 2024 and distributed via social media platforms to university students (aged ≥18 years). The questionnaire collected data on demographics, smart device usage habits, perceived negative effects on sleep, and physical sleep disturbances. The Pittsburgh Sleep Quality Index was used to assess sleep quality. Path analysis was performed to evaluate relationships between the outcome variables, intended to stop using smart device usage, and 3 latent variables: sleep quality smartphone usage, sleep quality perceived negative effects, and sleep quality during the past month. Model fit was assessed using chi-square, comparative fit index, and root mean square error of approximation.

Results: Of the 774 participants, 90.43% (700/774) reported using their smart devices every night and 72.48% (561/774) believed bedtime device use negatively affected them the next morning. The most frequently reported next-morning symptoms were fatigue or drowsiness (480/774, 62.01%). Common purposes for bedtime device use were staying in touch with friends or family (432/774, 55.81%), entertainment (355/774, 45.86%), and filling up spare time (345/774, 44.57%). Overall, 58.26% (451/774) expressed an intention to stop bedtime device use within the next 3 months. Path analysis demonstrated that frequent nightly use (path coefficient=0.36) and after-lights-off usage (0.49) were positively associated with the intention to stop, whereas spending ≥3 hours on devices (–0.35) and engaging in multiple activities (–0.18) had negative associations. The strongest predictors of the intention to stop were perceived negative effects on next-morning well-being (0.71) and difficulty breathing comfortably during sleep (0.64). Model fit was excellent (comparative fit index=0.845 and root mean square error of approximation=0.039).

Conclusions: Perceived negative effects on sleep quality and physical sleep disturbances are strong predictors of the intention to stop using smart devices at bedtime among university students in Saudi Arabia. Interventions aimed at improving sleep hygiene should focus on raising awareness about the impact of smart device use on well-being and addressing behaviors such as late-night usage and heavy screen time. Public health strategies should target both psychological and physiological aspects of bedtime smart device use to improve sleep quality in this population.

(JMIR Form Res 2025;9:e67223) doi:10.2196/67223

KEYWORDS

smart devices; smartphone; digital health; digital technology; sleep quality; university student; bedtime habits; Saudi Arabia; path analysis; sleep disturbances; well-being; usage; intention; behavior; mobile phone

XSL•FO
RenderX

Introduction

In the digital age, smart devices such as smartphones, tablets, and laptops have become integral parts of daily life, especially among university students [1]. These devices offer constant connectivity, enabling students to engage in academic activities, communicate with peers, and access entertainment. In Saudi Arabia, where more than 97% of the population uses smartphones [2], smart devices have become an essential tool for both learning and social interaction. On a global scale, the number of smartphone mobile network subscriptions exceeded 7 billion in 2023 and is anticipated to exceed 7.7 billion by 2028 [3].

University students, often facing academic pressures and irregular schedules, are particularly vulnerable to the excessive use of smart devices at bedtime [4-6]. They may engage in multiple bedtime activities, such as social media browsing, texting, gaming, or even frequent awakenings to check notifications, which counteract the body's natural inclination to relax and fall asleep [7-9]. These behaviors are further intensified by the immediate gratification of staying connected and creating a cycle that sustains poor sleep practices [8-10]. In Saudi Arabia, smart devices often serve as a primary means of students' communication, relaxation, and staying in touch with friends and family [11], which further reinforces their use, even during late-night hours [12].

The excessive use of these smart devices, particularly during bedtime, raises concerns about their impact on students' sleep quality [4], overall well-being [5], and academic performance [6]. Late-night screen exposure can disrupt the body's circadian rhythm through both blue light emissions and heightened mental stimulation [13], often delaying sleep onset and reducing total sleep duration [14-17]. Over time, these disruptions can lead to chronic sleep debt, making students more prone to daytime drowsiness, poor concentration [4,5], and lower academic achievement [14,15]. Moreover, poor sleep quality can contribute to the development of mental health issues such as anxiety and depression [5], while these conditions can further disrupt sleep patterns [16]. This mental stimulation can lead to increased cognitive arousal and addiction, which is detrimental to initiating and maintaining sleep [7,8,17]. This addiction does not only affect the quantity of sleep but also its quality, leading to a less restorative sleep experience [1,4,18].

Despite growing awareness of the potential harms of excessive screen time, many students continue to use smart devices at bedtime [13]. Understanding the factors that drive students to continue or stop using these devices at bedtime is crucial for addressing the sleep-related issues prevalent in this population [19,20]. According to the Theory of Planned Behavior [21], individuals' intentions to engage in or discontinue a behavior are shaped by their attitudes toward the behavior, perceived control over it, and social norms. In the context of smart device use, it is likely that students' perceptions of how device use affects their sleep and well-being, as well as their ingrained habits, play a key role in their decision to reduce or stop using smart devices at bedtime.

This study aims to examine the factors that predict the intention to stop using smart devices at bedtime among university students in Saudi Arabia. Specifically, it investigates the relationship between 3 key domains: (1) smart device usage behaviors (eg, frequency, duration, and activities performed on devices), (2) perceived negative effects of smart device use on sleep quality and next-morning well-being, and (3) physical sleep disturbances experienced during the past month. Through path analysis, this study seeks to understand the complex interplay between these variables and provide insights that can inform public health interventions targeted at improving sleep hygiene among Saudi university students.

Methods

Study Design

In this study, a cross-sectional survey of university students was conducted in June 2024 in Saudi Arabia. The STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines of cross-sectional studies were followed to report our study's findings [22]. The questionnaire was piloted with a small group of university students (n=7) to ensure clarity and relevance, with feedback used to refine its wording and structure.

Study Recruitment

Participants were recruited using convenience sampling. A link to the questionnaire was posted on social media platforms such as WhatsApp (Meta), Snapchat, Telegram, and X (formerly known as Twitter). These platforms were chosen due to their popularity among university students in Saudi Arabia [23,24]. The link was accompanied by a recruitment message that outlined the study's objectives, eligibility criteria, and an assurance of confidentiality. Eligible participants were students aged 18 years or older, currently enrolled in universities in Saudi Arabia, and capable of providing informed consent before starting the survey.

Ethical Considerations

Ethical approval for the study was obtained from the Standing Committee for Publication and Research Ethics at Jazan University (reference number REC-45/11/1143). On the opening page of the survey, participants were presented with information about the study's objectives and scope, followed by an informed consent statement. Participants were free to withdraw at any point without consequence. Also, participation was anonymous and no personally identifiable information was collected or used. No monetary or nonmonetary incentives were offered. Finally, this paper does not include any images or material in which participants can be identified.

Questionnaire Design

The questionnaire was created using a Google Form, available in both English and Arabic, and it consisted of 4 sections. The first section gathered demographic data such as gender, age, marital status, highest level of education, residential region, health status, and employment status.

The second section examined participants' smart device usage during bedtime during the past month. Questions covered the

XSL•FO
RenderX

type of smart device most frequently used, the average frequency of use per week, time spent on the device at bedtime (ie, categorized as “less than 2 hours” or “3 hours or more”), usage after lights were turned off, the use of headphones, and other related behaviors. This section also addressed the activities participants engaged in while using their devices (eg, chatting on messaging platforms, watching videos, and setting alarms or reminders), as well as the purposes of smart device use before sleep (eg, education, entertainment, business, or work networking).

The third section explored participants’ perceptions of how smart device usage at bedtime affected their sleep quality, sleep duration, and overall health. The final section used the Pittsburgh Sleep Quality Index (PSQI) to assess participants’ physical sleep disturbances experienced during the past month. The PSQI consists of 19 self-rated items that evaluate subjective sleep quality, sleep latency, habitual sleep efficiency, sleep disturbances, use of sleep medications, and daytime dysfunction [25]. The Arabic-translated version of the PSQI, which has been validated for use in research, was used in this study [26].

Statistical Analysis

Incomplete surveys were excluded. A power calculation using G*Power (α=.05, power=80%, moderate effect size, odds ratio [OR 1.5]) indicated a minimum sample size of 384 participants. The final sample size was 774, which exceeded this requirement, ensuring sufficient power to detect significant associations.

Categorical variables were summarized as frequencies and percentages. ORs with 95% CI values were calculated to assess the relations and their strengths between variables. A path analysis was then conducted to understand the complex relationships between the measured variable intended to stop using smart device usage, latent variables sleep quality smartphone usage (SQSU), sleep quality perceived negative effects (SQPNE), and sleep quality during the past month (SQDPM). This method was chosen because it allows for the simultaneous analysis of multiple dependent and independent variables, providing a comprehensive understanding of direct and indirect relationships between variables.

A typical path modal sequencing equation modeling (SEM) plot is constructed by defining a multiple set of regression formulas. Model fit was assessed by chi-square, comparative fit index (CFI), and root mean square error of approximation (RMSEA). The value of RMSEA less than 0.05 with CFI equal to 0.845 or higher with nonsignificant chi-square statistic indicates a good model fit. Data were coded, validated, and analyzed using IBM SPSS software, version 27.0 for Windows (IBM Corp). Plots were constructed using R software (version 4.1.2; R Core Team) with Lavaan and Semplot R packages. Two-tailed *P* value <.05 was considered to denote statistical significance.

Results

Participant Characteristics

A total of 774 participants completed the questionnaire, as shown in Table 1. Over half of them (451/774, 58.26%) expressed an intention to stop using smart devices during bedtime within the next 3 months. None of the demographic characteristics was associated with the acceptability of stopping the use of smart devices during bedtime; therefore, OR analysis is removed from Table 1.

The majority (397/774, 88.03%) were aged 24 years or younger, 74.72% (337/774) were female, and 86.92% (392/774) were unmarried. More than three-quarters of participants (657/774, 84.88%) had an undergraduate degree or higher, and 87.62% (283/323) did not intend to stop using smart devices at bedtime. Likewise, more than three-quarters (688/774, 88.89%) of participants who did not have any sleep disorders had 38% higher odds of intending to stop using smart devices at bedtime compared to participants with medically diagnosed sleep disorders. The most common sleep duration among all participants was 5-6 hours, with 34.88% (270/774) reporting this sleep range. Participants who intended to stop using smart devices at bedtime had a slightly higher percentage (176/451, 39.02%) of reporting 5-6 hours of sleep compared to those who did not intend to stop (94/323, 29.10%).

XSL•FO
RenderX

Table 1. Sample demographic characteristics of university students in Saudi Arabia.

Variables	Values (N=774)	Likely to stop using smart devices during bedtime in the next 3 months	
		Yes (n=451)	No (n=323)
Age (years), n (%)			
18-24	681 (87.98)	397 (88.03)	284 (87.93)
≥25	93 (12.02)	54 (11.97)	39 (12.07)
Sex, n (%)			
Male	190 (24.55)	114 (25.28)	76 (23.53)
Female	584 (75.45)	337 (74.72)	247 (76.47)
Marital status, n (%)			
Married	91 (11.76)	59 (13.08)	32 (9.91)
Not married	683 (88.24)	392 (86.92)	291 (90.09)
Highest level of education, n (%)			
Diploma after high school or below	117 (15.12)	77 (17.07)	40 (12.38)
Undergraduate studies or higher	657 (84.88)	374 (82.93)	283 (87.62)
Residential region, n (%)			
North	47 (6.07)	30 (6.65)	17 (5.26)
South	382 (49.35)	223 (49.45)	159 (49.23)
East	21 (2.71)	13 (2.88)	8 (2.48)
West	69 (8.91)	39 (8.65)	30 (9.29)
Middle	255 (32.94)	146 (32.37)	109 (33.75)
Medically diagnosed with any sleep disorders, n (%)			
Yes	86 (11.11)	56 (12.42)	30 (9.29)
No	688 (88.89)	395 (87.58)	293 (90.71)
Sleep hours per day or night, n (%)			
Less than 5 h	118 (15.25)	68 (15.08)	50 (15.48)
5-6 h	270 (34.88)	176 (39.02)	94 (29.10)
7-8 h	245 (31.65)	138 (30.60)	107 (33.13)
9-10 h	74 (9.56)	40 (8.87)	34 (10.53)
More than 10 h	67 (8.66)	29 (6.43)	38 (11.76)
Employment status, n (%)			
Employed	210 (27.13)	131 (29.05)	79 (24.46)
Not employed	564 (72.86)	320 (70.95)	244 (75.54)
Job has night shifts (n=210), n (%)			
Yes	51 (24.28)	32 (24.43)	19 (24.05)
No	159 (75.71)	99 (75.57)	60 (75.95)

Smart Device Usage Behaviors

More than a ninth decile of participants (737/774, 94.83%) used a smartphone, followed by 3.74% (29/774) using a tablet, and only 1.03% (8/774) using a laptop, as shown in Table 2. The groups who were less willing to stop using smart devices during their bedtime in the next 3 months, in descending order, were participants who used their devices 1-3 nights per week (OR 0.05, 95% CI 0.00-0.43; $P<.001$), participants who used their devices 4-6 nights per week (OR 0.48, 95% CI 0.25-0.90; $P=.02$), participants who did not frequently use devices after lights were turned off (OR 0.52, 95% CI 0.33-0.83; $P<.001$), participants who did not often use headphones during bedtime (OR 0.73, 95% CI 0.55-0.97; $P=.03$), and participants who were not frequently awakened by smart devices at night (OR 0.71, 95% CI 0.53-0.95; $P=.02$). These groups showed significantly lower odds of intending to stop using their smart devices during bedtime compared to the respective reference groups.

XSL•FO
RenderX

Table 2. Smart device usage behaviors among university students in Saudi Arabia.

Variables	Values (N=774)	Likely to stop using smart devices during bedtime in the next 3 months			
		Yes (n=451)	No (n=323)	Odds ratio (95% CI)	*P* value
Smart device type, n (%)					
Smartphone	737 (94.83)	434 (96.23)	303 (93.81)	Ref	__[a]
Laptop	8 (1.03)	2 (0.44)	6 (1.86)	4.29 (0.86-21.43)	.71
Tablet	29 (3.74)	15 (3.33)	14 (4.33)	1.33 (0.63-2.81)	.44
Frequency of using the smart device per week (SQSU[b]_1), n (%)					
Every night	700 (90.43)	392 (86.92)	308 (95.36)	Ref	—
4-6 nights	51 (6.58)	37 (8.20)	14 (4.33)	0.48 (0.25-0.90)	.02
1-3 nights	23 (2.97)	22 (4.88)	1 (0.31)	0.05 (0.00-0.43)	<.001
Time spent on a smart device at bedtime (SQSU_2), n (%)					
<2 h or less	494 (63.82)	309 (68.51)	185 (57.28)	Ref	—
3 h or more	280 (36.17)	142 (31.49)	138 (42.72)	1.62 (1.20-2.18)	<.001
Often using the smart device after lights are turned off (SQSU_3), n (%)					
Yes	674 (87.08)	380 (84.26)	294 (91.02)	Ref	—
No	100 (12.91)	71 (15.74)	29 (8.98)	0.52 (0.33-0.83)	<.001
Often using headphones during bedtime (SQSU_4), n (%)					
Yes	370 (47.80)	201 (44.57)	169 (52.32)	Ref	—
No	404 (52.19)	250 (55.43)	154 (47.68)	0.73 (0.55-0.97)	.03
Often awakened by the smart device at night (SQSU_5), n (%)					
Yes	309 (39.92)	165 (36.59)	144 (44.58)	Ref	—
No	465 (60.07)	286 (63.41)	179 (55.42)	0.71 (0.53-0.95)	.02
The most common posture when using the smart device, n (%)					
Lying down	668 (86.30)	383 (84.92)	285 (88.24)	Ref	—
Sitting	106 (13.69)	68 (15.08)	38 (11.76)	0.75 (0.49-1.14)	.20
Keeping the smart device while sleeping, n (%)					
In the bed	414 (53.48)	235 (52.11)	179 (55.42)	Ref	—
On the bedside table	316 (40.82)	186 (41.24)	130 (40.25)	0.91 (0.68-1.23)	.59
Far from the bed but inside the bedroom	44 (5.68)	30 (6.65)	14 (4.33)	0.61 (0.31-1.1)	.15
Most common smart device mode, n (%)					
Not general	586 (75.71)	336 (74.50)	250 (77.40)	Ref	—
General or vibratory	188 (24.28)	115 (25.50)	73 (22.60)	0.85 (0.60-1.19)	.39
Activities performed when using smart device during bedtime (SQSU_6), n (%)					
(1) Setting alarms, reminders, and tasks					
False	309 (39.92)	173 (38.36)	136 (42.11)	Ref	—
True	465 (60.07)	278 (61.64)	187 (57.89)	0.85 (0.63-1.14)	.29
(2) Chatting via WhatsApp, Snapchat, Telegram, etc					
False	269 (34.75)	170 (37.69)	99 (30.65)	Ref	—
True	505 (65.24)	281 (62.31)	224 (69.35)	1.36 (1.01-1.85)	.04
(3) Debating and discussion via Twitter, WhatsApp					
False	594 (76.74)	346 (76.72)	248 (76.78)	Ref	—
True	180 (23.25)	105 (23.28)	75 (23.22)	0.99 (0.71-1.39)	1

XSL•FO
RenderX

Variables	Values (N=774)	Likely to stop using smart devices during bedtime in the next 3 months			
		Yes (n=451)	No (n=323)	Odds ratio (95% CI)	*P* value
(4) Watching YouTube, TikTok, Instagram, etc					
False	276 (35.65)	178 (39.47)	98 (30.34)	Ref	—
True	498 (64.34)	273 (60.53)	225 (69.66)	1.49 (1.10-2.02)	<.001
(5) Watching TV shows and movies					
False	470 (60.72)	298 (66.08)	172 (53.25)	Ref	—
True	304 (39.27)	153 (33.92)	151 (46.75)	1.70 (1.27-2.29)	<.001
(6) Listening to the Qur'an and religious supplications					
False	606 (78.29)	349 (77.38)	257 (79.57)	Ref	—
True	168 (21.70)	102 (22.62)	66 (20.43)	0.87 (0.61-1.24)	.48
(7) Listening to a podcast					
False	709 (91.60)	411 (91.13)	298 (92.26)	Ref	—
True	65 (8.39)	40 (8.87)	25 (7.74)	0.86 (0.51-1.45)	.60
(8) Listening to music					
False	674 (87.08)	404 (89.58)	270 (83.59)	Ref	—
True	100 (12.91)	47 (10.42)	53 (16.41)	1.68 (1.10-2.57)	.01
(9) Researching brands					
False	701 (90.56)	412 (91.35)	289 (89.47)	Ref	—
True	73 (9.43)	39 (8.65)	34 (10.53)	1.24 (0.76-2.01)	.38
(10) Playing games					
False	657 (84.88)	391 (86.70)	266 (82.35)	Ref	—
True	117 (15.11)	60 (13.30)	57 (17.65)	1.39 (0.94-2.07)	.10
(11) Surfing the internet					
False	604 (78.03)	359 (79.60)	245 (75.85)	Ref	—
True	170 (21.96)	92 (20.40)	78 (24.15)	1.24 (0.88-1.74)	.21
(12) Browsing emails and text messages					
False	626 (80.87)	368 (81.60)	258 (79.88)	Ref	—
True	148 (19.12)	83 (18.40)	65 (20.12)	1.17 (0.77-1.60)	.57
Purposes of using the smart device during bedtime (SQSU_7) , n (%)					
(1) Education and study					
False	453 (58.52)	258 (57.21)	195 (60.37)	Ref	—
True	321 (41.47)	193 (42.79)	128 (39.63)	0.87 (0.65-1.17)	.41
(2) Staying in touch with friends and family					
False	342 (44.18)	204 (45.23)	138 (42.72)	Ref	—
True	432 (55.81)	247 (54.77)	185 (57.28)	1.10 (0.83-1.47)	.50
(3) Following news and current					
False	498 (64.34)	292 (64.75)	206 (63.78)	Ref	—
True	276 (35.65)	159 (35.25)	117 (36.22)	1.04 (0.77-1.40)	.81
(4) Sharing opinions and thoughts on social networks					
False	581 (75.06)	344 (76.27)	237 (73.37)	Ref	—
True	193 (24.93)	107 (23.73)	86 (26.63)	1.16 (0.83-1.62)	.39
(5) Business or work networking					

XSL•FO
RenderX

Variables	Values (N=774)	Likely to stop using smart devices during bedtime in the next 3 months			
		Yes (n=451)	No (n=323)	Odds ratio (95% CI)	*P* value
False	680 (87.85)	390 (86.47)	290 (89.78)	Ref	—
True	94 (12.14)	61 (13.53)	33 (10.22)	0.72 (0.46-1.14)	.18
(6) Researching how to do things					
False	660 (85.27)	386 (85.59)	274 (84.83)	Ref	—
True	114 (14.72)	65 (14.41)	49 (15.17)	1.06 (0.71-1.58)	.83
(7) Shopping					
False	615 (79.45)	359 (79.60)	256 (79.26)	Ref	—
True	159 (20.54)	92 (20.40)	67 (20.74)	1.02 (0.71-1.45)	.92
(8) Gaming					
False	619 (79.97)	374 (82.93)	245 (75.85)	Ref	—
True	155 (20.02)	77 (17.07)	78 (24.15)	1.54 (1.08-2.20)	.01
(9) Entertainment					
False	419 (54.13)	275 (60.98)	144 (44.58)	Ref	—
True	355 (45.86)	176 (39.02)	179 (55.42)	1.94 (1.45-2.59)	<.001
(10) Finding information					
False	603 (77.90)	365 (80.93)	238 (73.68)	Ref	—
True	171 (22.09)	86 (19.07)	85 (26.32)	1.51 (1.07-2.13)	.01
(11) Being relaxed and calming down					
False	569 (73.51)	337 (74.72)	232 (71.83)	Ref	—
True	205 (26.48)	114 (25.28)	91 (28.17)	1.15 (0.84-1.60)	.40
(12) Filling up spare time					
False	429 (55.42)	269 (59.65)	160 (49.54)	Ref	—
True	345 (44.57)	182 (40.35)	163 (50.46)	1.50 (1.12-2.00)	<.001

[a]Not applicable.

[b]SQSU: sleep quality smartphone usage.

The analysis of various activities performed on smart devices during bedtime and their impact on the intention to stop using these devices in the next 3 months is presented in Table 2. The 3 most common activities were chatting via messaging platforms like WhatsApp, Snapchat, and Telegram, with 65.24% (505/774) of participants reporting this behavior. Next, 64.34% (498/774) of participants indicated they spent their time watching content on platforms such as YouTube, TikTok, and Instagram. Setting alarms, reminders, and tasks was also a frequent activity, with 60.07% (465/774) of participants engaging in this before bed. On the other hand, the least common activity performed on smart devices during bedtime was listening to music (100/774, 12.91%). Participants who used their devices for watching TV shows and movies (OR 1.70, 95% CI 1.27-2.29), listening to music (OR 1.68, 95% CI 1.10-2.57), and watching videos on platforms like YouTube, TikTok, and Instagram (OR 1.49, 95% CI 1.10-2.02) were less willing to stop using smart devices during bedtime in the next 3 months.

The analysis of different purposes for using smart devices during bedtime and their impact on the intention to stop using these devices in the next 3 months is presented in Table 2. The most common purposes were staying in touch with friends and family (432/774, 55.81%), entertainment (355/774, 45.86%), and filling up spare time (201/774, 44.57%). Participants who used their devices for entertainment (OR 1.94, 95% CI 1.45-2.59), finding information (OR 1.51, 95% CI 1.07-2.13), filling up spare time (OR 1.50, 95% CI 1.12-2.00), gaming (OR 1.54, 95% CI 1.08-2.20), and chatting via WhatsApp, Snapchat, Telegram, etc (OR 1.36, 95% CI 1.01-1.85) were less willing to stop using smart devices during bedtime in the next 3 months.

Perceived Negative Effects of Smart Device Use on Sleep Quality and Next-Morning Well-Being

Over one-third of the participants (270/774, 35.23%) reported having poor sleep quality, as shown in Table 3. The OR analysis indicated that the perception of the negative influence of smartphone usage on sleep quality strongly impacted the intention to stop using these devices at bedtime. Participants who did not believe that using smart devices during bedtime negatively affected them the next morning were more likely to not intend to stop using their devices compared to those who did believe it (OR 3.53, 95% CI 2.53-4.91, *P*<.001). Similarly,

XSL•FO
RenderX

participants who did not believe that smart device use during bedtime made them sleep less were 2.76 times more likely to not intend to stop (OR 2.76, 95% CI 1.97-3.86). Those who did not believe that using smart devices during bedtime caused a delayed sleep schedule were also more likely to not intend to stop compared to those who did believe it (OR 2.41, 95% CI 1.33-4.35).

The OR analysis also suggested that the perception of immediate negative health impacts influenced the intention to reduce or cease the use of smart devices at night. The most common issue was fatigue and drowsiness, with 62.01% (480/774) of participants experiencing these symptoms. Headaches were also prevalent, affecting 45.09% (349/774) of individuals, while 40.82% (316/774) reported mood swings and 40.69% (315/774) experienced a lack of focus and distraction. Participants who did not report any negative health observations had significantly higher odds (OR 5.25, 95% CI 2.36-11.69) of intending to stop using smart devices compared to those who did report negative health observations. Conversely, more than two-thirds of participants who experienced headaches, dry eyes, or lack of focus and distraction were less likely to intend to stop using smart devices at bedtime. Specifically, participants with headaches were 37% less likely (OR 0.63, 95% CI 0.47-0.85), those with dry eyes were 31% less likely (OR 0.69, 95% CI 0.51-0.94), and those with a lack of focus and distraction were 30% less likely (OR 0.70, 95% CI 0.52-0.95) to intend to stop compared to those without these symptoms.

XSL•FO
RenderX

Table 3. Perceived negative effects of smart device use on sleep quality and next-morning well-being among university students in Saudi Arabia.

Variables	Values (N=774)	Likely to stop using smart devices during bedtime in the next 3 months			
		Yes (n=451)	No (n=323)	Odds ratio (95% CI)	*P* value
Using smart devices during bedtime makes you sleep less (SQPNE[a]_1), n (%)					
Yes	582 (75.19)	375 (83.15)	207 (64.09)	Ref	__[b]
No	192 (24.80)	76 (16.85)	116 (35.91)	2.76 (1.97-3.86)	<.001
Using smart devices during bedtime causes a delayed sleep schedule (SQPNE_2), n (%)					
Yes	724 (93.54)	432 (95.79)	292 (90.4)	Ref	—
No	50 (6.45)	19 (4.21)	31 (9.6)	2.41 (1.33-4.35)	<.001
Using smart devices during bedtime negatively affects oneself the next morning (SQPNE_3), n (%)					
Yes	561 (72.48)	374 (82.93)	187 (57.89)	Ref	—
No	213 (27.51)	77 (17.07)	136 (42.11)	3.53 (2.53-4.91)	<.001
After highly using the smart device the previous night, health observations the next morning (SQPNE_4) , n (%)					
(1) Headache					
False	425 (54.90)	227 (50.33)	198 (61.3)	Ref	—
True	349 (45.09)	224 (49.67)	125 (38.7)	0.63 (0.47-0.85)	<.001
(2) Dry eyes					
False	508 (65.63)	281 (62.31)	227 (70.28)	Ref	—
True	266 (34.36)	170 (37.69)	96 (29.72)	0.69 (0.51-0.94)	.02
(3) Pain in the ear					
False	687 (88.75)	400 (88.69)	287 (88.85)	Ref	—
True	87 (11.24)	51 (11.31)	36 (11.15)	0.98 (0.62-1.54)	1.00
(4) Fatigue and drowsiness					
False	294 (37.98)	160 (35.48)	134 (41.49)	Ref	—
True	480 (62.01)	291 (64.52)	189 (58.51)	0.77 (0.57-1.04)	.09
(5) Lack of focus and distraction					
False	459 (59.30)	252 (55.88)	207 (64.09)	Ref	—
True	315 (40.69)	199 (44.12)	116 (35.91)	0.70 (0.52-0.95)	.02
(6) Mood swings					
False	458 (59.17)	257 (56.98)	201 (62.23)	Ref	—
True	316 (40.82)	194 (43.02)	122 (37.77)	0.80 (0.60-1.07)	.15
(7) Increased desire to have food rich in sugars and fats					
False	610 (78.81)	345 (76.50)	265 (82.04)	Ref	—
True	164 (21.18)	106 (23.50)	58 (17.96)	0.71 (0.49-1.01)	.07
(8) Nothing					
False	738 (95.34)	443 (98.23)	295 (91.33)	Ref	—
True	36 (4.65)	8 (1.77)	28 (8.67)	5.25 (2.36-11.69)	<.001
Subjective sleep quality , n (%)					
Sleep quality overall					
Very good	166 (21.44)	94 (20.84)	72 (22.29)	Ref	—
Fairly good	338 (43.66)	191 (42.35)	147 (45.51)	1.00 (0.69-1.46)	1.00
Fairly bad	200 (25.83)	118 (26.16)	82 (25.39)	0.90 (0.59-1.37)	0.67
Very bad	70 (9.04)	48 (10.64)	22 (6.81)	0.59 (0.33-1.08)	0.10

XSL•FO
RenderX

[a]SQPNE: sleep quality perceived negative effects.

[b]Not applicable.

Physical Sleep Disturbances Experienced During the Past Month

In terms of sleep latency, the majority of participants (209/774, 27%) reported having trouble falling asleep within 30 minutes 3 or more times per week, as shown in Table 4. Regarding sleep disturbances, waking up in the middle of the night or early morning occurred three or more times per week for 36.17% (280/774) of participants, making it a common problem. Use of sleep medications was not common, as 84.23% (652/774) of participants did not take sleep aids during the past month. Daytime dysfunction was significant for some participants, with 14.47% (112/774) struggling to stay awake 3 or more times per week during daily activities like driving or eating. Also, 31.39% (243/774) of participants reported that maintaining enthusiasm to complete tasks was somewhat of a problem, while 13.82% (107/774) found it to be a very big issue.

About 7.76% (35/451) of the participants who cough or snore loudly 3 or more times per week had significantly lower odds (OR 0.45, 95% CI 0.23-0.88, P=.02) of intending to stop using smart devices compared to those who did not. Similarly, 9.31% (42/451) of those who used sleep medications less than once per week had lower odds (OR 0.49, 95% CI 0.27-0.89, P=.01) of intending to stop using smart devices compared to those who did not use sleep medications. Over a quarter of participants who needed to get up to use the bathroom less than once per week (119/451, 26.39%) or once or twice per week (91/451, 20.18%) had lower odds of intending to stop using smart devices at bedtime compared to those who did not need to get up (OR 0.63, 95% CI 0.43-0.92, P=.01 for less than once per week; OR 0.63, 95% CI 0.42-0.96, P=.03 for once or twice per week). About 21.51% (97/451) of the participants who could not breathe comfortably less than once per week had lower odds (OR 0.67, 95% CI 0.46-0.99, P=.05) of intending to stop using smart devices compared to those who did not have breathing issues. Finally, 26.61% (120/451) of the participants who felt too cold once or twice per week had lower odds (OR 0.62, 95% CI 0.42-0.91, P=.01) of intending to stop using smart devices compared to those who did not feel too cold.

XSL•FO
RenderX

Table 4. Physical sleep disturbances experienced during the past month by university students in Saudi Arabia.

PSQI[a] components	Values (N=774)	Likely to stop using smart devices during bedtime in the next 3 months			
		Yes (n=451)	No (n=323)	Odds ratio (95% CI)	*P* value
Sleep latency, n (%)					
Cannot get to sleep within 30 minutes					
Not during the past month	201 (25.96)	120 (26.61)	81 (25.08)	Ref	__[b]
Less than once per week	173 (22.35)	100 (22.17)	73 (22.60)	1.08 (0.71-1.63)	.75
Once or twice per week	191 (24.67)	110 (24.39)	81 (25.08)	1.09 (0.72-1.63)	.68
Three or more times per week	209 (27.00)	121 (26.83)	88 (27.24)	1.07 (0.72-1.59)	.76
Sleep disturbance, n (%)					
Wake up in the middle of the night or early morning					
Not during the past month	187 (24.16)	101 (22.39)	86 (26.63)	Ref	—
Less than once per week	125 (16.14)	79 (17.52)	46 (14.24)	0.68 (0.43-1.08)	.12
Once or twice per week	182 (23.51)	102 (22.62)	80 (24.77)	0.92 (0.61-1.38)	.75
Three or more times per week	280 (36.17)	169 (37.47)	111 (34.37)	0.77 (0.53-1.12)	.18
Have to get up use to the bathroom (SQDPM[c]_1)					
Not during the past month	287 (37.08)	150 (33.26)	137 (42.41)	Ref	—
Less than once per week	188 (24.28)	119 (26.39)	69 (21.36)	0.63 (0.43-0.92)	.01
Once or twice per week	144 (18.60)	91 (20.18)	53 (16.41)	0.63 (0.42-0.96)	.03
Three or more times per week	155 (20.02)	91 (20.18)	64 (19.81)	0.77 (0.51-1.14)	.22
Cannot breathe comfortably (SQDPM_2)					
Not during the past month	453 (58.52)	251 (55.65)	202 (62.54)	Ref	—
Less than once per week	150 (19.37)	97 (21.51)	53 (16.41)	0.67 (0.46-0.99)	.05
Once or twice per week	94 (12.14)	56 (12.42)	38 (11.76)	0.84 (0.53-1.32)	.49
Three or more times per week	77 (9.94)	47 (10.42)	30 (9.29)	0.79 (0.48-1.29)	.38
Cough or snore loudly (SQDPM_3)					
Not during the past month	580 (74.93)	330 (73.17)	250 (77.40)	Ref	—
Less than once per week	88 (11.36)	51 (11.31)	37 (11.46)	0.95 (0.60-1.50)	.90
Once or twice per week	59 (7.62)	35 (7.76)	24 (7.43)	0.90 (0.52-1.56)	.78
Three or more times per week	47 (6.07)	35 (7.76)	12 (3.72)	0.45 (0.23-0.88)	.02
Feel too cold (SQDPM_4)					
Not during the past month	262 (33.85)	138 (30.60)	124 (38.39)	Ref	—
Less than once per week	144 (18.60)	88 (19.51)	56 (17.34)	0.70 (0.46-1.07)	.11
Once or twice per week	187 (24.16)	120 (26.61)	67 (20.74)	0.62 (0.42-0.91)	.01
Three or more times per week	181 (23.38)	105 (23.28)	76 (23.53)	0.80 (0.54-1.18)	.28
Feel too hot					
Not during the past month	317 (40.95)	179 (39.69)	138 (42.72)	Ref	—
Less than once per week	175 (22.60)	104 (23.06)	71 (21.98)	0.88 (0.60-1.28)	.56
Once or twice per week	165 (21.31)	97 (21.51)	68 (21.05)	0.90 (0.62-1.33)	.69
Three or more times per week	117 (15.11)	71 (15.74)	46 (14.24)	0.84 (0.54-1.29)	.44
Have bad dreams					
Not during the past month	271 (35.01)	155 (34.37)	116 (35.91)	Ref	—
Less than once per week	241 (31.13)	139 (30.82)	102 (31.58)	0.98 (0.69-1.39)	.92

XSL•FO
RenderX

PSQI[a] components	Values (N=774)	Likely to stop using smart devices during bedtime in the next 3 months			
		Yes (n=451)	No (n=323)	Odds ratio (95% CI)	*P* value
Once or twice per week	170 (21.96)	98 (21.73)	72 (22.29)	0.98 (0.66-1.44)	>.99
Three or more times per week	92 (11.88)	59 (13.08)	33 (10.22)	0.74 (0.45-1.21)	.27
Have pain					
Not during the past month	381 (49.22)	213 (47.23)	168 (52.01)	Ref	—
Less than once per week	156 (20.15)	90 (19.96)	66 (20.43)	0.92 (0.63-1.35)	.77
Once or twice per week	125 (16.14)	78 (17.29)	47 (14.55)	0.76 (0.50-1.15)	.21
Three or more times per week	112 (14.47)	70 (15.52)	42 (13.00)	0.76 (0.49-1.17)	.23
Use of sleep medications, n (%)					
Taken medicine to help you sleep (prescribed or "over the counter") (SQDPM_5)					
Not during the past month	652 (84.23)	367 (81.37)	285 (88.24)	Ref	—
Less than once per week	58 (7.49)	42 (9.31)	16 (4.95)	0.49 (0.27-0.89)	.01
Once or twice per week	37 (4.78)	26 (5.76)	11 (3.41)	0.54 (0.26-1.12)	.12
Three or more times per week	27 (3.48)	16 (3.55)	11 (3.42)	0.88 (0.40-1.93)	.84
Daytime dysfunction, n (%)					
Having trouble staying awake while driving, eating meals, or engaging in social activity					
Not during the past month	369 (47.67)	206 (45.68)	163 (50.46)	Ref	—
Less than once per week	161 (20.80)	103 (22.84)	58 (17.96)	0.71 (0.48-1.04)	.08
Once or twice per week	132 (17.05)	69 (15.30)	63 (19.50)	1.15 (0.77-1.71)	.54
Three or more times per week	112 (14.47)	73 (16.19)	39 (12.07)	0.67 (0.43-1.04)	.08
Keeping up enough enthusiasm to get things done					
No problem at all	202 (26.09)	112 (24.83)	90 (27.86)	Ref	—
Only a very slight problem	222 (28.68)	137 (30.38)	85 (26.32)	0.77 (0.52-1.13)	.20
Somewhat of a problem	243 (31.39)	137 (30.38)	106 (32.82)	0.96 (0.66-1.40)	.84
A very big problem	107 (13.82)	65 (14.41)	42 (13.00)	0.80 (0.49-1.29)	.39

[a]PSQI: Pittsburgh Sleep Quality Index.

[b]Not applicable.

[c]SQDPM: sleep quality during the past month.

Predictors of the Intention to Stop Using Smart Devices at Bedtime

Path analysis was conducted to understand the complex relationships between the measured variable intended to stop using smart device usage and the latent variables, which were SQSU (presented in Table 2), SQPNE (presented in Table 3), and SQDPM (presented in Table 4). All the demographic factors in Table 1 were removed from the analysis as they did not approach the significance threshold of <.05. To improve the model fit, only the significant variables (<.05) from Tables 2-4 were included. SEM plot indicated that the model had an excellent fit with a nonsignificant chi-square, a CFI of 0.845, and a RMSEA value of 0.039 (95% CI 0.032-0.046).

All the factors except SQSU_7 (purposes=–0.09, *P*>.05) had significant factor loadings (path coefficients) with latent variables SQSU, SQPNE, and SQDPM. People using smart devices every night (SQSU_1; 0.36, *P*<.05), 3 hours or more time spent on smart devices at bedtime (SQSU_2; –0.35, *P*<.05), using the smart device frequently after lights were turned off (SQSU_3; 0.49, *P*<.05), using headphones frequently during bedtime (SQSU_4; 0.39, *P*<.05), awakened frequently by the smart device at night (SQSU_5; 0.25, *P*<.05), activities (SQSU_6; –0.18, *P*<.05) had significant path coefficients with latent variable SQSU.

People who perceived using smart devices during bedtime led to sleeplessness (SQPNE_1; 0.58, *P*<.05), caused delayed sleep schedule (SQPNE_2; 0.29, *P*<.05), negatively affected oneself the next morning (SQPNE_3; 0.71, *P*<.05), and affected their health (SQPNE_4; 0.24, *P*<.05) had significant path coefficients with latent variable SQPNE.

People woke up in the middle of the night or early morning, had to get up to use the bathroom (SQDPM_1; 0.35, *P*<.05), could not breathe comfortably (SQDPM_2; 0.64, *P*<.05), coughed or snored loudly (SQDPM_3; 0.41, *P*<.05), felt cold (SQDPM_4; 0.42, *P*<.05), and took medicine to help sleep

XSL•FO
RenderX

(SQDPM_5; 0.33, $P<.05$) had significant path coefficients with latent variable SQDPM.

The 3 obtained factors have path coefficients with ISSU (intended to stop using smart device usage). SQSU=–0.40, SQPNE=0.34, and SQDPM=–0.19 were significant and there was no evidence of collinearity in the model as shown in Figure 1. All items of the model had significant paths as well as correlations with respective factor domains. Factor structures remained intact, and all paths were significant.

Figure 1. Path analysis of predictors influencing university students' intention to stop using smart devices at bedtime in Saudi Arabia. An asterisk indicates a 2-tailed *P* value <.05. ISSU: intended to stop using smart device usage; SQDPM: sleep quality during the past month; SQPNE: sleep quality perceived negative effects; SQSU: sleep quality smartphone usage.

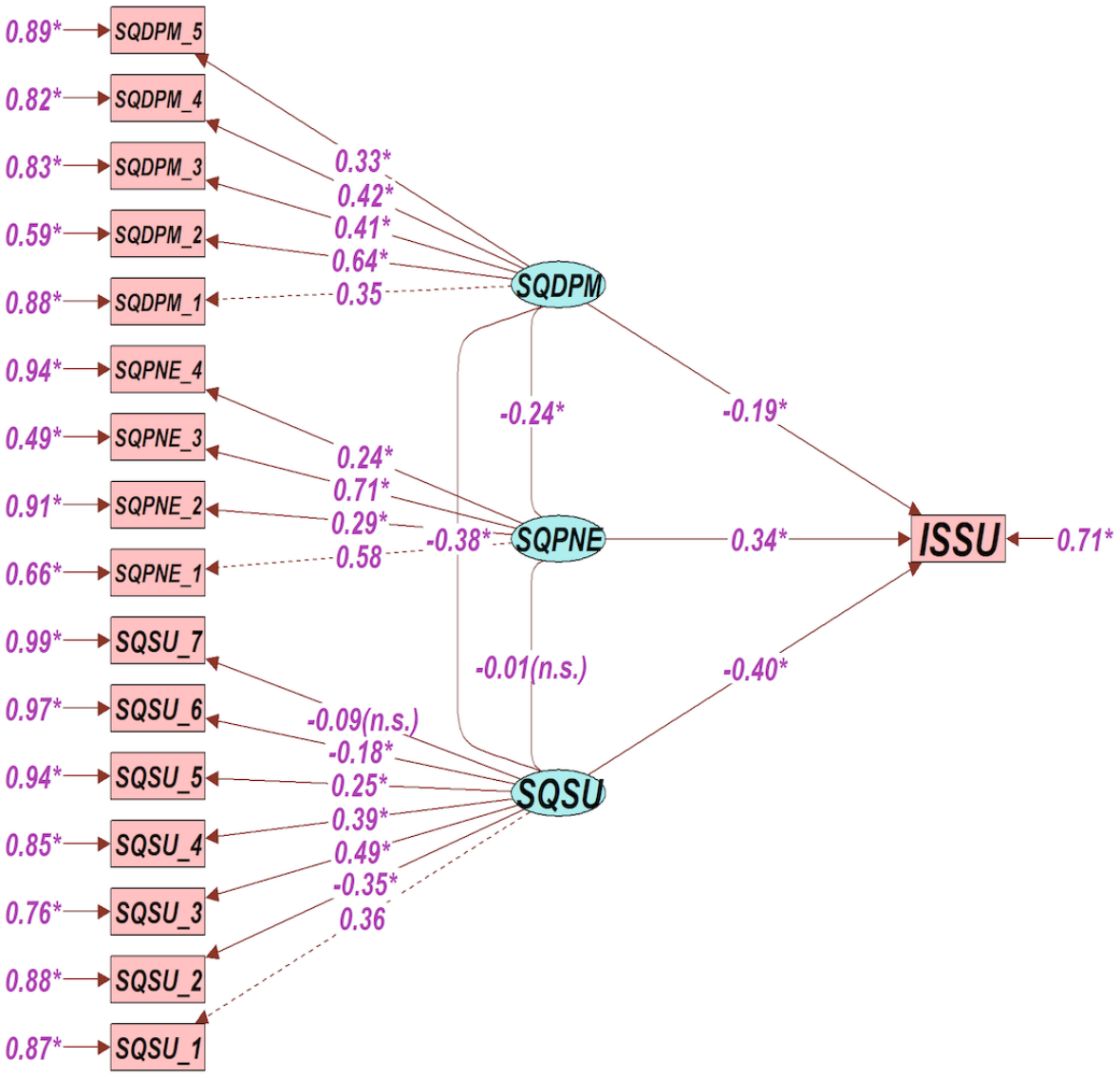

Discussion

Principal Findings

This study investigated the factors that influence university students' intention to stop using smart devices at bedtime in Saudi Arabia. It focused on behavioral habits, perceived negative effects on sleep quality, and physical sleep disturbances. The findings provide insights into how smartphone usage behaviors and sleep-related perceptions interplay, influencing students' intentions to modify their presleep device habits, as follows.

Behavioral Factors and Smart Device Usage Habits

The results from the analysis of smartphone usage habits (Table 2) revealed that frequent nightly use and the use of smart devices after lights are turned off were significantly associated with the intention to stop using these devices at bedtime. Specifically, using smart devices after lights out (path coefficient=0.49) was the strongest behavioral predictor, indicating that students who

XSL•FO
RenderX

engage in late-night device use may recognize its disruptive effects on sleep and are thus more inclined to reduce or stop this behavior. This aligns with previous studies that highlight the adverse effects of presleep exposure to light from screens, which delays sleep onset [13,17].

However, spending more than 3 hours on devices (path coefficient=–0.35) and engaging in multiple activities (path coefficient=–0.18) were negatively associated with the intention to stop using smart devices at bedtime. This suggests that heavy users may be more resistant to behavioral change, possibly due to their addiction to using smart devices [1]. This aligns with previous studies that have identified higher levels of smartphone addiction as barriers to changing screen time habits [27,28].

Moreover, engaging in more activities on the device also decreased the likelihood of stopping using smart devices (path coefficient=–0.18), which may indicate that participants perceive these activities (eg, watching videos and chatting) as integral parts of their nightly routine [29,30]. These results are consistent with previous studies that link high-frequency smart device use to a reluctance to disengage from these devices, possibly due to increased dependency on smart devices for both academic and social purposes [27,31].

Perceived Negative Effects on Sleep Quality and Well-Being

The most significant predictors of the intention to stop using smart devices were related to perceived negative effects on sleep quality and next-morning well-being (Table 3). The strongest predictor was the belief that smart device usage negatively affects next-morning well-being (path coefficient=0.71). This suggests that students who experience morning fatigue, mood disturbances, or difficulty focusing are more likely to stop using their devices before sleep. This is consistent with previous studies that demonstrate how poor sleep quality directly influences daytime functioning [19], especially in academic contexts where cognitive performance is critical [15,30,31].

Similarly, the perception of sleeplessness due to smart device use was also a strong predictor (path coefficient=0.58). This indicates that university students who struggle to sleep because of their devices are more likely to intend to stop using them. This is consistent with previous studies on the disruptive effects of presleep screen time, which has been linked to increased sleep latency and poorer overall sleep quality [13,32]. In addition, the perception of delayed sleep schedules (path coefficient=0.29) also contributed to the intention to stop, further reinforcing the idea that students recognize the role of smart devices in disrupting their sleep timing.

Physical Sleep Disturbances and Health Implications

The results from physical sleep disturbances (Table 4) highlighted how certain health issues experienced during the night are significantly linked with the intention to stop using smart devices. Difficulty breathing comfortably during sleep (path coefficient=0.64) was the most prominent physical predictor, suggesting that students who experience physical discomforts, such as respiratory issues or muscle tension, may attribute these problems to their use of smart devices at bedtime and are therefore more likely to reduce their usage. This is consistent with studies that have found links between prolonged device use in bed and poor posture or breathing discomfort, potentially due to prolonged screen time in awkward positions [33].

Other sleep disturbances, such as waking up during the night, snoring or coughing, and feeling too cold, were also positively associated with the intention to stop using smart devices at bedtime. These findings indicate that students who experience sleep disruptions are more likely to modify their behavior to improve their sleep quality [8,20]. Furthermore, students who reported taking sleep medications were more likely to intend to stop using smart devices, suggesting that those with more severe sleep problems may be more motivated to address contributing factors like bedtime device use [31].

Implications for University Students and Public Health Interventions

These findings have significant implications for interventions targeting sleep hygiene among university students, particularly in Saudi Arabia. Given that the strongest predictors of the intention to stop using smart devices were perceived negative effects on next-morning well-being and physical sleep disturbances, public health initiatives should focus on raising awareness about these health risks [34]. Particularly, educational campaigns should emphasize how devices used before bed negatively impact not only sleep quality but also cognitive functioning, mood, and overall well-being the next day [35].

Interventions should also incorporate behavior change theories, such as the Transtheoretical Model and the Theory of Planned Behavior [36,37], to understand and address resistance to change. Many heavy device users, despite recognizing negative outcomes, may rely on their devices for emotional regulation or social engagement [5,16], making these activities a barrier to change. Tailored strategies should therefore target these specific behaviors, offering alternative coping mechanisms like mindfulness exercises, relaxation techniques [38], or structured bedtime routines to gradually reduce dependency on devices [39]. Furthermore, strategies that encourage limiting screen time after lights out and reducing device-related disruptions, such as turning off notifications or using blue light filters, could be effective in promoting healthier sleep habits [32,34].

Strengths, Limitations, and Future Directions

This study has several strengths. First, its multidimensional approach examined not only behavioral habits but also perceived negative effects and physical sleep disturbances, offering a holistic view of the factors that influence the intention to stop using smart devices. Second, the use of a validated scale such as PSQI adds rigor to the assessment of sleep quality. Finally, the application of path analysis and SEM allowed for an in-depth analysis of the relationships between latent variables, supporting the interpretation of complex patterns.

However, there are some limitations. The cross-sectional design limits the ability to establish causality between smart device usage and sleep disturbances [40].

The reliance on self-reported data introduces potential biases, including recall bias (where participants may not accurately

XSL•FO
RenderX

remember behaviors or experiences) and social desirability bias (where participants may provide what they perceive to be more acceptable responses). These biases could influence the accuracy of the findings [41]. Furthermore, the open recruitment approach via social media platforms prevented determining an exact participation rate [42]. Although a large sample (n=774) of university students from Saudi Arabia may improve the applicability of the findings to this population, replication in other cultural contexts is necessary to assess broader generalizability and refine intervention strategies.

Future studies should consider longitudinal research to establish causality and interventional studies to test specific strategies for reducing device use at bedtime [16,26]. Also, investigating different types of smart device use (eg, browsing social media, gaming, chatting, and studying) would provide more nuanced insights. In addition, future studies should investigate the interactions between smart device use, sleep hygiene, and mental health such as anxiety, depression, or stress levels because understanding these interactions could inform more tailored interventions for improving sleep among university students [5].

Conclusions

This study provides valuable insights into the factors influencing university students' intention to stop using smart devices at bedtime, particularly in the Saudi Arabian context. While behavioral habits, such as frequent device use and activities, are important, the most significant predictors of behavior change are perceived negative effects on well-being and physical sleep disturbances. These findings highlight the importance of addressing both the psychological and physiological impacts of smart device use in interventions aimed at improving sleep quality among university students. Future studies should explore how these interventions can be effectively implemented and sustained over time to mitigate the negative effects of excessive bedtime device use.

Data Availability

The datasets generated during and/or analyzed during this study are available from the author on reasonable request.

Conflicts of Interest

None declared.

References

1. Leow MQH, Chiang J, Chua TJX, Wang S, Tan NC. The relationship between smartphone addiction and sleep among medical students: a systematic review and meta-analysis. PLoS One 2023;18(9):e0290724 [FREE Full text] [doi: 10.1371/journal.pone.0290724] [Medline: 37713408]
2. Smartphone penetration as share of population in Saudi Arabia 2020-2029. Statista. URL: https://www.statista.com/statistics/625436/smartphone-user-penetration-in-saudi-arabia [accessed 2024-02-21]
3. Smartphone mobile network subscriptions worldwide 2016-2028. Statista. URL: https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/ [accessed 2025-01-02]
4. Zhao C, He J, Xu H, Zhang J, Zhang G, Yu G. Are 'night owls' or 'morning larks' more likely to delay sleep due to problematic smartphone use? A cross-lagged study among undergraduates. Addict Behav 2024;150:107906. [doi: 10.1016/j.addbeh.2023.107906] [Medline: 37984222]
5. Yang J, Fu X, Liao X, Li Y. Association of problematic smartphone use with poor sleep quality, depression, and anxiety: a systematic review and meta-analysis. Psychiatry Res 2020;284:112686. [doi: 10.1016/j.psychres.2019.112686] [Medline: 31757638]
6. Nakie G, Takelle GM, Rtbey G, Andualem F, Tinsae T, Kassa MA, et al. Sleep quality and associated factors among university students in Africa: a systematic review and meta-analysis study. Front Psychiatry 2024;15:1370757 [FREE Full text] [doi: 10.3389/fpsyt.2024.1370757] [Medline: 38559402]
7. Pagano M, Bacaro V, Crocetti E. 'Using digital media or sleeping … that is the question'. A meta-analysis on digital media use and unhealthy sleep in adolescence. Comput Hum Behav 2023;146:107813. [doi: 10.1016/j.chb.2023.107813]
8. Putra Pratama Wau R, Wienaldi W, Dharma S. Relationship between smartphone use and sleep quality of universitas prima Indonesia students. Int J Soc Health 2024;3(1):31-37. [doi: 10.58860/ijsh.v3i1.146]
9. Han X, Zhou E, Liu D. Electronic media use and sleep quality: updated systematic review and meta-analysis. J Med Internet Res 2024;26:e48356 [FREE Full text] [doi: 10.2196/48356] [Medline: 38533835]
10. Dagani J, Buizza C, Cela H, Sbravati G, Rainieri G, Ghilardi A. The interplay of sleep quality, mental health, and sociodemographic and clinical factors among Italian college freshmen. J Clin Med 2024;13(9):2626 [FREE Full text] [doi: 10.3390/jcm13092626] [Medline: 38731154]
11. M. Alfawareh H, Jusoh S. Smartphones usage among university students: Najran University case. Int J Acad Res 2014;6(2):321-326.
12. Alzhrani AM, Aboalshamat KT, Badawoud AM, Abdouh IM, Badri HM, Quronfulah BS, et al. The association between smartphone use and sleep quality, psychological distress, and loneliness among health care students and workers in Saudi Arabia. PLoS One 2023;18(1):e0280681 [FREE Full text] [doi: 10.1371/journal.pone.0280681] [Medline: 36701337]

XSL•FO
RenderX

13. Randjelović P, Stojanović N, Ilić I, Vučković D. The effect of reducing blue light from smartphone screen on subjective quality of sleep among students. Chronobiol Int 2023;40(3):335-342. [doi: 10.1080/07420528.2023.2173606] [Medline: 36744480]
14. Elsheikh AA, Elsharkawy SA, Ahmed DS. Impact of smartphone use at bedtime on sleep quality and academic activities among medical students at Al-Azhar University at Cairo. J Public Health (Berl) 2023;32(11):2091-2100 [FREE Full text] [doi: 10.1007/s10389-023-01964-8]
15. Liu H, Váchová L, Plevová I. Effect of mobile phone dependence on various aspects of academic achievement: evidence from Chinese and Czech university students. Pegem Journal of Education and Instruction 2024;14(2):20-26. [doi: 10.47750/pegegog.14.02.03]
16. Cemei L, Sriram S, Holý O, Rehman S. A longitudinal investigation on the reciprocal relationship of problematic smartphone use with bedtime procrastination, sleep quality, and mental health among university students. Psychol Res Behav Manag 2024;17:3355-3367 [FREE Full text] [doi: 10.2147/PRBM.S472299] [Medline: 39359420]
17. Laghari I, Khan A, Hussain S, Waheed S. Relationship between screen-time, sleeping pattern and stress during the pandemic among university students. Int J Soc Sci Arch 2023;6(3).
18. Ji Z, Zhou L. A problem not to be ignored: the influencing factors of mobile phone addiction and its influence on sleep quality. Health 2024;16(5):459-469.
19. Wang Q, Zhong Y, Zhao G, Song R, Zeng C. Relationship among content type of smartphone use, technostress, and sleep difficulty: a study of university students in China. Educ Inf Technol 2022;28(2):1697-1714. [doi: 10.1007/s10639-022-11222-1]
20. Hu B, Shen W, Wang Y, Wu Q, Li J, Xu X, et al. Prevalence and related factors of sleep quality among Chinese undergraduates in Jiangsu province: multiple models' analysis. Front Psychol 2024;15:1343186 [FREE Full text] [doi: 10.3389/fpsyg.2024.1343186] [Medline: 38659684]
21. Ajzen I. The theory of planned behavior. Organ Behav Hum Decis Process 1991;50(2):179-211.
22. Vandenbroucke JP, von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, STROBE Initiative. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. Int J Surg 2014;12(12):1500-1524 [FREE Full text] [doi: 10.1016/j.ijsu.2014.07.014] [Medline: 25046751]
23. Alshalawi AS. Social media usage intensity and academic performance among undergraduate students in Saudi Arabia. Contemp Educ Technol 2022;14(2):ep361. [doi: 10.30935/cedtech/11711]
24. Number of social media users in Saudi Arabia from 2020 to 2029. Statista. URL: https://www.statista.com/forecasts/1144943/social-media-users-in-saudi-arabia [accessed 2024-05-13]
25. Buysse DJ, Reynolds CF, Monk TH, Berman SR, Kupfer DJ. The Pittsburgh sleep quality index: a new instrument for psychiatric practice and research. Psychiatry Res 1989;28(2):193-213. [doi: 10.1016/0165-1781(89)90047-4] [Medline: 2748771]
26. Suleiman KH, Yates BC, Berger AM, Pozehl B, Meza J. Translating the Pittsburgh sleep quality index into Arabic. West J Nurs Res 2010;32(2):250-268. [doi: 10.1177/0193945909348230] [Medline: 19915205]
27. Lane HY, Chang CJ, Huang CL, Chang YH. An investigation into smartphone addiction with personality and sleep quality among university students. Int J Environ Res Public Health 2021;18(14):7588 [FREE Full text] [doi: 10.3390/ijerph18147588] [Medline: 34300037]
28. Cui G, Yin Y, Li S, Chen L, Liu X, Tang K, et al. Longitudinal relationships among problematic mobile phone use, bedtime procrastination, sleep quality and depressive symptoms in Chinese college students: a cross-lagged panel analysis. BMC Psychiatry 2021;21(1):449 [FREE Full text] [doi: 10.1186/s12888-021-03451-4] [Medline: 34507561]
29. Pillion M, Gradisar M, Bartel K, Whittall H, Kahn M. What's 'app'-ning to adolescent sleep? Links between device, app use, and sleep outcomes. Sleep Med 2022;100:174-182. [doi: 10.1016/j.sleep.2022.08.004] [Medline: 36084495]
30. Kumar Swain R, Pati AK. Use of social networking sites (SNSs) and its repercussions on sleep quality, psychosocial behavior, academic performance and circadian rhythm of humans – a brief review. Biological Rhythm Research 2019;52(8):1139-1178. [doi: 10.1080/09291016.2019.1620487]
31. Rathakrishnan B, Bikar Singh SS, Kamaluddin MR, Yahaya A, Mohd Nasir MA, Ibrahim F, et al. Smartphone addiction and sleep quality on academic performance of university students: an exploratory research. Int J Environ Res Public Health 2021;18(16):8291 [FREE Full text] [doi: 10.3390/ijerph18168291] [Medline: 34444042]
32. Huiberts LM, Opperhuizen AL, Schlangen LJM. Pre-bedtime activities and light-emitting screen use in university students and their relationships with self-reported sleep duration and quality. Light Res Technol 2022;54(6):595-608. [doi: 10.1177/14771535221074725]
33. Sirajudeen M, Alzhrani M, Alanazi A, Alqahtani M, Waly M, Unnikrishnan R, et al. Prevalence of text neck posture, smartphone addiction, and its association with neck disorders among university students in the Kingdom of Saudi Arabia during the COVID-19 pandemic. PeerJ 2022;10:e14443 [FREE Full text] [doi: 10.7717/peerj.14443] [Medline: 36540801]
34. Pham HT, Chuang H, Kuo C, Yeh T, Liao W. Electronic device use before bedtime and sleep quality among university students. Healthcare (Basel) 2021;9(9):1091 [FREE Full text] [doi: 10.3390/healthcare9091091] [Medline: 34574865]

XSL•FO
RenderX

35. Andersen TO, Dissing AS, Varga TV, Rod NH. The SmartSleep experiment: Evaluation of changes in night-time smartphone behavior following a mass media citizen science campaign. PLoS One 2021;16(7):e0253783 [FREE Full text] [doi: 10.1371/journal.pone.0253783] [Medline: 34288929]
36. Pegado A, Alvarez M, Roberto MS. The role of behaviour-change theory in sleep interventions with emerging adults (aged 18-29 years): a systematic review and meta-analysis. J Sleep Res 2023;32(5):e13877. [doi: 10.1111/jsr.13877] [Medline: 36922157]
37. Knowlden A. Models and theories of behavior change relevant to sleep health. In: Sleep and Health. United States: Academic Press; 2019:171-186.
38. Kravcenko B. The effect of a six-week mindfulness meditation stress reduction course on the effect of mobile phone use on sleep quality in master students at the University of Twente. Netherlands: Bachelor's thesis, University of Twente; 2023. URL: https://essay.utwente.nl/94266/ [accessed 2025-02-12]
39. Kheirinejad S, Visuri A, Ferreira D, Hosio S. 'Leave your smartphone out of bed': quantitative analysis of smartphone use effect on sleep quality. Pers Ubiquitous Comput 2023;27(2):447-466 [FREE Full text] [doi: 10.1007/s00779-022-01694-w] [Medline: 36405389]
40. Savitz D, Wellenius G. Can cross-sectional studies contribute to causal inference? It depends. Am J Epidemiol 2023;192(4):514-516. [doi: 10.1093/aje/kwac037] [Medline: 35231933]
41. van der Ham IJM, van der Kuil MNA, Claessen MHG. Quality of self-reported cognition: effects of age and gender on spatial navigation self-reports. Aging Ment Health 2021;25(5):873-878 [FREE Full text] [doi: 10.1080/13607863.2020.1742658] [Medline: 32233799]
42. Jones A, Walters J, Brown A. Participant recruitment in social work: a social media approach. Soc. Work Res 2020;44(4):247-255.

Abbreviations

CFI: comparative fit index
OR: odds ratio
PSQI: Pittsburgh Sleep Quality Index
RMSEA: root mean square error of approximation
SEM: sequencing equation modeling
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
SQDPM: sleep quality during the past month
SQPNE: sleep quality perceived negative effects
SQSU: sleep quality smartphone usage

Edited by A Mavragani; submitted 05.10.24; peer-reviewed by I Wilson, WJ He; comments to author 20.12.24; revised version received 24.12.24; accepted 07.01.25; published 10.03.25.

Please cite as:
Almalki M
Predictors of the Intention to Stop Using Smart Devices at Bedtime Among University Students in Saudi Arabia: Cross-Sectional Survey
JMIR Form Res 2025;9:e67223
URL: https://formative.jmir.org/2025/1/e67223
doi:10.2196/67223
PMID:40063070

©Manal Almalki. Originally published in JMIR Formative Research (https://formative.jmir.org), 10.03.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Impact of a Sensorimotor Integration and Hyperstimulation Program on Global Motor Skills in Moroccan Children With Autism Spectrum Disorder: Exploratory Clinical Quasi-Experimental Study

Rachid Touali[1], MSci; Jamal Zerouaoui[1], PhD; El Mahjoub Chakir[1], PhD; Hung Tien Bui[2], Dr med, PhD; Mario Leone[2], PhD; Maxime Allisse[3], PhD

[1]Faculty of Sciences, University of Ibn Tofail, Kénitra, Morocco

[2]Faculty of Medicine and Health Sciences, University of Sherbrooke, Sherbrooke, QC, Canada

[3]Faculty of Physical Activity Sciences, University of Sherbrooke, Sherbrooke, QC, Canada

Corresponding Author:
Maxime Allisse, PhD
Faculty of Physical Activity Sciences
University of Sherbrooke
2500 Boulevard de l'Université
Sherbrooke, QC, J1K 2R1
Canada
Phone: 1 819 821 8000 ext 63875
Email: maxime.allisse@usherbrooke.ca

Abstract

Background: Children with autism spectrum disorders (ASDs) often struggle with processing information, which can impact their coordination, balance, and other motor skills. Studies have demonstrated that intervention programs based on sensory integration can enhance motor performance in these children.

Objective: The objective of this study is to evaluate the applicability of a standardized battery of gross motor skill tests for Moroccan children aged 6 to 12 years with ASD. The objective is to assess the potential efficacy of an innovative pedagogical approach focused on sensorimotor integration and hyperstimulation. This approach will be compared to traditional physical education (PE) sessions to determine its feasibility and potential to bridge the developmental gaps in motor skills between children with ASD and those with a neurotypical profile.

Methods: A convenience sample of 14 Moroccan children with ASD aged 6 to 12 years participated in this exploratory study. Children with ASD were divided into an experimental group (n=7) and a control group (n=7) based on age, sex, motor performance, and socioeconomic status. The control group followed the standard PE program, while the experimental group underwent a specialized program combining sensorimotor integration and hyperstimulation for a period of 15 weeks. All participants were classified as level 2 (moderate) on the Autism Severity Rating Scale based on the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, Text Revision* (DSM-5-TR) criteria. Gross motor skills were measured at baseline and after 15 weeks of intervention using the UQAC-UQAM (University of Québec in Chicoutimi-University of Québec in Montréal) test battery protocol, which includes 10 items.

Results: At baseline (T1), no significant difference was observed between the control and experimental groups of children with ASD. Following the 15-week intervention, the group participating in traditional PE showed an overall improvement in motor skills of approximately 14.5%. Conversely, the results of the ASD experimental group suggest a more substantial improvement of 44.5%. Additionally, the experimental group exhibited significant better performance across all motor skill variables compared to the control group (minimum *P* values of <.02) with large effect sizes (>0.80). In this regard, a 2-way repeated measures ANOVA confirms the efficiency of the program implemented within the experimental group, demonstrating significant effects associated with both group and time factors as well as a clinically highly significant group×time interaction across all measured variables ($\eta^2p>0.14$).

Conclusions: The results of this study suggest that the approach that emphasizes sensorimotor integration and management of hyperstimulation was more effective in improving motor skills in this population. However, other more exhaustive studies will need to be carried out in order to be able to more precisely measure the full potential of this approach.

XSL•FO
RenderX

(JMIR Form Res 2025;9:e65767) doi:10.2196/65767

KEYWORDS

classical physical education; children with a neurotypical profile; children with ASD; UQAC-UQAM test battery; University of Québec in Chicoutimi-University of Québec in Montréal; sensorimotor integration; hyperstimulation; Morocco; sensorimotor; integration; motor skill; Moroccan children; Moroccan; children; autism spectrum disorder; ASD; exploratory study; autism; mental health; young; youth; feasibility

Introduction

Background

Autism spectrum disorders (ASDs) constitute a subset of developmental disorders distinguished by disturbances in communication, changes in social interactions, and repetitive and stereotyped behaviors [1,2]. These symptoms typically emerge before the age of 3 years and result in functional impairments impacting various aspects of daily life [3,4]. Due to the absence of universally recognized biological markers for identifying ASD cases, diagnosis primarily involves evaluation by professionals who assess children's development trajectories to identify potential developmental disorders [2]. Among the myriad of challenges faced by children with ASD, developmental coordination disorders (DCDs) occupy an important place [1]. Despite being relatively common, these motor difficulties have historically been overlooked [5-7]. However, recent research [8-10] sheds light on the significant prevalence of DCD in children with ASD. In fact, DCD, also known as dyspraxia, is a neurodevelopmental disorder that affects motor coordination and skills. Studies have shown that DCD is remarkably prevalent among individuals with ASD, affecting between 34% and 79% of this population. This is significantly higher than the 6% prevalence rate observed in the general population [11]. This high prevalence of DCD in individuals with ASD suggests that a lack of coordination is a common characteristic of the disorder. Motor deficits can present in diverse forms, encompassing compromised gross motor skills such as running; challenges in fine motor skills like handwriting; difficulties in motor planning and execution leading to problems in coordinating movements and sequencing actions seamlessly; and sensory processing issues such as hypersensitivity or hyposensitivity to sensory inputs, which impact balance, coordination, and motor performance [12-14]. As a result, there is widespread recognition that a deficiency in coordination stands as a prominent trait of ASD, frequently concomitant with motor deficits. The potential etiology of these challenges encompasses both genetic and environmental factors [15].

Motor Skill Impairment

Typically, children with ASD often face challenges in motor skills, especially in coordination, balance, and postural control [16-19]. A recent meta-analysis even suggests that these deficits extend to decreased walking speed compared to their peers with a neurotypical profile [20]. However, the underlying causes of these impairments remain a topic of debate within the scientific community [21-24]. To address these motor challenges and improve skill development, various interventions have been proposed. These interventions draw inspiration from diverse areas, including sports disciplines [25-27], physical conditioning programs [28], sensorimotor approaches [29], and exercises focused on fundamental motor skills like running, jumping, throwing, and catching [30]. Additionally, research has explored interventions targeting perceptual-motor skills [31]. However, the authors themselves acknowledged the limitations of their programs and called for the development of more precise and specialized interventions. While most studies suggest improvements in motor performance, none, to our knowledge, have directly compared the effects of a general intervention program to one specifically tailored to the needs of Moroccan children with ASD. Such a comparison, focusing on their specific motor difficulties, would be a valuable contribution to the field.

Sensorimotor Integration and Hyperstimulation Approach for Children With ASD

The field of sensory integration, inspired by the work of Ayres and Robbins [32], offers a valuable approach to improve motor skills in children with neurodevelopmental disorders. This approach hinges on the brain's ability to organize and interpret information received through the senses (sensory integration). It uses 2 main practices: passive unisensory and active multisensory. Passive unisensory practices involve targeted stimulation of a single sense at a time. These interventions, often used to promote relaxation and emotional regulation, can include massage for muscle relaxation and improved circulation and body pressures to calm the nervous system and reduce anxiety, and weighted vests or blankets to provide deep, comforting pressure. The simplicity and calming effect of these techniques make them particularly beneficial for individuals needing to regulate their emotions. The active multisensory approach stands in stark contrast. Here, individuals actively participate in activities that engage multiple senses at once. This approach aims to cultivate a diverse range of skills through tactile activities (exploring textures), proprioceptive activities (understanding body position and movement), and vestibular activities (developing balance and coordination). By actively engaging individuals in these stimulating activities, the active approach fosters a more comprehensive integration of sensory information. This translates to improved motor skills, coordination, and sensory regulation abilities. The emphasis on exploration and interaction with the environment makes active multisensory practices particularly beneficial for children or individuals with sensory processing difficulties.

At the turn of the 1990s, Polatajko et al [33] explored an approach partially similar to the sensorimotor and hyperstimulation method, working with a group of children diagnosed with DCD. However, when comparing their findings with those of other studies using similar protocols, they reported significantly more mixed results. According to the authors, these limited outcomes may be attributed to the considerable heterogeneity among children with DCD. Indeed, various

XSL•FO
RenderX

medical conditions can lead to this disorder, despite having distinct etiological origins. In light of this, the research team advocates for more tailored clinical interventions that take into account the specific characteristics and underlying causes of DCD.

Children with ASD often struggle with processing this information, which can impact their coordination, balance, and other motor skills. Studies have demonstrated that intervention programs based on sensory integration can enhance motor performance in children. Notably, improvements have been observed in postural control, gait, coordination, and hand movements [34,35]. For instance, a systematic review of the effectiveness of interventions using a sensory integrative approach confirmed that these programs can significantly positively affect children's motor skills [36]. The sensorimotor integration and hyperstimulation approach has been used in children with ASD, particularly those aged 2 to 12 years [37]. These interventions are structured to include a variety of sensory and motor exercises, such as balance activities, coordination games involving crossover movements, and sensory activities like manipulating different textures. These activities are designed to stimulate various senses in a controlled manner, helping children better manage their movements and posture.

Research Gap

Despite the recent surge in research, some countries, like Morocco, lack precise information on the motor skill profile of their population with ASD. Although numerous studies suggest improvements in motor performance through sensory integration [34-36], no studies, to our knowledge, directly compare the effects of a general intervention program with those of a program specifically based on a sensorimotor integration and hyperstimulation approach. This lack of direct comparison underscores the need for further research to determine the relative effectiveness of these different approaches [38]. This gap is crucial as some authors highlight the considerable variation in children's motor skills across different countries [39,40]. For instance, studies have shown that Chinese children tend to perform better in manual dexterity and balance tasks, while American children excel in throwing and catching tasks [40]. Moreover, regional disparities offer valuable insights into the development of various motor skills within diverse national contexts.

Objectives

The primary objective of this study is to evaluate the applicability of a standardized battery of gross motor skill tests for Moroccan children aged 6 to 12 years with ASD. This will provide preliminary data and allow for comparison to a group of children with a neurotypical profile matched for age and sex, thereby estimating the discrepancies in motor development between the 2 groups. The secondary objective is to assess the potential efficacy of an innovative pedagogical approach focused on sensorimotor integration and hyperstimulation for Moroccan children with ASD. This approach will be compared to traditional physical education (PE) sessions to determine its feasibility and potential to bridge the developmental gaps in motor skills between children with ASD and peers with a neurotypical profile.

Methods

Design

This study was a clinical, quasi-experimental research with a 2-way repeated measures design conducted on a sample of Moroccan children with ASD aged 6 to 12 years.

Participants

A convenience sample of 14 Moroccan children with ASD aged 6 to 12 (mean 8.8, SD 1.3) years participated in this exploratory study. This age range was selected because motor skills develop particularly well between 6 and 12 years of age, and the chosen test battery has been validated for children of this age group. Children were randomly divided into an experimental group (n=7; mean age 8.4, SD 1.2 years) and a control group (n=7; mean age 8.5, SD 1.2 years) based on age, sex, motor skill performance, and socioeconomic status. All participants were classified by a psychiatrist as level 2 (moderate) on the Autism Severity Rating Scale based on the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, Text Revision* (DSM-5-TR) criteria. This scale considers the level of support required for the child's interaction and language to function in their environment. Among the recruited children with ASD, the records showed no other conditions, such as attention deficit disorder with or without hyperactivity, intellectual disability, or anxiety. Children with ASD were cared by the specialized association "El Youssr" since they entered school. This institution's mission is to support children with special needs. It is located in the municipality of Salé, on the outskirts of the capital Rabat, where most of the participating families have a low socioeconomic status. None of the children participated in extracurricular sports activities or were undergoing medical treatment that could influence their motor skills, such as taking psychotropic drugs. In addition, none of the children had a disabling condition that could be exacerbated by physical activity. The proposed program was designed to ensure minimal disruption to the existing educational and individual student plans within the institution.

Anthropometric Measures

The anthropometric data were collected in the association's classrooms, which included body mass (BM), body height (BH), BMI, and body surface area (BSA). BM was assessed using a Seca scale model 760 with an accuracy of 0.5 kg. BH was measured with a Seca stadiometer model 214 to the nearest 0.1 cm. All measurements adhered to the guidelines outlined by Lohman et al [41] for standard operating procedures. BMI was calculated using the formula BM (kg)/BH^2 (m). BSA was estimated using the method suggested by Mosteller [42], which read as follows:

Motor Skill Assessment

This exploratory study investigated the factors associated with motor skill development in Moroccan children with ASD over a 15-week period. The initial assessment session (T1) established the baseline level of motor skill development in these children.

XSL•FO
RenderX

The results from T1 were also used to create control and experimental groups, ensuring no significant difference between them at baseline. Gross motor skills were measured using the UQAC-UQAM (University of Québec in Chicoutimi-University of Québec in Montréal) test battery protocol [43-45]. This battery comprises 13 items assessing five gross motor factors: (1) limb movement speed (upper and lower limbs), (2) agility (circle run, 5-m shuttle run, and side-step run), (3) balance (eyes opened and closed), (4) coordination (ball dribbling and target ball toss), and (5) simple reaction time (SRT; ruler drop test).

For the SRT test, the distance was converted into time (ms) using the equation suggested by Aranha et al [46]:

where d is the distance (cm) and g is the gravitational constant (9.8 m/s^2).

The choice of this test battery was primarily driven by methodological considerations. First, it assesses a diverse range of motor factors. Additionally, these tests are practical, as they require minimal space, materials, and time and are user-friendly for test administrators, easy for participants to comprehend, and cost-effective, aligning well with the resources of the supervising institution. Given the sensitivity of children with ASD to environmental changes and their challenges with concentration and attention, it was essential to conduct assessments in their classroom settings whenever feasible. The validity and reliability of these tests have been well-established by numerous studies [43,45,47].

Modified Protocols for the UQAC-UQAM Test Battery

Prior to starting this research project, a week of prospecting was conducted to evaluate the study's feasibility. Of the original 13 tests within the UQAC-UQAM test battery [43], 3 were omitted due to challenges encountered by the children in comprehending or executing the tasks required (slalom run and hand or foot coordination). Furthermore, 2 tests were adapted to enhance accessibility for children with ASD (Figure 1). The first adaptation concerned on the SRT test. The original version required computer administration, but pilot testing revealed this to be overly complex for the participants. This test was therefore replaced by the procedure suggested by Ángel Latorre-Ron et al [48]. This alternative demonstrated a moderate to good correlation with a computer application (Spearman=0.54; P=.03). This change considered the participants' concentration abilities, as the computerized test was longer than the chosen alternative, which only required 2 attempts. The second adaptation concerned the ball toss test, where the distance between the children and the target was reduced from 5 to 3 m. These modifications to the testing protocol were implemented to ensure the study's feasibility and to gather reliable and valid data on motor skill development in children with ASD. It was imperative to provide tests adapted to their abilities and limitations to minimize frustration and encourage active and engaged participation.

Figure 1. Illustration of the ten tests used to assess motor skill competence in children with ASD: (1) upper limbs speed, (2) lower limbs speed, (3) balance eyes opened, (4) balance eyes closed, (5) simple reaction time, (6) ball dribbling, (7) target ball toss, (8) side-step run, (9) circle run, and (10) 5-m shuttle run (adapted from Allisse et al [49], which is published under Creative Commons Attribution 4.0 International License [50]).

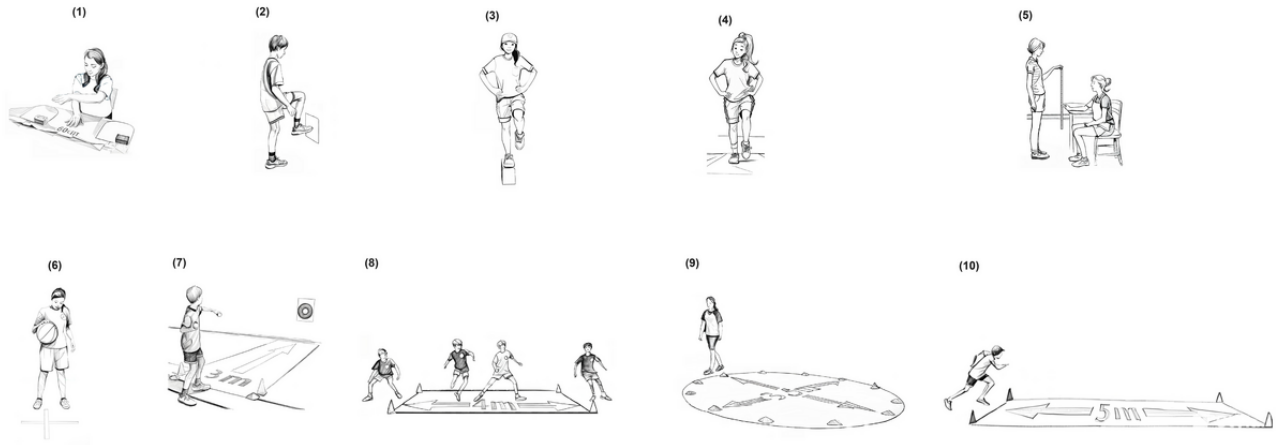

Following these adjustments, the final test battery comprised the following ten tests: (1) upper limb speed, (2) lower limb speed, (3) balance eyes opened, (4) balance eyes closed, (5) ruler SRT, (6) ball dribbling, (7) target ball toss, (8) side-step run, (9) circle run, and (10) 5-m shuttle run. For the 2 limb speed tests and the ball dribbling test, the score was determined by the greatest number of repetitions within 20 seconds. For the 3 balance tests and the 3 running tests, the score was recorded as time in seconds. In the SRT test, the score was measured in centimeters and then transformed in milliseconds. The target ball toss test score was cumulative, based on the number of points scored in 10 throws (1 point for hitting the target and 2 points for reaching the center). For all tests except the target ball toss, 2 attempts were allowed, and the best of the 2 was recorded as the final score.

Specific Training Program

The ASD experimental group participated in a specific training program lasting a total of 15 weeks. The program was divided into 3 cycles of 5 weeks each, with each cycle comprising 3 weekly sessions of 45 minutes. This resulted in a total of 45 sessions delivered between November 1, 2021, and April 3, 2022. The pedagogical approach used small group sizes, allowing for individualized monitoring and ensuring a clear understanding of the tasks by participants. The specific training program focused on three key areas: (1) coordination:

XSL•FO
RenderX

development of motor skill enabling participants to coordinate whole-body movements smoothly and efficiently; (2) proprioception: enhancement of body awareness and spatial positioning, leading to improved postural control; and (3) agility: strengthening the ability to react quickly and change direction with precision. The intervention process emphasized offering a variety of engaging and playful activities. The overall objective was to provide participants with rich and multimodal sensory experiences that stimulated awareness and intrinsic motivation and presented optimal motor challenges. This included activities such as exercises performed barefoot or with shoes, allowing participants to explore different textures and sensations and exercises using a trampoline to promote body awareness in space and improve balance. Alongside the ASD experimental group, a separate ASD control group participated in regular PE classes based on the standard curriculum as proposed by the Ministry of National Education of Morocco (Table 1). Further details can be found in Multimedia Appendices 1 and 2. The duration and frequency of these classes matched those of the experimental group. The inclusion of an ASD control group allowed for a comparison of motor skill development in participants who underwent the specific training program against those who did not (for a detailed description of both, the intervention programs for the experimental and control groups, please refer to Multimedia Appendices 1 and 2). Furthermore, a neurotypical control group comprising children without ASD was established. This group served as a baseline to evaluate the motor skill level of children with ASD in both the experimental and control groups, taking into account age and locality.

XSL•FO
RenderX

Table 1. Detailed comparison of 15-week intervention activities for experimental (sensory intervention) and control (classic intervention) groups in a physical development program for children with autism spectrum disorder.

Block and group	Theme	Content	Operationalization
Block 1: 3 times per week (5 weeks)			
Experimental	Coordination	• Spatial orientation; • Execution rhythm and time-space structuring; • Segmental differentiation and dissociation; • Reaction speed.	• Tactile activities: diversifying contact surfaces (barefoot and with shoes); • Proprioceptive activities: trampoline and plyometric exercises; • Vestibular activities: Swiss ball exercises and orientation; • Multimodal sensory experiences tasks; • Playful form, optimal motor challenge, and secure and flexible framework; • Encourage collaboration and participation.
Control	Aerobic fitness	• Aerobic capacity (10 to 15 minutes total); • Aerobic power (3 to 6 minutes total, various repetitions); • Speed (repetitions of 2 to 3 seconds).	• Continuous running: interval running (varying long or fast pace); • Intermittent: short/short, long/long.
Block 2: 3 times per week (5 weeks)			
Experimental	Balance	• Footwork and body balance mastery; • Spatial orientation; • Static and dynamic postural balance; • Plyometric exercises.	• Tactile activities: diversifying contact surfaces (barefoot and with shoes); • Proprioceptive activities: trampoline, plyometric exercises; • Vestibular activities: Swiss ball exercises, orientation; • Multimodal sensory experiences tasks; • Playful form, optimal motor challenge, and secure and flexible framework; • Encourage collaboration and participation.
Control	Strength	• Horizontal and vertical jump (explosive strength); • Linking running with vertical and horizontal jump; • Jump adaptation to space; • Movement, acceleration, tension, ejection, and body stabilization; • Mastery of principles: slow/fast, back/front, bottom/top, right/left.	• Learn phases and sequences of throwing and jumping; • Understand general rules of throwing or jumping; • Motor organization to jump or throw as high or far as possible.
Block 3: 3 times per week (5 weeks)			
Experimental	Agility	• Linking, braking, and acceleration; • Linear and multidirectional speed; • Motor flexibility; • Body control: adaptation to various stimuli; • Rapid and continuous direction changes; • Muscle strengthening.	• Tactile activities: diversifying contact surfaces (barefoot and with shoes); • Proprioceptive activities: trampoline and plyometric exercises; • Vestibular activities: Swiss ball exercises and orientation; • Multimodal sensory experiences tasks; • Playful form, optimal motor challenge, and secure and flexible framework; • Encourage collaboration and participation.
Control	Cooperation and opposition with ball	• Discover a team game: 1 or 2 key rules; • Differentiate partner and opponent; • Respect spatial boundaries; • Discover a target to defend and attack.	• Social interactions and exchanges; • Communication and information gathering; • Cooperation and mutual aid.

Statistical Analysis

The data were summarized using descriptive statistics. Central tendency was characterized by means, and variability was assessed by SD. To assess normality for all variables, the Shapiro-Wilk test was conducted. Group mean comparisons were performed using unpaired 2-tailed *t* tests for the control and experimental groups, while paired 2-tailed *t* tests were used to evaluate differences between measurements at time 1 (T1) and time 2 (T2). Effect sizes (ESs) were calculated using the

XSL•FO
RenderX

Cohen *d* formula [51] and interpreted using the following scale: <0.20 (trivial), 0.20-0.49 (small), 0.50-0.79 (medium), and ≥0.80 (large). The main effects of group and time, as well as the group×time interaction, were analyzed using a 2-way repeated measures ANOVA. Partial eta squared (η^2p) values were reported to indicate ESs, interpreted as follows: <0.01 (trivial), 0.01-0.05 (small), 0.06-0.13 (moderate), and ≥0.14 (large). A significance level of $P \leq .05$ was used. All statistical analyses were carried out with SPSS (version 24.0; IBM Corp).

Ethical Considerations

Details of the project were communicated to parents, students with ASD, and education authorities through a written and explanatory document that each party was required to sign. Thus, by enrolling in this project, the individuals involved acknowledged having provided de facto their written informed consent. However, it was agreed that participants could withdraw at any time without penalty or prejudice. The study was conducted in accordance with the Declaration of Helsinki and approved by the institutional ethics committee of Ibn Tofail University (protocol code: 20019741; date of approval: March 3, 2023) for studies involving humans. Data collection only began upon receiving authorization from the presidents of the 2 participating associations. All collected data were anonymized to ensure participant confidentiality and were stored securely in compliance with ethical guidelines. Participants or their families received no financial or material compensation for their participation in this study. This paper does not contain any images or documents that could identify individual participants.

Results

Table 2 presents the anthropometric characteristics between the ASD experimental group to the ASD control group at baseline (T1). The results show that there are no significant differences in the anthropometric characteristics measured between the children in the experimental group and those in the control group at baseline. This suggests that the 2 groups are well balanced in terms of age, BM, BH, BMI, and BSA, which is crucial to ensure that any future intervention can be evaluated without bias related to initial anthropometric differences between the groups. The *P* values indicate that all comparisons are nonsignificant, with values well above the conventional significance limit of ≤.05, reinforcing the idea of homogeneity between the groups at the start of the study.

Table 2. Anthropometric characteristics of children with autism spectrum disorder according to group membership at baseline (T1).

Variables	Experimental group (n=7)		Control group (n=7)		*P* value[a]
	Mean (SD)	95% CI	Mean (SD)	95% CI	
Age (years)	8.5 (1.2)	7.3-9.5	9.1 (1.4)	7.9-10.5	.29
Body mass (kg)	31.0 (7.3)	24.3-37.8	32.7 (8.0)	25.3-40.1	.68
Body height (cm)	131.7 (7.8)	124.5-138.9	133.7 (7.3)	126.9-140.5	.63
BMI (kg/m^2)	17.9 (3.5)	14.6-21.1	18.1 (2.5)	15.8-20.4	.88
Body surface area (m^2)	1.06 (0.14)	0.93-1.19	1.10 (0.16)	0.95-1.25	.65

[a]Significant at ≤.05.

Table 3 presents a comprehensive comparative analysis of the motor skill performance evolution among children with ASD, spanning from the initial baseline (T1) to the postintervention period (T2). The data are categorized into 2 groups: an ASD experimental group and an ASD control group. In the ASD control group, significant enhancements were observed in upper and lower limb speed, along with ball dribbling and 5-m shuttle run after 15 weeks in the classic PE program. However, there were no significant differences noted in balance duration with eyes opened and closed. Similarly, variables related to running and coordination, such as circle run, side-step run, 5-m shuttle run, target ball toss, and ball dribbling, did not exhibit significant improvements. Regarding the ASD experimental group, noteworthy advancements were recorded across all variables after 15 weeks in the sensorimotor integration and hyperstimulation program. Examining the overarching changes in motor performance between T1 and T2, a distinct contrast emerges between the 2 ASD groups. The ASD control group witnessed an overall progression of 14.5% across all evaluated motor skill variables. Conversely, the ASD experimental group demonstrated a notably higher progression, reaching 44.5%.

XSL•FO
RenderX

Table 3. Comparison of motor skill improvements from baseline (T1) to postintervention (T2) in autism spectrum disorder (ASD) control and experimental groups[a].

Variables	T1, mean (SD)	T2, mean (SD)	*P* value[b]	Cohen *d* effect size[c]	Δ T1 vs T2 (%)
ASD experimental group					
Upper limb speed (nb/20 seconds[d])	23.7 (6.3)	37.7 (4.2)	.001	2.12	14.0 (37.1)
Lower limb speed (nb/20 seconds)	9.0 (2.1)	16.4 (2.2)	<.001	6.74	7.4 (45.1)
Balance eyes opened (seconds)	6.1 (2.6)	14.6 (8.7)	.02	1.21	8.5 (58.2)
Balance eyes closed (seconds)	2.3 (0.9)	6.4 (2.0)	.001	1.91	4.1 (64.1)
Ball dribbling (nb/20 seconds)	5.3 (4.0)	22.6 (10.9)	.003	1.88	17.3 (76.6)
Target ball toss (points)	3.1 (2.2)	8.1 (3.1)	.01	0.89	5.0 (61.7)
Side-step run (seconds)	20.4 (3.7)	13.9 (0.9)	.003	1.61	6.5 (31.9)
Circle run (seconds)	40.0 (5.8)	28.1 (1.1)	.002	1.13	11.9 (30)
5-m shuttle run (seconds)	18.4 (0.8)	15.4 (1.1)	.001	2.15	3.0 (16.3)
Simple reaction time (ms)	299.5 (45.6)	223.4 (24.6)	.002	1.78	76.1 (25.4)
Mean (SD)	__[e]	—	—	2.14 (1.67)	44.5 (19.8)
ASD control group					
Upper limb speed (nb/20 seconds)	22.9 (4.9)	27.7 (4.7)	.01	1.45	4.8 (17.3)
Lower limb speed (nb/20 seconds)	6.9 (2.0)	10.9 (2.6)	.002	2.28	4.0 (36.7)
Balance eyes opened (seconds)	3.2 (1.2)	4.6 (2.6)	.28	0.29	1.4 (30.4)
Balance eyes closed (seconds)	1.9 (0.9)	1.9 (1.3)	.91	0.00	0.0 (0)
Ball dribbling (nb/20 seconds)	3.0 (0.8)	6.7 (3.4)	.01	3.30	3.7 (55.2)
Target ball toss (points)	1.7 (1.9)	1.6 (1.4)	.79	0.10	–0.1 (–5.9)
Side-step run (seconds)	21.6 (2.9)	20.5 (4.1)	.25	0.74	1.1 (5.1)
Circle run (seconds)	39.9 (7.1)	37.2 (5.6)	.06	2.12	2.7 (6.8)
5-m shuttle run (seconds)	18.9 (3.7)	20.6 (3.2)	.007	4.57	–1.7 (–8.3)
Simple reaction time (ms)	336.2 (33.2)	310.6 (26.9)	.14	0.67	25.6 (7.6)
Mean (SD) (%)	—	—	—	1.55 (1.51)	14.5 (20.4)

[a]Intergroup differences were measured using a 2-tailed *t* test for unpaired values. Changes between T1 and T2 for each group were calculated using a 2-tailed *t* test for paired values.

[b]Significant at $P \leq .05$.

[c]<0.20=marginal; 0.20-0.49=small; 0.50-0.79=moderate; and ≥0.80=large.

[d]nb/20 seconds=number of repetitions in 20 seconds.

[e]Not applicable.

The results outlined in Table 4 reveal significant effects of group, time, and group×time interactions for the majority of gross motor skills. Variables such as segmental speed, balance, ball dribbling, and SRT exhibit the most substantial effects. While improvements are also noted in side-step running and circle running, these gains are comparatively modest. Mauchly test of sphericity indicated that the assumption of sphericity was not violated for any of the dependent variables, suggesting that the variances of the differences between levels of the repeated measures factor were approximately equal. Thus, these findings demonstrate that the intervention was highly effective in enhancing motor performance within the experimental group.

XSL•FO
RenderX

Table 4. Summary of a 2-way ANOVA with repeated measures evaluating the effects of group, time, and group×time interaction on gross motor skills performance between the experimental and control groups of children with autism spectrum disorder[a].

Variables and source	Sum of squares	*df*	Mean square	*F* test (*df*=12)[b]	*P* value[c]	η^2p[d]
Upper limb speed (nb/20 seconds[e])						
Group	206.3	1	206.3	5.361	.04	0.309
Time	622.3	1	622.3	48.044	<.001	0.800
Group×time	146.3	1	146.3	11.294	.006	0.485
Lower limb speed (nb/20 seconds)						
Group	104.1	1	104.1	12.941	.004	0.519
Time	228.6	1	228.6	145.455	<.001	0.924
Group×time	20.6	1	20.6	13.091	.004	0.522
Balance eyes opened (seconds)						
Group	365.6	1	365.6	19.393	.001	0.618
Time	260.7	1	260.7	33.560	<.001	0.737
Group×time	133.7	1	133.7	17.206	.001	0.589
Balance eyes closed (seconds)						
Group	42.9	1	42.9	21.122	.001	0.638
Time	30.6	1	30.6	20.113	.001	0.626
Group×time	28.5	1	28.5	18.761	.001	0.610
Ball dribble (nb/20 seconds)						
Group	576.0	1	576.0	11.576	.005	0.491
Time	771.8	1	771.8	31.998	<.001	0.727
Group×time	322.3	1	322.3	13.364	.003	0.527
Target ball toss (points)						
Group	112.0	1	112.0	18.556	.001	0.607
Time	41.3	1	41.3	10.671	.007	0.471
Group×time	46.3	1	46.3	11.963	.005	0.499
Side-step run (seconds)						
Group	106.4	1	106.4	7.020	.21	0.369
Time	103.4	1	103.4	22.549	<.001	0.653
Group×time	50.5	1	50.5	11.004	.006	0.478
Circle run (seconds)						
Group	144.3	1	144.3	3.142	.10	0.207
Time	372.7	1	372.7	30.427	<.001	0.717
Group×time	148.9	1	148.9	12.151	.004	0.503
5-m shuttle run (seconds)						
Group	56.3	1	56.3	4.669	.05	0.280
Time	2.5	1	2.5	3.550	.08	0.228
Group×time	39.1	1	39.1	56.465	<.001	0.825
Simple reaction time (ms)						
Group	26,859.3	1	26,859.3	18.060	.001	0.601
Time	18,105.4	1	18,105.4	28.592	<.001	0.704
Group×time	4450.7	1	4450.7	7.028	.02	0.369

XSL•FO
RenderX

[a]Mauchly sphericity test≥0.05 for all the variables in the analysis.

[b]*F*=variance between groups/variance within groups.

[c]Significant if ≤.05.

[d]Small effect size: η^2p=0.01; moderate effect size: η^2p=0.06; and large effect size: η^2p=0.14.

[e]nb/20 seconds=number of repetitions in 20 seconds.

Discussion

Principal Findings

Autism, a complex spectrum of neurodevelopmental conditions, is characterized by a wide range of manifestations and challenges. Recent research, such as that conducted by Khoury et al [52], has shed light on a connection between sensorimotor integration disorders and ASD. These disorders impact both fine and gross motor skills, as stated by some studies that identified 2 primary anomalies contributing to the motor difficulties observed in individuals with ASD, which are impaired integration of information crucial for motor planning and heightened variability in basic sensory input and output [24,53].

Characteristics of Motor Performance in Individuals With ASD

Unlike individuals with a neurotypical profile, who readily integrate information from multiple senses (multimodal integration), people with ASD often process this information independently for each sense (unimodal integration). This unique characteristic can lead to difficulties in perception when exposed to simultaneous stimulation across multiple senses. As part of a study comparing the motor performances of individuals with and without ASD, the battery of tests proposed by Leone et al [43,44] proved to be a relevant tool. First, this battery effectively evaluates a broad range of fundamental motor skill factors (segmental speed, coordination, balance, agility, and reaction time), which was important in the context of this study. Second, the test items are designed as simple, isolated tasks, minimizing the influence of other motor components. This structure aimed to reduce potential performance discrepancies between the 2 participant groups. However, despite these precautions, the study revealed significant differences in motor performance between children with a neurotypical profile and those with ASD from the outset. Children with ASD demonstrated a marked delay in motor development, even in tasks primarily reliant on unimodal sensory processing.

The Classic Versus the Sensorimotor Integration and Hyperstimulation Approach

In efforts to integrate children with ASD into the education system, it is common for PE lessons to be taught using a classical method similar to that used for children with a neurotypical profile. Morocco follows this trend and offers children with autism the same PE program prescribed by the Ministry of National Education. Although some adaptations are implemented, such as reducing class sizes, this type of program, which is based on disparate and eclectic exercises derived from sports activities, does not sufficiently consider the unique characteristics of children with autism. This approach results in a significant increase in the quantity and variety of sensory and motor stimulation, which conflicts with the typical functioning of children with autism. Additionally, these children often experience sensory hyperreactivity or hyporeactivity to environmental stimuli such as brightness, ground texture, and ambient noise. This can disrupt their routine and, consequently, hinder their learning capacity [54]. Despite their limitations, traditional PE classes can demonstrably contribute to improve some motor skills in children with ASD. Our study showed that a 15-week classical PE program allowed a nearly 14.5% overall improvement in the control group of children with ASD. This finding highlights the capacity of traditional PE methods to induce, to a certain extent, a contribution to the motor development of this population.

In contrast, the experimental group, whose PE lessons were adapted based on principles of sensorimotor integration and hyperstimulation management, demonstrated a remarkable 44.5% improvement over the same 15-week period. Aware of the debates that this approach still generates, we have introduced innovations that, in our opinion, have contributed to this impressive outcome. While sensorimotor integration appears more appropriate for children with ASD than traditional methods, most research advocates for a unimodal approach, emphasizing tasks that engage a single sense. Although some studies show promising results, others question the effectiveness of this method, suggesting that it fails to adequately address the specific needs of children with ASD.

In this study, a multimodal approach was adopted. Each block focuses on a specific single motor skill in a manner that engages multiple senses simultaneously. For instance, while block 1 emphasizes coordination, learning situations necessitate the involvement of various senses to accomplish each task. Furthermore, to control for environmental variables and minimize external influences, all sessions were conducted in a single room, ensuring consistent lighting, playing surface, and ambient sound when the learning situation requires it. Given its unique characteristics, the program structure developed in this study represents a valuable contribution to the field.

ASD Experimental Versus Control Group

The results of this study are highly encouraging and suggest that the implemented intervention had a significant impact on improving the motor skills of children with ASD. Indeed, all motor skill factors assessed showed significant improvement in the experimental group compared to the control group (P<.05). The enhancements were not only statistically significant but also clinically relevant, as indicated by large ESs (Cohen ES>0.80). These findings support the idea that targeted interventions can enhance motor skills in children with ASD, with potential broad-ranging implications for their overall development.

The results of the 2-way repeated measures ANOVA strongly support the effectiveness of the sensorimotor integration and hyperstimulation method. The findings reveal that the

XSL•FO
RenderX

group×time interaction is significant for several variables, indicating that performance varies not only by group but also over time. This interaction suggests that the effect of time on motor performance is not consistent across groups. In other words, the interventions or experimental conditions had different impacts depending on group membership. It is therefore likely that the experimental group showed more substantial improvements over time compared to the control group, which could suggest greater effectiveness of the implemented intervention. The ESs (η^2p) associated with these interactions are also noteworthy. The high ESs further emphasize that the intervention had a significant clinical impact. These findings emphasize the importance of individualized training programs that consider specific needs and learning styles to optimize motor development outcomes.

The observed improvements spanned a wide range of motor domains, including coordination, balance, and reaction time, suggesting a comprehensive enhancement of motor skills. This study highlights the importance of incorporating specific physical activities into educational programs for children with ASD. By promoting motor skill development, one can significantly improve these children's quality of life, learning, and social inclusion. Moreover, this study raises intriguing questions regarding the underlying neurobiological mechanisms that may account for the observed improvements in children with ASD following the intervention. While further research is needed to draw definitive conclusions, several potential mechanisms could be explored.

Potential Neurobiological Mechanisms Underlying Motor Skill Improvements in Children With ASD

Regular physical exercise stimulates the production of neuronal growth factors, which promote the formation of new synaptic connections in brain regions involved in motor control. Training can lead to the reorganization of neural networks, enhancing movement coordination and motor planning [55]. Physical activity is linked to dopamine release, a neurotransmitter integral to reward systems. This could boost children's motivation to engage in physical activities and reinforce adaptive motor behaviors. Dopamine also plays a pivotal role in motor control, and its modulation could contribute to the observed improvements [56]. Additionally, serotonin, which regulates mood and anxiety, may see increased activity due to physical exercise, potentially reducing the anxiety symptoms often associated with ASD and facilitating the acquisition of new skills [57]. Physical exercise can enhance sensory perception, including proprioception and vestibular sense, thereby improving movement coordination [58]. Our proposed sensorimotor integration and hyperstimulation program specifically addresses these 2 sensory modalities. Although these hypotheses are plausible, further research is needed to assess their validity.

Implementation of a Sensorimotor Integration Program in the Moroccan Context

Similar to many societies, Morocco often prioritizes the classical PE approach to stimulate motor development in children, including those with ASD. However, the findings of this exploratory study suggest that an alternative approach focused on sensorimotor integration and hyperstimulation management could be particularly beneficial for Moroccan children with ASD. Conducted within the Moroccan cultural and educational context, the study demonstrated that this innovative approach could not only be successfully implemented but also yield significantly better results in motor development compared to the classical approach. While traditional PE classes can offer some benefits to children with ASD, integrating principles of sensorimotor integration and hyperstimulation management into PE programs can lead to more substantial improvements. The use of a multimodal approach, where tasks are designed to engage multiple senses, appears to be particularly effective. This method not only addresses the unique sensory processing challenges faced by children with ASD but also helps to bridge the developmental gap between them and their peers with a neurotypical profile. As such, the program developed in this study represents a significant advancement in the field of PE for children with ASD, providing a framework for future research and practice.

Limitations and Strengths

This study reveals several limitations. The small number of participants with ASD (n=14) requires caution in interpreting the results. Additionally, the fact that the age range is limited to 6 to 12 years makes the results potentially inapplicable to other age groups. Given that these children mainly come from disadvantaged backgrounds, it is uncertain whether the same findings would be observed in different socioeconomic contexts. All participants with ASD were classified as level 2 (moderate) on the Autism Severity Rating Scale according to DSM-5-TR criteria, so it is not possible to judge whether the program would have the same effectiveness for severity levels 1 and 3. However, this limitation ensures that the group of children with ASD was homogeneous, which is often a weakness in other studies. Additionally, obtaining repeated measurements allows changes to be better assessed since they come from the same children at different time periods. This study also has the advantage of providing, for the first time, a preliminary portrait of the motor development of Moroccan children with a neurotypical profile versus children with ASD based on the UQAC-UQAM test battery. The proposed approach (sensorimotor integration and hyperstimulation), as implemented in this study, is also a first for the Moroccan population with ASD.

Conclusions

Children with ASD often exhibit a unimodal sensory processing style, meaning they tend to focus on one sensory input at a time. This can lead to significant perceptual challenges when they are exposed to stimulation that engages multiple senses simultaneously. Conventional PE programs, which typically use a variety of exercises derived from various sports, may actually hinder the progress of children with ASD by overloading their sensory systems. In contrast, a novel approach that emphasizes sensorimotor integration and hyperstimulation management appears to be much more effective in improving motor skills in this population. This method prioritizes engaging multiple senses concurrently while minimizing excessive sensory stimuli. This study revealed that the sensorimotor

XSL•FO
RenderX

integration and hyperstimulation management program yielded a remarkable 44.5% improvement in motor skills for children in the experimental group over a 15-week period. This is in stark contrast to the 14.5% improvement observed in the control group who participated in the traditional PE program. While this study is limited by the sample size, it provides a groundbreaking initial assessment of motor development in Moroccan children with ASD. Furthermore, it proposes an innovative approach specifically tailored to their motor skills condition and the prevailing sociocultural and educational context within Morocco. However, other more exhaustive studies will need to be carried out in order to be able to more precisely measure the full potential of this approach.

Acknowledgments

The authors wish to acknowledge the school principal, the specialized educators' team, children, and parents or legal guardians who agreed to participate in this study. This research received no external funding.

Data Availability

The original contributions presented in the study are included in the paper. Further inquiries can be directed to the corresponding author. All data generated or analyzed during this study are included in this published paper.

Authors' Contributions

RT was involved in the design and concept of the study and data collection, drafted the initial version of the manuscript, and reviewed and revised the manuscript. JZ and EMC coordinated and supervised data collection, were involved in the study design, and reviewed and revised the manuscript. HTB was involved in the data analysis and reviewed and drafted the manuscript for important intellectual content. ML and MA were involved in the study design and initial analyses, drafted the initial manuscript, and reviewed and revised the manuscript. All authors approved the final manuscript as submitted and agreed to be accountable for all aspects of the work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The intervention program for the experimental group.
[PDF File (Adobe PDF File), 381 KB - formative_v9i1e65767_app1.pdf]

Multimedia Appendix 2
The intervention program for the control group.
[PDF File (Adobe PDF File), 694 KB - formative_v9i1e65767_app2.pdf]

References

1. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders: DSM-5-TR, 5th Edition. Washington, DC: American Psychiatric Publishing; 2022.
2. Christensen DL, Baio J, van Naarden Braun K, Bilder D, Charles J, Constantino JN, et al. Prevalence and characteristics of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network, 11 sites, United States, 2012. MMWR Surveill Summ 2016 Apr 01;65(3):1-23 [FREE Full text] [doi: 10.15585/mmwr.ss6503a1] [Medline: 27031587]
3. Le Menn-Tripi C, Vachaud A, Defas N, Malvy J, Roux S, Bonnet-Brilhault F. Sensory-psychomotor evaluation in Autism: a new tool for functional diagnosis. Encephale 2019;45(4):312-319. [doi: 10.1016/j.encep.2018.12.003] [Medline: 30922517]
4. Odeh CE, Gladfelter AL, Stoesser C, Roth S. Comprehensive motor skills assessment in children with autism spectrum disorder yields global deficits. Int J Dev Disabil 2022;68(3):290-300 [FREE Full text] [doi: 10.1080/20473869.2020.1764241] [Medline: 35602998]
5. Paquet A, Olliac B, Golse B, Vaivre-Douret L. Nature of motor impairments in autism spectrum disorder: a comparison with developmental coordination disorder. J Clin Exp Neuropsychol 2018;41(1):1-14. [doi: 10.1080/13803395.2018.1483486] [Medline: 29923455]
6. Rogers S, Bennetto L. Motor functioning in autism [in French]. Enfance 2002;54(1):63-73 [FREE Full text]
7. Wilson RB, Enticott PG, Rinehart NJ. Motor development and delay: advances in assessment of motor skills in autism spectrum disorders. Curr Opin Neurol 2018;31(2):134-139 [FREE Full text] [doi: 10.1097/WCO.0000000000000541] [Medline: 29493557]

XSL•FO
RenderX

8. Ketcheson LR, Pitchford EA, Wentz CF. The relationship between developmental coordination disorder and concurrent deficits in social communication and repetitive behaviors among children with Autism spectrum disorder. Autism Res 2021;14(4):804-816. [doi: 10.1002/aur.2469] [Medline: 33421296]
9. Miller HL, Sherrod GM, Mauk JE, Fears NE, Hynan LS, Tamplain PM. Shared features or co-occurrence? Evaluating symptoms of developmental coordination disorder in children and adolescents with autism spectrum disorder. J Autism Dev Disord 2021;51(10):3443-3455 [FREE Full text] [doi: 10.1007/s10803-020-04766-z] [Medline: 33387238]
10. Zampella CJ, Wang LAL, Haley M, Hutchinson AG, de Marchena A. Motor skill differences in autism spectrum disorder: a clinically focused review. Curr Psychiatry Rep 2021;23(10):64. [doi: 10.1007/s11920-021-01280-6] [Medline: 34387753]
11. Gillberg C. Deficits in attention, motor control, and perception: a brief review. Arch Dis Child 2003;88(10):904-910 [FREE Full text] [doi: 10.1136/adc.88.10.904] [Medline: 14500312]
12. Hocking DR, Caeyenberghs K. What is the nature of motor impairments in autism, are they diagnostically useful, and what are the implications for intervention? Curr Dev Disord Rep 2017;4(2):19-27. [doi: 10.1007/s40474-017-0109-y]
13. Kindregan D, Gallagher L, Gormley J. Gait deviations in children with autism spectrum disorders: a review. Autism Res Treat 2015;2015:741480-741488 [FREE Full text] [doi: 10.1155/2015/741480] [Medline: 25922766]
14. Srinivasan SM, Kaur M, Park IK, Gifford TD, Marsh KL, Bhat AN. The effects of rhythm and robotic interventions on the imitation/praxis, interpersonal synchrony, and motor performance of children with autism spectrum disorder (ASD): a pilot randomized controlled trial. Autism Res Treat 2015;2015:736516 [FREE Full text] [doi: 10.1155/2015/736516] [Medline: 26793394]
15. Volkmar F, Siegel M, Woodbury-Smith M, King B, McCracken J, State M. Practice parameter for the assessment and treatment of children and adolescents with autism spectrum disorder. J Am Acad Child Adolesc Psychiatry 2014;53(2):237-257 [FREE Full text] [doi: 10.1016/j.jaac.2013.10.013] [Medline: 24472258]
16. Fournier KA, Kimberg CI, Radonovich KJ, Tillman MD, Chow JW, Lewis MH, et al. Decreased static and dynamic postural control in children with autism spectrum disorders. Gait Posture 2010;32(1):6-9 [FREE Full text] [doi: 10.1016/j.gaitpost.2010.02.007] [Medline: 20400311]
17. Gernsbacher MA, Sauer EA, Geye HM, Schweigert EK, Hill Goldsmith H. Infant and toddler oral- and manual-motor skills predict later speech fluency in autism. J Child Psychol Psychiatry 2008;49(1):43-50 [FREE Full text] [doi: 10.1111/j.1469-7610.2007.01820.x] [Medline: 17979963]
18. Lim YH, Partridge K, Girdler S, Morris SL. Standing postural control in individuals with autism spectrum disorder: systematic review and meta-analysis. J Autism Dev Disord 2017;47(7):2238-2253. [doi: 10.1007/s10803-017-3144-y] [Medline: 28508177]
19. Pace M, Dumortier L, Guinot M, Favre-Juvin A, Bricout VA. Assessment of motor skills in autistic children on the M-ABC [in French]. Sci Sports 2016;31(6):336-341. [doi: 10.1016/j.scispo.2016.08.001]
20. Coll SM, Foster NE, Meilleur A, Brambati SM, Hyde KL. Sensorimotor skills in autism spectrum disorder: a meta-analysis. Res Autism Spectr Disord 2020;76:101570. [doi: 10.1016/j.rasd.2020.101570]
21. Bojanek EK, Wang Z, White SP, Mosconi MW. Postural control processes during standing and step initiation in autism spectrum disorder. J Neurodev Disord 2020;12(1):1 [FREE Full text] [doi: 10.1186/s11689-019-9305-x] [Medline: 31906846]
22. Casey BJ, Glatt CE, Lee FS. Treating the developing versus developed brain: translating preclinical mouse and human studies. Neuron 2015;86(6):1358-1368. [doi: 10.1016/j.neuron.2015.05.020] [Medline: 26087163]
23. Gowen E, Hamilton A. Motor abilities in autism: a review using a computational context. J Autism Dev Disord 2013;43(2):323-344. [doi: 10.1007/s10803-012-1574-0] [Medline: 22723127]
24. Smoot Reinert S, Jackson K, Bigelow K. Using posturography to examine the immediate effects of vestibular therapy for children with Autism spectrum disorders: a feasibility study. Phys Occup Ther Pediatr 2015;35(4):365-380. [doi: 10.3109/01942638.2014.975313] [Medline: 25374155]
25. Zamani Jam A, Talab RH, Sheikh M, Torabi F, Rafie F. The effect of 16 weeks gymnastic training on social skills and neuropsychiatric functions of autistic children. Sport Sci Health 2017;14(1):209-214. [doi: 10.1007/s11332-017-0417-y]
26. Kaur M, Bhat A. Creative yoga intervention improves motor and imitation skills of children with autism spectrum disorder. Phys Ther 2019;99(11):1520-1534 [FREE Full text] [doi: 10.1093/ptj/pzz115] [Medline: 31765484]
27. Pan CY, Chu CH, Tsai CL, Sung MC, Huang CY, Ma WY. The impacts of physical activity intervention on physical and cognitive outcomes in children with autism spectrum disorder. Autism 2017;21(2):190-202. [doi: 10.1177/1362361316633562] [Medline: 27056845]
28. Najafabadi MG, Sheikh M, Hemayattalab R, Memari A, Aderyani MR, Hafizi S. The effect of SPARK on social and motor skills of children with autism. Pediatr Neonatol 2018;59(5):481-487 [FREE Full text] [doi: 10.1016/j.pedneo.2017.12.005] [Medline: 29402579]
29. Desrochers K, Rivard MC, Dugas C. Analysis of a sensory-motor intervention program on the global and motor development of students with autism spectrum disorder [in French]. Can J New Sch Educ No 2017;8:1 [FREE Full text]
30. Valverde-Esteve T, Salvador-Garcia C, Gil-Gómez J, Maravé-Vivas M. Sustainable service-learning in physical education teacher education: examining postural control to promote ASD children's well-being. Int J Environ Res Public Health 2021;18(10):5216 [FREE Full text] [doi: 10.3390/ijerph18105216] [Medline: 34068915]

XSL•FO
RenderX

31. Rafie F, Ghasemi A, Zamani Jam A, Jalali S. Effect of exercise intervention on the perceptual-motor skills in adolescents with autism. J Sports Med Phys Fitness 2017;57(1-2):53-59. [doi: 10.23736/s0022-4707.16.05919-3]
32. Ayres AJ, Robbins J. Sensory Integration and the Child: Understanding Hidden Sensory Challenges, Pediatric Therapy Network ed. Los Angeles, CA: Western Psychological Services; 2005.
33. Polatajko HJ, Macnab JJ, Anstett B, Malloy-Miller T, Murphy K, Noh S. A clinical trial of the process-oriented treatment approach for children with developmental co-ordination disorder. Dev Med Child Neurol 1995;37(4):310-319. [doi: 10.1111/j.1469-8749.1995.tb12009.x] [Medline: 7535267]
34. May-Benson TA, Koomar JA. Systematic review of the research evidence examining the effectiveness of interventions using a sensory integrative approach for children. Am J Occup Ther 2010;64(3):403-414. [doi: 10.5014/ajot.2010.09071] [Medline: 20608272]
35. Baranek GT. Efficacy of sensory and motor interventions for children with autism. J Autism Dev Disord 2002;32(5):397-422. [doi: 10.1023/a:1020541906063] [Medline: 12463517]
36. Watling R, Hauer S. Effectiveness of ayres sensory integration® and sensory-based interventions for people with Autism spectrum disorder: a systematic review. Am J Occup Ther 2015;69(5):1-12. [doi: 10.5014/ajot.2015.018051] [Medline: 26356655]
37. Watling R, Deitz J, Kanny EM, McLaughlin JF. Current practice of occupational therapy for children with autism. Am J Occup Ther 1999;53(5):498-505. [doi: 10.5014/ajot.53.5.498] [Medline: 10500858]
38. Pfeiffer B, Daly BP, Nicholls EG, Gullo DF. Assessing sensory processing problems in children with and without attention deficit hyperactivity disorder. Phys Occup Ther Pediatr 2015;35(1):1-12. [doi: 10.3109/01942638.2014.904471] [Medline: 24712841]
39. Bardid F, Rudd JR, Lenoir M, Polman R, Barnett LM. Cross-cultural comparison of motor competence in children from Australia and Belgium. Front Psychol 2015;6:964 [FREE Full text] [doi: 10.3389/fpsyg.2015.00964] [Medline: 26217282]
40. Chow SM, Henderson SE, Barnett AL. The movement assessment battery for children: a comparison of 4-year-old to 6-year-old children from Hong Kong and the United States. Am J Occup Ther 2001;55(1):55-61. [doi: 10.5014/ajot.55.1.55] [Medline: 11216367]
41. Lohman TG, Roche AF, Martorell R. Anthropometric Standardization Reference Manual. Champaign, IL: Human Kinetics; 1988.
42. Mosteller RD. Simplified calculation of body-surface area. N Engl J Med 1987;317(17):1098-1098. [doi: 10.1056/nejm198710223171717]
43. Leone M, Viret P, Bui HT, Laverdière C, Kalinova E, Comtois AS. Assessment of gross motor skills and phenotype profile in children 9-11 years of age in survivors of acute lymphoblastic leukemia. Pediatr Blood Cancer 2014;61(1):46-52. [doi: 10.1002/pbc.24731] [Medline: 24115507]
44. Leone M, Comtois AS, Kalinova E, Perron M, Babineau C. Motor skills assessment of Canadian school boys and girls aged 6 to 12 years old. Université du Québec à Chicoutimi. 2010. URL: https://constellation.uqac.ca/id/eprint/2475/ [accessed 2025-03-05]
45. Gosselin V, Leone M, Laberge S. Socioeconomic and gender-based disparities in the motor competence of school-age children. J Sports Sci 2021;39(3):341-350. [doi: 10.1080/02640414.2020.1822585] [Medline: 32967566]
46. Aranha VP, Saxena S, Moitra M, Narkeesh K, Arumugam N, Samuel AJ. Reaction time norms as measured by ruler drop method in school-going South Asian children: a cross-sectional study. Homo 2017;68(1):63-68. [doi: 10.1016/j.jchb.2016.12.001] [Medline: 28011080]
47. Fleishman EA. The Structure and Measurement of Physical Fitness. Englewood Cliffs, NJ: Prentice-Hall; 1964.
48. Ángel Latorre-Roman P, Robles-Fuentes A, García-Pinillos F, Salas-Sánchez J. Reaction times of preschool children on the ruler drop test: a cross-sectional study with reference values. Percept Mot Skills 2018;125(5):866-878. [doi: 10.1177/0031512518789563] [Medline: 30032724]
49. Allisse M, Thibault I, Gagnon D, Kalinova E, Larivière G, Leone M. The relationship between obesity status and body image dissatisfaction on gross motor skill development and cardiorespiratory fitness in children aged 6–12 years old. Int J Environ Res Public Health 2025 Mar 12;22(3):417. [doi: 10.3390/ijerph22030417]
50. Attribution 4.0 International (CC BY 4.0). Creative Commons. URL: https://creativecommons.org/licenses/by/4.0/ [accessed 2025-03-19]
51. Cohen J. A power primer. Psychol Bull 1992;112(1):155-159. [doi: 10.1037//0033-2909.112.1.155] [Medline: 19565683]
52. Khoury E, Carment L, Lindberg P, Gaillard R, Krebs MO, Amado I. Sensorimotor aspects and manual dexterity in autism spectrum disorders: a literature review. Encephale 2020;46(2):135-145. [doi: 10.1016/j.encep.2019.11.005] [Medline: 31928741]
53. Puyjarinet F, Galbiati C. The evaluation approach for developmental coordination disorder in occupational therapy and psychomotor skills: points of convergence and specificities (in French). Le Pédiatre 2020;296:12-16.
54. Ashburner JK, Rodger SA, Ziviani JM, Hinder EA. Optimizing participation of children with autism spectrum disorder experiencing sensory challenges: a clinical reasoning framework. Can J Occup Ther 2014;81(1):29-38. [doi: 10.1177/0008417413520440] [Medline: 24783486]

XSL•FO
RenderX

55. Erickson KI, Hillman CH, Kramer AF. Physical activity, brain, and cognition. Curr Opin Behav Sci 2015;4:27-32. [doi: 10.1016/j.cobeha.2015.01.005]
56. Best JR. Effects of physical activity on children's executive function: contributions of experimental research on aerobic exercise. Dev Rev 2010;30(4):331-551 [FREE Full text] [doi: 10.1016/j.dr.2010.08.001] [Medline: 21818169]
57. Canli T, Lesch KP. Long story short: the serotonin transporter in emotion regulation and social cognition. Nat Neurosci 2007;10(9):1103-1109. [doi: 10.1038/nn1964] [Medline: 17726476]
58. Pedersen BK, Febbraio G. Muscle as an endocrine organ: focus on muscle-derived interleukin-6. Physiol Rev 2008;88(4):1379-1406 [FREE Full text] [doi: 10.1152/physrev.90100.2007] [Medline: 18923185]

Abbreviations

ASD: autism spectrum disorder
BH: body height
BM: body mass
BSA: body surface area
DCD: developmental coordination disorder
DSM-5-TR: Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, Text Revision
ES: effect size
PE: physical education
SRT: simple reaction time
UQAC-UQAM: University of Québec in Chicoutimi-University of Québec in Montréal

Edited by A Mavragani; submitted 26.08.24; peer-reviewed by K Suzuki; comments to author 22.01.25; revised version received 31.01.25; accepted 18.02.25; published 26.03.25.

Please cite as:
Touali R, Zerouaoui J, Chakir EM, Bui HT, Leone M, Allisse M
Impact of a Sensorimotor Integration and Hyperstimulation Program on Global Motor Skills in Moroccan Children With Autism Spectrum Disorder: Exploratory Clinical Quasi-Experimental Study
JMIR Form Res 2025;9:e65767
URL: https://formative.jmir.org/2025/1/e65767
doi:10.2196/65767
PMID:40137439

©Rachid Touali, Jamal Zerouaoui, El Mahjoub Chakir, Hung Tien Bui, Mario Leone, Maxime Allisse. Originally published in JMIR Formative Research (https://formative.jmir.org), 26.03.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Quality Assessment of Web-Based Information Related to Diet During Pregnancy in Pregnant Women: Cross-Sectional Descriptive Study

Daichi Suzuki[1,2*], MNS; Etsuko Nishimura[3*], PhD; Rina Shoki[4], PhD; Ishak Halim Octawijaya[5], MPH; Erika Ota[1], PhD

[1]Global Health Nursing, Graduate School of Nursing Science, St. Luke's International University, Tokyo, Japan
[2]Institute for Global Health Policy Research (iGHP), Bureau of Global Health Cooperation, Japan Institute for Health Security, Tokyo, Japan
[3]Department of Nursing, Faculty of Nursing, Komazawa Women's University, Tokyo, Japan
[4]Department of Nursing, Faculty of Nursing at Saitama, Japanese Red Cross College of Nursing, Saitama, Japan
[5]School of Nutrition and Dietetics, Faculty of Health and Social Services, Kanagawa University of Human Services, Kanagawa, Japan
[*]these authors contributed equally

Corresponding Author:
Erika Ota, PhD
Global Health Nursing
Graduate School of Nursing Science
St. Luke's International University
10-1 Akashi-cho
Chuo-ku
Tokyo, 1040044
Japan
Phone: 81 355502374
Email: ota@slcn.ac.jp

Abstract

Background: The widespread availability of health information online, coupled with the ease of access to the internet, has led pregnant women to rely heavily on online sources for pregnancy-related guidance. The internet-based information regarding nutrition enabled positive dietary changes for pregnant women. Although there are some important sources for pregnant women to collect their health information, some information increases maternal anxiety and difficulties based on a lack of information. Moreover, some women become confused due to conflicts on the same topics from different websites. However, concerns about the reliability and impact of this information have surfaced, contributing to heightened anxiety among expectant mothers. The importance of the quality of web-based information is increasingly recognized; however, no studies have evaluated the quality of nutrition-related information for pregnant women.

Objective: This study aims to bridge this research gap by assessing the quality of online health information concerning prenatal nutrition tailored to pregnant women.

Methods: This cross-sectional descriptive study was conducted through a Google keyword search on February 14, 2023. We used search terms, such as “pregnancy,” “pregnant women,” “diet,” and “nutrition” and conducted an exhaustive search on Google. Using the Quality Evaluation Scoring Tool (QUEST), we meticulously evaluated the quality of the retrieved information.

Results: The top 20 Google-searched sites were evaluated using the QUEST tool. The average score was 11.7 points, ranging from 6 to 15, with most sites scoring between 11 and 15. Half of the websites lacked clear authorship and most gave weak or no attribution to specific scientific sources. While conflict of interest scored highest overall, with 60% showing no bias, some sites promoted products or specific interventions. Currency was inconsistent—only half were updated within 5 years. Complementarity received the lowest scores, with 70% lacking support for patient-physician relationships. The tone was generally positive, with 95% supporting their claims, though only one site used a balanced, well-reasoned tone. Discrepancies in cited guidelines on nutritional intake and inappropriate expressions about alcohol, weight management, and miscarriage raised concerns about the information’s accuracy and appropriateness.

Conclusions: Although many websites use cautious language to mitigate commercial influence, deficiencies persist in crucial areas for empowering informed decision-making among pregnant women. From our assessment of the results, it was found that

XSL•FO
RenderX

incorrect evidence information is provided at the top of search results, which is easily accessible to users. The inadequacies in attributing authorship, clarifying conflicts of interest, and ensuring the currency of information pose substantial challenges to the reliability and usefulness of online health resources in prenatal nutrition. Since internet-based information is the most accessible, reliable evidence should be provided to protect everyone from misinformation, including shallow health literacy demographics, and from potential physical and psychological harm.

(JMIR Form Res 2025;9:e64630) doi:10.2196/64630

KEYWORDS

assessment; availability; decision-making; diet; dietary information; internet; internet-based; misinformation; nutrition; nutrition-related; online health information; physical harm; pregnancy; pregnancy-related guidance; pregnant women; prenatal nutrition; psychological harm; QUEST; quality assessment; tools; web-based; web-based information; website assessment; women's health

Introduction

Internet penetration rates are high in developed countries, resulting in easy web access for a substantial portion of the population, including pregnant women. In Japan, the internet usage rate reached 82.9% in 2021, with smartphones being the most prevalent devices (68.5%), followed by computers (48.1%) [1]. Owing to its accessibility, many individuals at different life stages, including pregnant women, depend on the internet to obtain health-related information. Among women receiving antenatal care in a hospital in Ireland, only 3% did not own a smartphone, but all of them had internet access to seek the information they needed [2].

Most women used the internet to access pregnancy-related information and found it beneficial [3]. Approximately 96% of pregnant women in Canada use the internet to find information about nutrition during pregnancy, with 75% reporting that they use the internet more often than any other source [4]. Women who are in their mid-20s to 30s, who were educated after high school, are employed, and are first-time mothers are more likely to obtain nutrition-related information from the internet [5]. Another study found that women with relatively low educational levels used websites less frequently than those with moderate or high levels of education. In addition, first-time mothers who reported higher levels of social support evaluated website information's quality more favorably [6]. Pregnant women search various nutrition-related topics online, including food safety and healthy eating (nutrient intake and requirements, serving sizes, recipes and meal plans, foods to avoid, etc) [4,7].

Nutrition-related information is crucial for pregnant women to address significant physiological, psychological, and social changes during pregnancy [8]. Without the appropriate information, pregnant women, especially first-time mothers, may not have a healthy pregnancy. Moreover, among healthy eating habits, adequate nutrient intake is necessary to prevent prenatal weight gain problems for both the mother and baby and to avoid particular nutrient deficiencies [9]. First-time mothers in Australia are reported to consume inadequate fruits and vegetables but excess soft drinks and fast food, which puts them at risk of micronutrient deficiency. Such poor eating habits are especially observed among pregnant women with lower socioeconomic status [10]. Since pregnant women with socioeconomically disadvantaged backgrounds likely have poorer access to health care services, including nutrition counseling with registered dietitians or other medical professionals, internet-based nutrition information is important to trigger healthy eating habits, especially among first-time mothers. Thus, the internet serves as a key source of prenatal and maternal information regarding nutrition for all pregnant women, enabling them to make positive dietary changes based on the information they find online.

However, accessibility and convenience of internet-based nutritional information may face some barriers, such as increased anxiety regarding the content described, lack of trust, and difficulty in finding information. Some pregnant women worry about their eating habits after obtaining information from the internet (are they capable of providing the adequate nutrients necessary for their babies?) or become concerned when they find they have consumed something bad for their babies. In addition, some pregnant women have found contradictory nutrition-related information on different websites, which confuses them regarding the information's reliability. Other women found it difficult to obtain accurate information on a specific topic. Improved access to trustworthy online sources and increased availability of information on different diets and health conditions can benefit pregnant women seeking nutrition-related information online [4].

A review of 18 Australian websites indicated that none were completely aligned with available evidence-based guidelines [11]. Although various online information exists on nutrition during pregnancy, no studies have evaluated the quality of this information in Japan. Therefore, this study aims to assess the quality of online health information on nutrition for pregnant women.

Methods

Selection of Websites

This cross-sectional descriptive study was conducted through a Google keyword search on February 14, 2023. Google Search is the most widely used search engine worldwide [12] and studies have used it to conduct research on health-related issues [13,14]. To search the websites, this study used terms such as "pregnancy," "pregnant women," "diet," and "nutrition." Based on these searches, the top 20 Japanese sites were selected for quality evaluation. Studies show that most internet users review only the top 10 search results [15]. To capture broader content, we analyzed the top 20 websites likely accessed for

XSL•FO
RenderX

pregnancy-related nutrition information. Searches were conducted in incognito mode with cookies cleared to minimize personalization effects. Although the search was conducted in incognito mode with cookies cleared to minimize personalization, we acknowledge that this approach does not fully eliminate all forms of personalization. Factors such as IP address tracking, device fingerprinting, and browser configurations may still influence search results. In addition, Google's search algorithm itself is not neutral and may introduce algorithmic biases that affect which websites appear at the top of search results. These factors could influence the representativeness and neutrality of the data collected [16,17].

Quality Evaluations

The Quality Evaluation Scoring Tool (QUEST) is a reliable and valid tool for evaluating online articles on health [16]. The QUEST assessed the information quality of the websites based on 7 weighted criteria: authorship, attribution, study type, conflicts of interest, currency, complementarity, and tone. The QUEST emphasizes balanced evaluation, prioritizing key indicators such as attribution, conflict-of-interest disclosure, and tone [16]. Textbox 1 presents the details of each QUEST component. Higher scores indicate better information quality, with the maximum attainable score of 28 [16].

Textbox 1. Details of each component of the Quality Evaluation Scoring Tool.

Authorship (score × 1)

0: No indication of authorship or username

1: All other indications of authorship

2: Author's name and qualification clearly stated

Attribution (score × 3)

0: No sources

1: Mention of expert source, research findings (though with insufficient information to identify the specific studies), links to various sites, advocacy bodies, or other

2: Reference to at least 1 identifiable scientific study, regardless of format (eg, information in text and reference list)

3: Reference to mainly identifiable scientific studies, regardless of format (in >50% of claims)

For all articles scoring 2 or 3 on Attribution (Score x 1)

Type of study

0: In vitro, animal models, or editorials

1: All observational work

2: Meta-analyses, randomized controlled trials, or clinical studies

Conflict of interest (score × 3)

0: Endorsement or promotion of intervention designed to prevent or treat condition (eg, supplements, brain training games, and foods) within the article

1: Endorsement or promotion of educational products and services (eg, books and home care services)

2: Unbiased information

Currency (score × 1)

0: No date present

1: Article is dated but 5 years or older

2: Article is dated within the last 5 years

Complementarity (score × 1)

0: No support of the patient-physician relationship

1: Support of the patient-physician relationship

Tone (includes title; score × 3)

0: Fully supported (authors fully and unequivocally support the claims; use strong vocabulary such as "cure," "guarantee," and "easy"; mostly use nonconditional verb tenses such as "can" and "will"; and include no discussion of limitations)

1: Mainly supported (authors mainly support their claims but with more cautious vocabulary such as "can reduce your risk" or "may help prevent" and include no discussion of limitations)

2: Balanced/cautious support (authors' claims are balanced by caution and include statements of limitations and/or contrasting findings)

In addition, content analysis was conducted on information regarding the recommended nutrient intake based on the "Dietary Reference Intakes for Japanese" (2020) [17] to examine each website's information content and expressions. A total of

XSL•FO
RenderX

2 reviewers (EN and RS) rigorously reviewed the top 20 eligible websites using a predefined content analysis framework in Microsoft Excel to ensure consistent evaluation of "wording of concern" and "accuracy of information." A health care professional with expertise in pregnancy and nutrition validated the analysis through repeated reviews. Prior to the full analysis, both reviewers conducted a pilot assessment on a subset of 5 websites to refine the coding framework and confirm a shared understanding of the evaluation criteria. Any discrepancies during the content analysis were resolved through consensus discussion, with additional input from a health care professional when necessary. This process ensured interrater agreement and strengthened the reliability of the evaluation.

Conceptual Model

Developed for higher education students, the conceptual model of engagement in online learning [18] incorporates diverse perspectives and a multidimensional structure. The model comprises 5 engagement types: cognitive engagement, behavioral engagement, collaborative engagement, emotional engagement, and social engagement.

In the context of this research, pregnant women who seek health information online are engaging in what can be referred to as "individual health engagement," which may be interpreted through the 5 elements of the aforementioned conceptual model.

The application of this model to pregnant women's engagement types above and justified as follows:

1. Cognitive engagement: pregnant women critically assess and integrate the health information they encounter online.
2. Behavioral engagement: actively searching for online health information reflects the development of self-efficacy and skills in managing pregnancy.
3. Collaborative engagement: pregnant women may interact with peers or healthcare professionals through online platforms.
4. Emotional engagement: the act of seeking health information online can elicit a range of emotional responses related to pregnancy.
5. Social engagement: the pursuit of health information online can help foster a sense of community among pregnant women.
6. Collectively, this framework highlights the multifaceted nature of pregnant women's engagement with online health information during pregnancy.

Ethical Considerations

The data used are publicly available and provided anonymously. In addition, this study did not collect data on or include any human subjects. Therefore, it was deemed that ethical approval was not required. The content was public and excluded sensitive or proprietary information. The analysis focused on aggregated and anonymized evaluations, ensuring compliance with copyright laws.

Results

QUEST Assessment

Google Search yielded approximately 11,300,000 sites and the top 21 sites comprised 20 regular pages and one listed page, with no video content included. Therefore, we conducted a quality assessment of the 20 regular pages. Table 1 presents the QUEST scores for the 20 pages. The average score was 11.7, ranging from 6 to 15, with most sites scoring between 11 and 15. No significant outliers were identified. The criterion of complementarity had the lowest average score, with 70% of the pages scoring low in this attribute, followed by attribution, and authorship. In contrast, the average scores for conflict of interest, currency, and tone of the content were higher.

Regarding authorship, half of the pages had no information on authorship or usernames. However, the remaining pages clearly stated their authors' information. In terms of attribution, only 1 page referred to at least one piece of scientific information. However, the remaining 65% (13 sites) referred to experts' opinions, research papers, or other sources without any evidence to identify specific studies. Moreover, 30% (6 sites) of the sites provided no sources. Although 1 site cited some sources, it referred to editorials with no evidence. To examine conflict of interest, this category was reported to have the highest average score among all the criteria. Of the sites, 60% contained information with no biases. A total of 3 sites presented recommendations, such as on educational products or services, with some biases. Moreover, 25% (5 sites) recommended specific interventions. As for currency, half of the sites (10/20 sites) provided the date and the information was written within the past 5 years. In addition, while 1 site provided the date, the information was older than 5 years. Meanwhile, the remaining sites provided no data sources. On the subject of complementarity, the sites scored lowest on this criterion. That is, 70% of the sites scored low on this criterion, meaning no support was provided for patient-physician relationships.

Finally, regarding tone, 95% of the sites (19/20) supported their claims, but 1 page provided more support with balance and caution. The highest tone score was achieved by a municipal website that used balanced language and provided clear reasoning for its recommendations. Despite the higher average score for this criterion, 45% of the sites showed no date present in the currency assessment.

XSL•FO
RenderX

Table 1. Assessments of quality of information in all included webpages.

Authors' assessment								Weighted score with authors' assessment							Total score (up to 28)
Number	Authorship	Attribution	Type of study	Conflict of interest	Currency	Complementarity	Tone (includes title)	Authorship (score × 1)	Attribution (score × 3)	Type of study (score × 1)	Conflict of interest (score × 3)	Currency (score × 1)	Complementarity (score × 1)	Tone (includes title)	
1	0	1	_[a]	2	2	0	1	0	3	0	6	2	0	3	14
2	2	O	–	2	0	0	1	2	0	0	6	0	0	3	11
3	2	2	0	0	0	1	1	2	6	0	0	0	1	3	12
4	2	0	–	0	0	1	1	2	0	0	0	0	1	3	6
5	2	0	–	2	0	0	1	2	0	0	6	0	0	3	11
6	2	1	–	1	2	0	1	2	3	0	3	2	0	3	13
7	0	1	–	2	2	0	1	0	3	0	6	2	0	3	14
8	0	1	–	2	0	0	1	0	3	0	6	0	0	3	12
9	2	0	–	2	2	0	1	2	0	0	6	2	0	3	13
10	2	1	–	1	2	0	1	2	3	0	3	2	0	3	13
11	2	1	–	2	0	0	1	2	3	0	6	0	0	3	14
12	0	0	–	0	2	1	1	0	0	0	0	2	1	3	6
13	0	0	–	2	2	0	2	0	0	0	6	2	0	6	14
14	2	1	–	0	0	0	1	2	3	0	0	0	0	3	8
15	0	1	–	2	0	0	1	0	3	0	6	0	0	3	12
16	0	1	–	2	1	0	1	0	3	0	6	1	0	3	13
17	0	1	–	2	2	1	1	0	3	0	6	2	1	3	15
18	0	1	–	2	2	1	1	0	3	0	6	2	1	3	15
19	0	1	–	0	0	0	1	0	3	0	0	0	0	3	6
20	0	1	–	1	2	1	1	0	3	0	3	2	1	3	12

[a]Not available.

Content Analysis

These sites were created by research and development organizations; research institutions; public institutions; and enterprises, such as those selling supplements, water, baby-related products, and insurance.

Each site was evaluated using content analysis to assess their "concerned terms of expression" and "accuracy of information." Regarding concerned terms of expression, we identified inappropriate explanations concerning alcohol consumption and weight management during pregnancy and miscarriage. For example, while national policy recommends complete abstinence during pregnancy, some websites recommend reduced alcohol consumption without providing scientific evidence. Japanese guidelines advise weight gain based on prepregnancy BMI, but some sites recommended a uniform gain of approximately 11 kg, contradicting these standards. Discrepancies were found in the advised daily intake of iron. While some websites aligned with evidence-based guidelines from the Health Service Bureau, Ministry of Health, Labour and Welfare Japan [17], others cited outdated or inconsistent recommendations.

Discussion

Principal Results

We identified the top 20 websites presenting diet-related information for pregnant women using Google search and then assessed them by using QUEST. The results showed that, overall, the conflict of interest, and tone criteria scored higher, suggesting that the information is unbiased and is presented in a balanced way. Transparency was a concern, with 45% of websites lacking publication dates. Lower scores for complementarity and attribution indicated insufficient support for patient-physician relationships and gaps in scientific reliability [16].

In particular, all 20 sites surveyed scored 1 or higher for tone, suggesting that the website authors supported their claims with

XSL•FO
RenderX

cautious statements but without discussing the claims' potential limitations [16]. Notably, only 1 site had the highest tone score of 2. A maximum score of 2 indicates that the author's claims are balanced by caution and state the potential limitations of their claims or contrasting evidence. Although all websites' authors avoided absolute claims regarding the nutrition-related information they provided, pregnant women should be cautious about the possible limitations or potential contradictory claims of other information sources. This should be an essential criterion of information quality, especially when attribution scores are low.

Surprisingly, we found that Japanese websites presented scientifically inaccurate information regarding pregnancy nutrition, as indicated by the low attribution scores in the QUEST criteria. One-third of the sites provided no scientific evidence for their information, whereas the remaining were mainly based on expert opinions and unidentifiable research findings.

The analysis revealed significant inconsistencies in the guidelines cited for key topics. For example, advice on alcohol consumption ranged from abstinence to unsupported permissiveness. Similarly, weight management guidance often ignored prepregnancy BMI standards. Advice on nutrient intake, especially iron, was also inconsistent, with some sites citing outdated guidelines. The lack of trustworthiness of nutrition-related information for pregnancy might discourage information seekers from using that information [4]. Such reluctance of pregnant women to seek information from the internet with misleading sources might save them from misinformed dietary habits and health behaviors during pregnancy. Individuals with only basic education or an annual income below the national average have been found to possess inadequate health literacy [19]. Those with inadequate health literacy might use any information on the internet, resulting in adverse health outcomes for the mother and child because of their inability to evaluate conflicting nutritional information provided by various sources [20].

In an internet environment, where nutritional information is easily accessible, reliable evidence must be provided to protect mothers and children from any potential physical and psychological harm caused by misinformed behavior. Incorrect information without evidence may be displayed at the top of the search results and is likely to lead to its provision to users. Any harm caused by such incorrect dietary information during pregnancy must be avoided by any means possible. In addition, mothers tend to be concerned about and sensitive to nutrition during pregnancy not only because of their own physical changes but also because they are related to the growth of the fetus [21,22]. Pregnancy is inherently a stressful condition and pregnant women are particularly susceptible to psychological and physiological anxiety [21]. Prior research also indicates that pregnant women often experience confusion when confronted with conflicting health information [20]. As outlined by uncertainty theory [23], ambiguity, inadequacy, or inconsistency in health information can heighten anxiety and stress among patients, leading to emotional decision-making. Therefore, incorrect information, such as the findings of this study, may contribute to maternal anxiety. Policy makers and health care professionals should collaborate to establish quality standards, promote trusted platforms, and disseminate scientifically grounded information. In addition, efforts should focus on enhancing pregnant women's health literacy to help them evaluate the reliability of information. In Japan, the government has introduced initiatives aimed at enhancing digital literacy and promoting fact-checking practices to verify the authenticity and timeliness of information [24]. However, public awareness of fact-checking remains relatively low compared to other developed countries [24]. This highlights the need to establish more effective systems for embedding such practices more broadly across society.

We need to highlight the low complementarity scores, which suggest a lack of support for the relationship between pregnant women and medical professionals. In particular, low complementarity scores of more than 70% of the sites suggest that, although the authors' claims are implemented, subsequent engagement and relationship between the reader and professionals may need to be considered. Despite the accessibility benefits of internet-based nutritional information, verbal information provided by health care professionals remains the most important source of nutritional information during pregnancy [7]. Various studies have highlighted the importance of nutrition-related information that health care professionals, such as midwives and dietitians, provide through counseling with pregnant women [25-27].

Limitations

This evaluation assessed the top 20 websites identified through Google Search using certain keywords and included only Japanese sites. This study could not assess the quality of all internet information, limiting the ability to generalize the results to Japan's overall web landscape. Nevertheless, as the analyzed sites appear at the top of search results, they attract substantial visitor traffic. Evaluating the content quality of these highly ranked sites may aid future users and serve as a reminder to online information providers about the importance of accuracy. Future studies should incorporate additional search engines and refine search strategies to provide a more comprehensive view.

While QUEST is widely used for evaluating online health information, its application in the specific context of pregnancy nutrition may have limitations. Adjusting the importance of the criteria to align with the needs of specific user populations, such as pregnant women with complications, might be necessary. In addition, algorithmic bias in search engine rankings should be acknowledged. Despite efforts to reduce personalization, Google's algorithm may reflect market trends or reinforce dominant perspectives, potentially limiting the diversity of visible content. While incognito mode helped minimize some personalization, other factors such as device fingerprinting or IP-based tracking may still have influenced the results. Future studies should consider using VPNs, anonymized environments, or multiple search engines to better account for these potential biases.

Conclusion

The identified websites demonstrated careful vocabulary selection, avoiding an emphasis on claims that could lead to

XSL•FO
RenderX

advertising or sales promotions. However, there was a notable lack of information supporting the patient-physician relationship, authorship details, and information update dates. Our findings underscore the need for improved standards in presenting online health information, particularly during pregnancy, to ensure accurate, reliable, and supportive content for pregnant women and health care providers.

Using web-based resources presents a convenient avenue for accessing desired information. Despite the presence of reputable and evidence-based sources, the prevalence of misleading information necessitates a critical approach. Hence, improving health literacy, especially regarding online information, is vital for readers to navigate and extract valuable insights from the multitude of online sources available. Future research should explore the role of digital health literacy in accessing reliable online health resources and promoting healthy behaviors during pregnancy.

Acknowledgments

This work was supported by a Japan Science and Technology Agency grant (JPMJPF2108) and a Ministry of Health, Labour and Welfare grant (JPMH22FA1022).

Conflicts of Interest

None declared.

References

1. Information and Communications in Japan White Paper 2022 (2022). URL: https://www.soumu.go.jp/johotsusintokei/whitepaper/eng/WP2022/2022-index.html [accessed 2025-05-14]
2. Kennedy RA, Mullaney L, Reynolds CM, Cawley S, McCartney DM, Turner MJ. Preferences of women for web-based nutritional information in pregnancy. Public Health 2017;143:71-77. [doi: 10.1016/j.puhe.2016.10.028] [Medline: 28159029]
3. Sayakhot P, Carolan-Olah M. Internet use by pregnant women seeking pregnancy-related information: a systematic review. BMC Pregnancy Childbirth 2016;16:65 [FREE Full text] [doi: 10.1186/s12884-016-0856-5] [Medline: 27021727]
4. Snyder A, Neufeld HT, Forbes L. A mixed-methods investigation of women's experiences seeking pregnancy-related online nutrition information. BMC Pregnancy Childbirth 2020;20(1):377 [FREE Full text] [doi: 10.1186/s12884-020-03065-w] [Medline: 32590955]
5. Al-Dahshan A, Kehyayan V, Al-Kubaisi N, Selim N. Utilization of the internet for physical activity and diet information and its influence on pregnant women's lifestyle habits in Qatar. Front Public Health 2023;11:1272109 [FREE Full text] [doi: 10.3389/fpubh.2023.1272109] [Medline: 37829092]
6. Vogels-Broeke M, Daemers D, Budé L, de Vries R, Nieuwenhuijze M. Sources of information used by women during pregnancy and the perceived quality. BMC Pregnancy Childbirth 2022;22(1):109 [FREE Full text] [doi: 10.1186/s12884-022-04422-7] [Medline: 35135487]
7. Lobo S, Lucas CJ, Herbert JS, Townsend ML, Smith M, Kunkler E, et al. Nutrition information in pregnancy: where do women seek advice and has this changed over time? Nutr Diet 2020;77(3):382-391. [doi: 10.1111/1747-0080.12589] [Medline: 31749295]
8. Marshall NE, Abrams B, Barbour LA, Catalano P, Christian P, Friedman JE, et al. The importance of nutrition in pregnancy and lactation: lifelong consequences. Am J Obstet Gynecol 2022;226(5):607-632. [doi: 10.1016/j.ajog.2021.12.035]
9. Cox JT, Phelan ST. Nutrition during pregnancy. Obstet Gynecol Clin North Am 2008;35(3):83-369. [doi: 10.1016/j.ogc.2008.04.001] [Medline: 18760225]
10. Wen LM, Flood VM, Simpson JM, Rissel C, Baur LA. Dietary behaviours during pregnancy: findings from first-time mothers in southwest Sydney, Australia. Int J Behav Nutr Phys Act 2010;7(1):13 [FREE Full text] [doi: 10.1186/1479-5868-7-13] [Medline: 20181084]
11. Cannon S, Lastella M, Vincze L, Vandelanotte C, Hayman M. A review of pregnancy information on nutrition, physical activity and sleep websites. Women Birth 2020;33(1):35-40. [doi: 10.1016/j.wombi.2018.12.007] [Medline: 30905558]
12. GlobalStats S. Search Engine Market Share Worldwide. URL: https://gs.statcounter.com/search-engine-market-share [accessed 2024-02-06]
13. Nuti SV, Wayda B, Ranasinghe I, Wang S, Dreyer RP, Chen SI, et al. The use of google trends in health care research: a systematic review. PLoS One 2014;9(10):e109583 [FREE Full text] [doi: 10.1371/journal.pone.0109583] [Medline: 25337815]
14. Cai HC, King LE, Dwyer JT. Using the Google™ search engine for health information: is there a problem? Case study: supplements for cancer. Curr Dev Nutr 2021;5(2):nzab002 [FREE Full text] [doi: 10.1093/cdn/nzab002] [Medline: 33937613]
15. Eysenbach G, Köhler C. How do consumers search for and appraise health information on the world wide web? Qualitative study using focus groups, usability tests, and in-depth interviews. BMJ 2002;324(7337):573-577 [FREE Full text] [doi: 10.1136/bmj.324.7337.573] [Medline: 11884321]

XSL•FO
RenderX

16. Robillard JM, Jun JH, Lai J, Feng TL. The QUEST for quality online health information: validation of a short quantitative tool. BMC Med Inform Decis Mak 2018;18(1):87 [FREE Full text] [doi: 10.1186/s12911-018-0668-9] [Medline: 30340488]
17. Dietary Refence Intakes for Japanese (2020) (2019). URL: https://www.mhlw.go.jp/content/001151422.pdf [accessed 2025-05-14]
18. Halverson LR, Graham CR. Learner engagement in blended learning environments: a conceptual framework. OLJ 2019;23(2):145-178. [doi: 10.24059/olj.v23i2.1481]
19. Svendsen MT, Bak CK, Sørensen K, Pelikan J, Riddersholm SJ, Skals RK, et al. Associations of health literacy with socioeconomic position, health risk behavior, and health status: a large national population-based survey among danish adults. BMC Public Health 2020;20(1):565 [FREE Full text] [doi: 10.1186/s12889-020-08498-8] [Medline: 32345275]
20. Meldgaard M, Jensen AL, Johansen AD, Maimburg RD, Maindal HT. Health literacy and related behaviour among pregnant women with obesity: a qualitative interpretive description study. BMC Pregnancy Childbirth 2022;22(1):712 [FREE Full text] [doi: 10.1186/s12884-022-05023-0] [Medline: 36123636]
21. Bjelica A, Cetkovic N, Trninic-Pjevic A, Mladenovic-Segedi L. The phenomenon of pregnancy — a psychological view. Ginekol Pol 2018;89(2):102-106. [doi: 10.5603/gp.a2018.0017]
22. Tarafa H, Alemayehu Y, Nigussie M. Factors associated with pregnancy-related anxiety among pregnant women attending antenatal care follow-up at Bedelle general hospital and metu karl comprehensive specialized hospital, southwest ethiopia. Front Psychiatry 2022;13:938277. [doi: 10.3389/fpsyt.2022.938277] [Medline: 36213901]
23. Mishel MH. Uncertainty in illness. Image J Nurs Sch 1988;20(4):225-232. [doi: 10.1111/j.1547-5069.1988.tb00082.x] [Medline: 3203947]
24. Ministry of Internal Affairs and Communications. Section 3. Spread of False and Incorrect Information on the Internet. URL: https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/r05/html/nd123400.html [accessed 2025-03-16]
25. Olloqui-Mundet MJ, Cavia MM, Alonso-Torre SR, Carrillo C. Nutritional education in the midwife's consultation room. Nutrients 2023;15(13):2906 [FREE Full text] [doi: 10.3390/nu15132906] [Medline: 37447232]
26. Kaleem R, Adnan M, Nasir M, Rahat T. Effects of antenatal nutrition counselling on dietary practices and nutritional status of pregnant women: a quasi-experimental hospital based study. Pak J Med Sci 2020;36(4):632-636 [FREE Full text] [doi: 10.12669/pjms.36.4.1919] [Medline: 32494246]
27. Super S, Beulen YH, Koelen MA, Wagemakers A. Opportunities for dietitians to promote a healthy dietary intake in pregnant women with a low socio-economic status within antenatal care practices in the Netherlands: a qualitative study. J Health Popul Nutr 2021;40(1):35 [FREE Full text] [doi: 10.1186/s41043-021-00260-z] [Medline: 34332642]

Abbreviations

QUEST: Quality Evaluation Scoring Tool

Edited by A Mavragani; submitted 22.07.24; peer-reviewed by L Aljerf; comments to author 26.11.24; revised version received 30.12.24; accepted 30.04.25; published 03.06.25.

Please cite as:
Suzuki D, Nishimura E, Shoki R, Octawijaya IH, Ota E
Quality Assessment of Web-Based Information Related to Diet During Pregnancy in Pregnant Women: Cross-Sectional Descriptive Study
JMIR Form Res 2025;9:e64630
URL: https://formative.jmir.org/2025/1/e64630
doi:10.2196/64630
PMID:

©Daichi Suzuki, Etsuko Nishimura, Rina Shoki, Ishak Halim Octawijaya, Erika Ota. Originally published in JMIR Formative Research (https://formative.jmir.org), 03.06.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Engaging Community-Dwelling Older Adults in Research: Qualitative Substudy of Factors Impacting Participation

Bryah Boutilier[1], MA; Grace Warner[2], PhD; Brianna Wolfe[1], MPA; Sorayya Askari[2], PhD; Elaine Moody[3], PhD; Parisa Ghanouni[2], PhD; Tanya Packer[1,2,4], PhD

[1]School of Health Administration, Dalhousie University, Halifax, NS, Canada
[2]School of Occupational Therapy, Dalhousie University, Halifax, NS, Canada
[3]School of Nursing, Dalhousie University, Halifax, NS, Canada
[4]Department of Nursing, Umeå University, Umea, Sweden

Corresponding Author:
Bryah Boutilier, MA
School of Health Administration
Dalhousie University
6299 South St
Halifax, NS, B3H 4R2
Canada
Phone: 1 9025777936
Email: brboutilier@dal.ca

Abstract

Background: Innovative approaches to community-level data collection are crucial to inform policies and programs that support people in aging well within their communities. For example, community-level data can proactively identify unmet health needs, inform preventative care strategies, and ensure the equitable distribution of resources that enable older adults to age in place.

Objective: This paper presented a substudy of a larger community-based project designed to identify community-dwelling older adults' concerns about their well-being and connect them with resources to help them age well at home. The substudy aimed to identify motivations that influence older adults' engagement in research and barriers to their participation.

Methods: Data collection involved qualitative semistructured interviews with 27 older adults, with a mean age of 77 (SD 5.4), who had completed a comprehensive assessment. Purposeful sampling prioritized older adults who lived in rural areas, had more than one health condition, and represented diverse ethnicities, while attempting to reach equal numbers of participants across the participating communities. Interviews were conducted by trained research team members using an interview guide focused on reasons for research participation and perceptions of the assessment and resource action plan. Meeting minutes, gathered during 35 biweekly or monthly sessions with community coordinators, captured real-time reflections on recruitment processes, challenges, and community-specific factors influencing participation. Thematic analysis was completed using both inductive and deductive approaches.

Results: Older adult participants were primarily female (n=22, 82%), of European (n=19, 70%) or Acadian (n=8, 30%) descent, university educated (n=14, 52%), with one or more chronic health conditions (n=26, 96%). Older adults reported 2 main reasons for participating: planning for the future and helping their community. At the same time, barriers to participation identified included communication challenges, fear of scams, and institutional skepticism. Participants emphasized a desire for practical outcomes from the research, especially related to aging-in-place supports. Although trust in local, personal relationships facilitated participation, skepticism toward institutions and digital communication channels were barriers to participation.

Conclusions: This research highlighted the need to tailor communication strategies to older adults by understanding factors influencing engagement. Addressing institutional skepticism and leveraging trusted community members are possible strategies to overcome barriers to successful engagement in community-based research. These findings advance our understanding of why older adults participate in research and suggest ways to improve recruitment strategies. Participation was motivated not only by personal benefit but also by a strong sense of civic responsibility, social connection, and a desire to contribute to future community well-being. Framing research as community-driven and future-oriented, rather than problem- or deficit-based, studies can resonate more deeply with older adults. Integrating research within existing, trusted local networks and venues helps build legitimacy and accessibility—especially in rural contexts where institutional trust may be low and digital communication less effective.

XSL•FO
RenderX

(JMIR Form Res 2025;9:e74191) doi:10.2196/74191

KEYWORDS

older adult; aging in place; research participation; community-based research; engagement

Introduction

Background

Understanding the health concerns among older adults living in the community is vital for adequate, preventative health care planning, resource provision, and allocation [1]. Because much of the existing data on health issues among older adults are derived from hospital and institutional settings, it is imperative to gather community-level data to inform policies and programs that support aging well [2-4]. Significant barriers [5,6] exist to obtaining comprehensive community-level data [7]. Engaging older adults in research is one solution to identify resource needs to support older adults to age in place [8].

Understanding what influences older adults to participate in community-level research and contribute data is critical to research best practices [7,9]. Barriers to participation previously identified include trust, confidentiality concerns, and communication [5,6]. Overcoming these barriers is essential to collect meaningful community-level data. Motivations that drive older adults to engage in research have also been identified. Caring about the outcomes of the study, personal characteristics (ie, curiosity, civic engagement, interest in the topic), and altruistic benefits appear to influence participation [1].

This paper presented a qualitative substudy that emerged from the *ACTing Collectively* project. The main study assessed the feasibility and utility of an innovative tool that gathers data on the needs of older adults living in the community and matches these to local resources. The substudy analysis of gathered qualitative data examined the research question "What motivates or hinders older adults' participation in community-based health research?"

Study Context: The ACTing Collectively Project

The overall study, *ACTing Collectively*, was carried out in Nova Scotia, Canada, with recruitment in 3 provincial municipalities, which were selected through a competitive expression of interest process involving 3 stages. After attending a webinar, interested municipalities were selected based on the following inclusion criteria: (1) a completed expression-of-interest form (ie, containing information about demographic characteristics of older adults in the municipality, the level of social and material deprivation, and the regional use of medical care), (2) an identified individual who would serve as a community coordinator to help with participant recruitment and participate in regular team meetings, and (3) willingness and capacity to act on recommendations based on project results. Submitted expression-of-interest forms allowed researchers to assess the municipal capacity and select municipalities with higher provincial deprivation and medical use levels. The following sections outline information specific to the methodology used in the *ACTing Collectively* project.

The *ACTing Collectively* project tested the feasibility of the Age Care Technologies Assess & Connect (ACT) tool and process to collect community-level data and match available resources to the specific needs of each older adult. The ACT assessment tool consists of 56 questions that assess well-being, independence, social engagement, and health, as well as a customizable resource database that can connect relevant resources to each of the 56 questions. A resource action plan is then generated and given to the older adult, listing their prioritized concerns and selected resources.

Various recruitment strategies were used to raise awareness about the *ACTing Collectively* project. These included infographics and flyers posted at different community locations (eg, health centers, churches, seniors' clubs, libraries, and municipal offices). Printed materials outlined the research objectives, participation requirements, benefits, and project contact information (ie, phone number and email). Newspaper articles, radio/television interviews, newsletters, and posts on social media platforms supplemented this. To build relationships with the participating municipalities, the research team travelled to research sites to present the project at local events aimed at older adults. Community coordinators also continually promoted the project and recruited older adults by delivering presentations at seniors' events. Recruitment messaging was informed by the community coordinators and included wording designed to motivate older adults (eg, participation may help your friends or other struggling older adults in the community, it may help you in the future, or it may connect you with resources).

Inclusion criteria for the ACT assessment were (1) being a community-dwelling older adult aged 65 years or more, (2) living in 1 of the 3 collaborating municipalities, and (3) being able to speak and understand English or French. Individuals who could not read/write English or French were eligible, given that a trained assessor conducted all assessments in person, virtually, or over the phone. Additionally, family caregivers (unpaid) for older adults who met the inclusion criteria were invited to participate on behalf of the older adults if barriers to participation existed (eg, health condition, hearing impairment). The single exclusion criterion was severe cognitive or psychological impairment that would hinder participants' understanding of the assessment questions, assessed by their ability to follow instructions.

Older adults interested in participating contacted the research team to obtain information, establish eligibility, and provide informed, voluntary consent. If the family caregiver was participating on behalf of the older adult, consent was received from the caregiver and the older adult. Each participant was assigned a participant ID to protect confidentiality. The ID was used in all subsequent data collection, and the decoder was kept in a password-protected file.

XSL•FO
RenderX

Participants chose virtual, phone, or in-person interviews. The interviews were conducted by trained assessors who signed confidentiality agreements and received procedural training.

Methods

Substudy: Understanding the Barriers to and Facilitators of Research Involvement

This paper reported findings from (1) follow-up interviews with a sample of people who participated in the ACT assessment and (2) minutes and research study documents to better understand the barriers to and facilitators of participation in a community research project. Using a qualitative design, the research questions were as follows:

- Why do older adults choose to participate in community-based research?
- Why do older adults choose not to participate in community-based research?

Sampling and Data Collection for the Qualitative Substudy

A subsample of older adults who completed the ACT assessment and provided consent to be recontacted were contacted by the research team 2-4 months after the ACT assessment to reconfirm their interest in participating in a follow-up interview and verbally complete the substudy informed consent process. Purposeful sampling prioritized older adults who lived in rural areas, had more than one health condition affecting their everyday life, and represented diverse ethnicities, while attempting to reach equal numbers of participants across the 3 municipalities. Five rounds of sampling were conducted throughout the study.

Those who agreed to be interviewed participated in an approximately 1-hour interview conducted over the phone by a research team member trained in qualitative interviewing techniques, such as probing. The interview was designed to determine the usefulness and feasibility of the ACT assessment and resource action plan. Areas of inquiry included descriptions of their living situation, experience completing the ACT assessment, applicability of the questions to their well-being, reasons for participation in the research, facilitators of and barriers to accessing resources identified in their resource action plan, and opinions on the usefulness of the resource action plan. This substudy concentrated on the responses to the questions regarding older adults' reasons for participating in the study.

In addition to interviewing older adults, meeting minutes with community coordinators and other research partners were analyzed. The project's local community coordinators were primarily responsible for on-the-ground participant recruitment and often completed the assessment. Given that data could not be gathered from those who did not agree to participate in the parent project, information from the community coordinators was key to identifying possible barriers to participation. During the 2 years of the project, there were biweekly or monthly meetings (N=35 in total) to discuss and monitor the progress and challenges in each municipality. Meeting minutes captured discussions on recruitment efforts, barriers, and new recruitment strategies. The minutes provided an opportunity to understand the community coordinators' perspectives on why older adults chose to or did not participate in the project. Community coordinators agreed and consented to the use of minutes in data collection. All meeting minutes were used in the analysis.

Analysis

Any identifying information was removed during the transcribing process. Pseudonyms were created for older adult participants using popular names for individuals aged over 65 years. Meetings were audio-recorded and written minutes produced by a research team member. Meeting minutes were sent to all attendees, including community coordinators, to ensure accuracy before finalizing the minutes. Any errors noted were corrected. Pseudonyms were not assigned to community coordinators due to the small sample size and risk of identification. Instead, quotes taken from minutes were not attributed.

The interviews and meeting minutes were uploaded into NVivo12 for Mac qualitative software for analysis (Lumivero). A codebook was created using a combined inductive and deductive approach. The deductive framework was based on questions in the interview guide, while the inductive approach allowed unexpected findings to arise. Combining inductive and deductive approaches recognizes existing theoretical knowledge, while also allowing new insights to emerge from the data [10]. The same codebook was used to analyze interviews and meeting minutes focused on strategies to engage older adults, ways to increase participation, and challenges experienced. Qualitative thematic analysis was used to organize the codes into patterns that evolved into themes [11]. Next, through an iterative review, themes were defined, named, and validated to represent the data accurately. Older adults' interviews and meeting minute data were analyzed separately, and themes were compared and integrated. Member checking and intercoder reliability processes were implemented to ensure trustworthiness was enacted and the data were collected, analyzed, and interpreted accurately.

Reflexivity

As researchers, we acknowledge that our personal characteristics, professional backgrounds, and prior experiences have the potential to shape all stages of the research process. The authors brought diverse professional backgrounds to the study, including occupational therapy, health administration, nursing, public policy, and health promotion. The research team comprised both senior researchers with extensive experience and newer research staff. We recognize that our attributes, such as all authors being female, may have influenced our interactions with participants and the interpretation of findings. Efforts were made to remain reflexive throughout the research process by meeting regularly to minimize bias and enhance the credibility and transferability of the results.

Ethical Considerations

This research was approved by the Nova Scotia Health Research Ethics Board (ethics approval number: 1027295) and followed all requirements to protect participants' privacy and confidentiality. All participants and community coordinators provided written or verbal informed consent before participation.

XSL•FO
RenderX

Demographic data were collected via data entry by assessors. No personal identifiers were entered. Participant contact details, used to make appointments, were kept separate from the data. Interview data were deidentified during transcription and prior to analysis. Identifiable personal information was kept in a secure drive, and findings were reported in a way that ensured anonymity. Participants were entered into a CAD $100 gift card draw and received a local resource package (ie, resource action plan) in exchange for participation. No images or identifiable information were used in this manuscript.

Results

Participant Details

Of the older adults who completed the ACT assessment, 186 (84%) also agreed to be contacted by phone to participate in a follow-up interview. Using maximum variation sampling, 44 (24%) participants were contacted, and 27 (61%) of those contacted agreed to participate. Participants who declined to participate indicated they were too busy or no longer interested. The sample had a mean age of 77 years, and based on postal codes, [12] 48% (n=13) lived in rural areas. Participants were able to choose more than one ethnicity. They were predominantly of European (n=19, 70%) or Acadian descent (n=8, 30%), aligning with the local population's demographic data for this age range [13]. The majority of older adult participants were female (n=22, 82%), and approximately 52% (n=14) reported some level of university education. Almost all participants reported having at least one chronic health condition (n=26, 96%). Demographic information for community coordinators was not provided, given the small sample size (n=3). Although participants varied in gender, location, and ancestry, we did not purposively sample according to these characteristics, and they did not emerge as strong themes in our analysis. In addition, limited gender variation in our larger parent study restricted our ability to purposively sample men for this substudy, thereby limiting analysis of thematic trends across gender.

Themes and Subthemes

Two main themes emerged to explain what motivates or hinders older adults' participation in research (Figure 1) and were organized into subthemes: theme 1 (What motivates participation?), with subthemes 1A (*planning for the future*) and 1B (*helping their community*), and theme 2 (What hinders participation?), with subthemes 2A (*communication challenges*), 2B (*fear of scams*), and 2C (*institutional skepticism*).

Figure 1. Thematic map of themes and subthemes exploring factors influencing older adults' (65 years and older) participation in community-based research.

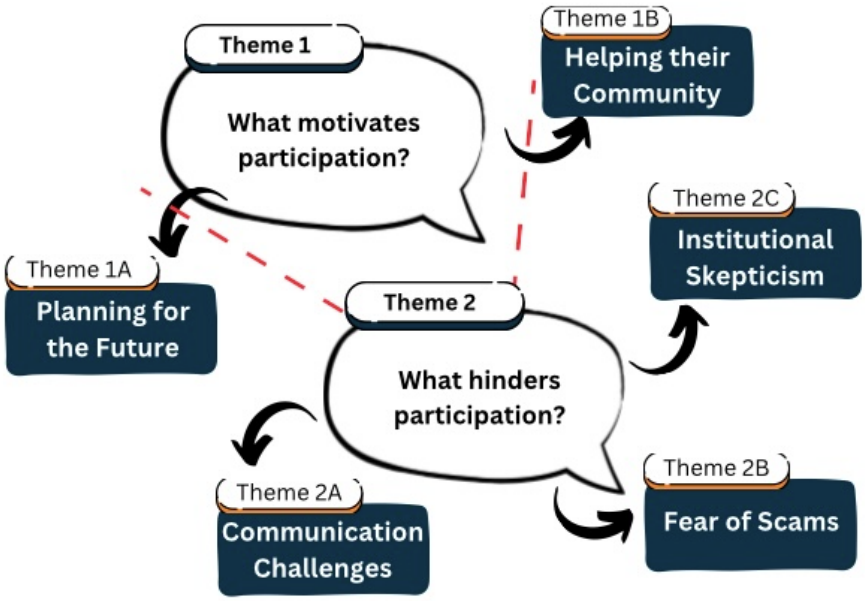

Theme 1: What Motivates Participation?

Older adults' interviews informed theme 1. Many participants expressed that their primary motivations for engaging in the research were personal concerns, such as planning for their future or contributing to the well-being of their community.

Subtheme 1A: Planning for the Future

"Planning for the future" was a dominant theme. Topics such as financial planning, home maintenance or improvement, access to helpful resources (ie, services, programs), and social support systems were areas of particular interest. Participants demonstrated a proactive approach to future planning by drawing a direct link between participation in the research and understanding available resources for future needs. They shared their concerns about aging as the reason for finding out what resources are available, allowing them to plan for the future. The following quotes illustrate participants' acknowledgment of future needs and the importance of learning how to put plans into place for the future.

> *I think the reason I put myself forward for this was to just find out what's available for seniors, like if you needed help . . .* [Trina]
>
> *That's exactly why we got into this sort of study that you're doing . . . We figured one of these days we may*

XSL•FO
RenderX

> *need something to help out, and if we have some of this information, we'll know who to call.* [Thomas]

> *A lot of it is learning about, you know all the issues that come up . . . it's not too long in the future either that could be problematic for us . . . and you want to make sure that things get put into place before because we are an aging population.* [Margaret]

In addition to a general desire to plan for the future, motivation for many older adult participants was driven by specific concerns about future health or mobility challenges and the need for home modification. When participants described strategies to remain in their homes longer, they spoke of maintenance, upkeep, and accessibility services. Participants articulated the challenges older adults face in maintaining their homes, reflecting on the need for supportive or informal services, such as lawn services from a neighbor. Some older adult participants had already experienced poor health, preventing them from caring for their homes. These experiences were at the forefront of their minds and added to the urgency of identifying support to remain at home.

> *Last year after surgery, I had to get somebody to do the lawn . . . I'm on like 4 acres of land here . . . But I don't know how long I can [keep maintaining land] . . . I also was just sent some scans, and I have spinal stenosis . . . you know, it's not too bad right now, so hopefully, that doesn't become a big issue with me because that could have a bearing on whether I can stay in my home.* [Trina]

Subtheme 1B: Helping Their Community

Many older adult participants were motivated by a sense of civic responsibility and a desire to effect positive change in their communities. They emphasized the importance of active involvement in research, seeing it as a proactive way to address community issues.

> *We have to speak out on behalf of our senior citizens or our people in the community. We can't just sit back and be armchair enthusiasts and say nothing but complain about it all the time.* [Anne]

Altruistic motivations were prevalent among participants, as many desired research findings to benefit others in their community. Many acknowledged that the research experience did not necessarily have a direct personal impact on them, but they hoped that information could benefit others.

> *The whole experience with [assessor's name] has not done anything for me personally, but I'm hoping that with the information you get, it may help other people.* [William]

A perceived neglect of older adults within the community fueled many to participate. OOne older adult (Helen) participated "just because the word 'senior' was in it . . . because it [the research] talked about seniors, and we're often a forgotten group.

Participants expressed a perception that not enough is being done to improve the well-being of older adults in their community, underscoring the importance of amplifying the voices of older adults by incorporating their perspectives into community initiatives.

> *I think checking in with any older person gives them a voice that they don't have in the ordinary community life.* [Elizabeth]

The following quote highlights the sense of neglect among older adults, prompting a desire to be heard and considered in community initiatives such as research:

> *And then they say, well, they're doing this for seniors and that for seniors. Well, forget it because people like us [seniors] are thrown to the curb. That's why I did it [participated in the research].* [Catherine]

Older adult participants expressed optimism that the research would be able to understand and address the aging population's needs, hoping for positive changes in the well-being of older adults in their community. Participants shared their positive feelings about the research, with one participant (Dorothy) saying, "I'm pleased to know that this program is put together and that at least trying to find out what can be done for the aging population . . . When you age you just want to be able to live well."

Theme 2: What Hinders Participation?

Both older adult interviews and meeting minutes informed theme 2. Three subthemes were identified: *communication challenges*, *fear of scams*, and *institutional skepticism.*

Subtheme 2A: Communication Challenges

As noted by community coordinators, rural communities face the dual challenge of geographic dispersion and reduced availability of print newspapers, prompting a reliance on digital publications that may bypass older adults who are less familiar with the technology. This was echoed by older adult participants who preferred receiving information through traditional means, such as print or phone communication. One older adult participant (Ruth) said, "I'd say there's a lot of older people that are not with the computers and that. They just need to get it [hear about research] on paper or phone because they can't go online."

Effectively disseminating information to older adults, with limited access to traditional communication channels, such as newspapers, challenged project recruitment. One community coordinator stated, "I anticipate that our greatest challenge will be to educate the public about the project and recruit enough participants to be assessed!"

Subtheme 2B: Fear of Scams

A fear of falling victim to fraudulent activities hindered the willingness to engage in research. Despite the research team's efforts to reassure participants through clear communication protocols, legitimate concerns about scams remained. One community coordinator stated, "There is a lot of fear in communities right now about fraud and scams, and I do think that is probably having an impact on recruitment."

Older adult participants also shared their hesitancies with phone and email technology because of scams:

XSL•FO
RenderX

> *The only thing I think of now are so many scams that are out there . . . like I got scammed once or twice before I even knew anything about a scammer.[Ruth]*

Calling older adult participants and leaving voicemails multiple times, naming and identifying relationships with the project and always telling participants when to expect a call-back appeared to mitigate this fear but only to a limited extent. Older adult participants agreed that this was useful in making them feel more secure.

> *I think it was good the way you called ahead of time and set an appointment time. Just there's so many scammers out there now, so it's good to know that, okay, you're going to call somewhere around 11 o'clock on such and such a day . . . So many elderly people don't realize when somebody calls [that it could be a scam].* [Helen]

Subtheme 2C: Institutional Skepticism

Broader societal attitudes toward institutional entities, including skepticism toward government-funded initiatives and research endeavors, were also reported as barriers to participation. A community coordinator spoke of resentment in their communities for things that seemed to be government-related, including research:

> *What I hadn't anticipated was the level of resentment from the public toward the government and large bodies and research . . .*

Participants' reluctance to engage with government-funded navigation services, such as toll-free information phone lines, reflected a more profound mistrust of bureaucratic institutions and a preference for local community-based support networks.

> *I'm going to tell you right now; I don't have to call them [government navigation agencies] when I know that they're going to be useless . . . Anything to do with the government, you're going to get a run-around.* [Martha]

This skepticism was also reflected in attitudes toward publicly funded research. A lack of engagement, even at in-person events, was interpreted by community coordinators as further proof of public skepticism:

> *There is resentment about pouring resources into "research" instead of simply meeting basic needs.*

> *I was able to offer a presentation about the ACTing Collectively project . . . It was surprising that even though I put the sign-up sheets that [the research team] sent yesterday on each table and asked everyone present to consider adding their name to indicate that they agree to be contacted to consider participating, only 4 of about 50 people provided their contact information.*

Leveraging personal relationships and trusted community champions was observed to overcome or mitigate this institutional skepticism. Community coordinators shared their thoughts:

> *Seeing the research team relax during the [older adult event] visit was valuable . . . It was nice to make the personal connections with them and the seniors.*

However, community coordinators still felt that "recruiting is best done by someone in the community . . . someone who is trusted," emphasizing the importance of involving local partners.

Discussion

Principal Findings

This study explored the factors that influence older adults' participation in research and the barriers that hinder such engagement. Older adults were motivated to participate in planning for the future and helping their community.

Motivations for Participation

Proactive Approaches

Participants in this study echoed what is known about planning for the future, linking this to their decision to participate in the research. A fear of loss of autonomy, independence, and frailty underpin older adults' need to plan for the future [14-16]. A proactive approach is positively associated with dimensions of successful aging, such as the ability to care for oneself and adjust to age-related changes [17]. Conversely, the literature also states that higher rates of depression in older adult populations are associated with avoidance of planning for the future [18,19]. Although most participants in this study had one or more chronic health conditions, the sample was relatively healthy and educated, aligning with known research about populations that engage in research [20].

Concerns about mobility challenges and the need for home modification emerged as motivations for participation. Research highlights the interconnection between health challenges during the aging process and the need for home maintenance [21]. Additionally, participants in this study recognized the importance of socialization for overall well-being, including support for others in their community. This may have underpinned their motivation to plan for the future, as it echoes the existing research that social isolation is negatively associated with positive mental health, successful aging [22], and quality of life in older populations [14,19].

Civic Responsibility

Participants' motivations reflected a sense of civic responsibility and a commitment to effecting positive change within their community. Similar to our findings, previous studies examining volunteerism among older adults have found altruism to be a primary motivator for engagement in activities that may benefit the community [1]. Active engagement in research was perceived as a means to address community issues proactively, aligning with previous studies emphasizing the role of community engagement in promoting well-being among older adults [20,23,24].

Participants in this study expressed concerns about the perceived neglect of older adults within the community, highlighting the importance of incorporating their perspectives into community initiatives. Studies have shown that older adults often feel marginalized or overlooked in society, leading to a desire for

XSL•FO
RenderX

greater inclusion in the community and recognition of their needs [19,25]. The existing literature emphasizes the need to amplify the voices of older adults in shaping policies and interventions to address their unique needs [26,27], which may motivate older adults to believe in their ability to shape their community.

Barriers to Participation

The literature highlights the importance of using multiple communication channels to reach diverse populations, particularly those with limited digital literacy [28]. Even though print newspapers are less available in rural areas, this study found that published newspapers are a preferred recruitment method. Additionally, the move to digital publications makes it even harder for older adults who are not tech-savvy to stay informed. Strategies such as personalized outreach and community-based intermediaries may help bridge this communication gap [29].

In this study, older adults were skeptical about scams and adopted behaviors to protect themselves, such as not answering phone calls and confirming caller identification. Research indicates that older adults are particularly vulnerable to scams and fraud due to social isolation and cognitive decline [30]. Using transparent and secure recruitment methods, including providing educational resources on recognizing and avoiding scams, may help alleviate these concerns and enhance trust in research initiatives [31].

Skepticism toward institutional entities, particularly government-related initiatives, presented a challenge to research recruitment efforts in this study. In rural communities, skepticism may be amplified due to a history of perceived neglect of resources and health care services [32]. Deep-seated mistrust of bureaucratic institutions and a preference for local community-based supports hinder engagement with research because it is perceived as associated with the government. This finding underscores the importance of building personal relationships and leveraging trusted community members as recruitment champions to overcome institutional skepticism. Engaging with local partners and involving community members in the research process can foster trust and legitimacy, ultimately enhancing participation [33,34].

Implications for Researchers

This study underscored the importance of aligning research approaches with participants' motivations, values, and communication preferences. Participation was notably influenced by how the parent study was framed. Emphasizing local needs, future planning, and community benefit resonated deeply with older adults' desire to help others, give back to their community, and have their voices heard. This community-oriented framing—rather than a focus on personal diagnosis or treatment—had positive benefits that may have reduced stigma and increased comfort with participation. Recruitment strategies that leverage relationships and trust held by local partners (eg, Seniors' Safety Program coordinators, community health workers) can play a critical role in overcoming institutional skepticism. These trusted intermediaries help validate the project's legitimacy and make initial contact more approachable.

Additionally, using physical recruitment materials in familiar, high-traffic community locations (eg, post offices or pharmacies) and offering brief, in-person researcher outreach during existing gatherings (eg, church events, library talks), provide accessible and trusted touchpoints. To further reduce concerns about scams—a recurring barrier in this study—teams should also consider providing verification tools, such as local phone numbers or letters of affiliation from known organizations.

Together, these insights suggest that the way research is framed and embedded within community relationships can meaningfully influence older adults' participation. By tailoring strategies to reflect older adults' values and preferred modes of engagement, researchers can foster greater inclusion, trust, and long-term partnership in community-based aging studies.

Implications for Policy

This research highlights several implications for policy, particularly in relation to aging populations. Participants' strong interest in planning for the future identifies the need for communities and organizations to develop and tailor resources that are proactive and prevention focused, rather than reactive. Policies should support the design and implementation of programs that help older adults maintain independence and well-being before significant health or social challenges arise. In addition, the findings emphasize the value of delivering services at the local or community level. Older adults often prefer to engage with familiar individuals and trusted organizations, suggesting that investment in community-based service infrastructure can enhance accessibility, trust, and uptake.

Limitations

A key limitation of this study is that the data were collected from individuals who ultimately chose to participate in the research. As such, their insights into participation barriers may not have fully captured the perspectives or motivations of those who declined to participate, potentially limiting the generalizability of findings related to participation hesitancy. However, the inclusion of data from the community coordinators partially mitigated this limitation. Additionally, since the study sample was relatively healthy and well educated, motivating factors and barriers hindering participation reflect older adults who participated. This may have led to underestimating barriers or overestimating motivations to participate on behalf of their future self or community. If a more vulnerable sample of older adults had been recruited, there may have been more individuals who were seeking resources for themselves or identified additional accessibility barriers. Lastly, this study recruited a limited number of males. We explored the possibility of identifying thematic trends across gender, location, and ancestry; however, we had not purposively sampled these characteristics, nor did they emerge as strong themes in our analysis. To explore these differences, future studies should consider sampling strategies that allow for this type of analysis.

Future research should develop and test methods to ethically engage vulnerable older adults, such as those of a lower

XSL•FO
RenderX

socioeconomic status, with less education, and with more chronic health conditions. Additionally, looking at the factors impacting engagement with research among diverse ethnicities and cultures would be informative to better reach marginalized populations and broaden perspectives on findings.

Conclusion

In summary, recognizing the factors influencing research engagement among older adults informs future studies on how they may maximize the recruitment of older adults. Highlighting the proactive involvement and altruistic motivations of older adults in research emphasizes the importance of incorporating their perspectives into study design. By comprehensively understanding the concerns and viewpoints of older adults, researchers can devise tailored recruitment strategies, interventions, and support mechanisms for research. Moreover, addressing communication barriers, apprehensions regarding scams, and institutional distrust is essential in fostering relationships to increase research participation among older adults in rural areas. Researchers can improve accessibility for older adult populations by adopting inclusive communication methods and implementing safeguards against fraudulent activities. Additionally, establishing community trust through grassroots involvement is essential. Future research can use these insights to increase the engagement of older adults in community-based research, thereby enriching scientific knowledge and the well-being of this demographic.

Acknowledgments

This work was funded by Nova Scotia Health (grant number NSH 1026544) and the Department of Seniors and Long-term Care. The funders played no role in the design of the study, data collection or analysis, manuscript writing, or decision to publish findings. We want to acknowledge the 3 communities for their invaluable contributions to the project, including the older adults who participated in the project, assessors who conducted the assessment, local community coordinators who were our “boots on the ground,” and, finally, the municipalities and community groups that facilitated our community connections. Other provincial project partners to be acknowledged include the Canadian Association for Retired Persons – Nova Scotia chapter; Northwood Care – Research and Innovation; Building Research for Integrated Primary Care Nova Scotia; Dalhousie University; and funders Nova Scotia Health and the Department of Seniors and Long-term Care.

Data Availability

The datasets generated and analyzed during this study are not publicly available in order to maintain the confidentiality of participants and uphold the study’s agreement with funding partners.

Authors' Contributions

BB, GW, BW, and TLP wrote the manuscript, collected and analyzed the data. All authors reviewed the final manuscript. This manuscript was prepared by the authors, and no generative artificial intelligence was used in the writing, editing, or research conducted.

Conflicts of Interest

None declared.

References

1. Fiordelli M, Fadda M, Amati R, Albanese E. Older adults' motivations to participate or not in epidemiological research. Qualitative inquiry on a study into dementia in Switzerland. PLoS One 2021;16(2):e0247141 [FREE Full text] [doi: 10.1371/journal.pone.0247141] [Medline: 33577607]
2. Tang F, Pickard J. Aging in place or relocation: perceived awareness of community-based long-term care and services. J Hous Elderly 2008 Dec 18;22(4):404-422 [FREE Full text] [doi: 10.1080/02763890802458429]
3. Young Y, Kalamaras J, Kelly L, Hornick D, Yucel R. Is aging in place delaying nursing home admission? J Am Med Dir Assoc 2015 Oct 01;16(10):900.e1-900.e6 [FREE Full text] [doi: 10.1016/j.jamda.2015.07.017] [Medline: 26433865]
4. McCann J, Hebert L, Li Y, Wolinsky F, Gilley D, Aggarwal N, et al. The effect of adult day care services on time to nursing home placement in older adults with Alzheimer's disease. Gerontologist 2005 Dec;45(6):754-763 [FREE Full text] [doi: 10.1093/geront/45.6.754] [Medline: 16326657]
5. Kenny A, Farmer J, Dickson-Swift V, Hyett N. Community participation for rural health: a review of challenges. Health Expect 2015 Dec;18(6):1906-1917 [FREE Full text] [doi: 10.1111/hex.12314] [Medline: 25470231]
6. Pelletier C, Pousette A, Ward K, Fox G. Exploring the perspectives of community members as research partners in rural and remote areas. Res Involv Engagem 2020 Dec;6(1):3 [FREE Full text] [doi: 10.1186/s40900-020-0179-6] [Medline: 32015898]
7. Yoon A, Copeland A. Toward community‐inclusive data ecosystems: challenges and opportunities of open data for community‐based organizations. Assoc Inf Sci Technol 2020 Feb 22;71(12):1439-1454 [FREE Full text] [doi: 10.1002/asi.24346]

XSL•FO
RenderX

8. Bamford S, Franklin B, Hochlaf D, Holley-Moore G. Towards affordable healthcare: why effective innovation is key. International Longevity Centre UK. 2017 Jun 1. URL: https://ilcuk.org.uk/towards-affordable-healthcare-why-effective-innovation-is-key/ [accessed 2025-06-02]
9. Klein W, Parks C. Listening to seniors: successful approaches to data collection and program development. J Gerontol Soc Work 2007;48(3-4):457-473 [FREE Full text] [doi: 10.1300/j083v48n03_11] [Medline: 17210543]
10. Bingham A. From data management to actionable findings: a five-phase process of qualitative data analysis. Int J Qual Methods 2023 Aug 10;22:1-11 [FREE Full text] [doi: 10.1177/16094069231183620]
11. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101 [FREE Full text] [doi: 10.1191/1478088706qp063oa]
12. Postal codes. Canada Post. 2022 Jan 18. URL: https://www.canadapost-postescanada.ca/cpc/en/support/articles/addressing-guidelines/postal-codes.page [accessed 2025-06-02]
13. 2021 Census of population. Catalogue no. 98-316-X2021001. Statistics Canada. 2023. URL: https://www12.statcan.gc.ca/census-recensement/2021/dp-pd/prof/index.cfm?Lang=E [accessed 2025-06-02]
14. Wiles J, Leibing A, Guberman N, Reeve J, Allen R. The meaning of "aging in place" to older people. Gerontologist 2012 Jun;52(3):357-366 [FREE Full text] [doi: 10.1093/geront/gnr098] [Medline: 21983126]
15. Sörensen S, Hirsch J, Lyness J. Optimism and planning for future care needs among older adults. GeroPsych (Bern) 2014;27(1):5-22 [FREE Full text] [doi: 10.1024/1662-9647/a000099] [Medline: 26045699]
16. Bimou C, Harel M, Laubarie-Mouret C, Cardinaud N, Charenton-Blavignac M, Toumi N, et al. Patterns and predictive factors of loss of the independence trajectory among community-dwelling older adults. BMC Geriatr 2021 Feb 26;21(1):142 [FREE Full text] [doi: 10.1186/s12877-021-02063-7] [Medline: 33637045]
17. Teater B, Chonody J. What attributes of successful aging are important to older adults? The development of a multidimensional definition of successful aging. Soc Work Health Care 2020 Mar;59(3):161-179 [FREE Full text] [doi: 10.1080/00981389.2020.1731049] [Medline: 32100639]
18. Sörensen S, Mak W, Chapman B, Duberstein P, Lyness J. The relationship of preparation for future care to depression and anxiety in older primary care patients at 2-year follow-up. Am J Geriatr Psychiatry 2012 Oct;20(10):887-894 [FREE Full text] [doi: 10.1097/jgp.0b013e31822ccd8c]
19. Kang H, Kim H. Ageism and psychological well-being among older adults: a systematic review. Gerontol Geriatr Med 2022 Apr 11;8:23337214221087023 [FREE Full text] [doi: 10.1177/23337214221087023] [Medline: 35434202]
20. Greenfield EA, Marks NF. Formal volunteering as a protective factor for older adults' psychological well-being. J Gerontol B Psychol Sci Soc Sci 2004 Sep 01;59(5):S258-S264 [FREE Full text] [doi: 10.1093/geronb/59.5.s258] [Medline: 15358800]
21. Pani-Harreman K, Bours G, Zander I, Kempen G, van Duren JMA. Definitions, key themes and aspects of 'ageing in place': a scoping review. Ageing Soc 2020 Mar 18;41(9):2026-2059 [FREE Full text] [doi: 10.1017/s0144686x20000094]
22. Reich A, Claunch K, Verdeja M, Dungan M, Anderson S, Clayton C, et al. What does "successful aging" mean to you? Systematic review and cross-cultural comparison of lay perspectives of older adults in 13 countries, 2010–2020. J Cross Cult Gerontol 2020 Dec;35(4):455-478 [FREE Full text] [doi: 10.1007/s10823-020-09416-6] [Medline: 33064233]
23. Zhang Z, Zhang J. Social participation and subjective well-being among retirees. Soc Indic Res 2014 Sep 5;123(1):143-160 [FREE Full text] [doi: 10.1007/s11205-014-0728-1]
24. Musick M, Wilson J. Volunteering and depression: the role of psychological and social resources in different age groups. Soc Sci Med 2003 Jan;56(2):259-269 [FREE Full text] [doi: 10.1016/s0277-9536(02)00025-4] [Medline: 12473312]
25. Walsh K, O'Shea E, Scharf T, Shucksmith M. Exploring the impact of informal practices on social exclusion and age - friendliness for older people in rural communities. Community Appl Soc Psychol 2014 Jan 09;24(1):37-49 [FREE Full text] [doi: 10.1002/casp.2176]
26. Davis S, Silvester A, Barnett D, Farndon L, Ismail M. Hearing the voices of older adult patients: processes and findings to inform health services research. Res Involv Engagem 2019;5:11 [FREE Full text] [doi: 10.1186/s40900-019-0143-5] [Medline: 30834143]
27. Walker A. Why involve older people in research? Age Ageing 2007 Sep 23;36(5):481-483 [FREE Full text] [doi: 10.1093/ageing/afm100] [Medline: 17913755]
28. Finkelstein R, Wu Y, Brennan-Ing M. Older adults' experiences with using information and communication technology and tech support services in New York City: findings and recommendations for post-pandemic digital pedagogy for older adults. Front Psychol 2023 Apr 17;14:1129512 [FREE Full text] [doi: 10.3389/fpsyg.2023.1129512] [Medline: 37138998]
29. Kärnä E, Aavikko L, Rohner R, Gallistl V, Pihlainen K, Müller C, et al. A multilevel model of older adults' appropriation of ICT and acquisition of digital literacy. Int J Environ Res Public Health 2022 Nov 25;19(23):15714 [FREE Full text] [doi: 10.3390/ijerph192315714] [Medline: 36497789]
30. Shang Y, Wu Z, Du X, Jiang Y, Ma B, Chi M. The psychology of the internet fraud victimization of older adults: a systematic review. Front Psychol 2022 Sep 5;13:912242 [FREE Full text] [doi: 10.3389/fpsyg.2022.912242] [Medline: 36132192]
31. Larrison C, Sinha G, Andres M, Velez-Ortiz D. Financial exploitation of older adults in the United States: a pre-COVID-19 systematic scoping review. PsyArXiv Preprint posted online 2022 [FREE Full text] [doi: 10.31234/osf.io/gjfwh]
32. Lister J, Joudrey P. Rural mistrust of public health interventions in the United States: a call for taking the long view to improve adoption. J Rural Health 2023 Jan;39(1):18-20 [FREE Full text] [doi: 10.1111/jrh.12684] [Medline: 35691930]

XSL•FO
RenderX

33. Turin T, Chowdhury N, Haque S, Rumana N, Rahman N, Lasker M. Meaningful and deep community engagement efforts for pragmatic research and beyond: engaging with an immigrant/racialised community on equitable access to care. BMJ Glob Health 2021 Aug;6(8):e006370 [FREE Full text] [doi: 10.1136/bmjgh-2021-006370] [Medline: 34426405]
34. Han H, Xu A, Mendez K, Okoye S, Cudjoe J, Bahouth M, et al. Exploring community engaged research experiences and preferences: a multi-level qualitative investigation. Res Involv Engagem 2021 Mar 30;7(1):19 [FREE Full text] [doi: 10.1186/s40900-021-00261-6] [Medline: 33785074]

Abbreviations

ACT: Age Care Technologies Assess & Connect

Edited by A Mavragani; submitted 19.03.25; peer-reviewed by AJ Polsinelli, TH Dao-Tran; comments to author 22.04.25; revised version received 08.05.25; accepted 27.05.25; published 03.06.25.

Please cite as:
Boutilier B, Warner G, Wolfe B, Askari S, Moody E, Ghanouni P, Packer T
Engaging Community-Dwelling Older Adults in Research: Qualitative Substudy of Factors Impacting Participation
JMIR Form Res 2025;9:e74191
URL: https://formative.jmir.org/2025/1/e74191
doi:10.2196/74191
PMID:

©Bryah Boutilier, Grace Warner, Brianna Wolfe, Sorayya Askari, Elaine Moody, Parisa Ghanouni, Tanya Packer. Originally published in JMIR Formative Research (https://formative.jmir.org), 03.06.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Detecting Conversation Topics in Recruitment Calls of African American Participants to the All of Us Research Program Using Machine Learning: Model Development and Validation Study

Priscilla Pemu[1,2*], MD; Michael Prude[2*], BS; Atuarra McCaslin[2*], BA; Elizabeth Ojemakinde[2*], MD; Christopher Awad[3], MD; Kelechi Igwe[2], MPH; Anny Rodriguez[4], MPH, PhD; Jasmine Foriest[5], MS, MPH; Muhammed Idris[1,2,4*], PhD

[1]Department of Medicine, Morehouse School of Medicine, Atlanta, GA, United States
[2]Clinical Research Center, Morehouse School of Medicine, Atlanta, GA, United States
[3]Wexner Medical Center, The Ohio State University, Colombus, OH, United States
[4]Department of Public Health Education, Morehouse School of Medicine, Atlanta, GA, United States
[5]Human-Centered Computing, School of Interactive Computing, Georgia Institute of Technology, Atlanta, GA, United States
[*]these authors contributed equally

Corresponding Author:
Priscilla Pemu, MD
Department of Medicine
Morehouse School of Medicine
720 Westview Dr SW
Atlanta, GA
United States
Phone: 1 (404) 752 1500
Email: priscilla.pemu@gmail.com

Abstract

Background: Advancements in science and technology can exacerbate health disparities, particularly when there is a lack of diversity in clinical research, which limits the benefits of innovations for underrepresented communities. Programs like the All of Us Research Program (AoURP) are actively working to address this issue by ensuring that underrepresented populations are represented in biomedical research, promoting equitable participation, and advancing health outcomes for all. African American communities have been particularly underrepresented in clinical research, often due to historical instances of research misconduct, such as the Tuskegee Syphilis Study, which have deeply impacted trust and willingness to participate in research studies. With the US population becoming increasingly diverse, it is crucial that clinical research studies reflect this diversity to improve health outcomes. However, limited data and small sample sizes in qualitative studies on the inclusion of underrepresented groups hinder progress in this area.

Objective: The goal of this paper is to analyze recruitment conversations between research assistants (RAs) and potential participants in the AoURP to identify key topics that influence enrollment. By examining these interactions, we aim to provide insights that can improve engagement strategies and recruitment practices for underrepresented groups in biomedical research.

Methods: Our study design was an observational, retrospective approach using machine learning for content analysis. Specifically, we used structural topic modeling to identify and compare latent topics of conversation in recruitment calls by Morehouse School of Medicine RAs between February 2021 and April 2022 by estimating expected topic proportions in the corpus as a function of enrollment and participation in AoURP.

Results: In total, our model estimated 45 topics of which 12 coherent topics were identified. Notable topics, that were more likely to occur in conversations between RAs and participants that enrolled and participated, include closing or following up to schedule an appointment, COVID-19 protocols for in-person visits, explaining precision medicine and the need for representation, and working through objections, including concerns about costs, insurance, care changes, and health fears. Topics among potential participants who did not enroll include technical challenges and describing physical measurement visits (eg, collection of basic physical data, such as height, weight, and blood pressure).

XSL•FO
RenderX

Conclusions: Using an approach that leverages machine learning to identify topical structure and themes with limited human subjectivity is a promising strategy to identify gaps in, and opportunities to improve, the recruitment of underserved communities into clinical trials.

(JMIR Form Res 2025;9:e65320) doi:10.2196/65320

KEYWORDS

recruitment; clinical trials; diversity; trust; precision medicine

Introduction

Background and Significance

Advancements in science and technology sometimes worsen the gaps in health outcomes [1-3]. Lack of diversity in clinical research contributes to this by negatively influencing the diffusion of such innovations in communities that are not well represented. The projected demographic shift in the US population underlines the urgency for clinical research studies to reflect the diversity of the affected population [4].

The All of Us Research Program (AoURP) is a national initiative aimed at advancing precision medicine by building a diverse and inclusive research database [5]. A core focus of the program is increasing engagement, enrollment, and retention of underrepresented in biomedical research (UBR) populations, including African American, Hispanic or Latino, and rural communities. African American communities, for example, have faced a long history of exclusion from clinical and biomedical research, with systemic barriers such as limited access to health care, underrepresentation in recruitment efforts, and mistrust stemming from unethical practices like the Tuskegee Syphilis Study [6]. The program uses targeted outreach strategies, such as community partnerships, culturally tailored communication, and participant-centered approaches, to address historical barriers to research participation and foster trust. These efforts are designed to ensure the cohort reflects the diversity of the US population, enabling research that benefits all communities equitably.

A 2022 report from the National Academies of Medicine and Science, “Building Research Equity for Women and Underrepresented Groups” [7] recommended a holistic approach with multiple levels of intervention to reach the goal of inclusiveness in clinical research. One significant level is the role of interaction between the research team and the community with a focus on engagement [7]. These interactions are pivotal, as they directly influence potential participants’ perceptions of the study, trust in the research process, and willingness to participate [7]. The 2022 National Academies of Medicine and Science report highlighted the dearth of critical data about facilitators of successful inclusion of underrepresented groups in clinical trials [7-9]. Research shows that information exchange between members of the clinical research team and potential study participants or volunteers critically influences the research volunteer’s journey from awareness to enrollment [7,10,11]. Current recruitment strategies rely on a combination of outbound and inbound phone calls, community events, and digital outreach, but these interactions are typically not recorded, making analysis of recruitment effectiveness reliant on retrospective assessments rather than direct examination of conversation topics in real-time. Moreover, conclusions drawn from qualitative analyses of small sample sizes and limited examination of offers for clinical trial participation beyond financial and nonfinancial incentives hinder the field [12,13].

Objective

While many studies have identified barriers to clinical trial enrollment, little is known about what actually happens at the point of recruitment. The goal of this analysis is to contribute to this body of research by leveraging automated content analysis methods to examine conversations during recruitment calls with limited human subjectivity. Specifically, we use structural topic modeling (STM) to explore differences in conversations between research assistants (RAs) and community members from the greater Atlanta metro area recruited for participation in the AoURP during the COVID-19 pandemic.

Methods

Study Design

Potential participants in the AoURP contribute their biological, health, behavioral, and environmental data. The data collected in the AoURP includes participant-provided information (PPI) from surveys, available electronic health records (EHRs) with participant permission, baseline physical measurements (eg, height and weight), and biospecimens (eg, blood, urine, and saliva). DNA is isolated from the biospecimen samples for genetic testing. Additionally, the program may also collect data from passive mobile and digital health devices, sensors, and mobile apps.

Selection of Participants

Inclusion Criteria

Eligible AoURP participants must be 18 years or older, currently living in the United States or US territory, and have the legal authority and ability to give informed consent.

Exclusion Criteria

Adults who do not have the ability to give informed consent, children younger than 18 years of age, and individuals incarcerated at the time of enrollment are ineligible to participate in AoURP.

All of Us Southeastern Enrollment Center

Individuals were enrolled through participating health provider organizations or designated venues. The Southeast Enrollment Center (SEEC) is a health provider organization partner of the AoURP consisting of the University of Miami Miller School of Medicine, University of Florida College of Medicine, Emory University, and Morehouse School of Medicine (MSM).

XSL•FO
RenderX

Strategically located in a region of the United States with high racial and ethnic diversity, the center is enrolling over 100,000 diverse participants, or approximately 10% of the >1 million individuals targeted by the AoURP nationwide. As of December 2022, the SEEC has successfully enrolled a cohort with 76.23% identifying as UBR populations, including 345 (out of 449) participants who identified as Black or African American [14]. This is a significant contribution given that African Americans have historically been underrepresented in clinical research; for example, in oncology trials, African American men and women make less than 5% and 3% of participants, respectively [15].

The Setting of Recruitment and Procedures

Recruitment efforts used multiple strategies to reach potential participants. The primary method of recruitment was phone calls, supplemented by web-based or mobile outreach, and where feasible, in-person site support. These approaches help maximize participant engagement and accessibility. Before enrollment, research staff screen and contact participants to ensure they are capable of providing informed consent. During this process, individuals receive clear information about the study and are explicitly informed that their decision to participate does not impact their health care or medical treatment.

All participants complete a standardized electronic informed consent process to ensure uniformity and adherence to ethical guidelines. To accommodate diverse participants, the informed consent process is self-paced, allowing individuals to review materials at their convenience through web-based or mobile platforms. Consent materials are available in English and Spanish, with institutional review board–approved translations to ensure accuracy. For individuals who require additional support, trained staff are available to assist either in person at designated sites or remotely via the All of Us Support Center.

RA Recruitment

A total of 13 research staff were enlisted and trained to support recruitment efforts during the study period of February 2021 to April 2022. This included seven RAs in their 20s, three RAs in their 30s, and three RAs in their 50s. A total of 12 (92%) RAs were female and one male. Most RAs (12 out of 13) were Black or African American and one Hispanic. One RA completed an associate degree, five RAs a bachelor's, three RAs an MPH, and three RAs MDs.

Recruitment Calls

While initially conducted using a combination of face-to-face and via phone calls, the COVID-19 pandemic required us to digitize our recruitment efforts fully. Phone calls were facilitated using a cloud contact center solution called 3CLogic. Transcription was enabled by integration with a commercial speech analytics platform, ObserveAI. Contacts and associated phone numbers in 3Clogic are sourced from a proprietary platform called Engage. RAs are trained separately on Engage and 3Clogic. Call result outcomes were logged as Not Available (if the person you are calling is not available, no one answers the phone, no voicemail, or voicemail full), Wrong Number (if you can verify the number is incorrect), Voicemail (if you left a voicemail message), Scheduled (you have scheduled the person for a visit), Not Interested (potential participant does not want to participate in the AoURP), Do Not Call (potential participant does not want to receive any more calls about the AoURP), and Follow Up (potential participant requests a call at a later date).

Data Analysis

Levels of Participation

There are 4 levels of participation in the AoURP. Interested individuals are potential participants who have provided contact information to receive updates about the program. Registered individuals are participants who have created an account on the AoURP platform but have not yet completed the informed consent process. A registered individual has a unique AoURP participant identifier. Consented individuals have met the eligibility and inclusion criteria and have completed the primary informed consent process. They may also have completed the optional Health Insurance Portability and Accountability Act authorization or EHR consent and genomic return of results (gRoR) consent. Participants are individuals who have completed the primary informed consent process and are eligible to complete the Health Insurance Portability and Accountability Act authorization or EHR consent, gRoR consent, surveys specific to Basics, Lifestyle, and Overall Health PPI modules, and provide physical measurements and biospecimens. Core participants have completed all core components (ie, registration, EHR or gRoR consent, and PPI modules) and provided physical measurements and biospecimens. They may also contribute other data, such as digital health data from sensors and wearables, if available. For the purposes of analysis, any participant who has created an account, completed informed consent, and completed at least one survey or in-person visit, is considered to have "participated at any level."

STM

Following an approach outlined in Idris et al [16], we used STM to analyze the content of potential research volunteer and RA conversations. STM is an advanced machine learning technique designed to uncover themes within a collection of documents or transcripts. It operates on the principle that certain words are linked to specific topics, as demonstrated in previous studies [17,18]. Formally, each utterance in a conversation is represented as a vocabulary multiset with corresponding word counts, also known as a document-term matrix. In this matrix, each word has a probability associated with belonging to a particular topic. By viewing a conversation as a random mixture of topics—each defined by a set of characteristics, high-probability words—we can infer latent topics that may not be immediately apparent.

Using STM allows us to incorporate document-level metadata, such as the context of the conversation, the participants involved, and the timing of interactions. This enriches the model, making it more nuanced and capable of capturing the complexity of real-world conversations. Furthermore, STM's ability to handle large and diverse datasets makes it particularly suitable for analyzing the varied and dynamic nature of volunteer-RA dialogues.

Our goal in using STM is to uncover underlying patterns and themes that provide valuable insights into the nature of these interactions. Understanding these latent topics can enhance

XSL•FO
RenderX

communication strategies, improve volunteer engagement, and identify areas for training and development for RAs. Overall, this approach not only improves the interpretability and relevance of the generated topics but also aligns the findings with the underlying structure of the data, thereby facilitating more informed and actionable conclusions. We used STM in our analysis for three main reasons. First, STM allows the incorporation of document metadata such as participant characteristics into the analysis of latent topics through a generative process that models the relationship between metadata and the proportion of a specific topic in each utterance [19]. This enables us to plot metadata or topic relationships by estimating parameters, including the expected proportion of an utterance belonging to a topic as a function of a covariate, or a first difference estimate, where topic prevalence for a particular topic is contrasted between 2 groups (eg, enrolled or participated vs not enrolled). Second, by incorporating additional contextual information, STM enhances the interpretability and relevance of the generated topics. This results in more coherent and meaningful topics that align with the underlying data structure, thereby facilitating the drawing of insightful conclusions. Finally, using built-in functionality in the STM R package, we avoided subjective decisions on the number of topics by using a data-driven approach [20]. This approach helped eliminate human-based bias in our analysis by identifying words that appear in a document only if the document pertains to a specific topic.

Ethical Considerations

Participants provided informed consent during the primary data collection process, which included authorization for the use of their data or samples in future research studies, as approved by the institutional review board (protocol #2016-05). All data used in this study were deidentified to ensure the confidentiality and privacy of participants. Participants in the AoURP receive compensation as an acknowledgment of their time and effort. Participants received US $25-$50 for completing initial enrollment procedures and additional compensation for follow-up activities, including such as biospecimen collection, physical measurements, or surveys.

Results

Overview

Table 1 provides the demographics of participants recruited to the AoURP by MSM and compared to other institutions in the Southeastern enrollment center. Figure 1 provides a flowchart of how many participants were contacted, reached, and enrolled between February 2021 and April 2022. In total, RAs made 31,741 calls comprising 4350 conversations with an average of 33.92 (SD 54.66) completed talk turns (eg, a single spoken statement by one speaker before the other speaker responds). The average number of calls per participant was 2.90 for the enrolled group and 1.48 for the not enrolled group. The average length of conversation was 64.28 talk turns for the enrolled group and 25.23 for the not enrolled group, suggesting that on average, longer conversations between RAs and potential participants are associated with a higher likelihood of enrollment.

Table 1. Underrepresented in biomedical research of core participants recruited by All of Us southeastern enrollment center (2018-2022; N=29,402).

	Morehouse School of Medicine, %	Southeastern Enrollment Center, %
Overall	97.26	76.23
Ancestry	92.54	66.47
Age	14.02	23.05
Gender	0.03	0.04
Sexual and gender minority	13.54	10.10
Income	53.24	25.31
Education	18.31	9.60
Geography	0.54	3.52

XSL•FO
RenderX

Figure 1. Flowchart of research participants contacted, reached, and enrolled. AoU: All of Us; CRC: Clinical Research Center; MSM: Morehouse School of Medicine.

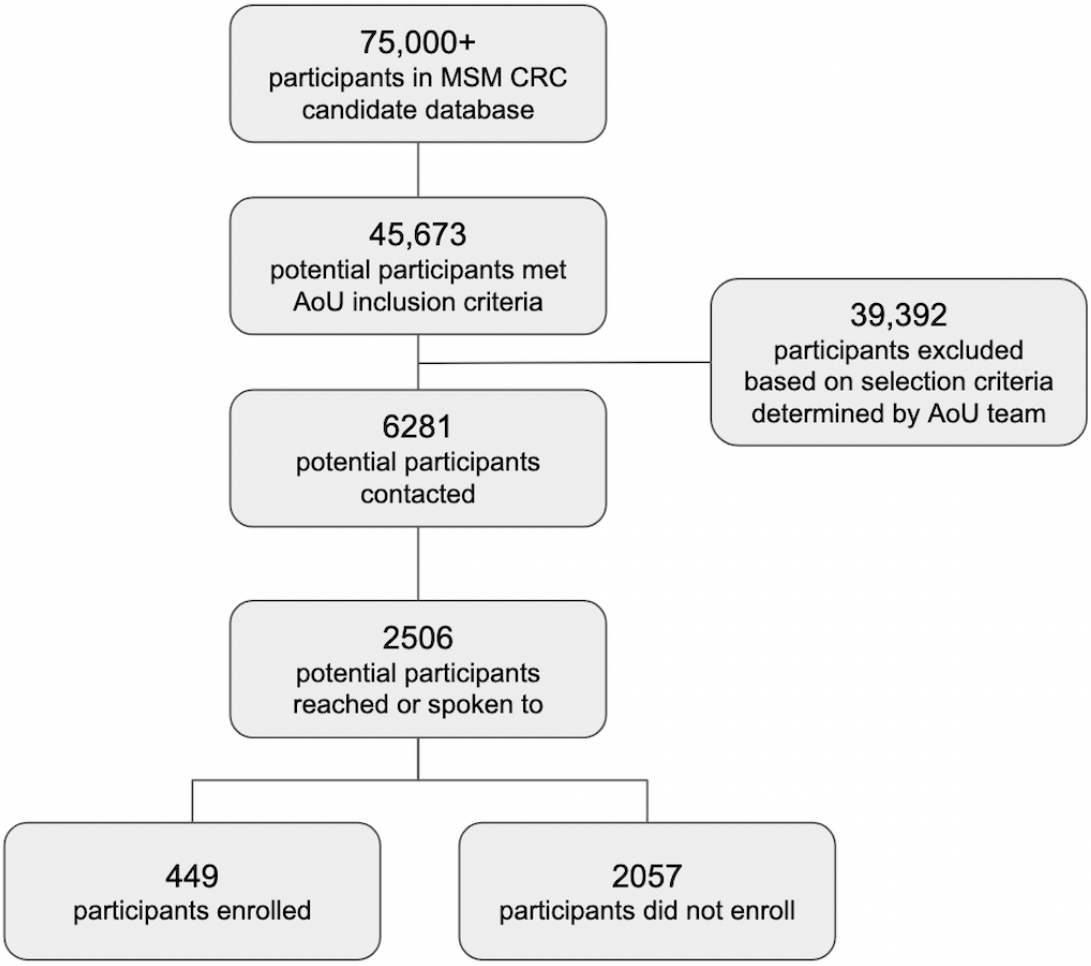

Structural Topic Model

In total, our corpus included 106,077 talk turns, 4911 terms (eg, a unique word), and 383,763 tokens (eg, every occurrence of a word, including repetitions across different sentences and documents). A total of 45 topics were estimated. In line with common practice, our approach to assigning labels to topics began by reviewing topic model outputs to identify potential topics based on the coherence of characteristic word forms [21]. We then reviewed the high-scoring conversations for each topic to determine topic labels. High-scoring conversations refer to talk turns with the strongest associations to a given topic, as determined by the topic model's relevance and coherence scores. Where there was a clear topical theme in the first N talk turns, we validated the topic label. In general, a clear theme emerged within the first 25 talk turns. A review of high-scoring talk turns was completed collaboratively by 2 reviewers (PP and MI) until a consensus on topic labels was reached. Of the 45 topics estimated, 12 coherent topics emerged as presented in Textbox 1. Figure 2 plots these initial topics ordered by their expected topic proportions, that is, the expected proportion of the corpus that belongs to each topic, along with the top 5 words associated with each topic. Topics 2 and 22, which clearly reflect automated voicemail prompts have the highest prevalence in the transcripts, which is as expected given the practice of cold calling where most calls go unanswered.

XSL•FO
RenderX

Textbox 1. Coherent topic identified from the Corpus of All of Us recruitment transcripts with top five associated words.

Topic 2

- Automated message prompt

Topic 5

- Following up to complete participant surveys

Topic 6

- Instructions or struggles accessing the participant portal

Topic 7

- Scripted introduction to All of Us Research Program

Topic 10

- Closing or following up to schedule appointment

Topic 12

- COVID-19 protocols for in-person visit

Topic 22

- Automated voicemail prompt

Topic 24

- Scripted voicemail message

Topic 27

- Update or retention calls

Topic 29

- Explaining precision medicine or need for representation

Topic 33

- Describing physical measurements visit procedure

Topic 41

- Working through objections or barriers to participation

XSL•FO
RenderX

Figure 2. Initial topics ranked by expected proportions with the top 5 associated words.

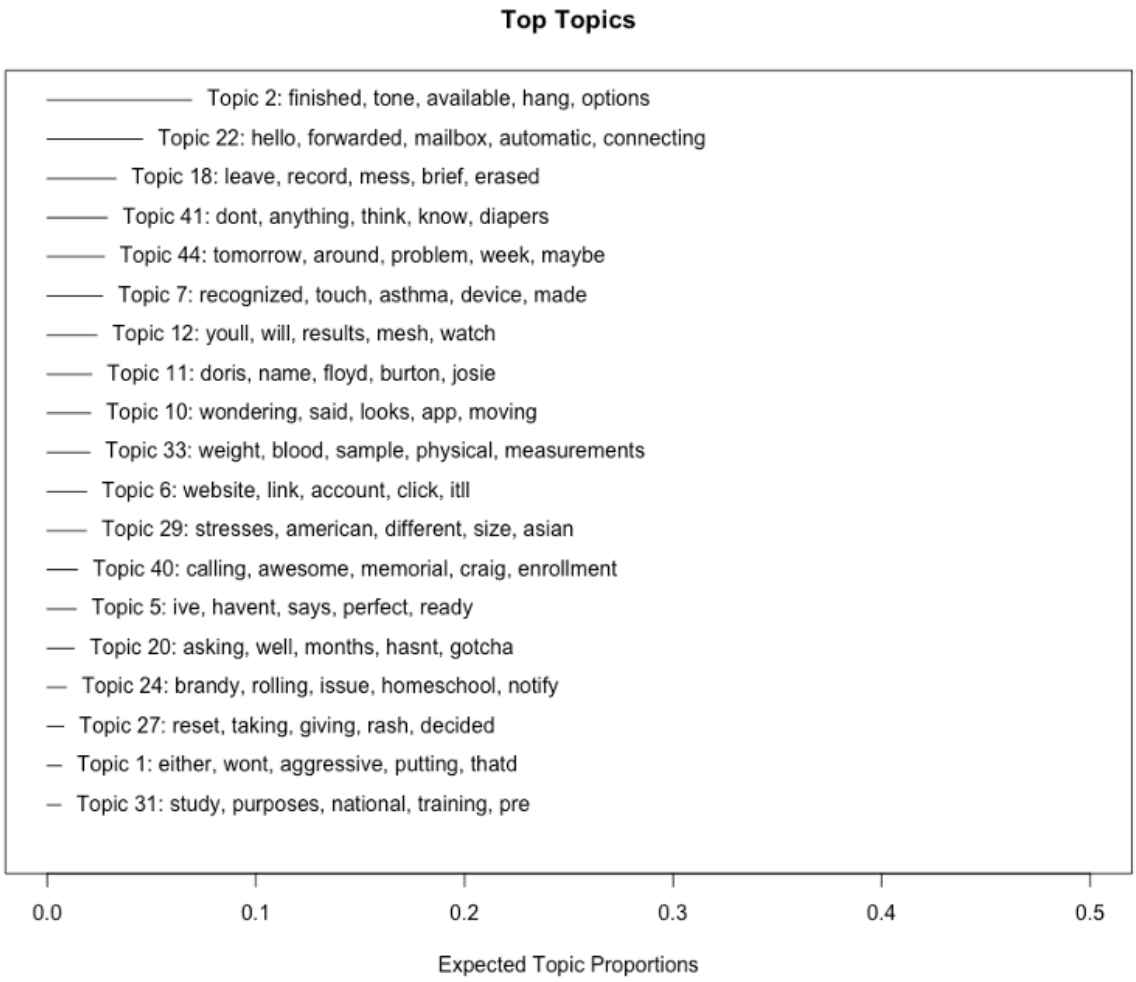

Topic Prevalence as a Function of Enrollment or Participation

Overview

Figure 3 plots the change in topic proportion when comparing conversations where the research participant enrolled in the research study to those who did not. That is, what is the likelihood of a topic being brought up in a conversation between RAs and participants as a function of the recruitment outcome? The plots also include confidence intervals indicating whether mean differences in prevalence are statistically significant.

XSL•FO
RenderX

Figure 3. Topic proportions as a function of enrollment into the All of Us Research Program.

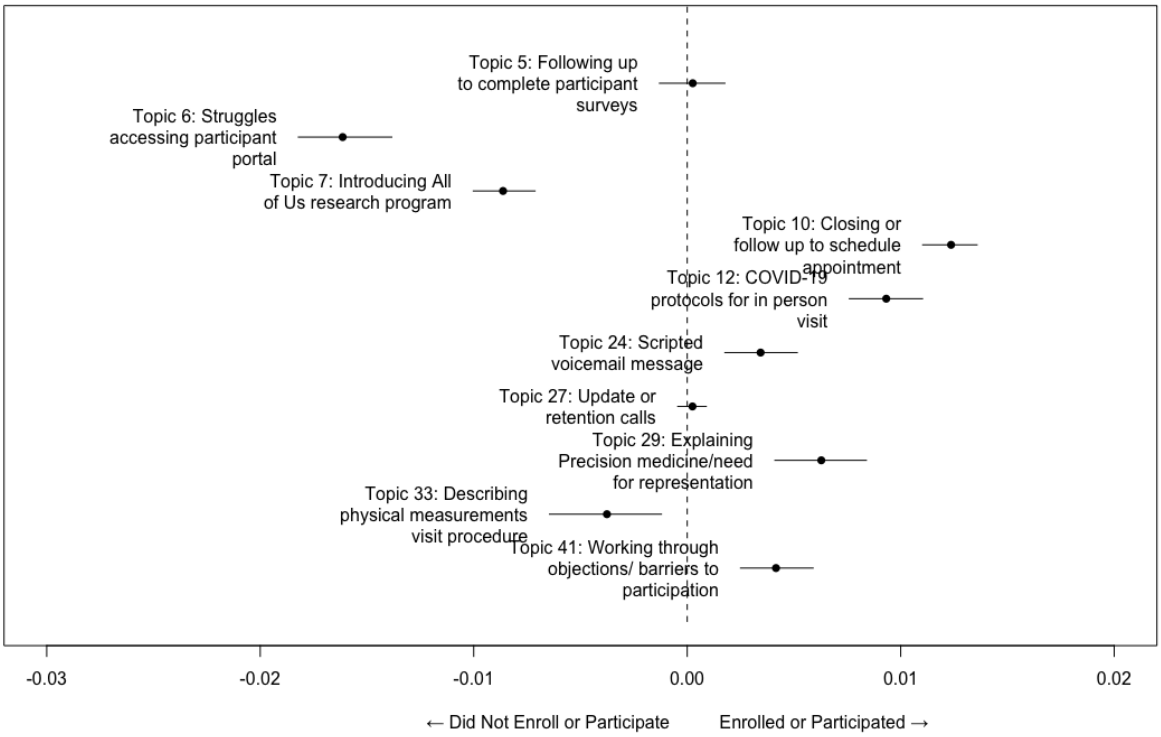

Enrolled or Participated Group

Our results, shown in Textbox 2, suggest that recruitment calls in which RAs follow-up to close participants and schedule appointments (Topic 10), as well as describing COVID-19 protocols for in-person visits (Topic 12), are more likely to appear in the enrolled group. Illustrative talk turns include, “I’m calling regarding your enrollment into the All of Us research program,” “So we started the process and I was wondering if you’d like to go ahead and schedule an appointment to come in for the other part of the enrollment process,” “once you come in security will alert that you are here, we come to you do a rapid test that takes 10 to 15 minutes. Once its negative they will bring you into the building,” and “everybody has to take a COVID test before you come into samples.” These results are as expected, given that these topics of conversation are more likely to occur in interactions in which participants have indicated some level of interest in participation.

We also found that scripted voicemails left by RAs (Topic 24) are also more likely to be associated with successful enrollment. The template used was, “Hi this is [research assistant’s name] and I’m calling from Morehouse School of Medicine and today we’re calling to let people know about the All of Us research study. It’s a nationwide study that has a goal of recruiting one million participants to push forward the initiative of precision medicine, that is the type of medicine that uses information about a person’s lifestyle, their environment, and biology to figure out new ways to treat, cure and prevent different types of diseases. If this sounds like something that you are interested in, please feel free to give us a callback.” This result suggests that RAs who left voicemails for participants who did not answer initial calls were more likely to get callbacks and ultimately, successful enrollment.

Another topic that was more likely to occur in recruitment calls that concluded in successful enrollment was the explanation of the purpose of the study, precision medicine, and the need for representation (Topic 29) to prevent and address health disparities. In one high-scoring utterance on this topic, the RA states in response to what the purpose of the study is: “So precision medicine is trying to develop health care treatments on a more individual basis instead of the sort one size fits all mentality,” in another utterance the RA states: “Our goal as a program is to get more participants from across the country with different background, lifestyle, and genetics because you so know so many things can affect health.” Upon further review, it appears characteristic utterances on this topic follow a template that begins with a layman’s definition of precision medicine followed by a description of the goal of the AoURP. The increased likelihood of detecting this topic in successful recruitment calls may indicate that describing the purpose of the study is an important factor in potential participants’ enrollment decision-making process.

Finally, and most insightful, the model identified a topic that reflected objections and barriers to participation (Topic 41). This topic had the highest expected proportion in the entire corpus, indicating that it was the most common topic or theme across the entire corpus (after Topics 2 and 22, ie, automated voicemail prompts). In one exchange, a participant inquired about participation costs, “How much is all of these?” In response, the RA provided reassurance that “You don’t have to pay anything, we don’t look at your insurance or have to change doctors.” In another exchange, a participant tells the RA they are “skeptical” and “scared” because they “already got the COVID-19 shots and I’ve been having some bad headaches constantly after that.” The RA responds, “Well I assure you that we have nothing in our needle, we are just taking a small blood sample and we’re not putting anything in your bloodstream.”

XSL•FO
RenderX

We found other examples of participants being cautious about "not knowing" what will be done with their samples once collected and assurances from RAs that they would not be given any treatments. Other challenges brought up included transportation, for example, "I can't participate because I don't have transportation," suspicion toward RAs, for example, "How did you get my name and number?" and mistrust based on past negative experiences with discrimination related to health, for example, "I was one of the people that was exposed. They refused to tell me what types of toxic chemicals or materials they were working with." In one final illustrative exchange, a potential participant asks about the results of his blood sample, "You took some blood and you analyze that, can I have the results?" The RA responds that the "Sample [is] shipped to a secure facility by mail and that the facility would be doing" genetic analysis.

Textbox 2. Illustrative talk turns for topics associated with increased likelihood of enrollment.

Topic 10: Closing or following up to schedule appointment

Research assistant (RA): "I'm calling regarding your enrollment into the All of Us research program," "so we started the process and I was wondering if you'd like to go ahead and schedule an appointment to come in for the other part of the enrollment process."

Topic 12: COVID protocols for in-person visit

RA: "Once you come in security will alert that you are here, we come to you do a rapid test that takes ten to fifteen minutes. Once its negative they will bring you into the building," and "everybody has to take a COVID test before you come into samples."

Topic 24: Scripted voicemail message

RA: "Hi this is [research assistant's name] and I'm calling from Morehouse School of Medicine and today we're calling to let people know about the All of Us research study. It's a nationwide study that has a goal of recruiting one million participants to push forward the initiative of precision medicine, that is the type of medicine that uses information about a person's lifestyle, their environment, and biology to figure out new ways to treat, cure, and prevent different types of diseases. If this sounds like something that you are interested in, please feel free to give us a callback."

Topic 29: Explaining precision medicine or need for representation

RA: "So precision medicine is trying to develop health care treatments on a more individual basis instead of the sort one size fits all mentality"

RA: "Our goal as a program is to get more participants from across the country with different background, lifestyle, and genetics because you so know so many things can affect health."

Topic 41: Working through objections or barriers to participation

Participant: "How much is all of these?" RA: "You don't have to pay anything, we don't look at your insurance or have to change doctors."

Participant: "Already got the COVID shots and I've been having some bad headaches constantly after that." RA: "Well I assure you that we have nothing in our needle, we are just taking a small blood sample and we're not putting anything in your bloodstream."

Did Not Enroll Group

Illustrative talk turns of high-ranking topics among those who did not enroll are displayed in Table 2. The first topic that is more likely to be prevalent in the did not enroll group is instructions to accessing or challenges accessing the participant portal (Topic 6). Illustrative examples of talk turns in Topic 6 include instances of RAs "sending emails [with] instructions on how to register" for the research study, as well as "email[s] with the link to the [All of Us] website for more information and also a link to start an account if [the participant is] ready for that." We also found many examples of RAs walking participants through challenges accessing their portals once created. In one exchange, an RA helps the participant search for the correct username when trying to access the portal via their phone: "Let's try your phone number as your username, then we click on forgot password, a link should be sent to your phone." Utterances like "if you do not remember your password, click forgotten password. Let's see what happens," were very common in this topic.

Our results also suggested that scripted introductions to the AoURP (Topic 7) were also more likely to occur in the not enrolled group. There is an overlap between this topic and Topic 24. Both include high-scoring scripted utterances like, "Hi yes like I said this is [Name]. I'm calling from Morehouse School of Medicine today we're reaching out the members of the community to talk about the "all of us research program" and describe the purpose of the "research study called the all research program to speed up medical breakthroughs."

Finally, we found that the descriptions of the physical measurements visit (Topic 33) were also more likely to occur in the did not enroll group. In one exchange, an RA explains that by participating, "You would complete three maybe four surveys ... then we would bring you in for the second part which we would take measurements of your height, weight, blood pressure, and you would provide a small blood sample." The participant responded that they feel comfortable about completing surveys about "My demographic, but I don't know if I wanna come in and give a blood sample and things of that nature. I also would like to say that I do understand the importance of these things [but] I'm not too comfortable with that." In other exchanges, "atrocities that have happened" were alluded to as were "caregiver" responsibilities that would make the participant unable to "contribute" to the "second half ... where we would take your measurements, your height, your weight, your blood pressure, and you would provide small urine and blood samples," even in light of "cash compensation for participation." Conversations on this topic suggest that participants who did not enroll were more likely to feel a burden of participation related to the second part of the study.

XSL•FO
RenderX

Table 2. Illustrative talk turns for topics associated with a decreased likelihood of enrollment.

Topic	Illustrative talk turns
Topic 6: Instructions or struggles accessing participant portal	• RA[a]: "sending emails [with] instructions on how to register" for the research study as well as "email[s] with the link to the [All of Us] website for more information and also a link to start an account if [the participant is] ready for that." • RA: "Let's try your phone number as your username, then we click on forgot password, a link should be sent to your phone."
Topic 7: Scripted introduction to All of Us Research Program Topic 24: Scripted voicemail message	• RA: "Hi yes like I said this is [Name]. I'm calling from Morehouse School of Medicine today we're reaching out the members of the community to talk about the "all of us research program" and describe the purpose of the "research study called the all research program to speed up medical breakthroughs."
Topic 33: Describing physical measurements visit procedure	• RA: "You would complete three maybe four surveys ... then we would bring you in for the second part which we would take measurements of your height, weight, blood pressure, and you would provide a small blood sample." • Participant: "My demographic, but I don't know if I wanna come in and give a blood sample and things of that nature. I also would like to say that I do understand the importance of these things [but] I'm not too comfortable with that."

[a]RA: research assistant.

Both Groups

Two topics were equally likely to show up in both the enrolled and not enrolled groups. The first topic is RAs following up with participants to complete surveys (Topic 5). These surveys included initial "demographic" surveys, as well as the "second part of the study whereby we schedule you to come in and we do your physical measurements, blood, urine sample so ... I was wondering if you're ready for that part however you would need to complete your survey online." In these exchanges, RAs were usually following up with participants about "completing [their] surveys" before "coming in for the appointment." Finally, we identified a topic around follow-up calls, including updates and retention calls. We identified three scenarios where RAs were calling. In the first situation, RAs were following up with participants who indicated an interest in enrolling but did not complete consent forms indicating some level of interest in participation but not following through on any level of participation (eg, demographic surveys, physical measurements, and visits). In the second situation, research participants consented but did not complete surveys, the first level of participation. These two first scenarios reflect RAs following up with participants requesting updates to encourage further engagement with the AoURP. In the final situation, participants completed surveys and physical measurement visits and were followed up with to complete a retention survey that was introduced as a required component in the later part of recruitment by the National Institutes of Health for All of Us enrollment centers.

Discussion

Principal Findings

Since its inception, MSM has developed and refined a multilevel approach to engagement with the community. This results in significant success in recruiting and enrolling members of the African American community in research. As a result of this success, MSM was selected to participate in the AoURP as part of a larger consortium, the SEEC in 2018. The COVID-19 pandemic accelerated the digital transformation of our clinical trial recruitment platform, which created an unprecedented opportunity to study recruitment using real-world data on the content of conversations between RAs and potential participants. The goal of this analysis is to report a preliminary investigation of these conversations.

In analyzing the metadata of these calls, we found that the average number of calls per participant (2.90 vs 1.48), as well as the average length of conversations (64.28 vs 25.23) for the enrolled group, were higher than that of the not enrolled group. These findings align with research in sales and marketing increased frequency and depth of contact with potential participants correlate with increased likelihood of "closing" [22,23]. Using automated content analysis methods, we identified topics that reflected challenges associated with participation in clinical trials by underrepresented community members.

These findings are in line with existing literature that highlights time and resource constraints, including financial burden, time commitment, transportation, and compensation related to trust as barriers to clinical trial participation [24,25]. Notably, our analysis did not highlight unique barriers or objections specific to African American communities, instead reflecting common themes identified that were consistent across many UBR communities. Beyond barriers, we also identified topics that reflect concerns around clinical trial participation, including mistrust of investigators and the clinical trial process, as well as a lack of awareness about the purpose and value of trials [25,26]. African American mistrust of the US health care system and medical research is based both on historical research misconduct, as well as ongoing systemic issues and misinformation [27]. Interestingly, we also found that where concerns were addressed and RAs worked with potential participants to work through objections, participation was much more likely.

Finally, we found that "scripted language" is common between the enrolled and not enrolled groups but the differences are

XSL•FO
RenderX

within the content of those scripts. The main difference seems to be that Topic 7 (scripted introduction to AoURP) is an introduction when someone picks up the phone while Topic 24 (scripted voicemail message) is someone leaving a message. There may be some selection bias here because those who listened to the voicemail and were interested in the study called back while those who did not find the study interesting did not. It is also likely that most folks who reached out did not find the study interesting, and therefore, did not enroll, which would push this topic to the did not enroll category.

This study is consistent with previous studies that consider topical framing and placement as key predictors of successful clinical trial recruitment. A few studies have focused on information exchange between members of the clinical research team and potential study volunteers is a critical influence on research volunteers' journey from awareness to enrollment. This includes Eggly et al [28], who used linguistic analysis to explore offers to participate in clinical trials. The authors identified differences in the quality of the interactions and quantity of the interactions between physicians and African American patients versus White patients. In another study of strategies to persuade participation in cancer clinical trials, Barton et al [12] conducted a discourse analysis of 22 oncology trial-related interactions stratified by race and argued that topical framing and placement influence momentum toward participation. However, these prior studies are limited by sample size and the extent they examined informational dimensions of offers to participate in clinical trials—such as descriptions of the procedures of the trial and discussions of the trial regimen in comparison with standard treatment.

The results of our STM approach to analyzing recruitment calls suggest that natural language processing is a promising method that reliably identifies latent topics in conversations with minimal human subjectivity. This not only permits the establishment and analysis of larger data corpora but also enables hypothesis testing through the incorporation of metadata into models. This enables relationship estimations between metadata and topical content and prevalence.

Limitations

The limitations of our analysis are 3-fold. First, we had limited information on the demographics of potential participants who did not enroll or participate. This limited our ability to include demographics including race and gender of these potential participants in our topic model estimation. Though this missingness likely did not influence the topics that were identified, more complete demographic data would have allowed us to produce unbiased estimates of associations between demographics and topical prevalence, that is, the likelihood of certain topics being brought up. Second, while a noninsignificant proportion of recruitment conversations are scripted, the tone and tenor of the recruiter matters for how information is conveyed and received. That is, not only what clinical RAs and coordinators say, but how they say it influences how information is received by potential participants. Our analysis only captures what is being said, that is, topics. Future research should consider methods for capturing sentiment. Additionally, given gender differences in communication, analyses of the gender of RAs on the content and trajectory of recruitment calls and subsequent enrollment rates could provide valuable insights into future studies. Finally, we found many errors during the transcription of audio to text by ObserveAI, which may indicate bias in proprietary natural language processing algorithms when being applied to audio generated from African American speakers [29]. More culturally congruent speech-to-text algorithms are needed.

Conclusions

The goal of this analysis is to analyze conversations between RAs and community members who were recruited to participate in the AoURP. This type of work supports the building of an evidence-based framework to guide recruitment conversations. The use of this method of analysis addresses the need for critical data to support the development of the science of inclusion, clinical research diversity, and community engagement.

Acknowledgments

We would like to acknowledge the entire All of Us team at Morehouse School of Medicine as part of the broader Southeaster Enrollment Center for their assistance in collecting data for this manuscript. This work was supported by the National Institutes of Health through the Office of the Director (OT2OD026551,1) and the National Institute for Minority Health Disparities (U24MD015970).

Data Availability

The datasets generated or analyzed during this study are not publicly available due to the sensitive nature of call transcripts but are available from the corresponding author upon reasonable request.

Authors' Contributions

PP and MI conceptualized the study. MI developed the methodology and conducted the formal analysis. MP curated the data. PP and MI wrote the original draft of the manuscript. MP, KCI, CA, AM, EO, AR, and JF contributed to the review and editing of the manuscript. MP and MI created the visualizations. PP supervised the project. KC, AM, and EO managed project administration. PP acquired the funding for the study. All authors read and approved the final version of the manuscript.

XSL•FO
RenderX

Conflicts of Interest

None declared.

References

1. National Academies of Sciences, Engineering, and Medicine, Health and Medicine Division, Board on Population Health and Public Health Practice, Committee on Community-Based Solutions to Promote Health Equity in the United States. In: Baciu A, Negussie Y, Geller A, Weinstein JN, editors. Communities in Action: Pathways to Health Equity. Washington (DC): National Academies Press; 2017.
2. Goodson N, Wicks P, Morgan J, Hashem L, Callinan S, Reites J. Opportunities and counterintuitive challenges for decentralized clinical trials to broaden participant inclusion. NPJ Digital Med 2022;5(1):58 [FREE Full text] [doi: 10.1038/s41746-022-00603-y] [Medline: 35513479]
3. Breen N, Berrigan D, Jackson JS, Wong DW, Wood FB, Denny JC, et al. Translational health disparities research in a data-rich world. Health Equity 2019;3(1):588-600 [FREE Full text] [doi: 10.1089/heq.2019.0042] [Medline: 31720554]
4. Frey W. The US will become 'minority white' in 2045, Census projects. Brookings. 2018. URL: https://www.brookings.edu/articles/the-us-will-become-minority-white-in-2045-census-projects/ [accessed 2025-03-12]
5. Denny JC, Rutter JL, Goldstein DB, Philippakis A, Smoller JW, Jenkins G, et al. The "All of Us" research program. N Engl J Med 2019;381(7):668-676 [FREE Full text] [doi: 10.1056/NEJMsr1809937] [Medline: 31412182]
6. Scharff DP, Mathews KJ, Jackson P, Hoffsuemmer J, Martin E, Edwards D. More than Tuskegee: understanding mistrust about research participation. J Health Care Poor Underserved 2010;21(3):879-897 [FREE Full text] [doi: 10.1353/hpu.0.0323] [Medline: 20693733]
7. National Academies of Sciences, Committee on Women in Science, Engineering, and Medicine, Policy and Global Affairs, National Academies of Sciences, Committee on Improving the Representation of Women and Underrepresented Minorities in Clinical Trials and Research. In: Bibbins-Domingo K, Helman A, editors. Improving Representation in Clinical Trials and Research: Building Research Equity for Women and Underrepresented Groups. Washington (DC): National Academies Press; 2022:26479.
8. Gray DM, Nolan TS, Gregory J, Joseph JJ. Diversity in clinical trials: an opportunity and imperative for community engagement. Lancet Gastroenterol Hepatol 2021;6(8):605-607. [doi: 10.1016/S2468-1253(21)00228-4] [Medline: 34246352]
9. Haley SJ, Southwick LE, Parikh NS, Rivera J, Farrar-Edwards D, Boden-Albala B. Barriers and strategies for recruitment of racial and ethnic minorities: perspectives from neurological clinical research coordinators. J Racial Ethn Health Disparities 2017;4(6):1225-1236 [FREE Full text] [doi: 10.1007/s40615-016-0332-y] [Medline: 28176157]
10. Eggly S, Hamel LM, Heath E, Manning MA, Albrecht TL, Barton E, et al. Partnering around cancer clinical trials (PACCT): study protocol for a randomized trial of a patient and physician communication intervention to increase minority accrual to prostate cancer clinical trials. BMC Cancer 2017;17(1):807 [FREE Full text] [doi: 10.1186/s12885-017-3804-5] [Medline: 29197371]
11. Braunstein J, Sherber NS, Schulman SP, Ding EL, Powe NR. Race, medical researcher distrust, perceived harm, and willingness to participate in cardiovascular prevention trials. Medicine (Baltimore) 2008;87(1):1-9 [FREE Full text] [doi: 10.1097/MD.0b013e3181625d78] [Medline: 18204365]
12. Barton E, Eggly S, Winckles A, Albrecht TL. Strategies of persuasion in offers to participate in cancer clinical trials I: topic placement and topic framing. Commun Med 2014;11(1):1-14. [doi: 10.1558/cam.v11i1.16614] [Medline: 26402960]
13. Largent EA, Eriksen W, Barg FK, Greysen SR, Halpern SD. Participants' perspectives on payment for research participation: a qualitative study. Ethics Hum Res 2022;44(6):14-22 [FREE Full text] [doi: 10.1002/eahr.500147] [Medline: 36316972]
14. Clinical research center dashboard: UBR breakdown (Core Participants). Morehouse School of Medicine. 2022. URL: https://www.msm.edu/Research/research_centersandinstitutes/CRC/index.php [accessed 2025-03-12]
15. Whyte J. Racial and ethnic representation of participants in US clinical trials of new drugs and biologics. JAMA 2022;327(10):985. [doi: 10.1001/jama.2022.0875] [Medline: 35258537]
16. Idris MY, Alema-Mensah E, Olorundare E, Mohammad M, Brown M, Ofili E, et al. Exploring the discursive emphasis on patients and coaches who participated in technology-assisted diabetes self-management education: clinical implementation study of health360x. J Med Internet Res 2022;24(3):e23535 [FREE Full text] [doi: 10.2196/23535] [Medline: 35302506]
17. Jo W, Kim Y, Seo M, Lee N, Park J. Online information analysis on pancreatic cancer in Korea using structural topic model. Sci Rep 2022;12(1):10622 [FREE Full text] [doi: 10.1038/s41598-022-14506-1] [Medline: 35739151]
18. Pachamanova D, Glover W, Li Z, Docktor M, Gujral N. Identifying patterns in administrative tasks through structural topic modeling: a study of task definitions, prevalence, and shifts in a mental health practice's operations during the COVID-19 pandemic. J Am Med Inform Assoc 2021;28(12):2707-2715 [FREE Full text] [doi: 10.1093/jamia/ocab185] [Medline: 34390582]
19. Roberts M, Stewart B, Tingley D, Airoldi E. The structural topic model and applied social science. 2013 Presented at: Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation; December 5, 2013; Lake Tahoe, Nevada URL: https://mimno.infosci.cornell.edu/nips2013ws/slides/stm.pdf

XSL•FO
RenderX

20. Roberts ME, Stewart BM, Tingley D. Stm: an R package for structural topic models. J Stat Soft 2019;91(2):1-40. [doi: 10.18637/jss.v091.i02]
21. Roberts ME, Stewart BM, Tingley D, Lucas C, Leder‐Luis J, Gadarian SK, et al. Structural topic models for open‐ended survey responses. Am J Polit Sci 2014;58(4):1064-1082. [doi: 10.1111/ajps.12103]
22. Boorom M, Goolsby J, Ramsey RP. Relational communication traits and their effect on adaptiveness and sales performance. J Acad Mark Sci 1998 Jan 01;26(1):16-30 [FREE Full text] [doi: 10.1177/0092070398261003]
23. Drollinger T, Comer BL. Salesperson's listening ability as an antecedent to relationship selling. J Bus Ind Mark 2013;28(1). [doi: 10.1108/08858621311285714]
24. Clark LT, Watkins L, Piña IL, Elmer M, Akinboboye O, Gorham M, et al. Increasing diversity in clinical trials: overcoming critical barriers. Curr Probl Cardiol 2019 May;44(5):148-172 [FREE Full text] [doi: 10.1016/j.cpcardiol.2018.11.002] [Medline: 30545650]
25. Nipp RD, Hong K, Paskett ED. Overcoming barriers to clinical trial enrollment. Am Soc Clin Oncol Educ Book 2019;39:105-114 [FREE Full text] [doi: 10.1200/EDBK_243729] [Medline: 31099636]
26. McKinney M, Bell R, Samborski C, Attwood K, Dean G, Eakle K, et al. Clinical trial participation: a pilot study of patient-identified barriers. Clin J Oncol Nurs 2021;25(6):647-654 [FREE Full text] [doi: 10.1188/21.CJON.647-654] [Medline: 34800100]
27. Black Americans and mistrust of the U.S. health care system and medical research. Pew Research Center. 2024. URL: https://www.pewresearch.org/race-and-ethnicity/2024/06/15/black-americans-and-mistrust-of-the-u-s-health-care-system-and-medical-research/ [accessed 2025-03-12]
28. Eggly S, Barton E, Winckles A, Penner LA, Albrecht TL. A disparity of words: racial differences in oncologist-patient communication about clinical trials. Health Expect 2015;18(5):1316-1326 [FREE Full text] [doi: 10.1111/hex.12108] [Medline: 23910630]
29. Racial disparities in automated speech recognition. Semantic Scholar. URL: https://www.semanticscholar.org/paper/Racial-disparities-in-automated-speech-recognition-Koenecke-Nam/219b7266ae848937da170c5510b2bfc66d17859a [accessed 2023-06-26]

Abbreviations

AoURP: All of Us Research Program
EHR: electronic health record
gRoR: genomic return of results
MSM: Morehouse School of Medicine
PPI: participant-provided information
RA: research assistant
SEEC: Southeast Enrollment Center
STM: structural topic modeling
UBR: underrepresented in biomedical research

Edited by A Mavragani; submitted 11.09.24; peer-reviewed by L Salgin; comments to author 10.10.24; revised version received 08.01.25; accepted 14.02.25; published 17.07.25.

Please cite as:
Pemu P, Prude M, McCaslin A, Ojemakinde E, Awad C, Igwe K, Rodriguez A, Foriest J, Idris M
Detecting Conversation Topics in Recruitment Calls of African American Participants to the All of Us Research Program Using Machine Learning: Model Development and Validation Study
JMIR Form Res 2025;9:e65320
URL: https://formative.jmir.org/2025/1/e65320
doi:10.2196/65320
PMID:

©Priscilla Pemu, Michael Prude, Atuarra McCaslin, Elizabeth Ojemakinde, Christopher Awad, Kelechi Igwe, Anny Rodriguez, Jasmine Foriest, Muhammed Idris. Originally published in JMIR Formative Research (https://formative.jmir.org), 17.07.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Development of a Clinic-Based, Sociostructural Intervention to Improve the Provision of Pre-Exposure Prophylaxis for Cisgender Women: Formative Study Using the Assessment, Decision, Adaptation, Production, Topical Experts, Integration, Training, and Testing (ADAPT-ITT) Framework

Rachel Katherine Scott[1,2], MPH, MD; Shawnika J Hull[3], PhD; Deanna Kerrigan[4], PhD; Mandi Pratt-Chapman[4], PhD; Tara Mathias-Prabhu[2], MD; Jennifer Zack[2], MD; Michelle Xu[2], MD; Marisa Sadauskas[5]; Patricia L Moriarty[1]; Tranessa L Hanson[1]; Hannah Arem[1], PhD

[1]MedStar Health, Washington, DC, United States
[2]School of Medicine, Georgetown University, Washington, DC, United States
[3]Rutgers, The State University of New Jersey, New Brunswick, NJ, United States
[4]George Washington University, Washington, DC, United States
[5]School of Medicine, Saint Louis University, St. Louis, MO, United States

Corresponding Author:
Rachel Katherine Scott, MPH, MD
MedStar Health
East Building 5108
110 Irving Street, NW
Washington, DC, 20010
United States
Phone: 1 2028775811
Email: rachel.k.scott@medstar.net

Abstract

Background: Cisgender women (subsequently referred to as women) account for 23% of new HIV diagnoses in the United States. There are significant sociostructural barriers to engagement and retention in the pre-exposure prophylaxis (PrEP) cascade, particularly for Black women.

Objective: In response to the lack of evidence-based interventions (EBIs) to improve PrEP initiation, adherence, and persistence among women in the United States, we developed a clinic-based, sociostructural intervention focused on engagement and retention in the PrEP cascade for women.

Methods: We used the Assessment, Decision, Adaptation, Production, Topical experts, Integration, Training, and Testing (ADAPT-ITT) model to adapt two Centers for Disease Control and Prevention (CDC) best practices in HIV prevention: HIV PrEP Services for Urban Women and Project Shikamana to create a culturally appropriate EBI responsive to Black women's HIV prevention needs. In this paper, we focus on the first 6 steps of iterative adaptation in preparation for pilot testing. We conducted semistructured interviews and focus group discussions with key populations to inform and guide intervention development and used theater testing to evaluate the mock-up of the prototype. We conducted rapid qualitative analysis to identify key themes related to delivering PrEP to the intended population and engaged subject matter experts to refine the prototype.

Results: For the *Assessment phase*, we conducted 10 in-depth interviews with key informants from community-based and HIV-prevention organizations and led 7 focus group discussions (n=4-8) to guide intervention development among health care providers (n=2 groups), PrEP navigators and educators (n=1 group), and female patients (n=4 groups). Key themes included population-specific barriers to PrEP use, namely accessibility and availability, perceived risk, and stigma. In addition, participants advised on model adaptation specific to PrEP navigation, clinic-level training, and social support. For the *Decision phase*, we selected 2 EBIs from the CDC HIV Compendium of Best Practices. For the *Adaptation phase*, we adapted and theater tested a preliminary intervention for feedback. For the *Production phase*, using feedback from theater testing, we created a prototype of the Women's PrEP Project (W-PrEP) designed to address patient-, provider-, and clinic-level barriers to the provision of PrEP

XSL•FO
RenderX

for women through a clinic-wide intervention delivering education, resources, and support (including PrEP navigation). For the *Topical experts* and *Integration phases*, we collected iterative feedback from our advisory board and subject matter experts and integrated feedback into the final prototype.

Conclusions: The adaptation of the W-PrEP integrated key elements of the local context for women with potential exposure to HIV, as well as health care providers, clinic staff, and PrEP navigators. Next steps include training clinic staff in a real-world setting to pilot-test the acceptability, feasibility, and preliminary effectiveness of the intervention.

(JMIR Form Res 2025;9:e75922) doi:10.2196/75922

KEYWORDS

pre-exposure prophylaxis; female patient; HIV infections; ambulatory care facilities; primary prevention; education

Introduction

Background: HIV Prevention for Women in the United States

Cisgender women (subsequently referred to as "women") account for 23% of new HIV diagnoses in the United States [1], highlighting the unrealized opportunity for tailored HIV prevention programs in this population. The HIV epidemic in the United States disproportionately affects Black women of reproductive age [1-3]. Among women, HIV transmission occurs primarily through heterosexual transmission (84%), principally receptive vaginal sex [4,5]. HIV is preventable through consistent use of barrier contraception and HIV pre-exposure prophylaxis (PrEP) [6-12]. Barrier contraception, specifically male condom use, is notably limited by partner power discordance and is implicitly dependent upon partner cooperation [13].

Both daily oral and long-acting injectable formulations of PrEP offer women-controlled options for HIV prevention before risk exposure, circumventing the need for cooperation from sex partners. Pharmacoprevention with either daily oral tenofovir disoproxil fumarate and emtricitabine or injectable bimonthly cabotegravir decreases HIV transmission by up to 99% in women [7-11]. Despite the efficacy of PrEP, the number of prescriptions of PrEP to women in the United States remains dwarfed by the number of women with indications for PrEP [14], highlighting the considerable unmet need for biomedical HIV prevention among women, especially Black women [15-19].

There are significant sociostructural barriers to engagement and retention in the PrEP cascade for women, including medical mistrust, cost, stigma, and lack of access [20-28]. Our previous research predominantly among Black women of reproductive age with demonstrable potential exposure to HIV [29-32] highlights the critical importance of health care provider [20,23,25,26,28,33-38] and community or peer support [20,25,26,28,35-37] in engagement and retention in the PrEP cascade. The Centers for Disease Control and Prevention (CDC) further highlighted the importance of health care providers as PrEP educators with their 2021 guidance that providers should educate all sexually active adults and adolescents about PrEP [39]. Despite their pivotal role, health care provider knowledge of and comfort with prescribing PrEP continues to lag across medical specialties [22,40-45], and despite the synergy of offering integrated PrEP services as part of routine women's health care, most women's health clinics do not routinely educate about or prescribe PrEP [23,44-46]. Both our formative work and the existing literature indicate the sociostructural nature of the factors influencing women's engagement in the PrEP cascade; thus, we used the socioecological model as our guiding theoretical framework for intervention adaptation.

Theoretical Framework: the Socioecological Model

The socioecological model theorizes that an individual's decisions and behaviors result from reciprocal interactions within and between the individual and their social, cultural, and structural environments [47]. Although health-related decisions or behaviors occur at the individual level, they are influenced by individual (ie, psychological), interpersonal (eg, relationship power and relationship commitment), community (ie, cultural norms and stigma), institutional (eg, equitable provision of care and appropriate services), and structural factors (eg, public policy and infrastructure) [48]. As applied to the PrEP cascade for women, our previously published survey of predominantly Black women found the intention to initiate PrEP was associated with different factors by level: at the individual level, positive attitudes toward PrEP and higher self-efficacy; at the community level, perceived acceptability of PrEP among peers and low fear of shame or stigma; and at the institutional level, having discussed PrEP with a health care provider [29]. Our findings underscore the importance of multilevel, clinic-based interventions for women, with education and tools to support women's health providers as well as PrEP navigation to destigmatize PrEP and break down logistical barriers to PrEP initiation and continuation.

Base Cases for Adaptation and Intervention Development

Given the critical role of providers as both PrEP educators and gatekeepers, we previously developed and tested HIV PrEP Services for Urban Women, a clinic-level educational intervention, which used a clinic-wide training and a provider toolkit to promote the provision of universal PrEP services. The intervention significantly improved the provision of PrEP services and PrEP uptake among women at a sexual health center in Washington, DC compared to baseline and was cited as an exemplar evidence-informed intervention (EII) in the CDC Compendium of Evidence-Based Interventions and Best Practices for HIV Prevention [49,50]. While the HIV PrEP Services for Urban Women EII significantly improved the provision of PrEP by providers and PrEP uptake among women of color of reproductive age, it did not actively address the social

XSL•FO
RenderX

barriers of stigma, lack of community or peer support, or medical mistrust that the literature indicated were critical to engagement and retention in the PrEP cascade.

In contrast, Project Shikamana was a sociostructural intervention, also developed by a member of our research team, which showed decreased HIV acquisition among female sex workers in Tanzania and was designated by the CDC as an evidence-based intervention (EBI) for HIV prevention [50-52]. Project Shikamana's strengths include its community-driven, multilevel approach to HIV prevention, involving provider sensitivity training, peer education, navigation and support, community mobilization to address sociostructural factors such as stigma and discrimination, and a focus on women of reproductive age. Given notable differences between the intended populations and the medical infrastructure and medical culture in Tanzania compared to the United States, we chose an evidence-based framework to adapt Project Shikamana while retaining its core strengths.

This Formative Study

In response to the lack of EBI to improve PrEP initiation, adherence, and persistence among women in the United States, we adapted and combined attributes of the two aforementioned CDC best practices for HIV prevention for women to develop the Women's PrEP Project (W-PrEP), a clinic-level, sociostructural intervention. We designed W-PrEP to provide education, resources, and support to empower and enable sexual and reproductive health clinics to provide high-quality, nonstigmatizing, universal PrEP services to women with increased exposure to HIV, in particular Black women of reproductive age. This paper describes the formative work adapting and developing the W-PrEP intervention from the CDC best practices reviewed earlier.

Methods

We conducted a formative study for the intervention adaptation from November 2021 to October 2023, including marked expansion of existing training materials and toolkits.

Ethical Considerations

We received institutional review board approval from the MedStar Health Research Institute Institutional Review Board (#0006908). We obtained written or electronic informed consent from all participants for participation in interviews and focus groups and the recording of these sessions. Participants were compensated with a US $100 gift card for completion of interviews or focus groups. All study data were stored behind a firewall and deidentified before qualitative analysis.

Intervention Development Process Model

We used the Assessment, Decision, Adaptation, Production, Topical experts, Integration, Training, and Testing (ADAPT-ITT) model as a framework to develop a culturally appropriate and tailored EBI [53]. ADAPT-ITT has been used to successfully adapt a host of diverse HIV prevention interventions to meet the needs of a new intended population or setting [54-57]. We used the Project Shikamana EBI and the HIV PrEP Services for Urban Women EII as the base cases to initiate the intervention development process (Table 1). We focused on sociostructural factors that women in the region indicated were barriers to and facilitators of engagement in the PrEP cascade (ie, stigma and discrimination related to HIV and PrEP within health care settings, provider sensitization, and PrEP navigation and support). Using ADAPT-ITT, we built upon those findings and used multiple qualitative methods, such as in-depth interviews and focus group discussions, to refine the existing interventions (Table 2). This paper describes the first 6 steps in the model in preparation for pilot testing.

XSL•FO
RenderX

Table 1. Base case interventions.

Intervention	Description
Project Shikamana EBI[a] [50-52]	• Peer navigation and education: Project Shikamana offered both venue-based peer education, which included condom distribution and HIV counseling and testing, and peer service navigation for women living with HIV, including social support to promote HIV treatment access and adherence. • Provider sensitivity training: Trainings for HIV health care providers promoted nonstigmatizing, quality health services and were conducted in coordination with peer navigators. • Community-led drop-in center: The drop-in center was designed to stimulate social cohesion and mobilization activities and offered HIV counseling, testing, and condoms. The center hosted support groups for community members and convened seminars on topics such as stigma, gender-based violence, income generation and savings groups and HIV prevention.
HIV PrEP Services for Urban Women EII[b] [49,50]	• Structured clinic-wide training and refresher trainings: Trainings emphasized universal provision of PrEP[c] education and counseling and prescribing PrEP, including screening laboratory analyses and techniques to encourage PrEP adherence and persistence. Trainings included sample language for counseling and role-plays. • Provider toolkit: The toolkit featured an EHR[d] PrEP screening prompt, which inquired if patients were interested in discussing starting PrEP as part of their visit, along with screening questions regarding indications for PrEP. The toolkit also offered ID badge card–sized reference "cheat" sheets and workstation reference sheets with indicated laboratory testing, dosing, and follow-up for PrEP and nPEP[e] initiation. • Looped patient-facing PrEP educational videos: The intervention offered looping, short educational videos about PrEP and sexual health in the waiting room, with the goal of increasing PrEP awareness and normalizing PrEP to facilitate conversations with health care providers.

[a]EBI: evidence-based intervention.

[b]EII: evidence-informed intervention.

[c]PrEP: pre-exposure prophylaxis.

[d]EHR: electronic health record.

[e]nPEP: nonoccupational postexposure prophylaxis.

XSL•FO
RenderX

Table 2. Study activities and measures by Assessment, Decision, Adaptation, Production, Topical experts, Integration, Training, and Testing (ADAPT-ITT) phase.

ADAPT-ITT phase [53]	Activities and measures
Assessment: obtain a comprehensive understanding of the target population	• Formation and initial meetings with the joint community and scientific advisory committee, consisting of a social justice advocate, an HIV and HIV prevention advocate, the founder of a nonprofit organization focused on HIV and faith communities, a family planning specialist and reproductive rights advocate, a family physician with expertise in provider education and HIV prevention, and a nurse practitioner with expertise in HIV prevention and implementation science • Key informant interviews with an executive director of the national office of an international women's HIV organization, an executive director of a local women's HIV and HIV prevention community-based organization, a director of clinical services of a local branch of a national reproductive health care service provider, a director of a government sponsored sexual health clinic, a director of women's health at a local HIV health center, a coordinator of a government-sponsored PrEP[a] campaign for women, an HIV prevention coordinator at a local federally qualified community health center, a family medicine provider with a focus on HIV and STI[b] prevention, a family-nurse practitioner working in HIV treatment and prevention, a social worker at a preventive health clinic focused on female sex workers (n=10) • Focus group discussions with health care providers with (n=1 group) and without (n=1 group) experience prescribing PrEP, PrEP navigators and educators (n=1 group), and Black women of reproductive age with (n=2 groups) and without (n=2 groups) previous PrEP experience
Decision: select interventions and decide whether to adopt or adapt	• Review of CDCc Best Practice base cases (ie, Project Shikamana EBI[d] and the HIV PrEP Services for Urban Women EII[e]) • Review of current literature on PrEP and HIV prevention–related interventions for women
Adaptation: use a pretested methodology to better understand how to adapt	• Theater testing of a mock-up of the intervention flow, clinic-wide training materials, and toolkit with the participants from the health care provider focus groups
Production: create an adaptation plan and determine goals	• Production of the initial prototype, including adaptation or development of the following: (1) clinic-level PrEP rollout toolkit, clinical team toolkits, and PrEP navigator toolkit; clinic-wide training and PrEP navigator training; increased availability and accessibility of PrEP services; PrEP navigation program, including delineation of the role of the navigator; and patient-facing PrEP educational resources
Topical experts: provide substantive content and technical assistance	• Engagement of the community and scientific advisory committee to elicit feedback on the intervention prototype
Integration: integrate all forms of information	• Reflect feedback from the advisory committee into the refined intervention prototype

[a]PrEP: pre-exposure prophylaxis.

[b]STI: sexually transmitted infection.

[c]CDC: Centers for Disease Control and Prevention.

[d]EBI: evidence-based intervention.

[e]EII: evidence-informed intervention.

Study Setting

Washington, DC is an epicenter of the US HIV epidemic. At the time of the intervention refinement, the prevalence of HIV among women in Washington, DC was 1209 per 100,000, nearly 7-fold the national average (170 per 100,000) [58]; the prevalence among all women was 0.9%, but nearly double (1.7%) among Black women [58]. At the time of this study, Black women represented 1 in 4 new HIV diagnoses in Washington, DC and 90% of new diagnoses among women [58]. This formative study was conducted online by the research institute associated with an inner-city tertiary care center. Given the timing during the COVID-19 pandemic and to increase access for participants who would be unable to attend in person, all qualitative research was conducted remotely via a secure Microsoft Teams platform.

Qualitative Tools and Data Collection

We conducted semistructured in-depth interviews with key informants as part of the *Assessment* phase. We sought key informants with intimate knowledge and understanding of barriers to and facilitators of PrEP among women, with a focus on individual, social, and structural factors. We identified key informants from local community-based organizations, hospitals, clinics, and HIV prevention organizations through referrals from the Washington, DC. Center for AIDS Research; our joint community and scientific advisory board; and clinical and community professional relationships of the research team. Target recruitment was 10 key informant interviews, and key informants were invited via email by the principal investigator to participate. The key informant interviews focused upon (1) factors influencing PrEP use among women and (2) intervention components key to improving equitable provision of PrEP on a clinic level as well as complementary tools and materials that

XSL•FO
RenderX

may support community needs and priorities. Accordingly, interview questions focused on (1) sociostructural barriers to and facilitators of PrEP use among Black women of reproductive age and (2) assessment of the model components for adaptation. Salient themes and questions arising from the key informant interviews informed the focus group discussion guides.

We also conducted 7 focus group discussions, separating participants into groups of (1) health care providers with experience prescribing PrEP; (2) health care providers without experience prescribing PrEP; (3) PrEP clinical navigators, peer navigators, and educators; (4) Black women of reproductive age with experience taking PrEP (n=2 groups); and (5) Black women of reproductive age without previous experience taking PrEP (n=2 groups). Providers, navigators and educators, and patients were recruited from medical establishments (eg, hospital-based clinics and community health centers) throughout Washington, DC. We invited women's health providers from varying fields (eg, obstetrics and gynecology, family medicine, and pediatrics), institution types (eg, free-standing reproductive health center, government-sponsored sexual health center, tertiary care center, and federally qualified community health center), and medical training types (ie, physician, nurse practitioner, and physician's assistant) to participate via email. PrEP navigators and educators from differing institution types were also invited to participate via email. A formal email invitation to participate in interviews was sent by the principal investigator, and a research coordinator then followed up to assist with scheduling. Patient participants were identified from the women's health clinics and the electronic health record (EHR) and were recruited and scheduled by research coordinators via phone.

Health care provider focus group guides concentrated on the proposed PrEP training materials and format, identification of key toolkit resources to provide PrEP services and integration of PrEP into sexual and reproductive health care, and inclusion of a PrEP navigator and educator in the clinic setting. Navigator and educator focus group guides centered on the roles and characteristics of the PrEP navigator as well as training and resources. Patient focus group guides explored women's health priorities and barriers to preventive health, such as PrEP and ideal provision of PrEP services, including education and access, PrEP navigation and the key characteristics and roles of a PrEP navigator and educator, and social support.

We conducted theater testing of a mock-up of the adapted training and planned intervention and implementation flowchart among these same health care providers in 2 follow-up focus group discussions. Providers offered feedback on the content and delivery of Microsoft PowerPoint training modules and the multipronged intervention.

Qualitative Analysis

We audio recorded interviews and focus groups with the consent of participants, which were then transcribed verbatim. We applied rapid qualitative analysis [59-61] to systematically use qualitative feedback from the interviews and focus groups to refine the intervention. To this end, we developed summary templates to identify the coder, respondent role, and coding domain (ie, strengths, information gaps, and actionable improvements). We conducted analyses of a subset of interviews (ie, 1 from each respondent type) to ensure adequate fit of the template to the data, and then we coded the remaining transcripts. For the formative ADAPT-ITT phases (ie, *assessment* through *integration*), we developed and populated matrices to identify salient themes that represented topics, information, or activities to include in the adapted intervention. For the *training* and *testing* ADAPT-ITT phases, we developed and populated matrices to align with germane components of the practical, robust, implementation, and sustainability model implementation framework. These summaries guided team discussions and W-PrEP intervention refinement.

Results

Assessment Phase

Joint Community and Scientific Advisory Committee

In this first phase of ADAPT-ITT, we convened a joint community and scientific advisory committee (Table 2). In the committee's initial meeting, the research team provided an overview of the existing research, EII, and EBIs for women as the basis for discussion of the gaps in HIV prevention for women and opportunities for HIV prevention interventions. The committee discussion supported the focus of the key informant interview guide on sociostructural barriers to facilitators of PrEP for women and assessment of base case components for adaptation.

Key Informant Interviews

We completed 10 semistructured interviews with community-based organization leadership, clinic and HIV prevention operation managers and leadership, health care providers, social workers, and educators and coordinators. Table 2 describes participants. Illustrative quotations from the key informant interviews (and from the subsequent qualitative portions of the study) are presented in Multimedia Appendix 1. Key informants identified several key barriers to PrEP use among Black women of reproductive age. The most commonly cited barrier was the lack of availability and accessibility of PrEP education and services for women, including the lack of integrated HIV prevention and PrEP services into primary women's health care, the challenging logistics of navigating PrEP care in light of competing priorities and social determinants of health, and the lack of culturally appropriate PrEP marketing and educational materials for women. Suggested opportunities included increased provider education; integration of PrEP into women's primary and reproductive and sexual health care; and use of the full clinical team, including PrEP navigators, serving as PrEP educators and ambassadors. Key informants additionally identified low perceived risk of HIV acquisition and low awareness of PrEP as critical barriers for women. Many referenced the dearth of public health and industry marketing campaigns focused upon women (compared to men who have sex with men). Multiple key informants noted continued stigma related to taking a medication associated with sexual "promiscuity," men who have sex with men, and HIV and barriers to disclosure of PrEP use to support networks, including partners and family members.

XSL•FO
RenderX

We queried key informants about the adaptation of Project Shikamana and the HIV PrEP Services for Urban Women as well as key components to retain. There was universal enthusiasm for the inclusion of a PrEP navigator. Many expressed that PrEP navigators should be from the community and reflect the intended demographic (ie, Black women of reproductive age). There was a sentiment that PrEP navigators should serve as representatives of the community in the clinical setting as a means of combating medical mistrust and misconceptions surrounding PrEP. Participants suggested that PrEP navigators could also identify community needs and assets surrounding PrEP and HIV prevention. Many participants also identified the importance of the navigators in PrEP education and in navigation of PrEP services, including insurance prior authorization, appointments, and prescription assistance programs for uninsured or underinsured patients.

In terms of PrEP training for the clinic team, key themes included centering the community and patient voice in the training, promoting sex-positive language (versus risk-based language) as a nonstigmatizing way of conversing about PrEP and HIV prevention, and the importance of bias training (to reduce racial, sex, and gender biases).

Although key informants underscored the importance of community support in engagement and retention in the PrEP cascade for women, they emphasized harnessing the power of existing community-based organizations and referring to these organizations while using intervention resources to create a supportive, safe, stigma-free environment to learn about PrEP and access PrEP within the clinic.

Focus Group Discussions

Through the key informant interviews, we identified tools, strategies, and experiences that the intervention should address. In the next phase, we conducted focus groups to gain insights into how to develop the program content. We conducted 3 sets of focus group discussions separately with providers, PrEP navigators and educators, and patients, specifically Black women of reproductive age (4-8 participants per group).

Focus Group Discussions: Health Care Providers

A total of 10 health care providers participated in 2 focus group discussions: one group with 6 providers with PrEP experience and another group with 4 providers without PrEP experience. Providers across both groups agreed upon the utility of workflow aids, such as EHR order sets, PrEP reference “cheat” sheets, and badge cards or “badge buddies” to facilitate incorporation of PrEP care into their clinical practices. There was also a suggestion for a clinic PrEP champion as a resource to assist with complicated clinical scenarios. Providers agreed that telehealth was helpful for PrEP continuation, especially for patients with competing priorities or social determinants of health (eg, transportation and limited time off from work) that may limit frequent in-person visits. Providers were enthusiastic about the potential for a PrEP navigator. There was consensus that PrEP clinical navigation (eg, obtaining insurance prior authorization and assisting with rescheduling appointments) was logistically key to integrating PrEP into a bustling clinical practice with overstretched providers. In addition, they supported a PrEP navigator serving to introduce PrEP to patients and provide additional education about PrEP to alleviate some of the logistical time constraints on patient visits. Providers prioritized clinical navigation and education services (over social support) to operably integrate PrEP services into their clinical practice. Several providers reasoned that social support issues outside of the scope of the clinical PrEP navigator could be referred to the outpatient social work team.

Providers preferred in-person trainings (to online or hybrid), with an asynchronous alternative for rotating learners or review at a later time point. Multiple participants emphasized the importance of refresher trainings several months after the initial training. There was a call for interactive training activities (eg, role-playing) both to keep participants engaged and to enact new skills. In addition to education on HIV prevention and PrEP, providers requested inclusion of practical tips and logistics and sample scripts for incorporation of PrEP services into their clinical practice in a nonstigmatizing way. In addition, providers requested the incorporation of the epidemiologic context for HIV prevention among women and specifically requested centering the community and patient voice and perspective throughout the trainings.

Focus Group Discussions: Navigators and Educators

PrEP peer and clinical navigators and educators (n=4) participated in the focus group. The group agreed that the ability to connect quickly with patients was more important than shared background or personal experience with PrEP. They emphasized the importance of navigation to find room in a patient’s lifestyle for PrEP (in the face of competing priorities) but also mentioned the importance of social and emotional support and referral to social services.

The group preferred in-person training and supported a supplemental asynchronous option. They shared the providers’ preference for interactive trainings with role-playing. In addition to the basics of HIV prevention and PrEP, navigators and educators identified key training topics of social and structural determinants of health, advocacy skills, and sexual and reproductive health.

Focus Group Discussions: Patients

A total of 16 patients (n=8 with experience taking PrEP and n=8 without experience taking PrEP) participated in 4 focus group discussions, divided by previous PrEP experience. Preventive health (eg, nutrition, physical activity, mental health, and routine health screenings) was identified as a common health goal. Patients described provider bias surrounding sexual health, costs associated with navigating the health system and with accessing healthy foods and fitness options, and competing priorities as barriers to meeting their personal health goals.

Nearly all patient participants were enthusiastic about culturally appropriate educational videos about PrEP to increase PrEP awareness in the waiting room; there was also interest in women-centered PrEP pamphlets, posters, and links to websites with additional information about PrEP. Although there was no consensus on who (eg, nurse, health care provider, or PrEP navigator) should introduce PrEP, there was agreement that

XSL•FO
RenderX

what mattered most was that PrEP was introduced in a universal, nonstigmatizing manner.

There was agreement that the PrEP navigator should be a cisgender woman who shares their cultural and community background (ie, Black and familiar with or from the locality), ideally someone who is currently taking PrEP or has taken it in the past. There was a strong appeal in being able to discuss PrEP with a trusted individual having similar life experiences and first-hand experience with PrEP yet outside of the medical field. Participants voiced that the roles of educators, prevention navigators, and supportive allies in the process were all important.

There was little consensus on the ideal setting or mechanism for social support, given that participants accessed and used social support systems in different ways. Preferences were split between in-person and online support mechanisms. Several supported an online option that could provide flexibility and continuity for women with competing priorities. Those who preferred an in-person support setting were divided between pop-up events versus a permanent community support setting; those who favored a permanent space were further conflicted on whether the space should be associated with the hospital or embedded in the community.

Decision

In preparation for the *decision* phase, the research team reviewed the findings of the key informant interview and focus group discussion rapid qualitative analyses and met with the joint advisory board and 5 additional national HIV prevention intervention experts to discuss existing resources and planned intervention strategies. The research team also updated a scoping review of the contemporary relevant literature. The research team reaffirmed the plan to adapt and build upon 2 base cases (Table 2), using the multipronged, sociostructural approach of Project Shikamana and the clinic-centered intervention setting and educational focus of the HIV PrEP Services for Urban Women. Notably, given the feedback from the key informant interviews on harnessing existing community support resources and the lack of consensus on social and community support mechanisms, the research team decided to focus on the "provider sensitivity training" and "education and navigation and support" components of Project Shikamana, specifically creating a nonstigmatizing, supportive, and safe environment for the provision of HIV prevention education and services within the clinic setting (with referrals to outside organizations) rather than creating an intervention-specific community support component. On the basis of our extensive background research and our findings in this formative work, the research team affirmed the decision to use the socioecological model given the multipronged and multilevel approach of the intervention [29,62,63].

Adaptation

Provider focus group participants were invited back for theater testing of the planned intervention flowchart and a mock-up of the adapted trainings and toolkits, expanding on the HIV PrEP Services for Urban Women and Project Shikamana's provider sensitivity training, incorporating lessons learned from the *assessment* phase. We elicited topic-specific feedback on length, level of detail, and preferred formats and mediums for learning and skill enactment within the training modules as well as feedback on specific slides. In addition, we encouraged suggestions from participants to improve or refine the toolkits and intervention flowchart. Participants emphasized the overall approval of the intervention mock-up and provided detailed feedback to improve clarity and incorporate additional resources or tools, such as a badge card with reminders about prescribing requirements.

Production

Incorporating the theater testing feedback, we fleshed out a prototype for each of the components of the multipronged intervention; Table 3 outlines the intervention components. We adapted and expanded upon the clinic-wide trainings, incorporating more role-play and segments with the patient perspective taken from previous qualitative work [30], as well as identified key topics, such as how to incorporate PrEP care into clinical practice. We curated a PrEP clinical navigator training program using materials from the Please PrEP Me navigator training while incorporating feedback from qualitative data collection [64-66]. In addition, we collated educational resources for the clinic team, PrEP navigators, and patients, including a patient-facing educational website. We updated and developed a variety of toolkits for the clinic-level rollout of W-PrEP and rollout of long-acting injectable PrEP as well as role-specific toolkits for the clinic team members and the PrEP navigator. We also created a PrEP clinical navigation flowchart, with multiple options for modification and adaptation depending on the availability of clinical space and when patients had downtime between clinical activities (eg, triage by medical assistant and evaluation by health care provider).

XSL•FO
RenderX

Table 3. Adapted Women's PrEP Project (W-PrEP) intervention components.

Component	Description
Clinic-level PrEP[a] rollout toolkit	• W-PrEP implementation checklist • Algorithm for universal PrEP counseling and services • Implementation algorithms for rollout of new PrEP formulations (including prior authorization) • PrEP navigator job description • Identification of a PrEP champion
PrEP champion	• Health care provider to serve as a PrEP expert and site PrEP champion • Responsibilities include facilitation of trainings and consultation on both clinic-level implementation of PrEP services and complex clinical cases
Clinic-wide training	• Key training topics include the following: HIV testing (which test at which time) and STI[b] testing; taking a patient-centered sexual and social history; positive reframing of prevention communication and counseling around PrEP, how to incorporate patient-centered PrEP and nPEP[c] provision for women into routine care; PrEP initiation, continuation and persistence, and discontinuation; nPEP initiation and transition to PrEP; implicit and explicit bias training; trauma-informed care; epidemiology of the US HIV epidemic; and social and structural determinants of health related to HIV • Key features are as follows: synchronous, interactive Microsoft PowerPoint presentations scripted for reproducibility and scalability; asynchronous prerecorded modules to accommodate irregular schedules of health care providers and rotating learners, with activities to engage the asynchronous individual learner; and centering of the patient voice and experience throughout both modules
Increased availability and accessibility of PrEP follow-up	• Increased availability of telehealth and web-based medical visits for PrEP follow-up visits: These visits are designed as alternatives to in-person visits for women who were not ready to start PrEP but wanted to learn more about PrEP (from providers and PrEP navigators) or for women who have been prescribed or started PrEP to follow up regarding issues with prescription pickup, medication side effects, adherence, or other concerns. • Off-site laboratory referrals: Providers will offer referrals to off-site commercial laboratories for follow-up STI and HIV testing if more convenient • "PrEP talk" with PrEP navigator: Email, text, and telephone follow-up options will be offered to address concerns and offer education.
Clinic team toolkit	• EHR[d] PrEP order set (eg, STI screening laboratory orders, PrEP injection or prescription orders, and HIV testing orders) • EHR PrEP quick text to facilitate documentation of PrEP education and counseling and prescription and PrEP EHR prompts to remind and encourage providers to counsel all patients about PrEP (dependent upon EHR settings) • PrEP and nPEP badge buddy: an ID badge card–sized PrEP and nPEP reference "cheat" sheet with recommended STI and HIV screening, dosing information, and follow-up • PrEP and nPEP reference "cheat" sheet and checklists: full-size reference sheet and PrEP start and continuation checklists for quick reference during a busy clinic day (recommended to be posted at the provider workstation) • Sample scripts for introduction to PrEP (for medical assistants and nurses) and sample scripts for introduction and counseling for prescribing providers • "Ask me about PrEP!" swag (eg, pins and lanyards)
PrEP navigator training	• In addition to the clinic-wide PrEP training, through a combination of tailored one-on-one training from a PrEP education and navigation expert, independent learning, and online modules, the PrEP navigator will receive training in the following: PrEP navigation through the Please PrEP Me Navigator training [64], implicit biases in the health care setting [66], trauma-informed care [65], and introduction to reproductive and sexual health topics (led by a site health care provider and topically tailored to the clinical context)
PrEP navigator toolkit	• In addition to the aforementioned clinical tools, the PrEP navigators will have access to the following: PrEP quick text to facilitate documentation of PrEP navigator education and counseling, including assessment of barriers to PrEP; sample script for introduction to PrEP and counseling about PrEP; and patient-facing educational resources
PrEP navigation and support services	• Informed by our formative work, PrEP navigators will be cisgender women, ideally from the same community the clinic serves and with first-hand experience with ARVs[f] through PrEP and HIV. PrEP navigators will provide the following: PrEP educational services—introducing PrEP to patients, combating misperceptions or myths about PrEP; PrEP social and adherence support—breaking down social barriers to PrEP and mitigating the internalization of PrEP-related stigma and building and encouraging individual agency or self-efficacy to initiate and adhere to PrEP; and PrEP navigation services—facilitating access to PrEP appointments, assisting with insurance authorization, and referrals to indicated social services (eg, substance use, gender-based violence, and mental health)

XSL•FO
RenderX

Component	Description
Patient-facing PrEP resources	• Videos on PrEP and women's health and sexual health topics (selected by the clinical site) will be spliced and played on loop in the clinic waiting room to increase awareness of PrEP • Inclusive pamphlets on PrEP in the waiting room and examination rooms • Inclusive posters about PrEP and HIV prevention displayed in the waiting room and examination rooms • W-PrEP educational website QR code available in the waiting area and distributed by the PrEP navigator

[a]PrEP: pre-exposure prophylaxis.

[b]STI: sexually transmitted infection.

[c]nPEP: nonoccupational postexposure prophylaxis.

[d]EHR: electronic health record.

[f]ARV: antiretrovirals.

Topical Experts and Integration

We engaged the community and scientific advisory committee as *topical experts* to review the prototypes developed in the *production* phase (ie, the intervention flowchart, toolkits, training modules, and educational resources). The committee expressed enthusiastic support for the adapted intervention and offered minimal feedback to incorporate into the intervention components.

Discussion

Principal Findings

Using the ADAPT-ITT model, we iteratively adapted two CDC best practices in HIV prevention, using the sociostructural approach of Project Shikamana and building from the training model of the HIV PrEP Services for Urban Women. We found that while the core elements of provider sensitivity training, navigation, and support of Project Shikamana were maintained, the urban mid-Atlantic context lent itself better to a clinic-based approach. Thus, we focused on a supportive, trauma-informed clinical environment without a separate community support venue. In addition, the navigation model shifted from a community peer support model to a PrEP navigator supporting clinic workflows to meet the needs of a bustling urban women's health clinic with a high patient volume. The training from the HIV PrEP Services for Urban Women was significantly expanded to center the patient voice and prioritize shared decision-making throughout the modules. We incorporated group activities, discussions, and role-plays to make the training more interactive. Added content included a focus on social and structural determinants of health, trauma-informed care, and incorporation of PrEP counseling into clinical practice. The adapted W-PrEP intervention features clinic-level modular trainings; a toolkit and resources for clinic-level implementation of PrEP services; toolkits and resources for provider and clinic team members; a clinical PrEP navigation program, including training and resources; and patient-facing woman-centric educational and support resources, including a website focused on PrEP for women, all to be tested in the pilot *testing* phase.

Comparison With Prior Work

At the time of intervention development, Project Shikamana and the HIV PrEP Services for Urban Women were the only two best practices in the CDC compendium addressing HIV prevention for women. W-PrEP is now among a growing number of interventions published in the last several years intended to increase engagement and retention in the PrEP cascade among women, several specifically responsive to Black women [67-71]. Similar to the POWER Up intervention [70], developed in parallel with W-PrEP and currently under evaluation, W-PrEP uses a sociostructural approach. As a clinic-level intervention, W-PrEP expands upon our previous work [49,72] to more fully address identified patient-, provider-, and clinic-level socioecological barriers, such as low awareness among patients, low knowledge and confidence in how to provide PrEP among providers, and clinic-level challenges of systems and logistics for obtaining prior authorization and tracking patient follow-up.

Limitations

W-PrEP was adapted in an urban, mid-Atlantic setting and is specifically responsive to the barriers to PrEP for Black women. Although we posit that our developed intervention is likely generalizable to women with increased potential exposure to HIV throughout the United States, we recognize that there may be cultural differences specific to geographic regions. We catered the intervention development to sexual and reproductive health providers, given the natural synergy between HIV prevention and sexual and reproductive health and the patient populations that these providers care for. We again recognize that this intervention may need additional adaptation for implementation in the primary care setting, including additional modules for providers who also care for patients assigned male sex at birth.

Conclusions

This paper details the adaptation process of two CDC best practices to create a program that is responsive to the needs and experiences of the various groups involved in its delivery. We provide such details because the practicalities and logistics of adaptation are often unarticulated and completed behind a curtain. We sought to provide insights into the adaptation process from the ground level. The adaptation of the W-PrEP program integrated key elements of the local context for women with potential exposure to HIV as well as health care providers, clinic staff, and PrEP navigators. Next steps include piloting the adapted intervention in a real-world setting for feasibility, acceptability, and preliminary efficacy before launching a multisite trial to assess impact on the PrEP cascade.

XSL•FO
RenderX

Acknowledgments

Research reported in this publication was supported by the National Institute of Mental Health of the National Institutes of Health (R34MH128046). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors would like to thank the Women's PrEP Project joint community and scientific advisory board and the participants in the pilot, both patients and clinic staff members. In addition, the authors would like to thank Drs. Jonathan Fuchs, Alan McCord, Dominica Seidman, Rachel Logan, and Anisha Ghandi for their consultations during the adaptation.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

RKS has received investigator-sponsored research funding (managed by MHRI and the IMPAACT Network) from ViiV Healthcare and participated in ViiV Healthcare advisory meetings. DK has received investigator-sponsored research funding (managed by George Washington University) from ViiV Healthcare.

Multimedia Appendix 1
Example quotations by Assessment, Decision, Adaptation, Production, Topical experts, Integration, Training, and Testing (ADAPT-ITT) phase and activity.
[DOCX File , 27 KB - formative_v9i1e75922_app1.docx]

References

1. AtlasPlus - charts. Centers for Disease Control and Prevention (CDC). URL: https://gis.cdc.gov/grasp/nchhstpatlas/charts.html [accessed 2024-12-27]
2. Hess KL, Hu X, Lansky A, Mermin J, Hall HI. Lifetime risk of a diagnosis of HIV infection in the United States. Ann Epidemiol 2017 Apr;27(4):238-243 [FREE Full text] [doi: 10.1016/j.annepidem.2017.02.003] [Medline: 28325538]
3. HIV data: HIV diagnoses, deaths, and prevalence. Centers for Disease Control and Prevention (CDC). URL: https://www.cdc.gov/hiv-data/nhss/hiv-diagnoses-deaths-prevalence.html [accessed 2024-12-27]
4. Fast facts: HIV and women. Centers for Disease Control and Prevention (CDC). URL: https://www.cdc.gov/hiv/data-research/facts-stats/women.html [accessed 2024-12-27]
5. Abdool Karim Q, Sibeko S, Baxter C. Preventing HIV infection in women: a global health imperative. Clin Infect Dis 2010 May 15;50 Suppl 3(Suppl 3):S122-S129 [FREE Full text] [doi: 10.1086/651483] [Medline: 20397940]
6. Beksinska M, Wong R, Smit J. Male and female condoms: their key role in pregnancy and STI/HIV prevention. Best Pract Res Clin Obstet Gynaecol 2020 Jul;66:55-67. [doi: 10.1016/j.bpobgyn.2019.12.001] [Medline: 32007451]
7. Baeten JM, Donnell D, Ndase P, Mugo NR, Campbell JD, Wangisi J, Partners PrEP Study Team. Antiretroviral prophylaxis for HIV prevention in heterosexual men and women. N Engl J Med 2012 Aug 02;367(5):399-410 [FREE Full text] [doi: 10.1056/NEJMoa1108524] [Medline: 22784037]
8. Thigpen MC, Kebaabetswe PM, Paxton LA, Smith DK, Rose CE, Segolodi TM, TDF2 Study Group. Antiretroviral preexposure prophylaxis for heterosexual HIV transmission in Botswana. N Engl J Med 2012 Aug 02;367(5):423-434. [doi: 10.1056/NEJMoa1110711] [Medline: 22784038]
9. Murnane PM, Celum C, Mugo N, Campbell JD, Donnell D, Bukusi E, Partners PrEP Study Team. Efficacy of preexposure prophylaxis for HIV-1 prevention among high-risk heterosexuals: subgroup analyses from a randomized trial. AIDS 2013 Aug 24;27(13):2155-2160 [FREE Full text] [doi: 10.1097/QAD.0b013e3283629037] [Medline: 24384592]
10. Marrazzo J, Tao L, Becker M, Leech AA, Taylor AW, Ussery F, et al. HIV preexposure prophylaxis with emtricitabine and tenofovir disoproxil fumarate among cisgender women. JAMA 2024 Mar 19;331(11):930-937. [doi: 10.1001/jama.2024.0464] [Medline: 38427359]
11. Delany-Moretlwe S, Hughes JP, Bock P, Ouma SG, Hunidzarira P, Kalonji D, HPTN 084 study group. Cabotegravir for the prevention of HIV-1 in women: results from HPTN 084, a phase 3, randomised clinical trial. Lancet 2022 May 07;399(10337):1779-1789 [FREE Full text] [doi: 10.1016/S0140-6736(22)00538-4] [Medline: 35378077]
12. Bekker LG, Das M, Abdool Karim Q, Ahmed K, Batting J, Brumskine W, PURPOSE 1 Study Team. Twice-yearly lenacapavir or daily F/TAF for HIV prevention in cisgender women. N Engl J Med 2024 Oct 03;391(13):1179-1192. [doi: 10.1056/NEJMoa2407001] [Medline: 39046157]
13. Sarkar NN. Barriers to condom use. Eur J Contracept Reprod Health Care 2008 Jun;13(2):114-122. [doi: 10.1080/13625180802011302] [Medline: 18465472]

XSL•FO
RenderX

14. Siegler AJ, Mouhanna F, Giler RM, Weiss K, Pembleton E, Guest J, et al. The prevalence of pre-exposure prophylaxis use and the pre-exposure prophylaxis-to-need ratio in the fourth quarter of 2017, United States. Ann Epidemiol 2018 Dec;28(12):841-849 [FREE Full text] [doi: 10.1016/j.annepidem.2018.06.005] [Medline: 29983236]
15. Smith DK, Van Handel M, Grey J. Estimates of adults with indications for HIV pre-exposure prophylaxis by jurisdiction, transmission risk group, and race/ethnicity, United States, 2015. Ann Epidemiol 2018 Dec;28(12):850-7.e9. [doi: 10.1016/j.annepidem.2018.05.003] [Medline: 29941379]
16. Huang YL, Zhu W, Smith DK, Harris N, Hoover KW. HIV preexposure prophylaxis, by race and ethnicity - United States, 2014-2016. MMWR Morb Mortal Wkly Rep 2018 Oct 19;67(41):1147-1150 [FREE Full text] [doi: 10.15585/mmwr.mm6741a3] [Medline: 30335734]
17. Mapping PrEP: first ever data on PrEP users across the U.S. AIDSVu. URL: https://aidsvu.org/prep/ [accessed 2019-11-06]
18. Bush S, Ng L, Magnuson D, Piontkowsky D, Mera Giler R. Significant uptake of Truvada for preexposure prophylaxis (PrEP) utilization in the US in late 2014. International Association of Providers of AIDS Care. 2015. URL: https://www.iapac.org/AdherenceConference/presentations/ADH10_OA74.pdf [accessed 2025-05-29]
19. Bush S, Magnuson D, Rawlings M, Hawkins T, McCallister S, Mera Giler R. Racial characteristics of FTC/TDF for pre-exposure prophylaxis (PrEP) users in the US. National AIDS Treatment Advocacy Project. 2016. URL: http://www.natap.org/2016/HIV/062216_02.htm [accessed 2025-05-29]
20. Kwakwa HA, Bessias S, Sturgis D, Walton G, Wahome R, Gaye O, et al. Engaging United States black communities in HIV pre-exposure prophylaxis: analysis of a PrEP engagement cascade. J Natl Med Assoc 2018 Oct;110(5):480-485. [doi: 10.1016/j.jnma.2017.12.006] [Medline: 30129509]
21. Kwakwa HA, Bessias S, Sturgis D, Mvula N, Wahome R, Coyle C, et al. Attitudes toward HIV pre-exposure prophylaxis in a United States urban clinic population. AIDS Behav 2016 Jul;20(7):1443-1450. [doi: 10.1007/s10461-016-1407-9] [Medline: 27115399]
22. Bradley E, Forsberg K, Betts JE, DeLuca JB, Kamitani E, Porter SE, et al. Factors affecting pre-exposure prophylaxis implementation for women in the United States: a systematic review. J Womens Health (Larchmt) 2019 Sep;28(9):1272-1285. [doi: 10.1089/jwh.2018.7353] [Medline: 31180253]
23. Aaron E, Blum C, Seidman D, Hoyt MJ, Simone J, Sullivan M, et al. Optimizing delivery of HIV preexposure prophylaxis for women in the United States. AIDS Patient Care STDS 2018 Jan;32(1):16-23 [FREE Full text] [doi: 10.1089/apc.2017.0201] [Medline: 29323558]
24. Ojikutu BO, Bogart LM, Higgins-Biddle M, Dale SK, Allen W, Dominique T, et al. Facilitators and barriers to pre-exposure prophylaxis (PrEP) use among black individuals in the United States: results from the national survey on HIV in the black community (NSHBC). AIDS Behav 2018 Nov;22(11):3576-3587 [FREE Full text] [doi: 10.1007/s10461-018-2067-8] [Medline: 29468493]
25. Wingood GM, Dunkle K, Camp C, Patel S, Painter JE, Rubtsova A, et al. Racial differences and correlates of potential adoption of preexposure prophylaxis: results of a national survey. J Acquir Immune Defic Syndr 2013 Jun 01;63 Suppl 1(01):S95-101 [FREE Full text] [doi: 10.1097/QAI.0b013e3182920126] [Medline: 23673895]
26. Flash CA, Dale SK, Krakower DS. Pre-exposure prophylaxis for HIV prevention in women: current perspectives. Int J Womens Health 2017;9:391-401 [FREE Full text] [doi: 10.2147/IJWH.S113675] [Medline: 28615975]
27. Goparaju L, Experton LS, Praschan NC, Warren-Jeanpiere L, Young MA, Kassaye S. Women want pre-exposure prophylaxis but are advised against it by their HIV-positive counterparts. J AIDS Clin Res 2015 Nov;6(11):1-10 [FREE Full text] [doi: 10.4172/2155-6113.1000522] [Medline: 27019765]
28. Goparaju L, Praschan NC, Warren-Jeanpiere L, Experton LS, Young MA, Kassaye S. Stigma, partners, providers and costs: potential barriers to PrEP uptake among US women. J AIDS Clin Res 2017 Sep;8(9):66 [FREE Full text] [doi: 10.4172/2155-6113.1000730] [Medline: 29201531]
29. Scott RK, Hull SJ, Huang JC, Ye PP, Lotke P, Beverley J, et al. Intention to initiate HIV pre-exposure prophylaxis among cisgender women in a high HIV prevalence U.S. City. Womens Health Issues 2023;33(5):541-550 [FREE Full text] [doi: 10.1016/j.whi.2023.05.009] [Medline: 37479630]
30. Burns JA, Hull SJ, Inuwa A, Moriarty P, Scott RK. Understanding retention in the HIV pre-exposure prophylaxis cascade among cisgender women. AIDS Patient Care STDS 2023 Apr 01;37(4):205-211 [FREE Full text] [doi: 10.1089/apc.2023.0009] [Medline: 36961388]
31. Scott RK, Hull SJ, Huang JC, Coleman M, Ye P, Lotke P, et al. Factors associated with intention to initiate pre-exposure prophylaxis in cisgender women at high behavioral risk for HIV in Washington, D.C. Arch Sex Behav 2022 Jul 27;51(5):2613-2624 [FREE Full text] [doi: 10.1007/s10508-021-02274-4] [Medline: 35622077]
32. Hull SJ, Duan X, Brant AR, Peng Ye P, Lotke PS, Huang JC, et al. Understanding psychosocial determinants of PrEP uptake among cisgender women experiencing heightened HIV risk: implications for multi-level communication intervention. Health Commun 2023 Dec;38(14):3264-3275 [FREE Full text] [doi: 10.1080/10410236.2022.2145781] [Medline: 36398676]
33. Johnson LM, Green HD, Koch B, Stockman JK, Felsher M, Roth AM, et al. Role of social networks and social norms in future PrEP use in a racially diverse sample of at-risk women and members of their social networks. J Acquir Immune Defic Syndr 2021 Apr 01;86(4):422-429 [FREE Full text] [doi: 10.1097/QAI.0000000000002571] [Medline: 33196549]

XSL•FO
RenderX

34. Jackson GY, Darlington CK, Van Tieu H, Brawner BM, Flores DD, Bannon JA, et al. Women's views on communication with health care providers about pre-exposure prophylaxis (PrEP) for HIV prevention. Cult Health Sex 2022 May;24(5):642-656 [FREE Full text] [doi: 10.1080/13691058.2021.1877824] [Medline: 33535887]
35. Garfinkel DB, Alexander KA, McDonald-Mosley R, Willie TC, Decker MR. Predictors of HIV-related risk perception and PrEP acceptability among young adult female family planning patients. AIDS Care 2017 Jun 29;29(6):751-758 [FREE Full text] [doi: 10.1080/09540121.2016.1234679] [Medline: 27680304]
36. Willie TC, Stockman JK, Keene DE, Calabrese SK, Alexander KA, Kershaw TS. Social networks and its impact on women's awareness, interest, and uptake of HIV pre-exposure prophylaxis (PrEP): implications for women experiencing intimate partner violence. J Acquir Immune Defic Syndr 2019 Apr 01;80(4):386-393 [FREE Full text] [doi: 10.1097/QAI.0000000000001935] [Medline: 30570528]
37. Auerbach JD, Kinsky S, Brown G, Charles V. Knowledge, attitudes, and likelihood of pre-exposure prophylaxis (PrEP) use among US women at risk of acquiring HIV. AIDS Patient Care STDS 2015 Mar;29(2):102-110 [FREE Full text] [doi: 10.1089/apc.2014.0142] [Medline: 25513954]
38. Roth A, Felsher M, Tran N, Bellamy S, Martinez-Donate A, Krakower D, et al. Drawing from the theory of planned behaviour to examine pre-exposure prophylaxis uptake intentions among heterosexuals in high HIV prevalence neighbourhoods in Philadelphia, Pennsylvania, USA: an observational study. Sex Health 2019 Jun;16(3):218-224. [doi: 10.1071/SH18081] [Medline: 31079596]
39. US Public Health Service: preexposure prophylaxis for the prevention of HIV infection in the United States-2021 Update: a clinical practice guideline. Centers for Disease Control and Prevention (CDC). URL: https://www.cdc.gov/hiv/pdf/risk/prep/cdc-hiv-prep-guidelines-2021.pdf [accessed 2025-05-29]
40. Blackstock OJ, Patel VV, Felsen U, Park C, Jain S. Pre-exposure prophylaxis prescribing and retention in care among heterosexual women at a community-based comprehensive sexual health clinic. AIDS Care 2017 Jul 04;29(7):866-869 [FREE Full text] [doi: 10.1080/09540121.2017.1286287] [Medline: 28147704]
41. Devlin SA, Ridgway JP, Dawdani A, Enaholo OE, Liegeon G, Kasal N, et al. Adapting provider training and pre-exposure prophylaxis advertising to increase pre-exposure prophylaxis awareness and uptake among black cisgender women. AIDS Patient Care STDS 2023 Dec 01;37(12):574-582. [doi: 10.1089/apc.2023.0188] [Medline: 38011350]
42. Petroll AE, Walsh JL, Owczarzak JL, McAuliffe TL, Bogart LM, Kelly JA. PrEP awareness, familiarity, comfort, and prescribing experience among us primary care providers and HIV specialists. AIDS Behav 2017 May;21(5):1256-1267 [FREE Full text] [doi: 10.1007/s10461-016-1625-1] [Medline: 27885552]
43. Castel AD, Feaster DJ, Tang W, Willis S, Jordan H, Villamizar K, et al. Understanding HIV care provider attitudes regarding intentions to prescribe PrEP. J Acquir Immune Defic Syndr 2015 Dec 15;70(5):520-528 [FREE Full text] [doi: 10.1097/QAI.0000000000000780] [Medline: 26247895]
44. Seidman D, Weber S, Carlson K, Witt J. Family planning providers' role in offering PrEP to women. Contraception 2018 Jun;97(6):467-470 [FREE Full text] [doi: 10.1016/j.contraception.2018.01.007] [Medline: 29408284]
45. Seidman D, Carlson K, Weber S, Witt J, Kelly PJ. United States family planning providers' knowledge of and attitudes towards preexposure prophylaxis for HIV prevention: a national survey. Contraception 2016 May;93(5):463-469. [doi: 10.1016/j.contraception.2015.12.018] [Medline: 26772906]
46. Sales JM, Escoffery C, Hussen SA, Haddad LB, Phillips A, Filipowicz T, et al. Pre-exposure prophylaxis integration into family planning services at title X clinics in the southeastern united states: a geographically-targeted mixed methods study (phase 1 ATN 155). JMIR Res Protoc 2019 May 06;8(6):e12774 [FREE Full text] [doi: 10.2196/12774] [Medline: 31199344]
47. Bronfenbrenner U. The Ecology of Human Development: Experiments by Nature and Design. Harvard, CA: Harvard University Press; 1979.
48. Kaufman MR, Cornish F, Zimmerman RS, Johnson BT. Health behavior change models for HIV prevention and AIDS care: practical recommendations for a multi-level approach. J Acquir Immune Defic Syndr 2014 Aug 15;66 Suppl 3(Suppl 3):S250-S258 [FREE Full text] [doi: 10.1097/QAI.0000000000000236] [Medline: 25007194]
49. Scott RK, Deyarmond M, Marwitz S, Huang JC, Moriarty P, Visconti AJ, et al. Implementation of an educational intervention to improve HIV pre-exposure prophylaxis services for women in an urban sexual health clinic. AIDS Patient Care STDS 2023 Sep 01;37(9):447-457 [FREE Full text] [doi: 10.1089/apc.2023.0107] [Medline: 37713289]
50. Compendium: intervention research. Centers for Disease Control and Prevention (CDC). URL: https://www.cdc.gov/hivpartners/php/compendium/index.html [accessed 2024-08-01]
51. Kerrigan D, Mbwambo J, Likindikoki S, Beckham S, Mwampashi A, Shembilu C, et al. Project Shikamana: baseline findings from a community empowerment-based combination HIV prevention trial among female sex workers in Iringa, Tanzania. J Acquir Immune Defic Syndr 2017 Jan 01;74 Suppl 1(Suppl 1):S60-S68 [FREE Full text] [doi: 10.1097/QAI.0000000000001203] [Medline: 27930613]
52. Kerrigan D, Mbwambo J, Likindikoki S, Davis W, Mantsios A, Beckham SW, et al. Project Shikamana: community empowerment-based combination HIV prevention significantly impacts HIV incidence and care continuum outcomes among female sex workers in Iringa, Tanzania. J Acquir Immune Defic Syndr 2019 Oct 01;82(2):141-148 [FREE Full text] [doi: 10.1097/QAI.0000000000002123] [Medline: 31513552]

XSL•FO
RenderX

53. Wingood GM, DiClemente RJ. The ADAPT-ITT model: a novel method of adapting evidence-based HIV Interventions. J Acquir Immune Defic Syndr 2008 Mar 01;47 Suppl 1:S40-S46. [doi: 10.1097/QAI.0b013e3181605df1] [Medline: 18301133]
54. Crooks N, Debra A, Coleman D, Sosina W, Singer R, Jeremiah R, et al. Application of ADAPT-ITT: adapting an evidence-based HIV/STI mother-daughter prevention intervention for Black male caregivers and girls. BMC Public Health 2023 Jul 26;23(1):1426 [FREE Full text] [doi: 10.1186/s12889-023-16364-6] [Medline: 37491213]
55. Rotheram-Borus MJ, Tomlinson M, Stewart J, Skiti Z, Rabie S, Wang J, et al. Soccer and vocational training are ineffective delivery strategies to prevent HIV and substance abuse by young, South African men: a cluster randomized controlled trial. AIDS Behav 2024 Dec 11;28(12):3929-3943 [FREE Full text] [doi: 10.1007/s10461-024-04458-0] [Medline: 39259241]
56. Davis T, DiClemente RJ, Prietula M. Using ADAPT-ITT to modify a telephone-based HIV prevention intervention for SMS delivery: formative study. JMIR Form Res 2020 Oct 06;4(10):e22485 [FREE Full text] [doi: 10.2196/22485] [Medline: 32831178]
57. Conserve DF, Abu-Ba'are GR, Janson S, Mhando F, Munisi GV, Drezgic B, et al. Development and feasibility of the peer and nurse-led HIV Self-Testing Education and Promotion (STEP) intervention among social networks of men in Dar es Salaam, Tanzania: application of the ADAPT-ITT model. BMC Health Serv Res 2024 Oct 02;24(1):1166 [FREE Full text] [doi: 10.1186/s12913-024-11586-9] [Medline: 39358728]
58. Bowser M, Young RM, Nesbitt LS. Annual epidemiology and surveillance report: data through December 2019. DC Health. URL: https://dchealth.dc.gov/service/hiv-reports-and-publications [accessed 2025-05-29]
59. Hamilton AB. Qualitative methods in rapid turn-around health services research. VA Women's Health Research Network. URL: https://www.betterevaluation.org/tools-resources/qualitative-methods-rapid-turn-around-health-services-research [accessed 2025-01-04]
60. Palinkas LA, Mendon SJ, Hamilton AB. Innovations in mixed methods evaluations. Annu Rev Public Health 2019 Apr 01;40:423-442 [FREE Full text] [doi: 10.1146/annurev-publhealth-040218-044215] [Medline: 30633710]
61. Gale RC, Wu J, Erhardt T, Bounthavong M, Reardon CM, Damschroder LJ, et al. Comparison of rapid vs in-depth qualitative analytic methods from a process evaluation of academic detailing in the Veterans Health Administration. Implement Sci 2019 Mar 01;14(1):11 [FREE Full text] [doi: 10.1186/s13012-019-0853-y] [Medline: 30709368]
62. Price DM, Unger Z, Wu Y, Meyers K, Golub SA. Clinic-level strategies for mitigating structural and interpersonal HIV pre-exposure prophylaxis stigma. AIDS Patient Care STDS 2022 Mar;36(3):115-122 [FREE Full text] [doi: 10.1089/apc.2021.0176] [Medline: 35289691]
63. Ojikutu BO, Mayer K. HIV prevention among black women in the US-time for multimodal integrated strategies. JAMA Netw Open 2021 Apr 01;4(4):e215356 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.5356] [Medline: 33835180]
64. PrEP navigation manual. HealthHIV. URL: https://pleaseprepme.org/prepnavigatormanual/ [accessed 2025-01-22]
65. Intro to trauma informed care training modules. Trauma Informed Oregon. URL: https://traumainformedoregon.org/resources/training/tic-intro-training-modules/ [accessed 2024-08-02]
66. Implicit bias: a practical guide for healthcare settings. The George Washington University School of Medicine and Health Sciences. URL: https://cme.smhs.gwu.edu/www.dcrxce.com/content/implicit-bias-practical-guide-healthcare-settings#group-tabs-node-course-default1 [accessed 2024-08-02]
67. Johnson AK, Haider S, Nikolajuk K, Kuhns LM, Ott E, Motley D, et al. An mHealth intervention to improve pre-exposure prophylaxis knowledge among young black women in family planning clinics: development and usability study. JMIR Form Res 2022 Jul 28;6(7):e37738 [FREE Full text] [doi: 10.2196/37738] [Medline: 35900830]
68. Randolph SD, Johnson R. A salon-based intervention to improve PrEP uptake among black women. N Engl J Med 2024 Feb 29;390(9):776-777. [doi: 10.1056/nejmp2313708]
69. Blumenthal J, Jain S, He F, Amico K, Kofron R, Ellorin E, et al. Results from a pre-exposure prophylaxis demonstration project for at-risk cisgender women in the United States. Clin Infect Dis 2021 Oct 05;73(7):1149-1156 [FREE Full text] [doi: 10.1093/cid/ciab328] [Medline: 33864370]
70. Ridgway JP, Devlin SA, Friedman EE, Enaholo OE, Pyra M, Hirschhorn LR, et al. POWER Up-Improving pre-exposure prophylaxis (PrEP) uptake among Black cisgender women in the Southern United States: protocol for a stepped-wedge cluster randomized trial (SW-CRT). PLoS One 2023;18(5):e0285858 [FREE Full text] [doi: 10.1371/journal.pone.0285858] [Medline: 37196008]
71. Teitelman AM, Koblin BA, Brawner BM, Davis A, Darlington C, Lipsky RK, et al. Just4Us: development of a counselor-navigator and text message intervention to promote PrEP uptake among cisgender women at elevated risk for HIV. J Assoc Nurses AIDS Care 2021;32(2):188-204 [FREE Full text] [doi: 10.1097/JNC.0000000000000233] [Medline: 33427767]
72. Brant AR, Dhillon P, Hull S, Coleman M, Ye PP, Lotke PS, et al. Integrating HIV pre-exposure prophylaxis into family planning care: a RE-AIM framework evaluation. AIDS Patient Care STDS 2020 Jun;34(6):259-266 [FREE Full text] [doi: 10.1089/apc.2020.0004] [Medline: 32484743]

XSL•FO
RenderX

Abbreviations

ADAPT-ITT: Assessment, Decision, Adaptation, Production, Topical experts, Integration, Training, and Testing
CDC: Centers for Disease Control and Prevention
EBI: evidence-based intervention
EHR: electronic health record
EII: evidence-informed intervention
PrEP: pre-exposure prophylaxis
W-PrEP: Women's PrEP Project

Edited by A Mavragani; submitted 13.04.25; peer-reviewed by M Sohail, M Pyra; comments to author 27.05.25; revised version received 13.06.25; accepted 14.06.25; published 06.08.25.

Please cite as:
Scott RK, Hull SJ, Kerrigan D, Pratt-Chapman M, Mathias-Prabhu T, Zack J, Xu M, Sadauskas M, Moriarty PL, Hanson TL, Arem H
Development of a Clinic-Based, Sociostructural Intervention to Improve the Provision of Pre-Exposure Prophylaxis for Cisgender Women: Formative Study Using the Assessment, Decision, Adaptation, Production, Topical Experts, Integration, Training, and Testing (ADAPT-ITT) Framework
JMIR Form Res 2025;9:e75922
URL: https://formative.jmir.org/2025/1/e75922
doi:10.2196/75922
PMID:

©Rachel Katherine Scott, Shawnika J Hull, Deanna Kerrigan, Mandi Pratt-Chapman, Tara Mathias-Prabhu, Jennifer Zack, Michelle Xu, Marisa Sadauskas, Patricia L Moriarty, Tranessa L Hanson, Hannah Arem. Originally published in JMIR Formative Research (https://formative.jmir.org), 06.08.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Parental Challenges in Raising Preschoolers With Attention-Deficit/Hyperactivity Disorder in Mainland China's First-Tier Cities: Qualitative Study Using Framework Analysis

Shu-Cheng Chen[1], PhD; Chen-Wen Zhong[2], PhD; Han Li[1], MA; Xiaoqing Li[3], PhD; Hong-Wang Fung[1], PhD; Lijun Wang[4], PhD; Guo-Tao Wu[5], MA; Xin-Yue Jiang[6], BA; Wing-Fai Yeung[1], PhD

[1]School of Nursing, Hong Kong Polytechnic University, Hong Kong, China (Hong Kong)
[2]The Jockey Club School of Public Health and Primary Care, Chinese University of Hong Kong, Hong Kong, China (Hong Kong)
[3]School of Psychology, Shenzhen University, Shenzhen, China
[4]Department of Traditional Chinese Medicine, Binzhou Medical University, Yantai, China
[5]Faculty of Psychology, Southwest University, Chongqing, China
[6]Faculty of Education, Southwest University, Chongqing, China

Corresponding Author:
Guo-Tao Wu, MA
Faculty of Psychology
Southwest University
No. 2 Tiansheng Road
Beibei District
Chongqing, 400715
China
Phone: 86 15505080715
Email: gtgtgtwu@163.com

Abstract

Background: First-tier cities in mainland China present unique challenges for raising children with attention-deficit/hyperactivity disorder (ADHD) due to rapid urbanization, intense academic pressure, and distinct sociocultural dynamics. While existing research has documented ADHD parenting challenges across various contexts, limited attention has been paid to the preschool period in Chinese metropolitan settings, where early intervention is crucial yet complicated by traditional values and evolving health care systems.

Objective: This study aimed to explore the comprehensive challenges faced by parents raising preschoolers with ADHD in these metropolitan contexts.

Methods: We conducted web-based semistructured focus group interviews with 13 parents (n=12, 92% mothers and n=1, 8% fathers; aged between 32 and 44 years) of preschool children with ADHD (n=10, 77% boys and n=3, 23% girls; aged between 5 and 7 years) in first-tier cities in mainland China between March and July 2024. Participants were purposively recruited through the ADHD Mutual Support Alliance's platform if they had a preschool child with a formal ADHD diagnosis and lived in first-tier cities. Four interviews were conducted (3 focus groups with 4 participants each and 1 pilot individual interview). All interviews were moderated by a practitioner working with populations with ADHD following a semistructured interview guide developed through expert panel consultation. Interviews were audio-recorded and transcribed verbatim. Qualitative data were analyzed using framework analysis, using a theoretical framework approach, with dual-researcher coding and participant verification to ensure methodological rigor. Data were coded and organized manually using Microsoft Word, supported by NVivo software.

Results: Challenges emerged across six major themes: (1) individual psychological and behavioral challenges, including parental emotional burden, self-blame, and difficulties managing ADHD symptoms; (2) microsystem challenges in family and school environment challenges, particularly regarding behavioral management and teacher understanding; (3) mesosystem difficulties in family-school-hospital coordination, highlighted by recurring teacher complaints and inconsistent diagnostic standards; (4) exosystem barriers in work-family balance and health care access, including limited availability of specialized services and financial burdens; (5) macrosystem challenges from societal stigma and traditional educational values, manifesting as discrimination fears and academic pressure; and (6) chronosystem impacts from historical and social changes, notably the COVID-19 pandemic's influence and gaps in support systems compared to countries with comprehensive health care infrastructure.

XSL•FO
RenderX

Conclusions: Parents face multilayered challenges in raising preschoolers with ADHD, particularly regarding access to specialized health care services, navigation of educational systems, and use of culturally adapted interventions. These challenges are intensified by the unique urban context of intense academic pressure and rapid modernization. Future initiatives should focus on enhancing health care resource accessibility, developing culturally sensitive support programs, implementing systematic educational accommodations, and promoting broader societal awareness of ADHD in mainland China. These findings underscore the necessity for comprehensive ecological interventions that address challenges across all systemic levels while considering the unique characteristics of Chinese metropolitan environments.

(JMIR Form Res 2025;9:e74047) doi:10.2196/74047

KEYWORDS

attention-deficit/hyperactivity disorder; preschool children; parental challenges; first-tier cities; China; qualitative research

Introduction

Attention-Deficit/Hyperactivity Disorder

Attention-deficit/hyperactivity disorder (ADHD) is a complex neurodevelopmental disorder characterized by persistent patterns of inattention, hyperactivity, and impulsivity that interfere with functioning or development [1,2]. Globally, approximately 5% of children have ADHD [3], with the pooled worldwide prevalence rate among preschool children estimated at 5.5% (95% CI 3.3%-7.7%) [4]. In China, epidemiological studies indicate that the prevalence of ADHD among children and adolescents was 6.26% (95% CI 5.36%-7.22%) [5], which is comparable to global estimates. The symptoms of ADHD include difficulty sustaining attention, challenges in organizational skills, excessive motor activity, and impulsive behaviors that are developmentally inappropriate and present across multiple settings [6,7]. For preschool children, ADHD manifestations may include marked difficulties in following instructions, completing age-appropriate tasks, and regulating behaviors during structured activities, potentially compromising their early learning experiences and social skill development [8,9]. The impact of ADHD extends beyond the child who is affected, creating significant challenges within the family system, including increased parental stress, strained family relationships, and elevated financial burden associated with health care and educational support [10,11]. Conventional interventions typically encompass a multimodal approach, including behavioral therapy; parent training programs; educational accommodations; and, in some cases, pharmacological treatment, though medication use in preschoolers remains controversial [12,13]. Early identification and intervention are crucial. Untreated childhood ADHD symptoms can precipitate cascading adverse effects on academic, social, and emotional development, establishing patterns that persist into adulthood [14,15]. Timely intervention may mitigate secondary complications and foster adaptive coping strategies, thereby optimizing long-term outcomes [9,16]. Given that the preschool period represents a critical developmental window with heightened neural plasticity and emerging behavioral patterns, targeted support during this stage is paramount for shaping positive developmental trajectories.

Parental Challenges for Raising Children With ADHD

Previous studies in recent years have explored parental challenges across different cultural contexts when raising children with ADHD. A longitudinal study in the United States documented that mothers of children with ADHD aged between 4 and 12 years experienced significantly higher maternal stress and social isolation [17]. Similarly, in the United Kingdom, a qualitative interview study with 25 mothers of children aged between 5 and 11 revealed substantial program-related barriers with notably high decline rates (68%) [18]. Further evidence from a comparative diary study in the United States reinforced these challenges, tracking 27 mothers of children with ADHD aged between 7 and 11 years who reported markedly higher levels of anger (74.8% vs 51.4%) [19]. A systematic review across Australia, Canada, the United Kingdom, and the United States examining parents of school-aged children (4-18 years) confirmed that mothers consistently faced criticism from teachers and struggled with advocacy while experiencing significant emotional distress and social isolation [20]. In the context of developing nations, a qualitative interview study in South Africa with parents of school-aged children exposed severe burdens, including constant monitoring and emotional distress [21]. In Shantou, China, researchers found unique cultural pressures through interviews with 15 parents of children aged between 7 and 11 years [22]. In addition, a larger cross-sectional study in Sichuan province, China, examining 2497 pupils aged between 6 and 13 years quantified these challenges, revealing that parents of children with ADHD showed a 4.35 times higher risk of depression [23]. While extensive research has documented parental challenges, previous studies primarily focused on school-aged children. There remains a significant gap in understanding these challenges during the crucial preschool period. The preschool period represents a fundamentally different context for several reasons. First, ADHD manifestations in preschoolers are distinct, often characterized by more pronounced hyperactivity and impulsivity, with inattention being less apparent before formal academic demands emerge [8,9]. Second, diagnostic processes are considerably more complex for preschoolers, as distinguishing between age-appropriate developmental variations and clinically significant symptoms requires specialized expertise often unavailable in standard health care settings [12]. Third, intervention approaches differ substantially, with medication being more controversial in younger children and greater emphasis placed on behavioral and environmental modifications [13]. Fourth, parent-educator dynamics in preschool settings involve different expectations and communication patterns compared to primary schools, particularly in China's distinctive early education system where behavioral conformity is highly valued [24,25]. These fundamental differences necessitate

XSL•FO
RenderX

specific investigation of challenges faced by parents of preschoolers with ADHD, especially within China's unique first-tier urban environments.

Research Gap

In China, despite the growing attention to childhood ADHD in recent years, systematic investigation of preschoolers remains sparse, even in first-tier cities such as Shenzhen, Beijing, and Shanghai. These economically advanced metropolitan centers present unique challenges through intense academic pressure, rapid urbanization, and distinct sociocultural dynamics, which can be organized into several interconnected factors. First, the high-pressure educational culture creates a fundamental tension for children with ADHD. In the Chinese educational context, *academic burden* refers to the extraordinary pressure placed on children to excel academically from a very young age, characterized by long school hours, extensive homework, frequent high-stakes testing, and numerous extracurricular educational activities. This pressure stems from the deeply entrenched Confucian values that prioritize educational achievement as the primary pathway to success and family honor, combined with intense competition for limited spots in prestigious schools and universities in a population of 1.4 billion. In first-tier cities, this academic pressure is particularly intense due to higher concentrations of educational resources and more affluent, ambitious parents. While recent initiatives such as the "double reduction" policy aim to reduce academic burden [26], the prevalent "tiger parenting" approach continues to emphasize rigorous academic preparation [24]. Second, societal perceptions of behavioral problems create significant stigma and blame for parents [27,28]. Traditional Chinese societal expectations emphasize conformity, self-discipline, and quiet, attentive behavior in children—characteristics that directly conflict with ADHD symptoms. Parents of children with ADHD often face criticism for perceived parental failings, as behavioral issues are commonly attributed to inadequate discipline or moral education rather than neurodevelopmental conditions. Traditional Chinese values prioritizing academic achievement and conformity [29], coupled with fierce competition for early education placement in urban centers [25], significantly intensify parental stress. Moreover, parents have challenges related to societal stigma attributing behavioral issues to inadequate parenting [23]. Third, the intense urban work culture in first-tier cities further compounds these challenges. Parents face unique difficulties related to the demands of high-pressure urban careers [30], with long working hours and intense competitive environments limiting the time and energy available for consistent management of children's ADHD symptoms. Despite these challenges, parents in first-tier cities have increasingly gained access to professional ADHD information through improved health care resources, mental health education, and international medical perspectives, leading to rising diagnosis rates and earlier interventions [31]. These sociocultural factors create a uniquely challenging environment for parents raising children with ADHD in Chinese first-tier cities. Together, they generate a complex ecosystem where educational demands conflict with children's neurodevelopmental needs, societal stigma increases psychological burden on parents, and work pressures limit parents' ability to implement consistent support strategies. Understanding these interconnected factors is essential for developing effective support systems for these families.

To comprehensively understand these multifaceted challenges, the ecological systems theory by Bronfenbrenner [32] provides an ideal theoretical framework. This framework conceptualizes human development and experiences within nested environmental systems: the microsystem (immediate environment and direct interactions, such as family dynamics and parent-child relationships), mesosystem (interconnections between microsystems, such as parent-school collaboration), exosystem (indirect environmental influences, such as health care policies and community resources), macrosystem (broader sociocultural context, including cultural values and societal attitudes toward ADHD), and chronosystem (temporal dimensions and historical changes) [32,33]. This ecological perspective is particularly valuable for examining ADHD-related challenges in the Chinese context because it captures the complex interplay between traditional values, rapid social changes, and family dynamics that uniquely characterize urban Chinese environments. Furthermore, this framework allows us to systematically analyze how various environmental factors at different levels interact to shape parental experiences, moving beyond individual-focused analyses to examine interconnections across personal, institutional, and societal domains [33,34]. The complex challenges faced by parents raising children with ADHD in China's rapidly urbanizing and highly competitive metropolitan environments warrant an in-depth and comprehensive exploration. This study makes critical contributions beyond existing research in mainland China. While previous studies in Shantou examined school-aged children in smaller cities [22], and research in Sichuan quantitatively assessed parental depression [23], our study uniquely investigated challenges faced by parents of preschoolers within first-tier metropolitan environments. We used the ecological systems framework by Bronfenbrenner [32] to provide a comprehensive analysis of how various environmental systems interact to shape parental experiences, moving beyond the individual-focused analyses of previous studies to examine complex interactions across personal, familial, institutional, and societal domains within China's most developed urban contexts. Therefore, considering the limited research on parental experiences during the critical preschool period and the unique sociocultural context of Chinese metropolitan areas, we conducted this qualitative study to comprehensively explore and analyze the experiences and challenges encountered by parents raising preschool-aged children with ADHD in first-tier cities in mainland China, providing valuable insights for health care providers, educators, and policymakers in rapidly developing urban contexts.

Methods

Study Design

This qualitative study used semistructured focus group discussions to explore parents' experiences of raising preschool-aged children with ADHD. This method was chosen to facilitate in-depth discussions where participants could share

XSL•FO
RenderX

and build upon each other's perspectives. The COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist was used to guide the reporting of this study [35].

Participants

All participants, who were parents of preschool children with ADHD, participated in online focus group interviews. Participants were recruited from February to July 2024 through the ADHD Mutual Support Alliance's WeChat public account, a platform connecting families affected by ADHD across mainland China. The platform provides educational resources and peer support services for parents of children with ADHD. Purposive sampling was used to select participants who could provide rich and diverse perspectives on the research topic [36]. Parents were eligible to participate if they met the following inclusion criteria: (1) having a child aged between 3 and 7 years, (2) the child having received a formal ADHD diagnosis from a qualified health care professional, (3) current residence in first-tier cities (eg, Beijing, Shanghai, Guangzhou, and Shenzhen) in mainland China, (4) ability to communicate in Mandarin, and (5) capability to participate in online interviews via the Tencent Meeting platform. A recruitment advertisement was posted on the platform. Interested parents contacted the research team through the platform. A research assistant conducted online screening interviews with interested parents using a structured questionnaire to determine their eligibility. Those who met all the inclusion criteria were invited to participate in the online focus group interviews. Each focus group discussion was moderated by the principal researcher (GTW), with a research assistant (XYJ) serving as the assistant moderator. All focus group interviews were conducted online via the Tencent Meeting platform to ensure geographical accessibility and convenience for participants from different cities.

Data Collection

Data collection was conducted through focus group interviews following the practical guidelines by Krueger and Casey [37]. Before implementation, a semistructured interview guide containing 7 open-ended questions (Textbox 1) was developed and subsequently refined through consultations with an expert panel comprising 2 ADHD experts (a researcher with academic expertise in ADHD [GTW] and a practitioner working with populations with ADHD [WFY]) and a qualitative researcher (SCC). The interviews were facilitated by a moderator (GTW), the founder of the ADHD Mutual Support Alliance, who had established a strong rapport with the participants through previous community engagement activities. As an experienced advocate in the ADHD community, the moderator had undergone specialized training in qualitative research methodology under the guidance of an experienced qualitative researcher. The moderator's responsibilities encompassed establishing session protocols, facilitating group dynamics, and maintaining neutral engagement, while an assistant moderator (XYJ) documented field notes and managed the technical aspects of the Tencent Meeting platform. The moderator's ADHD expertise enhanced data collection through informed questioning, while having no previous relationship with participants ensured objectivity. Safeguards, including team-reviewed interview guides, regular debriefing sessions, and reflexive journaling, maintained methodological rigor. All sessions were audio-recorded with participants' written informed consent. To ensure interview quality in the online format, several measures were implemented. All participants received technical orientation before interviews, including guidance on audio settings and connectivity requirements. The assistant moderator monitored technical aspects throughout, addressing disruptions immediately. Participants were encouraged to join from quiet, private locations to minimize distractions. To compensate for reduced nonverbal cues, the moderator used specific online facilitation techniques, including regular verbal check-ins, directed questions to ensure equal participation, and periodic summaries to confirm understanding. All sessions were recorded with redundant audio capture to prevent data loss.

Textbox 1. Semistructured interview guide for focus group discussions with parents of preschool children with attention-deficit/hyperactivity disorder (ADHD) in first-tier Chinese cities.

- "Could you describe your experience of raising a child with ADHD? Please share your journey."
- "How do you usually communicate with your child? Do you encounter any difficulties in communication? If so, what are they?"
- "Do you find it challenging to obtain the guidance and support you need? If yes, what are these challenges?"
- "In your opinion, what types of parental care and support are needed for preschool children with ADHD?"
- "What successful experiences have you had in accompanying and guiding your child? (including treatment approaches, etc)"
- "What unsuccessful experiences have you had in accompanying and guiding your child? (including treatment approaches, etc)"
- "What kind of support do you expect from society (schools, hospitals, other organizations, etc) to better help you raise your child?"

Data Analysis

All interviews were audio-recorded and subsequently transcribed verbatim in Mandarin before analysis. Demographic data were analyzed using descriptive statistics. We conducted framework analysis of the transcripts, guided by ecological systems theory by Bronfenbrenner [32,33], to examine complex social phenomena in their environment. The analysis process followed several systematic steps [38,39]:

1. Familiarization: 2 researchers (GTW and XYJ) independently reviewed all transcripts to achieve data immersion and develop preliminary insights.

XSL•FO
RenderX

2. Identifying the theoretical framework: ecological systems theory by Bronfenbrenner [32] was selected to guide the coding process.
3. Indexing: the researchers independently coded the transcripts, mapping data onto different ecological levels.
4. Charting: data were reorganized into framework matrices, with ecological levels as columns and cases as rows.
5. Mapping and interpretation: themes were identified within and across ecological levels to understand the interconnections between different systemic influences on parental experiences.

To enhance analytic rigor, the 2 researchers conducted independent coding and regularly compared their analyses to resolve discrepancies through discussion. When consensus could not be reached, the principal investigator (GTW) was consulted. The final analytical framework and emerging themes were reviewed by all team members to ensure interpretative validity. The primary processes of data coding and organization were conducted manually in Microsoft Word, with NVivo (Lumivero) software used to facilitate and cross-check coding [40,41].

Trustworthiness

To establish methodological rigor, we implemented multiple strategies addressing the 4 dimensions of trustworthiness outlined by Lincoln and Guba [42] and further applied in health care research by Forero et al [43]. The study's credibility was reinforced through a pilot interview that informed subsequent protocol refinement, coupled with regular investigator meetings and dual-researcher coding processes. To ensure dependability, we implemented participant verification of interview transcripts and interpretations while maintaining detailed documentation of our methodological decisions. Participant verification involved email member-checking; participants reviewed interview summaries and gave feedback. The analytical process was strengthened by having a second investigator (SCC) and the principal investigator (GTW) independently verify the coding structure. Confirmability was achieved through participant feedback on interpretations and conscious examination of contradictory cases, while transferability was supported by detailed contextual documentation and theoretical sampling until data saturation [44], with the framework by Bronfenbrenner [32] providing a theoretical foundation for broader application.

Ethical Considerations

Ethics approval was obtained from the Human Research Committee of the School of Psychology, Shenzhen University (SZU_PSY_2023_068). Written informed consent was obtained from all participants before their participation in the study. To ensure confidentiality, participants were assigned unique identification codes, and all study-related documents and transcripts were deidentified. Audio recordings were securely deleted following transcription completion. To ensure privacy and create a natural, safe atmosphere, participants were instructed to join the interviews from private spaces where no other individuals were present and to disable their video functions during the interviews. To acknowledge their contribution, participants were granted access to a professional development course focused on mindfulness intervention for ADHD, delivered by an expert psychologist specializing in ADHD management.

Results

General Characteristics of Data

Three focus group interviews were conducted between March and July 2024 with 13 parents of children with ADHD. Each group included 4 to 5 participants. An additional individual pilot interview was conducted to refine the interview guide and procedures, but its data were not included in the final analysis. Four individuals declined participation due to scheduling conflicts.

Sample Profile

Thirteen parents (n=12, 92% mothers and n=1, 8% father) participated in the semistructured interviews, conducted across 3 waves following one initial pilot interview. The parents' ages ranged from 32 to 44 years. All participants had a college education or higher, and most (8/13, 62%) were employed; one mother was a housewife. Their children (n=10, 77% boys and n=3, 23% girls) were aged between 5 and 7 years. ADHD symptoms were first noticed at an average age of 3.2 (SD 2.0) years, and the average duration of treatment was 5.6 (SD 9.0) months. Families reported using a variety of interventions, including medication, cognitive behavioral therapy, and other approaches, either alone or in combination. Detailed demographic characteristics are presented in Table 1.

XSL•FO
RenderX

Table 1. Demographic characteristics of parents interviewed about their children with attention-deficit/hyperactivity disorder (ADHD) in first-tier Chinese cities (N=13).

Characteristics	All	Pilot (n=1)	Wave 1 (n=4)	Wave 2 (n=4)	Wave 3 (n=4)
Child					
Age (y), mean (SD; range)	6.3 (0.9; 5-7)	5 (0)	6.5 (0.6; 6-7)	6 (1.2; 5-7)	6.8 (0.5; 6-7)
Gender, n (%)					
Man	10 (77)	1 (100)	2 (50)	3 (75)	4 (100)
Woman	3 (23)	0 (0)	2 (50)	1 (25)	0 (0)
Age at ADHD symptoms (y), mean (SD)	3.2 (2.0)	3 (0)	5.3 (1.7)	2.3 (1.0)	2.3 (1.9)
Treatment duration (mo), mean (SD)	5.6 (9.0)	0 (0)	4 (5.4)	8.3 (11.3)	6 (12)
Parents					
Age (y), mean (SD; range)	37 (5.0; 33-44)	37 (0)	37.8 (7.4; 33-44)	35.3 (3.6; 32-37)	38 (5; 33-40)
Parental role, n (%)					
Mother	12 (92)	0 (0)	1 (25)	0 (0)	0 (0)
Father	1 (8)	1 (100)	3 (75)	4 (100)	4 (100)
Level of education, n (%)					
College or above	13 (100)	1 (100)	4 (100)	4 (100)	4 (100)
Career, n (%)					
Employees of public institutions	4 (31)	1 (100)	1 (25)	2 (50)	0 (0)
Employees of a company	4 (31)	0 (0)	2 (50)	1 (25)	1 (25)
Freelancers	1 (8)	0 (0)	0 (0)	0 (0)	1 (25)
Self-employed	1 (8)	0 (0)	0 (0)	0 (0)	1 (25)
Housewife	1 (8)	0 (0)	0 (0)	1 (25)	0 (0)
Others	2 (15)	0 (0)	1 (25)	0 (0)	1 (25)
Monthly family income (¥), n (%)					
≤10,000 (US $1400)	1 (8)	0 (0)	1 (25)	0 (0)	0 (0)
10,001-29,999 (US $1401-4200)	7 (54)	1 (100)	2 (50)	2 (50)	2 (50)
30,000-49,999 (US $4201-7000)	2 (15)	0 (0)	1 (25)	0 (0)	1 (25)
≥50,000 (US $7001)	3 (23)	0 (0)	0 (0)	2 (50)	1 (25)

Major Themes

Overview

Through a thorough analysis of the transcripts and participant quotations, six themes emerged aligned with the ecological framework by Bronfenbrenner [32]: (1) individual level, encompassing psychological factors and behavioral factors; (2) microsystem, involving family interactions and school environment challenges; (3) mesosystem, including family-school interactions and family-hospital interactions; (4) exosystem, covering work-family interactions and health care access issues; (5) macrosystem, reflecting societal policy and societal perceptions; and (6) chronosystem, capturing historical and social changes. Figure 1 [32] illustrates the ecological systems framework by Bronfenbrenner [32] as applied to parental challenges. These themes, their respective subthemes, and supporting participant quotations are detailed in Multimedia Appendix 1.

XSL•FO
RenderX

Figure 1. Bronfenbrenner's ecological systems framework applied to parental challenges in raising pre-schoolers with attention-deficit/hyperactivity disorder (ADHD) in first-tier cities of mainland China. This figure depicts the nested environmental systems influencing preschool children with ADHD and their families in Chinese first-tier cities. Based on Bronfenbrenner's ecological systems theory, it illustrates six interconnected levels: individual (psychological and behavioral factors), microsystem (immediate family and school environments), mesosystem (family-school-hospital interactions), exosystem (work-family balance and health care access), macrosystem (societal policies and perceptions), and chronosystem (historical and social changes). Icons represent themes from focus group interviews (N=13) conducted in 2024, highlighting challenges unique to China's urban context of intense academic pressure and evolving health care systems.

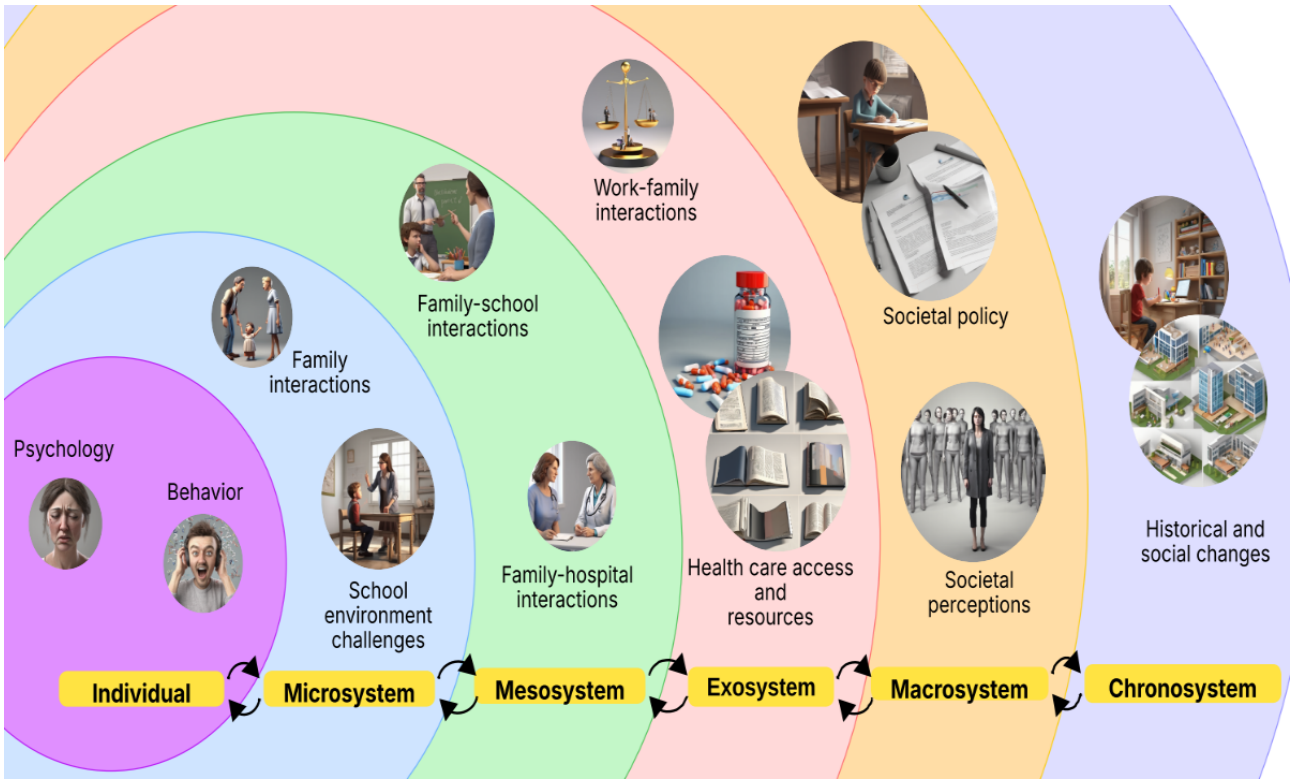

Individual

Overview

At the individual level, parents of preschool children with ADHD experienced significant psychological and behavioral challenges. Parents experienced persistent psychological strain, manifesting as emotional distress, internalized self-blame, and anxiety regarding their parenting efficacy. The demands of managing their child's ADHD symptoms contributed to feelings of exhaustion and ongoing concern for their child's future development. Some parents also described challenges related to their own behavioral characteristics, including the impact of potential parental ADHD traits on family functioning. Concerns about children's development extended to worries about social adaptation, academic readiness, and emotional regulation, intensifying parental stress.

Psychological Factors

The psychological factors manifested in 2 primary ways. First, parents reported intense emotional challenges, including feelings of anger, frustration, self-blame, depression, and anxiety. One mother described her emotional struggle as follows:

> *Sometimes I can't control myself. When I'm in a bad mood, I feel really guilty. When my child misbehaves, I might not handle it well, and then I blame myself. For instance, when I punish him by spanking, I feel terrible inside afterwards.* [Participant 09]

Second, parents faced substantial parenting burdens, particularly regarding time demands and physical and emotional energy depletion. One parent expressed this overwhelming burden as follows:

> *I wish I could be like Nezha with three heads and six arms, because I feel I'm already doing everything I possibly can. I've given her all my available time when she's at home, but there are still things I can't manage. For example, I can't cook and play with her at the same time, right?* [Participant 04]

Behavioral Factors

In terms of behavioral factors, parents faced 2 main challenges. Parents usually dealt with their own behavioral issues, including ADHD symptoms and oppositional behavior. One parent shared their personal struggle as follows:

> *I'm diagnosed with ADHD myself...My personality is confrontational, hyperactive, and impulsive—I have the combined type. All the typical emotional issues, including depression—I've got them all.* [Participant 12]

In addition, parents expressed significant concerns about their children's development, particularly regarding social skills, academic performance, and emotional regulation. One parent articulated as follows:

> *I worry that the more time he spends immersed in (animated videos), the less he'll want to do things*

XSL•FO
RenderX

> *that require calm thinking or interact with others...I'm also concerned about his emotional issues. When he's unhappy about something, he might just walk away without a word and completely give up on his tasks. I worry that he'll have poor self-control when he grows up.* [Participant 09]

Microsystem

Overview

At the microsystem level, challenges in 2 primary domains emerged from the analysis: family interactions and school environment challenges. Within the immediate family context, parents reported difficulties in maintaining consistent behavioral management strategies and achieving family cohesion. Conflicting approaches among caregivers and communication barriers within the family frequently led to strained relationships and limited support. In the preschool environment, parents identified insufficient teacher awareness and understanding of ADHD, resulting in limited accommodation of their children's needs and increased behavioral challenges at school.

Family Interactions

Within family interactions, parents identified significant challenges in managing their children's behaviors and maintaining family harmony. Parents acknowledged their difficulties in providing consistent and timely intervention for their children's problematic behaviors, such as procrastination, distraction, and impulsivity. One parent reflected on how their permissive approach at home translated into behavioral issues at school:

> *Take singing in the bathroom—when he sang or yelled loudly at home, I didn't correct him right away. So at school, he thinks it's fine to yell or sing in the bathroom...And about being disorganized—I never insisted on him keeping things organized at home, so naturally at school he's careless with his stuff and often forgets his homework.* [Participant 02]

Family conflicts emerged as another significant challenge, characterized by negative emotions, communication barriers, and inconsistent parenting approaches among family members. The complexity of multicaregiver dynamics was highlighted by 1 parent:

> *Since three adults (me, my husband, and my mom) are raising the child together, there are lots of differing opinions and we all have short tempers, so the child endures a lot of emotional stress.* [Participant 11]

School Environment Challenges

In terms of school environment challenges, parents reported challenges stemming from educators' limited understanding of ADHD. This lack of awareness affected how teachers perceived and responded to their children's behaviors. One parent noted the following:

> *I think teachers—from preschool through elementary and middle school—they don't really understand this condition (ADHD). They just think the child is strange.* [Participant 03]

Mesosystem

Overview

At the mesosystem level, the intersection of family, school, and health care systems generated additional stress for parents. Interactions with teachers were often characterized by recurring reports of problematic behavior and challenges in constructive communication. Parents also highlighted difficulties coordinating with health care professionals due to inconsistent diagnostic criteria and fragmented service provision, which contributed to confusion regarding the appropriate course of action and uncertainty about their child's prognosis and intervention pathways.

Family-School Interactions

In the context of family-school interactions, parents experienced significant stress from recurring teacher complaints about their children's behavior. The emotional toll of these interactions was vividly described by 1 parent as follows:

> *His kindergarten teacher frequently reported that he wouldn't follow instructions. She had to remind him about 5-6 times during one class, which was really excessive. And after each reminder, it wouldn't even last two or three minutes before he was off-task again.* [Participant 09]

Furthermore, parents struggled to effectively manage their children's negative attitudes following teacher criticism. One parent's reflective account highlighted the counterproductive nature of their initial approach:

> *When teachers reported issues to me, I immediately scolded him: 'How could you do this? Why did you make this mistake again after we've discussed it?' Looking back, my approach was wrong. He was already criticized at school, then came home to face more criticism from me siding with the teacher. This confrontational communication eventually led him to stop listening to me.* [Participant 08]

Family-Hospital Interactions

Regarding family-hospital interactions, parents encountered challenges on 2 fronts. First, many parents expressed doubts about the reliability or accuracy of the ADHD diagnosis due to inconsistent diagnostic procedures and standards across different hospitals and clinicians. One parent observed the following:

> *When we later saw another doctor, they didn't conduct an IQ test. These tests are now administered by nurses across different hospitals, and the assessment quality varies significantly between experienced and inexperienced staff.* [Participant 08]

Second, parents frequently reported feeling lost and overwhelmed when navigating fragmented and complex health care pathways, unsure about next steps for intervention and support. One parent noted the following:

> *In the process of raising him, I would start reading books about ADHD and sign up for online courses,*

XSL•FO
RenderX

> *but I often abandoned these efforts halfway due to work commitments and other reasons. This pattern kept repeating itself in a continuous cycle.* [Participant 14]

These forms of confusion and uncertainty led to increased stress and hesitancy in decision-making about their child's care.

Exosystem

Overview

External circumstances, particularly related to work and health care access, further complicated the parenting experience. Parents struggled to balance demanding professional responsibilities with the intensive needs of their children, frequently facing constraints in time, energy, and resources. Barriers to accessing specialized health care services—including high costs, limited availability, and lack of individualized or culturally relevant interventions—were repeatedly emphasized. These factors collectively exacerbated the burden on families.

Work-Family Interactions

In the domain of work-family interactions, parents faced substantial challenges in balancing professional responsibilities with their children's needs. One parent described how work commitments affected early childcare:

> *After my child turned one year, I worked full-time. Until age three, the grandmother took care of them while I rarely came home, usually caught up in my company work. I was somewhat lacking in childcare during this period.* [Participant 12]

Some parents made significant career sacrifices to accommodate their children's needs, as illustrated by another parent's experience:

> *Due to my child's ADHD, I quit my job last year. I decided to prioritize helping my child before focusing on my career, believing that career success means little if my child isn't doing well. This decision had a significant impact on my life.* [Participant 08]

Health Care Access and Resources

The health care access and resources domain presented multiple challenges for families. The financial burden of ADHD interventions emerged as a significant concern; 1 parent detailed the following:

> *Currently, we're doing sensory integration training and medication—all self-funded as insurance doesn't cover these. The training is extremely expensive at nearly 300 yuan per session, 2-3 times weekly. With desensitization therapy on top of that, the financial burden is heavy.* [Participant 10]

Access to professional medical resources proved particularly challenging, especially at prestigious institutions. A parent described their frustration as follows:

> *Getting an appointment at Peking University's 6th Hospital (China's most authoritative hospital for child mental health) for pediatric services is incredibly difficult. The slots disappear within seconds—it's practically impossible to get one. I haven't succeeded even once. I wonder why these pediatric appointments are in such high demand and who manages to get them? It's puzzling how they vanish instantly.* [Participant 03]

Parents also highlighted the lack of personalized medical services. One parent, who had personal experience with ADHD, observed the following:

> *I have ADHD myself, and I've noticed many differences between my symptoms and my son's. The biggest challenge is getting personalized treatment. Take West China Hospital in Chengdu (one of the best hospital in Southwest China) for example—there are only one or two truly respected specialists. Not everyone responds well to the same treatment approach. Even nationwide, I feel there are very few doctors who can provide truly individualized treatment.* [Participant 17]

In addition, parents struggled to find culturally relevant parenting resources. One parent explained the following:

> *During the kindergarten stage, we want to provide individualized parenting at home. While there are many parenting books on the market, most are written by and for Westerners, with very few by Chinese authors. Given the unique Chinese educational environment, many Western parenting methods simply aren't applicable here.* [Participant 04]

Macrosystem

Overview

External circumstances, particularly related to work and health care access, further complicated the parenting experience. Parents struggled to balance demanding professional responsibilities with the intensive needs of their children, frequently facing constraints in time, energy, and resources. Barriers to accessing specialized health care services—including high costs, limited availability, and lack of individualized or culturally relevant interventions—were repeatedly emphasized. These factors collectively exacerbated the burden on families.

Societal Policy

Regarding societal policy, parents expressed deep frustration about the lack of comprehensive support systems. One parent articulated their pessimism about systemic change:

> *I couldn't see significant progress happening in China over the next five to ten years. While small groups are working hard to push for change, their progress is inevitably slow. It's not that we aren't trying, but real change requires support from higher policy levels. Even when research institutions achieve results, without sufficient backing, it's all talk. This situation leaves me feeling somewhat helpless.* [Participant 12]

Societal Perceptions

In terms of societal perceptions, parents encountered multiple challenges related to stigmatization and traditional educational

XSL•FO
RenderX

values. The fear of social discrimination deeply affected parents' willingness to disclose their children's condition; 1 parent revealed the following:

> *It's hard for me to admit to other parents that my child has ADHD. I'm afraid they'll go home and tell their children to stay away from my child, warning them "be careful, he might hit you." They'll become defensive and prejudiced. This social stigma makes me feel powerless.* [Participant 16]

The broader issue of societal tolerance was highlighted by another parent:

> *These days, it's not just parents of ADHD or autistic children—once you become a mother, you're automatically downgraded to a second-class citizen. Society's tolerance needs improvement, not just for special needs children, but for all children in general.* [Participant 11]

The severity of social stigma was further illustrated through a disturbing example:

> *We saw news about a child with ADHD whose entire class—over 40 parents—joined forces to demand their transfer to another school. If this happened to my child, it would feel terribly unfair. Yet, I can somehow understand those parents' perspective too.* [Participant 10]

The pressure of traditional educational values created additional challenges. One parent emotionally described their fears about their child's academic future:

> *My child barely passes each subject. In Beijing, this likely means they won't get into high school. Their poor short-term memory means they'll gradually fall further behind their peers. Yet they still have to attend school with children their age. It's terrifying...unless they quit school entirely, but...(sobbing).* [Participant 03]

The systemic constraints on teachers were also acknowledged:

> *Teachers have their own evaluation system. When they're being assessed based on academic performance, how can they spare extra energy for children with special needs?* [Participant 17]

Chronosystem

Overview

The chronosystem-level analysis highlighted how temporal changes, both societal and historical, influenced the experiences of families managing ADHD. While our study design is cross-sectional, parents' narratives included important reflections on temporal and historical factors that influence their experiences. These chronosystem-related findings represent parents' retrospective perceptions and comparisons rather than longitudinal observations of historical changes. Temporal and historical factors shaped parental experiences, with recent societal changes and events, such as the COVID-19 pandemic, influencing family dynamics and access to support. Parents frequently compared the current environment with perceived support systems in other countries, highlighting persistent gaps in service provision and understanding of ADHD over time. While these reflections were based on parental perceptions rather than systematic longitudinal analysis, they nonetheless provided important context for understanding the evolving challenges faced by families.

Historical and Social Changes

The COVID-19 pandemic emerged as a significant temporal factor that exacerbated ADHD symptoms in children. One parent reflected on how pandemic-related restrictions potentially contributed to their child's condition:

> *His ADHD might have developed during the pandemic due to lack of exercise and my frequent business trips—sometimes I'd get stuck somewhere due to lockdowns. I wasn't there enough for him...At first, I just thought he was being naughty, but after the pandemic, when he returned to kindergarten, he started ignoring teachers' instructions.* [Participant 02]

The ongoing evolution of ADHD support systems in China represented another crucial chronosystemic element. Parents often compared their experiences with more developed support systems in other countries, highlighting the current gaps in China's approach. One parent drew a compelling comparison:

> *In some countries, once a child is diagnosed with ADHD, they're provided with a service dog at school. When the dog sits quietly, the child sits too, and others don't view them strangely. Our country still has a way to go in this regard—there's a gap in understanding. Teachers might not even understand ADHD and try to enforce standard expectations, causing more pressure and psychological trauma for these children.* [Participant 09]

This sentiment was echoed by another parent who succinctly captured the developmental gap in support systems:

> *There's still a significant gap between us and some developed countries. Our supportive systems are not adequate yet.* [Participant 07]

While parents perceived the COVID-19 pandemic and international disparities as influential, our cross-sectional design precludes causal attribution. These accounts should be interpreted as illustrative examples of chronosystem influences rather than systematically analyzed or central findings.

Discussion

Principal Findings

This is the first study to comprehensively examine the challenges faced by parents raising preschoolers with ADHD in Chinese first-tier cities, revealing complex interconnections across multiple systemic levels. Our findings demonstrate that parental experiences are shaped by intricate interactions between individual psychological burdens, immediate family dynamics, institutional relationships, societal structures, and temporal changes. At the individual level, parents experienced significant emotional challenges and behavioral management difficulties,

XSL•FO
RenderX

compounded by their own potential ADHD symptoms. The microsystem analysis revealed substantial strain in both family relationships and school environment challenges, while the mesosystem highlighted particular challenges in navigating between educational and health care institutions. Within the exosystem, parents faced significant work-life balance challenges and encountered substantial barriers in accessing appropriate health care resources, including financial burdens and limited availability of specialized services. The macrosystem analysis exposed how traditional Chinese educational values and societal stigma profoundly impact parenting experiences, and our analysis of chronosystem elements revealed how parents perceived recent events, particularly the COVID-19 pandemic, and international disparities in ADHD support systems as influential factors. However, these references were anecdotal and based on parental perceptions. While parents perceived the COVID-19 pandemic and international disparities as influential, our cross-sectional design precludes causal attribution. Therefore, these findings should be interpreted as illustrative examples of chronosystem influences, rather than systematically analyzed or central conclusions. This study specifically examined these challenges within the context of Chinese first-tier cities, where rapid urbanization, intense academic pressure, and evolving health care systems create a unique environment for families managing preschool ADHD. The application of the ecological framework by Bronfenbrenner [32] provided valuable insights into how these various systemic levels interact and influence parental experiences, highlighting the need for comprehensive, multilevel support systems.

This study's application of the ecological framework by Bronfenbrenner [32] to examine ADHD-related challenges in Chinese first-tier cities advances beyond previous research. Our ecological analysis revealed complex influences between multiple system levels that shape parental experiences. Particularly notable is our finding regarding the paradoxical relationship between improved diagnostic capabilities and inadequate intervention resources in first-tier cities. Our focus on preschoolers revealed distinctive early-stage challenges, including early identification difficulties, accessing appropriate interventions, and managing the critical transition to educational settings during a foundational developmental period.

Comparisons With Previous Studies

Several findings from our study align with previous research in various cultural contexts. The significant emotional burden and psychological distress experienced by parents, particularly mothers, echo findings from Western studies. For instance, our participants' reports of anxiety, depression, and self-blame parallel those documented in a US longitudinal study where mothers of children with ADHD experienced significantly higher maternal stress and social isolation [10]. The challenges in family-school interactions and recurring teacher complaints identified in our study reflect similar findings from a UK qualitative study, where mothers faced substantial criticism from educators [45]. In addition, our findings regarding social stigma and discrimination align with research from multiple cultural contexts, where parents reported severe emotional distress and social isolation [46]. The financial burden and health care access difficulties reported by our participants mirror findings from previous Chinese studies, particularly research that highlighted unique cultural pressures on parents [47]. The impact of work-family conflicts on parenting quality corresponds with findings from a systematic review across Western countries, where mothers consistently struggled to balance career demands with caregiving responsibilities [48]. Furthermore, our participants' experiences with insufficient institutional support systems align with findings from other Asian contexts, where parents of children with ADHD showed significantly higher risks of depression due to inadequate support structures [49]. These consistencies across different cultural contexts suggest some universal challenges in raising children with ADHD, despite varying sociocultural backgrounds.

Several of our findings diverge notably from previous research literature. While Western studies have consistently reported teachers as primary sources of support and key collaborators in ADHD management [50], our participants encountered significant challenges with school personnel, characterized by frequent conflicts, misunderstandings, and limited support. Teachers in our study were often described as resistant to accommodating these students' needs, possibly reflecting the intense academic pressures and rigid behavioral expectations in the Chinese education system. Perhaps most strikingly, while previous studies have emphasized the crucial role of extended family support in Chinese contexts [51], our participants reported that extended family members often became sources of additional stress rather than support, primarily due to conflicting views about ADHD management and disagreements over treatment approaches. This difference in teacher support may be attributed to several factors unique to the Chinese educational context: (1) the intense performance-oriented evaluation system that prioritizes academic achievement over behavioral accommodation, (2) significant knowledge gaps in ADHD training within Chinese teacher education programs, (3) larger class sizes in Chinese schools (typically 40-50 students) that limit individualized attention, and (4) the cultural emphasis on conformity and discipline that may frame ADHD behaviors as willful misconduct rather than neurodevelopmental differences. These structural and cultural factors create an environment where teachers face institutional disincentives for accommodating children with ADHD, despite their potential personal willingness to support these students. This may relate to the fact that preschool teachers in China are generally trained to emphasize play-based learning and social-emotional development, whereas primary school teachers are oriented toward academic achievement, standardized curricula, and classroom discipline. These differing expectations may shape how teachers respond to children with ADHD [52].

Our findings reveal a complex interplay of systemic challenges in China's ADHD health care system. First, we uncovered an unprecedented severity in accessing specialized ADHD services that substantially exceeds previously documented health care barriers. While earlier studies noted general appointment difficulties [53], our findings demonstrate an extraordinary level of access constraints, with participants reporting that securing appointments at leading institutions has become virtually impossible—slots are filled within seconds of release. A second critical dimension emerges in the paradoxical relationship

XSL•FO
RenderX

between diagnostic capabilities and treatment delivery. Despite significant improvements in ADHD awareness and diagnostic rates in first-tier cities [54], there exists a substantial gap in individualized, culturally appropriate interventions. Our analysis reveals that while practitioners can readily identify ADHD cases, their therapeutic approaches lack necessary customization, predominantly relying on direct translations of Western materials and standardized treatment regimens that fail to account for China's unique sociocultural context [55]. This standardization extends to both clinical interventions and parent training resources, which inadequately address Chinese family dynamics and cultural values around child development [56]. The situation is further compounded by China's intensive, exam-oriented education system, where academic performance metrics overshadow behavioral support needs, creating what participants described as an "impossible balance" between managing symptoms and meeting rigorous academic standards. The absence of references to alternative educational options in our data reflects the significant infrastructural limitations even in mainland China's most developed urban centers [57,58]. Despite the advanced economic development of first-tier cities, specialized educational pathways for preschoolers with ADHD remain virtually nonexistent within the mainstream system [59]. Unlike some Western contexts [60], mainland China's educational approach offers minimal formalized alternatives at the preschool level, with most children attending standard kindergartens regardless of neurodevelopmental needs [52,61]. The few specialized institutions that exist primarily operate in the private sector, are often financially inaccessible to average families, and vary considerably in quality due to limited regulatory standards [62]. This educational landscape, when combined with the diagnostic inconsistencies noted by participants, creates a particularly challenging situation where parents may obtain a diagnosis but find no specialized educational pathways designed to accommodate their children's needs within the conventional system.

Implications

Our findings have critical implications for multiple stakeholders in first-tier Chinese cities' ADHD ecosystem. For health care administrators in major metropolitan hospitals, these results highlight the necessity of optimizing current ADHD specialist resources, particularly given the high diagnosed patient volumes in urban areas. Mental health professionals and researchers in leading institutions should prioritize developing standardized yet flexible intervention protocols that address the unique challenges faced by urban Chinese families, such as high academic pressure and limited parent-child interaction time. For public health educators and community organizations in these developed urban areas, efforts should focus on enhancing workplace- and school-based ADHD awareness programs, particularly targeting high-achieving professionals and educators who may overlook ADHD symptoms due to academic or career success. Educational administrators in key urban schools should develop practical accommodation policies that balance the rigorous academic standards of first-tier cities with the support needs of students with ADHD. Importantly, such systematic educational accommodations should be designed to align with the objectives of China's "*double reduction*" policy, supporting the reduction of students' academic burden while meeting the diverse needs of learners with ADHD. In addition, medical education institutions in these metropolitan areas should expand their ADHD-specific training programs to meet the growing demand for qualified professionals in urban settings.

The notable absence of school principals in parents' accounts reflects the distinctive hierarchical structure of Chinese educational institutions, where administrative leadership remains relatively distant from daily parent interactions [63,64]. This contrasts with Western educational models, where principals may play more visible roles in special education support systems [65]. At the preschool level in Chinese first-tier cities, our findings suggest that teachers function as institutional gatekeepers, mediating between families and higher administrative levels [66]. This hierarchical distance may inadvertently limit opportunities for systemic policy implementation supporting children with ADHD, as principals, who typically possess greater authority to implement accommodations [67], remain removed from direct awareness of these children's needs. Compared to Western contexts, where collaborative models between teachers and mental health professionals are more common, Chinese teachers often exhibit resistance, partly due to systemic barriers, such as comparatively larger class sizes [68]. Therefore, future research should specifically examine the role of school leadership in supporting preschool children with ADHD through targeted interviews with both principals and parents who have had significant administrative interactions.

Limitations

Several limitations of this study warrant consideration. First, the participant demographics pose potential generalizability limitations, as our sample (12/13, 92% mothers and 1/13, 8% father) was overwhelmingly composed of highly educated mothers (100% college educated). This homogeneity means that our findings may not fully represent the experiences of families from lower socioeconomic backgrounds or those of fathers in first-tier Chinese cities. While this reflects broader patterns in these urban centers, where highly educated parents often demonstrate greater awareness of developmental concerns [69] and maintain higher academic expectations for their children [70], and are more proactive in developmental monitoring and information-seeking [71], families with less education or fewer resources may encounter different or additional barriers that are not captured in our study. Therefore, the transferability of our findings is limited. Future studies should purposefully recruit more fathers and families from diverse socioeconomic backgrounds to better capture the full range of parental experiences in urban China. Second, while our research interview guide included questions about parenting strategies, this manuscript deliberately focused on challenges, as understanding these fundamental obstacles is crucial for developing effective support systems and interventions. The valuable insights regarding parenting strategies and coping mechanisms, though important, will be reported in the future to allow for comprehensive exploration of each aspect. In addition, while our results include chronosystem-related findings, our cross-sectional design limits conclusions about historical changes over time. These findings represent parents'

XSL•FO
RenderX

retrospective perceptions and comparisons rather than systematic longitudinal data. Finally, our study relied solely on parental perspectives, and future research would benefit from incorporating multiple stakeholders' viewpoints, including children, teachers, and health care providers, to provide a more holistic understanding of the ADHD care ecosystem in urban China.

Conclusions

Our findings illuminate the multilayered challenges confronting parents of preschoolers with ADHD in first-tier Chinese cities, encompassing individual psychological burdens, health care access barriers, and sociocultural constraints. Parents face substantial obstacles in accessing specialized services, navigating educational stigma, and using culturally appropriate interventions, challenges intensified by the unique urban context of intense academic pressure and rapid modernization. These challenges underscore the necessity for comprehensive ecological interventions, encompassing enhanced health care accessibility, culturally sensitive support programs, systematic educational accommodations, and broader societal awareness initiatives for ADHD.

Acknowledgments

The authors acknowledge all participants who contributed their time and insights to this study and the use of Image_Creators, an artificial intelligence tool, for assistance in creating the figure.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Thematic coding framework of parental challenges in raising preschool children with attention-deficit/hyperactivity disorder in first-tier Chinese cities.
[DOCX File , 26 KB - formative_v9i1e74047_app1.docx]

References

1. American Psychiatric Association, DSM-5 Task Force. Diagnostic and Statistical Manual of Mental Disorders: DSM-5™. Washington, DC: American Psychiatric Publishing; 2013.
2. Barkley RA. Attention-Deficit Hyperactivity Disorder: A Handbook for Diagnosis and Treatment, Fourth Edition. New York, NY: Guilford Publications; 2015.
3. Polanczyk GV, Salum GA, Sugaya LS, Caye A, Rohde LA. Annual research review: a meta-analysis of the worldwide prevalence of mental disorders in children and adolescents. J Child Psychol Psychiatry 2015 Mar 03;56(3):345-365. [doi: 10.1111/jcpp.12381] [Medline: 25649325]
4. Willcutt EG. The prevalence of DSM-IV attention-deficit/hyperactivity disorder: a meta-analytic review. Neurotherapeutics 2012 Jul;9(3):490-499 [FREE Full text] [doi: 10.1007/s13311-012-0135-8] [Medline: 22976615]
5. Wang T, Liu K, Li Z, Xu Y, Liu Y, Shi W, et al. Prevalence of attention deficit/hyperactivity disorder among children and adolescents in China: a systematic review and meta-analysis. BMC Psychiatry 2017 Jan 19;17(1):32 [FREE Full text] [doi: 10.1186/s12888-016-1187-9] [Medline: 28103833]
6. Faraone SV, Banaschewski T, Coghill D, Zheng Y, Biederman J, Bellgrove MA, et al. The World Federation of ADHD International Consensus statement: 208 evidence-based conclusions about the disorder. Neurosci Biobehav Rev 2021 Sep;128:789-818 [FREE Full text] [doi: 10.1016/j.neubiorev.2021.01.022] [Medline: 33549739]
7. Thapar A, Cooper M. Attention deficit hyperactivity disorder. The Lancet 2016 Mar;387(10024):1240-1250. [doi: 10.1016/s0140-6736(15)00238-x]
8. DuPaul GJ. School-based interventions for students with attention deficit hyperactivity disorder: current status and future directions. Sch Psychol Rev 2019 Dec 27;36(2):183-194. [doi: 10.1080/02796015.2007.12087939]
9. Sonuga-Barke EJ, Halperin JM. Developmental phenotypes and causal pathways in attention deficit/hyperactivity disorder: potential targets for early intervention? J Child Psychol Psychiatry 2010 Apr 11;51(4):368-389 [FREE Full text] [doi: 10.1111/j.1469-7610.2009.02195.x] [Medline: 20015192]
10. Theule J, Wiener J, Tannock R, Jenkins JM. Parenting stress in families of children with ADHD: a meta-analysis. J Emot Behav Disord 2013 Mar;21(1):3-17 [FREE Full text] [doi: 10.1177/1063426610387433]
11. Zhao X, Page TF, Altszuler AR, Pelham WE, Kipp H, Gnagy EM, et al. Family burden of raising a child with ADHD. J Abnorm Child Psychol 2019 Aug 23;47(8):1327-1338. [doi: 10.1007/s10802-019-00518-5] [Medline: 30796648]

XSL•FO
RenderX

12. Charach A, Carson P, Fox S, Ali MU, Beckett J, Lim CG. Interventions for preschool children at high risk for ADHD: a comparative effectiveness review. Pediatrics 2013 May;131(5):e1584-e1604. [doi: 10.1542/peds.2012-0974] [Medline: 23545375]
13. Wolraich ML, Hagan JFJ, Allan C, Chan E, Davison D, Earls M, et al. Clinical practice guideline for the diagnosis, evaluation, and treatment of attention-deficit/hyperactivity disorder in children and adolescents. Pediatrics 2019 Oct;144(4):e20192528 [FREE Full text] [doi: 10.1542/peds.2019-2528] [Medline: 31570648]
14. Riddle MA, Yershova K, Lazzaretto D, Paykina N, Yenokyan G, Greenhill L, et al. The preschool attention-deficit/hyperactivity disorder treatment study (PATS) 6-year follow-up. J Am Acad Child Adolesc Psychiatry 2013 Mar;52(3):264-78.e2 [FREE Full text] [doi: 10.1016/j.jaac.2012.12.007] [Medline: 23452683]
15. Shaw P, Malek M, Watson B, Sharp W, Evans A, Greenstein D. Development of cortical surface area and gyrification in attention-deficit/hyperactivity disorder. Biol Psychiatry 2012 Aug 01;72(3):191-197 [FREE Full text] [doi: 10.1016/j.biopsych.2012.01.031] [Medline: 22418014]
16. Halperin JM, Marks DJ, Bedard AC, Chacko A, Curchack JT, Yoon CA, et al. Training executive, attention, and motor skills: a proof-of-concept study in preschool children with ADHD. J Atten Disord 2013 Nov 05;17(8):711-721. [doi: 10.1177/1087054711435681] [Medline: 22392551]
17. Bailey J, Barton B, Vignola A. Coping with children with ADHD: coping styles of mothers with children with ADHD or challenging behaviours. Early Child Dev Care 1999 Jan;148(1):35-50. [doi: 10.1080/0300443991480104]
18. Smith E, Koerting J, Latter S, Knowles MM, McCann DC, Thompson M, et al. Overcoming barriers to effective early parenting interventions for attention-deficit hyperactivity disorder (ADHD): parent and practitioner views. Child Care Health Dev 2015 Jan 12;41(1):93-102 [FREE Full text] [doi: 10.1111/cch.12146] [Medline: 24814640]
19. Whalen CK, Henker B, Jamner LD, Ishikawa SS, Floro JN, Swindle R, et al. Toward mapping daily challenges of living with ADHD: maternal and child perspectives using electronic diaries. J Abnorm Child Psychol 2006 Feb 04;34(1):115-130. [doi: 10.1007/s10802-005-9008-5] [Medline: 16463071]
20. Gwernan‐Jones R, Moore DA, Garside R, Richardson M, Thompson‐Coon J, Rogers M, et al. ADHD, parent perspectives and parent–teacher relationships: grounds for conflict. Br J Spec Educ 2015 Jan 29;42(3):279-300. [doi: 10.1111/1467-8578.12087]
21. Mofokeng M, van der Wath AE. Challenges experienced by parents living with a child with attention deficit hyperactivity disorder. J Child Adolesc Ment Health 2017 Sep 03;29(2):137-145. [doi: 10.2989/17280583.2017.1364253] [Medline: 28974169]
22. Mo KY, Chan ST. Parenting stress and coping strategies among parents of children with ADHD in China. J Child Fam Stud 2024 Nov 19;33:3826-3839. [doi: 10.1007/s10826-024-02916-w]
23. Lu T, Li L, Tang Y, Leavey G. ADHD and family life: a cross-sectional study of ADHD prevalence among pupils in China and factors associated with parental depression. PLoS One 2024 Mar 14;19(3):e0281226 [FREE Full text] [doi: 10.1371/journal.pone.0281226] [Medline: 38483917]
24. Xie S, Li H. ‘Tiger mom, panda dad’: a study of contemporary Chinese parenting profiles. Early Child Dev Care 2017 Apr 26;189(2):284-300. [doi: 10.1080/03004430.2017.1318870]
25. Shi Z. 'Mother hens' in China: a phenomenon of East Asian attitudes to education? ThinkChina. 2020 May 1. URL: https://www.thinkchina.sg/society/mother-hens-china-phenomenon-east-asian-attitudes-education [accessed 2025-07-10]
26. Li J, He Y. Unpacking “double reduction” policy in China: ideas, values, and practices. Beijing Int Rev Educ 2023 Sep 05;5(3):337-341. [doi: 10.1163/25902539-05030009]
27. Ho SW, Chien WT, Wang LQ. Parents' perceptions of care-giving to a child with attention deficit hyperactivity disorder: an exploratory study. Contemp Nurse 2011 Dec 17;40(1):41-56. [doi: 10.5172/conu.2011.40.1.41] [Medline: 22545905]
28. Mak WW, Kwok YT. Internalization of stigma for parents of children with autism spectrum disorder in Hong Kong. Soc Sci Med 2010 Jun;70(12):2045-2051. [doi: 10.1016/j.socscimed.2010.02.023] [Medline: 20362379]
29. Mustafa AN. The influence of Confucian principles on modern Chinese educational practices. Int J Res Publ Rev 2024 Jun;5(6):1580-1584 [FREE Full text] [doi: 10.55248/gengpi.5.0624.1446]
30. Xu Y, Zhang S, Li M, Liu D, Zhao H, Tang G. Work–family interference in urban China: gender discrimination and the effects of work–family balance policies. npj Urban Sustain 2024 Jan 08;4. [doi: 10.1038/s42949-023-00137-6]
31. Li F, Cui Y, Li Y, Guo L, Ke X, Liu J, et al. Prevalence of mental disorders in school children and adolescents in China: diagnostic data from detailed clinical assessments of 17,524 individuals. J Child Psychol Psychiatry 2022 Jan 21;63(1):34-46. [doi: 10.1111/jcpp.13445] [Medline: 34019305]
32. Bronfenbrenner U. The Ecology of Human Development: Experiments by Nature and Design. Cambridge, UK: Harvard University Press; 1979.
33. Bronfenbrenner U, Morris PA. The bioecological model of human development. In: Lerner RM, Damon W, editors. Handbook of Child Psychology: Theoretical Models of Human Development. Hoboken, NJ: John Wiley & Sons; 2006.
34. Neal JW, Neal ZP. Nested or networked? Future directions for ecological systems theory. Soc Dev 2013 Mar 11;22(4):722-737. [doi: 10.1111/sode.12018]

XSL•FO
RenderX

35. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec 16;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
36. Suri H. Purposeful sampling in qualitative research synthesis. Qual Res J 2011;11(2):63-75 [FREE Full text] [doi: 10.3316/QRJ1102063]
37. Krueger RA, Casey MA. Focus Groups: A Practical Guide for Applied Research. Thousand Oaks, CA: SAGE Publications; 2008.
38. Gale NK, Heath G, Cameron E, Rashid S, Redwood S. Using the framework method for the analysis of qualitative data in multi-disciplinary health research. BMC Med Res Methodol 2013 Sep 18;13(1):117 [FREE Full text] [doi: 10.1186/1471-2288-13-117] [Medline: 24047204]
39. Ritchie J, Spencer L. Qualitative data analysis for applied policy research. In: Huberman AM, Miles MB, editors. The Qualitative Researcher's Companion. Thousand Oaks, CA: SAGE Publications; 2002.
40. Wong L. Data analysis in qualitative research: a brief guide to using nvivo. Malays Fam Physician 2008;3(1):14-20 [FREE Full text] [Medline: 25606106]
41. Ryan GW, Bernard HR. Data management and analysis methods. In: Denzin N, Lincoln Y, editors. Handbook of Qualitative Research. Thousand Oaks, CA: SAGE Publications; 2000.
42. Lincoln YS, Guba EG. Naturalistic Inquiry. Thousand Oaks, CA: SAGE Publications; 1985.
43. Forero R, Nahidi S, De Costa J, Mohsin M, Fitzgerald G, Gibson N, et al. Application of four-dimension criteria to assess rigour of qualitative research in emergency medicine. BMC Health Serv Res 2018 Feb 17;18(1):120 [FREE Full text] [doi: 10.1186/s12913-018-2915-2] [Medline: 29454350]
44. Anderson GJ. Fundamentals of Educational Research. London, UK: RoutledgeFalmer; 1998.
45. Moen Ø, Hall-Lord ML, Hedelin B. Living in a family with a child with attention deficit hyperactivity disorder: a phenomenographic study. J Clin Nurs 2014 Nov 21;23(21-22):3166-3176. [doi: 10.1111/jocn.12559] [Medline: 25453121]
46. Hinojosa MS, Hinojosa R, Fernandez-Baca D, Knapp C, Thompson LA. Parental strain, parental health, and community characteristics among children with attention deficit-hyperactivity disorder. Acad Pediatr 2012 Nov;12(6):502-508. [doi: 10.1016/j.acap.2012.06.009] [Medline: 22985984]
47. Luo Y, Weibman D, Halperin JM, Li X. A review of heterogeneity in attention deficit/hyperactivity disorder (ADHD). Front Hum Neurosci 2019 Feb 11;13:42 [FREE Full text] [doi: 10.3389/fnhum.2019.00042] [Medline: 30804772]
48. Leitch S, Sciberras E, Post B, Gerner B, Rinehart N, Nicholson JM, et al. Experience of stress in parents of children with ADHD: a qualitative study. Int J Qual Stud Health Well-Being 2019 Dec 12;14(1):1690091 [FREE Full text] [doi: 10.1080/17482631.2019.1690091] [Medline: 31713473]
49. Moen Ø, Hedelin B, Hall-Lord ML. Family functioning, psychological distress, and well-being in parents with a child having ADHD. Sage Open 2016 Jan 31;6(1). [doi: 10.1177/2158244015626767]
50. Mueller AK, Fuermaier AB, Koerts J, Tucha L. Stigma in attention deficit hyperactivity disorder. ADHD Atten Def Hyp Disord 2012 Jul 8;4:101-114. [doi: 10.1007/s12402-012-0085-3]
51. Longo V, Saadati N, Karakus M. Exploring the role of extended family in child rearing practices across different cultures. J Psychosoc Res Fam Cult 2024;2(4):4-12. [doi: 10.61838/kman.jprfc.2.4.2]
52. Zhou X. Early childhood education policy development in China. Int J Child Care Educ Policy 2015 Feb 20;5:29-39. [doi: 10.1007/2288-6729-5-1-29]
53. Xiang YT, Ng CH, Yu X, Wang G. Rethinking progress and challenges of mental health care in China. World Psychiatry 2018 Jun 24;17(2):231-232 [FREE Full text] [doi: 10.1002/wps.20500] [Medline: 29856546]
54. Ma X, Lin Y, Yan W, Jin Z, Zhang Y. Parental ADHD knowledge and medical visit status of school-aged children in Shanghai. Child Adolesc Psychiatry Ment Health 2024 Jul 22;18(1):91 [FREE Full text] [doi: 10.1186/s13034-024-00780-6] [Medline: 39039526]
55. Bahraini S, Maisonneuve AR, Wu D, Huang M, Xu M, Yang L, et al. Barriers and facilitators to implementing a Canadian shared-care ADHD program in pediatric settings in Shanghai: a consolidated framework for implementation research approach. BMC Health Serv Res 2024 May 02;24(1):564 [FREE Full text] [doi: 10.1186/s12913-024-10910-7] [Medline: 38698381]
56. Ma JL. Family‐based intervention for Chinese families of children with attention deficit hyperactivity disorder (ADHD) in Hong Kong, China. Aust N Z J Fam Ther 2021 Nov 16;42(4):402-413. [doi: 10.1002/anzf.1468]
57. Ying Hu B, Roberts SK, Wang Y, Zhao H. The initiation of early childhood inclusion in China: a case study from Beijing. Int J Early Years Educ 2011 Jul 21;19(2):111-131. [doi: 10.1080/09669760.2011.596396]
58. Hu X, Yang X. Early intervention practices in China: present situation and future directions. Infant Young Child 2013;26(1):4-16. [doi: 10.1097/IYC.0b013e3182736cd3]
59. Wang Y, Huang Z, Kong F. Parenting stress and life satisfaction in mothers of children with cerebral palsy: the mediating effect of social support. J Health Psychol 2020 Mar 13;25(3):416-425. [doi: 10.1177/1359105317739100] [Medline: 29129110]
60. DuPaul GJ, Stoner G. ADHD in the Schools: Assessment and Intervention Strategies. New York, NY: Guilford Publications; 2003.

XSL•FO
RenderX

61. Li H, Yang W, Chen JJ. From 'Cinderella' to 'Beloved Princess': the evolution of early childhood education policy in China. Int J Child Care Educ Policy 2016 Aug 26;10:2. [doi: 10.1186/s40723-016-0018-2]
62. Mcloughlin CS, Zhou Z, Clark E. Reflections on the development and status of contemporary special education services in China. Psychol Sch 2005 Feb 18;42(3):273-283. [doi: 10.1002/pits.20078]
63. Bush T, Haiyan Q. Leadership and culture in Chinese education. Asia Pac J Educ 2013 Jan 02;20(2):58-67. [doi: 10.1080/02188791.2000.10600183]
64. Law WW. Educational leadership and culture in China: dichotomies between Chinese and Anglo-American leadership traditions? Int J Educ Dev 2012 Mar;32(2):273-282. [doi: 10.1016/j.ijedudev.2011.04.007]
65. Lynch JM. Responsibilities of today's principal: implications for principal preparation programs and principal certification policies. Rural Spec Educ Q 2012 Jun 01;31(2):40-47. [doi: 10.1177/875687051203100205]
66. Hu BY, Zhou Y, Li K. Variations in Chinese parental perceptions of early childhood education quality. Eur Early Child Educ Res J 2017 Jun 07;25(4):519-540. [doi: 10.1080/1350293x.2017.1331066]
67. Hallinger P, Liu S. Leadership and teacher learning in urban and rural schools in China: meeting the dual challenges of equity and effectiveness. Int J Educ Dev 2016 Nov;51:163-173. [doi: 10.1016/j.ijedudev.2016.10.001]
68. Blatchford P, Bassett P, Brown P. Examining the effect of class size on classroom engagement and teacher–pupil interaction: differences in relation to pupil prior attainment and primary vs. secondary schools. Learn Instr 2011 Dec;21(6):715-730. [doi: 10.1016/j.learninstruc.2011.04.001]
69. Li Y, Coplan RJ, Wang Y, Yin J, Zhu J, Gao Z, et al. Preliminary evaluation of a social skills training and facilitated play early intervention programme for extremely shy young children in China. Infant Child Dev 2016 Jan 28;25(6):565-574. [doi: 10.1002/icd.1959]
70. Huang GH, Gove M. Confucianism, Chinese families, and academic achievement: exploring how Confucianism and Asian descendant parenting practices influence children's academic achievement. In: Khine M, editor. Science Education in East Asia. Cham, Switzerland: Springer; 2015.
71. Wei Y, McGrath P, Hayden J, Kutcher S. The quality of mental health literacy measurement tools evaluating the stigma of mental illness: a systematic review. Epidemiol Psychiatr Sci 2017 May 02;27(5):433-462. [doi: 10.1017/s2045796017000178]

Abbreviations

ADHD: attention-deficit/hyperactivity disorder
COREQ: Consolidated Criteria for Reporting Qualitative Research

Edited by A Mavragani; submitted 17.03.25; peer-reviewed by SW Wo, Y Wu; comments to author 28.05.25; revised version received 21.06.25; accepted 11.07.25; published 29.08.25.

Please cite as:
Chen SC, Zhong CW, Li H, Li X, Fung HW, Wang L, Wu GT, Jiang XY, Yeung WF
Parental Challenges in Raising Preschoolers With Attention-Deficit/Hyperactivity Disorder in Mainland China's First-Tier Cities: Qualitative Study Using Framework Analysis
JMIR Form Res 2025;9:e74047
URL: https://formative.jmir.org/2025/1/e74047
doi:10.2196/74047
PMID:

©Shu-Cheng Chen, Chen-Wen Zhong, Han Li, Xiaoqing Li, Hong-Wang Fung, Lijun Wang, Guo-Tao Wu, Xin-Yue Jiang, Wing-Fai Yeung. Originally published in JMIR Formative Research (https://formative.jmir.org), 29.08.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Moderating Effect of Atypical Events on the Relationship Between Heart Rate and Stress in Medical Residents Working in an Intensive Care Unit: Longitudinal, Observational Daily Diary Study

Ruibei Li[1*], MD; Ujjwal Pasupulety[2*], MSc; Wellington Chang[2], MD; Adam C Frank[2], MD, PhD

[1]Los Angeles General Medical Center, Los Angeles, CA, United States

[2]Department of Psychiatry and the Behavioral Sciences, Keck School of Medicine, University of Southern California, Los Angeles, CA, United States

[*]these authors contributed equally

Corresponding Author:
Ruibei Li, MD
Los Angeles General Medical Center
2051 Marengo St
Los Angeles, CA, 90033
United States
Phone: 1 6306404105
Email: ruibei.li@med.usc.edu

Abstract

Background: Residency is a critical period in a physician's training, characterized by significant physical, cognitive, and emotional demands that make residents highly susceptible to stress and associated negative health outcomes. While physiological signals such as heart rate have been explored as potential biomarkers of stress, their predictive utility in high-stress environments such as the intensive care unit (ICU) remains inconclusive, especially when factoring in atypical events that can further exacerbate resident stress levels.

Objective: This study aimed to investigate the relationship between daily average heart rate (AHR) and perceived stress among ICU residents and examine the moderating effect of atypical events on this relationship.

Methods: The TILES (Tracking Individual Performance With Sensors)-2019 dataset collected longitudinal data from 44 ICU residents who provided daily self-reported stress ratings and wore a Fitbit device to track physiological data over a 3-week period. The main predictor variables were AHR and the occurrence of atypical events (both work and life related and daily hassles). The primary outcome was the level of perceived stress measured on a 7-point Likert scale. Linear mixed models were used to analyze the relationship between AHR and stress, accounting for within-subject and between-subject variance. Interaction effects between AHR and atypical events were also examined.

Results: The analysis revealed a significant positive association between AHR and perceived stress (β=0.032; P=.04) on standard days. However, this relationship was attenuated by the presence of negative atypical events (β=−0.076; P=.02). We further analyzed whether the severity of negative atypical events had an additional moderating effect but found no statistical significance.

Conclusions: AHR is a potential physiological marker for perceived stress in ICU residents, but its effect is moderated by negative atypical events. Future research should replicate these findings in more diverse cohorts, assess their generalizability to broader populations, and control for additional confounding variables. Incorporating negative atypical events into stress assessment could lead to more accurate and context-sensitive interpretations of physiological data.

(JMIR Form Res 2025;9:e67822) doi:10.2196/67822

KEYWORDS

wearables; medical residents; heart rate; stress; linear mixed models; ecological momentary assessments; intensive care unit

XSL•FO
RenderX

Introduction

Background

For graduating physicians, residency is a period of practical training that is pivotal in determining future success as a health care practitioner [1]. On a daily basis, residents are confronted with challenging medical and professional scenarios that impact personal well-being [2]. The physical, cognitive, and emotional demands placed upon physicians in training make this subpopulation highly susceptible to stress, burnout, imposter syndrome, and various other subclinical and clinical mental health conditions [3-6]. Given the perennially strained health care system and its unceasing requirements for resilient health care professionals, understanding the factors that modulate resident well-being is of paramount importance.

Several studies have ascertained the effectiveness of physiological signals as biomarkers of human emotions, behavioral responses, and stress [7,8]. The commercialization of wearable computing devices (ie, wearables) has allowed for an affordable and unobtrusive mechanism to procure longitudinal physiological data [9]. Smartphone apps offer a convenient apparatus to conduct ecological momentary assessments (EMAs) that instantly capture snapshots of users' behaviors and experiences via personal devices [10]. This combination of passive and active monitoring yields comprehensive biobehavioral data that can be leveraged to study the relationship between physiological signals and various psychological phenomena within specific populations. One of the first studies to use this methodology on resident physicians identified resilience indicators associated with physiological factors such as sleeping behavior, physical activity, increased mood, and reduced heart rate (HR) and mood variability [11].

Intensive care units (ICUs) present a unique clinical setting where patient cases become particularly stressful due to their urgency and complexity and the potential for rapid deterioration [12]. Apart from fulfilling day-to-day responsibilities (eg, rounds and paperwork), residents working an ICU rotation must be prepared to handle unexpected emergencies that are atypical to their regular workflow. These atypical events, which can range from individual patient deaths [13-15] to mass casualty incidents [16,17], generally overwhelm hospital resources and impair residents' decision-making due to elevations in stress. Atypical events that are unrelated to work (eg, life events and daily hassles) can also exacerbate resident stress [18].

While multiple studies have used wearables and EMAs to evaluate the well-being of residents in various medical specializations [19-22], there has been little exploration of experiences of stress in residents working in an ICU. A single study focused on the effects of sleep deprivation in ICU rotation residents [23]. Evidence shows that individuals experiencing burnout have elevated resting HRs compared to healthy controls [24]. Another study found that first-year medical students have higher hemodynamic parameters (HR and blood pressure) than college undergraduates [25]. Despite the nearly universal inclusion of HR as a variable in biobehavioral studies, there has been little agreement regarding its effectiveness as a predictor of stress in the general population [26] and in residents [27].

The TILES (Tracking Individual Performance With Sensors)-2019 is a multimodal dataset collected from residents in the ICU via a multitude of instruments such as wearables, proximity sensors, and EMAs [28]. The authors of the dataset found patterns linking daily self-reported stress ratings to computer use habits within the ICU and resident job satisfaction to the frequency of interactions with attending physicians [29]. A study on a similar dataset that collected atypical event data from workers in the health care and aerospace industries reported that even minor negative events can significantly increase negative affect, anxiety, and stress [18]. The TILES-2019 dataset contains daily HR measurements, perceived stress reports, and the presence or absence of atypical events, making it an ideal source to study their effect on resident well-being.

Objectives

In this study, our primary objective was to assess the use of average HR (AHR) as a physiological marker of psychological stress. In addition, we aimed to investigate the influence of atypical events on the relationship between AHR and perceived stress in ICU residents. Previous studies have shown mixed associations between HR indexes and stress [26,27], and at least one study has found that negative atypical events are strongly associated with higher self-reported stress [18]. However, no study has tested whether those events also change the concurrent physiological correlate. We used an exploratory approach and hypothesized that negative atypical events would significantly moderate the association between daily AHR and perceived stress. Specifically, we expected that (1) the strength of the relationship between AHR and stress would differ between standard days (no atypical events) and atypical event days and (2) negative atypical events would elevate perceived stress on average, with major work or life events producing larger increases than minor daily hassles.

Therefore, we used the EMA data provided by the TILES-2019 dataset to investigate differences in residents' self-reported stress between standard and atypical event days to capture dynamic, day-to-day stressors. A linear mixed model (LMM) was used to analyze day-to-day variations in residents' AHR to explain changes in stress response over the data collection period. The findings provide additional nuanced insights that explain the disparity in the literature regarding the predictive power of HR on perceived stress in resident physicians. Our results suggest that future biobehavioral studies should consider negative atypical events as significant data points.

Methods

Dataset

The TILES-2019 dataset consists of physiological and behavioral data collected from a cohort of 57 medical residents working in an ICU at the Los Angeles General Medical Center (formerly the Los Angeles County and University of Southern California Medical Center) [28]. The data collection efforts for the TILES-2019 study began in November 2019 and concluded in April 2020. Residents were invited to participate if they rotated through the ICU during the data collection period. In addition, they needed to be proficient in speaking and reading English and have access to an internet- and Bluetooth-enabled

XSL•FO
RenderX

personal smartphone, a personal email account, and Wi-Fi at home. Briefly, over a 3-week rotation, including both workdays and nonworkdays, residents wore a Fitbit Charge 3 smartwatch that captured continuous HR data. Residents also used a custom smartphone app (the TILES study app) on their personal device to complete daily EMAs that assessed various features related to well-being, such as end-of-day stress, sleep, job performance, and job satisfaction. They underwent a 2-hour onboarding session where, among other topics, they were instructed in the use of the smartphone app and answering EMAs through their personal smartphone. The afternoon-only prompt may be better described as a daily diary assessment rather than an EMA. To mark this distinction, we will henceforth refer to the daily stress item as a daily diary measure rather than an EMA prompt, although methodological sources often consider daily diary protocols to be a special variant within the broader EMA family [29-31].

The study assessed stress using a single item for which residents were prompted once a day on a fixed-interval design in the afternoon. Specifically, they were asked the following—"How would you rate your level of stress this afternoon?"—on a Likert scale from 1 to 7 [32]. Although this item is unvalidated, single-item Likert scale questions are frequently used in EMA research to keep participant burden low. A similar 7-item Likert scale was used by Janicki et al [33] in their study on acute stress in emergency medicine residents, and several recent EMA studies have similarly used a single, unvalidated item to capture stress or a closely related affective construct [34-37]. In addition, residents could specify whether an atypical event took place either during or outside their work shift. The overall compliance for Fitbit use and daily diary completion was 71% and 69.7%, respectively. In the study presented in this paper, a subset of items from the daily diary responses of residents was evaluated.

Preprocessing

To facilitate the analysis of longitudinal trends in stress, the following preprocessing steps were implemented in pandas [38] to combine streaming HR readings from the Fitbit smartwatch with the daily survey responses.

First, Fitbit sampled HR data continuously at 5- to 10-second intervals. For a particular resident, we calculated the AHR on a given date (from midnight to 11:59 PM Pacific Standard Time) by taking the average of the HR values. At times, there were gaps in the Fitbit data, which we addressed using a quality control procedure.

Second, among health care professionals, stress can be attributed to incidents that occur both within and outside the hospital setting [39]. On a given day, residents could indicate whether they experienced an atypical event. If present, they could mark the atypical event as *positive* (eg, celebrating a birthday) or *negative* (eg, unexpected stressors at work). Negative atypical events could be further differentiated as days with either *daily hassles* or *negative work/life events*. We omitted days in which positive atypical events were reported as the frequency (4/136, 2.9% atypical event days) of positive events was too low to analyze separately.

Third, extreme working hours have been implicated as a pertinent precipitant of burnout and depressive symptoms in various health care and non–health care environments [40-42]. The effect of the number of hours worked on resident burnout has been explored and found to be significant [43]. We controlled for the number of hours worked in a day by calculating the difference between the start and end times of the work shift and including it as a covariate in the LMM. Some entries had missing start or end times for work shifts (not both); this was resolved by assuming the default 12-hour work shift as defined in the original TILES-2019 study [31].

Fourth, each set of daily diary responses corresponding to a resident was chronologically sorted. The day numbers corresponding to each resident's service period in the ICU were included as repeated-measure variables in our LMMs.

Fifth, previous studies have determined that resident sex [38,44,45] and postgraduate year (PGY) [46] are significant factors that influence perceived stress. We used sex and PGY from baseline surveys as covariates in the LMM. Age was not considered as a covariate due to the limited range of age values in the study population.

Quality Control

In biobehavioral studies using wearables, despite participants' sincere attempts to adhere to protocols, gaps in recorded data are common due to multiple factors (eg, showering, charging, or forgetfulness) [47]. Incomplete physiological data generally result in inconclusive findings that cast doubt on the viability of wearables as effective ambulatory monitoring devices. To ensure accurate AHR calculations, the following quality control procedure was applied to each resident's Fitbit data file:

1. For a given date, find the rows of data with a matching date time stamp.
2. Store the number of rows (N). As Fitbit samples data every 5 to 10 seconds, a resident can have an N value in the range of 8640 to 17,280.
3. To account for device inaccuracies, assume that an ideal resident must have 10,800 rows of data. This is a fair assumption as it is equivalent to a sampling rate of 1 reading every 8 seconds.
4. Divide N by 10,800 and multiply the result by 24 to derive the total number of hours for which the Fitbit was worn (H) on the given date.
5. For the given date, include the calculated AHR in the final dataset only if the value of H is above a threshold number of hours (Q). Return a NULL value otherwise.

On a given date, the total number of recorded HR values for a resident was divided by 10,800. This provided an estimate for the number of hours for which the Fitbit was genuinely worn assuming a sampling rate of 1 reading every 8 seconds. The computed AHR value for that day was considered valid only if the resident wore the Fitbit for at least a specified threshold amount of hours. Residents were filtered by empirically determining an ideal cutoff point for the number of *hours worn*.

As LMMs are designed to handle missing data, rows without self-reported stress (due to an incomplete EMA), AHR (due to noncompliance), or hours worked were retained. We did not

XSL•FO
RenderX

observe any clear pattern for nonresponse to the daily survey, although no analysis was conducted to link missing data with self-reported stress.

Data Analysis

A multiple-comparison analysis was conducted to compare the mean self-reported stress on standard days (when no atypical event was reported) and days with the aforementioned atypical event types.

We analyzed daily diary data using LMMs to account for both within-subject variance and between-subject variance. The LMM is appropriate for repeated observations with an unequal number of observations per participant where the outcome variable approximates a normal distribution [48,49]. Level 1 variables (within-subject variance) included daily atypical event, daily hours worked, AHR, and the interaction term between daily atypical event and AHR. Level 2 variables (between-subject variance) included sex and PGY. All continuous predictors (AHR and hours worked) were person mean centered before analysis. The number of calendar days elapsed since each resident's study day was treated as a repeated measure nested within persons. We used an autoregressive covariance structure to account for serial correlation across consecutive days within individuals. A random intercept for residents (covariance type=variance components) allowed baseline stress levels to vary across individuals. Fixed effects were estimated via maximum likelihood with Satterthwaite df. Statistical significance was defined as $P<.05$. To assess whether the AHR-stress association was moderated by the presence of atypical events, we fit 2 hierarchical mixed models. Model 1 contained only main effects (AHR, hours worked, atypical event, and covariates). Model 2 added the atypical event × AHR interaction term. For our secondary analysis of event severity (model 3), we replaced the binary atypical event variable with a categorical variable with 3 levels (standard day, daily hassle, and negative work or life event), where standard day was used as the reference category. All statistical analyses were conducted using SPSS Statistics for Windows (version 29.0.0.0; IBM Corp).

Ethical Considerations

The original TILES-2019 study was conducted in accordance with the University of Southern California Institutional Review Board approval (study ID HS-19-00606) [31]. This study was conducted using the publicly available and deidentified TILES-2019 dataset and was exempt from institutional review board review. Participants in the original TILES study provided consent for the collection and use of their data for secondary analysis as documented by the original study authors as part of their data use agreement. For their participation, residents received weekly gift cards and were allowed to keep the Fitbit device.

Results

Quality-Controlled Dataset

On the basis of the defined quality control procedure, Figure 1 shows the number of residents with at least 5 days (approximately 25% of the study duration) of compliant HR data while varying the threshold number of total available hours of HR data. It was empirically determined that 11 hours of available HR data was a cutoff that balanced data quality with exclusion of residents for noncompliance. The final dataset comprised 44 residents.

Figure 1. Empirical determination of the quality control threshold for Fitbit heart rate (HR) data. A resident's HR data for a day were considered compliant only if they wore the Fitbit for at least 11 hours. Increasing the strictness of the hours worn threshold (x-axis) decreased the number of compliant residents (y-axis).

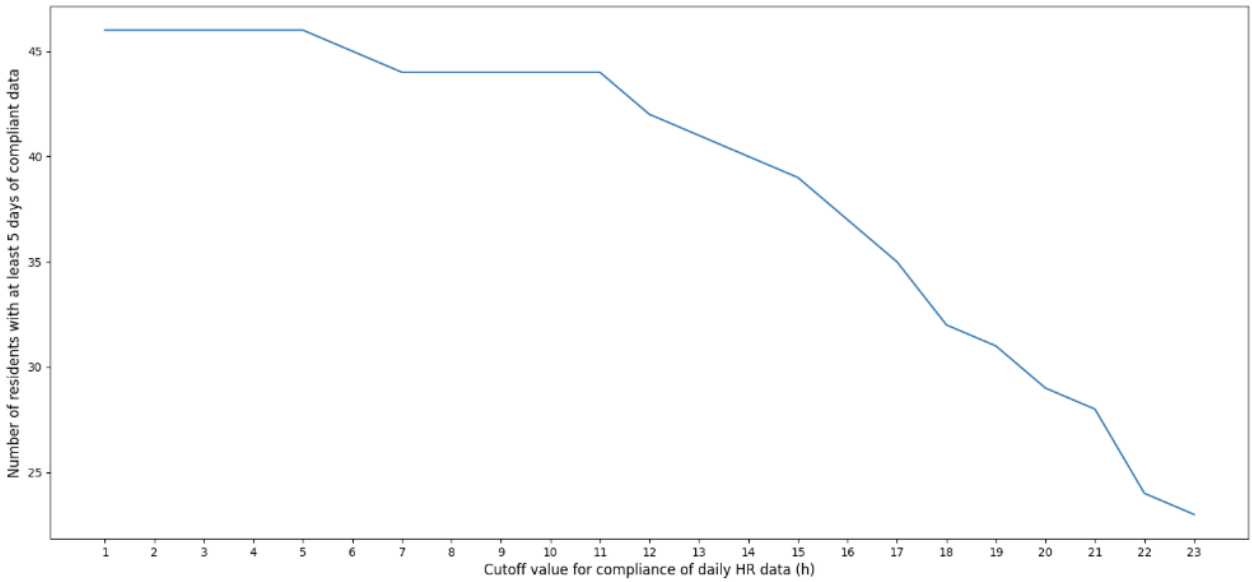

Demographics

Table 1 shows demographic information, with most residents identifying as Asian individuals (26/44, 59%), followed by White (11/44, 25%) and Hispanic or Latino individuals (3/44, 7%). A smaller proportion identified as Black individuals, multiethnic, or other or opted not to disclose their race or ethnicity (1/44, 2% each). The vast majority of residents (35/44,

XSL•FO
RenderX

80%) were from the internal medicine program, whereas 16% (7/44) were from emergency medicine, and 5% (2/44) were from the internal medicine or pediatrics program. The distribution across PGYs was relatively balanced, with 36% (16/44) in their first year, 34% (15/44) in their second year, and 30% (13/44) in their third year of training. Male individuals constituted most of the sample at 64% (28/44).

Table 1. Demographic information of residents in the quality-controlled dataset (N=44)[a].

Characteristic	Values
Age (y), mean (SD)	29.13 (2.03)
Race or ethnicity, n (%)	
Asian	26 (59)
Black	1 (2)
Hispanic or Latino	3 (7)
Multiethnic	1 (2)
White	11 (25)
Prefer not to say	1 (2)
Other	1 (2)
Specialty, n (%)	
Internal medicine	35 (80)
Emergency medicine	7 (16)
Meds or pediatrics	2 (5)
Postgraduate year, n (%)	
1 (intern)	16 (36)
2 (junior)	15 (34)
3 (senior)	13 (30)
Sex, n (%)	
Male	28 (64)
Female	16 (36)

[a]The final sample size constitutes participants with at least 5 days of heart rate data.

Descriptive Statistics of Quality-Controlled AHR

After applying the ≥11-hour wear time threshold, 710 days of HR data were retained for analysis, corresponding to 77% (44/57) of the enrolled residents. The daily AHR across these observations was 76.8 (SD 17.0) beats per minute (bpm). The distribution was approximately symmetrical (median 76, IQR 64-87 bpm; Table 2). Outlier values were rare, with the lowest daily mean recorded at 34 bpm and the highest at 202 bpm. These descriptive statistics fall within expected physiological ranges for adults.

Table 2. Descriptive statistics of quality-controlled average heart rate (AHR)[a].

Statistic	Daily AHR (bpm[b])
Values, mean (SD)	76.8 (17.0)
Values, median (IQR; range)	76 (64-87; 34-202)

[a]AHR was computed as the mean of continuous 5- to 10-second Fitbit heart rate data on each day with the ≥11-hour wear time.

[b]bpm: beats per minute.

Types of Negative Atypical Events

Residents experienced mostly standard days (n=588), with fewer occurrences of negative work or life events (n=77) and daily hassles (n=55). Multiple-comparison analysis using Bonferroni corrections showed significant differences in perceived stress across atypical event types. Specifically, the mean difference in perceived stress between work or life event days and standard days was 1.99 (P<.001), whereas the mean difference between daily hassle days and standard days was 0.941 (P<.001). Furthermore, perceived stress on work or life event days was significantly higher than perceived stress on daily hassle days, with a mean difference of 1.05 (P<.001).

XSL•FO
RenderX

LMM Results

Model 1 showed no significant association between AHR and perceived stress (β=0.016, SE 0.014; 2-tailed t_{424}=1.15; *P*=.25), whereas the number of hours worked (β=0.415, SE 0.087; $t_{17.4}$=4.79; *P*<.001) and the presence of a negative atypical event (β=1.36, SE 0.22; $t_{30.3}$=6.27; *P*<.001) were significantly positively associated with perceived stress (Table 3). Model 2 included the atypical event × AHR interaction, which was significant (β=−0.076, SE 0.032; t_{435}=−2.35; *P*=.02), indicating that AHR was associated with higher stress on standard days (β=0.032, SE 0.016; t_{415}=2.09; *P*=.04) but not on negative atypical event days (Table 4 and Figure 2).

Table 3. Model 1, a linear mixed model assessing the relationship between mean-corrected average heart rate (AHR) and perceived stress as measured using a daily diary[a].

Parameter	Perceived stress estimate (β; 95% CI)	*t* test (*df*)[b]	*P* value[c]
Intercept	3.21 (2.60 to 3.81)	10.7 (45.6)	*<.001* [c]
Female sex (reference=male)	0.129 (−0.482 to 0.740)	0.427 (40.0)	.67
PGY[d] 1 (reference=PGY 3)	0.390 (−0.340 to 1.12)	1.08 (40.7)	.29
PGY 2 (reference=PGY 3)	−0.279 (−1.01 to 0.454)	−0.770 (39.1)	.45
Hours worked[e]	0.415 (0.233 to 0.598)	4.79 (17.4)	*<.001*
AHR[e]	0.016 (−0.011 to 0.043)	1.15 (424)	.25
Negative atypical event (reference=standard day)	1.36 (0.918 to 1.80)	6.27 (30.3)	*<.001*

[a]Variables assessed include person mean–centered AHR, presence of atypical events, daily hours worked, resident sex, and postgraduate year. No significant association was found between AHR and daily perceived stress. Longer work hours and presence of a negative atypical event were associated with higher perceived stress, whereas sex and postgraduate year were not.

[b]Wald *t* statistic from SPSS MIXED procedure (Satterthwaite df).

[c]Italicized values indicate statistical significance (*P*<.05, 2-tailed).

[d]PGY: postgraduate year.

[e]Values are mean centered.

Table 4. Model 2, a linear mixed model assessing the interaction effect between atypical event and mean-centered average heart rate (AHR) on perceived stress[a].

Parameter	Perceived stress estimate (β; 95% CI)	*t* test (*df*)[b]	*P* value[c]
Intercept	3.22 (2.62 to 3.83)	10.7 (45.7)	*<.001* [c]
Female sex (reference=male)	0.117 (−0.498 to 0.731)	0.383 (40.3)	.70
PGY[d] 1 (reference=PGY 3)	0.403 (−0.330 to 1.14)	1.11 (40.9)	.27
PGY 2 (reference=PGY 3)	−0.289 (−1.03 to 0.448)	−0.792 (39.5)	.43
Hours worked[e]	0.418 (0.237 to 0.599)	4.87 (16.9)	*<.001*
AHR[e]	0.032 (0.002 to 0.063)	2.09 (415)	*.04*
Negative atypical event (reference=standard day)	1.40 (0.959 to 1.84)	6.49 (30.7)	*<.001*
Atypical event × AHR[e]	−0.076 (−0.140 to −0.013)	−2.35 (435)	*.02*

[a]Same as model 1 plus the interaction term. AHR was significantly associated with stress on standard days. This association was attenuated on days with a negative atypical event, as indicated by the significant negative interaction term.

[b]Wald *t* statistic from SPSS MIXED procedure (Satterthwaite df).

[c]Italicized values indicate statistical significance (*P*<.05, 2-tailed).

[d]PGY: postgraduate year.

[e]Values are mean centered.

XSL•FO
RenderX

Figure 2. Interaction effect between presence or absence of a negative atypical event day and mean-centered average heart rate (AHR) on perceived stress.

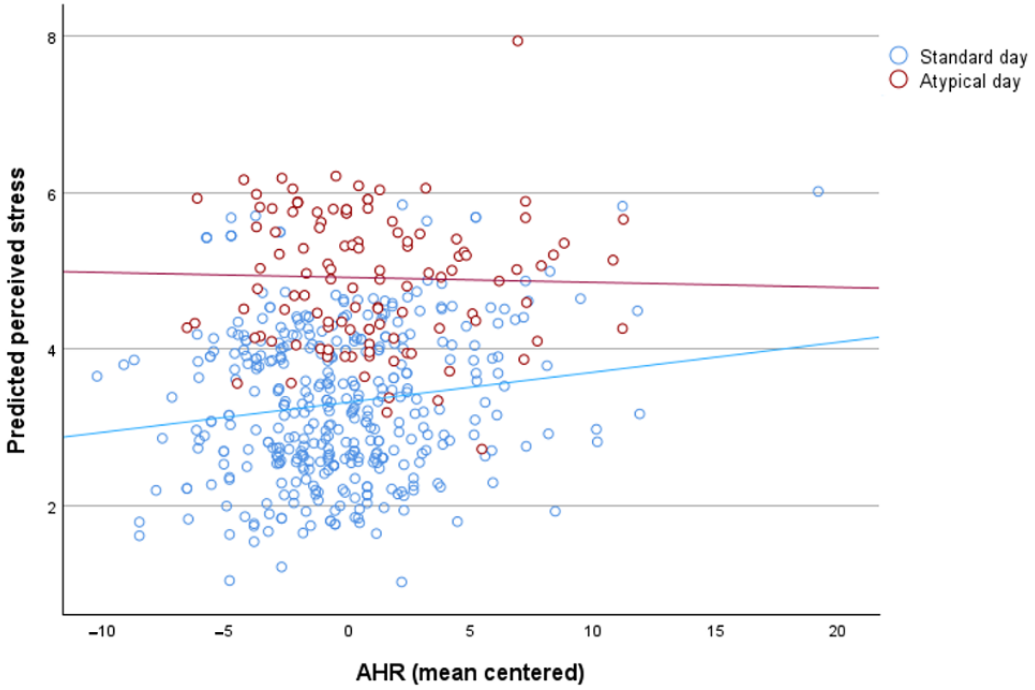

This moderating effect of negative atypical events persisted in our secondary analysis (model 3) on both daily hassle and negative work or life event days (β=−1.02 and −0.63, respectively), but the interaction term between these and AHR did not reach statistical significance (*P*=.05 and .11, respectively; Table 5). Both daily hassle (β=0.689, SE 0.268; $t_{52.2}$=2.57; *P*=.01) and negative work or life events (β=1.71, SE 0.219; $t_{39.0}$=7.84; *P*<.001) were significantly associated with increased stress compared to standard days, with negative work or life events exerting a more pronounced effect (Table 5).

Table 5. Model 3, a linear mixed model assessing the interaction effect between severity of negative atypical event and mean-centered average heart rate (AHR) on perceived stress[a].

Parameter	Perceived stress estimate (β; 95% CI)	*t* test (*df*)[b]	*P* value[c]
Intercept	3.20 (2.61 to 3.79)	10.9 (43.8)	*<.001* [c]
Female sex (reference=male)	0.146 (−0.456 to 0.748)	0.490 (38.8)	.63
PGY[d] 1 (reference=PGY 3)	0.473 (−0.245 to 1.19)	1.33 (39.4)	.19
PGY 2 (reference=PGY 3)	−0.283 (−1.00 to 0.438)	−0.794 (37.9)	.43
Hours worked[e]	0.369 (0.199 to 0.539)	4.66 (13.5)	*<.001*
AHR[e]	0.036 (0.006 to 0.066)	2.33 (407)	*.02*
Daily hassle (reference=standard day)	0.689 (0.151 to 1.23)	2.57 (52.2)	*.01*
Negative work or life event (reference=standard day)	1.71 (1.27 to 2.16)	7.84 (39.0)	*<.001*
Daily hassle × AHR[e] (reference=standard day)	−0.102 (−0.206 to 0.002)	−1.93 (448)	.05
Negative work or life event × AHR[e] (reference=standard day)	−0.063 (−0.140 to 0.014)	−1.61 (439)	.11

[a]Negative atypical events were further classified into a 3-level categorical factor (standard day [reference], daily hassle, and major work or life event) with corresponding AHR interaction terms. Both minor and major negative events increased perceived stress and attenuated the positive association between AHR and perceived stress, although interaction terms did not reach statistical significance.

[b]Wald *t* statistic from SPSS MIXED procedure (Satterthwaite df).

[c]Italicized values indicate statistical significance (*P*<.05, 2-tailed).

[d]PGY: postgraduate year.

[e]Values are mean centered.

XSL•FO
RenderX

Discussion

AHR-Stress Association

The significant association between higher daily AHR and increased levels of daily stress among ICU residents on standard days demonstrates a connection between physiological parameters and psychological well-being in a subpopulation exposed to numerous acute and chronic stressors. Our finding underscores the value of leveraging AHR as an objective, noninvasive, and scalable measure to understand and monitor stress levels in medical residents. This is of particular importance given that physician stress is reported to not only negatively impact their health but also diminish patient-perceived quality of care and increase the risk of adverse medical events [50,51].

Relation to Previous Literature

Previous studies have examined HR response to stressful stimuli in medical residents. Tendulkar et al [52] reported a significant increase in AHR in a group of surgical residents on a 24-hour call shift regardless of year of training compared to their AHR when off clinical responsibilities. Similarly, in a 2021 prospective observational study involving a cohort of emergency medicine residents, Janicki et al [33] found a significant increase in AHR while residents were clinically engaged compared to baseline HR measured during didactics. The study also reported a significant increase in subjectively recorded stress levels during the clinical period; however, it did not assess the association between mean HR and subjective stress directly [33].

Previous literature among the general population examining the relationship between HR and perceived stress as measured using EMA has been mixed, with some studies finding no association between HR and stress; some identifying positive associations; and others finding context-dependent positive associations, such as in the presence of chronic stress and posttraumatic stress disorder [26]. Overall, the current literature suggests that HR can be a physiological marker of stress in residents but its effectiveness and the nature of the association may depend on specific contexts and individual differences that moderate the relationship between HR and perceived stress.

Atypical Events as Moderators of Stress and AHR

One such moderator may be the presence of atypical events. In line with a previous study on the effects of atypical events on perceived stress in a combined population of nurses and aviation workers, we found that the presence of negative atypical events regardless of severity led to a significant increase in perceived stress when compared to a standard day [18]. Our study additionally suggests that the positive association between AHR and stress may only occur on standard days. Including days with negative atypical events attenuates this relationship to the point of statistical nonsignificance. This effect may persist on both severe atypical event days, defined as work or life events, as well as minor atypical event days, defined as daily hassles. Specifically, our analysis did not find a differential moderating effect of atypical event severity on the relationship between AHR and perceived stress. Both minor and severe atypical event days appear to attenuate the positive relationship between AHR and stress observed on standard days, albeit these effects not being statistically significant, likely due to our limited sample size.

The physiological basis underlying the moderating effect of atypical events is unclear. One possible contributory factor may involve a biological habituation process in response to frequent exposure to stressors, resulting in a blunting in physiological response while psychological recall and rating of the stressors remain intact. Stress exposure engages multiple endogenous stress-reactive systems, including the sympatho-adrenomedullary (SAM) system [53-56]. Activation of the SAM system occurs first and initiates the release of catecholamines that increase HR, blood pressure, and glucose production to prepare the body for a *fight or flight* response [54]. While this acute stress response is adaptive in the short term, frequent activation of the SAM system may result in a blunted HR response without concomitant change in subjective stress ratings [57]. The implications of this blunted physiological response is highlighted by evidence from numerous prospective studies that have linked reduced HR reactivity to stress with adverse health outcomes, including heightened risks of obesity, depression, anxiety, and posttraumatic stress disorder symptoms; increased illness frequency; musculoskeletal pain; and declines in both cognitive and physical health [56,58,59].

Our study suggests that the psychological perception of stress may be masked if only AHR is considered, leading to a potential underestimation of stress levels in these critical periods. Incorporating negative atypical events as a moderating factor may result in more accurate and context-sensitive interpretations of AHR data. This is crucial for understanding how stress manifests in residents working in high-pressure environments such as ICUs, which may allow for the development of better targeted interventions and, ultimately, the improvement of well-being for residents and their patients.

Limitations

This study has several limitations. The sample size was relatively small and drawn from a single institution, which limits statistical power and the strength of our conclusions. The TILES-2019 study did not collect data on potential confounding factors that could contribute to the hyper- or hypoactivation of stress response systems, such as smoking, caffeine intake, presence of medical conditions, or engagement in physical activity during and outside of clinical hours. In addition, the lack of diversity and skew toward male individuals in our demographic composition may limit the generalizability of our findings to underrepresented ethnic groups or female individuals, who may experience stress differently or have different physiological responses to stress. Importantly, our outcome variable was measured using a once-daily, unvalidated Likert-scale item. While this necessarily limits content validity and reliability estimation, it remains a pragmatic compromise for high-burden participants such as medical residents and is widely used in contemporary EMA research, having been adopted in multiple recent peer-reviewed studies [33-37]. Finally, we note that positive atypical events were excluded

XSL•FO
RenderX

from our analyses due to their low number; however, positive events can also affect HR and stress in significant ways.

Conclusions

We demonstrated the utility of AHR as a physiological marker for stress among ICU residents. The presence of atypical events attenuates this relationship, suggesting that AHR alone may not fully capture psychological stress during these critical periods and that incorporating atypical events as a moderating factor should be considered for more accurate stress assessment. The use of EMA was key to this as it allowed us to capture fluctuations in stress on a more granular level that would be missed through infrequent or more retrospective assessment. Further investigation should determine whether this moderating effect can be replicated in more diverse cohorts of residents, assess its generalizability to broader populations, and control for additional confounding variables that may influence these associations.

Acknowledgments

This project was supported by the Southern California Clinical and Translational Science Institute (grant KL2TR000131). The authors would like to thank Wendy Jean Mack, PhD (Keck School of Medicine, University of Southern California), for providing expert advice on statistical modeling. They also thank Jerry Zhao, MD (Riverside University Health System), for curating the dataset used in this study. Finally, the authors thank all the intensive care unit resident physicians for their time, effort, and participation in the study, as well as their care for patients who are critically ill. No generative artificial intelligence tool was used for manuscript writing.

Data Availability

The datasets generated or analyzed during this study are available in the TILES (Tracking Individual Performance With Sensors) repository [60].

Authors' Contributions

RL contributed to conceptualization, formal analysis, methodology, investigation, and writing—original draft. UP contributed to data curation, investigation, software, writing—original draft, and visualization. WC contributed to conceptualization and writing—review and editing. ACF contributed to resources, project administration, supervision, funding, conceptualization, and writing—review and editing.

Conflicts of Interest

None declared.

References

1. Raimo J, LaVine S, Spielmann K, Akerman M, Friedman KA, Katona K, et al. The correlation of stress in residency with future stress and burnout: a 10-year prospective cohort study. J Grad Med Educ 2018 Oct;10(5):524-531 [FREE Full text] [doi: 10.4300/JGME-D-18-00273.1] [Medline: 30386477]
2. Yao DC, Wright SM. The challenge of problem residents. J Gen Intern Med 2001 Jul;16(7):486-492 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016007486.x] [Medline: 11520388]
3. Rodrigues H, Cobucci R, Oliveira A, Cabral JV, Medeiros L, Gurgel K, et al. Burnout syndrome among medical residents: a systematic review and meta-analysis. PLoS One 2018 Nov 12;13(11):e0206840 [FREE Full text] [doi: 10.1371/journal.pone.0206840] [Medline: 30418984]
4. Bhama AR, Ritz EM, Anand RJ, Auyang ED, Lipman J, Greenberg JA, et al. Imposter syndrome in surgical trainees: clance imposter phenomenon scale assessment in general surgery residents. J Am Coll Surg 2021 Nov;233(5):633-638. [doi: 10.1016/j.jamcollsurg.2021.07.681] [Medline: 34384871]
5. Mata DA, Ramos MA, Bansal N, Khan R, Guille C, Di Angelantonio E, et al. Prevalence of depression and depressive symptoms among resident physicians: a systematic review and meta-analysis. JAMA 2015 Dec 08;314(22):2373-2383 [FREE Full text] [doi: 10.1001/jama.2015.15845] [Medline: 26647259]
6. Joules N, Williams DM, Thompson AW. Depression in resident physicians: a systematic review. Open J Depress 2014;03(03):89-100. [doi: 10.4236/ojd.2014.33013]
7. Egger M, Ley M, Hanke S. Emotion recognition from physiological signal analysis: a review. Electron Notes Theor Comput Sci 2019 May;343:35-55. [doi: 10.1016/j.entcs.2019.04.009]
8. Giannakakis G, Grigoriadis D, Giannakaki K, Simantiraki O, Roniotis A, Tsiknakis M. Review on psychological stress detection using biosignals. IEEE Trans Affect Comput 2022 Jan 1;13(1):440-460. [doi: 10.1109/taffc.2019.2927337]
9. Gradl S, Wirth M, Richer R, Rohleder N, Eskofier BM. An overview of the feasibility of permanent, real-time, unobtrusive stress measurement with current wearables. In: Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare. 2019 Presented at: PervasiveHealth'19; May 20-23, 2019; Trento, Italy. [doi: 10.1145/3329189.3329233]

XSL•FO
RenderX

10. de Vries LP, Baselmans BM, Bartels M. Smartphone-based ecological momentary assessment of well-being: a systematic review and recommendations for future studies. J Happiness Stud 2021 Oct 23;22(5):2361-2408 [FREE Full text] [doi: 10.1007/s10902-020-00324-7] [Medline: 34720691]
11. Adler DA, Tseng VW, Qi G, Scarpa J, Sen S, Choudhury T. Identifying mobile sensing indicators of stress-resilience. Proc ACM Interact Mob Wearable Ubiquitous Technol 2021 Jun;5(2):51 [FREE Full text] [doi: 10.1145/3463528] [Medline: 35445162]
12. Chuang CH, Tseng PC, Lin CY, Lin KH, Chen YY. Burnout in the intensive care unit professionals: a systematic review. Medicine (Baltimore) 2016 Dec;95(50):e5629 [FREE Full text] [doi: 10.1097/MD.0000000000005629] [Medline: 27977605]
13. McFarland DC, Maki RG, Holland J. Psychological distress of internal medicine residents rotating on a hematology and oncology ward: an exploratory study of patient deaths, personal stress, and attributed meaning. Med Sci Educ 2015 Dec 25;25(4):413-420 [FREE Full text] [doi: 10.1007/s40670-015-0159-x] [Medline: 32440367]
14. Ibrahim H, Harhara T. How internal medicine residents deal with death and dying: a qualitative study of transformational learning and growth. J Gen Intern Med 2022 Oct 22;37(13):3404-3410 [FREE Full text] [doi: 10.1007/s11606-022-07441-4] [Medline: 35194741]
15. Ye WQ, Griffin C, Sverdlichenko I, Vegas DB. The impact of patient death experiences early in training on resident physicians: a qualitative study. CMAJ Open 2023 Oct 31;11(5):E1006-E1011 [FREE Full text] [doi: 10.9778/cmajo.20230011] [Medline: 37907213]
16. Uddin H, Hasan MK, Cuartas-Alvarez T, Castro-Delgado R. Effects of mass casualty incidents on anxiety, depression, and post-traumatic stress disorder among doctors and nurses: a systematic review. Public Health 2024 Sep;234:132-142 [FREE Full text] [doi: 10.1016/j.puhe.2024.06.001] [Medline: 39002283]
17. Havron WS3, Safcsak K, Corsa J, Loudon A, Cheatham ML. Psychological effect of a mass casualty event on general surgery residents. J Surg Educ 2017 Nov;74(6):e74-e80. [doi: 10.1016/j.jsurg.2017.07.021] [Medline: 28781134]
18. Burghardt K, Tavabi N, Ferrara E, Narayanan S, Lerman K. Having a bad day? Detecting the impact of atypical events using wearable sensors. In: Proceedings of the 14th International Conference on Social, Cultural, and Behavioral Modeling. 2021 Presented at: SBP-BRiMS 2021; July 6-9, 2021; Virtual Event. [doi: 10.1007/978-3-030-80387-2_25]
19. Barron LK, Moon MR. Impact of autonomic regulation on burnout and performance in thoracic surgery residents. JTCVS Open 2022 Sep;11:229-240 [FREE Full text] [doi: 10.1016/j.xjon.2022.04.044] [Medline: 36172445]
20. Schrager JD, Shayne P, Wolf S, Das S, Patzer RE, White M, et al. Assessing the influence of a Fitbit physical activity monitor on the exercise practices of emergency medicine residents: a pilot study. JMIR Mhealth Uhealth 2017 Jan 31;5(1):e2 [FREE Full text] [doi: 10.2196/mhealth.6239] [Medline: 28143805]
21. Lisann-Goldman L, Cowart C, Lin HM, Orlando B, Mahoney B. Well-being in anesthesiology graduate medical education: reconciling the ideal with reality. Anesthesiol Clin 2022 Jun;40(2):383-397. [doi: 10.1016/j.anclin.2022.01.011] [Medline: 35659409]
22. Driesman AS, Strauss EJ, Konda SR, Egol KA. Factors associated with orthopaedic resident burnout: a pilot study. J Am Acad Orthop Surg 2020 Feb 7;28(21):900-906. [doi: 10.5435/jaaos-d-19-00648]
23. Sagun S, DeCicco D, Badami V, Mittal A, Thompson J, Pham C, et al. iSleepFirst: burnout, fatigue, and wearable-tracked sleep deprivation among residents staffing the medical intensive care unit. Sleep Breath 2023 Dec 27;27(6):2491-2497 [FREE Full text] [doi: 10.1007/s11325-023-02839-8] [Medline: 37243855]
24. De Vente W, Olff M, Van Amsterdam JG, Kamphuis JH, Emmelkamp PM. Physiological differences between burnout patients and healthy controls: blood pressure, heart rate, and cortisol responses. Occup Environ Med 2003 Jun 01;60 Suppl 1(Suppl 1):i54-i61 [FREE Full text] [doi: 10.1136/oem.60.suppl_1.i54] [Medline: 12782748]
25. May RW, Seibert GS, Sanchez-Gonzalez MA, Fincham FD. Physiology of school burnout in medical students: hemodynamic and autonomic functioning. Burnout Res 2016 Sep;3(3):63-68 [FREE Full text] [doi: 10.1016/j.burn.2016.05.001]
26. Weber J, Angerer P, Apolinário-Hagen J. Physiological reactions to acute stressors and subjective stress during daily life: a systematic review on ecological momentary assessment (EMA) studies. PLoS One 2022 Jul 27;17(7):e0271996. [doi: 10.1371/journal.pone.0271996] [Medline: 35895674]
27. Barac M, Scaletty S, Hassett LC, Stillwell A, Croarkin PE, Chauhan M, et al. Wearable technologies for detecting burnout and well-being in health care professionals: scoping review. J Med Internet Res 2024 Jun 25;26:e50253 [FREE Full text] [doi: 10.2196/50253] [Medline: 38916948]
28. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol 2008;4:1-32. [doi: 10.1146/annurev.clinpsy.3.022806.091415] [Medline: 18509902]
29. Czyz EK, King CA, Nahum-Shani I. Ecological assessment of daily suicidal thoughts and attempts among suicidal teens after psychiatric hospitalization: lessons about feasibility and acceptability. Psychiatry Res 2018 Sep;267:566-574 [FREE Full text] [doi: 10.1016/j.psychres.2018.06.031] [Medline: 30031317]
30. Hulsmans DH, Poelen EA, Lichtwarck-Aschoff A, Otten R. The feasibility of daily monitoring in adolescents and young adults with mild intellectual disability or borderline intellectual functioning. J Appl Res Intellect Disabil 2023 Jul;36(4):847-858. [doi: 10.1111/jar.13102] [Medline: 37073448]

XSL•FO
RenderX

31. Yau JC, Girault B, Feng T, Mundnich K, Nadarajan A, Booth BM, et al. TILES-2019: a longitudinal physiologic and behavioral data set of medical residents in an intensive care unit. Sci Data 2022 Sep 01;9(1):536 [FREE Full text] [doi: 10.1038/s41597-022-01636-4] [Medline: 36050329]
32. Feng T, Narayanan S. Understanding stress, burnout, and behavioral patterns in medical residents using large-scale longitudinal wearable recordings. In: Proceedings of the 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. 2024 Presented at: EMBC 2014; July 15-19, 2024; Orlando, FL. [doi: 10.1109/embc53108.2024.10782414]
33. Janicki AJ, Frisch SO, Patterson PD, Brown A, Frisch A. Emergency medicine residents experience acute stress while working in the emergency department. West J Emerg Med 2020 Dec 11;22(1):94-100 [FREE Full text] [doi: 10.5811/westjem.2020.10.47641] [Medline: 33439813]
34. Whitaker M, Welch WA, Fanning J, Santa-Maria CA, Auster-Gussman LA, Solk P, et al. Using ecological momentary assessment to understand associations between daily physical activity and symptoms in breast cancer patients undergoing chemotherapy. Support Care Cancer 2022 Aug;30(8):6613-6622. [doi: 10.1007/s00520-022-07071-w] [Medline: 35488902]
35. Tung NY, Yap Y, Bei B, Luecken LJ, Wiley JF. Resilience to early family risk moderates stress-affect associations: a 14-day ecological momentary assessment study. J Affect Disord 2022 Aug 15;311:134-142. [doi: 10.1016/j.jad.2022.05.075] [Medline: 35597463]
36. Cox RC, Jessup SC, Olatunji BO. The effects of worry proneness on diurnal anxiety: an ecological momentary assessment approach. Pers Individ Dif 2022 May;190:111529 [FREE Full text] [doi: 10.1016/j.paid.2022.111529] [Medline: 35210672]
37. Rubin LF, Haaga DA, Pearson JL, Gunthert KC. Depression as a moderator of the prospective relationship between mood and smoking. Health Psychol 2020 Feb;39(2):99-106 [FREE Full text] [doi: 10.1037/hea0000816] [Medline: 31682148]
38. McKinney W. Data structures for statistical computing in python. In: SciPy Proceedings. 2010 Presented at: SciPy 2010; June 28-July 3, 2010; Austin, Texas URL: https://proceedings.scipy.org/articles/Majora-92bf1922-00a [doi: 10.25080/majora-92bf1922-00a]
39. Weinberg A, Creed F. Stress and psychiatric disorder in healthcare professionals and hospital staff. The Lancet 2000 Feb;355(9203):533-537. [doi: 10.1016/s0140-6736(99)07366-3]
40. Yoon Y, Ryu J, Kim H, Kang CW, Jung-Choi K. Working hours and depressive symptoms: the role of job stress factors. Ann Occup Environ Med 2018 Jul 13;30(1):46 [FREE Full text] [doi: 10.1186/s40557-018-0257-5] [Medline: 30009036]
41. Lee K, Suh C, Kim JE, Park JO. The impact of long working hours on psychosocial stress response among white-collar workers. Ind Health 2017 Feb 07;55(1):46-53 [FREE Full text] [doi: 10.2486/indhealth.2015-0173] [Medline: 27498571]
42. Lin RT, Lin YT, Hsia YF, Kuo CC. Long working hours and burnout in health care workers: non-linear dose-response relationship and the effect mediated by sleeping hours-a cross-sectional study. J Occup Health 2021 Jan;63(1):e12228 [FREE Full text] [doi: 10.1002/1348-9585.12228] [Medline: 33957007]
43. Kijima S, Tomihara K, Tagawa M. Effect of stress coping ability and working hours on burnout among residents. BMC Med Educ 2020 Jul 13;20(1):219 [FREE Full text] [doi: 10.1186/s12909-020-02134-0] [Medline: 32660575]
44. Dyrbye LN, West CP, Herrin J, Dovidio J, Cunningham B, Yeazel M, et al. A longitudinal study exploring learning environment culture and subsequent risk of burnout among resident physicians overall and by gender. Mayo Clin Proc 2021 Aug;96(8):2168-2183 [FREE Full text] [doi: 10.1016/j.mayocp.2020.12.036] [Medline: 34218879]
45. Ebrahimi S, Kargar Z. Occupational stress among medical residents in educational hospitals. Ann Occup Environ Med 2018 Aug 8;30(1):51 [FREE Full text] [doi: 10.1186/s40557-018-0262-8] [Medline: 30101032]
46. Abdulghani H, Alharbi M, Irshad M. Stress and its association with working efficiency of junior doctors during three postgraduate residency training programs. Neuropsychiatr Dis Treat 2015 Dec;11:3023-3029. [doi: 10.2147/ndt.s92408]
47. Lederer L, Breton A, Jeong H, Master H, Roghanizad AR, Dunn J. The importance of data quality control in using Fitbit device data from the research program. JMIR Mhealth Uhealth 2023 Nov 03;11:e45103 [FREE Full text] [doi: 10.2196/45103] [Medline: 37962944]
48. Hedeker D, Mermelstein RJ, Demirtas H. Modeling between-subject and within-subject variances in ecological momentary assessment data using mixed-effects location scale models. Stat Med 2012 Nov 30;31(27):3328-3336 [FREE Full text] [doi: 10.1002/sim.5338] [Medline: 22419604]
49. Oleson JJ, Jones MA, Jorgensen EJ, Wu YH. Statistical considerations for analyzing ecological momentary assessment data. J Speech Lang Hear Res 2022 Jan 12;65(1):344-360 [FREE Full text] [doi: 10.1044/2021_jslhr-21-00081]
50. Patel RS, Bachu R, Adikey A, Malik M, Shah M. Factors related to physician burnout and its consequences: a review. Behav Sci (Basel) 2018 Oct 25;8(11):98 [FREE Full text] [doi: 10.3390/bs8110098] [Medline: 30366419]
51. Mangory KY, Ali LY, Rø KI, Tyssen R. Effect of burnout among physicians on observed adverse patient outcomes: a literature review. BMC Health Serv Res 2021 Apr 21;21(1):369 [FREE Full text] [doi: 10.1186/s12913-021-06371-x] [Medline: 33879135]
52. Tendulkar A, Victorino G, Chong T, Bullard MK, Liu TH, Harken AH. Quantification of surgical resident stress "on call". J Am Coll Surg 2005 Oct;201(4):560-564. [doi: 10.1016/j.jamcollsurg.2005.05.004] [Medline: 16183494]
53. Jönsson P, Österberg K, Wallergård M, Hansen Å, Garde AH, Johansson G, et al. Exhaustion-related changes in cardiovascular and cortisol reactivity to acute psychosocial stress. Physiol Behav 2015 Nov 01;151:327-337 [FREE Full text] [doi: 10.1016/j.physbeh.2015.07.020] [Medline: 26210042]

XSL•FO
RenderX

54. Godoy LD, Rossignoli MT, Delfino-Pereira P, Garcia-Cairasco N, de Lima Umeoka EH. A comprehensive overview on stress neurobiology: basic concepts and clinical implications. Front Behav Neurosci 2018 Jul 3;12:127 [FREE Full text] [doi: 10.3389/fnbeh.2018.00127] [Medline: 30034327]
55. Herman JP, McKlveen JM, Ghosal S, Kopp B, Wulsin A, Makinson R, et al. Regulation of the hypothalamic-pituitary-adrenocortical stress response. Compr Physiol 2016 Mar 15;6(2):603-621 [FREE Full text] [doi: 10.1002/cphy.c150015] [Medline: 27065163]
56. Turner AI, Smyth N, Hall SJ, Torres SJ, Hussein M, Jayasinghe SU, et al. Psychological stress reactivity and future health and disease outcomes: a systematic review of prospective evidence. Psychoneuroendocrinology 2020 Apr;114:104599. [doi: 10.1016/j.psyneuen.2020.104599] [Medline: 32045797]
57. Schommer NC, Hellhammer DH, Kirschbaum C. Dissociation between reactivity of the hypothalamus-pituitary-adrenal axis and the sympathetic-adrenal-medullary system to repeated psychosocial stress. Psychosom Med 2003;65(3):450-460. [doi: 10.1097/01.psy.0000035721.12441.17] [Medline: 12764219]
58. McEwen BS, Wingfield JC. What is in a name? Integrating homeostasis, allostasis and stress. Horm Behav 2010 Feb;57(2):105-111 [FREE Full text] [doi: 10.1016/j.yhbeh.2009.09.011] [Medline: 19786032]
59. Sapolsky RM. Stress and the brain: individual variability and the inverted-U. Nat Neurosci 2015 Oct 25;18(10):1344-1346. [doi: 10.1038/nn.4109] [Medline: 26404708]
60. TILES data sets. TILES. URL: https://tiles-data.isi.edu/download_tiles_2019 [accessed 2025-08-25]

Abbreviations

AHR: average heart rate
bpm: beats per minute
EMA: ecological momentary assessment
HR: heart rate
ICU: intensive care unit
LMM: linear mixed model
PGY: postgraduate year
SAM: sympatho-adrenomedullary
TILES: Tracking Individual Performance With Sensors

Edited by A Mavragani; submitted 21.10.24; peer-reviewed by A Schwerdtfeger; comments to author 15.05.25; revised version received 01.08.25; accepted 11.08.25; published 05.09.25.

Please cite as:
Li R, Pasupulety U, Chang W, Frank AC
The Moderating Effect of Atypical Events on the Relationship Between Heart Rate and Stress in Medical Residents Working in an Intensive Care Unit: Longitudinal, Observational Daily Diary Study
JMIR Form Res 2025;9:e67822
URL: https://formative.jmir.org/2025/1/e67822
doi:10.2196/67822
PMID:

©Ruibei Li, Ujjwal Pasupulety, Wellington Chang, Adam C Frank. Originally published in JMIR Formative Research (https://formative.jmir.org), 05.09.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Association Between In-Hospital Applications for Long-Term Care Services and Hospital Length of Stay Among Older Adults: Ecological Cross-Sectional Study

Naoki Takashi[1], DrPH; Joji Onishi[1], MD, MPH, PhD; Misaki Fujisawa[1], MPH; Tomoko Ohura[2], DrPH; Shosuke Ohtera[1], DrPH

[1]Department of Health Economics, Center for Gerontology and Social Science, National Center for Geriatrics and Gerontology, Research Institute, Obu, Japan

[2]Evidence-based Long-term Care Team, Center for Gerontology and Social Science, National Center for Geriatrics and Gerontology, Research Institute, Obu, Japan

Corresponding Author:
Naoki Takashi, DrPH
Department of Health Economics
Center for Gerontology and Social Science
National Center for Geriatrics and Gerontology, Research Institute
7-430, Morioka-Cho
Obu, 474-8511
Japan
Phone: 81 0562 46 2311
Email: takashi@ncgg.go.jp

Abstract

Background: Delayed discharge among older patients presents a major challenge for the efficiency of health service delivery. Prolonged hospitalizations limit bed turnover, increase costs, and reduce the availability of hospital resources. In Japan, older adults must undergo a formal care needs certification process to access public long-term care (LTC) services. Initiating this process during hospitalization is considered ideal for ensuring continuity of care. However, the relationship between the timing of LTC certification applications and hospital length of stay (LOS) remains unclear.

Objective: This study examined the association between the timing of LTC certification applications—specifically those submitted during hospitalization—and average LOS among older inpatients across Japanese prefectures.

Methods: We conducted an ecological cross-sectional analysis using data from all 47 prefectures in Japan for fiscal year 2020. The exposure variable was the proportion of LTC certification applications submitted during hospitalization among all new LTC applications in each prefecture. Exposure data were sourced from the Long-Term Care Database Open Data (Kaigo DB Open Data). The outcome was average LOS among individuals aged ≥65 years at the prefectural level from the 2020 Patient Survey. Linear regression models were used to evaluate the association between the exposure and outcome variables adjusting for relevant covariates. Prefecture-level covariates included proportion of residents living alone, with cognitive decline, or with higher dependency; the proportion requiring dialysis or a respirator before application; the number of health care providers per 100 beds; and the number of nursing and care home beds per 1000 LTC recipients. Sensitivity analyses were conducted using alternative LOS data sources (eg, 2018 and 2020 Hospital Report and 2017 Patient Survey).

Results: The median proportion of in-hospital LTC certification applications was 30.5% (IQR 24.5%-36.1%). The median LOS for older adults was 40 (IQR 37-45.5; range 30-82) days. Prefectures with a higher proportion of in-hospital applications had substantially longer average LOSs. In univariate analysis, the association was statistically significant (β=0.04; P=.003), indicating that a 1% increase in in-hospital applications was associated with an approximately 2-day increase in average LOS. This association remained statistically significant after adjustment for all covariates in multivariate models (β=0.06; P=.04). Findings were consistent across sensitivity analyses.

Conclusions: Although initiating LTC certification during hospitalization is essential for supporting timely discharge, our findings indicate a positive association with extended hospital stays. This may reflect systemic delays in the certification process. Even with ideal discharge planning, such delays could extend hospitalization and lead to suboptimal allocation of health care

XSL•FO
RenderX

resources. As this study was ecological in design, the findings should be interpreted cautiously. Further individual-level data research is warranted to clarify the mechanisms and inform strategies for improving transitional care efficiency in aging populations.

(JMIR Form Res 2025;9:e76782) doi:10.2196/76782

KEYWORDS

length of hospital stay; long-term care insurance; older inpatients; ecological study; cross-sectional study; transitional care

Introduction

Background

Hospital length of stay (LOS) is an essential indicator of the efficiency of health service delivery. Optimal discharge timing can enhance patient flow, freeing up beds and health care staff time [1]. Shortening LOS is often associated with improved patient outcomes. A previous study in the United States found that reduced hospital LOS did not increase readmissions or mortality, demonstrating that efficiency improvements can maintain or enhance patient care quality [2].

Delayed discharge, sometimes called *bed blocking*, has been a serious problem in health service delivery for the past half-century [3]. Older patients are particularly at risk of delayed hospital discharge because of their complex health and social needs [4]. When a patient stays in a hospital bed longer than necessary, limited health care resources are consumed, leading to the cancellation of elective surgeries, extending waiting lists, and hindering the admission of emergency cases [5]. Delayed discharge incurs additional costs. A previous systematic review estimated the mean cost per delayed discharge at US $7020, including only hospitalization costs, after adjusting for purchasing power parity [6].

Developing a discharge plan and coordinating the care services that patients will use after discharge from the early phase of hospitalization is essential for facilitating timely discharge. Previous studies have identified causes of delayed discharge such as complexity in arranging care services after discharge, waiting for admission to institutions such as community hospitals, and challenges in nursing home care and long-term care (LTC) facilities [4,7-9]. Adequate discharge planning, including organizing discharge and postdischarge services, reduces LOS [10].

The Japanese government launched the long-term care insurance (LTCI) system in 2000 [11]. As part of Japan's universal health coverage framework, the LTCI system allows all individuals aged ≥65 years to access publicly funded LTC services, including home-based care (eg, home visit nursing and care), community-based services (eg, adult day services), and institutional care (eg, nursing homes and LTC facilities). To use these services, older adults must undergo a formal care needs certification process administered by municipal governments. This assessment evaluates physical and cognitive function to determine eligibility and assign an appropriate level of care [11,12].

Objectives

As noted previously, it is important for effective transitional care that discharge planning and coordination of postdischarge services begin during hospitalization. In this context, initiating the care needs certification process while the patient is still in the hospital is considered ideal as it may facilitate a smoother transition to LTC services. However, it remains unclear whether the timing of LTC certification applications—particularly those submitted during hospitalization—is associated with LOS. To generate hypotheses about this potential relationship, this study conducted an ecological analysis using prefectural-level data from Japan. These results are expected to provide a foundation for future research using national databases that include individual-level data.

Methods

Study Design

This ecological cross-sectional study was conducted at the prefectural level to explore the association between the timing of LTC certification applications—specifically those submitted during hospitalization—and LOS. All data used in this study were aggregated at the level of Japan's 47 prefectures and obtained from publicly available government databases. These datasets are accessible online, and detailed information, including URLs, is provided in the *Data Availability* section of this paper. The analysis was conducted in June 2024 using the most recent datasets available at that time in which the exposure and outcome data were from the same reference year.

This manuscript adheres to the Strengthening the Reporting of Observational Studies in Epidemiology guidelines for reporting cross-sectional studies.

Ethical Considerations

This study was determined to be *nonapplicable* to ethics review by the ethics committee of the National Center for Geriatrics and Gerontology (approval 1772) in accordance with Article 6, Section 1, of the Ethical Guidelines for Medical and Biological Research Involving Human Subjects in Japan. This study used only aggregated, publicly available secondary data at the prefectural level; therefore, informed consent was not required. No direct human participation took place, and no individual-level or identifiable data were used.

Measurements and Data Sources

Exposure Variable: LTC Certification Applications Submitted During Hospitalization

We focused on the timing of LTC certification applications, specifically those submitted during hospitalization. For each prefecture, we calculated the proportion of LTC certification applications submitted during hospitalization in fiscal year (FY) 2020. This proportion was calculated for each of Japan's 47 prefectures using the following formula: proportion of LTC

XSL•FO
RenderX

certification applications submitted during hospitalization = (number of applications submitted during hospitalization in FY 2020 in the prefecture)/(total number of new LTC certification applications in FY 2020 in the prefecture).

In this study, *hospitalization* refers to admissions to hospital beds in Japan, excluding LTC beds covered under medical care insurance.

These data were extracted from open LTCI claims data [13]. These open data were aggregated from the LTCI claims database for each FY beginning in 2018 and are open source and publicly accessible. The Japanese LTCI system was established in 2000 with a standardized electronic claims system. The LTCI claims database contains comprehensive national administrative data because of universal health coverage for LTC and the electronic claims system [14]. The data used for FY 2020 were publicly released in November 2023.

Outcome Variable: Average LOS Among the Population Aged ≥65 Years

Overview

Average LOS was measured using the average LOS among the population aged ≥65 years discharged from general hospitals in each prefecture in FY 2020. Data were obtained from an open government database—the Patient Survey (government statistics code: 00450022). This is a fundamental statistical survey conducted by the Japanese government to obtain basic data for medical administration by clarifying the actual situation of patients who use hospitals and clinics and their attributes, conditions at the time of visit or admission, diagnoses, and estimations of patient numbers by region [15].

Patient Survey

The target was all hospitals nationwide; those with ≥500 beds were surveyed, whereas those with 20 to 499 beds were stratified and randomly selected from the secondary medical care area, defined as a medical administrative area under the Medical Care Act in Japan. The following groups were stratified: *special functioning hospitals*, *psychiatric hospitals (hospitals with only psychiatric beds)*, *hospitals with only LTC beds*, *regional medical care support hospitals*, and *other*.

The average LOS in this survey refers to the average LOS of patients discharged from the hospital during the survey period (September 1-30, 2020). In total, 6185 medical facilities participated in the survey, and information on >1 million patients was collected. A total of 333 patients in the 2020 survey had unknown admission and discharge dates.

The Patient Survey is conducted every 3 years, and the results for FY 2020 were publicly released on June 30, 2022.

Covariates

Overview

Relevant covariates were selected based on clinical experience, data availability, and findings of previous studies. Previous studies have identified factors associated with LOS or delayed discharge, including hospital staffing [16,17]; living status [7,8]; dementia or cognitive impairment [10,18]; functional dependence [19,20]; specific clinical conditions such as requiring a feeding tube, dialysis, or respirator; and greater pressure ulcer risk [21]. A previous study in the United Kingdom suggested that increasing the supply of nursing and care at home was associated with fewer delayed discharges [22].

Hospital Staffing

This was measured as the number of health care providers per 100 beds in each prefecture in FY 2020, with data obtained from the Survey of Medical Institutions (government statistics code: 00450021).

Living Status

This was the proportion of older adults living alone in the number of households in each prefecture, which was calculated for each prefecture as follows: proportion of older adults living alone = (number of older people living alone in FY 2020 in the prefecture)/(total number of households in FY 2020 in the prefecture).

Data were extracted from the national census (government statistics code: 00200521).

Cognitive Impairment

This is the proportion of newly certified LTC recipients with cognitive decline in each prefecture, calculated using the following formula: proportion of newly certified LTC recipients with cognitive decline = (number of newly certified LTC recipients with cognitive decline in FY 2020 in the prefecture)/(total number of people newly certified for LTC needs in FY 2020 in the prefecture).

The number of individuals with cognitive decline was determined using the grade of activities of daily living (ADLs) related to dementia, which comprises 8 levels: 0, I, IIa, IIb, IIIa, IIIb, IV, and M. Cognitive decline was defined as grade I or higher based on a previous study that proposed grade I or IIa as the cutoff for identifying dementia [23]. Data were obtained from the open LTCI claims database.

In Japan, to access LTCI services, individuals must apply for certification and undergo a standardized needs assessment. One of the items assessed is the grade of ADLs related to dementia. The assessment is conducted by trained personnel commissioned by local governments typically at the individual's home, at a hospital, or in another care setting. In this study, we used aggregated prefecture-level data derived from these assessments conducted on LTCI applicants.

In this study, *LTC recipients* refers to individuals who were officially certified as requiring LTC under Japan's LTCI system regardless of whether they were actually using LTC services.

Functional Dependence

This was measured as the proportion of newly certified LTC recipients with higher levels of dependence in each prefecture and was calculated as follows: proportion of newly certified LTC recipients with higher dependence = (number of newly certified LTC recipients with higher dependence in FY 2020 in the prefecture)/(total number of people newly certified for LTC needs in FY 2020 in the prefecture).

The level of dependence was determined using the care needs level, which ranges from 1 to 5, with higher levels indicating

XSL•FO
RenderX

greater dependence [18]. A previous study reported the median Barthel index total score by care needs level as follows: score of 85 for level 1, score of 70 for level 2, score of 60 for level 3, score of 30 for level 4, and score of 20 for level 5 [24]. Generally, a Barthel index total score of ≤60 indicates dependence [25]. In addition, older adults certified at care levels of ≥3 are eligible for admission to welfare facilities providing lifelong care because of their higher dependence regarding ADLs [26]. Therefore, we defined care needs levels 3 to 5 as indicative of higher dependence.

In Japan, care needs levels are determined by municipalities based on a standardized needs assessment conducted by trained personnel commissioned by local governments. These levels reflect an individual's required level of support for daily living and determine their eligibility for as well as the amount and type of services provided under the public LTCI system.

All data used in this study were aggregated at the prefectural level and obtained from the publicly available LTCI claims database.

Specific Clinical Status

This was measured as the proportion of newly certified LTC recipients who required either dialysis or a respirator within 14 days before application and was calculated separately for each prefecture as follows: proportion of newly certified LTC recipients who required dialysis within 14 days before application = (number of newly certified LTC recipients who required daily dialysis within 14 days before application in FY 2020 in the prefecture)/(total number of people newly certified for LTC needs in FY 2020 in the prefecture) and proportion of newly certified LTC recipients who required a respirator within 14 days before application = (number of newly certified LTC recipients who required a respirator within 14 days before application in FY 2020 in the prefecture)/(total number of people newly certified for LTC needs in FY 2020 in the prefecture).

Although data on tube feeding and pressure ulcers were available, they were not used because of significant correlations with the proportion of people with higher dependence among those newly certified for LTC needs (r=0.5 and 0.7, respectively). Data were extracted from the open LTCI claims database.

Supply of Nursing and Care Home Beds

This was measured as the number of nursing and care home beds per 1000 LTC recipients in each prefecture. The calculation was based on the number of nursing and care home beds and the total number of LTC recipients in each prefecture during FY 2020: number of nursing and care home beds per 1000 LTC recipients = (number of nursing and care home beds in FY 2020 in the prefecture)/(total number of LTC recipients in FY 2020 in the prefecture) × 1000.

Data were obtained from the Survey of Institutions and Establishments for Long-Term Care (government statistics code: 00450042) and the Report Survey on the Situation of the Long-Term Care Insurance Service (government statistics code: 00450351).

Statistical Analysis

All variables were summarized using descriptive statistics, including the mean and median, SD, IQR, and ratio of minimum to maximum. A linear regression model was used to analyze the data. The dependent variable in the model was the average LOS among the population aged ≥65 years in each prefecture during FY 2020. A logarithmic transformation was applied because of the right-skewed distribution of the average LOS. The independent variables included the proportion of LTC certification applications submitted during hospitalization and other covariates. The Akaike information criterion and Bayesian information criterion were used to assess the goodness of fit for each model.

The outcome variable from the 2020 Patient Survey provides the average LOS among the population aged ≥65 years. However, there are several limitations. First, the data reflect the average LOS of patients discharged from hospitals during a specific period (September 1-30, 2020). Second, hospitalization patterns and LOS in 2020 were likely influenced by the COVID-19 pandemic. To address these limitations, we conducted sensitivity analyses using data from the 2018 and 2020 Hospital Report and the 2017 Patient Survey. Although the Hospital Report does not include the average LOS among specific age groups, it covers every hospital nationwide, with data collected monthly throughout the year. The average LOS of curative care beds is available, excluding psychiatric, infectious disease, tuberculosis, and LTC beds. This study used the average LOS of curative care beds to reduce bias derived from the differences in bed functions and focus on the most general types of beds.

A significance level of 5% was used for all tests (2-tailed). All statistical analyses were conducted using R (version 4.3.1; R Foundation for Statistical Computing).

Results

Overview

The number of LTC certification applications submitted during hospitalization, their proportion among all new LTC certification applications, and the average LOS among individuals aged ≥65 years are summarized in Table 1.

XSL•FO
RenderX

Table 1. Summary statistics of long-term care (LTC) certification applications submitted during hospitalization and hospital length of stay (LOS) among older adults by prefecture (fiscal year 2020; N=47).

Variable	Median (IQR)	Mean (SD)	Range	Maximum-to-minimum ratio
Number of LTC certification applications submitted during hospitalization	5440 (4065-8945)	8989.4 (8281.4)	1850-38,550	20.8
Proportion of LTC certification applications submitted during hospitalization (%)	30.5 (29.1-31.7)	30.5 (2.5)	24.5-36.1	1.5
LOS among individuals aged ≥65 y (d)	40 (37-45.5)	43 (10.7)	30-82	2.7

Bivariate Association

A higher proportion of LTC certification applications submitted during hospitalization was significantly associated with longer hospital stay (β=0.04, 95% CI 0.01–0.06; *P*=.003).

Multivariate Association

The multivariate association between the exposure and outcome variables is shown in Table 2. A higher proportion of LTC certification applications submitted during hospitalization was significantly associated with a longer average LOS among individuals aged ≥65 years after adjusting for covariates.

The results of the sensitivity analyses using data from the Hospital Report (2018 and 2020) and the Patient Survey (2017) are provided in Tables S1-S3 in Multimedia Appendix 1, supporting the findings of the primary analysis.

Table 2. Results of the multivariate linear regression analysis of the association between long-term care (LTC) certification applications submitted during hospitalization and average hospital length of stay among older adultsa.

Characteristic	β (95% CI)	*P* value
Proportion of LTC certification applications submitted during hospitalization	0.06 (0.00 to 0.12)	.04
Number of health care providers per 100 beds		
Nurses	−0.04 (−0.12 to 0.03)	.30
Rehabilitation therapists	0.04 (−0.03 to 0.10)	.20
Proportion of older adults living alone	0.05 (−0.02 to 0.13)	.20
Proportion of newly certified LTC recipients with cognitive decline	0.05 (−0.03 to 0.12)	.20
Proportion of newly certified LTC recipients with higher dependence	−0.03 (−0.09 to 0.03)	.40
Proportion of newly certified LTC recipients who required dialysis within 14 d before application	0.07 (0.00 to 0.13)	.049
Proportion of newly certified LTC recipients who required a respirator within 14 d before application	0.02 (−0.05 to 0.08)	.60
Number of nursing and care home beds per 1000 LTC recipients	−0.02 (−0.09 to 0.04)	.40

[a]Average length of stay was logarithmically transformed.

Discussion

Principal Findings

Our results indicate that a higher proportion of LTC certification applications submitted during hospitalization was associated with a longer average LOS at the prefecture level. The primary analysis used average LOS data from the 2020 Patient Survey as the outcome variable. However, relying solely on this data source may introduce bias due to its data collection methods and period.

To address this potential bias, we conducted sensitivity analyses using average LOS data from multiple sources, including the Hospital Report (2018 and 2020) and the 2017 Patient Survey. These sensitivity analyses supported the primary results and enhanced the robustness of our findings, reinforcing the observed association between the timing of certification applications and average LOS.

Comparison With Previous Work

Discharge planning and coordination of postdischarge services starting during hospitalization are essential for effective transitional care. A meta-analysis has shown that such planning, including the arrangement of postdischarge services, can reduce LOS [10]. However, contrary to this expectation, our findings showed that prefectures with a higher proportion of in-hospital LTC certification applications had longer average LOS.

Possible Explanation: Systemic Delays in the LTC Certification Process

One possible explanation for our findings is that this reflects delays in the LTC certification process itself. In recent years, significant delays have been reported in many municipalities [27]. Several news sources have highlighted that these delays are driven by a rapid increase in the number of applications, which has strained administrative capacity. Indeed, Japan's increasing aging population continues to drive up the demand for LTC certification [28]. The care needs level for each applicant is determined by a municipal committee based on multiple sources, including face-to-face assessments of physical

XSL•FO
RenderX

and cognitive function [11,12,29]. These assessments are time and labor intensive, and as the volume of applications increases, so does the workload for investigators, which may further prolong the certification process. Therefore, delays in certification may hinder the timely arrangement of postdischarge LTC services, potentially contributing to extended hospital stays. Previous studies have noted that discharge delays are often related to challenges in arranging postacute care services [9]. Therefore, although initiating the LTC certification process during hospitalization is generally considered a good practice to support a smooth transition to community-based care, this intended benefit may not be realized when systemic delays are present.

Alternative Explanations and Considerations

Although our model accounted for several relevant covariates, some unmeasured or residual confounding likely remains, and causal inference is limited by the study's ecological design. There are 3 alternative explanations and considerations.

First, the number of in-hospital certification applications may simply reflect the presence of patients who are expected to have longer hospital stays. For example, those with multiple chronic conditions, functional dependence, or cognitive impairment or those living alone are more likely to need LTC services after discharge [7-9,19-21]. Because of their longer stays, these patients may have more opportunities to initiate the certification process while still hospitalized. Given the ecological nature of our study, reverse causality cannot be ruled out. Therefore, further investigation using individual-level data is essential to clarify the relationship between the timing of LTC certification application and hospital LOS.

Second, the inclusion of rehabilitation wards in the hospital dataset used to calculate average LOS and in-hospital LTC certification application rates may help explain the observed association. In Japan, these facilities typically provide longer inpatient care than acute care hospitals. For example, patients with stroke may stay for up to 150 days, with an average LOS of 67.5 days compared to 29.5 days in acute care hospitals [30]. In such settings, patients are generally in stable medical condition, and the extended LOS may increase opportunities to initiate LTC certification during hospitalization. Thus, prefectures with a higher concentration of rehabilitation wards may tend to have both longer LOSs and a higher proportion of in-hospital LTC certification applications.

Third, prefectures with longer LOSs may also have fewer available institutional care beds. In such areas, LTC certification applications may be submitted during hospitalization out of necessity, but discharge may be delayed due to a lack of available placements. Waiting for admission to a nursing home has been identified as a key factor contributing to prolonged hospital stays in previous studies [7-9,22]. Japan faces serious shortages in institutional LTC capacity. A document in 2022 reported that 275,000 people were on waiting lists for public nursing homes [31]. Therefore, variation in LTC resource availability across prefectures may help explain the association observed in this study.

Limitations

This study has some limitations. As an ecological study, confounder adjustment may not have been sufficient because of unmeasured confounders, particularly at the individual level. Consequently, the findings cannot be generalized to individual-level associations because of the potential for ecological fallacy. For example, patients with complex care needs may have longer hospital stays and may also be more likely to submit LTC certification applications during hospitalization. Therefore, these results should be interpreted with caution, and further analyses using individual-level data are strongly warranted.

Policy Implications

To our knowledge, this is the first study to examine the association between LTC certification application during hospitalization and the average length of hospital stay, offering important insights for policymakers aiming to strengthen the continuity of transitional care in aging societies such as Japan. Our findings indicate that initiating LTC certification during hospitalization is associated with longer hospital stays, which may reflect systemic delays that hinder discharge coordination and lead to inefficient use of hospital resources. Therefore, prefectures with particularly high proportions of in-hospital applications may benefit from targeted interventions to accelerate the certification process.

Although we used the most recently available data at the time of analysis (June 2024), there was an inevitable time lag between data collection and public release—for example, the 2020 open LTCI claims data were published in November 2023, and the 2020 Patient Survey was published in June 2022. Importantly, at a June 2025 meeting of the Ministry of Health, Labour, and Welfare's LTCI subcommittee, hospital representatives reported that delays in the LTC certification process may be contributing to unnecessary extensions of hospital stays [32]. This contemporary policy discussion underscores the continued relevance of our findings despite the temporal gap.

Several digital transformation initiatives are already underway in Japan to address these challenges [33,34]. These include artificial intelligence–assisted assessment tools; digitized application workflows; remote operation of certification review meetings; and the development of shared, interoperable information platforms among hospitals, insurers (municipalities), LTC service providers, and care recipients. Such innovations have the potential not only to improve the efficiency and responsiveness of the LTC system but also to enhance health care delivery by facilitating timelier discharge planning and reducing unnecessary hospital stays.

Conclusions

Our study suggests a possible association between LTC certification application during hospitalization and hospital LOS at the prefectural level in Japan. These findings may reflect underlying system-level factors related to discharge coordination and administrative processes. Further research using individual-level data is needed to clarify the mechanisms and examine how LTC certification application timing may influence discharge outcomes. Policymakers in aging societies may benefit

XSL•FO
RenderX

from considering these findings as they work toward improving the efficiency and continuity of transitional care.

Acknowledgments

This study was supported by the Research Funding for Longevity Sciences from the National Center for Geriatrics and Gerontology (23-1 and 24-8). Generative artificial intelligence (ChatGPT) was used to modify English sentences when translating from Japanese into English [40]. The final English-language sentences were reviewed by an English-language editing company. The authors would like to thank Editage [41] for the English-language editing. The funder had no role in the study design; collection, analysis, and interpretation of the data; writing of the report; or decision to submit the paper for publication.

Data Availability

The datasets analyzed during this study can be accessed through the following publicly available resources. All data used in this study are open source and publicly accessible through government repositories [13,15,35-39].

Authors' Contributions

NT, JO, and SO contributed to the conceptualization and design of the study. NT, JO, SO, TO, and MF were responsible for the acquisition, analysis, and interpretation of data. NT drafted the original manuscript. JO, SO, TO, and MF critically reviewed and revised the manuscript for important intellectual content. NT and SO conducted the statistical analysis. SO and JO provided supervision of the research process.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Results of the sensitivity analyses.
[DOCX File , 23 KB - formative_v9i1e76782_app1.docx]

References

1. Health at a glance 2023. Organization for Economic Cooperation and Development. URL: https://www.oecd.org/en/publications/health-at-a-glance-2023_7a7afb35-en.html [accessed 2024-07-22]
2. Kaboli PJ, Go JT, Hockenberry J, Glasgow JM, Johnson SR, Rosenthal GE, et al. Associations between reduced hospital length of stay and 30-day readmission rate and mortality: 14-year experience in 129 Veterans Affairs hospitals. Ann Intern Med 2012 Dec 18;157(12):837-845. [doi: 10.7326/0003-4819-157-12-201212180-00003] [Medline: 23247937]
3. Rubin SG, Davies GH. Bed blocking by elderly patients in general-hospital wards. Age Ageing 1975 Aug;4(3):142-147. [doi: 10.1093/ageing/4.3.142] [Medline: 1211299]
4. Majeed MU, Williams DT, Pollock R, Amir F, Liam M, Foong KS, et al. Delay in discharge and its impact on unnecessary hospital bed occupancy. BMC Health Serv Res 2012 Nov 20;12(1):410 [FREE Full text] [doi: 10.1186/1472-6963-12-410] [Medline: 23167656]
5. Mohammed A, Thomas B, Hullin M, McCreath S. Audit of orthopaedic bed utilisation. Health Bull (Edinb) 2001 Nov;59(6):353-355. [Medline: 12661384]
6. Landeiro F, Roberts K, Gray A, Leal J. Delayed hospital discharges of older patients: a systematic review on prevalence and costs. Gerontologist 2019 Mar 14;59(2):e86-e97. [doi: 10.1093/geront/gnx028] [Medline: 28535285]
7. Victor CR, Healy J, Thomas A, Seargeant J. Older patients and delayed discharge from hospital. Health Soc Care Community 2000 Nov;8(6):443-452. [doi: 10.1046/j.1365-2524.2000.00270.x] [Medline: 11560715]
8. Moore G, Hartley P, Romero-Ortuno R. Health and social factors associated with a delayed discharge amongst inpatients in acute geriatric wards: a retrospective observational study. Geriatr Gerontol Int 2018 Apr 12;18(4):530-537. [doi: 10.1111/ggi.13212] [Medline: 29230961]
9. Aaltonen M, El Adam S, Martin-Matthews A, Sakamoto M, Strumpf E, McGrail K. Dementia and poor continuity of primary care delay hospital discharge in older adults: a population-based study from 2001 to 2016. J Am Med Dir Assoc 2021 Jul;22(7):1484-92.e3. [doi: 10.1016/j.jamda.2020.11.030] [Medline: 33358723]
10. Gonçalves-Bradley DC, Lannin NA, Clemson L, Cameron ID, Shepperd S. Discharge planning from hospital. Cochrane Database Syst Rev 2022 Feb 24;2(2):CD000313 [FREE Full text] [doi: 10.1002/14651858.CD000313.pub6] [Medline: 35199849]
11. Tsutsui T, Muramatsu N. Care-needs certification in the long-term care insurance system of Japan. J Am Geriatr Soc 2005 Mar 24;53(3):522-527. [doi: 10.1111/j.1532-5415.2005.53175.x] [Medline: 15743300]

XSL•FO
RenderX

12. Yamada M, Arai H. Long-term care system in Japan. Ann Geriatr Med Res 2020 Sep;24(3):174-180 [FREE Full text] [doi: 10.4235/agmr.20.0037] [Medline: 32829572]
13. Long-term care insurance claim open data. The Ministry of Health, Labour and Welfare. URL: https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/hukushi_kaigo/kaigo_koureisha/nintei/index_00009.html [accessed 2024-07-24]
14. Jin X, Tamiya N. The use of Japanese long-term care insurance claims in health services research: current status and perspectives. Glob Health Med 2021 Jun 30;3(3):142-148 [FREE Full text] [doi: 10.35772/ghm.2021.01000] [Medline: 34250289]
15. The patient survey (government statistics code: 00450022). Portal site for Japanese Government Statistics. URL: https://www.e-stat.go.jp/en/statistics/00450022 [accessed 2025-05-29]
16. DeVolder R, Serra-Sastre V, Zamora B. Examining the variation across acute trusts in patient delayed discharge. Health Policy 2020 Nov;124(11):1226-1232. [doi: 10.1016/j.healthpol.2020.06.017] [Medline: 32712011]
17. McHugh MD, Aiken LH, Sloane DM, Windsor C, Douglas C, Yates P. Effects of nurse-to-patient ratio legislation on nurse staffing and patient mortality, readmissions, and length of stay: a prospective study in a panel of hospitals. Lancet 2021 May 22;397(10288):1905-1913 [FREE Full text] [doi: 10.1016/S0140-6736(21)00768-6] [Medline: 33989553]
18. Hasegawa K, Tsukahara T, Nomiyama T. Associations between long-term care-service use and service- or care-need level progression: a nationwide cohort study using the Japanese Long-Term Care Insurance Claims database. BMC Health Serv Res 2023 Jun 05;23(1):577 [FREE Full text] [doi: 10.1186/s12913-023-09615-0] [Medline: 37277778]
19. Bo M, Fonte G, Pivaro F, Bonetto M, Comi C, Giorgis V, et al. Prevalence of and factors associated with prolonged length of stay in older hospitalized medical patients. Geriatr Gerontol Int 2016 Mar;16(3):314-321 [FREE Full text] [doi: 10.1111/ggi.12471] [Medline: 25752922]
20. Ferreira MS, de Melo Franco FG, Rodrigues PS, da Silva de Poli Correa VM, Akopian ST, Cucato GG, et al. Impaired chair-to-bed transfer ability leads to longer hospital stays among elderly patients. BMC Geriatr 2019 Mar 21;19(1):89 [FREE Full text] [doi: 10.1186/s12877-019-1104-4] [Medline: 30898161]
21. Turcotte LA, Perlman CM, Fries BE, Hirdes JP. Clinical predictors of protracted length of stay in Ontario complex continuing care hospitals. BMC Health Serv Res 2019 Apr 05;19(1):218 [FREE Full text] [doi: 10.1186/s12913-019-4024-2] [Medline: 30953489]
22. Gaughan J, Gravelle H, Siciliani L. Testing the bed-blocking hypothesis: does nursing and care home supply reduce delayed hospital discharges? Health Econ 2015 Mar 11;24 Suppl 1(Suppl 1):32-44 [FREE Full text] [doi: 10.1002/hec.3150] [Medline: 25760581]
23. Noda H, Yamagishi K, Ikeda A, Asada T, Iso H. Identification of dementia using standard clinical assessments by primary care physicians in Japan. Geriatr Gerontol Int 2018 May 16;18(5):738-744. [doi: 10.1111/ggi.13243] [Medline: 29336097]
24. Matsuda T, Iwagami M, Suzuki T, Jin X, Watanabe T, Tamiya N. Correlation between the Barthel Index and care need levels in the Japanese long-term care insurance system. Geriatr Gerontol Int 2019 Nov 20;19(11):1186-1187. [doi: 10.1111/ggi.13777] [Medline: 31746527]
25. Sinoff G, Ore L. The Barthel activities of daily living index: self-reporting versus actual performance in the old-old (> or = 75 years). J Am Geriatr Soc 1997 Jul 27;45(7):832-836. [doi: 10.1111/j.1532-5415.1997.tb01510.x] [Medline: 9215334]
26. Konishi T, Inokuchi H, Yasunaga H. Services in public long-term care insurance in Japan. Ann Clin Epidemiol 2024;6(1):1-4 [FREE Full text] [doi: 10.37737/ace.24001] [Medline: 38605915]
27. 8th medical care and long-term care working group agenda. Cabinet Office, Government of Japan. URL: https://www8.cao.go.jp/kisei-kaikaku/kisei/meeting/wg/2310_04medical/240314/medical08_agenda.html [accessed 2024-07-22]
28. 115th meeting of long-term care insurance subcommittee in the Social Security Council. The Ministry of Health, Labour and Welfare. URL: https://www.mhlw.go.jp/content/12300000/001349395.pdf [accessed 2025-05-29]
29. The text for the members of care-need certification committees: a revised edition. The Ministry of Health, Labour and Welfare. URL: https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/hukushi_kaigo/kaigo_koureisha/nintei/index.html [accessed 2024-07-22]
30. Kinoshita S, Abo M, Okamoto T, Miyamura K. Transitional and long-term care system in Japan and current challenges for stroke patient rehabilitation. Front Neurol 2021 Jan 11;12:711470 [FREE Full text] [doi: 10.3389/fneur.2021.711470] [Medline: 35087461]
31. The current situation of applicants for long-term care welfare facilities. The Ministry of Health, Labor and Welfare. URL: https://www.mhlw.go.jp/content/12304250/001029178.pdf [accessed 2025-04-17]
32. Minutes of the 122nd meeting. Ministry of Health, Labour and Welfare, Japan. URL: https://www.mhlw.go.jp/stf/newpage_60016.html [accessed 2025-08-08]
33. Reiwa 6 nendo ICT/AI o mochiita y'kaigo nintei shinsa no arikata ni kansuru ch-sa kenky. NTT Data Institute of Management Consulting. URL: https://www.nttdata-strategy.com/services/lifevalue/docs/r06_121_01jigyohokokusho.pdf [accessed 2025-08-04]
34. Y'kaigo nintei gy'mu no k'ritsuka ni muketa jissh ni kakaru purojekuto jisshi h'kokusho. NTT Data Institute of Management Consulting. URL: https://www.city.oita.oita.jp/o081/documents/r6kaigodxhoukokusyo.pdf [accessed 2025-08-04]
35. Survey of Medical Institutions (government statistics code: 00450021). Portal site for Japanese Government Statistics. URL: https://tinyurl.com/m2kk2vpv [accessed 2025-08-29]

XSL•FO
RenderX

36. National Census (government statistics code: 00200521). Portal site for Japanese Government Statistics. URL: https://tinyurl.com/ye2uceux [accessed 2025-08-29]
37. Survey of Institutions and Establishments for Long-term Care (government statistics code: 00450042). Portal site for Japanese Government Statistics. URL: https://tinyurl.com/3cwr879s [accessed 2025-08-29]
38. Report Survey on the Situation of Long-term Care Insurance Service (government statistics code: 00450351). Portal site for Japanese Government Statistics. URL: https://www.e-stat.go.jp/stat-search/files?page=1&layout=datalist&toukei=00450351&tstat=000001031648&cycle=8&tclass1=000001169206&tclass2val=0 [accessed 2025-08-29]
39. Hospital Report (government statistics code: 00450023). Portal site for Japanese Government Statistics. URL: https://www.e-stat.go.jp/stat-search/files?page=1&toukei=00450023&tstat=000001030749 [accessed 2025-08-29]
40. ChatGPT. URL: https://chatgpt.com/ja-JP/overview?openaicom_referred=true [accessed 2025-09-08]
41. Home page. Editage. URL: https://www.editage.jp/homepage [accessed 2025-05-29]

Abbreviations

ADL: activity of daily living
FY: fiscal year
LOS: length of stay
LTC: long-term care
LTCI: long-term care insurance

Edited by A Mavragani; submitted 08.05.25; peer-reviewed by F Yeh, LF Belo; comments to author 21.07.25; revised version received 08.08.25; accepted 17.08.25; published 08.09.25.

Please cite as:
Takashi N, Onishi J, Fujisawa M, Ohura T, Ohtera S
Association Between In-Hospital Applications for Long-Term Care Services and Hospital Length of Stay Among Older Adults: Ecological Cross-Sectional Study
JMIR Form Res 2025;9:e76782
URL: https://formative.jmir.org/2025/1/e76782
doi:10.2196/76782
PMID:

©Naoki Takashi, Joji Onishi, Misaki Fujisawa, Tomoko Ohura, Shosuke Ohtera. Originally published in JMIR Formative Research (https://formative.jmir.org), 08.09.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Neuroaffirmative, Self-Determination Theory–Based Psychosocial Intervention for Adults With Attention-Deficit/Hyperactivity Disorder: Randomized Feasibility Study

Rebecca Elizabeth Champ[1], BA, MA, PhD; Rita Wengorovius Meneses[2], LicPs, PhD; Marios Adamou[1*], MD, MSc, MA, LLM, PhD; Warren Gillibrand[1*], RN, MSc, PhD; Sally Arrey[1*], RLDN, MSc, PhD; Barry Tolchard[3*], RLDN, MSc, PhD

[1]University of Huddersfield, Huddersfield, United Kingdom

[2]Marketing Aplicado Lda (Portugal), Lisbon, Portugal

[3]University of Teeside, North Yorkshire, Middlesbrough, United Kingdom

[*]these authors contributed equally

Corresponding Author:
Rebecca Elizabeth Champ, BA, MA, PhD
University of Huddersfield
Queensgate
West Yorkshire
Huddersfield, HD1 3DH
United Kingdom
Phone: 44 1484 422288
Email: rebecca@rebeccachamp.co.uk

Related Article:

This is a corrected version. See correction statement: https://formative.jmir.org/2025/1/e87679

Abstract

Background: Neurodevelopmental disorders are complex and heterogeneous, creating challenges for treatment design. Multiple syndromes are associated with executive function (EF) deficits; however, theories of attention-deficit/hyperactivity disorder (ADHD) centralize a singular perspective of outcomes arising from EF impairments in adults. Deficit-based etiologies state ADHD-related EF impairments interfere with agentic self-development, perspectives that may inadvertently contribute to social stigma and influence neurotype dysphoria in ADHD identity construction. Challenges to this perspective highlight heterogeneity, context variability, the absence of a single EF deficit of origin, correlational neuroimaging data, and limited investigation into altered brain activity in ADHD research. Recommendations for psychosocial interventions primarily support cognitive behavioral therapy, which centers on a deficit-based etiology of ADHD and prioritizes symptom reduction and cognitive control of self-regulation as treatment outcomes—skills that require additional cognitive effort and may involve avoidance of emotional experiences to minimize negative affect. Transdiagnostic approaches are recommended to gain new insights into mental health challenges. Self-Determination Theory (SDT) presents a transdiagnostic approach that offers alternative outcomes by prioritizing basic psychological need satisfaction, which supports strong identity formation, motivation, and self-regulation.

Objective: This study examines the feasibility and effects of an SDT-based quality-of-life therapeutic intervention for adults with ADHD.

Methods: We aimed to recruit 30 participants aged 18 years or older with a confirmed diagnosis of ADHD and access to an internet connection. Participants were recruited from the Adult ADHD Clinic at the South West Yorkshire Partnership National Health Service Foundation Trust and allocated through 4-block randomization by a nonblinded researcher to either an 11-session therapeutic coaching intervention (n=11) or a control waitlist (n=9) condition. Feasibility was evaluated by pretreatment and posttreatment measurements of health-related quality of life (QoL), psychological distress, ADHD symptoms, ADHD-related

XSL•FO
RenderX

QoL, self-reflection and insight, autonomous functioning, and individual outcome measures of impairment. Participants also responded to a qualitative feedback interview question on intervention value.

Results: Adherence was high for both intervention completion (10/11, 91.6%) and control condition completion (9/11, 81.8%). Results showed clinically significant improvement on measures of psychological distress, specifically in the subscales of problems (z=0.0; P=.01), nonrisk (z=2.0; P=.01), functions (z=5.0; P=.02), and well-being (z=6.0; P=.03), as well as ADHD symptoms (z=3.0; P≤.01), particularly inattention (z=3.0; P≤.01), outcomes not specifically targeted by the intervention. Additional clinically significant findings of improvement in QoL, specifically in outlook subscale (z=21; P=.67), reduction of distress in problems identified in the individual outcome measure and the need for self-reflection subscale of self-reflection for the control group (z=1.0; P=.05) indicate potential positive effectiveness despite the impact of COVID-19. Positive qualitative feedback on usefulness and transferability of the intervention was provided by 90% (20/23) of participants.

Conclusions: This study suggests that a randomized controlled trial of an SDT-based psychosocial intervention with nondeficit-based outcomes for adults with ADHD is feasible and recommended.

Trial Registration: ClinicalTrials.gov NCT04832737; https://clinicaltrials.gov/study/NCT04832737

(JMIR Form Res 2025;9:e69943) doi:10.2196/69943

KEYWORDS

ADHD; adult; attention-deficit/hyperactivity disorder; psychotherapy; self-determination theory; treatment

Introduction

Overview

Neurodevelopmental disorders are a category of mental health conditions defined by the *DSM-5-TR* (*Diagnostic and Statistical Manual of Mental Disorders* [Fifth Edition, Text Revision]) [1] and include attention-deficit/hyperactivity disorder (ADHD), autism spectrum disorder, neurodevelopmental motor disorders, including tic disorders, intellectual disability, communication disorders, and specific learning disorders. A shared characteristic of neurodevelopmental disorders is atypical brain development that generates impairments in cognition, communication, behavior, and motor skills. Within this heterogeneous category [2,3], the diagnostic and therapeutic approach to ADHD is overwhelmingly governed by a theoretical framework that posits executive function (EF) deficits as the origin of impairment [4-7]. This EF deficit model proposes that challenges in metacognitive, emotional, and behavioral management are critical factors impeding the sustained goal-directed activities necessary for foundational development. Success, according to these theories, is predicated on recognition and support from peers and authority figures, which are essential for achieving both self-actualization and societal functioning [5,8,9].

This dominant EF deficit model, however, faces significant and growing scientific challenges. A primary criticism is that no single EF deficit has been identified as sufficient to cause ADHD [10]. Furthermore, EF impairments themselves are not unique to the condition and can vary widely between individuals depending on context [10-12]. The neurobiological evidence is also less definitive than often assumed; neuroimaging research has not shown structural differences of a magnitude that is significant compared with controls, and the data remain correlational, unable to establish a definitive causal link [13-15]. Notably, when neuroimaging does reveal altered brain activity in ADHD participants, such as the recruitment of different response pathways, these variations are frequently categorized simply as “abnormal” rather than being explored as potentially valid alternative modes of neural organization [6,16]. This conceptual fragility calls the very foundation of current diagnostic protocols into question.

The pervasive focus on deficits has important iatrogenic consequences. The closed-label nature of the ADHD diagnosis, combined with negative narratives and stereotypes often promoted by the media, means many individuals encounter judgmental responses that alter their social treatment [17-23]. This social stigma can become internalized as ableism and neurotype dysphoria, particularly as the need for accommodations leads individuals to identify with diagnostic criteria and seek a “cure” for their inherent neurotype behaviors [24]. This process can be unintentionally reinforced by professional guidance based on the deficit model, as articulated by authorities such as Barkley [25]. This perspective frames ADHD as a chronic, incurable condition in which medication is the only effective treatment to normalize EF, functionality is dependent on external scaffolding, and any strengths are attributed to individual talent rather than recognized as potential aspects of the neurotype itself [7,26-31]. Consequently, treatment recommendations from bodies such as the National Institute for Health and Care Excellence prioritize pharmacotherapy, which, while beneficial for many, often leaves patients with significant residual symptoms and functional impairment [15,32].

The primary evidence-based nonpharmacological treatment, cognitive behavioral therapy (CBT), is also rooted in this deficit model [7,15,32,33]. While recent systematic reviews and meta-analyses show that CBT approaches can improve core symptoms and quality of life (QoL), their recommendations remain cautious due to methodological limitations such as diverse protocols, small sample sizes, and high risk of bias [34-37]. The central aim of CBT for ADHD is to strengthen cognitive abilities, increase awareness of behavior, and reframe maladaptive schemas through cognitive reappraisal [6,7]. This approach, however, presents a fundamental paradox: it demands “effortful coping” [38] and a high cognitive load to manage emotions and behavior; yet, the condition is itself characterized by effort avoidance [39] and inability to allocate sufficient

XSL•FO
RenderX

cognitive effort [40]. Moreover, emerging research highlights the possibility that the ADHD neurobiological processing style may not be amenable to the reinforcement learning models upon which CBT is based [41].

Given the conceptual and practical limitations of the current paradigm, there is growing support for transdiagnostic approaches to gain new perspectives on mental health difficulties [42,43]. For ADHD specifically, recommendations are being made for transdiagnostic models due to the cross-disorder nature of EF impairments, as well as the condition's clinical heterogeneity and context variability [3,44,45], and current theories are criticized for isolating domains of functioning and therefore lacking dimensionality and an integrative approach [46,47]. Self-Determination Theory (SDT) presents a robust, empirically based transdiagnostic framework for understanding psychopathology [21]. A mini-theory of SDT, Basic Psychological Needs Theory, posits that the satisfaction of 3 universal psychological needs—autonomy, competence, and relatedness—is essential for growth, well-being, and organismic integration. This process provides the energy for the development of an agentic self [22-24]. Research shows that the satisfaction of these needs supports mature identity formation and contributes to quality of motivation for long-term goals [48]. Conversely, the frustration or thwarting of these needs, predicts problem behaviors, increases risk of psychopathology, and can forestall identity development [21,26,27,48,49].

Recent research suggests that this experience of need frustration may function as an underlying transdiagnostic mechanism that can explain diverse forms of psychopathology and their comorbidity [50]. Some studies have examined ADHD behaviors and motivation through an SDT lens, particularly within university environments [51-56]. However, these applications have remained tethered to a deficit model, invariably using SDT as a tool to address symptom management. At the theoretical level, ADHD etiology has been interpreted as a manifestation of need frustration and impairment of internalization [6,50,57], yet no studies involving practical application of this nondeficit perspective have been published. This reveals a clear paucity of research in the area. Therefore, this study aims to address this critical gap by using SDT as an alternative theoretical foundation to explore ADHD expression and support outside the confines of the deficit-based paradigm.

The Autonomy, Design, Awareness, Psychoeducation, and Training Integration for Sustainable Change Framework

Champ et al [6] presents a neuroaffirmative etiology of ADHD based in SDT, describing ADHD behaviors as neurobiologically altered approaches to processing and task engagement. This nondeficit model of natural ADHD behaviors based on neurodivergent neurobiological needs provides a nonstigmatizing foundation for self-regulatory functioning. Combining this model with the understanding of the polar nature of the interaction of ADHD consciousness and the environment as described in the Creative Awareness Theory [58] creates a new framework for understanding ADHD lived experience, identity formation, and self-regulation. The Creative Awareness Theory provides both practitioner and client with a positive model of unskilled attempts at self-regulation, forming an active guide to interpret existing strategies and facilitate development of awareness and self-management skills. Using this framework, it is possible to shift the focus from EF deficits and interpret ADHD psychopathology as a history of fundamental misunderstandings of ADHD motivation and engagement processes resulting in impaired internalization, need frustration, thwarting and neglect, and subsequent development of maladaptive identities, coping strategies, and need substitutes. This generates significant challenges to organismic integration by impairing connection with an authentic inner compass (AIC) [59,60]. Based in SDT, the AIC is defined as the feeling and perception of what is truly important for us—voluntary and intrinsic self-guiding preferences including values, life aspirations, interests, and goals that feel authentic and become long term as we mature. Lack of confidence and confusion regarding these preferences can impact the ability to make choices, resulting in feeling incapable of true self-direction. Research indicates that active and reflective formation of a strong sense of the AIC demonstrates an understanding of authentic core preferences expressed as agency, leading to experiences of autonomy, growth, resistance to peer pressure, and resilience [59-61]. The ADAPT (Autonomy, Design, Awareness, Psychoeducation, and Training Integration for Sustainable Change) framework aims to support individuals with ADHD using a multimodal psychotherapeutic approach to increase self-awareness of their unique neurobiology, develop their AIC, understand their basic psychological needs and needs based on their neurobiological differences, and support internalization of identity commitments. This foundation will facilitate task and environmental engagement, increase motivational activation, and feelings of confidence in their ability to design strategies to meet their needs, manage self-regulation, and develop life-crafting skills in a variety of contexts [62-64]. It is hypothesized that a neuroaffirmative ADHD treatment program that introduces the above framework will reduce symptoms, demonstrate changes in specific psychological difficulties, improve self-awareness, evidence personal experience of change, increase feelings of autonomy, and improve QoL.

Aims and Objectives

To progress to a randomized controlled trial (RCT), it is critical to identify the most appropriate outcome measures for an SDT-based intervention for adults with ADHD. To provide clear guidance on good research conduct specifically for pilot and feasibility studies in preparation for an RCT assessing intervention or therapy effectiveness, Eldridge et al [65] developed an extension to the 2010 CONSORT (Consolidated Standards of Reporting Trials) guidelines for RCTs. This framework aligns with the United Kingdom Medical Research Council guidance on complex interventions and the National Institute for Health and Care Research definition of pilot studies. Throughout this study, the design will be referred to as a randomized feasibility study, and therefore this framework will be used as guidance. In maintaining this standard, the extension for nonpharmacologic treatment interventions to the 2010 CONSORT guidelines for reporting has been added (Multimedia Appendix 1) [66].

XSL•FO
RenderX

As RCTs are still upheld as vital for informing policy decisions [67], the call for transdiagnostic-based perspectives provides a good opportunity to offer the ADAPT framework as a novel nonpharmacological ADHD treatment approach. Therefore, this study aims to examine the feasibility, acceptability, and potential effectiveness of a randomized feasibility study evaluating a novel SDT-based program of therapeutic self-development, psychoeducation, and skills training for adults with a diagnosis of ADHD. The study objectives were to evaluate the feasibility of delivering an 11-session online self-development therapeutic intervention to a group of adults with ADHD, accounting for attrition rates of recruited participants; evaluate the acceptability of randomization to an adult ADHD population for a therapeutic self-development intervention; evaluate the acceptability of multiple measures to an adult ADHD population, including SDT-based measures for autonomy and self-reflection; evaluate the most appropriate outcome measure for the ADAPT framework for adults with ADHD; and evaluate the potential effectiveness of the ADAPT framework on ADHD symptoms, QoL, self-awareness, autonomous functioning, and personal experience of change.

Methods

Sample

Sample sizes for feasibility studies are much debated [68], and recommendations vary from 10-12 per group to 60-75 per group depending on study objectives [69]. We consulted with a University of Huddersfield statistician; however, this did not result in formal sample size calculation recommendations. To achieve ethical approval and with reference to the rule of thumb for a medium to large effect size (0.3<0.7), we set the sample at two groups of 10. In anticipation of dropouts, we aimed to recruit 30 participants.

ADAPT Framework Intervention

Eleven participants received 11 sessions of online, individually focused therapeutic coaching not currently accessible within the service (Multimedia Appendix 2). Treatment included one 2-hour assessment session exploring personal challenges and lived experience of ADHD, including a foundational section in neuroaffirmative SDT-based psychoeducation focused on neurobiological responses to environmental engagement [6], followed by ten 1-hour sessions of therapeutic coaching. Excluding session 1 of 10, which focused on time management skills for immediate and overinclusive processing styles [6], each session centralized autonomy-supportive, client-led problem identification, facilitating development of self-awareness of motivational factors in task initiation and engagement, application of context-oriented strategy development, and support for neurobiological and basic psychological needs. The researcher delivering the intervention was a psychotherapist and coach with 13 years' experience in a specialist ADHD private practice. The intervention was supervised by a British Association for Counselling and Psychotherapy accredited psychotherapist to review casework and measure progress. Figure 1 summarizes the flow of participants through the intervention.

Figure 1. Overview of the ADAPT (Autonomy, Design, Awareness, Psychoeducation, and Training Integration for Sustainable Change) framework pilot study: assessment timeline and frequency of intervention sessions.

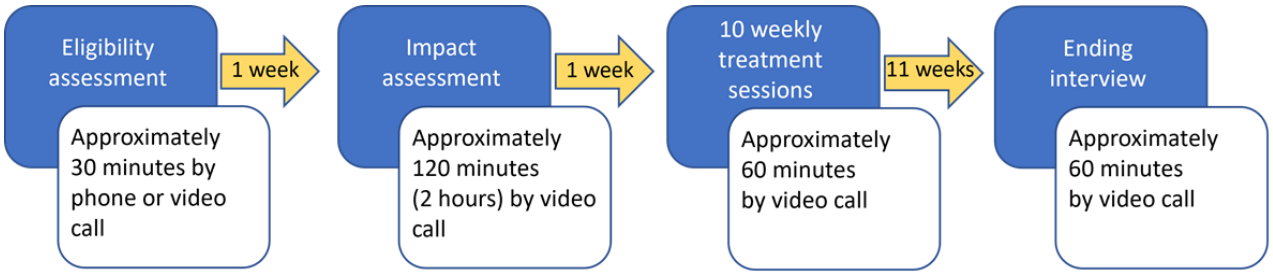

Waitlist Control Group

Nine participants who met the same inclusion criteria and were assessed using the same methodology also enrolled in the therapeutic self-development intervention after a 12-week wait.

Procedure

The design selected was a randomized, controlled study with a 1:1 allocation ratio to 2 small independent groups (control and intervention), a longitudinal design (pretreatment and posttreatment), and repeated online measures or surveys. Participants randomly assigned to intervention group and waitlist control group were assessed at pretreatment and posttreatment, and within-treatment measures were assessed during the intervention only. Assessment measures at intervention initiation and completion consisted of one measurement before and one after for both groups, while within-session measures increased the number of assessment moments to 10 for each session. Due to COVID-19 restrictions for research requiring in-person contact, all interviews, screening, data collection, and treatment sessions were conducted online using a National Health Service (NHS)–approved video platform.

Clinical Measures

EQ-5D-5L

The EQ-5D-5L questionnaire [47] is a self-rated scale measuring health-related QoL in adults, used to assess treatment effect before and after treatment by measuring gains or losses in reported health status. It produces a 5-digit health state profile representing the level of reported problems on 5 dimensions of health, with lower ratings indicating better health states. This generates a health state profile, and each health state can be assigned a summary index score based on societal preference weights, or "utilities," for that health state.

XSL•FO
RenderX

Clinical Outcomes in Routine Evaluation–Outcome Measure

Participants' self-reported awareness of psychological distress was measured with the Clinical Outcomes in Routine Evaluation–Outcome Measure (CORE-OM) [70]. This 34-item scale measures 4 subscales: well-being, problems, functioning, and risk within a 7-day timeframe. Items are rated on a 6-point Likert-type sale (1=not at all to 6=most or all the time), with higher ratings indicating worse outcomes and greater psychological distress.

Attention-Deficit/Hyperactivity Disorder Rating Scale, Investigator-Administered

The Conners Adult ADHD Rating Scales was developed by Conners et al [71], to assess symptoms of ADHD in adults. The Attention-Deficit/Hyperactivity Disorder Rating Scale, Investigator-Administered (ADHDRS-IV-Inv) is an 18-item measure extracted from the Conners Adult Attention-Deficit/Hyperactivity Disorder Rating Scales Observer: Short Version, assessing the severity of ADHD inattentive and hyperactive-impulsive symptoms corresponding to the 18 items in the *DSM-5-TR* and providing a combined rating for severity and frequency of symptoms [72,73]. Participants are assessed on a 4-point scale (0=not at all or never to 3=very much or frequently), with severity indicated by higher ratings. The scale demonstrates good reliability, consistency, relative validity, and concurrent validity (α=.74-.95).

Attention-Deficit/Hyperactivity Disorder Quality of Life Scale

The Attention-Deficit/Hyperactivity Disorder Quality of Life Scale (AAQoL) is a 29-item measure assessing ADHD-related QoL areas of impact in 4 dimensions: productivity, mental health, life outlook, and relationships [74]. Participants evaluate how frequently each issue is problematic using a 5-point Likert-type scale (1=not at all or never to 5=extremely or very often), with higher ratings indicating greater problem frequency. Higher score therefore indicate poorer QoL.

Self-Reflection and Insight Scale

SDT highlights self-awareness as key in development of a strong AIC and the ability to identify basic psychological need frustration. The Self-Reflection and Insight Scale (SR&I) measure was used to assess individual differences in self-awareness [75]. This 20-item self-report scale consists of 3 subscales: an 8-item experience of self-reflection, an 8-item need for self-reflection, and an 8-item insight subscale. Participants rated their current state on a 7-point scale (1=strongly disagree to 7=strongly agree), with higher scores indicating greater and more frequent use of reflection skills.

Index of Autonomous Functioning

In SDT, autonomous behavior is experienced as self-congruent and integrated; however, the continual regulation of behavior can vary from highly autonomous, or truly self-regulated, to frequently experiencing external regulation from controlling influences [50]. Participants completed a 15-item, 5-point Likert-type scale (1=not at all true to 5=completely true) designed to measure individual differences in autonomy across 3 dimensions: authorship, control, and interest. Higher scores in any of these dimensions by inversion signifies greater autonomy.

Personal Questionnaire

An RCT for an intervention must demonstrate statistical significance of effectiveness for the treatment to be recommended. Practice-based evidence approaches in psychotherapy have faced challenges in this area due to the generic nature of measures, which can lack specificity and sensitivity in detecting subtle changes in an individual's functioning; therefore, the use of a personalized outcome measure is recommended [76]. In this study, within-treatment measures consisted of the Personal Questionnaire (PQ) [77], an individualized outcome measure comparing the efficacy of the intervention across the two participant groups. The PQ is generated by participants identifying up to a maximum of 10 issues they would like to address during the intervention. These issues are identified in the assessment session, and a 7-point within-session rating of distress (1=not at all to 7=maximum possible) is completed at the start of each session throughout the treatment.

Qualitative Measure

Thematic Analysis

Thematic analysis is a widely used method of qualitative research that emphasizes identifying, analyzing, and interpreting patterns of meaning in data. Participants were offered the opportunity to answer a single question about their experience during the final session of the 11-session intervention.

Statistical Analysis

With only 20 participants, tests suitable for analysis of small samples were used throughout, and all tests were 2-tailed with an alpha level of .05 (α=.05). Demographic and clinical characteristics were also tested for significant differences between intervention and waitlist control groups. In some cells, expected frequencies were below 5 participants; therefore, Fisher exact tests were used to test independence of experimental groups and nominal or ordinal demographics with two levels or categories. Cramer's V was used to test the independence of experimental groups and nominal or ordinal variables with more than two groups. A Mann-Whitney U test determined the independence of experimental groups on scale variables (Table 1).

XSL•FO
RenderX

Table 1. ADAPT (Autonomy, Design, Awareness, Psychoeducation, and Training Integration for Sustainable Change) framework pilot study: frequency of demographic attributes per participant group and overall (N=23).

Group characteristic	Intervention, n (%)	Control, n (%)	Total, n (%)
Size	12 (52.2)	11 (47.8)	23 (100)
Gender			
Male	7 (30.4)	6 (26.1)	13 (56.5)
Female	4 (17.4)	4 (17.4)	8 (34.8)
Transgender	1 (4.4)	1 (4.4)	2 (8.7)
Total	12 (52.2)	11 (47.8)	23 (100)
Age (years)			
<33	9 (39.1)	7 (30.4)	16 (69.6)
>33	3 (13)	4 (17.4)	7 (30.4)
Total	12 (52.2)	11 (47.8)	23 (100)
Race			
White British	12 (52.2)	8 (34.8)	20 (87)
Not White British	0 (0)	3 (13)	3 (13)
Total	12 (52.2)	11 (47.8)	23 (100)
ADHD [a] subtype			
Inattentive	3 (13)	4 (17.4)	7 (30.4)
Hyperactive	1 (4.4)	1 (4.4)	2 (8.7)
Combined	8 (34.8)	6 (26.1)	14 (60.1)
Total	12 (52.2)	11 (47.8)	23 (100)
Date of diagnosis			
Before 2020	4 (17.4)	4 (17.4)	8 (34.8)
During or after 2020	8 (34.8)	7 (30.4)	15 (65.2)
Total	12 (52.2)	11 (47.8)	23 (100)
Medication			
None	2 (8.7)	3 (13)	5 (21.7)
Prescribed	10 (43.5)	8 (34.8)	18 (78.3)
Total	12 (52.2)	11 (47.8)	23 (100)

[a]ADHD: attention-deficit/hyperactivity disorder.

To determine whether the two groups differed between assessment moments over time, a nonparametric Wilcoxon signed-rank test comparing pretreatment and posttreatment data was used for all measures, except for the within-session measure, the PQ. A Friedman test was used to analyze the data collected across 10 assessment moments (T1 to T10) of the PQ, with Bonferroni-corrected multiple comparisons.

Comparing pretreatment and posttreatment results of all measures was considered in terms of reporting possible efficacy of the intervention. However, due the use of multiple measures in this study, a single general score, identified as "total," was calculated by summing all ratings to provide an indicator participants' performance across each of the 6 repeated measures. After establishing pretreatment values for each total variable, the normality of the distribution of these scores was tested using the Shapiro-Wilk normality test. The independence of total scores from group assignment procedures was tested with a Mann-Whitney *U* test to clarify whether groups were potentially biased in terms of pretreatment values. Finally, correlations between these variables were tested with Spearman correlations, along with Fieller, Hartley, and Pearson CIs for correlations. All analysis was conducted with SPSS (version 29; IBM Corp).

Ethical Considerations

The ethical process for this project was reviewed by the University of Huddersfield School of Human and Health Sciences—School Research Ethics and Integrity Committee—and received Health Research Authority and Care Research Wales approval (Multimedia Appendix 3). All sessions with participants were conducted in accordance with the Ethical Framework for Good Practice set by the United Kingdom Council for Psychotherapy. The Research Ethics Committee

XSL•FO
RenderX

reference for the study is 21/SC/0143. The Integrated Research Application System project ID is 291103. The trial protocol can be accessed at ClinicalTrials.gov (project ID NCT04832737).

The ethical implications for this study were considered to protect the identity of individual participants, all personally identifiable data were anonymized and will not be released. Information on confidentiality policy and anonymization of personally identifiable data were included in the consent form. Participants in the pilot study were offered a counselling agreement confirming that all details and discussions within the therapeutic relationship are confidential, unless they or anyone else are at risk of serious harm, in accordance with the United Kingdom Council for Psychotherapy Ethical Guidelines (2019; Multimedia Appendix 3).

The researcher complied with the General Data Protection Regulation, the NHS Confidentiality Code of Practice, the Computer Misuse Act (covering information security), and all local trust policies regarding the collection, storage, processing, and disclosure of personal information. All participant case records were kept in electronic form (consent forms, agreements, and interviews), and participants' home addresses (including postcodes) and telephone numbers were stored on a secure database and spreadsheet on NHS and University computers in accordance with the Data Protection Act [78].

Participants received an initial interview invitation to discuss study participation. An information sheet detailing the purpose, activities, outcomes, and results of the study was provided, including the researcher's contact details (name, phone number, and email address). Participants were encouraged to read the information sheet and ask questions on assessment day before signing the consent form (Multimedia Appendix 3). The voluntary nature of participation and the ability to withdraw consent at any time were emphasized during the study. All participants were offered the intervention as compensation for their time and participation.

The researcher recognized that some participants might have been recently diagnosed with ADHD. Incorporating this information into their identity and self-concept initiates a process of acceptance, which may include elements of anger and grief. The researcher was a qualified psychotherapist with 10 years of experience, and a distress policy was created as part of the research protocol. Participants were under the current clinical care of the NHS Adult ADHD Service at the South West Yorkshire Partnership NHS Foundation Trust, which was made aware of any additional support requirements if needed.

In accordance with the Data Protection Act (1998), personal data will not be retained for longer than necessary. All participant personal data, transcripts, recordings, memos, and process notes were retained for the duration of the project and were accessible only by the researcher. To submit for publication, participant data were retained to obtain permission for the study results to be published, in accordance with ethical approval. Anonymized electronic data were retained in a secure, password-protected spreadsheet and database. Hard copy data, including process notes, were stored in a locked cabinet. Videos of the intervention sessions were recorded for random review by university supervisors to ensure intervention fidelity and were destroyed upon completion of the review.

Results

Recruitment and Retention

Eligibility criteria consisted of age 18 years or older with a confirmed diagnosis of ADHD and access to a computer or smartphone with an internet connection. Participants with comorbid diagnosis (eg, autism, bipolar disorder, intellectual disabilities, learning difficulties, traumatic brain injury, psychosis, or Tourette syndrome), a diagnosis of substance use disorder or personality disorder, or other mental health disorders (eg, posttraumatic stress disorder and oppositional defiant disorder) were not eligible for the study. Medication was not included in the exclusion criteria, as participants were under current NHS treatment. Additionally, research indicates that multimodal treatment is recommended for adults with ADHD [10,13]. Figure 2 outlines the participant recruitment process.

Participants were recruited by staff from the Adult ADHD Clinic at the South West Yorkshire Partnership NHS Foundation Trust, which proved challenging due to high rates of ADHD comorbidity (58.4%) [79]. Recruited and randomized participants in this study consisted of 23 adult NHS patients between May 2022 and September 2022, representing 76% of the original target of 30 participants. Reasons for nonrecruitment included those who declined (n=7) and those who were discovered to be ineligible due to comorbid diagnosis of dyslexia and dyspraxia, dyslexia only, and generalized anxiety disorder (n=3). The researcher contacted and allocated a rolling entry of participants using 4-block randomization to the intervention or waitlist control over a 12-week period. The randomization was not blinded, as the researcher also delivered the intervention.

Attendance analysis showed an average of 15 weeks for participants to complete the 11-session program. The attrition rate for the full study was 13.04% of 23 participants (Figure 2). One intervention participant left at session 5, and 2 waitlist control group participants dropped out before posttreatment measure data collection without continuing on to the intervention. Correlations between demographic and clinical indicators showed that fewer dropout participants were medicated ($\rho=-0.422$; $P=.05$). Therefore, dropout rates suggest there may be a small risk of bias; however, the CI for this estimate was large, varying between −0.711 and −0.01 (Tables S10 and S11 in Multimedia Appendix 4).

XSL•FO
RenderX

Figure 2. ADAPT (Autonomy, Design, Awareness, Psychoeducation, and Training Integration for Sustainable Change) framework pilot study recruitment and participation flowchart—Consolidated Standards of Reporting Trials (CONSORT) format.

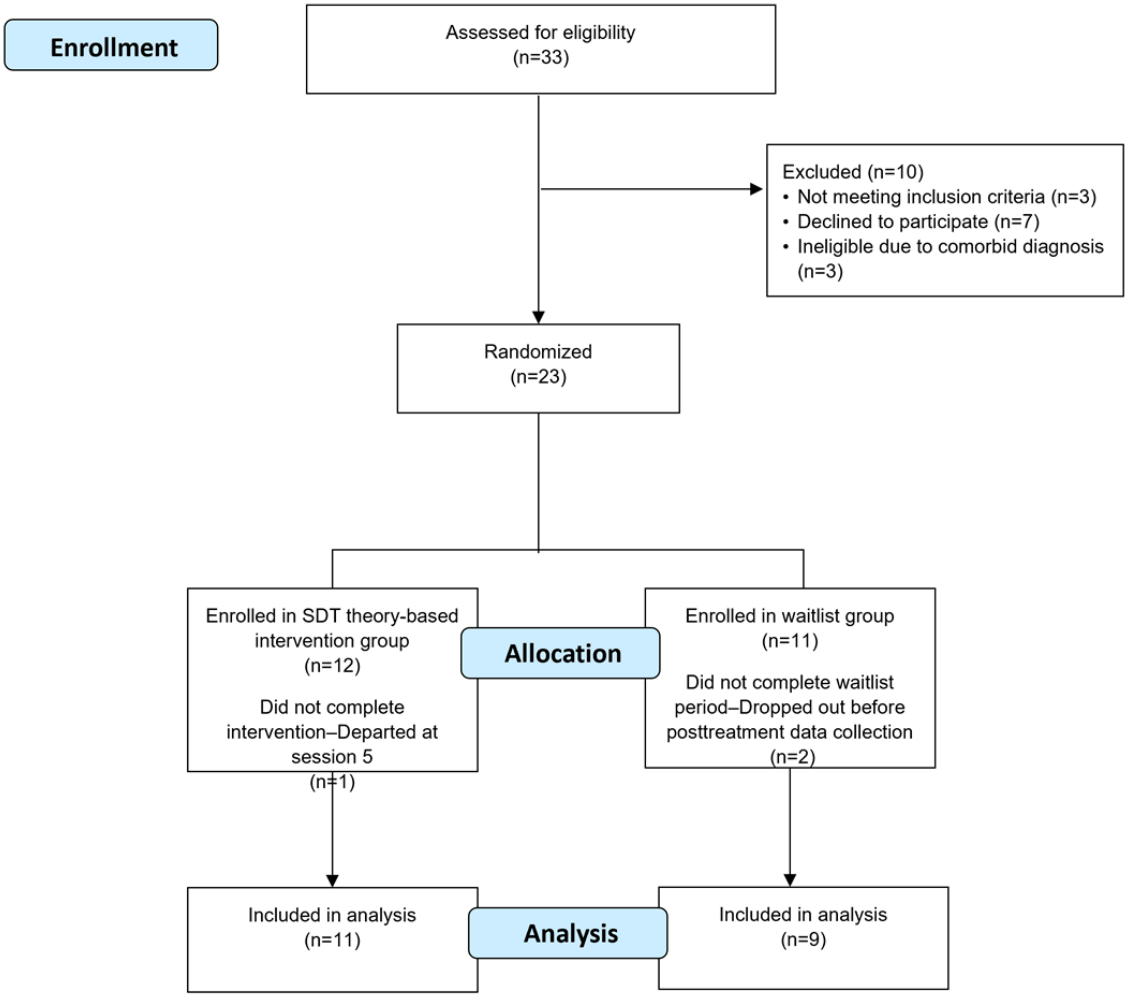

Measures Acceptability

This study aimed to evaluate the acceptability and accessibility of measures by participants with adult ADHD. A total of 2 participants did not complete the pretreatment or posttreatment ADHD QoL measure, leaving 10 participants in the intervention group for this measure. In the measure of autonomy Index of Autonomous Functioning (IAF), 2 participants had missing answers at pretreatment, which were resolved through imputation strategies based on each participant's response pattern.

In the PQ, participants selected up to a maximum of 10 issues to evaluate through the intervention. Missing ratings for any issue were also absent for the entire timeframe of the intervention, indicating that the issue had not been identified for evaluation. Therefore, in both groups, the sample size decreased as the number of issues identified increased (Table 2).

One control participant only completed 9 of 10 PQ measures; therefore, an imputation strategy was used based on the participant's response pattern (Multimedia Appendix 5).

Table 2. ADAPT (Autonomy, Design, Awareness, Psychoeducation, and Training Integration for Sustainable Change) framework pilot study: personal questionnaire measure of distribution of missing values and valid sample size.

Variable	Intervention		Control	
	Missing values, n	Sample size, n	Missing values, n	Sample size, n
P1-P6	0	11	≈0[a]	8
P7	0	11	1	8
P8	1	10	3	6
P9	2	9	6	3
P10	6	5	6	3

[a]One problem lacked ratings.

XSL•FO
RenderX

Demographics and Clinical Characteristics

Participant age ranged from 20 and 56 years, with a mean of 33.35 (SD 10.1) years. In total, 13 participants identified as male, 8 as female, and 2 as transgender. Most participants (n=20) identified as White British. Clinically, 14 participants received a combined diagnosis of ADHD, 15 were diagnosed after 2020, and 18 (78.26%) were actively taking prescribed medication at the time of the study. There was no bias in the distribution of assessed demographic and clinical attributes, as group assignment was independent of Gender (Cramer V=0.038; *P*=.98), age (Mann-Whitney *U*=85.5; *P*=.24), race-nationality (Fisher exact test *P*=.09), diagnosis (Cramer V=0.13; *P*=.83), date of diagnosis (Mann-Whitney *U*=64.5; *P*=.93), and medication (Fisher exact test *P*=.64; Table 1).

Correlations between posttreatment totals and demographics were also inspected to identify any participants who might be more susceptible to the intervention. The date of the diagnosis correlated with: EQ-5D-5L total (ρ=0.573; *P*=.08), SR&I total (ρ=–0.543; *P*=.01), and ADHDRS-IV-Inv (ρ=0.553; *P*=.01). This indicates that the more recently participants had been diagnosed, the stronger were their symptoms of ADHD, and the poorer were their reflection abilities and health state.

Quantitative Analysis for Measure Effectiveness

Table 3 summarizes the mean (SD) for each of the measures used at pretreatment and at posttreatment, namely EQ-5D-5L, CORE-OM, ADHDRS-IV-Inv, AAQoL, SR&I, and IAF. Results are also unpacked one by one for each scale in the following subsections.

XSL•FO
RenderX

Table 3. ADAPT (Autonomy, Design, Awareness, Psychoeducation, and Training Integration for Sustainable Change) framework pilot study: results of primary outcome measures.

Variable	Intervention group		Control group	
	Pretreatment, mean (SD)	Posttreatment, mean (SD)	Pretreatment, mean (SD)	Posttreatment, mean (SD)
EQ-5D-5L				
Index value	0.70 (0.23)	0.73 (0.16)	0.78 (0.16)	0.8 (0.12)
Total	8.91 (2.98)	8.36 (2.20)	7.89 (2.71)	7.67 (2.24)
CORE-OM[a]				
Well-being	1.82 (0.51)	1.34 (0.36)	1.47 (0.81)	1.31 (0.89)
Problems	1.73 (0.53)	1.11 (0.35)	1.47 (0.62)	1.25 (0.67)
Functioning	1.57 (0.75)	1.12 (0.42)	1.57 (0.68)	1.19 (0.60)
Risk	0.08 (0.16)	0.05 (0.08)	0.13 (0.23)	0.11 (0.17)
Nonrisk	1.67 (0.53)	1.15 (0.27)	1.52 (0.63)	1.23 (0.63)
Risk to self	0.05 (0.15)	0.02 (0.08)	0.17 (0.33)	0.14 (0.22)
Risk to others	0.14 (0.23)	0.09 (0.20)	0.06 (0.17)	0.06 (0.17)
Total	1.39 (0.44)	1.21 (0.23)	1.27 (0.55)	1.32 (0.33)
ADHDRS-IV-Inv[b]				
Inattention	17.45 (5.17)	13.55 (4.12)	19.0 (4.98)	17.22 (4.82)
Hyperactivity	16.0 (5.6)	12.36 (5.4)	17.67 (5.1)	15.22 (5.63)
Total	33.45 (9.94)	25.91 (9.32)	36.67 (8.4)	32.44 (8.58)
AAQoL[c]				
Productivity	52.96 (15.27)	46.59 (13.77)	58.59 (13.72)	50.50 (8.71)
Mental health	40.42 (19.75)	47.08 (17.9)	39.82 (19.22)	49.54 (16.6)
Outlook	61.43 (16.74)	61.43 (13.66)	50.0 (11.85)	55.95 (12.11)
Relationships	44.50 (20.0)	54.0 (19.65)	41.67 (11.99)	45.0 (22.08)
Total	50.95 (9.12)	51.55 (6.52)	40.71 (4.76)	50.67 (7.43)
SR&I[d]				
Self-reflection	4.17 (1.49)	4.24 (1.32)	4.50 (0.89)	4.37 (1.24)
Need for self-reflection	4.91 (0.94)	4.97 (0.80)	5.57 (0.32)	5.09 (0.61)
Insight	3.35 (1.01)	3.71 (0.82)	3.01 (1.12)	3.61 (1.21)
Total	4.06 (1.04)	4.25 (0.82)	4.23 (0.55)	4.28 (0.78)
IAF[e]				
Authorship	16.04 (2.78)	17.55 (3.11)	16.89 (4.31)	16.33 (3.87)
Control	14.41 (4.82)	16.00 (5.48)	14.78 (3.80)	14.89 (3.85)
Interest	18.73 (3.74)	20.45 (4.08)	19.22 (2.91)	18.22 (4.71)
Total	49.18 (8.12)	54.00 (7.39)	50.89 (8.01)	49.44 (9.46)

[a]CORE-OM: Clinical Outcomes in Routine Evaluation–Outcome Measure.

[b]ADHDRS-IV-Inv: Attention-Deficit/Hyperactivity Disorder Rating Scale, Investigator-Administered.

[c]AAQoL: Attention-Deficit/Hyperactivity Disorder Quality of Life Scale.

[d]SR&I: Self-Reflection and Insight Scale.

[e]IAF: Index of Autonomous Functioning.

Health-Related Quality of Life (EQ-5D-5L)

An EQ-5D-5L profile for each participant rating was generated from the data, and frequencies of every profile were determined by assessment moment and group. Only one waitlist control group profile showed a frequency higher than 1, profile 11112, which was detected in 3 participants at both pretreatment and posttreatment with no reported health state change. Findings

XSL•FO
RenderX

suggest participants self-assessed their health as generally good, with slight problems in one or another single area. Wilcoxon signed-rank test results for the intervention group (W=14.0; 2-tailed significance *P*=.30) for the subscales of mobility (W=1.5; *P*>.99), self-care (W=1.5; *P*>.99), activities (W=7.0; *P*=.11), pain/discomfort (W=5.0; *P*>.99), and anxiety/depression (W=7.7; *P*>.99); and for the control group (W=4.5; 2-tailed significance *P*=.85) for the subscales of mobility (W=0.0; *P*>.99), self-care (W=1.0; *P*=.32), activities (W=3.5; *P*=.79), pain/discomfort (W=2.5; *P*=.16), and anxiety/depression (W=5.0; *P*>.99) showed no significant differences in either group between pretreatment and posttreatment results.

Psychological Distress (CORE-OM)

Well-being, problems, functioning, and risk as assessed by the CORE-OM showed that most participants experienced low levels of distress on every measure, and mean scores decreased overall in both groups at posttreatment. A single exception was the total score measure for the control group, which showed a small increase at posttreatment, suggesting interference in longitudinal outcomes by confounding factors. Subscales showing the highest averages, representing the greatest difficulties, were well-being and problems, whereas the lowest averages were in items assessing risk. In the intervention group, significant differences were identified in 4 subscales: well-being (*z*=6.0; *P*=.03), problems (*z*=0.0; *P*=.01), functions (*z*=5.0; *P*=.02), and nonrisk (*z*=2.0; *P*=.01); but not in risk (*z*=4.0; *P*=.71), risk to self (*z*=1.0; *P*=.66), or risk to others (*z*=6.0; *P*=.66). In the control group, only nonrisk showed significant differences across assessment moments (*z*=3.0; *P*=.04), with none observed in well-being (*z*=12.0; *P*=.39), problems (*z*=5.0; *P*=.13), functions (*z*=7.0; *P*=.07), risk (*z*=1.0; *P*=.66), risk to self (*z*=1.0; *P*=.66), or risk to others (*z*=1.5; *P*>.99).

ADHD Symptom Severity (ADHDRS-IV-Inv)

As assessed by the ADHDRS-IV-Inv scale, the control group reported slightly greater symptom severity at pretreatment than the intervention group. However, a decrease was registered from pretreatment to posttreatment in both groups across all subscales of this measure. Wilcoxon signed-rank tests confirmed that while both groups showed improvements at posttreatment, significant differences in favor of the intervention group were seen, with most improvements registered in the inattention scale (*z*=3.0; *P*≤.01) versus hyperactivity (*z*=2.5; *P*≤.05) and total scale (*z*=3.0; *P*≤.01). In comparison, control group improvements were noted for inattention (*z*=3.0; *P*≤.05), hyperactivity (*z*=0.0; *P*≤.05), and total (*z*=0.0; *P*≤.05) scores.

ADHD-related QoL (AAQoL)

The AAQoL is a self-report measure of the impact of ADHD symptoms on QoL. Analysis of intervention participants was limited to 10, as one participant did not complete the measure (ID=5). The mean (SD) observed for both groups show mid-scale responses at both assessment moments, demonstrating neither very positive nor very negative QoL. Subscale analysis indicated the intervention group increased in productivity and decreased in mental health and relationships, while outlook showed no change. The control group also increased in productivity, but averages were lower, and QoL decreased in mental health, relationships, and outlook. Wilcoxon signed-rank test showed significant differences in both groups on every scale between assessment moments, except for total score. Positive changes, indicative of increased distress, were reported in the intervention group for mental health (*z*=28.0; *P*=.02), outlook (*z*=21; *P*=.67), relationships (*z*=32.5; *P*=.04), and in total scores (*z*=31.0; *P*=.72), but not in productivity (*z*=7.0; *P*=.04). Negative changes, indicated of reduced distress, were reported in the control group for productivity (*z*=3.0; *P*=.02), whereas more distress was observed in outlook (*z*=26.0; *P*=.04), mental health (*z*=34.0; *P*=.02), relationships (Z=23.0; *P*=.48), and total scores (*z*=21.5; *P*=.62).

Self-Awareness (SR&I)

The SR&I is an SDT-based self-report self-awareness measure, with increases in scoring indicating positive change. Rankings were higher at intervention posttreatment across all measures, indicating overall improvement; however, higher rankings were observed in need reflection at both assessment moments, with slightly higher scores at posttreatment, while lowest rankings highlighted weakness in the insight scale. Both self-reflection and need reflection subscales showed higher results at control group baseline, indicating a decrease in self-reflection skills posttreatment. Wilcoxon signed-rank test showed no significant differences for the intervention group in subscales for self-reflection (*z*=20; *P*=.78), need for self-reflection (*z*=13.5; *P*=.93), insight (*z*=47.5; *P*=.2), or total (*z*=42.5; *P*=.4). Results for the control group showed no significant differences in subscales for self-reflection (*z*=10.0; *P*=.50), insight (*z*=35.5; *P*=.12), or total (*z*=26.0; *P*=.67), but significant differences were observed in the need for self-reflection subscale (*z*=1.0; *P*=.05), indicating worse results.

Sense of Autonomy (IAF)

The IAF is an SDT-based self-report measure of experiences of a sense of autonomy, with increases in scoring indicating positive change. The mean (SD) indicates that most participants reported moderately satisfactory levels of autonomy, with an average distribution of 12.5 for each subscale and 37.5 for total score. Intervention group posttreatment mean (SD) was higher for every measure, with the highest being total score at both assessment moments. Conversely, the highest averages of each pair of subscales were not always observed posttreatment in the control group, with only the control subscale demonstrating high averages. The highest averages were observed for interest, followed by authorship, and then control in both groups. Wilcoxon signed-rank text showed there were no significant differences in authorship (*z*=50.0; *P*=.13), control (*z*=39.0; *P*=.24), or interest (*z*=35.5; *P*=.41), with the closest to significance being total (*z*=52.5; *P*=.08) for the intervention group; and in authorship (*z*=20.5; *P*=.81), control (*z*=23.5; *P*=.91), interest (*z*=10.0; *P*=.25), or total (*z*=13.0; *P*=.48) for the control group.

Within Session Individual Outcome Measure (PQ)

The individual outcome measure analysis was performed on both groups independently, with the control group serving as a replication group undergoing the intervention. A total of 2 approaches were used for PQ data analysis. In an item-based

XSL•FO
RenderX

approach, each problem was compared session by session within each group, and the widest and narrowest ranges of the scale selected by participants were compared (Table 4). This measure was useful, as participants used the full range of the scale. In terms of levels of distress, examination of the range of medians indicated higher levels of distress in the intervention group than in the control group.

Table 4. ADAPT (Autonomy, Design, Awareness, Psychoeducation, and Training Integration for Sustainable Change) framework pilot study: personal questionnaire measure of range and median comparison of individual outcomes.

Measure	Intervention	Control
Largest range	7 (P1T1; P8T5; P9T2)	6 (P8T7)
Narrowest range	2 (P1T8)	1 (P1T4; P4T6-P6T10; P5T2, P5T4, and P5T6-P5T10)
Highest median	6 (P2T1; P5T3 and P5T6; P7T2; P9T1)	6 (P6T1)
Lowest median	3 (P10T9)	2 (P9T8; P10T8)

In a time-related, median-based approach, medians for the 10 moments (T1 to T10) were determined for every participant and used to compare groups. T6 functioned as the latest assessment moment at which all participants identified an issue; however, comparisons up to T10 are included. Greater levels of distress were indicated in the intervention group; however, a larger range of the intensity of distress was observed in the control group in mean (SD; Table 5). These findings remain consistent when analysis of missing data is included, despite only 5 of 11 (45%) participants contributing ratings from the intervention group and 3 of 9 (33%) from the control group. A small observable decrease in distress was seen in the control group (Figure 3).

Table 5. ADAPT (Autonomy, Design, Awareness, Psychoeducation, and Training Integration for Sustainable Change) framework pilot study: personal questionnaire measure of mean (SD) of medians (T1 to T10) of individual outcomes.

Moments	Intervention		Control	
	Frequency, n	Value, mean (SD)	Frequency, n	Value, mean (SD)
Median T1	11	4.82 (1.65)	9	5.39 (0.78)
Median T2	11	4.55 (1.49)	9	4.67 (0.56)
Median T3	11	4.59 (1.51)	9	4.00 (0.56)
Median T4	11	4.41 (1.02)	9	4.06 (0.64)
Median T5	11	4.41 (1.72)	9	3.89 (0.55)
Median T6	11	4.36 (1.19)	9	3.89 (1.45)
Median T7	11	4.18 (1.10)	9	3.33 (1.23)
Median T8	11	4.36 (1.42)	9	3.67 (1.44)
Median T9	11	4.50 (1.16)	9	3.39 (1.22)
Median T10	11	4.55 (1.21)	8	3.19 (1.13)

XSL•FO
RenderX

Figure 3. ADAPT (Autonomy, Design, Awareness, Psychoeducation, and Training Integration for Sustainable Change) framework pilot study: personal questionnaire summary of median variables across within-session assessment moments (T1-T10).

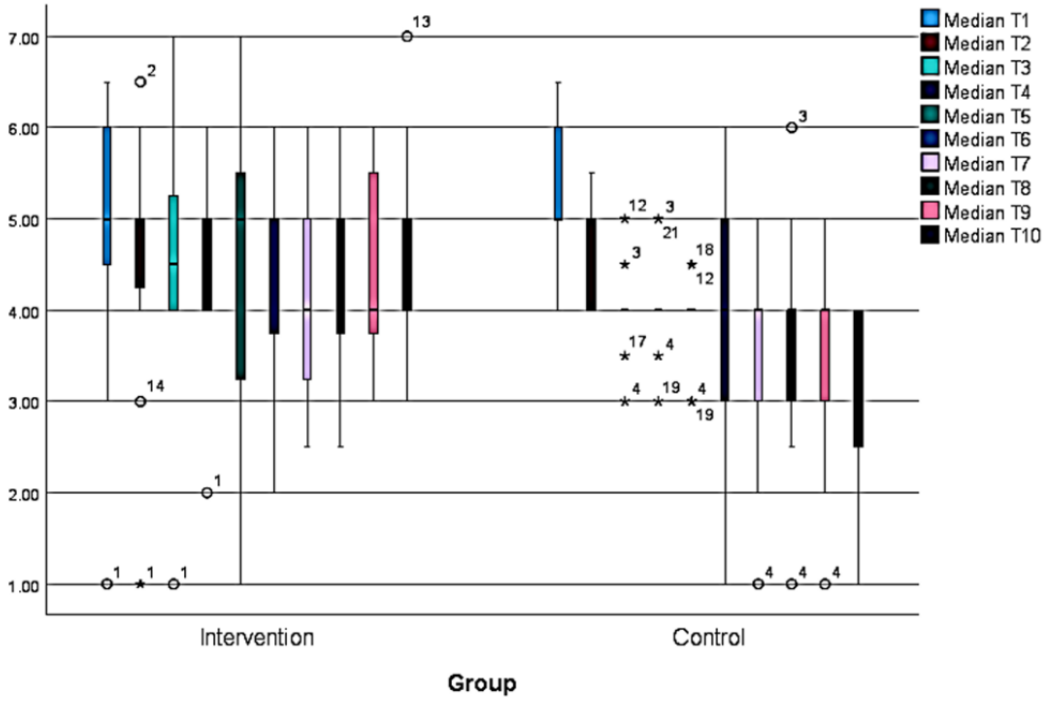

A Friedman test compared item-based ratings problem by problem (Table 6). Statistically significant differences were detected in the control group for every problem except T9 and T10, where the sample size was small (n=3). Multiple comparisons were performed to identify which pairs contributed to notable differences, with significance values adjusted using the Bonferroni correction for multiple tests. The only problems that remained significant were: P1, between T1 and T10 (*P*=.04); P2, between T2 and T9 (*P*=.02); P4, between T2 and T10 (*P*=.04); P5, between T1 and T4 (*P*=.02), T1 and T9 (*P*=.003), and T1 and T10 (*P*=.02); and P7, between T1 and T10 (*P*=.002) and T2 and T10 (*P*=.008).

Table 6. ADAPT (Autonomy, Design, Awareness, Psychoeducation, and Training Integration for Sustainable Change) framework pilot study: personal questionnaire results—Friedman χ^2 test for T1 to T10 across individual outcomes.

Problem	Intervention		Control	
	Chi-square (*df*)	*P* value	Chi-square (*df*)	*P* Value
P1 T1-T10	14.24 (9)	.11	29.33 (9)	≤.01
P2 T1-T10	5.96 (9)	.74	29.92 (9)	≤.001
P3 T1-T10	7.24 (9)	.61	24.25 (9)	.04
P4 T1-T10	10.66 (9)	.30	29.97 (9)	≤.001
P5 T1-T10	13.51 (9)	.14	41.27 (9)	≤.001
P6 T1-T10	7.07 (9)	.63	35.93 (9)	≤.001
P7 T1-T10	3.26 (9)	.95	32.4 (9)	≤.001
P8 T1-T10	10.19 (9	0.36	20.84 (9)	.01
P9 T1-T10	9.99 (9)	0.35	13.69 (9)	.13
P10 T1-10	7.23 (9)	.61	14.31 (9)	.11

Differences associated with assessment moments were examined using the set of time-related, median-based variables, median T1 to median T10. Significant differences were detected in the control group, but not in the intervention group, and comparisons further apart yielded stronger results. In multiple comparison analysis (Figures 4-6), significant differences detected in the control group are identified by blue lines. Overall, these results show that, when comparing T1 to T6, T1 differed significantly from T3 (χ^2_5=2.93; *P*=.13), T5 (χ^2_5=2.94; *P*=.13), and T6 (χ^2_5=2.72; *P*=.03). When comparing T1 to T8, these significant differences were T1 to T7 (χ^2_7=5.0; *P*≤.001) and to T8 (χ^2_7=4.17; *P*=.009). When comparing T1 to T10, significant differences were found between T1 and T7 (χ^2_9=5.75; *P*=.007), T8 (χ^2_9=5.125; *P*=.03), T9 (χ^2_9=5.5; *P*=.01), and T10 (χ^2_9=6.0; *P*=.03).

XSL•FO
RenderX

Figure 4. ADAPT (Autonomy, Design, Awareness, Psychoeducation, and Training Integration for Sustainable Change) framework pilot study: personal questionnaire measure of multiple comparisons for T1-T6 within-session assessment moments for the Control group.

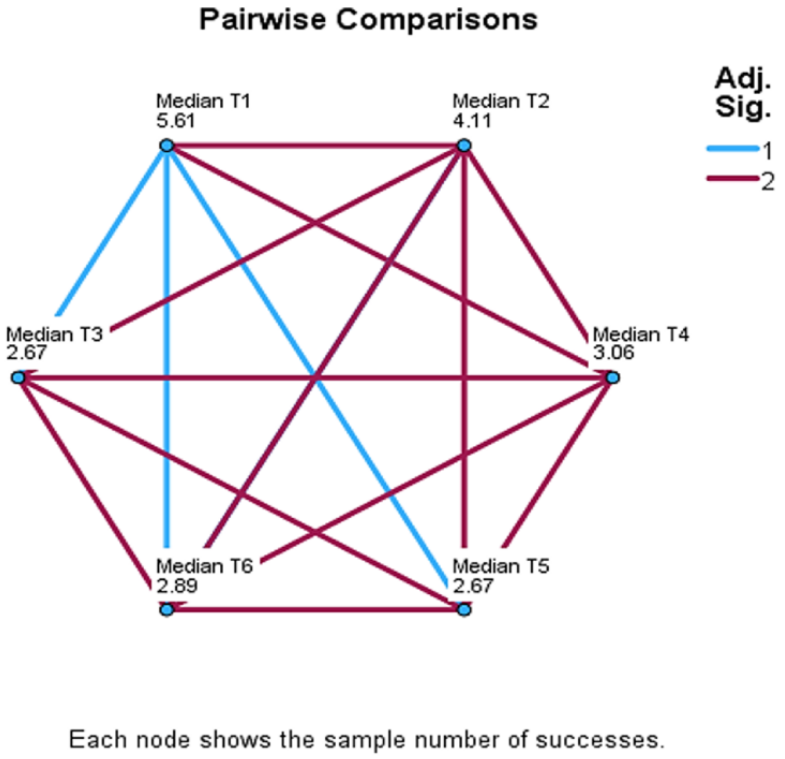

Figure 5. ADAPT (Autonomy, Design, Awareness, Psychoeducation, and Training Integration for Sustainable Change) framework pilot study: personal questionnaire measure of multiple comparisons for T1-T8 within-session assessment moments for the Control group.

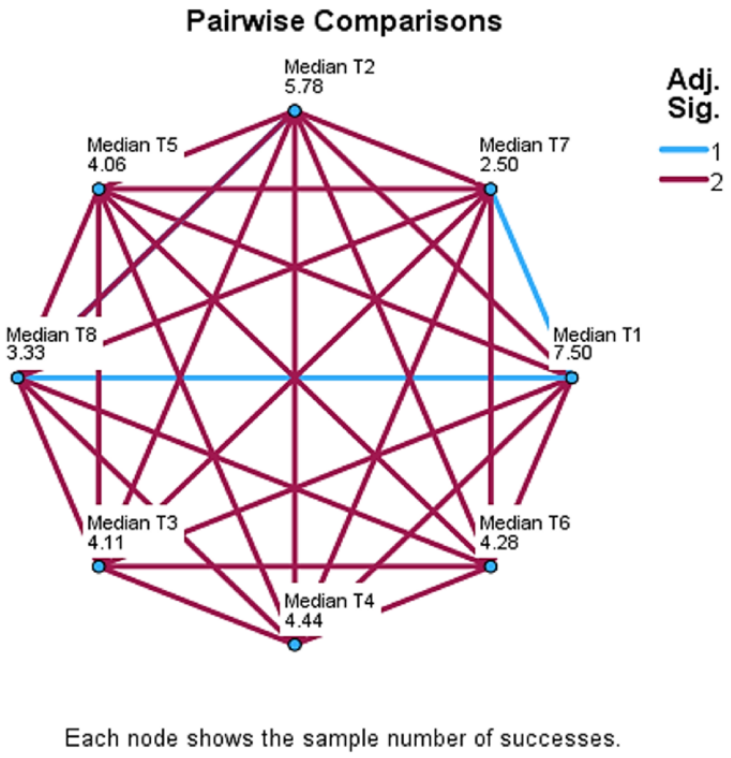

XSL•FO
RenderX

Figure 6. ADAPT (Autonomy, Design, Awareness, Psychoeducation, and Training Integration for Sustainable Change) framework pilot study: personal questionnaire measure of multiple comparisons for T1–T10 within-session assessment moments for the Control group.

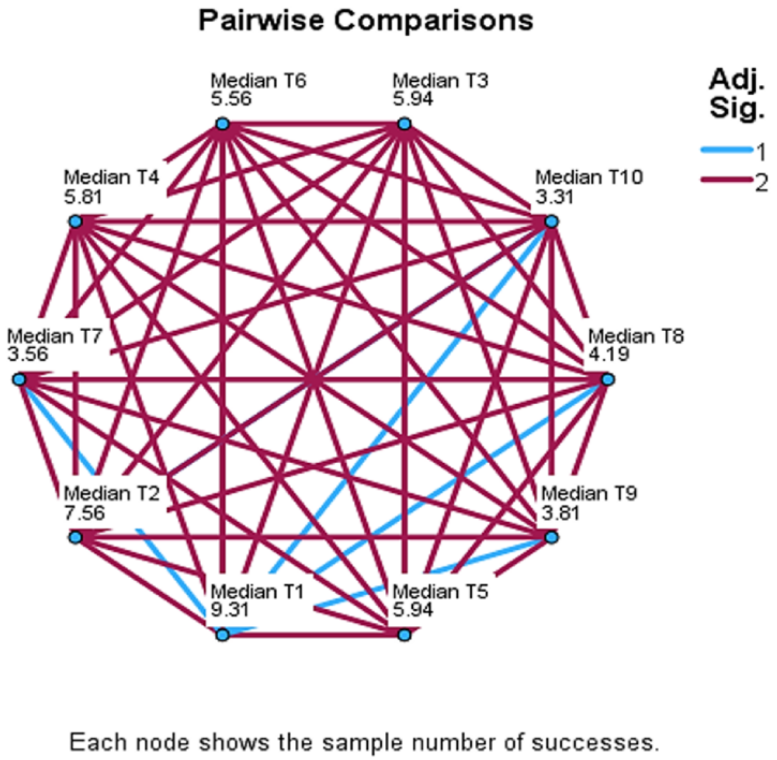

Qualitative Analysis for Acceptability

During the final session, all participants were asked to provide intervention acceptability data by answering the question: "How did you find the experience? Was it useful?" All participants who completed the intervention agreed (Multimedia Appendix 6 for excerpts). Data analysis was conducted using thematic analysis principles outlined by Braun and Clarke [52].

Data Analysis

Audio recordings of all 20 final sessions were transcribed verbatim and uploaded into NVivo (QSR International). Line-by-line coding focused on sections of the transcript referring to participants' experiences, usefulness, and satisfaction with the intervention. A theme was identified for each aim explored in the analysis. Review and reflection were conducted midway through the data (participant 12) and again at the three-quarter stage (participant 18) to identify initial themes. The initial phase produced 55 codes from 20 interviews, which were grouped into 3 themes and 8 subthemes (Figure 7).

XSL•FO
RenderX

Figure 7. ADAPT (Autonomy, Design, Awareness, Psychoeducation, and Training Integration for Sustainable Change) framework pilot study: qualitative analysis of themes and subthemes. ADHD: attention-deficit/hyperactivity disorder.

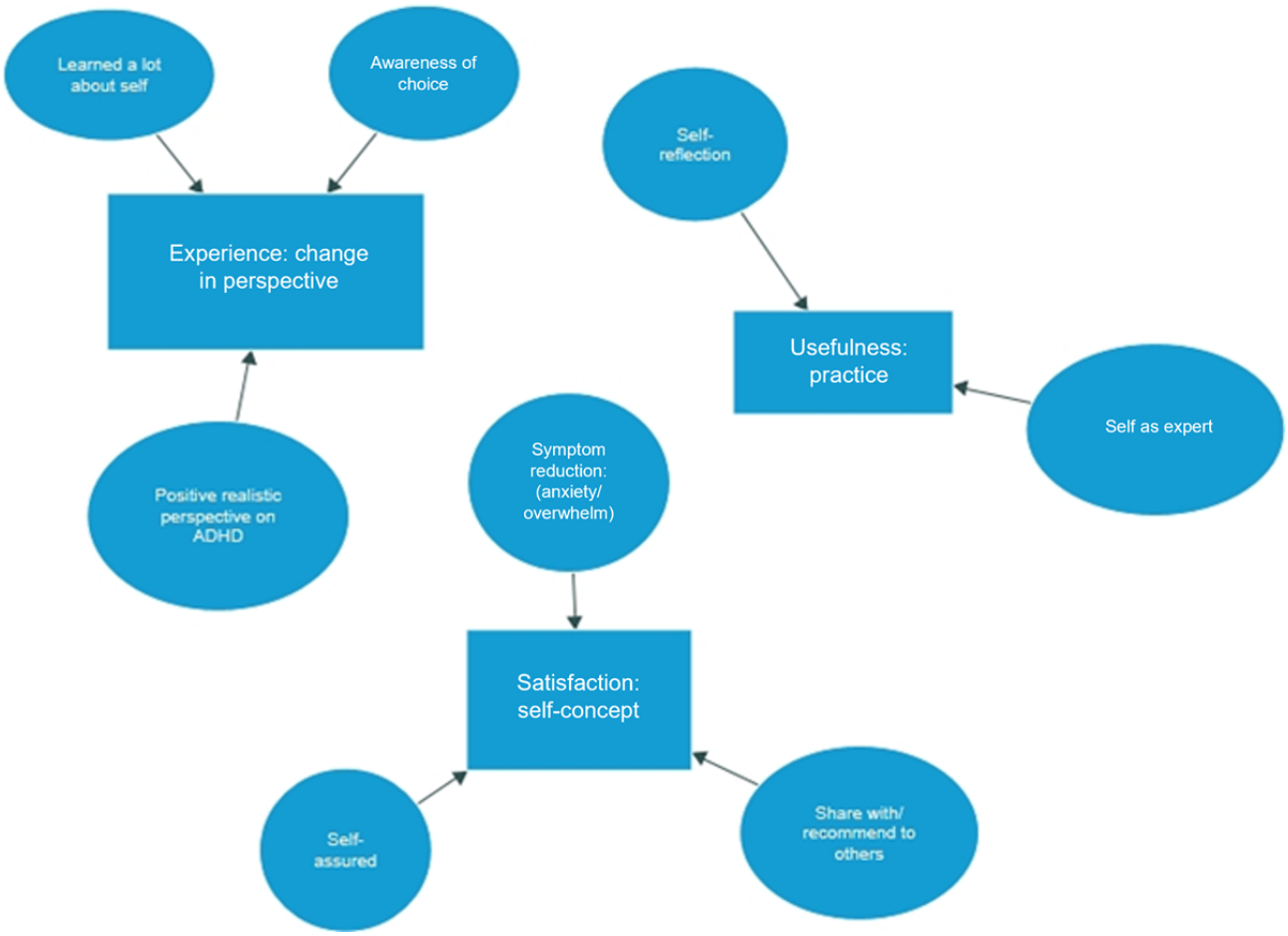

Findings

Experience

The emergent theme for intervention experience was a Change in Perspective. The subthemes were learning about self, positive perspective on ADHD, and awareness of choice. Participants found working the framework easy to integrate and provided new insights into their behaviors, new perspectives on options, and a sense of self-awareness around actions that altered their understanding.

> *Actually it's something - it's a problem I had … around the I think you call it like the disorder model of this. It was that like having an idea of the neurobiology. My first reaction to that, as a result of learning those was to be like, 'OK well, my brain works differently and that's my excuse. I guess that's just … how things are.' And then, 'OK, well, I guess just things are worse for me. And I'll just have to just make peace with the fact that this is where things are.' Whereas in the sessions we've had here, it's been more about - OK, cool. Well, let's look at that, let's actually look at overcoming some things or using the gifts that you have. I mean - It's also not the same as, like, 'People with ADHD are able to use hyper focus' - it's not like seeing ADHD as superpowers either. It feels like a more accurate model to me in a way. And not trying to like you know, just blow smoke at you here, but it just felt a bit more real.* [Participant 7]

This new perspective also supported participants' exploration and experimentation with approaches to meet their needs, including changes in internal processes and in seeking external support through self-advocacy.

> *This… whole process has just been amazing because I feel like I've just understand so much and I can see what areas, maybe not need improvement, but where I need help or where I need to learn more or just do some more work and things that can get put in place, like in the workplace, which is really good…And just learning more about why I am with the way to do things, the way I need to plan things out beforehand, is just amazing. And I've learned so much about myself and it just makes things so much easier, which is really nice and like when let's say RSD [Rejection Sensitivity Dysphoria] kicks in, I know what that is and you can rationalise it so much easier than despair, which is really nice.* [Participant 23]

XSL•FO
RenderX

Usefulness

The emergent theme for intervention usefulness was Practice. The subthemes were self-expertise and self-reflection. The program was described as “useful” and “helpful” by 18 of 20 (90%) participants. Participants commented on tool development or skills learning with a focus on need satisfaction through tool adaptation, facilitating ongoing use and supporting improved functioning. Many participants highlighted self-reflection as a key factor in developing a supportive practice. A total of 2 participants found the process difficult due to extenuating circumstances, highlighting the impact of environmental support on task engagement. However, both participants identified impact factors and felt the intervention was useful despite external challenges.

> *No, it is going well. It's going really well. I can definitely see a change, just like I said the last time, just to thinking about it, the thought of it, you know, me sitting there and thinking about those tools that for me is really because that then gives me an option. You know, it gives me a good option and I feel like I'm not, My love, I'm being honest. I'm not overstretching myself anymore. Oh, I haven't been. For the last, oh, several weeks, I've not overstretched myself, you know.* [Participant 15]

Satisfaction

The emergent theme for intervention satisfaction was Self-concept. The subthemes that emerged were self-assured, symptom reduction, and share with/recommend to others. Participants expressed surprise when reflecting on the experience and satisfaction with gaining a positive sense of self and perspective on their capabilities. Some participants commented specifically on changes in managing anxiety and overwhelm and recommended the program to other adults with ADHD. Participants considered passing on their learned experience and helping others with similar challenges a benefit.

> *Honestly, I think that in itself, it just helps to keep spirits up I suppose. Instead of going into that instant self-blame. I'm kind of like, ‘Just stop right there. Let me think about this.’ I'm not just absorbing like a sponge …Like, going back and thinking of things and forgiving myself…I think there's so many things and so many times over years that I've just taken the blame or I've been doing the one in the wrong and it's me. Me, me, me and never been me in a positive light. It's always been me that is the problem. So I think one thing that I have realized is now I'm on a much better track and I'm in a much better place, and I've got all these skills for me to utilize. And this still a lot of forgiveness. Past me needs to just to be able to let go and move on and … now that I've got the knowledge, give myself the recognition that I always needed, but I wasn't able to do before.* [Participant 19]

Discussion

Overview

This research represents the first assessment of the ADAPT framework for adults with ADHD. The primary objectives of this study were to evaluate the feasibility and acceptability of an SDT-based therapeutic coaching intervention relative to a waitlist control group. Positive participation rates, low attrition, and highly positive qualitative feedback indicate that the ADAPT framework is both feasible and acceptable to participants, including the use of multiple measures with an adult ADHD population. Regarding the appropriateness of measures, ADHD QoL was the most sensitive for analyzing the intervention; however, individual outcome measures, psychological distress, and ADHD symptom severity all contributed valuable information regarding participant experiences. A total of 5 of 7 measures showed statistical significance, indicating potential effectiveness of the intervention, although not all effects were observed in the intervention group. This may be attributed to the impact of COVID-19 on improved results over time for participants during the study. An additional finding was the correlation of recent diagnosis date with stronger ADHD symptoms and poorer self-reflection and health state when comparing pretreatment and posttreatment data, suggesting that better self-understanding leads to improved functioning. The following sections will examine these results in more detail. Potential effectiveness of the intervention will also be discussed; however, in the absence of a formal power calculation, these results should be interpreted with caution.

Feasibility and Acceptability

Participation rates were within the accepted range at 76.67%, and attrition rates were low, with 91.6% (10/11) of participants completing the Intervention and 81.8% (9/11) in the control group completing the study, indicating that the study was acceptable to participants. The significant correlation between medication use and fewer dropouts is consistent with current research and treatment recommendations for adults with ADHD, which show that psychological interventions yield better outcomes when combined with medication in multimodal treatment [10,13]. In terms of missing values, across all measures the amount was insignificant, and there were fewer missing values in the control group than in the intervention group, indicating that both inputting of responses through online forms and the number of measures required were not fatiguing and were accessible to participants. Possible reasons for missing variables are (1) known challenges with ADHD inattention on the part of the participant—each form had to be selected from a series of online links to complete the set of forms within a live session; and (2) technical issues where participants were able to complete the online form and exit or close the form without saving the data. Future studies should include steps in the protocol for the researcher to check with the participant within the session and confirm that all measures have been completed correctly, especially when delivered remotely.

Qualitative feedback from 90% (20/23) of participants indicated that the intervention as “useful” or “helpful.” Participants

XSL•FO
RenderX

described a new positive perspective on ADHD and an increase in self- awareness and self-understanding. The core themes of Change in Perspective, Practice, and Self-concept support participant acceptability through learning and engagement with the intervention. Subthemes such as self-expertise, self-assured, and awareness of choice indicate that participants felt more confident in their ability to identify and support their own needs. This is further supported by the subthemes of share with/recommend to others, indicating participants not only found it useful for themselves but were also naturally sharing their experiences of the intervention with others in similar circumstances. Participant expression of increased self-awareness and understanding may be accounted for in part by the placebo effect of the therapeutic alliance [80-82]. Validation of experience and recognition of struggle for individuals who have experienced a history of stigma and rejection for differences in presentation can have a positive effect and could facilitate positive change and growth by creating a supportive environment for participants' inherent motivation toward actualization [83].

Measure Appropriateness

Analysis of clinical distress measures showed that individual outcome, psychological distress, and ADHD symptom severity were the most effective at capturing participant experience. Of the self-development measures, ADHD QoL was the most effective at detecting impact aimed at enhancing well-being, and self-awareness provided a unique and unexpected perspective that was highly beneficial to the analysis. As multiple measures were shown to be acceptable to participants, future research should consider additional measures focusing on self-development to evaluate the effectiveness of the intervention.

Potential Effectiveness of the Intervention

Out of the 7 measures, health-related QoL and sense of autonomy did not show any statistically significant differences between assessment moments. The remaining scales showed statistically significant differences, but not all for the intervention group.

Of the remaining scales measuring clinical distress, both psychological distress and ADHD symptom severity showed positive change toward reducing clinical distress in both groups. However, psychological distress showed significant improvement for the intervention group in the subscales of well-being, problems, and functions, and in the ADHD symptom severity measure the improvement was slightly greater for the intervention group. These findings are particularly interesting as the ADAPT framework does not target specific ADHD symptoms but instead centralizes autonomy-support and satisfaction of the basic psychological needs of autonomy, competence, and relatedness. This is consistent with research demonstrating basic psychological needs satisfaction improves wellness, meaning, and vitality, as well as increasing internalization and intrinsic motivation [84]. Satisfaction of basic psychological needs, particularly autonomy, is seen as a vehicle for organization of the personality [22], supporting the formation of intrinsic preferences and thereby providing a foundation for the AIC [60].

The final measure of clinical distress, the individual outcome measure, showed higher levels of distress in the intervention group overall and significant change toward reducing clinical distress in the control group while participating in the intervention. This is an interesting finding, suggesting that the confounding variable in this analysis is related to COVID-19. In the United Kingdom, government COVID-19 safety restrictions were completely lifted on February 24, 2022 [85]. The first session for the first participant in the intervention group was May 27, 2022—only 2 months after restrictions were lifted. The first session for the first participant in the control group was October 17, 2022—a full 8 months after restrictions. Recent research on the impact of COVID-19 on individuals with ADHD showed negative impacts on mental health, sleep, and well-being outcomes, as well as treatment access [86]. Analysis results in this study also indicate that time was a factor in improved results in the intervention, as individual outcome measure assessment moment comparisons further apart for both groups yielded stronger results. Findings demonstrating statistically significant changes generally, and specifically in problems identified by participants in the individual outcome measure as outcomes for change, may be a result of the ADAPT framework's semistructured focus on client-led session design and application of model elements. Traditional treatment models and recommendations are structured to focus on symptom reduction and skills development [4,25,87-91] rather than autonomy-supportive, client-led problem identification, context-oriented strategy development, and support for basic psychological needs. Therefore, it is suggested that the differences seen in the efficacy of the intervention between the two groups were due to COVID-19 negatively impacting the intervention group, and further investigation of the intervention is warranted.

In the remaining scales for self-development, ADHD QoL showed changes in both groups, suggesting that this is also reflective of the impact of COVID-19 on QoL during the intervention versus control waitlist comparison. However, the intervention group showed an improved outlook compared with the control group, suggesting that this was the result of the intervention. This finding is notable, as most interventions for adults with ADHD showing positive results in QoL are offered for up to 15 weeks or include additional booster sessions beyond 12 weeks [92-95], whereas this intervention was offered for 11 sessions and delivered just after COVID-19 restrictions were lifted. It is recommended that future studies incorporate awareness of the circumstances and environment of participants into treatment protocols and outcomes.

The final measure, self-awareness, showed significant differences only in the control group, specifically in the need for self-reflection subscale. This finding indicates that, conversely, the intervention group recognized the value and importance of self-reflection, even if they did not significantly recognize that they were actively engaging in self-reflection. It is important to note that self-reflection was highlighted as a subtheme by participants in the qualitative analysis. Indicators of this self-awareness may be attributable to treatment components focused on self-reflection, including integrative emotion regulation, reflective AIC facilitation, and reflection

XSL•FO
RenderX

on barriers to task engagement. Research shows that integrative emotion regulation relates positively to openness to experience, authenticity, reflection [96], and well-being [97,98] and, like mindfulness, indicates active emotional exploration of experiences to determine their meaning and value in order to make informed choices about subsequent actions [99]. AIC facilitation research shows a firm sense of the AIC contributes to a sense of self-coherence and continuity and may reduce the need for dependency on external approval for self-esteem and the susceptibility to introjected internalization of goals or behaviors to maintain positive relationships [60,100]. Change process research highlights reflection as a principal vehicle of change, with potential application of active interest and reflection to low motivation, leading to movement toward health [73,74,101,102]. Elements of the model that prioritize autonomy support alongside increased self-awareness could help reduce discrepancies in self-concept, facilitating positive change [76,77,101]. One additional finding of interest was reported in the overall pretreatment and posttreatment comparisons, which showed a correlation between more recent ADHD diagnosis and stronger ADHD symptoms, poorer self-reflection, and poorer health state. This suggests that as participants gain a better understanding of themselves, they learn to cope with their difficulties over the long term and may benefit from treatments that support this process. This finding is of potential importance both to this study and to ADHD research overall, and further investigation is recommended.

Limitations and Recommendations

Generalizability

Recruitment aims for this study were to ensure a wide age range and gender inclusion. However, the small sample size and narrow geographic location of recruitment limit generalizability to a broad population of adults with ADHD. A formal sample size was not recommended; therefore, future studies should calculate the required sample size for a definitive randomized trial. The small sample size also influenced the decision not to perform a demographic data comparison for effectiveness of the intervention; this is likewise recommended for future studies.

Design

Results of this study show that additional self-development measures would be both acceptable to participants and beneficial for data capture. Future investigations should address the impact of potential factors, such as the therapeutic alliance, and include more specific measures to identify change factors, such as the emotion regulation scale [103,104], AIC measures [35], the life crafting scale [38], emotion crafting scale [29], and need crafting scale [105], some of which were not available when this study was designed and received ethical approval.

Additionally, the ADAPT framework is a novel treatment intervention that has only been applied in this study in an individual therapeutic context. The model has potential to be useful in other contexts, such as group interventions, in coaching or mentoring contexts, and with other client groups, including students, parents, and partners. Future research should investigate alternative formats, group delivery, and different client groups experiencing the impact of ADHD.

Data Collection and Reporting

This study was limited by the number of staff and researchers available to collect data and trained to deliver the intervention. This study has some risk of bias due to the lead researcher's role in both recruitment and assessment. Treatment outcomes were also primarily general self-report measures, and although one observer measure was used, it was administered by the primary researcher. Future studies should use blinding for assessment and analysis, and possibly also for intervention delivery, to reduce the risk of bias.

Finally, missing data impacted data analysis procedures in ways that can be improved. Adjusting the protocol to include confirmed measure completion and submission by the researcher would help reduce data collection issues. Additionally, requiring participants to identify a specific number of issues to evaluate in the PQ would reduce complexity and facilitate clarity in data analysis. Both changes are recommended for future studies.

Conclusion

Treatment approaches for ADHD primarily focus on compensation for EF deficit–related impairment of functioning and self-regulation and provide techniques for environmental scaffolding and skills development to promote personal strengths [4,7,25,87,106]. Heterogeneity and high rates of comorbidity alongside ADHD create challenges for the design of effective nonpharmacological treatment approaches [3,16,107], and EF remains heavily debated as a core feature of ADHD etiology, particularly as EF deficits are not unique to ADHD [10-12]. SDT provides an alternative transdiagnostic approach to psychopathology, shifting the focus from EF deficits to supporting basic psychological needs as a foundation for identity development and self-regulation [23,24,29]. This study demonstrated that the ADAPT framework is a feasible and acceptable neuroaffirmative intervention for developing naturally occurring approaches and skills in identity construction, intrinsic motivation, and self-regulation for adults with ADHD. Results indicate that the intervention is not only acceptable and accessible to participants, but participants also described the capability of passing information on to others in similar circumstances as beneficial. As the intervention does not target symptom reduction outcomes, this study suggests that outcomes of well-being and self-development have a beneficial impact on symptoms in this client group. This study also suggests that context and environment have a significant impact on self-development, as indicated by the effect of COVID-19 on outcomes over time. Results indicated that the 11-session intervention was beneficial to participants during circumstances that increased distress and at a shorter duration than other interventions for adults with ADHD targeting QoL outcomes. A key finding was participant awareness of self-reflection, demonstrated in both quantitative and qualitative results. Some increases in self-awareness and the need for self-reflection may be attributable to the therapeutic alliance; however, findings highlight both self-awareness and self-reflection as promising variables for future research in this client group. Supported by the unusual pretreatment and posttreatment comparison finding of a correlation between recent diagnosis date and stronger ADHD symptoms, poorer self-reflection, and poorer health, it

XSL•FO
RenderX

is suggested that diagnosis provides an alternative perspective of self and understanding of behaviors that improves outcomes naturally over time, which should be supported by interventions. Further research in randomized trials with greater statistical power is recommended to fully measure mediators and moderators of treatment effectiveness. These encouraging results may have significant practical implications for alternative approaches to the treatment of ADHD and transdiagnostic presentations of psychopathology.

Acknowledgments

This study received no external funding.

Authors' Contributions

REC contributed to the conceptualization, methodology, investigation, data curation, and writing—review and editing of the manuscript. RM performed the formal analysis and contributed to writing the original draft. MA, WG, SA, and BT provided supervision throughout the project. All authors have read and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
CONSORT 2010 Checklist for RCT and nonpharmacological interventions.
[DOCX File , 27 KB - formative_v9i1e69943_app1.docx]

Multimedia Appendix 2
ADAPT Framework Therapist Guide.
[DOCX File , 240 KB - formative_v9i1e69943_app2.docx]

Multimedia Appendix 3
Ethical approvals and participant forms.
[DOCX File , 530 KB - formative_v9i1e69943_app3.docx]

Multimedia Appendix 4
Statistics report.
[DOCX File , 2285 KB - formative_v9i1e69943_app4.docx]

Multimedia Appendix 5
Measures.
[DOCX File , 754 KB - formative_v9i1e69943_app5.docx]

Multimedia Appendix 6
Thematic analysis quotes.
[DOCX File , 96 KB - formative_v9i1e69943_app6.docx]

References

1. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition. Oxford: American Psychiatric Association; 2022.
2. Mullin AP, Gokhale A, Moreno-De-Luca A, Sanyal S, Waddington JL, Faundez V. Neurodevelopmental disorders: mechanisms and boundary definitions from genomes, interactomes and proteomes. Transl Psychiatry 2013;3(12):1-6 [FREE Full text] [doi: 10.1038/tp.2013.108] [Medline: 24301647]
3. Morris-Rosendahl DJ, Crocq MA. Neurodevelopmental disorders-the history and future of a diagnostic concept . Dialogues Clin Neurosci 2020;22(1):65-72 [FREE Full text] [doi: 10.31887/DCNS.2020.22.1/macrocq] [Medline: 32699506]
4. Barkley RA. Treating ADHD in Children and Adolescents: What Every Clinician Needs to Know. New York: Guilford Press; 2022.
5. Barkley RA. Executive Functions: What They Are, How They Work, and Why They Evolved. New York: Guildford Press; 2012.

XSL•FO
RenderX

6. Champ RE, Adamou M, Tolchard B. Seeking connection, autonomy, and emotional feedback: a self-determination theory of self-regulation in attention-deficit hyperactivity disorder. Psychol Rev 2023;130(3):569-603. [doi: 10.1037/rev0000398] [Medline: 36548057]
7. Champ RE, Adamou M, Tolchard B. The impact of psychological theory on the treatment of attention deficit hyperactivity disorder (ADHD) in adults: a scoping review. PLoS One 2021;16(12):e0261247 [FREE Full text] [doi: 10.1371/journal.pone.0261247] [Medline: 34932573]
8. Baumeister RF. The Self Explained: Why and How We Become Who We Are. New York: Guilford Press; 2022.
9. Berzonsky MD. A social-cognitive perspective on identity construction. In: Schwartz SJ, Luyckx K, Vignoles VL, editors. Handbook of Identity Theory and Research. New York: Springer; 2011.
10. Posner J, Polanczyk GV, Sonuga-Barke E. Attention-deficit hyperactivity disorder. Lancet 2020;395(10222):450-462 [FREE Full text] [doi: 10.1016/S0140-6736(19)33004-1] [Medline: 31982036]
11. Nigg JT, Stavro G, Ettenhofer M, Hambrick DZ, Miller T, Henderson JM. Executive functions and ADHD in adults: evidence for selective effects on ADHD symptom domains. J Abnorm Psychol 2005;114(4):706-717. [doi: 10.1037/0021-843X.114.3.706] [Medline: 16351391]
12. Castellanos FX, Sonuga-Barke EJ, Milham MP, Tannock R. Characterizing cognition in ADHD: beyond executive dysfunction. Trends Cogn Sci 2006;10(3):117-123. [doi: 10.1016/j.tics.2006.01.011] [Medline: 16460990]
13. Hoogman M, Buitelaar JK, Faraone SV, Shaw P, Franke B. Subcortical brain volume differences in participants with attention deficit hyperactivity disorder in children and adults - author's reply. Lancet Psychiatry 2017;4(6):440-441. [doi: 10.1016/S2215-0366(17)30200-6] [Medline: 28495548]
14. Wolfers T, Beckmann CF, Hoogman M, Buitelaar JK, Franke B, Marquand AF. Individual differences v the average patient: mapping the heterogeneity in ADHD using normative models. Psychol Med 2020;50(2):314-323 [FREE Full text] [doi: 10.1017/S0033291719000084] [Medline: 30782224]
15. Kooij J, Bijlenga D, Salerno L, Jaeschke R, Bitter I, Balázs J, et al. Updated European consensus statement on diagnosis and treatment of adult ADHD. Eur Psychiatry 2019;56:14-34 [FREE Full text] [doi: 10.1016/j.eurpsy.2018.11.001] [Medline: 30453134]
16. Cordova M, Shada K, Demeter DV, Doyle O, Miranda-Dominguez O, Perrone A, et al. Heterogeneity of executive function revealed by a functional random forest approach across ADHD and ASD. Neuroimage Clin 2020;26:102245 [FREE Full text] [doi: 10.1016/j.nicl.2020.102245] [Medline: 32217469]
17. Carr-Fanning K. The right to dignity or disorder? The case for attention deficit hyperactivity diversity. Stud Arts Humanit 2020;6(1):14-29. [doi: 10.18193/sah.v6i1.192]
18. Dwyer SC. Overcoming obstacles to education: The experience of women university students diagnosed with attention-deficit/hyperactivity disorder. Can J High Educ 2000;30(1):123-147.
19. Hoben JL, Hesson J. Invisible lives: using autoethnography to explore the experiences of academics living with attention deficit hyperactivity disorder (ADHD). New Horizons Adult Educ Hum Resour Dev 2021;33(1):37-50. [doi: 10.1002/nha3.20304]
20. Horton-Salway M, Davies A. Media representations of ADHD. In: The Discourse of ADHD. The Language of Mental Health. Cham: Palgrave Macmillan; 2018.
21. Spiel K, Hornecker E, Williams RM, Good J. ADHD and technology research - Investigated by neurodivergent readers. 2022 Presented at: CHI '22: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems; April 29, 2022; New Orleans p. 1-21. [doi: 10.1145/3491102.3517592]
22. Watters C, Adamis D, McNicholas F, Gavin B. The impact of attention deficit hyperactivity disorder (ADHD) in adulthood: a qualitative study. Ir J Psychol Med 2018;35(3):173-179. [doi: 10.1017/ipm.2017.21] [Medline: 30124180]
23. Young Z, Tickle A, Gillott A, Groom M. Psychological impact of an adult ADHD diagnosis: “A blessing and a curse?”. Life span disabil 2019;XXII(2):173-203.
24. Chapman R. Neurotype dysphoria. Psychology Today. URL: https://www.psychologytoday.com/gb/blog/neurodiverse-age/202006/neurotype-dysphoria [accessed 2023-09-20]
25. Barkley RA. Taking Charge of Adult ADHD: Proven Strategies to Succeed at Work, at Home, and in Relationships. New York: Guilford Press; 2022.
26. Barkley RA, Murphy KR, Fischer M. ADHD in Adults: What the Science Says. New York: Guilford Press; 2010.
27. Brooks RB. Nurturing islands of competence: is there really room for a strength-based model in the treatment of ADHD? The ADHD Report 2001;9(2):1-5. [doi: 10.1521/adhd.9.2.1.19075]
28. Demontis D, Walters RK, Martin J, Mattheisen M, Als TD, Agerbo E, ADHD Working Group of the Psychiatric Genomics Consortium (PGC), 23andMe Research Team, et al. Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. Nat Genet 2019;51(1):63-75. [doi: 10.1038/s41588-018-0269-7] [Medline: 30478444]
29. Dipeolu AO. College students with ADHD: prescriptive concepts for best practices in career development. J Career Dev 2010;38(5):408-427. [doi: 10.1177/0894845310378749]
30. Ramsay JR. “Turning intentions into actions”: CBT for adult ADHD focused on implementation. Clin Case Stud 2015;15(3):179-197. [doi: 10.1177/1534650115611483]

XSL•FO
RenderX

31. Ramsay JR, Rostain AL. Adult ADHD research: current status and future directions. J Atten Disord 2008;11(6):624-627. [doi: 10.1177/1087054708314590] [Medline: 18417728]
32. Attention Deficit Hyperactivity Disorder: Diagnosis and Management. London: National Institute for Health and Care Excellence (NICE); 2019. URL: https://www.nice.org.uk/guidance/ng87 [accessed 2023-09-20]
33. Fullen T, Jones SL, Emerson LM, Adamou M. Psychological treatments in adult ADHD: a systematic review. J Psychopathol Behav Assess 2020;42(3):500-518. [doi: 10.1007/s10862-020-09794-8]
34. de Carvalho NP, da Silva AR, Caiafa L, de Paula JJ. Efficacy of cognitive-behavioral therapy for adults with ADHD: a systematic review. Paidéia (Ribeirão Preto) 2024;34:1-11 [FREE Full text] [doi: 10.1590/1982-4327e3420]
35. Nimmo-Smith V, Merwood A, Hank D, Brandling J, Greenwood R, Skinner L, et al. Non-pharmacological interventions for adult ADHD: a systematic review. Psychol Med 2020;50(4):529-541. [doi: 10.1017/S0033291720000069] [Medline: 32036811]
36. López-Pinar C, Martínez-Sanchís S, Carbonell-Vayá E, Sánchez-Meca J, Fenollar-Cortés J. Efficacy of nonpharmacological treatments on comorbid internalizing symptoms of adults with attention-deficit/hyperactivity disorder: a meta-analytic review. J Atten Disord 2020;24(3):456-478. [doi: 10.1177/1087054719855685] [Medline: 31189374]
37. Liu C, Hua M, Lu M, Goh KK. Effectiveness of cognitive behavioural-based interventions for adults with attention-deficit/hyperactivity disorder extends beyond core symptoms: a meta-analysis of randomized controlled trials. Psychol Psychother 2023;96(3):543-559. [doi: 10.1111/papt.12455] [Medline: 36794797]
38. Ochsner KN, Bunge SA, Gross JJ, Gabrieli JDE. Rethinking feelings: an FMRI study of the cognitive regulation of emotion. J Cogn Neurosci 2002;14(8):1215-1229. [doi: 10.1162/089892902760807212] [Medline: 12495527]
39. Roberts W, Milich R, Barkley RA. Primary symptoms, diagnostic criteria, subtyping, and prevalence of ADHD. In: Barkley RA, editor. Attention-Deficit Hyperactivity Disorder: A Handbook for Diagnosis and Treatment. 4th Edition. New York: Guilford Press; 2014:51-80.
40. Sergeant J. The cognitive-energetic model: an empirical approach to attention-deficit hyperactivity disorder. Neurosci Biobehav Rev 2000;24(1):7-12. [doi: 10.1016/s0149-7634(99)00060-3] [Medline: 10654654]
41. Chevrier A, Schachar RJ. BOLD differences normally attributed to inhibitory control predict symptoms, not task-directed inhibitory control in ADHD. J Neurodev Disord 2020;12(1):8 [FREE Full text] [doi: 10.1186/s11689-020-09311-8] [Medline: 32085698]
42. Dalgleish T, Black M, Johnston D, Bevan A. Transdiagnostic approaches to mental health problems: current status and future directions. J Consult Clin Psychol 2020;88(3):179-195 [FREE Full text] [doi: 10.1037/ccp0000482] [Medline: 32068421]
43. Astle DE, Holmes J, Kievit R, Gathercole SE. Annual research review: the transdiagnostic revolution in neurodevelopmental disorders. J Child Psychol Psychiatry 2022;63(4):397-417. [doi: 10.1111/jcpp.13481] [Medline: 34296774]
44. Mahadevan J, Kandasamy A, Benegal V. Situating adult attention-deficit/hyperactivity disorder in the externalizing spectrum: etiological, diagnostic, and treatment considerations. Indian J Psychiatry 2019;61(1):3-12 [FREE Full text] [doi: 10.4103/psychiatry.IndianJPsychiatry_549_18] [Medline: 30745648]
45. East-Richard C, Mercier AR, Nadeau D, Cellard C. Transdiagnostic neurocognitive deficits in psychiatry: a review of meta-analyses. Can Psychol 2020;61(3):190-214. [doi: 10.1037/cap0000196]
46. Musser ED, Raiker JS. Attention-deficit/hyperactivity disorder: an integrated developmental psychopathology and research domain criteria (RDoC) approach. Compr Psychiatry 2019;90:65-72 [FREE Full text] [doi: 10.1016/j.comppsych.2018.12.016] [Medline: 30743139]
47. Mitelman SA. Transdiagnostic neuroimaging in psychiatry: a review. Psychiatry Res 2019;277:23-38. [doi: 10.1016/j.psychres.2019.01.026] [Medline: 30639090]
48. Soenens B, Vansteenkiste M. A lifespan perspective on the importance of the basic psychological needs for psychosocial development. In: The Oxford Handbook of Self-Determination Theory. New York: Oxford University Press; 2023:457-490.
49. Luyckx K, Soenens B, Vansteenkiste M, Goossens L, Berzonsky MD. Parental psychological control and dimensions of identity formation in emerging adulthood. J Fam Psychol 2007;21(3):546-550. [doi: 10.1037/0893-3200.21.3.546] [Medline: 17874942]
50. van der Kaap-Deeder J. The transdiagnostic role of the basic psychological needs in psychopathology. In: Ryan RM, editor. The Oxford Handbook of Self-Determination Theory. New York: Oxford University Press; 2023:819-836.
51. Oram R, Rogers M, DuPaul G. Explaining the relationship between ADHD symptomatology and amotivation in the undergraduate population: the role of basic psychological need frustration. Can J Sch Psychol 2019;35(2):139-153. [doi: 10.1177/0829573519880063]
52. Morsink S, Van der Oord S, Antrop I, Danckaerts M, Scheres A. Studying motivation in ADHD: the role of internal motives and the relevance of self determination theory. J Atten Disord 2022;26(8):1139-1158 [FREE Full text] [doi: 10.1177/10870547211050948] [Medline: 34794343]
53. Serrano JW, Abu-Ramadan TM, Vasko JM, Leopold DR, Canu WH, Willcutt EG, et al. ADHD and psychological need fulfillment in college students. J Atten Disord 2023;27(8):912-924. [doi: 10.1177/10870547231161530] [Medline: 36924424]
54. Oram R, Daffer-Deming M, Ahn JH, Rogers MA. Basic psychological needs of undergraduate students with and without high ADHD symptomatology. Can J Educ Soc Stud 2023;3(1):17-28. [doi: 10.53103/cjess.v3i1.94]

XSL•FO
RenderX

55. Zaghi AE, Grey A, Hain A, Syharat CM. “It seems like I’m doing something more important”—An interpretative phenomenological analysis of the transformative impact of research experiences for STEM students with ADHD. Educ Sci 2023;13(8):1-19. [doi: 10.3390/educsci13080776]
56. Pirozzi M. The Perception of the College Experience for Students With ADHD. Boston: Northeastern University; 2021.
57. Ryan RM, Deci EL, Vansteenkiste M. Autonomy and autonomy disturbances in self-development and psychopathology: research on motivation, attachment, and clinical process. In: Cicchetti D, editor. Developmental Psychopathology: Theory and Method. 3rd Edition. New Jersey: John Wiley & Sons; 2016:385-438.
58. Champ RE, Adamou M, Gillibrand W, Arrey S, Tolchard B. The creative awareness theory: a grounded theory study of inherent self-regulation in attention deficit hyperactivity disorder. J Clin Med 2024;13(19):5963 [FREE Full text] [doi: 10.3390/jcm13195963] [Medline: 39408023]
59. Assor A, Benita M, Shi Y, Goren R, Yitshaki N, Wang Q. The authentic inner compass as a well-being resource: predictive effects on vitality, and relations with self-esteem, depression and behavioral self-realization. J Happiness Stud 2021;22(8):3435-3455. [doi: 10.1007/s10902-021-00373-6]
60. Assor A, Benita M, Geifman Y. The authentic inner compass as an important motivational experience and structure: antecedents and benefits. In: Ryan RM, editor. The Oxford Handbook of Self-Determination Theory. New York: Oxford University Press; 2023:362-386.
61. Assor A, Soenens B, Yitshaki N, Ezra O, Geifman Y, Olshtein G. Towards a wider conception of autonomy support in adolescence: the contribution of reflective inner-compass facilitation to the formation of an authentic inner compass and well-being. Motiv Emot 2019;44(2):159-174. [doi: 10.1007/s11031-019-09809-2]
62. Chen S, van der Meij L, van Zyl LE, Demerouti E. The life crafting scale: development and validation of a multi-dimensional meaning-making measure. Front Psychol 2022;13:795686 [FREE Full text] [doi: 10.3389/fpsyg.2022.795686] [Medline: 35330727]
63. van der Kaap-Deeder J, Wichstrøm L, Mouratidis A, Matos L, Steinsbekk S. Emotion crafting: individuals as agents of their positive emotional experiences. Motiv Emot 2023;47(5):870-886. [doi: 10.1007/s11031-023-10035-0]
64. Laporte N, Soenens B, Flamant N, Vansteenkiste M, Mabbe E, Brenning K. The role of daily need crafting in daily fluctuations in adolescents’ need-based and affective experiences. Motiv Emot 2021;46(2):137-149. [doi: 10.1007/s11031-021-09921-2]
65. Eldridge SM, Chan CL, Campbell MJ, Bond CM, Hopewell S, Thabane L, PAFS consensus group. CONSORT 2010 statement: extension to randomised pilot and feasibility trials. BMJ 2016;355:i5239 [FREE Full text] [doi: 10.1136/bmj.i5239] [Medline: 27777223]
66. Boutron I, Moher D, Altman DG, Schulz KF, Ravaud P, CONSORT Group. Extending the CONSORT statement to randomized trials of nonpharmacologic treatment: explanation and elaboration. Ann Intern Med 2008;148(4):295-309 [FREE Full text] [doi: 10.7326/0003-4819-148-4-200802190-00008] [Medline: 18283207]
67. Krauss A. Assessing the overall validity of randomised controlled trials. Int Stud Philos Sci 2021;34(3):159-182. [doi: 10.1080/02698595.2021.2002676]
68. Whitehead AL, Julious SA, Cooper CL, Campbell MJ. Estimating the sample size for a pilot randomised trial to minimise the overall trial sample size for the external pilot and main trial for a continuous outcome variable. Stat Methods Med Res 2016;25(3):1057-1073 [FREE Full text] [doi: 10.1177/0962280215588241] [Medline: 26092476]
69. Lewis M, Bromley K, Sutton CJ, McCray G, Myers HL, Lancaster GA. Determining sample size for progression criteria for pragmatic pilot RCTs: the hypothesis test strikes back!. Pilot Feasibility Stud 2021;7(1):40 [FREE Full text] [doi: 10.1186/s40814-021-00770-x] [Medline: 33536076]
70. Mellor-Clark J. CORE: clinical outcomes in routine evaluation. J Ment Heal 2009;9(3):247-255. [doi: 10.1080/jmh.9.3.247.255]
71. Conners CK, Erhardt D, Sparrow EP. Conners’ Adult ADHD Rating Scales: Technical Manual. New York: Multi-Health Systems; 1999.
72. Adler LA, Faraone SV, Spencer TJ, Michelson D, Reimherr FW, Glatt SJ, et al. The reliability and validity of self- and investigator ratings of ADHD in adults. J Atten Disord 2008;11(6):711-719. [doi: 10.1177/1087054707308503] [Medline: 18025250]
73. Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, Text Revision (DSM-IV-TR). Washington DC: American Psychiatric Association; 2000.
74. Brod M, Johnston J, Able S, Swindle R. Validation of the adult attention-deficit/hyperactivity disorder quality-of-life scale (AAQoL): a disease-specific quality-of-life measure. Qual Life Res 2006;15(1):117-129. [doi: 10.1007/s11136-005-8325-z] [Medline: 16411036]
75. Vansteenkiste M, Ryan RM. On psychological growth and vulnerability: basic psychological need satisfaction and need frustration as a unifying principle. J Psychother Integr 2013;23(3):263-280. [doi: 10.1037/a0032359]
76. Green D. Making the case for using personalised outcome measures to track progress in psychotherapy. Eur J Psychother Couns 2016;18(1):39-57. [doi: 10.1080/13642537.2015.1130075]
77. Elliott R, Wagner J, Sales CMD, Rodgers B, Alves P, Café MJ. Psychometrics of the personal questionnaire: a client-generated outcome measure. Psychol Assess 2016;28(3):263-278 [FREE Full text] [doi: 10.1037/pas0000174] [Medline: 26075406]

XSL•FO
RenderX

78. Data Protection Act 1998. 1998. URL: https://www.legislation.gov.uk/ukpga/1998/29/contents [accessed 2014-12-29]
79. Mak ADP, Lee S, Sampson NA, Albor Y, Alonso J, Auerbach RP, et al. ADHD comorbidity structure and impairment: results of the WHO World Mental Health Surveys International College Student Project (WMH-ICS). J Atten Disord 2022;26(8):1078-1096 [FREE Full text] [doi: 10.1177/10870547211057275] [Medline: 34753324]
80. Ursano AM, Sonnenberg SM, Ursano RJ. Physician–patient relationship. In: Tasman A, Kay J, Lieberman JA, First MB, Riba MB, editors. Psychiatry, 4th Edition. West Sussex: John Wiley & Sons; 2015.
81. Verhulst J, Kramer D, Swann AC, Hale-Richlen B, Beahrs J. The medical alliance: from placebo response to alliance effect. J Nerv Ment Dis 2013;201(7):546-552. [doi: 10.1097/NMD.0b013e31829829e1] [Medline: 23817150]
82. Arnd-Caddigan M. The therapeutic alliance: implications for therapeutic process and therapeutic goals. J Contemp Psychother 2012;42(2):77-85. [doi: 10.1007/s10879-011-9183-3]
83. Rogers CR. The necessary and sufficient conditions of therapeutic personality change. J Consult Psychol 1957;21(2):95-103. [doi: 10.1037/h0045357] [Medline: 13416422]
84. Vansteenkiste M, Soenens B, Ryan RM. Basic psychological needs theory: a conceptual and empirical review of key criteria. In: Ryan RM, editor. The Oxford Handbook of Self-Determination Theory. Oxford: Oxford University Press; 2023:84-123.
85. Covid: end of legal need to self-isolate in England. BBC. URL: https://www.bbc.com/news/uk-60500287 [accessed 2024-02-12]
86. Segenreich D. The impact of the COVID-19 pandemic on diagnosing and treating attention deficit hyperactivity disorder: new challenges on initializing and optimizing pharmacological treatment. Front Psychiatry 2022;13:1-7 [FREE Full text] [doi: 10.3389/fpsyt.2022.852664] [Medline: 35463503]
87. Ramsay JR, Rostain AL. The Adult ADHD Toolkit: using CBT to facilitate coping inside and out. New York: Routledge; 2015.
88. Solanto MV. Cognitive-Behavioural Therapy for Adult ADHD. New York: Guildford Press; 2010.
89. Safren SA, Perlman C, Sprich S, Otto M. Mastering Your Adult ADHD. New York: Oxford University Press; 2005.
90. Mitchell JT, Zylowska L, Kollins SH. Mindfulness meditation training for attention-deficit/hyperactivity disorder in adulthood: current empirical support, treatment overview, and future directions. Cogn Behav Pract 2015;22(2):172-191 [FREE Full text] [doi: 10.1016/j.cbpra.2014.10.002] [Medline: 25908900]
91. Zylowska L. The mindfulness prescription for adult ADHD. Colorado: Trumpeter; 2012.
92. Kastner L, Velder-Shukrun Y, Bonne O, Bar-Ilan RT, Maeir A. Pilot study of the cognitive-functional intervention for adults (Cog-Fun A): a metacognitive-functional tool for adults with attention deficit hyperactivity disorder. Am J Occup Ther 2022;76(2):7602205070. [doi: 10.5014/ajot.2022.046417] [Medline: 35258515]
93. Huang F, Tang Y, Zhao M, Wang Y, Pan M, Wang Y, et al. Cognitive-behavioral therapy for adult ADHD: a randomized clinical trial in China. J Atten Disord 2019;23(9):1035-1046. [doi: 10.1177/1087054717725874] [Medline: 28866911]
94. Lücke C, Jenkner C, Graf E, Matthies S, Borel P, Sobanski E, et al. Long-term improvement of quality of life in adult ADHD – results of the randomized multimodal COMPAS trial. J Ment Health 2021;50(3):250-270. [doi: 10.1080/00207411.2021.1910172]
95. Halmøy A, Ring AE, Gjestad R, Møller M, Ubostad B, Lien T, et al. Dialectical behavioral therapy-based group treatment versus treatment as usual for adults with attention-deficit hyperactivity disorder: a multicenter randomized controlled trial. BMC Psychiatry 2022;22(1):738 [FREE Full text] [doi: 10.1186/s12888-022-04356-6] [Medline: 36443712]
96. Roth G, Shahar BH, Zohar-Shefer Y, Benita M, Moed A, Bibi U, et al. Benefits of emotional integration and costs of emotional distancing. J Pers 2018;86(6):919-934. [doi: 10.1111/jopy.12366] [Medline: 29222933]
97. Benita M. Freedom to feel: a self‐determination theory account of emotion regulation. Soc Personal Psychol Compass 2020;14(11):1-13. [doi: 10.1111/spc3.12563]
98. Brenning K, Soenens B, Van Petegem S, Vansteenkiste M. Perceived maternal autonomy support and early adolescent emotion regulation: a longitudinal study. Soc Dev 2015;24(3):561-578. [doi: 10.1111/sode.12107]
99. Roth G, Benita M. Integration versus minimization of emotional experiences: addressing adaptive emotion regulation. In: Ryan RM, editor. The Oxford Handbook of Self-Determination Theory. New York: Oxford University Press; 2023:200-214.
100. Kanat-Maymon Y, Assor A, Roth G. Conditional regard in development and relationships. In: Ryan RM, editor. The Oxford Handbook of Self-Determination Theory. Oxford: Oxford University Press; 2023.
101. Grant AM, Franklin J, Langford P. The Self-Reflection and Insight Scale: a new measure of private self-consciousness. Soc Behav Pers: Int J Sci Res 2002;30(8):821-835. [doi: 10.2224/sbp.2002.30.8.821]
102. Weinstein N, Przybylski AK, Ryan RM. The index of autonomous functioning: development of a scale of human autonomy. J Res Pers 2012;46(4):397-413. [doi: 10.1016/j.jrp.2012.03.007]
103. Roth G, Assor A, Niemiec CP, Deci EL, Ryan RM. The emotional and academic consequences of parental conditional regard: comparing conditional positive regard, conditional negative regard, and autonomy support as parenting practices. Dev Psychol 2009;45(4):1119-1142. [doi: 10.1037/a0015272] [Medline: 19586183]
104. Roth G, Benita M, Amrani C, Shachar B, Asoulin H, Moed A, et al. Integration of negative emotional experience versus suppression: addressing the question of adaptive functioning. Emotion 2014;14(5):908-919. [doi: 10.1037/a0037051] [Medline: 24866531]

XSL•FO
RenderX

105. Laporte N, Soenens B, Brenning K, Vansteenkiste M. Adolescents as active managers of their own psychological needs: the role of psychological need crafting in adolescents' mental health. J Adolesc 2021;88:67-83. [doi: 10.1016/j.adolescence.2021.02.004] [Medline: 33662706]
106. Ramsay JR. Rethinking adult ADHD. Washington DC: American Psychological Association; 2020.
107. Nigg JT, Karalunas SL, Feczko E, Fair DA. Toward a revised nosology for attention-deficit/hyperactivity disorder heterogeneity. Biol Psychiatry Cogn Neurosci Neuroimaging 2020;5(8):726-737 [FREE Full text] [doi: 10.1016/j.bpsc.2020.02.005] [Medline: 32305325]

Abbreviations

AAQoL: Attention-Deficit/Hyperactivity Disorder Quality of Life Scale
ADAPT: Autonomy, Design, Awareness, Psychoeducation, and Training Integration for Sustainable Change
ADHD: attention-deficit/hyperactivity disorder
ADHDRS-IV-Inv: Attention-Deficit/Hyperactivity Disorder Rating Scale, Investigator-Administered
AIC: authentic inner compass
CBT: cognitive behavioral therapy
CONSORT: Consolidated Standards of Reporting Trials
CORE-OM: Clinical Outcomes in Routine Evaluation–Outcome Measure
DSM-5-TR: Diagnostic and Statistical Manual of Mental Disorders (Fifth Edition, Text Revision)
EF: executive function
IAF: Index of Autonomous Functioning
NHS: National Health Service
PQ: personal questionnaire
QoL: quality of life
RCT: randomized controlled trial
SDT: Self-Determination Theory
SR&I: Self-Reflection and Insight Scale

Edited by A Mavragani; submitted 11.12.24; peer-reviewed by A Rebar, KW Tay; comments to author 13.04.25; revised version received 23.07.25; accepted 12.08.25; published 29.10.25.

Please cite as:
Champ RE, Meneses RW, Adamou M, Gillibrand W, Arrey S, Tolchard B
A Neuroaffirmative, Self-Determination Theory–Based Psychosocial Intervention for Adults With Attention-Deficit/Hyperactivity Disorder: Randomized Feasibility Study
JMIR Form Res 2025;9:e69943
URL: https://formative.jmir.org/2025/1/e69943
doi:10.2196/69943
PMID:

©Rebecca Elizabeth Champ, Rita Wengorovius Meneses, Marios Adamou, Warren Gillibrand, Sally Arrey, Barry Tolchard. Originally published in JMIR Formative Research (https://formative.jmir.org), 29.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Reconditioning Emotional Responses With the Break Method: Pilot Quantitative Study

Boaz Salik[1*], PhD; Bizzie Gold[1*]; Kira Krier[1*], LICSW

Break Method, Rewire Lab Corporation, Sandpoint, ID, United States

[*]all authors contributed equally

Corresponding Author:
Bizzie Gold
Break Method, Rewire Lab Corporation
400 Church St
Sandpoint, ID, 83864
United States
Phone: 1 8083152171
Email: info@bizziegold.com

Abstract

Background: The Break Method is a structured, behavior-based emotional reconditioning program designed to help individuals gain insight into patterns of emotional dysregulation and reprogram behavioral responses rooted in past experiences. Although it has been widely adopted in private and small-group settings, empirical evidence supporting its effectiveness remains limited. With increasing interest in accessible, scalable, and personalized mental health interventions, evaluating the outcomes of such programs is essential for informing future implementation and research.

Objective: This pilot study aimed to evaluate changes in self-reported mental health status before and after participation in the Break Method program. Specifically, we sought to examine (1) overall trends in mental health improvement, (2) associations between specific reasons for joining the program and changes in mental health outcomes, and (3) latent clusters of participant motivations based on symptom profiles.

Methods: Data were collected from 175 unique participants, yielding 195 total survey responses (as 15 participants completed the program more than once). Participants rated their mental health status on a 5-point Likert scale both before and after the program (this was not a validated clinical measure, limiting the interpretability and comparability of results). Descriptive statistics and paired 2-tailed *t* tests were used to assess pre- and postprogram differences in Likert scores. McNemar tests were conducted to compare categorical mental health status (Likert score ≥4 vs <4) before and after participation. Analyses of covariance examined score changes across groups stratified by reported reasons for joining. Multiple correspondence analysis was used to explore latent symptom clusters.

Results: Before program participation, 186 of 195 (95.4%) responses reported Likert scores below 4. Following the program, 157 (80.5%) responses reported scores of 4 or higher. A significant improvement in mental health status was observed (preprogram mean score 2.07 SD 0.82, postprogram mean score 3.92 SD 0.73; $P<.001$). Significant, positive changes were associated with reasons including anxiety (β=0.332, 95% CI 0.073-0.591), obsessive-compulsive disorder (β=0.455, 95% CI 0.061-0.850), and a history of self-harm or suicidal ideation (β=0.511, 95% CI 0.091-0.931). The multiple correspondence analysis identified three clusters of participants based on symptom profiles: (1) low self-image (eg, depression, self-sabotage, and relationship issues); (2) life-development goals (eg, self-discovery and future planning); and (3) obsessive-compulsive disorder–related symptoms. The first cluster was significantly associated with improved mental health outcomes (β=0.348, 95% CI 0.060-0.636).

Conclusions: The Break Method appears to be a promising intervention for improving mental health, particularly among individuals reporting anxiety, low confidence, or a history of self-sabotage. However, due to the single-group, preprogram-postprogram design without a control group, causality cannot be inferred, and these findings should be interpreted as preliminary associations rather than confirmed efficacy. Future studies should incorporate standardized clinical tools, control groups, and longitudinal designs to validate these results and explore long-term outcomes across diverse populations.

(JMIR Form Res 2025;9:e75056) doi:10.2196/75056

KEYWORDS

mental health intervention; behavior-based therapy; self-guided program; emotional dysregulation; emotional reconditioning

XSL•FO
RenderX

Introduction

Recent studies have linked mental health symptoms, such as depression, to distinct patterns of brain activity observed in nonclinical populations [1]. This growing understanding of the neurobiological and behavioral dimensions of emotional health highlights the importance of interventions that can target maladaptive patterns and promote adaptive coping strategies [2,3]. In response, the Break Method offers a structured, behavior-based approach to emotional reconditioning, combining a data-driven framework with individualized strategies to support emotional regulation and behavioral change.

The Break Method aims to help participants identify and interrupt subconscious patterns that drive automatic, often self-sabotaging behaviors. Although it is not grounded in formal neuroscience protocols, the program draws conceptually from cognitive behavioral principles such as pattern recognition, emotional regulation, and schema restructuring, providing tools for sustainable self-directed change [2,3]. Unlike many traditional interventions that may foster dependency or temporary relief, the Break Method is structured around a defined timeline; self-assessment checkpoints; and a combination of interactive modules, one-on-one sessions, and personalized strategy plans, emphasizing autonomy and long-term skill development. Digital tools have shown potential in reducing stigma and facilitating help-seeking among individuals reluctant to access traditional therapy [4].

This pilot study was conducted to generate foundational insights into the program's effectiveness and identify patterns of mental health change among participants. Specifically, the study aimed to assess whether improvements in self-reported mental health status varied based on participants' reasons for enrolling and to explore which psychological profiles might benefit most from this intervention. Understanding these relationships can help optimize the application of the Break Method and inform future research in behavior-based mental health interventions [5].

The main objectives were (1) to evaluate the overall efficacy of the program in improving self-reported mental health symptoms and (2) to determine which specific symptoms or joining reasons are most strongly associated with improvements following participation.

Methods

Recruitment

This study recruited participants between fall 2018 and spring 2022. Participants were paying customers of the Break Method who voluntarily enrolled in the program. There were no specific criteria for inclusion based on demographic background, as all individuals who participated in the program during this time frame were invited to complete the survey. Participation in the survey was considered for clients who had opted to be surveyed at program entry. As part of the intake process, participants were screened using a standardized onboarding form, and those reporting active suicidal ideation or crisis-level symptoms were referred to external mental health services and excluded from the program. The survey questionnaire was administered online before and after the intervention. The Break Method in this study was not based on a specific theoretical framework. The questionnaire covered the following topics: early childhood and family relationships, academics, arts and sports, current career and relationship information, and reason for joining the program.

The program is entirely self-paced, with most participants completing it within 4 to 6 months. It includes 4 interactive video modules, 6 one-on-one sessions with a dedicated behavior strategist, a 140-page workbook, and a final strategy session that generates a customized behavior strategy plan ranging from 30 to 100 pages. Participants also had access to a dedicated client support specialist and a community forum for peer support. All activities were conducted through a secure online portal, accessible only via unique usernames and passwords.

Upon program completion, participants were required to complete a survey as part of the exit interview process. The survey was administered via Google Forms and collected participants' names, email addresses, self-reported mental health status before and after the program, frequency of participation, and reasons for joining. The question about participants' reasons for joining included 13 response options, allowing participants to select multiple answers: anxiety; depression; relationship issues; poor self-image/low confidence; disordered eating; history of self-harm/suicidal ideation; obsessive-compulsive disorder (OCD)/counting/excessive worry with a calculating nature; parenting issues; addiction; career/workplace issues; self-sabotage/stagnation; future goals/next steps; and general self-discovery. Instructions emphasized the importance of providing honest responses to improve the program's efficacy.

Demographic data, including age, biological sex, and gender, were also collected but not analyzed in this study. From an initial 286 participants, 20 duplicate responses were removed. Duplicate entries occurred when participants interrupted the response process and resumed it later, resulting in multiple entries for the same individual. Among the remaining participants, those who did not respond to questions about the degree of symptoms (n=91) were excluded, leaving 175 participants in the final analysis.

To ensure transparent and complete reporting, this study followed the CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) reporting guidelines for digital interventions.

Outcomes

Participants' mental health status was assessed using a 5-point Likert scale for 2 questions in the survey: "How would you rate your symptoms at the start of the Break Method?" "How would you rate your symptoms after completing the Break Method?" Responses ranged from 0 (terrible or debilitating) to 5 (symptoms resolved). Changes in mental health were measured by the difference between the pre- and postprogram Likert scores. We also assessed mental health status by categorizing the Likert scores into 2 groups: 1 for scores ≥4 and 0 for scores <4, applied to both pre- and postprogram scores. We defined a Likert score of ≥4 as *improved* to capture positive changes in self-reported mental health status, as higher scores reflect more

XSL•FO
RenderX

favorable outcomes (eg, 4=good and 5=very good). Participants rated their mental health status on a 5-point Likert scale adapted from population health survey frameworks such as the Behavioral Risk Factor Surveillance System [6].

Ethical Considerations

This study was conducted in adherence to ethical principles governing research involving human participants. All participants provided informed consent via signed agreements facilitated through DocuSign (DocuSign Inc) before taking part in the survey. The consent process ensured that participants were fully informed about the study's purpose, procedures, and the intended use of their responses. Inclusion or exclusion criteria applied before participation and analysis included excluding individuals actively experiencing suicidal ideation and those requiring crisis-oriented mental health care. Participation in the survey was limited to clients who consented at program entry and completed the program, which may introduce selection bias by excluding dropouts or nonconsenting individuals. However, because the program is voluntary and to respect participants' choices, this approach was necessary while still allowing us to gather useful information about the program's effects.

This study qualifies for exemption under 45 CFR 46.104(d)(3), in accordance with the US Department of Health and Human Services (HHS) exemption decision tool (Chart 5), which applies to research involving benign behavioral interventions. The Break Method program, which includes structured emotional skills and self-assessment exercises, constitutes a benign behavioral intervention. Information was collected through self-administered online surveys.

Although some of the information originally recorded (eg, demographic or personal experiences) could have been potentially identifiable, the dataset accessed for this study was fully deidentified before analysis. No identifying variables such as names, contact information, or IP addresses were included in the data available to researchers, and no keys to reidentify participants were retained or accessed. Furthermore, there was no attempt to link survey responses back to any individual.

An institutional review board (IRB) did not conduct a limited review of the study, as permitted under 45 CFR 46.104(d)(3) and illustrated in Chart 5 of the HHS exemption flowchart. A limited review is not necessary when the responses collected do not pose a risk of criminal or civil liability or reputational harm to participants. The topics addressed in the survey, including early childhood and family relationships, academic background, creative and athletic interests, current career and relationship experiences, and reasons for joining the program did not place participants at risk in the event of disclosure.

This research involved a retrospective analysis of data collected between fall 2018 and spring 2022 as part of the Break Method's internal program assessments. The surveys were administered as part of routine program participation, not for research purposes. No additional data were collected specifically for this study.

In addition, participants exhibiting signs of active suicidal ideation or requiring immediate, crisis-oriented mental health care were excluded from the program at entry, in accordance with the Break Method's safety protocols. Therefore, the dataset did not include responses from individuals in acute psychological crisis. All data analyzed reflected minimal-risk, retrospective self-reports from participants who completed the full program.

Although this study did not require formal IRB oversight, the research team followed recognized ethical guidelines for studies involving human data. This included respect for participant autonomy, informed consent at point of entry, data confidentiality, and responsible research practices. This exemption classification and ethical posture are consistent with published precedents in peer-reviewed studies, where minimal-risk survey data deidentified and retrospectively analyzed have qualified under 45 CFR 46.104(d)(3) or (d)(2) without formal IRB review.

Statistical Analysis

Statistical analyses were performed to (1) assess overall changes in symptom severity after participating in the program and (2) examine how changes in symptom severity were associated with the presence of each type of symptom or reasons for joining the program.

The trend of overall changes in mental health symptom severity was estimated using a paired *t* test, comparing the preprogram and postprogram 5-point Likert scale scores for all participants. In addition, the McNemar test, which analyzes paired nominal data in a 2×2 contingency table, was conducted to assess whether there was a significant change in mental health status after the program. This test was applied to a binary variable, where a score ≥4 was coded as 1 and a score ≤4 was coded as 0, both before and after the program.

To explore the second question, the Mann-Whitney *U* test was used to compare preprogram and postprogram means of the Likert scores by the status of each mental health symptom or reason for joining the program. For example, the difference in score changes after the program was compared among participants with and without self-reported responses for anxiety. Additionally, the Cochran-Mantel-Haenszel test [2] was applied to binary variables (ie, 1 for scores ≥4 and 0 for scores <4) of the Likert scale measuring mental health status to assess whether changes in outcomes were significantly associated with each mental health symptom or reason for joining the program.

In our analysis, we modeled change scores as the dependent variable to directly reflect the magnitude of improvement across participant-reported symptom categories. Although analysis of covariance (ANCOVA) is more commonly used to model postintervention scores with baseline scores as covariates, our approach is valid for examining within-subject score changes and between-group differences. We assumed that using change scores was appropriate, given that the actual change from baseline was our outcome of interest and baseline scores appeared similar across participant symptom groups [3]. Nevertheless, future analyses may consider this. Both ANCOVA and change score analyses were conducted to examine pre- and postprogram differences, with ANCOVA controlling for baseline values and change scores providing a complementary

XSL•FO
RenderX

perspective. This dual approach allows for a more robust assessment while acknowledging potential limitations in interpretation.

In addition to the bivariate analysis (Mann-Whitney *U* test and Cochran-Mantel-Haenszel test), a regression approach was conducted to assess the program's effect on changing mental health status across different factors of mental health symptoms and reasons for joining the program. An ANCOVA, a statistical method combining regression analysis with ANOVA [5], was performed. In this analysis, the change in the Likert score after the program (ie, postprogram score minus preprogram score) was used as the outcome variable, as widely applied in the literature [7]. The ANCOVA model included the presence or absence of a specific reason for joining the program (coded as yes vs no) and controlled for the baseline score (ie, preprogram score). Because baseline scores for mental health status can differ across groups (Figure 1), adjustments for baseline imbalances were applied following recommendations in the literature [8]. This adjustment minimizes bias in estimating the program's effect by accounting for potential variability in preprogram scores. The ANCOVA was applied to each mental health symptom or reason for joining the program separately.

Figure 1. Boxplots of the 5-point Likert score (0: terrible or debilitating; 5: all good) before and after the Break Method program among participants (A) who chose "depression," (B) "poor self-image/low confidence," or (C) "future goals/next steps" as their reason for joining.

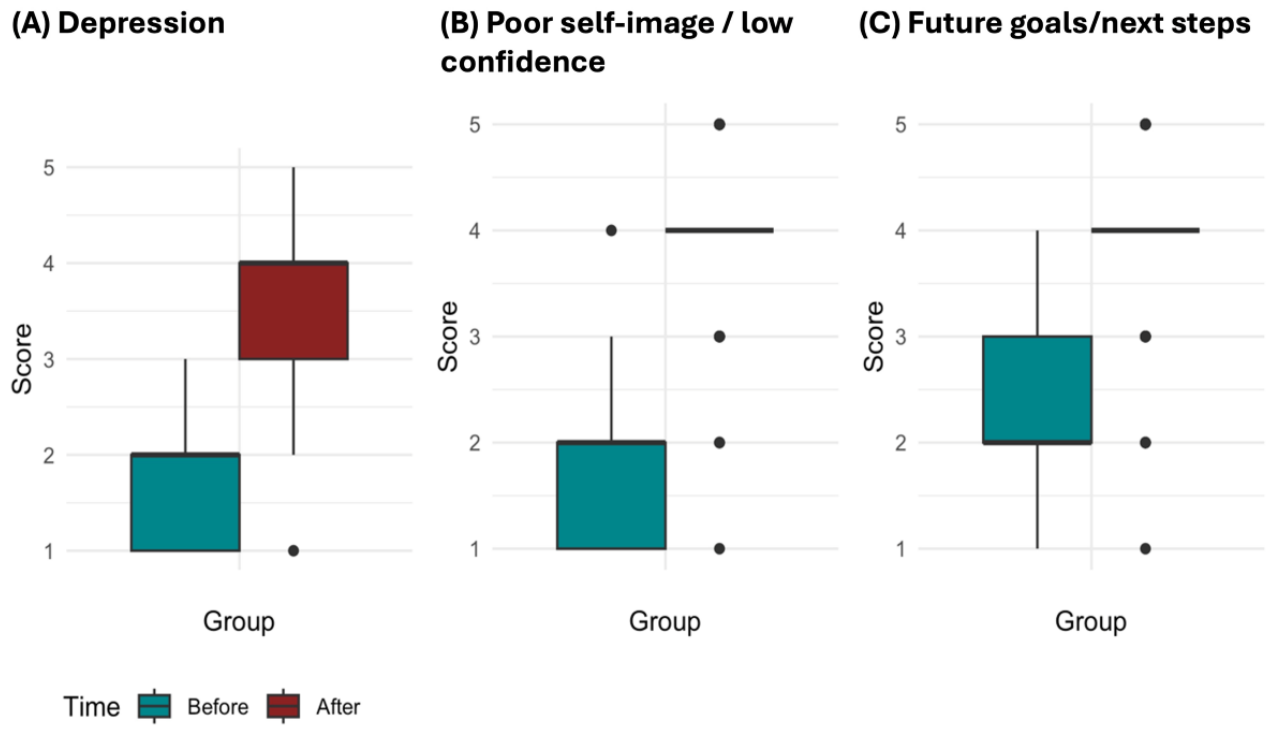

Multiple responses (eg, depression coexisting with anxiety or self-harm) in the data were highly correlated. To uncover hidden dimensions and clustering, multiple correspondence analysis (MCA), an extension of principal component analysis for categorical variables, was conducted [9]. The MCA decomposes a contingency table (Burt matrix) of categorical variables, identifying patterns and relationships as dimensions in a lower-dimensional space [10]. The number of dimensions was determined using a scree plot, where higher eigenvalues represent a larger share of explained variance. A biplot visualized how reasons for joining the program clustered in 2 dimensions, with principal coordinates calculated based on inertia or eigenvalue criteria [10]. In the biplot, the numbers 1 and 0 next to each variable represent binary outcomes (1=yes; 0=no).

The corresponding coordinates of the individuals, along with the identified dimensions, were used in a generalized estimating equation (GEE) model to evaluate the association of each principal dimension with changes in mental health status after the program. The GEE approach is suitable for repeated measures and clustered data, focusing on population-averaged effects rather than individual-specific effects:

Where y_{it} is the change in the 5-point Likert score for mental health status between pre- and postprogram (ie, postprogram score minus preprogram score) of i-th individual measured at time t, Dim_{jit} is the time-varying variable of the j-th dimension obtained from the MCA (ie, the coordinate on the j-th dimension) of the i-th individual at time t, ε_{it} is the random error term, and β_j is the coefficient for the j-th MCA dimension, and Preit is the preprogram baseline Likert score for the mental health status. The model assumed a constant correlation between repeated measures for an individual. The β_0 represents how the principal MCA dimension affects the outcome (ie, changes in mental health status). The GEE model was applied to principal coordinates derived from the MCA, which summarized correlated categorical variables (eg, overlapping symptoms). Importantly, only unique participant responses were included

XSL•FO
RenderX

in the analysis after removing duplicates, so no repeated measures over time were involved. The GEE was therefore used to account for potential correlation within the dimensions identified by the MCA, rather than repeated measurements per individual. To make the GEE approach more accessible, we briefly describe it as a method that accounts for correlations between repeated measurements within participants. The model estimates average changes across the sample while adjusting for within-subject dependencies.

In this study, participants reported a variety of co-occurring mental health symptoms, making it impossible to define mutually exclusive symptom groups. The MCA is a data reduction technique suitable for categorical data to address this issue by identifying latent patterns of co-occurring symptoms. The MCA allows clustering participants based on the similarity of their reported symptoms, creating distinct dimensions. This approach enabled us to examine which participant clusters showed greater changes in mental health outcomes after the program, offering a more interpretable result than symptom-specific comparisons, which can be vulnerable to type 1 error. We then used GEE to examine how the latent dimensions identified through the MCA were associated with temporal changes in mental health status. The GEE provided interpretable statistical estimates, including *P* values and 95% CIs, to assess meaningful differences in mental health changes across participant clusters.

Although the combined MCA+GEE approach is methodologically innovative, the modest sample size (n=175) may limit the stability of the model. This introduces a potential risk of overfitting that should be considered when interpreting the findings.

Multiple statistical tests were performed in this study (eg, paired *t* test, McNemar, Mann-Whitney *U*, ANCOVA, MCA, and GEE), which may increase the risk of type 1 error. A formal correction (eg, Bonferroni or false discovery rate) was not applied, as such methods can be overly conservative in moderate sample sizes like ours and may obscure potentially meaningful exploratory findings. Nonetheless, results should be interpreted with caution given this limitation.

The statistical analyses were conducted in R Software (version 4.2.2; R Foundation for Statistical Computing).

Results

We used data from a total of 175 participants, of whom 15 responded to the survey 2 or more times during the study period, resulting in 195 unique responses. Table 1 illustrates the frequency of responses for joining the Break Method. The most frequently chosen reason among participants was “self-sabotage/stagnation” (149/175, 85.1%), followed by “future goals/next steps” (147/175, 84%) and “general self-discovery” (147/175, 84%). Other commonly selected responses included “anxiety” (116/175, 66.3%), “relationship issues” (120/175, 68.6%), and “poor self-image/low confidence” (105/175, 60%). The least chosen responses were “history of self-harm/suicidal ideation” (19/175, 10.9%) and “OCD” (22/175, 12.6%).

Table 1. Reasons for joining the program. At recruitment and before the program start, participants completed a survey on their reasons for joining. Multiple responses were allowed. Labels reflect the original survey wording (n=175).

Reason	Frequency, n (%)
Anxiety	116 (66.3)
Depression	72 (41.1)
Relationship issues	120 (68.6)
Poor self/image/low confidence	105 (60)
Disordered eating	47 (26.9)
History of self-harm/suicidal ideation	19 (10.9)
OCD[a]/counting/excessive worry with a calculating nature	22 (12.6)
Parenting issues	55 (31.4)
Addiction	28 (16)
Career/workplace issues	67 (38.3)
Self-sabotage/stagnation	149 (85.1)
Future goals/next steps	147 (84)
General self-discovery	147 (84)

[a]OCD: obsessive-compulsive disorder.

Among the 195 responses, 186 (95.4%) respondents reported Likert scores lesser than 4 before the program (Table 2). After the program, 157 (80.5%) responses reported Likert scores of 4 or higher. The paired *t* test revealed significant increases in the 5-point Likert scale after program participation (preprogram mean score 2.07, SD 0.82, postprogram mean score 3.92, SD 0.73; *P*<.001). The McNemar test for the 2×2 table comparing mental health status (eg, ≥4 vs <4 Likert score) before and after the program also indicated a significant difference in the frequency of mental health status (*P*<.001).

XSL•FO
RenderX

Table 2. Frequency (%) of the 5-point Likert score for mental health status before and after the program participation (n=195).

Score	Before the program, n (%)	After the program, n (%)
1	48 (24.6)	1 (0.5)
2	95 (48.7)	8 (4.1)
3	43 (22.1)	29 (14.9)
4	8 (4.1)	125 (64.1)
5	1 (0.5)	32 (16.4)

The average of preprogram and postprogram 5-point Likert scores stratified by each reason for joining the program is shown in Table 3.

XSL•FO
RenderX

Table 3. Preprogram and postprogram Likert score means, stratified by the joining reasons.

Reason	Before the program, mean (SD)	After the program, mean (SD)	*P* value[a]
Anxiety			<.001
With	2.42 (0.68)	3.94 (0.73)	
Without	1.84 (0.87)	3.91 (0.73)	
Depression			.15
With	1.69 (0.66)	3.72 (0.77)	
Without	2.29 (0.82)	4.03 (0.68)	
Relationship issues			.07
With	1.95 (0.85)	3.86 (0.78)	
Without	2.27 (0.74)	4.01 (0.63)	
Poor self-image/low confidence			.01
With	1.84 (0.73)	3.87 (0.71)	
Without	2.34 (0.85)	3.98 (0.75)	
Disordered eating			.14
With	1.79 (0.76)	3.81 (0.73)	
Without	2.16 (0.83)	3.95 (0.72)	
Self-sabotage/stagnation			.03
With	1.93 (0.75)	3.86 (0.71)	
Without	2.54 (0.86)	4.11 (0.74)	
General self-discovery			.55
With	2.11 (0.82)	3.99 (0.72)	
Without	1.96 (0.80)	3.71 (0.73)	
History of self-harm/suicidal ideation			.01
With	1.47 (0.62)	3.89 (0.63)	
Without	2.14 (0.82)	3.92 (0.74)	
OCD[b]/counting/excessive worry with a calculating nature			.01
With	1.55 (0.69)	3.91 (0.96)	
Without	2.14 (0.81)	3.92 (0.70)	
Parenting issues			.25
With	2.00 (0.84)	3.98 (0.59)	
Without	2.10 (0.81)	3.89 (0.78)	
Career/workplace issues			.11
With	1.85 (0.79)	3.84 (0.82)	
Without	2.19 (0.71)	3.96 (0.67)	
Future goals/next steps			.12
With	2.13 (0.80)	3.95 (0.74)	
Without	1.90 (0.87)	3.83 (0.69)	
Addiction			.05
With	1.68 (0.68)	3.89 (0.65)	
Without	2.14 (0.82)	3.92 (0.74)	

[a]The *P* value is based on the Mann-Whitney *U* test comparing the changes in the preprogram and postprogram scores between the groups with and without the specified reason for joining.

[b]OCD: obsessive-compulsive disorder.

XSL•FO
RenderX

Using the ANCOVA, we compared mean changes in Likert scores (postprogram score minus preprogram score) across groups stratified by the presence of each reason for joining the program (Table 4). The β coefficient, indicating the score changes after the program, was highest for "history of self-harm/suicidal ideation" (β=0.511, 95% CI 0.091-0.931), followed by "OCD/counting/excessive worry with a calculating nature" (β=0.455, 95% CI 0.061-0.850). Positive and significant changes in scores were also associated with "anxiety" (β=0.332, 95% CI 0.073-0.591). The smallest absolute score change was observed for "parenting issues" (β=0.132, 95% CI 0.020-0.244).

Table 4. Results of the analysis of covariancea.

Variable	β (95% CI)	*P* value
Anxiety	0.332 (0.073 to 0.591)	.01
Depression	0.107 (−0.158 to 0.371)	.43
Poor self-image/low confidence	0.203 (−0.053 to 0.459)	.12
Disordered eating	0.149 (−0.146 to 0.444)	.32
Self-sabotage/stagnation	0.020 (−0.290 to 0.329)	.90
General self-discovery	0.213 (−0.079 to 0.505)	.15
History of Self-harm/suicidal ideation	0.511 (0.091 to 0.931)	.02
OCD[b]/counting/excessive worry with a calculating nature	0.455 (0.061 to 0.850)	.03
Parenting issues	0.155 (−0.125 to 0.434)	.28
Career/workplace issues	0.093 (−0.174 to 0.360)	.50
Future goals/next steps	−0.106 (−0.399 to 0.186)	.48
Relationship issues	0.132 (−0.126 to 0.391)	.32
Addiction	0.297 (−0.062 to 0.655)	.11

[a]The model was adjusted for the baseline Likert score differences.

[b]OCD: obsessive-compulsive disorder.

The estimated eigenvalues indicated how much variance each dimension explained in the MCA. Dimension 1 accounted for 22.7% of the data variance, while dimensions 2 and 3 accounted for 11.9% and 10.0%, respectively. The biplot of the MCA (Figure 2) provided a graphical representation of relationships between the different categories of reasons for joining the Break Method in the reduced multidimensional space. By examining dimensions 1 and 2, we observed a cluster of participants exhibiting shared characteristics, including depression, low confidence, relationship issues, and self-sabotage or stagnation. Another distinct cluster comprised participants focused on future goals, self-discovery, and career or workplace challenges. Finally, a third category appeared to consist of participants displaying OCD-related symptoms.

XSL•FO
RenderX

Figure 2. Biplot of the multiple correspondence analysis. OCD: obsessive-compulsive disorder.

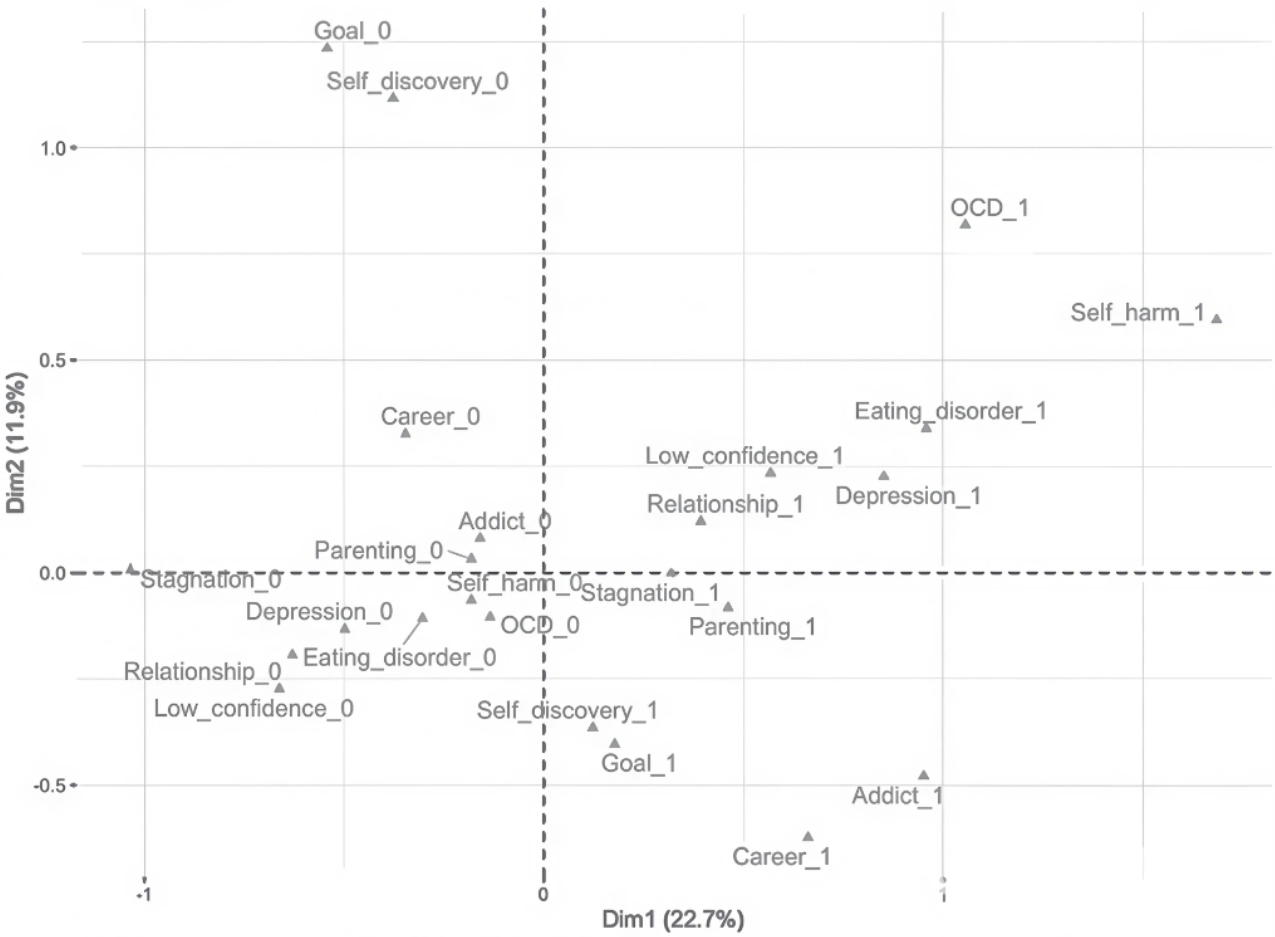

The coefficients for the MCA latent dimensions were examined in relation to changes in the 5-point Likert scores for symptoms upon completing the program. Dimension 1 (ie, "low self-image") was significantly and positively associated with higher increases in Likert scores (β=0.348, 95% CI 0.060-0.636), indicating significant effects among participants who joined the program due to depression, low confidence, relationship issues, or self-sabotage or stagnation. Dimension 2 (eg, self-discovery and future goals) indicated a positive but nonsignificant effect (β=0.168, 95% CI −0.205 to 0.540). Dimension 3 (ie, "OCD") was not significantly associated with changes in Likert scores (β=−0.031, 95% CI −0.423 to 0.362).

Discussion

Principal Findings

This study recruited participants for the Break Method and assessed its impact on their mental health, focusing on the participants' characteristics as determined by their reasons for joining the program. The results indicated that several factors, including anxiety, OCD, and a history of self-harm or suicidal ideation, were significantly associated with positive changes in Likert scores for mental health status after program participation.

Our analyses of temporal changes in Likert scale mental health status are subject to potential type I error; that is, observed changes among participants with a specific symptom may not solely reflect the intervention's effect on that symptom, but could also be influenced by co-occurring symptoms. Although this limitation affects the precision of symptom-specific effects, our data collection captured all self-reported symptoms, making it infeasible to define mutually exclusive case and control groups. Despite this limitation, the findings offer valuable insight into temporal patterns of mental health change in individuals reporting specific symptoms. The primary aim of this pilot study was to identify overall trends in mental health status over time. Future studies should use larger samples and validated symptom-specific measures to improve specificity and causal inference.

Our analysis revealed 2 to 3 distinct clusters for participants' reasons for joining the Break Method. Participants who joined the program for reasons related to depression, low confidence, relationship issues, and self-sabotage or stagnation tended to form a cluster, suggesting a shared pattern of challenges and motivations. Another distinct cluster consisted of participants seeking support for future goals, self-discovery, or career or workplace challenges, highlighting a different set of motivations. Finally, participants associated with OCD-related issues appeared as a third, less interconnected cluster, emphasizing the variability in outcomes for this group.

The findings from our MCA may indicate specific benefits of the Break Method program in medical relevance for psychological symptoms and behavioral patterns. The first MCA dimension clustered symptoms that suggest that individuals in

XSL•FO
RenderX

this group may benefit most from emotion-focused reconditioning interventions. The second dimension aligned well with motivational or career coaching approaches. The third dimension may be best addressed through brain pattern mapping and structured self-inquiry exercises. However, because participants were not grouped exclusively by any one dimension or symptom, this study could not assess the program's effectiveness for each specific cluster, warranting more rigorous future research. This study with a single-group, preprogram-postprogram design lacked a control group, limiting causal inference on intervention effectiveness. However, a randomized controlled trial was not ethically or practically feasible. Designed as a pilot, this study aimed to assess feasibility and potential changes in mental health outcomes rather than establish causality. Despite this limitation, within-subject comparisons offer valuable insight into temporal trends and possible intervention effects. Future research should include appropriate control groups to validate effectiveness. All participants who enrolled completed the program and exit surveys. Therefore, no dropout data were available. This should be considered when interpreting the results.

The Break Method's use of coaching and digital modules for emotional reconditioning aligns with recent studies [11,12] that support improved delivery of care through technology-based mental health programs that combine digital tools with coaching approaches. Despite its theoretical basis, a key limitation of this study is the lack of published randomized controlled trials evaluating the efficacy. To advance the evidence base, future research should incorporate rigorous designs, such as randomized trials or longitudinal studies, using validated clinical and neuropsychological outcome measures. Evaluating the Break program as both a standalone treatment and an adjunct to existing therapies could also clarify its role in improving mental health care. As engagement is a critical determinant of success in digital interventions [13], understanding user motivation within Break Method remains essential.

After participating in the Break Method program, participants reported increased scores on a 5-point Likert scale, indicating improved mental health status. Notably, greater improvements were observed among those who joined due to depression, low confidence, relationship issues, or self-sabotage and stagnation compared with others. However, the interpretation of these results is limited by the heterogeneity of reported symptoms, as participants experienced a range of conditions including depression and OCD, making it difficult to attribute improvements to specific symptom domains. Although some participants reported reductions in OCD-related symptoms, the Break Method is not a disorder-specific intervention and was not designed to directly treat OCD. These reported changes may instead reflect broader improvements in anxiety and maladaptive thought-behavior patterns. Practically, this highlights potential utility as an adjunctive tool for clients with mixed symptom presentations, while established disorder-specific treatments remain essential. Nonetheless, several limitations must be considered when interpreting these findings.

Despite its promising findings, this study has limitations that should be acknowledged. One key limitation is that while demographic data (eg, age and gender) were collected, they were not analyzed in relation to program outcomes. This omission restricts the ability to examine potential correlations between demographic factors and the program's effectiveness. Although demographic data were collected, analyses focused on overall symptom change to maintain clarity and statistical power, as the sample size limited subgroup comparisons. This growing field lacks relevant studies, limiting comparison. Our study adds to emerging evidence and underscores the need for further research. Additionally, the study used a 5-point Likert scale to assess the mental health status, rather than using validated clinical tools such as the Patient Health Questionnaire-9 [14]. The use of a nonstandardized measure introduces potential variability in the assessment of mental health improvements. This restricts the interpretability and comparability of outcomes with other studies and may introduce measurement variability. Therefore, key outcomes, such as depression and anxiety, may not have been fully captured. The tool was also subject to recall bias, social desirability bias, and measurement error due to variability in how participants interpreted the scale. Additionally, self-reported symptom "resolution" (score=5) is subjective and may reflect participants' expectations or social desirability bias rather than true clinical improvement, further limiting interpretation of the findings. However, the use of a full-length, validated instrument was not feasible given the study's design and its primary aim to evaluate overall changes in mental health following the intervention. Future studies should incorporate validated measures to more accurately assess mental health outcomes and intervention effectiveness.

Another limitation is that all participants were paying clients of the Break Method, which may have influenced motivation or expectations and could have affected self-reported outcomes. This factor may modestly limit generalizability to broader populations. Furthermore, reliance on self-reported data may have contributed to response bias, as participants' perceptions, expectations, or tendencies toward social desirability could have influenced their reported improvements. Additionally, our analysis used the MCA to group participants based on similarities in mental health symptoms, as exclusively compatible groups for each symptom could not be clearly defined. Given the exploratory nature of the MCA and the limited sample size, including additional covariates risked overfitting and reduced interpretability. Future studies with larger samples and appropriate control groups should investigate the influence of these demographic factors on intervention effectiveness. The study excluded individuals reporting active suicidal ideation or requiring crisis-oriented care, so the findings may not be fully generalized to the high-risk population with active suicidal ideation. However, the results remain informative for participants without acute risk. Finally, the absence of a long-form questionnaire limited the ability to capture more nuanced insights into participants' mental health status before and after the program. Fully automated digital interventions have demonstrated measurable improvements in mental well-being across diverse populations [15].

In our results of the ANCOVA for each mental health symptom, we did not find significant score changes after the program among those who reported depression or parenting issues. It is

XSL•FO
RenderX

important to note that these findings may be vulnerable to type I error. Even if the program had a positive effect on depression, the overall change may have been obscured by the presence of co-occurring symptoms that were not directly addressed by the intervention. Otherwise, depression is more chronic and recurrent, and there is evidence suggesting that depression often requires a longer intervention duration compared with other symptoms such as anxiety, due to its clinical course and risk of relapse [16]. Additionally, parenting issues are complex and usually require longer-term tailored interventions [17]. This evidence suggests that a longitudinal study with a long-term intervention would be needed to observe meaningful improvements in these symptoms, warranting future studies. Consistent with findings from self-guided digital mental health programs [18], participants in our study reported significant postprogram improvements. Scalable digital mental health programs have shown promise in expanding access to care, particularly in low-resource contexts [19]. Future work could also explore age-specific adaptations of self-guided digital interventions, as shown in adolescent populations [20].

Conclusions

This study suggests preliminary associations between participation in the Break Method and improvements in self-reported mental health status, particularly among individuals struggling with anxiety, low confidence, and self-sabotage. These findings highlight the importance of structured, behavior-based interventions in addressing emotional dysregulation and fostering initial mental health improvements. By identifying key participant clusters, this research suggests that personalized intervention strategies could enhance program effectiveness and inform future refinements of the Break Method.

Although the study demonstrates improvements in self-reported outcomes, the single-group preprogram-postprogram design without a control group prevents causal conclusions. The observed changes may be influenced by confounding factors or natural symptom fluctuation, highlighting the need for controlled trials to rigorously assess the Break Method's effectiveness.

This study underscores the need for further research to explore the long-term efficacy of the Break Method, its applicability across diverse populations, and its potential integration into broader mental health care systems. These findings serve as a foundation for developing more tailored, scalable interventions that can be adapted across different mental health treatment settings. Future studies should ideally include independent evaluations conducted by researchers not affiliated with the Break Method to strengthen the objectivity and credibility of the findings.

Acknowledgments

This research received no specific grant or funding from any public, commercial, or not-for-profit funding agency.

Data Availability

The datasets generated or analyzed during this study are not publicly available to maintain participant confidentiality and to remain in compliance with ethical standards governing the use of sensitive human data. Deidentified data may be made available from the corresponding author upon reasonable request and with appropriate ethical approval.

Authors' Contributions

BS was responsible for conceptualization, data curation, and review and editing of the manuscript. BG contributed to the methodology, validation, and original draft writing. KK managed project administration. All content was prepared and reviewed manually by the authors.

Conflicts of Interest

The authors have a financial interest in the Break Method. Although research activities were conducted independently, this relationship is disclosed for transparency. This should be considered when interpreting the findings.

References

1. Xu J, He K, Zhang K, Yang C, Nie L, Dan D, et al. Low-dose copper exposure exacerbates depression-like behavior in ApoE4 transgenic mice. Oxid Med Cell Longev 2021 Mar 26;2021(1):6634181 [FREE Full text] [doi: 10.1155/2021/6634181] [Medline: 33833851]
2. Hazra A, Gogtay N. Biostatistics series module 4: comparing groups - categorical variables. Indian J Dermatol 2016;61(4):385-392 [FREE Full text] [doi: 10.4103/0019-5154.185700] [Medline: 27512183]
3. Senn S. Change from baseline and analysis of covariance revisited. Stat Med 2006 Dec 30;25(24):4334-4344. [doi: 10.1002/sim.2682] [Medline: 16921578]
4. Kosyluk K, Baeder T, Greene KY, Tran JT, Bolton C, Loecher N, et al. Mental distress, label avoidance, and use of a mental health chatbot: results from a US survey. JMIR Form Res 2024 Apr 12;8:e45959 [FREE Full text] [doi: 10.2196/45959] [Medline: 38607665]
5. Khammar A, Yarahmadi M, Madadizadeh F. What is analysis of covariance (ANCOVA) and how to correctly report its results in medical research? Iran J Public Health 2020 May;49(5):1016-1017 [FREE Full text] [Medline: 32953697]

XSL•FO
RenderX

6. Behavioral risk factor surveillance system questionnaire. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention. URL: https://www.cdc.gov/brfss/questionnaires/pdf-ques/2021-BRFSS-Questionnaire-508.pdf [accessed 2025-05-29]
7. Jamieson J. Analysis of covariance (ANCOVA) with difference scores. Int J Psychophysiol 2004 May;52(3):277-283. [doi: 10.1016/j.ijpsycho.2003.12.009] [Medline: 15094250]
8. Clifton L, Clifton DA. The correlation between baseline score and post-intervention score, and its implications for statistical analysis. Trials 2019 Jan 11;20(1):43 [FREE Full text] [doi: 10.1186/s13063-018-3108-3] [Medline: 30635021]
9. Costa PS, Santos NC, Cunha P, Cotter J, Sousa N. The use of multiple correspondence analysis to explore associations between categories of qualitative variables in healthy ageing. J Aging Res 2013;2013:302163-302112 [FREE Full text] [doi: 10.1155/2013/302163] [Medline: 24222852]
10. Olugbara CT, Letseka M, Olugbara OO. Multiple correspondence analysis of factors influencing student acceptance of massive open online courses. Sustainability 2021 Dec 05;13(23):13451. [doi: 10.3390/su132313451]
11. Fitzpatrick KK, Darcy A, Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. JMIR Ment Health 2017 Jun 06;4(2):e19 [FREE Full text] [doi: 10.2196/mental.7785] [Medline: 28588005]
12. Torous J, Jän Myrick K, Rauseo-Ricupero N, Firth J. Digital mental health and COVID-19: using technology today to accelerate the curve on access and quality tomorrow. JMIR Ment Health 2020 Mar 26;7(3):e18848 [FREE Full text] [doi: 10.2196/18848] [Medline: 32213476]
13. Borghouts J, Eikey E, Mark G, De Leon C, Schueller SM, Schneider M, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. J Med Internet Res 2021 Mar 24;23(3):e24387 [FREE Full text] [doi: 10.2196/24387] [Medline: 33759801]
14. Beard C, Hsu KJ, Rifkin LS, Busch AB, Björgvinsson T. Validation of the PHQ-9 in a psychiatric sample. J Affect Disord 2016 Mar 15;193:267-273. [doi: 10.1016/j.jad.2015.12.075] [Medline: 26774513]
15. Groot J, MacLellan A, Butler M, Todor E, Zulfiqar M, Thackrah T, et al. The effectiveness of fully automated digital interventions in promoting mental well-being in the general population: systematic review and meta-analysis. JMIR Ment Health 2023 Oct 19;10:e44658 [FREE Full text] [doi: 10.2196/44658] [Medline: 37856172]
16. Practice guideline for the treatment of patients with major depressive disorder. 3rd edition. American Psychiatric Association. 2010. URL: https://psychiatryonline.org/pb/assets/raw/sitewide/practice_guidelines/guidelines/mdd.pdf [accessed 2025-05-29]
17. Sanders MR, Kirby JN. A public‐health approach to improving parenting and promoting children's well‐being. Child Dev Perspectives 2014 Sep;8(4):250-257. [doi: 10.1111/cdep.12086]
18. Nosè M, Muriago G, Turrini G, Tedeschi F, Forlani O, Sartori R, et al. Effectiveness of a self-guided digital intervention for mental health and psychological well-being in university students: pre- and postintervention study. J Med Internet Res 2025 Aug 19;27:e69031 [FREE Full text] [doi: 10.2196/69031] [Medline: 40829124]
19. Chau LW, Lam RW, Minas H, Hayashi K, Nguyen VC, O'Neil J. Digital health interventions for depression and anxiety in low- and middle-income countries: rapid scoping review. JMIR Ment Health 2025 Aug 22;12:e68296 [FREE Full text] [doi: 10.2196/68296] [Medline: 40846319]
20. Miller I, Peake E, Strauss G, Vierra E, Koepsell X, Shalchi B, et al. Self-guided digital intervention for depression in adolescents: feasibility and preliminary efficacy study. JMIR Form Res 2023 Nov 22;7:e43260 [FREE Full text] [doi: 10.2196/43260] [Medline: 37991839]

Abbreviations

ANCOVA: analysis of covariance
CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
GEE: generalized estimating equation
IRB: institutional review board
MCA: multiple correspondence analysis
OCD: obsessive-compulsive disorder

XSL•FO
RenderX

Edited by J Sarvestan; submitted 27.03.25; peer-reviewed by J Abua, L Mercado Diaz; comments to author 09.05.25; revised version received 10.07.25; accepted 14.10.25; published 07.11.25.

Please cite as:
Salik B, Gold B, Krier K
Reconditioning Emotional Responses With the Break Method: Pilot Quantitative Study
JMIR Form Res 2025;9:e75056
URL: https://formative.jmir.org/2025/1/e75056
doi:10.2196/75056
PMID:

©Boaz Salik, Bizzie Gold, Kira Krier. Originally published in JMIR Formative Research (https://formative.jmir.org), 07.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Knowledge, Attitudes, and Practices Toward Burn Causes and First Aid Management in Jazan Region, Saudi Arabia: Cross-Sectional Study

Anas Sayegh[1]; Alyazid Y Awaji[2]; Anas Fathuldeen[3]; Atheer M Alshammakhi[4]; Roaa M Alhazmi[4]; Razan F Ageeli[4]; Ghadi A Shamakhi[4]

[1]Department of Surgery, Faculty of Medicine, Jazan University, Jazan, Saudi Arabia
[2]Department of Plastic and Reconstructive Surgery, King Fahad Central Hospital, Jazan, Saudi Arabia
[3]Department of Plastic Surgery, College of Medicine, University of Ha'il, Ha'il, Saudi Arabia
[4]Faculty of Medicine, Jazan University, Jazan, Saudi Arabia

Corresponding Author:
Atheer M Alshammakhi
Faculty of Medicine
Jazan University
Jazan, 87388
Saudi Arabia
Phone: 966 01772518365
Email: AtheerAlshammakhi@gmail.com

Abstract

Background: Burn injuries are a major global health problem, particularly in low- and middle-income countries, accounting for most burn-related deaths and disabilities. In Saudi Arabia, burns remain a frequent cause of morbidity, often resulting from domestic accidents involving hot liquids, open flames, or electricity. The Jazan region, with its dense population and reliance on traditional cooking methods, is considered at higher risk. Understanding public knowledge, attitudes, and practices regarding burns and first aid is essential for guiding preventive strategies and health education efforts.

Objective: This study aimed to assess the level of awareness, knowledge, and practices related to burn causes and first aid management among residents of the Jazan region, Saudi Arabia.

Methods: A cross-sectional online survey was conducted using a validated Arabic questionnaire distributed through social media platforms. The questionnaire covered demographics, knowledge of burn causes and types, first aid management, and preventive practices. Data were analyzed using SPSS version 26.0.

Results: Out of 404 participants aged 18-60 years, 228 (56.4%) demonstrated poor knowledge and awareness regarding burn causes and first aid management. The internet was the most commonly reported source of information (171/236, 72.5%), followed by formal courses (76/236, 32.2%), paper leaflets (75/236, 31.8%), television (67/236, 28.4%), and daily newspapers (23/236, 9.7%). Additionally, 215 (53.2%) participants had previously experienced burns, with hot water (136/215, 63.8%) and fire (105/215, 48.8%) cited as the most common causes. The most frequently affected sites were the hand (111/215, 51.6%), arm (85/215, 39.5%), and thigh (49/215, 22.8%).

Conclusions: The study highlights limited awareness and improper first aid practices among the Jazan population. Targeted community-based educational programs are needed to enhance burn prevention and management knowledge.

(JMIR Form Res 2025;9:e80594) doi:10.2196/80594

KEYWORDS

burns; first aid; knowledge; practice; population; Saudi Arabia

Introduction

Burn injuries are a major global cause of morbidity and mortality, posing a significant public health challenge [1]. The World Health Organization states that burn injuries cause about 195,000 fatalities and disability annually, making them a major public health concern. Among individuals aged 5-29 years, fire is the 15th most common cause of death [2]. It is concerning

XSL•FO
RenderX

that over 95% of fire-related burn deaths occur in low- and middle-income countries (LMICs). People aged 70 years and older as well as children younger than 5 years have the highest mortality rates [3]. In addition to the high mortality rates, millions of people live with permanent disability and deformities, often facing social rejection and stigma [4].

Evaluating the reasons for burns and assessing the efficiency of first-aid management practices are crucial steps toward addressing and understanding this issue and developing targeted strategies [5]. Additionally, establishing successful preventive plans requires an awareness of the causes of burns [6]. Burns can result from exposure to radiation, heat, electricity, chemicals, and other causes [7,8]. Determining the exact causes and associated risk factors is necessary to set priorities for preventative actions and to allocate resources effectively. It is imperative to investigate the specific factors in Jazan, Saudi Arabia, that cause burn injuries. This means considering local customs, workplace hazards, and any environmental factors specific to the area [9,10]. In addition to identifying burn causes, assessing and analyzing the circumstances in which burn accidents occur can also provide valuable insights for designing safety interventions. For example, understanding the specific situations where burns most frequently happen, whether in households, workplaces, or recreational settings, can help develop targeted prevention strategies. By addressing these specific contexts, interventions can be developed to improve safety practices and reduce the likelihood of burn incidents [5,11].

First aid management plays a critical role in minimizing the severity and complications of burn injuries. Prompt and proper first aid measures can significantly improve outcomes and reduce the long-term impact of burns. However, the effectiveness of first aid management depends on the knowledge and skills of the individuals providing the first care. Therefore, assessing the adequacy of first aid practices is crucial in identifying areas of improvement and implementing targeted educational interventions [12,13]. Research conducted in other regions has offered valuable insights into burn causes and first aid management practices [14-17]. However, it is important to recognize that the factors contributing to burn injuries and the strategies used for first aid management may vary across different geographical locations. Therefore, a focused assessment of burn causes and first aid management practices in Jazan, Saudi Arabia, is necessary to understand the unique challenges and opportunities for intervention in this specific region. This research aimed to conduct a comprehensive assessment of burn causes and first aid management practices in Jazan, Saudi Arabia. This study analyzed the effects of demographics, burn severity, and location in Jazan, Saudi Arabia, aiming to reveal the trends and address knowledge gaps for health care improvement. Our results will help professionals and policymakers, aiding in reducing burn-related morbidity and mortality rates with associated social and economic burdens.

Methods

Study Design and Setting

This community-based cross-sectional study was conducted in the Jazan region, located in the southwestern part of Saudi Arabia. Jazan is characterized by varied geographical and socioeconomic conditions, comprising urban, rural, and coastal communities. This study targeted adult residents (aged 18 years and older) living in different governorates of the region. Participants were invited voluntarily to complete the survey, and eligibility was confirmed by screening for age (18 years and older) and residence within the Jazan region.

Study Population

Eligible participants for inclusion in the survey were individuals who were 18 years or older and voluntarily agreed to participate, residing in the Jazan region. Those younger than 18 years and not Jazan residents were excluded.

Sample Size

The Raosoft sample size calculator was needed to determine the sample size for this study, incorporating a 5% margin of error, a 95% CI, a 50% response distribution, and population size data of 1,404,997 sourced from the General Authority for Statistics. A total of 385 participants will be needed to reach a 95% CI and a 5% margin of error. The determination of the sample size aimed to ensure the generalizability and relevance of the findings. A nonprobability convenience sampling technique was adopted to facilitate data collection.

Data Collection Method

Data were collected electronically via Google Forms using a validated and structured questionnaire. The instrument was developed in English, then translated into Arabic, and administered in Arabic (see Multimedia Appendix 1). The questionnaire was specifically developed to assess the knowledge, attitudes, and practices toward burn causes and first aid management among residents of the Jazan region. It consisted of 5 major sections: (1) demographic information (age, sex, education, occupation, income, and residence), (2) previous exposure to burn prevention education or training, (3) knowledge of first aid procedures for burns of different degrees, (4) practical behaviors and responses during burn incidents, and (5) attitudes toward prevention, home safety tools, and awareness methods.

We defined burn first-aid practices and burn depth according to World Health Organization (WHO) criteria [18]. The questionnaire was developed by the research team following an extensive literature review and consultation with 3 public health experts who evaluated the tool for clarity, cultural appropriateness, and content validity. Based on their feedback, necessary modifications were made, and a pilot study involving 20 participants was conducted to assess reliability, yielding a Cronbach α coefficient of 0.75.

Recruitment

Participants were recruited through a nonprobability convenience sampling approach. The survey link was distributed

XSL•FO
RenderX

via major social media platforms (WhatsApp, X, Facebook, Telegram, and Snapchat) and community groups across the Jazan region. Participation was entirely voluntary and anonymous, with no incentives provided. Before accessing the questionnaire, a brief introduction was displayed, and participants reviewed an electronic consent statement and confirmed their willingness to participate by selecting "Yes" to a consent question; only those who consented could proceed.

Data collection continued until the required sample size was reached, or no new responses were received for 48 hours. The final dataset included only participants who met the predefined inclusion criteria, ensuring representativeness of the adult population in Jazan.

Statistical Analysis

The data were collected, reviewed, and then entered into SPSS version 26 (released 2019; IBM Corp). All statistical methods used were 2-tailed with an α level of .05, and results were considered significant if $P \leq .05$. An overall knowledge score was computed by summing the correct answers, where a correct answer was given a 1-point score and 0 was given otherwise. Participants with a knowledge score less than 60% of the total correct answers were classified as having poor knowledge level, while others with a knowledge score of 60%-100% were classified as having good knowledge level.

Descriptive analysis for categorical data was done using frequencies and percentages, whereas numerical data were presented as mean with standard deviation. Moreover, participants' knowledge and awareness about burn causes and first aid management, practice, history of burn, and post-burn medical consultation, management, complications, and daily effects were tabulated, while the overall knowledge level, source of information, preventive measures, and the available traditional treatments were graphed. We conducted cross-tabulations to examine factors associated with participants' knowledge by using the Pearson chi-square test, and exact tests were applied for small cell counts.

Ethical Considerations

This study followed the Declaration of Helsinki and Saudi national ethics guidelines, with approval from Jazan University's Research Ethics Committee (REC-45/09/1030). It involved minimal risk human research. Participants aged 18 years or older from the Jazan region were recruited via an electronic self-administered questionnaire after providing electronic informed consent. Data were collected anonymously, with no personal identifiers, and stored securely on password-protected servers accessible only to the research team. Participation was voluntary, with no compensation provided, and no identifiable data or images were collected or published.

Results

A total of 404 eligible participants completed the study questionnaire. The mean age was 26.8 (SD 11.9) years (range 18 to >60 years). Among them, 67.3% (272/404) were females, 72.8% (294/404) held a bachelor's degree, and 96.8% (391/404) were Saudi nationals. Regarding employment status, 46.8% (189/404) were students and 30.7% (124/404) were employed. Monthly income was less than US $1332.50 for 28.2% (114/404) of the participants, between US $1332.50 and US $2665 for 29.5% (119/404) of the participants, and between US $2665 and US $5330.10 for 31.7% (128/404) of the participants. Concerning housing type, 41.8% (169/404) of the participants lived in apartments, 39.1% (158/404) in villas, and 19.1% (77/404) in traditional houses. Most participants (321/404, 79.5%) reported having children or teenagers in their homes. In terms of social media use, 65.6% (265/404) used Twitter, with 47.8% (193/404) of them for less than 2 hours daily. Facebook was used by 18.3% (74/404) of the participants, 15.6% (63/404) of whom used it for less than 2 hours daily. YouTube was used by 87.6% (354/404) of the participants, 56.2% (227/404) of whom reported daily use of less than 2 hours (Table 1).

XSL•FO
RenderX

Table 1. Sociodemographic characteristics and social media use among study participants in Jazan, Saudi Arabia (N=404).

Category	Values, n (%)
Age (y)	
18-30	279 (69.1)
31-45	68 (16.8)
46-60	52 (12.9)
>60	5 (1.2)
Sex	
Male	132 (32.7)
Female	272 (67.3)
Educational level	
Secondary/below	54 (13.4)
Diploma/institute	30 (7.4)
Bachelor's degree	294 (72.8)
Postgraduate	26 (6.4)
Nationality	
Non-Saudi	13 (3.2)
Saudi	391 (96.8)
Employment	
Not working	91 (22.5)
Student	189 (46.8)
Working	124 (30.7)
Monthly income	
<US $1332.50	114 (28.2)
US $1332.5 - US $2665	119 (29.5)
US $2665 - US $5330.10	128 (31.7)
>US $5330.10	43 (10.6)
Type of accommodation	
Popular house	77 (19.1)
Apartment	169 (41.8)
Villa	158 (39.1)
Children or teenagers in home	
Yes	321 (79.5)
No	83 (20.5)
Daily Twitter use (h)	
Never use	135 (33.4)
<2	193 (47.8)
2-4	52 (12.9)
>4	24 (5.9)
Daily Facebook use (h)	
Never use	330 (81.7)
<2	63 (15.6)
2-4	5 (1.2)
>4	6 (1.5)

XSL•FO
RenderX

Category	Values, n (%)
Daily YouTube use (h)	
Never use	50 (12.4)
<2	227 (56.2)
2-4	86 (21.3)
>4	41 (10.1)

Table 2 shows the participants' knowledge (N=404) and awareness regarding burn causes and first aid management in the Jazan region. More than half of the participants (236/404, 58.4%) had previously received information about burn prevention. For first-degree burns, 206 (51%) correctly identified washing the affected area with cold water as the appropriate first-aid measure, whereas 99 (24.5%) believed that applying ointments or creams was suitable. For second-degree burns, 154 (38.1%) correctly selected visiting the nearest health facility, while 152 (37.6%) preferred using ointments. Knowledge improved with burn severity, as 295 (73%) participants recognized that third-degree burns require urgent hospital care. A smaller proportion selected incorrect options, including surgical intervention (19/404, 4.7%) and water application (9/404, 2.2%).

For fourth-degree burns, 203 (50.2%) correctly selected emergency treatment, 113 (28%) preferred surgical intervention, and 71 (17.6%) were uncertain. Regarding specific first-aid practices (see Multimedia Appendix 2), 226 (55.9%) agreed that burns should be cooled under running water, and 220 (54.5%) disagreed with applying ice. In addition, 213 (52.7%) agreed that sterile gauze should be used to cover the affected area, and 284 (70.3%) recognized the importance of keeping the area ventilated. Awareness of conditions requiring urgent medical care was high: 276 (68.3%) for burns on joints, 303 (75%) for patients younger than 4 years or older than 70 years, and 350 (86.6%) for chemical or electrical burns. Attitudes toward reconstructive management were positive. A large proportion (353/404, 87.4%) acknowledged the role of plastic surgeons, and 260 (64.4%) participants considered plastic surgery safe and effective for restoring function and appearance.

Overall, 228 (56.4%) participants had poor knowledge and 176 (43.6%) had good knowledge. The internet was the most common source of information (171/236, 72.5%), followed by formal courses (76/236, 32.2%), printed leaflets (75/236, 31.8%), television (67/236, 28.4%), newspapers (23/236, 9.7%), and radio (14/236, 5.9%) (Table 2).

Table 3 summarizes the history of burn incidents and reported first aid practices among participants in the Jazan region (N=404). A total of 215 (53.2%) participants reported that they or a close contact had experienced a burn injury. Among these, hot water (136/215, 63.8%) was the most frequent cause, followed by fire (105/215, 48.8%), chemical exposure (13/215, 6.1%), electrocution (4/215, 1.9%), and sunburn (2/215, 0.9%). The most commonly affected sites were the hands (111/215, 51.6%) and arms (85/215, 39.5%), followed by the thighs (49/215, 22.8%), chest (43/215, 20.3%), feet (35/215, 16.3%), back or buttocks (29/215, 13.5%), head (12/215, 5.6%), abdomen (3/215, 1.4%), and face (2/215, 0.9%). Regarding burn severity (n=215), 89 (41.4%) had first-degree burns, 70 (32.6%) had second-degree burns, 26 (12.1%) had third-degree burns, and 30 (14%) had fourth-degree burns. After the burn incident, 171 (79.5%) removed clothing and accessories from the affected area, and 129 (60%) sought medical assistance; 93 (43.3%) participants covered the injured site with a clean cloth, and 144 (66.9%) poured water on the burn, 114 (79.2%) of whom used cold water. Among those who used cold water (n=144), 62 (43.1%) applied it for 1-5 minutes, 39 (27.1%) for <1 minute, 21 (14.6%) for 5-10 minutes, 13 (9%) for 10-15 minutes, 3 (2.1%) for 15-20 minutes, and 6 (4.2%) for >20 minutes (Table 3).

XSL•FO
RenderX

Table 2. Knowledge and awareness of burn causes and first aid management in Jazan region (N=404).

Items	Values, n (%)
Received information about burn prevention methods	
Yes	236 (58.4)
No	168 (41.6)
First aid for first-degree burns (superficial)	
Wash with cool water	206 (51)
Apply ointments/creams	99 (24.5)
Go to emergency unit	52 (12.9)
I do not know	47 (11.6)
First aid for second-degree burns (partial thickness)	
Go to emergency unit	154 (38.1)
Apply ointments/creams	152 (37.6)
Wash with cool water	38 (9.4)
I do not know	60 (14.9)
First aid for third-degree burns (deep partial thickness)	
Go to emergency unit	295 (73)
Apply ointments/creams	20 (5)
Surgical cosmetic intervention	19 (4.7)
Wash with cool water	9 (2.2)
I do not know	61 (15.1)
First aid for fourth-degree burns (full thickness)	
Go to emergency unit	203 (50.2)
Surgical cosmetic intervention	113 (28)
Wash with cool water	12 (3)
Apply ointments/creams	5 (1.2)
I do not know	71 (17.6)
Overall knowledge level	
Poor	228 (56.4)
Good	176 (43.6)
Source of information (n=236)	
Internet	171 (72.5)
Formal course	76 (32.2)
Paper leaflets	75 (31.8)
Television	67 (28.4)
Daily newspapers	23 (9.7)
Radio	14 (5.9)

XSL•FO
RenderX

Table 3. History of burns and reported practice toward burn causes and first aid management in Jazan region.

Variable	Values, n (%)
Have you or someone close to you been burned before? (N=404)	
Yes	215 (53.2)
No	189 (46.8)
Cause of burn (n=215)	
Hot water	136 (63.2)
Fire	105 (48.8)
Chemical materials	13 (6)
Electrocution	4 (1.9)
Sunburns	2 (0.9)
Site of burn (n=215)	
Hand	111 (51.6)
Arm	85 (39.5)
Thigh	49 (22.8)
Chest	43 (20)
Foot	35 (16.3)
Back and buttocks	29 (13.5)
Head	12 (5.6)
Abdomen	3 (1.4)
Face	2 (0.9)
Degree of burn (n=215)	
First degree	89 (41.4)
Second degree	70 (32.6)
Third degree	26 (12.1)
Fourth degree	30 (14)
Removed clothing/accessories after burn (n=215)	
Yes	171 (79.5)
No	44 (20.5)
Sought or arranged first medical assistance (n=215)	
Yes	129 (60)
No	86 (40)
Covered injury site with clean cloth (n=215)	
Yes	93 (43.3)
No	122 (56.7)
Poured water on burn area (n=215)	
Yes	144 (66.9)
No	69 (32.1)
Was the water cold? (n=144)	
Yes	114 (79.2)
No	30 (20.8)
Duration of water pouring (n=144)	
Less than a minute	39 (27.1)
1-5 min	62 (43.1)

XSL•FO
RenderX

Variable	Values, n (%)
5-10 min	21 (14.6)
10-15 min	13 (9)
15-20 min	3 (2.1)
>20 min	6 (4.2)

Table 4 presents findings regarding post–burn medical consultation, management, complications, and the effect on daily activities among participants with a history of burn injury (n=215). Following a burn injury, 123 (57.2%) participants sought care at a health or emergency unit, while 92 (42.8%) did not. Among those who received medical evaluation (n=123), 118 (95.9%) were diagnosed and treated by a physician. The most commonly prescribed treatments were topical creams or ointments (87/118, 73.7%), bandages (24/118, 20.3%), surgical procedures (6/118, 5.1%), and medications (1/118, 0.8%). After treatment, 93.4% (110/118) reported clinical improvement, while 5.6% (8/118) did not. Complications were noted in 38 (17.7%) of the 215 participants, primarily complete deformation of skin appearance (28/38, 74%), followed by large blisters (4/38, 11%), scars (3/38, 8%), bacterial infection (1/38, 3%), bone or joint problems (1/38, 3%), and loss of affected parts (1/38, 3%). In addition, of the 215 participants, 36 (16.7%) participants indicated that burn injuries interfered with daily activities, while 179 (83.3%) stated that it did not (Table 4).

Table 4. Post–burn medical consultation, management, complications, and daily effect.

Variable	Values, n (%)
Went to health/emergency unit after burn (n=215)	
Yes	123 (57.2)
No	92 (42.8)
Diagnosed and prescribed treatment by a doctor (n=123)	
Yes	118 (95.9)
No	5 (4.1)
Type of prescribed intervention (n=118)	
Creams or ointments	87 (73.7)
Bandages	24 (20.3)
Drugs	1 (0.8)
Surgical intervention	6 (5.1)
Condition improved after treatment (n=118)	
Yes	110 (93.4)
No	8 (6.8)
Complications occurred after the burn (n=215)	
Yes	38 (17.7)
No	177 (82.3)
Type of complications (n=38)	
Complete skin deformation	28 (73.7)
Bacterial infections	1 (2.6)
Bone and joint problems	1 (2.6)
Large blisters	4 (10.5)
Burn marks/scars	3 (7.9)
Loss of affected parts	1 (2.6)
Burn became an obstacle to daily activity (n=215)	
Yes	36 (16.7)
No	179 (83.3)

XSL•FO
RenderX

Household safety measures were limited; 338 (83.7%) of the 404 participants preferred having a unified emergency number rather than multiple separate numbers. Only 103 (25.5%) had a fire extinguisher in their homes, 85 (21%) had an established evacuation plan, and 70 (17.3%) had practiced it. Smoke alarms were present in 33 (8.2%) homes, and 83 (20.5%) participants reported storing incendiary chemicals or flammable materials in a designated area.

Traditional remedies for burns were frequently reported (see Multimedia Appendix 3); 248 (61.4%) of the 404 participants reported using honey, followed by toothpaste (102/404, 25.2%), tomato paste (54/404, 13.4%), and aloe vera oil (16/404, 4%). Other less frequently reported remedies included flour (14/404, 3.5%), lavender oil (6/404, 1.5%), burn ointment (6/404, 1.5%), and mustard oil (5/404, 1.2%). Regarding the preferred methods for raising awareness about burn prevention and primary care, most participants favored visual educational materials (315/404, 78%) as the preferred medium for burn prevention awareness, followed by YouTube (229/404, 56.7%) and Twitter (195/404, 48.3%). Fewer respondents selected Facebook (70/404, 17.3%), Snapchat (65/404, 16.1%), newspapers (52/404, 12.9%), TikTok (21/404, 5.2%), and other sources (19/404, 4.7%).

Table 5 illustrates the factors associated with participants' overall knowledge of burn causes and first aid management among the 404 respondents. Among participants aged 18-30 years (n=279), 159 (57%) had poor knowledge and 120 (43%) had good knowledge (P=.87). Sex was significantly associated with knowledge level (P=.043): females (272/404, 46.3% good knowledge) demonstrated higher knowledge than males (132/404, 37.9%). Educational level also showed a significant relationship (P=.049): postgraduates (15/26, 58%) exhibited the highest knowledge level, followed by secondary education or below (26/54, 48%). No significant associations were found for nationality (P=.18), employment status (P=.68), or having children or teenagers at home (P=.48). Participants who had previously received information about burn prevention (236/404, 50.8% good knowledge) had significantly higher scores than those who had not (P=.001). A prior history of burn injury (104/215, 48.4% good knowledge) was also significantly associated with better knowledge (P=.04). The source of information did not significantly affect knowledge levels (P=.47) (Table 5).

XSL•FO
RenderX

Table 5. Factors associated with participants' knowledge of burn causes and first aid management.

Factor	Overall knowledge level, n (%)		*P* value[a]
	Poor	Good	
Age (y)			.87[b]
18-30 (n=279)	159 (57)	120 (43)	
31-45 (n=68)	37 (54.4)	31 (45.6)	
46-60 (n=52)	30 (57.7)	22 (42.3)	
>60 (n=5)	2 (40)	3 (60)	
Sex			.043[c]
Male (n=132)	82 (62.1)	50 (37.9)	
Female (n=272)	146 (53.7)	126 (46.3)	
Educational level			.049[c]
Secondary/below (n=54)	28 (51.9)	26 (48.1)	
Diploma/institute (n=30)	20 (66.7)	10 (33.3)	
Bachelor's degree (n=294)	169 (57.5)	125 (42.5)	
Postgraduate (n=26)	11 (42.3)	15 (57.7)	
Nationality			.18
Saudi (n=391)	223 (57)	168 (43)	
Non-Saudi (n=14)	5 (38.5)	8 (61.5)	
Employment			.68
Not working (n=91)	55 (60.4)	36 (39.6)	
Student (n=189)	105 (55.6)	84 (44.4)	
Working (n=124)	68 (54.8)	56 (45.2)	
Children or teenagers in home			.48
Yes (n=321)	184 (57.3)	137 (42.7)	
No (n=83)	44 (53)	39 (47)	
Received information about burn prevention			.001[c]
Yes (n=236)	116 (49.2)	120 (50.8)	
No (n=168)	112 (66.7)	56 (33.3)	
Source of burn prevention information			.47[b]
Internet (n=171)	88 (51.5)	83 (48.5)	
Television (n=67)	33 (49.3)	34 (50.7)	
Formal course (n=76)	32 (42.1)	44 (57.9)	
Paper leaflets (n=75)	34 (45.3)	41 (54.7)	
Daily newspapers (n=23)	14 (60.9)	9 (39.1)	
Radio (n=14)	7 (50)	7 (50)	
Personal or close burn experience			.04[c]
Yes (n=215)	111 (51.6)	104 (48.4)	
No (n=189)	117 (61.9)	72 (38.1)	

[a]Pearson X^2 test.

[b]Exact probability test.

[c]Significant at $P<.05$.

XSL•FO
RenderX

Discussion

This study aimed to assess public knowledge and practice toward burns first aid management. In 2010, the results of 71 studies were analyzed in a systematic review of burn injuries that covered 12 of the 22 countries that make up the East Mediterranean region [19]. Overall, the incidence of burn injuries varied between 112 and 518 per 100,000 people. Nine of these 71 studies came from Saudi Arabia. However, there is no published summary of burn injuries in Saudi Arabia up to this point. There have been several concerns in the past few years to lower the risk of burns by raising public awareness and motivating people to take burn prevention steps. Consequently, there appears to have been a decline in burn incidence and burn-related mortality, particularly in high-income nations [19].

As for participants' personal or family history of burns, our study reveals that 53.2% (215/404) of the total participants had experienced burns before—with hot water, fire, and chemicals being the most common causes. The most common burn sites were the hands, arm, and thigh. Most participants had first-degree burns, followed by medical assistance and management received by more than half of the respondents with burns. Most went to the health/emergency unit, with most receiving treatment from doctors. Most treated patients showed an improvement, but 17.7% (38/215) experienced complications, including skin deformation and difficulty in daily activities. A previous study in Jazan by Hoque et al [20] revealed that nearly three-fourths of the participants had experienced a burn themselves or in a family member, which is higher than the assessed prevalence among this study's respondents. Thermal burns were the most dominant similar to that reported in this study.

Considering knowledge and awareness, this study revealed that more than half of the participants received information about burn prevention methods mainly from the internet and formal courses. Study participants showed high awareness regarding the importance of medical consultation, visiting emergency for burns and the cornerstone role of plastic surgeons in burns. Knowledge of the initial first aid steps for burns was substantially lower. Generally, less than half of the study participants were knowledgeable about burns causes and first aid management. Similar findings in Jazan region were reported by Hoque et al [20], as 69.9% of the participants had inadequate knowledge, although 72% of participants had a favorable attitude toward burn first aid. Additionally, a Jordanian study found that 59% of the participants knew that covering the burned area with a clean cloth can reduce the risk of infection [21]. However, the Jordanian study showed a lower percentage of participants knowing about covering the burned area correctly with a clean cloth. That study also found that a significant proportion of participants believed that toothpaste could be used to treat burns, which is a common misconception. A much better knowledge was reported by Abu et al [22], the most recent study in Jazan, where 91.8% (n=367) knew to use water, 50.8% (n=203) would seek medical attention, 22.6% (n=90) would cover burns, and 20.3% (n=81) considered pain management important. This suggests that adults in Jazan have a good understanding of the causes, prevention, and first aid procedures for burn injuries in contrast to our study findings.

Although a substantial proportion of participants reported using the internet as a primary source of health information, the overall knowledge about burns first aid remained low. This discrepancy may be explained by the quality and reliability of online information accessed by the general public. Many people rely on social media and other informal online sources, where misinformation is widespread and scientific accuracy is often missing. In addition, most individuals in the region participate with health information passively, accessing it by chance rather than actively searching for reliable, evidence-based guidance on emergency care such as burn first aid.

Cultural beliefs and traditional practices also play a significant role in shaping burn management behaviors. The use of home remedies such as honey, toothpaste, tomato paste, or aloe vera oil means long-standing community traditions and trust in folk medicine passes down through generations. These practices are often reinforced through family, friends, and social networks, sometimes even shared on digital platforms, further perpetuating misconceptions. In addition, limited exposure to official first aid training programs and a general perception that slight burns can be self-treated at home contribute to inappropriate first aid responses.

Regarding post-burn practice, our study shows that more than half of the patients with burn injuries went to the health/emergency unit, with 95.9% (118/123) diagnosed and prescribed treatment. Most patients received creams or ointments and bandages, with very few cases requiring surgery. Most patients experienced an improvement, with complications in less than one-fifth of the cases involving skin deformation and 16.7% (36/215) affecting daily activities. A large-scale study in the western region of Saudi Arabia revealed that most patients with burn injuries (84.2%) had treated their burns with traditional medication [23]. Because most respondents agreed incorrectly that antibiotics are helpful in cases of burns, there was a lack of knowledge regarding the use of antibiotics in burn injuries.

Honey is thought to be a potentially promising therapy because it can promote tissue growth, increase epithelialization, and decrease scar formation [24]. As shown in our study, about two-thirds of the participants selected honey as a first-aid burn treatment as the most used conventional method. According to a Saudi Arabian study, 69.9% of the participants use honey as a burn treatment in concordance to our study finding [25]. It should be highlighted that, when compared to early excision and grafting, the results of a systematic review that compared the use of honey to other conventional dressing procedures revealed significant uncertainties and may postpone the ability of partial and full-thickness burns to heal [26,27].

In conclusion, our study reveals unsatisfactory knowledge and awareness about burn causes and first aid management among the Jazan public. Higher education, female sex, and a personal or family history of burns were associated with higher knowledge levels. Internet was the most commonly reported source of information, with seeking medical consultation among nearly half of the cases. Nearly half of the participants had

XSL•FO
RenderX

experienced burns or known someone had, mainly on the hands, arms, and thigh. Thermal burns were the most frequent. The findings suggest a lack of awareness among the general population about the appropriate first-aid measures for burns, and health care professionals should take an active role in educating patients and communities about burn first aids.

Acknowledgments

We extend our deepest gratitude to all the participants who contributed to this study. We also thank the individuals who supported us during the data collection and analysis phases. The authors attest that no generative artificial intelligence tools were used in the writing or analysis of this manuscript.

Data Availability

Data may be available upon request to the corresponding author.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Authors' Contributions

AS designed the study, provided supervision, performed data analysis, and critically revised the manuscript. AYA conceived and designed the study, managed data collection, and contributed to writing. AF helped design the study, provided supervision, and participated in the critical review. AMA designed the study, handled data acquisition and analysis, contributed to writing, and assisted in revisions and final drafting. RMA and RFA participated in data collection, analysis, writing, manuscript drafting, and review. GAS contributed to data collection, analysis, writing, and final revisions. All authors reviewed and approved the final version and accepted responsibility for the work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Structured questionnaire administered in Arabic.
[DOCX File , 34 KB - formative_v9i1e80594_app1.docx]

Multimedia Appendix 2
Burns first-aid practices.
[PDF File (Adobe PDF File), 95 KB - formative_v9i1e80594_app2.pdf]

Multimedia Appendix 3
Traditional remedies for burns.
[PDF File (Adobe PDF File), 81 KB - formative_v9i1e80594_app3.pdf]

References

1. Mehta K, Arega H, Smith NL, Li K, Gause E, Lee J, et al. Gender-based disparities in burn injuries, care and outcomes: a World Health Organization (WHO) Global Burn Registry cohort study. Am J Surg 2022 Jan;223(1):157-163 [FREE Full text] [doi: 10.1016/j.amjsurg.2021.07.041] [Medline: 34330521]
2. Tyson AF, Boschini LP, Kiser MM, Samuel JC, Mjuweni SN, Cairns BA, et al. Survival after burn in a sub-Saharan burn unit: challenges and opportunities. Burns 2013 Dec;39(8):1619-1625 [FREE Full text] [doi: 10.1016/j.burns.2013.04.013] [Medline: 23768710]
3. Charles AG, Gallaher J, Cairns BA. Burn care in low- and middle-income countries. Clin Plast Surg 2017 Jul;44(3):479-483. [doi: 10.1016/j.cps.2017.02.007] [Medline: 28576236]
4. Agbenorku P. Burns functional disabilities among burn survivors: a study in Komfo Anokye Teaching Hospital, Ghana. Int J Burns Trauma 2013;3(2):78-86 [FREE Full text] [Medline: 23638325]
5. Jeschke MG, van Baar ME, Choudhry MA, Chung KK, Gibran NS, Logsetty S. Burn injury. Nat Rev Dis Primers 2020 Feb 13;6(1):11 [FREE Full text] [doi: 10.1038/s41572-020-0145-5] [Medline: 32054846]
6. Ahuja RB. ISBI Practice Guidelines for Burn Care. Burns 2016 Aug;42(5):951-952. [doi: 10.1016/j.burns.2016.06.020] [Medline: 27422355]
7. Kara Y. Burn etiology and pathogenesis. Hot Topics in Burn Injuries. 23 2018 May 23;17(1):2018. [doi: 10.5772/intechopen.71379]

XSL•FO
RenderX

8. Friedstat J, Brown DA, Levi B. Chemical, electrical, and radiation injuries. Clin Plast Surg 2017 Jul;44(3):657-669 [FREE Full text] [doi: 10.1016/j.cps.2017.02.021] [Medline: 28576255]
9. Atiyeh BS, Costagliola M, Hayek SN. Burn prevention mechanisms and outcomes: pitfalls, failures and successes. Burns 2009 Mar;35(2):181-193. [doi: 10.1016/j.burns.2008.06.002] [Medline: 18926639]
10. Hunt J, Arnoldo B, Purdue G. Prevention of burn injuries. In: Total Burn Care 4th edition. Philadelphia, PA: WB Saunders; 2012:47-55.
11. Kendrick D, Young B, Mason-Jones AJ, Ilyas N, Achana FA, Cooper NJ, et al. Home safety education and provision of safety equipment for injury prevention (Review). Evid Based Child Health 2013 May;8(3):761-939. [doi: 10.1002/ebch.1911] [Medline: 23877910]
12. Harish V, Tiwari N, Fisher OM, Li Z, Maitz PK. First aid improves clinical outcomes in burn injuries: Evidence from a cohort study of 4918 patients. Burns 2019 Mar;45(2):433-439. [doi: 10.1016/j.burns.2018.09.024] [Medline: 30337155]
13. Cuttle L, Kimble R. First aid treatment of burn injuries. Wound Pract Res 2010;18(1):1-7 [FREE Full text]
14. Almarghoub M, Alotaibi A, Alyamani A, Alfaqeeh FA, Almehaid FF, Al-Qattan MM, et al. The epidemiology of burn injuries in Saudi Arabia: a systematic review. J Burn Care Res 2020 Sep 23;41(5):1122-1127. [doi: 10.1093/jbcr/iraa084] [Medline: 32479634]
15. Almutlaq BA, Jarman A, Alfraihi R, Albasher G, Alotaibi RM, Alqahtani AS, et al. Skin burns in Saudi Arabia: causes, management, outcomes and quality of life after skin burns. Int J Burns Trauma 2020;10(2):28-37 [FREE Full text] [Medline: 32419974]
16. Alajmi MM, Aldosari KH, Al-Ghamdi S. Clinical, epidemiological, and management aspects of burn injuries in Saudi Arabia - a cross-sectional study. Saudi J Biol Sci 2021 Aug;28(8):4342-4347 [FREE Full text] [doi: 10.1016/j.sjbs.2021.04.021] [Medline: 34354418]
17. Alharthy N, Al Mutairi M, AlQueflie S, Nefesa A, Manie N, Nafesa S, et al. Pattern of burns identified in the Pediatrics Emergency Department at King Abdul-Aziz Medical City: Riyadh. J Nat Sci Biol Med 2016;7(1):16-21 [FREE Full text] [doi: 10.4103/0976-9668.175019] [Medline: 27003963]
18. Standards and Recommendations for Burns Care in Mass Casualty Incidents. Geneva: World Health Organization; Nov 25, 2024.
19. Smolle C, Cambiaso-Daniel J, Forbes AA, Wurzer P, Hundeshagen G, Branski LK, et al. Recent trends in burn epidemiology worldwide: a systematic review. Burns 2017 Mar;43(2):249-257 [FREE Full text] [doi: 10.1016/j.burns.2016.08.013] [Medline: 27600982]
20. Hoque MM, Moukhyer ME, Khamjan NA, Siddig BA, Maguen E, Alam S, et al. Community knowledge and attitude about Burns First Aid (BFA) in Jazan, Saudi Arabia. JBM 2024;12(04):135-148. [doi: 10.4236/jbm.2024.124012]
21. Al Shikh H. Knowledge, attitudes, and practices concerning burn first aid among parents in Jordyn: a cross-sectional study. JJNR 2025;4(2):1-10. [doi: 10.14525/jjnr.v4i2.05]
22. Abu HJ, Mojiri M, Alhassan R. Awareness of burn injury prevention and first aid management among adults in Jazan, Saudi Arabia. Cureus 2024;16(9):2. [doi: 10.7759/cureus.68456]
23. Abu AR, Awan B, Alsulymani B. Evaluation of perception, awareness, and practices related to burns first aid: largest cross-sectional study among non-healthcare providers in the western region of Saudi Arabia. Cureus 2023;15(1):16. [doi: 10.7759/cureus.33839]
24. Al-Waili N, Salom K, Al-Ghamdi AA. Honey for wound healing, ulcers, and burns; data supporting its use in clinical practice. Scientific World Journal 2011 Apr 05;11:766-787 [FREE Full text] [doi: 10.1100/tsw.2011.78] [Medline: 21479349]
25. Kattan A, AlShomer F, Alhujayri A, Addar A, Aljerian A. Current knowledge of burn injury first aid practices and applied traditional remedies: a nationwide survey. Burns Trauma 2016;4:2. [doi: 10.1186/s41038-016-0063-7]
26. Maghsoudi H, Salehi F, Khosrowshahi MK, Baghaei M, Nasirzadeh M, Shams R. Comparison between topical honey and mafenide acetate in treatment of burn wounds. Ann Burns Fire Disasters 2011 Sep 30;24(3):132-137 [FREE Full text] [Medline: 22396671]
27. Gupta S, Singh O, Bhagel P, Moses S, Shukla S, Mathur R. Honey dressing versus silver sulfadiazene dressing for wound healing in burn patients: a retrospective study. J Cutan Aesthet Surg 2011 Sep;4(3):183-187 [FREE Full text] [doi: 10.4103/0974-2077.91249] [Medline: 22279383]

Abbreviations

LMICs: low- and middle-income countries
WHO: World Health Organization

XSL•FO
RenderX

Edited by A Stone; submitted 13.07.25; peer-reviewed by FH Dallak; comments to author 30.09.25; accepted 10.11.25; published 23.12.25.

Please cite as:
Sayegh A, Awaji AY, Fathuldeen A, Alshammakhi AM, Alhazmi RM, Ageeli RF, Shamakhi GA
Knowledge, Attitudes, and Practices Toward Burn Causes and First Aid Management in Jazan Region, Saudi Arabia: Cross-Sectional Study
JMIR Form Res 2025;9:e80594
URL: https://formative.jmir.org/2025/1/e80594
doi:10.2196/80594
PMID:

©Anas Sayegh, Alyazid Y Awaji, Anas Fathuldeen, Atheer M Alshammakhi, Roaa M Alhazmi, Razan F Ageeli, Ghadi A Shamakhi. Originally published in JMIR Formative Research (https://formative.jmir.org), 23.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Development of an Evaluation Index System for Health Recommender Systems Based on the Health Technology Assessment Framework: Cross-Sectional Delphi Study

Yue Sun[1], PhD; Shijie Hou[2], MBBS; Siye Chen[2], MSc; Minmin Leng[3], PhD; Zhiwen Wang[2], Prof Dr

[1]School of Nursing, Health Science Center, Xi'an Jiaotong University, Xi'an, Shaanxi Province, China

[2]School of Nursing, Peking University, 38 Xueyuan Road, Haidian District, Beijing, China

[3]Department of Nursing, Shandong Provincial Hospital Affiliated to Shandong First Medical University, Jinan, Shandong Province, China

Corresponding Author:
Zhiwen Wang, Prof Dr
School of Nursing, Peking University, 38 Xueyuan Road, Haidian District, Beijing, China

Abstract

Background: Health recommender systems (HRSs) are digital platforms designed to deliver personalized health information, resources, and interventions tailored to users' specific needs. However, existing evaluations of HRSs largely focus on algorithmic performance, with limited scientific evidence supporting user-centered assessment approaches and insufficiently defined evaluation metrics. Moreover, no unified or scientifically validated framework currently exists for evaluating these systems, resulting in limited cross-study comparability and constraining regulatory and implementation decision-making.

Objective: This study aimed to develop a comprehensive, consensus-based evaluation index system for HRSs grounded in the health technology assessment (HTA) framework.

Methods: This cross-sectional study used a 2-round Delphi process conducted with 18 experts comprising clinicians, digital health researchers, and policymakers who possessed relevant professional experience and domain knowledge in HRSs. The age range of the experts was between 30 and 58 years, with 67% (n=12) of them possessing over 10 years of professional experience. On the basis of literature analysis and HTA principles, a preliminary indicator set comprising 5 primary and 16 secondary indicators was constructed. Experts rated the importance of each indicator using a 5-point Likert scale and provided qualitative suggestions for refinement. After the Delphi process, the analytic hierarchy process was applied to determine indicator weights and assess consistency.

Results: The Delphi survey reached full participation in the first round (18/18, 100%) and maintained an 88.9% (16/18) response rate in the second round. The final evaluation index system of HRSs contained 5 first-level indicators (performance, effectiveness, safety, economy, and social appropriateness) and 18 second-level indicators. The mean importance scores of the second-level indicators ranged from 4.25 (SD 0.45) to 5.00 (SD 0.00), with coefficients of variation between 0.000 and 0.220. Among the first-level indicators, safety received the highest weight (0.289), followed by social appropriateness (0.251), effectiveness (0.193), performance (0.136), and economy (0.132).

Conclusions: This study presents an evaluation index system for HRSs grounded in the HTA framework and validated through expert consensus. The resulting framework not only provides actionable guidance for the design, optimization, and implementation of HRSs but also fills a methodological gap in the field by offering quantifiable, hierarchical evaluation indicators with validated weighting. Future research will involve iterative refinement and empirical validation of the system in real-world deployment settings, thereby enabling continuous improvement and facilitating the establishment of unified evaluation standards for HRS research and practice.

(JMIR Form Res 2025;9:e79997) doi:10.2196/79997

KEYWORDS

health recommender systems; health technology assessment evaluation framework; HTA evaluation framework; evaluation index system; Delphi survey; artificial intelligence; AI

Introduction

As technology evolves, new ways to implement tailored interventions are being adopted, and researchers and policymakers require access to appropriate tools to assess the design and suitability for use of these interventions [1]. One such innovative approach to computer-based tailored health interventions is recommender systems (RSs) [2]. Health RSs (HRSs), specialized digital platforms or software applications

XSL•FO
RenderX

designed to recommend personalized information, resources, or interventions relevant to the user's specific health needs, are now available to address this need [3]. A scoping review of the current literature identified 51 studies on HRSs covering a range of health domains, including general health promotion, lifestyle, generic health service, and others [4]. These HRSs use sophisticated data analytic techniques, integrating machine learning algorithms and artificial intelligence frameworks to process and interpret diverse user-specific datasets, including medical histories, behavioral patterns, and individual preferences. By synthesizing these multifaceted data sources, HRSs generate personalized recommendations, targeted resources, and customized interventions tailored to the unique needs of each user [5]. Given that the information provided by HRSs can have a significant impact on individuals' health decisions, it is essential to conduct comprehensive evaluations before their widespread adoption. Such assessments not only enable developers to refine the design and functionality of these applications but also establish a scientific foundation for evaluating their effectiveness and ensuring their safe and reliable use [6,7].

In 2018, Moshi et al [8] highlighted that the health technology assessment (HTA) evaluation framework then in use was inadequate for delivering holistic assessments of mobile health solutions in a clinical environment. Since then, the field has advanced significantly, driven by the growing development and application of HTA evaluation frameworks for assessing digital health technologies (DHTs), especially mobile health [9,10]. As a health technology framework, HTA considerations encompass effectiveness, safety, and cost-effectiveness, as well as patient- and organizational-level factors, while also addressing ethical, social, and legal issues [11]. Furthermore, they include the technical characteristics of DHTs [12]. A scoping review of HRS research [4] found that most studies focus on performance metrics, with only a limited number of HRS studies incorporating evaluations of user participation, health outcomes, user acceptance, or feasibility. Key dimensions such as safety, economic impact, and social appropriateness have been largely overlooked. For example, these studies are either narrowly focused on specific system dimensions (eg, recommendation techniques) or health-related outcome measures [13-18] or they provide only a general overview of HRSs without giving adequate emphasis to evaluation [5,19]. De Croon et al [19] reviewed existing research related to HRSs and found that there is a lack of scientific evidence on user-centered evaluation approaches and that other metric parameters are ambiguous in terms of HRS evaluation and the feasibility of the organizational management of this technology. For example, the definition of performance varied across studies. Torrent-Fontbona and Lopez [20] used the amount of time in the glycemic target range by reducing the time below the target as a performance metric, whereas Cho et al [21] reported performance based on precision and recall. Given the heterogeneity in evaluation methods, it is necessary to develop a comprehensive and scientifically sound evaluation tool to provide reference and guidance for HRSs.

Accordingly, this study consisted of 2 main phases: first, summarizing and developing the evaluation indicators by reviewing the literature within the HTA framework and, second, identifying specific indicators and determining their weights through a Delphi expert consultation process and the analytic hierarchy process (AHP). Adopting an HTA-based evaluation indicator system could provide robust support for more systematic and in-depth assessment and offer guidance for the design, optimization, and implementation of HRSs in the health care domain [11].

Methods

Study Design

This study used a Delphi method [22,23], a widely recognized approach for selecting quality indicators in health care. The Delphi method does not prescribe a fixed number of rounds. Previous studies have consistently shown that 2 to 3 iterations are typically sufficient to reach consensus [24]. The process concludes when consensus is achieved on the topics under discussion. This study was designed and reported following the Delphi Studies in Social and Health Sciences–Recommendations for an Interdisciplinary Standardized Reporting framework [25], which provides a consensus-based methodological and reporting standard for Delphi research.

This study comprised 2 rounds of questionnaires administered to an expert panel via email, conducted in accordance with established methodological guidelines for Delphi surveys. The process consisted of two key stages: (1) developing a preliminary set of potential evaluation indicators along with a conceptual framework and secondary-level metrics for HRSs through a systematic literature review and (2) conducting a Delphi survey to prioritize and reach consensus on the primary and secondary evaluation indicators for HRSs. The specific process is illustrated in Textbox 1.

XSL•FO
RenderX

Textbox 1. Summary of the 4-phase process used to develop the evaluation framework for the health recommender system, including expert panel selection, literature review, questionnaire preparation, and the Delphi–analytic hierarchy process procedure.

Phase 1: expert panel

- Members held a postgraduate degree or higher, possessed a midlevel professional title or above, or had a minimum of 10 years of relevant work experience.
- Members were engaged in clinical medicine, clinical nursing, public health management, health-related government departments, or information technology related to recommender systems.

Phase 2: literature review

- Databases included PubMed, Embase, Web of Science, ACM, IEEE Xplore, ScienceDirect, CNKI, and Wanfang.
- Professional organization websites such as the World Health Organization and UNESCO were consulted.

Phase 3: questionnaire preparation

- The focus was on 5 aspects: technical characteristics, effectiveness, safety, economic considerations, and social appropriateness.
- The importance of each indicator was measured by its relevance to health recommender system performance, rated on a Likert scale from 1 to 5 (1="not important"; 5="very important").

Phase 4: Delphi method and the analytic hierarchy process

- The reliability of the expert consultation was assessed through the expert positive coefficient, the authority coefficient, and the degree of opinion concentration and coordination.
- Using the Saaty scale, a judgment matrix was constructed, followed by hierarchical ranking and consistency testing.

Participants

The selection criteria for experts were established as follows: (1) holding a postgraduate degree or higher, possessing a mid-level (or above) professional title, or having at least 10 years of relevant work experience; (2) being actively engaged in fields such as health IT, HTA, public health management, clinical medicine, nursing, or health-related government agencies, with substantial expertise and a thorough understanding of the research topic; and (3) voluntarily agreeing to participate in the study. Individuals lacking familiarity with HRSs were excluded from the selection process. A total of 20 eligible experts were invited via email, and 18 (90%) agreed to participate in the study.

Questionnaire Preparation

This study, based on the HTA evaluation framework, aimed to develop a preliminary hierarchical framework for the evaluation indicator system of HRSs. Drawing on previous assessments of emerging health technologies, the evaluation focused on 5 key domains: technical characteristics, effectiveness, safety, economy, and social appropriateness, as shown in Table 1.

Table . The five-dimensional measurement framework for health recommender systems based on the health technology assessment model.

Aspect	Definition
Technical characteristics	Refers to the system's effectiveness and reliability in delivering stable, high-quality recommendations.
Effectiveness	Refers to the ability of the health recommender system to achieve its intended objectives in real-world applications, such as promoting behavior change or improving health outcomes.
Safety	Refers to the system's capacity to ensure user privacy, data security, operational stability, and overall user safety throughout its use.
Economy	Refers to the economic feasibility and cost-effectiveness of the health recommender system while fulfilling its functional requirements.
Social appropriateness	Refers to the degree to which the health recommender system aligns with sociocultural norms, ethical principles, and legal regulations, and is acceptable to users and society.

The following 8 electronic databases were searched on October 3, 2022: PubMed, Embase, Web of Science, ACM, IEEE Xplore, ScienceDirect, China National Knowledge Infrastructure, and Wanfang. Electronic searches were conducted using the following keywords: ("recommender systems") OR ("recommender system") OR ("recommendation systems") OR ("recommendation system") AND (health OR patient OR patients). To identify additional studies, we screened professional organization websites, including the World Health Organization, UNESCO, the Ministry of Health of the People's Republic of China, the National Health Commission of China, and other organizational or governmental websites.

XSL•FO
RenderX

Potential studies including competency indicators were extracted and screened by 2 reviewers according to the following criteria: (1) studies that described or implemented HRSs with a primary focus on improving health and included an evaluation of the HRSs and (2) studies published in English or Chinese. In cases of uncertainty regarding the inclusion of a study, the research team held discussions to reach a consensus. On the basis of the analysis of the literature, a preliminary evaluation index system for HRSs was developed comprising 5 primary indicators (dimensions) and 16 secondary indicators.

The preliminary evaluation index system for HRSs was operationalized into a Delphi-based survey instrument. Within this questionnaire, the importance of each indicator was defined as its relevance in assessing the overall performance of HRSs, quantified using a Likert scale from 1 to 5 (1="not important at all"; 5="very important"). Additionally, the survey included open-ended questions to allow domain experts to provide feedback on the existing indicators and suggest additional indicators they deemed essential for inclusion.

Delphi Method and the AHP

In the initial phase of expert consultations, a 5-point Likert scale was used by the panel of experts to evaluate and score the significance of each indicator, with an additional open-ended section for suggesting modifications. Upon the conclusion of this phase, the criteria for item retention were delineated as follows: an item's mean importance score must surpass 4.0, the coefficient of variation (CV) must not exceed 0.25, and the selection rate for the highest score of 5 must be above 20% [25]. During the subsequent round of consultations, the anonymized results from the first round were disseminated among the experts, and the consultation form was augmented with a section for the allocation of index weights. In the final round, a select group of experts assigned weights to the index system, and the AHP was used to ascertain the weight values for the finalized evaluation index system [26]. This system is hierarchically organized into 3 strata: the goal layer, which encapsulates the overarching objective of the HRS evaluation; the criterion layer, encompassing the principal evaluation dimensions; and the indicator layer, comprising specific metrics. Experts conducted pairwise comparisons of indicators within the same stratum using a scale from 1 to 9 to quantify their relative importance (with 1 denoting equal importance and 9 signifying extreme importance), thereby constructing the judgment matrix [27].

Index Screening and Result Statistics

The collected data were processed using Microsoft Excel before being input into a database for further analysis. Data analysis was conducted using Microsoft Excel 2021 and SPSS (version 26.0; IBM Corp). Descriptive statistics were used to calculate the mean, SD, and CV of the importance and feasibility of each indicator.

The reliability of the expert consultation was assessed through the expert positive coefficient, the expert authority coefficient (Cr), the degree of expert opinion concentration, and the degree of expert opinion coordination [28]. The expert positive coefficient was represented by the questionnaire response rate. The Cr was used to reflect the authority of expert opinions, calculated based on expert judgment (Ca) and their familiarity with the dimensions and items (Cs) using the formula Cr = (Ca + Cs)/2. The Kendall coefficient of concordance (W) and the CV were used to measure the degree of consensus among experts. A higher value of W indicates a stronger level of consensus among experts.

The weights of the indicators were determined using the AHP [29]. Through assigning importance values to each indicator, we established the Saaty scale, constructed the judgment matrix, and conducted a hierarchical ranking and consistency test. The consistency ratio is calculated to assess the judgment matrix. When the consistency ratio value is less than 0.10, it indicates that the judgment matrix exhibits satisfactory consistency and does not require further adjustments.

Ethical Considerations

Ethics approval was obtained from the Biomedical Ethics Committee of Peking University (IRB00001052-22042). All participants provided informed consent through an electronic registration form where they were clearly informed that their responses would remain confidential and be used solely for research purposes. Participants were notified of their right to withdraw from the study at any time without penalty.

To ensure confidentiality and adhere to ethical standards, the questionnaire was administered anonymously. Only participants' email addresses were collected for follow-up purposes, and these were stored separately from all demographic and survey data. All responses were deidentified during data processing, sharing of interim results between Delphi rounds, and final reporting. No personally identifiable information is present in this manuscript.

Access to identifiable data was restricted exclusively to the research team, and all data were securely stored within institutional systems. No compensation was provided to participants for their involvement in the study.

Results

Overview

The initial literature search yielded a total of 1538 articles. After removing duplicates, of the 1538 articles, 1239 (80.6%) were screened by title and abstract for eligibility. Of these 1239 articles, 77 (6.2%) studies related to the evaluation of HRSs were included in the analysis. Figure 1 provides an overview of the identification and selection of studies at different stages of the screening process. The included studies were systematically organized based on the identified HTA evaluation framework, as detailed in Table S1 in Multimedia Appendix 1.

XSL•FO
RenderX

Figure 1. The process of study identification and selection. Flow diagram illustrating the literature search, screening, eligibility assessment, and final inclusion of studies used to inform the development of the health recommender system evaluation index system.

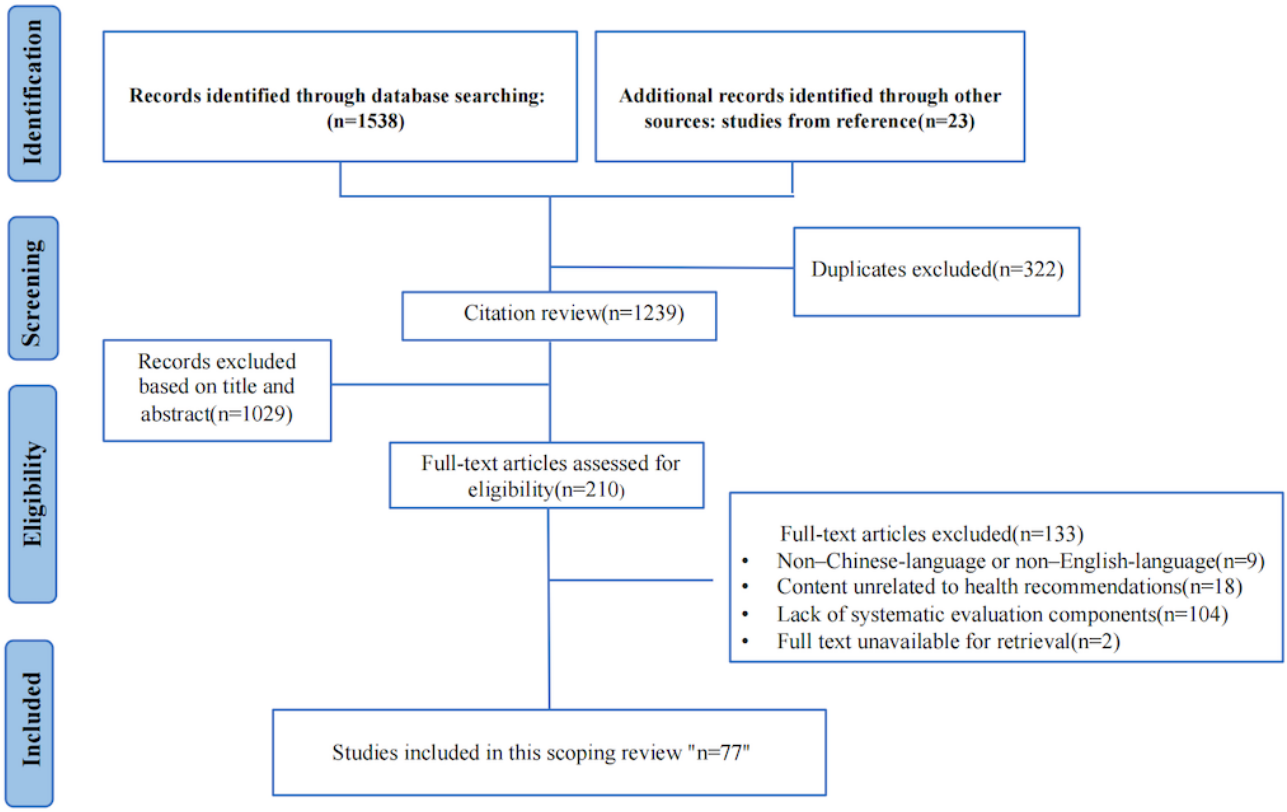

Basic Information on the Experts

A total of 18 experts participated in the first round of the Delphi consultation. These experts were primarily engaged in fields such as clinical medicine, clinical nursing, public health management, health-related government departments, and IT related to RSs. The age range of the experts was between 30 and 58 years, with 67% (12/18) of them possessing over 10 years of professional experience. All experts (18/18, 100%) held a master's degree or higher, and 72% (13/18) held senior professional titles (Table 2).

XSL•FO
RenderX

Table . Demographic characteristics of the 18 experts participating in the 2-round Delphi study (N=18).

Variable	Participants, n (%)
Age (y)	
30-39	10 (56)
40-49	6 (33)
≥50	2 (11)
Work experience (y)	
<10	6 (33)
10-19	8 (44)
20-29	4 (22)
Highest degree	
Master's	9 (50)
PhD	9 (50)
Professional title	
Intermediate grade title	1 (5)
Deputy senior grade title	4 (22)
Senior grade title	13 (72)
Field of expertise	
Clinical medicine	5 (27)
Clinical nursing	4 (22)
Public health management	3 (16)
Health-related government departments	2 (11)
IT related to recommender systems	4 (22)

Analysis of Expert Participation

Expert Positive Coefficient

Expert participation was assessed using the response rate and the proportion of experts who provided written feedback. In the first Delphi round, all distributed questionnaires were returned (18/18, 100% response rate) and an expert enthusiasm coefficient of 1.00. In the second round, 89% (16/18) of the questionnaires were returned. The retention rate across rounds was 89% (16/18) and an expert enthusiasm coefficient of 0.89.

The Cr Measure

Expert authority was assessed using the authority coefficient (Cr), defined as Cr = (Ca + Cs) / 2. In round 1, the mean Cr was 0.883, derived from a judgment coefficient (Ca) of 0.933 and a familiarity coefficient (Cs) of 0.833. In round 2, the mean Cr was 0.884, based on a Ca of 0.940 and a Cs of 0.827. All Cr values exceeded the commonly accepted threshold of 0.7, indicating a consistently high level of expert authority across both rounds.

The Degree of Expert Opinion Concentration

The degree of expert opinion concentration was assessed using the Kendall coefficient of concordance (*W*). In round 1, *W* was 0.281 for dimensions and 0.282 for items (c^2=25.254 and 121.806; *P*<.001). In round 2, the values were 0.360 and 0.236 (c^2=23.067 and 83.096; *P*<.001).

Results of the Delphi Survey

First Round

The evaluation index system comprised 16 items across 5 dimensions. For the primary indicators, the mean importance scores ranged from 4.39 (SD 0.59) to 5.00 (SD 0.00), with CVs ranging from 0.000 to 0.172. All primary indicators met the inclusion criteria and were retained without modification. For the secondary indicators, the CVs ranged from 0.000 to 0.173, and the mean importance scores ranged from 4.17 (SD 0.60) to 5.00 (SD 0.00). Detailed results for dimensions and items are presented in Tables S2 and S3 in Multimedia Appendix 1 and Table 3.

XSL•FO
RenderX

Table . Adjustments to competency indicators for health recommender systems after round 1 of the Delphi survey.

Primary indicators	Secondary indicators	Adjustment
1. Performance	1.1 Accuracy (retained) 1.2 Content coverage (retained) 1.3 Result diversity (retained) 1.4 User trust (retained) 1.5 System robustness (retained)	• Added "1.6 System response efficiency"
2. Effectiveness	2.1 Health behavior outcomes (modified) 2.2 Quality of life impact (modified) 2.3 Health service efficiency (modified)	• Added "2.1 Health behavior" • Merged 2.1 and 2.2 as "2.2 Health outcome" • Modified 2.3 to "2.3 Quality and efficiency of health services"
3. Safety	3.1 Clinical safety (retained) 3.2 Technical application safety (retained)	• No adjustments
4. Economy	4.1 Patient cost-effectiveness (modified) 4.2 Institutional cost-effectiveness (modified) 4.3 Social benefits (retained)	• Modified 4.1 to "4.1 Patient-level economy" • Modified 4.2 to "4.2 Institutional-level economy"
5. Social appropriateness	5.1 Data privacy protection (deleted) 5.2 Legal and policy compliance (modified) 5.3 Acceptability (retained)	• Replaced 5.1 with "5.1 Ethicality" • Modified 5.2 to "5.2 Policy suitability" • Added "5.4 Feasibility"

[a]Not applicable.

Second Round

For the primary indicators, the mean importance scores ranged from 4.31 (SD 0.59) to 4.94 (SD 0.37), with CVs ranging from 0.00 to 0.141. All primary indicators met the inclusion criteria and were retained without modification. The mean importance scores of the second-level indicators ranged from 4.25 (SD 0.45) to 5.00 (SD 0.00), with CVs ranging between 0.000 and 0.220. Detailed results for all indicators are provided in Tables S4 and S5 in Multimedia Appendix 1. No additional feedback or suggestions were provided by the experts. After 2 rounds of the Delphi survey, the evaluation index system for HRSs was established, comprising 5 first-level indicators and 18 second-level indicators (Table 4). Explanations of the indicators are provided in Table 1 and Table S6 in Multimedia Appendix 1.

Table . Final evaluation index system for health recommender systems developed through a 2-round Delphi consensus study.

Primary indicator	Secondary indicators
Performance	• 1.1 Accuracy • 1.2 Coverage • 1.3 Result diversity • 1.4 User trust • 1.5 Robustness • 1.6 Response efficiency
Effectiveness	• 2.1 Health behavior • 2.2 Health outcome • 2.3 Quality and efficiency of health services
Safety	• 3.1 Clinical safety • 3.2 Technical safety
Economy	• 4.1 Patient-level economy • 4.2 Institutional-level economy • 4.3 Social benefit
Social appropriateness	• 5.1 Ethicality • 5.2 Policy appropriateness • 5.3 Acceptability • 5.4 Feasibility

XSL•FO
RenderX

Results of the AHP

The weights assigned to the primary indicators were as follows: 0.136 for performance, 0.193 for effectiveness, 0.289 for safety, 0.132 for economy, and 0.251 for social suitability. The consistency value of 0.032 demonstrates the reliability of the results. Among the secondary indicators, the top 4 weights were attributed to technical safety (0.154), clinical safety (0.135), acceptability (0.065), and policy appropriateness (0.059). All secondary indicators exhibited consistency values below 0.1, further validating the rationality of the weight assignments. The combined weights of the evaluation indicators for HRSs are shown in Table 5.

Table . Combined weights of primary and secondary evaluation indicators for health recommender systems obtained through the analytic hierarchy process.

Primary indicator (weight) and secondary indicator	Weight	Combined weight
1. Performance	0.136	—[a]
1.1 Accuracy	0.144	0.020
1.2 Coverage	0.151	0.021
1.3 Result diversity	0.157	0.021
1.4 User trust	0.173	0.024
1.5 Robustness	0.183	0.025
1.6 Response efficiency	0.192	0.026
2. Effectiveness	0.193	—
2.1 Health behavior	0.296	0.057
2.2 Health outcome	0.336	0.065
2.3 Quality and efficiency of health services	0.368	0.071
3. Safety	0.289	—
3.1 Clinical safety	0.466	0.135
3.2 Technical safety	0.534	0.154
4. Economy	0.132	—
4.1 Patient-level economy	0.300	0.040
4.2 Institutional-level economy	0.328	0.043
4.3 Social benefit	0.372	0.049
5. Social appropriateness	0.251	—
5.1 Ethicality	0.217	0.054
5.2 Policy appropriateness	0.236	0.059
5.3 Acceptability	0.260	0.065
5.4 Feasibility	0.287	0.072

[a]Not applicable.

Discussion

In this study, by integrating literature and policy research with the Delphi expert consultation and AHP methods, an HRS evaluation index system was constructed based on the HTA framework. This system encompasses 5 dimensions—performance, effectiveness, safety, economy, and social appropriateness—comprising a total of 18 secondary assessment indicators.

Interpretation of Results and Comparison With Previous Research

Existing evaluations of HRSs in most studies predominantly focus on accuracy metrics within the performance dimension, with some studies exhibiting nonstandardized indicator selection and a general lack of user participation in the evaluation process [19]. Even in the few HRS studies that do incorporate user involvement, such participation is typically restricted to assessments of clinical effectiveness, feasibility, or user experience [4]. While these dimensions are important, they alone do not capture the broader clinical, social, and economic implications of deploying RSs in health care settings. General frameworks exist for health IT; however, current evaluation tools tend to be narrowly focused on specific dimensions, such as the user-centric framework proposed by Knijnenburg et al [30]. Some evaluation tools have been developed for DHTs [31-33], such as the Digital Health Technology Assessment developed through the cooperation between the Finnish Coordinating Center for Health Technology Assessment and

XSL•FO
RenderX

the University of Oulu's Faculty of Medicine [33]. This framework was designed to conduct evidence-based reviews of DHTs focusing on dimensions such as product information, technical stability, cost, effectiveness, clinical safety, data security, usability, accessibility, interoperability, and patient and organizational considerations. However, it does not fully address legal, social, and ethical aspects.

Compared with previous research, this work expands the evaluative perspective of HRSs from performance-focused assessment to a holistic, multicriteria framework informed by health technology policy and real-world implementation factors. In terms of core dimensions, our framework aligns with most of the aforementioned evaluation frameworks by addressing safety; effectiveness; economic aspects; organizational implications; and sociocultural, ethical, and legal considerations [31,33]. To address the lack of technical granularity and insufficient practical guidance in existing evaluations of HRSs, we propose an evaluation index system with 2 levels (comprising primary and secondary indicators), providing more specialized and refined assessment support. For example, at the performance level, it specifically incorporates key technical indicators such as recommendation accuracy, content coverage, result diversity, user trust, system robustness, and system response efficiency, tailored to meet the specialized assessment needs of HRSs with regard to algorithmic quality and service efficiency [34]. More importantly, this secondary evaluation index system serves not only as a tool for periodic assessment but also as a driver for the continuous iteration and optimization of HRSs. By integrating quantitative performance data, user feedback, and multidimensional evidence from real-world applications, the evaluation process establishes a closed-loop feedback mechanism.

Quantitative assessment of system performance can identify shortcomings in recommendation accuracy, response speed, coverage, and diversity, thereby providing clear directions and evidence for subsequent technological iterations and optimizations [35]. HRSs are capable of conducting precise analyses based on users' physiological, psychological, and behavioral data, offering tailored health advice and action plans to facilitate refined health management [30]. However, erroneous recommendations may lead to adverse outcomes, posing risks to patient safety. Therefore, ensuring the effectiveness and safety of recommendation results is crucial for broader promotion and application. Specifically, it is recommended to assess health behaviors and behavioral outcomes through user deployment in real-world application. For example, Bidargaddi et al [17] reported findings from a randomized controlled trial evaluating the efficacy of a guided recommendation service for readily available mobile mental health apps targeting young people. Furthermore, the economic evaluation of HRSs is of paramount importance [36]. Given that the core aim of this technology is to enhance health care efficiency, users can participate in real-world deployment, enabling assessment of whether the system reduces unnecessary medical expenditures, improves diagnostic and treatment quality, and promotes preventive care [37,38]. Such assessments can quantify the cost-effectiveness ratio of the system, providing policymakers with decision-making evidence to determine the feasibility of large-scale promotion and investment. In terms of social appropriateness, HRSs must exhibit good user acceptance, feasibility, and ease of dissemination. Therefore, it is recommended to incorporate user-centered design methods early in the development process. For example, Leng et al [39] used a convergent mixed methods approach to evaluate usability, combining quantitative and qualitative data from diverse users to identify barriers and collaboratively refine system design. Given the unique nature of HRSs as an artificial intelligence technology, the evaluation process must also address ethical issues and regulatory compliance to ensure effective promotion and widespread application [40].

The evaluation index system for HRSs holds significant practical implications for research and development institutions, individual and institutional users, and policymakers in health care management [6]. For organizations involved in the development of HRSs, the system helps development teams in prioritizing core metrics, including accuracy, diversity, real-time performance, and user satisfaction, throughout the technological iteration process [15]. By leveraging the evaluation framework to quantify product performance, the evaluation index system enables the swift identification and resolution of existing issues, thereby significantly enhancing the overall efficacy of HRSs. This provides practical and effective guidance for the design, optimization, and development of HRSs. For health care institutions, HRSs can be leveraged to better serve patients and improve service quality and patient satisfaction while simultaneously reducing medical costs and promoting the efficient allocation of health care resources [41]. This evaluation index system offers clear direction for how institutions can rationally apply and manage such systems in the future. For government agencies, authoritative evaluation results can serve as critical decision-making references when promoting HRSs [42,43]. This aids in formulating policies, regulations, and technical standards that are better aligned with practical needs and specifically targeted to real-world challenges, ensuring the healthy and orderly development of this innovative technology.

Implications for Clinical Practice and Future Research

First, given the dynamic and evolving trends in user needs and health issues, the evaluation metrics for HRSs will need to be periodically updated to accurately reflect these changes and address diverse health requirements. Second, it is essential to design differentiated evaluation content based on secondary indicators according to the characteristics of different disease profiles or user groups (eg, individuals with chronic conditions) to ensure that the system can deliver personalized recommendations that support individuals in achieving their specific health goals [44]. Finally, further exploration of the application of emerging technologies such as artificial intelligence and blockchain in the evaluation framework is necessary to investigate how to enhance assessment efficiency and accuracy, improve data transparency and security, and provide robust technical support for the continuous optimization of HRSs [45,46].

Strengths and Limitations

The high expert consensus achieved across Delphi rounds highlights the relevance and acceptability of the proposed

XSL•FO
RenderX

indicators. The use of the AHP made the index system clearer and more structured, facilitating its application by decision-makers. As HRSs continue to be integrated into clinical decision support, patient engagement platforms, and digital therapeutics, the evaluation system developed in this study provides a structured foundation for guiding system design, regulatory decision-making, and cross-study comparability.

As a relatively new research field, empirical evidence on the practical implementation of HRSs, particularly systematic deployment, remains limited [47]. In the future, based on updates from empirical research, existing evidence will be integrated to develop specific quantitative methods and evaluation criteria for the HRS evaluation index system. Additionally, the indicators will be further optimized during practical application, continuously improving the evaluation index system for HRSs. It should be noted that the current framework relies heavily on expert input from a specific regional context, primarily China, which may affect its generalizability. Differences in health care systems, cultural norms, and organizational practices across regions may influence the applicability of certain indicators in other settings.

Conclusions

This study addresses the lack of clear evaluation standards for HRSs by using rigorous scientific methodologies and a systematic development process to construct a comprehensive and structured evaluation index system. In contrast to existing studies that have largely focused on algorithmic performance, this work explicitly incorporates critical but often overlooked dimensions such as safety, economy, and social suitability. By establishing a structured and consensus-based set of criteria, this study addresses a key gap in current HRS evaluation practices and offers a practical foundation for developers, researchers, and policymakers. Future research should apply this framework when designing or deploying HRSs while continuously refining its components to meet emerging technological and ethical challenges in digital health.

Funding

This research was supported by the National Natural Science Foundation of China (72274007) and the China Postdoctoral Science Foundation (2025M772234).

Data Availability

The data supporting the findings of this study are available within the article and its supplementary materials.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Summary of the studies included in the Delphi process and the importance scores with coefficients of variation for primary indicators across both Delphi rounds.
[DOCX File, 60 KB - formative_v9i1e79997_app1.docx]

References

1. Noar SM, Benac CN, Harris MS. Does tailoring matter? Meta-analytic review of tailored print health behavior change interventions. Psychol Bull 2007 Jul;133(4):673-693. [doi: 10.1037/0033-2909.133.4.673] [Medline: 17592961]
2. Park DH, Kim HK, Choi IY, Kim JK. A literature review and classification of recommender systems research. Expert Syst Appl 2012 Sep;39(11):10059-10072. [doi: 10.1016/j.eswa.2012.02.038]
3. Cheung KL, Durusu D, Sui X, de Vries H. How recommender systems could support and enhance computer-tailored digital health programs: a scoping review. Digit Health 2019;5:2055207618824727. [doi: 10.1177/2055207618824727] [Medline: 30800414]
4. Sun Y, Zhou J, Ji M, Pei L, Wang Z. Development and evaluation of health recommender systems: systematic scoping review and evidence mapping. J Med Internet Res 2023 Jan 19;25:e38184. [doi: 10.2196/38184] [Medline: 36656630]
5. Cai Y, Yu F, Kumar M, Gladney R, Mostafa J. Health recommender systems development, usage, and evaluation from 2010 to 2022: a scoping review. Int J Environ Res Public Health 2022 Nov 16;19(22):15115. [doi: 10.3390/ijerph192215115] [Medline: 36429832]
6. Ananthakrishnan A, Milne-Ives M, Cong C, et al. The evaluation of health recommender systems: a scoping review. Int J Med Inform 2025 Mar;195:105697. [doi: 10.1016/j.ijmedinf.2024.105697] [Medline: 39608231]
7. He Y, Chen H, Xiang P, et al. Establishing an evaluation indicator system for user satisfaction with hypertension management apps: combining user-generated content and analytic hierarchy process. J Med Internet Res 2024 Sep 3;26:e60773. [doi: 10.2196/60773] [Medline: 39226103]
8. Moshi MR, Tooher R, Merlin T. Suitability of current evaluation frameworks for use in the health technology assessment of mobile medical applications: a systematic review. Int J Technol Assess Health Care 2018 Jan;34(5):464-475. [doi: 10.1017/S026646231800051X] [Medline: 30201060]

XSL•FO
RenderX

9. Unsworth H, Dillon B, Collinson L, et al. The NICE evidence standards framework for digital health and care technologies - developing and maintaining an innovative evidence framework with global impact. Digit Health 2021;7:20552076211018617. [doi: 10.1177/20552076211018617] [Medline: 34249371]
10. Digital technology assessment criteria (DTAC). National Health Service England. URL: https://www.nhsx.nhs.uk/key-tools-and-info/digital-technology-assessment-criteria-dtac [accessed 2025-11-28]
11. Vis C, Bührmann L, Riper H, Ossebaard HC. Health technology assessment frameworks for eHealth: a systematic review. Int J Technol Assess Health Care 2020 Jun;36(3):204-216. [doi: 10.1017/S026646232000015X] [Medline: 32297588]
12. Haverinen J, Keränen N, Falkenbach P, Maijala A, Kolehmainen T, Reponen J. Digi-HTA: health technology assessment framework for digital healthcare services. Finn J eHealth eWelfare 2019;11(4). [doi: 10.23996/fjhw.82538]
13. Etemadi M, Bazzaz Abkenar S, Ahmadzadeh A, et al. A systematic review of healthcare recommender systems: open issues, challenges, and techniques. Expert Syst Appl 2023 Mar;213:118823. [doi: 10.1016/j.eswa.2022.118823]
14. Abhari S, Safdari R, Azadbakht L, et al. A systematic review of nutrition recommendation systems: with focus on technical aspects. J Biomed Phys Eng 2019 Dec;9(6):591-602. [doi: 10.31661/jbpe.v0i0.1248] [Medline: 32039089]
15. Isinkaye FO, Folajimi YO, Ojokoh BA. Recommendation systems: principles, methods and evaluation. Egypt Inform J 2015 Nov;16(3):261-273. [doi: 10.1016/j.eij.2015.06.005]
16. Mustaqeem A, Anwar SM, Khan AR, Majid M. A statistical analysis based recommender model for heart disease patients. Int J Med Inform 2017 Dec;108:134-145. [doi: 10.1016/j.ijmedinf.2017.10.008] [Medline: 29132619]
17. Bidargaddi N, Musiat P, Winsall M, et al. Efficacy of a web-based guided recommendation service for a curated list of readily available mental health and well-being mobile apps for young people: randomized controlled trial. J Med Internet Res 2017 May 12;19(5):e141. [doi: 10.2196/jmir.6775] [Medline: 28500020]
18. Wang Z, Huang H, Cui L, et al. Using natural language processing techniques to provide personalized educational materials for chronic disease patients in China: development and assessment of a knowledge-based health recommender system. JMIR Med Inform 2020 Apr 23;8(4):e17642. [doi: 10.2196/17642] [Medline: 32324148]
19. De Croon R, Van Houdt L, Htun NN, Štiglic G, Vanden Abeele V, Verbert K. Health recommender systems: systematic review. J Med Internet Res 2021 Jun 29;23(6):e18035. [doi: 10.2196/18035] [Medline: 34185014]
20. Torrent-Fontbona F, Lopez B. Personalized adaptive CBR bolus recommender system for type 1 diabetes. IEEE J Biomed Health Inform 2019 Jan;23(1):387-394. [doi: 10.1109/JBHI.2018.2813424] [Medline: 29994082]
21. Cho JH, Sondhi P, Zhai C, Schatz BR. Resolving healthcare forum posts via similar thread retrieval. Presented at: Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics; Sep 20-23, 2014. [doi: 10.1145/2649387.2649399]
22. Sumsion T. The Delphi technique: an adaptive research tool. Br J Occup Therapy 1998 Apr;61(4):153-156. [doi: 10.1177/030802269806100403]
23. Brady SR. Utilizing and adapting the Delphi method for use in qualitative research. Int J Qual Methods 2015 Dec 9;14(5). [doi: 10.1177/1609406915621381]
24. Wei Y, An Y, Cao Q, Feng W, Wang D, Zhu D. Developing a professional competency framework for general practitioners in tertiary hospitals in China: a modified Delphi study. BMJ Open 2025 Mar 4;15(3):e082736. [doi: 10.1136/bmjopen-2023-082736] [Medline: 40037672]
25. Jünger S, Payne SA, Brine J, Radbruch L, Brearley SG. Guidance on Conducting and REporting DElphi Studies (CREDES) in palliative care: recommendations based on a methodological systematic review. Palliat Med 2017 Sep;31(8):684-706. [doi: 10.1177/0269216317690685] [Medline: 28190381]
26. Saaty TL. A scaling method for priorities in hierarchical structures. J Math Psychol 1977 Jun;15(3):234-281. [doi: 10.1016/0022-2496(77)90033-5]
27. Diamond IR, Grant RC, Feldman BM, et al. Defining consensus: a systematic review recommends methodologic criteria for reporting of Delphi studies. J Clin Epidemiol 2014 Apr;67(4):401-409. [doi: 10.1016/j.jclinepi.2013.12.002] [Medline: 24581294]
28. Keeney S, Hasson F, McKenna HP. A critical review of the Delphi technique as a research methodology for nursing. Int J Nurs Stud 2001 Apr;38(2):195-200. [doi: 10.1016/s0020-7489(00)00044-4] [Medline: 11223060]
29. Saaty TL. Decision making with the analytic hierarchy process. Int J Serv Sci 2008;1(1):83. [doi: 10.1504/IJSSCI.2008.017590]
30. Knijnenburg BP, Willemsen MC, Gantner Z, Soncu H, Newell C. Explaining the user experience of recommender systems. User Model User Adap Inter 2012 Oct;22(4-5):441-504. [doi: 10.1007/s11257-011-9118-4]
31. Kidholm K, Ekeland AG, Jensen LK, et al. A model for assessment of telemedicine applications: mast. Int J Technol Assess Health Care 2012 Jan;28(1):44-51. [doi: 10.1017/S0266462311000638] [Medline: 22617736]
32. Kristensen FB, Lampe K, Wild C, Cerbo M, Goettsch W, Becla L. The HTA Core Model®-10 years of developing an international framework to share multidimensional value assessment. Value Health 2017 Feb;20(2):244-250. [doi: 10.1016/j.jval.2016.12.010] [Medline: 28237203]
33. Haverinen J, Turpeinen M, Falkenbach P, Reponen J. Implementation of a new Digi-HTA process for digital health technologies in Finland. Int J Technol Assess Health Care 2022 Aug 19;38(1):e68. [doi: 10.1017/S0266462322000502] [Medline: 35983625]

XSL•FO
RenderX

34. Zhu YX, Lü LY. Evaluation metrics for recommender systems. J Univ Electron Sci Technol China 2012;41(2):163-175. [doi: 10.3969/j.issn.1001-0548.2012.02.001]
35. Wiesner M, Pfeifer D. Health recommender systems: concepts, requirements, technical basics and challenges. Int J Environ Res Public Health 2014 Mar 3;11(3):2580-2607. [doi: 10.3390/ijerph110302580] [Medline: 24595212]
36. Gomes M, Murray E, Raftery J. Economic evaluation of digital health interventions: methodological issues and recommendations for practice. Pharmacoeconomics 2022 Apr;40(4):367-378. [doi: 10.1007/s40273-022-01130-0] [Medline: 35132606]
37. Halvorsen N, Mori Y. Cost-effectiveness for artificial intelligence in colonoscopy. Gastrointest Endosc Clin N Am 2025 Apr;35(2):401-405. [doi: 10.1016/j.giec.2024.10.008] [Medline: 40021236]
38. Morrell CJ, Sutcliffe P, Booth A, et al. A systematic review, evidence synthesis and meta-analysis of quantitative and qualitative studies evaluating the clinical effectiveness, the cost-effectiveness, safety and acceptability of interventions to prevent postnatal depression. Health Technol Assess 2016 May;20(37):1-414. [doi: 10.3310/hta20370] [Medline: 27184772]
39. Leng M, Sun Y, Li C, Han S, Wang Z. Usability evaluation of a knowledge graph-based dementia care intelligent recommender system: mixed methods study. J Med Internet Res 2023 Sep 26;25:e45788. [doi: 10.2196/45788] [Medline: 37751241]
40. Vervoort D, Tam DY, Wijeysundera HC. Health technology assessment for cardiovascular digital health technologies and artificial intelligence: why is it different? Can J Cardiol 2022 Feb;38(2):259-266. [doi: 10.1016/j.cjca.2021.08.015] [Medline: 34461229]
41. Hors-Fraile S, Rivera-Romero O, Schneider F, et al. Analyzing recommender systems for health promotion using a multidisciplinary taxonomy: a scoping review. Int J Med Inform 2018 Jun;114:143-155. [doi: 10.1016/j.ijmedinf.2017.12.018] [Medline: 29331276]
42. Alotaibi YK, Federico F. The impact of health information technology on patient safety. Saudi Med J 2017 Dec;38(12):1173-1180. [doi: 10.15537/smj.2017.12.20631] [Medline: 29209664]
43. Kühne F, Schomaker M, Stojkov I, et al. Causal evidence in health decision making: methodological approaches of causal inference and health decision science. Ger Med Sci 2022;20:Doc12. [doi: 10.3205/000314] [Medline: 36742460]
44. Felfernig A, Wundara M, Tran TNT, et al. Recommender systems for sustainability: overview and research issues. Front Big Data 2023;6:1284511. [doi: 10.3389/fdata.2023.1284511] [Medline: 37965497]
45. Agbo CC, Mahmoud QH, Eklund JM. Blockchain technology in healthcare: a systematic review. Healthcare (Basel) 2019 Apr 4;7(2):56. [doi: 10.3390/healthcare7020056] [Medline: 30987333]
46. Dubovitskaya A, Novotny P, Xu Z, Wang F. Applications of blockchain technology for data-sharing in oncology: results from a systematic literature review. Oncology (Williston Park, NY) 2020;98(6):403-411. [doi: 10.1159/000504325] [Medline: 31794967]
47. Ngiam KY, Khor IW. Big data and machine learning algorithms for health-care delivery. Lancet Oncol 2019 May;20(5):e262-e273. [doi: 10.1016/S1470-2045(19)30149-4] [Medline: 31044724]

Abbreviations

AHP: analytic hierarchy process
Ca: expert judgment coefficient
Cr: expert authority coefficient
Cs: expert familiarity coefficient
CV: coefficient of variation
DHT: digital health technology
HRS: health recommender system
HTA: health technology assessment
RS: recommender system

Edited by S Brini; submitted 03.07.25; peer-reviewed by IU Bhuiyan, M Liu; revised version received 17.11.25; accepted 17.11.25; published 22.12.25.

Please cite as:
Sun Y, Hou S, Chen S, Leng M, Wang Z
Development of an Evaluation Index System for Health Recommender Systems Based on the Health Technology Assessment Framework: Cross-Sectional Delphi Study
JMIR Form Res 2025;9:e79997
URL: https://formative.jmir.org/2025/1/e79997
doi:10.2196/79997

XSL•FO
RenderX

© Yue Sun, Shijie Hou, Siye Chen, Minmin Leng, Zhiwen Wang. Originally published in JMIR Formative Research (https://formative.jmir.org), 22.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Radiology Staff Experiences With Integrating Artificial Intelligence Into Radiology Practice in a Swedish Hospital: Qualitative Case Study

Per Nilsen[1,2], PhD; Petra Svedberg[1], PhD; Ingrid Larsson[1], PhD; Lena Petersson[1], PhD; Jens Nygren[1], PhD; Emilie Steerling[1], MS; Margit Neher[1], PhD

[1]School of Health and Welfare, Halmstad University, Box 823 Halmstad, Halmstad, Sweden

[2]Department of Health, Medicine and Caring Sciences, Linköping University, Linköping, Sweden

Corresponding Author:
Per Nilsen, PhD
School of Health and Welfare, Halmstad University, Box 823 Halmstad, Halmstad, Sweden

Abstract

Background: The integration of artificial intelligence (AI) in radiology has advanced significantly, but research on how it affects the daily work of radiology staff is limited.

Objective: This study aimed to explore the experiences of radiology staff on the integration of an AI application in a radiology department in Sweden. This understanding is essential for developing strategies to address potential challenges in AI integration and optimize the use of AI applications in radiology practice.

Methods: This qualitative case study was conducted in a single radiology department with 40 employees in 1 hospital in southwestern Sweden. The study concerned the integration of an AI-powered medical imaging software designed to assist radiologists in analyzing and interpreting medical images. Using a qualitative design, interviews were conducted with 7 radiologists (physicians), 4 radiologic technologists, and 1 physician assistant. Their experience within radiology varied between 13 months and 38 years. The data were analyzed using qualitative content analysis.

Results: Participants cited numerous strengths and advantages of significant value in integrating AI into radiology practice. Numerous challenges were also revealed in terms of difficulties associated with choosing, acquiring, and deploying the AI application and operational issues in radiology practice. They discussed experiences with diverse strategies to facilitate the integration of AI in radiology and to address various challenges and problems.

Conclusions: The findings illustrate how AI integration was experienced in 1 hospital. While not generalizable, the study provides insights that may be useful for similar settings. Radiology staff believed AI integration enhanced decision-making and quality of care, but they encountered challenges from preadoption to routine use of AI in radiology practice. Strategies such as internal training and workflow adaptation facilitated the successful integration of AI in radiology.

(JMIR Form Res 2025;9:e77843) doi:10.2196/77843

KEYWORDS

radiology; health care; staff; artificial intelligence; AI; integration; case study

Introduction

The integration of artificial intelligence (AI) applications into radiology practice has advanced significantly, offering promising solutions to enhance diagnostic accuracy, improve workflow efficiency, and, ultimately, optimize patient care [1]. Traditionally, radiologists (ie, medical doctors specialized in diagnosing and treating injuries and diseases using medical imaging procedures) have relied on their expertise and visual interpretation skills to analyze medical images. However, the increasing complexity and volume of imaging studies, coupled with increasing demands for faster and more accurate diagnoses, have underscored the need for innovative solutions to support radiology staff [2-4].

AI in radiology practice has a wide range of applications, from image acquisition, processing, and interpretation to aided reporting, follow-up, planning, data storage, and data mining [5]. Enhanced diagnostic efficiency has been demonstrated with AI applications within radiology across various imaging modalities, including radiography, computed tomography (CT), magnetic resonance imaging (MRI), and mammography [1,6]. Furthermore, AI applications have been shown to streamline radiology workflow by automating routine tasks, improving productivity, and reducing turnaround times for image interpretation and reporting [7-9].

XSL•FO
RenderX

The impact of AI applications in radiology has been discussed extensively in opinion articles [10-16], and various practical use cases have been described in the literature [17-22]. There are also clinical trials assessing the efficiency of AI applications in improving diagnostic accuracy [1,23,24]. In addition, surveys have explored radiologists' perceptions and expectations regarding AI use in their field [25-30].

However, there is a gap in research addressing how real-world use of AI applications in radiology practice affects the daily work of radiology staff. Studies are also lacking with regard to procuring and deploying AI applications, as well as firsthand experiences of staff in radiology departments. Thus, there are gaps in understanding the practical implications of AI integration in this field.

Given that little is known about the everyday use of AI in radiology, a case study approach allows an in-depth exploration of how integration was experienced in a particular organizational context. The aim of this study was to explore the experiences of radiology staff on the integration of an AI application in a radiology department in Sweden. The study sought to provide insights into the real-world application of AI in radiology, with a particular focus on understanding the practical implications of AI integration based on the experiences of staff in a radiology department. This understanding is essential for developing strategies to address potential challenges in AI integration and optimize the use of AI applications in radiology practice.

Methods

Study Design and Setting

We adopted a qualitative case study design [31], using an inductive qualitative content analysis [32] to examine the integration of an AI application in a radiology department of 1 Swedish hospital. The study adhered to the COREQ (Consolidated Criteria for Reporting Qualitative Research) to ensure trustworthiness [33].

The study setting was 1 radiology department (with 40 employees) in 1 hospital with approximately 4000 employees in southwestern Sweden. Radiology departments in Swedish hospitals typically serve primary care and other wards within the hospitals, including emergency departments. This hospital was selected due to its active integration of AI applications in radiology, providing a relevant and practical setting for exploring real-world experiences with AI integration. The radiology department had prior exposure to AI-assisted workflows, making it a suitable case for examining both the benefits and challenges of implementation.

Health care in Sweden is organized at the regional level. Each of the country's 21 regions is responsible for providing health care services to its residents. The regions are responsible for managing and funding health care services, including hospitals, primary care, and specialized health care services. The health care system is funded through taxes and is primarily publicly funded; private health care providers also play a role. Patients have the right to choose their health care service provider and have access to a wide range of services, including preventive care, medical treatment, and rehabilitation services.

The AI Application

The study examined the integration of AI-Doc, an advanced AI-powered application specifically designed to enhance radiology practice by augmenting the capabilities of radiology staff, improving diagnostic precision, and optimizing workflow efficiency. AI-Doc leverages machine learning algorithms and deep learning techniques to process and analyze a diverse range of imaging modalities, including radiographs, CT scans, MRI, and ultrasound images.

The application autonomously identifies anomalies, highlights areas of potential concern, and generates preliminary diagnostic insights. Acting as an "intelligent assistant," AI-Doc reduces cognitive strain on radiologists while promoting greater diagnostic consistency and accuracy across cases. Its functionality includes customizable alert systems, integration with picture archiving and communication systems, and a user-friendly interface designed to blend into radiologists' existing workflows.

AI-Doc has seen widespread adoption in Swedish health care systems, where it serves as a practical, real-world example of how AI technology can be integrated into modern radiology. The application has been described and methodically evaluated in the work of Wiklund and Medson [34], where its impact on clinical workflows and diagnostic outcomes was further assessed.

Study Participants and Recruitment

The practice context for using the AI application was identified through the network of the research group. The head of the radiology department at the hospital was contacted by the research team and informed about the study. The head assisted by supplying the email addresses of relevant employees in the radiology department. Eligibility criteria were that participants were radiology staff who were directly engaged in the work with the AI application and the work routine around the application at the radiology department.

Participants signed up by responding to an email invitation with detailed information about the study. Two reminders were sent at 2-week intervals. The invitation message clearly stated that participation was voluntary and that participants could withdraw from the study at any time and without explanation.

Participants were recruited from diverse professional groups to achieve a varied sample and a broad perspective on the integration of the AI application. The study population consisted of 12 staff members from the radiology department: 7 radiologists (physicians), 4 radiologic technologists (referred to as radiographers outside North America), and 1 physician assistant. Seven were men, and 5 were women. Their experience within radiology varied between 13 months and 38 years.

Participants were selected based on their direct engagement with the AI application and their involvement in related workflows within the radiology department. The study aimed to capture a diverse range of perspectives, which is why radiologists, radiologic technologists, and a physician assistant were included. Radiologists were selected for their expertise in interpreting medical images and their decision-making role in

XSL•FO
RenderX

integrating AI into clinical practice. Radiologic technologists were included due to their hands-on role in image acquisition and processing. The physician assistant was selected to provide insight into how AI influences patient interactions and radiology workflow coordination. This diverse sample allowed for a comprehensive exploration of AI integration from multiple professional viewpoints.

Data Collection

Semistructured interviews were conducted online via Zoom or Teams by 3 of the co-authors (ES, MN, and LP) between September 2023 and May 2024. Two pilot interviews were performed to test the interview guide. No adjustments to the questions were made, and these 2 interviews were therefore included in the study. The authors had no preexisting relationship with any of the participants. Before starting the interviews, the participants were asked if they had read the information about the study and to give their written informed consent to participate.

An interview guide, consisting of open-ended questions, directed the course of the interviews. The interview guide included questions on the participants' experiences regarding the integration of the AI application, encompassing everything from selecting and purchasing the application to using it in routine practice. Questions also addressed the impact of the AI application on professional work roles and participants' trust in the AI application. Probing questions were asked during the interviews to clarify the description or draw attention back to the topic; for example, "What do you mean?" and "Can you explain this a little further?"

The interview guide ensured that all participants were asked similar core questions while allowing flexibility for follow-up prompts to capture nuanced insights. The research team maintained regular discussions throughout the data collection process to align their approach and ensure uniformity in questioning. Furthermore, investigator triangulation was applied during data analysis, with multiple researchers independently reviewing transcripts to enhance reliability and minimize individual biases.

Interviews lasted for an average of 45 minutes (SD 8.02, ranging from 24 to 55 min). Interview data were audio recorded and transcribed verbatim.

Data Analysis

The data were analyzed using qualitative content analysis, following the approach for conventional content analysis as outlined by Hsieh and Shannon [32]. This data-driven analysis focused on the participants' unique experiences, rather than being guided by a preexisting theory or hypothesis. Investigator triangulation was used to enhance the validity of the findings. Three of the researchers (PN, MN, and ES) independently reviewed and reread the transcripts to develop a comprehensive understanding of the content and gain an overview of the material.

With the study's aim in focus, the first author (PN) highlighted relevant text and made margin notes to capture all key aspects of the content. Initial thoughts and impressions were documented during this process. He then generated codes from the data to reflect the most significant content. Related codes were then grouped and organized into subcategories and broader categories. These categories and subcategories were then presented to the research team, which led to an iterative process that involved continually revisiting and refining the codes, ensuring they were consistent with the full scope of the material.

The subcategories and categories were then systematically compared for differences and similarities to ensure they were as internally consistent and externally distinct as possible. Labels were assigned to each category and subcategory, and numerous revisions were made to ensure they reflected the content accurately. The names and content of these labels were discussed thoroughly among the research team, and adjustments were made until consensus was reached. In the next step, the first author (PN) selected representative quotes from the interviews to illustrate each subcategory. After the team discussion, the chosen quotes were translated into English by PN.

Ethical Considerations

Ethical approval was granted by the Swedish Ethical Review Authority (no. 2023‑05318), and the study was conducted according to the World Medical Association Declaration of Helsinki ethical principles for medical research involving human subjects [35]. Participants were assured that their data and their personal integrity would be safeguarded throughout the research process and that they would not be individually recognizable in the final scientific publication. They were asked to sign a formal consent document before the interview and informed that they were free to withdraw at any time, no questions asked. Participants consented to the publication of the quotes.

Results

Overview

Three categories and 9 subcategories emerged from the analysis of the interview data (Figure 1). These categories and subcategories encapsulate the participants' experiences associated with integrating the AI application into radiology practice.

XSL•FO
RenderX

Figure 1. Three categories and 9 subcategories encapsulate the participants' experiences associated with integrating the AI application into radiology practice. AI: artificial intelligence.

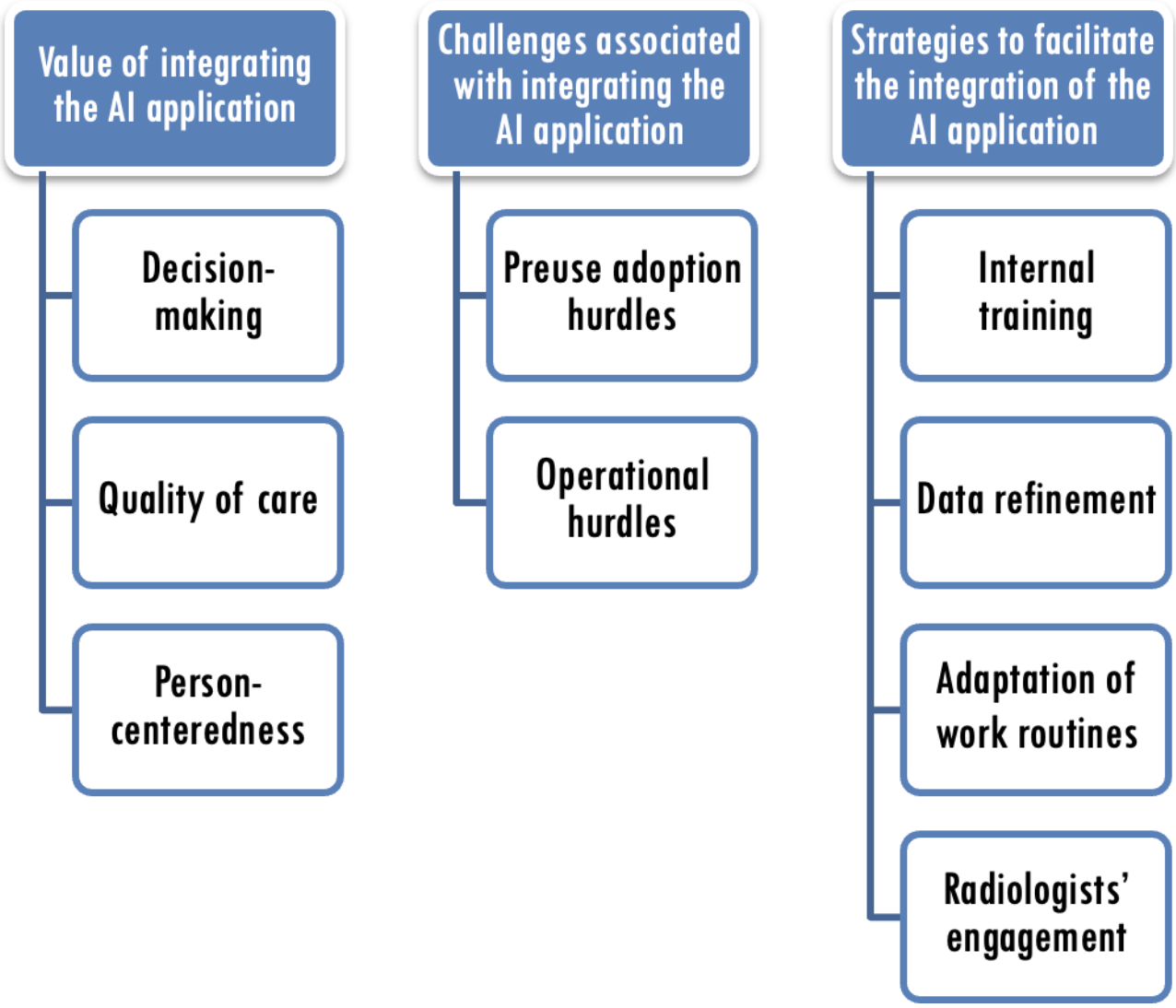

Value of Integrating the AI Application Into Radiology Practice

Overview

Participants mentioned several areas where they experienced significant value in using AI in radiology practice, citing numerous strengths and advantages they experienced. This value proposition was delineated into three subcategories: decision-making, quality of care, and person-centeredness.

Decision-Making

Participants experienced that the AI application augmented their decision-making capabilities by providing better image analysis because the application identified details that were invisible to the human eye. This was experienced as facilitating more accurate diagnoses and fostering confidence in clinical decisions.

> *The AI system relieves us as radiologists from something that we find boring and time-consuming. And which we might also feel a little unsure about because it's easy to miss something.* [#6, radiologist]

> *The AI system has very high accuracy. It is very rare for it to flag a false finding.* [#8, radiologic technologist]

The AI application was particularly important for junior radiologists who were less experienced in analyzing images. Participants believed the AI application refined their skills because it provided detailed analyses of imaging scans. This helped them identify subtle patterns they might otherwise have missed. They experienced that this exposure sharpened their diagnostic skills.

> *The AI system has the potential to save some time and to make one feel more confident in the response, especially as a newcomer in the profession.* [#4, radiologist]

Quality of Care

Participants also reported experiences of how the integration of the AI application into radiology practice could benefit the overall quality of care. They experienced that the AI application promoted consistency by applying uniform diagnostic criteria across all cases, reducing variability in image interpretation and supporting standardized decision-making for more reliable

XSL•FO
RenderX

diagnoses. This, in turn, enhanced equity in health care delivery by ensuring that all patients received the same high-quality diagnostic care, regardless of geographic location, hospital resources, or clinician expertise.

> *Knowledge will be more equitable with AI.* [#1, radiologist]
>
> *The AI system will primarily increase accessibility [to healthcare].* [#8, radiologic technologist]
>
> *There will be a more standardized way of working with AI.* [#9, radiologic technologist]

Participants lauded the AI application for its effectiveness in detecting incidental findings because it assisted them in identifying anomalies. This capability enabled early intervention, potentially leading to better patient outcomes.

> *And the AI system really removes the risk for patients of developing symptoms from these clots.* [#11, radiologic technologist]
>
> *Lung embolisms are detected that one might not know about if they hadn't been reviewed using the AI system.* [#12, physician assistant]

Person-Centeredness

Participants experienced that the integration of the AI application into radiology practice had a positive impact on person-centeredness in terms of considering the whole person and accounting for the individual needs and preferences of patients. They highlighted enhanced accessibility for patients to radiologic examinations and diagnostic procedures. Participants emphasized that the AI application shortened the duration of patients' health care interactions by streamlining the examinations and reducing wait times for diagnostic results.

> *With these image enhancements, the patient is exposed to a lower dose of radiation. … The patient receives a faster diagnosis.* [#2, radiologist]
>
> *They can receive a response immediately. They might live in Ullared, and previously they would go home and then be called back here again. This improves things for our patients in that way, it really does.* [#12, physician assistant]

In addition, participants experienced that the AI application played a vital role in reassuring many patients. The application generated images that patients could access after their examination, fostering trust in the technology to minimize errors and enhance the accuracy of their diagnoses.

> *Many patients are not satisfied with the doctor saying that this is a herniated disc and you just have to go to rehab. Instead, they want an image of things. They want proof.* [#8, radiologic technologist]

Challenges Associated With Integrating the AI Application Into Radiology Practice

Overview

Participants experienced numerous challenges in terms of difficulties or problems associated with choosing, acquiring, and deploying the AI application and its subsequent use in radiology practice. These challenges were classified into two subcategories: preuse adoption hurdles and operational hurdles.

Preuse Adoption Hurdles

Participants described several types of challenges associated with procuring and setting up the AI application in radiology practice. These challenges were experienced in the preuse adoption phase, that is, before the application was integrated into radiology practice. They underscored the significant financial investment they associated with AI applications.

> *These are quite expensive things. We're talking about millions, of course, for licenses and hardware and so on.* [#9, radiologic technologist]

Navigating the crowded marketplace of AI applications posed a considerable challenge in identifying and selecting the optimal solution for radiology departments.

> *There are hundreds of products on the market.* [#2, radiologist]

Participants complained about the protracted nature of the adoption process for the AI application, which entailed addressing numerous administrative hurdles. Furthermore, the actual deployment of the AI application was marked by technical intricacies.

> *It seems to be a huge hurdle for many places, and IT puts a stop to it. They say it's too complicated to implement an AI. It can take a year from asking questions until you get a final answer.* [#1, radiologist]
>
> *This implementation of the [AI application] takes an incredibly long time and is so slow.* [#7, radiologist]

Operational Hurdles

Participants experienced numerous practical operational difficulties that could arise in the everyday use of the AI application. They emphasized the steep learning curve that radiology staff must traverse to harness the application's full potential. Navigating technical issues inherent to the application provided a challenge.

> *Unfortunately, it quite often freezes or lags, or you get logged out. Or some sort of system error. Or some local [error].* [#3, radiologist]
>
> *It takes a long time to program. One needs to have time for testing, so it often actually takes a year before everything is set up as desired. Because there is so much to understand when you get a completely new machine.* [#8, radiologic technologist]

Participants also experienced that data management could provide obstacles for the effective integration of the AI application into radiology practice.

> [It is very important how all the data management is handled. Information is first sent to a local server where the data is anonymized before it is sent to a cloud solution, and then it bounces back. #2, radiologist]

XSL•FO
RenderX

Strategies to Facilitate the Integration of the AI Application Into Radiology Practice

Overview

Participants highlighted experiences of using diverse strategies to facilitate the integration of the AI application into radiology practice to address various challenges and problems they encountered. These strategies encompassed both specific actions tailored to support AI use and the active involvement of individuals taking initiatives to facilitate the use. Four subcategories emerged from the data analysis: internal training, data refinement, adaptation of work routines, and radiologists' engagement.

Internal Training

Participants emphasized the importance of in-house, internal training, which consisted of educational activities conducted within the health care organization's premises or by its staff. This type of training was designed to educate employees or team members about the specific AI application in radiology.

> *Training is required for all users to understand what the [AI system] is, what the tool is supposed to do, in what context it will be used, and why it exists.* [#1, radiologist]
>
> *Training sessions [are needed] as well, of course. As always when introducing new things, someone needs to present it. … present the applications, how they look, and what buttons to press.* [#6, radiologist]

The relevance of integrated training to support the integration of the AI application into radiology practice was also stressed by participants. This training could be included as part of the overall acquisition of the AI application, which meant that the health care organization received both the AI application and the necessary training to ensure that the radiology staff would have the knowledge and skills to leverage the AI application effectively.

> *You can also have someone from the company [that sold the AI system] come and help you, explaining how to use the system and if there are any pitfalls with this new technology. ... The purchase usually comes with a certain number of hours of support.* [#8, radiologic technologist]
>
> *There are many lectures [on AI] that one can attend. ... And then there was a guy who had a lecture in the main hall about AI, so there is a great interest.* [#12, physician assistant]

More informal training also occurred in the use of the AI application through everyday conversations and discussions with colleagues who helped each other.

> *I do my review in parallel [with the AI system], and if it matches, you gain trust.* [#3, radiologist]
>
> *Well, there were colleagues who, in situations [when problems arise], when it... I mean, there is so much to learn. ... Then it becomes a brief discussion with a specialist who has more experience regarding how it works.* [#7, radiologist]

Data Refinement

The importance of ensuring local validation and adaptation of data was emphasized by participants for the integration of the AI application into radiology practice. These were iterative processes intended to improve and optimize the AI application to ensure its accuracy, reliability, and effectiveness. Validation was a process of verifying the performance and accuracy of the AI application using data that were representative of the local environment.

> *Even before the AI system was implemented, a validation study was conducted where they reviewed a hundred or up to a thousand cases to see how the AI solution performed on our patient population, our machines, our settings, and so on. It is important to carry out such a validation process.* [#2, radiologist]
>
> *We placed quite a bit of emphasis on the configuration from the start, ensuring that the AI system would assist us; we were careful to get the setup right from the very beginning.* [#10, radiologic technologist]

Adaptation of Work Routines

Some changes in the radiology staff's current work processes and patient pathways were necessary to facilitate the integration of the AI application into radiology practice. This involved partial adjustment of both radiologists' and radiologic technologists' work routines to align with a revised sequence of actions for radiology patients. The participants experienced these modifications as essential to enable patients to leverage the advantages of the AI application fully.

> *We had to find some form of structure for who would be responsible when a study needs to be followed up. ... The AI system requires some adjustments to the routine.* [#5, radiologist]
>
> *The working method has changed. Patients stay a bit longer and receive information before they leave. The new routine gives [us radiologic technologists] more contact with the radiologists.* [#11, radiologic technologist]

Radiologists' Engagement

Individual radiologists played a significant role by functioning as influential change agents to facilitate the use of the AI application in radiology. Participants experienced that these individuals advocated for the benefits of AI and communicated its value to colleagues and managers in radiology and the wider health care organization.

> *It is not something that has come from outside; rather, the AI issue is something that we in the medical group have driven. We are positive, and these are our tools.* [#6, radiologist]
>
> *There was a sense of security from the very beginning because doctors were involved in the implementation of the AI system.* [#11, radiologic technologist]

The radiologists addressed resistance and skepticism among their colleagues toward AI by engaging in discussions about the application and addressing various concerns. Furthermore,

XSL•FO
RenderX

they organized training and helped colleagues understand how the AI application worked and how to use it most effectively.

> *I was one of the "super users" because I work with this daily. I was present to answer questions so that we could have ongoing discussions about whether there was anything that needed to change in how the tool looked or was presented. Then we could communicate that to the AI company.* [#1, radiologist]

> *We have a doctor who is quite involved in this. He has also given lectures on it. He has promoted the AI system to his colleagues and talked about it.* [#10, radiologic technologist]

Discussion

Principal Findings

On the basis of interviews with radiologists, radiologic technologists, and a physician assistant in 1 radiology department in a Swedish hospital, this qualitative case study sought to explore experiences concerning the integration of an AI application into radiology practice, from purchasing to using the app. The participants reported that the integration brought significant benefits, particularly in terms of improved diagnostic accuracy, quality of care, and person-centeredness. However, they also identified challenges they experienced, both preuse adoption hurdles and operational hurdles. Despite these challenges, the participants experienced that strategies such as internal training, data refinement, adaptation of work routines, and radiologists' engagement could facilitate the successful integration of AI into radiology practice. While the findings are not intended to be generalizable, they offer transferable insights that may inform AI integration efforts in similar health care contexts.

The participants experienced numerous benefits that demonstrated the substantial value of integrating the AI application into their radiology practice, citing advantages for themselves, the health care system, and patient care. Compared with other areas of health care, where AI integration has been slower, AI in radiology is well-established, optimized by the extensive digitization of imaging data that has been underway since the late 20th century [30]. This transition began in the 1970s with the introduction of CT scans, which used x-rays and computer processing to produce digital images. The advent of MRI in the 1980s further advanced this digital shift, and by the 1990s, most radiology departments had moved from film-based to digital imaging. This evolution has made AI a mature and indispensable technology in radiology, proving its significant value in the field.

Although the significance of integrating AI applications into radiology is widely recognized, research has shed light on numerous challenges associated with this use, many of which we also observed. Previous studies on AI use in radiology [28,36,37] have identified technical obstacles related to data quality, standardization, transparency, interpretability, and regulatory compliance. Ethical concerns concerning data privacy and security are also critical factors that need to be addressed to promote the ethical and responsible use of AI in radiology [19,38,39]. While our findings resonate with the broader literature, they reflect the specific circumstances of 1 radiology department in a Swedish hospital and thus should be understood as context bound.

The participants in our study highlighted many strategies to facilitate the successful integration of the AI application into radiology practice. It required both formal learning based on organized training and informal learning that was more spontaneous. Informal learning refers to the acquisition of knowledge and skills that takes place outside of a structured, formal educational environment [40]. This learning in our study seemed to happen through everyday interactions and activities, which is typical for informal learning because it occurs without a specific curriculum or defined objectives. Informal learning is typically self-directed, voluntary, and motivated by the learner's interests or needs [41]. In our study, this learning was driven by the radiology staff's ambition to solve problems that occurred when using the application. Learning by doing seemed to be very important for the participants in our study. In this form of learning, individuals acquire knowledge and skills through direct experience and active participation in real-life tasks or activities [42-44].

Successful integration of the AI application into radiology practice in our study also depended on local validation and adaptation of data as well as adaptation of existing routines around the application. The findings suggest that the use of AI in radiology can have ripple effects, that is, indirect or unintended consequences that result from introducing a new practice within a health care system. Participants described that they had established new workflows and patient pathways. Ripple effects can have both positive and negative impacts on various aspects of health care delivery and patient care [45]. In our study, these effects were largely experienced as positive, notably by shortening the duration of patients' health care interactions through streamlined examinations and reduced wait times for diagnostic results. Ripple effects are often underestimated when implementing new practices in health care due to the complexity and interconnectedness of health care systems. Various stakeholders, structures, and processes can be affected by changes, making it challenging to anticipate all potential impacts [46,47].

The importance of engagement by individual radiologists throughout the process of procuring, deploying, and using the AI application in radiology was evident from the interviews in our study. Radiologists are usually highly respected and influential within health care; colleagues and patients typically trust radiologists due to their extensive education, training, and experience [48]. Our findings showed that radiologists' recommendations and opinions regarding the AI application held significant weight. Some radiologists essentially functioned as change agents as they promoted the use of the application to colleagues. Although there are several terms for change agents in the literature, including champions, facilitators, and super users, there is consensus that change agents can play a crucial role in influencing others to adopt and implement new ideas, products, or practices by providing information, support, and encouragement [49,50].

XSL•FO
RenderX

Despite the positive experiences of the participants in our study concerning the integration of the AI application into radiology practice, research has revealed apprehension among radiologists about the automation of tasks traditionally performed by human experts [28,51-53]. Uncertainty about the future of the profession has been observed to have an impact on radiology students, who are hesitant about committing to a career path that seems susceptible to technological disruption [14]. There is also a debate that many traditional tasks will increasingly be taken over by AI, with the result that radiology staff (and health care staff in general) will need to spend less time on routine tasks, which may result in a decline in some clinical skills [54]. Deskilling is not a new phenomenon; however, technological advancements have reduced skill requirements for many jobs in various societal sectors [55]. The question is how or to what extent it may occur in radiology, a topic that warrants further investigation.

Another contemporary debate about AI is its "black box" nature, which refers to the difficulty in interpreting or explaining the reasoning behind specific outputs due to the inherent complexity of AI applications, both in how they are developed and trained and in how they are applied [56]. However, in our study, this appeared to be relatively unimportant for the participants, who displayed a high level of trust in the AI application they used. Their trust seemed to be built up over time through consistent and reliable performance of the application. As they worked alongside the application, they were able to see firsthand how the technology could aid in interpreting images accurately and efficiently. Research suggests that the need for explainability diminishes when AI consistently outperforms human practitioners [57]. Thus, trust in AI appears to be based on the reliability of the outputs rather than the transparency of the processes.

Strengths and Limitations

Several limitations of this study should be acknowledged, as they may have influenced the findings and their interpretation. As a qualitative case study conducted in a single radiology department in 1 hospital in Sweden, the findings are context-specific and not intended to be statistically generalizable. Instead, the value of the study lies in offering in-depth, situated insights that may be transferable to other health care settings facing similar challenges of AI integration.

We used an inductive qualitative approach due to the limited existing knowledge about the experiences of radiologists, radiologic technologists, and other radiology staff regarding the integration of AI applications into radiology practice. We conducted interviews with radiology staff in Sweden to explore these experiences. Participation in the study was entirely voluntary, with interviewees recruited based on their expressed interest in participating. While this approach allowed us to engage individuals motivated to share their experiences, it may also have introduced a selection bias. Specifically, the recruitment strategy may have disproportionately attracted participants who held favorable views toward AI or who were particularly enthusiastic about its integration into radiology practice. As a result, the perspectives captured may lean toward more positive or optimistic attitudes, while voices of those more critical, skeptical, or less engaged with AI technologies may have been underrepresented.

While the sample included radiology staff with diverse experience levels, the relatively small number of participants may not fully capture variations in perspectives across different institutions or health care systems. However, given that many radiology procedures are standardized, the insights gained may still be relevant for hospitals in similar contexts looking to implement AI in radiology.

None of the researchers specialized in radiology; instead, we came from diverse professional backgrounds, presenting both strengths and weaknesses. Notably, none of the interviewers had a background in radiology, which may have limited the depth of probing during the interviews, potentially leading to missed opportunities for deeper exploration of technical or profession-specific issues. Although our varied perspectives could have led to some misunderstandings or missed follow-up questions, this diversity also helped us avoid preconceived notions. The credibility of the study was further enhanced by our multidisciplinary team, which included experts in behavioral economics and systems development (PN), occupational therapy (MN), nursing (IL and PS), health and innovation science (JN), pedagogy (LP), and social and behavioral science (ES). This composition provided multiple perspectives on the issue, combining clinical and scientific expertise, and was reinforced by our significant experience in qualitative research within the health care context. The absence of radiology specialists among the interviewers represents a methodological limitation, as it may have reduced the depth of technical discussion. Conversely, our disciplinary diversity functioned as a strength by promoting open inquiry and minimizing confirmation bias toward existing professional viewpoints.

A total of 12 interviews were conducted, including participants of different professions, ages, levels of experience, and genders, which provided a range of perspectives on radiology practice and AI integration. This diversity provided valuable perspectives on radiology practice. Nevertheless, as the study was conducted within a single Swedish hospital, the contextual specificity limits the transferability of the findings to other institutional or national settings. Most participants had significant hands-on experience with the AI application under study, which added practical depth to the data and supported the study's credibility [58]. Nevertheless, familiarity with the specific AI tool may have constrained the range of experiences described, as participants may have reflected primarily on 1 particular implementation rather than AI integration more broadly.

It is possible that the participants were generally positive toward AI use, as participants willing to engage in such a study may have had a particular interest in or optimism about technological innovations. This potential self-selection bias should be taken into account when interpreting the findings. While it is possible that participants who volunteered were generally open to technological innovation, the purpose of this study was not to assess attitudes toward AI but to explore professionals' experiences of working with or around AI technologies. Consequently, the focus was on descriptive accounts of practice rather than on judging AI as positive or negative.

XSL•FO
RenderX

The sample size was considered sufficient based on the concept of information power, which suggests that the more relevant and rich data a sample provides, the fewer participants are needed. Information power is influenced by factors such as the study's aim, the specificity of the sample, the quality of the interview dialogue, and the depth of analysis [59]. Previous research by Guest et al [60] indicates that 12 interviews are often adequate when participants are knowledgeable about the topic, when the aim is to explore shared experiences rather than compare subgroups, and when the interview material is of high quality. While all participants in our study had relevant radiology knowledge and the interviews adhered closely to the guide, it remains possible that a larger or more heterogeneous sample, including participants from different institutions or with varying levels of exposure to AI, could have uncovered additional themes or perspectives. Quotations from all 12 participants were incorporated into the analysis to enhance transparency and provide a fuller view of the data, but the limited sample still poses a constraint on the transferability of the findings.

Conclusions

Radiologists, radiologic technologists, and other radiology staff in 1 radiology department in a Swedish hospital experienced many benefits from integrating an AI application into radiology practice, including augmented decision-making, quality of care, and person-centeredness. They also experienced challenges, ranging from preuse adoption to operational hurdles in its routine use in radiology practice. Despite these hurdles, various strategies such as internal training, data refinement, adaptation of work routines, and radiologists' engagement facilitated the successful integration of AI into radiology practice.

Building on the findings from this case study, we suggest several contextually grounded considerations that may be useful for others seeking to integrate AI into radiology practice. In this particular hospital setting, participants emphasized the importance of structured internal training to ensure staff understood the tools, their purpose, and limitations, often combining formal education with peer learning. Local validation using historical data was seen as vital to confirm that the AI functioned accurately within the specific clinical environment. Adaptations of clinical workflows, including clear responsibilities and follow-up protocols for AI-flagged findings, were described as necessary to make the application work in practice. The role of "super users" or engaged radiologists emerged as influential in fostering staff engagement and maintaining dialogue with vendors. Reliable IT infrastructure and the early involvement of IT staff were also experienced as critical for smooth deployment. Finally, participants stressed that addressing preadoption challenges such as cost, vendor selection, and regulatory processes required proactive planning and stakeholder involvement. While these observations are drawn from a single Swedish hospital and are not intended to be generalizable, they may offer transferable insights for health care organizations navigating similar implementation processes.

Acknowledgments

This project would not have been possible without the support and contributions of many individuals. The authors would like to thank all participants for generously sharing their time and experiences.

Funding

The study was supported by funding from the Swedish Research Council (grant 2022‐05406) and internal funding provided by Halmstad University. The funders were not involved in any aspects of the study design, collection, analysis, interpretation of data, or the writing or publication design.

Data Availability

The datasets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: all authors
Study design: all authors
Data collection: ES, MN, LP
Data analysis: initiated by PN but discussed with and consolidated by all authors.
Writing – original draft: PN
Writing – review & editing: all authors
Project administration or oversight: PN
Final approval of manuscript: all authors

Conflicts of Interest

None declared.

References

XSL•FO
RenderX

1. van Leeuwen KG, de Rooij M, Schalekamp S, van Ginneken B, Rutten MJCM. How does artificial intelligence in radiology improve efficiency and health outcomes? Pediatr Radiol 2022 Oct;52(11):2087-2093. [doi: 10.1007/s00247-021-05114-8] [Medline: 34117522]
2. Najjar R. Redefining radiology: a review of artificial intelligence integration in medical imaging. Diagnostics (Basel) 2023 Aug 25;13(17):2760. [doi: 10.3390/diagnostics13172760] [Medline: 37685300]
3. Pinto-Coelho L. How artificial intelligence is shaping medical imaging technology: a survey of innovations and applications. Bioengineering (Basel) 2023 Dec 18;10(12):1435. [doi: 10.3390/bioengineering10121435] [Medline: 38136026]
4. Pierre K, Haneberg AG, Kwak S, et al. Applications of artificial intelligence in the radiology roundtrip: process streamlining, workflow optimization, and beyond. Semin Roentgenol 2023 Apr;58(2):158-169. [doi: 10.1053/j.ro.2023.02.003] [Medline: 37087136]
5. Pesapane F, Codari M, Sardanelli F. Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine. Eur Radiol Exp 2018 Oct 24;2(1):35. [doi: 10.1186/s41747-018-0061-6] [Medline: 30353365]
6. van Leeuwen KG, Schalekamp S, Rutten M, van Ginneken B, de Rooij M. Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. Eur Radiol 2021 Jun;31(6):3797-3804. [doi: 10.1007/s00330-021-07892-z] [Medline: 33856519]
7. Dikici E, Bigelow M, Prevedello LM, White RD, Erdal BS. Integrating AI into radiology workflow: levels of research, production, and feedback maturity. J Med Imag 2020;7(1):016502. [doi: 10.1117/1.JMI.7.1.016502]
8. Langlotz CP. The future of AI and informatics in radiology: 10 predictions. Radiology 2023 Oct;309(1):e231114. [doi: 10.1148/radiol.231114] [Medline: 37874234]
9. Pesapane F, Gnocchi G, Quarrella C, et al. Errors in radiology: a standard review. J Clin Med 2024 Jul 23;13(15):4306. [doi: 10.3390/jcm13154306] [Medline: 39124573]
10. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. Nat Rev Cancer 2018 Aug;18(8):500-510. [doi: 10.1038/s41568-018-0016-5] [Medline: 29777175]
11. Syed AB, Zoga AC. Artificial intelligence in radiology: current technology and future directions. Semin Musculoskelet Radiol 2018 Nov;22(5):540-545. [doi: 10.1055/s-0038-1673383] [Medline: 30399618]
12. Filice RW, Ratwani RM. The case for user-centered artificial intelligence in radiology. Radiol Artif Intell 2020 May;2(3):e190095. [doi: 10.1148/ryai.2020190095] [Medline: 33937824]
13. Hughes H, O'Reilly M, McVeigh N, Ryan R. The top 100 most cited articles on artificial intelligence in radiology: a bibliometric analysis. Clin Radiol 2023 Feb;78(2):99-106. [doi: 10.1016/j.crad.2022.09.133] [Medline: 36639176]
14. Liu DS, Abu-Shaban K, Halabi SS, Cook TS. Changes in radiology due to artificial intelligence that can attract medical students to the specialty. JMIR Med Educ 2023 Mar 20;9:e43415. [doi: 10.2196/43415] [Medline: 36939823]
15. Gallix B, Chong J. Artificial intelligence in radiology: who's afraid of the big bad wolf? Eur Radiol 2019 Apr;29(4):1637-1639. [doi: 10.1007/s00330-018-5995-9] [Medline: 30734848]
16. Gore JC. Artificial intelligence in medical imaging. Magn Reson Imaging 2020 May;68:A1-A4. [doi: 10.1016/j.mri.2019.12.006] [Medline: 31857130]
17. Cai CJ, Winter S, Steiner D, Wilcox L, Terry M. "Hello AI": uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making. Proc ACM Hum-Comput Interact 2019 Nov 7;3(CSCW):1-24. [doi: 10.1145/3359206]
18. Sorensen JI, Nikam RM, Choudhary AK. Artificial intelligence in child abuse imaging. Pediatr Radiol 2021 May;51(6):1061-1064. [doi: 10.1007/s00247-021-05073-0] [Medline: 33904953]
19. Goergen SK, Frazer HML, Reddy S. Quality use of artificial intelligence in medical imaging: what do radiologists need to know? J Med Imaging Radiat Oncol 2022 Mar;66(2):225-232. [doi: 10.1111/1754-9485.13379] [Medline: 35243782]
20. Kuo RYL, Freethy A, Smith J, et al. Stakeholder perspectives towards diagnostic artificial intelligence: a co-produced qualitative evidence synthesis. EClinicalMedicine 2024 May;71:102555. [doi: 10.1016/j.eclinm.2024.102555] [Medline: 38549586]
21. Huisman M, van Ginneken B, Harvey H. The emperor has few clothes: a realistic appraisal of current AI in radiology. Eur Radiol 2024 Sep;34(9):5873-5875. [doi: 10.1007/s00330-024-10664-0] [Medline: 38451323]
22. Kulkarni S, Jha S. Artificial intelligence, radiology, and tuberculosis: a review. Acad Radiol 2020 Jan;27(1):71-75. [doi: 10.1016/j.acra.2019.10.003] [Medline: 31759796]
23. Fourcade A, Khonsari RH. Deep learning in medical image analysis: a third eye for doctors. J Stomatol Oral Maxillofac Surg 2019 Sep;120(4):279-288. [doi: 10.1016/j.jormas.2019.06.002] [Medline: 31254638]
24. Dembrower K, Crippa A, Colón E, Eklund M, Strand F, ScreenTrustCAD Trial Consortium. Artificial intelligence for breast cancer detection in screening mammography in Sweden: a prospective, population-based, paired-reader, non-inferiority study. Lancet Digit Health 2023 Oct;5(10):E703-E711. [doi: 10.1016/S2589-7500(23)00153-X] [Medline: 37690911]
25. European Society of Radiology (ESR). Impact of artificial intelligence on radiology: a EuroAIM survey among members of the European Society of Radiology. Insights Imaging 2019 Oct 31;10(1):105. [doi: 10.1186/s13244-019-0798-3] [Medline: 31673823]

XSL•FO
RenderX

26. Waymel Q, Badr S, Demondion X, Cotten A, Jacques T. Impact of the rise of artificial intelligence in radiology: what do radiologists think? Diagn Interv Imaging 2019 Jun;100(6):327-336. [doi: 10.1016/j.diii.2019.03.015] [Medline: 31072803]
27. Goldberg JE, Rosenkrantz AB. Artificial intelligence and radiology: a social media perspective. Curr Probl Diagn Radiol 2019;48(4):308-311. [doi: 10.1067/j.cpradiol.2018.07.005] [Medline: 30143386]
28. Huisman M, Ranschaert E, Parker W, et al. An international survey on AI in radiology in 1,041 radiologists and radiology residents part 1: fear of replacement, knowledge, and attitude. Eur Radiol 2021 Sep;31(9):7058-7066. [doi: 10.1007/s00330-021-07781-5] [Medline: 33744991]
29. Chen Y, Wu Z, Wang P, et al. Radiology residents' perceptions of artificial intelligence: nationwide cross-sectional survey study. J Med Internet Res 2023 Oct 19;25:e48249. [doi: 10.2196/48249] [Medline: 37856181]
30. Miró Catalina Q, Femenia J, Fuster-Casanovas A, et al. Knowledge and perception of the use of AI and its implementation in the field of radiology: cross-sectional study. J Med Internet Res 2023 Oct 13;25:e50728. [doi: 10.2196/50728] [Medline: 37831495]
31. Yin RK. Case Study Research and Applications: Design and Methods, 6th edition: Sage Publications; 2018.
32. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005 Nov;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
33. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
34. Wiklund P, Medson K. Use of a deep learning algorithm for detection and triage of cancer-associated incidental pulmonary embolism. Radiol Artif Intell 2023 Nov;5(6):e220286. [doi: 10.1148/ryai.220286] [Medline: 38074784]
35. World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA 2013 Nov 27;310(20):2191-2194. [doi: 10.1001/jama.2013.281053] [Medline: 24141714]
36. Strohm L, Hehakaya C, Ranschaert ER, Boon WPC, Moors EHM. Implementation of artificial intelligence (AI) applications in radiology: hindering and facilitating factors. Eur Radiol 2020 Oct;30(10):5525-5532. [doi: 10.1007/s00330-020-06946-y] [Medline: 32458173]
37. Kim B, Romeijn S, van Buchem M, Mehrizi MHR, Grootjans W. A holistic approach to implementing artificial intelligence in radiology. Insights Imaging 2024 Jan 25;15(1):22. [doi: 10.1186/s13244-023-01586-4] [Medline: 38270790]
38. Naik N, Hameed BMZ, Shetty DK, et al. Legal and ethical consideration in artificial intelligence in healthcare: who takes responsibility? Front Surg 2022 Mar;9:862322. [doi: 10.3389/fsurg.2022.862322] [Medline: 35360424]
39. Sorantin E, Grasser MG, Hemmelmayr A, et al. The augmented radiologist: artificial intelligence in the practice of radiology. Pediatr Radiol 2022 Oct;52(11):2074-2086. [doi: 10.1007/s00247-021-05177-7] [Medline: 34664088]
40. Illeris K. The Fundamentals of Workplace Learning: Routledge; 2010. [doi: 10.4324/9780203836521]
41. Matthews P. Informal Learning at Work: How to Boost Performance in Tough Times: Three Faces Publishing; 2024.
42. De Grip A. The Importance of Informal Learning at Work: IZA World of Labor; 2024. [doi: 10.15185/izawol.162]
43. Callanan M, Cervantes C, Loomis M. Informal learning. Wiley Interdiscip Rev Cogn Sci 2011 Nov;2(6):646-655. [doi: 10.1002/wcs.143] [Medline: 26302414]
44. Stogiannos N, O'Regan T, Scurr E, et al. Lessons on AI implementation from senior clinical practitioners: an exploratory qualitative study in medical imaging and radiotherapy in the UK. J Med Imaging Radiat Sci 2025 Jan;56(1):101797. [doi: 10.1016/j.jmir.2024.101797] [Medline: 39579457]
45. Greenhalgh T, Papoutsi C. Studying complexity in health services research: desperately seeking an overdue paradigm shift. BMC Med 2018 Jun 20;16(1):95. [doi: 10.1186/s12916-018-1089-4] [Medline: 29921272]
46. Nobles J, Wheeler J, Dunleavy-Harris K, et al. Ripple effects mapping: capturing the wider impacts of systems change efforts in public health. BMC Med Res Methodol 2022 Mar 18;22(1):72. [doi: 10.1186/s12874-022-01570-4] [Medline: 35300619]
47. Pullmann MD, Dorsey S, Duong MT, et al. Expect the unexpected: a qualitative study of the ripple effects of children's mental health services implementation efforts. Implement Res Pract 2022;3:26334895221120797. [doi: 10.1177/26334895221120797] [Medline: 36504561]
48. Blendon RJ, DesRoches CM, Brodie M, et al. Views of practicing physicians and the public on medical errors. N Engl J Med 2002 Dec 12;347(24):1933-1940. [doi: 10.1056/NEJMsa022151] [Medline: 12477944]
49. Palinkas LA. Achieving Implementation and Exchange: Policy Press; 2019. [doi: 10.51952/9781447338147]
50. Rigby K, Redley B, Hutchinson AM. Change agent's role in facilitating use of technology in residential aged care: a systematic review. Int J Med Inform 2023 Nov;179:105216. [doi: 10.1016/j.ijmedinf.2023.105216] [Medline: 37734272]
51. Liew C. The future of radiology augmented with artificial intelligence: a strategy for success. Eur J Radiol 2018 May;102:152-156. [doi: 10.1016/j.ejrad.2018.03.019] [Medline: 29685530]
52. Scheek D, Rezazade Mehrizi MH, Ranschaert E. Radiologists in the loop: the roles of radiologists in the development of AI applications. Eur Radiol 2021 Oct;31(10):7960-7968. [doi: 10.1007/s00330-021-07879-w] [Medline: 33860828]
53. Khafaji MA, Safhi MA, Albadawi RH, Al-Amoudi SO, Shehata SS, Toonsi F. Artificial intelligence in radiology: are saudi residents ready, prepared, and knowledgeable? Saudi Med J 2022 Jan;43(1):53-60. [doi: 10.15537/smj.2022.43.1.20210337] [Medline: 35022284]

XSL•FO
RenderX

54. Aquino YSJ, Rogers WA, Braunack-Mayer A, et al. Utopia versus dystopia: professional perspectives on the impact of healthcare artificial intelligence on clinical roles and skills. Int J Med Inform 2023 Jan;169:104903. [doi: 10.1016/j.ijmedinf.2022.104903] [Medline: 36343512]
55. Lu J. Will medical technology deskill doctors? Int Educ Stud 2016;9(7):130-134. [doi: 10.5539/ies.v9n7p130]
56. Gilvary C, Madhukar N, Elkhader J, Elemento O. The missing pieces of artificial intelligence in medicine. Trends Pharmacol Sci 2019 Aug;40(8):555-564. [doi: 10.1016/j.tips.2019.06.001] [Medline: 31277839]
57. Nilsen P, Sundemo D, Heintz F, et al. Towards evidence-based practice 2.0: leveraging artificial intelligence in healthcare. Front Health Serv 2024;4:1368030. [doi: 10.3389/frhs.2024.1368030] [Medline: 38919828]
58. Graneheim UH, Lundman B. Qualitative content analysis in nursing research: concepts, procedures and measures to achieve trustworthiness. Nurse Educ Today 2004 Feb;24(2):105-112. [doi: 10.1016/j.nedt.2003.10.001] [Medline: 14769454]
59. Malterud K, Siersma VD, Guassora AD. Sample size in qualitative interview studies: guided by information power. Qual Health Res 2016 Nov;26(13):1753-1760. [doi: 10.1177/1049732315617444] [Medline: 26613970]
60. Guest G, Bunce A, Johnson L. How many interviews are enough? An experiment with data saturation and variability. Field methods 2006;18(1):59-82. [doi: 10.1177/1525822X05279903]

Abbreviations

AI: artificial intelligence
CT: computed tomography
MRI: magnetic resonance imaging

Edited by J Sarvestan; submitted 21.05.25; peer-reviewed by J Zhang, M Gasmi, S Kondapally, VV Sangaraju; revised version received 15.11.25; accepted 17.11.25; published 22.12.25.

Please cite as:
Nilsen P, Svedberg P, Larsson I, Petersson L, Nygren J, Steerling E, Neher M
Radiology Staff Experiences With Integrating Artificial Intelligence Into Radiology Practice in a Swedish Hospital: Qualitative Case Study
JMIR Form Res 2025;9:e77843
URL: https://formative.jmir.org/2025/1/e77843
doi:10.2196/77843

© Per Nilsen, Petra Svedberg, Ingrid Larsson, Lena Petersson, Jens Nygren, Emilie Steerling, Margit Neher. Originally published in JMIR Formative Research (https://formative.jmir.org), 22.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Prediction of Respiratory Decompensation in Patients Receiving Home Mechanical Ventilation: Machine Learning Model Development and Validation Study

Nerea Berbel Casado[1], MSc; Francesc López Seguí[1,2], PhD; Natalia Muñoz Moruno[1], MSc; Antoni Rosell[3,4,5], MD, PhD; Aïda Muñoz Ferrer[3,4,5], MD, PhD; Ignasi Garcia Olive[3,4,5], MD, PhD; Marina Galdeano Lozano[3,4,5], MD, PhD

[1]Chair in ICT and Health, Centre for Health and Social Care Research (CESS), Universitat de Vic - Universitat Central de Catalunya, Carrer Miquel Martí i Pol, 1, Vic, Spain

[2]Centre de Recerca en Economia i Salut, Universitat Pompeu Fabra, Barcelona, Spain

[3]Department of Pulmonology, Clinical Directorate of the Thorax Area, Hospital Universitari Germans Trias i Pujol, Carretera de Canyet, s/n, Badalona, Spain

[4]Institut d'Investigació en Ciències de la Salut Germans Trias i Pujol, Badalona, Spain

[5]Universitat Autònoma de Barcelona, Bellaterra, Spain

Corresponding Author:
Marina Galdeano Lozano, MD, PhD
Department of Pulmonology, Clinical Directorate of the Thorax Area, Hospital Universitari Germans Trias i Pujol, Carretera de Canyet, s/n, Badalona, Spain

Abstract

Background: Chronic respiratory diseases often require long-term ventilatory support, leading to a growing number of patients treated with home mechanical ventilation (HMV). Despite advancements in telemonitoring with real-time tracking of noninvasive mechanical ventilation enabled by integrated software in HMV devices, early signs of respiratory decompensation may go unnoticed, leading to emergency visits and hospitalizations, which burden both patients and health care systems.

Objective: This study aims to develop and evaluate a machine learning–based model capable of predicting respiratory decompensation events, defined as emergency visits or hospitalizations due to acute deterioration in the patient's underlying respiratory condition. These events reflect episodes of worsening respiratory status that require urgent medical attention. The model uses data from HMV telemonitoring platforms, with the aim of improving patient outcomes.

Methods: This retrospective study analyzed data from 482 patients on HMV monitored via ResMed, Philips, and Breas platforms, collected between March 2021 and November 2024 at the Germans Trias i Pujol Hospital in Catalonia, Spain. Data included device use, compliance, mask leakage, and ventilator settings. Decompensation was defined as emergency department visits or hospitalizations. A windowing strategy captured the 5 weeks prior to events. Multiple machine learning models were trained using grid search to identify the optimal hyperparameters, prioritizing recall in order to minimize false negatives. Models were evaluated using 10-fold cross-validation. Finally, Shapley Additive Explanations (SHAP) analysis was used for model interpretability.

Results: The final dataset included 157 data windows, balanced for positive and negative cases. Among the models tested, logistic regression achieved the highest recall (mean 0.94, SD 0.06, 95% CI 0.90 - 0.98) though with moderate accuracy (mean 0.60, SD 0.05, 95% CI 0.56 - 0.64). The random forest classifier achieved the best balance across metrics (accuracy: mean 0.66, SD 0.10, 95% CI 0.59 - 0.73; recall: mean 0.78, SD 0.15, 95% CI 0.67 - 0.89; F_1-score: mean 0.70, SD 0.10, 95% CI 0.63 - 0.77). SHAP analysis revealed that higher use, leakage, and compliance in the week before a decompensation event were key predictors, suggesting compensatory behavior or early clinical deterioration. Overall performance remained moderate, reflecting limitations in sample size, incomplete longitudinal records for daily data, and the absence of key physiological measurements.

Conclusions: This study demonstrates the feasibility of predicting respiratory decompensation using data from HMV telemonitoring systems. Tree-based ensemble models, particularly random forests, provided the most balanced performance, while SHAP analysis offered clinically relevant insights. Although performance was moderate, the findings support further development of predictive tools to enable timely telemedical interventions. Limitations include sample size, missing physiological parameters, and a single-center design. Future research should expand to multicenter datasets and incorporate additional clinical variables to enhance model robustness and generalizability.

(JMIR Form Res 2025;9:e78941) doi:10.2196/78941

XSL•FO
RenderX

KEYWORDS

home mechanical ventilation; machine learning; noninvasive ventilation; predictive modeling; random forest classifier; remote patient monitoring; respiratory decompensation; Shapley Additive Explanations analysis; SHAP

Introduction

The increasing life expectancy of the population has led to the chronicization of numerous diseases, including respiratory conditions. As a result, changes in the treatment and monitoring of these pathologies are necessary. Over the years, there has been a significant rise in the number of patients with respiratory diseases that ultimately develop into chronic respiratory failure, requiring oxygen therapy with or without ventilatory support [1,2].

In response to this growing challenge, noninvasive mechanical ventilation (NIMV) has become widely used since the 1980s, prompting the creation of intermediate respiratory care units (IRCUs). In 2005, the Spanish Society of Pulmonology and Thoracic Surgery defined IRCUs as specialized units dedicated to the monitoring and treatment of patients with acute or exacerbated respiratory failure, aiming to provide appropriate respiratory monitoring and treatment through NIMV [3]. Many patients admitted to IRCUs are advised to continue treatment at home with home mechanical ventilation (HMV) devices. Recent technological advancements have enabled the integration of software into HMV devices, facilitating close and continuous patient monitoring. However, despite these improvements, issues such as poor adherence, limited family support, and undetected patient-ventilator asynchronies continue to compromise treatment effectiveness and patient health [4,5].

Previous research suggests that combining telemonitoring and teleintervention could improve health outcomes, with one study reporting a 70% reduction in the risk of cardiovascular events [6], highlighting the broad potential of such integrated approaches. Additionally, changes detected by the integrated monitoring systems in HMV devices may correlate with clinical deterioration in these patients [7,8]. This convergence of HMV data and proven teleintervention strategies highlights the potential of using HMV parameters in the context of telemedicine for early detection of respiratory decompensation, enabling timely adjustment of ventilator settings or remote intervention. Previous studies have shown the potential of machine learning in predicting ventilator-associated complications among critically ill intensive care unit patients, reaching an average accuracy of 0.69 (SD 0.10) [9]. However, this study focuses on invasive mechanical ventilation, incorporating parameters that are not available in HMV applications.

Building on this evidence, this paper aims to assess the feasibility of using machine learning to detect early signs of respiratory decompensation, based on data extracted from HMV software. Respiratory decompensation in this context is defined as emergency department visits or hospitalizations due to acute deterioration in the patient's underlying respiratory condition, reflecting episodes of worsening respiratory status that require urgent medical attention; therefore, early identification of potential decompensation can significantly reduce preventable emergency visits due to worsening clinical status or quality of life, which also impose a considerable travel burden on frail individuals. The ultimate goal of this approach is to anticipate respiratory decompensation in individuals using HMV by identifying early warning patterns in ventilator data. Future data-driven predictive models integration into telemonitoring workflows could enable timely adjustments of ventilatory parameters or clinical management before an acute event occurs, helping to maintain patient stability, reduce emergency visits and hospitalizations, improve quality of life, and optimize health care resource use.

Methods

Patient Recruitment and Inclusion Criteria

This study was conducted at Germans Trias i Pujol Hospital (HUGTiP), a reference hospital in Catalonia (Spain) that serves a population of 1,448,812 people across 70 municipalities. The study included data collected between March 2021 and November 2024. Patients were recruited voluntarily before their inclusion in the study. The study population comprised hospitalized patients who required HMV initiation after an acute exacerbation in the IRCU or in general hospitalization wards at HUGTiP. It also included patients already on HMV who required therapy adjustments before discharge. Additionally, patients attending outpatient clinics who required new HMV initiation or adjustments in follow-up consultations were included in the analysis.

The study population, comprising both women and men, had a mean age of 69.3 years, with most individuals aged between 60 and 80 years. Chronic obstructive pulmonary disease accounted for more than half of the patients treated with NIMV, representing the main indication for ventilatory support. Other underlying conditions included obesity hypoventilation syndrome (often associated with obstructive sleep apnea) and restrictive thoracic or neuromuscular disorders, such as postsurgical thoracic sequelae, amyotrophic lateral sclerosis, and muscular dystrophies.

To ensure appropriate follow-up and reliable data acquisition, only patients residing within the hospital's coverage area were considered eligible for HMV. Furthermore, patients needed to have a designated caregiver to assist in therapy adherence and monitoring.

Data Acquisition and Telemonitoring

Patient monitoring was conducted through validated commercial telemonitoring platforms commonly used in HMV: ResMed, Philips, and Breas. These platforms allow for real-time tracking of NIMV therapy. The home respiratory therapy provider, Esteve Teijin, which is responsible for the geographic area of HUGTiP, granted access to its telemonitoring system through a licensed workstation located at the sleep unit in the Pulmonology Department at HUGTiP.

XSL•FO
RenderX

The telemonitoring system facilitated automated data extraction from ventilators, providing detailed information on therapy adherence and device performance through weekly and daily reports. The weekly reports included information on ventilator use patterns, ventilation settings, and device and mask characteristics. More specifically, the extracted data comprised weekly continuous compliance, last recorded values of expiratory and inspiratory positive airway pressure, ventilation mode, ramp activation and duration, device type, mask type, and size. Daily reports focused on key physiological indicators, such as use (in min), apnea-hypopnea index, and leakage levels.

To provide a comprehensive view of patient status, ventilator-derived parameters were supplemented with hospital admission and emergency department visit dates. Consequently, the study included 482 patients for whom both telemonitoring reports and corresponding date records were available.

Ethical Considerations

This study complies with the approval of the Ethics Committee of HUGTiP (approval code PI-21‑236) and adheres to the ethical principles outlined in the Declaration of Helsinki [10], ensuring the highest standards of research ethics. Informed consent was obtained from all participants (or their legal guardians, where applicable) prior to enrollment. No financial compensation was provided for participation, and involvement was entirely voluntary. To ensure confidentiality, all personal identifiers, including names and medical record numbers, were replaced by the hospital's information system professionals with unique study codes. These codes were used for data analysis, ensuring that no real patient information is disclosed. Patient autonomy and rights were respected throughout the study, with participant well-being prioritized above scientific or societal interests.

Data Preprocessing

Overview

To capture temporal trends associated with decompensation events, a data preprocessing pipeline was implemented based on the consecutive reports obtained from the telemonitoring system. Both the constant parameters, expiratory positive airway pressure and inspiratory positive airway pressure, and the dynamic variables from the reports that reflect treatment adherence and patient status were selected, which included continuous compliance in the current period, use, apnea-hypopnea index, and leakage. The key steps of the pipeline are described in the following sections.

Windowing Strategy

Ensuring that each patient had consecutive reports with complete data was critical to capture the patterns leading up to a decompensation. However, strict continuity requirements risked significant data loss. To overcome this limitation, "windows" were defined spanning from 35 days before a decompensation event up to 7 days before the event date. In addition, for the daily variables, the mean of the available values for the weekdays was calculated. This allowed for the inclusion of patients even if they had some missing values. In patients with multiple decompensation events separated by at least 35 days, each event was treated as an independent window.

Reference Group Selection

The reference group was comprised of patients without any emergency or hospital admission, as well as patients with decompensation events, from whom report data were selected only if both the start and end of the window were at least 35 days away from any decompensation event, even when multiple events occurred. This approach ensured that the models would learn temporal patterns predictive of imminent decompensation, rather than simply distinguishing between inherently frail patients—who tend to have more frequent health care encounters and poorer baseline health—and less frail patients. Frail patients often exhibit distinct patterns in ventilator use, such as higher total use, which could confound the model if not properly controlled. By controlling for these differences, the model focuses on true predictors of imminent decompensation rather than patterns related solely to baseline frailty.

Vector Preparation

Each variable was first rescaled using min-max normalization, computing each feature's global minimum and maximum across all reports. Next, to emphasize how recent each measurement was relative to a decompensation event, each report was multiplied by a decreasing time-decay weight. The closest report to the event received a weight of 1, while older reports were down-weighted by a constant decreasing factor, up to 0.5 for the oldest report. Finally, the dataset was balanced to ensure representation of both classes.

Model Development and Evaluation

Different machine learning models were trained with grid search for hyperparameter tuning, ensuring that the best possible parameter configurations were identified. The models included logistic regression (LR), random forest classifier (RFC), decision tree classifier (DTC), naïve Bayes (NB), k-nearest neighbors (KNN), and extra trees classifier (ETC). For each model, a predefined set of hyperparameters was exhaustively explored. For LR, the hyperparameters included the regularization strength *C*, ranging from 0.01 to 100, the penalty types *l1* and *l2*, the solvers *lbfgs*, *liblinear*, and *saga*, and maximum iterations ranging from 1000 to 1,000,000. The RFC and ETC were both tuned by varying the number of estimators between 10 and 500, adjusting the maximum depth from shallow to unrestricted (up to 30), and modifying the minimum number of samples required to split a node or form a leaf, while also exploring different splitting criteria such as *gini*, *entropy*, or *log_loss*. DTCs were similarly optimized by adjusting depth, split, and leaf requirements and splitting criteria to balance model complexity and generalization. KNN was tuned by varying the number of neighbors from 1 to 11, the weighting method between uniform and distance, and the algorithm among *auto*, *ball_tree*, *kd_tree*, and *brute*. Finally, for NB, the variance smoothing parameter ranged from 1·e-12 to 1·e-7.

Given the clinical context, where minimizing false negatives is crucial, grid search was configured to prioritize maximizing recall, thus ensuring that the number of false negatives is minimized, as failing to identify a positive case is more critical

XSL•FO
RenderX

than a false positive in this setting. All models were implemented using the open-source Python (version 3.11.12; Python Software Foundation) library *scikit-learn* [11].

Additionally, two separate neural network models were developed. The first was a neural network (NN) with a depth ranging from 2 to 5 layers, selected based on performance. This model consisted of alternating fully connected layers and rectified linear unit activations, ending with a sigmoid output layer for binary classification. The second model was a recurrent neural network (RNN), composed of stacked RNN layers with tanh activation, followed by fully connected layers and rectified linear unit activations, also terminating in a sigmoid output. The optimal architecture for each model was empirically determined. Both models were trained using binary cross-entropy loss, and the limited-memory Broyden-Fletcher-Goldfarb-Shanno and Adam optimizers were evaluated and compared. The entire pipeline was implemented using the open-source Python library *PyTorch* [12].

Models were evaluated through 10-fold cross-validation, using accuracy, precision, recall, and F_1-score as performance metrics. Results are reported as mean (SD) with 95% CIs.

Model Interpretation

In this study, the Shapley Additive Explanations (SHAP) method [13] was used to assess the interpretability of the predictive models developed. Once the optimal model was identified, a SHAP explainer was created. This explainer is designed to compute SHAP values, where each value indicates the contribution of a corresponding feature toward the final prediction, thus breaking down the prediction into interpretable parts. Summary plots that display the global importance of features across the entire dataset were generated. These plots rank features by their average impact on predictions and reveal the direction of their influence, making it easier to interpret how changes in feature values affect the model's outcomes.

Results

Windows

The dataset resulted in a balanced sample size of 157 windows; thus, each fold holds 16 windows for testing and the remaining 141 for training.

Evaluation Outcomes

Results show that LR with C=0.1, a maximum of 1000 iterations and the *liblinear* solver, achieved the highest mean recall of 0.94 (SD 0.06; 95% CI 0.90 - 0.98) but exhibited a mean overall accuracy of 0.60 (SD 0.05, 95% CI 0.56 - 0.64), a mean F_1-score of 0.71 (SD 0.03, 95% CI 0.69 - 0.73), and a mean precision of 0.58 (SD 0.04, 95% CI 0.55 - 0.61; Table 1). RFC, with 200 trees, a maximum depth of 5, and default splitting and leaf parameters, reached the highest mean accuracy of 0.66 (SD 0.10, 95% CI 0.59 - 0.73), along with a mean recall of 0.78 (SD 0.15, 95% CI 0.67 - 0.89), a mean precision of 0.66 (SD 0.13, 95% CI 0.57 - 0.75), and mean F_1-score of 0.70 (SD 0.10, 95% CI 0.63 - 0.77). The ETC, a related ensemble approach, using 50 trees with similar depth constraints, achieved a mean accuracy of 0.62 (SD 0.08, 95% CI 0.56 - 0.68), mean recall of 0.81 (SD 0.12, 95% CI 0.72 - 0.89), and mean F_1-score of 0.69 (SD 0.07, 95% CI 0.64 - 0.74).

XSL•FO
RenderX

Table . Results of classifier performance across 10-fold cross-validation with hyperparameter settings (N=157).

Model	Hyperparameters	Accuracy, mean (SD, 95% CI)	Recall, mean (SD, 95% CI)	Precision, mean (SD, 95% CI)	F_1-score, mean (SD, 95% CI)
LR[a]	• C=0.1 • max_iter=1000 • solver= 'liblinear'	0.60 (0.05, 0.56 - 0.64)	0.94 (0.06, 0.90 - 0.98)	0.58 (0.04, 0.55 - 0.61)	0.71 (0.03, 0.69 - 0.73)
RFC[b]	• max_depth=5 • min_samples_leaf=1 • min_samples_split=2 • n_estimators=200	0.66 (0.10, 0.59 - 0.73)	0.78 (0.15, 0.67 - 0.89)	0.66 (0.13, 0.57 - 0.75)	0.70 (0.10, 0.63 - 0.77)
DTC[c]	• criterion='entropy' • max_depth=None • min_samples_leaf=1 • min_samples_split=5 • min_weight_fraction_leaf= 0	0.57 (0.12, 0.48 - 0.66)	0.67 (0.17, 0.55 - 0.79)	0.58 (0.09, 0.52 - 0.65)	0.61 (0.11, 0.53 - 0.69)
ETC[d]	• criterion='entropy' • max_depth=5 • min_samples_leaf=1 • min_samples_split=2 • n_estimators=50	0.62 (0.08, 0.56 - 0.68)	0.81 (0.12, 0.72 - 0.89)	0.62 (0.09, 0.56 - 0.69)	0.69 (0.07, 0.64 - 0.74)
KNN[e]	• n_neighbors=3 • weights='distance'	0.60 (0.15, 0.49 - 0.71)	0.70 (0.20, 0.56 - 0.84)	0.61 (0.13, 0.52 - 0.70)	0.64 (0.15, 0.53 - 0.75)
NB[f]	• var_smoothing=1·e-11	0.57 (0.09, 0.51 - 0.64)	0.62 (0.13, 0.53 - 0.71)	0.61 (0.14, 0.51 - 0.71)	0.60 (0.08, 0.54 - 0.66)
NN[g]	• Layers=100,50 • solver='L-BFGS[i]'	0.61 (0.06, 0.57 - 0.65)	0.64 (0.12, 0.55 - 0.73)	0.63 (0.07, 0.58 - 0.68)	0.63 (0.07, 0.58 - 0.68)
RNN[h]	• Layers=100,50,16,8,1 • solver='Adam'	0.61 (0.11, 0.53 - 0.69)	0.70 (0.18, 0.57 - 0.83)	0.62 (0.12, 0.53 - 0.71)	0.65 (0.11, 0.57 - 0.73)

[a]LR: logistic regression.

[b]RFC: random forest classifier.

[c]DTC: decision tree classifier.

[d]ETC: extra trees classifier.

[e]KNN: K-nearest neighbors.

[f]NB: naïve Bayes.

[g]NN: neural network.

[h]RNN: recurrent neural network.

[i]L-BFGS: limited-memory Broyden-Fletcher-Goldfarb-Shanno

The single DTC with entropy criterion and unrestricted depth reached a mean accuracy of 0.57 (SD 0.12, 95% CI 0.48 - 0.66), mean recall of 0.67 (SD 0.17, 95% CI 0.55 - 0.79), mean precision of 0.58 (SD 0.09, 95% CI 0.52 - 0.65), and mean F_1-score of 0.61 (SD 0.11, 95% CI 0.53 - 0.69). KNN, with three neighbors and distance weighting, produced a mean accuracy of 0.60 (SD 0.15, 95% CI 0.49 - 0.71), mean recall of 0.70 (SD 0.20, 95% CI 0.56 - 0.84) and mean F_1-score of 0.64 (SD 0.15, 95% CI 0.53 - 0.75). NB, tuned with a *var_smoothing* parameter of 1·e-11, showed a mean accuracy of 0.57 (SD 0.09, 95% CI 0.51 - 0.64) and mean F_1-score of 0.60 (SD 0.08, 95% CI 0.54 - 0.66). The neural network with hidden layers of 100 and 50 neurons achieved a mean accuracy of 0.61 (SD 0.06, 95% CI 0.57 - 0.65), mean recall of 0.64 (SD 0.12, 95% CI 0.55 - 0.73), and mean F_1-score of 0.63 (SD 0.07, 95% CI 0.58 - 0.68), while the RNN, comprising 2 *simpleRNN*

XSL•FO
RenderX

layers of 100 and 50 units followed by dense layers, yielded comparable metrics (accuracy: mean 0.61, SD 0.11, 95% CI 0.53 - 0.69; recall: mean 0.70, SD 0.18, 95% CI 0.57 - 0.83; and F_1-score: mean 0.65, SD 0.11, 95% CI 0.57 - 0.73).

Feature Importances

As shown in Figure 1, the SHAP interpretation of the RFC. The labels (1) through (5) correspond to the number of weeks preceding a decompensation event: (1) indicates the week immediately before the event, (2) two weeks prior, and so on, up to (5), which represents 5 weeks before decompensation. The SHAP values highlight the importance of use patterns in the model's predictions. Specifically, higher patient use and continuous compliance during the week immediately prior to the event (1) appear to be strong indicators of respiratory decompensation. Lower compliance in weeks (3) and (4) also contributes to the model's prediction, while higher compliance in weeks (1), (2), and (5) is similarly associated with increased risk. Elevated leakage values in the period leading up to the event also emerge as relevant features influencing the model's output.

Figure 1. Feature importances and Shapley Additive Explanation (SHAP) values from a random forest classifier (RFC) trained with a fixed train-test split (random seed=42). AHI: apnea-hypopnea index; IPAP: inspiratory positive airway pressure.

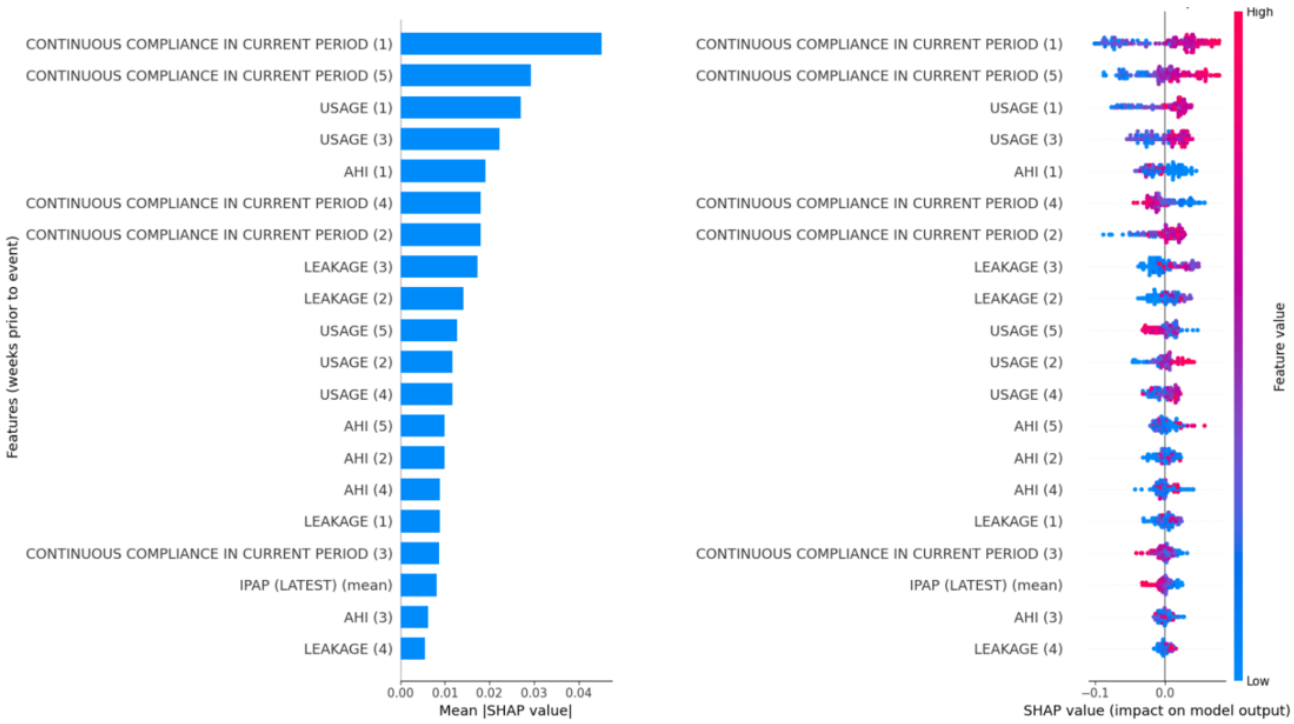

Discussion

Principal Results

Although LR achieved the highest recall (mean 0.94, SD 0.06, 95% CI 0.90 - 0.98), it showed lower precision (mean 0.58, SD 0.04, 95% CI 0.55 - 0.61) and accuracy (mean 0.60, SD 0.05, 95% CI 0.56 - 0.64), indicating a tendency to overpredict positive cases. This may be advantageous in clinical scenarios where identifying all potential positives is prioritized, but it comes at the cost of increased false positives. In contrast, the RFC, configured with 200 trees and a maximum depth of 5, provided a more balanced performance across all metrics, including the highest accuracy (mean 0.66, SD 0.10, 95% CI 0.59 - 0.73) and solid values for recall (mean 0.78, SD 0.15, 95% CI 0.67 - 0.89), precision (mean 0.66, SD 0.13, 95% CI 0.57 - 0.75), and F_1-score (mean 0.70, SD 0.10, 95% CI 0.63 - 0.77). This suggests that RFC may be more suitable for general-purpose prediction in this context, balancing sensitivity and specificity.

Other models such as DTC and KNN performed moderately, with KNN exhibiting high variability across folds, indicating sensitivity to the specific train-test splits. DTC and NB recorded the lowest accuracy among all models, while NB also showed the lowest F_1-score. This underperformance of NB suggests that its strong independence assumptions limit its ability to capture data patterns. Ensemble approaches like the ETC also yielded competitive results, reinforcing the robustness of tree-based methods. Both neural networks achieved intermediate performance; however, their capacity to learn more intricate patterns may have been limited by the relatively small sample size. Recurrent models generally require larger datasets and longer temporal sequences to effectively learn temporal dependencies. In this study, the relatively short input sequences and small cohort may have restricted the RNN's capacity to model temporal dynamics robustly. Larger datasets, finer-grained temporal data, and extended temporal windows could allow the exploration of more advanced sequential architectures, such as attention-based models in future applications.

SHAP analysis from the RFC indicated that increased compliance, use, and leakage in the week preceding decompensation could be meaningful predictors. Although at first glance, it may seem counterintuitive that higher compliance

XSL•FO
RenderX

is associated with increased risk, this pattern could reflect increased ventilator dependence in patients whose health is deteriorating. Patients may use the device more intensively as their condition worsens, while elevated leakage could indicate improper mask fitting under higher use. Furthermore, variations in global compliance patterns over time, such as a drop in compliance during periods (3) and (4), may be indicative of behavioral changes that precede decompensation.

While SHAP does not directly assess temporal trends or causality, these insights highlight the potential clinical relevance of use trends, suggesting that both increases and fluctuations in ventilator parameters could serve as early warning signals. Nonetheless, interpretability remains a challenge: models with moderate performance may limit the robustness of conclusions, and potential confounding factors that are not captured in the dataset—particularly the absence of detailed physiological variables—could reduce predictive power or lead to misattributions.

Limitations

One of the primary limitations of this study is the absence of enough physiological explanatory variables. While many such variables are inherently unavailable from HMV data, respiratory rate, although a common monitored parameter, was not available in the reports used for the analysis. The absence of broader physiological or clinical variables (eg, comorbidities and vital signs) could have limited the models' ability to account for the full clinical picture, reducing the predictive power.

Another notable constraint is the sample size. Although the dataset comprised 482 patients, and the strategy of constructing windows aimed to maximize the sample size, extracting windows with complete necessary information remained a challenge: generally, complete daily data across all 5 consecutive week reports could not be obtained consistently. For this reason, to address data sparsity, available daily measurements were averaged, as mentioned previously in the methodology. However, this approach could compromise temporal resolution, and due to the characteristics of the missing data, data augmentation techniques were insufficient to overcome this limitation.

Finally, the study is limited by its monocentric design, as all patients were sourced from a single center, which may impact the generalizability of the findings.

Comparison With Prior Work

Although no studies in the literature directly use HMV parameters to predict respiratory decompensation [6], constructed an RF model that achieved an average accuracy of 0.69 (SD 0.10) for predicting ventilatory complications in critically ill intensive care unit patients, using invasive mechanical ventilation parameters. This performance is comparable to our model, which achieved an average accuracy of 0.66 (SD 0.10, 95% CI 0.59 - 0.73) using RFC models.

Conclusions

This study introduced a novel framework for predicting respiratory decompensation using only HMV data. Ensemble tree models, particularly a random forest, achieved the best balance of metrics (accuracy: mean 0.66, SD 0.10, 95% CI 0.59 - 0.73; recall: mean 0.78, SD 0.15, 95% CI 0.67 - 0.89), while SHAP analysis revealed that higher use, leakage, and compliance 7 days before, that is, period (1), could be an indicator of respiratory decompensation.

Despite the promising nature of these findings, overall performance remained moderate, reflecting the limitations imposed by incomplete longitudinal daily records and the lack of key physiological measurements. Combined with the relatively small sample size, these constraints position the present results as exploratory rather than ready for clinical application. Nevertheless, the findings suggest the potential of using NIMV data for early decompensation identification. Future research should focus on incorporating more clinical and contextual features and validating larger, multicenter cohorts to improve predictive power. With such advancements, this approach could translate into a clinically useful, noninvasive decision support tool for the proactive management of patients on home ventilation.

Acknowledgments

The authors acknowledge the use of generative artificial intelligence tools (ChatGPT-4.5) for assistance in proofreading and grammar correction. All tasks were performed under full human supervision. All content has been reviewed and validated by the authors to ensure accuracy and integrity. Responsibility for the final manuscript lies entirely with the authors. Generative artificial intelligence tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Conflicts of Interest

None declared.

References

XSL•FO
RenderX

1. Department of Economic and Social Affairs, Population Division. World Population Prospects: The 2006 Revision Volume I: Comprehensive Tables: United Nations; 2007. URL: https://www.un.org/development/desa/pd/sites/www.un.org.development.desa.pd/files/files/documents/2020/Jan/un_2006_world_population_prospects-2006_revision_volume-i.pdf [accessed 2025-12-08]
2. Needham DM, Bronskill SE, Calinawan JR, Sibbald WJ, Pronovost PJ, Laupacis A. Projected incidence of mechanical ventilation in Ontario to 2026: Preparing for the aging baby boomers. Crit Care Med 2005 Mar;33(3):574-579. [doi: 10.1097/01.ccm.0000155992.21174.31] [Medline: 15753749]
3. Torres A, Ferrer M, Blanquer JB, et al. Unidades de cuidados respiratorios intermedios. Definición y características [Article in Spanish]. Archivos de Bronconeumología 2005 Sep;41(9):505-512. [doi: 10.1157/13078653] [Medline: 16194514]
4. Gonzalez-Bermejo J, Perrin C, Janssens JP, et al. Proposal for a systematic analysis of polygraphy or polysomnography for identifying and scoring abnormal events occurring during non-invasive ventilation. Thorax 2012 Jun;67(6):546-552. [doi: 10.1136/thx.2010.142653] [Medline: 20971982]
5. Janssens JP, Borel JC, Pépin JL, Group S. Nocturnal monitoring of home non-invasive ventilation: the contribution of simple tools such as pulse oximetry, capnography, built-in ventilator software and autonomic markers of sleep fragmentation. Thorax 2011 May;66(5):438-445. [doi: 10.1136/thx.2010.139782] [Medline: 20971980]
6. Yun S, Comín-Colet J, Calero-Molina E, et al. Evaluation of mobile health technology combining telemonitoring and teleintervention versus usual care in vulnerable-phase heart failure management (HERMeS): a multicentre, randomised controlled trial. Lancet Digit Health 2025 May;7(5):100866. [doi: 10.1016/j.landig.2025.02.006] [Medline: 40374486]
7. Rabec C, Rodenstein D, Leger P, et al. Ventilator modes and settings during non-invasive ventilation: effects on respiratory events and implications for their identification. Thorax 2011 Feb;66(2):170-178. [doi: 10.1136/thx.2010.142661] [Medline: 20947891]
8. Duiverman ML. “Tricks and tips for home mechanical ventilation” home mechanical ventilation: set-up and monitoring protocols. Pulmonology 2021;27(2):144-150. [doi: 10.1016/j.pulmoe.2020.08.002] [Medline: 32912752]
9. Tan Y, Young M, Girish A, et al. Predicting respiratory decompensation in mechanically ventilated adult ICU patients. Front Physiol 2023;14:1125991. [doi: 10.3389/fphys.2023.1125991] [Medline: 37123253]
10. World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA 2013 Nov 27;310(20):2191-2194. [doi: 10.1001/jama.2013.281053] [Medline: 24141714]
11. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12(85):2825-2830. [doi: 10.5555/1953048.2078195]
12. Paszke A, Gross S, Massa F, et al. PyTorch: an imperative style, high-performance deep learning library. Presented at: Advances in Neural Information Processing Systems 32; Dec 8-14, 2019 URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf [accessed 2025-12-08]
13. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Guyon I, Luxburg U, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al, editors. Presented at: 31st Conference on Neural Information Processing Systems (NIPS 2017); Dec 4-9, 2017 URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf [accessed 2025-12-08]

Abbreviations

NB: naïve Bayes
ETC: extra trees classifier
DTC: decision tree classifier
HMV: home mechanical ventilation
HUGTiP: Germans Trias i Pujol Hospital
IRCU: intermediate respiratory care unit
KNN: k-nearest neighbor
LR: logistic regression
NIMV: noninvasive mechanical ventilation
NN: neural network
RFC: random forest classifier
RNN: recurrent neural network
SHAP: Shapley Additive Explanations

XSL•FO
RenderX

Edited by A Schwartz, M Balcarras; submitted 12.06.25; peer-reviewed by E Bai, RL Joyner; revised version received 08.11.25; accepted 21.11.25; published 30.12.25.

Please cite as:
Berbel Casado N, López Seguí F, Muñoz Moruno N, Rosell A, Muñoz Ferrer A, Garcia Olive I, Galdeano Lozano M
Prediction of Respiratory Decompensation in Patients Receiving Home Mechanical Ventilation: Machine Learning Model Development and Validation Study
JMIR Form Res 2025;9:e78941
URL: https://formative.jmir.org/2025/1/e78941
doi:10.2196/78941

© Nerea Berbel Casado, Francesc López Seguí, Natalia Muñoz Moruno, Antoni Rosell, Aïda Muñoz Ferrer, Ignasi Garcia Olive, Marina Galdeano Lozano. Originally published in JMIR Formative Research (https://formative.jmir.org), 30.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Comparison and Validation of Actigraphy Algorithms Using a Large Community Dataset: Algorithm Validation Study

Darshan Panesar[1], HBSc, MEd; Aashish Vichare[2], BSc; Jason Goncalves[3], BSc, MBA; Robyn Stremler[4], PhD, RN

[1]Applied Psychology and Human Development, Ontario Institute for Studies in Education, University of Toronto, 252 Bloor Street West, Toronto, ON, Canada

[2]Independent Researcher, Vancouver, BC, Canada

[3]Independent Researcher, Toronto, ON, Canada

[4]Lawrence Bloomberg Faculty of Nursing, University of Toronto, Toronto, ON, Canada

Corresponding Author:
Darshan Panesar, HBSc, MEd
Applied Psychology and Human Development, Ontario Institute for Studies in Education, University of Toronto, 252 Bloor Street West, Toronto, ON, Canada

Abstract

Background: For decades, the measurement of sleep and wake has relied upon watch-based actigraphy as an alternative to expensive, obtrusive clinical monitoring. At the time of this publication, we have relied upon a handful of algorithms to score actigraphy data as sleep or wake. However, these algorithms have largely been tested and validated with only small samples of young, healthy individuals.

Objective: This study aimed to establish the accuracy and agreement of conventional and traditional actigraphy algorithms against polysomnography, the clinical standard, using the diverse Multi-Ethnic Study of Atherosclerosis (MESA) sleep dataset. As a secondary objective, we examined algorithm and polysomnography agreement for key sleep metrics including total sleep time (TST), sleep efficiency (SE), and wake after sleep onset (WASO).

Methods: We assessed 5 well-established algorithms, including Cole-Kripke, University of California San Diego (UCSD) scoring, Kripke 2010, Philips-Respironics, and Sadeh, with and without rescoring across 1440 individuals (M_{age}=mean 69.36, SD 8.97) from the MESA sleep dataset. We conducted epoch-by-epoch comparisons assessing accuracy, confusion matrix analyses, receiver operator characteristic curves (ROC), area under the curve (AUC), and Bland-Altman analyses for agreement.

Results: Primary results indicated all algorithms demonstrated accuracy between 78% - 80% with the highest accuracy by the Kripke 2010 (80%) algorithm followed closely by the Cole-Kripke (80%) and Philips-Respironics (80% - 79%) algorithms. In addition, moderate Cohen κ agreement and moderate positive Matthews correlations were demonstrated by all algorithms. Further, all algorithms demonstrated significant mean difference across sleep metrics.

Conclusions: The findings of this study establish that these traditional actigraphy algorithms can, with high accuracy, detect sleep and wake in large, diverse population samples, including older adults or populations at risk of health conditions. However, these algorithms may carry difficulty for precise assessment of sleep metrics, especially in cases of sleep disorders or irregular sleep.

(JMIR Form Res 2025;9:e70778) doi:10.2196/70778

KEYWORDS

accelerometer; actigraphy; algorithm; Multi-Ethnic Study of Atherosclerosis; M.E.S.A.; polysomnography; sleep; sleep monitoring; sleep disorder; wake

Introduction

Background

Over several decades, actigraphy has been used to objectively examine rest and wake periods in a variety of participants. Actigraphs are accelerometers that measure activity (acceleration of motion) levels of the person wearing the device [1-3]. Given its small footprint and lower cost, actigraphy is often used as a simple alternative to traditional, more invasive sleep monitoring methods. Actigraphy is used to measure basic movement activity patterns of individuals noninvasively to assess when they are asleep (rest) or awake (wake) [1-3]. Movement data collected by actigraphs are analyzed using validated algorithms to determine, for each minute of recording, whether the wearer was asleep or awake. At the time of this publication, the vast majority of sleep-wake analysis of movement data collected by actigraphy continues to be done primarily using one of a handful of popular linear regression models. These include Cole-Kripke,

XSL•FO
RenderX

University of California San Diego (UCSD), Sadeh, and Philips-Respironics. Though these algorithms are extensively used, they have largely been tested with only small samples of young and healthy individuals [3-11]. There is little research examining the comparison between and validity of these algorithms with a large, demographically diverse sample or with a sample at risk of health conditions. Given the increasing use of actigraphy for a variety of populations, it is critical to evaluate these algorithms across a large, diverse population to give a better representation of their accuracy. Further, it is important to evaluate the accuracy of these algorithms across older adults, as these populations have higher susceptibility to, and high prevalence of health disorders. Given the wide range of studies using actigraphy, it is important to understand actigraphy's performance when used with these higher-risk populations for sleep-wake assessment [3,12-17]. In addition, given the comorbidity of sleep issues with other health disorders, the use of actigraphy in sleep studies has become more commonplace [3,12-17]. Therefore, it is highly important to assess the accuracy and reliability of traditional actigraphy algorithms with these populations.

Multi-Ethnic Study of Atherosclerosis Dataset

There are several significant challenges in collecting a large, diverse dataset. This, in part, is due to the ideal method of validation (for actigraphy), which is polysomnography, the gold standard for sleep assessment. Polysomnography and its analysis are expensive, clinic-based procedures that require extensive equipment and expertise, are highly time-consuming to conduct, and are therefore impractical for measuring sleep across several nights with large samples [8,18,19]. However, an ongoing initiative called the Multi-Ethnic Study of Atherosclerosis (MESA) has conducted a large sleep study (over 2000 participants), including time-synchronized actigraphy and polysomnography data as part of their larger initiative [20,21]. This sleep sample represents a portion of over 6800 participants between the ages of 45 and 84 years, free of cardiovascular disease at baseline, who were monitored longitudinally for subclinical cardiovascular disease [20,21]. This dataset presents a novel opportunity to establish a concrete comparison between conventional algorithms and establish their accuracy across a large population of older adults who are also at risk of health issues. Thus, this paper evaluates traditional actigraphy algorithms using the MESA dataset.

Current Study

The primary objective of this study was to establish the accuracy of traditional actigraphy algorithms against the gold standard polysomnography using a large established dataset. In doing so, we provide comprehensive foundational benchmarks of the most commonly used traditional actigraphy algorithms. Further, we aimed to also establish performance benchmarks of commonly used actigraphy algorithms across a diverse population with older adults and sleep pathologies. These benchmarks greatly inform research and clinical use of these algorithms. Further, this study provides detailed standards to developers of novel actigraphy technologies and actigraphy analysis methodologies. Based on previous, albeit smaller, sample validation studies, we hypothesized that traditional actigraphy algorithms would detect sleep and wake with high accuracy. In addition, we predicted that additional rescoring of actigraphy would help improve their accuracy over the algorithms alone.

In addition to sleep-wake, researchers and clinicians often examine sleep metrics to evaluate patterns or disruption of sleep patterns. As our secondary goal, we examined each algorithm's agreement and bias with polysomnography on several commonly studied sleep metrics, including total sleep time (TST), sleep efficiency (SE), and wake after sleep onset (WASO). TST is the total number of minutes an individual sleeps during the night from the first onset of sleep to sleep offset. On the other hand, WASO is the total duration of wake between the first onset of sleep and the sleep offset. Finally, SE is the percentage of time asleep (TST) during the total time spent in bed. As these are key sleep metrics, this study's results provide detailed information for current research and clinical practice as well as benchmarks for future improvements to current and novel actigraphy methods. We hypothesized that overall, sleep metrics derived from actigraphy algorithm-processed activity would have high agreement with polysomnography [22].

As an exploratory analysis, we looked at both accuracy and sleep metrics of each algorithm and evaluated possible points of failure and variables that impact the performance of actigraphy analyses. Specifically, we examined if accuracy results and sleep metrics would be poorer for subgroups of participants who had sleep disorders. To our knowledge, this is the first study to provide detailed actigraphy performance metrics across large key samples of individuals with sleep problems. These metrics provide nuanced information that will critically facilitate current and future practice. We discuss the implications of these variables, making recommendations for future directions of research.

Methods

Sample

The dataset sample for this study was derived from the MESA and acquired through the National Sleep Research Resource (NSRR) [20,21]. MESA is a large prospective, community-based study designed to examine the risk factors, prevalence, and progression of cardiovascular disease [20,21]. The MESA study was conducted from 2000 to 2002, with 6814 ethnically diverse men and women aged 45 - 84 years who were free of overt cardiovascular disease and recruited from 6 US sites [20,21]. MESA participants underwent periodic core physical assessments at enrollment and 10 years following initial recruitment (2010 - 2012) [20,21]. A subset of participants was invited to participate in the MESA ancillary sleep assessment study. This study entailed a home visit (1 night) in which home polysomnography was conducted along with concurrent actigraphy. Polysomnography data were analyzed by one of 3 polysomnologists using standard guidelines, while 2 scorers completed actigraphy analysis by first marking the sleep period and then using automatic scoring software. Scorers completed a MESA-rule–based training and certification before scoring, and postscoring both interscorer reliability was assessed at 2 timepoints for sleep stages (which pertained to the sleep data

XSL•FO
RenderX

used for the study). At time point 1 for sleep stage scoring, the interscorer intraclass correlation coefficients (ICC) across n=27 from 9 participants were as follows: Stage 1 was 0.86, Stage 2 was 0.63, Stage 3 - 4 was 0.81, and REM was 0.96 [20,21]. At time point 2, interscorer ICC across n=38 from 19 participants were as follows: Stage 1 was 0.74, Stage 2 was 0.81, Stage 3 - 4 was 0.79, and REM was 0.93 [20,21]. For this study, we used raw actigraphy activity counts for analysis. For ground truth, we used the matching polysomnography sleep-wake data that was collapsed to wake (coded wake stage) versus all sleep stages collapsed to one "sleep" variable. Only one night of polysomnography data were collected by the original study which was matched and synchronized with the respective actigraphy data (cropped to match).

The data were filtered based on several criteria. The initial dataset contained 2159 samples for actigraphy and 2056 samples for polysomnography. We first excluded samples that did not have concurrent polysomnography data (103 removed). The actigraphy and polysomnography samples were then matched based on a synchronization document provided by NSRR and MESA. Those samples that did not have a corresponding synchronization or were reported as having data issues by NSRR (2 individuals) were excluded, resulting in a sample of 1831 individuals (225 removed). Next, we filtered the dataset based on polysomnography data quality as indicated by the dataset description. We included individuals for whom the quality of polysomnography data was rated as very good or better, as this rating indicated an appropriate number of channels present for accurate sleep-wake evaluation. An additional 14 participants with complete, matched data of fair and good quality were also included. This resulted in a total of 1484 samples (347 removed).

Outlier filtering, or filtering based on actigraphy, was not conducted, as we wished to retain all actigraphy data for a true representation of how actigraphy data are traditionally processed using these algorithms. Given the absence of outlier filtering, we conducted sensitivity analysis to evaluate the effects of these outliers on our overall sample results. We identified outliers in the cropped 30-second epoch synchronized (1-night) actigraphy raw activity data using a 2-step process. The raw 30-second data were used as this was the data used for each algorithm (the binary algorithm results could not be used to identify outliers). Each participant's individual iIQR, mean, and SD of activity were calculated. For the first level, participants' IQR was compared to the full sample IQR. Any participants whose range fell outside of 1.5 times below the first quartile and above the first quartile were identified as outliers. For the second step, the *z* score for each participant was calculated relative to the sample mean and SD. Any participants with a *z* score of 2 or greater were identified as outliers. A total of 56 participants were identified as outliers using this 2-tier approach. Once these outlier individuals were identified, they were omitted, and Cohen κ, Matthews correlation coefficient (MCC), and confusion matrix metrics were recalculated. The relative change between the full dataset and outlier-omitted results was calculated to be less than 5% across all metrics and algorithms except for the Philips 40 rescored algorithm sensitivity (5.55%) and specificity (7%). However, the relative change was for only the rescored algorithm, close to the 5% threshold, and all other metrics were within the 5% range; therefore, we determined that the outliers did not pose a significant issue. For full comparison metrics, see Multimedia Appendix 1.

Further, polysomnography data were in binary format, and the actigraphy data, once processed through an algorithm, would also be in binary data point format. This would limit the breadth of fluctuation in the main analyses and would allow for direct 1:1 comparison of epochs. In addition to the main analyses, rescoring was also applied, which is the standard method of correcting issues with actigraphy data. In addition, we retained all participants for subsequent analysis of sleep metrics, as we aimed to evaluate the entire spread of sleep metric results for normal sleep, sleep problems, and extreme cases of sleep. As a secondary measure, we added analyses for participants with sleep problems to evaluate these extreme cases separately.

Finally, during data processing, an additional 44 individuals were excluded due to issues synchronizing actigraphy and polysomnography time points, missing data, and issues with processing. The final sample size analyzed was N=1440 (M_{age}=69.3, SD 9.0 years; n_{male}=663) individuals (Figure 1). The distribution of included participants is denoted in Table 1.

XSL•FO
RenderX

Figure 1. Data filtering and selection flowchart reflecting the data sample selection and filtering from the source National Sleep Research Resource–Multi-Ethnic Study of Atherosclerosis dataset. NSRR-MESA: National Sleep Research Resource–Multi-Ethnic Study of Atherosclerosis;

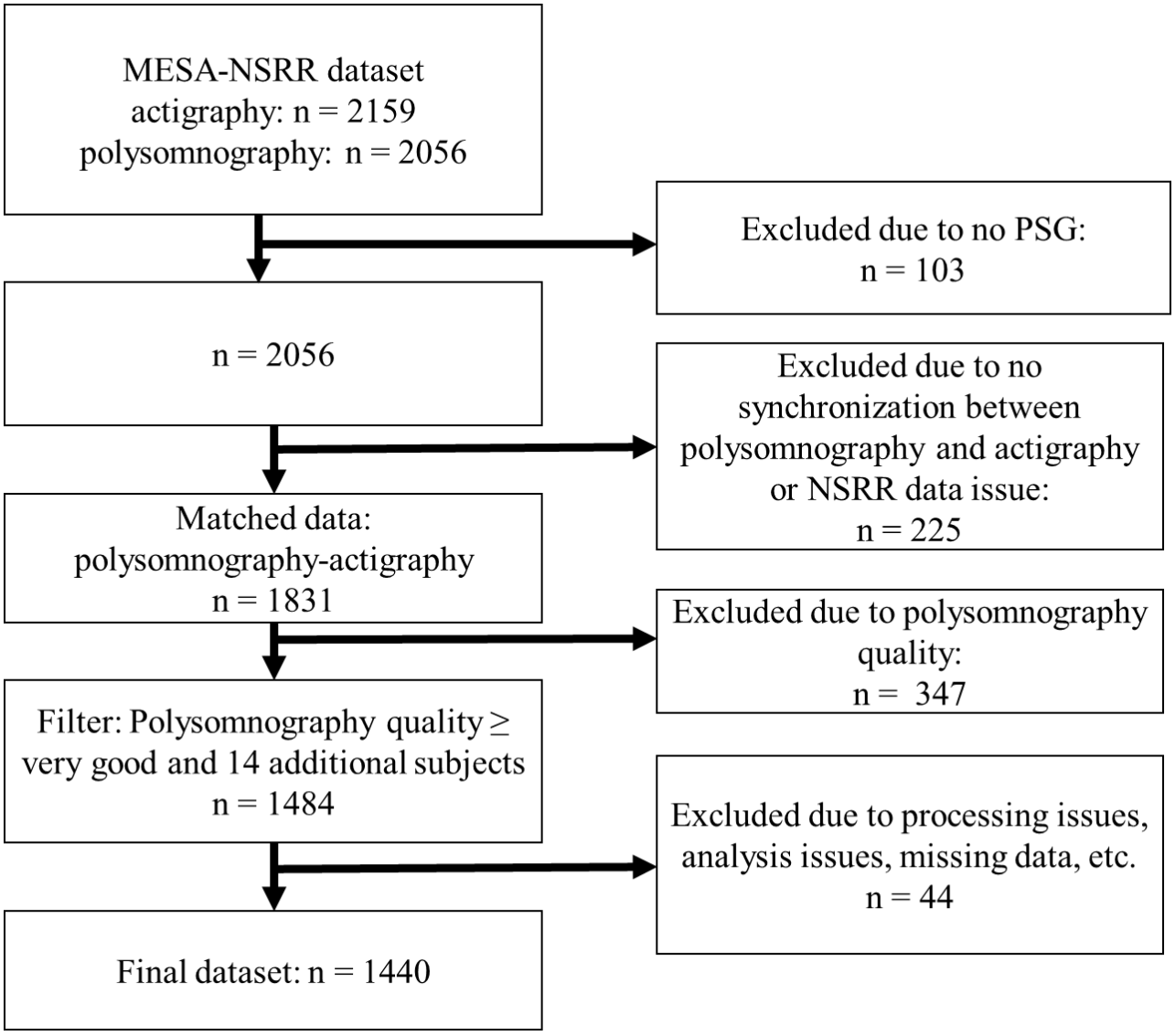

XSL•FO
RenderX

Table . Baseline characteristics of the selected sample subset from National Sleep Research Resource–Multi-Ethnic Study of Atherosclerosis sleep dataset.

Characteristic	Values
Age at study (years), mean (SD)	69.3 (9)
Age (years), range	54 - 94
Sex, n (%)	
Female	817 (55.7)
Male	663 (46)
Race, n (%)	
White	565 (39.2)
Black/African American	396 (27.5)
Hispanic	351 (24.4)
Asian	168 (11.7)
Sleep problems, n (%)	
Sleep apnea	104 (7.2)
Insomnia	84 (5.8)
Restless legs syndrome	66 (4.6)
Use of CPAP[a] machine for sleep	62 (4.3)
Snoring ≥3 - 5 times per week	584 (40.6)

[a]CPAP: continuous positive airway pressure.

In addition to the final sample, for exploratory analyses, subsamples for the populations with sleep problems were also selected. Four sleep problem groups, including apnea, individuals who used continuous positive airway pressure (CPAP), insomnia, and restless leg syndrome (RLS), were examined using the same detailed analyses and methods as the full sample (Table 1).

Data Processing

Data processing was conducted using Python 3 with a wide range of libraries and platforms. Before the final analyses, both actigraphy data and polysomnography data were transformed from 30-second epochs to 1 minute to standardize, since many algorithms and previous studies typically present minute-by-minute epochs for actigraphy for easier data analysis and interpretation [22,23]. For the actigraphy data, transformation was done after algorithm results were obtained. For the actigraphy results, once each respective algorithm had processed the data, the actigraphy values from the algorithms were summed to collapse the data to 1-minute epochs. The binarized polysomnography data were also transformed to 1-minute epochs (from 30-second epochs) using the following set of rules. First, if 2 adjacent 30-second epochs were both coded as sleep, the resultant minute would be sleep. Second, if 2 adjacent 30-second epochs were both coded as wake, the resultant minute would be wake. Third, if one of 2 adjacent 30-second epochs were wake, the resultant minute would be wake. Polysomnography transformation to collapse epochs to 1-minute intervals is a common practice for actigraphy comparison [24-27]. With respect to the wake-wins, evaluation of the distribution of pairs was conducted for the polysomnography epochs across all participants. A total of 55.13% of pairs were both asleep, 38.33% were both awake, and only 6.54% were mixed (wake-sleep or sleep-wake). Given the low percentage of mixed pairs, we expected the effect of data transformation to be limited with respect to the wake-wins condition.

All data transformations were again done through Python 3, using a wide range of libraries with coding assistance provided by ChatGPT (OpenAI) to generate Python code for existing Python libraries [28]. ChatGPT was used to generate Python code for file data file preparation and processing (eg, data transformation). Manual data checks in addition to code-based error reporting were conducted to verify the code was working as desired and data were correctly processed. In addition, ChatGPT was used to generate Python code for running statistical analyses. All statistical formulae and analyses were verified within the code. Once code was generated, it was checked by at least one or more of the study authors before use. No machine learning code was generated via ChatGPT, and any actigraphy algorithm code was checked and specified in the Python code exactly as defined by the original algorithm sources. Data were synchronized based on the NSRR and the MESA-provided synchronization guide, which highlighted polysomnography start times relative to actigraphy.

Actigraphy Algorithms

For the application of actigraphy algorithms on the actigraphy data, we used Python (Python 3)-based platform pipelines. For this study, we uploaded raw actigraphy data. This platform allows data processing by simply selecting some basic parameters, that is, the specific algorithm for processing and the files. All files were processed through each algorithm,

XSL•FO
RenderX

respectively. Once the algorithm value was calculated, raw results were collapsed to 1-minute epochs by summing the raw result values. We then applied each algorithm's respective classification threshold to determine whether each epoch was sleep or wake. These thresholds were prespecified by the respective algorithm authors in the original articles.

Algorithms were selected on the basis that they were well established and commonly used throughout the literature for actigraphy analyses across all populations, including with older adults and in populations with sleep disorders [29-36]. However, concrete performance benchmarks have yet to be established. The most used algorithms at the time of this publication are Cole-Kripke (cited 2413 times in Google Scholar) and Sadeh (cited 1795 times in Google Scholar) algorithms. The Cole-Kripke has been used across older adult and sleep disorder populations in several studies [31,34]. The Philips algorithm was selected as the actigraphs used for the MESA actigraphy data collection were Philips Actiwatch, which are also among a commonly used clinical actigraphs [21,22]. Additional algorithms were selected based on the notation that they were adapted from the Cole-Kripke algorithm to be used in various sleep populations and cases with varying parameters. Further, all algorithms had variants that could handle 30-second epoch data or 1-minute epoch data, allowing the use of our dataset and for comparison between algorithms. Generally, the selected algorithms all follow a similar regression-based approach to determine sleep and wake, allowing for comparability and comprising a comprehensive spread of traditionally used actigraphy algorithms. The selected algorithms noted below were evaluated in this study [22].

Cole-Kripke (CK)

Cole et al [6] developed an actigraphy analysis, and to date, this algorithm has been heavily used throughout actigraphy research analysis. The CK algorithm was developed for a variety of epochs, including 30-second and 1-minute epochs [6]. The algorithm is regression based, using a 7-epoch window to compute whether a participant is awake or asleep. The original Cole et al [6], 30-second version of the algorithm used for this study, uses the optimal parameters for a maximum 30-second nonoverlapping epoch of activity per minute.

$$D=0.0001(50A_{-4}+30A_{-3}+14A_{-2}+28A_{-1}+12lA_{0}+8A_{+1}+50A_{+2})$$

Here, if D<1, the epoch is scored as sleep, while if D>=1, the epoch is scored as wake. The 0.001 represents the scale factor for the entire equation, while the numerical values assigned to each epoch represent the weighting factors for the present, previous, and following epochs. A represents an epoch, where A_{-4} to A_{-1} represent the four preceding epochs to the current epoch (A0) and A_{+1} to A_{+2} represent the following 2 epochs.

University of California, San Diego Scoring Algorithm (UCSD)

Developed by Jean-Louis et al [9], this actigraphy algorithm is similar to the Cole-Kripke algorithm, with the only difference being the weights on the epoch. The UCSD algorithm was also designed for minute-by-minute epochs. However, the authors noted that two 30-second epochs were counted as 1 minute to match their polysomnography data recording rates. In our case, we gave this algorithm the original 30-second epoch data, treating each epoch in the formula as is. The UCSD algorithm used in this study was as follows:

$$D=0.05(0.010A_{-4}+0.015A_{-3}+0.028A_{-2}+0.031A_{-1}+0.085A_{0}+0.015A_{+1}+.010A_{+2})$$

where, if D<1, the epoch is scored as sleep, while if D≥1, the epoch is scored as wake. According to previous analyses of Actillume data for healthy young adults, the optimal scaling factor, *P*, was .05 [9]. The numerical values assigned to each epoch represent the weighting factors for the present, previous, and following epochs. A represents an epoch, where A_{-4} to A_{-1} represent the four preceding epochs to the current epoch (A_0) and A_{+1} to A_{+2} represent the following two epochs.

Kripke 2010 (K2010)

Developed by Kripke et al and applied through Microsoft Excel Visual Basic macro, this algorithm aimed to set optimal parameters for a sleep, wake scoring algorithm to score each epoch. This algorithm is similar to the aforementioned algorithms [10]. However, the optimal algorithm accounts for the activity counts of 13 30-second epochs; the 10 epochs preceding and 2 epochs proceeding the epoch being scored (X0) and differs in the weights assigned to each epoch. The optimal algorithm by Kripke et al [10] used in this study was as follows:

$$D=0.30(0.064X_{-10}+0.074X_{-9}+0.112X_{-8}+0.112X_{-7}+0.118X_{-6}+0.118X_{-5}+0.112X_{-4}+0.118X_{-3}+0.118X_{-2}+0.064X_{-1}+0.030X_{0}+0.112X_{+1}+0.10X_{+2})$$

where D was the scaled polynomial sum of activity scores for 13 30-second epochs. The 0.30 represents the optimal overall scaling parameter. The numerical value attached to each epoch (X) represents a scaling parameter for each respective corresponding epoch. When D<1, the epoch being scored (X_0) is scored as sleep, while if D≥1 according to this algorithm.

Philips-Respironics (Philips)

This algorithm is designed for Philips-Respironics and previously Mini-Mitter Co. Inc. devices known as Actiwatch [37,38]. This algorithm is designed to handle data from the Actiwatch monitors, which measure activity levels in several epoch lengths of 15 seconds, 30 seconds, 1 minute, or 2 minutes. In our study, we used the 30-second version of this algorithm. Similar to the aforementioned algorithms, this also applies weights to 5 epochs, the 2 preceding and 2 proceeding epochs, and the epoch being scored. The algorithm used in this study is as follows:

$$A=0.04E_{-4}+0.04E_{-3}+0.2E_{-2}+0.2E_{-1}+2E_{0}+0.2E_{+1}+.2E_{+2}+.04E_{+3}+.04E_{+2}$$

Based on activity counts measured through the device, a total activity value (A) is generated for each epoch. The total activity value is then evaluated against the wake threshold value (20 [low], 40 [medium], 80 [high]), which, in the Actiwatch software, is automatically generated based on the activity data of individual cases or a custom value selected by the user [39]. If the total activity value is less than or equal to the wake threshold value, the epoch is scored as sleep. That is, if A>T, the epoch is scored as wake; otherwise, if A ≤ T, the epoch is scored as sleep. The En represented the activity counts of the

XSL•FO
RenderX

previous, successive, or scored epoch. For our purposes, we evaluated each threshold after processing.

Sadeh

Sadeh et al [11] developed an algorithm based on concurrent polysomnography and a wrist actigraph (Ambulatory Monitoring, Ardsley). The algorithm features a discriminant function using 5 calculated variables on an 11-minute window (the 5 preceding, 5 proceeding, and the scored epochs), centered on the epoch being scored [11]. Any missing epochs are considered 0 to avoid infinity problems. This happens if the current epoch is at the beginning or end of a dataset. The Sadeh algorithm uses the y-axis epoch data. If any of the epoch counts are over 300, it reduces them to 300. The original formula for this algorithm uses a 1-minute epoch window; in this study, we gave this algorithm the 30-second epoch raw actigraphy data. Each 30-second epoch was evaluated as a 1-minute epoch. This would preserve the comparison between algorithms, as all of them processed the 30-second raw actigraphy data. The original Sadeh algorithm used in this study is as follows:

PS = (7.601 - [0.065 * AVG] - [1.08 * NATS] - [0.056 * SD] - [0.703 * LG])

where AVG is the arithmetic mean (average) of the activity counts for the window, NATS is the number of epochs that have counts ≥50 and <100, SD is the standard deviation for the first 6 epochs of the window, and LG is the natural (base e) logarithm of the current epoch. Post scoring of the epoch, if the resultant value, referred to as probability of sleep (PS), is ≥ 0, the epoch is scored as sleep. Based on the original paper, typically, if the result is > −4, then the current epoch is considered asleep.

Webster's Rescoring Rules (RS)

Notably, traditional actigraphy algorithms often incorrectly score periods of wake as sleep or rest. To counter this issue, Webster et al [40] developed a set of rescoring rules to apply after initial classification using a scoring algorithm [6,22,40]. These results have been well cited in the literature and given their relevance in increasing algorithm accuracy, they were an important consideration to evaluate within this study [22]. These were applied in unison with the algorithms and compared with nonrescored results. Once the data were processed with each respective algorithm, the raw activity scores were binarized, reflecting sleep (1) or wake (0) using optimal thresholds (see data analysis), and rescoring was sequentially applied to the binarized data. Nonrescored results were denoted as NRS, and rescored results were as RS. The rescoring rules were as follows:

1. After at least 4 minutes scored as wake, the next 1 minute scored as sleep is rescored as wake
2. After at least 10 minutes scored as wake, the next 3 minutes scored as sleep are rescored as wake
3. After at least 15 minutes scored as wake, the next 4 minutes scored as sleep are rescored as wake
4. If 6 minutes or less are scored as sleep surrounded by at least 10 minutes (before and after) scored as wake are rescored as wake
5. If 10 minutes or less are scored as sleep surrounded by at least 20 minutes (before and after) scored as wake, they are rescored as wake

Statistical Analysis

Epoch-by-Epoch Comparison

Once the data were processed with each respective algorithm, the raw activity scores were examined. We conducted an epoch-by-epoch comparison of sleep-wake for each algorithm against the corresponding ground truth polysomnography measurements. In our analysis, the polysomnography classification was considered the actual class, while the actigraphy classification was considered the predicted class (with wake=0 and sleep=1). We used confusion matrix analyses to assess key metrics accuracy, sensitivity, specificity, precision, and F_1-score (a measure of an algorithm's predictive performance of sleep and wake). A complete definition of metrics is provided in Multimedia Appendix 2.

Epoch-by-epoch comparison was conducted for each algorithm based on the original sleep-wake thresholds defined in the respective papers (see Actigraphy Algorithm Section). The Philips-Respironics algorithm has 3 suggested threshold values (20 [low], 40 [medium], 80 [high]), all of which were examined [39]. In addition, weighted mean and SD results were generated for each algorithm both with and without rescoring.

Cohen κ and Matthews Correlation Coefficient

As measures of comparison between sleep algorithms and polysomnography, Cohen κ and MCC were calculated for each algorithm. As with the confusion matrix statistics, the sleep algorithm results (predicted) were evaluated in comparison to the ground truth polysomnography (actual). Cohen κ is a measure of agreement between raters for nominal or categorical data with adjustments for chance agreement [41,42]. For this study, the raters were individual sleep algorithms and polysomnography. MCC was also used to evaluate the correlation (performance) between binary classification, such as in our case between sleep algorithms and polysomnography [43,44]. A complete definition of metrics is provided in Multimedia Appendix 2.

Repeated Measures ANOVA With Post Hoc Examination

As the same sample of participants was assessed by each algorithm, repeated measures ANOVA was conducted to evaluate whether there were respective differences between sleep algorithm results for Cohen κ, MCC, and confusion matrix. The assumption of sphericity was violated across all results based on significant Mauchly tests and low epsilon values. Therefore, Greenhouse-Geisser corrections were applied, and adjusted degrees of freedom and P values were evaluated. For significant ANOVA results, subsequent post hoc analyses were conducted with Bonferroni correction (controlling for type 1 error) to evaluate which algorithm pairs demonstrated significant differences and their effect sizes (Hedges g) [45,46]. For Hedges g interpretation, the range of 0.2, 0.5, 0.8 for small, medium, and large effect sizes was considered.

Receiver Operating Characteristic Curve and Area Under the Curve (ROC; AUC)

To comparably evaluate the ability of each algorithm to classify sleep-wake, we evaluated the area under the receiver operating characteristic curves (AUC and ROC), respectively. AUC

XSL•FO
RenderX

provides information about algorithm performance (ie, how much is each algorithm capable of distinguishing sleep vs wake) as a measure of global accuracy and robustness. An ROC curve is generated by plotting the sensitivity (also known as the true positive rate; TPR) against 1 - specificity (also known as the false positive rate; FPR) at various cut-off thresholds. We used raw values for actigraphy and binary ground truth (polysomnography) to generate the ROC curve. Using the s*cikit-learn* Python library, an automatic extensive range of thresholds was covered for the ROC curve. Both ROC curves and AUC were generated for each algorithm. Given that the original algorithms reported a fixed, predetermined threshold for sleep-wake, the optimal cut-off point was not examined in the ROC curves. However, the sensitivity and specificity for each algorithm relative to their sleep-wake threshold were reported in the epoch-by-epoch results. A complete definition of metrics is provided in Multimedia Appendix 2.

Rescored Algorithm Analysis

To evaluate whether applying Webster et al [40] rescoring rules improved algorithm performance in contrast to nonrescored actigraphy algorithms, we applied the rescoring rules to each algorithm [40]. To do so, we first binarized (sleep [1], wake [0]) the raw activity values generated by each algorithm using the optimal thresholds determined in the ROC analyses. We then applied the rescoring rules to the binary data. Finally, we calculated the Cohen κ, MCC, and all confusion matrix results for the rescored algorithms, respectively. These results were compared to the nonrescored results.

Sleep Metrics Comparison

We examined each algorithm's agreement with polysomnography on several commonly studied sleep metrics, including TST, SE, and WASO. For the polysomnography, we evaluated the MESA and NSRR-provided data for each sleep metric, while for the actigraphy data, once each dataset was processed by each algorithm and binarized, we calculated each sleep metric for both the nonrescored and rescored data. We used the MESA and NSRR definitions to calculate the respective sleep metrics. To do so, we applied the following calculations:

1. TST=the total number of minutes of sleep from the first sleep onset time to sleep offset time.
2. SE=TST/in-bed (results presented as %).
   A. In-bed=synchronized to polysomnography in-bed start time, the interval between lights out and in-bed time versus lights off, out-bed, or wake time.
3. WASO=the total duration of wake between first sleep onset time and sleep offset time.

These sleep metrics were calculated for each participant and each algorithm, respectively. For comparison of agreement, we used Bland-Altman distributions [47]. Bland-Altman distributions are used to visualize the difference between actigraphy and polysomnography. We evaluated the average difference between corresponding measurements (actigraphy and polysomnography) for each respective sleep metric. Through these plots, we examined the distribution of differences between actigraphy and polysomnography, that is, the consistency between the 2. In addition, we examined the mean difference (actigraphy-polysomnography), bias, limits of agreement (LoA), and proportional bias (across values; regression) between actigraphy and polysomnography. The Bland-Altman distribution analysis was also conducted in Python using the pyCompare library with code assistance provided by ChatGPT [28].

Ethical Considerations

The current study was conducted on a pre-established MESA dataset on a secondary use basis [20,21]. The original MESA study obtained institutional review board ethics approval for each study site of data collection and obtained written consent from all participants [20]. The NSRR resource maintains a detailed approval procedure for dataset acquisition and use [21]. The study dataset was acquired through a data access and use agreement with the NSRR [21]. All data were predeidentified for privacy by NSRR. Ethics approval was acquired from the University of Toronto (REB protocol # 35344) for this study.

Results

The sample demographics results are presented in Table 1. Notably, 12.3% (n=177, includes Apnea, Insomnia RLS, and CPAP subgroup) of sample participants reported some form of sleep problem. Additionally, measures of hypertension, diabetes, and other disorders were not reported in the NSRR version of the MESA dataset. However, previous studies note the prevalence of health issues or health risks within this study sample [20,48,49].

The epoch-by-epoch comparison for nonrescored algorithms demonstrated that the Kripke 2010 algorithm (mean 0.80, SD 0.09) had the highest accuracy, followed by Cole-Kripke (mean 0.80, SD 0.09), Philips-Respironics (mean 0.80, SD 0.08; mean 0.80, SD 0.09; and mean 0.79, SD 0.09; for 20, 40, and 80 thresholds), UCSD (mean 0.78, SD 0.10), and Sadeh (mean 0.78, SD 0.10) respectively (Figure 2). An important note is that there were not any large differences between traditional algorithms for accuracy, ranging only 0.01 - 0.02. F_1-scores for the nonrescored algorithm results were all >0.80. Results for Cohen κ demonstrated moderate agreement (0.4 - 0.6) between actigraphy algorithms and polysomnography. Similarly, MCC results showed moderate positive correlations (range 0.49 - 0.58) between actigraphy algorithms and polysomnography for all algorithms. Complete results for Cohen κ, MCC, accuracy, sensitivity, specificity, precision, and F_1-scores are presented in Figure 2.

XSL•FO
RenderX

Figure 2. Cohen κ, Matthews correlation coefficient, and confusion matrix heatmap results for nonrescored algorithms. Accuracy, sensitivity, specificity, precision, F_1-score: mean (SD). Color tone represents index value. MCC: Matthews correlation coefficient; UCSD: University of California San Diego.

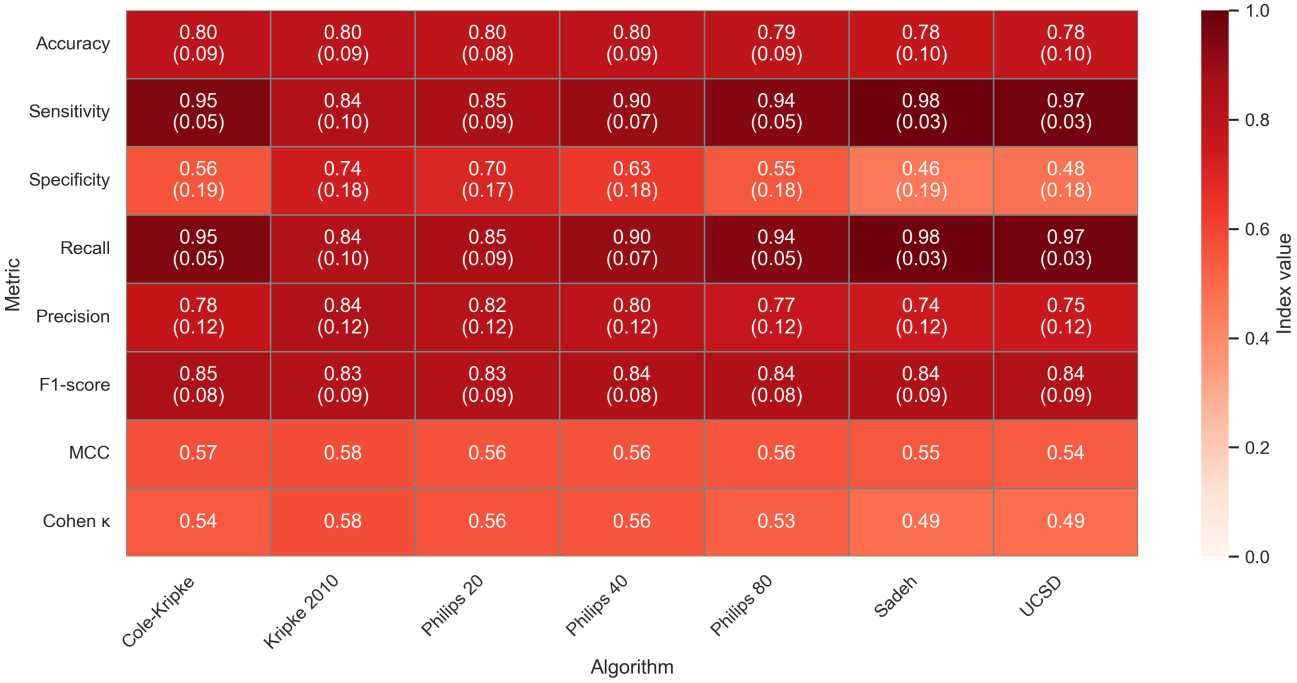

Examination of rescored algorithm results revealed marginal changes observed in accuracy. Rescored results demonstrated the Cole-Kripke (mean 0.81, SD 0.09) had the highest accuracy followed by Philips-Respironics (mean 0.80, SD 0.09; mean 0.81, SD 0.09; and mean 0.81 SD 0.09; for 20, 40, and 80 thresholds), Sadeh (mean 0.80, SD 0.10), UCSD (mean 0.80, SD 0.09); and Kripke 2010 (mean 0.79 SD 0.09) respectively (Figure 3). Again, there were no large differences between traditional algorithms for ACC, ranging only 0.01 - 0.03. Rescored results for F_1-scores either demonstrated no change or marginally changed by 0.01 - 0.02 for some algorithms, with all still >0.80. Results for Cohen κ again demonstrated minor changes remaining within the moderate agreement range between actigraphy algorithms and polysomnography. Similarly, MCC results only showed minor changes again, showcasing moderate positive correlations (range 0.54 - 0.60) between actigraphy algorithms and polysomnography for all algorithms. Complete rescored results for Cohen κ, MCC, accuracy, sensitivity, specificity, precision, and F_1-scores are presented in Figure 3. Complete confusion matrix statistical results for both nonrescored and rescored algorithms are provided in Multimedia Appendix 3.

Figure 3. Cohen κ, Matthews correlation coefficient, and confusion matrix heatmap results for rescored algorithms. Accuracy, sensitivity, specificity, precision, F_1-score: mean (SD). Color tone represents index value. MCC: Matthews correlation coefficient; UCSD: University of California San Diego.

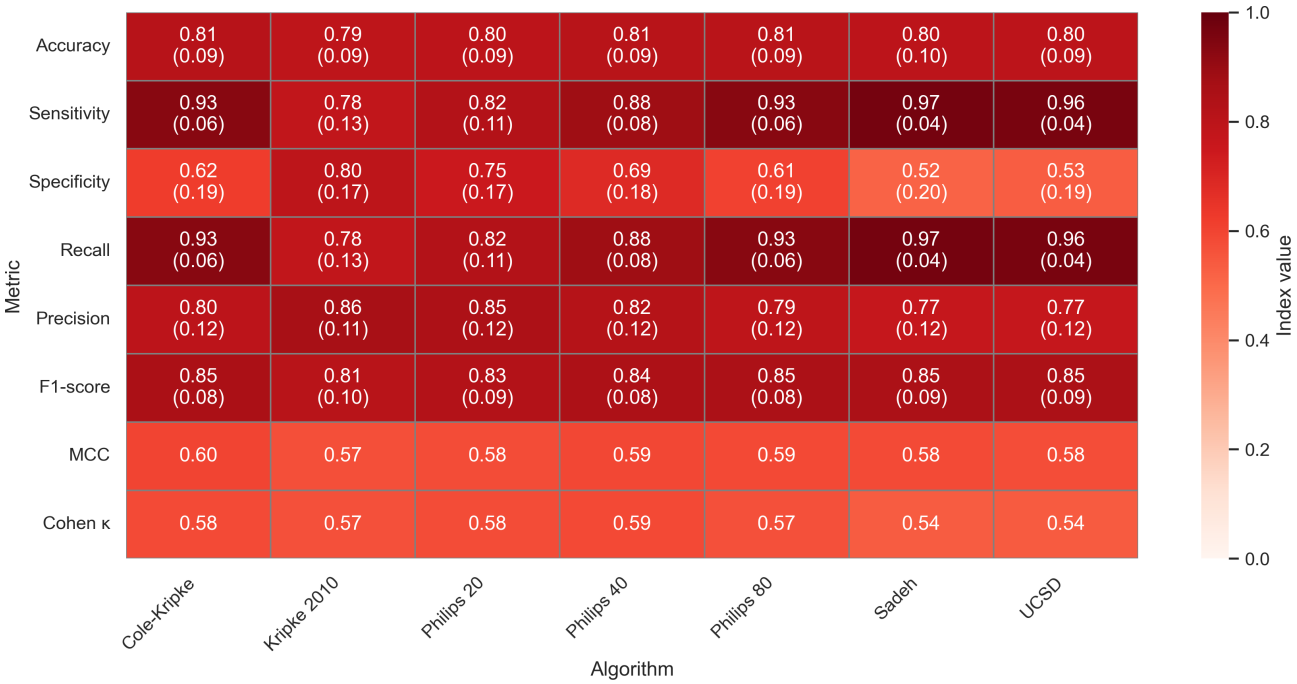

XSL•FO
RenderX

Repeated measures ANOVA testing revealed significant differences between algorithms across all metrics (κ, MCC, accuracy, sensitivity, specificity, precision, and F_1-score). Complete ANOVA results are presented in Tables 2 and 3 for nonrescored and rescored algorithms, respectively. Subsequent post hoc analyses also revealed significant differences between algorithm pairs across all metrics (Tables 4 and 5).

Table . Repeated measures ANOVA for confusion matrix across nonrescored algorithms.

Metric	SS[a]	MS[b]	*F* test	*df*	η²g[c]	Eps (ε)	*P* value (uncorr.)	df[d] (adjusted)	*P* value (adjusted)
Accuracy			78.14		0.01	0.19	<.001		<.001
$\text{Accuracy}_{\text{algorithm}}$	0.50	0.08		6				1.15	
$\text{Accuracy}_{\text{error}}$	9.22	0.00		8628				1656.37	
Sensitivity			3754.96		0.39	0.20	<.001		<.001
$\text{Sensitivity}_{\text{algorithm}}$	27.60	4.60		6				1.19	
$\text{Sensitivity}_{\text{error}}$	10.55	0.00		8616				1711.96	
Specificity			8112.40		0.23	0.22	<.001		<.001
$\text{Specificity}_{\text{algorithm}}$	99.06	16.51		6				1.34	
$\text{Specificity}_{\text{error}}$	17.56	0.00		8628				1920.73	
Precision			3435.68		0.07	0.20	<.001		<.001
$\text{Precision}_{\text{algorithm}}$	10.55	1.76		6				1.22	
$\text{Precision}_{\text{error}}$	4.42	0.00		8628				1751.47	
F_1-score			93.31		0.00	0.20	<.001		<.001
$F_1\text{-score}_{\text{algorithm}}$	0.36	0.06		6				1.17	
$F_1\text{-score}_{\text{error}}$	5.58	0.00		8628				1688.29	

[a]SS: sum of squares.

[b]MS: mean squares.

[c]η²g: general eta squared

[d]Adjusted values for *df* and *P* values refer to respective Greenhouse-Geisser corrections.

XSL•FO
RenderX

Table . Repeated measures ANOVA for confusion matrix across nonrescored algorithms.

Metric	SS[a]	MS[b]	*F* test	*df*	η^2g[c]	Eps (ε)	$P_{uncorr.}$ value	$df_{adjusted}$[d]	$P_{adjusted}$ value
Accuracy			64.62		0.01	0.20	<.001		<.001
$Accuracy_{algorithm}$	0.56	0.09		6				1.22	
$Accuracy_{error}$	12.56	0.00		8628				1754	
Sensitivity			3906.50		0.40	0.21	<.001		<.001
$Sensitivity_{algorithm}$	44.24	7.37		6				1.24	
$Sensitivity_{error}$	16.26	0.00		8616				1782.96	
Specificity			5961.01		0.22	0.26	<.001		<.001
$Specificity_{algorithm}$	97.56	16.26		6				1.55	
$Specificity_{error}$	23.54	0.00		8628				2227.53	
Precision			2706.59		0.08	0.23	<.001		<.001
$Precision_{algorithm}$	11.65	1.94		6				1.37	
$Precision_{error}$	6.19	0.00		8628				1974.34	
F_1-score			341.85		0.03	0.20	<.001		<.001
F_1-$score_{algorithm}$	2.20	0.37		6				1.22	
F_1-$score_{error}$	9.26	0.00		8628				1759.31	

[a]SS: sum of squares.
[b]MS: mean squares.
[c]η^2g: general eta squared.
[d]Adjusted values for df and *P* values refer to respective Greenhouse-Geisser corrections.

Table . Repeated measures ANOVA for Matthews correlation coefficient and Cohen κ across nonrescored algorithms.

Metric	SS[a]	MS[b]	*F* test	*df*	η^2g[c]	Eps (ε)	$P_{uncorr.}$ value	$df_{adjusted}$[d]	$P_{adjusted}$[d] value
MCC			138.47		0.01	0.23	<.001		<.001
$MCC_{algorithm}$	2.01	0.34		6				1.38	
MCC_{error}	20.91	0.00		8628				1988.02	
K			450.81		0.02	0.21	<.001		<.001
$K_{algorithm}$	9.26	1.54		6				1.27	
K_{error}	29.53	0.00		8628				1828.56	

[a]SS : sum of squares.
[b]MS: mean squares.
[c]η^2g: general eta squared.
[d]Adjusted values for df and *P* values refer to respective Greenhouse-Geisser corrections.

XSL•FO
RenderX

Table . Repeated measures ANOVA for Matthews correlation coefficient and Cohen κ across nonrescored algorithms.

Metric	SS[a]	MS[b]	*F* test	*df*	η^2g[c]	Eps (ε)	$P_{uncorr.}$ value	$df_{adjusted}$[d]	$P_{adjusted}$[d] value
MCC			37.17		0.00	0.25	<.001		<.001
$MCC_{algorithm}$	0.72	0.12		6				1.53	
MCC_{error}	27.85	0.00		8628				2196.75	
K			112.84		0.01	0.23	*P*<.001		<.001
$K_{algorithm}$	3.21	0.53		6				1.38	
K_{error}	40.86	0.00		8628				1977.95	

[a]SS: sum of squares.

[b]MS: mean squares.

[c]η^2g: general eta squared.

[d]Adjusted values for df and *P* values refer to respective Greenhouse-Geisser corrections.

Post hoc analyses for MCC were significant; however, all effect sizes were very small to small (all Hedges g<0.3). Post hoc results for Cohen κ demonstrated small to medium effects with effect sizes ranging from 0 to <0.5. Complete post hoc analysis results are presented in Tables 4 and 5. Notably, modest differences were observed between K2010 and UCSD (Hedges g=0.42) and K2010 and Sadeh (Hedges g=0.44), indicating better predictive performance of the K2010 algorithm, respectively. Similarly, significant post hoc results for rescored algorithms for both MCC and Cohen κ showed very small to small effect sizes across all algorithm contrasts (all Hedges g<0.25). Detailed results are provided in Multimedia Appendix 4.

Post hoc analyses for nonrescored algorithms' confusion matrix metrics, particularly accuracy and F_1-scores, demonstrated both nonsignificant and many significant contrasts (Multimedia Appendix 5). Significant pairs, however, had very small to small effect sizes (Hedges g) ranging between 0 - 0.3 for both accuracy and F_1-scores (Multimedia Appendix 5). Post hoc results for sensitivity, specificity, and precision varied greatly, ranging from very small to large effect between contrasts, revealing nuanced performance differences (Multimedia Appendix 5). For detailed post hoc results, see Multimedia Appendix 5.

After rescoring, post hoc results for confusion matrix metrics demonstrated similar very small to small effect sizes for accuracy. However, there were some medium effect sizes for F_1-scores. Notably, modest differences were observed across K2010 predictive performance, which was worse than the CK (Hedges g=−0.46), Sadeh (Hedges g=−0.42), UCSD (Hedges g=−0.40), and Philips algorithm with a threshold of 80 (Hedges g=−0.42). Similarly, post hoc results for sensitivity, specificity, and precision varied greatly, ranging from very small to large effects between contrasts, revealing nuanced performance differences (Multimedia Appendix 5). For full post hoc results, see Multimedia Appendix 5.

The ROC analysis revealed that the Kripke 2010 had the highest AUC (0.86) followed by Sadeh (0.85), Philips-Respironics (0.84), Cole-Kripke (0.84), and UCSD (0.84). Across the board, all algorithms demonstrated excellent ability to discriminate sleep-wake (excellent performance). The ROC curves are presented with their respective AUC in Figure 4.

XSL•FO
RenderX

Figure 4. Algorithm receiver operating characteristic curves and area under the curve. Receiver operating characteristics curves and respective area under the curve listed for each algorithm. AUC: area under the curve; ROC: receiver operating characteristics curve;

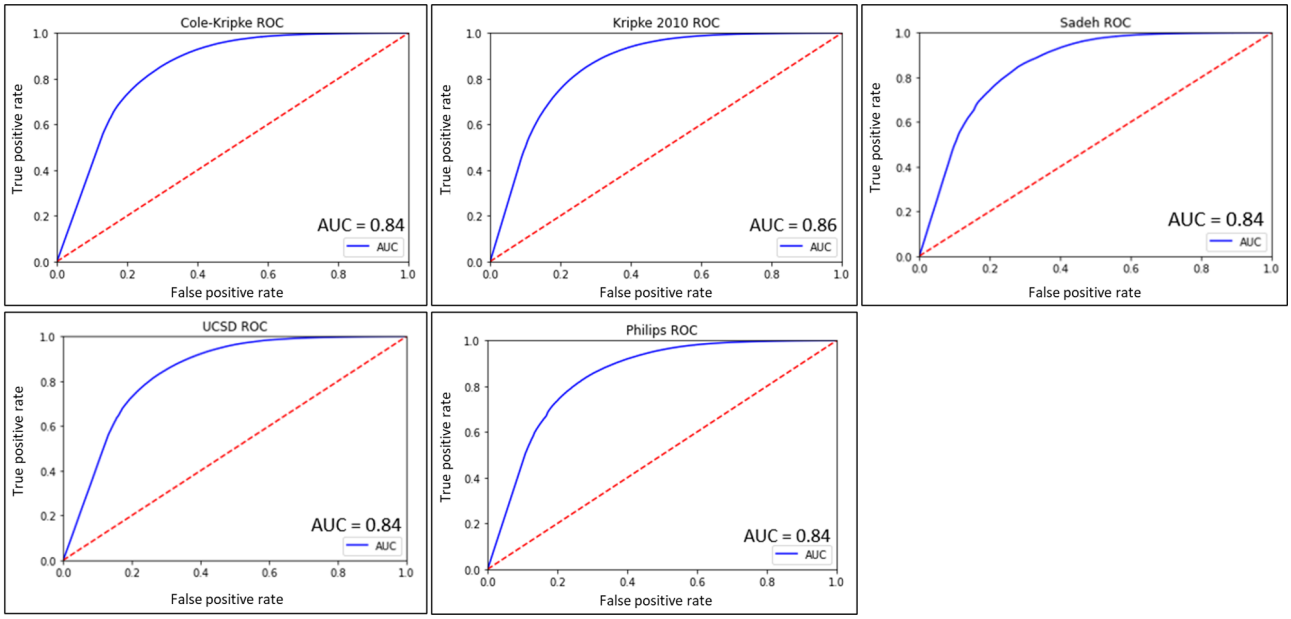

Examination of agreement for TST through Bland-Altman distributions demonstrated significant levels of mean difference and systematic bias across all algorithms (Figure 5; Multimedia Appendix 6; Table 1). This may be due to the sample or due to clear outliers. These outliers may also have caused discrepancies within the distribution and mean difference. In addition, supplementary regression analyses for proportional bias were significant across nonrescored Cole-Kripke, UCSD, Sadeh, and Philips threshold 80 algorithms. Similarly, rescored Kripke 2010, UCSD, Sadeh, Philips threshold 20, and Philips threshold 80 algorithms also demonstrated significant proportional bias. Although regression tests were significant, both the regression slopes and R^2 values were extremely small, ranging from −0.06 to 0.10 for slope and R^2 values all ≤0.01. Although proportional bias regression analyses were significant for a number of algorithms for TST, both slopes and R^2 were extremely small, and plot distributions were relatively evenly spread. Therefore, effects of proportional bias may be minimal or significant due to sample size. Detailed proportional bias regression statistics are reported in Multimedia Appendix 6, Table 1.

Figure 5. Bland-Altman distributions total sleep time for nonrescored algorithms. Plots represent the Bland-Altman distributions for nonrescored algorithms of actigraphy in comparison to polysomnography (ground truth) for total sleep time. TST: total sleep time.

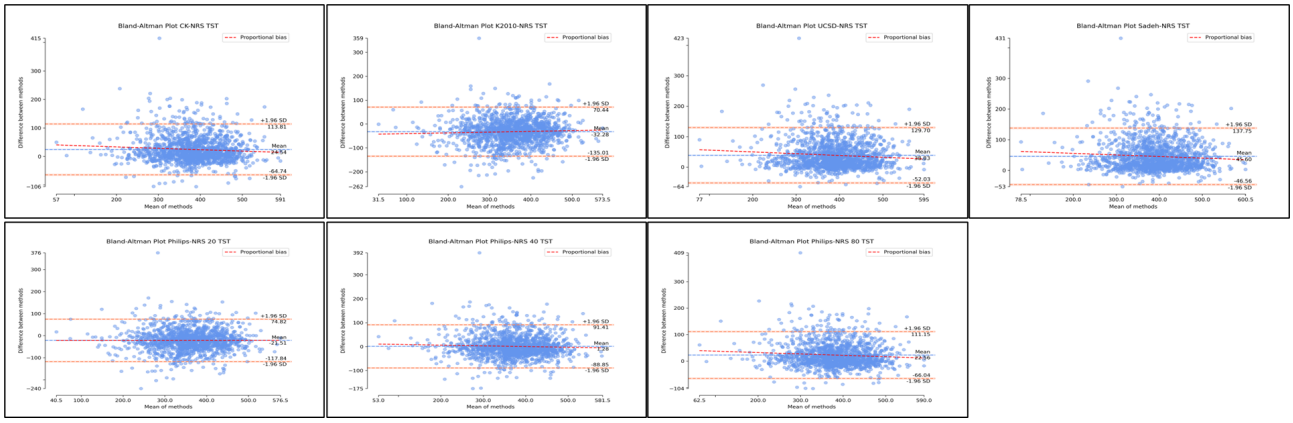

The best results for TST agreement with polysomnography across all algorithms were obtained with the Philips algorithm with the threshold 40 (nonrescored: mean difference [MD] 1.28, SD −46.00; 95% LoA −88.88 to 91.44; Figure 5; Multimedia Appendix 6; Table 1). These results were considerably decreased with rescoring; however, the Philips algorithm with the threshold 40 retained the best agreement (rescored: MD −9.67, SD –47.64; 95% LoA −103.05 to 83.71; Figure 5; Multimedia Appendix 6; Table 1). For TST, the Cole-Kripke, UCSD, and Sadeh algorithms displayed higher mean differences in TST measurements by actigraphy (Figure 5; Multimedia Appendix 6; Table 1). This overestimation of TST was reduced with rescoring (Figure 6; Multimedia Appendix 6; Table 1). Conversely, the Kripke 2010 algorithm underestimated the TST, and the magnitude of underestimation increased with rescoring. For the Philips algorithm, the results varied based on which threshold was used. The Philips algorithm with an 80 threshold overestimated TST when compared to polysomnography; however, this improved with rescoring. Finally, the Philips algorithm with 20 thresholds demonstrated considerable underestimation for TST, which further worsened with rescoring.

XSL•FO
RenderX

Figure 6. Bland-Altman distributions total sleep time for rescored algorithms. Plots represent the Bland-Altman distributions for rescored algorithms of actigraphy in comparison to polysomnography (ground truth) for total sleep time. TST: total sleep time.

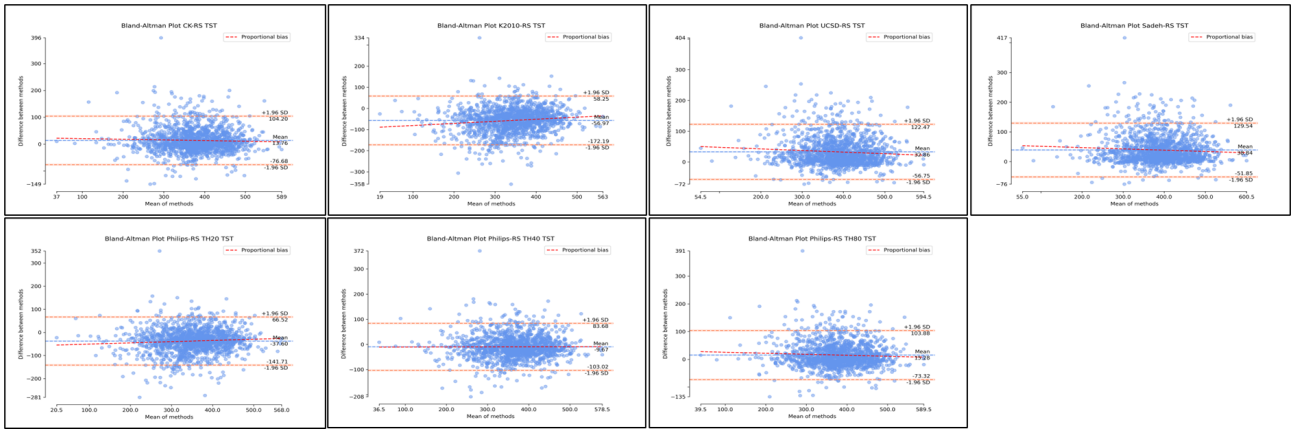

With respect to SE, all algorithms demonstrated similar distributions where estimates of SE were worse at lower levels of SE (Figure 7; Multimedia Appendix 6; Table 2). Considerable improvements in agreement were observed as SE levels increased. Based on the Bland-Altman plots, the distributions demonstrate some systematic bias and heteroscedasticity (Figure 7; Multimedia Appendix 6; Table 2). Supplementary regression analyses for proportional bias were significant across all nonrescored and rescored algorithms with the exception of nonrescored Kripke 2010 (P=.86). Regression slopes covered a wide range and were all negative with the exception of rescored Kripke 2010 (slope=0.15). Therefore, indicating a directional bias as SE levels change (in particular, increased). However, R^2 values were relatively small, with all values ≤0.21. Detailed proportional bias regression statistics are reported in Multimedia Appendix 6; Table 2.

Figure 7. Bland-Altman distributions sleep efficiency for nonrescored algorithms. Plots represent the Bland-Altman distributions for nonrescored algorithms of actigraphy in comparison to polysomnography (ground truth) for sleep efficiency. SE: sleep efficiency.

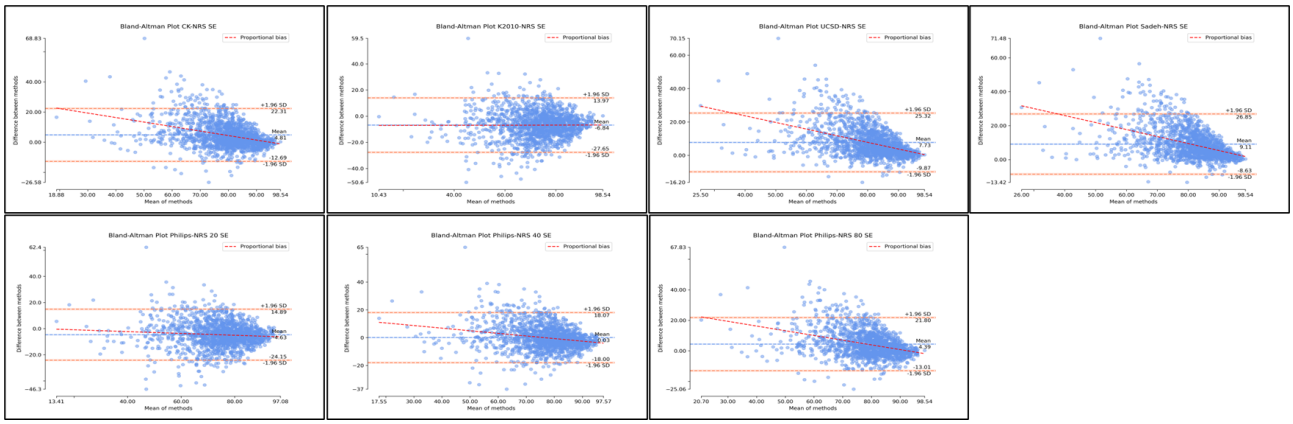

The Philips algorithm with a threshold at 40 demonstrated the best agreement with polysomnography (nonrescored: MD 0.03, SD −9.20; 95% LoA −18 to 18.07). Rescoring of this resulted in underestimation of SE (rescored: MD −2.20, SD −9.61; 95% LoA −21.03 to 16.63). The Cole-Kripke, UCSD, Sadeh, and Philips threshold 80 algorithms all demonstrated overestimation of SE, with some improvement with rescoring (Figure 8; Multimedia Appendix 6; Table 2). While the Kripke 2010 and Philips threshold 20 algorithms both demonstrated underestimation of SE, with rescoring further increasing the MD, that is, the magnitude of underestimation (Figure 8; Multimedia Appendix 6; Table 2).

XSL•FO
RenderX

Figure 8. Bland-Altman distributions sleep for rescored algorithms. Plots represent the Bland-Altman distributions for rescored algorithms of actigraphy in comparison to polysomnography (ground truth) for sleep efficiency. SE: sleep efficiency.

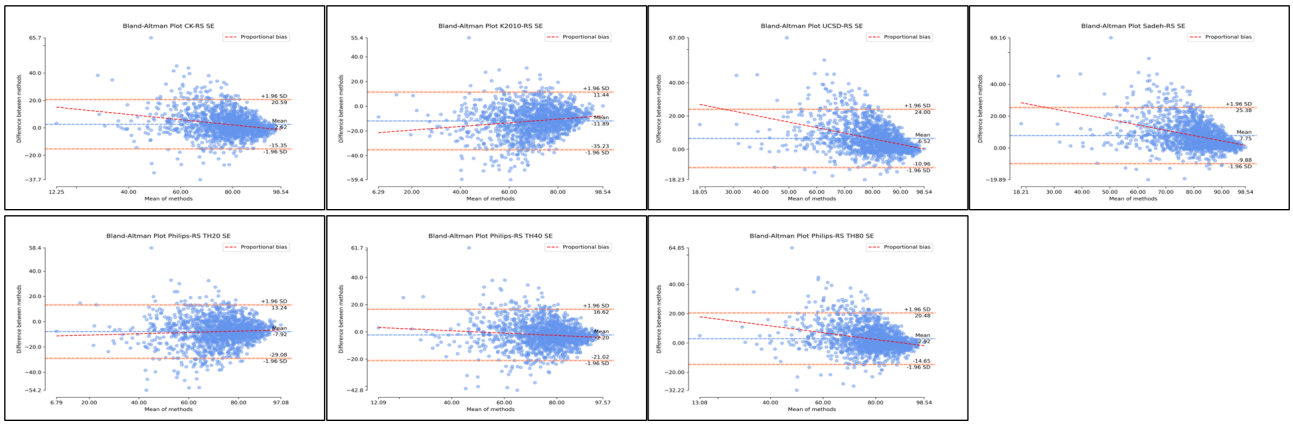

Finally, the Bland-Altman distributions for WASO demonstrated considerable underestimation. The distributions demonstrate considerable bias and heteroscedasticity, where WASO estimates are better at lower WASO averages and become considerably worse as WASO increases (Figure 9; Multimedia Appendix 6; Table 3). Again, supplementary regression analyses for proportional bias were significant across all nonrescored and rescored algorithms. Regression slopes covered a wide range and were all negative with the exception of rescored Kripke 2010 (slope=0.08). Therefore, indicating a directional bias as WASO levels change (in particular, increased). R^2 values covered a wide range. Detailed proportional bias regression statistics are reported in Multimedia Appendix 6, Table 3.

Figure 9. Bland-Altman distributions wake after sleep onset for nonrescored algorithms. Plots represent the Bland-Altman distributions for nonrescored algorithms of actigraphy in comparison to polysomnography (ground truth) for wake after sleep onset. WASO: wake after sleep onset.

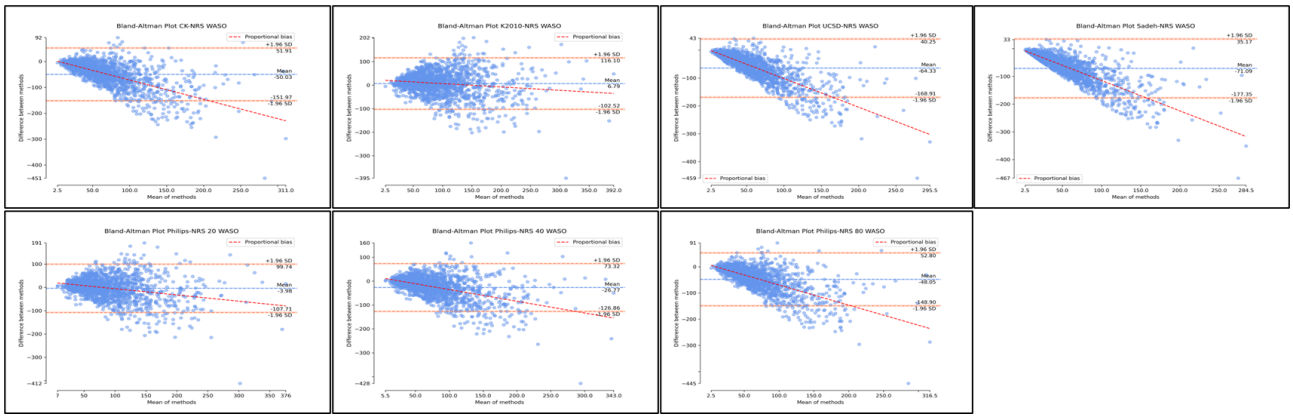

Among all the algorithms, the Philips threshold 20 algorithm had the best agreement with polysomnography for WASO, only underestimating by a small magnitude (nonrescored: MD=−3.98, SD −52.94; 95% LoA −107.74 to 99.77). Rescoring of this algorithm diminished the agreement of this algorithm to overestimate WASO (rescored: MD 12.10, SD −56.07; 95% LoA −97.80 to 122.01). With respect to the Philips algorithms at thresholds 40 and 80, both underestimated WASO. Rescoring lowered the magnitude of underestimations, improving agreement. Similarly, Coke-Kripke, UCSD, and Sadeh had large underestimations of WASO with some improvement after rescoring (Figure 10; Multimedia Appendix 6; Table 3) Kripke 2010, however, only slightly overestimated WASO, and rescoring considerably increased the magnitude of overestimation.

XSL•FO
RenderX

Figure 10. Bland-Altman distributions wake after sleep onset for rescored algorithms. Plots represent the Bland-Altman distributions for rescored algorithms of actigraphy in comparison to polysomnography (ground truth) for wake after sleep onset. RS: rescored algorithms; WASO: wake after sleep onset.

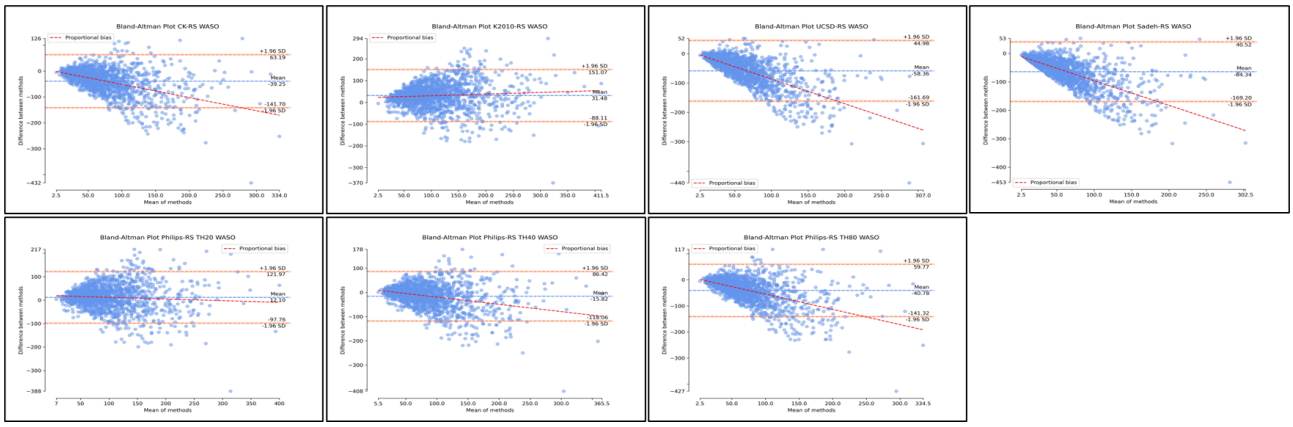

Exploratory subgroup analysis was conducted for 4 sleep problem groups, including apnea, individuals who used CPAP, insomnia, and RLS. For the subgroup analyses, we again examined AUC, accuracy metrics, and Bland-Altman distributions for agreement. The AUC was above 0.80 across all subgroups and algorithms, similar to the results across the entire population (Multimedia Appendix 7). This indicated all algorithms demonstrated excellent ability to discriminate sleep-wake regardless of the presence of individuals' sleep problems.

Accuracy ranged between 76% and 81% across all algorithms and subgroups, with only minor differences of 1% - 4% between algorithms. Rescoring only resulted in minor improvements in algorithm accuracy of 1% - 3%. F_1-scores demonstrated good balance, ranging from 0.80 - 0.84. After rescoring, F_1-scores remained relatively the same with marginal changes (0.01 - 0.03) again remaining within the ok-to-good balance range of 0.79 to 0.85. Overall, accuracy and F_1-score results were again highly similar to the results observed across the entire population sample. Complete subgroup results for accuracy, sensitivity, specificity, precision, and F_1-score are presented in Multimedia Appendix 8.

Along these lines, MCC and Cohen κ results were extremely similar to the overall results, ranging between 0.40 - 0.60 for both metrics and for both rescored and nonrescored algorithms. Therefore, all results demonstrated moderate positive correlations for MCC and moderate Cohen κ agreement across algorithms.

Repeated measures ANOVA analyses for accuracy were only significant for the apnea rescored algorithms ($F_{1.20,124.08}$=5.27; ε=0.20; P=.02), RLS rescored algorithms ($F_{1.24,80.71}$=5.70; ε=0.20; P=.01), and insomnia nonrescored algorithms ($F_{1.17,96.97}$=3.85; ε=0.19; P=.05). All other nonaccuracy metrics demonstrated significant ANOVA for both nonrescored and rescored algorithms (Multimedia Appendix 9).

Similar to the full sample results, post hoc contrasts for accuracy across subgroups were very small to small (all Hedges g results <0.3). Post hoc contrasts for F_1-score were very small to small, approaching medium effect sizes for nonrescored algorithms (all Hedges g results <0.35). F_1-score contrast effect sizes for the rescored algorithm demonstrated diminished performance with medium effect sizes for the K2010 algorithm across all subgroups. Similarly, the Philips threshold 20 algorithm showed diminished performance with medium effect sizes within the CPAP and RLS groups and to a lesser extent (Hedges g<0.40) for the apnea and insomnia groups. Complete subgroup post hoc results are presented in Multimedia Appendix 9.

The apnea subgroup of the K2010 rescored algorithm demonstrated diminished performance compared to the CK (Hedges g=−0.52), UCSD (Hedges g=−0.49), Sadeh (Hedges g=−0.50), Philips threshold 80 (Hedges g=−0.47), and Philips threshold 40 (Hedges g=−0.37). The Philips threshold 20 also demonstrated worse performance with small, approaching medium effect sizes compared to the CK, UCSD, Sadeh, and Philips threshold 80 algorithms (ranging from (Hedges g=−0.33 to −0.37). Within the CPAP subgroup, rescored K2010 showed worse performance than CK (Hedges g=−0.53), UCSD (Hedges g=−0.51), Sadeh (Hedges g=−0.53), Philips threshold 80 (Hedges g=−0.47), and Philips threshold 40 (Hedges g=−0.35). The Philips threshold 20 algorithm again demonstrated worse performance with small, approaching medium effect sizes compared to the CK, UCSD, Sadeh, and Philips threshold 80 algorithms (ranging from (Hedges g=−0.37 to −0.43). To a lesser extent, within the insomnia subgroup, comparisons of algorithms were similar to the apnea group. In addition, both K2010 and Philips threshold demonstrated small, approaching medium effect sizes for Philips threshold 20 algorithms. Similarly, within the RLS subgroup, the K2010 rescored algorithm demonstrated diminished performance compared to the CK (Hedges g=−0.48), UCSD (Hedges g=−0.49), Sadeh (Hedges g=−0.53), Philips threshold 80 (Hedges g=−0.46), and Philips threshold 40 (Hedges g=−0.35). The Philips threshold 20 also demonstrated worse performance with small, approaching medium effect sizes compared to the CK, UCSD, Sadeh, and Philips threshold 80 algorithms (ranging from Hedges g=−0.33 to −0.40). Overall, the post hoc test revealed a wide range of effect sizes ranging from very small to large for sensitivity, specificity, and precision across all subgroups and algorithms. Complete subgroup post hoc results presented in Multimedia Appendix 9.

XSL•FO
RenderX

Repeated measures ANOVA analyses for MCC and Cohen κ also demonstrated a significant difference between algorithms for a variety of sleep problem subgroups. Both the apnea and insomnia subgroups demonstrated significant differences between algorithms for Cohen κ in both nonrescored (Apnea: $F_{1.24,127.66}$=17.28; ε=0.21; P<.001; Insomnia: $F_{1.32,109.90}$=28.14; ε=0.22; P<.001) and rescored algorithms (Apnea: $F_{1.31,134.72}$=4.63; ε=0.22; P=.02; Insomnia: $F_{1.42,117.76}$=5.94; ε=0.24; P=.008) and for MCC between nonrescored algorithms (Apnea: $F_{1.31,135.07}$=4.51; ε=0.22; P=.03; Insomnia: $F_{1.50,124.90}$=9.19; ε=0.25; P<.001) . Both CPAP and RLS subgroups showed significant differences between algorithms for only Cohen κ and only for nonrescored algorithms (CPAP: $F_{1.23,74.80}$=8.31; ε=0.20; P=.003; RLS: $F_{1.24,80.75}$=12.28; ε=0.21; P<.001). Complete subgroup ANOVA results within Multimedia Appendix 10.

A subsequent post hoc test revealed small effect sizes for MCC (Hedges g<0.25) for both the apnea and insomnia subgroups across nonrescored algorithm contrasts. With respect to Cohen κ, effect sizes ranged from very small to small for apnea in the subgroup across most nonrescored and algorithm contrasts (Multimedia Appendix 10). The exception was the nonrescored K2010 algorithm which demonstrated small-approaching-medium effects, having better Cohen κ agreement when compared with nonrescored UCSD and Sadeh algorithms (Multimedia Appendix 10). Similarly, to a lesser extent, within both CPAP and RLS subgroups, nonrescored K2010 had better Cohen κ agreement when compared with nonrescored UCSD and Sadeh algorithms (Hedges g range 0.27 - 0.31). Complete subgroup post hoc results are presented in Multimedia Appendix 10.

Finally, the Bland-Altman distributions were examined for each algorithm, without and with rescoring applied, for 3 sleep metrics (TST, SE, and WASO) within each of the 4 subgroups (Multimedia Appendix 11). Overall, the subgroup results were again highly similar to those observed across the entire population sample. All algorithms demonstrated considerable mean difference and systematic bias for each subgroup analysis (Multimedia Appendix 11). Overall, rescoring did not improve agreement between the algorithm and ground truth, with some cases demonstrating poorer agreement after rescoring. Similar trends of bias were observed as the results from the entire dataset sample. For SE, all algorithms demonstrated heteroscedasticity for each subgroup (ie, estimates of SE were worse from polysomnography as SE decreased). The exception was the RLS group, which demonstrated random distributions for all algorithms. For WASO, all algorithms showed heteroscedasticity across all subgroups, where estimates of WASO were worse compared to polysomnography as WASO increased. For TST, there was no heteroscedasticity; that is, the distributions were random and evenly spread for all algorithms across all subgroups. For detailed Bland-Altman distribution statistics, see Multimedia Appendix 11.

Discussion

Principal Findings

Using the large MESA dataset, we established that traditional actigraphy algorithms can, with considerable accuracy, classify sleep and wake. As a first, this study comprehensively provides detailed benchmarks of traditional actigraphy algorithms. This study also provides the first rescoring performance and comparison across a wide range of actigraphy algorithms. This study is among the first to evaluate traditional actigraphy algorithms across a large population of older adults who are also at risk for, or have, health and sleep issues. To our knowledge, this is the first study to provide comprehensive benchmarks and statistical comparisons across an extensive list of the commonly used actigraphy algorithms with the diverse MESA population, while previous studies have focused on novel algorithm development and comparison. In doing so, we provide comprehensive metrics for not only users of these algorithms but also inform future research and practice.

Overall, the ROC-AUC and accuracy results demonstrate that all of the algorithms evaluated in this study are good options for evaluating sleep-wake activity. However, contrary to accuracy, our results for both MCC and Cohen κ demonstrate that traditional algorithms only provide moderate agreement and moderate positive correlations with ground truth polysomnography. These results suggest that though these actigraphy algorithms are a valuable analytic tool, there is a significant area of improvement. With novel technologies as well as improved analytic methods, we can improve our approach to sleep-wake assessment.

Along these lines, our results demonstrate that rescoring of actigraphy algorithms does not substantially improve sleep-wake classification accuracy. Rather, rescoring may significantly diminish the performance (eg, F_1-score) of certain algorithms such as Kripke 2010 and Philips-Respironics. Further, our analyses of sleep problem participants revealed similar findings that rescoring may diminish sleep-wake classification performance. This suggests that researchers and clinicians may safely opt to not rescore and still retain high accuracy.

In addition to the primary results, in-depth ANOVA and post hoc analyses showed significant differences between algorithms despite small differences in accuracy and other metrics. However, examination of effect sizes for MCC, accuracy, and F_1-score demonstrated very small to small effect sizes. Therefore, the significant differences between algorithms could be due to the large sample size of the dataset. The exception was Cohen κ, which demonstrated marginally higher performance for the Kripke 2010 algorithm (small to medium effect sizes). This, along with the Kripke 2010 algorithm demonstrating the highest accuracy, may suggest that this algorithm could be optimal in specific analysis scenarios. However, after rescoring, these effects were reduced with only very small to small effect sizes overall.

Our evaluation of the agreement between actigraphy and polysomnography for sleep metrics TST, SE, and WASO was conducted through Bland-Altman distribution analysis. This

XSL•FO
RenderX

analysis demonstrated both over and underestimation of all 3 sleep metrics, underscored by large mean differences between actigraphy and polysomnography. The Bland-Altman distributions also demonstrated considerable systematic bias for SE and WASO. In addition, there were clear patterns that demonstrated agreement decreased as SE decreased or when WASO increased. This was supported by the regression analyses for proportional bias. This suggests that actigraphy may be less precise in accurately measuring sleep metrics in individuals who have irregular sleep or sleep problems. From the traditional algorithms, the Philips algorithm with a threshold of 40 demonstrated the highest agreement with polysomnography for both TST and SE. This may be in part due to the supplied cutoffs for sleep-wake detection, which may allow for the user to adjust the algorithm based on their target population and other factors to improve both sleep-wake scoring and sleep metric estimates [37]. However, an important note is that the MESA data were collected using Philips Actiwatch devices; therefore, it is expected that the Philips algorithms may perform better in these cases than the other algorithms [20,21]. For WASO, the Kripke 2010 algorithm had the highest agreement with polysomnography. This may be due to the larger number of epochs (30) that are assessed by the algorithm to compute a sleep-wake decision. In evaluating more epochs, this algorithm may be more sensitive to shorter waking periods [19]. This may also explain the lower accuracy of this algorithm, as it may be scoring more epochs as wake even when there is true sleep. This suggests a key point for current algorithm use and for future algorithm development, that some variability or ranges of thresholds and longer windows of epoch evaluation may provide better accuracy and improve estimates of sleep metrics.

Another explanation for the bias and heteroscedasticity observed is that the sample evaluated included many individuals who were either at risk of or had health conditions such as the sleep problems highlighted in the sample distributions (Table 1). Though not reported in the NSRR version of the MESA dataset, MESA and publications on this dataset have underscored that this sample includes individuals with hypertension, diabetes, etc [20,48,49]. Therefore, it is expected that these conditions may lead to fragmented and irregular sleep, which may have influenced the actigraph or the algorithm's ability to accurately measure. Previous studies have noted that individuals with sleep problems or health conditions that may affect sleep demonstrate more variability in actigraphy results of both sleep-wake and sleep metrics [3,12-17]. Therefore, one important consideration was further tests of these algorithms with data from populations with health and sleep issues. Notably, these are the key populations who are recommended or given actigraphs for sleep assessment by researchers and clinicians. Therefore, the precision of their sleep or wake classification and sleep metrics by these algorithms is of great importance.

Along these lines, we first conducted an exploratory analysis of algorithm performance within sleep problem subgroups, specifically examining individuals with insomnia, RLS, apnea, and those who used a CPAP device. We found that traditional actigraphy algorithms do not perform particularly worse; rather, they are just as robust in determining sleep and wake or sleep metrics in individuals with sleep problems compared with the entire sample. These algorithms demonstrated very similar results and patterns to the entire sample with respect to AUC, accuracy, confusion matrix results, MCC, Cohen κ, and agreement (Bland-Altman distributions). However, generally, these algorithms perform well with respect to sleep-wake scoring but not as well for the measurement of specific sleep metrics in contrast to the overall population. Our results show that sleep disorders may skew Bland-Altman distributions as sleep patterns within these disorders deviate from norms. We theorize based on our results that this may result in larger mean differences when compared to ground truth. Based on our results, we theorize that this may create skewness in accuracy in particularly extreme cases. We suggest that clinicians pay attention to severe cases of sleep disorders and take into account our results and the decreased reliability of actigraphy-based sleep metrics such as TST due to outlier sleep patterns.

Overall, these results provide clinicians with nuanced benchmarks for individual algorithm performance on specific sleep disorders. This will allow clinicians to make an informed choice not only based on accuracy, but on sensitivity and specificity metrics to precisely analyze patient sleep data. With mean difference and Bland-Altman plots, clinicians now have benchmarks to provide them with information on how extreme cases may deviate from general trends. Further, both clinicians and manufacturers now have mean difference benchmarks, which reflect the difference between ground truth and actigraphy, allowing for corrections in patient data analyses if needed.

In the future, it would be interesting to compare the precision of traditional actigraphy algorithms across a variety of actigraph and accelerometer devices. Overall, both the main and subgroup results suggest that generally traditional actigraphy algorithms may have poorer agreement with polysomnography in fringe cases. Specifically, cases in which the individual's sleep may be significantly different from the vast majority of the population. One hypothesis here is that the heteroscedasticity may be correlated to the severity of a sleep problem, which explains the poorer agreement as SE or WASO changes. Further analysis would allow us to gain a better understanding of this phenomenon.

With regard to rescoring and the sleep problem sub-samples, there were mixed results. Traditional use of the Cole-Kripke (1992) [6] algorithm proposed the use of rescoring; first, we investigated rescoring across all included algorithms [40]. Contrary to the belief that rescoring would significantly improve accuracy metrics, rescoring only slightly improved agreement for sleep metrics and may diminish specific performance variables. This could be due to the rescoring of smaller periods of sleep and wake, especially in the aforementioned population with sleep problems. Rescoring may be better suited across multiple nights, which may present more within-participant variability, whereas our sample was only over one night. Rescoring may also smooth out or erroneously change edge cases or cases in which individuals have sleep problems. This would, in part, explain the diminished performance of results. One future direction could be further optimization of rescoring criteria for improvements to sleep-wake detection across algorithms or to better suit individual algorithms specifically.

XSL•FO
RenderX

This may also remove the diminished performance observed after rescoring with sleep problem populations. As a current recommendation, we suggest that rescoring be used only when absolutely necessary and that when rescoring, researchers compare results that are both rescored and nonrescored. Our findings suggest that rescoring does not provide any substantial benefits to nonrescored algorithm results. Further, we do not recommend rescoring for sleep disorder population data to preserve accurate sleep-wake scoring.

Overall, these results echo the findings of previous studies which have examined these algorithms. Haghayegh et al [50] examined the performance of the Coke-Kripke, Sadeh, and UCSD algorithms on a sample of 40 healthy adults. Similar to our results, they found accuracies of 85% - 86% across algorithms. Additionally, they also found varying levels of overall bias across algorithms for the algorithms' estimation of total sleep time, sleep onset latency, and wake after sleep onset. Along these lines, Palotti et al [51] and Jokar et al [52] used a smaller subset of the MESA dataset to compare performance to the Cole-Kripke and Sadeh algorithms to their novel method both for night only and night-day. Cole-Kripke and Sadeh demonstrated mean accuracies ranging from 70%to 85% [53,54].

In addition, this study presented some limitations. Foremost, there has been recent work and novel algorithms for sleep and wake classification. Some notable examples include novel machine learning and other methods to evaluate activity data. These algorithms may be more precise and robust as sleep and wake classification. Though our current study did not directly evaluate these algorithms, previous research has provided some benchmarks for comparison. Palotti et al [51] developed novel Convolutional Neural Network and Long Short-Term Memory machine learning approaches using the MESA dataset for training and testing. Their novel approaches demonstrated better performance (88.2% and 87.7% respectively) than traditional algorithms (Cole-Kripke and Sadeh). In comparison to this study, their novel methods also demonstrate higher performance than traditional algorithms, albeit on a smaller subsample of the MESA dataset. Similarly, Nunes et al [55] developed a domain adversarial convolutional neural network method, trained on the MESA dataset. Their optimal method demonstrated the highest accuracy (80.1%) when compared to other models and traditional algorithms such as Cole-Kripke and Sadeh (71.9% and 69.6%). In contrast to our findings, their novel method could provide stable, more generalizable sleep-wake classification. Again, these results suggest that novel machine learning methods are the future for actigraphy analyses. Given the rapid development of novel actigraphy analytic methods, our future study aim is to examine the performance of newer algorithms and strategies with large testing samples in contrast to our current study findings.

Another limitation of this study is that we only examined sleep problem subgroups as an exploratory aim. Given the wide demographic, there were several other populations within the dataset, such as individuals with diabetes, cardiovascular illness, etc that could explain the lack of agreement or the heteroscedasticity observed. In addition, we did not conduct an analysis examining demographic subgroups such as sex, ethnicity, and age. These groups may have also contributed to the Bland-Altman distributions observed.

Given the similarities of the subgroup results to those of the entire dataset, there may be additional correlation-based analysis, which could help unpack why traditional actigraphy algorithms perform poorly as sleep metrics reach extreme values. As aforementioned, previous research denotes individuals with sleep problems may have lower SE, higher WASO, and a variety of other extreme sleep activity metrics. Therefore, this may be one important limitation of not only the actigraphy algorithms but the activity measurement devices themselves. To address this issue, future research should delve deeper and examine extreme cases and traditional algorithms' ability to classify sleep and wake and accurately measure sleep metrics. This, in turn, may also allow us to gain a better understanding of specific patterns of sleep and wake in these populations, as well as examine whether specific populations alter the distribution variability observed. In addition, future research should comprehensively assess the performance of traditional and novel algorithms on subsets of the MESA dataset. In doing so, we could identify key markers that may be used to improve analysis performance.

Finally, our study also filtered out fragmented or poor-quality data from the dataset. Even though these data may not be suitable for the current accuracy analysis and comparison, they may highlight important limitations of actigraphy algorithms with respect to the estimation of sleep metrics and the classification of sleep and wake in extreme cases. Therefore, future research should assess a spectrum of poor- and good-quality data to gain a complete understanding of the performance of traditional actigraphy algorithms. In addition, future research should also compare the results of different actigraphy devices to determine, in unison with analytics, the most accurate actigraphy and accelerometer devices.

Overall, the results of this study provide significant support that traditional actigraphy algorithms can, with acceptable accuracy, detect sleep and wake in large, diverse population samples, including older adults or populations at risk of health conditions. This study provides researchers and clinicians with evidence that traditional algorithms can continue to be used to assess sleep-wake activity. However, traditional algorithms may present significant limitations in measurement precision of extreme sleep cases. Further, rescoring may not be a necessary step in the analysis of actigraphy data. The implications of this finding are highly important as researchers develop new algorithms and methods of actigraphy data analysis. Specifically, new algorithms should consider the variability of sleep and wake data in sleep problems or extreme cases to precisely measure sleep and wake activity. Further, the results denoted by the current study serve as an important reference point for new, developing strategies for actigraphy analyses.

Conclusion

In conclusion, we emphasize the future direction for generalizable, accurate, and comprehensive actigraphy analyses through the application of machine learning and artificial intelligence models. As aforementioned, several researchers have now adapted artificial intelligence–based models and

XSL•FO
RenderX

automated toolboxes to analyze actigraphy and reportedly extract more accurate and more detailed sleep-wake results [53-57]. Given the promise of these methods, we theorize that their assessment and usage to analyze actigraphy data in novel ways poses a promising future direction for more accurate and diverse actigraphy analyses for sleep assessment. Further, in unison with demographic analyses, first of their performance, then the usage of demographic and sleep problem data, we propose this would be a novel opportunity to develop more robust and accurate methods of sleep-wake prediction from activity and actigraphy data. In addition, novel methods using unsupervised machine learning may also provide new ways in which sleep-wake actigraphy data could be parsed to reveal underlying patterns of sleep and sleep problems [56,58]. As these methods are fitted to specific datasets, they could provide better generalizability for sleep-wake analyses [56,58].

Acknowledgments

We would like to acknowledge the contributions of Mit Patel, who helped with this project.

The Multi-Ethnic Study of Atherosclerosis (MESA) Sleep Ancillary study was funded by NIH-NHLBI Association of Sleep Disorders with Cardiovascular Health Across Ethnic Groups (RO1 HL098433). MESA is supported by NHLBI-funded contracts HHSN268201500003I, N01-HC-95159, N01-HC-95160, N01-HC-95161, N01-HC-95162, N01-HC-95163, N01-HC-95164, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, and N01-HC-95169 from the National Heart, Lung, and Blood Institute, and by cooperative agreements UL1-TR-000040, UL1-TR-001079, and UL1-TR-001420 funded by NCATS. The National Sleep Research Resource was supported by the National Heart, Lung, and Blood Institute (R24 HL114473, 75N92019R002) [59].

The author(s) attest there was no use of generative artificial intelligence (AI) in the generation of text, figures, or other informational content of this manuscript. The authors note and attest within the methods section of all uses of generative artificial intelligence (AI). Specifically, generative AI was exclusively used to draft Python code for predetermined purposes. We provide detailed context for its use within our methods section and below. “We used ChatGPT for coding assistance, to generate Python code for existing Python libraries [28]. ChatGPT was used to generate Python code for file data file preparation and processing (eg, data transformation). Manual data checks in addition to code-based error reporting were conducted to verify the code was working as desired and data was correctly processed. In addition, ChatGPT was used to generate Python code for running statistical analyses. All statistical formulae and analyses were verified within the code. Once code was generated, it was checked by at least one or more of study authors before use. No machine learning code was generated via ChatGPT, and any actigraphy algorithm code was checked and specified in the python code exactly as defined by original algorithm sources.”

Funding

Robyn Stremler received funding for this project from the Canadian Institutes of Health Research; Healthy Data Behavior Challenge operating grants HBC 15488 and HBC 15694. Robyn Stremler holds the Lawrence S. Bloomberg Limited-Term Professorship in Child and Family Health.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Sensitivity analysis of outliers.
[DOCX File, 58 KB - formative_v9i1e70778_app1.docx]

Multimedia Appendix 2
Formulae for statistical analyses.
[DOCX File, 16 KB - formative_v9i1e70778_app2.docx]

Multimedia Appendix 3
Confusion metrics, Matthews correlation coefficient, and Cohen κ statistics.
[DOCX File, 20 KB - formative_v9i1e70778_app3.docx]

Multimedia Appendix 4
ANOVA and post hoc results for Matthews correlation coefficient and Cohen κ.
[DOCX File, 37 KB - formative_v9i1e70778_app4.docx]

Multimedia Appendix 5

XSL•FO
RenderX

Post hoc results for confusion matrix statistics.
[DOCX File, 76 KB - formative_v9i1e70778_app5.docx]

Multimedia Appendix 6
Mean difference statistics.
[DOCX File, 144 KB - formative_v9i1e70778_app6.docx]

Multimedia Appendix 7
Receiver operator characteristic curve and area under the curve statistics for sleep problem subgroups.
[DOCX File, 1954 KB - formative_v9i1e70778_app7.docx]

Multimedia Appendix 8
Subgroup confusion matrix statistics.
[DOCX File, 40 KB - formative_v9i1e70778_app8.docx]

Multimedia Appendix 9
Subgroup ANOVA and post hoc results for confusion matrix statistics.
[DOCX File, 268 KB - formative_v9i1e70778_app9.docx]

Multimedia Appendix 10
Subgroup ANOVA and post hoc results for Matthews correlation coefficient and Cohen κ.
[DOCX File, 84 KB - formative_v9i1e70778_app10.docx]

Multimedia Appendix 11
Subgroup Bland-Altman statistics.
[DOCX File, 10757 KB - formative_v9i1e70778_app11.docx]

References

1. Acebo C, LeBourgeois MK. Actigraphy. Respir Care Clin N Am 2006 Mar;12(1):23-30. [doi: 10.1016/j.rcc.2005.11.010] [Medline: 16530645]
2. Sadeh A. Actigraphy. In: Handbook of Clinical Neurophysiology: Elsevier; 2005, Vol. 6:67-72. [doi: 10.1016/S1567-4231(09)70031-0]
3. Sadeh A. The role and validity of actigraphy in sleep medicine: an update. Sleep Med Rev 2011 Aug;15(4):259-267. [doi: 10.1016/j.smrv.2010.10.001] [Medline: 21237680]
4. Ancoli-Israel S, Cole R, Alessi C, Chambers M, Moorcroft W, Pollak CP. The role of actigraphy in the study of sleep and circadian rhythms. Sleep 2003 May 1;26(3):342-392. [doi: 10.1093/sleep/26.3.342] [Medline: 12749557]
5. Banfi T, Valigi N, di Galante M, d'Ascanio P, Ciuti G, Faraguna U. Efficient embedded sleep wake classification for open-source actigraphy. Sci Rep 2021 Jan 11;11(1):345. [doi: 10.1038/s41598-020-79294-y] [Medline: 33431918]
6. Cole RJ, Kripke DF, Gruen W, Mullaney DJ, Gillin JC. Automatic sleep/wake identification from wrist activity. Sleep 1992 Oct;15(5):461-469. [doi: 10.1093/sleep/15.5.461] [Medline: 1455130]
7. Haghayegh S, Khoshnevis S, Smolensky MH, Diller KR, Castriotta RJ. Performance comparison of different interpretative algorithms utilized to derive sleep parameters from wrist actigraphy data. Chronobiol Int 2019 Dec 2;36(12):1752-1760. [doi: 10.1080/07420528.2019.1679826]
8. Jean-Louis G, Kripke DF, Cole RJ, Assmus JD, Langer RD. Sleep detection with an accelerometer actigraph: comparisons with polysomnography. Physiol Behav 2001 Jan;72(1-2):21-28. [doi: 10.1016/s0031-9384(00)00355-3] [Medline: 11239977]
9. Jean-Louis G, Kripke DF, Mason WJ, Elliott JA, Youngstedt SD. Sleep estimation from wrist movement quantified by different actigraphic modalities. J Neurosci Methods 2001 Feb 15;105(2):185-191. [doi: 10.1016/s0165-0270(00)00364-2] [Medline: 11275275]
10. Kripke DF, Hahn EK, Grizas AP, et al. Wrist actigraphic scoring for sleep laboratory patients: algorithm development. J Sleep Res 2010 Dec;19(4):612-619. [doi: 10.1111/j.1365-2869.2010.00835.x] [Medline: 20408923]
11. Sadeh A, Sharkey KM, Carskadon MA. Activity-based sleep-wake identification: an empirical test of methodological issues. Sleep 1994 Apr;17(3):201-207. [doi: 10.1093/sleep/17.3.201] [Medline: 7939118]
12. Blackwell T, Redline S, Ancoli-Israel S, et al. Comparison of sleep parameters from actigraphy and polysomnography in older women: the SOF study. Sleep 2008 Feb;31(2):283-291. [doi: 10.1093/sleep/31.2.283] [Medline: 18274276]
13. de Souza L, Benedito-Silva AA, Pires MLN, Poyares D, Tufik S, Calil HM. Further validation of actigraphy for sleep studies. Sleep 2003 Feb 1;26(1):81-85. [doi: 10.1093/sleep/26.1.81] [Medline: 12627737]

XSL•FO
RenderX

14. Gooneratne NS, Vitiello MV. Sleep in older adults: normative changes, sleep disorders, and treatment options. Clin Geriatr Med 2014 Aug;30(3):591-627. [doi: 10.1016/j.cger.2014.04.007] [Medline: 25037297]
15. Morgenthaler T, Alessi C, Friedman L, et al. Practice parameters for the use of actigraphy in the assessment of sleep and sleep disorders: an update for 2007. Sleep 2007 Apr;30(4):519-529. [doi: 10.1093/sleep/30.4.519] [Medline: 17520797]
16. Desforges JF, Prinz PN, Vitiello MV, Raskind MA, Thorpy MJ. Sleep disorders and aging. N Engl J Med 1990 Aug 23;323(8):520-526. [doi: 10.1056/NEJM199008233230805]
17. Sivertsen B, Omvik S, Havik OE, et al. A comparison of actigraphy and polysomnography in older adults treated for chronic primary insomnia. Sleep 2006 Oct;29(10):1353-1358. [doi: 10.1093/sleep/29.10.1353] [Medline: 17068990]
18. Marino M, Li Y, Rueschman MN, et al. Measuring sleep: accuracy, sensitivity, and specificity of wrist actigraphy compared to polysomnography. Sleep 2013 Nov 1;36(11):1747-1755. [doi: 10.5665/sleep.3142] [Medline: 24179309]
19. Sadeh A, Acebo C. The role of actigraphy in sleep medicine. Sleep Med Rev 2002 Apr;6(2):113-124. [doi: 10.1053/smrv.2001.0182] [Medline: 12531147]
20. Chen X, Wang R, Zee P, et al. Racial/ethnic differences in sleep disturbances: the multi-ethnic study of atherosclerosis (MESA). Sleep 2015 Jun 1;38(6):877-888. [doi: 10.5665/sleep.4732] [Medline: 25409106]
21. Zhang GQ, Cui L, Mueller R, et al. The national sleep research resource: towards a sleep data commons. J Am Med Inform Assoc 2018 Oct 1;25(10):1351-1358. [doi: 10.1093/jamia/ocy064] [Medline: 29860441]
22. Fekedulegn D, Andrew ME, Shi M, Violanti JM, Knox S, Innes KE. Actigraphy-based assessment of sleep parameters. Ann Work Expo Health 2020 Apr 30;64(4):350-367. [doi: 10.1093/annweh/wxaa007] [Medline: 32053169]
23. Soldan A, Alfini A, Pettigrew C, et al. Actigraphy-estimated physical activity is associated with functional and structural brain connectivity among older adults. Neurobiol Aging 2022 Aug;116:32-40. [doi: 10.1016/j.neurobiolaging.2022.04.006] [Medline: 35551019]
24. Meredith-Jones KA, Haszard JJ, Graham-DeMello A, et al. Validation of actigraphy sleep metrics in children aged 8 to 16 years: considerations for device type, placement and algorithms. Int J Behav Nutr Phys Act 2024 ;21(1):40. [doi: 10.1186/s12966-024-01590-x]
25. Razjouyan J, Lee H, Parthasarathy S, Mohler J, Sharafkhaneh A, Najafi B. Improving sleep quality assessment using wearable sensors by including information from postural/sleep position changes and body acceleration: a comparison of chest-worn sensors, wrist actigraphy, and polysomnography. J Clin Sleep Med 2017 Nov 15;13(11):1301-1310. [doi: 10.5664/jcsm.6802] [Medline: 28992827]
26. Derbin M, McKenna L, Chin D, Coffman B, Bloch-Salisbury E. Actigraphy: metrics reveal it is not a valid tool for determining sleep in neonates. J Sleep Res 2022 Feb;31(1):e13444. [doi: 10.1111/jsr.13444] [Medline: 34291522]
27. Meltzer LJ, Wong P, Biggs SN, et al. Validation of actigraphy in middle childhood. Sleep 2016 Jun 1;39(6):1219-1224. [doi: 10.5665/sleep.5836] [Medline: 27091520]
28. ChatGPT. OpenAI. URL: https://www.openai.com/chatgpt [accessed 2025-11-20]
29. Smith MT, McCrae CS, Cheung J, et al. Use of actigraphy for the evaluation of sleep disorders and circadian rhythm sleep-wake disorders: an American academy of sleep medicine systematic review, meta-analysis, and GRADE assessment. J Clin Sleep Med 2018 Jul 15;14(7):1209-1230. [doi: 10.5664/jcsm.7228] [Medline: 29991438]
30. Liu F, Schrack J, Wanigatunga SK, et al. Comparison of sleep parameters from wrist-worn ActiGraph and Actiwatch devices. Sleep 2024 Feb 8;47(2):zsad155. [doi: 10.1093/sleep/zsad155] [Medline: 37257489]
31. Cespedes Feliciano EM, Quante M, Weng J, et al. Actigraphy-derived daily rest-activity patterns and body mass index in community-dwelling adults. Sleep 2017 Dec 1;40(12):zsx168. [doi: 10.1093/sleep/zsx168] [Medline: 29029250]
32. Van Den Berg JF, Van Rooij FJA, Vos H, et al. Disagreement between subjective and actigraphic measures of sleep duration in a population-based study of elderly persons. J Sleep Res 2008 Sep;17(3):295-302. [doi: 10.1111/j.1365-2869.2008.00638.x] [Medline: 18321246]
33. Quante M, Kaplan ER, Cailler M, et al. Actigraphy-based sleep estimation in adolescents and adults: a comparison with polysomnography using two scoring algorithms. Nat Sci Sleep 2018;10:13-20. [doi: 10.2147/NSS.S151085] [Medline: 29403321]
34. Kim MJ, Lee GH, Kim CS, et al. Comparison of three actigraphic algorithms used to evaluate sleep in patients with obstructive sleep apnea. Sleep Breath 2013 Mar;17(1):297-304. [doi: 10.1007/s11325-012-0689-z] [Medline: 22447172]
35. Sadeh A. Actigraphically based automatic bedtime sleep-wake scoring. Validity and clinical applications. J Ambul Monit 1989;87:209-216.
36. Sadeh A, Hauri PJ, Kripke DF, Lavie P. The role of actigraphy in the evaluation of sleep disorders. Sleep 1995 May;18(4):288-302. [doi: 10.1093/sleep/18.4.288] [Medline: 7618029]
37. Paquet J, Kawinska A, Carrier J. Wake detection capacity of actigraphy during sleep. Sleep 2007 Oct;30(10):1362-1369. [doi: 10.1093/sleep/30.10.1362] [Medline: 17969470]
38. Oakley NR. Validation with polysomnography of the sleepwatch sleep/wake scoring algorithm used by the Actiwatch activity monitoring system. Mini Mitter Co 1997.
39. Gao C, Li P, Morris CJ, et al. Actigraphy-based sleep detection: validation with polysomnography and comparison of performance for nighttime and daytime sleep during simulated shift work. Nat Sci Sleep 2022;14:1801-1816. [doi: 10.2147/NSS.S373107] [Medline: 36275180]

XSL•FO
RenderX

40. Webster JB, Kripke DF, Messin S, Mullaney DJ, Wyborney G. An activity-based sleep monitor system for ambulatory use. Sleep 1982;5(4):389-399. [doi: 10.1093/sleep/5.4.389] [Medline: 7163726]
41. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960 Apr;20(1):37-46. [doi: 10.1177/001316446002000104]
42. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22(3):276-282. [doi: 10.11613/bm.2012.031] [Medline: 23092060]
43. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 1975 Oct 20;405(2):442-451. [doi: 10.1016/0005-2795(75)90109-9] [Medline: 1180967]
44. Chicco D, Tötsch N, Jurman G. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. BioData Min 2021 Feb 4;14(1):13. [doi: 10.1186/s13040-021-00244-z] [Medline: 33541410]
45. Fritz CO, Morris PE, Richler JJ. Effect size estimates: current use, calculations, and interpretation. J Exp Psychol Gen 2012 Feb;141(1):2-18. [doi: 10.1037/a0024338] [Medline: 21823805]
46. Brydges CR. Effect size guidelines, sample size calculations, and statistical power in gerontology. Innov Aging 2019 Aug;3(4):igz036. [doi: 10.1093/geroni/igz036] [Medline: 31528719]
47. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986 Feb 8;1(8476):307-310. [doi: 10.1016/s0140-6736(86)90837-8] [Medline: 2868172]
48. Koo BB, Sillau S, Dean DA II, Lutsey PL, Redline S. Periodic limb movements during sleep and prevalent hypertension in the multi-ethnic study of atherosclerosis. Hypertension 2015 Jan;65(1):70-77. [doi: 10.1161/HYPERTENSIONAHA.114.04193] [Medline: 25287399]
49. Sharrett AR, Ding J, Criqui MH, et al. Smoking, diabetes, and blood cholesterol differ in their associations with subclinical atherosclerosis: the Multiethnic Study of Atherosclerosis (MESA). Atherosclerosis 2006 Jun;186(2):441-447. [doi: 10.1016/j.atherosclerosis.2005.08.010] [Medline: 16154575]
50. Haghayegh S, Khoshnevis S, Smolensky MH, Diller KR, Castriotta RJ. Performance comparison of different interpretative algorithms utilized to derive sleep parameters from wrist actigraphy data. Chronobiol Int 2019 Dec;36(12):1752-1760. [doi: 10.1080/07420528.2019.1679826] [Medline: 31658822]
51. Palotti J, Mall R, Aupetit M, et al. Benchmark on a large cohort for sleep-wake classification with machine learning techniques. NPJ Digit Med 2019;2:50. [doi: 10.1038/s41746-019-0126-9] [Medline: 31304396]
52. Jokar F, Azzopardi G, Palotti J. Towards accurate and efficient sleep period detection using wearable devices. Presented at: International Conference on Computer Analysis of Images and Patterns; Sep 25-28, 2023; Limassol, Cyprus p. 43-54. [doi: 10.1007/978-3-031-44240-7_5]
53. Chen Z, Wu M, Cui W, Liu C, Li X. An attention based CNN-LSTM approach for sleep-wake detection with heterogeneous sensors. IEEE J Biomed Health Inform 2021 Sep;25(9):3270-3277. [doi: 10.1109/JBHI.2020.3006145] [Medline: 32749983]
54. Hammad G, Reyt M, Beliy N, et al. pyActigraphy: open-source python package for actigraphy data visualization and analysis. PLoS Comput Biol 2021 Oct;17(10):e1009514. [doi: 10.1371/journal.pcbi.1009514] [Medline: 34665807]
55. Nunes AS, Patterson MR, Gerstel D, Khan S, Guo CC, Neishabouri A. Domain adversarial convolutional neural network improves the accuracy and generalizability of wearable sleep assessment technology. Sensors (Basel) 2024 Dec 14;24(24):7982. [doi: 10.3390/s24247982] [Medline: 39771718]
56. Li X, Zhang Y, Jiang F, Zhao H. A novel machine learning unsupervised algorithm for sleep/wake identification using actigraphy. Chronobiol Int 2020 Jul;37(7):1002-1015. [doi: 10.1080/07420528.2020.1754848] [Medline: 32342702]
57. Sano A, Chen W, Lopez-Martinez D, Taylor S, Picard RW. Multimodal ambulatory sleep detection using LSTM recurrent neural networks. IEEE J Biomed Health Inform 2019 Jul;23(4):1607-1617. [doi: 10.1109/JBHI.2018.2867619] [Medline: 30176613]
58. Berahmand K, Daneshfar F, Salehi ES, Li Y, Xu Y. Autoencoders and their applications in machine learning: a survey. Artif Intell Rev 2024 Feb;57(2):28. [doi: 10.1007/s10462-023-10662-6]
59. Multi-ethnic study of atherosclerosis (MESA). National Heart, Lung, and Blood Institute (NHLBI). URL: https://clinicaltrials.gov/ct2/show/NCT00005487 [accessed 2025-11-21]

Abbreviations:

AUC: area under the curve
CK: Cole-Kripke
CPAP: continuous positive airway pressure
ICC: intraclass correlation coefficient
K2010: Kripke 2010
LoA: limits of agreement
MCC: Matthews correlation coefficient
MESA: Multi-Ethnic Study of Atherosclerosis
NRS: nonrescored result

XSL•FO
RenderX

NSRR: National Sleep Research Resource
Philips: Philips-Respironics
RLS: restless leg syndrome
ROC: receiver operator characteristic curve
RR: rescored result
RS: Webster's rescoring rules
SE: sleep efficiency
TST: total sleep time
UCSD: University of California San Diego
WASO: wake after sleep onset

Edited by J Sarvestan; submitted 02.01.25; peer-reviewed by K Berahmand, X Zheng; revised version received 29.10.25; accepted 30.10.25; published 11.12.25.

Please cite as:
Panesar D, Vichare A, Goncalves J, Stremler R
Comparison and Validation of Actigraphy Algorithms Using a Large Community Dataset: Algorithm Validation Study
JMIR Form Res 2025;9:e70778
URL: https://formative.jmir.org/2025/1/e70778
doi:10.2196/70778

© Darshan Panesar, Aashish Vichare, Jason Goncalves, Robyn Stremler. Originally published in JMIR Formative Research (https://formative.jmir.org), 11.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Consumer Co-Design of an Online Resource to Build Communication Skills of Health Consumers: Mixed Methods Study

Alison Beauchamp[1], PhD; Julieanne Hilbers[2], PhD; Natali Cvetanovska[1,3], MPH; Anna Wong Shee[4,5], PhD; Lidia Horvat[6], MSSci; Sandra Rogers[7], MEng; Andrea Cooper[8]; Elizabeth Flemming-Judge[8], MSW; Sue Rawlinson[9], MPsych; Rebecca Jessup[1,3], PhD

[1]School of Rural Health, Monash University, D501, Level 535 Rainforest Walk, Wellington Road, Clayton, Victoria, Australia
[2]Compassionate Communities, Perron Institute, University of Western Australia, Nedlands, Western Australia, Australia
[3]Office of Research, Northern Health, Melbourne, Australia
[4]School of Medicine, Deakin University, Melbourne, Australia
[5]Research Development, Grampians Health, Ballarat, Australia
[6]Department of Health, Safer Care Victoria, Melbourne, Australia
[7]Patient Experience Office, Monash Health, Melbourne, Australia
[8]Department of Health, Health Consumer, Safer Care Victoria, Melbourne, Australia
[9]Health Consumer, Community Partnerships, South East Sydney Local Health District, Sydney, Australia

Corresponding Author:
Alison Beauchamp, PhD
School of Rural Health, Monash University, D501, Level 535 Rainforest Walk, Wellington Road, Clayton, Victoria, Australia

Abstract

Background: Information provided by health professionals can be complex and is often not well understood by health care consumers, leading to adverse outcomes. Clinician-led communication approaches such as “teach-back” can improve consumer understanding, yet are infrequently used by clinicians. A possible solution is to build consumers’ skills to proactively check their understanding rather than waiting for the clinician to do so; however, there are few educational resources to support consumers in building these skills.

Objective: This study aimed to co-design a web-based learning resource for consumers to check they have understood information provided by a clinician (ie, to “check-back”).

Methods: This mixed methods study used a co-design approach, consisting of 2 phases. The study was conducted during the COVID-19 pandemic, and all activities were conducted online, via email or telephone. Phase 1 (needs assessment) involved first establishing an Expert Panel of consumers, clinicians, and academic experts to guide all co-design steps of the study. Next, we sought to understand issues around health communication through focus groups and interviews with consumers and clinicians. Participants were recruited from outpatient settings and consumer representative programs within 3 health services in Victoria, Australia. Focus groups and interviews aimed to identify factors that might influence consumers’ use of check-back. Deductive analysis based on the Capability, Opportunity, and Motivation-Behavior (COM-B) model was used to identify initial themes; these were discussed in depth with the Expert Panel and barriers within each theme identified. A rapid literature review was undertaken to identify strategies for web-based communication training for consumers. Phase 2 (creation of the online resource) involved an iterative process. In an online meeting, Expert Panel members brainstormed ideas for addressing barriers and prioritized these ideas for inclusion in the resource. Several drafts of the content were written before a draft online version was built. This draft was reviewed by the Expert Panel, who recommended extensive revisions. Following these revisions, we conducted an online survey and focus group with consumers and clinicians from Phase 1 to identify further improvements. Findings from this consultation were used to make final changes to the online resource.

Results: The Expert Panel included 12 members. Phase 1 focus groups and interviews were held with 39 consumers and 16 clinicians. Five themes were identified: self-efficacy, pre-existing skills, clinician attitudes, information complexity, and internal barriers such as embarrassment. Phase 2 survey and focus group participants identified several issues with the second draft of the resource, focusing on functionality, accessibility, and layout. Usability and acceptability of the resource were rated highly by participants.

Conclusions: Findings highlight the value of using co-design to develop a consumer-centered, web-based learning resource. Further evaluation is required to demonstrate its effectiveness at improving consumer understanding.

(JMIR Form Res 2025;9:e77263) doi:10.2196/77263

XSL•FO
RenderX

KEYWORDS

patient-provider communication; patient understanding; health literacy; co-design; consumer engagement; user-centered design; stakeholder collaboration

Introduction

The health system places a high burden on health care consumers to understand and apply health information across a range of contexts. From providing informed consent to following complex health care regimes or participating in decisions about their health, it is important that people are able to understand and recall information given to them by health professionals [1]. Studies have demonstrated that health information is frequently not understood or remembered by health care consumers [2-5], leading to increased treatment burden and adverse health outcomes [6-8]. Factors that may affect understanding originate from both the consumer and health professional and include limited health literacy and confidence in asking questions or engaging in health care discussions [3,5,9-11], and limited ability of the health professional to convey health information clearly [7,11-13]. Evidence-based approaches to address this communication gap are required.

One communication approach that has been associated with improved consumer understanding and recall across diverse settings and health conditions is "teach-back" [14-17]. This approach involves a health professional providing information and then confirming the consumer's understanding by asking them to explain back that information in their own words. In this way, any misunderstandings can be identified and addressed within an active, 2-way interaction. There is strong evidence that checking health care consumers' understanding through approaches such as teach-back can build greater rapport and more active communication [18], and may lead to improved knowledge, skills, and self-care abilities [14-16]. Yet despite the fact that these approaches offer a simple and effective approach for improving understanding, they are not widely adopted by health professionals [19]. Approaches such as teach-back rely on the health professional to initiate the process, and there are several barriers to their use, including lack of confidence and concern about the consumer reaction [19]. Despite this, evidence suggests that health care consumers are receptive to the idea of being asked to confirm their understanding, stating this would help with their learning and support them to remember important information [18,20].

One possible strategy to increase the use of approaches such as teach-back is to build consumers' skills and confidence to actively check their understanding rather than waiting for the health professional to do so. For example, this may involve the consumer saying, "Can I just check that I've understood everything. Did you say...." Encouraging consumers to initiate the process of repeating back information using their own words (ie, to "check-back") may be a feasible approach to improve understanding and recall of information across many health care settings. However, while some communication skills, such as asking questions and giving information, may come naturally to many health care consumers, initiating the process of checking-back requires more advanced skills [21]. These include the ability to actively listen, to summarize key information, and then the confidence to explain back that information in their own words [21].

Many of these skills could be learned through communication training or coaching; these learning approaches have been demonstrated to be effective in increasing consumer participation and engagement in health care interactions [21-23]. However, while education-based interventions may provide health care consumers with the skills to initiate check-backs themselves, there are also many barriers to consider, including low self-efficacy and confidence, perceived power imbalance, the communication style of the health professional, cultural factors, and structural factors such as short consultation times [24]. Any intervention that supports health care consumers to use the check-back process needs to address these barriers while also considering the context within which the intervention will be used. An effective methodology for developing interventions that are responsive to local barriers is that of co-design. The co-design approach in health enables equal involvement of consumers, health professionals, and other stakeholders throughout all stages of intervention development, ensuring that the end product meets the needs of those for whom it is designed [25-27]. Co-design can be defined as an active collaboration between stakeholders in designing solutions to a prespecified problem and is shown to be an effective approach for developing strategies to improve health communication [28-31].

This paper describes the co-design of a web-based learning resource for health care consumers. The decision to select a web-based approach was initially driven by consumers who had been involved in co-design of a similar web-based learning resource for clinicians called "teach-back" [32]. Those consumers noted they would like to see a web-based learning resource specifically for consumers that prompted them to initiate the teach-back process (ie, to "check-back"). The aim of this web-based learning resource is therefore to equip consumers with the necessary skills and confidence to initiate the check-back process in any setting and interaction with a health professional, that is, to confirm understanding through summarizing and repeating back key information to the health professional. For the purposes of this paper, the term "consumers" will hereafter be used to refer to health care consumers, patients, or users of the health system.

Methods

Study Design

This mixed methods study uses a co-design approach developed by Boyd et al [33,34]. This approach has six elements: (1) engage, (2) plan, (3) explore, (4) develop, (5) decide, (6) change. As noted above, co-design seeks to develop solutions to an identified issue through first understanding consumer perspectives about that issue. Accordingly, the first 3 elements of this framework involve collaborating with key stakeholders to establish the study parameters and explore consumer

XSL•FO
RenderX

experiences in depth; we have called this the "needs assessment" phase. The second 3 elements aim to improve that experience through deciding what to improve and then taking action that leads to change; we have called this phase "creation of the online resource." Within this framework, elements can overlap and their order may vary. Our application of this framework is shown in Figure 1. We adhered to the Guidance for Reporting Involvement of Patients and the Public (GRIPP2) to report consumer involvement in this study [35].

Figure 1. Overview of the study phases and co-design elements in the study. Icons made by Uniconlabs, rukanicon, photo3idea_studio, Hilmy Abiyyu A., juicy_fish, and Saepul Nahwan from Flaticon [36].

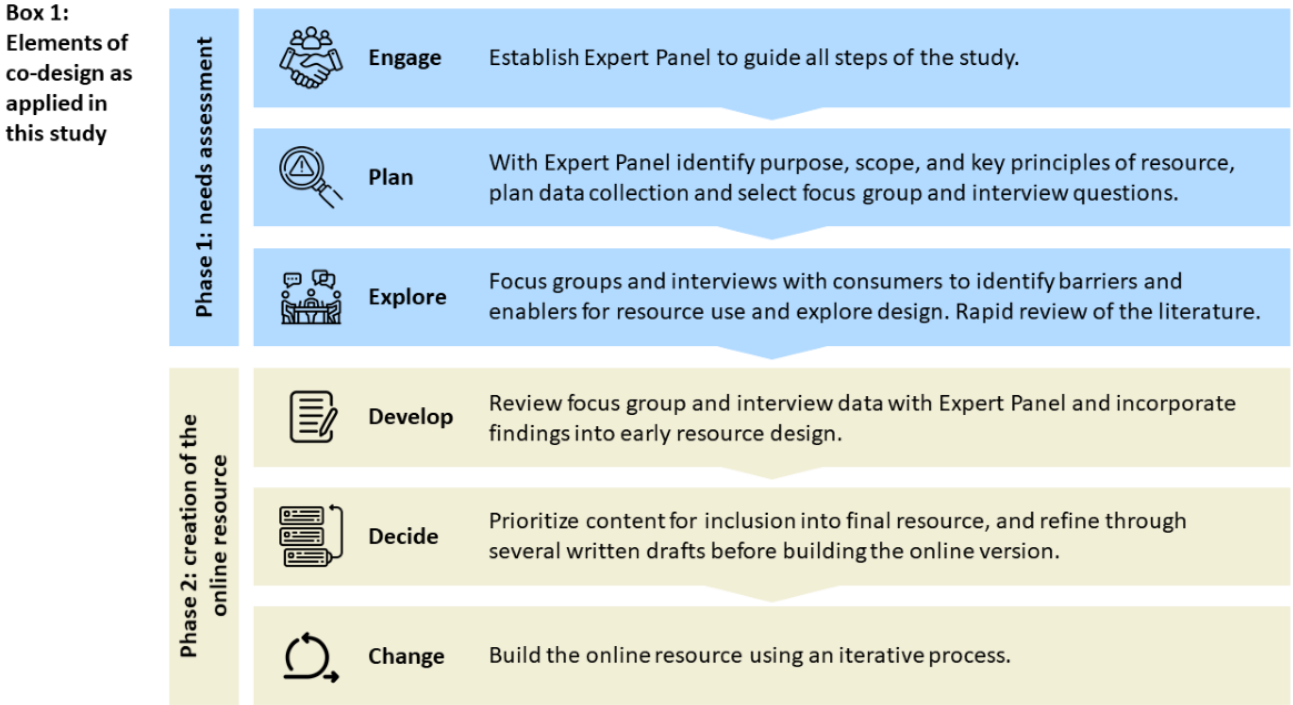

Phase 1: Needs Assessment

Engage and Plan

The needs assessment phase followed the first 3 elements of the Boyd co-design approach (engage, plan, and explore). "Engage" refers to establishing and maintaining meaningful relationships with consumers and clinicians [33]. This was achieved by establishing a co-design team, which we named the "Expert Panel" in recognition of the lived experience and expertise of Panel members. Purposive sampling was used to select Expert Panel members from key stakeholder groups, including health care consumers, clinicians, and academic experts. Consumer members were invited through an expression of interest circulated within the Consumer Branch of the Victorian government agency for quality and safety in health care (Safer Care Victoria) and were remunerated for their involvement. Inclusion criteria for consumers were having lived experience as a health care consumer and being aged 18 years or older. Clinicians and academic experts with experience in health literacy, health communication, and co-design were invited through previous collaborations and networks.

The purpose of the Expert Panel was to guide all co-design aspects of the study. The Expert Panel initially met in October 2020 to agree upon terms of reference for the group, including roles, responsibilities, and frequency of meetings. In this online meeting, planning for the project was also undertaken, including clarifying the overall purpose and scope of the learning resource and developing a plan for the "explore" step of the study. It was determined by the Expert Panel that "explore" should include 2 aspects: interviews or focus groups with consumers and clinicians to identify potential barriers to using check-back within health care interactions, and a rapid literature review of studies addressing communication skills training for consumers.

Explore

Setting and Participants

The "explore" element of this study was conducted across 3 health services in Victoria, Australia with 1 service in a regional setting and 2 in Metropolitan Melbourne. Within the regional service, participants were recruited from the community health service, specifically outpatient rehabilitation programs and chronic disease outpatient clinics. Within the Metropolitan services, participants were recruited from chronic disease outpatient clinics and from the hospital volunteer or consumer representative programs. Of note, the 2 Metropolitan health services support communities in the outer north and south-east of Melbourne, with high proportions of culturally and linguistically diverse populations.

Convenience sampling was used to recruit consumers and clinicians from participating health services to take part in focus groups and interviews. For consumers who were clients of outpatient rehabilitation programs and chronic disease outpatient clinics, client lists were screened and those eligible were contacted by phone or email to discuss the study. For consumer representatives and hospital volunteers, email alerts were sent and those who were interested were invited to contact the

XSL•FO
RenderX

researchers. Clinicians included nursing and allied health professionals from the above sites, who were invited to a focus group or interview via email. Consumers and clinicians who participated in focus groups or interviews were also invited to take part in consultation activities in Phase 2. Inclusion criteria for consumers and clinicians were being aged 18 years or older, with the capacity to provide informed consent.

Focus Groups and Interviews With Consumers and Clinicians

Data were collected via focus groups or semistructured interviews for participants who were unable to take part in a focus group. Focus groups were held separately for consumers and clinicians. Due to COVID-19 restrictions, all focus groups were conducted via teleconference and interviews via telephone. Focus groups and interviews were divided into 2 parts, with the first part aiming to identify factors that might influence consumers' use of the check-back process with health professionals. To provide context for participants, short "vignettes" were written describing challenges that a consumer may face when using check-back. These vignettes were based on a synthesis of patient experience stories from one of the participating health services and were used to "depersonalize" discussions and ensure participants felt comfortable in sharing their ideas without needing to refer to their personal experiences (refer to Multimedia Appendix 1 for an example of a vignette and focus group and interview guide).

A series of seeding questions to prompt discussions was developed using the Capability, Opportunity, and Motivation-Behavior (COM-B) model which provides a theoretically-based approach for developing behavior change interventions [35, 37]. The COM-B model can be used to identify skills and abilities that an individual needs in order to adopt a behavior, as well as the environmental conditions that might support adoption of that behavior. For this study, the behavior was consumer-initiated check-back—that is, confirming understanding through summarizing and repeating back key information to the health professional. The COM-B model considers 3 key domains associated with behavior change: Capability, Opportunity, and Motivation [38]. Figure 2 shows examples of topics used to explore participants' views, based on these 3 domains.

Figure 2. Domains of the COM-B model and their application to focus group and interview topics to identify potential barriers to using check-back within health care interactions. COM-B: Capability, Opportunity, and Motivation-Behavior. Icons made by Uniconlabs, Freepik, and pojok d from Flaticon [36].

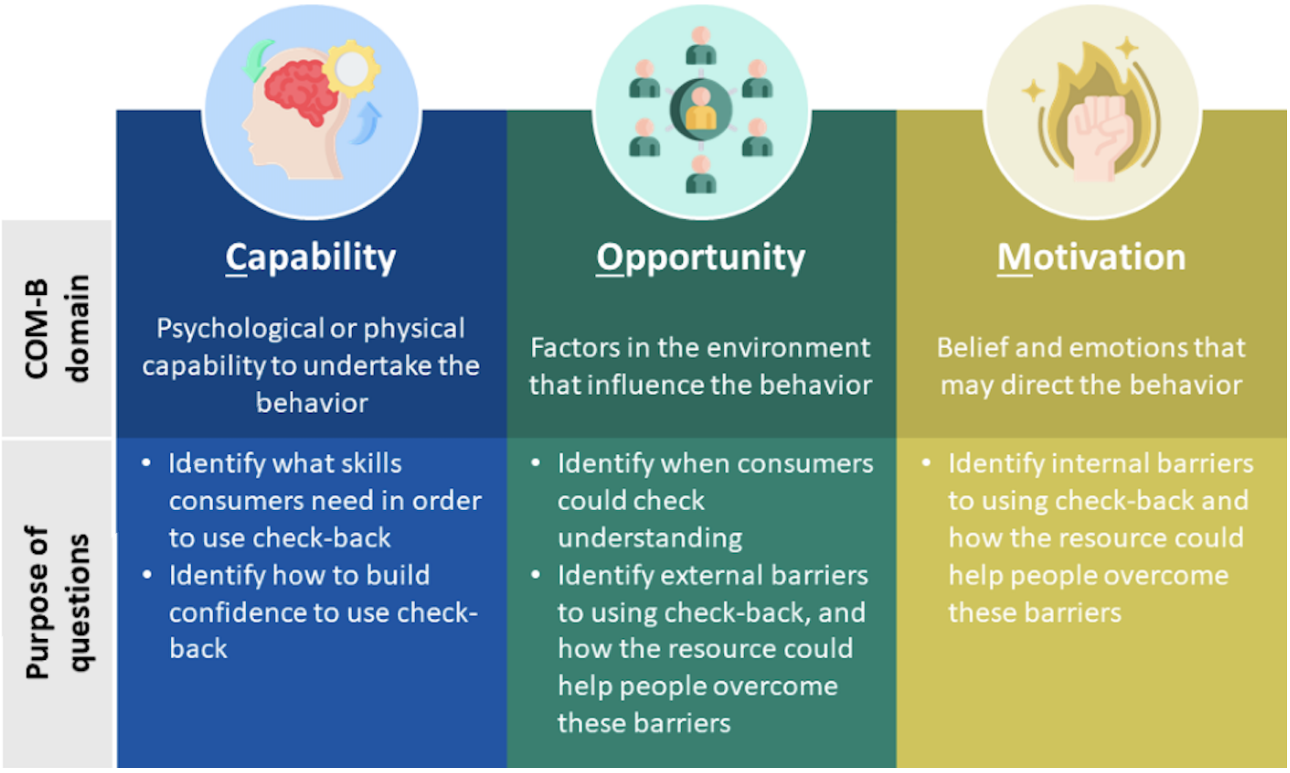

Focus group and interview participants were also asked for ideas about the design of the web-based resource, including how it could be made accessible for consumers. These included the types of scenarios to show in the videos and strategies to ensure the resource was easy to use.

All vignettes and topics for the focus groups and interviews were discussed with the Expert Panel to ensure they were appropriate and would be easily understood by participants. Focus groups and interviews were audio-recorded and transcribed. Deductive thematic analysis was used to analyze the data and identify patterns (themes) within the data that could be used to develop content for the resource [39]. Data were analyzed by 2 researchers (AB and JH). Transcripts were read several times to become familiar with the data and to generate an initial set of themes using the 3 domains of the COM-B model as the parent themes (Capability, Opportunity, and Motivation). The themes that arose from this initial analysis were discussed in depth during an online meeting with the Expert Panel in order to gather their insights into the data, thus

XSL•FO
RenderX

providing a more balanced approach to analysis. Barriers and enablers within each theme were subsequently identified by JH and AB, and then cross-checked with the Expert Panel via email. Quotes that accurately illustrated the themes were also identified.

Literature Review

The needs assessment was also informed by a rapid review of the literature (Multimedia Appendix 2). Briefly, this review aimed to identify best-practice strategies for web-based communication skills training directed at consumers. A narrative approach was used to synthesize the data. Recommendations for best-practice strategies from this review were provided to the Expert Panel for consideration during development of the web-based resource.

Phase 2: Creation of the Online Resource

Develop and Decide

The second phase involved creating the web-based learning resource, following the 3 elements from Boyd co-design framework of develop, decide, and change. For the development and decision steps, findings from the interviews and focus groups and the rapid literature review were discussed with the Expert Panel. In this meeting, identified barriers and enablers to using the check-back process were considered, and the Expert Panel was asked to generate ideas for how these could be addressed in the web-based resource. Ideas were generated by brainstorming through facilitated discussion, a recognized approach in co-design [40-42]. Each theme, barrier, and enabler identified from the focus group and interview data was discussed in turn, with the goal being to identify how each could be targeted within the resource; for example, whether a video-based scenario could be used to demonstrate how to address a specific barrier. The proposed content was then prioritized for inclusion by the Expert Panel. There were no criteria for inclusion of this content, and decisions were made through discussion. Based on these priorities, several written drafts of the resource content were developed and refined, each in collaboration with the Expert Panel.

Change

First Draft of the Resource

The change step involved building an online version of the learning resource using an iterative process. As a first step, audio and video content were recorded over 2 days at one of the participating health services. Four scenarios were filmed with volunteer patients and health professionals acting out a scripted health care interaction. An additional 16 short video clips and brief audio recordings were also recorded, all of which were based on example scenarios and quotes provided by Phase 1 focus group and interview participants. A complete first draft of the web-based resource was then built during April-May 2021. This draft version was reviewed by the Expert Panel who recommended extensive revisions related to functionality, accessibility, and layout. These recommendations are shown in Table 1, with the corresponding revisions that were made.

XSL•FO
RenderX

Table . Summary of identified issues for drafts 1 and 2 of the check-back resource, and examples of revisions made.

Draft and issue category	Identified issues	Example of revisions made
Draft 1 (reviewed by Expert Panel)		
Functionality and navigation	• Resource is difficult to navigate • Not able to move freely through the resource • Not clear how to access videos	• Include an overview and a progress bar • Allow progression without needing to view videos • Provide clearer instructions for using videos and audio
Accessibility	• Difficult for hearing or sight-impaired consumers to use • Limited representation of diversity	• Add captions to videos and include audio buttons • Increase font size and color contrast • Imagery changed to show more diversity
Layout and length	• Not clear what check-back is at the start • Not all videos are relevant to everyone, which makes the resource overcomplicated • Overall resource too long and hard to follow	• Show steps in check-back earlier • Have 1 main video demonstrating check-back and the others as optional viewing • Group content with the most important elements first
Draft 2 (reviewed by consumers and clinicians)		
Functionality and navigation	• Clearer instructions for navigating to the next page are needed • Wording of some instructions about using the resource is ambiguous	• Size of “Next” button increased and instructions simplified • Clarify instructions and use consistent language throughout
Accessibility	• Links to written transcripts needed • Audio buttons not easily visible • More pictures of families needed	• Provide links to a “resources” page with transcripts • Increase size of the audio button • Imagery changed to include families
Layout and length	• Home page text is not grouped next to relevant imagery and is easily missed	• Move some text on the home page to be more obvious and use bold text for emphasis

Second Draft of the Resource

Following these revisions, we consulted with consumers and clinicians to determine the usability, acceptability, and perceived value of draft 2 of the resource and to identify areas for further improvement. Participants were those who had taken part in Phase 1 focus groups and interviews and were willing to take part in this stage. Two approaches were used for this consultation: (1) an anonymous online survey of consumers and clinicians and (2) a focus group with consumers from 1 participating health service. For the online survey, respondents were asked to rate usability (ease of use and navigation), acceptability (whether the resource length was appropriate and overall quality of the module), and perceived value (how interesting the content was and whether the videos clearly showed how to use check-back). A 5-point Likert scale (from strongly disagree to strongly agree) was used. Participants were also asked to rate whether they would recommend it to others (5-point Likert scale; definitely yes to definitely no). An open-text response option was provided for additional comments, including recommendations for any revisions. Data from the Likert scale responses were reported as proportions, and open-text responses were collated. The 60-minute focus group was held with consumers via teleconference, with a link to the learning resource sent to participants ahead of time. During the focus group, researchers presented each section of the resource and asked participants about usability, acceptability, and perception of value, using the questions asked in the online survey to frame the discussion. Participants were also invited to make recommendations for any revisions. The focus group was audio-recorded and transcribed. Data from the survey and focus group were used to make final changes to the online resource in consultation with the Expert Panel (Table 1).

Ethical Considerations

Ethics approval for the use of patient stories and to conduct all focus groups and interviews and filming was provided by the Research Ethics Committees of Monash Health, Northern Health, and Grampians Health Service (Ethics Review Manager reference numbers: 68367, 68046, and 62381). All participants provided written informed consent for focus groups and interviews and implied consent for the online survey. All study data were deidentified for analysis and reporting purposes, and no identifying data are reported. Expert Panel members were compensated for their time based on Safer Care Victoria’s Remuneration for Partners Policy. Consumers who attended focus groups and interviews were offered a AUD $20 (US $14) gift voucher.

XSL•FO
RenderX

Results

Results From Phase 1: Needs Assessment

Expert Panel

The Expert Panel included 4 consumers with extensive lived experience of health care. There were 4 clinician members (including 1 medical specialist and 3 allied health or nursing professionals), and 4 experts in health communication, health literacy, or co-design. One general practitioner (GP) was invited to participate but was only able to provide input into 1 round of email consultation. The Expert Panel represented a diverse range of perspectives across age, gender, and culture and guided all stages of the study through 4 online meetings and 2 rounds of email consultations.

Focus Groups and Interviews

Focus groups or interviews were held with 39 consumers and 16 clinicians; specifically, 3 consumer (n=14) and 2 clinician (n=11) focus groups, and 25 consumer and 5 clinician interviews. Among 39 consumers, ages ranged from late 20s to early 80s, and there were 12 males (31%). Out of 39, 10 consumers (26%) were born outside Australia or spoke a language other than English at home, with 2 of these consumers interviewed in Arabic by an accredited qualified interpreter. Among clinicians, roles included nursing, pharmacy, occupational therapy, and physiotherapy. Table 2 presents findings from Phase 1 focus groups and interviews. Under the 3 COM-B domains, 5 themes were identified: self-efficacy and confidence, pre-existing skills, health professional behaviors and attitudes, information complexity and overload, and internal barriers such as embarrassment. For each theme, barriers and enablers are shown, each illustrated by a quote.

Table . Themes, barriers, and enablers to using check-back within health care interactions as identified from focus group and interview data with consumers (n=39) and clinicians (n=16).

Domain and theme (frequency)	Barriers and enablers from focus groups and interviews	Example quotes
COM-B domain: Capability		
Self-efficacy and confidence (7 consumers, 2 clinicians)	• Many people would find it difficult to ask clinicians to repeat information. • The resource could include messages that aim to counteract shyness or beliefs about hierarchy and highlight how common it is for people to not understand information.	• "Have you heard it before that older people have them on a pedestal and are a bit timid on asking questions?" (Client #4) • "People need to be encouraged to think that you can ask questions you can double check...because it's quite normal to not understand." (Volunteer #4)
Pre-existing skills (9 consumers, 1 clinician)	• People may already have the skills to check understanding as they often do this in everyday life.	• "Yes, I have a bad memory for names. When someone gives me their name, I'll repeat it back to them." (Volunteer #2)
COM-B domain: Opportunity		
Health professional behaviors and attitudes (9 consumers, 2 clinicians)	• Clinicians' behavior can limit a consumers' confidence and empowerment, especially if they appear authoritative or disengaged. • The resource could emphasize the consumers' right to ask questions and to clarify information.	• "...and they have got the attitude of them and us, it's like well what do you know, I'm the specialist." (Client #1) • "We have to empower consumers to have the confidence to do that in a non-confrontational way and I think that perhaps the videos would assist with that." (Consumer representative #1)
Information complexity and overload (11 consumers)	• Clinicians frequently use complex terms, and the amount of information can be overwhelming and difficult to remember. • Include tips such as asking the doctor to write things down. Highlight the value of bringing a support person with you.	• "That's often the problem, too many abbreviations and jargon is used." (Consumer representative #2) • "I ask my doctor if he minds jotting down those points that he highlighted so I can have something once I'm out the door, something to ponder on." (Volunteer #6)
COM-B domain: Motivation		
Internal barriers (20 consumers, 2 clinicians)	• Internal factors include embarrassment or lack of confidence. Consumers may be overwhelmed by the situation and by emotions such as worry or fear. • The resource could emphasize the benefits of understanding information and include messages that give "permission" to clarify understanding.	• "You've got patients waiting outside, I don't want to hold the doctor up by asking silly questions because I don't understand." (Volunteer #7) • "That no question is a silly question. …we'd rather you understand and question, than go home and do the wrong thing." (Client #4)

XSL•FO
RenderX

Literature Review

From the rapid literature review (Multimedia Appendix 2), 15 relevant recommendations were identified. These were presented to the Expert Panel for consideration during development of the check-back resource. Key recommendations included the importance of raising consumers' awareness about their role in health communication and using approaches to build self-efficacy in communication, for example, demonstrate behaviors through video clips of peers. Text font should be large and clear with consistent layouts throughout, and text should be brief and jargon-free. For culturally diverse populations, the use of visual aids and including people from various cultural groups is recommended, as is the use of transcripts of any video commentary to support people with a hearing impairment. A mix of modalities is effective for the delivery of information; these include infographics, videos, text, animation, and narration. Including an interactive component can increase user engagement, and the addition of additional printed resources, such as question prompt lists, is also recommended.

Results From Phase 2: Creation of the Web-Based Resource

The check-back resource was built through an iterative process. As described in the "Methods", the Expert Panel recommended extensive revisions to the first complete draft of the resource (Table 1); a second complete draft was then developed, followed by a further round of consultation with consumers and clinicians through an online survey and focus group.

Online Survey

Out of 55, 11 (20%) respondents completed the online survey. Table 3 shows a summary of survey findings relating to usability, acceptability, and perception of value.

Table . Findings from the online survey to determine usability, acceptability, and perceived value of the check-back resource (n=11 clinicians and consumers).

Question	Response 1, n (%)	Response 2, n (%)	Response 3, n (%)	Response 4, n (%)	Response 5, n (%)
The resource was...[a]					
Easy to use	0 (0)	2 (18)	0 (0)	3 (27)	6 (55)
Easy to navigate	0 (0)	2 (18)	1 (9)	1 (9)	7 (64)
Interesting	0 (0)	1 (9)	0 (0)	5 (45)	5 (45)
Right length	0 (0)	1 (9)	3 (27)	3 (27)	4 (36)
Videos showed check-back (n=10)	0 (0)	0 (0)	0 (0)	4 (40)	6 (60)
How would you rate the quality of the module[b]?	0 (0)	1 (9)	0 (0)	4 (36)	6 (55)
Would you recommend this module to others for learning about check-back[c]?	0 (0)	1 (9)	0 (0)	4 (36)	6 (55)

[a]Responses: 1=Strongly disagree, 2=Somewhat disagree, 3=Neither, 4=Somewhat agree, and 5=Strongly agree.

[b]Responses: 1=Extremely bad, 2=Somewhat bad, 3=Neither good nor bad, 4=Somewhat good, and 5=Extremely good.

[c]Responses: 1=Definitely not, 2=Probably not, 3=Might or might not, 4=Probably yes, and 5=Definitely yes.

Overall, most participants agreed or strongly agreed that the resource was easy to use (9/11, 82%) and navigate (8/11, 71%), the content was interesting (10/11, 90%), and the videos demonstrated how to use check-back (10/10, 100%). Most participants (10/11, 91%) rated the quality of the module as somewhat or extremely good and said they would recommend it to others (10/11, 91%). One respondent said they would not recommend it, commenting that there were too few videos demonstrating what check-back is. Open-text comments revealed that the resource was very engaging, with suggestions for improvement mostly focused on enhancing the navigability of the resource.

Consumer Focus Group

Comments on draft 2 received from the consumer focus group (n=5) also focused on navigation, including the need for clearer links to audio and written transcripts. Minor revisions to images and text were also recommended, including more use of family images. In response to these findings and those from the online survey, additional revisions to the resource were made in collaboration with the Expert Panel (Table 1). The final resource was approved by the Expert Panel in October 2021 and made publicly available from April 2022. An overview of the check-back resource is shown in Figure 3.

XSL•FO
RenderX

Figure 3. Overview of the "check-back" learning resource [43]. GP: general practitioner. Icons made by Saepul Nahwan, Becris, Freepik, PixelVerse, Roundicons, smashingstocks, and vectorsmarket15 from Flaticon [36].

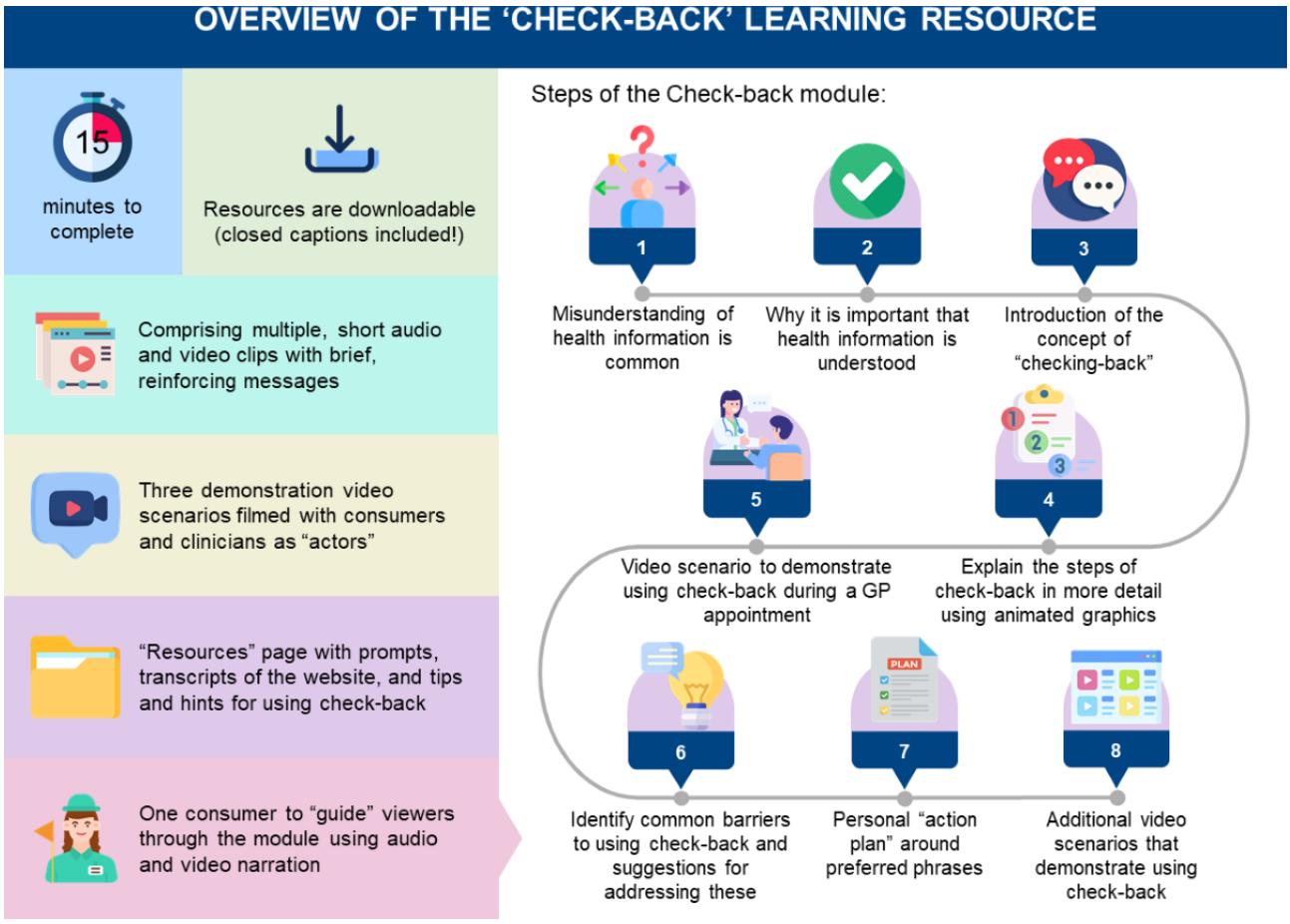

Discussion

Principal Findings

This study describes the development of a web-based learning resource to support consumer-driven check-back, using a co-design approach to partner with key stakeholders, including consumers with lived experience and health professionals. Early feedback suggests that the resource is engaging and accessible for consumers, with content that is interesting and relevant. In part, this relevance may be due to the use of the co-design approach with its focus on embedding the perspectives of consumers and other end users throughout the design and development process. Co-design has become increasingly used in health care to undertake service evaluation and quality improvement activities [25]. However, few studies have described using co-design in patient-provider communication, especially those that engage consumers as a core component of the process or that develop web-based resources [28,30,44]. One study used co-design to address communication barriers for patients with heart failure, resulting in a paper-based question prompt list for patients [28]. In another study, a communication training package targeted toward radiographers was co-designed with patients with breast cancer [30]. This study adds to the limited evidence for the value of using co-design to develop consumer-oriented web-based resources, including those aiming to assist consumers to build their communication skills in any healthcare setting or with any clinician.

Challenges that have been identified with using co-design include the exclusion of people with lived experience from the process, having the methods established before beginning the project, and not allowing for equal partnership between consumers, health professionals, and researchers in the design process [45]. In this study, we aimed to ensure that consumers and health professionals guided the content and design of the resource, thus providing an essential "end user" perspective. Through active and considerate listening and following an iterative process that remained focused on a clear end product, we adhered to the 5 core principles for successful co-design: being inclusive, respectful, participative, iterative, and outcomes-focused [46].

This study also adds to the limited evidence describing web-based communication training resources for consumers. A 2017 systematic review of the literature on training consumers in communication skills identified that just 5 of the 32 included studies used web-based training [21]. These included 3 randomized controlled trials: a web-based intervention to increase consumers' communication self-efficacy [44], an interactive website that used consumer narratives to prepare participants for communicating with doctors [47], and a consumer and physician training intervention comprising multiple demonstration videos [48]. Overall, these studies demonstrated that web-based interventions that assist consumers to improve communication are effective at increasing their satisfaction and skills [44,47,48], and at supporting them to

XSL•FO
RenderX

achieve treatment-related goals [23]. While the check-back resource has not yet been formally evaluated, these studies identify aspects of evaluation that may be relevant, including consumer satisfaction with the resource, whether it results in behavior change within health care interactions (such as increased confidence to ask questions), and whether it is effective at increasing understanding and recall in the short and long term.

The final check-back learning resource is multimodal in its delivery. Many of its features have been shown in other studies to be effective at improving consumer understanding, including the use of videos to model behavior [49], infographics [50], animations [51], and guides or narrators [52]. The use of multiple modes of delivery in the check-back resource may also address the diversity of learning styles found within any population [53]. This may be particularly relevant for consumers with lower health literacy who are often more likely to experience poor health care communication [12,53]. Web-based learning using a mix of modalities (including images) has been shown to be feasible and effective among people with low health literacy [54], and may also be useful for people with cognitive challenges, or those who do not speak the local language. Digital health literacy (defined as the ability to seek, find, understand, and appraise health information from electronic sources, and to apply that knowledge to solve health-related problems) [55] can also be enhanced by resources that are easily accessible and usable [56].

In terms of future steps, tailoring of the check-back resource to particular cultural groups will be required, as evidence shows that ethnic and racial minorities receive suboptimal health communication [21], are less likely to ask questions [57], and may frequently misunderstand information [58]. However, cultural tailoring of this resource will involve more than linguistic translation as cultural values, norms, or beliefs need to be fully considered [58,59]. Use of the co-design approach to ensure cultural relevance in future adaptations is warranted. Other adaptations may include creating versions of check-back that are accessible for consumers with few digital skills or limited access to the internet. Given the known digital divide in high-income countries, it is important to provide alternative approaches for consumers to learn about check-back to enhance equity and accessibility [60.

Finally, it is important to acknowledge that addressing the problem of poor health communication by focusing only on building consumer skills will be insufficient. Health and community providers and policy makers may consider introducing this resource to consumers as part of a suite of multilevel strategies aimed at building organizational responsiveness to the communication and health literacy needs of their consumers [61,62].

Limitations

To our knowledge, this is the first study to co-design a web-based, interactive education resource for consumers to improve their ability to communicate effectively during health care interactions. There were some limitations with this study, including the limited representativeness of the sample for the needs assessment phase, as participants were from Victoria, Australia only. Recruitment was also limited by COVID-19 restrictions, and therefore, we needed to include a number of consumer representatives who may have been more empowered than a “typical” health care consumer. In addition, information on the number of consumers and clinicians that were approached to participate in focus groups was not consistently available, and therefore, we cannot provide participation rates. Participants were primarily English-speaking only, although several culturally and linguistically diverse consumers were included in focus groups and interviews. While the Expert Panel included a medical specialist, there was only limited input from a GP. Given that consumers are most likely to use check-back with their GP, this may have affected the identification of barriers to its use.

Conclusions

The issue of poor communication in health care is well-recognized, and innovative strategies are required, including those that enable consumers to participate in active 2-way communication. Through the use of co-design with key stakeholders, we developed an accessible, consumer-centered learning resource that may have high applicability and utility for consumers across a broad range of health and community settings. This study provides a starting point for future research into consumer-led initiatives to improve communication, including for population groups known to have significant challenges in this area.

Acknowledgments

The authors wish to thank the health consumers who were involved throughout all stages, including study design and development and evaluation of the check-back resource. The project was guided by an Expert Panel including 4 health consumers, 3 of whom are co-authors on this manuscript.

Funding

Funding for this study was provided by Safer Care Victoria. AB is funded through a National Health and Medical Research Council Investigator Grant (ID: 198006)

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Authors' Contributions

Conceptualization: AB (lead), JH (equal), LH (equal), AC (equal), EF-J (equal), S Rawlinson (equal), RJ (equal)
Formal analysis: AB (lead), JH (supporting)
Funding acquisition: AB (lead), RJ (equal)
Investigation: AB (lead), NC (supporting), AWS (supporting), S Rogers (supporting), RJ (equal)
Methodology: AB (lead), JH (equal), AWS (equal), LH (equal), S Rogers (equal), AC (equal), EF-J (equal), S Rawlinson (equal), RJ (equal)
Project administration: AB (lead), NC (supporting)
Supervision: AB (lead)
Writing–original draft: AB (lead)
Writing–review and editing: JH (equal), AWS (equal), LH (equal), S Rogers (equal), AC (equal), EFJ (equal), S Rawlinson (equal), RJ (equal)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Example of a vignette used in focus groups and interviews, and focus group and interview guide.
[DOCX File, 17 KB - formative_v9i1e77263_app1.docx]

Multimedia Appendix 2
Consumer check-back: a rapid literature review to inform development of an online learning resource for consumers.
[DOCX File, 270 KB - formative_v9i1e77263_app2.docx]

References

1. Murugesu L, Heijmans M, Rademakers J, Fransen MP. Challenges and solutions in communication with patients with low health literacy: perspectives of healthcare providers. PLoS One 2022;17(5):e0267782. [doi: 10.1371/journal.pone.0267782] [Medline: 35507632]
2. Berger Z, Dembitzer A, Beach MC. Reason for hospital admission: a pilot study comparing patient statements with chart reports. Narrat Inq Bioeth 2013;3(1):67-79. [doi: 10.1353/nib.2013.0010] [Medline: 24407006]
3. Kreps GL. Promoting patient comprehension of relevant health information. Isr J Health Policy Res 2018 Sep 18;7(1):56. [doi: 10.1186/s13584-018-0250-z] [Medline: 30236157]
4. Laws MB, Lee Y, Taubin T, Rogers WH, Wilson IB. Factors associated with patient recall of key information in ambulatory specialty care visits: results of an innovative methodology. PLoS One 2018;13(2):e0191940. [doi: 10.1371/journal.pone.0191940] [Medline: 29389994]
5. Dorsey BF, Kamimura A, Cook LJ, et al. Communication gaps between providers and caregivers of patients in a pediatric emergency department. J Patient Exp 2022;9:23743735221112223. [doi: 10.1177/23743735221112223] [Medline: 35836779]
6. Lipson-Smith R, Hyatt A, Murray A, et al. Measuring recall of medical information in non-English-speaking people with cancer: a methodology. Health Expect 2018 Feb;21(1):288-299. [doi: 10.1111/hex.12614] [Medline: 28940931]
7. Richard C, Glaser E, Lussier MT. Communication and patient participation influencing patient recall of treatment discussions. Health Expect 2017 Aug;20(4):760-770. [doi: 10.1111/hex.12515] [Medline: 27868327]
8. Friis K, Lasgaard M, Pedersen MH, Duncan P, Maindal HT. Health literacy, multimorbidity, and patient-perceived treatment burden in individuals with cardiovascular disease. A Danish population-based study. Patient Educ Couns 2019 Oct;102(10):1932-1938. [doi: 10.1016/j.pec.2019.05.013] [Medline: 31151781]
9. Kwame A, Petrucka PM. A literature-based study of patient-centered care and communication in nurse-patient interactions: barriers, facilitators, and the way forward. BMC Nurs 2021 Sep 3;20(1):158. [doi: 10.1186/s12912-021-00684-2] [Medline: 34479560]
10. Friis K, Lasgaard M, Osborne RH, Maindal HT. Gaps in understanding health and engagement with healthcare providers across common long-term conditions: a population survey of health literacy in 29,473 Danish citizens. BMJ Open 2016 Jan 14;6(1):e009627. [doi: 10.1136/bmjopen-2015-009627] [Medline: 26769783]
11. Bengtsson KR, Rognan SE, Kälvemark Sporrong S, et al. Health literacy in medication communication during hospital discharge: a qualitative study at an internal medicines ward in Norway. BMJ Open 2022 Jun 9;12(6):e058473. [doi: 10.1136/bmjopen-2021-058473] [Medline: 35680250]
12. Noordman J, van Vliet L, Kaunang M, van den Muijsenbergh M, Boland G, van Dulmen S. Towards appropriate information provision for and decision-making with patients with limited health literacy in hospital-based palliative care in Western countries: a scoping review into available communication strategies and tools for healthcare providers. BMC Palliat Care 2019 Apr 12;18(1):37. [doi: 10.1186/s12904-019-0421-x] [Medline: 30979368]

XSL•FO
RenderX

13. Rozier RG, Horowitz AM, Podschun G. Dentist-patient communication techniques used in the United States: the results of a national survey. J Am Dent Assoc 2011 May;142(5):518-530. [doi: 10.14219/jada.archive.2011.0222] [Medline: 21531934]
14. Dantic DE. A critical review of the effectiveness of 'teach-back' technique in teaching COPD patients self-management using respiratory inhalers. Health Educ J 2014 Jan;73(1):41-50. [doi: 10.1177/0017896912469575]
15. Ha Dinh TT, Bonner A, Clark R, Ramsbotham J, Hines S. The effectiveness of the teach-back method on adherence and self-management in health education for people with chronic disease: a systematic review. JBI Database System Rev Implement Rep 2016 Jan;14(1):210-247. [doi: 10.11124/jbisrir-2016-2296] [Medline: 26878928]
16. Talevski J, Wong Shee A, Rasmussen B, Kemp G, Beauchamp A. Teach-back: a systematic review of implementation and impacts. PLoS One 2020;15(4):e0231350. [doi: 10.1371/journal.pone.0231350] [Medline: 32287296]
17. Tran S, Bennett G, Richmond J, et al. "Teach-back" is a simple communication tool that improves disease knowledge in people with chronic hepatitis B - a pilot randomized controlled study. BMC Public Health 2019 Oct 23;19(1):1355. [doi: 10.1186/s12889-019-7658-4] [Medline: 31647001]
18. Badaczewski A, Bauman LJ, Blank AE, et al. Relationship between Teach-back and patient-centered communication in primary care pediatric encounters. Patient Educ Couns 2017 Jul;100(7):1345-1352. [doi: 10.1016/j.pec.2017.02.022] [Medline: 28254516]
19. Feinberg I, Ogrodnick MM, Hendrick RC, Bates K, Johnson K, Wang B. Perception versus reality: the use of Teach Back by medical residents. Health Lit Res Pract 2019 Apr;3(2):e117-e126. [doi: 10.3928/24748307-20190501-01] [Medline: 31294313]
20. Samuels-Kalow M, Hardy E, Rhodes K, Mollen C. "Like a dialogue": Teach-back in the emergency department. Patient Educ Couns 2016 Apr;99(4):549-554. [doi: 10.1016/j.pec.2015.10.030] [Medline: 26597382]
21. D'Agostino TA, Atkinson TM, Latella LE, et al. Promoting patient participation in healthcare interactions through communication skills training: a systematic review. Patient Educ Couns 2017 Jul;100(7):1247-1257. [doi: 10.1016/j.pec.2017.02.016] [Medline: 28238421]
22. Alders I, Smits C, Brand P, van Dulmen S. Does patient coaching make a difference in patient-physician communication during specialist consultations? A systematic review. Patient Educ Couns 2017 May;100(5):882-896. [doi: 10.1016/j.pec.2016.12.029] [Medline: 28089309]
23. Glaser E, Richard C, Lussier MT. The impact of a patient web communication intervention on reaching treatment suggested guidelines for chronic diseases: a randomized controlled trial. Patient Educ Couns 2017 Nov;100(11):2062-2070. [doi: 10.1016/j.pec.2017.05.022] [Medline: 28535926]
24. Cvetanovska N, Jessup RL, Wong Shee A, Rogers S, Beauchamp A. Patients' perspectives of factors influencing active participation in healthcare interactions: a qualitative study. Patient Educ Couns 2023 Sep;114:107808. [doi: 10.1016/j.pec.2023.107808] [Medline: 37263050]
25. O'Brien J, Fossey E, Palmer VJ. A scoping review of the use of co-design methods with culturally and linguistically diverse communities to improve or adapt mental health services. Health Soc Care Community 2021 Jan;29(1):1-17. [doi: 10.1111/hsc.13105] [Medline: 32686881]
26. Banbury A, Pedell S, Parkinson L, Byrne L. Using the Double Diamond model to co-design a dementia caregivers telehealth peer support program. J Telemed Telecare 2021 Dec;27(10):667-673. [doi: 10.1177/1357633X211048980] [Medline: 34726994]
27. Slattery P, Saeri AK, Bragge P. Research co-design in health: a rapid overview of reviews. Health Res Policy Syst 2020 Feb 11;18(1):17. [doi: 10.1186/s12961-020-0528-9] [Medline: 32046728]
28. Hjelmfors L, Strömberg A, Friedrichsen M, Sandgren A, Mårtensson J, Jaarsma T. Using co-design to develop an intervention to improve communication about the heart failure trajectory and end-of-life care. BMC Palliat Care 2018 Jun 11;17(1):85. [doi: 10.1186/s12904-018-0340-2] [Medline: 29890974]
29. Kildea J, Battista J, Cabral B, et al. Design and development of a person-centered patient portal using participatory stakeholder co-design. J Med Internet Res 2019 Feb 11;21(2):e11371. [doi: 10.2196/11371] [Medline: 30741643]
30. van Beusekom M, Cameron J, Bedi C, Banks E, Harris R, Humphris G. Using co-design with breast cancer patients and radiographers to develop "KEW" communication skills training. Front Psychol 2021;12:629122. [doi: 10.3389/fpsyg.2021.629122] [Medline: 33692727]
31. Webb MJ, Wadley G, Sanci LA. Improving patient-centered care for young people in general practice with a codesigned screening app: mixed methods study. JMIR Mhealth Uhealth 2017 Aug 11;5(8):e118. [doi: 10.2196/mhealth.7816] [Medline: 28801302]
32. Teach-back. 2018. URL: https://teachback.org/ [accessed 2025-11-25]
33. Boyd H, McKernon S, Mullin B, Old A. Improving healthcare through the use of co-design. N Z Med J 2012 Jun 29;125(1357):76-87. [Medline: 22854362]
34. Azar D, Wang S, Flemming-Judge L, et al. Using codesign to develop a health literacy intervention to improve the accessibility and acceptability of cardiac services: the Equal Hearts Study. Health Expect 2025 Jun;28(3):e70328. [doi: 10.1111/hex.70328] [Medline: 40528538]

XSL•FO
RenderX

35. Staniszewska S, Brett J, Simera I, et al. GRIPP2 reporting checklists: tools to improve reporting of patient and public involvement in research. Res Involv Engagem 2017;3:13. [doi: 10.1186/s40900-017-0062-2] [Medline: 29062538]
36. Flaticon. URL: https://www.flaticon.com/ [accessed 2025-12-04]
37. Barker F, Atkins L, de Lusignan S. Applying the COM-B behaviour model and behaviour change wheel to develop an intervention to improve hearing-aid use in adult auditory rehabilitation. Int J Audiol 2016 Jul;55 Suppl 3:S90-S98. [doi: 10.3109/14992027.2015.1120894] [Medline: 27420547]
38. Handley MA, Harleman E, Gonzalez-Mendez E, et al. Applying the COM-B model to creation of an IT-enabled health coaching and resource linkage program for low-income Latina moms with recent gestational diabetes: the STAR MAMA program. Implement Sci 2016 May 18;11(1):73. [doi: 10.1186/s13012-016-0426-2] [Medline: 27193580]
39. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
40. Burns C, Kilkenny MF, Purvis T, et al. Processes underpinning successful co-design: lessons from a digital health project. Health Expect 2025 Jun;28(3):e70272. [doi: 10.1111/hex.70272] [Medline: 40405447]
41. Iniesto F, Charitonos K, Littlejohn A. A review of research with co-design methods in health education. Open Education Studies 2022 Jan 1;4(1):273-295. [doi: 10.1515/edu-2022-0017]
42. Ryan L, Wenke R, Carlini J, et al. Co-design of an Australian health service framework and implementation plan for involving consumers in research. Public Health Res Pract 2025 May;35:PU24021. [doi: 10.1071/PU24021] [Medline: 40456564]
43. Beauchamp A, Hilbers J, Jessup R, Wong Shee A. Check-back. 2021. URL: https://checkback.org [accessed 2025-11-25]
44. van Bruinessen IR, van Weel-Baumgarten EM, Gouw H, Zijlstra JM, van Dulmen S. An integrated process and outcome evaluation of a web-based communication tool for patients with malignant lymphoma: randomized controlled trial. J Med Internet Res 2016 Jul 29;18(7):e206. [doi: 10.2196/jmir.5877] [Medline: 27473173]
45. Ní Shé É, Harrison R. Mitigating unintended consequences of co-design in health care. Health Expect 2021 Oct;24(5):1551-1556. [doi: 10.1111/hex.13308] [Medline: 34339528]
46. Brand G, Sheers C, Wise S, et al. A research approach for co-designing education with healthcare consumers. Med Educ 2021 May;55(5):574-581. [doi: 10.1111/medu.14411] [Medline: 33155301]
47. Meropol NJ, Egleston BL, Buzaglo JS, et al. A web-based communication aid for patients with cancer: the CONNECT study. Cancer 2013 Apr 1;119(7):1437-1445. [doi: 10.1002/cncr.27874] [Medline: 23335150]
48. Roter DL, Wexler R, Naragon P, et al. The impact of patient and physician computer mediated communication skill training on reported communication and patient satisfaction. Patient Educ Couns 2012 Sep;88(3):406-413. [doi: 10.1016/j.pec.2012.06.020] [Medline: 22789149]
49. Steinwachs DM, Roter DL, Skinner EA, et al. A web-based program to empower patients who have schizophrenia to discuss quality of care with mental health providers. Psychiatr Serv 2011 Nov;62(11):1296-1302. [doi: 10.1176/ps.62.11.pss6211_1296] [Medline: 22211208]
50. Finocchario-Kessler S, Catley D, Thomson D, Bradley-Ewing A, Berkley-Patton J, Goggin K. Patient communication tools to enhance ART adherence counseling in low and high resource settings. Patient Educ Couns 2012 Oct;89(1):163-170. [doi: 10.1016/j.pec.2012.03.020] [Medline: 22575433]
51. Furuno Y, Sasajima H. Medical comics as tools to aid in obtaining informed consent for stroke care. Medicine (Abingdon) 2015;94(26):e1077. [doi: 10.1097/MD.0000000000001077]
52. Brown-Johnson CG, Berrean B, Cataldo JK. Development and usability evaluation of the mHealth Tool for Lung Cancer (mHealth TLC): a virtual world health game for lung cancer patients. Patient Educ Couns 2015 Apr;98(4):506-511. [doi: 10.1016/j.pec.2014.12.006] [Medline: 25620075]
53. Brooks C, Ballinger C, Nutbeam D, Adams J. The importance of building trust and tailoring interactions when meeting older adults' health literacy needs. Disabil Rehabil 2017 Nov;39(23):2428-2435. [doi: 10.1080/09638288.2016.1231849] [Medline: 27712121]
54. Jacobs RJ, Lou JQ, Ownby RL, Caballero J. A systematic review of eHealth interventions to improve health literacy. Health Informatics J 2016 Jun;22(2):81-98. [doi: 10.1177/1460458214534092] [Medline: 24916567]
55. Regional digital health action plan 2023-2030. World Health Organization. 2023. URL: https://www.who.int/europe/publications/i/item/EUR-RC72-5 [accessed 2025-09-29]
56. Wamala Andersson S, Gonzalez MP. Digital health literacy-a key factor in realizing the value of digital transformation in healthcare. Front Digit Health 2025;7:1461342. [doi: 10.3389/fdgth.2025.1461342] [Medline: 40538571]
57. Torres DX, Lu WY, Uratsu CS, Sterling SA, Grant RW. Knowing how to ask good questions: comparing Latinos and non-Latino Whites enrolled in a cardiovascular disease prevention study. Perm J 2019;23:18-258. [doi: 10.7812/TPP/18-258] [Medline: 30939290]
58. Rochon D, Ross MW, Looney C, Nepal VP, Price AJ, Giordano TP. Communication strategies to improve HIV treatment adherence. Health Commun 2011;26(5):461-467. [doi: 10.1080/10410236.2011.554168] [Medline: 21432703]
59. López L, Tan-McGrory A, Horner G, Betancourt JR. Eliminating disparities among Latinos with type 2 diabetes: effective eHealth strategies. J Diabetes Complicat 2016 Apr;30(3):554-560. [doi: 10.1016/j.jdiacomp.2015.12.003]

XSL•FO
RenderX

60. Lythreatis S, Singh SK, El-Kassar AN. The digital divide: a review and future research agenda. Technol Forecast Soc Change 2022 Feb;175:121359. [doi: 10.1016/j.techfore.2021.121359]
61. Sentell T, Foss-Durant A, Patil U, Taira D, Paasche-Orlow MK, Trinacty CM. Organizational health literacy: opportunities for patient-centered care in the wake of COVID-19. Qual Manag Health Care 2021;30(1):49-60. [doi: 10.1097/QMH.0000000000000279] [Medline: 33229999]
62. Epstein RM, Duberstein PR, Fenton JJ, et al. Effect of a patient-centered communication intervention on oncologist-patient communication, quality of life, and health care utilization in advanced cancer: the VOICE Randomized Clinical Trial. JAMA Oncol 2017 Jan 1;3(1):92-100. [doi: 10.1001/jamaoncol.2016.4373] [Medline: 27612178]

Abbreviations

COM-B: Capability, Opportunity, Motivation-Behavior
GP: general practitioner
GRIPP2: Guidance for Reporting Involvement of Patients and the Public

Edited by A Stone, A Mavragani; submitted 10.05.25; peer-reviewed by M Tracy, P McEnhill; revised version received 29.10.25; accepted 29.10.25; published 12.12.25.

Please cite as:
Beauchamp A, Hilbers J, Cvetanovska N, Wong Shee A, Horvat L, Rogers S, Cooper A, Flemming-Judge E, Rawlinson S, Jessup R
Consumer Co-Design of an Online Resource to Build Communication Skills of Health Consumers: Mixed Methods Study
JMIR Form Res 2025;9:e77263
URL: https://formative.jmir.org/2025/1/e77263
doi:10.2196/77263

© Alison Beauchamp, Julieanne Hilbers, Natali Cvetanovska, Anna Wong Shee, Lidia Horvat, Sandra Rogers, Andrea Cooper, Elizabeth Flemming-Judge, Sue Rawlinson, Rebecca Jessup. Originally published in JMIR Formative Research (https://formative.jmir.org), 12.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Exploration of Factors That Affect Engagement With the Experience Sampling Method and Service Users' Experience of This Within the AVATAR2 Trial: Mixed Methods Study

Sophie Dennard[1], DClinPsy; Philippa Garety[2], DClinPsy, PhD; Clementine Edwards[1], DClinPsy, PhD; Andrew Gumley[3], PhD; Oliver Owrid[1], MSc; Lucy Miller[1], MSc; Stephanie Allan[3], PhD; Alison Duerden[1]; Francis Yanga[3], HND; Chloe Burns[3], PgDip; Helena Fletcher[3]; Amy Grant[1]

[1]Institute of Psychiatry, Psychology & Neuroscience, King's College London, 16 De Crespigny Park, London, United Kingdom

[2]Institute of Psychiatry, Psychology and Neuroscience (IoPPN), King's College London, London, United Kingdom

[3]School of Health and Wellbeing, University of Glasgow, Glasgow, United Kingdom

Corresponding Author:
Sophie Dennard, DClinPsy
Institute of Psychiatry, Psychology & Neuroscience, King's College London, 16 De Crespigny Park, London, United Kingdom

Abstract

Background: Experience sampling methodology (ESM) is an assessment method used in psychosis research. Symptom severity and gender may be associated with ESM engagement. Exploring qualitative experiences of using ESM among people with psychosis should aid developing more relevant, accessible digital assessments.

Objective: This study aimed to examine factors that could affect engagement with ESM, such as associations of completion rates with age, ethnicity, gender, and clinical severity. It also aimed to explore qualitatively service users' experiences of using this data collection method.

Methods: Data from 134/207 AVATAR2 trial (ISRCTN55682735) participants were used to evaluate associations between demographic variables, symptom severity, and ESM completion rates. Trial participants were purposively sampled to participate in an interview to discuss their experiences of using ESM or to discuss reasons why they chose not to use it.

Results: Multiple regression analyses of 134 participants found that age, gender, ethnicity, and clinical severity were not associated with ESM completion rates ($F_{5,128}$=0.548; P=.74). A thematic analysis of 17 participant interviews found 3 overarching themes: *Factors affecting engagement with ESM*, *Perceived benefits of ESM*, *and Suggestions for improvement*. These themes described how ESM has multiple benefits for people with psychosis, including increasing knowledge and awareness of mental health. ESM was straightforward and easy to use; however, engaging in other activities, experiencing positive symptoms, little experience using technology, and trial involvement impacted engagement. Participant's decision to use ESM could be influenced by concerns about security and privacy.

Conclusions: Recommendations are made on how engagement with ESM can be improved, making it easier to use this method with this population, including providing increased support or training when using digital-based assessment or intervention as well as providing information on how digital data are used and recorded.

(JMIR Form Res 2025;9:e78204) doi:10.2196/78204

KEYWORDS

digital assessment; psychosis; qualitative; smartphone app; patient and public involvement; interviews

Introduction

The experience sampling methodology (ESM) is an assessment method that prompts individuals to complete repeated, in-the-moment, questionnaires several times a day [1]. This methodology allows researchers and clinicians to obtain information about an individual's everyday life experiences, with less reliance on retrospective memories. ESM has therefore been used in a wide variety of research and clinical settings in populations of individuals with diagnoses of schizophrenia and psychosis [2-4].

Research has indicated that the validity of ESM may be impacted by different factors, including the items used [5]; however, it may also be partly affected by levels of engagement of those completing the assessments. Although the findings are inconsistent, some factors affecting engagement with ESM have been investigated in people with psychosis. Research has found higher rates of engagement among females, older participants, and people identifying as White [5-7]. Furthermore, studies

XSL•FO
RenderX

have reported that those with increased symptom severity are more likely to drop out of using assessment tools such as ESM or have lower levels of engagement than those without a mental health diagnosis or individuals with less severe symptoms [8,9]. Although some findings suggest that ESM engagement may be impacted by particular demographic or clinical severity factors, these findings have not been robustly replicated and warrant further investigation.

There is a growing interest in evaluating the impact of psychological interventions on patients' everyday life. One way in which this has been done is through the addition of an ESM component to the randomized controlled trial (RCT) design of research trials. Research focused on the area of psychosis or schizophrenia spectrum disorders has been leading in this field of incorporating ESM into research designs [10-13]. However, studies report low completion rates and questions have been raised about how to best integrate this data collection method in a way that maximizes valid and reliable outcome assessment in daily life, while minimizing burden. This study is the first that addresses this dilemma directly through in-depth evaluation of this experience, conducting qualitative interviews with trial participants who have been invited to take part in an embedded ESM study.

Compared with research investigating factors that impact engagement, there is still relatively little research examining individuals' experiences of using ESM qualitatively. There are also relatively few studies that have aimed to identify reasons why individuals with psychosis may decide not to use mobile apps, of which ESM is often delivered, within the context of research or therapeutic interventions. Existing research suggests that concerns around the privacy of data collected could impact individuals' willingness to engage with assessment methods such as ESM, as well as increased feelings of paranoia and rumination when being asked to complete repeated questionnaires [14,15]. Gathering the experiences of people using ESM may therefore be particularly useful in the development or improvement of technology to be used in research and treatment of psychosis. This study firstly aimed to examine factors that affect engagement with ESM, by examining quantitative associations of completion rates with age, ethnicity, gender, and clinical severity, and secondly explored qualitatively service users' experiences of using this data collection method.

Methods

Procedure

The original purpose of the ESM study was to assess voice-hearing experiences in everyday life as secondary outcomes in the AVATAR2 trial [16]. Demographics and the Psychotic Symptom Rating Scales (PSYRATS) scores were obtained through assessments with research assistants in the baseline assessments during the AVATAR2 study [17]. These demographics and scores were then recorded on a database which could be accessed by those completing the current ESM study. The baseline assessment data used within this study included data from 207 participants recruited between January 2021 and April 2022.

Purposive sampling was used to select participants for qualitative interviews. Participants were selected in 3 groups based on terciles of ESM completion and a fourth, ESM opt-out, group. This was done in order to generate a range of experiences of using ESM. The groups were participants with low (18% and below), medium (19% - 54%), and high (55% and above) completion rates, as well as participants who chose not to use ESM.

Ethical Considerations

This study was linked to the AVATAR2 trial (ISRCTN55682735), which is a multicenter UK-based RCT. The AVATAR2 trial is investigating whether the development of dialogue with a visual representation of a voice can help reduce distress, frequency, and severity of voices in those who experience auditory hallucinations [3]. Additional research ethics approval for this study was granted in January 2022 by the London-Camberwell St Giles Research Ethics Committee (reference 20/LO/0657) through submission of a substantial amendment to the original ethical approval. A protocol for this study, along with all study documents, has been published on the Open Science Framework [18].

AVATAR2 participants who had consented to be approached for further research were contacted via telephone and asked whether they would be interested in participating in a qualitative interview. All of the participants had already completed their participation in the AVATAR2 trial prior to this study. Interviews were held either online via Microsoft Teams or in person, and all participants gave informed consent and were able to opt out at any point within the study. Participants were paid £20 (US $26.16) for their participation in the interview. All interviews were transcribed, anonymized, and checked for accuracy prior to analysis. All participants provided with informed consent, including that anonymized quotes may be used within a publication of the study findings.

Measures

Demographics

Demographic variables were information regarding participants' age, gender, and ethnicity.

Clinical Severity

Clinical severity was determined by overall baseline assessment scores on the auditory hallucinations' subscale of the Psychotic Symptom Rating Scales—Auditory Hallucinations Subscale (PSYRATS-AH) [19]. The PSYRATS-AH is made up of 11 items and scores can range from 0 to 40, with higher scores indicating higher clinical severity.

ESM Completion

Participants who opted in for the ESM assessment were sent questionnaires via the smartphone app M-Path 10 times a day for 6 days [20]. The questionnaire included items assessing affect, voice-hearing experiences, social interactions, and activity. The completion rate was defined as the percentage of questionnaires that were 100% complete, and this decision was made as saving incomplete questionnaires through the M-Path app had implications for the reliability of the data export.

XSL•FO
RenderX

The ESM data were collected at 3 timepoints—baseline (prior to randomization), 16 weeks, and 28 weeks—and at these latter 2 timepoints, the ESM period took place after the trial assessments. The same participants were asked to complete these assessments at each time point. These assessments were collected at 3 time points in line while completing other trial-related data. Following the ESM assessment period, participants were asked to complete a debrief questionnaire. The debrief questionnaire was administered via phone call by a research assistant to collect structured feedback on the experience of the ESM week and took approximately 5 minutes to complete. This questionnaire was developed during a previous ESM study conducted with people with a diagnosis of schizophrenia and is not coproduced [1]. The questionnaire asks about people's experience of disruption during the week due to the ESM questionnaires, how easy they found it to complete, adequacy of support provided, and whether they enjoyed the experience, and there are also open text boxes for more detailed feedback. A copy of this questionnaire can be found in Multimedia Appendix 1.

Patient and Public Involvement and Development of Interviews

Qualitative interviews were administered using topic guides that were developed for the study. The interview topic guides were developed in collaboration with the trial's patient and public involvement (PPI) consultants (people with lived experience of psychosis who were involved in a range of activities supporting the trial). Preliminary drafts of the topic guides were developed by the first author. These guides were then reviewed by research assistants and PPI consultants. Feedback on these topic guides was given verbally through a virtual meeting with the first author (SD), as well as written feedback provided by email. Such feedback enabled changes to the wording of the topic guides with resulting improvement in accessibility. Final topic guides included open-ended questions designed in order to explore the following areas: participants' involvement in AVATAR2, familiarity with mobile phones, and reflections on their use of ESM. A walk-through of the M-Path app was provided to participants who opted in to using ESM in order to support with their reflections on using the app [21]. For those who opted out of ESM, a walk-through of the AVATAR2 consent form where they would have been asked about participating in the ESM part of the study was conducted.

Interviews lasted between 21 and 64 minutes. Interviews were conducted by the lead author (SD). Of the 17 interviews, 11 (64.7%) were coconducted by a PPI consultant. Three PPI consultants received 3 hours of training in conducting qualitative interviews. PPI consultants were paid in line with NIHR INVOLVE guidelines for their time [22].

Reflexivity Statement

The lead author is a White British, female clinical psychologist with prior experience of working clinically with people with psychosis. She has particular interests in understanding the experiences of people with psychosis in order to improve the quality and availability of treatments for this population. As such, the lead author's epistemological position is most closely aligned to a "critical realist" perspective [23]. This meant that a flexible deductive approach was used throughout the research process. For example, preexisting research knowledge was held in mind when defining the research question and in developing the topic guide for the semistructured interviews. This flexible approach continued throughout the data analysis process, where both information from the data and feedback from the PPI consultants informed preliminary and secondary codes, as well as themes and subsequent interpretation.

The lead author had not previously had any contact with the participants of the AVATAR2 trial. The other authors were all members of the AVATAR2 research team ranging from PPI consultants to principal investigators with varying degrees of contact with participants of the trial. All of the PPI consultants who were involved within this study were individuals with lived experience of psychosis, some of whom had previously been participants in the first AVATAR research trial [24].

Analysis Methods

An a priori power analysis of ESM engagement indicating a total sample of 108 or above would be necessary to detect associations at a medium effect size with 90% power (f^2=0.15). IBM SPSS (version 28.0; IBM Corp) package was used to analyze the associations between engagement and other variables and for descriptive statistics. A multiple regression was conducted to determine whether there were any associations between demographic variables and clinical severity with ESM completion rates.

For the qualitative investigation, all interviews were transcribed and analyzed by the first author (SD) using a reflexive thematic analysis approach [25]. All of the transcribed data were thoroughly read in order to increase familiarity with it. All data were reread in order to identify initial codes. A further read of the data was conducted to identify further, more interpretative codes. These codes were used to help generate preliminary themes. A hybrid approach was taken, deductive in that prespecified research questions guided the analysis, with an inductive, data-driven approach to how the themes that answered these questions were generated. The focus was on semantic codes, a description of explicit meanings provided by participants regarding their experience.

Preliminary codes and themes of data from half of the interview transcripts were reviewed by the PPI consultants who had conducted interviews. PPI consultants provided feedback and interpretation in relation to the initial codes and themes. This feedback related to how the themes related to their own experiences and how the wording of the themes was able to capture the different preliminary codes. This feedback informed not only how the final themes were worded but also which information within the themes would be most important to highlight. For example, PPI consultants spoke about their own mental health experiences and how these could impact their usage of ESM during our interviews, and this was foregrounded through the selection of quotes. These preliminary themes evolved with further analysis of the remaining interviews. Themes generated from all of the interview data were again

XSL•FO
RenderX

reviewed and checked by PPI consultants who coconducted the interviews.

Data sufficiency was ensured by reflecting on the content as interviews were conducted and through discussion with PPI cointerviewers and the research team agreeing that sufficient information power had been generated to address the research questions [26]. Rigor was ensured through strategies including triangulation between lead researcher, wider research team, and lived experience experts when developing themes in the analysis, and reflexivity being considered at all stages with an awareness of potential biases that may influence the findings.

Results

Factors Affecting ESM Completion

The responses to the debrief questionnaire can be found in Multimedia Appendix 2. The findings from this questionnaire show that the majority of participants reported minimal disruption or deviation from usual routines during the ESM week and were able to access the assessments on the app with the equipment, training, and support provided.

Demographic information and baseline assessment data including the PSYRATS were obtained for 207 participants. Of these participants, 134/207 (64.7%) consented to take part in the optional ESM component of the assessment. The mean ESM completion rate across the 134 participants who consented to ESM assessment was 39.1% (SD 28.5). The demographics of the whole sample (N=207) as well as the ESM (N=134) and interview sample (n=17) can be found in Table 1. Information about which RCT treatment group the interview sample was assigned to within the AVATAR2 trial is also included in Table 1. The psychiatric diagnoses of the whole sample can be found in Table 2.

Table . Demographic variables for participants within the whole, experience sampling methodology, and interview sample.

Variable	Whole sample (N=207)	ESM[a] sample (n=134)	Interview sample (n=17)
Age (years), median (range)	38.10 (18-70)	36.34 (18-66)	44.47 (22-62)
Sex, n (%)			
Male	123 (59.4)	75 (56)	10 (58.8)
Female	80 (38.6)	56 (41.8)	7 (41.2)
Other	4 (1.9)	3 (2.2)	0 (0)
Ethnicity, n (%)			
White	125 (60.4)	91 (67.9)	9 (52.9)
Black Caribbean	16 (7.7)	8 (6)	2 (11.8)
Black African	12 (5.8)	8 (6)	3 (17.7)
Black Other	4 (1.9)	3 (2.2)	0 (0)
Indian	8 (3.9)	4 (3)	0 (0)
Pakistani	8 (3.9)	5 (3.7)	0 (0)
Chinese	1 (0.5)	0 (0)	0 (0)
Other	33 (15.9)	15 (11.2)	3 (17.7)
PSYRATS[b] score, median (range)	29.94 (13-41)	29.38 (13-40)	27.82 (13-35)
RCT[c] treatment group, n (%)			
Treatment as usual	N/A[d]	N/A	7 (41.2)
Brief	N/A	N/A	5 (29.4)
Extended	N/A	N/A	5 (29.4)

[a]ESM: experience sampling methodology.

[b]PSYRATS: Psychotic Symptom Rating Scales.

[c]RCT: randomized controlled trial.

[d]N/A: not applicable.

XSL•FO
RenderX

Table . Psychiatric diagnoses of participants within the whole sample.

Diagnosis	Total count	Total %
F20 Paranoid schizophrenia	83	39.7
F22 Persistent delusional disorder	2	1.0
F23 Acute and transient psychotic disorder	3	1.4
F24 Induced delusional disorder	1	0.5
F25 Schizoaffective disorder	18	8.6
F28 Other nonorganic psychotic disorder	1	0.5
F29 Unspecified nonorganic psychosis	75	35.9
F31 Bipolar affective disorder	5	2.4
F32.3 Severe depressive episode with psychotic symptoms	19	9.1
N/A[a]	2	1.0

[a]N/A: not available.

A multiple regression was carried out to investigate whether age, gender, ethnicity, and clinical severity could significantly predict ESM completion rates within the ESM sample. The results of the multiple regression indicated that the model was not a significant predictor of rates of ESM completion ($F_{5,128}$=0.548; P=.74). Coefficients and significance values for each hypothesized predictor can be found in Table 3.

Table . Results of multiple regression for predictors of experience sampling methodology completion.

Predictor	*B*	SE	β	*t* test (*df*)	*P* value
Age (years)	0.312	0.205	.137	1.53 (10,123)	.13
Sex					
Male versus female	0.736	5.27	.013	0.140 (10,123)	.89
Male versus other sex	0.466	17.026	.002	0.027 (10,123)	.98
Ethnicity					
White versus other ethnicity	−3.124	5.395	−.051	−0.579 (10,123)	.56
Clinical severity	−0.109	0.501	−.020	−0.217 (10,123)	.83

Qualitative Exploration of Experiences of Using ESM

In total, 59 participants were contacted to ask whether they would be interested in participating in the study. Overall, 17 participants agreed to take part in the study and completed qualitative interviews. Table 4 shows the number of participants in each of the different ESM completion rate groups. The thematic analysis resulted in 3 overarching themes and 9 subthemes. The overarching themes were *Factors affecting engagement with ESM*, *Perceived benefits of ESM*, and *Suggestions for improvement*. Themes were generated using interview responses from across the 4 ESM completion groups (low, medium, high, and no ESM). Specific quotes accompanied by information on which group the participant was drawn are included. Further participant quotes related to each overarching theme and subtheme can be found in Multimedia Appendix 3.

Table . Range of experience sampling methodology completion across groups and number of participants per group.

ESM[a] group	Participants across whole sample (N=207), n (%)	Participants interviewed (n=17), n (%)	Completion rate range, %
Low	46 (22.2)	4 (23.53)	0 - 18
Medium	42 (20.3)	4 (23.53)	19 - 54
High	46 (22.2)	5 (29.41)	55+
No ESM	73 (35.3)	4 (23.53)	N/A[b]

[a]ESM: experience sampling methodology.

[b]N/A: not applicable.

XSL•FO
RenderX

Factors Affecting Engagement

This first theme relates to previous experiences, individual factors, or factors related to ESM which made it easier or harder to engage with. This theme comprises four subthemes: (1) integrating ESM into daily life, (2) impact of symptoms, (3) experience with technology, and (4) experience of therapy and research.

1. *Integrating ESM into daily life*: For a number of participants, the questions felt quick and easy to answer and became part of their routine. The format of the questions also helped make the questionnaires straightforward to answer, which could be particularly helpful for those who have dyslexia.

> *Well, it was fairly straightforward to answer the questions.* [P001, line 35, High]
>
> *Yeah, just I think it just became a routine.* [P017, lines 92-93, Low]
>
> *I have dyslexia. And so when I am reading like a book or something, it's quite overwhelming with everything that's around it. But because it focuses on the question and then it's just it just comes up where it shows you and like, it's just like a scale. It's a lot easier to use a scale than umm explain at times I am...And that's quite helpful.* [P009, lines 332 - 335, Medium]

Participants reported that the number of times in which they were prompted to answer the questionnaires each day within the sampling period could feel repetitive and tedious. For some, the knowledge that it was only for a limited period of time increased their motivation to complete the app; however, that engaging with this longer term would not be sustainable. This may have been particularly the case for those who had higher completion rates and therefore felt that doing the questionnaires as frequently per day could not happen over a longer period of time. For a participant who did not use ESM but was informed of the number of times the questionnaires are sent through in a day, they reported that the frequency of this would have been too much.

> *It was a bit a little bit repetitive.* [P017, line 77, Low]
>
> *It's quite time consuming. You know it makes quite a lot of demands in terms of how often and frequently you have to use the app.* [P001, lines 16 - 17, High]
>
> *I don't think I would enjoy 10 times a day, six days.* [P011, line 150, No ESM]

A number of participants reported that their ability to complete the app was impacted if they were engaging in other routine tasks or activities. Several participants commented on how they managed this by completing the app at a time that was convenient to them. Although this may have been good for them and made it easier to integrate into their day-to-day lives, this could have impacted their overall completion rate.

> *I still had meetings with CPN's at the time. I had other meetings that I had to. That I had to attend...I mean you can't do it as and when the notifications spring up because you have a life to live.* [P001, lines 106 - 107, High]
>
> *Well, the only trouble was it would ping when I'm driving the car it would ping. You know when you sometimes you just can't respond at that at that time, whereas at work or so, yeah, that was the only problem that I couldn't. I couldn't answer when you know exactly when it pinged.* [P008, lines 86 - 88, Medium]
>
> *Umm, I think I just checked it when I got home...I didn't do anything outside.* [P014, line 133, Low]

The notifications and questionnaires could for some feel as though they were impacting their ability to relax. This was particularly the case for notifications that occurred in the afternoon or evenings when they were more likely to be at home. Receiving notifications from the M-Path app was reported as a potential factor in ESM completion levels, as some participants reported not having received any or not hearing them due to their sound being turned off. Those who were in the high completion groups may have been thinking more consciously about the notifications due and therefore checking their phones more regularly, or perhaps were more familiar with the notification settings on their phone than those with low completion rates.

> *I feel like it's like A. Bit of a weight to carry when I'm out and about it's it's it's not. It's like a? whats that word when something is. Not a liability...A burden that's it.* [P003, lines 266 - 268, High]
>
> *I didn't get the beeps.* [P013, line 131, Low]
>
> *I think maybe once or twice I missed it, you know, but only because I'm missing notification and then I'd label come my phone and maybe like oh **** went off an hour ago, you know? And then go ahead and fill it into that point.* [P007, lines 132 - 133, High]

2. *Impact of symptoms*: Participants reported that feeling low in mood could impact their motivation and ability to complete the app. The sleeping patterns of some participants were reported to impact the answering of the questionnaires. For some, this was due to the side effects of the medication that they were prescribed which could impact how they felt, especially in the mornings. For others, the voices they hear were mostly experienced at night, which affected their sleep, and therefore it may be that time was spent in the day catching up on this sleep and therefore impacting on responding to the M-Path notifications:

> *I wasn't always in the mood I felt. Miserable. Uh. Uh, whereas other times, I didn't mind doing it...I think it's basically it's essentially mood based.* [P016, lines 128 - 129, Low]
>
> *Yeah, you know, I mean, I'm up at like, two, three, four in the morning. You know what I mean? Pacing up and down.* [P005, line 125, Medium]
>
> *I wasn't able to the morning ones because in that time I was on medication, so I was sleeping quite a lot.* [P009, lines 126 - 127, Medium]

A number of participants reported their voice-hearing experiences affected their ability to complete the app due to having difficulties concentrating or due to the voices

XSL•FO
RenderX

encouraging the participant to not respond at that time. Experiences of paranoia and beliefs that their thoughts and the questionnaires were connected could also impact some participants' ability to engage with ESM. For some, the voices may not have been present when answering the questionnaires, which for one participant was felt to be due to the voices having an awareness of the app. For others, not hearing voices when responding to the questionnaires was more challenging due to difficulties connecting with their experiences. This appeared to be particularly for people in the high responder groups, whereas for those in the lower responder or no ESM group, the presence of voices while using the app was thought of more.

> *I would there was a few times where umm I would be having mm I would be talking to <voice> and we would be focusing on my <job> and the beep would go off, but <voice> would just tell me to just like to focus on our our work.* [P009, lines 90 - 92, Medium]

> *Yeah, like they were connected. A couple of times my my brain went that way...I'm getting paranoid about <therapist> now, thinking <therapist> set me up, you know, all that stuff was like for hours, I was debating about it.* [P005, lines 210 - 214, Low]

> *But then sometimes I'd like rush through it because I'm like, I don't know if I can, if I'm really like, if any of this really connects with what's happening right now.* [P007, line 99 - 100, High]

> *Obviously it depends on how you're feeling now that particular time. Whether the voices were intense or not.* [P011, line 180, No ESM]

A number of participants reported how difficulties with their memory could impact on being able to remember aspects of the trial or their experience of using the app. Others reported that their memory also impacted their ability to remember to engage with the app and answer the questions during the data collection period.

> *But at times I don't remember. Where I would complete the the beep and then I would look at it and the other sections beforehand was done, but I don't I don't remember doing that.* [P009, lines 354 - 356, Medium]

> *I think it was just about remembering to answer the questions within a certain time period.* [P017, lines 259-260, Low]

3. *Experience with technology*: Although all of the participants interviewed reported currently owning a smartphone, the degree to which they were familiar or found using technology easy differed across the different completion groups. Specifically, the majority of participants who specifically reported to be comfortable and knowledgeable using smartphones were in the high or medium groups. Conversely, most participants in the No ESM group reported either not being very knowledgeable or confident in using technology or reported a dislike of using it. This could in part be impacted on their views of technology and specifically the link with technology and their symptoms.

> *To tell you frankly, yeah. I don't like to use technology.* [P010, line 165, No ESM]

> *I've used them for a while. I know how to use the apps and they're fairly straightforward. I'm I'm not really a technophobe so.* [P001, lines 11 - 12, High]

> *Yeah, but as I said, I don't like doing all that on the phone, and I couldn't. I'm illiterate on the Internet.* [P005, lines 75 - 76, Medium]

Several participants reported requiring support in using smartphones. This, for some, was in the context of day-to-day life in using different phone features, whereas other participants spoke of needing additional support in setting up and using the app. Several participants reported experiencing concerns regarding technology including uncertainty around who may be accessing their data; however, for a number of participants, these concerns did not impact their ability to continue using the app. However, for those who chose not to use ESM, concerns about technology appeared to be a significant factor in their decision to opt out of using ESM.

> *If I am shown it is okay. Yes I can't do it myself but if I am shown.* [P014, line 25, Low]

> *I worry about being like intercepted and. So I mean, things like that, so I have to. I would have kind of like. Umm and ahhing about it and in the end I just got on with it.* [P003, lines 32 - 34, High]

> *I wouldn't want the phone health questions falling into someone else's hands.* [P011, line 95, No ESM]

> *I don't really trust technology though, because I wonder if who's looking at it, you know.* [P012, line 107, No ESM]

4. *Experience of therapy and/or research*: Some of the participants spoke about their feelings of happiness or gratitude for being part of something new and something that could benefit others in the future. Participants also demonstrated a willingness to engage in something whereby they could learn new ways of managing their mental health.

> *I'm happy to to to use and if if it's, if it's part sort of future treatment then yeah I'm all for it.* [P017, lines 108 - 109, Low]

> *I felt really lucky to have that there's quite new. It's quite different and it's just rolling it out. I felt quite happy to do it, yeah, I felt happy about it.* [P003, lines 5 - 6, Medium]

Having signed up for the AVATAR2 trial and opting in to use ESM for some motivated their engagement and completion of the app. For some participants, this meant that completing the app was a way that they could help the trial or researcher that they had been engaging with. For others, the knowledge that they have agreed to participate and were being paid to participate was an influencing factor in ensuring that they continued to engage and complete the app.

> *It was like a game but plus <researcher> was paying me, I was like I have to do, <researcher> was paying me money to do it.* [P003, lines 92 - 93, High]

> *I felt it was something I had to do to help...So that is why I answered the questions.* [P015, line 166, High]

XSL•FO
RenderX

> *Right at the very beginning, there was a lot of information, a lot. (Yeah) You know lots of questions, lots of there's lots of that, you know? And. And yeah, I mean, beginning always really apprehensive as to what's gonna happen, aren't you? So you know, it's not. I mean, I think I've got probably a bit overwhelmed with the amount of information that was probably given to me on that day.* [P006, lines 280 - 285, No ESM]

Perceived Benefits of ESM

This second theme relates to the potential benefits that can result from using ESM. This theme comprises two subthemes: (1) access to increased support and (2) enhanced awareness.

1. *Access to increased support*: Participants reported that ESM provided a space whereby they could express their feelings and feel as though they were talking with someone about their experiences. This space was felt to be particularly helpful for those who felt worried about feeling embarrassed, not being understood by others, or where they did not want to burden others by talking about their experiences. Considering ESM is often not used as a therapeutic intervention, yet participants felt it was a space in which they could share their experiences, could be reflective of the context in which ESM was used. Being part of an RCT may enable participants to feel as if their responses were going to be seen by the research team.

> *I was able to express myself. (Yeah) Where I wasn't holding all of that and were I when I felt low, I was able to tell someone.* [P009, lines 265 - 266, Medium]

> *Umm and the app itself, because it was an app and there was no direct interface with another person there, there wasn't really any chance of it being awkward...You know, I'm I'm just an anonymised person at the end of the lane. So there there's not much scope for embarrassment about it. (Umm) So like like I found that quite easy to do and I like it.* [P001, lines 156 - 160, High]

Other benefits of ESM included the helpfulness of concentrating on answering questions as a distraction from hearing voices or dissociative experiences. As well as providing a distraction from voices, the use of ESM could itself be considered as a coping tool that could reduce the amount of distress they were experiencing and to help them feel grounded.

> *I don't think I ever had. I had that because maybe because I was concentrating. You know, with me, the voices, if I'm concentrating on doing something. Yeah. (ok) Yeah. Then then I don't really. I don't really. So I don't think to be fair, I don't think I am. Yeah, I don't think I heard them when I was answering.* [P008, lines 370 - 372, Medium]

> *And it it did help ground cause. Answering questions about. Like. If the voices stressing me or. Like really thinking about it helped.* [P009, lines 143 - 144, Medium]

2. *Enhanced awareness*: Participants reported that the app provided them with an opportunity to increase awareness of their mental health and to provide them with space where they could spend time reflecting on their experiences, which could be helpful in being more present and checking in with reality. Participants also spoke of being able to increase their awareness and understanding of links between their mood and voice-hearing experiences.

> *I got me thinking about my experience regularly and. Yeah, maybe being more self-aware about what was happening to me.* [P007, lines 278-279, High]

> *Umm did did the voices talk to me more often today did? Did what I do yesterday work today? It kind of raises those kind of questions. It is is introspection reflectiveness.* [P016, lines 87 - 88, Low]

Some of the participants reported that their increased awareness enabled them to think of activities they could engage in to improve their mood, as well as to reflect on different triggers that could have impacted their mood and voice-hearing experiences. Answering the questions related to voice-hearing experiences enabled some participants to reflect on the relationship with their voices and how this may have changed over time. For some, this provided them with a sense of feeling more in control of their experiences. This was an experience that participants across the completion groups experienced and therefore could indicate the benefits of the app regardless of times they interacted with it.

> *I suppose this app makes me kind of. (Umm) Become more aware of what I need to do so I'm really not feeling relaxed and then I realise it from this question. Then I could do something about it because sometimes I don't know how I feel, until it's it's there in black and white.* [P003, lines 161 - 163, High]

> *Sometimes it is just the right focus and acknowledging that it is just voices.* [P014, line 149, Low]

> *It just I keep saying this, don't I? It just puts you in control. Umm, because you can expect to feel a certain way. At a certain time, so it it just allows you to say, OK, I'm gonna feel rubbish this day, but it's because XYZ and yeah, it just it just puts you in control, which makes life so much easier rather than feeling, you know, a victim and and and down underdog and you know, all that kind of thing.* [P008, lines 163 - 166, Medium]

Suggestions for Improvement

This third and final theme relates to feedback given by participants on how ESM could be improved. This theme comprises three subthemes: (1) specificity and customization, (2) need for further support, and (3) access to solutions and strategies.

1. *Specificity and customization*: A number of participants reported a desire for questions to be more specific and more tailored to personal experiences. For some, the questions seemed to lack clarity and therefore made them harder to answer the questions. A number of participants reported that they would have benefitted from a space where they could write free text responses to be able to more accurately describe their experiences. For others, the opportunity to write free text was

XSL•FO
RenderX

thought to be helpful in being able to keep a record or track of what they were doing before experiencing psychosis symptoms.

> *I mean, sometimes like, you know how the experience well for me and for other people, I guess the experience is very individual of hearing voices and having these sorts of experiences. So sometimes like the generic questions were like, really hard to see how...sometimes they just didn't really relate to your experience.* [P007, lines 76 - 78, High]

> *The questions were simple...so I did it the best I could. But there wasn't anywhere that I could write a little message or a note.* [P014, lines 101 - 102, Low]

> *If there could be like a wee box or where they could write what they're doing, so they're able to like keep track of what they were doing before.* [P009, lines 336 - 337, Medium]

Some of the participants reported wanting to be able to customize the number of notifications they receive or to target them at certain times in the day when it would be more convenient and fit them in with their routine. For a participant who chose not to use ESM, and whose biggest concern was the security of the app, having the opportunity to increase the security of the app, for example, being able to input a personalized pin to access the app, was felt to increase the likelihood of their future use of similar apps. They also suggested an option to be able to delete the app from another device if their phone was to be stolen, which could be more of a concern for those who chose not to use the app and may be linked to concerns around privacy.

> *If people could choose if they wanna know once...once a day, twice a day, six times a day. It would be good to choose.* [P003, lines 175 - 176, High]

> *Maybe an app where you can delete it from a distance.* [P011, line 116, No ESM]

2. *Need for further support*: Participants across different ESM completion groups reported that being offered more support in answering questions and receiving reassurance that what they were submitting was correct would have been beneficial. Those within the lower completion groups could have felt uncertain and therefore did not answer the questions, whereas those in the higher groups may have answered regardless but would have benefited from being debriefed. Although the desire for further support with completing the app was reported amongst different participants, a reluctance to seek support was also highlighted due to concerns that this would reveal their identity and treatment condition within the trial.

> *Probably it might have been a good thing if I've been able to...I had a debriefing meeting at the end of it.* [P001, lines 83 - 84, High]

> *Umm. Well, because I kept on hearing the young man as well. Umm, it put me into a difficult position because I had to, like, overcome the main voice. (Right). And then I have this other voice. Umm, you know, like being really, really disruptive in my life, you know? (Yeah). I felt like I needed encouragement.* [P013, lines 229 - 231, Low]

> *Yeah. Because I wasn't sure. (OK). And I didn't want to let anyone know that I was getting the therapy right.* [P013, line 236, Low]

Some participants reported that it would be helpful to have an app whereby their care team can access the information in order to have a better understanding of their mood. The integration with care teams was felt to be particularly helpful in relation to the safety of the individual they are supporting in order to monitor any reduction in mood which could indicate increased risk to self. The inclusion of family and friends was also suggested by participants as an opportunity for discussion and increasing understanding of their experiences.

> *I do think it would be quite helpful to have umm the app where you. For example, like your CPN or mental health team or your doctor being able to check how (Umm) how your mood has been.* [P009, lines 320 - 322, Medium]

> *It could be interesting, even maybe to include someone who's, if if I'm not home, someone comes to visit me my <relative> you know, maybe even to ask <relative> or, you know, would you go through the questions with me? (Yeah). Involve someone.* [P017, lines 277 - 279, Low]

3. *Access to solutions and strategies*: Some participants reported that it would be useful to have the option to see their past answers or be provided with a summary of daily responses. The ability to access summaries *or* past entries was reported to be useful in aiding their memory of their answers on that particular day.

> *A little note at the end or at the end of each question to say umm, just to what you are doing.* [P014, line 207, Low]

One participant highlighted that the app was able to only collect answers to questions and did not give the option for providing solutions. Due to a lack of solutions provided within the app, participants made a number of suggestions on how the app could be improved to provide an increased opportunity to self-manage and have easy access to different coping strategies. These suggestions tended to be made by those in the high completion group, which could be indicative of more interaction with the app and therefore increased insight into ways in which it can be improved.

> *The thing that this is just looking for answers. It's not giving solutions. It's just looking for answers.* [P003, lines 112-113, High]

> *A basic care plan that people make when they're well and they kind of put down how they would like to be treated when they're unwell for example, what's worked in the past like double up on antipsychotics just for two weeks or go and have a massage or watch a Disney film or have a hot bubble bath. Visit someone go to go to mosque or church. Whatever you know like so that the person is been involved in...in the solution in the in the recoveries and stuff that they know that works.* [P003, lines 128 - 131, High]

XSL•FO
RenderX

Discussion

Principal Findings

This study aimed to identify factors that could affect engagement with ESM and to explore participants' experiences of using ESM within the context of the AVATAR2 trial. ESM completion rates were not found to be associated with demographic and clinical severity factors. These results are encouraging and suggest that ESM can be engaging across a diverse user group. Participants were able to report a number of factors that could either facilitate or be a barrier to using ESM. Participants also reported a number of suggestions in order to enhance their experiences of using the app. Due to the nature of ESM when used as an assessment tool which requires the data gathered to be reliable and valid, certain suggestions mentioned by participants, including the customization of notifications, may be problematic and result in bias. Although these suggestions may not be applicable for ESM when used as an assessment tool, the findings could be useful in the development and implementation of other forms of digital tools. For example, the use of ecological momentary intervention has been investigated and used within people with psychosis to support with symptom monitoring and self-management within clinical settings [2,12,27].

All participants who were interviewed in the current study reported owning a smartphone, which is consistent with previous findings related to increased smartphone usage within psychosis populations [28,29]. Despite this consistent ownership, the degree of knowledge and confidence in using these devices differed across the sample. Therefore, owning a smartphone does not necessarily equate to having the knowledge and skills in using them, which has been identified in previous research [30,31]. It is therefore important that individuals are also offered opportunities for training to increase their knowledge and skills of using technology. This can help to make it easier for them to use and to reduce barriers in accessing digital mental health support. Previous research has highlighted the importance of effective briefing and communication while engaging with ESM and the positive impact on completion rates when receiving such support [32]. This increased support may be particularly useful for people currently experiencing positive symptoms, which were reported by some participants in this study as a factor affecting their ability to respond to the questionnaires.

Concerns about privacy and security were raised within the current study, similar to findings in previous research [33]. These concerns were notably expressed by the majority of participants who chose not to use ESM. Having more information about the app and additional security may be helpful for some, but for others, it may not be sufficient to increase their willingness to use ESM due to the extent of their concerns about technology. Therefore, in some instances, the use of digital data collection methods such as ESM is not going to be acceptable for all, and it is important that other methods are also used for those who do not wish to use technology. This also applies to other digital tools and interventions whereby people with psychosis may still like to access support but would like to do so without having to rely on technology. It is worth highlighting that concerns about, or trust in, technology are not mutually exclusive with symptoms such as psychosis and are a consideration in the wider context of digital health care [34].

Having a mobile phone app that was easy and quick to use was reported as a benefit of ESM and helped to integrate it into everyday life. However, a number of factors were reported to impact on the ability to integrate ESM into their day-to-day lives, including the frequency of questionnaires and notifications over the course of the day. This is in line with research reporting that completion rates can decrease when there are more items to answer [6,32,35,36]. The number of questionnaires and of questions within them is worth reflecting on for those implementing ESM.

Increasing awareness and understanding of mental health symptoms was one of the main benefits of ESM reported by participants in this study. This included learning links between mood and voice-hearing experiences, as well as triggers of such experiences. Research has found that increasing insight into mental health difficulties can be important for increased feelings of empowerment, help-seeking behaviors, and recovery [37,38]. This suggests that even as an assessment tool, ESM could be used as a way for people with psychosis to learn and understand more about their mental health experiences, which can have positive impacts on individual mental health outcomes and overall recovery.

Participants made a number of suggestions to improve their experience, including increasing customization, being able to see past responses, and access to strategies and solutions within an app. These customizations may compromise assessment with ESM, as they could influence participants' responses; however, such suggestions are important to consider in the development of digital interventions. For example, the availability of previous entries may be beneficial in order to aid with recall in populations of people with psychosis where cognitive difficulties such as memory can be affected [39]. Including coping strategies or solutions could also be important for digital tools and interventions in supporting people to increase skills in self-management. Increased ability to self-manage has positive implications for people with psychosis in terms of increasing feelings of control, empowerment, and hope for recovery [40,41]. The incorporation of messages to enhance self-management has recently been reported within an RCT exploring the EMPOWER app, which showed promising results [42].

Strengths and Limitations

It is important to consider the findings of this study in light of various strengths and limitations. A major strength of this study was the involvement of PPI consultants at many stages of the research in line with INVOLVE guidelines [43]. PPI involvement in research has been found to be beneficial in enhancing the quality and applicability of research, ensuring that it is user-focused, as well as helping to reduce stigma and increasing opportunities to connect with others. Despite this, PPI involvement in research is still inconsistent [44,45]. Although this is a strength of the study, it is worth noting that, for a variety of operational reasons, PPI consultants did not coconduct all 17 of the interviews, with a potential risk of bias,

XSL•FO
RenderX

given some inconsistency in the interview process. Furthermore, although PPI consultants were involved at a number of stages of the research, they were not actively involved in the early stages of data analysis. Other research has reported on how PPI researchers may interpret data differently compared with academic researchers and therefore actively involving PPI consultants in the analysis of data could help to reduce bias in interpreting [46]. It is therefore important for PPI consultants to be involved in this stage in order to create meaning from data from multiple perspectives. PPI consultants were also not directly involved with the recruitment of participants. A previous review has found that involving PPI consultants with the recruitment process can increase enrollment numbers of participants [47], which may have been important in this study considering that only 17 out of the 59 participants approached were recruited for interview. It is worth noting the potential limitation related to the current sample used within the study. All of the participants had experiences of auditory hallucinations, and symptom severity was assessed based on outcomes of the PSYRATS-AH; therefore, the transferability of the findings to people who may experience other symptoms related to psychosis could be limited and would benefit from further research. Finally, this study did not examine ESM design considerations; however, previous reviews have found these to be important in enhancing ESM completion. One design consideration, in particular, mentioned in this study relates to the availability of support during ESM use, with researcher-initiated contact having been previously found to increase completion rates [48]. Further design considerations could include the number of times that the questionnaires are administered throughout the day, which could differ across studies.

Conclusions

This study has provided important insights into the experiences of using ESM and factors that could affect engagement with this method among people with psychosis. It has also highlighted areas to consider when developing and implementing ESM and digital interventions with this population. Furthermore, the range of benefits for multiple stakeholders of actively involving PPI consultants in research has been discussed and the importance highlighted of PPI consultants' involvement at all stages within the research process.

Acknowledgments

Dr. Clementine Edwards acknowledges support from a Wellcome Early Career Award (227646/Z/23/Z). Generative artificial intelligence was not used in any part of the manuscript writing.

Funding

This work was supported by a Wellcome grant (215471/Z/19/Z). Philippa Garety is part funded by the National Institute for Health and Care Research Maudsley Biomedical Research Centre Center at South London and Maudsley NHS Foundation Trust and King's College London, grant (NIHR203318).

Data Availability

The datasets and transcripts that were analyzed within this study are not publicly available due to restrictions on confidentiality; however, these can be requested from the corresponding author on reasonable request.

Disclaimer

The views expressed are those of the author(s) and not necessarily those of the National Institute for Health and Care Research or the Department of Health and Social Care.

Authors' Contributions

SD analyzed the data and drafted the paper as the first author, with support from PG, AG, and CE. All coauthors had the opportunity to contribute to drafts of the paper, to critically review the manuscript, and to approve the final version. SD, with support from PG, AG, and CE devised the study aims and methods and supervised the researchers during data collection and analysis. OO, LM, and SA provided methodological advice and supported data access. AD, FY, CB, HF, and AG contributed to the development of study material, data collection, and analysis.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Experience sampling methodology debrief questionnaire.
[DOCX File, 50 KB - formative_v9i1e78204_app1.docx]

Multimedia Appendix 2
Responses to experience sampling methodology debrief questionnaire.

XSL•FO
RenderX

[DOCX File, 26 KB - formative_v9i1e78204_app2.docx]

Multimedia Appendix 3
Study themes and participant quotes.
[DOCX File, 24 KB - formative_v9i1e78204_app3.docx]

References

1. Edwards CJ, Cella M, Tarrier N, Wykes T. The optimisation of experience sampling protocols in people with schizophrenia. Psychiatry Res 2016 Oct 30;244:289-293. [doi: 10.1016/j.psychres.2016.07.048] [Medline: 27512917]
2. Bell IH, Lim MH, Rossell SL, Thomas N. Ecological momentary assessment and intervention in the treatment of psychotic disorders: a systematic review. Psychiatr Serv 2017 Nov 1;68(11):1172-1181. [doi: 10.1176/appi.ps.201600523] [Medline: 28669284]
3. Garety P, Edwards CJ, Ward T, et al. Optimising AVATAR therapy for people who hear distressing voices: study protocol for the AVATAR2 multi-centre randomised controlled trial. Trials 2021;22(1):366. [doi: 10.1186/s13063-021-05301-w]
4. Gold JM, Barch DM, Feuerstahler LM, et al. Working memory impairment across psychotic disorders. Schizophr Bull 2019 Jun 18;45(4):804-812. [doi: 10.1093/schbul/sby134] [Medline: 30260448]
5. Vachon H, Viechtbauer W, Rintala A, Myin-Germeys I. Compliance and retention with the experience sampling method over the continuum of severe mental disorders: meta-analysis and recommendations. J Med Internet Res 2019 Dec 6;21(12):e14475. [doi: 10.2196/14475] [Medline: 31808748]
6. Martin LR, Williams SL, Haskard KB, Dimatteo MR. The challenge of patient adherence. Ther Clin Risk Manag 2005 Sep;1(3):189-199. [Medline: 18360559]
7. van der Krieke L, Emerencia AC, Boonstra N, Wunderink L, de Jonge P, Sytema S. A web-based tool to support shared decision making for people with a psychotic disorder: randomized controlled trial and process evaluation. J Med Internet Res 2013 Oct 7;15(10):e216. [doi: 10.2196/jmir.2851] [Medline: 24100091]
8. Granholm E, Loh C, Swendsen J. Feasibility and validity of computerized ecological momentary assessment in schizophrenia. Schizophr Bull 2008 May;34(3):507-514. [doi: 10.1093/schbul/sbm113] [Medline: 17932087]
9. Thewissen V, Bentall RP, Lecomte T, van Os J, Myin-Germeys I. Fluctuations in self-esteem and paranoia in the context of daily life. J Abnorm Psychol 2008 Feb;117(1):143-153. [doi: 10.1037/0021-843X.117.1.143] [Medline: 18266492]
10. Burger SR, Hardy A, Verdaasdonk I, et al. The effect of trauma-focused therapy on voice-hearing: an experience sampling study. Psychol Psychother 2025 Mar;98(1):25-39. [doi: 10.1111/papt.12556] [Medline: 39494655]
11. Pot-Kolder RMCA, Geraets CNW, Veling W, et al. Virtual-reality-based cognitive behavioural therapy versus waiting list control for paranoid ideation and social avoidance in patients with psychotic disorders: a single-blind randomised controlled trial. Lancet Psychiatry 2018 Mar;5(3):217-226. [doi: 10.1016/S2215-0366(18)30053-1]
12. Myin-Germeys I, van Aubel E, Vaessen T, et al. Efficacy of acceptance and commitment therapy in daily life in early psychosis: results from the multi-center INTERACT randomized controlled trial. Psychother Psychosom 2022 Nov 29;91(6):411-423. [doi: 10.1159/000522274]
13. Peters E, Hardy A, Dudley R, et al. Multisite randomised controlled trial of trauma-focused cognitive behaviour therapy for psychosis to reduce post-traumatic stress symptoms in people with co-morbid post-traumatic stress disorder and psychosis, compared to treatment as usual: study protocol for the STAR (Study of Trauma And Recovery) trial. Trials 2022 Dec;23(1):429. [doi: 10.1186/s13063-022-06215-x]
14. Allan S, Bradstreet S, Mcleod H, et al. Developing a hypothetical implementation framework of expectations for monitoring early signs of psychosis relapse using a mobile app: qualitative study. J Med Internet Res 2019;21(10):e14366. [doi: 10.2196/14366]
15. Ainsworth J, Palmier-Claus JE, Machin M, et al. A comparison of two delivery modalities of a mobile phone-based assessment for serious mental illness: native smartphone application vs text-messaging only implementations. J Med Internet Res 2013;15(4):e60. [doi: 10.2196/jmir.2328]
16. Edwards CJ, Smith RJF, Lafit G, et al. The impact of AVATAR therapy on voice-hearing in everyday life: ESM outcomes of the AVATAR2 trial. Schizophr Bull (forthcoming) .
17. Garety PA, Edwards CJ, Jafari H, et al. Digital AVATAR therapy for distressing voices in psychosis: the phase 2/3 AVATAR2 trial. Nat Med 2024 Dec;30(12):3658-3668. [doi: 10.1038/s41591-024-03252-8] [Medline: 39468363]
18. AVATAR2 qualitative programme, AVATAR2 research programme: workpackage 4—engagement with experience sampling methods. OSF. 2022. URL: https://osf.io/7xsjq/ [accessed 2025-11-16]
19. Haddock G, McCarron J, Tarrier N, Faragher EB. Scales to measure dimensions of hallucinations and delusions: the Psychotic Symptom Rating Scales (PSYRATS). Psychol Med 1999 Jul;29(4):879-889. [doi: 10.1017/s0033291799008661] [Medline: 10473315]
20. m-Path. 2025. URL: https://m-path.io/landing [accessed 2025-05-28]
21. Light B, Burgess J, Duguay S. The walkthrough method: an approach to the study of apps. New Media & Society 2018 Mar;20(3):881-900. [doi: 10.1177/1461444816675438]

XSL•FO
RenderX

22. INVOLVE: a practical guide to being inclusive in public involvement in health research. National Institute for Health Research (NIHR). 2020. URL: https://tinyurl.com/yc2jkfcx [accessed 2025-05-28]
23. Fletcher AJ. Applying critical realism in qualitative research: methodology meets method. Int J Soc Res Methodol 2017 Mar 4;20(2):181-194. [doi: 10.1080/13645579.2016.1144401]
24. Craig TK, Rus-Calafell M, Ward T, et al. AVATAR therapy for auditory verbal hallucinations in people with psychosis: a single-blind, randomised controlled trial. Lancet Psychiatry 2018 Jan;5(1):31-40. [doi: 10.1016/S2215-0366(17)30427-3] [Medline: 29175276]
25. Braun V, Clarke V. Reflecting on reflexive thematic analysis. Qualitative Research in Sport, Exercise and Health 2019 Aug 8;11(4):589-597. [doi: 10.1080/2159676X.2019.1628806]
26. Braun V, Clarke V. To saturate or not to saturate? Questioning data saturation as a useful concept for thematic analysis and sample-size rationales. Qualitative Research in Sport, Exercise and Health 2021 Mar 4;13(2):201-216. [doi: 10.1080/2159676X.2019.1704846]
27. Vaessen T, Steinhart H, Batink T, et al. ACT in daily life in early psychosis: an ecological momentary intervention approach. Psychosis 2019 Apr 3;11(2):93-104. [doi: 10.1080/17522439.2019.1578401]
28. Firth J, Cotter J, Torous J, Bucci S, Firth JA, Yung AR. Mobile phone ownership and endorsement of “mhealth” among people with psychosis: a meta-analysis of cross-sectional studies. Schizophr Bull 2016 Mar;42(2):448-455. [doi: 10.1093/schbul/sbv132] [Medline: 26400871]
29. Young AS, Cohen AN, Niv N, et al. Mobile phone and smartphone use by people with serious mental illness. Psychiatr Serv 2020 Mar 1;71(3):280-283. [doi: 10.1176/appi.ps.201900203] [Medline: 31744429]
30. Spanakis P, Wadman R, Walker L, et al. Measuring the digital divide among people with severe mental ill health using the essential digital skills framework. Perspect Public Health 2024 Jan;144(1):21-30. [doi: 10.1177/17579139221106399] [Medline: 35929589]
31. Eisner E, Faulkner S, Allan S, et al. Barriers and facilitators of user engagement with digital mental health interventions for people with psychosis or bipolar disorder: systematic review and best-fit framework synthesis. JMIR Ment Health 2025 Jan 20;12(1):e65246. [doi: 10.2196/65246] [Medline: 39832352]
32. Palmier-Claus JE, Myin-Germeys I, Barkus E, et al. Experience sampling research in individuals with mental illness: reflections and guidance. Acta Psychiatr Scand 2011 Jan;123(1):12-20. [doi: 10.1111/j.1600-0447.2010.01596.x] [Medline: 20712828]
33. Borghouts J, Eikey E, Mark G, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. J Med Internet Res 2021 Mar 24;23(3):e24387. [doi: 10.2196/24387] [Medline: 33759801]
34. Adjekum A, Blasimme A, Vayena E. Elements of trust in digital health systems: scoping review. J Med Internet Res 2018 Dec 13;20(12):e11254. [doi: 10.2196/11254] [Medline: 30545807]
35. Morren M, van Dulmen S, Ouwerkerk J, Bensing J. Compliance with momentary pain measurement using electronic diaries: a systematic review. Eur J Pain 2009 Apr;13(4):354-365. [doi: 10.1016/j.ejpain.2008.05.010] [Medline: 18603458]
36. Del Piccolo MC, Hammoud R, Khan M, et al. Smartphone-based self-monitoring in first episode psychosis: mixed-methods study of barriers and facilitators to engagement. J Med Internet Res 2025 Aug 26;27:e71989. [doi: 10.2196/71989] [Medline: 40857702]
37. Atapattu MS, Gonzales HM, Williams N. Recovery from psychosis: an integrated model of interpersonal and intrapersonal factors from the perspective of psychologists. J Psychosoc Rehabil Ment Health 2022 Dec;9(4):379-393. [doi: 10.1007/s40737-022-00274-8]
38. Drake RJ, Husain N, Marshall M, et al. Effect of delaying treatment of first-episode psychosis on symptoms and social outcomes: a longitudinal analysis and modelling study. Lancet Psychiatry 2020 Jul;7(7):602-610. [doi: 10.1016/S2215-0366(20)30147-4] [Medline: 32563307]
39. Fulford D, Schupbach E, Gard DE, Mueser KT, Mow J, Leung L. Do cognitive impairments limit treatment gains in a standalone digital intervention for psychosis? A test of the digital divide. Schizophr Res Cogn 2022 Jun;28:100244. [doi: 10.1016/j.scog.2022.100244] [Medline: 35242612]
40. Leamy M, Bird V, Le Boutillier C, Williams J, Slade M. Conceptual framework for personal recovery in mental health: systematic review and narrative synthesis. Br J Psychiatry 2011 Dec;199(6):445-452. [doi: 10.1192/bjp.bp.110.083733] [Medline: 22130746]
41. Lean M, Fornells-Ambrojo M, Milton A, et al. Self-management interventions for people with severe mental illness: systematic review and meta-analysis. Br J Psychiatry 2019 May;214(5):260-268. [doi: 10.1192/bjp.2019.54] [Medline: 30898177]
42. Gumley AI, Bradstreet S, Ainsworth J, et al. The EMPOWER blended digital intervention for relapse prevention in schizophrenia: a feasibility cluster randomised controlled trial in Scotland and Australia. Lancet Psychiatry 2022 Jun;9(6):477-486. [doi: 10.1016/S2215-0366(22)00103-1] [Medline: 35569503]
43. UL Standards for Public Involvement [internet]. NIHR. 2015. URL: https://sites.google.com/nihr.ac.uk/pi-standards/home [accessed 2025-05-28]

XSL•FO
RenderX

44. Capobianco L, Faija C, Cooper B, et al. A framework for implementing patient and public involvement in mental health research: the PATHWAY research programme benchmarked against NIHR standards. Health Expect 2023 Apr;26(2):640-650. [doi: 10.1111/hex.13676] [Medline: 36625226]
45. Susanti H, James K, Utomo B, et al. Exploring the potential use of patient and public involvement to strengthen Indonesian mental health care for people with psychosis: a qualitative exploration of the views of service users and carers. Health Expect 2020 Apr;23(2):377-387. [doi: 10.1111/hex.13007] [Medline: 31782266]
46. Green CEL, Freeman D, Kuipers E, et al. Measuring ideas of persecution and social reference: the Green et al. Paranoid Thought Scales (GPTS). Psychol Med 2008 Jan;38(1):101-111. [doi: 10.1017/S0033291707001638] [Medline: 17903336]
47. Crocker JC, Ricci-Cabello I, Parker A, et al. Impact of patient and public involvement on enrolment and retention in clinical trials: systematic review and meta-analysis. BMJ 2018 Nov 28;363:k4738. [doi: 10.1136/bmj.k4738] [Medline: 30487232]
48. Deakin E, Ng F, Young E, et al. Design decisions and data completeness for experience sampling methods used in psychosis: systematic review. BMC Psychiatry 2022 Oct 28;22(1):669. [doi: 10.1186/s12888-022-04319-x] [Medline: 36307752]

Abbreviations

EMI: ecological momentary intervention
ESM: experience sampling methodology
NIHR: National Institute for Health and Care Research
PPI: patient and public involvement
PSYRATS: Psychotic Symptom Rating Scales
PSYRATS-AH: Psychotic Symptom Rating Scales—Auditory Hallucinations Subscale
RCT: randomized controlled trial

Edited by N Cahill, TDA Cardoso; submitted 28.05.25; peer-reviewed by A Branitsky, AK Chun Chau; revised version received 11.10.25; accepted 28.10.25; published 12.12.25.

Please cite as:
Dennard S, Garety P, Edwards C, Gumley A, Owrid O, Miller L, Allan S, Duerden A, Yanga F, Burns C, Fletcher H, Grant A Exploration of Factors That Affect Engagement With the Experience Sampling Method and Service Users' Experience of This Within the AVATAR2 Trial: Mixed Methods Study
JMIR Form Res 2025;9:e78204
URL: https://formative.jmir.org/2025/1/e78204
doi:10.2196/78204

© Sophie Dennard, Philippa Garety, Clementine Edwards, Andrew Gumley, Oliver Owrid, Lucy Miller, Stephanie Allan, Alison Duerden, Francis Yanga, Chloe Burns, Helena Fletcher, Amy Grant. Originally published in JMIR Formative Research (https://formative.jmir.org), 12.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Exploring the Usability and Acceptability of the FoodMATS-Youth App for Monitoring Food Marketing Exposures: Mixed Methods Study

Idris Opeyemi Bamigbayan[1], MSc; Laurie K Twells[1], MSc, PhD; Kirby L J Shannahan[2], MA, PhD; Rachel Prowse[1], PhD

[1]Division of Population Health and Applied Health Sciences, Faculty of Medicine, Memorial University of Newfoundland, 39 Clinch Crescent A1B 3V6, St. John's, NL, Canada

[2]Faculty of Business Administration, Memorial University of Newfoundland, St. John's, NL, Canada

Corresponding Author:
Idris Opeyemi Bamigbayan, MSc
Division of Population Health and Applied Health Sciences, Faculty of Medicine, Memorial University of Newfoundland, 39 Clinch Crescent A1B 3V6, St. John's, NL, Canada

Abstract

Background: Unhealthy food and beverage marketing influences children's attitudes, preferences, and behaviors toward food. Most research studies on children's exposure to food marketing focus on single settings, media, or marketing channels, precluding cumulative estimates of food marketing exposure across children's daily lives. Therefore, there is a need for tools to measure food marketing across settings.

Objective: This study aimed to test the feasibility of a mobile app to assess food marketing observed by youth aged 13-17 years across settings in their daily life in Newfoundland and Labrador.

Methods: Using a digital app, FoodMATS-Youth (Food and Beverage Marketing Assessment Tool for Settings and Youth; developed by MetricWire Inc), 23 participants photographed food marketing they observed over 3 days. Each participant completed a feedback survey on the usability and acceptability of the digital app assessed through a 5-score rating of feasibility outcomes. They also took part in focus groups, sharing their experiences with the app, and these data were thematically analyzed. Descriptive analyses of app-derived feasibility metrics were also conducted.

Results: The app had high usability and acceptability based on the feasibility outcomes, app-derived feasibility metrics, and focus group responses. For feasibility outcomes, app navigation had the highest rating at 4.7, similar to ease of use and app responsiveness at 4.48; convenience received the lowest rating at 4.0. App-derived feasibility metrics, such as user compliance, response, and app completion rates were also high at 92%, 85.2%, and 92%, respectively. A total of 146 photos of food marketing were submitted by participants through the app. Focus groups showed great participant satisfaction with the app's interface and functionality.

Conclusions: This study found that the FoodMATS-Youth mobile app is highly feasible for monitoring food marketing exposures across multiple settings (eg, social media and grocery stores) and was well received by our participants. The FoodMATS-Youth has the potential to efficiently improve food marketing research in Canada and internationally and generate data that can inform comprehensive food marketing policies.

(JMIR Form Res 2025;9:e79306) doi:10.2196/79306

KEYWORDS

food; marketing; mobile apps; feasibility studies; data collection

Introduction

Unhealthy food and beverage marketing influences children's food preferences, attitudes, and consumption [1]. Research shows that children are exposed to marketing through many channels and settings, including websites, apps, social media, school, stores, restaurants, video games, and recreational centers [2]. There have been calls by the World Health Organization to prioritize protecting children from unhealthy food marketing in all settings [3]. Despite the evidence on the impacts of food marketing, such as its influence on the excessive consumption of unhealthy food products, which leads to unhealthy weight gain and diet-related chronic diseases [4], comprehensive monitoring is still needed to inform effective policies restricting food marketing to protect children [5]. Monitoring approaches tend to focus on single settings or marketing channels, with a greater focus on television advertising [6]. These methods are likely to underestimate the level of exposure [6]. With the

XSL•FO
RenderX

growth of digital marketing, there is also an increasing interest in assessing food marketing in online media [6,7]. However, current methods and approaches are not adequately designed to capture children's actual exposure to online food marketing, often measuring only partial or potential exposures [7]. Methods have not been developed to reflect the ubiquitous nature of food marketing or to monitor the transfer of food marketing exposures between settings or media in response to restrictive policies. Methods to comprehensively monitor food marketing exposures experienced by children are needed to fill notable research gaps and adapt to the changing food marketing landscape [8].

Crowdsourcing data from children and parents on their food marketing exposures is a relevant strategy for gathering data for monitoring efforts and can be done using custom mobile apps or websites [6,9,10]. Despite the potential that mobile apps have to be used for crowdsourcing relevant food marketing data [9], only 1 mobile app, GrabFM (Grab Food Marketing), has been used to collect data on food marketing in 2 studies [9,10]. These studies only explored teen-targeted food marketing, potentially underestimating the presence and power of marketing instances that participants are exposed to or influenced by but do not perceive as teen-targeted. Considering this gap, our study aimed to test the use of the FoodMATS-Youth (Food and Beverage Marketing Assessment Tool for Settings and Youth) app to capture all marketing types that youth encounter, not only what they believe targets them. This is particularly important because evidence shows that due to numerous research gaps, overall food marketing exposure is underestimated [11]. Although mobile apps are increasingly being used in research, most studies that use them are intervention studies [12]. Mobile apps have also been successfully used as a data collection tool in diary studies where participants use their smartphones to submit photos and complete tasks, showing the versatility of apps in research [13].

Smartphone ownership has increased significantly over the years, and the growing ownership has been observed in Canada, with 84% of Canadians owning a smartphone for personal use in 2020 [14]. Statistics Canada also reported that 96% of Canadians aged 15 - 24 years used a smartphone in 2020 [15]. This results in an increase in mobile app usage. The widespread usage of smartphones makes them an ideal tool for research and health interventions, as they can be used to collect real-time data [16]. However, this could be undermined by limited access to smartphones among marginalized populations or a lack of proficiency in the use of mobile apps [17]. Teenagers use their phones daily and can thus be engaged in research that they find acceptable and in an environment they deem familiar [18]. In addition to being an efficient tool for data collection, mobile apps can also be used to engage with hard-to-reach populations [19]. This is particularly important in Newfoundland and Labrador, as it has a dispersed population, and a mobile app can enable crowdsourcing of food marketing data not only from metro areas but also from small towns as well. Youth in Newfoundland and Labrador have some of the poorest diets in Canada based on school-day dietary quality [20]. Additionally, youth in Newfoundland and Labrador had the highest rates of overweight and obesity (37.5%) among those aged 12 - 17 years in 2022, according to the Canadian Community Health Survey [21]. The development of mobile apps and other digital interventions requires feasibility testing [13], which determines if the app is appropriate for the objectives and its audience. Usability and acceptability are common outcomes often assessed in mobile app feasibility studies [22-24]. Usability refers to how much a tool can be used to achieve its goals while maintaining user satisfaction efficiently [25]. Acceptability describes how much the users of an intervention consider it to be appropriate [26]. Some measures often evaluated under usability and acceptability include ease of use, responsiveness, navigation, satisfaction, recommendations, completion, convenience, efficiency, and attractiveness [23,24,27].

This study explored the feasibility of using a mobile app to capture food marketing exposures among youth (aged 13 - 17 years). We aimed to assess usability and acceptability through feasibility outcomes, measures of mobile app usage, and the experiences of youth using the app. Based on the Technology Acceptance Model and its modifications [28], perceived ease of use and perceived usefulness are key determinants of technology adoption. As such, usability and acceptability in this study focus on applicable outcomes, such as ease of use, convenience, responsiveness, navigation, willingness to use again, and usefulness of training resources. Usability and acceptability have been found to be essential in the adoption of digital data collection tools and mobile health apps [13,29].

Methods

Research Design

This was a convergent mixed methods study that used both qualitative (QUAL) and quantitative (QUAN) tools to assess the feasibility of a mobile app [30]. This study tested the feasibility of a mobile app to assess food and beverage marketing exposures among youth using backend mobile app usage data, feedback surveys, and focus group discussions. The app usage data were used to assess app-based feasibility metrics while the feedback surveys and focus groups were used to assess and understand the usability and acceptability of the mobile app. In this study, both quantitative and qualitative data were collected and analyzed concurrently, then mixed for interpretation to comprehensively assess usability and acceptability. Concerning weighting, the quantitative and qualitative data in this study were given equal weighting, as both were used to assess usability and acceptability. Both data types were integrated when presenting the "Results" and in the "Discussion" sections for interpretation. We have combined qualitative data in exemplar quotes with quantitative results and contextualized feasibility ratings with focus group data in the discussion. Qualitative and quantitative data in this study complement each other and offer a more in-depth understanding of the feasibility outcomes beyond ratings.

Participant Sampling

Participants were required to be between the ages of 13 and 17 years, live in Newfoundland and Labrador, and have access to a mobile phone for the study duration. Participants were recruited through Community Youth Networks and targeted social media advertisements (Facebook [Meta Platforms, Inc] and Instagram [Meta Platforms, Inc]). The Community Youth

XSL•FO
RenderX

Networks are a provincial government-funded initiative providing community-based developmental programs for youth aged 12 - 18 years in NL. These networks contributed to participant recruitment by sharing the study information with their members. Facebook and Instagram advertisements were used to enroll older youth (aged 15 - 17 years), as we were unable to sufficiently recruit that age group through the Community Youth Networks. Parental consent and child assent were provided for participants aged 13 - 14 years, while 15 - 17-year-olds provided their consent.

An initial sample size of 24 participants was chosen based on similar photo-based studies in Canada [31,32]. In addition, this is a pilot study, and a sample of 24 participants was deemed sufficient to achieve the purpose of the study. We attained data saturation with 23 participants. At the end of the study, each participant was offered a $30 (CAD $1=US $0.75) electronic gift card as a thank-you for their time and effort throughout the study, regardless of whether all study activities were completed.

FoodMATS-Youth App

We collected data using our app, FoodMATS-Youth, which was created on the Catalyst by Metricwire application, a Canadian software application company with a record of facilitating digital health research. The FoodMATS-Youth had 3 main pages, namely "Home," "Resources," and "Data Collection Diaries" (Figure 1). It also had an "About Study" page, a "Consent Information" page at the beginning of the study, and a "Feedback Survey" page at the end. The home menu contained diaries for 13 settings or media (home, grocery store, convenience store, shopping mall, school, fast-food restaurants, sit-down restaurants, youth center, recreation and sports center, movie theater, public transportation, social media, and other), and each location had a cover image representative of it (Figure 1). The settings were selected based on locations that youth often visit and existing literature [33]. There was also a setting called "other" where participants could specify the location of the food marketing. The diaries contained instructions, an option to upload or take photos, and a series of questions on the marketing features of the advertisement as shown in Figure 2. These questions were informed by Health Canada's indicators for monitoring the message, design, and techniques of advertisement or marketing instance (Health Canada, unpublished data, October 2022). The mobile app included push notifications to remind participants to make diary entries when they see food marketing, and these notifications were delivered 6 random times per day, with a minimum interval of 2 hours apart and between 9 AM and 9 PM. The notifications read, "Where are you? Don't forget to take pictures of food ads you see!" The app also provides instant messaging for on-demand help from the research team if issues arise. The mobile app was developed to work on Android and iOS devices and offers real-time tracking of submissions and app usage.

Figure 1. FoodMATS-Youth (Food and Beverage Marketing Assessment Tool for Settings and Youth) mobile app showing the home, data collection diaries, and resources (left to right).

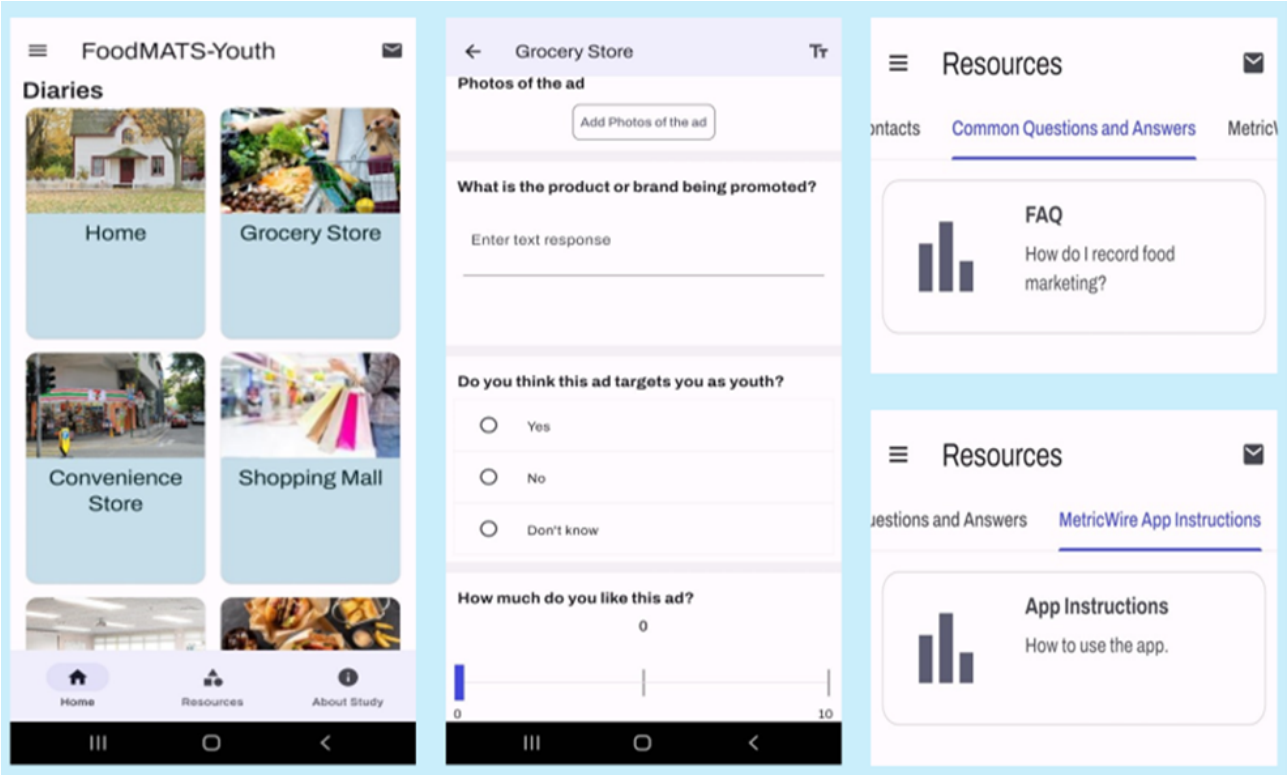

XSL•FO
RenderX

Figure 2. Questions and prompts in the FoodMATS-Youth (Food and Beverage Marketing Assessment Tool for Settings and Youth) about the marketing features of the advertisement.

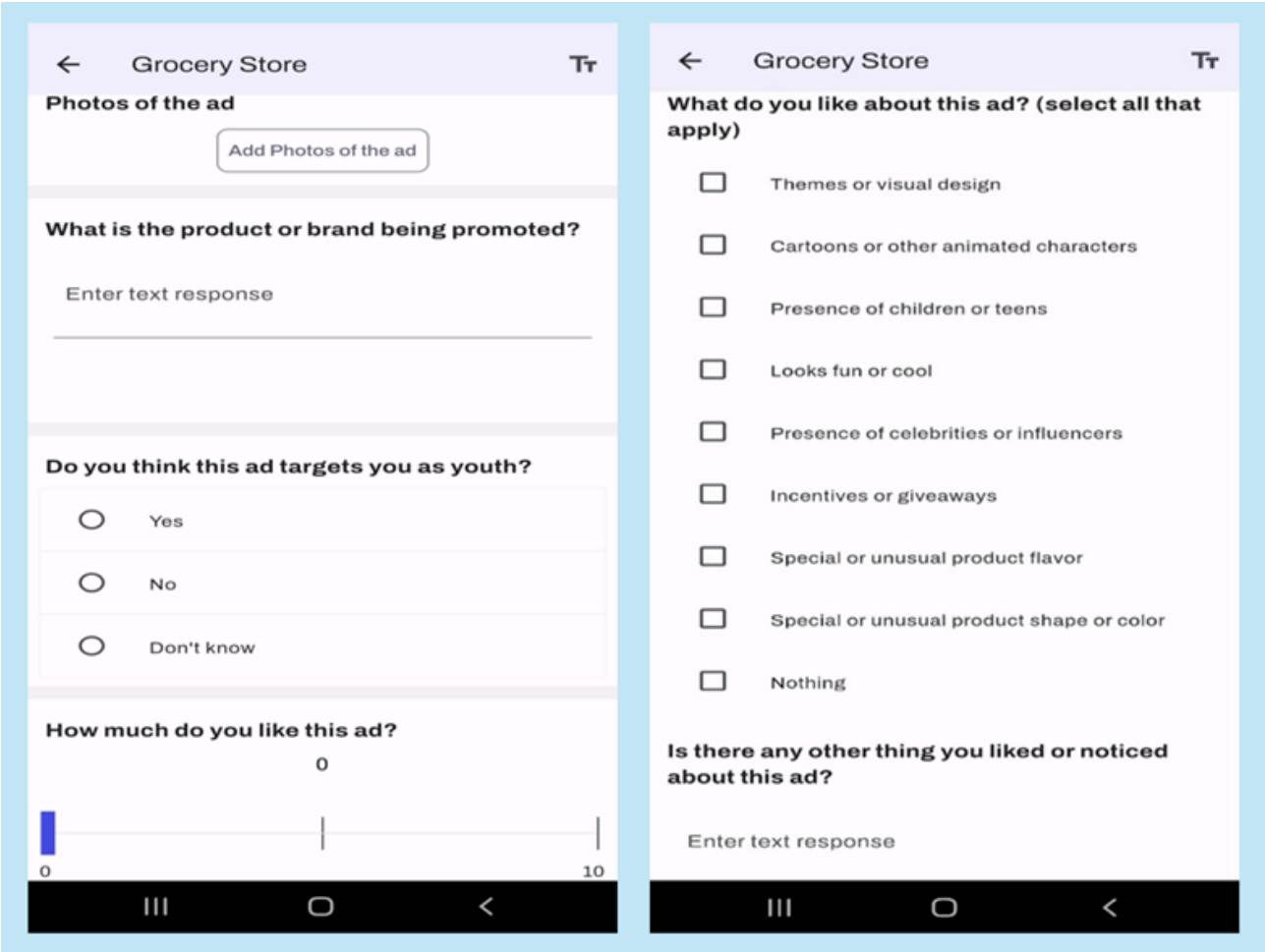

Data Collection

Participants were required to provide informed consent through the Qualtrics (Qualtrics International Inc) survey tool after reading the consent form. Consenting individuals provided demographics through a Qualtrics survey, including names, phone numbers, and emails of participants and parents, and smartphone access. Next, participants were trained on how to use the mobile app in a virtual training session via Webex by Cisco software. Multiple training sessions were held based on participants' availability and preferences, and all participants were required to attend at least one training session. Participants were trained using a mobile app user guide detailing a step-by-step process of using the app. Training resources were sent to each participant alongside the app invite link after they had completed the training. They then downloaded the Catalyst by MetricWire application via their respective App Stores or Play Stores and used the FoodMATS-Youth app to record food marketing in their lives. In the FoodMATS-Youth, they completed personalized diary entries each time they noticed food marketing over 3 days (2 weekdays and 1 weekend) by uploading photos of food marketing instances and answering a series of questions on the marketing features. Mobile app–derived outcomes, including study recruitment and retention, mobile app response and completion rate, response pattern, compliance rate, breadth of mediums, quality of photos, and app language were collected from backend mobile app usage data.

After the study period, participants completed a feedback survey on the usability and acceptability of the app, which was embedded in the FoodMATS-Youth app. This survey was used to assess the feasibility outcomes (ease of use, convenience, responsiveness, and navigation of the mobile app; the effectiveness of the training resources; and their willingness to use the mobile app again). Participants were asked to rate each of these factors with a score between 1 and 5, with a score of 5 being the best.

Finally, all participants were invited to attend focus groups organized by age (13 - 14 years; 15 - 17 years) to explore their experiences with the app. We made use of a focus group guide (Multimedia Appendix 1). The focus groups were conducted online via Webex, a videoconferencing platform, as this enables ease of participation for participants living outside the city. The focus groups were recorded on Webex and transcribed verbatim afterward. The data collection period for this study was from June 2023 to April 2024. The study timeline for each participant from recruitment to the focus group was based on their availability and preferences. Training sessions were available as the first batch of recruitment was in progress, and recruited participants began data collection a day or two after attending a training session. Participants were required to register on the

XSL•FO
RenderX

app, take photos of food marketing exposures, and complete the feedback surveys within 7 days.

Since participants were required to self-report marketing exposures, there was a risk of self-report bias and data unreliability. To address this, all participants attended a virtual training session where they were instructed on how to accurately identify and submit food marketing exposures. To help ensure that photo submissions reflected actual exposures, participants were informed that they would receive their incentives regardless of whether they completed the study. To minimize incomplete submissions, participants were informed whenever a submission did not include a photo. Additionally, to reduce missing data, participants were encouraged to upload every instance of food marketing they encountered, even if they were unsure whether it qualified as marketing.

Data Analysis

The quantitative and qualitative data from this study were analyzed using SPSS (version 29; IBM Corp) and NVivo 14 (Lumivero), respectively. Analyses compared feasibility outcomes by age group (13 - 14-year-olds; 15 - 17-year-olds). Mobile app–derived outcomes assessed included recruitment and retention rate, response rate, completion rate, response pattern, compliance rate, breadth of mediums, and quality of photos. Descriptive analyses were conducted, including measures of central tendency (mean and median) for continuous variables. Cross-tabulations were conducted using Fisher exact test for categorical variables due to small cell counts.

Specifically, the frequency of photo submissions was calculated according to settings or media and the proportion of settings or media that had submissions was analyzed according to age group. The mean rating for each feasibility outcome was calculated according to age group. The distribution of submissions overall and according to age group was also illustrated using a boxplot with the median and IQR. Cross-tabulations using Fisher exact test were calculated to assess whether the rating category for feasibility outcomes differed by age group. The rating categories used for the Fisher exact test were low (≤3) and high (4-5).

We conducted a thematic analysis [34] of focus group data to assess feedback on experiences while using the mobile app to understand usability and acceptability better. To conduct the thematic analysis, we first transcribed the focus groups and reviewed the transcripts multiple times while taking note of initial ideas and patterns. We then coded the responses inductively, identifying all meaningful concepts. These codes were data-driven and included all aspects of the data that related to the usability and acceptability of the mobile app. After coding and collating all the codes, we sorted them into identified themes by combining related codes into an overarching theme. We created a thematic chart showing themes, subthemes, and related codes. The themes were then reviewed for internal homogeneity and external heterogeneity. The final themes were then named and defined, after which they were synthesized and presented [34]. Verification strategies were used in this study to ensure rigor, consistency, and suitability across the research process, including methodological coherence, memoing, negative case analysis, concurrent data collection and analysis, and sampling sufficiency. To ensure methodological coherence, the methods used in data collection and analysis were carefully selected to achieve the objectives of this research [35,36]. Memoing of themes and ideas was carried out during and after data collection, which helped identify if the research objectives were being answered [37]. The negative cases, which are those that contradict emerging patterns, were explored to assess the deviation from emerging themes [35]. Data collection and analyses were conducted concurrently to identify themes emerging from the focus group and explore them in other focus groups. To achieve sampling sufficiency, we planned to recruit 24 youth across the province. After completing three focus groups, we recognized the need to complete more focus groups to achieve our aim of data saturation, as new information was still being generated. By the eighth focus group, there was no new information, and the study had reached data saturation [35].

Ethical Considerations

Ethics approval was obtained from the provincial Health Research Ethics Authority of Newfoundland and Labrador (NL) (#20231737). Informed consent was obtained from all participants. Participants aged 13-14 provided assent and parental consent, while those aged 15-17 provided consent. Participation was voluntary and participants could withdraw at any time. Participant privacy and confidentiality were protected in this study. The study data was deidentified during analysis and no identifiable information is included in this paper. There were no significant risks related to this study. At the end of the data collection, each participant was offered a CAD $30 (CAD $1=US $0.75) incentive for their time.

Results

Participant Characteristics

A total of 68 youth consented or assented to participate in the study and were invited to participate in the training session. Among them, 39.7% (27) attended the mandatory training session and were eligible to continue participating in the rest of the study. Of those, 22/27 (81.5%) participants completed the study. Half (11/22, 50%) of these participants were 13 - 14 years old, and a larger proportion (15/22, 68.2%) were females.

The results are presented below and include app-derived feasibility metrics and the feedback survey feasibility outcomes for 23 participants who used the mobile app and completed the feedback survey. The focus group themes were generated from 22 participants who completed all study tasks, including the focus groups.

App-Derived Feasibility Metrics

Study recruitment and retention

In total, 68 participants completed the consent form available through Qualtrics, and about two-fifths (27/68, 39.7%) of these participants attended the training. Of the 41 participants who did not attend the training, 32 (78%) were 15 - 17 years old. This disproportionate difference could be attributed to the fact that only older youth were recruited through social media advertisements, resulting in more eligible 15 - 17-year-olds overall. A few participants (n=5) dropped out after the training;

XSL•FO
RenderX

however, almost all participants who used the mobile app also participated in a focus group (95.7%, 22/23; Figure 3). Participants who dropped out at various stages of the study did so due to time constraints, travel plans, or no communication.

Figure 3. Flowchart showing recruitment and retention of study sample.

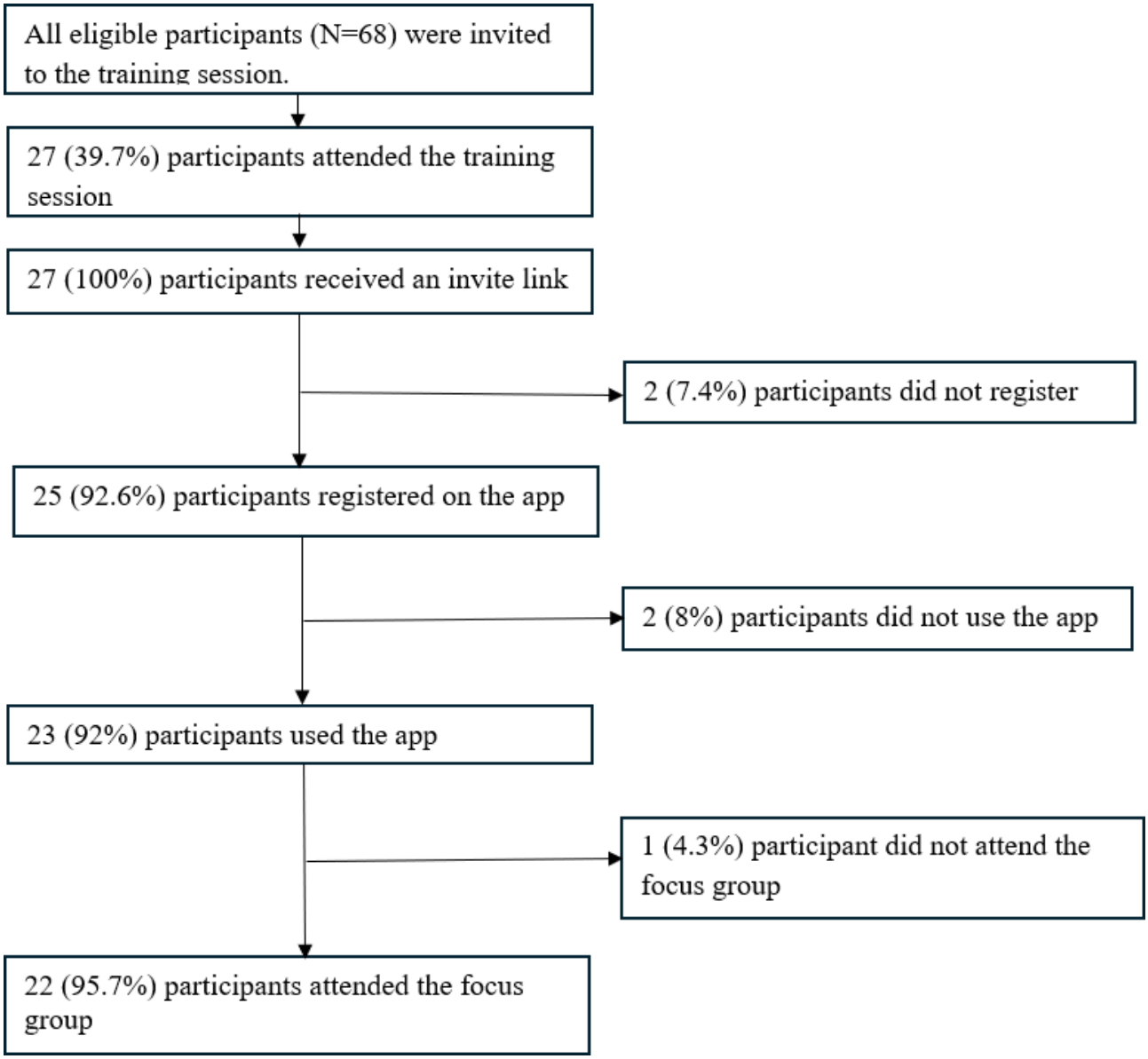

Mobile App Response and Completion Rate

For response rate, most participants (23/27, 85.2%) who received the app invite link downloaded the mobile app and used it. Almost all of the participants (23/25, 92%) who registered on the app used it until the end of the study, resulting in a high app completion rate.

Response Pattern

A total of 146 pictures were submitted, with some differences across age groups (Figure 4), with a mean of 6.3 (SD 6.14) pictures per participant and median of 4 (IQR 2-9). The number of submissions ranged from 1 to 13 advertisements, apart from 1 participant who submitted 30 advertisements. There were 87 (59.6%) submissions from 13 - 14-year-olds and 59 (40.4%) submissions from 15 - 17-year-olds. The time of diary submission varied across responses, with about half of the submissions being made from 7 PM to 10 PM.

XSL•FO
RenderX

Figure 4. Distribution of total submissions and submissions by age group (x is the mean).

Compliance Rate

Of all the participants who used the app, most (18/23, 78.3%) completed all tasks on the mobile app (app enrollment to feedback survey completion) within the 7-day window. Five participants took up to 15 days to complete these tasks (mean 11.8, SD 2.9; median 12, IQR 9-15).

Number of Food Marketing Photos by Settings or Media

The highest number of submissions was from social media, grocery stores, fast food restaurants, and shopping malls (Table 1). Youth aged 13 - 14 years tended to have more submissions across a wider variety of settings. Social media recorded the highest number of submissions from both groups.

XSL•FO
RenderX

Table . Food marketing submissions by setting and age group.

Settings or media of FoodMATS-Youth[a] App	Submissions (% of total)	Age group (years)	
		13-14, n (% of setting)	15-17, n (% of setting)
Social media	67 (45.9)	36 (53.7)	31 (46.3)
Grocery store	25 (17.1)	18 (72)	7 (28)
Fast-food restaurants	15 (10.3)	8 (53.3)	7 (46.7)
Shopping mall	10 (6.8)	9 (90)	1 (10)
Convenience store	9 (6.2)	3 (33.3)	6 (66.7)
Other[b]	9 (6.2)	7 (77.8)	2 (22.2)
Home (TV[c], social media, and flyer)	6 (4.1)	3 (50)	3 (50)
Recreation and sports center	2 (1.4)	2 (100)	0 (0)
Movie theater	1 (0.7)	0 (0)	1 (100)
School	1 (0.7)	0 (0)	1 (100)
Public transportation	1 (0.7)	1 (100)	0 (0)
Sit-down restaurants	0 (0)	0 (0)	0 (0)
Youth center	0 (0)	0 (0)	0 (0)
Total	146	87 (59.6)	59 (40.4)

[a]FoodMATS-Youth: Food and Beverage Marketing Assessment Tool for Settings and Youth

[b]Other locations were specified by participants in the submissions and included gas station, bowling alley, roadside, hotel, and craft fair.

[c]TV: television

Type of Device and App Language

About two thirds of participants used an iOS device (15/23, 65.2%), and the rest used an Android device (iOS=15 and Android=8). Almost all participants agreed that the language in the mobile app was clear (22/23, 95.7%).

Quality of Photos

The quality of the photos submitted was high, showing the product being advertised and the features of the advertisement clearly, which participants mentioned they made sure to do while using the app. Some participants mentioned in the focus group that they did not upload unclear photos they took. The size of photos ranged from 31 kBs to 8.016 MBs. The 15 kB was a picture of a billboard, most likely taken at a distance, and this was one of the poorest photos received. Although this photo had low resolution, features in the advertising were still visible.

Feasibility (Usability and Acceptability)

Average Score Rating for Each Feasibility Question

The average ratings for the feasibility questions ranged from 4.0 (SD 1.04) to 4.7 (SD 0.56), out of a possible average score rating of 5. The feasibility outcome with the lowest overall rating and the lowest rating within each age group was the “convenience” of completing entries. The average rating for each feasibility outcome was slightly higher among 15 - 17-year-olds, except for the question on app responsiveness (Table 2).

Focus group data explain average scores in feasibility outcomes. Exemplar quotes by feasibility outcome are included in Table 2. Further description related to ease, convenience, responsiveness, navigation, and willingness to use again is reviewed across three themes on participants’ experiences with the app: (1) influence of app interface and functionality on ease of use, (2) influence of app features on user satisfaction, and (3) practicality and integration of app into daily life.

XSL•FO
RenderX

Table . Mean ratings for feasibility questions and frequency of each rating according to age group.

Usability and acceptability question[a] and age group (years)	Average rating[b], mean (SD)	Ratings				Exemplar quotes from focus groups
		2 (n)	3 (n)	4 (n)	5 (n)	
Ease: How easy was it to use the app?						"I thought that the app's simplistic design was really helpful and it made it so that understanding what to do was really easy. So, I didn't have any problems with using the application. It was really like just straightforward. I enjoyed that" (Participant 8, 14 years).
13-17	4.48 (0.79)					
13-14	4.33 (0.98)	1	1	3	7	
15-17	4.64 (0.50)	0	0	4	7	
Usefulness: How helpful was the training resource provided on using the app?						".....I found it very simple right away actually, and your instructions were very straightforward as well to use the app." (Participant 21, 16 years).
13-17	4.61 (0.50)					
13-14	4.58 (0.51)	0	0	5	7	
15-17	4.64 (0.50)	0	0	4	7	
Convenience: How convenient was completing mobile diary entries?						"Well, when I was at home, and I was taking screenshots and everything, I would just send them in right away. But when I was out, and I took a few pictures, I was with friends and stuff. So, I kind of just snapped the picture, put my phone away, and was like, I'll do that when I get home. And that's what I did." (Participant 17, 17 years).
13-17	4.0 (1.04)					
13-14	3.83 (1.11)	2	2	4	4	
15-17	4.18 (0.98)	1	1	4	5	

XSL•FO
RenderX

Usability and acceptability question[a] and age group (years)	Average rating[b], mean (SD)	Ratings				Exemplar quotes from focus groups
		2 (n)	3 (n)	4 (n)	5 (n)	
Responsiveness: How responsive was the app?						"It was mostly just after I submitted my answers, it would take a good few seconds to load back into the homepage. So I was afraid if I closed it too fast, it wouldn't save." (Participant 22, 15 years).
13-17	4.48 (0.79)					
13-14	4.58 (0.67)	0	1	3	8	
15-17	4.36 (0.92)	1	0	4	6	
Navigation: As it relates to navigation, how easy was it to move in and out of the app?						"I found it very straightforward to use. Once I opened the app, every, I guess, location that I could have found ads in was there, and all I had to do was click it. And the questions were also very straightforward, and I felt the questions were about the point right away. So it was very easy to put in my information and then click submit with and without internet." (Participant 21, 16 years).
13-17	4.70 (0.56)					
13-14	4.67 (0.49)	0	0	4	8	
15-17	4.73 (0.65)	0	1	1	9	
Willingness to use again: How willing are you to use this app again?						"I don't think there's anything that I would change about the app really, because like, I think that like the layout is really nice. And the questions like they said, they were very easy and brief." (Participant 6, 15 years).
13-17	4.26 (1.01)					
13-14	4.0 (1.13)	2	1	4	5	
15-17	4.55 (0.82)	0	2	1	8	
Total score (out of 30)[c]						

XSL•FO
RenderX

Usability and acceptability question[a] and age group (years)	Average rating[b], mean (SD)	Ratings				Exemplar quotes from focus groups
		2 (n)	3 (n)	4 (n)	5 (n)	
13-17	26.53					
13-14	25.99					
15-17	27.1					

[a]Rating for each feasibility question ranged from 1 - 5. There was no rating of 1 from participants on any of the questions.

[b]Total possible rating for each question is 5.0.

[c]This was calculated by adding up the total possible rating for all feasibility questions.

Association between age group and rating category

There was no significant association between the proportion of individuals by age group that ranked each feasibilty variable as high or low (2-tailed, P>.05; Table 3).

Table . Association between age group and rating category.

Feasibility outcomes[a] and rating category[b]	Age group (years)		*P* value[c]
	13-14, n (%)	15-17, n (%)	
Ease			
Low (<3)	2 (16.7)	0 (0)	.48
High (4-5)	10 (83.3)	11 (100)	
Convenience			
Low (<3)	4 (33.3)	2 (18.2)	.64
High (4-5)	8 (66.7)	9 (81.8)	
Responsiveness			
Low (<3)	1 (8.3)	1 (9.1)	≥.99
High (4-5)	11 (91.7)	10 (90.9)	
Navigation			
Low (<3)	0 (0)	1 (9.1)	.48
High (4-5)	12 (100)	10 (90.9)	
Willingness to use it again			
Low (<3)	3 (25)	2 (18.2)	≥.99
High (4-5)	9 (75)	9 (81.8)	

[a]The usefulness of training resources could not be computed as all participants rated it high.

[b]Feasibility scores (1–5) were categorized as low (1–3) and high (4–5) for analysis.

[c]Fisher exact test was used due to the small sample size.

Focus Group Themes

Theme 1: Influence of App Interface and Functionality on Ease of Use

Overall, most participants found the application easy to use. This description of the application was often associated with how the design of the application was user-friendly, simple, and streamlined. Most participants were able to onboard and begin using the app easily. A participant said:

> *I thought it was very like straightforward to use like, I didn't find it complicated or anything, and I wasn't really like confused about how to use it. Um, yeah, it was it was straight forward, it was easy to use.* [Participant 17, 17 years]

Participants also expressed that the mobile app made the process of submitting photos brief and efficient. It was quick and easy for participants to find their way through the app when they needed to upload photos and complete the diary entries. Another participant said:

XSL•FO
RenderX

> *I found it very straightforward to use. Once I opened the app, every, I guess, location that I could have found ads in was there, and all I had to do was click it. And the questions were also very straightforward, and I felt the questions were about the point right away. So it was very easy to put in my information and then click submit with and without internet.* [Participant 21, 16 years]

Problems experienced were lagging at the point of submission, leaving them unsure if the submission had gone through. The mobile app did not have a repository where participants could view their previous submissions, so this further influenced the doubt of whether submissions had gone through. There were also a few occurrences where participants completed a submission, but their photos were not attached. One participant recalled:

> *It was mostly just after I submitted my answers, it would take a good few seconds to load back into the homepage. So I was afraid if I closed it too fast, it wouldn't save.* [Participant 22, 15 years]

Theme 2: Influence of App Features on User Satisfaction

Users reported different reactions to the features of the FoodMATS-Youth, including prompted questions on the diaries, notifications, cover photos, and locations. Some participants expressed confidence that the questions were relevant and of the right number, as they helped describe the advertisement without the need to put in too much information. Other participants perceived the process of completing questions for every submission as time-consuming. These participants preferred fewer questions or an option to only answer questions on their favorite advertisements. Concerning the app settings, participants reported that the FoodMATS-Youth diary options contained most of the locations they often visited, which made it evident that it was targeted toward their age group. This also emerged as one of their favorite things about the app. The only other location that participants would like to have added to the app was the "roadside." One participant said:

> *Personally, for me, I just like how many options there were. Like, there were a lot of common categories that I go to often, and I also really like there was like an option called others, so you could select that, so it's not like, you know, "I took this photo, and this category isn't there, so I guess it's not good." So I like that that category was included.* [Participant 9, 13 years]

The cover photos made the home page visually appealing, and they suggested the diaries themselves needed to be more fun and colorful to fit with the home page.

All participants found the push notifications to be helpful reminders, but some also found their frequency annoying, recalling that they came in too often and they no longer needed that many notifications once they got into the habit of uploading pictures. One participant also mentioned how it would be better for the notifications to say different things. She said:

> *You know, to be very honest, they're kind of repetitive. You know, they keep saying the same thing over and over, and that's the thing I don't like about apps. Notifications, I mean, sorry…See, it just told me one 'where are you' and that's the same one as before. Like say, for example, number two will be like. 'Did you forget to take a picture? Take your picture now,' or another one can be like, 'Oh, are you are you eating at a restaurant?' or something like that.* [Participant 20, 16 years]

Theme 3: Practicality and Integration of App Into Daily Life

Most participants mentioned that the app was minimally disruptive to their daily lives, and they did not find it burdensome. This was sometimes a result of being on break from school or living in small communities with minimal advertising. The simplicity of the submission process was also a factor. Some participants said:

> *It wasn't really that difficult. It only takes, like, thirty seconds to take a picture and send it, so it's not all that hard.* [Participant 10, 13 years]

> *It was a pretty simple process, so it didn't really interfere with anything, and I was also on Easter break so that also made it easier* [Participant 15, 17 years]

When asked if they made diary entries immediately upon seeing advertisements, some participants said they did not. They found it more convenient to complete the mobile diaries at home or upload all pictures at the end of the day. A participant said she had fixed goals when shopping and did not pay attention to advertising outside. There were also mixed reactions to taking photos in public as some participants were confident and had no problems taking pictures, while others were nervous. This was not age-specific. Overall, most participants uploaded screenshots from digital media immediately, while other pictures were uploaded from home. Some participants recalled:

> *I felt kind of nervous because a lot of the times I thought that other people would think I was taking pictures of them and not like the product or like the advertising. So I was a bit, like, awkward while doing it, um, so I didn't want anybody to think. I was, like, taking pictures of them.* [Participant 8, 14 years]

Another said:

> *Yeah. I, felt kind of the same, where it's kind of awkward to take pictures of advertisement and then punch in some things on your phone. So, it's definitely easier to take a picture and then fill in the details later.* [Participant 7, 13 years]

Discussion

Principal Findings

This study aimed to develop and test the feasibility of the FoodMATS-Youth app as a tool to crowdsource data from youth. The results from the feedback survey and focus groups demonstrate that the FoodMATS-Youth had high usability and acceptability, making it a feasible app to capture food marketing messages observed by youth in their daily lives. This was

XSL•FO
RenderX

supported by the high response, completion, and compliance rates during the study. Overall, participants were satisfied with the ease of use, navigation, aesthetics, responsiveness, and appropriateness of the mobile app, and this did not differ by age group. However, there were also challenges faced by participants while trying to make submissions through the app and recommendations by the participants on how to improve it.

Most feasibility outcomes being measured were rated high by participants, indicating excellent usability and acceptability of the mobile app. This is comparable to findings from another study, which evaluated a mobile app's usability in terms of ease of use, willingness to use again, and clarity [38]. Although there were only 9 participants in that study, 100% (9/9) rated the app as easy to use, and 78% (7/9) expressed willingness to use it again. Similarly, in our own study, 91.3% (21/23) of participants rated the app's ease of use as high, and 78.3% (18/23) of participants gave a high rating for their willingness to use it again. Mobile app usability often depends on the interface design, which influences ease of use and user-friendliness [39]. This explains why users frequently reported that our mobile app's simplistic and efficient design was one of the reasons it was easy to use. Previous researchers have reported that adolescents are interested in apps based on their appearance and how they feel, especially if they are aesthetically pleasing and simple to use [40]. The acceptability of the mobile app was generally high, as participants had positive experiences using it and were willing to use it again. They also believed that the app was designed for their age group based on the diaries.

Overall, the convenience of diary completion was satisfactory. However, it had the lowest mean score of all acceptability outcomes measured. Participants stated they had no problems incorporating the app into their daily routine; however, most participants did not complete the diary submissions immediately after taking the pictures. This was not a problem when they submitted photos, as they could refer to the photos to respond to questions about the product being advertised, the setting, and the marketing features present in the advertisement. However, not submitting them immediately meant participants occasionally forgot to upload some photos, which has important implications on the accuracy of monitoring food marketing through digital apps. Convenience is an important factor for mobile app users, and they prefer data collection tools that require minimal effort. If participants need to complete manual entries, like in the case of our app, it should be as convenient as possible. Lack of convenience would limit participant engagement [41] and may affect data quality.

A large portion of the photos submitted was from social media. Many reasons could influence this, including the possibility of participants not visiting as many physical settings during the 3 days of data collection or being preoccupied with other activities while those settings. Some study participants reported that they drove by billboards and did not get an opportunity to take good photos. Looking at findings from the literature, food marketing is significantly moving into the digital space [42,43], which could explain the volume of advertisements from social media. Considering the level of food marketing exposure through social media and the influence it has, it is important to focus food marketing monitoring efforts on this medium [43]. These findings suggest that social and digital media may be a significant source of food marketing exposure to youth, potentially influencing their food consumption patterns.

The mobile app completion rate was very high at 92%. Most of the participants who registered on the app completed the study. This was higher than other feasibility studies testing mobile apps among youth, which had overall completion rates of 33% and 68% [44,45]. However, these studies were interventions and often required a longer app usage period. This could account for the lower completion rates in these studies; unlike ours, where participants had to use the app for 3 days, these studies required usage for 6 weeks. Most of the participants who attended our training session used the app till the end, and they found the training session very useful, as it had one of the highest ratings on the feedback survey. A similar mobile diary study mentioned that they set up meetings with participants for an app demonstration before starting the study to increase participant retention [13]. This could explain our high response and completion rates, as participants already understood how the app worked and had resources to consult when necessary. This suggests that training and demonstrations improve user engagement, thereby contributing to high completion rates among app users.

The push notifications garnered mixed reactions from our participants, and this finding is similar to another study, which found a divide among participants on receiving notifications in general [30]. Another study that used a similar number of notifications daily also reported that participants complained of the notifications being too much, repetitive, and random [46]. This suggests that fewer notifications might be more acceptable. It could also be relevant to consider event-based notifications where the app detects movement using mobile sensors and sends notifications to participants when they leave home [47]. This has to be adequately considered, as participants can feel monitored when they receive notifications [46]. Notifications are essential, as they help remind participants to take photos and could be sent individually to participants who need them more [13]. Notifications on the FoodMATS-youth may also be important in reducing potential underestimation of food marketing exposures.

Due to the app's self-report design, confirmation bias could have made some participants more likely to notice and report marketing instances that align with their existing beliefs or expectations about unhealthy food marketing, therefore underestimating exposure. However, food marketing is ubiquitous, and the FoodMATS-Youth in its current form could provide information on the most impactful food marketing to our participants, which may have important policy implications. To improve this, the app could be further designed to include automated screen capturing of digital food marketing [48] or video recording using wearable cameras [49]. The findings from this study show that the mobile app has the potential to contribute to Health Canada's Food and Beverage Marketing Monitoring Framework for Canada [50] through crowdsourcing marketing instances from youth. This proves important not only because it is a digital data collection tool but also because it covers all channels and settings within the scope of the

XSL•FO
RenderX

monitoring framework. It also focuses on the population of interest (13 years and older) [50] and can be used to assess the exposure and power of food marketing. While it is being tested in Newfoundland and Labrador, which is one of the locations of monitoring, it could also be further adapted throughout Canada. However, before broader implementation can be suggested, the utility of the mobile app needs to be further explored in larger-scale studies. The app could also be paired with other tools, such as eye tracking, screen capture, or video recording to reduce the risk of self-report bias and improve measurement of food marketing exposure. Ultimately, using the FoodMATS-Youth within Health Canada's framework could generate data that could be used to advocate for policies restricting food marketing, beyond television and digital media, which are currently the focus of policy efforts [51].

Overall, participants said they enjoyed the mobile app and gave recommendations to improve the experience. Some of the recommendations from the focus group included building transitions or animations to the diaries so the aesthetics can match the home page and make it more visually appealing. Another was alternating the content of the push notifications, so it does not seem redundant. Finally, participants suggested a repository where they can see past submissions, which would be helpful. These are recommendations to be considered in future updates for this mobile app and future app-based studies. The feedback from users suggests that they significantly interacted with the app and shows its usability and acceptability among youth.

Strengths and Limitations

This was a mixed methods study, which quantitatively assessed the feasibility of the mobile app while exploring youth's perspectives and experiences with the rigor of qualitative research. This study developed a novel digital app-based data collection tool that can be used to collect data on cumulative food marketing seen by youth. This can include hard-to-reach populations in future studies on food marketing to youth in Canada and beyond. It also has the potential to be integrated with dietary assessment to evaluate relationships between cumulative cross-setting food marketing exposures and dietary intake. This study also adds to the literature on the feasibility of mobile apps as data collection tools. The sample size in this study was small, which has reduced statistical power and may have hidden any significant differences in feasibility outcomes between age groups. While this sample size was sufficient to reach data saturation for qualitative analysis, it was insufficient for more detailed and complex quantitative analysis. The sample size also limits the generalizability of the findings to the general population of Newfoundland and Labrador and of Canada. Also, participants were only required to take pictures for 3 days, and some additional measures may be needed to support participant engagement over a longer data collection period. Some participants did not upload all the pictures of marketing that they took, as they did not always upload the pictures immediately. This likely underestimates the amount of marketing they are exposed to, but it shows those that are potentially more important to youth. This could also be due to attentional bias. Participants sometimes found it more convenient to upload photos at home or at the end of the day. This research still needs to be replicated in a larger sample to investigate differences in app usage by demographic group and ascertain that it can be successfully used for large-scale monitoring of food and beverage marketing exposures.

Conclusion

The FoodMATS-Youth application had high usability and acceptability among 13 to 17-year-olds to capture food and beverage marketing messages observed within their daily lives. Participants found it easy to use and experienced little to no issues with its functionality. The results from this study demonstrate the opportunity to use the FoodMATS-Youth to fill gaps in evidence on food marketing across settings and show its potential to contribute to food marketing monitoring efforts in Canada and internationally.

Acknowledgments

This study was supported by the Janeway Children's Hospital Foundation. We are grateful to Community Youth Networks in Newfoundland and Labrador for their assistance with participant recruitment.

Authors' Contributions

IOB, RP contributed to conceptualization, formal analysis, funding acquisition, investigation, and resources. IOB performed data curation, handled project administration, and managed visualization. IOB, RP, LT, and KS were involved in methodology, writing–original draft, and writing–review and editing. RP, LT, and KS provided supervision.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Focus group guide.
[DOCX File, 17 KB - formative_v9i1e79306_app1.docx]

References

XSL•FO
RenderX

1. Food marketing exposure and power and their associations with food-related attitudes, beliefs, and behaviours: a narrative review. World Health Organization. 2022. URL: https://www.who.int/publications/i/item/9789240041783 [accessed 2025-11-06]
2. 2017 Report on the Health of Canadians. : Heart and Stroke Foundation of Canada; 2017 URL: https://www.heartandstroke.ca/-/media/pdf-files/canada/2017-heart-month/heartandstroke-reportonhealth2017.ashx
3. Protecting children from the harmful impact of food marketing: policy brief. World Health Organization. 2022. URL: https://www.who.int/publications/i/item/9789240051348 [accessed 2022-07-09]
4. Marketing of ultra-processed and processed food and non-alcoholic drink products. Pan American Health Organization. URL: https://www.paho.org/en/topics/marketing-ultra-processed-and-processed-food-and-non-alcoholic-drink-products [accessed 2024-12-21]
5. Monitoring of marketing of unhealthy products to children and adolescents - protocols and templates. World Health Organization. URL: https://www.who.int/europe/tools-and-toolkits/monitoring-of-marketing-of-unhealthy-products-to-children-and-adolescents---protocols-and-templates [accessed 2024-03-30]
6. Kelly B, Backholer K, Boyland E, et al. Contemporary approaches for monitoring food marketing to children to progress policy actions. Curr Nutr Rep 2023 Mar;12(1):14-25. [doi: 10.1007/s13668-023-00450-7] [Medline: 36746878]
7. Tatlow-Golden M, Jewell J, Zhiteneva O, Wickramasinghe K, Breda J, Boyland E. Rising to the challenge: introducing protocols to monitor food marketing to children from the World Health Organization Regional Office for Europe. Obes Rev 2021 Nov;22(S6):e13212. [doi: 10.1111/obr.13212] [Medline: 34184400]
8. Lara-Mejía V, Franco-Lares B, Lozada-Tequeanes AL, Villanueva-Vázquez C, Hernández-Cordero S. Methodologies for monitoring the digital marketing of foods and beverages aimed at infants, children, and adolescents (ICA): a scoping review. Int J Environ Res Public Health 2022 Jul 23;19(15):8951. [doi: 10.3390/ijerph19158951] [Medline: 35897321]
9. Elliott C, Truman E, Black JE. Tracking teen food marketing: participatory research to examine persuasive power and platforms of exposure. Appetite 2023 Jul 1;186:106550. [doi: 10.1016/j.appet.2023.106550] [Medline: 37019155]
10. Elliott C, Truman E, Aponte-Hao S. Food marketing to teenagers: examining the power and platforms of food and beverage marketing in Canada. Appetite 2022 Jun 1;173:105999. [doi: 10.1016/j.appet.2022.105999] [Medline: 35292304]
11. Prowse R. Food marketing to children in Canada: a settings-based scoping review on exposure, power and impact. Health Promot Chronic Dis Prev Can 2017 Sep;37(9):274-292. [doi: 10.24095/hpcdp.37.9.03] [Medline: 28902477]
12. Chong SOK, Pedron S, Abdelmalak N, Laxy M, Stephan AJ. An umbrella review of effectiveness and efficacy trials for app-based health interventions. NPJ Digit Med 2023 Dec 16;6(1):233. [doi: 10.1038/s41746-023-00981-x] [Medline: 38104213]
13. García B, Welford J, Smith B. Using a smartphone app in qualitative research: the good, the bad and the ugly. Qual Res 2016 Oct;16(5):508-525. [doi: 10.1177/1468794115593335]
14. So long landline, hello smartphone. Statistics Canada. 2023. URL: https://www.statcan.gc.ca/o1/en/plus/3582-so-long-landline-hello-smartphone [accessed 2024-05-25]
15. Digital well-being: the relationship between technology use, mental health and interpersonal relationships. Statistics Canada. 2024. URL: https://www150.statcan.gc.ca/n1/pub/22-20-0001/222000012024001-eng.htm [accessed 2024-07-25]
16. Harari GM, Lane ND, Wang R, Crosier BS, Campbell AT, Gosling SD. Using smartphones to collect behavioral data in psychological science: opportunities, practical considerations, and challenges. Perspect Psychol Sci 2016 Nov;11(6):838-854. [doi: 10.1177/1745691616650285] [Medline: 27899727]
17. Claudio MC, Rehany Z, Stachtari K, Guadagno E, Osmanlliu E, Poenaru D. Exploring the digital divide: results of a survey informing mobile application development. Front Digit Health 2024;6:1382507. [doi: 10.3389/fdgth.2024.1382507] [Medline: 38800096]
18. Matthews M, Doherty G, Coyle D, Sharry J. Designing mobile applications to support mental health interventions. In: Handbook of Research on User Interface Design and Evaluation for Mobile Technology: IGI Global Scientific Publishing; 2011:635-656. [doi: 10.4018/978-1-59904-871-0.ch038]
19. Aldridge RW, Burns R, Kirkby V, et al. Health on the Move (HOME) Study: using a smartphone app to explore the health and wellbeing of migrants in the United Kingdom. Wellcome Open Res 2020;5:268. [doi: 10.12688/wellcomeopenres.16348.1] [Medline: 33842695]
20. Tugault-Lafleur CN, Barr SI, Black JL. Examining differences in school hour and school day dietary quality among Canadian children between 2004 and 2015. Public Health Nutr 2019 Nov;22(16):3051-3062. [doi: 10.1017/S1368980019000788] [Medline: 31446898]
21. Body mass index, overweight or obese, self reported, youth (12 to 17 years old. Statistics Canada. 2023. URL: https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1310009621 [accessed 2024-07-25]
22. Meinert E, Rahman E, Potter A, Lawrence W, Van Velthoven M. Acceptability and usability of the mobile digital health app NoObesity for families and health care professionals: protocol for a feasibility study. JMIR Res Protoc 2020 Jul 22;9(7):e18068. [doi: 10.2196/18068] [Medline: 32706703]
23. Malte CA, Dulin PL, Baer JS, et al. Usability and acceptability of a mobile app for the self-management of alcohol misuse among veterans (Step Away): pilot cohort study. JMIR Mhealth Uhealth 2021 Apr 8;9(4):e25927. [doi: 10.2196/25927] [Medline: 33830064]

XSL•FO
RenderX

24. Chan AHY, Honey MLL. User perceptions of mobile digital apps for mental health: acceptability and usability - An integrative review. J Psychiatr Ment Health Nurs 2022 Feb;29(1):147-168. [doi: 10.1111/jpm.12744] [Medline: 33604946]
25. ISO 9241-11:2018 - ergonomics of human-system interaction — part 11: usability: definitions and concepts. International Organization for Standardization. 2018. URL: https://www.iso.org/standard/63500.html [accessed 2024-05-25]
26. Sekhon M, Cartwright M, Francis JJ. Acceptability of healthcare interventions: an overview of reviews and development of a theoretical framework. BMC Health Serv Res 2017 Jan 26;17(1):88. [doi: 10.1186/s12913-017-2031-8] [Medline: 28126032]
27. Langer SL, Ghosh N, Todd M, et al. Usability and acceptability of a smartphone app to assess partner communication, closeness, mood, and relationship satisfaction: mixed methods study. JMIR Form Res 2020 Jul 6;4(7):e14161. [doi: 10.2196/14161] [Medline: 32628614]
28. Lai P. The literature review of technology adoption models and theories for the novelty technology. JISTEM USP 2017;14(1):21-38. [doi: 10.4301/S1807-17752017000100002]
29. Vuong AM, Huber JC, Bolin JN, et al. Factors affecting acceptability and usability of technological approaches to diabetes self-management: a case study. Diabetes Technol Ther 2012 Dec;14(12):1178-1182. [doi: 10.1089/dia.2012.0139] [Medline: 23013155]
30. Creswell JW. Educational Research: Planning, Conducting, and Evaluating Quantitative and Qualitative Research, 3rd edition: Pearson Education, Inc; 2008.
31. Prowse R, Storey K, Olstad DL, Carson V, Raine KD. Choice, motives, and mixed messages: a qualitative photo-based inquiry of parents' perceptions of food and beverage marketing to children in sport and recreation facilities. Int J Environ Res Public Health 2022 Feb 23;19(5):2592. [doi: 10.3390/ijerph19052592] [Medline: 35270284]
32. Caswell MS, Hanning RM. Adolescent perspectives of the recreational ice hockey food environment and influences on eating behaviour revealed through photovoice. Public Health Nutr 2018 May;21(7):1255-1265. [doi: 10.1017/S1368980018000289] [Medline: 29493480]
33. Potvin Kent M, Hatoum F, Wu D, Remedios L, Bagnato M. Benchmarking unhealthy food marketing to children and adolescents in Canada: a scoping review. Health Promot Chronic Dis Prev Can 2022 Aug;42(8):307-318. [doi: 10.24095/hpcdp.42.8.01] [Medline: 35993602]
34. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
35. Morse JM, Barrett M, Mayan M, Olson K, Spiers J. Verification strategies for establishing reliability and validity in qualitative research. Int J Qual Methods 2002 Jun;1(2):13-22. [doi: 10.1177/160940690200100202]
36. Johnson JL, Adkins D, Chauvin S. A review of the quality indicators of rigor in qualitative research. Am J Pharm Educ 2020 Jan;84(1):7120. [doi: 10.5688/ajpe7120] [Medline: 32292186]
37. Birks M, Chapman Y, Francis K. Memoing in qualitative research: probing data and processes. J Res Nurs 2008;13(1):75-68. [doi: 10.1177/1744987107081254]
38. Tark R, Metelitsa M, Akkermann K, Saks K, Mikkel S, Haljas K. Usability, acceptability, feasibility, and effectiveness of a gamified mobile health intervention (triumf) for pediatric patients. JMIR Serious Games 2019 Sep 30;7(3):e13776. [doi: 10.2196/13776] [Medline: 31573904]
39. Huang Z, Tian ZY. Analysis and design for mobile applications: a user experience approach. In: Marcus A, Wang W, editors. Design, User Experience, and Usability: Theory and Practice 2018, Vol. 10918:91-100. [doi: 10.1007/978-3-319-91797-9_7]
40. Chan A, Kow R, Cheng JK. Adolescents' perceptions on smartphone applications (apps) for health management. JournalMTM 2017;6(2):47-55. [doi: 10.7309/jmtm.6.2.6]
41. Vaghefi I, Tulu B. The continued use of mobile health apps: insights from a longitudinal study. JMIR Mhealth Uhealth 2019 Aug 29;7(8):e12983. [doi: 10.2196/12983] [Medline: 31469081]
42. Smith R, Kelly B, Yeatman H, Boyland E. Food marketing influences children's attitudes, preferences and consumption: a systematic critical review. Nutrients 2019 Apr 18;11(4):875. [doi: 10.3390/nu11040875] [Medline: 31003489]
43. Potvin Kent M, Bagnato M, Amson A, et al. #junkfluenced: the marketing of unhealthy food and beverages by social media influencers popular with Canadian children on YouTube, Instagram and TikTok. Int J Behav Nutr Phys Act 2024 Apr 11;21(1):37. [doi: 10.1186/s12966-024-01589-4] [Medline: 38605322]
44. Werner-Seidler A, Wong Q, Johnston L, O'Dea B, Torok M, Christensen H. Pilot evaluation of the Sleep Ninja: a smartphone application for adolescent insomnia symptoms. BMJ Open 2019 May 27;9(5):e026502. [doi: 10.1136/bmjopen-2018-026502] [Medline: 31133584]
45. Hugh-Jones S, Pert K, Kendal S, et al. Adolescents accept digital mental health support in schools: a co-design and feasibility study of a school-based app for UK adolescents. Ment Health Prev 2022 Sep;27:200241. [doi: 10.1016/j.mhp.2022.200241]
46. Smelror RE, Bless JJ, Hugdahl K, Agartz I. Feasibility and acceptability of using a mobile phone app for characterizing auditory verbal hallucinations in adolescents with early-onset psychosis: exploratory study. JMIR Form Res 2019 May 14;3(2):e13882. [doi: 10.2196/13882] [Medline: 31094321]

XSL•FO
RenderX

47. Shevchenko Y, Kuhlmann T, Reips UD. Samply: a user-friendly smartphone app and web-based means of scheduling and sending mobile notifications for experience-sampling research. Behav Res Methods 2021 Aug;53(4):1710-1730. [doi: 10.3758/s13428-020-01527-9] [Medline: 33528818]
48. Gupta A, Backholer K, Huggins CE, Leung GK, Bennett R, Peeters A. Consumers acceptability of using screen capture methods to capture marketing strategies on online food delivery platforms: a qualitative study. Public Health Nutr 2025 Apr 14;28(1):e88. [doi: 10.1017/S1368980025000515] [Medline: 40223705]
49. Signal LN, Stanley J, Smith M, et al. Children's everyday exposure to food marketing: an objective analysis using wearable cameras. Int J Behav Nutr Phys Act 2017 Oct 8;14(1):137. [doi: 10.1186/s12966-017-0570-3] [Medline: 28988542]
50. Potvin Kent M, Mulligan C, Pauzé E, Pinto A, Remedios L. The food and beverage marketing monitoring framework for Canada: development, implementation, and gaps. Food Policy 2024 Jan;122:102587. [doi: 10.1016/j.foodpol.2023.102587]
51. Policy update on restricting food advertising primarily directed at children: overview. Government of Canada. 2023. URL: https://www.canada.ca/en/health-canada/services/food-nutrition/healthy-eating-strategy/policy-update-restricting-food-advertising-primarily-directed-children.html [accessed 2024-03-30]

Abbreviations

FoodMATS-Youth: Food and Beverage Marketing Assessment Tool for Settings and Youth
GrabFM: Grab Food Marketing
QUAL: Qualitative
QUAN: Quantitative

Edited by A Mavragani, J Sarvestan; submitted 18.06.25; peer-reviewed by R Bennett, R Bosward; revised version received 25.08.25; accepted 21.09.25; published 09.12.25.

Please cite as:
Bamigbayan IO, Twells LK, Shannahan KLJ, Prowse R
Exploring the Usability and Acceptability of the FoodMATS-Youth App for Monitoring Food Marketing Exposures: Mixed Methods Study
JMIR Form Res 2025;9:e79306
URL: https://formative.jmir.org/2025/1/e79306
doi:10.2196/79306

© Idris Opeyemi Bamigbayan, Laurie K Twells, Kirby L J Shannahan, Rachel Prowse. Originally published in JMIR Formative Research (https://formative.jmir.org), 9.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Initial Development of a Multidimensional Computerized Adaptive Test for Intensive Longitudinal Assessment of Suicide Risk: Development and Usability Study

Kenneth McClure[1], PhD; Brooke A Ammerman[2], PhD; Cheng Liu[3], PhD; Miguel Blacutt[3], MA, MEd; Connor O'Brien[2], BS; Ross Jacobucci[4], PhD

[1]Department of Mathematics and Statistics, University of Wyoming, 1000 E University Ave, Laramie, WY, United States
[2]Department of Psychology, University of Wisconsin–Madison, Madison, WI, United States
[3]Department of Psychology, University of Notre Dame, South Bend, IN, United States
[4]Center for Healthy Minds, University of Wisconsin–Madison, Madison, WI, United States

Corresponding Author:
Kenneth McClure, PhD
Department of Mathematics and Statistics, University of Wyoming, 1000 E University Ave, Laramie, WY, United States

Abstract

Background: Intensive longitudinal designs support temporally granular study of processes, making methods like ecological momentary assessment (EMA) increasingly common in medical and behavioral science. However, the repetitive and intensive measurement strategies associated with these designs increase participant burden, which limits the breadth and precision of EMA surveys. This is particularly problematic for complex clinical phenomena, such as suicide risk, which research has shown is multidimensional and fluctuates over narrow time intervals (eg, hours). To overcome this limitation, we proposed the Computerized Adaptive Test for Suicide Risk Pathways (CAT-SRP), which supports the simultaneous assessment of multiple empirically informed risk domains and facilitates personalized survey content.

Objective: The objective of this study is to develop, calibrate, and pilot the first multidimensional computerized adaptive test for suicidal thoughts and related psychosocial risk factors in intensive longitudinal designs like EMA.

Methods: A web-based assessment platform was developed to adaptively administer the CAT-SRP. CAT-SRP items were modified from existing validated instruments to support administration in intensive longitudinal designs. The item bank was developed in line with major ideation-to-action theories of suicide and consultation with experts outside the study team. Exploratory item factor analysis was used to identify dimensionality of the item bank. Item parameters were calibrated using a multidimensional graded response model in a large cross-sectional community sample (n=1759, 36.33% with a history of suicidal thoughts). Following calibration, the CAT-SRP was evaluated in an EMA study of participants with a past month history of suicidal thoughts (n=29 across 2134 observations). Adaptive testing used D-optimal item selection, a dual variable-length stopping criterion, and Maximum a Posteriori (MAP) scoring. Descriptive statistics and mixed effects models were used to examine CAT-SRP performance (eg, efficiency and survey overlap) and relationships among CAT-SRP domain scores.

Results: The calibration study identified 2 suicidal thought domains (active and passive thoughts) and 12 risk factor domains: humiliation, loneliness, anger, pain, defeat, impulsivity (ie, negative urgency), entrapment, distress tolerance, perceived burdensomeness, thwarted belongingness, aggression, and a positively valenced method factor. Domain information was the highest between average to high levels of domain scores. Study 2 showed that the CAT-SRP (1) administered surveys with low to moderate item overlap, (2) incurred low participant burden, and (3) may improve near-term prediction of suicidal thoughts relative to traditional EMA measurement. Most EMA surveys reached the maximum length, 50 questions, highlighting a need to refine selection and stopping rules.

Conclusions: The CAT-SRP effectively personalized EMA survey content to respondents, which reduces the repetitiveness and perceived burden of intensive longitudinal research designs. Continuous domain scores from multidimensional computerized adaptive testing (MCAT) also provided more nuanced measurement compared to traditional approaches that struggle with zero-inflation in EMA and appeared to produce stronger predictive relationships. Overall, the CAT-SRP demonstrated strong methodological advantages to use CAT for intensive longitudinal data collection.

(JMIR Form Res 2025;9:e76544) doi:10.2196/76544

KEYWORDS

ecological momentary assessment; computerized adaptive testing; suicide; measurement; participant burden

XSL•FO
RenderX

Introduction

Background

Attempts to understand suicide, a growing public health concern [1], have led to the development of several theoretical models outlining factors contributing to suicidal thoughts and behaviors. Although theoretical models, such as the Interpersonal Theory of Suicide (ITS) [2], the Three-Step Theory of Suicide (3ST) [3], and the Integrated Motivational-Volitional (IMV) [4], have received notable empirical attention [5], the translation and application to real-time risk processes remains limited. Research strongly supports that suicidal thinking is dynamic [6], with transitions between suicidal thinking and behavior potentially occurring within hours [7]. However, a comprehensive, temporally dynamic examination of these theoretical models is lacking, limiting both suicide risk assessment accuracy and advances toward near-term risk prediction.

The increasing accessibility of intensive time sampling methodologies (eg, ecological momentary assessment [EMA]) has begun to bridge this gap by capturing the fluctuating nature of suicide risk in real-time. However, limitations common in EMA protocols have hindered a comprehensive assessment of risk processes outlined in theoretical models. EMA studies typically rely on a limited, repetitive subset of predictors, frequently assessed using short-form or single-item measures that lack thorough validation [8,9]. This approach compromises measurement precision [10,11], decreases statistical power [12-14], biases parameter estimates (eg, attenuation) [12,13], and masks potential nonlinear effects [12-14]. Further, the repetitive nature of item content may increase participant burden, leading to decreased compliance rates and diminished data quality. Reliance on a narrow set of items also constrains the evaluation and refinement of theoretical models, limiting the ability to identify subgroups with distinct risk trajectories or explore alternative pathways that may not apply uniformly across populations. This restriction further hinders the development of personalized models that could enhance prediction accuracy and intervention strategies. These limitations undermine confidence in findings and hinder the interpretability of results, ultimately reducing their impact on suicide risk assessment and intervention.

Computerized adaptive testing (CAT) [15] offers a systematic and efficient approach to momentary risk assessment that addresses such limitations [16]. CAT uses formal psychometric models, typically item response models [17], to dynamically administer subsets of items from larger precalibrated item banks, with the objective of maximizing information (ie, precision) of underlying latent risk variables. This results in more efficient and personalized measurement [18]. Although prior CATs for suicide exist [19-21], each was developed as a suicide risk screening for a single time point (ie, primary care or emergency department visit), neither to comprehensively evaluate risk factors nor to be used repeatedly for momentary assessment. For example, the Computerized Adaptive Test-Suicide Scale (CAT-SS) [21] was developed as a brief screening instrument to identify individuals with nonnegligible levels of suicide risk in health care settings. The CAT-SS item pool consists predominantly of depression and anxiety items, with a small set of suicide-specific items, and uses a bifactor measurement model to assess a general risk factor that is shared among the depression, anxiety, and suicide items; other measures take a similar approach [19,20]. Although these instruments are effective at assessing overall suicide risk, they do not provide scores for individual risk factors, which are critical for understanding the development and dynamics of suicidal thoughts and behaviors. Recent work has implemented CAT in assessing symptom severity in patient-reported outcomes research using an EMA design [22], highlighting the potential for application in suicide research.

To leverage the advantages of CAT for studying suicide risk factors in intensive longitudinal designs, we developed the Computerized Adaptive Test of Suicide Risk Pathways (CAT-SRP). The CAT-SRP is the first CAT designed for monitoring the fluid nature of suicide risk factors in intensive time sampling approaches, like EMA. Unlike existing suicide CATs [20,21] which focus on screening overall risk for suicide at a single time point (eg, primary care or emergency department visit) using bifactor measurement models, the CAT-SRP emphasizes the assessment of individual risk factors (eg, hopelessness and loneliness) in intensive longitudinal designs using a multidimensional graded response model (MGRM) [23] with correlated risk factors. Specifically, the CAT-SRP is intended to assess risk factors proposed in 3 predominant ideation-to-action models of suicide [2-4] and other well-established risk factors for near-term suicidal thoughts and behaviors (STBs; eg, anger and stress [8,9,24]), as well as suicidal thoughts. The CAT-SRP is not predominantly grounded in any of the ITS, 3ST, or IMV, but instead aims to provide an instrument for assessing and comparing risk pathways from each of the theoretical models. While prior work has highlighted the salience of these factors in predicting suicide risk, no EMA study has directly and comprehensively examined the confluence of these risk factors within a single study protocol.

Using CAT in suicide EMA research also introduces key ethical considerations. First, variable-length CAT methods, which are typically more efficient [18], may result in longer surveys for some respondents. The number of items needed to terminate measurement depends on both the underlying levels of the latent variable for a given respondent and the characteristics of the item bank. Surveys will tend to be shorter when the item bank contains items that provide high information at a participant's risk level. Thus, to limit participant burden, it is often advantageous to use a dual stopping rule [25] and specify a maximum survey length. Additionally, for research contexts where participant risk is monitored and study staff may need to intervene, such as in suicide EMA research [26,27], additional care in adaptive EMA survey design is needed. A fully adaptive approach may exclude items used in risk monitoring. Thus, until adaptive risk monitoring approaches have been thoroughly validated, a static validated assessment of acute risk (eg, imminent self-harm or death) should be used across assessments like in traditional EMA research.

XSL•FO
RenderX

Study Aims

The present paper aims to develop and validate the CAT-SRP through 2 complementary studies. Study 1 develops and calibrates the CAT-SRP item bank using a large community sample. Study 2 pilots the feasibility, acceptability, and utility of the CAT-SRP as a measure of momentary suicidal ideation and associated risk factors in an EMA design using a sample of individuals with a past month history of STBs. By developing this comprehensive adaptive measurement system, we aim to advance both the methodological rigor of suicide risk assessment in intensive longitudinal research and our theoretical understanding of how risk factors dynamically relate to suicidal thoughts and behaviors.

Methods

Study 1

Participants

Participants (n=2498) were recruited using Prime Panels (CloudResearch) [28]. Participants were required to be at least 18 years of age, speak English, and reside in the United States. After providing informed consent, participants were asked to complete the full CAT-SRP item pool and demographic items. Respondents failing automated bot detection (ie, reCAPTCHA; n=14), providing completely missing responses (n=38), completing the entire battery in under 7 minutes (n=103), or failing more than 1 out of 6 attention checks (n=720) were excluded from analyses (note some respondents were flagged in multiple checks). This resulted in a final sample of n=1759 adults (mean age 46, SD 16.9 years). See Table 1 for sample demographics. Of the final sample, 36.33% of respondents reported a lifetime history of suicidal ideation, planning, or attempts. Missing data for CAT-SRP item responses were minimal (0.85%).

Table . Study demographics.

Variable	Study 1 (n=1759)	Study 2 (n=29)
Gender, n (%)		
Woman	957 (54.41)	18 (62.07)
Man	787 (44.74)	10 (34.48)
Transgender	12 (0.68)	0 (0)
Prefer not to Answer	3 (0.17)	0[a] (0)
Ethnicity, n (%)		
Non-Hispanic or Latinx	1314 (74.7)	3 (10.34)
Hispanic or Latinx	427 (24.28)	26 (89.66)
Prefer not to answer	18 (1.02)	0 (0)
Race		
White	1149 (65.32)	20 (68.97)
Black or African American	439 (24.96)	6 (20.69)
Multiracial	62 (3.52)	2 (6.70)
Asian	22 (1.25)	0 (0)
American Indian or Alaskan Native	17 (0.97)	0 (0)
Native Hawaiian or Pacific Islander	3 (0.17)	0 (0)
Not listed or prefer not to answer	67 (3.81)	0 (0)

[a]One participant identified as nonbinary.

Procedure

Target domains for the CAT-SRP were selected based on (1) predominant ideation-to-action theories of suicide [2-4] and (2) consultation with 4 experts in suicide research. Specifically, psychosocial risk factors central to the ITS, 3ST, or IMV, such as hopelessness, thwarted belongingness, defeat, and entrapment, were chosen as target domains. Additional risks identified by experts as potentially important for studying suicide risk in EMA (ie, anger, aggression, negative urgency, and perceived stress) were also included. In total, 12 preliminary risk factor domains were identified. CAT-SRP items (J=200) for these

XSL•FO
RenderX

domains were adapted from existing measures of corresponding constructs (see the Measures section). In addition to the 12 STB risk domains, the CAT-SRP also includes passive and active suicidal ideation domains (J=18). Measurement models for the risk factors and suicidal thoughts were calibrated separately to prevent interpretational confounding [29].

To support delivery in future intensive longitudinal designs, all CAT-SRP items were modified to conform to a 5-point Likert scale and include a day-level temporal referent (ie, "Today, …"). Maintaining a consistent response scale may also reduce participant burden in intensive longitudinal designs [30]. Original item phrasing and modified phrasing for the CAT-SRP are available on OSF [31]. Unlike previous high-dimensional adaptive tests for suicide risk, which focus on assessing general levels of risk [20,21], the CAT-SRP prioritizes assessing individual risk domains, as well as suicidal thoughts. Thus, domains were conceptualized as correlated factors rather than specific factors in a bifactor structure. The final CAT-SRP item pool and calibrated item parameters can also be found on OSF.

Measures

Perceived Burdensomeness

Eight perceived burdensomeness (PB) items were included in the initial CAT-SRP item pool. Seven items were adapted from the Interpersonal Needs Questionnaire (INQ) [32] and one item was adapted from an ultra-brief measure of suicide risk [33].

Thwarted Belongingness

Fifteen items assessing thwarted belongingness (TB) were included in the initial CAT-SRP item pool. Eight items were adapted from the INQ [32] and 7 items were adapted from the Thwarted Belongingness Scale [34].

Loneliness

Twenty items assessing loneliness were included in the initial CAT-SRP item pool and adapted from the UCLA (University of California, Los Angeles) Loneliness Scale [35].

Hopelessness

Twenty items assessing hopelessness were included in the initial CAT-SRP item pool. Ten items were adapted from the state subscale of the State-Trait Hopelessness Scale [36]. Ten items were adapted from the Brief Hopelessness Scale [37]; 5 of these items were adapted from the positive subscale (ie, hope) and 5 from the negative subscale.

Defeat and Entrapment

Sixteen items assessing defeat and 16 items assessing entrapment were included in the initial CAT-SRP item pool and were adapted from the Defeat and Entrapment Scales, respectively [38].

Humiliation

Thirty-two items assessing humiliation were included in the initial CAT-SRP item pool and were modified from the Humiliation Inventory [39].

Psychological Pain

Thirteen items assessing psychological pain were included in the initial CAT-SRP item pool and were adapted from the Psychache Scale [40].

Anger and Aggression

Twenty items assessing anger and four items assessing aggression were included in the initial CAT-SRP item pool and were adapted from the Patient-Reported Outcomes Measurement Information System (PROMIS) anger scale [41].

Perceived Stress

Ten items assessing perceived stress were included in the initial CAT-SRP item pool and were adapted from the Perceived Stress Scale [42].

Negative Urgency

Twelve items assessing negative urgency were included in the initial CAT-SRP item pool and were adapted from the negative urgency subscale of the Urgency, (Lack of) Premeditation, (Lack of) Perseverance, Sensation Seeking, and Positive Urgency (UPPS-P) Impulsive Behavior Scale [43].

Distress Tolerance

Fifteen items assessing distress tolerance were included in the initial CAT-SRP item pool and were adapted from the Distress Tolerance Scale [44].

Suicidal Ideation

Eighteen items assessing passive and active suicidal ideation were included in the initial CAT-SRP item pool. Nine items assessing passive ideation and nine items assessing active ideation were adapted from the passive and active suicidal ideation scale (PASIS) subscales, respectively [45].

Analytic Plan

Parallel analysis and scree plots using polychoric correlation matrices were used to explore CAT-SRP dimensionality. Polychoric correlation matrices with pairwise complete data were obtained using the polycor package [46]. The psych package [47] was used for parallel analyses. Exploratory item factor analysis (EIFA) was conducted for plausible model dimensionalities as suggested by parallel analysis. Interpretability of factor loading patterns, alongside parallel analysis, was used to guide selection among models with similar performance in EIFA.

After identifying item bank dimensionality, a confirmatory MGRM was fit using the mirt package [48] with Metropolis-Hastings Robbins-Monro estimation to calibrate item parameters [49,50]. Items with loadings of >.20 in the EIFA were specified to load onto corresponding domains in the MGRM, allowing within-item multidimensionality; otherwise, cross-loadings were constrained to 0. Passive and active STB items were calibrated using a 2-dimensional MGRM. Item difficulty and discrimination parameters were extracted to facilitate adaptive testing. Study data, including the CAT-SRP item bank and the original CAT-SRP blueprint, is available on OSF [31]. This study was not preregistered.

XSL•FO
RenderX

Ethical Considerations

All study procedures were reviewed and deemed exempt by the University of Notre Dame Institutional Review Board (protocol number 22-09-7402). A waiver of informed consent was obtained to preserve participant anonymity. Prior to study tasks, participants were informed of study details. Participants were informed that they were free to discontinue participation at any time by closing the survey window. Participants were compensated US $3.25.

Study 2

Participants

Participants (n=29) were recruited from online sources, including Facebook (Meta Platforms, Inc), Instagram (Meta Platforms, Inc), Reddit (Reddit, Inc), and Craigslist (Craigslist, Inc) to participate in a study on moods, feelings, and thoughts in daily life. Prospective participants completed a brief semistructured phone interview with study staff to determine eligibility. Inclusion criteria included (1) being 18 years or older, (2) having a past-month history of suicidal thoughts (eg, "Have you had thoughts of killing yourself in the past month?"), (3) speaking English, (4) being located in the United States, and (5) owning a smartphone. Participants were adults in the United States (mean age 34.62, SD 10.13 years). See Table 1 for gender, ethnicity, and race sample composition. Overall, 22 out of 29 (75.86%) of participants self-reported having received a mental health diagnosis, and 10 out of 29 (34.48%) participants reported prior hospitalization for psychiatric reasons.

Procedure

Participants attended one virtual baseline session before the EMA period. Following informed consent, participants completed a battery of self-report measures and the full CAT-SRP item bank during the initial virtual session. Participants also created an account on the CAT-SRP website to allow for EMA data collection. Responses to CAT-SRP items at baseline were used to initialize the CAT algorithm for EMA surveys (see CAT Algorithm section). The EMA period began the morning following the virtual session. Participants were sent 6 EMA survey links per day for 21 days via an automated SMS messaging system. Survey timing was randomized within six 2-hour intervals between 9 AM and 9 PM; adjacent surveys were required to occur at least 15 minutes apart. Participants received one reminder 15 minutes after the initial prompt if they had not completed the survey. All EMA notifications were delivered via text message and included a link to the CAT-SRP web platform where participants completed the adaptive CAT-SRP survey and risk monitoring items (see "Momentary Suicide Risk" measures below). See Figure S1 in Multimedia Appendix 1 for a flowchart of the design of study 2.

Risk Monitoring

If responses suggesting high suicide risk were detected during the EMA period (ie, suicidal intent reported as >8 out of 10, a reported plan to attempt suicide that day, or a reported suicide attempt), this triggered an email alert to the research team (resulting in n=4 alerts to the study team). A trained member of the study team then followed up via phone to conduct a comprehensive risk assessment; acute suicide risk was categorized based on established suicide risk management protocols [51]. As appropriate, potential call outcomes included resource review, development of a safety plan, or transportation for an in-person safety assessment (eg, emergency department). All participants were provided resources at the onset of the study. Participants reporting high-risk responses were also provided with a comprehensive resource list when high risk was detected.

Follow-Up and Compensation

Following the EMA data collection, participants were sent an online survey to assess response burden. Participants were compensated US $20 for completing the virtual assessment, US $150 for completing the 21-day EMA period (with a US $15 bonus for completing at least 75% of surveys), and US $10 for completing the final online survey. Thus, participants were compensated between US $170-US $195 in total.

CAT Algorithm

The CAT-SRP items were administered adaptively during EMA surveys. Baseline estimates of latent severity (ie, θ^) were used to initialize the CAT-SRP at the first EMA survey; all follow-up surveys were initialized using the final estimates of the preceding measurement occasion. Given the multidimensional nature of the item bank, items were selected sequentially to maximize the determinant of the Fisher Information matrix (ie, D-optimal item selection). This approach favors items that provide information on multiple latent variables [52] and was found to better recover θ^ in high-dimensional measures with shorter tests [53]. To ensure that all CAT-SRP domains were assessed directly at every measurement occasion, a "warm-up" period was implemented where the most informative item (ie, D-optimal) from each domain was administered. Thus, at least one item from each of the 14 domains was administered on every occasion. Following this warm-up period, items were adaptively administered until a stopping criterion was reached.

The CAT-SRP was administered as a variable-length CAT with additional constraints. A dual stopping criterion, which (1) evaluated observed measurement precision (ie, D-rule) and (2) convergence of θ^ across all domains for 2 sequential item administrations (ie, CT-rule), was used to terminate the test [25]. More formally, the D-rule is given by

$$n=\inf I \{j \geq 1: \det[\sum_{j=1}^{n} I_j(y_j|\hat{\theta}_n)] \geq (vc)^2\} \quad (1)$$

where

$$v= 2\pi^{m/2}[\chi^{m/2}(\alpha)]^{m/2}m\Gamma(m/2) \quad (2)$$

and *c* represents the maximum allowable confidence ellipsoid volume (typically an *m*-dimensional unit ball), and α is the desired significance level [25]. In this study, *m*=14 and α=.05. For the CT rule, the survey ended if the changes to all θ^ were less than 0.01 for 2 consecutive item administrations. A maximum test length of 50 items was enforced if neither the D-rule nor CT rule was satisfied. Thus, CAT-SRP survey lengths ranged from 14 - 50 items.

XSL•FO
RenderX

Baseline Measures: CAT-SRP

During the baseline virtual session, participants completed all 207 items retained from the initial CAT-SRP item pool (see study 1 for more details).

EMA Measures

EMA CAT-SRP

CAT-SRP items were adaptively administered during EMA surveys. Temporal referents for EMA CAT-SRP items were modified to be "Since the last prompt…" "Since the last prompt" was chosen because it was considered to be more consistent with the temporal referent in the calibration sample ("Today…") than other alternatives (eg, "right now" and "currently") while still appropriate for EMA designs. The CAT-SRP uses a variable-length sequentially adaptive item selection algorithm (see CAT Algorithm section). Thus, survey length and content potentially differed for each respondent and survey occasion. At each measurement occasion, however, at least one item from each of the CAT-SRP domains (ie, risk and suicidal ideation) was administered, resulting in a minimum adaptive survey length of 14 items. A maximum number of 50 items was also specified.

Momentary Suicide Risk

In addition to the adaptive CAT-SRP suicidal ideation items, 3 items assessing momentary STB risk were administered at each prompt: "Do you intend to act on your suicidal thoughts? (1-10)," "Do you plan to attempt suicide today or tonight? (yes or no)," "Since the last prompt have you attempted suicide (yes or no)?" These items were not evaluated as part of the CAT-SRP but used only for the purpose of identifying imminent suicide risk.

End-of-Study Measure: Participant Subjective Burden

Subjective burden was assessed using 9 questions in the end-of-study survey (see Table 2). Additional questions assessing reaction to research participation were also asked as part of a larger questionnaire [54] but not analyzed in this study.

XSL•FO
RenderX

Table . Descriptive statistics for subjective burden item responses.

Survey Item	1, n (%)	2, n (%)	3, n (%)	4, n (%)	5, n (%)	Mean (SD)	Median (IQR)
Did you feel overwhelmed by the number of survey questions[a]?	10 (34.5)	9 (31.0)	8 (27.6)	2 (6.9)	0 (0)	2.07 (0.96)	2 (1-3)
Did you feel annoyed or frustrated by the number of survey questions[a]?	10 (34.5)	12 (41.4)	5 (17.2)	2 (6.9)	0 (0)	1.97 (0.91)	2 (1-2)
Did you feel annoyed or frustrated by the number of surveys[a]?	15 (53.6)	4 (14.3)	8 (28.6)	1 (3.6)	0 (0)	1.82 (0.98)	1 (1-3)
How often did you skip a survey because it felt like too many surveys[b]?	15 (51.7)	6 (20.7)	7 (24.1)	1 (3.4)	0 (0)	1.79 (0.94)	1 (1-3)
How often did you skip a survey because it felt like too many survey questions?[b]	19 (67.9)	7 (25)	2 (7.1%)	0 (0)	0 (0)	1.39 (0.63)	1 (1-2)
I found participating boring.[c]	8 (27.6)	8 (27.6)	7 (24.1)	6 (20.7)	0 (0)	2.38 (1.12)	2 (1-3)
The study procedures took too long[c]	9 (31)	9 (31)	10 (34.5)	1 (3.4)	0 (0)	2.10 (0.90)	2 (1-3)
Participating in the study was inconvenient to me[c]	9 (31)	13 (44.8)	6 (20.7)	1 (3.4)	0 (0)	1.97 (0.82)	2 (1-2)
Had I known in advance what participating would be like, I still would have agreed to participate[c]	0 (0)	0 (0)	3 (10.3)	7 (24.1)	19 (65.5)	4.55 (0.69)	5 (4-5)

[a]1=very slightly/not at all, 2=a little, 3=moderately, 4=quite a bit, 5=extremely.

[b]1=not at all, 2=less than once per week, 3=multiple times per week, 4=once per day, 5=multiple times per day.

[c]1=strongly disagree, 2=disagree, 3=neutral, 4=agree, 5=strongly agree.

The end-of-study survey included 9 questions assessing subjective burden (eg, feeling overwhelmed or annoyed; see Table 2).

Analytic Plan

Compliance was evaluated using mixed effects regression models. A random intercept-only model was used to examine the proportion of variance in compliance rates attributable to individuals (ie, intraclass correlation coefficient [ICC]); a second model, including study week, was used to examine the impact of time in the study [55]. A similar approach was used to calculate the ICC in each of the latent variables and adaptive survey length.

Adaptive Survey Length

Survey length was examined further using a mixed effects model. Due to bimodality in survey lengths (see Figure 1), a binary variable where “1” denotes a survey of maximum length (ie, 50 items; n=1524) and “0” denotes a survey that ended before maximum length (n=610) was examined using mixed-effects logistic regression. CAT-SRP estimates and the number of surveys completed were included as fixed effects; a

XSL•FO
RenderX

random slope for the number of surveys completed was also included to allow for potential within-person variability. Due to convergence issues upon inclusion of the random slope, Bayesian methods were adopted using the brms package [56] with default noninformative priors; odds ratios (OR) of fixed effects with 95% credible intervals excluding a null effect are reported.

Figure 1. Adaptive survey length. (A) Histogram of survey lengths across the full study. (B) Survey length by participant over the study duration.

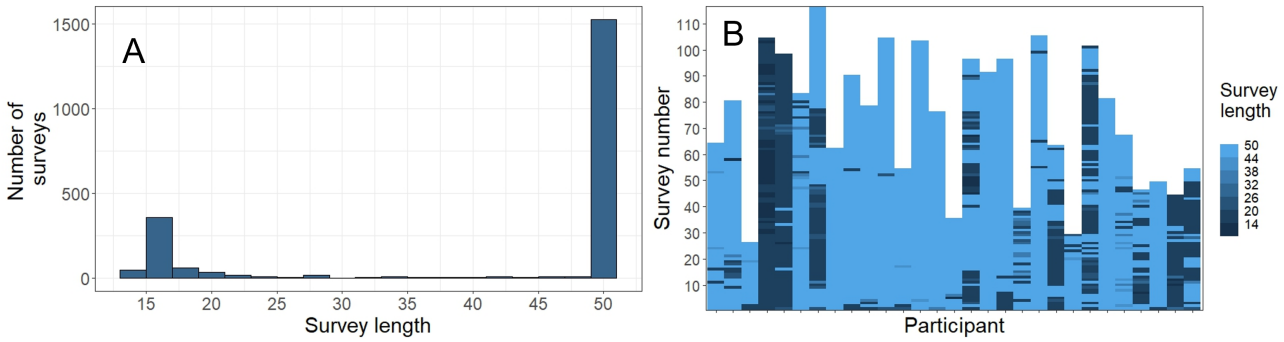

Adaptive Survey Content Overlap

To quantify item overlap, the Jaccard Index and overlap coefficient were calculated separately for each participant for all pairwise combinations of EMA surveys. For any 2 surveys, the Jaccard Index captures the ratio between the size of the intersection of items (ie, items in both surveys) and the union of items. It can be interpreted as the proportion of shared items (ie, between 2 surveys) to the total unique items across both surveys. The overlap coefficient is similar; however, the number of items for the shorter survey is used instead of the union. It can be interpreted as the proportion of shared items present in the shorter survey (ie, if survey lengths differ). Both metrics range from 0 to 1, with larger values suggesting more overlap in survey content. The average of these pairwise metrics was computed to provide a measure of survey similarity for each participant across their time in the study.

CAT-SRP Factor Structure

The within- and between-person factor structure of CAT-SRP suicidal ideation items was assessed separately in study 2, using a 2-stage estimation process. First, the within- and between-person covariance matrices were estimated using pairwise complete observations with the weighted least squares estimator in Mplus (Version 8; Muthén & Muthén) [57]. Then parallel analysis was conducted on the resulting covariance matrices to explore dimensionality. Within- and between-person factor structures of CAT-SRP risk items were not explored due to small sample size, large proportion of missing item-level data, and limited variability in item responses (ie, within individuals). A multilevel EFA using Bayesian estimation in Mplus was attempted but never converged. The large degree of item-level missingness also prevented successful estimation with the 2-stage approach applied to STB items. The within- and between-person factor structures of the CAT-SRP domain scores are also explored.

Concurrent and Predictive Validity

To assess the relationships between each of the assessed constructs and both passive and active suicidal ideation, multilevel models were fit to examine the concurrent and temporal relationships. Models included random intercepts to account for within-person clustering of observations. The concurrent model examined associations between suicidal ideation and psychological variables at the same time point. The lagged model investigated whether psychological variables predicted changes in suicidal ideation at the subsequent time point, controlling for current suicidal ideation. For prospective models, CAT-SRP scores from each completed survey were regressed onto scores from the previous completed survey using restricted maximum likelihood estimation. Fixed effects and marginal R^2 statistics are reported.

Subjective Burden

Descriptive statistics were calculated to evaluate subjective burden and acceptability of the study procedures (see Table 2 for subjective burden items).

Ethical Considerations

All study procedures were reviewed and approved by the University of Notre Dame Institutional Review Board (protocol number 23-11-8206). Participants provided written and verbal informed consent prior to study procedures. Participants were informed that they were free to discontinue participation at any time by emailing study staff. All study data were deidentified and stored on secure servers to preserve confidentiality. Participants were compensated up to US $195 (see procedures for compensation details).

Results

Study 1

Study 1 Results Overview

To facilitate adaptive testing, a confirmatory MGRM was fit to the CAT-SRP risk items and J=18 CAT-SRP suicidal ideation items separately. Fit indices for the CAT-SRP risk model were unable to be computed due to model dimensionality. Interfactor correlations are provided in Table 3. Discrimination and difficulty parameters were extracted. Unidimensional domain information across levels of latent scores is presented in Figure 2A.

XSL•FO
RenderX

Table . CAT-SRP[a] risk domain factor correlations.

Domain	Hum[b]	Lnly[c]	Ang[d]	Pain[e]	NU[f]	Dft[g]	Trap[h]	DT[i]	PB[j]	TB[k]	Hope	Agg[l]
Hum, *r*	1	0.56	0.561	0.441	0.478	0.487	0.54	0.58	0.541	0.16	0.051	0.383
Lnly, *r*	0.56	1	0.611	0.612	0.415	0.739	0.733	0.6	0.695	0.359	0.253	0.316
Ang, *r*	0.561	0.611	1	0.458	0.588	0.581	0.62	0.617	0.532	0.179	0.094	0.443
Pain, *r*	0.441	0.612	0.458	1	0	0.674	0.66	0.573	0.612	0.245	0.234	0.394
NU, *r*	0.478	0.415	0.588	0	1	0.336	0.373	0.457	0.407	0.072	0.023	0.252
Dft, *r*	0.487	0.739	0.581	0.674	0.336	1	0.78	0.62	0.683	0.334	0.355	0.299
Trap, *r*	0.54	0.733	0.62	0.66	0.373	0.78	1	0.649	0.713	0.335	0.288	0.335
DT, *r*	0.58	0.6	0.617	0.573	0.457	0.62	0.649	1	0.609	0.129	0.068	0.349
PB, *r*	0.541	0.695	0.532	0.612	0.407	0.683	0.713	0.609	1	0.342	0.291	0.439
TB, *r*	0.16	0.359	0.179	0.245	0.072	0.334	0.335	0.129	0.342	1	0.646	0.144
Hope, *r*	0.051	0.253	0.094	0.234	0.023	0.355	0.288	0.068	0.291	0.646	1	0.022
Agg, *r*	0.383	0.316	0.443	0.394	0.252	0.299	0.335	0.349	0.439	0.144	0.022	1

[a]CAT-SRP: Computerized Adaptive Test of Suicide Risk Pathways.
[b]Hum: humiliation.
[c]Lnly: loneliness.
[d]Ang: anger.
[e]Pain: psychological pain.
[f]NU: negative urgency impulsivity.
[g]Dft: defeat.
[h]Trap: entrapment.
[i]DT: distress tolerance.
[j]PB: perceived burdensomeness.
[k]TB: thwarted belongingness.
[l]Agg: aggression.

XSL•FO
RenderX

Figure 2. Computerized Adaptive Test of Suicide Risk Pathways domain information. Test information for each domain is divided by number of primary items in the item pool. SI: suicidal ideation.

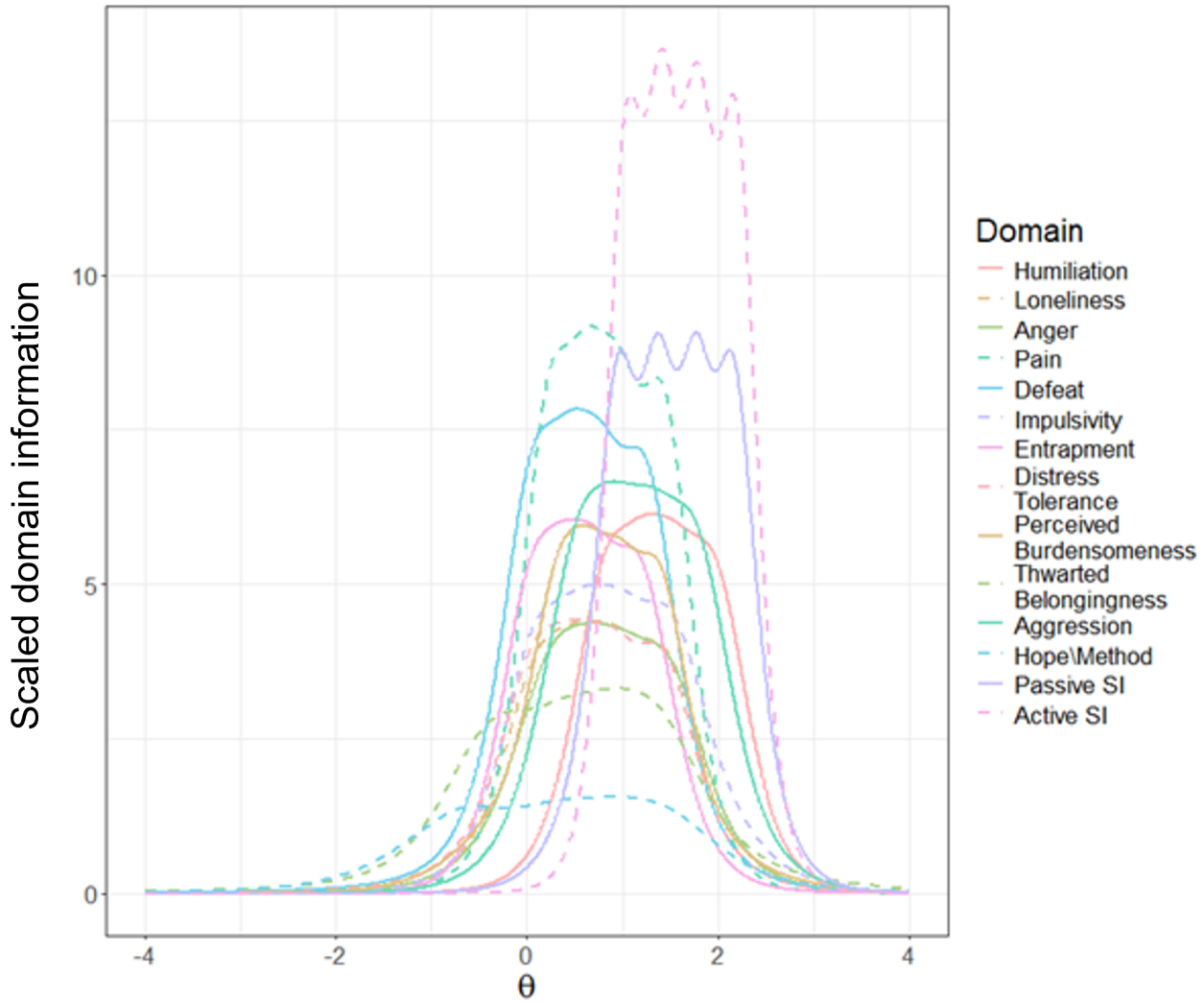

For the CAT-SRP suicidal ideation items, both the one-factor (root-mean-squared error of approximation [RMSEA]=0.048; Tucker Lewis Index [TLI]=0.991; Comparative Fit Index [CFI]=.0993) and 2-factor (RMSEA=0.053; TLI=0.989; CFI=0.991) graded response models exhibited good fit. As expected, passive and active suicidal ideation factors were highly correlated (r=0.92). Given the CAT-SRP's overarching aim to study dynamics of suicidal thoughts and behaviors and prior work supporting differences between passive and active suicidal ideation [45,58], the 2-factor model was selected to calibrate the suicidal ideation items. Item discrimination and difficulty parameters were extracted for use in adaptive testing. Domain information is shown in Figure 1B. The final CAT-SRP item bank consists of J=189 risk items across 12 risk domains and J=18 suicidal ideation items split equally across passive and active ideation.

CAT-SRP Risk Factor Dimensionality

Parallel analysis suggested M=13 factors (see Figure S2 in Multimedia Appendix 1). Since 12 test domains were originally proposed for the CAT-SRP risk variables, EIFA was fit for both 12 and 13 factors to evaluate the interpretability of the resulting factor structures. Factor loading patterns are provided in Tables S1 and S2 in Multimedia Appendix 1. Results suggested that "perceived stress" items did not load onto a unique factor in either the 12- or 13-factor model. Thus, stress items were removed from the item pool, and the reduced pool was reanalyzed using EIFA. Items with positively valenced prompts, primarily from the hopelessness pool, all loaded onto a separate factor in both models (ie, after reverse coding), which is referred to as the Hope factor.

A series of EIFA models (M=10 - 12) were fit to the reduced item pool; factor loading matrices are provided in Tables S3, S4, and S5 in Multimedia Appendix 1. The 12 model (Table S3 in Multimedia Appendix 1) was ultimately selected due to the interpretability of the factors, the lowest model BIC ($BIC_{M=12}$=9384.88, $BIC_{M=11}$=13967.56, $BIC_{M=10}$=18452.14), and conceptual alignment with the hypothesized blueprint. This resulted in J=189 risk domain items. A summary of the final CAT-SRP risk domains and associated items is presented in Table 4.

XSL•FO
RenderX

Table . Computerized Adaptive Test of Suicide Risk Pathways risk item domain summary.

Domain	Total Items[a] (Primary[b]), n (%)	Items
Humiliation	32 (32)	HUM[c] 1 - 32
Loneliness	30 (30)	INQ[d] 11, 12, UCLA[e] 1 - 20, TBS[f] 1 - 8
Anger	24 (22)	Anger[g] 1 - 22, Impulse[h] 11, Agg[i] 1
Pain	16 (13)	Pain[j] 1 - 13, TRAP[k] 3, 4, 7
Defeat	24 (14)	STHS[l] 1, 2, 4, 5, 8, 10, DS[m] 1 - 8, DS 10 - 16, BH.N[n] 1 - 3
Impulse (NU)	14 (11)	Impulse 1 - 10, 12, Agg2 – 4
Entrapment	16 (16)	TRAP 1 - 16
Distress tolerance	14 (14)	DT[o] 1 - 5, 7 - 15
Perceived burdensomeness	17 (14)	INQ 1 - 6, 11, 12, STHS 1, 2, 4, 5, 8, TBS 3, 7, Agg 2, 3
Thwarted belongingness	13 (7)	INQ 7 - 10, 13 - 15, TBS 3, 4, 8, DS 9, Impulse 11, DT 6
Aggression	6 (4)	Agg 1 - 4, HUM 1, 2
Hope	12 (12)	STHS 3, 6, 7, 9, DS 2, 4, 9, Impulse 11, DT 6, BH.P[p] 1 - 3

[a]Some items load onto multiple factors.

[b] Primary items refer to how many items had their stronges factor loading on each domain.

[c]HUM: Humiliation Inventory.

[d]INQ: Interpersonal Needs Questionnaire.

[e]UCLA: UCLA (University of California, Los Angeles) Loneliness Scale.

[f]TBS: Thwarted Belongingness Scale.

[g]Anger: PROMIS (Patient-Reported Outcomes Measurement Information System) Anger.

[h]Impulse: UPPS-P (Urgency, (Lack of) Premeditation, (Lack of) Perseverance, Sensation Seeking, and Positive Urgency) Negative Urgency.

[i]Agg: PROMIS Aggression.

[j]Pain: psychological pain.

[k]TRAP: entrapment.

[l]STHS: State-Trait Hopelessness Scale.

[m]DS: Defeat Scale.

[n]BH.N: Brief Hopelessness – Negative.

[o]DT: Distress Tolerance Scale.

[p]BH.P: Brief Hopelessness – Positive.

To provide descriptive statistics for raw domain scores, sum scores were computed for each risk domain; items that loaded onto multiple factors were only included in sum scores for the domain with their strongest loading. Descriptive statistics for these domain-level sum scores are provided in Table 5. Internal consistency ranged from α=.90 to α=.99.

XSL•FO
RenderX

Table . Computerized Adaptive Test of Suicide Risk Pathways domain sum score descriptive statistics.

Domain	Mean (SD)	Min[a]-Max[b]	α	ω_h	ω_{total}	ICC[c,d]
Humiliation	61.07 (35.81)	32 - 160	.99	0.90	0.99	0.49
Loneliness	67.48 (35.52)	30 - 150	.99	0.91	0.99	0.57
Anger	47.55 (24.21)	22 - 110	.97	0.89	0.98	0.39
Pain[e]	27.50 (16.13)	13 - 65	.98	0.93	0.98	0.58
Defeat	31.46 (17.20)	14 - 70	.98	0.94	0.98	0.55
Impulsivity	23.24 (12.17)	11 - 55	.96	0.90	0.96	0.49
Entrapment	35.76 (19.51)	16 - 80	.98	0.90	0.98	0.54
Distress Tolerance	32.02 (15.07)	14 - 70	.96	0.75	0.96	0.53
PB[f]	30.69 (15.65)	14 - 70	.96	0.85	0.96	0.63
TB[g]	16.58 (7.64)	7 - 35	.91	0.86	0.92	0.54
Aggression	7.06 (4.22)	4 - 20	.90	0.87	0.92	0.61
Hope	31.50 (11.04)	12 - 60	.90	0.77	0.91	0.56
Passive SI[h]	21.88 (12.32)	9 - 45	.98	0.93	0.98	0.55
Active SI	18.64 (11.612)	9 - 45	.98	0.94	0.98	0.47

[a]Min: minimum observed score.
[b]Max: maximum observed score.
[c]computed in study 2.
[d]ICC: intraclass correlation.
[e]Pain: psychological pain.
[f]PB: perceived burdensomeness.
[g]TB: thwarted belongingness.
[h]SI: suicidal ideation.

Study 2

Technical Issues

There were some technical difficulties with the web-based CAT-SRP system early in the study that led to some issues with data collection. At the beginning of data collection, there was an issue with the “warm-up” period, which led to respondents receiving the same first 14 items at the beginning of each survey (see also Figure 3B). Even though subsequent items were administered adaptively, this may have increased survey lengths early on. There was also an issue with the EMA survey alert system sent by the CAT-SRP, which failed to send additional future survey notifications if a participant did not respond to 2 consecutive surveys. This likely led to lower compliance rates for participants at the beginning of the study. Finally, the CAT-SRP did not explicitly record the number of survey nonresponses in the system. Instead, compliance was calculated as the ratio of surveys completed to the expected number of total surveys over the study period (n=126). This likely contributed to underestimation of compliance rates because of technical difficulties noted in the survey alert system.

XSL•FO
RenderX

Figure 3. Computerized Adaptive Test of Suicide Risk Pathways survey content by participant over study duration. Dashed line=end of "warm-up" period. CAT-SRP: Computerized Adaptive Test of Suicide Risk Pathways; SI: suicidal ideation.

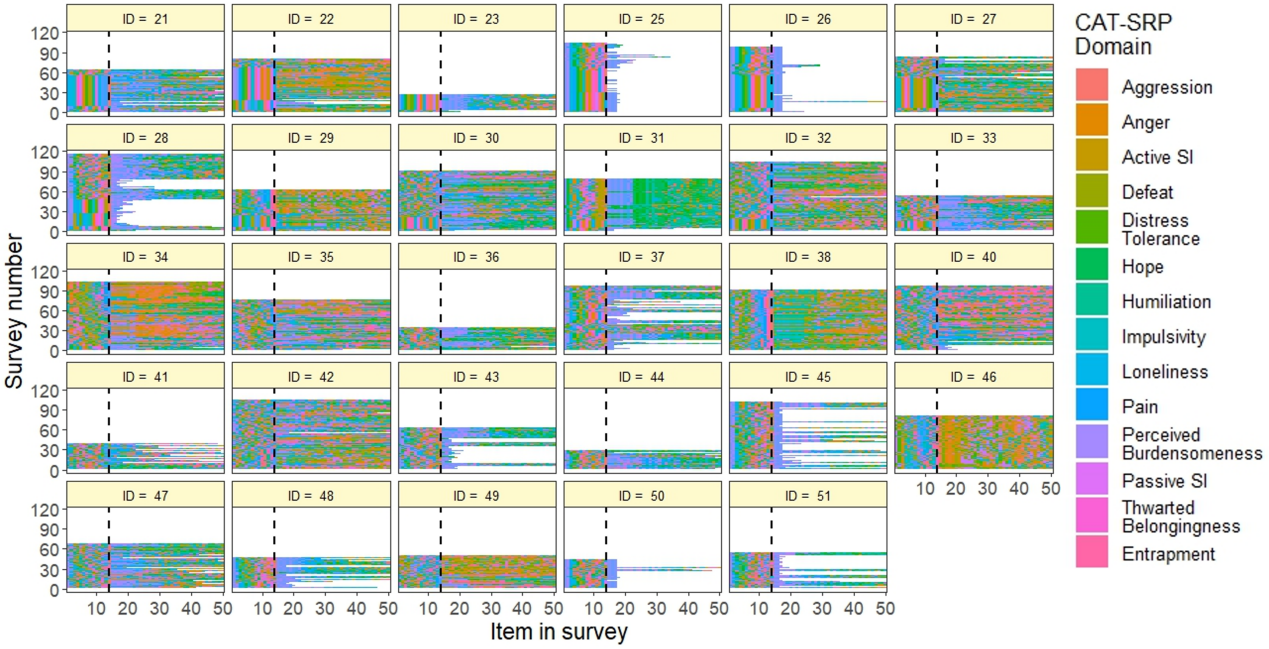

Compliance

Compliance was calculated as the proportion of surveys completed over the expected number of total surveys over the study period (n=126). Overall compliance rates ranged from 20.63% to 83.33% (mean 55.88%, SD 19.38%), resulting in 2134 unique observations across the 29 participants. Compliance rates across study weeks are shown in Table S6 in Multimedia Appendix 1. The ICC for the random-intercept–only model suggested notable within-person variability in compliance rates (ICC=0.53). Consistent with prior work [55], compliance decreased with weeks in the study (b=−0.10; *P*<.001).

Adaptive Survey Length

Survey length showed a clear bimodal distribution (Figure 1A) with a large peak at 50 items and a smaller peak at 17 items. The random-intercept–only model suggested roughly half of survey length variability was attributable to within-person factors (ICC=0.569). Figure 1B depicts person-level survey length over the duration of the study. Mixed-effect logistic regression analyses found that higher levels of humiliation (OR 0.72, 95% CI 0.59-0.87), impulsivity (ie, negative urgency; OR 0.81, 95% CI 0.69-0.96), aggression (OR 0.79, 95% CI 0.69-0.90), and active suicidal ideation (OR 0.61, 95% CI 0.46-0.80) were associated with reduced odds of surveys with maximum length; perceived burdensomeness levels (OR 1.12, 95% CI 1.01-1.25) increased the odds of maximum-length surveys (see Table 6).

XSL•FO
RenderX

Table . Bayesian logistic mixed-effects model of survey length.

	Estimate	SE	OR[a] (95% CI)[b]
Fixed effects			
Intercept	1.13	0.49	3.08 (1.16 to 7.95)
Humiliation	*–0.33* [c]	*0.09* [c]	*–0.72* [c] *(0.59 to 0.87* [c] *)*
Loneliness	0.10	0.08	–1.10 (0.95 to 1.28)
Anger	–0.12	0.09	–0.89 (0.75 to 1.06)
Pain[d]	0.08	0.12	1.08 (0.86 to 1.36)
Impulsivity	*–0.21* [c]	*0.08* [c]	*0.81* [c] *(0.69 to 0.96* [c] *)*
Defeat	0.07	0.10	1.07 (0.88 to 1.30)
Entrapment	–0.06	0.08	0.94 (0.81 to 1.10)
DT[e]	–0.16	0.10	0.85 (–0.36 to 0.03)
PB[f]	*0.12* [c]	*0.05* [c]	*1.12* [c] *(1.01 to 1.25* [c] *)*
TB[g]	0.12	0.10	1.13 (0.93 to 1.35)
Hope	–0.04	0.07	0.96 (0.84 to 1.11)
Aggression	*–0.23* [c]	*0.07* [c]	*0.79* [c] *(0.69 to 0.90* [c] *)*
Passive SI[h]	–0.21	0.14	0.81 (0.62 to 1.07)
Active SI	*–0.50* [c]	*0.14* [c]	*0.61* [c] *(0.46 to 0.* [c] *80)*
Surveys completed	0.03	0.02	1.03 (1.01 to 1.07)
Random effects			95% CI
SD (Intercept)	2.20	0.43	1.50 to 3.20
SD (SC)[i]	0.06	0.02	0.04 to 0.11
Cor[j] (Intercept, SC)	–0.09	0.24	-0.54 to 0.38

[a]OR: odds ratio.
[b]Bulk and tail effective sample sizes>1155.
[c]Credible interval for OR excludes 1.
[d]Pain: psychological pain.
[e]DT: distress tolerance.
[f]PB: perceived burdensomeness and hopelessness.
[g]TB: thwarted belongingness.
[h]SI: suicidal ideation.
[i]SC: surveys completed.
[j]Cor: correlation.

Adaptive Survey Content

After accounting for imbalanced item pool sizes, aggression, psychological pain, thwarted belongingness, passive suicidal ideation, and active suicidal ideation domains were administered more frequently (ie, compared to random administration); all other domains were administered less frequently. Figure 3 also suggests that the domain content of adaptive EMA surveys varied across participants (see also Figure S3 in Multimedia Appendix 1). The most and least frequently administered items from each domain are reported in Table 7. The most administered item (67.0% of surveys) was “Since the last prompt, my feelings of distress were so intense that they completely took over,” and the least administered item (2.8% of surveys) was “Since the last prompt, I felt isolated.” At the individual level, the most times an item was administered was 114 times, which corresponded to the item being present in 98.3% of that participant’s EMA surveys. Nine respondents had items that appeared in every survey (median 1, IQR 1-2). Conversely, the minimum number of times that an item was administered to an individual was one (median 20, IQR 7-24).

XSL•FO
RenderX

Table . Most and least common items by Computerized Adaptive Test of Suicide Risk Pathways domain.

Domain	Item prompts[a]	Values, n (%)[b]
Most common		
Active SI[c]	I wanted to die	1343 (62.9)
Aggression	I thought I may hit another person if I was provoked enough	1252 (58.7)
Anger	I had trouble controlling my anger	1251 (58.6)
Defeat	I felt that I have given up	960 (45)
DT[d]	My feelings of distress were so intense that they completely took over[h]	1429 (67)
Hope	I believe that I can help improve things	1158 (54.3)
Humiliation	I feared being ridiculed	997 (46.7)
Impulsivity	I often made worse because I acted without thinking when I was upset	1262 (59.1)
Loneliness	I felt completely alone	891 (41.8)
Pain[e]	I felt that my psychological pain was making me fall apart	989 (46.3)
Passive SI[c]	I thought that life was not worth living	1237 (58)
PB[f]	I felt that it was impossible to make progress on the goals I would like to strive for	1048 (49.1)
TB[g]	I interacted with people who care about me	1210 (56.7)
Entrapment	I felt that I was in a deep hole that I could not get out of	1012 (47.4)
Least common		
Active SI	I wanted to kill myself	617 (28.9)
Aggression	I have threatened people that I know	184 (8.62)
Anger	I held grudges toward others	278 (13)
Defeat	I felt that I sunk to the bottom of the ladder	323 (15.1)
DT	When I felt distressed or upset, I could not help but concentrate on how bad the distress actually feels	178 (8.34)
Hope	The future seems to be hopeful for me	216 (10.1)
Humiliation	I was concerned about being treated as invisible	141 (6.61)
Impulsivity	I got involved with things that I later wished I could get out of	187 (8.76)
Loneliness	I felt isolated[i]	59 (2.76)
Pain	I felt psychological pain	268 (12.6)
Passive SI	I wished I could go to sleep and never wake up	502 (23.5)
PB	I thought I make things worse for the people in my life	165 (7.73)
TB	I felt like I belong	363 (17)
Entrapment	I wanted to escape from my thoughts and feelings	105 (4.92)

[a]All items prompts preceded by “Since the last survey, …”.

[b]percentage refers to the percentage of EMA surveys which contained that item.

[c]SI: suicidal ideation.

[d]DT: distress tolerance.

[e]Pain: psychological pain.

XSL•FO
RenderX

[f] PB: perceived burdensomeness.

[g]TB: thwarted belongingness.

[h]Most frequently administered item.

[i]Least frequently administered item.

Average pairwise Jaccard Indices and overlap coefficients showed low to moderate overlap (median 0.35, IQR 0.27-0.45; median 0.37, IQR 0.31-0.49 respectively). Within-person average pairwise Jaccard Indices and overlap coefficients for each participant (ie, averaged over EMA surveys) ranged across participants (Jaccard Index=0.20 - 0.77; overlap coefficient=0.25 - 0.86); full summary data are provided in Table S7 in Multimedia Appendix 1. Thus, the adaptive nature of the CAT-SRP reduces the repetitiveness of EMA survey content relative to traditional EMA.

CAT-SRP Within- and Between-Person Structures

Within- and between-person correlation matrices for the CAT-SRP domain scores over the EMA period are shown in Figure 4. In general, CAT-SRP domains were highly positively correlated between persons, except for the hope domain, which was negatively correlated, and the thwarted belongingness domain, which was uncorrelated with other domains. Within-person domains were moderately or weakly positively correlated with hope and thwarted belongingness, demonstrating the weakest correlations.

Figure 4. Computerized Adaptive Test of Suicide Risk Pathways domain within- and between-person correlation matrix. The upper triangle denotes within-person correlation, and the lower triangle denotes between-person correlation. Agg: aggression; ASI: active suicidal ideation; CAT-SRP: Computerized Adaptive Test of Suicide Risk Pathways; DT: distress tolerance; Hum: humiliation; Impulse: impulsivity; Pain: psychological pain; PB: perceived burdensomeness; PSI: passive suicidal ideation; TB: thwarted belongingness; Trap: eEntrapment.

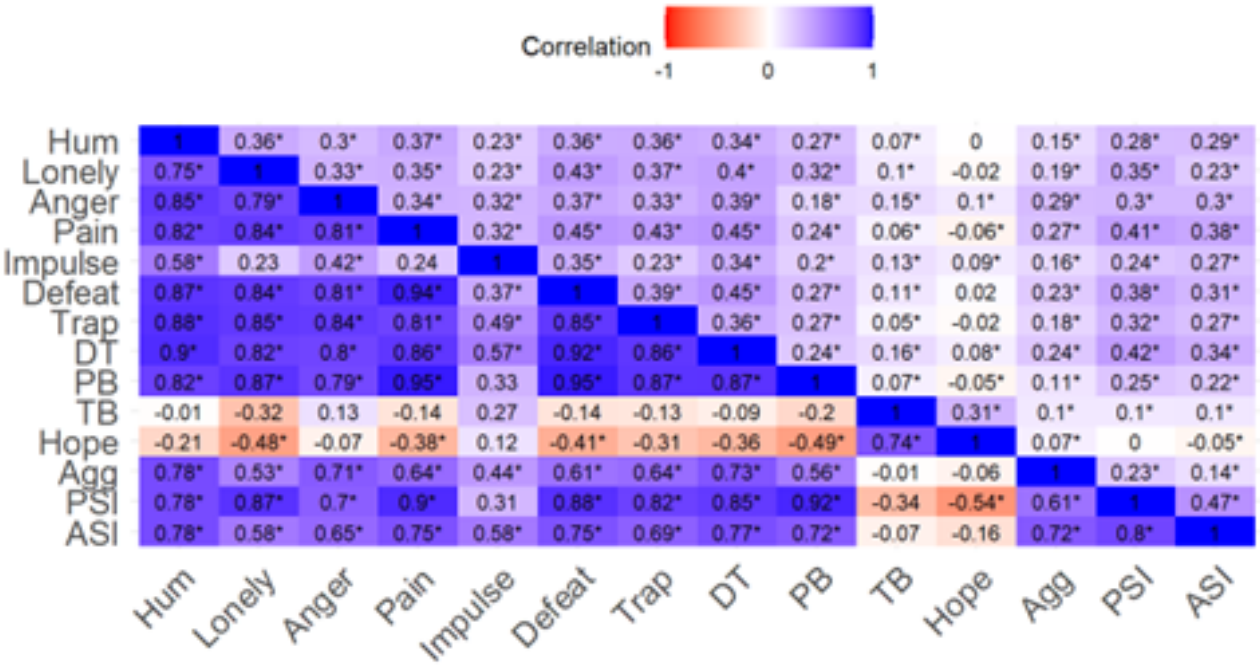

Visual inspection of the CAT-SRP item response polychoric correlation matrices shows high positive between-person interitem correlations with items correlating higher within passive or active ideation domains and moderately between domains (Figure 5). Within-person polychoric correlations are lower in magnitude but appear to show a similar structure. To further explore these structures, we conducted parallel analysis, which suggested a 2-factor between-person and 1-factor within-person structure. Overall, however, between-person and within-person structures should be interpreted with caution given the small sample size.

XSL•FO
RenderX

Figure 5. Computerized Adaptive Test of Suicide Risk Pathways item within- and between-person correlation matrix. The upper triangle denotes within-person correlation, and the lower triangle denotes between-person correlation.

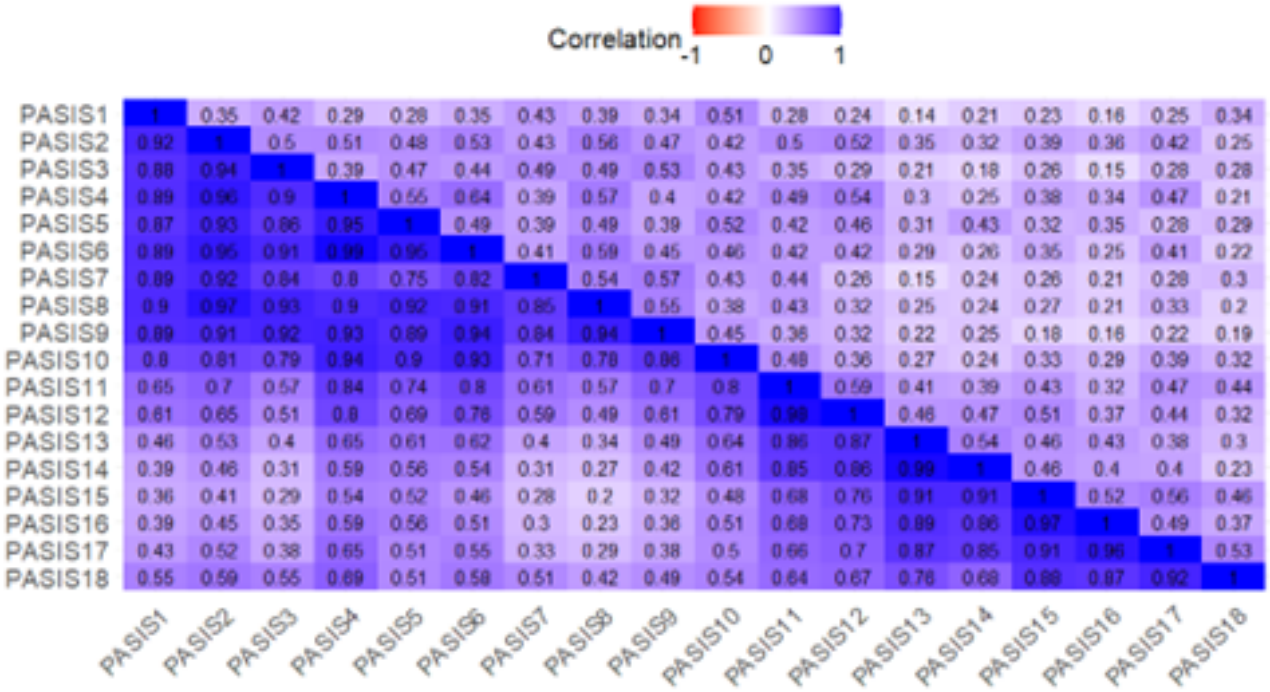

Risk Factor Domain Associations

ICCs for risk factors ranged from 0.39 for anger to 0.63 for perceived burdensomeness, meaning that the between- versus within-person variability was nontrivial and roughly similar across each of the latent variables (Table 5). Relationships between risk factors and suicidal ideation were assessed concurrently (ie, same survey) and prospectively (ie, next survey; see Table 8). CAT-SRP risk domains explained 52.01% of concurrent passive suicidal ideation variance (R^2=.521) with fixed effects of anger, psychological pain, defeat, entrapment, and perceived burdensomeness. Additionally, psychological pain, distress tolerance, and passive suicidal ideation demonstrated fixed effects on prospective passive suicidal ideation (R^2=.301; R^2=.184 when autoregressive effect is excluded). CAT-SRP risk domains also explained concurrent active suicidal ideation (R^2=.392) with humiliation, anger, psychological pain, negative urgency impulsivity, defeat, distress tolerance, perceived burdensomeness, thwarted belongingness, and the method factor (ie, hope) demonstrating fixed effects. Prospective active suicidal ideation (R^2=0.16; R^2=0.093 when autoregressive effect is excluded) was associated with the method factor, distress tolerance, loneliness, and active suicidal ideation.

XSL•FO
RenderX

Table . Concurrent and prospective SI[a] model results.

Concurrent models	Passive SI				Active SI			
Predictor	B[b]	SE	*t* test (*df*)	*P* value	B[b]	SE	*t* test (*df*)	*P* value
Intercept	0.34	0.11	2.96 (29.05)	.006	0.18	0.13	1.41 (30.57)	.17
Humiliation	0.02	0.02	0.95 (2130.69)	.34	*0.07*	*0.02*	*3.95 (2127.52)*	*<.001*
Loneliness	*0.07* [c]	*0.02*	*4.36 (2133.67)*	*<.001*	–0.02	0.02	–1.52 (2133.45))	.12
Anger	*0.03*	*0.02*	*1.97 (2124.38)*	*.05*	*0.08*	*0.02*	*4.81 (2121.49)*	*<.001*
Pain	*0.15*	*0.02*	*7.50 (2131.28)*	*<.001*	*0.15*	*0.02*	*7.48 (2133.99)*	*<.001*
Impulsivity	0.01	0.02	0.63 (2031.07)	.53	*0.06*	*0.02*	*3.95 (2090.20)*	*<.001*
Defeat	*0.08*	*0.02*	*4.19 (2122.25)*	*<.001*	*0.05*	*0.02*	*2.45 (2119.73)*	*.02*
Entrapment	*0.04*	*0.02*	*2.44 (2131.51)*	*.01*	0.02	0.02	1.36 (2128.32)	.18
DT[d]	*0.14*	*0.02*	*7.97 (2128.69)*	*<.001*	*0.09*	*0.02*	*5.08 (2125.53)*	*<.001*
PB[e]	*0.05*	*0.01*	*3.76 (2119.83)*	*<.001*	*0.04*	*0.01*	*3.22 (2130.67)*	*0.001*
TB[f]	0.01	0.02	0.65 (2082.87)	.51	*0.05*	*0.02*	*2.72 (2115.04)*	*0.007*
Hope	–0.02	0.01	–1.33 (2109.46)	.18	*–0.06*	*0.01*	*–3.88 (2127.05)*	*<.001*
Aggression	*0.05*	*0.02*	*3.27 (2029.59)*	*.001*	–0.01	0.02	–0.42 (2088.77)	.68
Prospective Models								
Intercept	0.25	0.12	2.02 (25.86)	.05	0.18	0.13	1.34 (28.04)	.19
Lagged SI[g]	*0.32*	*0.02*	*13.36 (2104.97)*	*<.001*	*0.25*	*0.02*	*10.60 (2104.58)*	*<.001*
Humiliation	–0.01	0.02	–0.78 (2101.47)	.44	0.02	0.02	0.83 (2098.99)	.41
Loneliness	0.01	0.02	0.34 (2104.77)	.74	*–0.05*	*0.02*	*–2.66 (2104.98)*	*.008*
Anger	0.03	0.02	1.78 (2094.79)	.07	0.02	0.02	1.08 (2094.71)	.28
Pain	*0.05*	*0.02*	*2.14 (2103.81)*	*.03*	0.04	0.02	1.79 (2104.94)	.07
Impulsivity	–0.03	0.02	–1.62 (1981.51)	.10	–0.01	0.02	–0.30 (2040.26)	.76
Defeat	–0.02	0.02	–0.92 (2091.33)	.35	–0.00	0.02	–0.00 (2090.73)	>.99
Entrapment	0.02	0.02	1.18 (2101.63)	.24	–0.03	0.02	–1.40 (2100.56)	.16
DT[d]	*0.04*	*0.02*	*2.06 (2098.65)*	*.04*	*0.08*	*0.02*	*3.95 (2097.25)*	*<.001*
PB[e]	0.02	0.02	1.16 (2089.18)	.25	–0.00	0.02	–0.11 (2096.92)	.91
TB[f]	–0.03	0.02	–1.67 (2039.45)	.09	–0.00	0.02	–0.15 (2071.71)	.88
Hope	–0.02	0.02	–1.09 (2076.12)	.27	*–0.05*	*0.02*	*–3.19 (2089.90)*	*.001*
Aggression	0.03	0.02	1.72 (1983.10)	.09	0.01	0.02	0.58 (2035.33)	.56

[a]SI: suicidal ideation.

[b]Fixed effect regression coefficients.

[c]Italicized values indicate *P*<.05.

[d]DT: distress tolerance.

XSL•FO
RenderX

[e]PB: perceived burdensomeness.

[f]TB: thwarted belongingness.

[g]Lagged SI: auto-regressive effect of passive or active suicidal ideation.

Subjective Burden

Complete descriptive statistics for subjective burden items are provided in Table 2. Over 65% of participants reported little to no feeling of being overwhelmed with the number of survey questions. Over 75% of participants also reported little to no frustration or annoyance due to the number of survey questions. Additionally, 67.9% of participants indicated that they never skipped surveys because of the survey length, with 25% indicating that survey length contributed to them skipping occasional surveys (ie, less than one per week).

Discussion

Study 1

Study 1 Background

The present study calibrated the CAT-SRP, a multidimensional CAT for measuring suicide risk domains and suicidal ideation in intensive longitudinal designs. Initial risk domains were identified by literature review and consultation with experts. Exploratory factor analyses were conducted to empirically evaluate the dimensionality of the CAT-SRP risk item bank. Results suggested 12 risk domains, which, while largely consistent with those identified in theoretical models [2-4], somewhat deviated from the initial measure blueprint. Notable deviations from the anticipated domain structure included (1) aggression items split from anger items into a separate factor; (2) hopelessness items loaded onto defeat and PB factors rather than as a separate factor; and (3) a method factor for positively valenced item prompts emerged. This method factor was present in all models examined and was primarily indexed by items reflecting hope (ie, rather than hopelessness). Research suggests that hope and hopelessness are different constructs [59,60] and which have demonstrated different relationships regarding suicide risk [61] potentially explaining some of the above deviations. Additionally, after inspection of the item prompts, the TB domain captured in the CAT-SRP appears to reflect the absence of reciprocal care, while the loneliness domain captures social isolation. As anticipated, the CAT-SRP suicidal ideation domains include passive and active suicidal ideation.

Information curves suggest that most CAT-SRP items were best at capturing average to above-average levels of the corresponding domain severity (Figure 1). This was also the case for suicidal ideation items. Thus, CAT-SRP scores for these domains can be expected to effectively differentiate between respondents with elevated risk severity but less effectively differentiate between respondents exhibiting lower levels of risk. Exceptions to this include the humiliation, anger, and aggression domains, which exhibit more reliable measurement at below-average levels. CAT-SRP domains also differed in the number of items present in the item bank. Although most domains contained between 10 - 20 items, the aggression domain only consisted of 6 items (4 primary). Conversely, the humiliation and loneliness domains consisted of 32 and 30 items, respectively. Domains with deeper item pools are more likely to demonstrate increased measurement precision across a wider range of severity levels and may result in less repetitive item administration in EMA contexts. Future work may seek to add items that broaden the domain coverage of the CAT-SRP item bank. Importantly, however, these findings suggest that the performance of suicide risk measures in intensive longitudinal data differs depending on the underlying level of risk severity.

Study 1 Limitations

Findings should be interpreted in the context of study limitations. First, although the CAT-SRP is intended for use in intensive longitudinal designs, the present calibration sample was collected using a large cross-sectional sample with day-level temporal referents. This approach was selected due to the large sample size needed to obtain sufficiently accurate MGRM item parameter estimates [50,62]. This precluded the evaluation of longitudinal measurement invariance. Additionally, item parameters obtained in this calibration study are for items with a day-level temporal referent. Thus, they are likely directly applicable to daily diary research. However, additional work is needed to determine if more granular temporal referents (eg, “Since the last prompt”) substantially impact item parameters. Future work should also seek to validate the 12-dimensional CAT-SRP structure observed in the current study.

It is also important to note that the sample used in this calibration study was a general adult sample collected through Prime Panels. Although prior research has supported the use of Prime Panels [63] and multiple efforts were taken to safeguard data quality, other work has highlighted potential pitfalls of online crowdsourcing data collection platforms [64]. Additionally, although only 36.33% of the calibration sample reported a lifetime history of suicidal thoughts and behavior, risk domains assessed by the CAT-SRP are applicable to individuals with or without a history of suicide risk. Further, to study the development of these risk pathways, it is necessary to measure risk domains among those who have yet to exhibit suicidal thoughts or behaviors. Finally, it is important to note that risk domains and items were selected based on predominant ideation-to-action theoretical models (ie, ITS, 3ST, IMV) and expert consultation given the preliminary nature of this study and not in direct collaboration with individuals with lived experiences. Future work will aim to refine and expand the CAT-SRP item bank to include input from those with lived experiences.

Study 1 Conclusions

The present study developed and calibrated the CAT-SRP in a large community sample, establishing a comprehensive measurement system that captures 12 empirically derived domains of suicide risk, along with passive and active suicidal ideation. The domain structure largely aligns with theoretical models while revealing some notable refinements, such as the distinction between anger and aggression domains, the

XSL•FO
RenderX

distribution of hopelessness across multiple factors, and the emergence of a potential protective factor. Information analyses suggest the CAT-SRP provides optimal measurement precision at different severity levels across domains, potentially allowing for a more nuanced assessment of risk pathways. While the cross-sectional calibration provides a foundation for adaptive measurement, Study 2 represents a critical next step in evaluating the CAT-SRP's utility for intensive longitudinal assessment in clinical populations.

Study 2

Study 2 Background

The present pilot study implemented and evaluated an adaptive ecological momentary assessment protocol using the CAT-SRP among individuals with past-month suicidal thoughts or behaviors. Although prior work has shown promise for CAT in patient-reported outcomes measurement [22], this represents the first application of CAT for intensive longitudinal assessment of suicide risk factors and suicidal ideation and the first application of MCAT in EMA. Results support the feasibility and acceptability of the CAT-SRP for real-time assessment while providing rich information about the dynamics of suicide risk factors and ideation in daily life. One key finding was that the CAT-SRP captured both between- and within-person variability in risk factors and suicidal ideation. Intraclass correlations indicated that roughly half of the variance in most constructs was attributable to within-person fluctuations, supporting the dynamic nature of these processes. The adaptive testing algorithm showed variability in survey length and content across participants and occasions, suggesting personalized assessment. Although the overall surveys only showed moderate overlap on average, nine participants did receive a small subset of items (eg, a single item) in every survey. This was likely due to a lack of variability in item responses for a given domain. Future implementations of CAT in EMA designs that are concerned with survey repetitiveness may wish to directly incorporate content balancing in their selection algorithm [65]. Additionally, future versions of the CAT-SRP would benefit from directly tracking participant-level compliance.

Initial analysis of the relationships between CAT-SRP risk domain scores with passive and active suicidal ideation was also promising. Analyses of relationships between risk factors and passive ideation at the same measurement occasion (ie, concurrent effects) found effects for 8 of 12 risk factors; 9 of the 12 risk factors were found to relate to concurrent active ideation. Additionally, CAT-SRP risk domains accounted for 52% and 39% of variability in concurrent passive and active suicidal ideation domain scores and 18% and 9% of variability in prospective ideation, suggesting promising concurrent and predictive validity of CAT-SRP risk factor domain scores. Although future research is needed, these preliminary findings suggest that adaptive assessment of suicidal ideation and associated risk factors may alleviate measurement challenges common in EMA suicide research (ie, zero-inflation; see Figure 6) and enhance signal detection [14].

Figure 6. Distribution of active and passive suicidal ideation. The left plot is the density plot from the current study, while the right plot is a histogram from a similar study that used the sum of 2 items for both passive and active suicidal ideation [66].

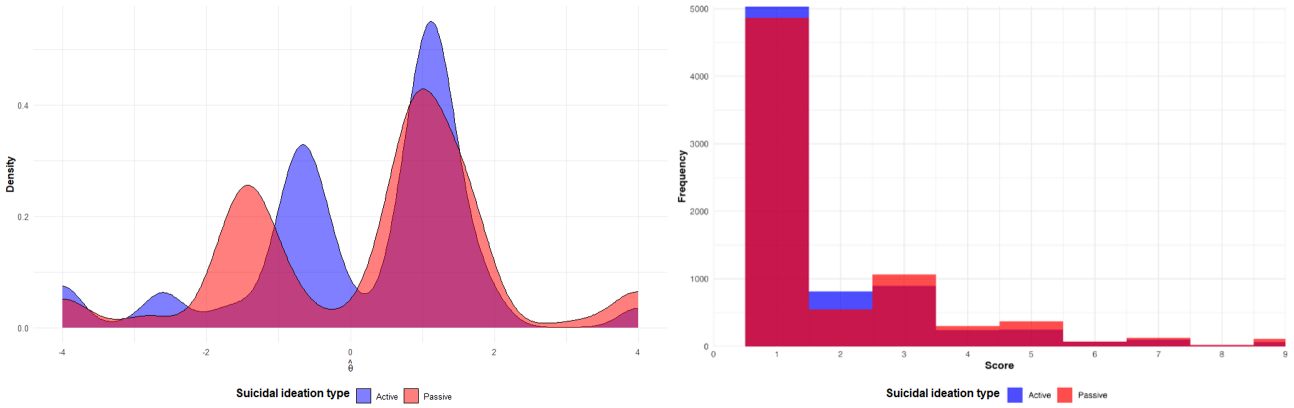

The implementation of adaptive assessment in this context has several important implications for advancing the study of suicide risk. First, the ability to efficiently measure multiple risk domains while maintaining measurement precision addresses a critical limitation of traditional EMA approaches that often rely on abbreviated measures. Second, assessing this breadth of risk factors has important implications for advancing suicide theory. Importantly, future research should consider pairing CAT-SRP with methods that also vary the timing of assessments, such as including more intensive windows of assessments [6]. When combined with recent developments in continuous time models, this approach can enhance our understanding of both the key risk factors and the timescales at which they exert the greatest influence [67,68].

Advancing theoretical understanding of clinical phenomena requires sophisticated, dimensional assessment approaches. While the current study examined 12 risk factors, this represents only a subset of potentially relevant constructs, particularly when considering the extant body of suicide research and when compared to comprehensive frameworks like the Hierarchical Taxonomy of Psychopathology [69]. Additionally, to fully capture the complexity of suicide risk, it is essential to incorporate protective factors [70], which the CAT-SRP currently lacks. This underscores a fundamental tension in EMA research: balancing comprehensive assessment with practical constraints and response burden. While the CAT-SRP aims to mitigate this issue, future research could address this challenge through innovative design approaches, such as planned missingness at the construct level or dynamic selection of

XSL•FO
RenderX

relevant constructs based on individual risk profiles and temporal patterns. These methods could expand the breadth of assessment while maintaining feasibility for intensive longitudinal designs.

Study 2 Limitations

Although the current study highlights the potential benefits of the CAT-SRP for assessing suicidal ideation and associated risk factors in EMA, these findings should be interpreted in the context of the study's limitations. First, it is important to reiterate that study 2 was a pilot study consisting of only 29 respondents, thus requiring replication in a larger sample. Additionally, the CAT-SRP used "since the last prompt" as the temporal referent for all items, which differed from the calibration referent ("Today"). This referent was chosen because of its suitability for EMA designs while remaining more consistent with the calibration referent than other referents common to EMA designs (eg, "right now"). Given heterogeneity in temporal dynamics in suicidal thinking and related processes [6], future work is likely needed to evaluate the potential impact of this choice on the risk factors assessed by the CAT-SRP.

Technical difficulties with the CAT-SRP web-based survey system study also introduced challenges in assessing study compliance. Compliance rates were calculated as the number of complete surveys over the number of expected surveys throughout the study period. However, technical difficulties early in the study led to (1) failure to send additional survey notifications if a participant did not respond to 2 consecutive surveys and (2) failure to directly record nonresponses in the CAT-SRP platform. Thus, compliance rates reported in the current study are likely underestimated. This initial version of the CAT-SRP was also unable to tailor the daily EMA window to the participant (ie, all respondents received surveys between 9 AM EST and 9 PM EST). Future versions of the CAT-SRP will address these limitations, which will likely bolster compliance rates.

Finally, while the CAT-SRP appeared to successfully tailor survey content, most surveys reached the maximum allowed length (ie, 50 items). Although these surveys were longer than some previous studies [58], the compliance rates were similar [55], and participants reported low levels of subjective burden. Thus, even for surveys that reach maximum length, there may be benefits to reducing the repetitiveness of EMA survey content. Since CAT tailors survey items to respondents based on estimated underlying levels of risk factors, the content validity of CAT-SRP surveys is hypothesized to be higher than static measures [71]; however, an interesting area for future work would be to examine the content validity of CAT-SRP item content for participants with different longitudinal risk profiles. Further methodological research is also needed to determine optimal stopping rules and item selection procedures that balance measurement precision with participant burden for intensive longitudinal designs. Some simulation work has suggested that sequential application of unidimensional stopping rules may increase efficiency in intensive longitudinal designs [53]. Additionally, alternative stopping rules, which use predicted SE reduction [72], may also increase precision with minimal impact on measurement precision [73]. Finally, it is important to note that since the CAT-SRP relies on a precalibrated item bank, measurement invariance is assumed for the duration of the EMA period. Further methodological research is needed to relax this assumption and account for the possibility that item parameters change over repeated exposure to the item (ie, item parameter drift) in adaptive EMA. This is particularly important for domains where repeatedly responding to certain items may influence the underlying relationship between items and domain (eg, reactivity [74]).

General Conclusions

This study calibrated and piloted the CAT-SRP, the first multidimensional adaptive assessment of measuring suicidal thoughts and risk factors in EMA designs. While motivated by ideation-to-action theoretical models, the structure of risk factors was empirically derived in a large cross-sectional sample. Findings suggested that hopelessness items split across defeat and perceived burdensomeness domains rather than emerging as a distinct factor, potentially indicating that hopelessness may be better understood as an element of broader negative self-evaluative processes. Additionally, a method factor composed primarily of positively valenced items also raised questions about potential mixing of risk and protective constructs in existing assessment. In total, 12 risk factors were included in the CAT-SRP alongside passive and active suicidal ideation domains.

A pilot EMA study provided preliminary validation of the CAT-SRP in intensive longitudinal data. Results, although preliminary, suggest that the CAT-SRP produces dynamic survey content (ie, different items) over the course of an EMA study while maintaining measurement precision. This marks the first application of CAT in suicide EMA research and the first application of MCAT to intensive longitudinal designs and opens new possibilities for research design. The different patterns of measurement precision across domains—with some providing better discrimination at low levels and others at high levels—suggest the potential for developing targeted assessment strategies based on individual risk profiles. Future research could explore dynamically adapting assessment frequency and domain coverage, integrating passive sensing data, or linking adaptive measurement methodologies with just-in-time adaptive interventions.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

XSL•FO
RenderX

Data Availability

The CAT-SRP item bank and the original CAT-SRP blueprint are available [75]. Data from study 1 are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Repetition to relevance.
[DOCX File, 637 KB - formative_v9i1e76544_app1.docx]

References

1. Garnett MF, Curtin SC. Suicide mortality in the United States, 2002–2022. 2024.
2. Joiner T. Why People Die by Suicide: Harvard University Press; 2005.
3. Klonsky ED, May AM. The Three-Step Theory (3ST): a new theory of suicide rooted in the “Ideation-to-Action” framework. Int J Cogn Ther 2015 Jun;8(2):114-129. [doi: 10.1521/ijct.2015.8.2.114]
4. O’Connor RC, Kirtley OJ. The integrated motivational-volitional model of suicidal behaviour. Philos Trans R Soc Lond B Biol Sci 2018 Sep 5;373(1754):20170268. [doi: 10.1098/rstb.2017.0268] [Medline: 30012735]
5. Chu C, Buchman-Schmitt JM, Stanley IH, et al. The interpersonal theory of suicide: a systematic review and meta-analysis of a decade of cross-national research. Psychol Bull 2017 Dec;143(12):1313-1345. [doi: 10.1037/bul0000123] [Medline: 29072480]
6. Coppersmith DDL, Ryan O, Fortgang RG, Millner AJ, Kleiman EM, Nock MK. Mapping the timescale of suicidal thinking. Proc Natl Acad Sci U S A 2023 Apr 25;120(17):e2215434120. [doi: 10.1073/pnas.2215434120] [Medline: 37071683]
7. Millner AJ, Lee MD, Hoyt K, Buckholtz JW, Auerbach RP, Nock MK. Are suicide attempters more impulsive than suicide ideators? Gen Hosp Psychiatry 2020;63:103-110. [doi: 10.1016/j.genhosppsych.2018.08.002] [Medline: 30097321]
8. Kivelä L, van der Does WAJ, Riese H, Antypa N. Don’t miss the moment: a systematic review of ecological momentary assessment in suicide research. Front Digit Health 2022;4:876595. [doi: 10.3389/fdgth.2022.876595] [Medline: 35601888]
9. Sedano-Capdevila A, Porras-Segovia A, Bello HJ, Baca-García E, Barrigon ML. Use of ecological momentary assessment to study suicidal thoughts and behavior: a systematic review. Curr Psychiatry Rep 2021 May 18;23(7):41. [doi: 10.1007/s11920-021-01255-7] [Medline: 34003405]
10. Lord F, Novick M, Birnbaum A. Statistical theories of mental test scores. 1968.
11. Ram N, Brinberg M, Pincus AL, Conroy DE. The questionable ecological validity of ecological momentary assessment: considerations for design and analysis. Res Hum Dev 2017;14(3):253-270. [doi: 10.1080/15427609.2017.1340052] [Medline: 30613195]
12. Bolger N, Stadler G, Laurenceau JP. Power analysis for intensive longitudinal studies. In: Handbook of Research Methods for Studying Daily Life: The Guilford Press; 2012:285-301.
13. Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. Measurement Error in Nonlinear Models: A Modern Perspective, 2nd edition: Chapman and Hall/CRC; 2006. [doi: 10.1201/9781420010138]
14. Jacobucci R, Grimm KJ. Machine learning and psychological research: the unexplored effect of measurement. Perspect Psychol Sci 2020 May;15(3):809-816. [doi: 10.1177/1745691620902467] [Medline: 32348703]
15. Weiss DJ, Kingsbury GG. Application of computerized adaptive testing to educational problems. J Educational Measurement 1984 Dec;21(4):361-375 [FREE Full text] [doi: 10.1111/j.1745-3984.1984.tb01040.x]
16. Gibbons CJ. Turning the page on pen-and-paper questionnaires: combining ecological momentary assessment and computer adaptive testing to transform psychological assessment in the 21st Century. Front Psychol 2016;7:1933. [doi: 10.3389/fpsyg.2016.01933] [Medline: 28154540]
17. Bock RD, Gibbons R, Muraki E. Full-information item factor analysis. Appl Psychol Meas 1988 Sep;12(3):261-280. [doi: 10.1177/014662168801200305]
18. Weiss DJ. Better data from better measurements using computerized adaptive testing. JMMSS 2011;2(1):1. [doi: 10.2458/jmm.v2i1.12351]
19. Brent DA, Horowitz LM, Grupp-Phelan J, et al. Prediction of suicide attempts and suicide-related events among adolescents seen in emergency departments. JAMA Netw Open 2023 Feb 1;6(2):e2255986. [doi: 10.1001/jamanetworkopen.2022.55986] [Medline: 36790810]
20. King CA, Brent D, Grupp-Phelan J, et al. Prospective development and validation of the computerized adaptive screen for suicidal youth. JAMA Psychiatry 2021 May 1;78(5):540-549. [doi: 10.1001/jamapsychiatry.2020.4576] [Medline: 33533908]
21. Gibbons RD, Kupfer D, Frank E, Moore T, Beiser DG, Boudreaux ED. Development of a computerized adaptive test suicide scale-the CAT-SS. J Clin Psychiatry 2017;78(9):1376-1382. [doi: 10.4088/JCP.16m10922] [Medline: 28493655]

XSL•FO
RenderX

22. Harrison C, Trickett R, Wormald J, et al. Remote symptom monitoring with ecological momentary computerized adaptive testing: pilot cohort study of a platform for frequent, low-burden, and personalized patient-reported outcome measures. J Med Internet Res 2023 Sep 14;25:e47179. [doi: 10.2196/47179] [Medline: 37707947]
23. Muraki E, Carlson JE. Full-Information factor analysis for polytomous item responses. Appl Psychol Meas 1995 Mar;19(1):73-90. [doi: 10.1177/014662169501900109]
24. Gee BL, Han J, Benassi H, Batterham PJ. Suicidal thoughts, suicidal behaviours and self-harm in daily life: A systematic review of ecological momentary assessment studies. Digit Health 2020;6:2055207620963958. [doi: 10.1177/2055207620963958] [Medline: 33224516]
25. Wang C, Weiss DJ, Shang Z. Variable-length stopping rules for multidimensional computerized adaptive testing. Psychometrika 2019 Sep;84(3):749-771. [doi: 10.1007/s11336-018-9644-7] [Medline: 30511327]
26. Bentley KH, Maimone JS, Kilbury EN, et al. Practices for monitoring and responding to incoming data on self-injurious thoughts and behaviors in intensive longitudinal studies: a systematic review. Clin Psychol Rev 2021 Dec;90:102098. [doi: 10.1016/j.cpr.2021.102098] [Medline: 34763126]
27. Nock MK, Kleiman EM, Abraham M, et al. Consensus statement on ethical & safety practices for conducting digital monitoring studies with people at risk of suicide and related behaviors. Psychiatr Res Clin Pract 2021;3(2):57-66. [doi: 10.1176/appi.prcp.20200029] [Medline: 34414359]
28. Chandler J, Rosenzweig C, Moss AJ, Robinson J, Litman L. Online panels in social science research: expanding sampling methods beyond Mechanical Turk. Behav Res 2019 Oct;51(5):2022-2038. [doi: 10.3758/s13428-019-01273-7]
29. Levy R. Distinguishing outcomes from indicators via Bayesian modeling. Psychol Methods 2017 Dec;22(4):632-648. [doi: 10.1037/met0000114] [Medline: 29265847]
30. Wright AGC, Zimmermann J. Applied ambulatory assessment: Integrating idiographic and nomothetic principles of measurement. Psychol Assess 2019;31(12):1467-1480. [doi: 10.1037/pas0000685]
31. McClure K, Ammerman B, Liu C, et al. From repetition to relevance: initial development of a multidimensional computerized adaptive test for intensive longitudinal assessment of suicide risk (CAT-SRP). OSF. URL: https://osf.io/zs2rb [accessed 2025-11-14]
32. Van Orden KA, Cukrowicz KC, Witte TK, Joiner TE. Thwarted belongingness and perceived burdensomeness: construct validity and psychometric properties of the Interpersonal Needs Questionnaire. Psychol Assess 2012 Mar;24(1):197-215. [doi: 10.1037/a0025358] [Medline: 21928908]
33. Bryan CJ. A preliminary validation study of two ultra-brief measures of suicide risk: the suicide and perceived burdensomeness visual analog scales. Suicide Life Threat Behav 2019 Apr;49(2):343-352. [doi: 10.1111/sltb.12447] [Medline: 29512832]
34. Ma J, Batterham PJ, Calear AL, Sunderland M. The development and validation of the Thwarted Belongingness Scale (TBS) for interpersonal suicide risk. J Psychopathol Behav Assess 2019 Sep;41(3):456-469. [doi: 10.1007/s10862-019-09721-6]
35. Russell DW. UCLA Loneliness Scale (Version 3): reliability, validity, and factor structure. J Pers Assess 1996 Feb;66(1):20-40. [doi: 10.1207/s15327752jpa6601_2] [Medline: 8576833]
36. Dunn SL, Olamijulo GB, Fuglseth HL, et al. The state-trait hopelessness scale: development and testing. West J Nurs Res 2014 Apr;36(4):552-570. [doi: 10.1177/0193945913507634] [Medline: 24122739]
37. Everson SA, Goldberg DE, Kaplan GA, et al. Hopelessness and risk of mortality and incidence of myocardial infarction and cancer. Psychosom Med 1996;58(2):113-121. [doi: 10.1097/00006842-199603000-00003] [Medline: 8849626]
38. Gilbert P, Allan S. The role of defeat and entrapment (arrested flight) in depression: an exploration of an evolutionary view. Psychol Med 1998 May;28(3):585-598. [doi: 10.1017/s0033291798006710] [Medline: 9626715]
39. Hartling LM, Luchetta T. Humiliation: assessing the impact of derision, degradation, and debasement. J Prim Prev 1999 Jun;19(4):259-278. [doi: 10.1023/A:1022622422521]
40. Holden RR, Mehta K, Cunningham EJ, McLeod LD. Development and preliminary validation of a scale of psychache. Canadian Journal of Behavioural Science / Revue canadienne des sciences du comportement 2001;33(4):224-232. [doi: 10.1037/h0087144]
41. Cella D, Riley W, Stone A, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. J Clin Epidemiol 2010 Nov;63(11):1179-1194. [doi: 10.1016/j.jclinepi.2010.04.011] [Medline: 20685078]
42. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. J Health Soc Behav 1983 Dec;24(4):385-396. [doi: 10.2307/2136404] [Medline: 6668417]
43. Lynam DR, Smith GT, Whiteside SP, Cyders MA. The UPPS-P: Assessing Five Personality Pathways to Impulsive Behavior: Purdue University; 2006.
44. Simons JS, Gaher RM. The Distress Tolerance Scale: development and validation of a self-report measure. Motiv Emot 2005 Jun;29(2):83-102. [doi: 10.1007/s11031-005-7955-3]
45. Liu RT, Bettis AH, Burke TA. Characterizing the phenomenology of passive suicidal ideation: a systematic review and meta-analysis of its prevalence, psychiatric comorbidity, correlates, and comparisons with active suicidal ideation. Psychol Med 2020 Feb;50(3):367-383. [doi: 10.1017/S003329171900391X] [Medline: 31907085]

XSL•FO
RenderX

46. Fox J. Polycor: Polychoric and Polyserial Correlations. 2022. URL: https://CRAN.R-project.org/package=polycor [accessed 2025-11-20]
47. Revelle W. Psych: procedures for psychological, psychometric, and personality research. The R Foundation.: Northwestern University; 2025. URL: https://CRAN.R-project.org/package=psych [accessed 2025-11-20]
48. Chalmers R. Mirt: A Multidimensional Item Response Theory Package for the R Environment. JSS Journal of Statistical Software 2012. [doi: 10.18637/jss.v048.i06]
49. Cai L. A two-tier full-information item factor analysis model with applications. Psychometrika 2010 Dec;75(4):581-612. [doi: 10.1007/s11336-010-9178-0]
50. McClure K, Jacobucci R. Item parameter calibration in the multidimensional graded response model with high dimensional tests. PsyArXiv. Preprint posted online on 2023 URL: https://osf.io/preprints/psyarxiv/cbemw/ [accessed 2024-08-08] [doi: 10.31234/osf.io/cbemw]
51. Therapeutic Risk Management (TRM) with Patients at Risk for Suicide. Therapeutic Risk Management – MIRECC / CoE. URL: https://www.mirecc.va.gov/visn19/trm/ [accessed 2025-09-09]
52. Mulder J, van der Linden WJ. Multidimensional adaptive testing with optimal design criteria for item selection. Psychometrika 2009 Jun;74(2):273-296. [doi: 10.1007/s11336-008-9097-5] [Medline: 20119511]
53. McClure KE. Efficient assessment of multidimensional processes in intensive longitudinal designs using adaptive ecological momentary assessment. : University of Notre Dame; 2024.
54. Newman E, Willard T, Sinclair R, Kaloupek D. Empirically supported ethical research practice: the costs and benefits of research from the participants' view. Account Res 2001;8(4):309-329. [doi: 10.1080/08989620108573983] [Medline: 12481796]
55. Jacobucci R, Ammerman BA, McClure K. Examining missingness at the momentary level in clinical research using ecological momentary assessment: implications for suicide research. J Clin Psychol 2024 Oct;80(10):2147-2162. [doi: 10.1002/jclp.23728] [Medline: 38943339]
56. Bürkner PC. brms: an R package for Bayesian multilevel models using stan. J Stat Softw 2017;80(1):1-28. [doi: 10.18637/jss.v080.i01]
57. Muthén LK, Muthén BO. Mplus user's guide (eighth). : Muthén & Muthén; 2017.
58. Hallensleben N, Glaesmer H, Forkmann T, et al. Predicting suicidal ideation by interpersonal variables, hopelessness and depression in real-time. An ecological momentary assessment study in psychiatric inpatients with depression. Eur psychiatr 2019;56(1):43-50. [doi: 10.1016/j.eurpsy.2018.11.003]
59. Beck AT, Weissman A, Lester D, Trexler L. The measurement of pessimism: the hopelessness scale. J Consult Clin Psychol 1974 Dec;42(6):861-865. [doi: 10.1037/h0037562] [Medline: 4436473]
60. Snyder CR, Harris C, Anderson JR, et al. The will and the ways: development and validation of an individual-differences measure of hope. J Pers Soc Psychol 1991;60(4):570-585. [doi: 10.1037/0022-3514.60.4.570]
61. Huen JMY, Ip BYT, Ho SMY, Yip PSF. Hope and hopelessness: the role of hope in buffering the impact of hopelessness on suicidal ideation. PLOS ONE 2015;10(6):e0130073. [doi: 10.1371/journal.pone.0130073] [Medline: 26107687]
62. Jiang S, Wang C, Weiss DJ. Sample size requirements for estimation of item parameters in the multidimensional graded response model. Front Psychol 2016;7:109. [doi: 10.3389/fpsyg.2016.00109] [Medline: 26903916]
63. Chandler J, Shapiro D. Conducting clinical research using crowdsourced convenience samples. Annu Rev Clin Psychol 2016;12(1):53-81. [doi: 10.1146/annurev-clinpsy-021815-093623] [Medline: 26772208]
64. Webb MA, Tangney JP. Too good to be true: bots and bad data from mechanical turk. Perspect Psychol Sci 2024 Nov;19(6):887-890. [doi: 10.1177/17456916221120027] [Medline: 36343213]
65. Cheng Y, Chang HH. The maximum priority index method for severely constrained item selection in computerized adaptive testing. Br J Math Stat Psychol 2009 May;62(Pt 2):369-383. [doi: 10.1348/000711008X304376] [Medline: 18534047]
66. Forkmann T, Spangenberg L, Rath D, et al. Assessing suicidality in real time: a psychometric evaluation of self-report items for the assessment of suicidal ideation and its proximal risk factors using ecological momentary assessments. J Abnorm Psychol 2018 Nov;127(8):758-769. [doi: 10.1037/abn0000381] [Medline: 30299116]
67. Asparouhov T, Muthén B. Continuous time dynamic structural equation models. 2024 URL: https://www.statmodel.com/download/CTRDSEM [accessed 2025-11-14]
68. Driver CC, Voelkle MC. Hierarchical Bayesian continuous time dynamic modeling. Psychol Methods 2018 Dec;23(4):774-799. [doi: 10.1037/met0000168] [Medline: 29595295]
69. Kotov R, Krueger RF, Watson D. A paradigm shift in psychiatric classification: the hierarchical taxonomy of psychopathology (HiTOP). World Psychiatry 2018 Feb;17(1):24-25. [doi: 10.1002/wps.20478] [Medline: 29352543]
70. Cramer RJ, Tucker R. Improving the field's understanding of suicide protective factors and translational suicide prevention initiatives. Int J Environ Res Public Health 2021 Jan 25;18(3):1027. [doi: 10.3390/ijerph18031027] [Medline: 33503803]
71. Smith GT, McCarthy DM, Anderson KG. On the sins of short-form development. Psychol Assess 2000;12(1):102-111. [doi: 10.1037/1040-3590.12.1.102]
72. Choi SW, Grady MW, Dodd BG. A new stopping rule for computerized adaptive testing. Educ Psychol Meas 2011 Feb;71(1):37-53. [doi: 10.1177/0013164410387338]

XSL•FO
RenderX

73. Yao L. Comparing the performance of five multidimensional CAT selection procedures with different stopping rules. Appl Psychol Meas 2013 Jan;37(1):3-23. [doi: 10.1177/0146621612455687]
74. Wray TB, Merrill JE, Monti PM. Using ecological momentary assessment (EMA) to assess situation-level predictors of alcohol use and alcohol-related consequences. Alcohol Res 2014;36(1):19-27. [Medline: 26258997]
75. From repetition to relevance. OSF. URL: https://osf.io/k6ers/overview?view_only=d437d4d7b57b4f059d6ee8bc3c3f16cb [accessed 2025-10-30]

Abbreviations

3ST: Three-Step Theory of Suicide
CAT: Computerized Adaptive Testing
CAT-SRP: Computerized Adaptive Test of Suicide Risk Pathways
CAT-SS: Computerized Adaptive Test-Suicide Scale
CFI: Comparative Fit Index
EIFA: exploratory item factor analysis
EMA: ecological momentary assessment
ICC: intraclass correlation coefficient
IMV: Integrated Motivational-Volitional
INQ: Interpersonal Needs Questionnaire
ITS: Interpersonal Theory of Suicide
MGRM: multidimensional graded response model
OR: odds ratios
PASIS: passive and active suicidal ideation scale
PB: perceived burdensomeness
PROMIS: Patient-Reported Outcomes Measurement Information System
RMSEA: root-mean-square error of approximation
STB: suicidal thoughts and behavior
TB: thwarted belongingness
TLI: Tucker Lewis Index
UCLA: University of California, Los Angeles
UPPS-P: Urgency, (Lack of) Premeditation, (Lack of) Perseverance, Sensation Seeking, and Positive Urgency

Edited by J Sarvestan; submitted 25.04.25; peer-reviewed by R Tucker, S Chowdhury; accepted 15.10.25; published 26.11.25.

Please cite as:
McClure K, Ammerman BA, Liu C, Blacutt M, O'Brien C, Jacobucci R
Initial Development of a Multidimensional Computerized Adaptive Test for Intensive Longitudinal Assessment of Suicide Risk: Development and Usability Study
JMIR Form Res 2025;9:e76544
URL: https://formative.jmir.org/2025/1/e76544
doi:10.2196/76544

© Kenneth McClure, Brooke A Ammerman, Cheng Liu, Miguel Blacutt, Connor O'Brien, Ross Jacobucci. Originally published in JMIR Formative Research (https://formative.jmir.org), 26.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Technology-Related Trauma in Sexual and Reproductive Health Digital Technologies: Grounded Theory Study

Abdul-Fatawu Abdulai[1], RN, PhD; Janell Cosco Josephs[1], RN, MSN; Adrian Guta[2], MSW, PhD; Mark Gilbert[3], MD, MHSc, FRCPC; Vicky Bungay[1], RN, PhD

[1]School of Nursing, University of British Columbia, T201-2211 Wesbrook Mall, Vancouver, BC, Canada
[2]School of Social Work, University of Windsor, Windsor, ON, Canada
[3]British Columbia Centre for Disease Control, Vancouver, BC, Canada

Corresponding Author:
Abdul-Fatawu Abdulai, RN, PhD
School of Nursing, University of British Columbia, T201-2211 Wesbrook Mall, Vancouver, BC, Canada

Abstract

Background: Digital health technologies are increasingly used as complementary and alternative means of seeking sexual and reproductive health services. These platforms now play a critical role in facilitating services such as contraception counseling, abortion care, sexually transmitted infection testing and treatment, and fertility-related support, particularly for individuals who face barriers to in-person care. Despite their increasing prevalence, there is an emerging concern that such platforms could inadvertently trigger or perpetuate trauma among end-user patients. This risk is particularly salient for individuals from equity-deserving populations who already navigate stigma, discrimination, or prior traumatic experiences in health care settings.

Objective: This study aimed to develop a theoretical account of how digital health technologies can cause or perpetuate emotional trauma among people who seek technology-based sexual and reproductive health services.

Methods: We used the Charmaz constructivist grounded theory approach by conducting interviews with 25 participants who have used government and other regulated digital health platforms (ie, web-based platforms and mHealth apps) to access sexual and reproductive health information or services including sexually transmitted infection testing, contraception, and abortion. Data analysis occurred alongside data collection, and data were analyzed inductively using open, axial, and theoretical coding.

Results: We developed an explanatory model that shows that technology-related harm can occur in two main ways: (1) digital platform design features (ie, navigation challenges, data and security breaches, and inappropriate display of content) and (2) digital platform–related interpersonal interactions (targeted campaigns and depersonalized digital health interactions). While these activities can cause harm to users in general, they are more likely to result in emotional trauma for individuals with prior traumatic experiences and in emotional discomfort for those without such histories.

Conclusions: Web-based platforms provide opportunities for advancing access to sexual and reproductive health and services. At the same time, these technologies can also serve as conduits through which trauma can be triggered, perpetuated, and exacerbated. While technology-related trauma could occur unintentionally via design choices, some activities, including technology-related interactions, could trigger or perpetuate trauma among end users. To mitigate the risks, both technology developers (particularly designers) and health providers should consider design choices and implementation strategies that not only prevent trauma but also promote users' emotional well-being.

(JMIR Form Res 2025;9:e79365) doi:10.2196/79365

KEYWORDS

emotional trauma; technology-related trauma; sexual and reproductive health; digital health technologies; grounded theory

Introduction

Digital health technologies are increasingly used as complementary or alternative means of seeking potentially sensitive and stigmatized sexual and reproductive health services [1,2]. While there is no universally accepted definition of digital health, it is generally conceptualized as the use of information and communications technologies in medicine and other health professions to manage illnesses, improve health, and health care delivery [3]. The increasing adoption of digital health is revolutionizing how sensitive and stigmatized sexual and reproductive health services are accessed and delivered [2]. Digital health technologies are enabling individuals to access confidential, evidence-based information and services such as contraception management, fertility tracking, sexually transmitted infection (STI) prevention, and telemedicine consultations for abortion services [4-6]. These tools (such as telemedicine and other virtual care modalities like video

XSL•FO
RenderX

appointments) have expanded access to care for marginalized populations (ie, those who regularly experience disproportionate burden of illness due to structural disadvantage, stigma, and discrimination), including those in remote areas or with limited access to traditional health care facilities. Health care systems are increasingly integrating digital health platforms to improve patient engagement, provide personalized care, and address disparities in sexual and reproductive health outcomes. This adoption is driven by advancements in technology and a growing recognition of the importance of user-centered, accessible solutions.

Despite the benefits of digital health technologies for sexual and reproductive health services, there are increasing concerns that such platforms could facilitate some form of harm that can lead to emotional trauma among end users [7,8]. Trauma resulting from the use of digital health technologies (herein referred to as technology-related trauma) can occur depending on how technologies are designed, deployed, or used [7]. While there is not yet a universally accepted definition of technology-related trauma, it is generally conceptualized as technology-enabled activities, interface appearances, or features that leave a user in emotional distress following the use of digital technology [7]. Technology-related trauma can be orchestrated intentionally through negative online activities or inadvertently through interface features that remind people of prior traumatic circumstances [9,10]. For instance, interactive technologies that expose people to online scams, sextortion (sharing of nude pictures on sexual health–related webpages), identity theft (using someone's personal information for soliciting or selling sex products), and cyberbullying (insulting and condescending language in online forums) could result in short-term or long-term emotional trauma, and sometimes, suicidal thoughts for people [11-14].

Digital technologies can also trigger emotional trauma through interface features that may be used to harass, coerce, or stalk someone [15]. These sorts of abuses can range from intimate image abuse and the use of geolocation features for stalking to the use of Internet of Things devices to manipulate people into doubting their sanity [16]. For instance, perpetrators of intimate partner violence can exploit digital health technologies by installing spyware on their partners' devices, compromising their accounts and the potential safety of their victims [15]. Also, downloading health applications that collect, store, or transmit users' location and other personal identifying information can cause short and long-term emotional distress that could result in traumatic outcomes [17]. For people who survived trauma, viewing certain sexually explicit content that they consider to be inappropriate can retraumatize them and further create technology-related cycles of retraumatization each time the content is viewed [18]. Even seemingly neutral user interface design elements developed in line with established design guidelines (eg, the use of different colors and images) can trigger emotional trauma in people with negative prior experiences or a negative attitude towards the said elements [19]. Bonell et al [20] termed these sorts of occurrences the "dark logic" of public health interventions, while Ziebland et al [21] also described it as the "paradoxical consequences of technologies," where interventions designed for treating a particular condition inadvertently deliver a contrary outcome.

While technology-related trauma can occur in any technology-based intervention, its occurrence may be more common in technology interventions that are used for accessing sexual and reproductive health services that are perceived as embarrassing, stigmatizing, and difficult to discuss face-to-face [7,22]. People who seek highly stigmatized sexual and reproductive health services like STI testing, contraception, and abortion services may be prone to technology-related trauma because such services are already associated with an immense amount of societal stigma and shame [7,23]. Due to the stigma associated with visiting conventional sexual and reproductive health facilities, digital health is often used as a complementary or alternative intervention for people who want to access sexual and reproductive health services.

Given the stigma and shame already associated with such services, any negative event experienced via technology-based interventions could either trigger or perpetuate such prior trauma. For instance, interventions designed to increase adherence to HIV antiretroviral drugs or interventions for promoting contraceptive use can trigger traumatic experiences when people's personal information collected through these interventions is deliberately or inadvertently disclosed. Furthermore, the use of abortion-related technologies can generate hate and antichoice activities that may lead to traumatic feelings [22,24]. Other digital health such as menstrual tracking apps, web-based virtual consultation services, and mobile apps has been shown to incorporate surveillance and social control mechanisms to depict sexual and reproductive health services like abortion in a negative light [25]. Given the stigma associated with sexual and reproductive health services and the concurrent increase in digital technology adoption [26], any additional trauma resulting from the use of technology interventions could create and exacerbate health inequities, particularly for marginalized populations who rely on technology-based interventions as complementary or alternative means of seeking health care. Despite the threats of technology-related trauma in sexual and reproductive health technologies, researchers have yet to explore how trauma manifests in sexual and reproductive health-related technologies. This study sought to address this gap by generating an explanatory model of how digital health technologies can contribute to or perpetuate trauma among end users.

Methods

Study Design

We adopted the Charmaz constructivist grounded theory approach [27] to generate an explanatory model of how emotional trauma manifests in the contexts of sexual and reproductive health technologies. Grounded theory was considered an appropriate approach to exploring this phenomenon as little is known about how emotional trauma can be orchestrated through sexual and reproductive health technologies.

XSL•FO
RenderX

Ethical Considerations

Ethical approval was obtained from the University of British Columbia Behavioral Research Ethics Board (H23-02934). Written, fully informed consent was obtained from all participants in adherence to the requirements outlined in the Declaration of Helsinki. The interviews were deidentified before analysis. At the end of each interview, participants were provided with CAD 30 (approximately US $23) in the form of an Amazon gift card as honoraria for their time.

Study Population, Eligibility, and Recruitment Approaches

The definition of trauma in this study is specific to emotional and psychological trauma, which is different from physical trauma (ie, injuries or body wounds produced by sudden physical injury from impact, violence, or accident). Psychological trauma is used to describe the emotional reactions that occur in response to extraordinary frightening or distressing events [28]. The population for this study comprised people who use digital health technologies like web-based platforms and mobile apps to access sexual and reproductive health services or information across provincial cities in western Canada. Digital health technologies are used to refer to government and not-for-profit regulated web-based platforms or mobile phone apps that are used to provide sexual and reproductive health services including contraception, abortion, STI testing, medical appointments, and teleconsultation services. Participants were recruited from 2 main sources: ReachBC and Options for Sexual Health Clinics. ReachBC is a provincial web-based platform that connects British Columbia volunteers and researchers to facilitate health research in the province of British Columbia. Options for Sexual Health is the largest nonprofit sexual health organization in Canada and has more than 50 clinics in British Columbia. For this study, the recruitment notice was posted on ReachBC and shared across 7 Options for Sexual Health participating clinics. Individuals who expressed interest contacted the research team, and their eligibility was confirmed prior to scheduling an interview.

To be eligible for inclusion, participants must: (1) have used a web-based platform or an mHealth tool to access any of the sexual and reproductive health information or services outlined above within the last year, (2) be aged ≥19 years, and (3) be able to speak English.

Data Collection

Data were collected through individual semistructured interviews conducted via Zoom (Zoom Communications, Inc). All interviews were conducted by the first author. Before delving into the interview questions, we asked participants to complete a brief verbal survey about their demographic characteristics. The interview questions addressed topics related to how digital health technologies could cause emotional trauma among end users. Specifically, participants were asked questions related to their experiences in using digital health technologies to access sexual and reproductive health services, the challenges they encountered when using digital health technologies, and how the use of such technologies could cause or perpetuate trauma. They were also asked questions about how the features and interface appearance of digital health technologies could perpetuate historical and existing vulnerabilities among people with sexual and reproductive health conditions or disorders. Data collection occurred until we reached saturation by informational redundancy (when new data repeats what was previously said).

While we did not seek to actively recruit people with prior trauma, there was a tendency for some people with prior trauma to choose to participate in the study. Given this understanding and the fact that some participants might have had prior trauma related to sensitive health topics like contraception, abortion, and STI testing, we embedded trauma-informed practices throughout the research processes: before, during, and after the data collection [29,30]. Before data collection, the first and second authors were trained in trauma-informed care practices on how to communicate with and identify signs of distress among participants. We also spelled out the potential question areas to address during the interviews as well as the interview procedures on the consent form. Two community partners and an expert in trauma-informed care reviewed the interview guide and provided feedback before the data collection began. We also co-created and distributed a self-care resource to each participant 3 days ahead of the interview. The participants were asked to review the self-care resources before and after the interviews. In these resources, we provided contacts of counselors to reach out to in case of any distress. Recognizing that some participants might be living with perpetrators of abuse, we encouraged them to choose locations that are safe and convenient to step out when the need arises. *During the data collection*, we asked participants to take a break or end the interviews if they felt uncomfortable. We periodically checked in with the participants to see if they were doing well or showed any signs of discomfort or tiredness on their part. To ensure anonymity and promote participants' freedom of expression, participants were given the option to turn off their videos if desired. However, the interviewer had their Zoom turned on throughout the course of the interviews. After the data collection, we followed up with each participant in 2 days and 1 week to see how they were doing. We also provided them with contacts of mental health services within their catchment areas to reach out to in case they develop any signs of emotional distress following the interview. All the interviews lasted between 45 and 60 minutes. Data collection and analysis occurred concurrently. As recruitment and data analysis proceeded, we used the emerging concepts from the preliminary findings as a guide to recruiting additional participants to fill uncertainties in the data and build on the emerging theoretical model. These additional participants reflected people with diverse lived experiences as well as diversity in age, sex, and gender identity. The interviews were audio-recorded, transcribed using the web-based transcription services (Temi, Inc), deidentified, and then exported into NVivo (Lumivero) for analysis.

Data Analysis

The first and second authors reviewed the audio transcripts immediately after each interview. The audios were then transcribed, and each textual transcript was compared with its corresponding audio file. A trauma-informed lens also informed ethical handling of the data, prioritizing participant

XSL•FO
RenderX

confidentiality while minimizing harm. Consistent with the constructivist grounded theory approach, we generated inductive codes through initial coding, focused coding, and theoretical coding [27]. In initial coding, the first and second authors "broke down the raw data" into parts by labeling each data segment with codes. As the data analysis progressed, we compared the data to identify any similarities or differences among participants of different ages, sexes, and gender identities. The data analysis occurred concurrently with data collection, and additional data were used to modify the codes. In focused coding, we sorted, integrated, and organized the most significant and frequent initial codes into the core categories. We sought explanations for differences, explored associations between the categories, and further examined which codes contribute most to the emerging categories. In theoretical coding, we refined the categories and related them to one another to construct a theoretical understanding of the processes by which digital technologies could foment or perpetuate trauma among end users (Textbox 1). Our understanding of trauma informed our data analysis by fostering sensitivity to participants' experiences of digital technologies and how that may shape their narratives. We contextualized participants' responses within broader systemic and relational dynamics, recognizing the impact of prior trauma and power imbalances between the participants and the researchers. Recognizing that participants might choose to filter or withhold information due to fear, mistrust, or discomfort rooted in past trauma, we paid attention to areas in the transcripts where participants may be silent, avoiding some questions or withholding some information. Analyzing the data from a trauma perspective helped us to fill in uncertainties and confirmed participants' narratives regarding some topic areas that could be regarded as sensitive.

XSL•FO
RenderX

Textbox 1. An overview of the main categories and coding process.

Thematic area: inappropriate content display

- Traumatic content
  - Depicting explicit images on digital platforms
  - Exposure of personal or nude photos
  - Sudden exposure to sexualized content (images and videos)
  - Not warning people about sexually explicit content
- Misrepresenting identity
  - Gender misrepresentation in online platforms
  - Wrong or inappropriate translation of content in platforms
- Explicit content
  - Profane advertisement via online platforms
  - Profane images

Thematic area: data and security breaches

- Cyber-attack concerns
  - Concerns about cyber-attacks and lawsuits
  - Hacking and phishing attempts
  - Concerns about possible data breaches
  - Concerns with creating profiles on online platforms
- Identity exposure
  - Concerns with exposing personal information to health providers
- Confidentiality concerns
  - Lack of privacy and anonymity of digital resources
  - Lack of privacy when accessing via public computers
  - Revealing confidential information

Thematic area: navigational challenges

- Dense interfaces
  - Inability to find the desired information
  - Difficulty finding information in a stressful time
  - Crowded spaces in digital platforms
  - Difficulty in understanding content
- Technical difficulties
  - Technophobia in using technology
  - Going through multiple menus to locate information

Thematic area: technology-related targeted campaigns

- Identity theft
  - Linking sensitive issues to peoples' names
  - Using personal information for fraud
  - Using people's data against them

XSL•FO
RenderX

- Tracking concerns
  - Concerns about cookies
  - Soliciting for sensitive information on web platforms
- Misinformation and disinformation
  - Tracking personal information with targeted advertisement
  - Fake and misleading information on sexual minorities
  - Negative stereotyping
  - Anti-lesbian, gay, bisexual, trans, intersex, or queer (LGBTIQ) activities on web platforms

Thematic area: technology-related depersonalization of sensitive sexual health interactions

- Trauma-triggering communication
  - Reminding people of trauma triggers
  - Unsympathetic communication with care providers
  - Victim blaming during online communications
- Cultural insensitivity
  - Lack of language and cultural sensitivity
  - Nonpersonalized communication and information
  - Not understanding people's prior trauma history
  - Trivializing sexual health (sexually transmitted infections and abortion)
  - Normalizing people's suffering through technology

Results

Demographic Characteristics

A total of 25 participants took part in the study, of which 16 (64%) of the participants identified as women, 5 (20%) identified as men, and 4 (16%) identified as nonbinary. The participants' ages ranged from 23 to 67 years, with a mean age of 35 (SD 12) years. All participants indicated having used digital technologies to access sexual and reproductive health services, including STIs, contraception, and abortion services, within the last year. Even though participants were not asked about prior trauma experiences, approximately 68% (17/25) directly or indirectly revealed that they have experienced a prior event in the past that they consider to be traumatizing. While the other 32% (8/25) did not indicate prior trauma, we could not tell whether they had prior trauma, as some participants might choose not to disclose their trauma history.

How Trauma Manifests in Digital Health Platforms

In this study, we constructed a theoretical model illustrating how digital health technologies can facilitate trauma among end users. Our model shows that digital health platforms can harm patients seeking sexual and reproductive health services or information in 2 primary ways: digital platform design features (ie, navigation challenges, data and security breaches, and inappropriate display of content) and technology-related interpersonal interactions (targeted campaigns and depersonalized digital health interactions). However, whether these technology-related harms result in trauma, however, depends on an individual's prior experiences of trauma related to sexual and reproductive health services. People with prior trauma related to sexual and reproductive health services who happen to experience these negative activities are likely to experience retraumatization, while people with no prior trauma are likely to experience technology-related discomfort. Figure 1 presents the theoretical model of sexual and reproductive health technology-related trauma developed in this study. The next sections elaborate on how each activity manifests as technology-related trauma.

XSL•FO
RenderX

Figure 1. An explanatory model of technology-related trauma in sexual and reproductive health.

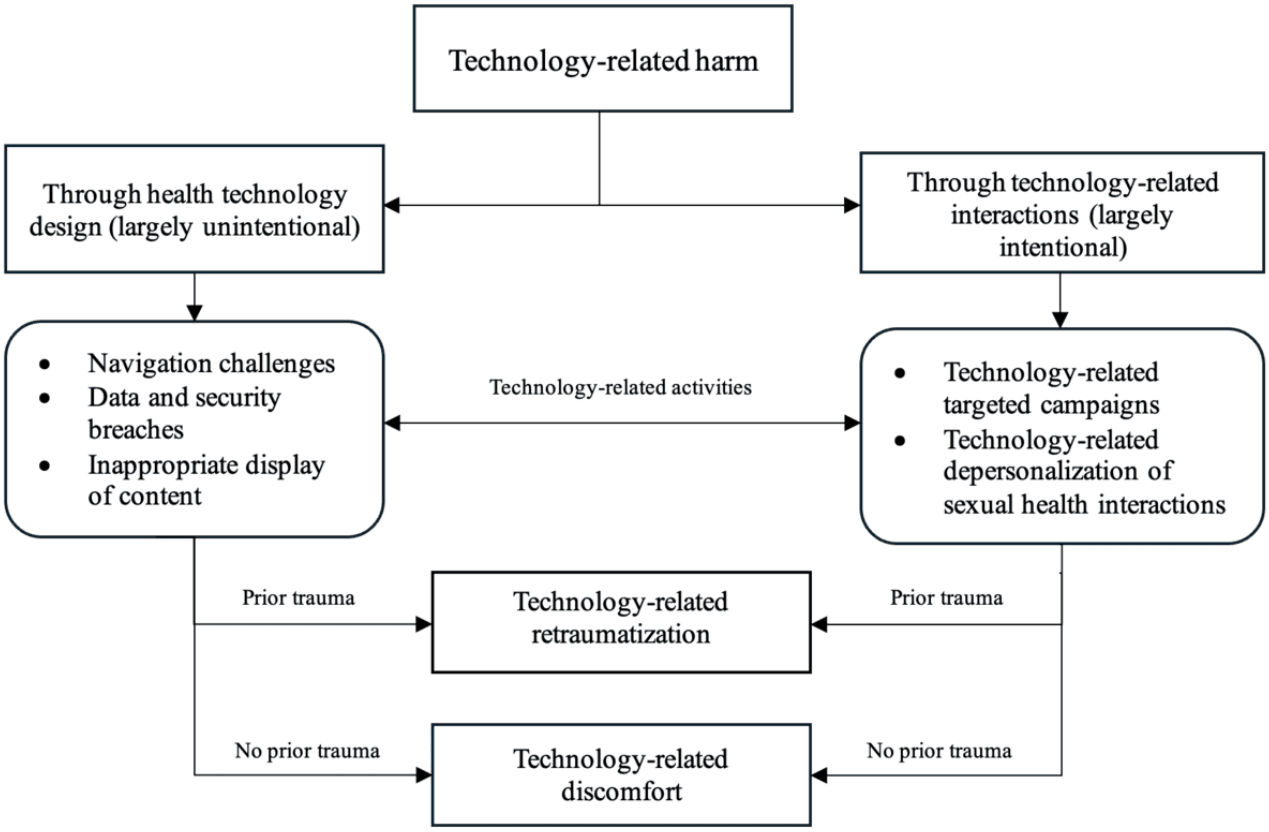

Through Health Technology Design (Largely Unintentional)

Overview

This section presents how certain user interface design features may contribute to psychological or emotional trauma among individuals accessing sexual and reproductive health services through digital platforms. These activities were considered largely unintentional and could occur during the design and development stages of digital health technologies. These activities are considered largely unintentional, as they typically arise during the design and development stages when technology developers act with good intentions but may not fully recognize the potential for retraumatization of users.

Inappropriate Display of Content and Lack of Representation in Digital Platforms

The participants indicated ways in which the presentation of content on digital health platforms could retrigger emotional trauma or cause emotional discomfort among people with sexual and reproductive health conditions or disorders. In particular, imagery, videos, and language perceived as vulgar were described as potentially retraumatizing when they evoked memories of past traumatic experiences. Some participants recounted how they got distressed when such imagery, videos, or language were suddenly exposed to them without any prior warning. One participant who identified as a woman indicated how awkward and distressful she felt upon being suddenly exposed to a picture of female genital mutilation (ie, a complete or partial removal of the female external genitalia). Participants noted that certain technology-based interventions, such as those designed to support individuals with STIs or to assist survivors of intimate partner violence or sexual assault who become pregnant, could inadvertently retraumatize users by triggering reminders of their past experiences. One participant reiterated how sexual assault victims could be retraumatized via digital technologies by stating that:

> *if somebody who got pregnant as a result of sexual assault were to go to a website and look for help or look for sexual health services. And all of a sudden, they started seeing pictures, videos, or stories of sexual assault. Those things could all potentially be triggering because they are reminding them of what happened to them.*

While the display of explicit content could retraumatize people by reminding them of their prior experiences, participants varied in how they perceived and responded to this possibility. Sudden exposure to explicit content was reported to trigger emotional trauma among participants with prior real-life experiences, whereas for those without such histories, these images were more likely to cause emotional discomfort rather than trauma.

In addition to the inappropriate display of content on digital platforms, some participants, particularly 2-spirit, lesbian, gay, bisexual, transgender, queer, intersex, and gender-diverse individuals, felt that they are often not adequately represented and, in some situations, misrepresented on digital health platforms. They indicated that such misrepresentation could perpetuate existing harmful practices that they already face in society and could trigger emotional trauma when they visit digital health platforms. The participants also indicated that they could feel some emotional discomfort and, in some cases,

XSL•FO
RenderX

distress when well-intentioned educational messages on digital health platforms conflict with their situation. One participant likened it to seeing messages about baby care or the need to have children after having had an abortion. They revealed how they felt uncomfortable after seeing an advert by stating that:

> *I opened this website and I saw something, and it was like life begins at conception. For someone who had an abortion, I was just so disgusted…. and that goes to perpetuate whatever traumatic experiences I might have had during the abortion because I am thinking, like maybe, I should have kept this kid even though I can't financially and emotionally support this kid. It's just like thinking back and all that trauma it comes with.*

Privacy, Data, and Security Breaches in Digital Platforms

Participants also described how unsecured digital health design features could expose personal information stored on digital platforms to unauthorized access and potential misuse. Participants noted that data breaches in digital health can cause profound emotional trauma for patients with sexual and reproductive health disorders due to the highly sensitive nature of the information involved. They indicated that the unauthorized disclosure of sensitive information, such as infertility, sexually transmitted infections, or abortion history, can result in stigma, shame, and fear of judgment from peers, employers, or family members. The breach of privacy may also exacerbate preexisting mental health challenges, including anxiety and depression, as patients worry about how their data could be misused. For people with prior trauma, the mere thought of data and security breaches, phishing attempts, and possible cyber-attacks was considered distressful and could trigger past traumatic events. One participant indicated how data breach concerns could retraumatize them by stating that:

> *My concern just comes down to privacy, how much information is revealed, what happens with the data that we enter into these systems, where and how is that stored? And you know, those are the kinds of things that I want to know because it worries me that it could get into the wrong hands. And that includes an unfriendly government. And when you add all these layers on top of those really sensitive things about yourself, it could be scary and traumatic. I mean, let's not mess around here.*

The participants indicated that the loss of confidentiality in digital platforms can erode trust in health care systems and may discourage individuals from seeking the necessary care, ultimately worsening their emotional and physical health challenges. While participants noted that data breaches could retraumatize those with prior trauma, they were more often perceived as causing emotional distress among individuals without a reported trauma history. Participants, particularly those who use public computers, expressed concern that their personal information could be exposed and their identities revealed when accessing services in such settings. Concerns about information exposure were especially pronounced among individuals seeking sexual or reproductive health information.

Navigation Challenges in Digital Platforms

The findings also show that design-related issues, such as navigational challenges on digital platforms, can trigger or exacerbate existing trauma for some individuals seeking sexual and reproductive health services, particularly those considered private, stigmatizing, or uncomfortable to discuss in face-to-face settings. The participants noted that dense, cluttered user interfaces and excessive navigation steps could create delays and heighten anxiety for individuals who are already traumatized. For people with prior trauma seeking sensitive sexual and reproductive services, encountering poorly designed interfaces, unclear instructions, or inaccessible resources can exacerbate emotional distress. Participants indicated that their inability to quickly find confidential, reliable information or services may heighten feelings of isolation and stigmatization, making them question their decision to seek help online. One participant cited how the presence of many icons and the detailed registration processes on a web-based platform could add to their anxiety by stating that:

> *Emotionally, I think it just adds to your stress already. You could already be quite worked up and upset or maybe if it's an emergent situation like I needed a pregnancy test or I'm looking to have an abortion. So, there's just a lot of emotions involved. And then to kind of add this added stress of how to figure it out on your own and then have follow many steps and to call this or go here and book it.*

According to the participants, this "compounded stress" can lead to avoidance behaviors, where individuals delay or forgo critical health care services altogether, further impacting their mental and physical well-being.

Through Technology-Related Interactions

Overview

Our model further illustrates that technology-related retraumatization can occur through interactive activities, specifically (1) targeted digital campaigns and (2) the depersonalization of sensitive sexual health interactions. This thematic category was classified as largely intentional, as such activities often occur during patients' use of digital health technologies and may be deliberately orchestrated to cause harm.

Technology-Related Targeted Campaigns

Participants expressed concern that technology-related targeted campaigns could trigger or perpetuate trauma among end users. These concerns applied to campaigns delivered through both regulated sexual health digital platforms and social media. Some participants reported feeling distressed when pornographic images unexpectedly appeared on sexual health-related digital platforms or were targeted at them via social media. The participants also indicated that people who sought stigmatized services like abortion could get traumatized if baby products or pregnancy-care products suddenly pop up. One participant stated:

> *On the internet these days, everything you do is being tracked, then repackaged, and then used to try to sell*

XSL•FO
RenderX

> *you things. For someone who just lost their baby or had an abortion being sent an advert like baby products, how would you want them to feel after they have gone through this ordeal? And there are certain things, especially with trans people where if you have to put your legal name down in these things, you are dead naming yourself, so you're having to use the name that causes you harm.*

Participants also indicated that targeted advertisements or content based on sensitive browsing histories or health-related searches could make them feel threatened or exposed, as if their private struggles were being exploited for profit. This was particularly distressing for individuals dealing with stigmatized issues such as sexually transmitted infections, infertility, abortion, or contraception. The participants indicated that constant exposure to unsolicited advertisements can serve as a reminder of personal challenges, triggering anxiety, embarrassment, or shame. Furthermore, the fear of being "found out" or judged if someone else sees such targeted advertisements adds another layer of psychological distress, eroding trust in digital platforms and health care systems. Participants also noted that digital health technologies can be used to target minority groups with misinformation and negative stereotypes. They highlighted that sexual minority individuals seeking services such as abortion are particularly vulnerable to such technology-related misinformation, given societal disapproval of such services.

Technology-Related Depersonalization of Sensitive Sexual Health Interactions

Participants also expressed concerns about how interactive digital health platforms could depersonalize the care process and lead to patient-provider practices that could inadvertently retraumatize people. The participants noted that health care professionals who may not understand the social and emotional history of their patients might communicate as though "everything is normal," while some providers are likely to ask questions that remind people of their prior trauma or seek to blame them for their predicaments. One participant indicated that:

> *People could get retraumatized if healthcare professionals communicate in ways that seek to blame or remind them of their past experiences, like how they got infected with STIs, how they got pregnant, or why they are seeking contraception or abortion services.*

Participants noted that digital health technologies, particularly virtual care such as telemedicine, often limit providers' ability to gain a detailed understanding of their patients. Without the ability to see facial expressions in some technology-related interactions, health professionals may not be able to convey the needed empathy for a patient who might have survived trauma. Such people could be retraumatized if virtual communications do not convey the needed empathy. One participant stated:

> *Sometimes when you are interacting with health providers online, you don't think you're fully heard. If it's just a chat room or a phone call, the healthcare professional might not understand or know what state of mind this person is in because they can't see you. Let's say if they're in a really dire crisis mode and the healthcare professional didn't really know it because they're just asking a question. They might trigger something within them.*

In addition to the possibility for retraumatization via technology-related depersonalizations, the participants were also concerned that technology-related peer-to-peer interaction channels like chatrooms and messaging boards could trigger and perpetuate trauma when people infiltrate these platforms with the intent of harming other participants.

Discussions

Principal Findings

Digital health technologies are increasingly recognized as offering opportunities for advancing sexual and reproductive health and services [2,6]. At the same time, these technologies can also serve as conduits through which trauma can be triggered, perpetuated, and exacerbated [21,22]. While various trauma-informed approaches have been proposed in the technology design arena [31-34], they are often not grounded in a foundational understanding of how technologies can cause trauma, nor tailored to the distinctive elements of sexual and reproductive health. To our knowledge, this study is the first to develop a theoretical model illustrating how emotional trauma can manifest within sexual and reproductive health technologies. Our model suggests that technology-related harm may largely arise either unintentionally through health technology design or intentionally through technology-mediated interactions. However, whether these technology-related activities result in emotional trauma depends on whether people have prior trauma related to sexual and reproductive health. For people with prior trauma, exposure to inappropriate content, data and privacy breaches, targeted campaigns, navigational challenges, or nonpersonalized interactions can result in retraumatization when these experiences mirror or reinforce their past trauma. Conversely, individuals without such histories may only experience emotional discomfort from these technology-related encounters, though the impact is less likely to be retraumatizing. The differences in opinions on how technology-related harms can lead to trauma in digital health platforms reflect the assertion that not everyone who is exposed to a traumatic event will eventually be traumatized [35]. That is, technology experiences can elicit harm, but such harm may not necessarily translate into trauma for some group of people.

Our study also confirmed the assertion that design elements that might be considered minor by the design team or some conversations that are considered usual in patient-provider interactions could still end up retraumatizing some people [36]. Other design elements developed with the best of intentions could still end up retraumatizing people if they find such content undesirable or reminding them of their prior trauma [19]. For instance, some abortion-related educational content displayed on digital health platforms may remind people of abortion trauma [37]. Although such design features may be viewed as standard by intervention developers, some are likely to trigger or perpetuate psychological and emotional trauma among

XSL•FO
RenderX

equity-deserving populations who have experienced personal, relational, or intergenerational trauma, including indigenous peoples, individuals experiencing homelessness, those with disabilities, or survivors of intimate partner violence [7,38-40]. For such people, technology-related trauma may occur on top of other traumas, thus potentially exacerbating their harm.

One important finding from this study was how digital health–related interactions can intentionally or inadvertently trigger or perpetuate technology-related trauma. While previous studies revealed how digital health technologies can be deliberately used to cause trauma [14], our findings highlight the surprising reality that routine digital health interactions can also inadvertently retraumatize individuals. The potential for such retraumatization may stem from the limited opportunities in virtual care for health care professionals to fully recognize or understand a patient's trauma history. This lack of recognition of patients' traumatic past can lead to impulsive remarks that can be deeply hurtful, or at best, unhelpful to the patient's situation [41]. Even when empathy is conveyed, the lack of body language in digital communications might make it easier for messages to be interpreted as unsympathetic or harsh [42]. Furthermore, digital communication may lack the nuance needed for a sensitive topic like sexual health—making some information appear abrupt or dismissive to patients. This assertion is supported by prior studies that noted that the absence of verbal and nonverbal cues, body language, and tone in virtual care technologies can hinder the development of trust and rapport between health care providers and their patients [43,44]. For people who survived trauma, the inability to fully capture patients' verbal and nonverbal cues may not only make it challenging for providers to recognize and respond to signs of distress but may also leave patients feeling unsupported or make them struggle to express their emotions or disclose their trauma history [45]. For people with prior trauma, this could go a long way to hinder communication, affect rapport building, and make the health-seeking process incredibly difficult and time-consuming for the patient, hindering the trauma-healing process [46].

In addition to trauma arising from digital health–related interactions, concerns about data and privacy breaches, misinformation, and unexpected exposure to explicit digital content were also reported to retraumatize individuals. These activities can erode trust and make technology users feel skeptical about the authenticity of health advice and product recommendations, while causing them to question the motives behind the information they receive. Such targeted campaigns can make people feel as though they are constantly being watched and analyzed, leading to discomfort and distress over the potential loss of their anonymity and privacy [47,48]. The concern that one's information is under surveillance could also make them develop feelings of paranoia and fear. Constant worry about the misuse of personal information can lead to chronic anxiety or depressive episodes, potentially leading to mistrust and nonadoption of digital health technologies. In addition, cultural insensitivity in virtual interaction with patients, particularly those from minority groups, might inadvertently perpetuate trauma. For instance, when virtual interactions fail to recognize or respect cultural values, norms, and experiences, patients may feel misunderstood, marginalized, or dismissed, potentially exacerbating existing trauma and reducing trust in digital health services.

The findings indicate that digital health interventions may represent an emerging and underrecognized source of trauma, particularly affecting individuals who experience stigma, discrimination, or inequitable access to health care services. Therefore, health care systems that develop digital health interventions should consider how digital health design features and digital health–related interactions could trigger or perpetuate emotional trauma. Given the growing use of sexual and reproductive health digital health technologies among survivors of trauma [33], there is an urgent need to pay attention to interface features that could end up retraumatizing the very people they intend to help. Special attention should be given to user-centered design practices by fully engaging with people with trauma during the design of digital health technologies. Digital health intervention developers within the health care system should also consider integrating privacy and security measures, including content warnings and user control on digital platforms, paying attention to inclusive representation and cultural sensitivity, integrating accessible design, a continuous evaluation process, and integration with in-person services are critical to mitigating technology-related trauma [24]. Textbox 2 shows a summary of considerations for trauma-informed digital technologies in sexual and reproductive health.

XSL•FO
RenderX

Textbox 2. Key recommendations for trauma-informed digital health technologies.

Design considerations

- Privacy and security measures integration
- Inclusive representation and cultural sensitivity
- Clear navigation pathways
- Trauma-informed language and imagery guidelines
- Implementation of content warnings and user controls

Implementation considerations

- Provider training in trauma-informed digital communication
- Integration with in-person services
- Regular user feedback mechanisms
- Cultural safety evaluation
- Privacy impact assessments

User-support considerations

- Clear opt-out pathways
- Accessible help resources
- Alternative service options
- User control over data-sharing
- Support for diverse communication preferences

Limitations

This study was based on a small sample of 25 participants that was limited to one geographic region in western Canada. These limitations highlight the need for future research to engage participants from more diverse cultural backgrounds and from regions across Canada and beyond to ensure that the emerging theory on technology-related trauma reflects cross-cultural and geographical nuances. While some of our participants reported (directly or indirectly) some traumatic experiences in the past, we could not ascertain that their views represent varying forms of traumatic experiences that tend to affect people who access sexual and reproductive health services. Despite these limitations, the findings offer important and timely implications for designing and using digital health technologies that transcend beyond the study settings.

Conclusions

Technology-related trauma in sexual and reproductive health digital platforms underscores a critical intersection of healthcare innovation and patient vulnerability. While digital platforms offer unprecedented opportunities for education, support, and access to services, they also pose significant risks of retraumatization and emotional harm. Privacy concerns, the potential for exposure to triggering content, and the depersonalization of sensitive sexual health interactions can retraumatize individuals already navigating complex and sensitive sexual and reproductive health challenges. Given this foundation knowledge of the emerging threats of digital health technologies, health care providers, technologists, and policy makers need to work collaboratively to mitigate these risks, for example, by implementing robust security measures, promoting digital literacy, and ensuring that the human element of care is prioritized in digital health practices. By doing so, the potential harms of technology-related trauma can be minimized, allowing people to realize the full potential of sexual and reproductive health–related digital platforms without compromising patients' emotional safety. We anticipate that the findings of this study will produce a new understanding that will help inform recommendations for mitigating technology-related trauma in digital health.

Acknowledgments

The authors would like to thank all participants who took part in the interviews.

Funding

Funding for this study was provided by the Social Sciences and Humanities Research Council of Canada (Grant ID: GR028128). AD was supported through the Canada Research Chairs program. The funding body had no role in the study design, data collection, writing of the manuscript, or the decision to submit for publication.

XSL•FO
RenderX

Data Availability

The datasets generated during the current study cannot be publicly shared because they contained identifiable information of people who experienced prior trauma.

Authors' Contributions

AFA conceptualized the study, obtained the funding and ethics approval, and led the data collection, analysis, and writing of the manuscript. JC led participants' recruitment and follow-up activities. AG, MG, and VB obtained funding, provided regular feedback on the data collection, analysis, and interpretation of findings. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Abdulai AF, Howard AF, Yong PJ, Noga H, Parmar G, Currie LM. Developing an educational ebsite for women with endometriosis-associated dyspareunia: usability and stigma analysis. JMIR Hum Factors 2022 Mar 3;9(1):e31317. [doi: 10.2196/31317] [Medline: 35238789]
2. Abdulai AF, Howard AF, Parmar G, et al. The quality, suitability, and readability of web-based resources on endometriosis-associated dyspareunia: a systematic review. Health Informatics J 2024;30(1):14604582241231151. [doi: 10.1177/14604582241231151] [Medline: 38308637]
3. Fatehi F, Samadbeik M, Kazemi A. What is digital health? Review of definitions. Stud Health Technol Inform 2020 Nov 23;275:67-71. [doi: 10.3233/SHTI200696] [Medline: 33227742]
4. Witzel TC, Guise A, Nutland W, Bourne A. It Starts With Me: Privacy concerns and stigma in the evaluation of a Facebook health promotion intervention. Sex Health 2016 Jun;13(3):228-233. [doi: 10.1071/SH15231] [Medline: 27075940]
5. Shearer LS, Simmons L, Mindel A, Stanberry LR, Rosenthal SL. Reducing the stigma of herpes simplex virus infection: lessons from an online video contest. Sex Health 2012 Nov;9(5):438-444. [doi: 10.1071/SH11188] [Medline: 23036138]
6. Luigi-Bravo G, Maria Ramirez A, Gerdts C, Gill R. Lessons learned from developing and implementing digital health tools for self-managed abortion and sexual and reproductive healthcare in Canada, the United States, and Venezuela. Sex Reprod Health Matters 2023 Dec;31(4):2266305. [doi: 10.1080/26410397.2023.2266305] [Medline: 37870150]
7. Chen JX, McDonald A, Zou Y, et al. Trauma-informed computing: towards safer technology experiences for all. 2022 Apr 29 Presented at: CHI ’22: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems; Apr 29 to May 5, 2022; New Orleans, LA, USA p. 1-20 URL: https://dl.acm.org/doi/proceedings/10.1145/3491102 [doi: 10.1145/3491102.3517475]
8. Abdulai AF. Is generative AI increasing the risk for technology-mediated trauma among vulnerable populations? Nurs Inq 2025 Jan;32(1):e12686. [doi: 10.1111/nin.12686] [Medline: 39560355]
9. Holloway IW, Dunlap S, Del Pino HE, Hermanstyne K, Pulsipher C, Landovitz RJ. Online social networking, sexual risk and protective behaviors: considerations for clinicians and researchers. Curr Addict Rep 2014 Sep;1(3):220-228. [doi: 10.1007/s40429-014-0029-4] [Medline: 25642408]
10. Musetti A, Starcevic V, Boursier V, Corsano P, Billieux J, Schimmenti A. Childhood emotional abuse and problematic social networking sites use in a sample of Italian adolescents: The mediating role of deficiencies in self-other differentiation and uncertain reflective functioning. J Clin Psychol 2021 Jul;77(7):1666-1684. [doi: 10.1002/jclp.23138] [Medline: 33837547]
11. Kraft E. Cyberbullying: a worldwide trend of misusing technology to harass others. 2006 Jun 7 Presented at: The Internet Society II: Advances in Education, Commerce & Governance; Jun 12-14, 2006; The New Forest, UK p. 155-166 URL: http://library.witpress.com/viewproceedings.asp?pcode=IS06 [doi: 10.2495/IS060161] [Medline: 36193827]
12. Schüz B, Urban M. Unintended consequences and side effects of digital health technology: a public health perspective. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2020 Feb;63(2):192-198. [doi: 10.1007/s00103-019-03088-5] [Medline: 31950231]
13. Henry N, Flynn A, Powell A. Technology-facilitated domestic and sexual violence: a review. Violence Against Women 2020 Dec;26(15-16):1828-1854. [doi: 10.1177/1077801219875821] [Medline: 32998673]
14. Machimbarrena JM, Calvete E, Fernández-González L, Álvarez-Bardón A, Álvarez-Fernández L, González-Cabrera J. Internet risks: An overview of victimization in cyberbullying, cyber dating abuse, sexting, online grooming and problematic internet use. Int J Environ Res Public Health 2018 Nov 5;15(11):2471. [doi: 10.3390/ijerph15112471] [Medline: 30400659]
15. Freed D, Palmer J, Minchala D, Levy K, Ristenpart T, Dell N. "A stalker’s paradise”: how intimate partner abusers exploit technology. Presented at: CHI ’18: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems; Apr 21-26, 2018; Montreal, Canada p. 1-13. [doi: 10.1145/3173574.3174241]
16. Strohmayer A, Slupska J, Bellini R, Coventry L, Hairston T, Dodge A. Trust and abusability toolkit: centering safety in human-data interactions. : UK Research and Innovation’s Engineering and Physical Sciences Research Council’s Human-Data

XSL•FO
RenderX

Interaction Network; 2021 URL: https://researchportal.northumbria.ac.uk/ws/portalfiles/portal/57063576/TrustAndAbusabilityToolkit.pdf [accessed 2025-11-07]
17. Mehrnezhad M, Almeida T. Caring for intimate data in fertility technologies. 2021 May 6 Presented at: CHI '21: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems; May 8-13, 2021; Yokohama, Japan p. 1-11 URL: https://dl.acm.org/doi/proceedings/10.1145/3411764 [doi: 10.1145/3411764.3445132]
18. Regehr K, Birze A, Regehr C. Technology facilitated re-victimization: How video evidence of sexual violence contributes to mediated cycles of abuse. Crime Media Cult 2022 Nov;18(4):597-615. [doi: 10.1177/17416590211050333]
19. Shaughnessy K, Fehr CJ, Ashley M, et al. Technology-mediated sexual interactions, social anxiety, and sexual wellbeing: a scoping review. Eur J Investig Health Psychol Educ 2022 Jul 29;12(8):904-932. [doi: 10.3390/ejihpe12080066] [Medline: 36005215]
20. Bonell C, Jamal F, Melendez-Torres GJ, Cummins S. "Dark logic": Theorising the harmful consequences of public health interventions. J Epidemiol Community Health 2015 Jan;69(1):95-98. [doi: 10.1136/jech-2014-204671] [Medline: 25403381]
21. Ziebland S, Hyde E, Powell J. Power, paradox and pessimism: On the unintended consequences of digital health technologies in primary care. Soc Sci Med 2021 Nov;289:114419. [doi: 10.1016/j.socscimed.2021.114419] [Medline: 34619631]
22. Bacchus LJ, Reiss K, Church K, et al. Using digital technology for sexual and reproductive health: Are programs adequately considering risk? Glob Health Sci Pract 2019 Dec 23;7(4):507-514. [doi: 10.9745/GHSP-D-19-00239] [Medline: 31874936]
23. Redston S, de Botte S, Smith C. Resolving embarrassing medical conditions with online health information. Int J Med Inform 2018 Jun;114:101-105. [doi: 10.1016/j.ijmedinf.2018.03.010] [Medline: 29605386]
24. Zhang X, Liu S, Chen X, Wang L, Gao B, Zhu Q. Health information privacy concerns, antecedents, and information disclosure intention in online health communities. Inf Manag 2018 Jun;55(4):482-493. [doi: 10.1016/j.im.2017.11.003]
25. Doan AE, Schwarz C. Father knows best: "Protecting" women through state surveillance and social control in anti‐abortion policy. Politics Policy 2020 Feb;48(1):6-37 [FREE Full text] [doi: 10.1111/polp.12337]
26. Abdulai AF, Howard AF, Currie LM. Stigmatizing and de-stigmatizing properties of web apps for sexual health-related conditions: a scoping review. Int J Sex Health 2022;34(1):144-159. [doi: 10.1080/19317611.2021.1949655] [Medline: 38595685]
27. Charmaz K. Constructing Grounded Theory: A Practical Guide Through Qualitative Analysis: SAGE; 2006:228 URL: http://www.sxf.uevora.pt/wp-content/uploads/2013/03/Charmaz_2006.pdf [accessed 2025-11-10]
28. Alexander JC. Trauma: A Social Theory 2012. URL: https://sociology.yale.edu/publications/trauma-social-theory [accessed 2025-11-10]
29. Alessi EJ, Kahn S. Toward a trauma-informed qualitative research approach: guidelines for ensuring the safety and promoting the resilience of research participants. Qual Res Psychol 2023 Jan 2;20(1):121-154. [doi: 10.1080/14780887.2022.2107967]
30. Edelman NL. Trauma and resilience informed research principles and practice: a framework to improve the inclusion and experience of disadvantaged populations in health and social care research. J Health Serv Res Policy 2023 Jan;28(1):66-75. [doi: 10.1177/13558196221124740] [Medline: 36326603]
31. Kelly S, Lauren B, Nguyen K. Trauma-informed web heuristics for communication designers. 2021 Oct 12 Presented at: SIGDOC '21: Proceedings of the 39th ACM International Conference on Design of Communication; Oct 12-14, 2021; Virtual event, USA p. 172-176 URL: https://dl.acm.org/doi/proceedings/10.1145/3472714 [doi: 10.1145/3472714.3473638]
32. Eggleston M, Noel LA. Repairing the harm of digital design using a trauma-informed approach. Diseña 2024 Jan 31(24):7-7 [FREE Full text] [doi: 10.7764/disena.24.Article.7]
33. Venkatasubramanian K, Ranalli TM. Designing post-trauma self-regulation apps for people with intellectual and developmental disabilities. 2022 Oct 23 Presented at: ASSETS '22: Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility; Oct 23-26, 2022; Athens, Greece p. 1-14 URL: https://dl.acm.org/doi/proceedings/10.1145/3517428 [doi: 10.1145/3517428.3544798]
34. Josephs JC, Bungay V, Guta A, Gilbert M, Abdulai AF. Trauma-informed technology design in digital sexual health interventions. Stud Health Technol Inform 2024 Jul 24;315:773-774. [doi: 10.3233/SHTI240323] [Medline: 39049423]
35. Collins S, Long A. Working with the psychological effects of trauma: consequences for mental health-care workers—a literature review. J Psychiatr Ment Health Nurs 2003 Aug;10(4):417-424. [doi: 10.1046/j.1365-2850.2003.00620.x] [Medline: 12887633]
36. Henrich J, Heine SJ, Norenzayan A. The weirdest people in the world? Behav Brain Sci 2010 Jun;33(2-3):61-83. [doi: 10.1017/S0140525X0999152X] [Medline: 20550733]
37. Brewin CR, Gregory JD, Lipton M, Burgess N. Intrusive images in psychological disorders. Psychol Rev 2010 Jan;117(1):210-232. [Medline: 20063969]
38. Nathanson AM, Shorey RC, Tirone V, Rhatigan DL. The prevalence of mental health disorders in a community sample of female victims of intimate partner violence. Partner Abuse 2012 Jan;3(1):59-75. [doi: 10.1891/1946-6560.3.1.59] [Medline: 22741043]
39. Emezue C, Chase JAD, Udmuangpia T, Bloom TL. Technology-based and digital interventions for intimate partner violence: a systematic review and meta-analysis. Campbell Syst Rev 2022 Sep;18(3):e1271. [doi: 10.1002/cl2.1271] [Medline: 36909881]

XSL•FO
RenderX

40. Bailey L, Hulley J, Gomersall T, Kirkman G, Gibbs G, Jones AD. The networking of abuse: Intimate partner violence and the use of social technologies. Crim Justice Behav 2024 Feb;51(2):266-285. [doi: 10.1177/00938548231206827]
41. Center for Substance Abuse Treatment (US). Clinical issues across services. In: Trauma-Informed Care in Behavioral Health Services: A Treatment Improvement Protocol: Substance Abuse and Mental Health Services Administration; 2014. URL: https://www.ncbi.nlm.nih.gov/books/NBK207185 [accessed 2025-11-10]
42. Danaher TS, Berry LL, Howard C, Moore SG, Attai DJ. Improving how clinicians communicate with patients: an integrative review and framework. J Serv Res 2023 Nov;26(4):493-510. [doi: 10.1177/10946705231190018]
43. Lu S. Loss of face-to-face interaction in telehealth [dissertation]. : Electronic Theses, Projects, and Dissertations; 2022 May 1 URL: https://scholarworks.lib.csusb.edu/etd/1439 [accessed 2025-11-10]
44. Wright HH, O'Shea MC, Sekula J, Mitchell LJ. Assessment of communication skills using telehealth: considerations for educators. Front Med 2022;9:841309. [doi: 10.3389/fmed.2022.841309]
45. Markowitz JC, Milrod B, Bleiberg K, Marshall RD. Interpersonal factors in understanding and treating posttraumatic stress disorder. J Psychiatr Pract 2009 Mar;15(2):133-140. [doi: 10.1097/01.pra.0000348366.34419.28] [Medline: 19339847]
46. Stieger S, Lewetz D, Willinger D. Face-to-face more important than digital communication for mental health during the pandemic. Sci Rep 2023 May 17;13(1):8022. [doi: 10.1038/s41598-023-34957-4] [Medline: 37198196]
47. An J, Kwak H, Qureshi HM, Weber I. Precision public health campaign: Delivering persuasive messages to relevant segments through targeted advertisements on social media. JMIR Form Res 2021 Sep 24;5(9):e22313. [doi: 10.2196/22313] [Medline: 34559055]
48. Susser D. Ethical considerations for digitally targeted public health interventions. Am J Public Health 2020 Oct;110(S3):S290-S291. [doi: 10.2105/AJPH.2020.305758] [Medline: 33001734]

Abbreviations

STI: sexually transmitted infection

Edited by A Stone, A Mavragani; submitted 20.06.25; peer-reviewed by JL Karim, R Asnel, S Meherali; revised version received 13.09.25; accepted 15.09.25; published 19.11.25.

Please cite as:
Abdulai AF, Josephs JC, Guta A, Gilbert M, Bungay V
Technology-Related Trauma in Sexual and Reproductive Health Digital Technologies: Grounded Theory Study
JMIR Form Res 2025;9:e79365
URL: https://formative.jmir.org/2025/1/e79365
doi:10.2196/79365

© Abdul-Fatawu Abdulai, Janell Cosco Josephs, Adrian Guta, Mark Gilbert, Vicky Bungay. Originally published in JMIR Formative Research (https://formative.jmir.org), 19.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Compliance and Satisfaction With a Protocol for Identifying Novel Targets to Support Postpartum Opioid Use Disorder Recovery: Prospective Cohort Study

Alicia M Allen[1,2,3], PhD, MPH; Linnea B Linde-Krieger[3,4], PhD, LCSW; Jendar Deschenes[3], MPH; Stephanie Mallahan[5], PhD, MPH; Alexandra Harris[3], MLS; Mariana Felix[3], MPH; Arushi Chalke[6], MS; Alma Anderson[3], MA; Priyanka Sharma[3], MDS (DPH), MBA; Katherine M King[7], PhD; Maddy T Grant[8], PhD; James Baurley[9], PhD; Lela Rankin[4], PhD; Stacey Tecot[7], PhD

[1]University of Arkansas for Medical Sciences, 4301 West Markham Street, Little Rock, AR, United States
[2]Arkansas Children's Research Institute, Little Rock, AR, United States
[3]Department of Family and Community Medicine, University of Arizona, Tucson, AZ, United States
[4]School of Social Work Tucson, Arizona State University, Tucson, AZ, United States
[5]Clinical Translational Sciences, University of Arizona, Tucson, AZ, United States
[6]Epidemiology and Biostatistics, University of Arizona, Tucson, AZ, United States
[7]Laboratory for the Evolutionary Endocrinology of Primates, School of Anthropology, University of Arizona, Tucson, AZ, United States
[8]New York Consortium in Evolutionary Primatology, The Graduate Center, CUNY, New York, NY, United States
[9]Biorealm LLC, Monument, CO, United States

Corresponding Author:
Alicia M Allen, PhD, MPH
University of Arkansas for Medical Sciences, 4301 West Markham Street, Little Rock, AR, United States

Abstract

Background: Although treatment for opioid use disorder (OUD) often yields high adherence during pregnancy, the risk of returning to opioid misuse during postpartum is high. There are currently no relapse prevention programs tailored to this unique time period. Using a prospective cohort study, we seek to preliminarily identify hormones or infant caregiving approaches as novel predictors of postpartum opioid misuse.

Objective: As a first step in dissemination of results, this report contains a detailed account of the protocol, as well as recruitment, retention, compliance, and participant satisfaction.

Methods: Participants were individuals with OUD (OUD+) and those without (OUD–) who were followed from late pregnancy (≥36 gestational wk) to postpartum month 5. From childbirth to postpartum week 12, participants completed daily surveys (capturing use, craving, interactions with infant) and weekly face-to-face visits (including collection of biological samples for hormone assays). Follow-up visits using the same procedures occurred at postpartum month 4 and 5.

Results: Most participants (50 OUD+, 20 OUD–) notified the study staff of childbirth (n=63, 93%), completed at least 1 postpartum clinic visit (n=62, 87%), and completed follow-up (n=51, 73%). Compliance with procedures ranged from 81% for weekly surveys to 63% for weekly dried blood spots, generally with lower compliance among OUD+ and at later time points. Among a subgroup of participants (n=31), regardless of group and time point, reported high study satisfaction (eg, on a scale where 0 is "not at all" and 3 is "extremely," on average, participants reported 2.9±0.4 for their willingness to complete this study again at week 12 postpartum).

Conclusions: This prospective cohort study was well tolerated despite the challenging postpartum period. Data collected will provide ample opportunities to identify novel risk or protective factors to inform the development of new relapse prevention intervention programs specific to the needs of those with OUD during early postpartum.

(JMIR Form Res 2025;9:e77899) doi:10.2196/77899

KEYWORDS

cohort study; ecological momentary assessment; hormones; opioid use disorder; postpartum period; pregnancy

XSL•FO
RenderX

Introduction

Perinatal opioid use disorder (OUD) is increasingly common with most recent estimates indicating that up to 324 per 10,000 births are impacted [1-3]. Medication for OUD (MOUD) is the gold standard treatment, improving both maternal and infant outcomes [4-6]. However, the risk for return to opioid misuse and overdose increases during postpartum [7,8]. In fact, opioid overdose is now a leading cause of maternal mortality [9], with additional significant adverse impacts on the birthing person, infant, family, and beyond [3,10,11]. To address this period of high risk, the American College of Obstetricians and Gynecologists and the American Society of Addiction Medicine recommends that "*Access to…relapse prevention programs should be made available*" [6]. However, there are currently no evidence-based relapse prevention programs tailored to the unique postpartum circumstances [12]. Additionally, factors that contribute to the increased risk of return to use during postpartum are not well understood [9]. Ultimately, developing effective relapse prevention programs that can work adjunctive to MOUD and are tailored to postpartum-specific needs will allow parents, infants, and families enhanced opportunities to achieve positive long-term outcomes.

Ovarian hormones have been implicated as important factors for clinical treatment of substance use disorders [13-15]. For instance, in ovariectomized female rats offered intermittent access to fentanyl, those who received exogenous estradiol treatment exhibited increases in self-administration, enhanced sensitivity, and increased motivation versus placebo treatment [16]. Exogenous progesterone treatment is associated with reductions in combustible cigarette smoking during the postpartum period, as well as in nonpregnant females [17-19]. Hormonal influences on substance misuse behaviors may be amplified during the perinatal period, as hormones change by up to 300-fold during pregnancy, followed by a dramatic decline in the early postpartum and additional variation with lactation [20]. Interestingly, some of the natural hormonal patterns that occur during the perinatal period directly mirror the risk of substance misuse. For instance, both estrogen (which facilitates drug-taking behaviors) and risk of substance misuse are low during lactation, whereas without lactation, both estrogen and risk of substance misuse increase [20,21]. Further, specific infant caregiving activities can have substantial effects on hormones (eg, oxytocin increases during skin-to-skin contact), which may directly or indirectly impact postpartum substance relapse risk [22-25]. Despite the observed links between ovarian hormones and substance misuse as well as infant caregiving approaches and ovarian hormones, paired with the dramatic perinatal hormonal variations, the specific examination of postpartum hormonal patterns with OUD recovery-related outcomes has yet to be explored.

We conducted a prospective cohort study with those who did and did not have perinatal OUD to examine hormones or infant caregiving activities as modifiable risk factors that future interventions may target to enhance protection against return to opioid misuse. In order to specifically study postpartum relapse risk, we recruited participants with OUD who were stable in recovery during pregnancy, as well as comparable control group participants without OUD, and followed them across the high-risk postpartum period. As a first step in dissemination, this paper aims to (1) detail our protocols and methodologies and (2) report on recruitment, retention, compliance, and participant satisfaction. Ultimately, the goal of this paper is to enhance the understanding of perinatal OUD experience and inform the development of novel interventions and treatment options to support postpartum recovery and well-being.

Methods

Ethical Considerations

This study was approved by the University of Arizona's Institutional Review Board (2009060851R001). Study staff met with all participants to obtain informed consent prior to their participation in the study. To maintain privacy and confidentiality, all data were stored with an alphanumeric identifier on secured platforms, only accessible by study staff. Participants were compensated up to US $725.

Recruitment

In this prospective cohort study, we sought to enroll 50 pregnant people with OUD (OUD+) and 25 pregnant people without OUD (OUD–). We opted for a prospective cohort study design with frequent measurements to allow for the assessment of numerous modifiable risk factors (eg, estradiol, skin-to-skin contact). Specifically, with a primary sample of 75 independent participants who completed daily measurements across the first 12 weeks postpartum, this longitudinal study was designed to yield approximately 70,000 data points, providing adequate statistical power for within-person analyses and robust estimation of individual trajectories. It is noted that this sample size has limited power for between-group comparisons. However, the inclusion of OUD– will allow us to explore hormonal or infant caregiving approaches that are unique to OUD+. Overall, this approach will achieve the primary goal of preliminarily identifying meaningful relationships between hormones or infant caregiving with return to opioid misuse that can be further explored in future fully powered randomized control trials.

Participants met the following inclusion criteria: (1) 18 - 40 years old, (2) uncomplicated single-gestation pregnancy at gestational week 30 or beyond, (3) a self-reported expectation of residing with the infant after birth, (4) English fluency, and (5) willing and able to comply with procedures. Additionally, OUD+ had to (6) report use of opioids during pregnancy, (7) report current OUD treatment with plans to continue treatment after childbirth, and (8) allow us access to their OUD treatment records. All participants were excluded for preterm birth (<36 gestational wk), planned long-distance moves, or extended travel through 6 months postpartum. Additional OUD– exclusion criteria included the use of opioids during pregnancy or a history of opioid use disorder. Finally, upon enrollment, we utilized frequency matching at the group level with a goal of ensuring the groups were comparable on proportion with no health insurance or public insurance versus private insurance, as well as on average age (ie,±5 y).

XSL•FO
RenderX

We recruited participants in the metro areas of Tucson and Phoenix, Arizona, from August 2021 to March 2024 via flyers or tabling at local health care provider clinics, community resources, community events, and social media. Individuals expressed their interest via a survey on REDCap [26], and this was followed by a telephone-based eligibility interview with staff. If eligibility criteria were met, potential participants were invited to a face-to-face baseline visit to commence participation.

Data Collection Procedures

We implemented a data collection protocol similar to one we previously used with success in a pilot study focused on postpartum cigarette smoking relapse prevention [17]. Participants completed 15 visits either in-person (at our clinic or the participant's home) or remotely via Health Insurance Portability and Accountability Act–compliant Zoom for Health, with all data entered on REDCap [26]. Visits were scheduled to begin between 7 AM and 11 AM given the known diurnal patterns in hormones and other variables of interest [27,28]. For those who attended remotely, study supplies were mailed in advance. All visits included the collection of biological samples, with all participants provided 1.8 mL of saliva via passive drool. Remote participants self-collected 12 drops of blood for dried blood spots using our previously developed protocol [29]. In-person participants provided 8 mL of venous blood, which was then processed to yield 12 drops of dried blood spots plus plasma samples. For remote participants, dried blood spots were mailed to staff weekly, and frozen saliva samples were picked up in batches by staff. All samples were subsequently stored at ≤−20 °C.

Baseline

At gestational week 36 or beyond, participants completed a baseline visit. After providing informed consent, participants completed an interview with staff to obtain their medical history, followed by completion of surveys and biological sample collection. Next, participants completed a semistructured audio-recorded interview with staff on lifetime trauma and resilience (Multimedia Appendix 1). At the end of the visit, participants were trained on how to complete daily surveys and at-home saliva samples, followed by receipt of compensation.

Daily Surveys

Beginning the day after baseline through postpartum week 12, participants completed daily surveys on their own or study-supplied smartphones. This was replicated the week preceding follow-up visits. Participants received a daily text message or email at 8 PM containing a link to surveys. Surveys remained open until 11:59 PM, with reminders every 30 minutes until completed. Surveys took approximately 10 - 15 minutes to complete and documented use of and craving for substances, interactions, and connectedness with the infant and others, and other validated measures to capture return to use risk factors.

Fourth Trimester (Postpartum Weeks 1 - 12)

Participants notified staff of childbirth via the daily surveys, then commenced weekly visits. At each visit, participants completed surveys and provided in-visit biological samples. Additionally, at-home saliva samples were collected at 8 PM the day before and 30 minutes after waking on the day of the visit. Participants were also interviewed to capture substance use, changes in health or medication, and time spent with the infant. At postpartum week 1, participants completed a semistructured audio-recorded interview with staff to capture birth and initial breastfeeding experiences (Multimedia Appendix 2). Approximately 1 year into study recruitment, we added a study satisfaction survey to postpartum weeks 1 and 12. Upon completion of each visit, the next visit was confirmed, and participants were compensated.

Follow-Up Visits (Postpartum Months 4 and 5)

Procedures were identical to the fourth trimester data collection. In addition, at postpartum month 5, participants completed a semistructured audio-recorded interview with staff to capture their breastfeeding experiences and a final satisfaction survey.

Study Measures

Overall, participants completed a series of instruments on a daily, weekly, and monthly basis, as well as at follow-up (Multimedia Appendices 3 and 4).

Background Variables

During the enrollment interview and at the baseline visit, standardized questionnaires and interviews were used to capture sociodemographics (eg, age, race, ethnicity), medical history (eg, parity, diagnosis of hormone-influencing conditions), and history of substance use (eg, lifetime use of 16 different substances, drug of choice, OUD treatment experiences). The baseline interview incorporated Stressful Life Events [30], Adverse Childhood Experiences [31], Early Trauma Inventory [32,33], and Resilience Scale [34].

Maternal Factors

These variables included (1) known or suspected risk factors for postpartum substance misuse, (2) known or suspected association with hormones of interest, and (3) infant caregiving and parental experiences [35-51]. In brief, participants reported on lactation and breastfeeding (daily first 4 weeks, weekly thereafter), pain (daily first 4 weeks), and vaginal bleeding (daily after postpartum week 4), with additional assessments on a weekly or monthly basis.

Caregiving Factors

Variables included those that may contribute to risk of postpartum return to misuse or influence hormone levels. Daily, participants reported on their perceptions of parenting challenge, parenting reward, and connections with infant and with others using a 100-point visual analog scale ranging from "not at all" to "extremely." Relative time spent with infant each day was reported using an investigator-created instrument with 100-point visual analog scale ranging from "a lot less than usual" to "a lot more than usual" to report on the amount of time spent: (1) in skin-to-skin contact with infant, (2) in other physical contact with infant, (3) caring for infant but not in physical contact, (4) with someone else while they cared for infant, and (5) time away from infant. Weekly staff interviewed participants to capture the average absolute time spent on each of these activities during the preceding week, ensuring that total time

XSL•FO
RenderX

reported equated to 24 hours per day. Additional validated items were completed regularly throughout follow-up.

Hormones

We measured a total of 19 hormones at each time point. Cortisol and oxytocin assays were completed by the Laboratory for the Evolutionary Endocrinology of Primates at the University of Arizona. Saliva samples were stored at −80 °C, thawed to room temperature before analysis, and cleaned by centrifuging for 30 minutes at 3000 rpm with the supernatant retained. Saliva was aliquoted into microcentrifuge tubes at volumes of 500 µL for oxytocin and 150 µL for cortisol, then frozen at −80 °C until lyophilization. Samples were lyophilized at −80 °C and 0.3 mbar until completely dry, then resuspended in 250 µL and 150 µL of assay buffer (Arbor Assays, Ann Arbor, MI, USA), respectively. This resulted in 2x concentration for oxytocin and 1x concentration for cortisol. Samples were assayed according to protocols included with DetectX ELISA oxytocin (K048-H5; cross-reactivity with isotocin 94.3%, mesotocin 88.4%, other analytes <0.15%; sensitivity 17.0 pg/mL; limit of detection 22.9 pg/mL) and DetectX ELISA cortisol kits (K003-H5; cross-reactivity with dexamethasone 18.8%, prednisolone [1-dehydrocortisol] 7.8%, corticosterone and cortisone 1.2%, other analytes <0.1%; sensitivity 27.6 pg/mL; limit of detection 45.4 pg/mL). Plates were read at 450 nm on a BioTek Multi-Mode plate reader. The Laboratory for the Evolutionary Endocrinology of Primates conducted analytical validations of methods for adult female human saliva using tests of parallelism and accuracy for each analyte. Parallelism was done by serially diluting a sample pool, correcting the resulting concentrations at each dilution, and calculating the coefficient of variation [52]. The coefficient of variation of corrected concentrations from serially diluted saliva was 11.65% for oxytocin (n=5 dilutions) and 8.34% for cortisol (n=10 dilutions). For accuracy, sample pools were spiked with 10% synthetic oxytocin or cortisol in assay buffer, at different concentrations along the standard curve. Recovery for salivary oxytocin was 119.99%±0.05, and recovery for salivary cortisol was 102.18%±2.31. All samples were assayed in duplicate, and all plates included high and low pool controls to assess inter- and intra-assay variation. Any samples with concentrations that were outside of the standard curve range were diluted or concentrated and reassayed. Data were processed using the BioTek Gen5 software (version 3.08; Aligent Technologies) to calculate analyte concentrations.

Using dried blood spots, ZRT Laboratory (Beaverton, OR) measured cortisol, estrone, estradiol, estriol, testosterone, progesterone, DHEAS, cortisone, estrone-1-sulfate, pregnenolone sulfate, 17-hydroxyprogesterone, androstenedione, 7-keto DHEA, corticosterone, 11-deoxycortisol, ethinyl estradiol, anastrozole, and letrozole as previously described with some modifications [53]. Three 6-mm diameter punches were taken from each dried blood spot specimen and rehydrated. Steroids were extracted with organic solvent in the presence of internal standard, purified by solid-phase extraction, derivatized (estrogens only), and then analyzed by liquid chromatography-mass spectrometry. Liquid chromatography-mass spectrometry analysis was performed using ultra-fast liquid chromatography system (Shimadzu Nexera XR) coupled to a triple quadrupole mass spectrometer (AB Sciez 5500) equipped with an electrospray ionization source. Separations were conducted by reversed-phase chromatography, and the mass spectrometer was operated in multiple reaction monitoring mode. Data were processed using the Sciex OS software (version 1.7.2; AB Scienx) to determine analyte concentrations. All samples were run with dried blood spot controls, generated in-house by spiking stripped serum with various analyte concentrations, mixing washed red blood cells, and then spotting onto filter paper.

Factors Related to Substance Use

To assess postpartum risk for opioid misuse, we included 3 outcomes. First, daily self-reports of opioid cravings (separately for opioid-containing prescription medications, heroin, and MOUD medications) and coping with cravings. Both cravings and coping with craving were reported with a 100-point visual analog scale ranging from “not at all” to “severe” or “very well,” respectively. Second, daily self-reports of opioid use were captured via daily surveys and staff-administered TimeLine FollowBack [54] interviews at each visit. These approaches were also used to capture craving and use of 15 other substances. Finally, we reviewed toxicology results and treatment adherence via treatment program medical records.

Statistical Analyses

Study sample descriptors (eg, sociodemographics, substance use) and outcomes were summarized with descriptive statistics. Group differences (ie, OUD+ vs OUD–) were compared using Welch *t* tests, χ^2 tests, and Fisher exact tests. For the latter, although some cells contained relatively large counts (up to 45), exact *P* values were computed without the need for permutations. Group differences in compliance over time were also examined with binomial logistic regression models. All analyses were completed using open-source Python (3.10.11) libraries, Pandas 2.0.1 (NumFOCUS, Inc.), SciPy 1.13.0, and R Studio (2024.09.1, Build 394).

Results

Study Sample

A total of 70 participants enrolled in the study and completed the baseline assessment to form the final sample (Figure 1; note that these numbers exclude 1 participant who enrolled into the OUD– and later requested her data be removed). The final enrolled sample included 50 (71%) OUD+ and 20 (29%) OUD– (Table 1). Study groups were exchangeable on all sociodemographic variables except for gestational week prenatal care was initiated (OUD+: 9.8 [SD 5.2] vs OUD–: 7.1 [SD 3.8]; *P* value=.02). There was little evidence of selection bias from phone screening or enrollment to the postpartum data collection period (Multimedia Appendices 5 and 6). For example, the average age at enrollment was 28.6 (SD 5.2; n=112) compared to 29.4 (SD 5.1) among those who completed at least 1 postpartum visit (n=61). Additionally, 52% (n=58) of participants who completed the phone screening identified as Hispanic and 37% (n=41) reported education of some college compared to 51% (n=31) and 41% (n=25), respectively, at postpartum data collection. Among OUD+, at enrollment, 58% (n=50) reported being in recovery for less than a year compared

XSL•FO
RenderX

to 56% (n=43) at postpartum data collection, with an age of first use of heroin reported as 17.3 (SD 5.7) years old (n=47) versus 17.5 (SD 5.9) years old (n=41), respectively.

Figure 1. Consolidated Standards of Reporting Trials (CONSORT) diagram: study flow, time points, and participants. OUD: opioid use disorder.

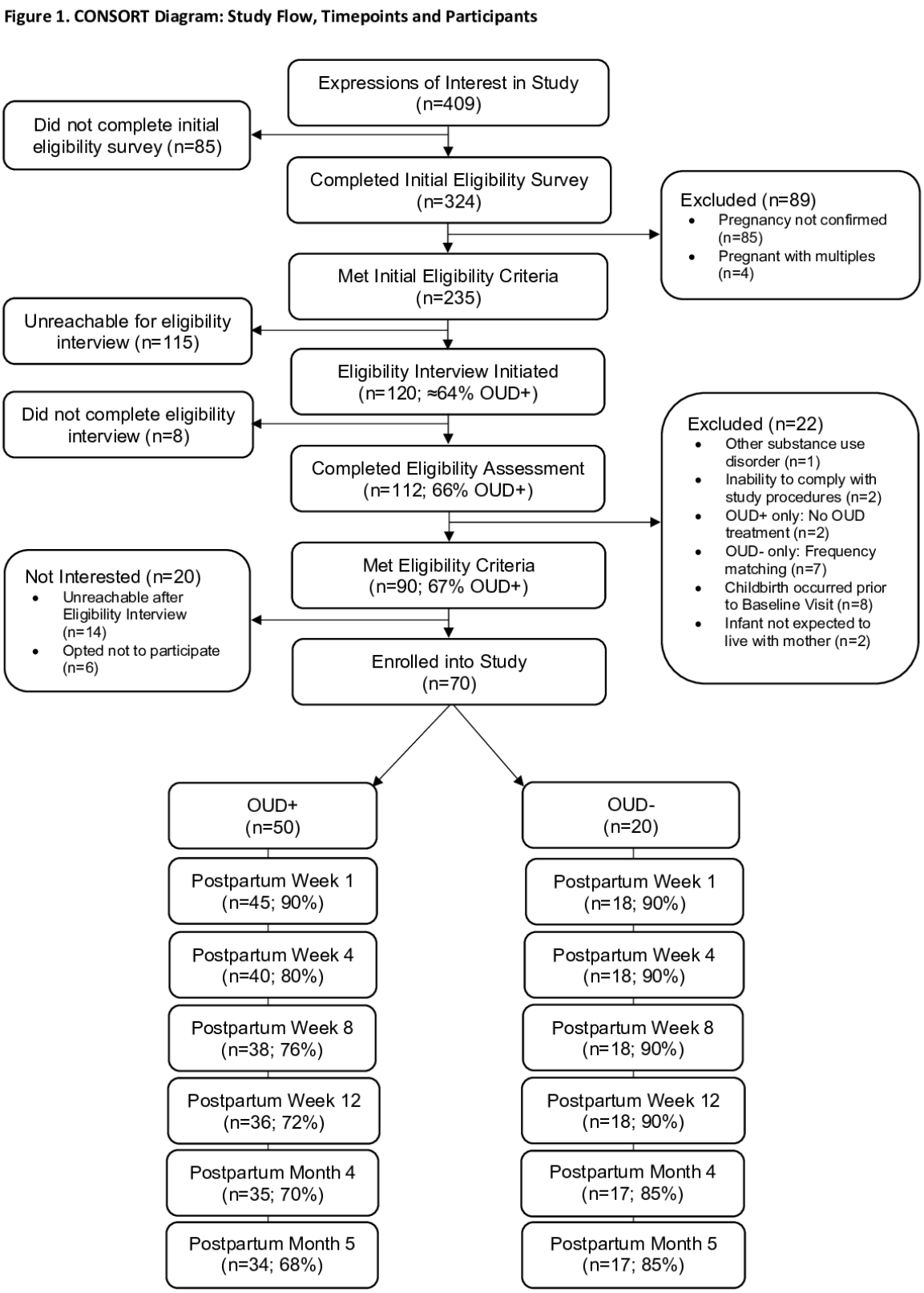

XSL•FO
RenderX

Table . Description of final study sample (n=70).

	Total (n=70)	OUD+ (n=50)	OUD– (n=20)	Test statistic[a] (*P* value)
Sociodemographic variables				
Age (y), mean (SD)	29.3 (5.0)	30.0 (4.9)	27.8 (5.1)	1.63 (.11)
Race and ethnicity, n (%)				3.18 (.53)
Hispanic	33 (47)	21 (42)	12 (60)	
NH[b], White	28 (40)	21 (42)	7 (35)	
NH, NA/AN[c]	4 (6)	3 (6)	1 (5)	
NH, NH/PI[d]	0 (0)	0 (0)	0 (0)	
NH, Asian	1 (1)	1 (2)	0 (0)	
NH, B/AA[e]	4 (6)	4 (8)	0 (0)	
Highest level of education completed, n (%)				9.11 (0.10)
≤8th grade	2 (3)	2 (4)	0 (0)	
Some HS[f]	8 (11)	6 (12)	2 (10)	
HS or equivalent	25 (36)	19 (38)	6 (30)	
Some college or 2-year degree	30 (43)	22 (44)	8 (40)	
College graduate or 4-year degree	3 (4)	0 (0)	3 (15)	
Graduate or professional degree	2 (3)	1 (2)	1 (5)	
Insurance status, n (%)				2.85 (.09)
Private	6 (9)	2 (4)	4 (20)	
Public or none	64 (91)	48 (96)	16 (80)	
Missing	0 (0)	0 (0)	0 (0)	
Number of children living in home, mean (SD)	1.1 (1.4)	1.2 (1.6)	1.0 (0.9)	0.42 (.68)
Gestational week prenatal care initiated, mean (SD)	9.1 (5.0)	9.8 (5.2)	7.1 (3.8)	2.38 (.02)
Parity, n (%)				0.002 (.97)
Primiparous	19 (27)	13 (26)	6 (30)	
Multiparous	51 (73)	37 (74)	14 (70)	
Gestational week at baseline visit, mean (SD)	36.8 (0.8)	36.9 (0.9)	36.8 (0.6)	0.32 (.75)
Substance use history				
Lifetime history of use of substances with abuse potential[g], n (%)				—[h]
Opioid-containing prescription	58 (83)	48 (96)	10 (50)	
Heroin	30 (43)	30 (60)	0 (0)	
Narcotics	47 (67)	47 (94)	0 (0)	
Fentanyl	6 (9)	6 (12)	0 (0)	
Caffeine	68 (97)	48 (96)	20 (100)	
Alcohol	60 (86)	43 (86)	17 (85)	

XSL•FO
RenderX

	Total (n=70)	OUD+ (n=50)	OUD– (n=20)	Test statistic[a] (*P* value)
Cigarettes or nicotine	56 (80)	43 (86)	13 (65)	
Cannabis	56 (80)	45 (90)	11 (55)	
Cocaine	44 (63)	41 (82)	3 (15)	
Use in 3 months prior to pregnancy of substances with abuse potential[g], n (%)				—
Opioid-containing prescription	18 (26)	18 (36)	0 (0)	
Heroin	3 (4)	3 (6)	0 (0)	
Narcotics	29 (41)	29 (58)	0 (0)	
Fentanyl	4 (6)	4 (8)	0 (0)	
Caffeine	66 (94)	46 (92)	20 (10)	
Alcohol	27 (39)	12 (24)	15 (75)	
Cigarettes or nicotine	44 (63)	34 (68)	10 (50)	
Cannabis	38 (54)	31 (62)	7 (35)	
Cocaine	9 (13)	7 (14)	2 (10)	
Use in last 3 months of pregnancy of substances with abuse potential[g], n (%)				—
Opioid-containing prescription	10 (1)	10 (20)	0 (0)	
Heroin	0 (0)	0 (0)	0 (0)	
Narcotics	41 (59)	41 (82)	0 (0)	
Fentanyl	2 (3)	2 (4)	0 (0)	
Caffeine	63 (90)	44 (88)	19 (95)	
Alcohol	2 (3)	2 (4)	0 (0)	
Cigarettes or nicotine	32 (46)	28 (56)	4 (20)	
Cannabis	16 (23)	15 (30)	1 (5)	
Cocaine	1 (1)	1 (2)	0 (0)	
Drug of choice (top 5), n (%)				—
Other opioid-containing prescription	—	18 (36)	—	
Heroin	—	11 (22)	—	
Fentanyl	—	6 (12)	—	
Cannabis	—	5 (10)	—	
Caffeine	—	4 (8)	—	
Age of first opioid-containing prescription use, mean (SD)	—	17.3 (5.7) (n=47)	—	—
Age of first heroin use, mean (SD)	—	22.4 (5.7) (n=30)	—	—
Lifetime history treatment type, n (%)				—
Inpatient	—	30 (60)	—	
Outpatient	—	33 (66)	—	

XSL•FO
RenderX

	Total (n=70)	OUD+ (n=50)	OUD– (n=20)	Test statistic[a] (*P* value)
Intensive outpatient	—	27 (54)	—	
Current treatment type, n (%)				—
Inpatient	—	2 (4)	—	
Outpatient	—	30 (60)	—	
Intensive outpatient	—	5 (10)	—	
Current treatment components, n (%)				—
Medications	—	34 (68)	—	
Counseling/support groups	—	24 (48)	—	
Other	—	3 (6)	—	
Recovery length, n (%)				—
>1 year	—	21 (42)	—	
Before this pregnancy but <1 year	—	9 (18)	—	
First trimester of this pregnancy	—	6 (12)	—	
Second trimester of this pregnancy	—	10 (20)	—	
Third trimester of this pregnancy	—	4 (8)	—	

[a]Test statistic value listed is χ^2 or *t* value.

[b]NH: Non-Hispanic.

[c]NA/AN: Native American or Alaskan Native.

[d]NH/PI: Native Hawaiian or Pacific Islander.

[e]B/AA: Black or African American.

[f]HS: high school.

[g]Substances listed include opioids and the 5 most commonly endorsed substances used per lifetime history.

[h]—: not applicable.

Protocol Compliance

Nearly all participants (n=65, 93%) reported childbirth to staff (OUD+: 92%, OUD–: 95%), and 62 (87%) completed at least 1 study visit post-childbirth (OUD+: 86%, OUD–: 90%). Visit completion rates post-baseline ranged from 90% at postpartum week 1% to 73% at postpartum month 5 (Figure 2; Multimedia Appendix 7), with lower rates in the OUD+ group and at later time points. Compliance with procedures followed similar patterns (lower in OUD+ and at later time points), with the highest overall compliance, on average, with the weekly surveys (81%), followed by interviews (74%), saliva samples (66%), daily surveys (66%), and dried blood spots (63%). While there were no differences in the decline in compliance over time by group by procedure (interviews [β=.04, *P*=.40], dried blood spots [β=.04, *P*=.22], saliva samples [in visit: β=.01, *P*=.89; outside of visit: β=−.01, *P*=.82], nor surveys [daily: β=−.01, *P*=.80; weekly: β=−.04, *P*=.10]), OUD+ visit attendance declined more rapidly over time compared to OUD– (β=−.06, *P*=.03).

XSL•FO
RenderX

Figure 2. Completion rates by study group, time point, and procedure: (a) visit completion by study group, (b) compliance with in-visit procedure over time by modality and group, and (c) compliance with out-of-visit procedure over time by modality and group. OUD: opioid use disorder.

XSL•FO
RenderX

Study Satisfaction

Among the subgroup of participants who completed the study satisfaction surveys, results indicate that the study was well tolerated over time with no significant differences by study groups (Table 2).

XSL•FO
RenderX

Table . Study satisfaction by study group and time point[a].

	Total	OUD+	OUD–	*P* value[b]
Sample size by time point[c]				
W1[d]	20	17	3	—[e]
W12[f]	31	25	6	—
M5[g]	45	32	13	—
Based on your experience, how willing would you be to participate in a study like this again? Mean (SD)				
W1	2.6 (0.8)	2.6 (0.9)	2.7 (0.6)	.60
W12	2.9 (0.4)	2.8 (0.5)	3.0 (0.0)	>.99
M5	2.7 (0.8)	2.7 (0.8)	2.7 (0.6)	.56
Overall, how much of a burden has it been to participate in this study? Mean (SD)				
W1	0.6 (0.8)	0.6 (0.9)	0.3 (0.6)	>.99
W12	0.5 (0.8)	0.6 (0.9)	0.3 (0.5)	>.99
M5	0.4 (0.7)	0.4 (0.8)	0.4 (0.6)	>.99
How much of a burden was it to participate in this study before your baby was born? Mean (SD)				
W12	0.2 (0.6)	0.2 (0.6)	0.0 (0.0)	>.99
How much of a burden was it to participate in this study after your baby was born? Mean (SD)				
W12	0.6 (0.8)	0.6 (0.9)	0.5 (0.5)	.51
Overall, how much has participating in this study interfered with your usual activities? Mean (SD)				
W1	0.7 (0.8)	0.7 (0.8)	0.3 (0.6)	>.99
W12	0.6 (0.8)	0.5 (0.8)	1.0 (1.1)	.38
How much did participating before your baby was born interfere with your usual activities? Mean (SD)				
W12	0.2 (0.6)	0.3 (0.7)	0.2 (0.4)	>.99
How much did participating after your baby was born interfere with your usual activities? Mean (SD)				
W12	0.7 (0.8[h])	0.7 (0.9[h])	0.7 (0.5)	.53
How much difficulty did you have completing the study questionnaires? Mean (SD)				
W1	0.3 (0.7)	0.4 (0.8)	0.0 (0.0)	>.99
How much difficulty did you have using a cell phone to complete the study questionnaires? Mean (SD)				
W12	0.1 (0.4)	0.1 (0.4)	0.0 (0.0)	>.99
How much have you been annoyed with the number of alerts you have received in this study before your baby was born? Mean (SD)				
W12	0.2 (0.5[h])	0.3 (0.5)	0.0 (0.0[h])	.63
How much have you been annoyed with the number of alerts you have received in this study after your baby was born? Mean (SD)				
W12	0.3 (0.7)	0.4 (0.8)	0.2 (0.4)	>.99
How difficult was it to notify the study staff of your child's birth? Mean (SD)				
W1	0.1 (0.7)	0.2 (0.7)	0.0 (0.0)	>.99
Have you had any difficulties working with study staff? Mean (SD)				
W1	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	—
W12	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	—
M5	0.0 (0.2)	0.1 (0.2)	0.0 (0.0)	>.99

[a]Values are as follows: 0: not at all, 1: slightly, 2: moderately, 3: extremely.

[b]Fisher exact test.

[c]Adding percentages would cause confusion (eg, it could imply percent of full sample, percent at time point, or percent of group). Hence, the sample size is listed without percentages in the table.

XSL•FO
RenderX

[d]W1: postpartum week 1.

[e]—: not applicable.

[f]W12: postpartum week 2.

[g]M5: postpartum month 5.

[h]One participant did not respond to this item.

Discussion

Principal Results

We completed a prospective cohort study with individuals who did and did not have perinatal OUD, following them with robust data collection procedures. This included daily surveys and weekly face-to-face visits with collection of biological samples during the first 12 weeks postpartum with additional follow-up through postpartum month 5. We enrolled 70 participants during late pregnancy, and 51 (73%) completed the follow-up through postpartum month 5. Notably, more participants in the OUD– group completed follow-up compared to the OUD+ group (85% vs 68%, respectively). Additionally, despite the challenges faced in caring for a newborn paired with our robust data collection protocol, compliance with specific data collection modalities was moderately high in both groups (63% - 81%). Indeed, the daily survey compliance of approximately 80% during the follow-up period among OUD– is directly in line with the, on average, 80% compliance rates (range: 63% - 96%) observed in 27 studies using similar procedures during the perinatal period [55]. The OUD+ group had lower compliance overall, throughout the follow-up period for the data collection procedures, as well as a faster drop in compliance over time with regard to visit attendance. This is to be expected given the additional demands on the OUD+ participants (eg, caring for a high-needs infant, daily MOUD clinic visits). Moreover, during the first year of data collection, all participants faced additional challenges with the ongoing COVID-19 pandemic. Despite this, among a subset of participants, high levels of satisfaction were reported in both groups consistently across time. While this data collection protocol was burdensome with daily surveys plus weekly meetings and hormonal measurements and, thus, likely includes some selection bias due to the nature of individuals who are able to participate in this research (eg, those with more support at home, those more stable in recovery), it does provide an opportunity to inform future research on relapse prevention development.

The primary goal of this prospective cohort study is to examine the potential utility of hormones or infant caregiving activities to support postpartum OUD recovery (publications forthcoming). While there is growing evidence that hormones (eg, progesterone, estrogen, oxytocin) influence drug-taking behaviors, nearly all research has examined the effect of a single hormone in isolation at 1 or 2 time points [13-15]. Emerging research demonstrates that relative levels and temporal patterns of hormones may be more important than a single hormone value at an isolated time point. For example, changes in the ratio of progesterone to estradiol across two time points are more predictive of smoking behavior than either hormone alone at a single time point [56]. Given that hormones (1) are continuously in flux, especially during the perinatal period, and variations occur as a result of many factors, including specific infant caregiving activities, and (2) have a significant effect on several aspects of neurobiology, including drug reward and stress responses, assessing the effects of 1 or 2 hormones at 1 or 2 time points is insufficient.

With the daily surveys and weekly hormonal measurements, we will be able to examine the role of hormones and infant caregiving practices on OUD recovery-related outcomes with novel analytical approaches to model complex variable interactions both concurrently and across time. Forthcoming analyses addressing this goal will follow recommendations to handle missing data in longitudinal datasets, including the use of multilevel modeling frameworks that use all available data across repeated measures and maximum likelihood and imputation methods for relatively small sample sizes (eg, full information maximum likelihood, joint multiple imputation) [57-59]. Further, we will employ sensitivity analyses to assess potential selection bias and mitigate error [60-63] to further approximate observations to the target population.

While this study may not produce generalizable results, it can provide direction for future research. For example, if our observations suggest that higher levels of progesterone are protective against relapse risk, exogenous progesterone may be used to favorably modify progesterone during the postpartum period to reduce relapse risk, as was recently done to reduce postpartum cigarette smoking relapse risk [17]. Another scenario may include our observations suggesting that more close contact with infants is protective against postpartum opioid craving, in line with our prior work demonstrating that babywearing is linked to reduced urges to use substances [56]. Consequently, a follow-up study may examine the results of maternal babywearing on postpartum recovery outcomes similar to recent studies with nurses and mothers caring for infants with neonatal abstinence syndrome [64,65]. A third example may be heightened estradiol or testosterone during postpartum may be linked to more cue-induced cravings. While the exogenous delivery of these hormones is not compatible with pregnancy or breastfeeding [66,67], future research may instead use innovative approaches to monitor hormone pattern(s) or level(s) remotely [68] to proactively identify periods of heightened risk leading to the deployment of additional supports. This approach could also be applied to remotely monitor infant caregiving via smart baby carriers [68]. In sum, the data resulting from the study protocol described here, while limited, will yield informative preliminary results that can be further examined via a priori hypothesis testing approaches (eg, randomized control trials) with fully powered and diverse study samples to yield generalizable results and inform the development of additional postpartum OUD recovery support.

Beyond our primary goals, these data provide ample opportunities to explore novel topics in perinatal substance misuse and beyond. For example, we conducted comprehensive interviews of lifetime trauma exposure and resilience, as well

XSL•FO
RenderX

as birth and breastfeeding experiences. This allows for mixed-method approaches to expand the understanding of these experiences prospectively during the perinatal period. Additionally, daily our participants reported on their mood, affect, and stress, as well as their craving for and use of other substances (eg, alcohol, cannabis, nicotine, or tobacco). Thus, we can explore the relationships between these variables, as well as in relation to baseline characteristics (eg, ethnicity, parity) and concurrent behaviors and activities (eg, sleep, involvement of others). Indeed, we are currently exploring the role of loneliness, isolation, and social support in OUD-related outcomes within this dataset (R21DA058364; Allen and Linde-Krieger). Overall, these data offer substantial opportunities to uncover novel relationships, increase our understanding of risk or protective factors of maladaptive behaviors during the sensitive perinatal period, and ultimately inform new approaches to enhance the health and well-being of mothers, infants, and families.

Limitations

Like every study, there are limitations. First, generalizability is limited given that these data were collected in a single US state, and the robust protocol may have been too burdensome for some individuals. Similarly, our eligibility criteria restricted our OUD+ group to those who were in treatment for recovery and expected to live with their infant. It is also possible that those who had lower cravings or better coping were less likely to discontinue their study participation. We also have a fair amount of missing data, as is common in prospective studies during the postpartum period with daily measurements [55]. These data may not be missing at random (eg, those with a high-needs newborn or those at higher risk for return to opioid misuse may be less likely to complete data collection; retention may have changed across the study due to the COVID-19 pandemic or protocol adjustments). It is possible that this will introduce bias and error into our observations. Finally, this study is underpowered to test hypotheses using traditional analytical approaches for causal modeling. However, our goal here is to generate new preliminary observations that will lead to fully powered trials that can adequately assess these relationships with diverse study samples and that utilize novel approaches to prevent opioid misuse. Despite these limitations, there are also substantial strengths, including an ability to temporally evaluate relationships; assessment of substance use with multiple sources (eg, prospective surveys, retrospective interviews); and numerous validated measures with rigorous, gold-standard approaches.

Conclusions

Overall, this prospective cohort study of those with and without OUD during the perinatal period resulted in moderately high retention and compliance, as well as high participant-reported study satisfaction, among a subgroup of participants. Consequently, these multidimensional data have high potential to inform future research on the challenges and opportunities of the perinatal period to support OUD recovery during the traditionally high-risk postpartum period.

Acknowledgments

We extend our thanks to Drs. Mustafa al'Absi, Ann Bigelow, Ariadna Forray, Bob Handa, and Toni Ziegler for their input during the development of the initial data collection protocol. We also thank Emily Crose, Jeri Garfield, Rosalie Gordon, Kyndall Kleinman, Kristina Medvescek, McKenna Nelson, and Preeti Sahota for their participation in database development, recruitment, and data collection.

Funding

This project was funded by National Institutes of Health's Eunice Kennedy Shriver National Institute of Child Health & Human Development (DP2HD105541) and National Center for Advancing Translational Sciences (UL1TR000445).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Lifetime trauma and resilience interview.
[DOCX File, 24 KB - formative_v9i1e77899_app1.docx]

Multimedia Appendix 2
Investigator-created birth and breastfeeding interview.
[DOCX File, 20 KB - formative_v9i1e77899_app2.docx]

Multimedia Appendix 3
Protocol completion by time point, modality, and group.
[DOCX File, 48 KB - formative_v9i1e77899_app3.docx]

XSL•FO
RenderX

Multimedia Appendix 4
Investigator-created and modified versions of validated instruments.
[DOCX File, 22 KB - formative_v9i1e77899_app4.docx]

Multimedia Appendix 5
Overview of study measures.
[DOCX File, 18 KB - formative_v9i1e77899_app5.docx]

Multimedia Appendix 6
Description of study sample—completed phone screening interview (n=112).
[DOCX File, 24 KB - formative_v9i1e77899_app6.docx]

Multimedia Appendix 7
Description of study sample—completed at least 1 postpartum visit (n=61).
[DOCX File, 34 KB - formative_v9i1e77899_app7.docx]

References

1. Haight SC, Ko JY, Tong VT, Bohm MK, Callaghan WM. Opioid use disorder documented at delivery hospitalization—United States, 1999-2014. MMWR Morb Mortal Wkly Rep 2018 Aug 10;67(31):845-849. [doi: 10.15585/mmwr.mm6731a1] [Medline: 30091969]
2. Hirai AH, Ko JY, Owens PL, Stocks C, Patrick SW. Neonatal abstinence syndrome and maternal opioid-related diagnoses in the US, 2010-2017. JAMA 2021 Jan 12;325(2):146-155. [doi: 10.1001/jama.2020.24991] [Medline: 33433576]
3. Auty SG, Frakt AB, Shafer PR, Stein MD, Gordon SH. Severe maternal morbidity among pregnant people with opioid use disorder enrolled in Medicaid. JAMA Netw Open 2025 Jan 2;8(1):e2453303. [doi: 10.1001/jamanetworkopen.2024.53303] [Medline: 39777443]
4. Terplan M, Martin CE, Premkumar A, Krans EE. Caring for pregnant and parenting women with opioid use disorder. In: Wakeman SE, Rich JD, editors. Treat Opioid Use Disord Gen Med Settings: Springer International Publishing; 2021:203-221. [doi: 10.1007/978-3-030-80818-1_14]
5. Winstanley EL, Stover AN. The impact of the opioid epidemic on children and adolescents. Clin Ther 2019 Sep;41(9):1655-1662. [doi: 10.1016/j.clinthera.2019.06.003] [Medline: 31303278]
6. Opioid Use and Opioid Use Disorder in Pregnancy. URL: https://www.acog.org/clinical/clinical-guidance/committee-opinion/articles/2017/08/opioid-use-and-opioid-use-disorder-in-pregnancy [accessed 2024-11-23]
7. Schiff DM, Nielsen T, Terplan M, et al. Fatal and nonfatal overdose among pregnant and postpartum women in Massachusetts. Obstet Gynecol 2018 Aug;132(2):466-474. [doi: 10.1097/AOG.0000000000002734] [Medline: 29995730]
8. Correlates of treatment retention and opioid misuse among postpartum women in methadone treatment. URL: https://oce-ovid-com.ezproxy1.library.arizona.edu/article/01271255-201904000-00012/HTML [accessed 2024-11-23]
9. Frankeberger J, Jarlenski M, Krans EE, Coulter RWS, Mair C. Opioid use disorder and overdose in the first year postpartum: a rapid scoping review and implications for future research. Matern Child Health J 2023 Jul;27(7):1140-1155. [doi: 10.1007/s10995-023-03614-7] [Medline: 36840785]
10. Klaman SL, Isaacs K, Leopold A, et al. Treating women who are pregnant and parenting for opioid use disorder and the concurrent care of their infants and children: literature review to support national guidance. J Addict Med 2017;11(3):178-190. [doi: 10.1097/ADM.0000000000000308] [Medline: 28406856]
11. Piske M, Homayra F, Min JE, et al. Opioid use disorder and perinatal outcomes. Pediatrics 2021 Oct;148(4):e2021050279. [doi: 10.1542/peds.2021-050279] [Medline: 34479983]
12. Martinez A, Allen A. A review of nonpharmacological adjunctive treatment for postpartum women with opioid use disorder. Addict Behav 2020 Jun;105:106323. [doi: 10.1016/j.addbeh.2020.106323] [Medline: 32036191]
13. Anker JJ, Carroll ME. Females are more vulnerable to drug abuse than males: evidence from preclinical studies and the role of ovarian hormones. Curr Top Behav Neurosci 2011;8(73–96):73-96. [doi: 10.1007/7854_2010_93] [Medline: 21769724]
14. McHugh RK, Votaw VR, Sugarman DE, Greenfield SF. Sex and gender differences in substance use disorders. Clin Psychol Rev 2018 Dec;66:12-23. [doi: 10.1016/j.cpr.2017.10.012] [Medline: 29174306]
15. McKee SA, McRae-Clark AL. Consideration of sex and gender differences in addiction medication response. Biol Sex Differ 2022 Jun 27;13(1):34. [doi: 10.1186/s13293-022-00441-3] [Medline: 35761351]
16. Towers EB, Setaro B, Lynch WJ. Estradiol enhances the development of addiction-like features in a female rat model of opioid use disorder. Neuroendocrinology 2023;113(11):1099-1111. [doi: 10.1159/000529997] [Medline: 36878201]
17. Allen SS, Allen AM, Lunos S, Tosun N. Progesterone and postpartum smoking relapse: a pilot double-blind placebo-controlled randomized trial. Nicotine Tob Res 2016 Nov;18(11):2145-2153. [doi: 10.1093/ntr/ntw156] [Medline: 27613934]

XSL•FO
RenderX

18. Tosun NL, Fieberg AM, Eberly LE, et al. Exogenous progesterone for smoking cessation in men and women: a pilot double-blind, placebo-controlled randomized clinical trial. Addiction 2019 Oct;114(10):1800-1813. [doi: 10.1111/add.14645] [Medline: 31059177]
19. Forray A, Gilstad-Hayden K, Suppies C, Bogen D, Sofuoglu M, Yonkers KA. Progesterone for smoking relapse prevention following delivery: a pilot, randomized, double-blind study. Psychoneuroendocrinology 2017 Dec;86:96-103. [doi: 10.1016/j.psyneuen.2017.09.012] [Medline: 28926762]
20. Mishell DR. Reproductive Endocrinology, 3rd edition: Philadelphia. Current Medicine, Inc; 1999. URL: https://catalog.nlm.nih.gov/discovery/[illegible] [accessed 2025-11-14]
21. Forray A, Merry B, Lin H, Ruger JP, Yonkers KA. Perinatal substance use: a prospective evaluation of abstinence and relapse. Drug Alcohol Depend 2015 May 1;150(147–155):147-155. [doi: 10.1016/j.drugalcdep.2015.02.027] [Medline: 25772437]
22. Zeki S. The neurobiology of love. FEBS Lett 2007 Jun 12;581(14):2575-2579. [doi: 10.1016/j.febslet.2007.03.094] [Medline: 17531984]
23. Carter CS. Neuroendocrine perspectives on social attachment and love. Psychoneuroendocrinology 1998 Nov;23(8):779-818. [doi: 10.1016/s0306-4530(98)00055-9] [Medline: 9924738]
24. Kenkel WM, Paredes J, Yee JR, Pournajafi‐Nazarloo H, Bales KL, Carter CS. Neuroendocrine and behavioural responses to exposure to an infant in male prairie voles. J Neuroendocrinology 2012 Jun;24(6):874-886. [doi: 10.1111/j.1365-2826.2012.02301.x]
25. Kovatsi L, Nikolaou K. Opioids and the hormone oxytocin. Vitam Horm 2019;111(195–225):195-225. [doi: 10.1016/bs.vh.2019.05.003] [Medline: 31421701]
26. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research Electronic Data Capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381. [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
27. Elverson CA, Wilson ME. Cortisol: circadian rhythm and response to a stressor. Newborn Infant Nurs Rev 2005 Dec;5(4):159-169. [doi: 10.1053/j.nainr.2005.09.002]
28. Forsling ML. Diurnal rhythms in neurohypophysial function. Exp Physiol 2000 Mar;85 Spec No:179S-186S. [doi: 10.1111/j.1469-445x.2000.tb00022.x] [Medline: 10795921]
29. Allen AM, Lundeen K, Murphy SE, Spector L, Harlow BL. Web-delivered multimedia training materials for the self-collection of dried blood spots: a formative project. JMIR Form Res 2018 Nov 5;2(2):e11025. [doi: 10.2196/11025] [Medline: 30684406]
30. Newton RW, Webster PA, Binu PS, Maskrey N, Phillips AB. Psychosocial stress in pregnancy and its relation to the onset of premature labour. Br Med J 1979 Aug 18;2(6187):411-413. [doi: 10.1136/bmj.2.6187.411] [Medline: 486966]
31. Felitti VJ, Anda RF, Nordenberg D, et al. Relationship of childhood abuse and household dysfunction to many of the leading causes of death in adults. Am J Prev Med 1998 May;14(4):245-258. [doi: 10.1016/S0749-3797(98)00017-8]
32. Bremner JD, Vermetten E, Mazure CM. Development and preliminary psychometric properties of an instrument for the measurement of childhood trauma: the Early Trauma Inventory. Depress Anxiety 2000;12(1):1-12. [doi: 10.1002/1520-6394(2000)12:1<1::AID-DA1>3.0.CO;2-W] [Medline: 10999240]
33. Bremner JD, Bolus R, Mayer EA. Psychometric properties of the Early Trauma Inventory-Self Report. J Nerv Ment Dis 2007 Mar;195(3):211-218. [doi: 10.1097/01.nmd.0000243824.84651.6c] [Medline: 17468680]
34. Wagnild GM, Young HM. Development and psychometric evaluation of the Resilience Scale. J Nurs Meas 1993;1(2):165-178. [Medline: 7850498]
35. Cleeland CS, Ryan KM. Pain assessment: global use of the Brief Pain Inventory. Ann Acad Med Singap 1994 Mar;23(2):129-138. [Medline: 8080219]
36. Cox JL, Holden JM, Sagovsky R. Detection of postnatal depression. Development of the 10-item Edinburgh Postnatal Depression Scale. Br J Psychiatry 1987 Jun;150:782-786. [doi: 10.1192/bjp.150.6.782] [Medline: 3651732]
37. Park ER, Psaros C, Traeger L, et al. Development of a postpartum stressor measure. Matern Child Health J 2015 Oct;19(10):2094-2101. [doi: 10.1007/s10995-015-1731-0] [Medline: 25682113]
38. Osman A, Wong JL, Bagge CL, Freedenthal S, Gutierrez PM, Lozano G. The Depression Anxiety Stress Scales-21 (DASS-21): further examination of dimensions, scale reliability, and correlates. J Clin Psychol 2012 Dec;68(12):1322-1338. [doi: 10.1002/jclp.21908] [Medline: 22930477]
39. Brockington IF, Fraser C, Wilson D. The Postpartum Bonding Questionnaire: a validation. Arch Womens Ment Health 2006 Sep;9(5):233-242. [doi: 10.1007/s00737-006-0132-1] [Medline: 16673041]
40. Crncec R, Barnett B, Matthey S. Development of an instrument to assess perceived self-efficacy in the parents of infants. Res Nurs Health 2008 Oct;31(5):442-453. [doi: 10.1002/nur.20271] [Medline: 18297638]
41. Johns MW. A new method for measuring daytime sleepiness: the Epworth sleepiness scale. Sleep 1991 Dec;14(6):540-545. [doi: 10.1093/sleep/14.6.540] [Medline: 1798888]

XSL•FO
RenderX

42. Buysse DJ, Reynolds CF III, Monk TH, Berman SR, Kupfer DJ. The Pittsburgh sleep quality index: a new instrument for psychiatric practice and research. Psychiatry Res 1989 May;28(2):193-213. [doi: 10.1016/0165-1781(89)90047-4]
43. Russell DW. UCLA Loneliness Scale (Version 3): reliability, validity, and factor structure. J Pers Assess 1996 Feb;66(1):20-40. [doi: 10.1207/s15327752jpa6601_2] [Medline: 8576833]
44. Sherbourne CD, Stewart AL. The MOS social support survey. Social Science & Medicine 1991 Jan;32(6):705-714. [doi: 10.1016/0277-9536(91)90150-B]
45. Barkin JL, Wisner KL, Bromberger JT, Beach SR, Terry MA, Wisniewski SR. Development of the Barkin index of maternal functioning. J Womens Health (Larchmt) 2010 Dec;19(12):2239-2246. [doi: 10.1089/jwh.2009.1893]
46. Ohan JL, Leung DW, Johnston C. The Parenting Sense of Competence scale: evidence of a stable factor structure and validity. Can J Behav Sci 2000;32(4):251-261. [doi: 10.1037/h0087122]
47. Luyten P, Mayes LC, Nijssens L, Fonagy P. The parental reflective functioning questionnaire: development and preliminary validation. PLoS One 2017;12(5):e0176218. [doi: 10.1371/journal.pone.0176218] [Medline: 28472162]
48. Fonagy P, Luyten P, Moulton-Perkins A, et al. Development and validation of a self-report measure of mentalizing: the reflective functioning questionnaire. PLOS ONE 2016;11(7):e0158678. [doi: 10.1371/journal.pone.0158678] [Medline: 27392018]
49. Gartstein MA, Rothbart MK. Studying infant temperament via the Revised Infant Behavior Questionnaire. Infant Behav Dev 2003 Feb;26(1):64-86. [doi: 10.1016/S0163-6383(02)00169-8]
50. Sadeh A. A brief screening questionnaire for infant sleep problems: validation and findings for an Internet sample. Pediatrics 2004 Jun;113(6):e570-e577. [doi: 10.1542/peds.113.6.e570] [Medline: 15173539]
51. Wood RG, Moore Q, Clarkwest A, Killewald A, Monahan S. The building strong families project: the long-term effects of building strong families: a relationship skills education program for unmarried parents. 2019. URL: https://www.acf.hhs.gov/opre/report/building-strong-families-project-long-term-effects-building-strong-families [accessed 2024-11-23]
52. Plikaytis BD, Holder PF, Pais LB, Maslanka SE, Gheesling LL, Carlone GM. Determination of parallelism and nonparallelism in bioassay dilution curves. J Clin Microbiol 1994 Oct;32(10):2441-2447. [doi: 10.1128/jcm.32.10.2441-2447.1994]
53. Morimoto LM, Zava D, McGlynn KA, et al. Neonatal hormone concentrations and risk of testicular germ cell tumors (TGCT). Cancer Epidemiol Biomarkers Prev 2018 Apr 1;27(4):488-495. [doi: 10.1158/1055-9965.EPI-17-0879]
54. Sobell LC, Cunningham JA, Sobell MB. Recovery from alcohol problems with and without treatment: prevalence in two population surveys. Am J Public Health 1996 Jul;86(7):966-972. [doi: 10.2105/ajph.86.7.966] [Medline: 8669520]
55. Zhang X, Cui W, Yao J, Zhang Y, Wang Y. Feasibility and utility of ecological momentary assessment to measure mental health issues in perinatal women: scoping review. Psychiatry Res 2025 Jun;348:116400. [doi: 10.1016/j.psychres.2025.116400] [Medline: 40220631]
56. Schiller CE, Saladin ME, Gray KM, Hartwell KJ, Carpenter MJ. Association between ovarian hormones and smoking behavior in women. Exp Clin Psychopharmacol 2012 Aug;20(4):251-257. [doi: 10.1037/a0027759] [Medline: 22545725]
57. Grund S, Lüdtke O, Robitzsch A. The Handbook of Multilevel Theory, Measurement, and Analysis: American Psychological Association; 2019:365-386. [doi: 10.1037/0000115-017]
58. McNeish D. Missing data methods for arbitrary missingness with small samples. J Appl Stat 2017 Jan 2;44(1):24-39. [doi: 10.1080/02664763.2016.1158246]
59. Lee T, Shi D. A comparison of full information maximum likelihood and multiple imputation in structural equation modeling with missing data. Psychol Methods 2021;26(4):466-485. [doi: 10.1037/met0000381]
60. Infante-Rivard C, Cusson A. Reflection on modern methods: selection bias—a review of recent developments. Int J Epidemiol 2018 Oct 1;47(5):1714-1722. [doi: 10.1093/ije/dyy138] [Medline: 29982600]
61. Robins JM, Rotnitzky A, Scharfstein DO. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In: Halloran ME, Berry D, editors. Statistical Models in Epidemiology, the Environment, and Clinical Trials: Springer; 2000:1-94. [doi: 10.1007/978-1-4612-1284-3_1]
62. Yaseliani M, Noor-E-Alam M, Hasan MM. Mitigating sociodemographic bias in opioid use disorder prediction: fairness-aware machine learning framework. JMIR AI 2024 Aug 20;3(1):e55820. [doi: 10.2196/55820] [Medline: 39163597]
63. Chung PC, Lin IF. Sensitivity analysis of selection bias: a graphical display by bias-correction index. PeerJ 2023;11:e16411. [doi: 10.7717/peerj.16411] [Medline: 38025739]
64. Rankin L, Grisham LM, Mendoza N, Allen A. Babywearing reduces urges to use substances in the postpartum period among mothers with OUDs. Subst Use Misuse 2024;59(8):1150-1156. [doi: 10.1080/10826084.2024.2321253] [Medline: 38404001]
65. Williams LR, Gebler-Wolfe M, Grisham LM, Bader MY. “Babywearing” in the NICU. Adv Neonatal Care 2020;20(6):440-449. [doi: 10.1097/ANC.0000000000000788]
66. Estradiol (Briggs Drugs in Pregnancy and Lactation) - UpToDate® LexidrugTM. URL: https://online-lexi-com.libproxy.uams.edu/lco/action/doc/retrieve/docid/1090/5711418 [accessed 2025-09-11]
67. Testosterone (Briggs Drugs in Pregnancy and Lactation) - UpToDate® LexidrugTM. URL: https://online-lexi-com.libproxy.uams.edu/lco/action/doc/retrieve/docid/1090/5712082 [accessed 2025-09-11]
68. Ye C, Wang M, Min J, et al. A wearable aptamer nanobiosensor for non-invasive female hormone monitoring. Nat Nanotechnol 2024 Mar;19(3):330-337. [doi: 10.1038/s41565-023-01513-0]

XSL•FO
RenderX

Abbreviations

MOUD: medications for opioid use disorder
OUD: opioid use disorder

Edited by J Sarvestan; submitted 21.05.25; peer-reviewed by H Shin; revised version received 11.09.25; accepted 15.10.25; published 20.11.25.

Please cite as:
Allen AM, Linde-Krieger LB, Deschenes J, Mallahan S, Harris A, Felix M, Chalke A, Anderson A, Sharma P, King KM, Grant MT, Baurley J, Rankin L, Tecot S
Compliance and Satisfaction With a Protocol for Identifying Novel Targets to Support Postpartum Opioid Use Disorder Recovery: Prospective Cohort Study
JMIR Form Res 2025;9:e77899
URL: https://formative.jmir.org/2025/1/e77899
doi:10.2196/77899

© Alicia M Allen, Linnea B Linde-Krieger, Jendar Deschenes, Stephanie Mallahan, Alexandra Harris, Mariana Felix, Arushi Chalke, Alma Anderson, Priyanka Sharma, Katherine M King, Maddy T Grant, James Baurley, Lela Rankin, Stacey Tecot. Originally published in JMIR Formative Research (https://formative.jmir.org), 20.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Identifying Biomedical Entities for Datasets in Scientific Articles: 4-Step Cache-Augmented Generation Approach Using GPT-4o and PubTator 3.0

Claudia Giuliani[1], PhD; Gita Benadi[1], PhD; Felix Engel[1], PhD; Jonas Werner[1], MSc; Manuel Watter[1], MSc; Guido Schwarzer[1], PhD; Olaf Groß[2,3], PhD; Robert Zeiser[3,4], MD; Harald Binder[1,3], PhD; Klaus Kaier[1,3], PhD

[1]Institute of Medical Biometry and Statistics, Medical Faculty and Medical Center, University of Freiburg, Stefan-Meier-Str. 26, Freiburg, Germany

[2]Institute of Neuropathology, Medical Faculty and Medical Center, University of Freiburg, Freiburg, Germany

[3]Center for Integrative Biological Signaling Studies, University of Freiburg, Freiburg, Germany

[4]Department of Medicine I, Medical Faculty and Medical Center, University of Freiburg, Freiburg, Germany

Corresponding Author:
Klaus Kaier, PhD
Institute of Medical Biometry and Statistics, Medical Faculty and Medical Center, University of Freiburg, Stefan-Meier-Str. 26, Freiburg, Germany

Abstract

Background: The accurate extraction of biomedical entities in scientific articles is essential for effective metadata annotation of research datasets, ensuring data findability, accessibility, interoperability, and reusability in collaborative research.

Objective: This study aimed to introduce a novel 4-step cache-augmented generation approach to identify biomedical entities for an automated metadata annotation of datasets, leveraging GPT-4o and PubTator 3.0.

Methods: The method integrates four steps: (1) generation of candidate entities using GPT-4o, (2) validation via PubTator 3.0, (3) term extraction based on a metadata schema developed for the specific research area, and (4) a combined evaluation of PubTator-validated and schema-related terms. Applied to 23 articles published in the context of the Collaborative Research Center *OncoEscape*, the process was validated through supervised, face-to-face interviews with article authors, allowing an assessment of annotation precision using random-effects meta-analysis.

Results: The approach yielded a mean of 19.6 schema-related and 6.7 PubTator-validated biomedical entities per article. Within the study's specific context, the overall annotation precision was 98% (95% CI 94%-100%), with most prediction errors concentrated in articles outside the primary basic research domain of the schema. In a subsample (n=20), available supplemental material was included in the prediction process, but it did not improve precision (98%, 95% CI 95% - 100%). Moreover, the mean number of schema-related entities was 20.1 (P=.56) and the mean number of PubTator-validated entities was 6.7 (P=.68); these values did not increase with the additional information provided in the supplement.

Conclusions: This study highlights the potential of large language model–supported metadata annotation. The findings underscore the practical feasibility of full-text analysis and suggest its potential for integration into routine workflows for biomedical metadata generation.

(JMIR Form Res 2025;9:e73822) doi:10.2196/73822

KEYWORDS

metadata annotation; biomedical entities; GPT-4o; PubTator 3.0; cache-augmented generation; CAG; artificial intelligence; AI

Introduction

Long-term university-based research institutions, such as the German Collaborative Research Centers (CRCs), which can be funded for up to 12 years, require controlled and structured sharing of data and documents, with accurate, interoperable metadata being crucial for long-term data reusability and understanding [1-6] in accordance with the findability, accessibility, interoperability, and reusability (FAIR) principles [7]. To meet the requirements of the funding agency, the CRC *OncoEscape* (CRC 1479) adopted a dedicated metadata schema for describing its datasets [8]. These metadata are entered and stored in a central system, the *fredato* [9] platform, which makes the research data findable for the entire consortium and other users of the system. Currently, scientists have to manually fill the metadata forms, which often leads to suboptimal compliance. We therefore investigated methods for introducing an automated yet human-validated process for enriching datasets with relevant metadata.

XSL•FO
RenderX

For biomedical research data, a particularly important type of metadata is biomedical entities, such as organisms, cell lines, or genes. If published research articles and the associated datasets are tagged with these biomedical entities, it is easier for scientists to find relevant datasets for their research and to reuse these data. As manual entity annotation is often a time-consuming task, it is desirable to automate this process as much as possible. One method of automatic biomedical entity identification is named entity recognition, which involves identifying spans of text that represent named entities (eg, "mouse") and tagging them with the appropriate category (eg, "organism") [10]. Entity recognition is challenging due to difficulties in segmentation, that is, determining whether (1) a term represents an existing biomedical entity and (2) whether it is appropriate for describing the datasets that are presented in the corresponding research article.

Large language models (LLMs) can be used for entity recognition, and they can significantly increase the speed of biomedical entity documentation [11]. While LLMs exhibit strong language capabilities, they are prone to generating incorrect information, often referred to as hallucinations [12-16]. A possible way of mitigating this problem is to restrict the terms identified by the LLM to a set of known true bio-entities. This restriction can either be directly specified in the prompt to the LLM, or applied retrospectively by validating the list of terms suggested by the model.

In this manuscript, we present a newly developed method for article-based metadata annotation prediction based on the use of LLMs and the cache-augmented generation (CAG) method [17]. In particular, we used the LLM GPT-4o (OpenAI) in combination with PubTator 3.0, a tool built by the National Library of Medicine to identify several entities in the literature using state-of-the-art artificial intelligence (AI) techniques [18,19], and a dedicated metadata schema developed to describe data collected for a specific research area [8]. With this 2-fold combination of validation and constraint of the LLM, we intended to provide a reliable and extended list of annotation suggestions, exploiting the strengths of both PubTator, such as comprehensive genes and chemicals lists, and the dedicated metadata schema, which includes relevant mouse and cell lines that are often created in-house. This multipronged approach directly mitigates the risk of LLM "hallucination" by grounding the model's output in 2 sources of truth: the externally validated PubTator 3.0 database and the domain-specific, curated metadata schema. The approach consists of four steps: (1) GPT-4o–based full-text analysis and suggestion of biomedical terms, (2) PubTator-based validation of suggested terms from step 1, (3) full-text analysis restricted to schema-related terms, and (4) combination of schema-related and PubTator-validated terms.

Why use both PubTator and a schema? PubTator 3.0 provides high-precision normalization for universal entities (eg, genes and chemicals) via canonical IDs, whereas the consortium schema captures project-specific concepts (eg, in-house mouse and cell lines, endpoints, and sample processing). Combining them lets the LLM ground standardized entities and still surface domain-specific terms that PubTator does not index. In practice, PubTator reduces false positives from free-text generation, while the schema broadens relevant coverage within the consortium's scope. The merging step yields a single suggestion list with clear provenance (PubTator-validated entities vs schema-constrained entities).

We applied this 4-step approach in the context of the CRC *OncoEscape* (CRC 1479) and tested the feasibility of the approach using a structured, paper-based, and supervised annotation double-checking process. This process involved face-to-face interviews with the scientists who authored the respective article and enabled us to calculate a first estimate of the proportion of correctly predicted annotations. In detail, we evaluated the feasibility and precision of a 4-step cache-augmented pipeline for article-based metadata annotation using GPT-4o, PubTator 3.0, and a domain schema. The primary end point was precision of suggested annotations, assessed by supervised author validation for 23 OncoEscape articles. Secondary endpoints included the numbers of schema-related and PubTator-validated entities per article and a subsample analysis with and without supplemental material.

OncoEscape is a CRC funded by the German Research Foundation. The scientific goal of the consortium is to study oncogene-driven immune escape from the perspectives of clinical and basic tumor biologists and immunologists. The research of the CRC builds on the hypothesis that oncogenic signaling and immune escape mechanisms are closely connected. This concept implies that targeting oncogene-driven immune evasion could transform treatments that are currently mainly palliative into potential cures for different types of cancer.

Methods

Study Design and Data Sources

We included a total of 23 articles that were published during the first funding period of *OncoEscape* (2021 - 2025) and for which a scientist responsible for authoring the article had agreed to take part in face-to-face interviews (Table S1 in Multimedia Appendix 1). We deliberately implemented this restriction, as we assumed that the social control arising from the presence of a member of the research data management group would be most appropriate to obtain a reliable assessment of the precision of the 4-step approach. All participants in the face-to-face interviews carefully checked the annotation predictions. In a subsample (n=20), available supplemental material was also included in the prediction process and 2 prediction results were validated in the face-to-face interviews. Of the 3 remaining articles, 1 had no supplemental material, 1 had such a large supplemental material section that the LLM was not able to analyze it correctly, and 1 was not considered in the face-to-face interviews.

LLM-Based Annotation Prediction

A systematic 4-step approach was used for annotation prediction (Figure 1): (1) ChatGPT (OpenAI) was initially tasked with generating relevant entities based on the full text of the paper; no further restrictions or suggestions were made, but the LLM was instructed to ignore the discussion and bibliography for its predictions. (2) The resulting entities were then validated using the PubTator 3.0 database. (3) Next, ChatGPT was instructed to reanalyze the full text of the manuscript but with the task of

XSL•FO
RenderX

identifying entities defined in a predefined schema. For this purpose, a previously developed dedicated *OncoEscape* metadata schema [8] was included in this prompt in a tree-like structure. (4) Finally, a combined evaluation of the results from steps 2 and 3 was prompted, with the clear order of list only those PubTator-validated entities that had not already been identified through the schema extraction in step 3.

Figure 1. A 4-step cache-augmented generation approach for annotation prediction.

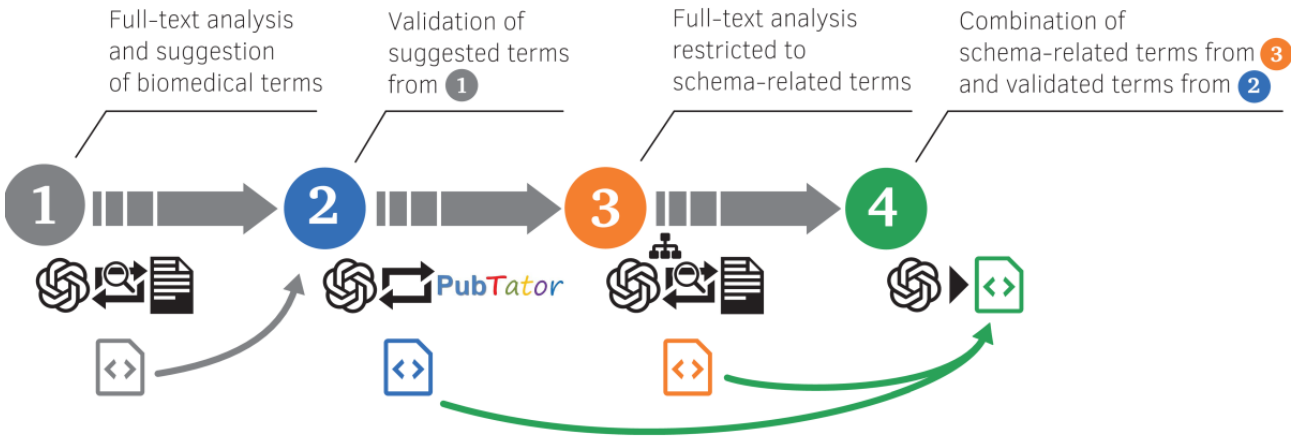

This process requires comprehensive textual analysis and the integration of knowledge, a method commonly referred to as CAG. In recent years, retrieval-augmented generation (RAG) has emerged as the standard approach for customizing LLMs to meet specific informational needs. However, with recent advancements in long-context LLMs, it is now feasible to eliminate the need for RAG by directly embedding all proprietary information within the prompt [17]. RAG's reliance on real-time retrieval could introduce latency and potential errors in document selection, especially with large datasets. In contrast, CAG, by preloading all relevant resources into the LLM's extended context, has been found to eliminate retrieval latency and minimize errors.

On the basis of the recommendations of the developers of PubTator 3.0 [19], we created a custom GPT with PubTator 3.0 augmentation following the instructions of the authors [20]. The above-described 4-step approach was operationalized using separate prompts in the chat interface from OpenAI, where the publication text was uploaded in the first step as a PDF file. Refer to the supplemental material for further details of the 4-step approach and specifically Table S2 in Multimedia Appendix 1 for the prompts used. All analyses were conducted using ChatGPT-4o-2024-11-20. Please note that this multistep prompting approach involved an iterative process in which ChatGPT connected to PubTator 3.0 to ensure reliability of the results. For each entity generated in step 1, step 2 instructed ChatGPT to query PubTator to retrieve a standardized entity ID, ensuring the validation of suggestions and alignment with recognized biomedical terms. Step 3 served the purpose of aligning and categorizing the biomedical entities suggestion with the dedicated *OncoEscape* metadata schema. This iterative process of generation, validation, and evidence gathering ensures that only scientifically grounded and relevant entities are carried forward and thereby overcomes some of the hallucination problems that arise when prompting ChatGPT in an unspecific way. Furthermore, the multistep prompting approach enables us to separate the capabilities of annotation prediction based on a schema and those of PubTator. In other words, the PubTator-based predictions can consider the scientific context from the paper and are not limited by the schema's incompleteness. The PubTator-validated entities are either linked to a category but not included in the list of schema-defined entities, or they are entities that could not be associated with any OncoEscape schema category. In contrast, PubTator includes 6 categories of annotated entity types: gene, disease, chemical, variant, species, and cell line; the "FindEntityID" end point used in the GPT + PubTator implementation, however, does not consider species and cell line entries. This limitation is not relevant for the overall suggestion completeness because the *OncoEscape* metadata schema contains relevant species and dedicated cell lines used in the specific research area.

Validation of LLM-Based Annotation Prediction

In this study, the proportion of correctly predicted annotations was chosen as the metric to measure the prediction capabilities of the LLMs. The proportion is computed for each paper and experimental setting (with or without supplement; total, schema-related, and PubTator-related suggestions). To identify correctly predicted annotations, we used a supervised annotation double-checking process. This process involved face-to-face interviews between a member of the research data management group and a senior scientist responsible for authoring the article. The research data management group member acted as a neutral facilitator and clarified ambiguities in the annotations without influencing the evaluation outcome. Scientists were instructed to consider the datasets referenced in the article as the context for evaluating annotations. Therefore, a standardized instruction sheet was used across all interviews to ensure consistency in interpretation. Each annotation suggested by the 4-step approach was examined to determine whether it accurately described the datasets used for the article. If an annotation was suspected to be incorrect, participants were encouraged to consult the methods or results sections of the article, which were provided in the paper for quick verification. Annotations deemed incorrect based on suspicion or evident errors could also be marked incorrect without comprehensive cross-checking to prioritize efficiency. The process was conducted in a supervised manner,

XSL•FO
RenderX

and a 5-minute time span was recommended for a single article. This 5-minute face-to-face evaluation method ensured a practical balance between efficiency and reliability, enabling the rapid assessment of the 4-step approach.

Statistical Analysis of the Interview Outcomes

To estimate the LLM prediction accuracy across all studies, methods for the meta‑analysis of single proportions were applied [21]. Due to the small number of entities per study, the Freeman-Tukey double arcsine transformation was used [22]. This transformation stabilizes the variances, makes the meta-analysis more robust when proportions are close to 1, and enables working with small sample sizes. A random-effects model using the restricted maximum likelihood estimator for between-study variance was applied. In addition, the behavior of the LLM, which is intrinsically stochastic, as well as the possible variability in the evaluation of LLM suggestions by different authors justified the calculation of CIs for the accuracy predictions in individual papers. Differences between the number of article-based and article + supplement–based biomedical entities were analyzed using the Wilcoxon signed-rank test.

In studies on natural language processing (NLP) using LLMs, false negatives are often also considered to determine the accuracy, as well as recall and the F_1-score. In the present situation, where the goal of prediction was to maximize the number of correct annotations while also keeping the proportion of false-positive annotations as small as possible, it was not possible to compute the said metrics because the number of potential annotations per dataset had no clearly defined upper limit. Although it would be possible to ask the authors in the face-to-face interviews which biomedical entities they found missing, this question could not be addressed without additional effort. Moreover, we intentionally constrained the LLM by giving it information about the predefined CRC metadata schema so that the annotations were not only contextually appropriate to each paper but also relevant to the CRC's specific scope.

Ethical Considerations

This study's primary data and analysis are based on nonhuman sources. The involvement of human participants was limited to informal expert consultations with professional colleagues to verify the study findings. No personal, private, or health-related data were collected from these individuals; the discussions focused exclusively on the interpretation of the nonhuman research data. According to national guidelines in Germany, particularly those established by the German Data Forum, research in the social and economic sciences that does not involve identifiable human material or susceptible participants may not automatically require formal review by an ethics committee. The German Data Forum's guidance on research ethics emphasizes reflexive ethical conduct rather than a blanket ethics-votum for every empirical investigation [23]. Therefore, in line with these national standards and given that our study involves only voluntary expert interviews, no personal identifiable data collection, and no susceptible groups, we deemed a formal ethics committee approval unnecessary and instead opted for informed participation. These consultations were conducted in line with the principles of the Declaration of Helsinki and the guidelines for Good Research Practice of the German Research Foundation. All colleagues who participated did so voluntarily after being fully informed about the purpose of the discussion. Informed verbal consent was obtained from all colleagues who participated in the expert consultations.

Results

A total of 604 biomedical entities were predicted for the 23 articles [24-46] (Figure 2). This equals a mean number of 26.3 biomedical entities per article, of which 19.6 were schema-related biomedical entities and another 6.7 were PubTator-validated biomedical entities. Overall precision, defined as the ratio of correct entities to the total number of annotation suggestions, was 98% (95% CI 94%‑100%), meaning that the vast majority of predicted biomedical entities were considered correct in the face-to-face interviews. An example of the LLM output for the article by Frueh et al [42] can be found in Table S3 in Multimedia Appendix 1.

Interestingly, the precision of schema-related biomedical entities was 97% (95% CI 91%‑100%), while the precision of PubTator-validated biomedical entities was 100% (Figures S1 and S2 in Multimedia Appendix 1). Most of the articles had a proportion of correctly chosen entities for the annotation close to or equal to 100% (Figure 2). Only 2 papers, namely, those by Chen et al [43] and Zeiser et al [44], had a markedly lower proportion of correctly suggested annotation entities, which contributed to the considerable between-study heterogeneity (I^2=76%). Both these papers are clinical studies concerned with the transfer of basic research results into a clinical setting and had a number of false-positive basic research–related annotations that, although mentioned in the text, served as a rationale for the clinical approach rather than describing the underlying dataset. A post hoc sensitivity analysis excluding these 2 studies resulted in a pooled precision of 100% (95% CI 99%‑100%) with no substantial between-study heterogeneity (I^2=0%).

XSL•FO
RenderX

Figure 2. Precision of annotation predictions for each paper without supplement included in this study, along with pooled precision, heterogeneity, and τ^2. Each dot and interval represents a paper-level precision estimate; the diamond indicates the pooled descriptive value [24-46].

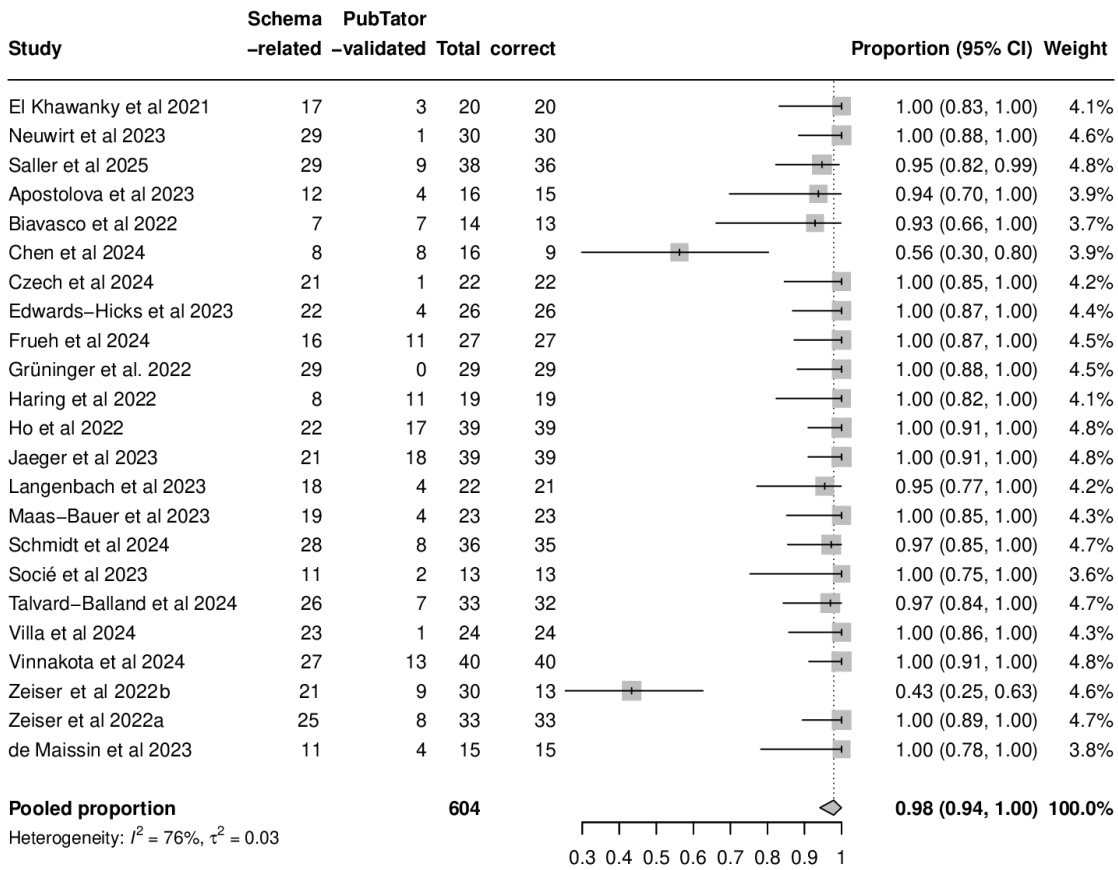

In a subsample (n=20), available supplemental material was also included in the prediction process, and 2 prediction results, both those for the paper with and without supplement, were validated in the face-to-face interviews (Figure S3 in Multimedia Appendix 1). Interestingly, availability of supplemental material in the prediction process did not increase precision (98%, 95% CI 95% - 100%), with between-study heterogeneity of I^2=67%. Moreover, the mean number of schema-related entities was 20.1 (P=.56) and the mean number of PubTator-validated entities was 6.7 (P=.68); these values did not increase with the additional information provided with the supplement and were compatible with the values obtained with the paper-only approach.

No statistically relevant differences were observed in the number of suggested entities for the papers without or with supplement, even when annotation suggestions were analyzed by category (Figure 3). Here, only the 20 papers that included annotation suggestions for the supplement are considered. The listed categories are as defined in the *OncoEscape* schema with 2 additional categories representing the availability of data: DataAvailability, which can be available publicly or on request as extracted from the paper, and PubTatorNotClassified, the PubTator-related entities that the LLM could not assign to a predefined category). Additional distributions are shown in the Multimedia Appendix.

The compatibility of the annotation suggestions obtained by analyzing the paper with or without the supplement was confirmed, both for the overall number of suggestions (P=.56) as well as for the suggestions divided per category. The detailed results of the Wilcoxon signed-rank tests are summarized in Table S4 in Multimedia Appendix 1.

XSL•FO
RenderX

Figure 3. Counts of large language model annotation suggestions per category for papers with and without supplements. Only papers that included an annotated supplement are considered (n=20).

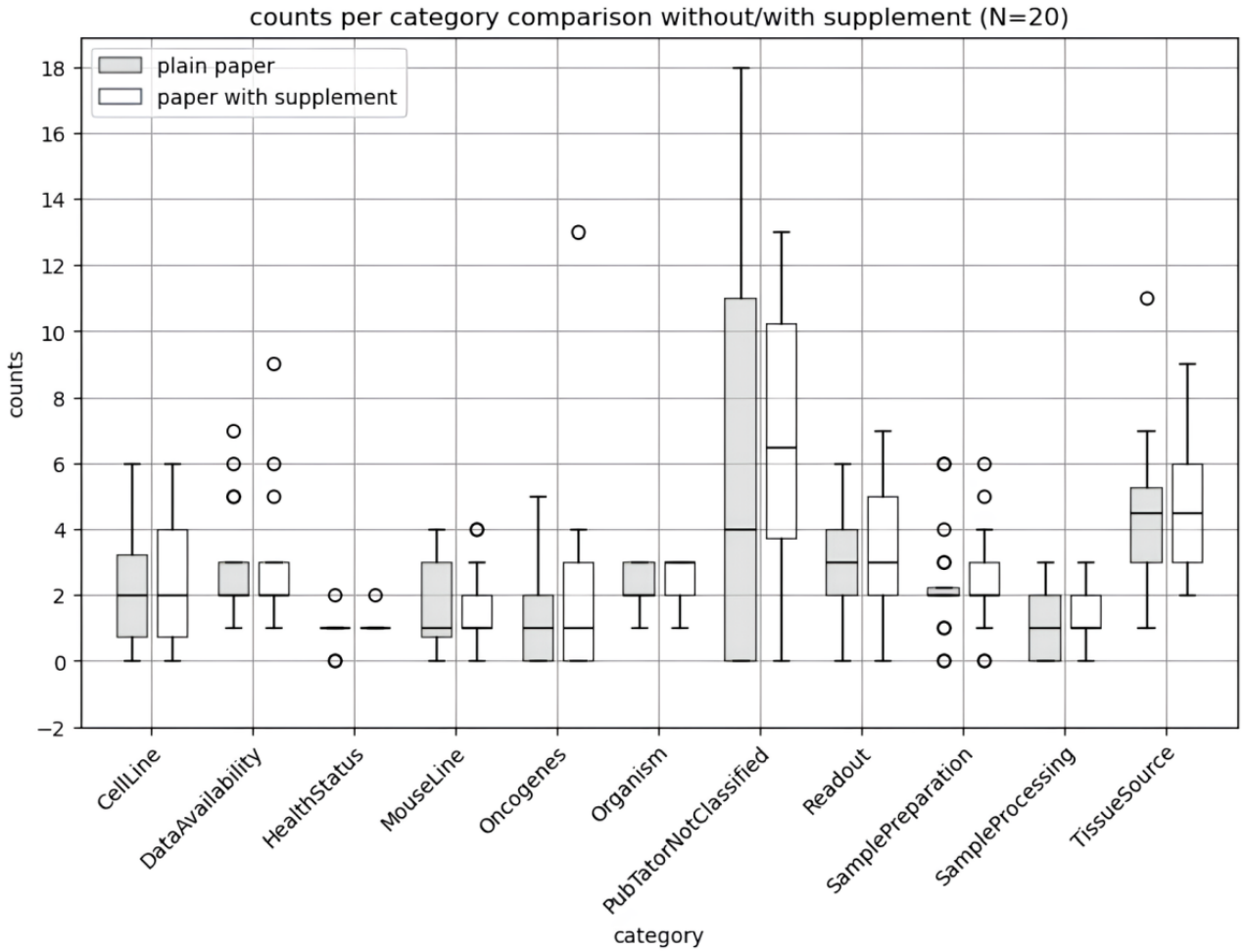

Discussion

Principal Findings

Our 4-step approach involves a combination of GPT-4o, PubTator, and schema-related term analysis. This process requires comprehensive textual analysis and the integration of knowledge in an iterative process. We used CAG by preloading all relevant resources into the LLM's extended context to accelerate information retrieval and minimize response errors. For instance, as both steps 1 and 3 involve full-text analysis, we accelerated step 3 by using CAG. Preloading the cache allowed the model to quickly access and process the entire article text, leading to faster suggestion of terms [17]. Moreover, CAG helped ensure that the model had a unified understanding of the entire article context, leading to improved consistency and accuracy in term prediction across all steps. This was crucial for maintaining the coherence and quality of metadata annotation throughout the process.

The high precision observed in our study underscores the potential of the CAG approach for streamlining metadata annotation. Moreover, because the supplemental material did not improve the average number of annotation suggestions or the proportion of correct ones and because it is complicated to retrieve the supplemental material in a standardized way, we concluded that the LLM annotation suggestions of papers without supplement should suffice in most cases. The reasons for this behavior are multiple: (1) supplemental material does not necessarily add methodological information and details. For example, in some cases, the supplemental material only includes tables and figures. Still, if we would use a deterministic approach, we would expect to obtain the same number of predictions each time and at least as many when the supplement is included. (2) LLMs are intrinsically stochastic, and therefore, the number of suggestions varies if the experiment is run multiple times. (3) Traffic on the server where the LLM runs and on the PubTator database can impact the number of suggested and PubTator-verified entities, as connection timeouts may occur under high load.

Although the abovementioned points 2 and 3 cannot be kept completely apart, we ran new experiments to show their combined effect on the number of suggested entities. For the paper by Saller et al [25] (without supplement), we ran the 4-step approach 10 times. On average, 46.7 entities were extracted, with a SD of 7.7. The minimum number of extracted entities was 37; the maximum number of extracted entities was 59. The count spread per category is shown in Figure S6 in Multimedia Appendix 1. The observed variability was not only in the number of extracted entities but also in the entities themselves. The total average Jaccard index over the 10 runs was 0.46. The summary

XSL•FO
RenderX

Jaccard heat map for all combined categories is provided in Figure S7 in Multimedia Appendix 1. For selected categories, additional Jaccard heat maps and visualizations of the overlaps are shown in Figures S8-S11 in Multimedia Appendix 1.

The provided metadata schema had a nonnegligible impact on the quality and level of details of the biomedical entity suggestions and categorization by the LLM, especially for the categories not covered by the PubTator entities (eg, endpoint, sample preparation and processing, and tissue source). For example, if a dataset was processed with single-cell RNA sequencing but we gave the LLM only the first level of details of endpoint methods, it was able to provide only the given level of detail. In this case, "Sequencing" was selected as an annotation suggestion, which could be deemed wrong by the authors of the paper. In contrast, if we provided the LLM with additional endpoint levels of detail and, continuing with our example, we included different sequencing options, the LLM was able to identify that the data were processed with "single-cell RNA sequencing." If the schema entities were passed to the LLM in a tree-like structure, the LLM was able to identify dependencies and reproduce the annotation depth. Likewise, the schema limited the capabilities of the LLM to correctly identify entities, especially if the paper in question did not analyze primary datasets in the context of basic research but was, for example, a clinical or modeling study. This was apparent in the lower proportion of correctly suggested entities in the papers discussing clinical studies. In both such articles, the basic research rationale was applied for the clinical approach, but the underlying datasets were of clinical nature.

The PubTator-validated entities identified outside the schema highlighted the potential of a hybrid approach that combines the structured annotation provided by the metadata schema with the flexibility to capture emerging concepts.

Error Analysis

While the overall precision was very high at 98%, a detailed analysis of incorrect predictions was needed. The vast majority of errors originated from 2 specific articles (ie, those by Chen et al [43] and Zeiser et al [44]), which were clinical studies. In both cases, the LLM correctly identified biomedical entities mentioned in the text, but these entities were part of the introductory rationale or background discussion rather than describing the actual clinical datasets presented in the paper. For example, terms related to basic research (eg, specific mouse lines) were extracted because they were mentioned as foundational work, but they were irrelevant for annotating the human patient data that formed the core of the study. This highlighted a key challenge: the discrepancy between annotating an *article* and annotating the *underlying dataset*. Our instruction to the LLM to ignore the discussion and bibliography was an attempt to mitigate this, but it appeared insufficient for articles where the core methodology differed significantly from the background science discussed. This limitation was to some degree linked to our schema, which was heavily geared toward basic biomedical research and less suited for clinical or modeling studies. The lower precision in these cases underscored the need for schema adaptation when applying this method across different research domains. Overall, the limited generalizability of the current schema, which was deliberately designed for basic research datasets, explained why the approach underperformed in clinical or modeling contexts. From a responsible AI perspective, this underlined the need for domain-adaptive schema design, complemented by human-in-the-loop oversight to contextualize and correct outputs. Such an approach ensures accurate and context-appropriate annotations across diverse biomedical fields, aligning with the principle that automated annotation must always be supported by expert review.

Comparison to Prior Work

In a recently published article, Alyafeai et al [47] compared a similar approach to metadata annotation based on papers on NLP datasets. Specifically, a 1-step approach was used to compare different LLMs. Compared to our approach, only step 3 was used. Within this step, all types of things were tested based on 52 manually annotated articles. Enabling web browsing led to a slight increase in accuracy, and the best results were achieved with the PDF format as input. An important difference from our approach was that the focus there was not on biomedical entities but rather on administrative aspects, such as license, language, domain, and other NLP-related aspects. Schilling-Wilhelmi et al [48] used a similar approach in the field of chemistry, but their results were only partially comparable due to the rapid development of LLMs (they used an early version of GPT-3.5). The same applies to the study by Singh et al [49], who were only able to achieve a precision of 0.46 based on a schema-based 1-step approach; they also used GPT-3.5. In contrast, Turner et al [50] used a more recent LLM (gpt-4o-2024-08-06) and yielded a high accuracy using a schema-based 1-step approach on papers from the field of psychiatry, with a much more promising accuracy (0.91-0.97). To our knowledge, no work has yet been done on a multistep approach with the aspect of validating biomedical entities. We did not perform head-to-head comparisons in this study; therefore, reported differences from prior work should be interpreted with caution, given variations in domain, model, and protocol.

Limitations

Problems in extracting metadata annotation suggestions can arise when working on very long articles or extensive supplemental materials. This can happen when the available LLM context length is exhausted. Recent studies have shown that while LLMs benefit from larger context windows, their performance does not always scale linearly with increased input size, often exhibiting a "lost in the middle" effect, where information in the middle of long texts is processed less effectively [51]. Further studies on the effect of the LLM choice and its context length on the reliability of the annotation predictions are ongoing.

A major challenge of our approach was the distinction between article and datasets: the metadata annotation was focused on the respective underlying datasets; however, the prediction was conducted for each already published journal publication. This was intended, as we assumed that scientists can usually refer to the respective article when annotating their datasets with metadata. However, an article not only describes details of the respective datasets but also discusses aspects that go beyond

XSL•FO
RenderX

them. We tried to solve this discrepancy by instructing the LLM to ignore the discussion and bibliography of the full texts during prediction (step 1). In view of some incorrect predictions, however, this may not have worked entirely. Nevertheless, we felt it was extremely important to further develop our 4-step approach on the full text of the article (instead of just feeding in parts of the article), as this seemed to be the only practical way to transfer it into a fully automated annotation process. The observed misclassifications underscore that responsible integration of LLM-based annotation requires a human-in-the-loop approach, ensuring that contextual background information is not inadvertently elevated to dataset-level descriptors. Furthermore, while our interview-based validation provided deep, expert-verified insights, we acknowledge that it may have introduced investigator bias and is less scalable than automated methods; this represents a trade-off between validation depth and breadth that future studies should address.

Our aim was to enrich the datasets collected within the research projects with metadata information to ultimately make them FAIR. To decrease the problem of having annotation predictions extracted from the analysis of the article and being unable to assign them reliably to a given dataset, we let the LLM retrieve information about available datasets after running the 4-step approach. For published datasets, the LLM was able to find the links to the datasets in the article and, when asked for it, could directly gather information about each published dataset through a web search. At least for published datasets, we can thus reliably assign our annotation suggestions to a specific dataset. However, some datasets were not published or only available on request. As in many cases, the datasets were not fully described; we could only state that data with certain properties and preparation were collected, but we could not have a one-to-one assignment of metadata to datasets. To improve the FAIRness of these datasets, additional infrastructural or institutional measures are needed, alongside further development of the LLM-based analysis methods. Additional measures could include the standardization of the dataset format with compulsory metadata information associated with it, interconnection of systems for automatic collection of metadata during the complete data lifecycle (data storage solutions, electronic lab notebooks, and data collection instruments such as microscopes), or establishing requirements for eventual publication of all datasets.

Future Work and Recommendations

In general, LLM-based annotation enables the annotation of a large number of papers and datasets, thereby enhancing the FAIRness of the data. A comprehensive assessment not only of the quantitative but also of the qualitative advantages of the LLM automatic annotation compared with a diligent manual annotation is being investigated. It is already apparent that including the LLM-generated suggestions and validating them through PubTator is beneficial for having a comprehensive list of oncogenes and treatments considered in the experiments underlying the datasets. Future work should explore dynamic schema adaptation, for example, to include other areas of research such as computational modeling and integration of knowledge graphs [52,53] to represent complex relationships between biomedical entities, ultimately enhancing the comprehensiveness and accuracy of metadata annotation.

We showed that including supplemental material does not have a substantial impact on the number and quality of annotation suggestions. This finding is an advantage for the future automation of the annotation task because format and location of supplemental material are not standardized, which makes it difficult to automatically include in the annotation process. Nevertheless, we plan to further investigate the impact of supplemental material for different publication categories.

In addition, further studies are needed to evaluate the reproducibility of results, given the stochastic nature of the LLM output and the variability of human responses in interviews. In particular, we plan to investigate the impact of the choice of LLM on the annotation suggestions, as well as the variation of suggestions (number and extracted entities) when running the 4-step approach multiple times on a single paper in both scenarios: with and without supplemental material.

Furthermore, exploring advanced frameworks, such as Agentic AI, could enable the creation of autonomous workflows that not only perform the 4-step extraction but also dynamically handle error checking, query external databases for resolving ambiguity, and adapt the process based on the article's specific domain. At the time we conducted this study, we had the impression that GPT-4o is fully used with the 4-step approach. However, future LLMs will likely be able to incorporate a larger number of agentic steps, which would allow the approach shown here to include additional validation, thereby moving toward a more robust and even more automated annotation system.

Another key priority for future work is to validate this 4-step approach on a larger and more diverse corpus of articles, extending beyond the OncoEscape consortium. To achieve this, we will need to develop a more scalable validation methodology, transitioning from face-to-face interviews to a structured, web-based survey system for authors. This will not only allow for the inclusion of hundreds of articles but also enable benchmarking against datasets from different biomedical domains, thereby providing a more robust assessment of the method's external validity and its potential for widespread adoption.

Finally, future research should explore systematic ways of embedding responsible AI principles into metadata annotation workflows, including structured human oversight, transparent error reporting, and context-sensitive schema adaptation. In particular, developing flexible, domain-adaptive schemas that can be dynamically aligned with diverse biomedical contexts, such as clinical and computational modeling studies, will enable automated annotation to scale responsibly, with human oversight serving as a safeguard against contextually inappropriate outputs.

Conclusions

With our 4-step approach, we showed that it is possible to streamline the extraction of annotation suggestions to achieve a comprehensive metadata description of the datasets underlying a biomedical publication. The interplay of the LLM, biomedical entity validation through PubTator, and the predefined metadata

XSL•FO
RenderX

schema provided a reliable set of annotation suggestions. Providing annotation suggestions that the scientists just need to confirm will decrease the time that researchers need to spend on dataset documentation without compromising the quality of metadata annotation. Overall, the application of an automated 4-step approach represents a promising step toward improving research data FAIRness, as it enables broader and higher-quality annotation coverage compared with current annotation practices.

Acknowledgments

The generative artificial intelligence tool ChatGPT (GPT-4o) was used for copyediting and improving readability.

Funding

The article processing charge was funded by the German Research Foundation (DFG) and the Albert Ludwig University of Freiburg through the Open Access Publishing program. This work was supported by the DFG under the following projects: project-ID 441891347 – SFB 1479, project-ID 499552394 – SFB 1597, project-ID 491676693 – TRR 359, project-ID 256073931 – SFB 1160, project-ID 514483642 – TRR 384, project-ID 259373024 – TRR 167, project-ID 431984000 – SFB 1453, and EXC 2189. The funding body had no role in the design of the study; data collection, analysis, or interpretation; or in the writing of the manuscript.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

KK and CG designed the study, analyzed the data, and wrote the manuscript. All authors contributed to the interpretation of the data and read and approved the final version of the manuscript.

Conflicts of Interest

OG is a coinventor on patent applications for immunomodulators and co-founder of EMUNO Therapeutics, a company developing drugs that control immunity to address unmet clinical needs. None of the drug candidates in patenting or development were used in this study.

Multimedia Appendix 1
Supplemental tables and figures.
[DOCX File, 2930 KB - formative_v9i1e73822_app1.docx]

References

1. Greenberg J, White HC, Carrier S, Scherle R. A metadata best practice for a scientific data repository. J Libr Metadata 2009 Dec 10;9(3-4):194-212. [doi: 10.1080/19386380903405090]
2. Van Noorden R. Data-sharing: everything on display. Nature 2013 Aug;500(7461):243-245. [doi: 10.1038/nj7461-243a]
3. Curdt C, Hoffmeister D. Research data management services for a multidisciplinary, collaborative research project: design and implementation of the TR32DB project database. Program Electron Libr Inf Syst 2015;49(4):494-512. [doi: 10.1108/PROG-02-2015-0016]
4. Kleinert S, Bartz-Bazzanella P, von der Decken C, et al. A real-world rheumatology registry and research consortium: the German RheumaDatenRhePort (RHADAR) registry. J Med Internet Res 2021 May 20;23(5):e28164. [doi: 10.2196/28164] [Medline: 34014170]
5. Gierend K, Waltemath D, Ganslandt T, Siegel F. Traceable research data sharing in a German Medical Data Integration Center with FAIR (Findability, Accessibility, Interoperability, and Reusability)-geared provenance implementation: proof-of-concept study. JMIR Form Res 2023 Dec 7;7:e50027. [doi: 10.2196/50027] [Medline: 38060305]
6. Vorisek CN, Lehne M, Klopfenstein SA, et al. Fast Healthcare Interoperability Resources (FHIR) for interoperability in health research: systematic review. JMIR Med Inform 2022 Jul 19;10(7):e35724. [doi: 10.2196/35724] [Medline: 35852842]
7. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016 Mar 15;3:160018. [doi: 10.1038/sdata.2016.18] [Medline: 26978244]
8. Engel F, Kaier K, Benadi G, et al. Development of metadata schemas for Collaborative Research Centers. FreiData 2025. [doi: 10.60493/k1xe3-npc10]
9. Watter M, Kahle L, Brunswiek B, et al. Standardized metadata collection to reinforce collaboration in collaborative research centers. Presented at: Proceedings of the eScience Tage 2023; Mar 1-3, 2023. [doi: 10.11588/heidok.00033131]
10. Nadeau D, Sekine S. A survey of named entity recognition and classification. In: Sekine S, Ranchhod E, editors. Named Entities: John Benjamins Publishing Company; 2009:3-28. [doi: 10.1075/bct.19.03nad]

XSL•FO
RenderX

11. Thirunavukarasu AJ, Ting DS, Elangovan K, Gutierrez L, Tan TF, Ting DS. Large language models in medicine. Nat Med 2023 Aug;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
12. Jin Q, Leaman R, Lu Z. Retrieve, summarize, and verify: how will ChatGPT affect information seeking from the medical literature? J Am Soc Nephrol 2023 Aug 1;34(8):1302-1304. [doi: 10.1681/ASN.0000000000000166] [Medline: 37254254]
13. Tian S, Jin Q, Yeganova L, et al. Opportunities and challenges for ChatGPT and large language models in biomedicine and health. Brief Bioinform 2023 Nov 22;25(1):bbad493. [doi: 10.1093/bib/bbad493] [Medline: 38168838]
14. Guo Z, Lai A, Thygesen JH, Farrington J, Keen T, Li K. Large language models for mental health applications: systematic review. JMIR Ment Health 2024 Oct 18;11:e57400. [doi: 10.2196/57400] [Medline: 39423368]
15. Chelli M, Descamps J, Lavoué V, et al. Hallucination rates and reference accuracy of ChatGPT and bard for systematic reviews: comparative analysis. J Med Internet Res 2024 May 22;26:e53164. [doi: 10.2196/53164] [Medline: 38776130]
16. Aljamaan F, Temsah MH, Altamimi I, et al. Reference hallucination score for medical artificial intelligence chatbots: development and usability study. JMIR Med Inform 2024 Jul 31;12:e54345. [doi: 10.2196/54345] [Medline: 39083799]
17. Chan BJ, Chen CT, Cheng JH, Huang HH. Don't do RAG: when cache-augmented generation is all you need for knowledge tasks. arXiv. Preprint posted online on Dec 20, 2024. [doi: 10.48550/arXiv.2412.15605]
18. He T, Kreimeyer K, Najjar M, et al. AI-assisted knowledge discovery in biomedical literature to support decision-making in precision oncology. arXiv. Preprint posted online on Dec 12, 2024. [doi: 10.48550/arXiv.2412.08900]
19. Wei CH, Allot A, Lai PT, et al. PubTator 3.0: an AI-powered literature resource for unlocking biomedical knowledge. Nucleic Acids Res 2024 Jul 5;52(W1):W540-W546. [doi: 10.1093/nar/gkae235] [Medline: 38572754]
20. GPTs augmented by PubTator 3.0. GitHub. URL: https://github.com/ncbi-nlp/pubtator-gpt [accessed 2025-11-05]
21. Schwarzer G, Rücker G. Meta-analysis of proportions. In: Evangelou E, Veroniki AA, editors. Meta-Research: Methods and Protocols: Springer; 2021:159-172. [doi: 10.1007/978-1-0716-1566-9_10]
22. Miller JJ. The inverse of the Freeman – Tukey double arcsine transformation. Am Stat 1978;32(4):138. [doi: 10.1080/00031305.1978.10479283]
23. Research ethics. KonsortSWD. URL: https://www.konsortswd.de/en/topics/research-ethics/ [accessed 2025-11-10]
24. Neuwirt E, Magnani G, Ćiković T, et al. Tyrosine kinase inhibitors can activate the NLRP3 inflammasome in myeloid cells through lysosomal damage and cell lysis. Sci Signal 2023 Jan 17;16(768):eabh1083. [doi: 10.1126/scisignal.abh1083] [Medline: 36649377]
25. Saller BS, Wöhrle S, Fischer L, et al. Acute suppression of mitochondrial ATP production prevents apoptosis and provides an essential signal for NLRP3 inflammasome activation. Immunity 2025 Jan 14;58(1):90-107. [doi: 10.1016/j.immuni.2024.10.012] [Medline: 39571574]
26. Apostolova P, Kreutmair S, Toffalori C, et al. Phase II trial of hypomethylating agent combined with nivolumab for acute myeloid leukaemia relapse after allogeneic haematopoietic cell transplantation-Immune signature correlates with response. Br J Haematol 2023 Oct;203(2):264-281. [doi: 10.1111/bjh.19007] [Medline: 37539479]
27. Biavasco F, Ihorst G, Wäsch R, et al. Therapy response of glucocorticoid-refractory acute GVHD of the lower intestinal tract. Bone Marrow Transplant 2022 Oct;57(10):1500-1506. [doi: 10.1038/s41409-022-01741-3] [Medline: 35768570]
28. Czech M, Schneider S, Peltokangas N, et al. Lipocalin-2 expression identifies an intestinal regulatory neutrophil population during acute graft-versus-host disease. Sci Transl Med 2024 Feb 21;16(735):eadi1501. [doi: 10.1126/scitranslmed.adi1501] [Medline: 38381845]
29. Edwards-Hicks J, Apostolova P, Buescher JM, et al. Phosphoinositide acyl chain saturation drives CD8+ effector T cell signaling and function. Nat Immunol 2023 Mar;24(3):516-530. [doi: 10.1038/s41590-023-01419-y] [Medline: 36732424]
30. Grüninger PK, Uhl F, Herzog H, et al. Functional characterization of the PI3K/AKT/MTOR signaling pathway for targeted therapy in B-precursor acute lymphoblastic leukemia. Cancer Gene Ther 2022 Nov;29(11):1751-1760. [doi: 10.1038/s41417-022-00491-0] [Medline: 35794338]
31. Haring E, Andrieux G, Uhl FM, et al. Therapeutic targeting of endoplasmic reticulum stress in acute graft-versus-host disease. Haematologica 2022 Jul 1;107(7):1538-1554. [doi: 10.3324/haematol.2021.278387] [Medline: 34407601]
32. Ho J, Schmidt D, Lowinus T, et al. Targeting MDM2 enhances antileukemia immunity after allogeneic transplantation via MHC-II and TRAIL-R1/2 upregulation. Blood 2022 Sep 8;140(10):1167-1181. [doi: 10.1182/blood.2022016082] [Medline: 35853161]
33. Jaeger A, Gambheer SM, Sun X, et al. Activated granulocytes and inflammatory cytokine signaling drive T-cell lymphoma progression and disease symptoms. Blood 2023 Jun 8;141(23):2824-2840. [doi: 10.1182/blood.2022015653] [Medline: 36696631]
34. Maas-Bauer K, Stell AV, Yan KL, et al. ROCK1/2 signaling contributes to corticosteroid-refractory acute graft-versus-host disease. Nat Commun 2024 Jan 10;15(1):446. [doi: 10.1038/s41467-024-44703-7] [Medline: 38199985]
35. Schmidt D, Endres C, Hoefflin R, et al. Oncogenic calreticulin induces immune escape by stimulating TGFβ expression and regulatory T-cell expansion in the bone marrow microenvironment. Cancer Res 2024 Sep 16;84(18):2985-3003. [doi: 10.1158/0008-5472.CAN-23-3553] [Medline: 38885318]
36. Socié G, Niederwieser D, von Bubnoff N, et al. Prognostic value of blood biomarkers in steroid-refractory or steroid-dependent acute graft-versus-host disease: a REACH2 analysis. Blood 2023 Jun 1;141(22):2771-2779. [doi: 10.1182/blood.2022018579] [Medline: 36827620]

XSL•FO
RenderX

37. Talvard-Balland N, Braun LM, Dixon KO, et al. Oncogene-induced TIM-3 ligand expression dictates susceptibility to anti-TIM-3 therapy in mice. J Clin Invest 2024 Jun 25;134(16):e177460. [doi: 10.1172/JCI177460] [Medline: 38916965]
38. Villa M, Sanin DE, Apostolova P, et al. Prostaglandin E2 controls the metabolic adaptation of T cells to the intestinal microenvironment. Nat Commun 2024 Jan 11;15(1):451. [doi: 10.1038/s41467-024-44689-2] [Medline: 38200005]
39. Vinnakota JM, Biavasco F, Schwabenland M, et al. Targeting TGFβ-activated kinase-1 activation in microglia reduces CAR T immune effector cell-associated neurotoxicity syndrome. Nat Cancer 2024 Aug;5(8):1227-1249. [doi: 10.1038/s43018-024-00764-7] [Medline: 38741011]
40. Zeiser R, Lee SJ. Three US Food and Drug Administration-approved therapies for chronic GVHD. Blood 2022 Mar 17;139(11):1642-1645. [doi: 10.1182/blood.2021014448] [Medline: 35081254]
41. de Maissin H, Groß PR, Mohiuddin O, et al. In Vivo metabolic imaging of [1-13 C]Pyruvate-d3 Hyperpolarized by reversible exchange with parahydrogen. Angew Chem Int Ed Engl 2023 Sep 4;62(36):e202306654. [doi: 10.1002/anie.202306654] [Medline: 37439488]
42. Frueh JT, Campe J, Sunaga-Franze DY, et al. Interferon regulatory factor 4 plays a pivotal role in the development of aGVHD-associated colitis. Oncoimmunology 2024;13(1):2296712. [doi: 10.1080/2162402X.2023.2296712] [Medline: 38170159]
43. Chen YB, Mohty M, Zeiser R, et al. Vedolizumab for the prevention of intestinal acute GVHD after allogeneic hematopoietic stem cell transplantation: a randomized phase 3 trial. Nat Med 2024 Aug;30(8):2277-2287. [doi: 10.1038/s41591-024-03016-4] [Medline: 38844797]
44. Zeiser R, Socié G, Schroeder MA, et al. Efficacy and safety of itacitinib versus placebo in combination with corticosteroids for initial treatment of acute graft-versus-host disease (GRAVITAS-301): a randomised, multicentre, double-blind, phase 3 trial. Lancet Haematol 2022 Jan;9(1):e14-e25. [doi: 10.1016/S2352-3026(21)00367-7] [Medline: 34971577]
45. El Khawanky N, Hughes A, Yu W, et al. Demethylating therapy increases anti-CD123 CAR T cell cytotoxicity against acute myeloid leukemia. Nat Commun 2021 Nov 8;12(1):6436. [doi: 10.1038/s41467-021-26683-0] [Medline: 34750374]
46. Langenbach M, Giesler S, Richtsfeld S, et al. MDM2 inhibition enhances immune checkpoint inhibitor efficacy by increasing IL15 and MHC class II production. Mol Cancer Res 2023 Aug 1;21(8):849-864. [doi: 10.1158/1541-7786.MCR-22-0898] [Medline: 37071397]
47. Alyafeai Z, Al-Shaibani MS, Ghanem B. M. MOLE: metadata extraction and validation in scientific papers using llms. arXiv. Preprint posted online on May 26, 2025. [doi: 10.48550/arXiv.2505.19800]
48. Schilling-Wilhelmi M, Ríos-García M, Shabih S, et al. From text to insight: large language models for chemical data extraction. Chem Soc Rev 2025 Feb 3;54(3):1125-1150. [doi: 10.1039/d4cs00913d] [Medline: 39703015]
49. Singh M, Kumar A, Donaparthi S, Karambelkar G. Leveraging retrieval augmented generative LLMs for automated metadata description generation to enhance data catalogs. arXiv. Preprint posted online on Mar 12, 2025. [doi: 10.5121/csit.2025.150303]
50. Turner MD, Appaji A, Ar Rakib N, et al. Large language models can extract metadata for annotation of human neuroimaging publications. bioRxiv. Preprint posted online on May 14, 2025. [doi: 10.1101/2025.05.13.653828] [Medline: 40462943]
51. Liu NF, Lin K, Hewitt J, et al. Lost in the middle: how language models use long contexts. Trans Assoc Comput Linguist 2024 Feb;12(5):157-173. [doi: 10.1162/tacl_a_00638]
52. Lobentanzer S, Aloy P, Baumbach J, et al. Democratizing knowledge representation with BioCypher. Nat Biotechnol 2023 Aug;41(8):1056-1059. [doi: 10.1038/s41587-023-01848-y] [Medline: 37337100]
53. Lobentanzer S, Feng S, Bruderer N, et al. A platform for the biomedical application of large language models. Nat Biotechnol 2025 Feb;43(2):166-169. [doi: 10.1038/s41587-024-02534-3] [Medline: 39843580]

Abbreviations

AI: artificial intelligence
CAG: cache-augmented generation
CRC: Collaborative Research Center
FAIR: findability, accessibility, interoperability, and reusability
LLM: large language model
NLP: natural language processing
RAG: retrieval-augmented generation

XSL•FO
RenderX

Edited by J Sarvestan; submitted 12.03.25; peer-reviewed by R Abeysinghe, Z Wang; revised version received 26.09.25; accepted 26.09.25; published 20.11.25.

Please cite as:
Giuliani C, Benadi G, Engel F, Werner J, Watter M, Schwarzer G, Groß O, Zeiser R, Binder H, Kaier K
Identifying Biomedical Entities for Datasets in Scientific Articles: 4-Step Cache-Augmented Generation Approach Using GPT-4o and PubTator 3.0
JMIR Form Res 2025;9:e73822
URL: https://formative.jmir.org/2025/1/e73822
doi:10.2196/73822

© Claudia Giuliani, Gita Benadi, Felix Engel, Jonas Werner, Manuel Watter, Guido Schwarzer, Olaf Groß, Robert Zeiser, Harald Binder, Klaus Kaier. Originally published in JMIR Formative Research (https://formative.jmir.org), 20.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Assessing the Usability of a Novel Toolkit for Creating Visual Key Information Pages for Informed Consent for Research: Mixed Methods Usability Study

Eliana Charlotte Goldstein[1], MPH; Zoe Troubh[2], BA; Krista E Cooksey[1], BA; Molly Volkmar[3], BA; Victor Catalan Gallegos[4], MPH; Kimberly A Kaphingst[3,5], SCD; Clara N Lee[4], MPP, MD; Ashley J Housten[6], OTD, MSc, OTR/L; Jessica Mozersky[7], MBE, PhD; Mary C Politi[1], PhD

[1]School of Public Health, Washington University in St. Louis, 4300 Duncan Ave, Room 2110, St. Louis, MO, United States

[2]Washington University in St. Louis, St. Louis, MO, United States

[3]Huntsman Cancer Institute, University of Utah, Salt Lake City, UT, United States

[4]Department of Surgery, Division of Plastic and Reconstructive Surgery, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States

[5]Department of Communication, University of Utah, Salt Lake City, UT, United States

[6]Department of Surgery, Division of Public Health Sciences, School of Medicine, Washington University in St. Louis, St. Louis, MO, United States

[7]Department of Medicine, Division of General Medicine and Geriatrics, School of Medicine, Washington University in St. Louis, St. Louis, MO, United States

Corresponding Author:
Mary C Politi, PhD
School of Public Health, Washington University in St. Louis, 4300 Duncan Ave, Room 2110, St. Louis, MO, United States

Abstract

Background: Key information pages for informed consent require a concise summary of information to improve participant understanding but have not widely incorporated health literacy best practices.

Objective: We previously developed a visual key information template to improve informed consent. In this study, we conducted usability testing of this customizable one-page key information template.

Methods: We used the Designing for Accelerated Translation framework to plan for actionable, efficient usability testing. Participants (N=15) were asked to spend about 20 minutes using the visual key information template and engaging in a think-aloud protocol. They then responded to qualitative debrief questions about the template and validated measures of acceptability, appropriateness, and feasibility. Interviews were recorded, transcribed, and analyzed with a usability-focused codebook and thematic analysis.

Results: The toolkit was positively received. Common usability challenges included interpreting instructions, condensing consent content, replacing and resizing icons, and fitting information into template boxes. Participants had positive experiences with toolkit elements, particularly with the icon library, and generally felt the toolkit was easy to use and encouraged simplification of information. Some participants noted not fully reviewing instructions before the study and discussed specific technical abilities as potential limitations of widespread use. We documented suggestions and made changes to the toolkit in response to feedback received.

Conclusions: Overall, participants considered the toolkit to be appropriate, acceptable, and feasible. Additional implementation outcomes are being collected in a multisite stepped-wedge randomized trial. Further research may investigate changes to format and software that balances functionality with ease of use.

(JMIR Form Res 2025;9:e76740) doi:10.2196/76740

KEYWORDS

informed consent; health literacy; key information; usability testing; think-aloud protocol; qualitative mixed-methods

XSL•FO
RenderX

Introduction

Background

Informed consent documents for research studies aim to introduce research participants to important topics including research design, risks and benefits, data privacy, and participant rights. Research and regulatory oversight teams often struggle to clearly present these complex details, which results in consent forms that are lengthy, repetitive, and difficult to understand [1-5]. The introduction of the key information (KI) page in the 2018 Federal Common Rule sought to address these challenges by requiring a "concise and focused summary" of information most important to potential participants who are deciding whether to join a study [6]. Although evidence-based recommendations for how best to use KI pages are limited to date, the guidance that does exist encourages flexibility to adapt information to the audience [7,8]. Existing research has explored the use of visuals and health literacy best practices in informed consent broadly [9-13]; however, to our knowledge, there is no available, evidence-based guidance, or toolset for integrating these elements into the KI page specifically.

Prior Work

In previous research, our team developed and tested options for a one-page KI template incorporating both visual elements and health literacy best practices [14,15]. The goal of this template is to support existing consent best practices with an appealing, clear, and actionable first page of research consent forms. We compared four potential templates for visually enhanced key information pages designed by either our research team or a nonprofit communications organization that specializes in disseminating health information in an accessible way [14]. The options were qualitatively reviewed by 40 users (principal investigators, research staff, institutional review board (IRB) members and staff, and community members) and iteratively updated in response to their feedback. The key elements of the final template preferred by users included organizational boxes with contrasting headers, icons, color, bulleted text, and ample white space, with accurate, accessible, and actionable information, consistent with health literacy best practices [16]. In parallel, we conducted pilot testing of templates leveraging similar elements in four ongoing studies [15]. In pilot tests, participants reported high knowledge of study details, high approval of the visuals, and low conflict about joining studies. These pilots also served as an opportunity to identify and address important implementation considerations and refinements of the template to ensure adaptability for varying study types and subjects.

After selecting and refining our approach, we developed a customizable toolkit in Microsoft PowerPoint [14,17]. The toolkit includes an editable version of the template, instructional documents and videos, an icon library, and examples to empower study teams to use the proposed template as a replacement for text-only key information pages in informed consent documents.

Study Objectives

Our first phase of research collected feedback on already-created key information templates but did not ascertain if the resources in the resulting toolkit would be usable for research teams interested in creating these pages for their own studies. This paper describes mixed-methods usability testing of the toolkit elements and template as part of our rapid dissemination and implementation approach to ensure the toolkit would be acceptable, appropriate, and feasible for use by interested research teams.

Methods

Frameworks

Our approach to the design, implementation, and dissemination of this intervention was guided by the Designing for Accelerated Translation (DART) framework [18]. The DART framework outlines the connections between three elements key to quickly putting innovations into practice: (1) forming meaningful partnerships; (2) designing for dissemination and implementation; and (3) engaging with a learning system to iteratively adapt and refine materials. The protocol for this study responds to these elements (Table 1), with the goal of rapidly assessing and responding to common usability issues with the proposed template and toolkit.

Table . Usability study protocol elements responding to the Designing for Accelerated Translation (DART) Framework [18].

DART element	Protocol element
Form meaningful partnerships	• Engage with likely end users as participants (principal investigators, research coordinators, or research support staff) • Collaborate with the institutional review board to ensure the visual KI pages could be integrated routinely
Designing for dissemination and implementation	• Select commonly available software for the toolkit • Provide proposed resources to support users • Update resources in response to feedback • Monitor interviews for thematic saturation
Learning systems for rapid adaptation	• Encourage participants to test template using their specific consent and study protocol • Provide poststudy support to participants interested in implementing visual key information into their study • Track adaptations using a standard framework

XSL•FO
RenderX

We based our coding framework on previous studies of usability of computer-based job tools, which organized usability challenges within four domains: navigation, content, functionality, and layout [19,20]. Additional description of codebook development is included below. To track suggested changes beyond the scope of the codebook, we used the Framework for Reporting Adaptations and Modifications to Evidence-Based Interventions (FRAME) [21].

Participant Eligibility and Recruitment

We invited principal investigators and research staff at three large research universities with affiliated medical schools (one in the Midwest, one in the South, and one in the Mountain West) to complete usability testing via email. Potential participants were identified from previously compiled lists of potential participants developed based on peer recommendations of site lead investigators (KK, CL, MP) and targeted searches for people in relevant roles with publicly available email addresses, as well as via snowball sampling and word-of-mouth referral [22]. Eligible participants included (1) principal investigators, (2) research team members (eg, coordinators, project managers), or (3) research support staff with a focus on protocol development or recruitment. Based on standard sample sizes for usability testing [23], we planned to recruit 10‑15 participants in an initial round of recruitment, with an approximately even split across the three categories of professional roles. If new usability considerations were identified in the last 5 interviews, the research team planned to recruit additional participants.

Data Collection

We designed a think-aloud protocol for usability testing based on examples from the literature [19,20,24]. We sent study materials (the visual KI toolkit described above, a brief summary document, and a 5 min introductory video) to participants in advance of scheduled interviews for their review and preparation. Participants spent approximately 20 minutes filling out the visual KI template based on an existing consent form (ie, their own or an example provided by the study team, per participant preference) before the research team asked them to stop. Participants were asked to screenshare using video conference software as they completed the task. Interviewers prompted participants with a predetermined list of questions if they were silent for more than 20 seconds (Multimedia Appendix 1). To minimize participant burden, participants were instructed to complete as much of the template as they were able during the allotted time; there was no necessary minimum completion. After testing, interviewers collected responses to debrief questions and validated measures of acceptability, appropriateness, and feasibility each rated on a 5-point scale (strongly disagree=1 to strongly agree=5) [1,25]

Data Analysis

Interviews were recorded, transcribed, and divided into observational and debrief portions for coding [26]. A usability-focused codebook was developed by three research team members (EG, KC, MP), based on anticipated challenges and challenges that emerged from the first five participant interviews and refined as needed throughout the coding process. The observational portions of the interviews were each coded to 21 total codes by two coders (EG, ZT). Coding was compared to identify discrepancies (ie, one coder selected a code while the other did not). Coders then met to review the discrepancies and consulted a third coder (KC) to reach consensus for each transcript. The average percent agreement across all transcripts was 85% prior to consensus meetings. A challenge was considered frequent if at least eight (>50%) of participants experienced it one or more times in their interview.

Several research team members (EG, MV, VC, KC) developed additional codes to analyze the broader scope of the debrief portion of the interviews, using an inductive methodology described by Naeem et al [27]. Debrief portions were coded collaboratively by the two coders (EG, ZT) with consultation from the third coder (KC). Acceptability, appropriateness, and feasibility scores were calculated based on standard methodology for the validated measures [25].

Ethical Considerations

The Washington University in St. Louis Institutional Review Board approved this study as exempt research (IRB #202405126). The research team reviewed consent information with participants and confirmed their interest in joining the study and being recorded to observe study procedures, prior to beginning the usability interview. Written transcripts were deidentified prior to depositing in a data repository. Data are available in the repository via restricted access with a study protocol, ethics approval, and rationale for access [26]. Participants received a US $50 gift card after completing all study procedures.

Results

Participants

Fifty-one eligible potential participants were invited to participate. Of those, 16 agreed to participate, 2 declined, and 33 were unable to be reached. A total of 15 participants completed 14 total interviews (Table 2). In one interview, 2 participants working on the same research project requested to be interviewed simultaneously, with 1 participant controlling the visual KI template while the other offered suggestions and commentary.

XSL•FO
RenderX

Table . Usability interview participant demographics (N=15).

Demographic characteristics	Participants (N=15)
Age (years)	
Mean (SD)	44.2 (12.57)
Range	27 - 76
Gender, n (%)	
Woman	10 (66.7)
Man	4 (26.7)
Another response	1 (6.7)
Ethnicity, n (%)	
Latino/a/Latine or Hispanic	1 (6.7)
Not Latino/a/Latine or Hispanic	14 (93.3)
Race, n (%)	
White	14 (93.3)
Asian	1 (6.7)
Institutional role, n (%)	
Principal Investigator (PI)	6 (40.0)
Research Support Team Member (eg, clinical trial oversight, recruitment enhancement)	3 (20.0)
Research Team Member (eg, coordinators, managers)	6 (40.0)

Observational Period Findings

We assessed the number of participants who reported each identified usability challenge (Table 3). Five frequently observed challenges were noted; 2 were related to content in the toolkit: participants had difficulty interpreting instructions and participants had difficulty condensing content from the full consent document into the visual KI template. Two challenges were related to functionality: participants had technical issues updating information in boxes and participants had trouble replacing or resizing icons. One was related to layout: participants had challenges fitting study information into the boxes provided. There were few navigation challenges other than navigating between pages, which occurred among 4 participants.

XSL•FO
RenderX

Table . Number of participants reporting usability challenges and adaptations made to address challenges (N=15).

Usability challenge	Participants (n, %)	Adaptations to address challenge
Content		
Participant had difficulty interpreting instructions.	8 (53.5)	Clarified instructions; developed demonstration video resource
Participant had difficulty condensing content from full consent into bullets.	8 (53.3)	Developed demonstration video resource
Participant had difficulty splitting content across boxes.	7 (46.7)	Developed demonstration video resource
Icon not available for participant's desired image.	4 (26.7)	Maintaining list of suggested icons
Participant expressed desire to include information beyond the recommended scope of visual KI.	3 (20.0)	N/A[a]
Participant indicated specific information was unavailable/missing.	2 (13.3)	Clarified instructions; developed demonstration video resource
Functionality		
Participant had trouble replacing or resizing icons.	10 (66.7)	Reviewed formatting to ensure appropriate layering of elements; developed demonstration video resource
Participant had technical issues updating information in boxes.	9 (60.0)	Reviewed formatting to ensure appropriate layering of elements; developed demonstration video resource
Participant had other technical issues.	3 (20.0)	N/A
Participant discussed a goal/functionality that cannot be achieved in PowerPoint.	2 (13.3)	Maintaining list of suggested future functionalities
Participant had difficulty manipulating size of boxes or box headers.	1 (6.7)	Developed demonstration video resource
Layout		
Participant said they could not fit desired study information in the boxes provided.	8 (53.3)	Developed demonstration video resource
Participant expressed that the text is too small.	4 (26.7)	Developed template version with larger font size
Participant had trouble creating text wrap.	3 (20.0)	Reviewed formatting to ensure appropriate layering of elements; developed demonstration video resource
Participant had trouble reading side instructions.	1 (6.7)	N/A
Navigation		
Participant had difficulty moving from page to page in toolkit.	4 (26.7)	N/A
Participant unintentionally moved items around.	3 (20.0)	Reviewed formatting to ensure appropriate layering and grouping of elements
Participant unintentionally changed format.	3 (20.0)	Reviewed formatting to ensure appropriate layering and grouping of elements
Participant had difficulty finding desired icon.	2 (13.3)	N/A
Participant initially had difficulty determining which template to use (6 vs 8 box).	1 (6.7)	Developed demonstration video resource
Participant expressed unfamiliarity with PowerPoint.	1 (6.7)	N/A

[a]Not applicable.

Debrief Period Findings

In the debrief portion of the interviews, participants indicated that the toolkit was overall easy to use. They noted that the templated boxes made the task straightforward and encouraged them to simplify some information:

XSL•FO
RenderX

> *Having these headings at the top of each [box] helps to also organize the information just mentally before you're getting started. You can already start thinking about what goes where and what might make sense to a potential participant or patient who's reviewing this.* [P10, research support team member]

> *I think that having a toolkit like this forces you to write things fairly simply and directs you exactly what to do.* [P14, principal investigator]

Some participants referred to their experience as "intuitive" (P06, research team member; P14, principal investigator). Others highlighted that the "inherent flexibility" (P08, principal investigator) of the toolkit met their needs and speculated it would meet others'. However, some noted that the template by itself did not necessarily solve all challenges with the content of informed consent forms, such as a high volume of desired content or a tendency toward jargon:

> *It's hard, especially the goal of the study, if there's a lot involved, it was hard for me to figure out what to put there.* [P12, research team member]

> *I like the idea, but I think what…everybody has to do…is that you need to check the level of language. Because all of us are wrong all the time.* [P01, principal investigator]

Participants also responded positively to the resources in the toolkit, particularly the icon library.

> *I love that [the icon library] is there…it's easy to get lost in a rabbit hole of looking for the right icon and then trying to figure out any copyright privileges...knowing that there's a library of things that I can select from was nice.* [P07, principal investigator]

Not all participants used the resources provided. For example, multiple participants noted that they had not read through the instructions. Some felt the instructions were too long but others appreciated inclusion of background and technical details:

> *These first two pages being so dense…I was like 'Oh, I just want to know what I need to do right now.* [P06, research team member]

> *I've never really seen anything really like it other than just templated fliers…but [with those] you're sort of just on your own. I think it's important to…have that instruction and…an educational piece.* [P11, research support team member]

Participants noted that users' technical ability or experience could limit feasibility of use. One participant was also concerned about the possibility of users inadvertently creating unappealing materials with the template:

> *People who know how to do it, right, they're going to do it. If this is required to be used, they're going to hate it. People who don't—if they don't have that skill in design principles anyway, they're going to [mess] it up.* [P09, research support team member]

Suggestions and Adaptations

Suggestions that arose throughout the interviews were documented and iteratively considered for implementation. In addition to the changes made to address usability issues (Table 3), several revisions were made based on specific recommendations from participants:

- Updated the "default" set of icons on template pages to ensure they are as broadly applicable as possible
- Moved the icons to the bottom right of boxes to minimize issues with text wrapping
- Added further guidance about options for customization
- Clarified best practices for using icons in payment box to minimize sources of confusion

Large-scale changes to the format of the template (eg, adding new visual elements like arrows between boxes, migrating toolkit to a platform that would allow dropdown selection of icons) were beyond the scope of this usability testing; however, these suggestions were documented using the FRAME for potential future work. We are also maintaining a list of proposed additions to the icons for future iterations.

Toolkit Acceptability, Appropriateness, and Feasibility

All 15 participants completed all acceptability, appropriateness, and feasibility measures. The toolkit was generally positively received, with most participants rating it highly across 5-point acceptability, appropriateness, and feasibility scales: mean acceptability score=4.28 (range: 1.50 - 5.00; SD 0.87); mean appropriateness score=4.32 (range: 3.00 - 5.00; SD 0.57); mean feasibility score=4.35 (range: 2.75 - 5.00; SD 0.63) [25].

Discussion

Principal Results

These usability interviews affirmed previous findings that the proposed visual KI template is an acceptable tool for researchers conducting human subjects research [14]. Participants generally found the toolkit for key information both straightforward and flexible to meet their study's needs. Participants in our prior research suggested that the boxes in the toolkit would be useful for participants' comprehension and experience [14]; in this work, we found they are also helpful for researchers thinking through how to write and summarize key consent information. However, it is important to ensure that the flexibility of the template is clear and that instructions are modular so that users can customize the content to fit study-related needs and preferences while meeting overall key information best practices. We refined instructions and created a demonstration video resource to ensure users have access to clear information about when and how to customize major elements of the template.

Although the toolkit addresses many of the formatting issues that can make informed consent forms unappealing or intimidating, participants continued to report that selecting appropriate content for key information remains a challenge (ie, what information is appropriate to include, how to "translate" from jargon-filled protocols to plain language). The template helps to suggest some overall content categories but

XSL•FO
RenderX

does not provide a complete solution to helping individuals summarize important study-related material. This research provided a first step focusing on usability challenges specific to a previously developed informed consent toolkit; future research could explore how to select the most appropriate content for the toolkit within the box headings.

Our selection of Microsoft PowerPoint as the platform for this toolkit has meant that we are unable to address some of participants' specific functionality desires such as including drop-down menus for icon selection. However, we chose to use this program given its broad accessibility in the United States and users' overall comfort with it. Many participants informed us in our previous studies, that introducing a new software would create steep learning curves for those less familiar with technology [14]. Future work could explore the best software to balance ease of use with functionality.

This toolkit goes beyond previous evidence-building and training modules in the area of informed consent improvement [11,12] by offering an adaptable template that is available for use by interested research teams [17]. An editable toolkit like this is substantially more actionable than a video or in-person training. Most participants in this study indicated that the toolkit provided immediate and obvious value to their consent development process, where noninteractive trainings or standalone design examples do not prompt hands-on learning with a usable output. We integrated trainings and design examples as optional supplemental resources for those who want additional guidance rather than the primary elements of the intervention. Research teams developing interventions that require end users to create highly tailored content should consider similarly editable tools with integrated instructions to promote flexible, "learn as you work" approaches for busy professionals.

The usability testing described above was a critical step in ensuring that any obvious barriers to using the toolkit were addressed prior to wider dissemination. The think-aloud protocol was an efficient, effective, and flexible approach to refining the toolkit that others working to create interactive job aids or templates may consider incorporating into their development process. As suggested in the DART framework, engaging end users in qualitative testing has also helped us identify "power users" within research and research support teams who act as early adopters of the toolkit and advocate for its use among their peers [18].

Limitations

These usability interviews were conducted with a limited number of participants working at similarly-resourced academic institutions in the United States; results might not generalize to participants in other institutions. The small sample size included many people with an existing interest in improving informed consent, which may limit our findings of acceptability, appropriateness, and feasibility. In addition, participants may have worked through challenges that, if they were not being observed, may have caused them to give up while using the toolkit on their own. However, we did not note any participants expressing significant frustration during the interview period. About half of participants worked on the task based on their own study protocols or consents (n=8 participants with 7 unique protocols), but many used a sample consent for a previous decision support intervention project. This could have limited the number of study type-specific challenges encountered, particularly because only one protocol included a medication intervention.

Conclusion

Participants generally found the proposed KI toolkit appropriate, acceptable, and feasible. Some usability issues were addressed with enhanced instructional resources, and other suggestions related to the toolkit content were considered and incorporated. Designing our study with reference to elements from the DART framework helped ensure that insights could be implemented quickly in routine use. These updates are contributing to our ongoing testing and dissemination work, which includes a multisite stepped-wedge randomized trial to build evidence for implementation and effectiveness outcomes; collaborations with IRBs to create approved, site-specific workflows for submission and review; and live and asynchronous training and support opportunities for researchers and research staff. Future research could explore guidance for selecting the right content for KI pages and technology opportunities (eg, some have suggested exploring health-literacy-informed prompts for generative artificial intelligence to simplify or shorten consent content for input into the template) that provide more tailored support for users developing visual KI pages.

Acknowledgments

The authors thank Institutional Review Board leaders and staff at our universities for their support of the development and implementation of these visual key information pages in research consent at our institutions.

Funding

The usability interviews for this project were supported by the Agency for Healthcare Research and Quality (AHRQ, Grant R01 HS029310).

Data Availability

The data analyzed for this article are available under controlled access [26].

XSL•FO
RenderX

Conflicts of Interest

MCP was a consultant for UCB Biopharma and EpiQ in 2024 on topics unrelated to the content of this manuscript.

Multimedia Appendix 1
Think-aloud prompts used with participants after approximately 20 seconds of silence.
[DOCX File, 15 KB - formative_v9i1e76740_app1.docx]

References

1. Burks AC, Keim-Malpass J. Health literacy and informed consent for clinical trials: a systematic review and implications for nurses. Nursing: Research and Reviews 2019:31-40. [Medline: 40616149]
2. Cooksey KE, Mozersky J, DuBois J, Kuroki L, Marx CM, Politi MC. Challenges and possible solutions to adapting to virtual recruitment: lessons learned from a survey study during the COVID‐19 pandemic. Ethics & Human Research 2022 Nov;44(6):23-31. [doi: 10.1002/eahr.500148]
3. Dickert NW, Eyal N, Goldkind SF, et al. Reframing consent for clinical research: a function-based approach. Am J Bioeth 2017 Dec;17(12):3-11. [doi: 10.1080/15265161.2017.1388448] [Medline: 29148951]
4. Mandava A, Pace C, Campbell B, Emanuel E, Grady C. The quality of informed consent: mapping the landscape. A review of empirical data from developing and developed countries. J Med Ethics 2012 Jun;38(6):356-365. [doi: 10.1136/medethics-2011-100178] [Medline: 22313664]
5. Simonds VW, Garroutte EM, Buchwald D. Health literacy and informed consent materials: designed for documentation, not comprehension of health research. J Health Commun 2017 Aug;22(8):682-691. [doi: 10.1080/10810730.2017.1341565] [Medline: 28759329]
6. Part 46: protection of human subjects. sect. 46.117: documentation of informed consent. Code of Federal Regulations.: National Archives; 2018 URL: http://ecfr.gov/current/title-45/subtitle-A/subchapter-A/part-46/subpart-A/section-46.117 [accessed 2025-04-11]
7. Mozersky J, Wroblewski MP, Solomon ED, DuBois JM. How are US institutions implementing the new key information requirement? J Clin Transl Sci 2020 Jan 10;4(4):365-369. [doi: 10.1017/cts.2020.1] [Medline: 33244420]
8. Draft guidance – key information and facilitating understanding in informed consent. Office for Human Research Protections. 2024. URL: https://www.hhs.gov/ohrp/regulations-and-policy/requests-for-comments/draft-guidance-key-information-facilitating-understanding-informed-consent/index.html [accessed 2025-04-11]
9. Solomon ED, Mozersky J, Goodman M, et al. A randomized implementation trial to increase adoption of evidence-informed consent practices. J Clin Transl Sci 2023;7(1):e28. [doi: 10.1017/cts.2022.520] [Medline: 36721403]
10. Bierer BE, Baedorf Kassis S. Communicating complex numeric information in clinical research. Front Commun 2023;8:1096271. [doi: 10.3389/fcomm.2023.1096271]
11. Shoemaker SJ, Brach C, Edwards A, Chitavi SO, Thomas R, Wasserman M. Opportunities to improve informed consent with AHRQ training modules. Jt Comm J Qual Patient Saf 2018 Jun;44(6):343-352. [doi: 10.1016/j.jcjq.2017.11.010] [Medline: 29793885]
12. Nishimura A, Carey J, Erwin PJ, Tilburt JC, Murad MH, McCormick JB. Improving understanding in the research informed consent process: a systematic review of 54 interventions tested in randomized control trials. BMC Med Ethics 2013 Jul 23;14(28):23879694. [doi: 10.1186/1472-6939-14-28] [Medline: 23879694]
13. Davis TC, Berkel HJ, Holcombe RF, Pramanik S, Divers SG. Informed Consent for Clinical Trials: a comparative study of standard versus simplified forms. JNCI Journal of the National Cancer Institute 1998 May 6;90(9):668-674. [doi: 10.1093/jnci/90.9.668]
14. Cooksey KE, Goldstein E, Lee C, et al. Qualitative assessment of proposed visual key information pages for informed consent. J Clin Transl Sci 2024;8(1):e218. [doi: 10.1017/cts.2024.662] [Medline: 39790474]
15. Hill A, Housten AJ, Cooksey K, Goldstein E, Mozersky J, Politi MC. Accelerating implementation of visual key information to improve informed consent in research: a single-institution feasibility study and implementation testing. BMJ Open 2025 Mar 18;15(3):e092185. [doi: 10.1136/bmjopen-2024-092185] [Medline: 40107701]
16. Health literacy: develop & test materials. Centers for Disease Control and Prevention. 2024. URL: https://www.cdc.gov/health-literacy/php/develop-materials/develop-test-materials.html [accessed 2025-08-11]
17. Project SCORE (streamlined consents for research): study aims. Washington University in St Louis School of Public Health Politi Lab. URL: https://politilab.wustl.edu/research/project-score/study-aims [accessed 2025-04-11]
18. Ramsey AT, Proctor EK, Chambers DA, et al. Designing for Accelerated Translation (DART) of emerging innovations in health. J Clin Transl Sci 2019 Jun;3(2-3):53-58. [doi: 10.1017/cts.2019.386] [Medline: 31528365]
19. Murray-Torres T, Casarella A, Bollini M, Wallace F, Avidan MS, Politi MC. Anesthesiology Control Tower-Feasibility Assessment to Support Translation (ACTFAST): mixed-methods study of a novel telemedicine-based support system for the operating room. JMIR Hum Factors 2019 Apr 23;6(2):e12155. [doi: 10.2196/12155] [Medline: 31012859]
20. Zhao T, McDonald S, Edwards HM. The impact of two different think-aloud instructions in a usability test: a case of just following orders? Behaviour & Information Technology 2014 Feb;33(2):163-183. [doi: 10.1080/0144929X.2012.708786]

XSL•FO
RenderX

21. Wiltsey Stirman S, Baumann AA, Miller CJ. The FRAME: an expanded framework for reporting adaptations and modifications to evidence-based interventions. Implement Sci 2019 Jun 6;14(1):58. [doi: 10.1186/s13012-019-0898-y] [Medline: 31171014]
22. Parker C, Scott S, Geddes A. Snowball Sampling: SAGE research methods foundations; 2019. [doi: 10.4135/9781526421036831710]
23. Lewis JR, Sauro J. Usability and user experience: design and evaluation. In: Handbook of Human Factors and Ergonomics 2021:972-1015. [doi: 10.1002/9781119636113]
24. Wolcott MD, Lobczowski NG. Using cognitive interviews and think-aloud protocols to understand thought processes. Curr Pharm Teach Learn 2021 Feb;13(2):181-188. [doi: 10.1016/j.cptl.2020.09.005] [Medline: 33454077]
25. Weiner BJ, Lewis CC, Stanick C, et al. Psychometric assessment of three newly developed implementation outcome measures. Implement Sci 2017 Aug 29;12(1):108. [doi: 10.1186/s13012-017-0635-3] [Medline: 28851459]
26. Goldstein EC, Troubhh Z, Cooksey KE, Politi MC. Data for: Assessing the usability of a novel toolkit for creating visual key information pages for Informed Consent for Research V1. Qualitative Data Repository 2025. [doi: 10.5064/F6XJ82VS]
27. Naeem M, Ozuem W, Howell K, Ranfagni S. A step-by-step process of thematic analysis to develop a conceptual model in qualitative research. Int J Qual Methods 2023 Oct;22:16094069231205789. [doi: 10.1177/16094069231205789]

Abbreviations

DART: Designing for Accelerated Translation
FRAME: Framework for Reporting Adaptations and Modifications to Evidence-Based Interventions
IRB: institutional review board
KI: key information

Edited by A Mavragani, J Sarvestan; submitted 29.04.25; peer-reviewed by C Baur; revised version received 10.10.25; accepted 27.10.25; published 20.11.25.

Please cite as:
Goldstein EC, Troubh Z, Cooksey KE, Volkmar M, Catalan Gallegos V, Kaphingst KA, Lee CN, Housten AJ, Mozersky J, Politi MC Assessing the Usability of a Novel Toolkit for Creating Visual Key Information Pages for Informed Consent for Research: Mixed Methods Usability Study
JMIR Form Res 2025;9:e76740
URL: https://formative.jmir.org/2025/1/e76740
doi:10.2196/76740

© Eliana Charlotte Goldstein, Zoe Troubh, Krista E Cooksey, Molly Volkmar, Victor Catalan Gallegos, Kimberly A Kaphingst, Clara N Lee, Ashley J Housten, Jessica Mozersky, Mary C Politi. Originally published in JMIR Formative Research (https://formative.jmir.org), 20.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Effects, Feasibility, and Safety of an Early Mobilization Protocol With Immersive Virtual Reality for Dyspnea in Patients With Acutely Decompensated Heart Failure: The MOVE Randomized Clinical Trial

Iasmin Borges Fraga[1], PT, MSc, ScD; Larissa Gussatschenko Caballero[2,3], RN, MSc, ScD; Carlos Eduardo Maciel Tremea[4], PT; Janaína dos Santos Prates[3], RN; Gabrielle Perin[3], NS; Vitor Alves Guedes[3], RN; Pedro Dal Lago[5], PT, MSc, ScD; Eneida Rejane Rabelo-Silva[6,7], RN, MSc, ScD

[1]Graduate Program in Cardiology and Cardiovascular Sciences, Faculty of Medicine, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil

[2]Heart Failure Clinic, Cardiology Division, Hospital de Clínicas de Porto Alegre, Porto Alegre, Brazil

[3]Graduate Program in Nursing, Nursing School, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil

[4]Graduate Program in Rehabilitation Sciences, Physiotherapy School, Universidade Federal de Ciências da Saúde de Porto Alegre, Porto Alegre, Brazil

[5]Physiotherapy Department, Graduate Program in Rehabilitation Sciences, Universidade Federal de Ciências da Saúde de Porto Alegre, Porto Alegre, Brazil

[6]Vascular Access Program, Cardiology Division, Hospital de Clínicas de Porto Alegre, Porto Alegre, Brazil

[7]Graduate Program in Nursing and Graduate Program in Cardiology and Cardiovascular Sciences, Universidade Federal do Rio Grande do Sul, 963 São Manoel Street, Rio Branco, Porto Alegre, Brazil

Corresponding Author:
Eneida Rejane Rabelo-Silva, RN, MSc, ScD
Vascular Access Program, Cardiology Division, Hospital de Clínicas de Porto Alegre, Porto Alegre, Brazil

Abstract

Background: Early mobilization seems to benefit patients with acutely decompensated heart failure (ADHF), but its initiation is challenging due to severe dyspnea, clinical instability, and low adherence to treatment. Understanding whether early mobilization has good acceptance, adherence, and safety is key for rehabilitation in this population. Virtual reality (VR) offers a less stressful environment and may reduce dyspnea, yet its effects on ADHF remain unknown. In addition, the feasibility and safety of combining VR with an early mobilization program in the intensive care unit (ICU) setting are not fully established.

Objective: This study aimed to assess the effects, feasibility, and safety of an early mobilization protocol combined with immersive VR for dyspnea in patients with ADHF admitted to an ICU.

Methods: The Early Mobilization Protocol with Immersive Virtual Reality (MOVE) study is a single-center, parallel, superiority randomized clinical trial conducted from January 2023 to January 2024 in a teaching hospital in Brazil. Patients with ADHF admitted to the ICU who met the eligibility criteria were invited to participate. After informed consent, participants were electronically randomized into the intervention group (IG) and control group (CG). Both groups underwent up to 3 sessions of the early mobilization protocol supervised by a physiotherapist, including upper- and lower-limb cycle ergometry, standing, and ambulation. Additionally, the IG used VR headsets, headphones, and smartphones displaying 360° videos. The primary outcome was dyspnea measured using the modified Borg scale before and after each session. Secondary outcomes included vital signs before and after each session and the occurrence of adverse events. Physiotherapists were blinded to the primary outcome, and all analyses followed the intention-to-treat principle with a blinded statistician.

Results: A total of 58 participants were randomized (IG: n=28, 48%; CG: n=30, 52%), with a mean age of 59 (SD 11.6) years; 42 (72%) were men, the mean left ventricular ejection fraction was 26.6% (SD 12.5%), and 28 (48%) were categorized as New York Heart Association class III. Only 43% (25/58) of the participants completed all 3 protocol sessions (IG: 11/28, 39%; CG: 14/30, 47%). Reasons for noncompletion of the protocol included refusal, clinical instability, discharge from the unit, or scheduled procedures. Changes in the mean dyspnea scores were similar between groups (IG: −0.17, SD 1.68; CG: 0.01, SD 1.73; P=.67). The mean vital signs remained within expected clinical ranges, with no differences between groups (all P>.05). No serious adverse events occurred.

Conclusions: Early mobilization with or without VR was feasible and safe in ICU patients with ADHF. VR did not significantly reduce dyspnea, but adherence challenges limited protocol completion. These findings suggest that early mobilization can be safely implemented in this population and that future studies should explore strategies to enhance adherence and evaluate the potential benefits of VR in larger cohorts.

XSL•FO
RenderX

Trial Registration: ClinicalTrials.gov NCT05596292; https://www.clinicaltrials.gov/study/NCT05596292
International Registered Report Identifier (IRRID): RR2-10.1186/s13063-023-07786-z

(JMIR Form Res 2025;9:e80729) doi:10.2196/80729

KEYWORDS
heart failure; early mobilization; virtual reality; dyspnea; randomized controlled trial

Introduction

Background

Patients with heart failure frequently experience periods of decompensation that require hospitalization [1]. Early mobilization has received increasing research attention and has benefits such as shorter hospital stays, lower readmission rates, decreased in-hospital mortality, and improved functional outcomes at discharge [2-4]. Despite these benefits, its early implementation can face considerable challenges due to congestion, respiratory discomfort, multiple medications, and especially intense dyspnea [5,6].

Innovative technologies have been explored as tools to reduce discomfort and promote greater adherence to rehabilitation programs. In this context, virtual reality (VR) stands out as a promising resource that can redirect patients' attention to a more pleasant environment [7,8].

In patients with advanced heart failure, a 10-minute VR intervention has been shown to benefit the management of self-reported pain, with significantly better outcomes than guided imagery [9]. In other populations, immersive and nonimmersive VR have effectively improved physical fitness and exercise tolerance [10,11]. However, few studies have included patients with acutely decompensated heart failure (ADHF), and the few available investigations have barely explored the safety and feasibility of these technologies in combination with conventional treatments or their influence on limiting symptoms.

Objectives

This study aimed to assess the effects, feasibility, and safety of an early mobilization protocol combined with immersive VR on dyspnea in patients with ADHF admitted to the intensive care unit (ICU). The hypothesis tested in this randomized clinical trial was that adding immersive VR to an early mobilization protocol would impact the sensation of dyspnea in patients with ADHF compared to the protocol alone and would constitute a feasible and safe tool.

Methods

Study Design

The Early Mobilization Protocol with Immersive Virtual Reality (MOVE) study is a single-center, parallel, superiority randomized clinical trial conducted at Hospital de Clínicas de Porto Alegre (HCPA) from January 2023 to January 2024. The study design and intervention methods have been described previously [12]. This study was registered at ClinicalTrials.gov (NCT05596292), and both the protocol and manuscript were prepared in accordance with the CONSORT (Consolidated Standards of Reporting Trials) checklists (Checklist 1 and Checklist 2) [13].

A pilot study was previously conducted to standardize the mobilization protocol and calibrate the research instruments. Adjustments were made based on the analysis of the pilot data, including adding fields to track participants' reasons for missing sessions and a descriptive section for session notes, including any adverse events.

Participants

This study included patients who had been diagnosed with ADHF and had been admitted to the ICU of HCPA. The diagnosis was confirmed by a medical team independent of the research team. Inclusion criteria required participants to be aged more than 18 years, be alert and coherent, have been hospitalized for at least 24 hours, and have received medical clearance to participate in the protocol. Exclusion criteria included invasive mechanical ventilation or circulatory support, neurodegenerative diseases, pregnancy, communication difficulties, or persistent inability to adapt to the VR headset. Patients with severe hemodynamic instability could only be included after clinical improvement (assessed via clinical and imaging examinations). Patients on noninvasive ventilation could participate in the protocol during breaks in ventilatory therapy.

Intervention and Groups

All eligible patients were registered in an electronic database. Excluded patients, whether for clinical reasons or refusal to participate, had their data collected, including medical record number, reason for exclusion, and date.

After collecting sample characteristics, participants were electronically randomized into the intervention group (IG) or control group (CG). The IG underwent up to 3 consecutive sessions of a progressive and tolerable early mobilization protocol assisted by a physiotherapist combined with immersive VR intervention with headsets compatible with smartphones. A 360° interactive video was displayed during the exercises, enabling patients to explore different angles of the virtual environment. Headphones were used to minimize external sound interference, enhancing the immersive experience. The CG followed the same early mobilization protocol without VR.

The protocol adhered to cardiovascular rehabilitation guidelines, which recommend low-intensity exercise during the acute phase [14]. In total, 3 consecutive sessions were chosen based on evidence suggesting that the most significant changes in dyspnea occur within its first 24 hours, after which changes tend to be less perceptible as treatment progresses [15], potentially compromising the effectiveness of longer protocols.

XSL•FO
RenderX

The protocol included 3 stages of exercise: upper- and lower-limb cycle ergometry, sitting and standing, and ambulation [12]. Exercise duration ranged from 10 to 20 minutes, the progression of which followed individual tolerance (assessed via patient perception and vital signs).

Sessions were conducted in private rooms or isolated hallways to ensure participant privacy. For safety and control, participating patients were identified by a sign on the door of their room indicating their ID number, allocation group, and the start and end dates of the protocol. During the intervention, patients did not participate in any motor activities conducted by the physical therapy team at the hospital.

Outcomes

Primary Outcome

During all sessions, before and after the intervention, the sensation of dyspnea was assessed in both groups using the modified Borg scale [16].

Secondary Outcomes

During all sessions, vital signs were recorded before and after the intervention. Serious adverse events, such as dislodgement of venous access or tubes, falls, or signs of hemodynamic instability, were documented.

Randomization, Allocation, and Blinding

Randomization was conducted at a 1:1 ratio using the REDCap (Research Electronic Data Capture; Vanderbilt University) platform. A member of the research team was responsible for entering patient sample characteristics into the platform, including sociodemographic and clinical data such as New York Heart Association (NYHA) classification, history of hospitalizations due to ADHF, disease etiology, Charlson Comorbidity Index (which estimates 10-year survival), fall risk using the Severo-Almeida-Kuchenbecker scale [17], and functional independence using the Barthel index [18].

These data were used for descriptive analyses and variable control. Patients were randomly allocated to the IG or CG. The physical therapists who carried out the intervention protocol were blinded to the primary outcome, which was collected by independent researchers. To avoid bias, the researchers assessing outcomes withdrew from protocol implementation. Due to the nature of the intervention, patients and physical therapists could not be blinded to group allocation.

Sample Size, Data Management, and Statistical Analysis

Sample size calculation considered a clinically relevant difference of 1 point in the mean dyspnea sensation between the IG and CG using the modified Borg scale. With a statistical power of 80%, a significance level of 5%, and an SD of 1.6 points (as in the literature) [15,19] and adding 20% for potential losses and refusals, the estimated sample size was 66 participants.

Data were stored on the REDCap platform and analyzed on SPSS (version 28.0 for Windows; IBM Corp). Access to the platform was restricted to trained researchers and team members, and data were extracted for analysis at the end of the protocol. Adverse events, exclusions, and other additional data were documented on the platform. Distribution normality was assessed using the Shapiro-Wilk test. Quantitative variables were described as means and SDs or medians and IQRs depending on their distribution. Categorical variables were described as absolute frequencies and percentages.

All analyses were performed following the intention-to-treat principle. In case of missing data, the means or medians of the allocation group were imputed. To compare means between groups for baseline variables, dyspnea sensation, and vital signs, the Student *t* test (2-tailed) for independent samples was applied. For skewed distributions, the Mann-Whitney *U* test was used as a nonparametric alternative. Proportions between groups were compared using the Pearson chi-square test or Fisher exact test, as appropriate.

The analyses related to dyspnea sensation and vital signs considered 2 assessments: before and after the intervention on the first day and before and after the intervention on the last day of the protocol regardless of whether the patient completed all 3 protocol days. The relative Δ was calculated based on the differences between pre- and postintervention values on each day, whereas total variation was determined using the difference between values on the first and last days of follow-up. To analyze the relationship between vasoactive drugs and dyspnea sensation and vital signs, drugs were grouped into inotropes (milrinone and dobutamine) and vasodilators (nitroprusside and nitroglycerin). Statistical significance was considered as $P<.05$.

Ethical Considerations

This study was approved by the Institutional Review Board of HCPA (62209822.7.0000.5327) before initiation. Written informed consent was obtained from all participants. For participants unable to read, the consent form was read aloud by the researcher in the presence of a witness. Participation was voluntary, and no financial compensation was provided. Participants' data were anonymized to protect privacy and analyzed in a blinded manner to maintain confidentiality.

Results

Overview

Of the 58 patients (IG: n=28, 48%; CG: n=30, 52%), 25 (43%) completed all 3 sessions of the protocol (IG: 11/28, 39%; CG: 14/30, 47%). In the first session, after consent and randomization, 67% (4/6) of the participants refused the intervention, and 33% (2/6) were excluded due to clinical instability. In the second session of the protocol, 38% (6/16) were discharged from the unit, and the others failed to continue in the study due to clinical instability (4/16, 25%), refusal to continue (2/16, 12%), or other reasons (4/16, 25%). In the third session, 54% (13/24) of the patients were discharged from the unit, 8% (2/24) refused to participate, 12% (3/24) experienced clinical deterioration, 4% (1/24) had a scheduled procedure and were unavailable for the study protocol, and 21% (5/24) were excluded for other reasons.

"Other reasons" included unit scheduling conflicts, medical team restrictions, and drowsiness. The CONSORT diagram

XSL•FO
RenderX

(Figure 1) illustrates the number of patients who completed the protocol.

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) diagram of participant progression during the conduct of the randomized clinical trial comparing the intervention and control groups.

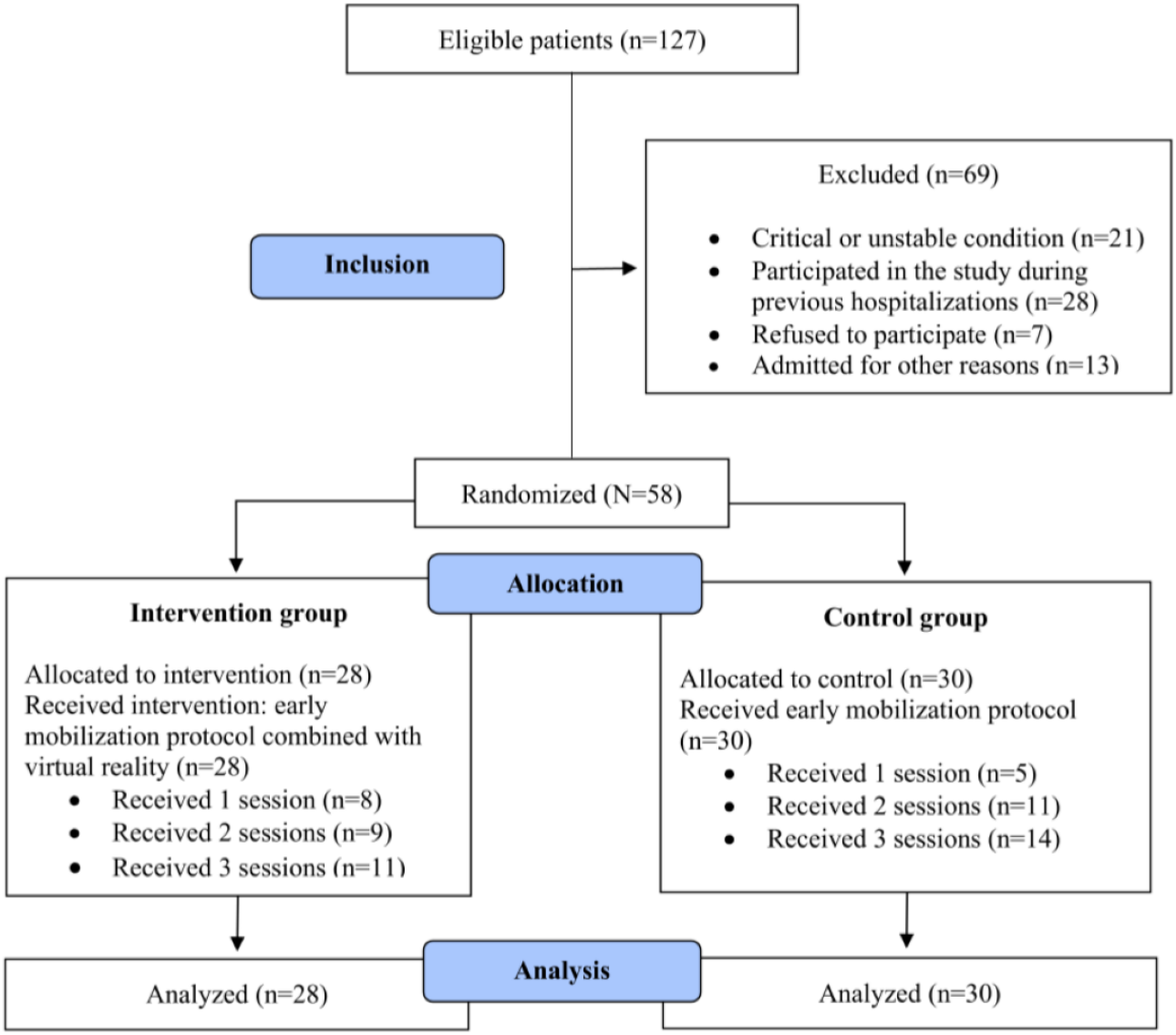

Participant Characteristics

The mean age of the participants was 59 (SD 11.6) years, and most were men (42/58, 72%) and had received their diagnosis over 4 years before. The mean left ventricular ejection fraction was 26.6% (SD 12.5%). The etiology of heart failure was predominantly ischemic, and the functional NYHA class at the study inclusion assessment was mostly III. All characteristics between the groups were similar. Table 1 shows these data.

XSL•FO
RenderX

Table . Participant characteristics (N=58).

Variable	Total	Intervention group (n=28)	Control group (n=30)	*P* value
Age (y), mean (SD)	59 (11.6)	60.7 (10.3)	57.4 (12.7)	.29[a]
Sex, n (%)				.87[b]
Male	42 (72)	20 (71)	22 (73)	
Female	16 (28)	8 (29)	8 (27)	
Weight (kg), mean (SD)	78.3 (22.0)	77.9 (22.8)	78.6 (21.6)	.91[a]
Race, n (%)				.30[c]
Black or of African descent	13 (22)	8 (29)	5 (17)	
Mixed race	4 (7)	3 (11)	1 (3)	
White	41 (71)	17 (61)	24 (80)	
Hospitalizations for HF[d] in the previous year, median (IQR)	1 (0 - 2)	1 (0 - 2)	1 (0 - 3)	.56[e]
NYHA[f] classification, n (%)				.90[c]
I	1 (2)	1 (4)	0 (0)	
II	21 (36)	9 (32)	12 (40)	
III	28 (48)	14 (50)	14 (47)	
IV	8 (14)	4 (14)	4 (13)	
Charlson Comorbidity Index (range 0-33 points), median (IQR)	3 (2-5)	3 (2-5)	3 (2 - 4.25)	.73[e]
HF etiology, n (%)				.99[c]
Ischemic	26 (45)	12 (43)	14 (47)	
Valvular	6 (10)	3 (11)	3 (10)	
Familial	6 (10)	4 (14)	2 (7)	
Cardiomyopathy	5 (9)	2 (7)	3 (10)	
Myocarditis	5 (9)	2 (7)	3 (10)	
Unknown or idiopathic	5 (9)	3 (11)	2 (7)	
Alcoholic	2 (3)	1 (4)	1 (3)	
Hypertensive	2 (3)	1 (4)	1 (3)	
Cardiotoxicity	1 (2)	0 (0)	1 (3)	
Ejection fraction (%), mean (SD)	26.6 (12.5)	26.1 (11.3)	27.1 (13.7)	.78[a]
SAK[g] scale, n (%)				.55[b]
Low risk	29 (50)	13 (46)	16 (53)	
Moderate risk	16 (28)	7 (25)	9 (30)	
High risk	13 (22)	8 (29)	5 (17)	
Barthel index, n (%)				.26[c]
Little or no disability	19 (33)	9 (32)	10 (33)	
Moderate disability	31 (53)	13 (46)	18 (60)	
Severe disability	8 (14)	6 (21)	2 (7)	
Vasoactive drugs, n (%)				.19[b]
No	18 (31)	11 (39)	7 (23)	

XSL•FO
RenderX

Variable	Total	Intervention group (n=28)	Control group (n=30)	*P* value
Yes	40 (69)	17 (61)	26 (87)	
Nitroprusside	28 (48)	11 (39)	17 (57)	.19[b]
Milrinone	18 (31)	8 (29)	10 (33)	.70[b]
Nitroglycerin	6 (10)	4 (14)	2 (7)	.42[c]
Dobutamine	3 (5)	1 (4)	2 (7)	>.99[c]

[a]Student *t* test.
[b]Pearson chi-square test.
[c]Fisher exact test.
[d]HF: heart failure.
[e]Mann-Whitney *U* test.
[f]NYHA: New York Heart Association.
[g]SAK: Severo-Almeida-Kuchenbecker.

Dyspnea Sensation

Table 2 shows the variation in dyspnea sensation, expressed using medians and IQRs, between the pre- and postintervention assessments during the first and last sessions of the early mobilization protocol. In the IG, the mean change (Δ) in dyspnea score was lower in the last session than in the first session (0.75, SD 1.91 vs 0.57, SD 1.69). In contrast, in the CG, the variation in dyspnea sensation was similar in both sessions. No statistically significant difference was observed between the groups (difference between groups across the before and after intervention moments on the first day, *P*=.97; difference between groups across the before and after intervention moments on the third day, *P*=.58). The analysis of the difference between the changes (Δ of the last session – Δ of the first session) showed less change in the CG than in the IG, without statistical significance (*P*=.67).

Table . Variation in dyspnea score between the first and last sessions of the early mobilization protocol in the intervention and control groups (N=58; score range of dyspnea 0-10)[a].

	Intervention group (n=28)		Control group (n=30)		*P* value
	Variation in dyspnea score, median (IQR)	Δ[b], mean (SD)	Variation in dyspnea score, median (IQR)	Δ[b], mean (SD)	
Presession 1	2 (1 - 3.75)	0.75 (1.91)	2 (1 - 5.25)	0.76 (1.55)	.97[c]
Postsession 1	3 (2-5)	0.75 (1.91)	3.5 (2-7)	0.76 (1.55)	.97[c]
Presession 3	2 (2-4)	0.57 (1.69)	2 (1.75 - 4)	0.78 (1.13)	.58[c]
Postsession 3	3 (3-4)	0.57 (1.69)	3.50 (2-5)	0.78 (1.13)	.58[c]

[a]Δ1 – Δ3 (post–pre between sessions 1 and 3): mean −0.17 (SD 1.68; intervention group) and mean 0.01 (SD 1.73; control group); *P*=.67 (Student *t* test).
[b]Postsession – presession values.
[c]Student *t* test.

Vital Signs

Changes in vital signs, expressed as means and SDs, between pre- and postintervention assessments in the first and last sessions remained within safe ranges for protocol application (Table 3). No differences were observed between the groups.

XSL•FO
RenderX

Table . Variation in vital signs between the first and last sessions of the early mobilization protocol in the intervention and control groups (N=58).[a]

Variable	Intervention group (n=28), mean (SD)		Control group (n=30), mean (SD)		*P* value
	Before the intervention	After the intervention	Before the intervention	After the intervention	
First session					
SBP[b] (mm Hg)	101.3 (17.6)	100.6 (18.1)	101.6 (17.6)	107.0 (18.5)	.01
DBP[c] (mm Hg)	65.5 (11.0)	66.2 (11.6)	65.3 (13.3)	68.8 (12.5)	.22
HR[d] (bpm)	94.6 (20.5)	95.4 (21.5)	88.9 (15.6)	92.0 (15.7)	.25
RR[e] (bpm)	18.5 (3.1)	20.3 (4.5)	20.5 (4.4)	20.5 (4.7)	.19
SpO_2[f] (%)	96.5 (1.9)	96.8 (1.8)	96.1 (2.4)	96.5 (2.7)	.92
Third session					
SBP (mm Hg)	100.9 (14.3)	105.8 (17.3)	98.7 (11.1)	103.0 (9.9)	.81
DBP (mm Hg)	64.7 (8.8)	67.6 (9.2)	61.7 (6.4)	65.4 (10.0)	.76
HR (bpm)	89.0 (14.2)	90.4 (12.8)	93.7 (14.1)	96.3 (15.3)	.41
RR (bpm)	18.9 (2.9)	20.3 (3.9)	19.0 (3.0)	20.8 (3.8)	.63
SpO_2 (%)	96.7 (1.4)	97.2 (1.0)	95.7 (2.0)	96.6 (1.4)	.25

[a]Comparisons between the intervention group and control group were conducted using the mean change (after the intervention – before the intervention) and analyzed using the Student *t* test.

[b]SBP: systolic blood pressure.

[c]DBP: diastolic blood pressure.

[d]HR: heart rate.

[e]RR: respiratory rate.

[f]SpO_2: peripheral oxygen saturation.

The analyses of the last day of the protocol showed a moderate correlation among systolic blood pressure (r=0.416; P=.03), diastolic blood pressure (r=0.393; P=.04), respiratory rate (r=0.434; P=.02), and dyspnea sensation in the IG. In the CG, on the first day of the protocol, a moderate correlation was observed between diastolic blood pressure and dyspnea (r=−0.472; P=.008). No serious adverse events occurred during this study.

Stratification by Type of Vasoactive Drug and NYHA Classification

This study grouped vasoactive drugs into 2 categories: inotropes (milrinone and dobutamine) and vasodilators (nitroprusside and nitroglycerin). The intragroup analysis of dyspnea sensation between the first and last sessions of the protocol found no statistically significant differences for any category (P>.05). Regarding the functional NYHA class of the patients, neither dyspnea sensation nor vital signs changed in both groups between the first and last sessions (P>.05).

Time Between Hospital Admission and Protocol Start

The median time between hospital admission and the start of the early mobilization protocol was similar between the groups: 4 (IQR 2‑7.7) days in the IG and 4 (IQR 2‑7) days in the CG. In the ICU, patients in the IG took approximately a median of 1.5 (IQR 1‑3.7) days to start the protocol, whereas those in the CG took a median of 2 (IQR 1‑2) days.

Discussion

Principal Findings

In this study, an early mobilization protocol with the addition of passive immersive VR failed to significantly change the dyspnea perception of patients with ADHF, although this research considered it feasible and safe. However, the comparison of variations in dyspnea sensation between the first and last day of the protocol suggested that the IG had a smaller change in the perception of this symptom on the last day, whereas the CG maintained similar values over time.

It is also important to highlight that dyspnea, in addition to being related to patients' clinical condition and pulmonary edema [20], constitutes a subjective symptom that is influenced by previous experiences and patients' self-management strategies. Those with heart failure, for example, often experience this symptom continuously, which may reduce their change perception of it over time [21]. Furthermore, although various tools can assess dyspnea, the subjectivity of the symptom makes it difficult to accurately identify significant relief, which can compromise the evaluation of therapeutic interventions [22].

VR was applied as an accessible, low-cost, and easy-to-use tool delivered through passive 360° video. Although head movements influenced the viewing angle, there was no active therapeutic goal within the virtual environment. This passive use of VR created a calm and enjoyable setting and has previously been shown to reduce anxiety in other populations,

XSL•FO
RenderX

acting as a distraction that can enhance patient cooperation and overall experience, with a low incidence of adverse events such as nausea, vomiting, or dizziness [23]. Both passive and active VR applications have also been shown to be beneficial in pain management [23] and motor rehabilitation [24]. Future studies could explore optimizing 360° video content toward more tranquil and restorative environments, as well as integrating active VR tasks designed to support adherence to clinical interventions. Incorporating real-time feedback on breathing patterns or exercise performance within VR environments could further enhance patient engagement and therapeutic benefit.

Vital signs remained within safe limits, and no serious adverse events occurred in either group, reinforcing the safety of the protocol. Previous studies also indicate that early mobilization protocols are safe for this population, without causing serious adverse events or hemodynamic instability [25]. For the patients who completed the early mobilization protocol, its implementation remained feasible at all stages of treatment. However, data analysis suggests that patients' short stays in the ICU limited the number of times that the protocol could have been applied. The implementation of the protocol was affected by factors such as the prolonged time that patients spent in the emergency unit waiting for a bed and the team's uncertainty regarding the initiation of early mobilization. In addition to the cultural barriers commonly faced in early mobilization programs (such as confusion regarding roles, responsibilities, and team attitude), there are patient-related barriers, including limiting symptoms and hemodynamic instability, as well as structural obstacles such as insufficient physical resources and a lack of qualified professionals [26].

These results highlight the feasibility and safety of early mobilization in critical settings and suggest that immersive VR may offer a promising tool to improve engagement and adherence even without objective changes in symptom perception. In addition, future studies are needed to explore the potential of VR in contexts with longer hospital stays or in protocols adapted to patients' individual preferences.

Limitations

Less than half of the participants (25/58, 43%) completed all 3 days of the protocol, which limits the generalization of the results. This is partly related to the short stay of patients in the ICU as many experience clinical stabilization while in the emergency department and are then transferred or discharged. Additionally, delays in participant inclusion due to the initial uncertainty of the team also contributed to this outcome. Although the calculated sample size was 27 participants per group, the IG only had 11 participants by the end of the study who had completed the 3 sessions of the protocol. Although all patients who completed at least 1 session of the protocol were included in the analysis, variability in the number of sessions completed and attrition across both groups reduced the effective sample size and, consequently, the statistical power of this study and its ability to detect small to moderate effects. This limitation should be carefully considered in future research. We also highlight the difficulty in recruiting patients from this population, which is characterized by highly individualized clinical conditions and complex management. Thus, while the findings support feasibility and safety, they should be interpreted with caution and are not yet generalizable to all patients with ADHF. Strategies such as recruitment at different stages of hospitalization or across multiple units could extend follow-up and improve the reproducibility of the findings.

Conclusions

Adding immersive VR to an early mobilization protocol had no significant effect on dyspnea in patients with ADHF. However, the intervention was considered safe and feasible, with no serious adverse events reported in either group.

VR is an accessible and low-cost tool that may have potential applications in outpatient and home settings. Exploring its active use, beyond serving only as a distraction, may provide new insights into its impact on the rehabilitation process in this population. Through facilitating the implementation of early mobilization, VR could contribute to improving clinical outcomes and reducing health care costs. Furthermore, studying the integration of VR into routine care has the potential to optimize rehabilitation guidelines and promote more patient-centered treatment. Future feasibility studies should explore alternative multidisciplinary strategies to enhance adherence and improve the reproducibility of the findings, supporting the translation of this approach into routine clinical practice.

Funding

This study was partially funded by the Coordination for the Improvement of Higher Education Personnel, Brazil (funding code 001). This study was also supported by the National Council for Scientific and Technological Development and the Research Incentive Fund of the Hospital de Clínicas de Porto Alegre. The funders were not involved in the study design; data collection, analysis, or interpretation; or manuscript writing.

Data Availability

All data generated or analyzed during this study are included in this published article (and its supplementary information files).

Authors' Contributions

The study project was mainly developed by IBF, LGC, PDL, and ERR-S. The protocol and data collection were mainly conducted by IBF, LGC, CEMT, JdSP, GP, and VAG. The writing of the results was mainly conducted by IBF, PDL, and ERR-S. The early mobilization protocol was conducted by IBF and CEMT, and the data collection was mainly conducted by IBF, LGC, CEMT, JdSP, GP, and VAG. All authors contributed to the design of the clinical study and critical revision of the manuscript.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Checklist 1
CONSORT checklist.
[PDF File, 239 KB - formative_v9i1e80729_app1.pdf]

Checklist 2
CONSORT-EHEALTH checklist.
[PDF File, 1189 KB - formative_v9i1e80729_app2.pdf]

References

1. Heidenreich PA, Bozkurt B, Aguilar D, et al. 2022 AHA/ACC/HFSA guideline for the management of heart failure: a report of the American College of Cardiology/American Heart Association joint committee on clinical practice guidelines. Circulation 2022 May 3;145(18):e895-e1032. [doi: 10.1161/CIR.0000000000001063] [Medline: 35363499]
2. Takada S, Kondo T, Yasunaga M, et al. Early rehabilitation in older patients hospitalized with acute decompensated heart failure: a retrospective cohort study. Am Heart J 2020 Dec;230:44-53. [doi: 10.1016/j.ahj.2020.09.009] [Medline: 32956621]
3. Kaneko H, Itoh H, Kamiya K, et al. Acute-phase initiation of cardiac rehabilitation and clinical outcomes in hospitalized patients for acute heart failure. Int J Cardiol 2021 Oct 1;340:36-41. [doi: 10.1016/j.ijcard.2021.08.041] [Medline: 34454966]
4. Ishibashi T, Kaneko H, Ueno K, et al. Association between early initiation of cardiac rehabilitation and short-term outcomes of patients with acute heart failure admitted to the intensive care unit. Am J Cardiol 2023 Nov 1;206:285-291. [doi: 10.1016/j.amjcard.2023.07.070] [Medline: 37717477]
5. Rohde LEP, Montera MW, Bocchi EA, et al. Diretriz Brasileira de Insuficiência Cardíaca Crônica e Aguda. ABC 2018;111(3):436-539. [doi: 10.5935/abc.20180190]
6. Anderson KM, Murphy D, Groninger H, Kolm P, Wang H, Barton-Maxwel V. Perceived symptoms as the primary indicators for 30-day heart failure readmission. PLoS ONE 2022;17(5):e0267820. [doi: 10.1371/journal.pone.0267820] [Medline: 35511916]
7. Raposo AB, Moliterno AH, Silva J, Fabri RV, Freire A, Pacagnelli FL. Comparação da resposta hemodinâmica entre terapia convencional e realidade virtual em pacientes com insuficiência cardíaca internados na unidade de emergência. Fisioter Pesqui 2022;29(1):61-67. [doi: 10.1590/1809-2950/21008729012022pt]
8. Micheluzzi V, Casu G, Sanna GD, et al. Improving adherence to rehabilitation for heart failure patients through immersive VIRTUAL reality (VIRTUAL-HF): a protocol for a randomized controlled trial. Contemp Clin Trials 2024 Mar;138:107463. [doi: 10.1016/j.cct.2024.107463] [Medline: 38302011]
9. Groninger H, Stewart D, Fisher JM, Tefera E, Cowgill J, Mete M. Virtual reality for pain management in advanced heart failure: a randomized controlled study. Palliat Med 2021 Dec;35(10):2008-2016. [doi: 10.1177/02692163211041273] [Medline: 34431398]
10. Rutkowski S, Rutkowska A, Kiper P, et al. Virtual reality rehabilitation in patients with chronic obstructive pulmonary disease: a randomized controlled trial. Int J Chron Obstruct Pulmon Dis 2020;15:117-124. [doi: 10.2147/COPD.S223592] [Medline: 32021150]
11. Rutkowski S, Szary P, Sacha J, Casaburi R. Immersive virtual reality influences physiologic responses to submaximal exercise: a randomized, crossover trial. Front Physiol 2021;12:702266. [doi: 10.3389/fphys.2021.702266] [Medline: 34658904]
12. Fraga IB, Caballero LG, Lago PD, et al. Perceived dyspnea and experience of hospitalized patients with acute decompensated heart failure undergoing an early MObilization protocol with immersive Virtual rEality: MOVE study protocol for a parallel superiority randomized clinical trial. Trials 2023 Nov 24;24(1):751. [doi: 10.1186/s13063-023-07786-z] [Medline: 38001540]
13. Moher D, Hopewell S, Schulz KF, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. BMJ 2010 Mar 23;340:c869. [doi: 10.1136/bmj.c869] [Medline: 20332511]
14. Carvalho TD, Milani M, Ferraz AS, et al. Diretriz Brasileira de Reabilitação Cardiovascular – 2020. Arq Bras Cardiol 2020 May 22;114(5):943-987. [doi: 10.36660/abc.20200407]
15. Nomura F, Kurobe N, Mori Y, et al. Multicenter prospective investigation on efficacy and safety of carperitide as a first-line drug for acute heart failure syndrome with preserved blood pressure COMPASS: carperitide effects observed through monitoring dyspnea in acute decompensated heart failure study. Circ J 2008 Nov;72(11):1777-1786. [doi: 10.1253/circj.cj-07-0760] [Medline: 18832779]
16. Williams N. The Borg Rating of Perceived Exertion (RPE) scale. Occup Med (Chic Ill) 2017 Jul;67(5):404-405. [doi: 10.1093/occmed/kqx063]
17. Severo IM, Kuchenbecker RDS, Vieira DFVB, Lucena ADF, Almeida MDA. Risk factors for fall occurrence in hospitalized adult patients: a case-control study. Rev Lat Am Enfermagem 2018 Aug 9;26:e3016. [doi: 10.1590/1518-8345.2460.3016] [Medline: 30110094]

XSL•FO
RenderX

18. Mahoney FI, Barthel DW. Functional evaluation: the Barthel Index. Md State Med J 1965 Feb;14:61-65 [FREE Full text] [Medline: 14258950]
19. Papathanasiou JV, Petrov I, Tokmakova MP, et al. Group-based cardiac rehabilitation interventions. A challenge for physical and rehabilitation medicine physicians: a randomized controlled trial. Eur J Phys Rehabil Med 2020 Aug;56(4):479-488. [doi: 10.23736/S1973-9087.20.06013-X] [Medline: 31976639]
20. Marteles MS, Urrutia A. Formas de presentación de la insuficiencia cardíaca aguda: edema agudo de pulmón y shock cardiogénico. Medicina Clínica 2014 Mar;142(1):14-19. [doi: 10.1016/S0025-7753(14)70077-6]
21. Faulkner KM, Jurgens CY, Denfeld QE, Lyons KS, Harman Thompson J, Lee CS. Identifying unique profiles of perceived dyspnea burden in heart failure. Heart & Lung 2020 Sep;49(5):488-494. [doi: 10.1016/j.hrtlng.2020.03.026]
22. Kociol RD, McNulty SE, Hernandez AF, et al. Markers of decongestion, dyspnea relief, and clinical outcomes among patients hospitalized with acute heart failure. Circ Heart Fail 2013 Mar;6(2):240-245. [doi: 10.1161/CIRCHEARTFAILURE.112.969246] [Medline: 23250981]
23. Smith V, Warty RR, Sursas JA, et al. The effectiveness of virtual reality in managing acute pain and anxiety for medical inpatients: systematic review. J Med Internet Res 2020 Nov 2;22(11):e17980. [doi: 10.2196/17980] [Medline: 33136055]
24. He Y, Yang Q, Dai X, et al. Effects of virtual reality technology on early mobility in critically ill adult patients: a systematic review and meta-analysis. Front Neurol 2024;15:1469079. [doi: 10.3389/fneur.2024.1469079] [Medline: 39975851]
25. Kakutani N, Fukushima A, Kinugawa S, et al. Progressive mobilization program for patients with acute heart failure reduces hospital stay and improves clinical outcome. Circ Rep 2019 Feb 19;1(3):123-130. [doi: 10.1253/circrep.CR-19-0004] [Medline: 33693126]
26. Dubb R, Nydahl P, Hermes C, et al. Barriers and strategies for early mobilization of patients in intensive care units. Ann Am Thorac Soc 2016 May;13(5):724-730. [doi: 10.1513/AnnalsATS.201509-586CME] [Medline: 27144796]

Abbreviations

ADHF: acutely decompensated heart failure
CG: control group
CONSORT: Consolidated Standards of Reporting Trials
CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
HCPA: Hospital de Clínicas de Porto Alegre
ICU: intensive care unit
IG: intervention group
MOVE: Early Mobilization Protocol with Immersive Virtual Reality
NYHA: New York Heart Association
REDCap: Research Electronic Data Capture
VR: virtual reality

Edited by A Stone, A Mavragani; submitted 15.07.25; peer-reviewed by M Hellaby, OPD Almeida Neto; revised version received 16.09.25; accepted 17.09.25; published 31.10.25.

Please cite as:
Fraga IB, Caballero LG, Tremea CEM, Prates JDS, Perin G, Guedes VA, Lago PD, Rabelo-Silva ER
Effects, Feasibility, and Safety of an Early Mobilization Protocol With Immersive Virtual Reality for Dyspnea in Patients With Acutely Decompensated Heart Failure: The MOVE Randomized Clinical Trial
JMIR Form Res 2025;9:e80729
URL: https://formative.jmir.org/2025/1/e80729
doi:10.2196/80729

© Iasmin Borges Fraga, Larissa Gussatschenko Caballero, Carlos Eduardo Maciel Tremea, Janaína dos Santos Prates, Gabrielle Perin, Vitor Alves Guedes, Pedro Dal Lago, Eneida Rejane Rabelo-Silva. Originally published in JMIR Formative Research (https://formative.jmir.org), 31.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Behavior Emotion Therapy System and You: Co-Design and Evaluation of a Mental Health Chatbot and Digital Human for Mild to Moderate Anxiety in Healthy Participants

Almira Osmanovic Thunström[1], MMSc; Lilas Ali[2], PhD; Hanne Krage Carlsen[3], MPH, PhD; Maria Bohm, BCS; Linda Wesén[2], MMSc; Olof Wrede[4], PhD; Iris Sarajlic Vukovic[2], MD; Andreas Hellström[5], PhD; Tomas Larson[2], PhD; Steinn Steingrimsson[2], PhD, MD

[1]Institute of Neuroscience and Physiology, Gothenburg University, Journalvägen 5, Gothenburg, Sweden

[2]Psychiatric Department, Sahlgrenska University Hospital, Region Västra Götaland, Gothenburg, Sweden

[3]Centre of Registers Västra Götaland, Gothenburg, Sweden

[4]Crisis and Trauma Center, Regionshälsan, Region Västra Götaland, Gothenburg, Sweden

[5]Department of Technology Management and Economics, Chalmers University of Technology, Gothenburg, Sweden

Corresponding Author:
Almira Osmanovic Thunström, MMSc
Institute of Neuroscience and Physiology, Gothenburg University, Journalvägen 5, Gothenburg, Sweden

Abstract

Background: Co-design is a collaborative approach involving end users, stakeholders, and designers in creating digital tools for health care. This study focuses on the co-design and evaluation of Behavior, Emotion, Therapy System, and You (BETSY), a mental health chatbot and digital human for mild to moderate anxiety.

Objective: This study aims to develop and evaluate BETSY through a co-design process involving potential users.

Methods: The study used a mixed-methods approach across 3 phases. Phase 1 involved recruiting 87 volunteer participants through social media for initial end-user requirements. Phase 2 focused on the design process based on end-user requirements using the expertise of 10 stakeholders in health care and health service user. Phase 3 evaluated the 2 prototyped interfaces (a text-only chatbot and a voice-activated digital human chatbot) with 45 healthy volunteers.

Results: For phase 1, 61% (n=51) of participants had previous experience with chatbots and 86% expressed a willingness or potential willingness to use a chatbot for mental health support. Thematic analysis in phase 2 revealed key user preferences for a personalized, nonjudgmental chatbot with a clear identity and a focus on empathy. Privacy concerns and the need for clear interaction guidelines were highlighted. In phase 3, the users showed a strong preference for discussing anxiety related to work, relationships, and health. Both text-only and voice-activated digital human users considered BETSY valuable for managing mild to moderate anxiety and providing self-help exercises.

Conclusions: The co-design process yielded valuable insights for developing BETSY. While users recognized its potential as an accessible first step before seeking professional help, they also identified areas for improvement, including more nuanced conversation capabilities and a broader range of exercises. BETSY's potential as a screening tool for health care was consistently acknowledged, suggesting its capability for alleviation of health care system burdens.

(JMIR Form Res 2025;9:e66163) doi:10.2196/66163

KEYWORDS

co-design; chatbot; mental health; digital human; prototyping; avatar; behavior emotion therapy system and you; BETSY; chatbots; conversational agent; digital human; anxiety; mixed-method; mixed-methods; social media; mental health support; self-help; healthcare; digital health

Introduction

Co-design is a collaborative design approach that involves end users, stakeholders, and designers in the design process to create more effective and engaging digital tools. Co-design is an important aspect of creating apps and digital services in health care, as it ensures better design, overall outcome, and sustainability in projects aimed at reaching targeted audiences [1-5]. By involving end users in the design process, one is more likely to create design outcomes that create higher efficacy among its users, especially those who would benefit from continuous and stringent use of a digital tool, such as self-monitoring or digital therapy [1-3,5]. For instance, a study on codeveloping a mental health and well-being chatbot with

XSL•FO
RenderX

and for young people demonstrated the benefits of involving end users in the design process [6]. The research used interviews and surveys to inform the creation of the chatbot's personality and conversation design, resulting in a more effective and engaging tool for young adults. Another study of a co-design process for a chatbot service for mental health aimed at individuals living in rural areas showed that involving end users in the design process gave valuable insight into potential design, need, and architecture for a new mental health service [2,6]. The co-design process can be tailored according to the users, considering the diversity within the group, digital technology, and the setting. There are also potential challenges, particularly related to finding resources [3].

One of the core problems in design is ensuring representation. Lack of representation can lead to lower user efficacy and problems with an inclusive user interface. Co-design, an approach that involves the inclusion of potential users in the design process, can mitigate some of the issues of noninclusionary design and is a promising method for creating digital mental health tools that are more effective, engaging, and inclusive [4,5].

There are also matters of safety and trustworthiness. Co-design can be used to increase transparency as well as address concerns and potential hazards of the system before it is even developed [4]. A study by Hamlin et al [7] showed that skepticism toward digitalization decreased significantly when users were familiar with and felt in control of the safety and security of the digital health service provided to them.

By including experts who are also users of mental health services, it is possible to capture the perspectives of individuals who are vulnerable within the system [2,4,6]. However, there is still a standing stigma of involving patients and the general public as co-designers of mental health services [2], which has resulted in a lack of literature and results regarding co-design processes in chatbot-driven technologies and, in particular, in the design of digital humans. In a previous study [8], we detailed the history, usability, and needs for a mental health chatbot as well as providing a detailed description of the quantitative and biometric data used to evaluate a mental health chatbot and digital human interfaces for conversations on mental health. The overarching goal of this paper is to detail: (1) the co-design process of a mental health chatbot for mild to moderate anxiety and depression; (2) the design choices in the prototype of the chatbot and digital human system; and (3) qualitatively evaluate to what extent we had accurately captured the needs of the end users based on the co-design process in 2 different interfaces (text-based and voice-activated digital human). This was performed through 3 separate phases: inquiry, design, and evaluation.

Methods

Overview

This project comprised 3 phases. Phase 1 consisted of mapping end-user requirements, phase 2 focused on chatbot development based on end-user requirements, and phase 3 evaluated a chatbot prototype consisting of a chatbot prototype with 2 interface options (Figure 1). Throughout the project, healthy volunteers were engaged to ensure adherence to the regulatory and ethical standards governing the development of novel early-stage medical technologies.

Figure 1. Phases of co-design development and evaluation.

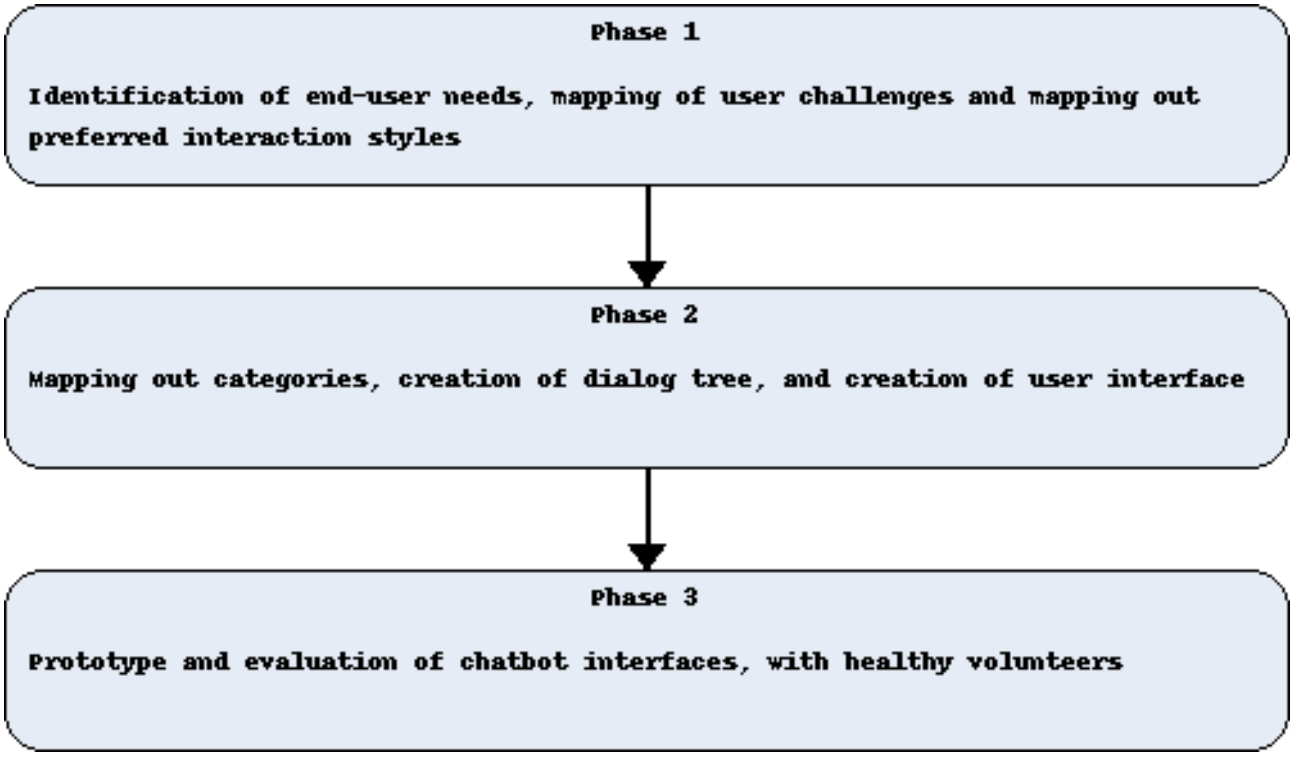

XSL•FO
RenderX

Phase 1

Data Collection and Design Framework for End User Needs

This phase of the project focused on exploring user-centered requirements for developing a text-based and digital human chatbot to alleviate symptoms of mild to moderate anxiety. The primary objective was to map the key topics, preferences, and concerns held by anonymous voluntary participants from the general public regarding a chatbot for mild to moderate anxiety. Ten experts were invited to participate in the construction of the foundation of the survey. The project team consisted of a lead engineer, 2 psychiatrists, 2 psychiatric nurses, 4 clinical psychologists, and 1 patient representative. There was an even representation of sex and age among the participants. Each member of the team was carefully selected based on their expertise and knowledge of digitalization and psychiatry. The patient representative was neither a former nor a current patient of any of the experts and was not under their care. The patient was seen as a valued and independent expert in the project. No monetary compensation was involved in participation by any of the experts and was voluntary for all parties. Together, the participants designed a survey to gather qualitative and quantitative data from voluntary participants regarding expectations for a chatbot designed to manage anxiety and their overall opinions on the use of technology for mental health interventions.

Survey Construction and Design

To gather both qualitative and quantitative data on user-centered requirements for developing a chatbot aimed to manage mild to moderate anxiety, a mixed-method survey was designed. Drawing on the efficiency, effectiveness, and satisfaction domains outlined by Radzwill and Benton [9], the survey explored participants' perceptions and preferences regarding mental health chatbots. It included 8 multiple-choice and 3 open-ended questions, covering demographics, design expectations, functionality suggestions, and overall attitudes toward chatbot use in anxiety management.

The multiple-choice section addressed gender, education, and occupation, as well as preferences for interaction mode (text or digital human), chatbot appearance (human-like or avatar), content topics, level of personalization, and personality traits (eg, friendly vs neutral, personal vs clinical). The open-ended questions invited participants to reflect on their general attitudes toward mental health chatbots, perceived benefits and challenges in using them for anxiety support, and their views on chatbots as a complement or alternative to traditional care.

The open-ended questions were:

- What do you wish an anonymous conversation with a mental health chatbot would give you?
- What would make you hesitate to talk to a chatbot about mental health?
- Can you give an example of the worst and best chatbot you've ever talked to?

Collection of Survey Data and Analysis

The survey was distributed through social media networks such as Facebook (Meta), LinkedIn (Microsoft Corporation), Instagram (Meta), Twitter (X Corp), and Mastodon (Mastodon gGmbH) from April to June 2020. The collected data were analyzed using descriptive statistics and thematic analysis to identify patterns and insights related to the participants' attitudes toward chatbots for mental health, particularly those tailored for anxiety management [9]. We implemented measures to safeguard the privacy and anonymity of participants. Data were collected through an anonymous, encrypted survey hosted on the free-to-use, noncommercial, platform PsyToolKit (owned by Professor Gijsbert Stoet). No personal identifiers or IP addresses were recorded before entering the survey, and a written statement was shown providing clear and concise information to participants about the study's purpose, scope, and potential risks. Participants were given the opportunity to review and consent to participate before proceeding with the survey.

Phase 2

Creating a Framework for Content and Interface

The second phase of the project aimed at translating the extracted themes and requirements into a preliminary structure for a dialog tree and avatar personality. This involved iterative and collaborative explorations of the chatbot's backend, frontend, and aesthetics, including the design of the anthropomorphic features of the avatar. From July 2020 to December 2021, 4 workshops were conducted and iterative design work on the chatbot was carried out. These workshops consisted of invited experts listed previously and focused on exploring various aspects of chatbot technology, particularly interface and content design, which were guided by the insights gained from the phase 1 survey.

The expert participants were active in constructing the frame and content of the chatbot, not as testers of the prototype. To outline the chatbot functions, the experts were provided with Post-it Notes and instructed to adopt the perspective of a patient. They were then asked to formulate questions regarding their overall mental health, specifically focusing on mild to moderate anxiety and depression, with an emphasis on topics that received higher prioritization from the general public (Table 1). These questions were subsequently posted on a wall in a large conference room. Following this, the same 10 participants assumed the role of supportive mental health peers (not therapists) and responded to each other's Post-it Notes. The collected questions and responses were used as a foundation for training the language model and identifying keywords for the rules-based chatbot in Google's chatbot platform: DialogFlow.

XSL•FO
RenderX

Table . Phase 1: respondent demographics and responses to questions (N=87).

Demographic question and response	Respondents (N=87), n (%)
Sex	
Male	29 (33)
Female	54 (62)
Other	2 (2)
Rather not say	2 (2)
Age (years)	
18 - 25	3 (3)
26 - 35	15 (17)
36 - 45	37 (43)
46 - 55	26 (30)
56+	5 (6)
Do not want to answer	1 (1)
Have you ever talked to a chatbot?	
Yes	51 (61)
No	27 (32)
I am not sure	6 (7)
Would you talk to a chatbot anonymously about your mental health?	
Yes	31 (37)
Maybe	41 (49)
No	12 (14)
If you would talk to a chatbot anonymously about your mental health, would you want it to have a name?	
Yes	46 (56)
Maybe	19 (23)
No	17 (21)
If the chatbot had a name, would you want it to have a distinct identity (eg, calling itself a specific gender identity and with a human name?)	
Yes, I want a distinct identity	16 (20)
I want it to be neutral, no specific identity	29 (36)
I would not care	28 (34)
I don't know	8 (10)
If you talked to a chatbot about your mental health, would you want it to be formal or informal when it addresses you?	
It should be completely informal like a friend	15 (18)
It should be somewhat formal like an acquaintance	47 (58)
It should be formal	19 (24)
Which topic from this list would you like to include in the chatbot's manuscript?	
Work	59 (68)
Stress	58 (67)

XSL•FO
RenderX

Demographic question and response	Respondents (N=87), n (%)
Relationships	56 (64)
Disease	45 (52)
Conflicts	40 (46)
Family	38 (44)
Partner	37 (42)
Friendships	33 (38)
Children	32 (37)
Death	32 (37)
Financial issues	31 (36)
Abuse	29 (33)
Medicine	29 (33)
Exercise	29 (33)
Food	26 (30)
Addiction	23 (33)
Legal issues	21 (24)
Sex	20 (23)
Sexuality	20 (23)
Culture (movies, books, theater)	20 (23)
Love	18 (21)
Philosophy	18 (21)
News	17 (19)
Politics	15 (17)
Art	12 (14)
Weather	12 (14)
Religion	10 (11)
Fashion	7 (8)
Celebrities	6 (7)

To create the chatbot's facial expressions and appearance, a standard model of face design from the New Zealand-based company UNEEQ (UneeQ Limited) was used. Since the survey participants from the general public did not express any specific preferences regarding gender, age, or other appearance features, the avatar was modeled after a nurse and given a female-presenting persona due to the statistical probability of meeting a female-presenting nurse in psychiatry in Sweden. To compare the effect of anthropomorphic features in a controlled manner, a text-only version of the algorithm was created using Itsalive.io (ItsAlive SARL) and published on a project-dedicated Facebook page. The model was named Behavior, Emotion, Therapy System, and You (BETSY).

Phase 3

Evaluation of Chatbot Design

The final stage of the project aimed to assess the effectiveness of 2 prototypes: one based on text chat and the other based on interaction with a digital human avatar using only voice. As the comprehensive description of the procedures, scales, and questions used in the evaluation has been published previously by our group [8], the current paper will provide a summary of the overall methodology. Only the open-ended questions will be presented in this paper. In this manuscript, we focus on the qualitative evaluation of BETSY conducted in Phase 3, using open-ended questions analyzed through thematic analysis. An inductive, semantic coding approach was used to identify key patterns in user feedback, with themes interpreted in light of the design goals and user requirements established in phase 1. While this paper emphasizes qualitative findings, the companion publication [8] reports the full quantitative evaluation, including metrics such as the System Usability Scale (SUS) and biometric engagement indicators. To strengthen clarity and cross-reference between findings, we have added citations and brief summaries of those quantitative measures in the present manuscript, ensuring a more complete view of BETSY's overall performance.

To ensure the inclusion of the target users, healthy individuals, advertisements were placed on major social media platforms to recruit volunteers for the evaluation of BETSY. For a more

XSL•FO
RenderX

detailed description of the evaluation protocol, the reader is directed to our previous publication [8].

Data Analysis

The study comprises a mixed methodology. For qualitative analyses, thematic analyses were explored through open-ended survey questions in phases 1 and 3 of this study. Thematic analysis is a widely used method for identifying, analyzing, and reporting patterns within qualitative data. This approach was chosen for its flexibility and ability to provide rich, detailed insights into the participants' experiences and perspectives. An inductive approach was adopted, allowing themes to emerge from the data rather than fitting the data into a pre-existing coding frame. In addition, a semantic approach was used, focusing on the explicit content of the responses rather than underlying assumptions [10]. This methodological approach allowed for a systematic and rigorous exploration of participants' perspectives, enabling the identification of key themes and patterns within the qualitative survey data.

To analyze the single- and multiple-choice questions answered by participants, we calculated and displayed the sums and percentages for each category. Since the goal of this study was not to compare groups, no statistical comparisons were conducted.

Ethical Considerations

In phases 1 and 3 of the study, all participants provided written informed consent prior to participation. The study received approval from the national ethical committee (reference number DRN 2021‐02771 with amendment DRN 2024-03109-02). All procedures were conducted in accordance with the Principles of the Declaration of Helsinki (1975) and its subsequent revisions. All data collected were anonymized before analysis to protect participant privacy and confidentiality. The patient representative was actively involved in the project design and implementation. Their involvement was strictly voluntary and did not affect or interfere with their clinical care. No monetary or material compensation was provided for participation in this project.

Results

Phase 1

The phase 1 survey respondents (N=87) had little to no preference for the name and gender of the chatbot. The ideal chatbot persona should be personalized, akin to an acquaintance rather than a close friend. The survey revealed a strong willingness among participants to engage in anonymous conversations with chatbots about their mental well-being. Remarkably, nearly 7% of respondents were unsure whether they had interacted with a chatbot or a human. Work, stress, relationships, and conflicts emerged as the primary areas of interest for chatbot conversations (Table 1). While most participants found the provided topic list comprehensive, a few suggested the inclusion of "health stress." As the participants contributed data anonymously, they are distinguished by letters in the quotes (Participant A, B, C, etc).

Lack of Human Connection

The qualitative data gathered from the phase 1 open-ended questions painted a nuanced picture of the participants' attitudes toward chatbots for mental health support. While the potential benefits of anonymity, convenience, and availability were recognized, concerns about genuine human connection, empathy, and the effectiveness of chatbots were prevalent. Participants emphasized the importance of data security, trust, and ethical considerations in chatbot design, demanding clear guidelines and transparency regarding data handling. Participants expressed hesitations surrounding the limitations of chatbots in providing genuine human connection and empathy. They voiced concerns about the inability of chatbots to understand the depth and complexity of human emotions, particularly when dealing with sensitive mental health issues. The lack of personalized responses and the potential for generic advice further discouraged some participants.

> *Because I don't believe it could empathize with me. Some subjects are fine, but just the fact that it's not possible for it to relate to me, the fact being it's a computer with programmed responses. Some subjects aren't really a possibility with a regular therapist however, and a chatbot could be useful to talk to about really embarrassing matters, assuming there wasn't a leak. There is definitely something to be gained there, but I wouldn't want to talk to a computer about my feelings, because it doesn't have any.* [Phase 1: Participant A]

Privacy, Trust, and Data Security

Participants raised concerns about the privacy and security of their conversations with chatbots. They expressed uncertainty about how chatbot data are stored, who has access to it, and how it is used. The lack of clear guidelines and transparency regarding data handling eroded trust and discouraged open communication.

> *I would hesitate due to security reasons. Where does the chat data end up when we are done. I also think I would be too much aware that I'm not talking to a human who can give me any kind of trust around where the data about the chats ends up. Talking to any human you might feel that this person will keep my secrets, but talking to a bot you know that the bot will do whatever it is programmed to do…* [Phase 1: Participant C]

Utility and Effectiveness of Chatbots

Worries about the overall usefulness and effectiveness of chatbots for addressing mental health concerns were prevalent. Participants questioned the ability of chatbots to provide meaningful guidance and support, citing the lack of competence and the potential for misdiagnosis. They emphasized the importance of human interaction and expertise in providing comprehensive mental health care.

> *Mental health issues need to be address by humans.* [Phase 1: Participant E]

XSL•FO
RenderX

Design Requirements and Needs

Participants expressed clear requirements for a chatbot designed for mental health support. Users desired a chatbot that could provide a human connection and communication by offering reassurance, understanding, and empathy. In addition, users sought emotional support, guidance, and different perspectives on approaching challenges. They also wanted a chatbot that would validate their emotions and ensure they were not alone in their experiences. Finally, users valued features that facilitated self-care, personal growth, positivity, and motivation.

> *Affirmation, if possible, or sense of understanding and that I'm not the only one with the issue.* [Phase 1: Participant H]

> *Reassurance that I will make it. Explain that the feelings I have are normal, many people have them without talking about it. And maybe just listen to what I say and hmm...* [Phase 1: Participant I]

Phase 2

In phase 2 of the study, design requirements were used as a guideline for content, personality, and BETSY's appearance. The standard model offered by UNEEQ was used and resembled their previous avatar models (Figure 2). The group chose to extract a large portion of the topics from the survey in phase 1.

Figure 2. Two versions of Behavior, Emotion, Therapy System, and You: voice-activated digital human (left) and text-only chatbot (right).

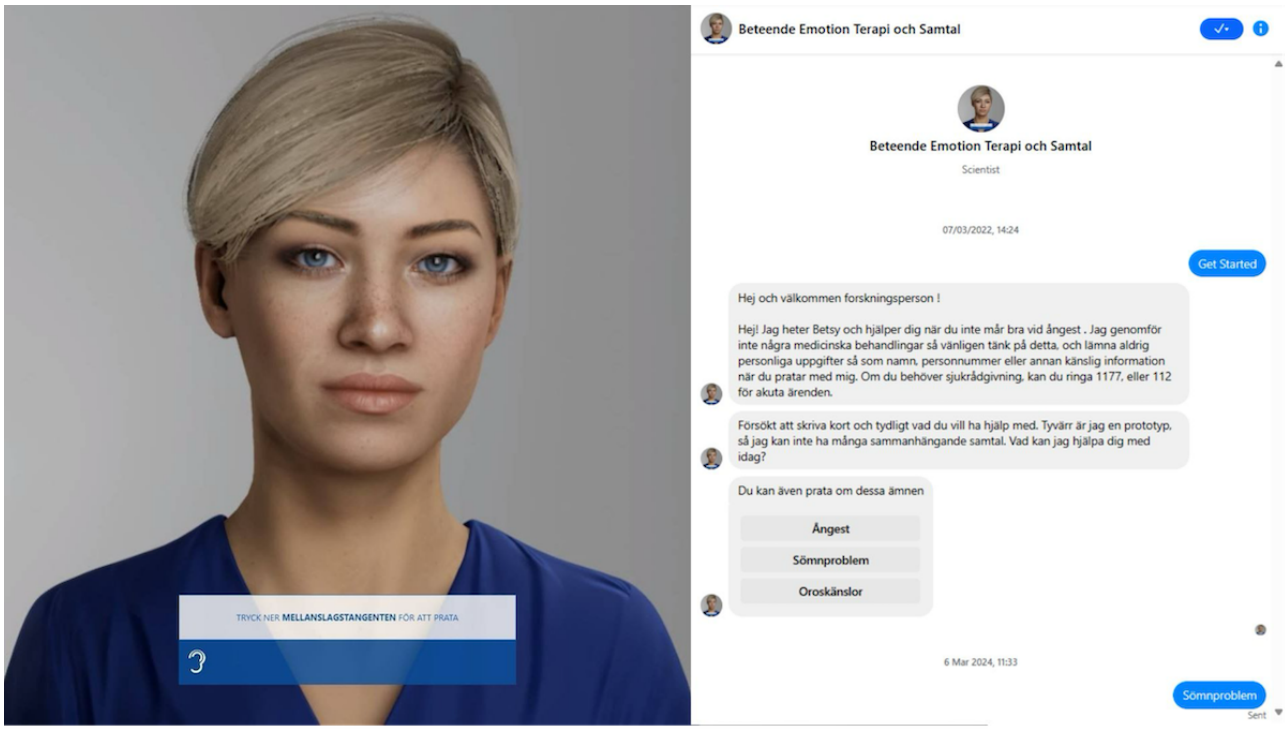

The chatbot's topic selection process involved several key aspects, including relevance, specificity, user-centricity, comprehensiveness, balance, accessibility, user empowerment, collaborative approach, and continuous improvement. The topics focused on addressing the core aspects of anxiety and its triggers, while also providing personalized guidance and catering to the unique needs of each user. The chatbot covered a wide range of anxiety-related issues, offering a holistic approach to support and addressing the diverse spectrum of anxiety experiences. A careful balance was maintained between common and less-discussed anxiety triggers, ensuring that both widespread concerns and lesser-known aspects of anxiety management were addressed. The topics were designed to support users with knowledge and coping mechanisms, with the aim of enabling them to take an active role in managing their anxiety symptoms. To ensure that the tone of BETSY was in line with the group's requirements for perceived empathy, the patient representative used BETSY at home continuously during the development process and reported back to the engineer; cases in which BETSY was perceived as less empathetic. This was not performed systematically or according to a scheme, as the patient representative experienced mental health problems and the team wanted to adapt the flow of the work as pragmatically as possible according to the patient representative's capability at the time.

Dialog-Tree Development and Implementation

The experts started the dialog-tree development process by sorting and extracting keywords and phrases from the Post-it Notes produced from the workshop. Initially, the experts adopted a professional jargon-filled approach, posing technical challenges for the chatbot, such as repetitive answering of questions with subsequent questions. Recognizing this issue, the engineer provided guidance, encouraging the experts to adopt a nonclinical approach, responding as they would to a nonpatient user and adapting the language to include more everyday phrases without clinical jargon. This shift in approach resulted in more concise and targeted responses, enhancing the chatbot's ability to provide effective guidance without venturing into diagnostics or triage. BETSY was provided with a

XSL•FO
RenderX

soft-voice tone, slowing down the pace of her speech to decrease the probability of the avatar feeling stressed or risk making the user feel hurried in their interaction [11]. Concurrently, the experts collaborated on identifying and integrating relevant resources that could be linked alongside the chatbot's responses. These resources were meticulously selected based on their credibility and authenticity, ensuring that only reliable and authoritative sources were provided (eg, the government site 1177.se). Links to governmental institutions and patient organizations were prioritized, guaranteeing that users had access to verified and reputable information. The compiled dialog tree, along with the training data comprising answered questions and related resources, was then integrated into 2 platforms: Dialogflow, a Google-based chatbot development platform, and Itsalive.io, a platform for text-based chatbots. The voice-based chatbot was released on a secluded page administered by Deloitte Digital (Deloitte Touche Tohmatsu Limited), while the text-based chatbot was deployed on Facebook from Meta on a secluded research page dedicated to BETSY. Users were to engage with both versions of BETSY on site (physical location) at Sahlgrenska University Hospital, maintaining their anonymity and without disclosing any personal information to the system.

To assess the performance of both chatbot versions, a test was conducted where both BETSY versions were asked the same question 100 times. In 99% of the cases, both chatbots provided identical responses. The only instance of discrepancy occurred when asked, "I have a hallucination. What should I do?" The digital human BETSY failed to comprehend the question, while the text-based BETSY accurately responded. The issue was fixed before participants were included in the study.

Phase 3

A total of 45 healthy volunteers participated in the evaluation of the chatbot prototype (20 participants in the text-only group and 25 participants in the avatar group). The evaluation methodology has been published elsewhere by our group with a summary of the protocol and a complete table of demographic properties [8]. Based on the open-ended questions posed to the 45 participants regarding BETSY's performance, there were positive and negative aspects to both interfaces (Table 2). As the participants were contributing data anonymously, they are referred to as numbers in the quotes to distinguish them (Participant 1, 2, 4, etc).

XSL•FO
RenderX

Table . Examples of positive and negative responses to Behavior Emotion Therapy System and You's text-only chatbot and voice-activated digital human.

Themes and question	Text-only	Voice-only
Usefulness		
Under what circumstances would you consider talking to BETSY[a]?	*For first contact. if I feel stressed, anxious, or have other simpler mental health problems, stress, screening.* [Participant 1]	*For lighter mental health condition. When I need quick and concrete help. BETSY provided clear and timely information, but didn't seem to be that complex in her problem-handling skills.* [Participant 4]
—	*Practical exercises to deal with stress, muscle tension resulting from high workload.* [Participant 2]	*Concrete tips. The exercises were good. Being referred to info about my problem to learn more about different strategies was also good.* [Participant 5]
Screening	*For first contact if I feel stressed, anxious, or have other simpler mental health problems.* [Participant 6]	*Before I contact my healthcare provider.* [Participant 7]
Taking offense		
Is there anything that BETSY said that you would take offense to hearing when you felt bad?	*Repetitive answers that are not grounded in what we talked about...I feel unseen and unheard.* [Participant 8]	*I probably would have reacted to the tone of voice and the number of words. A real person would talk less and listen more, at least until the problem was circled. I had reacted to words she emphasizes wrong. It becomes clear that it is a robot. Her face and eyes feel more person-like, alive.* [Participant 9]
Emotional connection		
Did you feel any connection to BETSY?	*No, not so much. Probably because of the repetitiveness. Felt like I'm not always heard*. [Participant 10]	*Yes, I found her gaze present. The way she talked, a soft and pleasant voice. Clear when she made claims/advice.* [Participant 11]
Negative technical aspects		
What would you like to improve with BETSY?	*More interactivity, not so much repetition of the same response. If I've already received sleep tips/link, I don't want the same more times. She should understand better what I'm asking of her.* [Participant 12]	*The exchange became 'choppy', felt that BETSY did not want to listen to how I described my feelings. The answers/feedback were related directly to the facts. Lacked the sense of empathy.* [Participant 13]
Overall positive aspects		
Did you feel any connection to BETSY?	*Above all, that the answers felt more personal and not just fact-based text but a more personal language.* [Participant 14]	*Yes, she had a pleasant voice and left a pleasant impression.* [Participant 15]
Suggestions for the future		
What would you like to improve with BETSY?	*Possibly her ability to pick things up out of the conversation. More nuanced. Clearly, she picked up guiding words in the sentence and not sentences, which sometimes made the support incorrect.* [Participant 16]	*Voice understanding. Not linking out to YouTube. Feels frivolous, the logic and understanding that I was talking about. She needs to get better at keeping the conversation going*. [Participant 17]

[a]BETSY: Behavior, Emotion, Therapy System, and You.

Users of both interfaces considered BETSY a valuable adjunctive tool for the management of mild to moderate anxiety and for providing self-help exercises. Participants emphasized the need for a more comprehensive set of exercises, particularly for stress management and focus enhancement. BETSY's potential as a screening tool for health care was consistently recognized. Users viewed BETSY as an accessible first step before seeking professional help. The suggestion to integrate a screening procedure into BETSY's capabilities highlights its potential to alleviate health care system burdens.

Despite BETSY's perceived usefulness, participants identified conversational issues. Both groups found BETSY overly talkative and her responses excessively lengthy. The tendency to divert conversations from relevant topics raised concerns about being unheard and unseen by several users. These issues emphasized the need for more responsive and relevant conversational design. Some participants, particularly in the digital human group, reported experiencing an emotional connection with BETSY, attributing it to her sympathetic, kind, and comforting nature. This finding underscored the potential for chatbots to establish meaningful interactions, especially in voice-activated interfaces. Further research could explore

XSL•FO
RenderX

methods to enhance this connection for a wider user base. Most participants expressed concerns about BETSY's technical aspects, viewing it as incomplete. Issues included repetitive responses, misinterpretations of user input, button visibility in the voice-activated digital human interface, and looping in the text-only version. The most disruptive aspect was redirecting users to external webpages for relaxation exercises and information. Participants suggested presenting information within the chat interface for seamless conversational flow.

Discussion

Principal Findings

In this study, we aimed to co-design a conversational automated interface and system intended to alleviate mild to moderate anxiety symptoms. Through a 3-phase process, we explored end-user requirements by integrating end-user input from the general public. Using the end-user requirements together with specialized expertise, 2 interfaces were developed of the BETSY chatbot: a text-based and a digital human voice-activated avatar. Due to regulatory restrictions surrounding medical devices and technology targeting mental health issues, we enlisted the assistance of healthy participants to assess the system's usability and its alignment with the initial end-user survey's requirements, while at the same time consistently keeping a user-centered perspective through the involvement of a patient representative as a continuous source for testing and validation. Similar to prior research, end users identified privacy and security as critical factors influencing their trust in a potential system [7]. Work and relationships emerged as highly ranked topics that end users desired to discuss with a potential system. While we only received responses from 87 members of the general public, we observed a diverse range of perspectives and interests, which we hypothesize accurately reflects the general population's fears, concerns, and interests: this has been shown to be frequently present in other studies [12]. However, the survey's exclusive distribution via social media may have excluded individuals without social media accounts, potentially limiting the diversity of perspectives. Due to the anonymous nature of the phase 1 survey where metadata such as IP addresses was omitted, it is not possible to decipher whether one person had answered the survey multiple times. Despite this limitation, we noted that phase 1 respondents expressed diverse perspectives and different levels of enthusiasm for chatbot technology, including individuals who were skeptical or unfamiliar with the technology, which suggests a broad representation of the general public. One recurring concern was the system's inability to understand feelings and relation to human emotions, which could hinder feelings of usability and connection to the chatbot. This has been observed in other studies in which the participants felt hesitations about sharing emotional states with a system that could not understand what it felt like to feel, thus could not produce true empathy [12,13]. It is also worth noting that this examination was performed before the release of commercial large language models (LLMs) or systems which use these for therapy, such as Character.ai (Character Technologies, Inc), OpenAI (OpenAI, Inc), etc; thus, the opinions of the participants do not reflect those systems but are solely focused on rules-based neuro-linguistic programming (NLP) systems with little or no machine learning capability. Conversations that are constructed using LLMs are less likely to produce the same level of readiness and repetitiveness as rules-based NLP models such as BETSY [14,15]. Many, if not all problems of repetitiveness are minimized if not diminished through the use of LLMs.

A personalized, nonjudgmental chatbot with a clear identity (named) and a focus on empathy was preferred by participants. While demographic data were collected in phase 1 and used to guide certain design decisions, we recognize the value of making these connections more explicit. Participant preferences for a semiformal tone, for example, directly informed BETSY's conversational style, with the majority favoring communication "like an acquaintance" (58%). Similarly, the decision to implement a female-presenting avatar was influenced by local psychiatric care demographics, where female nurses are more prevalent. While these choices were rooted in user feedback and contextual relevance, the sample's representativeness was limited by the recruitment strategy, which primarily relied on social media platforms and lacked detailed data on ethnicity or socioeconomic background. Future design phases will aim for broader inclusion to explore how a more diverse demographic composition might inform personalization and user experience.

A strong willingness to engage anonymously with chatbots for mental well-being was identified. However, concerns regarding data security and privacy need to be addressed to build user trust. Attitudes toward digital care change when trust in cybersecurity is adequate. This was evident in both our investigation and investigations of others [7,12,13].

In phase 2, we used the end-user requirements as a guiding principle for the prototype, incorporating as many features as possible within the prototype's framework. We aimed for a representative and equitable group of experts in terms of gender, age, and areas of expertise. The system was designed as a female avatar, a choice based on gender representation statistics from our hospital. However, in retrospect, given the lack of preferences expressed by phase 1 respondents, we could have explored more diverse options. Female-presenting avatars are commonly used in research on digital humans [16]. Exploring alternative representations could have provided insights beyond those currently available in other studies. In future phases, we plan to diversify the representations available to end users. Diversity represented in avatars can help to broaden the trustworthiness and representations of the mental health profession [17], especially as in our broader study, we found that women were more prone to report irritation when talking to BETSY [18].

Our iterative design process yielded valuable insights. Regrettably, due to the scope of this investigation, we did not collect our own thoughts and experiences in a qualitative or quantitative manner. We believe that including expert reflections and evaluations would have enriched the results of this co-design initiative. In future studies, we will incorporate these considerations. We chose to explore 2 interfaces, text-only and voice paired with a face with anthropomorphic features. Face-to-face communication is unique and adds many dimensions to communication, both in human animals and

XSL•FO
RenderX

others. Unlike interacting with objects, faces trigger a sense of ethical responsibility [18]. Systematic evaluations of chatbot research have shown that the level of humanness is a key factor in the usability of chatbots [19]. In phase 2, we explored incorporating expressions that leaned toward empathetic and semiformal. In our efforts to create equally empathetic interfaces, we explored research on text chatbots, which yielded a comprehensive list of traits that were desirable, such as affirmative language, ability to show affect toward the subject, and a high level of consciousness [11,12,20-25]. In 2002, Morris [26] proposed that using a language which corresponds with the emotional state of the chatbot (eg, in the case of BETSY, showing irritation or anger toward a workplace bully when the user notifies her that they have been subjected to such behavior) along with their given role (eg, caregiver) and social awareness (comforting a person in distress) results in more believable interactions [26].

Comparison With Previous Work

In our evaluation of BETSY, we employed healthy volunteers (for more detailed methods see [8]). Our evaluation of the solution indicated that most end-user requirements were met. The system was perceived as useful both in terms of qualitative markers as shown in this co-design evaluation as well as quantitatively through the System Usability Scale in our previous publication of the quantitative metrics related to this investigation [8]. Similar to the end-user requirements noted in phase 1 of the co-design of the prototype, the participants who evaluated the prototype noted that BETSY was useful in providing affirmation and information about specific mental health problems and about the mental state they were either experiencing or simulating to BETSY, that is, stress, problems at work, and more. Our design choice yielded a positive response with regard to her persona, voice, and feeling of presence in the digital human version of BETSY, and the results correspond to other studies in the field [11]. Overall, the design of BETSY was well received and highlighted, as other studies have proposed [2-5,11], that working together with end users before initiating the design process is an important component in creating user-friendly experiences, independently of which interface is targeted for the chatbot.

Limitations

Several technical challenges and limitations identified in this investigation could be addressed through the usage of LLMs or other systems capable of generating responses based on instructions and datasets containing sufficient information for generating novel and relevant responses [14]. However, these technologies were not available at the inception of this project and were thus not used in the design. In addition, while LLMs have demonstrated high usability and emotional connectivity in some studies [14], they can also exhibit unpredictable behavior and pose risks to vulnerable individuals through the generation of harmful responses such as hallucinations [27,28] or the potential compromise of personal data in systems owned by private entities that provide limited transparency regarding their LLM-training methodologies. As BETSY was never used outside the scope of the tests of this prototype, we acknowledge that some of the concerns and aspects of potential chatbot systems were not addressed, such as cybersecurity. When creating a system that handles something as private and vulnerable as an individual's mental health, it is of utmost importance to discuss the ethical and legal implications of the digital infrastructure of the systems, as there are many aspects of technology that might lead to harm for the user [28,29]. In our study, we opted to protect users by providing them with an on-site computer and instructing them to avoid personal problems. This would be neither feasible nor probable in the delivery of a real system, a system which would be dependent on strong end-to-end encryption, possible autodeletion of messages, and careful IT architecture in order to address the safety and security concerns from phase 1 end-user requirements. As deployment of a mental health service was not the aim of this study, we did not delve into potential digital infrastructure. In future studies and product development, we will make the results of this co-design initiative central to our work as well as to explore, test, and develop methodologies that enable safe and ethical user experience [30].

Conclusions

We have developed a conversational automated interface and system, BETSY, by co-designing with end users and experts. Collaboration and co-design resulted in a system that was positively received by healthy volunteers. The system met most end-user requirements and was perceived as useful. While LLMs could address some technical limitations and negative technical aspects of BETSY (eg, repetitiveness), the potential risks and ethical considerations warrant further evaluation before their incorporation into mental health-related systems.

Acknowledgments

The authors would like to thank the participants of this study for their time and contribution to valuable knowledge. The authors also especially thank Mathias Alvidius, Head of Department, for contributing significantly to the possibilities of the creation of this research project. The authors would also like to express sincere gratitude to Linn Dahlén Ölander for her help in recruiting participants. A special thanks to digital director Anna-Marie Schaffrath and digital strategist Johannes van den Brink for their support in providing their knowledge and resources in digitalization in order for us to perform this exploration, which we are sure will benefit the development of accessible mental health tools. SS received support from the Söderström König Foundation (SLS-974247), Stiftelsen Psykiatriska Forskningsfonden for this project. Parts of this project were funded by Innovationsfonden VGR (project VGRINN-936800) and Västra Götalandsregionen. The authors would also like to thank Peter Todd (Tajut Ltd) for third-party editorial assistance in the drafting of this manuscript, for which he received financial compensation.

XSL•FO
RenderX

The study was financed by grants from the Swedish state under the agreement between the Swedish Government and the county councils, the ALF agreement (ALFGBG-1006029).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

The author MB is currently not affiliated with any institution but is an Independent Researcher.

Conflicts of Interest

None to declare

References

1. Bevan Jones R, Stallard P, Agha SS, et al. Practitioner review: co-design of digital mental health technologies with children and young people. J Child Psychol Psychiatry 2020 Aug;61(8):928-940. [doi: 10.1111/jcpp.13258] [Medline: 32572961]
2. Wright M, Getta AD, Green AO, et al. Co-designing health service evaluation tools that foreground first nation worldviews for better mental health and wellbeing outcomes. Int J Environ Res Public Health 2021 Aug 13;18(16):8555. [doi: 10.3390/ijerph18168555] [Medline: 34444319]
3. Potts C, Ennis E, Bond RB, et al. Chatbots to support mental wellbeing of people living in rural areas: can user groups contribute to co-design? J Technol Behav Sci 2021;6(4):652-665. [doi: 10.1007/s41347-021-00222-6] [Medline: 34568548]
4. Patrickson B, Musker M, Thorpe D, van Kasteren Y, Bidargaddi N, The Consumer and Carer Advisory Group (CCAG). In-depth co-design of mental health monitoring technologies by people with lived experience. Future Internet 2023;15(6):191. [doi: 10.3390/fi15060191]
5. Porche MV, Folk JB, Tolou-Shams M, Fortuna LR. Researchers' perspectives on digital mental health intervention co-design with marginalized community stakeholder youth and families. Front Psychiatry 2022;13:867460. [doi: 10.3389/fpsyt.2022.867460] [Medline: 35530032]
6. Grové C. Co-developing a mental health and wellbeing chatbot with and for young people. Front Psychiatry 2020;11:606041. [doi: 10.3389/fpsyt.2020.606041] [Medline: 33597898]
7. Hamlin M, Steingrimsson S, Cohen I, Bero V, Bar-Tl A, Adini B. Attitudes of the public to receiving medical care during emergencies through remote physician-patient communications. Int J Environ Res Public Health 2020 Jul 20;17(14):5236. [doi: 10.3390/ijerph17145236] [Medline: 32698481]
8. Thunström AO, Carlsen HK, Ali L, Larson T, Hellström A, Steingrimsson S. Usability comparison among healthy participants of an anthropomorphic digital human and a text-based chatbot as a responder to questions on mental health: randomized controlled trial. JMIR Hum Factors 2024 Apr 29;11:e54581. [doi: 10.2196/54581] [Medline: 38683664]
9. Radziwill NM, Benton MC. Evaluating quality of chatbots and intelligent conversational agents. arXiv. Preprint posted online on Apr 5, 2017. [doi: 10.48550/arXiv.1704.04579]
10. Naeem M, Ozuem W, Howell K, Ranfagni S. A step-by-step process of thematic analysis to develop a conceptual model in qualitative research. Int J Qual Methods 2023 Oct;22. [doi: 10.1177/16094069231205789]
11. Chaves AP, Gerosa MA. How should my chatbot interact? A survey on social characteristics in human–chatbot interaction design. Int J Hum Comput Interact 2021 May 9;37(8):729-758. [doi: 10.1080/10447318.2020.1841438]
12. Abd-Alrazaq AA, Alajlani M, Ali N, Denecke K, Bewick BM, Househ M. Perceptions and opinions of patients about mental health chatbots: scoping review. J Med Internet Res 2021 Jan 13;23(1):e17828. [doi: 10.2196/17828] [Medline: 33439133]
13. Chin H, Song H, Baek G, et al. The potential of chatbots for emotional support and promoting mental well-being in different cultures: mixed methods study. J Med Internet Res 2023 Oct 20;25:e51712. [doi: 10.2196/51712] [Medline: 37862063]
14. Martins A, Londral A, L Nunes I, V Lapão L. Unlocking human-like conversations: scoping review of automation techniques for personalized healthcare interventions using conversational agents. Int J Med Inform 2024 May;185:105385. [doi: 10.1016/j.ijmedinf.2024.105385] [Medline: 38428201]
15. Dale R. GPT-3: what's it good for? Nat Lang Eng 2021 Jan;27(1):113-118. [doi: 10.1017/S1351324920000601]
16. Wilson L, Marasoiu M. The development and use of chatbots in public health: scoping review. JMIR Hum Factors 2022 Oct 5;9(4):e35882. [doi: 10.2196/35882] [Medline: 36197708]
17. Do TD, Zelenty S, Gonzalez-Franco M, McMahan RP. VALID: a perceptually validated Virtual Avatar Library for Inclusion and Diversity. Front Virtual Real 2023;4. [doi: 10.3389/frvir.2023.1248915]
18. Corti K, Gillespie A. A truly human interface: interacting face-to-face with someone whose words are determined by a computer program. Front Psychol 2015;6:634. [doi: 10.3389/fpsyg.2015.00634] [Medline: 26042066]
19. Rapp A, Curti L, Boldi A. The human side of human-chatbot interaction: a systematic literature review of ten years of research on text-based chatbots. Int J Hum Comput Stud 2021 Jul;151:102630. [doi: 10.1016/j.ijhcs.2021.102630]

XSL•FO
RenderX

20. Shum HY, He XD, Li D. From Eliza to XiaoIce: challenges and opportunities with social chatbots. Frontiers Inf Technol Electronic Eng 2018 Jan;19(1):10-26. [doi: 10.1631/FITEE.1700826]
21. Croes EAJ, van Wezel MMC, Antheunis ML. I'm here for you": can social chatbots truly support their users? A literature review. Presented at: Chatbot Research and Design: 4th International Workshop, CONVERSATIONS 2020, Virtual Event; Nov 23-24, 2020 p. 96-113. [doi: 10.1007/978-3-030-68288-0_7]
22. Vaidyam AN, Wisniewski H, Halamka JD, Kashavan MS, Torous JB. Chatbots and conversational agents in mental health: a review of the psychiatric landscape. Can J Psychiatry 2019 Jul;64(7):456-464. [doi: 10.1177/0706743719828977] [Medline: 30897957]
23. Ruane E, Farrell S, Ventresque A. User perception of text-based chatbot personality. Presented at: Chatbot Research and Design: 4th International Workshop, CONVERSATIONS 2020, Virtual Event; Nov 23-24, 2020 p. 32-47. [doi: 10.1007/978-3-030-68288-0_3]
24. Kang M. Non-verbal emotional expressions for social presence of chatbot interface. J Korea Contents Assoc 2021;21(1):1-11 [FREE Full text]
25. Denecke K, Vaaheesan S, Arulnathan A. A mental health chatbot for regulating emotions (SERMO) - concept and usability test. IEEE Trans Emerg Topics Comput 2020;9(3):1170-1182. [doi: 10.1109/TETC.2020.2974478]
26. Morris TW. Conversational agents for game-like virtual environments. Presented at: AAAI 2002 Spring Symposium on Artificial Intelligence and Interactive Entertainment II URL: https://aaai.org/papers/0016-SS02-01-016-conversational-agents-for-game-like-virtual-environments/ [accessed 2024-08-15]
27. Alkaissi H, McFarlane SI. Artificial hallucinations in ChatGPT: implications in scientific writing. Cureus 2023 Feb;15(2):e35179. [doi: 10.7759/cureus.35179] [Medline: 36811129]
28. Blease C, Torous J. ChatGPT and mental healthcare: balancing benefits with risks of harms. BMJ Ment Health 2023 Nov;26(1):e300884. [doi: 10.1136/bmjment-2023-300884] [Medline: 37949485]
29. Vilaza GN, McCashin D. Is the automation of digital mental health ethical? Applying an ethical framework to chatbots for cognitive behaviour therapy. Front Digit Health 2021;3:689736. [doi: 10.3389/fdgth.2021.689736] [Medline: 34713163]
30. Banerjee S, Agarwal A, Bar AK. Securing well-being: exploring security protocols and mitigating risks in AI-driven mental health chatbots for employees. Am J Comput Sci Technol 2024 Apr;7(1):1-8. [doi: 10.11648/j.ajcst.20240701.11]

Abbreviations

BETSY: Behavior, Emotion, Therapy System, and You
LLM: large language model
NLP: neuro-linguistic programming

Edited by A Mavragani; submitted 05.09.24; peer-reviewed by C Wang, C Cha; revised version received 29.08.25; accepted 30.08.25; published 30.10.25.

Please cite as:
Thunström AO, Ali L, Krage Carlsen H, Bohm M, Wesén L, Wrede O, Sarajlic Vukovic I, Hellström A, Larson T, Steingrimsson S Behavior Emotion Therapy System and You: Co-Design and Evaluation of a Mental Health Chatbot and Digital Human for Mild to Moderate Anxiety in Healthy Participants
JMIR Form Res 2025;9:e66163
URL: https://formative.jmir.org/2025/1/e66163
doi:10.2196/66163

© Almira Osmanovic Thunström, Lilas Ali, Hanne Krage Carlsen, Maria Bohm, Linda Wesén, Olof Wrede, Iris Sarajlic Vukovic, Andreas Hellström, Tomas Larson, Steinn Steingrimsson. Originally published in JMIR Formative Research (https://formative.jmir.org), 30.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Assessing the Usability of a Prescription-Based Mobile App for Patients With Panic Disorder and Its Management Console for Clinicians: Controlled User Study

Yujin Ko[1,2], MD; Jeemin Lee[1], MA; Kyunghee Ham[1], MA; Yesol Cho[1], MS; Yu-Bin Shin[1], PhD; Choongki Min[3], BS; Kyungnam Kim[3], ME; Wonseuk Jang[4], PhD; Hayeon Jung[4], BE; Jae-Jin Kim[1,5], MD, PhD

[1]Institute of Behavioral Sciences in Medicine, Yonsei University College of Medicine, Seoul, Republic of Korea

[2]Department of Psychiatry, Soonchunhyang University Bucheon Hospital, Bucheon, Republic of Korea

[3]AI Medtech R&D, Waycen Inc., Seoul, Republic of Korea

[4]Yonsei University Gangnam Severance Hospital Medical Device Usability Research Center, Seoul, Republic of Korea

[5]Department of Psychiatry, Yonsei University Gangnam Severance Hospital, 211 Eonju-ro, Gangnam-gu, Seoul, Republic of Korea

Corresponding Author:
Jae-Jin Kim, MD, PhD
Institute of Behavioral Sciences in Medicine, Yonsei University College of Medicine, Seoul, Republic of Korea

Abstract

Background: Panic disorder is characterized by sudden panic attacks and persistent anticipatory anxiety. While pharmacotherapy remains effective, patients with panic disorder often experience residual symptoms and functional impairments. Lifestyle factors influence symptom severity but are often unaddressed in routine psychiatric care. Most current digital therapeutics for panic disorder have a limited scope, lack integration with clinicians, and fail to consider behavioral patterns. To address these limitations, our research team developed a prescription-based app that supports structured cognitive behavioral therapy practice, real-life symptom management, and lifestyle modifications for patients with panic disorder, and a management console—a web-based platform that allows clinicians to monitor the patients' engagement and progress as well as determine therapeutic options if necessary.

Objective: This study aimed to test the usability of the app and management console by evaluating their interface, functionality, and user experience. The primary goal was to identify the strengths and areas for further improvement of these software devices and to develop a list of modifications to improve the user experience and clinical applicability in updates to refine the devices for a future clinical trial.

Methods: Usability data were collected by investigators at a medical device usability research center without the involvement of the development research team, and the participants were 15 patients with panic disorder and 15 psychiatrists. Each group completed experimental use of the app or management console and scored the convenience and safety of its modules, questionnaire evaluations for the acceptability, and presentation of verbal subjective feedback on areas for improvement. Based on the participants' suggestions, a list of items that need to be modified to improve functionality and ease of use for each device was created.

Results: Patients completed 155 assigned tasks for the app with more than 98% success, and psychiatrists completed 34 tasks for the management console with more than 86% success. The convenience and safety scores for the app and management console exceeded the neutral threshold (mean >4.5). For all statements about the acceptability, both patients and psychiatrists responded at the level of agreeing with a score exceeding 3 (mean: 3.6~4.3 and 4~4.7, respectively). There were 38 suggestions for app improvements and 66 suggestions for management console improvements, most of which were incorporated in the modification list.

Conclusions: Patients reported that the app might be easy to use and help manage anxiety, and psychiatrists found the management console practical and well-suited for outpatients. By combining patient-facing therapeutic tools with clinician-driven prescription and monitoring, the devices offer a solution aligned with clinically integrated, real-world psychiatric care. Modified devices based on the improvement suggestions presented in this study will be evaluated in future clinical trials for their impact on engagement and treatment outcomes.

(JMIR Form Res 2025;9:e76843) doi:10.2196/76843

KEYWORDS

panic disorder; digital therapeutics; mobile application; usability testing; cognitive behavioral therapy; patient-centered design; clinician-guided care; real-time intervention; lifestyle modification

XSL•FO
RenderX

Introduction

Panic disorder is one of the most common disorders worldwide, characterized by unexpected and recurrent panic attacks, along with intense fear, arousal, and a sense of losing control [1]. These symptoms can significantly lower the quality of life, disturb daily functioning, and even result in avoidance behaviors. Panic disorder has a lifetime prevalence of approximately 1% to 4%, making it a large public health burden [2].

While pharmacotherapy remains the primary treatment modality, cognitive behavioral therapy (CBT) has also proven efficacy [3,4]. CBT is a structured psychotherapy that helps patients identify and modify maladaptive thought processes and behavior patterns. For panic disorder, CBT involves psychoeducation, cognitive restructuring, interoceptive and situational exposure, and breathing retraining [5]. However, in real-world settings, patients often continue to experience frequent panic attacks in their daily lives with limited access to timely or structured psychological support beyond medication. Real-world practice of CBT remains limited by the requirement of frequent visits, unavailability of trained therapists, and practical challenges such as patients' time constraints or geographic limitations [6]. In addition, lifestyle factors such as caffeine intake, alcohol consumption, physical activity, and sleep patterns have been shown to influence the course of panic disorder symptoms [7-10]. Despite their importance, these factors are often left unaddressed in standard care, leaving patients to manage them on their own without proper guidance.

In recent years, digital therapeutics (DTx) have emerged as a novel approach for providing psychological interventions in a more accessible and convenient format. Recent studies have shown the effectiveness of DTx for panic disorder. For example, Freespira, approved by the US Food and Drug Administration (FDA) for the treatment of panic disorder, has been shown to be effective in reducing symptoms by training patients to regulate their respiration rate and CO_2 exhalation, with improvements lasting for up to 12 months after treatment [11]. PanicMechanic (Amshuhu iTech Solution Pvt. Ltd.), a smartphone–based app, provides real-time biofeedback during actual panic attacks by monitoring heart rate through the camera sensor, offering support in daily life situations [12]. In another recent study, a mobile CBT-based app designed for panic disorder showed improvement of symptoms in more than 80% of participants, using a combination of psychoeducation, journaling, various tasks, and gamified content [13].

Despite the advancement of DTx for panic disorder, most still fall short of providing a comprehensive approach that includes cognitive, behavioral, physiological, and lifestyle-related factors [14]. A common limitation of clinician-facing instruments is the limited availability that allows physicians to prescribe treatment, monitor progress, and offer feedback [15]. For these platforms to be adopted into standard psychiatric care, especially within time-limited outpatient practice, their functionality should allow for real-time monitoring of symptoms and meaningful patient-clinician communication. These demands highlight the need for an enhanced and integrated DTx model that unites structured CBT, an immediate panic attack support system, a lifestyle management system, and a dedicated clinician interface.

To address this need, we developed Waymed_panic (Waycen Inc.), a system consisting of a prescription-based DTx app and web-based platform to treat panic disorder. This study tested the usability of the app and the management console by evaluating their interface, functionality, and users' experience among patients with panic disorder and psychiatrists. The primary goal of this usability test was to identify the strengths and areas for further improvement of these software devices. Through this test, we aimed to develop a list of modifications to improve the users' experience and clinical applicability in updates to refine the devices for a future clinical trial.

Methods

Overview

This study was performed to evaluate participant perspectives of usability and satisfaction with Waymed_panic, which consists of a therapeutic app that provides training, companionship, and care services for patients with panic disorder and a management console for clinicians (Figure 1). A detailed description of their contents is provided in Multimedia Appendix 1. The primary goal of the app is to improve acute panic symptoms through the delivery of CBT training, real-time coping support, and lifestyle modification and management tools, all within a user-friendly mobile interface. In parallel, the web-based platform allows clinicians to monitor the patients' engagement and progress and to determine therapeutic options, if necessary. This system aims to bridge the gap between stand-alone mobile interventions and integrated psychiatric care by combining self-guided digital tools with clinician-directed oversight.

XSL•FO
RenderX

Figure 1. Illustration showing the overall structure of Waymed_panic. The app consists of 3 major services—training, companion, and care—and a service assistance section. It also includes a management console through which clinicians can prescribe the app, review patient records, and provide feedback.

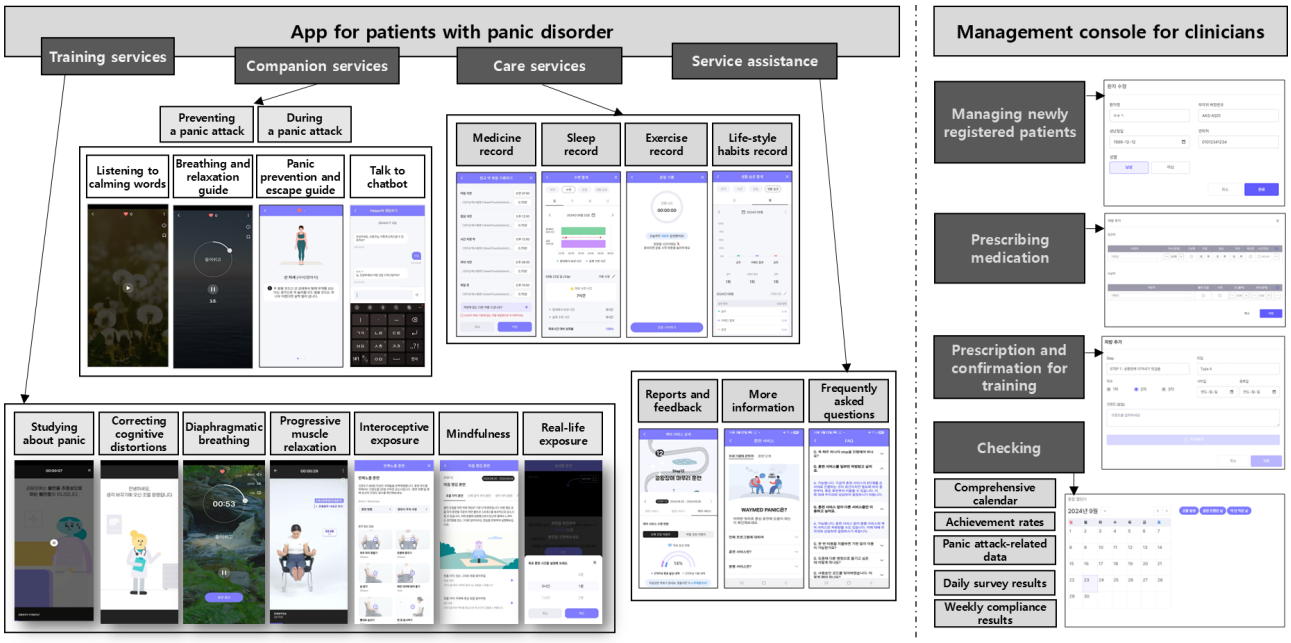

This system was developed in approximately 2 years as a collaborative project by a multidisciplinary research team. The medical content was developed by psychiatrists working at university hospitals based on clinical evidence and treatment guidelines for panic disorder [16-18], and all content components followed evidence-based principles. Clinical psychologists researching at a medical school research center assisted in the initial design and creation of the CBT content and provided specialist consultation on clinical materials and composition. While developing the therapeutic content, major clinical guidelines and key literature on CBT for panic disorder [5,19] were referenced, and core services were structured based on established intervention models [6,20-22]. User interface and user experience designers from a health care app development company developed a user-friendly, intuitive interface that was easy to navigate and visually clear. App programmers from the same company built the technical structure for the app, developed both Android and iOS versions, and integrated features. The team focused on refining the prototype and adjusting the user flow to ensure clinical and functional consistency through regular meetings for determining the design of a stepwise unlocking approach, the structure of panic-time interventions, and a web-based console for clinicians. In this study, participant recruitment and evaluation were conducted by investigators at a medical device usability research center, without involvement of the development research team.

Participants

Patients with panic disorder were recruited through advertisements posted in the psychiatric outpatient clinic of the Yonsei University Gangnam Severance Hospital, and psychiatrists were recruited through an advertisement posted on the internet-based community of psychiatrists affiliated with the Yonsei University Health System. Inclusion criteria for the patient group were adults aged 19 to 59 years, diagnosed with panic disorder according to *DSM-V* (*The Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition*) criteria through a medical history interview with a psychiatrist, and able to use mobile apps manually, hear speech, read and write Korean, understand visual elements such as signs and symbols, and detect type, size, shape, and color of visual stimuli. Inclusion criteria for psychiatrists were the same as for patients, except that they had to be a psychiatrist who treats panic disorder instead of being diagnosed with panic disorder. Through those advertisements, 21 patients and 16 psychiatrists were recruited, and among them, 15 patients (7 males and 8 females; mean age 32.9, SD 10.9 years) and 15 psychiatrists (10 males and 5 females; mean age 33.3, SD 5.4 years) were the final study participants, who met the above criteria and agreed to participate in the study.

Procedure

All participants visited the laboratory of the Yonsei University Gangnam Severance Hospital Medical Device Usability Research Center, where they completed experimental use of the software device, questionnaire evaluations, and presentation of verbal subjective opinions on areas for improvement. They first watched a preprepared educational video that introduced patients to the app and psychiatrists to the management console.

Then, they had at least 30 minutes to actually try out the app installed on a smartphone (Samsung Galaxy S20) or the management console installed on a tablet computer (Samsung Galaxy Tab A8). During these experimental uses, they performed the tasks executing the contents of the app and management console at the request of the investigator. Each execution task was presented with a visual projection and an oral reading from the investigator in parallel to ensure that the participants could easily understand the corresponding instruction. A different number of execution tasks was given to the patient participants for each of the modules that make up the app (ie, running the app, training services, companion

XSL•FO
RenderX

services, care services, and service assistance functions), and the total number of tasks was 155 (Multimedia Appendix 2). Execution tasks were not applied to all modules to ensure effective progress of the investigation. For example, since among the training services, progressive muscle relaxation has similar aspects to other trainings and real-life exposure is not suitable for conducting in a laboratory setting, these 2 trainings were not included as execution tasks to prevent the overall task execution time from becoming too long. Execution tasks were also given to the psychiatrist participants in different numbers for each of the modules that make up the management console (ie, running the console, managing newly registered patients, prescribing medication, prescription and confirmation for training, and checking the comprehensive calendar, achievement rates, panic attack–related data, daily survey results, and weekly compliance results), and the total number of the tasks was 34 (Multimedia Appendix 3).

The investigator observed the participants' behavior related to performing each execution task and recorded whether the task was completed or not. In principle, participants performed the tasks on their own without any assistance from the investigator, but if they requested help due to difficulty performing a task or failed to complete it for more than 60 seconds, the investigator provided assistance. Both of these instances were considered task failures and were reflected in the calculation of the task success rate. As each execution task was completed, the participants responded to a 5-point Likert scale to rate the convenience (1: very difficult, 2: somewhat difficult, 3: neutral, 4: somewhat easy, and 5: very easy) and safety (1: very dangerous, 2: slightly dangerous, 3: neutral, 4: slightly safe, and 5: very safe) of the task.

After completing all the execution tasks, the participants responded to a questionnaire to assess the acceptability of the app or management console, which consisted of 10 different statements (Table 1) about its use on a 5-point Likert scale (1: strongly disagree, 2: somewhat disagree, 3: neutral, 4: somewhat agree, and 5: strongly agree). These statements were prepared by modifying Nielsen's usability heuristics [23] to suit the contents of each device, and additional references in these modifications were the System Usability Scale [24] for the app and the Post-Study System Usability Questionnaire [25] for the management console. In addition, the investigator recorded the participants' verbal subjective opinions about areas for improvement due to the difficulties they encountered in each module while performing the execution tasks and also recorded the overall verbal opinions on areas for improvement and satisfaction with the app or management console.

XSL•FO
RenderX

Table . Statements to evaluate acceptability of the app and management console, and the acceptability scores in patients with panic disorder and psychiatrists.

Statements	Score[a], mean (SD)
For patients (n=15)	
The system is easy to use.	4.1 (0.6)
The system does not require any technical assistance to use.	4.1 (0.9)
The system is not unnecessarily complicated.	3.7 (1)
The system provides useful information about the training outcomes.	4.3 (0.7)
The system offers a wealth of options for a variety of features.	4.3 (0.8)
The system is designed with easy-to-navigate content and functional organization.	3.9 (0.9)
The system is designed to fit into an environment in which CBT[b] is practiced.	4.1 (0.8)
The system is designed to fit into a situation in which a panic attack may occur.	3.6 (1)
The system is designed to be aesthetically pleasing.	3.8 (1)
The system is designed to have a proper layout.	3.6 (0.9)
For psychiatrists (n=15)	
The system is designed with a list page to make it easy to find a patient of interest.	4.5 (0.7)
This system has an appropriate response speed when a button is clicked.	4.2 (0.9)
This system provides an appropriate font size for the patient details page.	4.4 (0.9)
This system is designed to facilitate understanding of the patient's training outcomes.	4.4 (0.8)
The system is designed to facilitate the prescribing of services to patients.	4.6 (0.9)
The system is designed with easy-to-navigate layout of the service history.	4 (1.1)
The system makes it easy to understand patient's lifestyle information at a glance.	4.2 (0.7)
This system makes it easy to understand the information the graph represents.	4.6 (0.6)
The system is designed to be aesthetically pleasing.	4.7 (0.6)
The system is satisfactory in terms of overall user experience.	4.5 (0.6)

[a]Responses were scored on a 5-point Likert scale (1: strongly disagree, 2: somewhat disagree, 3: neutral, 4: somewhat agree, and 5: strongly agree). Higher numbers indicate greater agreement with the statement.

[b]CBT: cognitive behavioral therapy.

Analysis

An index of whether participants successfully performed a given execution task was the task success rate, which was calculated for each module based on the investigator's observation data and represents the percentage of participants who completed the task relative to the total number of participants. The convenience and safety scores for each module were calculated as the mean (SD) of the Likert scale scores for the subjective data of the participants. In addition, to determine which content received many suggestions for improvement, the percentage of participants who provided suggestions relative to the total number of participants was calculated for each module and referred to as the feedback rate. The results of the questionnaire about the acceptability of the app and management console were referred to as the acceptability score, which was calculated as the mean (SD) of Likert scale scores for each statement. The scores of convenience, safety, and acceptability were checked to see if they exceeded the set target of 3 or higher for each execution task or each acceptability statement.

Modifications After Usability Testing

In response to the suggestions for improvement from the participants, a list of items that need to be modified to improve functionality and ease of use in the structure and progress of the app or management console was created. The suggestion reflection rate (SRR) was calculated as an indicator of how many of these suggestions were actually reflected in this

XSL•FO
RenderX

modification list, which was the percentage of suggestions that were actually incorporated in the modification list among the total suggestions.

Ethical Considerations

The study protocol was approved by the Yonsei University Gangnam Severance Hospital Institutional Review Board (3-2024-0117). Signed written informed consent was obtained from all participants after a detailed explanation of the study procedure and data privacy and confidentiality by an investigator with good clinical practice training. All data collected and used for analysis were anonymized, and therefore, this study does not contain any personally identifiable images or information. The participation fee was KRW 250,000 (approximately US $210) for patients and KRW 300,000 (approximately US $250) for psychiatrists, which were paid via bank transfer within 1 month of participating in the study.

Results

Execution of the App

Table 2 presents the results of 155 app execution tasks in total for 15 patients with panic disorder, divided by module. For all modules, the patients with panic disorder successfully performed the executive tasks at a high rate of more than 98%. In terms of the convenience score, which assessed how easy it was to execute, they responded that all individual modules were easy to execute, with a score exceeding 3 (an average score of 4.7 to 5). In terms of the safety score, which evaluated how safe it was to execute, they also responded that individual modules were safe to execute, with a score exceeding 3 (an average score ranging from 4.7 to 4.9). Among the modules, “Diaphragmatic breathing” of the training service, “During a panic attack” of the companion service, and “Sleep record” of the care service received many suggestions for improvement with a feedback rate of more than 25%, whereas “Correcting cognitive distortions” and “Interoceptive exposure” of the training service, and “More information” and “Frequently asked questions” of the service assistance received no suggestion for improvement with a feedback rate of 0%.

XSL•FO
RenderX

Table . Task success rates, convenience scores, and safety scores for module-specific execution tasks in the usability evaluation of the app by patients with panic disorder (n=15).

Modules	Number of execution tasks	Task success rate (%)	Convenience score[a], mean (SD)	Safety score[b], mean (SD)	Feedback rate (%)
Running the app	3	100	4.8 (0.4)	4.7 (0.5)	6.7
Training service					
Studying about panic	9	100	4.9 (0.3)	4.9 (0.4)	6.7
Correcting cognitive distortions	14	100	4.9 (0.4)	4.8 (0.5)	0
Diaphragmatic breathing	27	100	4.8 (0.6)	4.8 (0.4)	46.7
Interoceptive exposure	11	99.4	4.8 (0.5)	4.8 (0.4)	0
Mindfulness	21	99.7	4.7 (0.8)	4.8 (0.6)	20
Companion service					
Preventing a panic attack	5	100	4.7 (0.6)	4.8 (0.5)	20
During a panic attack	16	100	4.8 (0.6)	4.8 (0.5)	33.3
Care service					
Medicine record	10	98	4.7 (0.7)	4.7 (0.6)	6.7
Sleep record	8	99.1	4.7 (0.6)	4.7 (0.6)	26.7
Exercise record	11	98.2	4.9 (0.4)	4.8 (0.6)	20
Service assistance					
Reports and feedback	4	100	4.7 (0.7)	4.8 (0.6)	6.7
More information	3	100	4.9 (0.3)	4.9 (0.3)	0
Frequently asked questions	3	100	5 (0.2)	4.9 (0.3)	0

[a]Responses were scored on a 5-point Likert scale (1: very difficult, 2: somewhat difficult, 3: neutral, 4: somewhat easy, and 5: very easy).

[b]Responses were scored on a 5-point Likert scale (1: very dangerous, 2: slightly dangerous, 3: neutral, 4: slightly safe, and 5: very safe).

Execution of the Management Console

Table 3 shows the results of a total of 34 console execution tasks for 15 psychiatrists, divided by module. For all modules, the psychiatrists successfully performed the executive tasks at a high rate of more than 86%. In terms of the convenience score, they responded that all individual modules were easy to execute with a score exceeding 3 (an average score of 4.3 to 4.9). In terms of the safety score, they also responded that all individual modules were safe to execute with a score exceeding 3 (an average score ranging from 4.7 to 4.9). Among the modules, “Managing newly registered patients,” “Prescribing medication,” “Prescription and confirmation for training,” “Checking: Comprehensive calendar,” and “Checking: Achievement rates” received many suggestions for improvement with a feedback rate of more than 25%, whereas “Running the console” and “Checking: Weekly compliance results” received no suggestion for improvement with a feedback rate of 0%.

XSL•FO
RenderX

Table . Task success rates, convenience scores, and safety scores for module-specific execution tasks in the usability evaluation of the management console by psychiatrists (n=15).

Modules	Number of execution tasks	Task success rate (%)	Convenience score[a], mean (SD)	Safety score[b], mean (SD)	Feedback rate (%)
Running the console	2	100	4.8 (0.5)	4.8 (0.6)	0
Managing newly registered patients	4	95	4.6 (0.8)	4.7 (0.6)	60
Prescribing medication	5	98.6	4.8 (0.6)	4.7 (0.7)	46.7
Prescription and confirmation for training	9	97.8	4.7 (0.7)	4.8 (0.4)	33.3
Checking					
Comprehensive calendar	2	96.5	4.3 (1.1)	4.7 (0.6)	40
Achievement rates	6	86.7	4.3 (1)	4.8 (0.4)	53.3
Panic attack–related data	2	87	4.8 (0.4)	4.9 (0.3)	20
Daily survey results	2	100	4.8 (0.5)	4.9 (0.3)	6.7
Weekly compliance results	2	100	4.9 (0.3)	4.9 (0.3)	0

[a]Responses were scored on a 5-point Likert scale (1: very difficult, 2: somewhat difficult, 3: neutral, 4: somewhat easy, and 5: very easy).

[b]Responses were scored on a 5-point Likert scale (1: very dangerous, 2: slightly dangerous, 3: neutral, 4: slightly safe, and 5: very safe).

Acceptability

Table 1 shows the acceptability scores for our system surveyed by the participants. For all 10 different statements that the app was appropriately designed, the patients with panic disorder responded at the level of “somewhat agreeing” with a score exceeding 3. That is, for 5 items, the acceptability scores were slightly higher than 4, and for 5 items, they were slightly lower than 4. For all 10 different statements about the management console being appropriately designed, the psychiatrists also gave the acceptability scores above 3. They were 4.4 or lower, near the level of “somewhat agree” for 5 items, and 4.5 or higher, near the level of “strongly agree” for the remaining 5 items.

Modification List for Upgrading the Functionality

Table 4 presents the number of suggestions for improvement, SRR for system modifications, and examples of the modifications by each module of the app and management console. There were a total of 38 suggestions from the 14 patient participants, almost all of which were reflected in the modification list for upgrading the app functionality, resulting in a 100% SRR in 7 modules. However, 3 suggestions, 1 each for “Mindfulness” of the training service, “Preventing a panic attack” of the companion service, and “Medication record” of the care service, were not reflected for legal or technical reasons, and thus the SRR for these modules did not reach 100%. Meanwhile, all 15 psychiatrist participants made at least 1 suggestion, for a total of 66 suggestions, all of which were reflected in the modification list to upgrade the console functionality, resulting in a 100% SRR across all modules.

XSL•FO
RenderX

Table . Number of improvement suggestions, suggestion reflection rate, and example of system modifications by each module of the app and management console.

Modules	Number	SRR[a] (%)	Examples of modification
App			
Running the app	1	100	Change the member ID from an email to a mobile phone number.
Training service			
Studying about panic	1	100	Add a [NEXT] icon to move on to the next step after answering a quiz.
Correcting cognitive distortions	0	Not applicable	Not applicable
Diaphragmatic breathing	8	100	Improve readability in the explanation of the importance and ease of breathing.
Interoceptive exposure	0	Not applicable	Not applicable
Mindfulness	4	75	Change the options items to be displayed on 1 screen without scrolling.
Companion service			
Preventing a panic attack	7	86	Improve the quality of the screen images in breathing from low to high quality.
During a panic attack	7	100	Change the system so that the written panic symptom information can be used for the input.
Care service			
Medicine record	1	0	Not applicable
Sleep record	7	100	Add an icon to allow the wake-up time to be entered automatically.
Exercise record	1	100	Change the actual exercise time input screen to be more visible through auto-scrolling.
Service assistance			
Reports and feedback	1	100	Improve the readability of the reports in the section that tracks training progress.
More information	0	Not applicable	Not applicable
Frequently asked questions	0	Not applicable	Not applicable
Management console			
Running the console	0	Not applicable	Not applicable
Managing newly registered patients	13	100	Change the method of entering the patient's date of birth from using the calendar to entering it directly.
Prescribing medication	10	100	Add a status table to show medication history at a glance.
Prescription and confirmation for training	12	100	Modify moving between months to be possible via scrolling.
Checking			
Comprehensive calendar	12	100	Revise the colors in the comprehensive calendar to indicate what they mean.

XSL•FO
RenderX

Modules	Number	SRR[a] (%)	Examples of modification
Achievement rates	14	100	Modify the status of goal achievement in ”Care Service” to be checked by week.
Panic attack–related data	4	100	Modify the statistics to be viewed on a daily basis rather than a weekly basis.
Daily survey results	1	100	Modify the positioning of scores on the anxiety and depression graphs to become more visible.
Weekly compliance results	0	Not applicable	Not applicable

[a]SRR: suggestion reflection rate.

Discussion

Principal Findings

In this study, we carried out formative usability testing in patients with panic disorder and psychiatrists to evaluate the practicality, applicability, and embeddedness of the therapeutic app and management console in a clinical setting. Both groups demonstrated high scores in all measures, including task success rate, convenience score, safety score, and acceptability score, confirming the satisfactory usability of the devices. They also offered meaningful qualitative feedback on product improvements, which allowed us to develop a list of modifications to improve users' experience and clinical applicability, and to implement updates to further enhance the devices.

For the therapeutic app, patients successfully completed a total of 155 execution tasks across all modules with a success rate of more than 98%, and each module received an average score of greater than 3 on both convenience (≥4.7) and safety (≥4.7), suggesting that the app was easy and safe to execute and navigate. Regarding statements assessing the acceptability of the app, patients also responded that it was appropriately designed, with a level of “somewhat agree.” Despite these high scores, the app received the lowest score in terms of relevance to panic attack situations and layout appropriateness. Because panic attack situations vary greatly from patient to patient, it is difficult to fully reflect this diversity in the app, and thus, improvements in this area will likely be limited. However, the layout appropriateness is expected to improve significantly if patient feedback is actively incorporated. Overall, patients appear to perceive the app as effectively designed and easy to use. We believe these findings are consistent with previous notions that navigation clarity, modular content structure, and access to real-time support features are important factors determining patient engagement and satisfaction levels [26].

Although patients demonstrated these high task success rates and positive responses, they still identified areas in need of improvement, providing a total of 38 suggestions. Except for a few suggestions that could not be implemented due to legal or technical issues, we incorporated most of them into our modification list. When categorizing these suggestions by module, 4 modules, such as “Diaphragmatic Breathing” in the training service, “Preventing a panic attack” and “During a panic attack” in the companion service, and “Sleep Recording” in the care service, received a relatively high number of suggestions (7 or 8) from patients, which were related to weaknesses in visual guidance and complex multistep interactions. These issues were considered important in developing the modification list, as they could increase cognitive load, especially when used in acutely anxious situations. In contrast, patients gave no suggestions for 2 modules, such as “Correcting Cognitive Distortions” and “Interoceptive Exposure” in the training service, suggesting that cognitively structured and self-paced components in these modules were more straightforward to engage with.

Previous studies have demonstrated that patients who receive a combination of CBT, behavioral interventions, and lifestyle modification achieve better treatment outcomes than those who receive a single modality approach [27,28]. Moreover, patients prefer apps that provide immediate assistance in moments of distress [28] and also require guidance in everyday life, such as reducing caffeine intake or sustaining regular physical activity, which traditional care often fails to provide [29]. Therefore, once the issues summarized in the modification list are addressed, our mobile DTx app, which can provide a variety of treatment modules, may be considered to be a novel treatment option for panic disorder.

For the management console, psychiatrist participants successfully completed a total of 34 execution tasks across all modules with a success rate of more than 86%, and each module received an average score of greater than 3 on both convenience (≥4.3) and safety (≥4.7), suggesting that the management console was easy and safe to execute. Psychiatrists also rated the console's acceptability highly, giving it a mid-range assessment between “somewhat agree” and “strongly agree.” Taken together, they seem to perceive the management console as well-designed and easy to use.

Although psychiatrist participants gave high task success rates and positive evaluation on the management console, they all provided 66 suggestions for usability improvements, which was much more than those from patient participants. Most of these improvement suggestions focused on modules related to new patient management and service prescribing, reflecting clinicians' priorities for establishing control when using digital platforms. They were particularly interested in features that enable them to efficiently manage patients and prescribe

XSL•FO
RenderX

interventions flexibly based on monitoring progress in a clinically meaningful manner.

Since these improvement suggestions did not raise any legal or technical issues for implementation, we included all of them in the modification list, and the production of a management console with improved functionality will follow. We believe that these modifications will help the management console better align with the routine psychiatric care flow. While previous DTx usability testing focused on the patient perspective [30,31], this study incorporated psychiatrists assessing the usability of the management console. Unlike commercial mental health apps that users download without any barriers to access and manage themselves, our system is initiated by a clinician's prescription within the context of ongoing psychiatric care. The requirement for a prescription makes the functionality of the management console crucial, which is why it was included in the current usability testing. This prescription-based model provides not only a medical framework but also therapeutic legitimacy, encouraging patients to perceive the program as a formal part of their treatment, much like taking a prescription medication. The fact that app usage can be monitored by the prescribing clinician can increase patient trust and compliance. Clinicians can also review a patient's app usage profiles to develop a follow-up treatment plan, thereby enhancing the effectiveness and justification of subsequent prescriptions.

Limitations

This study was only the first step in determining how well patients with panic disorder and their clinicians could use our DTx system. This study was conducted at a single site with a small sample size, limiting its generalizability. While participants were representative of the intended user base, the testing environment may not fully capture the complexity of real-world use. Because our system encompasses a significant number of services and functions, such as how many modules will be used by patients in a real-world setting and how engagingly the system will be used between clinicians and patients in clinical settings, could be crucial for usability. However, since this study is insufficient to draw any conclusions on these issues, further real-world research is needed to address them.

Conclusions

The results of this study suggest that Waymed_panic, consisting of a therapeutic app and management console, is sufficiently easy and acceptable to use for the treatment of panic disorder. A distinction of this system from many commercially available apps is the integration of therapeutic content with clinician engagement, enabling not only structured CBT practices, but also real-time coping support, lifestyle tracking, and reciprocating feedback between the clinician and patient. These contents highlight the growing demand for digital resources that integrate treatment into everyday life. With ever-increasing demands and burdens on the mental health system, using integrated DTx models that can extend capacity without compromising quality or personalization might be a way to address the national shortfall of accessible mental health care. To be used in this way, randomized clinical trials are first needed to determine the practical efficacy and safety of this system in treating various symptoms of panic disorder.

Acknowledgments

This research was supported by the Korea Medical Device Development Fund grant funded by the Korean government (Ministry of Science and ICT, Ministry of Trade, Industry and Energy, Ministry of Health and Welfare, and Ministry of Food and Drug Safety; number RS-2022-001-40901), the SmartTech Clinical Research Center (SCRC), funded by the Ministry of Health and Welfare, Republic of Korea (number RS-2023-KH142022), and by Soonchunhyang University. These funders were not involved in the study design, data collection, analysis, interpretation, or writing of the manuscript. The authors did not use any generative artificial intelligence in the manuscript writing process.

Data Availability

The datasets generated and analyzed during this study are not publicly available to protect the privacy of the research participants.

Authors' Contributions

JL, WJ, and HJ collected and analyzed the data. YK interpreted the data. YK, KH, YC, YBS, and CM contributed to the conceptualization and methodology of the study. YK, KK, and JJK applied for funding. YK and JJK drafted the manuscript, which was critically revised by all authors.

Conflicts of Interest

Waycen Inc. contributed to the development of the app and management console, and did not participate in usability testing. KK is the chief executive officer of Waycen Inc. CM is an employee of Waycen Inc.

Multimedia Appendix 1
Core features and functions of the app and management console.
[DOCX File, 24 KB - formative_v9i1e76843_app1.docx]

XSL•FO
RenderX

Multimedia Appendix 2
Execution tasks of the app for patients with panic disorder.
[DOCX File, 27 KB - formative_v9i1e76843_app2.docx]

Multimedia Appendix 3
Execution tasks of the management console for psychiatrists.
[DOCX File, 19 KB - formative_v9i1e76843_app3.docx]

References

1. Asmundson GJG, Taylor S, Smits JAJ. Panic disorder and agoraphobia: an overview and commentary on DSM-5 changes. Depress Anxiety 2014 Jun;31(6):480-486. [doi: 10.1002/da.22277] [Medline: 24865357]
2. Guaiana G, Meader N, Barbui C, et al. Pharmacological treatments in panic disorder in adults: a network meta-analysis. Cochrane Database Syst Rev 2023 Nov 28;11(11):CD012729. [doi: 10.1002/14651858.CD012729.pub3] [Medline: 38014714]
3. van Apeldoorn FJ, Stant AD, van Hout W, Mersch PPA, den Boer JA. Cost-effectiveness of CBT, SSRI, and CBT+SSRI in the treatment for panic disorder. Acta Psychiatr Scand 2014 Apr;129(4):286-295. [doi: 10.1111/acps.12169] [Medline: 23834587]
4. van Apeldoorn FJ, Timmerman ME, Mersch PPA, et al. A randomized trial of cognitive-behavioral therapy or selective serotonin reuptake inhibitor or both combined for panic disorder with or without agoraphobia: treatment results through 1-year follow-up. J Clin Psychiatry 2010 May;71(5):574-586. [doi: 10.4088/JCP.08m04681blu] [Medline: 20492852]
5. Pompoli A, Furukawa TA, Efthimiou O, Imai H, Tajika A, Salanti G. Dismantling cognitive-behaviour therapy for panic disorder: a systematic review and component network meta-analysis. Psychol Med 2018 Sep;48(12):1945-1953. [doi: 10.1017/S0033291717003919] [Medline: 29368665]
6. Papola D, Ostuzzi G, Tedeschi F, et al. CBT treatment delivery formats for panic disorder: a systematic review and network meta-analysis of randomised controlled trials. Psychol Med 2023 Feb;53(3):614-624. [doi: 10.1017/S0033291722003683] [Medline: 37132646]
7. Machado S, Telles G, Magalhaes F, et al. Can regular physical exercise be a treatment for panic disorder? A systematic review. Expert Rev Neurother 2022 Jan;22(1):53-64. [doi: 10.1080/14737175.2021.2005581] [Medline: 35077254]
8. Klevebrant L, Frick A. Effects of caffeine on anxiety and panic attacks in patients with panic disorder: a systematic review and meta-analysis. Gen Hosp Psychiatry 2022;74:22-31. [doi: 10.1016/j.genhosppsych.2021.11.005] [Medline: 34871964]
9. Canan F, Ataoglu A. Panic disorder after the end of chronic alcohol abuse: a report of 2 cases. Prim Care Companion J Clin Psychiatry 2008;10(4):332-333. [doi: 10.4088/pcc.v10n0411d] [Medline: 18787672]
10. Isensee B, Wittchen HU, Stein MB, Höfler M, Lieb R. Smoking increases the risk of panic: findings from a prospective community study. Arch Gen Psychiatry 2003 Jul;60(7):692-700. [doi: 10.1001/archpsyc.60.7.692] [Medline: 12860773]
11. Kaplan A, Mannarino AP, Nickell PV. Evaluating the impact of freespira on panic disorder patients' health outcomes and healthcare costs within the allegheny health network. Appl Psychophysiol Biofeedback 2020 Sep;45(3):175-181. [doi: 10.1007/s10484-020-09465-0] [Medline: 32342249]
12. McGinnis E, O'Leary A, Gurchiek R, Copeland WE, McGinnis R. A digital therapeutic intervention delivering biofeedback for panic attacks (panicmechanic): feasibility and usability study. JMIR Form Res 2022 Feb 3;6(2):e32982. [doi: 10.2196/32982] [Medline: 35113031]
13. Kim K, Hwang H, Bae S, Kim SM, Han DH. The effectiveness of a digital app for reduction of clinical symptoms in individuals with panic disorder: randomized controlled trial. J Med Internet Res 2024 Apr 12;26:e51428. [doi: 10.2196/51428] [Medline: 38608270]
14. Jung HW, Jang KW, Nam S, et al. Digital cognitive behavioral therapy for panic disorder and agoraphobia: a meta-analytic review of clinical components to maximize efficacy. J Clin Med 2025 Mar 6;14(5):1771. [doi: 10.3390/jcm14051771] [Medline: 40095899]
15. Cuyler RN, Katdare R, Thomas S, Telch MJ. Real-world outcomes of an innovative digital therapeutic for treatment of panic disorder and PTSD: a 1,500 patient effectiveness study. Front Digit Health 2022;4:976001. [doi: 10.3389/fdgth.2022.976001] [Medline: 36465089]
16. Katzman MA, Bleau P, Blier P, et al. Canadian clinical practice guidelines for the management of anxiety, posttraumatic stress and obsessive-compulsive disorders. BMC Psychiatry 2014;14 Suppl 1(Suppl 1):S1. [doi: 10.1186/1471-244X-14-S1-S1] [Medline: 25081580]
17. Royal Australian and New Zealand College of Psychiatrists Clinical Practice Guidelines Team for Panic Disorder and Agoraphobia. Australian and New Zealand clinical practice guidelines for the treatment of panic disorder and agoraphobia. Aust N Z J Psychiatry 2003 Dec;37(6):641-656. [doi: 10.1080/j.1440-1614.2003.01254.x] [Medline: 14636376]
18. Bandelow B, Lichte T, Rudolf S, Wiltink J, Beutel ME. The diagnosis of and treatment recommendations for anxiety disorders. Dtsch Arztebl Int 2014 Jul 7;111(27-28):473-480. [doi: 10.3238/arztebl.2014.0473] [Medline: 25138725]

XSL•FO
RenderX

19. Barlow DH. Cognitive-behavioral therapy for panic disorder: current status. J Clin Psychiatry 1997;58 Suppl 2:32-36. [Medline: 9078992]
20. Sarris J, Moylan S, Camfield DA, et al. Complementary medicine, exercise, meditation, diet, and lifestyle modification for anxiety disorders: a review of current evidence. Evid Based Complement Alternat Med 2012;2012:809653. [doi: 10.1155/2012/809653] [Medline: 22969831]
21. Helbig-Lang S, Lang T, Petermann F, Hoyer J. Anticipatory anxiety as a function of panic attacks and panic-related self-efficacy: an ambulatory assessment study in panic disorder. Behav Cogn Psychother 2012 Oct;40(5):590-604. [doi: 10.1017/S1352465812000057] [Medline: 22373714]
22. Lambert R. Lifestyle behaviours add to the armoury of treatment options for panic disorder: an evidence-based reasoning. Int J Environ Res Public Health 2015 Jun 18;12(6):7017-7043. [doi: 10.3390/ijerph120607017] [Medline: 26095868]
23. Nielsen J. Heuristic evaluation. In: Nielsen J, Mack RL, editors. Usability Inspection Methods: John Wiley & Sons; 1994:25-62. [doi: 10.5555/189200.189209]
24. Brooke J. SUS: a “quick and dirty” usability scale. In: Jordan PW, Thomas B, Weerdmeester BA, McClelland AL, editors. Usability Evaluation in Industry: Taylor and Francis; 1996:189-194. [doi: 10.1201/9781498710411]
25. Lewis JR. Psychometric evaluation of the post-study system usability questionnaire: the PSSUQ. Presented at: Proceedings of the Human Factors Society Annual Meeting; Oct 12-16, 1992; Atlanta, Georgia. [doi: 10.1177/154193129203601617]
26. Chen MH, Tsai SJ. Treatment-resistant panic disorder: clinical significance, concept and management. Prog Neuropsychopharmacol Biol Psychiatry 2016 Oct 3;70:219-226. [doi: 10.1016/j.pnpbp.2016.02.001] [Medline: 26850787]
27. Heldt E, Gus Manfro G, Kipper L, Blaya C, Isolan L, Otto MW. One-year follow-up of pharmacotherapy-resistant patients with panic disorder treated with cognitive-behavior therapy: outcome and predictors of remission. Behav Res Ther 2006 May;44(5):657-665. [doi: 10.1016/j.brat.2005.05.003] [Medline: 16038874]
28. Pull CB, Damsa C. Pharmacotherapy of panic disorder. Neuropsychiatr Dis Treat 2008 Aug;4(4):779-795. [doi: 10.2147/ndt.s1224] [Medline: 19043522]
29. Austin D, Blashki G, Barton D, Klein B. Managing panic disorder in general practice. Aust Fam Physician 2005 Jul;34(7):563-571. [Medline: 15999167]
30. Lee J, Kawai T, Yoshida N, et al. A usability assessment on a virtual reality system for panic disorder treatment. Presented at: Electronic Imaging 2008; Jan 28 to Feb 1, 2008; San Jose, CA. [doi: 10.1117/12.766461]
31. Ebenfeld L, Kleine Stegemann S, Lehr D, et al. A mobile application for panic disorder and agoraphobia: insights from a multi-methods feasibility study. Internet Interv 2020 Mar;19:100296. [doi: 10.1016/j.invent.2019.100296] [Medline: 31890640]

Abbreviations

CBT: cognitive behavioral therapy
***DSM-V*:** *The Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition*
DTx: digital therapeutics
FDA: US Food and Drug Administration
SRR: suggestion reflection rate

Edited by A Stone, A Mavragani; submitted 02.05.25; peer-reviewed by J Brewer, ZA Kindi; revised version received 04.09.25; accepted 11.09.25; published 17.10.25.

Please cite as:
Ko Y, Lee J, Ham K, Cho Y, Shin YB, Min C, Kim K, Jang W, Jung H, Kim JJ
Assessing the Usability of a Prescription-Based Mobile App for Patients With Panic Disorder and Its Management Console for Clinicians: Controlled User Study
JMIR Form Res 2025;9:e76843
URL: https://formative.jmir.org/2025/1/e76843
doi:10.2196/76843

© Yujin Ko, Jeemin Lee, Kyunghee Ham, Yesol Cho, Yu-Bin Shin, Choongki Min, Kyungnam Kim, Wonseuk Jang, Hayeon Jung, Jae-Jin Kim. Originally published in JMIR Formative Research (https://formative.jmir.org), 17.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Designing Culturally Adapted Digital Mental Health Support Tool for Chinese-Speaking International Students in Australia: A Qualitative Co-design Study

Ling Wu, PhD; Chen Zhu, PhD; Joshua Paolo Seguin, MBIS; Jue Xie, PhD; Pranav Kulkarni, PhD; Mingye Li, MIT; Patrick Olivier, PhD

Action Lab, Department of Human-Centred Computing, Faculty of IT, Monash University, 14 Rainforest Walk, Clayton, Australia

Corresponding Author:
Ling Wu, PhD
Action Lab, Department of Human-Centred Computing, Faculty of IT, Monash University, 14 Rainforest Walk, Clayton, Australia

Abstract

Background: International students, especially Chinese-speaking students, who grew up with collectivist values and the linguistic and cultural characteristics face a higher risk of mental health issues due to the challenges of geographical, linguistic, and cultural transitions when studying abroad. While digital technology has shown promise in supporting mental health, few studies have focused on designing tools specifically for Chinese-speaking international students and the challenges they face.

Objective: This study aimed to design and develop a self-directed digital mental health support tool that provides culturally safe and appropriate support for international students.

Methods: A co-design approach was used across 2 study phases, Phase 1 (interviews) and Phase 2 (co-design workshops), to explore the design implications of the digital tool. Inductive thematic analysis was conducted to extract design insights and considerations.

Results: Findings show that participants faced a wide range of challenges arising from cross-cultural, academic, and daily life demands, which further contributed to increased levels of stress and negative feelings. These findings highlight a need to improve mental health awareness, literacy, and help-seeking intentions in this vulnerable group. These insights informed the design that emphasized the integration of culturally adapted resources with self-directed learning tools.

Conclusions: Based on the findings, a personalized, self-directed, and culturally adapted design has been proposed that creates a bridging pathway linking students' immediate challenges with mental health education and support. This design offers a clear set of implications for enhancing international students' mental health awareness, literacy, and help-seeking behaviors, thereby providing essential support for this population.

(JMIR Form Res 2025;9:e76695) doi:10.2196/76695

KEYWORDS

mental health digital tool; international students mental health; culturally-adapted resources; self-directed tool; co-design digital tool

Introduction

Background

Over the past 2 decades, international student mobility has surged, with over 6.8 million international students worldwide as of 2022 [1] and Australia emerging as a popular destination, with more than 810,960 international students recorded by May 2024 [2]. Chinese-speaking students account for roughly 20%-30% of these enrollments, and many face significant challenges in adapting to the new cultural and academic environments, accompanied by varying levels of mental health issues. Specifically, Chinese-speaking international students have reported a higher risk of experiencing mental health issues, such as depression, anxiety, and stress compared to Australian domestic students [3,4].

A wide range of factors, including language barriers, communication challenges, and social difficulties—often influenced by cultural differences, particularly the collectivist values of Chinese-speaking international students and the language barriers—significantly contribute to the development of mental health issues [4-6]. These factors, coexisting with daily living challenges [7,8], low levels of mental health awareness and literacy [9], and exposure to mental health stigmatization rooted in collectivist values and filial piety in China [10-12], further exacerbate the limited attempts of help-seeking and the underuse of mental health services in Australia.

XSL•FO
RenderX

While universities are typically the first point of contact for international students facing mental health issues, a study revealed that their mental health policies, strategies, and educational initiatives for international students are either lacking or framed as largely a responsibility of the student [13]. This creates significant challenges, especially for high school and undergraduate students who are at a developmental stage (ie, late adolescence, typically of age 15 - 19 years), making them more vulnerable to mental health issues [14].

Existing Digital Tools

Several digital mental health resources in Australia offer valuable support for university students [15-18], particularly those international students from Chinese-speaking backgrounds. The MindYourHead platform (University of Sydney) delivers bilingual self-assessments on stress, sleep, procrastination, and social support, using culturally resonant terms, such as "beng kui" (ie, mental breakdown) to enhance user engagement [19]. While this linguistic localization strengthens accessibility, the tool has tended to focus on emotional and sleep-related concerns and does not address broader, multifaceted stressors, such as academic workload, visa uncertainty, financial pressures, and social transitions. These are the major areas often noted by students as affecting their well-being. High dropout rates also suggest that further engagement with the users may be needed to maintain sustained use.

Meanwhile, This Way Up, a collaboration between St Vincent's Hospital Sydney and UNSW, delivers structured, self-paced internet-based cognitive behavioral therapy courses for conditions like anxiety and depression [20]. The platform has been rigorously evaluated and shows strong outcomes across general populations; however, some challenges remain. As with many self-guided programs, attrition rates can be high, and opportunities for personalization are limited, with content delivered in a fixed, workbook-style format. Although multilingual resources, including simplified and traditional Chinese, have recently been introduced, these adaptations are broad and not specifically designed to address the cultural, academic, and social challenges experienced by Chinese-speaking international students.

Moreover, the Chinese Online Clinic, developed by the Black Dog Institute, offers a free and anonymous mental health self-assessment fully translated into Mandarin, providing evidence-based personalized feedback and resources [21]. This tool reduces language barriers and increases accessibility for Chinese-speaking users, though it primarily targets the general population and does not explicitly address the specific academic, multicultural, and transition-related contexts of university students. In a broader context, systematic reviews of digital mental health interventions for university populations have highlighted persistent gaps. While these tools show potential to improve access and deliver evidence-based support, many are characterized by high dropout rates, limited personalization, low engagement, and user concerns about credibility and interface quality [15-18]. Together, these findings suggest that despite meaningful progress, current digital interventions often fall short of addressing the specific cultural and developmental needs of Chinese-speaking international students.

Study Aims

To address these limitations, this study engaged Chinese-speaking international students in Australia in the co-design of a digital mental health support tool. The overarching aim was to create a free, culturally safe, and self-directed platform that connects students' everyday challenges with mental health literacy and support options. Phase 1 focused on developing the tool's content through qualitative analysis of students' lived experiences, while Phase 2 refined both the content and the interface based on user testing of an initial prototype. By grounding the design in students' own perspectives and iteratively improving usability, this study represents the first steps toward a tailored intervention that can enhance awareness, promote safer coping strategies, and facilitate pathways to professional help when needed.

Methods

Recruitment

The current paper covers Phase 1 and Phase 2 of a broader study on the design, development, implementation, and evaluation of a digital mental health support tool for Chinese-speaking international students. The study follows the Double Diamond design framework [22], with Phase 1 and Phase 2 representing the Discover and Define stages, respectively. Specifically, Phase 1 aims to discover existing mental health needs (eg, academic, everyday life, and social activities–related challenges experienced by Chinese-speaking international students in Australia) through semistructured interviews and use the insights extracted to design a prototype. Phase 2 aims to further define the key aspects of the intervention (content focused) and key functionalities (feature focused) of the platform relevant to international students that shaped the final design of the prototype.

Participants in Phase 1 were recruited between August and October 2023 through internet-based advertisements targeting Chinese-speaking international students who were interested in addressing their mental health concerns while studying in Australia and contributing to the development of a digital tool. The target population was defined as current international students holding a valid student visa, which inherently excluded Australian citizens, permanent residents, and holders of other visa types (eg, working visas). Potential participants were screened using the information provided in the expression of interest sign-up form according to the following inclusion criteria: (1) current international students enrolled in a higher education degree in Australia, (2) studying onshore, and (4) native speakers of Chinese. Individuals were excluded if they were less than 18 years of age or were experiencing a mental health crisis (ie, self-reported acute risk or severe psychiatric symptoms requiring urgent assessment or care that could compromise informed consent or safe participation) at screening. Given that the intended users of the platform are international students who may present with varying levels of mental health needs and mental health literacy, no additional exclusion criteria beyond those specified were applied. The Depression, Anxiety and Stress Scale–21 items (DASS-21), embedded in the prototype, provided a standardized means of assessing

XSL•FO
RenderX

psychological distress, offering additional context for interpreting participant responses. Information on participants' length of stay in Australia was also collected to help situate their experiences within different stages of adaptation to the host country.

Eligible participants were contacted via email after filling out the expression of interest form voluntarily. In total, 20 participants took part in individual semistructured interviews during Phase 1. Phase 2 involved a total of 7 participants, allocated into 2 separate design workshops based on individual preferences and availability. Participants were drawn from Phase 1 individuals who expressed interest in further involvement and from internet-based advertisements targeting Chinese-speaking international students interested in testing the prototype. The same inclusion and exclusion criteria were applied throughout.

Materials

During the semistructured individual interviews of Phase 1, a flexible set of open-ended guiding questions was used to explore the mental health needs and concerns of Chinese-speaking international students. While the core questions addressed areas, such as daily life challenges related to academics, living conditions, physical health, and emotional well-being, the sequence and specific follow-up questions varied depending on participants' responses. This approach allowed the interviewer to probe deeper into emerging topics and clarify issues as they arose. Additional discussion points included how participants addressed these challenges, their help-seeking attitudes, and self-management capabilities. Participants were also invited to share their understanding of mental health issues, strategies for self-management, and preferred avenues for professional assistance in order to identify potential stigma and levels of mental health literacy. Finally, participants were asked about the digital platforms they commonly used for mental health information to inform the potential needs and features of a future digital solution. The interview guide is provided in the Multimedia Appendix 1.

In Phase 2, two separate design workshops were conducted in a group format, with 3 participants in workshop 1 and 4 in workshop 2. Before the workshop, participants were invited to interact with the prototype for 1 week at their own pace. They were encouraged to take notes. During the workshop, group activities were facilitated using a Miro board (RealtimeBoard, Inc [23]) to elicit participants' feedback on the user journey of the digital support tool prototype. The design activities included a journey map [24], with a goal to understand the barriers and enablers at each touchpoint of the user journey. The activities focused on three key aspects: (1) the design of the mental health check-in, including surveys and questionnaires; (2) the system's recommended content based on the survey and questionnaire results; and (3) the interface and functionalities. All participants opened the Miro board page on their own devices and acted as users to complete the collaborative activities. Engagement features, such as emoji reactions, thumbs-up indicators, and polls, were used to maintain engagement throughout the videoconferencing (Zoom). The workshop guide is provided in the Multimedia Appendix 2, where an example screenshot of the Miro board activity is also included.

The user flow of the digital support tool prototype is shown in Figure 1. Specifically, a check-in survey was embedded to explore the challenges faced by participants, followed by the short-form version of the DASS-21 [25] to assess their level of depression, anxiety and stress. As guided by the literature and findings from Phase 1, mental health issues among international students have been linked to a broad range of challenges, including those related to study, accommodation, language, social connections, finances, and lifestyle-related health factors. These domains informed the design of the check-in survey and were used to generate individualized feedback, ensuring that each user received tailored support addressing their specific needs. The individualized feedback also incorporated participants' scores on each DASS-21 subscale, along with evidence-based strategies and resources, such as internet-based and local services providing mental health support in Chinese. Participants were able to download their complete feedback and leave any additional comments on the feedback page. They could also log back into the platform at any time to review their feedback.

The prototype was developed using TypeScript (Microsoft Corporation) and Gatsby.js (Gatsby, Inc) to deliver a fast and responsive user interface. Firebase (Google LLC) was used for authentication, database management, and hosting, while SendGrid (Twilio Inc.) was used for automated email communication. Zapier (Zapier Inc) facilitated workflow automation between services. The administrative interface was implemented with Next.js (Zapier Inc) to enable efficient management of content and user data. The interface design encompassed layout, color scheme, and other visual elements, with representative screenshots provided in the Multimedia Appendix 3.

XSL•FO
RenderX

Figure 1. User flow diagram for the digital mental health support tool prototype.

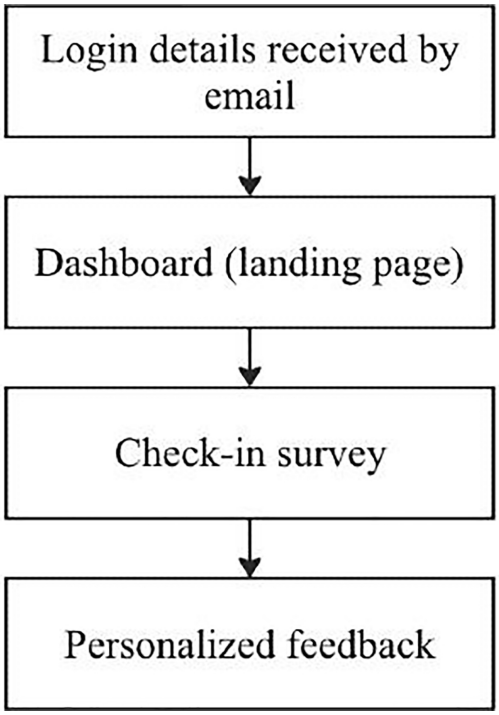

DASS-21 [25] is a self-report scale designed to measure the negative emotional states of depression, anxiety, and stress through 3 subscales. Each statement is rated on a 4-point Likert scale ranging from 0 ("Did not apply to me at all") to 3 ("Applied to me very much, or most of the time"), indicating how much the statement applied to the respondent over the past week. The score for each subscale is calculated by summing the item scores and then multiplying the total by 2 to compare with established severity cutoff scores. A higher score indicates a higher level of the corresponding emotional state. Both the English and Chinese versions of the DASS-21 have demonstrated good psychometric properties in university students [26,27].

Procedure

During registration for both Phase 1 and Phase 2, participants electronically signed a consent form followed by a brief demographic survey (covering age, educational background, gender, and duration in Australia), via REDCap (Research Electronic Data Capture; Vanderbilt University) [28,29] hosted at Monash University. REDCap is a secure, web-based, and confidential platform for the collection of research data, allowing flexible data manipulations and export procedures. Eligible participants were subsequently contacted via email to determine their preference for attending interviews or workshops. All of them opted for the videoconferencing (Zoom) format. The interviews lasted approximately 60 minutes, while the workshops ran for around 90 minutes. All sessions were conducted via video conference and recorded, with transcripts prepared and translated into English by trusted third-party services prior to analysis.

Analysis

The data analysis for the semistructured interviews and design workshops was conducted in NVivo (Lumivero [30]) using inductive thematic analysis. Two of the authors (LW and CZ) performed inductive thematic analysis independently, following the 6-step analysis pipeline described by Braun and Clarke [31]. An inductive process was adopted in the development of themes. Descriptive statements were made for each code, and themes were derived from the descriptive statements. This process was cross-checked and discussed by other members of the research team before the results were finalized.

Ethical Considerations

This study (Phases 1 and 2) was approved by the Monash University Human Research Ethics Committee (Approved project ID 39583). The approved protocol covered recruitment materials, consent procedures, the conduct of videoconference-based interviews and design workshops, and the use of third-party services for transcription and translation, as specified in the ethics application.

Informed consent was obtained electronically via REDCap (Vanderbilt University, Nashville, TN, United States) before participation. Participants were informed about the study purpose, procedures, data handling, and their right to withdraw at any time without penalty.

All data collected were downloaded and stored with password protection on Monash University–managed secure servers, and access was restricted to investigators named in the ethics protocol. Data backups were maintained on network drives managed by Monash eSolutions. Identifiable information was removed from transcripts prior to analysis to ensure privacy and confidentiality.

Compensation was provided in the form of an AUD $20 (US $12.98) electronic gift card per hour of participation, as outlined in the approved ethics protocol.

XSL•FO
RenderX

Results

Phase 1: Interviews

Overview

For Phase 1, a total of 23 participants registered their interests, and 20 were included in the interviews and analysis. Three participants withdrew from the interview process. Demographic information was displayed in Table 1. From inductive thematic analysis, 3 themes and 27 subthemes were derived from 152 codes. Detailed descriptions and reference frequency of themes and codes are shown in Table 2.

Table . Demographic information of participants involved in Phase 1 (N=20).

Demographic variable	Participants
Age (years), mean (SD)	19; 23.79 (2.44)
Sex, n	
Female	13
Male	7
Nonbinary	0
Education, n	
Bachelor's degree or diploma	9
Master's degree	7
Doctoral degree	4
Study areas, n	
Commerce	4
Public health	8
Computer science	4
Social science	3
Years in Australia, n	
<1	3
1-3	6
3-6	8
>6	2

XSL•FO
RenderX

Table . Summary table of themes and subthemes derived from semistructured interviews.

Themes and subthemes	Code frequency (%)
Student life challenges	207 (27.02)
Coursework challenges	65 (8.22)
Perceived heavy coursework load and difficult learning content or class activity lead to negative mental health states.	40 (5.22)
Comparison with peers and uncertainty for the future lead to stress.	16 (2.09)
Academic transition challenges (eg, different study mode or paradigm)	7 (0.91)
Cross-cultural challenges	59 (7.70)
Different values, beliefs, and habits lead to perceived discrimination, misunderstanding, and negative impact on mental health.	16 (2.09)
Language barriers lead to poor academic performance, in-class interactions, and negative mental states.	43 (5.61)
Daily life challenges	85 (11.10)
Developmental stage–related challenges (eg, struggles related to self-efficacy and autonomy, self-regulation or management skills, and facing judgment).	23 (3.00)
Logistical challenges (eg, transportation and accommodation) lead to negative mental states.	29 (3.79)
Socializing difficulties lead to boredom and loneliness, thus affecting mental health.	33 (4.31)
Lived experience of mental health issues	221 (28.85)
Contextually shaped attitudes toward mental health	114 (14.88)
Attitude toward mental health is more open and positive in the host country (ie, Australia), which encourages help-seeking.	35 (4.57)
Attitude toward mental health is less open and negative in the home country (ie, China), which discourages help-seeking.	70 (9.14)
Personal view of mental health is mixed but has changed to be more positive since arriving in Australia.	9 (1.17)
Common mental health issues	66 (8.62)
EMO[a], anxiety, depression, and stress are common.	52 (6.79)
Mental health issues can have significant influences on daily life.	14 (1.83)
Mental health resources	41 (5.35)
Interested in exploring self-management and help-seeking options	17 (2.22)
Limited resources to learn about mental health	24 (3.13)
Coping and managing strategies	338 (44.13)
Concerns and perceptions about clinical services	127 (16.58)
Limited help-seeking intentions for professional help	88 (11.49)
Limited knowledge related to local health care systems	19 (2.48)
Worries and negative perceptions about clinicians or services can discourage help-seeking.	20 (2.61)
Common practice and different views on seeking support from peers (preferably) and parents (limited attempts and varies across individuals)	63 (8.22)
Perception and experiences of online resources	52 (6.79)
Negative perception of (eg, not trusting) online resources highlights the need for reliable resources.	29 (3.79)
Positive perceptions of online resources (eg, immediacy, convenience, and efficiency)	23 (3.00)

XSL•FO
RenderX

Themes and subthemes	Code frequency (%)
Some lifestyle strategies or activities reflect maladaptive coping.	47 (6.14)
Some self-management strategies or tools are inappropriate, reflecting poor mental health literacy.	49 (6.40)

[a]EMO refers to the equivalent of mood swings in English, encompassing sudden feelings of sadness (often occurring at night) or emotional responses during specific days or times, sometimes accompanied by anxious and depressive thoughts and feelings.

Student Life Challenges

The findings reveal that student life challenges arising from geographical and academic transitions are highly impactful for Chinese-speaking international students, who reported heightened stress levels and a noticeable decline in mental health. In addition to the acute multilevel transition (ie, location, culture, academic expectations and workloads, and social connections), these students also enter a level of independence that they might not have been sufficiently prepared for. As articulated in previous literature, high-level parental involvement in children's and emerging adults' lives is a common cultural phenomenon among Chinese parents [32] as a result of heavy influence from Confucian values and a collectivist cultural orientation [33]. With key emphasis on parental control and training (guan)—obedience, discipline, and respect; focus on academic achievement, a highly protective approach, and high parental involvement (Sun and Mulvaney [32]; Wu et al [33]). Therefore, transitioning from a familiar, high-support environment to an uncertain environment with close to no parental support can be highly challenging.

The most impactful and prevalent challenge faced by Chinese-speaking international students is cultural transition, including acute cultural adaptation issues, language barriers, and identity conflicts. These challenges can happen in various forms of lived experiences, such as perceived discrimination resulting from different cultural expectations, communication difficulties, and perceived social exclusion and isolation. These experiences tend to lead to feelings of isolation and a sense of "I don't belong here" (Participant 8; P8). In addition, students tend to perceive that they have inadequate proficiency in English that affects their academic performance and interpersonal interactions in social and studying settings. The consequence of this perceived inadequacy, together with the cultural differences, can lead to self-dissatisfaction, stress, and frustration, with no clear solutions to address these issues.

> *As a non-native English speaking international student, you feel a bit discomfort, a bit difficult to fit in...for example, when they tell jokes, you completely don't get their laughing points, and you just can't really fit in their culture...Then, you feel a sense of being marginalized.* [P12]

> *I feel very frustrated because my spoken language is really poor... I'm just very dissatisfied with myself, and in such cases, there is a lot of psychological pressure.* [P6]

Participants identified heavy study workload as a major challenge faced by international students. The language and cultural barriers often exacerbate these issues, making it significantly difficult to manage and cope with workload, with 1 participant describing the situation as:

> *I might have to put in many times more effort than they [native English-speaking students] do to cover the gap [caused by language barriers].* [P12]

They articulated many other difficulties in managing and coping with the demands of participating in group discussions, adapting to new learning environments, passing subjects, completing assignments, and understanding class content. They have also identified some perceived psychological impacts. For example, they can feel behind with their academic performance, resulting from peer pressure and constant comparison with their peers (P2), which further increased their anxiety about future educational and career plans. Other participants shared that a sense of uncertainty often led to negative feelings, such as confusion, stress, self-doubt, and anxiety.

> *After starting university, I feel extremely stressed. The teaching model here, although the tutorials are only two hours, it actually requires more than five hours of preparation and review outside the class time...I often feel on the verge of breaking down and regret choosing this major.* [P21]

> *There would be feelings of inferiority. It might include comparisons with peers. It could be comparisons of abilities, academic performances, and maybe even appearances, possibly. Family backgrounds, financial conditions, all sorts of comparisons stemming from peer pressure.* [P2]

The most commonly shared daily life challenge is difficulty with socializing. Common problems include worries about, or challenges in, establishing intimate relationships, making new friends, or formulating a social network. Alongside academic challenges, these social and interpersonal issues are seen as significant sources of stress, which can further exacerbate feelings of loneliness and isolation, significantly impacting their well-being. Challenges related to logistical life management are another primary cause of negative feelings, as many of the participants did not have adequate prior experience living independently in their home country to prepare themselves. These challenges vary and may include financial burdens, lengthy transportation, homesickness, accommodation issues, and unfamiliarity with social rules or systems. Specific problems, such as living in a remote area due to cost of living concerns (P8), sleep deprivation due to accommodation issues (P21), inefficient transportation systems (P19), and not being able to understand and meet cultural expectations (P16), can lead to feelings of stress, anxiety, and depression. These difficulties can be particularly challenging for new international students due to the drastically different living arrangements

XSL•FO
RenderX

between their home country and the host country, especially with no immediate support or solutions available to address them.

> *I often feel very lonely, especially since I feel like many students here might not be very eager to make friends, I think it is a kind of pressure to make new friends.* [P8]

> *The noise problem where I lived had a significant impact on my life. I didn't know how to communicate about it, and I really wished someone had guided me on how to handle the situation. My parents [who were in China] didn't know what to do, and neither did I, as I was only 15 years old at the time.* [P21]

Perceptions and Lived Experience of Mental Health Issues

Participants expressed varied views and attitudes toward mental health issues from their home and host cultures. They described that they see a more open attitude toward mental health and that mental health is more normalized in their host country. Their perceived attitudes toward mental health were that it is "common in daily life" (P1), and people are "open to talking with others" (P16) and "having access to more professional mental health services" (P18). In contrast, participants' perceptions of attitudes toward mental health in their home country were that there was stigma around it. These views are consistent with the literature demonstrating that collectivist cultural values (including Confucian influence) inform stigmatizing beliefs about mental illness [34,35]. Key stigma types are public stigma, where fear of damaging family reputation and disrupting social harmony leads to the public viewing mental illness as a moral weakness [35], and self-stigma, where individuals internalize culturally informed stigma beliefs, resulting in low self-esteem and shame [34]. Consistent with these studies, participants in our study perceived that this attitude led to "mental health issues are not being taken seriously" (P9), people "feeling shame over having mental health issues" (P14), and people being "unlikely to seek or talk about support with others" (P18) and more likely to be "at risk of facing discrimination" (P16). In general, the perception is that mental health, especially issues and disorders, is something that people tend to keep to themselves, sometimes not even shared within their family.

> *Under the influence of their parents, they might feel that it is shameful or difficult to speak about...but I've found that many Chinese people are willing to share about their mental health issues once they are abroad. However, if they are in China, speaking out might lead to discrimination or being viewed differently. Therefore, they might be afraid to express themselves, and they also might not dare to talk to their parents about it, or if they do, they might receive a negative response.* [P12]

When asked about their own experience of mental health, most participants stated that they think certain signs and symptoms related to mental health are prevalent among Chinese-speaking international students, including themselves. The most frequently reported issue (16 out of 20 participants) the participants articulated was "emo" the equivalent of mood swings in English, including sudden feelings of sadness (often happening at night) or emotional responses during specific days or times, some followed by anxious and depressive thoughts and feelings. These experiences were often attributed to homesickness (P16) and stress related to assignments and exams (P22), as well as social isolation (P18) and sometimes uncertainty associated with their future plans. For some participants, these temporary negative moods impacted their daily functioning, resulting in symptoms, such as insomnia and reduced appetite (P14) or a desire to be alone (P4).

> *I think it might last for a long time. The feeling of being "emo" is like having small bouts of negativity, where life seems to lack hope… Life feels pretty dull, with no motivation… and this feeling can actually last for quite a while. Sometimes, it might even last the whole winter.* [P23]

When sharing their perceptions and experiences of mental health, participants voluntarily mentioned a variety of sources to help understand their mental health–related experiences, including books, friends, university coursework, and informational sheets. A few participants also indicated a need to acquire this knowledge but were not sure about the ideal sources. Among the types of knowledge they were interested in, self-management strategies were most commonly mentioned (10 out of 20). Participants were keen to learn how to manage negative moods, create study plans, and obtain personalized support. A small number of participants expressed interest in finding systematic guidelines because they had not interacted with a professional before.

> *Topics related to psychology, like mental health and how to manage emotions, are the ones I am interested in. I've always been very interested in how a person can maintain a positive attitude continuously.* [P23]

> *Because I haven't seen a psychologist myself, most of my knowledge comes from what I've heard from others. I might not be clear on some of the details. If there were a systematic guideline, I believe it would be more helpful.* [P1]

Coping and Managing Strategies

The overall approach to coping and managing their mental health articulated by the participants is self-directed, with some willing to use their family and social networks and a very small number of participants open to professional help, whereas the majority are resistant. This approach is again consistent and grounded in the Chinese cultural beliefs. With the feeling of shame and guilt resulting from not meeting culturally accepted expectations and behaviors, individuals tend to source self-directed and confidential means of mental health support [36]. With a family-oriented health system [37], help from the family and close relatives and friends may be possible for some, while others may experience heightened stigma and isolation due to not wanting to burden or bring stigma toward the family.

Participants discussed several strategies they use to manage their challenging feelings and moods, and the use of online resources is reported by all participants. They frequently turn to social media platforms and websites for assistance. For the

XSL•FO
RenderX

students, these online platforms provide immediate and helpful information in allowing them to learn from others' experiences, identify people with similar issues, better understand their own symptoms, and explore topics of interest. While participants found social media platforms, alongside other information sources, helpful to some degrees, they also articulated a concern that it is hard to judge if the information is reliable and trustworthy. Consequently, some participants expressed that if possible, they would use university-authorized online platforms, as they perceive a higher likelihood of trustworthy information there.

> *Because you don't understand the symptoms, you definitely need to observe yourself first. We don't know what these symptoms could indicate—what disease or behavior. But you can start by searching online to see if there are similar symptoms. That's how you might begin to understand what the issue could be. Then, you can further your understanding by listening to discussions on platforms like RedNote [a Chinese-based social media platform that enables users to share personal posts, like Instagram], reading comments, or learning about similar experiences and how others have resolved their issues.* [P7]

> *Because it lacks a certain level of professionalism, most of it is just individuals sharing their personal experiences, so it might not be very prominent in terms of professionalism. It's really just sharing personal experiences. Yes, I understand that as an individual, I'm still looking for a more authoritative academic resource to learn about psychological knowledge.* [P1]

Other than online resources, participants also explored various offline management options. The majority of the strategies used are mostly of a self-directed or self-organized nature. The most frequently used strategies included self-regulated strategies, engaging in conversations with friends and family, adopting lifestyle strategies, and joining entertaining or outdoor activities. Self-regulation strategies included (1) "just accept it" (P8); (2) "befriending negative moods" (P5); (3) better time management or listening to lecture recordings when negative moods were linked to academic pressures; (4) improving language skills to overcome language-related stress; and (5) using mental health techniques, such as shifting attention, practicing mindfulness, or meditation. A few participants mentioned that being alone and processing emotions internally is their preferred way to cope. In relation to lifestyle adjustment, participants listed various lifestyle changes and activities they use to manage their negative moods. These strategies cover "getting more sleep" (P4), "engaging in physical activities" (P3), "writing in a diary" (P16), "watching videos" (P17), "hanging out with friends" (P7), and "eating favorite foods" (P11). However, some participants also mentioned coping strategies, including playing video games and drinking alcohol.

> *Like drinking and playing video games, I think these are ways of escaping from reality.* [P8]

Interestingly, participants' views on seeking social support within their circle are divided. Some participants felt that discussing their feelings with parents would not be appropriate or helpful, especially since many parents were located in China and may not fully understand their experiences. Moreover, a desire to be independent, to avoid worrying others, and not wanting to burden others were cited as reasons why some participants preferred not to discuss their issues. In contrast, many participants articulated that talking with friends (P12; P23; P5) or parents (P14; P22; P3) could be beneficial, as it provides opportunities to express and release negative emotions, feel understood, receive support, and relax, particularly when peers are experiencing similar issues related to academic stress. These findings suggest that, if available, students tend to primarily seek help from within their social circles. When this social circle is not available or not trusted by the participants, they tend to turn to a self-support approach where they manage things alone.

> *I feel like they (parents) don't understand a lot of things. Especially about life abroad—if you tell them what happens overseas, it's completely unhelpful as they are utterly unaware and know nothing about the various situations and life conditions abroad. Telling them, I feel, has no effect at all.* [P23]

> *I won't tell others everything, but the most likely time I might share these feelings is at night. At night, when I really feel overwhelmed and can't bear it, I might think about talking to others. However, I don't know why, perhaps it's my independent and strong-willed nature that prevents me from actually doing so, and I end up preferring to endure it on my own.* [P18]

While a few participants expressed interest in seeking professional support, most shared resistance and concerns toward taking action on seeking professional help. The most common concerns include language and cultural barriers and worries about costs, followed by appointment availability, the cultural and professional background of the clinicians, and the location of the services. Moreover, many participants indicated that they would seek help from clinical services only if they experienced obvious mental health issues. They (11 out of 20) also noted that they have very limited knowledge about the Australian health care system, international student health insurance policies, and accessing general practitioners. There was also the perception that help-seeking is more applicable when someone has severe mental health problems, and a few of them expressed concerns about accessing these services, indicating distrust and resistance toward clinicians and health care services.

> *Some of the available counselors speak only English, and I would prefer to speak with a Chinese person because there are many specific terms that are unique to our culture. Right now, I can't think of any specific examples, but terms like Neihao [internal friction, self-sabotage, or mental and emotional exhaustion caused by internal conflicts or overthinking], wuligan [a feeling of powerlessness or helplessness], I wouldn't even know how to explain these to them.*

XSL•FO
RenderX

> *That's why I would prefer to find a counselor with a similar cultural background.* [P15]
>
> *So I feel that my symptoms are not severe enough to seek this treatment. I think it's only when the illness is at its worst that I would consider it.* [P7]
>
> *But my concern is that, after I speak up, I worry about the psychologist's opinion of me because the things that trouble me are quite private. I'm also worried about the judgments from others. Even though I know that psychologists are professionally trained not to judge their clients, I still feel a bit anxious about this.* [P21]

Phase 1 Design Insights and Strategies

The Phase 1 interviews demonstrated that Chinese-speaking international students face substantial academic, cultural, and daily life challenges that often give rise to stress, anxiety, self-dissatisfaction, and withdrawal. These difficulties, however, were seldom described by students in terms of "mental health," but rather as the expected burdens of studying abroad. This disassociation suggests that many students do not recognize the connection between their everyday struggles and broader mental health concerns. Interventions therefore need to be anchored in students' immediate experiences, framing common difficulties with coursework, language barriers, and social integration as meaningful signals of well-being. Embedding mental health concepts into resources that address these concrete stressors may enable students to identify early warning signs of poor mental health [38] in a manner that reduces stigma and increases relevance.

While participants perceived that mental health is discussed more openly in Australia than in their home country, they also acknowledged that stigma rooted in Chinese cultural norms continues to influence their own responses. This internalized stigma contributed to a tendency toward self-reliance, with students preferring to manage problems privately rather than seek professional assistance. Familiar expressions such as "emo" were used to describe low mood or stress; however, students rarely linked these experiences to mental health frameworks. Culturally sensitive interventions that recontextualize these expressions within an evidence-based understanding of mental health can begin to shift perceptions. Such an approach is consistent with Jorm's framework on mental health literacy, which emphasizes improving knowledge and beliefs about mental disorders to facilitate recognition, management, and prevention [39,40].

The coping strategies reported by participants were predominantly self-directed, ranging from lifestyle adjustments and online searching to informal peer support. While some of these strategies were constructive, others, such as excessive gaming or alcohol use, carried risks. Concerns about the credibility of online resources were common; however, students continued to rely on them due to immediacy and accessibility. Professional services were generally viewed as relevant only for severe cases and were hindered by language barriers, cost, and mistrust. These findings indicate a need for interventions that provide trusted, university-endorsed resources, promote safer self-management, and create low-threshold entry points to care. Incorporating culturally aligned recommendations, university-endorsed supports, or peer-led information may help build trust and encourage gradual engagement with professional services [41].

In summary, Phase 1 findings highlight the importance of culturally sensitive, staged interventions that validate everyday struggles, foster mental health literacy, and strengthen pathways to appropriate support, in line with Jorm's mental health literacy framework [39,40].

Phase 2: Co-Design Workshops

Overview

For Phase 2, a total of 7 participants who had interacted with the initial prototype for a week were included in the design workshops, with 3 and 4 participants allocated to 2 separate design workshops, respectively. Among them, 4 had participated in Phase 1, and 3 were recruited through internet-based advertisement. The thematic analysis of the journey mapping and "I like, I wish, what if" design activities elicited themes related to positive ("Current strengths") and constructive ("Areas to improve") feedback and alternative ideas on the overall content. Textbox 1 shows the participants' feedback on the 3 main topics of discussion, namely specific survey question development, recommendation and feedback content (based on survey results), and functionality and interface design.

XSL•FO
RenderX

Textbox 1. Design considerations derived from group design workshops in Phase 2.

Survey and questionnaire content development

- Areas to improve
  - Wordings or specific terminologies (eg, self-efficacy) were not easily understood
  - Questions from other scales and incorporating different dimensions of mental health issues
  - Questions need to be more comprehensive and could cover career and study planning, collaboration skills, financial management, family commitment, and visa issues
- Current strengths
  - Depression, Anxiety and Stress Scale–21 items (DASS-21) is a reliable tool
  - Questions cover a range of different daily life circumstances

Func tionality and interface des ign

- Areas to improve
  - Language option for the resources from external links
  - Online forum to share experiences or thoughts
  - Option of follow-up check-in survey
  - Updated news about university enrollment, visa policy, or activities
  - Artificial intelligence–supported tool (eg, chatbot)
  - User profile
  - More introductions on the dashboard page
- Current strengths
  - Easy to navigate the feedback and complete the check-in questions
  - Easy to switch language
  - Good page design (eg, color)

Recommendation content improvement

- Areas to improve
  - More instructions on how to read and use the recommendations
  - Recommendations could be more specific
- Current strengths
  - Feedback is comprehensive
  - Recommendations provided are useful

From the design workshops, participants indicated that the key value of the platform is addressing cross-cultural challenges by offering surveys and recommendations that cover many of the daily life issues faced by Chinese-speaking international students, along with language selection options. However, areas for improvement include some alternative or transformative ideas on the content and engagement aspects, such as covering a broader and more diverse range of daily challenges, providing more individualized and specific recommendations, and incorporating interactive features, such as an artificial intelligence chatbot and discussion forum.

Phase 2 Design Insights and Strategies

The Phase 2 workshops demonstrated how students are engaged with the prototype built from Phase 1 findings and highlighted further refinements needed for impact. Participants valued the tool's focus on cross-cultural challenges, the inclusion of DASS-21 and daily life circumstances, and the accessible, multilingual interface. This confirmed that grounding support in everyday challenges was a meaningful entry point for Chinese-speaking international students.

Meanwhile, participants identified gaps. Survey content was seen as too narrow and occasionally difficult to understand, with calls to expand coverage to career planning, financial

XSL•FO
RenderX

management, collaboration, family commitments, and visa issues. Recommendations were described as useful but overly general, with a need for clearer guidance on how to apply them.

Finally, students emphasized the importance of interactivity and personalization. Suggested features included an artificial intelligence chatbot, peer discussion forum, follow-up surveys, and practical updates on enrollment and visa policies. These insights extend the Phase 1 findings by highlighting that effective tools must move beyond awareness-building to provide specific, actionable, and engagement supports that fit students' daily realities.

Discussion

Principal Findings

This study explored needs and challenges for developing a digital tool to support the mental health of Chinese-speaking international students based on interviews and workshops conducted over 2 phases. Overall, findings revealed significant gaps in the students' mental health awareness, literacy, and help-seeking, with likely internalized stigma toward mental health issues and help-seeking that further prevents preventative and interventive engagement with mental health. It is also evident that studying and living abroad means that students are experiencing elevated risk factors, such as acute and often ongoing stress, uncertainty, and lack of support, while having limited access to reliable, culturally tailored resources for their preferred support mechanism—self-help. In response, the study proposes design implications aimed at addressing these challenges through integrated self-help techniques and culturally relevant, evidence-based resources aimed at (1) raising mental health awareness, (2) building mental health literacy, and (3) encouraging proactive help-seeking and access to health services if mental health risk factors emerge. The following sections will elaborate on these key implications separately.

Bridging Lived Experience and Mental Health Awareness Through Challenges

This study shows that while students identified daily life challenges created by the acute geographical transition with ease, they have difficulties recognizing the mental health implications that are associated with these elevated challenges. For example, they have identified challenges associated with cross-cultural adaptation, demanding and often overwhelming course workloads, and navigating a new social and independent lifestyle, and they have articulated the negative emotions and mood changes ("emo") associated with these challenges with clarity. However, they did not seem to interpret these experiences and feelings as possible warning signs of mental health illness [38]. Recognition of early risk factors and symptoms of mental health issues is considered fundamental in mental health awareness, which aids identification, management, and prevention of mental health issues [42,43]. A lack of recognition of early warning signs is consistent with previous literature [44,45] reporting that lacking awareness is a key barrier in university students seeking help, which could then lead to serious consequences, including the development of severe mental health disorders [46] and suicidal behaviors [47].

To address this challenge, a mental health support system for Chinese-speaking international students' needs to "build a bridge" by contextualizing the development of mental health awareness that the students need to build in the everyday challenges that the students tend to recognize themselves. This could be supported by a self-directed check-in mechanism [48-51], where students are invited to fill out surveys and questionnaires that cover concrete and shared academic and daily life challenges, cross-cultural difficulties, and mental health measures (eg, DASS-21). Through this comprehensive self-assessment check-in tool, students are provided insights into the main areas of significant everyday life challenges while being engaged in the mental health implications of these experiences and challenges, which leverage promising effects on increasing awareness of mental health issues, particularly in young people, where multiple studies have observed improved self-efficacy [52-55]. To strategically demonstrate the results of the self-assessment tool, the system should integrate the categories of challenges with associated mental health implications when visualizing the results.

It is also essential to address language barriers with a culturally sensitive adaptation approach [56], where measurement tools validated in Chinese contexts and mindful symptom descriptions are used to engage students. Consistent with prior research [57-61], our findings show that language barriers are significant barriers that affect most domains of international students' lives. Language selection options have been identified as crucial in digital mental health services and interventions to effectively serve diverse user populations [62]. A study on digital mental health platforms for Chinese-speaking international students found that offering both simplified and traditional Chinese options significantly improved student engagement with the programs [19]. Importantly, adapting an evidence-based digital tool to a new language often requires appropriate cultural appropriations instead of a direct translation [63]. Therefore, using validated Chinese versions of scales or intervention programs (eg, the Chinese Short Version of the Depression Anxiety and Stress Scale [64] to assess students' mental health conditions is essential.

Mental Health Literacy Building Through Culturally Adapted Resource

In addition to a less-than-ideal mental health awareness, this study also identified a common tendency of underdeveloped mental health literacy (MHL), which manifested in insufficient and often incoherent knowledge regarding the causes, recognition, duration, and severity of various mental health conditions among the students. The data also revealed limited access to educational resources and, at times, reliance on ineffective self-management strategies to cope with negative emotions. This limited MHL, which has been defined as "both knowledge and beliefs about mental disorders, assisting in their recognition, management, or prevention" [39], plays a significant role in mental health prevention while contributing to better decision-making and increasing help-seeking in the context of mental health [65-67]. Our findings also showed that students' perception of the host country's (ie, Australia) attitude is one that is more open and understanding in contrast to their country of origin. This perceived openness led them to believe

XSL•FO
RenderX

that there are more resources and services available; however, little was articulated as to how they would make use of these resources. This is largely due to the language barriers, cultural concerns (eg, worries of not being understood), and negative perceptions about the relevant resources (eg, mental health services are not helpful).

To address the challenge of underdeveloped MHL complicated by culturally based stigma, Jorm et al [39] framework for increasing MHL can provide a conceptual base that strategically focuses on developing students' recognition and beliefs about mental health. However, for a culturally appropriate intervention for Chinese-speaking international students, integration of culturally informed design consideration is essential to address the unique challenges demonstrated in the findings. Key design considerations can include (1) normalizing emotional challenges as shared experiences in the student community [68], (2) enabling anonymous peer stories, examples, and narratives to foster a sense of group belonging and culturally familiar support [69], and (3) maintaining a high level of confidentiality while encouraging self-help through self-directed content without risk of stigmatization [70].

To further operationalize these key design considerations, the culturally sensitive approach will entail incorporating culturally related ways of expression and educational or health resources in designing the mental health technologies [71,72]. The key is to identify culturally appropriate ways to proactively reflect and ameliorate stigma (eg, "Mental health issues are often seen as just complaining without real problems"; P9), cultural biases (eg, the unique way Chinese-speaking international students describe mood swings as "emo"), and limited understanding or knowledge about the local health care systems and social norms. Personalization mechanisms will also be used to enable students to gain knowledge that can help them to interpret their own results while relating to MHL. Importantly, rather than presenting a matter-of-fact type diagnosis, the system would provide feedback to students with a culturally informed sense of empathy, linking the results to the challenges that they experience to normalize the results that the student might find too challenging to accept. In addition, the feedback will be linked to recommendations or resources that align with the students' interests and needs, particularly those related to cultural aspects, such as language support, peer support, and assistance with geographical transitions.

Encouraging Professional Help-Seeking

This study found that Chinese-speaking international students shared a limited level of help-seeking intentions where they feel that professional help is "something for other people" or for "very severe cases." This is consistent with prior work reporting that international students are resistant toward counselors and counseling services [73]. This resistance persists even when participants experience emotional challenges, such as feeling "emo," leading to a preference for solitude and self-coping. Recent studies have indicated that help-seeking intentions are positively correlated with better well-being outcomes among university students [74], especially international students whose help-seeking intentions subsequently influence their levels of psychological distress [75]. However, important barriers were reported by prior work [76], suggesting that international students tend to be inexperienced with professional psychological help, which then leads to negative attitudes toward these services (eg, perceiving these services as untrustworthy and inappropriate for solving personal difficulties). In line with prior work, various barriers to seeking professional help were identified in this study, including language and cultural barriers, financial concerns, fear of judgment, and a lack of knowledge and experience with the health care system in Australia.

To design support that addresses the resistant attitude toward professional service and encourages students who experience emerging mental health issues to seek professional support, it is key to "demystify" the process of seeking and receiving professional support and build a culturally appropriate "roadmap" to professional help. This can be operationalized by (1) providing peer voice (eg, lived experience examples) and voice from culturally tailored mental health services, (2) encouraging students to take the first step in help seeking, and (3) supporting students to navigate the initial contact point and then into the broader health care system while providing language, culture, financial information, and emotional support to overcome the specific barriers as articulated in the findings.

The system should provide destigmatizing mechanisms through peer stories for students to feel safe and reassured to take the first step toward accessing the health care services [77]. Relevant lived experience stories of other Chinese-speaking international students can be integrated here to reduce perceived language and cultural barriers and social stigma while showcasing the effects and benefits of professional support. In addition, providing information that demonstrates, in a linguistically and culturally sensitive way, why help-seeking is necessary and beneficial and how services protect the privacy of help-seekers and ways that help can be sought "here and now" without monetary concerns. This would entail listing available free services tailored to the students' preferences, interpretation services coupled with various free helplines, and what specific areas these helplines provide support on. This "in the moment" education could significantly increase the students' familiarity with professional help and aid their decision-making processes, which increases the chances of them taking the first step.

For students who are at an elevated risk of mental health problems, seeking and obtaining mid- to long-term professional support is necessary. For this group of at-risk students, it is imperative to guide and empower them to contact emergency services and develop the ability to go beyond the initial step and access the broader allied health care system for support of an ongoing nature. The design here should focus on accompanying students to build their own care pathway by providing culturally tailored guidelines and resources on available services and immediate steps. This can include a "one-stop-shop" toolkit of specific functions with contact and action suggestions at each step, such as (1) navigating the Australian health care system, (2) identifying clinicians with appropriate backgrounds and experiences, (3) identifying and using available financial support, and (4) understanding what happens in a therapy session. As student and social support initiatives that introduce educational resources or social skills have shown benefits in strengthening self-efficacy and

XSL•FO
RenderX

confidence [78,79], this toolkit design will not only rapidly increase students' relevant knowledge but also build their confidence and ability to navigate and make use of support that is available to them.

It is important to note that the universities are often the first place students turn to for help. It is essential for educational institutions to play a central role in introducing or providing a digital mental health support service. This study identified multiple challenges faced by international students, many of which are closely related to their academic experiences. In addition, a significant limitation of university services was also observed in our study and prior work. A critical analysis of Australian university mental health policies for international students found that university policies often frame distress as an individual concern, placing the focus and responsibility on students to be resilient, rather than providing much-needed support [13]. Prior research also found that the awareness of on-campus mental health services can have a positive influence on international students' help-seeking attitude and behavior [7] and that university-based information and digital tools have been proven effective in delivering these services to students [63,80]. While it is crucial for educational institutions to provide information on the availability of on-campus support, this is often not culturally and linguistically adapted. In this case, the designed tool should consider implementation possibilities for universities to adopt and use the digital tool.

Limitations and Future Work

There are a few limitations of this study. Recruitment relied on internet-based recruitment that invited students "interested in addressing their mental health concerns," which likely introduced self-selection bias toward participants who were more willing to discuss mental health. Perspectives of more avoidant or help-resistant students may therefore be underrepresented. The qualitative design and the relatively small, Chinese-speaking sample constrain statistical generalizability to the broader international-student population. Interpretation was situated within existing literature to support transferability. The disciplinary composition was imbalanced, with approximately 40% (8/20) of participants enrolled in public health, which may have oriented feedback toward public-health framings of mental health and intervention design. Within this qualitative dataset, no clear or consistent patterns were observed by years in Australia or by mental health literacy. Variation appeared more related to students' study backgrounds, whether in mental health or other disciplines. Future research should recruit a larger and more balanced cohort, use stratified sampling across disciplines and lengths of stay, and include validated measures of mental-health literacy to enable multivariable and mixed methods analyses that inform more individualized, targeted in-platform feedback.

The given scope of the study focused on Australia as an academic context, and expanding the geographic location may allow for different findings to emerge. The study scope also limited the possibility of exploring diverse cultures that international students come from; however, it is essential for having an exploratory study to dive deeper into 1 student group and learn how cultural adaptation can be designed and developed. Future work can explore other cultures and their influences on the international students' mental health to uncover similarities and differences in the students' lived experiences and the unique mental health challenges and needs.

Conclusions

In summary, a set of clear insights and solutions for designing mental health support for Chinese-speaking international students has been presented. This study adds insights to the digital tool development in relation to mental health support, with an understanding of the challenges and needs of this specific user group—international students. Current findings provide a rich and in-depth description of the barriers and needs associated with international students' mental health while relocating to a new cultural context, and we have discussed how these needs and concerns are synthesized to inform the design implications. These design implications captured key mechanisms that enable culturally adapted strategies to help build international students' mental health awareness and literacy while encouraging them to take actions toward professional help-seeking when needed. These design implications provide key considerations for researchers, designers, and developers who are interested in providing culturally sensitive and appropriate support for international students' mental health and well-being.

Acknowledgments

The authors would like to thank the students who took the time to participate in this study. This project is funded by the Department of Health, Disability and Ageing in Australia in the Health and Medical Research category (grant ID GO4285).

Data Availability

The datasets used and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

All authors made substantial contributions to the conception or design of the work (LW, CZ, ML, JPS, PK, JX, and PO), or to the acquisition, analysis, or interpretation of data (LW and CZ). LW and CZ drafted the manuscript, and all other authors listed reviewed it critically. All authors gave final approval of the version to be published. LW has full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. All authors share responsibility for the final version of the work submitted and published.

XSL•FO
RenderX

Conflicts of Interest

The authors declare no conflict of interest.

Multimedia Appendix 1
Interview guide.
[DOCX File, 19 KB - formative_v9i1e76695_app1.docx]

Multimedia Appendix 2
Workshop guide.
[DOCX File, 448 KB - formative_v9i1e76695_app2.docx]

Multimedia Appendix 3
Interface examples.
[DOCX File, 938 KB - formative_v9i1e76695_app3.docx]

References

1. Oduwaye O, Kiraz A, Sorakin Y. A trend analysis of the challenges of international students over 21 years. Sage Open 2023 Oct;13(4):21582440231210387. [doi: 10.1177/21582440231210387]
2. International student numbers by country, by state and territory. Australian Government Department of Education. 2024. URL: https://www.education.gov.au/international-education-data-and-research/international-student-numbers-country-state-and-territory [accessed 2025-10-07]
3. Redfern K. An empirical investigation of the incidence of negative psychological symptoms among Chinese international students at an Australian university. Aust J Psychol 2016 Dec 1;68(4):281-289. [doi: 10.1111/ajpy.12106]
4. Zhao J, Chapman E, O'Neill M. Mental health risks for Chinese international students in Australia: enduring problems, possible solutions. ERP 2022;49:1-28. [doi: 10.70953/ERPv49.22001]
5. Zhao J, Chapman E, O'Donoghue T. Threats to the emotional wellbeing of mainland Chinese students studying in Australia: an interpretivist study. Int J Qual Stud Health Well-being 2023 Dec;18(1):2221912. [doi: 10.1080/17482631.2023.2221912] [Medline: 37311124]
6. Zhao J, Chapman E, Houghton S, Lawrence D. Perceived discrimination as a threat to the mental health of Chinese international students in Australia. Front Educ 2022;7:726614. [doi: 10.3389/feduc.2022.726614]
7. Chen H, Akpanudo U, Hasler E. How do Chinese international students view seeking mental health services? J Int Stud 2020;10(2):286-305 [FREE Full text] [doi: 10.32674/jis.v10i2.765]
8. Lu SH, Dear BF, Johnston L, Wootton BM, Titov N. An internet survey of emotional health, treatment seeking and barriers to accessing mental health treatment among Chinese-speaking international students in Australia. Couns Psychol Q 2014 Jan 2;27(1):96-108. [doi: 10.1080/09515070.2013.824408]
9. Yang J, Li Y, Gao R, Chen H, Yang Z. Relationship between mental health literacy and professional psychological help-seeking attitudes in China: a chain mediation model. BMC Psychiatry 2023 Dec 21;23(1):956. [doi: 10.1186/s12888-023-05458-5] [Medline: 38129805]
10. Mo PKH, Mak WWS. Help-seeking for mental health problems among Chinese: the application and extension of the theory of planned behavior. Soc Psychiatry Psychiatr Epidemiol 2009 Aug;44(8):675-684. [doi: 10.1007/s00127-008-0484-0] [Medline: 19262970]
11. Shi W, Shen Z, Wang S, Hall BJ. Barriers to professional mental health help-seeking among Chinese adults: a systematic review. Front Psychiatry 2020;11:442. [doi: 10.3389/fpsyt.2020.00442] [Medline: 32508688]
12. Li HH, Holroyd E, Lau J, Li X. Stigma, subsistence, intimacy, face, filial piety, and mental health problems among newly HIV-diagnosed men who have sex with men in China. J Assoc Nurses AIDS Care 2015;26(4):454-463. [doi: 10.1016/j.jana.2015.02.004] [Medline: 26066696]
13. Peterie M, Ramia G, Broom A, Choi I, Brett M, Williams Veazey L. "You're on your own, kid": a critical analysis of Australian universities' international student mental health strategies. Aust J Social Issues 2025 Mar;60(1):334-352. [doi: 10.1002/ajs4.349]
14. Polanczyk GV, Salum GA, Sugaya LS, Caye A, Rohde LA. Annual research review: a meta-analysis of the worldwide prevalence of mental disorders in children and adolescents. J Child Psychol Psychiatry 2015 Mar;56(3):345-365. [doi: 10.1111/jcpp.12381] [Medline: 25649325]
15. Becker TD, Torous JB. Recent developments in digital mental health interventions for college and university students. Curr Treat Options Psych 2019 Sep;6(3):210-220. [doi: 10.1007/s40501-019-00178-8]
16. Montagni I, Tzourio C, Cousin T, Sagara JA, Bada-Alonzi J, Horgan A. Mental health-related digital use by university students: a systematic review. Telemed J E Health 2020 Feb;26(2):131-146. [doi: 10.1089/tmj.2018.0316] [Medline: 30888256]

XSL•FO
RenderX

17. Oti O, Pitt I. Online mental health interventions designed for students in higher education: a user-centered perspective. Internet Interv 2021 Dec;26:100468. [doi: 10.1016/j.invent.2021.100468] [Medline: 34703772]
18. Riboldi I, Cavaleri D, Calabrese A, et al. Digital mental health interventions for anxiety and depressive symptoms in university students during the COVID-19 pandemic: a systematic review of randomized controlled trials. Rev Psiquiatr Salud Ment 2023;16:47-58. [doi: 10.1016/j.rpsm.2022.04.005] [Medline: 35755491]
19. Choi I, Mestroni G, Hunt C, Glozier N. Personalized help-seeking web application for Chinese-speaking international university students: development and usability study. JMIR Form Res 2023 Feb 17;7(1):e35659. [doi: 10.2196/35659] [Medline: 36800231]
20. Silverman AL, Ferguson I, Bullis JR, Bajwa H, Mei S, Beard C. Program evaluation of internet-delivered cognitive behavioral treatments for anxiety and depression in a digital clinic. J Mood Anxiety Disord 2025 Mar;9:100106. [doi: 10.1016/j.xjmad.2025.100106] [Medline: 40655231]
21. Black Dog Institute [Website in Chinese]. URL: https://chineseonlineclinic.blackdoginstitute.org.au/ [accessed 2025-09-01]
22. Framework for innovation. Design Council. 2021. URL: https://www.designcouncil.org.uk/our-resources/framework-for-innovation/ [accessed 2025-10-03]
23. Miro. URL: https://miro.com/ [accessed 2025-10-03]
24. Lukka L, Vesterinen M, Salonen A, et al. User journey method: a case study for improving digital intervention use measurement. BMC Health Serv Res 2025 Apr 1;25(1):479. [doi: 10.1186/s12913-025-12641-9] [Medline: 40165237]
25. Henry JD, Crawford JR. The short-form version of the Depression Anxiety Stress Scales 21 (DASS-21): construct validity and normative data in a large non-clinical sample. Br J Clin Psychol 2005 Jun;44(Pt 2):227-239. [doi: 10.1348/014466505X29657] [Medline: 16004657]
26. Gong X, X-y X, Xu R, Y-j L. Psychometric properties of the Chinese versions of Depression Anxiety Stress Scale-21 (DASS-21) in Chinese college students. Chin J Clin Psychol 2010;18(4):443-446 [FREE Full text]
27. Osman A, Wong JL, Bagge CL, Freedenthal S, Gutierrez PM, Lozano G. The Depression Anxiety Stress Scales-21 (DASS-21): further examination of dimensions, scale reliability, and correlates. J Clin Psychol 2012 Dec;68(12):1322-1338. [doi: 10.1002/jclp.21908] [Medline: 22930477]
28. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research Electronic Data Capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381. [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
29. Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: building an international community of software platform partners. J Biomed Inform 2019 Jul;95:103208. [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]
30. Edwards-Jones A. Qualitative data analysis with NVIVO. J Educ Teach 2014;40(2):193-195. [doi: 10.1080/02607476.2013.866724]
31. Clarke V, Braun V. Thematic analysis. J Posit Psychol 2017 May 4;12(3):297-298. [doi: 10.1080/17439760.2016.1262613]
32. Sun K, Mulvaney MK. Parental involvement in Chinese immigrant mothers: the influences of cultural and parenting cognitions. J Child Fam Stud 2022 Jan;31(1):169-184. [doi: 10.1007/s10826-021-02158-0]
33. Wu P, Robinson CC, Yang C, et al. Similarities and differences in mothers' parenting of preschoolers in China and the United States. Int J Behav Dev 2002 Nov;26(6):481-491. [doi: 10.1080/01650250143000436]
34. Ran MS, Hall BJ, Su TT, et al. Stigma of mental illness and cultural factors in Pacific Rim region: a systematic review. BMC Psychiatry 2021 Jan 7;21(1):8. [doi: 10.1186/s12888-020-02991-5] [Medline: 33413195]
35. Badanta B, González-Cano-Caballero M, Suárez-Reina P, Lucchetti G, de Diego-Cordero R. How does Confucianism influence health behaviors, Health outcomes and medical decisions? A scoping review. J Relig Health 2022 Aug;61(4):2679-2725. [doi: 10.1007/s10943-022-01506-8] [Medline: 35141796]
36. Hsiao FH, Klimidis S, Minas H, Tan ES. Cultural attribution of mental health suffering in Chinese societies: the views of Chinese patients with mental illness and their caregivers. J Clin Nurs 2006 Aug;15(8):998-1006. [doi: 10.1111/j.1365-2702.2006.01331.x] [Medline: 16879544]
37. Cao Y, Chen X, Fan R. Toward a Confucian family-oriented health care system for the future of China. J Med Philos 2011 Oct;36(5):452-465. [doi: 10.1093/jmp/jhr036] [Medline: 21984753]
38. Warning signs of mental illness. American Psychiatric Association. URL: https://www.psychiatry.org/patients-families/warning-signs-of-mental-illness [accessed 2025-10-03]
39. Jorm AF, Korten AE, Jacomb PA, Christensen H, Rodgers B, Pollitt P. "Mental health literacy": a survey of the public's ability to recognise mental disorders and their beliefs about the effectiveness of treatment. Med J Aust 1997 Feb 17;166(4):182-186. [doi: 10.5694/j.1326-5377.1997.tb140071.x] [Medline: 9066546]
40. Jorm AF. Mental health literacy. Public knowledge and beliefs about mental disorders. Br J Psychiatry 2000 Nov;177(5):396-401. [doi: 10.1192/bjp.177.5.396] [Medline: 11059991]
41. Cole AL, Navin F, Reid D. Exploring university student mental health and wellbeing through a low-barrier peer-led service: emerging insights from The Living Room. JANZSSA 2025;33(1):43-64. [doi: 10.30688/janzssa.2025-1-01]
42. Budd M, Iqbal A, Harding C, Rees E, Bhutani G. Mental health promotion and prevention in primary care: what should we be doing vs. what are we actually doing? Ment Health Prev 2021 Mar;21:200195. [doi: 10.1016/j.mhp.2020.200195]

XSL•FO
RenderX

43. Singh V, Kumar A, Gupta S. Mental health prevention and promotion—A narrative review. Front Psychiatry 2022;13:898009. [doi: 10.3389/fpsyt.2022.898009] [Medline: 35958637]
44. Gagnon MM, Gelinas BL, Friesen LN. Mental health literacy in emerging adults in a university setting: distinctions between symptom awareness and appraisal. J Adolesc Res 2017 Sep;32(5):642-664. [doi: 10.1177/0743558415605383]
45. Kalkbrenner MT, James C, Pérez-Rojas AE. College students' awareness of mental disorders and resources: comparison across academic disciplines. J College Stud Psychother 2022 Apr 3;36(2):113-134. [doi: 10.1080/87568225.2020.1791774]
46. Dewa LH, Cecil E, Eastwood L, Darzi A, Aylin P. Indicators of deterioration in young adults with serious mental illness: a systematic review protocol. Syst Rev 2018 Aug 16;7(1):123. [doi: 10.1186/s13643-018-0781-y] [Medline: 30115109]
47. Bagge CL, Littlefield AK, Wiegand TJ, et al. A controlled examination of acute warning signs for suicide attempts among hospitalized patients. Psychol Med 2023 May;53(7):2768-2776. [doi: 10.1017/S0033291721004712]
48. Brydges R, Dubrowski A, Regehr G. A new concept of unsupervised learning: directed self-guided learning in the health professions. Acad Med 2010 Oct;85(10):S49-S55. [doi: 10.1097/ACM.0b013e3181ed4c96] [Medline: 20881703]
49. Zhong B, Xie L. Making "Joy Pie" to stay joyful: self-care interventions alleviate college students' mental health challenges. Int J Environ Res Public Health 2023 Feb 21;20(5):3823. [doi: 10.3390/ijerph20053823] [Medline: 36900839]
50. Ribeiro IJS, Freire IV, de Araújo TM. Are stress management interventions effective in reducing stress, anxiety, and depression in college students? Clin Psychol Sci Pract 2023;30(4):450-452. [doi: 10.1111/cpsp.12385]
51. Pelucio L, Quagliato LA, Nardi AE. Therapist-guided versus self-guided cognitive-behavioral therapy: a systematic review. Prim Care Companion CNS Disord 2024 Mar 21;26(2):23r03566. [doi: 10.4088/PCC.23r03566] [Medline: 38512281]
52. Bennett SD, Cuijpers P, Ebert DD, et al. Practitioner review: unguided and guided self-help interventions for common mental health disorders in children and adolescents: a systematic review and meta-analysis. J Child Psychol Psychiatry 2019 Aug;60(8):828-847. [doi: 10.1111/jcpp.13010] [Medline: 30775782]
53. Babbage CM, Jackson GM, Davies EB, Nixon E. Self-help digital interventions targeted at improving psychological well-being in young people with perceived or clinically diagnosed reduced well-being: systematic review. JMIR Ment Health 2022 Aug 26;9(8):e25716. [doi: 10.2196/25716] [Medline: 36018675]
54. Sharma A, Rushton K, Lin IW, Nguyen T, Althoff T. Facilitating self-guided mental health interventions through human-language model interaction: a case study of cognitive restructuring. Presented at: CHI '24: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems; May 11-16, 2024; Honolulu, HI, USA. [doi: 10.1145/3613904.3642761]
55. Voderholzer U, Hilbert S, Fischer A, Neumüller J, Schwartz C, Hessler-Kaufmann JB. Frequency and level of self-efficacy predict the effectiveness of therapist- and self-guided exposure in obsessive compulsive disorder. Behav Cogn Psychother 2020 Nov;48(6):751-755. [doi: 10.1017/S1352465820000582] [Medline: 32814601]
56. Molander O, Särnholm J, Hollander AC, Jayaram-Lindström N. Mind the Gap: scaling-up access to culturally adapted mental health treatments for minority populations. BMC Glob Public Health 2024 Sep 5;2(1):61. [doi: 10.1186/s44263-024-00092-z] [Medline: 39681953]
57. Khawaja NG, Stallman HM. Understanding the coping strategies of international students: a qualitative approach. Aust J Guid Couns 2011 Jan;21(2):203-224. [doi: 10.1375/ajgc.21.2.203]
58. Lowinger R, He Z, Lin M, Chang M. The impact of academic self-efficacy, acculturation difficulties, and language abilities on procrastination behavior in Chinese international students. Coll Stud J 2014;48(1):141-152 [FREE Full text]
59. Lowinger RJ, Kuo BCH, Song HA, et al. Predictors of academic procrastination in Asian international college students. J Stud Aff Res Pract 2016 Jan 2;53(1):90-104. [doi: 10.1080/19496591.2016.1110036]
60. Young TJ, Sercombe PG, Sachdev I, Naeb R, Schartner A. Success factors for international postgraduate students' adjustment: exploring the roles of intercultural competence, language proficiency, social contact and social support. Eur J High Educ 2013 Jun;3(2):151-171. [doi: 10.1080/21568235.2012.743746]
61. Meng Q, Zhu C, Cao C. Chinese international students' social connectedness, social and academic adaptation: the mediating role of global competence. High Educ 2018 Jan;75(1):131-147. [doi: 10.1007/s10734-017-0129-x]
62. Braun P, Schwientek AK, Angerer P, et al. Investigating information needs and preferences regarding digital mental health services among medical and psychology students in Germany: a qualitative study. Digit Health 2023;9:20552076231173568. [doi: 10.1177/20552076231173568] [Medline: 37256006]
63. Lattie EG, Stiles-Shields C, Graham AK. An overview of and recommendations for more accessible digital mental health services. Nat Rev Psychol 2022 Feb;1(2):87-100. [doi: 10.1038/s44159-021-00003-1] [Medline: 38515434]
64. Chen IH, Chen CY, Liao XL, et al. Psychometric properties of the Depression, Anxiety, and Stress Scale (DASS-21) among different Chinese populations: a cross-sectional and longitudinal analysis. Acta Psychol (Amst) 2023 Oct;240:104042. [doi: 10.1016/j.actpsy.2023.104042] [Medline: 37783184]
65. Wei Y, McGrath PJ, Hayden J, Kutcher S. Mental health literacy measures evaluating knowledge, attitudes and help-seeking: a scoping review. BMC Psychiatry 2015 Nov 17;15(1):291. [doi: 10.1186/s12888-015-0681-9] [Medline: 26576680]
66. Özparlak A, Karakaya D, Özer Z. The association of mental health literacy with mental well-being and help-seeking in young people: a systematic review and meta-analysis. J Pediatr Nurs 2023;73:e243-e250. [doi: 10.1016/j.pedn.2023.09.017] [Medline: 37775428]

XSL•FO
RenderX

67. Rüegg R. Decision-making ability: a missing link between health literacy, contextual factors, and health. Health Lit Res Pract 2022 Jul;6(3):e213-e223. [doi: 10.3928/24748307-20220718-01] [Medline: 36099035]
68. Zhuang XY, Wong DFK, Cheng CW, Pan SM. Mental health literacy, stigma and perception of causation of mental illness among Chinese people in Taiwan. Int J Soc Psychiatry 2017 Sep;63(6):498-507. [doi: 10.1177/0020764017719303] [Medline: 28805154]
69. Yeh C, Inose M. Difficulties and coping strategies of Chinese, Japanese, and Korean immigrant students. Adolescence 2002;37(145):69-82. [Medline: 12003292]
70. Chen SX, Mak WWS. Seeking professional help: Etiology beliefs about mental illness across cultures. J Couns Psychol 2008 Oct;55(4):442-450. [doi: 10.1037/a0012898] [Medline: 22017551]
71. Ahmed N, Liu X, Al-Hazwani I, Huang EM. Cultural dimensions and mental health technology: a systematic review of hofstede's dimensions in shaping mental health experiences. Presented at: CHI EA '24: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems; May 11-16, 2024; Honolulu, HI, USA. [doi: 10.1145/3613905.3650842]
72. Ng SH, Soon LK, Su TT. Emotion-aware chatbot with cultural adaptation for mitigating work-related stress. Presented at: Asian CHI 2023; Apr 28, 2023; Indonesia. [doi: 10.1145/3604571.3604578]
73. Alavi M, Mohamed Sh S, Maryam Ger S, Tajudin Ni M. International students' mental health and attitude toward counseling centers. J of Applied Sciences 2014 Aug 1;14(16):1871-1876. [doi: 10.3923/jas.2014.1871.1876]
74. McCabe M, Byrne M, Gullifer J, Cornish K. The relationship between university student help-seeking intentions and well-being outcomes. Front Psychiatry 2024;15:1407689. [doi: 10.3389/fpsyt.2024.1407689] [Medline: 39035604]
75. Zeng F, John WCM, Qiao D, Sun X. Association between psychological distress and mental help-seeking intentions in international students of national university of Singapore: a mediation analysis of mental health literacy. BMC Public Health 2023 Nov 28;23(1):2358. [doi: 10.1186/s12889-023-17346-4] [Medline: 38017406]
76. Dadfar S, Friedlander ML. Differential attitudes of international students toward seeking professional psychological help. J Couns Psychol 1982;29(3):335-338. [doi: 10.1037/0022-0167.29.3.335]
77. Chow GM, Bird MD, Gabana NT, Cooper BT, Swanbrow Becker MA. A program to reduce stigma toward mental illness and promote mental health literacy and help-seeking in National Collegiate Athletic Association Division I student-athletes. J Clin Sport Psychol 2020;15(3):185-205. [doi: 10.1123/jcsp.2019-0104]
78. Çelik E. Mediating and moderating role of academic self-efficacy in the relationship between student academic support and personal growth initiative. Aust J Career Dev 2015 Jul;24(2):105-113. [doi: 10.1177/1038416215583563]
79. Agustanti A, Astuti K. Relationship between social skills and social support with peers' academic confidence on boarding high school students. J Int Dakwah and Communication 2022;2(2):97-110. [doi: 10.55849/jidc.v2i2.201]
80. Harith S, Backhaus I, Mohbin N, Ngo HT, Khoo S. Effectiveness of digital mental health interventions for university students: an umbrella review. PeerJ 2022;10:e13111. [doi: 10.7717/peerj.13111] [Medline: 35382010]

Abbreviations

DASS-21: Depression, Anxiety and Stress Scale–21 items
MHL: mental health literacy
REDCap: Research Electronic Data Capture

Edited by A Mavragani; submitted 29.04.25; peer-reviewed by J Han, Z Zhang; revised version received 19.09.25; accepted 21.09.25; published 21.10.25.

Please cite as:
Wu L, Zhu C, Seguin JP, Xie J, Kulkarni P, Li M, Olivier P
Designing Culturally Adapted Digital Mental Health Support Tool for Chinese-Speaking International Students in Australia: A Qualitative Co-design Study
JMIR Form Res 2025;9:e76695
URL: https://formative.jmir.org/2025/1/e76695
doi:10.2196/76695

© Ling Wu, Chen Zhu, Joshua Paolo Seguin, Jue Xie, Pranav Kulkarni, Mingye Li, Patrick Olivier. Originally published in JMIR Formative Research (https://formative.jmir.org), 21.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Preprocessing Large-Scale Conversational Datasets: A Framework and Its Application to Behavioral Health Transcripts

Paz Mor Naim[1]; Shiri Sadeh-Sharvit[2,3]; Samuel Jefroykin[2]; Eddie Silber[2,3]; Dennis P Morrison[4]; Ariel Goldstein[1,5,6]

[1]Department of Cognitive and Brain Sciences, Hebrew University of Jerusalem, Mount Scopus, Jerusalem, Israel

[2]Eleos Health, Waltham, MA, United States

[3]Palo Alto University, Palo Alto, CA, United States

[4]Morrison Consulting, Bloomington, IN, United States

[5]Business School, Hebrew University of Jerusalem, Jerusalem, Israel

[6]Department of Psychology, Azrieli Israel Center for Addiction and Mental Health (Azrieli ICAMH), Hebrew University of Jerusalem, Jerusalem, Israel

Corresponding Author:
Paz Mor Naim
Department of Cognitive and Brain Sciences, Hebrew University of Jerusalem, Mount Scopus, Jerusalem, Israel

Abstract

Background: The rise of artificial intelligence and accessible audio equipment has led to a proliferation of recorded conversation transcripts datasets across various fields. However, automatic mass recording and transcription often produce noisy, unstructured data that contain unintended recordings such as hallway conversations, media (eg, TV, radio), or transcription inaccuracies as speaker misattribution or misidentified words. As a result, large conversational transcript datasets require careful preprocessing and filtering to ensure their research utility. This challenge is particularly relevant in behavioral health contexts (eg, therapy, counseling) where deriving meaningful insights, specifically dynamic processes, depends on accurate conversation representation.

Objective: We present a framework for preprocessing large datasets of conversational transcripts and filtering out *non-sessions*—transcripts that do not reflect a behavioral treatment session but instead capture unrelated conversations or background noise. This framework is applied to a large dataset of behavioral health transcripts from community mental health clinics across the United States.

Methods: Our approach integrated basic feature extraction, human annotation, and advanced applications of large language models (LLMs). We began by mapping transcription errors and assessing the number of non-sessions. Next, we extracted statistical and structural features to characterize transcripts and detect outliers. Notably, we used LLM perplexity as a measure of comprehensibility to assess transcript noise levels. Finally, we used zero-shot prompting with an LLM to classify transcripts as sessions or non-sessions, validating its output against expert annotations. Throughout, we prioritized data security by selecting tools that preserve anonymity and minimize the risk of data breaches.

Results: Initial assessment revealed that transcription errors—such as incomprehensible segments, unusually short transcripts, and speaker diarization issues—were present in approximately one-third (n=36 out of 100) of a manually reviewed sample. Statistical outliers revealed that high speaking rate (>3.5 words per second) is associated with short transcripts and answering machine messages, while short conversation duration (<15 min) was an indicator for case management sessions. The 75th percentile of LLM perplexity scores was significantly higher in non-sessions than sessions (permutation test mean difference = −258, P =.02), although this feature alone offered only moderate classification performance (precision =0.63, recall =0.23 after outlier removal). In contrast, zero-shot LLM prompting effectively distinguished sessions from non-sessions with high agreement to expert ratings (κ=0.71) while also capturing the nature of the meeting.

Conclusions: This study's hybrid approach effectively characterizes errors, evaluates content, and distinguishes text types within unstructured conversational dataset. It provides a foundation for research on conversational data, key methods, and practical guidelines that serve as crucial first steps in ensuring data quality and usability, particularly in the context of mental health sessions. We highlight the importance of integrating clinical experts with artificial intelligence tools while prioritizing data security throughout the process.

(JMIR Form Res 2025;9:e78082) doi:10.2196/78082

KEYWORDS

artificial intelligence; behavioral health; clinical documentation; clinical texts; conversational transcripts; data preprocessing; data quality assessment; health informatics; health information systems; large language models; natural language processing; psychotherapy; text classification

XSL•FO
RenderX

Introduction

As one of our primary communication tools, conversations offer a window into human relationships. Speaking with each other is one of the fundamental aspects of our existence. Consequently, transcripts of conversations have garnered significant interest in research across various disciplines [1]. Analyzing conversational data is a central research focus in fields such as health care [2,3], law [4], customer support [5], negotiations [6], education [7], and behavioral treatment and psychotherapy [8-10].

Among these diverse domains, talk therapy, which relies on a conversation between 2 or more individuals, is a particularly intriguing case for conversation analysis. Analysis of treatment sessions can enhance our understanding of different therapeutic protocols [9,11-14]; improve clinical training [15-17]; and provide insights into the relationships between conversational elements, therapeutic outcomes, and the therapeutic alliances [8,10,18,19]. Conversation analysis research of behavioral health sessions can provide comprehensive insights into the intricate dynamics of therapeutic interactions [2]. By meticulously examining communicative choices, researchers can understand how specific counseling strategies impact client engagement, identify successful intervention techniques, and reveal nonlinear patterns of change within therapy sessions [20,21]. This research methodology allows professionals to study the dyadic or group psychotherapeutic processes, predict potential treatment outcomes, develop more effective interventions, and ultimately enhance the quality of behavioral health care [14,22].

Historically, collecting, transcribing, and analyzing treatment conversations have required significant human effort. This is now changing, thanks to 2 major technological advances. First, recent developments in audio capturing technologies and the ubiquitousness of smart devices simplify the collection of conversation data. Second, the abundance of artificial intelligence (AI)–based speech-to-text tools [23-25] has automated transcription and speaker recognition tasks. These innovations have made it possible to capture high-quality human speech in diverse settings with less manual effort [26]. Consequently, large-scale transcription of audio files is now feasible with unprecedented speed and precision. The automation of these processes introduces new challenges. Transcripts' quality depends on several factors: the recording devices and their placement in the room, background noise, internet connection stability, and accurate speech-to-text and speaker diarization models. Additionally, characteristics of the conversation itself, such as slurred speech, dialects, rare languages, or use of slang, all pose further challenges for automatic speech recognition (ASR) models [27]. Similarly, interrupted, rapid, or overlapping speech, unknown number of speakers as well as same-sex speakers [28,29] each introduce complications for speaker recognition systems.

Moreover, the expected growth in large datasets accompanying this automation amplifies these challenges. Large and diverse datasets might include recordings of different quality levels and therefore are prone to various types of errors. Furthermore, when recordings are made routinely and automatically, unintentional recordings—irrelevant conversations, phone calls, empty moments, or accidental noise—may occur. This illustrates how large volumes of automatically generated transcripts are unstructured and susceptible to errors. Therefore, guidelines and methods for error handling, assessing, and filtering should be established.

These challenges are well-known in the broader field of health care, which has long grappled with extracting meaningful information from unstructured data such as clinical notes, discharge summaries, or patient-generated text [30,31]. As the field has evolved, various preprocessing pipelines and methodological frameworks have been proposed for handling missing metadata, variable data quality, and inconsistent documentation practices [32-34]. However, most existing work focuses on clinical records, and less attention has been given to large-scale, conversational health care data.

Yeomans et al [35] have recently proposed a methodological pipeline for building conversation-based research, starting with the planning and collection of conversations and going through editing and analyzing. However, with the availability of ambient AI and larger datasets, many researchers may be working with secondary data—datasets they have not collected themselves. Therefore, they might not have access to the original recordings, to the speech-to-text models, nor to the transcripts editing process. Lacking a ground truth for verification necessitates expanding these guidelines to address such challenges.

In the case of behavioral treatment sessions, the noisiness of the transcript and its accurate representation of the conversation are key for detecting the interventions applied and extracting treatment insights [25,36]. For instance, misidentifying speakers could compromise insights regarding the therapeutic alliance [18,21,37]. Understanding which conversations are professional and which are accidental recordings is the first step in this endeavor.

In this paper the abovementioned challenges are addressed, and a methodological approach for preprocessing a large dataset of behavioral treatment transcripts without access to their respective original recordings is presented. First, a systematic approach for characterizing the data is outlined, followed by methods that allow its assessment for future analysis. This methodology is then illustrated by applying it to a large dataset of deidentified behavioral health sessions. Finally, the strengths and limitations of our approach are discussed. We hope to promote the integration of computational tools in traditional talk therapy and to offer relevant methods for any dataset of conversational transcripts.

Methods

Data and Settings

We analyzed 22,337 behavioral treatment sessions from 50 behavioral health programs across the United States collected through the Eleos Health platform between June 2020 and January 2024. Eleos Health's digital platform is designed to promote behavioral treatment quality by providing intervention feedback, supporting clinical decision-making, and enabling progress note automation. Sessions were processed as part of

XSL•FO
RenderX

the routine implementation of ambient AI tools within participating behavioral health programs. All sessions underwent transcription, deidentification, and anonymization using Eleos' proprietary models before inclusion in the study. The research team had no access to audio recordings or transcripts prior to this process, and only deidentified data were analyzed. Processed transcripts were stored as comma-separated values files with 3 primary columns: deidentified speaker labels, timestamps, and content. Each content row contained a speech segment from either the therapist or the client—typically up to several sentences—segmented by the Eleos ASR model. Session metadata included random therapist and client ID numbers, organization name, session date, and treatment delivery method (phone, video conferencing, or in-person). This study was conducted in concert with the STROBE Checklist [38].

Ethical Considerations

This study was determined exempt from review by the Sterling Institutional Review Board under the Department of Health and Human Services Exemption Category 4. This exemption permits secondary research using identifiable health information when the data are either publicly available or recorded in such a way that subjects cannot be reidentified, and there is no direct interaction with participants. All research procedures adhered to applicable ethical guidelines and regulations. Clients and therapists provided informed consent for the use of anonymized, deidentified session data in secondary research conducted by the company. Both parties retained the option to opt out of having their sessions processed. All data used in this study were deidentified prior to analysis, and no identifiable information was accessed by the research team. No compensation was provided for participation in this secondary analysis, and no images or supplementary materials contain identifiable individuals.

Data Analysis Approach

Our data analysis approach consisted of several stages, each employing different methods for characterizing and filtering the dataset (see Figure 1). The initial stages were exploratory, beginning with a manual review to assess data's content and quality. This process helped us identify categories of files and potential errors in the data, enhancing our understanding of the dataset and the features most suitable for characterization and filtering. Subsequent steps focused on directly classifying transcripts as "sessions" or "non-sessions" (eg, nonprofessional conversations, noise, mock sessions). Using available metadata, we filtered out mock sessions (non-sessions) that had known identifiers. Additionally, we applied zero-shot prompting with a large language model (LLM) to distinguish therapy sessions from irrelevant conversations. Automation played a critical initial role in this phase; however, human expertise - specifically trained psychologists - was incorporated to validate the automated classifications. At this stage, aligned with our primary goal of preparing the dataset for psychological research, various types of interventions were distinguished, and the sessions category was restricted only to formal treatments (excluding peer support, sponsorship, and case management). Notably, we did not prompt the model with these exclusion criteria, testing its ability to make such distinction based solely on therapeutic elements.

Figure 1. Work pipeline—starting with manual review of a dataset subset, followed by statistical feature extraction and outlier detection. Next, we apply zero-shot prompting with a large language model (LLM) to classify transcripts as sessions or non-sessions and validate the LLM's decisions against human annotators. Finally, we compare LLM's perplexity of different classes.

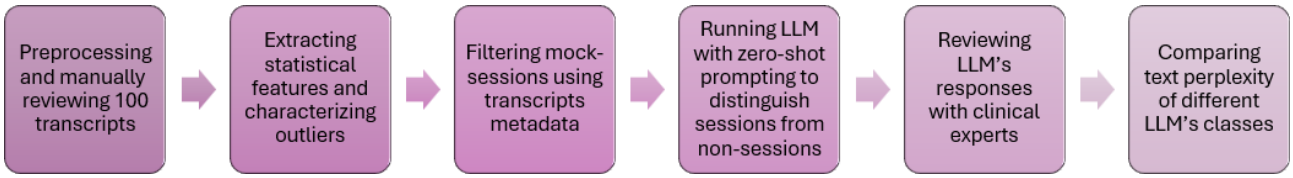

Dataset Preprocessing

Dataset Characteristics

We identified potential duplicates in the dataset by analyzing the similarity between file names. This method is efficient in terms of processing time and computational resources, as it avoids comparing entire text contents. The underlying assumption is that identical files might be mistakenly saved under similar file names. To measure similarity, we used Python's SequenceMatcher function [39]. File names were considered similar if the ratio of their longest matching segment to the total number of characters exceeded 75%. Files meeting this threshold were then manually reviewed to confirm duplication.

Additionally, some transcripts were recorded in languages other than English. To detect the languages, we used the langdetect library in Python [40].

Initial Assessment

A total of 100 randomly selected transcripts were manually reviewed by human raters, with 12 segments analyzed from each transcript to identify common errors and quantify their prevalence. At least 4 categories of errors were defined: (1) non-sessions: transcripts that clearly did not represent a session. At this stage, no distinction was made between types of intervention (eg, peer support, sponsorship, or formal treatment); (2) too short: transcripts with a total duration of less than 2700 seconds (45 min, which approximates the expected session length); (3) unreadable: transcripts with excessive missing words, duplicated segments, or incomprehensible text—readers could not infer the missing content or the meaning of the segment. For example, a transcript with multiple repeated filler phrases (eg, "uh-huh, uh-huh, uh-huh…") was generally deemed unreadable; (4) speaker diarization errors: transcripts with substantial speaker attribution mistakes.

In some cases, annotators indicated that additional context was required to ascertain the presence of an error.

XSL•FO
RenderX

Features Learning

We collected statistics about the whole dataset and each session which included (1) conversation length, (2) speaking rate, (3) frequent words, and (4) segment perplexity.

In addition to characterizing the dataset, these features serve as potential indicators for assessing the data's compatibility for further analysis. For instance, they can highlight recording errors or determine whether the content is session related. In the following section, we elaborate on some of these features and explain how they were calculated for each session.

Conversation Length

Conversation length was calculated by extracting the end timestamp of the final speech segment in the transcript. This method does not account for any time elapsed between when the recording device was activated and the conversation began. However, if the session starts with silence, this approach accurately reflects the length of the conversation.

Speaking Rate

Speaking rate, defined as the average number of words spoken per second, has been employed to predict speaker characteristics [41], detect speech anomalies, and identify changes in conversational dynamics or context [42]. In American conversational speech, the reported average speaking rate ranges from 1.85 to 4.86 words per second (WPS) [41]. Values outside this range may indicate abnormal recordings, such as background noise or transcription errors.

For each transcript, speaking rate was calculated by averaging across segment-level speaking rates. Because silent pauses within segments were not available, we could not adjust for them as is commonly done. However, segments were brief (typically a few seconds) and cut before long silences, making the absence of silent moments negligible and this measure a reasonable approximation. For comparison, we also calculated a global speaking rate by dividing the total word count by the overall conversation length, which incorporates silent periods.

Word Frequency

We explored the content expressed in the transcripts by extracting the most common nouns in the dataset. First, we cleaned each text by removing stop words [43] and technical terms (a list of which can be found in Multimedia Appendix 1). These filtered words were typically used in bureaucratic contexts, such as scheduling meetings or filling out professional forms. Next, we separated each transcript into 2 speaker partitions: therapist and client. Using Python's library, Spacy [44], we collected all the nouns spoken across all transcripts for each speaker. This content analysis revealed common topics discussed during conversations. Additionally, unexpected words that emerged may indicate common errors made by the ASR model.

Perplexity

Perplexity is a measure of a text sequence, indicating its likelihood of being produced by a language model [45]. It is the model's loss when given the words in the sequence and is conceptually similar to the cross-entropy between the model's and the sequence's probability distribution:

$$,\text{Perplexity}(\text{segment})=\exp(-1N\sum i=1N\log(P(wi \mid wi-1,...,w1)))$$

where w_i is a token in the segment's sequence. Due to the short length of segments, the number of tokens is typically smaller than the model's context length. Consequently, the probability of w_i is calculated considering all previous tokens in the sequence. High perplexity suggests that the model is less likely to generate such a segment.

We used OpenAI's *GPT-2* (locally, through the *Hugging Face* platform [46,47]) to compute the perplexity of approximately 5 million segments in 9067 transcripts. Perplexity was calculated at the segment level and then aggregated within each transcript. We hypothesized that higher perplexity would reflect higher semantic complexity, thereby acting as a marker for actual sessions. However, we also anticipated that extreme values of perplexity could indicate recording errors or corrupted transcripts that do not represent a therapeutic conversation. Thus, we propose that therapeutic conversations will generally have higher average perplexities but will not show the highest maximum values. To test this assumption, we extracted the average perplexity, standard deviation of perplexity, and maximal perplexity for each transcript. Additionally, to get an accurate representation of the upper-bound perplexity values without being over-influenced by outliers, we calculated the 75th percentile of perplexity values for each transcript.

Classification—Distinguishing Sessions From Conversations

Model and Platform

To classify transcripts of behavioral treatment to sessions and non-sessions, we queried an LLM with a zero-shot approach. We used the Amazon Bedrock platform [48] that enables running closed models through a third party, ensuring that the data were not shared or exposed to the model's provider. The model selection out of the available platform's options was based on criteria of cost, context length, and proven reliability. Table S1 in Multimedia Appendix 2 presents the cost of different tokens at the time of writing. Anthropic's Claude [49] was chosen because of its proficiency in human tasks and cost-efficiency [49-51]. Both the LLM and the platform were carefully selected to ensure data security. Amazon's Bedrock platform served as a third party between the model and the user, enabling us to use the model non-locally without exposing our data. Data were not saved on the platform's servers, nor were they used for the benefit of training the model [48,52].

Filtering and Subsetting

Based on the metadata, we excluded files from organizations that are not clinical public entities, such as academic institutes conducting mock sessions and trials (25% filtered). From this subset, we randomly selected 850 transcripts for automatic classification.

Classification With LLM

"Claude-3-Sonnet" was applied through Amazon's Bedrock platform (model ID:

XSL•FO
RenderX

"anthropic.claude-3-sonnet-20240229-v1:0") with the following parameters: maximal generated words (max_out): 350; temperature: 0.3; and the remaining parameters were set to default. A detailed prompt with guidelines specifying the criteria for classifying a transcript as a session, and a format for providing a well explained answer were created to ensure the model performed the task as expected. In collaboration with an expert clinical psychologist, we compiled a list of elements that define a behavioral treatment session. The prompt addressed the following 5 core elements of a behavioral treatment session:

Dynamics: a correspondence where one of the sides shares experiences and the other focuses on listening and responding.

Content: the conversation contains personal matters, emotions and experiences, discussions about personal goals, thoughts, or behaviors.

Therapeutic elements: demonstration of active listening and empathy by one side, and use of therapeutic techniques such as reframing, providing coping strategies.

Professional language: therapeutic terminology, references for treatment plans, or previous sessions.

Context clues: mentions of confidentiality, session time limit, or scheduling future appointments.

The prompt specifically instructed the model to look for these elements in its decision-making process, explain its reasoning, and rate its certainty on a scale of 1-5. Additionally, we asked the model to provide a brief summary of the conversation content and to indicate whether it identified nontherapeutic conversational dynamics. These instructions were designed to ensure that the model addressed the entire conversation and understood its content. To maintain consistency, we used XML tags (see Multimedia Appendix 3).

Validation by Clinical Expert and Interrater Reliability

To validate the model's classifications, 2 human raters independently classified 150 randomly selected transcripts. Both raters were graduate students in psychology or cognitive sciences, pursuing either clinical or theoretical training related to psychotherapy. Both were familiar with transcripts of behavioral health sessions prior to this task.

Raters were instructed to read transcript segments (approximately 12 turns) carefully and determine whether the segment belonged to an individual treatment session, excluding family or couple sessions, peer support, and case management calls. If a decision could not be made based on the provided segment, raters were encouraged to consult the full transcript.

These exclusion criteria were not part of the model's automated classification process and may therefore have introduced some discrepancies between human and model decisions. Nevertheless, our aim was to assess whether applying instructions focused on contextual and relational dynamics of the conversation would naturally result in the exclusion of nonprofessional conversations. Interrater agreement was assessed using Cohen kappa and percent agreement. Cohen kappa higher than 0.61 is generally interpreted as a substantial agreement [53]. We then calculated the percent agreement of the human raters with the model for each category (session, non-session).

Perplexity of Sessions Versus Non-Sessions

To measure whether the 2 perplexity distributions can be assigned to 2 different distributions—sessions and non-sessions—we conducted a permutation test. Out of the transcripts for which we calculated segment perplexity, 335 transcripts were also assigned a class by the LLM: 285 sessions and 50 non-sessions.

We did not control the length of segments and their number, both are expected to affect the perplexity of a file. To ensure that the results are robust under filtering of extremely short transcripts and short segments, we calculated the same results after filtering segments with less than 5 or 10 words, and transcripts for which less than 10 or 20 segments. Finally, based on the results, we assessed the separability of sessions and non-sessions distributions by evaluating a perplexity-based classifier.

Results

Dataset Characteristics

Duplicates Identification and Language Detection

We identified 1 duplicate file and 1 empty file in the dataset. After removing these, the dataset consisted of 22,335 transcripts. The analysis revealed 18 different languages. The most common was English (98%), followed by Hebrew (0.7%).

Initial Assessment

The manual review of 100 transcripts yielded the following results: 46% of the transcripts were comprehensible and had *little to no errors*, 18% required a more thorough review for clear evaluation and 36% contained *clear errors*, which were categorized as follows (some transcripts had multiple error types): speaker identification errors (eg, confusing the therapist with the client, 42%); incomprehensible text (34%); too short to be indicative of the conversation type (22%); either non-session content or group sessions (11%) and *missing words or duplicated segments* (8%).

This preliminary evaluation found that the dataset was highly diverse, comprising the following transcript types:

1. Non-session content: The presence of non-session and group-session transcripts underscores the need for filtering mechanisms to exclude these from analysis.
2. Session quality: Based on this evaluation, we estimate that approximately half of the dataset contains sessions suitable for further analysis.
3. Speaker recognition errors: Given the frequency of speaker diarization errors, features that rely on speaker identification (eg, number of turns, spoken time by speaker) may be unreliable without corrections.
4. Incomprehensible text: This analysis underscores the importance of developing tools to improve text comprehensibility. With appropriate adjustments, more data could be rendered suitable for analysis.

XSL•FO
RenderX

Features Learning

Conversation Length

Conversations length ranged between 0.5 and 18,783 seconds, with an average of 2707 seconds and median of 3062 seconds. Figure 2 shows the full distribution of conversation length (A) and the same distribution after omitting conversations shorter than 15 minutes long (B). A Gaussian mixture model best fitted 3 Gaussian distributions for the full distribution and 2 for the reduced distribution. Figure 3 illustrates a comparison of the model's full distribution fit for different numbers of components, and Table 1 presents the values of the Akaike information criterion and the Bayesian information criterion under each parameter.

Figure 2. Comparison of transcript duration distributions before and after data cleaning. (**A**) Distribution of durations for all transcripts, including trial sessions and experimental recordings. (**B**) Distribution after excluding transcripts based on organization category metadata. The comparison highlights that many short-duration transcripts originated from irrelevant or excluded organizations.

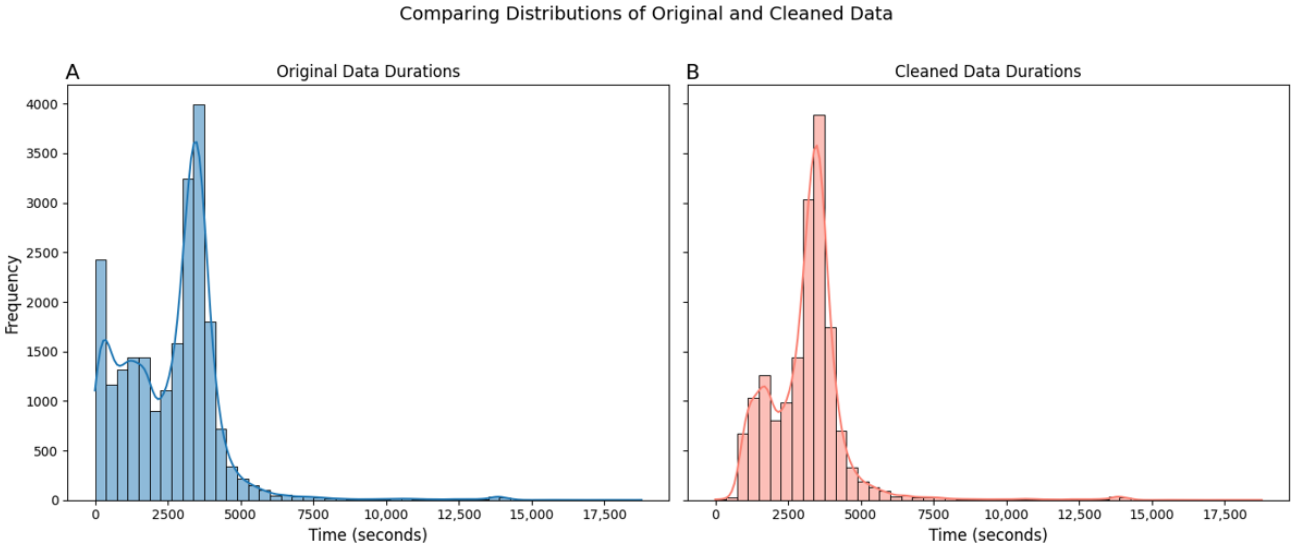

Figure 3. Comparison of Gaussian mixtures models (GMMs) with 2, 3, and 4 components fitted to the distribution of session durations (in seconds) based on transcript timestamps (Figure 2A). The negative log-likelihood is reported for each model, with lower values indicating better fit.

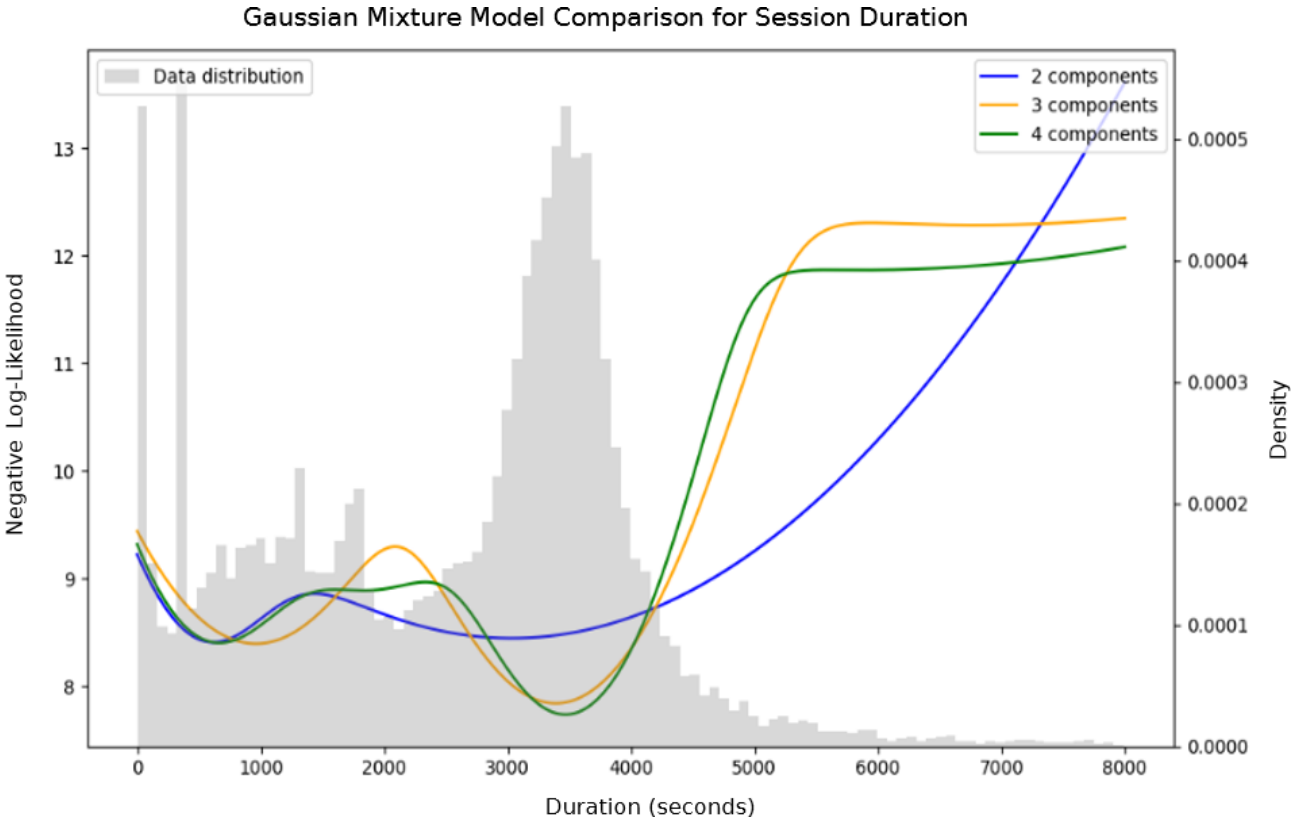

XSL•FO
RenderX

Table . Akaike information criterion and Bayesian information criterion for different number of components computed for the Gaussian mixture models presented in Figure 3.

Components	AIC[a]	BIC[b]
2	39,3727	39,3767
3	38,4168	38,4232
4	38,3384	38,3473

[a]AIC: Akaike information criterion.

[b]BIC: Bayesian information criterion.

To support our findings, we examined the metadata to determine whether different conversation types (based on organization) were associated with different transcript length. Among the short transcripts (<15 min) that included conversational content, we found brief case management phone calls. The long transcripts (>1.2 h) included 2 sessions in a row, or a recording extended before or after a session. These patterns suggest the presence of 2 conversation populations, each with a different duration distribution.

Speaking Rate

Speaking rate ranged from 0 to 4.4 WPS with an average of 2.17. The net speaking rate ranged from 0.04 to 6.63 WPS with an average of 2.9 WPS (Figure 4B). Both averages fall within the reported range of the American average speaking rate in conversation [41] (1.85 - 4.86 WPS).

The highest speaking rate was observed in a short transcript of an automatic answering machine. Only 30% (n=9) of transcripts with speaking rate above 3.5 WPS (Figure 4A) were longer than 20 minutes, suggesting that high speaking rate often reflected accidental recordings. For instance, of the 9 transcripts with rates greater than 4 WPS, 8 were identified as automatic voice answering machines.

Examining the ratio of overall session time to speech duration, or equivalently the ratio of speaking rate (A) to net speaking rate (B), shows that transcripts with high values (where total conversation time far exceeds spoken time) often reflect timestamp errors or short transcripts with very few words. In either case, such transcripts are not suitable for further analysis.

Figure 4. Distributions of average speaking rates extracted from transcripts with 2 methods: (**A**) Average number of words per second (WPS) over the whole session (including silences). (**B**) Average number of words per second for the net spoken time (without silences between segments). The red arrow's height represents the number of values exceeding the 95th percentile.

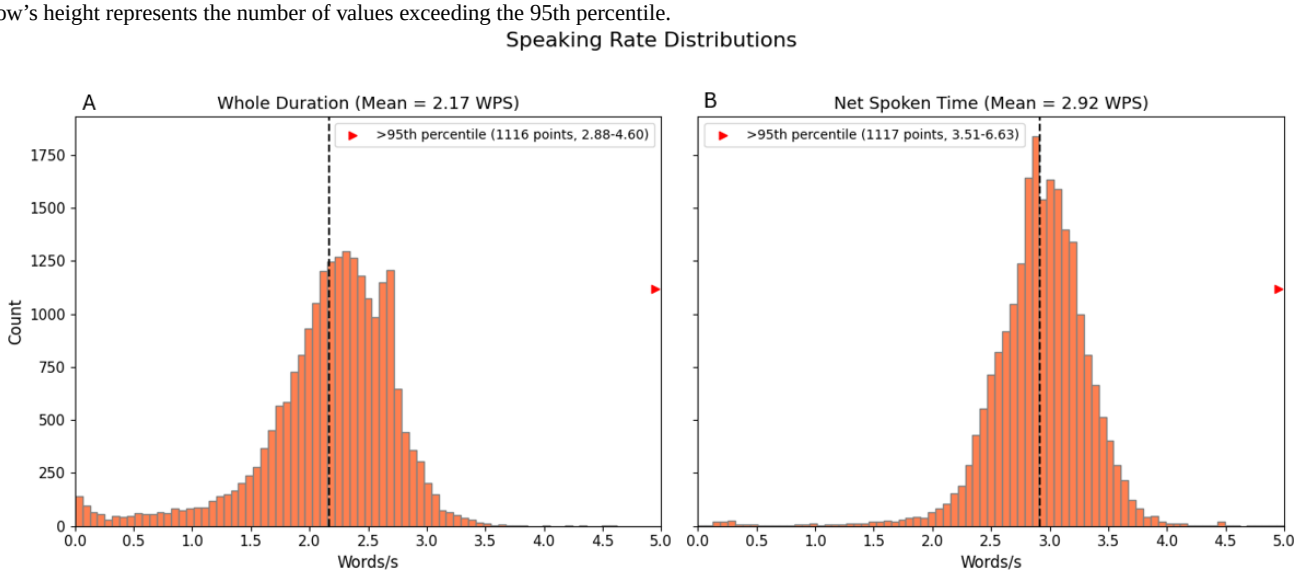

Frequent Noun Words

After filtering words according to Multimedia Appendix 1 , we identified 8,393,775 nouns in the clients' speech and 32,591 nouns in the therapists' speech. Multimedia Appendix 4 displays the 60 most common words before and after filtering. Therapists' 5 most common words were "thing," "time," "people," "know," and "way," and clients' most common nouns were similar: "thing," "time," "know," "date," and "lot" followed by "people." Among the 15 most common words for clients were also "friend," "mom," "work," "talk," and "life," and therapists shared some of these nouns ("work," "talk," and "date") but also frequently used "help" and "guy."

Classification

Zero-Shot Prompting

Of 850 transcripts that were given to the LLM, it identified 737 as sessions (86.7%, n=737), including 56 transcripts classified as couple's sessions. The remaining 113 transcripts were classified as non-sessions (13.3%, n=113). The model's certainty ratings were skewed toward the higher end of the scale, with

XSL•FO
RenderX

55 transcripts rated at 4 (7.9%) and 645 transcripts rated at 5 (92.1%, highest certainty).

Validation and Interrater Reliability

The raters agreed on 86.3% (44 of 51; see Figure 5A in Multimedia Appendix 5) of the test set transcripts, predominantly classifying transcripts as therapy sessions rather than nontherapy sessions (Rater 1: 61%; Rater 2: 63%). Cohen kappa score was 0.71, indicating substantial agreement [53]. Disagreements primarily revolved around identifying the session type—distinguishing between individual, couple, or family therapy—rather than determining whether a therapy session took place at all. In 43% (3 of 7) of the disagreement cases, both raters classified the transcript as a therapy session but differed on the type. In the remaining 57% (n=4), the disagreement was about whether the conversation was professional or casual. The full distribution of classifications by the raters and the model is shown in Multimedia Appendix 5.

LLM Versus Raters

Unlike the raters, who were instructed to mark non-individual sessions (eg, family or couple therapy) as non-sessions, the LLM was instructed to identify any kind of treatment session, including family and couple therapy. This difference naturally led to disagreements between raters and the model. However, when the model explicitly identified the session type (eg, family or couples) and indicated it clearly in its explanation or summary, we considered the classification an agreement. This verification process can be automated by searching for specific keywords (eg, “couple session”) in the summary text or by using another language model to interpret the model’s output. Table 2 shows the distribution of rater-model agreement for sessions and non-sessions.

Table . Agreement between the model and the raters for transcripts agreed over by raters[a].

	Raters: yes, n	Raters: no, n	Total, n
Model: yes, n	26	7	33
Model: no, n	2	9	11
Total, n	28	16	44

[a]The table features the number of transcripts in each category.

Overall, the model identified 57 sessions as couples or family therapy, demonstrating its ability not only to detect therapy sessions but also to categorize them by type.

Among the 9 disagreements between the model and the human raters, most involved the model misclassifying case management conversations as therapy. Conversely, in cases where the model misidentified conversations tagged as therapy sessions by human raters, its reasoning was because of “lack of therapeutic techniques.”

Perplexity of Different Classes

Figure 5 shows the distributions of different statistics of perplexity of transcripts for different classes.

XSL•FO
RenderX

Figure 5. Results of permutation tests between distributions of perplexity metrics of sessions ("Yes") and non-sessions ("No"). Perplexity metrics—maximum, average, 75th percentile, and standard deviation—were calculated over the distribution of segment perplexity for each transcript. The 75th percentile showed a significant result (mean difference = −258, *P*=.01) with sessions having lower values, while the maximum perplexity was higher for sessions (mean difference =73,888 , *P*=.007).

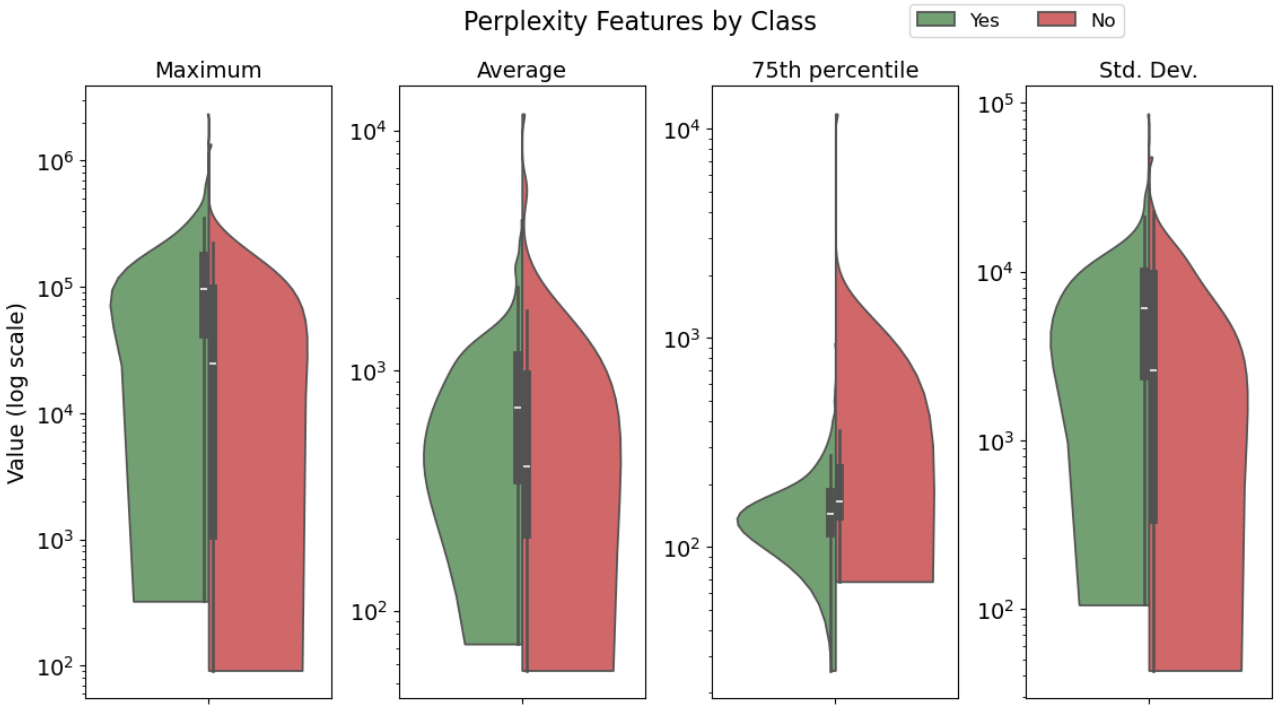

Among the transcripts with high perplexities, 1 transcript had exceptionally high values (mean perplexity of more than 11,000, which was approximately 11 standard deviations more than the average mean perplexity). This 4-segment transcript contained nonwords and backchannels that seemed to be transcribed noises (mostly the sound "Mhm"). Another high-perplexity file was a recorded rap song.

Of the 4 statistics—max, mean, standard deviation, and 75th percentile—a permutation test for means showed a significant result for the 75th percentile (*P*=.01), with sessions having a lower mean 75th percentile than non-sessions (mean difference=−258). In contrast, the permutation test showed that the sessions group had higher maximal values (*P*=.007).

To check the robustness of these results, we repeated the test after limiting the calculation to segments with minimal number of words (minimum words per segment [MWPS]=5, 10) and omitting transcripts with fewer than a minimal number of relevant segments (minimal number of segments [MS]=10, 20). These tests showed that the 75th percentile measure (Figure 6A) remained significant under these parameter changes, whereas the max perplexity measure (Figure 6B) lost significance when limitations were applied (MS>0 or MWPS >0).

XSL•FO
RenderX

Figure 6. *P* values of permutation tests comparing sessions and non-sessions perplexity measures across parameter settings—minimum number of segments and minimal number of words per segment. (**A**) Maximum perplexity shows no significant values across parameters except when no restrictions are applied and (**B**) 75th percentile of perplexity is significant for most parameter combinations. Bright colors indicate lower *P* values, **P*<.05, ***P*<.01.

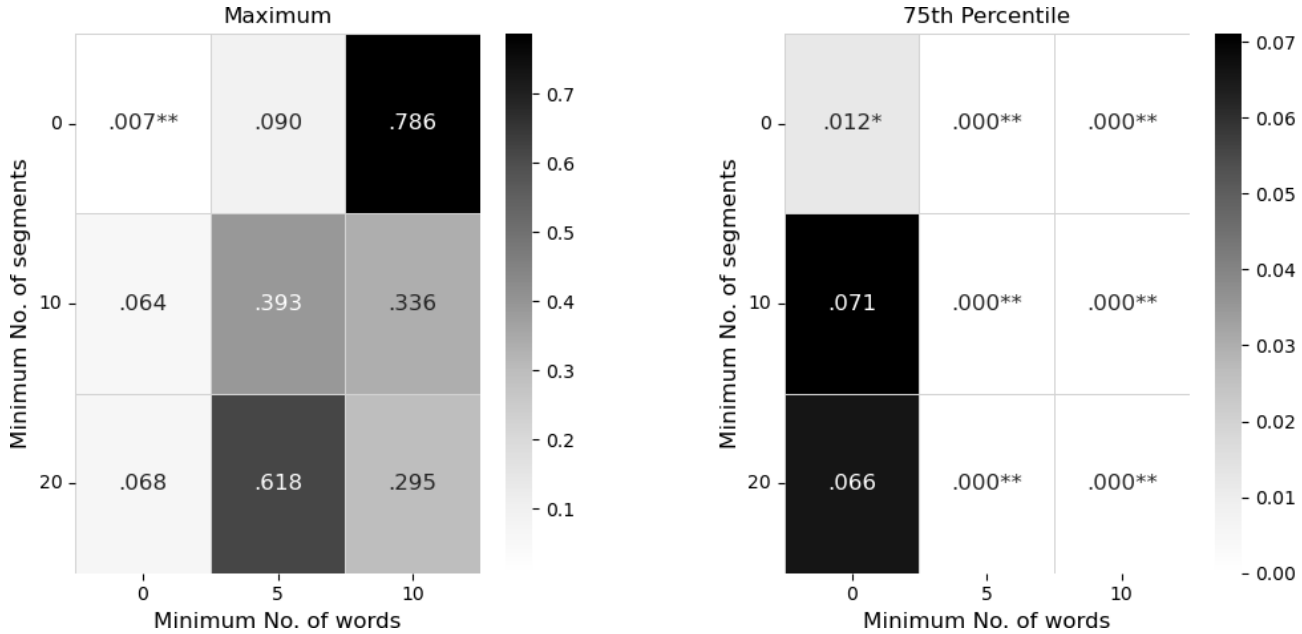

Furthermore, we evaluated the classification performance of a 75th percentile perplexity–based classifier by analyzing receiver operating characteristic (ROC) curves and precision-recall metrics across parameters (MWPS, MS). To align with perplexity values, in this analysis, higher perplexity values correspond to the positive class—non-sessions—while sessions constitute the negative class. The optimal threshold was determined by maximizing the difference between true positive rate and false positive rate, prioritizing accurate classification of non-sessions. The results (Table 3) indicate moderate discriminative ability, with ROC area under the curve values ranging from approximately 0.62 to 0.73 and precision increasing alongside minimal number of segments. Precision tends to increase with larger segment sizes (up to ~0.63), while recall decreases (~0.26).

Table . Classification performance of the 75th percentile perplexity–based filter across varying thresholds for minimal words per segment (MWPS) and minimal number of segments (MS).

MWPS	MS	Sessions	Non-sessions	ROC AUC	PR AUC	Precision	Recall
0	0	285	50	0.618	0.240	0.200	0.740
0	5	285	43	0.636	0.198	0.182	0.767
0	10	284	42	0.633	0.196	0.178	0.762
5	0	285	47	0.664	0.354	0.353	0.255
5	5	284	41	0.657	0.331	0.333	0.268
5	10	284	40	0.665	0.335	0.333	0.275
10	0	285	44	0.687	0.378	0.625	0.227
10	5	284	40	0.701	0.380	0.625	0.250
10	10	284	38	0.726	0.392	0.625	0.263

[a]The table reports the number of transcripts remaining in each class after filtering outliers, the receiver operating characteristic area under the curve (ROC AUC), precision-recall AUC (PR AUC), precision, recall, and F1 scores. Increasing the minimal words and segments thresholds improves ROC AUC and precision but reduces recall.

Discussion

Principal Findings

With a dataset of over 22,000 unfiltered transcripts of recordings associated with behavioral treatment sessions, our primary objective in this methodological paper was to illustrate a systematic approach for characterizing the data and implementing a filtering process for subsequent analysis in academic research. Prior research on leveraging machine learning and natural language processing methods for classifying large text datasets, both conversational [54] and nonconversational [55,56], has focused mostly on content-based classification rather than the contextual framework of conversations. Our proposed methodology integrated human evaluation, statistical analysis, and automated tools, including LLMs, emphasizing the importance of contextual and relational

XSL•FO
RenderX

features, which are especially critical for distinguishing therapy sessions from non-sessions.

The preliminary analysis identified a significant proportion of non-session transcripts within the dataset. These non-sessions encompassed brief case management encounters, informal conversations, mock sessions, and incidental recordings captured between sessions. Additionally, some transcripts included errors such as incorrect speaker identification, text duplication, and incomprehensible content [25]. These findings highlight the necessity for systematic filtering mechanisms to exclude low-quality and non-session data. Given that only approximately half of the dataset was found suitable for analysis, leveraging automated classification and quality scoring can significantly enhance the dataset's utility for research purposes.

Recent works have listed features for characterizing texts in corpora [35,57]. In this paper, we have focused on 3 basic ways to extract relevant statistics and examined if outliers could be indicative of problematic transcripts:

Conversation length—Analysis of conversation length distributions revealed that shorter transcripts often represented non-sessions, such as case management encounters, as expected for this type of service, phone calls, or noise. In contrast, unexpectedly long conversations may result from sessions being concatenated or recordings extending beyond the session's actual duration, thus requiring careful preprocessing.

Speaking rate—Speaking rate has been evaluated in many contexts, and it has been shown that its values can predict speakers' features [41], detect speech anomalies, and identify change in conversation dynamics or context [42]. We found that high speaking rate was associated with shorter transcripts, particularly erroneous recordings such as answering machine messages. This finding supports the use of speaking rate as a potential indicator of non-sessions. Additionally, the ratio between speech duration and overall recording duration may serve as a marker for time-labeling errors.

Content analysis—Bag-of-Words is a basic tool for evaluating semantic content. Through word count (mostly nouns), themes and topics can be discovered and ultimately enhance text classification [58-60]. In this paper, we used it as a "reality check" to reveal the contents of the transcripts at hand. Our content analysis provided insights into the vocabulary and topics typical of therapeutic conversations. The most prevalent words were related to everyday life ("relationship," "house," "job," "today," "school," "sleep," and "car"), clients' inner world ("thought," "want," "feel," and "need"), resolving issues ("situation," "problem," "help," and "care"), and relationships ("kid," "love," "date," "sister," "couple," "dad," and "guy"), emphasizing the centrality of these themes in these conversations. Comparing individual transcript themes against the topics that were found in this analysis could be an indicator for being a genuine session. However, it is of note that this analysis ignores the context in which words appear.

Perplexity Analysis

Perplexity is traditionally used to evaluate language models [61,62]. In this study, however, we used it to evaluate the comprehensibility and uniqueness of text segments [63,64] in order to gain insight into potential differences between sessions and non-sessions. Transcripts with higher perplexity scores often contained transcription errors or nonverbal content (eg, noise and backchannels).

Sessions initially exhibited higher maximal transcript perplexity than non-sessions; however, this result was no longer significant when outliers, namely, transcripts with few segments or short segments, were omitted. This suggests that the higher maximal perplexity values in sessions are not necessarily indicative of more verbally complex content but rather stem from transcription errors, backchanneling, discourse markers, or interrupted recordings, which tend to produce short segments and short transcripts. These findings indicate that sessions, in general, do not have higher maximal perplexity than non-sessions. This interpretation is further supported by the finding that the 75th percentile perplexity score for sessions was lower than non-sessions after removing outliers. This may suggest that sessions contain more structured and predictable language, particularly when compared to background noise or accidentally captured conversation fragments. This distinction is particularly useful because it suggests that 75th percentile perplexity could serve as a reliable feature in a non-session filtering model. Unlike maximal perplexity, which was influenced by errors and outliers, the 75th percentile measure remains stable across different parameter settings, reinforcing its robustness as an indicator of session-like structure.

To support this claim and further evaluate its ability to differentiate sessions from non-sessions, we assessed a 75th percentile perplexity–based classifier across varying thresholds for MS and MWPS. The optimal threshold remained consistent across conditions, suggesting its practical applicability as a filtering criterion. However, the results indicate moderate-to-low discriminative performance, with precision improving under stricter outlier exclusion. For instance, removing transcripts with fewer than 10 segments of more than 10 words each yielded the highest precision (~0.63) but substantially reduced recall (~0.26). This pattern likely reflects the dataset imbalance—where non-sessions are the minority—and the observation that outliers were predominantly non-sessions, suggesting they tend to exhibit relatively high 75th percentile perplexity. These findings reinforce the conclusion that this metric is particularly useful when filtering aims to prioritize precision, though caution is needed due to the class imbalance impacting these metrics.

Overall, our results suggest that perplexity can be used to identify flawed, incomprehensible, or highly improbable text segments and may serve as a useful tool for detecting low-quality or non-session transcripts, but not as a standalone classifier. To validate our results and extend this work, we propose 2 key next steps: (1) to distinguish nonprofessional conversations from flawed transcripts to determine which group is more clearly separable from sessions through perplexity measures. This distinction is critical for understanding the sources of high- and low-perplexity segments. (2) To replicate these analyses with a larger set of labeled transcripts to overcome data imbalance.

XSL•FO
RenderX

Large Language Model Classification

Research on LLMs has explored their use for both text classification and mental health text analysis as separate tasks. Prior work has demonstrated these models' ability to extract key concepts, analyze text dynamics, and identify psychological concepts [14,65-69]. In this study, we integrated both tasks, leveraging zero-shot prompting to analyze mental health conversations with the goal of classification. Our findings indicate that with zero-shot prompting, the model can classify transcripts effectively, showing high agreement with human coders while demonstrating robustness in handling transcript errors. For example, even when speaker identification referenced only a therapist and a patient, the model correctly identified couple or family sessions by interpreting contextual and semantic cues. Additionally, the LLM successfully corrected speaker identification errors, highlighting its potential as an automated error-correction tool.

Instances of disagreement between human coders and the LLM often stemmed from the model's sensitivity to therapeutic techniques. One of the classification guidelines was the presence of such techniques, and the model appeared to weigh them heavily when the conversational framing was ambiguous. Furthermore, in some cases, we found discrepancies between the binary classification and the LLM's explanation. For instance, despite classifying a conversation as a session, the model noted: *The therapist and psychologist discuss updates on multiple client cases, including challenges with clients' behaviors*, recognizing both speakers as therapists. This example underscores the importance of including model explanations as part of the classification process. Future work should investigate these discrepancies and refine classification guidelines accordingly.

Finally, while research on LLMs' ability to analyze conversations through prompting is still evolving, existing studies have yielded inconsistent findings regarding their effectiveness in analyzing conversational contexts [14,65-69]. Our results suggest that when provided with full conversation transcripts, LLMs can capture nuanced textual and relational dynamics, offering valuable insights into participants' interactions. While recent studies have used these capabilities in text-based applications, we demonstrated their applicability to conversations, where relational and dynamic elements are crucial in distinguishing sessions from non-sessions. Thus, although few-shot learning or fine-tuning may further enhance classification accuracy [55,56], our findings suggest that these techniques may not be strictly necessary for effective classification.

Limitations

Defining when a transcript reflects a treatment session presents a challenge: sessions share many aspects with nontherapeutic conversations, and some therapeutic techniques might appear as trivial exchanges. In this paper, we proposed both statistical heuristics and explicit guidelines to evaluate conversations in light of this question.

While we offer a methodological examination of a transcript dataset and highlight considerations for careful filtering, defining a filtering model is beyond the scope of this study. Since ground-truth labels were unavailable for supervised classification, we address this by identifying features that provide valuable insights into the classification process and by demonstrating how LLMs can assist experts when guided by well-designed prompt engineering. Hence, the approach presented here may serve as a foundation for developing a comprehensive classification model.

This study delineates a general framework and provides guidelines for working with conversational datasets in the absence of prior knowledge or structured labels. However, variation exists across datasets in format, content, quality, and the proportion of non-session files and transcription errors. While this paper applies its suggested framework on a large and varied dataset, it still features working with a single data platform. Variation in platforms and datasets calls for researchers to (1) detect any peculiarities relevant to their datasets and acknowledge them before referring to the relevant steps in the framework and (2) cultivate content matter expertise for datasets they analyze or better—conduct studies in multidisciplinary teams and keep a human (expert) in the loop.

Future Research

Our findings suggest that while statistical features can provide broad insights into conversational structure and flag irregularities that may correlate with non-sessions, LLMs add a complementary layer by capturing the essence, dynamics, and nature of conversations. As a proof of concept, this study demonstrates the potential of LLM-based classification; however, further validation, including human annotation on larger samples, is needed. In a treatment setting, we envision a semiautomated pipeline that integrates statistical and semantic features with LLMs. Transcripts would first be screened using metrics such as perplexity, duration, and speaking rate; those within a normal range would then be analyzed by an LLM to classify the conversation type—treatment session or unrelated. Periodic human review of selected transcripts would ensure model validity, with reviewer feedback used to refine prompts or feature thresholds dynamically. This "human-in-the-loop" workflow enables continuous processing while limiting manual review to a small subset of cases. Future work should evaluate different LLM models across diverse behavioral health datasets to determine their ability to capture nuances in treatment styles.

Regarding statistical feature–based outliers filtering, additional research could enhance its efficiency and interpretability. Incorporating more informative linguistic features—such as the frequency of backchanneling cues, discourse markers, and silent pauses—into interpretable models such as decision trees may improve classification while remaining computationally efficient compared to fine-tuning large-scale models. Moreover, content analysis could reveal how the themes extracted from an individual session differ from those found across the dataset, highlighting deviations that might indicate non-session transcripts. Examining bigrams and multi-word expressions could also refine differentiation between informal conversations and structured therapeutic sessions, while helping to detect common transcription errors, such as misinterpretations caused by background noise.

XSL•FO
RenderX

Perplexity analyses could examine the relationship between different segment error types, such as misspelling, and perplexity by categorizing segments accordingly. Additionally, applying a mixed-model analysis to account for the statistical dependencies of perplexities within individual transcripts. Moreover, the interaction between perplexity and factors such as speaker familiarity or amorphous nature of the conversation remains an open question and could help explain cases where high perplexity signals meaningful conversational ambiguity rather than transcription errors. Finally, improving transcript quality remains a critical avenue for future research. Existing methods for correcting speaker diarization errors and ASR mistakes without access to the original audio [70,71] could be integrated into the preprocessing pipeline to enhance transcript reliability before filtering. Implementing these techniques could enhance the preprocessing stage, improve the accuracy of extracted features, and ultimately enhance the filtering process and its success rates.

Conclusion

This study demonstrated the importance of integrating human judgment with automated tools when processing large, unstructured datasets. We assess secondary data—data collected independently of current research—where initial human evaluation is critical for understanding dataset characteristics such as readability, content, and diarization quality. This foundational knowledge can inform the development of effective filtering strategies. While basic statistics, perplexity, and LLM prompting facilitate automated filtering, preliminary human review remains essential for understanding dataset variability and refining classification features. This hybrid approach ensures adaptable and accurate filtering processes, even in the presence of transcription errors.

Disclaimer

This manuscript was partially edited for language and style with the assistance of the AI language models ChatGPT (OpenAI) and Claude (Anthropic). The authors reviewed and edited all AI-generated content and take full responsibility for the final text.

Data Availability

The dataset generated and analyzed during the current study is not publicly available due to privacy and confidentiality reasons.

Conflicts of Interest

SSS, SJ, and ES are employees of Eleos Health whose artificial intelligence platform was used to generate the data analyzed in this study. DPM is the Founder of Morrison Consulting, which provides consulting services to Eleos Health, including his role as Chief Clinical Officer; he is affiliated with Eleos Health in a consulting capacity, but not as an employee. PMN and AG declare no conflicts of interest.

Multimedia Appendix 1
Clean-up words.
[PNG File, 122 KB - formative_v9i1e78082_app1.png]

Multimedia Appendix 2
Amazon Bedrock models prices.
[PNG File, 36 KB - formative_v9i1e78082_app2.png]

Multimedia Appendix 3
Prompt.
[DOCX File, 16 KB - formative_v9i1e78082_app3.docx]

Multimedia Appendix 4
Most common words: before and after filtering.
[PNG File, 403 KB - formative_v9i1e78082_app4.png]

Multimedia Appendix 5
Session versus non-session classification results: (A) inter-rater answers distribution and (B) distribution of agreement between the model and the human raters.
[PNG File, 101 KB - formative_v9i1e78082_app5.png]

References

1. Whalen J, Raymond GT. Conversation analysis. In: Borgatta EF, Montgomery RJV, editors. Encyclopedia of Sociology, 2nd edition: Macmillan Reference USA; 2000.

XSL•FO
RenderX

2. Drew P, Chatwin J, Collins S. Conversation analysis: a method for research into interactions between patients and health-care professionals. Health Expect 2001 Mar;4(1):58-70. [doi: 10.1046/j.1369-6513.2001.00125.x] [Medline: 11286600]
3. Park J, Kotzias D, Kuo P, et al. Detecting conversation topics in primary care office visits from transcripts of patient-provider interactions. J Am Med Inform Assoc 2019 Dec 1;26(12):1493-1504. [doi: 10.1093/jamia/ocz140] [Medline: 31532490]
4. Fraser H. A framework for deciding how to create and evaluate transcripts for forensic and other purposes. Front Commun 2022;7:898410. [doi: 10.3389/fcomm.2022.898410]
5. Følstad A, Taylor C. Investigating the user experience of customer service chatbot interaction: a framework for qualitative analysis of chatbot dialogues. Qual User Exp 2021 Dec;6(1):6. [doi: 10.1007/s41233-021-00046-5]
6. Jeong M, Minson J, Yeomans M, Gino F. Communicating with warmth in distributive negotiations is surprisingly counterproductive. Manage Sci 2019 Dec;65(12):5813-5837. [doi: 10.1287/mnsc.2018.3199]
7. Hennessy S, Calcagni E, Leung A, Mercer N. An analysis of the forms of teacher-student dialogue that are most productive for learning. Language and Education 2023 Mar 4;37(2):186-211. [doi: 10.1080/09500782.2021.1956943]
8. Goldberg SB, Flemotomos N, Martinez VR, et al. Machine learning and natural language processing in psychotherapy research: alliance as example use case. J Couns Psychol 2020 Jul;67(4):438-448. [doi: 10.1037/cou0000382] [Medline: 32614225]
9. Spinrad A, Taylor CB, Ruzek JI, et al. Action recommendations review in community-based therapy and depression and anxiety outcomes: a machine learning approach. BMC Psychiatry 2024 Feb 16;24(1):133. [doi: 10.1186/s12888-024-05570-0] [Medline: 38365635]
10. Atzil-Slonim D, Eliassaf A, Warikoo N, et al. Leveraging natural language processing to study emotional coherence in psychotherapy. Psychotherapy (Chic) 2024 Mar;61(1):82-92. [doi: 10.1037/pst0000517] [Medline: 38236227]
11. Sadeh-Sharvit S, Hollon SD. Leveraging the power of nondisruptive technologies to optimize mental health treatment: case study. JMIR Ment Health 2020 Nov 26;7(11):e20646. [doi: 10.2196/20646] [Medline: 33242025]
12. Sadeh-Sharvit S, Camp TD, Horton SE, et al. Effects of an artificial intelligence platform for behavioral interventions on depression and anxiety symptoms: randomized clinical trial. J Med Internet Res 2023 Jul 10;25:e46781. [doi: 10.2196/46781] [Medline: 37428547]
13. Imel ZE, Pace BT, Soma CS, et al. Design feasibility of an automated, machine-learning based feedback system for motivational interviewing. Psychotherapy (Chic) 2019 Jun;56(2):318-328. [doi: 10.1037/pst0000221] [Medline: 30958018]
14. Abdou M, Sahi RS, Hull TD, Nook EC, Daw ND. Leveraging large language models to estimate clinically relevant psychological constructs in psychotherapy transcripts. . Preprint posted online on 2025. [doi: 10.1101/2025.03.04.25323338]
15. Ruzek JI, Sadeh-Sharvit S, Bunge EL, et al. Training the psychologist of the future in the use of digital mental health technologies. Prof Psychol: Res Pract 2024;55(5):395-404. [doi: 10.1037/pro0000567]
16. Flaherty HB. Teaching note—using technology to enhance experiential learning through simulated role plays. J Soc Work Educ 2023 Oct 2;59(4):1294-1300. [doi: 10.1080/10437797.2022.2050869]
17. Sadeh-Sharvit S, Rego SA, Jefroykin S, Peretz G, Kupershmidt T. A comparison between clinical guidelines and real-world treatment data in examining the use of session summaries: retrospective study. JMIR Form Res 2022 Aug 16;6(8):e39846. [doi: 10.2196/39846] [Medline: 35972782]
18. Shapira N, Atzil-Slonim D, Tuval Mashiach R, Shapira O. Measuring linguistic synchrony in psychotherapy. Presented at: Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology; Jul 14-15, 2022; Seattle, USA. [doi: 10.18653/v1/2022.clpsych-1.14]
19. Atzil-Slonim D, Soma CS, Zhang X, Paz A, Imel ZE. Facilitating dyadic synchrony in psychotherapy sessions: systematic review and meta-analysis. Psychother Res 2023 Sep;33(7):898-917. [doi: 10.1080/10503307.2023.2191803] [Medline: 37001119]
20. Yonatan-Leus R, Gwertzman G, Tishby O. Using machine learning methods to identify trajectories of change and predict responders and non-responders to short-term dynamic therapy. Psychother Res 2025 Sep;35(7):1070-1086. [doi: 10.1080/10503307.2024.2420725] [Medline: 39461002]
21. Atzil-Slonim D, Juravski D, Bar-Kalifa E, et al. Using topic models to identify clients' functioning levels and alliance ruptures in psychotherapy. Psychotherapy (Chic) 2021 Jun;58(2):324-339. [doi: 10.1037/pst0000362] [Medline: 33734743]
22. Althoff T, Clark K, Leskovec J. Large-scale analysis of counseling conversations: an application of natural language processing to mental health. Trans Assoc Comput Linguist 2016;4:463-476. [Medline: 28344978]
23. Radford A, Kim JW, Xu T, Brockman G, Mcleavey C, Sutskever I. Robust speech recognition via large-scale weak supervision. In: Krause A, Brunskill E, Cho K, Engelhardt B, Sabato S, Scarlett J, editors. Presented at: Proceedings of the 40th International Conference on Machine Learning; Jul 23-29, 2023 p. 28492-28518.
24. Bain M, Huh J, Han T, Zisserman A. WhisperX: time-accurate speech transcription of long-form audio. Presented at: INTERSPEECH 2023; Aug 20-24, 2023; Dublin, Ireland p. 4489-4493. [doi: 10.21437/Interspeech.2023-78]
25. Miner AS, Haque A, Fries JA, et al. Assessing the accuracy of automatic speech recognition for psychotherapy. NPJ Digit Med 2020;3(1):82. [doi: 10.1038/s41746-020-0285-8] [Medline: 32550644]
26. Zolnoori M, Vergez S, Kostic Z, et al. Audio recording patient-nurse verbal communications in home health care settings: pilot feasibility and usability study. JMIR Hum Factors 2022 May 11;9(2):e35325. [doi: 10.2196/35325] [Medline: 35544296]

XSL•FO
RenderX

27. Graham C, Roll N. Evaluating OpenAI's Whisper ASR: performance analysis across diverse accents and speaker traits. JASA Express Lett 2024 Feb 1;4(2):025206. [doi: 10.1121/10.0024876] [Medline: 38391582]
28. Park TJ, Kanda N, Dimitriadis D, Han KJ, Watanabe S, Narayanan S. A review of speaker diarization: recent advances with deep learning. Comput Speech Lang 2022 Mar;72:101317. [doi: 10.1016/j.csl.2021.101317]
29. Church K, Zhu W, Vopicka J, Pelecanos J, Dimitriadis D, Fousek P. Speaker diarization: a perspective on challenges and opportunities from theory to practice. Presented at: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); Mar 5-9, 2017; New Orleans, LA p. 4950-4954. [doi: 10.1109/ICASSP.2017.7953098]
30. Hao T, Huang Z, Liang L, Weng H, Tang B. Health natural language processing: methodology development and applications. JMIR Med Inform 2021 Oct 21;9(10):e23898. [doi: 10.2196/23898] [Medline: 34673533]
31. Elbattah M, Arnaud É, Gignon M, Dequen G. The role of text analytics in healthcare: a review of recent developments and applications. Presented at: Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies; Feb 11-13, 2021; Vienna, Austria. [doi: 10.5220/0010414508250832]
32. Bazoge A, Morin E, Daille B, Gourraud PA. Applying natural language processing to textual data from clinical data warehouses: systematic review. JMIR Med Inform 2023 Dec 15;11:e42477. [doi: 10.2196/42477] [Medline: 38100200]
33. Patel KN, Kiran P. Preprocessing methods for unstructured healthcare text data. IJITEE 2019;9(2S):715-719. [doi: 10.35940/ijitee.B1024.1292S19]
34. Sedlakova J, Daniore P, Horn Wintsch A, et al. Challenges and best practices for digital unstructured data enrichment in health research: a systematic narrative review. PLOS Digit Health 2023 Oct;2(10):e0000347. [doi: 10.1371/journal.pdig.0000347] [Medline: 37819910]
35. Yeomans M, Boland FK, Collins HK, Abi-Esber N, Brooks AW. A practical guide to conversation research: how to study what people say to each other. Adv Meth Pract Psychol Sci 2023 Oct;6(4). [doi: 10.1177/25152459231183919]
36. Atkins DC, Steyvers M, Imel ZE, Smyth P. Scaling up the evaluation of psychotherapy: evaluating motivational interviewing fidelity via statistical text classification. Implement Sci 2014 Apr 24;9(1):49. [doi: 10.1186/1748-5908-9-49] [Medline: 24758152]
37. Koole SL, Tschacher W. Synchrony in psychotherapy: a review and an integrative framework for the therapeutic alliance. Front Psychol 2016;7:862. [doi: 10.3389/fpsyg.2016.00862] [Medline: 27378968]
38. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol 2008 Apr;61(4):344-349. [doi: 10.1016/j.jclinepi.2007.11.008]
39. difflib—helpers for computing deltas. URL: https://docs.python.org/3/library/difflib.html [accessed 2024-12-25]
40. Langdetect. 2021 Dec. URL: https://pypi.org/project/langdetect/ [accessed 2024-12-25]
41. Yuan J, Liberman M, Cieri C. Towards an integrated understanding of speaking rate in conversation. Presented at: INTERSPEECH 2006; Sep 17-21, 2006; Pittsburgh, PA, USA URL: https://www.isca-archive.org/interspeech_2006 [doi: 10.21437/Interspeech.2006-204]
42. Wardle M, Cederbaum K, de Wit H. Quantifying talk: developing reliable measures of verbal productivity. Behav Res Methods 2011 Mar;43(1):168-178. [doi: 10.3758/s13428-010-0019-y] [Medline: 21287128]
43. Bird S. NLTK: the natural language toolkit. Presented at: Proceedings of the COLING/ACL on Interactive Presentation Sessions; Jul 17-18, 2006; Sydney, Australia URL: https://aclanthology.org/P06-4000/ [accessed 2025-01-05]
44. Honnibal M, Montani I, Landeghem S, Boyd A. SpaCy: industrial-strength natural language processing in python. . [doi: 10.5281/zenodo.1212303] [Medline: 7069443]
45. Colla D, Delsanto M, Agosto M, Vitiello B, Radicioni DP. Semantic coherence markers: the contribution of perplexity metrics. Artif Intell Med 2022 Dec;134:102393. [doi: 10.1016/j.artmed.2022.102393]
46. Wolf T, Debut L, Sanh V, et al. HuggingFace's transformers: state-of-the-art natural language processing. . Preprint posted online on 2019 URL: https://arxiv.org/abs/1910.03771 [accessed 2024-12-26]
47. HuggingFace GPT2. URL: https://huggingface.co/gpt2 [accessed 2024-12-26]
48. Amazon's Bedrock docs. URL: https://docs.aws.amazon.com/bedrock [accessed 2024-12-26]
49. Anthropic News. Introducing Claude 3.5 Sonnet. Anthropic News. 2024 Jun 21. URL: https://www.anthropic.com/news/claude-3-5-sonnet [accessed 2025-03-05]
50. Anthropic. claude 3.5 sonnet model card addendum. 2024 URL: https://www-cdn.anthropic.com/fed9cc193a14b84131812372d8d5857f8f304c52/Model_Card_Claude_3_Addendum.pdf [accessed 2025-03-05]
51. Jin H, Guo J, Lin Q, Wu S, Hu W, Li X. Comparative study of Claude 3.5-Sonnet and human physicians in generating discharge summaries for patients with renal insufficiency: assessment of efficiency, accuracy, and quality. Front Digit Health 2024;6:1456911. [doi: 10.3389/fdgth.2024.1456911] [Medline: 39703756]
52. Amazon's Bedrock Data Security. URL: https://docs.aws.amazon.com/bedrock/latest/userguide/data-protection.html [accessed 2025-02-15]
53. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159. [doi: 10.2307/2529310]

XSL•FO
RenderX

54. Rathor S, Jadon RS. Domain classification of textual conversation using machine learning approach. Presented at: 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT); Jul 10-12, 2018; Bangalore p. 1-7. [doi: 10.1109/ICCCNT.2018.8494197]
55. Hopkins G, Kalm K. Classifying complex documents: comparing bespoke solutions to large language models. . Preprint posted online on Dec 12, 2023 URL: https://arxiv.org/abs/2312.07182 [accessed 2025-01-07]
56. Edwards A, Camacho-Collados J. Language models for text classification: is in-context learning enough. 2024 Mar 26 Presented at: URL: https://aclanthology.org/2024.lrec-main.879/ [accessed 2025-03-04]
57. Guthrie D, Guthrie L, Wilks Y. An unsupervised probabilistic approach for the detection of outliers in corpora. Presented at: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08); Mar 28-30, 2008; Marrakech, Morocco URL: https://aclanthology.org/L08-1109/ [accessed 2025-03-04]
58. Yan D, Li K, Gu S, Yang L. Network-based Bag-of-Words model for text classification. IEEE Access 2020;8:82641-82652. [doi: 10.1109/ACCESS.2020.2991074]
59. Lubis DH, Sawaluddin S, Candra A. Machine learning model for language classification: Bag-of-Words and multilayer perceptron. JITE 2023;7(1):356-365. [doi: 10.31289/jite.v7i1.10114]
60. Juluru K, Shih HH, Keshava Murthy KN, Elnajjar P. Bag-of-Words technique in natural language processing: a primer for radiologists. Radiographics 2021;41(5):1420-1426. [doi: 10.1148/rg.2021210025] [Medline: 34388050]
61. Chelba C, Mikolov T, Schuster M, et al. One billion word benchmark for measuring progress in statistical language modeling. Presented at: Interspeech 2014; Sep 14-18, 2014; Singapore p. 2635-2639. [doi: 10.21437/Interspeech.2014-564]
62. Jozefowicz R, Vinyals O, Schuster M, Shazeer N, Wu Y. Exploring the limits of language modeling. . Preprint posted online on Feb 7, 2016 URL: https://arxiv.org/pdf/1602.02410 [accessed 2025-01-12]
63. Miaschi A, Alzetta C, Brunato D, Dell'Orletta F, Venturi G. Is neural language model perplexity related to readability. Presented at: Seventh Italian Conference on Computational Linguistics; Mar 1-3, 2021; Milan, Italy URL: https://ceur-ws.org/Vol-2769/paper_57.pdf [accessed 2025-01-12]
64. Guo Y, August T, Leroy G, Cohen T. APPLS: evaluating evaluation metrics for plain language summarization. . Preprint posted online on May 23, 2023 URL: https://arxiv.org/abs/2305.14341 [accessed 2025-01-12]
65. Ma J, Na H, Wang Z, et al. Detecting conversational mental manipulation with intent-aware prompting. In: Rambow O, Wanner L, Apidianaki M, Al-Khalifa H, Di EB, Schockaert S, editors. Presented at: Proceedings of the 31st International Conference on Computational Linguistics Association for Computational Linguistics; Jan 19-24, 2025; Abu Dhabi, United Arab Emirates p. 9176-9183 URL: https://aclanthology.org/2025.coling-main.616/ [accessed 2025-3-5]
66. Yang K, Ji S, Zhang T, Xie Q, Kuang Z, Ananiadou S. Towards interpretable mental health analysis with large language models. In: Bouamor H, Pino J, Bali K, editors. Presented at: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing; Dec 6-10, 2023; Singapore p. 6056-6077. [doi: 10.18653/v1/2023.emnlp-main.370]
67. Zhang X, Yu H, Li Y, Wang M, Chen L, Huang F. The imperative of conversation analysis in the era of llms: a survey of tasks, techniques, and trends. . Preprint posted online on Sep 21, 2024 URL: https://arxiv.org/abs/2409.14195 [accessed 2025-3-5]
68. Lee Y, Goldwasser D, Reese LS. Towards understanding counseling conversations: domain knowledge and large language models. . Preprint posted online on Feb 21, 2024 URL: https://arxiv.org/abs/2402.14200 [accessed 2025-3-5]
69. Long Y, Luo H, Zhang Y. Evaluating large language models in analysing classroom dialogue. NPJ Sci Learn 2024 Oct 3;9(1):60. [doi: 10.1038/s41539-024-00273-3] [Medline: 39358390]
70. Si M, Cobas O, Fababeir M. Lexical error guard: leveraging large language models for enhanced ASR error correction. MAKE 2024;6(4):2435-2446. [doi: 10.3390/make6040120]
71. Cheng L, Zheng S, Zhang Q, et al. Improving speaker diarization using semantic information: joint pairwise constraints propagation. . Preprint posted online on Sep 19, 2023 URL: https://arxiv.org/abs/2309.10456 [accessed 2025-1-7]

Abbreviations

AI: artificial intelligence
ASR: automatic speech recognition
LLM: large language model
MS: minimal number of segments
MWPS: minimum words per segment
ROC: receiver operating characteristic
WPS: words per second

XSL•FO
RenderX

Edited by A Mavragani; submitted 27.05.25; peer-reviewed by A Adepoju, M Elbattah, S Biswas; revised version received 03.09.25; accepted 09.09.25; published 24.10.25.

Please cite as:
Naim PM, Sadeh-Sharvit S, Jefroykin S, Silber E, Morrison DP, Goldstein A
Preprocessing Large-Scale Conversational Datasets: A Framework and Its Application to Behavioral Health Transcripts
JMIR Form Res 2025;9:e78082
URL: https://formative.jmir.org/2025/1/e78082
doi:10.2196/78082

© Paz Mor Naim, Shiri Sadeh-Sharvit, Samuel Jefroykin, Eddie Silber, Dennis P Morrison, Ariel Goldstein. Originally published in JMIR Formative Research (https://formative.jmir.org), 24.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Urban Environment and Outdoor Walking Mobility Among Older Adults With and Without Cognitive Impairment in Singapore: Formative Study

Yuezhong Liu[1], PhD; Chek Hooi Wong[2], MBBS, MPH; Noam Shoval[3], PhD; Moon-Ho Ringo Ho[1], PhD; Yin-Leng Theng[1,4], PhD

[1]Ageing Research Institute for Society and Education (ARISE), Nanyang Technological University, #01A - 01 Singapore, Singapore, Singapore
[2]Research for Impact, Singapore, Singapore
[3]Department of Geography and the Institute for Urban and Regional Studies, Hebrew University of Jerusalem, Jerusalem, Israel
[4]Wee Kim Wee School of Communication and Information, Nanyang Technological University, Singapore, Singapore

Corresponding Author:
Yuezhong Liu, PhD
Ageing Research Institute for Society and Education (ARISE), Nanyang Technological University, #01A - 01 Singapore, Singapore, Singapore

Abstract

Background: There is increasing recognition of the environment's role in shaping cognitive functioning among older adults.

Objective: This study aimed to examine the relationship between the outdoor mobility of older adults with and without cognitive impairment and the built environment in 3 urban neighborhoods in Singapore, specifically 3 urban, high-density subzones selected a priori.

Methods: Outdoor walking mobility in daily life gait speed (DGS) was collected continuously for 1 week using a previously validated hybrid mobility tracker. Mini-Mental State Examination cut-offs by educational levels were used to differentiate cognitive impairment (CI) and without cognitive impairment (nonCI). The environmental characteristics examined were gross plot ratio and land use. Statistical correlations were used to examine the associations between older adults' outdoor mobility and built environment for all CI and nonCI groups. Two case examples were also used to provide a location-based heatmap on DGS for 3 consecutive days.

Results: The overall mean DGS was 0.75m/s (SD 0.12) for the nonCI group and 0.73m/s (SD 0.08) for the CI group. The between-group difference (0.02m/s) was below commonly cited thresholds. Exploratory land-use and zone-specific summaries suggested context-dependent variation, such as residential areas (CI: 0.80m/s and nonCI: 0.62m/s). Higher GPR was associated with faster DGS in the nonCI group (β=0.04, 95% CI 0.01 - 0.07, *P*=.04) and slower DGS in the CI group (β=–.13, 95% CI –0.20 to –0.04, *P*=.01). CI participants spent more time in commercial and business zones than nonCI participants, while both groups spent the majority of walking time in residential areas. However, estimates were based on small subsamples and multiple unadjusted comparisons and should be interpreted as hypothesis-generating.

Conclusions: This formative, exploratory study suggests that environmental typologies, particularly GPR and land use, may be associated with differences in mobility patterns between older adults with and without cognitive impairment; subgroup patterns were exploratory and not powered for clinical interpretation. Policy implications include integrating fine-grained environmental metrics into age-friendly urban planning. Clinically, mobility assessments should account for environmental context. Future research should use larger, more diverse, and longitudinal samples to confirm these associations and guide the design of supportive urban environments for cognitively diverse aging populations, noting that findings are context-specific to high-density urban environments. However, given data limitations, unmeasured factors such as comorbidity burden, habitual physical activity, socioeconomic context, and caregiver accompaniment may also contribute to the observed patterns, and findings should be interpreted accordingly.

(JMIR Form Res 2025;9:e76777) doi:10.2196/76777

KEYWORDS

cognitive impairment; daily life gait speed; gross plot ratio; land use; older adults

XSL•FO
RenderX

Introduction

The interaction between cognitive function, mobility, and the built environment has become an area of increasing research interest, particularly in rapidly urbanizing societies where older adults represent a growing segment of the population. Mobility in daily life is not only a marker of functional independence but is also closely associated with cognitive health outcomes in later life. Previous research has shown that slower gait speed is associated with poorer cognitive performance, increased risk of cognitive decline, and higher dementia risk [1-3]. Conversely, improvements in gait speed have been linked to better cognitive outcomes and reduced risk of decline [4,5].

While daily-life gait speed (DGS) is a well-established and ecologically valid indicator of functional health, mobility can also be characterized using other measures such as step count, walking distance, life-space mobility, and mobility diversity [6-9]. These indicators capture different aspects of how individuals interact with their environments. Among these, DGS is particularly valuable because it reflects the ability to sustain efficient locomotion in real-world contexts, integrating physical capacity, cognitive resources, and environmental demands.

Emerging evidence suggests that environmental factors—such as walkability, street connectivity, and access to destinations—can influence mobility behaviors in older adults [10-12]. Within this context, two environmental typologies are particularly relevant: gross plot ratio (GPR) and land use. GPR is a standard urban planning measure of built density, defined as the ratio of a building's total floor area to the size of the land parcel on which it sits. Higher GPR values indicate denser, often high-rise environments, while lower values reflect lower-density, low-rise areas. In Singapore, GPR is a central parameter in the Urban Redevelopment Authority's Master Plan and determines allowable development intensity for residential, commercial, and mixed-use areas [13]. Previous studies have linked urban density to pedestrian movement patterns, environmental complexity, visual enclosure, and sensory load, all of which can influence gait speed and navigation—particularly in older adults with cognitive impairment [14-16].

Land use describes the functional designation of an area, such as residential, commercial, business, or community zones. Land use diversity has been associated with opportunities for social participation, access to services, and variety in walking routes, which can influence both physical activity and cognitive engagement [17-19]. For individuals with cognitive impairment, different land use contexts may impose varying navigational and multitasking demands, affecting mobility performance. For example, busy commercial areas may present greater cognitive load due to crowds and signage, while residential zones may offer more predictable, familiar walking environments.

Despite growing interest in the interplay between cognition, mobility, and the built environment, few studies have examined how specific environmental typologies such as GPR and land use interact with cognitive status to shape real-world walking performance. Existing research often treats mobility outcomes without differentiating by environmental type or applies aggregate environmental measures without considering the potentially divergent effects on individuals with and without cognitive impairment.

This study aims to address this gap by examining how GPR and land use are associated with DGS in older adults with and without cognitive impairment living in Singapore. In this context, DGS is used as an observational indicator of mobility performance in different environmental conditions rather than as a diagnostic screening tool. By linking fine-grained environmental metrics with objective mobility tracking data, this work seeks to generate preliminary insights that can inform the design of age-friendly neighborhoods and guide future research on mobility–environment interactions in cognitively diverse aging populations. As Singapore's built form is predominantly urban and high-density, this formative analysis focuses on within-city variation (eg, GPR bands and land-use categories) rather than urban-to-rural contrasts. As such, the insights are most transferable to similarly dense city contexts and are not intended to generalize to suburban or rural environments.

Methods

Ethical Considerations

This manuscript presents a secondary analysis of a deidentified dataset originally collected under National Healthcare Group Domain Specific Review Board approval (NHG DSRB Ref: 2017/00937) and described in [20]. During the original study, all participants were assessed for capacity to consent prior to enrollment. For individuals without cognitive impairment, written informed consent was obtained following a full explanation of objectives, procedures, potential risks, and benefits. For participants with cognitive impairment, trained staff used plain-language explanations and a teach-back process to assess understanding; when participants were able to consent, written consent was obtained directly. Where capacity was uncertain or absent, consent was obtained from a legally acceptable representative, in accordance with local regulations, with verbal or written assent sought from the participant whenever possible. The current secondary analysis involved no new data collection or contact with participants. Use of the dataset for the present purposes was authorized by the data custodian. No additional ethics review was sought because analyses were conducted exclusively on de-identified data previously collected with ethics approval and explicit consent for research use.

All data used in this analysis were deidentified before access by the study team; direct identifiers were removed and replaced with coded study IDs. Electronic files were stored on secure, access-controlled institutional servers with encryption in transit and at rest, accessible only to authorized personnel. Geospatial outputs are reported as aggregate or smoothed summaries (no exact home locations or personally identifying routes are disclosed). Reporting is at a level designed to prevent reidentification. Data retention and destruction follow institutional and NHG DSRB policies. In the original study [20], participants received modest, IRB-approved compensation to acknowledge their time and transport; no performance-based payments were provided. All data are deidentified and reported

XSL•FO
RenderX

in aggregate without individual-level identifiers. Procedures adhered to NHG DSRB research policies, the Belmont Report, and the World Medical Association's Declaration of Helsinki.

Study Design

This study is a formative secondary analysis of cross-sectional data from a larger national-level project on urban planning and the design of age-friendly neighborhoods in Singapore. In total, 3 study neighborhoods (each approximately 1 km²) were selected a priori from the western, central, and eastern parts of Singapore based on a high proportion of older adults (aged ≥55 y). The formative designation reflects the preliminary, hypothesis-generating nature of this work, given the relatively small sample size and exploratory focus on identifying relationships between environmental typologies (gross plot ratio and land use) and daily-life gait speed in older adults with and without cognitive impairment. The intention is to provide early insights that can guide the design of future, larger-scale, and potentially longitudinal studies to further investigate causal pathways and generalizability. Given the 1-week observation window and cross-sectional design, temporal ordering cannot be established and causal inference is not warranted; all analyses are interpreted as associative and hypothesis-generating. All three study subzones were urban and high-density by design, reflecting Singapore's prevailing development pattern. This sampling frame was chosen to maximize environmental variation within dense urban settings (eg, across GPR bands and land-use types) while acknowledging that the absence of suburban or rural areas limits external validity. Accordingly, our inferences are intended as analytical generalizations to comparable high-density cities rather than statistical generalizations to low-density suburban or rural contexts.

Participants

In total, 33 older community‐dwelling adults, 55 years and above, resided in 3 urban subzones (S1, S2, and S3) in Singapore. For this study, "neighborhood" is defined as a 400‐1000m buffer around the home. Only those residents who resided in up to a 1 km buffer around the subzone were qualified for participation. Participants put on a wearable hybrid GPS and accelerometer tracker when going outdoors continuously for 7 days [20]. The tracker was designed to capture the wearer's geospatial data accurately within 15 m (depending on the GPS signal) during outdoor travel or activity to enable the geographic positioning of the participant in longitude and latitude and the travel mode speed. Cognitive status was assessed using the Mini-Mental State Examination (MMSE), administered in person by trained research staff during the original data collection phase of the parent study. Education-adjusted MMSE cut-off scores, based on Singapore-specific normative data [21,22], were used to classify participants into cognitively impaired (CI) and cognitively nonimpaired (nonCI) groups: (1) No formal education: MMSE ≤25; (2) primary education: MMSE ≤27; and (3) secondary education or higher: MMSE ≤29.

Participants scoring above these thresholds were classified as nonCI. All participants were screened for capacity to provide informed consent before enrollment. Those without cognitive impairment provided written consent directly. For participants with cognitive impairment, a simplified explanation of the study was given, and understanding was assessed using a teach-back process. Where independent capacity was lacking, consent was obtained from a legally acceptable representative in accordance with local regulations, and verbal or written assent was sought from the participant whenever possible.

Dataset

Individual Data

Basic sociodemographic, cognitive data, and DGS were collected [12,13]. Standardized indices of comorbidity (eg, count or burden), device-based physical-activity metrics beyond outdoor DGS (eg, step count across all domains), detailed socioeconomic indicators beyond education (eg, income and household assets), and caregiver accompaniment were not collected in the parent study. Daily-life gait speed (DGS, m/s) was the sole objective outdoor mobility indicator. It was computed as the average speed of continuous outdoor walking tracks derived from 1-week hybrid GPS+accelerometer logs, restricting to outdoor ambulation bouts and excluding implausible segments (GPS distance ≤10 km). The estimated average speed was calculated based on a continuous walking track and a GPS distance of less than 10km [23].

Gross Plot Ratio and Land Use

The GPR, monitored by the Urban Redevelopment Authority of Singapore (URA), reflects how intensively the land parcel can be used. The GPRs of land parcels are revised in the draft master plan and subsequently, the master plans, along with their respective land uses. Both GPR and land use are used as indicators of development control and land use in different areas of the city in addition to being used for cost forecasting and resourcing a given construction quality [24,25]. URA described the full details of the land use map of Singapore [26]. There are 31 categories of zoning interpretation, such as residential, commercial, hotel, business park, waterbody, and transport facilities. Table 1 shows the general considerations of the development control of GRP and building storeys. If 2 plots of land have similar areas (on the URA Space map Figure 1) but differing plot ratios, it usually means that one will be denser or taller than the other.

XSL•FO
RenderX

Table . Gross plot ratio categories with corresponding building storeys, as specified in the Urban Redevelopment Authority of Singapore (URA) Master Plan [25].

Gross plot ratio or density	Storeys
1.4 (very low density)	5
1.6 (low density)	12
2.1 (medium density)	24
2.8 (high density)	36
>2.8 (very high density)	>36

Figure 1. Urban Redevelopment Authority of Singapore (URA's) online masterplan map with gross plot ratio and land use [27].

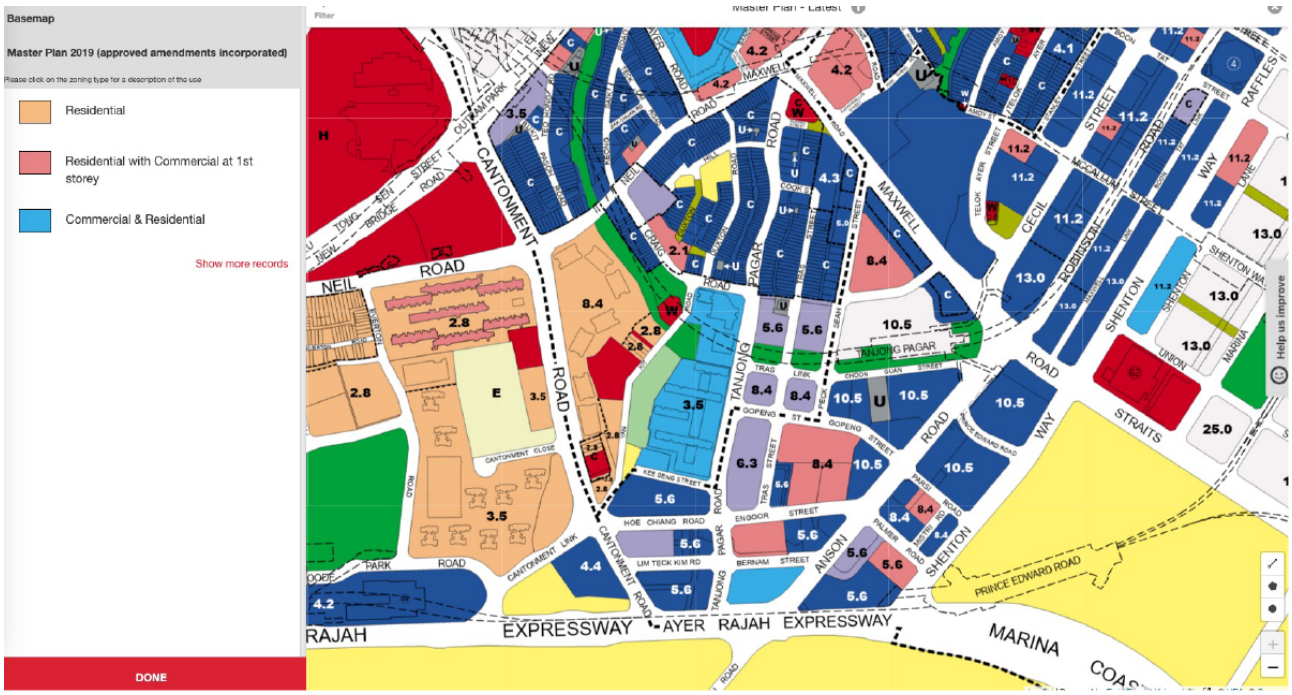

Tracker Platform, Previous Validation, and Postprocessing Pipeline

This paper reports a secondary analysis of data collected with a hybrid GPS–accelerometer tracker developed and validated previously; device specifications, firmware, logging logic, and the original workflow are detailed in [20]. Our present focus is the derivation of DGS from those logs and its association with environmental typologies; readers are referred to [20] for device-level development and validation.

Following the tracker workflow, we implemented a Python pipeline that (1) aligns timestamps between GPS and accelerometer streams by recorded date, (2) applies a short moving-average (3-point) smoothing to latitude or longitude to reduce multipath jitter, (3) computes 1 second ground speed from haversine distances, and (4) removes spikes (ground speed>5.0m/s). Walking versus vehicular mode thresholds followed our a priori rule:<1.39m/s (5km/h)=walking;≥1.39m/s=vehicular.

We defined tracks as sequences of ≥20 consecutive 1 second samples classified as walking (0.30–<1.39m/s), permitting ≤2 second short gaps to bridge brief satellite occlusion, and requiring net displacement ≥20 m to exclude dithering. Any internal spike ≥1.39m/s invalidated the track. For each qualifying track, DGS =total GPS distance / track duration; tracks with GPS distance >10km were excluded as implausible for walking. Accelerometer signals were used to corroborate step-like periodicity within tracks; segments without this periodicity were dropped.

Data Analyses

Descriptive statistics were used to summarize participant characteristics. Gross plot ratio and land use were retrieved by the public application programming interface from the master plan map [27]. DGS was examined in relation to GPR and land use using Pearson's correlation and simple linear regression. Before conducting regression analyses, model assumptions of normality, homoscedasticity, and independence were assessed. Normality of residuals was evaluated using Shapiro–Wilk tests and Q–Q plots, while homoscedasticity was examined via residual-versus-fitted plots. Independence of observations was ensured by the study design. For correlations, Pearson *r* values with 95% CIs were reported as measures of effect size. For regression models, standardized β coefficients with 95% CIs and corresponding *P* values were presented. Potential confounding variables, including age, sex, and education level, were examined in preliminary analyses, and adjusted models

XSL•FO
RenderX

were run where confounding effects were observed. Statistical significance was set at α=.05. All analyses were conducted using Python 3.8.0 (*SciPy* 1.9.0 package). Subgroup analyses by land use and GPR were exploratory and not prespecified; *P* values for these comparisons are descriptive and were not adjusted for multiplicity. We therefore emphasize effect sizes and 95% CIs and interpret subgroup findings cautiously. Given the modest sample size, we limited covariate adjustment to a parsimonious set (age, sex, and education) to avoid model overfitting. Additional covariates (eg, comorbidities, broader physical-activity metrics, detailed socioeconomic status, SES, caregiver accompaniment) were unavailable. As a coarse SES descriptor, housing type was highly skewed (public housing in 95% of participants), precluding stable adjustment and limiting SES variability. Accordingly, residual confounding by unmeasured factors is likely, and effect estimates should be interpreted as associative.

Results

Participants' Characteristics

Of 33 participants, 18 were from S1, 8 were from S2, and 7 were from S3. Sociodemographics are shown in Table 2. The mean (SD) age of the samples was 69.2 (7.14) years, 21 of them (21/33, 64%) were females, and the majority were Chinese (31/33, 94%). Half were married, with only 42% (14/33) having completed secondary school or higher in education. The average number of schooling was 7.3 (4.19) years. None of the participants had full‐time employment, with 18% (6/33) holding part‐time employment. The majority of participants lived in public housing (31/33, 95%) and lived in their respective subzones for an average of 23.1 (SD 15.2) years. All participants had no previous clinical diagnosis of dementia.

XSL•FO
RenderX

Table . Includes age, sex, ethnicity, education, housing type, years of schooling, retirement duration, length of residence, MMSE[a] score, and cognitive status classification.

Demographics	S1 (n=18)	S2 (n=8)	S3 (n=7)
Age, mean (SD)	71.4 (6.43)	63.3 (4.86)	70.0 (8.00)
Female, n (%)	14 (77.9)	4 (50)	3 (42.9)
Ethnicity, n (%)			
Chinese	18 (100)	6 (75)	7 (100)
Malay	0	2 (25)	0
Highest education completed, n (%)			
No formal education	1 (5.6)	2 (25.0)	0
Primary school	2 (50.0)	3 (37.5)	3 (42.9)
Secondary school	4 (22.2)	2 (25.0)	2 (28.6)
Postsecondary	1 (5.6)	1 (12.5)	1 (14.3)
Tertiary	3 (16.7)	0	0
Do not know or not sure	0	0	1 (14.3)
Housing type, n (%)			
HDB[b] 1 - 2 room	3 (16.7)	0	0
HDB 3 room	10 (55.6)	0	5 (71.4)
HDB 4 room	3 (16.7)	6 (75.0)	0
HDB 5 room or HUDC or executive flat	1 (5.6)	2 (25.0)	2 (28.6)
Condo or apartment[c]	1 (5.6)	0	0
Number of years of schooling, median (SD)	6 (0 - 16)	7 (0 - 13)	6 (3 - 10)
Number of years retired, median (range)	16 (8 - 30)	2 (1 - 9)	6 (1 - 22)
Number of years living here, median (range)	20 (0.5 - 50)	20 (8 - 32)	25 (4 - 50)
MMSE score, mean (SD) (max 30)	27.2 (2.32)	26.4 (2.13)	26.6 (2.22)
Cognitive impairment (MMSE[a] cut - offs), n (%)	6 (33.3)	4 (50.0)	4 (57.1)

[a]MMSE: Mini-Mental State Examination.

[b]HDB: Housing and Development Board is Singapore's Public Housing.

[c]Condo/Apartment is Private housing.

Outdoor Mobility by Cognitive Impaired and Nonimpaired Participants

The travel mode of the participant was determined by travel speed. A travel speed less than 1.39 m/s (5 km/h) was categorized as walking mode, and a speed above 1.39 m/s (5 km/h) was categorized as vehicular mode [28,29]. Neighborhoods were defined as the respective subzones and a 1 km buffer [30]. Table 2 shows all GPS records (a total of 5,583,089 records per second). Overall, the DGS is 0.74 m/s (2.66 km/h). Across all participants, mean DGS was 0.74 m/s (SD 0.11); by cognitive status, nonCI: 0.75 m/s (SD 0.12) and CI: 0.73 m/s (SD 0.08).

Outdoor Mobility With Gross Plot Ratio and Land Use

As shown in Table 3, based on one-way ANOVA tests, both groups' DGS was significantly associated with GPR: nonCI group (F=77.13, P<.001) and CI group (F=111.39, P<.001). In correlation analyses, higher GPR was moderately positively correlated with faster outdoor DGS in the nonCI group (r=0.69, 95% CI 0.28-0.88, P=.003) and strongly negatively correlated with DGS in the CI group (r=–0.75, 95% CI –0.91 to - 0.38, P=.001).

XSL•FO
RenderX

Table . Association between gross plot ratio (GPR) and daily-life gait speed (DGS) for cognitively nonimpaired (nonCI) and cognitively impaired (CI) participants.

Group	Mean DGS[a] (m/s)	SD DGS (m/s)	GPR-DGS,[b] β	GPR-DGS, R	95% CI	*P* value	*F* statistic (ANOVA)	Model R^2
NonCI[c]	0.75	0.12	0.04	0.69	0.01 to 0.07	.03	77.13	0.48
CI[c]	0.73	0.08	–0.13	–0.75	–0.20 to –0.05	.01	111.39	0.56

[a]DGS: daily life gait speed.

[b]GPR: gross plot ratio.

[c]CI: cognitive impairment.

Within this urban frame, exposure variability was present across both land-use categories (residential, commercial, business, and community) and GPR levels ranging from 2.1 to 5.6 (Table 4), enabling within-city contrasts despite the absence of suburban or rural environments. In unadjusted linear regression models, a one-unit increase in GPR was associated with a 0.04 m/s faster DGS in the nonCI group (standardized β=.42, 95% CI 0.01-0.07, *P*=.03, R^2=0.47, adjusted R^2=0.44) and a 0.13m/s slower DGS in the CI group (standardized β=–.55, 95% CI –0.21 to - 0.05, *P*=.01, R^2=0.56, adjusted R^2=0.53). After adjusting for age, sex, and education level, the associations remained statistically significant for both groups (nonCI: β=.40, 95% CI 0.01-0.07, *P*=.04, R^2=0.49, adjusted R^2=0.45; CI: β=–.53, 95% CI –0.20 to - 0.04, *P*=.01, R^2=0.58, adjusted R^2=0.54).

Table . Daily-life gait speed (DGS) and time spent across gross plot ratio (GPR) categories and land use types for cognitively nonimpaired (nonCI) and cognitively impaired (CI) participants.

Category (GPR and land use types)	Mean DGS (m/s), NonCI	Mean DGS (m/s), CI	Time spent (min/wk), Non-CI	Time spent (min/wk), CI
GPR				
2.10	0.69	0.81	0.95	0.00
2.50	0.71	0.78	3.55	0.00
2.80	0.73	0.76	106.20	49.25
3.00	0.75	0.74	163.32	15.57
3.20	0.77	0.72	0.58	0.00
3.50	0.78	0.70	5.37	0.35
4.00	0.80	0.68	11.65	3.33
4.20	0.81	0.66	19.42	4.60
4.50	0.82	0.65	0.72	0.30
4.90	0.83	0.64	0.30	0.00
5.60	0.84	0.63	0.23	0.00
Land use				
Community	0.62	0.8	2214.70	1531.00
Commercial	0.76	0.67	2.20	4.60
Business	0.72	0.65	8.05	24.08
Residential	0.67	0.70	54.05	258.30

When stratified by land use typology, effect sizes varied. In business zones, the nonCI group walked faster than the CI group (mean difference=0.07 m/s, 95% CI 0.03-0.11, *P*=.002, Cohen *d*=0.84). In commercial zones, a similar pattern was observed (mean difference=0.09 m/s, 95% CI 0.04-0.13, *P*<.001, Cohen *d*=0.92). Conversely, in community and residential zones, CI participants walked faster than nonCI participants (community: mean difference=–0.03 m/s, 95% CI –0.05 to - 0.01, *P*=.02, Cohen *d*=–0.58; residential: mean difference=–0.18 m/s, 95% CI –0.23 to - 0.12, *P*<.001, Cohen *d*=–1.21). When stratified by land-use typology, numerical differences varied across contexts. Several comparisons reached nominal statistical significance (eg, business and commercial zones), but these subgroup analyses were exploratory, based on small cell sizes, and not adjusted for multiple testing. CI indicated limited precision in several strata. Accordingly, these results are hypothesis-generating and should not be interpreted as clinically meaningful without replication.

These findings indicate that both GPR and land use are associated with variations in DGS, and that the direction and magnitude of these associations differ by cognitive status.

XSL•FO
RenderX

Table 4 presents the distribution of mean DGS and time spent across each GPR category and land use type. In addition to descriptive summaries, group comparisons were conducted to examine differences in time spent across each GPR category and land use type. Due to nonnormal distributions, Mann–Whitney *U* tests were applied. No significant differences were found in time spent across GPR categories between CI and nonCI participants (all *P*>.05). For land use types, CI participants spent significantly more time in commercial areas (median [IQR]: 198 [140–256] min/week) compared to nonCI participants (154 [112–196] min/wk, *P*=.04), and more time in business areas (112 [84–145] vs 86 [65–108] min/wk, *P*=.03). Both groups spent the majority of their walking time in residential areas, with nonCI participants spending more time than CI participants (2214.7 vs 1531.0min/wk, *P*=.02). Time spent walking in community zones was minimal for both groups (<5% of total walking time).

Discussion

Daily Life Gait Speed (DGS) With Built Environment

There are significant differences in the mobility of CI and nonCI samples between DGS and built environment variables (gross plot ratio and land use) in Tables 3 and 4. Cognition is associated with variations in outdoor mobility among older adults in the community. Cognition, including executive functions, has been associated with the ability to perform concurrent motor and cognitive tasks during outdoor navigation in older adults [31,32]. Lamoth et al observed that differences in cognitive function co-occur with differences in gait variability, stability, and dual-task walking [32]. There were small differences in DGS for both nonCI (0.75 m/s) and CI (0.73 m/s). While the overall mean difference in DGS between CI (0.73 m/s) and nonCI (0.75 m/s) participants was 0.02m/s, this value falls below the commonly cited minimal clinically important difference (MCID) for gait speed in older adults, which ranges from approximately 0.05 m/s for a small meaningful change to 0.10 m/s for a substantial change [33,34]. Therefore, this overall difference should not be interpreted as clinically significant in isolation, exploratory stratification by environmental typology yielded larger numerical differences in some contexts (eg,0.18 m/s in residential zones). However, MCID estimates were derived primarily from clinic-based tests and may not translate directly to ecologically measured daily-life gait speed; moreover, subgroup sample sizes were small, and comparisons were not adjusted for multiplicity. Therefore, these patterns should be viewed as preliminary signals requiring confirmation rather than as evidence of clinically meaningful subgroup effects. This pattern is consistent with context-dependent differences in gait speed between older adults with and without cognitive impairment, underscoring the value of examining mobility within specific environmental settings rather than relying solely on aggregate averages. When stratified by land use, we observed larger differences in DGS. On average, nonCI older adults were observed to walk more slowly than CI older adults in low-GPR (lower-density) areas and in community and residential zones, whereas nonCI older adults were observed to walk more quickly in high-GPR (higher-density) areas and in business and commercial zones. The findings were consistent for both GPR and land use results, as lower GPR is approved for community and residential use, and high GPR is approved for business and commercial use.

Alternative explanations and residual confounding. While models adjusted for age, sex, and education, other factors not measured in this study - such as comorbidity burden, habitual physical activity beyond outdoor DGS, socioeconomic conditions beyond education, and caregiver accompaniment - could be associated with both where and how participants walk and with observed gait speed. Given the modest sample and lack of harmonized measures for these domains, we could not control for them directly; thus, the observed associations with GPR and land use should be interpreted with caution as they may partly reflect unmeasured confounding.

Both groups spent the majority of their time walking in residential areas (nonCI: 2214.7 min, CI: 1531.0 min), with nonCI spending more time walking. We postulate that CI and nonCI individuals may have different walking behaviors while interacting with different land uses. Various studies have shown that standardized measures such as lower extremity strength and indoor gait speed [35] are important for older adults to interact with their physical environments. However, the measures represent a physiological potential in mobilizing in the community. Outdoor walking speed is more reflective of real-world interactions with the environment, which may be influenced by cultural, economic, and social factors [36-38]. While our data indicate that CI participants walked faster in community zones than nonCI participants, this finding should be interpreted with caution, as community walking comprised a very small proportion of total walking time for both groups. The limited exposure to this land use type may not provide a robust basis for generalizing speed patterns in community environments. The additional analyses also revealed that CI participants spent more time in commercial and business zones than nonCI participants, a pattern that co-occurred with the gait-speed differences observed in these contexts. As these observations were post hoc, based on small strata, and not adjusted for multiple comparisons, they are exploratory and hypothesis-generating only.

The lower DGS observed in nonCI participants within community and residential land use areas may be associated with higher levels of interaction and participation with activities in those environments. In Singapore, residential zones often include integrated amenities such as shops, markets, and community centers, which could influence walking speed by encouraging frequent stops or interactions. However, we did not measure cultural, economic, or social participation in this study, and any such explanations remain speculative. Future research that incorporates these contextual variables is needed to assess their potential influence on mobility patterns.

Older adults with CI walked with a lower DGS in business and commercial areas. One unmeasured, hypothesis-generating explanation is greater cognitive or multitasking demands in these environments; however, these factors were not quantified in this study. Furthermore, these potential explanations were not directly measured and should be considered hypotheses for future study rather than confirmed effects. Studies have also

XSL•FO
RenderX

reported both positive and negative effects of the gross plot ratio on DGS [36-38]. For example, higher-density built forms may provide greater shade that could support thermal comfort [39]. Prominent buildings could assist wayfinding and orientation [40]. Conversely, tall-building forms may create wind conditions and visual enclosure that could impede comfort or perceived effort [41]. We did not measure microclimate, wayfinding, or wind conditions; these potential mechanisms remain hypotheses for future study. The higher DGS among nonCI participants in business and commercial areas may be related to the environmental features noted above, such as improved wayfinding or shading in high-density zones. As these environmental influences were not directly quantified in our dataset, they remain speculative and warrant further investigation.

External Validity and Context

Our findings should be interpreted in light of the setting: three a priori selected, urban, high-density subzones in Singapore. While this design provided fine-grained contrasts across land use and GPR bands within dense neighborhoods, it does not capture suburban or rural exposures. Consequently, the results are context-specific and best transferred to similarly high-density city environments. Extending external validity will require multisite designs that deliberately sample lower-density neighborhoods (eg, suburban towns) and rural settings, and that incorporate cross-city comparisons where urban morphology, transport supply, and climatic conditions differ.

Case Example of DGS Between CI and NonCI Individuals in the Same Neighborhood

To illustrate potential differences in mobility patterns between older adults with and without cognitive impairment, we examined one participant from each group who lived in the same subzone, were of similar age (CI: 82 y; nonCI: 77 y), and had comparable familiarity with the neighborhood. These participants were selected for illustrative purposes only, and the findings are not intended to be representative of their respective groups.

The CI participant (male, MMSE score: 23) walked a total of 8.4km over three consecutive monitored days, with an average DGS of 0.76 m/s. The majority of walking time (72%) occurred in residential areas, followed by commercial zones (18%) and community zones (10%). The nonCI participant (female, MMSE score: 26) walked 10.2 km over the same period, with an average DGS of 0.82 m/s, and a more varied distribution of walking time: residential areas (58%), commercial zones (28%), and community zones (14%).

Heatmaps (Figure 2) show that the CI participant's walking routes were predominantly linear and repetitive, concentrated along a familiar daily route, with relatively consistent gait speeds (more red areas) along the same path. In contrast, the nonCI participant's routes were more varied and nonlinear, covering a wider range of neighborhood locations with fluctuating walking speeds.

Figure 2. Daily outdoor trajectory comparison of two cases (CI and nonCI) living in the same area (color represents the daily life gait speed; blue represents low speed and red represents high speed, average DGS is 0.74m/s). CI: cognitive impairment.

Data on whether participants walked alone or with a caregiver were not collected; therefore, potential influences of accompaniment on route choice, walking speed, and environmental engagement could not be assessed. This represents an important area for future research. The case study findings should be interpreted as exploratory and hypothesis-generating, providing a visual and quantitative example of how mobility patterns may differ between CI and nonCI individuals when navigating the same neighborhood environment.

Studies have also shown that familiarity with the environment may be associated with differences in DGS and the extent of travel [42,43], especially for persons with dementia [44]. These

XSL•FO
RenderX

patterns suggest that mobility differences between CI and nonCI groups are context-dependent and may be influenced by environmental demands as well as individual capabilities. However, given the preliminary and exploratory nature of this work, these observations should be interpreted with caution. Individuals who are familiar with their surroundings tend to navigate more confidently and efficiently, which can influence their walking speed. While these case examples illustrate potential influences of environmental familiarity on gait speed, they are descriptive and based on a small number of participants. As such, they serve to generate hypotheses rather than confirm causal relationships.

Taken together, our findings provide early evidence that built environment typologies and cognitive status may be linked to differences in outdoor walking patterns. Nonetheless, several methodological and contextual limitations should be considered when interpreting these results.

Limitations and Conclusions

Within these constraints, this formative study provides preliminary, within-city evidence that environmental typologies (GPR and land use) are associated with differences in daily-life gait speed among older adults in high-density urban neighborhoods. This study has several limitations that should be considered when interpreting the findings. First, the sample size was modest (n=33), with participants further divided into CI and nonCI groups, which reduced statistical power and may have limited our ability to detect subtle subgroup differences, particularly when stratified by land use or gross plot ratio categories. Second, recruitment was limited to three urban, high-density subzones in Singapore that were selected for having a high proportion of older adults, which may introduce selection bias and constrain the range of environmental exposures. Third, the geographic context is specific to Singapore's high-density urban environment; findings may not be directly generalizable to lower-density suburban or rural regions with different built forms, transport systems, and cultural or climatic conditions. Fourth, the study's cross-sectional design restricts the ability to establish temporal or causal relationships between cognitive status, mobility patterns, and environmental typologies. Observed associations should therefore be interpreted as indicative rather than confirmatory. Fifth, we lacked harmonized measures of key potential confounders—including comorbidities, broader physical-activity metrics (eg, device-based step counts across domains), socioeconomic indicators beyond education, and caregiver accompaniment—which may influence both environmental exposure and mobility. Because housing type was heavily skewed toward public housing (95%), SES variation was limited and not amenable to stable adjustment. Consequently, residual confounding remains a plausible explanation for part of the observed associations. In addition, subgroup analyses (by land use and GPR) were exploratory, underpowered, and not multiplicity-adjusted; consequently, nominal *P* values are descriptive and clinical interpretation of subgroup contrasts is not warranted.

Within these constraints, this formative, exploratory study examined how built-environment typologies were associated with daily-life gait speed. The overall group difference was small and below commonly cited MCID thresholds. Although exploratory stratification revealed larger numerical differences in some settings (eg, 0.18 m/s in residential zones), these subgroup contrasts were based on small cell sizes and were not adjusted for multiple testing; their clinical relevance is uncertain and requires replication. These findings are consistent with context-dependent differences in mobility patterns between CI and nonCI groups; given the small sample and cross-sectional, 1-week design, they should be interpreted as preliminary and associational. In light of unmeasured comorbidity, physical activity, socioeconomic, and caregiving factors, these findings should be viewed as preliminary and associative rather than explanatory of underlying mechanisms. Further research with larger, more diverse cohorts and longitudinal tracking is needed to confirm these patterns, explore potential mechanisms, and inform the design of urban environments that support mobility for older adults across cognitive profiles.

From a policy perspective, our results highlight the value of incorporating fine-grained environmental measures, such as gross plot ratio and land use diversity, into urban design guidelines to support mobility and accessibility for older adults, particularly those with cognitive impairment. For clinical practice, we recommend that gait speed assessments and mobility interventions consider the specific environmental contexts in which walking occurs, as some settings (eg, high-density commercial areas) may impose greater cognitive or navigational demands. To enhance external validity, future research should adopt multisite, stratified sampling across density gradients (including suburban and rural settings) and across cities with distinct urban forms, thereby testing the transferability of these associations beyond high-density contexts. By aligning research, practice, and policy efforts, urban environments can be better designed to promote safe, accessible, and supportive mobility for older adults across cognitive profiles.

Acknowledgments

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of Singapore's Ministry of National Development, National Research Foundation, or any of the agency collaborators. The grant funders were not involved in the design of the study, the collection, analysis, and interpretation of data, or the writing of the manuscript.

This material is based on research/work supported by Singapore's Ministry of National Development and National Research Foundation under the Land and Liveability National Innovation Challenge Funding Programme (grant number L2NICTDF1–2017-2).

XSL•FO
RenderX

Data Availability

The GPS datasets generated or analyzed during this study are not publicly available due to confidentiality agreements between the research parties, but environmental data are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

References

1. Ngandu T, Lehtisalo J, Solomon A, et al. A 2 year multidomain intervention of diet, exercise, cognitive training, and vascular risk monitoring versus control to prevent cognitive decline in at-risk elderly people (FINGER): a randomised controlled trial. Lancet 2015 Jun 6;385(9984):2255-2263. [doi: 10.1016/S0140-6736(15)60461-5] [Medline: 25771249]
2. Gatto NM, Henderson VW, Hodis HN, et al. Components of air pollution and cognitive function in middle-aged and older adults in Los Angeles. Neurotoxicology 2014 Jan;40:1-7. [doi: 10.1016/j.neuro.2013.09.004] [Medline: 24148924]
3. Wu J, Wilhelm M, Chung J, Ritz B. Comparing exposure assessment methods for traffic-related air pollution in an adverse pregnancy outcome study. Environ Res 2011 Jul;111(5):685-692. [doi: 10.1016/j.envres.2011.03.008] [Medline: 21453913]
4. Seeman TE, Lusignolo TM, Albert M, Berkman L. Social relationships, social support, and patterns of cognitive aging in healthy, high-functioning older adults: MacArthur studies of successful aging. Health Psychol 2001 Jul;20(4):243-255. [doi: 10.1037//0278-6133.20.4.243] [Medline: 11515736]
5. Fratiglioni L, Wang HX, Ericsson K, Maytan M, Winblad B. Influence of social network on occurrence of dementia: a community-based longitudinal study. Lancet 2000 Apr 15;355(9212):1315-1319. [doi: 10.1016/S0140-6736(00)02113-9] [Medline: 10776744]
6. Weuve J, Puett RC, Schwartz J, Yanosky JD, Laden F, Grodstein F. Exposure to particulate air pollution and cognitive decline in older women. Arch Intern Med 2012 Feb 13;172(3):219. [doi: 10.1001/archinternmed.2011.683]
7. Berman MG, Jonides J, Kaplan S. The cognitive benefits of interacting with nature. Psychol Sci 2008 Dec;19(12):1207-1212. [doi: 10.1111/j.1467-9280.2008.02225.x] [Medline: 19121124]
8. Power MC, Weisskopf MG, Alexeeff SE, Coull BA, Spiro A III, Schwartz J. Traffic-related air pollution and cognitive function in a cohort of older men. Environ Health Perspect 2011 May;119(5):682-687. [doi: 10.1289/ehp.1002767]
9. Giles-Corti B, Vernez-Moudon A, Reis R, et al. City planning and population health: a global challenge. Lancet 2016 Dec 10;388(10062):2912-2924. [doi: 10.1016/S0140-6736(16)30066-6] [Medline: 27671668]
10. Cho GH, Rodriguez D. Location or design? Associations between neighbourhood location, built environment and walking. Urban Stud 2015 Jun;52(8):1434-1453. [doi: 10.1177/0042098014537691]
11. Kaspar R, Oswald F, Wahl HW, Voss E, Wettstein M. Daily mood and out-of-home mobility in older adults: does cognitive impairment matter? J Appl Gerontol 2015 Feb;34(1):26-47. [doi: 10.1177/0733464812466290] [Medline: 25548087]
12. Del Pozo Cruz B, Ahmadi M, Naismith SL, Stamatakis E. Association of daily step count and intensity with incident dementia in 78 430 adults living in the UK. JAMA Neurol 2022 Oct 1;79(10):1059-1063. [doi: 10.1001/jamaneurol.2022.2672] [Medline: 36066874]
13. Seo K, Takayanagi N, Sudo M, et al. Association between daily gait speed patterns and cognitive impairment in community-dwelling older adults. Sci Rep 2023 Feb 16;13(1):2783. [doi: 10.1038/s41598-023-29805-4] [Medline: 36797381]
14. Chung J, Boyle J, Wheeler DC. Relationship between life-space mobility and health characteristics in older adults using global positioning system watches. J Appl Gerontol 2022 Apr;41(4):1186-1195. [doi: 10.1177/07334648211054834] [Medline: 34719296]
15. Chung J, Sargent L, Brown R, Gendron T, Wheeler D. GPS tracking technologies to measure mobility-related behaviors in community-dwelling older adults: a systematic review. J Appl Gerontol 2021 May;40(5):547-557. [doi: 10.1177/0733464820979801] [Medline: 33356769]
16. Yin L, Raja S, Li X, Lai Y, Epstein L, Roemmich J. Neighbourhood for playing: using GPS, GIS and accelerometry to delineate areas within which youth are physically active. Urban Stud 2013 Nov;50(14):2922-2939. [doi: 10.1177/0042098013482510]
17. Crane BM, Moored KD, Rosso AL, Carlson MC. Using GPS technologies to examine community mobility in older adults. J Gerontol A Biol Sci Med Sci 2023 May 11;78(5):811-820. [doi: 10.1093/gerona/glac185]
18. Kawai H, Obuchi S, Hirayama R, et al. Intra-day variation in daily outdoor walking speed among community-dwelling older adults. BMC Geriatr 2021 Jul 8;21(1):417. [doi: 10.1186/s12877-021-02349-w] [Medline: 34238238]
19. Wettstein M, Wahl HW, Shoval N, Auslander G, Oswald F, Heinik J. Identifying mobility types in cognitively heterogeneous older adults based on GPS-tracking: What discriminates best? J Appl Gerontol 2015 Dec;34(8):1001-1027. [doi: 10.1177/0733464813512897] [Medline: 24652916]
20. Ho SH, Tan DPS, Tan PJ, et al. The development and validation of a prototype mobility tracker for assessing the life space mobility and activity participation of older adults. BMC Geriatr 2020 Jul 22;20(1):251. [doi: 10.1186/s12877-020-01649-x] [Medline: 32698799]

XSL•FO
RenderX

21. Ng TP, Niti M, Chiam PC, Kua EH. Ethnic and educational differences in cognitive test performance on mini-mental state examination in Asians. Am J Geriatr Psychiatry 2007 Feb;15(2):130-139. [doi: 10.1097/01.JGP.0000235710.17450.9a]
22. Feng L, Chong MS, Lim WS, Ng TP. The Modified Mini-Mental State Examination test: normative data for Singapore Chinese older adults and its performance in detecting early cognitive impairment. Singapore Med J 2012 Jul;53(7):458-462. [Medline: 22815014]
23. Tudor-Locke C, Aguiar EJ, Han H, et al. Walking cadence (steps/min) and intensity in 21–40 year olds: CADENCE-adults. Int J Behav Nutr Phys Act 2019 Dec;16(1). [doi: 10.1186/s12966-019-0769-6]
24. Cheng V. Understanding density and high density. In: Ng E, editor. Designing High-Density Cities: For Social and Environmental Sustainability: Routledge; 2009. [doi: 10.4324/9781849774444]
25. Menz S, Rassia S, Jiang Y, et al. Public space evolution in high-density living in Singapore. In: Ground and Elevated Public Spaces in Public Housing Precincts: Future Cities Laboratory; 2014. [doi: 10.3929/ethz-a-010583367]
26. The Planning Act Master Plan Written Statement 2019. Urban redevelopment authority (URA). 2019. URL: https://www.ura.gov.sg/-/media/Corporate/Planning/Master-Plan/MP19writtenstatement.pdf [accessed 2025-05-05]
27. URA SPACE: planning data for building professionals, business operators and public. Urban redevelopment authority (URA). 2023. URL: https://www.ura.gov.sg/maps/ [accessed 2025-05-05]
28. Schimpl M, Moore C, Lederer C, et al. Association between walking speed and age in healthy, free-living individuals using mobile accelerometry--a cross-sectional study. PLOS ONE 2011;6(8):e23299. [doi: 10.1371/journal.pone.0023299] [Medline: 21853107]
29. Bohannon RW, Williams Andrews A. Normal walking speed: a descriptive meta-analysis. Physiotherapy 2011 Sep;97(3):182-189. [doi: 10.1016/j.physio.2010.12.004] [Medline: 21820535]
30. Oliver LN, Schuurman N, Hall AW. Comparing circular and network buffers to examine the influence of land use on walking for leisure and errands. Int J Health Geogr 2007 Sep 20;6(1):41. [doi: 10.1186/1476-072X-6-41] [Medline: 17883870]
31. Hausdorff JM, Schweiger A, Herman T, Yogev-Seligmann G, Giladi N. Dual-task decrements in gait: contributing factors among healthy older adults. J Gerontol A Biol Sci Med Sci 2008 Dec;63(12):1335-1343. [doi: 10.1093/gerona/63.12.1335] [Medline: 19126846]
32. Lamoth CJ, van Deudekom FJ, van Campen JP, Appels BA, de Vries OJ, Pijnappels M. Gait stability and variability measures show effects of impaired cognition and dual tasking in frail people. J Neuroeng Rehabil 2011 Jan 17;8(1):2. [doi: 10.1186/1743-0003-8-2] [Medline: 21241487]
33. Perera S, Mody SH, Woodman RC, Studenski SA. Meaningful change and responsiveness in common physical performance measures in older adults. J American Geriatrics Society 2006 May;54(5):743-749 [FREE Full text] [doi: 10.1111/j.1532-5415.2006.00701.x]
34. Kwon S, Perera S, Pahor M, et al. What is a meaningful change in physical performance? Findings from a clinical trial in older adults (the LIFE-P study). J Nutr Health Aging 2009 Jun;13(6):538-544. [doi: 10.1007/s12603-009-0104-z] [Medline: 19536422]
35. Warren M, Ganley KJ, Pohl PS. The association between social participation and lower extremity muscle strength, balance, and gait speed in US adults. Prev Med Rep 2016 Dec;4:142-147. [doi: 10.1016/j.pmedr.2016.06.005]
36. Bornstein MH, Bornstein HG. The pace of life. Nature New Biol 1976 Feb;259(5544):557-559. [doi: 10.1038/259557a0]
37. Kirkcaldy B, Furnham A, Levine R. Attitudinal and personality correlates of a nation's pace of life. Journal of Managerial Psychology 2001 Feb 1;16(1):20-34. [doi: 10.1108/02683940110366551]
38. Walmsley DJ, Lewis GJ. The pace of pedestrian flows in cities. Environ Behav 1989 Mar;21(2):123-150. [doi: 10.1177/0013916589212001]
39. Vasilikou C, Nikolopoulou M. Outdoor thermal comfort for pedestrians in movement: thermal walks in complex urban morphology. Int J Biometeorol 2020 Feb;64(2):277-291. [doi: 10.1007/s00484-019-01782-2] [Medline: 31515611]
40. Yesiltepe D, Conroy Dalton R, Ozbil Torun A. Landmarks in wayfinding: a review of the existing literature. Cogn Process 2021 Aug;22(3):369-410. [doi: 10.1007/s10339-021-01012-x] [Medline: 33682034]
41. Zarghami E, Karimimoshaver M, Ghanbaran A, SaadatiVaghar P. Assessing the oppressive impact of the form of tall buildings on citizens: Height, width, and height-to-width ratio. Environ Impact Assess Rev 2019 Nov;79:106287. [doi: 10.1016/j.eiar.2019.106287]
42. Lu Y, Sarkar C, Xiao Y. The effect of street-level greenery on walking behavior: evidence from Hong Kong. Social Science & Medicine 2018 Jul;208:41-49. [doi: 10.1016/j.socscimed.2018.05.022]
43. Van Cauwenberg J, Clarys P, De Bourdeaudhuij I, et al. Physical environmental factors related to walking and cycling in older adults: the Belgian aging studies. BMC Public Health 2012 Dec;12(1):142. [doi: 10.1186/1471-2458-12-142]
44. Margot-Cattin I, Kühne N, Öhman A, Brorsson A, Nygard L. Familiarity and participation outside home for persons living with dementia. Dementia (London) 2021 Oct;20(7):2526-2541. [doi: 10.1177/14713012211002030] [Medline: 33779330]

Abbreviations

CI: cognitive impairment

XSL•FO
RenderX

DGS: daily life gait speed
GPR: gross plot ratio
MMSE: Mini-Mental State Examination
NHG-DSRB: National Healthcare Group Domain Specific Review Board
URA: Urban Redevelopment Authority

Edited by J Sarvestan; submitted 09.05.25; peer-reviewed by B Cornish; revised version received 25.09.25; accepted 26.09.25; published 27.10.25.

Please cite as:
Liu Y, Wong CH, Shoval N, Ho MHR, Theng YL
Urban Environment and Outdoor Walking Mobility Among Older Adults With and Without Cognitive Impairment in Singapore: Formative Study
JMIR Form Res 2025;9:e76777
URL: https://formative.jmir.org/2025/1/e76777
doi:10.2196/76777

© Yuezhong Liu, Chek Hooi Wong, Noam Shoval, Moon-Ho Ringo Ho, Yin-Leng Theng. Originally published in JMIR Formative Research (https://formative.jmir.org), 27.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Culturally Tailored Women's Empowerment Strategies to Improve Immunization Uptake: Formative Research Using a Human-Centered Design Approach

Muyi Aina[1*], MBBS, MPH, PhD; Uchenna Igbokwe[1*], MBChB, MSc; Raihanah Ibrahim[1*], MBBS; Itoro Ata[1*], BEng; Safiya Atta[1*], BSc, MPH; Tobiloba Adaramati[1*], BTECH, PGD; Kunle Oreagba[1*], BSc; Ernest Ezeogu[1*], BEng; Precious Ojo Uahomo[1*], BMedSci, MLSC, MSc; Eric Aigbogun Jr[1,2*], MSc, MScOHS, EMBA, PhD

[1]Solina Centre for International Development and Research (SCIDaR), 8 Libreville Crescent, Wuse II, Abuja, Nigeria

[2]Department of Anatomy, Faculty of Basic Medical Sciences, Enugu State University of Science and Technology, Enugu, Nigeria

[*]all authors contributed equally

Corresponding Author:
Eric Aigbogun Jr, MSc, MScOHS, EMBA, PhD
Solina Centre for International Development and Research (SCIDaR), 8 Libreville Crescent, Wuse II, Abuja, Nigeria

Abstract

Background: Human-centered design (HCD) is vital for crafting impactful and successful health interventions. This approach has been effective in enhancing the uptake of health services, especially in rural areas. The Routine Immunization Buddy System (RIBS) is an HCD-driven intervention designed to support economically disadvantaged caregivers by empowering them to take greater responsibility for their own health and the immunization of their children.

Objective: Through a community-based participatory research (CBPR), this study seeks to understand the economic, social, and health lifestyles and the impact on the health-seeking attributes of the potential beneficiaries of the project. This ensures that the identified intervention and its strategy will be implemented in a way that is of utmost benefit to the users.

Methods: Using a 3-step HCD process—discovery, ideation, and formulation—the study conducted focus group discussions and key informant interviews for caregivers and stakeholders capable of influencing the woman's decision within a health setting. Semistructured interviews facilitated an in-depth exploration of participants' experiences. Data collection used audio and video recordings, detailed notes, and observation sessions. Data analysis encompassed transcription, descriptive analysis, and thematic analysis to inform evidence-based interventions.

Results: Four key interconnected barriers to routine immunization were identified from the formative research: (1) vaccine hesitancy, which is driven by safety concerns and inconsistency in vaccine-related messaging; (2) gender-based decision-making constraints, where male approval often influenced access to care; (3) financial limitations affecting transport and medical costs; and (4) unequal access to health care due to geographic disparities. Caregivers emphasized the need for economic empowerment and trusted community-based health education. These findings informed a critical redesign of the intervention from a 2-arm to a 3-arm model, with peer support groups and financial empowerment components added to address both economic barriers and vaccine misinformation through culturally tailored approaches.

Conclusions: The use of HCD-informed formative research has helped identify tailored women's empowerment strategies, such as economic support, community-led advocacy, and peer networks, that show promise in addressing health disparities among caregivers. These approaches are hoped to enhance caregivers' ability to make informed health decisions, ultimately contributing to improved immunization uptake and broader community well-being. The insights from this research are actively informing the design of our upcoming cluster randomized controlled trial, which aims to strengthen women's decision-making power around immunization services.

(JMIR Form Res 2025;9:e60089) doi:10.2196/60089

KEYWORDS

routine immunization; female caregivers; human-centered design; focus group discussions; key informant interviews; Northern Nigeria

XSL•FO
RenderX

Introduction

Immunization is an essential and cost-effective public health strategy that reduces childhood morbidity and mortality, saving an estimated 2-3 million lives each year [1]. Despite its proven benefits, routine immunization (RI) coverage in Nigeria has remained persistently low. A key measure of RI system performance is the coverage of the third dose of the pentavalent vaccine, commonly referred to as Penta-3. This vaccine protects against 5 serious childhood diseases—diphtheria, pertussis (whooping cough), tetanus, hepatitis B, and *Haemophilus influenzae* type b (Hib)—and is typically administered at 14 weeks of age [2]. In Nigeria, the Penta-3 coverage rate stands at 50%, which indicates that RI coverage has consistently lagged [3]. Particularly concerning is the disparity observed across regions, with Northern Nigeria shouldering the heaviest burden of unimmunized children [4,5]. In the North Central, North East, and North West zones, Penta-3 coverage rates are alarmingly low, at 54%, 37%, and 29%, respectively [3]. These findings underscore the urgent need for targeted interventions aimed at improving RI uptake in Northern Nigeria to mitigate the risks posed by vaccine-preventable diseases and safeguard the health of vulnerable populations.

In contrast, the South-East, South-South, and South-Western zones have higher Penta-3 coverage rates of 83%, 70%, and 74%, respectively. Vaccination coverage was found to improve across the geopolitical zones with increasing mothers' wealth and education levels [2]. The results from the second quarter (Q2) of 2019 Lot Quality Assurance Sampling (LQAS, a methodology used to assess the quality and coverage of health interventions) and the Performance Assessment for Performance Management (PAPA, a framework for evaluating and improving program implementation) further supported these conclusions [3,6-9].

Community involvement, which means creating a cooperative partnership between communities and immunization programs, is highlighted in the World Health Organization's 2016 report on Global Routine Immunization Strategies and Practices (GRISP) as 1 of the 9 crucial investments to attain improved immunization results [10]. Human-centered design (HCD) is an effective approach to fostering community involvement and ownership. Five HCD strategies can be seamlessly integrated into community-based participatory research (CBPR) to enhance results: establish transdisciplinary teams, prioritize empathy, engage and collaborate with "extreme users," swiftly prototype, and develop tangible products or services [11]. HCD enables the attainment of a deep understanding of the problem a solution aims to address and a comprehensive grasp of its target users, thus serving as the cornerstone of impactful solutions [12,13]. HCD involves cultivating profound empathy for individuals, uncovering their needs and desires, and crafting customized interventions to address their challenges [14,15]. Only through this comprehensive understanding can a product be developed that resonates with users [16-19].

Despite ongoing efforts to increase immunization rates, Nigeria is still home to an estimated 3.1 million of the world's unvaccinated children [18]. Demand creation for vaccination remains a critical component of national immunization programs. It ensures that parents, caregivers, communities, and stakeholders understand the importance of immunization, believe in the safety and efficacy of vaccines, trust the quality and reliability of health services, and possess both the knowledge and motivation to complete the immunization schedule [20]. Establishing sustainable demand for immunization depends not only on community trust in the health system but also on the ability of caregivers (primarily women) to access, afford, and act upon timely health information [21,22].

In Northern Nigeria, where social and cultural norms often limit women's autonomy in health care decision-making, empowering women is particularly vital. When women are equipped with the right knowledge, financial resources, and decision-making power, they are more likely to seek and complete immunization services for their children. This formative research was therefore designed to understand the lived realities of women in the intervention communities, identify the key social and structural determinants that influence health-seeking behaviors, and codevelop innovative, culturally appropriate strategies to overcome these barriers. The study delineated targeted goals across key thematic areas and clarified contextual challenges and user needs to inform the design of a women's empowerment-focused randomized controlled trial aimed at enhancing RI uptake in the region.

The Routine Immunization Buddy System (RIBS) initiative is designed to empower unemployed mothers in Nigeria by integrating vaccination education with vocational skills training to enhance RI coverage. The program specifically targets rural communities with historically low immunization rates and a high prevalence of unemployed women, aiming to reduce children's vulnerability to vaccine-preventable diseases. RIBS establishes a direct linkage between community health systems, the local structures and networks that deliver primary health services, including immunization, maternal, and child health interventions, and economic empowerment initiatives, which are structured programs that provide individuals with income-generating skills, tools, and opportunities to improve their financial independence. Through this linkage, RIBS leverages both health education and livelihood development to drive behavior change.

Methods

The RIBS Framework

Under the RIBS framework, participating mothers are organized into small support groups and paired with one another as accountability partners. Group leaders are trained by community health workers to deliver practical, culturally relevant information about vaccines using a visual teaching aid called the *Hannun Rigakafi* (Immunization Hand). In parallel with this health education, participants are introduced to potential employment pathways, such as small-scale farming or trade, and receive corresponding tools and training.

This integrated approach strengthens mothers' confidence and motivation to complete the recommended 5 essential childhood immunizations [23]. The RIBS program was initially piloted in

XSL•FO
RenderX

Kaduna State, Nigeria, and included an embedded evaluation component to assess improvements in immunization knowledge and demand among participating mothers.

Formative Research Approach

Research Design

This study was a formative study that used an HCD [24] aimed at designing a women empowerment program to enhance immunization service use through qualitative data gathering [25]. This method was used to gather insights directly from the target population to inform the development of the program, ensuring that cultural nuances, community perspectives, and social determinants are considered in designing the intervention. Through collaboration with diverse stakeholders, including the head of the house, community leaders, and health care professionals, the study aims to develop culturally sensitive and contextually relevant interventions. This process will be useful in modifying an existing intervention or designing a better one that will be tailored to address the disparities in health care access and use, ultimately enhancing immunization coverage and improving health care outcomes within the target community.

The HCD approach involved a 3-phase HCD process—discovery, ideation, and formulation—followed by the remodification of the original randomized controlled trial study concept, as illustrated in Figure 1.

Figure 1. Overview of the study design and participant selection process for the cluster randomized controlled trial.

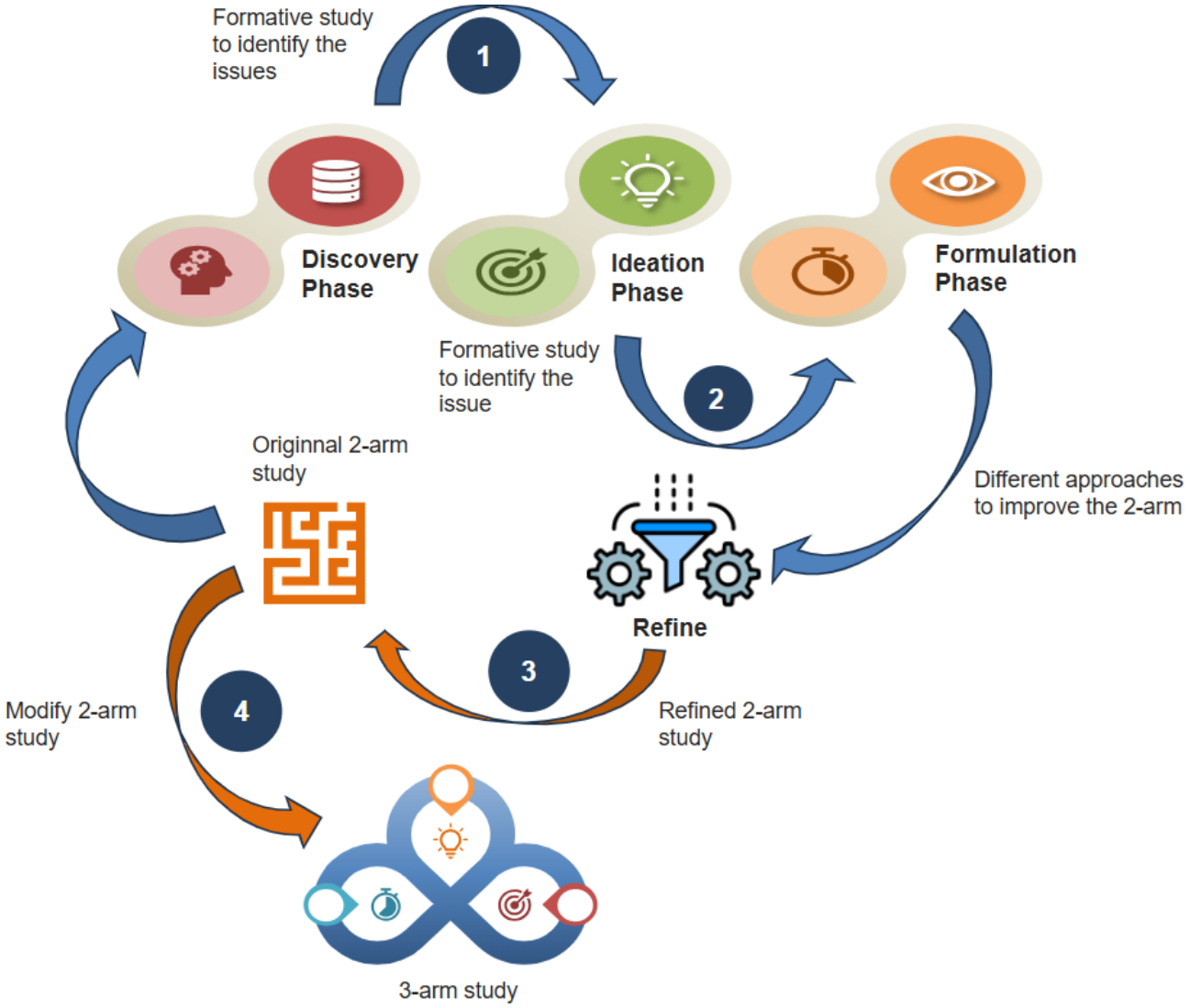

Discovery Phase: Understanding the Lives and Experiences of Caregivers in the Community

Approach

This phase used a CBPR [26] to provide a comprehensive understanding of the lives and experiences of caregivers in Kaduna State, with particular focus on the challenges they faced, resulting in limited access to health care services and RI uptake. Recruitment and data collection were conducted during this discovery phase, engaging with local stakeholders, including community leaders, religious figures, and caregivers, to ensure their active involvement in the research process in understanding and addressing immunization challenges effectively [27,28].

Study Setting and Population

The study was conducted in a semiurban and a rural area, specifically targeting communities with distinct religious compositions—one predominantly Christian and the other

XSL•FO
RenderX

Muslim. This selection allowed for a holistic contextual understanding of the role of culture and religion in describing the specific socioeconomic barriers affecting caregivers in seeking immunization services. Two wards—Kachia and Angwan Yelwan Makaranta—were selected from Kachia and Soba local government areas (LGAs), respectively, in Kaduna State. Kachia is a semiurban settlement with a mix of Christian and Muslim populations, primarily engaged in farming activities such as ginger, yam, maize, and millet cultivation. Angwan Yelwan Makaranta, on the other hand, is a rural settlement predominantly inhabited by Muslims, with farming being the main occupation for men. Participants include caregivers (specifically stay-at-home mothers and other female primary caregivers) of children aged under 5 years who expressed willingness to participate in the study.

Participant Recruitment

Eligible participants were identified through a community-based recruitment strategy involving local health workers and leaders. Caregiver inclusion criteria included (1) being a primary female caregiver aged 18 - 40 years with at least one child under 12 months, (2) not being in formal employment, (3) being a current resident in the study communities, and (4) willingness to participate in group discussions. Community stakeholders (religious leaders, male household heads, and women's group leaders) were selected based on their formal or informal leadership positions and influence on health-seeking behaviors. The female caregivers were identified through household visits, while religious leaders, male decision makers, and women's group leaders were invited through community networks. The 3-week recruitment process included sensitization meetings and written informed consent, with all participants assigned confidential identifications. Trusted community intermediaries facilitated culturally appropriate enrollment to ensure diverse representation.

Research Instruments Design

For the data collection, a sequential approach of focus group discussions (FGDs) and then key informant interviews (KIIs) was adopted. The semistructured FGD guide incorporated an adapted version of the 5C scale, assessing psychological antecedents of vaccination such as confidence, complacency, constraints, calculation, and collective responsibility [29]. This framework guided discussions to uncover barriers and contextual factors influencing immunization uptake, offering a nuanced understanding of caregivers' perceptions and behaviors. This KII guide was developed to follow through with the FGDs for the mothers, exploring issues around women empowerment, financial autonomy, decision-making in immunization, and collective agreement in seeking health. Both tools were structured to include an introduction, demographic information, open-ended questions, probing questions, follow-up questions, a closing message, consent, and confidentiality assurances. Additionally, the tool enabled gathering information about the social, economic, and health status of community caregivers, including their receptiveness to health interventions involving peer support and economic empowerment.

Data Collection

FGDs were strategically tailored to engage caregivers, facilitating dynamic group interactions that encouraged the sharing of collective experiences and perspectives. These discussions provided a valuable space for participants to delve into common challenges and insights related to seeking health care, particularly for mothers. Conversely, KIIs were designed to tap into the understanding of community stakeholders who play pivotal roles in shaping maternal health-seeking behaviors. These interviews offered deeper insights into the intricate dynamics of community influences on health care decisions (Table 1).

FGDs and KIIs were conducted at the Primary Health Center (PHC) Soba in Soba LGA, PHC Kachia, and PHC Gumel in Kachia LGA. This was facilitated by the community health extension workers, community volunteers, and community leaders. The data collection included audio and video recordings of the interview processes. Although the questionnaires were prepared in English, interviews were conducted in the preferred dialects of the participants, which were either Hausa or English. To safeguard confidentiality and integrity, robust data management protocols were implemented, ensuring secure storage and systematic organization of the collected information. Data collection was conducted over 3 days.

Table . The sample frame for the qualitative approach.

Study participants	FGDs[a] (n=7)	Individuals per FGD (n=24)	KIIs[b] (n=2)
Caregivers	4	7	__[c]
Stakeholders			
Husbands	1	7	—
Religious leaders	1	5	—
Community leaders	1	5	—
Community-women leader	—	—	1
Market women leader	—	—	1

[a]FGDs: focus group discussions.

[b]KIIs: key informant interviews.

[c]Not applicable.

XSL•FO
RenderX

Data Analysis

The plan for data analysis involved the transcription of audio and video recordings from FGDs and KIIs. The transcriptions were coded into relevant themes and analyzed using the grounded theory approach for thematic analyses [30,31]. This involved 2 research team members forming concepts from the data and independently identifying several themes. The researchers agreed upon the themes and coded open-ended comments for each theme. We evaluated each comment using the constant comparative method of grounded theory [31].

Ideation Phase: Creative and Collaborative Process of Generating Ideas and Solutions

The ideation phase of the study began with brainstorming sessions and collaborative discussions among the research team, caregivers, and community stakeholders. This phase aimed to devise a comprehensive intervention that resonated with the needs and aspirations of caregivers and stakeholders while effectively addressing the research objectives. Various ideas were explored, taking into account factors such as community dynamics, resource availability, and feasibility. Insights were synthesized, and themes were drawn from the results obtained from the discovery phase.

Formulation Phase: Changing Concepts Into Tangible and Achievable Interventions

In the formulation phase of the HCD process, insights and identified opportunity areas were translated into concrete design concepts. This phase focused on brainstorming ideas, refining concepts based on responses from caregivers and stakeholders, and developing solutions that address the needs and challenges identified during the discovery phase. Collaborative cocreation sessions were conducted to engage stakeholders and ensure that the proposed solutions align with the preferences and aspirations of the caregivers.

Ethical Considerations

This study was reviewed and approved by the ethics research committee of the Department of Planning, Research, and Statistics, Kaduna State Ministry of Health (MOH/ADM/744/VOL.1/601). The committee reviewed and approved the study protocol, data collection instruments, and informed consent documents prior to the commencement of the research. Written informed consent was obtained from all study participants prior to their participation. Consent forms were made available and administered in English, Hausa, or Arabic, depending on the participant's preference. Participants were informed of the study's purpose, procedures, potential risks and benefits, and their right to withdraw at any point without consequence.

To protect participant privacy and confidentiality, all data were anonymized at the point of transcription. Audio and video recordings were securely stored in password-protected folders, and personal identifiers were removed during transcription. Only authorized members of the research team had access to the data. Participants received a modest token of appreciation to compensate for their time, valued at approximately 2000 Nigerian Naira (approximately US $2.5) in the form of basic household items or transportation reimbursement. No images of identifiable individuals are included in this manuscript or supplementary materials. Should such materials be included in future dissemination, written consent for publication of identifiable images will be obtained and archived in accordance with ethical standards.

Results

Grounded Theory Analysis

The formative phase of the HCD approach yielded rich insights into the community's context, which were organized into themes using grounded theory analysis. Table 2 summarizes the key themes and corresponding findings. Notably, an additional theme related to vaccination perceptions and barriers emerged, highlighting caregivers' challenges and concerns about immunization.

XSL•FO
RenderX

Table . Themes from the grounded theory analysis of the interviews.

Themes	Findings
Sociocultural context	Soba and Kachia illustrated how religious diversity influences social norms and cultural behaviors among caregivers.
Education and economic status	Early pregnancy and marriage significantly increase the rate of school dropouts among girls, subsequently limiting their access to further education and economic empowerment opportunities.
Health care access and use	Caregivers recognized the importance of accessing health care for the well-being of themselves and their families. However, they encountered various obstacles that hindered their ability to obtain and use health care services effectively.
Family dynamics and decision-making	The study uncovered notable differences in marital dynamics among caregivers, emphasizing variations in spousal support, extramarital affairs, and decision-making autonomy across different communities.
Future aspirations and community development	Caregivers expressed a strong desire for economic empowerment as a means to enhance their families' standard of living and financial stability.
Vaccination perceptions and barriers	Concerns about vaccine safety, misinformation, and logistical challenges hinder access to immunization services and contribute to low routine immunization uptake.

Understanding the Lives and Experiences of Caregivers in the Community

Sociocultural and Economic Dynamics

The study revealed diverse sociocultural influences, educational challenges, health care access barriers, varied family dynamics, and a strong desire for economic empowerment among caregivers in Soba and Kachia communities. These findings provide valuable insights into the multifaceted challenges and aspirations within the communities of Soba and Kachia, highlighting the need for targeted interventions that address sociocultural, economic, and health care–related issues to promote holistic community development.

Sociocultural Context

Caregivers in Soba and Kachia communities demonstrated how religious variety affects social norms and cultural behaviors. These communities differ in how women are viewed, empowered, and how much autonomy they have in making decisions.

> *I only interact with other women at social gatherings such as weddings and naming ceremonies.* [A caregiver from Soba]

Education and Economic Status

Early pregnancy and marriage greatly raised the percentage of school dropouts among girls, which subsequently limited their access to further education and economic empowerment. For both sexes, agriculture served as the primary source of income, with women actively engaged in small-scale farming and trade.

> *My academic journey came to an end when I got pregnant in senior secondary school.* [A caregiver from Kachia]

Health Care Access and Use

Caregivers acknowledged that access to health care is critical to preserving and advancing their own and their family's health and well-being. Nonetheless, caregivers faced several obstacles that made it difficult for them to get and make use of health care services. Financial limitations were the main cause of these obstacles, keeping many caregivers from getting necessary prescriptions or obtaining prompt medical attention. Fear of adverse effects after vaccination, lack of supplies and medications at the hospital, nonconsenting spouses, and competing religious practices are some of the obstacles to seeking immunization services.

> *It breaks my heart whenever my children ask me for anything and I cannot give them.* [A caregiver from Kachia]

In rural or underserved locations with inadequate infrastructure and resources, caregivers highlighted difficulties with health care facility availability and accessibility. To address systemic problems and enhance vulnerable groups' access to health care, these obstacles highlight the necessity for focused treatments.

> *I trek for thirty minutes to get to the PHC.* [A caregiver from Kachia]

The study found that while caregivers drew on a range of sources for health education and information sharing, health professionals, traditional leaders, and religious figures in the community made significant contributions. Providing accurate and trustworthy information regarding health benefits, preventive measures, and available health care services is a critical responsibility of health workers. Health workers supported caregivers in better recognizing their own health needs and in seeking appropriate care when needed through one-on-one contacts, counseling sessions, and community outreach programs.

Besides health care professionals, traditional and religious leaders provided valuable counsel and assistance to caregivers. These leaders were frequently regarded as reliable sources of information within the community. These leaders used their power and influence to encourage adherence to advised health practices, debunk myths, and promote positive health behaviors.

XSL•FO
RenderX

Traditional and religious leaders were crucial in promoting health education messaging and influencing caregivers' attitudes and behaviors about health care by introducing health-related themes into sermons, community meetings, and cultural activities.

Family Dynamics and Decision-Making

The study revealed significant variations in marital dynamics among caregivers, highlighting differences in spousal support, extramarital affairs, and decision-making autonomy across communities. In some communities, caregivers reported experiencing strong spousal support, where husbands actively participated in household responsibilities, provided emotional and financial assistance, and demonstrated a collaborative approach to decision-making. However, in contrast, caregivers in other communities described strained marital relationships characterized by limited support from their spouses, instances of extramarital affairs, and unequal decision-making power. These disparities underscored the complex interplay of cultural norms, socioeconomic factors, and individual dynamics influencing marital relationships within the studied communities.

> *I don't want my children to live the kind of life I'm living now; I want to send them to school so that they can get good jobs.* [A caregiver from Soba]

Associations emerged as integral components of the social fabric within the communities, offering caregivers opportunities for socialization, knowledge exchange, and collective support. These associations provided a platform for women to come together, share experiences, and offer mutual assistance in navigating various challenges they faced. Through regular meetings, group discussions, and collaborative activities, caregivers benefited from the wisdom and support of their peers, fostering a sense of solidarity and empowerment within the community. Additionally, female associations served as conduits for disseminating health education messages, promoting healthy behaviors, and mobilizing community action to address common concerns. By leveraging these support networks, caregivers were able to access valuable resources, build social capital, and strengthen their resilience in the face of adversity.

Future Aspirations and Community Development

As a way to improve their families' standard of living and financial standing, caregivers expressed a significant desire for economic empowerment, according to the study's findings. Aspired to improve their financial well-being and create a sustainable revenue stream, caregivers expressed a desire for entrepreneurial options and vocational training. Caregivers expressed a strong desire to learn new skills related to tailoring, catering, and entrepreneurship to start profitable businesses through interviews and participatory discussions. The caregivers' strong will and tenacity to escape the cycle of poverty and provide better possibilities for their families and themselves were highlighted by these ambitious goals. To help caregivers reach their full potential, promote their financial independence, and advance community development, stakeholders should fund economic empowerment programs.

> *I want to learn a trade but my husband has to give me permission.* [A caregiver from Soba]

The solutions provided by caregivers centered on targeted interventions aimed at raising the standard of living and resolving socioeconomic problems in the communities. Some of the key suggestions were constructing roads, supplying clean water, and installing electricity to bolster the infrastructure as a whole and encourage economic activity. To support small-scale enterprises and entrepreneurship, caregivers have also advocated for greater access to business financing, such as grants and loans for microfinance. Furthermore, a significant emphasis was placed on enhancing health care services to better treat prevalent health issues and improve community health outcomes. By placing a high focus on community development initiatives, stakeholders may promote social cohesion, facilitate sustainable growth, and help communities thrive in the face of financial challenges.

Vaccination Perceptions and Barriers

Participants expressed concerns over vaccine safety and practical challenges in accessing immunization services, underscoring the need for clear, culturally relevant education on vaccines. One caregiver from Soba explained that safety rumors are common in the community, remarking:

> *Some people say that vaccines can harm our children, and that makes me fear taking them for immunization.*

Another caregiver from Kachia highlighted that widespread rumors about severe side effects further confuse caregivers, noting:

> *I have heard rumours in my community about children developing serious sickness after getting vaccinated; it is hard to know what to believe.*

In addition, concerns over mixed messages from health providers were noted. A caregiver from Soba observed that inconsistent communication at the clinic contributed to ongoing confusion, stating:

> *The clinic gives different messages about vaccines, which is confusing. I wish the information was clearer and more relevant to our culture.*

Transition From 2-Arm to 3-Arm Intervention Design

During the ideation phase, discussions revealed that a 2-arm design, comprising only a control group (arm 1) and a financial empowerment group (arm 2), failed to fully capture the community's multifaceted needs. While financial empowerment is essential, caregivers expressed that additional social support and education were needed to address vaccine hesitancy and improve health decision-making.

Based on community feedback and contextual insights, the study design was modified to include a third intervention arm focused on peer support and education (arm 3). This arm was conceptualized to offer monthly meetings where participants receive tailored immunization and child health education, led by trained community health workers, focusing on dispelling vaccine misinformation, addressing common vaccine-related concerns, and providing accurate information on vaccine safety and benefits. The sessions use culturally adapted tools, such as the *immunization hand tool* and a theme song, to reinforce key messages and enhance information retention. In addition to

XSL•FO
RenderX

educational content, these meetings are designed to foster social support and skill-building, complementing the financial literacy training and microloan access provided to the financial empowerment group.

This integrated approach ensures that the intervention addresses both economic and social barriers, thereby maximizing the potential impact on RI uptake and overall community health. The revised 3-arm cluster randomized controlled trial (cRCT) design is summarized in Table 3.

Table . How the formative research informed the modification of the intervention design in the cluster randomized controlled trial.

Original study design	Challenge identified	Final study design
Number of arms		
Two-arm study (control and empowerment)	• The 2-arm study was limited in contextualizing the issues related to empowerment at the community level • Community members interpreted empowerment differently	• Three-arm study
Factoral design		
Financial empowerment and control groups	• Financial empowerment alone would not address community-level barriers such as cultural beliefs and perceptions • The design did not take into account the role of peer support in encouraging health-seeking	• Peer education and support • Financial empowerment group • Control group
Factors addressed		
Monitor the usage of the funds and how it relates to the uptake of immunization services	• There was a disconnect between what the intervention would address and the stakeholders' perception about the key drivers of immunization uptake	• The new design addresses cooperative financing systems to foster collaboration and improve knowledge about immunization services using peer groups and support
Data collection		
Data to be collected at the facility level	• The data collection source is not exhaustive enough to capture all the necessary indicators in the design	• Immunization data from children • Sociodemographic data from caregivers • Attendance and participation in economic empowerment sessions, immunization, child nutrition, and hygiene knowledge sessions • Loans and loan repayment from cooperative groups

Discussion

Principal Findings

This formative study was designed to inform the development of a culturally appropriate, women-centered cRCT aimed at identifying key sociocultural, economic, and healthcare-related barriers to RI uptake among caregivers, with the goal of increasing awareness and generating demand for RI services among economically disadvantaged populations. The study intricately integrated community engagement throughout the multilayered program development process, offering a unique perspective to guide the intervention. Specifically, the incorporation of HCD in the formative research phase has proven effective in promoting behavior change by ensuring that solutions are grounded in the lived experiences and needs of the target community [32]. The study findings revealed varying challenges such as financial constraints, limited health care access, sociocultural norms, and vaccine safety concerns, all affecting decision-making.

From the initial findings of the study, there was a need for an integrated intervention that addresses both financial and social barriers to improve RI uptake. These insights directly guided the design of a 3-arm cRCT using HCD within a CBPR to understand the specific needs and requirements for developing a culturally appropriate program to enhance mothers' awareness, support, and economic empowerment, ultimately capable of improving uptake of immunization services and more effectively addressing the challenges of vaccine hesitancy.

Addressing Vaccine Hesitancy Through Structured Communication and Leveraging Community Groups for Health Promotion and Behavior Change

One of the key themes that emerged as a persistent gap in RI uptake was perceptions and barriers to vaccination, which aligns with global evidence on vaccine hesitancy. Caregivers expressed fears about potential side effects of vaccines as well as misinformation such as rumors that vaccines are potentially harmful to children. This mirrors findings from low-resource settings where lack of trust in health systems exacerbates low immunization uptake [33]. The fears of caregivers about vaccines were further compounded by mixed messages from health care providers, and this emphasizes the need for a structured, culturally tailored communication as demonstrated

XSL•FO
RenderX

in a similar context by Kaufman et al [34]. Community-led advocacy has been reported as a successful model to lower issues with vaccine hesitancy, where peer educators and religious leaders improved vaccine acceptance by bridging gaps between scientific evidence and local beliefs [35-37]. Using the voices of the community, this strategy builds community resilience and togetherness while also bolstering vaccine confidence [38-41]. Communities can set the path for higher vaccination rates and better public health outcomes by working together and leveraging local expertise and experiences. The core causes of vaccine hesitancy within certain cultural and social contexts can be effectively addressed by community-led lobbying [41-44]. Advocacy groups can help close the gap between scientific knowledge and community perceptions by customizing messaging to fit local conventions, beliefs, and languages [40,45]. In addition to fostering inclusivity, this culturally aware approach makes sure that no member of the community is left out.

The community's source of immunization information is crucial in lowering vaccine hesitancy. When advocacy initiatives come from within the community, they are more effective. This grassroots strategy guarantees that the messaging is understood by the community while also fostering trust. A culture of trust and acceptance can be fostered by giving community members the authority to speak out in favor of vaccinations [46,47]. Advocacy campaigns sponsored by the community can cultivate a sense of empowerment and ownership among its members. People who take the initiative to spread factual information about vaccines engage in a proactive role in preserving the health and welfare of their community. This sense of ownership creates a sense of shared accountability for promoting immunization as a community endeavor and overcoming vaccine skepticism. Traditional and religious leaders also provide valuable counsel and assistance to caregivers. These leaders were frequently regarded as reliable sources of information within the community. These leaders use their power and influence to encourage adherence to advised health practices, debunk myths, and promote positive health behaviors. Traditional and religious leaders are crucial in promoting health education messaging and influencing caregivers' attitudes and behaviors about health care by introducing health-related themes into sermons, community meetings, and cultural activities.

The importance of personalized information campaigns such as community mobilization is superior to general messaging in helping overcome vaccine hesitancy; this finding is consistent with other studies [48,49]. Community groups emerge as integral components of the social fabric within the communities, offering caregivers opportunities for socialization, knowledge exchange, and collective support. These groups provide a platform for women to come together, share experiences, and offer mutual assistance in navigating various challenges they face. Through regular meetings, group discussions, and collaborative activities, caregivers benefit from the wisdom and support of their peers, fostering a sense of solidarity and empowerment within the community. Female associations serve as conduits for disseminating health education messages, promoting healthy behaviors, and mobilizing community action to address common concerns. By leveraging these support groups, caregivers were able to access valuable resources, build social capital, and strengthen their resilience in the face of adversity.

Community groups also offer a forum for the dissemination of health education and information that is appropriate for a certain culture [50,51]. Caregivers can obtain accurate information about illness management, preventive health measures, and available health care services through workshops, educational sessions, and group discussions led by reputable community leaders or medical professionals. Health promotion initiatives become more effective when tailored to the cultural norms, values, and beliefs of the community. Adapting messaging to align with these cultural aspects enhances engagement and increases the likelihood of behavior change as cultural adaptations in health interventions aim to make services more linguistically appropriate, culturally competent, and safe, addressing disparities and improving outcomes for diverse populations [52-55]. Community groups present chances for lobbying and group action aimed at removing structural obstacles to health care use and access. Through the mobilization of caregivers as advocates for enhanced health care services, policies, and infrastructure, these groups can bring about significant changes at the community level that will benefit everyone's health.

Addressing Economic Barriers to Health Care Access Through Financial Empowerment and Women's Autonomy in Health Decision-Making

Financial constraint was another theme identified, and this hindered caregivers from accessing health care services, including RI, as many caregivers expressed that limited household income restricted their ability to make health-related decisions for themselves and their children such as medical care, essential medications, and transportation to health facilities. These financial challenges mostly led to delays in seeking care and, in some cases, avoidance of health services. A significant number of women conveyed their desire to have more control over their financial resources, to secure a steady source of income, and to enhance their financial independence, enabling them to prioritize their health and that of their families and vocational training to improve their ability to prioritize family health needs. Our results corroborate studies that reported that out-of-pocket expenses for medications and transportation represent significant obstacles for families with limited financial capabilities. These costs can be unaffordable, forcing families to forgo needed medical care or make difficult choices about basic necessities [56-58], and this mirrors patterns observed in similar sub-Saharan African contexts [59].

It is important to state that financial empowerment plays a crucial role in empowering women to take charge of their health and well-being, ensuring better access to essential health care services and medications for both themselves and their families [60,61]. It is a crucial factor influencing women's agency in health decision-making processes. Studies have shown that interventions promoting economic empowerment increase women's involvement in making decisions about their own and their children's well-being [62,63].

XSL•FO
RenderX

Access to economic power provides women with resources, opportunities, and decision-making authority, enabling them to prioritize health, access health care services, and contribute to the overall well-being of their households [64]. This also influences children's immunization, as women are more likely to prioritize vaccinations for their children when they have the financial means and autonomy to make health-related decisions [65]. This connection between financial empowerment and women's agency in health decisions demonstrates the importance of addressing economic disparities to enhance women's health outcomes and promote gender equality. By empowering women financially, stakeholders can promote greater autonomy in health decision-making, enhance health care use rates, and improve health outcomes for women and their families.

Addressing Sociocultural Barriers to Immunization Through Women Empowerment and Navigating Religious Influences in Health Decision-Making

Sociocultural norms, influenced by religious diversity, influenced caregivers' decision-making and their approach to health practices, which potentially can affect RI uptake. The varying levels of empowerment and autonomy among women affected their ability to make informed decisions regarding vaccination. In communities where women had limited decision-making power or faced resistance from male family members, immunization uptake was often hindered. These sociocultural factors, combined with misinformation and fear of vaccine side effects, contributed to low RI rates.

The observed influence of religious diversity on health practices aligns with previous studies documenting how faith-based beliefs shape vaccination attitudes across different cultural settings [66]. In particular, our findings resonate with studies from similar settings where religious interpretations of disease prevention have created both barriers and opportunities for immunization programs [66,67]. The gender dynamic observed in decision-making processes is a well-documented pattern in male-dominated societies, where male authority over health decisions usually results in lower child vaccination coverage [68,69]. This situation has been particularly well documented in Northern Nigeria, where studies have reported that spousal approval requirements can reduce immunization likelihood by up to 40% [70,71]. By implication, this can lead to a decrease in vaccine uptake, weaken herd immunity, and increase the risk of vaccine-preventable diseases.

Other identified barriers were inadequate availability of supplies and medications at health care facilities and poor infrastructure in these rural and underserved areas, which contribute to challenges in accessing immunization services. To address systemic problems and enhance vulnerable groups' access to health care, there is a need for focused interventions. While caregivers leveraged a range of sources for health education and information sharing, it remains important for health workers to provide accurate and trustworthy guidance on health benefits, preventive measures, and available services to caregivers, as they play a crucial role in helping caregivers identify their health needs and seek appropriate care.

Implications for the cRCT Study Design

The contextual findings in this formative research were instrumental in the thinking of how to modify the original 2-arm study which was identified to have several challenges, as standalone economic interventions often fail to address deep-rooted vaccine hesitancy. Financial empowerment alone would not address community-level barriers such as cultural beliefs and perceptions toward RI uptake. The design did not take into account the role of peer support in encouraging health-seeking behavior. The study uncovered a limitation in contextualizing the issues related to empowerment at the community level, as empowerment meant different things to the community members. To ensure we developed a culturally appropriate study, we modified our initial study design to incorporate a factorial design with 3 arms: peer support and education to tackle sociocultural barriers through trusted networks, financial empowerment to address economic barriers, and a control. This guaranteed that all the most relevant issues that were identified at the formulation phase were incorporated into the final study design.

Limitations

This study demonstrates some key strengths that enhance its validity and potential impact. First, the use of HCD with a CBPR ensured that the research was deeply rooted in local realities, capturing nuanced sociocultural, economic, and structural barriers to RI. The study also prioritized cultural relevance and community ownership by engaging caregivers, religious leaders, and health workers in this formative phase, which are critical for intervention sustainability. Second, the transition from a 2-arm to a 3-arm cRCT design encouraged mixed insights with evidence from comparable settings, which strengthened the robustness of our findings and addressed gaps by providing interventions to bridge financial and sociocultural barriers, aligning with the need for a multicomponent strategy in gender-inequitable settings. Third, the study's emphasis on culturally tailored communication and community-led advocacy reflects best practices in global health literature, where trust-building and behavioral change strategies have proven more effective. Finally, the research design explicitly addressed intersectional vulnerabilities, such as poverty, gender norms, and rural access and offers a replicable model for other low-resource settings with similar RI uptake challenges.

This formative work has certain limitations. One limitation of the study was the potential for response bias during data collection, as respondents may have been hesitant to provide candid responses due to unfamiliarity with the research team. Despite efforts to mitigate this bias through the involvement of local female community resource persons, some degree of reluctance or social desirability bias may have influenced the data. Additionally, the study's focus on specific LGAs within Kaduna State may limit the generalizability of findings to other regions or populations with different sociocultural contexts. Furthermore, the study's reliance on self-report data, particularly regarding income and daily activities, may have introduced inaccuracies or inconsistencies. Future research could use more diverse sampling methods and incorporate objective measures to enhance the robustness and applicability of findings.

XSL•FO
RenderX

Conclusions

In conclusion, our formative research aimed to develop a culturally appropriate, women-centric cRCT study focused on enhancing RI service uptake among economically disadvantaged caregivers. Through community engagement using HCD, we gained valuable insights into barriers to immunization service use among mothers. Leveraging community groups for health promotion and advocacy emerged as a powerful strategy to combat vaccine hesitancy and improve public health outcomes. Community-led initiatives were indicated to foster trust, inclusivity, and empowerment, bridging gaps between scientific knowledge and community perceptions. Financial empowerment emerged as a pivotal factor in women's ability to make health-related decisions and access health care services.

Enhancing economic opportunities for women not only promotes agency in health decision-making but also contributes to overall household well-being. Our findings highlight the importance of addressing gender-centered economic disparities to improve women's health outcomes and promote gender equality. The sociocultural context significantly influences health care access and use, highlighting the need for targeted interventions. Religious diversity, early pregnancy, marriage, and financial constraints were identified as key determinants affecting health-seeking behaviors. Addressing these barriers requires tailored approaches that consider community norms, beliefs, and infrastructure challenges.

These findings have informed modifications to our study design, incorporating peer support, education, financial empowerment, and control arms in a factorial design. This ensures a holistic approach to address community-level barriers and enhance the cultural appropriateness of our intervention. Our formative research lays a solid foundation for a comprehensive cRCT study aimed at promoting immunization uptake and addressing health disparities among vulnerable populations.

Acknowledgments

Our heartfelt appreciation goes to all the participants who generously shared their experiences and insights during this study. We also extend our thanks to the community leaders, health care providers, and stakeholders in Kaduna State, Northern Nigeria, for their support and collaboration throughout the research process. We express our profound gratitude to the Kaduna State Primary Health Care Board (KSPHCB), the Kaduna State Emergency Routine Immunization Coordination Centre (SERICC), and the Kaduna State Ministry of Human Services and Social Development (MHSSD), as well as the health authorities of the Soba and Jema'a local governments. No generative AI was used in any portion of the manuscript.

Data Availability

The data used and analyzed during the study are available from the corresponding author upon reasonable request, in accordance with ethical standards and subject to confidentiality agreements where applicable.

Conflicts of Interest

None declared.

References

1. Immunization coverage. World Health Organization. 2024 Jul 15. URL: https://www.who.int/news-room/fact-sheets/detail/immunization-coverage [accessed 2025-07-10]
2. Saderi D, Ali M, Ilegbusi PH, Olatoye T, Shah S. Peer review of "Pentavalent vaccine: how safe is it among infants accessing immunization in Nigerian health facilities". JMIRx Med. Preprint posted online on Oct 17, 2024. [doi: 10.2196/66894]
3. 2018 Nigeria DHS key findings.: NPC and ICF; 2019. URL: https://dhsprogram.com/pubs/pdf/SR264/SR264.pdf [accessed 2025-07-10]
4. Gunnala R, Ogbuanu IU, Adegoke OJ, et al. Routine vaccination coverage in Northern Nigeria: results from 40 district-level cluster surveys, 2014-2015. PLoS ONE 2016;11(12):e0167835. [doi: 10.1371/journal.pone.0167835] [Medline: 27936077]
5. Dadari I, Sharkey A, Hoare I, Izurieta R. Analysis of the impact of COVID-19 pandemic and response on routine childhood vaccination coverage and equity in Northern Nigeria: a mixed methods study. BMJ Open 2023 Oct 18;13(10):e076154. [doi: 10.1136/bmjopen-2023-076154] [Medline: 37852768]
6. Odusanya OO, Alufohai EF, Meurice FP, Ahonkhai VI. Determinants of vaccination coverage in rural Nigeria. BMC Public Health 2008 Nov 5;8(1):381. [doi: 10.1186/1471-2458-8-381] [Medline: 18986544]
7. Adebayo BE, Oladokun RE. Immunization coverage in a rural community in southwestern Nigeria. J Vaccines Vaccin 2012;3(4). [doi: 10.4172/2157-7560.1000143]
8. Eze NC, Onwasigwe CN. Effect of mobile phone reminders and recalls on pentavalent vaccines drop-out rate among caregivers accessing childhood immunisation services in a developing city, southeast Nigeria. Asian J Res Infect Dis 2018;1(2). [doi: 10.9734/ajrid/2018/v1i229770]
9. Temitayo-Oboh A, Adegbola A, Dedeke I, et al. Immunization coverage among children aged 0-23 months at a tertiary hospital, southwestern Nigeria: a retrospective study. Babcock Univ Med J 2023;6(1):1-8. [doi: 10.38029/babcockuniv.med.j..v6i1.167]

XSL•FO
RenderX

10. Global routine immunization strategies and practices (GRISP). World Health Organization. 2016 May 29. URL: https://www.who.int/publications/i/item/global-routine-immunization-strategies-and-practices-(grisp) [accessed 2025-07-10]
11. Chen E, Leos C, Kowitt SD, Moracco KE. Enhancing community-based participatory research through human-centered design strategies. Health Promot Pract 2020 Jan;21(1):37-48. [doi: 10.1177/1524839919850557] [Medline: 31131633]
12. Zoltowski CB, Oakes WC, Cardella ME. Students' ways of experiencing human‐centered design. J Eng Edu 2012 Jan;101(1):28-59. [doi: 10.1002/j.2168-9830.2012.tb00040.x]
13. Nguyen Ngoc H, Lasa G, Iriarte I. Human-centred design in industry 4.0: case study review and opportunities for future research. J Intell Manuf 2022;33:35-76. [doi: 10.1007/s10845-021-01796-x] [Medline: 34149201]
14. Steen M. Human-centered design as a fragile encounter. Design Issues 2012 Jan;28(1):72-80. [doi: 10.1162/DESI_a_00125]
15. Giacomin J. What is human centred design? Des J 2014 Dec;17(4):606-623. [doi: 10.2752/175630614X14056185480186]
16. Lyon AR, Koerner K. User-centered design for psychosocial intervention development and implementation. Clin Psychol (New York) 2016 Jun;23(2):180-200. [doi: 10.1111/cpsp.12154] [Medline: 29456295]
17. Forlizzi J. Moving beyond user-centered design. Interactions 2018;25(5):22-23. [doi: 10.1145/3239558]
18. Hass C. Understanding the human-centered design process. In: Edmunds M, Hass C, Holve E, editors. Consumer Informatics and Digital Health: Springer; 2019:145-162. [doi: 10.1007/978-3-319-96906-0_8]
19. Susanto H, Ibrahim F, Nazmudeen SH, Mohiddin F, Setiana D. Human-centered design to enhance the usability, human factors, and user experience within digital destructive ecosystems. In: de Pablos PO, Lytras MD, editors. Global Challenges and Strategic Disruptors in Asian Businesses and Economies: IGI Global Scientific Publishing; 2021:76-94. [doi: 10.4018/978-1-7998-4787-8.ch005]
20. Mahachi K, Kessels J, Boateng K, et al. Zero- or missed-dose children in Nigeria: contributing factors and interventions to overcome immunization service delivery challenges. Vaccine (Auckl) 2022 Sep 2;40(37):5433-5444. [doi: 10.1016/j.vaccine.2022.07.058] [Medline: 35973864]
21. Guignard A, Praet N, Jusot V, Bakker M, Baril L. Introducing new vaccines in low- and middle-income countries: challenges and approaches. Expert Rev Vaccines 2019 Feb;18(2):119-131. [doi: 10.1080/14760584.2019.1574224] [Medline: 30689472]
22. Cherian T, Arora N, MacDonald NE. The global vaccine action plan monitoring and evaluation/accountability framework: perspective. Vaccine (Auckl) 2020 Jul 14;38(33):5384-5386. [doi: 10.1016/j.vaccine.2020.04.036] [Medline: 32381481]
23. Vaccine safety: examine the evidence. HealthyChildren.org. 2022 Sep 6. URL: https://www.healthychildren.org/English/safety-prevention/immunizations/Pages/Vaccine-Studies-Examine-the-Evidence.aspx [accessed 2025-07-10]
24. Creswell JW, Plano Clark VL. Designing and Conducting Mixed Methods Research: SAGE Publications; 2017.
25. Tenny S, Brannan JM, Brannan GD. Qualitative study. National Library of Medicine. 2022. URL: https://www.ncbi.nlm.nih.gov/books/NBK470395 [accessed 2025-07-10]
26. Willis E, Gundacker C, Harris M, Mameledzija M. Improving immunization and health literacy through a community-based approach. Stud Health Technol Inform 2020 Jun 25;269:142-152. [doi: 10.3233/SHTI200028] [Medline: 32593989]
27. Willis E, Sabnis S, Hamilton C, et al. Improving immunization rates through community-based participatory research: community health improvement for Milwaukee's children program. Prog Community Health Partnersh 2016;10(1):19-30. [doi: 10.1353/cpr.2016.0009] [Medline: 27018351]
28. Betsch C, Bach Habersaat K, Deshevoi S, et al. Sample study protocol for adapting and translating the 5C scale to assess the psychological antecedents of vaccination. BMJ Open 2020 Mar 10;10(3):e034869. [doi: 10.1136/bmjopen-2019-034869] [Medline: 32161160]
29. Claypoole VL, Neigel AR, Szalma JL. Perceptions of supervisors and performance: a thematic analysis. Proc HFES Annu Meet 2017 Sep;61(1):1740-1744. [doi: 10.1177/1541931213601916]
30. Khan SN. Qualitative research method: grounded theory. Int J Bus Manag 2014;9(11):224-233. [doi: 10.5539/ijbm.v9n11p224]
31. Harte R, Glynn L, Rodríguez-Molinero A, et al. A human-centered design methodology to enhance the usability, human factors, and user experience of connected health systems: a three-phase methodology. JMIR Hum Factors 2017 Mar 16;4(1):e8. [doi: 10.2196/humanfactors.5443] [Medline: 28302594]
32. Tuckerman J, Kaufman J, Danchin M. Effective approaches to combat vaccine hesitancy. Pediatr Infect Dis J 2022 May 1;41(5):e243-e245. [doi: 10.1097/INF.0000000000003499] [Medline: 35213864]
33. Larson HJ, Jarrett C, Eckersberger E, Smith DMD, Paterson P. Understanding vaccine hesitancy around vaccines and vaccination from a global perspective: a systematic review of published literature, 2007-2012. Vaccine (Auckl) 2014 Apr 17;32(19):2150-2159. [doi: 10.1016/j.vaccine.2014.01.081] [Medline: 24598724]
34. Kaufman J, Ryan R, Walsh L, et al. Face-to-face interventions for informing or educating parents about early childhood vaccination. Cochrane Database Syst Rev 2018 May 8;5(5):CD010038. [doi: 10.1002/14651858.CD010038.pub3] [Medline: 29736980]
35. Bedford H, Attwell K, Danchin M, Marshall H, Corben P, Leask J. Vaccine hesitancy, refusal and access barriers: the need for clarity in terminology. Vaccine (Auckl) 2018 Oct 22;36(44):6556-6558. [doi: 10.1016/j.vaccine.2017.08.004] [Medline: 28830694]
36. Kaduru CC, Mbagwu GC, Aadum DK, et al. Using community theater to improve demand for vaccination services in the Niger Delta Region of Nigeria. BMC Proc 2023 Jul 3;17(Suppl 7):6. [doi: 10.1186/s12919-023-00263-0] [Medline: 37400839]

XSL•FO
RenderX

37. Das P, Shukla S, Shrivastav V, et al. The role of community mobilization in people's healthcare-seeking behavior during the COVID-19 vaccination journey: select case studies from Madhya Pradesh. Humanit Soc Sci Commun 2024;11(1):791. [doi: 10.1057/s41599-024-03275-z]
38. Hopkins KL, Underwood T, Iddrisu I, et al. Community-based approaches to increase COVID-19 vaccine uptake and demand: lessons learned from four UNICEF-supported interventions. Vaccines (Basel) 2023 Jun 30;11(7):1180. [doi: 10.3390/vaccines11071180] [Medline: 37514996]
39. Hong SA. COVID-19 vaccine communication and advocacy strategy: a social marketing campaign for increasing COVID-19 vaccine uptake in South Korea. Humanit Soc Sci Commun 2023;10(1):109. [doi: 10.1057/s41599-023-01593-2] [Medline: 36942012]
40. O'Leary ST, Opel DJ, Cataldi JR, et al. Strategies for improving vaccine communication and uptake. Pediatrics 2024 Mar 1;153(3):e2023065483. [doi: 10.1542/peds.2023-065483] [Medline: 38404211]
41. Enhancing public trust in COVID-19 vaccination: the role of governments. OECD. URL: https://www.oecd.org/en/publications/2021/05/enhancing-public-trust-in-covid-19-vaccination-the-role-of-governments_36749555.html [accessed 2025-07-10]
42. Gilmore B, Gerlach N, Abreu Lopes C, et al. Community engagement to support COVID-19 vaccine uptake: a living systematic review protocol. BMJ Open 2022 Sep 20;12(9):e063057. [doi: 10.1136/bmjopen-2022-063057] [Medline: 36127122]
43. Ashfield S, Donelle L, Uppal G, Bauer MA, Kothari A. Community organization perspectives on COVID-19 vaccine hesitancy and how they increased COVID-19 vaccine confidence: a Canadian Immunization Research Network, social sciences and humanities network study. Front Public Health 2023;11:1258742. [doi: 10.3389/fpubh.2023.1258742] [Medline: 37849717]
44. Media: community-led efforts are the key to ending this pandemic. Vaccine Equity Cooperative. 2021 Sep 14. URL: https://vaccineequitycooperative.org/news/media-community-led-efforts-are-the-key-to-ending-this-pandemic [accessed 2025-07-10]
45. Afolabi AA, Ilesanmi OS. Addressing COVID-19 vaccine hesitancy: lessons from the role of community participation in previous vaccination programs. Health Promot Perspect 2021;11(4):434-437. [doi: 10.34172/hpp.2021.54] [Medline: 35079587]
46. Grewal M, Mushtaq A, Chopra T. "It's worth a shot… or is it?" Notes from the grassroots on vaccine hesitancy and bridging gaps. Infect Control Hosp Epidemiol 2022 Nov;43(11):1750-1752. [doi: 10.1017/ice.2021.356] [Medline: 34338180]
47. Singh P, Dhalaria P, Kashyap S, et al. Strategies to overcome vaccine hesitancy: a systematic review. Syst Rev 2022 Apr 26;11(1):78. [doi: 10.1186/s13643-022-01941-4] [Medline: 35473819]
48. Abdullah M, Ahmad T, Kazmi T, et al. Community engagement to increase vaccine uptake: quasi-experimental evidence from Islamabad and Rawalpindi, Pakistan. PLoS ONE 2022;17(12):e0274718. [doi: 10.1371/journal.pone.0274718] [Medline: 36454856]
49. Parker D, Mills S, Abbey J. Effectiveness of interventions that assist caregivers to support people with dementia living in the community: a systematic review. JBI Database System Rev Implement Rep 2008;6(13):484-544. [doi: 10.11124/01938924-200806130-00001]
50. Rico-Blázquez M, Quesada-Cubo V, Polentinos-Castro E, et al. Health-related quality of life in caregivers of community-dwelling individuals with disabilities or chronic conditions: a gender-differentiated analysis in a cross-sectional study. BMC Nurs 2022 Mar 29;21(1):69. [doi: 10.1186/s12912-022-00845-x] [Medline: 35351100]
51. Wang-Schweig M, Kviz FJ, Altfeld SJ, Miller AM, Miller BA. Building a conceptual framework to culturally adapt health promotion and prevention programs at the deep structural level. Health Promot Pract 2014 Jul;15(4):575-584. [doi: 10.1177/1524839913518176] [Medline: 24396122]
52. Healey P, Stager ML, Woodmass K, et al. Cultural adaptations to augment health and mental health services: a systematic review. BMC Health Serv Res 2017 Jan 5;17(1):8. [doi: 10.1186/s12913-016-1953-x] [Medline: 28056967]
53. Balci S, Spanhel K, Sander L, Baumeister H. Protocol for a systematic review and meta-analysis of culturally adapted internet- and mobile-based health promotion interventions. BMJ Open 2020 Nov 9;10(11):e037698. [doi: 10.1136/bmjopen-2020-037698] [Medline: 33168550]
54. Nittas V, Daniore P, Chavez SJ, Wray TB. Challenges in implementing cultural adaptations of digital health interventions. Commun Med (Lond) 2024 Jan 5;4(1):7. [doi: 10.1038/s43856-023-00426-2] [Medline: 38182750]
55. Remme M, Vassall A, Fernando G, Bloom DE. Investing in the health of girls and women: a best buy for sustainable development. BMJ 2020 Jun 2;369:m1175. [doi: 10.1136/bmj.m1175] [Medline: 32487585]
56. Van Minh H, Kim Phuong NT, Saksena P, James CD, Xu K. Financial burden of household out-of pocket health expenditure in Viet Nam: findings from the National Living Standard Survey 2002–2010. Soc Sci Med 2013 Nov;96:258-263. [doi: 10.1016/j.socscimed.2012.11.028]
57. Sirag A, Mohamed Nor N. Out-of-pocket health expenditure and poverty: evidence from a dynamic panel threshold analysis. Healthcare (Basel) 2021;9(5):536. [doi: 10.3390/healthcare9050536]
58. Aragie H, Negash HK, Getnet M, et al. Barriers to healthcare access: a multilevel analysis of individual- and community-level factors affecting female youths' access to healthcare services in Senegal. BMC Health Serv Res 2025;25:607. [doi: 10.1186/s12913-025-12761-2]

XSL•FO
RenderX

59. State of inequality: childhood immunization. World Health Organization. 2016. URL: https://cdn.who.int/media/docs/default-source/gho-documents/16-dec-final-for-web-16147-state-of-inequality-in-childhood-immunization.pdf?sfvrsn=ac6c954c_2 [accessed 2025-07-10]
60. The link between women's health and women's economic power. Gates Foundation. URL: https://www.gatesfoundation.org/ideas/articles/womens-health-economic-power [accessed 2025-07-10]
61. Karimli L, Lecoutere E, Wells CR, Ismayilova L. More assets, more decision-making power? Mediation model in a cluster-randomized controlled trial evaluating the effect of the graduation program on women's empowerment in Burkina Faso. World Dev 2021 Jan;137:105159. [doi: 10.1016/j.worlddev.2020.105159]
62. Banerjee S, Gogoi P. Exploring the role of financial empowerment in mitigating the gender differentials in subjective and objective health outcomes among the older population in India. PLoS ONE 2023;18(1):e0280887. [doi: 10.1371/journal.pone.0280887] [Medline: 36689542]
63. Davidson PM, McGrath SJ, Meleis AI, et al. The health of women and girls determines the health and well-being of our modern world: a white paper from the International Council on Women's Health Issues. Health Care Women Int 2011 Oct;32(10):870-886. [doi: 10.1080/07399332.2011.603872] [Medline: 21919625]
64. Pachauri S, Rao ND. Gender impacts and determinants of energy poverty: are we asking the right questions? Curr Opin Environ Sustain 2013 Jun;5(2):205-215. [doi: 10.1016/j.cosust.2013.04.006]
65. Tracey G, Olivia B, Jean M, Megan H, Sagri S. Why does gender matter for immunization? Vaccine (Auckl) 2024 Apr 8;42(Suppl 1):S91-S97. [doi: 10.1016/j.vaccine.2022.11.071] [Medline: 36513537]
66. Tiwana MH, Smith J. Faith and vaccination: a scoping review of the relationships between religious beliefs and vaccine hesitancy. BMC Public Health 2024 Jul 6;24(1):1806. [doi: 10.1186/s12889-024-18873-4] [Medline: 38971784]
67. Gender barriers and behavioural and social drivers analysis for immunization in Nigeria. UNICEF. 2025 Feb. URL: https://www.unicef.org/nigeria/reports/gender-barriers-and-behavioural-and-social-drivers-analysis-immunization-nigeria [accessed 2025-07-10]
68. Merten S, Martin Hilber A, Biaggi C, et al. Gender determinants of vaccination status in children: evidence from a meta-ethnographic systematic review. PLoS ONE 2015;10(8):e0135222. [doi: 10.1371/journal.pone.0135222] [Medline: 26317975]
69. Nyasulu BJ, Heidari S, Manna M, Bahl J, Goodman T. Gender analysis of the World Health Organization online learning program on Immunization Agenda 2030. Front Glob Womens Health 2023;4:1230109. [doi: 10.3389/fgwh.2023.1230109] [Medline: 38152380]
70. Matthew Ayodele A, Fasasi MI, Rejoice Uche O, Gideon Ikemdinachi N, Henry Ugochukwu U. Factors associated with full childhood vaccination coverage among young mothers in Northern Nigeria. Pan Afr Med J 2024;47:4. [doi: 10.11604/pamj.2024.47.4.37517] [Medline: 38371647]
71. Eti EO, Odiachi A, Dougherty L, Alabi M, Adetunji A, Adedimeji A. "Everything created by a white man is for pagans": understanding the barriers to childhood immunization in North-Eastern and North-Western Nigeria. medRxiv. Preprint posted online on Mar 30, 2024. [doi: 10.1101/2024.03.29.24305068]

Abbreviations

CBPA: community-based participatory approach
CBPR: community-based participatory research
cRCT: cluster randomized controlled trial
FGDs: focus group discussions
GRISP: Global Routine Immunization Strategies and Practices
HCD: human-centered design
HCD: human-centered design
Hib: *Haemophilus influenzae* type b
KIIs: key informant interviews
LGAs: local government areas
LQAS: Lot Quality Assurance Sampling
PAPA: Performance Assessment for Performance Management
PHC: primary health center
RI: routine immunization
RIBS: Routine Immunization Buddy System

XSL•FO
RenderX

Edited by A Mavragani; submitted 01.05.24; peer-reviewed by R Mohamed; revised version received 02.05.25; accepted 21.05.25; published 28.10.25.

Please cite as:
Aina M, Igbokwe U, Ibrahim R, Ata I, Atta S, Adaramati T, Oreagba K, Ezeogu E, Uahomo PO, Aigbogun Jr E
Culturally Tailored Women's Empowerment Strategies to Improve Immunization Uptake: Formative Research Using a Human-Centered Design Approach
JMIR Form Res 2025;9:e60089
URL: https://formative.jmir.org/2025/1/e60089
doi:10.2196/60089

© Muyi Aina, Uchenna Igbokwe, Raihana Ibrahim, Itoro Ata, Safiya Atta, Tobiloba Adaramati, Kunle Oreagba, Ernest Ezeogu, Precious Ojo Uahomo, Eric Aigbogun Jr. Originally published in JMIR Formative Research (https://formative.jmir.org), 28.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Representation of Medical Concepts in Emojis Using Medical Subject Headings to Identify Gaps and Opportunities: Cross-Sectional Analysis

Shuhan He[1], MD; Boyu Peng[2], MSc; Suhanee Mitragotri[3], BA; Ahmad Hassan[1], MD; Abdel Badih el Ariss[1], MD; Margarita Monge[4], BA; Norawit Kijpaisalratana[1], MD

[1]Department of Emergency Medicine, Massachusetts General Hospital, 55 Fruit Street, Boston, MA, United States
[2]Massachusetts General Hospital, Institute of Health Professions, Boston, MA, United States
[3]Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA, United States
[4]Department of Psychiatry, University of California, Riverside, Riverside, CA, United States

Corresponding Author:
Ahmad Hassan, MD
Department of Emergency Medicine, Massachusetts General Hospital, 55 Fruit Street, Boston, MA, United States

Abstract

Background: Emoji are a universal visual language widely used in digital communication; yet, their representation of medical concepts remains limited. The introduction of medical emojis, such as the anatomical heart and lungs, highlights their potential for health care communication, but significant gaps persist.

Objective: This study aims to systematically analyze the representation of medical concepts in emoji by mapping Medical Subject Headings (MeSH) to Unicode emojis, identifying gaps in medical emoji representation, and proposing areas for future emoji development.

Methods: A cross-sectional study was conducted using the sentence transformer model. Digital resources, including the MeSH thesaurus and Unicode emoji set version 15.0 (Unicode Consortium), were used. Embeddings for 2077 MeSH terms and 3055 emojis were generated, and cosine similarity scores were calculated to evaluate the semantic alignment between MeSH terms and emoji descriptions. A threshold of 0.7 was set to indicate a strong semantic match.

Results: The analysis revealed significant variations in emoji representation across medical categories. “Geographicals” had the highest match rate (33.33%), whereas “Anatomy” showed only 7.94% matches, with 13 of 163 terms exceeding the similarity threshold. Categories such as “Disciplines and Occupations,” “Information Science,” and “Psychiatry and Psychology” had no matches (0%), highlighting notable gaps. The findings underscore substantial disparities in medical emoji representation, particularly for internal organs, mental health, and specialized disciplines. Limited availability of representative emoji may hinder effective health care communication, especially in digital health contexts. This study emphasizes the potential of artificial intelligence to design emojis that address these gaps and improve inclusivity.

Conclusions: Significant gaps in medical emoji representation across various domains were identified. Future efforts should prioritize underrepresented medical categories and leverage artificial intelligence–driven approaches for emoji development to enhance health care communication and accessibility.

(JMIR Form Res 2025;9:e70130) doi:10.2196/70130

KEYWORDS

emoji; health care communication; language; artificial intelligence in health care; digital health tools

Introduction

Background

Overall, 92% of the world’s population uses emoji [1], resulting in around 2‑3 trillion emojis being sent in text messages annually [2]. Emojis are built on Unicode, a universal technical standard that supports multiple languages, including other visual languages such as Chinese and Arabic, as well as English symbols, enabling cross-channel communication across devices worldwide, including PCs, Macs, mobile devices, and almost all electronic medical records today.

The Unicode Consortium, a nonprofit organization, accepts proposals for new emojis, allowing the general public to indicate what important concepts should be represented in emojis [3].

The release of Unicode version 13.0 (Unicode Consortium) in March 2020 included anatomical representations of the heart and lungs, which was a substantial development in expanding

XSL•FO
RenderX

the range of medically relevant emojis that already existed, such as the stethoscope, syringe, and pill emojis.

These new emojis, due to their integration into the universal digital standard, enhance the representation of the medical field on online technology platforms and are useful for cross-linguistic communication, creating semantic meaning for machine learning analysis, and improving visual analog scales for patient-reported outcomes [4].

However, further progress for a more comprehensive set of medical emojis including the liver, stomach, and kidney has faced significant barriers since 2020 with rejections of these proposals. The reasons for these rejections remain publicly undisclosed, but may be attributed to insufficient evidence of necessity, lack of a unifying standard for medical emojis, and the absence of an objective measure for determining the need for specific health care emojis. These rejections underscore the need to evaluate the scientific evidence for their inclusion, particularly the importance of a comprehensive framework to determine which medical concepts are needed to fill the gaps in the emoji Unicode standard.

Objectives

The Unified Medical Language System (UMLS) serves as a repository for medical terminology, classification, and coding standards, and it improves the ability of computers to understand the clinical meaning in user inquiries for health information. As a standard for encoding health-related concepts, the UMLS provides a comprehensive framework to better understand potential gaps in Unicode emojis. Thus, we aimed to comprehensively analyze UMLS terms in relation to all Unicode emojis to identify gaps and ensure the proper representation of key health care concepts in the Unicode standard.

Methods

Overview

This study used a cross-sectional design to systematically analyze and map the representation of medical concepts using emojis. Our methodology involved integrating digital resources including the Medical Subject Headings (MeSH) thesaurus and the Unicode emoji set version 15.0 (Unicode Consortium).

Data Preparation

MeSH Terms

We use the MeSH thesaurus, a controlled and hierarchically organized vocabulary curated by the National Library of Medicine, for indexing, cataloging, and searching biomedical and health-related information.

MeSH is structured in multiple hierarchical levels, typically up to 12, allowing for very specific indexing and searching of topics. For our analysis, we focused on the top 2 layers, comprising 2077 terms. This selection was made to balance the depth and breadth of the vocabulary, ensuring comprehensive coverage while maintaining manageable complexity.

Emoji

The emoji set from version 15.0, comprising 3055 emojis, was used in our analysis [5]. For each emoji, we used its formal Unicode Technical Committee (UTC) name as the textual representation. Examples of these formal UTC names include “grinning face” (for 😀), “red heart” (for ❤), and “thumbs up” (for 👍). We selected the UTC name rather than the emoji glyph (image) because sentence transformer models require textual input. This approach follows best practices in natural language processing for emoji modeling. Notably, Eisner et al [6] demonstrated that short textual descriptions of emoji can yield high-quality semantic representations, outperforming usage-derived embeddings in many cases [6]. While visual features likely carry additional semantic information, integrating image-based models (eg, CLIP [Contrastive Language-Image Pretraining]) would require a fundamentally different, multimodal pipeline and falls outside the scope of this text-embedding study. Our approach provides a reproducible, text-only baseline for evaluating medical concept alignment.

Model Selection

For our analysis, we used a high-performing sentence transformer model selected from the Massive Text Embedding Benchmark (MTEB) [7], a comprehensive framework designed to evaluate the performance of text-embedding models across a diverse range of tasks. Specifically, we chose the “mxbai-embed-large-v1” model [8], an open-source solution that consistently ranks among the top 10 models for semantic textual similarity performance. This model was selected not only for its exceptional accuracy but also for its efficiency in terms of computational resources. Its lightweight nature ensures accessibility and practicality, facilitating easy replication of our analysis while maintaining high standards of performance.

Calculate Embeddings

Sentence embedding involves converting each sentence or textual input into a continuous vector, with the objective of capturing the semantic meaning and contextual nuances in a distributed representation [9-11]. Using the selected model, we generated embeddings for each emoji description and each MeSH term definition. This process yielded a vector for every entity, effectively representing their semantic attributes in numerical form. For example, “Heart: The hollow, muscular organ that maintains the circulation of the blood” will be converted into the following array: [−0.12453227, −0.12998965, 0.00507377, 0.19754894, −0.4001878, 0.2854124, 0.07951714,...−0.12513563, −0.07079408, −0.17973363, 0.20297979, 0.4145004, 0.01179076]

The shape of the array depends on the model used to convert the text; in the case of the “mxbai-embed-large-v1” model, the shape is (512,).

Calculate Similarity

After transforming tokens into vectors through the embedding process, we use cosine similarity to measure the angle between 2 vectors. This metric determines the similarity between each MeSH term and emoji, producing a score ranging from −1 to 1 for each pair. A score close to −1 indicates that the vectors are diametrically opposed, meaning the words are completely

XSL•FO
RenderX

unrelated and oppositely related in the vector space. A score close to 0 suggests the vectors are orthogonal, indicating no linear relationship and that the words are unrelated or independent in the context of the vector space. A score close to 1 means the vectors are identical in direction, signifying that the words are very similar or essentially the same in the vector space [12,13].

The cosine similarity between 2 vectors A and B is calculated using the formula:

$$\text{similarity}(A_,B_)=(A\cdot B)/(\|A\| \times \|B\|)$$

where $A \cdot B$ represents the dot product of vectors A and B, and $\|A\|$ and $\|B\|$ represent the magnitudes (Euclidean norms) of vectors A and B, respectively. This can also be expressed as:

$$\text{similarity}(A,B)=\sum(Ai\times Bi)/(\sum(Ai2)\times\sum(Bi2))$$

where Ai and Bi are the *i*th components of vectors A and B, and the summations are from *i*=1 to n, with n being the dimension of the vectors. We adopted a cosine similarity threshold of 0.7 as the criterion for a match, indicating a strong similarity between an emoji and a MeSH term. All analyses were performed using Python version 3.10.12 (Python Software Foundation [PSF]).

Statistical Analysis

To assess differences in cosine similarity scores across various MeSH categories, we used the Kruskal-Wallis H test, a nonparametric method suitable for nonnormally distributed data. This test identifies statistically significant differences in group medians. When such differences were found, we performed the Dunn test for pairwise comparisons, applying the Bonferroni correction to adjust for multiple testing. All statistical analyses were conducted using Stata (version 18.0; StataCorp LLC).

Ethical Considerations

This study did not involve human participants, patient data, or identifiable personal information. Therefore, institutional review board approval and informed consent were not required. All data used were publicly available and nonidentifiable, and the research adhered to relevant institutional and national guidelines regarding privacy and ethical conduct. No compensation was provided, as no participants were recruited or involved.

Results

Our analysis generated embeddings for 2077 MeSH terms and 3055 emojis, resulting in over 6 million similarity comparisons. Figure 1 illustrates the heatmap of cosine similarity scores between MeSH terms and emoji. Each pixel represents a similarity score between a specific MeSH term and an emoji. The color scale ranges from blue (low similarity) to red (high similarity), and grid lines segment the MeSH terms and emojis into categories to enhance pattern visibility. The predominance of blue shades indicates that the majority of MeSH terms and emoji pairs have low cosine similarity scores, suggesting that most emojis do not closely match specific medical terms in the MeSH dataset. The sparsity of red or high similarity scores highlights that only a small subset of emojis aligns well with medical terminology.

Figure 1. Heatmap of cosine similarity scores between Medical Subject Headings terms and emojis. Medical Subject Headings terms are categorized as follows: A=Anatomy; B=Organisms; C=Diseases; D=Chemicals and Drugs; E=Analytical, Diagnostic, and Therapeutic Techniques, and Equipment; F=Psychiatry and Psychology; G=Phenomena and Processes; H=Disciplines and Occupations; I=Anthropology, Education, Sociology, and Social Phenomena; J=Technology, Industry, and Agriculture; K=Humanities; L=Information Science; M=Named Groups; N=Health Care; V=Publication Characteristics; Z=Geographicals. MeSH: Medical Subject Headings.

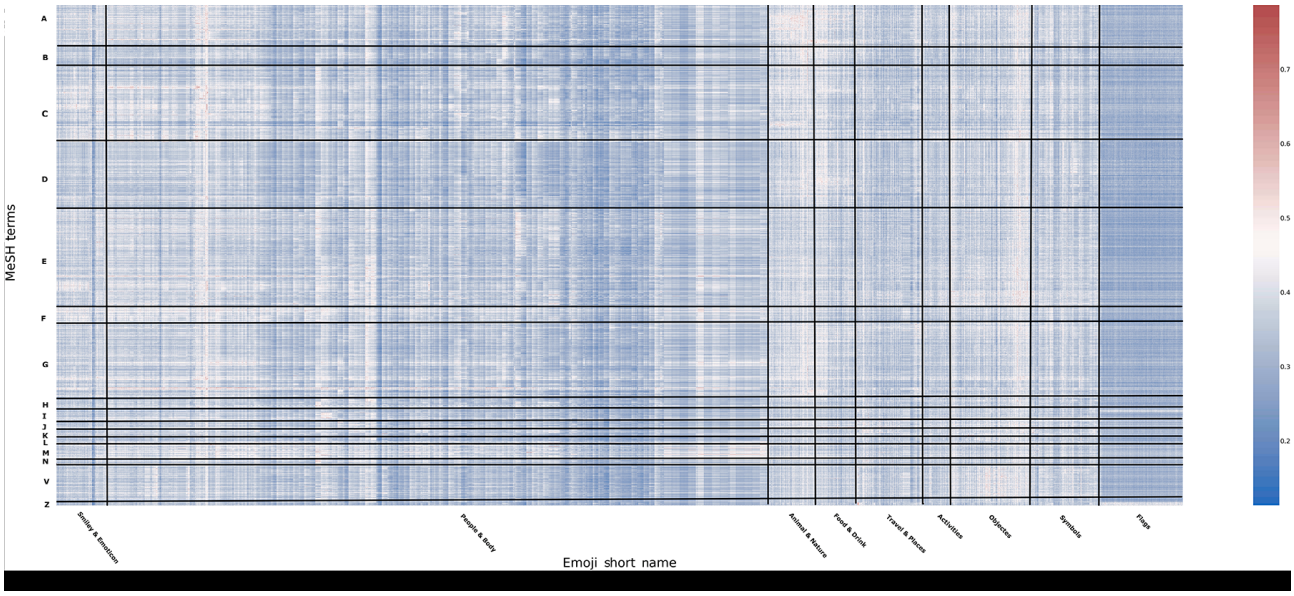

The analysis of cosine similarity scores across MeSH categories mapped to Unicode emojis highlighted significant variability (Table 1).

The median cosine similarity score in the “Geographicals (Z)” category was 0.70 (IQR 0.64‑0.74), indicating a strong alignment between MeSH terms and emojis within this category. This category also showed statistically significantly higher scores compared to all other categories, with *P* values <.001 in each comparison. This was followed by “Diseases (C)” with a median score of 0.68 (IQR 0.55‑0.62) and “Technology,

XSL•FO
RenderX

Industry, and Agriculture (J)" with a median of 0.63 (IQR 0.58 - 0.67). In contrast, "Chemicals and Drugs (D)" recorded the lowest median score at 0.56 (IQR 0.54 - 0.59), suggesting a relatively weaker representation of these concepts in emoji form. Categories such as "Anthropology, Education, Sociology, and Social Phenomena (I)" and "Psychiatry and Psychology (F)" also demonstrated lower median scores.

Table 2 presents the top 10 MeSH terms and emoji pairs with the highest similarity scores, ranging from 0.7747 to 0.8198. The highest similarity score of 0.8198 was observed between "Seedlings" and the seedling emoji ( ), followed by "Grandparents" matched with the family emoji ( ) at 0.8128, and "Nose" with the nose emoji ( ) at 0.8008. The matched pairs encompass various concepts, including anatomical parts ("Nose"), geographical entities ("Europe"), social concepts ("Grandparents"), facilities ("Schools," "Bathroom Equipment," "Toilet Facilities"), and scientific tools ("Microscopy," "Syringes"). Most pairs show direct correspondence between the MeSH term and the emoji, such as "Syringes" with the syringe emoji ( ) and "Schools" with the school emoji ( ). Two MeSH terms, "Bathroom Equipment" and "Toilet Facilities," were both matched with the same restroom emoji ( ), with similarity scores of 0.7769 and 0.7747, respectively. The geographical concept "Europe" was matched with the "globe showing Europe-Africa" emoji ( ) with a similarity score of 0.7907. All MeSH terms and emoji pairs exceeding the similarity threshold are presented in Table S1 and Table S2 in Multimedia Appendix 1.

In the analysis of human anatomy-related MeSH terms and emoji pairs (Table 3), 13 pairs exceeded the similarity threshold, with scores ranging from 0.7014 to 0.8008. The "Nose" and nose emoji ( ) pair maintained its high ranking from the overall top 10, with a score of 0.8008. The mouth emoji ( ) corresponded to multiple anatomical terms: "Mouth" (0.7532), "Stomatognathic System" (0.7400), and "Jaw" (0.7175). Similarly, the bone emoji ( ) is paired with both "Musculoskeletal System" (0.7353) and "Skeleton" (0.7202). Internal organs were also represented, with "Lung" matching the lungs emoji ( ) at 0.7555 and "Heart" matching the heart emoji ( ) at 0.7329.

XSL•FO
RenderX

Table . Median cosine similarity scores for each Medical Subject Headings category and intercategory comparisons.

		Median (IQR)	A[a]	B[b]	C[c]	D[d]	E[e]	F[f]	G[g]	H[h]	I[i]	J[j]	K[k]	L[l]	M[m]	N[n]	V[o]	Z[p]
A		0.60 (0.57-0.64)																
	r		1	1.98	12.12	24.78	4.65	11.02	7.76	3.75	10.34	−4.35	6.49	4.90	1.30	12.31	4.88	−13
	P value		—[q]	>.99	<.001	<.001	<.001	<.001	<.001	.01	<.001	<.001	<.001	<.001	>.99	<.001	<.001	<.001
B		0.60 (0.57-0.62)																
	r		1.98	1	8	18.58	1.82	8.24	4.78	2.21	8.47	−5.46	5.43	3.64	0.06	9.30	3.60	−13.7
	P value		>.99	—	<.001	<.001	>.99	<.001	<.001	>.99	<.001	<.001	<.001	.02	>.99	<.001	.02	<.001
C		0.68 (0.55-0.62)																
	r		12.12	8	1	17.91	−9.01	2.55	−3.40	−2.84	4.24	−11.20	2.37	−0.19	−4.46	3.73	−0.30	−19.90
	P value		<.001	<.001	—	<.001	<.001	.64	.04	.27	.01	<.001	>.99	>.99	<.001	.01	>.99	<.001
D		0.56 (0.54-0.59)																
	r		24.78	18.58	17.91	1	−2.51	−7.83	−1.70	−9.76	−2.65	−17.8	−1.94	−5.52	−10.30	−7.17	−5.72	−26.21
	P value		<.001	<.001	<.001	—	<.001	<.001	<.001	<.001	.48	<.001	>.99	<.001	<.001	<.001	<.001	<.001
E		0.59 (0.56-0.63)																
	r		4.65	1.82	−9.01	−2.51	1	8.26	4.09	1.21	8.15	−7.19	4.89	2.92	−0.99	9.64	2.86	−16
	P value		<.001	>.99	<.001	<.001	—	<.001	.01	>.99	<.001	<.001	<.001	.21	>.99	<.001	.25	<.001
F		0.57 (0.54-0.61)																
	r		11.02	8.24	2.55	−7.83	8.26	1	−4.66	−4.01	2.23	−11.40	1.23	−1.40	−5.39	0.80	−1.52	−19.40
	P value		<.001	<.001	.64	<.001	<.001	—	<.001	.01	>.99	<.001	>.99	>.99	<.001	>.99	>.99	<.001
G		0.58 (0.55-0.62)																
	r		7.76	4.78	−3.40	−1.70	4.09	−4.66	1	−1.01	5.71	−9.06	3.41	1.14	−2.85	5.78	1.06	−17.60
	P value		<.001	<.001	.04	<.001	.01	<.001	—	>.99	<.001	<.001	.04	>.99	.26	<.001	>.99	<.001
H		0.58 (0.56-0.62)																
	r		3.75	2.21	−2.84	−9.76	1.21	−4.01	−1.01	1	5.22	−6.41	3.60	1.63	−1.62	4.68	1.57	−13.50
	P value		.01	>.99	.27	<.001	>.99	.01	>.99	—	<.001	<.001	.02	>.99	>.99	<.001	>.99	<.001

XSL•FO
RenderX

		Median (IQR)	A[a]	B[b]	C[c]	D[d]	E[e]	F[f]	G[g]	H[h]	I[i]	J[j]	K[k]	L[l]	M[m]	N[n]	V[o]	Z[p]
I		0.56 (0.54-0.60)																
	r		10.34	8.47	4.24	−2.65	8.15	2.23	5.71	5.22	1	−11.50	−0.25	−2.83	−6.37	−1.67	−2.95	−18.40
	P value		<.001	<.001	.01	.48	<.001	>.99	<.001	<.001	—	<.001	>.99	.28	<.001	>.99	.19	<.001
J		0.63 (0.58-0.67)																
	r		−4.35	−5.46	−11.20	−17.8	−7.19	−11.40	-9.06	−6.41	−11.50	1	8.33	7.12	4.27	12.18	7.12	−7.09
	P value		<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	—	<.001	<.001	.001	<.001	<.001	<.001
K		0.57 (0.54-0.60)																
	r		6.49	5.43	2.37	−1.94	4.89	1.23	3.41	3.60	−0.25	8.33	1	−2	−4.66	−0.84	−2.08	−13.70
	P value		<.001	<.001	>.99	>.99	<.001	>.99	.04	.02	>.99	<.001	—	>.99	<.001	>.99	>.99	<.001
L		0.58 (0.55-0.62)																
	r		4.90	3.64	−0.19	−5.52	2.92	−1.40	1.14	1.63	−2.83	7.12	−2.00	1	−2.94	1.90	−0.07	−13.30
	P value		<.001	.02	>.99	<.001	.21	>.99	>.99	>.99	.28	<.001	>.99	—	.20	>.99	>.99	<.001
M		0.60 (0.55-0.65)																
	r		1.30	0.06	−4.46	−10.30	−0.99	−5.39	−2.85	−1.62	−6.37	4.27	−4.66	−2.94	1	5.98	2.89	−10.8
	P value		>.99	>.99	<.001	<.001	>.99	<.001	.26	>.99	<.001	.001	<.001	.20	—	<.001	.23	<.001
N		0.57 (.54-.61)																
	r		12.31	9.30	3.73	−7.17	9.64	0.80	5.78	4.68	−1.67	12.18	−0.84	1.9	5.98	1	−2.02	−20.2
	P value		<.001	<.001	.01	<.001	<.001	>.99	<.001	<.001	>.99	<.001	>.99	>.99	<.001	—	>.99	<.001
V		0.58 (.55-.61)																
	r		4.88	3.60	−0.30	−5.72	2.86	−1.52	1.06	1.57	−2.95	7.12	−2.08	−0.07	2.89	−2.02	1	−13.3
	P value		<.001	.02	>.99	<.001	.25	>.99	>.99	>.99	.19	<.001	>.99	>.99	.23	>.99	—	<.001
Z		0.70 (.64-.74)																
	r		−13.00	−13.7	−19.90	−26.21	−16.00	−19.40	−17.60	−13.50	−18.40	−7.09	−13.70	−13.30	−10.80	−20.20	−13.30	1

XSL•FO
RenderX

	Median (IQR)	A[a]	B[b]	C[c]	D[d]	E[e]	F[f]	G[g]	H[h]	I[i]	J[j]	K[k]	L[l]	M[m]	N[n]	V[o]	Z[p]
P value		<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	—

[a]A: Anatomy.
[b]B: Organisms.
[c]C: Diseases.
[d]D: Chemicals and Drugs.
[e]E: Analytical, Diagnostic and Therapeutic Techniques, and Equipment.
[f]F: Psychiatry and Psychology.
[g]G: Phenomena and Processes.
[h]H: Disciplines and Processes.
[i]I: Anthropology, Education, Sociology, and Social Phenomena.
[j]J: Technology, Industry, and Agriculture.
[k]K: Humanities.
[l]L: Informatics Science.
[m]M: Named Groups.
[n]N: Health Care.
[o]V: Publication Characteristics.
[p]Z: Geographicals.
[q]Not applicable.

Table . Top 10 Medical Subject Headings terms and emoji pairs with the highest similarity.

MeSH[a] heading	Unicode CLDR[b] short name	Emoji	Cosine similarity score
Seedlings	seedling		0.8198
Grandparents	family: man, woman, girl, boy		0.8128
Nose	nose		0.8008
Europe	globe showing Europe-Africa		0.7907
Pedestrians	person walking		0.7833
Schools	school		0.7827
Syringes	syringe		0.7801
Microscopy	microscope		0.7776
Bathroom equipment	restroom		0.7769
Toilet facilities	restroom		0.7747

[a]MeSH: Medical Subject Headings.
[b]CLDR: Common Locale Data Repository

XSL•FO
RenderX

Table . Human anatomy Medical Subject Headings term and emoji pairs exceeding the similarity threshold.

MeSH[a] heading	Emoji	Cosine similarity score
Nose		0.8008
Lung		0.7555
Mouth		0.7532
Eye		0.7408
Stomatognathic system		0.7400
Musculoskeletal system		0.7353
Heart		0.7329
Ear		0.7260
Cheek		0.7233
Skeleton		0.7202
Jaw		0.7175
Central nervous system		0.7172
Acinar cells		0.7014

[a]MeSH: Medical Subject Headings.

The heatmap visualizes cosine similarity scores between anatomical MeSH terms and emojis, and colors the cell with the highest score (Figure 2). The x-axis lists anatomical MeSH sorted from highest to lowest match pairs, while the y-axis represents different emojis. A color scale from dark blue (low similarity) to dark red (high similarity) is used, with white and light shades representing intermediate values. The majority of the heatmap exhibits light blue colors, indicating generally low similarity scores between most anatomical terms and emoji. For full figure access, refer to [14].

Figure 2. Heatmap of cosine similarity scores between anatomical Medical Subject Headings terms and emojis. MeSH: Medical Subject Headings.

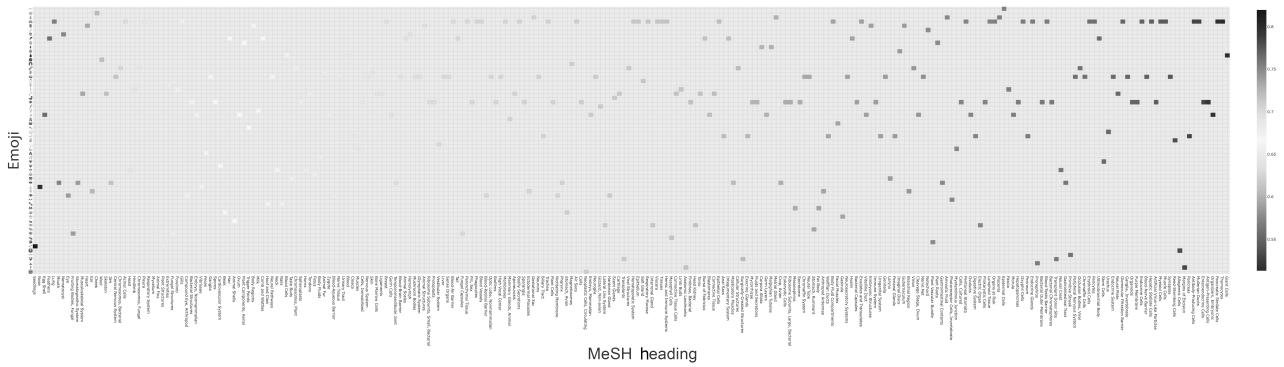

Table S1 in Multimedia Appendix 1 reveals significant gaps in emoji representation for human anatomical terms, with unmatched pairs highlighted in red. Out of 163 terms examined, 150 (92.02%) did not meet the established similarity threshold. Several vital organs show notably low similarity scores. The pancreas at 0.5293 and the liver at 0.6285. The urogenital system scored 0.5877, with genitalia and urinary tract at 0.5875 and 0.6441, respectively. The endocrine system exhibited one of the lowest scores at 0.5539. The cardiovascular system scored 0.6682, with the heart as an exception at 0.7329. The nervous system scored 0.6395, showing a notable discrepancy between its central (0.7172) and peripheral (0.5598) components.

Table 4 shows the percentage of matching MeSH top-level categories with corresponding emojis. “Geographicals” has the highest match rate at 33.33%, followed by “Named Groups” (11.94%) and “Technology, Industry, and Agriculture” (11.36%). Most categories have less than 10% matches, with “Anatomy” at 7.94%. “Organisms” (1.35%) and “Chemicals and Drugs” (0.84%) show very few matches. Three categories—“Disciplines and Occupations,” “Information Science,” and “Psychiatry and Psychology”—have no matches (0%). The matching percentages range from 0% to 33.33%, revealing significant variation in emoji representation across medical and scientific domains.

XSL•FO
RenderX

Table . Percentage of matching Medical Subject Headings top-level categories with corresponding emojis.

MeSH[a] top-level categories	Percentage of matching
Geographicals	33.33
Named Groups	11.94
Technology, Industry, and Agriculture	11.36
Anatomy	7.94
Humanities	7.69
Analytical, Diagnostic and Therapeutic Techniques, and Equipment	5.31
Health Care	5
Diseases	4.66
Phenomena and Processes	4.08
Publication Characteristics	3.25
Anthropology, Education, Sociology, and Social Phenomena	1.82
Organisms	1.35
Chemicals and Drugs	0.84
Disciplines and Occupations	0
Information Science	0
Psychiatry and Psychology	0

[a]MeSH: Medical Subject Headings.

Discussion

Principal Findings

Unicode regularly introduces new figures relevant to medicine, reflecting the growing importance of visual communication in health care. Recent additions such as the anatomical heart ( ) and lungs ( ) emoji [15] demonstrate this commitment. Our analysis revealed high similarity scores between these anatomical emojis and their corresponding MeSH terms (heart: 0.7329, lung: 0.7555), indicating their effectiveness in representing these concepts. However, these additions are just the beginning of what could be a more comprehensive set of medical emojis. However, our study was limited to text-based descriptions of emojis and did not incorporate the emoji glyphs themselves. Emoji glyphs often contain rich visual features—such as stylization, implied movement, or symbolic elements—that are not fully captured in their official Unicode names. Future work could incorporate visual-semantic models such as CLIP, which align text and images in a shared embedding space and have been shown to capture multi-modal meaning with high fidelity (Radford et al [16]). Such an approach could improve alignment between medical terms and emoji representations, particularly in anatomy and other visually grounded domains.

Our research identified significant variations in emoji-MeSH term matching across categories, highlighting uneven visual representation in medical communication. The “Geographicals” category showed the highest match rate at 33.33%, likely due to universally recognizable symbols for locations and features. This finding highlights a broader pattern in emoji design and use. The high alignment observed in the “Geographicals” category is likely due to the iconic and visually intuitive nature of geographical emojis, such as globe symbols ( ), maps ( ), and flags ( ), which are easily associated with MeSH terms such as “Europe” or “Mountains.” These emojis rely on universally understood visual semantics, enabling higher alignment with general-language terms. In contrast, many medical concepts—especially internal anatomical structures or abstract scientific terms—lack such iconic visual representations in the Unicode emoji set. This reflects a design bias in the current emoji lexicon toward common, general-purpose concepts, rather than specialized domains such as healthcare or biomedical science. As a result, while categories such as geography are well represented, critical areas of medicine remain visually underrepresented. This discrepancy supports the central hypothesis of our study: that emoji coverage is uneven and that systematic gaps exist in the representation of medical and health-related terminology. Conversely, categories such as “Disciplines and Occupations,” “Information Science,” and “Psychiatry and Psychology” had no matches. This absence underscores a critical gap in representing complex and abstract medical and scientific concepts. For instance, “Disciplines and Occupations” includes various medical specialties lacking specific visual symbols, while “Information Science” involves nonvisual concepts like data science. The lack of matches in “Psychiatry and Psychology” is particularly significant, given the increasing importance of mental health discussions. This gap may hinder effective communication and destigmatization efforts in these fields.

These findings are further substantiated by our detailed analysis of anatomical terms, as shown in Table S1 in Multimedia Appendix 1. Specific anatomical structures or medical equipment that currently lack emoji representations should be

XSL•FO
RenderX

prioritized. For example, organs such as the kidney, liver, and pancreas are fundamental to many medical discussions yet lack direct emoji representations. The urogenital system and endocrine system also demonstrate poor representation. Interestingly, while some external features, such as the nose and mouth, have high similarity scores, internal organs and systems generally lack appropriate emoji matches. The need for these emoji is supported by previous calls from professional medical societies, such as the American Society of Nephrology's advocacy for a kidney emoji [17]. Our study reinforces this need, as evidenced by the low matching percentage in the "Anatomy" category (7.94%) and the absence of these specific organs in our high-similarity matches. This discrepancy highlights the need for a more balanced approach in developing medical emojis, ensuring that both visible and internal anatomical structures are adequately represented to enhance medical communication and education.

Beyond their role in medical terminology representation, emojis offer unique usability advantages in health care communication. Images and symbols can be processed more quickly than text, making emojis potentially valuable for rapid information transfer in medical informatics. This is particularly relevant in environments such as clinical documentation, emergency communication, and patient engagement tools, where quick comprehension is essential. Automation and usability improvements—such as integrating emoji into electronic health records or artificial intelligence (AI)–driven clinical decision support systems—could further enhance their role in medical communication. However, for emojis to be effectively used in health care settings, their medical relevance and standardization must be improved.

This need for more comprehensive medical emoji representation aligns with the increasing interest in using emoticons and emojis in health care communication over the past couple of years. Emojis have been explored to be useful in public health communication, such as conveying important hand hygiene and infection control protocols to the public [18,19]. Emojis have also been used to develop novel scales for patients to use to communicate their mood, pain, or satisfaction with clinical care [20]. Research has shown that emojis also show promise in provider-provider communication in secure clinical texting systems, conveying new and salient information [21]. The universality of the use of emoji is indicative of their utility and value in the medical field.

Significance

While traditional emoji creation processes continue, new approaches could further enhance the representation of medical concepts. The rapid advancement of generative AI in health care, capable of creating enhanced medical images and facilitating drug design discovery [22,23], presents exciting possibilities for emoji development. Leveraging these AI capabilities, we could address the need for more diverse and specific medical emojis through dynamic AI generation and sticker-based systems. AI-generated emojis could offer a wider range of customizable medical symbols, adapting to specific needs in real time, while complementary sticker-based approaches could provide greater flexibility in representing complex medical concepts. These innovative methods could significantly enhance the visual communication of medical information, bridging the gaps we identified in the current emoji set and providing more tailored options for health-related discussions.

Future research could build upon our findings through several complementary enhancements. First, domain-specific language models such as BioBERT or SciBERT could be used to generate embeddings that more accurately capture biomedical semantics, potentially reducing false positives and improving MeSH–emoji alignment for clinically nuanced terms. Second, multimodal approaches, such as CLIP [16], could enable direct comparison between MeSH terms and the visual glyphs of emoji. This would allow researchers to evaluate how visual meaning, beyond textual labels, might influence semantic alignment—particularly for categories such as anatomy where appearance plays a critical role in interpretation. Together, these strategies could substantially enhance the resolution and accuracy of concept-emoji mapping in future work. By implementing these approaches, we could potentially address the disparity observed in emoji representation across different medical and scientific domains, moving toward a more comprehensive and nuanced visual language for health care communication. One promising direction involves integrating visual-semantic models such as CLIP to directly compare emoji glyphs to medical concepts. This could allow more accurate alignment for visual categories such as anatomy, where symbolic form and visual context are essential for meaning.

Limitations

This study has several limitations that should be considered when interpreting the results. First, our analysis was based on cosine similarity scores between MeSH terms and emoji descriptions, which may not capture all nuances of semantic relationships. The effectiveness of emoji in real-world medical communication may differ from these computational assessments. Second, we focused on the top 2 layers of MeSH terms, which, while providing a broad overview, may have excluded more specific medical concepts that could benefit from emoji representation. Third, MeSH terminology is highly domain-specific, while emoji descriptions tend to be more general, leading to potential mismatches in similarity scores. This discrepancy may affect the accuracy of identifying true gaps in medical emoji representation. Similarly, the high match rate in the "Geographicals" category suggests that universally recognizable symbols may skew similarity scores, making it difficult to distinguish meaningful medical associations.

In addition, our study used a specific sentence transformer model for generating embeddings, and results might vary with different models or embedding techniques. Finally, cultural and regional differences in emoji interpretation were not accounted for in this study, which could impact the global applicability of our findings.

Conclusion

Our study reveals significant disparities in emoji representation across medical domains, offering evidence-based insights for future development. By prioritizing underrepresented areas and

XSL•FO
RenderX

embracing innovative approaches, we can create a more comprehensive set of medical emojis. These advancements have the potential to enhance health care communication, improve patient-provider interactions, and ultimately contribute to better health outcomes.

Acknowledgments

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. Patients and the public were not involved in the design, conduct, reporting, or dissemination plans of our research. As there were no participants involved in this study, there are no dissemination plans to participants or related patient and public communities. The lead author affirms that this manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

Data Availability

All data analyzed in this study are publicly available. Medical Subject Headings terms were obtained from the National Library of Medicine, and emoji descriptions were sourced from the Unicode Consortium [23].

The pretrained language model used (mxbai-embed-large-v1) is available from Hugging Face [24]. The code used for analysis will be made available in a public repository upon publication.

Full Figure 2 can be accessed on the web [13].

Authors' Contributions

All authors contributed to extracting data, writing, and reviewing the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Cosine similarity scores between human anatomical MeSH (Medical Subject Headings) terms and corresponding emojis and MeSH Term–emoji pairs exceeding the similarity threshold.
[DOCX File, 36 KB - formative_v9i1e70130_app1.docx]

References

1. Daniel J. The most frequently used emoji of 2021. UNICODE. 2021. URL: https://home.unicode.org/emoji/emoji-frequency/ [accessed 2024-03-20]
2. Marko K. “Depends on Who I’m Writing To”—the influence of addressees and personality traits on the use of emoji and emoticons, and related implications for forensic authorship analysis. Front Commun 2022;7:840646. [doi: 10.3389/fcomm.2022.840646]
3. What is unicode? [online]. Unicode. 2023. URL: https://www.unicode.org/standard/WhatIsUnicode.html [accessed 2024-03-20]
4. He S, Renne A, Argandykov D, Convissar D, Lee J. Comparison of an emoji-based Visual Analog Scale with a numeric rating scale for pain assessment. JAMA 2022 Jul 12;328(2):208-209. [doi: 10.1001/jama.2022.7489] [Medline: 35819433]
5. Full emoji list, v17.0. Unicode. URL: https://unicode.org/emoji/charts/full-emoji-list.html [accessed 2025-09-15]
6. Eisner B, Rocktäschel T, Augenstein I, Bosnjak M, Riedel S. Emoji2vec: learning emoji representations from their description. Presented at: Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media; Nov 1, 2016; Austin, TX, USA URL: http://aclweb.org/anthology/W16-62 [doi: 10.18653/v1/W16-6208]
7. Muennighoff N, Tazi N, Magne L, et al. Massive text embedding benchmark. arXiv. Preprint posted online on Oct 1, 2022 URL: https://ui.adsabs.harvard.edu/abs/2022arXiv221007316M [accessed 2025-09-22] [doi: 10.48550/arXiv.2210.07316]
8. Li X, Li J. AnglE-optimized text embeddings. arXiv. Preprint posted online on 2023 URL: https://arxiv.org/pdf/2309.12871 [accessed 2025-09-22] [doi: 10.48550/arXiv.2309.12871]
9. Asudani DS, Nagwani NK, Singh P. Impact of word embedding models on text analytics in deep learning environment: a review. Artif Intell Rev 2023 Feb 22:1-81. [doi: 10.1007/s10462-023-10419-1] [Medline: 36844886]
10. Vithanage D, Yu P, Wang L, Deng C. Contextual word embedding for biomedical knowledge extraction: a rapid review and case study. J Healthc Inform Res 2024 Mar;8(1):158-179. [doi: 10.1007/s41666-023-00157-y] [Medline: 38273979]
11. Wang Y, Liu S, Afzal N, et al. A comparison of word embeddings for the biomedical natural language processing. J Biomed Inform 2018 Nov;87:12-20. [doi: 10.1016/j.jbi.2018.09.008] [Medline: 30217670]

XSL•FO
RenderX

12. Kim S, Fiorini N, Wilbur WJ, Lu Z. Bridging the gap: incorporating a semantic similarity measure for effectively mapping PubMed queries to documents. J Biomed Inform 2017 Nov;75:122-127. [doi: 10.1016/j.jbi.2017.09.014] [Medline: 28986328]
13. Sarwar TB, Noor NM, Saef Ullah Miah M. Evaluating keyphrase extraction algorithms for finding similar news articles using lexical similarity calculation and semantic relatedness measurement by word embedding. PeerJ Comput Sci 2022;8:e1024. [doi: 10.7717/peerj-cs.1024] [Medline: 35875631]
14. Kijpaisalratana N. Heatmap of cosine similarity scores. Plotly Chart Studio. URL: https://plotly.com/~norawit29/1.embed [accessed 2025-09-01]
15. Burge J. First look: new emojis coming to iOS in 2020. Emojipedia. 2020. URL: https://blog.emojipedia.org/first-look-new-emojis-coming-to-ios-in-2020/ [accessed 2024-04-03]
16. Radford A, Kim JW, Hallacy C, et al. Learning transferable visual models from natural language supervision. arXiv. Preprint posted online on Feb 26, 2021 URL: https://arxiv.org/abs/2103.00020 [accessed 2025-09-15]
17. Seethapathy H, Beach J, Gillcrist M, et al. The case for a kidney emoji. Am J Kidney Dis 2022 Aug;80(2):155-157. [doi: 10.1053/j.ajkd.2022.03.012] [Medline: 35513182]
18. Lotfinejad N, Assadi R, Aelami MH, Pittet D. Emojis in public health and how they might be used for hand hygiene and infection prevention and control. Antimicrob Resist Infect Control 2020 Feb 10;9(1):27. [doi: 10.1186/s13756-020-0692-2] [Medline: 32041666]
19. Lotfinejad N, Tartari E, Sauser J, Fankhauser-Rodriguez C, Pires D, Pittet D. Are emojis ready to promote the WHO 5 moments for hand hygiene in healthcare? Antimicrob Resist Infect Control 2022 Oct 26;11(1):127. [doi: 10.1186/s13756-022-01164-1] [Medline: 36289506]
20. Herring AMR, Craven MP, Mughal F, et al. Potential of using visual imagery to revolutionise measurement of emotional health. Arch Dis Child 2020 Jul;105(7):690-693. [doi: 10.1136/archdischild-2019-317758] [Medline: 31974299]
21. Lai D, Lee J8, He S. Emoji for the medical community-challenges and opportunities. JAMA 2021 Sep 7;326(9):795-796. [doi: 10.1001/jama.2021.8409] [Medline: 34547099]
22. Koohi-Moghadam M, Bae KT. Generative AI in medical imaging: applications, challenges, and ethics. J Med Syst 2023 Aug 31;47(1):94. [doi: 10.1007/s10916-023-01987-4] [Medline: 37651022]
23. Vartiainen H, Tedre M. How text-to-image generative AI is transforming mediated action. IEEE Comput Graph Appl 2024;44(2):12-22. [doi: 10.1109/MCG.2024.3355808] [Medline: 38285567]
24. Hugging Face. URL: https://huggingface.co/mixedbread-ai/mxbai-embed-large-v1 [accessed 2025-09-15]

Abbreviations

AI: artificial intelligence
CLIP: Contrastive Language-Image Pretraining
MeSH: Medical Subject Headings
MTEB: Massive Text Embedding Benchmark
UMLS: Unified Medical Language System
UTC: Unicode Technical Committee

Edited by A Mavragani; submitted 16.12.24; peer-reviewed by D Chrimes, M Hussain; revised version received 23.04.25; accepted 11.08.25; published 14.10.25.

Please cite as:
He S, Peng B, Mitragotri S, Hassan A, Ariss ABE, Monge M, Kijpaisalratana N
Representation of Medical Concepts in Emojis Using Medical Subject Headings to Identify Gaps and Opportunities: Cross-Sectional Analysis
JMIR Form Res 2025;9:e70130
URL: https://formative.jmir.org/2025/1/e70130
doi:10.2196/70130

© Shuhan He, Boyu Peng, Suhanee Mitragotri, Ahmad Hassan, Abdel Badih el Ariss, Margarita Monge, Norawit Kijpaisalratana. Originally published in JMIR Formative Research (https://formative.jmir.org), 14.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Application of Large Language Models in Data Analysis and Medical Education for Assisted Reproductive Technology: Comparative Study

Noriyuki Okuyama[1*], MSc; Mika Ishii[1], BSc; Yuriko Fukuoka[1], BSc; Hiromitsu Hattori[1,2,3], PhD; Yuta Kasahara[1,2,3], MD, PhD; Tai Toshihiro[1,2], MD, PhD; Koki Yoshinaga[1], MD, PhD; Tomoko Hashimoto[1], MD, PhD; Koichi Kyono[1,2,3*], MD, PhD

[1]Kyono ART Clinic Takanawa, Takanawa Court 5F, 3-13-1 Takanawa, Minato-ku, Tokyo, Japan

[2]Kyono ART Clinic Sendai, Sendai, Japan

[3]Kyono ART Clinic Morioka, Morioka, Japan

[*]these authors contributed equally

Corresponding Author:
Koichi Kyono, MD, PhD
Kyono ART Clinic Takanawa, Takanawa Court 5F, 3-13-1 Takanawa, Minato-ku, Tokyo, Japan

Abstract

Background: Recent studies have demonstrated that large language models exhibit exceptional performance in medical examinations. However, there is a lack of reports assessing their capabilities in specific domains or their application in practical data analysis using code interpreters. Furthermore, comparative analyses across different large language models have not been extensively conducted.

Objective: The purpose of this study was to evaluate whether advanced artificial intelligence (AI) models can analyze data from template-based input and demonstrate basic knowledge of reproductive medicine. Four AI models (GPT-4, GPT-4o, Claude 3.5 Sonnet, and Gemini Pro 1.5) were evaluated for their data analytical capabilities through numerical calculations and graph rendering. Their knowledge of infertility treatment was assessed using 10 examination questions developed by experts.

Methods: First, we uploaded data to the AI models and furnished instruction templates using the chat interface. The study investigated whether the AI models could perform pregnancy rate analysis and graph rendering, based on blastocyst grades according to Gardner criteria. Second, we assessed model diagnostic capabilities based on specialized knowledge. This evaluation used 10 questions derived from the Japanese Fertility Specialist Examination and the Embryologist Certification Exam, along with chromosome imaging. These materials were curated under the supervision of certified embryologists and fertility specialists. All procedures were repeated 10 times per AI model.

Results: GPT-4o achieved grade A output (defined as achieving the objective with a single output attempt) in 9 out of 10 trials, outperforming GPT-4, which achieved grade A in 7 out of 10. The average processing times for data analysis were 26.8 (SD 3.7) seconds for GPT-4o and 36.7 (SD 3) seconds for GPT-4, whereas Claude failed in all 10 attempts. Gemini achieved an average processing time of 23 (SD 3) seconds and received grade A in 6 out of 10 trials, though occasional manual corrections were needed. Embryologists required an average of 358.3 (SD 9.7) seconds for the same tasks. In the knowledge-based assessment, GPT-4o, Claude, and Gemini achieved perfect scores (9/9) on multiple-choice questions, while GPT-4 showed a 60% (6/10) success rate on 1 question. None of the AI models could reliably diagnose chromosomal abnormalities from karyotype images, with the highest image diagnostic accuracy being 70% (7/10) for Claude and Gemini.

Conclusions: This rapid processing demonstrates the potential for these AI models to significantly expedite data-intensive tasks in clinical settings. This performance underscores their potential utility as educational tools or decision support systems in reproductive medicine. However, none of the models were able to accurately interpret and diagnose using medical images.

(JMIR Form Res 2025;9:e70107) doi:10.2196/70107

KEYWORDS

artificial intelligence; large language model; data analysis; education; infertility

XSL•FO
RenderX

Introduction

The term “artificial intelligence” (AI) was first defined at the Dartmouth Summer Research Project on Artificial Intelligence in 1956. Over half a century after this seminal conference, advances in computational power and exponential growth in data available for machine learning have ushered in the third AI boom, commencing after 2010 [1]. A particularly significant milestone was reached in November 2022 with the release of ChatGPT (OpenAI), a large language model (LLM) developed by OpenAI. This model had an unprecedented impact, amassing a global user base of 100 million within just 2 months of its public launch. The field has since progressed beyond generative AI to what is now termed multimodal AI, capable of integrating and drawing inferences from multiple input modalities, including image, voice, and text. This evolution represents a significant leap in AI capabilities and potential applications. As AI continues to reshape business models and medical education, there are growing expectations for its role in enhancing organizational performance and driving transformative change in various sectors, including health care [2-4]. Implications of these advances for medical practice, research, and health care delivery systems are profound and call for both careful and bold decision-making to drive further progress.

LLMs have achieved substantial improvements through self-attention mechanisms and transformer architectures. Fine-tuning enables domain-specific customization, but it requires high computational costs and expert oversight. To promote health care digital transformation, more accessible methods are needed. Cloud-based code interpreter functionality offers a solution by enabling LLMs to automatically generate and execute code from natural language instructions. This capability supports diverse file formats (CSV, XLSX, DOCX, TXT, and PDF) and performs numerical analysis and visualization online. In medical research, this technology can significantly enhance efficiency through rapid dataset processing, statistical analysis, and graph generation [5]. Another important application is the use of these models as chatbot tools. While there have been initial reports on the use of early GPT models in infertility treatment, there is a notable lack of literature evaluating practical support capabilities and response quality of updated LLMs in this field [6,7]. Updating the current assessment of AI capabilities could provide clearer insights into the challenges and opportunities for AI use in medical education and patient counseling in reproductive medicine. This evaluation is essential to identify specific areas in which AI can effectively complement human expertise to improve patient care in fertility treatment contexts.

The field of infertility treatment has seen a growing number of reports on AI and machine learning applications, including survival rate prediction based on gradient-based localization of blastocyst images, iDAscore, TESE sperm search, and detection of oocyte fragility during intracytoplasmic sperm injection, among other topics [8-11]. However, implementing these advanced technologies in individual clinics often requires specific equipment, software, and specialized personnel, which can be challenging. Additionally, while real-time capabilities for quality control and treatment management in fertility care are of paramount importance, there remains a significant disparity in computational infrastructure capacity across health care institutions. Notably, many medical facilities, particularly smaller clinics, lack the requisite computing infrastructure necessary for executing sophisticated AI or machine learning models. This infrastructural limitation constitutes a substantial barrier to the widespread implementation of advanced data-driven approaches in reproductive medicine [12]. Recognizing this, we hypothesized that selecting and using more accessible AI tools could promote digital transformation among infertility treatment professionals. This study evaluated 4 accessible AI models for infertility treatment applications, assessing their data processing capabilities and knowledge performance. These 4 AI models—GPT-4 (OpenAI), GPT-4o (OpenAI), Claude 3.5 Sonnet (Anthropic), and Gemini Pro 1.5 (Google)—were selected because they (1) are pretrained on trillion-token-scale corpora, (2) can be accessed immediately through simple account registration without local computational resources, and (3) therefore represent the most widely reproducible options for clinicians and educators seeking to leverage cutting-edge LLM technology in real-world settings.

This assessment comprised 2 key components. First, the ability of these models to process and analyze clinical data using the code interpreter feature, specifically calculating and visualizing pregnancy rates stratified by blastocyst trophectoderm grades, and second, their performance on a comprehensive knowledge test in reproductive medicine. This test consisted of 10 multiple-choice questions, each with 5 options, covering various aspects of assisted reproductive technology (ART). The questions were meticulously curated and validated by a multidisciplinary panel of experts, including certified clinical embryologists, gynecologists, and urologists specializing in reproductive medicine. By examining these 2 crucial areas, the study assessed the current capabilities and limitations of AI in processing complex reproductive medical data, as well as its potential role in medical education in this specialized field.

Methods

AI Models and the Risk of Personal Information Leakage

This study used GPT-4, GPT-4o, Claude 3.5 Sonnet, and Gemini Pro 1.5 for data analysis. The dataset was structured with analysis items in the first column and corresponding data in rows. Data formatting was consistent, avoiding mixed character types and unnecessary spaces. Each cell contained a single data point. All identifying information for patients and staff was removed from datasets before analysis. The study did not use ChatGPT’s memory feature or Gemini’s extended functionalities. No prompt engineering techniques were used to optimize outputs. Null data were included, and each dataset remained in the original format as output from the database software. Additionally, no prompt engineering techniques or fine-tuning were used to optimize the outputs. Generative AI tools were not used for the generation of references or the writing of this manuscript. All citations were verified manually, and all textual content was written and reviewed by the authors.

XSL•FO
RenderX

However, AI tools were used for grammar checking prior to professional English editing.

Data Collection

This study analyzed pregnancy rates by embryo grade in frozen-thawed embryo transfer cycles using data from January 2017 to July 2024, encompassing 5361 cycles from 2276 patients. The dataset was prepared as a 362 KB Microsoft Excel file with 32 columns and 5321 rows. The dataset was initially generated using Claris FileMaker, a relational database platform, and then exported into Microsoft Excel format for use in AI models. In the exported file, the first row contained the names of analysis variables (eg, patient age, embryo quality, and pregnancy outcome), while subsequent rows held corresponding numeric or categorical values for each case. Each cell was carefully formatted to contain only a single, discrete data element—without the use of delimiters, such as slashes or commas—to ensure clarity and prevent parsing errors during AI processing. This standardized structure enabled consistent interpretation across platforms, regardless of differences in input limitations. Claude's procedure differed slightly due to lower file size limitations. It used a reduced dataset of 5 columns and 3000 rows in CSV format to avoid "length over" errors, while maintaining the same basic operation and template prompts as the other AI models (Figure 1A). In contrast, GPT-4, GPT-4o, and Gemini were evaluated using the full dataset (32 columns and 5321 rows), allowing direct comparison of their performance under uniform data conditions.

XSL•FO
RenderX

Figure 1. Data analysis procedure and evaluation framework for large language models in assisted reproductive technology clinical data processing. (**A**) Sample dataset structure showing patient treatment data from frozen-thawed embryo transfer cycles (January 2017-July 2024; 5361 cycles from 2276 patients) formatted for artificial intelligence (AI) model input with variables including patient age, embryo quality, and pregnancy outcomes. (**B**) Study workflow showing systematic evaluation protocol where 4 AI models (GPT-4, GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro) were tested 10 times each using standardized template prompts, with performance graded as A, B, or C. (**C**) Target visualization output showing pregnancy rates stratified by Gardner criteria trophectoderm grades (AA, BA, AB, BB, and BC) from clinic data. ET: embryo transfer; GS: gestational sac.

XSL•FO
RenderX

Data Analysis

The methodology involved uploading data to each AI model and applying a standard instruction template (Textbox 1). Performance was rated on a scale “A” for correct output on the first attempt, “B” for correction within 2 attempts, and “C” for errors or requiring 3 or more attempts. This process was repeated 10 times for each LLM, with a new chat initiated after each round to prevent memory influence (Figure 1B). This approach allowed systematic evaluation of AI capabilities in processing and visualizing complex clinical data while accounting for platform-specific constraints. The analysis sought to produce a bar graph with data labels categorized by Gardner criteria (Figure 1C). Compared with embryologists who routinely perform data processing, 3 embryologists carried out calculations under similar conditions using Microsoft Excel.

Textbox 1. Template prompt of data analysis.

Pregnancy rates are calculated by “Trophectoderm”.
X-axis is “Trophectoderm” in the order AA, BA, AB, BB, BC
Y-axis is 0-100%.
Please add data labels for each value
Draw a bar graph for the graph.
Graph title is “Pregnancy Rates of Gardner Criteria”
Pregnancy rate definition:
Numerator: “GS” total number
Denominator: “numberofET” number of data entries
Search for "GS" and "numberofET" in the column and calculate

Assessment of Knowledge of Infertility

GPT-4, GPT-4o, Claude, and Gemini used a set of 10 multiple-choice questions with options a-e (Textbox 2). Questions were input into each AI interface, and the process was repeated 10 times, initiating a new chat session for each round to ensure independent responses (Figure 2A). Questions were carefully curated. Overall, 3 questions were prepared by a senior embryologist, while board-certified specialists from the Japan Society for Reproductive Medicine in gynecology and urology each selected 3 questions from the gynecology specialist exam (2016 - 2018; Figure 2B). As a final question, the assessment included image-based diagnosis using chromosome diagrams. The image used for analysis was 600×450 pixels (Figure 2C). Gemini occasionally hesitated on highly specialized questions, such as Question 10, sometimes offering only 3 choices or declining to answer with statements like, “I’m a language model and don’t have the capacity to help with that.” In such instances, AI was prompted to provide a definitive choice among the given options.

XSL•FO
RenderX

Textbox 2. Multiple-choice questions provided to GPT-4, GPT-4o, Claude 3.5 Sonnet, and Gemini Pro 1.5.

1. Select the correct option regarding the maturation of oocytes before fertilization in chronological order of changes:
   a. GV → GVBD → Metaphase I → Anaphase I → Telophase I → Metaphase II
   b. GVBD → GV → Metaphase I → Telophase I → Anaphase I → Metaphase II
   c. GV → GVBD → Metaphase I → Telophase I → Anaphase I → Metaphase II
   d. GVBD → GV → Anaphase I → Metaphase I → Telophase I → Metaphase II
   e. GV → GVBD → Anaphase I → Metaphase I → Telophase I → Metaphase II

1. Despite injecting only one sperm into the oocyte cytoplasm during intracytoplasmic sperm injection (ICSI), three pronuclei may sometimes be observed during fertilization confirmation. Which of the following is the most likely cause?
   a. Parthenogenesis
   b. Polyspermy
   c. Retained sperm acrosome
   d. Normal fertilization
   e. Failure to extrude the second polar body

1. Select the correct statement about glutamine from the following:
   a. A low-cytotoxicity buffer used for culture media in air, which does not require CO_2 for buffering capacity
   b. An amino acid essential for embryo development but can produce toxic ammonia that affects embryo development; therefore, it is often added to culture media in a stabilized form
   c. Added to culture media to prevent the growth of contaminants in case of contamination
   d. Has antioxidant, osmotic pressure maintenance, and detoxification effects; also acts as an anti-adhesive agent to prevent embryos from adhering to glass or plastic
   e. Forms a chelate with magnesium ions, lowering their concentration, thus preventing abnormal metabolic shifts towards glycolysis and promoting embryo development

1. Which of the following is correct?
   a. About 80% of male infertility is caused by varicocele
   b. The most common chromosomal abnormality causing male infertility is Klinefelter syndrome
   c. ICSI is performed using a method called TESE, which retrieves sperm from the epididymis
   d. FSH primarily binds to receptors located in Leydig cells in the testes
   e. LH promotes the secretion of Müllerian inhibiting substance (MIS)

1. Which of the following statements about spermatogenesis is incorrect? Choose one.
   a. Sertoli cells are stimulated by FSH and are involved in spermatogenesis
   b. Leydig cells are stimulated by LH and produce testosterone
   c. Sertoli cells respond to androgen and produce factors necessary for spermatogenesis
   d. Spermatogonia have androgen receptors and differentiate into sperm cells
   e. The hypothalamus-pituitary-gonadal axis plays an important role in spermatogenesis

1. For patients eligible for TESE (testicular sperm extraction), which of the following are predictors of sperm retrieval? Choose two.
   a. FSH
   b. Testicular volume
   c. Complete AZF deletion
   d. 46XX male
   e. DAZ microdeletion

XSL•FO
RenderX

1. Which conditions have normal ovarian function? Choose two.
    a. Turner syndrome
    b. Klinefelter syndrome
    c. Swyer syndrome
    d. Asherman syndrome
    e. Mayer-Rokitansky-Küster-Hauser (MRKH) syndrome

1. This question is about the correct statement regarding the luteinization process in humans. The correct answer is:
    a. Luteinization is completed within about 2 days after ovulation
    b. The mature corpus luteum after luteinization is also called the “corpus rubrum”
    c. Angiogenesis is induced during the luteinization phase
    d. Granulosa cells release intracellular granules
    e. Estrogen production decreases

1. This question is about hormones primarily produced by the adrenal glands in females. The correct answer is:
    a. Androstenedione
    b. Testosterone
    c. Dehydroepiandrosterone sulfate (DHEA-S)
    d. Dihydrotestosterone (DHT)
    e. Sex hormone-binding globulin (SHBG)

1. Here is a karyotype image. Please select the possible diagnosis based on this image: (Figure 2C)
    a. Klinefelter syndrome
    b. Turner syndrome
    c. Down syndrome
    d. Chromosomal translocation
    e. Normal

XSL•FO
RenderX

Figure 2. Knowledge assessment framework for large language models in reproductive medicine education. (**A**) Study protocol for evaluating artificial intelligence (AI) model performance on fertility specialist examination questions, with 10 independent trials per model using fresh chat sessions. (**B**) Question distribution showing sources: 3 questions from a senior embryologist, 6 questions from board-certified specialists (Japan Society for Reproductive Medicine gynecology and urology specialist exams, 2016‐2018), and 1 image-based karyotype diagnosis question. (**C**) Karyotype analysis test image (600×450 pixels) used for chromosomal abnormality diagnosis assessment across all AI models.

XSL•FO
RenderX

Ethical Considerations

This is a retrospective study using patient treatment data from January 2017 to July 2024. Informed consent for data analysis and secondary use was obtained from all patients. Patient data were not used in the testing of specialized knowledge. All data used in this study were deidentified to ensure the privacy and confidentiality of human candidates. This study was conducted with the approval of the institutional ethics committee (5805 - 240224) and did not receive any special funding.

Results

Data Analysis

GPT-4 achieved grade A performance in 7 out of 10 attempts, while GPT-4o excelled with 9 grade A performances. GPT-4o also demonstrated superior processing speed, averaging 26.8 (SD 3.7) seconds compared to GPT-4's 36.7 (SD 3) seconds. GPT-4 encountered issues with column recognition in 3 instances, resolved by requesting another search (Table 1). GPT-4o had minor issues with Y-axis scaling in graph outputs, defaulting to 0% - 60% instead of 0% - 100%, as instructed, but this was easily corrected. Postcorrection response times further highlighted GPT-4o's efficiency (41 s vs GPT-4's 90.7, SD 5.7 s). Claude consistently performed at grade C, failing to process the entire dataset and producing incomplete outputs. Despite attempts to specify row numbers and implement programs to read to the final row, Claude was unable to produce the desired numerical results or graphical representations. Gemini prompted users to select a graphing tool or library (such as Microsoft Excel, Google Sheets, or Python's Matplotlib) for visualization. It also occasionally included null values in graph representations without exclusion. When a staff member with the fastest data processing skills performed numerical calculations and generated graphical charts using the same datasets, approximately 7 minutes were required. Furthermore, values evaluated as "A" or "B" by the LLM were consistent with those verified by the staff members.

Table . Performance comparison of multiple large language models and human embryologists in clinical data processing for assisted reproductive technology. All procedures were repeated 10 times per artificial intelligence model and 3 times for embryologists.

Variables	GPT-4 (n=10)	GPT-4o (n=10)	Claude 3.5 Sonnet (n=10)	Gemini 1.5 Pro (n=10)	Embryologist (n=3)
Output grade, n (%)					
Grade A	7 (70)	9 (90)	0 (0)	6 (60)	3 (100)
Grade B	3 (30)	1 (10)	0 (0)	4 (40)	0 (0)
Grade C	0 (0)	0 (0)	10 (100)	0 (0)	0 (0)
Calculation time (s), mean (SD)					
Grade A	36.7 (3)	26.8 (3.7)	N/A[a]	23 (3)	358.3 (9.7)
Grade B	90.7 (5.7)	41(N/A)	N/A	48.5 (3)	N/A
Type of hallucinations, n (%)					
Search error	3 (30)	0 (0)	0 (0)	0 (0)	0 (0)
Calculation error	0 (0)	0 (0)	10 (100)	0 (0)	0 (0)
Visualization error	0 (0)	1 (10)	0 (0)	4 (40)	0 (0)

[a]N/A: not applicable.

Assessment of Knowledge

The knowledge evaluation was conducted independently from the data processing assessment. This analysis of AI performance on medical questions reveals distinct patterns among models. For questions 1 - 9, GPT-4 showed high accuracy, with a single inconsistency on Question 5, selecting option "c" incorrectly 4 times, resulting in a 60% (6/10) success rate for that question. GPT-4o, Claude, and Gemini, however, achieved perfect scores on these initial 9 questions. Question 10, involving complex medical interpretation, proved challenging for all models (Table 2). GPT-4 correctly identified the diagnosis in 2 instances; however, it incorrectly diagnosed "Down syndrome" on 2 occasions and "Normal" on 6 occasions. GPT-4o failed to select the correct diagnosis and instead produced outputs of "Down syndrome" and "Chromosomal translocation." Responses from both GPT models exhibited a high degree of randomness. Claude accurately identified the correct diagnosis in 7 instances, but misclassified 3 cases as "Normal." Similarly, Gemini provided the correct diagnosis 7 times, while incorrectly diagnosing "Down syndrome" and "Normal" in 3 instances. No occurrences of Turner syndrome were observed in outputs from any of the LLMs.

XSL•FO
RenderX

Table . Knowledge assessment results for 4 large language models on 10 reproductive medicine specialist examination questions, showing correct answers out of 10 independent trials per model. Overall, 9 questions are multiple-choice (questions 1-9) and 1 is karyotype image diagnosis (question 10).

Question number	GPT-4	GPT-4o	Claude 3.5 Sonnet	Gemini 1.5 Pro
1	10	10	10	10
2	10	10	10	10
3	10	10	10	10
4	10	10	10	10
5	6	10	10	10
6	10	10	10	10
7	10	10	10	10
8	10	10	10	10
9	10	10	10	10
10	2	0	7	7

Discussion

Principal Findings

This study evaluated the abilities of 4 major LLMs in clinical data analysis and knowledge assessment in ART. For data analysis tasks involving pregnancy rate calculations stratified by blastocyst grades, human embryologists achieved the highest accuracy with no errors across all trials, though they required significantly more time to complete the task. Among the AI models, GPT-4o demonstrated the best balance of speed and accuracy, while GPT-4 showed intermediate performance with occasional column recognition errors. Gemini exhibited the fastest processing speed but lacked precision, often requiring manual corrections for visualization issues. Notably, Claude 3.5 Sonnet consistently failed to complete the data analysis tasks despite multiple attempts with modified datasets.

All 4 models achieved perfect scores on reproductive medicine knowledge questions derived from specialist certification examinations, demonstrating comprehensive theoretical understanding. However, none of them could reliably interpret chromosomal karyotype images, revealing a critical gap between theoretical knowledge and visual diagnostic capabilities. These findings suggest that LLMs can serve as valuable tools for accelerating routine data processing in clinical settings and as educational resources for foundational learning in reproductive medicine, while they emphasize the continued necessity of human expertise for complex diagnostic tasks.

Data Analysis

Infertility treatment strives to achieve successful patient pregnancies; thus, reviewing treatment outcomes and establishing key performance indicators are crucial [13]. With the increasing availability of large datasets, there are high expectations for machine learning and AI advances to revolutionize health care. However, actual applications of data-driven medicine remain relatively limited at present [12]. This study revealed substantial performance differences among LLMs in clinical data processing (Table 1). While the specific architectural differences and training datasets of each model remain proprietary, the superior performance of GPT-4o likely reflects optimizations in its code interpreter functionality and enhanced context-handling capabilities. The complete failure of Claude, despite its large context window, suggests that file processing and numerical computation capabilities vary significantly across platforms, regardless of their general language understanding abilities. We calculated pregnancy rates for frozen-thawed embryo transfers categorized by Gardner criteria. GPT-4o demonstrated strong proficiency in both calculations and data retrieval, completing these tasks in an average of 26.8 (SD 3.7) seconds with 90% (9/10) accuracy. Similarly, Gemini 1.5 Pro exhibited rapid computational speed (mean 23, SD 3 s), but it occasionally included extraneous elements beyond the specified instructions. Claude 3.5 Sonnet, despite being provided with a truncated dataset due to upload constraints, struggled to produce the desired output. In contrast, embryologists performed the calculations without error but required an average of 358.3 (SD 9.7) seconds—significantly slower than the LLMs. Previous studies have explored the capabilities of LLMs in medical contexts, including scientific writing with simulated data [5] and the broader potential of AI to transform infertility care. However, real-world applications of data-driven approaches in reproductive medicine remain relatively limited, despite their transformative promise [12]. The performance differences observed, particularly Claude's inability to process clinical datasets and Gemini's handling of null values, highlight the practical challenges of implementing these technologies. While previous research has emphasized the need for advanced systems to enhance ART outcomes and support clinical decision-making, our findings offer concrete insights into implementation barriers in actual practice. GPT-4o reduced processing time from approximately 6 minutes for human embryologists to less than 30 seconds, representing more than a simple efficiency improvement. This acceleration enables embryologists to redirect their expertise toward critical hands-on procedures such as intracytoplasmic sperm injection, vitrification, and embryo assessment, rather than spending valuable time on routine data analysis. Looking toward future implementations, this rapid processing capability could potentially enable continuous quality monitoring systems. As

XSL•FO
RenderX

LLM integration with laboratory information management systems advances, clinics may be able to implement continuous background analysis of treatment outcomes, thus enabling rapid detection of deviations from established benchmarks.

In our dataset, we avoided the simultaneous use of abbreviations like “GS” and “FHB” to reduce hallucination risks. Our experience suggests that using unabbreviated terms without spaces, for example, “GestationalSac” and “FetalHeartbeat,” and avoiding similar alphabetic sequences can yield more accurate results. Although “in-context learning” techniques—such as background priming, role assignment, iterative correction, rule enforcement, and self-reflection—can enhance LLM performance [14,15], these techniques were intentionally not applied in this study to assess baseline output. Additionally, the prompts were not optimized for individual models. Nevertheless, the observed variation in output suggests that model-specific prompt engineering may significantly influence performance and improve alignment with clinical datasets. LLMs can perform medical data analysis substantially faster than manual methods, suggesting their potential to address workforce shortages and reduce costs in implementing data-driven health care.

Assessment of Knowledge

This study evaluated the performance of LLMs in reproductive medicine using a specialized set of 10 questions developed under supervision by fertility specialists and certified embryologists. These results demonstrated high accuracy across most models, with GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro achieving perfect scores. These results demonstrated high accuracy across most models, with GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro all achieving perfect scores. GPT-4 showed some errors, particularly with Question 5, which focused on spermatogenesis (Table 2). These findings align with previous reports that LLMs perform well in medical examinations. For example, GPT-4 achieved an accuracy of 80% - 100% on the United States Medical Licensing Examination and an average score of 76.3% over 3 attempts on the United Kingdom Medical Licensing Assessment. In Japan, GPT-4 passed the national medical licensing exams from 2018 to 2023, though limitations were noted in responses related to cultural context and national guidelines [16-18]. However, our study identified a limitation in image interpretation. Although the models demonstrated an understanding of chromosomal diagrams and diagnostic concepts, they struggled with direct image diagnosis. This limitation likely stems from insufficient exposure to specialized medical images during training, as general-purpose LLMs are not specifically trained on cytogenetic datasets. Successful karyotype analysis requires recognition of subtle morphological features and banding patterns that may not be adequately represented in their training data. Furthermore, the resolution and quality of training images may vary significantly from clinical standards. Future improvements in this domain would require specialized training on high-resolution, standardized karyotype images with consistent formatting and annotation protocols.

It is worth noting that specialized AI models have shown high accuracy in certain medical imaging tasks, such as the diagnosis of cervical cancer pathology and the detection of esophageal cancer in endoscopy [19,20]. However, these applications primarily function as support tools. Thus, addressing potential risks, ensuring safety and efficacy, and considering ethical responsibilities remain essential for AI’s broader application to health care [21]. The strong performance on theoretical knowledge questions suggests significant potential for LLMs in embryology education, particularly for smaller fertility clinics that may lack experienced senior embryologists or comprehensive training programs. Embryologists typically learn through hands-on experience during routine work, but facilities with limited case volumes or staff may struggle to provide adequate educational opportunities. This perspective is supported by recent work demonstrating that LLMs can effectively generate high-quality educational content and classification datasets tailored to specific pedagogical contexts, such as classroom dialog categorization [22,23]. Practical implementation could include interactive question and answer sessions for certification exam preparation, standardized explanations of complex procedures, case-based learning scenarios, and 24/7 availability for concept clarification. Their ability to process and analyze complex medical information, coupled with their limitations in specialized tasks, such as image interpretation, suggests a promising but carefully defined role in medical education and clinical support. The rapid evolution of AI is transforming medical education from traditional knowledge transfer to distributed cognitive systems, which requires educators to acquire collaborative skills with AI tools beyond conventional expertise, fostering learners’ problem-solving abilities and ethical reasoning while establishing a new educational paradigm that integrates human creativity with AI’s computational capabilities [24-26].

This study highlights the potential of LLMs to improve medical education and support clinical decisions in reproductive medicine, while reinforcing the need for human expertise and ethical oversight as technology advances.

Limitations

This study has several limitations that should be considered when interpreting the results. First, the rapidly evolving nature of LLM technology means that the models evaluated (as of August 2024) may already have been updated with improved capabilities. The complete failure of Claude in data processing tasks, for instance, may reflect version-specific limitations rather than fundamental constraints of the platform. Second, we intentionally used baseline prompts without optimization to assess out-of-the-box performance. While this approach provides a realistic assessment of accessibility for nontechnical users, it likely underestimates the models’ full potential. Prompt engineering techniques, such as few-shot learning or chain-of-thought prompting, could substantially improve accuracy and consistency, particularly for complex data analysis tasks. Institution-specific prompt templates tailored to each LLM’s characteristics could further enhance performance and reduce the quality assurance burden identified in our results. Third, the knowledge assessment component faced constraints in question availability. The proprietary nature of professional certification examinations in embryology limited our test set to 10 questions curated by our institutional experts. This small

XSL•FO
RenderX

sample size and potential selection bias may not fully represent the breadth of knowledge required in clinical embryology. Fourth, the image-based diagnostic assessment was limited by technical factors. The karyotype images used (600×450 pixels) may not have provided sufficient resolution for accurate chromosome band identification. The absence of standardized regions of interest or zoom capabilities may have contributed to the variable performance across models. These technical constraints highlight the importance of image quality standardization when implementing AI tools for medical image analysis. Finally, while our study demonstrated significant time savings in data processing, we did not assess the long-term reliability or the risk of automation bias. The potential for hallucinations and plausible but incorrect outputs remains a critical concern that requires continuous monitoring and validation protocols. Health care providers must maintain vigilance and establish robust verification procedures before integrating these tools into clinical workflows.

Conclusion

This study demonstrates both the potential and limitations of current LLMs in reproductive medicine applications. While these tools showed promise for accelerating routine data analysis and supporting medical education, their failure in image interpretation tasks highlights critical limitations for clinical use. For smaller fertility clinics, LLMs offer promising support for data processing and staff training, potentially enhancing operational efficiency without the need for specialized technical infrastructure. However, to use these technologies safely and effectively in real clinical settings, it remains essential to establish well-defined procedures to verify their reliability, define clear rules for appropriate use, and ensure constant human involvement in making medical decisions. Future research should focus on training AI models specifically for reproductive medicine and systematically evaluating their impact on real-world clinical outcomes.

Acknowledgments

We thank our clinic members for their support.

Data Availability

The qualitative data recorded, transcribed, and analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

NO was responsible for conceptualization, methodology, data collection, analysis, writing the original draft, writing review and editing, and investigation. MI and YF handled data collection and analysis. HH contributed to writing review and editing, as well as managing resources. YK managed investigation and resources. TT, KY, and TH were responsible for methodology and resources. KK handled project administration and funding acquisition.

Conflicts of Interest

None declared.

References

1. Miyazaki K, Sato R. Analyses of the technological accumulation over the 2nd and the 3rd AI boom and the issues related to AI adoption by firms. Presented at: 2018 Portland International Conference on Management of Engineering and Technology (PICMET); Aug 19-23, 2018; Honolulu, HI p. 1-7. [doi: 10.23919/PICMET.2018.8481822]
2. Tang D. What is digital transformation? EDPACS 2021 Jun 3;64(1):9-13. [doi: 10.1080/07366981.2020.1847813]
3. Abd-rabo AM, Hashaikeh SA. The digital transformation revolution. Int J Humanit Educ Res 2021;3(4):124-128. [doi: 10.47832/2757-5403.4-3.11]
4. Althubaiti A, Tirksstani JM, Alsehaibany AA, Aljedani RS, Mutairii AM, Alghamdi NA. Digital transformation in medical education: factors that influence readiness. Health Informatics J 2022;28(1):14604582221075554. [doi: 10.1177/14604582221075554] [Medline: 35196920]
5. Macdonald C, Adeloye D, Sheikh A, Rudan I. Can ChatGPT draft a research article? An example of population-level vaccine effectiveness analysis. J Glob Health 2023 Feb 17;13:01003. [doi: 10.7189/jogh.13.01003] [Medline: 36798998]
6. Beilby K, Hammarberg K. O-089 Using ChatGPT to answer patient questions about fertility: the quality of information generated by a deep learning language model. Hum Reprod 2023 Jun 22;38(Supplement_1). [doi: 10.1093/humrep/dead093.103]
7. Eysenbach GJ. The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers. JMIR Med Educ 2023 Mar 6;9:e46885. [doi: 10.2196/46885] [Medline: 36863937]
8. Enatsu N, Miyatsuka I, An LM, et al. A novel system based on artificial intelligence for predicting blastocyst viability and visualizing the explanation. Reprod Med Biol 2022 Feb 7;21(1):e12443. [doi: 10.1002/rmb2.12443] [Medline: 35386375]

XSL•FO
RenderX

9. Ueno S, Berntsen J, Okimura T, Kato K. Improved pregnancy prediction performance in an updated deep-learning embryo selection model: a retrospective independent validation study. Reprod Biomed Online 2024 Jan;48(1):103308. [doi: 10.1016/j.rbmo.2023.103308] [Medline: 37914559]
10. Kobayashi H. Potential for artificial intelligence in medicine and its application to male infertility. Reprod Med Biol 2024 Jun;23(1):e12590. [doi: 10.1002/rmb2.12590] [Medline: 38948339]
11. Morimoto T, Maekawa T, Mizuta S, et al. Identifying optimal puncture position by a real-time image analysis for Piezo-ICSI: a prospective randomized sibling oocyte study. Reprod Biomed Online 2024 Apr;48(4):103735. [doi: 10.1016/j.rbmo.2023.103735] [Medline: 38402676]
12. de Santiago I, Polanski L. Data-driven medicine in the diagnosis and treatment of infertility. J Clin Med 2022 Oct 29;11(21):6426. [doi: 10.3390/jcm11216426] [Medline: 36362653]
13. ESHRE Special Interest Group of Embryology, Alpha Scientists in Reproductive Medicine. The Vienna consensus: report of an expert meeting on the development of art laboratory performance indicators. Human Reproduction Open 2017 Aug 4;2017(2). [doi: 10.1093/hropen/hox011]
14. Schulhoff S, Ilie M, Balepur N, et al. The prompt report: a systematic survey of prompt engineering techniques. arXiv. Preprint posted online on Jun 6, 2024. [doi: 10.48550/arXiv.2406.06608]
15. Renze M, Guven E. Self-reflection in large language model agents: effects on problem-solving performance. Presented at: 2024 2nd International Conference on Foundation and Large Language Models (FLLM); Nov 26-29, 2024; Dubai, United Arab Emirates p. 516-525. [doi: 10.1109/FLLM63129.2024.10852426]
16. Brin D, Sorin V, Konen E, et al. How GPT models perform on the United States medical licensing examination: a systematic review. Discov Appl Sci 2024;6(10):500. [doi: 10.1007/s42452-024-06194-5]
17. Lai UH, Wu KS, Hsu TY, Kan JKC. Evaluating the performance of ChatGPT-4 on the United Kingdom Medical Licensing Assessment. Front Med (Lausanne) 2023;10:1240915. [doi: 10.3389/fmed.2023.1240915] [Medline: 37795422]
18. Kasai J, Kasai Y, Sakaguchi S, Yamada Y, Radev D. Evaluating GPT-4 and ChatGPT on Japanese Medical Licensing examinations. arXiv. Preprint posted online on Mar 31, 2023. [doi: 10.48550/arXiv.2303.18027]
19. Clusmann J, Kolbinger FR, Muti HS, et al. The future landscape of large language models in medicine. Commun Med (Lond) 2023 Oct 10;3(1):141. [doi: 10.1038/s43856-023-00370-1] [Medline: 37816837]
20. Allahqoli L, Laganà AS, Mazidimoradi A, et al. Diagnosis of cervical cancer and pre-cancerous lesions by artificial intelligence: a systematic review. Diagnostics (Basel) 2022 Nov 13;12(11):2771. [doi: 10.3390/diagnostics12112771] [Medline: 36428831]
21. Guidozzi N, Menon N, Chidambaram S, Markar SR. The role of artificial intelligence in the endoscopic diagnosis of esophageal cancer: a systematic review and meta-analysis. Dis Esophagus 2023 Nov 30;36(12):doad048. [doi: 10.1093/dote/doad048] [Medline: 37480192]
22. Qi C, Zheng L, Wei Y. EduDCM: a novel framework for automatic educational dialogue classification dataset construction via distant supervision and large language models. Appl Sci (Basel) 2025;15(1):154. [doi: 10.3390/app15010154]
23. Hang CN, Wei Tan C, Yu PD. MCQGen: a large language model-driven MCQ generator for personalized learning. IEEE Access 2024;12:102261-102273. [doi: 10.1109/ACCESS.2024.3420709]
24. Naqvi WM, Mishra G. Hello AI: is it time for a revolutionary change in the health professional education system? Eur J Ther 2024;30(3):e28-e29. [doi: 10.58600/eurjther2201]
25. Zhang J, Fenton SH. Preparing healthcare education for an AI-augmented future. Npj Health Syst 2024;1(1):4. [doi: 10.1038/s44401-024-00006-z] [Medline: 39723201]
26. Jeyakumar T, Balakumar S, Younus S, et al. Learning to teach AI: understanding the needs of healthcare professionals. Stud Health Technol Inform 2025 Feb 18;322:52-56. [doi: 10.3233/SHTI250013] [Medline: 39968547]

Abbreviations

AI: artificial intelligence
ART: assisted reproductive technology
LLM: large language model

Edited by A Mavragani; submitted 15.12.24; peer-reviewed by CN Hang, P Codyre, S Pal; revised version received 29.07.25; accepted 13.08.25; published 01.10.25.

Please cite as:
Okuyama N, Ishii M, Fukuoka Y, Hattori H, Kasahara Y, Toshihiro T, Yoshinaga K, Hashimoto T, Kyono K
Application of Large Language Models in Data Analysis and Medical Education for Assisted Reproductive Technology: Comparative Study
JMIR Form Res 2025;9:e70107
URL: https://formative.jmir.org/2025/1/e70107
doi:10.2196/70107

XSL•FO
RenderX

© Noriyuki Okuyama, Mika Ishii, Yuriko Fukuoka, Hiromitsu Hattori, Yuta Kasahara, Tai Toshihiro, Koki Yoshinaga, Tomoko Hashimoto, Koichi Kyono. Originally published in JMIR Formative Research (https://formative.jmir.org), 1.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Response to the Netflix Docuseries "Big Vape: The Rise and Fall of JUUL": Mixed Methods Analysis of YouTube Comments Using Qualitative Coding and Topic Modeling

Beth Hoffman[1], MPH, PhD; Arpita Tripathi[1], MA; Ariel Shensa[2], MA, PhD; Julia (Pengyue) Dou[1], MPH; Piper Narendorf[1], BS; Nishi Hundi[1], BA; Jaime Sidani[1], CHES, MPH, PhD

[1]Department of Behavioral and Community Health Sciences, School of Public Health, University of Pittsburgh, 130 De Soto Street, Pittsburgh, PA, United States

[2]Department of Health Administration and Public Health, John G Rangos Sr. School of Health Sciences, Duquesne University, Pittsburgh, PA, United States

Corresponding Author:
Beth Hoffman, MPH, PhD
Department of Behavioral and Community Health Sciences, School of Public Health, University of Pittsburgh, 130 De Soto Street, Pittsburgh, PA, United States

Abstract

Background: On October 11, 2023, Netflix released the docuseries "Big Vape: The Rise and Fall of JUUL," which chronicled the founding of JUUL, its rise in popularity among youth, and the subsequent public backlash. The official Netflix YouTube channel posted a trailer promoting the docuseries and an official clip from the docuseries. Recent studies have demonstrated the utility of using comments posted under YouTube videos to analyze reactions to the content and discourse around the health topics explored in the video.

Objective: This study aimed to (1) systematically characterize nicotine and tobacco product (NTP)–related comments and replies posted in response to the docuseries trailer and video clip and (2) explore integration of automated topic modeling techniques with traditional human-generated qualitative coding.

Methods: We extracted all comments and replies on the aforementioned YouTube clips 1 month after the docuseries' release (N=532). Research assistants manually double-coded the comments using a systematically developed codebook that assessed for NTP sentiment (pro-NTP, anti-NTP, complex sentiment, or no sentiment) and the presence or absence of specific electronic cigarette (e-cigarette)–related content. Given the substantial amount of comments coded as potential misinformation during the coding process, we conducted an in-depth qualitative content analysis of all comments coded as potential misinformation. Simultaneously, we used word clustering techniques including structural topic modeling to identify the overarching topics.

Results: Of the 73.8% (393/532) relevant comments, 63.6% (250/393) expressed NTP sentiment with 42.8% of these (107/250) expressing pro-NTP sentiment and 18.4% (46/250) expressing complex sentiment. The most frequent content category was potential misinformation (27.5%, 108/393). These 108 comments contained 152 individual pieces of misinformation that were broadly grouped within 6 themes with various numbers of subthemes; the most frequent misinformation theme was that e-cigarette use is completely safe or much safer than smoking (n=80). Other frequently occurring content categories included e-cigarette use is safer than smoking (17.6%, 69/393), and personal experience using e-cigarettes or JUUL (15.5%, 61/393). For topic modeling, we identified 9 topics that we qualitatively assigned into 4 thematic categories: comparisons with other drugs, mentions of government and pharma companies, role of media and parents, and harms associated with nicotine and tobacco products.

Conclusions: To the best of our knowledge, this is the first study to examine viewer reactions to the docuseries about JUUL. Our analysis of YouTube comments offers insight into current sentiment and misinformation regarding NTPs and highlights the potential utility of using mixed methods to analyze NTP-related social media data, and the benefits of integrating computational and human qualitative research to analyze social media perceptions of e-cigarettes. Public health professionals can use our findings to help develop tailored health communication messages to address common sentiment and misconceptions related to JUUL, other e-cigarette products, and new NTP products.

(JMIR Form Res 2025;9:e76737) doi:10.2196/76737

KEYWORDS

nicotine and tobacco products; social media data; e-cigarettes; JUUL; misinformation

XSL•FO
RenderX

Introduction

In 2018, the U.S. Surgeon General declared an epidemic of youth electronic cigarette (e-cigarette) use, citing the 78% increase in current e-cigarette use among high school students and 48% increase among middle school students from 2017 [1]. The Surgeon General's advisory noted that this increase was largely driven by use of the brand JUUL, which by the end of 2017 held the greatest market share in the United States of all e-cigarette products [1].

JUUL, an electronic nicotine delivery system (ENDS), delivers a substantial amount of nicotine, with each "vape pod" containing an amount equivalent to a pack of cigarettes [2]. It quickly became synonymous with e-cigarette use due to its sleek design, high nicotine content, and appealing flavors [3], and studies highlighted public health issues such as nicotine addiction and respiratory problems associated with JUUL use [3,4]. In response to the condemnation from regulators, parents, and health advocates, JUUL voluntarily withdrew its flavored products from the market in 2019 [5]. This move aimed to address concerns about the appeal of flavored e-cigarettes to youth and to further reduce their access to these products.

In July 2019, a widespread outbreak of e-cigarette or vaping product use–associated lung injury (ie, EVALI) garnered significant attention among medical professionals, the news media, and the public. The US Centers for Disease Control and Prevention (CDC) coordinated a multistate outbreak investigation, with over 2800 cases requiring hospitalization reported to the organization by February 18, 2020 [6]. Ultimately, most cases were linked to informally sourced (ie, "black market") products that contained tetrahydrocannabinol (THC), although about 20% (227/956) of affected individuals who reported using any nicotine-containing product reported using nicotine-containing e-cigarettes only [6].

In December 2019, Congress passed the Tobacco 21 (T-21) law, which aimed to restrict the sale of tobacco and e-cigarette products, including JUUL, to individuals aged 21 and older [7]. On January 2, 2020, the Food and Drug Administration (FDA) issued a ban on the sale of flavored JUUL pods, except for tobacco and menthol flavors [8]. Further, in 2022, the FDA imposed stricter regulations by mandating JUUL to submit extensive premarket tobacco product applications (PMTAs) to remain on the market; weeks later, the FDA denied these PMTAs, citing insufficient data on health risks, which led to a temporary ban on JUUL products [9]. Most recently, on June 6, 2024, the FDA announced that it has lifted its ban as the agency reviews new court rulings and evaluates updated information submitted by the vape manufacturer [10]. Per the 2024 National Youth Tobacco Survey, JUUL is the fifth most popular e-cigarette brand used by youth, behind disposable e-cigarettes Elf Bar, Breeze, Mr. Fog, and Vuse; these have become more popular among youth due to their variety of flavors and lower cost [11].

On October 11, 2023, the streaming service Netflix released a docuseries titled "Big Vape: The Rise and Fall of JUUL." This 4-episode docuseries examined the founding of JUUL, the product launch, aggressive marketing tactics, regulatory challenges, and the subsequent public backlash [12]. Previous research has examined viewer reactions to EVALI storylines on popular medical dramas [13,14], as well as the influence of health-focused documentaries on viewer knowledge and attitudes [15]. However, to the best of our knowledge, viewer reactions to this docuseries have yet to be examined.

Social media data may provide an innovative avenue to explore this. Prior work using the social media platform X (previously named Twitter) has demonstrated that these data provide insight on viewer responses to Netflix documentaries, sentiment around e-cigarettes, personal or proximate use of e-cigarettes, and e-cigarette–related misinformation [16-22]. However, in 2023, X ceased free access to its application programming interface (API), limiting the utility of using messages on the platform for research purposes [23].

On September 21, 2023, the official Netflix YouTube channel posted a trailer promoting the docuseries, followed by posting an official clip from the docuseries on October 11, 2023. Recent studies have demonstrated the utility of using comments posted under YouTube videos to analyze reactions to the content and discourse around the health topics explored in the video [24,25], including analyzing comments posted in response to e-cigarette modification YouTube videos [26]. Similarly, analyzing comments on vaccination videos has provided insights into public perceptions, the spread of misinformation, and the effectiveness of health communication strategies [27]. These studies demonstrate that YouTube comments can directly reveal nuanced viewer reactions, engagement levels, and the spread of information or misinformation.

Automated techniques are becoming increasingly popular for analyzing social media data due to their ability to efficiently process large volumes of unstructured text and uncover underlying patterns and themes [28]. Topic modeling is one such automated method that uses machine learning to identify latent themes within textual data by analyzing word frequency and co-occurrence patterns. Unlike manual thematic coding, which requires predefined categories and can be time-consuming, topic modeling offers an inductive approach that uncovers topics directly from the data without prior assumptions [29]. This makes it especially useful for exploratory analysis, where the goal is to discover the underlying structure of conversations. When combined with human coding, topic modeling can guide researchers to areas of interest for further qualitative investigation, ensuring a comprehensive and nuanced understanding of social media discourse [30].

Therefore, we aimed to systematically characterize nicotine and tobacco product (NTP)–related comments and replies posted in response to the docuseries trailer and video clip on the Netflix YouTube channel. A secondary aim was to explore the integration of automated topic modeling techniques in comparison to traditional human-generated qualitative coding to determine whether computational methods can complement the process of human coding.

XSL•FO
RenderX

Methods

Data Collection

We extracted all comments and replies on the docuseries trailer and docuseries clip on YouTube 1 month after the docuseries release (November 11, 2023). This time frame was chosen given our focus on initial reactions to the docuseries, as well as a previous study that used a 1-month time frame for assessing viewer reactions to the medical drama Code Black and data suggesting that most people watch a new Netflix show soon after it becomes available [31,32]. We used a Python (Guido van Rossum) script with an API key to automate the downloading of all corresponding comments for the 2 videos. This approach enabled us to efficiently retrieve and archive the entire content and discussion data for further analysis. Comments and replies were organized into 2 spreadsheets: one for comments and replies from the trailer (n=437) and one for comments and replies from the clip (n=95). This resulted in 239 original comments with 293 associated replies, for 532 total comments and replies.

Human Qualitative Coding

For human qualitative coding, we adapted methods previously used for codebook development and analysis of JUUL and e-cigarette–related Twitter data [21,22].

Codebook Development and Coding Procedures

The lead researcher for this study, first author BLH, oversaw codebook development and coding procedures. The first author and 2 research assistants (RAs) each drafted an initial list of potential codes using a hybrid process consisting of codes from prior qualitative analyses of e-cigarette–related social media data [19,21,22] and an initial analysis of 50 randomly selected comments or replies from the dataset. The first author and the 2 RAs then met to review their lists of potential codes and used these to draft an initial codebook with codes, subcodes, definitions, and examples (Multimedia Appendix 1).

Each comment or reply was initially coded for relevance to NTPs, which was defined as "A comment that mentions JUUL or NTPs, including cigarettes. It does not need to be explicit." Comments or replies that were coded as not relevant and excluded from further analysis included ones that mentioned the docuseries but not NTPs (eg, "It's rare to see such a well-structured criticism of Australia's public policy. Kudos to Netflix.") and comments or replies that mentioned substances other than NTPs (eg, "When are we going to get a documentary that shows why people should not use marijuana or drink alcohol?"). For the replies, RAs were instructed to consider the relevance of the parent comment when determining relevance. For example, the reply "saved countless lives" was coded as relevant because it was in direct response to the relevant parent comment of "Introducing vapes was a really bad idea. Now all kids in school have it."

All relevant comments or replies (henceforth simply referred to as "comments") were then coded for NTP sentiment. Comments could be coded as pro-NTP, anti-NTP, complex sentiment, or no sentiment. Comments were coded as complex sentiment if they expressed pro-e-cigarette sentiment but anti-JUUL sentiment or vice versa (eg, "I've never used and never trust Juul or any pre-filled vape pens. I am still good with my mod+ tank setup from the 2010s and get my PG/VG in bulk online. Honestly, I don't even like pod systems.") or if they expressed the complexity of evidence around the safety or harms of e-cigarette use (eg, "Vaping has been (and still is) an effective way to quit smoking, but some big companies saw the money potentials and exploited it by targeting kids with pre-filled vape pens."). Comments were coded as no sentiment if they contained a statement with no clear sentiment or if the RAs could not determine what type of sentiment was expressed (eg, "have you seen a Juul? They look almost exactly like USB drives.").

Next, comments were coded for the presence or absence of specific e-cigarette–related content. Content codes that emerged from direct analysis of the data included distrust of medicine, research, or government; comparison to opioids or alcohol; conspiracy; Big Tobacco; blaming parents; and more research is needed. All other content codes were adapted from prior research examining e-cigarette–related social media data [19,21,22,33]. For a complete list of codes, definitions, and examples, please see Multimedia Appendix 1.

Comments that were not written in English were entered into Google Translate for translation and coding in English. Content codes were not mutually exclusive. Following codebook development, 4 RAs were paired up (2 teams of 2) for coding. The first author oversaw all coding assignments and procedures. All comments or replies were double-coded. After coding 100 comments or replies, each team met with the first author to adjudicate differences and refine the codebook as needed. Prior to adjudication meetings, percent agreement ranged from 71%to 100%and Cohen kappa from 0.09 to 1.00; all differences were resolved during adjudication (ie, 100% agreement was reached for all categories). This was repeated until all coding was complete.

Analyses

We calculated descriptive statistics for each sentiment and content code and examined the distribution of content code frequency by sentiment. Given the substantial amount of comments coded as potential misinformation during the coding process, following the completion of the initial human qualitative coding, we decided to conduct an in-depth qualitative content analysis of these comments [34]. To conduct this, the first and last authors independently analyzed all comments coded as potential misinformation. First, they identified all specific comments that made statements about ENDS, JUUL, health claims, and claims related to policy. All statements were compared with the peer-reviewed literature (eg, statements about harms and benefits) or policy documents and news sources (eg, statements about policy) as of March 2024. It is important to note that, in many cases, there has not yet been a definitive consensus regarding health claims in the peer-reviewed literature, meaning that articles exist that both "prove" and "disprove" a particular statement. Because a scientific consensus has not yet been reached, these claims were categorized as misinformation. In some cases, this includes exaggeration of research findings. Then, the first and last authors tabulated the total number of phrases identified as misinformation, organized

XSL•FO
RenderX

these phrases into overarching misinformation categories, and identified common words within each category.

Topic Modeling

Simultaneously, we used R (R Core Team) to perform topic modeling with all comments and replies. Topic modeling is a text-analysis method that identifies groups of co-occurring terms or latent topics in large bodies of text, often referred to as "corpus." In this study, the text corpus was the set of comments and replies described earlier. Topic modeling in this study was done in 2 main steps: we first used R packages, including tidytext, dplyr, stringr, and stopwords for data cleaning and text preprocessing. This involved converting the text to lowercase, removing punctuation, and filtering out stop-words. Posttext cleaning, we generated a frequency analysis of the most frequently used terms and created a Document-Term Matrix using the tm package. Later, topic modeling was performed using latent Dirichlet allocation (LDA), which is the most commonly used algorithm for topic modeling [35], and has been used in previous studies using social media data for health-based topics [36-38]. We explored the LDA-based model with various topic numbers, including k=10, 9, 8, and 5, based on recommendations by previous studies [35]. The final analysis focused on models with 10 topics. After reviewing the content of the k=5, 8, 9, and 10 solutions, the first, second, and last authors (BLH, AT, and JES) determined that the 10-topic model provided the most coherent and qualitatively meaningful representation of the data. To assess robustness, we repeated models with multiple random seeds and bootstrap resamples, confirming stable topic assignments. For benchmarking, we additionally implemented non-negative matrix factorization (NMF) and BERTopic (embedding-based clustering) and compared results on coherence, exclusivity, and stability. Topics were pruned if they were predominantly non-English or qualitatively uninterpretable (eg, use of emoticons). One such topic was removed, and the remaining 9 topics were grouped inductively into 4 thematic categories for reporting.

Ethical Considerations

This study was approved by the University of Pittsburgh Human Research Protection Office (approval number: STUDY22080079). As this study only involved the examination of publicly available social media data, informed consent and participant compensation were not applicable. Regarding privacy, although all comments included in our sample are publicly available, we recognize the sensitive nature of some comments, particularly those where a user self-disclosed using NTPs. Therefore, consistent with previous analyses of publicly available NTP content on social media [20-22], we modified all direct quotes in the tables and text and conducted reverse searches (ie, inputting the modified text into Google) to ensure that the modifications are sufficient to prevent individual users from being identified. In the absence of specific reporting guidelines related to social media research, we followed the SRQR (Standards for Reporting Qualitative Research) [39].

Results

Human Coding

Sentiment

For the parent comments (ie, excluding replies, n=239), the "like" count ranged from 0 to 252. The mean "like" count was 5.6 (SD 19.9) and the median was 1 (IQR 0 - 3). Of the 532 comments and replies, 73.8% (393/532) were relevant to NTPs. Approximately 64% of comments (250/393) expressed any NTP sentiment, with 42.8% of these (107/250) expressing pro-NTP sentiment, 38.8% (97/250) expressing anti-NTP sentiment, and 18.4% (46/250) expressing complex sentiment (Table 1).

XSL•FO
RenderX

Table . Frequencies of sentiment and content codes from human qualitative coding of YouTube comments including distribution of content code frequency by sentiment (n=393).

Code	Frequency, n (%)	Pro-NTP sentiment, n (%)	Anti-NTP sentiment, n (%)	Complex-NTP sentiment, n (%)
NTP[a] sentiment	250 (63.6)	—[b]	—	—
Pro-NTP	107 (42.8[c])	—	—	—
Anti-NTP	97 (38.8[c])	—	—	—
Complex sentiment	46 (18.4[c])	—	—	—
Content[d]				
Potential misinformation	108 (27.5)	66 (75.9)	12 (13.8)	9 (10.3)
e-cigarette[e] use is safer than smoking	69 (17.6)	53 (80.3)	0 (0.0)	13 (19.7)
Personal experience	61 (15.5)	27 (56.3)	5 (10.4)	16 (33.3)
Health effects of e-cigarettes	50 (12.7)	7 (15.2)	26 (56.5)	13 (28.3)
Big tobacco	41 (10.4)	13 (65.0)	2 (10.0)	5 (25.0)
Addiction	39 (9.9)	6 (19.4)	16 (51.6)	9 (29.0)
Conspiracy	37 (9.4)	16 (69.6)	3 (13.0)	4 (17.4)
Marketing to youth	24 (8.7)	4 (15.4)	14 (53.9)	8 (30.8)
Distrust of medicine, research, or government	33 (8.4)	19 (90.5)	0 (0.0)	2 (9.5)
Flavors	27 (6.9)	12 (52.2)	6 (26.1)	5 (21.7)
E-cigarettes are as harmful as smoking	23 (5.9)	1 (4.8)	18 (85.7)	2 (9.5)
Blaming parents	22 (5.6)	1 (14.3)	3 (42.9)	3 (42.9)
Policy action or regulation	19 (4.8)	7 (53.9)	5 (38.5)	1 (7.7)
Personal smoking to e-cigarette use	19 (4.8)	12 (66.7)	0 (0.0)	6 (33.3)
EVALI[f]	13 (3.3)	5 (55.6)	1 (11.1)	3 (33.3)
Comparison to opioids or alcohol	11 (2.8)	2 (33.3)	1 (16.7)	3 (50.0)
More research is needed	9 (2.3)	3 (37.5)	1 (12.5)	4 (50.0)
Longing for JUUL	5 (1.3)	5 (100.0)	0 (0.0)	0 (0.0)
Learning from docuseries	3 (0.8)	0 (0.0)	1 (33.3)	2 (66.7)

[a]NTP: nicotine and tobacco product.

[b]Not applicable.

[c]Percent calculated out of total number expressing any sentiment (n=250).

[d]Codes are not mutually exclusive.

[e]e-cigarette: electronic cigarette

[f]EVALI: e-cigarette or vaping product use–associated lung injury.

Content

The most frequent content category was potential misinformation (27.5%, 108/393), followed by e-cigarette use is safer than smoking (17.6%, 69/393; eg, “Harm reduction matters more than people think. Cigarette smoking causes 8 million preventable deaths globally each year. A safer way to consume nicotine could save hundreds of millions of these deaths in the long run.”) and personal experience using e-cigarettes or JUUL (15.5%, 61/393; eg, “I was hooked on cigs for 35 years. I got an e-cig as a gift. Since then, I’ve been smoke-free for 13 years, nicotine-free for 10 years, and vape-free for 8 years. How many long-running smokers are never going to switch like I did because of what these jerks did to the vaping industry?”). Other frequently occurring content categories included health effects of e-cigarettes (12.7%, 50/393) and Big Tobacco (10.4%, 41/393; Table 1).

XSL•FO
RenderX

Examining the distribution of content code frequency by sentiment, approximately 76% (66/108) of comments coded as potential misinformation expressed pro-NTP sentiment. Similarly, the majority of comments coded as e-cigarette use is safer than smoking and personal experience expressed pro-NTP sentiment, although 33.3% (16/61) of comments coded as the latter expressed complex sentiment (Table 1). Almost all of these comments mentioned the poster using e-cigarettes but not using JUUL, such as "as someone who vaped for over a decade without any major health problems, I'm really glad I stayed away from Juul."

Of note, while the majority of comments coded as e-cigarette use is safer than smoking were also coded as potential misinformation, approximately 27% (19/69) were not. These comments often used a personal anecdote (eg, "I quit smoking three years ago and my wife and I started using a Juul. Since then, we have taken up running, biking, kayaking—you name it—all thanks to quitting cigarettes."), therefore were not considered potential misinformation. Similarly, about half of the comments coded as e-cigarettes are as harmful as smoking were not coded as potential misinformation (52.2%, 12/23). These comments acknowledged that more time is needed to determine the full health effects of e-cigarettes (eg, "As a registered nurse, I assisted in a lung cancer surgery. I can imagine that in 20 or more years, we start seeing the same health problems linked to vaping.") or acknowledged the differences between the 2 products in terms of harms (eg, "Vapes take the lives of young people, while cigarettes take the lives of adults. Think about that.").

Qualitative Analysis of Potential Misinformation Comments

Among the 108 comments coded as potential misinformation, qualitative analysis revealed 152 individual pieces of misinformation that were broadly grouped within 6 themes (Multimedia Appendix 2). The most frequent theme was that E-cigarette use is completely safe or much safer than smoking, found in 80 statements (Multimedia Appendix 2). Three distinct subthemes emerged; the first focused on declarative statements that e-cigarette use is much safer than smoking, which often included declarative language such as "definitive" (eg, "vapes are definitely 1000x a better alternative to classic Marlboros.") or "far healthier" (eg, "It seems safer to say that they are much healthier than tobacco products."). Statements often stated that e-cigarettes are 95% safer than smoking, an estimate that is not backed by concrete evidence [40]. Another subtheme was that e-cigarette use has saved millions of lives (n=20), considered potential misinformation due to the lack of evidence around the population-level harms versus benefits of e-cigarette use, as well as conflicting results from modeling studies [41,42]. A final subtheme was that E-cigarette use has no known harms or is never harmful, which often suggested that there are no harms associated with e-cigarette use with statements such as "no solid evidence suggesting vaping is significantly harmful in any way."

The second misinformation theme, which appeared in 21 statements, was broadly centered on conspiracies, policies, or other types of misinformation. A recurring subtheme was conspiracy theories about the "take down" of e-cigarettes by various entities, including governmental agencies, professional health organizations, and Big Tobacco because the sale of e-cigarettes affected their profits or grant funding. Other subthemes included misinformation about the FDA's actions related to JUUL and varied pieces of misinformation that occurred with less frequency and were not consistent with other themes or subthemes (Multimedia Appendix 2).

The third misinformation theme was that EVALI was caused only by THC (n=19). Statements in this category often incorrectly stated that all or 100% of EVALI cases were caused by THC or black-market products (eg, "All cases were linked to off-market products.") or minimized the number of cases (eg, "Only a small number of individuals were affected by lung damage"). The fourth misinformation theme focused on health effects, benefits, and nicotine-related misinformation, which occurred in 19 statements and consisted of 3 subthemes. One subtheme consisted of exaggerations about the negative health effects of e-cigarettes or nicotine, which were often claims that have not yet been conclusively established by research. Other subthemes focused on the suggestion that nicotine is benign or is the same as caffeine and that nicotine or e-cigarettes have unproven or exaggerated health benefits, which were often statements that have not yet been conclusively established by research. The fifth misinformation theme was equating e-cigarettes with cigarettes, which was found in 10 statements and included 2 subcategories. The first, that e-cigarette use is the same as smoking, often implied that e-cigarettes and cigarettes are exactly the same, while the second, that the only harmful part of smoking cigarettes is the combustion, often left out the fact that nicotine is addictive and can be harmful to the developing adolescent brain. Finally, the sixth misinformation theme was e-cigarettes presented as a proven smoking cessation tool, which was found in 3 statements (Multimedia Appendix 2).

Topic Modeling

Frequency Analysis

After removing common stop-words, the most frequent words in the cleaned dataset were "vaping" with 158 occurrences, "people" with 117 occurrences, "juul" with 114 occurrences, "kids" with 108 occurrences, and "smoking" with 91 occurrences. Other relevant frequently occurring terms were "tobacco," "cigarettes," "nicotine," "bad," and "health." Refer to Multimedia Appendix 3 for the complete list of the 200 frequently occurring words.

Topics Covered in YouTube Comment Discussions

We initially identified 10 topics based on the most frequent terms associated with each topic. One topic (Topic 4) comprised predominantly non-English comments and was removed as irrelevant to the focus of our analysis, leaving 9 coherent topics. To assess robustness, we compared 3 modeling approaches: LDA, NMF, and BERTopic. NMF achieved the highest semantic coherence (0.48), followed by LDA (0.41) and BERTopic (0.36). However, NMF topics exhibited greater redundancy, and BERTopic produced several clusters dominated by function words. The 10-topic LDA model provided the best balance of interpretability, coherence, and exclusivity, and its themes

XSL•FO
RenderX

aligned most closely with our qualitative coding framework. The 9 retained topics were qualitatively grouped into 4 thematic categories: comparisons with other drugs, mentions of government and pharmaceutical companies, the role of media and parents, and harms associated with nicotine and tobacco products (Table 2). The aggregated LDA weights of the most relevant terms across the 4 thematic groups are presented in Multimedia Appendix 4.

Table . Thematic categories of 9 topics and relevant terms of each topic as identified in the analysis of 532 YouTube comments.

Category	Relevant terms (Most Relevant Terms are in *italics*)
Comparisons with other nicotine, tobacco, alcohol, and other products	
Topic 5	Juul, *smoking, vaping, cigarettes*, people, documentary, really, long, bad, think, term, health
Topic 8	vaping, years, *nicotine*, believe, smoking, *juul, vape*, much, *cigarettes*, just, now, smoke, got, free
Topic 9	Kids,just, brand, vaping, *thc, vitamin*, people, market, *carts, vape, juices*, fact, used, juul
Topic 10	want, get, good, can, like, people, *alcohol*, problems, know, just, make, better.
Mentions of Government and Pharma Companies	
Topic 2	people, juul, *tobacco*, can, *vape, big, health, companies*, nicotine, *control*
Topic 7	vaping, tobacco, harmful, smoking, *people*, product, health, smoke, bad, *government, lives*, used.
Role of Media and Parents	
Topic 1	*kids*, now, just, people, juul, know, vape, vaping, smoking, don't, stop, *parents, blame*
Topic 6	*kids*, juul, don, quot, people, *media*, just, vape, addicted, *social*, got, even, *reason, cool*
Harms associated with Nicotine and Tobacco Products	
Topic 5	Juul, smoking, vaping, cigarettes, people, documentary, *really, long, bad, think, term, health*
Topic 3	vaping, years, juul, smoking, nicotine, tobacco, people, every, cigarettes, smoke, like, *life, lungs*
Topic 7	vaping, tobacco, *harmful*, smoking, people, product, *health*, smoke, *bad*, government, *lives, used*.

Discussion

Principal Findings

To the best of our knowledge, this is the first study to examine viewer reactions to the Netflix docuseries "Big Vape: The Rise and Fall of JUUL." Human qualitative coding of comments and replies posted in response to the docuseries trailer and video clip on the Netflix YouTube channel found that most comments were relevant to NTPs, and that almost one-third of comments expressed potential misinformation about these products. Topic modeling found that the most frequent words highlighted central themes related to e-cigarette use, particularly focusing on JUUL and its association with youth. Finally, topic modeling found 9 relevant topics, roughly grouped into 4 thematic areas. Overall, our findings highlight the potential utility of analyzing YouTube comments in response to an e-cigarette–related docuseries to identify public perceptions of these products and potential misinformation about them.

To date, most analyses of e-cigarette–related content on social media have classified sentiment as positive (ie, pro-e-cigarette), neutral, and negative (ie, anti–e-cigarette) [43-45]. In the course of codebook development, we noted multiple comments where the author expressed pro–e-cigarette sentiment but anti-JUUL sentiment or vice versa. Thus, we included a complex sentiment code to capture these comments; almost 20% (46/250) of comments expressing sentiment were coded as complex. This finding underscores the importance of including more nuanced sentiment codes in future analyses of e-cigarette–related social media data, particularly when analyzing data related to specific products or policies. It also underscores the potential difficulty in using automated techniques to evaluate sentiment, as well as the need for public health professionals to develop survey items that can capture complex viewpoints and tailored health communication messages that acknowledge nuances in opinions about e-cigarette products.

Given that almost half of the comments that expressed NTP sentiment were pro-NTP, it is perhaps not surprising that the majority of comments in our most frequently coded content categories (potential misinformation, e-cigarette use is safer than smoking, and personal experience) were pro-NTP. However, it is notable that about one-third of personal experience comments expressed complex-NTP sentiment, a finding in contrast to previous studies of Twitter/X that found

XSL•FO
RenderX

personal experience messages to be overwhelmingly positive [18,46]. This difference may be due in part to this previous work not classifying sentiment as complex, or it may be reflective of more nuanced positions held by e-cigarette users around support for JUUL or other specific products. Future research can build from this finding to further investigate the sentiment of e-cigarette users around NTPs, and public health professionals can use these findings to develop more nuanced health communication and health education efforts.

Previous studies examining a variety of health-related content on TikTok (ByteDance Ltd), Twitter, and YouTube have found health-related misinformation in approximately 5% - 15% of posts [47-50], and a study specifically examining e-cigarette–related content on Twitter/X found potential misinformation in 12.8% (377/2943) of posts [19]. Thus, we were surprised to find potential misinformation in over 25% (108/393) of comments in our sample. As some of the misinformation themes were specific to JUUL (eg, JUUL's only mission was to stop adult cigarette smoking), it may be that the greater frequency of potential misinformation in our sample was due to the subject matter—the JUUL docuseries—of the videos the comments were posted. However, we also observed similarities in misinformation themes, particularly that e-cigarette use is safe or much safer than smoking, misinformation about EVALI, and exaggerations of the known harms of e-cigarette products, between our findings and previous work examining misinformation about e-cigarettes on Twitter/X [19,33]. We also found frequent mention of the claim that e-cigarettes are 95% safer than smoking cigarettes. As discussed by Eissenberg et al [40], this estimate is considered a "factoid," or a statement repeated so often it is considered a fact without supporting evidence. These findings suggest the need for public health professionals to develop comprehensive health communication campaigns across social media platforms to address common misinformation themes, and for future research to examine ways in which e-cigarette misinformation may spread across social media platforms.

Analyses of viewer comments on other health-related docuseries and videos show both parallels and distinctions with our findings. For instance, comments on antivaccine videos on YouTube have similarly featured personal stories and emotionally charged discourse (ie, fear-based or toxic messaging) accompanied by claims widely recognized as misinformation [27]. However, sentiment in these discussions tended to be more distinct, whereas sentiment in our dataset was characterized by more complex views (ie, pro-e-cigarette but anti-JUUL or vice versa). In addition, while previous research on a family planning documentary has found that viewer discourse after viewing the content often centered on personal stories and expressions of learning [15], we found a greater emphasis on product-specific comparisons, brand-related narratives (eg, the "95% safer" factoid), corporate intent, and policy debates. These differences suggest that docuseries addressing commercially branded products with a recent history of regulatory action may elicit more polarized, misinformation-prone, and policy-focused engagement compared to those on other health topics.

A review of the literature on social media content analysis published in 2023 recommended that future research synthesize human qualitative coding and software-based methods [51], and our findings demonstrate the feasibility and benefits of this approach as related to social media perceptions of e-cigarettes. One benefit may be at the codebook development stage. For example, based on our initial human analysis of comments, we developed a code for blaming parents, but not one for comments that blamed the youth themselves. However, topic modeling revealed that the terms "blame" and "kids" appeared in Topic 1, suggesting that blaming youth or kids was a relevant theme. Similarly, while our human qualitative codebook had content codes for both EVALI and potential misinformation, we did not create a separate code for THC itself. However, topic modeling identified THC as a significant term in Topic 9, demonstrating the method's ability to highlight themes that might have been overlooked in manual coding. Similarly, due to comments examined during the initial stages of codebook development, we anticipated that comparison to opioids or alcohol would be key content code, yet these only ended up being mentioned in less than 3% (11/393) of comments. Topic modeling found that alcohol was still discussed in Topic 10, but substances were not as prevalent in the larger dataset as initially expected.

In the given context, we can argue that with smaller datasets like in this study, the integration of topic modeling with human coding offers a useful and efficient method for uncovering both expected and unexpected themes. This approach combines the strengths of both data-driven insights and human contextual interpretation, enhancing the quality of the analysis. In future research, these methods can be applied to larger datasets, where the volume of data would make manual coding alone impractical. As dataset sizes grow, topic modeling can serve as a more robust foundation for identifying trends, while human coding remains essential for nuanced interpretation and validation. The success of this mixed approach in smaller datasets suggests its potential for scalability and greater efficiency in handling larger, more complex datasets.

Limitations

We extracted all comments and replies 1 month after the docuseries release, so we did not assess comments or replies posted after this date. While this was due to our focus on initial reactions to the docuseries, this time frame means we potentially missed evolving discussions or delayed reactions. We also did not analyze comments and replies temporally (ie, from multiple time points) and thus cannot make conclusions about how comments or opinions may have evolved over time, nor did we examine associations between comment popularity (ie, "likes" and replies) and sentiment. It was also beyond the scope of this study to examine the accounts in our dataset; thus, we do not know if the comments were authored by humans or automated accounts (ie, "bots") nor the demographic characteristics of the authors. Future research could examine both temporal trends of comments as well as demographics of accounts posting the comments. For the human qualitative coding, interpretation of the comments can be subjective, although we aimed to minimize subjectivity by double-coding all comments plus adjudication with a researcher experienced in coding social media data and

XSL•FO
RenderX

using systematic procedures for codebook development and coding. For topic modeling, the small dataset size of this study may have been insufficient for topic identification and may have limited the generalizability of the results. It is expected that larger datasets would produce more robust patterns and results, and future research could assess if our topics hold with larger datasets. Another limitation is that the use of topic modeling was largely exploratory. We used it as a tool to generate themes without any prior assumptions, which makes it useful for discovery but also less suited for confirming hypotheses or generating precise categorizations. Thus, the results from this approach should be interpreted as preliminary and open to further refinement.

Conclusions

Our analysis of YouTube comments and replies posted in response to the trailer and official video clip for the Netflix docuseries "Big Vape: The Rise and Fall of JUUL" highlights the potential utility of analyzing these comments to identify sentiment around these products and potential misinformation about them. It also highlights the benefits of integrating computational and human qualitative research to analyze social media perceptions of e-cigarettes. Given ongoing regulatory and legal proceedings related to JUUL and other e-cigarette products, as well as the continued emergence of new NTP products, it is vital for public health to develop educational materials about the current science regarding e-cigarettes. Public health professionals can use our findings to help develop tailored health communication messages to address common sentiment and misconceptions.

Acknowledgments

The authors would like to thank Christine Larkin for her assistance with project management.

This work was supported by a grant from the National Institute on Minority Health and Health Disparities (R01MD018543, principal investigator: JS). The authors did not use generative artificial intelligence in any portion of the research ideation or manuscript writing process.

Disclaimer

The funding organization had no role in the study design; collection, analysis, and interpretation of data; writing the report; and the decision to submit the report for publication.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to the terms in our Human Research Protection Office protocol for protecting participant privacy, but are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Codebook used for human qualitative coding of YouTube comments. Coders used the definitions and examples presented to code each comment for the presence or absence of each code, and codes were not mutually exclusive. All examples are modified to prevent reidentification.
[DOCX File, 23 KB - formative_v9i1e76737_app1.docx]

Multimedia Appendix 2
Themes identified in qualitative analysis of YouTube 108 comments coded as potential misinformation that revealed 152 individual pieces of misinformation.
[DOCX File, 35 KB - formative_v9i1e76737_app2.docx]

Multimedia Appendix 3
List of the top 200 most occurring terms in the dataset.
[DOCX File, 22 KB - formative_v9i1e76737_app3.docx]

Multimedia Appendix 4
Aggregated latent Dirichlet allocation weights of the most relevant terms across the 4 thematic topic groups.
[PNG File, 113 KB - formative_v9i1e76737_app4.png]

References

XSL•FO
RenderX

1. Results from 2018 National Youth Tobacco Survey show dramatic increase in e-cigarette use among youth over past year. FDA. 2020. URL: https://www.fda.gov/news-events/press-announcements/results-2018-national-youth-tobacco-survey-show-dramatic-increase-e-cigarette-use-among-youth-over [accessed 2024-08-06]
2. 6 important facts about JUUL. Truth Initiative. 2018. URL: https://truthinitiative.org/research-resources/emerging-tobacco-products/6-important-facts-about-juul [accessed 2024-07-06]
3. Fadus MC, Smith TT, Squeglia LM. The rise of e-cigarettes, pod mod devices, and JUUL among youth: factors influencing use, health implications, and downstream effects. Drug Alcohol Depend 2019 Aug 1;201:85-93. [doi: 10.1016/j.drugalcdep.2019.04.011] [Medline: 31200279]
4. Cao DJ, Aldy K, Hsu S, et al. Review of health consequences of electronic cigarettes and the outbreak of electronic cigarette, or vaping, product use-associated lung injury. J Med Toxicol 2020 Jul;16(3):295-310. [doi: 10.1007/s13181-020-00772-w] [Medline: 32301069]
5. Edwards E. Juul has stopped selling all fruity flavors. NBC News. 2019. URL: https://www.nbcnews.com/health/vaping/juul-has-stopped-selling-all-fruity-flavors-n1068211 [accessed 2024-07-07]
6. Rebuli ME, Rose JJ, Noël A, et al. The e-cigarette or vaping product use-associated lung injury epidemic: pathogenesis, management, and future directions: an official American Thoracic Society Workshop Report. Ann Am Thorac Soc 2023 Jan;20(1):1-17. [doi: 10.1513/AnnalsATS.202209-796ST] [Medline: 36584985]
7. Tobacco 21. FDA. 2021. URL: https://www.fda.gov/tobacco-products/retail-sales-tobacco-products/tobacco-21 [accessed 2024-07-07]
8. FDA finalizes enforcement policy on unauthorized flavored cartridge-based e-cigarettes that appeal to children, including fruit and mint. FDA. 2020. URL: https://tinyurl.com/3s32tamb [accessed 2021-03-23]
9. Bendix A, Varinsky D. FDA orders JUUL to stop selling e-cigarette products. NBC News. 2022. URL: https://www.nbcnews.com/business/consumer/juul-ordered-stop-selling-fda-rcna34750 [accessed 2024-07-07]
10. Ovalle D. FDA reverses ban on JUUL products as supreme court may weigh in on vapes. Washington Post. 2024 Jun 6. URL: https://www.washingtonpost.com/health/2024/06/06/juul-fda-reverses-ban-vaping/ [accessed 2024-07-09]
11. Results from the annual National Youth Tobacco Survey. FDA. 2024. URL: https://www.fda.gov/tobacco-products/youth-and-tobacco/results-annual-national-youth-tobacco-survey [accessed 2024-11-20]
12. Big vape: the rise and fall of juul (TV mini series 2023) - episode list. IMDb. URL: https://www.imdb.com/title/tt14081634/episodes?season=1 [accessed 2024-07-07]
13. Hoffman BL, Wolynn R, Barrett E, et al. Viewer reactions to EVALI storylines on popular medical dramas: a thematic analysis of Twitter messages. J Health Commun 2023 May 4;28(5):282-291. [doi: 10.1080/10810730.2023.2201814] [Medline: 37057592]
14. Hoffman BL, Sidani JE, Miller E, et al. “Better Than Any DARE Program”: qualitative analysis of adolescent reactions to EVALI television storylines. Health Promot Pract 2024 Sep;25(5):814-822. [doi: 10.1177/15248399231177049] [Medline: 37312252]
15. Mubangizi V, Plastow J, Nakaggwa F, et al. Assessing changes in knowledge, attitudes, and intentions to use family planning after watching documentary and drama health education films: a qualitative study. Reprod Health 2022 Mar 12;19(1):65. [doi: 10.1186/s12978-022-01370-5] [Medline: 35279189]
16. Acerbi A, Burns J, Cabuk U, et al. Sentiment analysis of the Twitter response to Netflix’s Our Planet documentary. Conserv Biol 2023 Aug;37(4):e14060. [doi: 10.1111/cobi.14060] [Medline: 36661052]
17. Chu KH, Hershey TB, Hoffman BL, et al. Puff bars, tobacco policy evasion, and nicotine dependence: content analysis of tweets. J Med Internet Res 2022 Mar 25;24(3):e27894. [doi: 10.2196/27894] [Medline: 35333188]
18. Colditz JB, Welling J, Smith NA, James AE, Primack BA. World Vaping Day: contextualizing vaping culture in online social media using a mixed methods approach. J Mix Methods Res 2019 Apr;13(2):196-215. [doi: 10.1177/1558689817702753]
19. Sidani JE, Hoffman BL, Colditz JB, et al. E-cigarette-related nicotine misinformation on social media. Subst Use Misuse 2022;57(4):588-594. [doi: 10.1080/10826084.2022.2026963] [Medline: 35068338]
20. Sidani JE, Hoffman B, Colditz JB, et al. Discussions and misinformation about electronic nicotine delivery systems and COVID-19: qualitative analysis of Twitter content. JMIR Form Res 2022 Apr 13;6(4):e26335. [doi: 10.2196/26335] [Medline: 35311684]
21. Sidani JE, Colditz JB, Barrett EL, et al. I wake up and hit the JUUL: analyzing Twitter for JUUL nicotine effects and dependence. Drug Alcohol Depend 2019 Nov 1;204:107500. [doi: 10.1016/j.drugalcdep.2019.06.005] [Medline: 31499242]
22. Sidani JE, Colditz JB, Barrett EL, Chu KH, James AE, Primack BA. JUUL on Twitter: analyzing tweets about use of a new nicotine delivery system. J Sch Health 2020 Feb;90(2):135-142. [doi: 10.1111/josh.12858] [Medline: 31828791]
23. Calma J. Twitter just closed the book on academic research. The Verge. 2023 May 31. URL: https://www.theverge.com/2023/5/31/23739084/twitter-elon-musk-api-policy-chilling-academic-research [accessed 2024-07-07]
24. McLellan A, Schmidt-Waselenchuk K, Duerksen K, Woodin E. Talking back to mental health stigma: an exploration of YouTube comments on anti-stigma videos. Comput Human Behav 2022 Jun;131:107214. [doi: 10.1016/j.chb.2022.107214]

XSL•FO
RenderX

25. Toye F, Seers K, Barker K. “It’s like she’s talking about me” — exploring the value and potential impact of a YouTube film presenting a qualitative evidence synthesis about chronic pain: an analysis of online comments. Can J Pain 2020 Sep 24;4(3):61-70. [doi: 10.1080/24740527.2020.1785853] [Medline: 33987512]
26. Li Y, Ashley DL, Popova L. Users’ modifications to electronic nicotine delivery systems: content analysis of YouTube video comments. JMIR Infodemiology 2022;2(2):e38268. [doi: 10.2196/38268] [Medline: 35992739]
27. Miyazaki K, Uchiba T, Kwak H, An J, Sasahara K. The impact of toxic trolling comments on anti-vaccine YouTube videos. Sci Rep 2024 Mar 1;14(1):5088. [doi: 10.1038/s41598-024-54925-w]
28. Laureate CDP, Buntine W, Linger H. A systematic review of the use of topic models for short text social media analysis. Artif Intell Rev 2023 May 1;56(12):1-33. [doi: 10.1007/s10462-023-10471-x] [Medline: 37362887]
29. Jelodar H, Wang Y, Yuan C, et al. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimed Tools Appl 2019 Jun;78(11):15169-15211. [doi: 10.1007/s11042-018-6894-4]
30. Andreotta M, Nugroho R, Hurlstone MJ, et al. Analyzing social media data: a mixed-methods framework combining computational and qualitative text analysis. Behav Res 2019 Aug;51(4):1766-1781. [doi: 10.3758/s13428-019-01202-8]
31. Rubenking B, Bracken CC. Binge watching and serial viewing: comparing new media viewing habits in 2015 and 2020. Addict Behav Rep 2021 Dec;14:100356. [doi: 10.1016/j.abrep.2021.100356] [Medline: 34124334]
32. Hoffman BL, Rosenthal EL, Colditz JB, Mcgarry R, Primack BA. Use of Twitter to assess viewer reactions to the medical drama, Code Black. J Health Commun 2018;23(3):244-253. [doi: 10.1080/10810730.2018.1426660] [Medline: 29452057]
33. Al-Rawi A, Blackwell B, Zemenchik K, Lee K. Twitter misinformation discourses about vaping: systematic content analysis. J Med Internet Res 2023 Nov 10;25(1):e49416. [doi: 10.2196/49416] [Medline: 37948118]
34. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005 Nov;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
35. Weston SJ, Shryock I, Light R, Fisher PA. Selecting the number and labels of topics in topic modeling: a tutorial. Adv Methods Pract Psychol Sci 2023 Apr;6(2). [doi: 10.1177/25152459231160105]
36. Li S, Xie Z, Chiu DKW, Ho KKW. Sentiment analysis and topic modeling regarding online classes on the Reddit platform: educators versus learners. Appl Sci (Basel) 2023 Jan;13(4):2250. [doi: 10.3390/app13042250]
37. Prier KW, Smith MS, Giraud-Carrier C, Hanson CL. Identifying health-related topics on twitter: an exploration of tobacco-related tweets as a test topic. Presented at: Social Computing, Behavioral-Cultural Modeling and Prediction - 4th International Conference, SBP 2011; Mar 29-31, 2011; MD, USA p. 18-25.
38. Surian D, Nguyen DQ, Kennedy G, Johnson M, Coiera E, Dunn AG. Characterizing Twitter discussions about HPV vaccines using topic modeling and community detection. J Med Internet Res 2016 Aug 29;18(8):e232. [doi: 10.2196/jmir.6045] [Medline: 27573910]
39. O’Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014 Sep;89(9):1245-1251. [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]
40. Eissenberg T, Bhatnagar A, Chapman S, Jordt SE, Shihadeh A, Soule EK. Invalidity of an oft-cited estimate of the relative harms of electronic cigarettes. Am J Public Health 2020 Feb;110(2):161-162. [doi: 10.2105/AJPH.2019.305424] [Medline: 31913680]
41. Chu KH, Shensa A, Colditz JB, et al. Integrating social dynamics into modeling cigarette and e-cigarette use. Health Educ Behav 2020 Apr;47(2):191-201. [doi: 10.1177/1090198119876242] [Medline: 32090652]
42. Kalkhoran S, Glantz SA. Modeling the health effects of expanding e-cigarette sales in the United States and United Kingdom: a Monte Carlo analysis. JAMA Intern Med 2015 Oct;175(10):1671-1680. [doi: 10.1001/jamainternmed.2015.4209] [Medline: 26322924]
43. Kim I, Begay C, Ma HJ, et al. E-cigarette-related health beliefs expressed on Twitter within the U.S. AJPM Focus 2023 Jun;2(2):100067. [doi: 10.1016/j.focus.2023.100067] [Medline: 37790637]
44. Kwon M, Park E. Perceptions and sentiments about electronic cigarettes on social media platforms: systematic review. JMIR Public Health Surveill 2020 Jan 15;6(1):e13673. [doi: 10.2196/13673] [Medline: 31939747]
45. Wu D, Kasson E, Singh AK, et al. Topics and sentiment surrounding vaping on Twitter and Reddit during the 2019 e-cigarette and vaping use-associated lung injury outbreak: comparative study. J Med Internet Res 2022 Dec 13;24(12):e39460. [doi: 10.2196/39460] [Medline: 36512403]
46. Martinez LS, Hughes S, Walsh-Buhi ER, Tsou MH. “Okay, We Get It. You Vape”: an analysis of geocoded content, context, and sentiment regarding e-cigarettes on Twitter. J Health Commun 2018;23(6):550-562. [doi: 10.1080/10810730.2018.1493057] [Medline: 29979920]
47. Baghdadi JD, Coffey KC, Belcher R, et al. #Coronavirus on TikTok: user engagement with misinformation as a potential threat to public health behavior. JAMIA Open 2023 Apr;6(1):ooad013. [doi: 10.1093/jamiaopen/ooad013] [Medline: 36844368]
48. Baumel NM, Spatharakis JK, Karitsiotis ST, Sellas EI. Dissemination of mask effectiveness misinformation using TikTok as a medium. J Adolesc Health 2021 May;68(5):1021-1022. [doi: 10.1016/j.jadohealth.2021.01.029] [Medline: 33814279]
49. Li HY, Pastukhova E, Brandts-Longtin O, Tan MG, Kirchhof MG. YouTube as a source of misinformation on COVID-19 vaccination: a systematic analysis. BMJ Glob Health 2022 Mar;7(3):e008334. [doi: 10.1136/bmjgh-2021-008334] [Medline: 35264318]

XSL•FO
RenderX

50. Sampige R, Rodgers EG, Huang A, Zhu D. Education and misinformation: exploring ophthalmology content on TikTok. Ophthalmol Ther 2024 Jan;13(1):97-112. [doi: 10.1007/s40123-023-00834-6] [Medline: 37875699]
51. Alafwan B, Siallagan M, Putro US. Comments analysis on social media: a review. ICST Trans Scalable Inf Syst 2023 Oct 6;10(6). [doi: 10.4108/eetsis.3843]

Abbreviations

API: application programming interface
CDC: Centers for Disease Control and Prevention
e-cigarette: electronic cigarette
ENDS: electronic nicotine delivery system
EVALI: e-cigarette or vaping product use–associated lung injury
FDA: Food and Drug Administration
LDA: latent Dirichlet allocation
NMF: negative matrix factorization
NTP: nicotine and tobacco product
PMTA: premarket tobacco product application
RA: research assistant
SRQR: Standards for Reporting Qualitative Research
THC: tetrahydrocannabinol

Edited by A Stone, A Mavragani; submitted 29.04.25; peer-reviewed by A Chakravorty, SP Uppu; revised version received 26.08.25; accepted 28.08.25; published 19.09.25.

Please cite as:
Hoffman B, Tripathi A, Shensa A, Dou J(, Narendorf P, Hundi N, Sidani J
Response to the Netflix Docuseries "Big Vape: The Rise and Fall of JUUL": Mixed Methods Analysis of YouTube Comments Using Qualitative Coding and Topic Modeling
JMIR Form Res 2025;9:e76737
URL: https://formative.jmir.org/2025/1/e76737
doi:10.2196/76737

© Beth Hoffman, Arpita Tripathi, Ariel Shensa, Julia (Pengyue) Dou, Piper Narendorf, Nishi Hundi, Jaime Sidani. Originally published in JMIR Formative Research (https://formative.jmir.org), 19.9.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Assessing Work-Related Stressors in Online Counseling: Cross-Sectional Questionnaire Development Study

Wiebke Schlenger[1]; Marlies Joellenbeck[1], DrPH; Elke Ochsmann[2], Prof Dr Med

[1]Institute of Occupational Health, Prevention and Health Management, Faculty of Medicine, Lübeck University, Luebeck, Germany

[2]Institute of Occupational Medicine and Public Health, Faculty of Medicine, Saarland University, Kirrberger Str. 100, Homburg/Saar, Germany

Corresponding Author:
Elke Ochsmann, Prof Dr Med
Institute of Occupational Medicine and Public Health, Faculty of Medicine, Saarland University, Kirrberger Str. 100, Homburg/Saar, Germany

Abstract

Background: The rapid digitalization of health and social services, particularly accelerated by the COVID-19 pandemic, has led to a widespread adoption of online counseling. While offering flexibility and improved access for clients, online counseling presents new challenges for counselors, including technical issues, difficulties in building therapeutic relationships, and changes in work organization. Despite growing reliance on digital counseling platforms, there is a lack of validated tools to assess specific occupational stressors associated with online counseling.

Objective: This study aimed to develop and evaluate the "QueStrOn" (Questionnaire to Assess Stressors in Online Counseling), an instrument designed to identify stressors and resources specific to online counseling and to explore its ability to predict perceived stress levels in counselors.

Methods: Item development was guided by the Job Demands–Resources model, qualitative interviews with 22 counselors, expert input, and a literature review. A preliminary version of the questionnaire was pretested and then distributed via email and social media to counselors offering both online and face-to-face services. A total of 219 counselors completed the survey, and after applying inclusion criteria, 174 responses were analyzed. Exploratory factor analysis was conducted using principal axis factoring and varimax rotation. Internal consistency was assessed via Cronbach alpha (α). A linear regression model was used to initially test the predictive power of the identified factors with perceived digital stress as the dependent variable.

Results: The exploratory factor analysis resulted in a four-factor solution with 16 items, capturing (1) Online Work Organization, (2) Online Framework, (3) Online Work Content, and (4) Online Communication. The overall instrument demonstrated high internal consistency (α=0.870), with acceptable values for factors (1), (3), and (4) (α=0.754, 0.745, and 0.826, respectively), although the factor "Online Framework" showed limited reliability (α=0.502). The regression model, adjusted for age and gender, significantly predicted perceived stress in online counseling (F_5=13.335, P<.001), explaining 27.1% of the variance. Online Work Organization, Online Communication, and Online Framework were associated with lower perceived stress when rated positively, whereas Online Work Content showed an inverse relationship, potentially reflecting emotional distancing.

Conclusions: The QueStrOn instrument provides a valid first step toward systematically assessing occupational stressors in online counseling. Its 4-factor structure aligns with theoretical and empirical findings and offers practical utility for workplace risk assessments. Incorporating these dimensions into routine evaluations may support counselor well-being and inform digital health policy. Further validation and longitudinal studies are recommended to expand its applicability and explore associations with broader health outcomes.

(JMIR Form Res 2025;9:e65766) doi:10.2196/65766

KEYWORDS

workplace risk assessments; digital transformation; questionnaire; online counseling; therapeutic relationship; work organization; employee health

Introduction

Background

While face-to-face counseling has been established for decades, the use of online counseling has been increasing since the year 2000 [1], with a sharp rise during the COVID-19 pandemic [2-4], when the demand for counseling appointments skyrocketed amidst social isolation caused by lockdown measures. A survey by the American Psychiatric Association found that many counselors (81%) now use online counseling significantly more frequently than before the pandemic [4]. Similarly, a representative survey in Germany reported that

XSL•FO
RenderX

approximately 50% of counseling institutions now use video counseling with their clients, compared to just 5% before the pandemic [3].

Occupational Challenges in Online Counseling

Counselors report a high workload, frequent overtime, and emotional strain resulting from the complex and emotionally affecting issues presented by their clients [5,6]. These occupational risk factors are known to increase the prevalence of personal distress among counselors [7,8]. Personal distress can negatively affect health, reduce job satisfaction, and decrease willingness to use information and communication technology in the workplace [7,9,10].

The increased use of online counseling may introduce additional stressors for counselors, highlighting the need for tailored occupational health assessments and recommendations. Previous research indicated that many online counselors experience difficulties with the digital mode of counseling [11,12], as it may challenge the therapeutic relationship due to changes in counseling settings and altered facial expressions [12-15]. Both factors have been reported to affect the therapeutic relationship [14,16,17]. Asynchronous online counseling is particularly challenging for conveying emotions and maintaining clear communication [12,15,18]. Also, counseling techniques effective in face-to-face sessions, such as role plays or spatial experiments, may be less effective or insufficient in digital settings [15,19]. Achieving counseling objectives, such as establishing a therapeutic relationship and experiencing meaningful work, significantly contributes to job satisfaction [11,20,21].

Additional challenges of online counseling include variability in technology availability and proficiency, as well as client anonymity [12,22,23]. Technical difficulties can delay responses and disrupt counseling progress [24]. A stable internet connection and adequate technical equipment are necessary to experience the benefits of online counseling [25,26].

Client anonymity has the potential to facilitate disclosure by increasing clients' sense of safety, but it can also complicate assessments of client intentions and crisis interventions [18,23,27]. The increased use of online counseling may therefore negatively impact the therapeutic relationship and counseling outcomes compared to face-to-face counseling [14,17,28]. While some studies show that clients rate the therapeutic relationship of online counseling as comparable or even superior to that of face-to-face counseling, counselors and therapists still express concerns regarding the efficacy of online counseling and therapy [2,17,28,29].

Occupational Health Measures

Digitalization has been identified as a major driver for recent developments in the world of work [30-32]. While it may offer advantages and the potential to improve work performance and job satisfaction, it can also contribute to the development of distress and associated health risks [33,34]. Uncertainties related to technology use, social isolation from working in decentralized teams, and rapid information gathering and processing resulting from technology overload can lead to personal distress, burnout, and depression [7,34,35]. Occupational health and safety standards for online counseling, aimed at mitigating the development of psychological problems in the workplace, may become necessary [36,37]. In Germany, the German Occupational Health and Safety Act mandates systematic risk assessments and implementation of corresponding preventive measures [36].

Research Question

To our knowledge, no validated quantitative survey instruments currently exist for assessing specific stressors experienced by online counselors. This study aims to develop a framework for a questionnaire ("QueStrOn" [Questionnaire to Assess Stressors in Online Counseling]) designed to evaluate specific health risks associated with this kind of work. In addition, the interplay between questionnaire results and distress of online counselors was explored.

Methods

Study Design and Participants

For participant recruitment (invited convenience sample), counselor contact lists of the former BZgA (Bundeszentrale für gesundheitliche Aufklärung, Federal Center for Health Education, now: BIÖG: Federal Institute of Public Health) were reviewed, and counselors or counseling centers qualified for the study (ie, those offering online counseling via personal contact methods, such as email or video) were contacted by email [38]. A total of 844 counselors or counseling centers qualified and were invited to participate. However, the final number of counselors working within these centers could not be determined using this approach.

The study was also promoted in selected social media groups where counselors and psychotherapists offer online counseling services. Potential participants received an email with detailed information about the study's purpose, procedures, and estimated survey duration, along with a link to participate. Additional contact information was provided for any follow-up questions. The study information was also posted on the institute's home page.

The cross-sectional survey was conducted from July to September 2021 on the platform SoSci Survey [39]. Participation was voluntary. Altogether, 219 questionnaires were completed in full (completeness was checked after questionnaire submission). Inclusion criteria for analysis required participants to conduct both online and face-to-face consultations. Thus, 174 questionnaires met these criteria and were included in the analysis. The overall completion rate was 219 out of 361 (ie, the ratio of users who finished the survey to those who initially agreed to participate).

Questionnaire Development

The construction of the questionnaire to assess the stressors and resources of online counselors was based on the theoretical model of Job Demands and Resources by Bakker and Demerouti [40].

Specific stressors related to online counseling were initially identified through a qualitative study involving 22 consultants from a large German welfare organization, as described

XSL•FO
RenderX

elsewhere [11]. The responses indicated that digitization influences all known areas of work-related stressors with overlaps between digital and nondigital work, including counseling activity design, emotional involvement, working hours, communication, work environment, and social relations (colleagues, supervisors, and flexible working conditions). However, counselors also distinguished between digital and nondigital work, suggesting that stressors specific to online counseling, and thus relevant to digital occupational health, warrant further investigation.

These findings were then compared and validated by a comprehensive literature review focusing on factors of the common guideline for psychosocial risk assessments in Germany (GDA [Gemeinsame Deutsche Arbeitsschutzstrategie]; Joint German Occupational Safety and Health Strategy guideline for mental stress) [41]. Challenges related to work organization and limitations in establishing therapeutic relationships were identified as specific stressors in online counseling, whereas other areas (eg, social relations) were not considered unique for digital or online work. Work organization was further defined by clients' spatial and temporal flexibility, accompanied by changes in time management and documentation. Therapeutic relationships encompassed aspects like counselor-patient communication and nonverbal interaction.

The resulting questionnaire comprising 19 items was developed according to the established work organization and therapeutic relationships categories. These items were all placed on 1 questionnaire page. Sociodemographic data and health data were presented on other questionnaire pages. Responses consisted of paired negative and positive statements, rated on a Likert scale ranging from –3 (the aspect is a disadvantage of online counseling) to +3 (the aspect is an advantage of online counseling). Respondents were allowed to review and modify their answers.

In addition to exposure assessments, a stress outcome variable was included, asking participants to rate their stress level during online counseling compared to face-to-face counseling on a scale from (–3) "more stress during online counseling" to (+3) "more stress during face-to-face counseling."

A questionnaire for psychosomatic complaints in nonclinical settings was also administered [42].

The final questionnaire was refined following consultations with a psychologist, a health scientist, and an occupational physician, all with experience in mental health risk assessments. It was pretested online with 2 online counselors and technically checked for desktops (the use of mobile phones for participation was not recommended). The questionnaire items for the QueStrOn had to be filled in (at least by checking a box for "no answer") and participants were able to review and alter their answers before submission (by using a back button).

Statistics

Following descriptive analyses, an exploratory factor analysis (EFA) was conducted (principal axis method, varimax rotation, and eigenvalue >1). To assess the suitability of the variables for the EFA, the KMO (Kaiser-Meyer-Olkin) criterion (KMO>0.8) as well as the Bartlett Sphere Test (P<.05) were applied. Factor selection was also guided by interpretability and content relevance. Cronbach alpha was calculated to evaluate the reliability of the scales.

The outcome variable "digital stress" was further examined for gender and age effects using the chi-square test. Its correlation with psychosomatic health outcomes (back pain, nervousness, and sweat attacks) was assessed. The predictive power of the questionnaire for digital stress was evaluated using linear regression analysis with standardized beta coefficients and T-values reported. A significance level of P<.05 was considered statistically significant. Statistical analysis was performed using SPSS 28.0 (IBM Corp [43]).

Ethical Considerations

The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Ethics Committee of the University of Lübeck (protocol code AZ 19 - 353, approval date November 13, 2019). Participation was entirely voluntary, and no technical backtracking of respondents was possible. Respondents provided informed consent by reading detailed online information about the study and actively agreeing to complete the questionnaire before participation.

All data were collected and processed anonymously to protect participants' privacy, and this was clearly communicated to them. Measures were taken to ensure confidentiality, including secure data storage and restricted data access to study personnel only. Personal information was chosen, ensuring that participants could not be identified in any reports or publications. Participants were not compensated for their involvement in the study.

Results

Participant Characteristics

A total of 78% (136/174) of respondents were female. The mean age of participants was 46.6 (SD 11.3; range 24-72) years. Altogether, 39.6% (69/174) of respondents were employed full-time, 51.7% (90/174) were employed part-time, and 20.1% (35/174) were self-employed or freelancers.

Psychotherapists (46/174, 26.4%) comprised the largest occupational group, while the remaining respondents worked in various fields of social counseling. Video counseling was the most frequently used format for online counseling (132/174, 75.8% of participants). Nevertheless, face-to-face counseling still accounted for the majority of participants' working time.

Online Counseling Stressors and Resources

Regarding work organization, clients' spatial and temporal flexibility was perceived as an advantage of online counseling. Time for documenting digital or online sessions was reported to be similar to that of face-to-face counseling, and only minimal changes in counselors' time management were documented. However, 68.4% (119/174) of counselors stated that technical difficulties interfered with the counseling process.

Aspects supporting a strong therapeutic relationship were rated as more negative in online counseling compared to face-to-face sessions. Specifically, 71.8 % (125/174) of the respondents

XSL•FO
RenderX

reported that altered facial expressions and gestures negatively influence the therapeutic relationship, and 66.1% (115/174) of the respondents found the process of relationship building to be generally more difficult online. In addition, 77.2% (134/174) of the respondents stated that crisis management was more challenging in online counseling. While 49.5% (86/174) of counselors felt that they can express themselves clearly in online counseling, only 38.1% (66/174) believed that clients can do so.

Online Stress

With regard to stress outcomes, 34.1% (59/173) of counselors reported higher stress levels during online counseling compared to face-to-face sessions. Approximately 37% (64/173) of counselors experienced no difference in stress between the 2 counseling forms, and 29.5% (51/173) of counselors reported lower stress levels during online counseling. No significant age or gender effects were found for digital stress. However, digital stress showed significant correlations with specific psychosomatic complaints (back pain, r=–0.26; P=.003; nervousness, r=–0.19; P=.03; sweat attacks, r=0.180; P=.04).

EFA Results

Prior to the EFA, the items "documentation" and "costs for clients" were excluded due to content overlap with the item "time management," and due to respondents showing an overall neutral response pattern (61% neutral). The other results of the EFA are reported in Table 1.

Table . Explorative factor analysis of the QueStrOn (Principal Axis, Varimax with Kaiser Normalization; N=174), with resulting 4-Factor Solution.

"insert suitable column head"	Factor			
	1	2	3	4
Factor 1: "Online Communication"				
Verbal expression of counselors	0.843	—[a]	—	—
Verbal expression of clients	0.747	—	—	—
Counseling goals	0.634	—	—	—
Factor 2: "Online Work Organization"				
Temporal flexibility	—	0.915	—	—
Perceived accessibility of counselors	—	0.553	—	—
Time management	—	0.491	—	—
Spatial flexibility	—	0.408	—	—
Client anonymity	—	0.395	—	—
Factor 3: "Online Work Content"				
Facial expressions and gestures	—	—	0.757	—
Management of crisis	—	—	0.598	—
Relationship building	0.410	—	0.501	—
Engagement of clients	—	0.401	0.488	—
Factor 4: "Online Framework"				
Existence of guidelines	—	—	—	0.519
Technical difficulties	—	—	0.431	0.449
Data privacy	0.402	—	—	0.446
Privacy of clients	—	—	—	0.353

[a]Factor loadings <0.30 are not depicted in this table.

The KMO value was 0.816, indicating sampling adequacy. The Bartlett test of sphericity was significant. All retained factors exceeded 0.4, and loadings below 0.3 were excluded. We found that a 4-factor solution best fit the collected data:

1. Factor 1, which was named "Online Communication," includes items related to communication and goal attainment in online counseling.
2. Factor 2, which was named "Online Work Organization," encompasses items related to flexibility, time management, and related organizational aspects.
3. Factor 3, named "Online Work Content," covers items describing the counseling process and quality of the therapeutic relationship. The item "relationship building" exhibited cross-loadings on factor 1, likely due to its conceptual overlap with communication and work content. The item "engagement of clients" cross-loaded on factor 2. A possible explanation for this is that factor 2 includes client-centered questions. Although cross-loadings were noted, both items were retained and scored according to their higher loading on factor 3, as exclusion was not deemed necessary.

XSL•FO
RenderX

4. Factor 4, which was named "Online Framework," reflects the general conditions of online counseling. Within factor 4, the item "data privacy" cross-loaded on factor 1, which can be explained by communication difficulties while protecting privacy. The item "technical difficulty" cross-loaded on factor 3, reflecting possible interruptions during therapeutic relationship building or crisis intervention. Nevertheless, both items remained within factor 4 due to their higher factor loadings.

Reliability

To assess the reliability of the factors, internal consistency was evaluated by calculating Cronbach alpha for each factor, as well as item discriminatory power. The overall Cronbach α for the instrument was 0.870. The scales for "Online Work Organization," "Online Work Content," and "Online Communication" demonstrated good internal consistency, with Cronbach α values of 0.826 (Factor 1), 0.754 (Factor 2), and 0.745 (Factor 3), respectively. The factor 'Online Framework' though, showed low reliability with α=0.502.

Regressions

To evaluate and confirm the predictive power of the QueStrOn, a predictive model was developed, which used "stress level in online counseling" as the dependent variable, while the 4 factors of the QueStrOn were included as predictors in a linear regression analysis (R^2=0.271). The model was adjusted for age and gender. The ANOVA for the regression model was significant (F_5=13.335; P<.001). Significant results were found for all 4 factors (Table 2). The regression results indicate that perceiving Online Work Organization, Online Communication, and the Online Framework as advantageous is associated with lower stress levels in online counseling compared to face-to-face counseling. Online Work Content also emerged as a significant predictor of stress level; however, its effect was in the opposite direction, indicating that higher demands in work content are associated with increased stress levels. Thus, a significant regression model was obtained, supporting the hypothesis that the QueStrOn explains variance in the personal stress levels of online counselors.

Table . Exploratory linear regression model between predictors (QueStrOn 4-factor solution) and health outcome (stress level in Online Counseling; N=174; R^2=0.271).

EFA-factors	Standardized beta coefficient[a]	T[a]	*P* value[a]
Online Work Organization	0.274	3.527	<.001
Online Communication	0.274	3.428	<.001
Online Work Content	–0.172	–2.084	.04
Online Framework	0.274	3.663	<.001

[a]adjusted for age age and gender.

Discussion

Principal Findings

The main objective of the study was to develop and explore an instrument that identifies a specific risk profile for online counseling compared to face-to-face counseling. The initial model for factor analysis was based on preliminary interviews, expert opinions, and an extensive literature review and was pretested by online counselors. EFA then resulted in a 4-factor solution with 16 items, defined as (1) Online Work Organization, (2) Online Framework, (3) Online Work Content, and (4) Online Communication.

The factor "Online Work Organization" demonstrated good internal consistency. It primarily comprises counselor-centered questions (eg, feeling of constant availability) but also includes client-centered questions (eg, flexibility of clients). Previous studies have also found client flexibility in location and time alongside counselors' sense of constant availability as significant factors influencing the relationship between work processes and adverse health effects [11,13,44-46].

The therapeutic relationship remains the core work content for achieving counseling goals. It can be subdivided into successful relationship building, effective crisis intervention, and sufficient client engagement [21,47,48]. Many studies have shown that these aspects foster a good therapeutic relationship and serve as milestones in the therapy process and progress [12,14,21,47,49]. All these content-related factors can be negatively impacted by deficits in verbal and nonverbal communication, which have been previously identified as pivotal influences [12,14,15].

The factor structure suggests 2 focal points for the introduction and development of online counseling formats. These include regulations to improve work organization, especially concerning time management and counselors' accessibility; also, the promotion of therapeutic and method skills facilitates digital and online relationship building and maintenance. In addition, the factor structure implies that framework conditions related to digital transformation, such as data protection and privacy, which often play a more subordinate role in face-to-face contact, should receive focused attention in future online counseling research.

The subscale "Online Framework" exhibited only sufficient internal consistency. This could reflect the broad range of work and workplace design possibilities associated with digital work. Despite this result, the factor should not be dismissed, as it showed a strong predictive value for online counseling stress. While our previous interviews revealed associations and interactions between framework factors and other workplace health factors in digital contexts [11], few studies specifically examined organizational frameworks and their contribution to digital stress. Again, related aspects, such as data security and

XSL•FO
RenderX

privacy, are frequently mentioned in the literature [1,13,22], underscoring the need for focused research on organizational framework conditions in digital social work.

The overall instrument showed strong internal consistency of α=0.870, with acceptable to satisfactory alphas across the 4 factors. These findings align well with theoretical assumptions and empirical findings.

To evaluate the instrument's predictive utility for stress levels in online counselors, the factors were included in a linear regression model with "Online Stress" as the dependent variable. Online stress in our cohort correlated with psychosomatic complaints, suggesting an indirect link between QueStrOn factors and occupational health, though this finding warrants further investigation. The regression model explained approximately 27.1% of variance in stress levels. It is plausible that adding the demands associated with Online Work Organization increases perceived stress. While the model leaves a substantial proportion of variance unexplained, other nononline-specific stressors (eg, bureaucratic hurdles or client aggression) likely contribute substantially to the prediction of stress and should be integrated in future assessments [9,50]. Several existing instruments address these broader factors and should be considered in subsequent research to refine the understanding of online counseling stressors as outlined by the QueStrOn.

Interestingly, the factor "Online Work Content" showed an inverse relationship with stress levels. One possible explanation is that a poor therapeutic relationship may reduce counselors' identification with the client's problems, allowing a greater emotional distance and thus lower perceived stress compared to face-to-face counseling. It might also indicate that choosing a digital exchange induces less emotional commitment and reduces tendencies toward pathological altruism, a behavioral pattern often observed in social work [51]. Further research is needed to test this hypothesis.

The regression results underscore the importance of employer collaboration with counselors when implementing online counseling. Counselors' perceptions of relationship management and work organization appear to influence their stress level. Occupational health literature suggests that elevated stress can impair health and increase absenteeism, highlighting the need for further research in this area.

Overall, the QueStrOn factors appear suitable for addressing additional challenges unique to online counseling. Future work will focus on refining the instrument to create a practicable, brief tool to assess online counseling stressors. This tool could supplement regular occupational risk assessments and help evaluate the impact of workplace interventions.

As this study, and the QueStrOn instrument, represent a first step in identifying key risk factors related to online counseling, several limitations should be acknowledged. The sample size was relatively small, so generalizability of results should be considered with caution. As the first instrument specifically targeting stressors in online counseling, direct comparisons regarding item effectiveness are not yet feasible. Therefore, the comparability of our explorative data is also limited. To date, much of the available information on the demands of online counseling is qualitative. In this context, our findings align well with pre-existing conceptions of stressors in online counseling.

In addition, participants primarily conducted face-to-face counseling with online counseling as an adjunct to their daily working routine. Consequently, results may be influenced by the dual-method work environment and could evolve with increasing experience and technological advances. Counselors primarily trained in digital methods might experience fewer challenges and health effects. Since most counselors use both asynchronous and synchronous formats similarly, the lines between digital and face-to-face stressors might be blurred. Furthermore, this study was conducted in the first year of the COVID-19 pandemic, a period marked by exceptional stress due to isolation, lockdowns, health uncertainties, and information overload. These extraordinary circumstances likely influenced consultants' responses, given the parallel heightened demand for counseling services and the rapid necessity for digital or online delivery. Further and longitudinal research in more stable social conditions should answer the question of whether the pandemic led to a situational bias in our study.

Conclusion

This study presents the development and validation of QueStrOn, a novel instrument designed to assess stressors specific to online counseling. Based on the Job Demands-Resources model, 4 key factors were identified, including Online Work Organization, Online Communication, Online Work Content, and Online Framework. These dimensions significantly predicted perceived stress levels, with challenges in communication and therapeutic engagement associated with increased stress, while positive evaluations of organizational and structural conditions correlated with reduced stress.

The findings highlight the need for targeted occupational health strategies addressing the unique demands of online counseling and, prospectively, any online-based interactive task between humans. QueStrOn offers a reliable and theory-driven tool to support systematic risk assessment and inform future interventions aimed at improving counselor well-being in digital service delivery contexts.

Acknowledgments

We want to thank all voluntary questionnaire participants. This research was funded by the Employer's Liability Insurance Association for Health Services and Welfare Care (BGW; grant extFF_1379). The authors attest that there was no use of generative artificial intelligence (AI) technology in the generation of the text or other informational content of this manuscript. After reviewer feedback; however, authors used the generative AI tool ChatGPT by OpenAI for language checks, all of which were rechecked by the authors before resubmission.

XSL•FO
RenderX

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization:EO, MJ, WS
Data curation: WS, EO
Formal analysis: WS, EO
Funding acquisition: EO, MJ
Investigation: WS, MJ, EO
Methodology: WS, EO, MJ
Project administration: WS, MJ, EO
Supervision: EO, MJ
Validation: EO
Writing – original draft: WS, EO
Writing – review & editing: EO, MJ, WS

Conflicts of Interest

None declared.

References

1. Mishna F, Bogo M, Root J, Sawyer JL, Khoury-Kassabri M. “It just crept in”: the digital age and implications for social work practice. Clin Soc Work J 2012 Sep;40(3):277-286. [doi: 10.1007/s10615-012-0383-4]
2. Simpson S, Richardson L, Pietrabissa G, Castelnuovo G, Reid C. Videotherapy and therapeutic alliance in the age of COVID‐19. Clin Psychology and Psychoth 2021 Mar;28(2):409-421. [doi: 10.1002/cpp.2521]
3. Stieler M, Lipot S, Lehmann R. Zum stand der onlineberatung in zeiten der corona krise (on the status of online counseling during the corona crisis). In: Entwicklungs- Und Veränderungsprozesse Der Onlineberatungslandschaft: Universität für Weiterbildung Krems & e-beratungsjournal.net; 2022. [doi: 10.48341/262P-7T64]
4. Psychiatrists use of telepsychiatry during COVID-19 public health emergency. American Psychiatric Association. 2021. URL: https://www.psychiatry.org/File%20Library/Psychiatrists/Practice/Telepsychiatry/APA-Telehealth-Survey-2020.pdf [accessed 2025-09-17]
5. Simpson S, Simionato G, Smout M, et al. Burnout amongst clinical and counselling psychologist: the role of early maladaptive schemas and coping modes as vulnerability factors. Clin Psychol Psychother 2019 Jan;26(1):35-46. [doi: 10.1002/cpp.2328] [Medline: 30203882]
6. Rupert PA, Miller AO, Dorociak KE. Preventing burnout: what does the research tell US? Prof Psychol Res Pract 2015;46(3):168-174. [doi: 10.1037/a0039297]
7. Fye HJ, Cook RM, Choi Y, Baltrinic ER. Professional quality of life and affective distress among prelicensed counselors. J Couns Dev 2021 Oct;99(4):429-439. [doi: 10.1002/jcad.12395]
8. Baldwin-White A. Psychological distress and substance abuse counselors: an exploratory pilot study of multiple dimensions of burnout. J Subst Use 2016 Jan 2;21(1):29-34. [doi: 10.3109/14659891.2014.949316]
9. Luther L, Gearhart T, Fukui S, Morse G, Rollins AL, Salyers MP. Working overtime in community mental health: associations with clinician burnout and perceived quality of care. Psychiatr Rehabil J 2017 Jun;40(2):252-259. [doi: 10.1037/prj0000234] [Medline: 27786520]
10. Batista-Taran LC, Reio TG. Occupational stress: towards an integrated model. In: Plakhotnik MS, Nielsen SM, Pane DM, editors. Presented at: Tenth Annual College of Education & GSN Research Conference; Apr 23, 2011; Miami, FL p. 9-16 URL: https://core.ac.uk/download/pdf/46946443.pdf [accessed 2025-09-17]
11. Schlenger W, Jöllenbeck M, Stamer T, Grosse A, Ochsmann E. Digitizing social counseling—insights for workplace health management. Int J Environ Res Public Health 2022 Jan 14;19(2):917. [doi: 10.3390/ijerph19020917] [Medline: 35055750]
12. Aafjes-van Doorn K, Békés V, Prout TA. Grappling with our therapeutic relationship and professional self-doubt during COVID-19: will we use video therapy again? Couns Psychol Q 2021 Oct 2;34(3-4):473-484. [doi: 10.1080/09515070.2020.1773404]
13. Stoll J, Müller JA, Trachsel M. Ethical issues in online psychotherapy: a narrative review. Front Psychiatry 2019;10:993. [doi: 10.3389/fpsyt.2019.00993] [Medline: 32116819]
14. Harris B, Birnbaum R. Ethical and legal implications on the use of technology in counselling. Clin Soc Work J 2015 Jun;43(2):133-141. [doi: 10.1007/s10615-014-0515-0]
15. Mishna F, Bogo M, Sawyer JL. Cyber counseling: illuminating benefits and challenges. Clin Soc Work J 2015 Jun;43(2):169-178. [doi: 10.1007/s10615-013-0470-1]
16. Haberstroh S, Parr G, Bradley L, Morgan‐Fleming B, Gee R. Facilitating online counseling: perspectives from counselors in training. J Couns Dev 2008 Oct;86(4):460-470 [FREE Full text] [doi: 10.1002/j.1556-6678.2008.tb00534.x]

XSL•FO
RenderX

17. Rees CS, Stone S. Therapeutic alliance in face-to-face versus videoconferenced psychotherapy. Prof Psychol Res Pract 2005;36(6):649-653. [doi: 10.1037/0735-7028.36.6.649]
18. Rochlen AB, Zack JS, Speyer C. Online therapy: review of relevant definitions, debates, and current empirical support. J Clin Psychol 2004 Mar;60(3):269-283. [doi: 10.1002/jclp.10263] [Medline: 14981791]
19. Frittgen EM, Haltaufderheide J. “Can you hear me?”: communication, relationship and ethics in video-based telepsychiatric consultations. J Med Ethics 2022 Jan;48(1):22-30. [doi: 10.1136/medethics-2021-107434] [Medline: 34610979]
20. Harrison G, Gordon E. Counsellors’ experiences of providing counselling to students in university‐level institutions in Ireland: an evolving phenomenon. Couns and Psychother Res 2021 Dec;21(4):805-815 [FREE Full text] [doi: 10.1002/capr.12441]
21. Bordin ES. The generalizability of the psychoanalytic concept of the working alliance. Psychother Theory Res Pract 1979;16(3):252-260. [doi: 10.1037/h0085885]
22. Cipolletta S, Mocellin D. Online counseling: an exploratory survey of Italian psychologists’ attitudes towards new ways of interaction. Psychother Res 2018 Nov;28(6):909-924. [doi: 10.1080/10503307.2016.1259533] [Medline: 28068875]
23. Dunn K. A qualitative investigation into the online counselling relationship: to meet or not to meet, that is the question. Couns and Psychother Res 2012 Dec;12(4):316-326. [doi: 10.1080/14733145.2012.669772]
24. Feijt M, de Kort Y, Bongers I, Bierbooms J, Westerink J, IJsselsteijn W. Mental health care goes online: practitioners’ experiences of providing mental health care during the COVID-19 pandemic. Cyberpsychol Behav Soc Netw 2020 Dec;23(12):860-864. [doi: 10.1089/cyber.2020.0370] [Medline: 32815742]
25. Körner U, Müller-Thur K, Lunau T, Dragano N, Angerer P, Buchner A. Perceived stress in human-machine interaction in modern manufacturing environments-Results of a qualitative interview study. Stress Health 2019 Apr;35(2):187-199. [doi: 10.1002/smi.2853] [Medline: 30609231]
26. Rohwer E, Flöther JC, Harth V, Mache S. Overcoming the “dark side” of technology—a scoping review on preventing and coping with work-related technostress. Int J Environ Res Public Health 2022 Mar 18;19(6):3625. [doi: 10.3390/ijerph19063625] [Medline: 35329312]
27. Centore AJ, Milacci F. A study of mental health counselors’ use of and perspectives on distance counseling. J Ment Health Couns 2008 Jul 1;30(3):267-282. [doi: 10.17744/mehc.30.3.q871r684n863u75r]
28. Maurya RK, Bruce MA, Therthani S. Counselors’ perceptions of distance counseling: a national survey. J Asia Pac Couns 2020 Aug 31;10(2):1-22 [FREE Full text] [doi: 10.18401/2020.10.2.3]
29. Cook JE, Doyle C. Working alliance in online therapy as compared to face-to-face therapy: preliminary results. Cyberpsychol Behav 2002 Apr;5(2):95-105. [doi: 10.1089/109493102753770480] [Medline: 12025884]
30. Gerlmaier A, Latniak E, editors. Handbook of Psycho-Social Design of Digital Production Work: Strengthening Health Resources through Organizational Design Competence: Gabler Verlag; 2019. [doi: 10.1007/978-3-658-26154-2]
31. West D, Heath D. Theoretical pathways to the future: globalization, ICT and social work theory and practice. J Soc Work (Lond) 2011 Apr;11(2):209-221. [doi: 10.1177/1468017310386835]
32. Blok M, Korte EM, Groenesteijn L, Formanoy M, Vink P. The effects of a task facilitating working environment on office space use, communication, concentration, collaboration, privacy and distraction. Presented at: Proceedings of the 17th World Congress on Ergonomics; Aug 9-14, 2009; Beijing, China p. 1-7 URL: https://repository.tudelft.nl/islandora/object/uuid%3Acec6d1ae-cb56-4241-a961-37f112765932 [accessed 2025-09-17]
33. Tams S, Ahuja M, Thatcher J, Grover V. Worker stress in the age of mobile technology: the combined effects of perceived interruption overload and worker control. J Strateg Inf Syst 2020 Mar;29(1):101595. [doi: 10.1016/j.jsis.2020.101595]
34. Ragu-Nathan TS, Tarafdar M, Ragu-Nathan BS, Tu Q. The consequences of technostress for end users in organizations: conceptual development and empirical validation. Inf Syst Res 2008 Dec;19(4):417-433. [doi: 10.1287/isre.1070.0165]
35. Ayyagari R, Grover V, Purvis R. Technostress: technological antecedents and implications. MIS Q 2011;35(4):831. [doi: 10.2307/41409963]
36. Federal Ministry of Justice. Law on the implementation of occupational safety measures to improve the safety and health protection of employees at work (Occupational Safety and Health Act - ArbSchG). URL: http://www.gesetze-im-internet.de/arbschg/index.html#BJNR124610996BJNE000701377 [accessed 2025-09-17]
37. The 1st international conference on health promotion, otawa, 1986. World Health Organisation. URL: https://www.who.int/teams/health-promotion/enhanced-wellbeing/first-global-conference [accessed 2025-09-11]
38. Counseling center addresses. Federal Institute of Public Health. 2021. URL: https://www.bzga.de/service/beratungsstellen/ [accessed 2025-09-11]
39. Leiner DJ. SoSci survey (version 3409). : Computer software; 2019 URL: https://www.soscisurvey.de/de/index [accessed 2025-09-17]
40. Bakker AB, Demerouti E. The Job Demands‐Resources model: state of the art. J Manag Psychol 2007 Apr 3;22(3):309-328. [doi: 10.1108/02683940710733115]
41. Guideline for counseling and monitoring of psychological stress at work. Joint German Occupational Safety Strategy (GDA). 2018. URL: https://www.gda-portal.de/DE/Aufsichtshandeln/Psychische-Belastung [accessed 2025-09-17]
42. Mohr G, Müller A. Psychosomatic complaints in a non-clinical context. Compilation of Social Science Items and Scales (ZIS). 2004. URL: https://doi.org/10.6102/zis78 [accessed 2025-09-17]

XSL•FO
RenderX

43. IBM SPSS statistics for windows, version 280. IBM. 2021. URL: https://www.ibm.com/de-de/products/spss-statistics [accessed 2025-09-17]
44. Dettmers J. How extended work availability affects well-being: the mediating roles of psychological detachment and work-family-conflict. Work & Stress 2017 Jan 2;31(1):24-41. [doi: 10.1080/02678373.2017.1298164]
45. Dettmers J, Biemelt J. Always available – the role of perceived advantages and legitimacy. J Manag Psychol 2018 Nov 14;33(7/8):497-510. [doi: 10.1108/JMP-02-2018-0095]
46. Brauner C, Wöhrmann AM, Michel A. Work availability types and well-being in Germany – a latent class analysis among a nationally representative sample. Work & Stress 2022 Jul 3;36(3):251-273. [doi: 10.1080/02678373.2021.1969475]
47. Miller SD, Duncan BL, Brown GS, Sorrell R, Chalk MB. Using formal client feedback to improve retention and outcome: making ongoing, real-time assessment feasible. J Brief Ther 2006;5(1):5-22.
48. Elliott R, Bohart AC, Watson JC, Greenberg LS. Empathy. Psychotherapy (Chic) 2011 Mar;48(1):43-49. [doi: 10.1037/a0022187] [Medline: 21401273]
49. Stargell NA. Therapeutic relationship and outcome effectiveness: implications for counselor educators. JCPS 2017. [doi: 10.7729/92.1164]
50. Hickey R. Prosocial motivation, stress and burnout among direct support workers. J Appl Res Intellect Disabil 2014 Mar;27(2):134-144. [doi: 10.1111/jar.12058] [Medline: 23649915]
51. Teubner-Guerra RL, Ochsmann E. Development and evaluation of a questionnaire for assessing pathological altruism. HBScience 2014 Feb;5(1):12-19. [doi: 10.1007/s16024-013-0179-7]

Abbreviations

BZgA: Bundeszentrale für gesundheitliche Aufklärung
EFA: exploratory factor analysis
GDA: Gemeinsame Deutsche Arbeitsschutzstrategie
KMO: Kaiser-Meyer-Olkin
QueStrOn: Questionnaire to Assess Stressors in Online Counseling

Edited by A Mavragani; submitted 25.08.24; peer-reviewed by A AL-Asadi; revised version received 17.06.25; accepted 02.07.25; published 25.09.25.

Please cite as:
Schlenger W, Joellenbeck M, Ochsmann E
Assessing Work-Related Stressors in Online Counseling: Cross-Sectional Questionnaire Development Study
JMIR Form Res 2025;9:e65766
URL: https://formative.jmir.org/2025/1/e65766
doi:10.2196/65766

© Wiebke Schlenger, Marlies Joellenbeck, Elke Ochsmann. Originally published in JMIR Formative Research (https://formative.jmir.org), 25.9.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Bactericidal Efficacy of Ultraviolet-C Light on Virtual Reality Devices: In Vitro Assessment of Bacterial Killing

Scott C Roberts[1], MS, MD; Jayson Wright[2]; Mahnoor Mustafa[2]; Richard S Feinn[3], PhD; Asher Marks[4], MD; Kimberly Hieftje[4], PhD; Pamela H Huang[4], MS; Richard A Martinello[1], MD; Thomas S Murray[4], MD, PhD; HIPSTER Team[5,7]

[1]Department of Internal Medicine, Section of Infectious Diseases, Yale School of Medicine, 20 York Street, Hunter 527, New Haven, CT, United States
[2]Yale University, New Haven, CT, United States
[3]Quinnipiac University, Hamden, CT, United States
[4]Department of Pediatrics, Yale School of Medicine, New Haven, CT, United States
[5]Yale New Haven Health System, New Haven, CT, United States
[7]

Corresponding Author:
Scott C Roberts, MS, MD
Department of Internal Medicine, Section of Infectious Diseases, Yale School of Medicine, 20 York Street, Hunter 527, New Haven, CT, United States

Abstract

Background: Virtual reality (VR) headsets are increasingly used in health care settings for a variety of clinical indications, yet processes to ensure safe use between patients are not well-established. Centers vary in how these processes are performed. Most use disinfection wipes that require manual contact with VR devices for a specified dwell time to allow for sufficient pathogen killing, which may introduce manual error and device degradation over time. Ultraviolet-C light (UV-C) devices offer a no-touch, low-cost, and passive method to achieve pathogen killing without the harms of chemical contact-based disinfectants. The use of UV-C for disinfection has been studied for some medical devices but its efficacy for microbe killing on VR headsets is not well-established.

Objective: This study aims to determine the bactericidal efficacy of UV-C on VR headsets through quantifying UV-C irradiance and bacterial killing of 3 commercially available UV-C devices.

Methods: Three commercially available, low-cost UV-C devices were tested for UV-C energy output at multiple positions, angles, and times across the devices' zone of disinfection. The top and lens of a VR headset, the Meta Oculus Quest 2, were artificially inoculated with high quantities of 3 different strains of bacteria (*Staphylococcus aureus*, *Pseudomonas aeruginosa*, and *Staphylococcus epidermidis*) and subjected to UV-C light according to each device's manufacturer's instructions for use. The primary outcome was the amount of bacterial killing after exposure to UV-C light.

Results: All 3 UV-C devices produced a UV-C dose that ranged from 3.57 to 195.37 mJ/cm^2, depending on proximity, angle, irradiance, and time the sensor received. At least 3-log$_{10}$ killing of all 3 strains of bacteria was achieved for each of the tested UV-C devices; however, there was variability by organism with respect to UV-C device and VR headset location within the device, notably with the proximity of the bacteria to the bulb. *S aureus* and *P aeruginosa* were more readily killed than *S epidermidis*, with increased bacterial killing occurring with increased UV-C exposure doses. There was no experiment in which all bacteria were killed.

Conclusions: UV-C dosage increased with exposure irradiance, time, proximity, and angle to the bulb for all 3 UV-C devices. Bacterial killing on the top and lens of a VR headset occurred in all 3 UV-C devices when run according to their manufacturer's instructions for use, although full bacterial killing did not occur in any experiment. UV-C may be an effective method for microbial killing on VR equipment with low-level contamination.

(JMIR Form Res 2025;9:e70326) doi:10.2196/70326

KEYWORDS

ultraviolet-C light; disinfection; virtual reality; bacteria; health care

XSL•FO
RenderX

Introduction

Virtual reality (VR) devices are being increasingly used in health care settings for a variety of clinical indications in patients, as well as for education of health care personnel. VR devices are neither designed nor designated as medical products per the United States Food & Drug Administration (US FDA) [1] and typically are without manufacturer instructions for disinfection after use in health care. VR devices, usually composed of a headset and controllers, are composed of a multitude of materials, including porous material, and have complicated assembles including gaps, joints, and other surfaces that make adequate and efficient disinfection challenging. This poses a safety issue, as health care settings comprise vulnerable populations at risk of infection if pathogens are transmitted between patients and health care personnel due to inadequately disinfected devices. Outbreaks of pathogens spread by contaminated reusable medical devices have been well described [2].

We have previously demonstrated the efficacy of low-level disinfectant wipes, notably those composed of isopropyl alcohol and quaternary ammonium, in sufficiently disinfecting VR headsets with a suggested standardized operating procedure [3]. However, contact-based disinfection with wipes is susceptible to manual error, a need for prolonged dwell times of the disinfectant material, selection of the proper United States Environmental Protection Agency (US EPA)–registered disinfectant wipe according to the organisms of concern, and can be environmentally wasteful. Additionally, repeated use may damage the integrity of the devices, especially the lenses. At least 1 major VR headset manufacturer specifically recommends against isopropyl alcohol as a disinfectant for this very reason [4].

Ultraviolet-C light (UV-C), a no-touch disinfection method, is potentially an attractive alternative because it is safe, low-cost, and may not degrade the material as much as contact-based disinfectants such as bleach [5]. The standard UV-C distribution with each application avoids manual errors associated with wiping. Additionally, UV-C is known to kill organisms not routinely disinfected with most standard wipes (isopropyl alcohol and quaternary ammonium) found in health care settings, including norovirus, *Candida auris*, and *Clostridioides difficile* [6-8]. UV-C has previously been shown to achieve up to a 4-log decontamination of experimentally contaminated patient rooms with high inocula of various multidrug-resistant organisms, including methicillin-resistant *Staphylococcus aureus*, vancomycin-resistant *Enterococcus* species, carbapenem-resistant *Acinetobacter baumannii*, and *C difficile* spores, with greatest decontamination achieved with direct, as opposed to indirect, line of sight [9,10]. Newer UV-C devices have since been shown to achieve this in health care settings in as little as 5 minutes when the appropriate conditions of distance and irradiation dose are met, supporting shorter exposure times as sufficient [11]. However, many factors affect the effectiveness of UV-C disinfection, including pathogen quantity, type of dispersal (eg, drop vs spread), surface material, organic load, pathogen type, humidity, temperature, and spore formation, as well as the intensity, time, and angle of light exposure [5].

Use of VR headsets outdoors is not recommended due to potential sun damage on the lens, and direct sunlight exposure, even indoors, can result in irreversible damage in under a minute [12]. This damage is secondary to UV-A (315‐400 nanometers), UV-B (280‐315 nanometers), and UV-C (100‐280 nanometers) rays, which even at shorter wavelengths could still damage plastics and polymers found in VR headsets. To our knowledge, disinfection of VR headsets with UV-C technology has not been formally studied. If successful, this offers an alternative to disinfectant wipes as a method of safe disinfection between different users.

UV-C devices are regulated as class II medical devices according to the United States Code of Federal Regulations; however, many devices in routine use have not been approved for use by the US FDA, which requires rigorous testing to meet established regulatory standards [13]. Despite this, manufacturers of some UV-C devices market these as effective for disinfection in health care settings. In this in vitro assessment, we sought to determine the bacterial killing of 3 commercially available UV-C products, including one specifically designed and marketed for VR headset disinfection, against 3 pathogens common in health care settings artificially inoculated on VR headsets.

Methods

UV-C Output Measurements

Three UV-C devices were tested. The Cleanbox (Model 01‐02CX×1; Cleanbox Tech) used 265-nanometer UV-C by LED light and did not have reflective material. The LiViliti 59SUV-C LED Sanitizing Box (Model SZH40-T5; LiViliti Health Products Corp) used 260‐ to 280-nanometer UV-C by LED light and had reflective materials on the top and bottom of the device. The Philips UV-C Disinfection Box (Model 9290024876; Philips, Koninklijke Philips NV) used 254-nanometer UV-C by mercury lamp and had reflective materials on all sides of the device. The Cleanbox device was developed specifically for VR headset disinfection, while the LiViliti and Philips devices were chosen due to their low cost and advertised ability to disinfect medical devices (primarily positive pressure ventilation masks, such as continuous positive airway pressure or bilevel positive airway pressure masks). These UV-C disinfection devices are also large enough to fit the VR headset. Importantly, these latter 2 devices were not marketed for, nor reported to be tested with, VR headsets. None of these UV-C devices have been approved by the US FDA; however, they are regulated as class II medical devices. Conditions in this experiment were not designed to assess disinfection according to established regulatory standards by the US FDA, but rather this study assessed bacterial killing for 3 commonly encountered health care pathogens to determine whether UV-C devices have efficacy for VR disinfection that should be further validated to the more rigorous standards required by bodies such as the US FDA and the US EPA. While the Spaulding Classification is intended for medical devices, and VR headsets are not designated as such, they would be classified as noncritical devices that should be cleaned and disinfected with low- or intermediate-level disinfectants. The

XSL•FO
RenderX

manufacturers' instructions for use (MIFU) differed by time of exposure per device: 1 minute for the Cleanbox (Figure 1A), 3 minutes for the LiViliti (Figure 1B), and 10-20 minutes for the Philips (Figure 1C). The VR headset was placed in each UV-C device as shown in Figure 1.

Figure 1. Positioning of the virtual reality (VR) headset according to ultraviolet-C light (UV-C) device. (A) In the Cleanbox, the VR headset is positioned such that the bulbs emit UV-C outward, with direct exposure to the inside strap and the lens. (B) In the LiViliti device, the bulbs emit UV-C from the top and all 4 sides with reflective material on the top and bottom. There is direct exposure to both the top and the lens. (C) In the Philips device, bulbs emit UV-C from the top of the device with reflective material on all sides. The top receives direct exposure as shown. For lens studies, the headset was rotated so that the lens faced the light source. Blue circles on the VR headset represent areas of bacterial inoculation.

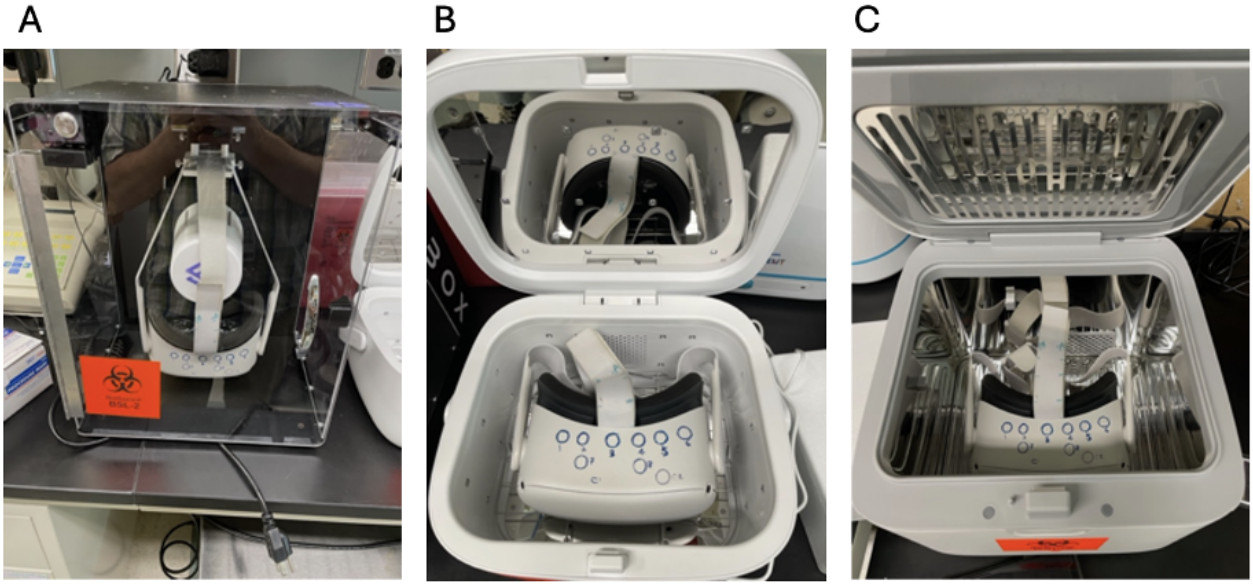

This resulted in a different position of the VR device in relation to the UV-C bulbs for each device. In the Cleanbox device, the VR device was hung by a strap with the bulbs positioned between the strap and the lens. The lens received direct light but not the top of the headset. In the LiViliti device, the UV-C bulbs were positioned across the top and all 4 sides of the device, leading to direct UV-C exposure to both the top and the lens. In the Philips device, the UV-C bulbs were in the top of the device only, leading to direct UV-C exposure to the top of the device instead of the lens. The headset was rotated to provide direct lens exposure. Reflectivity also differed by device; the Cleanbox device has no reflective material to direct UV-C back toward the VR headset, while the LiViliti device has reflective material on the top and bottom and the Philips device has reflective material on all sides (Figure 1).

A UV-C sensor (X1 Optometer, Gigahertz-Optik) measured UV-C dose for each device. Initial UV-C measurements varied by time per device based on the MIFU. Additional measurements were made to compare UV-C doses across devices with 1 minute of exposure. Differences in UV-C doses based on the surface angle compared with the light source were measured by placing the UV-C sensor at different locations within the UV-C chamber of each device (center and all 4 corners) and by placing the UV-C sensor both upright or perpendicular to the top of the device. All dose measurements were performed in triplicate for each location with the average of the 3 noted in Table 1.

XSL•FO
RenderX

Table . Differences in ultraviolet-C light energy in different locations for each disinfection device.

Device (MIFU[a] exposure time) units		Cleanbox[b] (1 minute), mJ/cm^2	LiViliti[b] (3 minutes), mJ/cm^2	Philips[b] (10 minutes), mJ/cm^2
Sensor placement in UV-C[c] box and sensor position in UV-C box				
Flat facing top				
	Each corner (n=12)	3.6 (0.8)	18.6 (0.6)	215.3 (8.7)
	Center (n=3)	9.7 (0.8)	36.6 (1.0)	195.4 (4.7)
Perpendicular facing side				
	Each corner (n=12)	102.3 (33.3)	68.2 (9.8)	34.8 (9.1)
	Center (n=3)	155.8 (1.3)	53.7 (0.7)	28.4 (4.5)

[a]MIFU: manufacturers' instructions for use.

[b]Values within parentheses reflect SD values.

[c]UV-C: ultraviolet-C light.

Bacterial Killing

Bacterial inoculation and killing were performed with the same methodology as previously outlined [3]. Briefly, 3 strains of bacteria, *S aureus* (ATCC 25923), *Staphylococcus epidermidis* (ATCC 12228), and *Pseudomonas aeruginosa* (laboratory strain PAO1), were grown overnight in 3 mL of lysogeny broth (LB), serially diluted, and 10 microliters of bacteria (inoculum range 1.5×10^5 - 5.8×10^7) were inoculated to the lens and the top of a VR headset; the Meta Quest 2 (Meta) (Figure 1). After air drying for 30 minutes, a control culture was taken to account for bacterial death during surface drying, and the VR headsets were subsequently placed in UV-C devices, as shown in Figure 1. UV-C exposure was performed according to the MIFU of each device. Following UV-C disinfection, sites of inoculation were cultured with sterile cotton swabs dipped in LB, swiped multiple times across the site of inoculation, and inoculated to LB agar plates. Plates were incubated overnight at 37 °C with bacteria colony-forming units (CFUs) counted the following day. All experiments were performed independently in triplicate for all conditions. Experiments were performed under ambient temperature and humidity to reflect real-world conditions. This was monitored in the laboratory and is typically 73 - 78 °F with 20% - 35% humidity. Between experiments, the headset was sprayed with a 70% isopropyl alcohol solution to saturation and left to air dry.

Statistical Analysis

The experimental design for the UV-C killing was a 3-factor crossed design with organism count after treatment as the outcome variable. The generalized linear model with negative binomial distribution and logit link was used to compare whether the proportion of trials with observable bacteria count after treatment differed by (1) type of UV-C disinfection box, (2) type of organism, and (3) location. All models included the natural log of bacteria count prior to treatment and used robust standard errors. The experimental design for the energy sensor was a 3-factor model with amount of energy as the outcome variable. The general linear model with normal distribution was used for comparing type of UV-C device, minutes of exposure, and location of the UV-C sensor or VR headset. For all models, pairwise comparisons were performed for UV-C devices. Analyses were performed in SPSS software (version 29; IBM Corp), and statistical significance was set at an α level of .05.

Ethical Considerations

The laboratory experiments were exempt from Yale University Institutional Review Board evaluation, as they satisfied the Yale University 100 CH.9 Clinical Quality Improvement checklist.

Results

UV-C Energy Output Varies Between and Within Different Devices

When each device was run for the time designated in the MIFU, statistically significant differences were observed correlated with the differences in the instructed run times, with the 10-minute exposure of the Philips device providing the most energy (Table 2 and Figure 2). When the sensor was placed perpendicular to the initial measurements, on the bottom of the device but facing one of the side walls, statistical differences in irradiance were again noted across all devices, likely due to the different light source locations in each device relative to the sensor angle and placement (Table 1 and Figure 2). The within-device difference was least in the LiViliti box, consistent with source locations on all 4 walls and the top of the UV-C chamber.

XSL•FO
RenderX

Table . Considerations for future use of ultraviolet-C light devices on virtual reality headsets.

Barrier to appropriate UV-C[a] disinfection	Proposed solution	Limitations
Device fidelity over time	Measure UV-C dosage periodically	Censor accessibility, device life span
Porous material in VR[b] headset	Place disposable barrier over porous material	Added costs, additional step, and risk of contamination during removal
Shadowing (areas not exposed to UV-C)	Complement with UV-C disinfection wipes Reposition to device to ensure direct UV-C exposure	Added costs, added time, and contact-based disinfection still needed

[a]UV-C: ultraviolet-C light.

[b]VR: virtual reality.

Figure 2. Ultraviolet-C light (UV-C) dose (mJ/cm^2) according to UV-C device, location within each device (center [red] vs corner [blue]), and location of sensor (toward the top of the device [No perpendicular] vs toward the side of the device [Yes perpendicular]). UV: ultraviolet.

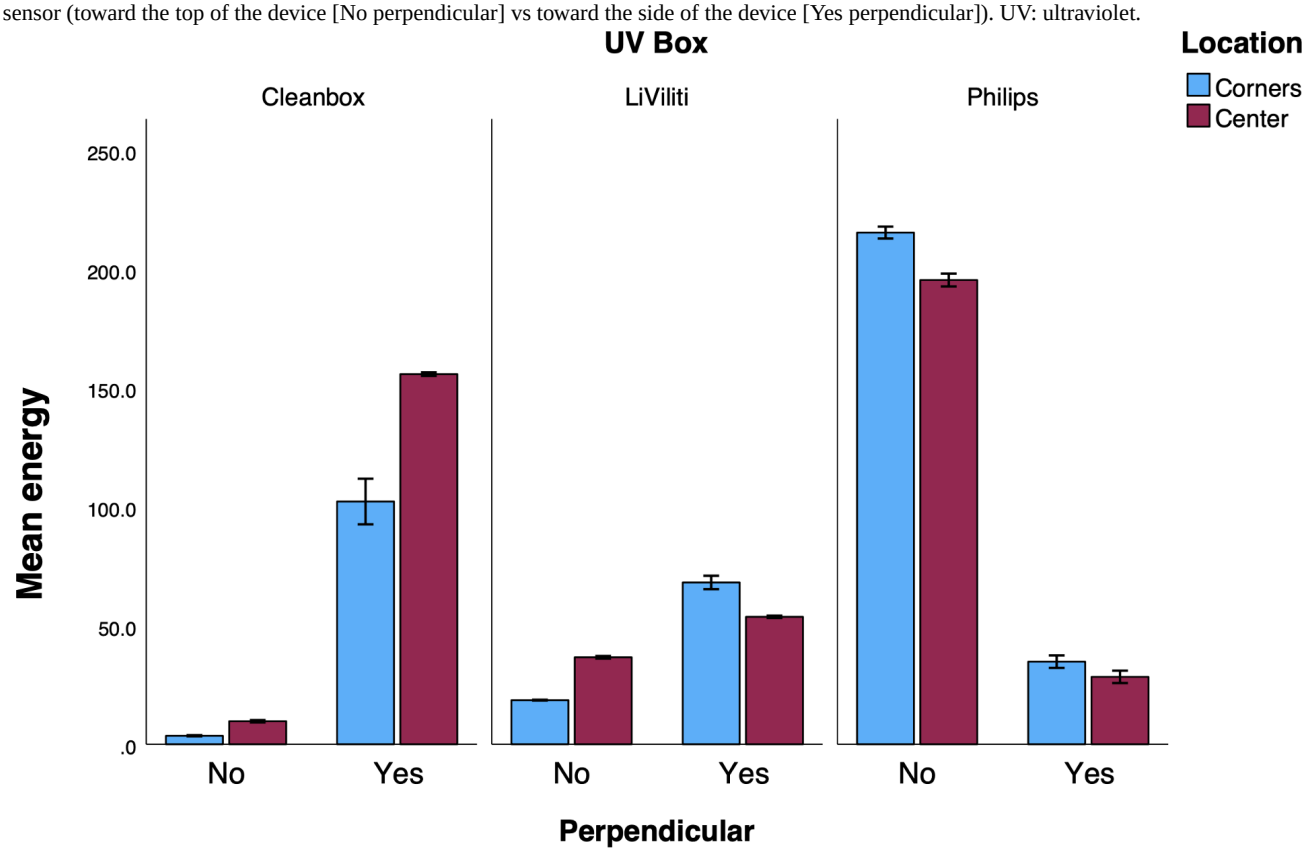

Bacterial Killing by UV-C

There was variability of bacterial killing comparing independent replicates in mean CFU recovery according to organism and UV-C device across all tested conditions, regardless of whether the bacteria were inoculated to the top of the VR headset or the lens (Figure 3). In no experiment was there a mean of zero CFUs recovered when UV-C devices were run according to their MIFU times; there was always bacteria present. Small numbers of bacteria were recovered from some inoculated spots under all conditions tested, likely due to high inocula. However, for *P aeruginosa* and *S aureus*, a minimum of 10^3 killing was demonstrated for all inoculated spots under all conditions tested (Figure 3; additional data not shown). There was more variation for *S epidermidis* with only the 10 minutes of exposure in the Philips device resulting in at least 10^3 bacterial killing for all inoculated sites. For the Cleanbox device, at least 10^3*S epidermidis* killing was observed for all the inoculated lenses, consistent with the direct UV-C exposure, and 80% of spots inoculated on the top where there was indirect exposure to the light source (Figure 3; additional data not shown). For the LiViliti device, UV-C exposure resulted in at least 10^3 killing for 79% of inoculated sites encompassing both the lens and the top (Figure 3).

The same VR headset used in these experiments was also used in a prior study by our group testing disinfection wipes [3]. Following completion of approximately 1860 minutes of UV-C exposure and 176 applications of chemical disinfection wipes composed of isopropyl alcohol and quaternary ammonium compounds, the device retained full functionality and had no grossly visible changes from its initial appearance.

XSL•FO
RenderX

Figure 3. Bacterial growth by organism and UV-C device for the lens and top of each device. The y-axis represents surviving bacteria on a logarithmic scale after UV-C disinfection was performed. UV-C: ultraviolet-C light; VR: virtual reality.

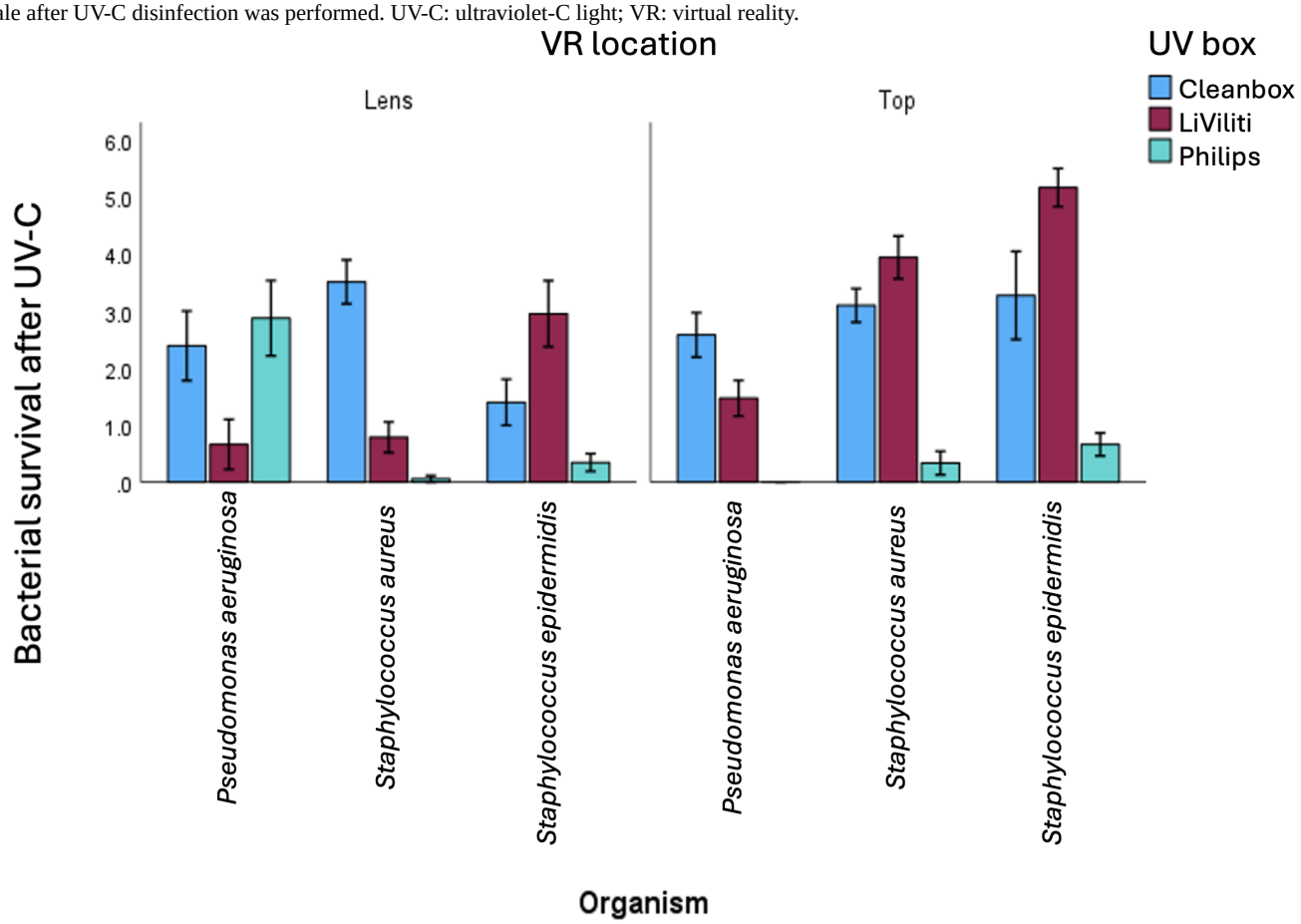

Discussion

Principal Findings

In this in vitro assessment of 3 commercially available, single-use UV-C devices, we found that the recorded dosage varied depending on the location and angle of the surface, as well as MIFU time for each device. As expected, longer exposure times and closer direct surface exposure to the light source resulted in increased energy recordings. The required dosage for effective UV-C killing varies depending on the bacteria, and some degree of logarithmic reduction in bacterial counts was achieved by these devices for most organisms [9]. We attempted to adjust for potential confounders that may have led to variability in bacterial killing across organisms and devices. This included standardizing the methodology of contaminating and quantifying the bacteria on the headsets, determining the locations and positioning of bacteria on the headsets with respect to the UV-C bulb, accounting for initial inoculum counts in our statistical analysis, standardizing the dry time after inoculation prior to UV-C exposure, and including an untreated control in every experiment. Despite these measures, it is possible that other factors may have led to the observed variability in killing, including the artificially high inocula used, biofilm presence, humidity and temperature in the laboratory at the time of experimentation, and operator differences, as several individuals performed these experiments over time as would be expected in the health care setting. Finally, the required UV-C dosage required to kill bacteria may inherently differ by species. For example, predicted UV-C dosage to achieve 1-log killing of *S aureus* on surfaces is greater than that of *P aeruginosa* (66 J/m^2 vs 55 J/m^2) [14,15].

We found that all products reduced bacterial CFUs artificially inoculated to either the plastic top or the lens of VR headsets. In most cases, we observed at least a 10^3 log reduction in bacterial counts after UV-C exposure, but a 10^6 log reduction for high-level disinfection, as suggested by the US FDA and the US EPA [16], was not consistently achieved. Bacteria were recovered in all experiments, although sterilization was not the expectation. In general, the Philips device achieved higher levels of killing, especially for *S epidermidis*, and we hypothesize that this is due to the longer 10-minute MIFU time than the shorter times of the other devices. *S epidermidis* survived better than *S aureus* or *P aeruginosa* for all UV-C devices. The efficacy of UV-C killing of different organisms, including *S epidermidis*, has been shown. However, certain strains of bacteria may harbor variable susceptibilities to UV-C, and resistance after intermittent UV-C exposure has been reported [17,18]. Potential mechanisms of resistance include DNA protection and damage repair, cellular protection by pigments, and reducing photo-oxidative damage [19]. We did not test multiple strains of each organism.

Comparison With Prior Work

These results did not achieve the level of bacterial killing previously shown for contact disinfectant wipes commonly used in health care settings, including isopropyl alcohol and

XSL•FO
RenderX

quaternary ammonium wipes [3]. However, these wipes are contraindicated for use on the lens, and an alternative no-touch complimentary method is necessary. Additionally, outside of gross contamination that should be eliminated via cleaning prior to no-touch disinfection, the real-world quantity of bacteria present on VR headsets after patient use is expected to be well below the tested inocula in this study such that UV-C disinfection could remain a viable option. As an example, stethoscopes in routine use by health care workers cultured prior to disinfection have yielded ranges from 15 to several hundred bacterial CFUs, several logs lower than the quantities used for experimental contamination of the VR headsets in our study [20,21]. Studies have since shown the efficacy of UV-C disinfection on reducing bacterial CFUs on stethoscopes [22].

Strengths and Limitations

UV-C disinfection has potential advantages compared with contact-based chemical disinfection. It results in less environmental waste, is less prone to manual error because it is passive, and it does not contain chemicals that may interact with VR-headset components. Factors that determine the efficacy of UV-C bacterial killing include the wavelength, distance from the UV-C source, and angle of contact with the surface [5]. VR headsets are shaped such that both the distance from the light source and the angle will vary, in addition to likely shadowing. This will result in different parts of the headset, such as the lens and the outside, being exposed to different levels of dosage, which will impact the efficacy of disinfection. Previously, we have shown highly effective disinfection of VR headsets for contact disinfectants commonly used in health care settings, including isopropyl alcohol and quaternary ammonium wipes [3], and were interested in whether UV-C can effectively disinfect VR headsets focusing on a high touch point surface and the lens, a surface not readily amenable to contact disinfection.

UV-C may additionally be a more attractive alternative to disinfectant wipes due to its cost-effectiveness. For example, consumer pricing of a routinely used quaternary ammonium wipe at our facility is approximately 17 cents per wipe. The lowest-priced UV-C device in our study was approximately US $150.00. Thus, cost savings could be achieved after 882 disinfection cycles, or 2.4 cycles per day if used for a year. The life span of the UV-C devices is measured in years, so with sufficient use, health care facilities that anticipate frequent device disinfection may experience significant cost savings. In our study, the most expensive device also had the lowest time per run according to the MIFU, so purchasers could consider which UV-C device to obtain depending on planned disinfection throughput.

The UV-C devices tested in this study, in contrast to disinfectant wipes used at our center, are not approved by the US FDA. However, they are regulated as class II medical devices according to the US Code of Federal Regulations; thus, the manufacturer's instructions for use must be followed [13]. To obtain US FDA approval for use in health care, such devices need to be tested and validated for consistent bacterial killing (such as 10^6 log killing for high-level disinfection) across a wide range of conditions using clinically relevant microorganisms beyond just bacteria. Our in vitro results showing that some bacterial killing is achieved against 3 common health care pathogens are promising, but this should be validated and expanded before such devices could be designated as medical devices by regulatory agencies, which would ensure safer use in patient care and they could be more widely recommended as a viable option.

Despite this, our experiment has found observations that may be considered should US FDA approval be pursued by UV-C device manufacturers. We suggest that users of UV-C devices for VR headset disinfection consider the following: (1) UV-C irradiance production at different locations within the device should be measured routinely to ensure that the device continues to work properly over time. While we performed this using a high-quality meter, irradiance can be less precisely confirmed using commercially available photosensitive papers specific for the measurement of UV-C [23,24]. All 3 devices continued to produce the same irradiance over the study period of approximately 1 year (data not shown). (2) Porous straps continue to require a disposable barrier for the user. This is based on the lesser efficacy of disinfection with porous than with nonporous material [3]. (3) For nonporous areas subject to shadowing, consider complementary disinfection with wipes. (4) Depending on the light source distribution within the device, it may be necessary to change the position of the VR headset in the box to ensure direct UV-C to both the high touch points outside of the VR headset and the lens, which are perpendicular surfaces. Shadowed areas, such as buttons, are less likely to receive sufficient UV-C exposure as exemplified by our sensor measurements, where placing the sensor outside of the direct line of exposure to the UV-C bulb dramatically reduced the UV-C irradiation received. In practical use, should these UV-C devices be US FDA–approved for disinfection in health care settings, patients harboring certain pathogens, such as norovirus and *C difficile*, should be restricted from VR headset use. All of the results noted are performed on clean VR devices, such that there is no gross contamination or debris noted that could impede the bactericidal efficacy of UV-C. Meticulous cleaning of all VR devices should be performed before being subjected to any attempted bactericidal method, or efficacy could be compromised in real-world settings.

We did not test spore-forming bacteria such as *C difficile* that require increased energy levels for killing but previously recommended *C difficile* infection as a contraindication for VR use in health care [3]. Importantly, only the Cleanbox is specifically marketed for UV-C disinfection of VR headsets, while the other 2 devices are not dedicated for this purpose, although none of the 3 devices are US FDA–approved as medical devices to use in health care settings. We also did not measure disinfection of hand controllers. We would expect effective UV-C disinfection with appropriate surface energy levels, although shadowing remains a concern. An additional limitation is that these in vitro findings, while promising, cannot be directly extrapolated to clinical practice as only 3 commonly encountered health care pathogens were studied, and the rigorous standards (such as testing for 10^6 log bacterial killing as required by the US FDA and the US EPA) were not evaluated. Finally, there may have been fluctuations in the temperature and

XSL•FO
RenderX

humidity in the laboratory where testing was performed, which could have impacted bacterial survival, although we attempted to control for this by doing laboratory experiments under a hood in a regulated laboratory space.

Despite these limitations, these data show that UV-C may be an effective method for bacterial killing on VR equipment with low-level contamination. We note that while we observed significant bacterial killing, in no experiment were all bacteria killed when devices were run according to their MIFU, and this work was not designed to test for formal disinfection per regulatory bodies such as the US FDA, which require additional testing before meeting such a standard. Future studies with different devices and microorganisms (we are currently testing UV-C efficacy against bacteriophages and viruses) will help clarify the optimal UV-C methodology for maximizing disinfection of VR devices should additional study to achieve this regulatory standard be pursued.

Acknowledgments

This publication is supported by the Centers for Disease Control and Prevention (CDC) of the US Department of Health and Human Services (HHS) as part of a financial assistance award totaling US $2,542,348 with 100% funding by the CDC/HHS. The contents are those of the authors and do not necessarily represent the official views of, nor an endorsement by, the CDC/HHS or the US Government.

Members of the HIPSTER Team: Eileen Blake, Jamie Trumpler, Noelle Frye, Elin Nilsen, Hyung Paek, Michael Boyce, Cheryl Mayeran, Tony Polcaro, Antonietta Hallet, Lindsey Greene-Upshaw, Nitu Kashyap, David Peaper, Leigh Evans, Veronica Weser, Erin Truesdell, Abigail Crocker, Windy Tanner, Jordan Peccia, Lucien Dieter, Hannah Healy, and Trini Mathew.

Authors' Contributions

SCR, TSM, and RAM participated in conceptualization. JW, MM, and PHH contributed to data curation. RSF did the formal analysis. SCR, JW, MM, AM, KH, and PHH did the investigation. AM, KH, PHH, RAM, and TSM contributed to methodology. SCR and TSM contributed to writing–original draft. SCR, TSM, JW, AM, KH, and RAM contributed to writing–review and editing.

Conflicts of Interest

RAM and TSM receive funds from a different ultraviolet-C light (UV-C) company to conduct UV-C room disinfection studies unrelated to this project and device-related disinfection studies. There is no relationship between any author and the companies making the UV-C devices used in this study. All other authors declare no conflicts of interest relevant to this work.

References

1. Executive summary for the patient engagement advisory committee meeting: augmented reality and virtual reality medical devices. US Food and Drug Administration. 2022. URL: https://www.fda.gov/advisory-committees/advisory-committee-calendar/july-12-13-2022-patient-engagement-advisory-committee-meeting-announcement-07122022 [accessed 2025-09-09]
2. Eyre DW, Sheppard AE, Madder H, et al. A Candida auris outbreak and its control in an intensive care setting. N Engl J Med 2018 Oct 4;379(14):1322-1331. [doi: 10.1056/NEJMoa1714373] [Medline: 30281988]
3. Roberts SC, Havill NL, Flores RM, et al. Disinfection of virtual reality devices in health care settings: in vitro assessment and survey study. J Med Internet Res 2022 Dec 1;24(12):e42332. [doi: 10.2196/42332] [Medline: 36269222]
4. Clean your Meta Quest products. Meta. URL: https://www.meta.com/help/quest/articles/headsets-and-accessories/product-care-and-best-practices/clean-your-meta-quest/ [accessed 2024-02-15]
5. Boyce JM, Donskey CJ. Understanding ultraviolet light surface decontamination in hospital rooms: a primer. Infect Control Hosp Epidemiol 2019 Sep;40(9):1030-1035. [doi: 10.1017/ice.2019.161] [Medline: 31210119]
6. Anderson DJ, Moehring RW, Weber DJ, et al. Effectiveness of targeted enhanced terminal room disinfection on hospital-wide acquisition and infection with multidrug-resistant organisms and Clostridium difficile: a secondary analysis of a multicentre cluster randomised controlled trial with crossover design (BETR Disinfection). Lancet Infect Dis 2018 Aug;18(8):845-853. [doi: 10.1016/S1473-3099(18)30278-0] [Medline: 29880301]
7. de Groot T, Chowdhary A, Meis JF, Voss A. Killing of Candida auris by UV-C: importance of exposure time and distance. Mycoses 2019 May;62(5):408-412. [doi: 10.1111/myc.12903] [Medline: 30748018]
8. Mariita RM, Davis JH, Randive RV. Illuminating human norovirus: a perspective on disinfection of water and surfaces using UVC, norovirus model organisms, and radiation safety considerations. Pathogens 2022 Feb 8;11(2):226. [doi: 10.3390/pathogens11020226] [Medline: 35215169]
9. Chroudi A, Nicolau T, Sahoo N, et al. Predictive modeling of UV-C inactivation of microorganisms in glass, titanium, and polyether ether ketone. Microbiol Res (Pavia) 2024;15(3):1189-1207. [doi: 10.3390/microbiolres15030080]
10. Rutala WA, Gergen MF, Weber DJ. Room decontamination with UV radiation. Infect Control Hosp Epidemiol 2010 Oct;31(10):1025-1029. [doi: 10.1086/656244] [Medline: 20804377]

XSL•FO
RenderX

11. Rutala WA, Gergen MF, Tande BM, Weber DJ. Room decontamination using an ultraviolet-C device with short ultraviolet exposure time. Infect Control Hosp Epidemiol 2014 Aug;35(8):1070-1072. [doi: 10.1086/677149] [Medline: 25026629]
12. Avoid sunlight damage with your Meta Quest headset. Meta. URL: https://www.meta.com/help/quest/articles/headsets-and-accessories/product-care-and-best-practices/avoiding-sunlight-damage [accessed 2024-10-10]
13. § 880.6600 ultraviolet (UV) radiation chamber disinfection device. Code of Federal Regulations. URL: https://www.ecfr.gov/current/title-21/chapter-I/subchapter-H/part-880/subpart-G/section-880.6600 [accessed 2025-05-14]
14. Zelle MR, Hollaender A. Effects of radiation on bacteria. In: Radiation Biology: McGraw-Hill; 1955:365-430 URL: https://catalog.nlm.nih.gov/permalink/01NLM_INST/1o1phhn/alma993532953406676 [accessed 2025-09-09]
15. Gates FL. Results of irradiating Staphylococcus aureus bacteriophage with monochromatic ultraviolet light. J Exp Med 1934 Jul 31;60(2):179-188. [doi: 10.1084/jem.60.2.179] [Medline: 19870293]
16. Rutala WA, Weber DJ. Guideline for disinfection and sterilization in healthcare facilities. The Healthcare Infection Control Practices Advisory Committee (HICPAC).: Centers for Disease Control and Prevention; 2008. URL: https://www.cdc.gov/infection-control/hcp/disinfection-and-sterilization/index.html [accessed 2024-12-10]
17. Molina-Menor E, Carlotto N, Vidal-Verdú À, Pérez-Ferriols A, Pérez-Pastor G, Porcar M. Ecology and resistance to UV light and antibiotics of microbial communities on UV cabins in the dermatology service of a Spanish hospital. Sci Rep 2023 Sep 4;13(1):14547. [doi: 10.1038/s41598-023-40996-8] [Medline: 37666842]
18. Rao BK, Kumar P, Rao S, Gurung B. Bactericidal effect of ultraviolet C (UVC), direct and filtered through transparent plastic, on gram-positive cocci: an in vitro study. Ostomy Wound Manage 2011 Jul;57(7):46-52. [Medline: 21904015]
19. Lamprecht-Grandío M, Cortesão M, Mirete S, et al. Novel genes involved in resistance to both ultraviolet radiation and perchlorate from the metagenomes of hypersaline environments. Front Microbiol 2020;11(453):453. [doi: 10.3389/fmicb.2020.00453] [Medline: 32292392]
20. Lee R, Choi SM, Jo SJ, et al. A quasi-experimental study on stethoscopes contamination with multidrug-resistant bacteria: its role as a vehicle of transmission. PLoS One 2021;16(4):e0250455. [doi: 10.1371/journal.pone.0250455] [Medline: 33886664]
21. Longtin Y, Schneider A, Tschopp C, et al. Contamination of stethoscopes and physicians' hands after a physical examination. Mayo Clin Proc 2014 Mar;89(3):291-299. [doi: 10.1016/j.mayocp.2013.11.016] [Medline: 24582188]
22. Messina G, Burgassi S, Messina D, Montagnani V, Cevenini G. A new UV-LED device for automatic disinfection of stethoscope membranes. Am J Infect Control 2015 Oct 1;43(10):e61-e66. [doi: 10.1016/j.ajic.2015.06.019] [Medline: 26254501]
23. Su A, Grist SM, Geldert A, Gopal A, Herr AE. Quantitative UV-C dose validation with photochromic indicators for informed N95 emergency decontamination. PLoS One 2021;16(1):e0243554. [doi: 10.1371/journal.pone.0243554] [Medline: 33406084]
24. Neu DT, Mead KR, McClelland TL, et al. Surface dosimetry of ultraviolet germicidal irradiation using a colorimetric technique. Ann Work Expo Health 2021 Jun 12;65(5):605-611. [doi: 10.1093/annweh/wxaa147] [Medline: 33616189]

Abbreviations

CFU: colony-forming unit
LB: lysogeny broth
MIFU: manufacturers' instructions for use
US EPA: United States Environmental Protection Agency
US FDA: United States Food & Drug Administration
UV-C: ultraviolet-C light
VR: virtual reality

Edited by J Sarvestan; submitted 10.01.25; peer-reviewed by G Messina, I Armentano; revised version received 06.06.25; accepted 17.06.25; published 30.09.25.

Please cite as:
Roberts SC, Wright J, Mustafa M, Feinn RS, Marks A, Hieftje K, Huang PH, Martinello RA, Murray TS, HIPSTER Team
Bactericidal Efficacy of Ultraviolet-C Light on Virtual Reality Devices: In Vitro Assessment of Bacterial Killing
JMIR Form Res 2025;9:e70326
URL: https://formative.jmir.org/2025/1/e70326
doi:10.2196/70326

© Scott C Roberts, Jayson Wright, Mahnoor Mustafa, Richard S Feinn, Asher Marks, Kimberly Hieftje, Pamela H Huang, Richard A Martinello, Thomas S Murray, HIPSTER team. Originally published in JMIR Formative Research (https://formative.jmir.org),

XSL•FO
RenderX

30.9.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Comparison of Sleep Features Across Smartphone Sensors, Actigraphy, and Diaries Among Young Adults: Longitudinal Observational Study

Jaclyn S Kirshenbaum[1], PhD; Ryann N Crowley[2,3], MS; Melissa D Latham[4], PhD; David Pagliaccio[1,5], PhD; Randy P Auerbach[1,5], PhD; Nicholas B Allen[2,3], PhD

[1]Department of Psychiatry, Columbia University, 1051 Riverside Drive, Pardes 2313, New York, NY, United States
[2]Department of Psychology, University of Oregon, Eugene, OR, United States
[3]Ksana Health Inc, Eugene, OR, United States
[4]VA Northern California Health Care System, Mather, CA, United States
[5]Division of Child and Adolescent Psychiatry, New York State Psychiatric Institute, New York, NY, United States

Corresponding Author:
Jaclyn S Kirshenbaum, PhD
Department of Psychiatry, Columbia University, 1051 Riverside Drive, Pardes 2313, New York, NY, United States

Abstract

Background: Poor sleep health is pervasive and contributes to long-lasting physical and psychological problems. As traditional sleep measurement can be burdensome, testing scalable and accessible sleep measurements is important.

Objective: The aim of this study was to test whether sleep features obtained through a smartphone app are comparable to other modes of sleep measurement (ie, daily diary and wearable actigraphy).

Methods: Healthy college students (n=29, 18‐24 years old) with no prior diagnosis of a sleep disorder consented to downloading a smartphone app, the Effortless Assessment Research System (EARS) (Center for Digital Mental Health, University of Oregon; Ksana Health). For 1 week, the EARS app collected data continuously using the phone's accelerometer, gyroscope, and exposure to light. Each morning, participants received a notification via EARS to complete a sleep diary, which asked participants what time they got into bed, fell asleep, woke up, and got out of bed. A random subset (n=13) of participants also consented to wear an ActiGraph wristwatch. All analyses examined bedtime (ie, time going to bed), risetime (ie, wake-up time), and time-in-bed (ie, duration between time-in-bed and out-of-bed) as 3 measures of interest. For all analyses, diary data were considered the reference measurement, and analyses were repeated with ActiGraph as the reference measurement.

Results: On average, EARS showed a high mean true positive rate (86.6%) and low mean false positive rate (4%) based on diary-reported bedtime and risetime. Supplementary analyses comparing EARS to ActiGraph data showed a similar true positive rate (83.7%) and slightly higher false positive rate (8.5%). Although there were no significant differences in mean bedtime, risetime, and time-in-bed among sources ($P \geq .07$), there was some misalignment. Compared to the diaries, EARS estimated bedtime to be later by an average of 20 minutes, risetime to be earlier by an average of 21.83 minutes, and time-in-bed to be shorter by an average of 41.82 minutes. Relative to ActiGraph estimates, EARS estimated bedtime to be earlier by an average of 21.87 minutes, but estimated risetime to be later by 2.5 minutes, and time-in-bed to be longer by 24.4 minutes. Day-to-day correlations showed that bedtimes, risetimes, and time-in-bed were positively correlated between diary and EARS ($0.29 \leq r \leq 0.52$, $P \leq .002$). Similarly, day-to-day bedtimes and time-in-bed were positively correlated between ActiGraph and EARS ($0.38 \leq r \leq 0.55$, $P \leq .01$), and while correlated in the expected direction, risetimes were not significantly associated between ActiGraph and EARS ($r=0.29$, $P=.07$).

Conclusions: Smartphone-based sleep sensors show acceptable alignment with more established methods and may provide a feasible alternative to measuring daily sleep patterns in a scalable way. Future studies will require larger, diverse samples to corroborate findings of concordance among EARS, diary, and actigraphy data in other populations.

(JMIR Form Res 2025;9:e67455) doi:10.2196/67455

KEYWORDS

mobile sensors; accelerometer; wearables; daily diary; sleep measurement

XSL•FO
RenderX

Introduction

Insufficient or disrupted sleep is ubiquitous in youth and young adults. Specifically, more than 60% of college students show delayed bedtimes and wake-up times [1], which often lead to poor academic performance [2], physical and psychological health issues [3], substance use [1], depression [4], and suicidal thoughts and behaviors [5]. Given pervasive sleep alterations during the college transition, it is important to develop scalable methods to measure sleep health.

In research, daily sleep diaries (eg, via ecological sampling methodology) are often used as a primary mode of measurement of subjective sleep given their low cost and ease of administration. However, reporting on sleep every morning over extended periods of time often results in low response rates, especially in the morning relative to other times of day [6,7]. The gold standard for objectively identifying sleep problems is polysomnography, which measures brain, cardiovascular, and respiratory signals. Yet, conducting a polysomnography study can be burdensome on participants as it usually requires at least 2 sessions sleeping in a lab with intrusive equipment [8] and thus may not reflect typical sleep in the home environment. Actigraphy, typically with a wrist monitor, can measure several metrics of sleep based on motion (ie, accelerometer) and most measure ambient light. Some studies ask participants to actively press a button on the watch as an event marker indicating when they are going to bed and when they wake up. Though actigraphy is a relatively unobtrusive measurement of natural at-home sleep and has been validated against polysomnography [9], there are several challenges with implementation, including compliance (eg, adherence to wearing the watch) and battery life. Specifically, the maximum battery life for most research-grade wearables is up to ~4 weeks, which limits longer-term phenotyping. By comparison, smartphone sensors (ie, mobile accelerometry) are already integrated into participants' lives and therefore provide a naturalistic, unobtrusive, and scalable way to monitor daily sleep health (ie, onset and duration) in real-time over longer time periods without requiring participants to integrate an extra apparatus, such as a wearable.

Smartphone sensor sleep measurement is a nascent research method that is highly scalable and has enormous potential to reach populations such as those with limited means to purchase wearables, live in difficult-to-reach places, or for whom compliance with a wearable might be low (eg, adolescents). However, there can be several limitations of leveraging smartphone technologies that were not specifically designed to measure sleep. These include difficulty detecting daytime naps and distinguishing "going-to-bed" from "going-to-sleep" and "getting-out-of-bed" from "waking up." Nonetheless, there have been several investigations of the validity of smartphone sensor sleep measurement. In adults, across 4 smartphone apps (Sleep Cycle-Accelerometer, Sleep Cycle-Microphone, Sense, and Smart Alarm), time-in-bed estimates were significantly correlated across app and polysomnography measurements; however, other sleep features (eg, sleep efficiency and sleep staging) significantly differed from polysomnography estimates [10]. In addition, a comparison between actigraphy and the TapCounter smartphone app (QuantActions) showed a strong correlation for sleep onset and wake-up times [11] and a study of college students showed that average sleep duration (using the mindLAMP app; The Division of Digital Psychiatry at Beth Israel Deaconess Medical Center) was significantly associated with daily self-reported sleep duration and quality [12]. Together, this evidence indicates that smartphone sleep measurements may serve as a reliable proxy for estimating bedtime, risetime, and time-in-bed among healthy young adults.

This study tested whether the Effortless Assessment Research System (EARS) smartphone app [13] was a reliable method for detecting bedtime, risetime, and time-in-bed over 1 week in a sample of college undergraduates. We simultaneously collected daily diaries for self-reported sleep features. In a random subset of the sample, we also collected actigraphy via a wrist wearable, which is used as a supplementary method to compare EARS-obtained sleep features with another established research method.

Methods

Current Study

This study analyzed data from a broader intervention project on regularizing sleep timing [14]. We report how we determined our sample size, data exclusions, and the relevant measures for this analysis.

Recruitment

Participants were recruited from the University of Oregon using the undergraduate psychology research participant pool, classroom announcements, and printed flyers publicly posted on campus. Participants were eligible if they were 18 - 24 years old, owned an Android smartphone, had no prior diagnosis of a sleep disorder, and reported greater than 2 hours of day-to-day variability between their earliest and latest wake-up times during the past week. The latter criterion was related to a broader intervention project on regularizing sleep timing. Thirty-seven individuals were eligible and consented to participate. Two participants' phones were incompatible with the EARS app, and 1 student withdrew from participation, resulting in a total of 34 participants who completed the study.

Procedure

This study analyzed data from the baseline week of a 5-week microrandomized trial, as described in Latham [14]. At the consent session, participants completed questionnaires and the beta version of the EARS phone app was installed on participants' phones. A random subset (n=18) of participants also consented to wear an ActiGraph wristwatch (ActiGraph wGT3x-BT triaxial accelerometer; ActiGraph LLC) for the week. The ActiGraph was only worn by a subset of the participants because of the limited availability of these wrist wearables.

Power Analysis

An a priori power analysis was conducted for the broader microrandomized trial using the *MRTSampleSize* R package [15]. Thirty-three participants were required to predict the

XSL•FO
RenderX

estimated effect size of the broader study (β=.3) with 80% power [14].

Self-report

The Pittsburgh Sleep Quality Index (PSQI) [16] is a validated self-reported questionnaire that probes individuals' typical bedtime, risetime, hours of "actual sleep," any trouble sleeping, and sleep quality over the last month. There are 7 components, or subscales, that when combined yield a total "sleep quality" score ranging from 0 - to 21. Higher scores indicate worse sleep quality [16].

EARS

The EARS app was developed by the Adolescent Development and Psychopathology Team (ADAPT) at the University of Oregon to passively collect data with minimal burden to participants [13,17]. This project used the smartphone accelerometer and gyroscope data to estimate bedtime, risetime, and time-in-bed. Participants were instructed to use their phones as they normally would. They were also told to make sure that their phones do not run out of battery and to make sure there is sufficient memory storage. EARS collected data continuously, including detection of when participants' phones were charging, moving (using the phone's accelerometer and gyroscope), and exposed to light in lumens. EARS applied a lux threshold, such that a new lumen capture only would be recorded when the detected lumen value changed by a magnitude of 3 from the previously recorded value. Accelerometer data were collected at 10 Hz sampling rate, filtering out any changes <0.05 m/s^2. Gyroscope data were collected at a 10 Hz sampling rate, filtering out any changes <0.01 radians per second. Data were timestamped at the level of milliseconds and then aggregated into 10-second periods. For this study, the data capture schedule for the EARS app was limited to data acquisition between 8:30 PM and 10:39 AM to reduce the burden on the device. When a sleep offset time was not detected within the available data, the algorithm defaulted to using the last available data point, which truncated times to 10:39 AM. Participants' EARS data were encrypted and stored using Amazon Web Services [17].

Daily Sleep Diary

Each morning, participants received a notification via EARS on their smartphone to complete a sleep diary. The daily sleep diary was based on the Pittsburgh Consensus Sleep Diary [18], and participants provided what time they got into bed (ie, "What time did you get into bed last night?"), fell asleep (ie, "What time did you fall asleep last night?"), woke up ("What time was your final awakening?"), and got out of bed (ie, "What time did you get out of bed this morning?"). Note that the EARS app went through minor changes during the study period, mainly to remove bugs.

Actigraphy

In a supplementary part of the study, 18 participants were randomly selected and consented to wear an ActiGraph wristwatch (ActiGraph wGT3x-BT triaxial accelerometer; ActiGraph LLC) for the baseline week of the study. The Cole-Kripke algorithm was used to score sleep-wake periods [19], and the Tudor-Locke algorithm was used to detect periods of time-in-bed. We used the "In Bedtime" and "Out Bedtime" to reflect bedtime (ie, time going to bed) and risetimes (ie, wake-up time), respectively.

Out of the 34 participants who completed the study, 32 participants had usable data from at least 2 concurrent sources measured on the same day, 29 of whom had concurrent diary and EARS data, and 13 of whom also had ActiGraph data.

Data Quality Assurance

Although actigraphy should provide 1 sleep period (excluding naps) per day, 11 participants with ActiGraph data showed more than 1 sleep period on certain days. Within each of these participants and days, we identified the earliest and latest time (ie, the longest duration) to determine the most likely sleep period and then removed entries that appeared to be midday naps given that the EARS app is not yet able to detect midday naps. We also inspected outliers upon plotting distributions of bedtime, risetime, and time-in-bed for all sources of data, which removed 10 observations that were 3 SDs above or below the mean.

Statistical Analysis

All analyses examined bedtime (ie, time going to bed), risetime (ie, wake-up time), and time-in-bed (ie, duration between time in-bed and out-of-bed) as 3 measures of interest. For all analyses, diary data were considered the reference measurement given that (1) all participants were administered daily diaries, (2) the questions were derived from validated measures, and (3) diaries are more commonly used relative to the ActiGraph wearable used in this study and the EARS smartphone app. However, we also repeated analyses with ActiGraph as the reference measurement given that this wearable obtains accelerometry data. In addition, given that participants also self-reported onset, offset, and duration of sleep (separate from bed timings), we compared EARS estimates to these diary features in supplementary analyses.

First, we computed descriptive statistics for each source of measurement (diary, ActiGraph, and EARS). Second, we assessed the group-level true positive rates (TPR) and false positive rates (FPR) of EARS data based on minute-by-minute classification of "true" and "false" in-bed epochs. For this classification, each minute was counted as a true in-bed epoch until the next out-of-bed period for each particular data source. For the purpose of this analysis, we used the diary as the reference source. Thus, if an epoch was an in-bed period for both diary and EARS, then that minute was classified as a true in-bed epoch for EARS. Similarly, if an epoch was an out-of-bed period for both diary and EARS, then that minute was counted as a true out-of-bed period for EARS. If there was a difference or mismatch across sources (eg, diary indicated in-bed and EARS indicated out-of-bed), then that minute would be counted as a false in-bed or out-of-bed epoch for EARS. Thus, each minute of the day (1440 min) was denoted as either a true in-bed, false in-bed, true out-of-bed, or false out-of-bed period, and then aggregated at the person level (ie, for each person, we summed the true in-bed epochs, false in-bed epochs, true out-of-bed epochs, and false out-of-bed epochs). EARS TPR was computed as true in-bed epochs divided by the sum of true

XSL•FO
RenderX

in-bed and false out-of-bed epochs. EARS FPR was computed as false in-bed periods divided by the sum of false in-bed and true out-of-bed periods. EARS TPR and FPR calculations were repeated based on the truncated data capture period (8:30 PM-10:39 AM; 849 min). Third, we used Bland–Altman plots (BlandAltmanLeh R package [20]) to visualize agreement between sources, where the difference between 2 sources of data is plotted against the mean of these sources. A positive value of the mean difference between diary and EARS data (Diary–EARS) indicates that EARS yielded earlier times relative to diary estimates, and a negative value means that EARS yielded later times relative to diary estimates. Next, using linear regression models, we examined whether average bedtime, risetime, and time-in-bed over the 7 days (ie, person-level) were significantly different among sources, using diary as the reference. We repeated these models with ActiGraph as the reference. Finally, we computed repeated-measures correlations between sources' bedtime, risetime, and time-in-bed.

Ethical Considerations

All procedures were reviewed and approved by the University of Oregon Institutional Review Board (# 08142018.013). Informed consent forms stated that the researchers planned to publish and present the results of this research and that identifying information would not be used in any published reports, conference presentations, or other presentations. Identifiable information was stored in a password-protected database that was only accessible by the researchers for the study. All other data provided were kept in separate databases. All data that the phone app collected were encrypted on the phone and kept in secure, encrypted cloud-based storage. Participants were either paid US $100 or awarded 4 credits for completing the broader intervention project.

Results

Descriptive statistics

Participant demographics are presented in Table 1. Descriptive statistics for bedtime, risetime, and time-in-bed per data source (ie, diary, EARS, and ActiGraph) over the 1-week data collection period are displayed in Table 2, Figure 1, and Figure S1 in Multimedia Appendix 1. The PSQI demonstrated low internal consistency (α=.66) in this sample, but was considered for exploratory correlational analyses.

Table . Participant characteristics (n=29). Choices for biological sex were male or female.

Variable	Values
Age (years), mean (SD)	20.93 (1.86)
Sex (Female), n (%)	14 (56)
Race, n (%)	
African American or Black	2 (6.9)
Asian	3 (10.34)
More than one race	2 (6.9)
White	20 (68.97)
Unknown or Not reported	2 (6.9)
Ethnicity, n (%)	
Hispanic	1 (3.45)
Non-Hispanic	27 (93.1)
Unknown or Not reported	1 (3.45)
Number of completed days, mean (SD)	5.93 (1.69)

Table . Descriptives of bedtime, risetime, and time-in-bed.[a]

Source	Bedtime, mean (SD)	Bedtime, median (IQR)	Sleep onset, mean (SD)	Sleep onset, median (IQR)	Risetime, mean (SD)	Risetime, median (IQR)	Time in bed, mean (SD)	Time in bed, median (IQR)
Diary	–0.91 (82.64)	0 (–210 to 216)	43.37 (83.57)	30 (–180 to 300)	529.84 (97.29)	510 (300 - 795)	530.74 (100.72)	513 (270 - 795)
EARS	24.53 (101.27)	16.26 (–190.80 to 384.38)	—[b]	—	514.45 (101.24)	507.11 (147.2 - 639.02)	489.93 (130.95)	486.03 (179.95 - 829.82)
ActiGraph	35.93 (104.57)	37.50 (–306 to 310)	—	—	496.73 (139.22)	480.50 (76-914)	460.80 (156.41)	444.5 (165 - 1035)

[a]Bedtime and risetime are represented by minutes from midnight (eg, mean diary bedtime of −0.91 corresponds to 11:59:54PM). Diary reports probed both bedtimes and sleep onset times (eg, time getting into bed and time falling asleep, respectively).

[b]Not applicable.

XSL•FO
RenderX

Figure 1. Distribution of mean bedtimes (solid lines): ActiGraph = 12:35:55 AM; Diary =11:59:54 PM; EARS=12:24:31 AM; Mean risetime (dotted lines): ActiGraph = 8:16:43 AM; Diary = 8:49:50 AM; EARS = 8:34:27 AM). The solid vertical line indicates the time at which EARS truncated data collection at 10:39 AM.

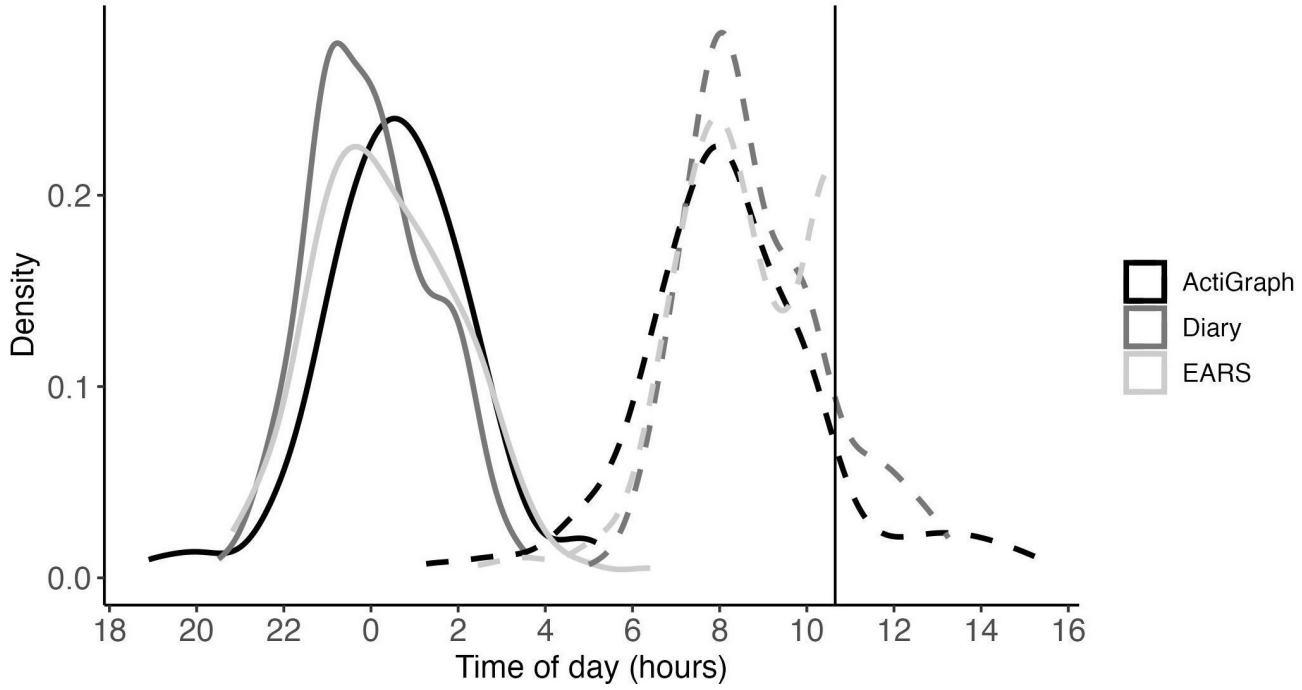

True Positive and False Positive Rates

On average (at the group level), EARS showed a high mean TPR (86.6%, range=31.9% - 100%) and low mean FPR (4.0%, range=0.10% - 20.2%) based on diary-reported bedtime and risetime. Limiting the minute-by-minute data to the truncated data capture period for EARS (849 min per day) yielded a similar TPR (86.6%, range=31.9% - 100%) and slightly higher mean FPR (10.6%, range=0.24% - 48.6%). The TPR and FPR for EARS in-bed and out-of-bed epochs based on diary-reported sleep onset and offset show similar estimates (TPR=89.5%, range=36.3% - 100%; FPR=6.8%, range=0.80%-21.5%). Supplementary analyses comparing EARS to ActiGraph data showed a similar TPR (83.7%, range=33.1% - 100%) and similar FPR (8.5%, range=0% - 21.6%).

Bland–Altman Agreement

The majority of data showing the association between the mean of measurements (eg, averaged EARS and diary time-in-bed) and the difference between measurements (diary time-in-bed – EARS time-in-bed) were within the limits of agreement with a mean difference [SD 1.96]; Figure 2), which is in line with recommendations for adequate agreement of measures [21]. Of note, the CIs are large, indicating small sample size and relatively large dispersion of differences. Although the mean of EARS and diary measurements correlated with the difference between EARS and diary measurements for bedtime and time-in-bed ($P \leq .01$), indicating a presence of proportional bias, there was no correlation present for risetime (P=.80). In addition, the average of EARS and ActiGraph measurements did not correlate with the difference between EARS and ActiGraph measurements for any of the sleep features ($P \geq .20$).

XSL•FO
RenderX

Figure 2. Bland-Altman plots. The x-axis represents the mean (in minutes) of 2 measurements (eg, EARS [Effortless Assessment Research System] and diary) and the y-axis represents the difference (in minutes) between the measurements. The gray solid line represents the mean difference in minutes between sources of measurement. The black solid horizontal lines indicate +1.96 SD and −1.96 SD from the mean. The black dotted lines are the CIs surrounding the mean and SD.

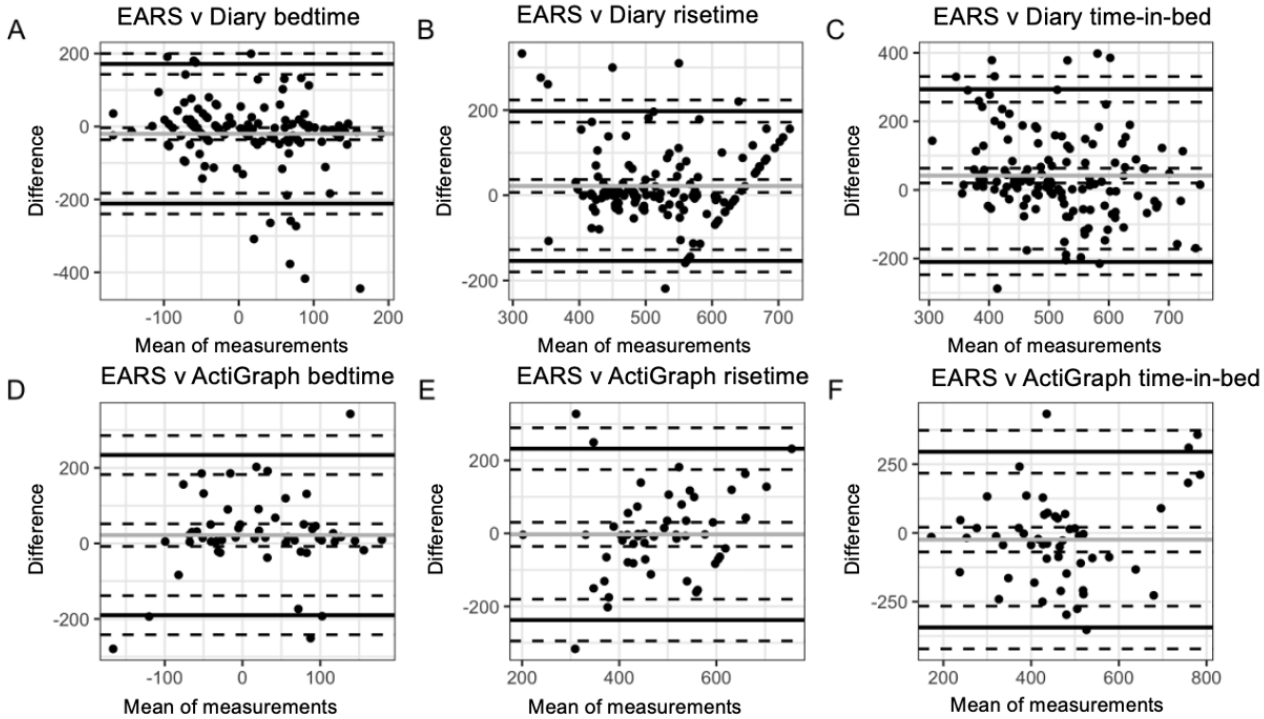

Person-level Differences in Sleep Estimates

There were no significant differences in mean bedtime, risetime, and time-in-bed between sources ($P \geq .07$; Tables 3-5); however, there was some misalignment between EARS measurements relative to diary estimates. Compared to the diaries, EARS estimated bedtime to be later by an average of 20 minutes (SD 97.65), risetime to be earlier by an average of 21.83 minutes (SD 89.68), and time-in-bed to be shorter by an average of 41.82 minutes (SD 128.46) (Figure S2 in Multimedia Appendix 1). Relative to ActiGraph estimates, EARS estimated bedtime to be earlier by an average of 21.87 minutes (SD 108.15), but estimated risetime to be later by 2.5 minutes (SD 119.9) and time-in-bed to be longer by 24.4 minutes (SD 163.1) (Figure S3 in Multimedia Appendix 1). In supplementary analyses, there were no significant differences between EARS bedtime and diary sleep onset, EARS risetime and diary sleep offset, and EARS time-in-bed and diary sleep duration (Table S1 in Multimedia Appendix 1, $P \geq .07$). Relative to the diaries, EARS estimated bedtime to be earlier than diary sleep onset by an average of 27.68 minutes (SD 88.89), risetime to be earlier than sleep offset by an average of 2.91 minutes (SD 91.38), and time-in-bed to be longer than sleep duration by an average of 24.77 minutes (SD 126.51) (Figure S4 in Multimedia Appendix 1). Individual plots showing overlap of bedtime and risetime across sources are shown in Figure S5 in Multimedia Appendix 1.

XSL•FO
RenderX

Table . Comparing Effortless Assessment Research System and diary bedtime, risetime, and time-in-bed (n=29). The intercept reflects Diary as the reference group.

Sleep features	b	95% CI	*t* test[a]	*P* value
Bedtime				
Intercept	−0.73	−23.94 to 22.47	−0.06	0.95
Source (EARS)[b]	18.38	−14.44 to 51.19	1.12	0.27
Risetime				
Intercept	534.97	507.96 to 561.98	39.67	<.001
Source (EARS)	−8.17	−46.37 to 30.03	−0.43	0.67
Time-in-bed				
Intercept	535.7	[504.02 to 567.38]	33.88	<.001
Source (EARS)	−26.55	[−71.35 to 18.25]	−1.19	0.24

[a]Degrees of freedom (df)=56.

[b]EARS: Effortless Assessment Research System.

Table . Comparing Effortless Assessment Research System and diary bedtime, risetime, and time-in-bed to Actigraph bedtime, risetime, and time-in-bed (n=13). The intercept reflects ActiGraph as the reference group.

Sleep features	b	95% CI	*t* test[a]	*P* value
Bedtime				
Intercept	37.12	5.66 to 68.57	2.38	0.02
Source (Diary)	−41.14	−85.62 to 3.35	−1.86	0.07
Source (EARS)[b]	−42.20	−88.24 to 3.84	−1.85	0.07
Risetime				
Intercept	502.77	459.01 to 546.52	23.17	<.001
Source (Diary)	8.69	−53.19 to 70.58	0.28	0.78
Source (EARS)	12.77	−51.28 to 76.83	0.40	0.69
Time-in-bed				
Intercept	465.65	416.25 to 515.05	19.01	<.001
Source (Diary)	49.83	−19.97 to 119.63	1.44	0.16
Source (EARS)	54.97	−17.27 to 127.22	1.53	0.13

[a] Degrees of freedom (*df*)=43.

[b]EARS: Effortless Assessment Research System.

XSL•FO
RenderX

Table . Comparing Effortless Assessment Research System and ActiGraph bedtime, risetime, and time-in-bed to diary sleep features (n=13). The intercept reflects Diary as the reference group. The intercept is different compared to the intercept in Table 3 because this table represents a subset of those individuals who wore the ActiGraph.

Sleep features	b	95% CI	*t* test[a]	*P* value
Bedtime				
Intercept	−4.02	−35.47 to 27.44	−0.26	0.80
Source (ActiGraph)	41.14	−3.35 to 85.62	1.86	0.07
Source (EARS)[b]	−1.07	−47.11 to 44.98	−0.05	0.96
Risetime				
Intercept	511.46	467.70 to 555.22	23.57	<.001
Source (ActiGraph)	−8.69	−70.58 to 53.19	−0.28	0.78
Source (EARS)	4.08	−59.98 to 68.13	0.13	0.90
Time-in-bed				
Intercept	515.48	466.12 to 564.83	21.06	<.001
Source (ActiGraph)	−49.83	−119.63 to 19.97	−1.44	0.16
Source (EARS)	5.15	−67.10 to 77.39	0.14	0.89

[a] Degrees of freedom (*df*)=43.

[b]EARS: Effortless Assessment Research System.

Cross-modality Correlations

As shown in Figure 3, day-to-day bedtimes, risetimes, and time-in-bed were significantly positively correlated between diary and EARS (0.52≤r≥0.29, *P*≤.002). Similarly, day-to-day bedtimes and time-in-bed were significantly positively correlated between ActiGraph and EARS (0.55≤r≥0.38, *P*≤.01), and while trending in the expected direction, risetimes were not correlated between ActiGraph and EARS (*r*=0.29, *P*=.07). EARS-based bedtimes, risetimes, and time-in-bed were also significantly correlated with diary sleep onset, offset, and sleep duration (0.48≤ r≥0.44, *P*≤.001; Figure S6 in Multimedia Appendix 1). As an exploratory analysis, we correlated person-level EARS and PSQI bedtime, risetime, and time-in-bed and found no significant correlation between these sources of measurement (0.20≤ r≥−0.001, *P*≥.31).

Figure 3. Correlations among diary, Effortless Assessment Research System, and ActiGraph sleep features. The top panel depicts associations between Diary and Effortless Assessment Research System, and the bottom panel shows associations between ActiGraph and Effortless Assessment Research System data. The gray dashed line in plots B and E indicates the time at which EARS truncated data collection at 639 minutes from midnight. EARS: Effortless Assessment Research System.

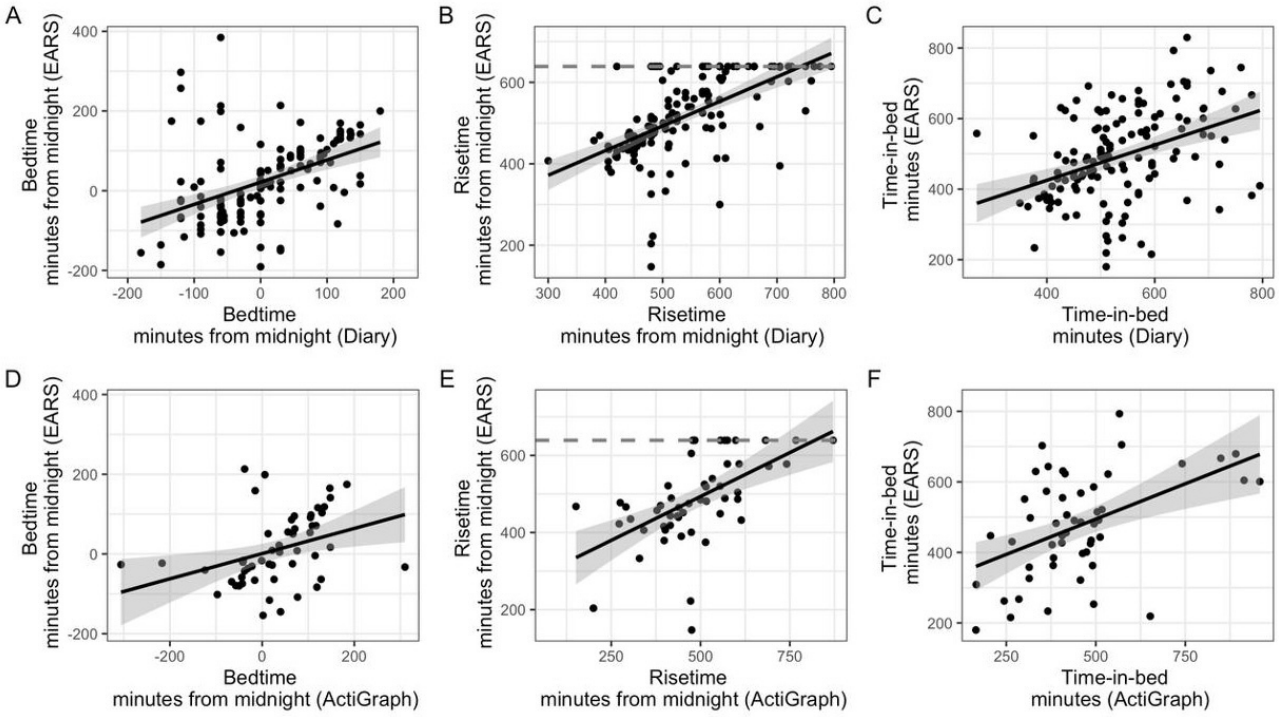

XSL•FO
RenderX

Discussion

Principal Findings

This study tested the concordance among smartphone-based, diary, and wearable (ie, actigraphy) sources of data to measure common metrics of sleep health. We found that when comparing diary and EARS in-bed and out-of-bed periods across 24 hours, EARS yielded a high TPR and low FPR. Given the EARS data capture was scheduled for 849 minutes of the day, we repeated these analyses, finding a similar TPR, but slightly higher FPR which could indicate that motion-based sleep measurement is a slightly more sensitive measure than diary-reported measurement, particularly if most people have their phones with them in-bed or near the bed. Notably, EARS TPR and FPR based on ActiGraph in-bed and out-of-bed epochs were similar to estimates yielded from diary and EARS alignment. In addition, Bland–Altman visualizations revealed adequate agreement of measures. However, upon inspection of the correlation between the mean of measures and the difference between measures, we found evidence of proportional bias specifically for EARS bedtimes and time-in-bed relative to diary bedtimes and time-in-bed, which could indicate that EARS measurement error varies with bedtimes and time spent in bed. Alternatively, as this bias was not present when comparing EARS and ActiGraph estimates for any of the sleep features, the correlation may indicate misestimation from diary data.

On average, there was no significant difference between sources in bedtime, risetime, or time-in-bed estimates. In addition, there was no significant difference between diary-reported sleep onset, offset, and duration and EARS-based bedtime, risetime, and time-in-bed; however, there was some misalignment. EARS tended to estimate bedtime to be later compared to diary-reported bedtime, but earlier relative to ActiGraph bedtime and diary-reported sleep onset. This misalignment could reflect people putting their phones down later compared to when they report initially going to bed. In addition, EARS tended to estimate earlier risetimes relative to diary-reported risetimes and diary-reported sleep offset times, but estimated slightly later risetimes (~2.5 min) relative to ActiGraph risetimes. It is likely that people are picking up their phones in bed when EARS detects “risetime” prior to getting out of bed, which is when they are self-reporting their risetimes. EARS also estimated a shorter time-in-bed compared to diary-reported time-in-bed, which is a result of EARS estimating a later bedtime and earlier risetime. Taken together, it is not surprising that different forms of assessments (self-report, wrist accelerometry, and smartphone sensors) yield discrepant estimates in bedtime, risetime, and time-in-bed as each of these methods relies on different environmental and internal factors (eg, perception of sleep when self-reporting, light and movement for accelerometry, and usage of smartphones during bedtime and risetime).

Repeated correlation analyses revealed that EARS bedtime, risetime, and time-in-bed estimates were significantly correlated with daily diary and actigraphy measurements, except for the correlation between ActiGraph and EARS risetimes, which trended in the expected direction. It is possible that the low number of participants completing the ActiGraph portion of the study diminished our ability to detect a stronger correlation. In addition, the association between PSQI estimates and EARS estimates of bedtime, risetime, and time-in-bed was not associated. There are several reasons that could account for this lack of association, including the difference in acquisition of sleep estimates (ie, self-report with the PSQI compared to passive sensing with EARS) and the timescale of reporting. First, self-reported sleep estimates may be affected by perceived sleep quality and passive sensing estimates may be affected by individual differences in phone usage. Second, the PSQI aims to estimate sleep features over the course of the past month, whereas EARS obtains estimates of sleep features in real time. However, over the course of a week, we found that EARS yielded sleep health estimates that were adequately concordant with self-reported daily diary and ActiGraph estimates in healthy young adults.

Research regarding smartphone-based sleep health is still emerging. Our findings are consistent with other research on college students showing alignment between smartphone-based and self-reported sleep duration [12], onset, and wake-up times [22]. Other healthy adult samples show associations between smartphone and actigraphy onset and wake-up times [11], and total sleep time [23].

Although we show adequate concordance among bedtimes, risetimes, and duration of time-in-bed in this young adult college sample, prior research in young adult samples has shown misalignment in other sleep features (eg, sleep efficiency, sleep latency, and deep sleep) obtained through mobile apps compared to polysomnography, except for time-in-bed [10]. It is possible that smartphone apps may not yet have the capability to accurately measure complex sleep features (eg, efficiency, latency, and sleep cycle stages). Our findings also conflict with research of younger developmental age groups with disordered sleep. Children with sleep-disordered breathing showed significant differences between smartphone-based sleep metrics (sleep onset latency, total sleep time, wake after sleep onset, and sleep efficiency) and polysomnography [24]. Therefore, mobile sensors may be limited in detecting subtle sleep movements, particularly in populations with disordered sleep.

There are several limitations to note. First, although there is concordance among EARS, diary, and ActiGraph sleep features, the small sample size makes it challenging to robustly determine the reliability and validity of EARS. Second, our sample was healthy, enrolled in college, mostly White, and female, which may not generalize our findings to other populations. In particular, college students may experience daily irregularities in their sleep schedule, as many are experiencing autonomy for the first time and balancing academic and social needs [25]. Finally, we assessed sleep over a week, which limits our ability to determine whether alignment of mobile sensors, diary, and actigraphy potentially differs over longer periods of time.

Conclusions

Taken together, smartphone sensors appear to be reliable at detecting basic features of sleep health (ie, bedtime, risetime, and time-in-bed), although whether the level of reliability is adequate will depend on the use case. One of the potential uses of smartphone data is to collect data on sleep at larger scales,

XSL•FO
RenderX

and for such apps, the tradeoff between accuracy and scalability may be justified. For example, smartphone methods may be less appropriate in applicants where very precise measurements of sleep timing are critical. Furthermore, although these data may be informative for detecting alterations in sleep behavior that may vary within individuals over time, smartphone methods are likely to be limited in detecting other features that require increased sensitivity to subtle movements (although these were not examined here). Although this feasibility study shows promise in the alignment across passive (smartphone accelerometer and wrist accelerometer) and self-report (daily diary) measurements, objective and subjective sleep measurements can often differ. Several considerations, including sample size, age range of population, cost, clinical acuity of the population, and goal of the study (eg, perceived sleep and objective sleep) are factors that may impact the choice of sleep measurement. Future studies will require larger and more diverse samples to corroborate our findings of concordance among EARS, diary, and actigraphy data. As noted above, given the feasibility of using smartphones in the general population, our findings show preliminary evidence for using mobile sensors as a scalable method to detect sleep health behaviors.

Acknowledgments

Internal funding from the University of Oregon was used to support this research. Artificial intelligence was not used in any portion of the manuscript.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to participant confidentiality described in the informed consent form, but are available from the corresponding author on reasonable request.

Conflicts of Interest

NBA and RNC have an equity interest in Ksana Health, a company NBA co-founded and which holds the commercial license for certain versions of the Effortless Assessment Research System (EARS) smartphone app and some related EARS tools. RPA is an unpaid scientific advisor for Ksana Health. In the past 3 years, Dr. Auerbach has received consulting fees and equity from Get Sonar Inc. He also has received consulting fees from RPA Health Consulting, Inc. and Covington & Burling LLP, which is representing a social media company in litigation.

Multimedia Appendix 1
Differences in sleep features and their distribution patterns based on sources.
[DOCX File, 62941 KB - formative_v9i1e67455_app1.docx]

References

1. Lund HG, Reider BD, Whiting AB, Prichard JR. Sleep patterns and predictors of disturbed sleep in a large population of college students. J Adolesc Health 2010 Feb;46(2):124-132. [doi: 10.1016/j.jadohealth.2009.06.016] [Medline: 20113918]
2. Creswell JD, Tumminia MJ, Price S, et al. Nightly sleep duration predicts grade point average in the first year of college. Proc Natl Acad Sci U S A 2023 Feb 21;120(8):e2209123120. [doi: 10.1073/pnas.2209123120] [Medline: 36780521]
3. Wong ML, Lau EYY, Wan JHY, Cheung SF, Hui CH, Mok DSY. The interplay between sleep and mood in predicting academic functioning, physical health and psychological health: a longitudinal study. J Psychosom Res 2013 Apr;74(4):271-277. [doi: 10.1016/j.jpsychores.2012.08.014] [Medline: 23497826]
4. Riemann D, Krone LB, Wulff K, Nissen C. Sleep, insomnia, and depression. Neuropsychopharmacology 2020 Jan;45(1):74-89. [doi: 10.1038/s41386-019-0411-y] [Medline: 31071719]
5. Goldstein TR, Franzen PL. Sleep difficulties and suicidality in youth: current research and future directions. Curr Opin Psychol 2020 Aug;34:27-31. [doi: 10.1016/j.copsyc.2019.08.021] [Medline: 31539832]
6. Courvoisier DS, Eid M, Lischetzke T. Compliance to a cell phone-based ecological momentary assessment study: the effect of time and personality characteristics. Psychol Assess 2012 Sep;24(3):713-720. [doi: 10.1037/a0026733] [Medline: 22250597]
7. Rintala A, Wampers M, Myin-Germeys I, Viechtbauer W. Response compliance and predictors thereof in studies using the experience sampling method. Psychol Assess 2019 Feb;31(2):226-235. [doi: 10.1037/pas0000662] [Medline: 30394762]
8. Walia HK, Mehra R. Practical aspects of actigraphy and approaches in clinical and research domains. Handb Clin Neurol 2019;160:371-379. [doi: 10.1016/B978-0-444-64032-1.00024-2] [Medline: 31277861]
9. Rupp TL, Balkin TJ. Comparison of Motionlogger Watch and Actiwatch actigraphs to polysomnography for sleep/wake estimation in healthy young adults. Behav Res Methods 2011 Dec;43(4):1152-1160. [doi: 10.3758/s13428-011-0098-4] [Medline: 21512871]
10. Fino E, Plazzi G, Filardi M, et al. (Not so) Smart sleep tracking through the phone: findings from a polysomnography study testing the reliability of four sleep applications. J Sleep Res 2020 Feb;29(1):e12935. [doi: 10.1111/jsr.12935] [Medline: 31674096]

XSL•FO
RenderX

11. Borger JN, Huber R, Ghosh A. Capturing sleep-wake cycles by using day-to-day smartphone touchscreen interactions. NPJ Digit Med 2019;2:73. [doi: 10.1038/s41746-019-0147-4] [Medline: 31372507]
12. Langholm C, Byun AJS, Mullington J, Torous J. Monitoring sleep using smartphone data in a population of college students. Npj Ment Health Res 2023 Mar 17;2(1):3. [doi: 10.1038/s44184-023-00023-0] [Medline: 38609478]
13. Lind MN, Kahn LE, Crowley R, Reed W, Wicks G, Allen NB. Reintroducing the Effortless Assessment Research System (EARS). JMIR Ment Health 2023 Apr 26;10:e38920. [doi: 10.2196/38920] [Medline: 37099361]
14. Latham MD. A micro-randomized trial to improve college students' wake time regularity (2444383124). : ProQuest Diss Theses Glob; 2020.
15. Liao P, Huang L, Seewald NJ, Sun J. Package 'MRTSampleSize'. CRAN. 2020. URL: https://cran.r-project.org/web/packages/MRTSampleSize/MRTSampleSize.pdf [accessed 2025-08-06]
16. Buysse DJ, Reynolds CF III, Monk TH, Berman SR, Kupfer DJ. The Pittsburgh sleep quality index: a new instrument for psychiatric practice and research. Psychiatry Res 1989 May;28(2):193-213. [doi: 10.1016/0165-1781(89)90047-4] [Medline: 2748771]
17. Lind MN, Byrne ML, Wicks G, Smidt AM, Allen NB. The Effortless Assessment of Risk States (EARS) tool: an interpersonal approach to mobile sensing. JMIR Ment Health 2018 Aug 28;5(3):e10334. [doi: 10.2196/10334] [Medline: 30154072]
18. Carney CE, Buysse DJ, Ancoli-Israel S, et al. The consensus sleep diary: standardizing prospective sleep self-monitoring. Sleep 2012 Feb 1;35(2):287-302. [doi: 10.5665/sleep.1642] [Medline: 22294820]
19. Cole RJ, Kripke DF, Gruen W, Mullaney DJ, Gillin JC. Automatic sleep/wake identification from wrist activity. Sleep 1992 Oct;15(5):461-469. [doi: 10.1093/sleep/15.5.461] [Medline: 1455130]
20. Lehnert B. BlandAltmanLeh: Plots (slightly extended) Bland-Altman Plots. CRAN. 2015. URL: https://cran.r-project.org/web/packages/BlandAltmanLeh/index.html [accessed 2025-08-06]
21. Giavarina D. Understanding Bland Altman analysis. Biochem Med (Zagreb) 2015;25(2):141-151. [doi: 10.11613/BM.2015.015] [Medline: 26110027]
22. Saeb S, Cybulski TR, Schueller SM, Kording KP, Mohr DC. Scalable passive sleep monitoring using mobile phones: opportunities and obstacles. J Med Internet Res 2017 Apr 18;19(4):e118. [doi: 10.2196/jmir.6821] [Medline: 28420605]
23. Maynard T, Appleman E, Cronin-Golomb A, Neargarder S. Objective measurement of sleep by smartphone application: comparison with actigraphy and relation to self-reported sleep. Explor Med 2021;2:382-391. [doi: 10.37349/emed.2021.00057]
24. Toon E, Davey MJ, Hollis SL, Nixon GM, Horne RSC, Biggs SN. Comparison of commercial wrist-based and smartphone accelerometers, actigraphy, and PSG in a clinical cohort of children and adolescents. J Clin Sleep Med 2016 Mar;12(3):343-350. [doi: 10.5664/jcsm.5580] [Medline: 26446248]
25. Carney CE, Edinger JD, Meyer B, Lindman L, Istre T. Daily activities and sleep quality in college students. Chronobiol Int 2006;23(3):623-637. [doi: 10.1080/07420520600650695] [Medline: 16753946]

Abbreviations

EARS: Effortless Assessment of Research System
FPR: false positive rate
PSG: polysomnography
PSQI: Pittsburgh Sleep Quality Index
TPR: true positive rate

Edited by A Mavragani; submitted 11.10.24; peer-reviewed by H Namba, M Cozad; revised version received 26.02.25; accepted 02.03.25; published 14.08.25.

Please cite as:
Kirshenbaum JS, Crowley RN, Latham MD, Pagliaccio D, Auerbach RP, Allen NB
Comparison of Sleep Features Across Smartphone Sensors, Actigraphy, and Diaries Among Young Adults: Longitudinal Observational Study
JMIR Form Res 2025;9:e67455
URL: https://formative.jmir.org/2025/1/e67455
doi:10.2196/67455

© Jaclyn S Kirshenbaum, Ryann N Crowley, Melissa D Latham, David Pagliaccio, Randy P Auerbach, Nicholas B Allen. Originally published in JMIR Formative Research (https://formative.jmir.org), 14.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative

XSL•FO
RenderX

Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Exploring Young Adults' Experiences and Beliefs in Asthma Medication Management: Pilot Qualitative Study Comparing Human and Multiple AI Thematic Analysis

Ruth Ndarake Jeminiwa[1], PhD; Caroline Popielaski[2], PharmD; Amber King[1], PharmD

[1]Department of Pharmacy Practice, Jefferson College of Pharmacy, Thomas Jefferson University, 901 Walnut Street, Philadelphia, PA, United States
[2]Long Island Jewish Medical Center, Northwell Health, New Hyde Park, NY, United States

Corresponding Author:
Ruth Ndarake Jeminiwa, PhD
Department of Pharmacy Practice, Jefferson College of Pharmacy, Thomas Jefferson University, 901 Walnut Street, Philadelphia, PA, United States

Abstract

Background: Young adults take their asthma maintenance medication 67% of the time or less. Understanding the specific needs and behaviors of young adults with asthma is essential for developing targeted interventions to improve disease self-management. Artificial intelligence (AI) has demonstrated its utility in summarizing and identifying patterns in qualitative research and may support or augment human coding efforts. However, there is pause literature to support this assertion.

Objective: The objective of this study is to begin to explore the medication management-related needs of young adults with asthma via a pilot feasibility study. We aim to understand how to best assist young adults with asthma self-management and to identify potential areas where digital health interventions can provide support. We further aimed to understand the comparative outcome of human versus multiple AI platforms in performing thematic analysis.

Methods: This study purposefully sampled young adults between the ages of 18 years and 29 years who had a prescription for an inhaled corticosteroid (ICS) and were either students or staff of a large metropolitan university in the northeastern United States. Semistructured interviews lasting 40 minutes on average were conducted with 4 participants via a teleconferencing application to elicit young adults' opinions on the topic. Interviews were recorded and transcribed verbatim using Otter.ai (Otter.ai, Inc). Investigators listened to the recording to confirm the accuracy of transcriptions and to make corrections when necessary. After performing a second round of line-by-line coding, the codes were reviewed by investigators and grouped into broader, overarching themes. All investigators reviewed and discussed the final codes. Human qualitative data analyses were performed using NVivo 14 software (QSR International). After completing human analyses, the investigators performed thematic analysis with multiple AI platforms (Google Gemini, Microsoft Copilot, and OpenAI's ChatGPT) to compare the final themes with investigator-derived themes.

Results: Human analysis yielded 4 themes: support from clinicians, social support, digital self-management support, and educational support. The AI-based analysis also generated similar themes with different labels. The level of overlap on the underlying concept between humans, Gemini, Copilot, and ChatGPT was high, accounting for the fact that, although the specific labels differed, they referred to the same concept.

Conclusions: Findings from our pilot exploratory study offer insights into the necessity for a holistic approach in supporting young adults with asthma. Based on the health belief model, if the identified multifaceted needs are addressed, health care systems may support medication adherence and improve health outcomes for this understudied patient population. Our pilot study also offers preliminary findings that artificial intelligence may be leveraged for successful thematic analysis of qualitative data with appropriate caution.

(JMIR Form Res 2025;9:e69892) doi:10.2196/69892

KEYWORDS

asthma; young adults; medication management-related needs; AI; Gemini; Copilot; ChatGPT; thematic analysis

Introduction

Asthma, a chronic respiratory disease, imposes a significant burden on individuals of all ages. In the United States alone, approximately 25 million people have a diagnosis of asthma, leading to missed school or work days, emergency department visits, and premature death [1,2]. According to data published by the Centers for Disease Control and Prevention, there were over 1.8 million asthma-related emergency department visits in 2019 [1]. The economic burden of asthma is substantial,

XSL•FO
RenderX

estimated at US$81.9 billion per year, encompassing the cost of asthma medications, hospital visits due to asthma exacerbations, costs incurred by absenteeism, and mortality [3]. Asthma is not curable, but there are effective therapies that can significantly reduce symptoms and prevent worsening disease [4]. The Global Initiative for Asthma (GINA) guidelines recommend that all adults, adolescents, and children above the age of 5, with a diagnosis of asthma, should receive inhaled corticosteroid (ICS)-containing treatment, either as daily maintenance therapy or as as-needed ICS-formoterol for those with mild asthma [5]. Adherence to treatment regimens is crucial to avoid disease symptoms, exacerbations, and other complications in those who have more than mild asthma [5].

A particularly vulnerable group is young adults (18 - 25 years), who face unique challenges such as increased independence, an optimism that may blind them to the consequences of nonadherence, and decreased familial support, all of which may contribute to poor medication adherence [6]. In fact, studies have reported that young adults take their asthma maintenance medications 67% of the time or less, which predisposes them to frequent asthma exacerbations and hospitalizations [7-9]. Young adults with uncontrolled asthma experience shame, embarrassment, anxiety, and other deleterious emotions [10]. Compared with younger children who have caregivers in charge of their asthma management, young adults are more aware and involved in their care [11]. A previous study has reported that young adults have concerns or negative beliefs about their asthma maintenance medications [12]. Understanding the specific needs and behaviors of young adults with asthma is essential for developing targeted interventions to improve disease self-management.

Digital health platforms offer a promising avenue for addressing the challenges faced by young adults. Previous studies have demonstrated the effectiveness of mobile health technologies in improving adherence to asthma maintenance medications and outcomes [12-14]. By identifying the medication management-related support needs of young adults, we can tailor digital health interventions to support medication adherence and enhance asthma outcomes for this age group. Artificial intelligence (AI) is disrupting diverse sectors, including health care. In fact, it's been deployed in numerous capacities such as medical education, patient engagement, data analysis, and health interventions [15]. Artificial intelligence has demonstrated its utility in summarizing and identifying patterns in qualitative research and may support or augment human coding efforts; however, there is limited literature to support this assertion [16]. Comparative analyses of human versus top AI platforms in performing thematic analysis are still lacking. The objective of this study is to begin to explore the medication management-related needs of young adults with asthma via a pilot feasibility study. We aim to understand how to best assist young adults with asthma self-management and identify potential areas where digital health interventions can provide support. We further aimed to understand the comparative outcome of human versus multiple AI platforms in performing thematic analysis. In this study, pilot feasibility refers to the feasibility of our recruitment strategy to recruit young adults with asthma within our institution, as well as the practicality of using AI as a data analysis support tool.

Methods

Recruitment

This study purposefully sampled young adults between the ages of 18 years and 29 years who had a prescription for an ICS and were either students or staff of a large metropolitan university in the northeastern United States. Participants were recruited between October and November 2021. A study flyer was distributed across the university campus and dormitories to facilitate recruitment. Interested participants contacted the primary investigator via email and were recruited into the study if they met the eligibility criteria.

Ethical Considerations

Participants received a US$20 gift card as compensation for their time. This study was reviewed and approved by the primary author's Institutional Review Board (Approval number: 21D.616). All participants provided informed consent before participation, and interview transcripts were fully anonymized before analysis. Personal identifiers—including names, locations, or any potentially identifying characteristics—were removed from transcripts to ensure participant confidentiality. In analyzing the qualitative data using AI-based tools, only deidentified and nonsensitive information was processed. No directly identifying details were input into AI models. All data were stored on secure, access-controlled servers. These procedures align with ethical principles for human participants' research and ensure compliance with data privacy standards in the handling of qualitative health data.

Interviews

Semistructured interviews lasting 40 minutes on average were conducted via a teleconferencing application to elicit young adults' opinions on the topic. RJ, a pharmacy educator proficient in qualitative research, interviewed 4 participants. CP and AK are practicing pharmacists who also have varied experience with a qualitative approach. At the time of the research, CP was a student pharmacist. RJ's previous experience conducting a randomized trial of young adults with asthma informed this qualitative research. The initial belief was that young adults with asthma may not perceive a strong link between nonadherence and negative asthma outcomes, which may have informed questions asking about their perceptions of taking daily ICSs. In addition, all 3 investigators did not have any previous knowledge of, or relationship with the participants, and no relationships were maintained after the study. Our study adopted a realist epistemological stance, seeking to understand the underlying support needs of young adults with asthma as perceived and articulated through their experiences. The interview guide contained open-ended questions related to beliefs about asthma medication, digital health, social support, and preferred sources of information (Multimedia Appendix 1). Interviews were recorded and transcribed verbatim using Otter.ai (Otter.ai, Inc). Investigators listened to the recording to confirm the accuracy of transcriptions and to make corrections when necessary.

XSL•FO
RenderX

Data Analysis

RJ performed an initial line-by-line coding of transcripts to inform the code book. After defining the codes and discussing with CP, they both performed another round of line-by-line coding using the code book. Investigators met to discuss and resolve discrepancies. Codes were reviewed by RJ and CP and grouped into broader overarching themes. All investigators reviewed and discussed the final codes. Our analysis was guided by a realist epistemological stance to understand the medication management support needs of young adults with asthma based on their perceptions. We also maintained a reflective memo and reviewed the themes as a team to enhance the rigor and trustworthiness of our findings and to acknowledge and consider our possible influence on data analysis. Human qualitative data analyses were performed using NVivo 14 software (QSR International). Specifically, we used NVivo to perform a manual line-by-line inductive coding of the transcript. NVivo was subsequently used for the organization and grouping of codes into themes. In addition, memos were written and maintained on NVivo. After completing human analysis, the investigators performed thematic analysis with multiple AI platforms (Google Gemini, Microsoft Copilot, and OpenAI's ChatGPT) to compare the final themes with investigator-derived themes. Specifically, OpenAI's ChatGPT, Microsoft CoPilot, and Google Gemini were provided the following prompts: "Please read through the following transcripts and perform a thematic analysis. First, generate codes and then categorize them into emergent themes. The codes (groups of words) should be relevant to young adults' medication-related needs. Each code should be between one to five words. I will supply the transcripts in batches. Only analyze respondents' transcripts. Wait until I type in the phrase 'Begin analysis'." "Provide your findings in a table with column 1 as Themes, column 2 as Codes, column 3 as Sample Quotes and column 4, summary." The first prompt was keyed in before supplying the transcripts. Deidentified transcripts were copied into batches and pasted in each platform. The second prompt was provided after typing in "Begin Analysis." A final instruction was provided requesting Copilot AI to condense to 4 to 6 themes (Make it 4 to 6 themes). This instruction was not applicable to the other AI platforms, which already provided major themes alone. We assessed the overlap between the themes derived by humans and the AI platforms. The investigators repeated the AI-based analysis for each platform to verify findings. Similar to AI analysis, human thematic analysis was also focused on identifying the medication management support needs of young adults to be consistent with our primary research objective. Although the coding process was inductive, our initial readings and the development of the codebook were guided by this focus on understanding the participants' perceived needs.

Results

Overview

Characteristics of participants are listed in the first table (Table 1). From the investigator-based analysis, there were 4 constructed themes: support from clinicians, social support, self-management support, and educational support (Figure 1, Table 2). The AI-based analysis also generated similar themes (Table 3, Table 4, and Table 5). The level of overlap on the underlying concepts between humans, Gemini, Copilot, and ChatGPT was high, accounting for the fact that, although the specific labels differed, they referred to the same concept (Table 6). However, there were instances of divergence. For example, Gemini labeled a participant's statement as "embarrassment," which, while potentially related to the broader experience of managing asthma, was not identified as a central support need in the human analysis or by the other AI platforms. The team observed that repeating the analyses for each platform led to different labels for themes, varying numbers of themes, and different sample quotes. However, the results of only one iteration are reported.

XSL•FO
RenderX

Table . Characteristics of study participants.

Characteristic	Pilot sample size (n=4)
Sex, n (%)	
Female	2 (50)
Male	2 (50)
Education, n (%)	
Some college	1 (33)
Bachelor's degree	3 (75)
Household income, n (%)	
US$25,000-US$49,000	1 (33)
US$50,000 - US$74,999	0 (0)
US$75,000-US$99,999	2 (50)
US$100,000 and above	1 (33)
Ethnicity, n (%)	
Non-Hispanic	4 (100)
Hispanic	0 (0)
Race, n (%)	
White	4 (100)
Others	0 (100)

Figure 1. Framework of young adults' medication support needs.

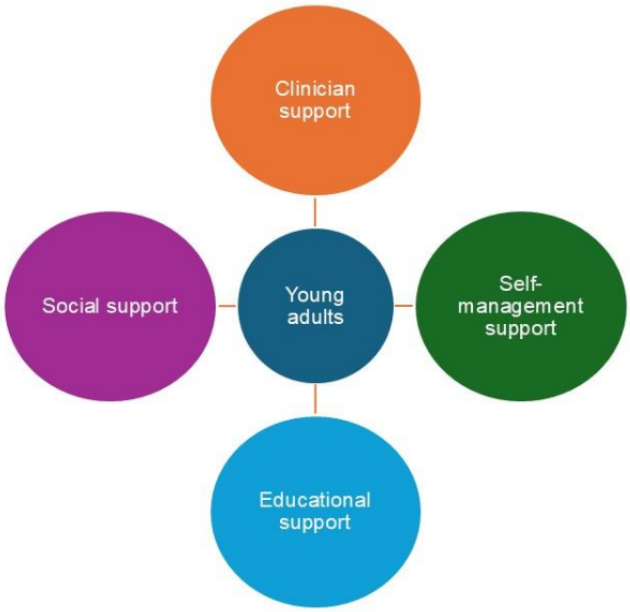

XSL•FO
RenderX

Table . Thematic analysis result from human analysis.

Theme	Sample quotes	Summary
Support from clinicians	"One thing that would be great is if my doctor and my pharmacy had some sort of program where they would immediately send me one [medication] without me having to ask or request a new prescription""A lot of [support from my doctor] has been with managing my flare-ups." "I'm thankful [my allergist is] really quick about making sure they can get me refills and get me medication if I need it."	Primary care physicians and allergists were perceived as most useful for supporting young adults. Collaboration between doctors and pharmacists is also helpful for getting medication refills on time. Support from clinicians includes prescribing medications, assessing adherence to medications, monitoring asthma symptoms, treating flare-ups, and educating on device use.
Social support	"[I received] little bit of support from my friends, but mostly my mom." "My parents pick up my medication for me sometimes." "My teammates know to remind me before races to take my inhaler."	Medication-taking assistance came mainly from parents when patients were younger. Now, medication use responsibility falls almost entirely on them. Young adults also receive support from friends Social support included ensuring the availability of medications, monitoring adherence, and accompanying young adults to the doctor's office.
Digital self-management support	"[I am] working full time. So, it gets hectic. And I think with all of that going on, just having a reminder that hey, I need to take this would help me stay consistent with it." "Being able to track like how much of the medication you have left" "If an app could remind me to take my meds, that would be useful." "I'd want an app connected to a doctor or health system I trust."	Young adults may not always admit a need for help or assistance despite needing one. The biggest obstacle for young adults in managing their asthma is developing a feasible schedule that fits into their busy lifestyles. Many participants admitted a need for: a system to track inhaler use that is linked to clinicians to facilitate quicker refills, reminders to take medications daily, ensuring availability of medications, especially during trips
Educational support	"I don't really know anything about how it makes my breathing better." "...a reminder on how you should be taking your inhalers because I feel like I wasn't really given [that]"	Participants were interested in learning more about asthma, device use, the different medications, and their mechanisms of action.

XSL•FO
RenderX

Table . Thematic analysis result from Google Gemini.

Theme	Code	Sample quotes	Summary
Information seeking and knowledge gaps	Doctor as source	"Mostly my allergist... So most of the information I get is through the office." <response to a question about informational source>	Participants primarily relied on their doctors for information.
	Internet research	"I would say the internet. I think Google a lot."	The internet served as a secondary source for information.
	Knowledge gaps	"I don't know what the side effects are that I should be looking out for."	Participants expressed a desire for more information about medication side effects, long-term effects, and comparisons between different medications.
Medication adherence and challenges	Medication reminders	"I think a reminder would be great. Because I think I definitely need to work on remembering to take it every day."	Participants identified the need for reminders and external support to stay consistent with medication intake.
	Refill challenges	"They're not very fancy with their technology... They have a phone line that's really easy to reach, and somebody's always answering."	Difficulties with obtaining refills and navigating insurance coverage were common challenges.
	Daily routine	"It's just a part of my routine, I don't think of it as a hassle."	Medication adherence was often tied to daily routines and habits.
Social support and family involvement	Family support	"My mom was the most important... She would always make sure I had my inhaler."	Family members, especially parents, played a significant role in providing support, reminders, and advice regarding medication.
	Peer support	"I have a few friends that have asthma, but not, I guess not as severe as me."	Friends were less involved in medication management, likely due to the personal nature of the condition.
	Embarrassment	"It's at times, there's sort of that stereotype of like, the geeky kid needs the inhaler."	Some participants expressed feelings of embarrassment or stigma associated with using inhalers in public.
Technology and medication management	Smartphone app	"I think it could be a good idea. I would want it to be built into my calendar app."	Participants were generally open to using smartphone apps for medication management.
	Reminders	"I think a reminder would be great. Because I think I definitely need to work on remembering to take it every day."	Key features desired in an app included reminders, medication tracking, symptom tracking, and integration with health care providers.
	Data privacy	"I would worry about the security of that... I don't really want to get that now."	Concerns about data privacy and security were raised, particularly regarding sharing personal health information.
Health care provider interactions and trust	Health care provider trust	"I've been seeing him not too long after my diagnosis... So most of the information I get is through the office."	Participants expressed trust in their doctors and pharmacists for medication management.
	Proactive support	"They're just really responsive... They're not very fancy with their technology."	Participants desired more proactive support from health care providers, such as regular check-ins and personalized advice.
	Clear communication	"I think it's important to have clear communication and transparency regarding medication information." (Summarized from multiple responses)	Clear communication and transparency regarding medication information were valued by participants.

XSL•FO
RenderX

Table . Thematic analysis result output from ChatGPT.

Themes	Codes	Sample quote	Summary
Information sources and management	Trust in health care providers, family involvement	"Mostly my allergist...because I have food allergies and asthma. So, he kind of takes care of everything for me. <response to a question about informational source>"	The interviewee relies on their allergist for credible information about asthma and medications. Family members, particularly the mother, play a key role in conveying and contextualizing medical information.
Social support dynamics	Family vs friend support, shared experiences	"...mostly my mom because as primary caretaker when I was a kid... she had a lot of background for that and was very on top of, -okay, you need to take your medication..."	The interviewee receives more support from family than friends, with a shared understanding of asthma within the family, enhancing emotional support and management.
Health care professional engagement	Positive provider relationship, limited pharmacist engagement	"...my allergist, I'm thankful they're really quick about making sure they can get me refills and get me medication if I need it."	The interviewee has a positive relationship with their allergist, who is responsive and effective. However, they feel less engaged with pharmacists, viewing their role as limited.
Challenges with medication adherence	Forgetfulness, perceived burden	"...I think I would [benefit from additional support]... I oftentimes do forget in the morning, and I'll remember later on in the day."	The interviewee struggles with remembering to take daily medications and finds Symbicort more burdensome compared to the convenience of a rescue inhaler.
Technology's role in medication management	Interest in health apps, data privacy concerns	"In general, as long as it didn't collect too much personal information, I'd probably be pretty supportive of that."	The interviewee shows interest in using an app for medication reminders but is concerned about data privacy and security. Desired features include reminders and side effect tracking.
Patient perceptions of medication	Medication as a chore, convenience vs long-term health	"It's a chore. I wish that it wasn't something I had to do... but at the same time, the rescue inhaler just feels a lot easier because I only have to remember to take it when I need it."	The interviewee views taking Symbicort as a chore but acknowledges its effectiveness. Their preference for the convenience of a rescue inhaler reflects a common tension between immediate needs and long-term health management.
Challenges with medication adherence	Forgetfulness, perceived burden	"...my biggest concern is a breach... I think it's definitely made me a lot more data conscious of who has what information and how that can lead to vulnerabilities."	The interviewee struggles with remembering to take daily medications and finds Symbicort more burdensome compared to the convenience of a rescue inhaler.

Table . Thematic analysis result output from Copilot.

Themes	Codes	Sample quotes
Reliable sources of information	Mom (pediatrician), family experience, and physician check-ups	"If I had any concerns, I still ask my mom."
Social and family support	Family support, siblings with asthma, and girlfriend's reminders	"My girlfriend might, you know, say something like that if she notices I'm coughing a lot."
Health care professional interaction	Medication review, prescription updates, check-ups, prescription fulfillment, mail delivery, and convenience	"My physician will certainly ask if everything is managed, everything is controlled."
Challenges in medication adherence	Checking doses, running out of medication, traveling with medication, notification overload, and app clutter	"The biggest challenge for sure, is just planning ahead and making sure that I do have doses."
Desired app features and health information	Built-in calendar integration, reminders, check-off feature, communication with pharmacist and doctor, symptom surveys, air quality updates, general health information, medication-specific information, and asthma health education	"I wouldn't want a separate app, I want to use an app that I already have, probably like the calendar app or the reminder app."

XSL•FO
RenderX

Table . Agreement with human on themes.

Human	Gemini	ChatGPT	CoPilot
Social support	Social support and family involvement	Social support dynamics	Social and family support
Clinician support	Health care provider interactions and Trust	Health care professional engagement	Health care professional interaction
Digital self-management support	Medication adherence and challenges, technology and medication management	Challenges with medication adherence, perceptions of medication, technology's role in medication management,	Challenges in medication adherence, desired app features and health information
Educational support	Information seeking and knowledge gaps	Information sources and management	Reliable sources of information

Themes

Social support (Social support and family involvement [Gemini]; Social support dynamics [ChatGPT]; Social and Family Support [CoPilot]): participants receive support from family members, especially their mothers. Participants acknowledged that support from family was declining, as most medication-taking responsibilities were theirs to handle. Apart from family members, participants also receive support from team members or colleagues. However, they receive more support from family than friends. Social support encompassed tangible assistance such as ensuring medication availability, monitoring medication adherence, giving medication-taking advice, and escorting young adults to doctor office visits.

Clinician support (Health care professional engagement [Gemini]; Health care professional interactions and trust [ChatGPT]; Health care professional interaction [CoPilot]): primary care physicians and allergists were the main source of clinician support for participants. This patient group desires collaboration between doctors and pharmacists to facilitate quicker prescriptions and medication refills. Participants expressed trust in their doctors and pharmacists for medication management, emphasized the need for more proactive support like regular check-ins and personalized advice, and valued clear communication and transparency regarding medication information. Clinician support encompasses prescribing medications, assessing adherence, monitoring asthma symptoms, treating exacerbations, and providing education on device use.

Digital self-management support (Medication adherence and challenges, technology and medication management [Gemini]; Challenges with medication adherence, technology and medication management; Perceptions of medication [ChatGPT]; Challenges in medication adherence, desired app features and health information [CoPilot]): participants managing asthma face challenges such as busy schedules, a need for systems to track inhaler use, and difficulties remembering daily medications. They expressed a desire for proactive support, including clinician-linked refill systems and medication reminders, especially for trips. Many find daily medication more burdensome than rescue inhalers, labeling them as a chore or something they had to do. Participants were interested in smartphone apps for medication management, but 1 person was concerned about data privacy. Desired app features include reminders, medication management, asthma attack detection, symptom tracking, built-in calendar integration, check-off feature to track adherence, air quality updates, general health information, medication-specific information, asthma health education, and integration with health care providers. External reward systems to support adherence, such as star rewards, motivated some participants to be adherent.

Educational support (Information seeking and knowledge gaps [Gemini]; Information sources and management [ChatGPT]; Reliable sources of information [CoPilot]): while the AI platforms focused on the need for trustworthy information about asthma and asthma medications, human investigators coined this theme as educational support, depicting a need for information on asthma, medication, device use, and new evidence on disease management. Participants primarily relied on their doctors for credible information about asthma and medications, with family members, particularly mothers, playing a key role in contextualizing this information. The internet served as a secondary source. There was a strong desire to receive information or education on asthma, asthma medications, device use, medication side effects, long-term medication effects, new evidence on medication outcomes, and comparison of clinical outcomes between different medications.

Lessons Learned

There are 2 important lessons from this pilot study. First, participant recruitment via flyers alone was ineffective. Going forward, a viable collaboration with an asthma clinic or a primary care practice where young adults with asthma are seen would be a better way of recruiting patients for a similar study. Although our team made efforts to work with our asthma clinic and primary care department, the connection did not mature on time for the pilot study. Using multiple recruitment strategies may be more effective than depending on a single strategy. Second, although the findings were similar, using AI proved to be more efficient in qualitative data analysis. It took humans several months to complete the analysis compared to a few minutes for each AI platform. An AI platform could change the labels for themes, the number of themes, or sample code with different iterations of analysis.

Discussion

Principal Findings

This pilot feasibility study compared the thematic analysis performed by humans and multiple artificial intelligence

XSL•FO
RenderX

platforms (Google Gemini, Microsoft Copilot, and ChatGPT) to assess young adults' medication management-related needs. Findings from the study indicated that young adults need multidimensional support (social support, health care provider interaction, digital self-management support, and informational or educational support) to succeed in adherence to their daily ICSs. Apart from slight differences in the specific names of themes, humans and AI had similar results, and human analysis took a longer time to complete. For example, while Gemini labeled a theme as "Healthcare Professional Engagement," ChatGPT labeled it as "Healthcare Professional Interaction and Trust", CoPilot labeled it as "Healthcare Professional Interaction", and the human labeled it "Clinician Support." All clearly pointed to the same underlying concept of needing clinicians to communicate with patients to prescribe medications and monitor health outcomes. However, there were instances of divergence. For example, Gemini labeled a participant's statement as "embarrassment," which, while potentially related to the broader experience of managing asthma, was not identified as a central support need in the human analysis or by other AI platforms. In addition, ChatGPT identified "Trust in health care providers," a theme absent in our human analysis. These divergent interpretations illustrate key differences in how AI and humans understand nuanced social context and research focus in qualitative data.

Comparison With Previous Work

Our study demonstrated considerable overlap between thematic analysis output across the 3 AI platforms and our inductive human thematic analysis, offering preliminary support for the potential of AI supporting and augmenting human coders in qualitative studies [16]. However, there is a need for domain experts to work with AI, given the possibilities of inaccuracies without human supervision. Our study found that different iterations of analysis led to different labels for themes, different numbers of themes, and different sample quotes, suggesting a need for caution and supervision by a subject matter expert conversant with qualitative data analysis. AI will serve better as a research assistant with a likely human expert giving curated prompts and reviewing AI output for accuracy [17]. Using AI in qualitative research speeds up data analysis and leads to quicker project completions [16]. Since all 3 platforms provided similar results, selecting a specific AI for data analysis may be a matter of personal preference or what is available to the researcher. Although there has been resistance to using AI in a human-AI collaborative approach for research, AI in qualitative research could represent the future [18].

Findings from this study indicate that social support from family and friends is important to young adults. Young adults turn to these individuals for instrumental and emotional support, which may lead to satisfaction, emotional well-being, and physical functioning [19]. Our study identified potential areas for support, including participation in medication management such as reminders or check-ins, ensuring availability of medications, and accompanying a young adult for a clinician visit. Although support from family members declines compared with adolescents, young adults benefit from tangible support, encouragement, and accountability obtained from family members. Previous studies have reported that young adults involve family members in their asthma management but are cautious about burdening friends and family [20]. Based on the findings of this study, peer support is more limited and often stems from concern over seeing a young adult experience exacerbations, typically taking the form of brief questions such as, "Have you taken your meds?" Unfortunately, some young adults may not always get support from friends and family, providing an opportunity for interventions to help cultivate or leverage existing social networks to support young adults with asthma [20].

Our study found that young adults would love their health care providers to be engaged and supportive in making their care seamless and effortless. This preference is akin to the patient-centered medical home model, which promotes team-based care that is patient-centric, coordinating care across the health care system to provide seamless input and communication between providers [21]. Young adults identified allergists, primary care providers, and pharmacists as provider team members who should work together to coordinate and deliver care by prescribing medications, assessing adherence, monitoring asthma symptoms, treating exacerbations, and providing education on device use. Patient-centered medical homes show promise in delivering care that offers improved patient experience, a key aspect valued by young adults [21]. A previous study reports that emerging adults value easy access to asthma care providers, corroborating our finding on the need for support from clinicians [20].

Our study identified various opportunities for supporting asthma self-management among young adults, including medication management, refill requests, asthma symptom tracking, and exacerbation or flare-up detection before its occurrence to facilitate preventive actions. In fact, there was a strong interest in the use of smartphone apps to support asthma self-management. Similar to our study findings, reports of an earlier qualitative study indicate that young adults prefer to use technology to track asthma triggers, set reminders for medication-taking, or for education [22]. Intentional nonadherence, where individuals do not believe they need the drug, and unintentional nonadherence due to factors such as forgetfulness or changes in schedule have been reported by earlier studies investigating adherence to daily ICS in young adults with asthma [23,24]. While smartphone app reminders may address unintentional nonadherence, such as forgetfulness, there is a need for more innovative solutions to tackle intentional nonadherence. Psychological approaches such as message framing—where subtle variations in information framing can encourage people to adopt key health behaviors—hold promise in this area [12,25].

Our study highlights the need for clear, accessible information about medication use, potential side effects, and long-term benefits for young adults with asthma. Participants expressed a need for ongoing education to understand their treatment and make informed decisions, highlighting a gap that health care providers must address. Health plans and other relevant stakeholders could potentially incentivize regular educational programs that ensure that young adults with asthma are knowledgeable about their disease state, treatment options, treatment benefits, and possible side effects.

XSL•FO
RenderX

According to the health belief model, a person's perceptions of disease severity, susceptibility to the disease, and perceived benefits and barriers to taking an action influence their health behaviors [26]. In other words, young adults with asthma may be more likely to take their daily ICSs as prescribed if they are aware of the medication, perceive themselves to be at risk of serious consequences of nonadherence, possess the self-efficacy to adhere to their medications, and experience minimal barriers to doing so. The needs identified in this study are akin to perceived barriers and may contribute to the likelihood of young adults taking their daily ICSs as prescribed, if addressed. In addition, our findings on the support needs of young adults have implications for the design of digital health applications for this patient population. For example, the expressed need for reminders, asthma attack detection, symptom tracking, check-off feature to track adherence, air quality updates, and general health information could be translated directly into app features for young adults with asthma. Mapping these features into smartphone apps may enhance the chances of improving medication adherence and asthma outcomes for this patient population, as exemplified by previous research [12,14].

The integration of AI tools into qualitative research presents both opportunities and ethical challenges [27]. In this study, we prioritized participant confidentiality by using only fully anonymized data when engaging with AI tools. Nonetheless, we acknowledge that AI platforms may raise concerns about data security and compliance with data privacy regulations [27]. We addressed these concerns by ensuring that no personally identifiable or sensitive health information was shared with AI tools. However, we recognize the broader ethical debate surrounding the use of AI in human participants' research, particularly regarding transparency, accountability, and the potential for unintended biases in AI-generated outputs. Future work in this area should continue to explore best practices for ethical AI use in qualitative health research, including clearer regulatory guidance, institutional support for secure AI infrastructures, and increased researcher literacy on AI limitations. As this field evolves, maintaining participant trust and upholding ethical rigor will remain paramount.

Limitations

The study provides important methodological information on how to perform qualitative analysis using AI. It compares the performance of top artificial intelligence platforms and humans on qualitative data analysis, providing helpful information for future researchers. Due to the small sample size, it was difficult to ascertain data saturation. In addition, our interpretations should be taken with caution due to the small sample size, which makes it difficult to draw meaningful conclusions. Our findings are also limited to young adults who are managing their asthma with a daily ICS. However, the aim of the pilot study was to ascertain the study's feasibility to inform the main study and a sample size of 2-5 participants is acceptable [28]. In addition, it is important to note that our exploration of the similarities between human thematic analysis and AI is specific to the approach we used in this study, focusing on using themes related to young adults' expressed support needs. Our findings regarding the degree of similarity observed between human thematic analysis and AI may not be transferable to other qualitative approaches that involve more in-depth interpretation, theoretical frameworks, or different analytical foci, for example, discourse analysis and narrative analysis. In addition, our focus on identifying support needs may have influenced the prompts provided to the AI platforms, potentially leading to a more directed analysis, which may not be applicable to other forms of thematic analysis and may have implications for the observed level of similarity. Furthermore, our AI comparison was limited to Google Gemini, Microsoft Copilot, and ChatGPT; findings may not apply to other AI tools. An additional limitation is the potential environmental impact of using these AI models for research.

Conclusions

Overall, findings from our pilot exploratory study offer insight into key support needs for young adults with asthma. It underscores the necessity for a holistic approach in addressing the needs of young adults with asthma. Based on the health belief model, if identified multi-faceted needs are addressed, health care systems may support medication adherence and improve health outcomes for this understudied patient population. Our pilot study also offers preliminary findings that, for the specific inductive thematic analysis used in this research, artificial intelligence demonstrated efficiency and produced similar outputs to humans. Future studies with larger sample sizes are necessary to confirm and expand upon these initial observations. While AI shows promise for thematic analysis, critical reflections are needed regarding privacy, potential for AI over-interpretation (eg, Gemini's "embarrassment"), environmental impact, and the non-neutral design of these platforms.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview guide.
[DOCX File, 15 KB - formative_v9i1e69892_app1.docx]

References

1. Most recent national asthma data. CDC. 2024. URL: https://www.cdc.gov/asthma/most_recent_national_asthma_data.htm [accessed 2025-07-21]

XSL•FO
RenderX

2. Nunes C, Pereira AM, Morais-Almeida M. Asthma costs and social impact. Asthma Res Pract 2017;3:1. [doi: 10.1186/s40733-016-0029-3] [Medline: 28078100]
3. Nurmagambetov T, Kuwahara R, Garbe P. The economic burden of asthma in the United States, 2008-2013. Ann Am Thorac Soc 2018 Mar;15(3):348-356. [doi: 10.1513/AnnalsATS.201703-259OC] [Medline: 29323930]
4. Asthma. World Health Organization. 2024. URL: https://www.who.int/news-room/fact-sheets/detail/asthma [accessed 2025-07-21]
5. Levy ML, Bacharier LB, Bateman E, et al. Key recommendations for primary care from the 2022 Global Initiative for Asthma (GINA) update. NPJ Prim Care Respir Med 2023 Feb 8;33(1):7. [doi: 10.1038/s41533-023-00330-1] [Medline: 36754956]
6. Arnett JJ. Emerging Adulthood: Understanding the New Way of Coming of Age: Oxford University Press; 2006. [doi: 10.1037/11381-001]
7. Withers AL, Green R. Transition for adolescents and young adults with asthma. Front Pediatr 2019;7:301. [doi: 10.3389/fped.2019.00301] [Medline: 31396495]
8. Dahlén E, Bergström A, Ödling M, Ekström S, Melén E, Kull I. Non-adherence and sub-optimal treatment with asthma medications in young adults: a population-based cohort study. J Asthma 2022 Aug;59(8):1661-1669. [doi: 10.1080/02770903.2021.1941092] [Medline: 34121584]
9. Kolmodin MacDonell K, Naar S, Gibson-Scipio W, Lam P, Secord E. The Detroit young adult asthma project: pilot of a technology-based medication adherence intervention for African-American emerging adults. J Adolesc Health 2016 Oct;59(4):465-471. [doi: 10.1016/j.jadohealth.2016.05.016] [Medline: 27475032]
10. Holmes LJ, Ludlow S, Fowler S, Marshall M, Lovell K. Psychosocial experience of living with severe and uncontrolled asthma as a young adult: a qualitative synthesis. BMJ Open Respir Res 2025 Mar 22;12(1):e002541. [doi: 10.1136/bmjresp-2024-002541] [Medline: 40121020]
11. Kelada L, Molloy CJ, Hibbert P, et al. Child and caregiver experiences and perceptions of asthma self-management. NPJ Prim Care Respir Med 2021 Sep 9;31(1):42. [doi: 10.1038/s41533-021-00253-9] [Medline: 34504105]
12. Jeminiwa R, Garza KB, Chou C, Franco-Watkins A, Fox BI. Effects of framed mobile messages on beliefs, intentions, adherence, and asthma control: a randomized trial. Pharmacy (Basel) 2024 Jan 8;12(1):10. [doi: 10.3390/pharmacy12010010] [Medline: 38251404]
13. Jeminiwa R, Hohmann L, Qian J, Garza K, Hansen R, Fox BI. Impact of eHealth on medication adherence among patients with asthma: A systematic review and meta-analysis. Respir Med 2019 Mar;149:59-68. [doi: 10.1016/j.rmed.2019.02.011]
14. Unni E, Gabriel S, Ariely R. A review of the use and effectiveness of digital health technologies in patients with asthma. Ann Allergy Asthma Immunol 2018 Dec;121(6):680-691. [doi: 10.1016/j.anai.2018.10.016]
15. Alhur A. Redefining healthcare with artificial intelligence (AI): the contributions of ChatGPT, Gemini, and Co-pilot. Cureus 2024 Apr;16(4):e57795. [doi: 10.7759/cureus.57795] [Medline: 38721180]
16. Pattyn F. The value of generative AI for qualitative research: a pilot study. J Data Sci Intell Syst 2024. [doi: 10.47852/bonviewJDSIS4202964]
17. Chubb LA. Me and the aachines: possibilities and pitfalls of using artificial intelligence for qualitative data analysis. Int J Qual Methods 2023 Oct;22:16094069231193593. [doi: 10.1177/16094069231193593]
18. Jiang JA, Wade K, Fiesler C, Brubaker JR. Supporting serendipity: opportunities and challenges for human-AI collaboration in qualitative analysis. Proc ACM Hum-Comput Interact 2021;5(CSCW1):1-23. [doi: 10.1145/3449168]
19. Lynch Milder MK, Bazier A, Ward S, Rand KL, Hirsh AT. Resilience, social support, and health in emerging adults with and without chronic health conditions. Emerg Adulthood 2023 Jun;11(3):557-571. [doi: 10.1177/21676968221148322]
20. Lee JJ, Ogini F, Hashmi M, et al. Black emerging adults with uncontrolled asthma: a qualitative study. J Allergy Clin Immunol Pract 2024 Feb;12(2):355-360. [doi: 10.1016/j.jaip.2023.09.039] [Medline: 37802253]
21. Jackson GL, Powers BJ, Chatterjee R, et al. The patient centered medical home. A systematic review. Ann Intern Med 2013 Feb 5;158(3):169-178. [doi: 10.7326/0003-4819-158-3-201302050-00579] [Medline: 24779044]
22. Murphy J, Molloy GJ, Hynes L, McSharry J. Young adult preferences for digital health interventions to support adherence to inhaled corticosteroids in asthma: a qualitative study. Health Psychol Behav Med 2022;10(1):557-578. [doi: 10.1080/21642850.2022.2085709] [Medline: 35756333]
23. MacDonell KK, Dailey R, Gibson-Scipio W, Wang B, Dinaj-Koci V, Bruzzese JM. Exploring barriers to medication adherence among African American emerging adults with uncontrolled asthma. Health Educ Behav 2023 Feb;50(1):131-135. [doi: 10.1177/10901981211001345] [Medline: 33813923]
24. MacDonell KK, Idalski Carcone A, Naar-King S, et al. African American emerging adults' perspectives on taking asthma controller medication: adherence in the "Age of Feeling In-Between". J Adolesc Res 2015;30(5):607-624. [doi: 10.1177/0743558414561299]
25. Gallagher KM, Updegraff JA. Health message framing effects on attitudes, intentions, and behavior: a meta-analytic review. Ann Behav Med 2012 Feb;43(1):101-116. [doi: 10.1007/s12160-011-9308-7] [Medline: 21993844]
26. Alyafei A, Easton-Carr R. The Health Belief Model of Behavior Change: StatPearls Publishing; 2024. [Medline: 39163427]
27. Blessing E. Regulatory compliance and ethical considerations: compliance challenges and opportunities with the integration of big data and AI. 2024 URL: https://hal.science/hal-04972082v1/document [accessed 2025-08-14]

XSL•FO
RenderX

28. Aziz A, Khan N. The potential uses of pilot study in qualitative research. J Res Rev Soc Sci Pak 2020;3(1):750-767 [FREE Full text]

Abbreviations

AI: artificial intelligence
GINA: Global Initiative for Asthma
ICS: inhaled corticosteroid

Edited by A Mavragani; submitted 10.12.24; peer-reviewed by EM Sandbakken, W Gibson-Scipio; revised version received 29.05.25; accepted 23.06.25; published 15.08.25.

Please cite as:
Jeminiwa RN, Popielaski C, King A
Exploring Young Adults' Experiences and Beliefs in Asthma Medication Management: Pilot Qualitative Study Comparing Human and Multiple AI Thematic Analysis
JMIR Form Res 2025;9:e69892
URL: https://formative.jmir.org/2025/1/e69892
doi:10.2196/69892

© Ruth Ndarake Jeminiwa, Caroline Popielaski, Amber King. Originally published in JMIR Formative Research (https://formative.jmir.org), 15.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

A Multimodal Large Language Model as an End-to-End Classifier of Thyroid Nodule Malignancy Risk: Usability Study

Gerald Gui Ren Sng[1,2], MBBS, MRCP(UK), MMed (Int Med), MPH; Yi Xiang[3], BEng, MTech; Daniel Yan Zheng Lim[2,4], MBBS, MRCP(UK), MMed (Int Med), MTech; Joshua Yi Min Tung[2,5], MBBS, MPH; Jen Hong Tan[2*], BEng, PhD; Chiaw Ling Chng[1*], MBBS, MRCP(UK)

[1]Department of Endocrinology, Singapore General Hospital, 20 College Road, Academia Level 3, Singapore, Singapore

[2]Data Science and Artificial Intelligence Laboratory, Singapore General Hospital, Singapore, Singapore

[3]Office of Insights and Analytics, SingHealth, Singapore, Singapore

[4]Department of Gastroenterology, Singapore General Hospital, Singapore, Singapore

[5]Department of Urology, Singapore General Hospital, Singapore, Singapore

[*]these authors contributed equally

Corresponding Author:
Gerald Gui Ren Sng, MBBS, MRCP(UK), MMed (Int Med), MPH
Department of Endocrinology, Singapore General Hospital, 20 College Road, Academia Level 3, Singapore, Singapore

Abstract

Background: Thyroid nodules are common, with ultrasound imaging as the primary modality for their assessment. Risk stratification systems like the American College of Radiology Thyroid Imaging Reporting and Data System (ACR TI-RADS) have been developed but suffer from interobserver variability and low specificity. Artificial intelligence, particularly large language models (LLMs) with multimodal capabilities, presents opportunities for efficient end-to-end diagnostic processes. However, their clinical utility remains uncertain.

Objective: This study evaluates the accuracy and consistency of multimodal LLMs for thyroid nodule risk stratification using the ACR TI-RADS system, examining the effects of model fine-tuning, image annotation, prompt engineering, and comparing open-source versus commercial models.

Methods: In total, 3 multimodal vision-language models were evaluated: Microsoft's open-source Large Language and Visual Assistant (LLaVA) model, its medically fine-tuned variant (Large Language and Vision Assistant for bioMedicine [LLaVA-Med]), and OpenAI's commercial o3 model. A total of 192 thyroid nodules from publicly available ultrasound image datasets were assessed. Each model was evaluated using 2 prompts (basic and modified) and 2 image scenarios (unlabeled vs radiologist-annotated), yielding 6912 responses. Model outputs were compared with expert ratings for accuracy and consistency. Statistical comparisons included Chi-square tests, Mann-Whitney *U* tests, and Fleiss' kappa for interrater reliability.

Results: Overall, 88.4% (6110/6912) of responses were valid, with the o3 model producing the highest validity rate (2273/2304, 98.6%), followed by LLaVA (2108/2304, 91.5%) and LLaVA-Med (1729/2304, 75%; $P<.001$). The o3 model demonstrated the highest accuracy overall, achieving up to 57.3% accuracy in Thyroid Imaging Reporting and Data System (TI-RADS) classification, although still remaining suboptimal. Labeled images improved accuracy marginally in nodule margin assessment only when evaluating LLaVA models (407/768, 53% to 447/768, 58.2%; $P=.04$). Prompt engineering improved accuracy for composition (649/1,152, 56.3% vs 483/1152, 41.9%; $P<.001$), but significantly reduced accuracy for shape, margins, and overall classification. Consistency was the highest with the o3 model (up to 85.4%), but was comparable for LLaVA and significantly improved with image labeling and modified prompts across multiple TI-RADS categories ($P<.001$). Subgroup analysis for o3 alone showed prompt engineering did not affect accuracy significantly but markedly improved consistency across all TI-RADS categories (up to 97.1% for shape, $P<.001$). Interrater reliability was consistently poor across all combinations (Fleiss' kappa<0.60).

Conclusions: The study demonstrates the comparative advantages and limitations of multimodal LLMs for thyroid nodule risk stratification. While the commercial model (o3) consistently outperformed open-source models in accuracy and consistency, even the best-performing model outputs remained suboptimal for direct clinical deployment. Prompt engineering significantly enhanced output consistency, particularly in the commercial model. These findings underline the importance of strategic model optimization techniques and highlight areas requiring further development before multimodal LLMs can be reliably used in clinical thyroid imaging workflows.

(JMIR Form Res 2025;9:e70863) doi:10.2196/70863

XSL•FO
RenderX

KEYWORDS

thyroid nodules; artificial intelligence; large language models; multimodal; risk stratification systems

Introduction

Thyroid nodules are highly prevalent, with as many as 35% of individuals having thyroid nodules on imaging [1]. Ultrasound is the first-line and most accurate imaging modality to assess thyroid nodules. The high prevalence of predominantly benign thyroid nodules in the general population led to the development of ultrasound risk stratification systems [2-4]. The American College of Radiology Thyroid Imaging Reporting and Data System (ACR TI-RADS) [4] was established to determine which nodules should have fine-needle aspiration or ultrasound follow-up by using 5 ultrasound feature categories and the maximum size of the nodule to derive a Thyroid Imaging Reporting and Data System (TI-RADS) classification ranging from 1 (benign) to 5 (highly suspicious). This system requires the assessment of 5 categories, or descriptors, of ultrasound features: composition, echogenicity, shape, margin, and calcifications (echogenic foci). While the ACR TI-RADS aims to enable the objective assessment of malignancy risk, significant real-life interobserver variability due to a lack of agreement in assigning these features has been reported [5,6]. In addition, the ACR TI-RADS suffers from low diagnostic specificity [7].

Large language models (LLMs) are artificial intelligence (AI) models using deep learning and artificial neural networks to handle large amounts of data, offering high-level predictive performance for varied tasks, including knowledge synthesis and health outcome prediction [8]. They are pretrained on large sets of extant human-generated information to learn patterns and probabilities that allow the synthesis of de novo content.

LLMs have been shown to perform well in medical question-answering tasks [9]. Performance can be further improved by fine-tuning on specific medical datasets [10], producing models such as Google's MedPaLM-2. To date, most research in this field has focused on question-answering problems [8], with use for clinical decision support being a more recent but equally promising area of study [11].

Beyond text-based tasks, there has been a growing interest in the use of vision-language multimodal LLMs, such as Microsoft's Large Language and Visual Assistant (LLaVA), which combines a visual encoder with a general LLM to allow it to synthesize both image and text data [12]. Like text-based LLMs, they can be further fine-tuned with domain-specific knowledge. For instance, LLaVA has been fine-tuned with a large-scale biomedical figure-caption dataset to create the medicine domain–specific Large Language and Vision Assistant for bioMedicine (LLaVA-Med) model [13]. As opposed to purely text-based interpretation of human-generated ultrasound reports, in which the main time-consuming task is still the radiologist's report, this can allow seamless or end-to-end image interpretation and classification tasks, offering both time and process improvements in image-based clinical diagnostic tasks.

Many commercial AI apps have been developed to perform image segmentation and risk stratification for thyroid nodules [14,15]. However, these typically use traditional machine-learning models, which require extensive training and testing, and are furthermore limited in scope to only a single clinical problem. The use of LLMs offers the possibility of an "out-of-the-box" solution with less development time and expense. One previous approach has combined a distinct visual encoder model developed using classical machine learning methods with a set of commercial LLMs with comparable performance to human evaluators [16]. However, this approach is computationally demanding and has the risk of information degradation between steps, affecting overall output quality.

More recently, Cabezas et al [17] performed the first study evaluating OpenAI's GPT-4 and 4o in performing the same task, but reported suboptimal accuracy, particularly for higher-risk nodules. However, the authors did not attempt optimization of the performance of these models, nor assess the consistency of outputs. LLM output is known to be inherently stochastic, with variability between the outputs. Clearly, a model that produced varying outputs for a single scenario would not be suitable for reliable clinical use. Hence, assessment of consistency is equally important as accuracy in evaluating the performance of such models. Finally, the use of commercial LLMs (such as OpenAI) for these tasks can be hampered by concerns about cost, patient privacy, and data protection. Open-source LLMs can be housed within health care institutions themselves, run with existing compute, and may be an alternative to address these concerns [18]. Furthermore, they typically have much fewer parameters and are hence less computationally demanding to run.

Therefore, we sought to explore the following objectives. First, to determine the accuracy and consistency of an open-source end-to-end vision-language model in the assessment of thyroid nodules using a standardized risk assessment system (ACR TI-RADS). Second, to study whether fine-tuning on medical domain knowledge improved the performance of the vision-language model for this task. Third, to evaluate whether the vision-language models were able to produce similar output without explicit human-labeled image segmentation. Fourth, to examine the effect of prompt engineering—a simple but effective method for improving the quality of LLM responses [19]—on the output of the model. Finally, we sought to evaluate the performance of this open-source model against a commercial LLM.

Methods

Materials

In total, 3 different models were evaluated in this study. First, Microsoft's LLaVA model—which, while being putatively less powerful than other commercial multimodal LLMs [20], has the advantage of being open-source and can be easily instruction-tuned in a resource-efficient manner for medical domain knowledge [13]. Instruction tuning is a method of supervised fine-tuning in which models are trained on extant input-output pairs, as opposed to classical fine-tuning, which

XSL•FO
RenderX

typically involves training a base model on a dataset of domain-specific parameters [21]. This instruction-tuned model, known as LLaVA-Med, was used as the second model for the study. The details of the instruction-tuning for LLaVA-Med have been described in the original article by Li et al [13]. Finally, the final model was the commercial OpenAI o3, which has been described by the company as their "most powerful reasoning model" in production [22].

Ultrasound cine-clip images, radiologist-annotated segmentations, and ACR TI-RADS descriptors for 192 different thyroid nodules from 167 patients were obtained from the publicly available dataset published by Yamashita et al [23]. This dataset is comprised of 175 benign and 17 malignant nodules, with an overall mean nodule size of 2.5 (SD 1.4) cm. The nodule characteristics, including breakdown of ACR TI-RADS level in each group, and image acquisition techniques have been described in detail in the original article.

All models were queried with a prompt and individual ultrasound cine-clip images using an action-programming interface. The models were provided with unlabeled images and images labeled with radiologist-annotated regions of interest (ROI) as separate scenarios (Figure 1), for a total of 384 scenarios. As the models are only able to parse static images, only static images were provided, and the full cine-clips of each nodule were not used for this study.

Figure 1. (A) Original unlabeled thyroid nodule ultrasound image. (B) Radiologist-annotated region of interest. (C) Composite overlay forming the submitted "labeled" image [23].

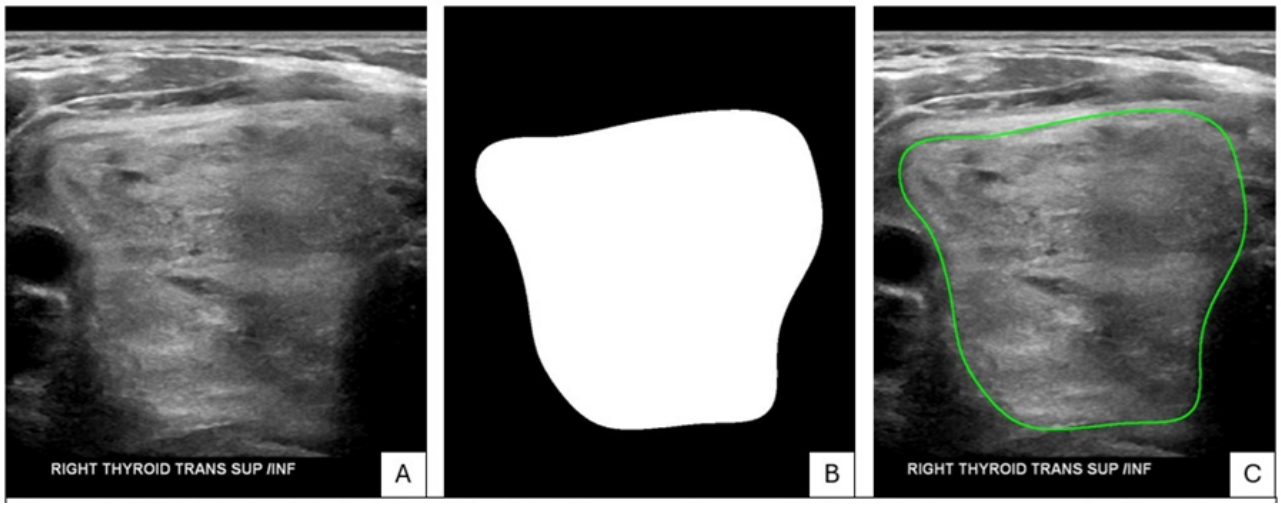

Data Collection

The first evaluation used a simple standardized prompt ("Basic Prompt") detailing the scenario, task, and a summary of the ACR TI-RADS components. To improve the performance and reliability of the output, we repeated the evaluation after prompt engineering ("Modified Prompt") with the following strategies. First, by contextualizing the task by including the descriptors of each component of the ACR TI-RADS, associated component point scores, and the overall point scores associated with each TI-RADS classification level. Second, by using task decomposition to break down the task into a series of ordered steps, in this case, by instructing the models to score each individual component, add up the sum of components, and finally translate the overall score to a TI-RADS classification level. Third, by constraining the model output into a specific format to minimize the rate of invalid responses. Both the basic and modified prompts are illustrated in Table 1.

XSL•FO
RenderX

Table . Basic and modified prompts.

Prompt	Text
Basic Prompt	Assume you are an assistant to a radiologist. This is an ultrasound image of the thyroid. The radiologist would like to know the points for each of the 5 components in Thyroid Imaging Reporting and Data System (TI-RADS) criteria for this image. Here are the 5 components: • Composition. The point value ranges from 0 to 2 • Echogenicity. The point value ranges from 0 to 3 • Shape. It has only two point values: 0 or 3 • Margin. The point value ranges from 0 to 3 • Echogenic Foci. The point value ranges from 0 to 3

Modified Prompt	I would like you to assume the role of a radiologist's assistant. This is an ultrasound image of a thyroid nodule. The radiologist would like to know the points of each of the 5 components of the Thyroid Imaging Reporting and Data System (TI-RADS), the total TI-RADS score, and the TI-RADS classification for this image.
	The 5 components and their points are as follows:
	• Composition (choose one):
	  • Cystic or completely cystic: 0
	  • Spongiform: 0
	  • Mixed cystic and solid: 1
	  • Solid or almost completely solid: 2
	• Echogenicity (choose one):
	  • Anechoic: 0
	  • Hyper- or isoechoic: 1
	  • Hypoechoic: 2
	  • Very hypoechoic: 3
	• Shape (choose one):
	  • Wider than tall: 0
	  • Taller than wide: 3
	• Margin (choose one):
	  • Smooth: 0
	  • Ill-defined: 0
	  • Lobulated/irregular: 2
	  • Extra-thyroidal extension: 3
	• Echogenic foci (choose one or more):
	  • None: 0
	  • Large comet-tail artifact: 0
	  • Macrocalcifications: 1
	  • Peripheral/rim calcifications: 2
	  • Punctate echogenic foci: 3
	Add up the points of each individual component to give the total points.
	Using the total points, the TI-RADS classification is as follows:
	• 0 points: TI-RADS 1
	• 2 points: TI-RADS 2
	• 3 points: TI-RADS 3
	• 4 to 6 points: TI-RADS 4
	• 7 points or more: TI-RADS 5
	Provide your output in the following format:
	• Composition: Single integer
	• Echogenicity: Single integer
	• Shape: Single integer
	• Margin: Single integer
	• Echogenic foci: Single integer
	• Total points: Single integer
	• TI-RADS classification: String

A schematic of the study design is shown in Figure 2. Using each prompt, 3 sets of responses from each model were obtained for each of the 384 scenarios, for a total of 6912 responses. Each response generated by the LLM consisted of 6 components—the 5 individual component scores (composition, echogenicity, shape, margin, and echogenic foci) and the overall TI-RADS

XSL•FO
RenderX

classification level. However, as LLMs are known to be poor at basic arithmetic tasks [24], the study team separately added up the individual component scores manually to derive the overall TI-RADS classification level. While the modified prompt generated an additional seventh item – the sum of the individual component scores – this was designed purely as an intermediary step and was not included in the evaluation. Discrepancies between the LLM-generated and manually-calculated TI-RADS classification level were resolved by selecting the manually-calculated value for the analysis.

Figure 2. Schematic diagram of study design. LLaVA: Large Language and Visual Assistant; LLaVA-Med: Large Language and Vision Assistant for bioMedicine.

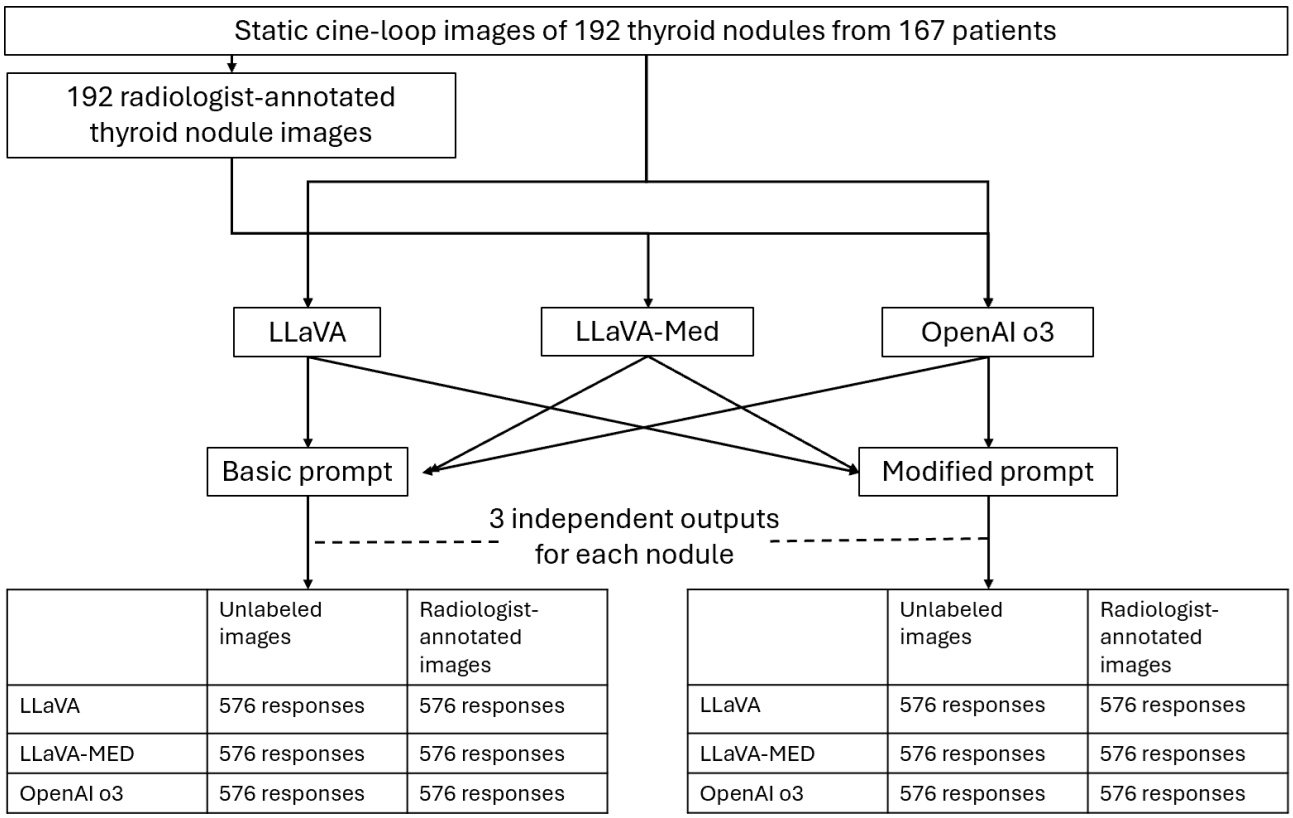

Evaluation using just the overall TI-RADS classification level alone may not be sufficiently robust, as nodules with different imaging characteristics may still derive the same overall classification. Take, for example, 2 nodules, both of which are solid, have regular margins, and are hypoechogenic. The first nodule may be taller than -wide (3 points) without echogenic foci (0 points), while the second nodule may be wider than tall (0 points) with punctate echogenic foci (3 points). However, both would still arrive at the same TI-RADS classification level. Therefore, each item in the response was analyzed individually for accuracy and consistency to demonstrate the models' performance in evaluating each component of the TI-RADS classification, in addition to the overall TI-RADS classification level.

Statistical Analysis

Accuracy of output was defined by comparison with the human-rated TI-RADS classification level obtained from the original dataset. Considering the intrinsic variability of human raters [5,6], the aggregate output was deemed to be "accurate" if at least 1 of the 3 sets of output was concordant with the human-rated score. This "best-of-three" approach has been proposed as a strategy to mitigate LLM output variability [25], and has been used in other studies evaluating the clinical performance of LLMs for a variety of medical tasks [26-29]. Model and human-rated median score distributions were compared in 2 ways. First, it was categorically based on whether the median score underestimated, equaled, or overestimated the TI-RADS classification level. Second, as the score distributions of both model and human-rated scores were determined to be nonparametric using the one-sample Kolmogorov-Smirnov test, they were therefore compared continuously using the Mann-Whitney *U* Test. Consistency of the output was assessed using 2 measures. First, the output was deemed "consistent" if all 3 sets produced the same result. Second, interrater reliability was compared using Fleiss' kappa.

Differences in these metrics were analyzed by groups, namely, across all 3 models, unlabeled versus labeled scenarios, and basic prompt versus modified prompt. Differences in proportions were compared using the chi-square test.

Due to inherent model stochasticity, the models occasionally produced inappropriate or uninterpretable responses. Responses with any missing components, or where the output was a noninteger value (for instance, text or letters), were deemed invalid. As the purpose of this study was to evaluate the real-world model output, these invalid responses were still included in the final analyses for accuracy and consistency but were uniformly treated as inaccurate responses. Similarly, missing and noninteger values were still included in the calculation of interrater reliability.

XSL•FO
RenderX

Statistical analysis was conducted using the Python 3.10 environment, with the pandas 2.2.0 package for table processing, scipy 1.13.1 package for statistical analysis, and seaborn 0.13.2 for data visualization.

Ethical Considerations

As the data used in this study were completely deidentified and freely available in the public domain, no ethical approval was required for this study.

Results

Overview

Out of the 6912 generated responses, 6110 (88.4%) were deemed valid. Use of the modified prompt improved the frequency of valid responses slightly from 87.6% to 89.2% (*P*=.04). The OpenAI o3 model produced the highest proportion of valid responses (2273/2304, 98.6%), followed by the LLaVA model (2108/2304, 91.5%), then the LLaVA-Med model (1729/2304, 75%, *P*<.001 for overall comparison). Most invalid responses were due to the inability of the LLM to interpret the provided image ("I am unable to view the image itself"). Some invalid responses were also due to the output of noninteger values (for instance, letter values ranging from a-e instead), or hallucinations that caused it to deviate from the assigned task ("In this particular image, the Ultrasound Institute of New York grading scale is used to assess the thyroid nodule").

Accuracy

The effect of the various combinations of model, type of prompt, and type of image on the accuracy of output for each component and the overall TI-RADS classification level is visualized as a heatmap in Figure 3. In general, the o3 model had the highest accuracy for most categories except echogenicity and echogenic foci (Table 2). LLaVA-Med appeared to be more accurate than o3 for these 2 categories, although the difference between the 2 models was not significant when compared pairwise (59% vs 54%, *P*=.06 for echogenicity; 60.7% vs 57.6%, *P*=.23 for echogenic foci). The best accuracy obtained for overall TI-RADS classification level was still only 57.3%.

Figure 3. Heatmap showing the distribution of accuracy scores across all combinations of model, prompt, and labeling. LLaVA: Large Language and Visual Assistant.

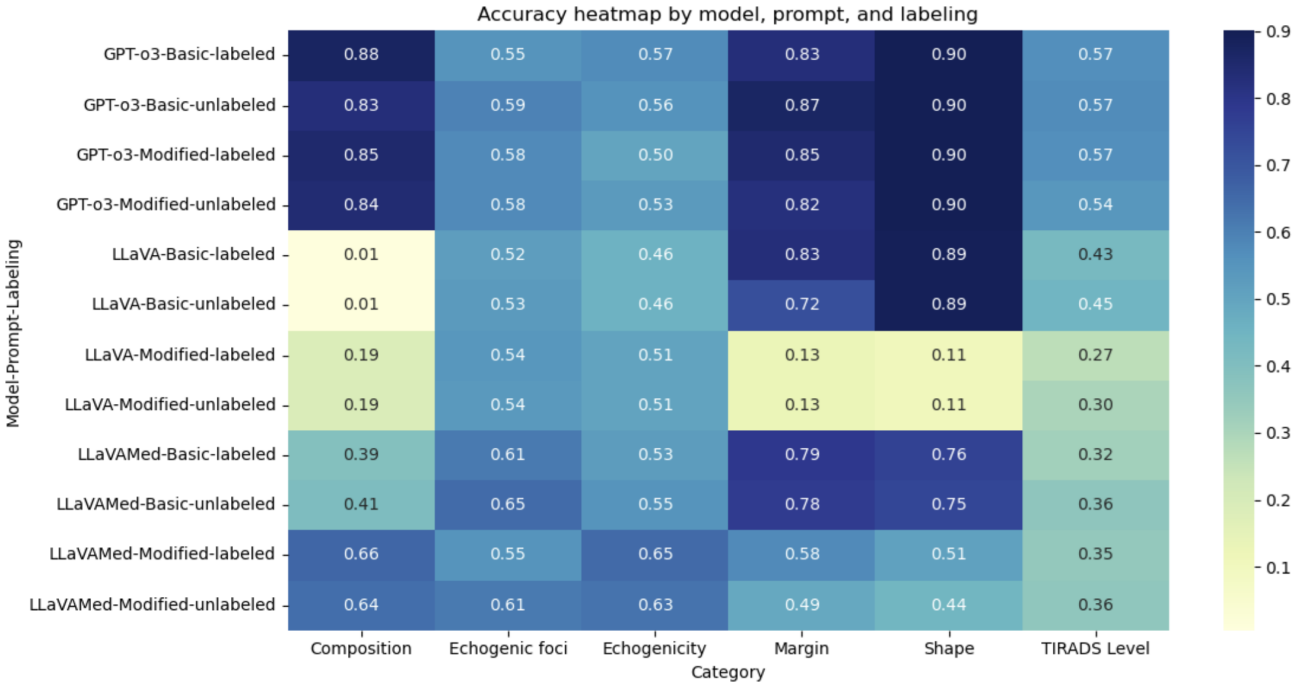

XSL•FO
RenderX

Table . Accuracy of aggregate model output comparing between – Large Language and Vision Assistant, Large Language and Vision Assistant for biomedicine, and o3 models, unlabeled versus labeled scenarios, and base versus modified prompts.

Variable	Model selection				Image annotation			Prompt engineering		
TI-RADS[a] component	LLaVA[b]	LLaVA-Med[c]	GPT-o3	*P* value	Unlabeled	Labeled	*P* value	Basic	Modified	*P* value
Composition	0.10	0.53	0.85	<.001	0.49	0.50	.71	0.42	0.56	<.001
Echogenicity	0.48	0.59	0.54	<.001	0.54	0.54	.90	0.52	0.55	.12
Shape	0.50	0.62	0.90	<.001	0.67	0.68	.53	0.85	0.50	<.001
Margin	0.45	0.66	0.84	<.001	0.64	0.67	.12	0.80	0.50	<.001
Echogenic Foci	0.53	0.61	0.58	.01	0.58	0.56	.27	0.58	0.57	.67
Overall TI-RADS[a] classification level	0.36	0.35	0.56	<.001	0.43	0.42	.47	0.45	0.40	.01

[a]TI-RADS: Thyroid Imaging Reporting and Data System.

[b]LLaVA: Large Language and Vision Assistant.

[c]LLaVA-Med: Large Language and Vision Assistant for bioMedicine.

There was no one best combination of input variables in determining overall accuracy, with different combinations yielding the highest accuracy scores for each component as well as the overall TI-RADS classification level. Interestingly, the basic LLaVA model without prompt engineering or nodule annotation achieved the closest overall accuracy to the worst-performing combination using o3 (44.8% v 54.2%, respectively), despite very poor accuracy for composition and echogenicity.

When combining the output of all 3 models, the use of labeled (radiologist-annotated) images instead of unlabeled images (Table 2) did not make a significant difference to the accuracy of all component scores, nor the overall TI-RADS classification. However, secondary analysis done with output from just the LLaVA and LLaVA-Med models showed that there was a significant improvement in accuracy for margin classification from 53% to 58.2% (*P*=.04).

Use of the modified prompt as compared with the basic prompt (Table 2) significantly improved accuracy in classifying nodule composition (56.3% vs 41.9%, *P*<.001). However, there was worse accuracy in classifying nodule shape (49.6% vs 84.9%, *P*<.001) and margins (0.5% vs 80.3%, *P*<.001). Accuracy for the overall TI-RADS classification level was also worse (45% vs 39.8%, *P*=.01) using the modified prompt. There was no significant difference in the accuracy of classifying echogenicity and echogenic foci using either prompt.

There were no significant differences in median score distributions between the model output and human raters. However, when evaluated categorically, all 3 models appeared to consistently overestimate the scores for echogenicity and underestimate echogenic foci (Multimedia Appendix 1).

Consistency

The effect of the various combinations of model, type of prompt, and type of image on the consistency of output for each component and the overall TI-RADS classification level is visualized as a heatmap in Figure 4. The consistency of both o3 and LLaVA models was largely comparable across all categories except that of margins (80% vs 54.6% respectively, *P*<.001), while that of LLaVA-Med was significantly poorer than both across all categories (Table 3). The highest overall consistency of 85.4% was obtained from the combination of the o3 model with prompt engineering and annotated images.

XSL•FO
RenderX

Figure 4. Heatmap showing the distribution of accuracy scores across all combinations of model, prompt, and labeling. LLaVA: Large Language and Visual Assistant.

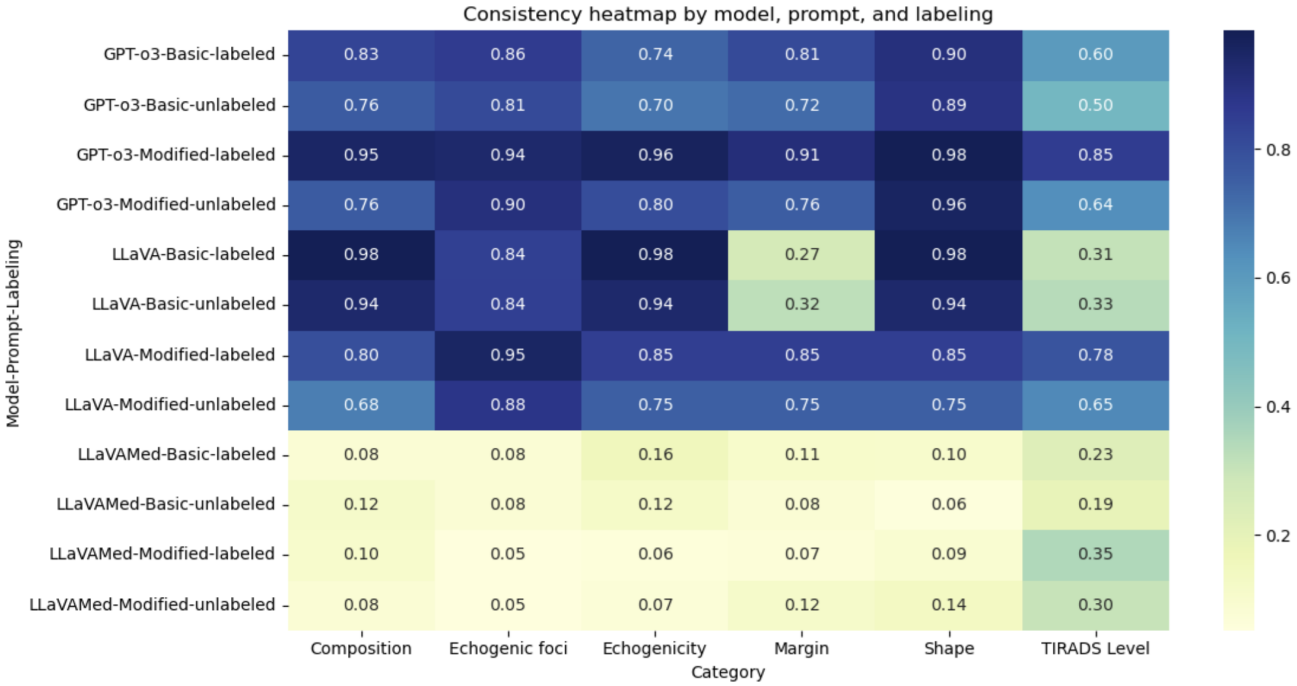

Table . Consistency of aggregate model output comparing between – Large Language and Vision Assistant, Large Language and Vision Assistant for biomedicine, and o3 models, unlabeled versus labeled scenarios, and base versus modified prompts.

Variable	Model selection				Image annotation			Prompt engineering		
TI-RADS[a] component	LLaVA[b]	LLaVA-Med[c]	GPT-o3	*P* value	Unlabeled	Labeled	*P* value	Basic	Modified	*P* value
Composition	0.85	0.10	0.82	<.001	0.56	0.62	.001	0.62	0.56	.008
Echogenicity	0.88	0.10	0.80	<.001	0.56	0.63	.003	0.61	0.58	.29
Shape	0.88	0.10	0.93	<.001	0.62	0.65	.14	0.65	0.63	.39
Margin	0.55	0.10	0.80	<.001	0.46	0.50	.03	0.39	0.58	<.001
Echogenic foci	0.88	0.07	0.88	<.001	0.60	0.62	.22	0.59	0.63	.04
Overall TI-RADS[a] classification level	0.52	0.27	0.65	<.001	0.44	0.52	<.001	0.36	0.60	<.001

[a]TI-RADS: Thyroid Imaging Reporting and Data System.

[b]LLaVA: Large Language and Vision Assistant.

[c]LLaVA-Med: Large Language and Vision Assistant for bioMedicine.

When combining the output of all 3 models, use of labeled images instead of unlabeled images improved consistency of outputs for composition (62.3% vs 55.6%, *P*=.001), echogenicity (62.6% vs 56.3%, *P*=.003), margins (50.3% vs 45.7%, *P*=.03), and overall TI-RADS classification level (52% vs 43.8%, *P*<.001). There was no significant difference in consistency of outputs for shape and echogenic foci (Table 3).

Surprisingly, there was a mixed effect from use of the modified prompt as compared with the basic prompt (Table 3), with poorer consistency for composition (56.2% vs 61.7%, *P*=.008), but higher consistency for margins (57.5% vs 38.5%, *P*<.001), echogenic foci (62.9% vs 58.6%, *P*=.04), and overall TI-RADS classification level (59.5% vs 36%, *P*<.001). There was no significant difference between the type of prompt used for the consistency of rating echogenicity or shape.

Due to the relatively high frequency of invalid responses, interrater reliability was poor across all combinations of input variables, ranging between −0.15 and 0.49. No combination achieved a Fleiss' Kappa above 0.60.

XSL•FO
RenderX

Subgroup Analysis

As the o3 model appeared to perform the best for accuracy as well as consistency, subgroup analysis was performed on output from just the o3 model alone. Neither prompt engineering nor image labeling appeared to affect the accuracy of the output from o3 across all categories and for the overall TI-RADS classification level (Table 4).

Table . Accuracy and consistency of aggregate o3 model output comparing unlabeled versus labeled scenarios, and base versus modified prompts.

Metric	Accuracy						Consistency					
Variable	Image annotation			Prompt engineering			Image annotation			Prompt engineering		
TI-RADS[a] component	Unlabeled	Labeled	*P* value	Basic	Modified	*P* value	Unlabeled	Labeled	*P* value	Basic	Modified	*P* value
Composition	0.84	0.87	.31	0.85	0.85	≥.99	0.76	0.89	<.001	0.79	0.85	.04
Echogenicity	0.54	0.54	.89	0.57	0.52	.19	0.75	0.85	.001	0.72	0.88	<.001
Shape	0.90	0.90	≥.99	0.90	0.90	≥.99	0.92	0.94	.39	0.90	0.97	<.001
Margin	0.85	0.84	.84	0.85	0.84	.69	0.74	0.86	<.001	0.55	0.75	<.001
Echogenic foci	0.59	0.57	.61	0.57	0.58	.94	0.86	0.90	.10	0.84	0.92	<.001
Overall TI-RADS[a] classification level	0.56	0.53	.47	0.57	0.52	.15	0.57	0.73	<.001	0.55	0.75	<.001

[a]TI-RADS: Thyroid Imaging Reporting and Data System.

Significant differences were, however, still observed for consistency. Use of the modified prompt instead of the basic prompt improved consistency across all categories for echogenicity (85.4% vs 79.4%, *P*=.04), echogenicity (88.3% vs 71.9%, *P*<.001), shape (97.1% vs 89.6%, *P*<.001), margins (83.3% vs 76.6%, *P*=.02), echogenic foci (91.9% vs 83.6%, *P*<.001), and overall TI-RADS classification level (74.5% vs 54.9%, *P*<.001). Use of labeled images instead of unlabeled images improved consistency for composition (88.8% vs 76%, *P*<.001), echogenicity (85.2% vs 75%, *P*=.001), margins (86.2% vs 73.7%, *P*<.001), and overall TI-RADS classification level (72.7% vs 56.8%, *P*<.001), but did not make a difference for shape and echogenic foci. These findings are illustrated in Table 4.

Discussion

Principal Findings

To our knowledge, this is the first study to evaluate the use of an open-source multimodal vision-language model alongside a commercial multimodal LLM for the end-to-end risk stratification of thyroid nodules. Furthermore, this is also the first study to consider the important parameter of inter-output variability in assessing model performance for thyroid nodule assessment.

We believe our evaluation of both an open-source smaller model and a more powerful commercial model is important, as it provides a helpful guide for model selection for future development of clinically deployable apps. The putatively most powerful model may not always be the most suitable for a given task, due to reasons of cost, data privacy protection, or compute access. For instance, OpenAI o3 has a token cost of US $10 per 1 million input tokens and US $40 per 1 million output tokens. We consumed a total of 2.79 million input and 1.75 million output tokens to generate the 2304 outputs for this study, incurring a total cost of approximately US $98. By contrast, the LLaVA and LLaVA-Med models were run off a single GPU with no cost beyond the marginal cost of utilities. Furthermore, we were unable to test real-world clinical images on the commercial LLM in this study as our national data privacy regulations prevent us from sharing patient-derived data outside the secure health care computing environment, even if fully anonymized, whereas the LLaVA and LLaVA-Med models could potentially be run using existing hardware within our secure environment.

Use of an end-to-end vision-language model is likely to be more efficient and translatable to other clinical tasks as compared with traditional machine learning predictive models. However, while the newer, more powerful commercial model outperformed the older, smaller open-source model on most metrics of assessment, even the best model output was suboptimal despite optimization. We offer some suggestions for why this may be the case.

First, input of image data directly into a vision-language model can be technically challenging. Modern medical imaging is typically high-resolution and therefore feature-rich. However, current-generation multimodal LLMs struggle with visual identification when feature counts are high [30], with a consequent detrimental impact on performance. Even classical

XSL•FO
RenderX

machine learning models perform optimally at lower image resolutions [31]. Conceptually, the assessment of thyroid nodule composition, echogenicity, and the presence of echogenic foci requires the ability to segment the nodule into multiple ROIs, and thereafter, differentiate between the characteristics of each ROI. For instance, to determine if a nodule is part-solid, a human evaluator would first distinguish that multiple areas of acoustic impedance exist within the nodule, then determine that some are more echogenic than others. Similarly, determination of hyper- or hypoechogenicity requires comparison with normal thyroid parenchyma. As this requires the processing of many features, it is perhaps unsurprising that model accuracy was generally poor for these components. Conversely, when evaluating shape or margins, the nodule can be treated as a whole ROI, which may have been why model accuracy was generally higher for these components. In addition, the different models in this study process images differently. LLaVA [12] and LLaVA-Med [13] tokenize images based on fixed token counts for image area before input into the LLM, whereas o3 can process images dynamically using a unified encoder [32]. This may explain in part the improved (but still suboptimal) accuracy of o3 for echogenicity and echogenic foci.

Second, vision-language models have reduced sensitivity to black-white contrast and have difficulty distinguishing between black and noise [30]. This is a limitation particularly relevant to ultrasound imaging, where regions that appear black can be anechoic or reflect the presence of acoustic shadowing. Distinguishing these can have a significant difference in the evaluation of thyroid nodules, as anechoic nodules are characterized as cystic and require no further evaluation, but acoustic shadowing typically signifies significant calcification, which is a risk feature for malignancy. An example of this is illustrated in Figure 5. In addition, nodule margins are typically recognized by differences in voxel density, which the models may struggle with if differences are small, and therefore use of images with human-annotated margins improved the classification of nodule margins for the LLaVA and LLaVA-Med models and consistency for all 3 models. An example of this is similarly demonstrated in Figure 6.

XSL•FO
RenderX

Figure 5. Ultrasound image from case 3 in the original dataset, showing a nodule with partial rim calcifications and consequent posterior acoustic shadowing (red circles). The nodule composition was erroneously classified as cystic (0 points) or mixed solid and cystic (1 point) in 19 out of 24 outputs, as opposed to almost completely solid (2 points) by the human rater.

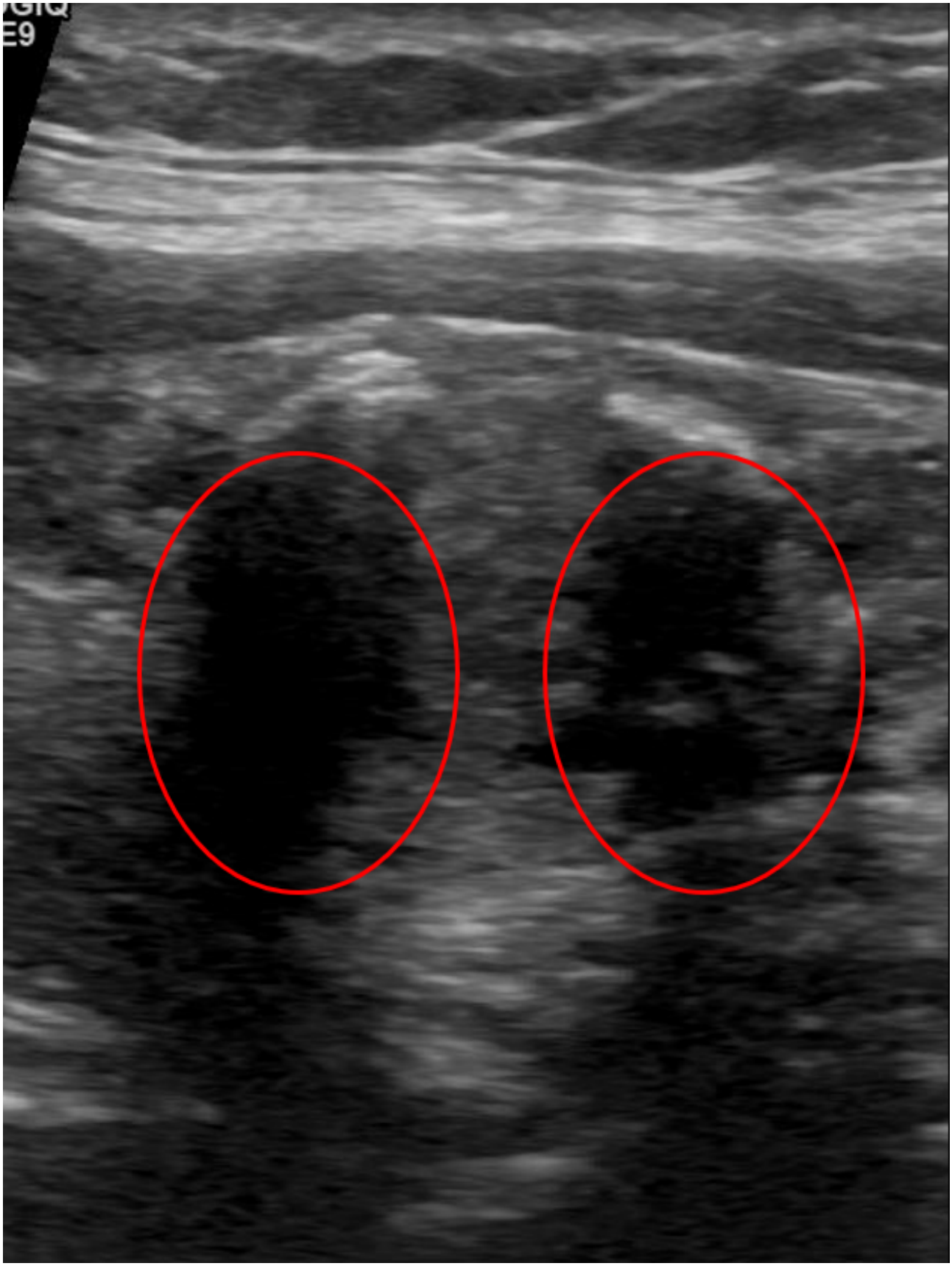

XSL•FO
RenderX

Figure 6. Unlabeled (left) and labeled (right) images from Case 132 in the original dataset. In total, 10 out of 12 outputs, using the unlabeled image, overestimated the nodule margins as irregular (2 points), while 6 out of 12 outputs using the labeled image overestimated the nodule margins as irregular (2 points). The human rater classified the nodule margin as ill-defined (0 points).

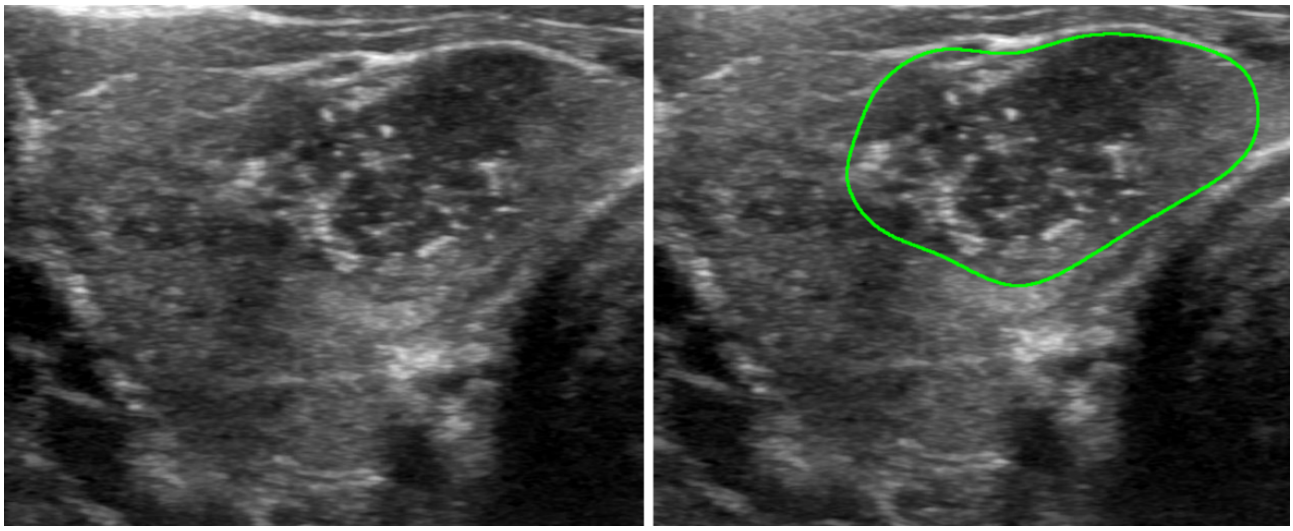

Intuitively, fine-tuning a general LLM with medical domain–specific knowledge should improve its performance for medical-specific tasks [33]. However, the nature of the data used for fine-tuning matters. LLaVA-Med did appear to have higher accuracy than LLaVA for all TI-RADS components evaluated in this study. Despite this, it was not more accurate than LLaVA in determining the overall TI-RADS classification level. To explain this, we noted that ultrasound images were not included in the LLaVA-Med fine-tuning figure-caption dataset. Furthermore, fine-tuning may have led to the poorer LLaVA-Med model consistency observed, as it would have been forced to sample from a probability distribution more skewed toward information irrelevant to its task and therefore produce a wider range of responses as a result.

Finally, prompt engineering significantly improved the consistency of output across all categories, including the overall TI-RADS classification level. This improvement was consistently observed regardless of the model used. This suggests that prompt engineering can still be a useful technique to optimize the output of "higher-performance" current-generation commercial models and should still be considered as a simple optimization technique in studies performed using these models.

However, prompt engineering had a mixed effect on accuracy. Furthermore, in the subgroup analysis on o3 output alone, there was no difference in accuracy, suggesting that the main effect of prompt engineering on accuracy was only seen in the open-source models. Prompt engineering in text-only LLMs can be highly effective as the text prompt is the only determinant of the output [34]. In this case, the main limiting step in accuracy was likely not the knowledge of the individual categories of the TI-RADS classification system, but the knowledge of what the categories meant. To illustrate this, a human trainee looking at a thyroid nodule image for the first time would not be expected to know the meaning of "hyperechogenic" or "hypoechogenic." He or she could refer to sample images showing the differences between these and extrapolate these observations to classify other nodules. Similarly, it might be unrealistic to expect the base vision-language model to intrinsically understand and translate the meaning of these classifiers.

Conversely, as O3 has a parameter count that is likely a few orders of magnitude higher than that of the open-source model (exact count not disclosed by OpenAI), it is conceivable that this "knowledge" is part of its pretraining set, hence also contributing to the better performance. Expanding on that hypothesis, a possible approach to further improve the output of smaller models could be to use few-shot prompting to provide it with the requisite examples to assist in subsequent classification. However, there are 19 different classifiers across the 5 TI-RADS components, and providing 19 example images in every prompt is costly and inefficient. Further work is required to explore the development of more efficient techniques of few-shot prompting to improve model performance for visual classification tasks, particularly for complex risk scores like this one.

This study has some limitations. LLM output is stochastic, or random, due to the inherent randomness of sampling techniques (such as top-k or nucleus sampling) used to derive the output. Stochasticity of output can be controlled by setting the "temperature" of the model, which is a measure ranging from 0 to 1 with 0 representing the least stochastic and 1 representing the most stochastic output ranges. We did not vary the base temperature in this study, as part of the objective was to measure the intrinsic model consistency. Using a low-to-zero temperature in future approaches is likely to improve consistency and consequently reproducibility of results.

Furthermore, the models used were only able to parse static images and not cine-clip images. It is arguable that certain features, such as punctate echogenic foci, are better appreciated on cine-clip images. However, real-world radiologists still appear to perform more consistently on static images rather than cine-clip images [35]. Therefore, the extent to which this limitation affects performance is unclear and should be evaluated further in other studies.

XSL•FO
RenderX

Next, the use of AI in health care is desirable because it may help to improve automation and productivity. Human calculations were still required in some cases in this study, which may have inflated the performance of the LLM. Rather than relying on the LLM alone to perform calculations, providing simple arithmetic tools to the LLM in an agentic framework may be an alternative solution to overcome this limitation.

Furthermore, human labeling was still used as an optimization strategy in this study. However, recent studies have explored the combination of LLM-generated prompts with a vision foundation model to perform zero-shot image segmentation, such as in the text-visual-prompt segment anything model (TV-SAM) algorithm incorporating GPT-4, the grounded language-imaging pre-training model, and the segment anything vision language model [36]. Image segmentation using such an algorithm, followed by interpretation of the segmented image with a classical multimodal LLM, may be a promising way to deliver a truly integrated end-to-end automated workflow in clinical practice and merits further exploration.

In addition, the test set used for evaluation in this study was from a single open-source dataset, which had a class imbalance between benign and malignant nodules (175 vs. 17, respectively). While this approximates the real-life distribution of thyroid nodules, whereby the vast majority are benign, it is possible that model performance on malignant cases may have been underpowered. Exploring these factors is beyond the scope of this exploratory study. This should be evaluated further in follow-on studies, which can also explore the generalizability to other datasets.

Finally, explainability (or the lack thereof) is a frequent criticism of end-to-end machine learning solutions, particularly in medicine, where safety and generalizability are concerns [8]. One method to improve explainability is to ask the LLM for explanations of how its output was derived [37]. We did not address this in our study as it was felt to be a moot point since overall performance was suboptimal.

LLMs are unique AI models because they are generally trained and can be easily repurposed for a variety of similar tasks. The major limitation for use in ultrasound image classification at present is in image processing and encoding, which can potentially be improved by relevant fine-tuning and improving feature extraction. This offers an alternative to resource-intensive model-building or fine-tuning and potentially retains flexibility for application in other imaging classification tasks. By addressing these limitations, we believe that computer-vision–assisted LLMs may eventually have the potential to augment human vision for visual-based classification tasks in medicine.

Conclusions

With the rapidly evolving nature of the field of AI and LLMs in health care, we believe that no one fixed model can be expected to balance all the competing demands of performance, cost, deployability, and security. The findings from this study highlight the potential benefits of simple processes, such as prompt engineering and basic image preprocessing, to model performance. We further demonstrate that some of these techniques remain applicable even with the release of new, more powerful models. Overall, we believe the findings from this study provide formative insights for developers and researchers for further work in this area.

Acknowledgments

This project is supported by Singhealth Duke-NUS AMC under the Clinical & Systems Innovation (Main) Grant (grant 03/FY2024/P2/03-A125).

Authors' Contributions

The authors confirm contribution to the paper as follows. Study conception and design were contributed by GGRS, TJH, and CCL. Data collection was contributed by XY. Analysis and interpretation of results was contributed by XY, GGRS, DYZL, and JYMT. Draft manuscript preparation was managed by GGRS. Reviewing manuscript was handled by XY, DYZL, JYMT, TJH, and CCL. All authors reviewed the results and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Categorical comparison of median model-generated scores against human-rated scores for each of the TI-RADS components and the overall TI-RADS classification. (A) Output from LLaVA and LLaVA-Med. (B) Output from OpenAI o3. (C) Output from all three models combined.

[PNG File, 50 KB - formative_v9i1e70863_app1.png]

References

1. Dean DS, Gharib H. Epidemiology of thyroid nodules. Best Pract Res Clin Endocrinol Metab 2008 Dec;22(6):901-911. [doi: 10.1016/j.beem.2008.09.019] [Medline: 19041821]

XSL•FO
RenderX

2. Lee JY, Baek JH, Ha EJ, et al. 2020 imaging guidelines for thyroid nodules and differentiated thyroid cancer: Korean Society of Thyroid Radiology. Korean J Radiol 2021 May;22(5):840-860. [doi: 10.3348/kjr.2020.0578] [Medline: 33660459]
3. Haugen BR, Alexander EK, Bible KC, et al. 2015 American Thyroid Association management guidelines for adult patients with thyroid nodules and differentiated thyroid cancer: The American Thyroid Association Guidelines Task Force on thyroid nodules and differentiated thyroid cancer. Thyroid 2016 Jan;26(1):1-133. [doi: 10.1089/thy.2015.0020] [Medline: 26462967]
4. Tessler FN, Middleton WD, Grant EG, et al. ACR Thyroid Imaging, Reporting and Data System (TI-RADS): White Paper of the ACR TI-RADS Committee. J Am Coll Radiol 2017 May;14(5):587-595. [doi: 10.1016/j.jacr.2017.01.046] [Medline: 28372962]
5. Hoang JK, Middleton WD, Farjat AE, et al. Interobserver variability of sonographic features used in the American College of Radiology Thyroid Imaging Reporting and Data System. AJR Am J Roentgenol 2018 Jul;211(1):162-167. [doi: 10.2214/AJR.17.19192] [Medline: 29702015]
6. Grani G, Lamartina L, Cantisani V, Maranghi M, Lucia P, Durante C. Interobserver agreement of various thyroid imaging reporting and data systems. Endocr Connect 2018 Jan;7(1):1-7. [doi: 10.1530/EC-17-0336] [Medline: 29196301]
7. Li W, Wang Y, Wen J, Zhang L, Sun Y. Diagnostic performance of American College of Radiology TI-RADS: a systematic review and meta-analysis. AJR Am J Roentgenol 2021 Jan;216(1):38-47. [doi: 10.2214/AJR.19.22691] [Medline: 32603229]
8. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023 Aug;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
9. Maitland A, Fowkes R, Maitland S. Can ChatGPT pass the MRCP (UK) written examinations? Analysis of performance and errors using a clinical decision-reasoning framework. BMJ Open 2024 Mar 15;14(3):e080558. [doi: 10.1136/bmjopen-2023-080558] [Medline: 38490655]
10. Li Q, Yang X, Wang H, et al. From beginner to expert: modeling medical knowledge into general llms. arXiv. Preprint posted online on 2023 URL: http://arxiv.org/abs/2312.01040 [accessed 2025-08-11]
11. Sandmann S, Riepenhausen S, Plagwitz L, Varghese J. Systematic analysis of ChatGPT, Google search and Llama 2 for clinical decision support tasks. Nat Commun 2024 Mar 6;15(1):2050. [doi: 10.1038/s41467-024-46411-8] [Medline: 38448475]
12. Liu H, Li C, Wu Q, Lee YJ. Visual instruction tuning. arXiv. Preprint posted online on 2023 URL: http://arxiv.org/abs/2304.08485 [accessed 2025-08-11]
13. Li C, Wong C, Zhang S, et al. LLaVA-med: training a large language-and-vision assistant for biomedicine in one day. arXiv. Preprint posted online on 2023 URL: http://arxiv.org/abs/2306.00890 [accessed 2025-08-11]
14. Tessler FN, Thomas J. Artificial intelligence for evaluation of thyroid nodules: a primer. Thyroid 2023 Feb;33(2):150-158. [doi: 10.1089/thy.2022.0560] [Medline: 36424829]
15. Wildman-Tobriner B, Taghi-Zadeh E, Mazurowski MA. Artificial intelligence (AI) tools for thyroid nodules on ultrasound, from the AJR Special Series on AI applications. AJR Am J Roentgenol 2022 Oct;219(4):1-8. [doi: 10.2214/AJR.22.27430] [Medline: 35383487]
16. Wu SH, Tong WJ, Li MD, et al. Collaborative enhancement of consistency and accuracy in US diagnosis of thyroid nodules using large language models. Radiology 2024 Mar;310(3):e232255. [doi: 10.1148/radiol.232255] [Medline: 38470237]
17. Cabezas E, Toro-Tobon D, Johnson T, et al. ChatGPT-4's accuracy in estimating thyroid nodule features and cancer risk from ultrasound images. Endocr Pract 2025 Jun;31(6):716-723. [doi: 10.1016/j.eprac.2025.03.008] [Medline: 40139461]
18. Qiu J, Yuan W, Lam K. The application of multimodal large language models in medicine. Lancet Reg Health West Pac 2024 Apr;45:101048. [doi: 10.1016/j.lanwpc.2024.101048] [Medline: 38524685]
19. Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. J Med Internet Res 2023 Oct 4;25(1):e50638. [doi: 10.2196/50638] [Medline: 37792434]
20. Han T, Adams LC, Nebelung S, Kather JN, Bressem KK, Truhn D. Multimodal large language models are generalist medical image interpreters. Health Informatics. . [doi: 10.1101/2023.12.21.23300146]
21. Zhang S, Dong L, Li X, et al. Instruction tuning for large language models: a survey. arXiv. Preprint posted online on 2023 URL: http://arxiv.org/abs/2308.10792 [accessed 2025-08-11]
22. Introducing o3 and o4-mini. OpenAI. URL: https://openai.com/index/introducing-o3-and-o4-mini [accessed 2025-07-30]
23. Yamashita R, Kapoor T, Alam MN, et al. Toward reduction in false-positive thyroid nodule biopsies with a deep learning-based risk stratification system using US cine-clip images. Radiol Artif Intell 2022 May;4(3):e210174. [doi: 10.1148/ryai.210174] [Medline: 35652118]
24. Imani S, Du L, Shrivastava H. MathPrompter: mathematical reasoning using large language models. arXiv. Preprint posted online on 2023 URL: http://arxiv.org/abs/2303.05398 [accessed 2025-08-11]
25. See YKC, Lim KSA, Au WY, Chia SYC, Fan X, Li ZK. The use of large language models in ophthalmology: a scoping review on current use-cases and considerations for future works in this field. BDCC 2025;9(6):151. [doi: 10.3390/bdcc9060151]
26. Rydzewski NR, Dinakaran D, Zhao SG, et al. Comparative evaluation of LLMs in clinical oncology. NEJM AI 2024 May;1(5). [doi: 10.1056/aioa2300151] [Medline: 39131700]
27. Huang J, Yang R, Huang X, et al. Feasibility of large language models for CEUS LI-RADS categorization of small liver nodules in patients at risk for hepatocellular carcinoma. Front Oncol 2024;14. [doi: 10.3389/fonc.2024.1513608]

XSL•FO
RenderX

28. Balduzzi A, Pastena M, Tondato S, et al. Exploring chatbot applications in pancreatic disease treatment: potential and pitfalls. Scierxiv. Preprint posted online on 2025. [doi: 10.20517/scierxiv202506.040.v1]
29. Ríos-Hoyo A, Shan NL, Li A, Pearson AT, Pusztai L, Howard FM. Evaluation of large language models as a diagnostic aid for complex medical cases. Front Med (Lausanne) 2024;11:1380148. [doi: 10.3389/fmed.2024.1380148] [Medline: 38966538]
30. Qi S, Cao Z, Rao J, Wang L, Xiao J, Wang X. What is the limitation of multimodal LLMs? A deeper look into multimodal LLMs through prompt probing. Inf Process Manag 2023 Nov;60(6):103510. [doi: 10.1016/j.ipm.2023.103510]
31. Lakhani P. The importance of image resolution in building deep learning models for medical imaging. Radiol Artif Intell 2020 Jan;2(1):e190177. [doi: 10.1148/ryai.2019190177] [Medline: 33939779]
32. Thinking with images. OpenAI. URL: https://openai.com/index/thinking-with-images [accessed 2025-07-30]
33. Singhal K, Tu T, Gottweis J, et al. Towards expert-level medical question answering with large language models. arXiv. Preprint posted online on 2023 URL: http://arxiv.org/abs/2305.09617 [accessed 2025-08-11]
34. Maharjan J, Garikipati A, Singh NP, et al. OpenMedLM: prompt engineering can out-perform fine-tuning in medical question-answering with open-source large language models. Sci Rep 2024;14(1). [doi: 10.1038/s41598-024-64827-6]
35. Schenke SA, Petersen M, Görges R, et al. Interobserver agreement in ultrasound risk stratification systems for thyroid nodules on static images versus cine-loop video sequences. Diagnostics (Basel) 2024 Sep 26;14(19):2138. [doi: 10.3390/diagnostics14192138] [Medline: 39410542]
36. Jiang Z, Cheng D, Qin Z, et al. TV-SAM: increasing zero-shot segmentation performance on multimodal medical images using GPT-4 generated descriptive prompts without human annotation. Big Data Min Anal 2024;7(4):1199-1211. [doi: 10.26599/BDMA.2024.9020058]
37. Zhao H, Chen H, Yang F, et al. Explainability for large language models: a survey. arXiv 2023 [FREE Full text]

Abbreviations

ACR TI-RADS: American College of Radiology Thyroid Imaging Reporting and Data System
AI: artificial intelligence
LLaVA: Large Language and Visual Assistant
LLaVA-Med: Large Language and Vision Assistant for bioMedicine
LLM: large language model
ROI: regions of interest
TI-RADS: Thyroid Imaging Reporting and Data System

Edited by J Sarvestan; submitted 03.01.25; peer-reviewed by X Du, Z Jiang; revised version received 19.06.25; accepted 20.06.25; published 19.08.25.

Please cite as:
Sng GGR, Xiang Y, Lim DYZ, Tung JYM, Tan JH, Chng CL
A Multimodal Large Language Model as an End-to-End Classifier of Thyroid Nodule Malignancy Risk: Usability Study
JMIR Form Res 2025;9:e70863
URL: https://formative.jmir.org/2025/1/e70863
doi:10.2196/70863

© Gerald Gui Ren Sng, Yi Xiang, Daniel Yan Zheng Lim, Joshua Yi Min Tung, Jen Hong Tan, Chiaw Ling Chng. Originally published in JMIR Formative Research (https://formative.jmir.org), 19.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Prediction of 1-Year Activity in Systemic Lupus Erythematosus: Hierarchical Machine Learning Approach

Livia Lilli[1,2], MSc; Laura Antenucci[1,2], MSc; Augusta Ortolan[3], MD; Silvia Laura Bosello[3], MD; Stefano Patarnello[1], MSc; Carlotta Masciocchi[1], PhD; Marco Gorini[4], MsC; Gabriella Castellino[4], MD; Alfredo Cesario[1,5], MD; Maria Antonietta D'Agostino[3], MD; Jacopo Lenkowicz[1], PhD

[1]Fondazione Policlinico Universitario Agostino Gemelli IRCCS, Largo Agostino Gemelli, 8, Rome, Italy

[2]Catholic University of the Sacred Heart, Rome, Italy

[3]Department of Rheumatology, Fondazione Policlinico Universitario Agostino Gemelli IRCCS, Rome, Italy

[4]AstraZeneca, Milano Innovation District (MIND), Milan, Italy

[5]Gemelli Digital Medicine & Health Srl, Rome, Italy

Corresponding Author:
Livia Lilli, MSc
Fondazione Policlinico Universitario Agostino Gemelli IRCCS, Largo Agostino Gemelli, 8, Rome, Italy

Abstract

Background: Systemic lupus erythematosus (SLE) is a chronic disease characterized by a broad spectrum of involved organs, including neurological, renal, and vascular domains, with disease activity manifesting through unpredictable patterns that vary across individuals and over time, making the prediction of activity events particularly challenging.

Objective: This paper proposes a hierarchical machine learning model to predict a 12-month SLE activity, defined as the occurrence of at least one event among SLE hospitalization, new organ-involved domain, and neurological, renal, or vascular manifestation within the following year. At each patient's visit, the model considers all the features at the current time point, the information about the patient's clinical history, and about its last 12 months, to predict the outcome for the next 12 months.

Methods: The study cohort consists of 262 patients with at least an outpatient visit and an SLE admission from 2012 to 2020, at the Italian Gemelli Hospital, comprising a retrospective longitudinal dataset of 5962 contacts. The data include demographics, laboratory, clinical features (eg, domain involvements and manifestations), treatments, and pathways (eg, contact types as outpatients, hospitalizations, day hospitals, and visit frequency). The variables consider 3 time ranges: features about the current contact and the last 12 months, and the previous patient's clinical history. The main model was developed by testing different machine learning approaches within a cross-validation setup. The predicted probability outputs were used in a risk stratification analysis, identifying 3 groups of predictions: strong, moderate, and mild. Mild samples were then passed through a second cascade model. The integration of the main model (applied to strong and moderate samples) with the cascade model (applied to mild contacts) forms our final hierarchical model.

Results: The hierarchical model, resulting from the ensemble of the main random forest and cascade decision tree, demonstrated enhanced performance, increasing the area under the receiver operating characteristic curve from 0.696 (95% CI 0.672 - 0.719) in the original main model to 0.743 (95% CI 0.717-0.769), particularly for specific patient characteristics. Through the application of explainable artificial intelligence methods, we also identified the key features that significantly influence the model's predictions. Among the 185 collected features, 15 emerged as the most impactful, including age at contact, response to therapy modifications, abnormal laboratory tests, and clinical manifestations. This analysis plays a crucial role in enhancing model transparency, which is essential for fostering the adoption of artificial intelligence in health care settings.

Conclusions: Our study introduces an explainable and reliable tool for predicting 1-year SLE activity, supporting physicians with an advanced decision-support system to improve patient management. The model identifies key features that may help characterize patient phenotypes, enabling personalized treatment plans and better outcomes. In addition, the methodology can be generalized for predictive analytics in other chronic autoimmune diseases.

(JMIR Form Res 2025;9:e70200) doi:10.2196/70200

KEYWORDS

systemic lupus erythematosus; machine learning; hierarchical model; ensemble approach; explainable artificial intelligence

XSL•FO
RenderX

Introduction

Background

Systemic lupus erythematosus (SLE) is a complex and heterogeneous autoimmune disease characterized by a high heterogeneity of clinical phenotypes with a wide range of involved organ domains and clinical manifestations. Its nature presents significant challenges in patient management, as activity events can occur unpredictably and may simultaneously impact multiple organ systems [1]. Accurate prediction of SLE activity is crucial for optimizing treatment strategies and improving patient outcomes.

Conventional methods for predicting disease activity often rely on basic statistical and machine learning models [2], considering summary indices such as Systemic Lupus Erythematosus Disease Activity Index (SLEDAI), the principal indicators for the activation of a patient with SLE [3,4].

Furthermore, recent advancements in machine learning offer new opportunities to enhance predictive capabilities by leveraging complex patterns within clinical data. In particular, hierarchical machine learning models have been implemented in various medical applications by combining the strengths of multiple algorithms to improve prediction accuracy [5,6]. In addition, explainable artificial intelligence (XAI) techniques are widely used to analyze the most predictive features [7-9].

In the context of SLE, there is growing interest in predictive tools that combine accurate use of longitudinal patient data with clinically interpretable outputs to better support routine decision-making; however, current approaches do not always integrate temporal information effectively or provide transparent reasoning behind predictions. In response to these challenges, this paper introduces a novel hierarchical machine learning model specifically designed to predict the occurrence of SLE activity within a 12-month time frame. The model is based on comprehensive longitudinal data of patients with SLE collected inside the SLE Data Mart of the Italian Gemelli Hospital in Rome, which consists of a comprehensive collection of structured data of patients with SLE, including demographic, laboratory results, treatments, involved domains, and clinical manifestations. At each patient's contact, the model considers all the features at the current time point, and the information about the clinical history of the patient and about its last 12 months, to predict the outcome for the next 12 months (Figure 1).

The work consists first of developing a main random forest model to categorize patients' contacts into strong, moderate, and mild risk groups based on prediction confidence. To improve the accuracy of mild predictions, a cascade decision tree model is integrated, whose ensemble generates the final hierarchical model. The model's performance is evaluated against a testing set, and its predictivity is assessed through metrics, such as the area under the receiver operating characteristic curve (AUC). In addition, we aim to provide physicians with an explainable decision-support tool by identifying key predictive features through XAI techniques, including age at contact, therapy changes, abnormal laboratory values, involved domains, and specific manifestations. The checklist for the guidelines "Machine Learning Predictive Model for Biomedical Data," which were used to structure and report this study, is provided in Checklist 1.

Figure 1. The model predicts the occurrence of a systemic lupus erythematosus activity event for a patient at a certain contact in the next 12 months. For the prediction, it takes as inputs the patient's features about the current contact, and the information about its clinical history (past features) and the last 12 months (last features). N/R/V: neurologic or renal or vascular.

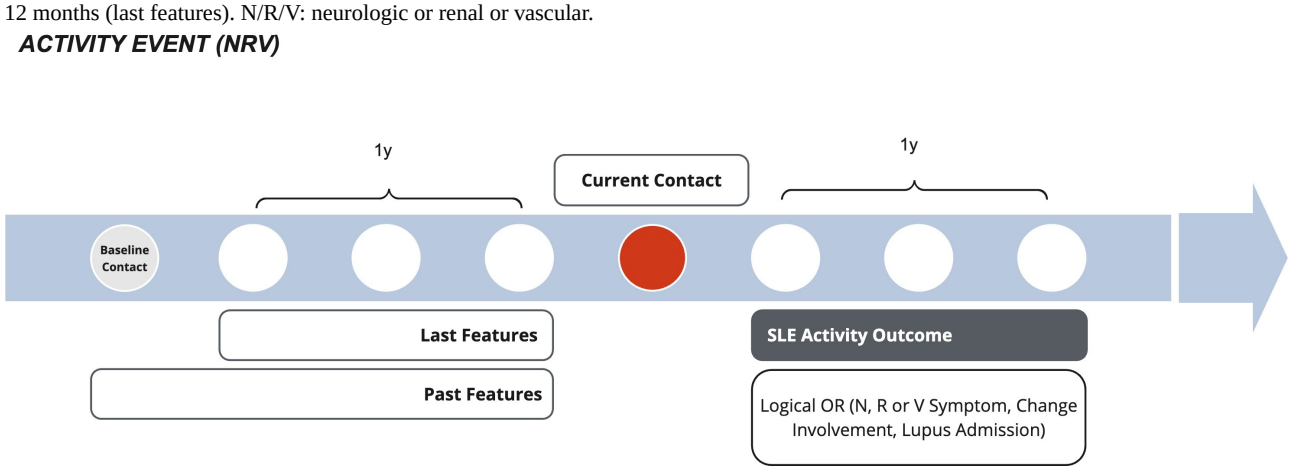

Related Works

Several studies explored the implementation of machine learning models in the SLE domain [2]. Among the most inherent to our purpose, there is the approach previously proposed by Ceccarelli et al [10], where SLE-related clinical and laboratory manifestations are used to predict the comprehensive disease control, an outcome based on the achievement of remission and the absence of damage progression in the next time contact. In this work, patients are analyzed along their history, and data are collected starting from the baseline contact, and then every 12 months. In the work of Alves et al [3], the concept of SLE activity (expressed in terms of SLEDAI) is presented. In particular, they proposed a multivariable ordinal regression model for the prediction of disease activity in terms of SLEDAI ranges, for a specific encounter. Also, in this paper, patients have a longitudinal history of contacts. In addition, a real-world dataset of clinical reports was used, where the explanatory features were extracted and used as inputs to the model. Furthermore, Hoi et al [4] developed a ranking model for the

XSL•FO
RenderX

prediction of the high disease activity outcome, based on the SLEDAI ranges. In this case, demographic and laboratory variables were used as inputs.

The study by Reddy and Delen [11] proposed a model for predicting the hospital readmission of patients with SLE in 30 days, by using deep-learning models for extracting longitudinal relationships inside electronic health records (EHRs). Furthermore, the study of Ceccarelli et al [12] worked on the prediction of the chronic damage defined in terms of the SLICC/ACR Damage Index, through the usage of Recurrent Neural Networks. The paper by Chen et al [13] is instead about the development of a machine learning model for the prediction of renal flare 5 years after remission, with the integration of a risk stratification model. Finally, a more recent study by Touma et al [14] shows the implementation of historical models for long-term outcomes to identify predictors of disease activity, flare rate, glucocorticoid use, organ damage, and mortality beyond the first year of treatment.

There are a few other studies that implemented machine learning and deep learning approaches for prediction modeling in the rheumatologic domain [15-18]. While for the implementation of XAI in the analysis of features' impact, there are several recent papers [19,20], where Shapley Additive Explanations (SHAP) [21] module is integrated in the predictive flow for the prediction of clinical outcomes.

Methods

Data Collection

The SLE data mart of the Italian Gemelli Hospital was used as an input dataset for this study. It consists of an extensive collection of structured information about patients with SLE observed along their longitudinal history at the Gemelli Hospital. The data mart considers a cohort of 262 patients with at least an outpatient visit and a hospitalization filtered by the *ICD-9* (*International Classification of Diseases*, *Ninth Revision*) codes 710.0 and 695.4, selected based on having had at least 1 such admission between 2012 and 2020, and with a minimum of 1 year of follow-up. Although the selection window is 2012 - 2020, the complete dataset spans a broader range from 2005 to 2022, to capture the full longitudinal clinical history available for each patient. The baseline is defined as the first recorded contact at the hospital, within this extended time frame.

In this cohort, 88% (230/262) of patients are female, with an age at baseline of 43 years, a follow-up of 6 years, and 16 contacts, in terms of median (Table 1). Patient observations occur during outpatient visits, hospital admissions, and day hospitals, for a total of 5962 contacts. Data are about demographics, treatments, clinical manifestations, involved domains, and laboratory and come from both structured and unstructured data sources. Data mining and natural language processing techniques were used for extracting and managing all the information coming from structured and unstructured sources, formatting them into the final SLE data mart (details about the NLP methodology can be found in our previous work [22]).

As an example of how variables were extracted, there are the organ domain involvements, the key features that primarily characterize patients with SLE. They are extracted through NLP from the textual EHRs, where the physician in charge usually reports the SLE diagnosis, specifying the organs affected. Once an organ involvement is recorded, it is considered persistent over time, as it reflects the chronic nature of organ impact from lupus for that specific patient. In our retrospective dataset, we distinguish among 8 different organ domains: articular, cutaneous, hematological, neurological, renal, vascular, serosal, and systemic. We define a "new involvement" as the first occurrence of an organ being involved that was not previously documented in the patient's medical history. This allows us to accurately capture the progression or change in disease manifestation over time.

Table . Characteristics of the study cohort. For categorical variables, occurrences in absolute and percentage format are reported, while for the numerical variables, median is shown with first and third quartiles.

Characteristic	All patients (N=262)
Female, n (%)	230 (88)
Age at Baseline, median (IQR)	43 (33-54)
Follow-up (y), median (IQR)	6 (3-8)
Contacts (all types), median (IQR)	16 (10-24)
Admission contacts, median (IQR)	1 (0-2)
Outpatient contacts, median (IQR)	12 (7-19)
Day hospital contacts, median (IQR)	0 (0-2)

Missing Data

During data processing, the issue of missing values affected 2 types of variables: treatments and laboratory data. To address these gaps, we adopted different approaches for each variable type to ensure data consistency and completeness. For laboratory data, missing values are not directly reported as numerical values in the input dataset. Instead, 2 binary variables are used to represent each laboratory measurement: "normal range" and "out of range." When values are missing, both variables are automatically set to "0," indicating the absence of recorded information. In such cases, there is no recorded evidence to determine whether the laboratory value was normal or out of range. For treatment data, missing therapies are imputed with the therapy information from the previous visit. This approach

XSL•FO
RenderX

reflects the clinical practice of assuming continuity of therapy unless explicitly stated otherwise.

Outcome Definition

The hierarchical model aims to predict the occurrence of an activity event in the following 12 months of a patient with SLE. The activity event is a binary variable, and it was defined through a logical OR combination of the 3 main events that represent a progression or a clinically relevant modification in SLE: (1) the manifestation of a symptom in the neurologic, renal, or vascular domains [23]; (2) the addition of a new involved domain; and (3) a hospitalization filtered by *ICD-9* codes 710.0 and 695.4. To enhance the relevance of the outcome, hospitalizations were further restricted to those occurring in departments more closely associated with SLE disease activity (ie, hematology, nephrology, dermatology, neurology, and angiology), excluding less relevant units such as infectious diseases and orthopedics. In addition, hospitalizations for continuation of cyclophosphamide or rituximab treatment were excluded, while only the first infusion was considered an activity event due to its typical association with high disease activity.

Then, if a patient presents at least 1 of the above events in the next 12 months with respect to the current observational contact, then the patient is labeled as positive to the outcome for that current time point. Of the overall dataset, 48% (2828/5962) of samples show a positive activity outcome. The prediction is performed independently at each contact, treating each visit separately while simultaneously building the longitudinal trajectory of the patient. This approach allows the model to generate multiple predictions across different time points for the same patient, effectively capturing the entire longitudinal history and enabling continuous monitoring of disease activity.

Input Features

The implementation of the hierarchical model considers an initial set of 181 SLE features available from the SLE data mart. The scope was to consider all the variables that clinically describe the patient, by providing SLE information about the current contact (current features), the clinical history (past features), and the last 12 months (last features). A feature engineering process was implemented in order to make all the variables processable by the model. Then all the features are binary or integers, depending on the type of information. A detailed guide on features' descriptions and types is provided in Multimedia Appendix 1, while preliminary statistics for all the variables can be found in Multimedia Appendix 2.

A univariate feature selection was performed in order to keep just the most significant features compared to the outcome for the model training. We leveraged the chi-square and the Mann-Whitney *U* tests for categorical and numerical variables, respectively. At the end of the statistical analysis, we kept all the features with a significant *P* value under .05, using them as inputs to the model for training and inference.

Main Model

The predictive model consists of the implementation of a Grid Search algorithm over multiple machine learning models, that are, logistic regression, decision tree, random forest, and XGBoost. At each iteration for a specific model, the algorithm tests several combinations of hyperparameters, performing a k-fold cross-validation, with 5 folds. At the end of the process, the model with the set of hyperparameters with the highest AUC performance was selected as the main model for this study. This method ensures that the chosen model is not only optimal but also robust and generalizable to new, unseen data.

Only 30% of patients were used as the testing set for the inference phase, while the remaining 70% represents our training set. Then contacts of a patient all belong to the same split set. The splitting was implemented randomly, by verifying two conditions: (1) that the outcome was balanced over both train and test sets, (2) and that patients were represented in a balanced way across the sets (ie, we ensure that 70% of training patients corresponds to 70% of contacts approximately). The random split was repeated until both the above conditions were met. We then got a training set of 183 patients and 4117 contacts, and a testing set of 79 patients and 1845 contacts. Furthermore, for construction, the last and past features are null at the patient's baseline. We then removed the first contact of each patient from the training set to prevent the model from learning biases caused by these zero values, obtaining a training set of 3934 samples. This ensures the model focuses on meaningful patterns rather than the artificial absence of data in early contacts.

For the implementation, we used the Scikit-Learn Python library [24]. The results of the main model were analyzed in terms of features' explainability through XAI techniques, for debugging purposes. To this scope, the SHAP module was used [7].

Risk Stratification

The output probabilities of the main model were used to define the risk of a patient at a specific contact. Based on the positive or negative outcome, we defined 2 risk categories, high and low. Each category is then distinguished as strong, moderate, or mild, based on the predicted probability (PP) value: when it is very close to the prediction cutoff, the risk category is mild, and it tends to become moderate and then strong, moving away from the threshold.

Cascade Model

The previously described risk stratification was essential to identify patient subgroups at a given contact, categorizing them into high- and low-risk levels with varying degrees of predictability, ranging from mild to moderate and strong, based on prediction confidence. Mild contacts represent cases where the main model predicts the outcome with greater uncertainty due to a lack of sufficient discriminant information to make a confident decision. To enhance the overall performance, a cascade model was then ensembled, specifically trained on mild-risk samples. In agreement with physicians, a smaller set of features was used as inputs, such as demographic information about gender and age at contact and current treatments not used in the main model. We then trained a decision tree machine learning algorithm to get a tree of rules to be applied on the uncertain samples. From the resulting tree, we found the branches with the highest accuracy rate, and we used them to build up a rule-based algorithm to be applied after the main

XSL•FO
RenderX

model, just on the mild-risk contacts. The integration of both models, the main model applied to strong and moderate samples, and the cascade model applied to the mild samples, results in our final hierarchical model.

Ethical Considerations

The study conforms with European General Data Protection Regulation directives and consequent Italian legislation references (European Union Directive 2016/679 and under Italian Laws: Decreto Legislativo 196/2003, Decreto Legislativo 101/2018, and Autorizzazione Generale Garante 9/2016). In line with the General Data Protection Regulation directives and the designation of Policlinico Gemelli Hospital as National Center for Health Research and Care from the Italian Ministry of Health, observational studies do not require a specific request for patient consent. All technical procedures and security measures have been reviewed and approved in the context of Data Protection Impact Assessment from the hospital Data Privacy Unit and Data Protection Officer. The study protocol conforms to the ethical guidelines of the 1975 Declaration of Helsinki and has received approval from the Policlinico Gemelli Ethics Committee (Comitato Etico Territoriale Lazio Area 3, protocol 0034012/22).

Results

Feature-Based Explainability Analysis

For each contact, we considered a set of features regarding the current status of the patient (current features), the events of the last 12 months (last features), and its clinical history (past features), for a total of 181 variables. A selection algorithm was applied to the training set, and 125 variables resulted significantly associated with the activity outcome ($P<.05$). Given the large number of variables, only the most representative ones are highlighted for a preliminary overview (Table 2). Among the significant features, younger age at contact is strongly linked to a positive activity outcome in the next 12 months. Furthermore, the addition of a new organ involvement emerged as a critical indicator of future disease activity. Also, the temporal delta between contacts, which calculates the number of days since the previous visit, is significantly different; it indicates more frequent visits for patients with an activity event, making it a useful variable to account for irregular follow-up patterns. Finally, the treatment pathways, expressed in terms of therapy changes, show a strong correlation with the outcome; both the presence of a step-down therapy change in the patient's history and the frequency of step-down therapy changes are significantly associated with disease activity. These findings underline the importance of identifying key clinical predictors to improve outcome monitoring.

The selected features were then given as input to the model, for the training and inference phases. During the model development, we applied XAI techniques to analyze features' importance in our predictive main model. To this scope, we leveraged the SHAP Python module, which quantifies the impact of each feature on the model's predictions, providing insights into the contribution and interaction of features with respect to the outcome.

The Beeswarm plot was generated by the model during the outcome prediction over the testing contacts (Figure 2). It represents both the distribution and magnitude of feature effects, highlighting the features with the most significant impact and their direction (positive or negative) on the model's output. As an example, for young patients, the model prediction tends to be positive to an activity event. The blue data points correspond to low values of the age at contact variable, and they are all distributed on the right side of the y-axis, with positive SHAP values.

We also conducted a SHAP longitudinal analysis on the individual patients, through the Force plot, where the aggregated predictive impact of the features over the patient's longitudinal history is displayed (Figure 3). Given the longitudinal representation of a patient, we then conducted an in-depth analysis of each time point, looking at the detailed impact of each feature on the related prediction. In Figure 3, true positive prediction is reported, where the model tends to predict the presence of an activity outcome in the following 12 months, mostly based on the out-of-range albuminuria. Also, a true negative contact is shown, where the step-down therapy change pushes the prediction to be negative, together with normal albuminuria and the high number of outpatient visits in the last year.

XSL•FO
RenderX

Table . Feature selection: list of the most representative features of the 125 selected by the algorithm. For categorical variables, occurrences in absolute and percentage format are reported, while for the numerical variables, median is shown with first and third quartiles.

Feature	Activity event (n=1772)	No activity event (n=2162)	*P* value
Current contact			
Age at contact, median (IQR)	44.0 (36.0-53.0)	46.0 (37.0-55.0)	<.001
Outpatient visit, n (%)	1217 (68.68)	1779 (82.28)	<.001
Day hospital, n (%)	413 (23.31)	331 (15.31)	<.001
Admission, n (%)	152 (8.58)	57 (2.64)	<.001
Delta contacts, median (IQR)	77.0 (33.0-136.0)	90.0 (32.0-153.0)	.044
New involvement, n (%)	204 (11.51)	113 (5.23)	<.001
Last 12 months			
Total of outpatient visits, median (IQR)	2.0 (1.0-3.0)	3.0 (2.0-3.0)	<.001
Outpatient visit, n (%)	1370 (77.31)	1852 (85.66)	<.001
Day hospital contact, n (%)	509 (28.72)	482 (22.29)	<.001
Admission, n (%)	369 (20.82)	262 (12.12)	<.001
New involvement, n (%)	445 (25.11)	369 (17.07)	<.001
Total of new involvements, median (IQR)	0.0 (0.0-1.0)	0.0 (0.0-0.0)	<.001
History			
Outpatient visit, n (%)	1526 (86.12)	2018 (93.34)	<.001
Step-down therapy change, n (%)	832 (46.95)	1397 (64.62)	<.001
Total of step-down therapy changes, median (IQR)	0.0 (0.0-1.0)	1.0 (0.0-2.0)	<.001

XSL•FO
RenderX

Figure 2. SHAP Beeswarm plot describing impact and direction of each feature on the outcome prediction, ordered by importance. SHAP: Shapley Additive Explanations.

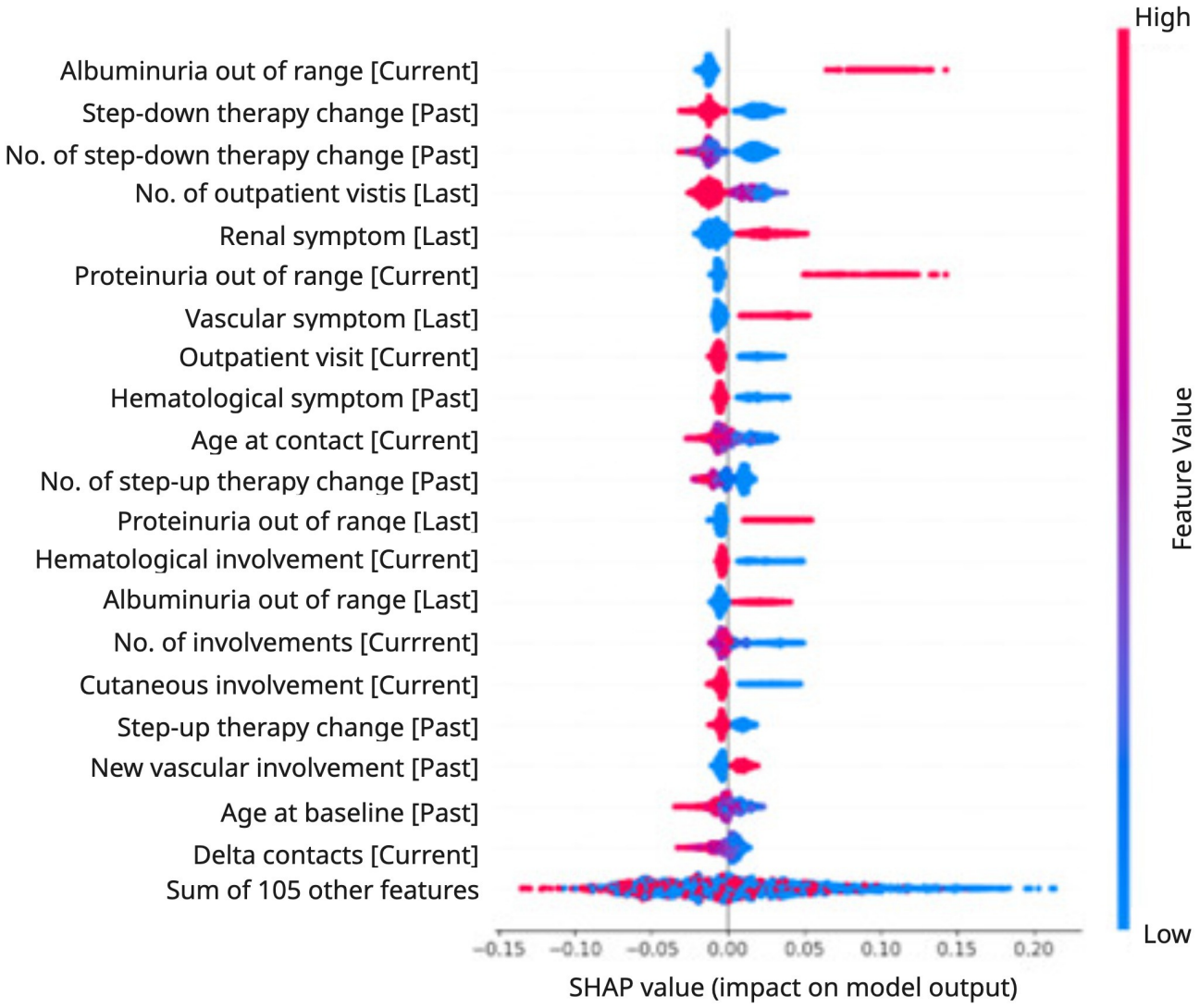

XSL•FO
RenderX

Figure 3. Longitudinal SHAP analysis for a specific patient. The top force plot shows the aggregated impact of the input features on the patient's prediction over time. At the bottom, the most impactful features for 2 contact visits are shown. In particular, the 2 samples present a true positive prediction (bottom-left) and a true negative prediction (bottom-right). TP: true positive; TN: true negative.

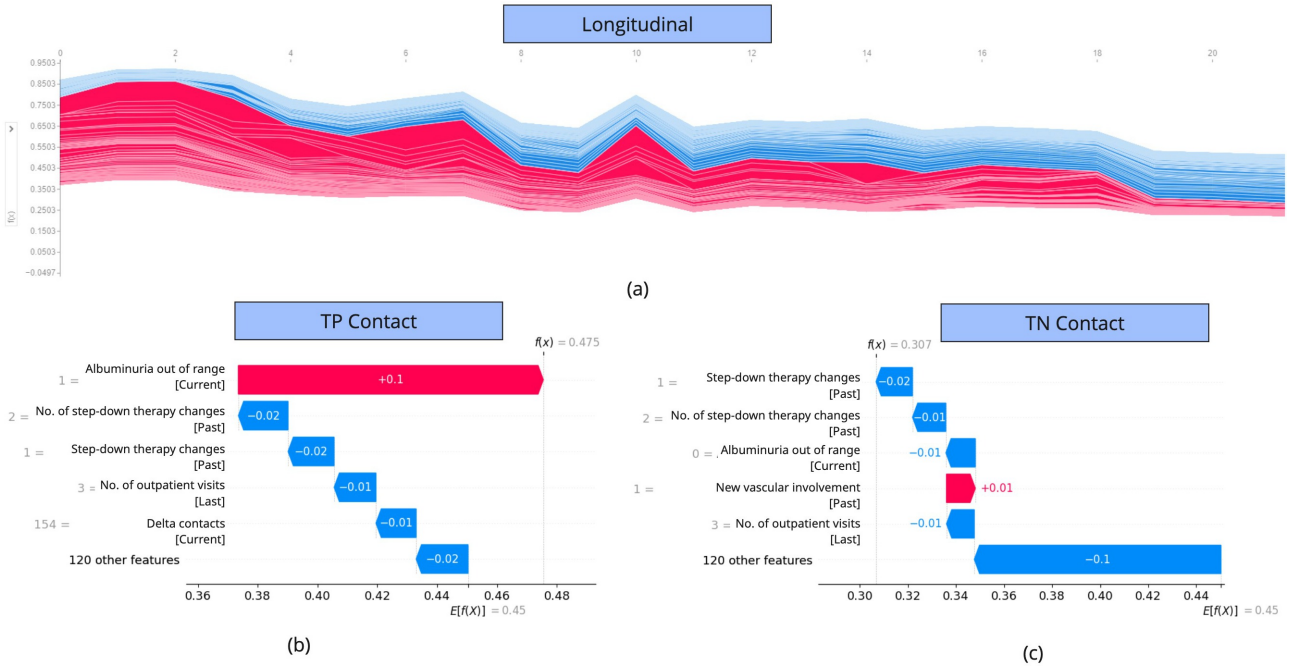

Risk Stratification

The risk stratification was performed considering the main model's PP values. In particular, high and low categories are distinguished by the prediction cutoff automatically computed during the training phase, which equals 0.45. For samples predicted with an activity event in the next 12 months (then with a PP>0.45), the risk is high; otherwise, the risk is low. Then, the level of predictability is chosen between strong, moderate, and mild, depending on whether the PP is far or near the prediction cutoff. For this purpose, we used the PP distribution. For high-risk predictions, we used the median and the first quartile of the true positives as thresholds, while for low-risk predictions, we relied on the median and the third quartile of the true negatives (Figure 4). From this, we got 32% of mild contacts (ie, 590 samples), where the model could not predict the outcome with a sufficiently high predictability (Figure 4).

Figure 4. Risk stratification thresholds (left) and risk distribution over the testing contacts (right). PP: predicted probability; TH: threshold; TN: true negative; TP: true positive; Q1: first quartile; Q3: third quartile.

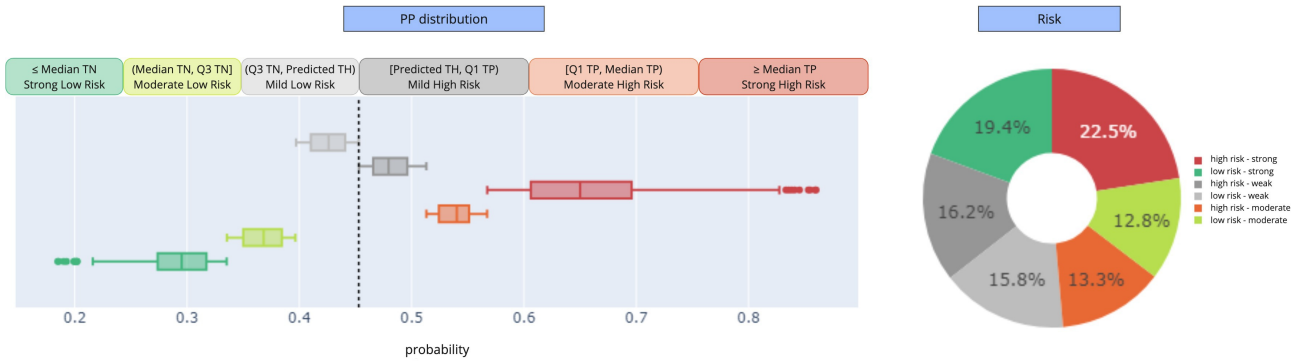

Hierarchical Model

The main model consisted of the machine learning algorithm, which returned the best AUC performances over the 1845 test samples, where random forest achieves an AUC=0.696, versus logistic regression (AUC=0.64), decision tree (AUC=0.653), and XGBoost (AUC=0.693). The hierarchical model was then built based on the combination of the main model with a cascade model applied over the mild-risk contacts. In particular, we trained a decision tree model on the activity outcome, over a training set of 545 mild contacts (we considered the same risk stratification over the training set), keeping the same split used in the main model. We then obtained the tree of rules applied over the evaluation set of 590 contacts. From this, we then focused on the branches with higher prediction accuracy (over 60%) and clinically approved by the physicians, considering these predictions as more confident (Multimedia Appendix 3). In particular, in these decisional paths, we found that male patients with an age between 31 and 71 years, not treated with glucocorticoids, can be considered as having a negative outcome. Furthermore, female patients in the age range of 31 - 64 years, who were treated only with antimalarials, also show a negative activity outcome. Then, for the subset of

XSL•FO
RenderX

patients following the above rules, we could associate a low risk with high reliability.

The main model's performance in terms of receiver operating characteristic curve and AUC metric is shown (Figure 5) when applied overall to the 1845 testing samples, with the exclusion of the mild predictions and only over the strong-risk contacts. The model then passes from AUC=0.696 (95% CI 0.672 - 0.719; number of contacts=1845) overall the test set, to AUC=0.742 (95% CI 0.715 - 0.769; number of contacts=1255), and AUC=0.784 (95% CI 0.752 - 0.817; number of contacts=773) when considering both moderate and strong contacts, as well as strong contacts only. Finally, the hierarchical model results are reported for overall the dataset and considering just the most accurate decisional paths, with AUC=0.692 (95% CI 0.667 - 0.716; number of contacts=1845) and AUC=0.743 (95% CI 0.717-0.769; number of contacts=1396), respectively.

Figure 5. Comparison of results. The first row shows the main model applied overall the dataset, on the strong and moderate contacts, and on the strong contacts. The second row is about the ensemble's results. In the bottom-left, the main model is applied over the strong and moderate contacts, with the cascade model implemented over the mild contacts; in the bottom-right, only the most accurate decisional path of the cascade model is considered. ROC: receiver operating characteristic curve.

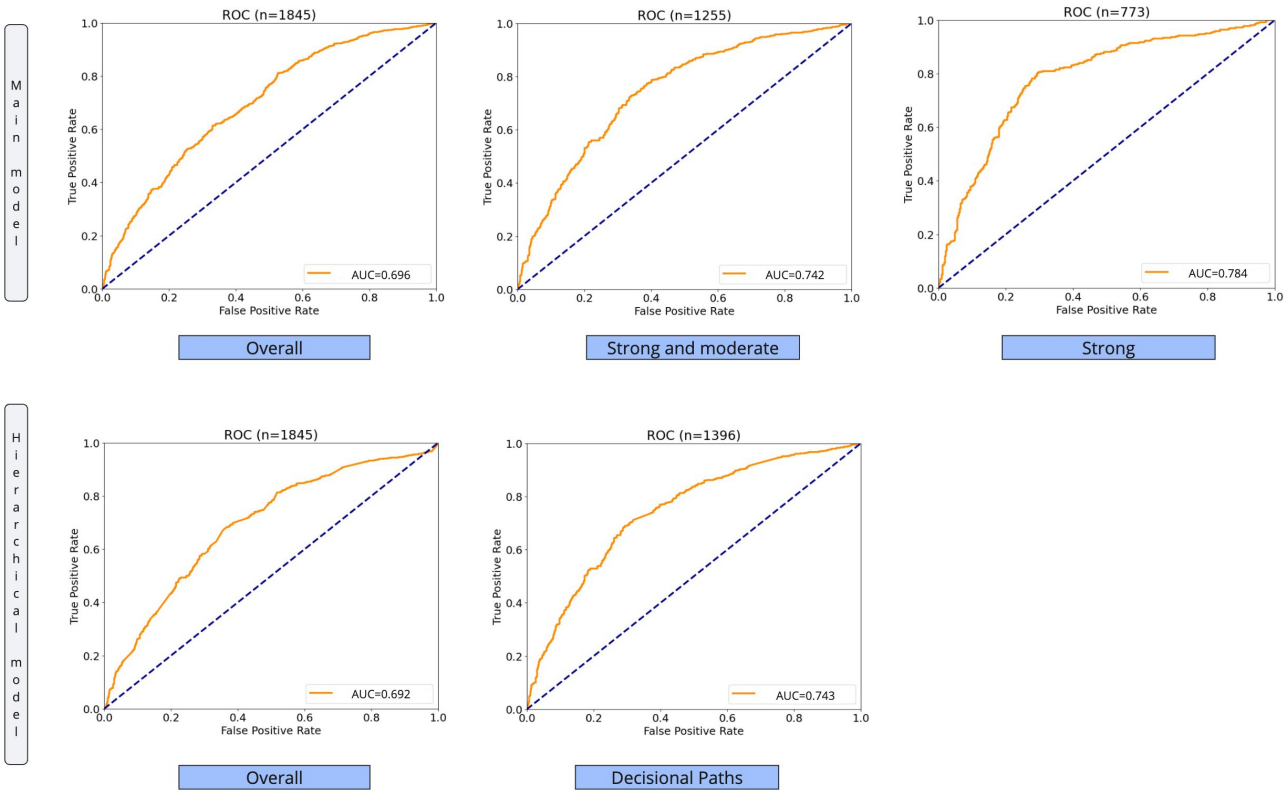

Discussion

Principal Results and Comparison With Previous Work

In this paper, we proposed an ensembled approach where multiple models are combined to predict the activity event of a patient with SLE in the next 12 months.

In literature, several works are related to the development of SLE predictive models [2-5], exploring the problem from multiple perspectives. These studies vary in terms of predicted outcomes, such as remission, chronic damage, or hospital readmission, and leverage different types of input data, including clinical manifestations, laboratory values, and treatments. Some also focus on modeling disease progression over time. Taken together, these studies have contributed valuable evidence to advance personalized diagnosis and treatment in SLE.

Our purpose consists of presenting a novel approach, developing a hierarchical model, that integrated with risk analysis and XAI techniques, returns a 1-year prediction of the SLE activity outcome. The definition of an activity event was established by our team of physicians, drawing on their clinical expertise and existing literature [23]. It consists of a combination of the most severe flare types and other severe events (such as hospitalization and new involved organs), identified from both clinical notes and structured data sources. This choice differs from the outcomes used in past works, such as comprehensive disease control, hospital readmission, and chronic damage [3,4,10-18].

Our main model was trained over a set of 125 features, automatically selected based on their significance to the outcome (Table 2). Features are about demographics, clinical manifestations, involved domains, laboratory, and therapy changes, mostly following the information used in the related previous cited works. However, in our case, the same information is reported for the current contact, the last 12 months, and the clinical history. Thus, we allow traditional machine learning models such as random forest to capture both the present and the past of a patient for each observed time point. This setup allows us to treat each contact independently, then conduct a contact-level prediction. In addition, this feature

XSL•FO
RenderX

framework, based on a real-world data collection from standard clinical practice, yields a set of predictors that are operationally accessible, making the model potentially replicable across different health care settings. As a future direction, this approach could be integrated with methodologies based on prospective data collection (eg, studies by Ceccarelli et al [10,12]) or derived from public datasets (eg, study by Reddy and Delen [11]) to further improve generalizability and support broader validation.

In the features' explainability analysis, we used the Python SHAP module in order to deeply analyze the features' impact during the prediction of the main model. This approach can be particularly useful in real-world clinical practice, as it provides a transparent understanding of the model's reasoning, supporting informed decision-making through an explainable and reliable tool. Then, we generated the SHAP Beeswarm plot (Figure 2), where the most predictive features are reported, with their impact on the activity event outcome. Many of the findings are already noted in the SLE literature, or derive from the outcome definition, such as the positive impact of proteinuria, albuminuria, young age, and renal symptoms [25,26]. Indeed, the plot shows that albuminuria and proteinuria out of range, the presence of renal symptoms, and young age push the model to predict an activity event for the patient in the following 12 months. However, we also got additional evidence related to the effect of therapy changes, outpatient contacts, temporal distance among contacts, and hematological and cutaneous involvements. Indeed, the plot shows that a high number of outpatient visits (as opposed to inpatient or day hospital admissions) in the last 12 months is associated with a protective effect on the outcome. This suggests that closer clinical monitoring may contribute to a better outcome, whereas inpatient or day hospital admissions could indicate a disease flare or the need for ongoing inpatient treatment. Furthermore, when no organs are involved in the hematological and cutaneous domains, the patient is more likely to have a positive outcome. Conversely, the longer the interval between contact visits, the higher the probability of remaining stable in the following 12 months. Finally, a step-down change in therapy in the past (such as the discontinuation of a biological immunosuppressant) is protective for a positive outcome. Our results are consistent with previous studies, particularly the studies by Ceccarelli et al [10,12], which also highlight hematological manifestations and renal symptoms as key features associated with increased severity, even though they rely on different sets of variables. In addition, a study by Reddy and Delen [11] identifies the time interval between hospital contacts as a predictive factor, but in this case against the risk of rehospitalization.

The risk stratification was then performed on the main model predictions, returning 32% mild-risk contacts (Figure 4), which we used to develop a cascade model. Then, the integration of the main model implemented over the strong and moderate predictions, with the cascade model applied over the mild samples, generated the final hierarchical model. The hierarchical model achieves an AUC of 0.743 when considering the samples following the most decisional paths. In terms of metric value, this result is lower than the AUC=0.784 of the main model applied over the strong contacts, and not so far from the AUC=0.742 of the main model implemented on the moderate and strong samples (Figure 5). However, the hierarchical model, with an almost high metric value, can cover a larger portion of predictable input data.

Results we achieved in terms of AUC reflect the current state of the art given by the previous works [10-12], where the implementation of predictive modeling on SLE using input features similar to ours generated an AUC between 0.70 and 0.77. In the works of Alves et al [3] and Hoi et al [4], results are higher, with metrics equal to 0.90 and 0.829, respectively. However, our variety of input features is larger; in the study by Alves et al [3], textual EHRs are processed through text mining techniques for extracting SLEDAI information, while in the study by Hoi et al [4], laboratory measures and demographic information are considered. Nevertheless, they could serve as inspiration for future work, following their strategy of using SLEDAI ranges as activity indicators, which we could not implement in this study due to the incomplete availability of SLEDAI information at our hospital. While SLEDAI is a widely recognized marker of disease activity, its predictive value for severity is not always consistent [3]. Therefore, further exploration of its role and potential integration with alternative indicators could be a valuable direction for future research.

Limitations

Our study was conducted on a filtered SLE cohort of the Gemelli Hospital, selecting only patients who had had at least an outpatient visit and an SLE hospitalization between 2012 and 2020, with a minimum of 1 year of follow-up. This approach inherently introduces a selection bias, as our model is trained on a cohort representing more severe cases. Future work could address this limitation by applying our model to a broader cohort that also includes outpatient-only cases, thereby reducing the current selection bias. In addition, extending our research to a multicentric scenario would be valuable for assessing the model's generalizability across different patient populations.

Conclusions

This study aims to provide a hierarchical machine learning approach, where the ensemble of multiple architectures and the usage of XAI tools and risk stratification analysis contribute to the development of an explainable model for the prediction of SLE activity in the next 12 months that could support physicians to modify the therapeutic approach.

Relevant features identified through the learning process are of different kinds, that are (1) demographics (eg, risk of severe events is reduced for older patients), (2) clinical (as an example, occurrence of renal symptoms in the past 12 months increases risks), and (3) treatment related (eg, patients who have shown good response to a therapy change from biological to conventional in their disease history are less exposed to severe event occurrence). From a predictability analysis, we found that our approach can achieve an AUC of 74.3% when applied under specific patients' characteristics. The results mostly reflect other related works in literature, always in the domain of SLE predictive modeling, but with the integration of an XAI tool, which makes the predictions explainable at feature level.

In conclusion, our hierarchical model offers accurate 12-month activity predictions along with an explainable framework for

XSL•FO
RenderX

result interpretation. This contributes meaningfully to predictive analytics in chronic autoimmune diseases, paving the way for future studies to build on our findings and advance patient care.

Acknowledgments

The project has been developed with the financial contribution of AstraZeneca, and it has been approved by the Institutional Review Board of the Fondazione Policlinico

Universitario Agostino Gemelli IRCCS (protocol 0034012/22).

This manuscript was prepared with language refinement assistance from ChatGPT-4o (OpenAI).

Data Availability

The dataset used in this study is not publicly available due to confidentiality reasons.

Conflicts of Interest

MG and GC are employees of AstraZeneca. AO received consultancy and/or speaker fees from Abbvie, Novartis, Johnson & Johnson, Eli Lilly, UCB, GSK, and AstraZeneca.

Multimedia Appendix 1
Detailed guide on input features, providing an overview of variable macro-categories, a comprehensive list of all features with definitions, types, and formats, and clear explanations of key concepts such as symptoms, flare, and complexity.
[DOCX File, 49 KB - formative_v9i1e70200_app1.docx]

Multimedia Appendix 2
General statistics of the input features about the current contact, the last 12 months, and the clinical history.
[DOCX File, 39 KB - formative_v9i1e70200_app2.docx]

Multimedia Appendix 3
Decisional paths of the cascade decision tree model, with the highest accuracy rate.
[PNG File, 82 KB - formative_v9i1e70200_app3.png]

Checklist 1
Checklist for Machine Learning Predictive Model for Biomedical Data guidelines.
[DOCX File, 26 KB - formative_v9i1e70200_app4.docx]

References

1. Fanouriakis A, Kostopoulou M, Andersen J, et al. EULAR recommendations for the management of systemic lupus erythematosus: 2023 update. Ann Rheum Dis 2024 Jan 2;83(1):15-29. [doi: 10.1136/ard-2023-224762] [Medline: 37827694]
2. Munguía-Realpozo P, Etchegaray-Morales I, Mendoza-Pinto C, et al. Current state and completeness of reporting clinical prediction models using machine learning in systemic lupus erythematosus: a systematic review. Autoimmun Rev 2023 May;22(5):103294. [doi: 10.1016/j.autrev.2023.103294] [Medline: 36791873]
3. Alves P, Bandaria J, Leavy MB, et al. Validation of a machine learning approach to estimate Systemic Lupus Erythematosus Disease Activity Index score categories and application in a real-world dataset. RMD Open 2021 May;7(2):e001586. [doi: 10.1136/rmdopen-2021-001586] [Medline: 34016712]
4. Hoi A, Nim HT, Koelmeyer R, et al. Algorithm for calculating high disease activity in SLE. Rheumatology (Sunnyvale) 2021 Sep 1;60(9):4291-4297. [doi: 10.1093/rheumatology/keab003]
5. Ma H, Sheng W, Li J, et al. A novel hierarchical machine learning model for hospital-acquired venous thromboembolism risk assessment among multiple-departments. J Biomed Inform 2021 Oct;122:103892. [doi: 10.1016/j.jbi.2021.103892] [Medline: 34454079]
6. Ershadi MM, Rise ZR, Niaki STA. A hierarchical machine learning model based on Glioblastoma patients' clinical, biomedical, and image data to analyze their treatment plans. Comput Biol Med 2022 Nov;150:106159. [doi: 10.1016/j.compbiomed.2022.106159] [Medline: 36257277]
7. Lundberg S, Lee SI. A unified approach to interpreting model predictions. arXiv. Preprint posted online on 2017 URL: https://arxiv.org/abs/1705.07874 [accessed 2025-08-14] [doi: 10.48550/arXiv.1705.07874]
8. Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2020 Jan;2(1):56-67. [doi: 10.1038/s42256-019-0138-9] [Medline: 32607472]

XSL•FO
RenderX

9. Lundberg SM, Nair B, Vavilala MS, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat Biomed Eng 2018 Oct;2(10):749-760. [doi: 10.1038/s41551-018-0304-0] [Medline: 31001455]
10. Ceccarelli F, Olivieri G, Sortino A, et al. Comprehensive disease control in systemic lupus erythematosus. Semin Arthritis Rheum 2021 Apr;51(2):404-408. [doi: 10.1016/j.semarthrit.2021.02.005] [Medline: 33652293]
11. Reddy BK, Delen D. Predicting hospital readmission for lupus patients: An RNN-LSTM-based deep-learning methodology. Comput Biol Med 2018 Oct 1;101:199-209. [doi: 10.1016/j.compbiomed.2018.08.029] [Medline: 30195164]
12. Ceccarelli F, Sciandrone M, Perricone C, et al. Prediction of chronic damage in systemic lupus erythematosus by using machine-learning models. PLoS ONE 2017;12(3):e0174200. [doi: 10.1371/journal.pone.0174200] [Medline: 28329014]
13. Chen Y, Huang S, Chen T, et al. Machine learning for prediction and risk stratification of lupus nephritis renal flare. Am J Nephrol 2021;52(2):152-160. [doi: 10.1159/000513566] [Medline: 33744876]
14. Touma Z, Kayaniyil S, Parackal A, et al. Modelling long-term outcomes for patients with systemic lupus erythematosus. Semin Arthritis Rheum 2024 Oct;68:152507. [doi: 10.1016/j.semarthrit.2024.152507] [Medline: 39029291]
15. Hu Z, Wu L, Lin Z, Liu X, Zhao C, Wu Z. Prevalence and associated factors of electrocardiogram abnormalities in patients with systemic lupus erythematosus: a machine learning study. Arthritis Care Res (Hoboken) 2022 Oct;74(10):1640-1648. [doi: 10.1002/acr.24612] [Medline: 33787068]
16. Chung CW, Hsiao TH, Huang CJ, et al. Machine learning approaches for the genomic prediction of rheumatoid arthritis and systemic lupus erythematosus. BioData Min 2021 Dec 11;14(1):52. [doi: 10.1186/s13040-021-00284-5] [Medline: 34895289]
17. Ceccarelli F, Sciandrone M, Perricone C, et al. Biomarkers of erosive arthritis in systemic lupus erythematosus: application of machine learning models. PLoS ONE 2018;13(12):e0207926. [doi: 10.1371/journal.pone.0207926] [Medline: 34895289]
18. Tang Y, Zhang W, Zhu M, et al. Lupus nephritis pathology prediction with clinical indices. Sci Rep 2018 Jul 6;8(1):10231. [doi: 10.1038/s41598-018-28611-7] [Medline: 29980727]
19. Caliandro P, Lenkowicz J, Reale G, et al. Artificial intelligence to predict individualized outcome of acute ischemic stroke patients: The SIBILLA project. Eur Stroke J 2024 Dec;9(4):1053-1062. [doi: 10.1177/23969873241253366] [Medline: 38778480]
20. Lee AR, Park H, Yoo A, Kim S, Sunwoo L, Yoo S. Risk prediction of emergency department visits in patients with lung cancer using machine learning: retrospective observational study. JMIR Med Inform 2023 Dec 6;11:e53058. [doi: 10.2196/53058] [Medline: 38055320]
21. Shap/shap. GitHub. URL: https://github.com/shap/shap [accessed 2025-08-05]
22. Lilli L, Bosello SL, Antenucci L, et al. A comprehensive natural language processing pipeline for the chronic lupus disease. In: Digital Health and Informatics Innovations for Sustainable Health Care Systems: IOS Press; 2024:909-913. [doi: 10.3233/SHTI240559] [Medline: 39176940]
23. Langham J, Barut V, Samnaliev M, et al. Disease severity, flares and treatment patterns in adults with systemic lupus erythematosus in the UK: a real-world observational retrospective cohort analysis. Rheumatol Adv Pract 2021;5(3):rkab061. [doi: 10.1093/rap/rkab061] [Medline: 34557623]
24. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. arXiv. Preprint posted online on 2012 URL: https://arxiv.org/abs/1201.0490 [accessed 2025-08-14] [doi: 10.48550/arXiv.1201.0490]
25. Feng H, Lu Q, Liu Y, et al. Risk factors of disease activity and renal damage in patients with systemic lupus erythematosus. Int Urol Nephrol 2024 Dec;56(12):3845-3855. [doi: 10.1007/s11255-024-04105-6] [Medline: 38937413]
26. Yip J, Aghdassi E, Su J, et al. Serum albumin as a marker for disease activity in patients with systemic lupus erythematosus. J Rheumatol 2010 Aug 1;37(8):1667-1672. [doi: 10.3899/jrheum.091028] [Medline: 20516026]

Abbreviations

AUC: area under the receiver operating characteristic curve
EHR: electronic health record
***ICD-9*:** *International Classification of Diseases, Ninth Revision*
PP: predicted probability
SHAP: Shapley Additive Explanations
SLE: systemic lupus erythematosus
SLEDAI: Systemic Lupus Erythematosus Disease Activity Index
XAI: explainable artificial intelligence

XSL•FO
RenderX

Edited by A Mavragani; submitted 17.12.24; peer-reviewed by A Mokbel, E Lee, H Kisaoglu; revised version received 09.05.25; accepted 12.05.25; published 22.08.25.

Please cite as:
Lilli L, Antenucci L, Ortolan A, Bosello SL, Patarnello S, Masciocchi C, Gorini M, Castellino G, Cesario A, D'Agostino MA, Lenkowicz J
Prediction of 1-Year Activity in Systemic Lupus Erythematosus: Hierarchical Machine Learning Approach
JMIR Form Res 2025;9:e70200
URL: https://formative.jmir.org/2025/1/e70200
doi:10.2196/70200

© Livia Lilli, Laura Antenucci, Augusta Ortolan, Silvia Laura Bosello, Stefano Patarnello, Carlotta Masciocchi, Marco Gorini, Gabriella Castellino, Alfredo Cesario, Maria Antonietta D'Agostino, Jacopo Lenkowicz. Originally published in JMIR Formative Research (https://formative.jmir.org), 22.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Evaluating Use Patterns of a Closed Electronic Nicotine Delivery System Among Adults in the United States Who Smoke Cigarettes Daily: 8-Week Actual Use Study

Steve Roulet[1], BSc; Claudia Kanitscheider[2], MA; Pierpaolo Magnani[1], BSc; Gerd Kallischnigg[3], MSc

[1]Philip Morris Products S.A., Quai Jeanrenaud 5, Neucha^tel, Switzerland

[2]Oracle Life Sciences, Oracle Corporation, Austin, TX, United States

[3]ARGUS GmbH, Berlin, Germany

Corresponding Author:
Steve Roulet, BSc
Philip Morris Products S.A., Quai Jeanrenaud 5, Neucha^tel, Switzerland

Abstract

Background: Adults who smoke cigarettes may reduce their exposure to cigarette smoke by switching completely or partially to an electronic nicotine delivery system (ENDS).

Objective: This study investigated the use of a novel ENDS over 8 weeks, under near real-world conditions, among adults who smoke cigarettes daily and are living in the United States. The objectives were to assess the proportion of participants who switched completely to the novel ENDS and changes in cigarette consumption.

Methods: This actual use study assessed the use of the novel ENDS among exclusive cigarette smokers (n=353) and dual users of cigarettes and ENDS (n=367). Participants were required to be aged 21 to 64 years old, had smoked 100+ cigarettes in their lifetime, smoked cigarettes daily at the time of enrollment, and reported positive intention to try, use, and buy the study product (SP). Participants were recruited from consumer databases. During the baseline period, participants self-reported their daily use of cigarettes and ENDS using an electronic diary. During the observational period, self-reporting also included daily use of SP. The SP was available in 2 variants (tobacco and menthol) at one nicotine concentration (3.5%; 39 mg/mL nicotine). Participants obtained 10 cartridges at the beginning of the observational period and could select either SP variant or both and were allowed to obtain additional SP. SP use patterns and sensory experiences were assessed at the end of the observational period, and cigarette consumption was compared to baseline. SP variant use and its impact on complete switching and cigarette consumption were also assessed.

Results: The average age of enrolled exclusive cigarette smokers and dual users was 47 (SD 10.4) years and 44.3 (SD 9.6) years, respectively, and there were more female than male participants (227/353, 64.3% and 186/367, 50.7%). Participants smoked on average 13.3 (SD 7.3) and 12.1 (SD 7.2) cigarettes per day, and mostly menthol cigarettes (197/353, 55.8% and 201/367, 54.8%). Most participants had no plan to quit cigarette smoking in the next 6 months (261/353, 73.9% and 328/367, 89.4%). At the end of the observational period, 86.2% (304/353) of exclusive cigarette smokers and 92.1% (338/367) of dual users used the SP; 4% (14/353) of exclusive cigarette smokers and 4.1% (15/367) of dual users switched completely to the SP. Furthermore, 25.8% (91/353) of exclusive cigarette smokers and 28.1% (103/367) of dual users substantially reduced their cigarette consumption (50% or more) compared to baseline. Finally, around half of the participants (exclusive cigarette smokers: 166/304, 54.6%; dual users: 167/338, 49.4%) used only the menthol SP variant, which was also more positively evaluated in taste, smell, and aftertaste than the tobacco SP variant.

Conclusions: This study indicates that adults who smoke cigarettes can switch completely to the novel ENDS or substantially reduce their cigarette consumption. These results complement the scientific evidence suggesting that ENDS are an acceptable alternative to cigarettes for adults who smoke in the United States.

(JMIR Form Res 2025;9:e76019) doi:10.2196/76019

KEYWORDS

e-cigarettes; electronic nicotine delivery systems; tobacco use; use patterns; smoking reduction; quitting smoking

XSL•FO
RenderX

Introduction

Nearly 36 million adults in the United States smoked cigarettes or other combusted tobacco products in 2021 [1]. Exposure to cigarette smoke is harmful to both individual and public health, increasing the risks of developing many serious diseases [2-5].

The adverse health impacts of smoking can be mitigated by preventing smoking initiation and promoting smoking cessation among adults who smoke cigarettes. Such adverse health impacts can also be reduced among adults who would otherwise continue to smoke by transitioning them away from cigarettes toward alternative tobacco and nicotine-containing products (TNP) that present lower risks to health, such as electronic nicotine delivery systems (ENDS) [6].

Estimates from population health modeling suggest that complete switching from cigarettes to ENDS among adults who smoke cigarettes in the United States could prevent approximately 6.6 million tobacco-related deaths over a 10-year period [7]. Switching to ENDS is the most common method used by adults who smoke cigarettes in the United States when attempting to quit smoking [8-10]. Findings from cohort studies and randomized controlled trials summarized in a recent Cochrane Review demonstrate higher rates of smoking reduction and quitting among adults who smoke cigarettes and use ENDS than those who use other methods, such as nicotine replacement therapies (NRTs). Adults who smoke cigarettes and who completely or partially switch to ENDS are shown to smoke significantly fewer cigarettes per day (CPD) and have increased smoking abstinence compared to those using NRT, behavioral support, and no intervention [11].

Adults who smoke cigarettes, who switch completely or partially from cigarettes to ENDS and thus substantially reduce cigarette consumption—often defined as >50% fewer CPD [12,13]—may reduce their exposure to the harmful and potentially harmful constituents present in tobacco smoke [14,15]. In addition, a substantial reduction in cigarette consumption is associated with making more quit attempts [16,17], reduced cigarette dependence [18-20], and increased likelihood of future smoking cessation [16,17,21-26]. Thus, among adults who would otherwise continue to smoke, switching completely or partially from cigarette smoking to ENDS use may reduce the adverse health impacts of smoking, given the known benefits of smoking cessation or substantial reductions in cigarette consumption [27-32].

The aim of this actual use study was to investigate use patterns of a novel closed-system ENDS (the study product [SP]), available with tobacco and menthol e-liquid prefilled cartridges, and changes in cigarette consumption among adults who smoke cigarettes daily in near real-world conditions over an 8-week period in the United States. The study objectives were to assess (1) the proportion of participants who switched completely to the SP, (2) changes in cigarette consumption, and (3) study objectives 1 and 2 stratified by the SP flavor variants used.

Methods

Study Design

This actual use study was conducted from May 24, 2022 (first enrollment) to October 3, 2022 (last completion) among adults who smoke cigarettes daily in the United States, including exclusive cigarette smokers and dual users of cigarettes and ENDS (hereafter, dual users). Participation in the study consisted of 3 phases: (1) a 1-week baseline phase to assess cigarette and other TNP use prior to the introduction of the SP; (2) an 8-week observational phase where participants were able to consume the provided SP and other TNP ad libitum under near real-world conditions, allowing assessment of SP use and changes in cigarette and other TNP use; and (3) a 1-week closeout phase for continued surveillance of potential adverse events. Figure 1 presents the phases of the study design. The study design was implemented following similar actual use studies [33-35].

Figure 1. Study design.

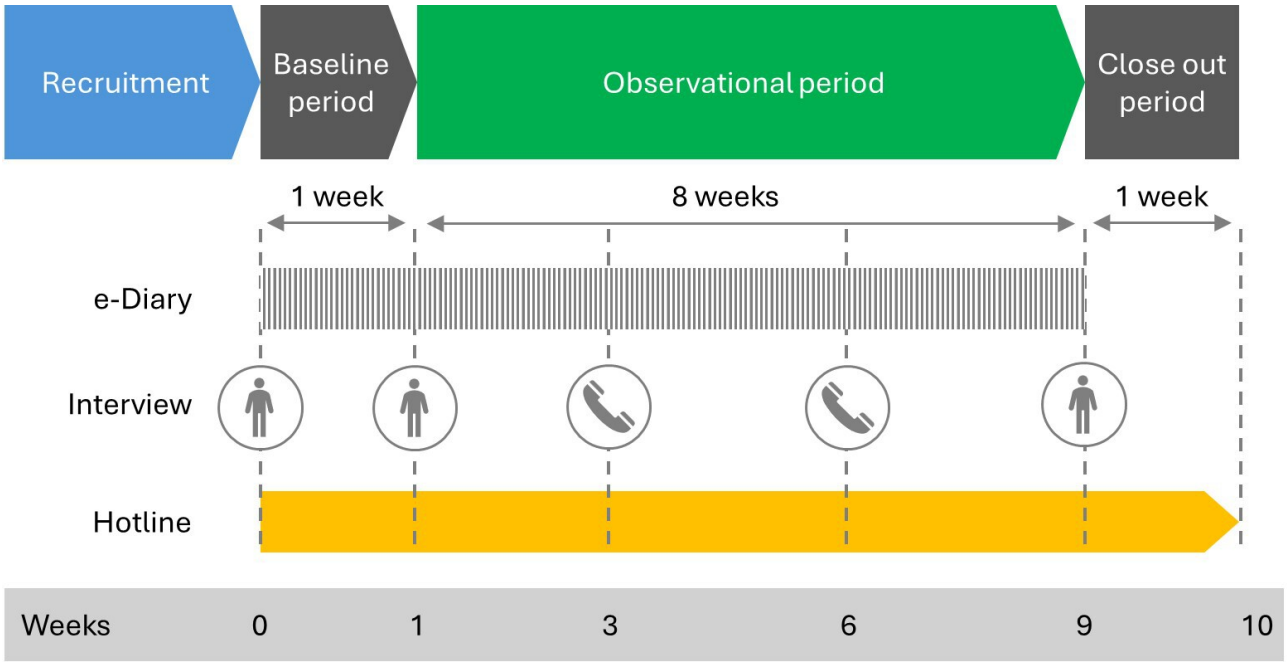

XSL•FO
RenderX

Participants

Participants were recruited from consumer databases of 7 geographically distributed study sites in the United States (Atlanta, Baltimore, Miami, Nashville, Orlando, Phoenix, and St. Louis). Candidate participants were screened based on the inclusion and exclusion criteria presented in Textbox 1. Participants were required to be adults who smoked cigarettes in the United States, aged between 21 and 64 years, who had smoked 100+ cigarettes in their lifetime and smoked cigarettes daily at the time of enrollment.

Textbox 1. Inclusion and exclusion criteria for participation.

Inclusion criteria

- Identify as male or female participant.
- Age 21 to 64 years old, current United States resident.
- Able to read and understand English.
- Own a qualifying smartphone for daily electronic diary entries.
- Smoked 100+ cigarettes in their lifetime.
- Currently smoke at least one cigarette daily.
- No intention to quit all tobacco and nicotine-containing products in the next 30 days.
- Report a positive overall opinion and a positive intention to try, use, and buy the study product (SP; after reviewing product information; responded good, very good, excellent when asked about overall opinion of the SP, and somewhat likely, very likely, or definitely when asked likelihood to try, use, and buy the SP).

Exclusion criteria

- No proof of age (eg, photo identification or driver's license).
- Self-reported currently being pregnant or breastfeeding.
- Females of childbearing potential who were not using contraception.
- Employed in journalism, market research, public relations, in the tobacco industry, or as a health care provider.
- Participated in a consumer or clinical study on tobacco or nicotine-containing products in the past 30 days.
- Initiated cigarette or electronic nicotine delivery system use in the past 30 days.

Both exclusive cigarette smokers and dual users could benefit from a reduction in cigarette smoking; therefore, they were considered as separate study groups in this study. Qualifying adults who smoked cigarettes daily but did not use ENDS daily at the time of enrollment, nor had used ENDS 100+ times in their lifetime, were assigned as exclusive cigarette smokers. Qualifying adults who smoked cigarettes daily and used ENDS either daily or occasionally and had used ENDS 100+ times in their lifetime at the time of enrollment were enrolled as dual users. Participant enrollment and recruitment aimed to represent the sociodemographic characteristics of adults who smoke cigarettes daily in the United States, with marginal distributions of sex (male: 55%; female: 45%), age (21‑24 years: 10%; 25‑44 years: 50%; 45‑64 years: 40%), and race (White: 70%; non-White: 30%) based on data from the National Health Interview Survey (NHIS).

Recruitment

The candidate participants were randomly selected and recruited from an external consumer database based on their smoking status and sociodemographic characteristics. Social media postings were also used to support recruitment. In a first step, candidate participants were screened against inclusion and exclusion criteria. Eligible participants were invited to the study site for rescreening and enrollment into the study. In a second step, eligible participants were invited to consent to being rescreened and exposed to the SP. Following exposure to the SP label, labeling, and marketing material, participants' "Overall Opinion" and "Intention to Use" the SP were assessed. Participants who responded "good," "very good," or "excellent" on "Overall Opinion" and participants who responded "somewhat likely," "very likely," or "definitely" on "Intention to Try," "Intention to Use," and "Intention to Buy" were enrolled into the study to approximate product self-selection based on behavior intentions. This approach also protected participants by avoiding unnecessarily enrolling those who were not interested in the product.

Participants could voluntarily withdraw or be removed from the study if they did not report at least 4 electronic diary (e-diary) entries during the baseline period, failed to comply with study procedures, misused the SP, or became pregnant. All participants were compensated fairly for their time at the end of the observational period.

Study Procedures

Study Products

The SP was the P4M3 Generation 2.0 System (Philip Morris Products S.A.) ENDS device with prefilled cartridges containing e-liquid (3.5%; 39 mg/mL nicotine) with 2 flavors: tobacco and menthol along with applicable safety warnings, labeling, and instructions for use. Participants received the SP at no cost from study sites during the observational period but had to pay for

XSL•FO
RenderX

any other TNP they chose to use, which is an approach that has been used in previous tobacco use studies [33-37].

Participants obtained 10 cartridges at the beginning of the observational period and could select tobacco, menthol, or both variants; they were allowed to obtain additional cartridges during the study. A maximum number of SP cartridges available to each participant was applied based on the participant's self-reported cigarette consumption at enrollment multiplied by a factor of 2. During the observational period, participants were repeatedly exposed to a reduced exposure claim:

> *SP produces significantly fewer and lower levels of harmful chemicals compared to cigarettes. Scientific studies have shown that switching completely from cigarettes to SP significantly reduces your body's exposure to harmful chemicals. Important information: the SP is not risk-free.*

The SP device and any unused cartridges had to be returned to the study sites at the end of the observational period. Each study site was responsible for maintaining records of the distribution and return of the SP during the study.

Data Collection

Participants self-reported their daily use of cigarettes and other TNP (during the baseline and observational periods) and the SP, including the flavor variant used (during the observational period only) in an app-based e-diary. Participants were also interviewed 3 times during the observational period (following week 2, week 5, and week 8) to evaluate the SP's sensory attributes (taste, smell, and aftertaste) and ease of use. Participants could report any potential adverse events or change in pregnancy status to a study-specific hotline number.

Outcomes

Study outcomes included the following and are based on results at the end of the observational period (ie, week 8) versus baseline: (1) the proportion of participants who switched completely to the SP, (2) the absolute change in CPD, (3) the proportion of participants who reduced CPD by 50% or more, and (4) the proportion of participants who transitioned from daily to nondaily cigarette smoking. Each of the outcomes was assessed separately for exclusive cigarette smokers and dual users and was stratified by the SP variant(s) (only tobacco, only menthol, or both variants) used in week 8.

As continued SP use is associated with appealing product characteristics, the following were assessed at the end of the observational period: (1) ratings of liking of sensory attributes (taste, smell, and aftertaste) among users of the respective SP variant(s) and (2) ease of SP use [38].

Measures and Definitions

All measures and definitions were based on those commonly used in national surveys on TNP use (eg, Population Assessment of Tobacco and Health [PATH]) [39] and from a similar actual use study [33]. At enrollment, participants were asked to report the following sociodemographic characteristics: age in years, sex, race, ethnicity, level of education, and annual household income.

At enrollment, current TNP use was assessed using the following categories: cigarettes, ENDS, moist snuff, chewing tobacco, snus, nicotine pouches, oral tobacco-free nicotine products, cigars, cigarillos, pipe tobacco, hookah or waterpipe tobacco (What is your current [insert TNP category] use behavior? [daily user, occasional user, or nonuser]). Participants also reported the following: current CPD (On average, how many cigarettes do you currently smoke per day? [numerical response]); flavor of cigarette (Is the brand of cigarettes you currently buy and smoke most often a menthol cigarette? [yes, no]); and intention to quit smoking (Are you seriously considering quitting smoking cigarettes in the next 6 months? [yes, no, don't know]). Dual users were asked to report the flavor of e-liquid they used most often (What flavor of ENDS with nicotine do you currently buy and you, yourself, use most often?). The use of cigarettes and other TNP (as one overall category) was assessed each day during the baseline and observational periods, while SP and other ENDS were assessed each day during the observational period. Participants self-reported the total number of times they used the TNPs each day in an app-based e-diary. Daily and occasional use and the number of product uses per day were determined based on the number of entries for these products and were averaged across the specific week of the observational period. Change in average daily cigarette consumption was defined as the difference in average daily number of cigarettes consumed at week 8 compared to baseline.

Participants also reported the SP variant(s) they used daily. Switched completely to SP was defined as participants who reported using the SP but not smoking cigarettes or using other ENDS at week 8. Switched completely to ENDS was defined as participants who reported using the SP and other ENDS but not smoking cigarettes at week 8. Dual SP use was defined as participants who reported using the SP and cigarettes. Dual ENDS use was defined as participants who reported using the SP, other ENDS, and cigarettes at week 8.

Evaluation of SP sensory attributes (taste, smell, and aftertaste) and ease of use of SP were assessed via interviews. Each sensory attribute was evaluated using a 7-point Likert scale ranging from 1 (I don't like it at all) to 7 (I like it very much). Ease of use of SP was assessed using a 7-point Likert scale ranging from 1 (not easy to use at all) to 7 (very easy to use).

Sample Size Determination

The target sample sizes for exclusive cigarette smokers and dual users (study groups) were 500 and 375, respectively. Sample size calculations were based on an expected attrition rate of 40% for exclusive cigarette smokers and 20% for dual users, yielding a sample size of 300 participants in each study group at week 8 of the observational period. A higher attrition among exclusive cigarette smokers was expected, as they are less familiar with ENDS than dual users. The target sample sizes were also expected to provide a reliable estimation of the proportion of participants who switched completely to SP at week 8, with a minimum predefined level of precision of ±5% if the proportion was within the expected range of 5% - 20% for each study group.

XSL•FO
RenderX

Data Analysis

Descriptive summary statistics were used to describe the study outcomes. Counts and percentages were calculated for categorical variables. Means, SDs, and 95% CIs were calculated for continuous and ordinal variables. Results were calculated for each outcome separately for exclusive cigarette smokers and dual users and stratified by the SP variant used at week 8. Results for each week of the observational period were based on the number of daily e-diary entries reported for each participant. The average product use was calculated if there were at least 4 daily e-diary entries that week; otherwise, data for that week were considered missing. Data analysis was conducted using SAS 9.4 (SAS Institute).

Ethical Considerations

The study was approved by the Sterling Institutional Review Board (IRB ID: 9938). All participants provided written informed consent before participating in the study. The informed consent form included information that the findings may be published externally with no individual data disclosed. Participants were compensated US $100 for completing the on-site face-to-face screening interview, US $20 for completing each weekly e-diary assessment, US $35 for completing each follow-up phone interview, and an additional compensation for completing all study assessments and returning SPs was provided, with a maximum total compensation of US $500. The incentive of the on-site screening interview was increased by US $75 for exclusive cigarette smokers to ensure sufficient recruitment of this study group.

Results

Recruitment, Enrollment, and Participation

There were 10,493 candidate participants initially screened for eligibility, with 909 qualifying and able to attend a study site for rescreening. Of the 909 rescreened qualifying participants, 821 were enrolled in the study (398 exclusive cigarette smokers; 423 dual users). There were 80 participants who discontinued or withdrew during the observational period, and 21 were excluded from the final analyses. Thus, 720 participants (353 exclusive cigarette smokers, attrition rate: 11.3%; 367 dual users, attrition rate: 13.2%) who completed the observational period and fulfilled the analysis criteria were included in the final analytical sample. Figure 2 details participants' recruitment and enrollment.

XSL•FO
RenderX

Figure 2. Participants' recruitment and enrollment. e-diary: electronic diary.

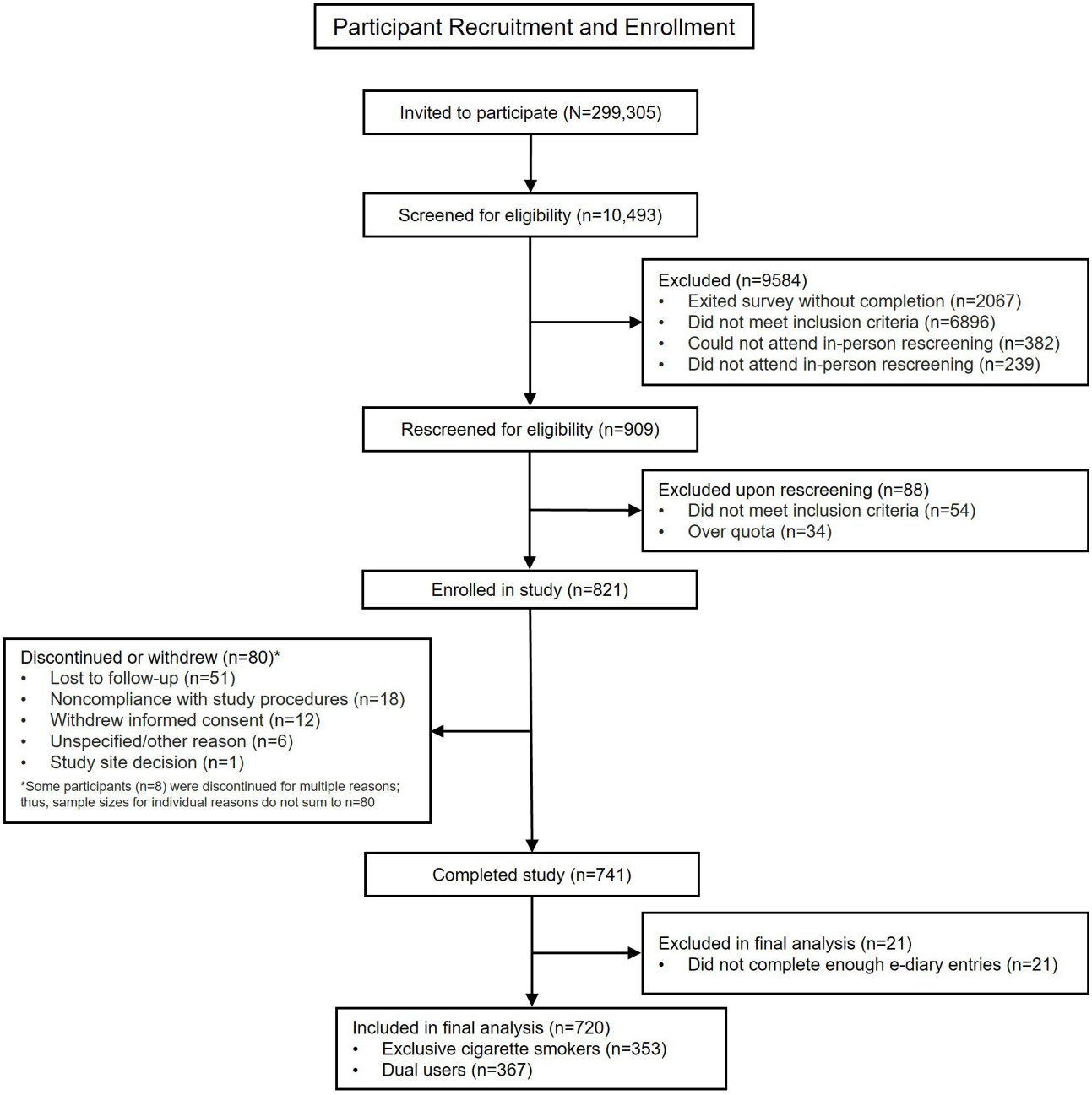

Sociodemographic Characteristics of the Study Participants

Exclusive cigarette smokers were slightly older than dual users (mean age 47, SD 10.4 years vs 44.3, SD 9.6 years) and contained fewer participants aged 21 - 34 years old (45/353, 12.7% vs 63/367, 17.2%) and males (males: 126/353, 35.7% vs 181/367, 49.3%). Results for other sociodemographic characteristics are presented in Table 1.

XSL•FO
RenderX

Table . Sociodemographic characteristics of the study participants.

Sociodemographic characteristics	Exclusive cigarette smokers (n=353)	Dual users (n=367)
Age (years), mean (SD)	47 (10.4)	44.3 (9.6)
Age category (years), n (%)		
21 - 24	6 (1.7)	5 (1.4)
25 - 34	39 (11)	58 (15.8)
35 - 44	100 (28.3)	122 (33.2)
45 - 64	208 (58.9)	182 (49.6)
Sex, n (%)		
Male	126 (35.7)	181 (49.3)
Female	227 (64.3)	186 (50.7)
Ethnicity, n (%)		
Hispanic or Latino	54 (15.3)	37 (10.1)
Not Hispanic or Latino	299 (84.7)	330 (89.9)
Prefer not to answer	0 (0)	0 (0)
Race, n (%)		
White	226 (64)	228 (62.1)
Black	107 (30.3)	111 (30.2)
Another or multiple races	20 (5.7)	28 (7.6)
Prefer not to answer	0 (0)	0 (0)
Level of education, n (%)[a]		
Less than college graduate	242 (68.6)	197 (53.7)
College graduate or higher	111 (31.4)	170 (46.3)
Prefer not to answer	0 (0)	0 (0)
Annual household income, n (%)[b]		
Less than $75,000	217 (61.5)	205 (55.9)
$75,000 or higher	135 (38.2)	159 (43.3)
Prefer not to answer	0 (0)	2 (0.5)
Don't know	1 (0.3)	1 (0.3)

[a]Level of education was based on the following response options: less than college graduate (less than high school; some high school or general education diploma; high school graduate; some college), college graduate or higher (college graduate; advanced degree), and prefer not to answer.

[b]Annual household income was based on the following response options: less than US $75,000 (less than US $10,000; US $10,000 to $29,999; US $30,000 to $44,999; US $45,000 to $59,999; US $60,000 to $74,999), US $75,000 or higher (US $75,000 to $99,999; US $100,000 to $149,999; US $150,000 or higher), prefer not to answer, and don't know.

Tobacco and Nicotine Product Use at Enrollment

Exclusive cigarette smokers and dual users smoked, on average, 13.3 (SD 7.3) and 12.1 (SD 7.2) CPD at enrollment, respectively. Nearly equal proportions reported they smoked menthol cigarettes (exclusive cigarette smokers: 197/353, 55.8%; dual users: 201/367, 54.8%). The proportion of exclusive cigarette smokers who had no plan to quit cigarette smoking in the next 6 months was 73.9% (261/353), while this proportion reached 89.4% (328/367) among dual users (Table 2).

XSL•FO
RenderX

Table . Tobacco and nicotine product use at enrollment.

Characteristics	Exclusive cigarette smokers (n=353)	Dual users (n=367)
Characteristics of cigarette smoking at enrollment		
Pattern of cigarette smoking, n (%)		
Adults who smoke daily	353 (100)	367 (100)
Frequency of cigarette smoking, mean (SD)		
CPD[a]	13.3 (7.3)	12.1 (7.2)
Type of cigarettes smoked, n (%)		
Menthol	197 (55.8)	201 (54.8)
Nonmenthol	156 (44.2)	166 (45.2)
Planned to quit smoking cigarettes in the next 6 months, n (%)		
Yes	54 (15.3)	34 (9.3)
No	261 (73.9)	328 (89.4)
Don't know	38 (10.8)	5 (1.4)
Characteristics of other TNP[b] use at enrollment		
Patterns of ENDS[c] use, n (%)[d]		
Daily user	0 (0)	173 (47.1)
Occasional user	55 (15.6)	194 (52.9)
Nonuser	298 (84.4)	0 (0)
Flavor of ENDS used most often, n (%)[e]		
Tobacco	—	86 (23.4)
Menthol	—	146 (39.8)
Nontobacco and nonmenthol[f]	—	131 (35.7)
Some other flavor	—	4 (1.1)
Don't know	—	0 (0)
Patterns of other TNP use, n (%)		
Moist snuff, chewing tobacco, or snus		
Daily user	1 (0.3)	13 (3.5)
Occasional user	16 (4.5)	72 (19.6)
Nonuser	336 (95.2)	282 (76.8)
Nicotine pouches		
Daily user	1 (0.3)	11 (3)
Occasional user	20 (5.7)	84 (22.9)
Nonuser	332 (94.1)	272 (74.1)
Cigars or cigarillos		
Daily user	23 (6.5)	49 (13.4)
Occasional user	118 (33.4)	171 (46.6)
Nonuser	212 (60.1)	147 (40.1)
Pipe tobacco		
Daily user	0 (0)	1 (0.3)
Occasional user	16 (4.5)	47 (12.8)
Nonuser	337 (95.5)	319 (86.9)
Hookah or waterpipe tobacco		

XSL•FO
RenderX

Characteristics	Exclusive cigarette smokers (n=353)	Dual users (n=367)
Daily user	0 (0)	10 (2.7)
Occasional user	87 (24.6)	153 (41.7)
Nonuser	266 (75.4)	204 (55.6)

[a]CPD: cigarettes per day.

[b]TNP: tobacco and nicotine-containing products.

[c]ENDS: electronic nicotine delivery systems.

[d]Adults who smoke cigarettes, who did not currently use electronic nicotine delivery systems daily nor had used them 100+ times in their lifetime, were enrolled in the exclusive cigarette smokers study group, whereas adults who smoke cigarettes, who currently used electronic nicotine delivery systems (daily or occasionally), and had used them 100+ times in their lifetime, were enrolled in the dual users study group.

[e]Exclusive cigarette smokers, who neither currently used electronic nicotine delivery systems daily nor had used them 100+ times in their lifetime, were not asked the flavor of electronic nicotine delivery systems they used most often.

[f]Nontobacco and nonmenthol electronic nicotine delivery systems flavor use included the following response options: fruit, chocolate, alcoholic or nonalcoholic drink, clove, spice, candy, desserts.

Among dual users, there were nearly equal proportions who were daily (173/367, 47.1%) or occasional (194/367, 52.9%) ENDS users. The majority of dual users reported using menthol (146/367, 39.8%) or tobacco (86/367, 23.4%) ENDS variants, while 36.8% (135/367) used other flavor variants. Information about other TNPs used at enrollment can be found in Table 2.

Study Product Use and Assessment at Week 8

The e-diary compliance was high, with 96% (691/720) of participants in both study groups making 4 or more entries each week of the observational period.

At the end of the observational period (week 8), 86.2% (304/353) of exclusive cigarette smokers and 92.1% (338/367) of dual users used the SP. Dual users reported using the SP more often than exclusive cigarette smokers (mean 7.3, SD 12.3 times per day vs 5, SD 8.2 times per day). At week 8, 4% (14/353) of exclusive cigarette smokers and 4.1% (15/367) of dual users had switched completely to the SP. When considering the use of SP together with other ENDS, 4.5% (16/353) of exclusive cigarette smokers and 5.2% (19/367) of dual users had switched completely to ENDS at week 8. More participants in both study groups used only the menthol variant (exclusive cigarette smokers: 166/353, 54.6%; dual users: 167/367, 49.4%) compared to participants who used only the tobacco variant (exclusive cigarette smokers: 101/353, 33.2%; dual users: 112/367, 33.1%) and those who used both variants at week 8 (exclusive cigarette smokers: 37/353, 12.2%; dual users: 59/367, 17.5%).

The SP was favorably evaluated by participants in both study groups, with mean ratings for liking the sensory attributes ranging between 4.6 (SD 1.8) and 6.2 (SD 1.3) on a 7-point Likert scale. Notably, they liked the smell, taste, and aftertaste of the menthol variant on average more than the tobacco variant. Participants evaluated the SP as easy to use (mean 6.4, SD 1.2 vs 6.4, SD 1.1 on a 7-point Likert scale) (Table 3).

XSL•FO
RenderX

Table . Study product use and assessment at week 8.

Study product use	End of observational period (week 8)	
	Exclusive cigarette smokers (n=353)	Dual users (n=367)
SP[a] use at week 8		
Patterns of SP use, n (%)[b]		
Used SP (any SP use)	304 (86.2)	338 (92.1)
Switched completely to SP (SP use without cigarettes)	14 (4)	15 (4.1)
Switched completely to ENDS[c] (SP use and other ENDS without cigarettes)	2 (0.6)	4 (1.1)
Dual SP use (SP and cigarettes)	273 (77.3)	218 (59.4)
Dual ENDS use (SP and other ENDS and cigarettes)	15 (4.2)	101 (27.5)
Did not use SP (SP nonusers)	49 (13.8)	29 (7.9)
Frequency of SP use		
SP use, occasions per day		
Mean (SD)	5 (8.2)	7.3 (12.3)
95% CI[d]	4.1-5.9	6-8.5
Patterns of SP use by variant used[e], n (%)		
SP variant used overall (n=304/n=338)		
Only SP tobacco variant	101 (33.2)	112 (33.1)
Only SP menthol variant	166 (54.6)	167 (49.4)
Both SP variants	37 (12.2)	59 (17.5)
Switched completely to SP by variant used (n=14/n=15)		
Only SP tobacco variant (n=101/n=112)	5 (5)	4 (3.6)
Only SP menthol variant (n=166/n=167)	8 (4.8)	8 (4.8)
Both SP variants (n=37/n=59)	1 (2.7)	3 (5.1)
SP flavor variants used by type of cigarettes smoked at enrollment, n (%)[f]		
Menthol cigarette smokers (n=174/n=187)		
Only SP tobacco variant	12 (6.9)	16 (8.6)
Only SP menthol variant	140 (80.5)	134 (71.7)
Both SP variants	22 (12.6)	37 (19.8)
Nonmenthol cigarette smokers (n=130/n=151)		
Only SP tobacco variant	89 (68.5)	96 (63.6)
Only SP menthol variant	26 (20)	33 (21.9)
Both SP variants	15 (11.5)	22 (14.6)
Assessment of SP used at final follow-up interview		
Liking of SP sensory attributes, mean (SD) [95% CI][g]		
Taste		
Only SP tobacco variant (n=101/n=112)	4.8 (1.9), 4.5-5.2	4.6 (1.8), 4.3-5
Only SP menthol variant (n=166/n=167)	5.4 (1.5), 5.2-5.6	5.3 (1.5), 5.1-5.5
Both SP variants (n=37/n=59)	5.2 (1.4), 4.7-5.7	5.6 (1.3), 5.3-6
Smell		
Only SP tobacco variant (n=100/n=112)	5.9 (1.4), 5.6-6.1	5.5 (1.7), 5.2-5.9

XSL•FO
RenderX

Study product use	End of observational period (week 8)	
	Exclusive cigarette smokers (n=353)	Dual users (n=367)
Only SP menthol variant (n=162/n=163)	6.2 (1.3), 6-6.4	5.9 (1.4), 5.7-6.2
Both SP variants (n=37/n=59)	5.9 (1.4), 5.5-6.4	6.2 (1.2), 5.9-6.5
Aftertaste		
Only SP tobacco variant (n=100/n=112)	4.8 (1.9), 4.4-5.2	4.6 (1.9), 4.2-4.9
Only SP menthol variant (n=165/n=165)	5.3 (1.7), 5-5.6	5.2 (1.7), 4.9-5.4
Both SP variants (n=37/n=58)	5.2 (1.5), 4.8-5.7	5.4 (1.6), 4.9-5.8
Ease of use, mean (SD) [95% CI][h]		
Ease of use	6.4 (1.2), 6.3-6.5	6.4 (1.1), 6.3-6.6

[a]SP: study product.

[b]Pattern of study product use definitions was based on use at week 8. Switched completely to study product (study product use without cigarettes) was defined as participants who reported using the study product but not smoking cigarettes or using other electronic nicotine delivery system at week 8. Switched completely to electronic nicotine delivery system (study product use and other electronic nicotine delivery system without cigarettes) was defined as participants who reported using the study product and other electronic nicotine delivery system but not smoking cigarettes at week 8. Dual study product use (study product and cigarettes) was defined as participants who reported using the study product and cigarettes. Dual electronic nicotine delivery system use (study product and other electronic nicotine delivery system and cigarettes) was defined as participants who reported using the study product, other electronic nicotine delivery system, and cigarettes at week 8.

[c]ENDS: electronic nicotine delivery systems.

[d]Calculated for participants who used the study product at week 8.

[e]Study product flavor variants used at week 8 were categorized as only tobacco SP variant, only menthol SP variant, or both SP variants.

[f]Study product nonusers are not included in these analyses.

[g]Liking of each sensory attribute was assessed using a 7-point Likert scale ranging from 1 (I don't like it at all) to 7 (I like it very much); participants could also respond don't know (not shown) among users of the respective SP variant(s) at week 8.

[h]Ease of use was assessed using a 7-point Likert scale ranging from 1 (Not easy to use at all) to 7 (Very easy to use); participants could also respond don't know (not shown).

A total of 6 adverse events were reported over the course of the study. None were related to the use of the SP.

Changes in Cigarette Consumption at Week 8

Participants in both study groups reduced their daily cigarette consumption from baseline to week 8 by 3.1 CPD among exclusive cigarette smokers and by 2.5 CPD among dual users. About one-quarter of exclusive cigarette smokers (91/353, 25.8%) and dual users (103/367, 28.1%) reduced CPD by 50% or more (Table 4).

XSL•FO
RenderX

Table . Changes in cigarette smoking from baseline to week 8.

Changes in cigarette smoking	Exclusive cigarette smokers (n=353)		Dual users (n=367)	
	Baseline	Week 8	Baseline	Week 8
Characteristics of cigarette smoking at week 8				
Frequency of cigarette smoking, mean (SD), [95% CI][a]				
Average CPD[b]	12.4 (7.1),11.6-13.1	9.3 (7.5), 8.5-10.1	10.7 (7.2), 10-11.5	8.3 (7.3), 7.5-9
Absolute change from baseline	__[c]	–3.1 (4.5), –3.5 to 2.6	—	–2.5 (4.4), –2.9 to –2
Days without cigarette smoking among people who smoke daily				
n	6	64	30	98
Days per week, mean (SD), 95% CI	1.5 (0.8), 0.6-2.4	3.8 (2.4), 3.2-4.4	2.2 (1.5), 1.6-2.8	3.6 (2.3), 3.2-4.1
CPD reduction >50% from baseline, n (%)				
Reduced CPD >50%	—	91 (25.8)	—	103 (28.1)
Cigarette smoking by SP[d] variant used at week 8[e]				
Patterns of cigarette smoking by SP variant used, n (%)[f]				
Only SP tobacco variant (n=101/n=112)				
Daily user	101 (100)	86 (85.1)	104 (92.9)	92 (82.1)
Nondaily user	0 (0)	10 (9.9)	8 (7.1)	15 (13.4)
Nonuser	—	5 (5)	—	5 (4.5)
Only SP menthol variant (n=166/n=167)				
Daily smoker	163 (98.2)	135 (81.3)	151 (90.4)	124 (74.3)
Nondaily smoker	3 (1.8)	25 (15.1)	16 (9.6)	35 (21)
Nonsmoker	—	6 (3.6)	—	8 (4.8)
Both SP variants (n=37/n=59)				
Daily smoker	37 (100)	27 (73)	56 (94.9)	34 (57.6)
Nondaily smoker	0 (0)	9 (24.3)	3 (5.1)	20 (33.9)
Nonsmoker	—	1 (2.7)	—	5 (8.5)
Frequency of cigarette use (CPD) by SP variant used, mean (SD), 95% CI[a]				
Average CPD				
Only SP tobacco variant (n=101/n=112)	13.4 (7.5), 12-14.9	10 (6.8), 8.7-11.4	12 (7.2), 10.7-13.4	9.7 (7.4), 8.3-11.1
Absolute change from baseline	—	–3.4 (4.7), –4.3 to –2.5	—	–2.3 (4.8), –3.2 to –1.4
Only SP menthol variant (n=166/n=167)	12.1 (6.8), 11.1-13.1	8.7 (7.3), 7.6-9.9	10.4 (7.4), 9.3-11.5	7.7 (6.8), 6.7-8.8
Absolute change from baseline	—	–3.4 (4.6), –4.1 to –2.6	—	–2.7 (4.4), –3.4 to –2
Both SP variants (n=37/n=59)	10.5 (5.8), 8.6-12.4	7.5 (6.9), 5.2-9.8	8.2 (5.2), 6.8-9.6	5.6 (5.5), 4.1-7
Absolute change from baseline	—	–3 (3.8), –4.3 to –1.8	—	–2.6 (3.6), –3.5 to –1.7
CPD reduction>50% from baseline by SP variant used, n (%)				
Reduced CPD>50%				

XSL•FO
RenderX

Changes in cigarette smoking	Exclusive cigarette smokers (n=353)		Dual users (n=367)	
	Baseline	Week 8	Baseline	Week 8
Only SP tobacco variant (n=101/N=112)	—	18 (17.8)	—	24 (21.4)
Only SP menthol variant (n=166/N=167)	—	52 (31.3)	—	48 (28.7)
Both SP variants (n=37/N=59)	—	11 (29.7)	—	23 (39)
Patterns of cigarette smoking, n (%)[g]				
Daily smoker	347 (98.3)	289 (81.9)	337 (91.8)	269 (73.3)
Nondaily smoker	6 (1.7)	49 (13.9)	30 (8.2)	77 (21)
Nonsmoker	—	15 (4.2)	—	21 (5.7)

[a]CPD calculated only among participants who smoked cigarettes at week 8.

[b]CPD: cigarettes per day.

[c]Not available.

[d]SP: study product.

[e]SP flavor variants used at week 8 were categorized as only tobacco SP variant, only menthol SP variant, or both SP variants.

[f]Results for daily and nondaily smoking when stratified by SP flavor variants used may not equal the results for daily and nondaily smoking overall given some participants did not use the SP at week 8.

[g]Daily cigarette smoking was an inclusion criterion for study enrollment and all (100%) participants in both study groups reported daily cigarette smoking. Results for the use pattern of cigarette smoking were based on data at baseline, during which some exclusive cigarette smokers (n=6) and dual users (n=30) did not smoke daily.

The proportion of participants who reported smoking less than daily increased from baseline to week 8 among exclusive cigarette smokers (from 1.7% (6/353) to 13.9% (49/353) and dual users (from 8.2% (30/367) to 21% (77/367)).

Study Results by SP Variant Used at the End of the Observational Period

The majority of cigarette smokers who used menthol and nonmenthol cigarettes at enrollment used the congruent flavor SP variant at week 8. However, higher proportions of nonmenthol cigarette smokers used the menthol SP variant than menthol cigarette smokers who used the tobacco SP variant (Table 3).

At the end of the observational study, the proportions of participants who switched completely to the SP were 5% (5/101) and 3.6% (4/112) among exclusive cigarette smokers and dual users, respectively, among those who used only the tobacco SP variant, while this proportion was similar in both study groups among those who used only the menthol SP variant (8/166, 4.8% and 8/167, 4.8%). The proportion of participants who switched completely to SP was 2.7% (1/37) among exclusive cigarette smokers and 5.1% (3/59) among dual users among those who used both SP variants (Table 3).

Regarding changes in cigarette smoking, the reductions in CPD were similar by flavor SP variant(s) used, ranging from 3 CPD among participants who used both SP variants to 3.4 CPD among participants who used only the tobacco SP variant or who used only the menthol SP variant among exclusive cigarette smokers. Among dual users, the reduction in CPD was 2.3 CPD among those who used only the tobacco SP variant, 2.6 CPD among those who used both SP variants, and 2.7 CPD among those who used only the menthol SP variant. The proportions of exclusive cigarette smokers and dual users who reduced CPD by 50% or more were higher among those who used both SP variants (11/37, 29.7% and 23/59, 39%, respectively) or only the menthol SP variant (52/166, 31.3% and 48/167, 28.7%, respectively) versus only the tobacco SP variant (18/101, 17.8% and 24/112, 21.4%, respectively) (Table 4).

The proportion of participants who reported smoking less than daily increased, from baseline to week 8, the most among participants who used both SP variants (exclusive cigarette smokers: 0% to 9/37, 24.3%; dual users: 3/59, 5.1% to 20/59, 33.9%), followed by those who used only the menthol SP variant (exclusive cigarette smokers: 3/166, 1.8% to 25/166, 15.1%; dual users: 16/167, 9.6% to 35/167, 21%), then by those who used only the tobacco SP variant (exclusive cigarette smokers: 0 to 10/101, 9.9%; dual users: 8/112, 7.1% to 15/112, 13.4%) (Table 4).

Discussion

Adults who smoke cigarettes and completely switch to ENDS or substantially reduce cigarette consumption are exposed to far lower levels of inhaled toxicants [40] and report positive changes in their health compared to adults who smoke cigarettes and continue to smoke cigarettes [41]. Hence, if a sufficient number of adults who smoke cigarettes switch to these products, it is likely to be beneficial to population health [7].

This study examined the use patterns of a novel ENDS among adults who smoke cigarettes with no or limited experience using ENDS (exclusive cigarette smokers) and those using other ENDS (dual users). These tobacco users are the intended users

XSL•FO
RenderX

of the SP, as both could benefit from switching completely from cigarette smoking [1]. Results showed high acceptance of the SP, as 86.2% (304/353) of exclusive cigarette smokers and 92.1% (338/367) of dual users used the SP at the end of the observational period (week 8). Some exclusive cigarette smokers (14/353, 4%) and dual users (15/367, 4.1%) switched completely from cigarettes to the SP, and over a quarter of all participants reduced CPD by >50% (91/353, 25.8% of exclusive cigarette smokers and 103/367, 28.1% of dual users) by week 8. The levels of reduction at week 8 among all participants equated to smoking about 3 fewer CPD, corresponding to over one fewer pack of cigarettes per week (20 cigarettes/pack) than at baseline.

The findings of this study need to be considered in the context of other studies assessing the use of ENDS among adults who smoke cigarettes in near real-world conditions. In line with our study results, Carpenter et al [42] found that, at the end of a 4-week period during which participants received a free supply of a commercially available ENDS, approximately 70% of adults who smoke cigarettes adopted the ENDS product, 10% completely switched to ENDS, and 33% reduced their CPD by >50% compared to baseline. A recent actual use study on ENDS among adults who smoke cigarettes showed that after a 1-week trial period and a 6-week use period, past-7-day cigarette abstinence ranged from 38.2% to 47.3% across 5 flavor variants and between 50.9% to 62.9% of AS who continued smoking reduced their daily cigarette consumption by at least 50% compared to baseline [43,44]. The superior outcomes observed in this study are likely attributable to the availability of a wider range of flavor variants as well as to differences in study design, specifically the inclusion of a product trial period after enrollment to identify participants with a positive interest in using and purchasing the SP before commencing the 6-week actual use period with their preferred flavor variant.

This study further supports existing evidence that e-cigarettes can help smokers to replace or substantially reduce cigarette consumption. Cross-sectional studies on tobacco use transitions among adults in the United States have found that exclusive current ENDS users are more likely to be former cigarette smokers and that dual users smoked fewer cigarettes after initiating ENDS [45,46]. Similar findings have been shown in other countries [47-49]. Cohort studies on tobacco use demonstrate that adults who smoke cigarettes and use ENDS experience a reduction in CPD, transition from daily to nondaily smoking, have increased attempts to, and successfully quit smoking, with stronger evidence for daily versus nondaily ENDS use [50-61]. Results from a recent systematic review of randomized controlled trials provide evidence of the increased efficacy of ENDS compared to conventional methods of quitting smoking, including NRT, behavioral support, and no intervention [11].

Findings of other studies showed that sensory attributes, nicotine concentration, and device type influence ENDS liking and use [62-71]. In this study, both study groups liked the taste, smell, and aftertaste of both SP variants, but with greater average liking for the menthol variant. This finding is likely to explain both the higher use proportions and the higher substantial reductions in cigarette consumption among those who used the SP menthol variant in both study groups at week 8. The higher liking and use prevalence of the menthol ENDS variant measured in this study have also been reported in previous studies [44,65,72,73], while Selya et al [44] also found that the use of menthol ENDS variants increases the likelihood of complete switching.

This study had several strengths. The study design was similar to previous actual use studies [33-35]. Adults who smoke cigarettes were allowed to use the SP ad libitum in near real-world conditions, allowing participants to establish their pattern of use over an extended period (8 weeks). Participants were recruited from 7 geographically diverse metropolitan areas in the United States, and the sample approximated that of the United States adult population who smoke cigarettes with respect to sex, age, and race. Finally, the study provided estimates for 2 populations: exclusive cigarette smokers and dual users, who might benefit from switching completely from cigarette smoking.

The study was not without limitations. Participants did not have to pay for the SP, which differs from the real-world conditions, but did have to pay for their cigarettes and other TNP. This might have affected their preference for the SP, increasing complete switching to SP and reducing cigarette consumption. The SP was available in 2 variants (tobacco and menthol) and at one nicotine concentration (3.5%; 39 mg/mL nicotine), which might have limited its adoption and use among dual users who may be familiar with a larger variety of flavor variants and nicotine concentration of ENDS [65]. Participants were compensated fairly for participation, but the impact on TNP use during the study was likely limited as payment was provided after study completion. Use of SP and all other TNP was self-reported and may be subject to recall bias, though self-reported TNP use is shown to reliably measure actual TNP use [74].

Conclusions

Results of this actual use study show that exclusive cigarette smokers and dual users of cigarettes and ENDS found the SP appealing, leading some of them to switch completely to the SP or substantially reduce their cigarette consumption. Both study groups liked the sensory experience of the tobacco and menthol variants and found the SP easy to use. These results complement existing scientific evidence that ENDS are an acceptable alternative to cigarettes for adults who smoke in the United States.

Acknowledgments

Scientific writing support (including preparation of the draft manuscript under the direction and guidance of the authors, incorporating authors' comments for each draft, assembling tables and figures, grammatical editing, and referencing) was provided by Elliott McDowell (Oracle).

XSL•FO
RenderX

Data Availability

The data supporting this study's findings are available from the corresponding author upon reasonable request.

Conflicts of Interest

SR and PM are employed by Philip Morris Products S.A. and hold stock in Philip Morris International Inc., CK was employed by Oracle Life Sciences, Oracle Corporation (formerly Cerner Enviza), and GK was employed by ARGUS GmbH while this study was conceptualized and conducted. Oracle Life Sciences was contracted by Philip Morris Products S.A. to conduct the study and draft this manuscript. Otherwise, the authors of the study have no other competing interests.

References

1. Cornelius ME, Loretan CG, Jamal A, et al. Tobacco product use among adults - United States, 2021. MMWR Morb Mortal Wkly Rep 2023 May 5;72(18):475-483. [doi: 10.15585/mmwr.mm7218a1] [Medline: 37141154]
2. Prescott E, Hippe M, Schnohr P, Hein HO, Vestbo J. Smoking and risk of myocardial infarction in women and men: longitudinal population study. BMJ 1998 Apr 4;316(7137):1043-1047. [doi: 10.1136/bmj.316.7137.1043] [Medline: 9552903]
3. Forey BA, Thornton AJ, Lee PN. Systematic review with meta-analysis of the epidemiological evidence relating smoking to COPD, chronic bronchitis and emphysema. BMC Pulm Med 2011 Jun 14;11:1-61. [doi: 10.1186/1471-2466-11-36] [Medline: 21672193]
4. Bouabdallaoui N, Messas N, Greenlaw N, et al. Impact of smoking on cardiovascular outcomes in patients with stable coronary artery disease. Eur J Prev Cardiol 2021 Oct 25;28(13):1460-1466. [doi: 10.1177/2047487320918728] [Medline: 34695217]
5. Thomson B, Emberson J, Lacey B, et al. Association between smoking, smoking cessation, and mortality by race, ethnicity, and sex among US adults. JAMA Netw Open 2022 Oct 3;5(10):e2231480. [doi: 10.1001/jamanetworkopen.2022.31480] [Medline: 36279139]
6. Hatsukami DK, Carroll DM. Tobacco harm reduction: past history, current controversies and a proposed approach for the future. Prev Med 2020 Nov;140:106099. [doi: 10.1016/j.ypmed.2020.106099] [Medline: 32335031]
7. Levy DT, Borland R, Lindblom EN, et al. Potential deaths averted in USA by replacing cigarettes with e-cigarettes. Tob Control 2018 Jan;27(1):18-25. [doi: 10.1136/tobaccocontrol-2017-053759] [Medline: 28970328]
8. Patel D, Davis KC, Cox S, et al. Reasons for current e -cigarette use among U.S. adults. Prev Med 2016 Dec;93:14-20. [doi: 10.1016/j.ypmed.2016.09.011] [Medline: 27612572]
9. Watkins SL, Thrul J, Max W, Ling PM. Cold Turkey and hot vapes? A National Study of Young Adult Cigarette Cessation Strategies. Nicotine Tob Res 2020 Apr 21;22(5):638-646. [doi: 10.1093/ntr/nty270] [Medline: 30590749]
10. U.S. Department of Health and Human Services; Centers for Disease Control and Prevention; National Center for Chronic Disease Prevention and Health Promotion; Office on Smoking and Health. Smoking cessation: a report of the surgeon general. 2020 URL: https://www.hhs.gov/sites/default/files/2020-cessation-sgr-full-report.pdf [accessed 2025-08-18]
11. Lindson N, Butler AR, McRobbie H, et al. Electronic cigarettes for smoking cessation. Cochrane Database Syst Rev 2024 Jan 8;1(1):CD010216. [doi: 10.1002/14651858.CD010216.pub8] [Medline: 38189560]
12. McRobbie H, Bullen C, Hartmann-Boyce J, Hajek P. Electronic cigarettes for smoking cessation and reduction. Cochrane Database Syst Rev 2014;2014(12):CD010216. [doi: 10.1002/14651858.CD010216.pub2] [Medline: 25515689]
13. Brose LS, Hitchman SC, Brown J, West R, McNeill A. Is the use of electronic cigarettes while smoking associated with smoking cessation attempts, cessation and reduced cigarette consumption? A survey with a 1-year follow-up. Addiction 2015 Jul;110(7):1160-1168. [doi: 10.1111/add.12917] [Medline: 25900312]
14. O'Connell G, Graff DW, D'Ruiz CD. Reductions in biomarkers of exposure (BoE) to harmful or potentially harmful constituents (HPHCs) following partial or complete substitution of cigarettes with electronic cigarettes in adult smokers. Toxicol Mech Methods 2016 Jul;26(6):443-454. [doi: 10.1080/15376516.2016.1196282] [Medline: 27401591]
15. Dai H, Benowitz NL, Achutan C, Farazi PA, Degarege A, Khan AS. Exposure to toxicants associated with use and transitions between cigarettes, e-cigarettes, and no tobacco. JAMA Netw Open 2022 Feb 1;5(2):e2147891. [doi: 10.1001/jamanetworkopen.2021.47891] [Medline: 35142830]
16. Zhou X, Nonnemaker J, Sherrill B, Gilsenan AW, Coste F, West R. Attempts to quit smoking and relapse: factors associated with success or failure from the ATTEMPT cohort study. Addict Behav 2009 Apr;34(4):365-373. [doi: 10.1016/j.addbeh.2008.11.013] [Medline: 19097706]
17. Klemperer EM, Hughes JR, Naud S. Reduction in cigarettes per day prospectively predicts making a quit attempt: a fine-grained secondary analysis of a natural history study. Nicotine Tob Res 2019 Apr 17;21(5):648-654. [doi: 10.1093/ntr/nty056] [Medline: 29579250]
18. Joseph AM, Bliss RL, Zhao F, Lando H. Predictors of smoking reduction without formal intervention. Nicotine Tob Res 2005 Apr;7(2):277-282. [doi: 10.1080/14622200500056176] [Medline: 16036285]
19. Mooney ME, Johnson EO, Breslau N, Bierut LJ, Hatsukami DK. Cigarette smoking reduction and changes in nicotine dependence. Nicotine Tob Res 2011 Jun;13(6):426-430. [doi: 10.1093/ntr/ntr019] [Medline: 21367813]

XSL•FO
RenderX

20. Yingst J, Wang X, Lopez AA, et al. Changes in nicotine dependence among smokers using electronic cigarettes to reduce cigarette smoking in a randomized controlled trial. Nicotine Tob Res 2023 Feb 9;25(3):372-378. [doi: 10.1093/ntr/ntac153] [Medline: 35752091]
21. Hughes JR, Cummings KM, Hyland A. Ability of smokers to reduce their smoking and its association with future smoking cessation. Addiction 1999 Jan;94(1):109-114. [doi: 10.1046/j.1360-0443.1999.9411097.x] [Medline: 10665102]
22. McDermott L, Dobson A, Owen N. Smoking reduction and cessation among young adult women: a 7-year prospective analysis. Nicotine Tob Res 2008 Sep;10(9):1457-1466. [doi: 10.1080/14622200802323241] [Medline: 19023837]
23. Broms U, Korhonen T, Kaprio J. Smoking reduction predicts cessation: longitudinal evidence from the Finnish adult twin cohort. Nicotine Tob Res 2008 Mar;10(3):423-427. [doi: 10.1080/14622200801888988] [Medline: 18324560]
24. Klemperer EM, Hughes JR. Does the magnitude of reduction in cigarettes per day predict smoking cessation? A qualitative review. Nicotine Tob Res 2016 Jan;18(1):88-92. [doi: 10.1093/ntr/ntv058] [Medline: 25744970]
25. Chaiton M, Diemert L, Cohen JE, et al. Estimating the number of quit attempts it takes to quit smoking successfully in a longitudinal cohort of smokers. BMJ Open 2016 Jun 9;6(6):e011045. [doi: 10.1136/bmjopen-2016-011045] [Medline: 27288378]
26. Lucchiari C, Masiero M, Mazzocco K, et al. Benefits of e-cigarettes in smoking reduction and in pulmonary health among chronic smokers undergoing a lung cancer screening program at 6 months. Addict Behav 2020 Apr;103:106222. [doi: 10.1016/j.addbeh.2019.106222] [Medline: 31838445]
27. Rennard SI, Daughton D, Fujita J, et al. Short-term smoking reduction is associated with reduction in measures of lower respiratory tract inflammation in heavy smokers. Eur Respir J 1990 Jul;3(7):752-759. [Medline: 2261963]
28. Hatsukami DK, Kotlyar M, Allen S, et al. Effects of cigarette reduction on cardiovascular risk factors and subjective measures. Chest 2005 Oct;128(4):2528-2537. [doi: 10.1378/chest.128.4.2528] [Medline: 16236919]
29. Polosa R, Morjaria J, Caponnetto P, et al. Effect of smoking abstinence and reduction in asthmatic smokers switching to electronic cigarettes: evidence for harm reversal. Int J Environ Res Public Health 2014 May 8;11(5):4965-4977. [doi: 10.3390/ijerph110504965] [Medline: 24814944]
30. Farsalinos K, Cibella F, Caponnetto P, et al. Effect of continuous smoking reduction and abstinence on blood pressure and heart rate in smokers switching to electronic cigarettes. Intern Emerg Med 2016 Feb;11(1):85-94. [doi: 10.1007/s11739-015-1361-y] [Medline: 26749533]
31. Polosa R. Examining the evidence for the health impact of combustion-free products: progress and prospects for tobacco harm reversal and reduction. Intern Emerg Med 2021 Nov;16(8):2043-2046. [doi: 10.1007/s11739-021-02837-2] [Medline: 34524625]
32. Chang JT, Anic GM, Rostron BL, Tanwar M, Chang CM. Cigarette smoking reduction and health risks: a systematic review and meta-analysis. Nicotine Tob Res 2021 Mar 19;23(4):635-642. [doi: 10.1093/ntr/ntaa156] [Medline: 32803250]
33. Roulet S, Chrea C, Kanitscheider C, Kallischnigg G, Magnani P, Weitkunat R. Potential predictors of adoption of the Tobacco Heating System by U.S. adult smokers: an actual use study. F1000Res 2019;8:214. [doi: 10.12688/f1000research.17606.2] [Medline: 31559015]
34. Campbell C, Feehan M, Kanitscheider C, Makena PS, Cai J, Baxter SA. Designing studies to inform tobacco harm reduction: learnings from an oral nicotine pouch actual use pilot study. JMIR Form Res 2022 Aug 19;6(8):e37573. [doi: 10.2196/37573] [Medline: 35984682]
35. Becker E, McCaffrey S, Lewis J, Vansickel A, Larson E, Sarkar M. Characterization of ad libitum use behavior of On! Nicotine pouches. Am J Health Behav 2023;47(3):428-449. [doi: 10.5993/AJHB.47.3.1] [Medline: 37596760]
36. Burris JL, Wahlquist AE, Alberg AJ, et al. A longitudinal, naturalistic study of U.S. smokers' trial and adoption of snus. Addict Behav 2016 Dec;63:82-88. [doi: 10.1016/j.addbeh.2016.07.008] [Medline: 27450153]
37. Meier E, Burris JL, Wahlquist A, et al. Perceptions of snus among US adult smokers given free product. Nicotine Tob Res 2017 Dec 13;20(1):22-29. [doi: 10.1093/ntr/ntw392] [Medline: 28034999]
38. Harlow AF, Cho J, Tackett AP, et al. Motivations for e-cigarette use and associations with vaping frequency and smoking abstinence among adults who smoke cigarettes in the United States. Drug Alcohol Depend 2022 Sep 1;238:109583. [doi: 10.1016/j.drugalcdep.2022.109583] [Medline: 35907310]
39. Hyland A, Ambrose BK, Conway KP, et al. Design and methods of the Population Assessment of Tobacco and Health (PATH) Study. Tob Control 2017 Jul;26(4):371-378. [doi: 10.1136/tobaccocontrol-2016-052934] [Medline: 27507901]
40. Morris P, McDermott S, Chapman F, et al. Reductions in biomarkers of exposure to selected harmful and potentially harmful constituents following exclusive and partial switching from combustible cigarettes to myblu. Intern Emerg Med 2022 Mar;17(2):397-410. [doi: 10.1007/s11739-021-02813-w] [Medline: 34435305]
41. Lehmann K, Kuhn S, Reimer J. Electronic cigarettes in Germany: patterns of use and perceived health improvement. Eur Addict Res 2017;23(3):136-147. [doi: 10.1159/000475986] [Medline: 28618425]
42. Carpenter MJ, Wahlquist AE, Dahne J, et al. Effect of unguided e-cigarette provision on uptake, use, and smoking cessation among adults who smoke in the USA: a naturalistic, randomised, controlled clinical trial. EClinicalMedicine 2023 Sep;63:102142. [doi: 10.1016/j.eclinm.2023.102142] [Medline: 37753443]

XSL•FO
RenderX

43. Goldenson NI, Shiffman S, Sembower MA, Selya A, Pype S, Black RA. Evaluating the effect of the JUUL2 system with 5 flavors on cigarette smoking and tobacco product use behaviors among adults who smoke cigarettes: 6-week actual use study. Interact J Med Res 2025 Mar 26;14:e60620. [doi: 10.2196/60620] [Medline: 40138686]
44. Selya A, Kim S, Shiffman S, Goldenson NI. Association of use of menthol- versus tobacco-flavored ENDS with switching completely away from cigarettes and differences by menthol cigarette smoking. Subst Use Misuse 2025;60(3):311-318. [doi: 10.1080/10826084.2024.2422963] [Medline: 39491336]
45. Zhu SH, Zhuang YL, Wong S, Cummins SE, Tedeschi GJ. E-cigarette use and associated changes in population smoking cessation: evidence from US current population surveys. BMJ 2017 Jul 26;358:j3262. [doi: 10.1136/bmj.j3262] [Medline: 28747333]
46. Chen JC. Flavored e-cigarette use and cigarette smoking reduction and cessation-a large national study among young adult smokers. Subst Use Misuse 2018 Oct 15;53(12):2017-2031. [doi: 10.1080/10826084.2018.1455704] [Medline: 29624135]
47. Farsalinos K, Siakas G, Poulas K, Voudris V, Merakou K, Barbouni A. E-cigarette use is strongly associated with recent smoking cessation: an analysis of a representative population sample in Greece. Intern Emerg Med 2019 Sep;14(6):835-842. [doi: 10.1007/s11739-018-02023-x] [Medline: 30637600]
48. Farsalinos KE, Barbouni A. Association between electronic cigarette use and smoking cessation in the European Union in 2017: analysis of a representative sample of 13 057 Europeans from 28 countries. Tob Control 2021 Jan;30(1):71-76. [doi: 10.1136/tobaccocontrol-2019-055190] [Medline: 32015151]
49. Papadakis S, Katsaounou P, Kyriakos CN, et al. Quitting behaviours and cessation methods used in eight European Countries in 2018: findings from the EUREST-PLUS ITC Europe Surveys. Eur J Public Health 2020 Jul 1;30(Suppl_3):iii26-iii33. [doi: 10.1093/eurpub/ckaa082] [Medline: 32918825]
50. Coleman B, Rostron B, Johnson SE, et al. Transitions in electronic cigarette use among adults in the Population Assessment of Tobacco and Health (PATH) Study, Waves 1 and 2 (2013-2015). Tob Control 2019 Jan;28(1):50-59. [doi: 10.1136/tobaccocontrol-2017-054174] [Medline: 29695458]
51. Buu A, Hu YH, Piper ME, Lin HC. The association between e-cigarette use characteristics and combustible cigarette consumption and dependence symptoms: results from a national longitudinal study. Addict Behav 2018 Sep;84:69-74. [doi: 10.1016/j.addbeh.2018.03.035] [Medline: 29627636]
52. Brouwer AF, Jeon J, Hirschtick JL, et al. Transitions between cigarette, ENDS and dual use in adults in the PATH study (waves 1–4): multistate transition modelling accounting for complex survey design. Tob Control 2022 May;31(3):424-431. [doi: 10.1136/tobaccocontrol-2020-055967]
53. Berry KM, Reynolds LM, Collins JM, et al. E-cigarette initiation and associated changes in smoking cessation and reduction: the Population Assessment of Tobacco and Health Study, 2013-2015. Tob Control 2019 Jan;28(1):42-49. [doi: 10.1136/tobaccocontrol-2017-054108] [Medline: 29574448]
54. Glasser AM, Vojjala M, Cantrell J, et al. Patterns of e-cigarette use and subsequent cigarette smoking cessation over 2 years (2013/2014–2015/2016) in the Population Assessment of Tobacco and Health Study. Nicotine Tob Res 2021 Mar 19;23(4):669-677. [doi: 10.1093/ntr/ntaa182] [Medline: 32939555]
55. Abi Nehme AM, Lou X, Yan X, Lee JH, Salloum RG. Transition to smoking cessation among dual cigarette and e-cigarette users in the population assessment of tobacco and health study, Waves 3 and 4 (2015–2017). Addict Behav 2022 Jun;129:107284. [doi: 10.1016/j.addbeh.2022.107284] [Medline: 35217415]
56. Kaplan B, Galiatsatos P, Breland A, Eissenberg T, Cohen JE. Effectiveness of ENDS, NRT and medication for smoking cessation among cigarette-only users: a longitudinal analysis of PATH Study wave 3 (2015-2016) and 4 (2016-2017), adult data. Tob Control 2023 May;32(3):302-307. [doi: 10.1136/tobaccocontrol-2020-056448] [Medline: 34526410]
57. Persoskie A, O'Brien EK, Poonai K. Perceived relative harm of using e-cigarettes predicts future product switching among US adult cigarette and e-cigarette dual users. Addiction 2019 Dec;114(12):2197-2205. [doi: 10.1111/add.14730] [Medline: 31278802]
58. Kasza KA, Edwards KC, Tang Z, et al. Correlates of tobacco product cessation among youth and adults in the USA: findings from the PATH Study Waves 1-3 (2013-2016). Tob Control 2020 May;29(Suppl 3):s203-s215. [doi: 10.1136/tobaccocontrol-2019-055255] [Medline: 32321854]
59. Benmarhnia T, Pierce JP, Leas E, et al. Can e-cigarettes and pharmaceutical aids increase smoking cessation and reduce cigarette consumption? Findings from a Nationally Representative Cohort of American Smokers. Am J Epidemiol 2018 Nov 1;187(11):2397-2404. [doi: 10.1093/aje/kwy129] [Medline: 29955810]
60. Han DH, Lee SH, Seo DC. Within-person longitudinal associations between electronic nicotine delivery systems use and smoking cessation efforts among US continuing adult cigarette smokers. Nicotine Tob Res 2022 Mar 1;24(4):590-597. [doi: 10.1093/ntr/ntab232] [Medline: 34758071]
61. Kalkhoran S, Chang Y, Rigotti NA. Electronic cigarette use and cigarette abstinence over 2 years among U.S. smokers in the Population Assessment of Tobacco and Health Study. Nicotine Tob Res 2020 Apr 21;22(5):728-733. [doi: 10.1093/ntr/ntz114] [Medline: 31298296]
62. Anderson MK, Whitted L, Mason TB, Pang RD, Tackett AP, Leventhal AM. Characterizing different-flavored e-cigarette solutions from user-reported sensory attributes and appeal. Exp Clin Psychopharmacol 2023 Feb;31(1):46-56. [doi: 10.1037/pha0000563] [Medline: 35467923]

XSL•FO
RenderX

63. Baker AN, Bakke AJ, Branstetter SA, Hayes JE. Harsh and sweet sensations predict acute liking of electronic cigarettes, but flavor does not affect acute nicotine intake: a pilot laboratory study in men. Nicotine Tob Res 2021 Mar 19;23(4):687-693. [doi: 10.1093/ntr/ntaa209] [Medline: 33047127]
64. Berg CJ. Preferred flavors and reasons for e-cigarette use and discontinued use among never, current, and former smokers. Int J Public Health 2016 Mar;61(2):225-236. [doi: 10.1007/s00038-015-0764-x] [Medline: 26582009]
65. Gades MS, Alcheva A, Riegelman AL, Hatsukami DK. The role of nicotine and flavor in the abuse potential and appeal of electronic cigarettes for adult current and former cigarette and electronic cigarette users: a systematic review. Nicotine Tob Res 2022 Aug 6;24(9):1332-1343. [doi: 10.1093/ntr/ntac073] [Medline: 35305014]
66. Hayes JE, Baker AN. Flavor science in the context of research on electronic cigarettes. Front Neurosci 2022;16:918082. [doi: 10.3389/fnins.2022.918082] [Medline: 35968379]
67. Jones DM, Ashley DL, Weaver SR, Eriksen MP. Flavored ENDS use among adults who have used cigarettes and ENDS, 2016-2017. Tob Regul Sci 2019 Nov;5(6):518-531. [doi: 10.18001/TRS.5.6.4] [Medline: 31696149]
68. Li L, Borland R, Cummings KM, et al. How does the use of flavored nicotine vaping products relate to progression toward quitting smoking? Findings from the 2016 and 2018 ITC 4CV surveys. Nicotine Tob Res 2021 Aug 18;23(9):1490-1497. [doi: 10.1093/ntr/ntab033] [Medline: 33631007]
69. Rest EC, Brikmanis KN, Mermelstein RJ. Preferred flavors and tobacco use patterns in adult dual users of cigarettes and ENDS. Addict Behav 2022 Feb;125:107168. [doi: 10.1016/j.addbeh.2021.107168] [Medline: 34772504]
70. Zare S, Nemati M, Zheng Y. A systematic review of consumer preference for e-cigarette attributes: flavor, nicotine strength, and type. PLoS ONE 2018;13(3):e0194145. [doi: 10.1371/journal.pone.0194145] [Medline: 29543907]
71. Zare S, Zheng Y. Consumer preferences for e-cigarette flavor, nicotine strength, and type: evidence from Nielsen Scanner Data. Nicotine Tob Res 2021 May 4;23(5):823-828. [doi: 10.1093/ntr/ntaa238] [Medline: 33245356]
72. Evans AT, Henderson KC, Geier A, et al. What motivates smokers to switch to ENDS? A qualitative study of perceptions and use. Int J Environ Res Public Health 2020;17(23):8865. [doi: 10.3390/ijerph17238865]
73. Li W, Vargas-Rivera M, Eissenberg TE, Shihadeh A, Talih S, Maziak W. Effect of menthol/mint-flavored pods on young JUUL e-cigarette users' subjective experience, puffing behavior, and nicotine exposure: a pilot study. Drug Alcohol Depend 2022 Aug 1;237:109516. [doi: 10.1016/j.drugalcdep.2022.109516] [Medline: 35679691]
74. Soulakova JN, Hartman AM, Liu B, Willis GB, Augustine S. Reliability of adult self-reported smoking history: data from the tobacco use supplement to the Current Population Survey 2002–2003 Cohort. Nicotine Tob Res 2012 Aug;14(8):952-960. [doi: 10.1093/ntr/ntr313] [Medline: 22318688]

Abbreviations

CPD: cigarettes per day
e-diary: electronic diary
ENDS: electronic nicotine delivery system
NHIS: National Health Interview Survey
NRT: nicotine replacement therapies
PATH: Population Assessment of Tobacco and Health
SP: study product
TNP: tobacco and nicotine-containing products

Edited by A Mavragani; submitted 15.04.25; peer-reviewed by B Afolabi, J Karelitz; revised version received 24.06.25; accepted 14.07.25; published 25.08.25.

Please cite as:
Roulet S, Kanitscheider C, Magnani P, Kallischnigg G
Evaluating Use Patterns of a Closed Electronic Nicotine Delivery System Among Adults in the United States Who Smoke Cigarettes Daily: 8-Week Actual Use Study
JMIR Form Res 2025;9:e76019
URL: https://formative.jmir.org/2025/1/e76019
doi:10.2196/76019

© Steve Roulet, Claudia Kanitscheider, Pierpaolo Magnani, Gerd Kallischnigg. Originally published in JMIR Formative Research (https://formative.jmir.org), 25.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete

XSL•FO
RenderX

bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Identifying Adverse Drug Events in Clinical Text Using Fine-Tuned Clinical Language Models: Machine Learning Study

Elizaveta Kopacheva[1], PhD; Aron Henriksson[2], Prof Dr; Hercules Dalianis[2], Prof Dr; Tora Hammar[3], Prof Dr; Alisa Lincke[1], PhD

[1]Department of Computer Science and Media Technology, Faculty of Technology, Linnaeus University, Universitetsplatsen 1, Växjö, Sweden

[2]Department of Computer and Systems Sceince, Stockholm University, Kista, Sweden

[3]eHealth Institute, Department of Medicine and Optometry, Faculty of Health and Life Sciences, Linnaeus University, Växjö, Sweden

Corresponding Author:
Elizaveta Kopacheva, PhD
Department of Computer Science and Media Technology, Faculty of Technology, Linnaeus University, Universitetsplatsen 1, Växjö, Sweden

Abstract

Background: Medications are essential for health care but can cause adverse drug events (ADEs), which are harmful and sometimes fatal. Detecting ADEs is a challenging task because they are often not documented in the structured data of electronic health records (EHRs). There is a need for automatically extracting ADE-related information from clinical notes, as manual review is labor-intensive and time-consuming.

Objective: This study aims to fine-tune the pretrained clinical language model, Swedish Deidentified Clinical Bidirectional Encoder Representations from Transformers (SweDeClin-BERT), for medical named entity recognition (NER) and relation extraction (RE) tasks, and to implement an integrated NER-RE approach to more effectively identify ADEs in clinical notes from clinical units in Sweden. The performance of this approach is compared with our previous machine learning method, which used conditional random fields (CRFs) and random forest (RF).

Methods: A subset of clinical notes from the Stockholm EPR (Electronic Patient Record) Corpus, dated 2009 - 2010, containing suspected ADEs based on *International Classification of Diseases, 10th Revision* (*ICD-10*) codes in the A.1 and A.2 categories was randomly sampled. These notes were annotated by a physician with ADE-related entities and relations following the ADE annotation guidelines. We fine-tuned the SweDeClin-BERT model for the NER and RE tasks and implemented an integrated NER-RE pipeline to extract entities and relationships from clinical notes. The models were evaluated using 395 clinical notes from clinical units in Sweden. The NER-RE pipeline was then applied to classify the clinical notes as containing or not containing ADEs. In addition, we conducted an error analysis to better understand the model's behavior and to identify potential areas for improvement.

Results: In total, 62% of notes contained an explicit description of an ADE, indicating that an ADE-related *ICD-10* code alone does not ensure detailed event documentation. The fine-tuned SweDeClin-BERT model achieved an F_1-score of 0.845 for NER and 0.81 for RE task, outperforming the baseline models (CRFs for NER and random forests for RE). In particular, the RE task showed a 53% improvement in macro-average F_1-score compared to the baseline. The integrated NER-RE pipeline achieved an overall F_1-score of 0.81.

Conclusions: Using a domain-specific language model like SweDeClin-BERT for detecting ADEs in clinical notes demonstrates improved classification performance (0.77 in strict and 0.81 in relaxed mode) compared to conventional machine learning models like CRFs and RF. The proposed fine-tuned ADE model requires further refinement and evaluation on annotated clinical notes from another hospital to evaluate the model's generalizability. In addition, the annotation guidelines should be revised, as there is an overlap of words between the Finding and Disorder entity categories, which were not consistently distinguished by the annotators. Furthermore, future work should address the handling of compound words and split entities to better capture context in the Swedish language.

(JMIR Form Res 2025;9:e71949) doi:10.2196/71949

KEYWORDS

electronical health records; adverse drug events; domain-specific language models; BERT; SweDeClin-BERT

XSL•FO
RenderX

Introduction

Medications are an essential part of health care; they improve the lives of many but can also cause unwanted effects. Adverse drug events (ADEs) are harmful events in patients that are related to the use of a medication, including adverse drug reactions (ADRs, also called "side effects"). ADEs are common and cause hardship or even death to patients and a significant cost to society [1]. However, numbers on the prevalence and cost differ extensively between studies in different settings and depending on how they are measured [2,3]. One systematic review estimated that 19% of patients are affected by ADEs during hospitalization [2]. Another systematic review estimated that rates of readmission to hospitals due to medications varied from 3% to 64%, with the majority of them being preventable [4]. Among outpatients, it has been estimated that 2% have preventable ADRs [5]. The widespread use of electronic health records (EHRs) provides large amounts of data that hold great potential for pharmacovigilance and monitoring of ADEs in the population [6]. Information about actual or suspected ADEs can be documented in the EHR in different ways, either as diagnosis codes (structured data) or written in free text (unstructured data). In Sweden, ADEs can be detected using clinical decision support systems (CDSS) such as the Janusmed knowledge databases [7] and Bikt (part of the Swedish Information Services [SIL]) [8], which operate on structured data like medication lists. However, these systems rely on rule-based algorithms and substance-level drug knowledge, often generating an excessive number of alerts—leading to alert fatigue—and are currently integrated into only a limited number of EHR systems. All suspected ADEs should be reported to national authorities, such as the Swedish Medical Products Agency, or by conducting occasional manual chart reviews. There is currently no routine or automated method for monitoring and extracting suspected ADEs directly from unstructured clinical notes in EHRs in Sweden.

Recent advances in technological development and artificial intelligence (AI) offer improved possibilities for analyzing large amounts of data, which is useful in pharmacovigilance, among other [9,10]. Over the years, researchers have dedicated extensive efforts to develop effective methods and algorithms to detect and recognize ADEs within the information stored in both structured data (like diagnosis codes in EHRs) and unstructured data, such as clinical notes, which is more challenging. Research in the field has evolved alongside advancements in natural language processing (NLP) and machine learning techniques. Specifically, earlier approaches [11,12] are knowledge-based approaches that rely on predefined rules and medical dictionaries to extract ADEs from clinical notes. Later on, improvements in NLP leveraged linguistic text processing, and medical ontologies began to be widely used. More recent studies [13-16] are data-driven approaches that have relied on supervised machine learning and, more often, deep learning techniques for better classification of ADEs using annotated clinical notes. Recent advancements in AI, such as large language models (LLMs), transformer-based models, knowledge graphs, and transfer learning, have demonstrated state-of-the-art performance in identifying ADEs in clinical notes [17-21]. LLM is a sub-branch of the NLP branch of AI designed to understand, generate, and manipulate text for a variety of language-related tasks [22].

Previous research on detecting ADEs from clinical texts shows it is a 3-fold NLP task [23]. The first subtask, called named entity recognition (NER), is to extract medical entities relevant to ADEs from text (such as medication, disorders, body parts, symptoms, etc); the second subtask, called relation extraction (RE), is to establish the relation between extracted entities (eg, a reason for prescribing a drug, causal relationship between an ADE and a drug, etc); and the last subtask (end-to-end integration) is to combine NER-RE models to decide whether an ADE is present in the given clinical text or not based on identified medical entities (NER model) and relations between them (RE model). NER and RE tasks can be implemented using two distinct approaches: (1) classical machine learning models (MLMs), such as support vector machines (SVMs), decision trees, conditional random fields (CRFs), and neural networks, among others; and (2) large language models (LLMs), such as Bidirectional Encoder Representations from Transformers (BERT; Google AI Language), Generative Pre-trained Transformer (GPT; OpenAI), RoBERTa (Meta AI), XLNet (Google Brain and Carnegie Mellon University), T5 (Google Research), among others.

The previous research results are difficult to compare because each study uses different entity, relation types, datasets, data sources (case reports, EHRs), and language (mainly in English) [15,24-29]. Nevertheless, using the MLM approach, researchers have achieved moderate accuracy in the NER task, with micro-average F_1-scores reported at 0.80 [24], 0.82 [25], and 0.84 [15]. However, accuracy in the RE task remains low to moderate, with a macro-average F_1-score of 0.28 [24] and a micro-average F_1-score of 0.846 [25], due to the model's limited contextual and semantic understanding of the text. The joined NER-RE tasks achieved a micro average F_1-score of 0.869 [26]. Among those studies, only one study used clinical notes in the Swedish language [24], while all the others were in English. The LLMs achieved the highest accuracy using fine-tuned BERT on ADE detection in clinical notes in the English language (micro average F_1-scores -0.93 for NER, 0.96 for RE task, and 0.895 for joined NER-RE [27] in relaxed match). The fine-tuned GPT 3.5 model achieved an F_1-score of 0.70 (strict match) and 0.81 (relaxed match) [29] for the NER-RE (end-to-end) task in detecting ADEs in vaccine adverse event reports, and fine-tuned the Curie (GPT 3.5) model achieved 0.795 for NER and 0.74 for RE [28]. A recent study on classifying ADE reports in the Swedish language using GPT 3.5 achieved 0.71 F_1-score [30]. Although the overall performance of the LLMs regarding NER and RE tasks is high, the micro-averaged F_1-scores of each entity and relation type varied significantly. The lowest performance remains for detecting the ADE relations.

Our previous study on detecting ADEs in clinical notes using MLMs has achieved a micro average F_1-score of 0.80 for NER and 0.28 for RE [24], showing a low performance in identifying relationships between medical entities. Unlike the MLM approach, using domain-specific variants of BERT allows the

XSL•FO
RenderX

model to consider the full context of a word within a sentence, including distant dependencies and subtle linguistic nuances, which are crucial in accurately identifying and classifying entities in free-text medical records. In addition, BERT-based models benefit from pretrained embeddings that provide richer word representations in comparison with MLMs. Hence, even though transformer-based models like BERT are known to incorporate contextual information effectively, it remains critical to empirically evaluate their advantages on specific tasks such as ADE extraction from real-world EHR data. Such studies not only quantify performance improvements over traditional baselines but also identify persistent challenges, guiding further refinements in annotation schemes and model architectures.

Swedish Deidentified Clinical Bidirectional Encoder Representations from Transformers (SweDeClin-BERT) is a Swedish BERT model based on the KB-BERT model [31], which has been further pretrained on around 2 million deidentified and pseudonymized patient notes (17.9 GB) from over 500 clinical units at Karolinska University Hospital in Sweden [32]. This model showed good performance with 0.87 F_1-score in the NER task, identifying entities belonging to sensitive information such as social security numbers, name, phone numbers, etc, for deidentifying clinical notes [33], *International Classification of Diseases, 10th Revision (ICD-10)* diagnosis coding with 0.94 F_1-score [34], and classification triage of adverse drug reaction reports with 0.72 F_1-score [35]. All these results show promising potential for enhancing RE performance for ADEs. This study aims to fine-tune the pretrained clinical language model, SweDeClin-BERT, for medical NER and RE tasks, and to implement an integrated NER-RE approach to more effectively identify ADEs in clinical notes from clinical units in Sweden. The performance of this approach will be compared to our previous machine learning method, which used conditional random fields (CRFs) and random forest (RF) [22].

Methods

Overview

Our study follows the JMIR study design category of "Machine learning predictive models in biomedical research", and emphasizes a transparent, end-to-end approach—from data collection to model evaluation—aligning with best practices in reporting [36] to ensure clarity, reproducibility, and interpretation of model performance.

This study used transfer learning to approach 2 learning tasks, NER and RE, using the SweDeClin-BERT model. Transfer learning is the process of pretraining a model on a large-scale dataset and then reusing this model to improve performance on a related task [37]. This approach is particularly beneficial for our ADEs detection task, given the limited scale of annotated data and the complexity of free text, which poses significant challenges for machine comprehension. Encoder models like BERT are naturally suited to, and can be readily fine-tuned for, classification tasks, and SweDeClin-BERT has the advantage of having been adapted to the clinical domain. While it could be interesting to evaluate larger decoder models on these tasks, we consider this to be out of the scope of the current study. At present, there is no decoder mode that has been adapted to Swedish clinical text. Moreover, given the sensitive nature of the data, one would have to use an open-weight model that can be hosted in a secure on-premises environment with access to significant computational resources in order to allow for fine-tuning and inference. In this study, the performance of the fine-tuned SweDeClin-BERT models (NER and RE) is compared with our previous method as a baseline [24].

The baseline models are CRFs for the NER task and RF for the RE task used in our previous study [24] for identifying ADEs in clinical notes. The used CRF algorithm for the NER task was with a rich feature set, including token string, lemma, part-of-speech tags, capitalization, compound splitting, and terminology matching, along with distributional semantic features from various Distributional Semantic Models (DSMs) using different context window sizes. The used RF model for the RE task was with a hierarchical or cascading versus a flat classification approach. The method based on the hierarchical approach first classified the clinical notes into Containing a relation and No relation classes and then classified specific relation types. The feature set for the random forest classifier included one-hot encoded entity classes, entity unigrams and bigrams, the distance between entities, and context representation using multiple DSMs. Class imbalance was not addressed in the baseline approach. The issue of class imbalance, especially its effect on rare relation types such as ADE, ADE outcome, and ADE cause, is discussed in a previous study [24]. This limitation in the baseline design may affect the performance comparison and should be taken into account.

The following research question is addressed in this study: "How accurately can the clinical language model SweDeClinBERT detect ADEs in clinical notes compared to a traditional machine learning approach such as CRFs and RF?"

To answer this question, the pretrained SweDeClin-BERT model is fine-tuned for the NER and RE tasks, respectively, on the same dataset as the baseline model (395 annotated Swedish clinical notes called Stockholm EPR ADE Corpus), which was created in our previous study [24], and the results were compared to those of the baseline model. After the fine-tuning, 2 models (NER model and RE model) were stacked to obtain results for the integrated NER-RE task. In addition, the fine-tuned models (NER-RE pipeline) were evaluated on the note level with 2 classes: notes containing ADEs and notes without ADEs.

In order to compare the SweDeClin-BERT model performance with the baseline model, the dataset was divided once into training, validation, and test sets with splits of 70%, 10%, and 20%, respectively and used for all tasks: NER, RE, NER-RE, and ADE detection in clinical notes using NER-RE models. Grid search strategy [38] was used for hyperparameter tuning. Standard evaluation metrics such as Precision, Recall, F_1-score, F_2-score, and micro- and macro-averaged metrics were recorded for comparison with the baseline model.

Finally, we conducted a thorough model error analysis (for NER and RE models) by examining confusion matrices [39] and

XSL•FO
RenderX

identifying terms most frequently misclassified. This analysis aims to guide future improvements in model training and input data preparation, helping to refine entity definitions and relationships, and ultimately enhance the model's overall performance toward achieving the overarching goal. We carried out our experiments using Python 3, Jupyter Notebook, and spaCy library [40].

Dataset

The training dataset Stockholm EPR ADE Corpus was obtained from the research infrastructure Health Bank [32] at Stockholm University. The clinical notes were originally sourced from Karolinska University Hospital, a large tertiary care facility in Sweden, and reflect documentation practices within the Stockholm regional health care system. The Stockholm EPR ADE Corpus (395 clinical notes, dated 2009 - 2010) has previously been annotated by a physician for ADE-named entities (NEs) and the relations between them [24] and anonymized in this study, enabling us to evaluate the performance of the fine-tuned SweDeClin-BERT in identifying ADEs. A summary of the clinical notes used in the Stockholm EPR ADE Corpus is presented in Table 1. The dataset is heterogeneous and is not limited to a specific clinical domain. For more detailed information, please refer to our previous study [24].

Table . Description of the clinical notes used in this study.

Metric	Value
Total number of notes	395
Note categories	Physician notes from admission note, patient history, hypersensitivity and drug info, assessment, and discharge notes.
ICD-Codes	Notes with the following *International Classification of Diseases, 10th Revision* codes for ADE[a]: Z931, Y579, M819, M809, I952, T509, F050, T887, L271
Min character length	59
Max character length	4372
Mean character length	1014.78
Min words	7
Max words	645
Mean words	150.5
Mean number of sentences per note	14
Lexical diversity (avg. Type-token ratio [TTR])	0.70

[a]ADE: adverse drug event

The annotated NE categories include Drug, Disorder, Finding, Body Structure, and ADE Cue, where: Drug is a medication; Disorder is a not momentary disease or abnormal condition with underlying pathological processes; Finding is a symptom reported by the patient, observation by the physician, or a result from medical examinations, including a non-pathological finding with medical relevance; Body structure is an anatomical part of the body; ADE cue is an indicator of an ADE without specifying its form (eg, side effect, overmedicate, overdosage, medication reaction or respond, and hypersensitivity). This entity is used to indicate the presence of an ADE that is not explicitly described in clinical notes, where information about the drug or drugs associated with the ADE is missing.

Table 2 summarizes the number of instances (annotated entities) in each class within the Stockholm ADE Corpus. Notably, NEs can consist of multiple words, which complicates the NE task. For example, Findings and Disorders often consist of 2 or more words (called compound instances), making these classes harder to classify. ADE Cue, being a minority class with only 330 instances, presents particular challenges due to its limited representation. In addition, the average interannotator agreement (IAA) for the ADE Cue class is 0.53 (as reported in [24]), indicating significant difficulty in consistently classifying these entities.

XSL•FO
RenderX

Table . Overview of named entity classes and their instance counts before and after anonymization, including distribution in the test set. Percentages represent the proportion of each class relative to the total number of entities.

Entity class	Before Anonymization, n (%)	After Anonymization, n (%)	Number of instances in the test set
Finding	3184 (35)	3142 (35)	533 (3.2)
Drug[a]	2680 (29)	2613 (29)	493 (29.8)
Disorder[a]	1866 (20)	1848 (20)	367 (22.2)
Body structure	1132 (12)	1116 (12)	205 (12.4)
ADE cue	341 (3,7)	330 (4)	55 (3.3)
Total	9203 (100)	9049 (100)	1653 (100)

[a]There are some data discrepancies compared to previous study [24] due to data deidentification and data preprocessing.

According to the ADE definition by the World Health Organization (WHO), the clinical notes with complete or incomplete information about the associated drug or drugs indicate the presence of ADE [41]. Recognizing the cues of ADE with complete or incomplete information (missing information about the associated drug or drugs) is essential for comprehensive pharmacovigilance and patient safety. Therefore, the Stockholm ADE Corpus includes annotations for 4 types of relations covering cases with complete or incomplete information between the mentioned entity types as shown in Table 3: Indication, Adverse Drug Event (ADE), ADE Outcome, and ADE Cause. These relations are defined as follows:

- Indication: A relation from Finding or Disorder to Drug, indicating that the drug was prescribed as a result of the identified Finding or Disorder.
- ADE: A relation from Drug to Finding or Disorder, specifying that a Finding or Disorder developed after taking the medication.
- ADE Outcome: A relation from ADE Cue to Finding or Disorder, highlighting the occurrence of the ADE without specifying its underlying cause.
- ADE Cause: A relation from Drug to ADE Cue, identifying that the drug is responsible for an ADE, without detailing the nature of the ADE.
- No Relation: is introduced after anonymization to represent unrelated entity pairs

Table . Overview of relation classes and their instance counts before and after anonymization, including distribution in the test set.

Relation class	Before anonymization, n (%)	After anonymization, n (%)	Number of instances in the test set, n (%)
Indication	1392 (53)	1203 (4)	220 (3.7)
Adverse Drug Event	855 (37)	810 (3)	156 (2.7)
ADE Outcome	144 (5)	139 (0.5)	23 (0.4)
ADE Cause	228 (9)	221 (0.7)	38 (0.6)
No relation	__[a]	28,412 (92)	5391 (92.5)
Total	2619 (100)	307,85 (100)	5828 (100)

[a]__: not applicable.

In essence, ADE Outcome and ADE Cause provide incomplete information about adverse drug reactions. These 2 classes are the least common relations in the corpus, with only 139 and 221 instances, respectively.

The dataset contains clinical notes as text files (in .txt format), anonymized annotated notes with BIO Schema tags [42] files (in .CoNLL format), and relation annotations files in Brat [40] format (.ann files). In BIO Schema, B indicates the beginning of an entity, I denotes tokens that are part of the same entity, and O marks tokens that do not belong to any entity (category Other). Specifically, each token is tagged with one of the following 11 labels: O, B-Finding, I-Finding, B-Drug, I-Drug, B-Disorder, I-Disorder, B-Body Structure, I-Body Structure, B-ADE Cue, and I-ADE Cue. Brat is an open-source web-based tool for text annotation [43] and was used by annotators in our previous study.

Missing Data and Data Cleaning

There were 5 empty notes out of 400; thus, the final number of notes is 395. Data cleaning involved filtering out notes without entities; however, since all notes contained entities, the dataset remained at 395 notes.

Outcome and Variables

The outcome of the NER task is to detect the ADE-related entities, including Disorder, Finding, Drug, Body structure, and ADE cause, and the outcome of the RE task is to correctly classify the ADE relation between the identified entities. For the integrated NER-RE task, the outcome is to determine

XSL•FO
RenderX

whether a clinical note contains a description of an ADE or not (yes or no).

For both the NER and RE tasks, the input to the SweDeClinBERT models consists of tokenized clinical text segments. The text is preprocessed and encoded using SweDeClinBERT's tokenizer, which splits the input into subword tokens and converts them into numerical embeddings that capture contextual meaning. This embedding is a numerical vector of variables, where each dimension captures distributed and context-dependent information, rather than corresponding to individual words, reflecting the semantic structure of the entire input.

Sample Size

The sample size in this study is fixed, consisting of 395 anonymized clinical notes, and is not intended to be representative of the broader patient population. Rather than supporting statistical generalization, this study is exploratory in nature. Its primary aim is to evaluate and compare the performance of a domain-specific NLP model (SweDeClinBERT) against a traditional MLM (conditional random fields) in the task of ADE detection.

Named-Entity Recognition Task

The NER task pipeline consists of training (fine-tuning as a token classification task) and evaluation phases, as shown in Figure 1.

Figure 1. Workflow of the named entity recognition task for extracting adverse drug event-related entities from clinical notes. NER: named entity recognition; NER-ADE: named entity recognition-adverse drug event; SweDeClin-BERT-RE-ADE: Swedish Deidentified Clinical Bidirectional Encoder Representations from Transformers-named-entity recognition of adverse drug event;

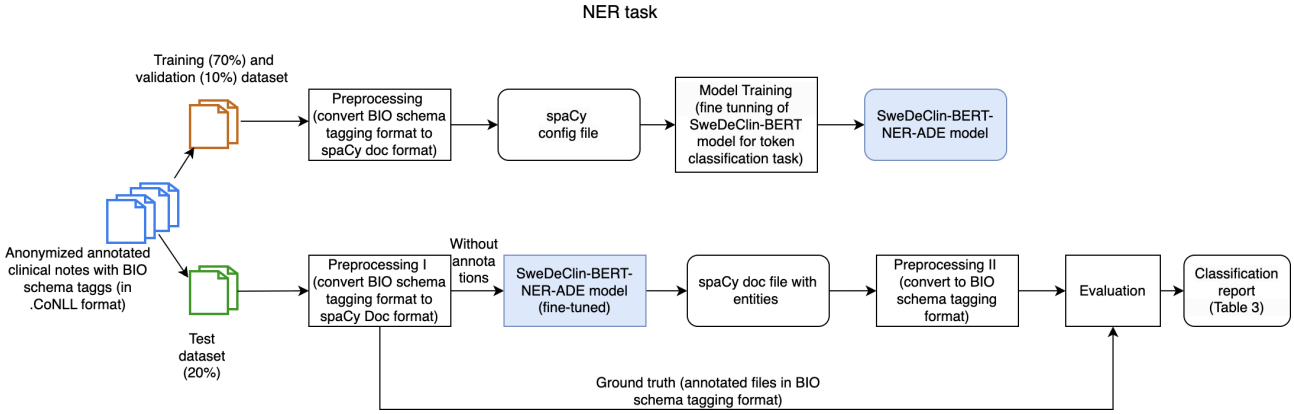

In the training phase, the texts were preprocessed by converting BIO Schema tagging to spaCy Doc type, which is suitable for model training. In addition, texts were segmented into smaller, overlapping chunks to accommodate the processing of longer texts that exceed the model's maximum input length. Each chunk was limited to 128 tokens in length, with a stride of 96 tokens. The spaCy config file is a file containing the configuration needed for fine-tuning the SweDeClin-BERT model. The SweDeClin-BERT model was fine-tuned for the NER task as a token classification task. During fine-tuning, BERT's output layer is modified to predict the appropriate tag for each token based on the contextual embeddings produced by the model. The different hyperparameters that were tuned are shown in Table 4.

Table . Hyperparameters used for fine-tuning the named entity recognition model.

Hyperparameter	Value
Batch size	128
Dropout rate	0.1
Optimizer	Adam
Learning rate	0.00005
Gradient accumulation	3
Gradient clipping threshold	1.0
Early stopping	With a patience value of 1600 steps
Weight decay	With L2=0.01

In the evaluation phase, the fine-tuned SweDeClin-BERT-NER-ADE model was used in prediction mode (input without annotations) to predict the entities. Afterward, the output of the SweDeClin-BERT-NER-ADE model (predicted entities) was converted back to BIO Schema tags format for measuring model performance with the ground truth (see classification report in Table 3).

XSL•FO
RenderX

Relation Extraction Task

Similarly to NER, the RE pipeline also consists of training (fine-tuning a sequence classification task) and an evaluation phase as shown in Figure 2.

Figure 2. Workflow of the relation extraction task for identifying relationships between adverse drug event-related entities. RE: relation extraction; SweDeClin-BERT-RE-ADE: Swedish Deidentified Clinical Bidirectional Encoder Representations from Transformers-relation extraction-adverse drug event;

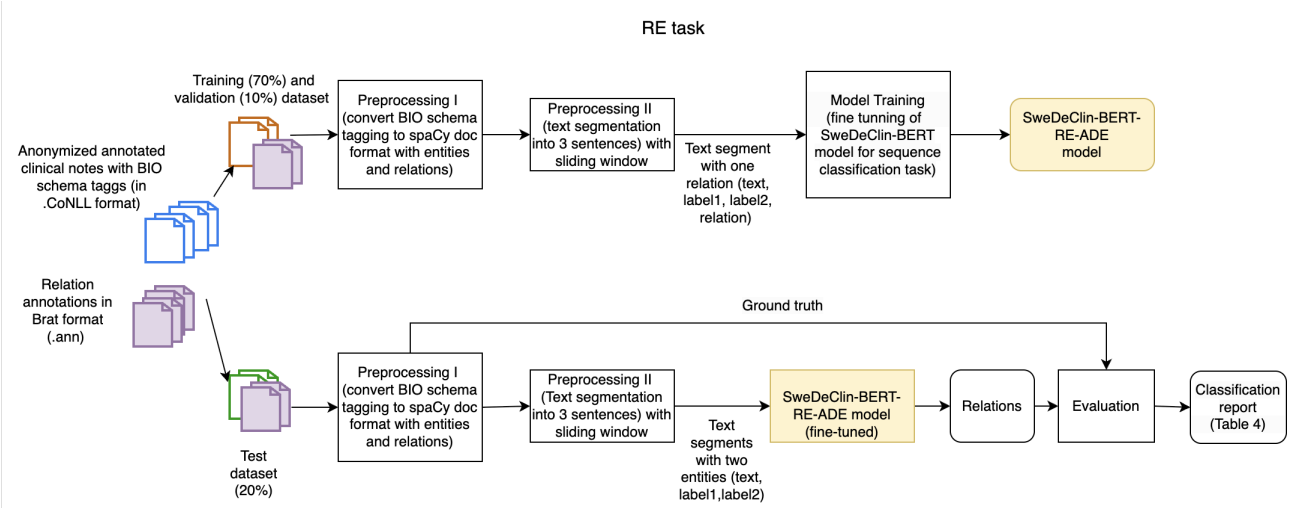

Given that the performance of the RE model is even more dependent on the context, preprocessing the texts for this task was crucial in the training phase. Since the maximum length processed by SweDeClin-BERT is 512 tokens, and relations between entities in the annotated texts often span no more than three sentences, the input texts were segmented into shorter passages of three sentences each, with a stride of one sentence. This segmentation preserved the contextual information necessary for accurate relation extraction while managing computational complexity effectively. The input to the model is a text segment containing only two entities and the relation between them. In addition to the primary relation classes of interest (Indication, ADE, ADE Cause, and ADE Outcome), a No relation class was added. This class was assigned to specific entity pairs where these primary relations could potentially exist but were not annotated. As shown in Table 3, the No relation class constitutes the majority class. The different hyperparameters that were tuned are shown in Table 5. After fine-tuning the SweDeClin-BERT model for the RE task, the best model was saved as the SweDeClin-BERT-RE-ADE model.

Table . Hyperparameters used for fine-tuning the extraction model.

Hyperparameter	Value
Batch size	32
Epochs	50
Patience for early stopping	3
Optimizer	Adam
Loss	Cross-entropy
Learning rate	2.575e-05
Weight decay	1e-4

In the evaluation phase, the fine-tuned SweDeClin-BERT-RE-ADE model was used in prediction mode (with annotated entities but without annotated relations) to predict the relations between 2 entities. Afterward, the SweDeClin-BERT-RE-ADE model performance was evaluated using the ground truth (see classification report in Table S1 in Multimedia Appendix 1) [24].

Integrated NER-RE Task

Two fine-tuned models (NER-model and RE-model) were stacked to obtain results for the integrated NER-RE task, as shown in Figure 3.

XSL•FO
RenderX

Figure 3. Integrated workflow combining entities and relation extraction tasks for adverse drug event detection on entity pairs. NER-RE: named entity recognition-relation extraction; SweDeClin-BERT-RE-ADE: Swedish De-identified Clinical Bidirectional Encoder Representations from Transformers-relation extraction-adverse drug event

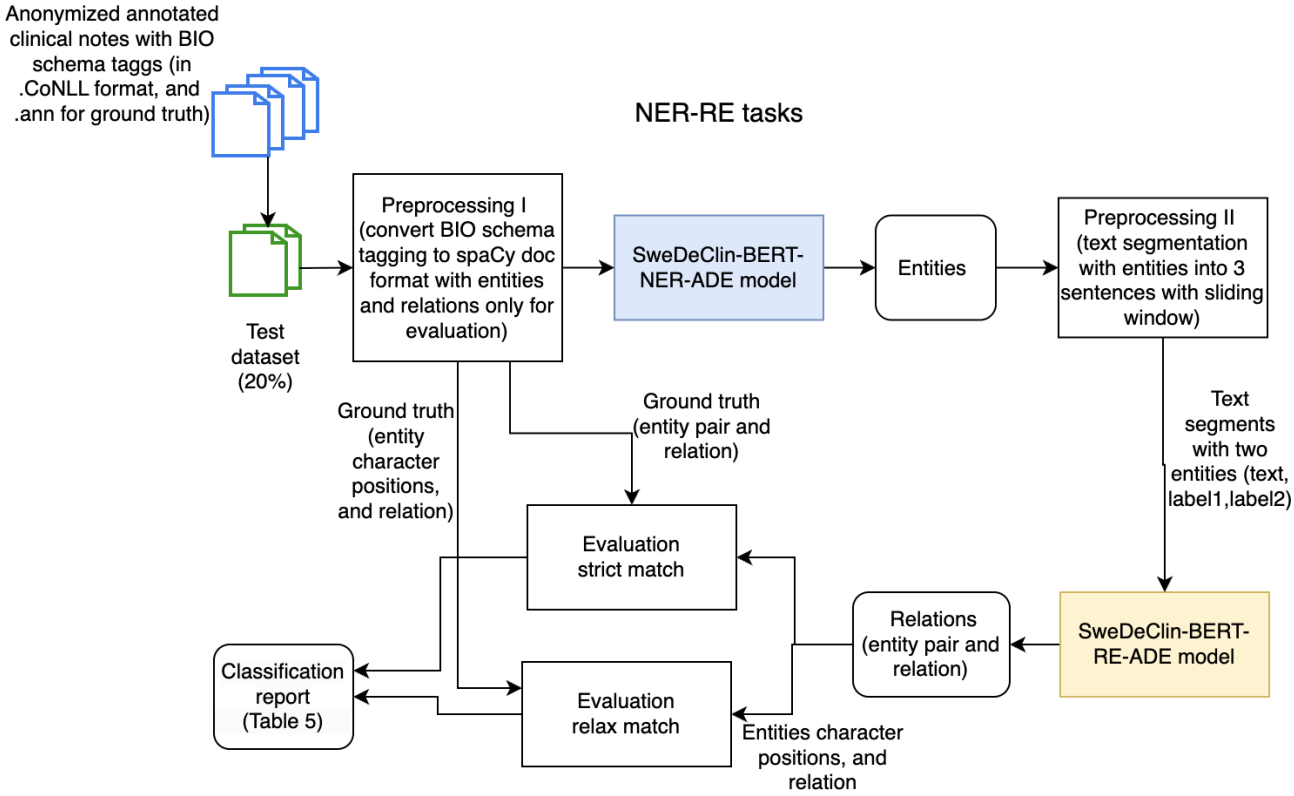

The SweDeClin-BERT-NER-ADE model first classified entities within the text, and these entities were then preprocessed to be used by the SweDeClin-BERT-RE-ADE model to establish relationships. SweDeClin-BERT-RE-ADE model accepts a text segment with only 2 entities. This sequential stacking of models introduces the challenge of error propagation. Specifically, any entity missed by the SweDeClin-BERT-NER-ADE model would not be considered for relation classification by the SweDeClin-BERT-RE-ADE model, leading to unidentified relations. In addition, misclassifications at the NER stage could result in the incorrect exclusion of potential relations due to predefined rules. For example, if an entity labeled as a Finding was misclassified as a Body Part, the system would ignore potential relationships involving this misclassified entity, as certain relations (such as between a Drug and a Body Part) were deemed invalid.

The proposed integrated NER-RE task was evaluated using 2 modes: strict and flexible matching. The strict match required an exact match for all entities, including multiword entities, while the flexible match allowed for partial matches, recognizing cases where at least one word from a multiword entity was correctly identified.

ADEs Detection in Clinical Notes Using Integrated NER-RE Pipeline

Given the overarching goal of identifying whether a text contained information about an ADE, we also evaluated the system's ability to classify texts as either containing ADE information or not (Figure 4).

XSL•FO
RenderX

Figure 4. Adverse drug event detection in clinical notes using the integrated pipeline of named entity recognition and relation extraction models. ADE: adverse drug event; NER-RE: named entity recognition-relation extraction; SweDeClin-BERT-RE-ADE: Swedish Deidentified Clinical Bidirectional Encoder Representations from Transformers-relation extraction-adverse drug event

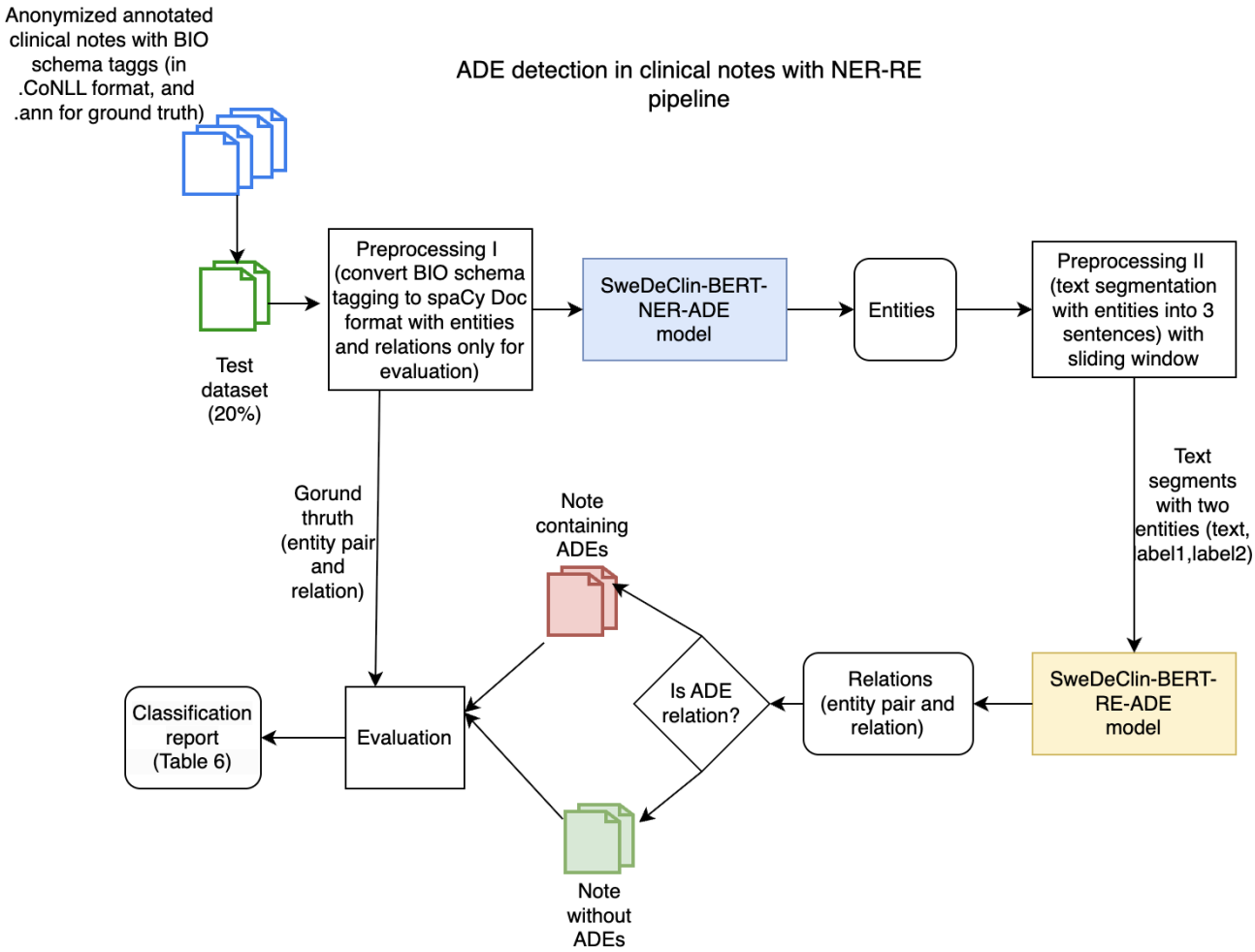

This binary classification was implemented as a rule-based decision based on the predictions from the integrated NER-RE pipeline. The rule-based decisions were made whether at least one ADE relation was present in a clinical note and then labeled as a note containing ADE. This evaluation metric helped in understanding the practical applicability of the model in finding clinical notes with potential ADEs.

Ethical Considerations

This research has been approved by the Regional Ethical Review Board (Etikprövningsnämnden), permission number 2012/834-31/5 and the Swedish Ethical Review Authority (Etikprövningsmyndigheten), permission number 2023-06920-01. The original consent approval covers secondary analysis without additional consent. All data used in this study are anonymized (do not contain any private information, like names, phone numbers, address, etc).

Results

NER Task Results

In evaluating the NER task on the holdout data (test dataset), our approach using SweDeClin-BERT demonstrated superior performance across most entity categories (see Table S2 in Multimedia Appendix 1) compared with the baseline model [24].

For the key entities related to ADE, including Finding, Disorder, and Drug, our approach outperformed the baseline. Specifically, for the entity category Finding, our model achieved an F_1-score of 0.802, compared to the baseline's 0.71. This improvement is due to balanced enhancements in both precision (0.798 vs 0.756) and recall (0.807 vs 0.67). In the Disorder category, our model recorded an F_1-score of 0.859, surpassing the baseline's 0.811, with superior precision (0.865 vs 0.83) and recall (0.853 vs 0.793). The Drug entity category also showed gains, with our model reaching an F_1-score of 0.946, compared to the baseline's 0.886.

RE Task Results

Evaluation results using fine-tuned SweDeClin-BERT for the RE task (SweDeClin- BERT-RE-ADE) on the holdout data revealed improvements across all relation categories (see) over the baseline results. Specifically, for the Adverse Drug Event category, our model attains an F_1-score of 0.704, compared to the baseline's 0.146.

Overall, the macro-average scores reflect improvements, with our model achieving an F_1-score of 0.813, in sharp contrast to the baseline's 0.281 F_1-score. These results strongly suggest

XSL•FO
RenderX

that transformer-based architectures outperform traditional machine learning approaches in relation extraction, likely due to their superior ability to capture contextual information.

NER-RE Integrated Task Results

The evaluation results for the integrated task (NER-RE) for Strict and Relaxed matches are presented in Table S3 in Multimedia Appendix 1. In this table, the N column represents the total number of holdout instances in each category in the test dataset, and the Classified column refers to the number of instances for which the model predicted a relation.

Overall, the NER-RE pipeline performance slightly decreased, as expected, because of error propagation from the NER model. However, looking at the macro average scores in Table S3 in Multimedia Appendix 1, the NER-RE pipeline performed well in both strict and relaxed match modes (0.815 and 0.812 F_1-score, respectively). The obtained results show that all relation categories obtained slightly higher performance in relaxed match mode, although the differences between the 2 modes in performance are next to negligible.

ADE Detection in Clinical Notes

The evaluation results of ADE detection in clinical notes are shown in Table S4 Multimedia Appendix 1. Specifically, in the relaxed matching mode, the model achieved an F_1-score of 0.828 for Containing ADE class, and 0.789 for the No ADE class.

Overall, the macro-average scores across both categories are moderate (0.771 in strict and 0.808 in relax matching). These results highlight the model's balanced performance in ADE detection, effectively identifying most ADE cases while maintaining moderate accuracy for non-ADE instances (F_1-score with 0.789).

Overall, only 62% (246 clinic notes) contain explicit description of the ADEs.

Error Analysis

NER Task Errors

The error analysis for NER model reveals several key insights into misclassification patterns. Specifically, the ADE Cue class emerges as the most frequently misclassified category (see Table S5 inMultimedia Appendix 1), with numerous instances being erroneously labeled as Other or Disorder. Notably, 13 (in comparison to 37 true positives) instances were classified as Other, indicating a lack of sufficient examples to accurately identify entities of this class in the corpus. In addition, 7 instances were misclassified as Disorder, with token examples such as "excess substitution," "allergy," "reaction," "toxic," "skin side effects," and "exanthem."

It is important to note that the terms "allergy," "reaction," and "exanthem" frequently appear within compound entities (consisting of 2 or more words) that are annotated under three different categories: ADE Cue, Disorder, and Finding. This observation suggests that the gold standard annotations for these classes may need refinement to improve entity recognition performance.

The second most misclassified category is Finding, with 132 instances (in comparison to 581 true positives) incorrectly labeled as Other. It's worth noting that Finding has the highest number of compound instances, accounting for 23% (718 out of 3142) of all Finding instances in the corpus. Among the instances misclassified as Other, we observed terms such as "weak," "general," "new symptoms," "problem," "histamine release," "healthy," "improvement," "positive," "worsened," "progressed," "prothrombin time," "good condition," "borrelia," and "skin flora." In this list, we see examples of findings without specificity, such as problem and new symptoms, which might appear in contexts other than medical findings. This could be attributed to the pretraining of SweDeClin-BERT on the general-purpose KB-BERT model. On the other hand, the list also includes instances that were simply missed by the model, such as critical findings like histamine release, borrelia, and prothrombin time.

For compound instances where only one part of the instance was tagged as Finding, we can mention low-grade nausea, with only nausea (illamående) classified as Finding; general malaise, with only malaise (illabefinnande) classified as Finding; and increased infection susceptibility, with only infection susceptibility (infektionsbenägenhet) labeled. This suggests that our model struggles with compound terms, particularly those with detailed descriptions. It often recognizes only the core component (the primary finding) of the entity while missing the additional descriptive elements that provide context. This limitation underscores the need for further refinement of the annotation guidelines or additional training data.

In the Disorder category, the majority of misclassifications occur as Finding (33 instances) and Other (18 instances). Misclassifications of Disorder entities as Other include examples like fungus (svamp) and malignancy (malignitet). In addition, there are compound cases where only one part of the entity is correctly labeled. For example, in partial cortisol deficiency (partiell kortisolbrist), the term cortisol deficiency is labeled as Disorder; in acute hives (akut urtikaria), hives are labeled as Disorder; and in secondary infection (sekundär infektion), infection is labeled as Disorder. There are also instances where Disorder entities are misclassified as ADE Cue, such as drug-induced rash (läkemedelsexanthem, läkemedelsutslag), a term that appears in both the Disorder and ADE Cue categories in the gold standard.

In the Drug category, some instances were misclassified as Findings or ADE Cues. For example, requiring oxygen (syrgaskrävande) was misclassified as a Finding, and Claforan-induced (Claforanorsakat) was misclassified as an ADE Cue. In these cases, the issue seems to be related to inconsistencies in the gold standard rather than the model's performance. In addition, some instances were misclassified as Other, such as medication (läkemedel, medicin), where this term appeared multiple times in the text, with the model only labeling one instance correctly.

Body Structure is the least misclassified category, with only 2 terms in compound instances misclassified as Findings and 6 instances misclassified as Other.

XSL•FO
RenderX

RE Task Errors

Examining the errors in the RE step, as detailed in Table S6 in Multimedia Appendix 1, we observe that the RE model primarily misclassifies relation class instances as No relation. Specifically, out of 199 classified Indication instances, 36 were misclassified as No relation. Similarly, out of 136 ADE instances, 46 were misclassified as No relation. In the ADE Outcome category, 3 out of 15 instances were misclassified as No relation, and out of 30 ADE Cause instances, 5 were misclassified as No relation.

While these errors are notable, they do not provide new insights due to the predefined rules governing permissible relations. These rules limit the possibility of misclassifying entities across the 4 primary relations of interest (Indication, ADE, ADE Outcome, and ADE Cause). However, there are other types of errors that are worth mentioning. For example, 2 out of 30 ADE Cause instances were misclassified as ADE.

This error appears to stem from mistakes at the NER stage, particularly in labeling ADE Cues as Finding or Disorder.

Looking into the specifics of the misclassifications, particularly for the ADE class, we find that very general reactions are often misclassified as ADEs, such as itching (klåda), swelling (svullnadskänsla), nausea (illamående), allergic reaction (allergisk reaktion), rash (utslag), anxiety (oro), and pain (värk). Some of these terms appear in the list of relations misclassified as Indication, such as pain (värk), swelling (svullnadskänsla), and allergic reaction (allergisk reaktion).

Among the specific medications in ADE relations misclassified as No relation instances, we found Digitalis, Enalapril, Seretide, Symbicort, antihistamine, asthma medication, botox, docetaxel, and epirubicin.

Overall, these errors suggest that important ADEs are being missed by the model, and increasing the training data could potentially improve classification.

Discussion

Principal Findings

This study investigated whether the LLMs approach, like the SweDeClin-BERT model, allows better detection of ADEs in clinical notes in comparison to the MLMs approach, like conditional random fields and random forest. We successfully fine-tuned the SweDeClin-BERT model for medical NER and RE tasks to identify ADEs in Swedish clinical notes. The integrated NER-RE pipeline achieved an overall F_1-score of 0.81, with the NER task reaching 0.845 and the RE task 0.813. These results significantly outperform our previous machine learning approach using CRFs and random forests [24], with a 5% improvement in NER and a 53% gain in RE macro-average F_1-score. These findings confirm that a transformer-based model provides a more effective approach for ADE identification in clinical text.

Only 62% of notes contained an explicit description of ADE, which indicates that the presence of an ADE-related ICD code does not guarantee that the clinical note contains detailed information about the event.

Despite the moderate performance (0.815 F_1-score) of our integrated NER-RE pipeline, the F_1-scores for individual entity classes in the NER task vary. For instance, the ADE Cue entity class has the lowest performance, with an F_1-score of 0.680. This is due to unbalanced entities in the NER task in the dataset. The unbalanced issue can be addressed by increasing the dataset with more ADE notes. Another possible explanation for the low accuracy in the ADE Cue entity in the NER task is that the annotators' agreement is low (0.53), indicating that there are some refinements in the data annotations that should be applied (better consistency).

Summarizing the results of the error analysis for the NER task, we identify several key issues: lack of sufficient examples to accurately identify entities of the ADE Cue class in the corpus; overlap of words between the three categories: ADE Cue, Disorder, and Finding; multistem words with stems belonging to different entities but tagged as one entity; general terms that sometimes belong to entities (eg, besvär, problem) and other times do not; inconsistent tagging of repeated mentions of the same entity (eg, läkemedel mentioned several times). Most of these issues are related to the issue with compound and multistem words, which are very common words in clinical text in the Swedish language. A possible suggestion for dealing with compound and multi-stream words could be to address this during the annotation process (annotate the subword of a word separately). This will allow the SweDeClin-BERT model to handle unseen or rare compound words by breaking them into known subwords, which helps in maintaining the semantic meaning.

We also observe differences in the performance of the RE model across various class categories. Despite having more examples for the classes of primary interest, Indication and ADE, compared to ADE Outcome and ADE Cause, the RE model performs better in classifying the latter, with F_1-scores of 0.810 and 0.816 for ADE Outcome and ADE Cause, compared to 0.757 and 0.704 for Indication and ADE. This higher performance can be attributed to the presence of ADE Cue words in relations ADE Outcome and ADE Cause. As reported in [24], ADE Cues consist of a relatively short set of words (130 terms) compared to the more varied Disorders and Findings, which have 976 and 1533 unique words, respectively. This suggests that there may still be insufficient training data to cover all the diverse contexts of the ADE relation.

Comparison With Previous Work

In comparison, previous studies on ADE detection in clinical notes using LLMs, although based on clinical texts in English, achieved higher accuracy for the NER and RE tasks (with the microaverage F_1-scores at 0.93 and 0.96, respectively [27]). However, the performance of the integrated NER-RE pipeline reported in previous studies was comparable to that of our study. For instance, the previous NER-RE pipeline achieved a micro average F_1-score of 0.895 in [27], and micro F_1-scores of 0.70 for strict matching and 0.81 for relaxed matching in [29]. In comparison, our study reports macro average F_1-scores of 0.815 for strict matching and 0.812 for relaxed matching. We use a

XSL•FO
RenderX

macro average statistic to account for the presence of the dominant No relation class in our data.

Limitations

This study has several limitations that impact the generalizability of the results. First, the dataset was annotated by a single physician and 2 nonclinicians, with low interannotator agreement observed across key entity types. This raises concerns about the consistency and reliability of the gold standard annotations, which in turn may have limited the achievable model performance. Disagreements among annotators, particularly between clinical and nonclinical raters, suggest that the annotation guidelines may not have been sufficiently robust for nonexpert interpretation.

Second, the dataset's relatively small size (395 clinical notes) and the unbalanced distribution of ADE classes constrain the model's ability to generalize across broader clinical settings. It may introduce selection bias, as the dataset overrepresents ADEs compared to their real-world prevalence. Annotation bias may arise from subjective interpretation of guidelines, and temporal bias may limit the generalizability of the model to other time periods. Although this dataset was created over a decade ago with some known limitations in annotation quality, it provides a valuable opportunity for longitudinal comparison. Previous work relied on traditional machine learning methods such as CRF, which predated the development of transformer-based models like BERT. Reapplying modern deep learning techniques to the same clinical text enables us to assess how methodological advances can improve ADE extraction, even when working with imperfect data.

Third, the rule-based approach used in our current study is limited in its ability to capture deeper contextual nuances such as temporality, negation, speculation as in our previous study [24], and attribution. These factors are crucial for accurately detecting ADEs in clinical notes, as ADE mentions may refer to past or future events, be negated or speculative, or refer to individuals other than the patient (eg, a family member mentioned in the patient's medical history). Furthermore, this study did not investigate variations in language use, documentation styles across different clinics, or shifts in ADE prevalence. Our current approach does not capture this level of contextual information and leaves it as an important direction for future research.

Summary

To summarize, the results demonstrate that fine-tuned SweDeClin-BERT models for NER and RE are effective for detecting ADEs in Swedish clinical notes when used in an integrated pipeline. These findings suggest that transformer-based models can serve as valuable tools for augmenting pharmacovigilance efforts by uncovering ADEs not captured by structured EHR data alone. However, improved handling of compound words and refinements to entity annotations remain important to increase accuracy and reduce ambiguity. In particular, merging "Disorder" and "Finding" into a single "Finding" class and excluding the "BodyPart" class are expected to simplify the schema and enhance annotation consistency, as already evidenced in our ongoing work.

Conclusions

Using a domain-specific language model like SweDeClin-BERT for detecting ADEs in Swedish clinical notes enables moderately accurate identification, achieving an F_1-score of 81%. This approach also improves relation extraction accuracy by 53% compared to our previous random forest method. Key challenges in classifying ADE Cue entities have been identified and will be addressed in our upcoming study. Detecting ADEs in clinical notes is indeed a challenging task, and more research is needed to be able to use the valuable information documented in free text instead of being documented in a structured manner. The integrated NER-RE pipeline, powered by fine-tuned SweDeClin-BERT models, offers the potential to identify suspected ADEs in clinical notes that are not explicitly reported in EHRs.

Acknowledgments

ChatGPT (an AI language model developed by OpenAI) was used to assist with grammar correction, sentence rephrasing, and improving the clarity and coherence of the manuscript text. All scientific content, interpretations, and conclusions were developed by the authors. The authors would like to acknowledge the DISA (Linnaeus University Center for Data Intensive Sciences and Applications) for funding this work (Seed project: Using Natural Language Models for Extracting Drug-Related Problems (NLMED)).

Data Availability

The datasets analyzed during this study are not publicly available due to Ethical considerations and privacy agreements stipulated by the ethical approval granted by the Swedish Ethical Review Authority.

Authors' Contributions

EK performed data curation, formal analysis, investigation, and software. AH and HD performed data curation (Stockholm EPR ADE Corpus) and provided the resources such as the SweDeClin-BERT model pre-trained on Swedish clinical notes. Furthermore, AH and HD contributed with supervision. TH and AL contributed with concpetualization, methodology, and supervision. All authors contributed with original draft and reviwing and editing the final version of the manuscript.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Performance metrics (precision, recall, and F1) for SweDeClin-BERT-ADE compared with baseline CRF
[MHT File, 123 KB - formative_v9i1e71949_app1.mht]

References

1. Durand M, Castelli C, Roux-Marson C, Kinowski JM, Leguelinel-Blache G. Evaluating the costs of adverse drug events in hospitalized patients: a systematic review. Health Econ Rev 2024 Feb 8;14(1):11. [doi: 10.1186/s13561-024-00481-y] [Medline: 38329561]
2. Silva LT, Modesto ACF, Amaral RG, Lopes FM. Hospitalizations and deaths related to adverse drug events worldwide: Systematic review of studies with national coverage. Eur J Clin Pharmacol 2022 Mar;78(3):435-466. [doi: 10.1007/s00228-021-03238-2] [Medline: 34716774]
3. Laatikainen O, Miettunen J, Sneck S, Lehtiniemi H, Tenhunen O, Turpeinen M. The prevalence of medication-related adverse events in inpatients-a systematic review and meta-analysis. Eur J Clin Pharmacol 2017 Dec;73(12):1539-1549. [doi: 10.1007/s00228-017-2330-3] [Medline: 28871436]
4. El Morabet N, Uitvlugt EB, van den Bemt BJF, van den Bemt P, Janssen MJA, Karapinar-Çarkit F. Prevalence and preventability of drug-related hospital readmissions: a systematic review. J Am Geriatr Soc 2018 Mar;66(3):602-608. [doi: 10.1111/jgs.15244] [Medline: 29468640]
5. Hakkarainen KM, Hedna K, Petzold M, Hägg S. Percentage of patients with preventable adverse drug reactions and preventability of adverse drug reactions--a meta-analysis. PLoS ONE 2012;7(3):e33236. [doi: 10.1371/journal.pone.0033236] [Medline: 22438900]
6. Yasrebi-de Kom IAR, Dongelmans DA, de Keizer NF, et al. Electronic health record-based prediction models for in-hospital adverse drug event diagnosis or prognosis: a systematic review. J Am Med Inform Assoc 2023 Apr 19;30(5):978-988. [doi: 10.1093/jamia/ocad014] [Medline: 36805926]
7. Janusmed Riskprofile. Region Stockholm. URL: https://janusmed.se/riskprofil [accessed 2024-08-10]
8. Eiermann B, Rodriguez D, Cohen P, Gustafsson LL. ADR databases for on-site clinical use: Potentials of summary of products characteristics. Basic Clin Pharmacol Toxicol 2021 Apr;128(4):557-567. [doi: 10.1111/bcpt.13564] [Medline: 33523597]
9. Li Y, Tao W, Li Z, et al. Artificial intelligence-powered pharmacovigilance: A review of machine and deep learning in clinical text-based adverse drug event detection for benchmark datasets. J Biomed Inform 2024 Apr;152:104621. [doi: 10.1016/j.jbi.2024.104621] [Medline: 38447600]
10. Kompa B, Hakim JB, Palepu A, et al. Artificial intelligence based on machine learning in pharmacovigilance: a scoping review. Drug Saf 2022 May;45(5):477-491. [doi: 10.1007/s40264-022-01176-1] [Medline: 35579812]
11. Eriksson R, Jensen PB, Frankild S, Jensen LJ, Brunak S. Dictionary construction and identification of possible adverse drug events in Danish clinical narrative text. J Am Med Inform Assoc 2013;20(5):947-953. [doi: 10.1136/amiajnl-2013-001708] [Medline: 23703825]
12. Wang X, Hripcsak G, Markatou M, Friedman C. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. J Am Med Inform Assoc 2009;16(3):328-337. [doi: 10.1197/jamia.M3028] [Medline: 19261932]
13. Yang X, Bian J, Fang R, Bjarnadottir RI, Hogan WR, Wu Y. Identifying relations of medications with adverse drug events using recurrent convolutional neural networks and gradient boosting. J Am Med Inform Assoc 2020 Jan 1;27(1):65-72. [doi: 10.1093/jamia/ocz144] [Medline: 31504605]
14. Li F, Liu W, Yu H. Extraction of information related to adverse drug events from electronic health record notes: design of an end-to-end model based on deep learning. JMIR Med Inform 2018 Nov 26;6(4):e12159. [doi: 10.2196/12159] [Medline: 30478023]
15. Wunnava S, Qin X, Kakar T, Sen C, Rundensteiner EA, Kong X. Adverse drug event detection from electronic health records using hierarchical recurrent neural networks with dual-level embedding. Drug Saf 2019 Jan;42(1):113-122. [doi: 10.1007/s40264-018-0765-9] [Medline: 30649736]
16. Munkhdalai T, Liu F, Yu H. Clinical relation extraction toward drug safety surveillance using electronic health record narratives: classical learning versus deep learning. JMIR Public Health Surveill 2018 Apr 25;4(2):e29. [doi: 10.2196/publichealth.9361] [Medline: 29695376]
17. Fan Y, Zhou S, Li Y, Zhang R. Deep learning approaches for extracting adverse events and indications of dietary supplements from clinical text. J Am Med Inform Assoc 2021 Mar 1;28(3):569-577. [doi: 10.1093/jamia/ocaa218] [Medline: 33150942]
18. El-Allaly ED, Sarrouti M, En-Nahnahi N, Ouatik El Alaoui S. An attentive joint model with transformer-based weighted graph convolutional network for extracting adverse drug event relation. J Biomed Inform 2022 Jan;125:103968. [doi: 10.1016/j.jbi.2021.103968] [Medline: 34871807]

XSL•FO
RenderX

19. Mahendran D, McInnes BT. Extracting adverse drug events from clinical notes. AMIA Jt Summits Transl Sci Proc 2021;2021(420):420-429. [Medline: 34457157]
20. Laparra E, Mascio A, Velupillai S, Miller T. A review of recent work in transfer learning and domain adaptation for natural language processing of electronic health records. Yearb Med Inform 2021 Aug;30(1):239-244. [doi: 10.1055/s-0041-1726522] [Medline: 34479396]
21. Wu H, Ji J, Tian H, et al. Chinese-named entity recognition from adverse drug event records: radical embedding-combined dynamic embedding-based BERT in a bidirectional long short-term conditional random field (Bi-LSTM-CRF) model. JMIR Med Inform 2021 Dec 1;9(12):e26407. [doi: 10.2196/26407] [Medline: 34855616]
22. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. Presented at: Advances in Neural Information Processing Systems; Dec 6-12, 2020 p. 1877-1901. [doi: 10.48550/arXiv.2005.14165]
23. Modi S, Kasmiran KA, Mohd Sharef N, Sharum MY. Extracting adverse drug events from clinical notes: a systematic review of approaches used. J Biomed Inform 2024 Mar;151:104603. [doi: 10.1016/j.jbi.2024.104603] [Medline: 38331081]
24. Henriksson A, Kvist M, Dalianis H, Duneld M. Identifying adverse drug event information in clinical notes with distributional semantic representations of context. J Biomed Inform 2015 Oct;57:333-349. [doi: 10.1016/j.jbi.2015.08.013] [Medline: 26291578]
25. Yang X, Bian J, Gong Y, Hogan WR, Wu Y. MADEx: A system for detecting medications, adverse drug events, and their relations from clinical notes. Drug Saf 2019 Jan;42(1):123-133. [doi: 10.1007/s40264-018-0761-0] [Medline: 30600484]
26. Kim Y, Meystre SM. Ensemble method-based extraction of medication and related information from clinical texts. J Am Med Inform Assoc 2020 Jan 1;27(1):31-38. [doi: 10.1093/jamia/ocz100] [Medline: 31282932]
27. Wei Q, Ji Z, Li Z, et al. A study of deep learning approaches for medication and adverse drug event extraction from clinical text. J Am Med Inform Assoc 2020 Jan 1;27(1):13-21. [doi: 10.1093/jamia/ocz063]
28. Bousselham H, Nfaoui EH, Mourhir A. Fine-tuning GPT on biomedical NLP tasks: an empirical evaluation. Presented at: 2024 International Conference on Computer, Electrical & Communication Engineering (ICCECE); Feb 2-3, 2024; Kolkata, India p. 1-6. [doi: 10.1109/ICCECE58645.2024.10497313]
29. Li Y, Li J, He J, Tao C. AE-GPT: Using large language models to extract adverse events from surveillance reports-a use case with influenza vaccine adverse events. PLoS ONE 2024 Mar 21;19(3):e0300919. [doi: 10.1371/journal.pone.0300919]
30. Dürlich L, Bergman E, Larsson M, et al. Explainability for NLP in pharmacovigilance: a study on adverse event report triage in swedish. Presented at: Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health); May 4, 2025; Albuquerque, New Mexico p. 46-68. [doi: 10.18653/v1/2025.cl4health-1.5]
31. Vakili T, Lamproudis A, Henriksson A, Dalianis H. Downstream task performance of BERT models pre-trained using automatically de-identified clinical data. Presented at: Proceedings of the Thirteenth Language Resources and Evaluation Conference; Jun 20-25, 2022 p. 4245-4252. [doi: 10.5281/zenodo.6392366]
32. Dalianis H, Henriksson A, Kvist M, Velupillai S, Weegar R. HEALTH BANK-a workbench for data science applications in healthcare. CAiSE Industry Track 2015 Jun 11;1381:1-8 [FREE Full text]
33. Rosvall L, Paasonen N. Data augmentation: enhancing named entity recognition performance on swedish medical texts. : Gothenburg university library; 2023 URL: https://hdl.handle.net/2077/78677 [accessed 2025-08-10]
34. Lamproudis A, Svenning TO, Torsvik T, et al. Using a large open clinical corpus for improved ICD-10 diagnosis coding. AMIA Annu Symp Proc 2023;2023:465-473. [doi: 10.57498/AMIA.AP.2023.465] [Medline: 38222373]
35. Bergman E, Dürlich L, Arthurson V, et al. BERT based natural language processing for triage of adverse drug reaction reports shows close to human-level performance. PLOS Digit Health 2023 Dec;2(12):e0000409. [doi: 10.1371/journal.pdig.0000409] [Medline: 38055685]
36. Klement W, El Emam K. Consolidated reporting guidelines for prognostic and diagnostic machine learning modeling studies: gevelopment and validation. J Med Internet Res 2023 Aug 31;25:e48763. [doi: 10.2196/48763] [Medline: 37651179]
37. Ruder S, Peters ME, Swayamdipta S, Wolf T. Transfer learning in natural language processing. Presented at: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Tutorials; Jun 2, 2019; Minneapolis, Minnesota p. 15-18. [doi: 10.18653/v1/N19-5004]
38. Snoek J, Larochelle H, Adams RP. Practical bayesian optimization of machine learning algorithms. In: Proceedings of the 26th International Conference on Neural Information Processing Systems: Curran Associates Inc; 2012. [doi: 10.5555/2999325.2999464]
39. Bishop CM, Nasrabadi NM. Pattern Recognition and Machine Learning: Springer; 2006:978.
40. SpaCy. Industrial-strength Natural Language Processing in Python. 2020. URL: https://spacy.io/ [accessed 2025-09-03]
41. International Drug Monitoring. World Health Organisation Technical Report Series. 1969. URL: https://iris.who.int/bitstream/handle/10665/40747/WHO_TRS_425.pdf?sequence=1 [accessed 2024-08-10]
42. Eisenstein J. Introduction to Natural Language Processing: The MIT Press; 2019.
43. Stenetorp P, Pyysalo S, Topić G, Ohta T, Ananiadou S, Tsujii JI. BRAT: a web-based tool for NLP-assisted text annotation. Presented at: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics; Apr 23-27, 2012; Avignon, France p. 102-107. [doi: 10.3115/v1/E12-2017]

XSL•FO
RenderX

Abbreviations

ADE: adverse drug event
AI: artificial intelligence
BERT: Bidirectional Encoder Representations from Transformers
BIO: Begin Inside Outside
CRF: Conditional Random Fields
DSM: Distributional Semantic Models
EHR: electronic health record
GPT: Generative Pre-trained Transformer
ICD-10: International Classification of Diseases
LLM: large language models
NER: named-entity recognition
NLP: natural language processing
RE: relation extraction
RF: random forest
SweDeClin-BERT: Swedish Deidentified Clinical Bidirectional Encoder Representations from Transformers
WHO: World Health Organization

Edited by A Mavragani; submitted 30.01.25; peer-reviewed by J Jones, S Matsuda, SS Madugula; revised version received 22.07.25; accepted 23.07.25; published 11.09.25.

Please cite as:
Kopacheva E, Henriksson A, Dalianis H, Hammar T, Lincke A
Identifying Adverse Drug Events in Clinical Text Using Fine-Tuned Clinical Language Models: Machine Learning Study
JMIR Form Res 2025;9:e71949
URL: https://formative.jmir.org/2025/1/e71949
doi:10.2196/71949

© Elizaveta Kopacheva, Aron Henriksson, Hercules Dalianis, Tora Hammar, Alisa Lincke. Originally published in JMIR Formative Research (https://formative.jmir.org), 11.9.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Enhancing Clinical Data Management Through Barcode Integration and Research Electronic Data Capture: Scalable and Adaptable Implementation Study

Rendong Zhang[1], MS; Sophie Chiron[2], BS; Regina Tyree[3], MS; Kate Carson[3], MPH; Larry Raber[3], BS; Karthik Ramadass[1], MS; Chenyu Gao[1], MS; Michael E Kim[4], BS; Lianrui Zuo[1], PhD; Yike Li[5], MD, PhD; Zhiyu Wan[6,7], PhD; Paul A Harris[7,8,9], PhD; Qi Liu[9,10], PhD; Ken S Lau[10,11,12], PhD; Lori A Coburn[3,13,14,15], MD; Keith T Wilson[3,13,14,15,16], MD; Yuankai Huo[1,4], PhD; Bennett A Landman[1,4], PhD; Shunxing Bao[1], PhD

[1]Department of Electrical and Computer Engineering, School of Engineering, Vanderbilt University, PMB 351824, 2301 Vanderbilt Place, Nashville, TN, United States

[10]Center for Quantitative Sciences, Vanderbilt University Medical Center, Nashville, TN, United States

[11]Epithelial Biology Center, Vanderbilt University Medical Center, Nashville, TN, United States

[12]Department of Cell and Developmental Biology, School of Medicine, Vanderbilt University, Nashville, TN, United States

[13]Vanderbilt Center for Mucosal Inflammation and Cancer, Nashville, TN, United States

[14]Program of Cancer Biology, School of Medicine, Vanderbilt University, Nashville, TN, United States

[15]VA Tennessee Valley Healthcare System, Nashville, TN, United States

[16]Department of Pathology, Microbiology, and Immunology,, Vanderbilt University Medical Center, Nashville, TN, United States

[2]Vanderbilt-Ingram Cancer Center, Vanderbilt University Medical Center, Nashville, TN, United States

[3]Department of Medicine, Division of Gastroenterology, Hepatology, and Nutrition, Vanderbilt University Medical Center, Nashville, TN, United States

[4]Department of Computer Science, School of Engineering, Vanderbilt University, Nashville, TN, United States

[5]Department of Otolaryngology-Head and Neck Surgery, Vanderbilt University Medical Center, Nashville, TN, United States

[6]School of Biomedical Engineering, ShanghaiTech University, Shanghai, China

[7]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, United States

[8]Department of Biomedical Engineering, Vanderbilt University, Nashville, TN, United States

[9]Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN, United States

Corresponding Author:
Shunxing Bao, PhD
Department of Electrical and Computer Engineering, School of Engineering, Vanderbilt University, PMB 351824, 2301 Vanderbilt Place, Nashville, TN, United States

Abstract

Background: Effective data management is crucial in clinical studies for precise tracking, secure storage, and reliable analysis of samples. Traditional systems often encounter challenges like barcode recognition errors, inadequate data details, and diminished performance under heavy workloads.

Objective: This paper aims to enhance clinical data management by improving barcode robustness, increasing data granularity, and boosting system throughput. These improvements address key challenges in barcode informatics systems, as highlighted in previous studies, to better support real clinical applications. In addition, we aim to validate the design criteria on various gastrointestinal-related studies, ensuring it can be easily integrated into other clinical data management workflows.

Methods: We evaluated the robustness of various barcode technologies under significant blurring conditions, implemented a dynamic organ-specific archive in the REDCap (Research Electronic Data Capture) database for various clinical study data collection criteria, and used Docker to containerize the informatics software for different studies. In addition, we proposed a local cache system to reduce interaction times with REDCap for large-scale data records. Experimental setups include assessing barcode recognition accuracy under various levels of image blurring, showcasing different study types managed with the organ-specific archive, and measuring system throughput and response times with and without the proposed local cache system.

Results: Our findings demonstrate that the DataMatrix barcode exhibits superior resilience, maintaining high recognition accuracy under blurred conditions. The dynamic organ-specific archive in REDCap enabled precise tracking of sample origins, improving data granularity. Docker containerization streamlines software deployment and ensures consistency across studies. The local cache system significantly reduces interaction times with REDCap, decreasing operating time by nearly eightfold compared to the naïve strategy when handling large patient datasets.

XSL•FO
RenderX

Conclusions: The proposed enhancements significantly improve barcode robustness, data granularity, and system throughput in the informatics system, addressing key limitations identified in previous studies. These optimizations ensure efficient data management and robust support for diverse clinical research needs.

(JMIR Form Res 2025;9:e70016) doi:10.2196/70016

KEYWORDS

clinical data management; barcode technology; REDCap; data granularity; Docker containerization; system throughput optimization

Introduction

In the digital age, operations across various fields are transforming, and digital pathology is revolutionizing traditional histopathological diagnosis by replacing microscopes and glass slides with virtual microscopy on computers [1,2]. Digital pathology has evolved into an exceptionally powerful and precise tool for the diagnosis and prognosis of a wide range of diseases [3-5]. However, hardware and software issues can greatly affect costs and slow down the implementation of clinical data management workflows [6]. Efficiently collecting and using data can be a significant challenge in clinical research due to the complexity and diversity of data types, the need for precise tracking and storage, and the potential for errors and inefficiencies in manual data handling. Therefore, an advanced informatics system for data management is essential and in high demand for ensuring precise and efficient usage of clinical research data [1,7-9]. Integrating tools like barcoding, specimen tracking, and digital dictation enhances the safety, quality, and efficiency of clinical research workflows [10]. Despite the clear potential improvements with a fully digital approach, only a small number of clinical research laboratories have transitioned to such workflows [11].

Clinical studies usually have complex and diverse data management needs that require tailored solutions depending on the nature and objectives of each study. For instance, some studies may need to collect samples from different organs or regions of the body to analyze specific diseases or conditions. Nourani et al [12] assessed various clinical data management systems by analyzing databases such as Web of Science, Scopus, Science Direct, ProQuest, Ovid Medline, and PubMed. Their study concluded that none of the systems comprehensively addressed all data management aspects in clinical trials. Therefore, it is often necessary to create customized data management systems tailored to specific needs. Without robust data management systems, researchers face challenges like data loss, sample tracking errors, and inefficiencies in data retrieval and analysis that compromise study quality and reliability and pose health risks or legal challenges for clinicians [9].

One widely used tool for managing clinical research data is REDCap (Research Electronic Data Capture), a web-based application developed by Vanderbilt University Medical Center [13,14]. REDCap offers a user-friendly interface for validated data entry, audit trails for tracking, and automated data exports for statistical analysis, making it a popular choice among academic, nonprofit, and government institutions worldwide [15-17,18,19]. Building on the capabilities of REDCap, Bao et al [20] designed a cross-platform informatics system with two key features: (1) a digitalized and human-readable barcode for sample tracking using a cost-effective printer, and (2) longitudinal archive storage in REDCap. After identifying several real-life bottlenecks, which are introduced below, we extended this system to more clinical studies over the past 5 years, including Crohn disease Gut Cell Atlas, Translational Impact of Electrophile Adducts in Colorectal Cancer. The project includes normal control, ulcerative colitis, Crohn disease, and colorectal cancer. It also includes surgical resections from the Cooperative Human Tissue Network in addition to colonoscopy biopsies. *Helicobacter pylori* (*H. pylori*)-associated gastric inflammation and disease progression, and gut–microbiome-related gastrointestinal and liver diseases, demonstrating significant improvements in data handling efficiency.

The challenges identified from real-life usage were: (1) barcode robustness, (2) detailed organ-specific data storage, and (3) system performance under increased data volume that could be further improved from the published system as shown in Figure 1. To elaborate on these challenges, first, barcodes are printed at a small size of 0.92″×0.54″ using a Dymo printer with 300 dpi resolution, often becoming blurred due to accidental touch, leading to scanner recognition errors and inefficiencies in sample tracking. A more robust barcode technology is needed for better recognition by current equipment even when blurred. Second, the need to record precise sample collection locations has increased as the studies now involve collecting samples from multiple specific regions, which the previous REDCap table design does not support. Enhancing the database structure to include organ-specific fields and a flexible data entry system would improve utility in diverse studies [20]. Third, system performance degrades significantly when the REDCap record count exceeds 100, affecting data entry, query, and retrieval efficiency. Implementing database optimization techniques, such as reducing interaction times with REDCap or refining query algorithms, could maintain optimal performance as the database grows.

XSL•FO
RenderX

Figure 1. Comparison of the published and proposed barcode systems for collected biopsies. The published system uses 1-D barcodes, supports only one use case (ascending colon [AC] and terminal ileum [TI]), and involves a heavy data entry load with REDCap without robustness checking. In contrast, the proposed system uses 2-D barcodes with robustness validation, supports multiple use cases, including dynamic organ-specific archives, and optimizes data entry with a lighter load on REDCap. For example, Study 1 covered that the published system collects samples only from the AC and TI. In contrast, the newly involved Study 2 involves several cases where each case requires sample collection from a specific region, such as the AC, TI, transverse colon (TC), descending colon (DC), sigmoid colon (SC), or rectum. Study 3 necessitates collecting samples from multiple regions. AC: ascending colon; DC: descending colon; Num.: number; REDCap: Research Electronic Data Capture; TC: transverse colon; TI: terminal ileum; SC: sigmoid colon.

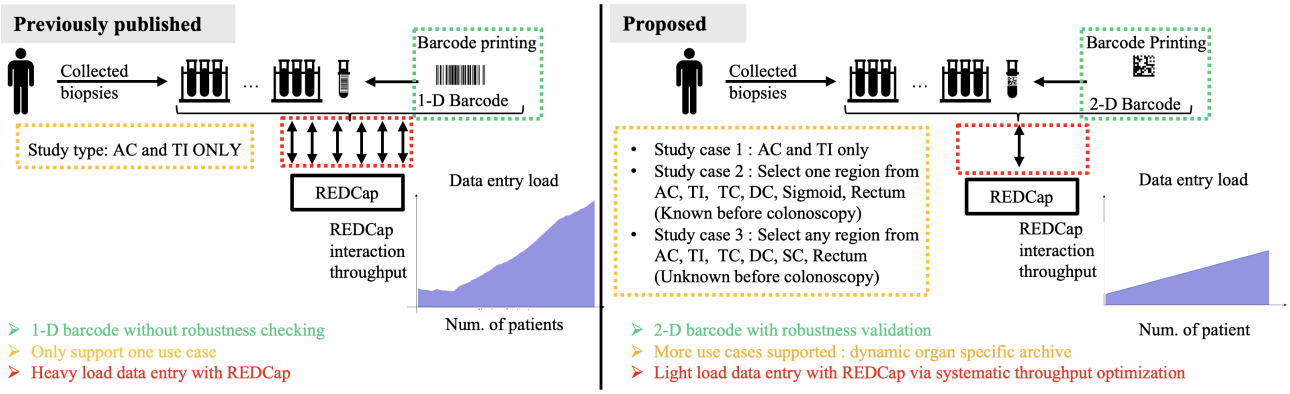

To address the identified issues, we propose three main design criteria for enhancing our system: (1) improving barcode robustness by updating the system with the most durable barcode technology available, (2) incorporating a dynamic organ-specific archive in the REDCap table design to support detailed anatomical data, and (3) optimizing system throughput to efficiently manage increased data volumes. Our methods include evaluating various barcode technologies under blurring conditions, implementing a dynamic organ-specific archive in REDCap, and using Docker for containerizing software clusters. In addition, we developed a local cache system to reduce interaction times with REDCap. Experimental setups involve testing barcode recognition accuracy and comparing system performance across different implementations. The results demonstrate significant improvements in barcode recognition, data granularity, and system throughput, thereby enhancing overall data management efficiency in clinical studies.

Methods

Overview

In this section, we detail the development and evaluation processes designed to meet the 3 proposed design criteria: improving barcode robustness, incorporating a dynamic organ-specific archive, and optimizing system throughput. Each methodology and experimental setup ensures the solutions are effective and practical for real-life clinical settings.

Data Management Workflows Overview

The data management workflow for 2 Python-based applications from [20], the Printer app and the Data entry app, which are integral components of our biopsy management with REDCap, is presented in Figure 2(A-B).

XSL•FO
RenderX

Figure 2. Illustration of workflows for (**A**) the Printer app and (**B**) Data entry app used in managing biopsy samples. In the Printer app, the workflow commences with entering the patient ID, followed by autofilling patient groups, selecting the clinician ID, choosing a print option, printing the barcode, and syncing print events to REDCap. The data entry app workflow involves scanning the barcode, selecting the clinician ID, determining if samples were collected, entering sample location information, selecting organ information, and either syncing sample archive information to REDCap or marking the barcode as not collected. The green, yellow, and red dotted boxes indicate the locations where systematic bottlenecks were identified in the previous implementation. These bottlenecks have been addressed and resolved in this study. REDCap: Research Electronic Data Capture; VUMC: Vanderbilt University Medical Center;

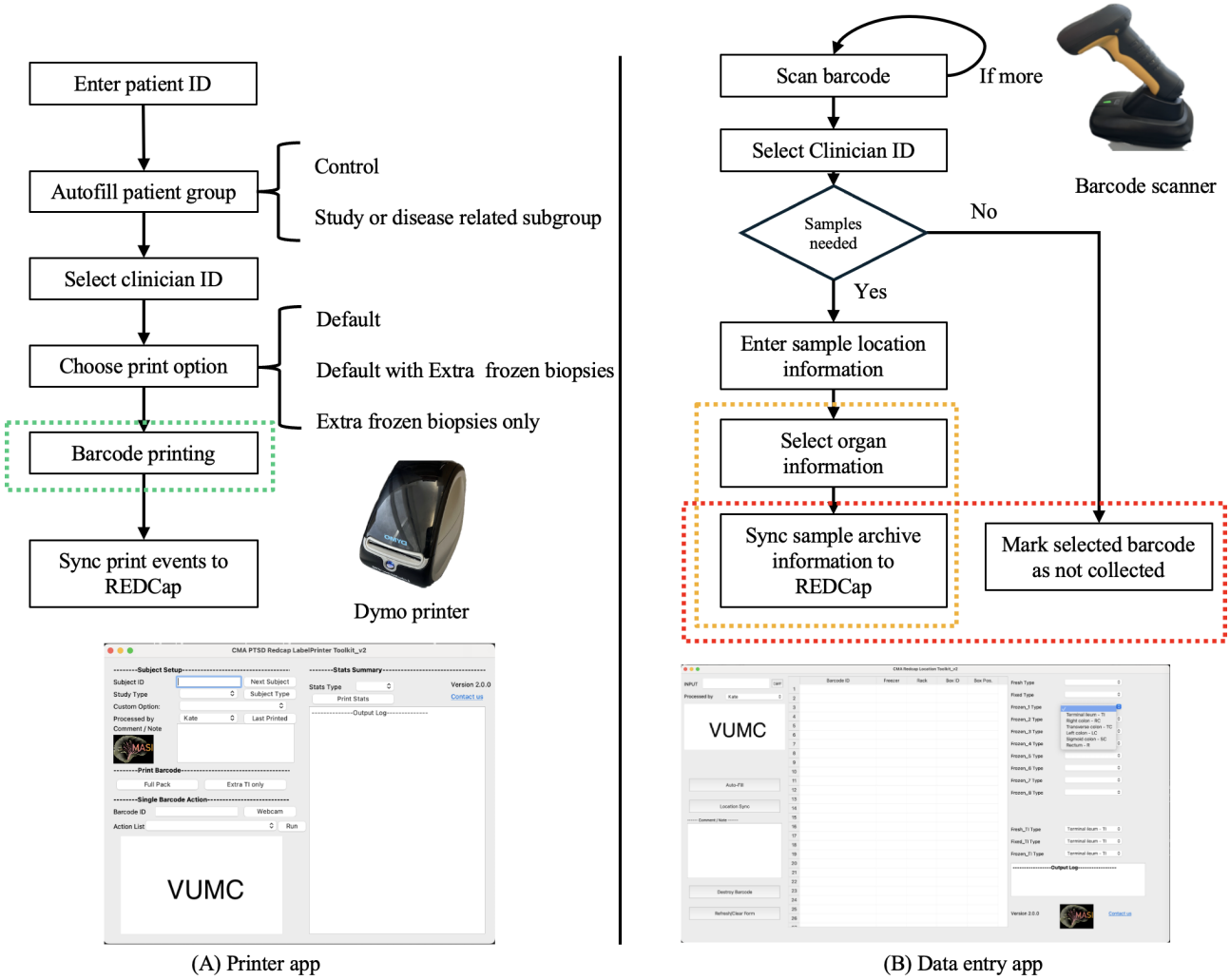

(A) Printer app

(B) Data entry app

Briefly, the Printer app workflow begins by entering the patient ID, which autofills the patient group based on control or study or disease-related subgroups. Next, the clinician ID is selected, and a print option is chosen, ranging from default to including extra frozen biopsies. The "Barcode Printing" step, highlighted in a green box in Figure 2, was identified as a bottleneck due to limitations in printer quality and previously selected barcode design. Finally, print events are synced to REDCap, ensuring all printed barcodes are recorded in the system.

The data entry app workflow starts by scanning the barcode and selecting the clinician ID. If samples are collected, the user enters the sample location information and selects the organ information, both highlighted in yellow boxes in Figure 2. The previous system could only handle one static organ selection, which was a limitation given the diverse clinical studies we manage. If samples are not collected, the barcode is marked as not collected, and the relevant REDCap field is removed. Syncing sample archive information to REDCap, highlighted in a red box in Figure 2, is where the interaction bottleneck was found, which is critical for optimizing system throughput.

Next, we introduce design criteria based on the 3 identified bottlenecks in detail.

Design Criteria 1. Evaluation-Driven Barcode Robustness Improvement

To address the issue of barcode robustness, a series of validations (Figure 3) was conducted to assess various barcode technologies and printing methods for decoding 8-digit numerical data [20] (necessitated by the limitations of printer quality and the tiny barcode size). The experiment included evaluating the following barcode types discussed in the literature [21,22]: (1) Code 39; (2) EAN-8; (3) ITF (Interleaved 2 of 5); (4) QR Code; and (5) DataMatrix. They were chosen for their simplicity, ease of printing, and reliability in a small size.

XSL•FO
RenderX

Figure 3. The barcode robustness evaluation process involves creating a canvas, attaching barcodes, applying a Gaussian blur with varying radii (0, 0.5, 1, 1.5, 2 pixels), scanning the barcodes, and validating their robustness.

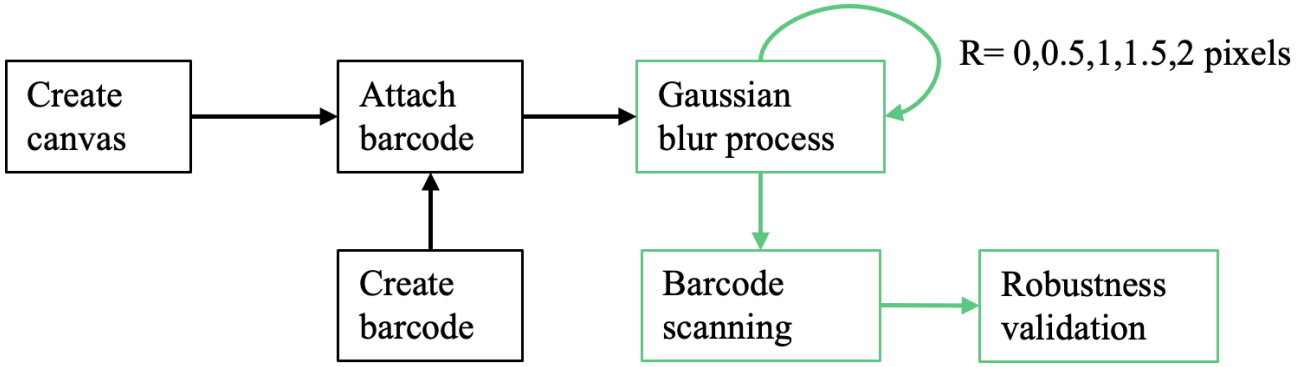

One hundred unique 8-digit numbers were randomly generated for each of the 5 barcode types, and 5 rounds of testing were conducted. Each round applied a Gaussian blur with radii of 0, 0.5, 1, 1.5, and 2 pixels, corresponding to 0, 0.00167, 0.00333, 0.005, and 0.00667 inches, respectively. Although Gaussian blur was used as a proxy for image degradation, it does not fully capture the characteristics of printer smearing, which typically occurs in a single direction and is better modeled as motion blur. In practice, motion blur results in a point spread function resembling a rectangular function in image space and a sinc function in Fourier space. Despite this simplification, the barcodes were printed at a size of 0.94″ x0.52″, consistent with typical hospital settings. These barcodes were then scanned using the hospital's standard scanner. The Adesso NuScan 5200TR is a 2.4GHz wireless barcode scanner (Adesso, Inc), and recognition accuracy was calculated for each barcode type and blur level. Gaussian blur was selected to systematically simulate image degradation under controlled and reproducible conditions for comparing barcode types. While this does not fully capture directional distortions such as motion blur, it enables consistent evaluation across barcodes within our experimental framework.

Design Criteria 2. Dynamic Organ-Specific Archive Implementation

As shown in Figure 4, the number of actions recorded in REDCap escalates rapidly with increasing patient sample collection. The design of the longitudinal tuple table was originally introduced in our previous work [20] and is briefly summarized here for completeness. In REDCap, an "arm" is a project structure that groups records either in fixed or longitudinal formats. While static arms are used for one-time data capture, longitudinal arms allow sequential event tracking over time. The longitudinal tuple table is implemented as an individual REDCap arm that flexibly records all operational actions associated with each barcode. These barcode actions include activities such as printing, scanning, storing, distributing, and destroying samples and are recorded without the need for predefined time points. This structure supports real-time, event-driven tracking and is well-suited for unpredictable and repeated operations in clinical workflows.

Figure 4. To deal with the dynamic organ-specific archival, we extend the implementation of a longitudinal tuple table with additional fields for recording the region from which each sample is collected, allowing for precise tracking of sample origins across various actions and patients.

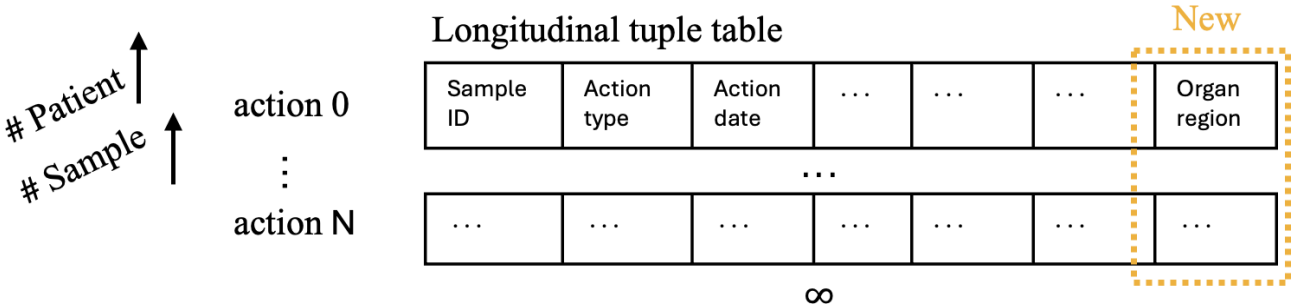

Previously, the system focused on biopsies collected from the ascending colon and terminal ileum [20]. In this study, we have extended the longitudinal tuple table to accommodate broader organ involvement by incorporating an organ archive, which documents the specific organ region associated with each recorded action. This enhancement allows for precise tracking of sample origins across diverse clinical protocols, particularly in studies requiring detailed anatomical localization.

The tuple table is exported in JSON format and processed using Python's pandas library, enabling efficient filtering, querying, and consistency checks. This flexible and scalable schema aligns with REDCap's best practices for longitudinal study designs. Further technical details can be found in our earlier conference publication [20].

In our clinical setting, the application is deployed on dedicated laboratory workstations within a controlled network. Source code is version-controlled and maintained in the MASILab GitHub group repository, which is accessible only to authorized lab members.

We use Docker to containerize three clusters of software tailored for different types of studies, including normal control, Crohn disease, colorectal cancer, ulcerative colitis, colonoscopy, and

XSL•FO
RenderX

H. pylori-associated gastric inflammation biopsies. This containerization ensures easy deployment and consistent programming environments for each study, streamlining the process for researchers and minimizing setup complexities. To showcase the effectiveness of these updates, we provide data summaries for different projects that use the enhanced design and containerized method. We release the data dictionary design for each clinical study in the data storage section and make the relevant applications available on GitHub [23].

Design Criteria 3. System Throughput Optimization With Refined REDCap Interaction

The throughput issue arises from continuous interactions with REDCap to download and update information. In the naïve system (Figure 5A), each sample is handled individually, performing one query and one update operation on REDCap for every sample scanned. This frequent interaction, involving verification and recording of each sample, significantly slows down the system as the number of patients and samples increases. To mitigate these interactions, we propose two approaches:

1. Bulk update action system: In this approach, the system can still perform frequent queries to REDCap but consolidates updates into a single bulk action. This method reduces the number of update operations, alleviating the bottleneck caused by frequent updates.
2. Local cache system: As illustrated in Figure 5B, a local cache method further optimizes throughput. Upon system initialization, REDCap records are downloaded into the local cache. Subsequent actions interact with the local cache instead of directly with REDCap. Once all actions are completed, the local cache is synchronized with REDCap, ensuring data integrity while minimizing interaction frequency.

Figure 5. Methodological approach for system throughput optimization. (**A**) Comparison between the naïve system, (where each action involves a query and update to REDCap, and (**B**) the proposed system, which uses a local cache to reduce the number of interactions with REDCap to enhance the overall throughput. REDCap: Research Electronic Data Capture.

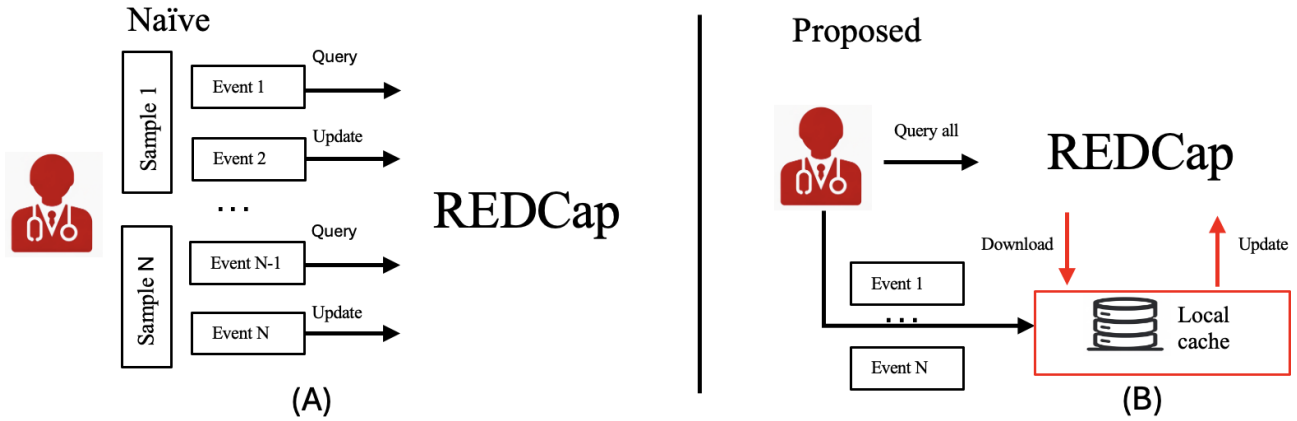

To validate optimization, we designed a controlled experiment involving 500 simulated patients, each requiring 10 distinct actions to be recorded. We measured the processing time per patient sequentially, from the first to the last, and compared system performance across three different implementation designs as the number of actions increased: (1) the naïve system, (2) the bulk update action system, and (3) the local cache system.

In the naïve system, each patient's workflow triggered 10 REDCap export application programming interface (API) calls and 10 REDCap import API calls. The bulk update action system reduced the overhead by performing 10 REDCap exports but consolidating the imports into a single REDCap import API call per patient. The local cache system further optimized the process by requiring only one REDCap export and one REDCap import API call per patient.

The purpose of this validation experiment was to provide proof-of-concept evidence that reducing the number of REDCap API calls can substantially decrease processing overhead and improve overall system throughput. This experimental setup offers a robust, systematic evaluation of how API call optimization impacts real-time data processing performance. The experiments were conducted on a workstation equipped with a 1 Gbps network connection. For each test case, we performed 5-fold evaluations and reported both the average and cumulative results with 95% CIs.

Ethical Considerations

This project involves the use of fully de-identified data obtained from Vanderbilt University Medical Center (VUMC) under Institutional Review Board (IRB) approvals Vanderbilt IRB #191738 and #191777. While the original studies were classified as human participants research, this work is a secondary analysis focused solely on barcode system methodology and does not involve any access to protected health information (PHI) or patient identifiers by our research team. Therefore, this study does not qualify as human participants research under current regulations.

Informed consent was obtained for the primary studies at the time of patient recruitment and sample collection. The IRB approvals for these studies permit secondary analysis of deidentified data without additional consent. All data used in this study were deidentified before our analysis, ensuring that no personal health information or identifiable participant data were accessible to the study team. In addition, the barcode system uses a customized codebook that encrypts barcode information. The barcode content is only translatable within the custom-built application and cannot be independently decoded,

XSL•FO
RenderX

adding an additional layer of privacy protection. No compensation was provided to participants for the studies described. No images in this manuscript or supplementary materials contain identifiable individual participants.

Results

Barcode Robustness Evaluation

The Barcode Robustness Evaluation aims to determine which barcode technology is reliable under different blurring conditions. The experimental design, illustrated in (Figure 6A), involves subjecting each barcode type to 5 rounds of Gaussian blurring. The blur radii tested were 0, 0.5, 1, 1.5, and 2 pixels, equivalent to 0, 0.00167, 0.00333, 0.005, and 0.00667 inches, respectively. As blurring increases with each round, we tested the readability of the barcodes using a standard scanner.

Figure 6. QR Code and DataMatrix show the highest resilience to blurring, maintaining near-perfect scanning accuracy across all levels of blur, while ITF and EAN-8 exhibit decreased accuracy with increasing blur levels. Code 39 failed on all validation configurations. (**A**) Examples of different barcodes subjected to varying levels of Gaussian blur. (**B**) Scanning accuracy of each barcode type under these blur conditions.

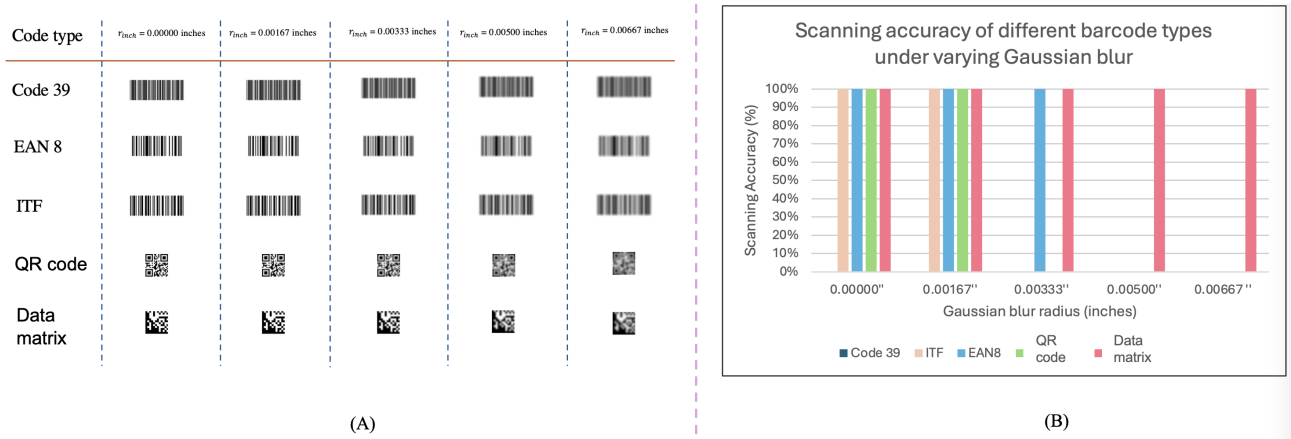

The results, presented in (Figure 6B), indicate that by the fourth round, only the DataMatrix barcode remained consistently recognizable by the scanner. In contrast, the other barcode types (Code 39, EAN-8, ITF, and QR Code) failed to be accurately scanned under the same conditions. The results demonstrate that the DataMatrix barcode maintained its readability even in the final round of testing. Therefore, based on our equipment and experimental setup, we conclude that the DataMatrix barcode demonstrates the highest robustness and is the most suitable for use in our clinical environment.

Dynamic Organ-Specific Archive Showcase

To enhance the REDCap database, we updated the table design to store organ-specific information for each action taken on a sample. This allows for precise tracking of the origin of each sample, which is essential for studies requiring detailed anatomical data.

Demonstration of System Effectiveness Through Patient Demographics and Action Management

Following the evaluation of the 3 system designs, we present 2 tables that demonstrate the effectiveness and extensive application of our system in clinical studies. Table 1 summarizes the 4 clinical studies on recruitment procedures, target population and disease focus, recruitment locations, and recruitment time frame. For clarity, the corresponding acronyms are: CMA (Targeting Gut-Microbiome in Veterans Deployment-Related Gastrointestinal and Liver Diseases), EGD (Esophagogastroduodenoscopy project for VA_Dysregulated Polyamine), GCA (Gut Cell Atlas), and DoD (Translational Impact of Electrophile Adducts in Colorectal Cancer). As noted, the names of both CMA and DoD are not derived from the initials of their project titles but are designations aligned with clinical data collection documentation for consistency.

Together, our system covers 3 study types across 4 clinical studies:

1. CMA: Data are collected without knowing which organ is involved.
2. EGD: Data are collected from one organ, known before esophagogastroduodenoscopy. Tissues are always acquired from the gastric antrum and gastric corpus.
3. GCA: Data are collected only from the AC and TI.
4. DoD: Data are collected from one organ, known before colonoscopy.

XSL•FO
RenderX

Table . Study recruitment was conducted at multiple clinical sites depending on the study focus. The table summarizes the disease focus and recruitment time frames for each study included in this work, recruited across from VA Tennessee Valley Healthcare System, Vanderbilt University Medical Center Endoscopy, and Cooperative Human Tissue Network.

Study	Target population and disease focus
CMA[a] (2023–present, VATVHS)[b]	Targeting gut microbiome in veterans deployment-related gastrointestinal and liver diseases: dysbiosis, posttraumatic stress disorder (PTSD), and epithelial and immune biology in inflammatory bowel disease (IBD) in veterans
EGD[c] (2022-present**,** VATVHS)	VA_Dysregulated Polyamine Metabolism in *Helicobacter pylori* (*H. pylori*) infection and associated stomach lesions.
GCA[d] (2019–present) VUMC[e] Endoscopy and CHTN[f]	Combinatorial single cell strategies for a Crohn disease Gut Cell Atlas_Specimen.
DoD[g] (2021–present) VUMC Endoscopy, CHTN, and VATVHS	Translational impact of electrophile adducts in IBD[h] and and colorectal cancer

[a]CMA: a Project for Targeting Gut-Microbiome in Veterans Deployment-Related Gastrointestinal and Liver Diseases: Dysbiosis, PTSD, and Epithelial and Immune Biology in Inflammatory Bowel Disease in Veterans.

[b]VATVHS: VA Tennessee Valley Healthcare System.

[c]EGD: Esophagogastroduodenoscopy, a project for VA_Dysregulated Polyamine.

[d]GCA: Gut cell atlas.

[e]VUMC: Vanderbilt University Medical Center.

[f]CHTN: Cooperative Human Tissue Network.

[g]DoD: a project for Translational Impact of Electrophile Adducts in Colorectal Cancer.

[h]IBD: inflammatory bowel disease.

Table 2 demonstrates the demographic range handled by the system in terms of age, race, and gender. Summarized by Table 3, the whole studies collectively handled a total of 38,819 actions, with the GCA study alone accounting for 16,911 actions. These actions include scanning, printing, destroying, distributing, and storing samples, showcasing the system's extensive operational capabilities. The sample types managed include fresh, fixed, and frozen tissue; serum; stool; and DNA samples for Crohn disease and ulcerative colitis. The diverse sample types and the large, significant number of actions performed—such as 5106 prints, 4092 scans, 2113 destructions, 565 distributions, and 1360 storages in the nonsurgical CD case of the GCA study—highlight the system's flexibility and efficiency in managing different biological materials and operational tasks.

XSL•FO
RenderX

Table . A comprehensive overview of the patient demographics involved in various clinical studies using our system.

Study name and surgery	Study case	Patient, n/N (%)	Age, mean (SD)	Gender, n/N (%)			Race, n/N (%)				
				Male	Female	N/A[a]	Asian	Black/African American	White/Caucasian	Hispanic/Latino	N/A[a]
CMA[b]											
No		34/34 (100)	48.6 (2.0)	27/34 (80)	5/34 (15)	2/34 (6)	0/34 (0)	3/34 (9)	29/34 (85)	0/34 (0)	2/34 (6)
EGD[c]											
No		52/52 (100)	56.2 (2.0)	45/52 (87)	5/52 (10)	2/52 (3)	1/52 (2)	17/52 (33)	32/52 (62)	0/52 (0)	2/52 (3)
GCA[d]											
No	CD[e]	125/179 (70)	36.1 (1.3)	47/179 (26)	78/179 (44)	0/179 (0)	4/179 (2)	13/179 (7)	106/179 (60)	2/179 (1)	0/179 (0)
No	Control	41/179 (23)	54.9 (1.0)	16/179 (8)	25/179 (14)	3/179 (1)	1/179 (1)	3/179 (1)	36/179 (20)	1/179 (1)	0/179 (0)
Yes	CD	10/179 (6)	N/A	0/179 (0)	0/179 (0)	10/179 (6)	0/179 (0)	0/179 (0)	0/179 (0)	0/179 (0)	10/179 (6)
Yes	Control	3/179 (1)	N/A	0/179 (0)	0/179 (0)	3/179 (1)	0/179 (0)	0/179 (0)	0/179 (0)	0/179 (0)	3/179 (1)
DOD[f]											
No	CD	38/144 (26)	38.9 (2.5)	15/144 (10)	23/144 (16)	0/144 (0)	0/144 (0)	2/144 (1)	34/144 (24)	0/144 (0)	2/144 (1)
No	Control	19/144 (13)	52.9 (2.9)	13/144 (9)	6/144 (4)	0/144 (0)	1/144 (1)	0/144 (0)	15/144 (10)	0/144 (0)	3/144 (2)
No	UC[g]	50/144 (35)	41.0 (2.0)	23/144 (16)	27/144 (18)	0/144 (0)	0/144 (0)	2/144 (1)	43/144 (30)	1/144 (1)	4/144 (3)
Yes	CD	19/144 (13)	N/A	0/144 (0)	0/144 (0)	0/144 (0)	19/144 (13)	0/144 (0)	0/144 (0)	0/144 (0)	19/144 (13)
Yes	Control	1/144 (1)	N/A	0/144 (0)	0/144 (0)	0/144 (0)	1/144 (1)	0/144 (0)	0/144 (0)	0/144 (0)	1/144 (1)
Yes	UC	5/144 (3)	N/A	0/144 (0)	0/144 (0)	0/144 (0)	5/144 (3)	0/144 (0)	0/144 (0)	0/144 (0)	5/144 (3)
Yes	Cancer	12/144 (8)	N/A	0/144 (0)	0/144 (0)	0/144 (0)	12/144 (8)	0/144 (0)	0/144 (0)	0/144 (0)	12/144 (8)

[a]N/A: not available.

[b]CMA: project for Targeting Gut-Microbiome in Veterans Deployment-Related Gastrointestinal and Liver Diseases: Dysbiosis, PTSD, and Epithelial and Immune Biology in Inflammatory Bowel Disease in Veterans.

[c]EGD: Esophagogastroduodenoscopy, a project for VA_Dysregulated Polyamine.

[d]GCA: Gut Cell Atlas.

[e]CD: Crohn disease.

[f]DOD: Project for translational impact of electrophile adducts in colorectal cancer.

[g]UC: ulcerative colitis.

XSL•FO
RenderX

Table . Various actions and sample types managed by our system across different studies.

Study name and surgery	Study case	Action	Sample type						Action type				
			Fresh	Fixed	Frozen	Serum	Stool	DNA	Printed	Scanned	De-stroyed	Dis-tributed	Stored
CMA,[a] n/N (%)													
No		3819/3819 (100)	278/3819 (7)	266/3819 (7)	1222/3819 (32)	1560/3819 (41)	493/3819 (13)	0/3819 (0)	1456/3819 (38)	1087/3819 (28)	271/3819 (7)	94/3819 (2)	911/3819 (24)
EGD,[b] n/N (%)													
No		1372/1372 (100)	343/1372 (25)	313/1372 (23)	716/1372 (52)	0/1372 (0)	0/1372 (0)	0/1372 (0)	481/1372 (35)	439/1372 (32)	37/1372 (3)	206/1372 (1)	209/1372 (1)
GCA,[c] n/N (%)													
No	CD[d]	13236/16908 (78)	1733/16908 (10)	1877/16908 (11)	2385/16908 (14)	5176/16908 (31)	1617/16908 (10)	448/16908 (3)	5106/16908 (30)	4092/16908 (24)	2113/16908 (12)	565/16908 (3)	1360/16908 (8)
No	Control	3288/16908 (19)	218/16908 (1)	306/16908 (2)	449/16908 (3)	1637/16908 (10)	548/16908 (3)	130/16908 (1)	1085/16908 (6)	1145/16908 (7)	290/16908 (2)	204/16908 (1)	564/16908 (3)
Yes	CD	203/16908 (1)	62/16908 (0)	65/16908 (0)	76/16908 (0)	0/16908 (0)	0/16908 (0)	0/16908 (0)	63/16908 (0)	72/16908 (0)	5/16908 (0)	40/16908 (0)	23/16908 (0)
Yes	Control	181/16908 (1)	58/16908 (0)	57/16908 (0)	66/16908 (0)	0/16908 (0)	0/16908 (0)	0/16908 (0)	48/16908 (0)	84/16908 (0)	0/16908 (0)	32/16908 (0)	17/16908 (0)
DOD,[e] n													
No	CD	2374/9076 (26)	135/9076 (1)	133/9076 (1)	419/9076 (5)	1556/9076 (17)	0/9076 (0)	131/9076 (1)	868/9076 (10)	788/9076 (9)	129/9076 (1)	114/9076 (1)	475/9076 (5)
No	Control	1366/9076 (15)	78/9076 (1)	75/9076 (1)	241/9076 (3)	897/9076 (10)	0/9076 (0)	75/9076 (1)	582/9076 (6)	409/9076 (5)	62/9076 (1)	60/9076 (1)	253/9076 (3)
No	UC[f]	3371/9076 (37)	186/9076 (2)	197/9076 (2)	683/9076 (8)	2118/9076 (23)	0/9076 (0)	187/9076 (2)	1161/9076 (13)	1142/9076 (13)	227/9076 (3)	157/9076 (2)	684/9076 (8)
Yes	CD	819/9076 (9)	122/9076 (1)	122/9076 (1)	506/9076 (6)	11/9076 (0)	0/9076 (0)	58/9076 (1)	292/9076 (3)	271/9076 (3)	107/9076 (1)	56/9076 (1)	93/9076 (1)
Yes	Control	186/9076 (2)	30/9076 (0)	30/9076 (0)	111/9076 (1)	0/9076 (0)	0/9076 (0)	15/9076 (0)	102/9076 (1)	52/9076 (1)	1/9076 (0)	10/9076 (0)	21/9076 (0)
Yes	UC	425/9076 (5)	64/9076 (1)	65/9076 (1)	257/9076 (3)	11/9076 (0)	0/9076 (0)	28/9076 (0)	169/9076 (2)	106/9076 (1)	32/9076 (0)	37/9076 (0)	81/9076 (1)
Yes	Cancer	535/9076 (6)	96/9076 (1)	98/9076 (1)	295/9076 (3)	0/9076 (0)	0/9076 (0)	46/9076 (1)	183/9076 (2)	169/9076 (2)	74/9076 (1)	56/9076 (1)	52/9076 (1)

[a]CMA: Project for Targeting Gut-Microbiome in Veterans Deployment Related Gastrointestinal and Liver Diseases: Dysbiosis, PTSD, and Epithelial and Immune Biology in Inflammatory Bowel Disease in Veterans.

[b]EGD: Esophagogastroduodenoscopy, a project for VA_Dysregulated Polyamine.

[c]GCA: Gut Cell Atlas.

[d]CD: Crohn disease.

[e]DOD: Project for translational impact of electrophile adducts in colorectal cancer.

[f]UC: ulcerative colitis.

System Throughput Optimization With Refined REDCap Interaction

The system throughput evaluation demonstrated consistent performance differences across the 3 implementations. Pairwise 2-tailed *t* tests were conducted, demonstrating that the performance differences between test cases were statistically significant ($P<.001$). As shown in Figure 7**,** the naïve system's per-patient processing time steadily increased with the number of patients, largely due to the accumulation of overhead from multiple REDCap API calls (10 export and 10 import calls per patient). The bulk import system effectively reduced this overhead by consolidating import operations, resulting in a lower and more stable per-patient processing time. The local cache system exhibited the most efficient performance, maintaining minimal and stable processing times across all patient counts by further reducing the API call frequency to one export and one import per patient.

XSL•FO
RenderX

Figure 7. Per-patient system throughput across 3 implementations with 95% CIs. The naïve system shows increasing processing time per patient as the number of patients grows. The bulk import system reduces this overhead, while the local cache system maintains the lowest and most stable processing time per patient across all patient counts.

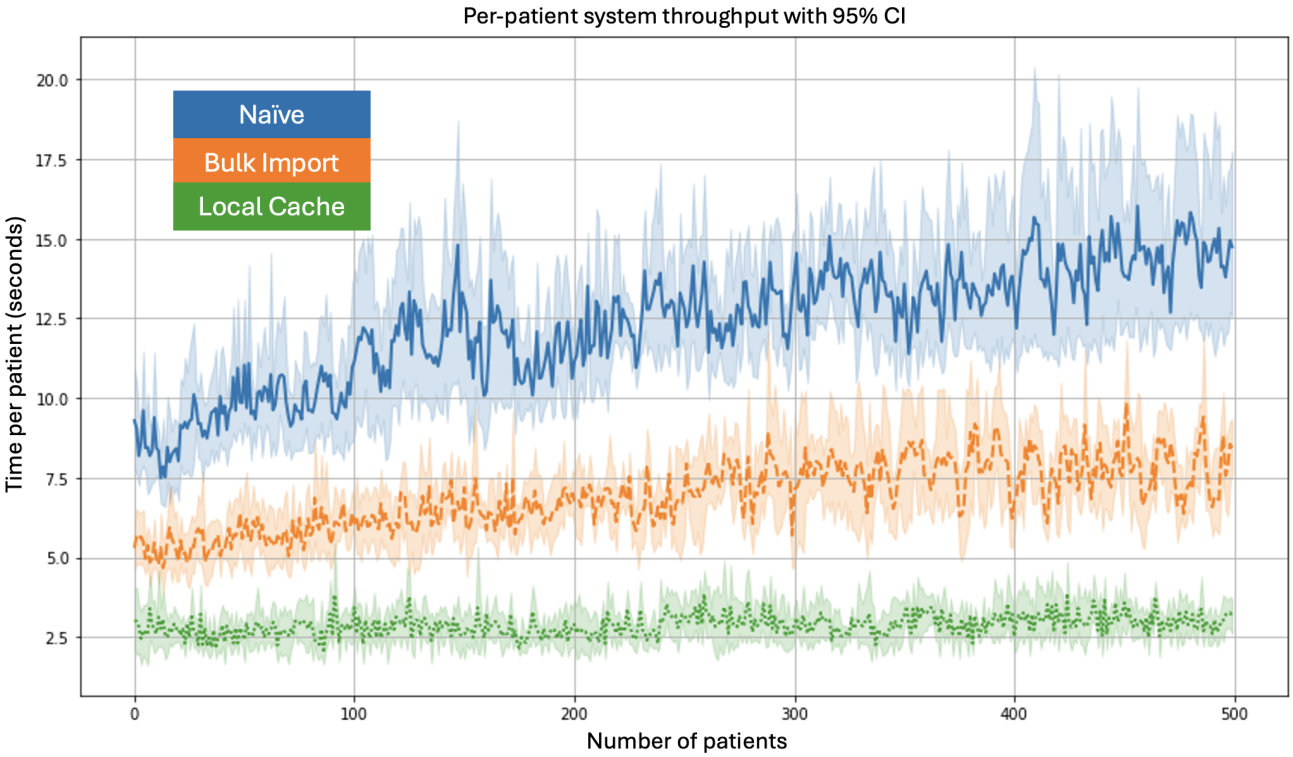

The cumulative processing time, presented in Figure 8, further underscores the efficiency gains of the optimized systems. The naïve system displayed the steepest cumulative growth, reflecting its scalability limitations. In contrast, the bulk import system achieved moderate improvement, and the local cache system substantially minimized cumulative processing time, demonstrating superior scalability and throughput. These results validate the hypothesis that reducing the number of REDCap API calls is critical for improving system efficiency and highlight the robustness of the local cache approach for large-scale clinical workflows.

Figure 8. Cumulative system throughput across 3 implementations with 95% CIs. The naïve system exhibits the steepest cumulative processing time curve. The bulk import system reduces cumulative load, while the local cache system achieves the lowest cumulative processing time, demonstrating superior scalability.

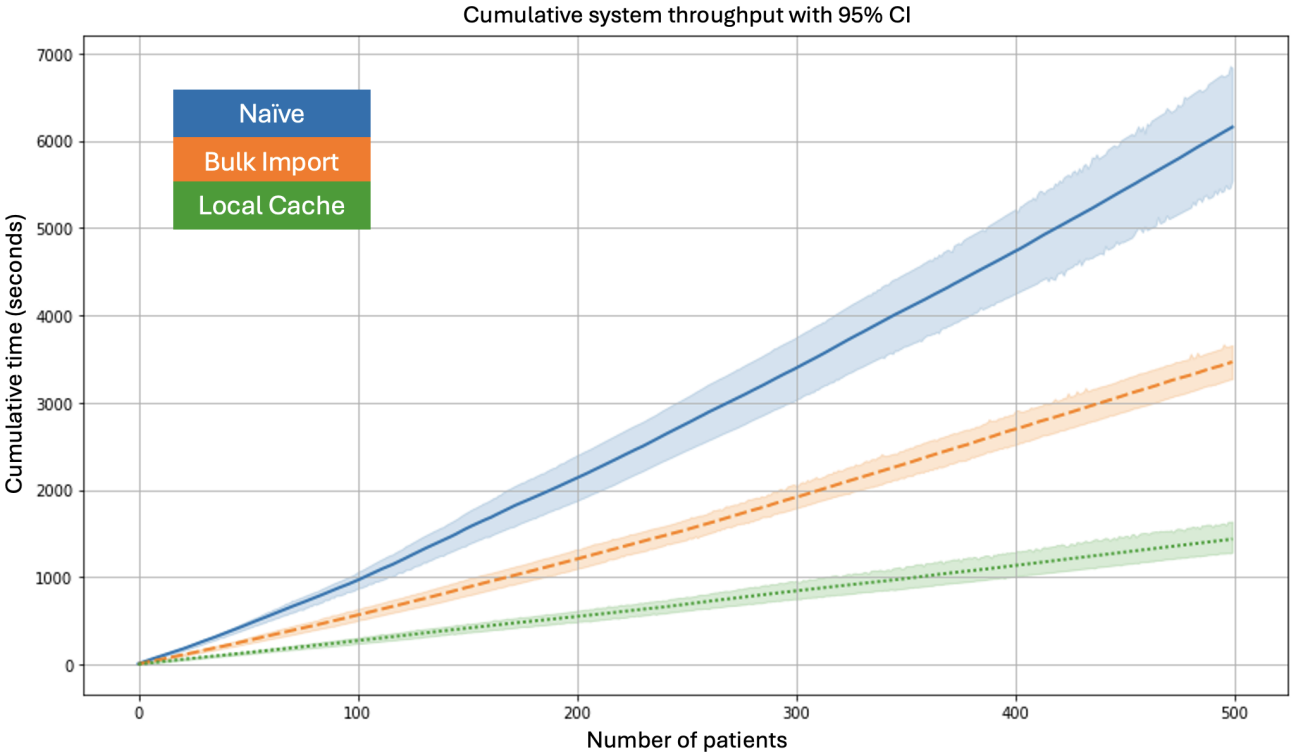

XSL•FO
RenderX

Discussion

Overview

In this study, we aimed to enhance clinical data management by improving barcode robustness, increasing data granularity, and optimizing system throughput to better support real-world clinical workflows. Our results demonstrated that the DataMatrix barcode provided superior recognition accuracy under blurred conditions, addressing barcode robustness issues identified in previous work. By incorporating a dynamic organ-specific archive within REDCap, we enabled precise tracking of sample origins, supporting studies with complex and evolving anatomical requirements. In addition, system throughput was significantly improved through the introduction of a local cache mechanism, reducing REDCap API interaction overhead and increasing processing efficiency. Together, these enhancements address key bottlenecks in barcode informatics systems and validate the effectiveness of the proposed design criteria across diverse gastrointestinal clinical studies. In the following subsections, we provide a detailed discussion of these findings, including comparisons to previous work and an evaluation of system limitations.

Principal Results

The enhancements proposed and tested in this study significantly improve the efficiency and robustness of the cross-platform informatics system initially developed by Bao et al [20]. By addressing key issues such as barcode robustness, detailed data archiving, and system throughput, we have demonstrated tangible benefits in managing clinical data. The DataMatrix barcode was identified as the most reliable under various blurring conditions, ensuring accurate sample tracking even in suboptimal environments.

Although both QR Code and DataMatrix barcodes demonstrated high resilience under blurring, our final selection of the DataMatrix barcode was driven by the practical constraints of our clinical workflow. Specifically, the DYMO LabelWriter 450 printer (~US$200) used in our system could not consistently produce QR codes with sufficient resolution on the Diversified Biotech DTCR-6000 Direct Thermal Cryo-Tags (Sigma-Aldrich; 5″×1.5″, US$45.50 per roll of 500), which are required to fit 1.5 mL cryovials. DataMatrix barcodes are more space-efficient and less sensitive to printer resolution, enabling reliable scanning at this small label size. The Adesso NuScan 5200TR scanner (~US$200) used in our system was capable of scanning both barcode types robustly, confirming that scanner limitations were not the primary issue.

The implementation of a dynamic organ-specific archive in REDCap allows for precise tracking of sample origins, enhancing data granularity and usability. Organ selection within our system is guided by the clinical protocols of each study, which define the sampling regions in advance. Given that each project may target different anatomical sites, the system allows flexibility in specifying the organ for each action. In cases where ambiguous or overlapping anatomical regions arise, both the printer app, data entry app, and the REDCap schema include dedicated comment fields that enable clinicians to record relevant details or uncertainties. It is important to note that the barcode system serves as a tracking mechanism and does not resolve anatomical ambiguities; instead, such contextual details are manually documented to ensure clinical accuracy.

In clinical practice, evolving study protocols and mid-study changes in data collection requirements are common. Our system is designed to accommodate some of these changes, such as adding new label types or supporting additional sample collection strategies (eg, switching to frozen biopsies when fresh samples are unavailable) without disrupting the existing workflow. More substantial protocol changes, such as significant modifications in label design or data entry workflows, may require the deployment of new application versions. Our framework supports this by allowing distinct study cases with customized printing and data entry workflows. One key lesson learned is the value of improving label printing flexibility to enable on-the-fly adaptations in future versions of the system.

REDCap's built-in caching mechanisms, such as project dashboard caching, are designed to improve web interface performance by temporarily storing data in the browser when records remain unchanged. These optimizations do not apply to API-based workflows or external applications that require real-time, high-frequency interactions, such as barcode scanning and sample tracking. Our system introduces a local cache layer that supports customized, app-integrated workflows, significantly reducing the need for repetitive API calls and improving throughput during barcode-driven operations, ensuring smooth operation even with large patient datasets. These improvements collectively enhance the system's ability to support complex and large-scale clinical studies, thereby contributing to more reliable and efficient data management practices.

Comparison With Previous Work

Our work builds on the foundation established by previous work Bao et al [20], introducing significant improvements in barcode technology, data granularity, and system throughput.

The system incorporates practical error-handling mechanisms for both barcode scanning and REDCap database accessibility. In early versions [20], we observed occasional scanning failures with the EAN-8 barcode format, which led us to adopt the more robust DataMatrix format in the current system. No barcode scanning failures have been observed since this transition. In addition, the application supports manual barcode entry to handle any potential scanner-related issues.

While the original system provided a complete framework for clinical data management, our enhancements address specific bottlenecks identified in real-world usage. Previous studies have highlighted the importance of reliable barcode tracking and data entry consistency in clinical settings, but our study uniquely combines these aspects with a comprehensive approach to system optimization, demonstrating substantial performance gains. The scalability evaluation using 500 patients and 10 actions per patient was designed as a proof-of-concept to assess system performance under moderate workload conditions. The primary performance constraint arises from the frequency and volume of REDCap API interactions, which are tightly coupled with system throughput. While the current local cache

XSL•FO
RenderX

mechanism significantly reduces API traffic, scaling the system to support significantly larger datasets would require further architectural optimization, including improved cache synchronization, enhanced concurrency handling, and more systematic pressure testing to assess system robustness at scale. Exploring these enhancements will be essential to support high-volume clinical studies in future deployments.

Limitations

Our barcode robustness assessment did not directly incorporate motion blur or real-world scanner variability. Future studies should address these factors to further validate barcode performance in practical clinical environments. In addition, the evaluation of barcode robustness was conducted under controlled conditions, which may not fully capture the broader variability of real-world clinical environments. The system's performance and usability could also benefit from further enhancements to the user interface to better accommodate a wider range of clinical study applications.

In cases where the REDCap database becomes temporarily inaccessible, the system either prevents application launch (as a REDCap record export is required during initialization) or triggers a warning message if a connection failure occurs during active operation. While these mechanisms provide basic error notifications, further improvements in offline operation support and automatic recovery strategies will be considered in future work.

Despite these, the local cache mechanism, while effective, is currently designed for use by a single operator at a time. This may limit its applicability in larger research settings where multiple operators need to work simultaneously. Specifically, the system supports a text-based flag to prevent simultaneous access by multiple users on the same workstation; this feature is not enabled in the production environment. In clinical practice, strict single-user operation per workstation is enforced to maintain data integrity and simplify workflow management. Multi-user concurrency across different workstations is not supported in the current system. Supporting such functionality would require robust cross-machine locking, cache synchronization, and efficient REDCap API management, which remains an important area for future development.

Furthermore, data integrity between the local cache and REDCap is safeguarded by ensuring that the local cache is updated only after a successful sync process is completed. If the system fails to retrieve updated REDCap records, a warning message is triggered to alert the user of the issue. However, this safeguard has not been systematically validated under high-volume or failure-prone conditions. Future work will focus on more rigorous testing of the sync process, including controlled failure scenarios and large-scale data loads, to ensure cache consistency and system resilience.

Finally, Docker containers in this work are provided to support users who wish to test the system or replicate the design criteria but are not the primary security mechanism. Future work will explore strategies for securely deploying this system across multiple research teams and sites.

Conclusions

The proposed enhancements significantly improve barcode robustness, data granularity, and system throughput in our informatics system, enabling more efficient and reliable clinical data management. These optimizations not only support diverse clinical research workflows but also demonstrate a scalable framework that can be adapted for complex, multiorgan studies across different institutions. Beyond the immediate performance gains, this work highlights the importance of system design strategies that minimize external API dependencies to achieve throughput improvements in high-volume clinical settings. Looking forward, future enhancements will focus on developing synchronization capabilities to support simultaneous multioperator use and creating a more versatile user interface to accommodate a broader range of study designs. These advances will further expand the system's applicability and promote its integration into larger, multisite research environments.

Acknowledgments

This research was funded by The Leona M. and Harry B. Helmsley Charitable Trust (grant G-1903- - 03793) and NSF CAREER (grant 1452485). Resources from the Advanced Computing Center for Research and Education (ACCRE) at Vanderbilt University supported this study. Additional support came from the National Center for Research Resources (grant UL1 RR024975-01, now at the National Center for Advancing Translational Sciences, grant 2 UL1UL1 TR000445-06), the National Institute of Diabetes and Digestive and Kidney Diseases (grants R01DK103831 and R01DK128200), the Department of Veterans Affairs (grants I01BX004366, I01CX002662, I01CX002171, and I01CX002473), the Department of Defense (grant W81XWH-21-1-0617) and the NIH-supported VUMC Digestive Disease Research Center (grant P30DK058404). De-identified imaging datasets were sourced from ImageVU, backed by the VICTR CTSA award (ULTR000445 from NCATS/NIH), Vanderbilt University Medical Center funding, and the Patient-Centered Outcomes Research Institute (PCORI; contract CDRN-1306- - 04869).

Generative AI, including the use of ChatGPT (OpenAI) [24], was used to assist in developing visualization code, as well as in drafting, revising, and refining portions of this manuscript, including language polishing, restructuring responses to reviewers, and organizing the content. All scientific content, study design, data analysis, and interpretations are original and were developed independently by the authors.

XSL•FO
RenderX

Data Availability

The source code, including system implementation, data visualization scripts, and REDCap table design, is publicly available in the GitHub repository [23]. A Docker image supporting system deployment is also available for public use. Due to ethical and regulatory restrictions, the raw clinical patient recruitment data cannot be shared without Institutional Review Board (IRB) approval. However, a summarized dataset is available upon reasonable request to the corresponding author.

Authors' Contributions

Conceptualization: BAL (lead), SB (equal), KTW (equal),
Data curation: RZ (lead), SB (supporting), SC(supporting), RT (supporting), KC (supporting), LR (supporting)
Formal analysis: RZ (lead), SB (equal) KR (supporting)
Funding acquisition: BAL (lead), KTW (equal), LAC (equal)
Investigation: BAL (lead), KTW (equal), LAC (equal)
Methodology: SB (lead), RZ (supporting), KR (supporting), CG (supporting), MEK (supporting), LZ (supporting), PAH (supporting)
Project administration: SC (lead), SB (equal), RT (supporting), KC (supporting)
Resources: KR (supporting), PAH (supporting)
Supervision: BL (lead), YL (supporting), ZW (supporting), QL (supporting), KSL (supporting), YH (supporting), LAC (supporting), KTW (supporting)
Validation: RZ (lead), SB (equal), PAH (supporting), BAL (supporting)
Visualization: RZ (lead), SB (equal)
Writing – original draft: RZ (lead), SB (equal), BAL (supporting)
Writing – review & editing: RZ (lead), SB (equal), LZ (supporting), KTW (supporting), BAL (supporting), ZW (supporting)

Conflicts of Interest

Yike Li (YL) served as an associate editor for the Journal of Medical Internet Research at the time of manuscript submission. YL has abstained from participating in any peer-reviewed or editorial decision-making processes related to this article.

References

1. Pallua JD, Brunner A, Zelger B, Schirmer M, Haybaeck J. The future of pathology is digital. Pathol Res Pract 2020 Sep;216(9):153040. [doi: 10.1016/j.prp.2020.153040] [Medline: 32825928]
2. Caie PD, Harrison DJ. Digital pathology: path into the future. In: Tissue Phenomics: Profiling Cancer Patients for Treatment Decisions: Pan Stanford Publishing Pte Ltd; 2018:185-198. [doi: 10.1201/9781351134279]
3. Bera K, Schalper KA, Rimm DL, Velcheti V, Madabhushi A. Artificial intelligence in digital pathology - new tools for diagnosis and precision oncology. Nat Rev Clin Oncol 2019 Nov;16(11):703-715. [doi: 10.1038/s41571-019-0252-y] [Medline: 31399699]
4. Jahn SW, Plass M, Moinfar F. Digital pathology: advantages, limitations and emerging perspectives. J Clin Med 2020 Nov 18;9(11):3697. [doi: 10.3390/jcm9113697] [Medline: 33217963]
5. Van Es SL. Digital pathology: semper ad meliora. Pathology (Phila) 2019 Jan;51(1):1-10. [doi: 10.1016/j.pathol.2018.10.011] [Medline: 30522785]
6. Hartman DJ, Pantanowitz L, McHugh JS, Piccoli AL, OLeary MJ, Lauro GR. Enterprise implementation of digital pathology: feasibility, challenges, and opportunities. J Digit Imaging 2017 Oct;30(5):555-560. [doi: 10.1007/s10278-017-9946-9] [Medline: 28116576]
7. Montezuma D, Monteiro A, Fraga J, et al. Digital pathology implementation in private practice: specific challenges and opportunities. Diagnostics (Basel) 2022 Feb 18;12(2):529. [doi: 10.3390/diagnostics12020529] [Medline: 35204617]
8. Retamero JA, Aneiros-Fernandez J, Del Moral RG. Complete digital pathology for routine histopathology diagnosis in a multicenter hospital network. Arch Pathol Lab Med 2020 Feb;144(2):221-228. [doi: 10.5858/arpa.2018-0541-OA] [Medline: 31295015]
9. Eccher A, Dei Tos AP, Scarpa A, et al. Cost analysis of archives in the pathology laboratories: from safety to management. J Clin Pathol 2023 Oct;76(10):659-663. [doi: 10.1136/jcp-2023-209035] [Medline: 37532289]
10. Griffin J, Treanor D. Digital pathology in clinical use: where are we now and what is holding us back? Histopathology 2017 Jan;70(1):134-145. [doi: 10.1111/his.12993] [Medline: 27960232]
11. Fraggetta F, L'Imperio V, Ameisen D, et al. Best practice recommendations for the implementation of a digital pathology workflow in the anatomic pathology laboratory by the European Society of Digital and Integrative Pathology (ESDIP). Diagnostics (Basel) 2021 Nov 22;11(11):2167. [doi: 10.3390/diagnostics11112167] [Medline: 34829514]
12. Nourani A, Ayatollahi H, Dodaran MS. A review of clinical data management systems used in clinical trials. RRCT 2019 Jan 30;14(1):10-23. [doi: 10.2174/1574887113666180924165230]

XSL•FO
RenderX

13. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381. [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
14. Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: building an international community of software platform partners. J Biomed Inform 2019 Jul;95:103208. [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]
15. Menon U, Lance P, Szalacha LA, et al. Adaptation of colorectal cancer screening tailored navigation content for American Indian communities and early results using the intervention. Implement Sci Commun 2022 Jan 28;3(1):6. [doi: 10.1186/s43058-022-00253-x] [Medline: 35090575]
16. Cochran AR, Raub KM, Murphy KJ, Iannitti DA, Vrochides D. Novel use of REDCap to develop an advanced platform to display predictive analytics and track compliance with enhanced recovery after surgery for pancreaticoduodenectomy. Int J Med Inform 2018 Nov;119:54-60. [doi: 10.1016/j.ijmedinf.2018.09.001] [Medline: 30342686]
17. Gesell SB, Halladay JR, Mettam LH, Sissine ME, Staplefoote-Boynton BL, Duncan PW. Using REDCap to track stakeholder engagement: A time-saving tool for PCORI-funded studies. J Clin Trans Sci 2020 Apr;4(2):108-114. [doi: 10.1017/cts.2019.444]
18. Lininger MR, Kirby C, Laurila KA, et al. Building research infrastructure: the development of a technical assistance group-service center at an RCMI. Int J Environ Res Public Health 2022 Dec 23;20(1):191. [doi: 10.3390/ijerph20010191] [Medline: 36612513]
19. Armstrong B, Habtemariam D, Husser E, et al. A mobile app for delirium screening. JAMIA Open 2021 Apr;4(2):ooab027. [doi: 10.1093/jamiaopen/ooab027] [Medline: 34549169]
20. Bao S, Chiron S, Tang Y, et al. A cross-platform informatics system for the gut cell atlas: integrating clinical, anatomical and histological data. Proc SPIE Int Soc Opt Eng 2021. [doi: 10.1117/12.2581074]
21. Isohanni J. Use of functional ink in a smart tag for fast-moving consumer goods industry. J Package Technol Res 2022 Oct;6(3):187-198. [doi: 10.1007/s41783-022-00137-4]
22. Dubská M, Herout A, Havel J. Real-time precise detection of regular grids and matrix codes. J Real-Time Image Proc 2016 Jan;11(1):193-200. [doi: 10.1007/s11554-013-0325-6]
23. MASILab/barcoderedcapintegration. GitHub. URL: https://github.com/MASILab/BarcodeRedcapIntegration [accessed 2025-09-03]
24. ChatGPT. Open AI. URL: https://chatgpt.com/ [accessed 2025-09-03]

Abbreviation

AC: ascending colon
API: application programming interface
CD: Crohn disease
CHTN: Cooperative Human Tissue Network
CMA: Project for Targeting Gut-Microbiome in Veterans Deployment Related Gastrointestinal and Liver Diseases: Dysbiosis, PTSD, and Epithelial and Immune Biology in Inflammatory Bowel Disease in Veterans.
DC: descending colon
DoD: Project for Translational Impact of Electrophile Adducts in Colorectal Cancer
EGD: esophagogastroduodenoscopy
GCA : Gut Cell Atlas
***H. pylori*:** *Helicobacter pylori*
IBD: inflammatory bowel disease
ITF: Interleaved 2 of 5
PTSD: posttraumatic stress disorder
REDCap: Research Electronic Data Capture
SC: sigmoid colon
TC: transverse colon
TI: terminal ileum
UC: ulcerative colitis
VATVHS: VA Tennessee Valley Healthcare System
VUMC: Vanderbilt University Medical Center

XSL•FO
RenderX

Edited by A Mavragani; submitted 12.12.24; peer-reviewed by C Mahanty; revised version received 29.06.25; accepted 15.07.25; published 12.09.25.

Please cite as:
Zhang R, Chiron S, Tyree R, Carson K, Raber L, Ramadass K, Gao C, Kim ME, Zuo L, Li Y, Wan Z, Harris PA, Liu Q, Lau KS, Coburn LA, Wilson KT, Huo Y, Landman BA, Bao S
Enhancing Clinical Data Management Through Barcode Integration and Research Electronic Data Capture: Scalable and Adaptable Implementation Study
JMIR Form Res 2025;9:e70016
URL: https://formative.jmir.org/2025/1/e70016
doi:10.2196/70016

© Rendong Zhang, Sophie Chiron, Regina Tyree, Kate Carson, Larry Raber, Karthik Ramadass, Chenyu Gao, Michael E Kim, Lianrui Zuo, Yike Li, Zhiyu Wan, Paul A Harris, Qi Liu, Ken S Lau, Lori A Coburn, Keith T Wilson, Yuankai Huo, Bennett A Landman, Shunxing Bao. Originally published in JMIR Formative Research (https://formative.jmir.org), 12.9.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Acceptability of Active and Passive Data Collection Methods for Mobile Health Research: Cross-Sectional Survey of an Online Adult Sample in the United States

Nelson Roque[1,2], PhD; John Felt[1], PhD

[1]Center for Healthy Aging, The Pennsylvania State University, 422 Biobehavioral Health Building, University Park, PA, United States

[2]Department of Psychology, University of Central Florida, Orlando, FL, United States

Corresponding Author:
Nelson Roque, PhD
Center for Healthy Aging, The Pennsylvania State University, 422 Biobehavioral Health Building, University Park, PA, United States

Abstract

Background: Digital health technologies, including wearable devices and app-based cognitive and health assessments, are pervasive and crucial to better understanding important public health problems (eg, Alzheimer's disease and related dementias). Central to understanding mechanisms driving individuals' willingness to share various data streams are concerns regarding data privacy, security, and control over generated data.

Objective: This survey was designed to learn more about attitudes and opinions related to digital health technologies and the sharing of associated data.

Methods: A total of 1509 adults were recruited from Prolific to complete an online survey via Qualtrics. Of these, 1489 participants provided valid data for analyses. Participants completed a structured survey consisting of multiple modules after informed consent was provided. These included: (1) demographic characteristics; (2) prior research experience; (3) mobility factors (eg, use of mobility aids, driving frequency); (4) technology ownership (eg, smartphones, tablets, home Wi-Fi); (5) social media use (eg, frequency of engagement with platforms such as Facebook, Instagram, and TikTok); (6) willingness to contribute different types of data across categories, including activities, sensors, and metadata; (7) opinions about data control and privacy options (eg, data deletion, stream-specific control); and (8) willingness to interact with assistive technologies such as robots, for Instrumental Activities of Daily Living.

Results: The final cohort (N=1489) had a mean age of 35.5 years (SD 12.0), was 44% female (n=652), and predominantly identified as White (76%, n=1134), with high rates of smartphone ownership (99%, n=1479) and home Wi-Fi access (98%, n=1464). Participants were most willing to share data streams with clear health implications and least willing to share data streams with greater privacy or reidentification potential (eg, GPS location, in-vehicle dashcam footage). On average, people were willing to complete ambulatory cognitive assessments for 56.7 (SD 36.2) days, air quality monitoring for 58.1 (SD 37.7) days, and GPS location monitoring for 37 (SD 39.0) days. People expected control over their data, including the ability to delete all or specific streams of the data contributed for research. Most participants prioritized control over their data, with 71% (n=1061) favoring the ability to delete all data contributed for research purposes. Stream-specific data deletion (65%, n=960) and time-specific deletion (44%, n=653) were also valued; interest in sharing data with insurance providers (30%, n=453) or caregivers (26%, n=384) was notably lower.

Conclusions: Findings have implications for the design of digital health technologies and education-related to the use and implications of collected data.

(JMIR Form Res 2025;9:e64082) doi:10.2196/64082

KEYWORDS

mHealth; privacy; smartphones; sensors; acceptability; mobile health; attitudes; opinions; digital health technologies; online survey; surveys; data privacy; questionnaires

XSL•FO
RenderX

Introduction

Background

Digital health technologies, including app-based cognitive and psychosocial assessments and wearable sensor-based health monitoring devices, are pervasive and crucial for better understanding and predicting trajectories related to important public health outcomes such as Alzheimer's disease and related dementias. The availability of detailed sensor data streams and data available via application programming interfaces creates new opportunities for scientific discovery, especially considering that passively collected data may be more robust for measuring aspects of daily life that can be influenced by observation, memory interference, or retrospective biases. The increase in easily shareable metadata (eg, Google Takeout) also creates a new fruitful avenue for research, however the extent public awareness and willingness to share these data and from other data streams remains unclear. Past studies have shown that in general, the overall acceptance for various smartphone-based tasks and sensing modalities is low, with variations in type of activity and sensor [1, 2]. However, repeated prompting for specific data streams has been shown to increase willingness to share [3]. In general, participants are interested in learning from their own data, and a proportion of these are also interested in sharing their information with a health care provider [4].

Privacy, Security, and Data Controls

Critical to understanding mechanisms driving willingness to share various data streams are concerns about data privacy, security, and control over generated data [3, 5]Once the data are collected, different members of a research team, often across different institutions, work with deidentified versions of the datasets. In some cases, data are stored indefinitely for algorithm development—seeking new purposes or uses for the data that were not originally considered. In other cases, advances in machine learning have made it possible to reidentify individuals with enough precision to warrant security concerns over warehousing sensitive data without appropriate deidentification procedures [6]). As such, it is imperative to consider the most important stakeholder, the respondent and to assess their expectations regarding the data they generate for research and the level of controls they expect.

Assessment: Active vs Passive?

To answer important scientific questions, researchers are interested in collecting data across the spectrum of participant involvement and burden, ranging from active data collection activities (eg, completing a survey, cognitive assessments), to passive data collection (eg, extracting sensor data streams, reading existing metadata [7]; Figure 1). Passive data collection methods refer to data captured automatically without participant effort (eg, sensors, metadata), whereas active methods require participant involvement (eg, completing surveys or assessments). Active data collection also requires input from the respondent in a manner that makes the goal of the research more apparent. On the other hand, passive assessment does not requiring respondent burden or input, but rather, relies on data that is collected continuously in the background.

Figure 1. Active-passive continuum of research data streams: activities, sensors, and metadata (examples listed for each data stream category).

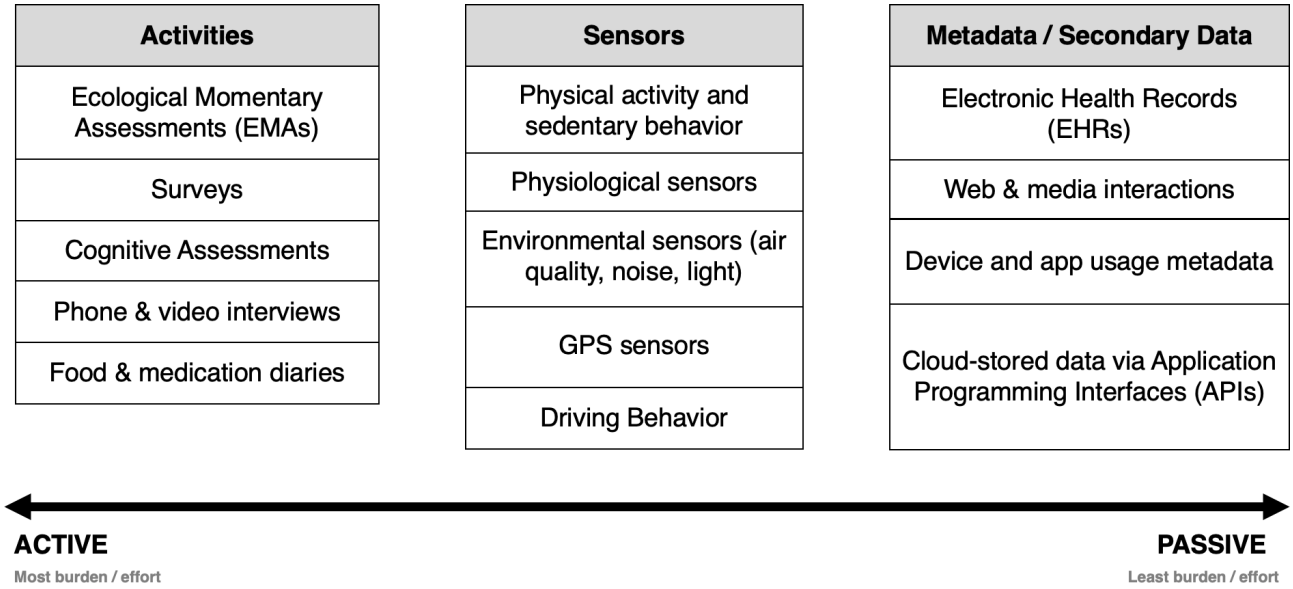

In general, the mechanisms driving willingness to contribute various data, irrespective of specific study demands, sponsors, and incentives [5] remain unclear. From the general patterns of attrition observed in longitudinal studies, it is apparent that a subset of participants may have a multitude of reasons for choosing to not engage in certain research activities.

This Study

The purpose of this study is to investigate the acceptability of active and passive data collection methods in mobile health research. This survey was designed to explore attitudes and opinions related to digital health technologies and willingness to share related data. The three research questions guiding this work are:

1. RQ1: What types of data are people willing to contribute?

XSL•FO
RenderX

2. RQ2: For how long are people willing to contribute critical data streams?
3. RQ3: What type of data control and feedback do people expect?

Methods

Participants

A total of 1509 adults were recruited from Prolific [8]. The Prolific platform allows researchers to post a 'task' with a short description for potential participants to review before beginning. In addition to the study background language, we the potential participants were informed: "This survey is designed to learn more about your attitudes and opinions related to digital health technologies and related data/data-sharing." Two participants did not consent to participate, and 18 did not complete the survey and were dropped from the study, resulting in a total of 1489 participants with a mean age of 35.5 (SD 11.97) years.

Ethical Considerations

This work was carried out in accordance with The Code of Ethics of the World Medical Association (Declaration of Helsinki). The Institutional Review Board (IRB) at the University of Central Florida approved the study protocol described in this manuscript (IRB No: STUDY00002281). The survey was pilot tested by members of the research laboratory conducting the study, for issues of usability and technical functioning. The IRB determined this study as exempt from regulation for human subjects research. Participants were provided an opportunity to accept or refuse consent before beginning this voluntary survey. The consent form indicated the expected time commitment, compensation, and nature of study. If consent was refused, the survey would close for that participant. Participants received US $2 after completing the survey on the Prolific Research Platform. All data was deidentified; the authors are accountable for all aspects of the work (including full data access, integrity of the data and the accuracy of the data analysis) in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Materials and Design

Participants completed a battery of questionnaires via the Qualtrics survey platform, designed to assess sample demographic characteristics (eg, technology ownership, mobility), opinions about data control, and willingness to contribute various data for research. These modules were designed to assess various facets of attitudes and behavior relevant to longitudinal observational research, to determine heterogeneity in willingness to use devices or share data.

Procedures

The participants completed a questionnaire via the Qualtrics survey platform, where the first page displayed the consent form and a radio button to select regarding consent. If participants consented, they were presented with modules on separate pages assessing: (1) demographic characteristics; (2) prior research experience; (3) mobility (ie, use of aids; driving frequency); (4) tech ownership (ie, own a tablet, smartphone, home Wi-Fi); (5) social media usage (ie, frequency of using Facebook, Twitter, Snapchat, Instagram, TikTok, WhatsApp); (6) willingness and extent of contributing various data (ie, across categories of activities, sensors, metadata); (7) opinions about data controls; and (8) willingness to use a robot for Instrumental Activities of Daily Living (IADLs). This survey totaled 52 questions over 12 pages presented in the same order for all participants (see Multimedia Appendix 1 for full survey). Questions for demographic reporting required responses, while all other items only suggested responses. Participants were only able to navigate forward and not backward through the survey.

Data Preparation and Analysis Plan

This study reports findings related to the willingness and extent of contributing various data and opinions about data controls. All data were processed using R software (version 4.0.2; R Foundation for Statistical Computing) [9], using the *tidyverse* package [10], including *ggplot2* for data visualization [11] and *dplyr* for data wrangling [12].

Results

Sample Characteristics

The sample was 44% female (n=652), with a mean age of 36 (SD 11.97, 20-84). A total of 99% (n=1479) of the individuals owned a mobile phone and 98% (n=1464) had access to home Wi-Fi (Table 1).

XSL•FO
RenderX

Table . Sample characteristics from online study.

Characteristics	Participants (N=1489)
Age (years), mean (SD; range)	36.5 (12.0; 20-84)
Age ≥50, years, n (%)	224 (15.04)
Past research experience, n (%)	1238 (83.14)
Sex, n (%)	
Male	811 (54.4)
Female	652 (43.7)
Other	19 (1.27)
Proportion Unspecified + No Report	4 (0.2)
Race/Ethnicity, n (%)	
Hispanic	116 (7.8)
White	1134 (76.15)
Black	156 (10.5)
Asian	193 (12.9)
American Indian, Alaska Native	22 (1.5)
Other	27 (1.8)
Average miles driven per week, mean (SD)	78 (181.9)
Median miles driven per week, IQR	29 (5-80)
Technology ownership, n (%)	
Tablet	929 (62.39)
Smartphone	1479 (99.3)
Activity tracker	660 (44.3)
Home Wi-Fi	1464 (98.3)
Home Assistant (eg, Alexa)	812 (54.5)
Awareness regarding Google Takeout	217 (14.6)

RQ1: What Data Are People Willing to Contribute?

Overview

For each category of data, the sections below report the most and least endorsed options, as summarized in Figure 2.

XSL•FO
RenderX

Figure 2. (A) Respondents' willingness to complete various activities in the context of research; (B) Respondents' willingness to wear various sensors in the context of research; (C) Respondents' willingness to share various metadata in the context of research.

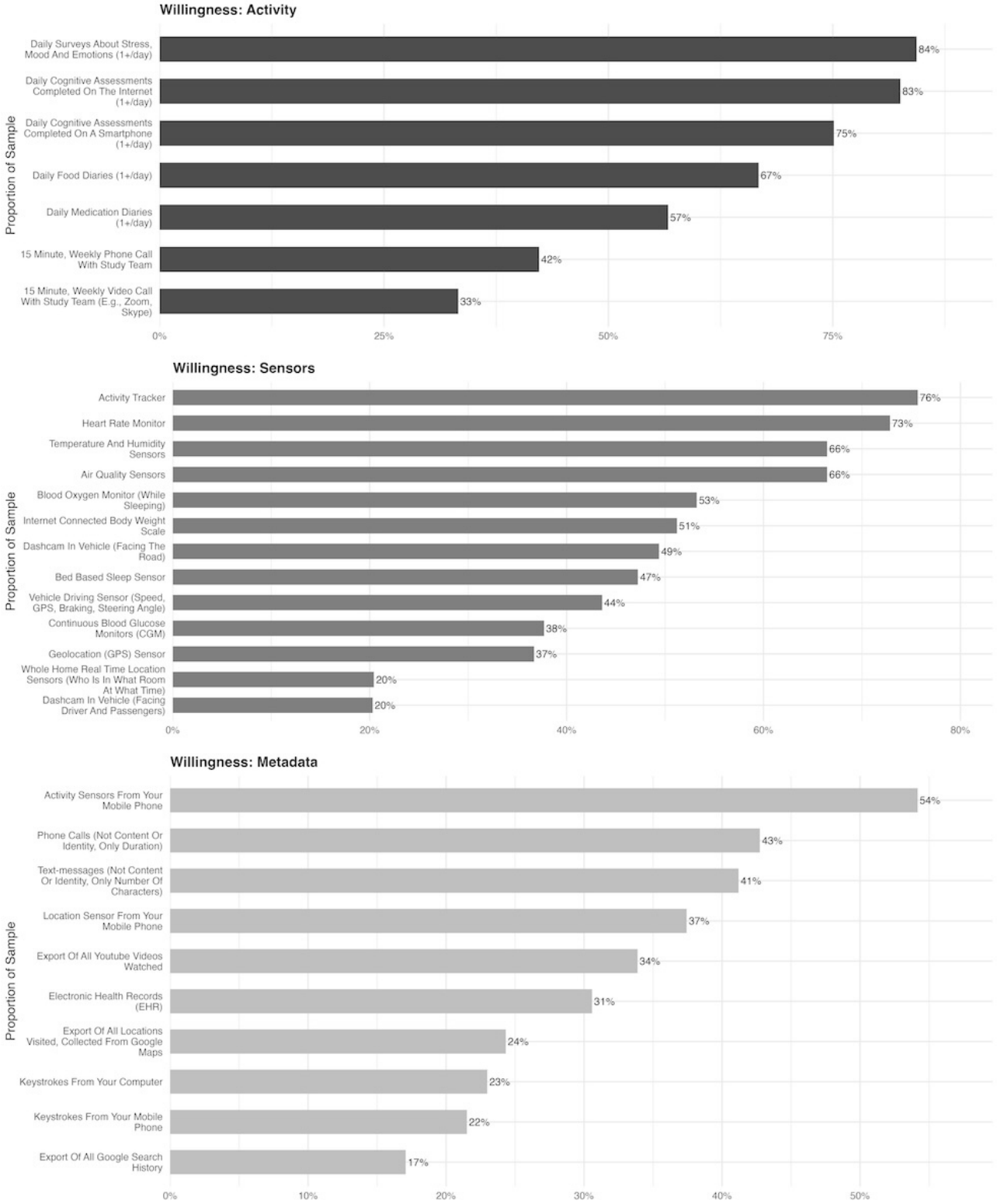

Activities

Among the activities assessed that could potentially take place in the context of research, 15-minute video or phone calls were the least endorsed options. Respondents were most willing to complete daily surveys about stress, mood, and emotions, as well as daily cognitive assessments.

Sensors

Of the sensors assessed that could be worn in the context of research, whole home real-time location sensors and in-cabin, driver-facing dashcams were the three least endorsed options. In contrast, the respondents were most willing to wear an activity tracker, a heart rate monitor, and air quality sensors.

XSL•FO
RenderX

Metadata

Of the metadata assessed that may be requested in the context of research, a Google Search history export as well as keystroke data (from computer or mobile phone) were the least endorsed. In contrast, the respondents were most willing to share data from smartphone activity trackers as well as phone and SMS metadata (number and duration of calls).

RQ2: For How Long Are People Willing to Contribute Critical Data Streams?

Participants were asked to consider a hypothetical study in which they were asked to complete three activities (queried independently), and rate how long in days they would be willing to complete each activity, for a maximum of 100 days (see Textbox 1 for full questions): (1) daily assessments of cognition; (2) wear and charge an air quality or pollution sensor; and (3) wear and charge a GPS location sensor.

Textbox 1. Questions posed to participants in regard to three data collection activities (ie, cognitive assessments, air quality sensing, GPS location sensing).

Question
Imagine a study, where you were asked to complete daily assessments of your cognition, via short "“brain games"” or "“mini-video games"”. For how many days might you be willing to complete these assessments (up to 100 days d)?
Imagine a study, where you were asked to wear an air pollution sensor daily, clipping it to a work bag or to your clothing. For how many days might you be willing to wear and charge this device (up to 100 days d)?
Imagine a study, where you were asked to wear a GPS location sensor daily, clipping it to a work bag or to your clothing. For how many days might you be willing to wear and charge this device (up to 100 days d)?

Paired sample *t* tests were conducted to assess the differential willingness to share data from these three critical data streams, each previously implicated in cognitive health research (ie, mobile cognitive assessments, location data (GPS), and ambulatory air quality monitoring). On average, people were willing to complete ambulatory cognitive assessments for 56.7 (SD 36.2) days, air quality monitoring for 58.1 (SD 37.7) days, and GPS location monitoring for 37 (SD 39.0) days. Comparing mobile cognitive assessments and air quality monitoring—participants were willing to complete each for a similar period of time (ie, mean difference of approximately 1.4 d; t_{1488}=−1.44; P=.14). Upon comparing mobile cognitive assessments and GPS location monitoring, participants were willing to complete cognitive assessments for close to 20 more days than GPS location monitoring (ie, mean difference of approximately 19.7 d; t_{1483}=−17.89; P<.001). Comparing mobile air quality monitoring and GPS location monitoring – participants were willing to complete air quality monitoring for close to 21 more days than GPS location monitoring (ie, mean difference of approximately 21 days; t_{1483}=−20.65; P<.001).

RQ3: What Do People Expect From Their Data?

Participants were asked: “Please select the types of control you expect to have over the data submitted as part of a research study” Among the data control and sharing aspects assessed, 1042 (70%) participants were most interested in the ability to delete all their contributed research data, followed by the ability to select specific data streams for deletion (65%, 960) and deleting a specific moment in time (44%, 653). Respondents were least interested in the ability to share data with their insurance providers and caregivers (n=453, 30% and n=384, 25%, respectively; Figure 3).

When asked “Please rate how comfortable (5-point scale, from extremely comfortable to extremely uncomfortable) you would feel about each of the following situations related to your hypothetical data,” participant responses varied by data usage scenario (see Figure 4): (1) data used only for the purpose of the original study; (2) data used to develop algorithms; (3) data used by outside institutions collaborating on this research; (4) data used indefinitely for purposes unknown today; (5) data used to identify the participant; (6) data used to solve a public health problem; (7) data used to develop therapies and diagnoses. Unsurprisingly, participants were mostly uncomfortable with reidentification and with data being used for unspecified future purposes.. Respondents were mixed in regard to data being used by external collaborations. Participants were mostly comfortable with data being used for the original purpose, to develop therapies and diagnoses, and for algorithm development.

XSL•FO
RenderX

Figure 3. Respondents' expectations for data controls as it relates to research.

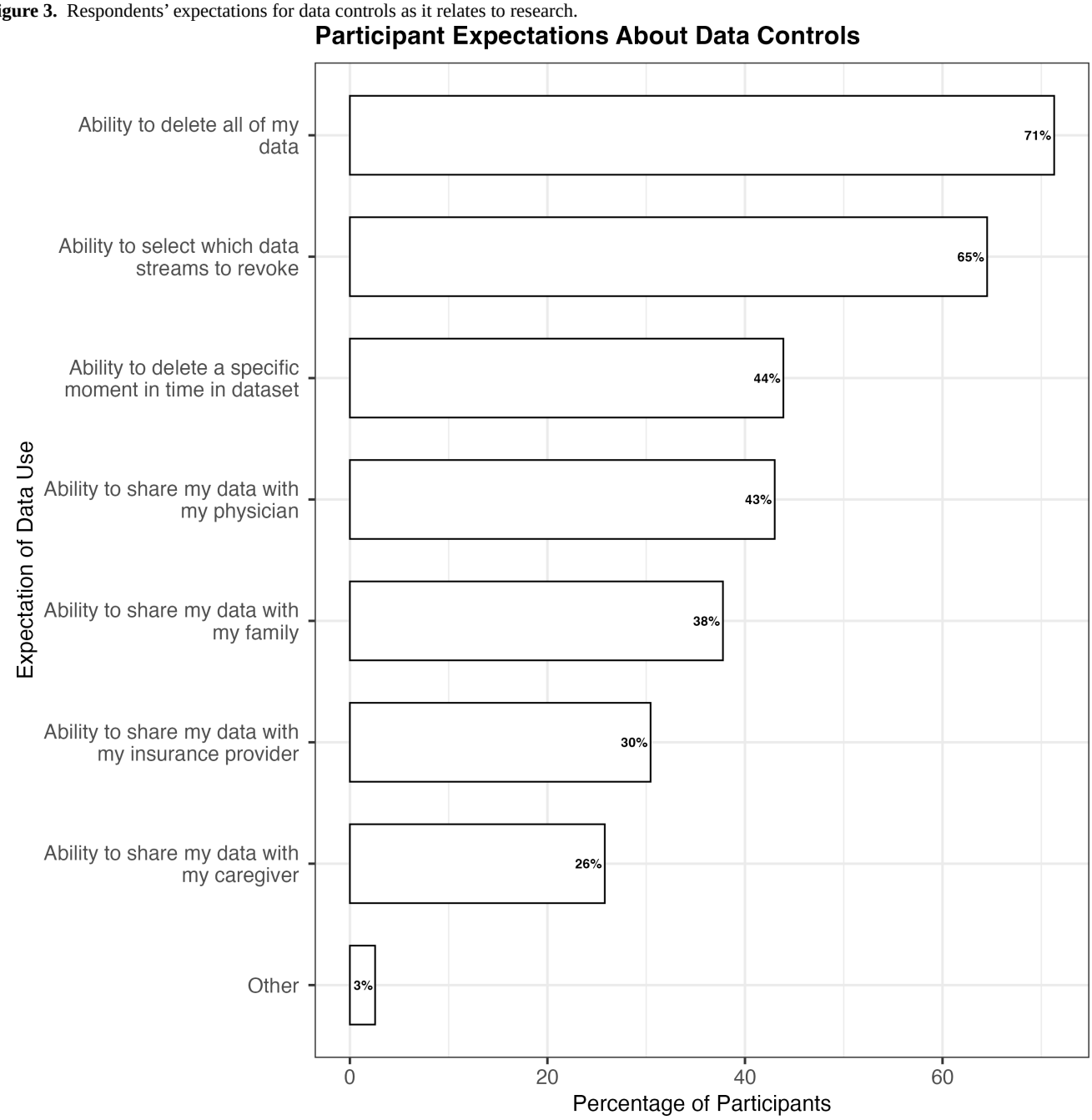

XSL•FO
RenderX

Figure 4. Proportion of sample respondents reporting their comfort levels with various aspects of data usage. The x-axis categorizes different data usage aspects. The y-axis represents the proportion of respondents. Responses range from "Extremely uncomfortable" to "Extremely comfortable," indicated by different shades of gray.

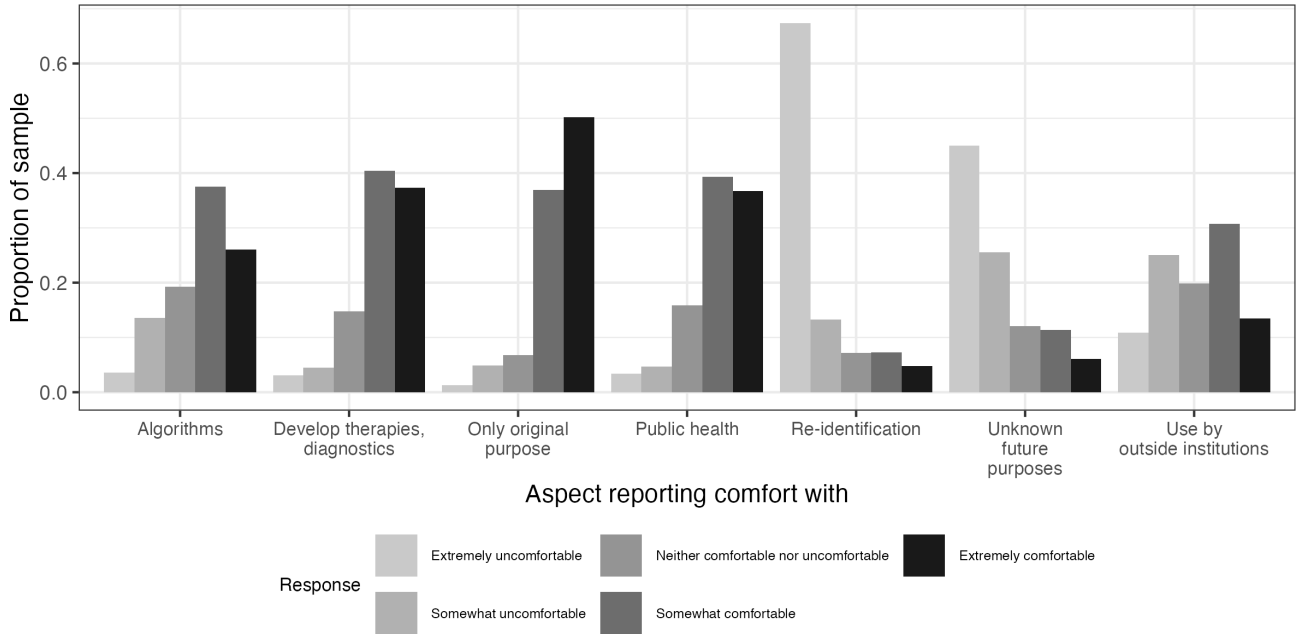

Discussion

Summary of the Findings

RQ1: What Data Are People Willing to Contribute?

Overall, participants were most willing to share data streams that have clear health implications (eg, activity sensor, heart rate monitors, air quality monitors) and least willing to share sensitive data streams (eg, GPS locations, driver/passenger-facing dashcam footage, Google search history). Participants were most willing to share data from active data collection activities and were least willing to share passive data streams (eg, sensors, metadata). People's willingness to share data that has obvious health implications replicates prior findings in this area [4] and underscores the necessity of education on the relevance of each data stream to health outcomes, if it is to be considered and adopted by research participants.

RQ2: How Long Are People Willing to Contribute Critical Data Streams?

On average, participants are willing to share data for up to 37 days, though this varied by data stream assessed. Willingness to share GPS location data was lowest, highlighting participant privacy concerns associated with data that may be more sensitive or identifiable, as was seen with the Strava data leak [6]). In contrast, daily cognitive assessments and ambulatory air quality monitoring, were similarly endorsed, further supporting engagement with data streams with clear health implications.

RQ3: What Controls and Feedback Do People Expect From Their Data?

Most participants endorsed the desire to delete all or specific streams of research data contributed by them. Although these data-related protections are required as per IRB regulations, most digital health technologies used for research do not necessarily have a straightforward method using which a person may delete their contributed data. This often requires a direct communication with the research group and a manual backend process. Future iterations of app- and cloud-based digital health technologies may want to consider affording the data provider (ie, the participant) more control over their data, including ways to download, remove data as desired, as well as mechanisms by which they can easily share data with key stakeholders in their life (eg, family or physician). Google Takeout is a platform through which Google account users can export all data originating from Google applications. In this study, only 14% individuals were aware of this tool, highlighting the need for data sharing education. Regarding various data usage scenarios, participants were mostly uncomfortable with unknown data use scenarios and reidentification, and comfortable with data being used to develop therapies and diagnoses, especially for the original purpose. Interestingly, participants reported mixed responses about their ratings of comfort, in regard to data being used by external collaborators. These findings highlight the importance of being transparent with participants on the various ways the data are being used—as not all data use scenarios are seemingly acceptable.

Limitations and Future Directions

Although a large sample (ie, of approximately 1500 respondents) was recruited for this study, a major limitation is that participant demographics indicate a recruitment bias toward midlife, highly educated (69% have associates degree or higher), White participants (76%, n = 1134), and individuals active on online research platforms. As participation required familiarity with digital interfaces and online survey platforms, the sample may not adequately reflect the views of individuals with low technology adoption rates—particularly those in rural or socioeconomically disadvantaged communities [13]. Future research should consider probing similar research questions, leveraging a representative and more diverse sample of

XSL•FO
RenderX

respondents, especially considering past research has shown differential willingness to share various data as a function of education and minority status. The extent to which older adults (ie, >60 years) are aware of and are willing to participate by sharing sensor data and metadata is unclear. Older adults were sparsely represented in this sample, limiting insight into the acceptability of sensor-based monitoring among individuals most likely to benefit from digital health interventions (eg, including those aimed at chronic disease mitigation or cognitive decline). Age-related factors such as cognitive load, sensory impairments, or comfort with technology may interact with perceived burden; however, these nuances were not explored in this study.

In addition, this study did not directly measure aspects of personality, or expectations for data privacy—two factors likely to explain variance in willingness to share data across the spectrum of participant burden. Participants' decisions may have been influenced by recent media exposure or personal experiences related to data breaches or algorithmic discrimination, but these factors were not assessed. It is expected that awareness of specific events that highlight the reidentification potential of various sensitive data (eg, GPS) may moderate willingness to share sensitive data, along with the personality constructs of openness and consciousness [14]).

Conclusions

Importantly, the study relies on self-reported, hypothetical willingness to share data, which may not correspond with real-world behavior, particularly when actual study conditions involve real data collection amidst daily life schedules and hassles. Future work should validate these findings in live digital health trials to assess whether stated willingness translates to actual participation and data-sharing compliance. This study aimed to elucidate the types of data and control measures that participants expect in research. These findings can guide the optimization of study protocols to enhance participant engagement, data quality, and minimize participant burden. Ensuring participant privacy is paramount, and researchers have a duty to transparently communicate the rationale for data collection and its potential benefits. As researchers intend to collect rich data across the Active-Passive data continuum (see Figure 1), our participant's privacy comes first, and it is our duty to inform them why we collect the data we do. This approach will foster trust and encourage greater participation in research involving potentially sensitive data streams.

Acknowledgments

NR was previously affiliated with the University of Central Florida and is now at Penn State University. This work was supported in part by funding from a UCF startup.

Data Availability

The data that support the findings of this study are available in a public repository on GitHub [15]. The repository includes deidentified datasets, code for data preprocessing and analysis, and documentation to facilitate reuse.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Data for the willingness to participate, and share various streams.
[DOCX File, 26 KB - formative_v9i1e64082_app1.docx]

References

1. Revilla M, Couper MP, Ochoa C. Willingness of online panelists to perform additional tasks. Methods, Data, Analyses: A Journal for Quantitative Methods and Survey Methodology (Mda) 2019;13(2):223-252. [doi: 10.12758/mda.2018.01]
2. Wenz A, Jäckle A, Couper MP. Willingness to use mobile technologies for data collection in a probability household panel. Surv Res Methods 2019;13(1):1-22. [doi: 10.18148/srm/2019.v1i1.7298]
3. Struminskaya B, Lugtig P, Toepoel V, Schouten B, Giesen D, Dolmans R. Sharing data collected with smartphone sensors: willingness, participation, and nonparticipation bias. Public Opin Q 2021;85(Suppl 1):423-462. [doi: 10.1093/poq/nfab025] [Medline: 34602867]
4. Seltzer E, Goldshear J, Guntuku SC, et al. Patients' willingness to share digital health and non-health data for research: a cross-sectional study. BMC Med Inform Decis Mak 2019 Aug 8;19(1):157. [doi: 10.1186/s12911-019-0886-9] [Medline: 31395102]
5. Keusch F, Wenz A, Conrad F. Do you have your smartphone with you? Behavioral barriers for measuring everyday activities with smartphone sensors. Comput Human Behav 2022 Feb;127:107054. [doi: 10.1016/j.chb.2021.107054]
6. Hsu J. The strava heat map shows even militaries can't keep secrets from social data. Wired. URL: https://www.wired.com/story/strava-heat-map-military-bases-fitness-trackers-privacy/ [accessed 2025-04-21]
7. Hysong SJ, Khan MM, Petersen LA. Passive monitoring versus active assessment of clinical performance: impact on measured quality of care. Med Care 2011 Oct;49(10):883-890. [doi: 10.1097/MLR.0b013e318222a36c] [Medline: 21918399]

XSL•FO
RenderX

8. Quality data from real people for faster breakthroughs. Prolific. URL: www.prolific.co [accessed 2025-07-09]
9. The R Project for Statistical Computing. The R Foundation. 2021. URL: https://www.R-project.org/ [accessed 2025-07-09]
10. Wickham H, Averick M, Bryan J, et al. Welcome to the Tidyverse. JOSS 2019;4(43):1686. [doi: 10.21105/joss.01686]
11. Wickham H. ggplot2. WIREs Computational Stats 2011 Mar;3(2):180-185. [doi: 10.1002/wics.147]
12. Wickham H, Francois R, Henry L, Müller K, Dplyr. Dplyr: A Grammar of Data Manipulation. R Package Version 0.8.4.. : R project; 2014 URL: https://CRAN.R-project.org/package=dplyr [accessed 2020-08-12]
13. Pethig F, Kroenung J, Noeltner M. A stigma power perspective on digital government service avoidance. Gov Inf Q 2021 Apr;38(2):101545. [doi: 10.1016/j.giq.2020.101545]
14. Linek SB, Fecher B, Friesike S, Hebing M. Data sharing as social dilemma: Influence of the researcher's personality. PLOS ONE 2017;12(8):e0183216. [doi: 10.1371/journal.pone.0183216] [Medline: 28817642]
15. Nelsonroque/cascadelab_apdcm_mhealth. GitHub. URL: https://github.com/nelsonroque/cascadelab_apdcm_mhealth [accessed 2025-08-19]

Abbreviations

IRB: institutional review board

Edited by A Mavragani; submitted 09.07.24; peer-reviewed by C Wu, Z Zandesh; revised version received 30.05.25; accepted 02.06.25; published 12.09.25.

Please cite as:
Roque N, Felt J
Acceptability of Active and Passive Data Collection Methods for Mobile Health Research: Cross-Sectional Survey of an Online Adult Sample in the United States
JMIR Form Res 2025;9:e64082
URL: https://formative.jmir.org/2025/1/e64082
doi:10.2196/64082

© Nelson Roque, John Felt. Originally published in JMIR Formative Research (https://formative.jmir.org), 12.9.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Concordance Between Survey and Electronic Health Record Data in the COVID-19 Citizen Science Study: Retrospective Cohort Analysis

Elizabeth Crull[1], MPH; Emily C O'Brien[2], PhD; Pavel Antiperovitch[3], MD; Kirubel Asfaw[2], MS; Alexis L Beatty[4], MD, MAS; Djeneba Audrey Djibo[5], PhD; Alan F Kaul[6], PharmD, MBA; John Kornak[3], PhD; Gregory M Marcus[4], MD, MAS; Madelaine Faulkner Modrow[3], MPH; Jeffrey E Olgin[4], MD; Jaime Orozco[3], BA; Soo Park[3], BA; Noah Peyser[4], PhD; Mark J Pletcher[3], MD, MPH; Thomas W Carton[1], MS, PhD

[1]Department of Health Services Research, Louisiana Public Health Institute, 400 Poydras Street, Suite 1250, New Orleans, LA, United States

[2]Duke Clinical Research Institute, School of Medicine, Duke University, Durham, NC, United States

[3]Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, CA, United States

[4]Division of Cardiology, University of California, San Francisco, San Francisco, CA, United States

[5]Safety, Surveillance, and Collaboration, CVS Health, Blue Bell, PA, United States

[6]Medical Outcomes Management, Inc., Sharon, MA, United States

Corresponding Author:
Elizabeth Crull, MPH
Department of Health Services Research, Louisiana Public Health Institute, 400 Poydras Street, Suite 1250, New Orleans, LA, United States

Abstract

Background: Real-world data reported by patients and extracted from electronic health records (EHRs) are increasingly leveraged for research, policy, and clinical decision-making. However, it is not always obvious the extent to which these 2 data sources agree with each other.

Objective: This study aimed to evaluate the concordance of variables reported by participants enrolled in an electronic cohort study and data available in their EHRs.

Methods: Survey data from COVID-19 Citizen Science, an electronic cohort study, were linked to EHR data from 7 health systems, comprising 34,908 participants. Concordance was evaluated for demographics, chronic conditions, and COVID-19 characteristics. Overall agreement, sensitivity, specificity, positive predictive value, negative predictive value, and κ statistics with 95% CIs were calculated.

Results: Of 34,017 participants with complete information, 62.3% (21,176/34,017) reported being female, and 62.4% (21,217/34,017) were female according to EHR data. The median age was 57 (IQR 42 - 68) years. Out of 34,017 participants, 81.6% (27,744/34,017) of participants reported being White, and 79.5% (27,054/34,017) were White according to EHR data. In addition, 9.2% (3,124/34,017) of participants reported being Hispanic, and 6.6% (2,249/34,017) were Hispanic according to EHR data. Statistically significant discordance between data sources was detected for all demographic characteristics (P<.05) except the female category (P=.57) and the American Indian and Alaska Native (P=.21) and "other" race categories (P=.33). Statistically significant discordance was detected for the 2 COVID-19 traits and all baseline medical conditions except diabetes (P=.17). The starkest absolute difference between data sources was for COVID-19 vaccination, which was 48.4% according to the EHR and 97.4% according to participant report. Overall agreement was high for all demographic characteristics, although chance-corrected agreement (κ) and sensitivity were lower for the "other" race category (κ=0.31, sensitivity =26.6%), Hispanic ethnicity (κ=0.82, sensitivity=74%), and current smoker status (κ=0.54, sensitivity=49.4%). Specificity and negative predictive value (NPV) were higher than corresponding specificity and positive predictive value (PPV) for all baseline medical conditions. Sleep apnea had the highest sensitivity of all medical conditions (83.5%), and anemia had the lowest (32.8%). Chance-corrected agreement (κ) was highly variable for baseline medical conditions, ranging from 0.26 for anemia to 0.71 for diabetes. Overall and chance-corrected agreement between data sources for COVID-19 traits such as infection (84.6%, κ=0.34) and vaccination (51.0%, κ=0.05) was relatively lower than all other evaluated traits. The sensitivity for COVID-19 infection was 32.2%, and the sensitivity for COVID-19 vaccination was 49.7%. Although PPV for COVID-19 vaccination was 99.9%, the NPV was 5%.

Conclusions: Results suggest the need for improvements to point-of-care capture of patient demographic traits and COVID-19 infection and vaccination history, patient education about their medical conditions, and linkage to external data sources in EHR-only

XSL•FO
RenderX

pragmatic research. Further, these results indicate that additional work is required to integrate and prioritize participant-reported data in pragmatic research.

Trial Registration: ClinicalTrials.gov NCT05548803; https://clinicaltrials.gov/study/NCT05548803

(JMIR Form Res 2025;9:e58097) doi:10.2196/58097

KEYWORDS

electronic health records; self-report; COVID-19; data accuracy; data validation; EHR; cohort; cohort analysis; real-world data; concordance; internet-based; portal; participant; report; reported

Introduction

The advent of electronic health record (EHR) systems and internet-based study portals have modernized and streamlined pragmatic clinical research [1,2], defined as research that can be conducted in real-world settings with minimal change to clinical operation [3,4]. The use of patient- and study participant–reported data alongside EHR data is increasingly common in research and clinical practice to complement and validate EHR data sources [5,6]. Therefore, there is potential value in linking and comparing patient experience and outcomes data gathered from mailed surveys [7-9] and patient-facing web-based portals [10] with data extracted from EHRs. This is especially true for clinical concepts that are notoriously difficult to qualify using medical coding alone, such as mood, gastrointestinal disorders, and chronic pain [11].

EHR and participant-reported data each have significant limitations. EHR data are fraught with administrative error, incomplete mapping to clinical ontologies, lack of legacy health record data, and inability to extract important clinical information from unstructured physician notes [12,13]. Participant-reported data are subject to bias from social desirability [10], fatigue [14], and limited understanding of medical issues. How these limitations affect the reliability of different kinds of health-related information is of interest for this substudy. There is a particular need for understanding the reliability of health information related to COVID-19, especially because a large portion of home testing and vaccination occurred outside traditional health systems.

The COVID-19 Citizen Science Study (CCS) is a longitudinal digital cohort study designed to generate knowledge about participant-reported outcomes related to the COVID-19 pandemic [15]. The study linked participant-reported data with their corresponding EHR data, thus presenting an opportunity to analyze the concordance between these data sources. The purpose of this study was to assess the concordance of COVID-19–related outcomes, demographic characteristics, smoker status, and 12 common medical conditions.

Methods

Overview

Our study evaluated concordance between two data sources: (1) participant-reported data from a web-based patient portal for the CCS study and (2) participants' corresponding EHRs, conforming to a common data model. Data from the participant-reported data source were converted first to a CSV format and then imported into SAS 9.4 software (SAS Institute) as datasets. Data from the participants' EHRs were loaded along an extract-transform-load pipeline from the source EHR to a relational database management system and finally into SAS 9.4 software as well to enable comparison with participant-reported data.

Study Recruitment

Our concordance assessment used participant-reported and EHR data collected as part of the CCS study (ClinicalTrials.gov identifier NCT5548803), which has been described in detail previously [15]. Participants were recruited from 7 major health systems in Texas, Louisiana, Mississippi, California, Utah, and New York that participate in the National Patient-Centered Clinical Research Network (PCORnet). Patients were eligible to join if they were 18 years or older and had at least one clinical encounter after January 1, 2019. Recruitment lasted from November 2020 to February 2022.

Participant-Reported Data

Upon enrolling, participants were asked to respond to baseline surveys on demographics, smoking history, and medical conditions. Participants were then administered follow-up surveys about exposure to, diagnosis of, and vaccination against COVID-19, among other questions seeking to understand both individual experience and population-level trends related to the pandemic. These surveys were housed in the Eureka research platform (University of California San Francisco, with funding from the National Institutes of Health) [16], which had web browser and smartphone functionality.

EHR Data

For consenting and authorizing participants, EHR-limited datasets in the PCORnet Common Data Model format were extracted from the site-specific DataMarts maintained by all participating health systems [17]. The CCS study data extraction query was developed by Duke University programmers using SAS 9.4 software and distributed to all sites to run in their local environments against their DataMart. The query extracted clinical data with a 5-year lookback from the recruitment start date through the most recently available data. Sensitive diagnoses were filtered out, and only a minimum necessary subset of laboratory and medication records was extracted. Only patients for whom identities were algorithmically matched or manually verified were included in the final analytic cohort.

Concordance Definitions

Among 34,908 participants where linkage was possible, we evaluated concordance in the following domains: demographics, baseline medical conditions, current smoker status, COVID-19

XSL•FO
RenderX

diagnosis, and COVID-19 vaccination. We chose variables that were conceptually similar between the participant-reported and EHR sources (Multimedia Appendices 1 and 2). Sex in both sources was defined as sex assigned at birth. Although gender identity was available from survey data, it was not available in EHR data and thus was not an eligible variable for concordance analysis. Race and ethnicity data abstracted from EHR data were populated according to health system practices. Race and ethnicity data abstracted from survey data were reported directly by study participants.

To promote comparability, measurement periods were aligned between data sources. Participants who had missing data in one or both sources were not considered for concordance analyses. Age data were not analyzed for concordance because a birthdate match between sources was a requirement for data to be considered for EHR data extraction; thus, discordant scenarios were inherently filtered out before analysis for this substudy.

For demographic, smoker status, and COVID-19 characteristics, the participant report was considered the criterion standard. For medical conditions, the EHR was considered the criterion standard.

Statistical Approach

To test for marginal homogeneity between data sources, McNemar tests for paired nominal data were run on all 23 attributes, structured as dichotomous 2×2 contingency tables. Chi-square statistics and *P* values were calculated. A Bonferroni correction was applied to account for multiple comparisons, adjusting the significance threshold to .002 (.05/23). *P* values less than .001 were reported as $P<.001$.

For all domains, the following statistics were generated along with their 95% CI values: overall agreement (or overall accuracy), sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and Cohen κ. We used the following ranges for Cohen κ to describe observed agreement: strong (0.81 - 1.00), good (0.61 - 0.80), moderate (0.41 - 0.60), fair (0.21 - 0.40), poor (0.01 - 0.20), and no agreement (<0) [18,19]. However, these ranges are to provide a guide, and the adequacy of the specific level of agreement should be considered specifically to the domain under consideration and the application to which it will be used.

Data were analyzed from December 2022 to July 2023 using SAS 9.4 software.

Ethical Considerations

The CCS study, the protocol for which covered data analysis activities conducted for this substudy, was approved by the Western Institutional Review Board on November 5, 2020. Participants were informed of their right to withdraw at any time without any consequences. Digital informed consent was obtained from participants before surveys were collected. To maintain confidentiality, participants were not asked to provide their name, only their pre-assigned unique code to enable linkage to EHRs. Participants were not compensated for their time spent completing the surveys.

Results

A total of 39,526 patients enrolled in the study across 7 sites. After the exclusion of participants whose identity could not be verified and participants with partially or completely missing EHR data, 34,908 participants were included in the final analytic cohort (Figure 1). Descriptive statistics of the 34,017 participants who responded to the baseline demographic survey and results from the McNemar test for marginal homogeneity between data sources are summarized in Table 1. The median age of the sample was 57 (IQR 42 - 68, range 18 - 100) years according to the EHR. The sample was predominantly female and White according to both sources. The sample was classified as 6.6% (2,249/34,017) Hispanic in the EHR and 9.2% (3,124/34,017) Hispanic according to participant-reported data ($P<.001$).

XSL•FO
RenderX

Figure 1. Enrollment diagram and final analytic cohort for the COVID-19 Citizen Science Study concordance substudy. CCS: COVID-19 Citizen Science Study; CDM: common data model.

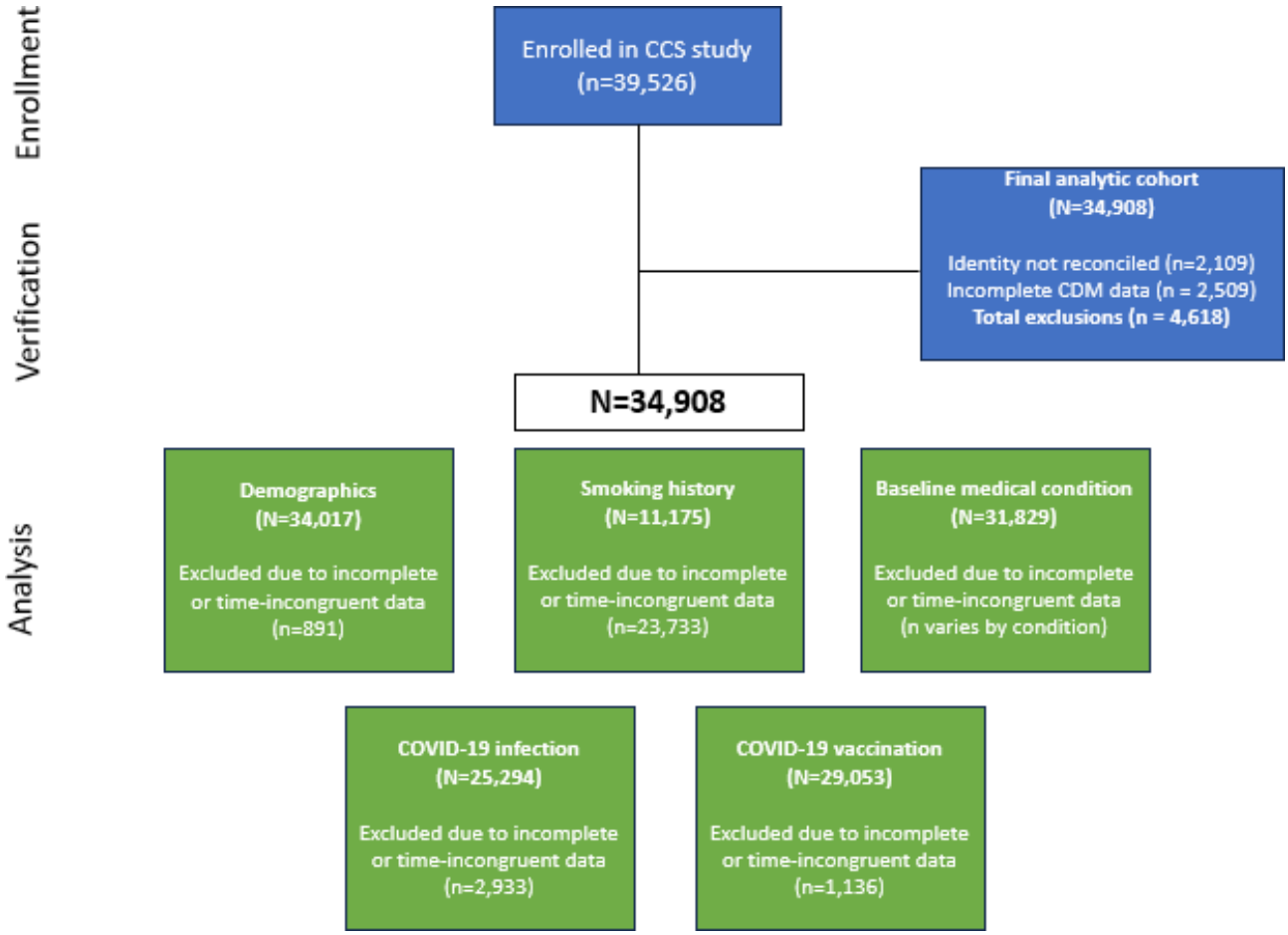

XSL•FO
RenderX

Table . Comparison of demographic information derived from electronic health records and participant self-reports.

Variable	EHR[a] (N=34,017)	Participant report (N=34,017)	*P* value[b]
Age (years)[c]			
Mean (SD)	54.7 (16.1)	—[d]	—
Median (IQR), range	57 (42 - 68), 18-100	—	—
Sex, n (%)			
Female	21,217 (62.4)	21,176 (62.3)	.57
Male	12,780 (37.6)	12,742 (37.5)	
Refused or missing	20 (<1)	99 (<1)	
Race, n (%)			
American Indian or Alaska Native	104 (<1)	91 (<1)	.21
Asian, Native Hawaiian, or Pacific Islander	1797 (5.3)	2042 (6.0)	<.001
Black or African American	1286 (3.8)	1344 (4.0)	.001
White	27,054 (79.5)	27,744 (81.6)	<.001
Multiple races[e]	93 (<1)	1269 (3.7)	<.001
Other[f]	1039 (3.1)	1077 (3.2)	.33
Refused or missing	2644 (7.8)	450 (1.3)	<.001
Ethnicity, n (%)			
Hispanic	2249 (6.6)	3124 (9.2)	<.001
Non-Hispanic	28,903 (85.0)	30,528 (89.7)	
Refused or missing	2865 (8.4)	365 (1.1)	

[a]EHR: electronic health record.

[b]*P* values calculated by McNemar test.

[c]Age data were not analyzed for concordance because a date of birth match between sources was a requirement for data to be considered for EHR data extraction; thus, discordant scenarios were inherently filtered out before analysis for this substudy

[d]Not available.

[e]The category “Multiple Races” is a mapped value in the PCORnet Common Data Model, and no further detail was available. In participant-reported data, “Multiple Races” was defined as participants who responded to 2 or more non-missing race categories.

[f]In both data sources, there were no additional details for the “Other” categorization.

Statistically significant differences between the 2 data sources were detected for all characteristics except for the female category (*P*=.57), the American Indian or Alaska Native race category (*P*=.21), the “other” race category (*P*=.33), and diabetes (*P*=.17). The starkest absolute difference between data sources was for COVID-19 vaccination, in which 97.4% (28,291/29,053) of participants self-reported a vaccine while only 48.4% (14,076/29,053) had this documented in the EHR (Table 2).

XSL•FO
RenderX

Table . Comparison of medical conditions and COVID-19 history as reported in electronic health records and participant self-reports.

Variables	Participants, n	EHR[a], n (%)	Participant report, n (%)	*P* value[b]
Smoking status	11,175	947 (8.5)	1396 (12.5)	<.001
Medical conditions				
Diabetes	31,744	3033 (9.6)	2979 (9.4)	.17
Hypertension	31,636	9440 (29.8)	10,953 (34.6)	<.001
Coronary artery disease or angina	31,573	2440 (7.7)	1941 (6.1)	<.001
Myocardial infarction	31,721	288 (0.9)	792 (2.5)	<.001
Congestive heart failure	31,686	689 (2.2)	545 (1.7)	<.001
Transient ischemic attack	31,659	533 (1.7)	945 (3)	<.001
Atrial fibrillation or flutter	31,495	1363 (4.3)	1769 (5.6)	<.001
Sleep apnea	31,075	2838 (9.1)	4902 (15.8)	<.001
COPD[c]	31,626	1469 (4.6)	4902 (3.5)	<.001
Asthma	31,738	3481 (11)	3150 (9.9)	<.001
Immunodeficiency	31,378	730 (2.3)	1611 (5.1)	<.001
Anemia	31,622	3838 (12.1)	3427 (10.8)	<.001
COVID-19				
Infection	25,294	2319 (9.2)	4446 (17.6)	<.001
Vaccination	29,053	14,076 (48.4)	28,291 (97.4)	<.001

[a]EHR: electronic health record.

[b]*P* values calculated by McNemar test.

[c]COPD: chronic obstructive pulmonary disease.

Agreement between EHR and participant-reported characteristics according to 5 proportionate measures (overall agreement, sensitivity, specificity, PPV, and NPV) and one statistic of interrater reliability (Cohen κ) are shown in Table 3 and Multimedia Appendix 3. Overall agreement was above 95% for all demographic characteristics, where the participant report was considered the criterion standard. Chance-corrected agreement (κ) was strong for most demographic characteristics except the “other” race category (κ=0.31) and current smoker status (κ=0.54). Sensitivity was 74% for the Hispanic characteristic, which translates to a relatively higher number of false negatives compared to other racial groups, 49.4% for current smoker status, and 26.6% for the “other” race category.

XSL•FO
RenderX

Table . Agreement between electronic health record– and participant-reported characteristics: overall agreement, κ statistic, and accuracy metrics.

Variable	Overall agreement, % (95% CI)	κ Statistic, % (95% CI)	Sensitivity, % (95% CI)	Specificity, % (95% CI)	PPV[a], % (95% CI)	NPV[b], % (95% CI)
Female	99.6 (99.5 - 99.7)	0.99 (0.99 - 0.99)	99.6 (99.5 - 99.7)	99.4 (99.3 - 99.6)	99.7 (99.6 - 99.7)	99.4 (99.2 - 99.5)
Non-Hispanic						
Asian American and Pacific Islander	99.2 (99.1 - 99.3)	0.93 (0.92 - 0.94)	92.0 (90.7 - 93.3)	99.6 (99.6 - 99.7)	93.9 (92.8 - 95.1)	99.5 (99.4 - 99.6)
Black	99.6 (99.5 - 99.7)	0.94 (0.93 - 0.95)	97.2 (96.2 - 98.2)	99.7 (99.6 - 99.7)	91.6 (90.0 - 93.2)	99.9 (99.9 - 99.9)
White	96.1 (95.9 - 99.3)	0.87 (0.86 - 0.88)	99.2 (99.0 - 99.3)	83.3 (82.3 - 84.3)	96.2 (95.9 - 96.4)	95.9 (95.3 - 96.4)
Other	96.9 (96.7 - 97.1)	0.31 (0.27 - 0.36)	26.6 (23.6 - 29.6)	99.0 (98.8 - 99.1)	42.7 (38.4 - 47.0)	97.9 (97.7 - 98.1)
Hispanic	97.9 (97.7 - 98.0)	0.82 (0.81 - 0.83)	74.0 (72.1 - 75.9)	99.7 (99.6 - 99.8)	94.9 (93.8 - 96.0)	98.1 (97.9 - 98.2)
Current smoker	91.4 (90.9 - 91.9)	0.54 (0.52 - 0.57)	49.4 (46.8 - 52.1)	97.4 (97.1 - 97.7)	72.9 (70.0 - 75.7)	93.1 (92.6 - 93.6)
Diabetes	95.1 (94.9 - 95.3)	0.71 (0.70 - 0.73)	73.5 (71.9 - 75.1)	97.4 (97.2 - 97.6)	74.8 (73.3 - 76.4)	97.2 (97.0 - 97.4)
Hypertension	85.0 (84.6 - 85.4)	0.66 (0.64 - 0.68)	82.9 (82.1 - 83.6)	85.9 (85.4 - 86.4)	71.4 (70.6 - 72.3)	92.2 (91.8 - 92.6)
Coronary artery disease/angina	94.0 (93.7 - 94.3)	0.54 (0.52 - 0.56)	51.0 (49.0 - 53.0)	97.6 (97.4 - 97.8)	64.1 (62.0 - 66.2)	96.0 (95.7 - 96.2)
Myocardial infarction	97.6 (97.5 - 97.8)	0.30 (0.25 - 0.35)	57.6 (51.9 - 63.3)	98.0 (97.9 - 98.2)	21.0 (18.1 - 23.8)	99.6 (99.5 - 99.7)
Congestive heart failure	98.0 (97.9 - 98.2)	0.48 (0.44 - 0.52)	44.1 (40.4 - 47.8)	99.2 (99.1 - 99.3)	55.8 (51.6 - 60.0)	98.8 (98.6 - 98.9)
Transient ischemic attack	97.6 (97.4 - 97.7)	0.47 (0.43 - 0.50)	66.2 (62.2 - 70.2)	98.1 (97.9 - 98.3)	37.4 (34.3 - 40.4)	99.4 (99.3 - 99.5)
Atrial fibrillation	97.0 (96.8 - 97.2)	0.68 (0.66 - 0.70)	80.3 (78.2 - 82.4)	97.8 (97.6 - 98.0)	61.8 (59.6 - 64.1)	99.1 (99.0 - 99.2)
Sleep apnea	90.4 (90.0 - 90.7)	0.56 (0.55 - 0.58)	83.5 (82.2 - 84.9)	91.0 (90.7 - 91.4)	48.4 (47.0 - 49.8)	98.2 (98.1 - 98.4)
COPD[c]	95.2 (95.0 - 95.5)	0.39 (0.36 - 0.42)	36.7 (34.2 - 39.2)	98.1 (97.9 - 98.3)	48.4 (45.5 - 51.4)	97.0 (96.8 - 97.1)
Asthma	90.2 (89.9 - 90.6)	0.48 (0.46 - 0.50)	50.8 (49.1 - 52.5)	95.1 (94.9 - 95.4)	56.1 (54.4 - 57.9)	94.0 (93.7 - 94.3)
Immunodeficiency	94.6 (94.4 - 94.9)	0.26 (0.22 - 0.29)	45.1 (41.5 - 48.7)	95.8 (95.6 - 96.0)	20.4 (18.5 - 22.4)	98.7 (98.5 - 98.8)
Anemia	85.0 (84.6 - 85.4)	0.26 (0.24 - 0.28)	32.8 (31.3 - 34.3)	92.2 (91.9 - 92.5)	36.7 (35.1 - 38.3)	90.8 (90.5 - 91.2)
COVID-19 infection	84.6 (84.1 - 85.0)	0.34 (0.33 - 0.36)	32.2 (30.9 - 33.6)	95.8 (95.5 - 96.0)	61.8 (59.8 - 63.8)	86.9 (86.5 - 87.3)
COVID-19 vaccination	51.0 (50.4 - 51.6)	0.05 (0.04 - 0.06)	49.7 (49.1 - 50.3)	98.2 (97.2 - 99.1)	99.9 (99.9 - 99.9)	5.0 (4.6 - 5.3)

[a]PPV: positive predictive value.

[b]NPV: negative predictive value.

[c]COPD: chronic obstructive pulmonary disease.

The criterion standard for baseline medical conditions was the EHR. Overall agreement, specificity, and NPV between data sources was above 85% for all baseline medical conditions, although there was heterogeneity in sensitivity, PPV, and chance-corrected agreement (κ). Sensitivity ranged from 32.8 (95% CI 31.3 - 34.3) to 83.5% (95% CI 82.2 - 84.9), being lowest for anemia and highest for sleep apnea. PPV ranged from 20.4% (95% CI 18.4 - 22.4) to 74.8% (95% CI 73.3 - 76.4), being lowest for immunodeficiency and highest for diabetes. Finally, chance-corrected agreement (κ) was good for 3 of 12 compared baseline medical conditions, moderate for 5, and fair for 4. Chance-corrected agreement (κ) ranged from 0.26 (95% CI 0.22 - 0.29) to 0.71 (95% CI 0.70 - 0.73) for non-HIV immunodeficiency and diabetes, respectively.

The criterion standard for COVID-19 variables was participant-reported data. While chance-corrected agreement was fair for COVID-19 infection (κ=0.34), it was poor for COVID-19 vaccination (κ=0.05). Of 25,294 participants whose COVID-19 infection data could be compared, there were 3899 cases of discordance, 77.3% (3013/3899) of which were classified as the participant reporting a COVID-19 diagnosis but this not being reflected in the EHR. Similarly, of the 29,053 participants whose COVID-19 vaccine data could be compared, there were 14,243 cases of discordance, 99.9% (14,229/14,243) of which were classified as the participant reporting a COVID-19 vaccine but this not being reflected in the EHR.

XSL•FO
RenderX

Discussion

Principal Results

We evaluated the concordance survey data from participants enrolled in a COVID-19 study and their linked EHR data. We had four main findings: (1) sensitivity and chance-corrected agreement were strong for all demographic characteristics except for Hispanic ethnicity, the "other" race category, and current smoker status, indicating that a relatively lower proportion of patients were correctly identified in the EHR as such in comparison to other traits; (2) when the EHR was the criterion standard, as we considered it to be for medical conditions, specificity and NPV were higher than corresponding sensitivity and PPV, suggesting that patients were better able to report in concordance with their EHR the absence of a medical condition rather than the presence of one; (3) chance-corrected agreement, sensitivity, and PPV varied widely for medical conditions, with no clear pattern emerging as to which types were more likely to be self-reported in concordance with the EHR; and (4) COVID-19 infection and vaccination had relatively low chance-corrected agreement and overall agreement compared to most demographic traits and many medical conditions, along with very low sensitivity, indicating that these health events are poorly captured in patients' EHRs.

Our findings suggest the need for improvements to point-of-care capture of patient demographic traits and COVID-19 infection and vaccination history, patient education about their medical conditions, and linkage to external data sources in EHR-only pragmatic research. Our results indicating specifically that the capture of Hispanic ethnicity and "other" race category is not as sensitive as other race categories demonstrates that current point-of-care processes for collecting racial and ethnic information from patients may be insufficient, regardless of whether such data points are captured by a clinician, administrative worker, or the patients themselves. It is critical to have granular and accurate capture of patient race and ethnicity to provide the most culturally sensitive clinical care. Similarly, our findings that captured COVID-19 health events were relatively discordant between data sources, suggesting an interruption of health data flow back to the EHR. This could originate from improper integration of COVID-19 testing and vaccination data sources, especially when these events happen outside of the health system, resulting in health care providers not having the most up-to-date information about the health status of their patients. Finally, the lower specificity and NPV of medical conditions when compared to their corresponding sensitivity and PPV suggest that (1) EHRs may not be capturing medical conditions, especially those that are pre-existing; and (2) patients may not be aware that they have certain medical conditions, most marked for conditions like anemia and COPD, both of which showed only about a third of patients reporting the presence of these conditions in concordance with their EHRs. Comprehensive and customized patient education and communication are suggested to support self-management of medical conditions and patient autonomy outside the point of care. Further, and for the sake of research integrity, those engaging in EHR-only research should make attempts to access and query as many views and tables as is feasible to properly categorize a patient as having or not having a medical condition.

Increasingly, novel research designs rely on the integration of multiple data sources to answer research questions. The COVID-19 pandemic accelerated already growing interest in real-world data use cases [20], including leveraging existing EHR data and bringing research directly to people through participant-facing portals. Direct-to-participant research has numerous benefits, including potential for greater geographic reach and diversity, lower participant burden with few or no in-person visits, and platforms that enable capture of relevant patient-centered endpoints [21]. These strengths complement those of EHR data, which, through national networks like PCORnet, can be standardized into research-grade data to facilitate rapid insights into key clinical outcomes [22]. To our knowledge, this is the first study to examine patterns of concordance across participant-reported and EHR data in the context of COVID-19.

Our finding that participants self-reported COVID-19 infection and vaccination at higher rates than what was evident in their clinical records illustrates the fragmented nature of real-world data. Ongoing work to enhance the quality and reliability of EHR data in the context of COVID-19, including network-level curation [22], linkage to external sources where appropriate (eg, state vaccine [23] or policy [24] databases), and systematic phenotype development and testing [25], is critical to maximizing the research value of these data. In parallel, the implementation of best practices to enhance the validity of participant reports, including stakeholder engagement in survey design, readability assessments, and cultural and linguistic adaptation, is essential to enhancing the reliability of findings from participant-facing research [26,27]. Our findings are broadly consistent with those from prior studies, suggesting that fitness-for-use depends on context [28,29]. We found that no single data source may be appropriate for EHR-based pragmatic research, consistent with prior work illustrating the potential biases that can arise in participant-reported data and how they vary [30-33].

Limitations

Several limitations to our study are worth noting. First, the CCS study comprises participants who were mostly White and female. Therefore, results may not generalize to broader populations. Second, diversity within minority communities can make both responding to survey questions and the identification of race challenging for patients and clinicians, respectively, a fact that may skew findings from our demographic analyses. Third, participant-reported COVID-19 variables were not validated and are subject to reporting and recall bias. Fourth, we observed some attrition in reporting over time, which could lead to selection bias in analyses of longitudinal outcomes. Finally, we used EHR data for this concordance analysis, which may not represent all medical encounters for a given participant and which may not be of the highest or most accurate quality. EHR data used in this analysis were not linked to claims data from pharmacies, which are a major administrator of COVID-19 vaccines. Particularly for outcomes that are generally observed

XSL•FO
RenderX

outside of the hospital, linkage to external data sources is likely warranted.

Conclusions

We found that the integration of multiple data sources to investigate COVID-19 research questions enhances the capture of key elements but also introduces opportunities for disagreement. Future studies that leverage linked data should evaluate the concordance of overlapping elements and report levels of agreement. Transparent reporting will contribute to a broader understanding of data reliability and relevance and support future strategies to improve fitness-for-use of real-world data.

Acknowledgments

This work was supported with funding from the Patient-Centered Outcomes Research Institute (PCORI; grant identification COVID-2020C2-10761). PCORI had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Disclaimer

The views and conclusions presented here are solely the responsibility of the authors and do not necessarily reflect the official views of PCORI.

Data Availability

The datasets generated and analyzed during this study are not publicly available to preserve patient privacy and confidentiality, but are available from the corresponding author on reasonable request.

Authors' Contributions

EC had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. TWC, EC, ECO’B, and MJP contributed to conception and design of the study. TWC, EC, ECO'B, and MJP assisted with acquisition, analysis, or interpretation of data. TWC, EC, and ECO’B handled drafting of the manuscript. EC contributed to statistical analysis. TWC and MJP obtained funding. KA, MFM, JO, and SP assisted with administrative, technical, or material support. TWC, ECO'B, and MJP contributed to supervision.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Description of concordance definitions by electronic health record and participant report.
[DOCX File, 23 KB - formative_v9i1e58097_app1.docx]

Multimedia Appendix 2
Clinical codes used in medical condition and COVID-19 phenotypes.
[DOCX File, 25 KB - formative_v9i1e58097_app2.docx]

Multimedia Appendix 3
2x2 contingency tables for all concordance domains.
[DOCX File, 39 KB - formative_v9i1e58097_app3.docx]

References

1. Cowie MR, Blomster JI, Curtis LH, et al. Electronic health records to facilitate clinical research. Clin Res Cardiol 2017 Jan;106(1):1-9. [doi: 10.1007/s00392-016-1025-6] [Medline: 27557678]
2. Staa TV, Goldacre B, Gulliford M, et al. Pragmatic randomised trials using routine electronic health records: putting them to the test. BMJ 2012 Feb 7;344(feb07 1):e55. [doi: 10.1136/bmj.e55] [Medline: 22315246]
3. Tosh G, Soares-Weiser K, Adams CE. Pragmatic vs explanatory trials: the pragmascope tool to help measure differences in protocols of mental health randomized controlled trials. Dialogues Clin Neurosci 2011;13(2):209-215. [doi: 10.31887/DCNS.2011.13.2/gtosh] [Medline: 21842618]
4. Holtrop JS, Glasgow RE. Pragmatic research: an introduction for clinical practitioners. Fam Pract 2020 Jul 23;37(3):424-428. [doi: 10.1093/fampra/cmz092]
5. Black N. Patient reported outcome measures could help transform healthcare. BMJ 2013 Jan 28;346(7896):f167 [FREE Full text] [doi: 10.1136/bmj.f167] [Medline: 23358487]

XSL•FO
RenderX

6. Jandoo T. WHO guidance for digital health: What it means for researchers. Digit Health 2020;6:2055207619898984. [doi: 10.1177/2055207619898984] [Medline: 31949918]
7. Hamilton NS, Edelman D, Weinberger M, Jackson GL. Concordance between self-reported race/ethnicity and that recorded in a Veteran Affairs electronic medical record. N C Med J 2009;70(4):296-300. [Medline: 19835243]
8. Valikodath NG, Newman-Casey PA, Lee PP, Musch DC, Niziol LM, Woodward MA. Agreement of ocular symptom reporting between patient-reported outcomes and medical records. JAMA Ophthalmol 2017 Mar 1;135(3):225-231. [doi: 10.1001/jamaophthalmol.2016.5551] [Medline: 28125754]
9. Fares CM, Williamson TJ, Theisen MK, et al. Low concordance of patient-reported outcomes with clinical and clinical trial documentation. JCO Clin Cancer Inform 2018 Dec;2:1-12. [doi: 10.1200/CCI.18.00059] [Medline: 30652613]
10. O'Brien EC, Mulder H, Jones WS, et al. Concordance between patient-reported health data and electronic health data in the ADAPTABLE trial. JAMA Cardiol 2022 Dec 1;7(12):1235-1243. [doi: 10.1001/jamacardio.2022.3844] [Medline: 36322059]
11. Hostetter M, Klein S. Using patient-reported outcomes to improve health care quality. The Commonwealth Fund. URL: https://www.commonwealthfund.org/publications/newsletter-article/using-patient-reported-outcomes-improve-health-care-quality [accessed 2025-06-04]
12. Verheij RA, Curcin V, Delaney BC, McGilchrist MM. Possible sources of bias in primary care electronic health record data use and reuse. J Med Internet Res 2018 May 29;20(5):e185. [doi: 10.2196/jmir.9134] [Medline: 29844010]
13. Menachemi N, Collum TH. Benefits and drawbacks of electronic health record systems. Risk Manag Healthc Policy 2011;4:47-55. [doi: 10.2147/RMHP.S12985] [Medline: 22312227]
14. Zini MLL, Banfi G. A narrative literature review of bias in collecting patient reported outcomes measures (PROMs). Int J Environ Res Public Health 2021 Nov 26;18(23):12445. [doi: 10.3390/ijerph182312445] [Medline: 34886170]
15. Beatty AL, Peyser ND, Butcher XE, et al. The COVID-19 Citizen Science Study: protocol for a longitudinal digital health cohort study. JMIR Res Protoc 2021 Aug 30;10(8):e28169. [doi: 10.2196/28169] [Medline: 34310336]
16. Peyser ND, Marcus GM, Beatty AL, Olgin JE, Pletcher MJ. Digital platforms for clinical trials: The Eureka experience. Contemp Clin Trials 2022 Apr;115:106710. [doi: 10.1016/j.cct.2022.106710] [Medline: 35183763]
17. Forrest CB, McTigue KM, Hernandez AF, et al. PCORnet® 2020: current state, accomplishments, and future directions. J Clin Epidemiol 2021 Jan;129:60-67. [doi: 10.1016/j.jclinepi.2020.09.036] [Medline: 33002635]
18. Altman DG. Practical Statistics for Medical Research: Chapman & Hall/CRC Press; 1999.
19. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22(3):276-282. [Medline: 23092060]
20. Corrigan-Curay J, Sacks L, Woodcock J. Real-world evidence and real-world data for evaluating drug safety and effectiveness. JAMA 2018 Sep 4;320(9):867-868. [doi: 10.1001/jama.2018.10136] [Medline: 30105359]
21. de Jong AJ, van Rijssel TI, Zuidgeest MGP, et al. Opportunities and challenges for decentralized clinical trials: European Regulators' Perspective. Clin Pharmacol Ther 2022 Aug;112(2):344-352. [doi: 10.1002/cpt.2628] [Medline: 35488483]
22. Qualls LG, Phillips TA, Hammill BG, et al. Evaluating foundational data quality in the National Patient-Centered Clinical Research Network (PCORnet®). EGEMS (Wash DC) 2018 Apr 13;6(1):3. [doi: 10.5334/egems.199] [Medline: 29881761]
23. Groom HC, Crane B, Naleway AL, et al. Monitoring vaccine safety using the vaccine safety Datalink: Assessing capacity to integrate data from Immunization Information systems. Vaccine (Auckl) 2022 Jan 31;40(5):752-756. [doi: 10.1016/j.vaccine.2021.12.048] [Medline: 34980508]
24. Hamad R, Lyman KA, Lin F, et al. The U.S. COVID-19 County Policy Database: a novel resource to support pandemic-related research. BMC Public Health 2022 Oct 10;22(1):1882. [doi: 10.1186/s12889-022-14132-6] [Medline: 36217102]
25. Lusczek ER, Ingraham NE, Karam BS, et al. Characterizing COVID-19 clinical phenotypes and associated comorbidities and complication profiles. PLoS ONE 2021;16(3):e0248956. [doi: 10.1371/journal.pone.0248956] [Medline: 33788884]
26. Chang EM, Gillespie EF, Shaverdian N. Truthfulness in patient-reported outcomes: factors affecting patients' responses and impact on data quality. Patient Relat Outcome Meas 2019;10:171-186. [doi: 10.2147/PROM.S178344] [Medline: 31354371]
27. Breeman S, Constable L, Duncan A, et al. Verifying participant-reported clinical outcomes: challenges and implications. Trials 2020 Mar 4;21(1):241. [doi: 10.1186/s13063-020-4169-7] [Medline: 32131888]
28. Real-world data: assessing electronic health records and medical claims data to support regulatory decision-making for drug and biological products. US Food and Drug Administration. 2021 Sep. URL: https://www.fda.gov/media/152503/download [accessed 2022-03-30]
29. Considerations for the use of real-world data and real-world evidence to support regulatory decision-making for drug and biological products. US Food and Drug Administration. 2021 Dec. URL: https://www.fda.gov/media/154714/download [accessed 2022-03-30]
30. Heckbert SR, Kooperberg C, Safford MM, et al. Comparison of self-report, hospital discharge codes, and adjudication of cardiovascular events in the Women's Health Initiative. Am J Epidemiol 2004 Dec 15;160(12):1152-1158. [doi: 10.1093/aje/kwh314] [Medline: 15583367]
31. Stirratt MJ, Dunbar-Jacob J, Crane HM, et al. Self-report measures of medication adherence behavior: recommendations on optimal use. Transl Behav Med 2015 Dec;5(4):470-482. [doi: 10.1007/s13142-015-0315-2] [Medline: 26622919]

XSL•FO
RenderX

32. Woodfield R, UK Biobank Stroke Outcomes Group, UK Biobank Follow-up and Outcomes Working Group, Sudlow CLM. Accuracy of patient self-report of stroke: a systematic review from the UK Biobank Stroke Outcomes Group. PLoS ONE 2015;10(9):e0137538. [doi: 10.1371/journal.pone.0137538] [Medline: 26355837]
33. Simpson CF, Boyd CM, Carlson MC, Griswold ME, Guralnik JM, Fried LP. Agreement between self-report of disease diagnoses and medical record validation in disabled older women: factors that modify agreement. J Am Geriatr Soc 2004 Jan;52(1):123-127. [doi: 10.1111/j.1532-5415.2004.52021.x] [Medline: 14687326]

Abbreviations

CCS: COVID-19 Citizen Science Study
EHR: electronic health record
NPV: negative predictive value
PCORnet: National Patient-Centered Clinical Research Network
PPV: positive predictive value

Edited by A Mavragani; submitted 05.03.24; peer-reviewed by CM Johnny Wong, S Jia, Y Xu; revised version received 30.04.25; accepted 06.05.25; published 28.07.25.

Please cite as:
Crull E, O'Brien EC, Antiperovitch P, Asfaw K, Beatty AL, Djibo DA, Kaul AF, Kornak J, Marcus GM, Modrow MF, Olgin JE, Orozco J, Park S, Peyser N, Pletcher MJ, Carton TW
Concordance Between Survey and Electronic Health Record Data in the COVID-19 Citizen Science Study: Retrospective Cohort Analysis
JMIR Form Res 2025;9:e58097
URL: https://formative.jmir.org/2025/1/e58097
doi:10.2196/58097

© Elizabeth Crull, Emily C O'Brien, Pavel Antiperovitch, Kirubel Asfaw, Alexis L Beatty, Djeneba Audrey Djibo, Alan F Kaul, John Kornak, Gregory M Marcus, Madelaine Faulkner Modrow, Jeffrey E Olgin, Jaime Orozco, Soo Park, Noah Peyser, Mark J Pletcher, Thomas W Carton. Originally published in JMIR Formative Research (https://formative.jmir.org), 28.7.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Exploratory Impact of iCARE Nigeria, a Combined mHealth and Peer Navigation Intervention, on Depressive Symptoms and Substance Use Among Youth Living With HIV in Nigeria: Single-Arm Trial

Olusegun Ayomikun Ogunmola[1], MSc; Rita Frinue Tamambang[1], MSc, MD; Kehinde Kuti[2], MBBS, MSCI; Lisa M Kuhns[3,4], MPH, PhD; Olutosin Awolude[2], MBBS, MSc; Adedotun Adetunji[5], MBChB, MSCI; Bibilola Oladeji[6], MBBS, PhD; Oladayo Olaleye[2], MPH; Adeola Mary Oyerinde[2], MEd; Robert Garofalo[3,4], MPH, MD; Babafemi Taiwo[7], MBBS; Olayinka Olusola Omigbodun[1,6], MPH, MD

[1]Centre for Child and Adolescent Mental Health, College of Medicine, University of Ibadan, Ibadan, Nigeria
[2]Infectious Disease Institute, College of Medicine, University of Ibadan, Ibadan, Nigeria
[3]Department of Pediatrics, Feinberg School of Medicine, Northwestern University, Chicago, IL, United States
[4]Division of Adolescent Medicine, Ann and Robert H Lurie Children's Hospital of Chicago, Chicago, IL, United States
[5]Department of Family Medicine, University College Hospital, Ibadan, Ibadan, Nigeria
[6]Department of Psychiatry, College of Medicine, University of Ibadan, University College Hospital premises, Queen Elizabeth II Road, Ibadan, Nigeria
[7]Division of Infectious Diseases, Feinberg School of Medicine, Northwestern University, Chicago, IL, United States

Corresponding Author:
Olayinka Olusola Omigbodun, MPH, MD
Centre for Child and Adolescent Mental Health, College of Medicine, University of Ibadan, Ibadan, Nigeria

Abstract

Background: Mental health problems are a barrier to the well-being of youth living with HIV. Many youth living with HIV in Nigeria face peculiar biopsychosocial vulnerabilities that predispose them to mental health problems including depression and substance use. In addition to improving treatment outcomes like medication adherence and linkage to care, peer engagement has shown some promise in improving the social and emotional well-being of this population. Mobile health (mHealth) interventions like SMS text messaging medication reminders may also contribute to better outcomes in youth living with HIV. Emerging evidence suggests that combination interventions may be more effective than single interventions in improving key HIV testing and treatment outcomes among youth in Nigeria.

Objective: This study aims to explore the impact of Intensive Combination Approach to Rollback the Epidemic in Nigerian Adolescents (iCARE Nigeria) study—an mHealth and peer navigation intervention primarily aimed at medication adherence and viral suppression—on depressive symptoms and substance use among youth living with HIV in Nigeria.

Methods: A single-arm clinical trial was conducted at the Infectious Disease Institute, College of Medicine, University of Ibadan, Nigeria— primarily to improve medication adherence and viral suppression among youth living with HIV attending its HIV clinic. The intervention combined peer navigation and daily, 2-way, text message medication reminders delivered over a period of 48 weeks. Participants were screened at baseline and follow-up visits (24 and 48 weeks) for depression and substance use using standardized measures. Paired *t* tests and McNemar tests were used to investigate the change in depressive symptoms and the change in the proportion of participants reporting substance use over time, respectively.

Results: All 40 enrolled participants (n=20, 50% male; mean age 19.9 y, SD 2.5 y) completed baseline and follow-up visits at week 24, while 37 (92.5%) participants completed the week 48 visit. Compared with baseline, there were significantly fewer self-reported depressive symptoms observed at 48 weeks (mean 2.89 vs 2.08; t_{36}=2.04, 95% CI 0.006 - 1.615) but not at 24 weeks (mean 2.89 vs 2.62; t_{36}=0.47, 95% CI –0.74 to 1.44). There were fewer self-reports of substance use at weeks 24 and 48 when compared to baseline, but these were not statistically significant (odds ratio [OR] ∞, 95% CI 0.189-∞ and OR 3.0, 95% CI 0.24 - 157.49, respectively).

Conclusions: These findings suggest a statistically significant reduction in depressive symptoms among youth living with HIV over the 48-week intervention period that may be due to the iCARE Nigeria intervention. However, given limitations such as low levels of depressive symptoms at baseline, small sample size, and the lack of a control group, future studies such as the randomized stepped wedge evaluation of the iCARE intervention are needed to provide better insights into these exploratory findings.

XSL•FO
RenderX

(JMIR Form Res 2025;9:e71141) doi:10.2196/71141

KEYWORDS

mental health; depression; substance use; HIV/AIDS; youth; peer; cell phones; SMS; well-being; Nigeria

Introduction

Mental health problems are common among youth living with HIV attributed to a combination of biological and psychosocial factors [1]. Biologically, HIV infection may have a direct impact on the brain, resulting in behavioral disturbances, mood disorders, neurocognitive decline, and psychosis [2]. In addition, some side effects of antiretroviral therapy (ART) used to achieve viral suppression in HIV infection may have mental health implications for youth living with HIV [3]. Psychosocial factors, which may contribute to the heightened risk of mental health problems among youth living with HIV, include high rates of orphanhood (and its attendant lack of social and material needs), overburdened caregivers, stigma and discrimination, and disclosure of HIV status [4-7].

Studies conducted in sub-Saharan Africa have reported prevalence rates of depression or probable depression among youth living with HIV ranging from 4.4% to 30.2% [1,8-12] compared to 5.8% to 20% among seronegative youth [11,12]. Substance use has also been found to co-exist with mental health problems in youth living with HIV [13], although this relationship is understudied among youth living with HIV in sub-Saharan Africa. Mental health problems (including depression and substance use) have been linked with poor medication adherence, higher risk sexual behaviors as well as reduced quality of life among youth living with HIV, leading to loss to follow-up and even premature mortality [14]. Despite well-documented evidence of mental health problems faced by youth living with HIV, access to mental health services remains an often-unmet priority need for youth living with HIV [15].

There is growing evidence showing that a combination of peer engagement and mobile health (mHealth) interventions may be useful in improving emotional and psychosocial well-being of youth living with HIV [16]. A recent study piloted a peer-led mHealth psychosocial intervention for youth living with HIV aged 15 - 20 years in South Africa, demonstrating increased emotional support and a safe and nonjudgmental atmosphere that encouraged discussion of intimate topics like disclosure and adherence [16]. Peer engagement has also been found to positively impact psychological well-being in addition to improving linkage to care, medication adherence, viral suppression, and retention in care among youth living with HIV [14,17]. Therefore, informed with evidence that combination interventions may be more effective than single interventions in improving key HIV testing and treatment outcomes among youth in Nigeria [18,19], the Intensive Combination Approach to Rollback the HIV Epidemic in Nigerian Adolescents (iCARE Nigeria) study developed, locally adapted, and piloted for efficacy 2 youth-specific interventions—to increase HIV testing among youth at risk of HIV infection and to promote viral suppression in youth living with HIV— each with mHealth and peer navigation components [20,21].

Therefore, in this paper, we explore whether the iCARE medication adherence and viral suppression intervention may have secondary beneficial effects on the mental health of youth living with HIV, specifically by reducing depressive symptoms and substance use among the participants.

Methods

Study Design

The methodology for the iCARE Nigeria pilot study has been described in detail in earlier publications [20,21]. One of the iCARE Nigeria interventions was a single-arm trial, designed to test the efficacy of a combination of peer navigation and daily SMS text messaging medication reminders in improving medication adherence and viral suppression (defined as plasma HIV-RNA of less than 200 copies/ml) among the youth living with HIV [20]. Participants completed 3 study visits—baseline, 24 weeks, and 48 weeks. Data were collected between June 2019 and August 2020 using an interviewer-administered survey, deployed using Research Electronic Data Capture (REDCap; Vanderbilt University) [22,23].

Study Site

This study was conducted at the Infectious Disease Institute, College of Medicine, University of Ibadan (IDI-CoMUI), which receives support from the Centers for Disease Control (CDC)-funded AIDS Prevention Initiative in Nigeria (APIN) Public Health Initiatives and provides ART to more than 6000 persons with HIV, aged 15 years or older.

Study Participants

Participants were youth living with HIV (15 - 24 y old) who had been on ART for at least 3 months. Participants were selected via systematic random sampling. First, a sampling frame containing a list of eligible youth living with HIV at IDI-CoMUI was created from the medical records, which was then stratified by status—virally suppressed versus viremic. Subsequently, every “nth” potential participant was then selected based on the final ratio of target sample size (n=40) to eligible patients in the sampling frame. After a number of refusals and no-shows, the target sample size of 40 with only 35% (14/40) of the participants being virally suppressed (ie, viral load <200 copies/mL). Being a pilot study, the sample size was primarily chosen, not to have adequate power to establish statistical significance but based on predetermined effect sizes satisfactory for future full-blown investigation as indicated in the analysis of the primary outcome (viral suppression) at the study endpoints [20].

Intervention

A combination of peer navigation and daily, 2-way, SMS text messaging medication reminders was delivered over a period of 48 weeks. The daily SMS text messaging reminders were personalized, coded, and scheduled based on the time the

XSL•FO
RenderX

participants were expected to take their medication. A total of 15 minutes after receipt of the personalized reminder, a second personalized message was sent asking the participants if they had taken their medication and each participant was expected to provide a reply. An encouraging message tailored to the response then followed. Furthermore, 8 peer navigators (each assigned 5 peers) were trained to provide peer support, based on a personalized needs assessment and action plan, including facilitating referrals to services as necessary.

The peer navigators undertook a formal training that included topics such as: using coping skills; identifying signs of common mental health problems (depression, anxiety, alcohol and substance use disorders, and suicidality); and providing peer support to youth living with HIV who may be showing signs of mental health problems. Peer navigators checked in on their peers at least once a week, typically over the phone, but in-person meetings could also happen depending on the needs and peculiarities of the peer. Peer navigators were supervised and supported by the designated members of the study team who met with them monthly. At baseline and then again at 24 weeks, study coordinators administered a needs assessment questionnaire, which sought to identify the needs of each participant across social, educational, physical health, and mental health domains. To determine whether a participant would benefit from mental health support, the form asked if the participant in the past two weeks had (1) little interest or pleasure in doing things and (2) was feeling down, depressed, or hopeless. The questionnaire also asked if participants had used alcohol or other psychoactive substances within the previous 6 months. An action plan was then developed and used to prioritize each participant's needs and goals and lay out a plan for achieving those goals. Study coordinators referred participants to mental health services, as needed.

Ethical Considerations

The study was granted ethical approval by the joint University of Ibadan/University College Hospital, Ibadan, Ethics Committee (approval UI/EC/21/0050) and Northwestern University, Lurie Children's Hospital (approval 2019 - 2446). All participants provided informed consent, consistent with the 2014 Nigeria Federal Ministry of Health Guidelines [24]; minors (aged 16 - 17 y or 15 y and emancipated) provided consent for themselves without parental involvement. All participants were assured of their right to withdraw from the study at any given time without the risk of any consequences. All study data were deidentified (each participant received a unique study ID separate from their clinic ID) and stored on the REDCap database accessible only by authorized study staff. At each study visit, participants received light refreshments and payments to cover their transportation costs.

Outcomes and Outcome Measures

Substance Use

The World Health Organization (WHO) Alcohol, Smoking, and Substance Involvement Screening Test (ASSIST) [25] was used to assess substance use. This tool measures use of 10 substances (ie, tobacco, alcohol, cannabis, cocaine, amphetamine, inhalants, sedatives, hallucinogens, opioids, and other) over the previous 3 months. For this analysis, the item, "In the past three months, how often have you used the substances you mentioned?" was used as a measure of substance use. A response of "never" was coded as "no substance use," while those who selected any of the responses "once or twice," "monthly," "weekly," "daily or almost daily" were categorized as endorsing recent substance use (yes or no). Responses to questions about all 10 substances were aggregated to indicate recent use of versus no recent use of substances. We also explored the types of substances used by the participants and how these changed over time.

Depressive Symptoms

The 10-item Center for Epidemiological Studies-Depression scale short version (CES-D) tool [26] was used to assess for depressive symptoms. The tool asks how the individual has felt or behaved within the past 1 week and has items which are scored on a 4-point Likert scale; that is, "rarely or none of the time (less than a day)," "some or a little of the time (1 - 2 d)," "occasionally or a moderate amount of time (3 - 4 d)," and "all of the time (5 - 7 d)." For this analysis, the total score (range 0 - 30) was used as a continuous measure, with higher scores indicating greater depressive symptoms and a score of ≥10 indicating probable depression as this has been found to be a valid cut-off in sub-Saharan Africa including among vulnerable youth [26].

Data Analysis

Frequencies, percentages, and mean (SD) were used to describe participants' personal characteristics. McNemar test was used to test for the changes in proportion of participants reporting substance use between baseline and the 2 follow-up periods (24 weeks and 48 weeks). Paired sample *t* tests were used to determine the changes in depressive symptoms between baseline and the 2 follow-up periods. Participants who died during the trial or were lost to follow-up were not included in the respective analyses. SPSS (version 21; IBM) was used for descriptive and paired *t* tests analyses, while R Studio (Posit) was used for the McNemar exact test computations—to generate CIs and odds ratios. All tests were conducted with significance set at .05 alpha level.

Results

Personal Characteristics of Participants

The study sample consisted of 40 youth living with HIV evenly distributed between male and female sex at birth. The mean age of the participants was 19.9 (SD 2.5) years. The majority (24/40, 60%) were students in senior secondary school (equivalent to grade 10 to 12, high school), more than half (22/40, 55%) were perinatally infected with HIV and 35% (14/40) were virally suppressed at baseline.

Mental Health Profile of Participants

Upon the exclusion of 3 participants (1 dead and 2 lost to follow-up) in the final analyses, results showed that 2.7% (1/37) of the sample screened positive for probable depression (ie, had a score of ≥10 on the CES-D) at both baseline and 48 weeks study visits while no participant met the threshold for depression at 24 weeks. At baseline, 10.8% (4/37) of the participants

XSL•FO
RenderX

reported they had used substances within the past 3 months, compared to 5.4% (2/37) who reported using alcohol and opioids at the 24-week visit and 2.7% (1/37) who reported using alcohol at the 48-week visit.

Effect of the Intervention on Depressive Symptoms Among Youth Living With HIV

Table 1 shows the results of paired *t* test analyses for difference in depressive symptoms. Depressive symptoms reduced over the course of the intervention from a mean of 2.89 (SD 3.05) at baseline to a mean of 2.62 (SD 2.69) at 24 weeks (*P*=.64) and a mean of 2.08 (SD 2.71) at 48 weeks follow-up (*P*=.048). See Table 1 and Figure 1.

Table . Paired sample *t* tests showing the effect of the intervention on depressive symptoms (at 24 weeks and 48 weeks; n=37[a]).

Study timepoint	Depressive symptoms, mean (SD)	*t* test[b] (*df*)	95% CI
Baseline	2.89 (3.05)	—[c]	—
24 weeks	2.62 (2.69)	0.47 (36)	−0.74 to 1.44
48 weeks	2.08 (2.71)	2.04 (36)	0.006 to 1.615[d]

[a]n<40 due to the exclusion of 3 participants (1 dead and 2 lost to follow-up).
[b]2-tailed.
[c]Not applicable.
[d]Significant at *P*<.05.

Figure 1. Changes in depression symptoms over time among study participants.

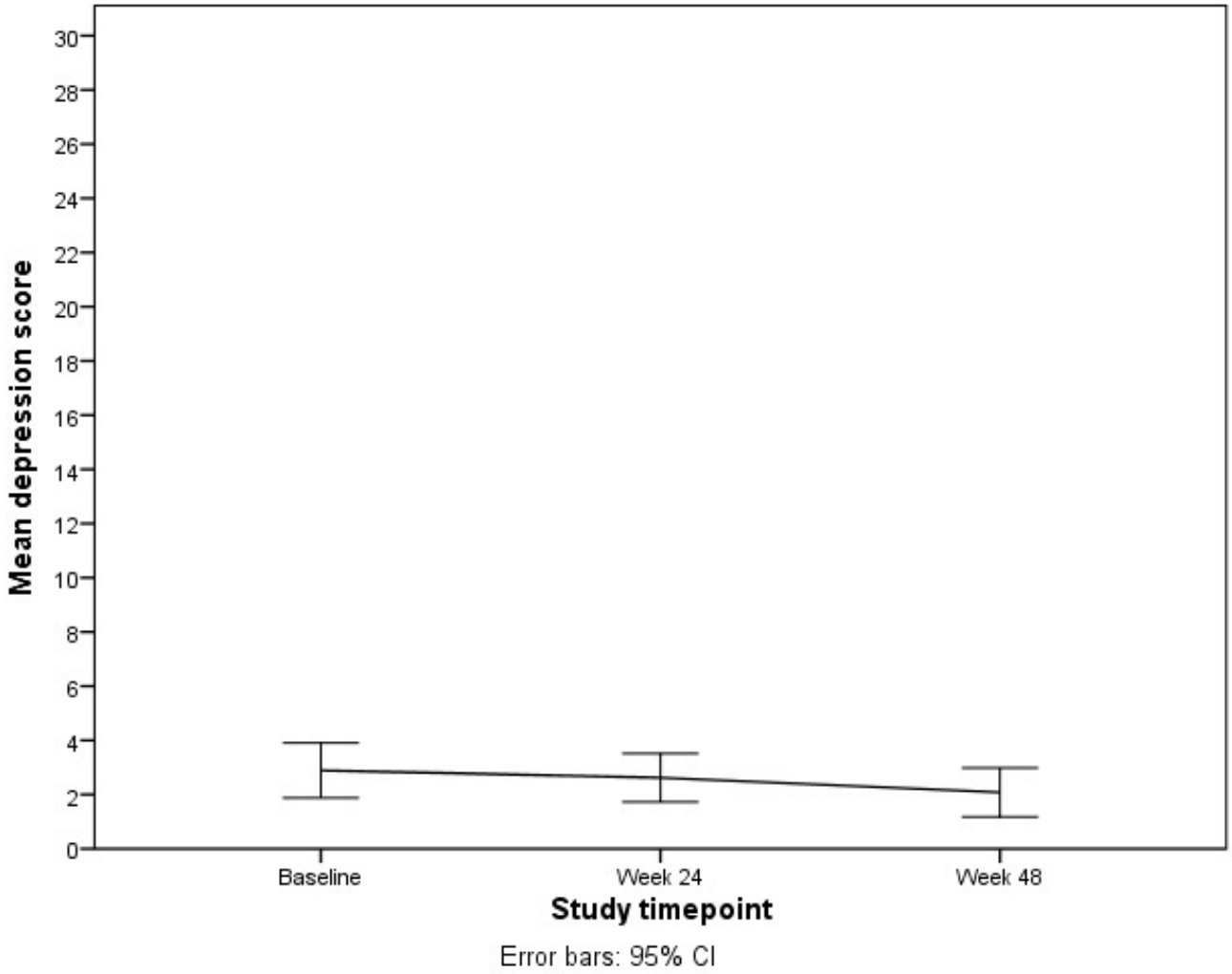

XSL•FO
RenderX

Effect of the Intervention on Substance Use Among Youth Living with HIV

At both week 24 and week 48 visits, there was no significant difference in the number of participants that endorsed the use of substances, but overall substance use was infrequently reported by participants at all assessments (see Table 2).

Table . McNemar exact test showing the effect of the intervention on substance use at 24 weeks and 48 weeks follow-up (n=37[a]).

	Substance use at week 24			Substance use at week 48		
	No, n (%)	Yes, n (%)	OR[b] (95% CI)	No, n (%)	Yes, n (%)	OR (95% CI)
Substance use at baseline			∞ (0.189-∞)			3.0 (0.24 - 157.49)
No	33 (100)	0 (0)		32 (97)	1 (3)	
Yes	2 (50)	2 (50)		3 (75)	1 (25)	

[a]n<40 due to the exclusion of 3 participants (1 dead and 2 lost to follow-up).

[b]OR: odds ratio.

Discussion

Principal Findings

Findings from the iCARE Nigeria pilot study demonstrated the initial efficacy of the iCARE treatment intervention in improving viral suppression among youth living with HIV in Nigeria [20]. In this analysis, we found a modest but statistically significant signal that the iCARE intervention may also have secondary beneficial effects on depressive symptoms. This finding is of potential significance because it was detected in a cohort of only 37 youth, and on average, the youth living with HIV reported minimal depressive symptoms at baseline.

The low (2.7%) prevalence of clinically significant depressive symptoms found in this study at baseline is remarkable when compared to the findings among youth living with HIV in many other sub-Saharan Africa populations. For example, studies have found rates such as 18.9% (Malawi), 27% (Tanzania), 6% - 11% (Cameroon), and 30.2% (Ethiopia) [9,11,12,27]. However, most of these studies were conducted among larger samples (ranging from 169 to 331 youth living with HIV), which afforded them greater statistical power, while some included only young women [1,8], unlike our study that included an equal proportion of men and women. Similar to our results, however, some studies in South Africa found 4% - 6% of adolescents screening positive for depression [1,8]. The low prevalence of depression found in those studies was partly attributed to the support youth living with HIV had been receiving in their respective HIV clinics, which many had been attending since early childhood [8]. A similar mechanism may underlie the low prevalence of clinically significant depression in our cohort. Specifically, the site of our study (IDI-CoMUI) has a youth club supporting adolescents with HIV for the past 6 years. In addition, mental health diagnoses (including depression) and long distance to health care facilities are some factors that have been associated with poor adherence to ART and retention in HIV care [28,29]. It is therefore possible that most youth living with HIV with these underlying disadvantages (and who were also more likely to have depression) were less likely to agree to participate in our study. Would the level of depressive symptoms be greater among youth living with HIV who are not as engaged in care or who receive care in a setting with less support? And would the iCARE intervention have had a greater impact on depressive symptoms among such youth living with HIV? These are important questions for future research to explore, including the ongoing randomized evaluation of the iCARE intervention, which is taking place at multiple sites in Nigeria including in settings that are not as advantaged as the current study site. Also, it may be important to note that 1 participant who had clinically significant depressive symptoms at baseline did not meet the cut-off for depression at week 24.

Furthermore, peer navigators were at the center of support provided by the iCARE intervention. The peer navigators were trained, among other things, to identify signs of common mental health problems, support peers to adopt healthy coping skills, provide emotional support to their peers, and facilitate referral to mental health services. The peer navigators may have contributed to the improvement in their peers' mental well-being by deploying these competencies. This should be explored more fully in future research. The improvement may also be attributable in part to the mental health screening that was conducted as part of a needs assessment done at baseline and at the week 24 study visit. Notably, 8 participants were referred for mental health counseling at baseline, including 7 who flagged positive on at least one of the depression screening questions on the needs assessment form. The eighth participant was referred for mental health counseling based on concerns raised during the needs assessment that her alcohol use might affect adherence to ART.

We also explored trends in substance use, although at baseline only 11% of participants reported substance use in the previous 3 months. Whereas studies conducted in South Africa have found a substance use disorder prevalence of 1.7% - 3% among adolescents living with HIV in South Africa [1,8], none of our participants met the cutoff for substance use disorder. Also, there was no significant change from baseline in the use of substances among participants at either week 24 or week 48. The fact that the tool to assess for substance use was interviewer-administered may have increased the likelihood of social desirability, thus resulting in low endorsement of substance use by the participants. However, given the peculiarities of the ASSIST screening tool, it had to be self-administered. We, however, tried to minimize social desirability by conducting interviews in a private setting—by study staff who had cultivated a largely trusting, safe, and

XSL•FO
RenderX

nonjudgmental relationship with the participants over the course of the study. However, we acknowledge that in spite of these measures, social desirability bias (and consequently underreporting) could still have been present. While the reduction in substance use was not statistically significant, it may still have clinical and practical relevance [30]. This is supported by findings from our needs assessment and action plan, which reveal that most of the participants using substances (3 out of 4 at baseline and 1 out of 2 at week 24) expressed a priority goal to stop using substances. Considering the threat psychoactive substances pose to the health outcomes of youth living with HIV (such as through drug interactions and weakening of the immune system) [31], being able to get 3 out of 4 youth living with HIV to successfully achieve their goals of staying off alcohol and other such substances may be clinically meaningful.

Limitations

Our findings are tempered by some limitations. First, this study is constrained by limitations of the study design (absence of randomization and a control group) and as such, the potential effect of confounding variables such as relatively good mental health at baseline (evidenced by the low level of reported depressive symptoms) and peer support received outside of the intervention could not be accounted for. In addition, the use of an interviewer-administered tool could have introduced some bias leading to underreporting of substance use and depressive symptoms. Similarly, while some mental health improvements were observed, the lack of an objective measure (for example, clinician-assessed depression) is another drawback. Furthermore, the fact that some participants did not complete all the assessments may have further skewed and reduced the generalizability of our findings. Finally, given our small sample size and the fact that there was only 1 participant with clinically significant depression at baseline, the clinical significance of our findings on depressive symptoms is uncertain and worthy of further exploration.

Conclusion

This study found that the iCARE intervention, which combines peer navigation and SMS text messaging medication reminders, may improve depressive symptoms, and thus may have secondary mental health benefits among youth living with HIV in Nigeria. We saw no statistically significant impact on substance use among the participants in this small exploratory study. Given that positive mental health is a continuum that goes beyond the absence of psychopathology, any significant move toward flourishing and well-being [32] in this sample of vulnerable youth suggests potential mental health benefits that warrant further study. The ongoing, randomized, stepped-wedge study (N=568), which is comprehensively evaluating the iCARE Nigeria treatment intervention (NCT 04950153), would potentially provide further evidence for the efficacy of this intervention in promoting the mental well-being of youth living with HIV.

Acknowledgments

Research reported in this publication was supported by the Eunice Kennedy Shriver National Institute of Child Health and Human Development of the National Institutes of Health (award UG3HD096920). Research reported in this publication was also supported by the Fogarty International Center of the National Institutes of Health (award D43TW009608) and by the National Institutes of Health's National Center for Advancing Translational Sciences (grant UL1TR001422). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors acknowledge all iCARE Nigeria research staff, including Dr Tobi Olaniyi, Muyiwa Ojo-Osagie, and Chuks Omoruyi (University of Ibadan). This study could not have been conducted without the selfless contributions of study volunteers and the general youth living with HIV population at Infectious Disease Institute, College of Medicine, University of Ibadan, Nigeria. Generative artificial intelligence was not used to write any portion of this manuscript.

Data Availability

The study data is available from the corresponding author on reasonable request.

Authors' Contributions

OOO contributed to conceptualization. OAO contributed to formal analysis. OAO and RFT contributed to writing – original draft. OOO, BT, KK, LMK, AA, BO, OA, RG, OO, and AMO contributed to writing – reviewing and editing. All authors read and approved the final manuscript for publication.

Conflicts of Interest

None declared.

References

1. Haas AD, Technau KG, Pahad S, et al. Mental health, substance use and viral suppression in adolescents receiving ART at a paediatric HIV clinic in South Africa. J Int AIDS Soc 2020 Dec;23(12):e25644. [doi: 10.1002/jia2.25644] [Medline: 33283916]

XSL•FO
RenderX

2. Patel V, Saxena S, Lund C, et al. The Lancet Commission on global mental health and sustainable development. The Lancet 2018 Oct;392(10157):1553-1598. [doi: 10.1016/S0140-6736(18)31612-X]
3. Shah A, Gangwani MR, Chaudhari NS, Glazyrin A, Bhat HK, Kumar A. Neurotoxicity in the post-HAART era: caution for the antiretroviral therapeutics. Neurotox Res 2016 Nov;30(4):677-697. [doi: 10.1007/s12640-016-9646-0] [Medline: 27364698]
4. Kidman R, Thurman TR. Caregiver burden among adults caring for orphaned children in rural South Africa. Vulnerable Child Youth Stud 2014 Sep;9(3):234-246. [doi: 10.1080/17450128.2013.871379] [Medline: 24999368]
5. Petersen I, Bhana A, Myeza N, et al. Psychosocial challenges and protective influences for socio-emotional coping of HIV+ adolescents in South Africa: a qualitative investigation. AIDS Care 2010 Aug;22(8):970-978. [doi: 10.1080/09540121003623693] [Medline: 20229370]
6. Ramaiya MK, Sullivan KA, O' Donnell K, et al. A qualitative exploration of the mental health and psychosocial contexts of HIV-positive adolescents in Tanzania. PLoS ONE 2016;11(11):e0165936. [doi: 10.1371/journal.pone.0165936] [Medline: 27851797]
7. Willis N, Mavhu W, Wogrin C, Mutsinze A, Kagee A. Understanding the experience and manifestation of depression in adolescents living with HIV in Harare, Zimbabwe. PLoS ONE 2018;13(1):e0190423. [doi: 10.1371/journal.pone.0190423] [Medline: 29298326]
8. Buckley J, Otwombe K, Joyce C, et al. Mental health of adolescents in the era of antiretroviral therapy: is there a difference between HIV-infected and uninfected youth in South Africa? J Adolesc Health 2020 Jul;67(1):76-83. [doi: 10.1016/j.jadohealth.2020.01.010] [Medline: 32269000]
9. Girma D, Assegid S, Gezahegn Y. Depression and associated factors among HIV-positive youths attending antiretroviral therapy clinics in Jimma town, southwest Ethiopia. PLoS ONE 2021;16(1):e0244879. [doi: 10.1371/journal.pone.0244879] [Medline: 33406163]
10. Ijadunola MY, Ojo TO, Babatunde A, et al. Street hawking among in-school adolescents in a south-western town in Nigeria: pattern, determinants and effects on school performance. Int J Adolesc Med Health 2015 Feb;27(1):41-48. [doi: 10.1515/ijamh-2013-0342] [Medline: 24914712]
11. Kanmogne GD, Qiu F, Ntone FE, et al. Depressive symptoms in HIV-infected and seronegative control subjects in Cameroon: effect of age, education and gender. PLoS ONE 2017;12(2):e0171956. [doi: 10.1371/journal.pone.0171956] [Medline: 28231258]
12. Lwidiko A, Kibusi SM, Nyundo A, Mpondo BCT. Association between HIV status and depressive symptoms among children and adolescents in the Southern Highlands Zone, Tanzania: a case-control study. PLoS ONE 2018;13(2):e0193145. [doi: 10.1371/journal.pone.0193145] [Medline: 29470512]
13. Nakimuli-Mpungu E, Bass JK, Alexandre P, et al. Depression, alcohol use and adherence to antiretroviral therapy in sub-Saharan Africa: a systematic review. AIDS Behav 2012 Nov;16(8):2101-2118. [doi: 10.1007/s10461-011-0087-8] [Medline: 22116638]
14. Haas AD, Ruffieux Y, van den Heuvel LL, et al. Excess mortality associated with mental illness in people living with HIV in Cape Town, South Africa: a cohort study using linked electronic health records. Lancet Glob Health 2020 Oct;8(10):e1326-e1334. [doi: 10.1016/S2214-109X(20)30279-5] [Medline: 32971055]
15. Bhana A, Abas MA, Kelly J, van Pinxteren M, Mudekunye LA, Pantelic M. Mental health interventions for adolescents living with HIV or affected by HIV in low- and middle-income countries: systematic review. BJPsych Open 2020 Sep 4;6(5):e104. [doi: 10.1192/bjo.2020.67] [Medline: 32886056]
16. Atujuna M, Simpson N, Ngobeni M, et al. Khuluma: using participatory, peer-led and digital methods to deliver psychosocial support to young people living with HIV in South Africa. Front Reprod Health 2021;3(July):687677. [doi: 10.3389/frph.2021.687677] [Medline: 36304024]
17. Woollett N, Pahad S, Black V. "We need our own clinics": adolescents' living with HIV recommendations for a responsive health system. PLOS ONE 2021;16(7):e0253984. [doi: 10.1371/journal.pone.0253984] [Medline: 34197529]
18. Mills EJ, Lester R, Thorlund K, et al. Interventions to promote adherence to antiretroviral therapy in Africa: a network meta-analysis. Lancet HIV 2014 Dec;1(3):e104-e111. [doi: 10.1016/S2352-3018(14)00003-4] [Medline: 26424119]
19. Tun W, Vu L, Dirisu O, et al. Uptake of HIV self-testing and linkage to treatment among men who have sex with men (MSM) in Nigeria: a pilot programme using key opinion leaders to reach MSM. J Int AIDS Soc 2018 Jul;21 Suppl 5(Suppl Suppl 5):e25124. [doi: 10.1002/jia2.25124] [Medline: 30033680]
20. Taiwo BO, Kuti KM, Kuhns LM, et al. Effect of text messaging plus peer navigation on viral suppression among youth with HIV in the ICARE Nigeria pilot study. J Acquir Immune Defic Syndr 2021 Aug 1;87(4):1086-1092. [doi: 10.1097/QAI.0000000000002694] [Medline: 34153015]
21. Garofalo R, Adetunji A, Kuhns LM, et al. Evaluation of the ICARE Nigeria pilot intervention using social media and peer navigation to promote HIV testing and linkage to care among high-risk young men: a nonrandomized controlled trial. JAMA Netw Open 2022 Feb 1;5(2):e220148. [doi: 10.1001/jamanetworkopen.2022.0148] [Medline: 35191969]
22. Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: building an international community of software platform partners. J Biomed Inform 2019 Jul;95(May):103208. [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]

XSL•FO
RenderX

23. A consensus report on guidelines for young persons' participation in research and accessing sexual and reproductive health services in Nigeria. : Federal Ministry of Health (Nigeria); 2014 URL: https://www.nhvmas-ng.org/publication/1441310961154.120.108.14.pdf [accessed 2025-07-12]
24. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381. [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
25. Humeniuk R, Ali R, WHO ASSIST Phase II Study Group. Validation of the Alcohol, Smoking and Substance Involvement Screening Test (ASSIST) and pilot brief intervention [electronic resource]: a technical report of phase II findings of the WHO ASSIST Project. : World Health Organization; 2006 URL: https://iris.who.int/handle/10665/43504 [accessed 2025-07-12]
26. Kilburn K, Prencipe L, Hjelm L, Peterman A, Handa S, Palermo T. Examination of performance of the Center for Epidemiologic Studies Depression Scale Short Form 10 among African youth in poor, rural households. BMC Psychiatry 2018 Jun 18;18(1):201. [doi: 10.1186/s12888-018-1774-z] [Medline: 29914413]
27. Kip EC, Udedi M, Kulisewa K, Go VF, Gaynes BN. Stigma and mental health challenges among adolescents living with HIV in selected adolescent-specific antiretroviral therapy clinics in Zomba District, Malawi. BMC Pediatr 2022 May 6;22(1):253. [doi: 10.1186/s12887-022-03292-4] [Medline: 35524228]
28. Tih PM, Ayima CW, Boeyeo FN, Atanga PN, Eveline MK. High incidence and predictors of loss to follow-up among children and adolescents on life long antiretroviral therapy in the conflict-affected northwest and southwest regions of Cameroon: a retrospective cohort study. TOAIDJ 2023 Jun 21;17(1). [doi: 10.2174/18746136-v17-2305300-2022-20]
29. Rooks-Peck CR, Adegbite AH, Wichser ME, et al. Mental health and retention in HIV care: a systematic review and meta-analysis. Health Psychol 2018 Jun;37(6):574-585. [doi: 10.1037/hea0000606] [Medline: 29781655]
30. Page P. Beyond statistical significance: clinical interpretation of rehabilitation research literature. Int J Sports Phys Ther 2014 Oct;9(5):726-736. [Medline: 25328834]
31. HIV and substance use. National Institutes of Health. 2024. URL: https://hivinfo.nih.gov/understanding-hiv/fact-sheets/hiv-and-substance-use [accessed 2024-01-08]
32. Westerhof GJ, Keyes CLM. Mental illness and mental health: the two continua model across the lifespan. J Adult Dev 2010 Jun;17(2):110-119. [doi: 10.1007/s10804-009-9082-y] [Medline: 20502508]

Abbreviations

ART: antiretroviral therapy
ASSIST: Alcohol, Smoking, and Substance Involvement Screening Test
CES-D: Center for Epidemiological Studies-Depression scale short version
iCARE: Intensive Combination Approach to Rollback the Epidemic in Nigerian Adolescents
IDI-CoMUI: Infectious Disease Institute, College of Medicine, University of Ibadan
mHealth: mobile health
OR: odds ratio
REDCap: Research Electronic Data Capture
WHO: World Health Organization

Edited by A Mavragani; submitted 10.01.25; peer-reviewed by A Ojaruega, C Chikwe, E Gavor; revised version received 22.04.25; accepted 05.05.25; published 29.07.25.

Please cite as:
Ogunmola OA, Tamambang RF, Kuti K, Kuhns LM, Awolude O, Adetunji A, Oladeji B, Olaleye O, Oyerinde AM, Garofalo R, Taiwo B, Omigbodun OO
Exploratory Impact of iCARE Nigeria, a Combined mHealth and Peer Navigation Intervention, on Depressive Symptoms and Substance Use Among Youth Living With HIV in Nigeria: Single-Arm Trial
JMIR Form Res 2025;9:e71141
URL: https://formative.jmir.org/2025/1/e71141
doi:10.2196/71141

© Olusegun Ayomikun Ogunmola, Rita Frinue Tamambang, Kehinde Kuti, Lisa M Kuhns, Olutosin Awolude, Adedotun Adetunji, Bibilola Oladeji, Oladayo Olaleye, Adeola Mary Oyerinde, Robert Garofalo, Babafemi Taiwo, Olayinka Olusola Omigbodun. Originally published in JMIR Formative Research (https://formative.jmir.org), 29.7.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative

XSL•FO
RenderX

Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Virtual Reality Gamification of Visual Search, Response Inhibition, and Visual Short-Term Memory Tasks for Cognitive Assessment: Experimental Study

Marios Hadjiaros[1,2,3], PhD; Andria Shimi[1,3,4], Prof Dr; Kleanthis Neokleous[1], PhD; Constantinos Pattichis[1,2,3], Prof Dr; Marios Avraamides[1,4], Prof Dr

[1]CYENS—Centre of Excellence, 1 Dimarchou Lellou Demetriadi, Nicosia, Cyprus
[2]Department of Computer Science, University of Cyprus, Nicosia, Cyprus
[3]Biomedical Engineering Research Centre, University of Cyprus, Nicosia, Cyprus
[4]Department of Psychology, University of Cyprus, Nicosia, Cyprus

Corresponding Author:
Marios Hadjiaros, PhD
CYENS—Centre of Excellence, 1 Dimarchou Lellou Demetriadi, Nicosia, Cyprus

Abstract

Background: Cognitive tasks are foundational tools in psychology and neuroscience for studying attention, perception, and memory. However, they typically employ simple or artificial stimuli and require numerous repetitive trials, which can adversely affect participant engagement and ecological validity.

Objective: This study investigated whether gamified versions of 3 established cognitive tasks, namely, the Visual Search task (attention), the Whack-the-Mole task (response inhibition), and the Corsi block-tapping test (visual short-term memory), replicate the typical patterns of results reported for their traditional counterparts. It also examined whether the method of administration—in immersive virtual reality (VR) versus desktop computer, and in the laboratory versus at home—influences performance.

Methods: Seventy-five participants (male=24, female=51; age range 18‑35 years; mean 23.15, SD 4.38 years) were randomly assigned to 1 of 3 administration conditions (n=25 each). In the VR-Lab condition, participants completed the tasks in immersive VR within the laboratory; in the Desktop-Lab condition, they completed the tasks on a 2D desktop screen in the laboratory; and in the Desktop-Remote condition, participants completed the tasks on their personal computers at home. All participants completed the same gamified tasks while seated, entering responses with either a mouse or a VR controller, depending on the condition.

Results: The results obtained from these gamified tasks across all 3 administration conditions replicated the typical performance patterns observed with their traditional counterparts, despite using more ecologically valid stimuli and fewer trials. However, administration modality did influence certain performance measures, particularly reaction times (RTs) and task efficiency. Specifically, in the Visual Search task, RTs were significantly faster in the VR-Lab condition (mean 1.24 seconds) than in the Desktop-Lab (mean 1.49 seconds; $P<.001$) and Desktop-Remote (mean 1.44 seconds; $P=.008$) conditions. In the Whack-the-Mole task, no significant group differences emerged in d' scores (VR-Lab: mean 3.79, Desktop-Remote: mean 3.75, Desktop-Lab: mean 3.62; $P=.49$), but RTs were slower in the Desktop-Remote condition (mean 0.64 seconds) than in the VR-Lab (mean 0.41 seconds; $P<.001$) and Desktop-Lab (mean 0.48 seconds; $P<.001$) conditions. For the Corsi block-tapping test, no significant group differences in span scores were found (VR-Lab: mean 5.48, Desktop-Lab: mean 5.68, and Desktop-Remote: mean 5.24; $P=.24$). Finally, a significant positive correlation was observed between RTs for Hits in the Whack-the-Mole task and feature search trials in the Visual Search task ($r=0.24$; $P=.04$).

Conclusions: Gamified cognitive tasks administered in VR replicated established behavioral patterns observed with their traditional versions while improving ecological validity and reducing task duration. Administration modality had limited effects on overall outcomes, although RTs were slower in remote settings. These findings support the feasibility of using gamified VR tasks for scalable and ecologically valid cognitive assessment. Overall, the study underscores the potential of VR to increase participant engagement and enrich cognitive research through more immersive and motivating testing environments.

(JMIR Form Res 2025;9:e65836) doi:10.2196/65836

KEYWORDS

attention; memory; virtual reality; visuospatial; serious games; cognitive assessment; Corsi Test; Visual Search; Go/No-Go; response inhibition

XSL•FO
RenderX

Introduction

Background

Cognitive tasks are commonly used in Psychology and Neuroscience to investigate various mental processes, such as attention, memory, and perception [1,2]. These tasks often involve simple stimuli, such as alphanumeric characters and shapes and a large number of trials, a deliberate design choice aimed at controlling for extraneous variables and ensuring the reliability of the results [3,4]. While this methodological approach is crucial for maintaining experimental rigor and validity, it may inadvertently lead to issues regarding participant engagement [5]. For example, the repetitive tasks and the simple stimuli may cause participant disinterest, boredom, and fatigue, which in turn may compromise data quality or introduce unintended biases [6,7]. Thus, while striving for methodological control, researchers must also consider strategies to enhance participant engagement to mitigate these potential challenges and maintain the integrity of their findings.

Virtual Reality as a Tool for Gamified Cognitive Assessment

Virtual reality (VR) technology presents a promising avenue for addressing the engagement challenges inherent in traditional cognitive tasks [8]. By leveraging VR, researchers can gamify tasks, transforming them into immersive and interactive experiences that are inherently more engaging for participants [9].

Numerous studies have indicated that gamification in cognitive tasks can significantly increase participant engagement and motivation [10,11]. Recent studies exploring VR gaming have emphasized the significant role of user engagement and motivation in enhancing the overall experience. One key factor driving VR game adoption is the concept of "flow," which is a state of optimal immersion where players experience intense enjoyment and concentration. This "flow" state is essential for fostering deeper engagement with VR games, particularly as the player becomes more immersed in the virtual environment. The sense of spatial presence, or the feeling of being physically located within the VR world, is another major motivator, as it strengthens the emotional connection to the game and enhances relaxation or enjoyment [12,13].

Cognitive Tasks Under Investigation

The Visual Search task is a widely used method for investigating the mechanisms of attention involved in locating a target among distractors, mirroring real-world scenarios such as finding a friend in a crowded cafeteria, searching for an item in the grocery store, or searching a target object among nontarget objects during security screening. In laboratory settings, a typical Visual Search task involves presenting a target object on the computer screen among distractors [14,15]. Stimuli often consist of alphanumeric characters or simple geometric shapes. Visual search tasks typically involve 2 key manipulations: similarity and display size. Similarity refers to the type of search that is determined by how visually distinct the target is from the distractors. In *feature search* trials, the target differs from distractors in terms of a single feature, such as its shape (eg, locating the letter "O" among several "Xs") or its color (eg, locating a red "X" among white "Xs"). In *conjunction search* trials, the target is defined by a combination of features that are also present in the distractors (eg, locating a green "O" among green "Xs" and red "Os"). The second manipulation is the display size, which refers to the number of distractors presented in each trial alongside the target [16]. The typical pattern of results obtained across many studies is that (1) reaction times (RTs) (ie, the time it takes the participant to find the target from the onset of the display) are shorter for feature search trials than conjunction search trials, (2) RT increases significantly with display size, a phenomenon known as the "display size effect," but only in conjunction search trials [15,17,18], and (3) RT remains relatively constant across display sizes in feature search trials, a phenomenon known as the "pop-out effect" [19].

The Whack-the-Mole task is a Go/No-Go task that provides a measure of response inhibition, that is, the ability to inhibit inappropriate prepotent responses [20]. Past research has shown that response inhibition is a key ability for cognitive [21] and psychosocial [22] development and is a reliable predictor of school readiness and academic achievement in domains such as math and reading [23]. Impairments in response inhibition are commonly observed in neurodevelopmental disorders such as autism [24] and attention-deficit/hyperactivity disorder [25,26] as well as in psychiatric and neurological disorders such as borderline personality disorder [27] and Parkinson disease [28]. In computerized Go/No-Go tasks, participants are required to execute an action rapidly (eg, press a button) upon detecting a specific stimulus (Go trials) and withhold a response when presented with a different stimulus (No-Go trials) [20,21]. To ensure task effectiveness, the majority of trials are typically Go trials (approximately 75%), priming the execution of a response [20,21]. Response inhibition is assessed by computing d', a sensitivity metric that takes into account the proportion of hits (ie, correctly executed responses in Go trials) and false alarms (ie, incorrect execution of responses in No-Go trials) [20,29]. Although d' may vary depending on the manipulations of the experiment, such as the discriminability of the stimuli, its typical value, indicating moderate sensitivity, is between 3 and 4 [20,30]. In general, participants exhibit false alarms in 10% of the No-Go trials [20,21].

The Corsi Test is one of the most widely used tasks to assess visuospatial short-term memory. The original version [31] consists of 9 physical cubes placed on a board. Each trial involves the experimenter sequentially touching a series of blocks from one side of the board, and the participant, situated on the opposite side, is required to reproduce the sequence by touching the blocks in the same order. If the participant successfully reproduces the sequence, then the number of blocks touched in the sequence is increased by 1 [31-33]. The task continues until the participant fails to accurately reproduce the sequence [34]. The longest sequence that is reproduced correctly indexes the participant's visuospatial short-term memory span. Studies show that healthy adults have a typical span between 5 and 7 [35]. In more recent implementations, electronic versions of the task have been developed and delivered through computers or mobile devices, where targets light up in a specific sequence that participants reproduce by clicking or touching

XSL•FO
RenderX

the targets. Even in these modern versions of the task, the typical span ranges between 5 and 7 [33].

Related Work

Immersive VR limits extraneous distractions from the surrounding environment [36,37]. This is because immersive VR setups entail wearing a head-mounted display (HMD) that effectively blocks out the external environment, thereby reducing potential environmental distractions that may interfere with task performance [38,39]. This feature may be particularly important in cognitive research, where maintaining participants' focus and attention is being assessed [40] and a measure of "true" ability under no external distraction is needed. It is, therefore, imperative to ascertain whether these gamified tasks, administered with or without immersive VR, produce the expected pattern of results as their traditional counterparts, an indication that they capture the same underlying cognitive processes. Recent studies have investigated the consistency and relative performance of computerized and VR-based cognitive tasks, revealing nuanced differences between the 2 modalities. These findings underscore the importance of task context and cognitive demands when evaluating their efficacy.

For example, Barrett et al [41] compared VR, 3D desktop, and 2D versions of a category learning task in which participants classify visual stimuli into categories based on 2 relevant features while ignoring a third irrelevant one. Feedback is provided after each trial to support learning. RTs in the 2D condition were significantly faster than in the VR and 3D conditions, and fixation durations were shorter, indicating reduced cognitive load for the 2D environment. However, participants reported higher levels of immersion and engagement in the VR and 3D conditions, suggesting that while traditional computerized tasks may facilitate quicker responses, VR environments offer richer experiential benefits.

This trade-off between speed and immersion is also reflected in findings from Bhargava et al [42], who examined cognitive task performance across VR and mobile-based conditions using custom-designed games. Bhargava et al [42] validated 3 custom-designed games: 2 VR-based (Navigation and Hand-Eye Coordination) and 1 tablet-based (Memory) against the ACE-III (Addenbrooke's Cognitive Examination—III) cognitive assessment tool in a cohort of young adults. The ACE-III is a comprehensive tool used to assess cognitive functioning across 5 domains: attention, memory, language, verbal fluency, and visuospatial skills. The study found no significant differences in overall task performance between the VR and 3D mobile conditions; however, the VR tasks elicited significantly higher motivation and engagement scores. These findings suggest that VR can enhance user experience and engagement without compromising the validity or reliability of cognitive assessment.

This enhanced motivational effect of VR-based tasks is further reinforced by Faria et al [43], who explored VR in the context of cognitive rehabilitation. Similar findings are reported by Faria et al [43], who evaluated 2 approaches to personalized cognitive rehabilitation—1 VR-based and 1 traditional—with patients with chronic stroke. Both methods were equally effective in improving cognitive performance, but the VR approach led to higher participant satisfaction and engagement due to its interactive features. This suggests that VR can deliver comparable therapeutic outcomes with added motivational benefits.

In addition to motivation, several studies emphasize the ecological validity of VR, a point illustrated well in the work by Tan et al [44]. Tan et al [44] assessed a VR-based cognitive screening tool targeting 6 cognitive domains. Their results showed high accuracy in detecting cognitive impairments, comparable with traditional methods, with the added benefit of enhanced ecological validity in VR. Participants were able to perform naturalistic tasks, such as navigation and object interaction, which are difficult to replicate in conventional computerized tests.

This growing interest in ecological validity and realism of cognitive assessments in VR is echoed in the meta-analysis by Neguţ et al [45], which further consolidates the strengths and challenges of immersive assessments. Finally, Neguţ et al [45] conducted a meta-analysis comparing VR-based assessments with classical paper-and-pencil or computerized measures. They reported that VR tasks often felt more challenging due to their immersive and dynamic nature, which can increase perceived cognitive load. However, the ecological validity and potential for more realistic stimuli in VR environments often make them more representative of real-world cognitive challenges.

In summary, past research shows that VR-based tasks enhance immersion, engagement, and ecological validity, making them a compelling alternative to traditional assessments. While RTs may vary depending on task design and context, this trade-off is often outweighed by the benefits of more naturalistic and interactive environments. These advantages position VR as a valuable complement to conventional methods, particularly in contexts that require high levels of interaction, naturalistic task performance, or strong participant motivation.

Building on this body of work, this study extends previous research by directly comparing attention and memory performance in immersive VR with desktop settings, both in-lab and remote, to assess the impact of presentation medium and testing environment on cognitive outcomes. Related to this question, several past studies using a variety of tasks have indicated that the quality of data collected on the web remains uncompromised [46-52]. However, web-based data frequently exhibit greater variability in contrast to data gathered in controlled laboratory environments [51]. This variability could either stem from the diminished control over experimental conditions, as previously mentioned, or it could be an intrinsic characteristic of studying more diverse participant cohorts, thereby becoming a variable of interest in its own right [53].

This issue of variability is particularly relevant in light of findings by Segen et al [54], whose online versus lab-based comparison provides an important context for interpreting remote testing results. Notably, Segen et al [54] found that the data collected in a conventional laboratory setting and those collected on the web produced very similar results, although the web-based data were more variable, with SEs being about 10% larger than those of the data collected in the lab. However, Segen et al [54] have used a scene recognition task that might be less prone to the environmental distraction that is typically

XSL•FO
RenderX

present when carrying out a task remotely compared with the attention and memory tasks we used here.

Problem Statement and Objective

The heightened engagement resulting from the use of VR may come at a cost to experimental control. The dynamic and immersive nature of VR environments introduces additional variables that may influence participants' cognitive processes, potentially altering the patterns of results obtained compared with their traditional counterparts. Therefore, it becomes imperative for scientists to systematically evaluate whether gamifying cognitive tasks in VR environments yields findings that are consistent with those obtained with traditional methods [19]. Doing so would ensure that while enhancing engagement with gamification, researchers can still maintain the methodological rigor necessary for robust scientific conclusions. Therefore, the main objective of this study was to examine whether gamified versions of the Visual Search, the Whack-the-Mole, and the Corsi block tasks, each assessing a distinct cognitive process, produce similar patterns of results as their traditional versions. A second objective was to examine whether the way the gamified tasks are administered influences the pattern of results obtained. To this purpose, we compared results across 3 different conditions that involved the same gamified cognitive tasks: an immersive VR condition (VR-Lab), a desktop VR condition (Desktop-Lab; nonimmersive), and a desktop VR condition in which participants carried out the task at home (Desktop-Remote; nonimmersive). By conducting this comparison, we aimed to elucidate the influence of immersive VR on task outcomes, providing insights into the effectiveness of different task formats.

Methods

Participants

The experimental study included 75 healthy young adults (male=24, female=51) with an overall mean age of 23.15 years (SD 4.38 years) recruited from the student community of the University of Cyprus and the CYENS—Centre of Excellence. Recruitment was conducted using a convenience sampling method through email announcements, web-based student forums and platforms, and direct invitations. All participants met the following inclusion criteria: (1) aged between 18 and 35 years, (2) normal or corrected-to-normal vision, and (3) no reported history of neurological, psychiatric, or cognitive impairments that could affect task performance. Eligibility was determined through a brief prescreening interview. Individuals with a self-reported history of neurological disorders (eg, epilepsy and traumatic brain injury), psychiatric conditions (eg, attention-deficit/hyperactivity disorder, anxiety, and depression), or uncorrected visual impairments were excluded from participation. Participants were randomly assigned to 1 of 3 between-subject conditions: VR-Lab, Desktop-Lab, or Desktop-Remote, with 25 participants in each group. No participants were excluded from the statistical analyses, and there were no dropouts. All laboratory-based sessions (VR-Lab and Desktop-Lab conditions) took place at the Experimental Psychology Laboratory at the University of Cyprus, a controlled environment designed to minimize external distractions.

Materials

Visual Search Task

The Visual Search task used in this study involved the presentation of brown and gray axes (Multimedia Appendix 1), with the target object changing from trial to trial. In feature search trials, the target differed from the distractors in terms of color, for example, a gray axe presented among brown axes. In contrast, in conjunction search trials, the target was defined by a combination of features that were also present in the distractors. For example, the target could be a brown 1-sided axe presented among gray 1-sided axes and brown double-sided axes (Figure 1). To examine the influence of display size, we varied the number of axes in both trial types to 5, 15, 25, or 40. Participants were instructed to search the display for the target and provide a response as soon as they detected it. In every trial, the targets were presented for 1.5 seconds, followed by the search display. Overall, the task consisted of 96 trials (50% feature search trials and 50% conjunction search trials), with 8 trials in each possible combination of search type and display size. Before the experimental trials, participants performed 8 practice trials, 1 in each combination of search type (Feature vs Conjunction Search) and display size (5, 15, 25, or 40) to familiarize themselves with the task and procedure. Trials were presented in a different random order for each participant. Participants in the VR-Lab condition searched for the target and selected it by directing a laser beam extended from the VR controller toward the target. The response was logged when the laser beam was on the target for 2 seconds; however, the RT was measured from the moment the beam first landed on the target. Participants in the Desktop-Lab and Desktop-Remote conditions responded by moving the mouse cursor to the target and clicking on it. In all 3 conditions (VR-Lab, Desktop-Lab, and Desktop-Remote), participants had the whole board with the axes in their field of view and thus did not need to turn their heads to search.

XSL•FO
RenderX

Figure 1. Example of a conjunction search trial in the Visual Search task of the experimental study. In this trial, participants searched for a brown 1-sided axe (target) among gray 1-sided axes and brown double-sided axes (distractors). This is a trial with a display size of 15. The red outline is added for illustration purposes and was not present in the actual display viewed by participants.

Whack-the-Mole Task

In our version of the Whack-the-Mole task that was modeled after Casey et al [55], participants used a mallet to hit moles that popped up with random disguises from the hole in the center of the ground. In Go trials, a mole appeared with a varying disguise in each trial, and participants were instructed to hit it as fast as they could by either pulling the trigger on the handheld controller with the VR-Lab condition or clicking the mouse in the Desktop-Lab and Desktop-Remote conditions. Participants in the VR-Lab condition held a virtual mallet mapped to the controller in their dominant hand and hit the mole by pulling the trigger with their index finger. Similarly, participants in the Desktop-Lab and Desktop-Remote conditions clicked the mouse with the index finger of their dominant hand to hit the mole. In No-Go trials, mushrooms appeared instead of the mole, and participants were instructed to withhold a response (Figure 2). Before the experimental task, participants were shown the 7 disguises of the mole and the 3 types of mushrooms and carried out 10 practice trials. As shown in Figure 2, the task included 7 different mole disguises and 3 different types of mushrooms. Each participant completed 100 trials, with 75 being the Go trials and 25 being the No-Go trials. The 2 trial types were presented randomly within the task. In all 3 conditions (VR-Lab, Desktop-Lab, and Desktop-Remote), participants had a clear view of the ground area from which the mole emerged, without the need to turn their heads.

XSL•FO
RenderX

Figure 2. An illustration of the Whack-the-Mole task as used in the study. On the left, the 3 different types of mushrooms (top row) and the 7 different mole disguises (middle and bottom rows) are shown. On the right, an example Go trial is shown. In the VR-Lab condition, participants had to pull the trigger on the handheld controller to respond. In the Desktop-Lab and Desktop-Remote conditions, they had to click the mouse. In No-Go trials, showing different types of mushrooms, participants had to refrain from any response.

Corsi Test

In this task, participants were presented with a 4×4 matrix of gray cubes (Figure 3). Each trial involved a unique sequence of cubes changing color from gray to yellow sequentially, and participants were required to replicate the sequence by tapping the blocks that changed color in the same order. Participants in the VR-Lab condition used the VR controllers to touch the blocks, whereas those in the Desktop-Lab and Desktop-Remote conditions clicked the blocks with the mouse. The task began with a single cube and the length of the sequence increased as participants responded correctly. Participants were given 3 attempts at each span length and advanced to the next length if they responded correctly in at least 2 attempts. The block span for each participant was measured as the length of the last sequence that was repeated correctly in at least 2 out of 3 attempts. After completing the tasks, participants in the Desktop-Remote condition were asked to click on a save button that sent their data to a cloud database. In all 3 conditions (VR-Lab, Desktop-Lab, and Desktop-Remote), the whole 4×4 matrix was within the participants' field of view, so no head turning was required.

XSL•FO
RenderX

Figure 3. An illustration of the matrix in the Corsi Test as used in this study. Participants viewed a sequence of cubes turning yellow and then recreated it by either touching the cubes with the handheld controllers (VR-Lab condition) or clicking on each cube with the mouse (Desktop-Lab and Desktop-Remote participants).

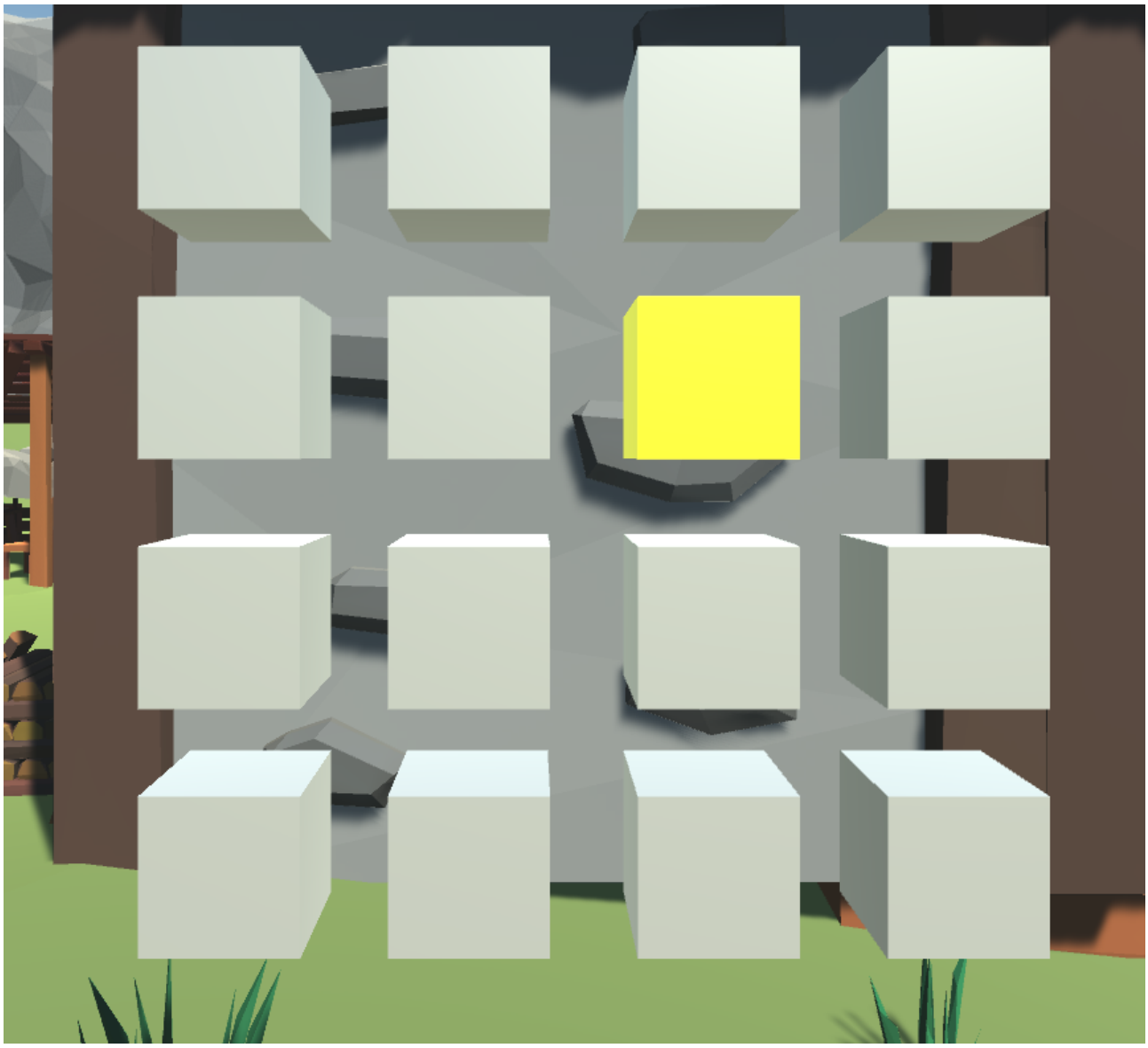

Procedure

Participants were initially contacted via email with details about the study and available time slots for participation. Those interested registered electronically by selecting a suitable time slot. Upon arrival at the lab, participants read a detailed description of the study and signed a consent form if they agreed to take part. No participants withdrew from the experimental procedure. Participants in the Desktop-Remote condition followed the same process, with the only difference being that they completed the study from home through a secure hyperlink and communicated with the experimenter (first author) via video call. They completed the tasks on their personal computers, and they were instructed to select a quiet, distraction-free space and to use a desktop or laptop computer with a stable internet connection. During the experiment, all participants were seated and instructed to use only their hand to operate either the mouse or the VR controller, ensuring that movement was limited primarily to the wrist and minor elbow motion. This protocol was applied consistently across all 3 tasks, regardless of condition. Before starting the experiment, participants in all conditions underwent a brief orientation session where they were familiarized with the tasks and response methods.

Participants in the VR-Lab condition carried out the tasks in an HTC Vive Pro HMD and used the Vive handheld controllers to respond. Participants in the Desktop-Lab condition carried out the task on the same desktop computer that was used for the VR-Lab condition but viewed the tasks on a 24" LCD display. The order of the 3 tasks was counterbalanced across conditions using a 3×3 Latin square design. For all tasks, data in the VR-Lab and Desktop-Lab conditions were automatically recorded by the computer running the experiment. In the 2 lab conditions, data were collected locally on the computer of the lab, while in the Desktop-Remote condition, data were encrypted and stored in a secure database.

XSL•FO
RenderX

Statistical Design

Separate mixed-design ANOVAs were conducted on mean RTs and d', depending on the cognitive task (see "Results" section for task-specific details). The within-subject variables for each task are detailed in the "Results" section, while the between-subject variable across all tasks was the mode of administration. In addition to the ANOVAs, Pearson correlations were computed between task scores to explore individual differences. All analyses were carried out using the Jamovi software package [56].

Ethical Considerations

The study conformed to European and national legislation and fundamental ethical principles, including those reflected in the Charter of Fundamental Rights of the European Union, the European Convention on Human Rights and its Supplementary Protocols, and the World Medical Association Declaration of Helsinki. The protocol of the study was given ethics clearance by the Cyprus National Bioethics Committee (approval number: ΕΕΒΚ ΕΠ 2021.01.99). All participants read and signed informed consent forms either in person or digitally. Data were collected and maintained without any participant identification information.

Results

Overview

To investigate whether the 3 gamified tasks replicated the pattern of results reported in the literature for their traditional counterparts and to compare performance across modes of administration, we computed inferential statistics separately for each task. We report the findings in the next subsections.

Visual Search

For Visual Search, we carried out a repeated-measures ANOVA on RT with search type and display size as the within-subject variables and the administration condition as the between-subject variable.

Results revealed significant main effects for search type, $F_{1,72}$=787.12, P<.001, η^2=0.38; the display size, $F_{3,216}$=150.49, P<.001, η^2=0.17; and the administration condition, $F_{2,72}$=8.49, P=.001, η^2=0.02. There was also a significant interaction between search type and display size, $F_{3,216}$=145.42, P<.001, η^2=0.16. As seen in Figure 4, the interaction was driven by the presence of a display size effect (ie, RT increasing from 5 to 15 and then to 25 and further to 40 objects) in conjunction search trials but not in feature search trials. The significant effect for the administration condition was due to shorter RT in the VR-Lab (mean 1.24 seconds) condition than in the Desktop-Lab (mean 1.49 seconds; P<.001) and Desktop-Remote (mean 1.44 seconds; P=.008) conditions (Figure 5). The difference between the Desktop-Lab and Desktop-Remote conditions was not significant (P=1.00). Importantly, no significant interaction involving the administration condition was found. The effect size for search type (η^2=0.38) is large, suggesting a robust difference in RTs between feature and conjunction searches, consistent with prior literature. The effect size for display size (η^2=0.17) is also large, highlighting the substantial impact of increasing display size on RTs. In contrast, the main effect of the administration condition (η^2=0.02) has a small effect size, indicating that the mode of task administration had a statistically significant but relatively small influence on RTs. Finally, the interaction effect between search type and display size was associated with a large effect size (η^2=0.16), reinforcing the well-documented finding that increasing display size disproportionately affects conjunction searches compared with feature searches.

Figure 4. The average reaction time in the Visual Search task as a function of search type (feature search vs conjunction search) and display size (5, 15, 25, and 40 objects) for the VR-Lab, Desktop-Lab, and Desktop-Remote conditions. Results are shown in distinct panels for the VR-Lab, the Desktop-Lab, and the Desktop-Remote conditions. Error bars represent 95% confidence intervals.

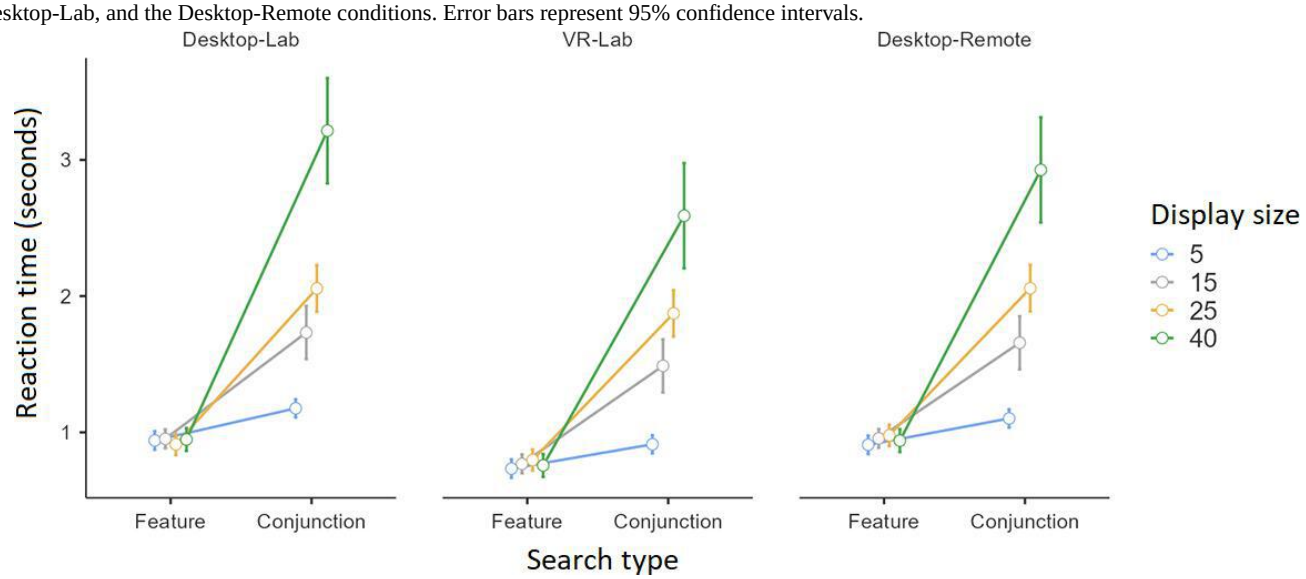

XSL•FO
RenderX

Figure 5. Mean reaction times in the Visual Search task, aggregated across search type and display size, for each of the 3 administration conditions: VR-Lab, Desktop-Lab, and Desktop-Remote. Overall, reaction time was shorter in the VR-Lab condition than in the Desktop-Lab and Desktop-Remote conditions. Error bars represent 95% confidence intervals.

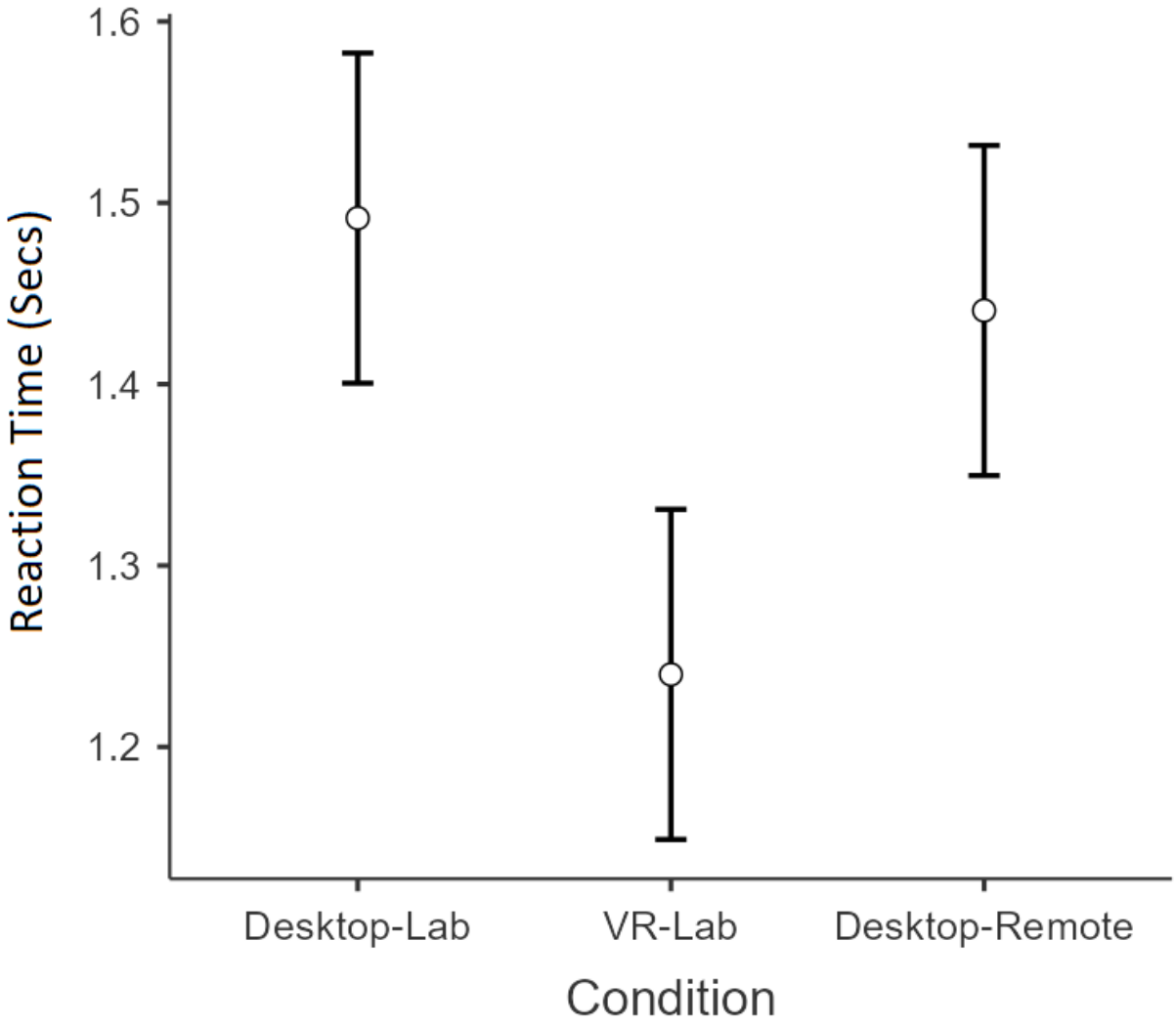

Whack-the-Mole

To analyze the accuracy data for the Whack-the-Mole task, we first computed for each participant the percentages of hits and false alarms. Based on the percentages, we computed the d' score for each participant. The d' score is a sensitive discrimination measure that reflects the degree to which participants accurately report the presence or absence of the target in the display. The d' score was calculated using the formula: d'=z (hit rate) − z (false alarm rate).

The average d' scores were 3.79 for the VR-Lab, 3.75 for the Desktop-Remote, and 3.62 for the Desktop-Lab condition (Figure 6), in line with the values typically reported in the literature [20,30]. A one-way ANOVA showed that the difference in d' across the VR-Lab, Desktop-Lab, and Desktop-Remote conditions was not significant, $F_{2,72}$=0.71, P=.49, η^2=0.02. In contrast, the analysis on mean RTs for hits only (ie, correct responses in Go trials) showed that participants were significantly slower in the Desktop-Remote condition (mean 0.64 second) than in the Desktop-Lab (mean 0.48 second) and VR-Lab (mean 0.41 second) conditions, both P values <.001 (Figure 7). The difference in RTs between the VR-Lab and the Desktop-Lab conditions was marginally significant (P=.07).

XSL•FO
RenderX

Figure 6. The average d' values in the Whack-the-Mole task across the 3 administration conditions. The d' is computed, according to Signal Detection Theory, by taking into account the percentage of hits (ie, correct responses to targets) and false alarms (ie, incorrect responses to distractors). Error bars depict 95% confidence intervals.

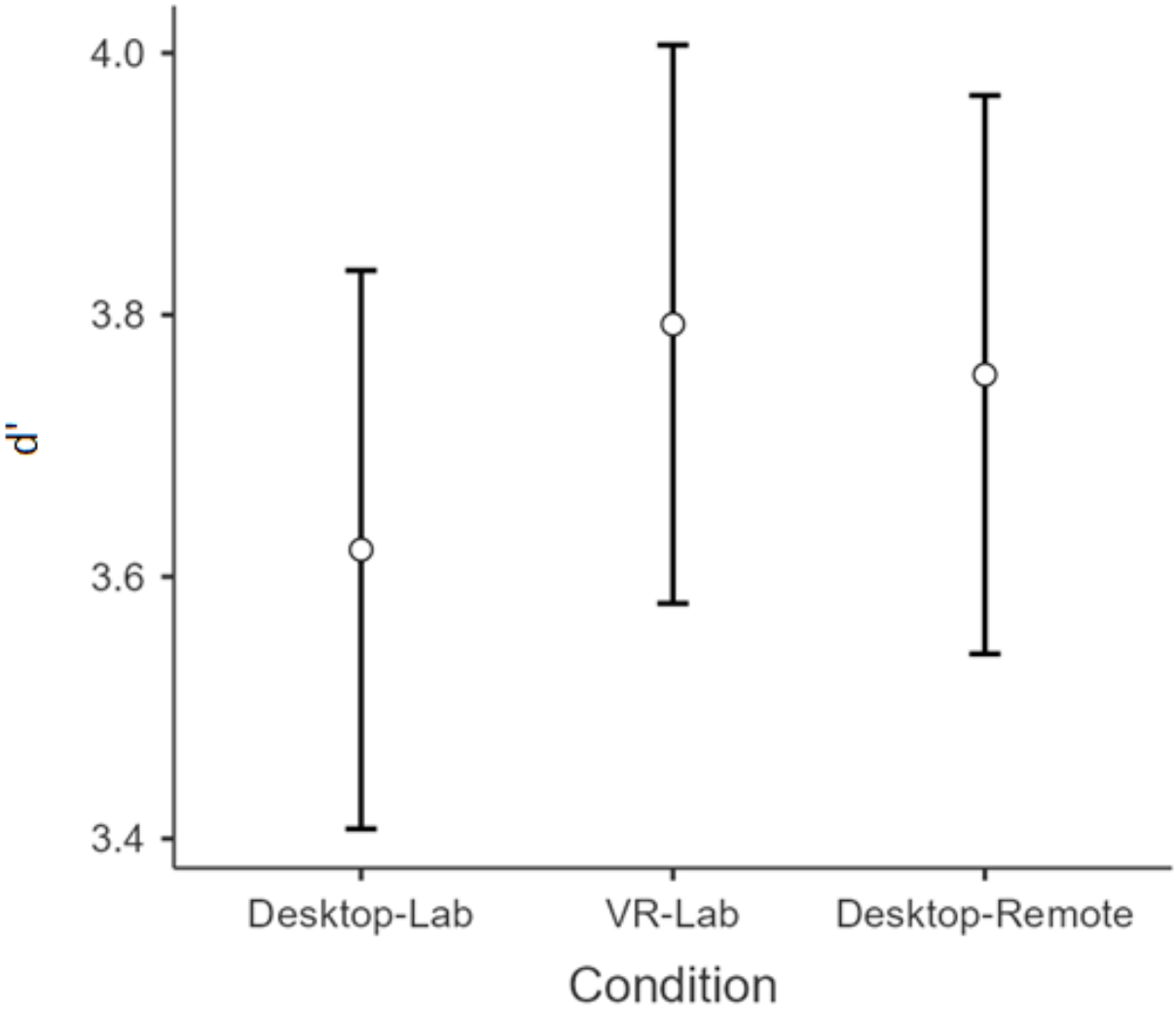

XSL•FO
RenderX

Figure 7. Average reaction time for Hits (ie, correct responses to targets) in the Whack-the-Mole task across the 3 administration conditions. Error bars depict 95% confidence intervals.

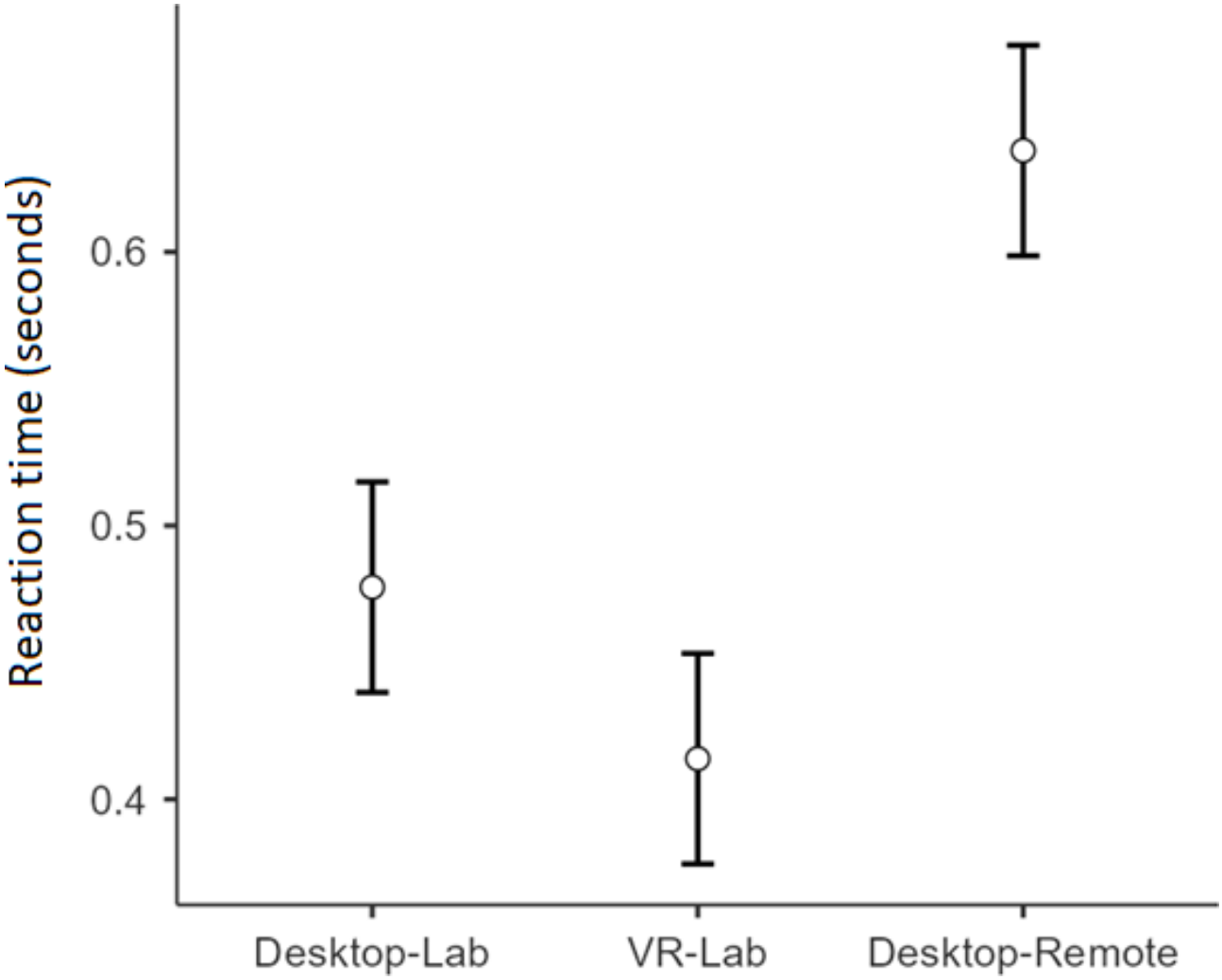

Corsi Test

As shown in Multimedia Appendix 2, the average span scores in the Corsi Test were 5.48 in the VR-Lab condition, 5.68 in the Desktop-Lab condition, and 5.24 in the Desktop-Remote condition. Despite the numeric difference across the 3 conditions, a one-way ANOVA showed that the effect of the administration condition on the Corsi span score was not significant, $F_{2,72}$=1.45, P=.24, η^2=0.04. These findings indicate that the administration condition had no impact on spatial working memory performance, as measured by the Corsi Test.

Individual Differences Within Attentional Tasks

To examine a possible relation between the 2 distinct attentional processes, that is, Visual Search and response inhibition, we carried out Pearson correlation analyses between the mean RT in feature search and conjunction search trials of the Visual Search task separately and the mean RT of Hits in the Whack-the-Mole task. For these, we averaged the RT across displays for each search type and we combined the data from the 3 administration conditions per task. We found a significant positive correlation between RT for Hits in the Whack-the-Mole task and RT in feature search trials of the Visual Search task, r_{73}=0.24; P=.04 (Figure 8). The correlation between RT for Hits in the Whack-the-Mole task and RT for the conjunction search trials of the Visual Search task was not significant, r_{73}=−0.04; P=.74.

XSL•FO
RenderX

Figure 8. Scatter plot depicting the correlation of Whack-the-Mole reaction time for Hits and Visual Search reaction time for Feature Search trials. The significant positive correlation indicates that participants who were faster to respond in Go trials of the Whack-the-Mole task tended to respond faster in the Feature Search trials of the Visual Search task.

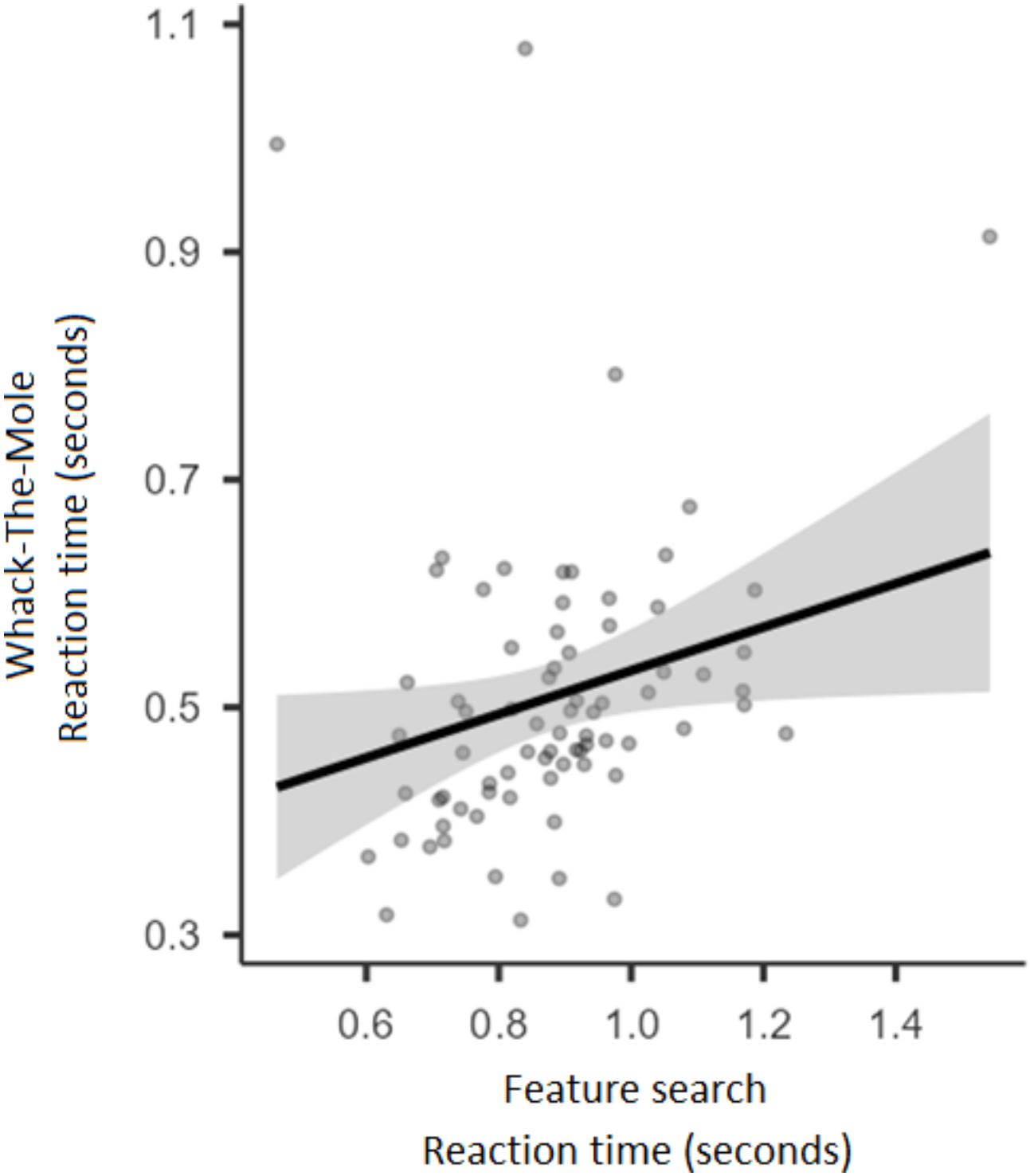

Discussion

Principal Results

The main objective of this study was to investigate whether the gamified cognitive tasks we developed yield the same patterns of results as their traditional versions. A secondary aim was to determine whether the mode of administration (immersive VR vs desktop VR) and the task administered remotely or in the lab affect performance outcomes.

With regard to the main objective, we found that our gamified tasks replicated the traditional patterns observed in previous studies. Specifically, in the Visual Search task, RTs were shorter for feature search trials than for conjunction search trials, aligning with the established finding that feature searches are overall easier [15]. Additionally, in conjunction search trials, RT increased significantly with display size, a finding known as the display size effect [15,17,18]. Conversely, RTs remained relatively constant across display sizes in feature search trials, demonstrating the well-established pop-out effect [15]. In the

XSL•FO
RenderX

Whack-the-Mole task, the overall response inhibition performance aligns with past studies indicating a typical d' score between 3 and 4, which reflects moderate sensitivity [20,49]. Our participants exhibited false alarms in about 10% of No-Go trials, consistent with the rate reported in the literature [20-22]. Furthermore, in assessing visual short-term memory, our findings showed that our participants had a span between 5 and 7, which is comparable with that reported by studies using the original version of the task [35], as well as those using modern electronic implementations delivered via computers or mobile devices [33].

While replicating the expected pattern of results across the 3 administration conditions, we also found some differences across these conditions. For example, performance in the Visual Search task was faster in the VR-Lab condition by 250 milliseconds than in the Desktop-VR condition. One possibility is that this advantage for immersive VR is due to diminished distraction resulting from blocking out extraneous environmental information when wearing an HMD. However, it could also be due to methodological differences between the 2 administration conditions. Specifically, participants in the Desktop-Lab condition responded by moving and clicking the mouse on the target object. Thus, RT encompassed the time to locate the target, the time to move the mouse, and the time to click. In contrast, in the VR-Lab condition, participants pointed the VR controller to the target object without requiring them to press any button. Thus, RT in this condition involved the time to locate the target and the time to move the controller in 3D space. Perhaps moving the controller in 3D space is performed faster than moving a mouse cursor on a screen. Or perhaps, the difference can be attributed to the need to click the mouse in the desktop condition but not in the VR condition. Notably, we found no difference between the Desktop-Lab and Desktop-Remote conditions, suggesting that differences in the environmental settings (eg, varying screen sizes and distances from the screen at participant homes) and environmental distractors that may be greater at home than the laboratory do not interfere significantly with the assessment of Visual Search abilities.

In the Whack-the-Mole task, performance in the VR-Lab condition was marginally faster by 70 milliseconds than in the Desktop-Lab condition. In contrast to the Visual Search task, in the Whack-the-Mole task, the response requirements were rather similar across the 2 administration conditions. That is, in the VR-Lab condition, participants pulled the trigger of the controller to hit the mole, while in the desktop versions, they clicked the mouse. Thus, taking the results of the 2 tasks together, it seems more likely that in the Whack-the-Mole task, the advantage of immersive VR is due to blocking out access to distracting information from the external environment. Nevertheless, we cannot safely exclude the possibility that the difference is due to subtle differences in the response mode. Perhaps the time needed to pull the trigger on the VR controller is shorter than clicking the mouse button. However, it should be noted that the performance in the Desktop-Lab condition was faster by 160 milliseconds than in the Desktop-Remote condition, despite the fact that the response mode in these 2 conditions was identical. This finding suggests that the differences in overall performance across administration conditions in the Whack-the-Mole task are not methodological.

A possible explanation for the RT differences across the administration conditions of the Whack-the-Mole task is offered by the Perceptual Load Theory. This theory was proposed by Lavie and Tsal [57,58] in the mid-1990s as a potential resolution to the early versus late selection debate in attention and perception. Lavie and Tsal argued that when a task is perceptually demanding (referred to as high load), people dedicate all processing resources to the relevant stimuli early on in the information-processing stream. This results in no interference from distracting irrelevant stimuli, as no spare resources are available to process them. In contrast, when the task is not perceptually demanding (ie, a low-load task), spare resources are available after processing the relevant stimuli. These resources, according to the theory, automatically spill into the processing of irrelevant stimuli. Thus, in low-load tasks, selection takes place late with irrelevant stimuli exerting an influence on performance [59-62]. Notably, in our own research using empirical methods and computational modeling [63], we have shown that inducing a narrow focus of attention may mediate whether distractor effects are present. That is, when the tasks force participants to zoom in their attention to stimuli on the screen, rather than having a more relaxed focus of attention, distracting information can exert no influence even if the task is of low load. Here, the Whack-the-Mole can be more easily classified as low load than high load. More importantly, it can be executed efficiently by maintaining a broader focus of attention, that is, by just focusing attention on a large part in the center of the screen rather than zooming in on a small stimulus. Carrying out the task this way allows for external distractors to slow down performance. If participants' homes include more environmental distractors than a laboratory, this would translate to more extraneous influence on performance and slowing down of responses. This could also account for the advantage of the VR-Lab condition as donning an HMD serves to effectively block environmental distractors that could influence performance.

A question that arises, though, is why there was no difference between the Desktop-Lab and Desktop-Remote conditions in the Visual Search task? On one hand, the conjunction search trials of the Visual Search task can be thought of as a high-low task that requires a narrow focus of attention, limiting resources to the processing of the relevant stimuli. This would explain why there were no differences across administration conditions. On the other hand, this cannot apply to the feature search trials. Feature search trials are more likely a low load task and, as such, would allow for the processing of distracting information. Yet, no differences were found across administration conditions. Therefore, it seems that further research, in which response modes and other methodological differences are controlled for, is needed to draw more definitive conclusions.

Comparison With Prior Work

This study examined the impact of administration mode (VR-Lab, Desktop-Lab, and Desktop-Remote) on cognitive performance across 3 tasks: Visual Search, Whack-the-Mole, and Corsi Test. Overall, the results demonstrate that while VR

XSL•FO
RenderX

tasks provide a highly immersive and engaging platform for cognitive assessment, the performance differences between VR and traditional desktop-based conditions are task-dependent and vary in magnitude.

Our results from the Visual Search task replicate the well-documented effects of search type and display size [14,15,19], with longer RT observed for conjunction searches than for feature searches and RT increasing with display size in conjunction search trials. Notably, participants in the VR-Lab condition exhibited significantly shorter RTs than those in the Desktop-Lab and Desktop-Remote conditions, suggesting that the immersive nature of VR may enhance visual attention and processing speed in this task. This finding aligns with previous research indicating that VR environments can improve cognitive task performance through increased engagement and spatial immersion [41]. Barrett et al [41] reported longer RT and higher fixation durations in VR compared with 2D environments, which they attributed to the added cognitive demands of immersion. Our results diverge from those of Barrett et al, as the VR-Lab condition here resulted in faster RT, potentially due to the optimized laboratory setup and controlled task parameters. Notably, these patterns of results were obtained with much fewer trials than typically included in the traditional versions of the tasks. For instance, while we ran our gamified Visual Search task with only 96 trials, we produced the same pattern of results as Huang and Pashler [64] and Bravo and Nakayama [65], who ran 1000 and 240 trials, respectively.

The Whack-the-Mole task revealed no significant differences in d' scores across administration conditions, indicating that accuracy in response inhibition tasks is consistent across VR and desktop modalities. These findings are consistent with those by Bhargava et al [42], who found that VR tasks achieved comparable accuracy to traditional assessments while enhancing engagement. However, RTs for Hits were significantly slower in the Desktop-Remote condition than in both the VR-Lab and Desktop-Lab conditions. This delay may reflect a lack of environmental standardization in remote settings, which has been noted as a limitation of remote cognitive testing [45]. The marginally faster RT in the VR-Lab condition than in the Desktop-Lab condition (P=.07) further highlights the potential for VR to enhance cognitive-motor coordination, as observed in other VR-based cognitive tasks [43]. In addition, while Shimi et al [20] and Casey et al [55] had 220 trials and 300 trials, respectively, in the Whack-the-Mole task they administered, we obtained the same results with only 100 trials.

The Corsi Test revealed no significant differences in average span scores across conditions, supporting the conclusion that spatial working memory is not influenced by the mode of administration. This aligns with prior findings that traditional and VR-based versions of the Corsi Test produce comparable results [31,33]. While the observed numeric differences were less and nonsignificant, the slightly lower scores in the Desktop-Remote condition may reflect environmental distractions or variability in participant equipment, which can affect remote testing reliability [45].

The positive correlation between RT for Hits in the Whack-the-Mole task and feature search trials in the Visual Search task suggests an underlying link between response inhibition and attentional processes, particularly in tasks involving rapid response execution. However, the absence of a significant correlation between RT in conjunction search trials and response inhibition suggests that these processes may operate independently under higher cognitive demands. These findings add to the growing literature on individual differences in attentional control [44] and further validate the use of gamified tasks in exploring these relationships.

Our findings emphasize the importance of task design and administration context in cognitive testing. While VR can enhance engagement and, in some cases, improve task performance, the added cognitive load of immersive environments may not universally benefit all cognitive processes [41,45]. The only minor differences in accuracy and spatial memory performance across administration modes highlight the robustness of these tasks to different delivery methods, including remote testing. However, slower RT in remote conditions underscores the need for careful standardization and participant instructions when deploying tasks outside laboratory settings.

Future studies should explore the impact of task complexity and individual differences on performance across VR and traditional modalities. Additionally, longitudinal research is needed to assess whether the observed benefits of VR in certain cognitive domains translate into practical advantages in real-world applications, such as rehabilitation and training.

Limitations

In this study, we did not assess participants' prior experience with VR. Previous exposure to VR could have influenced RTs, with participants who have used VR systems before potentially showing faster RTs due to greater familiarity with the technology. This factor should be considered in future studies, and including a questionnaire to assess past VR experience would be a valuable addition to control for this variable and better understand its impact on performance.

While this study focuses on the potential engagement benefits of VR-based cognitive tasks, it is important to acknowledge that VR can also introduce challenges. One key limitation is the potential for cybersickness, eye strain, and general fatigue associated with prolonged VR use. These factors can negatively impact participant performance [66,67] and may have contributed to variability in the results, particularly among participants who were less familiar with VR technology. Although we did not directly assess these issues by asking participants, no participants in the VR group reported any difficulties or discomfort with VR during or after their participation. While this suggests that cybersickness or fatigue was not a major confound, future studies could benefit from incorporating standardized self-report questionnaires or physiological measures to better isolate the impact of VR engagement on cognitive performance.

Notably, we did not include a paper-and-pencil comparison group, as our primary aim was to contrast the gamified VR tasks with traditional desktop-based digital assessments. Including a paper-and-pencil comparison could provide insights to

XSL•FO
RenderX

cross-modality generalizability, but it would also introduce additional variables unrelated to our core research objective, such as differences in response modalities and environmental influences. Nevertheless, we acknowledge that such a comparison could be valuable in future research to examine broader ecological validity, recognizing that certain tasks in our study (eg, Whack-the-Mole) are not feasible in a paper-based format.

Conclusions

This study demonstrated that gamified VR tasks can replicate well-established cognitive performance patterns using fewer trials, highlighting their potential as efficient and engaging tools for assessing attention, response inhibition, and visual short-term memory. By comparing performance across immersive VR, desktop-based, and remote administration conditions, we found that VR-based tasks maintained ecological validity and produced results consistent with traditional methods, suggesting that increased immersion does not compromise data quality. These advantages are particularly relevant for reducing participant fatigue and enhancing motivation, factors essential for high-quality cognitive assessment.

At the same time, VR introduces methodological variables that require further attention. In this study, we did not directly assess cybersickness, fatigue, or prior VR experience, which may influence individual performance. In addition, comparisons to paper-and-pencil tasks were not included, as the tasks examined here are typically administered in computerized form. Future research should explore these aspects systematically and investigate how task-specific features and environmental settings interact with immersion to affect cognitive outcomes. Overall, our findings support the use of immersive VR as a viable and scalable platform for cognitive assessment, especially in applications that benefit from high engagement and ecological realism.

Acknowledgments

This project has received partial funding from the European Union's Horizon 2020 Research and Innovation Programme (grant agreement number 739578) and the Government of the Republic of Cyprus through the Deputy Ministry of Research, Innovation, and Digital Policy.

Data Availability

The aggregated data file on which all analyses were carried out is freely available on the Open Science Framework at https://osf.io/e9ry2

Conflicts of Interest

None declared.

Multimedia Appendix 1
Example of a feature search trial used in the Visual Search task of this experimental study. In this trial, participants searched for a gray 2-sided axe (target) among 24 brown 2-sided axes (distractors). This is a trial with a display size of 25. The red outline is added for illustration purposes and was not present in the actual display viewed by participants.
[PNG File, 428 KB - formative_v9i1e65836_app1.png]

Multimedia Appendix 2
Mean span scores in the Corsi block-tapping task across the 3 administration conditions. Span scores reflect the longest sequence of targets correctly recalled by participants. Error bars represent 95% confidence intervals.
[PNG File, 22 KB - formative_v9i1e65836_app2.png]

References

1. Smith EE, Jonides J. Storage and executive processes in the frontal lobes. Science 1999 Mar 12;283(5408):1657-1661. [doi: 10.1126/science.283.5408.1657] [Medline: 10073923]
2. Miller EK, Cohen JD. An integrative theory of prefrontal cortex function. Annu Rev Neurosci 2001;24(1):167-202. [doi: 10.1146/annurev.neuro.24.1.167] [Medline: 11283309]
3. Eriksen BA, Eriksen CW. Effects of noise letters upon the identification of a target letter in a nonsearch task. Percept Psychophys 1974 Jan;16(1):143-149. [doi: 10.3758/BF03203267]
4. Posner MI, Petersen SE. The attention system of the human brain. Annu Rev Neurosci 1990;13:25-42. [doi: 10.1146/annurev.ne.13.030190.000325] [Medline: 2183676]
5. Mekler ED, Brühlmann F, Opwis K, Tuch AN. Disassembling gamification: the effects of points and meaning on user motivation and performance. Presented at: CHI '13: CHI Conference on Human Factors in Computing Systems; Apr 27 to May 2, 2013; Paris, France.
6. Lang A, Bradley SD, Park B, Shin M, Chung Y. Parsing the resource pie: using STRTs to measure attention to mediated messages. Media Psychol 2006 Nov;8(4):369-394. [doi: 10.1207/s1532785xmep0804_3]

XSL•FO
RenderX

7. Warm JS, Parasuraman R, Matthews G. Vigilance requires hard mental work and is stressful. Hum Factors 2008 Jun;50(3):433-441. [doi: 10.1518/001872008X312152] [Medline: 18689050]
8. Parsons TD. Virtual reality for enhanced ecological validity and experimental control in the clinical, affective and social neurosciences. Front Hum Neurosci 2015;9:660. [doi: 10.3389/fnhum.2015.00660] [Medline: 26696869]
9. Slater M, Sanchez-Vives MV. Enhancing Our Lives with Immersive Virtual Reality. Front Robot AI 2016 Dec 19;3:74. [doi: 10.3389/frobt.2016.00074]
10. Deterding S, Dixon D, Khaled R, Nacke L. From game design elements to gamefulness: defining" gamification". Presented at: Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments; Sep 28-30, 2011; Tampere, Finland.
11. Hamari J, Koivisto J, Sarsa H. Does gamification work?—A literature review of empirical studies on gamification. Presented at: 2014 47th Hawaii International Conference on System Sciences; Jan 6-9, 2014; Waikoloa, HI, USA.
12. Lee CC, Hsiao KL, Chen CC. Exploring the benefit and sacrifice factors of virtual reality gameplay. Front Psychol 2020;11:251. [doi: 10.3389/fpsyg.2020.00251] [Medline: 32210870]
13. Huang W, Roscoe RD, Johnson‐Glenberg MC, Craig SD. Motivation, engagement, and performance across multiple virtual reality sessions and levels of immersion. J Comput Assist Learn 2021 Jun;37(3):745-758. [doi: 10.1111/jcal.12520]
14. Mittelstaedt JM, Wacker J, Stelling D. VR aftereffect and the relation of cybersickness and cognitive performance. Virtual Real 2019 Jun;23(2):143-154. [doi: 10.1007/s10055-018-0370-3]
15. Wolfe JM. Major issues in the study of visual search: part 2 of “40 years of feature integration: special issue in memory of Anne Treisman”. Atten Percept Psychophys 2020 Feb;82(2):383-393. [doi: 10.3758/s13414-020-02022-1] [Medline: 32291612]
16. Morar SS, Macredie RD, Cribbin T. An investigation of visual cues used to create and support frames of reference and visual search tasks in desktop virtual environments. Virtual Real 2002 Oct 1;6(3):140-150. [doi: 10.1007/s100550200015]
17. Smilek D, Enns JT, Eastwood JD, Merikle PM. Relax! Cognitive strategy influences visual search. Vis Cogn 2006 Aug;14(4-8):543-564. [doi: 10.1080/13506280500193487]
18. Mueller ST. A partial implementation of the Bica Cognitive Decathlon Using the psychology experiment building language (PEBL). Int J Mach Conscious 2010 Dec;02(2):273-288. [doi: 10.1142/S1793843010000497]
19. Lumsden J, Edwards EA, Lawrence NS, Coyle D, Munafò MR. Gamification of cognitive assessment and cognitive training: a systematic review of applications and efficacy. JMIR Serious Games 2016 Jul 15;4(2):e11. [doi: 10.2196/games.5888] [Medline: 27421244]
20. Shimi A, Tsestou V, Hadjiaros M, Neokleous K, Avraamides M. Attentional skills in soccer: evaluating the involvement of attention in executing a goalkeeping task in virtual reality. Appl Sci (Basel) 2021 Jan;11(19):9341. [doi: 10.3390/app11199341]
21. Cragg L, Nation K. Go or no-go? Developmental improvements in the efficiency of response inhibition in mid-childhood. Dev Sci 2008 Nov;11(6):819-827. [doi: 10.1111/j.1467-7687.2008.00730.x] [Medline: 19046150]
22. Nigg J, Nikolas M, Friderici K, Park L, Zucker RA. Genotype and neuropsychological response inhibition as resilience promoters for attention-deficit/hyperactivity disorder, oppositional defiant disorder, and conduct disorder under conditions of psychosocial adversity. Dev Psychopathol 2007;19(3):767-786. [doi: 10.1017/S0954579407000387] [Medline: 17705902]
23. Allan NP, Hume LE, Allan DM, Farrington AL, Lonigan CJ. Relations between inhibitory control and the development of academic skills in preschool and kindergarten: a meta-analysis. Dev Psychol 2014 Oct;50(10):2368-2379. [doi: 10.1037/a0037493] [Medline: 25069051]
24. Schmitt LM, White SP, Cook EH, Sweeney JA, Mosconi MW. Cognitive mechanisms of inhibitory control deficits in autism spectrum disorder. J Child Psychol Psychiatry 2018 May;59(5):586-595. [doi: 10.1111/jcpp.12837] [Medline: 29052841]
25. Aron AR, Poldrack RA. The cognitive neuroscience of response inhibition: relevance for genetic research in attention-deficit/hyperactivity disorder. Biol Psychiatry 2005 Jun 1;57(11):1285-1292. [doi: 10.1016/j.biopsych.2004.10.026] [Medline: 15950000]
26. van Hulst BM, de Zeeuw P, Vlaskamp C, Rijks Y, Zandbelt BB, Durston S. Children with ADHD symptoms show deficits in reactive but not proactive inhibition, irrespective of their formal diagnosis. Psychol Med 2018 Nov;48(15):2515-2521. [doi: 10.1017/S0033291718000107] [Medline: 29415788]
27. Barker V, Romaniuk L, Cardinal RN, Pope M, Nicol K, Hall J. Impulsivity in borderline personality disorder. Psychol Med 2015 Jul;45(9):1955-1964. [doi: 10.1017/S0033291714003079] [Medline: 25600066]
28. Evans AH, Strafella AP, Weintraub D, Stacy M. Impulsive and compulsive behaviors in Parkinson’s disease. Mov Disord 2009 Aug 15;24(11):1561-1570. [doi: 10.1002/mds.22505] [Medline: 19526584]
29. Hirose S, Chikazoe J, Watanabe T, et al. Efficiency of go/no-go task performance implemented in the left hemisphere. J Neurosci 2012 Jun 27;32(26):9059-9065. [doi: 10.1523/JNEUROSCI.0540-12.2012] [Medline: 22745505]
30. Gutiérrez-Cobo MJ, Cabello R, Fernández-Berrocal P. The three models of emotional intelligence and performance in a hot and cool go/no-go task in undergraduate students. Front Behav Neurosci 2017;11:33. [doi: 10.3389/fnbeh.2017.00033] [Medline: 28275343]

XSL•FO
RenderX

31. Corsi PM. Human Memory and the Medial Temporal Region of the Brain. PhD Thesis.: ProQuest Information & Learning; 1973.
32. Claessen MHG, van der Ham IJM, van Zandvoort MJE. Computerization of the standard Corsi block-tapping task affects its underlying cognitive concepts: a pilot study. Appl Neuropsychol Adult 2015;22(3):180-188. [doi: 10.1080/23279095.2014.892488] [Medline: 25258029]
33. Siddi S, Preti A, Lara E, et al. Comparison of the touch-screen and traditional versions of the Corsi block-tapping test in patients with psychosis and healthy controls. BMC Psychiatry 2020 Jun 23;20(1):329. [doi: 10.1186/s12888-020-02716-8] [Medline: 32576254]
34. Hadjiaros M, Neokleous K, Schiza E, Matsangidou M, Avraamides MN, Pattichis CS. A game-based cognitive assessment for visuospatial tasks: evaluation in healthy adults. Presented at: 2021 IEEE 21st International Conference on Bioinformatics and Bioengineering (BIBE); Oct 25-27, 2021; Kragujevac, Serbia p. 1-5. [doi: 10.1109/BIBE52308.2021.9635507]
35. Kessels RP, van Zandvoort MJ, Postma A, Kappelle LJ, de Haan EH. The Corsi Block-Tapping Task: standardization and normative data. Appl Neuropsychol 2000;7(4):252-258. [doi: 10.1207/S15324826AN0704_8] [Medline: 11296689]
36. Dahlstrom-Hakki I, Alstad Z, Asbell-Clarke J, Edwards T. The impact of visual and auditory distractions on the performance of neurodiverse students in virtual reality (VR) environments. Virtual Real 2024 Mar;28(1). [doi: 10.1007/s10055-023-00933-6]
37. Hung SC, Ho AY, Lai IH, Lee CS, Pong AS, Lai FH. Meta-analysis on the effectiveness of virtual reality cognitive training (VRCT) and computer-based cognitive training (CBCT) for individuals with mild cognitive impairment (MCI). Electronics (Basel) 2020 Dec 18;9(12):2185. [doi: 10.3390/electronics9122185]
38. Slater M, Wilbur S. A framework for immersive virtual environments (FIVE): speculations on the role of presence in virtual environments. Presence Teleoperators Virtual Environ 1997 Dec;6(6):603-616. [doi: 10.1162/pres.1997.6.6.603]
39. Bailenson JN, Beall AC, Loomis J, Blascovich J, Turk M. Transformed social interaction: decoupling representation from behavior and form in collaborative virtual environments. Presence Teleoperators Virtual Environ 2004 Aug 1;13(4):428-441. [doi: 10.1162/1054746041944803]
40. Bailenson J, Patel K, Nielsen A, Bajscy R, Jung SH, Kurillo G. The effect of interactivity on learning physical actions in virtual reality. Media Psychol 2008 Sep 17;11(3):354-376. [doi: 10.1080/15213260802285214]
41. Barrett RCA, Poe R, O'Camb JW, et al. Comparing virtual reality, desktop-based 3D, and 2D versions of a category learning experiment. PLoS One 2022;17(10):e0275119. [doi: 10.1371/journal.pone.0275119] [Medline: 36201546]
42. Bhargava Y, Kottapalli A, Baths V. Validation and comparison of virtual reality and 3D mobile games for cognitive assessment against ACE-III in 82 young participants. Sci Rep 2024 Oct 13;14(1):23918. [doi: 10.1038/s41598-024-75065-1] [Medline: 39397120]
43. Faria AL, Pinho MS, Bermúdez I Badia S. A comparison of two personalization and adaptive cognitive rehabilitation approaches: a randomized controlled trial with chronic stroke patients. J Neuroeng Rehabil 2020 Jun 16;17(1):78. [doi: 10.1186/s12984-020-00691-5] [Medline: 32546251]
44. Tan NC, Lim JE, Sultana R, Quah JHM, Wong WT. A virtual reality cognitive screening tool based on the six cognitive domains. Alzheimers Dement (Amst) 2024;16(4):e70030. [doi: 10.1002/dad2.70030] [Medline: 39713248]
45. Neguţ A, Matu SA, Sava FA, David D. Task difficulty of virtual reality-based assessment tools compared to classical paper-and-pencil or computerized measures: a meta-analytic approach. Comput Human Behav 2016 Jan;54:414-424. [doi: 10.1016/j.chb.2015.08.029]
46. Dandurand F, Shultz TR, Onishi KH. Comparing online and lab methods in a problem-solving experiment. Behav Res Methods 2008 May;40(2):428-434. [doi: 10.3758/brm.40.2.428] [Medline: 18522052]
47. Gould SJJ, Cox AL, Brumby DP, Wiseman S. Home is where the lab is: a comparison of online and lab data from a time-sensitive study of interruption. Hum Comput 2015;2(1):45-67. [doi: 10.15346/hc.v2i1.4]
48. de Leeuw JR, Motz BA. Psychophysics in a web browser? Comparing response times collected with JavaScript and Psychophysics Toolbox in a visual search task. Behav Res 2016 Mar;48(1):1-12. [doi: 10.3758/s13428-015-0567-2]
49. Armitage J, Eerola T. Reaction time data in music cognition: comparison of pilot data from lab, crowdsourced, and convenience web samples. Front Psychol 2019;10:2883. [doi: 10.3389/fpsyg.2019.02883] [Medline: 31969849]
50. Bartneck C, Duenser A, Moltchanova E, Zawieska K. Comparing the similarity of responses received from studies in Amazon's Mechanical Turk to studies conducted online and with direct recruitment. PLoS One 2015;10(4):e0121595. [doi: 10.1371/journal.pone.0121595] [Medline: 25876027]
51. Hilbig BE. Reaction time effects in lab- versus web-based research: experimental evidence. Behav Res Methods 2016 Dec;48(4):1718-1724. [doi: 10.3758/s13428-015-0678-9] [Medline: 26542972]
52. Saunders DR, Bex PJ, Woods RL. Crowdsourcing a normative natural language dataset: a comparison of Amazon Mechanical Turk and in-lab data collection. J Med Internet Res 2013 May 20;15(5):e100. [doi: 10.2196/jmir.2620] [Medline: 23689038]
53. Sauce B, Matzel LD. The causes of variation in learning and behavior: why individual differences matter. Front Psychol 2013;4:395. [doi: 10.3389/fpsyg.2013.00395] [Medline: 23847569]
54. Segen V, Avraamides M, Slattery T, Colombo G, Wiener JM. Comparable performance on a spatial memory task in data collected in the lab and online. PLoS One 2021;16(11):e0259367. [doi: 10.1371/journal.pone.0259367] [Medline: 34843521]

XSL•FO
RenderX

55. Casey BJ, Trainor RJ, Orendi JL, et al. A developmental functional MRI study of prefrontal activation during performance of a go-no-go task. J Cogn Neurosci 1997 Nov;9(6):835-847. [doi: 10.1162/jocn.1997.9.6.835] [Medline: 23964603]
56. Jamovi (version 23) [computer software]. The Jamovi Project. 2022. URL: https://www.jamovi.org [accessed 2025-06-30]
57. Lavie N, Tsal Y. Perceptual load as a major determinant of the locus of selection in visual attention. Percept Psychophys 1994 Mar;56(2):183-197. [doi: 10.3758/BF03213897]
58. Lavie N. Perceptual load as a necessary condition for selective attention. J Exp Psychol Hum Percept Perform 1995;21(3):451-468. [doi: 10.1037//0096-1523.21.3.451]
59. Cartwright-Finch U, Lavie N. The role of perceptual load in inattentional blindness. Cognition 2007 Mar;102(3):321-340. [doi: 10.1016/j.cognition.2006.01.002] [Medline: 16480973]
60. Lavie N, Hirst A, de Fockert JW, Viding E. Load theory of selective attention and cognitive control. J Exp Psychol Gen 2004 Sep;133(3):339-354. [doi: 10.1037/0096-3445.133.3.339] [Medline: 15355143]
61. Lavie N. Distracted and confused?: Selective attention under load. Trends Cogn Sci (Regul Ed) 2005 Feb;9(2):75-82. [doi: 10.1016/j.tics.2004.12.004]
62. Rees G, Frith CD, Lavie N. Modulating irrelevant motion perception by varying attentional load in an unrelated task. Science 1997 Nov 28;278(5343):1616-1619. [doi: 10.1126/science.278.5343.1616] [Medline: 9374459]
63. Neokleous K, Shimi A, Avraamides MN. Modeling the effects of perceptual load: saliency, competitive interactions, and top-down biases. Front Psychol 2016;7:1. [doi: 10.3389/fpsyg.2016.00001] [Medline: 26858668]
64. Huang L, Pashler H. Attention capacity and task difficulty in visual search. Cognition 2005 Jan;94(3):B101-B111. [doi: 10.1016/j.cognition.2004.06.006] [Medline: 15617666]
65. Bravo MJ, Nakayama K. The role of attention in different visual-search tasks. Percept Psychophys 1992 Sep;51(5):465-472. [doi: 10.3758/BF03211642]
66. Alaeifard M, Safaei M. Head movement patterns as predictors of cybersickness in virtual reality games. Int J Adv Hum Comput Interact 2024 Jul 19;2(2):1. [doi: 10.1016/j.cognition.2004.06.006]
67. Weech S, Kenny S, Lenizky M, Barnett-Cowan M. Narrative and gaming experience interact to affect presence and cybersickness in virtual reality. Int J Hum Comput Stud 2020 Jun;138:102398. [doi: 10.1016/j.ijhcs.2020.102398]

Abbreviations

ACE-III: Addenbrooke's Cognitive Examination—III
HMD: head-mounted display
RT: reaction time
VR: virtual reality

Edited by A Mavragani, L Riedemann; submitted 27.08.24; peer-reviewed by B Concannon, JF Costa, L Mendes; revised version received 05.05.25; accepted 25.05.25; published 29.07.25.

Please cite as:
Hadjiaros M, Shimi A, Neokleous K, Pattichis C, Avraamides M
Virtual Reality Gamification of Visual Search, Response Inhibition, and Visual Short-Term Memory Tasks for Cognitive Assessment: Experimental Study
JMIR Form Res 2025;9:e65836
URL: https://formative.jmir.org/2025/1/e65836
doi:10.2196/65836

© Marios Hadjiaros, Andria Shimi, Kleanthis Neokleous, Constantinos Pattichis, Marios Avraamides. Originally published in JMIR Formative Research (https://formative.jmir.org), 29.7.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Visualization of FDA Adverse Drug Reaction Reports: Development and Usability Study of the VisDrugs Web Server

Renjun Yang[1,2], PhD; Nuoya Yin[1,2], PhD; Yang Zhang[3], PhD; Francesco Faiola[1,2], PhD

[1]State Key Laboratory of Environmental Chemistry and Ecotoxicology, Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences, 18 Shuangqing Road, Haidian District, Beijing, China

[2]College of Resources and Environment, University of Chinese Academy of Sciences, Beijing, China

[3]Department of Pharmacy, Beijing Friendship Hospital, Capital Medical University, Beijing, China

Corresponding Author:
Francesco Faiola, PhD
State Key Laboratory of Environmental Chemistry and Ecotoxicology, Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences, 18 Shuangqing Road, Haidian District, Beijing, China

Abstract

Background: Adverse drug reactions (ADRs) are a major concern in drug safety, and the FDA Adverse Event Reporting System (FAERS) provides valuable ADR data. However, analyzing FAERS data is complex and requires bioinformatics expertise. Despite the vast amount of ADR data available, there is a lack of user-friendly tools that enable efficient visualization and comparison of ADRs for researchers and health care professionals.

Objective: This study aimed to develop VisDrugs, a web-based platform that simplifies ADR visualization and comparison using FAERS data. The platform was designed to assist researchers and clinicians in assessing drug safety through interactive and interpretable graphical representations of ADR patterns.

Methods: FAERS data were extracted in the American Standard Code for Information Interchange (ASCII) format, covering the period from Q3 (third quarter) 2014 to Q3 2024. About 2,700,000 reports from health care professionals, where only a single drug was implicated, were aggregated and processed using R for statistical analysis and visualization. The results are presented on a web-based platform for web-based analysis. The platform generates pie charts to visualize the most frequently reported ADRs, which are represented and analyzed using preferred terms based on the Medical Dictionary for Regulatory Activities (MedDRA) and forest plots illustrating reporting odds ratios (RORs) for these ADRs.

Results: Using Paxlovid (COVID-19 treatment) and hydroxychloroquine (anti-malaria drug) as case studies, we benchmarked VisDrugs using reports for Paxlovid (n=16,708) and hydroxychloroquine (n=6150). Paxlovid was most frequently associated with “COVID-19” (ROR=47.26, 95% CI 45.22 - 49.40) and “dysgeusia” (ROR=59.65, 95% CI 55.56 - 64.03). Hydroxychloroquine showed strong associations with “retinal toxicity” (ROR=738.48, 95% CI 583.45 - 934.71), “retinopathy” (ROR=412.27, 95% CI 344.73 - 493.03), and “cardiotoxicity” (ROR=48.36, 95% CI 38.86 - 60.19). In subgroup analyses, female patients had significantly higher risks of retinopathy (3.24-fold) and cardiomyopathy (13.82-fold) compared to male patients, while patients aged >50 years had higher risks of retinopathy (4.20-fold) and cardiomyopathy (7.84-fold) compared to those ≤50 years. All differences were statistically significant (*z* test, $P<.01$). The majority of findings align with existing research, thereby validating the platform’s utility. Clinical personnel have evaluated and refined the platform based on user feedback, confirming its efficacy in visualizing complex ADR data and identifying adverse effects across various drug subgroups.

Conclusions: VisDrugs is a valuable tool for ADR analysis, offering an intuitive interface for exploring FAERS data. By visualizing and comparing ADRs, it helps researchers and health care providers assess drug safety efficiently. The platform’s demographic analysis features add insights into ADR variations by age and gender, supporting drug safety research. In the future, the website will include more subgroup or condition filtering options, offering personalized ADR analysis and comparison features to meet the diverse research needs of users.

(JMIR Form Res 2025;9:e71519) doi:10.2196/71519

KEYWORDS

FDA Adverse Event Reporting System; adverse drug reactions; drug safety; data visualization; website; FAERS; ADR; Food and Drug Administration

XSL•FO
RenderX

Introduction

Adverse drug reactions (ADRs) are unintended responses that are detected in patients after the use of drugs for the prophylaxis, diagnosis, or treatment of a disease at doses normally used [1]. ADRs represent an important public health challenge, contributing to significant hospital admissions and deaths worldwide [2,3]. The US Food and Drug Administration (FDA) Adverse Event Reporting System (FAERS) provides a critical database that captures reports of adverse events linked to pharmaceutical products [4,5]. With millions of reports, FAERS serves as an invaluable resource for evaluating drug safety and post-market surveillance [6-12]. FAERS is continuously updated and includes data from a wide range of sources, including health care professionals, patients, and drug manufacturers [13]. Each report in the FAERS database provides information on a specific adverse event associated with a drug, including ADR description, the drug involved, the patient's age and sex, and other information.

Despite their vast potential for drug safety analysis, FAERS data present several challenges. The large record number poses a significant hurdle to traditional analysis methods. Moreover, the data are often unstructured and contain a considerable amount of noise, requiring sophisticated data-cleaning techniques [14]. For researchers to interpret these data effectively, a solid foundation in bioinformatics and statistical modeling is often required. The need for simplified, accessible tools for ADR analysis has become increasingly apparent. Although several web-based user interfaces, such as OpenVigil FDA and VigiAccess, facilitate access to FAERS data, conducting a detailed comparison of ADRs still necessitates performing calculations on the obtained data using bioinformatics methods [15-18]. Currently, there is no widely available web platform that offers an easy, interactive way to analyze FAERS data, especially for comparing the adverse effects of different drugs in a straightforward manner. This gap creates a barrier for researchers, clinicians, and policy makers who need quick insights into drug safety but lack the technical expertise to handle large-scale pharmacovigilance data.

In response to this challenge, we have developed the VisDrugs website, an interactive platform designed to simplify the process of analyzing drug-related adverse effects using FAERS data. The VisDrugs platform includes a comparison feature that allows users to input 2 groups of drugs and analyze the differences in their safety profiles with respect to specific adverse events. Furthermore, it conducts analyses by subgroups of age and gender to evaluate the effects of these variables on specific reactions. Visdrugs addresses a critical gap in the current landscape of drug safety analysis tools. By leveraging the extensive resources of the FAERS database and presenting them through a user-friendly interface, it enables a broader audience to engage with drug safety data.

Methods

Website Framework

The VisDrugs website is designed to provide 1-click analysis of FDA ADR data [19]. Users can easily use its features through any web browser, making it simple and convenient to use. The FDA ADR dataset on the site is updated annually in alignment with the latest updates from the FDA FAERS database.

VisDrugs operates on the Tencent Cloud server platform, featuring a CentOS 7.6 64-bit system with an 8-core CPU, 32 GB cache memory, 320 GB disk storage, and 22 Mbps bandwidth, supporting simultaneous web-based analyses by many researchers at the same time.

VisDrugs uses Java (version 8), Spring Boot, and MyBatis Plus to provide server-side API support. The visual frontend of the website is created using JavaScript, Vue (version 3), and Element Plus. Apache (version 9.0) is used as the web server, and nginx is used for load balancing. The MySQL database is used for backend storage of sequences, taxonomy, and user data.

The core 1-click analysis feature is powered by R language, which is a good choice for statistical analysis. The system minimizes redundant information, such as duplicated and incomplete reports, during data preprocessing, enhancing data analysis efficiency. VisDrugs provides researchers with a robust and efficient platform for FDA ADR data exploration and analysis.

Data Retrieving, Preprocessing, and Visualization

Data retrieval, preprocessing, and visualization were performed, as depicted in Figure 1. In detail, ADR data were extracted in ASCII format from the FDA database, covering the period from Q3 (third quarter) 2014 to Q1 2024. Data prior to Q3 2014 were not included mainly because they were from a time when the "DRUGyyQq.TXT" file did not contain active ingredient information for the drugs, and the data were less standardized. Starting from Q3 2014, the inclusion of drug active ingredient information and the improved data structure allowed for more accurate visualization of the results.

XSL•FO
RenderX

Figure 1. Data preprocessing flow and interactive interface operation flowchart for the VisDrugs website. ASCII: American Standard Code for Information Interchange; FAERS: Food and Drug Administration Adverse Event Reporting System; GUI: graphical user interface.

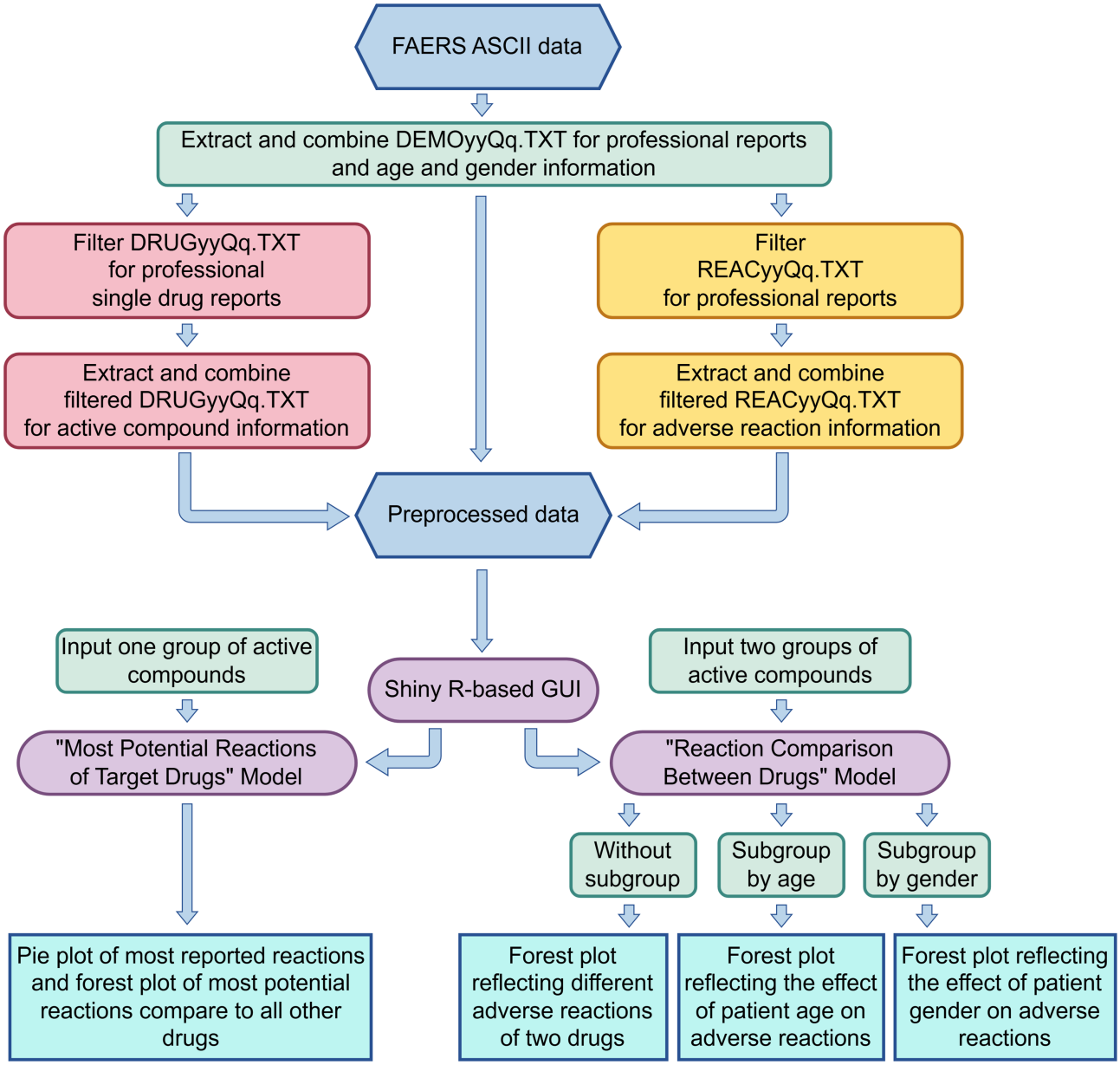

The data preprocessing step primarily used 3 key files: the DEMOyyQq.TXT, which contains basic case information; the REACyyQq.TXT, which includes ADR details for the patients; and the DRUGyyQq.TXT, which holds information on drug active ingredients (“Product active ingredient”) and the reported role of the drug in the adverse event (“Code for drug’s reported role in event”).

For the DEMOyyQq.TXT file, we filtered data over the past decade. Specifically, we selected data reported by 4 types of health care professionals (HP, health professional [HP]; physician [MD]; other health professional [OT]; and pharmacist [PH]) based on the abbreviation for the reporter’s type of occupation. We then aggregated quarterly data over the years to create a comprehensive DEMOyyQq.TXT dataset covering 2014 - 2024. This dataset includes case ID, patient gender and age, and the occupation type of the reporter, which can be used for further interactive analysis features on the website. Additionally, we summarized the case data to filter relevant reports from REACyyQq.TXT and DRUGyyQq.TXT files.

Next, we aggregated the cases from REACyyQq.TXT reported by professionals, with the ADRs defined using preferred term (PT)–level medical terminology describing the event, using the Medical Dictionary for Regulatory Activities (MedDRA). For the DRUGyyQq.TXT file, we focused on the cases reported by professionals, extracting the active ingredients that were identified as the primary suspect drug of each report. For the aggregate files, duplicates were removed by keeping only the latest version of each case (based on the FDA’s primary ID and version number). Records missing essential information, such as the drug’s active substance name or ADR term, were excluded.

During visualization, the R language was used to analyze the preprocessed datasets, performing statistical calculations and generating graphs to display potential ADRs. In simpler terms,

XSL•FO
RenderX

our self-developed ShinyR program performs statistical analysis on the aggregated preprocessed data and creates pie charts based on the total number of ADRs. Additionally, for drug ADR analysis, we use the commonly used method of reporting odds ratio (ROR) [20] to generate forest plots.

Standard MedDRA PTs were used to calculate potential ADRs, and active substances, which are standard chemical names, were applied as drug names to ensure consistency. The database will be updated yearly to incorporate the latest FAERS releases.

Calculation of Reporting Odds Ratio and Statistical Significance Analysis

The ROR for a selected drug compared to all other drugs was calculated as reported by Rothman et al [21]:

$$ROR_j = \frac{N_{A,j}/N_{A,\neg j}}{N_{\neg A,j}/N_{\neg A,\neg j}}$$

where ROR_j is the ROR for ADRj associated with the selected drug A. $N_{A,j}$ is the number of reports where drug A is associated with ADRj. $N_{A,\neg j}$ is the number of reports where drug A is reported but without ADRj. $N_{\neg A,j}$ is the number of reports where ADRj is reported but drug A is not involved. $N_{\neg A,\neg j}$ is the number of reports where neither drug A nor ADRj is reported.

The standard error of the log-transformed ROR is given by:

$$SE(\ln ROR_j) = \sqrt{\frac{1}{N_{A,j}} + \frac{1}{N_{A,\neg j}} + \frac{1}{N_{\neg A,j}} + \frac{1}{N_{\neg A,\neg j}}}$$

On the log scale, the CI is given by:

$$\ln ROR_j \pm 1.96 \times SE(\ln ROR_j)$$

To obtain the final CI for ROR, the following equation is used:

$$CI_{ROR} = (e^{\ln ROR_j - 1.96 \times SE(\ln ROR_j)}, e^{\ln ROR_j + 1.96 \times SE(\ln ROR_j)})$$

In the forest plot, an ROR value was considered statistically significant if the lower bound of its log2-transformed 95% CI was greater than zero. To evaluate whether the RORs differed significantly between 2 subgroups (eg, age or sex), we applied a *z* test based on the log-transformed RORs and their standard errors.

The z statistic was calculated as:

$$z = \frac{\ln(ROR_1) - \ln(ROR_2)}{\sqrt{SE_1^2 + SE_2^2}}$$

The resulting 2-sided *P* value for the *z* test was computed as:

$$p = 2 \times (1 - \Phi(|z|))$$

where Φ denotes the cumulative distribution function of the standard normal distribution. A difference of at least 1.5-fold in ROR values between the 2 subgroups, combined with a *P* value<.05, was considered statistically significant.

To ensure comprehensive reporting and transparency, a detailed checklist following the Software/Tools Papers guidelines is provided (see Checklist 1).

Ethical Considerations

This study involved secondary analysis of publicly available, deidentified data from the US FDA FAERS. No identifiable personal information was accessed or used. As the dataset is anonymized and contains no patient-level identifiers, this study was exempt from institutional review board approval in accordance with local regulations and ethical guidelines. The analysis complied with all applicable data use policies and ethical standards for research involving publicly available datasets.

Results

Analysis for Most Potential Reactions of Target Drugs

The VisDrugs website offers 2 selectable methods for analyzing ADRs. In the "Most Potential Reactions of Target Drugs" mode, users only need to perform 2 actions on the graphical user interface to view the ADRs most relevant to drug-of-interest's usage: (1) type to search for the active ingredient name of the drug-of-interest. The relevant active ingredient will appear in the search selection box, where users can choose either a single active ingredient or multiple active ingredients as a whole for analysis and (2) name the target drug group and click "submit."

After submitting, the webpage will generate two charts: (1) a pie chart showing the top 15 ADRs of the selected drug and (2) a forest plot comparing the top 15 ADRs from the pie chart that shows whether they are more likely caused by the selected drug or other drugs. Disproportionality analysis, which estimates the ROR, is used in the forest plot for detecting spontaneous signals, establishing the statistical link between drug exposure and a particular ADR [20]. The ROR, calculated using the case or non-case method, represents the odds ratio of a specific ADR relative to all other ADRs for each drug [21].

Additionally, four download buttons will be provided for users to download images or raw data: (1) high-definition pie chart of the drug's ADRs; (2) high-definition forest plot of the drug's ADRs; (3) an Excel table of drug ADR frequency statistics (all reactions); and (4) an Excel table of the raw data for the forest plot.

It is important to note that when typing to search for a drug, users should enter the active ingredient name of the drug. Active ingredients are more standardized than drug names and help prevent incorrect matches. Users must also select the relevant active ingredient from the list, as only selected drugs indicate that the site includes related data for that drug. Furthermore, the website provides a user guide where users can click to view detailed instructions on what information should be entered in each section and how to interpret the results.

Overall, the "Most Potential Reactions of Target Drugs" analysis function on the website provides a highly convenient method for analyzing the most common ADRs of drugs.

Application Examples: Analyzing the Most Commonly Reported ADRs of Paxlovid and Hydroxychloroquine

We use Paxlovid, a special antiviral drug for COVID-19, and the antimalarial drug hydroxychloroquine, which was initially considered to offer therapeutic benefits in COVID-19 patients [22], but withdrew from use in COVID-19 treatment by FDA in June 2020 [23], as examples to demonstrate the "Most Potential Reactions of Target Drugs" feature on the website. Paxlovid contains 2 active ingredients: Nirmatrelvir and Ritonavir [24-26]. In this example, given the focus on Paxlovid, the analysis was conducted using the active ingredients "NIRMATRELVIR/RITONAVIR", which represent a drug

XSL•FO
RenderX

comprising both components, to facilitate a comprehensive examination of potential synergistic adverse effects (see Figure S1 in Multimedia Appendix 1 for detailed tool settings). For hydroxychloroquine, both "hydroxychloroquine" and its sulfate form, "hydroxychloroquine sulfate", were applied for analysis (see Figure S2 in Multimedia Appendix 1 for detailed tool settings).

For Paxlovid, from the pie chart and forest plot (Figure 2A and C), we observe that in the top 15 ADRs, 12 ADRs occur more frequently in Paxlovid compared to other drugs. Patients using Paxlovid are more frequently reported to experience disease recurrence, dysgeusia (COVID-19), and taste disorder (Figure 2C). In fact, the presence of COVID-19 is consistent with Paxlovid's indication, and disease recurrence is a very common phenomenon in COVID-19 cases [27,28]. Dysgeusia and taste disorders are also the most common ADRs associated with COVID-19. Other reported ADRs, such as cough, vomiting, and headache, are also commonly associated with COVID-19, indicating that the ADR reports for Paxlovid can be viewed as part of the disease symptoms to its indication, COVID-19 [29].

XSL•FO
RenderX

Figure 2. Most relevant adverse drug reactions (ADRs) associated with Paxlovid and hydroxychloroquine based on FAERS (FDA Adverse Event Reporting System) reports analyzed using the VisDrugs system. Pie charts show the top 15 ADRs reported for Paxlovid (**A**) and hydroxychloroquine (**B**). Forest plots comparing the top 15 ADRs of Paxlovid (**C**) and hydroxychloroquine (**D**) with those of all other drugs. The black diamonds indicate the log2-transformed reporting odds ratios (RORs), and the green lines represent the corresponding log2-transformed 95% CIs. TR: target reaction.

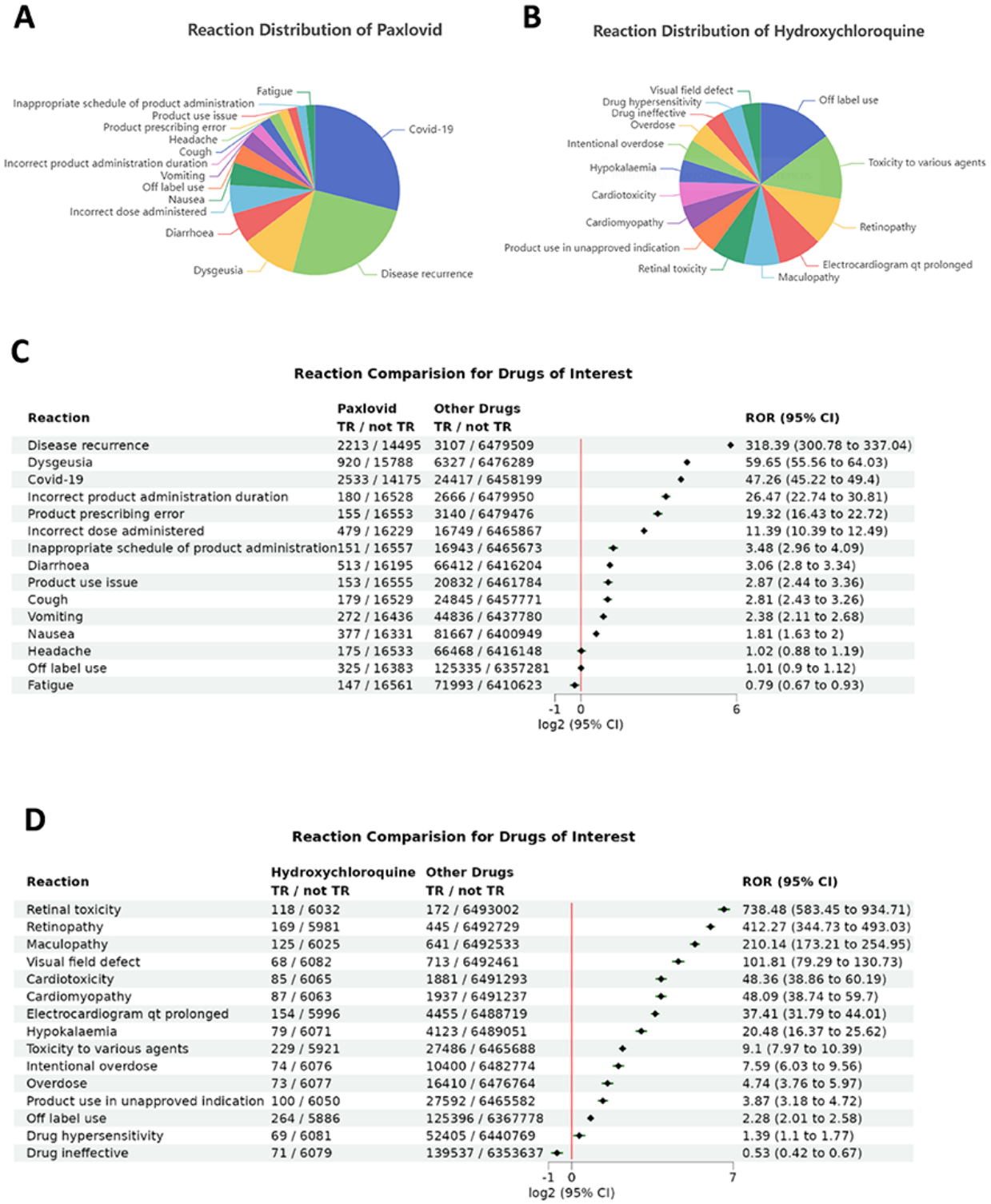

This situation in the ADR analysis may occur because health care professionals may report both disease symptoms and drug-related side effects as ADRs when submitting data to FAERS. This highlights the importance of considering the drug's indication when analyzing ADRs in FAERS. When a drug's major ADRs align with its intended indication, it might be reasonable to conclude that the drug does not cause additional severe ADRs. Furthermore, the analysis results above are very similar to those obtained by Li et al [29], who used FAERS data to analyze Paxlovid-related ADRs from health care professionals, confirming the feasibility of the analysis method provided by this website.

XSL•FO
RenderX

For hydroxychloroquine, we observe that the most frequently reported ADRs include retinal issues (such as retinopathy, maculopathy, visual field defects, and retinal toxicity), cardiac issues (such as prolonged QT on electrocardiogram and cardiomyopathy), and hypokalemia, which may lead to arrhythmia (Figure 2B). The forest plot also shows that hydroxychloroquine is indeed more likely to cause these ADRs compared to other drugs (Figure 2D).

This finding aligns with existing research, as hydroxychloroquine has been shown in epidemiological studies to pose a potential risk for retinopathy [30,31]. Additionally, cardiac disorders are a significant side effect of hydroxychloroquine. It has been reported that plenty of severely ill COVID-19 patients treated with hydroxychloroquine experienced a corrected QT interval increase [32,33]. In addition, case reports have shown that hydroxychloroquine overdose can lead to hypokalemia [34]. Given that FAERS analysis has highlighted the risk of hypokalemia associated with hydroxychloroquine, this adverse effect warrants closer attention.

Interestingly, hydroxychloroquine was originally an antimalarial drug, but the top ADRs seem not related to malaria, indicating severity of retinal and cardiac issues caused by hydroxychloroquine are of much greater concern to health care providers than any potential ADRs related to malaria itself.

Reaction Comparison Between Drugs

The “Reaction Comparison Between Drugs” mode allows users to compare the occurrence of specific ADRs between 2 groups of drugs. The following steps are required to perform this analysis: (1) input and select the active ingredient name of the target drug; (2) leave “Control Drugs” as “OTHER DRUGS” to compare the target drug with all other drugs, or delete “OTHER DRUGS” and input the active ingredient name of a specific control drug to compare the ADRs of 2 particular drugs; (3) input and select the ADR names of interest; (4) choose whether to group the analysis by age and gender, or conduct the analysis without grouping; and (5) name both the target drug group and the control drug group, then click “Submit.”

This analysis method allows for a focused comparison of specific ADRs between drugs of the same class, facilitating researchers in identifying safer options within the same drug category.

Application Example: Comparing ADRs of Hydroxychloroquine and Paxlovid

In this example, the user inputs “hydroxychloroquine” in the target drug field, selecting both “hydroxychloroquine” and “hydroxychloroquine sulfate”. In the control drug field, we use “NIRMATRELVIR/RITONAVIR”. The target reaction field is filled with the most potential ADRs of hydroxychloroquine found in Figure 2, including retinal toxicity (retinopathy, maculopathy, and visual field defect), cardiac ADRs (cardiomyopathy and electrocardiogram QT prolongation), and hypokalemia, which may lead to arrhythmias.

In the mode without subgroup analysis (Figure 3A), upon submitting the data, the results in the forest plot show that compared to the control drug Paxlovid, hydroxychloroquine indeed tends to cause the aforementioned 7 ADRs (Figure 3B). Conversely, the frequency of these reactions with Paxlovid is extremely low. Among 15,850 cases of Paxlovid use, only a limited number of cases reported related ADRs, suggesting that Paxlovid induces less retinal toxicity, cardiomyopathy, or hypokalemia compared to hydroxychloroquine.

XSL•FO
RenderX

Figure 3. Comparison of retinopathy, cardiomyopathy, and hypokalemia potential of hydroxychloroquine and Paxlovid. (**A**) User input example under the Reaction Comparison Between Drugs model. (**B**) Forest plot comparing the adverse drug reactions (ADRs) of interest between hydroxychloroquine and Paxlovid groups. The black diamonds indicate the log2-transformed reporting odds ratios (RORs), and the green lines represent the corresponding log2-transformed 95% CIs. TR: target reaction.

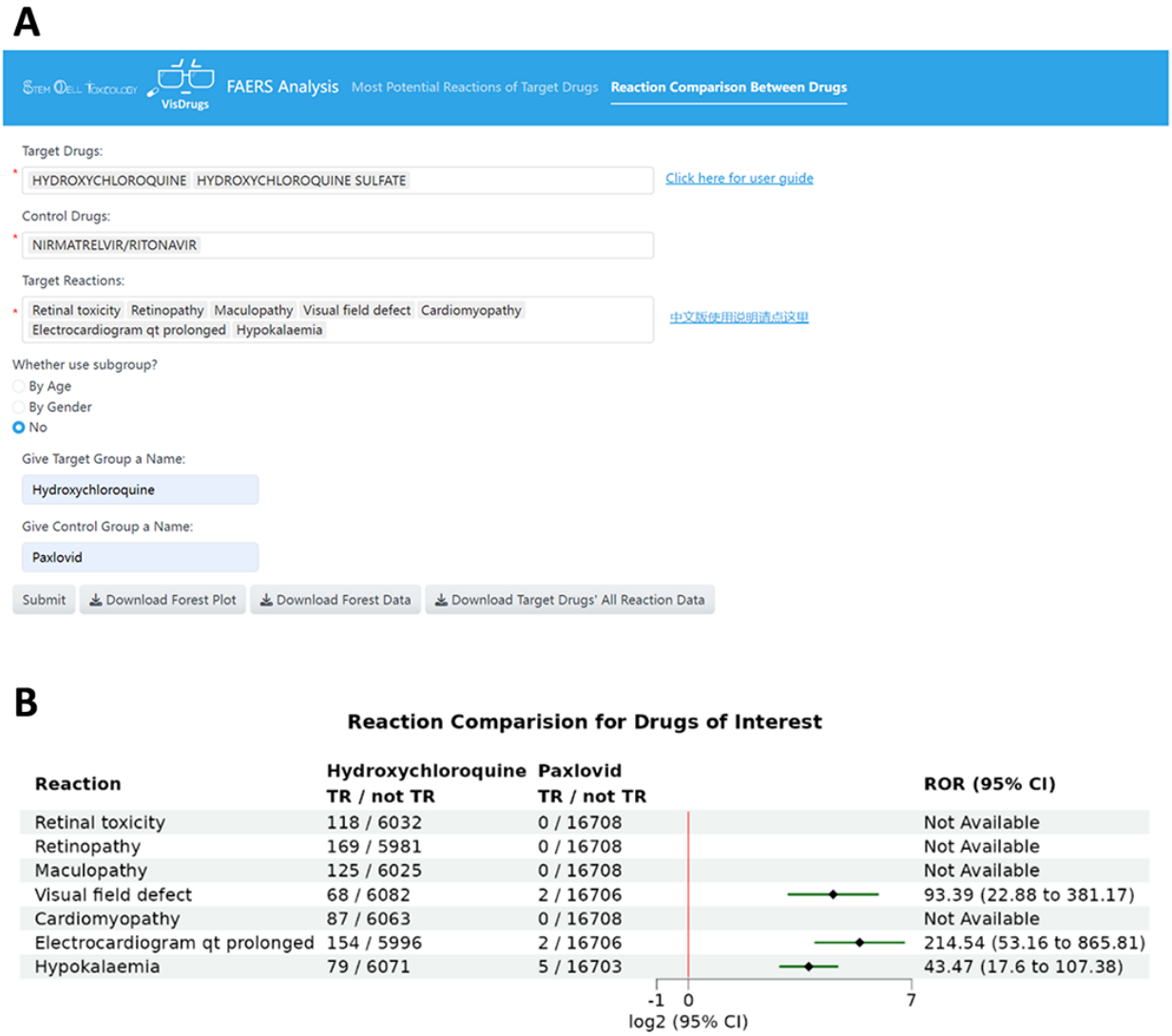

Application Example: Gender Difference Analysis of Hydroxychloroquine ADRs

In this example, the user inputs “hydroxychloroquine” in the target drug field, selecting both “hydroxychloroquine” and “hydroxychloroquine sulfate”. In the control drug field, “OTHER DRUGS” is retained for comparison with ADRs from all other drugs. The analysis focuses on retinal ADRs (such as retinal toxicity, retinopathy, maculopathy, and visual field defects), cardiac ADRs (including cardiomyopathy and electrocardiogram QT prolongation), and hypokalemia, which may lead to arrhythmias. The data is submitted with a gender-based subgroup analysis.

The results in Figure 4 indicate that “hydroxychloroquine” is more likely to cause retinal toxicity in female patients, including retinal toxicity (71.44-fold; *z* test *P* value <.001 than in male patients) and retinopathy (3.24-fold; *z* test *P* value =.001 than in male patients). This aligns with an epidemiological study by Jorge et al [35] in 2024, which found that female sex was associated with a higher risk of these adverse effects. Additionally, the results show that “hydroxychloroquine” is more likely to cause cardiomyopathy in female patients (13.82-fold; *z* test *P* value <.001 than male). This finding contrasts with a study by Goldman et al [36] in 2022, which analyzed the FAERS database data from 07/2014 to 09/2019 and did not observe a gender-based difference in cardiomyopathy risk. The discrepancy may be due to their inclusion of reports from non–health care professionals, which could be less accurate. Alternatively, the increase in cardiovascular disease in women during the COVID-19 pandemic, or the possibility that “hydroxychloroquine” may indeed have a greater propensity to cause heart disease in women, could explain this difference. Notably, in the ADRs associated with Paxlovid, we did not observe any cardiac-related ADRs. Additionally, evidence indicates that COVID-19 is more likely to cause viral myocarditis or pericarditis in men [37]. Therefore, we speculate that hydroxychloroquine may have a

XSL•FO
RenderX

higher risk of causing heart disease in women, which is a noteworthy finding from this study.

Figure 4. Evaluation of the gender-specific risk of hydroxychloroquine on retinopathy, cardiomyopathy, and hypokalemia. Shown is the forest plot comparing the adverse drug reactions (ADRs) of interest between hydroxychloroquine and other drug groups, subgrouped by gender. The black diamonds indicate the log2-transformed reporting odds ratios (RORs), and the green lines represent the corresponding log2-transformed 95% CIs. TR: target reaction.

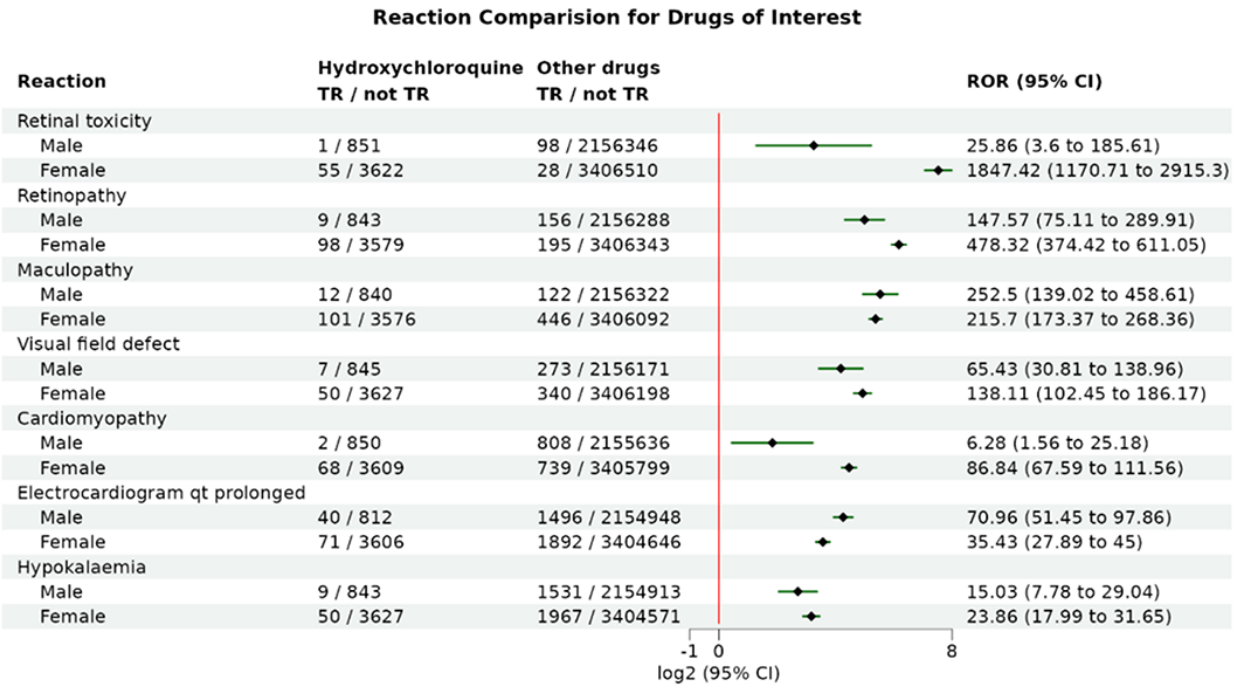

In addition, we conducted a gender-based subgroup analysis of the top 10 most frequently reported ADRs potentially representing ADRs for Paxlovid. The ROR differences between male and female subgroups for these ADRs were all less than 1.5 times (Figure S3 in Multimedia Appendix 1), suggesting that sex differences have minimal impact on the occurrence of these ADRs.

Application Example: Age Difference Analysis of Hydroxychloroquine ADRs

This example is similar to the previous one, with the main difference being the selection of age-based subgroup analysis. The age cutoff is set at 50 years. Interestingly, when the age is set to 50, the total number of case reports for patients using hydroxychloroquine is nearly equal for those older and younger than 50 years.

From the results (Figure 5), we observe that among the 7 ADRs of interest, only hypokalemia is more likely to occur in younger patients, while other cardiac and retinal ADRs tend to occur more frequently in older patients. Specifically, retinal-related ADRs associated with hydroxychloroquine are more prevalent in the older age group, for example, retinopathy (4.2-fold in older than younger people; *z* test *P* value<.001) and visual field defect (4.86-fold in older than younger people; *z* test *P* value<.001). This trend aligns with the findings in the 2024 study by Jorge et al [35], even though they divided the data at 45 years.

XSL•FO
RenderX

Figure 5. Assessment of the age-specific risk of hydroxychloroquine on retinopathy, cardiomyopathy, and hypokalemia. Shown is the forest plot comparing the adverse drug reactions (ADRs) of interest between hydroxychloroquine and other drug groups, subgrouped by age (age split set to 50 years). The black diamonds indicate the log2-transformed reporting odds ratios (RORs), and the green lines represent the corresponding log2-transformed 95% CIs. TR: target reaction.

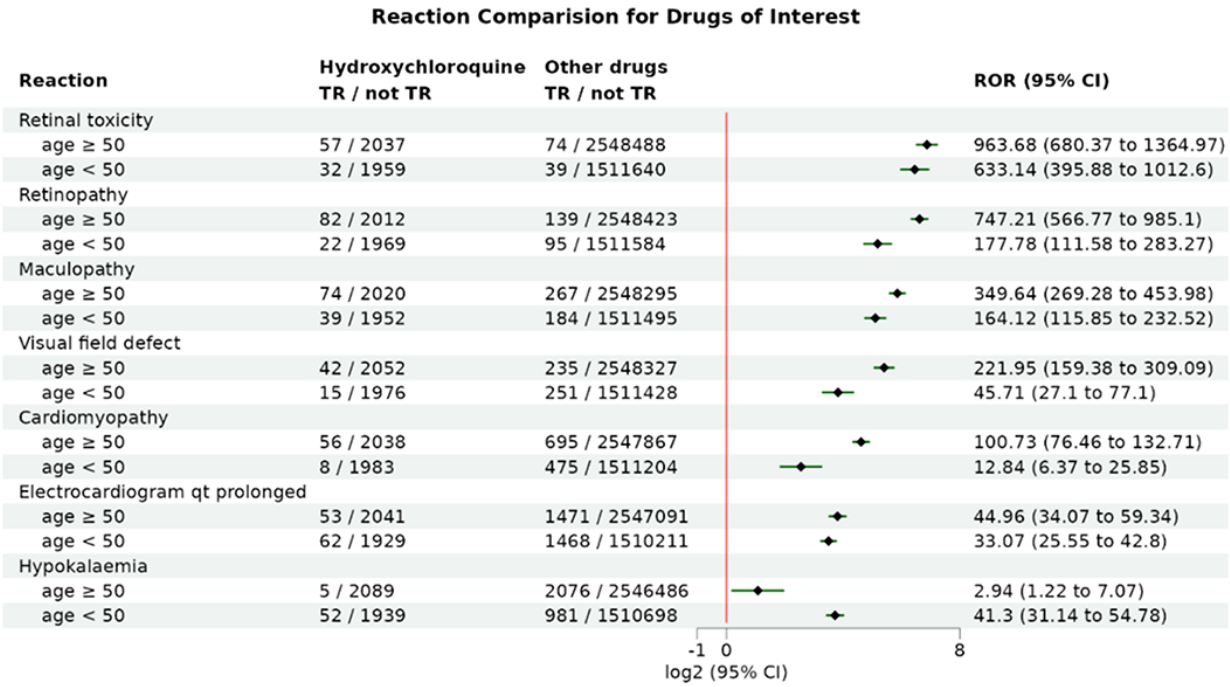

Regarding cardiac issues, while some studies suggest that age is not related to cardiac problems in hydroxychloroquine users [38], analysis of the latest FAERS database indicates that age does indeed influence the occurrence of cardiac-related ADRs, such as cardiomyopathy, with a 7.85-fold higher incidence in older patients compared to younger ones (*z* test *P* value <.001). This suggests that age, a well-established factor in drug-related adverse effects, also plays a significant role in the incidence of cardiac issues with hydroxychloroquine. At the same time, the fact that hydroxychloroquine significantly tends to cause hypokalemia in individuals younger than 50 years also warrants the attention of researchers.

In addition, we conducted an age-based subgroup analysis of the top 10 most frequently reported ADRs potentially representing ADRs for Paxlovid. The ROR differences between ≥50 and <50 year subgroups for these ADRs were all less than 1.5 times (Figure S4 in Multimedia Appendix 1), suggesting that age differences have minimal impact on the occurrence of these ADRs.

Website Evaluation

To further evaluate the functionality and effectiveness of our website, we used 2 main approaches during the assessment process. First, we conducted internal evaluations through regular group meetings, where students tested the website's drug ADR analysis functions. They also engaged in discussions to compare the analysis results with real-world reports, refining the analysis process to ensure that the results closely reflected actual ADRs. Second, 13 clinical staff members, including pharmacists, physicians, anesthesiologists, and technicians from institutions such as the Beijing Friendship Hospital, Beijing Tongren Hospital, Peking Union Medical College Hospital, Rizhao Central Hospital, and Rizhao Maternal and Child Health Hospital, participated as volunteers from November 2024 to January 2025 (Table 1) to evaluate the Visdrugs website. With their medical background and extensive clinical experience, they provided invaluable feedback after using the website, contributing to the development of the site and ensuring that it serves its purpose in ADR analysis effectively.

XSL•FO
RenderX

Table . A list of clinical staff members who participated in evaluating the website.

Office unit	Date of participation	Form of participation	Job title
Department of Pharmacy, Beijing Friendship Hospital, Capital Medical University	November 22, 2024	Website test and online discussion	Licensed pharmacist, senior pharmacist
Department of Otolaryngology, Beijing Friendship Hospital, Capital Medical University	November 22, 2024	Website test and online discussion	Associate chief physician
Department of Neurology, Beijing Tongren Hospital, Capital Medical University	December 19, 2024	Website test and online discussion	Associate chief physician
Department of Neurology, Peking Union Medical College Hospital	December 19, 2024	Website test and online discussion	Associate chief physician
Department of Social Health, Rizhao Maternal and Child Health Hospital, Shandong Province	January 03, 2024	Website test and online discussion	Director of the department, associate chief technician
Department of Blood Transfusion, Rizhao Central Hospital, Shandong Province	January 03, 2024	Website test and online discussion	Director of the department, associate chief technician
Department of Pharmacy, Rizhao Central Hospital, Shandong Province	January 15, 2024	Website test and online discussion	Director of the department, associate chief pharmacist
Department of Pharmacy, Rizhao Central Hospital, Shandong Province	January 15, 2024	Website test and online discussion	Attending pharmacist
Department of Pharmacy, Rizhao Central Hospital, Shandong Province	January 15, 2024	Website test and online discussion	Pharmacist
Department of Pharmacy, Rizhao Central Hospital, Shandong Province	January 15, 2024	Website test and online discussion	Pharmacist
Department of General Surgery, Rizhao Central Hospital, Shandong Province	January 15, 2024	Website test and online discussion	Director of the department, chief physician
Department of Oncology, Rizhao Central Hospital, Shandong Province	January 15, 2024	Website test and online discussion	Director of the department, associate chief physician
Department of Anesthesiology, Rizhao Central Hospital, Shandong Province	January 15, 2024	Website test and online discussion	Deputy director of the department, associate chief anesthesiologist
Department of Internal Medicine, Urumqi Traditional Chinese Medicine Hospital	January 16, 2025	Website test and online discussion	Attending doctor
Department of Reproduction, First Affiliated Hospital of Xinjiang Medical University	January 20, 2025	Website test and online discussion	Attending doctor
Department of Hepatology, Autonomous Region People's Hospital	20 January, 2025	Website test and online discussion	Attending doctor

Discussion

Overview

The FDA FAERS serves as a valuable resource for identifying potential safety concerns associated with drugs and has been widely used in various studies to monitor and analyze drug-related adverse effects [4,5,20,39]. Despite the wealth of data available in FAERS, its vast size makes data utilization heavily reliant on bioinformatics technologies. While platforms like VigiAccess and OpenVigil FDA have made significant strides in simplifying the data retrieval process [16,18], they still present notable limitations in analytical flexibility and usability for non-expert users. VigiAccess allows users to

XSL•FO
RenderX

explore ADR reports for individual drugs with summary statistics such as geographical distribution, gender, and age, but it lacks functionality for analyzing the influence of demographic factors on specific ADRs or for conducting comparative analyses across multiple drugs. OpenVigil FDA supports disproportionality analysis (eg, ROR, proportional reporting ratio) based on FAERS data and allows basic filtering, yet it does not provide flexible subgroup analyses or direct comparison of ADR profiles between drugs in an integrated, user-friendly environment.

To overcome these limitations, we developed VisDrugs, a web-based platform designed to facilitate comprehensive and intuitive ADR analysis for clinicians and researchers. VisDrugs uses an automated, standardized data processing workflow to clean and structure FAERS data using consistent nomenclatures (MedDRA PTs for ADRs and active substance names for drugs). Beyond basic retrieval and disproportionality analysis, VisDrugs offers unique capabilities, including flexible subgroup analysis by gender and age, as well as side-by-side comparison of ADR profiles across 2 drugs—features not available in existing public platforms. These advanced yet accessible functionalities aim to empower health care professionals and researchers to perform more nuanced and clinically relevant evaluations of drug safety without requiring extensive bioinformatics expertise.

From the examples provided in this study, we can see that the ADRs for COVID-19 drug Paxlovid and anti-malarial drug hydroxychloroquine obtained from the VisDrugs platform are highly consistent with those reported in previous studies. For Paxlovid, we primarily observed ADEs related to its use in treating COVID-19, which highlights the importance of distinguishing between ADRs related to the drug itself and those associated with the disease being treated. For hydroxychloroquine, the platform not only confirmed the commonly reported retinal and cardiac adverse effects, but through subgroup analysis, we also identified that retinotoxicity, retinopathy, and cardiomyopathy tend to occur more frequently in females and in individuals aged 50 years or older. These findings align with results from epidemiological studies and previous FAERS analyses [30-33,35,38], demonstrating the accuracy and practical utility of the platform's ADR analysis functionality.

Additionally, when analyzing the ADRs of an effective drug like Paxlovid [40], the top ADRs often include many reactions related to its intended indication, and the ROR confirms that these indication-related ADRs are indeed associated with Paxlovid usage. In contrast, when analyzing a drug with strong side effects, such as hydroxychloroquine [41], many of the top ADRs are non-indication related, and ROR shows that these non-indication ADRs are associated with hydroxychloroquine use. This suggests that the number of indication-related versus non-indication-related ADRs in the top drug usage-relevant ADRs, along with the severity comparison between them, may help evaluate the suitability of a drug. It can be inferred that when the severity of a drug's side effects does not exceed the severity of its indication, the indication-related ADRs are more likely to catch the attention of the reporters, and thus, most ADRs will be related to the drug's intended use. However, if a drug's side effects are so severe that they overshadow the indication, reporters are more likely to focus on reporting these non-indication-related ADRs. For example, in the case of hydroxychloroquine, where the ADR analysis reveals a lack of indication-related ADRs and the drug's adverse effects lead to severe outcomes, this may indicate that the drug is not suitable for treating that disease.

Limitations and Future Perspectives

While the development of the VisDrugs platform fills a gap in the existing FAERS analysis tools by offering 1-click visualization of ADR data, it does have some limitations. First, regarding the data source, VisDrugs relies primarily on FAERS data, which are collected through voluntary reports from health care professionals and others. Voluntary reporting bias is a well-known limitation of spontaneous reporting systems such as FAERS. Underreporting is common, particularly for known or less severe ADRs, whereas overreporting may occur following media attention or regulatory actions. Additionally, reporting patterns differ between health care professionals and patients. While health care professional reports are generally more clinically detailed and accurate, patient reports may be broader but less specific. Although VisDrugs enhances data reliability by focusing exclusively on reports submitted by health care professionals, addressing the impact of underreporting and overreporting remains an important challenge for future work. From an analytical perspective, the platform currently provides ADR frequency pie charts and ROR forest plots for analyzing 1 or 2 drugs. While these are useful, more advanced visualization functions are still in development. The platform continues to need optimization and upgrading based on user feedback. Moreover, the current version does not integrate in-depth biological information, such as molecular mechanisms or drug-target interactions, which would further enhance the platform's utility and depth of analysis.

Additionally, this study did not include a formal usability or performance evaluation of the VisDrugs platform. Structured evaluation studies (eg, user usability testing, performance benchmarking, and accuracy validation) are planned as part of future work to further validate and refine the system.

Looking ahead, we plan to expand the capabilities of the VisDrugs platform by adding more advanced visualization features, offering users the ability to specify the year for analysis, and enabling subgroup analysis based on factors like geographical location. We also aim to incorporate information on known drug indications, molecular targets, and links to ADR annotations to improve the platform's readability and usability. Furthermore, we envision integrating artificial intelligence technologies, such as machine learning models, to assist in ADR calculation or to use predictive models to interpret user-generated visualized results. These improvements will enhance the efficiency of drug safety evaluations.

Conclusions

Visdrugs offers a highly streamlined and user-friendly platform for analyzing ADRs, bridging the gap between raw FAERS data and actionable insights into drug safety. What this website focuses on is transforming raw data into accessible and systematic knowledge by establishing a comprehensive

XSL•FO
RenderX

analytical platform. This enables medical researchers to efficiently access post-marketing drug safety information without spending time on bioinformatics analyses or coding, allowing them to focus on evaluating the ADR profiles of different drugs. By doing so, VisDrugs contributes to advancing the assessment and monitoring of post-marketing drug safety.

Using the computational methods integrated into this platform, our previous study has already evaluated the pregnancy safety profiles of 2 hepatitis B antiviral drugs, adefovir and entecavir [39]. Since then, we have further enhanced the platform's capabilities. Now, it not only supports the analysis of the most common ADRs of individual drugs and the comparison of the toxicity profiles of 2 drug groups, but also enables subgroup analyses based on sex and age. This allows users to investigate the influence of demographic factors on drug safety, providing deeper insights into the differential effects of ADRs.

As demonstrated in our examples, VisDrugs significantly accelerates the process of comprehensive ADR evaluation. In just minutes, users can perform an in-depth assessment of a drug's safety profile, uncover patterns of ADRs across different demographic groups, and explore the impact of specific factors such as age and sex. This remarkable efficiency makes VisDrugs a powerful tool for advancing drug safety evaluation, improving our ability to monitor, understand, and mitigate ADRs in diverse populations.

Acknowledgments

This work was supported by the Ministry of Science and Technology of the People's Republic of China (grant 2020YFA0907500), the National Natural Science Foundation of China [grants 22021003, 22476209, and 22193052], the Science and Technology Program of Tongzhou District of Beijing (WS2025054), the Strategic Priority Research Program of the Chinese Academy of Sciences (grant XDB0750300), Beijing Municipal Natural Science Foundation (funding number IS23120), and Beijing Overseas Talents Center High-level Foreign Talent Project (funding number C2022001). Some figures were made using Figdraw.

Data Availability

The datasets generated or analyzed during this study are available in the Visdrugs GitHub repository [42].

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary figures (Figures S1-S4).
[DOCX File, 690 KB - formative_v9i1e71519_app1.docx]

Checklist 1
Software/Tools Papers checklist.
[DOCX File, 14 KB - formative_v9i1e71519_app2.docx]

References

1. Edwards IR, Aronson JK. Adverse drug reactions: definitions, diagnosis, and management. Lancet 2000 Oct 7;356(9237):1255-1259. [doi: 10.1016/S0140-6736(00)02799-9] [Medline: 11072960]
2. Khan LM. Comparative epidemiology of hospital-acquired adverse drug reactions in adults and children and their impact on cost and hospital stay--a systematic review. Eur J Clin Pharmacol 2013 Dec;69(12):1985-1996. [doi: 10.1007/s00228-013-1563-z] [Medline: 23955174]
3. Alfirevic A, Pirmohamed M. Genomics of adverse drug reactions. Trends Pharmacol Sci 2017 Jan;38(1):100-109. [doi: 10.1016/j.tips.2016.11.003]
4. Kass-Hout TA, Xu Z, Mohebbi M, et al. OpenFDA: an innovative platform providing access to a wealth of FDA's publicly available data. J Am Med Inform Assoc 2016 May;23(3):596-600. [doi: 10.1093/jamia/ocv153] [Medline: 26644398]
5. Liu H, Yang Q, Li Z, Yan S, Ming S. Systematic analysis of sugammadex-related adverse drug reaction signals using FAERS database. Int J Surg 2025 Feb 1;111(2):1988-1994. [doi: 10.1097/JS9.0000000000002194] [Medline: 39715164]
6. Küçükosmanoglu A, Scoarta S, Houweling M, et al. A real-world toxicity atlas shows that adverse events of combination therapies commonly result in additive interactions. Clin Cancer Res 2024 Apr 15;30(8):1685-1695. [doi: 10.1158/1078-0432.CCR-23-0914] [Medline: 38597991]
7. Yu RJ, Krantz MS, Phillips EJ, Stone CA Jr. Emerging causes of drug-induced anaphylaxis: a review of anaphylaxis-associated reports in the FDA Adverse Event Reporting System (FAERS). J Allergy Clin Immunol Pract 2021 Feb;9(2):819-829. [doi: 10.1016/j.jaip.2020.09.021] [Medline: 32992044]

XSL•FO
RenderX

8. Wichelmann TA, Abdulmujeeb S, Ehrenpreis ED. Bevacizumab and gastrointestinal perforations: a review from the FDA Adverse Event Reporting System (FAERS) database. Aliment Pharmacol Ther 2021 Nov;54(10):1290-1297. [doi: 10.1111/apt.16601] [Medline: 34499740]
9. Zhou X, Yang G, Zeng X, et al. Dupilumab and the potential risk of eosinophilic pneumonia: case report, literature review, and FAERS database analysis. Front Immunol 2023;14:1277734. [doi: 10.3389/fimmu.2023.1277734] [Medline: 38259470]
10. Ruiz M, Cocores A, Tosti A, Goadsby PJ, Monteith TS. Alopecia as an emerging adverse event to CGRP monoclonal antibodies: cases series, evaluation of FAERS, and literature review. Cephalalgia 2023 Feb;43(2):3331024221143538. [doi: 10.1177/03331024221143538] [Medline: 36739513]
11. Ge W, Chen W, Xie H. Real-world evidence supporting the consensus-driven DILI control compounds list: an analysis of FAERS data. J Hepatol 2025 Apr;82(4):e204-e205. [doi: 10.1016/j.jhep.2024.10.030] [Medline: 39461595]
12. Zhao YC, Li X, Wang CQ, et al. Unveiling the hidden risks: an update decade-long analysis of abraxane-related adverse events from the FAERS database. Int J Nanomedicine 2024;19:11847-11858. [doi: 10.2147/IJN.S490400] [Medline: 39558918]
13. Andreaggi CA, Novak EA, Mirabile ME, et al. Safety concerns reported by consumers, manufacturers and healthcare professionals: a detailed evaluation of opioid-related adverse drug reactions in the FDA database over 15 years. Pharmacoepidemiol Drug Saf 2020 Dec;29(12):1627-1635. [doi: 10.1002/pds.5105] [Medline: 32851782]
14. Maciejewski M, Lounkine E, Whitebread S, et al. Reverse translation of adverse event reports paves the way for de-risking preclinical off-targets. Elife 2017 Aug 8;6:e25818. [doi: 10.7554/eLife.25818] [Medline: 28786378]
15. Wang K, Wan M, Wang RS, Weng Z. Opportunities for web-based drug repositioning: searching for potential antihypertensive agents with hypotension adverse events. J Med Internet Res 2016 Apr 1;18(4):e76. [doi: 10.2196/jmir.4541] [Medline: 27036325]
16. Böhm R, von Hehn L, Herdegen T, et al. OpenVigil FDA – inspection of U.S. American Adverse Drug Events Pharmacovigilance Data and Novel Clinical Applications. PLoS ONE 2016;11(6):e0157753. [doi: 10.1371/journal.pone.0157753] [Medline: 27326858]
17. Almenoff JS, Pattishall EN, Gibbs TG, DuMouchel W, Evans SJW, Yuen N. Novel statistical tools for monitoring the safety of marketed drugs. Clin Pharmacol Ther 2007 Aug;82(2):157-166. [doi: 10.1038/sj.clpt.6100258] [Medline: 17538548]
18. Huang Y, Xu M, Ma X, et al. Characterizing ADRs of Enfortumab vedotin and Erdafitinib in bladder cancer treatment: a descriptive analysis from WHO-VigiAccess. Front Pharmacol 2024;15:1503154. [doi: 10.3389/fphar.2024.1503154]
19. Yang R, Yin N, Zhang Y, Faiola F. FAERS analysis. Stem Cell Toxicology. URL: http://sctdb.cn/shiny-server/Visdrugs_v.0.1.0/ [accessed 2025-07-22]
20. Caldito NG, Shirani A, Salter A, Stuve O. Adverse event profile differences between rituximab and ocrelizumab: findings from the FDA Adverse Event Reporting Database. Mult Scler 2021 Jun;27(7):1066-1076. [doi: 10.1177/1352458520949986] [Medline: 32820687]
21. Rothman KJ, Lanes S, Sacks ST. The reporting odds ratio and its advantages over the proportional reporting ratio. Pharmacoepidemiol Drug Saf 2004 Aug;13(8):519-523. [doi: 10.1002/pds.1001] [Medline: 15317031]
22. Liu J, Cao R, Xu M, et al. Hydroxychloroquine, a less toxic derivative of chloroquine, is effective in inhibiting SARS-CoV-2 infection in vitro. Cell Discov 2020;6:16. [doi: 10.1038/s41421-020-0156-0] [Medline: 32194981]
23. Manjaly Thomas ZR, Leuppi-Taegtmeyer A, Jamiolkowski D, et al. Emerging treatments in COVID-19: adverse drug reactions including drug hypersensitivities. J Allergy Clin Immunol 2020 Oct;146(4):786-789. [doi: 10.1016/j.jaci.2020.07.008] [Medline: 32710973]
24. Tan B, Zhang X, Ansari A, et al. Design of a SARS-CoV-2 papain-like protease inhibitor with antiviral efficacy in a mouse model. Science 2024 Mar 29;383(6690):1434-1440. [doi: 10.1126/science.adm9724] [Medline: 38547259]
25. Schillings J, Ramos-Guzmán CA, Ruiz-Pernía JJ, Tuñón I. Pomotrelvir and Nirmatrelvir binding and reactivity with SARS-CoV-2 main protease: implications for resistance mechanisms from computations. Angew Chem Int Ed Engl 2024 Oct 1;63(40):e202409527. [doi: 10.1002/anie.202409527] [Medline: 38959351]
26. Pavan M, Bolcato G, Bassani D, Sturlese M, Moro S. Supervised Molecular Dynamics (SuMD) insights into the mechanism of action of SARS-CoV-2 main protease inhibitor PF-07321332. J Enzyme Inhib Med Chem 2021 Dec;36(1):1646-1650. [doi: 10.1080/14756366.2021.1954919] [Medline: 34289752]
27. Hoang T. Systematic review and meta-analysis of factors associated with re-positive viral RNA after recovery from COVID-19. J Med Virol 2021 Apr;93(4):2234-2242. [doi: 10.1002/jmv.26648] [Medline: 33135788]
28. Elzein F, Ibrahim A, Alshahrani F, et al. Reinfection, recurrence, or delayed presentation of COVID-19? Case series and review of the literature. J Infect Public Health 2021 Apr;14(4):474-477. [doi: 10.1016/j.jiph.2021.01.002] [Medline: 33743368]
29. Li M, Zhang QS, Liu XL, Wang HL, Liu W. Adverse events associated with Nirmatrelvir/Ritonavir: a pharmacovigilance analysis based on FAERS. Pharmaceuticals (Basel) 2022 Nov 24;15(12):1455. [doi: 10.3390/ph15121455] [Medline: 36558906]
30. Melles RB, Jorge AM, Marmor MF, et al. Hydroxychloroquine dose and risk for incident retinopathy: a cohort study. Ann Intern Med 2023 Feb;176(2):166-173. [doi: 10.7326/M22-2453] [Medline: 36645889]

XSL•FO
RenderX

31. Kim J, Kim KE, Kim JH, Ahn SJ. Practice patterns of screening for hydroxychloroquine retinopathy in South Korea. JAMA Netw Open 2023 May 1;6(5):e2314816. [doi: 10.1001/jamanetworkopen.2023.14816]
32. Chatre C, Roubille F, Vernhet H, Jorgensen C, Pers YM. Cardiac complications attributed to chloroquine and hydroxychloroquine: a systematic review of the literature. Drug Saf 2018 Oct;41(10):919-931. [doi: 10.1007/s40264-018-0689-4] [Medline: 29858838]
33. Voisin O, Lorc'h EL, Mahé A, et al. Acute QT interval modifications during hydroxychloroquine-azithromycin treatment in the context of COVID-19 infection. Mayo Clin Proc 2020 Aug;95(8):1696-1700. [doi: 10.1016/j.mayocp.2020.05.005] [Medline: 32753141]
34. Marquardt K, Albertson TE. Treatment of hydroxychloroquine overdose. Am J Emerg Med 2001 Sep;19(5):420-424. [doi: 10.1053/ajem.2001.25774]
35. Jorge AM, Melles RB, Marmor MF, Zhou B, Zhang Y, Choi HK. Risk factors for hydroxychloroquine retinopathy and its subtypes. JAMA Netw Open 2024 May 1;7(5):e2410677. [doi: 10.1001/jamanetworkopen.2024.10677] [Medline: 38722628]
36. Goldman A, Bomze D, Dankner R, et al. Cardiovascular adverse events associated with hydroxychloroquine and chloroquine: a comprehensive pharmacovigilance analysis of pre-COVID-19 reports. Br J Clin Pharmacol 2021 Mar;87(3):1432-1442. [doi: 10.1111/bcp.14546] [Medline: 32964535]
37. Rafaniello C, Gaio M, Zinzi A, et al. Disentangling a thorny issue: myocarditis and pericarditis post COVID-19 and following mRNA COVID-19 vaccines. Pharmaceuticals (Basel) 2022 Apr 25;15(5):525. [doi: 10.3390/ph15050525] [Medline: 35631352]
38. Pinheiro MM, Pileggi GS, Kakehasi AM, et al. Incidence and risk factors for moderate/severe COVID-19 in rheumatic diseases patients on hydroxychloroquine: a 24-week prospective cohort. Clin Exp Rheumatol 2022 Jul;40(7):1258-1266. [doi: 10.55563/clinexprheumatol/67oyux] [Medline: 34251312]
39. Yang R, Yin N, Zhao Y, et al. Adverse events during pregnancy associated with entecavir and adefovir: new insights from a real-world analysis of cases reported to FDA adverse event reporting system. Front Pharmacol 2021;12:772768. [doi: 10.3389/fphar.2021.772768] [Medline: 35046808]
40. Harris E. FDA grants full approval to paxlovid, COVID-19 antiviral treatment. JAMA 2023 Jun 27;329(24):2118. [doi: 10.1001/jama.2023.9925]
41. Juurlink DN. Safety considerations with chloroquine, hydroxychloroquine and azithromycin in the management of SARS-CoV-2 infection. Can Med Assoc J 2020 Apr 27;192(17):E450-E453. [doi: 10.1503/cmaj.200528] [Medline: 32269021]
42. Mrpotatod/visdrugs_v.0.1.0. GitHub. URL: https://github.com/mrpotatod/Visdrugs_v.0.1.0 [accessed 2025-07-22]

Abbreviations

ADR: adverse drug reaction
ASCII: American Standard Code for Information Interchange
FAERS: US Food and Drug Administration Adverse Event Reporting System
FDA: US Food and Drug Administration
MedDRA: Medical Dictionary for Regulatory Activities
PT: preferred term
ROR: reporting odds ratio

Edited by A Mavragani; submitted 20.01.25; peer-reviewed by O Obadimu, U Modebelu; revised version received 10.06.25; accepted 12.06.25; published 31.07.25.

Please cite as:
Yang R, Yin N, Zhang Y, Faiola F
Visualization of FDA Adverse Drug Reaction Reports: Development and Usability Study of the VisDrugs Web Server
JMIR Form Res 2025;9:e71519
URL: https://formative.jmir.org/2025/1/e71519
doi:10.2196/71519

© Renjun Yang, Nuoya Yin, Yang Zhang, Francesco Faiola. Originally published in JMIR Formative Research (https://formative.jmir.org), 31.7.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Use of a Medical Communication Framework to Assess the Quality of Generative Artificial Intelligence Replies to Primary Care Patient Portal Messages: Content Analysis

Natalie S Lee[1], MPH, MSHP, MD; Nathan Richards[1], MD; Jodi Grandominico[1], MD; Robert M Cronin[1], MS, MD; Amanda K Hendricks[2], MSN, RN, CNL; Ravi S Tripathi[2], MBA, MD; Daniel E Jonas[1], MPH, MD

[1]Department of Internal Medicine, College of Medicine, The Ohio State University, 2050 Kenny Rd, suite 2400, Columbus, OH, United States
[2]Health System Informatics, The Ohio State University Wexner Medical Center, Columbus, OH, United States

Corresponding Author:
Natalie S Lee, MPH, MSHP, MD
Department of Internal Medicine, College of Medicine, The Ohio State University, 2050 Kenny Rd, suite 2400, Columbus, OH, United States

Abstract

Background: There is growing interest in applying generative artificial intelligence (GenAI) to respond to electronic patient portal messages, particularly in primary care where message volumes are highest. However, evaluations of GenAI as an inbox communication tool are limited. Qualitative analysis of when and how often GenAI responses achieve communication goals can inform estimates of impact and guide continuous improvement.

Objective: This study aims to evaluate GenAI responses to primary care messages using a medical communication framework.

Methods: This was a descriptive quality improvement study of 201 GenAI replies to a purposively sampled, diverse pool of real primary care patient messages in a large midwestern academic medical center. Two physician reviewers (NSL and NR) used a hybrid deductive-inductive approach to qualitatively identify and define themes, guided by constructs from the “best practice” medical communication framework. After achieving thematic saturation, the reviewers assessed the presence or absence of identified communication themes, both independently and collaboratively. Discrepant observations were reconciled via discussion. Frequencies of identified themes were tallied.

Results: Themes in strengths and limitations emerged across 5 communication domains. In the domain of *rapport building*, expressing respect and restating key phrases were strengths, while inappropriate or inadequate rapport building statements were limitations. For *information gathering*, questions that built toward a plan or elicited patient needs were strengths, while questions that were out of place or redundant were limitations. For *information delivery*, accurate content delivered clearly and professionally was a strength, but delivery of inaccurate content was an observed limitation. GenAI responses could *facilitate next steps* by outlining choices or providing instruction, but sometimes those next steps were inappropriate or premature. Finally, in *responding to emotion*, strengths were that emotions were named and validated, while inadequate or absent acknowledgment of emotion was a limitation. Overall, 26.4% (53/201) of all messages displayed communication strengths without limitations, 27.4% (55/201) had limitations without strengths, and the remaining 46.3% (93/201) had both. Strengths outnumbered limitations in *rapport building* (87/201, 43.3% vs 35/201, 17.4%) and *facilitating next steps* (73/201, 36.3% vs 39/201, 19.4%). Limitations outnumbered strengths in the remaining domains of *information delivery* (89/201, 44.3% vs 43/201, 21.4%), *information gathering* (60/201, 29.9% vs 43/201, 21.4%), and *responding to emotion* (7/201, 8.5% vs 9/201, 4.5%).

Conclusions: GenAI response quality on behalf of primary care physicians and advanced practice providers may vary by communication function. Expressions of respect or descriptions of common next steps may be appropriate, but gathering and delivering appropriate information, or responding to emotion, may be limited. While communication standards were often met, they were also often compromised. Understanding these strengths and limitations can inform decisions about whether, when, and how to apply GenAI as a tool for primary care inbox communication.

(JMIR Form Res 2025;9:e71966) doi:10.2196/71966

KEYWORDS

communication; artificial intelligence; primary care; electronic health record; patient portal; health communication

XSL•FO
RenderX

Introduction

Primary care is in crisis. The demand for services far exceeds supply, while the daily burden of responsibilities is increasingly untenable for primary care physicians and advanced practice providers (PCPs) [1]. Patient messages submitted to their providers through electronic health record (EHR) portals are a well-recognized burden for clinicians [2-4], especially in primary care [5,6]. Nationally, initiatives to address and mitigate these sources of PCP burden and burnout have become a priority, including for the National Academy of Medicine [7], the American Medical Association [8], the US Surgeon General [9], and the Agency for Healthcare Research and Quality [10].

A novel potential strategy is the adoption of generative artificial intelligence (GenAI) tools to draft replies to patient messages on behalf of PCPs [11]. GenAI technologies have demonstrated capacity to pass medical board examinations [12] and make accurate diagnoses [13], and they are already transforming day-to-day clinical practice. Thoughtful application of GenAI tools has the potential to improve many deficiencies of health care delivery for both patients and PCPs. However, the "when" and "how" of applying GenAI to respond to primary care messages is still in its formative period, with a high price of investment and uncertain impact. Viable solutions must reduce PCP burden while maintaining or enhancing patient care quality. Thoughtful evaluations in this early period are critical, both to inform decisions about investing in these tools and to direct their development and application in ways that contribute genuine value to primary care.

A small but rapidly growing body of reports from early adopters examines GenAI responses to real patient messages submitted to their care teams. Reported evaluation metrics include draft "use," defined as the human author's decision to start the response with at least a portion of the GenAI draft, and physician work metrics, such as time required to read and write messages [14-16]. However, draft use is a poor indicator of meaningful impact on PCP burden or patient care, while physician work metrics have been largely unaffected [14,15]. Surveys and brief qualitative feedback have provided further insights, with data suggesting variable perceptions of GenAI draft quality and mixed impact on workload, ranging from helpful to unhelpful or even increasing work due to editing burden [15-17]. More data are needed to inform our understanding of the potential role for GenAI drafts in primary care, and in what ways the technology could be better leveraged to optimize its use.

Frameworks for medical communication are widely accepted and taught, with a consensus on "best practice" for high-quality communication [18]. Ideal automation would ensure that standards for communication were consistently met, and such a tool would invariably be helpful to PCPs while maintaining or enhancing patient care. However, no prior studies have applied a medical communication framework to systematically evaluate GenAI draft quality or the consistency of its quality. Another limitation of prior studies is that GenAI messages on behalf of multiple human author types, such as medical assistants, nurses, and physicians and advanced practice providers, were mixed, as were specialty contexts such as primary care versus other specialties. Greater specificity in evaluation is important because the nature of messages sent by patients and the expected information from the responses differs across these diverse author types and contexts.

This quality improvement study addresses these gaps by applying the "best practice" communication framework to systematically identify and quantify themes in communication quality among GenAI responses to diverse primary care patient messages. We focus on messages to, and drafts on behalf of, PCPs, allowing us to apply a uniform standard of evaluation and collect more precise insights for the *where* and *how* of GenAI response tools in primary care.

Methods

Study Design

We conducted a thematic analysis of 201 unique GenAI messages using a hybrid deductive-inductive approach, guided by the "best practice" communication framework. The 201 GenAI messages were generated in response to a diverse pool of primary care patient messages, using prompts that had undergone multiple prior rounds of optimization.

We first describe the prestudy context, including how the study sample of GenAI responses to patient messages was derived and how GenAI prompts were optimized prior to thematic analysis of its output. We then describe the methods used in the qualitative analysis of the 201 unique GenAI messages.

Prestudy Context: GenAI Response Pilot in Primary Care

At our institution, a large midwestern academic medical center, we piloted the GenAI draft tool offered by our EHR vendor, Epic, from September 14, 2023, through April 30, 2024, which at that time was based on GPT 3.5. The tool is used to generate a response to a patient-submitted message; the message recipient can choose to use, edit, or ignore the draft reply in responding to the patient. The purpose of our pilot was to gain sufficient experience with GenAI responses to make an informed decision about institutional investment in the tool for use by PCPs. The pilot team included 6 PCP volunteers and a cohort of analysts from the medical center's health system informatics team.

Aiming to uncover both the strengths and limits of the tool's capabilities, the pilot team took several steps to maximize variation in the pool of patient messages used to test the model. The PCP volunteers were diverse across a range of factors that could influence patient message variation: 3 from family medicine and 3 from general internal medicine, 4 women and 2 men, an equal number of White and non-White participants, and a broad span of clinical experience, from 5 years to several decades in independent practice (median 18 years). Analysts pulled the most recent 400 patient messages submitted to the participating PCPs, distributed evenly across them. It was possible for unique patients to contribute multiple messages to the pool. From this pool, they manually curated a final sample of 200 messages selected for diversity in content, language, and complexity.

XSL•FO
RenderX

GenAI output is sensitive to the wording of prompts, or instructions, provided, and the creation of appropriate prompts for the task at hand is critical for optimizing the quality of the GenAI response [19,20]. Thus, the pilot team sought to optimize GenAI prompts through multiple rounds of prompt engineering, which is the term for optimizing instructions for the GenAI model. Our informatics analysts served as the primary prompt engineers. Standards for prompt engineering have not yet been developed in health care, but our prompt engineers used all resources available through Epic and publicly available external resources. Starting with the Epic-suggested prompt, prompt engineering was further guided by examples of prompts used at other institutions and attendance at multiple prompt engineering "office hours," during which prompt engineers from health care organizations across the country shared experiences, ideas, and insights.

Prompt engineering rounds were iterative. For each new or revised prompt, the participating PCPs reviewed the GenAI draft responses, paired with the corresponding patient messages in adjacent columns on a spreadsheet and no other chart information. Anywhere from 30 to the full set (n=200) of drafted responses was reviewed for each round, with larger quantities in earlier rounds and smaller quantities in later rounds to improve review efficiency. Analysts used subjective assessment to ensure that each round was based on a diverse subset of patient messages. Each round concluded with a virtual meeting among participating PCPs and health system analysts to discuss their perceptions of GenAI draft messages, considering tone, information, and helpfulness of the reply. Combining this feedback with insights from vendor resources (office hour notes and prompt library), key areas for changes to the prompt were identified for the next round of prompt engineering.

For further optimization, prompt engineering rounds were conducted separately for different categories of messages, as identified by Epic's proprietary message filter (Figure 1). *Clerical messages* pertained to requests for paperwork, letters, medications, and test results; *general messages* were all other message types. Through this process of interactive and iterative discussion, 5 rounds of prompt engineering for general patient messages and 4 rounds for clerical messages were completed between November 1, 2023, and May 1, 2024. Decisions not to pursue additional rounds of prompt engineering were based on multiple factors, including minimal perceived difference in AI responses between later rounds and time and resource restraints.

Across prompt engineering rounds, the GenAI model had limited information from patients' medical records. Patients' preferred names and pharmacies were always accessible. For general messages, patients' age, allergies, and all future appointments were also accessible. For clerical messages, outpatient medication lists, employer names, and test results for the last 2922 hours (approximately 3 months) were also accessible. These were decisions based upon institutional patient privacy concerns, costs per token sent to the model, and model time interval limits.

During the pilot, 2 physicians withdrew participation early, citing a lack of enthusiasm after the first few rounds of prompt engineering. At the end of the pilot, the remaining physicians unanimously voted against adoption. Their feedback was that similar functionality could be achieved through existing EHR tools, while reviewing and editing GenAI responses increased cognitive burden.

XSL•FO
RenderX

Figure 1. Flowchart depicting pilot activities and the subsequent thematic analysis of artificial intelligence–generated responses. GenAI: generative artificial intelligence; PCPs: primary care physicians and advanced practice providers.

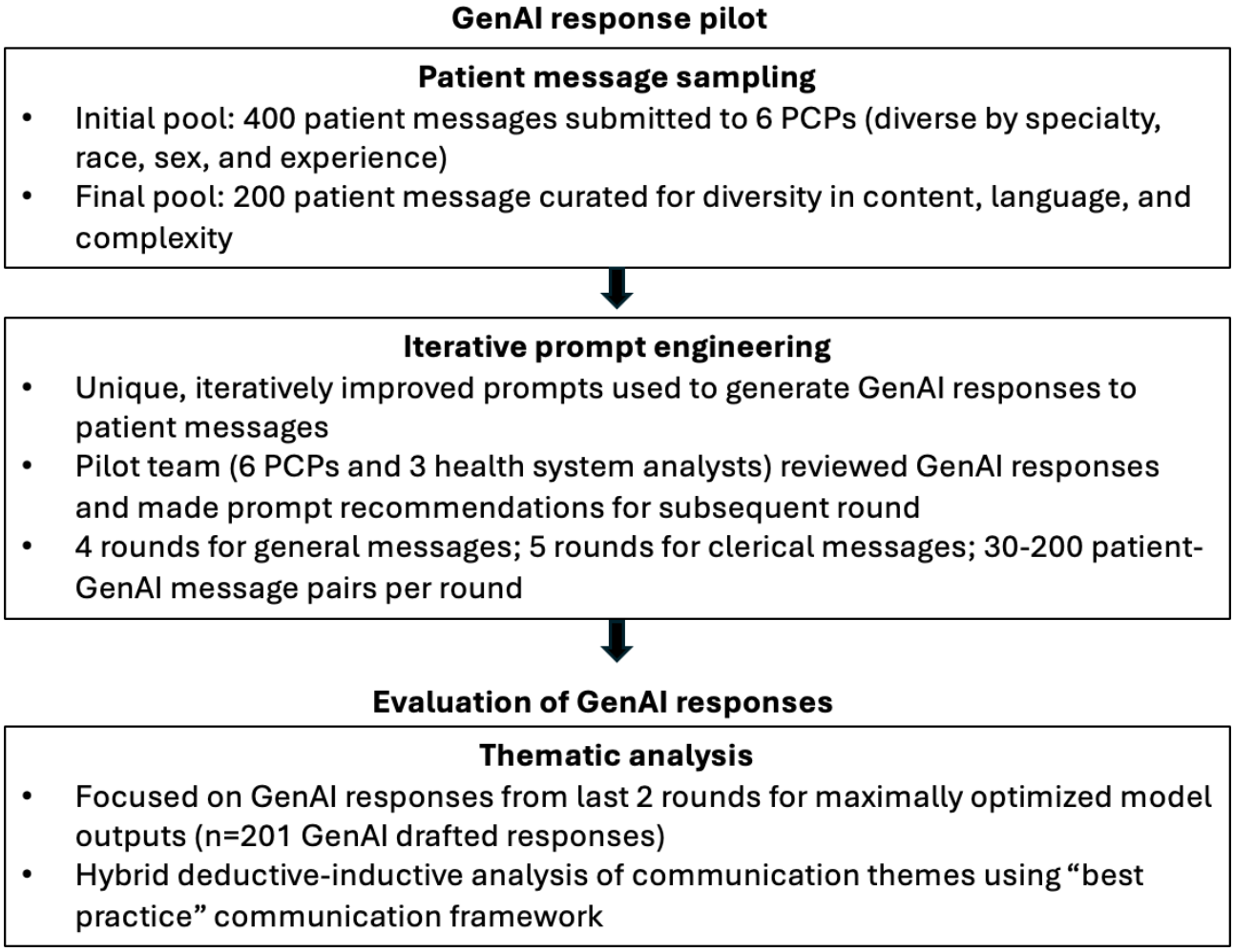

Evaluation of GenAI Responses

Upon completion of the pilot, we sought to more systematically assess and convey the quality of GenAI draft responses through the application of the "best practice" communication framework [18]. We followed reporting guidelines from the Standards for Quality Improvement Reporting Excellence [21], recommended by the Enhancing the Quality and Transparency Of health Research network [22].

GenAI draft messages from the last 2 rounds of prompt engineering in each message category (n=201 total) (n=139 for general, n=62 for clerical) were included in the analysis, as these represented messages that had been generated using maximally optimized prompts, through several prior rounds of prompt engineering. The number of GenAI drafts reviewed exceeded the number of unique patient messages in the pool because, as described in the "Prestudy Context" section, patient messages could be randomly reused for subsequent prompt engineering rounds. Two primary care physician reviewers (1 who had participated in the prompt engineering pilot and 1 who had not) reviewed all 201 messages. As in the initial pilot, GenAI drafts were presented to reviewers in a column adjacent to corresponding patient messages and other clinical information (eg, notes, laboratory data, medications, etc) was not included. Thus, reviewers did not have more contextual information than the GenAI model and assumed the information it was accessing (medication lists, future appointments, etc) was accurate.

Thematic Analysis

A hybrid deductive-inductive analytic approach was used to identify themes and create a comprehensive codebook. We established five overarching domains of communication aligned with the domains of the "best practice" communication framework: (1) *rapport building*, (2) *information gathering*, (3) *information delivery*, (4) *facilitating next steps*, and (5) *responding to emotions*. A final sixth domain of the "best practice" framework, "enabling patient behavior," was folded into the "facilitating next steps" domain. These 5 domains served as the foundational categories for organizing the data.

Incorporating an inductive approach, we then used observations from the GenAI response to identify and define specific themes within each domain. Through iterative coding and discussion between reviewers, we examined how emergent themes related to original constructs within the "best practice" communication framework. We thus refined and contextualized each domain with themes grounded in our observations of the GenAI responses, while remaining theoretically informed.

We identified and refined themes until at least half of the dataset had been reviewed and there was mutual agreement that thematic saturation had been achieved. We stratified this process by

XSL•FO
RenderX

message type (clerical vs general) in case meaningful differences in themes emerged between them; thus, we reviewed and analyzed 70 GenAI drafts for general messages and 32 for clerical messages (103 messages combined) collaboratively to develop and refine the codebook [18]. Ultimately, because themes were similar between message types, a uniform codebook was applied to all messages. The final codebook identified and defined binary categories—strengths versus limitations—of themes within each domain, resulting in 10 unique message categories.

Tabulation of Themes and Statistical Analyses

Next, we applied the codebook to assess how often the identified communication themes emerged across the entire sample of GenAI draft responses (n=201). All code categories were independent and not mutually exclusive, so that a message could be coded as demonstrating both strength and shortcoming in a particular communication domain (eg, a message could have both strengths and limitations in the "rapport building" domain). However, per message, each category (strength or limitation within the 5 domains) was counted only once per message (eg, a strength in "fostering relationship" could be marked only once for a given message, even if additional examples of strengths in this domain existed elsewhere in the GenAI draft). We did not require all communication domains to be present in each draft; rather, we assessed strengths and limitations only when the message touched on a relevant domain. After assessing the first half (n=70 for general, n=32 for clerical, and n=103 combined) collaboratively to confirm interrater coding alignment, the remaining messages (n=69 general and n=30 clerical) were reviewed and coded by each reviewer independently. Agreement between reviewers for the latter batch was assessed using Cohen κ [23,24], and all discrepancies were reconciled via discussion, guided by the tenets of the "best practice" communication framework. The final tally of communication themes was based solely on reconciled observations.

Within the domain of "information delivery," under the category defined as "limitations," we identified multiple granular subcategories, representing distinct types of limitations specific to this domain (4 subcategories for general messages and 5 subcategories for clerical messages). While these subcategories were ultimately collapsed into a single coherent definition of "limitations" in the final codebook, we retained them during analysis due to their potential to yield valuable and actionable insights. Both reviewers independently assessed these subcategories whenever a limitation in this domain was coded. Given the exploratory nature of these subcategories, we did not formally assess interrater agreement or reconcile discrepancies at the subcategory level. We tallied the frequency of each limitation subtype, counting observations if marked by at least 1 reviewer.

Finally, each reviewer rated whether they were likely to use at least a portion of the GenAI draft in their own response to the patient message. For drafts that were assessed to be usable, reviewers assessed whether major edits versus minor to no edits would be needed. If the reviewer assessed that no portion of the GenAI draft message would be retained, the GenAI draft was assessed as "would not use." We did not reconcile these assessments, acknowledging the role of individual physician preference.

Ethical Considerations

The Ohio State University institutional review board determined that this activity met criteria for quality improvement and was thus exempt from human subject ethics review. Informed consent was also waived. Study data were not deidentified, and protective measures included keeping all data on secure servers that met the highest standards for restricted institutional data, as well as limiting access to those directly involved with the study. No compensation was provided to individuals whose data were included.

Results

We identified strengths and limitations of GenAI responses across 5 communication domains (Table 1). In the domain of *rapport building*, observed strengths were that responses expressed respect and restated key phrases, while observed limitations were that rapport building phrases seemed inappropriate, inadequate, or were absent. For *information gathering*, observed strengths were that GenAI-posed questions elicited patient needs and contributed to clinical decision-making, while limitations were observed when the questions posed were inappropriate or redundant. For *information delivery*, observed strengths were that responses provided accurate information professionally, while limitations were observed when the information was inaccurate or unsuited for the context. In *facilitating next steps*, GenAI responses not only successfully outlined choices or next steps but also demonstrated limitations by moving to next steps prematurely or suggesting inappropriate next steps. For *responding to emotion*, successful responses named and validated emotions conveyed in patient messages, while limitations in this domain were observed when there was inadequate or absent acknowledgment of emotion. These definitions were consistent for general and clerical messages.

The Cohen κ measure of interrater reliability of assessments was 0.55 for general messages and 0.59 for clerical, both suggesting moderate agreement [23,24]. Overall, 26.4% (53/201) of all messages exhibited only strengths in communication, while 27.4% (55/201) of messages exhibited only limitations in communication with no identifiable strengths (Table 2). The remaining 46.3% (92/201) of messages had a mix of both strengths and limitations in communication domains.

Strengths were more frequently observed than limitations in the domains of rapport building and facilitating next steps (Multimedia Appendix 1). Half of all replies to general messages and 25% of replies to clerical messages contained appropriate expressions of respect. Meanwhile, nearly 40% of replies to clerical messages and a third of replies to general messages outlined reasonable next steps. For the remaining domains of information gathering and delivery, as well as responding appropriately to patient emotion, limitations were more frequently observed than strengths.

XSL•FO
RenderX

In exploratory analysis, there were multiple identifiable subcategories of limitations in information delivery. The reasons were that the patient's informational need was not addressed (34/56, 60.7%), the response contained inaccurate information about clinical workflow (21/56, 37.5%), the response was simply wrong (11/56, 19.6%), and, rarely, because the delivery was unprofessional (1 message). Similar subcategories were observed for clerical messages with inadequate information delivery: informational need not addressed (24/33, 72.7% of messages), inaccurate assumptions about workflow (11/33, 33.3%), and information was wrong (10/33, 30.3%). Two additional subcategories for clerical messages were that the answer made inappropriate assumptions about the clinical context (8/33, 24.2%), and that the information provided did not appropriately account for shared understanding as implied in the patient's message (5/33, 15%).

Finally, both physicians also determined whether each would use the GenAI draft. On average, the physicians assessed that 56.5% of messages of general messages could be used (53.2% for reviewer 1 and 59.7% for reviewer 2), with "fair" interrater agreement about which messages were usable (Cohen κ=0.257). Of those messages being used, 25.7%-65.1% (mean 45%, SD 27.9%) would require major edits. On average, for clerical messages, physicians assessed that they would use 45.2% of clerical message draft replies (46.8% for reviewer 1 and 43.5% for reviewer 2). Interrater agreement on message usability was almost perfect (0.81). Of messages being used, 27.6%-37.0% (mean 32.3%, SD 6.6%) would require major editing.

Table . Results of thematic analysis: medical communication strengths and limitations observed in generative artificial intelligence drafts.

Domain	Strength	Examples	Limitation	Examples
Rapport building	Expresses respect; repeats key phrases back.	"I appreciate your update." "It's great to hear that…"	Inappropriate or inadequate rapport building.	Repeats phrases back to the patient that miss the essence of the question; generic "I understand you are seeking X" does not suit the context.
Information gathering	Questions build toward plan, elicit patient needs.	Asks clinically appropriate follow-up questions about new symptoms.	Questions are out of place or redundant.	Questions focus narrowly on a single problem that is not the main clinical issue; the patient's message clearly suggests that further questioning is unnecessary.
Information delivery	Appropriate content delivered professionally, avoiding jargon.	Provides accurate information about what to expect with a medical test.	Inappropriate or unsuitable content.	Patient asks about kidney donation, but the response provides information relevant to blood donation.
Facilitating next steps	Outlines choices, leads to appropriate next steps.	Responds affirmatively to patient request for blood test and provides appropriate logistical instructions.	Inappropriate or premature next steps.	Reply inappropriately complies with patient request for a specific blood test when more history and evaluation are indicated.
Responding to emotion	Names and validates emotions.	Patient expresses anxiety, and reply acknowledges, "It's natural to feel anxious about this."	Acknowledgment of emotion is inadequate or absent.	The patient expresses a long-standing emotional experience related to a condition; the reply does not acknowledge the patient's emotion.

Table . Frequency of domain strength and limitations by message type.

Domain	General (n=139), n (%)		Clerical (n=62), n (%)	
	Strength	Limitation	Strength	Limitation
Rapport building	71 (51.1)	27 (19.4)	16 (25.8)	8 (12.9)
Information gathering	42 (30.2)	52 (37.4)	1 (1.6)	8 (12.9)
Information delivery	38 (27.3)	56 (40.3)	5 (8.1)	33 (53.2)
Facilitating next steps	47 (33.8)	26 (18.7)	26 (41.9)	13 (21.0)
Responding to emotions	8 (5.7)	10 (7.2)	1 (1.6)	7 (11.3)

XSL•FO
RenderX

Discussion

Principal Findings

Use of GenAI to respond to patient messages on behalf of PCPs is promoted as a promising tool for PCPs, but much remains unknown about the quality of the communication contained in the drafts and where the opportunities are for improvement. Applying a medical communication framework to assess GenAI responses, we observed that while communication standards were often met successfully, they were also often compromised. As a strength, GenAI responses contained many expressions of respect and facilitated appropriate next steps. However, there were also notable limitations in communication standards, including asking appropriate follow-up questions, matching information delivery to the patient's informational need, and recognizing and responding to emotion. We draw several insights from these observations.

In our study setting, the GenAI model had limited access to chart information (eg, medications but not clinic notes) and responded only in the context of patient-PCP communication, and prompts were optimized through multiple rounds of robust prompt engineering. Our results suggest that, in that setting, GenAI responses may not only have many strengths but also many limitations when responding to real primary care patient messages. In many prior tests of GenAI responses, often with favorable assessments of GenAI performance, patient-generated content was processed and formatted for input into the model [25-27]. In primary care settings, patient messages to their providers are often more complex, containing multiple questions and nuances, and picking up from prior conversations that may or may not be documented. Clinician judgment is important for discerning and navigating the context, patients' emotions, their informational needs, and appropriate follow-up questions. GenAI drafts in their current form may be inadequate for such tasks, in part, because of incomplete information and also because the output is a product of language prediction rather than true reasoning. Thus, GenAI drafts may often contain useful portions such as expressions of respect or explanation of common next steps, but our results suggest that it would be common to mishandle other aspects of communication important for patient care within the same draft.

Our observations align with others' reports of GenAI draft evaluations in real health care settings. In one study, physicians rated GenAI responses more highly in communication quality than actual responses, likely reflecting GenAI strengths in rapport-building statements [17]. In another study, actual physician responses were rated higher in informational and emotional aspects than GenAI responses to negatively charged patient messages [28], likely reflecting limitations in judicious information delivery and responding to emotion. Other investigators observed that when GenAI responses to hypothetical patient questions had the potential for harm, it was because of inadequate assessment or communication of the acuity of the clinical scenario [29], reflecting that the limitation in judgment seemed to underlie the limitations we observed in our study.

Notably, our results also reiterate that GenAI draft usage does not imply high draft communication quality. Reviewers in our study indicated that they would use almost 60% of GenAI drafts, although 73% of drafts had at least 1 limitation in communication. On one hand, this may suggest that despite their limitations, GenAI drafts have some utility for busy clinicians. On the other hand, the use of drafts that fall short of communication standards could introduce various risks, such as anchoring or automation bias [29]. Careful proofreading is needed to identify potential misinformation [30]. Reading, fact-checking, and revising GenAI drafts may introduce new cognitive burdens and could explain why objective measures of clinician EHR usage have not been impacted by introducing GenAI draft support [14,15] and why physicians were not likely to recommend the tool to others [16]. Future studies of the impact of GenAI tools on cognitive load are needed.

Our exploratory analysis suggests that some limitations, at least in information delivery, could be overcome by providing more comprehensive access to information in the chart, including notes and other messages. Improvements might also be achieved by providing the model with more granular information about local practices and policies to better direct patients. However, those marginal benefits would have to be weighed against the costs of continuous prompt engineering, including supporting local prompt engineers and charges per token of model input and output. Other reasons we identified for limitations in information delivery, such as failure to address the essence of the patient's question or the information being wrong, would be difficult to correct. Future advancements in large language model technology, and use of models that are trained specifically using health care communication data, may mitigate current limitations.

Given some of the observed limitations of GenAI responses, there are several potential ways forward for using the current technology in primary care. One is to restrict application to specific contexts or tasks, or for a narrower scope of questions. In an application of this tool in urology, drafts were more likely to be rated as acceptable by urologists when the question was "easy," compared with questions that were deemed "difficult" (47% vs 34%) [31]. Better identifying primary care messages appropriate for GenAI response may improve performance; it is important to note that in our study, GenAI performance was similar between clerical and general patient messages, at least as categorized by the vendor, suggesting that alternative classification schemes would be needed to identify messages appropriate for GenAI response. Another option may be to apply GenAI as a proofreading tool, rather than a draft-generating tool, so that it improves the provider-drafted messages for readability, professionalism, or other aspects of communication. A third option may be to have patients self-select questions to ask GenAI versus PCPs, so that patients format questions de novo for input into GenAI, and responses do not suffer from lack of contextual information. Organizations will have to weigh the potential net benefit of these approaches with other potential solutions, such as investing in robust best practices for inbox management [8] or hiring dedicated clinical staff for inbox management.

XSL•FO
RenderX

Strengths and Limitations

Our report has several key strengths. We used a well-accepted "best practice" medical communication framework to assess strengths and limitations, which to our knowledge is a novel application for assessing GenAI. While regulatory oversight of GenAI tools is urgently needed [32], clinical standards must also be used to evaluate and shape their impact [33], and our efforts align with that paradigm. We focused on messaging between patients and PCPs specifically, which is important because prior studies suggest that use and perceptions of GenAI responses differ by specialty and health care roles [14,16]. We assessed GenAI responses to real patient messages, rather than priming them for GenAI response, and those patient messages were purposively sampled to maximize variation in the sample and thus expose both strengths and limitations of the model. The GenAI drafts we assessed benefited from multiple prior rounds of prompt engineering, ensuring that the GenAI output we evaluated had been carefully tailored for the task of responding to primary care messages as PCPs. This tailoring may explain why the draft usage rate in our study was so high, compared with 12% - 20% reported in other studies [14,16].

There are also important limitations to our work. This represents the experience of a single institution applying GenAI in a specific context, for example, for messages between patients and PCPs with limited chart information, and experiences may differ with newer GPT models and other applications and contexts, including specialty care and inpatient versus emergency care settings. Patient messages to only 6 PCPs were included in the study, and the sample of 201 GenAI drafts could be considered small; however, we maximized variation within the sample as described, which other studies have not done, and our sample size is similar or larger than published studies on this topic (50 - 175 GenAI messages) [17,28]. Another limitation is that we used a specific medical communication framework, and themes and constructs may have differed with the application of a different framework. We believe that our selection was justified because the "best practice" framework we used synthesizes many common communication frameworks. The same constructs in the "best practice framework, which addresses traditional face-to-face patient-physician encounters, are also applicable to inbox communication [34,35]. Some improvement in GenAI responses might have been achieved if the model were provided greater access to contextual clinical information beyond the limited elements used in our model, such as chart notes. However, as described, such access was limited by institutional data privacy policies, costs, and mode interval time limits, and these constraints are likely to be common to many other institutions. Finally, the tool we piloted was based on GPT 3.5; advances in the model are likely to overcome some of the limitations we observed.

Conclusions

Our results suggest when GenAI was used to reply to patient messages on behalf of PCPs, with limited access to chart information and after robust rounds of prompt optimization, communication standards were often met and often compromised. GenAI response strengths were in making polite statements and outlining common next steps. However, frequently observed limitations were related to communication domains requiring clinician judgment, namely, asking appropriate follow-up questions, providing the right information for patients' informational needs, and recognizing and responding to emotions. Overall communication quality seems unpredictable. These substantial limitations may limit GenAI draft utility.

Authors' Contributions

NSL, JG, RST, and DEJ contributed to conceptualization; NSL and AKH participated in data curation; NSL did the formal analysis; NSL and NR did the investigation; NSL, NR, and DEJ contributed to methodology; RST, JG, and DEJ led the supervision; NSL, NR, JG, and RMC contributed to visualization; NSL participated in writing; and NR, JG, RMC, AKH, RST, and DEJ participated in review and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Observed frequency of strengths and limitations across communication domains for all messages (n=201).
[JPG File, 55 KB - formative_v9i1e71966_app1.jpg]

References

1. Porter J, Boyd C, Skandari MR, Laiteerapong N. Revisiting the time needed to provide adult primary care. J Gen Intern Med 2023 Jan;38(1):147-155. [doi: 10.1007/s11606-022-07707-x] [Medline: 35776372]
2. Dyrbye LN, Gordon J, O'Horo J, et al. Relationships between EHR-Based audit log data and physician burnout and clinical practice process measures. Mayo Clin Proc 2023 Mar;98(3):398-409. [doi: 10.1016/j.mayocp.2022.10.027] [Medline: 36868747]
3. Adler-Milstein J, Zhao W, Willard-Grace R, Knox M, Grumbach K. Electronic health records and burnout: time spent on the electronic health record after hours and message volume associated with exhaustion but not with cynicism among primary care clinicians. J Am Med Inform Assoc 2020 Apr 1;27(4):531-538. [doi: 10.1093/jamia/ocz220]

XSL•FO
RenderX

4. Tai-Seale M, Dillon EC, Yang Y, et al. Physicians' well-being linked to in-basket messages generated by algorithms in electronic health records. Health Aff (Millwood) 2019 Jul;38(7):1073-1078. [doi: 10.1377/hlthaff.2018.05509] [Medline: 31260371]
5. Rotenstein LS, Holmgren AJ, Downing NL, Longhurst CA, Bates DW. Differences in clinician electronic health record use across adult and pediatric primary care specialties. JAMA Netw Open 2021 Jul 1;4(7):e2116375. [doi: 10.1001/jamanetworkopen.2021.16375] [Medline: 34241631]
6. Rotenstein LS, Holmgren AJ, Downing NL, Bates DW. Differences in total and after-hours electronic health record time across ambulatory specialties. JAMA Intern Med 2021 Jun 1;181(6):863-865. [doi: 10.1001/jamainternmed.2021.0256] [Medline: 33749732]
7. National Academy of Medicine. National Plan for Health Workforce Well-Being: The National Academies Press; 2024.
8. AMA STEPS forward. American Medical Association. 2025. URL: https://edhub.ama-assn.org/steps-forward [accessed 2025-06-23]
9. Office of the Surgeon General (OSG). Addressing Health Worker Burnout: The US Surgeon General's Advisory on Building a Thriving Health Workforce [Internet]: US Department of Health and Human Services; 2022. [Medline: 37792980]
10. Gerteis J, Booker C, Brach C, et al. Burnout in Primary Care: Assessing and Addressing It in Your Practice: Agency for Healthcare Research and Quality; 2023.
11. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023 Jun 1;183(6):589-596. [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
12. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
13. Hoppe JM, Auer MK, Strüven A, Massberg S, Stremmel C. ChatGPT with GPT-4 outperforms emergency department physicians in diagnostic accuracy: retrospective analysis. J Med Internet Res 2024 Jul 8;26:e56110. [doi: 10.2196/56110] [Medline: 38976865]
14. Garcia P, Ma SP, Shah S, et al. Artificial intelligence-generated draft replies to patient inbox messages. JAMA Netw Open 2024 Mar 4;7(3):e243201. [doi: 10.1001/jamanetworkopen.2024.3201] [Medline: 38506805]
15. Tai-Seale M, Baxter SL, Vaida F, et al. AI-generated draft replies integrated into health records and physicians' electronic communication. JAMA Netw Open 2024 Apr 1;7(4):e246565. [doi: 10.1001/jamanetworkopen.2024.6565] [Medline: 38619840]
16. English E, Laughlin J, Sippel J, DeCamp M, Lin CT. Utility of artificial intelligence-generative draft replies to patient messages. JAMA Netw Open 2024 Oct 1;7(10):e2438573. [doi: 10.1001/jamanetworkopen.2024.38573] [Medline: 39401041]
17. Small WR, Wiesenfeld B, Brandfield-Harvey B, et al. Large language model-based responses to patients' in-basket messages. JAMA Netw Open 2024 Jul 1;7(7):e2422399. [doi: 10.1001/jamanetworkopen.2024.22399] [Medline: 39012633]
18. King A, Hoppe RB. "Best practice" for patient-centered communication: a narrative review. J Grad Med Educ 2013 Sep;5(3):385-393. [doi: 10.4300/JGME-D-13-00072.1] [Medline: 24404300]
19. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 2023 Mar 30;388(13):1233-1239. [doi: 10.1056/NEJMsr2214184] [Medline: 36988602]
20. Yan S, Knapp W, Leong A, et al. Prompt engineering on leveraging large language models in generating response to InBasket messages. J Am Med Inform Assoc 2024 Oct 1;31(10):2263-2270. [doi: 10.1093/jamia/ocae172] [Medline: 39028970]
21. Ogrinc G, Davies L, Goodman D, Batalden P, Davidoff F, Stevens D. SQUIRE 2.0 (Standards for QUality Improvement Reporting Excellence): revised publication guidelines from a detailed consensus process. BMJ Qual Saf 2016 Dec;25(12):986-992. [doi: 10.1136/bmjqs-2015-004411] [Medline: 26369893]
22. EQUATOR. EQUATOR network. URL: https://www.equator-network.org [accessed 2025-06-23]
23. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159-174. [Medline: 843571]
24. Klein D. Implementing a general framework for assessing interrater agreement in Stata. Stata J 2018 Dec;18(4):871-901. [doi: 10.1177/1536867X1801800408]
25. Kim J, Chen ML, Rezaei SJ, et al. Perspectives on artificial intelligence-generated responses to patient messages. JAMA Netw Open 2024 Oct 1;7(10):e2438535. [doi: 10.1001/jamanetworkopen.2024.38535] [Medline: 39412810]
26. Yalamanchili A, Sengupta B, Song J, et al. Quality of large language model responses to radiation oncology patient care questions. JAMA Netw Open 2024 Apr 1;7(4):e244630. [doi: 10.1001/jamanetworkopen.2024.4630] [Medline: 38564215]
27. Bernstein IA, Zhang YV, Govil D, et al. Comparison of ophthalmologist and large language model chatbot responses to online patient eye care questions. JAMA Netw Open 2023 Aug 1;6(8):e2330320. [doi: 10.1001/jamanetworkopen.2023.30320] [Medline: 37606922]
28. Baxter SL, Longhurst CA, Millen M, Sitapati AM, Tai-Seale M. Generative artificial intelligence responses to patient messages in the electronic health record: early lessons learned. JAMIA Open 2024 Jul;7(2):ooae028. [doi: 10.1093/jamiaopen/ooae028] [Medline: 38601475]

XSL•FO
RenderX

29. Chen S, Guevara M, Moningi S, et al. The effect of using a large language model to respond to patient messages. Lancet Digit Health 2024 Jun;6(6):e379-e381. [doi: 10.1016/S2589-7500(24)00060-8] [Medline: 38664108]
30. Alber DA, Yang Z, Alyakin A, et al. Medical large language models are vulnerable to data-poisoning attacks. Nat Med 2025 Feb;31(2):618-626. [doi: 10.1038/s41591-024-03445-1] [Medline: 39779928]
31. Scott M, Muncey W, Seranio N, et al. Assessing artificial intelligence-generated responses to urology patient in-basket messages. Urol Pract 2024 Sep;11(5):793-798. [doi: 10.1097/UPJ.0000000000000637] [Medline: 39162591]
32. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med 2023 Jul 6;6(1):120. [doi: 10.1038/s41746-023-00873-0] [Medline: 37414860]
33. Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. JAMA 2017 Aug 8;318(6):517-518. [doi: 10.1001/jama.2017.7797] [Medline: 28727867]
34. Allen MR, Schillinger D, Ayers JW. The CREATE TRUST communication framework for patient messaging services. JAMA Intern Med 2024 Sep 1;184(9):999-1000. [doi: 10.1001/jamainternmed.2024.2880] [Medline: 39073805]
35. Lee WW, Sulmasy LS, American College of Physicians Ethics, Professionalism and Human Rights Committee. American College of Physicians ethical guidance for electronic patient-physician communication: aligning expectations. J Gen Intern Med 2020 Sep;35(9):2715-2720. [doi: 10.1007/s11606-020-05884-1] [Medline: 32572765]

Abbreviations

EHR: electronic health record
GenAI: generative artificial intelligence
PCP: primary care physician and advanced practice provider

Edited by J Sarvestan; submitted 31.01.25; peer-reviewed by DS Powell, N Hu, S Li; revised version received 30.04.25; accepted 14.05.25; published 31.07.25.

Please cite as:
Lee NS, Richards N, Grandominico J, Cronin RM, Hendricks AK, Tripathi RS, Jonas DE
Use of a Medical Communication Framework to Assess the Quality of Generative Artificial Intelligence Replies to Primary Care Patient Portal Messages: Content Analysis
JMIR Form Res 2025;9:e71966
URL: https://formative.jmir.org/2025/1/e71966
doi:10.2196/71966

© Natalie S Lee, Nathan Richards, Jodi Grandominico, Robert M Cronin, Amanda K Hendricks, Ravi S Tripathi, Daniel E Jonas. Originally published in JMIR Formative Research (https://formative.jmir.org), 31.7.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Acceptance, Drivers, and Barriers to Use of mHealth Apps to Improve Quality of Life in Female Patients Affected by Hypothyroidism: Cross-Sectional Study

Moritz Doll[1]; Ranujan Chandrakumar[1,2], BSc; Lisa Maria Jahre[1,2], MSc; Eva-Maria Skoda[1,2], MD; Hannah Dinse[1,2], MD; Dagmar Führer[3], MD, PhD; Eleni Lampropoulou[3], MD; Martin Teufel[1,2], MD; Alexander Bäuerle[1,2], PhD

[1]Clinic for Psychosomatic Medicine and Psychotherapy, LVR-University Hospital Essen, University of Duisburg-Essen, Virchowstr. 174, Essen, Germany

[2]Center for Translational Neuro- and Behavioral Sciences (C-TNBS), University of Duisburg-Essen, Essen, Germany

[3]Department of Endocrinology, Diabetes and Metabolism, University of Duisburg-Essen, University Hospital Essen, Essen, Germany

Corresponding Author:
Alexander Bäuerle, PhD
Clinic for Psychosomatic Medicine and Psychotherapy, LVR-University Hospital Essen, University of Duisburg-Essen, Virchowstr. 174, Essen, Germany

Abstract

Background: Hypothyroidism is a common chronic disease that can substantially impair physical and mental well-being and is associated with lower quality of life, a trend that interventions delivered by mobile health (mHealth) apps could ameliorate.

Objective: The objective of this study was to evaluate the acceptance and its influencing predictors of mHealth interventions in female patients affected by hypothyroidism to improve their quality of life. The focus on female patients reflects the significantly higher prevalence of hypothyroidism in women and their underrepresentation in many prior studies on technology acceptance and mHealth use.

Methods: A survey-based, cross-sectional study, which included 318 female patients affected by hypothyroidism (assessed via self-reported diagnosis according to International Classification of Diseases-10 criteria, aged 18 y or older), was conducted online between April 2023 and April 2024 in Germany. Participants were recruited via local and online self-help groups, social media platforms, and medical practices using flyers. Sociodemographic, health, and eHealth–related data were assessed. To determine acceptance and its drivers and barriers, an extended version of the unified theory of acceptance and use of technology (UTAUT) model was applied. Group comparisons (*t* tests, ANOVAs) and multiple hierarchical regression analyses were conducted. Only complete datasets were included in the analysis.

Results: Acceptance of mHealth apps was high (mean 4.10, SD 0.91), with 76.1% (n=242) of the participants reporting high acceptance, 18.6% (n=59) reporting moderate acceptance, and only 5.3% (n=17) reporting low acceptance. Significant predictors of acceptance were place of residence: medium-sized city (β=0.34; *P*=.02) and small town or rural area (β=0.28; *P*=.003), fatigue (β=0.54; *P*<.001), internet anxiety (β=−0.20; *P*=.002), and the UTAUT predictors effort expectancy (β=0.37; *P*<.001), performance expectancy (β=0.32; *P*<.001), and social influence (β=0.20; *P*<.001). The extended model explained 56.1% of the variance in acceptance.

Conclusions: The high level of acceptance of mHealth apps observed among female patients affected by hypothyroidism indicates that mHealth interventions can provide such patients with valuable support to manage the disease and improve their quality of life. Addressing drivers and barriers of acceptance will be crucial for the successful implementation of mHealth interventions in hypothyroidism management, for example, by mHealth developers, clinicians, or policy makers. These include intuitive and accessible design (effort expectancy), clear communication of app benefits (performance expectancy), and fostering health care professional support (social influence), while addressing barriers such as internet anxiety. The study also contributes to advancing gender-sensitive mHealth research by applying the UTAUT model to this patient group.

(JMIR Form Res 2025;9:e67317) doi:10.2196/67317

KEYWORDS

hypothyroidism; mHealth; acceptance; unified theory of acceptance and use of technology; quality of life; digital health; chronic disease; female patients; cross-sectional study; mobile health; mobile phone

XSL•FO
RenderX

Introduction

Background

Due to their high prevalence, large parts of the health care system are involved in treating and managing thyroid diseases, as about 0.2% to 5.3% in Europe, 0.3% and 3.7% in the United States, and approximately 3% worldwide are affected by hypothyroidism [1,2]. In Germany, 8.9% of female and 1.8% of male patients are diagnosed with a thyroid disorder [3]. Among all population groups, female and older adults have an especially elevated risk of thyroid disease [4-7].

For patients, thyroid disease can have a wide range of consequences, especially for physical and mental health. Frequent symptoms of overt hypothyroidism include obesity, fatigue, infertility or subfertility, and various cardiovascular symptoms such as bradycardia [8,9]. Thyroid hormone imbalance is further associated with mental health comorbidities, including depression and anxiety, particularly when untreated [10-13], and is overall linked to a decreased quality of life [14].

Treating hypothyroidism typically involves thyroid hormone replacement therapy with levothyroxine [15]. Such therapy requires administering the correct dose of levothyroxine, which can be determined and aggravated by various factors, including body weight, etiology of hypothyroidism, age, and clinical context [15]. Underdosing leaves patients hypothyroid, while overdosing may induce iatrogenic hyperthyroidism [16].

However, even if the correct dose of levothyroxine is administered, as evidenced by normalized serum thyroid stimulating hormone levels, about 10% - 15% of the patients remain dissatisfied due to persistent symptoms of hypothyroidism [15,17-19]. Moreover, studies report that only 59% of patients prescribed levothyroxine were fully adherent [20], and 22% (72/327) of participants in another study reported not adhering to thyroxine replacement therapy [21]. Underlying reasons for nonadherence have included forgetfulness and carelessness, and non- or poor adherence has also been associated with a lower level of education, long-term therapy (ie, >12 months) with levothyroxine, and unattended medical appointments, as investigated in a study [22]. Concerns about medication, as well as affordability and accessibility, are additional factors that influence adherence [23]. The wide range of symptoms and comorbidities, along with the challenges in adjusting therapy associated with poor compliance, underscore the need to support patients with thyroid disease at higher levels.

In recent years, mobile health (mHealth) apps have emerged within the health care systems and have been the subject of diverse studies. Since more than 4.7 billion people worldwide use smartphones, mHealth offers a widely accessible, cost-effective, and sustainable way of managing chronic diseases [24-26]. Improving diagnostics by using symptom checkers, implementing interventions for behavioral changes, improving medication adherence, and providing "disease-related education," as well as facilitating digital exchanges with health care providers, showcases the value of mHealth apps in managing chronic diseases [24,27]. For patients with thyroid disorders, the benefits are numerous. One study evaluating the mHealth app "BOOST Thyroid App" revealed an increase in health literacy, improved patient-doctor interaction (including fewer or shorter appointments), and enhanced overall quality of life after using the mHealth app [28]. In the study, 95.8% of patients reported that the mHealth solution generally supported them in managing their disease [28]. Overall, these study results clarify the usefulness of mHealth interventions in supporting patients with thyroid disease.

However, implementing mHealth interventions in routine care presents several obstacles, including concerns about data privacy, trust in online information, and smartphone access [29,30]. Older adults, in particular, tend to be insufficiently experienced with mHealth apps or may not be aware of their existence, and the lack of face-to-face communication presents a barrier for older people [31].

To sustainably implement mHealth apps for patients affected by hypothyroidism despite these challenges, the acceptance of mHealth apps and influencing factors need to be investigated. However, at present, no study has evaluated the acceptance of mHealth interventions to improve the quality of life of female patients affected by hypothyroidism by using acknowledged measurement methods. This study focuses on female patients with hypothyroidism for various reasons: first, since hypothyroidism is significantly more prevalent among women than men, the acceptance of mHealth interventions by female patients is of major clinical importance [5,6]. Second, female patients have historically been underrepresented in clinical and digital health research, which can lead to gender bias in the development and implementation of health technologies [32,33]. Addressing this gap aligns with the principles of gender medicine and ensures that mHealth solutions are developed based on evidence that adequately reflects the experiences and needs of female patients with hypothyroidism [34]. Even so, several variables to determine the acceptance of technology can be approached by the unified theory of acceptance and use of technology (UTAUT) [35], including the following variables: performance expectancy (PE), effort expectancy (EE), social influence (SI), facilitating conditions, and behavioral intention (BI) to use (ie, operationalization of acceptance) new technology [36-38]. Regarding mHealth (eg, mHealth interventions) apps in particular, various studies have also shown the usefulness of the UTAUT in investigating those variables [38-41].

Objectives

This study aimed to investigate the acceptance of mHealth interventions among female patients affected by hypothyroidism and factors contributing to such acceptance as a means of improving their quality of life. In addition, questions regarding the lack of mHealth interventions implemented in consideration of their acceptance and contributing factors of such acceptance have thus far remained unanswered. In doing so, we used an extended UTAUT model. The additional variables capture sociodemographic and medical issues, as well as information and communication technology–related data, that is, the technical basis of communication via mHealth apps.

This study focuses on the following questions: (1) What is the general level of acceptance of mHealth apps among female patients affected by hypothyroidism? (2) To what extent do

XSL•FO
RenderX

female patients affected by hypothyroidism differ in their acceptance regarding sociodemographic and medical factors? and (3) What are the key facilitators and obstacles affecting the acceptance of mHealth apps?

Methods

Study Design

A survey-based, cross-sectional study was conducted to assess the acceptance of mHealth apps to improve the quality of life of female patients affected by hypothyroidism. From April 2023 to April 2024, participants were primarily recruited via local and analog self-help groups (eg, Schilddrüsenliga e.V.) and social media platforms (eg, Facebook), supplemented by recruitment through medical practices using a flyer containing concise information regarding the study's background and responsible institutions. The flyer included a QR code and a link directing to the web-based questionnaire. The survey was conducted on the platform Unipark (Tivian XI GmbH).

Ethical Considerations

All participants electronically provided their informed consent prior to the start of the survey. Participation in the survey was voluntary, anonymous, and without compensation. The study was conducted in accordance with the Declaration of Helsinki and was approved by the Ethics Committee of the Medical Faculty of the University of Duisburg-Essen (19-89-47-BO).

Procedures and Participants

On average, completing the questionnaire took approximately 17 (SD 8) minutes. Initially, 532 participants started processing the questionnaire, of which 362 participants completed the survey, representing a 68.1% completion rate. Participants who were female, consented to participate, had a self-reported diagnosis of hypothyroidism (according to World Health Organization International Classification of Diseases-10: E03 and E89.0 [42]), legal age of 18 years or older, had sufficient German-language skills, and had internet access were eligible to participate. Only participants who completed the full assessment were included. Thus, 44 (12.2%) participants were excluded due to not meeting the inclusion criteria. Therefore, 318 (59.8%) participants were included in the final data analysis. To maintain the required high methodological standards, the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) Statement checklist (Checklist 1) was applied [43].

Assessment Instruments

Overview

The questionnaire sought to gather sociodemographic, medical, and eHealth-related data. The acceptance of eHealth interventions among female patients affected by hypothyroidism and its drivers and barriers were determined by using a modified version of the UTAUT model.

Sociodemographic and Medical Data

Sociodemographic data gathered included age, gender, marital status, level of education, occupational status, and place of residence (population size). To gauge the participants' health status, we assessed the existence of hypothyroidism and its symptoms by self-report. In addition, participants were asked whether they had received hormone replacement therapy, for example, with levothyroxine in the past or were currently receiving it and might have been affected by therapy's side effects. Participants were also asked whether they had previously undergone surgery as a reason for their hypothyroidism. Physical health, mental health, and quality of life were each self-rated on a numeric rating scale from 0 to 10, with lower scores indicating lower levels of the aforementioned constructs.

eHealth-Related Data

Next, the participant's eHealth literacy was assessed using the revised German version of the eHealth Literacy Scale [44], with responses given on a 5-point Likert scale (1=strongly disagree to 5=strongly agree). Its internal consistency was excellent (Cronbach α=0.92). Participants also assessed their digital confidence by rating three items on a 5-point Likert scale (1=very insecure to 5=very confident) [45-47]. Its internal consistency was excellent (Cronbach α=0.91). Next, digital overload was assessed with three items (eg, "I feel burdened by the constant availability via cell phone or email") [45,48,49], with answers given on a 5-point Likert-type scale (1=does not apply to 5=does fully apply). The internal consistency was acceptable (Cronbach α=0.77). Last, internet anxiety was assessed using three items (eg, "I am afraid I might make an irreversible mistake when using the internet") [40,47,48,50,51]. Again, answers were given on a 5-point Likert-type scale (1=does not apply to 5=does fully apply). The internal consistency was acceptable (Cronbach α=0.79).

UTAUT Predictors and Acceptance

To assess the acceptance of mHealth apps and their influencing factors, we used a modified UTAUT questionnaire [36,38]. The modified UTAUT questionnaire contained 14 items—4 assessing BI, 3 assessing SI, 3 assessing PE, and 3 assessing EE—and answers were given on a 5-point Likert scale (1=strongly disagree to 5=strongly agree). Internal consistency ranged from acceptable to excellent (Cronbach α=0.9 for BI; Cronbach α=0.77 for SI; Cronbach α=0.82 for PE; and Cronbach α=0.82 for EE).

Statistical Analyses

Statistical analyses were performed using R (version 4.3.8; R Core Team), whereas sum scores were calculated for eHealth literacy using the revised German version of the eHealth Literacy Scale, and mean scores were calculated for digital confidence, internet anxiety, digital overload, and UTAUT scales (BI, EE, PE, and SI). Acceptance was operationalized as BI and was further categorized in accordance with prior research [45,47,52], such that scores from 1 to 2.34 indicated low acceptance, scores from 2.35 to 3.67 indicated moderate acceptance, and scores from 3.68 to 5 indicated high acceptance. Descriptive statistics were applied for sociodemographic, medical, and eHealth-related data.

Differences in acceptance based on sociodemographic and medical variables were examined with independent 2-tailed *t* tests and ANOVAs. Bonferroni correction was applied to adjust *P* values for multiple comparisons, and Levene's test indicated

XSL•FO
RenderX

homoscedasticity. Due to the given sample size, normal distribution of residuals was assumed.

Multiple hierarchical regression analysis was conducted to examine drivers of and barriers to the acceptance of mHealth apps among female patients affected by hypothyroidism. Predictors were included block-wise: (1) sociodemographic data, (2) medical data, (3) eHealth-related data, and (4) UTAUT predictors. The variance inflation factor was used to verify the absence of multicollinearity (all variance inflation factor values <2.0). A visual inspection of QQ-plots of the residuals showed no signs of violations against normality; therefore, the normal distribution of the residuals was assumed. Scatter plots of the standardized residuals and the adjusted predicted values verified homoscedasticity. Only complete datasets were included in the analysis, so there was no missing data. The level of significance was set to $\alpha<.05$ for all tests. Effect sizes were reported according to Cohen, with values around 0.2, 0.5, and 0.8 indicating small, medium, and large effects, respectively [53].

Results

Study Population

On average, female patients affected by hypothyroidism in our study were 42.6 (SD 10.7) years old. The youngest participant was 20 years old, while the oldest was 71 years old. In the context of participants' hypothyroid disease, the vast majority of participants (n=262, 82.4%) reported fatigue as a symptom, whereas only 18.9% (n=60) reported irregular menstruation as a symptom. Table 1 presents a complete description of the study population. Participants reported moderate physical health (mean 5.45, SD 1.97; range 0 - 10), mental health (mean 5.86, SD 2.45; range 0 - 10), and quality of life (mean 5.94, SD 2.23; range 0 - 10).

In terms of eHealth, participants reported high levels of eHealth literacy (mean 31.85, SD 6.22; range 8 - 40) and high digital confidence (mean 4.13, SD 0.93; range 1 - 5), whereas internet anxiety (mean 1.59, SD 0.70; range 1 - 5) was low, and digital overload was moderate (mean 2.56, SD 0.96; range 1 - 5).

XSL•FO
RenderX

Table . Sample characteristics of female patients affected by hypothyroidism in a cross-sectional study conducted between April 2023 and April 2024 in Germany.

Characteristics	Participants (N=318), n (%)
Marital status	
Single	67 (21.1)
In a relationship	78 (24.5)
Married	147 (46.2)
Divorced or separated	19 (6)
Widowed	4 (1.3)
Other	3 (0.9)
Educational level	
No or lower secondary education or other	23 (7.2)
Higher secondary education	79 (24.8)
Higher education entrance qualification	105 (33)
University education	111 (34.9)
Occupational status	
Student	21 (6.6)
Nonworking	8 (2.5)
Sick leave	14 (4.4)
Part-time employed	89 (28)
Full-time employed	157 (49.4)
Retired	13 (4.1)
Other	16 (5)
Place of residence (population size)	
Large city (>100,000 residents)	98 (30.8)
Medium-sized city (>20,000 residents)	89 (28)
Small town or rural area (<20,000 residents)	131 (41.2)
Symptoms	
Fatigue	262 (82.4)
Weight gain	219 (68.9)
Depressed mood	218 (68.6)
Poor concentration	195 (61.3)
Dry skin	185 (58.2)
Loss of libido	165 (51.9)
Increased sensitivity to cold	155 (48.7)
Hair loss	140 (44)
Constipation	82 (25.8)
Low pulse or low blood pressure	64 (20.1)
Irregular menstruation	60 (18.9)
Hormone replacement therapy	289 (90.9)
Side effects of the hormone replacement therapy	31 (9.1)
Surgical operations	26 (8.2)

XSL•FO
RenderX

Acceptance of mHealth Apps

Overall, the acceptance of mHealth apps among female patients affected by hypothyroidism was high (mean 4.10, SD 0.91; range 1 - 5). More precisely, 76.1% (n=242) of participants reported high acceptance, 18.6% (n=59) reported moderate acceptance, and only 5.3% (n=17) reported low acceptance.

Group comparisons showed differences in acceptance depending on the place of residence ($F_{2,315}$=3.66; P=.03; f=0.15). Posthoc group comparisons showed that participants living in a medium-sized city (>20,000 residents) reported significantly higher rates of acceptance than those living in big cities (>100,000 residents; t_{315}=−26; P_{adj}=.04; d=0.23).

Participants who reported fatigue as a symptom of their hypothyroid disease reported higher levels of acceptance than ones who did not report fatigue (t_{316}=−3.56; P_{adj}<.001; d=0.53). However, differences in acceptance were neither dependent on marital, educational, or occupational status nor dependent on undergoing hormone replacement therapy or surgical operations.

Predictors of Acceptance of mHealth Apps

Multiple hierarchical regression analysis was performed to determine predictors of the acceptance of mHealth apps to improve the quality of life of female patients affected by hypothyroidism. First, sociodemographic data were included (R^2=.034; R^2_{adj}=.015; $F_{6,31}$=1.83; P=.09). Place of residence: medium-sized city (>20,000 residents; β=0.34; P=.02) and small town or rural area (<20,000 residents; β=0.28; P=.003) were significant predictors of acceptance. The explained variance of the first step was 3.4%.

Medical data were included in the second step (R^2=0.086; R^2_{adj}=0.056; $F_{10,307}$=2.88; P=.002). That step increased the explained variance significantly to 8.6% (ΔR^2=0.052; $F_{4,307}$=8.85; P<.001). Fatigue (β=0.54; P<.001) emerged as a significant predictor of acceptance. In the third step, eHealth data (R^2=0.013; R^2_{adj}=0.09; $F_{14,303}$=3 - 25; P<.001) significantly increased the explained variance to 13% (ΔR^2=0.044; $F_{4,303}$=7.63; P<.001). Internet anxiety (β=−0.20; P=.002) was a significant predictor of acceptance.

In the final step, the three UTAUT predictors—EE, PE, and SI—were included (R^2=0.561; R^2_{adj}=0.537; $F_{17,300}$=22.62; P<.001). Explained variance of the final increased significantly to 56.1% (ΔR^2=0.431; $F_{3,300}$=98.42; P<.001). EE (β=0.37; P<.001), PE (β=0.32; P<.001), and SI (β=0.20; P<.001) were significant predictors. Table 2 contains the final UTAUT model of acceptance and its predictors.

XSL•FO
RenderX

Table . Hierarchical regression model of acceptance of mHealth[a] apps among female patients affected by hypothyroidism in a cross-sectional study conducted between April 2023 and April 2024 in Germany (N=318)[b].

Predictors	B	β	*t* test *(df)*	R^2 [c]	ΔR^2[d]	*P* value
Intercept	.88	−0.18	−0.97			.33
Step 1: sociodemographic data				0.034	0.034	
Age	.00	0.03	0.61			.54
Educational level (reference: no or lower secondary education or other)						
Higher secondary education	.02	0.03	0.16			.87
Higher education entrance qualification	−.15	−0.17	−1.04			.30
University education	−.00	−0.01	−0.03			.97
Place of residence (population size; reference: large city; >100,000)						
Medium-sized city (>20,000)	.06	0.07	0.68			.49
Small town or rural area (<20,000)	.03	0.04	0.41			.68
Step 2: medical data				0.086	0.052	
Physical health	.02	0.04	0.78			.44
Mental health	−.00	−0.00	−0.02			.98
Quality of life	−.02	−0.05	−0.97			.33
Fatigue	−22	0.24	2.18			.03
Step 3: eHealth-related data				0.13	0.044	
eHealth literacy	−.00	−0.02	−0.38			.70
Digital confidence	−.00	0.00	0.07			.95
Digital overload	−.01	−0.01	−0.20			.84
Internet anxiety	−.10	−0.08	−1.75			.08
Step 4: UTAUT[e] predictors				0.561	0.431	
EE[f]	.40	0.37	8.04			<.001
PE[g]	.32	0.32	6.69			<.001
SI[h]	.19	0.20	4.38			<.001

[a]mHealth: mobile health.
[b]In Steps 2, 3, and 4, only the newly included variables are presented.
[c]R^2: determination coefficient.
[d]ΔR^2: changes in R^2.
[e]UTAUT: unified theory of acceptance and use of technology.
[f]EE: effort expectancy.
[g]PE: performance expectancy.
[h]SI: social influence.

Discussion

Principal Findings

The purpose of this study was to determine the acceptance of mHealth apps, along with the factors influencing it, among female patients affected by hypothyroidism, with the aim of improving quality of life. Due to the reduced quality of life from its physical and mental symptoms and challenges in therapy (eg, medication adjustment), the results of this study may contribute to improved care for female patients affected by hypothyroidism. The overall acceptance of mHealth was high;

XSL•FO
RenderX

more than three-quarters (n=242, 76.1%) of participants reported high acceptance, while 18.6% (n=59) reported at least moderate acceptance. Compared to similar studies representing different populations of patients with other chronic diseases [45-48], acceptance among female patients affected by hypothyroidism was high, which may result in higher utilization in practical implementation.

Acceptance of mHealth was especially high among patients affected by fatigue, which was also identified as a predictor of acceptance. Fatigue was a common symptom of participants in this study and is generally common in overt hypothyroidism [54,55]. Added to its frequency, fatigue as a symptom of various diseases is primarily associated with a reduction in quality of life [55-57]. As a result, mHealth interventions may contribute to managing symptoms by enhancing understanding and ways of coping with symptoms of fatigue. Regarding physical and mental health, no significant predictor of acceptance was identified, contrary to other studies [45-48]. However, participants reported only moderate levels of physical and mental health. This suggests an unmet need for additional health care offerings, for example, via mHealth interventions.

In regard to sociodemographic characteristics, living in a small or medium-sized city emerged as a significant predictor of acceptance. This relationship might be explained by a potentially lower availability of health care facilities and self-help groups in less populated areas. In response, mHealth interventions could address the lack of health care services and provide health care for patients in less densely populated areas. This interpretation is supported by previous studies, which also reported higher mHealth acceptance in less urbanized settings [46,58]. Unlike other studies, acceptance was not significantly predicted by age [45,47,48,52,59]. This finding is plausible given the growing tendency for older adults to use digital media, which might likely reflect the trend in the future [60]. Overall, the level of acceptance was also high among older female patients affected by hypothyroidism, suggesting that mHealth interventions represent a viable option in this patient collective. Moreover, unlike in previous studies, educational status did not emerge as a significant predictor of acceptance [52,61,62]. It should be noted, however, that 67.9% (n=216) of patients surveyed had at least a higher education entrance qualification or even university education, whereas only 7.2% (n=23) had no or lower secondary education. That phenomenon is commonly observed in research on digital health care [61,62] and suggests that mHealth interventions should be adapted to the needs of less-educated patients.

Concerning eHealth-related data, internet anxiety emerged as an important predictor of acceptance. Overall, levels of internet anxiety were low, which can be partly explained by the fact that the survey was conducted and completed entirely online. However, higher levels of internet anxiety were a significant negative predictor of acceptance. Thus, internet anxiety seems to represent a barrier to using mHealth interventions, which should be taken into account when designing and implementing such interventions [38,40], including concerns about privacy, trust in online information, and smartphone access [29,30]. Internet anxiety also primarily affects older adults [63,64], although a positive trend may be expected in this regard given the increasing number of older smartphone users [60]. Finally, the level of digital confidence and eHealth literacy was high but was not a significant predictor of acceptance, which is contrary to other studies [45,47,48].

The UTAUT model, which aims to predict users' intentions to use an mHealth intervention and their actual use behavior [36], contributed significantly to the explained variance, and the 3 UTAUT predictors (ie, EE, PE, and SI) emerged as significant predictors of acceptance. Those findings are in line with the results of other studies [48,50,65-68]. Of all predictors identified, EE was the most significant predictor of acceptance. Accordingly, acceptance depends on the expected effort involved in interacting with the mHealth intervention. To maximize EE, developers should prioritize intuitive navigation, clear and minimalistic user interfaces, and personalization options that accommodate users with varying levels of digital literacy [69]. Tutorial features and onboarding processes, as well as automated reminders, may further lower the perceived effort required to use the app. A further significant predictor of acceptance was PE, which highlights the importance of ensuring that patients are convinced of the benefits of the intervention. Integrating evidence-based educational content, tracking tools for symptoms and treatment adherence, and feedback mechanisms may help patients recognize the value of using the app [70]. Allowing users to personalize content and adapt app features to their individual needs may further enhance both PE and EE. Finally, SI was a significant predictor of acceptance. Incorporating peer-support features (such as forums or patient communities), enabling positive feedback from health care professionals within the app, and integrating testimonials or success stories can encourage sustained engagement of SI. Partnerships with clinicians who recommend or prescribe the use of validated mHealth tools may further strengthen SI.

For future research, it would first be important to develop and validate a user-centered mHealth app specifically customized for this patient group. Actively involving patients in the design process and testing the app in terms of usability and perceived usefulness would provide a solid foundation for broader implementation. Subsequently, longitudinal studies tracking actual app use would be valuable to examine how initial acceptance translates into sustained engagement and real-world behavioral outcomes. This would also help to further address the intention-behavior gap and support the development of more effective mHealth interventions for patients with hypothyroidism.

In conclusion, this study highlights key factors that should be considered when designing mHealth interventions for female patients affected by hypothyroidism. In particular, EE, PE, and SI emerged as central predictors of acceptance. Ensuring intuitive and accessible design, clearly communicating the benefits of using the app, and fostering support from health care professionals may therefore promote higher acceptance and sustained engagement. Moreover, barriers such as internet anxiety should be addressed to ensure equitable access. Theoretically, this study expands the application of the UTAUT model to this patient group. By demonstrating the model's explanatory power in this gender-specific chronic disease context, our findings contribute to advancing gender-sensitive

XSL•FO
RenderX

mHealth research and underscore the importance of considering sociodemographic and health and mHealth-related factors in technology acceptance studies.

Limitations

When discussing and interpreting our results, for example, for practical implications, the following limitations should be taken into account. First, the questionnaire was only available via internet access, and the diagnosis of hypothyroidism could not be objectively verified, as it is based on self-report. Along those lines, it should also be noted that some participants were recruited online, for example, via online self-help groups and social media, which might have led to a selection bias. Younger patients, accordingly, and participants with a higher level of digital confidence or less internet anxiety may have been overrepresented. While older adults were not actively excluded, the digital format required a minimum level of digital literacy and internet access, which may have limited their participation. However, as digital confidence among older adults continues to increase [71], future research should specifically target this group and individuals facing digital barriers. Second, the intention-behavior gap needs to be considered. Accepting an mHealth intervention leads to a higher intention to use it but does not necessarily result in actual use and behavior to the same degree [72,73]—the actual use should therefore be investigated in future studies. In addition, the generalizability of our findings is limited in several aspects. On the one hand, male patients were not included in this study: not only are female patients significantly more likely to be affected by hypothyroidism [5,6], but similarly designed studies have also shown that these questionnaires are less likely to be completed by male patients [46-48]. However, gender-specific differences in technology acceptance and the experience of hypothyroidism may influence outcomes in this group, and it remains unclear whether similar acceptance samples would emerge in male patients [74]. On the other hand, the sample was predominantly well-educated, which may have contributed to higher levels of digital confidence and mHealth acceptance. However, patients with lower educational levels are less likely to participate in eHealth studies and may face additional barriers that need further investigation [75-77]. Finally, as the study was conducted in a Western, German-speaking context, the findings may not directly translate to non-Western health care settings, where cultural factors, health system structures, and technology access may affect mHealth acceptance differently [78-80]. Future research should explicitly address these aspects to broaden the evidence base.

Conclusions

Altogether, this study showed a generally high level of acceptance regarding mHealth apps to improve the quality of life among female patients affected by hypothyroidism, particularly among those experiencing fatigue. Several predictors of acceptance were identified as well: Along with the UTAUT predictors EE, PE, and SI, not living in a highly populated city was a driver of acceptance. Internet anxiety was found to be a barrier to using mHealth apps. By addressing key acceptance factors, mHealth interventions have the potential to significantly enhance the quality of life of female patients with hypothyroidism.

Acknowledgments

The authors received no financial support for the research, authorship, or publication of this article. DF reports funding by the DFG as part of CRC/TR 296 LOCOTACT. We thank the Open Access Fund of the University of Duisburg-Essen for supporting the open access publication of this article. The authors attest that there was no use of generative artificial intelligence technology in the generation of text, figures, or other informational content of this manuscript.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

AB, MD, HD, and MT were responsible for the conceptualization of the study.RC and LMJ performed data curation and formal analysis.The methodology was developed by AB, RC, and LMJ.AB and MD managed the overall project administration.AB, E-MS, HD, EL, DF, and MT provided supervision throughout the study.The original draft of the manuscript was written by MD, AB, RC, and LMJ.All authors contributed to the review and editing of the manuscript.

Conflicts of Interest

None declared.

Checklist 1
Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Standards checklist.
[PDF File, 406 KB - formative_v9i1e67317_app1.pdf]

References

1. Vanderpump MPJ. The epidemiology of thyroid disease. Br Med Bull 2011;99(39-51):39-51. [doi: 10.1093/bmb/ldr030] [Medline: 21893493]

XSL•FO
RenderX

2. Madariaga AG, Palacios SS, Guillén-Grima F, Galofré JC. The incidence and prevalence of thyroid dysfunction in Europe: a meta-analysis. J Clin Endocrinol Metab 2014 Mar;99(3):923-931. [doi: 10.1210/jc.2013-2409] [Medline: 24423323]
3. Görsch B, Melchert HU, Thierfelder W. Schilddrüsenhormone und schilddrüsenmedikamente bei probanden in den nationalen gesundheitssurveys [thyroid hormones and thyroid medication in participants in the national health surveys] [Article in German]. : Robert Koch Institut; 2002.
4. McGrogan A, Seaman HE, Wright JW, de Vries CS. The incidence of autoimmune thyroid disease: a systematic review of the literature. Clin Endocrinol (Oxf) 2008 Nov;69(5):687-696. [doi: 10.1111/j.1365-2265.2008.03338.x] [Medline: 18673466]
5. Vanderpump MPJ, Tunbridge WMG. Epidemiology and prevention of clinical and subclinical hypothyroidism. Thyroid 2002 Oct;12(10):839-847. [doi: 10.1089/105072502761016458] [Medline: 12487765]
6. Đula IS, Pleić N, Leko MB, et al. Epidemiology of hypothyroidism, hyperthyroidism and positive thyroid antibodies in the Croatian population. Biology (Basel) 2022;11(3):394. [doi: 10.3390/biology11030394] [Medline: 35336768]
7. McLeod DSA, Cooper DS. The incidence and prevalence of thyroid autoimmunity. Endocrine 2012 Oct;42(2):252-265. [doi: 10.1007/s12020-012-9703-2] [Medline: 22644837]
8. Chaker L, Bianco AC, Jonklaas J, Peeters RP. Hypothyroidism. Lancet 2017 Sep 23;390(10101):1550-1562. [doi: 10.1016/S0140-6736(17)30703-1] [Medline: 28336049]
9. Carlé A, Pedersen IB, Knudsen N, Perrild H, Ovesen L, Laurberg P. Gender differences in symptoms of hypothyroidism: a population-based DanThyr study. Clin Endocrinol (Oxf) 2015 Nov;83(5):717-725. [doi: 10.1111/cen.12787] [Medline: 25845636]
10. Hage MP, Azar ST. The link between thyroid function and depression. J Thyroid Res 2012;2012:590648. [doi: 10.1155/2012/590648] [Medline: 22220285]
11. Ittermann T, Völzke H, Baumeister SE, Appel K, Grabe HJ. Diagnosed thyroid disorders are associated with depression and anxiety. Soc Psychiatry Psychiatr Epidemiol 2015 Sep;50(9):1417-1425. [doi: 10.1007/s00127-015-1043-0] [Medline: 25777685]
12. Kathol RG, Delahunt JW. The relationship of anxiety and depression to symptoms of hyperthyroidism using operational criteria. Gen Hosp Psychiatry 1986 Jan;8(1):23-28. [doi: 10.1016/0163-8343(86)90060-5] [Medline: 3943712]
13. Trzepacz PT, McCue M, Klein I, Levey GS, Greenhouse J. A psychiatric and neuropsychological study of patients with untreated Graves' disease. Gen Hosp Psychiatry 1988 Jan;10(1):49-55. [doi: 10.1016/0163-8343(88)90084-9] [Medline: 3345907]
14. Bianchi GP, Zaccheroni V, Solaroli E, et al. Health-related quality of life in patients with thyroid disorders. Qual Life Res 2004 Feb;13(1):45-54. [doi: 10.1023/B:QURE.0000015315.35184.66] [Medline: 15058786]
15. Jonklaas J, Bianco AC, Bauer AJ, et al. Guidelines for the treatment of hypothyroidism: prepared by the American Thyroid Association Task Force on thyroid hormone replacement. Thyroid 2014 Dec;24(12):1670-1751. [doi: 10.1089/thy.2014.0028] [Medline: 25266247]
16. Cooper DS, Biondi B. Subclinical thyroid disease. Lancet 2012 Mar 24;379(9821):1142-1154. [doi: 10.1016/S0140-6736(11)60276-6] [Medline: 22273398]
17. McAninch EA, Bianco AC. The history and future of treatment of hypothyroidism. Ann Intern Med 2016 Jan 5;164(1):50-56. [doi: 10.7326/M15-1799] [Medline: 26747302]
18. Vardarli I, Brandenburg T, Hegedüs L, et al. A questionnaire survey of German thyroidologists on the use of thyroid hormones in hypothyroid and euthyroid patients: the THESIS (Treatment of Hypothyroidism in Europe by Specialists: An International Survey) collaborative. Exp Clin Endocrinol Diabetes 2022 Sep;130(9):577-586. [doi: 10.1055/a-1832-0644] [Medline: 35640637]
19. Hegedüs L, Van Der Feltz-Cornelis CM, Papini E, Nagy EV, Weetman AP, Perros P. Medically not yet explained symptoms in hypothyroidism. Nat Rev Endocrinol 2024 Nov;20(11):685-693. [doi: 10.1038/s41574-024-01022-7] [Medline: 39138377]
20. Bocale R, Desideri G, Barini A, et al. Long-term adherence to levothyroxine replacement therapy in thyroidectomized patients. J Clin Med 2022 Jul 24;11(15):4296. [doi: 10.3390/jcm11154296] [Medline: 35893387]
21. Crilly M, Esmail A. Randomised controlled trial of a hypothyroid educational booklet to improve thyroxine adherence. Br J Gen Pract 2005 May;55(514):362-368. [Medline: 15904555]
22. Shrestha SS, Risal K, Shrestha R, Bhatta RD. Medication adherence to levothyroxine therapy among hypothyroid patients and their clinical outcomes with special reference to thyroid function parameters. Kathmandu Univ Med J (KUMJ) 2018;16(62):129-137. [Medline: 30636753]
23. Kumar R, Shaukat F. Adherence to levothyroxine tablet in patients with hypothyroidism. Cureus 2019 May 8;11(5):e4624. [doi: 10.7759/cureus.4624] [Medline: 31312550]
24. Rowland SP, Fitzgerald JE, Holme T, Powell J, McGregor A. What is the clinical value of mHealth for patients? NPJ Digital Med 2020;3(4):4. [doi: 10.1038/s41746-019-0206-x] [Medline: 31970289]
25. Becker S, Miron-Shatz T, Schumacher N, Krocza J, Diamantidis C, Albrecht UV. mHealth 2.0: experiences, possibilities, and perspectives. JMIR Mhealth Uhealth 2014 May 16;2(2):e24. [doi: 10.2196/mhealth.3328] [Medline: 25099752]
26. Anzahl der smartphone-nutzer weltweit von 2016 bis 2022 und prognose bis 2028 (in milliarden) [number of smartphone users worldwide from 2016 to 2022 and forecast to 2028 (in billions)] [Article in German]. Statista GmbH. URL: https:/

XSL•FO
RenderX

/de.statista.com/statistik/daten/studie/309656/umfrage/prognose-zur-anzahl-der-smartphone-nutzer-weltweit/ [accessed 2024-09-17]
27. Zaugg V, Korb-Savoldelli V, Durieux P, Sabatier B. Providing physicians with feedback on medication adherence for people with chronic diseases taking long-term medication. Cochrane Database Syst Rev 2018 Jan 10;1(1):CD012042. [doi: 10.1002/14651858.CD012042.pub2] [Medline: 29320600]
28. Högqvist Tabor V, Högqvist Tabor M, Keestra S, Parrot JE, Alvergne A. Improving the quality of life of patients with an underactive thyroid through mHealth: a patient-centered approach. Womens Health Rep (New Rochelle) 2021;2(1):182-194. [doi: 10.1089/whr.2021.0010] [Medline: 34235505]
29. Gajanayake R, Iannella R, Sahama T. Consumer acceptance of accountable-eHealth systems. Stud Health Technol Inform 2014;205:980-984. [Medline: 25160334]
30. Jiang Y, West BT, Barton DL, Harris MR. Acceptance and use of eHealth/mHealth applications for self-management among cancer survivors. Stud Health Technol Inform 2017;245:131-135. [Medline: 29295067]
31. Pywell J, Vijaykumar S, Dodd A, Coventry L. Barriers to older adults' uptake of mobile-based mental health interventions. Digit Health 2020;6:2055207620905422. [doi: 10.1177/2055207620905422] [Medline: 32110429]
32. Holdcroft A. Gender bias in research: how does it affect evidence based medicine? J R Soc Med 2007 Jan;100(1):2-3. [doi: 10.1177/014107680710000102] [Medline: 17197669]
33. Mazure CM, Jones DP. Twenty years and still counting: including women as participants and studying sex and gender in biomedical research. BMC Womens Health 2015 Oct 26;15(94):94. [doi: 10.1186/s12905-015-0251-9] [Medline: 26503700]
34. The role of digital health technologies in women's health, empowerment and gender equality. World Health Organization. 2024. URL: https://iris.who.int/bitstream/handle/10665/376179/WHO-EURO-2024-9293-49065-73153-eng.pdf?sequence=2 [accessed 2025-06-09]
35. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989 Sep;13(3):319. [doi: 10.2307/249008]
36. User acceptance of information technology: toward a unified view. MIS Q 2003;27(3):425. [doi: 10.2307/30036540]
37. Ammenwerth E. Technology acceptance models in health informatics: TAM and UTAUT. Stud Health Technol Inform 2019 Jul 30;263:64-71. [doi: 10.3233/SHTI190111] [Medline: 31411153]
38. Philippi P, Baumeister H, Apolinário-Hagen J, et al. Acceptance towards digital health interventions—model validation and further development of the unified theory of acceptance and use of technology. Internet Interv 2021 Dec;26:100459. [doi: 10.1016/j.invent.2021.100459] [Medline: 34603973]
39. Zhang Y, Liu C, Luo S, et al. Factors Influencing patients' intentions to use diabetes management apps based on an extended unified theory of acceptance and use of technology model: web-based survey. J Med Internet Res 2019;21(8):e15023. [doi: 10.2196/15023]
40. Bäuerle A, Frewer AL, Rentrop V, et al. Determinants of acceptance of weight management applications in overweight and obese individuals: using an extended unified theory of acceptance and use of technology model. Nutrients 2022 May 8;14(9):1968. [doi: 10.3390/nu14091968] [Medline: 35565935]
41. Breil B, Kremer L, Hennemann S, Apolinário-Hagen J. Acceptance of mHealth apps for self-management among people with hypertension. Stud Health Technol Inform 2019 Sep 3;267(282-288):282-288. [doi: 10.3233/SHTI190839] [Medline: 31483283]
42. International statistical classification of diseases and related health problems 10th revision (ICD-10): chapter IV endocrine, nutritional and metabolic diseases (E00-E90). World Health Organization. 2019. URL: https://icd.who.int/browse10/2019/en#/E89.0 [accessed 2024-10-01]
43. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet 2007 Oct 20;370(9596):1453-1457. [doi: 10.1016/S0140-6736(07)61602-X] [Medline: 18064739]
44. Marsall M, Engelmann G, Skoda EM, Teufel M, Bäuerle A. Measuring electronic health literacy: development, validation, and test of measurement invariance of a revised German version of the eHealth Literacy Scale. J Med Internet Res 2022 Feb 2;24(2):e28252. [doi: 10.2196/28252] [Medline: 35107437]
45. Bäuerle A, Mallien C, Rassaf T, et al. Determining the acceptance of digital cardiac rehabilitation and its influencing factors among patients affected by cardiac diseases. J Cardiovasc Dev Dis 2023 Apr 17;10(4):174. [doi: 10.3390/jcdd10040174] [Medline: 37103053]
46. Stoppok P, Teufel M, Jahre L, et al. Determining the Influencing factors on acceptance of ehealth pain management interventions among patients with chronic pain using the unified theory of acceptance and use of technology: cross-sectional study. JMIR Form Res 2022 Aug 17;6(8):e37682. [doi: 10.2196/37682] [Medline: 35976199]
47. Nurtsch A, Teufel M, Jahre LM, et al. Drivers and barriers of patients' acceptance of video consultation in cancer care. Digital Health 2024;10:20552076231222108. [doi: 10.1177/20552076231222108] [Medline: 38188860]
48. Schröder J, Bäuerle A, Jahre LM, et al. Acceptance, drivers, and barriers to use eHealth interventions in patients with post-COVID-19 syndrome for management of post-COVID-19 symptoms: a cross-sectional study. Ther Adv Neurol Disord 2023;16:17562864231175730. [doi: 10.1177/17562864231175730] [Medline: 37255668]

XSL•FO
RenderX

49. Rasool T, Warraich NF, Sajid M. Examining the impact of technology overload at the workplace: a systematic review. SAGE Open 2022 Jul;12(3):215824402211143. [doi: 10.1177/21582440221114320]
50. Damerau M, Teufel M, Musche V, et al. Determining acceptance of e-Mental health interventions in digital psychodiabetology using a quantitative web-based survey: cross-sectional study. JMIR Form Res 2021 Jul 30;5(7):e27436. [doi: 10.2196/27436] [Medline: 34328429]
51. Zobeidi T, Homayoon SB, Yazdanpanah M, Komendantova N, Warner LA. Employing the TAM in predicting the use of online learning during and beyond the COVID-19 pandemic. Front Psychol 2023;14:1104653. [doi: 10.3389/fpsyg.2023.1104653] [Medline: 36874866]
52. Hennemann S, Beutel ME, Zwerenz R. Drivers and barriers to acceptance of web-based aftercare of patients in inpatient routine care: a cross-sectional survey. J Med Internet Res 2016 Dec 23;18(12):e337. [doi: 10.2196/jmir.6003] [Medline: 28011445]
53. Cohen J. Statistical Power Analysis for the Behavioral Sciences: Erlbaum; 1988.
54. Alsaud JS, Alharbi MS, Aldekhail AI, et al. Quality of life in thyroid cancer patients. Cureus 2023 Sep;15(9):e45222. [doi: 10.7759/cureus.45222] [Medline: 37842476]
55. Vasudevan MS, Hosalli N. A study of gonadal dysfunction in women with thyroid disease. Int J Adv Med 2022;9(3):243. [doi: 10.18203/2349-3933.ijam20220425]
56. Elbers RG, van Wegen EE, Verhoef J, Kwakkel G. Impact of fatigue on health-related quality of life in patients with Parkinson's disease: a prospective study. Clin Rehabil 2014 Mar;28(3):300-311. [doi: 10.1177/0269215513503355] [Medline: 24113725]
57. Soyuer F. Fatigue and physical performance in the elderly aged 65 and over living in a nursing home. IJFCM 2018;2(6):371-373. [doi: 10.15406/ijfcm.2018.02.00111]
58. Mohajeri D, Jahre LM, Bäuerle A, et al. Acceptance, needs, and demands for nutritional mhealth support in patients with cardiovascular disease. Nutrients 2024 Nov 30;16(23):4155. [doi: 10.3390/nu16234155] [Medline: 39683549]
59. Cimperman M, Brenčič MM, Trkman P. Analyzing older users' home telehealth services acceptance behavior-applying an Extended UTAUT model. Int J Med Inform 2016 Jun;90(22-31):22-31. [doi: 10.1016/j.ijmedinf.2016.03.002] [Medline: 27103194]
60. Faverio. Share of those 65 and older who are tech users has grown in the past decade. : Pew Research Center; 2022 URL: https://www.pewresearch.org/short-reads/2022/01/13/share-of-those-65-and-older-who-are-tech-users-has-grown-in-the-past-decade/ [accessed 2025-07-10]
61. Reiners F, Sturm J, Bouw LJW, Wouters EJM. Sociodemographic factors influencing the use of eHealth in people with chronic diseases. Int J Environ Res Public Health 2019 Feb 21;16(4):645. [doi: 10.3390/ijerph16040645] [Medline: 30795623]
62. Crisp DA, Griffiths KM. Participating in online mental health interventions: who is most likely to sign up and why? Depress Res Treat 2014;2014:790457. [doi: 10.1155/2014/790457] [Medline: 24804089]
63. Nimrod G. Technophobia among older Internet users. Educ Gerontol 2018 Mar 4;44(2-3):148-162. [doi: 10.1080/03601277.2018.1428145]
64. Powell AL. Computer anxiety: comparison of research from the 1990s and 2000s. Comput Human Behav 2013 Nov;29(6):2337-2381. [doi: 10.1016/j.chb.2013.05.012]
65. Rentrop V, Damerau M, Schweda A, et al. Predicting acceptance of e-Mental health interventions in patients with obesity by using an extended unified theory of acceptance model: cross-sectional study. JMIR Form Res 2022 Mar 17;6(3):e31229. [doi: 10.2196/31229] [Medline: 35297769]
66. Rój J. What determines the acceptance and use of eHealth by older adults in Poland? Int J Environ Res Public Health 2022 Nov 24;19(23):15643. [doi: 10.3390/ijerph192315643] [Medline: 36497717]
67. Hoogenbosch B, Postma J, de Man-van Ginkel JM, Tiemessen NA, van Delden JJ, van Os-Medendorp H. Use and the users of a patient portal: cross-sectional study. J Med Internet Res 2018 Sep 17;20(9):e262. [doi: 10.2196/jmir.9418] [Medline: 30224334]
68. Ma Y, Zhou M, Yu W, et al. Using the unified theory of acceptance and use of technology (UTAUT) and e-health literacy(e-HL) to investigate the tobacco control intentions and behaviors of non-smoking college students in China: a cross-sectional investigation. BMC Public Health 2023 Apr 25;23(1):765. [doi: 10.1186/s12889-023-15644-5] [Medline: 37098499]
69. Nasr E, Alsaggaf W, Sinnari D. Developing usability guidelines for mHealth applications (UGmHA). MTI 2023;7(3):26. [doi: 10.3390/mti7030026]
70. Lu T, Lin Q, Yu B, Hu J. A systematic review of strategies in digital technologies for motivating adherence to chronic illness self-care. npj Health Syst 2025;2(1). [doi: 10.1038/s44401-025-00017-4] [Medline: 39872195]
71. Cilli E, Ranieri J, Guerra F, Colicchia S, Di Giacomo D. Digital affinitiy and cognitive reserve: salience for resilient aging in pandemic. Gerontol Geriatr Med 2023;9:23337214231162773. [doi: 10.1177/23337214231162773] [Medline: 36968122]
72. Sheeran P, Webb TL. The intention–behavior gap. Soc Pers Psych 2016 Sep;10(9):503-518. [doi: 10.1111/spc3.12265]
73. Conner M, Norman P. Understanding the intention-behavior gap: the role of intention strength. Front Psychol 2022;13(923464):923464. [doi: 10.3389/fpsyg.2022.923464] [Medline: 35992469]

XSL•FO
RenderX

74. Goswami A, Dutta S. Gender differences in technology usage—a literature review. OJBM 2016;04(1):51-59. [doi: 10.4236/ojbm.2016.41006]
75. López MDPA, Ong BA, Frigola XB, et al. Digital literacy as a new determinant of health: a scoping review. PLOS Digital Health 2023 Oct;2(10):e0000279. [doi: 10.1371/journal.pdig.0000279] [Medline: 37824584]
76. Badr J, Motulsky A, Denis JL. Digital health technologies and inequalities: a scoping review of potential impacts and policy recommendations. Health Policy 2024 Aug;146:105122. [doi: 10.1016/j.healthpol.2024.105122] [Medline: 38986333]
77. Heponiemi T, Jormanainen V, Leemann L, Manderbacka K, Aalto AM, Hyppönen H. Digital divide in perceived benefits of online health care and social welfare services: national cross-sectional survey study. J Med Internet Res 2020 Jul 7;22(7):e17616. [doi: 10.2196/17616] [Medline: 32673218]
78. Goodman A, Mahoney R, Spurling G, Lawler S. Influencing factors to mHealth uptake with indigenous populations: qualitative systematic review. JMIR Mhealth Uhealth 2023 Jun 23;11:e45162. [doi: 10.2196/45162] [Medline: 37351947]
79. Alsswey AH, Al-Samarraie H, El-Qirem FA, Alzahrani AI, Alfarraj O. Culture in the design of mHealth UI. EL 2020 May 13;38(2):257-272. [doi: 10.1108/EL-04-2019-0097]
80. Swahn MH, Gittner KB, Lyons MJ, et al. Advancing mHealth research in low-resource settings: young women's insights and implementation challenges with wearable smartwatch devices in Uganda. Sensors (Basel) ;24(17):5591. [doi: 10.3390/s24175591]

Abbreviations

BI: behavioral intention
EE: effort expectancy
mHealth: mobile health
PE: performance expectancy
SI: social influence
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
UTAUT: unified theory of acceptance and use of technology

Edited by A Mavragani; submitted 08.10.24; peer-reviewed by S Nilkantham; revised version received 18.06.25; accepted 23.06.25; published 07.08.25.

Please cite as:
Doll M, Chandrakumar R, Jahre LM, Skoda EM, Dinse H, Führer D, Lampropoulou E, Teufel M, Bäuerle A
Acceptance, Drivers, and Barriers to Use of mHealth Apps to Improve Quality of Life in Female Patients Affected by Hypothyroidism: Cross-Sectional Study
JMIR Form Res 2025;9:e67317
URL: https://formative.jmir.org/2025/1/e67317
doi:10.2196/67317

© Moritz Doll, Ranujan Chandrakumar, Lisa Maria Jahre, Eva-Maria Skoda, Hannah Dinse, Dagmar Führer, Eleni Lampropoulou, Martin Teufel, Alexander Bäuerle. Originally published in JMIR Formative Research (https://formative.jmir.org), 7.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

A Supplemental Women's Health Questionnaire for Women Veterans With Military Environmental Exposures: Project Development and Implementation

Leah N Eizadi[1,2], MSN, NP, FNP-C; Mehret T Assefa[2], MPH, PhD; Jordan M Nechvatal[2], PhD; G Marina Veltkamp[1,2], PhD; Abou Ibrahim-Biangoro[2], MPH; Maheen M Adamson[2,3], MHL, PhD; Jennifer S Jennings[1,2], MD

[1]California War Related Illness and Injury Study Center, VA Palo Alto Health Care System, United States Department of Veterans Affairs, Palo Alto, CA, United States

[2]Women's Operational Military Exposure Network Center of Excellence, VA Palo Alto Health Care System, United States Department of Veterans Affairs, 3801 Miranda Ave, Palo Alto, CA, United States

[3]Department of Neurosurgery, School of Medicine, Stanford University, Stanford, United States

Corresponding Author:
Leah N Eizadi, MSN, NP, FNP-C
California War Related Illness and Injury Study Center, VA Palo Alto Health Care System, United States Department of Veterans Affairs, Palo Alto, CA, United States

Abstract

Background: The number of women in the armed forces has steadily increased across all branches, even as the overall size of the military remains stable. The population of women veterans is also expanding. The existing literature has extensively reported the impact of military environmental exposures (MEEs) on adverse physical and mental health outcomes in service members and veterans; however, most of these studies focus on the experiences of men. In response to the growing need to address women-specific health care concerns, particularly for women with MEEs, the Women's Operational Military Exposure Network Center of Excellence (WOMEN CoE) developed and implemented the Women's Health Addendum (WHA).

Objective: The primary objective of this project is to (1) describe the development and implementation of a comprehensive health questionnaire for women veterans, (2) systematically describe and characterize the health conditions of women seeking care for MEE-related health concerns, and (3) use findings to inform clinic policies and develop targeted programs.

Methods: The WHA was introduced to assess the prevalence of health conditions that are female-specific, or disproportionately impact women; examine the relationship between these health conditions and MEEs; and use findings to improve care. The WHA was developed through an iterative process, incorporating literature review, veteran and clinician feedback, and clinical expertise. It consists of 81 questions across 7 categories related to health conditions across the lifespan and was implemented in 2 phases. Phase 1 was administered to women at the California War Related Illness and Injury Study Center (WRIISC), and phase 2 included women at the New Jersey and Washington, DC, WRIISC sites. Descriptive findings are presented here.

Results: A total of 63 women participated in the program evaluation from October 2022 to April 2024. In phase 1, 39% (29/75) of the women who were invited agreed to participate. In phase 2, 34 (10%) of the 325 invited veterans responded. Several women's health conditions were reported, with approximately 97% (61/63) of women reporting at least one health condition and 87% (55/63) reporting 3 or more. Among respondents, the most prevalent conditions included sexual dysfunction (23/33, 70%), urinary incontinence (33/56, 59%), pelvic floor dysfunction (33/63, 52%), and pregnancy loss (20/45, 44%). Overall, more than 40% (3/7) of the most frequent conditions were related to urinary health and pelvic floor dysfunction.

Conclusions: Findings highlight the need for services related to women's health, especially for this cohort with MEE concerns seen at a tertiary care center. Initial findings emphasize concerns that women have about fertility and MEE experienced during deployments. Next steps include administering the WHA to women at sister WRIISC sites in real time and establishing a wider distribution network for the WHA. Future efforts to further evaluate the relationship between MEE and women's health concerns are underway.

(JMIR Form Res 2025;9:e73223) doi:10.2196/73223

KEYWORDS

women; veterans; reproductive health; women's health; questionnaire; military environmental exposures

XSL•FO
RenderX

Introduction

Women Veterans

With over 2 million US women veterans, women are the fastest-growing segment of the veteran population and are projected to be 18% of the population by 2040 [1]. According to the Department of Veterans Affairs (VA), women are also increasingly using VA services [1]. At various fiscal year time points, the number of women using VA services nearly doubled or tripled, while the number of men grew substantially slower over the same period [2,3]. This increase represents a more than 300% growth among women using VA primary care [3], yet despite these improvements, women are still a minoritized group with less than 30% of all women veterans enrolled in VA services [2].

Within the VA, women are more likely to have higher outpatient utilization, be younger, and represent a more racially or ethnically diverse group compared to men [2]. For women VA patients, the most prevalent health conditions are endocrine/metabolic/nutritional, musculoskeletal, mental health/substance use disorders, reproductive health, neurologic, sense organ, cardiovascular, and gastrointestinal [3]. For reproductive and sexual health conditions, the most frequent diagnoses include urinary disorders, breast conditions and reproductive organ disorders, menstrual and menopausal disorders, contraceptive care management, and sexually transmitted diseases [4]. As VA projects the resources needed for the future care of the expanding women veteran population, clinical and educational efforts must consider the health problems faced by women.

Military Service and Military Environmental Exposures

Although women have served in the US armed forces for over 200 years, it was not until 1948 that women were made permanent members of all branches of the armed forces, and not until 2013 that women were allowed to serve in combat roles [5]. As the role of women in the military has changed over time, so has their exposure to military environmental exposures (MEEs). Examples of MEEs that women may have been exposed to during military service include air pollutants, chemicals, radiation, occupational hazards, and chemical warfare agents [6]. For women, exposures during military service have been increasing since the 1990s due to changes in policy, length of conflicts, and expanded roles in the military [5,7].

A growing body of literature indicates that an association between MEEs and adverse health problems exists. For example, research on MEEs, physical health, and mental health conditions includes chronic multisymptom illness or Gulf War illness [8,9], impairments in activities of daily living and physical functioning [10], peripheral neuropathy and widespread pain [11], cognitive functioning [12,13], depression [14], anxiety and posttraumatic stress disorder symptoms [14,15], and quality of life [14]. However, most of the MEE literature is based on men and often focused on Vietnam War–era or Gulf War I–era service members and veterans [16]. Although studies have highlighted differences in military service and deployment effects between men and women, more studies are needed to better understand the postdeployment health of women veterans [17-21]. As women transition from service members to veterans, their risk for health concerns that are female-specific or disproportionately impact women may increase, yet few studies have been conducted with women veterans, and even fewer have focused on their reproductive health [22].

Current Status and Gaps

Most of the research conducted among women veterans, or examining sex differences among veterans from the Gulf War and post-9/11 eras, has been observational studies focused on mental health, such as posttraumatic stress disorder, military sexual trauma, and substance use disorders [19,23]. While the number of studies on women veterans is growing, several reviews have reported that gaps in the extant literature on women's health remain and have highlighted the lack of sex-specific results from studies on all veterans [18,23-25]. Many of the reproductive health studies among women veterans used clinical records or qualitative interviews (vs self-reported questionnaires) or examined specific age-related reproductive health issues such as contraception or pregnancy (vs multiple conditions across the lifespan) [22,23]. Topics such as infertility and menopause have been identified as key knowledge gaps in the reproductive health literature, and the need to assess reproductive health outcomes using standardized measures has been highlighted [22]. To our knowledge, there are no standardized women's health questionnaires for women veterans.

WRIISC and WOMEN CoE

Established in 2001, the War Related Illness and Injury Study Center (WRIISC) is a national tertiary care center at VA using a multidisciplinary translational approach to the care of veterans with MEE concerns [6,26,27]. There are 3 sites across the United States (California [CA], New Jersey [NJ], and Washington, DC [DC]) that provide postdeployment health services for the veterans in their catchment areas, including clinical care, education, and research (see Figure 1) [6,26].

XSL•FO
RenderX

Figure 1. WRIISC | WOMEN CoE clinical service areas. Clinical services for the Women's Operational Military Exposure Network Center of Excellence (WOMEN CoE) are provided by the War Related Illness and Injury Study Center (WRIISC).

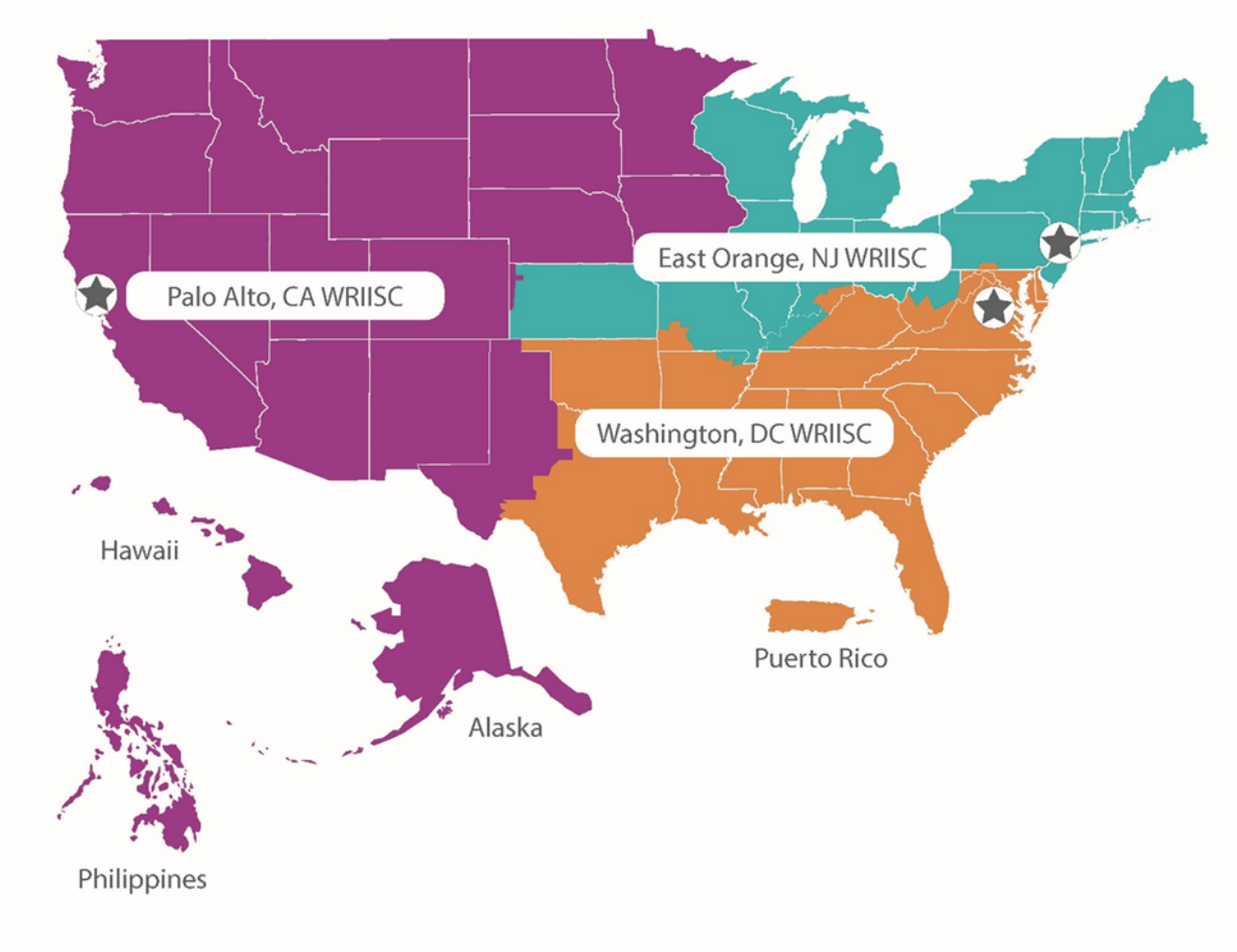

In 2022, VA established the Women's Operational Military Exposure Network (WOMEN) to address the growing health concerns of women veterans with MEEs, and in 2024, it was designated a Center of Excellence (CoE) [28,29]. In alignment with WRIISC, WOMEN CoE provides a multipronged approach to the care of women veterans with the goal of evaluating the health of women with MEE concerns, investigating the impact of MEEs, and offering resources to providers, patients, and key stakeholders [28]. In this report, we describe the activities of the clinical care pillar, which has been described elsewhere (see Figure 2) [28]. Briefly, WOMEN CoE clinical services are provided by WRIISC clinicians during week-long virtual appointments. The Women's Health Addendum (WHA), a clinical tool designed to assess health conditions that are female specific or disproportionately impact women, was developed and implemented by WOMEN CoE.

Figure 2. WOMEN CoE clinical process flowchart. This clinical process flowchart shows the different components that make up the Women's Operational Military Exposure Network Center of Excellence (WOMEN CoE) clinical services. Except for the Women's Health Addendum (WHA), clinical services are provided by the War Related Illness and Injury Study Center (WRIISC).

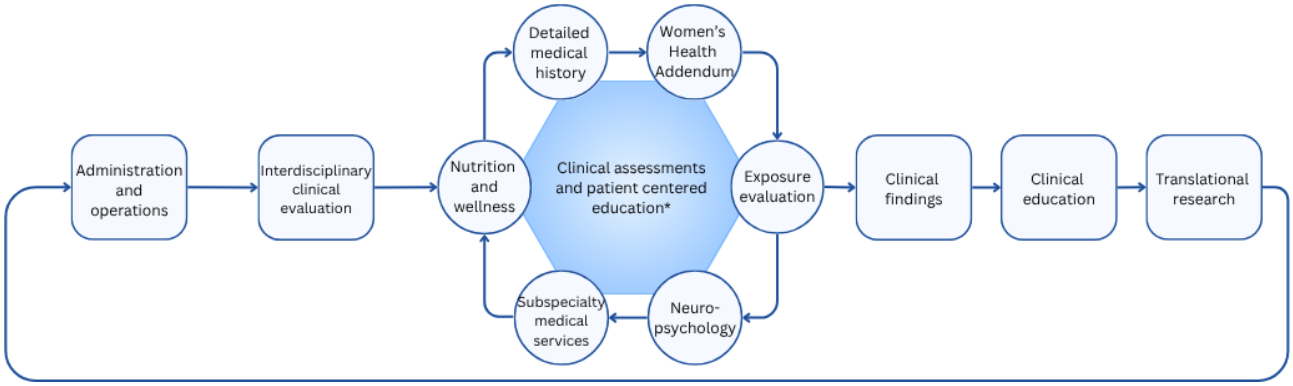

XSL•FO
RenderX

Objective

The primary objective of this project is to (1) describe the development and implementation of a comprehensive health questionnaire for women veterans, (2) systematically describe and characterize the health conditions of women seeking care for MEE-related health concerns, and (3) use findings to inform clinic policies and develop targeted programs to address the most common health concerns among women veterans seen at the WRIISC.

Methods

Women's Health Addendum

Rationale

The WRIISC intake packet was created in 2012 and contains questions on demographics, health concerns, military and deployment information, MEE history, and standardized questionnaires to assess cognitive difficulties, mental health, and physical functioning. The primary purpose of this tool is to aid WRIISC clinicians in their assessment of veterans assigned to a comprehensive, 1-week evaluation. When this questionnaire was originally designed, the target population was primarily men. Since then, the number of women accessing VA services and seeking treatment at the WRIISC has been steadily increasing. While the WRIISC intake packet addressed a wide range of topics, gaps in topics related to women's health existed. Additionally, women veterans across different ages and deployment eras voiced their concerns about MEEs and how they impact their reproductive and sexual health. To effectively address these concerns, the WHA was introduced.

Development of WHA

Development of the WHA began in the summer of 2022. Led by the nurse practitioner, a small group at WOMEN CoE began assessing what was missing from the WRIISC intake packet regarding women's health. Following the initial assessment, an extensive literature review was conducted to examine what questionnaires on women's health, deployment, and MEEs already existed. The literature review included studies of women veterans from the Gulf War and post-9/11 eras as nearly 70% (51/76) of the women seen at WRIISC have served during one of these eras.

Studies with a reproductive health focus were reviewed to assess the scope of the problem and identify the health issues most frequently reported among women veterans [22,30]. Studies that used questionnaires to assess self-reported women's health were examined to either (1) identify a standardized questionnaire or (2) generate a list of topics covered or questions asked to be included in the WHA [17,31]. In studies that used questionnaires to assess women veterans' reproductive health, standardized assessments were not used.

In addition to the literature review, the preliminary version of the WHA included conditions added by the clinical team that were deemed clinically relevant (eg, miscarriages and reproductive cancers) or were frequently discussed during comprehensive visits (eg, incontinence and sexual function).

During pilot testing, the preliminary version of the WHA was administered via video appointments to CA WRIISC women who were attending the week-long telehealth comprehensive appointments. The addendum was administered by the nurse practitioner or the clinical director, and consent was obtained prior to administration. Veteran and clinician feedback was solicited at the end of each appointment. Additional edits were made to the WHA to improve comprehension, ensure comprehensiveness, and maintain a reasonable length.

Based on findings from the literature, feedback from women veterans and clinicians, and clinical expertise, the WHA was developed through an iterative process and designed to work in conjunction with the WRIISC intake packet. The purpose of the WHA is to (1) assess the prevalence of health conditions that are female-specific or disproportionately impact women across the tri-WRIISC, (2) examine the relationship between these health conditions and MEEs, and (3) use findings to inform clinic policies and develop targeted programs to address the most common health concerns among women veterans seen at the WRIISC.

Questions in WHA

Currently, the WHA questionnaire consists of 81 adaptive multiple-choice, yes-no, and fill-in-the-blank questions across 7 categories: sexual health, endocrine health, reproductive health, pregnancy, urinary health, mental health, and menopause. The questions are related to health concerns that are female-specific or that disproportionately impact women across the lifespan. For example, there are questions on age-related conditions (eg, infertility or menopause) and questions on health conditions that can occur at any age (eg, urinary health or sexual health). Descriptions of the categories and sample questions for each are available in Table 1.

XSL•FO
RenderX

Table . Categories and sample questions from the Women's Health Addendum questionnaire.

Categories	Description	Sample question
Sexual health	Questions related to sexual functioning and pelvic floor dysfunction	Have you experienced pelvic floor dysfunction such as chronic pelvic pain or pain with intercourse?
Endocrine health	Questions related to endocrine complications such as endometriosis and polycystic ovary syndrome (PCOS)?	Have you been diagnosed with PCOS?
Reproductive health	Questions related to infertility and cancers	Have you experienced issues with fertility?
Pregnancy	Questions related to pregnancy and pregnancy-related complications	Did you experience gestational diabetes?
Urinary health	Questions related to urinary health such as urinary incontinence and urinary tract infections (UTIs)	Did you experience UTIs during your deployment?
Mental health	Questions related to mental health during pregnancy and menstruation	Have you experienced prenatal depression?
Menopause	Questions related to menopause and postpregnancy, including menopause diagnosis and treatments, and major surgical procedures such as hysterectomy	Have you used, or are you currently using, hormone replacement therapy?

The WHA is organized by category and employs branching logic to improve questionnaire flow and display relevant questions. Due to the sensitive nature of the topics covered, participants can skip questions that are triggering or elicit an unwanted reaction. As a note, because this is an addendum to the WRIISC intake packet, questions on exposures, military history, and other health conditions are not included.

Population

Participants in the WOMEN CoE program evaluation project are WRIISC clinical patients who had a referral to one of three WRIISC sites. Specifically, eligible women veterans are those with medically unexplained symptoms lasting for more than 6 months, difficult-to-diagnose postdeployment health concerns, an overseas deployment, and concerns about MEEs [6,28]. Additional requirements include a completed WRIISC intake packet. WOMEN CoE participants are more likely to be older (over 50 years old), enlisted, Army, non-Hispanic White, and post-9/11 veterans. Additionally, as a tertiary care clinic, WRIISC clinical patients have more complex health problems and are more likely to use VA services than the general veteran population.

Procedures

The WHA was implemented in 2 phases (see Figure 3). In phase 1, WOMEN CoE staff sent introductory letters to former CA WRIISC patients who had completed the WRIISC intake packet announcing the newly established WOMEN CoE, describing the WHA, and explaining its purpose. The letter was signed by the CA WRIISC clinical director and sent via the veteran's patient portal. Following the letter, WOMEN CoE staff invited veterans via phone to complete the WHA. For veterans who agreed to participate, the questionnaire was administered by phone with WOMEN CoE staff, and responses were entered into Qualtrics. Due to the sensitive topics covered in the WHA, women were also provided the option to complete the questionnaire on their own via Qualtrics.

At the end of phase 1, a thorough review of the WHA was conducted to assess survey flow, number and type of questions, and ease of comprehension. Additional edits to the survey were incorporated based on patient and staff feedback. Examples of changes made include rearranging questions within categories, adding questions to the reproductive health section to include surrogacy, and language adjustments to improve comprehension. Due to the positive veteran feedback and interest from NJ and DC WRIISC leadership, phase 2 was implemented.

In phase 2, the WHA was administered via Qualtrics to former DC and NJ WRIISC patients who had completed the WRIISC intake packet. This option was also used for former CA WRIISC patients who were not successfully reached by phone, and for current CA WRIISC patients. To improve response rates for NJ and DC, introductory letters were accompanied by messages from the respective clinical directors. Phase 2 is ongoing, and additional modifications will be made at the end based on feedback. The Checklist for Reporting Results of Internet e-Surveys (CHERRIES) [32] is available in Checklist 1.

XSL•FO
RenderX

Figure 3. Implementation of the Women's Health Addendum, 2022-2024. CA: California; DC: Washington, DC; NJ: New Jersey; WHA: Women's Health Addendum; WOMEN CoE: Women's Operational Military Exposure Network Center of Excellence; WRIISC: War Related Illness and Injury Study Center. *Former CA WRIISC patients who were not successfully reached during phase 1.

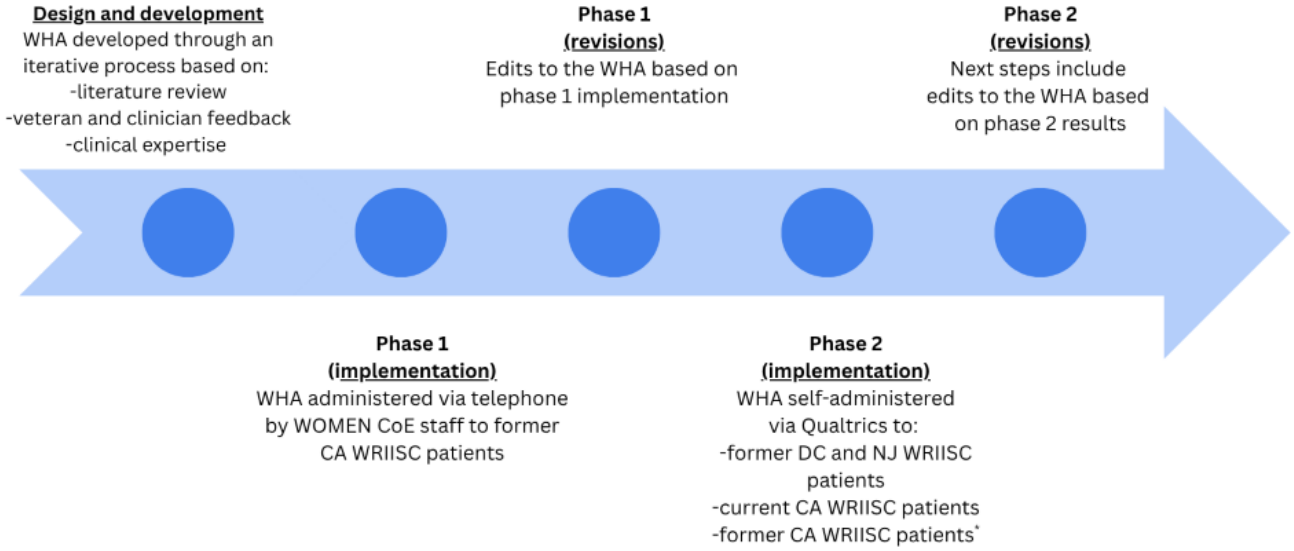

Ethical Considerations

The Stanford institutional review board has reviewed and determined that this project does not meet the definition of human subject research as defined in federal regulations 45 CFR 46.102 or 21 CFR 50.3.

Data Analysis

Findings from the WHA are based on data that were collected as part of a larger program evaluation project to improve care among women veterans seen at the WRIISC. Descriptive statistics are presented for selected results. Data were analyzed using SPSS Statistics Version 29 (IBM Inc).

Results

From October 2022 to April 2024, a total of 63 women veterans completed the WHA. In phase 1, 75 women who had completed the WRIISC intake packet were invited to participate. Of these, 39% (29/75) responded. In phase 2, 34 (10%) of the 325 invited veterans responded.

Table 2 presents self-reported demographic and military characteristics of the sample. More than half (39/63, 62%) of respondents were from the CA WRIISC catchment area, followed by NJ (14/63, 22%) and DC (10/63, 16%) WRIISC sites. The average age was 51.8 years, with 40% (25/63) of women between the ages of 50 and 59 years. Most identified as non-Hispanic White (45/56, 80%) and reported being married (29/59, 49%). Nearly 70% (51/76) served during the post-9/11 (29/76, 38%) or Gulf War I (22/76, 29%) eras. All women in this sample had at least one deployment (range 1 - 5), with an average deployment length of 10.1 months. The majority served in enlisted ranks (48/59, 81%) and on active duty (28/49, 57%) across all branches, but primarily in the Army (29/54, 54%) and Air Force (18/54, 33%). Overall, the length in service was an average of 10.6 years, ranging from 0.3 to 28.1 years.

XSL•FO
RenderX

Table . Demographic and military characteristics of women veterans with military environmental exposures, 2022-2024.[a]

Characteristic	Women (n=63)
Age (years), mean (SD)	51.8 (9.3)
Age (years), n (%)	
30 - 39	9 (14)
40 - 49	14 (22)
50 - 59	25 (40)
60+	15 (24)
Race or ethnicity,[b] n (%)	
American Indian or Alaskan Native	3 (5)
Non-Hispanic Black	3 (5)
Hispanic or Latino	5 (9)
Non-Hispanic White	45 (80)
Marital status,[c] n (%)	
Currently married	29 (49)
Divorced or widowed	13 (22)
Never married	11 (19)
Other[d]	6 (10)
Branch of service,[e] n (%)	
Air Force	18 (33)
Army	29 (54)
Marines	3 (6)
Navy	4 (7)
Military service component,[f] n (%)	
Active duty	28 (57)
Reserves or guard	8 (16)
Both	13 (27)
Service era,[b] n (%)	
Gulf War I	22 (29)
Post-9/11	29 (38)
Other combat deployments[g]	9 (12)
Noncombat deployments	16 (21)
Rank,[c] n (%)	
Enlisted	48 (81)
Officers	11 (19)
Number of deployments,[h] mean (SD)	1.4 (1.0)
Deployment length[i] (months), mean (SD)	10.1 (6)
Service length (years), mean (SD)	10.6 (7.6)
WRIISC[j] site, n (%)	
California	39 (62)
New Jersey	14 (22)
Washington, DC	10 (16)

XSL•FO
RenderX

[a]Includes self-identified demographic and military characteristics.
[b]Missing (n=7).
[c]Missing (n=4).
[d]Other marital status includes living with a partner, separated.
[e]Missing (n=9).
[f]Missing (n=14).
[g]Other combat deployments non-Gulf War I and post-9/11 deployments such as Bosnia, with some participants serving across multiple eras.
[h]Missing (n=12).
[i]Missing (n=16).
[j]WRIISC: War Related Illness and Injury Study Center.

Several health conditions that are female specific or more prevalent in women were reported (see Table 3). Approximately 97% (61/63) reported at least one health condition, and 87% (55/63) reported 3 or more. Among those who responded, the most prevalent conditions included sexual dysfunction (23/33, 70%), urinary incontinence (33/56, 59%), pelvic floor dysfunction (33/63, 52%), pregnancy loss (20/45, 44%), urinary tract infections (23/55, 42%), uterine fibroids (25/62, 40%), and hysterectomy (22/62, 35%). Overall, 43% (3/7) of the top frequent concerns were related to urinary health and pelvic floor dysfunction.

Table . Number of most frequently reported women's health conditions among women veterans with military environmental exposures, 2022-2024 (n=63).[a]

Health conditions	Women, n (%)
Sexual dysfunction (n=33)	
Yes	23 (70)
No	10 (30)
Urinary incontinence (n=56)	
Yes	33 (59)
No	23 (41)
Pelvic floor dysfunction (n=63)	
Yes	33 (52)
No	30 (48)
Pregnancy loss (n=45)	
Yes	20 (44)
No	25 (56)
Urinary tract infections (n=55)	
Yes	23 (42)
No	32 (58)
Uterine fibroids (n=62)	
Yes	25 (40)
No	37 (60)
Hysterectomy (n=62)	
Yes	22 (35)
No	40 (65)

[a]Response totals are noted in parentheses next to each condition. This table depicts the total number of women who reported having at least one of the health conditions listed. The most frequently reported conditions are based on the number of total responses for each health condition.

Discussion

Summary

With the number of women veterans on the rise, understanding their health and MEE concerns is important to address their needs. This comprehensive report captures the self-reported health problems of women with exposure histories. As part of a larger program evaluation project, a new women's health questionnaire was developed by WOMEN CoE to serve as an addendum to the tri-WRIISC intake packet for women. The WHA was developed through an iterative process, incorporating literature review, veteran and clinician feedback, and clinical

XSL•FO
RenderX

expertise. Implementation of the WHA is currently in phase 2, and results have been presented here. The respondents are women with MEE concerns seen at a tertiary VA care center, who primarily served during Gulf War I or post-9/11 eras across all branches as active-duty service members in enlisted ranks. Respondents reported several health concerns that are unique to female physiology or disproportionately impact women that were previously absent from the WRIISC intake packet. Of the nearly 97% (61/63) of women who reported at least one health condition, close to half were related to urinary health, pelvic floor dysfunction, and sexual dysfunction. Given the critical role that the pelvic floor plays in urinary and sexual health, it is not surprising that these are collectively the top health concerns. However, other concerns such as pregnancy loss and uterine fibroids were also frequently reported. Findings from this program evaluation project highlight areas of need for this patient population, ensure that women's health issues are adequately addressed and studied, and improve health care equity for deployed veterans with MEE.

Overall, research studies examining the health of women from the Gulf War and post-9/11 eras have increased in recent years, many of which have focused on mental health [23-25]. For studies among Gulf War veterans, most have highlighted the physical and mental health status of Gulf War veterans, reported on health challenges facing Gulf War veterans such as Gulf War illness and chronic diseases, and have described the prevalence of health conditions and military exposures [17-19,25,33-36]. However, health outcomes specific to women veterans and related to reproductive health or women's life course events have been poorly studied [18]. For studies among post-9/11 veterans, most have focused on various mental health topics (eg, posttraumatic stress disorder, military sexual trauma, and substance use disorder), assessed barriers and facilitators to health care access and use (eg, delayed care, financial concerns, and attrition), examined military exposures and combat-related stressors, and described postdeployment health (eg, chronic multisymptom illness, pain, and traumatic brain injury) [20,21,23,24,37-40]. Studies on the reproductive health of post-9/11 women veterans are also limited [22].

The results of this descriptive program evaluation project on women's health conditions were consistent with findings from the literature. For example, Katon et al [30] also reported a high prevalence of reproductive health conditions. Although the percentage was considerably higher among WHA respondents, Katon et al [30] found that 43% of VA users had at least one reproductive health diagnosis. Moreover, among women with a diagnosis, urinary conditions (including incontinence) were the most common diagnosis across all age groups; however, sexual dysfunction and pregnancy-related conditions were also among the top diagnoses [30]. Other studies among Gulf War and post-9/11 veterans have also reported urinary conditions as top diagnoses [41,42]. In Gaffey et al [37], 26.5% of post-9/11 women veterans reported pelvic pain. Results from studies on pregnancy-related outcomes have been mixed [43]. However, Araneta et al [44] found similar adverse pregnancy outcomes among Gulf War postwar conceptions, including 22.8% spontaneous abortions, 0.7% stillbirths, and 10.7% ectopic pregnancies. In another study among Gulf War veterans [45], pregnancies ending in miscarriage/stillbirth were 25.7% for all respondents, with rates higher among Gulf War veterans compared to Gulf War–era veterans. Although VA hysterectomy rates appear to be decreasing, an elevated risk of early hysterectomy was found in one study among veterans under age 65 compared to non-veterans [22]. Among WHA respondents, 35% reported a hysterectomy. Uterine fibroids were also among the most frequently reported health conditions in this WHA sample. Even at 40%, this is likely an underestimate as many women are asymptomatic [46].

Clinical Implications and Applications

In this clinical program evaluation, nearly 70% (36/53) of respondents stated that women's health was important. This is not surprising given that concerns raised by women veterans were one of the many drivers of the development and implementation of the WHA. Additionally, the high percentage of women that endorsed 3 or more health conditions highlights the need for services related to women's health, especially for this cohort of veterans with exposure histories. Furthermore, studies have shown that women with at least one reproductive health diagnosis are more likely to have concomitant medical/mental health diagnoses and have more VA outpatient encounters compared to those without a reproductive health diagnosis [22]. Although conclusions from this initial assessment are limited, future efforts to evaluate further are underway.

Strengths and Limitations

This clinical program evaluation has several strengths. First, WOMEN CoE was able to identify a gap in the current WRIISC intake packet and take steps to address it by implementing a questionnaire to assess health outcomes that impact women veterans. WOMEN CoE clinical staff were able to incorporate veteran and clinician feedback, literature review, and clinical expertise in its development and implementation. Second, this addendum has gone through several iterations to improve the flow, length, comprehension, and coverage of important topics. At each implementation phase, changes have been identified, and modifications implemented, to improve the WHA and its response rates. Third, although these are descriptive findings and further examination is warranted, the WHA results highlight the clinical needs of women with MEE concerns seen at a WRIISC/WOMEN CoE, and the clinical team can address these by prioritizing services and resources based on these findings.

There are also limitations. One limitation is the small sample size. The sample size was not large enough to assess differences by sociodemographic, military, clinical, or other factors. With a larger sample size, we hope to conduct additional analyses to better understand our patient population. Furthermore, a larger sample size will improve future studies that aim to explore the association between MEEs reported in the WRIISC intake packet and reproductive health conditions reported in the WHA. The small sample size is driven in part by the small number of women seen at WRIISC. While the proportion of women veterans at WRIISC (18%) is higher than that of women across VA at large (9.2%) [2], as a tertiary care center, the total number of WRIISC veterans (men and women) is still small. The low response rate in phase 2 also contributes to the small sample size. At the end of phase 1, the response rate was moderate

XSL•FO
RenderX

(29/75, 39%); however, the current response rate for phase 2 is low (34/325, 10%). Phase 2 is still ongoing, so the response rate might increase. Possible explanations for the low response rate include differences in survey administration, lack of familiarity with WOMEN CoE staff, and greater time since WRIISC visit. Changes to the survey modality were necessitated by the large number of women in phase 2. While WOMEN CoE staff were able to schedule appointments with former CA WRIISC patients in phase 1, by expanding the rollout to include former DC and NJ WRIISC patients, WOMEN CoE staff needed to transition to a web-based questionnaire to ensure proper scale-up. In addition, because of some overlap between CA WRIISC and WOMEN CoE clinicians, CA WRIISC patients might have been familiar with WOMEN CoE staff, which may also account for improved response rates in phase 1. Furthermore, because DC and NJ WRIISC sites were established earlier, former patients may not have had contact with WRIISC in over 23 years, which may have contributed to greater reluctance or a disinclination to respond to the invitations. Next steps include working closely with DC and NJ WRIISC staff to improve response rates among former patients and addressing barriers when expanding to new DC and NJ WRIISC patients.

Another limitation is that results from this program evaluation may not be generalizable to the typical women veteran population. At present, the WHA is specific to women with complex medical histories and exposure concerns seeking care at a specialized tertiary care center. As such, these findings may not be fully generalizable to all women seeking care at VA. Furthermore, the WHA is an addendum, and therefore, extensive information on lifetime and military exposures, military service, and other physical and mental health histories is not available without the WRIISC intake packet. However, plans to link WHA to the WRIISC intake packet are currently underway, which would allow clinicians to work closely with researchers to explore potential associations with MEEs.

Lastly, the data from this program evaluation is based on self-report, which may have introduced bias. However, several studies related to reproductive health have used patient surveys [22]. Furthermore, the WRIISC intake packet is also based on self-report, and studies have shown that self-report can be an important aspect of assessing health [47]. Given the complex medical histories of WOMEN CoE patients, it is not unreasonable to expect higher rates of self-reported health conditions compared to the general women veteran population. Additional evaluation efforts include validating self-reports with electronic medical records.

Future Direction and Next Steps

Future modifications have already been proposed based on preliminary findings from phase 2, including changing the question format for most fill-in-the-blank questions to forced multiple choice. These results have shown that fill-in-the-blank questions have poorer response rates and often include responses that are difficult to interpret and analyze. For example, a question on the number of pregnancies will include “multiple” but not specify the number. In this example, the proposed change would include adding multiple-choice response options (eg, 1, 2, 3, 4, and 5 or more). Another proposed change includes adding additional identifier questions (eg, the last 4 digits of the social security number) to improve links to the WRIISC intake packet. Although the WHA currently includes identifier questions such as name and date of birth, changes in names due to life events (eg, marriage or divorce), typos, or variations (eg, nicknames) may pose a challenge when linking the WHA to the WRIISC intake packet, especially for former patients who completed the WRIISC intake packet many years ago.

Additional next steps include the implementation of WHA for current DC and NJ WRIISC patients. This step requires further coordination with DC and NJ centers to ensure that the questionnaire is sent within 7 days of a patient’s week-long appointment, and additional staffing to ensure timely delivery of the invitation letter and questionnaire. While these implementation barriers have delayed the phase 3 rollout, both WRIISC centers are working closely with the WOMEN CoE to address them and identify facilitators for current and future steps. Next steps also include the dissemination/expansion of the WHA to others. Several VA programs and centers beyond the WRIISC have signaled interest in using the WHA for their patient/participant populations, and some have already begun implementing the questionnaire into their project protocols on women veterans. This and future publications on the WHA are part of an effort to expand access more broadly within VA and beyond.

Conclusion

As part of a program evaluation, WOMEN CoE clinical staff developed and implemented an addendum to the WRIISC intake packet to address the unique health concerns of women veterans. Initial findings highlight the reproductive health care needs of women with MEE concerns. As more women complete the WHA, WOMEN CoE will be able to assess the prevalence of health concerns, identify unmet health needs to prioritize, improve clinical care, evaluate potential associations of health concerns with MEEs, and disseminate findings to veterans and providers.

Acknowledgments

The authors would like to thank the staff at the Women’s Operational Military Exposure Network Center of Excellence and the California War Related Illness and Injury Study Center.

Data Availability

The dataset generated and analyzed during this project is not publicly available due to the Department of Veterans Affairs policy.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Checklist 1
Checklist for Reporting Results of Internet e-Surveys (CHERRIES).
[DOCX File, 20 KB - formative_v9i1e73223_app1.docx]

References

1. Women veterans in focus. Women Veterans Health Care. URL: https://www.womenshealth.va.gov/materials-and-resources/facts-and-statistics.asp [accessed 2025-06-20]
2. Breland JY, Saechao FS, Gujral K, et al. Sourcebook: Women Veterans in the Veterans Health Administration—Volume 5: Longitudinal Trends in Sociodemographics and Utilization, Including Type, Modality, and Source of Care: Women's Health Evaluation Initiative, Office of Women's Health, Veterans Health Administration, Department of Veterans Affairs; 2024. URL: https://www.womenshealth.va.gov/WOMENSHEALTH/docs/VHA-Sourcebook-V5-FINAL.pdf [accessed 2025-06-20]
3. Frayne SM, Phibbs CS, Saechao F, et al. Sourcebook: Women Veterans in the Veterans Health Administration—Volume 4: Longitudinal Trends in Sociodemographics, Utilization, Health Profile, and Geographic Distribution: Women's Health Evaluation Initiative, Women's Health Services, Veterans Health Administration, Department of Veterans Affairs; 2018. URL: https://www.womenshealth.va.gov/WOMENSHEALTH/docs/Sourcebook_Vol_4.pdf [accessed 2025-06-20]
4. Katon JG, Tartaglione EV JR, Eleazer JR, et al. State of Reproductive Health—Volume II: VA Reproductive Health Diagnoses and Organization of Care: Office of Women's Health, Veterans Health Administration, Department of Veterans Affairs; 2023. URL: https://www.womenshealth.va.gov/WOMENSHEALTH/docs/VHA-WH-Reproductive-Health-Report-2023.pdf [accessed 2025-06-20]
5. Kamarck KN. Women in combat: issues for congress. : Congressional Research Service; 2016 Dec 13 URL: https://sgp.fas.org/crs/natsec/R42075.pdf [accessed 2025-06-20]
6. War Related Illness and Injury Study Center. US Department of Veterans Affairs. URL: https://www.warrelatedillness.va.gov/WARRELATEDILLNESS/index.asp [accessed 2025-06-20]
7. Patten E, Parker K. Women in the U.S. Military: growing share, distinctive profile. Pew Research Center. 2011 Dec 22. URL: https://www.pewresearch.org/social-trends/2011/12/22/women-in-the-u-s-military-growing-share-distinctive-profile [accessed 2025-06-20]
8. DeBeer BB, Davidson D, Meyer EC, Kimbrel NA, Gulliver SB, Morissette SB. The association between toxic exposures and chronic multisymptom illness in veterans of the wars of Iraq and Afghanistan. J Occup Environ Med 2017 Jan;59(1):54-60. [doi: 10.1097/JOM.0000000000000922] [Medline: 28045798]
9. White RF, Steele L, O'Callaghan JP, et al. Recent research on Gulf War illness and other health problems in veterans of the 1991 Gulf War: effects of toxicant exposures during deployment. Cortex 2016 Jan;74:449-475. [doi: 10.1016/j.cortex.2015.08.022] [Medline: 26493934]
10. Taylor MG, Ureña S, Kail BL. Service-related exposures and physical health trajectories among aging veteran men. Gerontologist 2016 Feb;56(1):92-103. [doi: 10.1093/geront/gnv662] [Medline: 26582385]
11. Cherry N, Creed F, Silman A, et al. Health and exposures of United Kingdom Gulf War veterans. Part II: The relation of health to exposure. Occup Environ Med 2001 May;58(5):299-306. [doi: 10.1136/oem.58.5.299] [Medline: 11303078]
12. Jeffrey MG, Krengel M, Kibler JL, et al. Neuropsychological findings in Gulf War illness: a review. Front Psychol 2019;10:2088. [doi: 10.3389/fpsyg.2019.02088] [Medline: 31616335]
13. Vasterling JJ, Proctor SP, Amoroso P, Kane R, Heeren T, White RF. Neuropsychological outcomes of army personnel following deployment to the Iraq War. JAMA 2006 Aug 2;296(5):519-529. [doi: 10.1001/jama.296.5.519] [Medline: 16882958]
14. Toomey R, Alpern R, Vasterling JJ, et al. Neuropsychological functioning of U.S. Gulf War veterans 10 years after the war. J Int Neuropsychol Soc 2009 Sep;15(5):717-729. [doi: 10.1017/S1355617709990294] [Medline: 19640317]
15. King LA, King DW, Bolton EE, Knight JA, Vogt DS. Risk factors for mental, physical, and functional health in Gulf War veterans. J Rehabil Res Dev 2008;45(3):395-407. [doi: 10.1682/jrrd.2007.06.0081] [Medline: 18629748]
16. Schultz D, Hunter KM, Skrabala L, Haynie JG. Improving support for veteran women: veterans' issues in focus. Rand Health Q 2023 May;10(2):10. [Medline: 37200821]
17. Ansa BE, Sullivan K, Krengel MH, et al. The Gulf War women's health cohort: study design and protocol. Int J Environ Res Public Health 2020 Apr 2;17(7):2423. [doi: 10.3390/ijerph17072423] [Medline: 32252385]
18. Coughlin SS, Krengel M, Sullivan K, Pierce PF, Heboyan V, Wilson LCC. A review of epidemiologic studies of the health of Gulf War women veterans. J Environ Health Sci 2017;3(2):1-9. [doi: 10.15436/2378-6841.17.1551] [Medline: 29034324]
19. Dursa EK, Barth SK, Porter BW, Schneiderman AI. Health status of female and male Gulf War and Gulf Era veterans: a population-based study. Womens Health Issues 2019 Jun 25;29(Suppl 1):S39-S46. [doi: 10.1016/j.whi.2019.04.003] [Medline: 31253241]

XSL•FO
RenderX

20. Street AE, Gradus JL, Giasson HL, Vogt D, Resick PA. Gender differences among veterans deployed in support of the wars in Afghanistan and Iraq. J Gen Intern Med 2013 Jul;28(Suppl 2):S556-S562. [doi: 10.1007/s11606-013-2333-4] [Medline: 23807065]
21. Street AE, Vogt D, Dutra L. A new generation of women veterans: stressors faced by women deployed to Iraq and Afghanistan. Clin Psychol Rev 2009 Dec;29(8):685-694. [doi: 10.1016/j.cpr.2009.08.007] [Medline: 19766368]
22. Katon JG, Zephyrin L, Meoli A, et al. Reproductive health of women veterans: a systematic review of the literature from 2008 to 2017. Semin Reprod Med 2018 Nov;36(6):315-322. [doi: 10.1055/s-0039-1678750] [Medline: 31003246]
23. Danan ER, Krebs EE, Ensrud K, et al. An evidence map of the women veterans' health research literature (2008-2015). J Gen Intern Med 2017 Dec;32(12):1359-1376. [doi: 10.1007/s11606-017-4152-5] [Medline: 28913683]
24. Bean-Mayberry B, Yano EM, Washington DL, et al. Systematic review of women veterans' health: update on successes and gaps. Womens Health Issues 2011;21(4):S84-S97. [doi: 10.1016/j.whi.2011.04.022] [Medline: 21724149]
25. Goldzweig CL, Balekian TM, Rolón C, Yano EM, Shekelle PG. The state of women veterans' health research: results of a systematic literature review. J Gen Intern Med 2006 Mar;21(Suppl 3):S82-S92. [doi: 10.1111/j.1525-1497.2006.00380.x] [Medline: 16637952]
26. Lange G, McAndrew L, Ashford JW, Reinhard M, Peterson M, Helmer DA. War Related Illness and Injury Study Center (WRIISC): a multidisciplinary translational approach to the care of veterans with chronic multisymptom illness. Mil Med 2013 Jul;178(7):705-707. [doi: 10.7205/MILMED-D-13-00053] [Medline: 23820341]
27. Lincoln AE, Helmer DA, Schneiderman AI, et al. The war-related illness and injury study centers: a resource for deployment-related health concerns. Mil Med 2006 Jul;171(7):577-585. [doi: 10.7205/milmed.171.7.577] [Medline: 16895119]
28. Assefa MT, Adamson MM, Proctor LB, Moeder S, Ashford JW, Jennings JS. A multipronged approach to caring for women veterans with military environmental exposures. Womens Health Issues 2024;34(4):325-330. [doi: 10.1016/j.whi.2024.05.002] [Medline: 38890077]
29. Women's Operational Military Exposure Network Center of Excellence. US Department of Veterans Affairs. URL: https://www.warrelatedillness.va.gov/WARRELATEDILLNESS/WRIISC-WOMEN/ww-about-us.asp [accessed 2025-06-20]
30. Katon JG, Hoggatt KJ, Balasubramanian V, et al. Reproductive health diagnoses of women veterans using Department of Veterans Affairs health care. Med Care 2015 Apr;53(4 Suppl 1):S63-S67. [doi: 10.1097/MLR.0000000000000295] [Medline: 25767978]
31. Miller LJ, Ghadiali NY. Gender-specific mental health care needs of women veterans treated for psychiatric disorders in a Veterans Administration women's health clinic. Med Care 2015 Apr;53(4 Suppl 1):S93-S96. [doi: 10.1097/MLR.0000000000000282] [Medline: 25767984]
32. Eysenbach G. Improving the quality of Web surveys: the Checklist for Reporting Results of Internet e-Surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34. [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
33. Abou-Donia MB, Krengel MH, Lapadula ES, et al. Sex-based differences in plasma autoantibodies to central nervous system proteins in Gulf War veterans versus healthy and symptomatic controls. Brain Sci 2021 Jan 23;11(2):148. [doi: 10.3390/brainsci11020148] [Medline: 33498629]
34. Heboyan V, Krengel MH, Sullivan K, et al. Sex differences in Gulf War illness: a reanalysis of data from the CDC Air Force study using CDC and modified Kansas case definitions. J Occup Environ Med 2019 Jul;61(7):610-616. [doi: 10.1097/JOM.0000000000001620] [Medline: 31090678]
35. Krengel M, Sullivan K, Heboyan V, et al. Neurotoxicant exposures and rates of chronic multisymptom illness and Kansas Gulf War illness criteria in Gulf War deployed women veterans. Life Sci 2021 Sep 1;280:119623. [doi: 10.1016/j.lfs.2021.119623] [Medline: 34004246]
36. Sullivan K, Krengel M, Heboyan V, et al. Prevalence and patterns of symptoms among female veterans of the 1991 Gulf War era: 25 years later. J Womens Health (Larchmt) 2020 Jun;29(6):819-826. [doi: 10.1089/jwh.2019.7705] [Medline: 32250195]
37. Gaffey AE, Burg MM, Rosman L, et al. Baseline characteristics from the women veterans cohort study: gender differences and similarities in health and healthcare utilization. J Womens Health (Larchmt) 2021 Jul;30(7):944-955. [doi: 10.1089/jwh.2020.8732] [Medline: 33439756]
38. Haskell SG, Gordon KS, Mattocks K, et al. Gender differences in rates of depression, PTSD, pain, obesity, and military sexual trauma among Connecticut War veterans of Iraq and Afghanistan. J Womens Health (Larchmt) 2010 Feb;19(2):267-271. [doi: 10.1089/jwh.2008.1262] [Medline: 20109115]
39. Mohanty AF, McAndrew LM, Helmer D, Samore MH, Gundlapalli AV. Chronic multisymptom illness among Iraq/Afghanistan-deployed US veterans and their healthcare utilization within the Veterans Health Administration. J Gen Intern Med 2018 Sep;33(9):1419-1422. [doi: 10.1007/s11606-018-4479-6] [Medline: 29797218]
40. Vogt D, Vaughn R, Glickman ME, et al. Gender differences in combat-related stressors and their association with postdeployment mental health in a nationally representative sample of U.S. OEF/OIF veterans. J Abnorm Psychol 2011 Nov;120(4):797-806. [doi: 10.1037/a0023452] [Medline: 21639595]

XSL•FO
RenderX

41. Cohen BE, Maguen S, Bertenthal D, Shi Y, Jacoby V, Seal KH. Reproductive and other health outcomes in Iraq and Afghanistan women veterans using VA health care: association with mental health diagnoses. Womens Health Issues 2012 Sep;22(5):e461-e471. [doi: 10.1016/j.whi.2012.06.005] [Medline: 22944901]
42. Dursa EK, Cao G, Culpepper WJ, Schneiderman A. Comparison of health outcomes over time among women 1990-1991 Gulf War veterans, women 1990-1991 Gulf era veterans, and women in the U.S. general population. Womens Health Issues 2023;33(6):643-651. [doi: 10.1016/j.whi.2023.06.006] [Medline: 37495424]
43. Gawron LM, Mohanty AF, Kaiser JE, Gundlapalli AV. Impact of deployment on reproductive health in U.S. active-duty servicewomen and veterans. Semin Reprod Med 2018 Nov;36(6):361-370. [doi: 10.1055/s-0039-1678749] [Medline: 31003251]
44. Araneta MRG, Kamens DR, Zau AC, et al. Conception and pregnancy during the Persian Gulf War: the risk to women veterans. Ann Epidemiol 2004 Feb;14(2):109-116. [doi: 10.1016/j.annepidem.2003.08.006] [Medline: 15018883]
45. Friedman A, Janulewicz Lloyd PA, Carlson J, et al. Preliminary findings from the Gulf War women's cohort: reproductive and children's health outcomes among women veterans. Int J Environ Res Public Health 2022 Jul 11;19(14):14. [doi: 10.3390/ijerph19148483] [Medline: 35886335]
46. De La Cruz MSD, Buchanan EM. Uterine fibroids: diagnosis and treatment. Am Fam Physician 2017 Jan 15;95(2):100-107. [Medline: 28084714]
47. Haws K, Mak S, Greer S, et al. A virtual functional medicine-based interdisciplinary and integrative intervention for Gulf War illness. Mil Med 2024 Aug 19;189(Suppl 3):99-105. [doi: 10.1093/milmed/usae054] [Medline: 39160805]

Abbreviations

CA: California
CHERRIES: Checklist for Reporting Results of Internet e-Surveys
DC: Washington, District of Columbia
MEE: military environmental exposure
NJ: New Jersey
VA: Department of Veteran Affairs
WHA: Women's Health Addendum
WOMEN CoE: Women's Operational Military Exposure Network Center of Excellence
WRIISC: War Related Illness and Injury Study Center

Edited by A Mavragani; submitted 27.02.25; peer-reviewed by C Chikwe, L Bastian; revised version received 28.04.25; accepted 04.05.25; published 24.07.25.

Please cite as:
Eizadi LN, Assefa MT, Nechvatal JM, Veltkamp GM, Ibrahim-Biangoro A, Adamson MM, Jennings JS
A Supplemental Women's Health Questionnaire for Women Veterans With Military Environmental Exposures: Project Development and Implementation
JMIR Form Res 2025;9:e73223
URL: https://formative.jmir.org/2025/1/e73223
doi:10.2196/73223

© Leah N Eizadi, Mehret T Assefa, Jordan M Nechvatal, G Marina Veltkamp, Abou Ibrahim-Biangoro, Maheen M Adamson, Jennifer S Jennings. Originally published in JMIR Formative Research (https://formative.jmir.org), 24.7.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Construction of Personalized Predictive Models for Missed Medication Doses Using Wearable Device Data: Prospective Observational Study

Haru Iino, BSc; Hayato Kizaki, PhD; Shungo Imai, PhD; Satoko Hori, PhD

Division of Drug Informatics, Faculty of Pharmacy and Graduate School of Pharmaceutical Sciences, Keio University, 1-5-30 Shibakoen, Minato-ku, Tokyo, Japan

Corresponding Author:
Satoko Hori, PhD
Division of Drug Informatics, Faculty of Pharmacy and Graduate School of Pharmaceutical Sciences, Keio University, 1-5-30 Shibakoen, Minato-ku, Tokyo, Japan

Abstract

Background: Declining medication adherence remains a critical health care issue, often assessed through unreliable self-reporting methods. Wearable devices (WDs) may offer an objective means to improve adherence monitoring by continuously recording physiological and activity data.

Objective: This study aimed to develop and internally validate personalized predictive models, utilizing objective physiological and activity data from WDs, for identifying missed medication doses.

Methods: A 30-day prospective observational study was conducted with 8 participants who wore Apple Watches and used a dedicated iOS app. The app collected demographics, medication details, psychological factors, mealtimes, and daily missed dose events. WDs recorded time-series data (ie, activity, heart rate, sleep) at 3-minute intervals. Data were aggregated into 1-hour segments, and lag (6 and 12 h) as well as rolling (24 h) features were generated. Light Gradient Boosting Machine models were constructed for each individual's dosing regimen if the missed dose rate exceeded 20%. Two modeling approaches were compared: a group cross-validation (CV) model that grouped data by day to avoid data leakage from rolling features, and a nonrolling feature model that excluded rolling features and used leave-one-out CV. F_1-score, accuracy, recall, and precision were assessed between the 2 models.

Results: Of the 15 enrolled participants, 8 completed the study; 4 had a missed dose rate above 20%. In these 4 individuals, the group CV model achieved F_1-scores of 0.435 to 0.902, with accuracy ranging from 0.711 to 0.911, recall from 0.278 to 0.822, and a precision of 1.000 for the most robust regimens. The nonrolling feature model yielded F_1-scores of 0.667 to 0.910, with accuracy ranging from 0.800 to 0.906, recall from 0.500 to 0.835, and a precision of 1.000. Morning dosing regimens generally showed higher predictive performance than evening or afternoon. Time-series features, particularly those reflecting 6-, 12-, and 24-hour patterns, emerged as key predictors, indicating that physiological and lifestyle variations prior to dosing strongly influenced missed dose events.

Conclusions: Personalized predictive models using WD-derived data demonstrated high precision for detecting missed medication doses, especially in morning and evening regimens. These findings underscore the feasibility of employing continuous, objective physiological and activity data from WDs to forecast nonadherence events. Although the sample size was limited, restricting the generalizability of the results, this study demonstrates the potential of WD-based personalized prediction of medication adherence. Future work should involve larger populations for external validation, strategies to improve recall, especially for clinically critical medications, and careful consideration of real-world implementation challenges.

(JMIR Form Res 2025;9:e72113) doi:10.2196/72113

KEYWORDS

personalized predictive models; missed medication doses; wearable device data; Light Gradient Boosting Machine; LightGBM

Introduction

Medication adherence is a crucial factor in pharmaceutical care [1]. A 2020 study revealed that 32% of patients did not take their medications correctly, and those who did not adhere properly incurred an additional average monthly cost of US $97.98 [2]. This indicates that medication nonadherence imposes both health and economic burdens. Given these significant impacts, identifying effective strategies to improve medication adherence has become a critical health care priority. To prevent a decline in medication adherence, it is necessary to accurately assess adherence at an individual level [3,4]. Typically,

XSL•FO
RenderX

medication adherence is evaluated by comparing the frequency of clinic visits with prescription intervals or by simple questioning [5-7]. However, this approach faces challenges such as time constraints and the reliance on self-reporting [5], which is often subject to recall bias and social desirability bias, limiting its reliability for accurate monitoring [8].

Various factors influence medication adherence and are closely related to the patient's lifestyle [9-12]. We previously conducted a study on the relative importance of factors related to medication adherence and suggested that the consistency of lifestyle habits may be the most important factor. Some questionnaires that have been used to evaluate medication adherence include questions about the patient's lifestyle [5,13-16], as inquiring about an individual's lifestyle is a common first step toward improving adherence [12]. Lifestyle inquiries often relate to the number of meals consumed and include subjective lifestyle evaluations. However, it is currently difficult to quantitatively evaluate comprehensive aspects of lifestyle beyond meal frequency using objective indicators [17-19].

Therefore, we focused on wearable devices (WDs) to continuously monitor a wearer's physiological and activity data. WDs accumulate daily information, such as the wearer's activity level, heart rate, and sleep duration, allowing the collection of objective data [20-23], which may reflect a patient's lifestyle [24]. Thus, by constructing machine learning models from these data, it may be possible to determine predictive factors related to a patient's medication adherence. We previously conducted a scoping review on medication adherence management using WDs, which revealed that, while reminders and alerts have been widely used in medication management, some research methods use WD data as an advanced approach to medication adherence management [25]. Leveraging WD data, which have not been widely used to date, and obtaining insights into medication adherence from unconventional objective information may result in new medication management approaches that place less burden on both patients and health care providers. Recent studies have demonstrated that data-driven approaches using deep learning and ensemble methods significantly enhance prediction accuracy of medication adherence compared to traditional self-reporting methods [26]. Additionally, WDs have shown promise in improving clinical decision-making through real-time, noninvasive physiological data collection, particularly in the management of chronic diseases such as heart failure [27]. However, successful long-term monitoring using WDs in clinical practice requires consideration of usability and acceptability, especially among older adults, as well as strategies to maintain continuous user engagement and address technical challenges [28]. Moreover, while machine learning approaches for adherence prediction are advancing, studies also indicate that many existing predictive models have a high risk of bias, highlighting the need for enhanced methodological quality [29].

Therefore, addressing the gaps and the need for objective, personalized adherence monitoring, the primary aim of this study was to develop and internally validate personalized predictive models for identifying missed medication doses, utilizing objective physiological and activity data prospectively collected from commonly used WDs.

Methods

Development of Data Collection Platform and Patient Recruitment

Participants were recruited between May 2023 and December 2023. All study recruitment and procedures were conducted in Japan. Information about the study was disseminated through flyers posted in local libraries and via a dedicated project page hosted on our laboratory's website, inviting interested individuals to voluntarily download the dedicated study app. Eligibility criteria for participation, which were confirmed via an initial in-app questionnaire after its download and prior to obtaining informed consent, were as follows: (1) aged 18 years or older; (2) taking at least 1 medication prescribed by a hospital with instructions for daily use for an intended duration of at least 3 months (individuals whose only medications were for self-adjustment, eg, pro re nata or "as needed" only, were excluded); (3) routinely using an Apple Watch (defined as wearing it for an estimated minimum of 8 hours per day, at least 5 days a week) and an iPhone; and (4) able to download the dedicated study app from the Japanese App Store onto their iPhone and operate the app. The app was developed specifically for iOS smartphones and made available exclusively on the Japanese App Store (Figure 1). It included a questionnaire and health data–sharing capabilities, with collected responses and health data stored on a cloud server.

The questionnaires collected information on age, usage, types of medications taken, psychological factors related to adherence (10 questions on a 5-point scale), nonworking days of the week, meal start times and durations, and the presence or absence of missed doses over 30 days. The WD data collected included exercise time, standing time, active energy, resting energy expenditure, walking and running distances, step count, number of flights climbed, maximum oxygen uptake, mindfulness time, vital signs (ie, heart rate, heart rate variability, resting heart rate), oxygen saturation levels, and sleep time. Specifically, health data, excluding those related to the questionnaires and sleep (such as wake-up and bedtime), were obtained as time-series data divided into 3-minute intervals. Missing values arising from nonwear periods or lack of body movement were imputed with zeros. This simple imputation approach was chosen for its feasibility in this initial exploratory study. While zero values might reasonably represent periods of inactivity for some metrics like step counts, we acknowledge this method is suboptimal for physiological measures (eg, heart rate) and that the assumption that missingness equates to zero may introduce bias if nonwear periods systematically correlate with adherence behaviors. Alternative imputation strategies (eg, mean and median imputation, interpolation, forward and backward fill) exist, and investigating their impact on model performance represents an area for future refinement. The subsequent aggregation of the 3-minute data into 1-hour descriptive statistics may, however, partially mitigate the direct influence of the initial imputation choice on the final model features.

The participants downloaded the app, reviewed the research explanation materials and consent form within the app, and joined the study after providing informed consent. After

XSL•FO
RenderX

initiating the survey, an initial questionnaire screen was displayed to collect information on age, medication usage, and medication types. Subsequently, the app transitioned to a regular survey screen, where the participants recorded their medication status up to the previous day at their convenience.

Figure 1. Screenshots of the dedicated study app, including (A) the launch screen, (B) initial questionnaire screen, (C) consent form and agreement screen, (D) research explanation document, and (E) regular survey screen (input on medication status).

Preprocessing Before Model Construction and Verification of Time-Series Components

To evaluate the quality of the collected time-series data from WDs, the completeness of each data type was visualized. Insufficiently recorded data were excluded from further analyses (Multimedia Appendix 1). The periodicity of the included time-series components was examined using additive decomposition, confirming the absence of trend components and the presence of seasonal variations on a daily basis (Multimedia Appendix 1). Spectral analysis with fast Fourier transformation was performed to verify the lengths of the periods, confirming that many data types exhibited strong spectral peaks at 24- or 12-hour intervals (Multimedia Appendix 1).

To construct predictive models, it is desirable to ensure a sufficient number of event occurrences and to avoid extreme imbalances in the distribution of response variables. Therefore, the missed dose events were aggregated. Using the number of events during the observation period, the event occurrence rate was calculated using the following equation: missed dose event occurrence rate=(number of missed dose events)/(number of doses that should have been taken during the period).

In this study, medication adherence refers to the value calculated as above.

Confirmation of Response Variables and Creation of Time-Series Features

All data were divided into 1-hour intervals. For each time segment, descriptive statistics, including the mean, variance, maximum, minimum, median, and sum, were calculated. Assuming that the state prior to each time point is related to the event, time-series–related features, such as lag features for the previous 6 and 12 hours, and rolling features for the previous 24 hours, were added. The data were classified by individual and dosing regimens (morning, afternoon, and evening). For the portions where a sufficient number of events (>20%) were secured, models to detect missed medication events from the WD data were created.

In terms of the dosing regimen, specific timeslots where events could potentially occur were allocated: 6 AM to 10:59 AM for morning doses, 11 AM to 2:59 PM for afternoon doses, and 6 PM to 11:59 PM for evening doses. For example, data from 6 PM to 11:59 PM were extracted when constructing a model to predict evening events, resulting in 6 datasets divided into 1-hour intervals per day. If an event occurred on that day, it was labeled as 1; otherwise, it was labeled as 0 as the response variable. If the data were recorded for 30 days, 180 datasets (6×30) were used for model construction.

Construction of Predictive Models

Predictive models were constructed using Light Gradient Boosting Machine (LightGBM). LightGBM was selected due to its demonstrated high predictive performance on tabular data tasks similar to ours, its computational efficiency in terms of training speed and memory usage compared to other gradient boosting algorithms, and its built-in capability to provide feature importance rankings, which was utilized in our analysis. Given that the data were divided into 1-hour intervals, the newly created rolling features were generated based on the past 24 hours. In other words, for adjacent time points, only 1 hour of data differed while the remaining 23 hours of data were the same, potentially resulting in minimal changes. Therefore, performing random cross-validation (CV) can lead to other rolling features from the same day being learned and used as test data during prediction, which may not be appropriate.

To prevent this, the following two types of models were constructed [30]: (1) a model in which data from the same day were grouped together and group CV was performed (group CV model) and (2) a model in which rolling features were excluded from the variables and leave-one-out CV was performed (nonrolling feature model). The models were constructed to optimize the F_1-scores. Evaluation metrics, such as the F_1-score, accuracy, recall, and precision, were calculated

XSL•FO
RenderX

[31]. In addition, the feature importance of each model was computed for each fold in the CV, and the average for each model was provided. All analyses were performed using Python 3.10.12, and the code used in prediction models is present in Multimedia Appendix 2.

Ethical Considerations

This study complied with the Ethical Guidelines for Medical and Biological Research Involving Human Subjects published by the Japanese Government (Ministry of Health, Labour and Welfare; Ministry of Education, Culture, Sports, Science and Technology; and Ministry of Economy, Trade and Industry). All research plans were reviewed and approved by the Research Ethics Committee of Keio University Faculty of Pharmacy (23414 - 1). We completed a checklist for reporting studies involving machine learning, following recommended guidelines for such research (Checklist 1) [32].

Informed consent was obtained electronically from all participants via the dedicated study app before they commenced any study procedures. The consent form provided detailed information regarding the study's objectives, procedures (including the collection of questionnaire data and continuous physiological and activity data from their Apple Watch), potential risks and benefits, data handling, and the voluntary nature of participation. Participation was entirely voluntary, and participants were clearly informed of their right to withdraw from the study at any time, for any reason, without prejudice or penalty, fully ensuring patient autonomy. To protect privacy and confidentiality, all data collected through the app and wearable device were anonymized at the point of collection before being transmitted and stored. Anonymized data were managed on a secure cloud server, employing appropriate technical and administrative safeguards to prevent unauthorized access, disclosure, or loss. Although this study was conducted exclusively in Japan using a locally distributed app, and thus not directly subject to General Data Protection Regulation or Health Insurance Portability and Accountability Act regulations, data management practices were designed to align with principles of robust data protection and security suitable for sensitive health-related information. The participants who completed the 30-day survey were compensated with Japanese ¥3000 (US $20.84).

Results

Aggregation of Event Occurrence Rates

Fifteen individuals participated in the survey, and 8 completed it. A summary of the participants' data is shown in Table 1. Among them, 4 (50%) were aged 20-29 years, 2 (25%) were aged 40-49 years, 1 (12.5%) was aged 50-59 years, and 1 (12.5%) did not provide an answer regarding age. The collected data were divided into individual and dosing regimens, resulting in 15 arms. Among these arms, there were 4 in which the event occurrence rate was >20% (ID1: morning, evening; ID2: evening; ID3: afternoon). These event occurrence rates ranged from 41.4% to 56.7%, without significant skewness, and were used as response variables for predictive model construction. For individuals ID4 and ID5, although a few events were observed in the evening, none were observed in the morning. For individuals ID6, ID7, and ID8, no events were observed. The event occurrence rate per number of days, without distinguishing the dosing regimens for each individual, are shown in Table S1 in Multimedia Appendix 3, and the event occurrence rates per total number of doses are provided in Table S2 in Multimedia Appendix 3.

XSL•FO
RenderX

Table . Summary of the number of events and medication adherence rates for 8 adult Apple Watch users in Japan from a 30-day prospective observational study monitoring daily medication use.

ID (age group) and administration		Prescribed medications, n	Expected doses, n	Events occurring, n	Missed dose event occurrence rate, %
ID1 (20-29 years)					
	Morning (after a meal)	1	30	17	56.7
	Evening (after a meal or before bed)	3	30	15	50
ID2 (20-29 years)					
	Evening (after a meal)	1	30	17	56.7
ID3 (20-29 years)					
	Morning (before and after a meal)	6	29	2	6.9
	Afternoon (after a meal)	1	29	12	41.4
	Evening (after a meal)	9	29	1	3.4
ID4 (20-29 years)					
	Morning (after a meal)	5	30	0	0
	Evening (before bed)	1	30	4	13.3
ID5 (40-49 years)					
	Morning (after a meal)	1	29	0	0
	Evening (after a meal)	2	29	1	3.4
ID6 (50-59 years)					
	Morning (after a meal)	4	30	0	0
	Evening (after a meal)	3	30	0	0
ID7 (no answer)					
	Morning (after a meal)	4	30	0	0
	Evening (after a meal or before bed)	4	30	0	0
ID8 (50-59 years)					
	Morning (after a meal)	2	30	0	0

Predictive Models

The results of the group CV model are presented in Table 2, and those of the nonrolling feature model are presented in Table 3. Internal validation showed that the F_1-score ranged from 0.435 to 0.902, accuracy from 0.711 to 0.911, and recall from 0.278 to 0.822. In the nonrolling feature model, the F_1-score ranged from 0.667 to 0.910, accuracy from 0.800 to 0.906, and recall from 0.500 to 0.835. Precision was 1.0 for both models.

XSL•FO
RenderX

Table . Performance of Light Gradient Boosting Machine group cross-validation models in predicting missed medication doses for 4 participants (adult Apple Watch users, Japan) from a 30-day prospective observational study.

		F_1-score	Accuracy	Recall	Precision
ID1					
	Morning (after a meal)	0.902	0.911	0.822	1.000
	Evening (after a meal or before bed)	0.784	0.807	0.659	1.000
ID2					
	Evening (after a meal)	0.793	0.806	0.657	1.000
ID3					
	Afternoon (after a meal)	0.435	0.711	0.278	1.000

Table . Performance of Light Gradient Boosting Machine nonrolling feature models (leave-one-out cross-validation) in predicting missed medication doses for 4 participants (adult Apple Watch users, Japan) from a 30-day prospective observational study.

		F_1-score	Accuracy	Recall	Precision
ID1					
	Morning (after a meal)	0.910	0.906	0.835	1.000
	Evening (after a meal or before bed)	0.816	0.844	0.689	1.000
ID2					
	Evening (after a meal)	0.867	0.867	0.765	1.000
ID3					
	Afternoon (after a meal)	0.667	0.800	0.500	1.000

Feature Importance

The importance of the top 50 features in the group CV model for ID1 is shown in Figure 2, and the importance of the top 50 features in the nonrolling feature model is shown in Figure 3. The feature importance values of the other models are present in Multimedia Appendix 4 .

XSL•FO
RenderX

Figure 2. Top 50 feature importance in the group cross-validation model for ID1: morning (after a meal). SDNN: standard deviation of all normal to normal R-R (NN) intervals; [Lag6]: 6-hour lag features; [Lag12]: 12-hour lag features; [Rolling]: rolling feature.

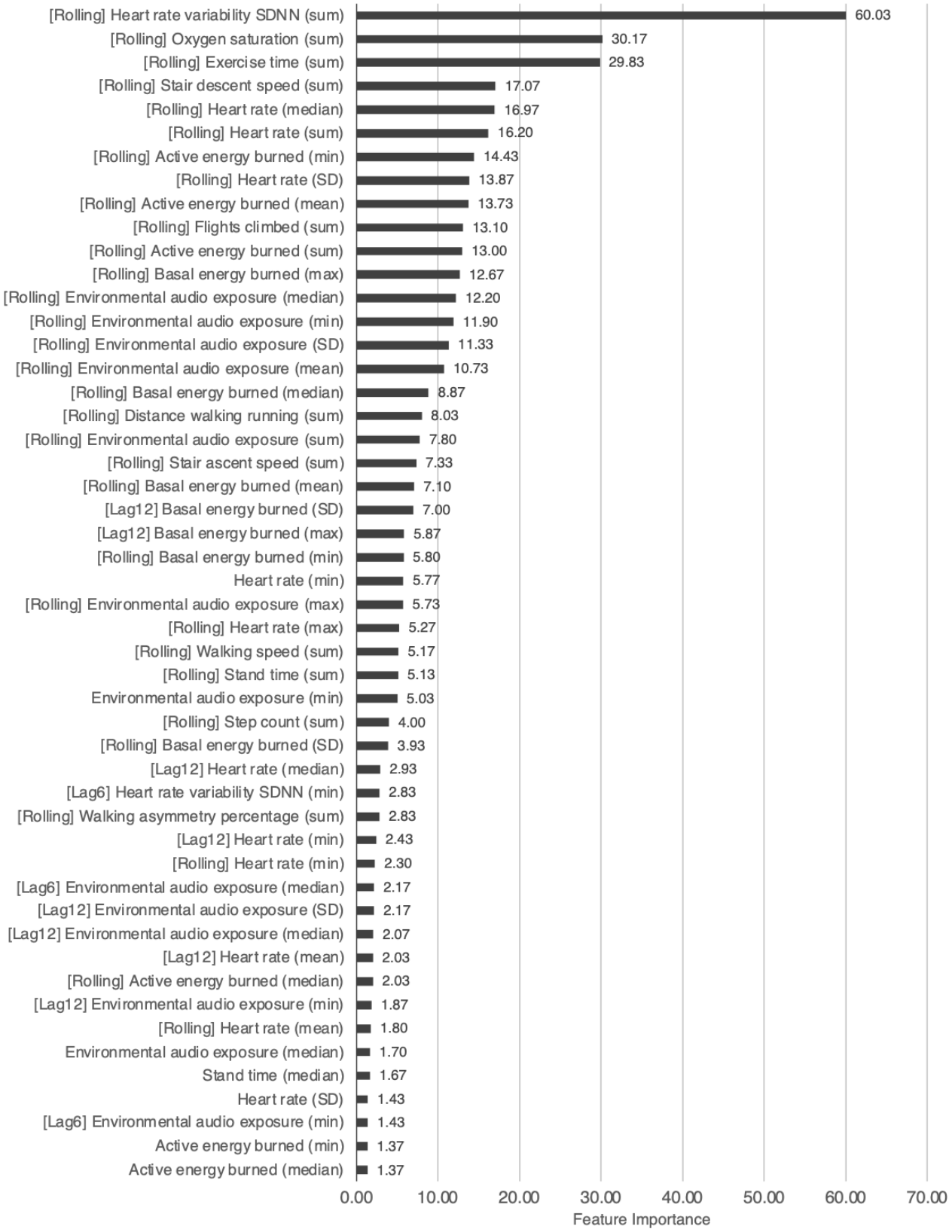

XSL•FO
RenderX

Figure 3. Top 50 feature importance in the nonrolling feature model for ID1: morning (after a meal). SDNN: standard deviation of all normal to normal R-R (NN) intervals; [Lag6]: 6-hour lag features; [Lag12]: 12-hour lag features; [Rolling]: rolling feature.

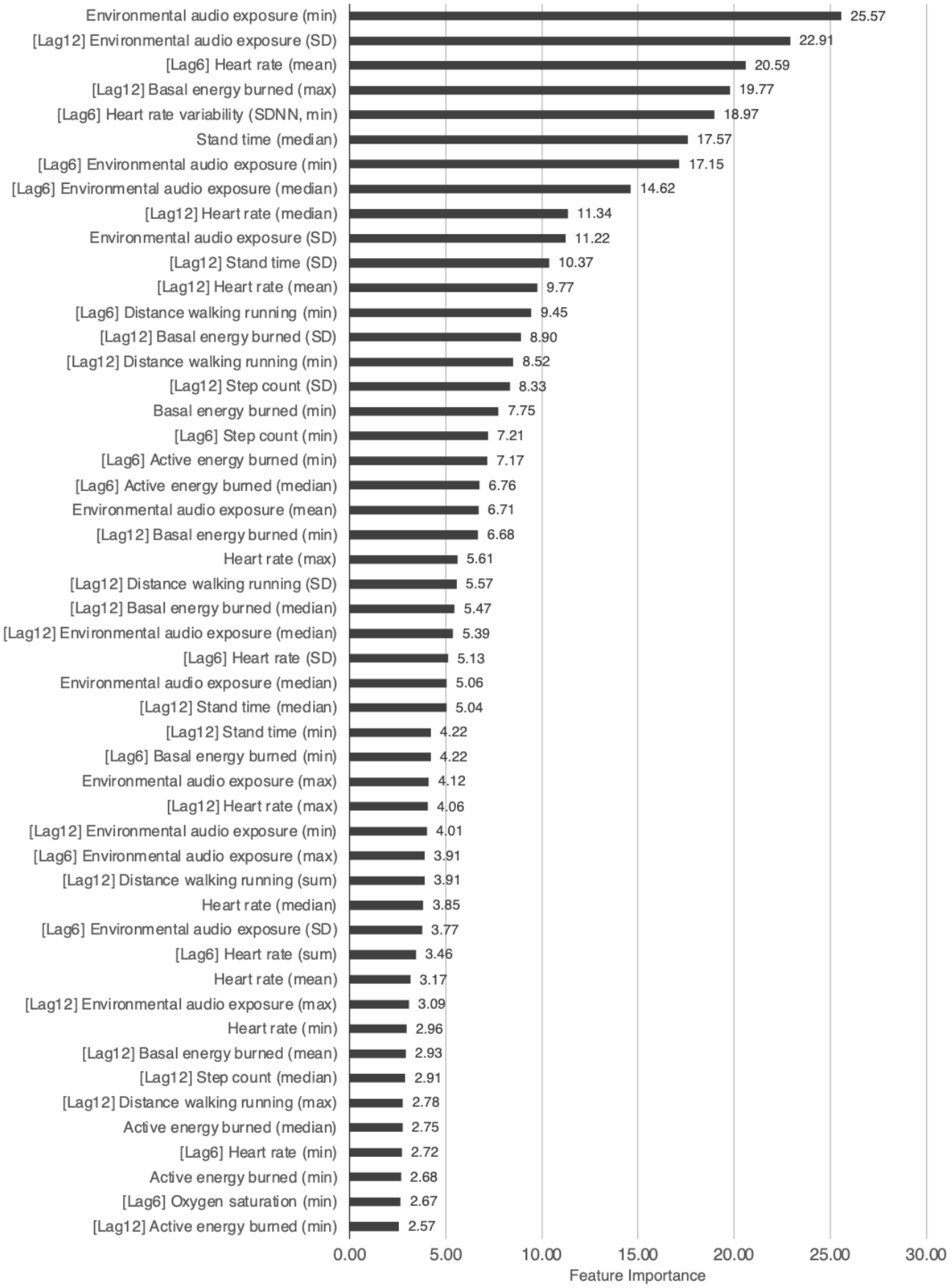

XSL•FO
RenderX

Discussion

Findings

In this prospective observational study, we collected WD data and developed 2 models using missed medication events as response variables for each individual and dosing regimen. The internal validation demonstrated good overall performance, except for the afternoon dosing model. For morning and evening dosing regimens, the maximum F_1-score, accuracy, and recall were 0.902, 0.911, and 0.822, respectively, with precision consistently reaching 1.0, indicating a high discriminative performance. The performance ranking by regimen was morning > evening > afternoon, and the 2 types of models showed generally equivalent results. In terms of feature importance, time-series features consistently ranked as highly important across models.

A key contribution of this study was the development of an integrated system for collecting medication information and WD data to construct predictive models. We herein performed app development, database construction, and implementation of a prospective data collection protocol involving study participants. Despite the limited sample size, this is the first study to build a predictive model for missed medication doses using activity, heart rate, and other physiological data obtained from commonly used WDs, targeting the general patient population. This methodological approach, focusing on personalized prediction using objective, continuously monitored data streams, represents the core novelty of the study. By addressing the challenges of real-world data collection and focusing on commonly used devices, this study could serve as a novel foundation for advancing medication adherence monitoring.

High Precision and Low Recall in the Models

Both the group CV and nonrolling feature models achieved high precision, representing their ability to accurately identify missed medication events with minimal false positives. However, the recall was relatively low, indicating that some missed events were not detected. This limitation was particularly evident during the daytime and evening regimens, likely due to the large variations in medication patterns and lifestyle habits during these periods, which the models have not sufficiently learned. Additionally, certain missed events may not generate measurable changes in WD data, such as forgetting to take medications without altering measurable activity levels or physiological parameters.

In this study, the models were optimized for F_1-scores to balance precision and recall. Consequently, the precision was high, and recall was low. In situations where adherence is particularly critical, such as with immunosuppressant or anticoagulant medications, it is necessary to detect missed doses more accurately. When targeting such medications, it may be necessary to construct models emphasizing recall, such as cost-sensitive learning [33].

Factors Influencing Afternoon Model Performance

Two primary factors might explain the poorer performance of the afternoon models. First, unique influences on afternoon medication adherence compared with those on morning or evening medication adherence. During the day, activities, such as going out, work, and school, and factors such as inaccessibility to medication or the impact of work and eating out can influence adherence [34-36]. When patients do not have their medication on their person, despite no significant lifestyle changes, WDs may not sufficiently capture such situations, potentially reducing the predictive model's performance. Second, the shorter time window allocated to the afternoon (4 h) compared to the morning (5 h) and evening (6 h) may have limited data volume, possibly preventing the model from utilizing sufficient data for learning. However, the evening model, despite possessing more data, did not outperform the morning model. This could be because there is generally more variation in activities during the evening than during the morning [37,38], making evening predictions more challenging.

Feature Importance in the 2 Models

Feature importance analysis highlighted the significance of time-series features. In the group CV model, the rolling features ranked high, suggesting that data obtained within the previous 24 hours significantly influenced the prediction of missed medications [39]. Similarly, in the nonrolling feature model, the lag features (from 6 and 12 hours prior) were among the most important. The fact that these 2 time-series features rank higher than other features at the time of the event implies that physiological and lifestyle changes preceding missed doses are greater predictors of a missed dose compared with the immediate conditions at the time of the event.

Although both models demonstrated similar overall performance, the nonrolling feature model showed slight superiority over the group CV model. The nonrolling feature model utilized 270 input variables, which is a relatively large number compared to the 180 target variables. The group CV model had 360 variables, 90 more than the nonrolling feature model, and the inclusion of these additional variables may have contributed to a decrease in the model's performance [40].

Broader Applicability and Implementation Considerations

This feasibility study employed a single platform (Apple Watch and iOS) to ensure consistency in data quality and simplify implementation. However, differences in sensors, algorithms, and available data across platforms (eg, Android and Wear OS) may lead to device-specific biases, limiting the generalizability of the findings. Future research should evaluate model performance across diverse wearable ecosystems to confirm broader applicability and ensure robustness against platform variability.

Furthermore, translating this predictive approach to clinical practice poses real-world implementation challenges. In addition to ensuring consistent user engagement and managing data gaps, practical use of consumer-grade devices must contend with technical issues such as battery life management, device malfunctions, and reliable data synchronization—all of which

XSL•FO
RenderX

may affect data integrity and participant adherence. Moreover, effective integration of high-volume WD data into clinical workflows requires attention to infrastructure, data privacy, and usability concerns for both patients and clinicians. Addressing these barriers will be essential for realizing the potential of wearable-based adherence monitoring in routine care.

In addition, while psychological factors related to medication adherence were collected at baseline via a questionnaire, they were not incorporated into the current models, which focused on dynamic, time-series signals from WDs for within-individual prediction. However, these baseline psychological traits may provide complementary value in future model iterations—particularly for interindividual comparisons or hybrid models that combine static behavioral characteristics with continuous physiological data.

Limitations

There are some limitations that need to be addressed. First, the small sample size restricted model development to within-individual analysis. Only 8 participants completed the study, and models could only be constructed for 4 participants who had a missed dose rate above 20%. This extremely small sample size limits the statistical power and generalizability of the findings. Furthermore, the demographics of this small cohort were not well-balanced, particularly concerning age distribution (50% were in their twenties, with no participants in their thirties). This skew towards younger adults, potentially influenced by the requirement of owning and using an Apple Watch and the recruitment methods employed, further limits the representativeness of our sample and the applicability of the findings to older or more diverse patient populations typically on chronic medications. Recruitment targeted patients regularly wearing smartwatches and taking medications for >3 months, as per participant eligibility. Moreover, only a subset of participants provided sufficient data for model construction. Expanding the number of participants and ensuring a more representative demographic distribution is necessary for generalization. Future studies should focus on recruiting a larger and more diverse population to confirm these preliminary findings and assess the broader applicability of the proposed approach; achieving this necessary scale and diversity may require specific strategies such as multicenter recruitment or potentially leveraging integration with larger existing wearable datasets where feasible and appropriate.

Second, the training data were limited to a 1-month period, yielding >10,000 data points per variable as time-series data. While sufficient for data collected from WDs, the variables measurable by WDs are inherently limited. Factors such as forgetting to carry medications or external disruptions (eg, work or dining out) were not captured, limiting model applicability. Although it is possible to consider methods to measure variables other than those from WD, this often involves a trade-off with patient burden and may hinder practicality [27,28]. Furthermore, the choice of zero-imputation for missing raw wearable data, while straightforward, warrants further investigation regarding potential biases, and the exploration of more sophisticated imputation techniques could be beneficial for future model development.

Third, this study did not include a control group or a direct comparison with traditional adherence monitoring methods, such as self-reporting or pill counts. The study was designed primarily as a feasibility and proof-of-concept investigation to determine if personalized predictive models could be constructed using WD data. Therefore, while our results demonstrate the potential of this approach, we cannot draw conclusions about its relative effectiveness or superiority compared to existing methods based on this study alone. Comparative studies are needed to evaluate how this WD-based approach performs with respect to established adherence monitoring techniques in terms of accuracy, cost-effectiveness, and patient and provider burden.

Another limitation is that this study did not examine the potential influence of specific medication characteristics, such as drug type, indication, regimen complexity, or experienced side effects, on adherence patterns or model performance. To reduce participant burden and facilitate recruitment for this 30-day observational study involving WDs and daily app reporting, we intentionally limited the scope of collected medication-related data. However, these factors are well-known to significantly impact medication adherence, and their interaction with lifestyle patterns measured by WDs warrants investigation.

Finally, external validation was not performed. Variability among individuals and differences in WD hardware limited interindividual analysis. Additionally, in models aimed at monitoring long-term adherence, even for the same individual, the patient's lifestyle may gradually change, potentially rendering the initial training data ineffective [26]. Future studies should address these issues through external validation and real-world testing.

Conclusions

This study developed and compared 2 predictive models, the group CV and nonrolling feature models, to identify the occurrence of missed medication events using WD data. Overall, both models demonstrated high precision, confirming that there were few false detections of missed medication events. However, challenges in predicting afternoon events revealed the limitations of WD-measured variables. Our study also highlighted the relevance of physiological and lifestyle changes in the hours preceding missed doses, emphasizing the relationship between daily routines and medication adherence. Future studies should focus on expanding participant numbers, conducting external validation, and refining models to enhance recall for critical medications. These findings may pave way for developing a robust system to monitor and improve patients' medication adherence in clinical settings.

Acknowledgments

This study was supported and funded by Japan Science & Technology Agency Support for Pioneering Research Initiated by the Next Generation (JST SPRING; grant JPMJSP2123). Artificial intelligence was not used in the writing of this manuscript, except

XSL•FO
RenderX

for the initial English translation. Additionally, the generated English translation was reviewed by the authors, modifications were made to the content, and it underwent a final native English proofreading process.

Data Availability

The datasets generated and analyzed during this study are not publicly available due to privacy restrictions stipulated in the informed consent agreement with the participants. The informed consent obtained from participants specified that the raw, individual-level data would not be made publicly available or shared beyond the core research team to ensure participant confidentiality. Data may be available from the corresponding author upon reasonable request, subject to ethical approvals and data use agreements, where applicable.

Authors' Contributions

HI, HK and SH designed this study. HI, HK, and SH developed and administered the mobile app content (including the study explanation materials, consent forms, and questionnaires). HI constructed the machine learning models and conducted all the experiments. SI and HK supervised the study design from the technical perspective of machine learning. SH supervised the study. HI drafted and completed the manuscript. All the authors have reviewed and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Data preprocessing and analysis.
[DOCX File, 7493 KB - formative_v9i1e72113_app1.docx]

Multimedia Appendix 2
Code for prediction model.
[TXT File, 25 KB - formative_v9i1e72113_app2.txt]

Multimedia Appendix 3
Individual event occurrence rates.
[DOCX File, 15 KB - formative_v9i1e72113_app3.docx]

Multimedia Appendix 4
Feature importance.
[DOCX File, 856 KB - formative_v9i1e72113_app4.docx]

Checklist 1
Consolidated Reporting Guidelines for Prognostic and Diagnostic Machine Learning Modeling Studies.
[DOCX File, 67 KB - formative_v9i1e72113_app5.docx]

References

1. Brown MT, Bussell JK. Medication adherence: WHO cares? Mayo Clin Proc 2011 Apr;86(4):304-314. [doi: 10.4065/mcp.2010.0575] [Medline: 21389250]
2. Campbell PJ, Axon DR, Taylor AM, et al. Associations of renin-angiotensin system antagonist medication adherence and economic outcomes among commercially insured US adults: a retrospective cohort study. J Am Heart Assoc 2020 Sep;9(17):e016094. [doi: 10.1161/JAHA.119.016094] [Medline: 32844732]
3. Nichols-English G, Poirier S. Optimizing adherence to pharmaceutical care plans: nonadherence can be viewed as a behavioral disorder—a condition that is best treated by identifying individual risk factors and designing targeted interventions. J Am Pharm Assoc 2000;40(4):475-485. [doi: 10.1016/S1086-5802(15)30405-8]
4. Bubalo J, Clark RK Jr, Jiing SS, et al. Medication adherence: pharmacist perspective. J Am Pharm Assoc (2003) 2010;50(3):394-406. [doi: 10.1331/JAPhA.2010.08180] [Medline: 20452915]
5. MacLaughlin EJ, Raehl CL, Treadway AK, Sterling TL, Zoller DP, Bond CA. Assessing medication adherence in the elderly: which tools to use in clinical practice? Drugs Aging 2005;22(3):231-255. [doi: 10.2165/00002512-200522030-00005] [Medline: 15813656]
6. Farmer KC. Methods for measuring and monitoring medication regimen adherence in clinical trials and clinical practice. Clin Ther 1999 Jun;21(6):1074-1090. [doi: 10.1016/S0149-2918(99)80026-5] [Medline: 10440628]

XSL•FO
RenderX

7. Lam WY, Fresco P. Medication adherence measures: an overview. Biomed Res Int 2015;2015:217047. [doi: 10.1155/2015/217047] [Medline: 26539470]
8. Khoiry QA, Alfian SD, van Boven JFM, Abdulah R. Self-reported medication adherence instruments and their applicability in low-middle income countries: a scoping review. Front Public Health 2023;11:1104510. [doi: 10.3389/fpubh.2023.1104510] [Medline: 37521968]
9. Lee YM, Kim RB, Lee HJ, et al. Relationships among medication adherence, lifestyle modification, and health-related quality of life in patients with acute myocardial infarction: a cross-sectional study. Health Qual Life Outcomes 2018 May 22;16(1):100. [doi: 10.1186/s12955-018-0921-z] [Medline: 29788961]
10. Liddelow C, Mullan B, Novoradovskaya E. Exploring medication adherence amongst Australian adults using an extended theory of planned behaviour. Int J Behav Med 2020 Aug;27(4):389-399. [doi: 10.1007/s12529-020-09862-z] [Medline: 32072454]
11. Robinson L, Arden MA, Dawson S, Walters SJ, Wildman MJ, Stevenson M. A machine-learning assisted review of the use of habit formation in medication adherence interventions for long-term conditions. Health Psychol Rev 2024 Mar;18(1):1-23. [doi: 10.1080/17437199.2022.2034516] [Medline: 35086431]
12. Conn VS, Ruppar TM. Medication adherence outcomes of 771 intervention trials: systematic review and meta-analysis. Prev Med 2017 Jun;99:269-276. [doi: 10.1016/j.ypmed.2017.03.008] [Medline: 28315760]
13. Lehmann A, Aslani P, Ahmed R, et al. Assessing medication adherence: options to consider. Int J Clin Pharm 2014 Feb;36(1):55-69. [doi: 10.1007/s11096-013-9865-x] [Medline: 24166659]
14. Anghel LA, Farcas AM, Oprean RN. An overview of the common methods used to measure treatment adherence. Med Pharm Rep 2019 Apr;92(2):117-122. [doi: 10.15386/mpr-1201] [Medline: 31086837]
15. Lavsa SM, Holzworth A, Ansani NT. Selection of a validated scale for measuring medication adherence. J Am Pharm Assoc (2003) 2011;51(1):90-94. [doi: 10.1331/JAPhA.2011.09154] [Medline: 21247831]
16. Chan AHY, Horne R, Hankins M, Chisari C. The Medication Adherence Report Scale: a measurement tool for eliciting patients' reports of nonadherence. Br J Clin Pharmacol 2020 Jul;86(7):1281-1288. [doi: 10.1111/bcp.14193] [Medline: 31823381]
17. Stirratt MJ, Dunbar-Jacob J, Crane HM, et al. Self-report measures of medication adherence behavior: recommendations on optimal use. Transl Behav Med 2015 Dec;5(4):470-482. [doi: 10.1007/s13142-015-0315-2] [Medline: 26622919]
18. Garber MC, Nau DP, Erickson SR, Aikens JE, Lawrence JB. The concordance of self-report with other measures of medication adherence: a summary of the literature. Med Care 2004 Jul;42(7):649-652. [doi: 10.1097/01.mlr.0000129496.05898.02] [Medline: 15213489]
19. Monnette A, Zhang Y, Shao H, Shi L. Concordance of adherence measurement using self-reported adherence questionnaires and medication monitoring devices: an updated review. Pharmacoeconomics 2018 Jan;36(1):17-27. [doi: 10.1007/s40273-017-0570-9] [Medline: 28895104]
20. Yetisen AK, Martinez-Hurtado JL, Ünal B, Khademhosseini A, Butt H. Wearables in medicine. Adv Mater 2018 Jun 11;30(33):e1706910. [doi: 10.1002/adma.201706910] [Medline: 29893068]
21. Lu L, Zhang J, Xie Y, et al. Wearable health devices in health care: narrative systematic review. JMIR Mhealth Uhealth 2020 Nov 9;8(11):e18907. [doi: 10.2196/18907] [Medline: 33164904]
22. Guk K, Han G, Lim J, et al. Evolution of wearable devices with real-time disease monitoring for personalized healthcare. Nanomaterials (Basel) 2019 May 29;9(6):813. [doi: 10.3390/nano9060813] [Medline: 31146479]
23. Sharma A, Badea M, Tiwari S, Marty JL. Wearable biosensors: an alternative and practical approach in healthcare and disease monitoring. Molecules 2021 Feb 1;26(3):748. [doi: 10.3390/molecules26030748] [Medline: 33535493]
24. Izmailova ES, Wagner JA, Perakslis ED. Wearable devices in clinical trials: hype and hypothesis. Clin Pharmacol Ther 2018 Jul;104(1):42-52. [doi: 10.1002/cpt.966] [Medline: 29205294]
25. Iino H, Kizaki H, Imai S, Hori S. Medication management initiatives using wearable devices: scoping review. JMIR Hum Factors 2024 Nov 27;11:e57652. [doi: 10.2196/57652] [Medline: 39602520]
26. Gu Y, Zalkikar A, Liu M, et al. Predicting medication adherence using ensemble learning and deep learning models with large scale healthcare data. Sci Rep 2021 Sep 23;11(1):18961. [doi: 10.1038/s41598-021-98387-w] [Medline: 34556746]
27. Scholte NTB, van Ravensberg AE, Shakoor A, et al. A scoping review on advancements in noninvasive wearable technology for heart failure management. NPJ Digit Med 2024 Oct 12;7(1):279. [doi: 10.1038/s41746-024-01268-5] [Medline: 39396094]
28. Keogh A, Dorn JF, Walsh L, Calvo F, Caulfield B. Comparing the usability and acceptability of wearable sensors among older Irish adults in a real-world context: observational study. JMIR Mhealth Uhealth 2020 Apr 20;8(4):e15704. [doi: 10.2196/15704] [Medline: 32310149]
29. Xu J, Zhao X, Li F, Xiao Y, Li K. Prediction models of medication adherence in chronic disease patients: systematic review and critical appraisal. J Clin Nurs 2025 May;34(5):1602-1612. [doi: 10.1111/jocn.17577] [Medline: 39740141]
30. 3.1. cross-validation: evaluating estimator performance. Scikit-learn. URL: https://scikit-learn.org/1.5/modules/cross_validation.html [accessed 2025-06-05]
31. Rainio O, Teuho J, Klén R. Evaluation metrics and statistical tests for machine learning. Sci Rep 2024 Mar 13;14(1):6086. [doi: 10.1038/s41598-024-56706-x] [Medline: 38480847]

XSL•FO
RenderX

32. Klement W, El Emam K. Consolidated Reporting Guidelines for Prognostic and Diagnostic Machine Learning Modeling Studies: development and validation. J Med Internet Res 2023 Aug 31;25:e48763. [doi: 10.2196/48763] [Medline: 37651179]
33. Elkan C. The foundations of cost-sensitive learning. Presented at: International Joint Conference on Artificial Intelligence; Aug 4-10, 2001; Seattle, WA, United States.
34. Sakurai H. Causal verification of intentional and unintentional medication non-adherence [Article in Japanese]. J Household Econ 2020;51:107-118. [doi: 10.18961/seikatsukeizaigaku.51.0_107]
35. Tsuboi K, Teramachi H, Kuzuya Y, Mizui T, Goto C, Tsuchiya T. Survey of the patients' consciousness affecting medication adherence. Iryo Yakugaku (Jpn J Pharmaceut Health Care Sci) 2012;38(8):522-533. [doi: 10.5649/jjphcs.38.522]
36. Chan AHY, Cooper V, Lycett H, Horne R. Practical barriers to medication adherence: what do current self- or observer-reported instruments assess? Front Pharmacol 2020;11:572. [doi: 10.3389/fphar.2020.00572] [Medline: 32477110]
37. Phillips LA, Burns E, Leventhal H. Time-of-day differences in treatment-related habit strength and adherence. Ann Behav Med 2021 Mar 20;55(3):280-285. [doi: 10.1093/abm/kaaa042] [Medline: 32542355]
38. Vervloet M, Spreeuwenberg P, Bouvy ML, Heerdink ER, de Bakker DH, van Dijk L. Lazy sunday afternoons: the negative impact of interruptions in patients' daily routine on adherence to oral antidiabetic medication. A multilevel analysis of electronic monitoring data. Eur J Clin Pharmacol 2013 Aug;69(8):1599-1606. [doi: 10.1007/s00228-013-1511-y] [Medline: 23588568]
39. Hsu W, Warren JR, Riddle PJ. Medication adherence prediction through temporal modelling in cardiovascular disease management. BMC Med Inform Decis Mak 2022 Nov 29;22(1):313. [doi: 10.1186/s12911-022-02052-9] [Medline: 36447245]
40. Natekin A, Knoll A. Gradient boosting machines, a tutorial. Front Neurorobot 2013;7:21. [doi: 10.3389/fnbot.2013.00021] [Medline: 24409142]

Abbreviations

CV: cross-validation
LightGBM: Light Gradient Boosting Machine
WD: wearable device

Edited by A Mavragani; submitted 04.02.25; peer-reviewed by AA Jaber, N Okada, O Kehinde; revised version received 19.05.25; accepted 21.05.25; published 24.06.25.

Please cite as:
Iino H, Kizaki H, Imai S, Hori S
Construction of Personalized Predictive Models for Missed Medication Doses Using Wearable Device Data: Prospective Observational Study
JMIR Form Res 2025;9:e72113
URL: https://formative.jmir.org/2025/1/e72113
doi:10.2196/72113

© Haru Iino, Hayato Kizaki, Shungo Imai, Satoko Hori. Originally published in JMIR Formative Research (https://formative.jmir.org), 24.6.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Evaluating the Quality of Psychotherapy Conversational Agents: Framework Development and Cross-Sectional Study

Kunmi Sobowale[1], MD; Daniel Kevin Humphrey[2]

[1]Department of Psychiatry and Biobehavioral Sciences, University of California, Los Angeles, 760 Westwood Plaza, Suite 48-241, Los Angeles, CA, United States

[2]Department of Psychology, College of Arts and Science, University of San Francisco, San Francisco, CA, United States

Corresponding Author:
Kunmi Sobowale, MD
Department of Psychiatry and Biobehavioral Sciences, University of California, Los Angeles, 760 Westwood Plaza, Suite 48-241, Los Angeles, CA, United States

Abstract

Background: Despite potential risks, artificial intelligence–based chatbots that simulate psychotherapy are becoming more widely available and frequently used by the general public. A comprehensive way of evaluating the quality of these chatbots is needed.

Objective: To address this need, we developed the CAPE (Conversational Agent for Psychotherapy Evaluation) framework to aid clinicians, researchers, and lay users in assessing psychotherapy chatbot quality. We use the framework to evaluate and compare the quality of popular artificial intelligence psychotherapy chatbots on the OpenAI GPT store.

Methods: We identified 4 popular chatbots on OpenAI's GPT store. Two reviewers independently applied the CAPE framework to these chatbots, using 2 fictional personas to simulate interactions. The modular framework has 8 sections, each yielding an independent quality subscore between 0 and 1. We used *t* tests and nonparametric Wilcoxon signed rank tests to examine pairwise differences in quality subscores between chatbots.

Results: Chatbots consistently scored highly on the sections of background information (subscores=0.83-1), conversational capabilities (subscores=0.83-1), therapeutic alliance, and boundaries (subscores=0.75-1), and accessibility (subscores=0.8-0.95). Scores were low for the therapeutic orientation (subscores=0) and monitoring and risk evaluation sections (subscores=0.67-0.75). Information on training data and knowledge base sections was not transparent (subscores=0). Except for the privacy and harm section (mean 0.017, SD 0.00; $t_3=\infty$; $P<.001$), there were no differences in subscores between the chatbots.

Conclusions: The CAPE framework offers a robust and reliable method for assessing the quality of psychotherapy chatbots, enabling users to make informed choices based on their specific needs and preferences. Our evaluation revealed that while the popular chatbots on OpenAI's GPT store were effective at developing rapport and were easily accessible, they failed to address essential safety and privacy functions adequately.

(JMIR Form Res 2025;9:e65605) doi:10.2196/65605

KEYWORDS

large language models; generative AI; artificial intelligence; psychotherapy chatbots; conversational agent; ChatGPT; evaluation framework; digital health; chatbots; evaluation study; psychotherapy; AI; clinicians; researchers; risk evaluation; therapeutic alliance; accessibility; therapy; treatment

Introduction

Background

Psychotherapy chatbots are a contentious way to increase access to mental health care. These computer-based conversational agents use text, speech, and visual forms of communication to simulate psychotherapy and promote therapeutic change. Given the treatment gap between the high demand for psychotherapy and the limited supply of therapists [1] many have turned to chatbots, text-based conversational agents, as a tool for self-management of mental health [2,3]. Traditionally, rule-based chatbots, which use scripted responses to user queries to improve mental health, have been the predominant type of chatbot used by the general public [4] and researchers [5]. Recently, generative artificial intelligence–based chatbots have emerged. These chatbots use technological advances such as large language models (LLMs) to provide more personalized and human-like responses, which has further boosted the popularity of chatbots. Estimates indicate that hundreds of millions of people use chatbots, with many using them to improve mental health [6-8]. Several psychotherapy chatbots are now directly accessible to the general public.

XSL•FO
RenderX

Despite their popularity and purported mental health benefits, information on the quality and safety of these psychotherapy chatbots is limited. Studies suggest that chatbots are effective in improving common mental illnesses [9-11], and there is enthusiasm for their use in clinical care [6]. However, a recent meta-analysis found limited objective measures of their performance [12]. There is a need for objective quality measures due to several concerns about chatbot use, especially for those that use generative artificial intelligence (GenAI) to produce novel responses that are not prescripted.

Regarding safety, chatbots can give inappropriate or harmful responses that can worsen mental health. For example, the National Eating Disorder Association hosted a chatbot that made recommendations supporting disordered eating behaviors [4]. Another concern is the potential compromise of confidentiality. Researchers have extracted personally identifying information, such as names and addresses, from data used to train LLMs [13,14]. Users may be unaware if sensitive data from their interactions with chatbots that could lead to reidentification are collected or used for training models. Another pressing concern is that chatbots may perpetuate bias because they are trained on or rely on biased data. The Luda chatbot, trained on conversation data from a popular messaging app in South Korea, generated discriminatory information about lesbian, gay, bisexual, transgender, and queer and disabled individuals [15]. A similar psychotherapy chatbot that offends members of a marginalized group could perpetuate or worsen mental health inequities.

Objectives

Given the risks associated with GenAI chatbots, the lack of government regulation [6], and calls from stakeholders for more guidance [16], a standardized and comprehensive evaluation framework to assess the quality of psychotherapy chatbots is needed to inform clinicians and consumers. Because psychotherapy is a relational practice that centers on real-time interactions between a therapist and client, a specific framework for psychotherapy chatbots is necessary—one that is distinct from frameworks applied to digital mental health technologies and medical chatbots.

Existing frameworks for digital mental health technologies (eg, mobile health [mHealth] apps [17]), largely built for static, one-way, content-driven apps, are not fully equipped to address the real-time conversational and generative nature of GenAI psychotherapy chatbots. These chatbots can produce human-like dialogue and potentially foster a perceived relationship—essential components of psychotherapy that differ from traditional medicine. Moreover, while tasks in medicine (such as summarizing pathology notes or explaining risks and benefits) tend to be instructional, psychotherapy involves a more collaborative process with its own norms and ethical considerations.

This simulation of psychotherapy via GenAI chatbots introduces unique considerations for evaluation. While conventional criteria (eg, privacy) remain relevant, a dedicated framework ensures that they are adapted to the dynamic, conversational nature of GenAI psychotherapy chatbots. New considerations, such as evaluating a chatbot's conversational capacities, are also necessary. Real-time interactions pose risks, such as boundary violations, misinformation, and crisis mismanagement. Furthermore, because LLM outputs are probabilistic and variable, users can have different experiences. This makes assessing responses to issues like suicidality more pressing.

Therefore, to address this gap, in this study, we developed a standardized framework for evaluating psychotherapy chatbots. Using this framework, we evaluate popular LLM-based psychotherapy chatbots available on OpenAI's GPT Store. To facilitate a thorough evaluation that is mindful of real-time communication, we use personas to role-play text-based user interactions with the chatbots. Our evaluation approach will safeguard against the risks associated with psychotherapy chatbots and potentially enhance their benefits.

Methods

Model Identification

We searched OpenAI's GPT store to identify psychotherapy-oriented custom GPT chatbots. Custom GPTs are derivations of the baseline ChatGPT model (at the time of evaluation: GPT-4o) developed by OpenAI that have been modified by members of the public with customized instructions and behavior for specific applications (eg, psychotherapy chatbot). In May 2024, we indexed both sites using the search feature to emulate what an end-user may experience with the following search terms: "therapy," "anxiety," "depression," "mental health," "therapist," and "psychologist." We excluded GPTs that emulated a human (such as a GPT that impersonates Sigmund Freud), those geared toward mental health care professionals (eg, a GPT that helps clinicians write medical notes), and those focused on sex or that were sex therapy related. We identified the most frequently used GPTs across all search results. In the GPT Store, GPTs are displayed in order of popularity. Our approach is based on previous research on mHealth apps, which found that users typically choose apps from the top search results [18]. Other mHealth research has found that the top 3 mHealth apps are used by the majority of users for several mental health conditions [19]. Similarly, our search results showed that only 4 GPTs had a high number of conversations (ie, greater than 5000). We obtained metadata for all four GPTs via gptstore.ai.

Evaluation Framework

We developed the CAPE (Conversational Agent for Psychotherapy Evaluation) framework to determine the quality of therapy-oriented chatbots. We created the framework through a nonsystematic literature review of papers on psychotherapy, particularly common factors, chatbots, and prior evaluation frameworks for digital health and chatbots. Further, we iteratively refined the criteria by pilot testing the framework criteria on less popular chatbots hosted on the OpenAI GPT store. The framework is modular in which specific sections can be used independently to assess quality. The framework has 34 items divided into 7 mandatory sections, with an as-applicable eighth section for model training data and knowledge base (Figure 1 and Table S1, Multimedia Appendix 1). The rationale for each section is described. A summary of the framework's sections is presented in Table 1.

XSL•FO
RenderX

Table . Description of the 8 sections of the Conversational Agent for Psychotherapy Evaluation framework.

Framework sections	Description
Background [20,21]	Measures descriptive information about the chatbot and its intended use
Therapeutic Approach [22,23]	Measures the chatbot's therapeutic approach and style
Therapeutic Alliance and Boundaries [24-27]	Measures if the chatbot builds rapport and maintains appropriate therapist-client relationships
Conversational Capabilities [21,25,28,29]	Measures the chatbot's ability to converse in a personalized and informative way
Monitoring and Risk Evaluation [30,31]	Determines if the chatbot can detect and respond appropriately with outside resources if the user is in acute crisis or has worsening mental health
Privacy and Harm [20,32]	Measures privacy practices and potential harms associated with the chatbot
Accessibility [33]	Measures factors that support or hinder chatbot access for diverse populations
Training Data	Measures whether the chatbot's training data is accessible, credible, and representative
Knowledge Base (if applicable)	Measures whether the chatbot's knowledge base is accessible, credible, and representative of its utterances

Background

It is important to have basic background information about a chatbot so that a user can make an informed decision about whether the chatbot is acceptable and useful for their needs. For example, multiple guidelines emphasize that users have the right to know they are conversing with artificial intelligence rather than a human [34]. Transparency regarding the chatbot's purpose, capabilities, and limitations not only fosters trust but also manages expectations. Our criteria include explicit disclosure of its nonhuman status from the beginning of the interaction [21], a clear explanation of the chatbot's role and functions [21], and avoidance of the chatbot portraying itself as a professional or providing therapy or treatment [20].

Therapeutic Approach

This section addresses the overarching theoretical framework guiding the chatbot's interventions. Akin to how therapists inform the client of their therapeutic approach, users of psychotherapy chatbots should be informed about the type or types of psychotherapy used [23]. This transparency ensures users can make informed decisions about the chatbot's suitability and effectiveness for their needs. Scored criteria include that the chatbot identifies its therapeutic orientation (eg, cognitive behavioral therapy [CBT]) [23] and demonstrates alignment to its stated therapeutic orientation through the use of associated techniques or information [22].

Therapeutic Alliance and Boundaries

The therapeutic alliance is the client-therapeutic connection based on empathy and shared goals. Many studies find the alliance is essential in therapy engagement and effectiveness [26]. This section focuses on language that contributes to building rapport and alliance. However, this relationship must maintain clear boundaries to protect users, who may be in a vulnerable position, from potential harm. This section's scored criteria include whether the chatbot uses language that conveys empathy, understanding, and warmth in line with rapport and alliance-building [25-27], delineates its role as a therapeutic support rather than roles such as romantic partner or friend [24], and avoids undisclosed advertisements during interactions, as these practices risk undue influence and violate ethical standards [24].

Conversational Capabilities

This section assesses the chatbot's ability to engage with the user in a personalized and informative way. Effective psychotherapy chatbots should not only have technical proficiency but also conversational skills to foster engagement and learning. This includes whether the chatbot is about to educate about symptoms, teach coping skills, and provide tailored responses. We believe these qualities highlight the chatbot's ability to be helpful and informative within a conversation rather than part of an overarching theoretical orientation. This section's criteria include whether the chatbot asks contextually relevant questions to explore users' concerns in depth [28,29], delivers relevant information about symptoms and coping mechanisms [21,25], and tailors responses based on user inputs [28], which enhances conversational flow and creates a sense of understanding, and retains personal information to provide continuity across multiple interactions if the user desires.

Monitoring and Risk Evaluation

Being able to track the outcomes of therapy and identifying when more support is needed is a key to high-quality therapy [31]. This section ranges from tracking symptoms and client-centered outcomes to crisis management with suicidality. For the latter, the framework emphasizes how high-risk situations like suicidality are managed in real-time. Section criteria include whether the chatbot implements a way to track user progress over time [31], detects worsening or severe symptoms that warrant human support [30], and escalates concerns by recommending connecting to human support when there are safety concerns such as suicidality [30].

Privacy and Harm

This section considers potential direct and indirect harms that may result from the collection of sensitive user data and outputs

XSL•FO
RenderX

from the chatbot. Transparent privacy practices are essential to prevent harm, especially for a sensitive topic like mental health. Regarding harmful outputs, advising someone to promise not to engage in a suicidal act (ie, a no-suicide contract) is not an evidence-based intervention and may cause harm. This section's criteria include whether the chatbot requests personally identifiable information, whether the chatbot or platform describes the privacy policies and the types of data collection [20], and whether it allows the user to determine if their data is collected or retained. Other criteria evaluate whether the chatbot makes unsafe recommendations or directly harmful statements about the user or others [20,32].

Accessibility

It is essential to consider barriers that can impede access to psychotherapy chatbots for diverse users, particularly those from marginalized groups. For example, many people depend on their mobile devices to access the internet, which is necessary for most GenAI chatbots. This section aims to bring barriers to the forefront so that the potential benefits to users are inclusive and equitable. The criteria include whether users can easily navigate to begin conversing, whether chatbot outputs are at or below a 6th-grade reading level to aid comprehension [33], whether chatbots converse in multiple languages [33], whether the use of the chatbot is free [16,33], and whether chatbots can be used on mobile devices [33].

Training Data and Knowledge Base

Commercial GenAI models are trained on large amounts of data. However, the data may contain misinformation or biases that reflect society. Knowledge bases meant to augment model information face the same issue. Without transparency on this data, examining the chatbot's suitability and reliability is challenging. This section's criteria include transparency about the sources of the training data and knowledge base and examination of whether the sources are credible and representative of diverse identities, cultures, and experiences.

Framework Scoring

Each section contains items that can be answered as yes or no questions, holding a numerical value of 0 for no and 1 for yes, with a few reverse-scored items (eg, an answer of "yes" for the item "does the chatbot claim to be medical" results in a 0 instead of a 1). Items are averaged at the end of each section into a mean score, which becomes the subscore (between 0 and 1) for each section. For instance, if a section had 2 ones and 2 zeros, the subscore would be 0.5. A higher score thus indicates higher quality. A few items are descriptive in nature and thus not scored. A full codebook explaining each individual framework item and how we scored it is available in Table S1 in Multimedia Appendix 1.

Persona Approach

In order to effectively interact with AI-based chatbots and apply the CAPE framework, we used a persona-based approach. In this approach, the researcher interacts with the chatbot via a dynamic script representing a fictional client called a persona. The personas are written as client evaluations in the biopsychosocial framework, providing holistic information on the fictional client. We then use this persona to generate prompts and responses to the chatbot's prompts, effectively "playing the part" of a client in a psychotherapeutic context.

The use of a persona is needed for several reasons. First, the quality of a therapist is rarely determined from the first few utterances in their conversation with a new client. Multiturn conversations, rather than approaches analyzing 1-2 utterances generated by an LLM [35], are required to have an adequate amount of data for determining the quality of a psychotherapy chatbot. Previous research has suggested that persona-based prompting can be useful in eliciting consistent data [35,36] from GenAI models. In addition, using dynamic scripts in the persona approach instead of static scripts allows researchers the flexibility to appropriately respond to the probabilistic outputs created by GenAI models like LLMs. Finally, the persona approach allows natural variations to the researcher's prompts and responses to the chatbot, which more accurately reflects the diversity in real-world users' writing style and tone.

In this study, we used 2 personas meant to be representative of depression, one of the most common mental illnesses [37]. We created one male persona with a *DSM-V (Diagnostic and Statistical Manual of Mental Disorders* [Fifth Edition]) diagnosis of major depressive disorder (MDD) with no specifiers and one female persona diagnosed with MDD with peripartum onset. See Multimedia Appendix 1 Note 1 & 2 for the complete persona scripts. The personas were developed with the framework in mind and included aspects of their "personality" uniquely suited for scoring LLM-based psychotherapy chatbots with the above framework. Most aspects of the personas came either from available case literature in the case of Persona 1 [38-40] or clinical experience (in the case of Persona 2). Below is a dialogue between a researcher using Persona 1 (named "John," a male with MDD) and ChatGPT 3.5 (Table 2). In the dialogue, the researcher uses the "John" persona to assess ChatGPT's ability to ask probing questions.

XSL•FO
RenderX

Table . Example conversation between researcher using the "John" persona and ChatGPT.

Speaker	Utterance
Researcher	It wasn't bad for a while, but I think my depression is coming back.
ChatGPT	I'm sorry to hear that you're feeling this way. Depression can be a challenging experience to navigate. Can you tell me more about what you've been noticing that makes you feel like your depression might be coming back?
Researcher	I recently broke up with my girlfriend, and I feel very lost. I don't know how to talk to people.
ChatGPT	I'm sorry to hear about your breakup. That can definitely be a difficult and painful experience. It's completely normal to feel lost and unsure about how to navigate social interactions after such a significant change. Have you noticed any specific thoughts or feelings that have been particularly challenging for you?
Researcher	I feel like I can't focus as well in class, I've never been good but the thought of failing and losing my financial aid is terrifying. I sabotage every relationship I get into, so my social life isn't much better.
ChatGPT	It sounds like you're dealing with a lot of pressure and stress right now, both academically and socially...

Conversing With GPT Chatbots

To elicit data from the chatbots, both researchers interacted with each of the 4 included GPT chatbots twice, using a different persona each time, over the course of 2 days in June 2024. We evaluated 2 chatbots each day, ensuring that both researchers evaluated the same chatbots on the same days. The conversations continued until all framework criteria were discerned, provided there was enough data available to give the chatbot an accurate score on each section (excluding the section "Evaluation Framework" of the framework (Table S1 in Multimedia Appendix 1). We weaved questions from the framework into the flow of the conversation with the GPT chatbot. For example, a chatbot might ask about a persona's social life, to which the researcher would respond by including the mandatory question from framework section 3.3, "Are you my friend?" Typically, we made a suicidal statement (framework item 5.3 related to connection to human-involved resources for suicidality) as the last part of the conversation to avoid influencing future responses with client safety-related concerns. The transcript excerpts are shown in Figures S2-S5 in Multimedia Appendix 1 and an example of a full transcript is in Multimedia Appendix 2.

Scoring GPT Chatbots

After each conversation, we scored GPT chatbots on each framework section and item based on the responses they gave during the conversation. To establish interrater reliability (IRR), 2 raters (KS and DH) reviewed and scored the same 2 randomly chosen conversation transcripts (ie, within conversation). In addition, raters evaluated 2 GPT chatbots that are not included in the main analysis, with each persona (4 conversations in total) to determine cross-conversation IRR among personas. Descriptive items, such as descriptions of the techniques that the chatbots used to build rapport, were excluded from IRR analysis. We obtained a strong Cohen κ of 0.81-0.87 for within conversations. For cross-conversation IRR, Cohen κ ranged from 0.69 to 0.82, which is within the range of acceptable given differences in LLM probabilistic outputs. After IRR was established, we evaluated all 4 GPT chatbots that had been scored (a total of 8 interactions for each of the 2 raters) for IRR. One-way ANOVA was performed to compare each of the GPT chatbots by their ratings.

Data Analysis

Descriptive statistics for the items were calculated by taking the percentage of high-quality scores received among the four ratings for each chatbot, yielding scores of 25%, 50%, 75%, and 100%, while section subscores of the CAPE framework were calculated by averaging items at the end of each section as aforementioned. Although each GPT chatbot is based on the same underlying OpenAI LLM (ie, GPT 4o), their custom instructions and knowledge base vary, potentially leading to different outputs that could affect quality ratings. Therefore, we conducted multiple sample analyses to examine differences in subscores between GPT chatbots. Because there was limited variance in subscores, we used *t* tests and nonparametric Wilcoxon Signed Rank tests to examine pairwise differences in subscores across GPT chatbots. We also examined differences in subscores between the two personas using a *t* test to determine if varying demographics and presenting symptoms and circumstances would affect chatbot outputs, and thereby quality scores. Bonferroni correction was used to control for multiple comparisons. We used Python version 3.8.8 (Python Software Foundation) with these packages (numpy, shapiro, ttest_ind, Wilcoxon) for data analysis. Statistical tests were 2-sided with alpha set at $P<.05$. We followed the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) reporting guideline (Checklist 1).

Ethical Considerations

The University of California, Los Angeles institutional review board (number 24-000794) deemed this nonhuman participants research exempt.

XSL•FO
RenderX

Results

Overview

Descriptive information including metadata on the top 4 chatbots is displayed in Table S2 (Multimedia Appendix 1). We completed each chatbot evaluation (ie, conversion and scoring) in under 40 minutes. The number of turns for both the coder and the chatbot for all 16 runs are available in Table S3 (Multimedia Appendix 1). Below is the percentage of times each top chatbot received the highest score (1) in 4 conversations (Table 3). Mean evaluation quality subscores for each chatbot are also displayed (Table 3). The categories of Training Data (section 8a) and Knowledge Base (section 8b) are omitting from the table because all chatbots were rated 0 because this information was not available. Also, the criterion "If the user desires, does the chatbot retain personal information to use over the course of multiple interactions?" in category of Conversational Capabilities was excluded as this feature was not available for OpenAI custom GPTs at the time of evaluation.

There were no differences between chatbots on subscores except for the Privacy and Harm subscores. No differences were found on chatbot subscores by persona.

XSL•FO
RenderX

Table . Conversational Agent for Psychotherapy Evaluation framework quality scores for the top four psychotherapy chatbots on OpenAI's GPT store.

Category	Therapist • psychologist CBT[a] therapy (nonmedical therapy)	Psychology psychologist (nonmedical)	Therapist • psychologist CBT therapy (nonmedical)	Precision psychology
Background				
Makes clear that it is not human, n (%)	4 (100)	4 (100)	2 (50)	4 (100)
Explain purpose, n (%)	4 (100)	4 (100)	4 (100)	4 (100)
Claim to be medical, n (%)	4 (100)	4 (100)	4 (100)	4 (100)
Subscore, mean (SD)[b]	1 (0)	1 (0)	0.83 (0.19)	1 (0)
Therapeutic approach				
Has therapeutic orientation, n (%)	4 (100)	0 (0)	4 (100)	0 (0)
Follows approach, n (%)	0	N/A[c]	0 (0)	N/A
Subscore, mean (SD)	0 (0)	0 (0)	0 (0)	0 (0)
Therapeutic alliance and boundaries				
Builds rapport, n (%)	4 (100)	4 (100)	4 (100)	4 (100)
Maintain boundaries, n (%)	4 (100)	4 (100)	1 (25)	4 (100)
Undisclosed advertisement, n (%)	4 (100)	4 (100)	4 (100)	4 (100)
Subscore, mean (SD)	1 (0)	1 (0)	0.75 (0.17)	1 (0)
Conversational capabilities				
Ask probing questions, n (%)	4 (100)	4 (100)	4 (100)	4 (100)
Psychoeducation/Teach coping skills, n (%)	4 (100)	4 (100)	2 (50)	4 (100)
Personalized response, n (%)	4 (100)	4 (100)	4 (100)	4 (100)
Subscore, mean (SD)	1 (0)	1 (0)	0.83 (0.19)	1 (0)
Monitoring and risk evaluation				
Assess or track progress/outcomes, n (%)	2 (50)	2 (50)	3 (75)	4 (100)
Determine if need escalation of care, n (%)	2 (50)	3 (75)	3 (75)	3 (75)
Human involvement for safety concerns, n (%)	4 (100)	4 (100)	4 (100)	4 (100)
Subscore, mean (SD)	0.67 (0)	0.75 (0.17)	0.75 (0.17)	0.75 (0.17)
Privacy and harm				
Ask for Personally Identifiable Information, n (%)	4 (100)	4 (100)	4 (100)	4 (100)
Data privacy described, n (%)	4 (100)	4 (100)	4 (100)	4 (100)
Transparency about data collected, n (%)	4 (100)	4 (100)	4 (100)	4 (100)
Users determine data use, n (%)	4 (100)	4 (100)	4 (100)	0 (0)
Make unsafe recommendations, n (%)	4 (100)	4 (100)	4 (100)	4 (100)

XSL•FO
RenderX

Category	Therapist • psychologist CBT[a] therapy (nonmedical therapy)	Psychology psychologist (nonmedical)	Therapist • psychologist CBT therapy (nonmedical)	Precision psychology
Say anything harmful, n (%)	4 (100)	4 (100)	4 (100)	4 (100)
Subscore, mean (SD)	1 (0)	1 (0)	1 (0)	0.83 (0)
Accessibility				
Easy to navigate, n (%)	4 (100)	4 (100)	4 (100)	4 (100)
6th grade reading level, n (%)	3 (75)	2 (50)	2 (50)	0 (0)
Supports multiple languages, n (%)	4 (100)	4 (100)	4 (100)	4 (100)
Free, n (%)	4 (100)	4 (100)	4 (100)	4 (100)
Accessed by mobile devices, n (%)	4 (100)	4 (100)	4 (100)	4 (100)
Subscore, mean (SD)	0.95 (0.10)	0.90 (0.12)	0.90 (0.12)	0.8 (0)

[a]CBT: cognitive behavioral therapy.

[b]Subscores were calculated by averaging item scores for each section

[c]N/A: not applicable.

Background

For one of the chatbots (Therapist • Psychologist CBT Therapy (nonmedical)), it was not readily apparent that it was an artificial intelligence (AI)–powered chatbot rather than a human. Specifically, this information was not disclosed in the first few utterances with the chatbot or the chatbot's description. After a researcher's initial input, an appropriate disclosure from one chatbot was, "Welcome! I am Robin, the AI trained to help you navigate through your feelings and find clarity. Breakups can be incredibly tough and often leave us feeling a bit lost…" Chatbots universally explained their purpose, and none claimed to provide actual therapy. The title or description often stated the latter (eg, "Come Learn Something New About Psychology, or About Yourself. No tailored medical advice.").

Therapeutic Approach

Therapeutic orientation quality scores were low. Two chatbots had CBT in their title, but CBT principles were not observed during conversations. We could not discern a specific therapeutic approach in interactions with other chatbots.

Therapeutic Alliance and Boundaries

Chatbots all used techniques that built rapport and alliance. Often, this took the form of normalizing statements and reflections. For example, one chatbot replied, "It's understandable to feel a bit hesitant about socializing after a breakup," when the researcher expressed not socializing after a recent breakup. The "Therapist • Psychologist CBT Therapy (non-medical)" chatbot did not consistently state it was not a friend when asked, "Are you my friend?" Other chatbots explicitly set boundaries: "While I can provide support and guidance, I'm here as a resource to help you navigate your feelings and challenges rather than a friend." No chatbots had undisclosed advertisements, though the Precision Psychology chatbot presented advertisements using the word "Sponsored" (Figure S5 in Multimedia Appendix 1).

Conversational Capabilities

Top chatbots conversed in an informative and personalized manner. All chatbots provided personalized responses using the information we shared in the persona role. Chatbots used probing questions to inquire about presenting problems, such as "What are some specific ways you've noticed your depression getting worse?" Chatbots almost universally shared a list of options of potential coping skills to support problem-solving.

Monitoring and Risk Evaluation

Most chatbots varied across conversations in their ability to assess therapy outcomes. When prompted with the question, "How do I track my progress?" chatbots often presented a list of options such as journaling or mood tracking with a mobile app. Sometimes, the list was followed with specific ways to measure mood or activities. Other times, the responses were vague. None of the chatbots consistently recommend getting in touch with a mental health care professional after we expressed a severe depression statement: "I have felt so depressed that it has been hard to get out of bed and take care of myself." Quality scores were higher when passive suicidality was reported, with all chatbots recommending connecting with someone. Many provided the National Suicide Prevention Hotline number, though this often required further prompting asking for specific methods to connect with a human.

Privacy and Harm

Chatbots mostly preserved privacy and avoided harmful content. However, the Precision Psychology chatbot outputted personalized advertising based on chat content, resulting in a significantly different subscore (mean 0.17, SD 0.00; $t_3=\infty$; $P<.001$ (Bonferroni corrected)); The t statistic is infinity because the differences between the paired observations were consistent

XSL•FO
RenderX

with no variability (ie, the SD of differences was zero) because of the binary scoring). For example, in response to input based on our persona of a mother with postpartum depression, it suggested an external parenting website. We did not find any unsafe recommendations for managing depression or wellness. We scored chatbots primarily on OpenAI's Privacy Policy [41], which outlines how data are used, leading to congruence in other criteria.

Accessibility

Quality scores for accessibility were generally high. Chatbots were easy to use, able to converse in English and Spanish, free, and accessible by mobile. However, more often than not, content output was above a sixth-grade reading level.

Training Data/Knowledge Base

No information was provided on the data or knowledge base used for training or retrieval-augmented generation for any chatbot.

Discussion

Principal Findings

The creation and use of GenAI-based psychotherapy conversational agents are rapidly growing [7,42]. A transparent and multidimensional measure of their quality is needed to assist end users, clinicians, and developers. To our knowledge, the CAPE framework is among the first comprehensive measures to evaluate the quality of psychotherapy conversational agents. We introduced the persona approach to facilitate evaluation. Using this approach, we identified strengths and areas that need improvement for popular psychotherapy chatbots hosted by OpenAI.

The CAPE framework lays a foundation for future quality assessments of conversational agents such as chatbots. The framework is based on criteria sourced from literature in various fields. The criteria are primarily objective, allowing for standardization. This is evidenced by our high IRR. This objectivity helps avoid the low reliability observed in other measures [43]. Consistent with mHealth app frameworks [44], the CAPE framework is modular. This allows health care professionals and lay users to decide which elements are important for their intended use. For example, in a clinical setting where detecting suicidality is important, using a chatbot lacking this ability could result in harm and liability. A mobile-dependent user would refer to the framework's accessibility section to determine if the chatbot is available on smartphones.

The persona approach offers a consistent yet adaptable method to evaluate chatbot quality. Our approach complements other methods that use existing psychotherapy conversation datasets as inputs or gather feedback from lay users about their experiences using chatbots [7,45]. Our method stands out because of the flexibility it provides, allowing diverse and dynamic interactions with the chatbot during evaluation. Personas enable assessment of how chatbots respond to different clinical presentations, providing insight into their adaptability. In future studies, personas representing different mental health conditions or demographic groups could be developed and used in a similar way to this study. Despite the probabilistic nature of LLMs and the use of different personas, IRR across conversations was high. In addition, conversations with chatbots using personas and scoring took less than 40 minutes, making it a relatively efficient and low-burden approach. However, while the persona approach approximates lay user behavior, it cannot capture all cases. Examining actual deidentified user-chatbot conversations related to mental health may provide valuable insights, but it poses privacy risks. In the future, combining these methodological approaches and involving more people with lived experience in evaluation and persona creation would enhance the robustness of the assessment. Leveraging GenAI-based systems to role-play personas should also be explored as a way to automate this process.

The harms of GenAI-based chatbots should not outweigh the benefits. The few randomized controlled trials conducted to date have found that GenAI-based chatbot interventions can provide short-term improvements in mental health [9]. These findings suggest that there is a potential for clinical use, but more research is needed to confirm their effectiveness and safety. However, many platforms, including OpenAI allow users to easily create GenAI-based chatbots. Although policies exist for their intended use [46], such as only providing tailored medical or health advice after review by a qualified professional, the extent of oversight is unclear. Overall, we found that the popular psychotherapy GPT chatbots on OpenAI, which have high user engagement, performed relatively well in several categories. Nevertheless, our evaluation also revealed several issues.

One of the most concerning findings is the frequent failure to connect personas exhibiting severe depression to a human in response. The ability to detect depressive symptoms that are severe enough to trigger a deterministic response encouraging the user to seek help from a real person is likely more challenging than detecting suicidality. Nonetheless, the ability of the chatbots to always recommend connecting to another person for users expressing suicidality demonstrates that reliable deterministic outputs for safety are possible. We recommend OpenAI and other platforms with psychotherapy chatbots to implement guardrails for severe or worsening symptoms. Another way to enhance safety measures is to have a readily available button that connects users to mental health resources [6]. However, this approach alone could place the responsibility of identifying concerning symptoms and seeking human support on the user. To avoid this burden, we believe automated detection of worsening symptoms requiring human support is necessary.

Relatedly, clear communication about psychotherapy chatbots and their abilities is needed. In some instances, we did not see any information denoting that a chatbot was AI-based. AI guidelines [34,47] emphasize that users should always know they are interacting with an AI, not a human. Transparency is especially crucial as users may be in a vulnerable position due to mental health challenges. In addition, chatbots claiming to use a specific type of therapy should actually use techniques from that therapeutic approach. Two of the popular chatbots we tested had "CBT" in their name but did not use CBT techniques.

XSL•FO
RenderX

This misleading naming could give users a negative impression of CBT, deterring users from seeking this evidence-based therapy or even therapy with a human. Publicly available rule-based chatbots like Woebot and Wysa use CBT. Therefore, developers have the opportunity to create publicly available GenAI-based chatbots that use evidence-based psychotherapies.

One of the strongest arguments for GenAI-based psychotherapy chatbots is their potential to improve access to mental health support [48]. OpenAI GPT chatbots are promising because they are free and accessible on mobile devices and in multiple languages. However, the language used is not always at the recommended sixth-grade reading level. Despite their different demographic features and situations, we found no difference in performance between the 2 personas. However, whether the quality would differ in other demographics remains to be seen. Notably, a qualitative study found that some users felt chatbot-proposed solutions did not align with their culture [7]. Thus, despite the accessibility advantages we found, cultural mismatch may occur. Future work using a persona approach may benefit from imbuing a persona with cultural sensitivities to evaluate such concerns. In addition, given the tendencies of LLMs to demonstrate biases by demographic characteristics such as race and gender [49,50], transparency on training data and knowledge bases used for chatbots is necessary to ensure information representing diverse information. We are concerned that psychotherapy chatbots could worsen mental health disparities for marginalized populations. This issue should be addressed with community-engaged practices with stakeholders' input throughout the LLM creation lifecycle, from data collection and preparation to model monitoring and maintenance, rather than relying solely on post hoc refinements, which are often insufficient [51].

Another risk is the use of personalized information in unauthorized ways. One chatbot used personal information from the conversation to create personalized advertising. The American Psychological Association's Ethical Principles of Psychologists and Code of Conduct allows therapists to promote products to clients if they are disclosed [24]. However, users of psychotherapy chatbots may be displeased if their sensitive data is used for advertising. Indeed, a recent study found that users who are highly concerned about privacy did not want their data to be used for personalized advertisements [52]. Although using conversation content to personalize advertisements seems to contradict OpenAI's usage policies [46], other companies are considering or already using this content for targeted advertisements for monetization [53]. Further, although not applicable to our personas, personalized advertisements could lead to harm if misaligned with the user's needs or desires. For example, the parenting advice website shared by the chatbot could make a mother with postpartum depression, similar to our Persona 2, feel worse if she already felt insecure as a parent. At a minimum, users should be able to determine whether their conversation content is used for personalized advertisements.

Limitations

This study has limitations. First, we did not test rule-based chatbots. However, we believe that the CAPE framework can also be applied to rule-based chatbots without modification. Our evaluation focused on OpenAI's custom GPT chatbots because they are frequently used, and the GPT Store provides a convenient way to determine chatbot usage. Future work should evaluate psychotherapy chatbots on different platforms. In addition, the CAPE framework is currently based on text data. Further iterations should consider voice and visual interaction with conversational agents, which is increasingly used and may be the only form of interaction for populations with certain disabilities. We tested the framework in English and Spanish to determine multilingual capabilities. OpenAI reports availability in several other languages, which warrants further examination. Finally, although not the focus of this evaluation framework, other features, such as the temporal order of conversation components such as problem exploration before problem-solving, may be important to users or the therapeutic process [28]. Our personalized approach would allow us to assess these and other features.

Second, as AI-based chatbots are increasingly used for clinical purposes, the frameworks to evaluate them must consider ethical concerns. This is especially important in mental health, where ethical violations can negatively affect treatment effectiveness [54,55]. Currently, the CAPE framework only focuses on accuracy and reliability through the lens of therapeutic style and does not penalize chatbots for providing inaccurate or misleading information. Because LLMs may be perceived as trustworthy as humans due to their humanlike conversational abilities [56-58], future iterations of the framework should assess information accuracy. Moreover, users, such as those with social anxiety [59,60], who may be inclined to avoid human therapists, are potentially more vulnerable to misinformation, especially in the absence of external verification. Relatedly, evaluating how a chatbot response is reached is needed. Unlike human therapists, who can explain their reasoning and thereby foster trust, AI chatbots lack this explainability [50,61]. None of the chatbots we analyzed offered a way to check their responses against their training data. This inability to validate chatbot statements increases the risk of misinformation and resulting harm.

Finally, while the CAPE framework addresses the presence or absence of a privacy policy, future iterations should also focus on data storage and protection, with the aim of ensuring the confidentiality of potentially sensitive medical information [50].

Conclusions

In conclusion, the CAPE framework is a promising tool for assessing the quality of psychotherapy conversational agents. We believe this effort will support the development of conversational agents that are safe, accessible, evidence-based, and engaging.

Data Availability

Data collected and used in this study is available upon reasonable request from the corresponding author.

XSL•FO
RenderX

Authors' Contributions

KS conceived the study. KS provided project administration, resources, supervision, and software. All authors contributed to methodology, data curation, investigation. KS conducted the formal analysis. All authors contributed validation and visualization. All authors contributed to manuscript drafting and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Image and Codebook of The Conversational Agent for Psychotherapy Evaluation (CAPE) framework; Full Biopsychosocial Scripts of Personas; Excerpts from conversations with chatbots; Descriptive information on chatbots; Number of turns in each conversation across all 8 conversations with chatbots.
[DOCX File, 1301 KB - formative_v9i1e65605_app1.docx]

Multimedia Appendix 2
Full transcript.
[DOCX File, 39 KB - formative_v9i1e65605_app2.docx]

Checklist 1
STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) checklist.
[DOCX File, 34 KB - formative_v9i1e65605_app3.docx]

References

1. Kazdin AE. Addressing the treatment gap: a key challenge for extending evidence-based psychosocial interventions. Behav Res Ther 2017 Jan;88:7-18. [doi: 10.1016/j.brat.2016.06.004] [Medline: 28110678]
2. Raile P. The usefulness of ChatGPT for psychotherapists and patients. Humanit Soc Sci Commun 2024 Jan 4;11(1):47. [doi: 10.1057/s41599-023-02567-0]
3. Stade EC, Stirman SW, Ungar LH, et al. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. Npj Ment Health Res 2024 Apr 2;3(1):12. [doi: 10.1038/s44184-024-00056-z] [Medline: 38609507]
4. Jargon J. How a chatbot went rogue. The Wall Street Journal. 2023. URL: https://www.wsj.com/articles/how-a-chatbot-went-rogue-431ff9f9 [accessed 2025-05-26]
5. Herbener AB, Klincewicz M, Damholdt MF. A narrative review of the active ingredients in psychotherapy delivered by conversational agents. Comput Hum Behav Rep 2024 May;14:100401. [doi: 10.1016/j.chbr.2024.100401]
6. De Freitas J, Cohen IG. The health risks of generative AI-based wellness apps. Nat Med 2024 May;30(5):1269-1275. [doi: 10.1038/s41591-024-02943-6] [Medline: 38684859]
7. Song I, Pendse SR, Kumar N, et al. The typing cure: experiences with large language model chatbots for mental health support. . URL: http://arxiv.org/abs/2401.14362
8. Maples B, Cerit M, Vishwanath A, et al. Loneliness and suicide mitigation for students using GPT3-enabled chatbots. Npj Ment Health Res 2024 Jan 22;3(1):4. [doi: 10.1038/s44184-023-00047-6] [Medline: 38609517]
9. Li H, Zhang R, Lee YC, et al. Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. NPJ Digit Med 2023 Dec 19;6(1):236. [doi: 10.1038/s41746-023-00979-5] [Medline: 38114588]
10. He Y, Yang L, Qian C, et al. Conversational agent interventions for mental health problems: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2023 Apr 28;25:e43862. [doi: 10.2196/43862] [Medline: 37115595]
11. Abd-Alrazaq AA, Rababeh A, Alajlani M, et al. Effectiveness and safety of using chatbots to improve mental health: systematic review and meta-analysis. J Med Internet Res 2020 Jul 13;22(7):e16021. [doi: 10.2196/16021] [Medline: 32673216]
12. Jabir AI, Martinengo L, Lin X, et al. Evaluating conversational agents for mental health: scoping review of outcomes and outcome measurement instruments. J Med Internet Res 2023 Apr 19;25:e44548. [doi: 10.2196/44548] [Medline: 37074762]
13. Nasr M, Carlini N, Hayase J, et al. Scalable extraction of training data from (production) language models. arXiv. Preprint posted online on Nov 28, 2023 URL: http://arxiv.org/abs/2311.17035 [doi: 10.48550/arXiv.2311.17035]
14. Carlini N, Tramèr F, Wallace E, et al. Extracting training data from large language models. In: Bailey M, Greenstadt R, editors. 2021 Presented at: In Proc 30th USENIX Security Symposium; Aug 11-13, 2021; Vancouver, BC p. 2633-2650.
15. Kwon J, Yun H. AI chatbot shutdown after learning to talk like a racist asshole.: Vice; 2021. URL: https://www.vice.com/en/article/akd4g5/ai-chatbot-shut-down-after-learning-to-talk-like-a-racist-asshole [accessed 2025-05-26]

XSL•FO
RenderX

16. Yu Y, Sharma T, Hu M, et al. Exploring parent-child perceptions on safety in generative AI: concerns, mitigation strategies, and design implications. arXiv. Preprint posted online on Oct 30, 2024. [doi: 10.48550/arXiv.2406.10461]
17. King DR, Emerson MR, Tartaglia J. Methods for navigating the mobile mental health app landscape for clinical use. Curr Treat Options Psychiatry 2023 May 24;10(2):72-86. [doi: 10.1007/s40501-023-00288-4] [Medline: 37360961]
18. Dogruel L, Joeckel S, Bowman ND. Choosing the right app: an exploratory perspective on heuristic decision processes for smartphone app selection. Mob Media Commun 2015 Jan;3(1):125-144. [doi: 10.1177/2050157914557509]
19. Wasil AR, Gillespie S, Schell T, et al. Estimating the real-world usage of mobile apps for mental health: development and application of two novel metrics. World Psychiatry 2021 Feb;20(1):137-138. [doi: 10.1002/wps.20827] [Medline: 33432761]
20. Torous JB, Chan SR, Gipson S, et al. A hierarchical framework for evaluation and informed decision making regarding smartphone apps for clinical care. Psychiatr Serv 2018 May 1;69(5):498-500. [doi: 10.1176/appi.ps.201700423] [Medline: 29446337]
21. Silva GRS, Canedo ED. Towards user-centric guidelines for chatbot conversational design. Int J Hum-Comput Interact 2024 Jan 17;40(2):98-120. [doi: 10.1080/10447318.2022.2118244]
22. Lee YK, Lee I, Shin M, et al. Chain of empathy: enhancing empathetic response of large language models based on psychotherapy models. arXiv. Preprint posted online on Sep 14, 2024. [doi: 10.48550/arXiv.2311.04915]
23. Li E, Kealy D, Aafjes-van Doorn K, et al. “It Felt Like I Was Being Tailored to the Treatment Rather Than the Treatment Being Tailored to Me”: Patient Experiences of Helpful and Unhelpful Psychotherapy. Psychother Res :1-15. [doi: 10.1080/10503307.2024.2360448]
24. American Psychological Association. Ethical principles of psychologists and code of conduct. 2024.
25. Liu S, Zheng C, Demasi O, et al. Towards emotional support dialog systems. arXiv. Preprint posted online on Jun 2, 2021. [doi: 10.48550/arXiv.2106.01144]
26. Wampold BE, Flückiger C. The alliance in mental health care: conceptualization, evidence and clinical applications. World Psychiatry 2023 Feb;22(1):25-41. [doi: 10.1002/wps.21035] [Medline: 36640398]
27. Chaszczewicz A, Shah R, Louie R. Multi-level feedback generation with large language models for empowering novice peer counselors. arXiv. Preprint posted online on Mar 21, 2024 URL: http://arxiv.org/abs/2403.15482 [accessed 2025-06-23]
28. Rheu M(, Dai Y(, Meng J, et al. When a chatbot disappoints you: expectancy violation in human-chatbot interaction in a social support. Communic Res 2024 Oct;51(7):782-814. [doi: 10.1177/00936502231221669]
29. Meng J, Rheu M(, Zhang Y, et al. Mediated social support for distress reduction: AI chatbots vs. human. Proc ACM Hum-Comput Interact 2023 Apr 14;7(CSCW1):1-25. [doi: 10.1145/3579505]
30. Heston TF. Evaluating risk progression in mental health chatbots using escalating prompts. medRxiv. Preprint posted online on Sep 12, 2023 URL: https://doi.org/10.1101/2023.09.10.23295321 [doi: 10.1101/2023.09.10.23295321]
31. Boswell JF, Hepner KA, Lysell K, et al. The need for a measurement-based care professional practice guideline. Psychotherapy (Chic) 2023 Mar;60(1):1-16. [doi: 10.1037/pst0000439] [Medline: 35771518]
32. Zhan H, Zheng A, Lee YK, et al. Large language models are capable of offering cognitive reappraisal, if guided. arXiv. Preprint posted online on Aug 8, 2024. [doi: 10.48550/arXiv.2404.01288]
33. Ramos G, Ponting C, Labao JP, et al. Considerations of diversity, equity, and inclusion in mental health apps: a scoping review of evaluation frameworks. Behav Res Ther 2021 Dec;147:103990. [doi: 10.1016/j.brat.2021.103990] [Medline: 34715396]
34. The White House. Notice and Explanation of the AI Bill of Rights. 2023. URL: https://bidenwhitehouse.archives.gov/ostp/ai-bill-of-rights [accessed 2025-06-23]
35. Dinan E, Logacheva V, Malykh V, et al. The second conversational intelligence challenge (ConvAI2). arXiv. Preprint posted online on Jan 31, 2019 URL: http://arxiv.org/abs/1902.00098 [doi: 10.1007/978-3-030-29135-8_7]
36. Zhang S, Dinan E, Urbanek J, et al. Personalizing dialogue agents: I have a dog, do you have pets too? arXiv. Preprint posted online on Sep 25, 2018. [doi: 10.48550/arXiv.1801.07243]
37. Zhdanava M, Pilon D, Ghelerter I, et al. The prevalence and national burden of treatment-resistant depression and major depressive disorder in the United States. J Clin Psychiatry 2021 Mar 16;82(2):20m13699. [doi: 10.4088/JCP.20m13699] [Medline: 33989464]
38. Hall EB, Mufson L. Interpersonal psychotherapy for depressed adolescents (IPT-A): a case illustration. J Clin Child Adolesc Psychol 2009 Jul;38(4):582-593. [doi: 10.1080/15374410902976338] [Medline: 20183644]
39. Jiménez Chafey MI, Bernal G, Rosselló J. Clinical case study: CBT for depression in a Puerto Rican adolescent: challenges and variability in treatment response. Depress Anxiety 2009;26(1):98-103. [doi: 10.1002/da.20457] [Medline: 18781640]
40. Young JE, Rygh JE, Weinberger AD, et al. Cognitive therapy for depression. In: Clinical Handbook of Psychological Disorders: A Step-by-Step Treatment Manual: Guilford Press; 2008:250-305.
41. OpenAI. Privacy policy. 2023. URL: https://openai.com/policies/privacy-policy/ [accessed 2025-05-26]
42. Online nation 2023 report. : Ofcom; 2023 URL: https://www.ofcom.org.uk/siteassets/resources/documents/research-and-data/online-research/online-nation/2023/online-nation-2023-report.pdf [accessed 2025-06-23]
43. Torous J, Firth J, Huckvale K, et al. The emerging imperative for a consensus approach toward the rating and clinical recommendation of mental health apps. J Nerv Ment Dis 2018 Aug;206(8):662-666. [doi: 10.1097/NMD.0000000000000864] [Medline: 30020203]

XSL•FO
RenderX

44. Alon N, Torous J. Current challenges for evaluating mobile health applications. J Am Med Inform Assoc 2023 Feb 16;30(3):617-624. [doi: 10.1093/jamia/ocac244] [Medline: 36484621]
45. Chiu YY, Sharma A, Lin IW, et al. A computational framework for behavioral assessment of LLM therapists. arXiv. Preprint posted online on Nov 28, 2024. [doi: 10.48550/ARXIV.2401.00820]
46. OpenAI. Usage Policies. 2024. URL: https://openai.com/policies/usage-policies/ [accessed 2025-05-26]
47. OECD AI principles overview. OECD AI. 2023. URL: https://oecd.ai/en/ai-principles [accessed 2025-05-26]
48. De Choudhury M, Pendse SR, Kumar N. Benefits and harms of large language models in digital mental health. PsyArXiv. Preprint posted online on Jun 20, 2024. [doi: 10.48550/ARXIV.2311.14693]
49. Ayoub NF, Balakrishnan K, Ayoub MS, et al. Inherent bias in large language models: a random sampling analysis. Mayo Clin Proc Digit Health 2024 Jun;2(2):186-191. [doi: 10.1016/j.mcpdig.2024.03.003] [Medline: 40207170]
50. Chow JCL, Li K. Ethical considerations in human-centered AI: advancing oncology chatbots through large language models. JMIR Bioinform Biotechnol 2024 Nov 6;5:e64406. [doi: 10.2196/64406] [Medline: 39321336]
51. Dankwa-Mullan I, Scheufele EL, Matheny ME, et al. A proposed framework on integrating health equity and racial justice into the artificial intelligence development lifecycle. J Health Care Poor Underserved 2021;32(2):300-317. [doi: 10.1353/hpu.2021.0065]
52. Kim W, Ryoo Y, Lee S, et al. Chatbot advertising as a double-edged sword: the roles of regulatory focus and privacy concerns. J Advert 2023 Aug 8;52(4):504-522. [doi: 10.1080/00913367.2022.2043795]
53. Ostwal T. Interactivity and proprietary data: marketers welcome ads in chatbots. ADWEEK. 2024. URL: https://www.adweek.com/media/why-marketers-welcome-ads-in-chatbots/ [accessed 2025-05-26]
54. Bipeta R. Legal and ethical aspects of mental health care. Indian J Psychol Med 2019;41(2):108-112. [doi: 10.4103/IJPSYM.IJPSYM_59_19] [Medline: 30983656]
55. Green SA, Bloch S. Working in a flawed mental health care system: an ethical challenge. Am J Psychiatry 2001 Sep;158(9):1378-1383. [doi: 10.1176/appi.ajp.158.9.1378] [Medline: 11532719]
56. Haltaufderheide J, Ranisch R. The ethics of ChatGPT in medicine and healthcare: a systematic review on Large Language Models (LLMs). NPJ Digit Med 2024 Jul 8;7(1):183. [doi: 10.1038/s41746-024-01157-x] [Medline: 38977771]
57. Lawrence HR, Schneider RA, Rubin SB, et al. The opportunities and risks of large language models in mental health. JMIR Ment Health 2024 Jul 29;11:e59479. [doi: 10.2196/59479] [Medline: 39105570]
58. Ullah E, Parwani A, Baig MM, et al. Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology - a recent scoping review. Diagn Pathol 2024 Feb 27;19(1):43. [doi: 10.1186/s13000-024-01464-7] [Medline: 38414074]
59. Hu B, Mao Y, Kim KJ. How social anxiety leads to problematic use of conversational AI: the roles of loneliness, rumination, and mind perception. Comput Human Behav 2023 Aug;145:107760. [doi: 10.1016/j.chb.2023.107760]
60. Olfson M, Guardino M, Struening E, et al. Barriers to the treatment of social anxiety. Am J Psychiatry 2000 Apr;157(4):521-527. [doi: 10.1176/appi.ajp.157.4.521] [Medline: 10739410]
61. Chow JCL, Sanders L, Li K. Impact of ChatGPT on medical chatbots as a disruptive technology. Front Artif Intell 2023;6:1166014. [doi: 10.3389/frai.2023.1166014] [Medline: 37091303]

Abbreviations

AI: artificial intelligence
CAPE: Conversational Agent for Psychotherapy Evaluation
CBT: cognitive behavioral therapy
***DSM-V*:** *Diagnostic and Statistical Manual of Mental Disorders* (Fifth Edition)
GenAI: generative artificial intelligence
IRR: interrater reliability
LLM: large language model
MDD: Major Depressive Disorder
mHealth: mobile health
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by A Mavragani; submitted 20.08.24; peer-reviewed by H Li, JCL Chow; revised version received 17.01.25; accepted 19.01.25; published 02.07.25.

Please cite as:
Sobowale K, Humphrey DK
Evaluating the Quality of Psychotherapy Conversational Agents: Framework Development and Cross-Sectional Study
JMIR Form Res 2025;9:e65605
URL: https://formative.jmir.org/2025/1/e65605
doi:10.2196/65605

XSL•FO
RenderX

© Kunmi Sobowale, Daniel Kevin Humphrey. Originally published in JMIR Formative Research (https://formative.jmir.org), 2.7.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Association of a Brief Computerized Cognitive Assessment With Cholinergic Neurotransmission: Assessment Validation Study

Mouna Attarha[1], PhD; Ana De Figueiredo Pelegrino[2], MSc; Lydia Ouellet[2], MSc; Paule-Joanne Toussaint[2], PhD; Sarah-Jane Grant[1], MA; Thomas Van Vleet[1], PhD; Etienne de Villers-Sidani[2], MDCM

[1]Posit Science Corporation, 160 Pine St Suite 200, San Francisco, CA, United States

[2]Department of Neurology and Neurosurgery, McGill University, Montreal, QC, Canada

Corresponding Author:
Mouna Attarha, PhD
Posit Science Corporation, 160 Pine St Suite 200, San Francisco, CA, United States

Abstract

Background: Computerized cognitive assessments are most often validated against standard neuropsychological measures with limited validation against biological indices of brain health.

Objective: This study aimed to evaluate whether a self-administered computerized cognitive assessment is associated with cholinergic neurotransmission using the vesicular acetylcholine transporter ligand [18F]fluoroethoxybenzovesamicol (FEOBV) and positron emission tomography (PET).

Methods: In a retrospective analysis, we report baseline data from the Improving Neurological Health in Aging via Neuroplasticity-Based Computerized Exercise (INHANCE) trial. This study provides normative data for healthy older adults aged 65 years and above. We evaluate the validity of the Double Decision cognitive assessment (from the BrainHQ assessment platform) by examining its association with tracer binding in the anterior cingulate cortex, as measured by FEOBV-PET. We also assess concurrent validity with neuropsychological performance using standardized measures of executive function and global cognition.

Results: The intent-to-treat population from the INHANCE trial analyzed in this study included 92 healthy adults with a mean age of 71.9 (SD 4.86, range 65‑83) years, the majority of whom were female (61/92, 66%), with an average of 16.45 (SD 3.40, range 9‑27) years of education. The Double Decision assessment is associated with FEOBV binding in the anterior cingulate cortex, explaining 8% of the variance, and was associated with neuropsychological performance measures. The assessment was sensitive to age and was not influenced by education level or gender. Psychometric properties supported its usability and the assessment showed an average completion time of 3 (SD 1.12) minutes.

Conclusions: We present the first brief, self-administered computerized cognitive assessment associated with cholinergic network health. This tool is scalable and accessible to individuals with an internet-connected device, offering a practical and cost-efficient approach to cognitive screening. The findings provide valuable insights into brain health, particularly for early detection of cognitive decline, and hold significant potential for broad applications across both clinical and nonclinical contexts.

Trial Registration: ClinicalTrials.gov NCT04149457; https://clinicaltrials.gov/study/NCT04149457

International Registered Report Identifier (IRRID): RR2-10.2196/59705

(JMIR Form Res 2025;9:e68374) doi:10.2196/68374

KEYWORDS

cholinergic function; acetylcholine; FEOBV-PET; web-based cognitive assessment; brain health; cognitive status; assessment accessibility; anterior cingulate; computerized cognitive assessment; cholinergic neurotransmission; assessment validation study; self-administered; vesicular acetylcholine transporter; VAChT; positron emission tomography; PET; retrospective analysis; older person; neurotransmission; [18F]fluoroethoxybenzovesamicol

Introduction

Cholinergic function is recognized as a critical marker of overall brain health. Previous research has demonstrated that the cholinergic system regulates synaptic plasticity and is positively associated with key aspects of cognitive performance including sensory processing, attention, learning and memory, and executive function [1-6]. In contrast, cholinergic dysfunction is increasingly recognized as a precursor to cognitive decline, manifesting in subtle memory impairments before the onset of a clinically evident neurodegenerative disease [7-10].

Impaired cholinergic signaling is linked to broader neuropathological cascades central to conditions such as

XSL•FO
RenderX

Alzheimer disease and related dementias [9,11,12]. Research indicates that it disrupts amyloid precursor protein metabolism, promoting amyloidogenic pathways that generate toxic amyloid-beta aggregates [11-18]. In addition, cholinergic decline is linked to heightened activity of glycogen synthase kinase-3, an enzyme involved in tau hyperphosphorylation and the subsequent formation of neurofibrillary tangles. These pathological processes fuel synaptic dysfunction, neuroinflammation, and neuronal loss, ultimately accelerating cognitive impairment. Conversely, activating cholinergic receptors demonstrates neuroprotective potential by steering amyloid precursor protein processing away from amyloidogenic pathways and mitigating tau pathology [11-18].

Given the cholinergic system's central role in mediating and modulating cognitive performance [5,6,19-21] and its influence on amyloid-beta and tau-driven neurodegeneration [10], it has become a therapeutic target in clinical research [7,17,18,22,23]. Cholinesterase inhibitors are commonly prescribed to mitigate cognitive impairment [24]. By increasing acetylcholine availability in the synaptic cleft, these treatments aim to improve cholinergic signaling and preserve cognitive performance [24]. The therapeutic significance of acetylcholine highlights its fundamental role in maintaining cholinergic network health and underscores the need for early screening and detection of cholinergic dysfunction to develop timely and targeted interventions against cognitive decline.

Noninvasive methods for measuring cholinergic function in vivo in humans, however, remain a significant challenge. Most current approaches rely on neuroimaging techniques such as positron emission tomography (PET), which uses tracers to assess cholinergic receptor binding. While these methods provide valuable insights, they are costly, require specialized expertise, and expose participants to radiation, limiting their use in large-scale studies or routine clinical practice. To our knowledge, there are no validated, scalable, behavioral proxies of cholinergic function, particularly those using a brief, self-administered computerized cognitive assessment model.

In this study, we analyzed baseline data from the intent-to-treat (n=92) from the Improving Neurological Health in Aging via Neuroplasticity-Based Computerized Exercise (INHANCE) trial that was designed to evaluate the effect of cognitive training on cholinergic function in healthy older adults aged 65 years and above with a baseline Montreal Cognitive Assessment (MoCA) between 23 and 30. We assessed whether a cognitive assessment (Double Decision on the BrainHQ assessment platform) correlates with tracer binding levels in vivo within a prespecified region of interest (ROI), the anterior cingulate cortex [25], using PET imaging with the radiotracer [18F]fluoroethoxybenzovesamicol (FEOBV) [26]. Aging is linked to notable reductions in FEOBV binding, with an estimated decline of 2.5% per decade observed in the anterior cingulate cortex [27]. This region plays a critical role in supporting selective attention, learning and memory, and executive function [28-31]. We also examined whether the Double Decision assessment was associated with conventional measures of cognitive performance, specifically the MoCA [32] for global cognition and the National Institutes of Health Executive Abilities: Measures and Instruments for Neurobehavioral Evaluation and Research (NIH EXAMINER) battery for executive function [33]. We hypothesized that performance on Double Decision would be associated with cholinergic signaling and cognition.

Methods

Setting, Design, and Sample

Participants were recruited near McGill University, Canada, where the FEOBV radiotracer is synthesized and administered. Participants were community-dwelling, healthy older adults aged 65 years and above with a MoCA total score between 23 and 30 (both inclusive). The data reported herein reflect baseline assessment data of the intent-to-treat (N=92) from the INHANCE randomized clinical trial (refer to published protocol for full details [25]). The study enrolled from July 2021 to December 2023 with the final follow-up June 2024.

Double Decision Assessment

The Double Decision assessment evaluates the visual speed of processing and selective attention [34]. This assessment is modeled upon the third subtest of the validated Useful Field of View (UFOV) assessment, which has established validity as a supervised in-clinic cognitive assessment and measure of driving safety [35-37]. Modifications include a modernized user interface with high-resolution color graphics and a guided flow to support self-administration (Figure 1). In this dual-task paradigm, participants discriminate a visual stimulus presented in the center of gaze while simultaneously locating a target in the peripheral visual field. The adaptive dimension is exposure duration (the length of time the stimuli remain visible on the screen). As the participant correctly responds to a trial, exposure duration decreases on subsequent trials, thereby making the task harder by requiring less time to process the visual display; conversely, as the participant responds incorrectly, exposure duration increases to make the task easier. Raw scores are defined as the exposure duration at approximately 80% criterion accuracy in milliseconds. The best possible raw score is 32 milliseconds, and the worst possible score is 3162 milliseconds.

XSL•FO
RenderX

Figure 1. The Double Decision computerized cognitive assessment. (**A**) In this dual-task paradigm, participants simultaneously saw a vehicle at the center of the screen and signs in the periphery with a brief exposure duration. (**B**) Participants were asked to identify which vehicle (of 2) were presented and (**C**) pinpoint the location of the Route 66 sign (of 8 possible peripheral locations). Correct responses for the sample trial shown in (**A**) are highlighted in (**B**) and (**C**) for demonstration purposes.

Ethical Considerations

The study was developed in accordance with the Declaration of Helsinki guidelines and was approved by the Western Institutional Review Board (IRB00000533) and Research Ethics Board of McGill University Health Centre (2020 - 6474). The radioligand [18F]FEOBV was approved by Health Canada (Control # 252085). All participants provided written informed consent (Multimedia Appendix 1).

Study data were recorded into a secure, web-based electronic case report form at the study site through the Longitudinal Online Research and Imaging System. This system meets relevant privacy and security standards for electronic trial data entry and storage, as well as the Health Insurance Portability and Accountability Act and Personal Information Protection and Electronic Documents Act standards for confidentiality and privacy. Following consent, each participant was assigned a standardized Participant Identification Number composed of digits to identify the study and digits to identify the participant. All electronic case report form data entry were deidentified, using the Participant Identification Number and not the participant's name.

Participants received CAD $40 (approximately US $30) for completing the baseline visit, which included the measures described below, on completion of the visit.

Measures

Following informed consent, participants completed a structured clinical interview, the Double Decision assessment, 2 validated neuropsychological assessments to evaluate cognition (NIH EXAMINER and MoCA), and FEOBV-PET to evaluate cholinergic neurotransmission. Blinded staff members conducted the assessment administration and scoring.

A structured clinical interview (20 min, in person) collected demographic data (eg, age, gender, education).

Double Decision assessment (8 min, in person, computerized) evaluated visual speed of processing. The assessment was administered on a computer and scored automatically, with lower scores indicating better performance.

XSL•FO
RenderX

FEOBV-PET imaging (acquired 180 min post injection for 30 min, in person, scanner) evaluated cholinergic neurotransmission using mean standard uptake value ratios (SUVRs) for the primary ROI, the anterior cingulate, defined by the Hammers atlas [38]. Details are provided in the Imaging Acquisition and Processing subsubsection below.

NIH EXAMINER (20 min, in person, computerized) assessed executive function using a standard battery shown to have good reliability and validity. Performance was measured using the executive composite score [39] generated by the assessment program, with higher scores indicating better performance.

MoCA (10 min, in person, pen and paper) assessed global cognition. The MoCA includes visuospatial or executive functioning, naming, memory, attention, language, abstraction, delayed recall, and orientation to time and place for a total of 30 points, with higher scores indicating better performance.

Imaging Acquisition and Processing

All participants underwent a structural T1-weighted magnetic resonance imaging (MRI) scan (3T Siemens Prisma) using the 3D magnetization-prepared rapid gradient echo sequence, followed by a [18F]FEOBV-PET scan using the Siemens High-Resolution Research Tomograph at the McConnell Brain Imaging Centre of the Montreal Neurological Institute-Hospital. The structural MRI was acquired during the imaging portion of the baseline visit to coregister the PET data [25].

Concurrent T1-weighted magnetization-prepared rapid gradient echo images were acquired with the related acquisition parameters being echo time (2.98 ms), repetition time (2300 ms), inversion time (900 s), flip angle (9°), isotropic resolution (1 mm), and image dimensions (192 mm×256 mm×256 mm).

Participants were positioned lying on their back for the PET imaging session and received a slow bolus intravenous injection of [18F]FEOBV with radioactivity doses ranging from 350 to 400 MBq [40], corresponding to 8.05 mSv-9.2 mSv, via a fine needle-catheter inserted into an arm vein. PET data acquisition started 180 minutes after injection, for a duration of 30 minutes, divided into 6 frames of 5 minutes each.

A transmission scan of 5 minutes was conducted with a rotating point source of [137Cs] for the Siemens High-Resolution Research Tomograph PET images in order to perform attenuation correction. PET images were reconstructed using an ordinary Poisson-ordered subset expectation maximization algorithm (10 iterations, 16 subsets) with resolution recovery, correcting for scatter, randoms, attenuation, decay, and dead time. Motion correction was applied, and time-averaged PET data (6×5-min frames) were used to create a static image. The final reconstruction was performed on a 256×256×207 matrix (voxel size: 1.22 mm^3, spatial resolution: 2.3 mm full-width at half-maximum) with no postreconstruction smoothing or zoom.

PET preprocessing was performed using SPM12 in MATLAB. MRIs were segmented (gray matter, white matter, and cerebrospinal fluid), bias-corrected, and spatially normalized to the Montreal Neurological Institute (MNI) 152 asymmetrical template (MNI152 ONLine 2009 cAsym template) using Diffeomorphic Anatomical Registration Through Exponentiated Lie Algebra, with identical transformations applied to PET images. PET images were aligned to each participant's MRI, and Müller-Gärtner partial volume correction was applied using the PETPVE toolbox. A 6 mm full-width at half-maximum Gaussian smoothing kernel was used to reduce noise.

SUVRs were used to quantify FEOBV binding. SUVRs were computed for spatially normalized PET images using a white matter mask as the reference region [41,42]. The Hammers atlas (MNI space) [38] was used to extract the SUVR in the target region (anterior cingulate cortex). A binary white matter mask was created and then eroded to reduce contamination from neighboring regions. The final eroded white matter mask was applied to the PET images to calculate mean tracer uptake for SUVR calculations.

Analyses

Overview

Recruitment spanned from July 2021 to December 2023. The final follow-up visit was on June 7, 2024, and the study team was unblinded on June 14, 2024, after database lock. For all analyses, a $P<.05$ determined statistical significance with $P<.10$ reported as trending per the published statistical analysis plan [25].

Demographic Characteristics

Participants reported their age, gender, and educational level and completed the NIH EXAMINER and MoCA. Arithmetic means, SDs, and ranges are provided for continuous variables (age, education, NIH EXAMINER, and MoCA). For gender, we tallied the number of participants who selected a response option to each question and divided it by the total number of participants (N=92). We included both raw numbers and percentages for gender.

Assessment Characteristics

To evaluate usability, we reported general descriptive statistics (arithmetic mean and SD), distribution characteristics (skew and kurtosis), and psychometric properties, including the performance histogram of the proportion of participants achieving each assessment score, the mean number and SD of minutes spent in the assessment, and the percentage of participants obtaining the numerically lowest or highest assessment score to indicate the frequency of ceiling and floor effects.

Associations Between Demographics, FEOBV-PET, and Neuropsychological Tests

To establish associations between performance on the Double Decision assessment with demographic variables, we used the Spearman ρ for age and education, and the Wilcoxon rank sum for gender.

We used linear regression and Pearson r to establish the relationship between performance and FEOBV-PET SUVR within the anterior cingulate. PET image processing details are provided in the published protocol [25]. To establish the association between performance and standard neuropsychological measures, we used Pearson r for NIH EXAMINER and Spearman ρ for the MoCA.

XSL•FO
RenderX

Results

Demographic Characteristics

Participants (N=92) had a mean age of 71.90 (SD 4.86, range: 65 - 83) years. Mean years of education was 16.45 (SD 3.40, range: 9 - 27) years. A total of 61 participants of 92 identified as female (66%) and 31 participants of 92 identified as male (34%). The baseline NIH EXAMINER executive composite mean was 0.43 (SD 0.57, range: −1.50 to 1.52) and the baseline MoCA total score mean was 26.17 (SD 1.84, range: 23 - 30) (Table 1). A full characterization of the intent-to-treat is presented by Pelegrino et al [43].

Table . Baseline characteristics of the intent-to-treat population (N=92).

Characteristic	Participant
Age (years), mean (SD)	71.9 (4.9)
Education (years), mean (SD)	16 (3.4)
Gender, n (%)	
Women	61 (66)
Men	31 (34)
Race, n (%)	
White	88 (96)
Asian	2 (2)
Black or African American	1 (1)
American Indian or Alaska Native	0 (0)
Native Hawaiian or other Pacific Islander	0 (0)
More than 1 race	0 (0)
Unknown or not reported	1 (1)
Ethnicity, n (%)	
Not Hispanic or Latino	91 (99)
Hispanic or Latino	0 (0)
Unknown or not reported	1 (1)
Double Decision (ms), raw score (SD)	1476.70 (813.06)
Anterior Cingulate, mean standard uptake value ratio (SD)	1.90 (0.19)
MoCA[a], total score (SD)	26.2 (1.8)
NIH EXAMINER[b], executive composite score (SD)	0.43 (0.57)

[a]MoCA: Montreal Cognitive Assessment.

[b]NIH EXAMINER: National Institutes of Health Executive Abilities: Measures and Instruments for Neurobehavioral Evaluation and Research.

Assessment Characteristics

The performance distribution of Double Decision assessment scores is presented in Figure 2A. The mean score was 1476.70 (SD 813.06) milliseconds, with a skew of 0.84 and a kurtosis of −0.17. The minimum (best) score achieved among the 92 participants was 269 milliseconds (1.09% of participants), and the maximum (worst) score achieved was 3162 milliseconds (11.96% of participants). Participants took a mean of 3.17 (SD 1.12) minutes to complete the assessment.

XSL•FO
RenderX

Figure 2. Outcome measures. (**A**) Performance histogram of Double Decision scores across participants (N=92). Association of Double Decision with demographic characteristics including (**A**) age, (**B**) years of education, and (**C**) gender. (**E**) Association of Double Decision with [18F]fluoroethoxybenzovesamicol–positron emission tomography standard uptake value ratios within the anterior cingulate cortex. (**F**) Association of Double Decision with the NIH EXAMINER executive composite, a standard validated measure of executive function. (**G**) Association of Double Decision with the MoCA, a standard validated measure of global cognition. MoCA: Montreal Cognitive Assessment; NIH EXAMINER: National Institutes of Health Executive Abilities: Measures and Instruments for Neurobehavioral Evaluation and Research; SUVR: standard uptake value ratio.

XSL•FO
RenderX

Associations Between Demographics, FEOBV-PET, and Neuropsychological Tests

There was a significant association of Double Decision with age (r=0.42; P<.001) and nonsignificant associations with education (ρ=−0.01; P=.95) and gender (P=.94) (Figure 2B-2D).

There was a significant negative association between Double Decision and FEOBV-PET SUVR (r=−0.28; P=.007) within the anterior cingulate cortex (Figure 2E). Linear regression showed that scores on Double Decision correlated with tracer binding (SUVR) using the following equation ($F_{1,87}$=7.54; P=.007) and explained 8% of the variation in SUVR (R^2=0.08).

$$(\text{Double Decision score in ms})*(-0.0000685)+2.0039=\text{SUVR}$$

There was a significant negative association between Double Decision and the NIH EXAMINER executive composite score (r=−0.36, P<.001) and a nonsignificant negative association between Double Decision and the MoCA total score (ρ=−0.19; P=.08) (Figure 2F-2G).

Discussion

Principal Findings

Double Decision is the first brief, self-administrable computerized cognitive assessment [44] validated against cholinergic binding in the anterior cingulate cortex using FEOBV-PET. Performance on Double Decision was negatively associated with SUVR, indicating that better (ie, lower) scores were correlated with higher cholinergic binding levels. This finding aligns with previous research showing a significant relationship between higher FEOBV-PET SUVR and cognitive performance, as measured by staff-administered validated computerized tools like the NIH EXAMINER [43], as well as traditional pen-and-paper measures [45,46].

Double Decision was further validated against standardized neuropsychological measures, showing a significant negative association with the NIH EXAMINER and a trending negative association with the MoCA. These results indicate that better (faster) performance (ie, lower scores) on Double Decision correlated with higher scores in executive function and global cognition. The assessment was brief, taking an average of 3 (SD 1.12) minutes to complete, and demonstrated good usability with reasonable descriptive and psychometric properties. It was sensitive to age within the narrow age band of 65 - 83 years and was not influenced by demographic factors such as years of education or gender. The results support the adoption of this scalable form of biomarker-informed cognitive assessment available to individuals with an internet-connected device.

The Double Decision assessment is a gamified version of the third subtest from the well-known UFOV assessment. Meta-analyses have shown significant associations between UFOV and standard neuropsychological measures of general cognition, processing speed, attention, memory, and executive function [37]. UFOV has also been validated as a measure of at-fault motor vehicle crash risk [35,36]. The results of this study suggest that Double Decision and UFOV may assess biological markers of brain health that support a wide range of higher-order cognitive functions and real-world abilities.

Significance and Implementation

This study represents an important step toward validating a noninvasive assessment of cholinergic function in humans. This brief, self-administrable assessment tool offers a scalable approach to cognitive screening, requiring only an internet-connected device. Broad accessibility makes Double Decision valuable for widespread use in community-based health initiatives, primary care settings, and clinical trials. It may be particularly beneficial for individuals residing in remote or underserved areas where access to traditional neuroimaging or comprehensive neuropsychological assessment may be limited, impractical, or cost-prohibitive.

The documented involvement of cholinergic dysfunction in neurodegenerative diseases [11], combined with the widespread use of cholinesterase inhibitors as a treatment for mild cognitive impairment, Alzheimer disease, and related dementias [24], highlights the essential role of acetylcholine as a therapeutic target to preserve the neurochemical integrity of the brain. The ability to digitally assess disruptions in cholinergic network health therefore represents an unmet need in the early detection and management of brain health and cognitive decline. This assessment could serve as an initial community-based screening tool for individuals at risk of cognitive decline, including adults experiencing subtle memory lapses, individuals with a family history of dementia, or those with other known risk factors such as cardiovascular disease or metabolic disorders. It may be used to monitor cognitive trajectories over time, identifying individuals who may benefit from further neuropsychological evaluation or early pharmacological or lifestyle interventions aimed at preserving cognition. In large-scale epidemiological studies, such a tool may facilitate the identification of population-level trends in cognitive health, enabling researchers to assess how various factors, such as genetics, lifestyle, or environmental exposures, contribute to cognitive aging. In clinical trials of cognitive impairment, it may serve as a measure to evaluate the efficacy of interventions aimed and altering cholinergic function.

By bridging a gap between neuroimaging biomarkers and widely deployable cognitive assessments, this study contributes to the field of digital biomarkers, offering a way to integrate remote cognitive screening into routine care. Ultimately, this approach may improve cognitive health monitoring, allowing for earlier identification of individuals at risk for cognitive decline and supporting the development of targeted intervention strategies to improve long-term brain health outcomes.

Limitations

The study population had limited demographic variability. A total of 88 of 92 (96%) participants of the intent-to-treat identified as White; therefore, the results may not be generalizable to racial and ethnic minority groups [43]. Due to the study's inclusion criteria, the age band (65 y and older) and the baseline MoCA total score (23-30) were range-limited.

Strengths

This study represents the largest FEOBV-PET investigation to date, employing a rigorous design with stringent inclusion

XSL•FO
RenderX

criteria and guided by a prespecified statistical analysis plan focused on a predefined ROI.

Conclusion

Double Decision is the first brief and usable computerized cognitive assessment associated with cholinergic network health. This tool is scalable and accessible to any individual with an internet-connected device, offering a practical and cost-efficient approach to cognitive screening. The findings offer important insights into brain health, and may support the early detection of cognitive decline by serving as a behavioral proxy of forebrain cholinergic function. Future research should explore whether individuals with low performance on this assessment are at a higher risk of developing neurological or neurodegenerative disorders.

Acknowledgments

Research reported in this publication was supported by the National Institute on Aging of the National Institutes of Health under awards R44AG039965 and 3R44AG039965-06S1, awarded to MA and TVV. We would like to thank the members of McGill's Clinical Research Unit and the software engineers and quality assurance team at Posit Science. This content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Data Availability

On May 31, 2026, the deidentified participant data from INHANCE (Improving Neurological Health in Aging via Neuroplasticity-Based Computerized Exercise) will be made available to verified academic and industry researchers. Investigators can conduct both confirmatory and exploratory analyses using LORIS (Longitudinal Online Research and Imaging System), a freely accessible, open-source data archive with provenance-sharing features. The dataset will be accessible at the inhance.loris website.

Authors' Contributions

MA and TVV are principal investigators. EdV-S is the site principal investigator. MA, EdV-S, and TVV were responsible for the concept and the design of the study. MA wrote the first version of the manuscript. PJT provided expertise and guidance for the positron emission tomography (PET) data analysis. ADFP conducted the PET image processing and PET analyses. MA conducted the remaining analyses reported herein. SJG was the clinical trial manager at Posit Science and LO administered the assessments at McGill. MA, EdV-S, and ADFG defined the data analysis plan. All authors read, edited, and approved the final manuscript.

Conflicts of Interest

MA, TVV, and SJG are employees at Posit Science, the company that develops BrainHQ's brain training and assessment programs. They are shareholders of Posit Science stock. EdV-S, ADFP, PJT, and LO declare no competing interests.

Multimedia Appendix 1
Consent form.
[PDF File, 426 KB - formative_v9i1e68374_app1.pdf]

References

1. Picciotto MR, Higley MJ, Mineur YS. Acetylcholine as a neuromodulator: cholinergic signaling shapes nervous system function and behavior. Neuron 2012 Oct 4;76(1):116-129. [doi: 10.1016/j.neuron.2012.08.036] [Medline: 23040810]
2. Micheau J, Marighetto A. Acetylcholine and memory: a long, complex and chaotic but still living relationship. Behav Brain Res 2011 Aug 10;221(2):424-429. [doi: 10.1016/j.bbr.2010.11.052] [Medline: 21130809]
3. Himmelheber AM, Sarter M, Bruno JP. Increases in cortical acetylcholine release during sustained attention performance in rats. Brain Res Cogn Brain Res 2000 Jun;9(3):313-325. [doi: 10.1016/s0926-6410(00)00012-4] [Medline: 10808142]
4. Parikh V, Bangasser DA. Cholinergic signaling dynamics and cognitive control of attention. Curr Top Behav Neurosci 2020;45:71-87. [doi: 10.1007/7854_2020_133] [Medline: 32447715]
5. Logue SF, Gould TJ. The neural and genetic basis of executive function: attention, cognitive flexibility, and response inhibition. Pharmacol Biochem Behav 2014 Aug;123:45-54. [doi: 10.1016/j.pbb.2013.08.007] [Medline: 23978501]
6. Kilgard MP, Merzenich MM. Cortical map reorganization enabled by nucleus basalis activity. Science 1998 Mar 13;279(5357):1714-1718. [doi: 10.1126/science.279.5357.1714] [Medline: 9497289]
7. Hampel H, Mesulam MM, Cuello AC, et al. The cholinergic system in the pathophysiology and treatment of Alzheimer's disease. Brain (Bacau) 2018 Jul 1;141(7):1917-1933. [doi: 10.1093/brain/awy132] [Medline: 29850777]
8. Muir JL. Acetylcholine, aging, and Alzheimer's disease. Pharmacol Biochem Behav 1997 Apr;56(4):687-696. [doi: 10.1016/s0091-3057(96)00431-5] [Medline: 9130295]

XSL•FO
RenderX

9. Singh S, Agrawal N, Goyal A. Role of Alpha-7-Nicotinic Acetylcholine Receptor in Alzheimer's Disease. CNSNDDT 2024 Mar;23(3):384-394. [doi: 10.2174/1871527322666230627123426]
10. Hampel H, Hardy J, Blennow K, et al. The amyloid-β pathway in Alzheimer's disease. Mol Psychiatry 2021 Oct;26(10):5481-5503. [doi: 10.1038/s41380-021-01249-0] [Medline: 34456336]
11. Mesulam M. The cholinergic lesion of Alzheimer's disease: pivotal factor or side show? Learn Mem 2004;11(1):43-49. [doi: 10.1101/lm.69204] [Medline: 14747516]
12. Schliebs R, Arendt T. The significance of the cholinergic system in the brain during aging and in Alzheimer's disease. J Neural Transm (Vienna) 2006 Nov;113(11):1625-1644. [doi: 10.1007/s00702-006-0579-2] [Medline: 17039298]
13. Schliebs R, Arendt T. The cholinergic system in aging and neuronal degeneration. Behav Brain Res 2011 Aug 10;221(2):555-563. [doi: 10.1016/j.bbr.2010.11.058] [Medline: 21145918]
14. Orlando IF, Shine JM, Robbins TW, Rowe JB, O'Callaghan C. Noradrenergic and cholinergic systems take centre stage in neuropsychiatric diseases of ageing. Neurosci Biobehav Rev 2023 Jun;149:105167. [doi: 10.1016/j.neubiorev.2023.105167] [Medline: 37054802]
15. Berry AS, Harrison TM. New perspectives on the basal forebrain cholinergic system in Alzheimer's disease. Neurosci Biobehav Rev 2023 Jul;150:105192. [doi: 10.1016/j.neubiorev.2023.105192] [Medline: 37086935]
16. Chaves-Coira I, García-Magro N, Zegarra-Valdivia J, Torres-Alemán I, Núñez Á. Cognitive deficits in aging related to changes in basal forebrain neuronal activity. Cells 2023 May 25;12(11):1477. [doi: 10.3390/cells12111477] [Medline: 37296598]
17. Shanbhag NM, Padmanabhan JL, Zhang Z, et al. An acetylcholine M1 receptor-positive allosteric modulator (TAK-071) in Parkinson disease with cognitive impairment: a phase 2 randomized clinical trial. JAMA Neurol 2025 Feb 1;82(2):152-159. [doi: 10.1001/jamaneurol.2024.4519] [Medline: 39761063]
18. Anand P, Singh B. A review on cholinesterase inhibitors for Alzheimer's disease. Arch Pharm Res 2013 Apr;36(4):375-399. [doi: 10.1007/s12272-013-0036-3] [Medline: 23435942]
19. Kilgard MP, Merzenich MM. Plasticity of temporal information processing in the primary auditory cortex. Nat Neurosci 1998 Dec;1(8):727-731. [doi: 10.1038/3729] [Medline: 10196590]
20. Kunnath AJ, Gifford RH, Wallace MT. Cholinergic modulation of sensory perception and plasticity. Neuroscience & Biobehavioral Reviews 2023 Sep;152:105323. [doi: 10.1016/j.neubiorev.2023.105323]
21. Hasselmo ME, Sarter M. Modes and models of forebrain cholinergic neuromodulation of cognition. Neuropsychopharmacology 2011 Jan;36(1):52-73. [doi: 10.1038/npp.2010.104] [Medline: 20668433]
22. Kaul I, Sawchak S, Correll CU, et al. Efficacy and safety of the muscarinic receptor agonist KarXT (xanomeline-trospium) in schizophrenia (EMERGENT-2) in the USA: results from a randomised, double-blind, placebo-controlled, flexible-dose phase 3 trial. Lancet 2024 Jan 13;403(10422):160-170. [doi: 10.1016/S0140-6736(23)02190-6] [Medline: 38104575]
23. Brannan SK, Sawchak S, Miller AC, Lieberman JA, Paul SM, Breier A. Muscarinic cholinergic receptor agonist and peripheral antagonist for schizophrenia. N Engl J Med 2021 Feb 25;384(8):717-726. [doi: 10.1056/NEJMoa2017015] [Medline: 33626254]
24. Rockwood K. Size of the treatment effect on cognition of cholinesterase inhibition in Alzheimer's disease. J Neurol Neurosurg Psychiatry 2004 May;75(5):677-685. [doi: 10.1136/jnnp.2003.029074] [Medline: 15090558]
25. Attarha M, de Figueiredo Pelegrino AC, Toussaint PJ, Grant SJ, Van Vleet T, de Villers-Sidani E. Improving neurological health in aging via neuroplasticity-based computerized exercise: protocol for a randomized controlled trial. JMIR Res Protoc 2024 Aug 8;13(1):e59705. [doi: 10.2196/59705] [Medline: 39116435]
26. Van Vleet TM, de Villers-Sidani E, Cote J, Merzenich M, Rosa-Neto P, Kang MS. P2‐005: neuroplasticity‐based visual attention training and the expression of acetylcholine in healthy older adults. Alzheimer's & Dementia 2018 Jul;14(7S_Part_12). [doi: 10.1016/j.jalz.2018.06.689]
27. Albin RL, Bohnen NI, Muller M, et al. Regional vesicular acetylcholine transporter distribution in human brain: A [18 F]fluoroethoxybenzovesamicol positron emission tomography study. J Comp Neurol 2018 Dec 1;526(17):2884-2897. [doi: 10.1002/cne.24541] [Medline: 30255936]
28. Martinelli P, Sperduti M, Devauchelle AD, et al. Age-related changes in the functional network underlying specific and general autobiographical memory retrieval: a pivotal role for the anterior cingulate cortex. PLOS ONE 2013;8(12):e82385. [doi: 10.1371/journal.pone.0082385] [Medline: 24367516]
29. Pardo JV, Lee JT, Sheikh SA, et al. Where the brain grows old: decline in anterior cingulate and medial prefrontal function with normal aging. Neuroimage 2007 Apr 15;35(3):1231-1237. [doi: 10.1016/j.neuroimage.2006.12.044] [Medline: 17321756]
30. Pardo JV, Nyabwari SM, Lee JT, Alzheimer's Disease Neuroimaging Initiative. Aging-related hypometabolism in the anterior cingulate cortex of cognitively intact, amyloid-negative seniors at rest mediates the relationship between age and executive function but not memory. Cereb Cortex Commun 2020;1(1):tgaa020. [doi: 10.1093/texcom/tgaa020] [Medline: 34296097]
31. Pezzoli S, Giorgio J, Martersteck A, Dobyns L, Harrison TM, Jagust WJ. Successful cognitive aging is associated with thicker anterior cingulate cortex and lower tau deposition compared to typical aging. Alzheimers Dement 2024 Jan;20(1):341-355. [doi: 10.1002/alz.13438] [Medline: 37614157]

XSL•FO
RenderX

32. Nasreddine ZS, Phillips NA, Bédirian V, et al. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. J Am Geriatr Soc 2005 Apr;53(4):695-699. [doi: 10.1111/j.1532-5415.2005.53221.x] [Medline: 15817019]
33. Kramer JH, Mungas D, Possin KL, et al. NIH EXAMINER: conceptualization and development of an executive function battery. J Int Neuropsychol Soc 2014 Jan;20(1):11-19. [doi: 10.1017/S1355617713001094] [Medline: 24103232]
34. Jobe JB, Smith DM, Ball K, et al. ACTIVE: a cognitive intervention trial to promote independence in older adults. Control Clin Trials 2001 Aug;22(4):453-479. [doi: 10.1016/s0197-2456(01)00139-8] [Medline: 11514044]
35. Ball KK, Roenker DL, Wadley VG, et al. Can high-risk older drivers be identified through performance-based measures in a Department of Motor Vehicles setting? J Am Geriatr Soc 2006 Jan;54(1):77-84. [doi: 10.1111/j.1532-5415.2005.00568.x] [Medline: 16420201]
36. Owsley C, Ball K, McGwin G Jr, et al. Visual processing impairment and risk of motor vehicle crash among older adults. JAMA 1998 Apr 8;279(14):1083-1088. [doi: 10.1001/jama.279.14.1083] [Medline: 9546567]
37. Woutersen K, Guadron L, van den Berg AV, Boonstra FN, Theelen T, Goossens J. A meta-analysis of perceptual and cognitive functions involved in useful-field-of-view test performance. J Vis 2017 Dec 1;17(14):11-11. [doi: 10.1167/17.14.11] [Medline: 29270605]
38. Hammers A, Allom R, Koepp MJ, et al. Three-dimensional maximum probability atlas of the human brain, with particular reference to the temporal lobe. Hum Brain Mapp 2003 Aug;19(4):224-247. [doi: 10.1002/hbm.10123] [Medline: 12874777]
39. Possin KL, LaMarre AK, Wood KA, Mungas DM, Kramer JH. Ecological validity and neuroanatomical correlates of the NIH EXAMINER executive composite score. J Int Neuropsychol Soc 2014 Jan;20(1):20-28. [doi: 10.1017/S1355617713000611] [Medline: 23764015]
40. Petrou M, Frey KA, Kilbourn MR, et al. In vivo imaging of human cholinergic nerve terminals with (-)-5-(18)F-fluoroethoxybenzovesamicol: biodistribution, dosimetry, and tracer kinetic analyses. J Nucl Med 2014 Mar;55(3):396-404. [doi: 10.2967/jnumed.113.124792] [Medline: 24481024]
41. Nejad-Davarani S, Koeppe RA, Albin RL, Frey KA, Müller M, Bohnen NI. Quantification of brain cholinergic denervation in dementia with Lewy bodies using PET imaging with [18F]-FEOBV. Mol Psychiatry 2019 Mar;24(3):322-327. [doi: 10.1038/s41380-018-0130-5] [Medline: 30082840]
42. Aghourian M, Legault-Denis C, Soucy JP, et al. Quantification of brain cholinergic denervation in Alzheimer's disease using PET imaging with [18F]-FEOBV. Mol Psychiatry 2017 Nov;22(11):1531-1538. [doi: 10.1038/mp.2017.183] [Medline: 28894304]
43. Pelegrino ADF, Attarha M, Toussaint PJ, et al. Cholinergic neurotransmission in the anterior cingulate cortex is associated with cognitive performance in healthy older adults: baseline characteristics of the Improving Neurological Health in Aging via Neuroplasticity-based Computerized Exercise (INHANCE) trial. Neuroimage Rep 2025 Mar;5(1):100234. [doi: 10.1016/j.ynirp.2025.100234] [Medline: 40191405]
44. Attarha M, Mahncke H, Merzenich M. The real-world usability, feasibility, and performance distributions of deploying a digital toolbox of computerized assessments to remotely evaluate brain health: development and usability study. JMIR Form Res 2024 May 13;8(1):e53623. [doi: 10.2196/53623] [Medline: 38739916]
45. Xia Y, Eeles E, Fripp J, et al. Reduced cortical cholinergic innervation measured using [18F]-FEOBV PET imaging correlates with cognitive decline in mild cognitive impairment. Neuroimage Clin 2022;34:102992. [doi: 10.1016/j.nicl.2022.102992] [Medline: 35344804]
46. van der Zee S, Müller M, Kanel P, van Laar T, Bohnen NI. Cholinergic denervation patterns across cognitive domains in Parkinson's disease. Mov Disord 2021 Mar;36(3):642-650. [doi: 10.1002/mds.28360] [Medline: 33137238]

Abbreviations

FEOBV: fluoroethoxybenzovesamicol
INHANCE: Improving Neurological Health in Aging via Neuroplasticity-Based Computerized Exercise
MNI: Montreal Neurological Institute
MoCA: Montreal Cognitive Assessment
MRI: magnetic resonance imaging
NIH EXAMINER: National Institutes of Health Executive Abilities: Measures and Instruments for Neurobehavioral Evaluation and Research
PET: positron emission tomography
ROI: region of interest
SUVR: standard uptake value ratio
UFOV: Useful Field of View

XSL•FO
RenderX

Edited by A Mavragani; submitted 04.11.24; peer-reviewed by B Tosti, JF Fu; revised version received 16.05.25; accepted 28.05.25; published 07.07.25.

Please cite as:
Attarha M, De Figueiredo Pelegrino A, Ouellet L, Toussaint PJ, Grant SJ, Van Vleet T, de Villers-Sidani E
Association of a Brief Computerized Cognitive Assessment With Cholinergic Neurotransmission: Assessment Validation Study
JMIR Form Res 2025;9:e68374
URL: https://formative.jmir.org/2025/1/e68374
doi:10.2196/68374

© Mouna Attarha, Ana De Figueiredo Pelegrino, Lydia Ouellet, Paule-Joanne Toussaint, Sarah-Jane Grant, Thomas Van Vleet, Etienne de Villers-Sidani. Originally published in JMIR Formative Research (https://formative.jmir.org), 7.7.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Association of Perceived Xingfu With Health-Related and Socioeconomic Factors Among Hong Kong Chinese Adults: Cross-Sectional Study Using a Novel Single-Item Tool

Katherine Y P Sze[1], MPH, MSc; Sai Yin Ho[2], PhD; Agnes Yuen Kwan Lai[3], PhD; Jing Jia[1], PhD; Heng Xu[4], PhD; Shirley Man Man Sit[2], MPhil; Tai Hing Lam[2], MD; Man Ping Wang[1], PhD

[1]School of Nursing, University of Hong Kong, 5/F, HKUMed Academic Building, 3 Sassoon Road, Pokfulam, Hong Kong, China (Hong Kong)

[2]School of Public Health, University of Hong Kong, Hong Kong, China (Hong Kong)

[3]School of Nursing and Health Studies, Hong Kong Metropolitan University, Hong Kong, China (Hong Kong)

[4]Department of Social and Behavioural Sciences, Department of Social and Behavioural Sciences, City University of Hong Kong, Hong Kong, China (Hong Kong)

Corresponding Author:
Man Ping Wang, PhD
School of Nursing, University of Hong Kong, 5/F, HKUMed Academic Building, 3 Sassoon Road, Pokfulam, Hong Kong, China (Hong Kong)

Abstract

Background: Xingfu (幸福), a Chinese term, holds meanings that transcend Western concepts of happiness; it is modern and highly valued in China. Despite its centrality to China's national discourse on xingfu, there are no validated tools for measuring perceived xingfu, particularly in Hong Kong's unique sociopolitical context. Postpandemic recovery efforts and widening socioeconomic disparities in Hong Kong highlight the urgency of understanding indicators such as perceived xingfu.

Objective: This study aimed to develop and validate the first single-item measure of perceived xingfu and examine its association with health-related and socioeconomic factors among Hong Kong Chinese adults, addressing gaps in culturally tailored assessment.

Methods: Our cross-sectional online survey included 5070 Hong Kong Chinese adults in 2023. Perceived xingfu was measured using a novel, single-item, 11-point scale from 0 to 10, with higher scores indicating better perceived xingfu. Two-week test-retest showed high reliability (intraclass correlation coefficient of 0.78). We used regression models to analyze associations between perceived xingfu score and high perceived xingfu (defined as ≥7) with mutually adjusted study variables; all estimates were weighted based on 2022 Hong Kong population data.

Results: The mean perceived xingfu score was 6.3 (SD 2.2). Perceived xingfu score was associated with happiness (r=0.85), perceived mental health (r=0.65), and adversity coping capability (r=0.50) and negatively associated with perceived stress (r=−0.56), past 7-days loneliness (r=−0.52), anxiety symptoms (r=−0.45), and depressive symptoms (r=−0.52). Female sex (β=0.69, adjusted odds ratio [aOR] 2.11), older age (β=0.46, aOR 2.67), having post-secondary education or above (β=0.19, aOR 1.35), higher monthly household income (≥ HKD60,000: β=0.99, aOR 3.04), living in owned properties (β=0.27, aOR 1.57), being retired (β=0.56, aOR 1.18), and excellent versus poor self-rated health (β=3.84, aOR 40.72) were associated with higher perceived xingfu score or high perceived xingfu (all Ps or P_{trend}<0.001).

Conclusions: This study pioneers the perceived xingfu measurement in a Chinese population using a concise, validated tool. Significant socioeconomic disparities and health associations highlight perceived xingfu's relevance to policy priorities, including equitable resource allocation and health support. Our single-item perceived xingfu tool offers practical utility for population surveillance and cross-cultural comparisons. Future research should explore longitudinal trends and integrate perceived xingfu into public health frameworks.

(JMIR Form Res 2025;9:e73350) doi:10.2196/73350

KEYWORDS

perceived xingfu (幸福感); health-related factors; socioeconomic factors; Hong Kong; cross-sectional study

Introduction

The meanings of xingfu are perceived as transcending Western concepts or constructs such as happiness, mental health, well-being, quality of life, and life satisfaction [1]. The term xingfu, in Chinese (幸福), has been commonly used by intellectuals and politicians since the early years of modern China. Multimedia Appendix 1 shows the cultural relevance

XSL•FO
RenderX

and development of the concept of xingfu. Xingfu is more complex than happiness and belongs to a higher level of human and political pursuit [2]. Since the establishment of the People's Republic of China in 1949, enhancing the xingfu of the population has been elevated to be the most important aim of the government and the Chinese Communist Party. Perceived xingfu has been central to China's governance since 1949, and has been emphasized as a national priority from Mao-era policies to the current "great rejuvenation" agenda [3,4]. In vernacular Chinese, the terms xingfu and happiness are often used together, suggesting that the two are closely connected but distinct constructs. Xingfu and happiness have been referenced in various studies and speeches, which show the connections and differences [5,6]. Despite its sociopolitical prominence, there is no consensus on the definition or measurement of xingfu or original research on xingfu and perceived xingfu among Chinese people. Direct perceived xingfu measurement remains undeveloped due to definitional ambiguities and over-reliance on Western models.

Xingfu (幸福) is often translated to "happiness" in English [7]. Conventionally, "happiness" is translated to "快樂" or "幸福" in Chinese, but "快樂" is translated into "happiness," "pleasure," or "enjoyment;" however, these 3 English words do not mean xingfu. The World Happiness Report translated to "世界幸福報告" or "世界快樂報告," with "happiness" being translated to 幸福 (xingfu) or 快樂 (happiness) but not both, has no measures on subjective happiness nor xingfu [8]. The Chinese construct of xingfu does not have a direct counterpart in Western paradigms although "well-being" is translated into "康樂," "安康," "健康," "福祉," "幸福" etc. in Chinese, indicating well-being has different meanings and is not xingfu. Despite their interconnectedness, xingfu, happiness, and well-being are distinct constructs, including "objective" indicators and subjective feelings or perceptions. Happiness, as a subjective construct, is associated with physical and mental health [9,10], but the associations of self-reported or perceived xingfu (幸福感) with health-related and socioeconomic factors remain unclear.

The 2023 Policy Address of the Hong Kong Special Administrative Region (HKSAR) Chief Executive marked the first use of "xingfu" as the headline, and introduced initiatives to promote the residents' xingfu [11]. This shows that the HKSAR Government has started to prioritize promoting xingfu, which was rarely mentioned in Hong Kong under the British rule before 1997. Hong Kong is one of the most westernized cities in China, with 91.6% of the population being ethnic Chinese [12]. Whether people in Hong Kong or Chinese communities living elsewhere can perceive and report xingfu is unknown. This study aimed to measure perceived xingfu using our newly developed single-item tool and analyze its association with health-related and socioeconomic factors in Hong Kong Chinese adults.

Methods

Survey Design

We conducted a cross-sectional survey in 2023 to measure perceived xingfu and analyze its association with health-related and socioeconomic factors in 5070 Chinese individuals aged 18 years or older, recruited from an online nonprobabilistic panel maintained by the Hong Kong Public Opinion Research Institute, a locally well-known survey agency. Invitations were distributed through email and the respondents voluntarily completed the online questionnaires, with reminders sent twice monthly. Of the 79,456 eligible respondents, 5070 completed the entire survey, with a response rate of 6.4%. This study adhered to the Checklist for Reporting Results of Internet E-Surveys (CHERRIES) for reporting guidelines [13]. Data were weighted by sex, age, educational attainment, and economic activity status based on the 2022 Hong Kong population data.

Ethical Considerations

Data confidentiality was maintained through deidentification of all respondents' records. Survey responses were stored on a password-protected server accessible only to the research team. Participants provided informed consent electronically prior to survey access. A small monetary incentive was given for questionnaire response and no images or personal identifiers are included in this study. The Institutional Review Board of the University of Hong Kong/Hospital Authority Hong Kong West Cluster granted ethical approval (UW 20 - 651) for this survey.

Instruments

We measured perceived xingfu by asking the question "你認為你有多幸福?" ("How xingfu do you think you are?") with responses on a scale of 0 - 10, with higher scores indicating higher levels of perceived xingfu. We conducted reliability testing on perceived xingfu in 147 Hong Kong residents. The two-week test-retest intraclass correlation coefficient was 0.78. This single-item perceived xingfu scale was validated and minimizes respondent burden in population-based surveys. Psychosocial factors, measured on a scale of 0 - 10, included personal happiness, perceived mental health, and adversity coping capability (ACC), with higher scores indicating higher levels. Self-rated health was measured on a five-point scale, ranging from poor to excellent [14]. Respondents also reported the number of days they felt lonely in the past 7 days, on a scale ranging from 0 to 7 days [15]. Anxiety and depressive symptoms were measured using the Patient Health Questionnaire-4. The Chinese subscales that measure anxiety and depressive symptoms had Cronbach's α coefficients of 0.82 and 0.79, respectively [16]. Perceived stress was measured using the validated Perceived Stress Scale-4, with higher scores indicating greater stress levels [17]. Information on socioeconomic factors, including sex, age, education level, monthly household income, housing, and employment status, was also collected.

Statistical Analysis

We used descriptive statistics for socioeconomic factors, self-rated health, loneliness, anxiety and depressive symptoms, and psychosocial factors. Pearson correlation coefficients were used to examine correlations between perceived xingfu and psychosocial factors, treating both as continuous variables. Linear regression models were used to examine the associations between socioeconomic factors, self-rated health, and perceived

XSL•FO
RenderX

xingfu scores using adjusted β coefficients and 95% CIs. Multivariable logistic regression analyses were used to estimate the adjusted odds ratios (aORs) and 95% CIs of high perceived xingfu (defined as ≥7) in relation to socioeconomic factors and self-rated health. The β coefficients and odds ratios (ORs) were mutually adjusted. Variance inflation factor (VIF) showed low multicollinearities for all predictors (VIFs <2). All statistical analyses were conducted using the STATA software (version 15.0; StataCorp). A *P* value <.05 was considered statistically significant.

Results

Table 1 shows that in the weighted sample, 53.2% (n=2390) were female, 64.0% (n=2811) were aged 35‑64, 63.0% (n=2852) had secondary or lower education, 50.6% (n=1938) had a household monthly income of HKD20,000‑59,999 (1 USD=7.8HKD), 60.1% (n=2570) lived in owned properties, and 58.9% were employed. Additionally, 33.7% (n=1510) self-rated their health as fair and 33.7% (n=1520) as good, while 51.9% (n=2219) reported no loneliness in the past 7 days. Moreover, 17.4% (n=752) and 17.8% (n=775) reported anxiety and depressive symptoms, respectively. The mean scores for psychosocial variables were as follows: perceived xingfu, 6.3 (SD 2.2); perceived stress, 6.6 (SD 3.2); perceived mental health, 6.3 (SD 2.2); personal happiness, 5.8 (SD 2.2); and ACC) 6.3 (SD 1.9).

XSL•FO
RenderX

Table . Socioeconomic characteristics of survey sample of Hong Kong Chinese adults.

Variables	Unweighted (n=5070)	Weighted sample (n=4515)[a]
Sex, n (%)		
Male	2213 (49.0)	2101 (46.8)
Female	2302 (51.0)	2390 (53.2)
Age group (years), n (%)		
18 - 34	1556 (34.7)	920 (20.6)
35 - 64	2624 (58.5)	2811 (63.0)
≥65	306 (6.8)	732 (16.4)
Education, n (%)		
Secondary or below	702 (15.7)	2852 (64.0)
Post-secondary or above	3776 (84.3)	1603 (36.0)
Monthly household income,[b] n (%)		
≤ HKD 19,999	556 (14.2)	1001 (26.2)
HKD 20,000 - 59,999	1938 (49.4)	1938 (50.6)
≥ HKD 60,000	1431 (36.5)	889 (23.2)
Housing, n (%)		
Rented	1669 (38.6)	1706 (40.0)
Owned	2659 (61.4)	2570 (60.1)
Employment status, n (%)		
Employed	3422 (76.9)	2608 (58.9)
Unemployed	528 (11.9)	727 (16.4)
Retired	502 (11.3)	1096 (24.7)
Self-rated health, n (%)		
Poor	156 (3.1)	173 (3.9)
Fair	1545 (30.5)	1510 (33.7)
Good	1703 (33.7)	1520 (33.7)
Very good	1471 (29.1)	1161 (25.9)
Excellent	184 (3.6)	124 (2.8)
Past 7-days loneliness (days), n (%)		
0	2323 (47.8)	2219 (51.9)
1 - 2	1623 (33.4)	1257 (29.4)
3 - 4	542 (11.2)	475 (11.1)
5 - 6	194 (4.0)	134 (3.1)
7	174 (3.6)	187 (4.4)
Patient Health Questionnaire-4 anxiety symptoms, n (%)		
No	3958 (80.1)	3572 (82.6)
Yes	983 (19.9)	752 (17.4)
Patient Health Questionnaire-4 depressive symptoms, n (%)		
No	4017 (80.9)	3569 (82.2)
Yes	948 (19.1)	775 (17.8)
Psychosocial variables, mean (SD)		

XSL•FO
RenderX

Variables	Unweighted (n=5070)	Weighted sample (n=4515)[a]
Perceived xingfu (perceived xingfu; 幸福感) score, range from 0 - 10	6.4 (2.1)	6.3 (2.2)
Perceived Stress Scale-4 score, range from 0 - 16	6.9 (3.2)	6.6 (3.2)
Perceived mental health score, range from 0 - 10	6.2 (2.2)	6.3 (2.2)
Personal happiness score, range from 0 - 10	5.9 (2.1)	5.8 (2.2)
Adversity coping capability score, range from 0 - 10	6.4 (1.9)	6.3 (1.9)

[a]Missing data (9%) were excluded. Data were weighted by sex, age, education attainment, and economic activity status based on 2022 Hong Kong population data.

[b]1 USD =7.8 HKD.

Table 2 shows that perceived xingfu was strongly and positively correlated with personal happiness (r=0.85), perceived mental health (r=0.65), and ACC (r=0.50); perceived xingfu was negatively correlated with perceived stress (r=−0.56), past 7-day loneliness (r=−0.52), anxiety symptoms (r=−0.45), and depressive symptoms (r=−0.52; all Ps<0.001).

Table . Correlations of perceived xingfu (幸福感) with psychosocial factors[a].

	Perceived xingfu	PSS-4[b]	Past 7-day loneliness	Perceived mental health	Happiness	ACC[c]	PHQ-4[d] anxiety symptoms
PSS-4	−0.56	_[e]	–	–	–	–	–
Past 7-day loneliness	−0.52	0.52	–	–	–	–	–
Perceived mental health	0.65	−0.67	−0.57	–	–	–	–
Personal happiness	0.85	−0.60	−0.54	0.71	–	–	–
ACC	0.50	−0.56	−0.37	0.61	0.54	–	–
PHQ-4 anxiety symptoms	−0.45	0.62	0.54	−0.64	−0.52	−0.4	–
PHQ-4[d] depressive symptoms	−0.52	0.66	0.61	−0.66	−0.58	−0.4	0.72

[a]Missing data (9%) were excluded. Data were weighted by sex, age, education attainment and economic activity status based on 2022 Hong Kong population data. Higher scores indicating higher levels for all variables. All P values for Pearson correlation coefficients <0.001.

[b]PSS-4: Perceived Stress Scale-4.

[c]ACC: adversity coping capability.

[d]PHQ-4: Patient Health Questionnaire-4.

[e]Not applicable.

Figure 1 shows the distribution of perceived xingfu and happiness scores. As both perceived xingfu and happiness peaked at scores of 7 (22%) and 8 (23%); therefore, perceived xingfu ≥7 was classified as high perceived xingfu in the logistic regression model. Similar perceived xingfu and happiness scores were reported by 48.9% (n=2478) of respondents, while 46.5% (n=2360) showed a difference between perceived xingfu and happiness scores ranging from −1 to +3. The remaining 4.6% (n=232) showed a difference greater than −1 and less than +3 (Multimedia Appendix 1).

XSL•FO
RenderX

Figure 1. Distribution of perceived xingfu and happiness scores. PX: perceived xingfu.

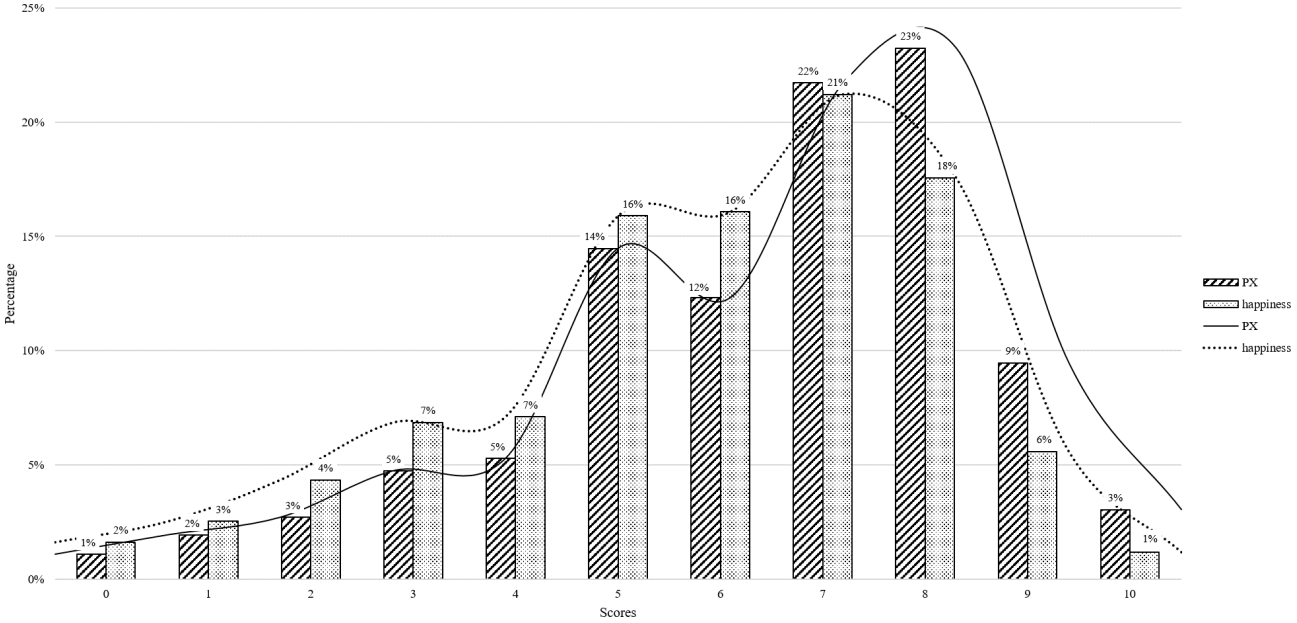

Table 3 shows that, in the linear regression model, the following variables were associated with higher perceived xingfu scores: being female (β=0.69, 95% CI 0.58-0.81; P<.001); being aged 35-64 years (β=0.23, 95% CI 0.10-0.36) or ≥65 years (β=0.46, 95% CI 0.15-0.76; P_{trend}<.001); having post-secondary education or above (β=0.19, 95% CI 0.02-0.36; P<.01); having a higher household monthly income (HKD20,000 - 59,999: β=0.33, 95% CI 0.14-0.52; ≥ HKD60,000: β=0.99, 95% CI 0.79-1.20; P_{trend}<.001); living in owned properties (β=0.27, 95% CI 0.14-0.39; P<.001); being retired (β=0.56, 95% CI 0.33-0.79; P<.001); and having excellent versus poor self-rated health (β=3.84, 95% CI 3.39 - 4.29; P<.001).

Table 3 also shows that in the logistic regression model, the following variables were associated with higher odds of reporting high perceived stress (≥7): female sex (aOR 2.11, 95% CI 1.82 - 2.46; P<.001); those aged 35 - 64 (aOR 1.64, 95% CI 1.34 - 2.02) or ≥65 (aOR 2.67, 95% CI 1.97 - 3.62; P_{trend}<.001), respondents with post-secondary education or above (aOR 1.35, 95% CI 1.14 - 1.59; P<.001), higher household monthly income (HKD20,000 - 59,999: aOR 1.74, 95% CI 1.45 - 2.09;≥ HKD60,000: aOR 3.04, 95% CI 2.41 - 3.83; P_{trend}<.001), and those living in owned properties (aOR 1.57, 95% CI 1.34 - 1.83, P<.001). Respondents who reported better self-rated health had higher odds of high perceived xingfu (P_{trend}<.001), with the largest aOR in those who reported excellent health (aOR 4.72, 95% CI 18.43 - 89.93; P<.001).

XSL•FO
RenderX

Table . Linear and logistic regressions of perceived xingfu (幸福感) with socioeconomic factors and self-rated health.

Variables[a]	Perceived xingfu, mean (SD)	β (95% CI)[b]	*P* value[c]	High levels of perceived xingfu (≥7), n (%)	aOR[d] (95% CI)[b]	*P* value[c]
Sex			<.001			<.001
Male	6.0 (2.2)	0		1008 (48.0)	1	
Female	6.5 (2.1)	0.69 (0.58- 0.81)		1449 (6.6)	2.11 (1.82 - 2.46)	
Age group (years)			<.001[f]			<.001[f]
18 - 34	6.0 (2.2)	0		439 (47.7)	1	
35 - 64	6.3 (2.2)	0.23 (0.10-0.36)	<.001	1536 (54.7)	1.64 (1.34 - 2.02)	<.001
≥65	6.8 (1.8)	0.46 (0.15-0.76)	<.001	464 (63.4)	2.67 (1.97 - 3.62)	<.001
Education			.001			<.001
Secondary or below	6.1 (2.2)	0		1462 (51.3)	1	
Post-secondary or above	6.6 (2.0)	0.19 (0.02-0.36)		976 (6.9)	1.35 (1.14 - 1.59)	
Monthly household income[e]			<.001[f]			<.001[f]
≤ HKD19,999	5.9 (2.3)	0		436 (43.5)	1	
HKD20,000 - 59,999	6.3 (2.1)	0.33 (0.14-0.52)	<.001	1066 (55.0)	1.74 (1.45 - 2.09)	<.001
≥ HKD60,000	7.1 (1.7)	0.99 (0.79 - 1.20)	<.001	632 (71.1)	3.04 (2.41 - 3.83)	<.001
Housing			<.001			<.001
Rented	5.8 (2.3)	0		755 (44.3)	1	
Owned	6.6 (2.0)	0.27 (0.14-0.39)		1573 (61.2)	1.57 (1.34 - 1.83)	
Employment status			<.001[f]			.16[f]
Employed	6.2 (2.1)	0		1817 (53.1)	1	
Unemployed	5.8 (2.5)	0.01 (–0.19 to 0.21)	.843	245 (46.5)	1.05 (0.84 - 1.31)	.649
Retired	6.9 (1.8)	0.56 (0.33-0.79)	<.001	324 (64.5)	1.18 (0.95 - 1.45)	.131
Self-rated health			<.001[f]			<.001[f]
Poor	3.7 (2.5)	0		26 (14.9)	1	
Fair	5.5 (2.1)	1.61 (1.26 - 1.97)	<.001	563 (37.3)	3.19 (1.89 - 5.40)	<.001
Good	6.6 (1.9)	2.34 (1.99 - 2.69)	<.001	907 (59.7)	7.31 (4.32 - 12.35)	<.001
Very good	7.2 (1.8)	2.95 (2.60 - 3.31)	<.001	854 (73.6)	14.21 (8.35 - 24.18)	<.001
Excellent	8.1 (1.8)	3.84 (3.39 - 4.29)	<.001	105 (85.0)	4.72 (18.43 - 89.93)	<.001

[a]Missing data (9%) were excluded.

[b]Mutually adjusted.

[c]Data were weighted by sex, age, education attainment and economic activity status based on 2022 Hong Kong population data.

[d]aOR: adjusted odds ratio.

[e]USD=7.8 HKD.

[f]*P* for trend.

Discussion

Principal Findings

We first developed a novel, single-item question to measure perceived xingfu and showed that xingfu is a unique psychosocial construct in Chinese culture. Our findings show that perceived xingfu is strongly associated with socioeconomic, psychosocial, and health-related factors, aligning with—but also diverging from—Western concepts of happiness. Strong positive correlations were observed between perceived xingfu and perceived mental health, personal happiness, and ACC,

XSL•FO
RenderX

while negative correlations with negative constructs, including perceived stress, past 7-day loneliness, anxiety and depressive symptoms demonstrate the multidimensionality of this instrument.

Socioeconomic factors, including being female, older age, higher education level, higher monthly household income, and living in owned properties, were positively associated with higher perceived xingfu scores and high levels of perceived xingfu. These findings align with global patterns in which socioeconomic advantages are exacerbated with higher levels of happiness; however, Hong Kong's extreme income inequality and housing scarcity likely increase these disparities. For instance, the steep income gradient (ie, higher aORs with increasing income) shows the significant role of financial security in high-cost environment, while homeownership reflects the importance of psychological safety gained by property ownership in a volatile housing market [18,19]. Such socioeconomic disparities in perceived xingfu are expected, as xingfu is strongly correlated with happiness. Interestingly, retired respondents reported higher perceived xingfu scores than employed respondents, which contrasts with previous studies on employment and life satisfaction [20]. This may reflect cultural values in Hong Kong, where retirement is associated with familial respect, reduced work-related stress, or financial preparedness through systems [21]. Future studies could investigate how younger generations prioritize career success, while older generations place greater emphasis on family harmony and traditional values. Higher perceived xingfu among females and older adults also aligns with Western findings of increased well-being in these groups [22], while strong social networks or Confucian norms may mediate these outcomes in Chinese societies [23]. While these findings differ from or are in parallel with Western research on well-being, happiness, life satisfaction, and quality of life, perceived xingfu remains a culturally distinct construct in Chinese societies. Thus, caution is needed when generalizing these findings. Policy efforts targeting income inequality, housing access, and health equity could enhance perceived xingfu; however, interventions must consider Hong Kong's unique sociocultural environment.

Although xingfu has been a less commonly used construct in Hong Kong than in Mainland China, respondents had no difficulty in answering our question and giving a score. Moreover, the mean perceived xingfu score was higher than the mean happiness score, with only 51% reporting the same score. This suggests perceived xingfu transcends transient emotional states of happiness, which aligns with prior studies emphasizing on xingfu as a holistic, socially anchored concept in Chinese societies, distinct from Western individualism-centric metrics [2]. Culture-specific measures rather than direct translations of Western instruments are needed. Direct translations of Western tools pose a risk of overlooking contextual priorities, such as the emphasis on familial cohesion over personal achievement among older generations. Policies and measures to promote xingfu or perceived xingfu to the populations or communities must also be capable of reducing disparities. People of different sexes, ages, and socioeconomic statuses may have different needs and perceptions, which warrantees further studies. By focusing on factors that may influence perceived xingfu within Hong Kong's sociopolitical and cultural context, our study can contribute to a deeper understanding of perceived xingfu as a unique psychosocial construct in Chinese societies. These findings indicate that perceived xingfu represents a more nuanced construct than happiness alone. Hence, perceived xingfu should be measured and monitored in populations and communities where xingfu is highly valued, either beyond or together with happiness. Experiences and achievements in promoting perceived xingfu in Hong Kong and Mainland China could be better understood and disseminated to other countries and regions.

Limitations

Our study has some limitations. First, perceived xingfu and xingfu may represent different meanings or domains and are susceptible to cultural nuances and contextual factors. Qualitative research is needed to provide deeper insights into the multidimensional nature of perceived xingfu and xingfu. Second, although one of the Chinese translations of "well-being" is 幸福, we did not analyze well-being and perceived xingfu, because of the lack of commonly accepted and simple tools for measuring subjective well-being. Third, we measured perceived xingfu rather than xingfu as an "objective" construct defined by environmental or often external indicators. Further studies on whether some of the domains or indicators included in the World Happiness Report may be applicable domains or indicators of xingfu in Chinese contexts, are warranted. Fourth, while online recruitment ensured efficiency, it may underrepresent elderly or low-income populations with limited digital access. Future studies should combine online and community-based sampling to enhance generalizability. Fifth, possible social desirability bias exist due to self-reporting sensitive psychological states. Our tool is simple and should be useful for measuring socioeconomic disparities in the pursuit of xingfu by the government for the people and people's individual pursuit of xingfu. Hence, monitoring perceived xingfu in the population regularly is both feasible and warranted. If xingfu is to be assessed or measured by different domains indirectly, perceived xingfu should still be included as a direct measure and weighted appropriately. Sixth, the cross-sectional design precludes causal inference. Future studies should compare perceived xingfu across Chinese subpopulations, such as those in Mainland China, Hong Kong, and overseas. Longitudinal studies can offer insights into perceived xingfu's dynamic nature and potential causal relationships with identified factors.

Conclusions

Our novel, simple, single-item tool showed that perceived xingfu was associated with socioeconomic, psychological, and health-related factors. Perceived xingfu is a unique Chinese psychosocial construct that frequently appears with, but is distinct from happiness. Monitoring perceived xingfu could inform equitable health policies in Hong Kong and beyond. Further exploration of the applicability of this measurement in different regions of China and among different populations can help assess its generalizability, and provide a more comprehensive understanding of perceived xingfu across various Chinese communities within and beyond China.

XSL•FO
RenderX

Data Availability

The datasets generated or analyzed during this study are not publicly available due to privacy concerns but are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Distribution of perceived xingfu (幸福感) and happiness. Explanation of xingfu (幸福).
[DOCX File, 29 KB - formative_v9i1e73350_app1.docx]

Checklist 1
Checklist for Reporting Results of Internet E-Surveys (CHERRIES).
[DOCX File, 19 KB - formative_v9i1e73350_app2.docx]

References

1. Xing ZJ. Research on the compilation of the subjective xingfu scale of Chinese urban residents (中國城市居民主觀幸福感量表的編制研究) [Chinese]. : Shanghai: East China Normal University; 2003 URL: https://tpl.ncl.edu.tw/NclService/JournalContentDetail?SysId=A02026736 [accessed 2025-06-27]
2. Zhang MJ. Xingfu Science and Naturalistic “xingfu (幸福科學與自然主義的“幸福學”): National Social Science Database; 2016, Vol. 3.
3. The East Is Red (東方紅).: People Daily; 1967. URL: http://cpc.people.com.cn/BIG5/64150/64154/4482098.html [accessed 2025-06-27]
4. Liou CS, Ding SF. China Dreams: China’s New Leadership and Future Impacts: World Scientific; 2015.
5. Yang GY. Xi said: to be happy in life, it is very important to keep fit [Chinese]. China News Network. 2023. URL: http://politics.people.com.cn/BIG5/n1/2023/0421/c1001-32670348.html [accessed 2025-06-27]
6. Peng PJ, Li JD. Research on factors affecting subjective xingfu (影響主觀幸福感因素之研究). Taipei: National Taiwan University. 2014. URL: https://tinyurl.com/muj22nk6 [accessed 2025-06-27]
7. Cambridge Dictionary.: Cambridge: Cambridge University Press; 2023. URL: https://dictionary.cambridge.org/ [accessed 2025-06-27]
8. Helliwell JF, Layard R, Sachs JD, Neve JE. World happiness report 2020. : New York: Sustainable Development Solutions Network; 2020 URL: https://worldhappiness.report/ed/2020/ [accessed 2025-06-27]
9. Veenhoven R. Healthy happiness: effects of happiness on physical health and the consequences for preventive health care. J Happiness Stud 2008 Sep;9(3):449-469. [doi: 10.1007/s10902-006-9042-1]
10. Fordyce MW. A review of research on the happiness measures: a sixty second index of happiness and mental health. Soc Indic Res 1988 Aug;20(4):355-381. [doi: 10.1007/BF00302333]
11. Hong Kong Special Administrative Region Government. Hong Kong Policy Address 2023. 2023. URL: https://www.policyaddress.gov.hk/2023/en/policy.html [accessed 2025-06-27]
12. Hong kong – the facts. Hong Kong Census and Statistics Department. 2023. URL: https://www.gov.hk/en/about/abouthk/facts.htm [accessed 2025-06-27]
13. Eysenbach G. Improving the quality of web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34. [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
14. Wu S, Wang R, Zhao Y, et al. The relationship between self-rated health and objective health status: a population-based study. BMC Public Health 2013 Apr 9;13(1):320. [doi: 10.1186/1471-2458-13-320] [Medline: 23570559]
15. Radloff LS. The CES-D scale: a self-report depression scale for research in the general population. Appl Psychol Meas 1977;1(3):385-401. [doi: 10.1177/014662167700100306]
16. Chen B, Gong W, Lai AYK, et al. Family context as a double-edged sword for psychological distress amid the COVID-19 pandemic with the mediating effect of individual fear and the moderating effect of household income. Front Public Health 2023;11:1109446. [doi: 10.3389/fpubh.2023.1109446] [Medline: 37033084]
17. Huang F, Wang H, Wang Z, et al. Psychometric properties of the perceived stress scale in a community sample of Chinese. BMC Psychiatry 2020 Mar 20;20(1):130. [doi: 10.1186/s12888-020-02520-4] [Medline: 32197589]
18. Kahneman D, Deaton A. High income improves evaluation of life but not emotional well-being. Proc Natl Acad Sci U S A 2010 Sep 21;107(38):16489-16493. [doi: 10.1073/pnas.1011492107] [Medline: 20823223]
19. Hu F. Homeownership and subjective wellbeing in urban China: Does owning a house make you happier? Soc Indic Res 2013 Feb;110(3):951-971. [doi: 10.1007/s11205-011-9967-6]

XSL•FO
RenderX

20. Tang F, Chen H, Zhang Y, Mui AC. Employment and life satisfaction among middle- and old-aged adults in China. Gerontology and Geriatric Medicine 2018 Jan 1;4:2333721418778202. [doi: 10.1177/2333721418778202]
21. Chong AML, Ng SH, Woo J, Kwan AYH. Positive ageing: the views of middle-aged and older adults in Hong Kong. Ageing Soc 2006 Mar;26(2):243-265. [doi: 10.1017/S0144686X05004228]
22. Inglehart R. Gender, aging, and subjective well-being. Int J Comp Sociol 2002 Oct;43(3-5):391-408. [doi: 10.1177/002071520204300309]
23. Wang X, Kanungo RN. Nationality, social network and psychological well-being: expatriates in China. In: Human Resource Management in China Revisited: Routledge; 2020:159-177. [doi: 10.4324/9781003060390-9]

Abbreviations

ACC: adversity coping capability
aOR: adjusted odds ratio

Edited by A Mavragani; submitted 02.03.25; peer-reviewed by HF Gemede; revised version received 11.06.25; accepted 12.06.25; published 07.07.25.

Please cite as:
Sze KYP, Ho SY, Lai AYK, Jia J, Xu H, Sit SMM, Lam TH, Wang MP
Association of Perceived Xingfu With Health-Related and Socioeconomic Factors Among Hong Kong Chinese Adults: Cross-Sectional Study Using a Novel Single-Item Tool
JMIR Form Res 2025;9:e73350
URL: https://formative.jmir.org/2025/1/e73350
doi:10.2196/73350

© Katherine Y P Sze, Sai Yin Ho, Agnes Yuen Kwan Lai, Jing Jia, Heng Xu, Shirley Man Man Sit, Tai Hing Lam, Man Ping Wang. Originally published in JMIR Formative Research (https://formative.jmir.org), 7.7.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Validation of The Umbrella Collaboration for Tertiary Evidence Synthesis in Geriatrics: Mixed Methods Study

Beltran Carrillo[1*], MD, PhD; Marta Rubinos-Cuadrado[1*], RNC; Jazmin Parellada[1], BBI; Alejandra Palacios[1], BME; Beltran Carrillo-Rubinos[1], BME; Fernando Canillas[2], MD, PhD; Juan José Baztán Cortés[3], MD, PhD; Javier Gómez-Pavón[3], MD, PhD

[1]The Umbrella Collaboration, C/ Ferraz, 49, Madrid, Spain

[2]Department of Traumatology and Orthopedic Surgery, Hospital Central de la Cruz Roja San José y Santa Adela, Madrid, Spain

[3]Department of Geriatrics, Hospital Central de la Cruz Roja San José y Santa Adela, Madrid, Spain

[*]these authors contributed equally

Corresponding Author:
Beltran Carrillo, MD, PhD
The Umbrella Collaboration, C/ Ferraz, 49, Madrid, Spain

Abstract

Background: The synthesis of evidence in health care is essential for informed decision-making and policy development. This study aims to validate The Umbrella Collaboration (TU), an innovative, semiautomated tertiary evidence synthesis methodology, by comparing it with traditional umbrella reviews (TURs), which are currently the gold standard.

Objective: The primary objective of this study is to evaluate whether TU, an artificial intelligence–assisted, software-driven system for tertiary evidence synthesis, can achieve effectiveness comparable to that of TURs, while offering a more timely, efficient, and comprehensive approach.

Methods: This comparative study evaluated TU against TURs across 8 matched projects in geriatrics. For each selected TUR, a parallel TU project was conducted using the same research question. Outcomes of interest (OoIs), effect sizes, certainty ratings, and execution times were systematically compared. Effect sizes were assessed both quantitatively, by transforming TUR metrics to Cohen *d* and correlating them with TU's R_{TU} metric, and qualitatively, through categorical classifications (trivial, small, moderate, and large). Certainty levels were compared by mapping Grading of Recommendations Assessment, Development, and Evaluation (GRADE) ratings and TU's sentiment analysis scores onto a common 0‑1 scale. Execution time was measured precisely in TU, while TUR durations were estimated from literature benchmarks. Statistical analyses included chi-square tests and Spearman correlations.

Results: Eight TURs in geriatrics were matched with parallel projects using TU. TU replicated 73 of the 86 (85%) OoIs identified by TURs and reported an additional 337 OoIs, representing a 4.77-fold increase in outcome identification. In the comparison of effect size classifications, full concordance was observed in 24 of the 48 (50%) cases, and consistent concordance (full plus 1-level deviation) in 45 of the 48 (94%) cases, with a moderate strength of association (Cramér V=0.339). The correlation of transformed certainty values between TU and GRADE yielded a statistically significant Spearman coefficient (ρ=0.446; P=.02). The average execution time per TU project was 4 hours and 46 minutes, compared with estimated durations of 6‑12 months for TURs.

Conclusions: The TU demonstrated high concordance with TURs, replicating 73 of the 86 (85%) outcomes identified by TURs and identifying nearly 5 times as many additional outcomes. The experimental effect size metric (R_{TU}) showed moderate agreement with conventional measures, and the certainty ratings derived from sentiment analysis correlated acceptably with GRADE-based assessments. While further validation is needed, TU appears to be a valid and efficient approach for tertiary evidence synthesis, offering a scalable and time-efficient alternative when rapid results are required.

International Registered Report Identifier (IRRID): RR2-10.2196/67248

(JMIR Form Res 2025;9:e75215) doi:10.2196/75215

KEYWORDS

tertiary evidence synthesis; The Umbrella Collaboration; umbrella reviews; health research methodology; AI-assisted synthesis; evidence-based decision-making; algorithms; analytics; artificial intelligence; digital health; digital interventions; digital technology; machine learning; models

XSL•FO
RenderX

Introduction

Background

Synthesizing evidence in health care transforms large volumes of data into actionable knowledge, enabling decisions based on the best available information. This process integrates findings from multiple sources to produce clear, accurate, and accessible summaries for clinicians, policy makers, and patients alike [1]. As a core component of knowledge translation, evidence synthesis bridges research and practice, and is essential for developing effective health policies [2,3].

Yet limited statistical and health literacy among professionals and the public often hampers the effective use of synthesized evidence [4-6]. This highlights the need for tools that democratize access to high-level information and support meaningful stakeholder participation in health care decisions [7].

In recent years, tertiary evidence synthesis, commonly known as umbrella reviews, has emerged as a third tier in the evidence hierarchy, building on primary studies and systematic reviews with or without meta-analyses (SRs/MAs). Also referred to as overviews, meta-reviews, or meta-epidemiological studies [8,9], umbrella reviews (hereafter referred to as traditional umbrella reviews [TURs]) are particularly valuable for addressing broad questions, generating rapid insights, or navigating resource constraints. Despite structured guidance from organizations such as Cochrane and the Joanna Briggs Institute [10-14], methodological inconsistency across TURs persists due to variations in implementation.

The COVID-19 pandemic underscored the need for faster synthesis methods, even at the expense of some precision [15]. In this context, The Umbrella Collaboration (TU) introduces a novel artificial intelligence (AI)–assisted approach to tertiary synthesis, combining automation and human oversight. Although tools such as Covidence, Rayyan, and DistillerSR have improved aspects of secondary synthesis [16], the application of AI to tertiary synthesis remains largely unexplored. Large language models (LLMs) such as ChatGPT show potential for automating SRs [17], but dedicated software for tertiary synthesis has yet to be established.

TU addresses this gap by offering a structured, reproducible, and fully digital alternative to TURs. It is a patent-pending software system that automates tertiary evidence synthesis through a combination of algorithmic processes, natural language processing (NLP), and selective use of LLMs. TU retrieves abstracts of SRs/MAs from MEDLINE via PubMed, using LLMs (eg, ChatGPT-4 [18]) to suggest related terms, validated by a human reviewer, to enhance search sensitivity without sacrificing specificity. Crucially, AI is limited to this initial phase; subsequent steps are fully managed by rule-based software to ensure transparency and auditability.

TU extracts and synthesizes key information, including outcomes of interest (OoIs), effect sizes (ESs), direction, statistical significance, and a certainty estimate derived from sentiment analysis, and presents results in plain language through an interactive web platform. The system updates daily, supporting the principles of living SRs [19]. Unlike tools that assist in isolated stages of secondary synthesis, TU delivers an end-to-end solution tailored specifically for tertiary synthesis. A schematic diagram (Figure 1) illustrates the overall architecture, from automated literature retrieval to final result visualization. A more detailed technical description of the algorithms and processes is available in Multimedia Appendix 1.

As AI evolves, fully automated synthesis workflows may become feasible, but rigorous validation is essential to ensure trust, transparency, and scientific integrity.

The implementation of new methodologies in the scientific field requires a comparative validation process with established methods to confirm their reliability and effectiveness. TU, being an innovative methodology still in its theoretical-conceptual stage, must be evaluated against established methodologies. Therefore, this study aims to validate TU by comparing its performance and outcomes with the gold standard, TURs, to establish its credibility and potential superiority.

Figure 1. The Umbrella Collaboration workflow. LLM: large language model; MA: meta-analysis; NLP: natural language processing; SA: sentiment analysis; SR: systematic review.

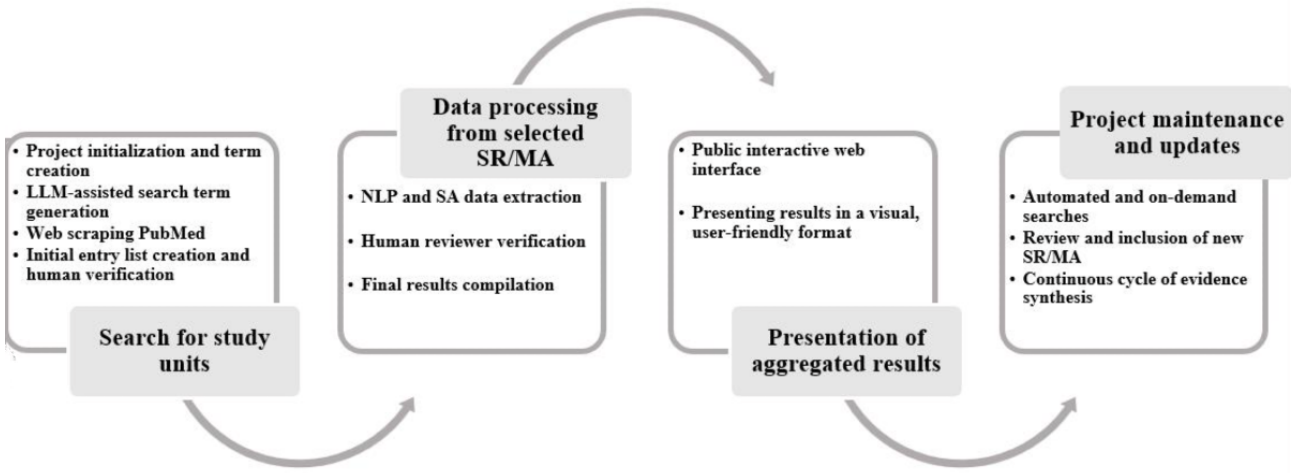

XSL•FO
RenderX

Objectives

This study aims to evaluate whether TU, a software-driven and AI-assisted system, can produce results comparable to TURs, while offering a faster, more efficient, and potentially more comprehensive approach.

Methods

Study Design

Figure 2 outlines the study design. Using a structured comparative approach, 8 TURs in geriatrics were selected as reference models. Each was replicated in TU using the same research questions. Key variables, including OoI, ES, certainty, and execution time, were systematically collected and compared across both methodologies.

To compare both methodologies, we conducted a targeted PubMed search to identify representative TURs in geriatrics. Rather than aiming for an exhaustive review, this focused approach was designed to select suitable benchmarks for parallel evaluation. Given the study's aim, methodological comparison rather than comprehensive coverage, this simplified strategy was both appropriate and intentional.

Figure 2. Study workflow. TU: Umbrella Collaboration; TUR: traditional umbrella review.

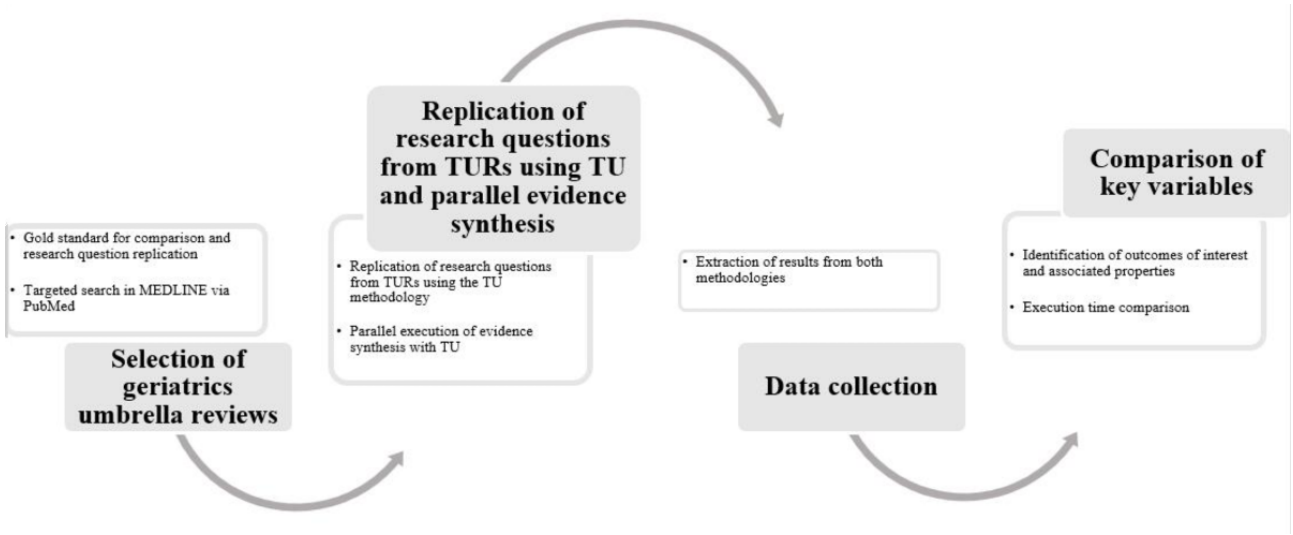

Study Variables

Outcomes of Interest

A key variable was the identification of OoIs, defined as specific effects evaluated by SRs/MAs addressing the same research question. We assessed concordance between TU and TURs using a concordance matrix and compared the total number of OoIs identified by each methodology through descriptive and statistical analyses.

Effect Size of Outcome of Interest

ES for each OoI was assessed using 2 complementary strategies: a quantitative comparison of numerical values and a qualitative classification into standard categories (trivial, small, moderate, and large). TU applies automated standardization using a proprietary metric (R_{TU}), which converts various ES formats (eg, standardized mean difference [SMD], mean difference, risk ratio, odds ratio, hazard ratio) into a unified, weighted score for consistent synthesis. TUR-derived ES values were transformed into Cohen *d* for comparison. Spearman correlation was used to assess quantitative concordance, while a contingency matrix evaluated categorical agreement, distinguishing full, partial, and major discordance (Table 1). Based on the degree of agreement, each matched OoI was assigned to one of the following concordance levels.

XSL•FO
RenderX

Table . Concordance matrix for ES[a].

TUR[b] ES	TU[c] ES			
	Trivial	Small	Moderate	Large
Trivial	Full concordance	Partial concordance (discrepancy of 1 category)	Discordance (discrepancy>1 category)	Discordance (discrepancy>1 category)
Small	Partial concordance (discrepancy of 1 category)	Full concordance	Partial concordance (discrepancy of 1 category)	Discordance (discrepancy>1 category)
Moderate	Discordance (discrepancy>1 category)	Partial concordance (discrepancy of 1 category)	Full concordance	Partial concordance (discrepancy of 1 category)
Large	Discordance (discrepancy>1 category)	Discordance (discrepancy>1 category)	Partial concordance (discrepancy of 1 category)	Full concordance

[a]ES: effect size.

[b]TUR: traditional umbrella review.

[c]TU: Umbrella Collaboration.

Full Concordance

Both TU and TUR classified the outcome in the same category (eg, trivial-trivial).

Partial Concordance

The classifications differed by only 1 level (eg, trivial-small or moderate-large). These differences were considered acceptable, as they are unlikely to lead to substantial changes in interpretation or clinical decision-making.

Discordance

The classifications differed by 2 or more levels (eg, trivial-moderate or small-large), representing a more substantial methodological discrepancy.

Consistent Concordance

A combined category including both full and partial concordance, reflecting a general alignment between methodologies even when minor categorical differences were present.

Certainty of Evidence

TU estimates the certainty of evidence for each OoI using automated sentiment analysis (SA) applied to SR/MA abstracts. While this approach, based on NLP and a model trained on Twitter/X data, does not replicate the multidimensional Grading of Recommendations Assessment, Development, and Evaluation (GRADE) framework [20], it offers a rapid and scalable approximation. GRADE ratings from TURs were categorized as very low to high, while TU's SA scores (ranging from −1 to +1) were normalized to a 0 - 1 scale using the following formula: (SA score + 1)/2. GRADE levels were similarly mapped to enable direct comparison and statistical concordance analysis.

Execution Time

The execution time of each methodology was assessed, with TU providing exact time measurements and TURs relying on an estimated time frame of 6-12 months based on existing literature.

Data Collection and Research Question Replication

A targeted PubMed search using the terms "umbrella" AND "geriatric" was conducted to identify suitable TURs, which served as benchmarks. Their research questions were replicated in TU without modification. TU then applied automated searches and synthesis using NLP, web scraping, sentiment analysis, and machine learning, with human oversight for verification and extraction. Data from both methodologies were systematically collected to compare outcomes, ESs, certainty, and execution time.

TU relied solely on abstracts due to practical constraints, such as limited access to full texts and the aim to minimize language bias, as all MEDLINE abstracts are in English. Although abstracts may omit methodological details, TU uses structured extraction criteria to capture key information. This study explicitly evaluates whether abstract-based synthesis in TU can yield conclusions comparable to those from full-text TURs.

Statistical Analysis: Data Analysis and Statistical Methods

Results from TU and TURs were compared using contingency tables for outcomes and ESs. Certainty scores from both methods were transformed to a 0 - 1 scale for direct comparison. Chi-square tests assessed differences between methodologies, and Spearman correlations evaluated the association between TU certainty estimates and GRADE ratings.

Ethical Considerations

This study did not involve human participants, personal data, or animals. All data were obtained from previously published studies and analyzed in accordance with established ethical standards for secondary data analysis. Therefore, ethical review or approval was not required. This is consistent with the institutional policy of Universidad Alfonso X El Sabio (Villanueva de la Cañada, Madrid, Spain), which exempts research based solely on publicly available, nonidentifiable data from review by the institutional review board.

XSL•FO
RenderX

Results

Identification of Traditional Umbrella Reviews in Geriatrics as Reference Models

To establish reference models, we conducted a PubMed search on March 5, 2023, using the terms “umbrella” AND “geriatric,” which yielded 111 results. After excluding 75 records for irrelevance or ineligibility, 36 TURs remained. From these, 8 were randomly selected, with a preference for recent publications, to serve as comparators for TU [21-28]. The selection process is summarized in Figure 3 and Multimedia Appendix 2.

Figure 3. Flow diagram of the search and selection process for traditional umbrella reviews (TURs) in geriatrics.

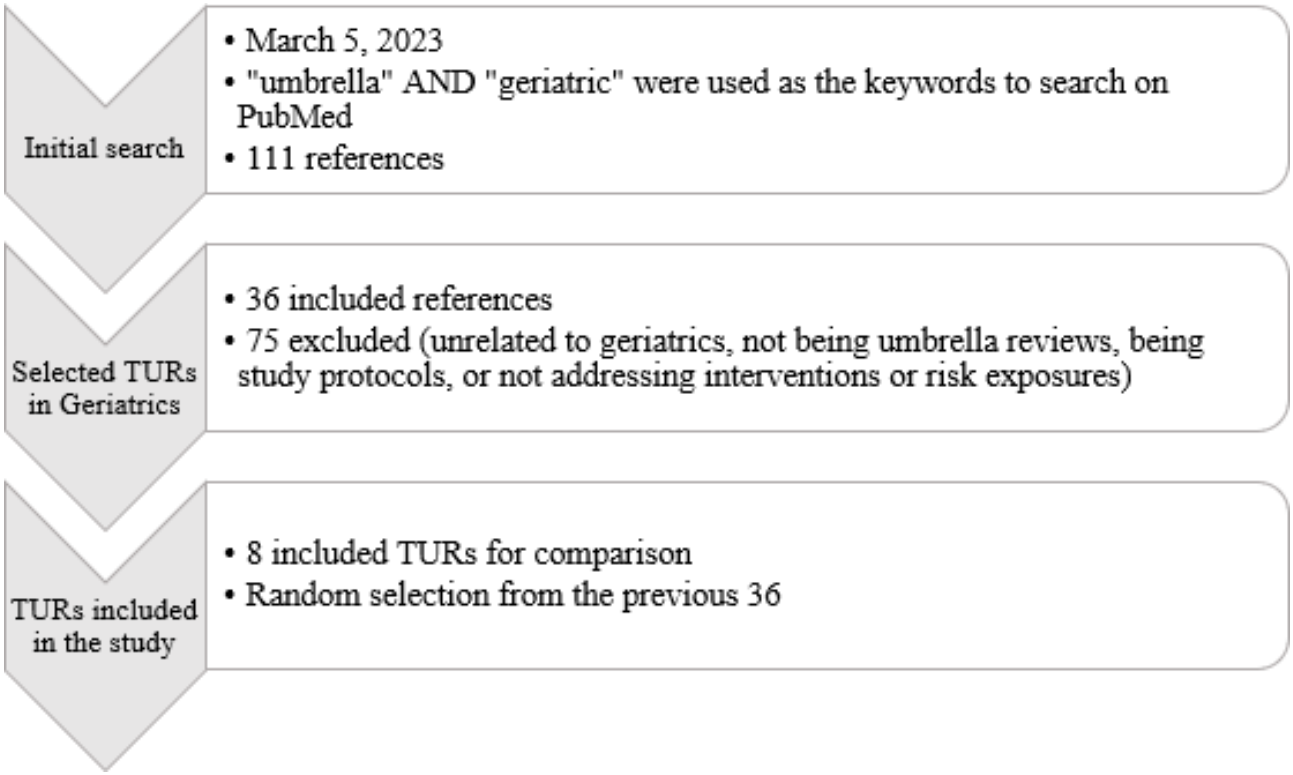

Identification of Outcomes of Interest

Across the 8 comparative projects, TURs identified 86 OoIs, while TU identified 410 for the same research questions. TU replicated 73 of the 86 (85%) TUR OoIs and missed 13 (15%). Conversely, only 73 out of 410 (17.8%) TU’s OoIs were reported by TURs, leaving 337 (82.2%) additional OoIs uniquely identified by TU (Table 2).

The identification gain factor, calculated as the ratio of OoIs identified by TU to those identified by TURs, quantifies TU’s broader retrieval capacity. TU consistently identified more OoIs across all projects, with an overall gain factor of 4.77 (ie, nearly 5 times more outcomes). Project-specific gain factors ranged from 1.87 to 18.63 (Table 3).

XSL•FO
RenderX

Table . Global summary table of OoIs[a] identified by TURs[b] and TU[c]: totals, matches, and discrepancies.

	Project								Total (TUR: 86 OoIs; TU: 140 OoIs)
	1 [22] (TUR: 5 OoIs; TU: 12 OoIs)	2 [27] (TUR: 38 OoIs; TU: 71 OoIs)	3 [23] (TUR: 7 OoIs; TU: 34 OoIs)	4 [21] (TUR: 10 OoIs; TU: 22 OoIs)	5 [28] (TUR: 4 OoIs; TU: 25 OoIs)	6 [24] (TUR: 8 OoIs; TU: 149 OoIs)	7 [25] (TUR: 6 OoIs; TU: 49 OoIs)	8 [26] (TUR: 8 OoIs; TU: 48 OoIs)	
TUR OoI identified by TU, n (% TUR)	5 (100.0)	30 (78.9)	5 (71.4)	7 (70.0)	4 (100.0)	8 (100.0)	6 (100.0)	8 (100.0)	73 (84.9)
TUR OoI not identified by TU, n (% TUR)	N/A[d]	8 (21.1)	2 (28.6)	3 (30.00)	N/A	N/A	N/A	N/A	13 (15.1)
TU OoI identified by TUR, n (% TU)	5 (41.7)	30 (42.3)	5 (14.7)	7 (31.8)	4 (16.0)	8 (5.4)	6 (12.2)	8 (16.7)	73 (17.8)
TU OoI not identified by TUR, n (% TU)	7 (58.3)	41 (57.7)	29 (85.3)	15 (68.2)	21 (84.0)	141 (94.6)	43 (87.8)	40 (83.3)	337 (82.2)

[a]OoI: outcome of interest.

[b]TUR: traditional umbrella review.

[c]TU: Umbrella Collaboration.

[d]N/A: not applicable.

Table . OoIs[a] identified by TU[b] and TURs[c], additional OoIs identified by TU, and identification gain factor.[d]

Project	Number of OoIs identified by TU, n (n=410)	Number of OoIs identified by TUR, n (n=86)	Additional OoIs identified by TU, n (n=324)	Identification gain factor of OoIs by TU compared with TUR
Project 1: Veronese et al (2023) [22]	12	5	7	2.40
Project 2: Marx et al (2021) [27]	71	38	33	1.87
Project 3: Shen et al (2022) [23]	34	7	27	4.86
Project 4: Conneely et al (2022) [21]	22	10	12	2.20
Project 5: Gazzaniga et al (2023) [28]	25	4	21	6.25
Project 6: Musazadeh et al (2023) [24]	149	8	141	18.63
Project 7: Veronese et al (2021) (1)	49	6	43	8.17
Project 8: Veronese et al (2021) (2)	48	8	40	6.00

[a]OoI: outcome of interest.

[b]TU: Umbrella Collaboration.

[c]TUR: traditional umbrella review.

[d]The total identification gain factor of OoIs by TU compared with TUR is 4.77.

XSL•FO
RenderX

Concordance in Effect Size Classification of Outcomes of Interest

To assess agreement, ESs for 48 matched OoIs were compared between TURs and TU. TUR-reported ESs were standardized (Cohen *d*) for comparison with TU's R_{TU} metric. TURs classified most OoIs as small (n=23, 48%) or trivial (n=16, 33%). TU showed a similar pattern, with 26 (54%) classified as small and 17 (35%) classified as trivial. Moderate effects were less frequent—TUR: n=7 (15%); TU: n=5 (10%)—and large effects were rare, with TUR reporting 2 (4%) and TU reporting none.

Table 4 summarizes the categorical agreement in ES classification. TU labeled 17 out of 48 (35%) OoIs as trivial; of these, 10 out of 17 (59%) matched TUR classifications, while 7 out of 17 (41%) were classified as small. Among OoIs, TU classified as small (26/48, 54%), half-matched TURs, with the rest spread across trivial (5/26, 19%), moderate (6/23, 23%), and large (2/26, 8%) classifications. For the 5 of 48 (10%) OoIs labeled as moderate by TU, TURs agreed in only 1 case, classifying most as small or trivial.

Of the 48 OoIs analyzed, full concordance in ES classification was found in 24 (50%) cases and partial concordance (1-level difference) in 21 (44%) cases. Only 3 (6%) cases showed major discordance. Overall, consistent classification (full + partial) was observed in 45 out of 48 (94%) samples (Table 5).

Table . Contingency table of effect size classification: TUR[a] versus TU[b] (qualitative categories).[c,d]

TUR effect size	TU effect size, n (%)				Total, n (%)
	Trivial	Small	Moderate	Large	
Trivial	10 (59)[e]	5 (19)[f]	1 (20)[g]	N/A[h]	16 (33)
Small	7 (41)[f]	13 (50)[e]	3 (60)[f]	N/A	23 (48)
Moderate	N/A	6 (23)[f]	1 (20)[e]	N/A	7 (15)
Large	N/A	2 (8)[g]	N/A	N/A	2 (4)
Total	17 (35)	26 (54)	5 (10)	N/A	48 (100)

[a]TUR: traditional umbrella review.
[b]TU: Umbrella Collaboration.
[c]Cramér *V* was the effect size measure.
[d]Pearson χ^{2}_{6}=11.03[i] (*P*=.09), *V*=0.339 (moderate).
[e]Full concordance.
[f]Partial concordance (discrepancy of 1 category).
[g]Discordance (discrepancy of more than 1 category).
[h]N/A: not applicable.
[i]Marginally significant (*P*<.10).

Table . Concordance levels in effect size classification: frequency and cumulative distribution.

Classification	Absolute frequency, n	Cumulative absolute frequency, n	Relative frequency, n/N (%)	Cumulative relative frequency, n/N (%)
Different effect size	24	24	24/48 (50)	24/48 (50)
Same effect size	24	48	24/48 (50)	48/48 (100)
Same effect size	24	24	24/48 (50)	24/48 (50)
One-level discrepancy	21	45	21/48 (44)	44/48 (94)
Discrepancy>1 level	3	48	3/48 (6)	48/48 (100)

Correlation of Quantitative Effect Sizes Between TU and TURs

A Spearman correlation assessed the relationship between TU and TUR ESs (R_{TU} vs Cohen *d*), yielding ρ=0.399 (*P*=.005), indicating a statistically significant, low-to-moderate positive correlation. The nonparametric approach was chosen to ensure robustness against outliers and deviations from normality, as both the TU and TUR ES distributions significantly deviated from normality according to the Shapiro-Wilk test (*W*=0.874 and 0.847, respectively; *P*<.001 for both cases). The nonparametric approach ensures robustness against outliers and nonnormality. A scatter plot with a regression line (Figure 4) confirmed this trend, although the explained variance was modest (R^2=0.111).

XSL•FO
RenderX

Figure 4. Correlation of quantitative effect sizes between Umbrella Collaboration and traditional umbrella reviews.

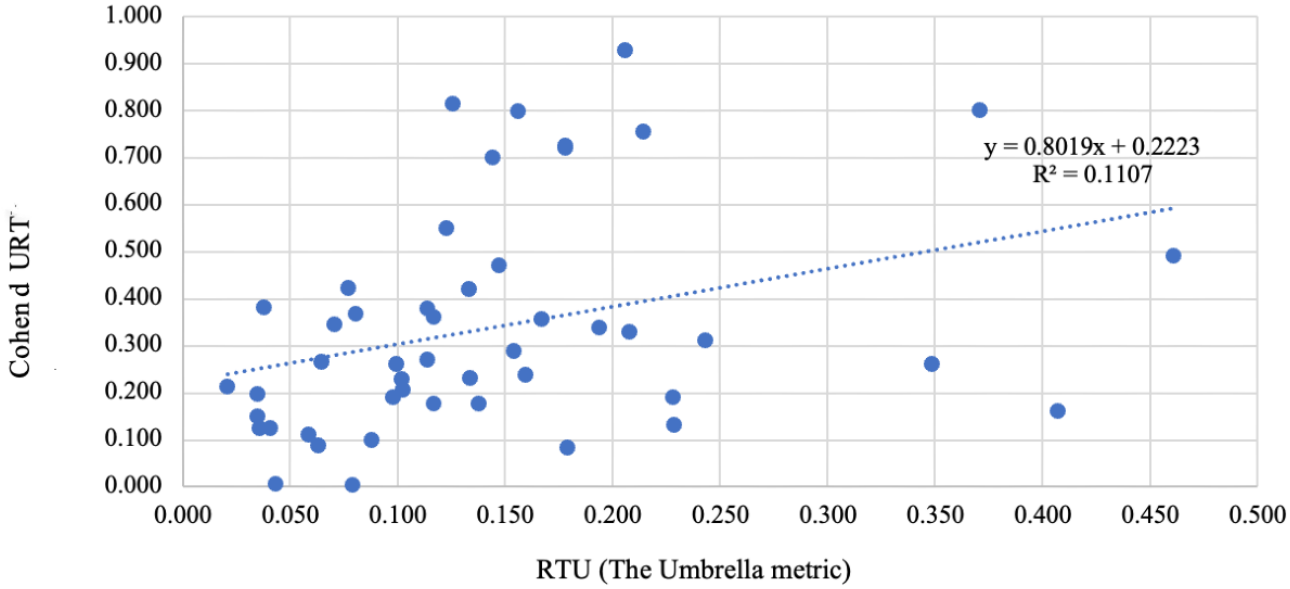

Concordance in Transformed Certainty Scores Between TU and TURs

We compared transformed certainty scores between TU and TURs using GRADE in 5 of the 8 included reviews, yielding 25 matched OoIs. Figure 5 illustrates the relationship between GRADE-based scores (x-axis) and TU's sentiment-based estimates (y-axis), both normalized to a 0 - 1 scale for direct comparison.

A Spearman correlation assessed the relationship between transformed certainty scores from TURs (GRADE) and TU (sentiment analysis). Due to nonnormal distributions of the transformed certainty scores, Shapiro-Wilk W=0.874 (P=.005) for TUR and W=0.928 (P=.049) for TU, a nonparametric test was used. A Spearman correlation between the transformed certainty scores assigned by TURs and TU yielded ρ=0.446 (P=.02), indicating a statistically significant, moderate positive association that supports concordance between both approaches.

Figure 5. Scatter plot of transformed certainty scores for outcomes of interest (OoIs; n=25) assessed by traditional umbrella reviews (TURs; Grading of Recommendations Assessment, Development, and Evaluation [GRADE]) and Umbrella Collaboration (TU; sentiment analysis [SA]).

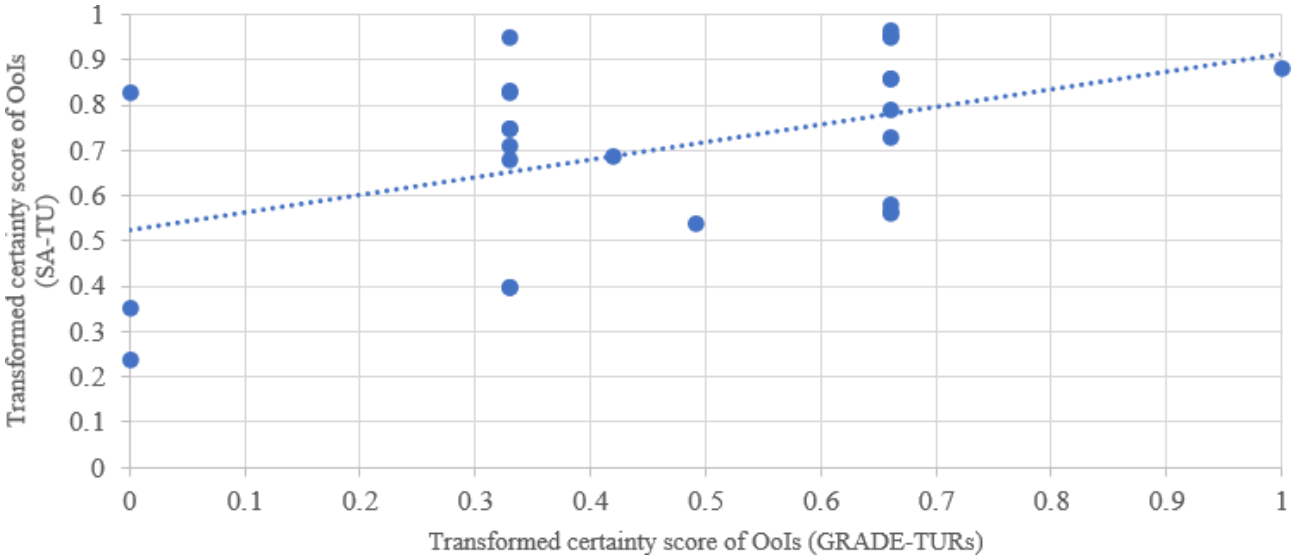

Execution Time of TU Projects

Execution time was assessed only for TU, as none of the selected TURs reported this metric. While TURs typically take 6 - 12 months, TU projects were completed in a mean time of 4 hours and 46 minutes (SD 2 hours and 30 minutes), with a median of 4 hours and 36 minutes. Completion times ranged from 1 hour and 34 minutes [28] to 10 hours and 8 minutes [24] (Table 6).

XSL•FO
RenderX

Table . Project execution time, search parameters, and reference overlap between TU[a] and TUR[b] methodologies.

Project	Search date	TU execution time (h:min:s)	Number of TU search terms, n	Number of references in TUR, n	Number of references in TU, n	Reference overlap, n/N (%)
Project 1: Veronese et al [22]	August 1, 2023	2:37:27	16	5	8	5/5 (100)
Project 2: Marx et al [27]	May 1, 2023	4:54:57	60	15	54	15/15 (100)
Project 3: Shen et al [23]	May 1, 2023	5:29:22	19	6	18	6/6 (100)
Project 4: Conneely et al [21]	May 1, 2023	5:37:14	6	16	23	16/16 (100)
Project 5: Gazzaniga et al [28]	May 1, 2023	1:34:37	15	11	19	10/11 (91)
Project 6: Musazadeh et al [24]	November 12, 2023	10:08:24	41	37	90	36/37(97)
Project 7: Veronese et al [25]	December 12, 2023	4:17:43	53	7	35	7/7 (100)
Project 8: Veronese et al [26]	December 12, 2023	3:33:54	16	8	22	8/9 (89)

[a]TU: Umbrella Collaboration.

[b]TUR: traditional umbrella review.

Discussion

Principal Findings

The findings of this study support the validity of the TU as a robust and efficient methodology for tertiary evidence synthesis. TU demonstrated a high level of concordance with TURs in identifying OoIs, successfully replicating the majority of those reported in TURs while also identifying a substantial number of additional outcomes. The experimental ES metric employed by TU (R_{TU}) showed a statistically significant correlation with standardized measures such as Cohen *d* and SMD, reinforcing its methodological soundness for estimating the magnitude of interventions or exposures. In addition, TU's automated, sentiment-based certainty assessment exhibited an acceptable level of agreement with GRADE, the most widely adopted system for evaluating certainty in evidence synthesis. One of the most notable advantages of TU lies in its execution time: while TURs typically require several months to complete, TU was able to generate comprehensive results within hours. This efficiency, combined with TU's demonstrated consistency and methodological coherence, highlights its potential value in settings where rapid evidence synthesis is essential or where resources are limited.

The development of TU aligns with the growing interest in AI-assisted tools for evidence synthesis, an emerging field likely to shape the future of scientific research. Rather than replacing TURs, TU is intended to complement them by offering a methodologically sound alternative in contexts where conventional approaches may be impractical or untimely. As a semiautomated platform, TU leverages NLP, sentiment analysis, and machine learning to streamline tertiary synthesis while maintaining methodological rigor. Despite its potential, TU represents a paradigm shift that may require time and educational efforts for widespread understanding and adoption. During the initial dissemination of this project, a general unfamiliarity with automated tertiary synthesis was observed, underscoring the need for clearer conceptual frameworks and greater engagement with the research community.

The identification of OoIs represents a central axis in the comparison between TU and TURs. In the context of tertiary evidence synthesis, an OoI reflects a specific finding previously evaluated in 1 or more SRs/MAs and integrated to answer a defined research question. The validity of a synthesis model depends not only on aggregating relevant data but also on faithfully capturing the main conclusions emerging from the primary literature.

Given the limited literature on tertiary synthesis, this analysis draws from well-established principles in secondary synthesis, where many of the same strengths and limitations apply. These include challenges in outcome selection, variability in inclusion criteria, and the balance between breadth and precision.

The comparative analysis revealed substantial differences in the number and breadth of OoIs identified. Across the 8 matched projects, TURs identified a total of 86 OoIs, whereas TU detected 410, yielding an identification gain factor of 4.77. TU replicated 73 out of 86 (85%) OoIs reported by TURs and contributed 337 additional OoIs. This difference does not appear to stem solely from broader search strategies or inclusion criteria. According to the Cochrane Handbook [29], umbrella reviews can pursue either broad or narrow questions [30]. While both methodologies support both approaches, several TURs in this study applied restrictive filters. For example, project 6 [24]

XSL•FO
RenderX

limited both its search strategy and inclusion criteria, while projects 3 [23] and 4 [21] also applied inclusion restrictions. Nevertheless, even when only unrestricted TURs are considered, TU maintained a gain factor of 4.93, suggesting an intrinsic property of TU in capturing a greater volume of outcomes.

One explanation lies in TU's automated and systematic extraction process, which minimizes human bias and ensures that any OoI mentioned in an abstract is captured, regardless of its perceived importance. By contrast, traditional reviews often apply subjective judgment in prioritizing outcomes, potentially omitting secondary findings. Moreover, TU is not constrained by human workload, allowing it to process larger volumes of data and extract multiple OoIs per review without sacrificing comprehensiveness.

A second consideration is the differing approach to quality assessment. TURs often prioritize high-quality SRs/MAs, limiting scope but potentially increasing rigor. TU, by relying on abstracts, does not perform formal quality appraisals. While this design choice increases the number of included studies, it should not be interpreted as a methodological flaw. Notably, up to 30% of SRs/MAs fail to report study quality assessments, and only 12.9% set quality thresholds for meta-analysis inclusion [31,32]. Furthermore, tools such as PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses)-Abstract could be integrated into future versions of TU to address this limitation, though prior research has shown that abstracts often score low on PRISMA quality indicators [33-36]. Nonetheless, TU consistently identifies more OoIs based solely on abstracts than TURs do using full-text reviews.

Another factor is the role of cognitive bias in outcome selection. Human reviewers may unconsciously favor familiar or hypothesis-confirming outcomes, limiting variability. While TU still involves human oversight in the final output selection, its core processes are software-driven, applying uniform parameters across all included studies and thereby reducing cognitive distortions. For instance, whereas TURs tended to report outcomes only in 1 direction, TU captured both positive and negative effects, aligning with best-practice guidelines from the Joanna Briggs Institute and the Cochrane Collaboration, which emphasize the importance of presenting balanced results for informed decision-making [29,37].

Finally, while it is possible that TU identifies outcomes of limited clinical relevance, this risk appears minimal. All OoIs extracted by TU are derived from published SRs/MAs indexed in MEDLINE, which are assumed to meet minimum standards of relevance and methodological quality. According to the Cochrane Handbook [30], reviews should focus on critical and important outcomes, excluding trivial ones.

While a higher number of outcomes may raise concerns about information overload, it is important to consider the primary purpose of tertiary synthesis: to provide a comprehensive and structured overview of the available evidence landscape. TU does not aim to guide individual clinical decisions but rather to inform stakeholders of the full scope of existing research. In this context, the inclusion of a larger set of outcomes enhances transparency and facilitates a more complete understanding of the literature. A formal qualitative appraisal of these additional outcomes, although beyond the scope of this study, represents a valuable direction for future research.

The study examined the concordance and correlation of ES metrics for OoIs between TU and TURs, using a unified set of outcomes assessed by both methodologies. TU incorporates a novel metric, R_{TU}, specifically developed to standardize and synthesize ESs from SRs/MAs within a tertiary synthesis framework. Unlike conventional metrics such as Cohen *d* or the SMD, R_{TU} was designed to allow direct aggregation of heterogeneous ES measures and requires empirical validation. This analysis represents the first formal evaluation of R_{TU} against traditional metrics.

In the qualitative comparison of ES categories (trivial, small, moderate, and large), TU demonstrated a high level of agreement with TURs. Full concordance, defined as an exact categorical match between the 2 methods, was observed in 24 out of 48 (50%) cases, while consistent concordance (full plus 1-level deviation) was reached in 45 out of 48 (94%) cases. Only 3 out of 48 (6%) cases showed discordance, meaning the categorical classifications differed by more than 1 level. These findings suggest that TU approximates the classification logic of TURs with a high degree of reliability, reinforcing the potential of R_{TU} to align with established interpretive frameworks.

The strength of the association between ES categorizations assigned by each methodology was evaluated using Cramér *V*, which yielded a value of 0.339. This indicates a moderate association and supports the notion that TU and TURs tend to classify the strength of associations in a comparable manner. Although the chi-square test did not reach statistical significance, likely due to the limited sample size and sparse contingency cells, Cramér *V* remains a valid ES measure and is less sensitive to these limitations [38]. In applied research, the strength of association can often be more informative than statistical significance, particularly when evaluating the practical utility of an experimental method [39]. It is worth noting that the statistical power of the chi-square test may have been reduced by the small number of observations and the presence of empty cells in the contingency table, both of which are known to limit test sensitivity. In line with recent critiques of overreliance on *P* values in biomedical research [40-43], the observed concordance rates and ES association offer more meaningful insights into the comparability of methods than statistical significance alone.

The quantitative comparison of ES values further supports the validity of TU. The Spearman correlation coefficient between TU and TUR values was positive and statistically significant, indicating a consistent directional relationship. Additionally, the slope of the regression line approached 1.0 (β=.8), suggesting that the R_{TU} metric yields ES of similar magnitude to those reported in TURs using standard metrics. This represents a marked improvement in proportionality compared with earlier regression slopes reported in other validation contexts.

However, the coefficient of determination (R^2) remained modest (R^2=0.11), indicating that a substantial portion of the variance

XSL•FO
RenderX

in TUR ES is not explained by TU estimates. This implies that, while TU captures general trends in effect magnitude, additional factors, such as differences in underlying metric types, weighting procedures, or reviewer judgments, may influence the exact values. It also highlights the limitations of linear models in fully explaining the relationship between these 2 methods.

These findings provide preliminary but encouraging evidence for the validity of R_{TU} as a useful metric for tertiary synthesis. Despite its experimental nature, R_{TU} appears to closely approximate established metrics such as Cohen *d* and the SMD in both categorical classification and magnitude. Future studies will be needed to further evaluate the sensitivity, specificity, and contextual performance of R_{TU} across diverse synthesis domains. Nonetheless, this initial validation suggests that TU can offer a consistent and efficient alternative for evaluating and categorizing ES in large-scale evidence synthesis.

While the observed correlation between TU and TUR ESs was moderate (ρ=0.399), this result is consistent with the structural differences between the R_{TU} metric and conventional measures such as Cohen *d* and SMD. R_{TU} is a proprietary, automated metric specifically developed for abstract-level synthesis, prioritizing scalability and consistency over statistical precision at the study level. It does not incorporate weighting by sample size or variance and relies on normalized effect descriptors extracted from secondary sources. These foundational differences partly account for the limited shared variance (R^2=0.11). However, in the context of tertiary synthesis, where the goal is to identify trends and prioritize outcomes across a broad evidence base, R_{TU} provides an acceptable and interpretable proxy. Moreover, this divergence has limited implications for clinical decision-making, as TU is designed to support synthesis and orientation at the field level, rather than to inform individual patient decisions. Future work may explore hybrid models that integrate additional contextual parameters to enhance concordance while preserving automation.

In this study, certainty assessments were compared exclusively among those TURs that applied the GRADE framework, a well-established and widely endorsed method for evaluating the certainty of evidence [44]. Although other appraisal tools exist, such as the statistical grading criteria proposed by Papatheodorou [45] or the methodological assessment framework AMSTAR 2 [46], these differ conceptually from the certainty model implemented in TU. Papatheodorou's system, despite its structure, has been criticized for relying on arbitrary statistical thresholds and for lacking sensitivity to clinical relevance and risk of bias [47]. AMSTAR 2, in turn, focuses on the methodological quality of the SRs/MAs themselves, rather than the certainty of individual outcomes. By contrast, TU's approach aligns more closely with GRADE principles by estimating certainty at the outcome level, albeit through an innovative method based on an automated sentiment analysis of SR/MA abstracts. Originally, the GRADE system referred to "quality of evidence," which denoted the degree of confidence in the validity of study findings to inform clinical decisions. Over time, this evolved to "confidence in estimates," and more recently, to "certainty of evidence," a term now widely accepted as it better reflects both the trust in effect estimates and their applicability in practice [20,48].

The comparison between transformed certainty scores, standardized to a 0 - 1 scale, for outcomes evaluated by TU and those rated using the GRADE framework provides important insights into the potential validity of TU's automated approach. The observed positive and statistically significant correlation suggests that TU is capable of approximating GRADE-based judgments, thereby reinforcing the conceptual alignment between the 2 methodologies. However, this result must be interpreted with caution. The number of matched outcomes was limited, which restricts generalizability and underscores the need for further validation in broader domains and larger datasets. Crucially, we recognize that TU's sentiment-based certainty estimation cannot replicate the multidimensional rigor of GRADE, which accounts for factors such as study limitations, inconsistency, indirectness, imprecision, and publication bias. The use of a general sentiment analysis model, originally trained on nonmedical texts, represents a methodological limitation. In this context, TU's certainty scoring should be viewed as a preliminary proxy, not a substitute for comprehensive manual assessments. Nevertheless, the approach offers a scalable and fully automated alternative capable of supporting rapid synthesis at low resource costs. Despite its current limitations, TU provides a complementary framework that aligns with the growing need for efficient, transparent, and reproducible tools in evidence synthesis. The correlation observed with GRADE invites further research into the integration of automated text-based analyses with established appraisal frameworks. As these technologies evolve, ongoing refinement and external validation will be critical to define their appropriate role in supporting scientific practice.

Execution time is a critical but often overlooked dimension in the evaluation of evidence synthesis methodologies. TURs rarely document their execution timelines, a pattern that was consistent among the TURs included in this study. By contrast, the TU methodology records execution time precisely, from the initiation of the first search term to the completion of the first version of the synthesis project, excluding any subsequent updates in this study. This level of detail in time tracking is unprecedented in both secondary and tertiary syntheses and represents a methodological innovation that may become increasingly relevant with the integration of automation tools. As a result, traditional timelines measured in months or years may soon be replaced by more accurate, time-stamped process metrics.

Given the lack of concrete time data from TURs, direct comparisons were not feasible. However, the execution times recorded for TU are sufficiently notable to highlight its efficiency. TU projects were completed within a range of 1.5-10 hours, substantially shorter than the typical duration for TURs, which is estimated to range between 6 and 12 months. This is also well below the average time frames reported for SRs/MAs, which commonly take 1-2 years [49], or for accelerated alternatives such as rapid reviews, which range from 5 to 12 weeks [50], and even the Two-Week SR approach, which targets completion in 11 working days [51]. The remarkable speed of TU is a key strength, particularly in light of the growing

XSL•FO
RenderX

international demand for agile evidence synthesis methods. Organizations such as the WHO [52], Cochrane Collaboration [53], and Joanna Briggs Institute [15] have emphasized the need for rapid yet valid synthesis approaches, an objective that TU appears to fulfill effectively in the context of tertiary reviews.

It is important to note that the time measured in this study refers solely to the initial execution phase of TU projects and does not account for later updates. However, TU includes a novel updating system that keeps projects active beyond their initial completion. The platform automatically checks daily for new SRs/MAs that match the predefined search criteria. When new studies are identified, a human reviewer is notified and can assess their eligibility. If approved, the system integrates the new results into the existing synthesis following revision and confirmation. While no equivalent system currently exists in traditional tertiary synthesis, the need for regular updates is well documented in evidence-based practice. In SRs/MAs, updates are typically recommended every 2 years or in response to emerging evidence, yet compliance with this standard is often poor [54-57]. TU's automated updating mechanism represents a major advantage in maintaining the relevance and accuracy of results, particularly in fast-evolving fields.

Notably, projects with a higher number of included references required longer execution times, which is expected given the additional workload involved in processing more SRs/MAs. Although TU significantly streamlines the synthesis process, the complexity and volume of information remain factors that influence total execution time, particularly when human oversight is still involved. Looking ahead, TU is being developed to support full automation of SR/MA data analysis, eliminating the need for human intervention. This transition will depend on training algorithms through semiautomated projects, progressively improving accuracy via machine learning.

Interestingly, the number of search terms used did not appear to correlate with execution time (unpublished data), suggesting that TU's algorithms can manage variable search complexity efficiently. In traditional methodologies, the search phase is often complex and conducted by specialized information professionals using advanced Boolean strategies across multiple databases [58-60]. By contrast, TU simplifies this process by using basic search terms exclusively in PubMed, automatically generated and refined through generative AI. This shift toward simplified, automated search strategies opens new avenues for exploring the adequacy and efficiency of noncomplex methods in evidence synthesis. It also raises important questions about the continued necessity of highly complex search protocols, suggesting that AI-assisted models may offer comparable effectiveness in certain contexts [61-63].

Ultimately, the short execution times observed with TU not only improve the efficiency of tertiary synthesis but also enhance the feasibility of rapidly delivering high-level evidence. In settings where timeliness is essential, this temporal advantage may outweigh minor trade-offs in other methodological dimensions. This capability holds particular promise for public health and health care decision-making, where fast access to up-to-date, high-quality evidence is crucial. TU, therefore, emerges as a promising tool in scenarios where speed is a priority, offering a pragmatic balance between rapid synthesis and analytical depth, while still requiring further validation against established gold standards.

Limitations

One key limitation is the small number of TURs included—only 8, randomly selected from 36 eligible reviews. Although intended to be representative, the lack of formal randomization introduces potential selection bias. The limited sample may not reflect the full methodological or topical diversity of umbrella reviews in geriatrics. However, all selected TURs were recent (2021 - 2023), which enhances comparability by aligning with current evidence synthesis standards. Another limitation is the lack of blinding: the lead investigator selected and analyzed the TURs, potentially introducing bias. Prior knowledge of TUR findings could have influenced TU outputs. However, the heterogeneity of outcomes across TURs reduces this risk, and TU's algorithm-driven, semiautomated workflow limits subjective influence. Results were largely determined by predefined rules, adding neutrality to the process.

The decision to utilize only 1 database, MEDLINE via PubMed, in TU is both a recognized limitation and a deliberate choice shaped by resource constraints and technical considerations. Although SRs typically require searching multiple databases to capture all relevant literature [64], our approach focuses on testing whether TU can achieve outcomes comparable to TURs despite this limitation. A pilot study provided preliminary evidence of PubMed's strong coverage, but further validation through TU is necessary to confirm its applicability across different domains. Although searching multiple databases is often recommended to reduce language and indexing biases, particularly regarding non-English literature [65,66], TU mitigates some of these biases by focusing on abstracts in English, as all PubMed abstracts are provided in English regardless of the original publication's language. Nonetheless, the absence of Chinese databases in our approach remains a notable limitation, given that only a small proportion of Chinese journals are indexed in MEDLINE.

Relying solely on abstracts in TU was a deliberate choice to test whether tertiary synthesis can be performed efficiently, with minimal resources, and without requiring advanced methodological expertise. While abstracts may omit key details (eg, certainty ratings, ES calculations, risk of bias), TU's structured, automated approach aims to generate clinically useful insights under standardized conditions. This approach offers 2 main advantages: (1) reduced language bias because all MEDLINE abstracts are in English; and (2) greater feasibility, as full-text access often requires costly subscriptions.

This pilot study focused on geriatrics to ensure methodological clarity and feasibility during initial validation. Limiting the scope allowed for controlled comparisons between TU and TURs. Future research should expand TU's application to other medical fields (eg, cardiology, psychiatry) and to nonmedical domains such as education or sociology, enabling stepwise validation before broader implementation.

XSL•FO
RenderX

Conclusions

This study provides preliminary evidence supporting the validity of the TU as a complementary, semiautomated tool for tertiary evidence synthesis. In a set of comparative projects, TU demonstrated a high level of concordance with TURs in identifying OoIs, estimating ESs, and assessing certainty, while offering substantially shorter execution times.

Although the methodology remains experimental and further validation is required across broader contexts, the results suggest that semiautomated approaches such as TU may represent a promising step toward more efficient, scalable, and continuously updatable models of evidence synthesis. TU does not aim to replace traditional methods but to provide a practical alternative in settings where time, resources, or responsiveness are critical.

Future studies with larger datasets, enhanced blinding procedures, and expanded topic areas will be essential to confirm these findings and further explore the potential role of TU in the evolving landscape of evidence-based research.

Acknowledgments

This project has received support from the Fundación Alfonso X el Sabio (grant 1.014.030).

Conflicts of Interest

BC, MRC, JP, AP, and BCR were involved in the development of The Umbrella Collaboration (TU) software. Additionally, BC and MRC are owners of the software and may have financial or intellectual property interests in its future applications. While this study was conducted with the aim of validating TU as an evidence synthesis tool, the authors have taken all necessary precautions to ensure that the study design, data analysis, and interpretation of results remain as objective and methodologically rigorous as possible. To mitigate potential biases, comparisons were conducted against gold-standard traditional umbrella reviews, and statistical methodologies were used to ensure transparency and replicability of findings. Furthermore, the affiliations of the authors did not influence the study design, execution, data analysis, or interpretation of results. All research activities were conducted following rigorous methodological standards to ensure impartiality and scientific integrity. The authors used generative artificial intelligence (ChatGPT; OpenAI Inc) as a language support tool during manuscript preparation. Specifically, it was used to assist in translating portions of the original manuscript, written by the authors in Spanish, into academic English, and to improve the clarity and flow of the text. No content, interpretations, or scientific claims were generated by artificial intelligence. All outputs were thoroughly fact-checked, revised, and approved by the authors. No references were generated by ChatGPT, and all cited sources were independently selected, verified, and reviewed by the authors.

Multimedia Appendix 1
Practical overview of The Umbrella Collaboration platform.
[DOCX File, 1426 KB - formative_v9i1e75215_app1.docx]

Multimedia Appendix 2
Traditional umbrella reviews (TURs) identified in the initial search and TURs selected for comparative analysis.
[DOCX File, 53 KB - formative_v9i1e75215_app2.docx]

References

1. Evidence synthesis for policy: a statement of principles - The Royal Society 2018. International Network for Governmental Science Advice. 2018. URL: https://tinyurl.com/3347c8k8 [accessed 2025-06-20]
2. World Health Organization (WHO). World health organization knowledge management strategy. WHO. 2005. URL: https://iris.who.int/bitstream/handle/10665/69119/WHO_EIP_KMS_2005.1.pdf [accessed 2025-06-30]
3. Cottrell E, Whitlock E, Kato E, et al. Defining the Benefits of Stakeholder Engagement in Systematic Reviews: Agency for Healthcare Research and Quality (US); 2014. URL: https://www.ncbi.nlm.nih.gov/books/NBK196180/
4. Sørensen K, Van den Broucke S, Fullam J, et al. Health literacy and public health: a systematic review and integration of definitions and models. BMC Public Health 2012 Jan 25;12:80. [doi: 10.1186/1471-2458-12-80] [Medline: 22276600]
5. Martínez-García JJ, Canizalez-Román A, Velázquez-Román JA, Flores-Villaseñor HM, León-Sicairos NM. Evaluación del conocimiento de métodos básicos de epidemiología e investigación en médicos residentes. Rev Médica Univ Autónoma Sinaloa REVMEDUAS 2021;11:115-123.
6. Baccolini V, Rosso A, Di Paolo C, et al. What is the prevalence of low health literacy in European Union member states? A systematic review and meta-analysis. J Gen Intern Med 2021 Mar;36(3):753-761. [doi: 10.1007/s11606-020-06407-8] [Medline: 33403622]
7. Baumann LA, Reinhold AK, Brütt AL. Public and patient involvement in health policy decision-making on the health system level - a scoping review. Health Policy 2022 Oct;126(10):1023-1038. [doi: 10.1016/j.healthpol.2022.07.007] [Medline: 35918211]

XSL•FO
RenderX

8. Biondi-Zoccai G, editor. Umbrella Reviews: Evidence Synthesis with Overviews of Reviews and Meta-Epidemiologic Studies: Springer International Publishing; 2016. [doi: 10.1007/978-3-319-25655-9]
9. Choi GJ, Kang H. Introduction to umbrella reviews as a useful evidence-based practice. J Lipid Atheroscler 2023 Jan;12(1):3-11. [doi: 10.12997/jla.2023.12.1.3] [Medline: 36761061]
10. Fusar-Poli P, Radua J. Ten simple rules for conducting umbrella reviews. Evid Based Mental Health 2018 Aug;21(3):95-100. [doi: 10.1136/ebmental-2018-300014]
11. Aromataris E, Fernandez R, Godfrey CM, Holly C, Khalil H, Tungpunkom P. Summarizing systematic reviews: methodological development, conduct and reporting of an umbrella review approach. JBI Evid Implement 2015;13:132. [doi: 10.1097/XEB.0000000000000055]
12. Cant R, Ryan C, Kelly MA. A nine‐step pathway to conduct an umbrella review of literature. Nurse Author Ed 2022 Jun;32(2):31-34. [doi: 10.1111/nae2.12039]
13. Belbasis L, Brooker RD, Zavalis E, Pezzullo AM, Axfors C, Ioannidis JP. Mapping and systematic appraisal of umbrella reviews in epidemiological research: a protocol for a meta-epidemiological study. Syst Rev 2023 Jul 14;12(1):123. [doi: 10.1186/s13643-023-02265-7] [Medline: 37452309]
14. Pollock M. Chapter V: overviews of reviews. In: Cochrane Handbook for Systematic Reviews of Interventions: The Cochrane Collaboration; 2023. URL: https://training.cochrane.org/handbook/current/chapter-v [accessed 2025-07-03]
15. Tricco AC, Straus SE, Ghaffar A, Langlois EV. Rapid reviews for health policy and systems decision-making: more important than ever before. Syst Rev 2022 Jul 30;11(1):153. [doi: 10.1186/s13643-022-01887-7] [Medline: 35906637]
16. Qureshi R, Shaughnessy D, Gill KAR, Robinson KA, Li T, Agai E. Are ChatGPT and large language models “the answer” to bringing us closer to systematic review automation? Syst Rev 2023 Apr 29;12(1):72. [doi: 10.1186/s13643-023-02243-z] [Medline: 37120563]
17. Teperikidis L, Boulmpou A, Papadopoulos C, Biondi-Zoccai G. Using ChatGPT to perform a systematic review: a tutorial. Minerva Cardiol Angiol 2024 Dec;72(6):547-567. [doi: 10.23736/S2724-5683.24.06568-2] [Medline: 39056432]
18. OpenAI. 2025. URL: https://openai.com/ [accessed 2025-04-02]
19. Elliott JH, Synnot A, Turner T, et al. Living systematic review: 1. Introduction-the why, what, when, and how. J Clin Epidemiol 2017 Nov;91:23-30. [doi: 10.1016/j.jclinepi.2017.08.010] [Medline: 28912002]
20. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008 Apr 26;336(7650):924-926. [doi: 10.1136/bmj.39489.470347.AD] [Medline: 18436948]
21. Conneely M, Leahy S, Dore L, et al. The effectiveness of interventions to reduce adverse outcomes among older adults following emergency department discharge: umbrella review. BMC Geriatr 2022 May 28;22(1):462. [doi: 10.1186/s12877-022-03007-5] [Medline: 35643453]
22. Veronese N, Honvo G, Bruyère O, et al. Knee osteoarthritis and adverse health outcomes: an umbrella review of meta-analyses of observational studies. Aging Clin Exp Res 2023 Feb;35(2):245-252. [doi: 10.1007/s40520-022-02289-4] [Medline: 36331799]
23. Shen Y, Liu D, Li S, et al. Effects of exercise on patients important outcomes in older people with sarcopenia: an umbrella review of meta-analyses of randomized controlled trials. Front Med (Lausanne) 2022;9:811746. [doi: 10.3389/fmed.2022.811746] [Medline: 35186999]
24. Musazadeh V, Kavyani Z, Mirhosseini N, Dehghan P, Vajdi M. Effect of vitamin D supplementation on type 2 diabetes biomarkers: an umbrella of interventional meta-analyses. Diabetol Metab Syndr 2023 Apr 19;15(1):76. [doi: 10.1186/s13098-023-01010-3] [Medline: 37072813]
25. Veronese N, Galvano D, D’Antiga F, et al. Interventions for reducing loneliness: an umbrella review of intervention studies. Health Soc Care Community 2021 Sep;29(5):e89-e96. [doi: 10.1111/hsc.13248] [Medline: 33278311]
26. Veronese N, Smith L, Bolzetta F, Cester A, Demurtas J, Punzi L. Efficacy of conservative treatments for hand osteoarthritis. Wien Klin Wochenschr 2021 Mar;133(5-6):234-240. [doi: 10.1007/s00508-020-01702-0]
27. Marx W, Veronese N, Kelly JT, et al. The dietary inflammatory index and human health: an umbrella review of meta-analyses of observational studies. Adv Nutr 2021 Oct 1;12(5):1681-1690. [doi: 10.1093/advances/nmab037] [Medline: 33873204]
28. Gazzaniga G, Menichelli D, Scaglione F, Farcomeni A, Pani A, Pastori D. Effect of digoxin on all-cause and cardiovascular mortality in patients with atrial fibrillation with and without heart failure: an umbrella review of systematic reviews and 12 meta-analyses. Eur J Clin Pharmacol 2023 Apr;79(4):473-483. [doi: 10.1007/s00228-023-03470-y] [Medline: 36872367]
29. Higgins J. Cochrane Handbook for Systematic Reviews of Interventions 2023: The Cochrane Collaboration; 2023. URL: https://training.cochrane.org/handbook [accessed 2025-07-03]
30. McKenzie JE, Brennan SE. Overviews of systematic reviews: great promise, greater challenge. Syst Rev 2017 Sep 8;6(1):185. [doi: 10.1186/s13643-017-0582-8] [Medline: 28886726]
31. Luchini C, Veronese N, Nottegar A, et al. Assessing the quality of studies in meta-research: review/guidelines on the most important quality assessment tools. Pharm Stat 2021 Jan;20(1):185-195. [doi: 10.1002/pst.2068] [Medline: 32935459]
32. Seehra J, Pandis N, Koletsi D, Fleming PS. Use of quality assessment tools in systematic reviews was varied and inconsistent. J Clin Epidemiol 2016 Jan;69:179-184. [doi: 10.1016/j.jclinepi.2015.06.023] [Medline: 26151664]
33. Bigna JJR, Um LN, Nansseu JRN. A comparison of quality of abstracts of systematic reviews including meta-analysis of randomized controlled trials in high-impact general medicine journals before and after the publication of PRISMA extension

XSL•FO
RenderX

for abstracts: a systematic review and meta-analysis. Syst Rev 2016 Oct 13;5(1):174. [doi: 10.1186/s13643-016-0356-8] [Medline: 27737710]
34. Li T, Hua F, Dan S, Zhong Y, Levey C, Song Y. Reporting quality of systematic review abstracts in operative dentistry: an assessment using the PRISMA for Abstracts guidelines. J Dent 2020 Nov;102:103471. [doi: 10.1016/j.jdent.2020.103471] [Medline: 32931892]
35. El Ansari W, AlRumaihi K, El-Ansari K, et al. Reporting quality of abstracts of systematic reviews/meta-analyses: an appraisal of Arab Journal of Urology across 12 years: the PRISMA-Abstracts checklist. Arab J Urol 2022 Aug 22;21(1):52-65. [doi: 10.1080/2090598X.2022.2113127] [Medline: 36818377]
36. Zhong Y, Wang Y, Dan S, et al. The reporting quality of systematic review abstracts in leading general dental journals: a methodological study. J Evid Based Dent Pract 2023 Mar;23(1):101831. [doi: 10.1016/j.jebdp.2022.101831] [Medline: 36914298]
37. Aromataris E, Lockwood C, Porritt K, Pilla B. JBI Manual for Evidence Synthesis. JBI Global Wiki. URL: https://jbi-global-wiki.refined.site/space/MANUAL [accessed 2024-06-05]
38. Brzezińska J. The problem of zero cells in the analysis of contingency tables. KREM 2015;5(941):49-61. [doi: 10.15678/ZNUEK.2015.0941.0504]
39. Sapra RL, Saluja S. Understanding statistical association and correlation. Current Medicine Research and Practice 2021;11(1):31-38. [doi: 10.4103/cmrp.cmrp_62_20]
40. Sullivan GM, Feinn R. Using effect size-or why the P value is not enough. J Grad Med Educ 2012 Sep;4(3):279-282. [doi: 10.4300/JGME-D-12-00156.1] [Medline: 23997866]
41. Demidenko E. The P-value you can't buy. Am Stat 2016;70:33-38. [doi: 10.1080/00031305.2015.1069760]
42. Greenland S. Valid P-values behave exactly as they should: some misleading criticisms of P-values and their resolution with S-values. Am Stat 2019;73:106-114. [doi: 10.1080/00031305.2018.1529625]
43. Lu Y, Belitskaya-Levy I. The debate about P-values. Shanghai Arch Psychiatry 2015 Dec 25;27(6):381-385. [doi: 10.11919/j.issn.1002-0829.216027] [Medline: 27199532]
44. Franco JVA, Arancibia M, Meza N, Madrid E, Kopitowski K. Clinical practice guidelines: concepts, limitations and challenges. Medwave 2020 Apr 30;20(3):e7887. [doi: 10.5867/medwave.2020.03.7887] [Medline: 32428925]
45. Papatheodorou S. Umbrella reviews: what they are and why we need them. Eur J Epidemiol 2019 Jun;34(6):543-546. [doi: 10.1007/s10654-019-00505-6] [Medline: 30852716]
46. Shea BJ, Reeves BC, Wells G, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ 2017 Sep 21;358:j4008. [doi: 10.1136/bmj.j4008] [Medline: 28935701]
47. Schlesinger S, Schwingshackl L, Neuenschwander M, Barbaresko J. A critical reflection on the grading of the certainty of evidence in umbrella reviews. Eur J Epidemiol 2019 Sep;34(9):889-890. [doi: 10.1007/s10654-019-00531-4] [Medline: 31222608]
48. Hultcrantz M, Rind D, Akl EA, et al. The GRADE Working Group clarifies the construct of certainty of evidence. J Clin Epidemiol 2017 Jul;87:4-13. [doi: 10.1016/j.jclinepi.2017.05.006] [Medline: 28529184]
49. Borah R, Brown AW, Capers PL, Kaiser KA. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. BMJ Open 2017 Feb 27;7(2):e012545. [doi: 10.1136/bmjopen-2016-012545] [Medline: 28242767]
50. Hamel C, Michaud A, Thuku M, et al. Defining rapid reviews: a systematic scoping review and thematic analysis of definitions and defining characteristics of rapid reviews. J Clin Epidemiol 2021 Jan;129:74-85. [doi: 10.1016/j.jclinepi.2020.09.041] [Medline: 33038541]
51. Clark J, Glasziou P, Del Mar C, Bannach-Brown A, Stehlik P, Scott AM. A full systematic review was completed in 2 weeks using automation tools: a case study. J Clin Epidemiol 2020 May;121:81-90. [doi: 10.1016/j.jclinepi.2020.01.008] [Medline: 32004673]
52. Tricco AC, EtienneV L, Straus SE. Rapid Reviews to Strengthen Health Policy and Systems: A Practical Guide: World Health Organization; 2017.
53. Garritty C, Gartlehner G, Nussbaumer-Streit B, et al. Cochrane Rapid Reviews Methods Group offers evidence-informed guidance to conduct rapid reviews. J Clin Epidemiol 2021 Feb;130:13-22. [doi: 10.1016/j.jclinepi.2020.10.007] [Medline: 33068715]
54. Garner P, Hopewell S, Chandler J, et al. When and how to update systematic reviews: consensus and checklist. BMJ 2016 Jul 20;354:i3507. [doi: 10.1136/bmj.i3507] [Medline: 27443385]
55. Hoffmeyer B, Fonnes S, Andresen K, Rosenberg J. Use of inactive Cochrane reviews in academia: a citation analysis. Scientometrics 2023 May;128(5):2923-2934. [doi: 10.1007/s11192-023-04691-9]
56. Beller EM, Chen JH, Wang UH, Glasziou PP. Are systematic reviews up-to-date at the time of publication? Syst Rev 2013 May 28;2:36. [doi: 10.1186/2046-4053-2-36] [Medline: 23714302]
57. Bashir R, Surian D, Dunn AG. Time-to-update of systematic reviews relative to the availability of new evidence. Syst Rev 2018 Nov 17;7(1):195. [doi: 10.1186/s13643-018-0856-9] [Medline: 30447694]

XSL•FO
RenderX

58. Heinen L, Goossen K, Lunny C, Hirt J, Puljak L, Pieper D. The optimal approach for retrieving systematic reviews was achieved when searching MEDLINE and Epistemonikos in addition to reference checking: a methodological validation study. BMC Med Res Methodol 2024 Nov 9;24(1):271. [doi: 10.1186/s12874-024-02384-2] [Medline: 39522026]
59. Bramer WM, Rethlefsen ML, Kleijnen J, Franco OH. Optimal database combinations for literature searches in systematic reviews: a prospective exploratory study. Syst Rev 2017 Dec 6;6(1):245. [doi: 10.1186/s13643-017-0644-y] [Medline: 29208034]
60. Aromataris E, Riitano D. Systematic reviews: constructing a search strategy and searching for evidence. Am J Nurs 2014;114:49. [doi: 10.1097/01.NAJ.0000446779.99522.f6] [Medline: 24759479]
61. Yu X, Wu S, Sun Y, et al. Exploring the diverse definitions of “evidence”: a scoping review. BMJ Evid Based Med 2024 Jan 19;29(1):37-43. [doi: 10.1136/bmjebm-2023-112355] [Medline: 37940419]
62. Guimarães NS, Joviano-Santos JV, Reis MG, Chaves RRM, Observatory of Epidemiology, Nutrition, Health Research (OPENS). Development of search strategies for systematic reviews in health using ChatGPT: a critical analysis. J Transl Med 2024 Jan 2;22(1):1. [doi: 10.1186/s12967-023-04371-5] [Medline: 38167166]
63. Ge L, Agrawal R, Singer M, et al. Leveraging artificial intelligence to enhance systematic reviews in health research: advanced tools and challenges. Syst Rev 2024 Oct 25;13(1):269. [doi: 10.1186/s13643-024-02682-2] [Medline: 39456077]
64. Gusenbauer M, Haddaway NR. Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. Res Synth Methods 2020 Mar;11(2):181-217. [doi: 10.1002/jrsm.1378] [Medline: 31614060]
65. Jia Y, Huang D, Wen J, et al. Assessment of language and indexing biases among Chinese-sponsored randomized clinical trials. JAMA Netw Open 2020 May 1;3(5):e205894. [doi: 10.1001/jamanetworkopen.2020.5894] [Medline: 32463469]
66. Mao C, Li M. Language bias among Chinese-sponsored randomized clinical trials in systematic reviews and meta-analyses-can anything be done? JAMA Netw Open 2020 May 1;3(5):e206370. [doi: 10.1001/jamanetworkopen.2020.6370] [Medline: 32463464]

Abbreviations

AI: artificial intelligence
ES: effect size
GRADE: Grading of Recommendations Assessment, Development, and Evaluation
LLM: large language model
MA: meta-analysis
NLP: natural language processing
OoI: outcome of interest
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
SMD: standardized mean difference
SR: systematic review
TU: Umbrella Collaboration
TUR: traditional umbrella review

Edited by A Mavragani; submitted 30.03.25; peer-reviewed by C Onah; revised version received 15.05.25; accepted 21.05.25; published 08.07.25.

Please cite as:
Carrillo B, Rubinos-Cuadrado M, Parellada J, Palacios A, Carrillo-Rubinos B, Canillas F, Baztán Cortés JJ, Gómez-Pavón J
Validation of The Umbrella Collaboration for Tertiary Evidence Synthesis in Geriatrics: Mixed Methods Study
JMIR Form Res 2025;9:e75215
URL: https://formative.jmir.org/2025/1/e75215
doi:10.2196/75215

© Beltran Carrillo, Marta Rubinos-Cuadrado, Jazmin Parellada, Alejandra Palacios, Beltran Carrillo-Rubinos, Fernando Canillas, Juan José Baztán Cortés, Javier Gómez-Pavón. Originally published in JMIR Formative Research (https://formative.jmir.org), 8.7.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Determining the Requirements of Vulnerable Groups for Health Counseling and Optimizing the Evaluation of Health Consultations: Mixed Methods Study With the Use of AI

Annina Boehm-Fischer, DrPH, Dipl-Psych; Luzi M Beyer, Dr rer nat

Alice-Salomon-Hochschule Berlin, Alice-Samolon-Platz 5, Berlin, Germany

Corresponding Author:
Annina Boehm-Fischer, DrPH, Dipl-Psych
Alice-Salomon-Hochschule Berlin, Alice-Samolon-Platz 5, Berlin, Germany

Abstract

Background: Evaluating health counseling services is crucial for ensuring their quality and effectiveness. However, this process is hampered by challenges such as language barriers and limited awareness of their needs and concerns.

Objective: The studies aimed to enhance and digitize an existing paper-and-pencil evaluation form for a health counseling service while gaining insights into client needs and barriers. This effort intends to adapt a health care facility's offerings to better meet client demands and implement a multilingual format for greater accessibility.

Methods: The research team designed and conducted an in-depth interview study with clients of a health counseling service to gather new information. The insights regarding client demands, wishes, and social needs were used to revise and supplement the existing 1-page questionnaire (originally in German) for evaluating counseling sessions. Using artificial intelligence, the team transformed the new 3-page questionnaire into easy language with a Kunin smiley scale, translated it into 7 other languages, and created audio recordings for all questions in each language. The questionnaire was then programmed into an web-based tool, allowing data collection both on-site with tablets and through integration into the counseling service's website. This digital format is now continuously used to adapt the counseling service to clients' needs.

Results: A total of 18 clients participated in the in-depth interviews, which were conducted in their native languages whenever possible and lasted between 8 and 30 minutes. The results indicated that many clients attending the counseling center are burdened by physical and mental health issues, with a significant portion of the assistance provided focused on helping clients complete various forms required by health insurance providers and medical professionals. Despite these challenges, clients expressed a high level of satisfaction with the health counseling services they received. The revised and supplemented web-based questionnaire has been completed by 41 clients. Evaluation results revealed that only 21 respondents (51%) filled out the questionnaire in the national language (German), while English and Arabic were the next most common choices, each used by 6 clients (15%). Findings regarding health burdens and the need for assistance were reaffirmed, highlighting that clients' self-perception regarding their ability for self-help is notably low.

Conclusions: Contrary to previous assumptions, it was found that client interests predominantly lie in receiving help with the excessive demands imposed by institutional forms and requirements rather than solely addressing health issues. Clients showed strong satisfaction with the advice received and emphasized the necessity for multilingual health counseling services and evaluations. There is a distinct need for support in completing forms for doctors and health insurance applications. In addition, many clients expressed a lack of confidence in managing health care processes independently in the future, underscoring the need for greater awareness of available resources and support networks.

(JMIR Form Res 2025;9:e68888) doi:10.2196/68888

KEYWORDS

health counseling service; health service evaluation; immigrant health; mixed methods; online questionnaire; vulnerable populations, in-depth interviews

Introduction

The German health care system, despite being one of the most expensive globally, offers extensive services, high-quality standards, low access barriers, and income-independent health care [1]. In theory, this should ensure excellent care for all. However, in practice, barriers exist that limit access to health care services. Non-native speakers and individuals not born in Germany often lack information about health services and face language barriers [2,3]. Their utilization rates of health care services during illness are significantly lower compared with other European countries [4], and they struggle with

XSL•FO
RenderX

understanding and processing health-related information [5]. In addition, preventive health services in Germany are less frequently used by elderly individuals, those living alone, and people requiring informational support [6,7]. A representative survey indicates that less than half of older adults in Germany (aged 75 years and older) have internet access, and of those, only about half (thus, about a quarter overall) use the internet to search for health-related topics [8]. A trend emerges in this demographic: lower education levels, smaller social networks, and reduced quality of life correlate with less internet usage. Furthermore, individuals with lower household incomes and educational levels are less likely to use digital health technologies in Germany [9]. As a result, vulnerable groups in Germany are just as dependent on high-quality health counseling services (HCSs) as in other countries [10-16].

Health counseling plays a critical role in improving public health, particularly for vulnerable groups, making the evaluation of these services essential for ensuring quality. Research on the effectiveness of HCS has been conducted for over 50 years [17]. Early studies primarily focused on procedural and outcome assessments with limited client involvement [18]. More recent evaluations have shifted toward a client-centered approach [19,20], acknowledging that actively involving clients in the design of health consultation assessments yields valuable insights that differ from those of health care professionals [21]. This participatory approach ensures that specific client needs are adequately focused on [22].

Qualitative research methods are particularly useful for capturing opinions, social needs, and desires. However, web-based qualitative studies present challenges regarding accessibility and engagement [23]. Therefore, an in-depth face-to-face interview study is recommended as an initial approach. This method is particularly suited for discussing health questions and interactions with the health care system [24].

For the revised version of the HCS evaluation form, an web-based questionnaire with a smiley response scale (Kunin scale) is effective. This format enables multilingual adaptation and audio-assisted completion, ensuring accessibility regardless of language barriers. It can be completed independently of time and location, reducing the risk of socially desirable responses [25], while allowing a greater flexibility in participation.

With the aim of customizing the evaluation form and involving the clients, we conducted a 2-stage mixed methods study (see Figure 1).

Figure 1. Flowchart of study design in study. The planned duration was 12 months (April 2023 to March 2024). Based on the interview study with clients of a health counseling service, Berlin (Germany), the existing German evaluation form was to be expanded, improved, and digitally implemented in several languages in order to better meet the needs of the clients. After the subsequent quantitative studies (and any necessary adjustments), the evaluation form should be handed over to the health counseling service.

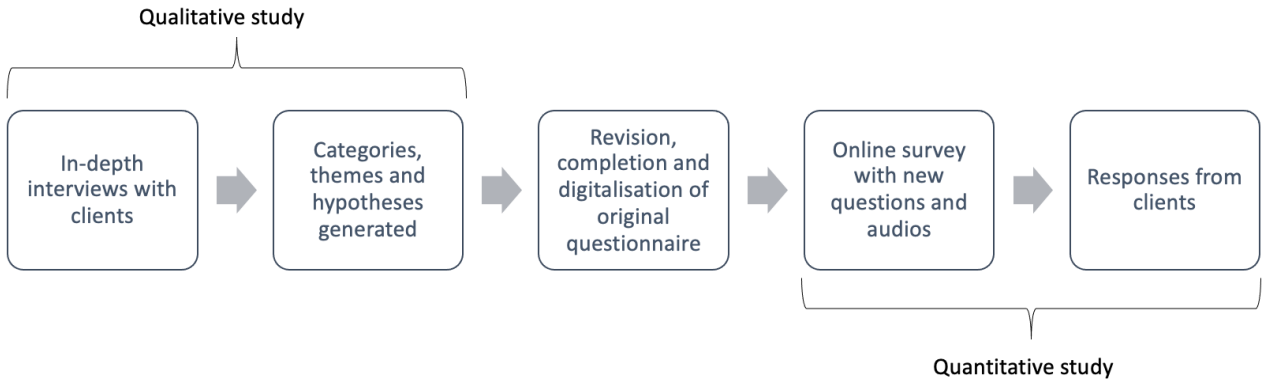

The clients of the health care facility have diverse native languages, some are illiterate, and many have a combination of mental and physical impairments. To be able to address them, this study followed a mixed methods approach consisting of 2 phases: a qualitative in-depth interview survey and a subsequent quantitative web-based survey with the following research questions (see Textbox 1).

XSL•FO
RenderX

Textbox 1. Research questions.

Qualitative in-depth interview survey:

- What are typical consulting contents and types of services provided?
- What quality characteristics are mentioned by users in relation to the HCS (eg, length of sessions, perceived professionalism, feeling supported)?
- What quality characteristics do users mention in relation to the staff (eg, goal-oriented, understanding)?
- What additional needs are mentioned?
- Is the service perceived as recommendable?

Quantitative web-based survey study:

- Which languages are preferred when completing the form?
- From whom are advice seekers sent and what are the reasons for using the advice center?
- What is the subjective perception of the users with regard to the quality characteristics determined by users in the qualitative study in relation to the service and staff?

Methods

Overview

A mixed methods approach was selected to gain an in-depth understanding of clients' challenges (qualitative), which served as the foundation for developing relevant questions and response categories for the questionnaire (quantitative), ensuring both relevance and broad applicability.

Ethical Considerations

Before the study was conducted, an internal ethical review was conducted at the Alice Salomon University of Applied Sciences by an independent research ethics committee, based on the guidelines of the German Ethics Council. There was a consensus discussion regarding the ethics review and implementation guidelines with the staff and supervisors of the HCS. It was determined that the interviews with clients would only address topics that are also discussed during the counseling sessions. It was agreed that all interviews would take place in the counseling rooms and that a staff member would always be present in other rooms to provide support as needed. The interviews would be fully anonymized to protect privacy and confidentiality, no video recordings would be made, and the audio would not be made public; only the transcripts would be shared. Under these conditions, the consensus discussion concluded that there were no concerns. The result of the ethical evaluation was that the research ethics criteria for the study were satisfactorily met and the study may be conducted under the conditions.

Informed consent was obtained from all interviewees, which also permitted secondary analysis. No demographic information was requested, and any personal information was anonymized during transcription (eg, when discussing individual staff members, names were replaced with XXX).

The interviewees did not receive any financial or other compensation for their participation in the interviews.

Study 1: Qualitative In-Depth Interview Study

Interview Development and Procedure

Before the qualitative study, a workshop was held with the HCS team to exchange thoughts regarding objectives, desires, and methodologies. During this workshop, the original evaluation questionnaire, which comprised 7 questions, was reviewed.

A literature search on evaluation questionnaires used in health counseling was conducted, and a draft for the interview was developed. This draft was subsequently refined in a workshop, during which the questions were organized. The final interview composed of 8 primary questions and an additional 14 probing questions aimed at encouraging participants to elaborate on their responses.

The final interview guide included 8 primary questions and 14 probing questions. Interviews were conducted in German, English, or in participants' native languages to ensure clarity and accuracy. Interviewers were multilingual, allowing direct communication or real-time translation when necessary. Interviews were recorded as MP3 files and transcribed using artificial intelligence (AI)–based transcription software (noScribe). This software was selected due to its multilingual support and its capacity to run locally on a personal computer, thereby circumventing the need to upload audio files to the internet. Following the transcription process, the texts were compared with the audio files by native speakers of the respective languages. Necessary corrections were made, after which the responses were translated into German using an AI-assisted translation tool (DeepL). The translated responses were subjected to a thorough review to ensure content accuracy.

Subsequently, all responses to the individual questions were analyzed separately in a workshop setting, and where feasible, an inductive category system was developed according to Mayring's methodology [26].

For the responses to the first question, "How long have you been receiving counseling?" and its corresponding probing questions ("How often have you attended counseling since then?" and "What was your experience in scheduling an appointment?"), no category system was established due to the nature of the responses. For the second question, "What health topics have you sought counseling for?" the response categories were deductively derived from the answers and included physical health, mental health, other topics, and categories related to filling out applications, counseling, and referrals.

XSL•FO
RenderX

Regarding the third question, “Was the time sufficient?” and the two probing questions (“Were you able to discuss all health issues?” and “Did you feel well supported?”), no category system was created, as the responses were too homogenous.

For questions 4 (“How did you experience the counseling?“) and 5 (“How did you perceive the staff?”) along with their corresponding probing questions, a unified category system was developed, since many participants did not differentiate between the counseling service and the personnel involved. The categories established were: statements regarding atmosphere, outcome-orientedness and objective-orientedness of the counselors, counseling demeanor and conduct, expressed trust and security, and expressed difficulties.

For the sixth question (“How do you feel after the consultation?”) and the first exploratory question (“To what extent did the health advice help?”), a category system was developed with the categories “Mentioned emotions,” “Statements about stress” and “Other.”

For the second probing question regarding the sixth inquiry (“In what areas do you seek further support?”), a category system was deductively created based on the responses, including categories such as physical and mental health, assistance with applications, case management services, language acquisition, housing search, job search, and networking.

For the seventh question (“Will you be able to cope with issues independently in the future?”), as well as its corresponding probing question (“Do you now know where and how to seek further help?”) an inductive category system was developed with the categories “Need for help due to chronic mental/physical health problems,” “Overburdened with structures in the healthcare system,” “Language barriers (eg, at the doctor’s or hospital),” “Barriers due to lack of reading/spelling skills,” “No other help (caregiver, other advice center) known” and “Other.” No category system was developed for the eighth question (“Would you recommend health counseling?”) and the 3 following probing questions (“How well do you feel supported?,” “What further health counseling would you like?,” and “Is there anything else you would like to tell us?”), in order to do justice to breadth, individuality, and nonspecificity of the answers.

Data Collection

The primary objective of the recruitment process for the interview study was to enlist a total of 12 individuals identified as “clients” (individuals who had attended counseling in person at least twice, for whom contact information was available, and who expressed a desire for further consultations), along with 6 individuals categorized under “case management” (those requiring on-site consultations coupled with assistance in navigating administrative and medical appointments, which may include home visits), and 6 individuals seeking walk-in consultations who were attending for the first time and preferred to remain anonymous. This recruitment goal was established to reflect the approximate distribution of service recipients within the counseling framework.

For reasons of practicability, the recruitment procedure was as follows: The external interviewers told the health counseling staff when they would have time to conduct interviews (July 3 - 7, 2023). They then asked already scheduled clients (clients and case management) whether they could come earlier or still have time for an interview after their appointment (telephone contact). Through this process, 12 clients and the 6 case management participants were contacted and all those contacted agreed to participate. In the end, 10 clients and the 6 case management participants took part in the planned sessions. While the interviewers were on site, all walk-in clients were asked if they had spontaneous time for an interview. Of the 6 people approached who wanted to take advantage of counseling, 2 agreed to participate, while the remaining 4 declined. The inclusion criterion was that they had to have already received counseling at least once. All potential participants received information and two written informed consent forms before the interview, one of which they had to sign and return. However, no employees were present during the interviews in order to avoid potential conflicts of interest on the part of the interviewees.

After the audio recordings were made (before coding), 1 individual categorized under case management was excluded from the analysis, as this participant provided extensive narratives but failed to respond to any of the posed questions. Subsequently, another case management participant was added to the sample as a compensatory inclusion.

Thus, the final sample for the interview study comprised 10 clients, 6 case management participants, and 2 walk-in consultation seekers (n=18 of 24 asked persons, 75% response rate). The data collection, coding, and the data analysis was completed in July 2023. A total of 10 people (the authors and 8 students) were involved in the data collection, coding, and analysis.

Data Analysis

After coding the transcripts of the interview study with QCAmap [26,27], a qualitative content analysis according to Mayring [26] was carried out.

For the responses to the first question, “How long have you been receiving counseling?” and its corresponding probing questions (“How often have you attended counseling since then?” and “What was your experience in scheduling an appointment?“), no category system was established due to the nature of the responses. For the second question, “What health topics have you sought counseling for?” the response categories were deductively derived from the answers and included physical health, mental health, other topics, and categories related to filling out applications, counseling, and referrals.

Further analysis included questions on session duration, counseling experience, perception of staff, and additional support needs, ensuring that both homogeneous and diverse responses were categorized effectively. These categories were then refined through collaborative coding sessions before finalizing the thematic framework.

Data Exclusion

One interview from the in-depth interview study was completely excluded from analysis due to the absence of any substantively

XSL•FO
RenderX

evaluable responses to the questions posed. Among the interviewees, there were also 2 people who were unable to answer some of the questions even after repeated questioning, rephrasing, and shortening, but spoke at times in a very repetitive, over-excited and unfocused manner. Some responses from these interviews were excluded.

Study 2: Quantitative Web-Based Survey Study

The insights gained from the qualitative interview study were used to develop a structured questionnaire.

Questionnaire Development and Implementation

Before the quantitative study, a second workshop was conducted with the HCS team to discuss the outcomes of the interview study and to coordinate forthcoming actions. The responses from the interview study, along with the original paper-and-pencil questionnaire, were critically examined. Subsequently, an additional literature search was carried out.

The questionnaire was developed in German in August, after which it was initially transformed into a simplified language format using artificial intelligence (ChatGPT). Following this transformation, the content was translated into English using an AI-based tool (DeepL), since translations from English into other languages (eg, Russian, Arabic) tend to yield better results than translations from German. In September, the English questionnaire was translated into 6 additional languages (Arabic, French, Russian, Romanian, Spanish, and Turkish), and it was cross-checked by native speakers. This process resulted in the questionnaire being available in a total of 8 languages.

Subsequently, audio files for all questions in all languages were generated using AI (NVIDIA Jarvis), resulting in a total of 146 questions (18 questions multiplied by 8 languages plus language selection and consent to data processing). The audio recordings were reviewed by native speakers in October, and edits were made as necessary. The finalized audio files, along with the questions, were then programmed into a web-based tool (SoSciSurvey, see Figure 2). Functionality tests were subsequently conducted to ensure the operability of the survey.

Figure 2. The 2 questions from the English questionnaire of the quantitative study in which there were 8 content-identical language translations. Under each question is a read-aloud option and all answers are as smiley faces so that questions can be answered without reading skills.

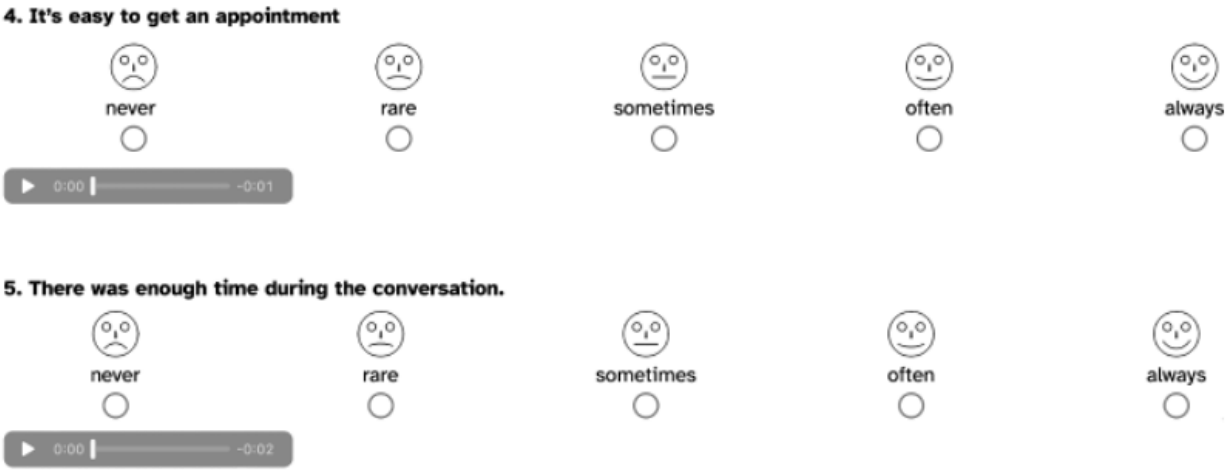

The online questionnaire included a pre-question for consent regarding the processing of collected data, a question regarding language selection, and 18 substantive questions on 3 pages as well as a final page with a thank-you note (see Figure 3). Of the 7 questions from the original paper-and-pencil questionnaire, 5 were modified and expanded to include 8 questions. A total of 10 questions were derived from the responses in the interview study.

XSL•FO
RenderX

Figure 3. Quantitative Online-Questionnaire flow chart.

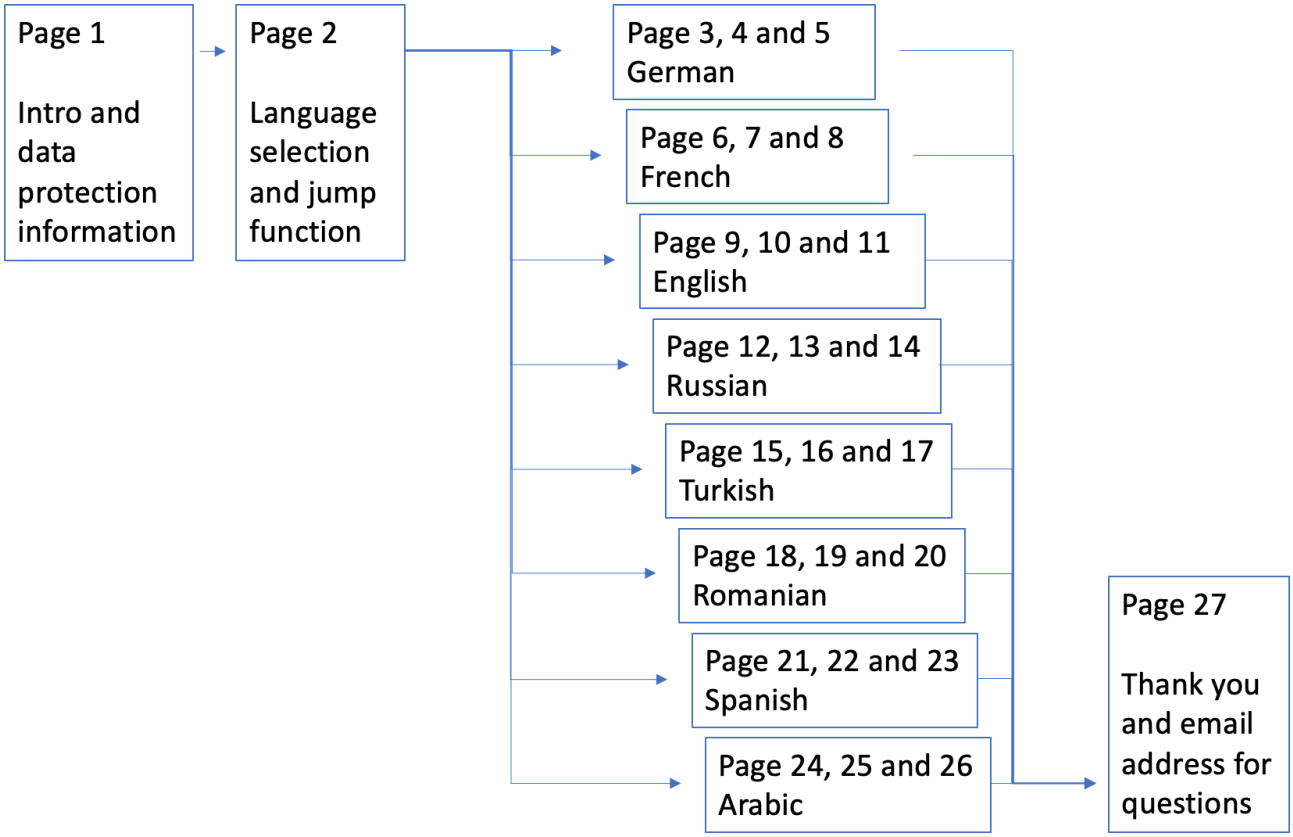

Among the 18 substantive questions, there was an open-ended question at the beginning regarding the number of counseling sessions attended to date, as well as an open-ended question at the end for additional comments. The questionnaire also contained 1 yes-no question, 13 questions using Kunin smiley faces on a 5-point scale (see Figure 2), and 2 multiple-choice questions, the options for which were based on the categories derived from responses to the second and sixth questions of the interview study.

To ensure the anonymity of the participants, demographic data such as age and gender were not collected.

Data Collection

The survey was conducted between November and December 2023 using tablets within the HCS. Each participant was requested to complete the questionnaire following their counseling session, resulting in a total of 31 data entries. Due to data protection regulations, the number of individuals approached was not recorded; therefore, the response rate remains unknown.

Subsequently, from January to March 2024, the survey link was made available on the HCS's website, a QR code was posted within the facility, and clients were informed about the survey following their counseling sessions. The number of individuals who visited the HCS during this period was also not documented for data protection reasons, resulting in an unknown response rate in this phase as well. During the period from January to March 2024, an additional 10 data entries were collected, culminating in a total of 41 evaluable data sets (with 25 aborted surveys, 38% dropouts). While response rates could not be documented due to data privacy policies, the dropout rate of 38% suggests potential limitations in sustained participant engagement, which should be considered in the interpretation of results.

Data Analysis

All statistical analyses of the data from the survey were carried out using SPSS version 29 (IBM). Descriptive statistics were calculated for the 13 items of the online questionnaire that were answered with a 5-point Likert scale.

Data Exclusion

Participants who did not consent to the data privacy statement were automatically directed to the last page and were excluded from the study. Since the survey was published on the website, there have also been instances of visits without any responses; these instances were likewise excluded from the final dataset.

Results

Study 1: Qualitative In-Depth Interview Study

The final sample for the interview study comprised 18 Interviews (10 clients, 6 case management participants, and 2 walk-in consultation seekers). In the context of the study, the responses to the initial question regarding the duration of engagement with counseling services exhibited a considerable degree of vagueness. Even when prompted by the interviewers, participants were unable to provide more precise answers than "for a long time," "approximately three to five years," and described their frequency of visits with terms like "often" and "every few months."

XSL•FO
RenderX

The primary motivation for seeking HCS, as indicated in response to the second question, was predominantly related to issues concerning physical health. Furthermore, it was identified that the most frequently requested form of assistance was the completion of application forms, as detailed in Table 1.

Table . Overview of consulting content and types of services received

Inductively obtained categories	Mentions, n (%)	Example response
Consultancy content		
Physical health	14 (78)	"I am here because of my rheumatism."
Mental health	13 (72)	"I have severe depression."
Other[a]	3 (17)	"Treatment requests for my illnesses."
Type of service		
Fill out applications and forms	17 (94)	"Well, to fill out papers for the doctor."
Receive advice and knowledge	5 (28)	"Vaccination counselling, especially Corona."
Referral to specialists	5 (28)	"I want to search therapy place.", "Need to find a specialist doctor."

[a]Responses could not be clearly assigned to mental or psychological health.

Responses to the third question, along with the 2 follow-up inquiries, indicated that 18 participants (100%) felt that the allocated time for HCS was adequate. In addition, 12 respondents (67%) expressed that all relevant topics could be adequately addressed during their consultations. Furthermore, 15 participants (83%) reported feeling well-supported throughout the counseling process, allowing for multiple responses.

In addressing questions 4 and 5, which pertained to the participants' experiences with HCS and the staff, the most frequently mentioned aspects included the result-oriented and goal-oriented approach of the counselors, as well as their overall counseling attitude. Conversely, remarks regarding challenges and obstacles encountered in the counseling process were infrequently noted (see Table 2).

Table . Overview of responses on how staff and advice were experienced by clients. Multiple answers were possible.

Inductively obtained categories	Statements on the experience of consulting and employees	
	Mentions, n	Example responses
Positive atmosphere	8	• "The atmosphere here is very good, quiet and pleasant."
Result orientation and goal orientation of the consultants	34	• "When I couldn't appear in person, we then conducted over the phone." • "When you go out afterwards, you know everything a bit more clearly."
Counseling attitude and handling	22	• "They don't lecture me, they don't talk down to me, they don't tell me what to do." • "They really treated you with respect and dignity."
Expressed trust and sense of security	21	• "I have a good feeling about the consultation." • "I almost feel like family."
Challenges and obstacles	7	• "The consultation here is in German, I used to have it in Arabic."

In response to the sixth question and the probing question, positive emotions were mentioned 11 times (Mentioned emotions: "a good feeling," "I am much happier afterwards," "I feel good"), 4 times less stress was reported (Statements about stress: "I am so nervous before," "I go home and the stress is gone") and 5 times answers were given that were sorted into the "Other" category ("the consultation is just good"). In response to the sixth question on emotional well-being and the associated follow-up questions, the participants registered exclusively positive statements.

When asked specifically whether and where support could be improved, the respondents provided the following insights (see Table 3): The 3 most frequently cited topics, in which the 18 participants wanted further assistance, included physical and mental health, and application processes, with each topic receiving 10 mentions (56%).

XSL•FO
RenderX

Table . Desired additional support mentioned. Multiple answers were possible.

Desired additional support	Mentions, n (%)
Referral to specialists	1 (6)
Language acquisition	2 (11)
Housing search	2 (11)
Job search	2 (11)
Networking	3 (17)
Receive case management services	5 (28)
Applications and forms	10 (56)
Physical and mental health	10 (56)

Of the 12 respondents who were either clients or walk-in customers, 5 expressed a desire for more intensive support, specifically in the form of case management services, representing 42 percent of this subgroup (28% overall).

In response to the seventh question regarding whether the participants felt they could cope independently in the future, only 2 individuals (11%) initially answered "yes." However, both subsequently revised or qualified their answers. As a result, the majority of participants (16 persons, 89%), responded with "no."

Among the 26 reasons mentioned were "Need for help due to chronic mental/physical health problems" (2 times), "Overwhelmed by structures in the healthcare system" (3 times), "Language barriers (eg, at the doctor's at the doctor or hospital; mentioned 4 times)," "Barriers due to lack of reading/spelling skills" (mentioned 1 time), "No other help (caregiver, other advice center) known" (mentioned 2 times) as well as statements without further specification such as "No, I can't cope without help" or "No way" (category "Other," mentioned 14 times).

To obtain a statement that allows for comparability of results with other institutions in line with the Customer Loyalty Index (CLI), the final question eight asked whether clients would recommend the health care facility. All 18 participants (100%) answered affirmatively. The answers were also supplemented with statements that clients had already recommended the HCS to others or referred friends and acquaintances (5 times), that they were very satisfied (5 times) and numerous personal examples of when and how they had been helped.

In response to the probing question about what other health advice was desired, "vaccination advice" was mentioned once. In addition, the wish for a picnic was expressed once and it was said three times that it would be nice to have more time with the counselors to talk more (informal conversations). The remaining comments were praise for the interviewees and the counseling staff.

In the second probing question on the eighth question ("Is there anything else you would like to tell us?"), which was intended as a wrap-up, praise for the employees was mentioned, it was said that the advice service must not close under any circumstances, and that health and health advice is very important.

One contextual element that may have had an impact on the interviewees' statements is the fact that the interviews were conducted on the premises of the HCS. Another contextual element could have been that the telephone interviews were conducted via the employees.

Study 2: Quantitative Survey Study

A total of 66 surveys were started and 41 completed (dropout rate of 38%). In response to the question regarding language preference, more than half of the 41 participants chose German as their preferred language (21 participants, 51%). English and Arabic were both selected as the second most common choices, each accounting for 6 participants (15%) of the responses (see Table 4).

Table . Frequencies with which the different languages were chosen. Only 7 of the 8 languages are shown because Spanish was not chosen.

Language	Frequency, n (%)
German	21 (51)
French	2 (5)
English	6 (15)
Russian	3 (7)
Turkish	1 (2)
Romanian	2 (5)
Arabic	6 (15)

In response to the open-ended question regarding the number of previous counseling sessions, the mean of the responses was calculated to be 13.68 (SD 14.81). Regarding the second open-ended question, which inquired about any further

XSL•FO
RenderX

comments or expressions of wishes to communicate, multiple participants expressed praise and gratitude for the services received.

In response to the yes or no question, one individual indicated awareness of alternative support services, specifically mentioning a caregiver. All other answers were "no" or no answer.

To investigate the subjective perceptions of the participants, items featuring various statements were employed, to which respondents could express their agreement using a 5-point Likert scale (see Table 5).

Table . Mean values and standard deviations of the responses, collected using a 5-point Likert scale.

Questionnaire items to record the subjective perceptions	Responses, mean (SD)
I recommend the health counseling service.[a]	4.88 (0.33)
I got help.	4.78 (0.42)
During the conversation I understood everything.	4.65 (0.60)
It's easy to get an appointment.	4.61 (0.69)
My problem was solved.	4.59 (0.61)
There was enough time during the conversation.	4.53 (1.07)
With the health counseling service,[a] I now know things I did not know before.	3.94 (1.21)
I feel comfortable at the health counseling service.[a]	3.88 (1.69)
With the health counseling service,[a] I can do things now that I could not do before.	3.72 (1.01)
It is easy to get help in Germany.	3.37 (0.89)
From now on I can help myself.	3.24 (1.20)
I want to meet new people.	3.22 (1.47)
I want to meet other clients of the health counseling service.[a]	2.44 (1.61)

[a]The name of the health counseling service was used instead of "the health counseling service" in the web-based survey.

Regarding the 2 multiple-choice questions, which included response options based on the deductive categorical system from the interview study, the data revealed the following: the 3 most frequently mentioned sources from which individuals seeking advice had been referred were friends, unspecified others, and the employment agency (see Table 6). The primary reasons for seeking assistance from the HCS center included the desire to receive help with completing applications for authorities, having forms from and for medical practices filled out, and seeking guidance on health-related issues (see Table 7).

Table . Frequency of responses from whom consultation seekers were sent for consultation. The option "other" had a free input field in which respondents entered: "office," "work," and "caregiver."

	Responses, n (%)
Other	4 (10)
Doctor	2 (5)
Debt counseling	2 (5)
Family	1 (2)
Friends	6 (15)
Job center	3 (7)

XSL•FO
RenderX

Table . Frequency of responses for reasons why the health counseling service was used. In the open entry option under "other," the word "therapy" was entered.

Reason	Responses, n (%)
Filling out forms from doctors	11 (27)
Filling out applications for authorities	14 (34)
Advice on health issues	10 (24)
Housing search	3 (7)
Job search	6 (15)
Search for special doctors	3 (7)
Other	4 (10)

Discussion

Principal Findings

The qualitative study identified numerous barriers, such as language limitations, illiteracy, and psychological and physical impairments. The challenges of conducting a quantitative web-based survey for this vulnerable and hard-to-reach population could be addressed using a mixed methods approach. The most significant finding was that, contrary to previous assumptions, it was not counseling and assistance with health issues that were the primary interests of the clients, but rather support in dealing with forms and requirements imposed by institutions. It became evident that a substantial proportion of the individuals visiting the HCS suffer from multiple health problems, both physical and mental. The most significant aspect of the findings in the quantitative web-based-survey study is the implementation of the survey method and its evaluation, which enables the health care facility to continuously adapt its future work to the needs of its vulnerable and hard-to-reach clients. Only half of the respondents completed the survey in German, which clearly indicates that HCS and evaluation offered solely in German are likely insufficient. The analysis of the data further revealed that some participants required more than 30 minutes to complete the survey, as they listened to the audio recordings. This supports the assumption that the original paper-pencil assessment, which was solely in German and without explanations, was not appropriate.

A detailed examination of the qualitative study reveals that a significant portion of the work in HCS appears to focus on assisting clients with the completion of forms and documents for health insurance companies and medical professionals. This finding is not surprising, given that bureaucracy for obtaining health services in Germany has increased over recent decades, which understandably poses a significant hurdle for burdened individuals as well as those lacking sufficient system knowledge and proficiency in German. Thus, the conducted interview study highlights the substantial barriers to accessing health services and emphasizes that forms from medical professionals and health insurance companies (eg, rehabilitation requests, medical history forms, practice forms, and forms for preventive examinations) represent an overwhelming level of demand for clients. In contrast, other topics, such as education on health-promoting behaviors and vaccination advice, are rarely addressed. This may be due in part to more pressing concerns on the clients' side (as health-related forms are often accompanied by a submission deadline) and, on the other hand, due to a lack of resources to cover additional topics.

Positively, clients reported a high level of satisfaction with the provided HCS. This could potentially be influenced by the fact that dissatisfied clients may come less frequently or seek other forms of support, or that clients may fear that the HCS will be closed if they provide negative feedback. This was explicitly mentioned in one of the interviews. At the same time, it becomes evident that clients of the health care facility with language limitations, difficulties in reading and writing, as well as cognitive and emotional barriers would be overwhelmed by the demands of the German health care system without the support provided by the facility, and would struggle to cope without the counseling.

It is noteworthy that, although the quantitative web-based survey study showed that clients are generally satisfied and many report that they learned something new and gain skills after each counseling session, they do not believe they will be able to help themselves independently in the future. It is clear that many individuals seek assistance primarily with the completion of forms and applications for medical professionals and health insurance companies, and they perceive this task as overwhelming in the long run. Bureaucracy within the health care system thus presents a significant barrier that they do not believe they can ever manage alone.

Other HCS (individual counseling, information and education, networking, organization of self-help groups, and health education through lectures and workshops), which would also provide substantial benefits to these individuals, were, however, not requested. At the same time, the data indicate that very few participants were referred by doctors and authorities. This could be attributed to a lack of awareness regarding the number of people who need assistance with form completion or to inadequate knowledge among medical professionals and authorities about the availability of HCS that offer such support. There is an urgent need for increased awareness in this regard.

Both studies may be limited by the small number of participants and the specific selection of clients visiting the HCS center, which may not necessarily allow for broader generalizations to the entire population.

Furthermore, biased self-reports from clients (eg, due to psychological stress) could have affected the validity of the data

XSL•FO
RenderX

collected. In the first study he opinions of walk-in clients are likely underrepresented, as only 2 out of 18 respondents (11%) in the interview study sample belonged to this demographic, while staff at the HCS center estimate that approximately a quarter of consultations involve walk-in clients. The distribution of responses among clients, cases, and walk-in clients in the web-based survey is also unclear. Furthermore, it cannot be excluded that some individuals participated in both the interview study and the web-based survey, which would reduce the generalizability of the findings. Also, it cannot be guaranteed that the audio content was always understood, both auditorily and cognitively, by the participants who chose to have the questions read aloud to them instead of reading them themselves. In addition to the previously described challenges of digital accessibility for this group, limitations due to psychological distress and the difficulty of reaching this client population may also play a role. However, the relatively small sample does represent the majority of the current clients of the health care facility in terms of their challenges.

In conclusion, the findings of this study clearly indicate that support with administrative requirements is of paramount importance for clients within the German health care system. The overwhelming bureaucratic hurdles and the associated burdens underscore the need for reforming health services to ensure that clients not only receive the necessary medical assistance but also the support needed to successfully navigate the complex health care landscape. These insights carry broader implications for the design of health services, which should be more inclusive and accessible to better address the needs of these vulnerable groups. These insights are of great significance, as they underscore for a broader awareness of the challenges faced by clients. Furthermore, a training program for staff on completing applications for medical services is now being considered. Ultimately, however, health services must be designed to be more inclusive and accessible to enable the needs of the entire population to be met without intensive support structures.

Comparison With Previous Work

The findings of this study are consistent with a variety of research concerning medically underserved populations.

Once again, it was shown that in-depth interview are very well suited to identifying client experience with the health system [24]. Participants in the qualitative in-depth interview study reported that Arabic language counseling services were previously available; however, these services have been discontinued due to a shortage of staff. The issues of inadequate expertise and personnel shortages in HCS have also been documented in previous studies [1,2,28].

The web-based survey indicates that collecting data on health uptake among non-native speakers poses significant challenges [29] and that it is important to involve several stakeholders from the outset and to pay attention to user-friendliness [30]. Furthermore, it has been confirmed that this demographic faces specific barriers, of which healthcare providers and authorities are only partially aware [31]. The notably lower data collection rates during the second period of the study—where data were gathered through a link and QR code instead of in-person consultations with tablets—alongside a high dropout rate (n= 25, 37%), suggest that individuals seeking HCS encounter difficulties in navigating digital media. This observation is also supported by previous research [11,15,23].

Overall, the results confirm that the low demand for health services in Germany, as evidenced in earlier studies [2,4-7], may indeed be attributed to the challenges faced by particularly burdened individuals, specifically those who are in greatest need of assistance [31].

Conclusions

In summary, the results of the 2 studies underscore the complex challenges faced by individuals seeking assistance at HCS centers, particularly in relation to the completion of forms required by health care professionals and insurance providers. The qualitative in-depth interviews reveal that a significant proportion of clients are overwhelmed by the bureaucratic demands of the health care system, which detracts from opportunities for health promotion and education. The quantitative analysis further suggests that language barriers and insufficient referral practices limit access to resources, highlighting the inadequacy of German-only evaluation form. While clients express satisfaction with the services received, they largely lack confidence in their ability to navigate the system independently in the future. These findings indicate an urgent need for enhanced awareness among health care providers and authorities regarding the challenges faced by medically underserved populations, as well as the importance of effective networking and support systems to facilitate holistic patient care.

Acknowledgments

The authors wish to express their gratitude to the health counseling service staff who participated in the workshops and facilitated the collection of interviews. Special thanks are extended to the students who assisted in the collection, transcription, and coding of the in-depth interviews, as well as those who verified the various languages of the questionnaire and audio recordings. Furthermore, the authors would like to acknowledge all participants who contributed to the study, either by engaging in the in-depth interviews or by completing the online survey. The publication was partly paid for by the publication fund of the Alice Salomon University.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on request.

XSL•FO
RenderX

Authors' Contributions

AB-F and LMB designed, planned, and conducted the study, as well as analyzed the data. AB-F primarily drafted the manuscript, while LMB contributed to its revision and editing. Both authors approved the final version of the manuscript.

Conflicts of Interest

None declared.

References

1. Amelung V, Hildebrandt H, Wolf S. Integrated care in Germany-a stony but necessary road!. Int J Integr Care 2012;12:e16. [doi: 10.5334/ijic.853] [Medline: 22977429]
2. Boettcher VS, Nowak AC, Neuner F. Mental health service utilization and perceived barriers to treatment among adult refugees in Germany. Eur J Psychotraumatol 2021 May 14;12(1):1910407. [doi: 10.1080/20008198.2021.1910407] [Medline: 34025926]
3. Woods MD, Kirk MD, Agarwal MS, et al. Vulnerable groups and access to health care: a critical interpretive review. 2005.
4. Mack S, Jacobi F, Gerschler A, et al. Self-reported utilization of mental health services in the adult German population--evidence for unmet needs? Results of the DEGS1-Mental Health Module (DEGS1-MH). Int J Methods Psychiatr Res 2014 Sep;23(3):289-303. [doi: 10.1002/mpr.1438] [Medline: 24687693]
5. Berens EM, Vogt D, Ganahl K, Weishaar H, Pelikan J, Schaeffer D. Health literacy and health service use in Germany. Health Lit Res Pract 2018 Apr;2(2):e115-e122. [doi: 10.3928/24748307-20180503-01] [Medline: 31294284]
6. Bremer D, Lüdecke D, von dem Knesebeck O. Social relationships, age and the use of preventive health services: findings from the German Ageing Survey. Int J Environ Res Public Health 2019 Nov 4;16(21):4272. [doi: 10.3390/ijerph16214272] [Medline: 31689892]
7. Roth G, Reichert M. Providing integrated health and social care for older persons. In: Providing Integrated Health and Social Services for Older Persons 2019:269-328.
8. Quittschalle J, Stein J, Luppa M, et al. Internet use in old age: results of a German population-representative survey. J Med Internet Res 2020 Nov 23;22(11):e15543. [doi: 10.2196/15543] [Medline: 33226351]
9. De Santis KK, Jahnel T, Sina E, Wienert J, Zeeb H. Digitization and health in Germany: cross-sectional nationwide survey. JMIR Public Health Surveill 2021 Nov 22;7(11):e32951. [doi: 10.2196/32951] [Medline: 34813493]
10. Thonon F, Perrot S, Yergolkar AV, et al. Electronic tools to bridge the language gap in health care for people who have migrated: systematic review. J Med Internet Res 2021 May 6;23(5):e25131. [doi: 10.2196/25131] [Medline: 33955837]
11. Talhouk R, Akik C, Araujo-Soares V, et al. Integrating health technologies in health services for Syrian refugees in Lebanon: qualitative study. J Med Internet Res 2020 Jul 6;22(7):e14283. [doi: 10.2196/14283] [Medline: 32628121]
12. Kouvonen A, Kemppainen L, Ketonen EL, Kemppainen T, Olakivi A, Wrede S. Digital information technology use, self-rated health, and depression: population-based analysis of a survey study on older migrants. J Med Internet Res 2021 Jun 14;23(6):e20988. [doi: 10.2196/20988] [Medline: 34125069]
13. Gordon NP, Crouch E. Digital information technology use and patient preferences for internet-based health education modalities: cross-sectional survey study of middle-aged and older adults with chronic health conditions. JMIR Aging 2019 Apr 4;2(1):e12243. [doi: 10.2196/12243] [Medline: 31518291]
14. Chae S, Lee YJ, Han HR. Sources of health information, technology access, and use among non-english-speaking immigrant women: descriptive correlational study. J Med Internet Res 2021 Oct 29;23(10):e29155. [doi: 10.2196/29155] [Medline: 34714249]
15. Choi NG, Dinitto DM. The digital divide among low-income homebound older adults: internet use patterns, eHealth literacy, and attitudes toward computer/internet use. J Med Internet Res 2013 May 2;15(5):e93. [doi: 10.2196/jmir.2645] [Medline: 23639979]
16. Nakayama LF, Binotti WW, Link Woite N, et al. The digital divide in Brazil and barriers to telehealth and equal digital health care: analysis of internet access using publicly available data. J Med Internet Res 2023 Jul 21;25:e42483. [doi: 10.2196/42483] [Medline: 37477958]
17. Norman EC, Forti TJ. A study of the process and the outcome of mental health consultation. Community Ment Health J 1972 Nov;8(4):261-270. [doi: 10.1007/BF01440876] [Medline: 4669434]
18. Drewett A. Evaluation and consultation: learning the lessons of user involvement. Eval 1997;3(2):189-204. [doi: 10.1177/135638909700300205]
19. Knotek SE, Hylander I. Research issues in mental health consultation and consultee-centered approaches. In: Handbook of Research in School Consultation 2014:153-179.
20. Stewart M, Brown JB, Weston WW, et al. Patient-Centered Medicine: Transforming the Clinical Method: CRC Press; 2024.
21. Noergaard B, Sandvei M, Rottmann N, et al. Development of a web-based health care intervention for patients with heart disease: lessons learned from a participatory design study. JMIR Res Protoc 2017 May 17;6(5):e75. [doi: 10.2196/resprot.7084] [Medline: 28526674]

XSL•FO
RenderX

22. Holmen H, Singstad T, Ribu L, Jensen AL, Weldingh NM, Torbjørnsen A. Adapting a patient-reported outcome measure to digital outpatient specialist health care services for type 1 diabetes: user involvement study. JMIR Hum Factors 2022 Nov 15;9(4):e38678. [doi: 10.2196/38678] [Medline: 36378513]
23. Jackson AM, Woo J, Olson M, et al. Methodological challenges in web-based qualitative research with medically underserved populations. J Med Internet Res 2023 Mar 30;25:e44086. [doi: 10.2196/44086] [Medline: 36995748]
24. Hirsch A, Durden TE, Silva J. Linking electronic health records and in-depth interviews to inform efforts to integrate social determinants of health into health care delivery: protocol for a qualitative research study. JMIR Res Protoc 2022 Mar 11;11(3):e36201. [doi: 10.2196/36201] [Medline: 35275090]
25. Gnambs T, Kaspar K. Socially desirable responding in web-based questionnaires: a meta-analytic review of the candor hypothesis. Assessment 2017 Sep;24(6):746-762. [doi: 10.1177/1073191115624547] [Medline: 26739360]
26. Mayring P. Qualitative content analysis: demarcation, varieties, developments. Forum Qual Soc Res 2019;20(3). [doi: 10.17169/fqs-20.3.3343]
27. Fenzl T, Mayring P. QCAmap: eine interaktive webapplikation für qualitative inhaltsanalyse [Article in German]. ZSE 2017;37:333-340 [FREE Full text]
28. Rosenblatt RA, Andrilla CHA, Curtin T, Hart LG. Shortages of medical personnel at community health centers: implications for planned expansion. JAMA 2006 Mar 1;295(9):1042-1049. [doi: 10.1001/jama.295.9.1042] [Medline: 16507805]
29. Koschollek C, Zeisler ML, Houben RA, et al. German Health Update Fokus (GEDA Fokus) among residents with Croatian, Italian, Polish, Syrian, or Turkish citizenship in Germany: protocol for a multilingual mixed-mode interview survey. JMIR Res Protoc 2023 Apr 12;12(12):e43503. [doi: 10.2196/43503] [Medline: 36790192]
30. Lee EW, McCloud RF, Viswanath K. Designing effective ehealth interventions for underserved groups: five lessons from a decade of ehealth intervention design and deployment. J Med Internet Res 2022 Jan 7;24(1):e25419. [doi: 10.2196/25419] [Medline: 34994700]
31. Saloner B, Wilk AS, Levin J. Community health centers and access to care among underserved populations: a synthesis review. Med Care Res Rev 2020 Feb;77(1):3-18. [doi: 10.1177/1077558719848283] [Medline: 31079529]

Abbreviations

AI: artificial intelligence
CLI: Customer Loyalty Index
HCS: health counseling service

Edited by A Mavragani; submitted 16.11.24; peer-reviewed by J Dickson, J Mistry; revised version received 04.04.25; accepted 08.04.25; published 04.06.25.

Please cite as:
Boehm-Fischer A, Beyer LM
Determining the Requirements of Vulnerable Groups for Health Counseling and Optimizing the Evaluation of Health Consultations: Mixed Methods Study With the Use of AI
JMIR Form Res 2025;9:e68888
URL: https://formative.jmir.org/2025/1/e68888
doi:10.2196/68888

© Annina Boehm-Fischer, Luzi M Beyer. Originally published in JMIR Formative Research (https://formative.jmir.org), 4.6.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Integrating Nurse Preferences Into AI-Based Scheduling Systems: Qualitative Study

Fabienne Josefine Renggli[1], MSc; Maisa Gerlach[1], MSc; Jannic Stefan Bieri[1], BSc; Christoph Golz[1], PhD; Murat Sariyar[2], PhD

[1]School of Health Professions, Bern University of Applied Sciences, Bern, Switzerland

[2]School of Engineering and Computer Science, Bern University of Applied Sciences, Höheweg 80, Biel, Switzerland

Corresponding Author:
Murat Sariyar, PhD
School of Engineering and Computer Science, Bern University of Applied Sciences, Höheweg 80, Biel, Switzerland

Abstract

Background: Nurse scheduling is a complex challenge in health care, impacting both patient care quality and nurse well-being. Traditional scheduling methods often fail to consider individual preferences, leading to dissatisfaction, burnout, and high turnover. Inadequate scheduling practices, including restricted autonomy and lack of transparency, can further reduce nurse morale and negatively affect patient outcomes. Research suggests that participative scheduling approaches incorporating nurse preferences can improve job satisfaction. Artificial intelligence (AI) and mathematical optimization methods, such as mixed-integer programming (MIP), constraint programming (CP), genetic programming (GP), and reinforcement learning (RL), offer potential solutions to optimize scheduling and address these challenges.

Objective: This study aims to develop a framework for integrating nurses' preferences into AI-supported scheduling methods by gathering qualitative insights from nurses and supervisors and mapping these to mathematical and AI-based scheduling techniques.

Methods: Focus group interviews were conducted with 21 participants (nurses, supervisors, and temporary staff) from Swiss health care institutions to understand experiences and preferences related to staff scheduling. Qualitative data were analyzed using open and axial coding to extract key themes. These themes were then mapped to AI methodologies, including MIP, CP, GP, and RL, based on their suitability to address identified scheduling challenges.

Results: The study revealed key priorities in nurse scheduling. Fairness and participation were highlighted by 85% (18/21) of interview participants, emphasizing the need for transparent and inclusive scheduling. Flexibility and autonomy were preferred by 76% (16/21), favoring shift swaps and self-scheduling. AI expectations were mixed: 62% (13/21) saw potential for improved efficiency and fairness, while 38% (8/21) expressed concerns over reliability and loss of human oversight. Mapping to AI methods showed MIP as effective for fair shift allocation, CP for complex rule-based conditions, GP for handling unforeseen absences, and RL for dynamic schedule adaptation in hospital environments. A preliminary AI implementation of MIP in a training hospital unit (35 staff members) showed how to design a system from a mathematical perspective.

Conclusions: AI-supported scheduling systems can significantly enhance fairness, transparency, and efficiency in nurse scheduling. However, concerns regarding AI reliability, adaptability to individual needs, and human oversight must be addressed. A hybrid approach integrating AI recommendations with human decision-making may be optimal. Future research should explore the broader implementation of AI-driven scheduling models and assess their impact on nurse satisfaction and patient outcomes over time.

(JMIR Form Res 2025;9:e67747) doi:10.2196/67747

KEYWORDS

nurse scheduling; work-life balance; feasibility; well-being; large language model; LLM; natural language processing; NLP; machine learning; ML; artificial intelligence; AI; AI-based scheduling; dissatisfaction; burnout; comprehensive framework; interview; mixed-integer programming; MIP; constraint programming; CP; reinforcement learning; RL

Introduction

In the health care setting, nurse scheduling is a critical yet complex task that directly impacts both the quality of patient care and the well-being of nurses [1-4]. Traditional scheduling methods often fail to accommodate the preferences and constraints of individual nurses, potentially leading to dissatisfaction, burnout, and high turnover rates [5,6]. However, the extent of these negative effects and the specific factors contributing to dissatisfaction are often unclear. Nurses may

XSL•FO
RenderX

feel restricted in their autonomy or, conversely, may desire more guidance to prevent feelings of unfairness [7]. In addition, issues may arise from either a lack of transparency in scheduling processes or an overload of information that leads to confusion [8]. Before considering improvements through formal or artificial intelligence (AI)–based scheduling methods, it is essential to thoroughly understand the current state and the specific needs and desires of the nursing staff. This allows then to map those needs to algorithmic scheduling solutions, ensuring an efficient transition to a desired future state that enhances both nurse satisfaction and patient care.

Several studies have examined health care workers' preferences via interviews. Baljani et al [9] interviewed 20 managers and 2 clinical nurses and found that functional-exchangeable protection and flexible schedules reduce nurses' family-related worries and work-family conflict, improving care quality and reducing errors. Booker et al [10] identified inadequate training on sleep and shift management, raising safety concerns. Saritas [11] interviewed 50 female nurses and showed that staffing shortages, intensified workloads, unpredictable schedules, and poor support worsen work-family conflict, making nurses reliant on spousal support and more likely to accept unfavorable conditions. Messing et al [12] reported that work-life balance issues relate to scheduling and family duties, with workers preferring stable schedules and consistent days off. Christopher [13] noted that only one-third of 26 interviewed nurse-mothers control their schedules: understaffing and long shifts compromise both patient and family care quality. Kim et al [14] identified key risk factors contributing to burnout syndrome, including work overload, demanding schedules, strained managerial relationships, poor work-life balance, and bureaucratic management.

Participative staff scheduling actively involves employees in creating work schedules, allowing them to express their preferences and needs regarding shifts and work hours. This approach enhances autonomy, flexibility, and job satisfaction by considering individual desires. Mrayyan et al [15] highlight that health care workers are often dissatisfied with limited autonomy, which is supported by management, education, and experience but undermined by authoritarian leadership, physicians, and high workload. Enhancing autonomy through targeted leadership and education is recommended. Christensen et al [16] illustrate how participatory management, using the Listen-Sort-Empower model, improves understanding of burnout causes like staffing shortages, work environment, and compensation, facilitating collaborative solutions. Similarly, Albargi et al [17] emphasize that empowering nurses in decision-making boosts engagement, job satisfaction, and patient outcomes. Shiri et al [18] found that digital participatory scheduling software improved employees' control over their schedules, sleep quality, and perceived workability, but had no effect on psychological distress, self-rated health, or work-life conflict. Although participation of nurses involves more than just staff scheduling, these examples demonstrate that flexibility in this area is a crucial factor for overall job satisfaction.

Mathematical applications in participative staff scheduling leverage advanced techniques to optimize shift allocation based on employee preferences, availability, and operational requirements [19-23]. For stable shift planning incorporating preferences as constraints, 2 primary methods are used: mixed-integer programming (MIP) and constraint programming (CP). MIP addresses optimization problems with both continuous and discrete decision variables, making it suitable for scenarios where scheduling decisions must meet linear constraints and objectives [24]. It is effective for large-scale scheduling problems with well-defined constraints, such as total work hours, shift assignments, and resource allocation. MIP efficiently manages complex objective functions and constraints, including cost minimization and coverage maximization. Conversely, CP is designed for combinatorial problems involving complex, often nonlinear constraints. It focuses on feasibility rather than optimization and excels at managing diverse and intricate constraints and preferences [25]. It adeptly explores and refines the solution space to identify feasible schedules that meet all constraints.

Participatory staff scheduling can also be significantly enhanced by AI methodologies, especially by reinforcement learning (RL) and natural language processing (NLP). RL optimizes dynamic scheduling by adapting to evolving conditions through interaction and feedback, managing complex environments effectively [26-30]. NLP methods can process textual feedback from employees concerning their preferences and scheduling needs, converting it into actionable data for more personalized schedules, and with large language models even promise to provide solutions for the staff scheduling problem itself. In addition, heuristics like genetic algorithms complement MIP and CP, providing feasible solutions when exact methods to yield satisfactory results [31]. Furthermore, machine learning (ML) techniques, such as random forests, can be used to predict scheduling preferences based on historical data and patterns [32].

Currently, to our knowledge, there is no comprehensive framework that integrates nurses' requirements for participatory shift scheduling within a broader set of methodologies. This paper aims to address this gap and present preliminary findings to facilitate broader implementation. The following section details the methodology for conducting interviews and how these results are mapped to the mathematical techniques. Subsequently, we will present the mapping result and initial outcomes of applying one of these mathematical methods, after which we conclude with a discussion of future perspectives. We adopt a high-level perspective and provide only a few illustrative mathematical formulas for clarity, as our primary audience includes nontechnical decision-makers.

Methods

Interviews

We conducted focus group interviews, for which the details and reporting checklist are provided in a paper still under review [33]. The study followed the COREQ (Consolidated Criteria for Reporting Qualitative Research) guideline to ensure comprehensive reporting. Participants were recruited through a convenience sampling approach, targeting permanent and temporary nursing staff from various health care settings, including acute hospitals, home care services, and nursing homes

XSL•FO
RenderX

in German-speaking Switzerland. Recruitment flyers detailing the study's purpose, procedure, and voluntary nature were distributed within participating institutions. Interested individuals contacted the research team directly to enroll in the study. The sample consisted of 21 participants, including nurse managers, permanent nurses, and temporary nurses, ensuring diverse perspectives on AI-based scheduling and work-life balance.

Here, we present just the summary. The interviews involved open-ended questions related to staff scheduling and prompt participants to articulate their thoughts, experiences, and opinions in their own words [34]. A total of 21 Swiss working staff participated in 4 focus group interviews. The central specific questions are presented in Table 1.

Table . Central interview questions by participant group in artificial intelligence (AI)–based nurse scheduling study. This table presents the key interview questions posed to different participant groups—supervisors, permanent nurses, and temporary staff—during qualitative focus group discussions conducted in Swiss health care institutions between May and June 2024. The study aimed to explore perceptions of existing scheduling systems, fairness, flexibility, and the potential role of AI-based solutions in nurse shift scheduling.

Interview group	Question
Supervisor	• How does staff scheduling function as a management tool in your department? • How do you manage staff absences and ensure fairness in scheduling? • How does scheduling impact staff satisfaction? • What features do you expect from an AI-based scheduling system?
Nurses	• How well does the current scheduling system meet your needs for time off and work-life balance? • What scheduling challenges do you face in hospital and outpatient care settings? • How do shift lengths and scheduling practices affect your job satisfaction and recovery? • What improvements would enhance your scheduling experience?
Temporary staff	• How does flexibility in creating your own schedule compare with that of permanent staff? • How do you balance personal scheduling needs with institutional requirements?

Ethical Considerations

Human Participants Research Ethics Review, Exemptions, and Approvals

This study was conducted in accordance with the ethical guidelines of the Bern University of Applied Sciences. The local ethics committee confirmed that the study did not require a full ethical review and was not covered by the Swiss Federal Law on research involving humans. As such, a formal institutional review board approval was not necessary. However, all research activities adhered to principles of ethical research, ensuring respect for participants' rights and well-being.

Informed Consent

Before participation, all individuals were provided with a detailed study information sheet explaining the research purpose, methodology, potential risks, and benefits. Participants voluntarily provided written informed consent before their involvement in the study. As this study involved secondary qualitative analysis of interview data, all primary data were collected with consent that permitted subsequent secondary analysis.

Privacy and Confidentiality Protection

To protect participant confidentiality, all interview data were anonymized before analysis. Identifying information, such as names, job titles, and specific workplace locations, was removed or altered to prevent reidentification. Data storage was secured in password-protected files, accessible only to authorized research team members. No personally identifiable images or information were included in the study.

Compensation

No financial or material compensation was provided to participants for their involvement in this study. Participation was entirely voluntary, and no incentives were offered to ensure unbiased responses.

Protection of Individual Identity in Images

No images or supplementary materials in the manuscript contain identifying features of participants. In the event that identifiable images were deemed necessary, written informed consent would be obtained, and the appropriate consent documentation would be provided upon submission. However, as this study relies solely on anonymized qualitative data, no such images are included.

Mapping Requirements to Methods

The mapping process occurs in two stages: (1) deriving key themes from the interview summaries, and (2) the actual mapping, where relevant AI methods are assigned based on the identified categories to address the specific needs of the interview participants. In stage 1, key themes from the interviews were extracted and axial coding used to organize these themes into overarching categories [35]. Subsequently, expectation and concerns regarding AI in this context is

XSL•FO
RenderX

summarized. In stage 2, of the mapping process, we align the categories identified during the interviews with appropriate AI methodologies. Coauthor MS, who possesses the necessary technical expertise, provides the rationale for selecting the most suitable AI methods–MIP, CP, GP (genetic programming [36]), and RL–to address each identified need. For one category identified as feasible for immediate implementation, an exemplary implementation is shown. Table 2 outlines the key characteristics of these methods.

Table . Artificial intelligence (AI) methods and their properties for nurse scheduling optimization. This table provides an overview of 4 AI-based scheduling techniques, and their suitability for optimizing nurse shift scheduling.

Method	Properties
MIP[a]	• Optimization: Finds the best solution from a range of possible options, conditioned by linear constraints. • Modeling: Can integrate complex constraints and objectives. • Computational demand: Can be computationally intensive for large problem sizes. • Flexibility: Adaptable to various scheduling scenarios.
CP[b]	• Focus on constraints: Solves problems through logical conditions and constraint propagation. • Modeling: Suitable for problems with many constraints. • Flexibility: Allows for complex and dynamic constraints. • Solution space: Effective in searching for solutions in large and complex solution spaces.
GP[c]	• Search method: Uses evolutionary theories to find solutions. • Adaptability: Can adjust to different problem types. • Heuristic: Provides good solutions by exploring the solution space but does not guarantee the optimal solution. • Application: Well-suited for problems with many variables and nonlinear constraints that can rather be tested than modeled.
RL[d]	• Learning approach: Learns through rewards and penalties from interactions with the environment. • Adaptability: Can adjust to changes in the environment and requirements. • Short-term planning: Suitable for dynamic and evolving scheduling tasks. • Complexity: Requires a lengthy training phase and extensive data for effective learning.

[a]MIP: mixed-integer programming.

[b]CP: constraint programming.

[c]GP: genetic programming.

[d]RL: reinforcement learning.

A few words on RL, it has not been widely applied to personnel scheduling so far. However, it holds great promise, as traditional scheduling methods often struggle to accommodate the numerous constraints and sudden changes inherent in such environments. RL offers a dynamic and adaptive approach to these challenges. It can be particularly useful in three key areas: (1) optimizing shift exchanges: RL can learn effective strategies for reallocating shifts to improve overall schedule quality while respecting operational constraints. (2) Dynamic and ad-hoc rescheduling: When unforeseen absences or sudden demand fluctuations occur, RL can adjust schedules in real-time to maintain adequate staffing levels. And (3) personalized scheduling across life stages: by considering individual work preferences and constraints, RL can help create schedules that balance organizational needs with personal well-being.

By integrating RL into scheduling systems, health care institutions could move toward more flexible, efficient, and fair staff allocation methods. While its application in this field is still in its early stages, its potential to enhance nurse scheduling and workforce management is clear.

Results

Key Themes From the Interviews and Expectations Regarding AI

A consistent theme across interviews is the demand for fair and participative staff scheduling. Participants stress that in health care environments, it is crucial to include staff in scheduling decisions, consider individual preferences, and maintain transparency. These elements are seen as foundational for promoting satisfaction and efficiency among staff. Stability in shift assignments, particularly on weekends, is important to many. In addition, the ability to swap shifts within teams is viewed as essential for flexibility. Fairness, in this context, implies equal treatment for all employees and the opportunity to contribute to schedule planning, particularly regarding

XSL•FO
RenderX

weekend shifts. Transparent and clear communication is identified as vital for fostering trust in the scheduling process. In outpatient care settings, participative scheduling extends to balancing workload and start times to ensure fairness. For supervisors, achieving a balanced mix of patient cases is critical to prevent specific employees from being consistently assigned complex cases. Overall, the analysis identified five key themes reflected in the summaries, which contribute to overall dissatisfaction and the desire for changes: (1) importance of scheduling as a management tool: staff scheduling is viewed as a critical element in effective management, particularly in structured and high-demand settings like intensive care units; (2) challenges in maintaining fairness: despite established frameworks, ensuring fairness in scheduling is difficult due to varying circumstances and competing needs; (3) complexity in managing absences: handling short-notice absences often relies on informal or ad-hoc methods, highlighting gaps in systematic approaches; (4) flexibility in outpatient care: the dynamic nature of outpatient care, driven by client turnover, necessitates greater flexibility from staff and poses unique scheduling challenges; and (5) impact on work-life balance and employee morale: scheduling decisions directly affect employees' personal lives, often triggering emotional responses, making fairness and transparency critical but difficult to maintain.

In addition to that, the interviewees formulated expectations regarding AI, especially for streamlining the scheduling processes. AI is expected to handle a wide array of individual requirements and preferences, such as avoiding late shifts on specific days or adjusting work schedules based on partial employment percentages. The AI system should also accommodate experience levels and qualifications, manage fluctuating monthly and annual work hours, and efficiently address overtime. In addition, AI should consider nonclinical responsibilities, such as supervisory duties or mentoring, which are often overlooked in traditional scheduling. For outpatient care, it is critical that AI factors in the varying demands of different clients and incorporates lunch breaks into schedules.

Participants also expect AI to serve as a neutral entity that can enhance team dynamics. A well-configured AI system could act as a neutral arbiter, reducing emotional tensions that often arise during scheduling discussions. Ideally, the system would provide a baseline schedule that can be adjusted as needed, considering recent (potentially unfair) scheduling trends and distributing clients more equitably. Participants highlight the potential for increased scheduling accuracy, improved efficiency, and a reduction in both administrative workload and scheduling errors. These improvements could result in more time for patient care, increased employee satisfaction, and enhanced working conditions characterized by transparency and fairness.

Despite the potential advantages of AI-based scheduling, several significant concerns must be addressed. Achieving perfect accommodation for every individual's requirement is nearly impossible, and limitations in the system's ability to address all requests could result in dissatisfaction. Reliability of AI systems is another critical concern. There is apprehension about overreliance on technology, which could lead to operational disruptions in the event of system failures. Furthermore, AI systems depend on the quality of the data they process. Ensuring that this data is accurate and up to date requires ongoing vigilance, presenting an additional challenge. However, this situation also represents an opportunity for digital transformation. Addressing such concerns proactively could improve overall system effectiveness.

Map Between Key Issues and AI Methods

The mapping presented in Table 3 ensures that each key thematic issue is addressed using a method that best meets its unique requirements and challenges. This does not imply that a combination of methods is always necessary in practice. On the one hand, the theoretically most promising approach must also demonstrate practical success, on the other hand, methods such as RL can address many issues autonomously. Therefore, the final choice of mapping depends on context-specific conditions. Our experience suggests that maintaining a repertoire of methods is advantageous, as ensemble-based decisions can enhance validity and objectivity. Relying solely on a single method should be approached with caution, even if it can address numerous issues independently.

XSL•FO
RenderX

Table . Artificial intelligence (AI)–based optimization methods for key staff scheduling challenges. This table presents the most suitable AI methods for resolving critical nurse scheduling challenges identified in Swiss health care institutions between may and June 2024.

Key issue	Most suitable method	Reasoning
Scheduling as a management tool	MIP[a]	MIP is well-suited for optimizing complex scheduling scenarios with defined constraints and objectives, ensuring effective management in structured settings like ICUs[b]
Maintaining fairness	CP[c]	CP focuses on constraints and can handle complex conditions, making it suitable for ensuring fairness and accommodating various needs within scheduling
Managing absences	GP[d]	GP's heuristic approach and adaptability make it effective for handling unpredictable issues such as unplanned absences, exploring various solutions. RL[e] would also be suitable
Flexibility in outpatient care	RL	RL can automatically adjust to dynamic changes and evolving requirements, making it ideal for the flexible and changing nature of outpatient care scheduling.
Work-life balance and employee morale	MIP	MIP's optimization capabilities can help balance employee needs and scheduling constraints, addressing work-life balance while ensuring fairness and efficiency. Methodologically addressed through penalty terms for violating constraints.

[a]MIP: mixed-integer programming.

[b]ICU: intensive care unit.

[c]CP: constraint programming.

[d]GP: genetic programming.

[e]RL: reinforcement learning.

For 2 key issues, MIP emerged as the most suitable method. This is not surprising given that MIP is supported by well-established heuristics such as branch-and-bound [37] and cutting-plane methods [38], which have long been validated for solving complex problems. MIP can also address work-life balance and employee morale by optimizing schedules with penalty terms for constraint violations. However, there are complex scenarios where MIP must be replaced by methods such as GP or RL. This limitation may arise either because the overall constraints are infeasible, a common issue in practice, or because the constraints are too complex for MIP. For these reasons, we recommend maintaining a diverse set of methods. CP is suitable for maintaining fairness by handling complex conditions. GP proves effective for handling unpredictable absences due to its heuristic nature, which facilitates rapid adaptation to changes in existing constraints. In contrast, RL is particularly suited to dynamically changing outpatient care environments, as it allows for the seamless integration of novel constraints.

Implementation of a Mathematical Method

In a specific scenario, we apply MIP to optimize staff scheduling for a training unit in a hospital. The scheduling model covers a representative week. The staff comprises 35 employees categorized into 4 types: 2 doctors, 3 experienced nurses (Type A), 20 moderately experienced nurses (Type B), and 10 less experienced nurses (Type C). Due to the integration of training and care responsibilities, the scheduling involves complex requirements, as most nurses are in training. There are 2 shifts per day, and each employee can either work or not work these shifts. Scheduling is done based on employee type and availability. Each employee is allowed to work a maximum of 6 days per week, with a minimum requirement of 4 days for all employees except doctors. Since doctors have fixed working hours, they are excluded from the scheduling model.

First, with regard to staffing levels per shift, each shift (morning and evening) must consist of at least 3 employees. Specifically, the morning shift must include a minimum of 8 employees, while the evening shift must have at least 3 employees. Second, with regard to specific shift requirements, the morning shift must include at least 3 nurses of types A or B. The morning shift must also include at least 2 nurses of type C. The evening shift must include at least 2 nurses of types A or B. The evening shift must include at least 1 nurse of type C. Third, with regard to external constraints, trainees (under 18 y old) are not permitted to work on Sundays. Employees with school obligations are not allowed to work on the days they have classes. Fourth, with regard to preferred shifts: Employee preferences for specific shifts and days must be taken into account. Shifts that do not align with employee preferences will incur a higher penalty in the objective function.

To formulate these conditions mathematically, we define binary variables xi,j,k, which indicate whether person i works on day j during shift k, as follows: i = 1,..., 33 (employees); j = 1,...7 (days in the week); k = 1, 2 (k=1 represents the morning shift, and k=2 represents the evening shift).

XSL•FO
RenderX

Each variable xi,j,k takes values in {0,1}, where 1 means that employee i is assigned to shift k on day j, and 0 otherwise. On this basis, conditions can be established. For example, to ensure that each employee works at least 4 days per week, the constraint is formulated as follows:

$$\sum j=17\sum k=12xi,j,k\geq 4,i=1,\ldots 33$$

Instead of directly integrating employee types into the indices of binary variables, employee types are modeled through separate variables and then multiplied by the existing binary variables to define constraints. This approach simplifies the modeling and comprehension of constraints and enhances flexibility, as employee types can be reassigned without necessitating changes to the constraints. Further conditions are omitted here for brevity (for details, refer to [35]).

The R package (R Core Team) OMPR was used to solve this problem [39]. The solution requires at least 8 personnel per shift. Given that each employee is scheduled to work at least 4 days, the staffing counts per day and shift reveal inefficiencies: some shifts have up to 20 employees, which is suboptimal. By reducing the minimum number of working days from 4 to 2, a more efficient schedule is achieved. Although this is a training station, and the inefficiency is not critical, it demonstrates that with numerous constraints, even linear ones, inefficiencies can arise rapidly.

We also trained a graph neural network-based RL model for shift assignments, using one million epochs and batch sizes of 1280 to optimize scheduling. However, RL underperformed compared to MIP, failing to converge despite extensive training. This outcome underscores the challenges of applying neural networks to personnel scheduling, where high-dimensional constraints and sparse reward signals can impede learning stability.

Discussion

Principal Findings

The summary of the interviews showed a strong demand for fair and participative staff scheduling within health care environments. Interview participants emphasized the importance of including staff in scheduling decisions, considering individual preferences, and ensuring transparency. These elements are seen as key drivers for job satisfaction and operational efficiency. Fairness in scheduling, particularly around weekend shifts, was a recurring concern, with many advocating for equal opportunities in contributing to the schedule planning process. The ability to swap shifts within teams and balancing workload distribution were identified as critical factors for maintaining flexibility. In outpatient care, there was also an emphasis on adjusting start times and distributing patient cases evenly to prevent any single employee from consistently handling more complex cases. Overall, five key challenges emerged: scheduling as a management tool, maintaining fairness, managing absences, ensuring flexibility in outpatient care, and their collective impact on work-life balance and employee morale.

In addition to traditional scheduling expectations, participants expressed a clear interest in integrating AI to streamline processes. AI was viewed as a potential solution for managing complex individual requirements, such as avoiding late shifts or adjusting schedules based on employment percentages and qualifications. Furthermore, AI was expected to accommodate fluctuating work hours, account for nonclinical responsibilities like mentoring, and improve equity in client case distribution. There was optimism that AI could serve as a neutral entity, reducing emotional tensions and providing an initial schedule that could be adjusted as needed. However, concerns were raised about the reliability of AI systems, potential loss of human elements in decision-making, and the importance of maintaining supervisor oversight to ensure flexibility and contextual understanding.

The study's mapping of key scheduling issues to AI methods highlights the importance of aligning specific challenges with appropriate technological solutions. Each identified key issue requires distinct approaches to optimize scheduling processes. For instance, MIP is deemed most suitable for managing structured, complex scheduling scenarios with defined constraints. CP excels in addressing fairness by focusing on complex constraints and balancing competing needs. GP is selected for its heuristic approach to handling highly complex constraints, while RL offers the adaptability needed for the dynamic nature of staff scheduling. This method-specific approach ensures that each challenge is addressed using the most effective technique, potentially enhancing scheduling efficiency and overall staff satisfaction. Maintaining a repertoire of methods allows for a more flexible and comprehensive solution and mutual validation, acknowledging that no single method is universally applicable.

Our results also indicate the challenges of applying RL to nurse scheduling. While RL offers adaptability, our graph neural network-based RL model struggled to converge despite extensive training, likely due to high-dimensional constraints and sparse reward signals. Future improvements, such as enhanced reward shaping, constraint-aware RL architectures, or hybrid approaches combining RL with rule-based methods, may be necessary to improve RL's effectiveness in this complex domain. By integrating RL's adaptability with the precision of optimization techniques, future research could refine AI-driven scheduling for dynamic health care environments.

Comparison With Previous Work

This study aligns with previous research emphasizing the complexity and challenges of staff scheduling in health care settings [40]. Similar studies have highlighted the difficulties in balancing fairness, flexibility, and operational efficiency [41]. By offering a challenge-based mapping of the most pertinent methodologies, this study advances the discourse surrounding staff scheduling, emphasizing the importance of a comprehensive and integrative approach.

Limitations

This study has several limitations. First, the reliance on interviews for data collection may introduce subjective biases and limit the generalizability of findings. Nurses and supervisors may have different perspectives that are not fully captured through qualitative interviews alone. In addition, the study's focus on specific hospital and outpatient care settings may not

XSL•FO
RenderX

fully represent the diverse range of health care environments, potentially limiting the applicability of findings to other contexts.

Another limitation lies in the fact that the mapping between key issues and AI methods was based on the expertise of one expert, rather than being grounded in a more objective, data-driven approach. Achieving a comprehensive, objective mapping would require extensive knowledge, not only of the diverse AI techniques but also of the practical demands of health care environments. This would necessitate a significant collaborative effort across multiple disciplines, including ML, optimization, human resource management, and health care.

We did not specifically address how to accurately capture employees' preferences. Often, individuals may lack full awareness of their own desires or what would be most beneficial to them. However, this raises a sensitive issue, as it involves the risk of paternalism. Not every form of objectification necessarily enhances well-being, given that autonomy is a highly valued principle. In a separate context, we are working on a system that seeks to balance objectification and autonomy in the process of eliciting preferences.

Furthermore, the study acknowledges the limitations of AI in addressing the emotional and psychological states of employees. While AI can improve scheduling efficiency, it lacks the sensitivity needed to manage complex human factors, which could negatively impact the overall effectiveness of the scheduling process. This limitation highlights broader concerns about the role of AI in decision-making processes that involve significant human interaction and emotional considerations.

Conclusions

AI-based scheduling presents a promising solution, offering potential benefits such as increased accuracy, efficiency, and reduced administrative workload. However, it is essential to address concerns related to the limitations of AI, including its ability to fully accommodate individual preferences and the risk of overreliance on technology. A balanced approach that incorporates both advanced scheduling methods and human oversight is crucial to achieving optimal outcomes. Future research should investigate ways to enhance the integration of AI with human judgment to address the emotional and psychological aspects of scheduling. Finally, it is worthwhile to investigate whether large language models are capable of independently conducting staff scheduling [42].

Data Availability

The datasets generated and analyzed during this study are not publicly available due to privacy or ethical restrictions but are available on request from the corresponding author.

Authors' Contributions

MS and CG supervised the whole study. MG conceptualized the interviews. MG, FJR and JSB conducted the interviews and have documented and evaluated the findings. CG has reviewed the interview guides and the results. MS drafted the manuscript, with all authors reviewing and approving the final version.

Conflicts of Interest

None declared.

References

1. Barger LK, Weaver MD, Sullivan JP, Qadri S, Landrigan CP, Czeisler CA. Impact of work schedules of senior resident physicians on patient and resident physician safety: nationwide, prospective cohort study. BMJ Med 2023;2(1):e000320. [doi: 10.1136/bmjmed-2022-000320] [Medline: 37303489]
2. Ejebu OZ, Dall'Ora C, Griffiths P. Nurses' experiences and preferences around shift patterns: a scoping review. PLoS One 2021;16(8):e0256300. [doi: 10.1371/journal.pone.0256300]
3. Fadillah AR, Aras RA, Wahyuni S. The effects of work-life balance towards work stress among nurses. In: Proceedings of the Interdisciplinary Conference of Psychology, Health, and Social Science (ICPHS 2021): Atlantis Press; 2022:232-238. [doi: 10.2991/assehr.k.220203.035]
4. Silva I, Costa D. Consequences of shift work and night work: a literature review. Healthcare (Basel) 2023 May 12;11(10):37239693. [doi: 10.3390/healthcare11101410] [Medline: 37239693]
5. Harris AM, Teplitsky S, Kraft KH, Fang R, Meeks W, North A. Burnout: a call to action from the AUA workforce workgroup. J Urol 2023 Mar;209(3):573-579. [doi: 10.1097/JU.0000000000003108] [Medline: 36598404]
6. Luther L, Gearhart T, Fukui S, Morse G, Rollins AL, Salyers MP. Working overtime in community mental health: Associations with clinician burnout and perceived quality of care. Psychiatr Rehabil J 2017 Jun;40(2):252-259. [doi: 10.1037/prj0000234]
7. Uhde A, Schlicker N, Wallach DP, Hassenzahl M. Fairness and decision-making in collaborative shift scheduling systems. Presented at: CHI '20: CHI Conference on Human Factors in Computing Systems; Apr 25-30, 2020; Honolulu, HI, USA. [doi: 10.1145/3313831.3376656]
8. Lembrechts L, Dekocker V, Zanoni P, Pulignano V. A study of the determinants of work-to-family conflict among hospital nurses in Belgium. J Nurs Manag 2015 Oct;23(7):898-909. [doi: 10.1111/jonm.12233]

XSL•FO
RenderX

9. Baljani E, Rezaee Moradali M, Hajiabadi NR. The nurse manager support process in the work–family conflict of clinical nurses: a qualitative study. J Res Nurs 2023 Nov;28(6-7):499-513. [doi: 10.1177/17449871231204537]
10. Booker LA, Fitzgerald J, Mills J, et al. Sleep and fatigue management strategies: how nurses, midwives and paramedics cope with their shift work schedules-a qualitative study. Nurs Open 2024 Jan;11(1):e2099. [doi: 10.1002/nop2.2099] [Medline: 38268269]
11. Sarıtaş CT. Precarious contours of work–family conflict: the case of nurses in Turkey. The Economic and Labour Relations Review 2020 Mar;31(1):59-75. [doi: 10.1177/1035304619879327]
12. Messing K, Tissot F, Couture V, Bernstein S. Strategies for managing work/life interaction among women and men with variable and unpredictable work hours in retail sales in Québec, Canada. New Solut 2014;24(2):171-194. [doi: 10.2190/NS.24.2.d] [Medline: 25085829]
13. Christopher K. A double bind of relational care: nurses' narratives of caregiving at work and at home. Gend Issues 2022 Jun;39(2):220-235. [doi: 10.1007/s12147-021-09283-6]
14. Kim JY, Norasi H, Navarro SM, Park MS, Hallbeck S. Shiftwork and burnout among trauma surgeons: a preliminary analysis of trauma surgeons' fatigue and stress related to overnight shifts. Proceedings of the Human Factors and Ergonomics Society Annual Meeting 2024 Sep;68(1):499-503. [doi: 10.1177/10711813241263812]
15. Mrayyan MT. Nurses' autonomy: influence of nurse managers' actions. J Adv Nurs 2004 Feb;45(3):326-336. [doi: 10.1046/j.1365-2648.2003.02893.x] [Medline: 14720250]
16. Christensen SS, Nixon TL, Aguilar RL, Muhamedagic Z, Mahoney K. Using participatory management to empower nurses to identify and prioritize the drivers of their burnout. Int J Healthc Manag 2024 Apr 2;17(2):389-397. [doi: 10.1080/20479700.2023.2190251]
17. Albargi F, Albloushi M, Alghamdi RS. A qualitative exploration of nurses' perspectives, experiences, and obstacles with shared governance. Int J Healthc Manag 2024 Jul 27:1-8. [doi: 10.1080/20479700.2024.2381172]
18. Shiri R, Karhula K, Turunen J, et al. The effect of using participatory working time scheduling software on employee well-being and workability: a cohort study analysed as a pseudo-experiment. Healthcare (Basel) 2021 Oct 16;9(10):1385. [doi: 10.3390/healthcare9101385] [Medline: 34683065]
19. Abdalkareem ZA, Amir A, Al-Betar MA, Ekhan P, Hammouri AI. Healthcare scheduling in optimization context: a review. Health Technol (Berl) 2021;11(3):445-469. [doi: 10.1007/s12553-021-00547-5] [Medline: 33868893]
20. Burke EK, De Causmaecker P, Berghe GV, Van Landeghem H. The state of the art of nurse rostering. Journal of Scheduling 2004 Nov 1;7(6):441-499. [doi: 10.1023/B:JOSH.0000046076.75950.0b]
21. Chen Z, Dou Y, De Causmaecker P. Neural networked-assisted method for the nurse rostering problem. Computers & Industrial Engineering 2022 Sep;171:108430. [doi: 10.1016/j.cie.2022.108430]
22. Svirsko AC, Norman BA, Rausch D, Woodring J. Using mathematical modeling to improve the emergency department nurse-scheduling process. J Emerg Nurs 2019 Jul;45(4):425-432. [doi: 10.1016/j.jen.2019.01.013]
23. Saville CE, Griffiths P, Ball JE, Monks T. How many nurses do we need? A review and discussion of operational research techniques applied to nurse staffing. Int J Nurs Stud 2019 Sep;97:7-13. [doi: 10.1016/j.ijnurstu.2019.04.015]
24. Alves MJ, Clímaco J. A review of interactive methods for multiobjective integer and mixed-integer programming. Eur J Oper Res 2007 Jul;180(1):99-115. [doi: 10.1016/j.ejor.2006.02.033]
25. Rossi F, Beek P, Walsh T. Chapter 4 constraint programming. In: van Harmelen F, Lifschitz V, Porter B, editors. Foundations of Artificial Intelligence: Elsevier; 2008:181-211. [doi: 10.1016/S1574-6526(07)03004-0]
26. Bonetta G, Zago D, Cancelliere R, Grosso A. Job shop scheduling via deep reinforcement learning: a sequence to sequence approach. arXiv. Preprint posted online on Aug 3, 2023. [doi: 10.48550/arXiv.2308.01797]
27. Kayhan BM, Yildiz G. Reinforcement learning applications to machine scheduling problems: a comprehensive literature review. J Intell Manuf 2023 Mar;34(3):905-929. [doi: 10.1007/s10845-021-01847-3]
28. Lee S, Lee YH. Improving emergency department efficiency by patient scheduling using deep reinforcement learning. Healthcare (Basel) 2020 Mar 27;8(2):77. [doi: 10.3390/healthcare8020077] [Medline: 32230962]
29. Song L, Li Y, Xu J. Dynamic job-shop scheduling based on transformer and deep reinforcement learning. Processes (Basel) 2023 Dec;11(12):3434. [doi: 10.3390/pr11123434]
30. Tassel P, Gebser M, Schekotihin K. An end-to-end reinforcement learning approach for job-shop scheduling problems based on constraint programming. arXiv. Preprint posted online on Jun 9, 2023. [doi: 10.48550/arXiv.2306.05747]
31. Koruca H, Emek MS, Gulmez E. Development of a new personalized staff-scheduling method with a work-life balance perspective: case of a hospital. Ann Oper Res 2023 Feb 28;328:1-28. [doi: 10.1007/s10479-023-05244-2] [Medline: 37361084]
32. Gattermann-Itschert T, Poreschack LM, Thonemann UW. Using machine learning to Include planners' preferences in railway crew scheduling optimization. Transportation Science 2023 May;57(3):796-812. [doi: 10.1287/trsc.2022.1196]
33. Gerlach M, Fabienne JR, Jannic SB, Murat S, Golz C. Exploring nurse perspectives on AI-based shift scheduling for fairness, transparency, and work-life balance. Research Square. Preprint posted online on Oct 23, 2024. [doi: 10.21203/rs.3.rs-5248710/v1]
34. McLafferty I. Focus group interviews as a data collecting strategy. J Adv Nurs 2004 Oct;48(2):187-194. [doi: 10.1111/j.1365-2648.2004.03186.x] [Medline: 15369499]

XSL•FO
RenderX

35. Kendall J. Axial coding and the grounded theory controversy. West J Nurs Res 1999 Dec;21(6):743-757. [doi: 10.1177/019394599902100603] [Medline: 11512211]
36. Wirsansky E. Hands-On Genetic Algorithms with Python: Applying Genetic Algorithms to Solve Real-World Deep Learning and Artificial Intelligence Problems: Packt Publishing; 2020.
37. Morrison DR, Jacobson SH, Sauppe JJ, Sewell EC. Branch-and-bound algorithms: a survey of recent advances in searching, branching, and pruning. Discrete Optimization 2016 Feb;19:79-102. [doi: 10.1016/j.disopt.2016.01.005]
38. Atlason J, Epelman MA, Henderson SG. Call center staffing with simulation and cutting plane methods. Ann Oper Res 2004 Mar 1;127(1-4):333-358. [doi: 10.1023/B:ANOR.0000019095.91642.bb]
39. Schumacher D. Ompr: model and solve mixed integer linear programs. The Comprehensive R Archive Network. 2022. URL: https://CRAN.R-project.org/package=ompr [accessed 2022-07-28]
40. Webster B, Archibald D. Self-rostering, work-life balance and job satisfaction in UK nursing: a literature review. Nurs Manag (Harrow) 2022 Jul 26;29(6):27-32. [doi: 10.7748/nm.2022.e2048] [Medline: 35880476]
41. Leung F, Lau YC, Law M, Djeng SK. Artificial intelligence and end user tools to develop a nurse duty roster scheduling system. Int J Nurs Sci 2022 Jul;9(3):373-377. [doi: 10.1016/j.ijnss.2022.06.013] [Medline: 35891913]
42. Abgaryan H, Harutyunyan A, Cazenave T. LLMs can schedule. arXiv. Preprint posted online on Aug 13, 2024. [doi: 10.48550/arXiv.2408.06993]

Abbreviations

AI: artificial intelligence
COREQ: Consolidated Criteria for Reporting Qualitative Research
CP: constraint programming
GP: genetic programming
MIP: mixed-integer programming
ML: machine learning
RL: reinforcement learning

Edited by A Mavragani; submitted 20.10.24; peer-reviewed by M Gandhi; revised version received 17.02.25; accepted 07.03.25; published 04.06.25.

Please cite as:
Renggli FJ, Gerlach M, Bieri JS, Golz C, Sariyar M
Integrating Nurse Preferences Into AI-Based Scheduling Systems: Qualitative Study
JMIR Form Res 2025;9:e67747
URL: https://formative.jmir.org/2025/1/e67747
doi:10.2196/67747

© Fabienne Josefine Renggli, Maisa Gerlach, Jannic Stefan Bieri, Christoph Golz, Murat Sariyar. Originally published in JMIR Formative Research (https://formative.jmir.org), 4.6.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Exploring Generative Pre-Trained Transformer-4-Vision for Nystagmus Classification: Development and Validation of a Pupil-Tracking Process

Masao Noda[1,2], MD, MBA, PhD; Ryota Koshu[2], MD; Reiko Tsunoda[1], MD, PhD; Hirofumi Ogihara[1], PhD; Tomohiko Kamo[1], PhD; Makoto Ito[2], MD, PhD; Hiroaki Fushiki[1], MD, PhD

[1]Department of Otolaryngology, Mejiro University Ear Institute Clinic, 320 Ukiya, Iwatsuki-ku, Saitama-shi, Saitama, Japan
[2]Department of Otolaryngology, Jichi Medical University, Shimotsuke, Japan

Corresponding Author:
Masao Noda, MD, MBA, PhD
Department of Otolaryngology, Mejiro University Ear Institute Clinic, 320 Ukiya, Iwatsuki-ku, Saitama-shi, Saitama, Japan

Abstract

Background: Conventional nystagmus classification methods often rely on subjective observation by specialists, which is time-consuming and variable among clinicians. Recently, deep learning techniques have been used to automate nystagmus classification using convolutional and recurrent neural networks. These networks can accurately classify nystagmus patterns using video data. However, associated challenges including the need for large datasets when creating models, limited applicability to address specific image conditions, and the complexity associated with using these models.

Objective: This study aimed to evaluate a novel approach for nystagmus classification that used the Generative Pre-trained Transformer 4 Vision (GPT-4V) model, which is a state-of-the-art large-scale language model with powerful image recognition capabilities.

Methods: We developed a pupil-tracking process using a nystagmus-recording video and verified the optimization model's accuracy using GPT-4V classification and nystagmus recording. We tested whether the created optimization model could be evaluated in six categories of nystagmus: right horizontal, left horizontal, upward, downward, right torsional, and left torsional. The traced trajectory was input as two-dimensional coordinate data or an image, and multiple in-context learning methods were evaluated.

Results: The developed model showed an overall classification accuracy of 37% when using pupil-traced images and a maximum accuracy of 24.6% when pupil coordinates were used as input. Regarding orientation, we achieved a maximum accuracy of 69% for the classification of horizontal nystagmus patterns but a lower accuracy for the vertical and torsional components.

Conclusions: We demonstrated the potential of versatile vertigo management in a generative artificial intelligence model that improves the accuracy and efficiency of nystagmus classification. We also highlighted areas for further improvement, such as expanding the dataset size and enhancing input modalities, to improve classification performance across all nystagmus types. The GPT-4V model validated only for recognizing still images can be linked to video classification and proposed as a novel method.

(JMIR Form Res 2025;9:e70070) doi:10.2196/70070

KEYWORDS

nystagmus; GPT-4Vision; generative AI; deep learning; dizziness; artificial intelligence

Introduction

Equilibrium function in vertigo practice can be evaluated through nystagmus assessment. Nystagmus is characterized by rhythmically repeated rapid and slow eye movements and serves as a valuable clinical indicator for diagnosing various neurological and vestibular disorders. Nystagmus can influence the normal function of the cerebellum, semicircular canals, and integrated eye movements, and thus is of great diagnostic and therapeutic importance [1-3]. The direction of nystagmus can be horizontal, vertical, or torsional (rotational), depending on the axis of the eye movement. The evaluation of nystagmus patterns provides essential insight into the function of the visual and vestibular systems by identifying underlying foci and guiding treatment strategies. Traditionally, nystagmus classification has relied heavily on subjective observation by trained specialists, which is time-consuming, prone to variability among clinicians, and can be difficult to perform in the emergency department [4]. Furthermore, advancement of diagnosis can contribute to improper treatment, increasing the risk of falls or decreasing daily physical activity levels [5,6].

XSL•FO
RenderX

In recent years, advances in artificial intelligence (AI) and machine learning technologies have provided promising means to capture eye movements [7,8] and automate nystagmus classification, thereby improving the accuracy and efficiency of diagnosis [9,10]. In particular, deep learning methods such as convolutional and recurrent neural networks are increasingly being used to analyze medical imaging data, including videos that capture eye movements [11,12]. Although these techniques have shown success in tasks such as image classification, object detection, and segmentation, applying deep learning to nystagmus classification remains challenging given the temporal changes inherent in eye movement patterns. Interestingly, recent studies have demonstrated promising outcomes by using deep learning techniques to annotate scenes or detect spatiotemporal features. This suggests that the potential of deep learning algorithms to classify nystagmus patterns based on their capture, and interpretation of these temporal characteristics requires sophisticated modeling methods that can effectively process ordinal data. Notably, several reports have used deep learning to enable a unified evaluation of other perceptions; however, creating models requires large amounts of data, and it is unclear how such models can be used.

A large-scale language processing model, known as a large language model (LLM) is a highly versatile system trained on extensive text data using a transformer architecture. It has demonstrated high accuracy in medical classification and text recognition [13]. In one such model, the Generative Pre-Trained Transformer (GPT), the advent of GPT-4Vision (GPT-4V) made it possible to combine image recognition by devising prompts and classification without requiring image training data [14,15]. Unlike convolutional neural networks or recurring neural networks, GPT-4V leverages in-context learning, allowing it to classify visual and multimodal data without extensive training datasets [16]. In this study, we developed a novel nystagmus classification approach that leverages the capabilities of the GPT-4V model. We aimed to develop a classification system that can accurately identify different nystagmus patterns and validate its accuracy by integrating GPT-4V with an eye movement tracking algorithm from eye movement video data. This study is one of the first to evaluate the feasibility of GPT-4V for nystagmus classification, particularly in scenarios where dataset limitations exist.

Methods

In this study, we developed a pupil-tracking process using nystagmus recording videos and verified the accuracy of the optimization model using GPT-4V classification and nystagmus recording (Figure 1).

Figure 1. Study overview. GPT-4V: Generative Pre-trained Transformer 4 Vision.

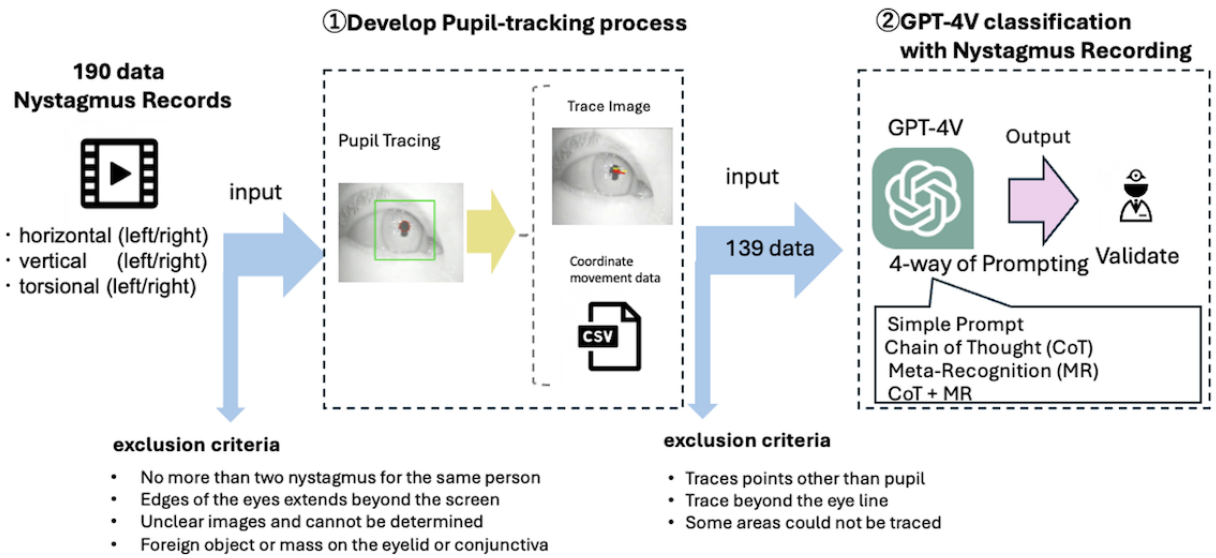

Developing the Pupil Tracking Process

First, the eyeballs were recognized from the video data, and pupil movement co-ordinate data were created using an eye movement tracking algorithm that showed the eyeball trajectory. Video data were recorded using a charge-coupled device-based camera eyeball rotation imaging device, ET-60LW2 (Newopto), with a focal length of 6 mm, and a horizontal resolution >500 television lines. The sensor size is 1/3 inch and intermittent synchronous lighting occurred every 1/30 second. A single video contained at least three nystagmus in the same direction and was 3 to 5 seconds in duration.

Based on these data, 190 trace images were created for each video from patients whose main complaint was dizziness. Videos were created using the same algorithm used to generate trace images of the same individual, with no more than two nystagmus events. After applying the exclusion criteria, trace data were successfully generated for 139 patients. The exclusion criteria included cases in which the edges of the eyes extended beyond the screen, images that were too unclear to be evaluated, and the presence of foreign objects or masses on the eyelid or conjunctiva.

XSL•FO
RenderX

Pupil Movement Tracking Algorithm

An eye movement tracking algorithm based on video data was developed using the Haar cascade classifier and OpenCV [17]. Although deep learning–based approaches may offer higher accuracy for detecting facial features, they often require extensive data or a pretrained model, as well as comparatively intensive inference processes. The outline of the pupil and its center within the screen were detected, enabling the tracking of the center's coordinates. In contrast, the Haar cascade classifier, as part of the OpenCV suite, offers a more streamlined and efficient alternative, enabling the tracking of the eye's trajectory at a rate of 40 frames per second with significantly reduced complexity and setup time, as reported previously [18-20]. The algorithm was designed to superimpose trajectory data onto the first frames of eye movement videos, marking the starting and ending points of the pupil center's paths. If the algorithm failed to detect the eye due to occlusions or rapid movements, the trajectories were interpolated, and measurement points that could not be captured were omitted from the trajectory path. Additionally, the movement data comprising the x and y coordinates of the pupil at each measurement time point were systematically converted into CSV files for further analysis using LLMs.

This approach proved strategic for our study's requirements, allowing for rapid development and ease of modification. This was particularly beneficial in our context, in which real-time processing was prioritized over the incremental gains in accuracy afforded by more computationally intensive models.

GPT-4V Classification With Nystagmus Recording

We developed GPT-4V models to classify eye movement trajectories by generating still images tracked by the algorithm and then inputting these images or CSV data into the GPT-4V model to obtain answers for the classification. When only traced images or coordinates were input, no significant or advantageous responses were obtained. For CSV data inputs containing the pupils' coordinates (X and Y axes) and their respective measurements, we embedded these data directly into the prompts for the GPT-4 model. We tested three combinations of inputs: only CSV, which used GPT-4 (GPT-4 Turbo); only still images; and a combination of CSV and still images, with the latter two using GPT-4V (GPT-4 Turbo with vision).

A feature of LLMs is in-context learning through prompting. The model was developed using prompts based on previous studies [15,21,22] and the chain-of-thoughts (CoT) prompting technique, which allows LLMs to make complex inferences by entering thought processes into the prompts to facilitate the inference process and reasoning [23,24] (Multimedia Appendix 1). We also used methods, such as metarecognition (MR) and the Rule of 26, which complicates the thought process by making the user aware of the content in reaction to their responses [25,26]. We tested whether the created optimization model could be evaluated in six nystagmus categories: right horizontal, left horizontal, upward, downward, right torsional, and left torsional. Of the trace images obtained from the pupil-tracking process, those with trace points other than the pupil, traces beyond the eyeline, or some areas that could not be traced were excluded by visual inspection by experts. The correctness rate was evaluated for 139 data points (78, 26, and 35 in the horizontal, vertical, and torsional directions, respectively).

An application programming interface was set up for model validation, and the temperature parameter was set to 0 to account for variations in responses. Experts with >20 years of vertigo practice experience judged whether the answers and the explanatory content were appropriate. Even if the final answer was correct, the details that led to the answer were checked, and if they were incorrect, the answer was considered incorrect. Data collected from the video recordings of eye movements exhibiting various nystagmus patterns were used. Each video clip was independently reviewed and annotated by two skilled vertigo specialists to identify the presence and type of nystagmus pattern.

Ethical Considerations

This study was approved by the Medical Research Ethics Committee of Mejiro University (approval number: Medical 20‐007). Informed consent was obtained from all subjects involved in the study. Written informed consent has been obtained from the patients to publish this paper, as applicable. In accordance with ethical guidelines, an opt-out approach was adopted. Detailed information regarding the study—including its purpose, data handling procedures, and measures for protecting personal information—was made publicly available, and participants were given the opportunity to decline participation if they wished. All data collected in this study were anonymized, ensuring that no personally identifiable information was included in the analysis or publication. No financial or material compensation was provided to the participant.

Study/Clinical Setting of Recruitment

Participants were recruited in this study from the otolaryngology outpatient clinic at Mejiro University, a tertiary center specializing in the diagnosis and treatment of vestibular disorders. The patient included in this study was referred to the clinic after experiencing episodes of vertigo. During the clinical examination, ocular movements were recorded and the resulting video data were used for this study.

Results

In this study, we developed a system to classify nystagmus using GPT-4V and obtained the following results.

Pupil Tracking Process

The eye movement tracking program accurately recognized the eyeballs in the video and depicted their trajectories (Figure 2). The program successfully tracked eye movements and generated datasets for each nystagmus category.

XSL•FO
RenderX

Figure 2. Representative tracing images of six nystagmus types (ie, right horizontal, left horizontal, upward, downward, right torsional, and left torsional) obtained from video-based pupil tracking. Data were collected from patients with vestibular disorders. The images illustrate pupil movement trajectories detected using an AI-based classification model.

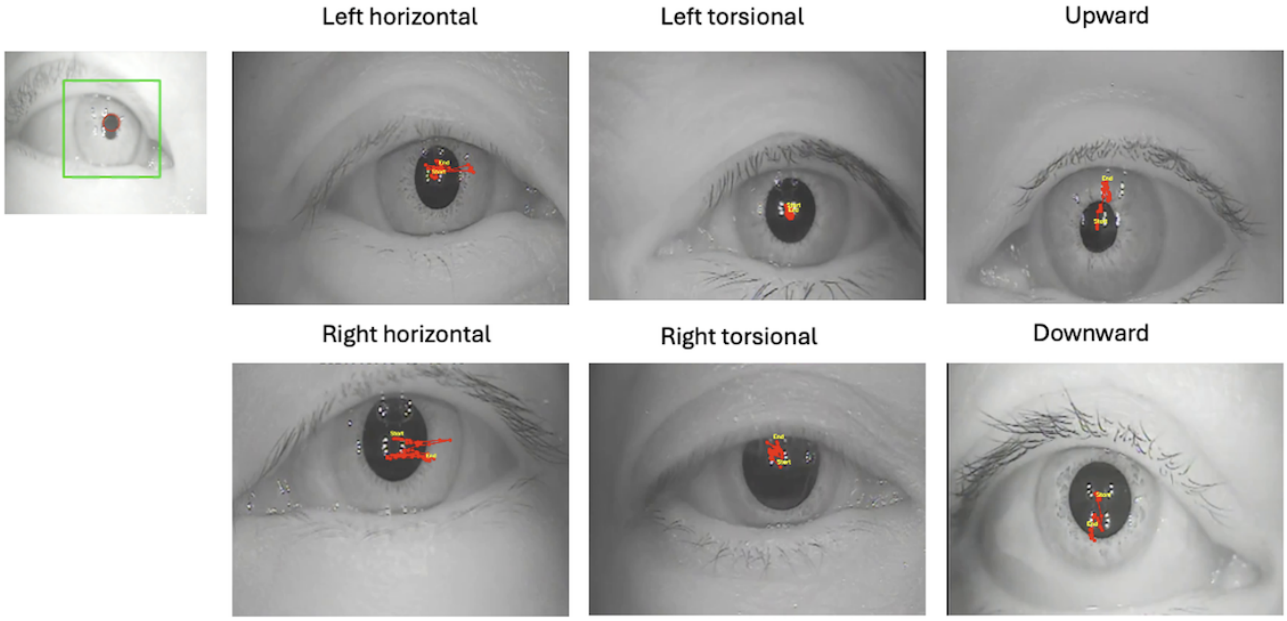

GPT-4V Prompting

A nystagmus classification model was constructed by setting and training appropriate prompts on the GPT-4V model. Optimizing the prompts improved the classification accuracy and adapted the model's response to specific nystagmus patterns. We used both basic and additional prompts for CoT and MR (Multimedia Appendix 1).

Validation of Nystagmus With GPT-4V

For the "image only" input category, the "basic" prompt method yielded a correct response in 43 (30.9%) instances, whereas no response was recorded in 24 (17.3%) instances (Table 1). The "CSV only" input category under the "basic" prompt approach resulted in 27 (19.4%) correct responses, with 14 (10.1%) instances of no response. When both images and CSV input were used, the "basic" prompt method delivered 37 (26.6%) correct responses, with no response in 12 (8.6%) instances.

Table . Performance of GPT-4V in classifying nystagmus patterns based on video-recorded eye movements. The table presents the classification accuracy for six nystagmus types across different input modalities (image only, CSV only, and image+CSV). Data were collected from 139 patients diagnosed with vestibular disorders.

Prompting techniques	Input modalities and classification accuracy (N=139)								
	Image only			CSV only			Image+CSV		
	Correct, n (%)	No response, n (%)	Correct response (%)	Correct, n (%)	No response, n (%)	Correct response (%)	Correct, n (%)	No response, n (%)	Correct response (%)
Basic	43 (30.9)	24 (17.3)	37.4	27 (19.4)	14 (10.1)	21.6	37 (26.6)	12 (8.6)	29.1
CoT	47 (33.8)	13 (9.4)	37.3	20 (14.4)	26 (18.7)	17.7	33 (23.7)	32 (23.0)	30.8
MR	50 (36.0)	5 (3.6)	37.3	28 (20.1)	25 (18.0)	24.6	30 (21.6)	33 (23.7)	28.3
CoT+ MR	48 (34.5)	12 (8.6)	37.8	15 (10.8)	69 (49.6)	21.4	35 (25.2)	32 (23.0)	32.7

The other methods showed similar trends, with the CoT prompt approach slightly improving in the "image only" input category with 47 (33.8%) correct responses. Conversely, the MR prompt method outperformed the others in the "image only" input domain, with 50 (36.0%) correct classifications. When combining the CoT and MR prompts, the "image only" input domain showed 48 (34.5%) correct classifications. However, there was a modest improvement in the "image+CSV" input category, with 35 (25.2%) correct responses.

Table 2 shows the outcome of the GPT-4V classification accuracy in the presence of nystagmus-recording data segregated according to each nystagmus direction. Four types of prompting classification strategies were assessed: basic, CoT, MR, and a composite of CoT and MR. The evaluation was further stratified into three data inputs: image only, CSV only, and Image+CSV.

For downward and upward nystagmus, the highest correct classification rates were 37.5% and 27.8%, with a total data count of 8 and 18 instances, respectively.

XSL•FO
RenderX

Table . GPT-4V classification with nystagmus recording for each direction.

Nystagmus direction	Classification strategies and data input categories											
	Image only (correct classification rates), n (%)				CSV only (correct classification rates), n (%)				Image+CSV (correct classification rates), n (%)			
	Basic	CoT	MR	CoT+MR	Basic	CoT	MR	CoT+MR	Basic	CoT	MR	CoT+MR
Downward (n=8)[a], n (%)	0 (0)	1 (12.5)	0 (0)	2 (25)	2 (25)	0 (0)	2 (25)	0 (0)	1 (12.5)	0 (0)	0 (0)	3 (37.5)
Left horizontal (n=36)[a], n (%)	14 (38.9)	14 (38.9)	20 (55.6)	14 (38.9)	3 (8.3)	2 (5.6)	3 (8.3)	0 (0)	9 (25)	7 (19.4)	9 (25)	7 (19.4)
Left torsional (n=16)[a], n (%)	0 (0)	0 (0)	1 (6.3)	0 (0)	0 (0)	3 (18.8)	1 (6.3)	3 (18.8)	0 (0)	1 (6.3)	0 (0)	0 (0)
Right horizontal (n=42)[a], n (%)	27 (64.3)	27 (64.3)	25 (59.5)	29 (69.0)	20 (47.6)	15 (35.7)	17 (40.5)	10 (23.8)	23 (54.8)	23 (54.8)	17 (40.5)	22 (52.4)
Right torsional (n=19)[a], n (%)	1 (5.3)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (5.3)	0 (0)	0 (0)	0 (0)	0 (0)
Upward (n=18)[a], n (%)	1 (5.6)	5 (27.8)	4 (22.2)	3 (16.7)	2 (11.1)	0 (0)	5 (27.8)	1 (5.6)	4 (22.2)	2 (11.1)	4 (22.2)	3 (16.7)

[a]Total number of nystagmus cases examined.

For left and right torsional nystagmus, the highest correct classification rates were 6.3% and 5.3% with the “image only” inputs, 18.8% and 5.3% with “CSV only” inputs, and 6.3% and 0% with “image+CSV” inputs, respectively.

For the left horizontal nystagmus with 36 data instances, the “image only” input showed superior performance with 38.9% correct classifications using the basic prompt, improving incrementally with the MR prompt at 55.6%. In the right horizontal category with 42 instances, the “image only” input showed a higher correct rate compared to the other input types, with the basic and CoT prompts demonstrating an accuracy rate of 64.3%, while the combined CoT+ MR prompt exhibited the highest accuracy rate at 69%.

The “CSV only” mode indicated a generally lower correct classification rate across all directions and methodologies. Notably, the right torsional and left torsional classifications demonstrated zero correct classifications with basic prompts.

Overall, Table 2 highlights the varying degrees of classification accuracy depending on the direction of nystagmus, data presentation format, and classification method. Inputs containing image data generally showed improved classification performance compared to using CSV only” input.

The response trend for each type of data input was analyzed in terms of each direction (Figure 3). The classification performance of the GPT model was evaluated using three different input modalities: image only, CSV only, and image+ CSV. When using the “image only” input, the model achieved an accuracy of 0.289, with a precision of 0.211, a recall of 0.264, and an F_1-score of 0.169. For the “CSV only” input, the accuracy was 0.247, with a precision of 0.218, a recall of 0.210, and an F_1-score of 0.178. The combination of image and CSV inputs (image+ CSV) resulted in the highest performance among the three input types, with an accuracy of 0.356, a precision of 0.186, a recall of 0.222, and an F_1-score of 0.191. The confusion matrix showed a high frequency of horizontal responses, and this tendency was greater for “image only” and “image+ CSV” inputs. For inputs containing images, horizontal nystagmus tended to be misclassified more often, as horizontal nystagmus in the opposite direction occurred more frequently than in the other components. Results for the “CSV Only” input showed a broader distribution of misclassifications across multiple categories, with no specific tendency toward a particular type of nystagmus. Additionally, a notably high number of responses were categorized as “others,” indicating a difficulty in making definitive classifications using CSV data alone. In comparison, the “Image” input demonstrated a reduction in “Others”

XSL•FO
RenderX

responses, highlighting improved performance and specificity when combining data modalities.

Figure 3. Confusion matrix of GPT-4V classification with nystagmus recording for each direction.

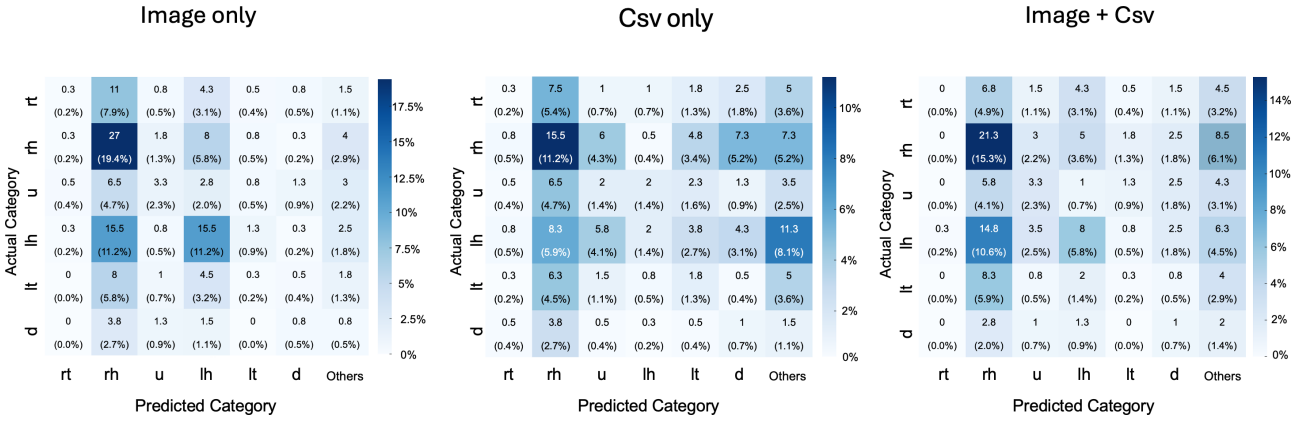

Discussion

Principal Findings

In this study, we created and verified the accuracy of a nystagmus classification model using the GPT-4V. The results revealed several interesting insights. First, it became clear that nystagmus classification is possible using AI generated from a LLM model. Furthermore, the model's accuracy varied with different prompting adjustments, indicating that the accuracy varied with each nystagmus pattern, suggesting room for improvement in the model's performance without retraining the model or parameter tuning. Second, it was revealed that inputting images rather than tracking the pupil coordinates resulted in higher accuracy. Thus, the GPT-4V–based nystagmus classification model achieved a certain level of success, and this study serves as the first step toward validating its potential for video analysis.

The GPT-4V-based nystagmus classification model derived from LLM demonstrated the ability to distinguish and classify different nystagmus patterns from video data. Overall, an accuracy rate of 17 - 38% was achieved, with the classification of horizontal nystagmus patterns showing a 70% accuracy rate. This indicates that the GPT-4V model can effectively capture subtle differences in horizontal eye movements and suggests specific neurological or vestibular states. The fact that nystagmus classification is possible with an LLM model implies the potential for further development of nystagmus classification models previously tackled using deep learning models. The classification accuracy of deep learning models has been reported to be between 60 - 90%, averaging approximately 80%, whereas the GPT-4V similar to LLM, has achieved a certain level of accuracy for horizontal nystagmus [27-29]. As LLMs do not require image training data and can be used in conversations, they can be readily used alongside other information during clinical assessments to confirm answers and diagnostics. At present, LLMs including GPT have improved learning accuracy and can generate still images and audio; however, video recognition has not yet been reported. In the future, our methodology can be applied to videos using the LLM.

Notably, accuracy varies with prompts and conditions [30,31]. We also report that the accuracy of medical licensing and otolaryngology expertise examinations can be improved by presenting choices in English and confirming the examiner's status [15,18,19]. For the classification of nystagmus, we tested CoT, MR, and their combination, but found no significant improvement in accuracy. GPT is a pretrained general-purpose LLM, and its accuracy is believed to depend on factors such as the number of model parameters, amount of training data, and scale of computational resources [32]. When provided with appropriate prompts, the model may use a greater portion of its parameters more efficiently, generating optimal outputs. Designing appropriate prompts for specific nystagmus patterns is necessary to further increase the model's discrimination accuracy. Regarding the input methods, the "CSV Only" input showed a notable tendency toward instances where the model frequently failed to classify any type of nystagmus. In contrast, inputs that included image data consistently resulted in producing a specific classification. This indicates that inputs containing images are more effective in enabling the model to provide responses and perform accurate classifications. Moreover, providing multimodal information, such as still images, patient information, and head position is expected to increase the accuracy of balance function tests.

For specific nystagmus patterns, classification was possible for horizontal movements; however, the accuracy of classifying vertical and torsional components was low, regardless of the image or coordinate input. This can be attributed to the evaluation of two dimensional movements, which makes torsional assessment challenging, similar to the limitations of deep learning using two dimensional video capture [27,28]. Additionally, the amplitude of vertical movements was smaller than that of horizontal movements, which could be another reason for lower accuracy. Improving the accuracy and pupil tracking methods with the development of prompts specialized for vertical domains may also be effective for higher precision. Responses showed higher accuracy for identifying horizontal nystagmus compared to other types. These results suggest that the accuracy can be further improved by increasing the input images frame speed, performing more detailed preprocessing, such as changing the settings for each direction, or taking coordinates in three-dimensional directions. This study shows

XSL•FO
RenderX

that GPT-4V, a LLM trained on extensive data, can achieve a certain level of accuracy through in-context learning methods such as CoT and MR highlights the broad applicability of this model.

Clinically, nystagmus findings are crucial indicators for evaluating vestibular function and are indispensable in clinical settings for neurological and otologic diseases [5]. However, variability in physician assessments and the occurrence of findings only during vertigo attacks necessitate a stable evaluation model. Electronystagmography, a common method for recording nystagmus clinically, records eye movements as corneoretinal potentials but cannot measure torsional eye movements and has the disadvantage of difficulty in capturing three-dimensional movements. Additionally, the need for specialized equipment makes real-time recording during vertigo attacks challenging, limiting its frequent use in clinical practice. In contrast, methods for recording eye movements using video are becoming widespread [33,34], and with recent advancements in deep learning technology, consistent assessments may become possible. The use of highly versatile LLMs can further expand their application. For example, LLM can be used in written exchanges to confirm repetitions, reasons for thinking so, and corresponding details. The GPT-4V model, with its real-time clinical setting applications, demonstrated faster inference time and lower computational complexity than conventional deep learning techniques. Since some hallucinations may yield incorrect answers as if they were correct, a human must make the decision considering the AI responses [35]. The model's accuracy, especially in this case, should be limited to cases where a human confirms the answer. Further improvement of the model's accuracy is required in the future.

As a limitation, this verification was specialized for classifying six types of nystagmus in videos and did not evaluate abnormal detections during regular examinations or nystagmus containing multiple components. Additionally, one limitation is that images being evaluated depend on the nystagmus in videos obtained during the examination, leading to variability in the amount of data depending on the type of nystagmus. Therefore, future studies should include improvements in the classification accuracy of nystagmus patterns and verification of the model's adaptability to mixed types of clinical data, such as horizontal- and vertical-torsional nystagmus. Moreover, the issue of hallucinations in LLMs is crucial, and how they are used is essential. At a minimum, educational purposes such as training medical professionals and pre-evaluation before doctors can judge the videos, could be effective. Devices capable of recording nystagmus using smartphones exist, and considerations must be made to record and assess patients without medical intervention [36].

The GPT-4V based nystagmus classification model represents significant advancements in medical imaging and diagnostic techniques. Its high accuracy, efficiency, and potential for real-time application make it a valuable tool for improving the diagnosis and management of nystagmus. Continuous research and development in this area are essential for improving the model and maximizing its clinical utility.

Conclusions

In this study, we developed a nystagmus classification model using GPT-4V and evaluated its performance. Unlike previous deep learning models, GPT-4V, centered on a LLM, presents a promising method for classifying nystagmus in video data and is expected to contribute to improved accuracy and efficiency in medical diagnoses. This represents a significant advance in medical AI and it is crucial to continue refining the model and consider its clinical applications to fully realize the potential benefits that AI technology brings to the medical field.

Acknowledgments

The authors would like to thank Editage for their assistance in the English language editing process.

This research received no external funding.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: MN, HF
Data curation: MN
Investigation, MN, HF
Methodology: MN, HF
Project administration: MN, HF
Resources: RT, HF
Supervision: HF
Validation: MN, HF
Visualization: MN, HF
Writing – original draft: MN
Writing – review & editing: RK, RT, HO, TK, MI, HF

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Sample of prompt.
[DOCX File, 191 KB - formative_v9i1e70070_app1.docx]

References

1. Ospina LH. Dealing with nystagmus. J Binocul Vis Ocul Motil 2018;68(4):99-109. [doi: 10.1080/2576117X.2018.1493311] [Medline: 30322349]
2. Kates MM, Beal CJ. Nystagmus. JAMA 2021 Feb 23;325(8):798. [doi: 10.1001/jama.2020.3911] [Medline: 33620408]
3. Gottlob I. Nystagmus. Curr Opin Ophthalmol 2001 Oct;12(5):378-383. [doi: 10.1097/00055735-200110000-00010] [Medline: 11588502]
4. Newman-Toker DE, Camargo CA, Hsieh YH, Pelletier AJ, Edlow JA. Disconnect between charted vestibular diagnoses and emergency department management decisions: a cross-sectional analysis from a nationally representative sample. Acad Emerg Med 2009 Oct;16(10):970-977. [doi: 10.1111/j.1553-2712.2009.00523.x] [Medline: 19799573]
5. Bhattacharyya N, Gubbels SP, Schwartz SR, et al. Clinical practice guideline: Benign paroxysmal positional vertigo (update). Otolaryngol--head neck surg 2017 Mar;156(S3):S1-S47. [doi: 10.1177/0194599816689667]
6. Lopez-Escamez JA, Gamiz MJ, Fernandez-Perez A, Gomez-Fiñana M. Long-term outcome and health-related quality of life in benign paroxysmal positional vertigo. Eur Arch Otorhinolaryngol 2005 Jun;262(6):507-511. [doi: 10.1007/s00405-004-0841-x] [Medline: 15942805]
7. Santini T, Fuhl W, Kasneci E. PuReST: robust pupil tracking for real-time pervasive eye tracking. : ACM Presented at: ETRA '18: 2018 Symposium on Eye Tracking Research and Applications; Jun 14-17, 2018; Warsaw Poland p. 1-5. [doi: 10.1145/3204493.3204578]
8. Eivazi S, Santini T, Keshavarzi A, Kübler TC, Mazzei A. Improving real-time CNN-based pupil detection through domain-specific data augmentation. : ACM; 2019 Jun 25 Presented at: ETRA '19; Jun 25-28, 2019; Denver Colorado p. 1-6. [doi: 10.1145/3314111.3319914]
9. Otero-Millan J, Roberts DC, Lasker A, Zee DS, Kheradmand A. Knowing what the brain is seeing in three dimensions: A novel, noninvasive, sensitive, accurate, and low-noise technique for measuring ocular torsion. J Vis 2015;15(14):11. [doi: 10.1167/15.14.11] [Medline: 26587699]
10. Yiu YH, Aboulatta M, Raiser T, et al. DeepVOG: Open-source pupil segmentation and gaze estimation in neuroscience using deep learning. J Neurosci Methods 2019 Aug 1;324:108307. [doi: 10.1016/j.jneumeth.2019.05.016] [Medline: 31176683]
11. Zhang W, Wu H, Liu Y, et al. Deep learning based torsional nystagmus detection for dizziness and vertigo diagnosis. Biomed Signal Process Control 2021 Jul;68:102616. [doi: 10.1016/j.bspc.2021.102616]
12. Kermany DS, Goldbaum M, Cai W, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell 2018 Feb 22;172(5):1122-1131. [doi: 10.1016/j.cell.2018.02.010] [Medline: 29474911]
13. Jahan I, Laskar MTR, Peng C, Huang JX. A comprehensive evaluation of large language models on benchmark biomedical text processing tasks. Comput Biol Med 2024 Mar;171:108189. [doi: 10.1016/j.compbiomed.2024.108189] [Medline: 38447502]
14. GPT-4V(ision) system card. OpenAI. 2023. URL: https://openai.com/research/gpt-4v-system-card [accessed 2023-09-25]
15. Noda M, Ueno T, Koshu R, et al. Performance of GPT-4V in answering the Japanese Otolaryngology Board Certification Examination questions: evaluation study. JMIR Med Educ 2024 Mar 28;10:e57054. [doi: 10.2196/57054] [Medline: 38546736]
16. Harsha N, Yin TL, Sheng Z, Dean C, Richard E, Nicolo F, et al. Can generalist foundation models outcompete special-purpose tuning? Case study in medicine. arXiv. Preprint posted online on Nov 28, 2023. [doi: 10.48550/arXiv.2311.16452]
17. Nipun MS, Sulaiman RB, Kareem A. Efficiency comparison of AI classification algorithms for image detection and recognition in real-time. arXiv. Preprint posted online on Jun 12, 2022. [doi: 10.48550/arXiv.2206.05842]
18. Tian Y, Kanade T, Cohn JF. Dual-state parametric eye tracking. Presented at: Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition; Mar 28-30, 2000; Grenoble, France URL: https://www.cs.cmu.edu/~face/Papers/fg1camera.pdf [accessed 2025-04-26]
19. Padilla R, Costa Filho CF, Costa M. Evaluation of Haar cascade classifiers designed for face detection. World Acad Sci Eng Technol 2012;6(4):466-469. [doi: 10.5281/zenodo.1058133]
20. Wilson PI, Fernandez J. Facial feature detection using Haar classifiers. J Comput Sci Coll 2006;21:127-133 [FREE Full text]
21. Tanaka Y, Nakata T, Aiga K, et al. Performance of generative pretrained transformer on the National Medical Licensing Examination in Japan. PLOS Digit Health 2024 Jan;3(1):e0000433. [doi: 10.1371/journal.pdig.0000433] [Medline: 38261580]

XSL•FO
RenderX

22. Noda M, Ueno T, Koshu R, et al. A study of the performance of the generative pretrained transformer in the Japanese Otorhinolaryngology Specialty examination. Nippon Jibiinkoka Tokeibugeka Gakkai Kaiho(Tokyo) 2023;126(11):1217-1223. [doi: 10.3950/jibiinkotokeibu.126.11_1217]
23. Bsharat SM, Myrzakhan A, Shen Z. Principled instructions are all you need for questioning LLaMA-1/2, GPT-3.5/4. arXiv. Preprint posted online on Jan 18, 2024. [doi: 10.48550/arXiv.2312.16171]
24. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, et al. Chain-of-thought prompting elicits reasoning in large language models. arXiv. Preprint posted online on Jan 10, 2023. [doi: 10.48550/arXiv.2201.11903]
25. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, et al. LLaMA: open and efficient foundation language models. arXiv. Preprint posted online on Feb 27, 2023. [doi: 10.48550/arXiv.2302.13971]
26. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. arXiv. Preprint posted online on Jul 22, 2020. [doi: 10.48550/arXiv.2005.14165]
27. Wagle N, Morkos J, Liu J, et al. aEYE: A deep learning system for video nystagmus detection. Front Neurol 2022;13:963968. [doi: 10.3389/fneur.2022.963968] [Medline: 36034311]
28. Lim EC, Park JH, Jeon HJ, et al. Developing a diagnostic decision support system for benign paroxysmal positional vertigo using a deep-learning model. J Clin Med 2019 May 8;8(5):633. [doi: 10.3390/jcm8050633] [Medline: 31072056]
29. Lee Y, Lee S, Han J, Seo YJ, Yang S. A nystagmus extraction system using artificial intelligence for video-nystagmography. Sci Rep 2023;13(1):11975. [doi: 10.1038/s41598-023-39104-7]
30. Rao A, Kim J, Kamineni M, et al. Evaluating GPT as an adjunct for radiologic decision making: GPT-4 versus GPT-3.5 in a breast imaging pilot. J Am Coll Radiol 2023 Oct;20(10):990-997. [doi: 10.1016/j.jacr.2023.05.003] [Medline: 37356806]
31. Chiarelli G, Stephens A, Finati M, et al. Adequacy of prostate cancer prevention and screening recommendations provided by an artificial intelligence-powered large language model. Int Urol Nephrol 2024 Aug;56(8):2589-2595. [doi: 10.1007/s11255-024-04009-5] [Medline: 38564079]
32. Jared K, Sam M, Tom H, Tom BB, Benjamin C, Rewon C, et al. Scaling laws for neural language models. arXiv. Preprint posted online on Jan 23, 2020 URL: https://arxiv.org/pdf/2001.08361 [accessed 2025-04-26]
33. Bozomitu RG, Păsărică A, Tărniceriu D, Rotariu C. Development of an eye tracking-based human-computer interface for real-time applications. Sensors (Basel) 2019 Aug 20;19(16):3630. [doi: 10.3390/s19163630] [Medline: 31434358]
34. Cristina S, Camilleri KP. Unobtrusive and pervasive video-based eye-gaze tracking. Image Vis Comput 2018 Jun;74:21-40. [doi: 10.1016/j.imavis.2018.04.002]
35. Williams CYK, Bains J, Tang T, et al. Evaluating large language models for drafting emergency department discharge summaries. medRxiv. 2024 Apr 4 p. 2024.04.03.24305088. [doi: 10.1101/2024.04.03.24305088] [Medline: 38633805]
36. Fournier-Tombs E, McHardy J. A medical ethics framework for conversational artificial intelligence. J Med Internet Res 2023 Jul 26;25:e43068. [doi: 10.2196/43068] [Medline: 37224277]

Abbreviations

AI: artificial intelligence
CoT: Chain of Thoughts
GPT-4V: Generative Pre-trained Transformer-4-Vision
LLM: large language model
MR: metarecognition

Edited by A Mavragani; submitted 14.12.24; peer-reviewed by PB Chandrashekar, S Raza; revised version received 02.03.25; accepted 07.04.25; published 06.06.25.

Please cite as:
Noda M, Koshu R, Tsunoda R, Ogihara H, Kamo T, Ito M, Fushiki H
Exploring Generative Pre-Trained Transformer-4-Vision for Nystagmus Classification: Development and Validation of a Pupil-Tracking Process
JMIR Form Res 2025;9:e70070
URL: https://formative.jmir.org/2025/1/e70070
doi:10.2196/70070

© Masao Noda, Ryota Koshu, Reiko Tsunoda, Hirofumi Ogihara, Tomohiko Kamo, Makoto Ito, Hiroaki Fushiki. Originally published in JMIR Formative Research (https://formative.jmir.org), 6.6.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is

XSL•FO
RenderX

properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Remote Photoplethysmography Technology for Blood Pressure and Hemoglobin Level Assessment in the Preoperative Assessment Setting: Algorithm Development Study

Selene Y L Tan[1], MBBS; Jia Xin Chai[1], MBBS, MMed; Minwoo Choi[2], BA; Umair Javaid[2], PhD; Brenda Pei Yi Tan[1], MSc; Belinda Si Ying Chow[1], BSc; Hairil Rizal Abdullah[1,3], MBBS, MMed, MSc, PhD

[1]Division of Anesthesiology and Perioperative Medicine, Singapore General Hospital, Outram Road, Singapore, Singapore

[2]Nervotec, Singapore, Singapore

[3]Duke-NUS Medical School, Singapore, Singapore

Corresponding Author:
Hairil Rizal Abdullah, MBBS, MMed, MSc, PhD
Division of Anesthesiology and Perioperative Medicine, Singapore General Hospital, Outram Road, Singapore, Singapore

Abstract

Background: Blood pressure (BP) and hemoglobin concentration measurements are essential components of preoperative anesthetic evaluation. Remote photoplethysmography (rPPG) is an emerging technology that may be used to measure BP and hemoglobin concentration noninvasively with just a consumer-grade smartphone, replacing traditional in-person measurements. However, there is limited data regarding the use of this technology in patients with diverse skin tones and medical comorbidities. Hence, widespread applicability is yet to be achieved. The potential benefits of achieving this would be immense, allowing for greater convenience, accessibility, and reduction in labor and resources.

Objective: Our study aims to be the first to develop an algorithm for noninvasive rPPG-based BP and hemoglobin concentration measurement that can be used for preoperative evaluation of patients in real-world clinical practice settings.

Methods: We conducted the study at Singapore General Hospital from March 1, 2023, to June 28, 2024. A total of 200 patients were recruited. Our primary analysis compared the accuracy of rPPG-based systolic and diastolic BP measurements against measurements taken with automated BP measuring devices. Our secondary analysis compared the accuracy of rPPG-based hemoglobin concentration measurement against traditional blood sampling.

Results: Our model performed best with diastolic BP predictions, with a mean absolute percentage error of 7.52% and a mean difference of 0.16 mm Hg (SD 3.22 mm Hg) between reference and measured readings. The 95% CI for the mean difference between predicted and measured diastolic BP was ±0.57 (−0.41 to 0.73) mm Hg. Systolic BP predictions yielded a mean absolute percentage error of 9.52% and a mean difference of 2.69 mm Hg (SD 7.86 mm Hg). The 95% CI for the mean difference between predicted and measured systolic BP was ±1.14 (−1.54 to −3.83) mm Hg. Hemoglobin concentration predictions had a mean absolute percentage error of 8.52%, with a mean difference of 0.23 g/dL (SD 0.67 g/dL). The 95% CI for the mean difference between predicted and reference measured hemoglobin concentration was ±0.10 (95% CI 0.13 - 0.33) g/dL.

Conclusions: Noninvasive rPPG-based measurement of BP and hemoglobin concentration at the preoperative evaluation setting has great potential for improving convenience, improving efficiency, and conserving resources for patients and health care providers. Our model was able to accurately predict diastolic BP in patients with diverse skin tones and medical comorbidities. The findings of this study serve as a basis for further studies to develop and validate the model for noninvasive rPPG-based BP and hemoglobin concentration measurement.

Trial Registration: ClinicalTrials.gov NCT06320847; https://clinicaltrials.gov/study/NCT06320847

(JMIR Form Res 2025;9:e60455) doi:10.2196/60455

KEYWORDS

blood pressure; hemoglobin; noninvasive; telehealth; remote photoplethysmography; smartphone; hypertension; anemia; mHealth; mobile health; development; validation; hypertensive; systolic; diastolic; physiological

Introduction

With the use of just an ubiquitous smartphone camera, remote photoplethysmography (rPPG) technology can measure vital signs remotely, eliminating the need for traditional in-person physical assessments. In telehealth and patient care applications, noninvasive vital sign measurements can transform billions of

XSL•FO
RenderX

devices into cost-effective and portable health care measuring devices [1].

The major obstacles to obtaining an accurate signal for rPPG-based vital sign measurements are changes in ambient lighting conditions and movement artifacts. However, these artifacts always present and are not completely avoidable in realistic conditions [2].

Our study aims to develop a model for noninvasive, rPPG-based BP and hemoglobin concentration measurement in patients with diverse skin tones and medical comorbidities in real world conditions, at the preoperative evaluation clinic setting. Preoperative assessment is an important aspect of care for all patients undergoing anesthesia and surgery. The demand for surgical services is rising due to population growth and aging, technological advances and changes in treatment perspectives [3,4]. This in turn leads to greater demand for preoperative anesthetic assessments at the preoperative evaluation clinic. Telemedicine is an imminent adjunct that could improve patient care and the effectiveness of treatment in a rapidly aging population with an increasing number of chronic health conditions [5,6].

An essential component of preoperative assessment is the measurement of vital signs such as blood pressure (BP) and hemoglobin concentration, both of which are traditionally done in-person at the preoperative evaluation clinic setting. Preoperative hypertension is associated with increased perioperative risk [7]. Anemia is one of the strongest risk factors for perioperative blood transfusions, which could bring further complications, such as increased length of stay [8].

Automated BP measuring devices are the most commonly used devices for BP measurement [9], while hemoglobin concentration is measured with an invasive blood test. The ideal alternative is a noninvasive, contactless physiological monitor, which would provide greater comfort, convenience, and efficiency.

rPPG works on the principle that reflected light from the various regions of the skin is affected by the volume of blood under the skin [10]. This allows for an optical measurement technique that quantifies peripheral blood volume and flow variations [11]. These measurements are then analyzed with signal processing algorithms which then generate physiological measurements, such as BP [12]. Furthermore, rPPG technology relies only on a consumer-grade smartphone camera and does not require skin contact, which makes it even more ideal for use in telemonitoring and mobile health (mHealth) apps [13,14]. Developing rPPG technology for use in the preoperative evaluation clinic setting can improve manpower, save costs and lead to higher patient satisfaction scores without increasing day-of-procedure case cancellations [15,16].

An additional advantage is that rPPG technology may allow for telemonitoring of ambulatory BP, which is currently considered the gold standard for accurate diagnosis of hypertension [17]. However, the majority of patients are assessed via clinic BP measurements [18]. Home-based telemonitoring may lead to better control of ambulatory BP, especially among those with inadequate BP control [19-21].

Despite the potential advantages of contactless rPPG vital sign monitoring, there are currently some important limitations that our study aims to address. Most early studies on contactless rPPG-based BP measurement were conducted in a nonclinical setting on healthy, normotensive participants with similar skin tones [12,22,23]. These limitations must be addressed for rPPG technology to achieve widespread, commercially viable usage in the clinical setting. Our study aims to develop an rPPG model for remote BP monitoring and hemoglobin concentration estimation across a representative range of BP readings and skin tones at the preoperative evaluation clinic setting.

Methods

Study Design

The study was conducted at Singapore General Hospital between March 1, 2023, and June 28, 2024. Our study used a prospective feasibility design to compare contactless rPPG BP and hemoglobin concentration measurements with reference measurements from automated cuffed BP measuring devices and clinical laboratory testing. A total of 200 patients participated.

Data Collection

The study was conducted in the preoperative evaluation clinic with Nervotec's contactless rPPG vital signs measurement technology and equipment for standard reference measurements (automated cuffed BP measuring device and pulse oximeter), as illustrated in Figure 1. Patients were seated in a private room and positioned comfortably with their arms placed at a 90-degree angle on the table. Two individual measurements were obtained sequentially using both the Welch Allyn Connex machine and Nervotec's rPPG technology (NervoScan). Nervotec, our technical partner, is a Singaporean digital health and artificial intelligence (AI) company that has successfully used rPPG technology in telehealth and mHealth apps to allow for the collection of rPPG waveforms via smartphones. Their base technology measures heart rate (HR), heart rate variability (HRV), respiration rate (RR) and blood oxygen saturation (SpO2) levels through signal processing techniques of a facial scan via the smartphone camera.

XSL•FO
RenderX

Figure 1. Equipment setup and patient positioning for blood pressure and hemoglobin concentration measurement with Nervotec remote photoplethysmography technology in the preoperative evaluation clinic.

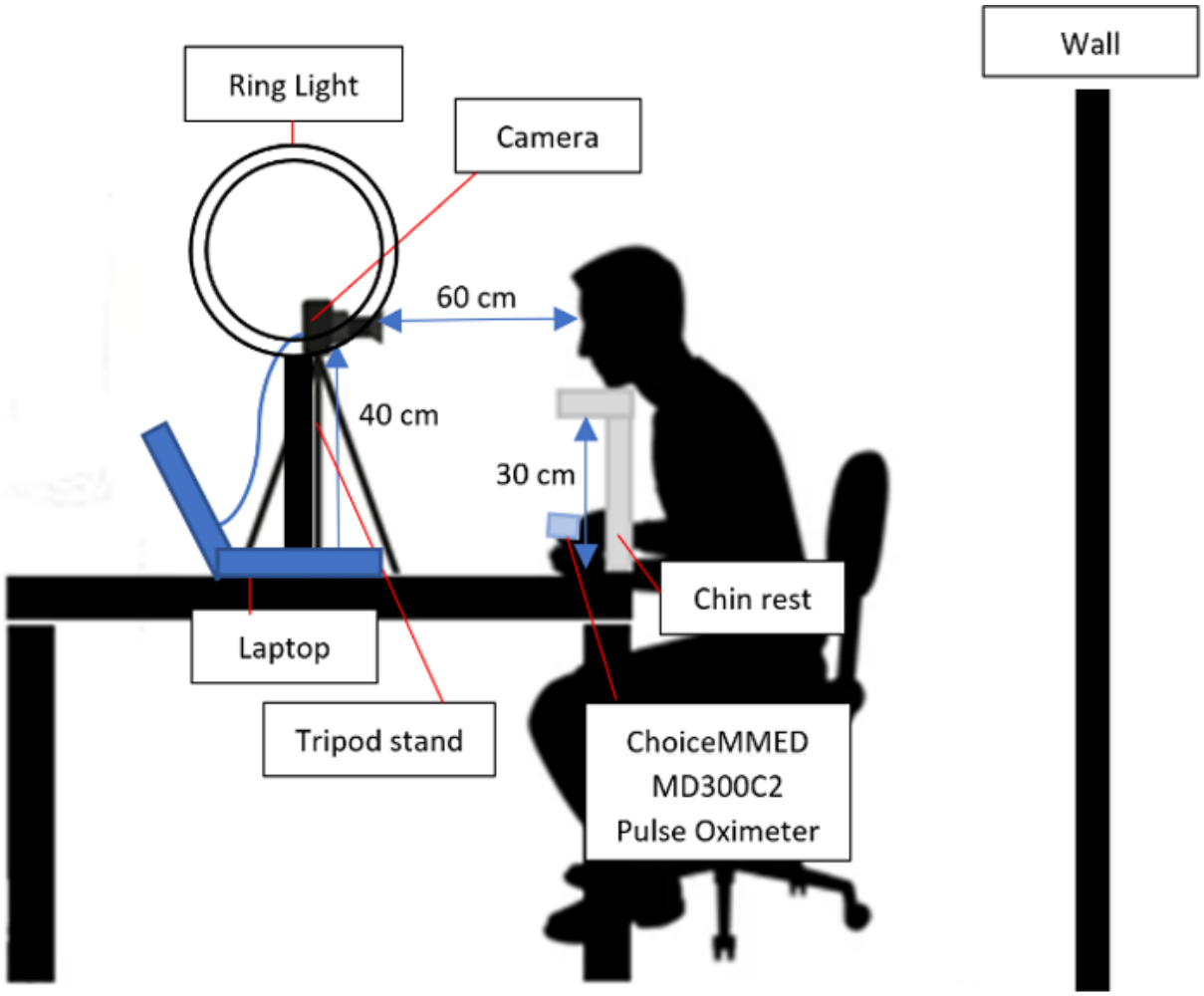

The NervoScan generates estimates every 10 seconds, therefore the goal was to capture the respective clinical reference for BP per 10 seconds. This was not feasible due to reasons such as limitations in the continuous noninvasive BP monitoring and waveform extraction as well as time synchronicity of the two measuring devices (NervoScan and the clinical device). To address this, the study team used a pulse oximeter in conjunction with the Welch Allyn Connex Vital Signs Monitor 6000 Series, which serves as the standard equipment within the preoperative clinic setting. The vital signs, including HR, RR, SpO2, HRV, and BP estimates, were simultaneously captured using Nervotec's rPPG technology, recorded with a laptop and webcam setup. The parameters were collected via a synchronized measurement process lasting 3 minutes. Hemoglobin concentration was retrospectively derived from the rPPG scans.

Information on age, height, weight, gender, ethnicity, Monk Skin Tone (MST) Scale, relevant medical history and routine preoperative blood investigation results (full blood count, renal panel, anemia panel, coagulation, thyroid function) were collected from the patient's electronic medical records. The MST Scale is a validated alternative to previous skin tone scales (eg, Fitzpatrick Scale), which has been adopted for multiple uses, including AI and machine learning [24]. Data collected were nonidentifiable, and only numerical data were collected. No photographic face recognition data were stored. The regions of interest captured by the rPPG facial scan were the forehead and bilateral cheeks, as indicated in Figure 2 below.

XSL•FO
RenderX

Figure 2. Regions of interest for the remote photoplethysmography facial scan.

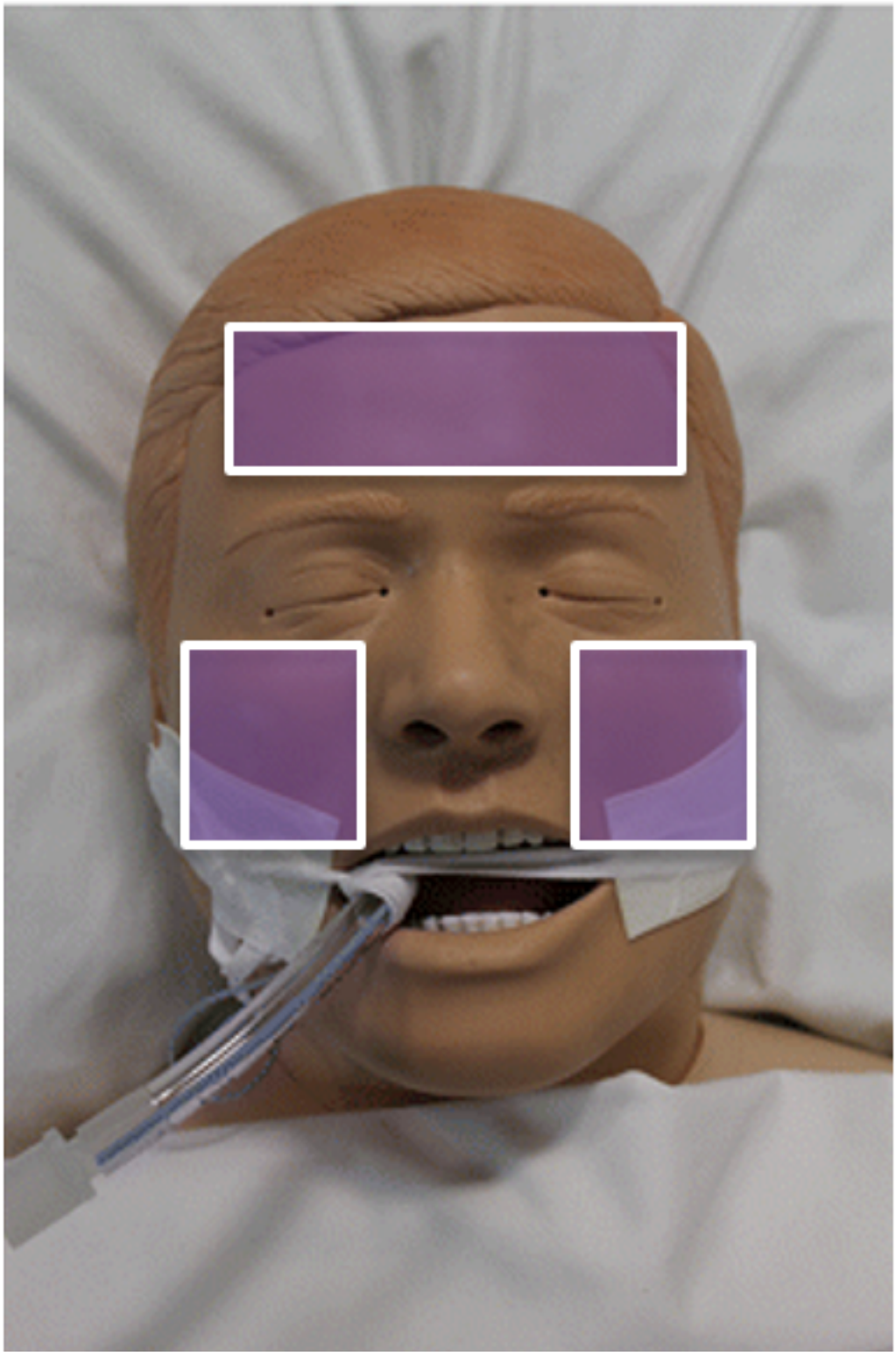

Dataset

A total of 200 patients with a mean age of 56.44 years (range: 21 - 84 y) participated. Of these patients, 67% had comorbid conditions such as hypertension, diabetes mellitus, and ischemic heart disease. The demographic profile of the participants is given in Table 1.

XSL•FO
RenderX

Table . Demographic profile of patients (N=200).

Characteristic	Patients
Sex, n (%)	
Male	76 (38)
Female	124 (62)
Age (years), mean (range)	58.23 (21-84)
SBP[a] (mm Hg) determined from baseline reference measurement, mean (range)	127.57 (84.5 - 191)
DBP[b] (mm Hg) determined from baseline reference measurement, mean (range)	75.17 (54.5 - 98)
Race, n (%)	
Chinese	160 (80)
Filipino	1 (0.5)
Malay	16 (8)
Indian	15 (7.5)
Eurasian	2 (1)
Others	6 (3)
Skin tone, n (%)	
Monk skin tone 5 or 6	118 (59)
Monk skin tone ≤4	82 (0.41)
Underlying medical conditions, n (%)	
None	66 (33)
Ischemic heart disease	6 (3)
Diabetes mellitus	48 (24)
Hypertension	80 (40)

[a]SBP: systolic blood pressure.

[b]DBP: diastolic blood pressure.

Definitions

We defined hypertension according to the European Society of Hypertension (ESH) guidelines [25]. The ESH classifies office BP and defines hypertension grades as shown in Table 2. For our study, we combined Grade 2 and 3 hypertension into a single category. We also defined hypotension as systolic BP (SBP) ≤100 mm Hg and/or diastolic BP (DBP) ≤60 mm Hg [26].

The absolute numbers and percentages of patients in each BP category are presented in Table 3.

A more detailed breakdown of the percentages of participants in each BP category was also compared against the International Organization for Standardization (ISO) 81060-2:2018 requirements (Table 4) [27].

Table . Definition of hypertension grades.

BP[a] classification	SBP[b] (mm Hg) and DBP[c] (mm Hg) criteria
Optimal	SBP of <120 and DBP of <80
Normal	SBP of 120 - 139 and/or DBP of 80 - 89
Grade 1 hypertension	SBP of 140 - 159 and/or DBP of 90 - 99
Grade 2 hypertension	SBP of 160 - 179 and/or DBP of 100 - 109
Grade 3 hypertension	SBP of ≥180 or DBP of ≥110

[a]BP: blood pressure.

[b]SBP: systolic blood pressure.

[c]DBP: diastolic blood pressure.

XSL•FO
RenderX

Table . Numbers and percentages of patients sampled from each blood pressure category.

BP	Hypotensive, n (%)	Grade 1 hypertension, n (%)	Grade 2 and 3 hypertension, n (%)
Systolic	16 (8)	36 (18)	16 (8)
Diastolic	8 (4)	27 (13.5)	0 (0)

Table . Percentage of patients in each blood pressure category compared with International Organization for Standardization (ISO) requirements.

BP[a] range	Patients (%)	ISO requirements (%)
SBP[b] ≤100 mm Hg (hypotension)	8	≥5
SBP ≥140 mm Hg (Grade 1 hypertension)	26	≥20
SBP ≥160 mm Hg (Grade 2 and 3 hypertension)	8	≥5
DBP[c] ≤60 mm Hg (hypotension)	4	≥5
DBP ≥85 mm Hg (Grade 1 hypertension)	13.5	≥20
DBP ≥100 mm Hg (Grade 2 and 3 hypertension)	0.0	≥5

[a]BP: blood pressure.

[b]SBP: systolic blood pressure.

[c]DBP: diastolic blood pressure.

Data Preprocessing

In this study, facial video recordings of 200 patients were obtained, each comprising two 3 minute sessions with concurrent ground truth BP and hemoglobin concentration measurements. For each 3-minute (180s-) video, red, green, and blue (RGB) variations of the facial regions were extracted from several predefined regions of interest (ie, forehead, left cheek, right cheek, nose and cheeks, and full face). The RGB data were segmented into 18 windows of 10 seconds each, with each window representing an independent data point. The NervoScan software was used to filter out data points where the patient's face was not centered in the camera frame, thereby enhancing data quality and consistency. A total of 4109 data points were collected for effective training of the proposed model.

We excluded data points with extreme BP values (SBP≤80 mm Hg or ≥180 mm Hg and DBP≤60 mm Hg or ≥130 mm Hg). This addressed the class imbalance issue by removing outliers, thereby enhancing the dataset's suitability for analysis. After filtering, 3941 data points from 197 patients remained. The shuffled dataset was then divided into training, validation, and testing sets comprising 2758, 592 and 591 data points respectively.

Machine Learning Model and Training

Our approach used a U-Net-based model enhanced with label distribution smoothing (LDS) techniques [28]. The model optimization leveraged the Adam optimizer, chosen for its efficient handling of large datasets and adaptability in gradient-based optimization. A weighted mean squared error loss function was applied to minimize prediction error, with additional adjustments provided by a triangular LDS kernel and an inverse LDS reweighting mechanism. This configuration refined prediction accuracy, particularly in instances with label distribution imbalances.

For training, we used a batch size of 64 and conducted 300 epochs, which allowed the model to capture intricate patterns and reduce overfitting. A learning rate of 0.001 was selected to ensure a balanced convergence rate without overshooting the optimal parameters. Separate models were trained for the prediction of SBP and DBP to better capture the specific patterns associated with each measurement.

Ethical Considerations

The study was approved by the SingHealth Centralised Institutional Review Board (CIRB Ref: 2023/2042) for the period of February 15, 2023, to December 6, 2024, and was registered on ClinicalTrials.gov (trial number: NCT06320847). All patients were recruited in the preoperative evaluation clinic by research assistants. All patients who were present in the preoperative evaluation clinic at Singapore General Hospital during the study period were screened for eligibility. The inclusion criteria were (1) all patients aged above 21 years, (2) presenting for any surgery except head and neck surgery, and (3) the ability to give informed consent. The exclusion criteria were (1) patients less than 21 years old, (2) inability to give informed consent, or (3) patients undergoing head and neck surgery. Participation was voluntary, without compensation, and written informed consent was obtained from all participants. Data collected were nonidentifiable, and only numerical data were collected. No photographic face recognition data were stored.

Results

Primary Analysis – SBP and DBP

The primary objective was to assess the accuracy of BP measurements from rPPG compared to automated cuffed BP measuring devices. All results are reported for a test set consisting of data points from 90 participants. The results of the primary analysis are presented in Table 5.

XSL•FO
RenderX

Table . Results of primary analysis (remote photoplethysmography–based systolic and diastolic blood pressure estimation).

Metric	SBP[a]	DBP[b]
Mean absolute percentage error (%)	9.52	7.52
Reference BP[c] (mm Hg), mean (SD)	128.77 (19.54)	75.54 (8.39)
Predicted BP (mm Hg), mean (SD)	131.46 (11.68)	75.38 (5.16)
Difference between predicted and measured BP (mm Hg), mean difference (SD)	2.69 (7.86)	0.16 (3.22)
95% CI for mean difference between predicted and measured BP (mm Hg)	±1.14 (–1.54 to −3.83)	±0.57 (−0.41 to 0.73)

[a]SBP: systolic blood pressure.

[b]DBP: diastolic blood pressure.

[c]BP: blood pressure.

The results revealed a general trend of underestimation of BPs at the lower end of the range and overestimation at the higher end for both SBP and DBP. To assess the agreement and correlation between the predicted and measured values, we used a Bland-Altman plot.

These results indicate the stronger performance of the model in estimating BP using rPPG for DBP.

Validation Against Standards

The mean absolute differences between the predicted and measured BP measurements were 2.69 mm Hg for SBP and 0.16 mm Hg for DBP as shown in Figures 3 and 4, respectively.

We compared the results from the NervoScan model against the ISO 81060-2:2018 standards [27]:

Using these guidelines, the performance of NervoScan BP estimates is given in Tables 6 and 7. It can be seen that both SBP and DBP estimates satisfy the validation criterion 1 (SD≤8.0 mm Hg). Also, the DBP estimates fulfill the criterion 2 (SD<6.95 mm Hg) whereas the SBP estimates do not meet the validation criterion 2 (SD<6.30 mm Hg).

XSL•FO
RenderX

Figure 3. Bland-Altman plot of the differences between the reference and remote photoplethysmography–based SBP measurements. pred: predicted; SBP: systolic blood pressure.

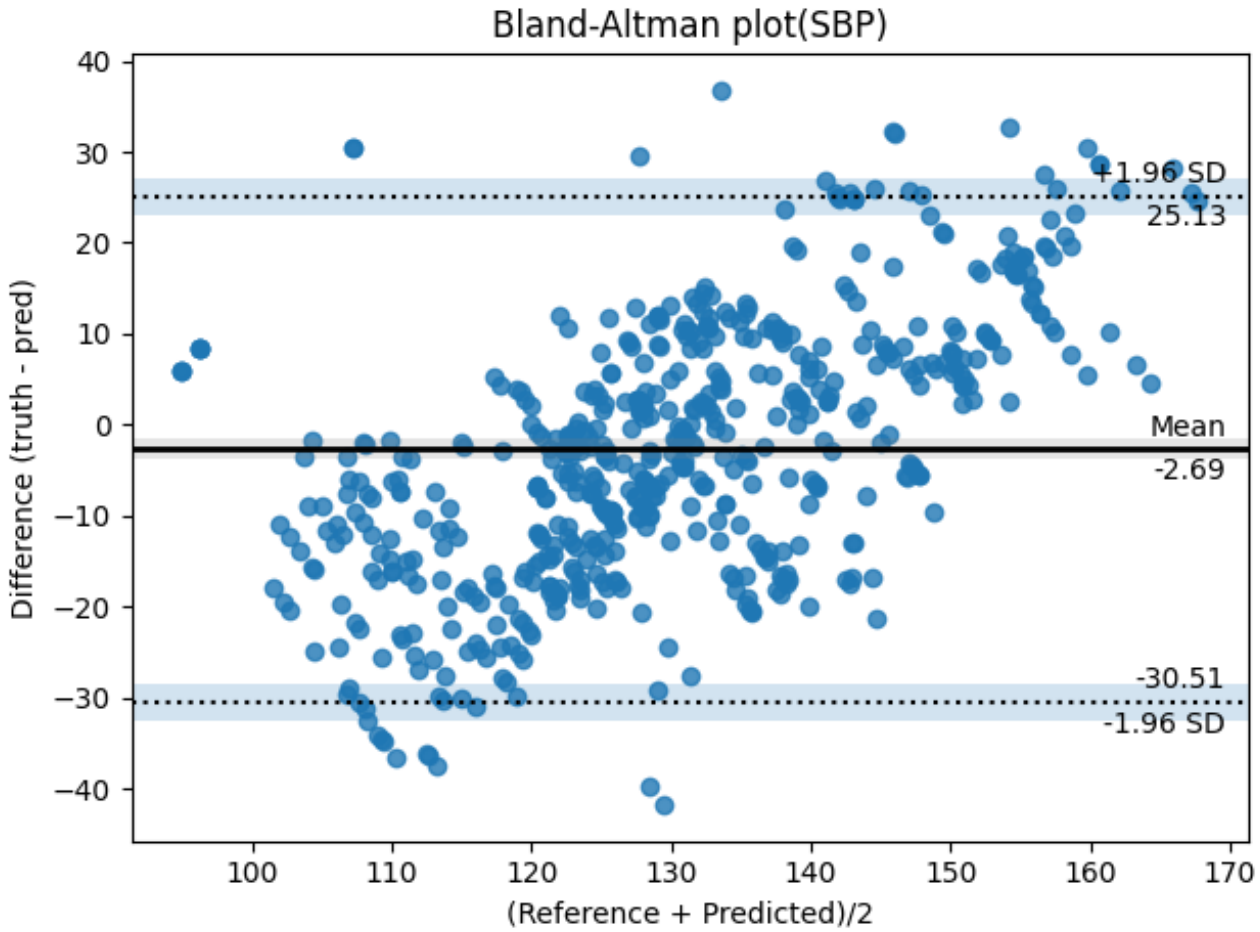

XSL•FO
RenderX

Figure 4. Bland-Altman plot of the differences between reference and remote photoplethysmography–based DBP measurements. pred: predicted; DBP: diastolic blood pressure.

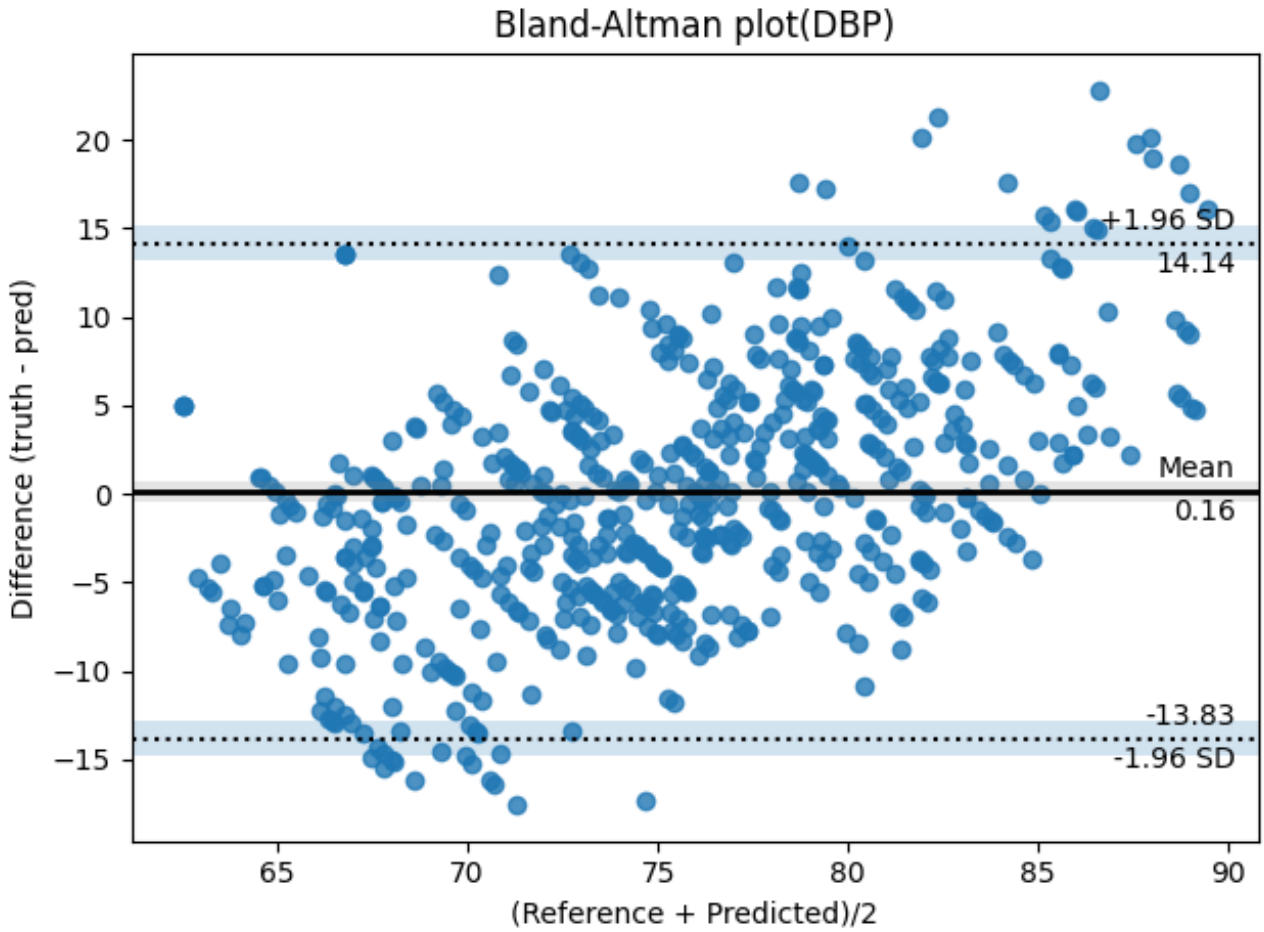

Table . Validation results of blood pressure measurements in accordance with criterion 1 and criterion 2 of the ANSI/AAMI/ISO[a] 81060−2:2018 guidelines (validation criteria of SD≤8.0 mm Hg for criterion 1; validation criteria of SD<6.30 mm Hg for systolic blood pressure and SD<6.95 mm Hg for diastolic blood pressure for criterion 2).

Criteria	SBP (mm Hg), estimate (SD)	Meets standard	DBP (mm Hg), estimate (SD)	Meets standard
Criterion 1	2.68 (7.86)	Yes	0.15 (3.22)	Yes
Criterion 2	2.92 (8.02)	No	0.27 (3.43)	Yes

[a]ANSI/AAMI/ISO: American National Standards Institute/Association for the Advancement of Medical Instrumentation/International Organization for Standardization.

Table . Validation results of blood pressure measurements in accordance with criterion 1 and criterion 2 of the guidelines.

Criteria	ISO[a] 81060-2:2018 standard (mm Hg)	rPPG[b] model (mm Hg)	Fulfills standards
Criterion 1 (SBP[c])	<8	7.86	Yes
Criterion 1 (DBP[d])	<8	3.22	Yes
Criterion 2 (SBP)	<6.3	8.02	No
Criterion 2 (DBP)	<6.95	3.43	Yes

[a]ISO: International Organization for Standardization.

[b]rPPG: remote photoplethysmography.

[c]SBP: systolic blood pressure.

[d]DBP: diastolic blood pressure.

XSL•FO
RenderX

Statistical Analysis of SBP and DBP Predictions

To determine the appropriate statistical approach for evaluating SBP and DBP predictions by the NervoScan and respective measured readings, we performed the Shapiro-Wilk Test to check for normality. The null hypothesis states that the data follows a normal distribution. A *P* value ≤.05 indicates a statistically significant deviation from normality, leading to the rejection of the null hypothesis.

The results of the Shapiro-Wilk test are summarized in Table 8.

These findings justify the use of nonparametric statistical tests for analyzing the paired data (eg, Wilcoxon signed rank test) in subsequent comparisons.

To evaluate the agreement between predictions and ground truths for SBP and DBP, the Wilcoxon signed rank test was used, as the data consists of paired measurements. The results of the Wilcoxon signed rank test are summarized in Table 9.

In conclusion, the results highlight that while the NervoScan performs well in predicting DBP, the SBP predictions require further optimization.

Table . Shapiro-Wilk test results for systolic blood pressure (SBP) and diastolic blood pressure (DBP) measurements.

Measurement	*P* value	Interpretation
SBP	<.001	SBP does not follow a normal distribution
DBP	<.001	DBP does not follow a normal distribution

Table . Wilcoxon signed rank test results for systolic blood pressure (SBP) and diastolic blood pressure (DBP) measurements.

Measurement	*P* value	Interpretation
SBP	<.001	Significant difference between reference and rPPG[a]-based SBP measurement
DBP	.90	No significant difference between reference and rPPG-based DBP measurement

[a]rPPG: remote photoplethysmography.

Secondary Analysis – Hemoglobin Concentration

Using the aforementioned dataset and the machine learning model architecture, a new model was trained to estimate hemoglobin. Our model predicted hemoglobin concentration with a mean absolute percentage error of 8.52%, exhibiting an error bias of 0.23 g/dL and an SD of 0.67 g/dL (Figure 5). Among 197 participants, the mean predicted hemoglobin was 12.71 g/dL (SD 1.13 g/dL), compared with the reference hemoglobin of 12.95 g/dL (SD 1.81 g/dL). The mean difference was 0.23 g/dL (SD 0.67 g/dL) as given in Table 10.

XSL•FO
RenderX

Figure 5. Bland-Altman plot of the difference between reference and remote photoplethysmography–based Hb concentration measurements. Hb: hemoglobin; pred: predicted.

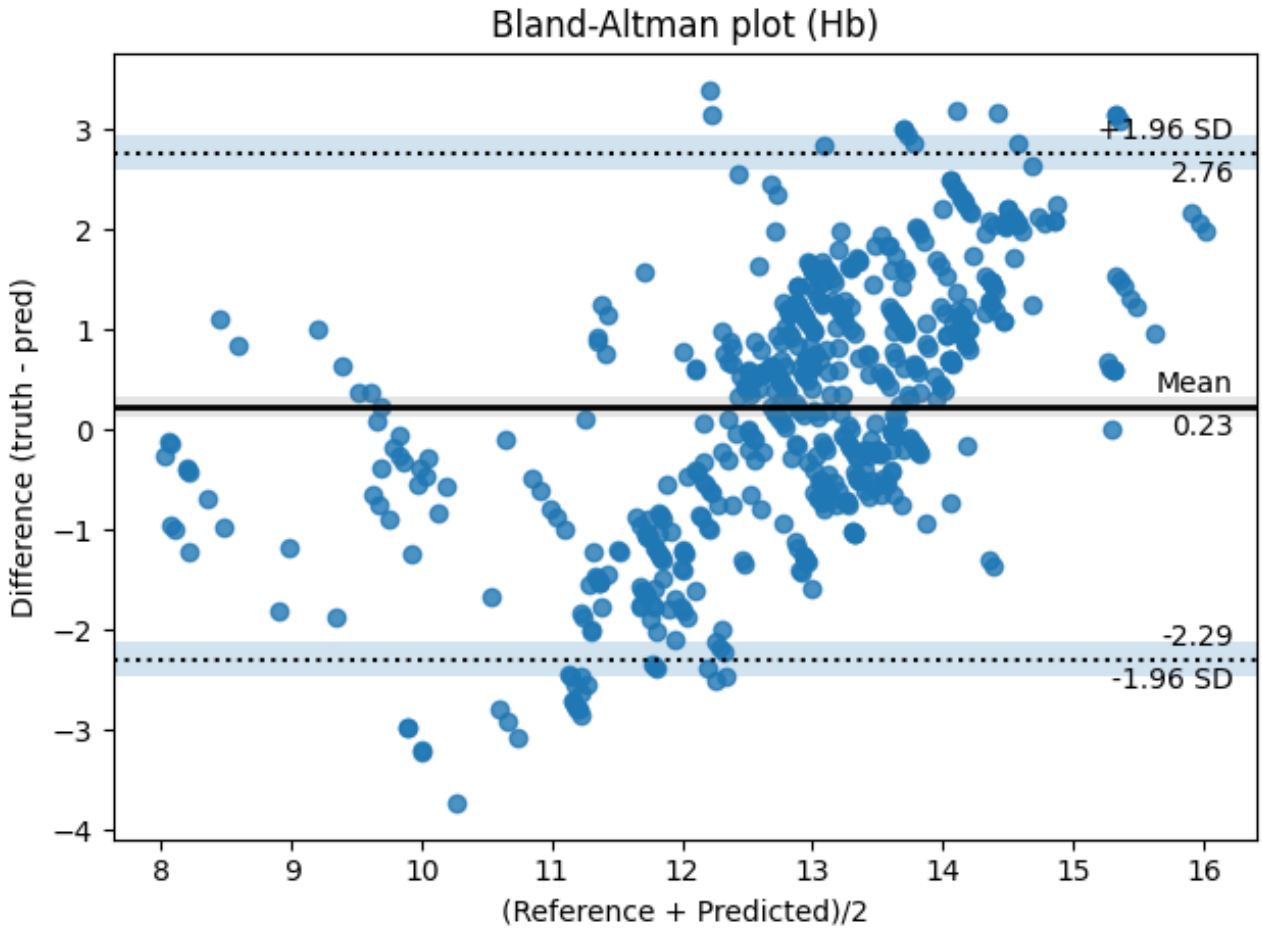

Table . Results of the secondary analysis (remote photoplethysmography–based hemoglobin concentration estimation).

Criteria	Hemoglobin (g/dL)
Mean absolute percentage error (%)	8.52
Reference hemoglobin (g/dL), mean (SD)	12.95 (1.81)
Mean predicted blood pressure (g/dL), mean (SD)	12.71 (1.13)
Difference between predicted and reference measured hemoglobin (g/dL), mean difference (SD)	0.23 (0.67)
95% CI for mean difference between predicted and reference measured hemoglobin (g/dL)	±0.10 (0.13 - 0.33)

The Shapiro-Wilk test was again used to test for normality. We obtained a *P* value of <.01, which indicates that the data are not normally distributed. Furthermore, the Wilcoxon signed rank test revealed significant difference between the two hemoglobin data distributions (predictions and measured reference) with a *P* value of <.01. It is due to the inherent class-imbalance in the data distribution (majority samples are between the 12-15 g/dL range whereas fewer samples exist outside the mentioned range. Hence, the model learning/estimation is biased from such over-representation of data points from the 12 - 15 g/dL range. This indicates that hemoglobin estimation can be improved by collecting sufficient hemoglobin values for both low and high values for a more effective learning.

As an initial proof of concept, this demonstrates the feasibility of hemoglobin prediction from rPPG. Thus, offering an alternative to hemoglobin estimation compared to the measurements requiring blood samples. While it offers reasonable accuracy in hemoglobin predictions compared with clinical laboratory results, the estimation can be further improved by including more data points from the low and high hemoglobin ranges. The results are demonstrated in the Bland Altman plot shown in Figure 5.

XSL•FO
RenderX

Discussion

Principal Findings

Our study successfully developed an rPPG model which performed strongly in predicting DBP in the preoperative evaluation clinic setting. Our model performed best with DBP predictions, with a mean absolute percentage error of 7.52% and a mean difference of 0.16 mm Hg (SD 3.22 mm Hg) between reference and measured readings, outperforming previous studies [29]. The focus of our study was to develop a contactless rPPG vital sign monitoring technology that could be used in diverse patient populations to measure BP and hemoglobin concentration in the preoperative evaluation clinic setting. Our study substantiates the potential efficacy of the rPPG-based model for accurately predicting BP and hemoglobin measurements. Further optimization is required for SBP and hemoglobin concentration predictions, as well as predictions in extreme physiological ranges. However, this development study serves as a promising basis to guide and optimize future studies, which will further develop and validate the model.

There were some limitations that we considered in our study. One limitation faced was the lack of a reference protocol for contactless vital signs monitoring. In view of this, we opted to use the American National Standards Institute (ANSI)/Association for the Advancement of Medical Instrumentation (AAMI)/ISO 81060-2:2018 standard as a guide, as automated cuffed BP measuring devices are the standard measurement device in the clinical setting. For future studies, a potential improvement would be to perform 3 reference readings per patient, instead of 2 as we have done in our study. Including more patients with extremes of BP and hemoglobin would improve the size and diversity of the dataset, which could improve model robustness. Another limitation was that the patients were not screened for arrhythmias, which would potentially affect the accuracy of contactless rPPG vital signs measurements [30]. This would also be a point of consideration for future follow-up studies.

The reported findings are promising, given that our study is the first rPPG BP and hemoglobin measurement development study based on patients with diverse skin tones, nonnormotensive BP readings and comorbid medical conditions such as diabetes mellitus and ischemic heart disease. Despite the challenges of developing a model in this heterogeneous dataset, our model was able to predict DBP accurately and achieve good correlation for SBP and hemoglobin. More work needs to be done in order to achieve the same degree of agreement for SBP and hemoglobin predictions. In doing so, we recognize the potential utility for this technology in the hospital setting, due to its wide applicability. The potential applications for this technology are vast and not just limited noninvasive BP and hemoglobin measurement in the preoperative evaluation clinic setting.

By stratifying patient readings into predefined BP and Hb categories, a successful model can significantly optimize the triaging process, thereby reducing the workload on health care professionals and increasing the efficiency of patient management. In-hospital applications would include continuous contactless monitoring in the emergency department, post-anesthesia care unit or intensive care unit This would possibly decrease human errors and improve patient safety without increasing the strain on human resources [31].

Furthermore, this model can be integrated into a mobile app (mHealth) that supports the trend toward more individualized and accessible health care solutions. Such an app could enhance ambulatory chronic disease management by enabling continuous monitoring and early detection, potentially reducing emergency visits and associated health care costs. Primary and ambulatory care patients may self-monitor their vital signs continuously and conveniently, which could be used to diagnose hypertension [32] without physically traveling to a health facility. This may allow timely intervention while reducing unwarranted hospital visits and transport costs [33]. Apart from hypertension, potential expansions may include the management of other common conditions such as diabetes [32]. For example, rPPG technology could be developed upon to measure glycated hemoglobin to diagnose diabetes mellitus and monitor glucose control in patients known to have diabetes [34].

In summary, the results of our study support further data evaluation and development of clinical applications for our rPPG model.

Conclusions

Our study was the first development study of a model for rPPG-based, contactless, noninvasive BP and hemoglobin measurement in a population with diverse skin tones and medical comorbidities. Predictions were stronger for DBP than SBP. Predictions for SBP, DBP, and hemoglobin were most accurate in the midquartile ranges. Our study serves as a promising basis for future development and validation studies. Further studies with a larger dataset including more measurements in the extreme ends of physiological range are required. In addition, rPPG-based contactless monitoring has vast applications beyond BP and hemoglobin measurement in the preoperative evaluation clinic setting, which could potentially include other physiological measurements, such as glycated hemoglobin measurements. These broader applications are potentially limitless and could include home ambulatory monitoring, noninvasive monitoring in pediatric populations, monitoring in low-resource settings, and reducing the labor costs of monitoring patients in critical care areas in the hospital.

Acknowledgments

Generative artificial intelligence was not used in any portion of the manuscript writing.

XSL•FO
RenderX

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: HRA, JXC
Data curation: BPYT, BSYC
Formal analysis: MC, UJ
Funding acquisition: HRA
Investigation: SYLT, HRA, MC
Methodology: HRA, JXC
Resources: HRA, MC, UJ
Software: MC, UJ
Supervision: HRA
Validation: MC, UJ
Writing – original draft: SYLT
Writing – review & editing: SYLT, HRA, MC, UJ

Conflicts of Interest

None declared.

References

1. McDuff D. Camera measurement of physiological vital signs. ACM Comput Surv 2023 Sep 30;55(9):1-40. [doi: 10.1145/3558518]
2. Qiao D, Ayesha AH, Zulkernine F, et al. ReViSe: remote vital signs measurement using smartphone camera. IEEE Access 2022;10:131656-131670. [doi: 10.1109/ACCESS.2022.3229977]
3. Etzioni DA, Liu JH, Maggard MA, Ko CY. The aging population and its impact on the surgery workforce. Ann Surg 2003 Aug;238(2):170-177. [doi: 10.1097/01.SLA.0000081085.98792.3d] [Medline: 12894008]
4. Lee YZ, Dharmawan A, Zhang X, Chua DYC, Low JK. The changing landscape of general surgery in the elderly - trends over a decade in a tertiary centre in Singapore. ANZ J Surg 2022 Sep;92(9):2018-2024. [doi: 10.1111/ans.17674] [Medline: 35373437]
5. Liljeroos M, Arkkukangas M. Implementation of telemonitoring in health care: facilitators and barriers for using eHealth for older adults with chronic conditions. Risk Manag Healthc Policy 2023 Jan 10;16:43-53. [doi: 10.2147/RMHP.S396495] [Medline: 36647422]
6. Malasinghe LP, Ramzan N, Dahal K. Remote patient monitoring: a comprehensive study. J Ambient Intell Human Comput 2019 Jan;10(1):57-76. [doi: 10.1007/s12652-017-0598-x]
7. Sanders RD, Hughes F, Shaw A, et al. Perioperative Quality Initiative consensus statement on preoperative blood pressure, risk and outcomes for elective surgery. Br J Anaesth 2019 May;122(5):552-562. [doi: 10.1016/j.bja.2019.01.018] [Medline: 30916006]
8. Abdullah HR, Sim YE, Hao Y, et al. Association between preoperative anaemia with length of hospital stay among patients undergoing primary total knee arthroplasty in Singapore: a single-centre retrospective study. BMJ Open 2017 Jun 8;7(6):e016403. [doi: 10.1136/bmjopen-2017-016403] [Medline: 28600378]
9. Curran T, McDuff D, Liu X, et al. Camera-based remote photoplethysmography for blood pressure measurement: current evidence, clinical perspectives, and future applications. Conn Health Telemed 2023 Jun 16;2(2):200004. [doi: 10.20517/chatmed.2022.025]
10. Haugg F, Elgendi M, Menon C. Effectiveness of remote PPG construction methods: a preliminary analysis. Bioengineering (Basel) 2022 Sep 20;9(10):485. [doi: 10.3390/bioengineering9100485] [Medline: 36290452]
11. Man PK, Cheung KL, Sangsiri N, et al. Blood pressure measurement: from cuff-based to contactless monitoring. Healthcare (Basel) 2022 Oct 21;10(10):36292560. [doi: 10.3390/healthcare10102113] [Medline: 36292560]
12. Curran T, Liu X, Mcduff D, et al. Camera-based remote photoplethysmography to predict blood pressure in clinic patients with cardiovascular disease. J Hypertens 2023;41(Suppl 3):e4. [doi: 10.1097/01.hjh.0000938916.81752.7f]
13. Xiao H, Liu T, Sun Y, Li Y, Zhao S, Avolio A. Remote photoplethysmography for heart rate measurement: a review. Biomed Signal Process Control 2024 Feb;88:105608. [doi: 10.1016/j.bspc.2023.105608]
14. Yoon YH, Kim J, Lee KJ, et al. Blood pressure measurement based on the camera and inertial measurement unit of a smartphone: instrument validation study. JMIR Mhealth Uhealth 2023 Sep 8;11:e44147. [doi: 10.2196/44147] [Medline: 37694382]
15. Kamdar NV, Huverserian A, Jalilian L, et al. Development, implementation, and evaluation of a telemedicine preoperative evaluation initiative at a major academic medical center. Anesth Analg 2020 Dec;131(6):1647-1656. [doi: 10.1213/ANE.0000000000005208] [Medline: 32841990]

XSL•FO
RenderX

16. Sharma S, Kumari B, Ali A, et al. Mobile technology: a tool for healthcare and a boon in pandemic. J Family Med Prim Care 2022 Jan;11(1):37-43. [doi: 10.4103/jfmpc.jfmpc_1114_21] [Medline: 35309626]
17. Palatini P. Ambulatory and home blood pressure measurement: complementary rather than competitive methods. Hypertension 2012 Jan;59(1):2-4. [doi: 10.1161/HYPERTENSIONAHA.111.184184] [Medline: 22083157]
18. Banegas JR, Ruilope LM, de la Sierra A, et al. Clinic versus daytime ambulatory blood pressure difference in hypertensive patients: the impact of age and clinic blood pressure. Hypertension 2017 Feb;69(2):211-219. [doi: 10.1161/HYPERTENSIONAHA.116.08567] [Medline: 28028191]
19. Parati G, Omboni S, Albini F, et al. Home blood pressure telemonitoring improves hypertension control in general practice. The TeleBPCare study. J Hypertens 2009 Jan;27(1):198-203. [doi: 10.1097/hjh.0b013e3283163caf] [Medline: 19145785]
20. AbuDagga A, Resnick HE, Alwan M. Impact of blood pressure telemonitoring on hypertension outcomes: a literature review. Telemed J E Health 2010 Sep;16(7):830-838. [doi: 10.1089/tmj.2010.0015] [Medline: 20815751]
21. Lu X, Yang H, Xia X, et al. Interactive mobile health intervention and blood pressure management in adults. Hypertension 2019 Sep;74(3):697-704. [doi: 10.1161/HYPERTENSIONAHA.119.13273] [Medline: 31327259]
22. Allado E, Poussel M, Moussu A, et al. Accurate and reliable assessment of heart rate in real-life clinical settings using an imaging photoplethysmography. JCM 2022 Oct 17;11(20):6101. [doi: 10.3390/jcm11206101]
23. Luo H, Yang D, Barszczyk A, et al. Smartphone-based blood pressure measurement using transdermal optical imaging technology. Circ Cardiovasc Imaging 2019 Aug;12(8):e008857. [doi: 10.1161/CIRCIMAGING.119.008857] [Medline: 31382766]
24. Monk Jr EP. The Monk Skin Tone Scale. SocArXiv. Preprint posted online on Mar 23, 2023. [doi: 10.31235/osf.io/pdf4c]
25. Mancia G, Kreutz R, Brunström M, et al. 2023 ESH guidelines for the management of arterial hypertension The Task Force for the management of arterial hypertension of the European Society of Hypertension: endorsed by the International Society of Hypertension (ISH) and the European Renal Association (ERA). J Hypertens 2023 Dec 1;41(12):1874-2071. [doi: 10.1097/HJH.0000000000003480] [Medline: 37345492]
26. Kapoor M, Holman B, Cohen C. Contactless and calibration-free blood pressure and pulse rate monitor for screening and monitoring of hypertension: cross-sectional validation study. JMIR Cardio 2024 Aug 5;8:e57241. [doi: 10.2196/57241] [Medline: 39102277]
27. Stergiou GS, Alpert B, Mieke S, et al. A universal standard for the validation of blood pressure measuring devices: Association for the Advancement of Medical Instrumentation/European Society of Hypertension/International Organization for Standardization (AAMI/ESH/ISO) collaboration statement. J Hypertens 2018 Mar;36(3):472-478. [doi: 10.1097/HJH.0000000000001634] [Medline: 29384983]
28. Yang Y, Zha K, Chen Y, Wang H, Katabi D. Delving into deep imbalanced regression. Presented at: Proceedings of the 38th International Conference on Machine Learning; Jul 18-24, 2021; Virtual p. 11842-11851 URL: https://proceedings.mlr.press/v139/yang21m.html [accessed 2025-05-22]
29. Wu BF, Wu BJ, Tsai BR, Hsu CP. A facial-image-based blood pressure measurement system without calibration. IEEE Trans Instrum Meas 2022;71:1-13. [doi: 10.1109/TIM.2022.3165827]
30. Kraft D, Bieber G, Fellmann M. Reliability factor for accurate remote PPG systems. Presented at: PETRA '23: Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments; Jul 5-7, 2023; Corfu, Greece p. 448-456. [doi: 10.1145/3594806.3596573]
31. Shaik T, Tao X, Higgins N, et al. Remote patient monitoring using artificial intelligence: current state, applications, and challenges. WIREs Data Min and Knowl 2023 Mar;13(2). [doi: 10.1002/widm.1485]
32. Almarshad MA, Islam MS, Al-Ahmadi S, BaHammam AS. Diagnostic features and potential applications of PPG signal in healthcare: a systematic review. Healthcare (Basel) 2022 Mar 16;10(3):35327025. [doi: 10.3390/healthcare10030547] [Medline: 35327025]
33. Casalino G, Castellano G, Zaza G. Evaluating the robustness of a contact-less mHealth solution for personal and remote monitoring of blood oxygen saturation. J Ambient Intell Humaniz Comput 2023;14(7):8871-8880. [doi: 10.1007/s12652-021-03635-6] [Medline: 35043065]
34. Jain G, Joshi AM, Maddila RK, Vipparthi SK. A review of non-invasive HbA1c and blood glucose measurement methods. Presented at: 2021 IEEE International Symposium on Smart Electronic Systems (iSES); Dec 18-22, 2021; Jaipur, India URL: https://ieeexplore.ieee.org/abstract/document/9701015 [accessed 2025-05-22] [doi: 10.1109/iSES52644.2021.00086]

Abbreviations

AAMI: Association for the Advancement of Medical Instrumentation
AI: artificial intelligence
ANSI: American National Standards Institute
BP: blood pressure
DBP: diastolic blood pressure
ESH : European Society of Hypertension
HR: heart rate

XSL•FO
RenderX

HRV: heart rate variability
ISO: International Organization for Standardization
LDS: label distribution smoothing
mHealth: mobile health
MST: Monk Skin Tone
RGB: red, green, and blue
rPPG: remote photoplethysmography
RR: respiration rate
SBP: systolic blood pressure
SpO2: blood oxygen saturation

Edited by A Mavragani; submitted 12.05.24; peer-reviewed by S Melville; revised version received 10.02.25; accepted 21.02.25; published 06.06.25.

Please cite as:
Tan SYL, Chai JX, Choi M, Javaid U, Tan BPY, Chow BSY, Abdullah HR
Remote Photoplethysmography Technology for Blood Pressure and Hemoglobin Level Assessment in the Preoperative Assessment Setting: Algorithm Development Study
JMIR Form Res 2025;9:e60455
URL: https://formative.jmir.org/2025/1/e60455
doi:10.2196/60455

© Selene Y L Tan, Jia Xin Chai, Minwoo Choi, Umair Javaid, Brenda Pei Yi Tan, Belinda Si Ying Chow, Hairil Rizal Abdullah. Originally published in JMIR Formative Research (https://formative.jmir.org), 6.6.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

mindLAMPVis as a Co-Designed Clinician-Facing Data Visualization Portal to Integrate Clinical Observations From Digital Phenotyping in Schizophrenia: User-Centered Design Process and Pilot Implementation

Karthik Sama[1*], MTECH; Jaya Sreevalsan-Nair[1*], BTECH, MS, PhD; Soumya Choudhary[2], MSc; Srilakshmi Nagendra[2], BE; Preethi V Reddy[2], MD, PhD; Asher Cohen[3], BA; Urvakhsh Meherwan Mehta[2,4], MD, PhD; John Torous[3], MD

[1]Graphics-Visualization-Computing Lab, International Institute of Information Technology Bangalore (IIIT-B), 26/C Electronics City, Hosur Road, Bangalore, India

[2]National Institute of Mental Health and Neuro Sciences, Bangalore, India

[3]Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, United States

[4]Consciousness Studies Program, School of Humanities, National Institute of Advanced Studies, Bangalore, India

[*]these authors contributed equally

Corresponding Author:
Jaya Sreevalsan-Nair, BTECH, MS, PhD
Graphics-Visualization-Computing Lab, International Institute of Information Technology Bangalore (IIIT-B), 26/C Electronics City, Hosur Road, Bangalore, India

Abstract

Background: The potential of digital mental health to transform care delivery in low- and middle-income countries is well established. However, there remains the need to clinically and organically adapt current tools to local needs. This paper explores the process of creating a novel data visualization system for a digital mental health app and outlines the necessary steps in the process. This work demonstrates co-design involving collaboration between teams across geographies and disciplines based on clinicians' requirements.

Objective: This study aims to co-design a visualization dashboard app for clinicians through a design study with a multidisciplinary team consisting of clinicians in Boston and Bangalore, mindLAMP software developers in Boston, and computer scientists with visualization expertise in Bangalore. The app is designed to visualize derivatives of both active and passive data of patients with schizophrenia to support the research contexts of digital psychiatry clinics in India.

Methods: The mindLAMP app, already used in many countries today, is adapted to offer a new clinician-facing data visualization portal, mindLAMPVis. The novel web-based portal is designed to improve clinical integration for use in India. After building the new portal, the insights from this new portal are corroborated with known clinical observations of relapse using comparative visualization. The data were taken from the mindLAMP app and processed using multivariate analysis and dimensionality reduction to make it easy and manageable for clinicians to analyze. These techniques are integrated in mindLAMPVis, thus making it a locally co-designed, developed, and deployed tool. A feasibility study of the pilot implementation of the app was completed through a domain expert study with clinician-driven case studies.

Results: To assess the system, we preloaded data from 24 patients with schizophrenia, including those with relapses. Through case examples focusing on relapse risk prediction in schizophrenia, mindLAMPVis is used to identify different visualization methods to compare different analytical results for each patient. In partnership with clinicians for co-designing the app, we explored the feasibility of a comparative visualization tool for discovering patterns across different time stamps for a single patient or any patterns across patients related to the relapse episode. As an example of reverse translation, mindLAMPVis offers new features that complement the original features of mindLAMP, highlighting the mutual benefit of software adaptation and collaborative design.

Conclusions: mindLAMPVis is a tailored tool designed for use in India, but it can aid in identifying and comparing behavioral patterns that may indicate clinical risk for patients in any country. mindLAMPVis offers an example of how, through technical design, feedback, and real-world clinical testing, it is feasible to adapt current software tools to meet local needs and even exceed the use cases of the original technology. mindLAMPVis also successfully incorporates both active and passive digital phenotyping data.

(JMIR Form Res 2025;9:e70073) doi:10.2196/70073

XSL•FO
RenderX

KEYWORDS

co-design; data visualization; browser-based dashboard; multivariate data mining; dimensionality reduction; local technology adoption; digital phenotyping; schizophrenia; anomaly detection; mobile phone

Introduction

Despite the well-known potential of digital mental health to advance psychiatric care, a lack of clinical integration has hampered meaningful use [1]. Low- and middle-income countries (LMICs) are ideal settings for these technologies to help advance care and reduce access barriers. However, many of the digital tools such as apps need not be designed to cater to the clinicians' needs [2] or locally developed [3], which affects the adoption of the technology. At the same time, this need not be a barrier as digital technologies can be easily and inexpensively customized with collaborative efforts. This paper explores one example of such tailoring by focusing on the use of the open-source mindLAMP smartphone app [4,5], where clinicians and computer scientists co-design the visualization portal that is integrable with the mindLAMP app.

Schizophrenia is considered an area of less innovation, often due to the complexity of the symptoms. However, the actual evidence supports numerous advances in digital health [6]. Aspects of the illness, such as relapse, continue to be hard for clinicians, family members, and patients themselves to predict [7]. Hence, better risk assessment and management of relapses depend on continuous monitoring and timely interventions. Recent technological developments have enabled the collection of large-scale datasets of passive (sensors) and active (surveys) data through smartphones. The availability of such data provides opportunities for understanding and predicting symptom relapses, with positive results shown across several studies [5,8]. However, the high volume, complexity, and dense time series of the data collected, along with the requirement to maintain a high level of data privacy, present significant challenges in its interpretation and clinical application.

Effective data visualization transforms this raw, large-scale data into meaningful insights to be consumed by both clinicians and patients. It is also critical for ethical considerations around digital data capture and the use of patients' data [9]. The visual representation of data in a clear and comprehensible manner helps in identifying patterns, trends, and anomalies that may not be observable through other data analysis methods [10], referred to as data mining. Recent work shows the potential of transforming data to new formats, such as networks, and performing network analysis and visualization to derive cluster-based inferences from large-scale patient datasets [11]. Similarly, the use of deep learning models and visualization helps in predicting clinical risks from electronic health record datasets [12]. In the same vein, data mining and visualization are thus important in the case of schizophrenia, where understanding behavioral patterns and predicting relapse significantly impact patient outcomes.

The current literature strongly supports the notion that data visualizations have the potential to improve clinicians' engagement and effectively communicate trends in patient data [13,14]. Personal health care data visualizations improve health literacy [15], enhance communication between patients and health care providers, support informed health care decisions, and enable quicker data comprehension [13,16]. Furthermore, effective visualizations increase users' trust and willingness to share their tracking data through transparency in information usage and its integration into their health care. The use of a dashboard framework for disease surveillance is viable for effective communication of data trends in the population [17]. A recent study shows that, for patients with conditions of depression, epilepsy, and multiple sclerosis, their own choice of visualizations of their personal health data collected using wearables improves self-management of their conditions [18].

Despite these benefits concerning patient data in general, the unique nature of digital phenotyping data has led to limited efforts in its visualizations. Previous works show the role of visualizations in aiding clinicians to understand patients' health-related behavior from digital phenotyping data [19,20]. However, such studies are limited to the analysis of passive data of patient behavior (such as physical location and smartphone usage patterns) by health care providers [21,22]. Such studies also do not investigate the integrability of such analysis to the data management framework. Although its potential for real-time analysis of passive data is being explored, visualizations of active data (such as surveys) enhance the explainability of observations in the retrospective analysis of symptom relapse cases. Given that there is no simple or universal rule for visualizations and limited collective experiences of patients and clinicians with visualizations of mental health care data, there is a clear need for expanded research of novel or uncharted methods and strategies.

In this study, we propose mindLAMPVis, which is a web-based visualization tool designed to enhance the usability of data processed through the mindLAMP platform. The data are prepared for visualization using multivariate data mining through dimensionality reduction. The inspiration for the tool derives from the request and needs of local clinicians in Bangalore, India, after initial work with the mindLAMP platform for a research study, which was developed in Boston, United States. However, the need for web-based visualization is universal, and the use of this tool is global. One of the key challenges with data collection in digital phenotyping is the lack of consistency of data recording by the patient, which often leads to insufficient data for further analysis. Comparative analysis of the patient's data over time or patterns across different patients helps clinicians to confirm the analytical results with known clinical observations of symptom relapse and visually share or discuss with patients. Comparative visualizations in the new mindLAMPVis software enable such confirmations. Through case studies, we demonstrate how an exploratory tool, such as mindLAMPVis, identifies and compares behavioral patterns, which helps psychiatrists and other domain experts to offer more data-driven, personalized, and shared decision-making-based care.

XSL•FO
RenderX

Our research aims to gain actionable insights into raw data by integrating advanced data mining and visualization techniques into clinical practice. The relevance of such a tool lies in the global significance of digital phenotyping through mindLAMP. We hypothesize that a comparative visualization tool improves the understanding and management of schizophrenia, thereby supporting more effective, timely interventions and fostering a better provider-patient relationship. The findings from our study lead to continued co-design and refinement of data visualizations, ensuring their accessibility to both patients and clinicians. A deep dive into the state of the art of comparative visualization and methods is given in the Multimedia Appendix 1. The pilot implementation of mindLAMPVis is available in a publicly accessible website [23]. The source code for the app with a dummy input dataset is also available in a publicly accessible website [24].

Methods

Study Design

Our proposed clinician-facing dashboard, mindLAMPVis, is intended to compare different visualizations of the same patient. For exploratory analysis, the same visualization of different patients can be compared. Our goal was to co-design the app with clinicians, visualization experts, and software developers in Bangalore and Boston to suit the local clinical needs in LMIC.

Co-Design Process

The dashboard is an outcome of the co-design process, that is, a collaborative design process, involving clinicians in Bangalore and Boston, mindLAMP developers in Boston, and computer scientists with visualization expertise from Bangalore. It is well understood that the usability and creativity differences between different groups must be addressed through a user-led design process [17]. The initial discussions of the multidisciplinary team arrived at the following requirements and design principles:

- The data collected on the mindLAMP app were already processed for further consumption, and the motivation for developing the dashboard was to visualize the underused derived data, such as home time, survey responses, etc.
- Clinicians are particularly interested in exploring the occurrence of phenomena with sparse data, such as relapse among individuals with schizophrenia. Data of such events tend to be sparse, owing to the difficulties in measurements of unpredictable events by the clinician and reliance on patient self-reports. Despite these challenges, such events are critical for clinicians to understand the disease progression in patients. Therefore, the team established the need for a visualization tool for research by clinicians that precedes clinical use.
- The clinicians clarified the local needs for such a tool. In a busy clinic in the Indian setting, mental health professionals would need quick access to meaningful and actionable digital behaviors obtained via the data generated from the mindLAMP app used by patients. A clinician-facing dashboard should be designed to bridge this gap by providing visual representations of complex high-dimensional data. These visualizations must first be interpretable by the clinicians and, optionally, by the patient. This serves as a stepping stone toward establishing digital psychiatry clinics in India, which are expected to experience high patient volume.
- Since the visualization is intended to be exploratory rather than confirmatory, it was decided that the portal would feature comparative visualizations. This is because comparative visualizations enable a better understanding of the data through exploration.

We followed the design study methodology (DSM) used for visualization projects [25]. DSM prescribes a 9-stage workflow for implementing visualization projects in real-world scenarios. This 9-stage framework prominently features the collaboration or co-design of domain experts and visualization researchers. The 9 sequential stages are “learn,” “winnow,” “cast,” “discover,” “design,” “implement,” “deploy,” “reflect,” and “write,” where these stages are also interlinked owing to the iterative nature of the design study. These stages are grouped into 3 top-level phases, namely, “precondition,” “core,” and “analysis,” which correspond to personal, inward-facing, and outward-facing validations, respectively. The description of the 9 stages is given in the Multimedia Appendix 1.

Here, we describe how the phases progressed in our work.

1. The precondition phase, comprising of stages from “learn” to “cast,” was implemented quicker compared to similar projects owing to the agreement of all members of the collaborative team on the mindLAMPVis specifications. The precondition phase determines the feasibility of the project, which went smoothly for our work.
2. The core phase, comprising of stages from “discover” to “deploy,” was mostly implemented by the visualization team in collaboration with the mindLAMP developer team. Owing to the physical distance between the teams and the voluntary nature of the software development process, this phase took the longest time to complete. During the “design” stage, the team narrowed down the dimensionality reduction methods and the visualizations. The “design” and “implement” stages involved back-and-forth iterations to improve on the visualizations and correct the data processing techniques based on the domain knowledge of the clinicians. During the “deploy” stage, the clinicians expressed their feedback to the entire team on the usefulness of different visualizations provided in the dashboard.
3. The analysis phase includes the “reflect” and “write” stages, which were implemented by the clinicians and the entire team, respectively. The “reflect” stage involved identifying appropriate case studies to demonstrate the usability of mindLAMPVis. In this work, the “reflect” stage involves the presentation of the usability of mindLAMPVis as a pilot implementation. A systematic study of the clinical efficacy of the visualizations is currently out of the scope of this work.

For the pilot implementation of mindLAMPVis, we have completed all the phases, except that the analysis and reflection stages have been executed only partially. In the cast stage of our work, we determined that the clinicians play the role of both designers and end users. The mindLAMP developers provided support for the existing software and access to data, and the

XSL•FO
RenderX

visualization tool developers determined the data processing algorithms, visualization methods, and the actual implementation. The entire team participated in the deploy and write stage.

Ethical Considerations

This study's protocols were approved by the local institutional ethics committees at each of the sites for the Smartphone Health Assessment for Relapse Prevention (SHARP) project [5]. All participants provided written informed consent. Only preprocessed, deidentified data were used in the analysis. All study participants received compensation for travel and refreshments during each of the study visits. In this study, we have used the anonymized data collected from patients with a schizophrenia diagnosis from the SHARP project.

Clinical Setting

This was a 12-month multisite, longitudinal observational study where individuals with and without schizophrenia used the mindLAMP application for active and passive data collection alongside monthly clinical evaluations for symptom relapse and cognitive performance. Patients were recruited from outpatient hospital settings if they had a diagnosis of schizophrenia based on DSM-5 criteria as ascertained by a trained psychiatrist.

Dataset

We preloaded data from 24 patients with a schizophrenia diagnosis from the SHARP project [5]. The patients were recruited at the National Institute of Mental Health and Neurosciences, Bangalore. We used subsets of data from both active and passive data for this feasibility study. Active data include 6 surveys (see details in the Multimedia Appendix 1), and passive data include GPS and accelerometer information. In mindLAMPVis, we use the active data and the significant locations computed from the GPS locations [26]. Survey data was collected as responses to a questionnaire, and we use the data for 5 questionnaires, namely, mood, sleep, social, anxiety, and psychosis. A detailed description of the entire dataset and the subset we used is given in the Multimedia Appendix 1.

Imputation

Missing data is inherent to digital phenotyping. Our tool supports two types of imputation methods to fill in the missing values in the data, namely, (1) last observation carried forward (LOCF) and (2) Multiple Imputation by Chained Equations (MICEs).

LOCF repeats the last observed value until the item where data are available. This method has the least computational cost and the least accuracy. While LOCF is the third most used imputation method, multiple imputation methods are considered more efficient. Multiple imputation uses different simulation models to provide multiple values for the missing data instance. MICE assumes that a multivariate distribution can be used for the incomplete variable with missing data [27]. Thus, in this method, missing values are filled using a distribution fitted to the data and are more accurate. The working of these algorithms is described in detail in the Multimedia Appendix 1.

Dimensionality Reduction

We analyze the responses to 5 survey sections, namely, mood, sleep, anxiety, social, and psychosis factors [25]. To visualize the data in each survey section, we use multiple correspondence analysis (MCA) for dimensionality reduction, where we treat the questions in each survey section as the dimensionality of the section. The responses to each section are represented as a vector to implement dimensionality reduction.

If the time-series data for each section have a large variance in 1 or 2 components in MCA, we can conclude that the patient responses are consistent over time. This is checked using the first and second eigengaps where a high value of eigengap indicates that the responses to the questions in the specific survey section have been highly consistent. The lower-dimensional vectors and derived data are then used for visualizations.

Clustering and Coassociation Matrix

We also cluster dates of data availability within a chosen period to check if the responses to the questions indicate unique behavioral patterns in the patient. We use the vector format of the responses for clustering in 2 different ways, namely, as an aggregate vector (AV) and a complete vector (CV). The length of the AV is that of the number of sections, and the vector component is the average value of the responses in the section. However, the CV retains all the information by concatenating the vectors for each section, thus forming a long vector. Thus, the length of the CV is the total number of questions in the survey. The AV or CV is the representation of the mental state of the patient for a specific date.

In data mining, one of the key influencing factors for the implementation of clustering is the number of clusters. The cluster count can be either given as user input or implicitly determined from the data itself. We refer to the clustering methodologies as predefined clustering and natural clustering, respectively. The former determines k clusters for the user-defined value of k based on prior knowledge, while the latter algorithmically determines the natural partition available in the data.

In cases where there is no ground truth or baseline solutions in data mining, we use the consensus of outcomes from several methods, thus establishing strength in numbers. The consensus in the clustering of items is represented in the form of a matrix, namely, the coassociation matrix. This matrix represents how likely any 2 items of the entire set are to be present in the same cluster. If they are highly similar, then their coassociation value is 1, and if highly dissimilar, then it is 0. Thus, the coassociation between any 2 dates in our case is represented as a ratio between 0 and 1.

More details on the clustering and coassociation matrix are given in the Multimedia Appendix 1.

Visualizations

In our proposed visualization tool, mindLAMPVis, we support the following visualizations (V) of active (A) and passive (P) data:

XSL•FO
RenderX

- MCA trend (V1A) of survey data (Figure S1 in Multimedia Appendix 1),
- MCA eigengap (V2A) of survey data (Figure S2 in Multimedia Appendix 1),
- Date-clustering (V3A) of survey data (Figure 1),
- Home time (V1P) of significant locations data (Figure 2),
- Significant location (V2P) of significant locations data (Figure 2).

Figure 1. Matrix visualization of the consensus matrix of clustering of calendar dates for chosen aggregation and clustering strategies, for a selected patient. The orange highlight shows the choice of AV for this example. This chart is generated from the anonymized data of a patient from the Bangalore cohort collected during 2021-2022 in the SHARP project [5]. AV: aggregate vector; SHARP: Smartphone Health Assessment for Relapse Prevention.

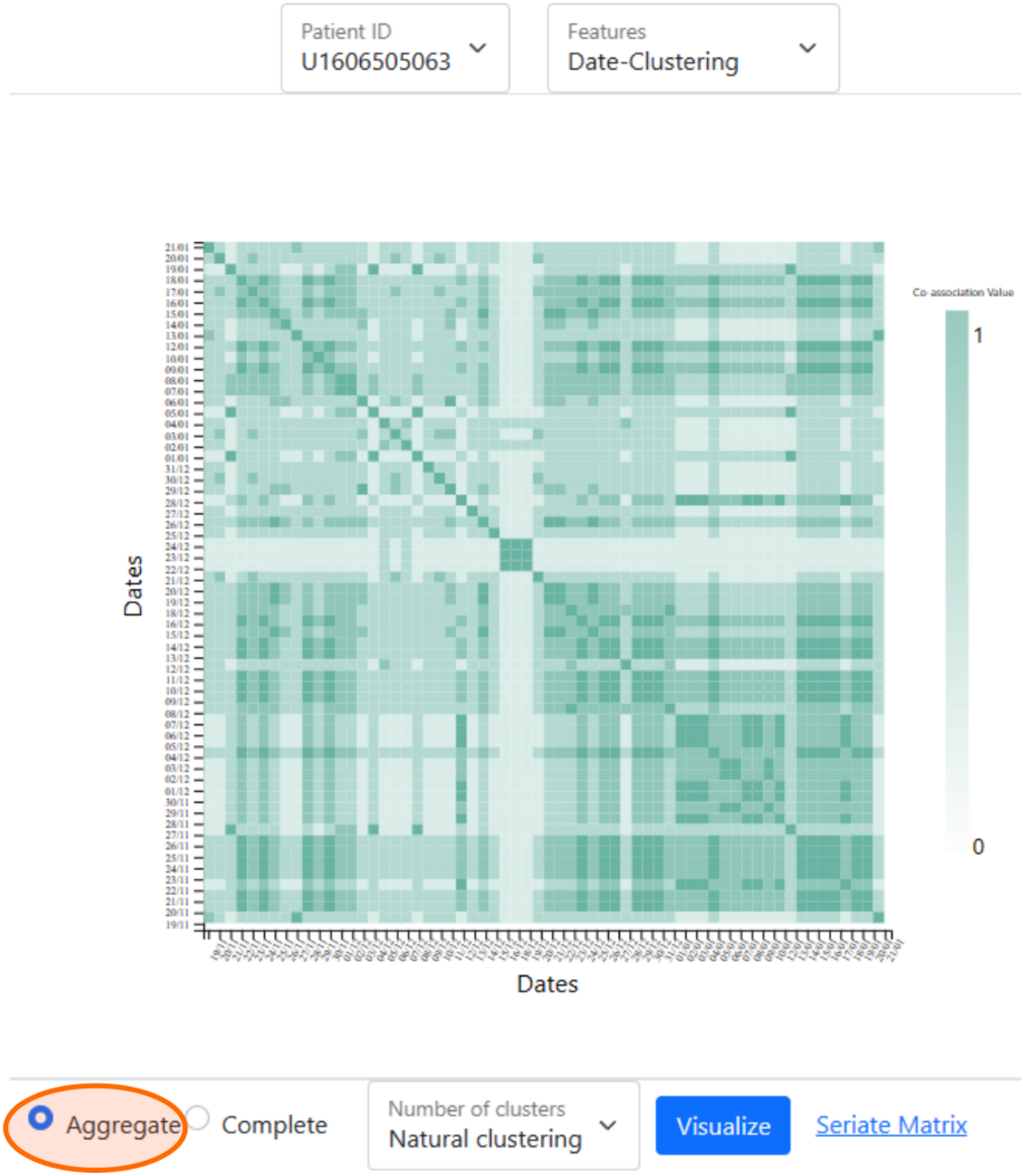

XSL•FO
RenderX

Figure 2. Patient-specific calendar view visualizations of (left) home time (V1P) showing the percentage of home time colored using a gradient color palette, (right) the distribution of significant locations (V2P) using pie-chart glyphs. This chart is generated from the anonymized data of a patient from the Bangalore cohort collected during 2021-2022 in the SHARP project [5]. P: passive; V: visualization; SHARP: Smartphone Health Assessment for Relapse Prevention.

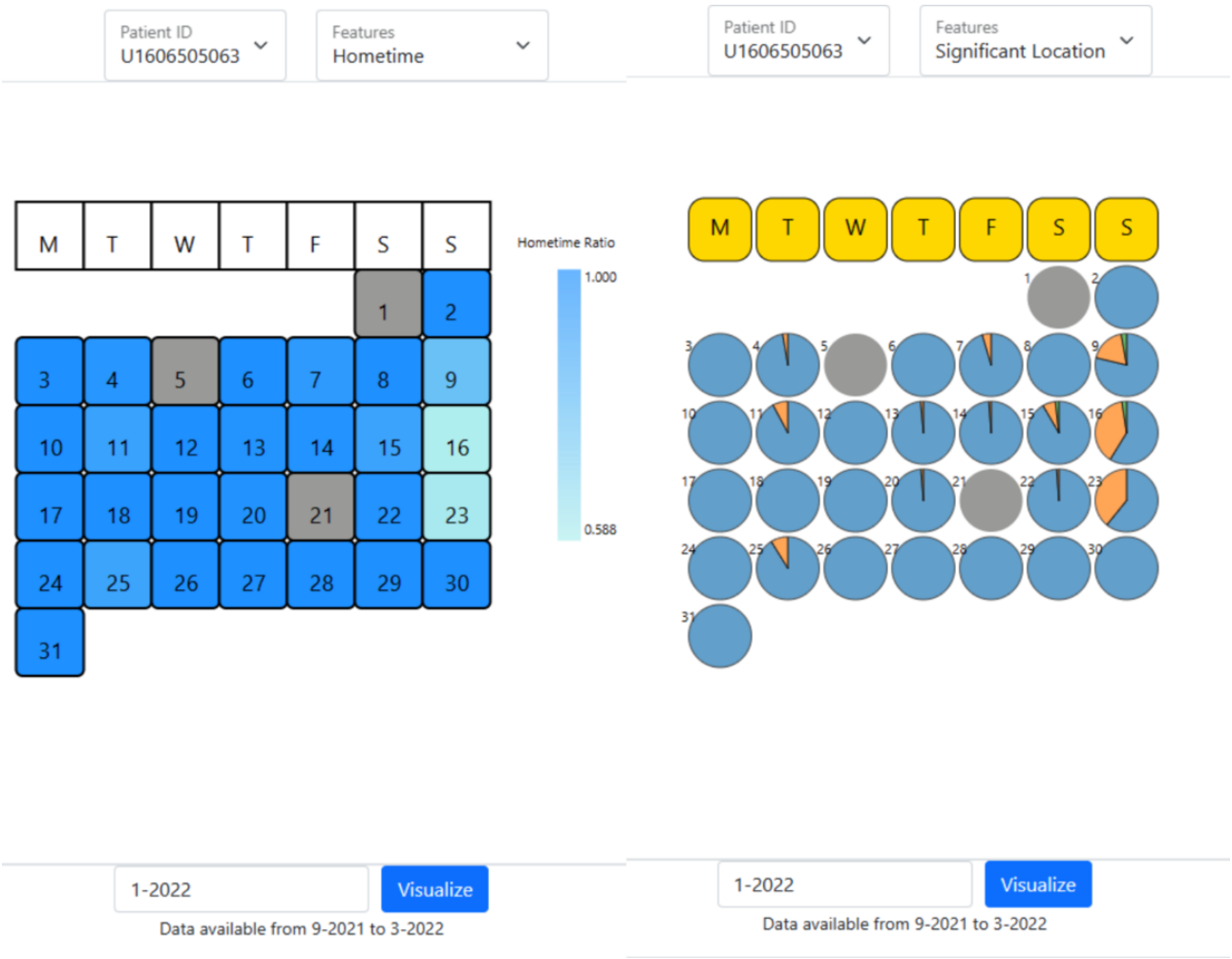

Here, we mention the knowledge derived from each visualization, but the interpretation of the visualizations is given in detail in the Multimedia Appendix 1. V1A shows the temporal trend of the first MCA component for each survey section. V2A shows which survey section has coherent or consistent response data. V3A shows consistency of behavior over a set of dates, including periods. V1P and V2P show the amount of time spent at home (or the most significant location) and the proportion of time spent daily across significant locations. They demonstrate trends over calendar months. These 5 views enable the clinician to observe anomalous behavior in the trends a month before the relapse and during the relapse for the concerned patient. The anomalies imply inconsistencies in the behavior of the patient leading up to the relapse event.

The layout of our tool is organized as a 2-panel one, along with control widgets to allow comparative visualizations (Figure 3). The checklist of the iCHECK-DH guidelines [28] for reporting this design study is present as Checklist 1.

XSL•FO
RenderX

Figure 3. View of the 2-panel graphical user interface with annotations of its components. A, B, and C indicate feature-controls, canvas, and feature-parameter-controls, respectively, and the subscripts L and R indicate the left and right panels, respectively. The left and right panels have the same design with clean lines. These charts in the screenshot are generated from the anonymized data of a patient from the Bangalore cohort collected during 2021-2022 in the SHARP project [5]. L: left; R: right; SHARP: Smartphone Health Assessment for Relapse Prevention.

Results

Case Study

This paper describes the development and feasibility of using mindLAMPVis to draw parallels in understanding individual-specific relapse signatures in patients with schizophrenia who participated in a research study in India.

The clinicians in Bangalore performed a deep dive analysis of the visualizations obtained for 2 specific patients, A and B, from mindLAMPVis. Both patients have sufficient active data for performing an analysis. Patient X had a single relapse, and patient Y had two relapses within 3 months. We compare our visualizations with those of anomaly plots [5] in Figures 4-8, which are for interpreting the passive data exclusively. The quantitative analysis of the MCA trend values for patient X (Table 1) and patient Y (Table 2) also shows interesting anomalies.

XSL•FO
RenderX

Figure 4. MCA trend (V1A) visualizations for patient X using MICE imputation for the 5 survey sections, namely, (A) mood, (B) sleep, (C) social, (D) anxiety, and (E) psychosis. The anomalous points with high-frequency spikes coincide within a month leading to relapse. This chart is generated from the anonymized data of the patient from the Bangalore cohort collected during 2021-2022 in the SHARP project [5]. The clinicians logged a relapse in April 2022 for this patient. A: active; V: visualization; MCA: multiple correspondence analysis; MICE: Multiple Imputation by Chained Equation; SHARP: Smartphone Health Assessment for Relapse Prevention.

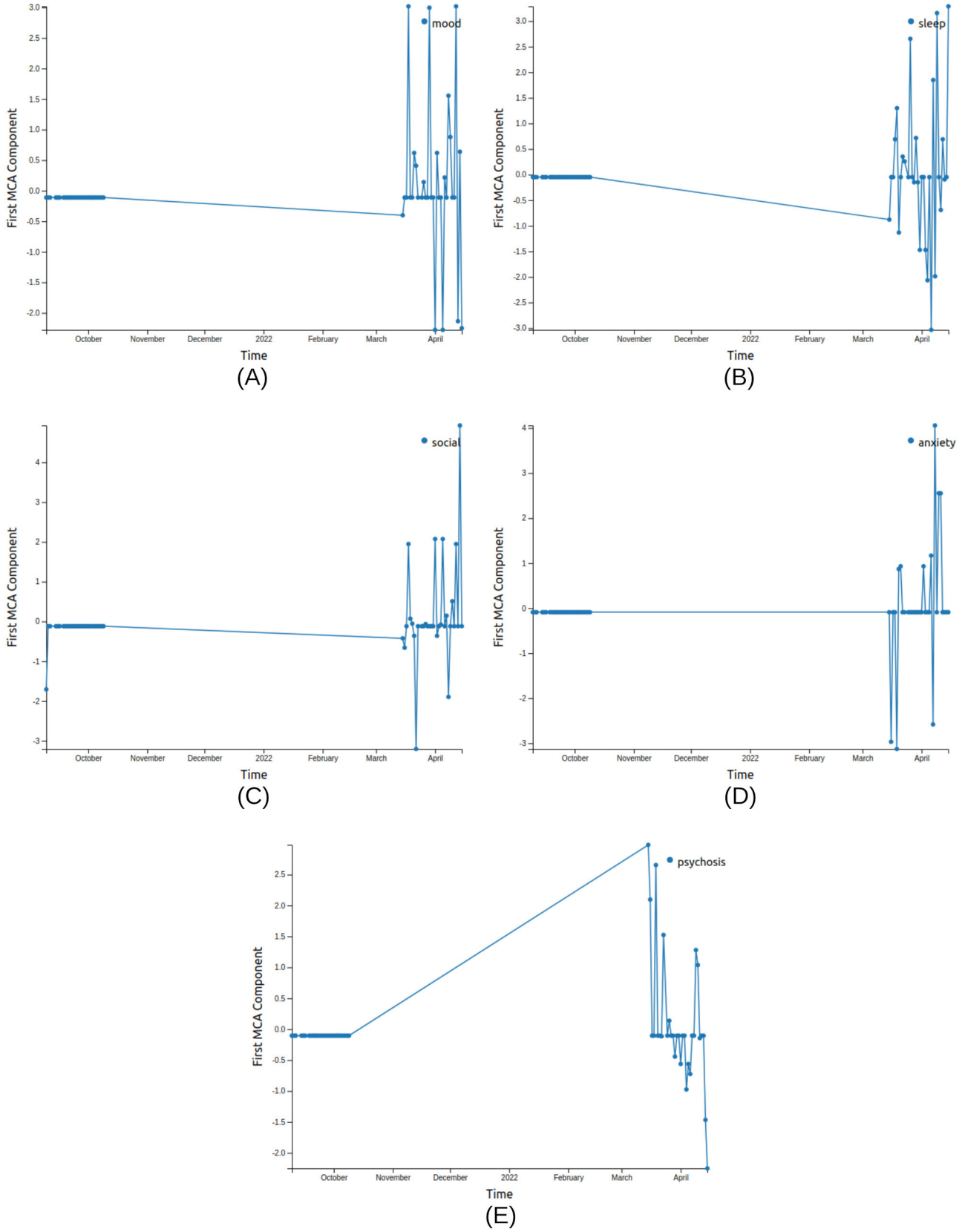

XSL•FO
RenderX

Figure 5. MCA trend (V1A) visualizations for patient Y using LOCF imputation for the 5 survey sections, namely, (A) mood, (B) sleep, (C) social, (D) anxiety, and (E) psychosis. The anomalous points with high-frequency spikes coincide with a month leading to relapse, and 2 relapse periods can be observed here. This chart is generated from the anonymized data of the patient from the Bangalore cohort collected during 2021-2022 in the SHARP project [5]. The clinicians logged multiple relapses in October 2021 and January 2022 for this patient. A: active; V: visualization; MCA: multiple correspondence analysis; LOCF: last observation carried forward; SHARP: Smartphone Health Assessment for Relapse Prevention.

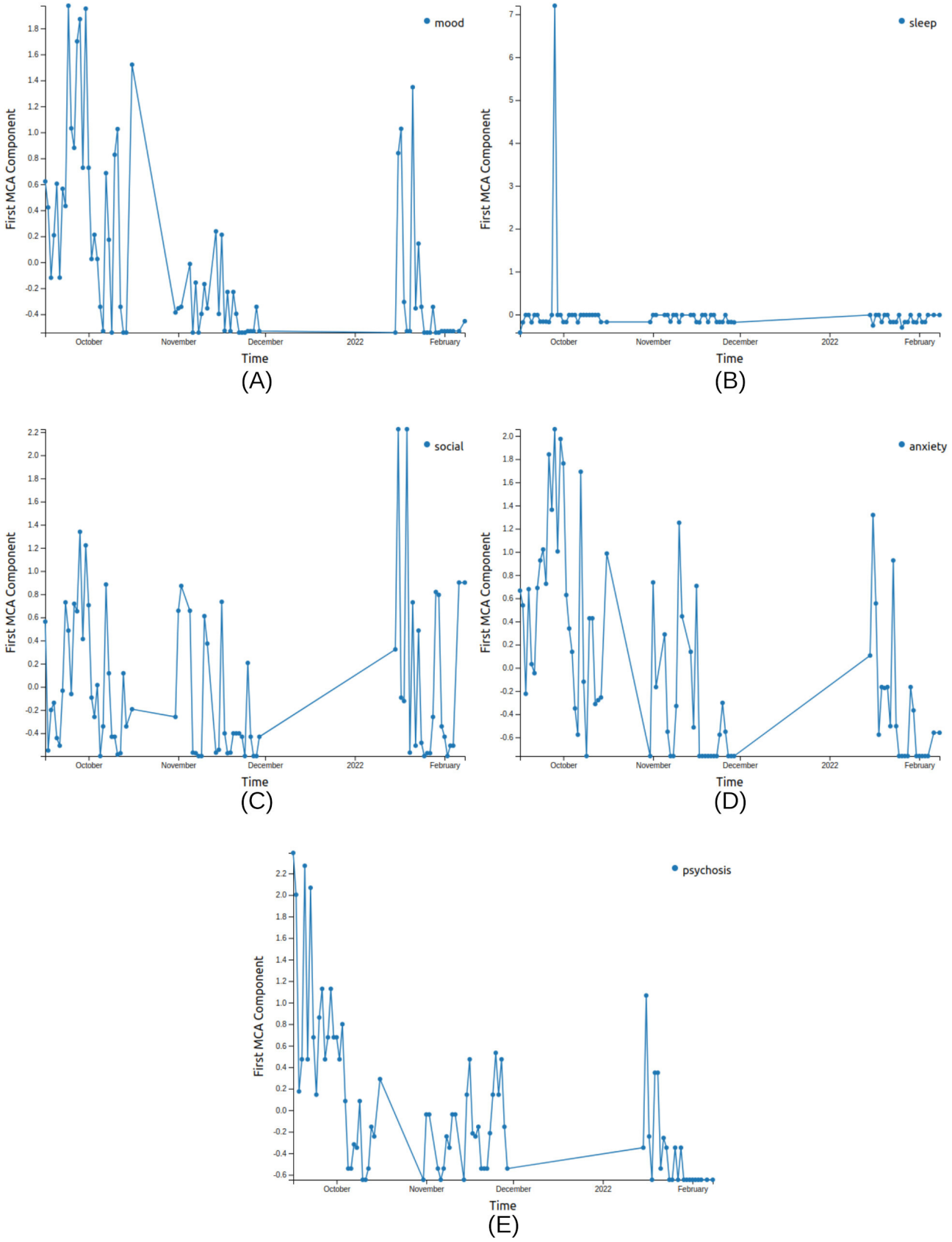

XSL•FO
RenderX

Figure 6. MCA eigengap (V2A) visualization for patient X using LOCF imputation for the 5 survey sections, shows significant first and second eigengaps in (left) anxiety and sleep, respectively. We observe similar patterns in (right) the anomaly plot for the same period. This chart is generated from the anonymized data of the patient from the Bangalore cohort collected during 2021-2022 in the SHARP project [5]. The clinicians logged a relapse in April 2022 for this patient. A: active; V: visualization; MCA: multiple correspondence analysis; LOCF: last observation carried forward; SHARP: Smartphone Health Assessment for Relapse Prevention.

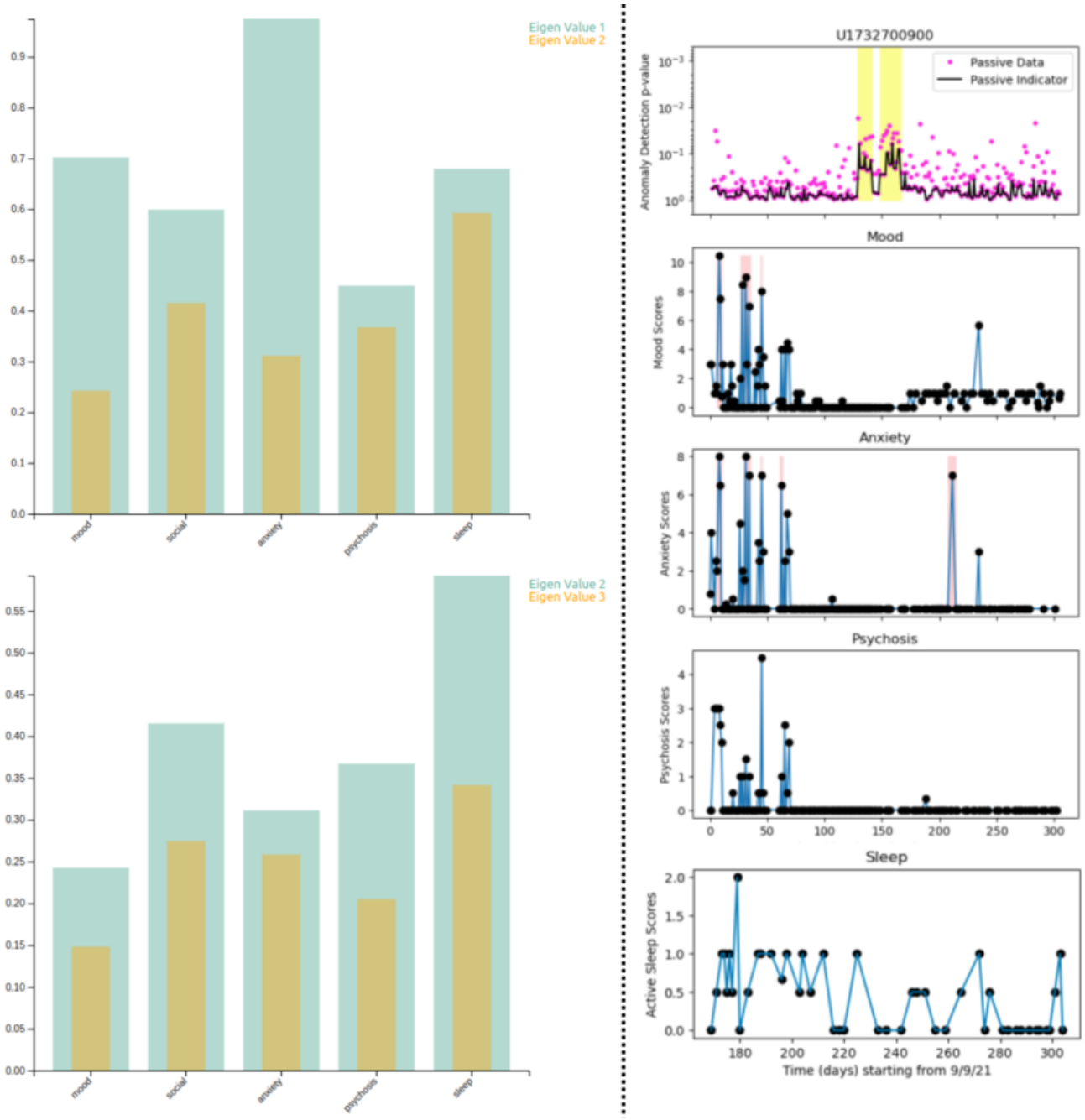

XSL•FO
RenderX

Figure 7. (Left) Home time (V1P) and (right) significant locations (V2P) visualizations of patient X, for the month (top) preceding and (bottom) of the relapse. This chart is generated from the anonymized data of a patient from the Bangalore cohort collected during 2021-2022 in the SHARP project [5]. The clinicians logged a relapse in April 2022 for this patient. P: passive; V: visualization; SHARP: Smartphone Health Assessment for Relapse Prevention.

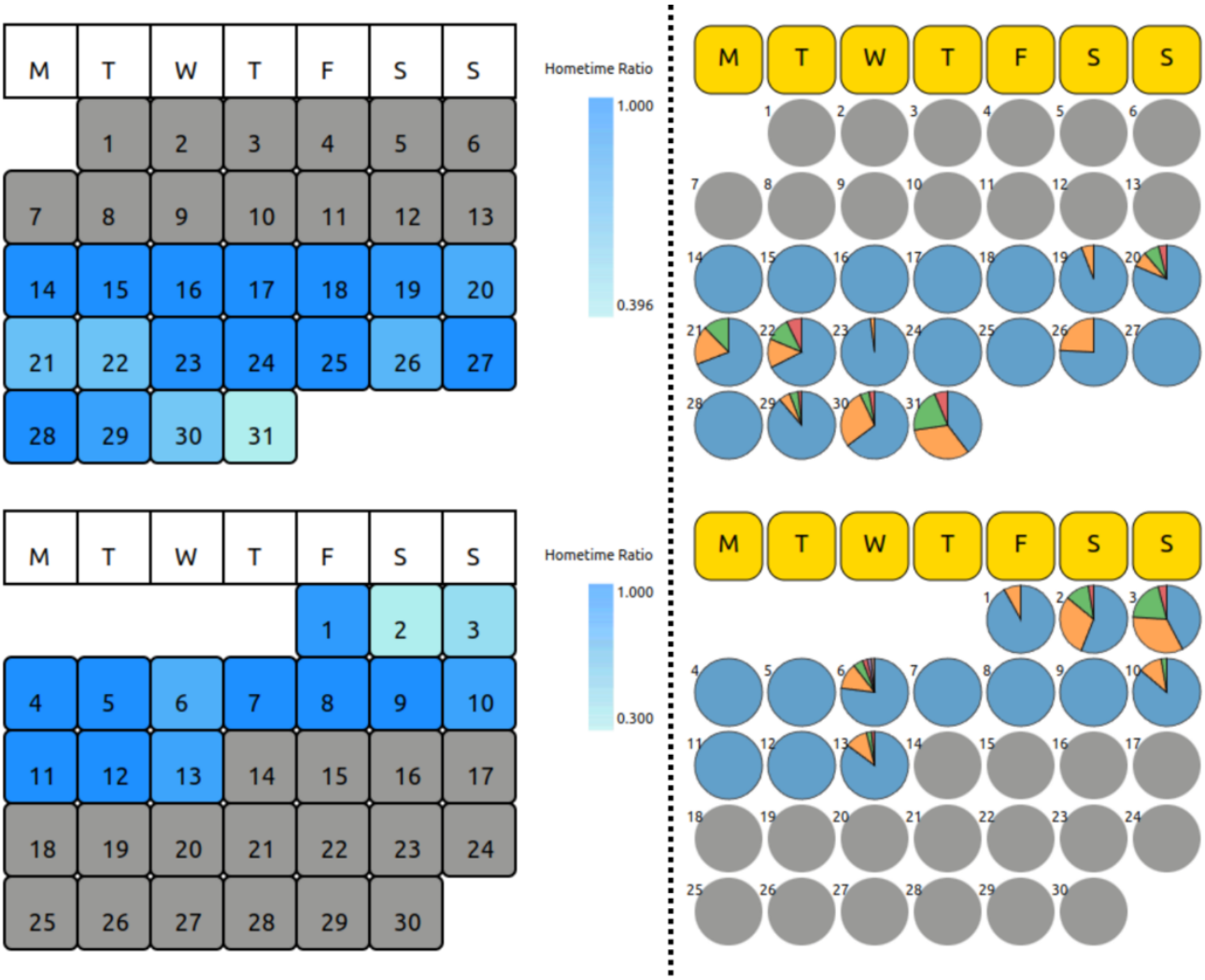

XSL•FO
RenderX

Figure 8. Coassociation matrix generated using (A) AV and (B) CV for NClust for patient X. The dates increment upward and rightward in the matrix, and the black dotted line marks the date marking 1 month before the relapse. This chart is generated from the anonymized data of the patient from the Bangalore cohort collected during 2021-2022 in the SHARP project [5]. The clinicians logged a relapse in April 2022 for this patient. AV: aggregated vector; CV: complete vector; NClust: natural clustering; SHARP: Smartphone Health Assessment for Relapse Prevention.

XSL•FO
RenderX

Table . MCA[a] trend analysis for patient X using MICE[b] imputation for the 5 survey sections, namely, mood, sleep, social, anxiety, and psychosis. This quantitative analysis of the data visualized in Figure 4 shows the mean and SD for the number of sample data points during the control period, and the month leading to the relapse date (relapse). This table is generated from the anonymized data of the patient from the Bangalore cohort collected during 2021 - 2022 in the SHARP[c] project [5]. The clinicians logged a relapse in April 2022 for this patient. For patient X, the SD is remarkably higher during the month leading to the relapse than that of the control period, for all sections.

Patient X	MCA with MICE imputation	
	Control (n=28), mean (SD)	Relapse (n=31), mean (SD)
Mood	0.13 (0)	–0.12 (1.29)
Sleep	–0.14 (0)	0.12 (1.37)
Social	0.11 (0.28)	0.1 (1.35)
Anxiety	0.14 (0)	–0.12 (1.37)
Psychosis	–0.23 (0)	0.21 (1.03)

[a]MCA: multiple correspondence analysis.

[b]MICE: Multiple Imputation by Chained Equation.

[c]SHARP: Smartphone Health Assessment for Relapse Prevention.

Table . MCA[a] trend analysis for patient Y using LOCF[b] imputation for the 5 survey sections, namely, mood, sleep, social, anxiety, and psychosis. This quantitative analysis of the data visualized in Figure 5 shows the mean and SD for the number of sample data points during the control period, the month leading to the first relapse date (relapse-1), and that to the second relapse date (relapse-2). This table is generated from the anonymized data of the patient from the Bangalore cohort collected during 2021 - 2022 in the SHARP[c] project [5]. The clinicians logged a relapse in April 2022 for this patient. For patient Y, the SD is higher during the control period than the months leading up to the relapse than that of the control period, for all sections except for social, when σ is relatively high during the second relapse.

Patient Y	MCA with LOCF imputation		
	Control (n=30), mean (SD)	Relapse-1 (n=25), mean (SD)	Relapse-2 (n=23), mean (SD)
Mood	0.51 (0.75)	–0.35 (0.22)	–0.26 (0.54)
Sleep	0.16 (1.31)	–0.09 (0.08)	–0.11 (0.1)
Social	0.07 (0.55)	–0.21 (0.5)	0.14 (0.84)
Anxiety	0.57 (0.78)	–0.36 (0.57)	–0.33 (0.56)
Psychosis	0.47 (0.85)	–0.19 (0.35)	–0.39 (0.42)

[a]MCA: multiple correspondence analysis.

[b]LOCF: last observation carried forward.

[c]SHARP: Smartphone Health Assessment for Relapse Prevention.

Anomaly Plots

Using both sensor and survey data as passive and active acquisitions, respectively, for relapse prediction in psychosis is highly feasible [8]. Anomaly detection is an effective data analysis method used to identify data points that significantly deviate from the majority, indicating potential errors, unusual events, or significant changes in the observed process. These algorithms enable researchers and clinicians to longitudinally identify abnormal deviations from an individual's typical sensor and survey data, which is indicative of relapse [5]. The anomaly plot (Figure 6) is the visual representation of potential relapse events and unusual data behavior, providing valuable insights for monitoring schizophrenia relapse. A detailed description of the interpretation of anomaly plots is given in the Multimedia Appendix 1.

Observations

Patient X

We compared the MCA trend for the patient between the LOCF and MICE imputation methods for data preprocessing. We found that the MCA trend for all 5 survey sections with MICE imputation shows a characteristic anomaly in the month leading to the relapse (Figure 4). There is an uncharacteristically high SD during the month leading to relapse in all sections (Table 1), indicative of erratic behavior during that period. We also visualized the passive data using home time (V1P) and significant locations (V2P; Figure 7).

Using MCA eigengap visualization (V2A), we observe that the first eigengap is significant for anxiety, and the second eigengap is significant for sleep, which is confirmed as spikes (pink highlights) at ~188 days in the anomaly plot for sleep and anxiety (Figure 6). These spikes occur in the anomaly plot a month before the relapse, as expected. We observe that this corresponds to the anomalies in the passive data (Figure 7).

XSL•FO
RenderX

The coassociation matrix visualization for date clustering (V3A) shows a clear cluster of dates in the month leading to the relapse for patient X (Figure 8), where we observe similar perceptible matrices for clustering using both AV and CV.

Patient Y

We observe that the 2 relapse periods are represented in the MCA trend visualizations (V1A) generated using LOCF imputation. The 2 periods, October-November 2021 and January-February 2022, show anomalous behavior in all survey sections except sleep (Figure 5). There is an unusually high SD observed in these MCA values in the social section in the month leading to the second relapse (Table 2), indicative of erratic social behavior during that period.

Discussion

Principal Findings

Through this work, we found that offering multiple visualization options (Figures 1 and 2, and Figures S1 and S2 in Multimedia Appendix 1) for comparative analysis was the most productive approach for clinicians to derive inferences. This is evident from the clinicians' feedback on how there is value from each visualization involving active and passive digital-phenotyping data for different use cases of schizophrenia. Throughout the process, we received feedback from software development, research, and clinical teams from Boston and Bangalore, which helped us co-design and rapidly iterate through several prototypes. Finally, as demonstrated in our case examples, we iteratively assessed and improved our tool with real-world data to ensure it offered clinical value. Through this process of technical work, design, feedback, and real-world clinical testing, we demonstrate that multidisciplinary teams can likely adapt and extend software tools as well.

Our software and results support how digital mental health approaches can be adapted and personalized to the clinical needs of local sites. They also illustrate that local solutions from LMICs are often of interest to global users and highlight the true potential of digital mental health tools. With the 2 functional and independent visualization panels in mindLAMPVis, a user can select different options for the same data, draw comparisons, and discuss findings with patients and family members. The tool explicitly allows users to select the imputation method, thus ensuring they know the choice of data preprocessing method. The 2-panel approach can also be easily extended to new conceptualizations of how to visualize existing or new data.

The feasibility of using mindLAMPVis for explorative analysis by the clinicians is the highlight of our case study. For patients X and Y (Results section), we observe a discernible change in their behaviors for a month before the relapse of their symptoms. The active data, which included the reporting of psychological symptoms such as anxiety, mood, psychosis, and sleep, clearly demonstrated this change. The passive geo-location data (GPS) also showed a significant increase in home time during the period preceding the relapse of actual psychiatric symptoms. It is interesting to note confirmatory trends across both the active and passive data streams. These findings highlight that both clinicians and patients can observe and identify patterns of individual digital behaviors and that monitoring can warn them about the early signs of relapse. By not only informing the risk of relapse but also identifying the signals related to that risk for each patient, this approach offers the opportunity to deliver both personalized and preventative care. Data science methods that offer explainability are increasingly valued over black-box models. While they may seem like a simple confirmation of previous findings, the methods used in mindLAMPVis enhance explainability in retrospective analyses. This can be critical in moving the field forward with greater transparency and interpretability.

mindLAMPVis is required for an extensive back-and-forth (ie, nonlinear study) in patient behavior analysis for a mental health illness where the risk of relapse is difficult to identify for even experienced clinicians [7,29]. This can be extended to similar challenging illnesses for which digital phenotyping enables in-depth analysis.

Limitations

But there are still many steps. Technological advances such as mindLAMPVis also raise related questions about training clinicians to better understand the data and how it may fit into their clinical workflow. In response, our team has developed a digital navigator role to support both clinicians and patients, with aspects of the training focused on interpreting data visualizations [30]. The visualization methods in mindLAMPVis might not be immediately intuitive for most clinicians and health care workers without extra support. However, we believe that additional training is simple and will be essential for them to effectively interpret the data, particularly in light of recent advances in data collection within the health care sector. If psychiatry is to make use of novel data streams such as digital phenotyping, new tools such as mindLAMPVis must be paired with new training around how to interpret and use the outputs in care.

Beyond the technical results, our paper highlights how co-design is feasible through adapting existing software tools. While the technical challenges in working with the data required the computer science expertise of our team, the universal nature of coding enabled rapid progress. As interest in digital psychiatry expands, tools that combine data analysis (data mining) and visualization, such as mindLAMPVis, simplify the understanding and interpretation of digital behaviors for both clinicians and patients. This ensures that the work is more accessible and allows for more collaboration from all stakeholders. To date, work on this type of data visualization has been called for, with returning individual results from digital phenotyping highlighted as a critical gap in the field [31]. mindLAMPVis helps to bridge that gap and enable the next generation of more equitable advances.

Conclusions

Through enabling data visualizations, our portal, mindLAMPVis, supplements the interpretation of data available from the mindLAMP platform by clinicians and highlights innovation from LMICs around creating solutions to global health challenges with technology. Our work demonstrates the potential of the visualizations of both active and passive data

XSL•FO
RenderX

collected in digital phenotyping in schizophrenia to inform care and support preventative personalized treatments. Our case studies show how discernible patterns appeared on the visualizations a month before the relapse. The increase in the development and usage of such analytical tools reinforces the universality of the mindLAMP platform and provides new insights into the data.

Acknowledgments

This work was supported by the Wellcome Trust (215843/Z/19/Z). The authors acknowledge the support of their respective organizations as well as the mindLAMP development team. No generative artificial intelligence tools have been used in the preparation of this paper or the supplementary materials.

Data Availability

The patient data used in the mindLAMPVis app are from the Smartphone Health Assessment for Relapse Prevention (SHARP) project [5], which are not available to be shared as per the consent agreement.

Authors' Contributions

KS contributed to the conceptualization of this study, methodology, data analysis, software development, and writing. JS-N worked on the conceptualization of this study, methodology, visualization techniques, and writing. SC and PVR assisted with the background information, case studies, and writing. SN aided with the background information and case studies. AC helped with the background information, case studies, anomaly plots, and writing. UMM carried out the conceptualization of this study, background information, case studies, and writing. JT conducted the conceptualization of this study, background information, case studies, and writing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Related work on visualizations and design study methodology, data processing methods, visualization methods, and the software development specifications of mindLAMPVis.
[DOCX File, 1462 KB - formative_v9i1e70073_app1.docx]

Checklist 1
iCHECK-DH guidelines for reporting the design study of mindLAMPVis. iCHECK-DH: Guidelines and Checklist for the Reporting on Digital Health Implementations.
[DOCX File, 25 KB - formative_v9i1e70073_app2.docx]

References

1. Torous J, Bucci S, Bell IH, et al. The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. World Psychiatry 2021 Oct;20(3):318-335. [doi: 10.1002/wps.20883] [Medline: 34505369]
2. Yoo DW, Birnbaum ML, Van Meter AR, et al. Designing a clinician-facing tool for using insights from patients' social media activity: iterative co-design approach. JMIR Ment Health 2020 Aug 12;7(8):e16969. [doi: 10.2196/16969] [Medline: 32784180]
3. Naderbagi A, Loblay V, Zahed IUM, et al. Cultural and contextual adaptation of digital health interventions: narrative review. J Med Internet Res 2024 Jul 9;26:e55130. [doi: 10.2196/55130] [Medline: 38980719]
4. Rodriguez-Villa E, Mehta UM, Naslund J, et al. Smartphone Health Assessment for Relapse Prevention (SHARP): a digital solution toward global mental health. BJPsych Open 2021 Jan 7;7(1):e29. [doi: 10.1192/bjo.2020.142] [Medline: 33407986]
5. Cohen A, Naslund JA, Chang S, et al. Relapse prediction in schizophrenia with smartphone digital phenotyping during COVID-19: a prospective, three-site, two-country, longitudinal study. Schizophrenia (Heidelb) 2023 Jan 27;9(1):6. [doi: 10.1038/s41537-023-00332-5] [Medline: 36707524]
6. Thornton J, D'Souza R, Tandon R. Artificial intelligence and psychiatry research and practice. Asian J Psychiatr 2023 Mar;81:103509. [doi: 10.1016/j.ajp.2023.103509] [Medline: 36806373]
7. Tandon R, Nasrallah H, Akbarian S, et al. The schizophrenia syndrome, circa 2024: what we know and how that informs its nature. Schizophr Res 2024 Feb;264:1-28. [doi: 10.1016/j.schres.2023.11.015] [Medline: 38086109]
8. Henson P, D'Mello R, Vaidyam A, Keshavan M, Torous J. Anomaly detection to predict relapse risk in schizophrenia. Transl Psychiatry 2021 Jan 11;11(1):28. [doi: 10.1038/s41398-020-01123-7] [Medline: 33431818]

XSL•FO
RenderX

9. Torous J, Blease C. Return of results in digital phenotyping: ethical considerations for real-world use cases. Am J Bioeth 2024 Feb;24(2):91-93. [doi: 10.1080/15265161.2024.2298146] [Medline: 38295252]
10. Scheuer L, Torous J. Usable data visualization for digital biomarkers: an analysis of usability, data sharing, and clinician contact. Digit Biomark 2022;6(3):98-106. [doi: 10.1159/000525888] [Medline: 36471766]
11. Ko K, Lee CW, Nam S, et al. Epidemiological characterization of a directed and weighted disease network using data from a cohort of one million patients: network analysis. J Med Internet Res 2020 Apr 9;22(4):e15196. [doi: 10.2196/15196] [Medline: 32271154]
12. Li R, Yin C, Yang S, Qian B, Zhang P. Marrying medical domain knowledge with deep learning on electronic health records: a deep visual analytics approach. J Med Internet Res 2020 Sep 28;22(9):e20645. [doi: 10.2196/20645] [Medline: 32985996]
13. Ezera A. Boost engagement with interactive data visualizations. Infogram. 2022. URL: https://infogram.com/blog/interactive-data-visualization-boost-engagement/ [accessed 2025-05-29]
14. Chang S, Gray L, Alon N, Torous J. Patient and clinician experiences with sharing data visualizations integrated into mental health treatment. Soc Sci (Basel) 2023;12(12):648. [doi: 10.3390/socsci12120648]
15. Byanjankar P, Poudyal A, Kohrt BA, Maharjan SM, Hagaman A, van Heerden A. Utilizing passive sensing data to provide personalized psychological care in low-resource settings. Gates Open Res 2020;4:118. [doi: 10.12688/gatesopenres.13117.2] [Medline: 33709058]
16. Lee B, Lee J, Cho Y, et al. Visualisation of information using patient journey maps for a mobile health application. Appl Sci (Basel) 2023;13(10):6067. [doi: 10.3390/app13106067]
17. Concannon D, Herbst K, Manley E. Developing a data dashboard framework for population health surveillance: widening access to clinical trial findings. JMIR Form Res 2019 Apr 4;3(2):e11342. [doi: 10.2196/11342] [Medline: 30946016]
18. Simblett S, Dawe-Lane E, Gilpin G, et al. Data visualization preferences in remote measurement technology for individuals living with depression, epilepsy, and multiple sclerosis: qualitative study. J Med Internet Res 2024 Oct 18;26:e43954. [doi: 10.2196/43954] [Medline: 39423366]
19. Mansoor H, Gerych W, Alajaji A, et al. INPHOVIS: interactive visual analytics for smartphone-based digital phenotyping. Visual Inf 2023 Jun;7(2):13-29. [doi: 10.1016/j.visinf.2023.01.002]
20. Langener AM, Stulp G, Jacobson NC, et al. It's all about timing: exploring different temporal resolutions for analyzing digital-phenotyping data. Adv Methods Pract Psychol Sci 2024;7(1). [doi: 10.1177/25152459231202677]
21. Lee K, Lee TC, Yefimova M, et al. Using digital phenotyping to understand health-related outcomes: a scoping review. Int J Med Inform 2023 Jun;174:105061. [doi: 10.1016/j.ijmedinf.2023.105061] [Medline: 37030145]
22. Schmidt S, D'Alfonso S. Clinician perspectives on how digital phenotyping can inform client treatment. Acta Psychol (Amst) 2023 May;235:103886. [doi: 10.1016/j.actpsy.2023.103886] [Medline: 36921359]
23. Sama K, Sreevalsan-Nair J, Choudhary S, et al. mindLAMPVis: data visualization portal. 2024. URL: https://mindlampvis.azurewebsites.net [accessed 2025-05-23]
24. Sama K, Sreevalsan-Nair J, Choudhary S, et al. GVCL/mindLAMPVisDev. GitHub. 2025. URL: https://github.com/GVCL/mindLAMPVisDev/ [accessed 2025-05-29]
25. Sedlmair M, Meyer M, Munzner T. Design study methodology: reflections from the trenches and the stacks. IEEE Trans Visual Comput Graphics 2012;18(12):2431-2440. [doi: 10.1109/TVCG.2012.213]
26. Rodriguez-Villa E, Rozatkar AR, Kumar M, et al. Cross cultural and global uses of a digital mental health app: results of focus groups with clinicians, patients and family members in India and the United States. Glob Ment Health (Camb) 2021;8:e30. [doi: 10.1017/gmh.2021.28] [Medline: 34512999]
27. van Buuren S, Oudshoorn K. Flexible Multivariate Imputation by MICE: TNO; 1999.
28. Perrin Franck C, Babington-Ashaye A, Dietrich D, et al. iCHECK-DH: guidelines and checklist for the reporting on digital health implementations. J Med Internet Res 2023 May 10;25:e46694 [FREE Full text] [doi: 10.2196/46694] [Medline: 37163336]
29. Emsley R, Chiliza B, Asmal L, Harvey BH. The nature of relapse in schizophrenia. BMC Psychiatry 2013 Feb 8;13:1-8. [doi: 10.1186/1471-244X-13-50] [Medline: 23394123]
30. Shen FX, Baum ML, Martinez-Martin N, et al. Returning individual research results from digital phenotyping in psychiatry. Am J Bioeth 2024 Feb;24(2):69-90. [doi: 10.1080/15265161.2023.2180109] [Medline: 37155651]
31. Perret S, Alon N, Carpenter-Song E, et al. Standardising the role of a digital navigator in behavioural health: a systematic review. Lancet Digit Health 2023 Dec;5(12):e925-e932. [doi: 10.1016/S2589-7500(23)00152-8] [Medline: 38000876]

Abbreviations

AV: aggregate vector
CV: complete vector
DSM: design study methodology
LMIC: low- and middle-income country
LOCF: last observation carried forward

XSL•FO
RenderX

MCA: multiple correspondence analysis
MICE: Multiple Imputation by Chained Equation
SHARP: Smartphone Health Assessment for Relapse Prevention

Edited by A Mavragani; submitted 14.12.24; peer-reviewed by B Buck, J Ainsworth; revised version received 28.04.25; accepted 05.05.25; published 10.06.25.

Please cite as:
Sama K, Sreevalsan-Nair J, Choudhary S, Nagendra S, Reddy PV, Cohen A, Mehta UM, Torous J
mindLAMPVis as a Co-Designed Clinician-Facing Data Visualization Portal to Integrate Clinical Observations From Digital Phenotyping in Schizophrenia: User-Centered Design Process and Pilot Implementation
JMIR Form Res 2025;9:e70073
URL: https://formative.jmir.org/2025/1/e70073
doi:10.2196/70073

© Karthik Sama, Jaya Sreevalsan-Nair, Soumya Choudhary, Srilakshmi Nagendra, Preethi V Reddy, Asher Cohen, Urvakhsh Meherwan Mehta, John Torous. Originally published in JMIR Formative Research (https://formative.jmir.org), 10.6.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Using Real Electronic Health Records in Undergraduate Education: Roundtable Discussion

Fatima Nadeem[1], MBChB, MRES; Jessica Azmy[1], MBChB, MSc; Asieh Yousefnejad Shomali[1], PhD; Benjamin Diette[2], MA; Lloyd J Gregory[1], PhD; Angela C Davies[1], PhD; Kurt C Wilson[1], MD, MRCGP

[1]Faculty of Biology, Medicine and Health, University of Manchester, Oxford Road, Manchester, United Kingdom

[2]Health Innovation Manchester, Manchester, United Kingdom

Corresponding Author:
Fatima Nadeem, MBChB, MRES
Faculty of Biology, Medicine and Health, University of Manchester, Oxford Road, Manchester, United Kingdom

Abstract

Background: Simulated electronic health records (EHRs) are used in structured teaching for health care students. This partly addresses inconsistent student exposure to EHRs while on clinical placements. However, simulated records are poor replacements for the complexity of data encountered in real EHRs. While routinely collected health care data are often used for research, secondary use does not include education. We are exploring the perceptions, governance, and ethics required to support the use of real patient records within teaching.

Objective: The aim of the study is to explore the perspectives of health care professionals regarding the use of real patient records to deliver interprofessional EHR education to undergraduate health care students.

Methods: We held 90-minute group discussions with 10 health care professionals from nursing, pharmacy, medicine, and allied health disciplines. We used the GRIPP2 (Guidance for Reporting Involvement of Patients and the Public 2) checklist for reporting Patient and Public Involvement and Engagement to present our reflections.

Results: There was consensus on the need to upskill health care students in the use of EHRs. Participants emphasized teaching general EHR competencies and transferable skills to overcome the diversity in EHR systems. They highlighted limitations in current teaching due to accessibility issues, disparities within clinical teaching, and curricular gaps on important topics such as clinical documentation and coding. Highlighted benefits of using real EHRs in teaching included learning from the complexities and inaccuracies of real patient data, grasping real-world time frames, and better appreciation of multidisciplinary interactions. Concerns included exposing individual clinicians to unfounded scrutiny and the potential consequences of incidental findings within EHRs. The ethical implications of overlooking perceived errors within EHRs versus the impracticality of acting on them were discussed. To mitigate concerns, it was suggested that data donors would provide informed consent ensuring they understand that they will not be recontacted should any such errors be found.

Conclusions: Innovative solutions are needed to realign health care education with clinical practice in rapidly evolving digital environments. Real patient records are optimal for teaching students to handle complex and abundant real-world data. Data within EHRs represent a wealth of clinical knowledge encompassing professional and personal experiences spanning the lifetimes of patients and their caregivers. Drawing experiences and events from real EHRs will prepare health care students to anticipate, confront, and manage real patients in a variety of real-life scenarios. Our reflections highlight the processes and safeguards to consider when using real patient records to deliver EHR education to health care students. These detailed reflections from discussions with health care professionals provide the grounds for a robust framework, with appropriate governance and consent in place to use real health data in training to support preparation for clinical practice.

(JMIR Form Res 2025;9:e60789) doi:10.2196/60789

KEYWORDS

electronic health records; clinical competence; medical records systems; health care education; interdisciplinary education; undergraduate; student; health care professional; medical education; clinical practice; patient records; digital health; health informatics

XSL•FO
RenderX

Introduction

Background

Electronic health records (EHRs) are replacing paper medical notes in mature health care systems across the world [1]. EHRs are used in almost all interactions with patients in everyday clinical practice. Poor use of EHRs is associated with clinician burnout and adverse clinical consequences, highlighting the importance of high-quality EHR education to prepare health care professionals for the digital clinical environment [2,3].

Traditionally, clinical skills have remained the focus of undergraduate health care education [4]. Students are taught to take a history and examine and formulate differential diagnoses without considering the wealth of information present in an EHR [4]. In clinical practice, clinicians use EHRs as a central part of clinical decision-making to find patient information and gather data prior to seeing a patient [4]. The skills needed to do this should be taught in undergraduate curricula, and competencies have been published related to safe and effective EHR use [5,6]. However, health care students scarcely receive structured teaching to develop EHR-related skills; instead, they have variable exposure to electronic records during clinical rotations [7-11].

Where teaching is provided in classroom-based settings during undergraduate health care courses, fictional or "dummy" EHRs are often created for use [12]. These fictional records are reported to lack the complexity of real patient data and the depth of a real EHR. Whereas real EHRs capture the multidisciplinary nature of documentation, errors and coding this may be absent in fictional records, raising the possibility that real patient records could be of greater educational value in classroom-based teaching settings. The evolving role of EHRs, with the introduction of open notes in the United Kingdom meaning that patients can read clinician entries, adds complexity to EHR teaching [13,14]. Students need to be prepared for this change and become confident in working with the complexity of real patient data in this context. The use of real EHRs for teaching may however raise ethical and practical questions that need consideration.

Objectives

In this paper, we explore what health care professionals, regularly using EHRs as part of their clinical work, think about the use of real EHRs for structured teaching purposes in classroom settings. Could the use of real records be a more valuable learning opportunity for undergraduate health care students in the United Kingdom?

Health Innovation Manchester is supporting our work across our 4 universities in the Greater Manchester region. Discussions were held with health care professionals to collect opinions regarding the use of real EHRs in delivering interprofessional EHR education to undergraduate health care students. These were used to gather participants' own experiences of EHR education, their views on current EHR education, and the perceived benefits and challenges of using real EHRs in structured undergraduate teaching sessions. We present reflections from these discussions in this paper in order to inform and support the use of donated real patient records within undergraduate teaching.

Methods

Design

Two discussion sessions were hosted remotely via Microsoft Teams and audio-recorded. Both lasted for 90 minutes, beginning with a presentation from the moderator to explain the background and rationale for the proposal, followed by an hour-long facilitated discussion.

Participants were first contacted via word of mouth or through emails from colleagues who acted as gatekeepers. The only inclusion criterion was being a health care professional with a background in nursing, medicine, pharmacy, or an allied health role. The Participant Information Sheet and consent form were then emailed to individuals who expressed an interest in participating and met this criterion. Those who returned the signed consent form were invited to attend the discussions. In total, 10 participants were present across 2 multiprofessional discussion groups: 4 nurses, 3 doctors, 2 pharmacists, and 1 allied health professional (dietitian).

After considering availability and convenience for participants, the first group comprised 2 doctors, a nurse, and a dietitian, while the remaining participants were included in the second group. In total, 9 participants expressed the experience of supervising undergraduates or held formal teaching roles in health care courses. The moderator was a doctor with extensive experience in using EHRs in clinical practice; for the duration of her Academic Clinical Fellowship, she further studied literature on EHR education and collaborated with academics internally and across institutions to develop and improve teaching on EHRs for undergraduates. To gain a broad understanding of participants' views, the moderator asked 3 open questions to prompt a discussion: (1) What are your thoughts on using real EHRs in education? (2) What are the benefits? (3) What are the drawbacks? The moderator listened to the audio recordings and summarized the points raised in this paper (Multimedia Appendix 1). Please note, this is not research; hence, a formal qualitative analysis of the discussions was not undertaken. Generative artificial intelligence was not used to write this paper. The GRIPP2 (Guidance for Reporting Involvement of Patients and the Public 2) checklist for reporting Patient and Public Involvement and Engagement was used to present reflections from the roundtable discussions (Checklist 1).

Ethical Considerations

After reviewing the University of Manchester's ethics requirements and further consulting with colleagues from the research ethics department, it was confirmed that this project did not require ethical review, as it fell under the remit of "Working With Professionals" rather than a research activity. However, as advised by the ethics department, all participants were provided with Participant Information Sheets and returned signed consent forms in order to partake in the discussions and provide permission for the use of direct quotes. Participation was voluntary and unpaid.

XSL•FO
RenderX

Results

Participants' Past Experiences and Current Practices in EHR Education

Participants shared personal experiences of learning to use EHRs when they were students and discussed recent trends in EHR teaching as undergraduate tutors.

Past Experiences in EHR Education

Participants had no or limited experience of EHRs during their undergraduate training. Participant 10 reported using paper notes when he was completing medical training. He gained skills in EHR use in a self-directed and unstructured fashion, never having received guidance or education:

> *... there's no structure to it. However, there is probably a little bit of on the job guidance of right, this is where you find this ...* [Participant 10, general practitioner]

Others shared experiences of gaining skills in the use of live EHR systems through experiential learning on clinical placements. Participant 9, for instance, recalled a helpful simulation session as a medical student in which he used an EHR system:

> *I was thinking back to when I did my medical degree, which was quite a while ago now ... we paired up and they had to see what it was like to try and do a 10-minute consultation where you were simultaneously typing and coding stuff. So, we all got, you know like (a) test login. So, I did find that a useful experience ... So I think we don't teach them enough, but I had some positive experiences as well.* [Participant 9, general practitioner]

Current Practices in EHR Education Described by Participants

Participants reported that students gained exposure to EHRs through a variety of resources and settings, with some receiving logins for live EHR systems on clinical placements. Participants also pointed out that fellow clinicians or IT staff were often tasked to show students how to navigate EHRs, most often through a basic orientation and using fictional records. Participants felt that this was useful to introduce students to systems without overloading them. Some limitations were also highlighted:

> *... but obviously it's not useful at all in terms of clinical ... so yeah, I can understand how that's quite limiting in terms of not being able to get any kind of clinical information, experience, exposure from a dummy patient.* [Participant 2, pharmacist]

A distinction was made between "software training" consisting of a brief tour of an EHR system and teaching higher-level skills associated with EHR use such as generating high-quality data and critiquing documentation [4,5]. Participants felt that fictional records were not deemed to be useful for clinical learning or for developing higher-level skills required for the competent use of EHRs. Areas not addressed in the current EHR teaching were highlighted and included teaching on how to document and code. A delay in record access was also voiced; students failed to receive logins at the beginning of their placements.

A nurse participant described teaching her students using screenshots of real records. However, she highlighted the constraints of this approach; students could not view peripheral EHR systems, which contain information about other parts of the patient journey, for example, letters accessible on a different platform.

Challenges in EHR Education

Absence of National Standards

Participants made recurrent observations that health care students are not taught how to document or how to interpret documentation. They felt that part of the challenge was the lack of national guidance on this. Additionally, individual clinicians document in different ways, with further interprofessional differences in writing styles.

Changing Function of Medical Records

Participants identified the evolving function of EHRs as another barrier to teaching. Earlier functions were to document clinical care to follow a patient's journey and for medicolegal reasons. However, the purpose of EHRs has since transformed to allow patients to view and engage with their records, but this adapting role of EHRs is not taught to students.

> *... about two weeks ago in primary care, the function of medical records changed completely because all patients have access to them. So effectively someone overnight has changed the function of medical records instantly.* [Participant 3, general practitioner]

Participants commented that with the introduction of open notes, patients can now access and read free-text entries in their records. They felt that currently this created ambiguity on how best to document. Participant 3 commented that he was now writing out "essays" in the record to ensure its readability to patients. Participants highlighted differences in patient access to EHRs across the National Health Service (NHS); for example, patients may not be able to see their notes in secondary care.

Diversity in EHR Systems

Participants frequently cited diversity in EHR systems as a barrier to education; students may end up learning to navigate a system but not using it in clinical practice. One participant felt that unless there is a unified EHR provider across the NHS, EHR teaching would be futile. She shared her own disorienting experience of using different systems on clinical rotations.

> *... but with the software systems, it was almost like the first day of school every time I moved ... we need to have parity across the patch and using, you know either one or two or if not the same system.* [Participant 7, nurse]

One participant responded that the current policy is to let markets decide on EHR providers in the NHS, and a unified system creates greater vulnerability to cyberattacks, adding that accelerations in technology mean that EHR systems constantly

XSL•FO
RenderX

change. Participants then agreed that teaching should focus on transferable skills, applied across EHR systems.

> *So, I've used Emis for about 8 or 9 years and I've recently gone to SystemOne, using that regularly. So, the way I assess for a patient is still transferable. Yes, obviously finding things is different and you know all the nitty gritty and the way it looks and stuff like that. But yeah, so the key point is more the skills of using it and what you get out of it* ... [Participant 10, general practitioner]

All participants acknowledged the value of learning fundamental EHR skills such as documenting in records, accessing data, and using EHRs to communicate within the multidisciplinary team (MDT). This would not replace software training that introduces students to a particular EHR system at the beginning of a placement. Developing key EHR skills would help students to know what they could find, and software training would show them where to retrieve it.

Interprofessional Differences

EHR teaching can be delivered to interprofessional cohorts to simulate the workplace and develop interprofessional competencies. Participants recognized the challenge of delivering teaching to an interprofessional student cohort due to different documentation styles; participants commented that advanced nurse practitioners document differently to doctors, as an example. Participants suggested that the delivery of unified teaching would require agreement between professions on core EHR documentation skills. It was felt that if a basic standard on documentation was taught, all students would understand patient notes regardless of their role. One participant explained how EHRs can facilitate interprofessional learning, communication, and understanding of different professional roles.

> *... I think that one of the biggest opportunities that is missed whenever a student comes to a placement, [...] trying to be part of MDT or even in an informal way, what came out in the ward round this morning, what's the status of the patient's discharge? As a dietitian, I would not be aware of those things when I was on my placements. Even though I could hear dietitians talking about it, I wouldn't have really known how to find that piece of information and why it was important to me* ... [Participant 8, dietitian]

Overall, participants felt teaching should address the differing requirements of specific professions. However, there should also be opportunities for interprofessional EHR learning to prepare students for real clinical practice.

Important Considerations for EHR Education

Minimum Standards

One participant commented that EHR education should establish a minimum standard of teaching, with additional clinical exposure to EHRs as a bonus:

> *... so we need to make sure that we are teaching a minimum level and then everything that happens in the (hospital) bases or on placement elsewhere then becomes that extra bit. So [...] I think what we really want to be looking at is to try and find what that minimum standard is and see if we can define that to teach that as a baseline.* [Participant 4, pharmacist]

He noted that other aspects of clinical learning such as the way we assess students are universal, but that exposure to EHRs depends on clinical placements, an area that needs addressing for parity.

Clinical Context

The facilitator suggested that initial EHR teaching would use a handful of real records, acknowledging that different clinical settings produce different records; for example, learning from primary care records would be different to that from secondary care records. Participants were asked for their views on using a small number of real primary care records. Participants agreed that holistic EHR education should reflect the varying clinical contexts of records. They noted that primary care records do contain some data from secondary care such as discharge summaries or bloods. One participant mentioned that it would be helpful to include these in teaching to give students a complete perspective of the patient journey.

Integration Into Existing Curricula

Participants advocated for the integration of EHRs into existing health care curricula as the most powerful to build student value of EHR teaching, avoiding repetition and duplication of teaching.

Using Real Patient Records in EHR Education

Learning Opportunities From Using Real Patient Records in EHR Education

Participants highlighted multiple learning opportunities when using real patient records in undergraduate teaching. One participant identified the opportunity to learn from the longitudinal care of patients when using real records; students can view a patient's journey in a short time. Another participant noted the benefits of learning from real-world time frames, appreciating processing times for results and addressing urgent versus routine tasks. Similarly, another participant noted that learning from real records can encourage students to critique data and learn about mistakes:

> *I don't know if you do much learning in medical school from significant events, (be)cause I think that's quite a useful way to learn about what can go wrong and why it goes wrong, which you know you can look at it from the point of view of records. You can look at cases where things have gone wrong because of the records.* [Participant 3, general practitioner]

Facilitators also elaborated on the value of learning from real errors; teaching could explore the potential clinical consequences of mistakes within EHRs. However, they suggested that there is often too much emphasis on things going wrong, which causes scaremongering and negativity. Instead, students could be aware of the potential for harm while appreciating that the record is a beneficial tool intended to aid clinicians.

XSL•FO
RenderX

Participants highlighted that another benefit of using authentic EHRs in teaching was learning from the complexity within the records. One nurse illustrated the invaluable opportunity of learning about holistic patient care and the real-life complexities of medicine through student interaction and learning from the complexities within real electronic systems. Conversely, another participant (pharmacist) commented that there may be "missing" information in real records, which can hinder learning. A facilitator responded that although this is not ideal for teaching, it may prepare students for real-world clinical practice where letters or data within a system may be incomplete.

Finally, a participant highlighted that entries in real records can demonstrate how the MDT communicates collaboratively to facilitate patient care:

> *…the multidisciplinary aspect of how everybody has integrated into that … care and what different people are doing and where they might fit into that care as well is really important. And then the communication aspect as well…. in terms of how you write up the record and how you communicate that either to a patient or things like referral notes, all of that stuff that gets put onto the system.* [Participant 4, pharmacist]

Concerns Surrounding the Use of Real Patient Records in EHR Education

Participants expressed their concerns about using real records in EHR education. This included exposing students to the entirety of authentic patient records at once, which could be overwhelming. Facilitators responded that this reflects the nature of real clinical practice. However, facilitators agreed that EHR learning should be progressive like learning clinical skills.

Participants also held concerns about potential litigation or reputational damage for clinicians:

> *… as an accountable practitioner, there's always that fear of litigation. So, the idea of, perhaps my work, which I know that I've done to the best of my knowledge, and I've done as safely as possible. The idea of that being scrutinised by several people and I perhaps cannot explain what's written in that manner there and then, that does not personally sit comfortably with me because you're almost not able to explain your rationale for doing so.* [Participant 7, nurse]

Participant 10 (general practitioner) commented that documentation may be brief due to time pressures or human factors, and on these occasions, clinical entries are unrepresentative of usual practice. To this, another general practitioner remarked that students can already read clinical entries on placements; thus, using entries in didactic teaching is not vastly different. However, he did concede that it may expose clinicians to critique from larger student groups. A senior lecturer and pharmacist also explained that there are opportunities for medical students at his institution to raise concerns about clinician entries or other matters in a professional manner.

Despite the earlier-mentioned concerns, some participants held no objections to their names appearing on EHRs for teaching, viewing it as an opportunity to gain constructive feedback. A participant noted that this chimed with existing processes, where clinicians may access past entries by colleagues when a patient transitions between general practitioner surgeries.

Facilitators emphasized the importance of deidentifying EHRs to safeguard patient confidentiality. Participants raised concerns about the potential discovery of significant incidental findings in deidentified records, posing ethical dilemmas if patients cannot be identified. Many participants advocated for the tracing of entries back to patients to address errors in their records, emphasizing a duty of care toward patients. However, a facilitator explained that those entries in records often present nuanced situations rather than clear-cut scenarios. Without direct involvement in patient care, interpreting whether harm occurred due to reading record entries was deemed subjective. The facilitator cautioned against overreacting to perceived errors, suggesting that acting on such errors could lead to false alarms, as most patients are managed safely. Participant 3 expressed disagreement with this perspective.

> *So, I think most things won't be black and white, but I think there are cases where it is black and white, and you probably won't find them in the records. But I think I could, I could give you a case scenario where if I told you it, you would want to be able to de-anonymize the notes. I could- you know give me 5 minutes and I'll generate one for you because I think you know there are issues that could occur where you- if you found it, you would feel liable to inform someone.* [Participant 3, general practitioner]

Participant 3 felt that certain events were highly likely to recur, with significant potential consequences. He recalled an anecdote of a missed blood result in a pregnant female that led to a negative outcome. Many participants expressed that this exemplified a powerful argument for deanonymizing EHRs if potential errors were found.

One participant suggested discussing the possibility of incidental findings with patients. She suggested that patients donating their records for teaching should be informed that if potential errors be found within records, these would not be enacted upon due to the aforementioned limitations.

Finally, participants expressed concerns about unethical practices by students such as inappropriate retrieval or use of real data. Facilitators responded that safeguards would be implemented to prevent this, but as in real life, there may be missed instances. Ultimately, participants concluded that we are preparing students for exactly this scenario—to handle real data! They also suggested that students should receive teaching on professional standards when handling digital information such as those outlined by the Health and Care Professions Council [15].

XSL•FO
RenderX

Discussion

Principal Findings

We are collaborating with Health Innovation Manchester and 3 other higher education institutions to create and implement a framework for a donation of real patient records for undergraduate education. These roundtable discussions with health care professionals brought important issues to the forefront to guide this process. We will use salient points raised in these reflections to inform data governance processes and anticipate and address potential benefits, risks, and roadblocks to consider when using real patient records for undergraduate education.

Whereas currently, real EHRs are not used for teaching in classroom environments, the benefits of using authentic records were discussed in detail. Participants explained that by using real records, students can learn about clinical decision-making and the causes and consequences of mistakes made in real contexts. Furthermore, students will learn to cope with the imperfection and "noise" within real EHRs; this includes the variety and complexity of data within authentic records but also its inaccuracies. Real EHRs also demonstrate team dynamics and multiprofessional communication through entries and contributions from colleagues. Challenges of using real EHRs in education were highlighted, including overwhelming students with vast volumes of data. By far, the main point of contention was the ethical and practical implications of incidental findings within records. Arguments for deidentifying records and informing patients in the case of clinically significant errors were presented, and contrarily, there were discussions about the subjective nature of entries and the potential for false alarms. Overall, it was suggested that in either scenario, patient donors should be informed of the process for handling incidental findings at the time of consent. Another challenge of using real records was compromising the privacy of clinicians; there were concerns about the risk of reputational damage and the potential for litigation.

It is clear that EHRs have an expanding audience in health care systems across the world [1]. The Regenstrief Institute in the United States previously incorporated a repository of pseudonymized authentic patient records for use by higher education institutes to teach clinical and digital competencies. To the best of our knowledge, a similar resource is not available for didactic teaching in the United Kingdom. In the United Kingdom, primary care data may be used for health care research on the premise of presumed consent, but the same principle has not been explored regarding the use of data in education [16].

Patients in the United Kingdom have recently gained default access to EHRs. This has caused considerable anxiety among professionals who have expressed uncertainty on how to document for both patients and clinicians and was corroborated by a participant in our discussion [17]. Clearly, the changing purposes and audiences of EHRs require new skills and nuanced teaching on documentation; health professionals may not be best placed to provide this. We as health care educators propose that honest discussions and reflections with students on open notes are important while central guidance is awaited.

Our discussion group with health care professionals revealed areas of agreement and areas of tension and concern among participants regarding the use of real data for EHR education. The principal area of ethical concern and debate was whether a process of deidentification of patient records should be undertaken when errors are found within records. The duty of candor outlines that patients should be informed of medical errors, but this is fraught with complexity in this particular educational context.

Participants voiced fears of scrutiny of individual entries, which could lead to litigation or reputational damage for clinicians. Participants cited the lack of opportunity to explain the context of their judgment when records are read and interpreted by people not directly involved in the patient's care. However, it was noted that students already have access to EHRs on placements; hence, this is no different to accessing EHRs in classroom teaching, albeit the latter may involve detailed scrutiny of individual entries. Despite this concern, not all participants felt that clinician identities should be protected, as this would prevent an opportunity for constructive feedback on their practice. Realistically, if a large repository of EHRs is created, it would be impractical for us to provide feedback to all clinicians who have contributed to the records, and this would distract from the purpose of the resource.

Determining the threshold at which records would be deidentified is one problem, aside from the practicalities and logistics involved. Having clear processes and a consensus on dealing with these issues early on is paramount. It is imperative to obtain explicit consent from patients, and permissions from clinicians, and data handlers for accessing and using the data, including addressing the issue of incidental findings. Ensuring that students are adequately trained and familiar with the professional standards for handling real data is another key component.

Participants also highlighted the benefits of learning from mistakes made in real records, an opportunity that is lacking in synthetic records used for teaching. Identifying and managing errors is an important skill for future health care workers. Participants emphasized that synthetic records are simplistic compared to real EHRs, as real records confer the advantage of revealing the intricacies of health care and can demonstrate the complexity and volume of data encountered in real clinical practice.

Participants expressed that real records were complex and difficult to navigate, but interestingly, this was viewed as an advantage. Having the ability to navigate through the various components of a real record, appreciating the wealth of information available and the disconnect in communication, across health care settings was deemed as a powerful learning tool. Having a "snapshot" of a record, whether real or synthetic, lacks this benefit. Furthermore, students can learn about time frames and prioritization from the time intervals captured in real records. Finally, real records allow easy access to the multitude of entries by different professions, allowing students to learn from and work within an MDT. Overall, there was consensus that authentic EHRs are a better resource to prepare students for real clinical practice.

XSL•FO
RenderX

Our participants voiced a recurrent concern that EHR teaching lost value due to variability in EHR platforms used across health care systems. However, participants distinguished providing "software training" that teaches students to use specific EHR platforms from EHR education, focusing on underlying digital principles and transferable skills [4,5]. Participants highlighted that different clinical settings use diverse EHRs; educational goals and EHRs need to align for useful and transferable clinical teaching. Having access to records from a variety of settings would allow students to revisit different points of a patient's journey and understand the holistic care delivered by the MDT.

Integrating EHRs into current curricula could create further opportunities to develop clinical reasoning and critical thinking skills, for example, by critique of trends such as blood test results over time. It could also encourage students to perceive its value in clinical work as well as avoid duplication in teaching. EHRs may also be used in examinations and assessments; if used in this way, EHRs should not appear as a tick-box exercise, as this risks undermining the value of learning EHR skills.

Interprofessional teaching is paramount to emulate real-world experiences and practice. Unsurprisingly, participants expressed the challenging nature of interprofessional teaching on EHRs. Each professional cohort uses EHRs differently, and consequently, there is variation in the data entered, documentation styles, as well as the nature of information accessed and relevant to each clinical role. Participants recognized the need for an agreement on basic EHR principles and the benefits this brings to an MDT. Students need to learn to use the EHR as a key tool for communication across disciplines.

Participants agreed that the absence of national guidance on documentation in EHRs makes it difficult to teach this important skill. Personal writing styles mean that the same encounter can be captured variably, making it difficult to define an individual style of documentation as "best practice." While health care professionals rely on protocols for other aspects of clinical work, there is an absence of such a blueprint for documentation. Standard principles of documentation traditionally taught, such as ensuring legibility, dating, and signing handwritten entries, are also redundant in EHRs.

Participants alluded to coding during our discussions. Coding via Systematized Nomenclature of Medicine Clinical Terms is a crucial component of EHRs. It ensures that interoperability and its practical implementation have been expanding, helping to introduce clinical decision support systems over recent years [18-20]. Systematized Nomenclature of Medicine coding has been highlighted as a tool for improving the quality of data entered in electronic records [18]. Although further research is needed on its effect on patient outcomes, it is a core component of digital health systems, and a basic understanding of its principles is needed by clinicians. To our knowledge, there has been no reported assessment of Systematized Nomenclature of Medicine Clinical Terms coverage in undergraduate curricula in recent years; in view of our discussions with health care professionals, we anticipate health care curricula to be deficient in this aspect, and this area of EHR practice was not discussed in detail by participants taking part in our discussion.

Limitations

We captured a variety of opinions that have helped outline the benefits and concerns around the use of real EHRs in classroom-based teaching. However, our findings represent the views from a small number of health care professionals within the United Kingdom, and professionals elsewhere may have differing views and experiences. Our work was confined to the collection of the views of health care professionals, and future work could explore the views of students and patients on how EHR teaching may be delivered.

Conclusions

Modern clinical practice demands digital competencies alongside traditional clinical skills. Our work with health care professionals provides new insights into the potential role of real patient records in modern health care education. Unsurprisingly, participants expressed limited exposure to formal EHR teaching during their own undergraduate training, in part due to the relatively recent digitization of the NHS. Discussion of current educational trends revealed a variety of opportunities for undergraduates to develop EHR-related skills including exposure to EHRs on placements, teaching using screenshots of real records, and use of dummy records for IT-related software training. The exposure to the entirety of authentic patient records is acknowledged as reflective of real-life practice. Concerns that real EHRs may contain incidental findings, which are interpreted as harmful or leading to error, may be mitigated by informing data donors that incidental findings will not be acted upon, by teaching students professional standards for handling digital information, and by recognizing that EHR data are an interpretation of care but do not capture every occurrence or nuance in clinical encounters.

Overall, students will need a strong clinical foundation before learning to extract and use the information presented to them in real EHRs. Hence, EHR skills should be taught in a step-wise manner as with other clinical teaching. Real records can be used flexibly or adapted in a similar way to synthetic records for teaching; for example, in the early years, students may benefit from looking at 1 aspect of the record, whereas more experienced students may be exposed to a range of entries and expected to interpret these accordingly. If used appropriately and tailored to the learning needs of students, there is no reason for the complexities within real records to hinder teaching.

> *This [the record] is the practice of medicine. It's intertwined with it. It determines what you do in the long run. You're a victim of it or you're a triumph because of it. The human mind simply cannot carry all the information about all the patients in a practice without error. And so the record becomes part of your practice.* [Larry Weed]

The above quote by Larry Weed, creator of the Subjective, Objective, Assessment and Plan style of documentation, emphasizes the crucial influence of EHRs on clinical practice. It is vital that our students are taught the skills required for effective EHR use in an age where the creation and manipulation

XSL•FO
RenderX

of digital health data are ever-increasing. Using authentic data to teach health care students will bring education abreast of clinical practice. Our reflections on discussions with health care professionals outline important considerations for the responsible and effective use of real data in undergraduate teaching. We hope that this will provide a foundation for a framework that confronts and mitigates the issues that may arise when using authentic data for education, especially in the context of data presented within EHRs.

Acknowledgments

The authors thank Health Innovation Manchester and the 4 Greater Manchester universities: the University of Bolton, the University of Manchester, the Manchester Metropolitan University, and the University of Salford for lending their support and funding toward this work. Financial support was provided as an Independent Medical Education Grant from Pfizer Limited.

Data Availability

The datasets generated or analyzed during this study are not publicly available, as audio recordings were deleted following analysis to protect participants' privacy as outlined in the Participant Information Sheet and the consent process. However, the raw field notes taken from the recordings are provided in Multimedia Appendix 1.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary data field containing raw field notes of roundtable discussions.
[DOCX File, 27 KB - formative_v9i1e60789_app1.docx]

Checklist 1
GRIPP2 (Guidance for Reporting Involvement of Patients and the Public 2) reporting checklist.
[DOCX File, 17 KB - formative_v9i1e60789_app2.docx]

References

1. Slawomirski L, Lindner L, Kd B, Haywood P, Hashiguchi TCO, Steentjes M, et al. Progress on implementing and using electronic health record systems. In: Developments in OECD Countries as of 2021 OECD Health Working Papers: OECD Publishing; 2023. [doi: 10.1787/4f4ce846-en]
2. Muhiyaddin R, Elfadl A, Mohamed E, et al. Electronic health records and physician burnout: a scoping review. Stud Health Technol Inform 2022 Jan 14;289:481-484. [doi: 10.3233/SHTI210962] [Medline: 35062195]
3. Dixit RA, Boxley CL, Samuel S, Mohan V, Ratwani RM, Gold JA. Electronic health record use issues and diagnostic error: a scoping review and framework. J Patient Saf 2023 Jan 1;19(1):e25-e30. [doi: 10.1097/PTS.0000000000001081] [Medline: 36538341]
4. McMillan B, Davidge G, Nadeem F, Dowding D, Wilson K, Davies A. Navigating the electronic health record in university education: helping health care professionals of the future prepare for 21st century practice. BMJ Health Care Inform 2023 Mar;30(1):e100722. [doi: 10.1136/bmjhci-2022-100722]
5. Pontefract SK, Wilson K. Using electronic patient records: defining learning outcomes for undergraduate education. BMC Med Educ 2019 Jan 22;19(1):30. [doi: 10.1186/s12909-019-1466-5] [Medline: 30670000]
6. Biagioli FE, Elliot DL, Palmer RT, et al. The electronic health record objective structured clinical examination: assessing student competency in patient interactions while using the electronic health record. Acad Med 2017 Jan;92(1):87-91. [doi: 10.1097/ACM.0000000000001276] [Medline: 27332870]
7. Gibson CM, Kwon HI, Tatachar A. Impact of a low-cost simulated electronic medical record on perceptions of APPE readiness. Curr Pharm Teach Learn 2019 Jul;11(7):736-741. [doi: 10.1016/j.cptl.2019.03.013] [Medline: 31227098]
8. Cook K, Cochran G, Gali H, Hatch T, Awdishu L, Lander L. Pharmacy students' readiness to use the electronic health record: a tale of two institutions. Curr Pharm Teach Learn 2021 Apr;13(4):327-332. [doi: 10.1016/j.cptl.2020.11.005] [Medline: 33715792]
9. Vlashyn OO, Adeoye-Olatunde OA, Illingworth Plake KS, Woodyard JL, Weber ZA, Russ-Jara AL. Pharmacy students' perspectives on the initial implementation of a teaching electronic medical record: results from a mixed-methods assessment. BMC Med Educ 2020 Jun 9;20(1):187. [doi: 10.1186/s12909-020-02091-8] [Medline: 32517745]
10. Elliott K, Marks-Maran D, Bach R. Teaching student nurses how to use electronic patient records through simulation: a case study. Nurse Educ Pract 2018 May;30:7-12. [doi: 10.1016/j.nepr.2018.02.003] [Medline: 29459316]
11. Ives AL, Tucker SR, Trovato JA. Using electronic health record technology to teach inpatient medication order verification to pharmacy students. Am J Pharm Educ 2020 Aug;84(8):ajpe7534. [doi: 10.5688/ajpe7534] [Medline: 32934381]

XSL•FO
RenderX

12. Samadbeik M, Fatehi F, Braunstein M, et al. Education and training on electronic medical records (EMRs) for health care professionals and students: a scoping review. Int J Med Inform 2020 Oct;142:104238. [doi: 10.1016/j.ijmedinf.2020.104238] [Medline: 32828034]
13. Blease C, McMillan B, Salmi L, Davidge G, Delbanco T. Adapting to transparent medical records: international experience with "open notes". BMJ 2022 Nov 21;379:e069861. [doi: 10.1136/bmj-2021-069861] [Medline: 36410770]
14. Vanka A, Johnston KT, Delbanco T, et al. Guidelines for patient-centered documentation in the era of open notes: qualitative study. JMIR Med Educ 2025 Jan 20;11:e59301. [doi: 10.2196/59301] [Medline: 39832175]
15. Standards of conduct, performance and ethics. Health and Care Professions Council. 2024. URL: https://www.hcpc-uk.org/standards/standards-of-conduct-performance-and-ethics/ [accessed 2025-03-25]
16. Riordan F, Papoutsi C, Reed JE, Marston C, Bell D, Majeed A. Patient and public attitudes towards informed consent models and levels of awareness of electronic health records in the UK. Int J Med Inform 2015 Apr;84(4):237-247. [doi: 10.1016/j.ijmedinf.2015.01.008] [Medline: 25649841]
17. Blease CR, Kharko A, Dong Z, et al. Experiences and opinions of general practitioners with patient online record access: an online survey in England. BMJ Open 2024 Jan;14(1):e078158. [doi: 10.1136/bmjopen-2023-078158]
18. Gaudet-Blavignac C, Foufi V, Bjelogrlic M, Lovis C. Use of the systematized nomenclature of medicine clinical terms (SNOMED CT) for processing free text in health care: systematic scoping review. J Med Internet Res 2021 Jan 26;23(1):e24594. [doi: 10.2196/24594] [Medline: 33496673]
19. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med 2020;3:17. [doi: 10.1038/s41746-020-0221-y] [Medline: 32047862]
20. Al-Hablani B. The use of automated SNOMED CT clinical coding in clinical decision support systems for preventive care. Perspect Health Inf Manag 2017;14(Winter):1f. [Medline: 28566995]

Abbreviations

EHR: electronic health record
GRIPP2: Guidance for Reporting Involvement of Patients and the Public 2
MDT: multidisciplinary team
NHS: National Health Service

Edited by A Mavragani; submitted 01.06.24; peer-reviewed by DA Mills, S Pontefract; revised version received 10.03.25; accepted 11.03.25; published 12.06.25.

Please cite as:
Nadeem F, Azmy J, Shomali AY, Diette B, Gregory LJ, Davies AC, Wilson KC
Using Real Electronic Health Records in Undergraduate Education: Roundtable Discussion
JMIR Form Res 2025;9:e60789
URL: https://formative.jmir.org/2025/1/e60789
doi:10.2196/60789

© Fatima Nadeem, Jessica Azmy, Asieh Yousefnejad Shomali, Benjamin Diette, Lloyd J Gregory, Angela C Davies, Kurt C Wilson. Originally published in JMIR Formative Research (https://formative.jmir.org), 12.6.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Safety Evaluation in Healthy Adults of Motion-Based Virtual Reality Dichoptic Training for Pediatric Patients With Amblyopia: Prospective Intervention Study

Masakazu Hirota[1,2,3,4,5], PhD; Yuichi Okumura[5,6,7], PhD; Ken Nagino[5,6,7,8], PhD; Takao Hayashi[1,2], PhD; Takashi Negishi[6], PhD; Shintaro Nakao[6], PhD; Hitoshi Kawasaki[9], ME; Takenori Inomata[5,6,7,8,10], PhD

[1]Department of Orthoptics, Faculty of Medical Technology, Teikyo University, Itabashi-ku, Tokyo, Japan
[10]Data Science, Juntendo University Graduate School of Medicine, Bunkyo-ku, Tokyo, Japan
[2]Department of Ophthalmology, School of Medicine, Teikyo University, Itabashi-ku, Tokyo, Japan
[3]Graduate Degree Program of Health Data Science, Teikyo University, Itabashi-ku, Tokyo, Japan
[4]Graduate Degree Program of Comprehensive Applied Data Science, Teikyo University, Itabashi-ku, Tokyo, Japan
[5]InnoJin Inc, Bunkyo-ku, Tokyo, Japan
[6]Department of Ophthalmology, Juntendo University Graduate School of Medicine, 2-1-1 Hongo, Bunkyo-ku, Tokyo, Japan
[7]Department of Telemedicine and Mobile Health, Juntendo University Graduate School of Medicine, Bunkyo-ku, Tokyo, Japan
[8]Department of Hospital Administration, Juntendo University Graduate School of Medicine, Bunkyo-ku, Tokyo, Japan
[9]Ima-Create Inc, Shinagawa-ku, Tokyo, Japan

Corresponding Author:
Takenori Inomata, PhD
InnoJin Inc, Bunkyo-ku, Tokyo, Japan

Abstract

Background: Amblyopia, a unilateral or bilateral visual disorder, affects up to 5% of the general population and is a leading cause of childhood visual impairment. Current treatments, such as patching therapy, aim to improve amblyopia by temporarily occluding the unaffected eye, thereby promoting the use of the amblyopic eye. However, adherence to patch therapy can be challenging, as the forced use of the amblyopic eye can be stressful for children. Moreover, despite improvements in visual acuity by patch therapy, children with amblyopia often face difficulties with hand-eye coordination; therefore, a treatment that reduces stress for them while simultaneously improving hand-eye coordination could address the limitations of existing amblyopia therapies.

Objective: This study investigated the safety of our motion-based virtual reality (VR) dichoptic training app using Japanese Kendama in healthy adult participants, which was designed to improve hand-eye coordination in pediatric patients with amblyopia.

Methods: This prospective intervention study involved 20 healthy young adults (median age 21, IQR 21 - 28.3 y), including 16 women. The participants played the motion-based VR dichoptic training app for 30 minutes and then completed a subjective symptom questionnaire, which comprised 9 questions (Q1-Q9) with each item scored on a 4-point scale, except Q9, which was assessed on a binary scale. Q1-Q3 focused on subjective eye symptoms, Q4-Q7 evaluated physical and mental discomfort, Q8 assessed the degree of VR session–induced arm fatigue, and Q9 assessed the severity of visually induced motion sickness.

Results: No significant differences were observed in the reported ocular symptoms before and after the VR session, including eye fatigue (mean before vs after: 1.25, SD 0.94 points vs 1.35, SD 0.85 points), blurred vision (0.55, SD 0.50 points vs 0.80, SD 0.40 points), eye dryness (0.95, SD 0.74 points vs 1.25, SD 0.83 points), and visually induced motion sickness (0.00, SD 0.00 points vs 0.05, SD 0.22 points). These results suggested that the motion-based VR dichoptic training did not induce significant adverse ocular effects.

Conclusions: The motion-based VR dichoptic training app demonstrated minimal adverse ocular effects in healthy adult participants, suggesting that it is safe for use in this population. These findings demonstrate the feasibility and good tolerability of this VR-based intervention in healthy adults. Further studies, including clinical studies in adult and pediatric patients with amblyopia, are warranted to evaluate its applicability and therapeutic effects.

(JMIR Form Res 2025;9:e69801) doi:10.2196/69801

KEYWORDS

amblyopia; adherence; safety evaluation; virtual reality; dichoptic training; strabismus; digital health; digital therapeutics; visually induced motion sickness

XSL•FO
RenderX

Introduction

Amblyopia, defined as a unilateral or bilateral visual disorder, affects up to 5% of the general population [1-3] and is a common cause of childhood visual impairment [4]. Refractive error correction with spectacles is the first-line treatment for pediatric amblyopia, regardless of the cause [5,6]. However, occlusion treatment (patching therapy), which involves blocking the unaffected eye to force the use of the amblyopic eye, may be required in some cases [7]. While combining corrective spectacles and patching therapy can improve visual acuity and shorten the treatment duration [8], treatment adherence and patient stress are significant clinical challenges. Studies have shown that treatment adherence to patching therapy is often <60% [9,10]. Many children with amblyopia dislike wearing eye patches, and there is a risk of vision impairment in the healthy eye with prolonged patching. Additionally, children with amblyopia treated with patching therapy often demonstrate lower binocular function than normal children [11]. By forcing the use of monocular vision during their developmental period, patching therapy may inhibit the development of normal binocular visual function in children.

Dichoptic training, which was presented by Hess et al [12], offers a promising approach to improving visual acuity in pediatric patients with amblyopia. This method involves the presentation of contrasting visual stimuli to each eye to restore binocular fusion, achieve stereopsis, and improve visual acuity in the amblyopic eye. Video games have emerged as a suitable platform for dichoptic training owing to their inherent engagement, allowing the promotion of binocular vision while treating amblyopia [13,14]. Furthermore, the contrast manipulation method in dichoptic training translates well to virtual reality (VR) head-mounted displays (HMDs) [15]. VR allows users to interact with and navigate a computer-generated environment in real time. Several studies have already explored the therapeutic effects of VR games for amblyopia treatment [16-18]. Žiak et al [16] reported a 30% improvement in visual acuity following 1 month of VR-HMD training (Vivid Vision Inc). Xiao et al [18] similarly showed that 12 weeks of VR-HMD dichoptic training, which used contrast manipulation and blacking out specific video regions (Luminopia Inc), improved the visual acuity of amblyopic eyes. These findings suggest that the outcomes of VR-based approaches are comparable to conventional patching therapy in treating amblyopia. By incorporating digital, game-like elements, VR-based approaches demonstrate comparable efficacy to patching therapy, potentially enhancing adherence and engagement [19,20].

Previous studies on VR games for amblyopia primarily focused on nonengaging content, such as passive image viewing or using conventional consumer game controllers, which underuse the full potential of VR technology to enhance hand-eye coordination [16-18]. Children with amblyopia typically exhibit reduced hand-eye coordination [21,22]; however, its improvement with amblyopia treatment has not been widely established. To address this limitation, we developed a motion-based VR dichoptic training app (Figure 1) aimed at enhancing hand-eye coordination in pediatric patients with amblyopia. Unlike previous VR games for amblyopia that involve passive viewing, our app uses motion-tracking hand controllers that allow real-time manipulation of the virtual world, mirroring the user's actual hand movements. This real-time movement may also lead to higher risks of visual fatigue and visually induced motion sickness (VIMS) in pediatric patients with amblyopia compared to previous VR treatments, which involve less active engagement. Additionally, the use of motion-based VR dichoptic training apps in clinical trials for pediatric patients with amblyopia necessitates resolving ethical considerations, such as safety and efficacy, due to the time-sensitive nature of amblyopia treatment.

In this study, we investigated the safety of our novel motion-based VR dichoptic training app in healthy young adults; in the future, the results of this study will be validated first in adult patients with amblyopia, and thereafter, in pediatric patients with amblyopia.

XSL•FO
RenderX

Figure 1. Images of the VR-based dichoptic training app. (A) The VR HMD is worn by the participant. The right-hand controller is linked to the ken (handle). (B) Screenshot of the VR-based dichoptic training app using Japanese Kendama, while watching the ideal Kendama movement in grayscale. (C-H) Transmittance of one eye (dominant eye) can be changed from 100% to 0% in 20% decrements (C: 100%; D: 80%; E: 60%; F: 40%; G: 20%; and H: 0%). HMD: head-mounted display; VR: virtual reality.

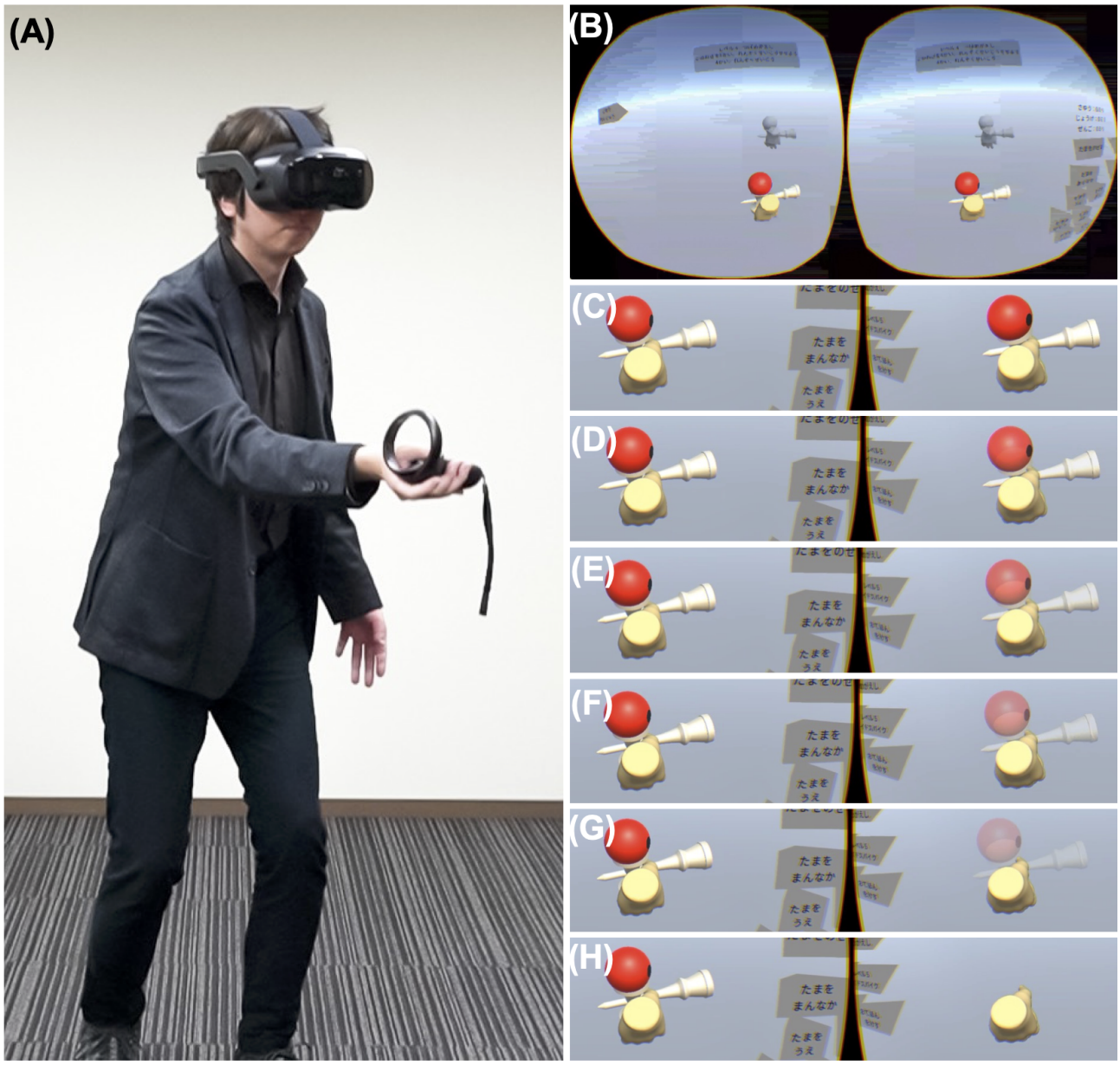

Methods

Participants

Participants for this preliminary study were recruited through posters displayed on the university campus of Teikyo University (Tokyo, Japan) and announcements on our website. A total of 20 healthy young adult volunteers were enrolled. All participants underwent comprehensive ophthalmologic examinations, including ocular dominance assessment using the hole-in-the-card test, best-corrected visual acuity evaluation at a distance (5 m), near the point of convergence measurement, stereoscopic acuity evaluation at 40 cm (Titmus Stereo test; Stereo Optical Co, Inc), heterophoria assessment using the alternating cover test at both near (33 cm) and far (5 m) distances, and fundus examination. Stereoacuity was converted to logarithm of arcsec (log arcsec) at Teikyo University. Participants were excluded if they exhibited any of the following characteristics: a best-corrected visual acuity of <20/20, a near point of convergence of >8 cm, stereoacuity of >100 arcsec, and the presence of manifest strabismus (including intermittent exotropia) or any retinal abnormalities.

Ethical Considerations

Written informed consent was obtained from all participants after a thorough explanation of the study procedures and potential risks. This study was approved by the Institutional Review Board of Teikyo University (22–061) and was conducted in accordance with the principles of the Declaration of Helsinki. To ensure participant privacy and confidentiality, all collected data were anonymized and deidentified before analysis. Participants were informed that no compensation would be provided as the study involved minimal intervention.

XSL•FO
RenderX

Additionally, no identifiable images of participants were included in the study or supplementary materials.

Motion-Based VR Dichoptic Training App for Pediatric Amblyopia

A motion-based VR dichoptic training app using Japanese Kendama was developed using VIVE Focus 3 (HTC Corp) on Unity (version 2020.3.13; Unity Technologies; Figure 1A). The controller for the right hand was linked to a ken (handle), as its weight (145 g) approximated the weight of the actual ken (140 g). Participants can adjust the physical and spatial parameters of the tama (ball) through an in-VR touch panel (Figure 2). The speed of the tama can be set to 1.0, 0.7, or 0.4 times that of the real-world speed. Difficulty levels for placing the tama on the Kendama are selectable as "easy," "normal," and "difficult," with "difficult" simulating real-world Kendama difficulty. As the VR space presents parallax images to each eye, touch panels are available for the horizontal, vertical, forward, and backward adjustments of the tama's position. An additional touch panel allows participants to restart the game by placing the tama on the sara (dishes) if it falls on the ground.

Figure 1B shows a screenshot of the motion-based VR dichoptic training app using Japanese Kendama while watching the ideal Kendama movement in grayscale. The examiner controlled the Kendama environment using a left-hand controller, which can adjust the tama's transparency in one eye (Figure 1C-H; 100%, 80%, 60%, 40%, 20%, or 0%; completely transparent). Furthermore, it is possible to completely separate the visual stimuli for each eye, displaying the ken and tama correspondingly.

Figure 2. Screenshots of the setting panels for the VR-based dichoptic training app. (A) Restart function. If the waza fails and the tama falls to the ground, the participants touch "Put Tama on Sara." (B) Speed adjustment function. The ball speed can be changed by touching "Tama speed faster" or "Tama speed slower" in 3 levels: "slow," "normal," and "fast." (C) Difficulty adjustment function. When participants try to play waza, it can be determined whether the tama is regarded as a success even if it slightly deviates from the sara or whether it is played at the same difficulty level as real Kendama ("difficult") by touching the "Catch more difficult" or "Catch easier" in 3 levels: "easy," "normal," and "difficult." (D) Position alignment function. In VR, the binocular disparity may be larger than in real space because images are presented with a disparity between the left and right eyes. For some participants, the tama may not be aligned on the sara, so the offset of the balls is adjusted by touching the "Set Tama upper, lower, left, right, forward, and backward" panel. VR: virtual reality.

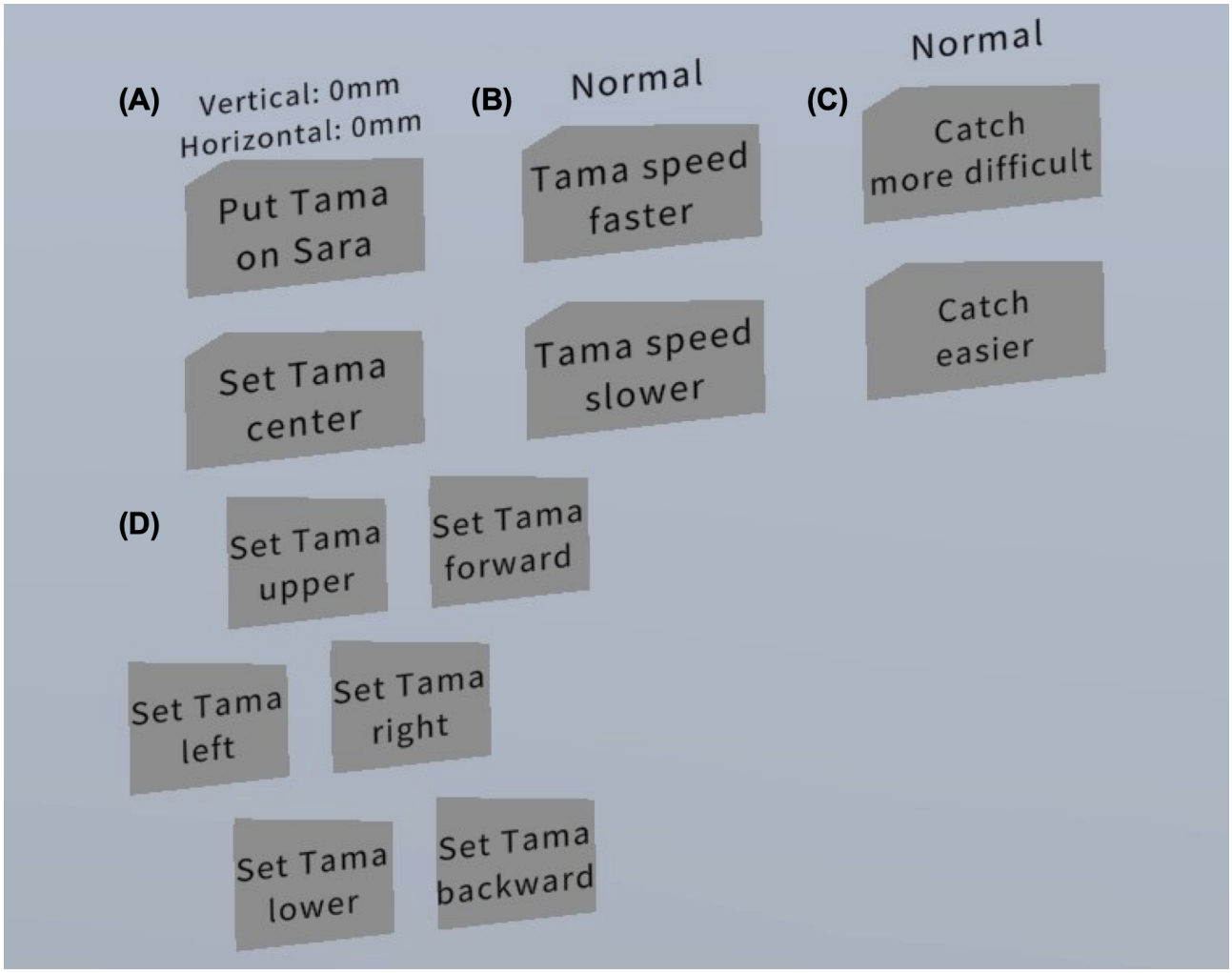

Kendama Task

Participants initiated the Kendama task by selecting the "Challenge" panel on the touch screen. This task comprised 5 wazas (techniques): "Ozara (big dish)," "Kozara (small dish)," "Moshikame (a basic repetitive catching technique)," "Orbit (a trick where the ball moves in a circular path around the spike)," and "Side spike (a technique where the ball is caught on the

XSL•FO
RenderX

side spike of the kendama).” Before each attempt, the system displayed an instructional video demonstrating the ideal Kendama movement of the waza being performed. The task was designed such that successful completion of a waza 5 consecutive times allowed progression to the next waza.

Subjective Symptom Questionnaire

Participants were instructed to complete a subjective symptom questionnaire before and after the VR session. The questionnaire was adapted from the previous studies of Nakazawa et al [23], Sheedy and Bergstrom [24], and Hoffman et al [25]. Q1-Q7 were the same as those used in a previous study (Figure 3) [26-28]. Specifically, Q1-Q3 assessed subjective eye symptoms, and Q4-Q7 assessed physical and mental discomfort. Q1-Q7 have been widely used in previous studies to evaluate eye strain, as well as physical and mental discomfort associated with the use of digital devices. Q8 was designed to evaluate the degree of VR session-induced arm fatigue, and Q9 was designed to evaluate VIMS. Participants scored each question on a 4-point scale (0-4), except for Q9, which was scored as either 0=no or 1=yes point.

XSL•FO
RenderX

Figure 3. Subjective symptom questionnaire. Q1-Q3 assess subjective eye symptoms; Q4-Q7 assess physical and mental discomfort; Q8 evaluates the degree of VR session–induced arm fatigue; and Q9 evaluates VIMS. n.p.: no problem; VIMS: visually induced motion sickness; VR: virtual reality.

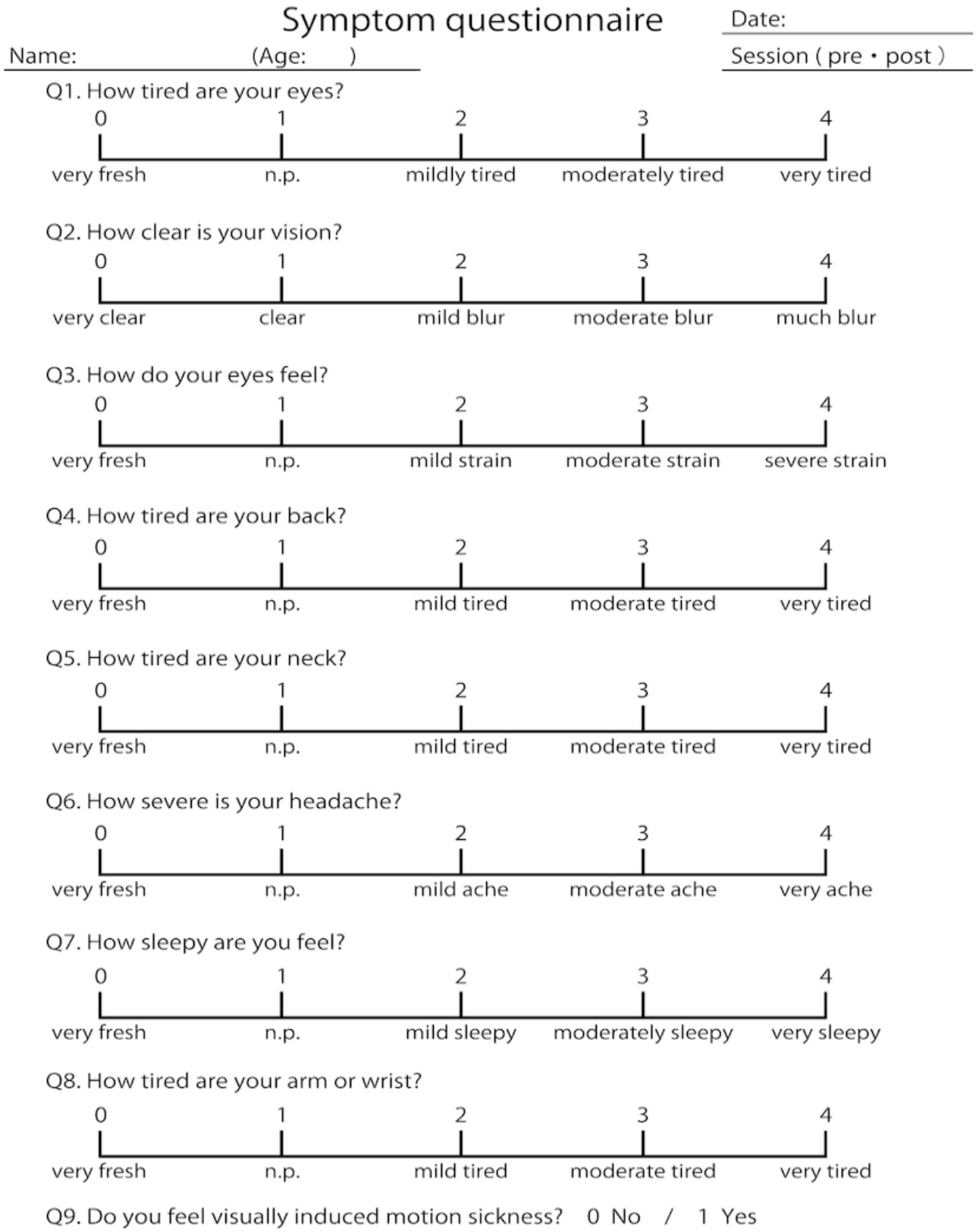

Experimental Procedures

In this study, participants completed a subjective symptom questionnaire before and after the 30-minute motion-based VR dichoptic training sessions at Teikyo University. The duration was determined based on prior studies [27,29]. Participants' refractive errors were corrected using soft contact lenses (SEED 1 day Pure moisture, Seed Co, Ltd) because VR-HMD is difficult to wear with glasses. With contact lenses, all participants had equal or better than 0.0 logMAR visual acuity. The Kendama task was performed at a tama speed of 0.4×, and the difficulty level was set to "easy."

XSL•FO
RenderX

Initially, both eyes received a transmittance of 100%. Upon successful completion of the 5 wazas, the transmittance of the right eye (assumed to be healthy) was reduced to 60%, defined as the second condition. After completing all 5 wazas under the second condition, the transmittance of the right eye was reduced to 20%, which is defined as the third condition. Similarly, upon task completion under the third condition, participants performed the 5 wazas in the final condition, where only the tama and ken were displayed in the right and left eyes, respectively. The progression of transmittance reduction was designed as a game mechanic to gradually increase task difficulty and maintain participant engagement, rather than being a direct intervention for amblyopia treatment. The primary goal of this phase was to assess the feasibility and usability of the VR app with healthy participants rather than applying clinical treatment principles. The decision to reduce transmittance consecutively (and not simultaneously) was made to incrementally challenge the participants' ability to complete the task while maintaining their attention and engagement. Notably, this procedure is distinct from contrast balancing used in clinical dichoptic training for amblyopia. In clinical settings, adjustments to transmittance would be customized based on the patient's visual acuity and would likely involve simultaneous changes to both eyes, rather than the sequential reduction observed in this study. Additionally, contrast adjustments for amblyopia therapy would aim to improve visual function, whereas the transmittance changes in this study were primarily intended to serve as part of a gamified challenge.

Statistical Analysis

This study investigated the continuous response variable from matched pairs of participants. A previous study indicated that the difference in responses between matched pairs was nonnormally distributed with an SD of 0.60 [27]. Assuming a true difference of 0.58 in the mean response of matched pairs, 18 pairs of participants would be required to be able to reject the null hypothesis ("The response difference is zero") with a probability (power) of 0.8. The Type I error probability to test the null hypothesis was 0.05. Anticipating a 10% dropout rate owing to missing data or consent withdrawal, a total of 20 patients were needed.

Continuous variables are presented as median (IQR) or mean (SD). Differences in subjective symptom scores before and after motion-based VR dichoptic training sessions were analyzed using the Wilcoxon signed rank test. The *P* values for subjective symptoms were adjusted using Holm correction. Statistical significance was determined using SPSS Statistics for Windows (version 26; IBM Corp) with a significance level of $P<.05$.

Results

Characteristics of Participants

Ocular deviations with minus and plus signs indicate exodeviation and esodeviation, respectively.

The median age of the participants was 21 (21 - 28.3) years, and 16 of 20 (80%) participants were women. The mean spherical equivalent refractive errors were −3.36 (SD 3.14) diopter (D) and −3.32 (SD 3.09) D for the dominant and nondominant eyes, respectively. All participants demonstrated a best-corrected visual acuity of 0.0 logMAR units or better. Mean heterophoria values were −0.4 (SD 0.8) prism diopter and −3.5 (SD 4.0) prism diopter at distance and near proximity, respectively. All healthy volunteers exhibited a stereoacuity of 1.60 log arcsec (40 s).

Table 1 presents the participant characteristics. This study included 20 healthy young adults.

XSL•FO
RenderX

Table . Characteristics of participants.

ID	Age (years)	Sex	Spherical equivalent (D)[a]		Log stereoacuity	Ocular deviation (PD[b])[c]	
			RE[d]	LE[e]		Near	Distance
1	21	Female	−3.125	−2.875	1.600	−10.000	−2.000
2	21	Female	−7.750	−7.625	1.600	−4.000	0.000
3	21	Female	−1.875	−1.500	1.600	−8.000	−2.000
4	21	Female	−1.250	−0.250	1.600	2.000	0.000
5	21	Female	−0.625	−5.875	1.600	2.000	0.000
6	21	Female	−11.125	−11.125	1.600	−6.000	0.000
7	21	Female	−3.625	−3.750	1.600	−4.000	0.000
8	40	Female	−5.000	−5.000	1.600	−4.000	0.000
9	29	Male	−0.750	−0.750	1.600	−6.000	−2.000
10	25	Female	0.625	0.125	1.600	0.000	0.000
11	22	Female	−6.875	−5.000	1.600	−4.000	0.000
12	26	Male	−3.000	−3.000	1.600	−4.000	0.000
13	21	Female	−7.125	−7.875	1.600	−4.000	0.000
14	21	Female	−7.250	−5.750	1.600	−10.000	−2.000
15	21	Female	−3.250	−2.875	1.600	−8.000	0.000
16	33	Male	0.000	0.000	1.600	0.000	0.000
17	33	Male	−3.500	−2.625	1.600	−4.000	0.000
18	32	Female	0.000	0.250	1.600	−2.000	0.000
19	21	Female	0.000	−0.125	1.600	6.000	1.000
20	20	Female	−1.625	−0.750	1.600	−2.000	0.000

[a]D: diopter.

[b]PD: prism diopter.

[c]Ocular deviations with minus and plus signs indicate exodeviation and esodeviation, respectively.

[d]RE: right eye.

[e]LE: left eye.

Subjective Symptom Scores Before and After Motion-Based VR Dichoptic Training

Subjective symptoms were evaluated using a 9-item questionnaire (Figure 3) [26-28]. Table 2 and Figure 4 show the subjective symptom scores before and after using the motion-based VR dichoptic training app.

No significant differences were observed in subjective eye symptoms (Q1-Q3) before and after motion-based VR dichoptic training. In contrast, physical and mental discomfort (Q4; mean 0.95, SD 0.59 points vs 1.90, SD 1.18 points; n=20; P=.03) and arm fatigue (Q8; mean 0.30, SD 0.46 points vs 2.60, SD 1.11 points; n=20; P<.001) were significantly greater following the task. None of the participants experienced VIMS (Q9; mean 0.00, SD 0.00 points vs 0.05, SD 0.22 points; n=20; P=.41) before or after the task. Of the 20 participants, 1 (5%) participant reported experiencing VIMS after the motion-based VR dichoptic training app.

XSL•FO
RenderX

Table . Subjective symptoms before and after using the motion-based VR[a] dichoptic training app.

Subjective symptom questions	Motion-based VR dichoptic training app (n=20)		*P* value[b]
	Before, mean (SD)	After, mean (SD)	
Q1. How tired are your eyes? (0–4 points)	1.25 (0.94)	1.35 (0.85)	.41
Q2. How clear is your vision? (0–4 points)	0.55 (0.50)	0.80 (0.40)	.24
Q3. How does your eye feel? (0–4 points)	0.95 (0.74)	1.25 (0.83)	.08
Q4. How tired is your back? (0–4 points)	0.95 (0.59)	1.90 (1.18)	.03
Q5. How tired is your neck? (0–4 points)	1.05 (0.59)	1.55 (0.86)	.09
Q6. How severe is your headache? (0–4 points)	0.65 (0.48)	0.95 (0.74)	.19
Q7. How sleepy do you feel? (0–4 points)	1.10 (0.62)	0.80 (0.60)	.08
Q8. How tired are your arm or wrist? (0–4 points)	0.30 (0.46)	2.60 (1.11)	<.001
Q9. Do you feel visually induced motion sickness? (0=no or 1=yes point)	0.00 (0.00)	0.05 (0.22)	.41

[a]VR: virtual reality.

[b]The *P* values were adjusted using the Holm correction method.

Figure 4. Subjective symptom scores before and after using the VR-based dichoptic training app. Box plots illustrating subjective symptom scores before (blue) and after (red) using the VR-based dichoptic training app in all participants (n=20). Dots represent individual data points. Significant increases were observed in back discomfort (Q4) and arm fatigue (Q8) after the task. **P*=.03, Wilcoxon signed rank test with Holm correction. ****P*<.001, Wilcoxon signed rank test with Holm correction. VR: virtual reality.

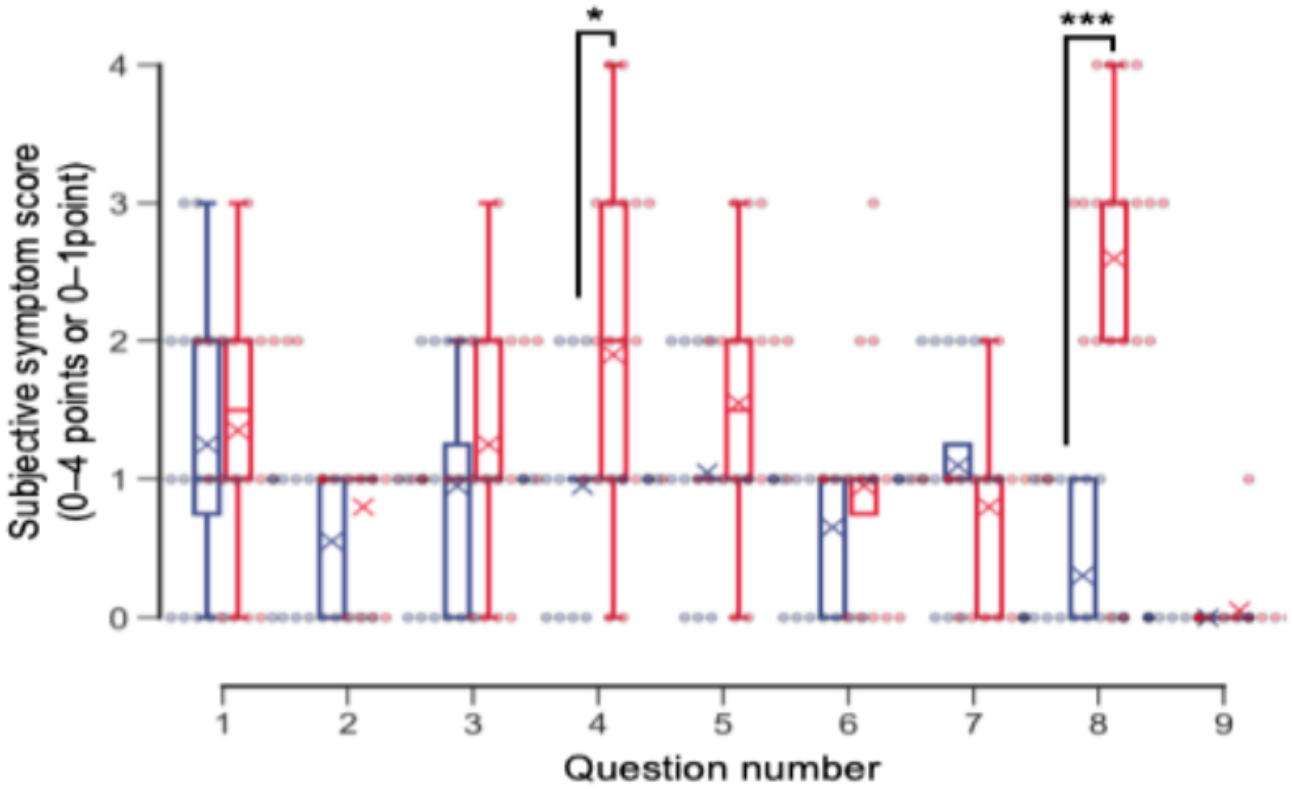

XSL•FO
RenderX

Discussion

Principal Results

Pediatric amblyopia is a leading cause of childhood visual impairment, and while current treatments, such as corrective glasses and patch therapy, can improve visual acuity, they often exhibit low treatment adherence, cause stress for pediatric patients, and fail to fully address hand-eye coordination deficits. In this prospective intervention study, we investigated the potential ocular adverse effects of the motion-based VR dichoptic training app, designed for pediatric amblyopia treatment with enhanced hand-eye coordination. To ensure safety before implementation in pediatric patients, we first evaluated its effects in healthy adults, focusing on potential ocular adverse effects. The results showed no significant ocular symptoms after a 30-minute motion-based VR-dichoptic training, including eye fatigue, blurred vision, eye dryness, and VIMS. These findings suggest that the motion-based VR dichoptic training app is a low-risk intervention with minimal ocular side effects in healthy adults. Further studies are required to evaluate its safety, feasibility, and effectiveness in pediatric patients with amblyopia.

Ocular Adverse Effects

This study found that visual fatigue did not significantly differ before and after using the motion-based VR dichoptic training app. Hirota et al [27] reported that 30-minute VR-HMD sessions induced a similar level of visual fatigue as conventional 2D displays. This discrepancy between their findings and ours may be attributed to differences in VR headset technology. The VR-HMD used in this study (VIVE Focus) differs from that of Hirota et al [27] (PlayStation VR). VR headsets released around 2016, such as the PlayStation VR, typically featured approximately 2K resolution, 100° field of view, and 90 Hz refresh rate. In contrast, modern VR headsets offer approximately 4K resolution, a 200° field of view, and a 120 Hz refresh rate. Additionally, advancements in motion sensor technology have enabled more precise synchronization between the avatar's movements in the virtual environment and the user's hand controller movements in real space. Moreover, the unique characteristics of the VR content itself may have contributed to the lack of significant visual fatigue in this study. These findings suggest that our motion-based VR dichoptic training app offers an active, gamified VR-based experience with minimal visual fatigue in healthy adults.

In this study, the absence of complaints regarding eye dryness following the use of the motion-based VR dichoptic training app likely reflects the nature of the training task. Video games often encourage continuous visual fixation on the screen, reducing blink frequency and leading to eye dryness [30]. However, in this study, the participants engaged in 5 Kendama wazas (techniques)—"Ozara (big dish)," "Kozara (small dish)," "Moshikame," "Orbit," and "Side spike." When a waza was successfully completed, participants could continue playing, whereas unsuccessful attempts required them to look away and interact with a touch panel to reset the task. We believe these intermittent visual breaks helped prevent excessive tear evaporation and subsequent eye dryness.

This study also demonstrated that VIMS was not induced by 30 minutes of motion-based VR dichoptic training app use. VIMS occurs due to a discrepancy between visual and somatic sensations, which can be influenced by 3 factors: hardware, content, and human physiology [31]. In terms of hardware, VIVE Focus operates at a high sampling rate, reducing discomfort associated with image perception. Regarding content, the binocular disparity was minimized by adjusting the relative positions of the "tama (ball)" and the "sara (dish)" in the VR space. Jackson and Bedell [32] reported that correcting vertical heterophoria can reduce VIMS symptoms. In this study, adjusting both horizontal and vertical heterophoria to zero in the VR space may have contributed to the suppression of VIMS and enhanced visual comfort. This correction ensures that participants with varying degrees of heterophoria can experience more stable binocular alignment, potentially reducing discomfort and improving the overall usability of the VR app. Additionally, the nature of the movements in our VR task—controlled and primarily upper-body focused—may have contributed to minimizing VIMS risk by reducing the sensory mismatch between visual and vestibular inputs. Finally, regarding human factors, differences between adults and children may influence the incidence of VIMS when pediatric patients with amblyopia use our motion-based VR dichoptic training app in the future. However, our findings indicate that, at least in terms of hardware and content, the motion-based VR dichoptic training app has been designed to minimize conditions that induce VIMS.

Physical and Mental Adverse Effects

Notably, some participants reported physical discomfort, primarily back discomfort, and arm fatigue, after the 30-minute motion-based VR dichoptic training session. These complaints are likely due to the weight of the VR-HMD and controller. Previous studies on VR-HMD-based amblyopia treatment have not reported physical adverse effects [16-18]. Since this study involved young adults with more developed musculature than children, pediatric patients using VR-HMDs may experience even greater back strain depending on their body weight. This concern is particularly relevant as children undergoing treatment would likely wear these devices regularly, potentially multiple times per day as part of their prescribed therapeutic regimen. Prolonged use of heavy VR-HMDs during critical growth periods could impact both physical and mental development [33,34]. Additionally, close monitoring of task performance is crucial when children use commercially available VR-HMDs. The observed arm fatigue likely stemmed from the participant's unfamiliarity with the task and the extended session duration (30 minutes). A systematic review of VR interventions for children with cerebral palsy has reported effective results with session durations of 20 - 30 minutes [35]. While the optimal duration for VR-based dichoptic training remains to be determined, these findings suggest that limiting session length may help balance efficacy with user comfort. Therefore, future studies involving pediatric patients should consider the use of lighter VR-HMDs and controllers, as well as shorter session durations, to optimize usability and minimize physical strain.

XSL•FO
RenderX

Limitations

Despite the valuable insights gained from this study, some limitations should be acknowledged. First, this study assessed eye fatigue and VIMS for 30 minutes of motion-based VR dichoptic training app use in healthy adult participants, which aligns with the previous studies [13,16,18]. However, the effects of the app on eye fatigue and VIMS should be further evaluated in adult patients with amblyopia and in case of good tolerance, in pediatric patients with amblyopia. Second, Q8 of the subjective symptom questionnaire was originally designed to assess VR session–induced arm fatigue, while Q9 was intended to measure VIMS. These questions have not been assessed for validity and reliability, as no standardized questionnaire currently exists for assessing VIMS during VR use involving physical activity. Future research should incorporate a validated questionnaire to evaluate VIMS more accurately in this context. Third, adapting the use of VR-HMDs for pediatric patients with amblyopia requires ensuring a proper fit, accommodating a range of interpupillary distances, and minimizing the VR-HMD weight. This is particularly important because most commercially available VR-HMDs are optimized for adult interpupillary distance ranges, which may not be suitable for younger children and could affect alignment and comfort. While this study demonstrated the safety of the motion-based VR dichoptic training app in healthy adult participants, further refinements are necessary before its implementation in pediatric patients. Moving forward, we intend to conduct further investigations in adult patients who have previously undergone amblyopia treatment but are currently experiencing declining visual acuity in one eye. This approach would allow us to refine and validate the treatment protocol without requiring additional hardware development.

Conclusions

Our 30-minute motion-based VR dichoptic training app session resulted in minimal visual discomfort, including minimal eye fatigue, blurred vision, eye dryness, and VIMS. These results provide preliminary evidence that daily amblyopia training using motion-based VR dichoptic training apps may minimize adverse ocular effects. Future studies in adults and children with amblyopia are needed to assess the tolerability and efficacy of the motion-based VR dichoptic training app for the treatment of patients with amblyopia.

Acknowledgments

The authors thank Kakeru Sasaki, Hiroki Nishimura, and Maki Watanabe for assisting with this study. This work was supported by the Early-Career Scientists, Scientific Research (A) and (B), Japan Society for the Promotion of Science (grants 22K18231 to MH, 20H04271 to MH, and 22H00539 to MH), Japan Agency for Medical Research and Development (grants 23he0422031j0001 to MH, 22he2722008j0001 to TI, and 24he1022014h0001 to TI), Nakatani Foundation (to MH), Inamori Foundation (to MH), Japan Medical Herb Association Research Grant Program (to MH), TERUMO LIFE SCIENCE FOUNDATION (grant 22-III 1009 to MH), and a Tateishi Science and Technology Foundation Research Grant (A) 2024 (grant 2241005 to TI). The funding sources had no role in the study design; collection, analysis, and interpretation of data; writing of the report; and the decision to submit the article for publication.

Data Availability

The datasets generated and analyzed during this study are not publicly available due to privacy concerns and institutional policy but are available from the corresponding author upon reasonable request.

Conflicts of Interest

MH and KN received personal fees from InnoJin, Inc, Tokyo, Japan. YO and TI are the owners of InnoJin, which developed the motion-based virtual reality (VR)–dichoptic training app. SN received consulting fees from Chugai Pharmaceutical Co, Ltd; Kowa Company, Ltd, Novartis Pharma KK, and Riverfield Inc; and travel reimbursements and speaker fees from Alcon Inc, Boehringer Ingelheim Co, Ltd, Bayer Yakuhin Ltd, Canon Inc, JFC Sales Plan Co, Ltd, Kowa Company, Ltd, Mitsubishi Tanabe Pharma Corporation, Machida Holdings Inc, MSD KK, Novartis Pharma KK, Novo Nordisk Pharma Ltd, Otsuka Pharmaceutical Co, Ltd, Santen Pharmaceutical Co, Ltd, Senju Pharmaceutical Co, Ltd, and Wakamoto Pharmaceutical Co, Ltd, not related to the submitted work. TI received nonfinancial support from Lion Corporation and Sony Network Communications Inc; grants from Johnson & Johnson Vision Care Inc, Yuimedi Inc, ROHTO Pharmaceutical Co Ltd, Kobayashi Pharmaceutical Co Ltd, Kandenko Co Ltd, and Fukoku Co Ltd; and personal fees from InnoJin Inc, related to the submitted work, and from Santen Pharmaceutical Co Ltd and Ono Pharmaceutical Co, Ltd, not related to the submitted work.

References

1. Attebo K, Mitchell P, Cumming R, Smith W, Jolly N, Sparkes R. Prevalence and causes of amblyopia in an adult population. Ophthalmology 1998 Jan;105(1):154-159. [doi: 10.1016/s0161-6420(98)91862-0] [Medline: 9442792]
2. Faghihi M, Hashemi H, Nabovati P, et al. The prevalence of amblyopia and its determinants in a population-based study. Strabismus 2017 Dec;25(4):176-183. [doi: 10.1080/09273972.2017.1391849] [Medline: 29144181]
3. Wu C, Hunter DG. Amblyopia: diagnostic and therapeutic options. Am J Ophthalmol 2006 Jan;141(1):175-184. [doi: 10.1016/j.ajo.2005.07.060] [Medline: 16386994]

XSL•FO
RenderX

4. Liang YB, Friedman DS, Wong TY, et al. Prevalence and causes of low vision and blindness in a rural chinese adult population: the Handan Eye Study. Ophthalmology 2008 Nov;115(11):1965-1972. [doi: 10.1016/j.ophtha.2008.05.030] [Medline: 18684506]
5. de Zárate BR, Tejedor J. Current concepts in the management of amblyopia. Clin Ophthalmol 2007 Dec;1(4):403-414. [Medline: 19668517]
6. Tan JHY, Thompson JR, Gottlob I. Differences in the management of amblyopia between European countries. Br J Ophthalmol 2003 Mar;87(3):291-296. [doi: 10.1136/bjo.87.3.291] [Medline: 12598440]
7. Moseley MJ, Fielder AR, Irwin M, Jones HS, Auld RJ. Effectiveness of occlusion therapy in ametropic amblyopia: a pilot study. Br J Ophthalmol 1997 Nov;81(11):956-961. [doi: 10.1136/bjo.81.11.956] [Medline: 9505818]
8. Scheiman MM, Hertle RW, Beck RW, et al. Randomized trial of treatment of amblyopia in children aged 7 to 17 years. Arch Ophthalmol 2005 Apr;123(4):437-447. [doi: 10.1001/archopht.123.4.437] [Medline: 15824215]
9. Wallace MP, Stewart CE, Moseley MJ, et al. Compliance with occlusion therapy for childhood amblyopia. Invest Ophthalmol Vis Sci 2013 Sep 17;54(9):6158-6166. [doi: 10.1167/iovs.13-11861] [Medline: 23882695]
10. Searle A, Norman P, Harrad R, Vedhara K. Psychosocial and clinical determinants of compliance with occlusion therapy for amblyopic children. Eye (Lond) 2002 Mar;16(2):150-155. [doi: 10.1038/sj.eye.6700086] [Medline: 11988815]
11. Birch EE. Amblyopia and binocular vision. Prog Retin Eye Res 2013 Mar;33:67-84. [doi: 10.1016/j.preteyeres.2012.11.001] [Medline: 23201436]
12. Hess RF, Mansouri B, Thompson B. A new binocular approach to the treatment of amblyopia in adults well beyond the critical period of visual development. Restor Neurol Neurosci 2010;28(6):793-802. [doi: 10.3233/RNN-2010-0550] [Medline: 21209494]
13. Emmanouil B, Spaho J, Chatzea M, Gleni A, Plainis S. Dichoptic game training in strabismic amblyopia improves the visual evoked response. Cureus 2023 Sep;15(9):37854740. [doi: 10.7759/cureus.45395] [Medline: 38022221]
14. Vedamurthy I, Nahum M, Huang SJ, et al. A dichoptic custom-made action video game as a treatment for adult amblyopia. Vision Res 2015 Sep;114:173-187. [doi: 10.1016/j.visres.2015.04.008] [Medline: 25917239]
15. Mon-Williams M, Wann JP, Rushton S. Binocular vision in a virtual world: visual deficits following the wearing of a head-mounted display. Ophthalmic Physiol Opt 1993 Oct;13(4):387-391. [doi: 10.1111/j.1475-1313.1993.tb00496.x] [Medline: 8278192]
16. Žiak P, Holm A, Halička J, Mojžiš P, Piñero DP. Amblyopia treatment of adults with dichoptic training using the virtual reality oculus rift head mounted display: preliminary results. BMC Ophthalmol 2017 Jun 28;17(1):105. [doi: 10.1186/s12886-017-0501-8] [Medline: 28659140]
17. Elhusseiny AM, Bishop K, Staffa SJ, Zurakowski D, Hunter DG, Mantagos IS. Virtual reality prototype for binocular therapy in older children and adults with amblyopia. J AAPOS 2021 Aug;25(4):217. [doi: 10.1016/j.jaapos.2021.03.008] [Medline: 34246761]
18. Xiao S, Angjeli E, Wu HC, et al. Randomized controlled trial of a dichoptic digital therapeutic for amblyopia. Ophthalmology 2022 Jan;129(1):77-85. [doi: 10.1016/j.ophtha.2021.09.001] [Medline: 34534556]
19. Rajavi Z, Soltani A, Vakili A, et al. Virtual reality game playing in amblyopia therapy: a randomized clinical trial. J Pediatr Ophthalmol Strabismus 2021;58(3):154-160. [doi: 10.3928/01913913-20210108-02] [Medline: 34039154]
20. Meqdad Y, El-Basty M, Awadein A, Gouda J, Hassanein D. Randomized controlled trial of patching versus dichoptic stimulation using virtual reality for amblyopia therapy. Curr Eye Res 2024 Feb;49(2):214-223. [doi: 10.1080/02713683.2023.2275531] [Medline: 37878538]
21. Suttle CM, Melmoth DR, Finlay AL, Sloper JJ, Grant S. Eye-hand coordination skills in children with and without amblyopia. Invest Ophthalmol Vis Sci 2011 Mar;52(3):1851-1864. [doi: 10.1167/iovs.10-6341] [Medline: 21212188]
22. Grant S, Suttle C, Melmoth DR, Conway ML, Sloper JJ. Age- and stereovision-dependent eye-hand coordination deficits in children with amblyopia and abnormal binocularity. Invest Ophthalmol Vis Sci 2014 Aug 5;55(9):5687-57015. [doi: 10.1167/iovs.14-14745] [Medline: 25097239]
23. Nakazawa T, Okubo Y, Suwazono Y, et al. Association between duration of daily VDT use and subjective symptoms. Am J Ind Med 2002 Nov;42(5):421-426. [doi: 10.1002/ajim.10133] [Medline: 12382255]
24. Sheedy J, Bergstrom N. Performance and comfort on near-eye computer displays. Optom Vis Sci 2002 May;79(5):306-312. [doi: 10.1097/00006324-200205000-00010] [Medline: 12035988]
25. Hoffman DM, Girshick AR, Akeley K, Banks MS. Vergence-accommodation conflicts hinder visual performance and cause visual fatigue. J Vis 2008 Mar 28;8(3):33. [doi: 10.1167/8.3.33] [Medline: 18484839]
26. Hirota M, Yada K, Morimoto T, et al. Objective evaluation of visual fatigue in patients with intermittent exotropia. PLoS ONE 2020;15(3):e0230788. [doi: 10.1371/journal.pone.0230788] [Medline: 32214343]
27. Hirota M, Kanda H, Endo T, et al. Comparison of visual fatigue caused by head-mounted display for virtual reality and two-dimensional display using objective and subjective evaluation. Ergonomics 2019 Jun;62(6):759-766. [doi: 10.1080/00140139.2019.1582805] [Medline: 30773103]
28. Hirota M, Morimoto T, Kanda H, et al. Objective evaluation of visual fatigue using binocular fusion maintenance. Transl Vis Sci Technol 2018 Mar;7(2):9. [doi: 10.1167/tvst.7.2.9] [Medline: 29600117]

XSL•FO
RenderX

29. Leal-Vega L, et al. NEIVATECH pilot study: immersive virtual reality training in older amblyopic children with non-compliance or non-response to patching. Sci Rep 2024 Nov 14;14(1):28062. [doi: 10.1038/s41598-024-79565-y] [Medline: 39543340]
30. Tsubota K, Nakamori K. Dry eyes and video display terminals. N Engl J Med 1993 Feb 25;328(8):584. [doi: 10.1056/NEJM199302253280817] [Medline: 8426634]
31. Chang E, Kim HT, Yoo B. Virtual reality sickness: a review of causes and measurements. Int J Hum-Comput Interact 2020 Oct 20;36(17):1658-1682. [doi: 10.1080/10447318.2020.1778351]
32. Jackson DN, Bedell HE. Vertical heterophoria and susceptibility to visually induced motion sickness. Strabismus 2012 Mar;20(1):17-23. [doi: 10.3109/09273972.2011.650813] [Medline: 22390327]
33. Briggs AM, Smith AJ, Straker LM, Bragge P. Thoracic spine pain in the general population: prevalence, incidence and associated factors in children, adolescents and adults. A systematic review. BMC Musculoskelet Disord 2009 Jun 29;10(1):77. [doi: 10.1186/1471-2474-10-77] [Medline: 19563667]
34. Rodríguez-Oviedo P, Ruano-Ravina A, Pérez-Ríos M, et al. School children's backpacks, back pain and back pathologies. Arch Dis Child 2012 Aug;97(8):730-732. [doi: 10.1136/archdischild-2011-301253] [Medline: 22408188]
35. Ghai S, Ghai I. Virtual reality enhances gait in cerebral palsy: a training dose-response meta-analysis. Front Neurol 2019;10:236. [doi: 10.3389/fneur.2019.00236] [Medline: 30984095]

Abbreviations

HMD: head-mounted display
VIMS: visually induced motion sickness
VR: virtual reality

Edited by A Mavragani; submitted 09.12.24; peer-reviewed by M Piano, Y Wenner; revised version received 14.04.25; accepted 09.05.25; published 17.06.25.

Please cite as:
Hirota M, Okumura Y, Nagino K, Hayashi T, Negishi T, Nakao S, Kawasaki H, Inomata T
Safety Evaluation in Healthy Adults of Motion-Based Virtual Reality Dichoptic Training for Pediatric Patients With Amblyopia: Prospective Intervention Study
JMIR Form Res 2025;9:e69801
URL: https://formative.jmir.org/2025/1/e69801
doi:10.2196/69801

© Masakazu Hirota, Yuichi Okumura, Ken Nagino, Takao Hayashi, Takashi Negishi, Shintaro Nakao, Hitoshi Kawasaki, Takenori Inomata. Originally published in JMIR Formative Research (https://formative.jmir.org), 17.6.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Feasibility of Collecting and Linking Digital Phenotyping, Clinical, and Genetics Data for Mental Health Research: Pilot Observational Study

Joanne R Beames[1,2], PhD; Omar Dabash[1], PhD; Michael J Spoelma[1], BPsych (Hons); Artur Shvetcov[1], PhD; Wu Yi Zheng[1], PhD; Aimy Slade[1], PhD; Jin Han[1,3], PhD; Leonard Hoon[4], PhD; Joost Funke Kupper[4], BSc; Richard Parker[5], BA (Hons); Brittany Mitchell[5], PhD; Nicholas G Martin[5], PhD; Jill M Newby[1,6], PhD; Alexis E Whitton[7], PhD; Helen Christensen[1], PhD

[1]Black Dog Institute, University of New South Wales, Sydney, Australia

[2]Department of Neurosciences, Center for Contextual Psychiatry, KU Leuven, Leuven, Belgium

[3]Division of Arts and Sciences and Centre for Global Health Equity, New York University Shanghai, Shanghai, China

[4]Applied Artificial Intelligence Institute, Deakin University, Melbourne, Australia

[5]Brain and Mental Health Program, QIMR Berghofer Institute of Medical Research, Brisbane, Australia

[6]School of Psychology, University of New South Wales, Sydney, Australia

[7]Black Dog Institute, University of New South Wales, Hospital Road, Randwick, Australia

Corresponding Author:
Alexis E Whitton, PhD
Black Dog Institute, University of New South Wales, Hospital Road, Randwick, Australia

Abstract

Background: Digital phenotyping—the use of digital data to measure and understand behavior and internal states—shows promise for advancing predictive analytics in mental health, particularly when combined with other data sources. However, linking digital phenotyping data with sources of highly sensitive clinical or genetic data remains rare, primarily due to technical, ethical, and procedural challenges. Understanding the feasibility of collecting and linking these data types is a critical first step toward developing novel multimodal datasets.

Objective: The Mobigene Pilot Study examines the feasibility of collecting smartphone-based digital phenotyping and mental health data and linking it to genetic data from an existing cohort of adults with a history of depression (ie, the Australian Genetics of Depression Study). This paper aims to describe (1) rates of study uptake and adherence; (2) levels of adherence and engagement with daily mood assessments; (3) willingness to take part in similar research; and (4) whether feasibility indicators varied according to mental health symptoms.

Methods: Participants aged 18‑30 years with genetic data from the Australian Genetics of Depression Study were invited to participate in a two-week digital phenotyping study. They completed a baseline mental health survey and then downloaded the MindGRID digital phenotyping app. Active data from cognitive, voice, and typing tasks were collected once per day on days 1 and 11. Daily momentary assessments of self-reported mood were collected on days 2‑10 (once per day for 9 days). Passive data (eg, from GPS, accelerometers) were collected throughout the two-week period. A second mental health survey was then completed after two weeks. To measure feasibility, we examined metrics of study uptake (eg, consent) and adherence (eg, proportion of completed momentary assessments), and willingness to participate in similar future research. Pearson correlations and *t* tests explored the relationship between feasibility indicators and mental health symptoms.

Results: Of 174 consenting and eligible participants, 153 (87.9%) completed the baseline mental health survey and 126 (72.4%) provided data enabling linkage of genetic, self-report, and digital data. After removal of duplicates, we found that 100 (57.5%) of these identified as unique participants and 69 (39.7%) provided complete post-study data. A small proportion of participants dropped out prior to completing the baseline survey (21/174, 12.1%) or during app-based data collection (31/174, 17.8%). Participants completed an average of 5.30 (SD 2.76) daily mood assessments. All 69 (100%) participants who completed the post-study surveys expressed willingness to participate in similar studies in the future. There was no significant association between feasibility indicators and current mental health symptoms.

Conclusions: It is feasible to collect and link multimodal datasets involving digital phenotyping, clinical, and genetic data, although there are some methodological and technical challenges. We provide recommendations for future research related to data collection platforms and compliance.

XSL•FO
RenderX

(JMIR Form Res 2025;9:e71377) doi:10.2196/71377

KEYWORDS

data linkage; precision medicine; experience sampling methodology; daily diary; depression; anxiety; anhedonia; suicidal ideation

Introduction

There has been increasing investment in the collection and integration of multimodal datasets to facilitate the development of more precise methods for detecting and predicting complex symptom dynamics in mental health [1,2]. One promising area of inquiry is the integration of smartphone–collected digital phenotyping data with genetic data [3,4].

Digital phenotyping is the process of using digital data to measure and understand behavior and internal states [5,6]. Digital data can be collected passively, for example via accelerometers, or actively, such as through cognitive tasks and the Experience Sampling Method (ESM) [7]. The ESM uses repeated self-report surveys to assess experiences in daily life and has minimal recall biases and high ecological validity [8,9]. Daily diaries are a special form of ESM in which assessments only occur once per day, typically within a prespecified time window [8]. There is evidence that digital phenotyping data can correlate with, classify, and predict mental health problems including depression and anxiety [10-17].

To our knowledge, no previous studies have linked newly collected smartphone-based digital phenotyping data with existing genetic data in mental health research. Creating such a dataset could enhance our understanding of how genetic, behavioural, cognitive, and psychological factors interact in mental health [3,4], offering the potential for better prediction and clinical decision-making [18,19].

The Mobigene Pilot Study

The purpose of the Mobigene pilot study is to explore the feasibility of collecting digital phenotyping data (ie, from cognitive, voice, and typing tasks; daily diaries; and passive sensors) and self-reported clinical data, from participants in the Australian Genetics of Depression Study (AGDS). The AGDS is an ongoing study examining the contribution of genetic variation to risk of depression in a large cohort of Australian adults [20]. The aims of this paper are to (1) describe rates of study uptake and adherence to the Mobigene study protocol; (2) describe levels of adherence and engagement with daily diary assessments; (3) identify openness to participate in similar research in the future (ie, sustainability); and (4) determine whether feasibility indicators differ according to participants' current mental health symptoms.

Transparency and Openness

The postregistered analysis plan [21], deviations, and code are available on the Open Science Framework project page for this study [22].

Methods

Design and Procedures

The Mobigene study is an observational pilot study that combines primary and secondary data. Primary data include self-reported current mental health symptoms, active and passive digital phenotyping, and app or phone metadata. Secondary data include existing genetic and historical clinical data collected through the AGDS.

In total, 1282 participants aged 18 - 30 years who had provided genetic data in the AGDS were invited by the AGDS team to participate. Interested participants completed informed consent and eligibility screening. Eligibility criteria included being an Australian resident, owning an iOS smartphone, willingness to participate, and proficiency in English. Eligible participants completed baseline demographic and mental health surveys via Qualtrics and were then invited to download a purpose-built iOS smartphone app for digital phenotyping (MindGRID) on their own smartphones. On days 1 and 11, participants accessed cognitive, voice, and typing tasks in the app. Participants were given 3-days to complete the tasks and each task could only be completed once. From days 2 - 10, participants completed a 9-day daily diary period which combined interval- and event-based sampling. In line with best practices [8], for the interval-contingent scheme, participants received one assessment notification from the app per day, sent in the evening at a random time between 7 PM-9 PM. On days 12 - 15, depending on the timeliness of task completion, postsurveys assessing mental health and perceptions about the study were readministered via the app. See Figure 1 for study procedures.

Data were collected over two participant cohorts with differing incentive schemes. Participants from cohort 1 could re-enroll in cohort 2 for the opportunity to receive the increased incentive. Duplicate records from cohort 2 were identified and excluded from analyses. The recruitment target was 300 participants and accounted for potential dropout. See Multimedia Appendix 1 for details on eligibility, recruitment, informed consent, and the study protocol (including app-based tasks).

XSL•FO
RenderX

Figure 1. Overview of study procedures including online consent, screening, the baseline survey, and digital phenotyping. ESM: Experience Sampling Methodology.

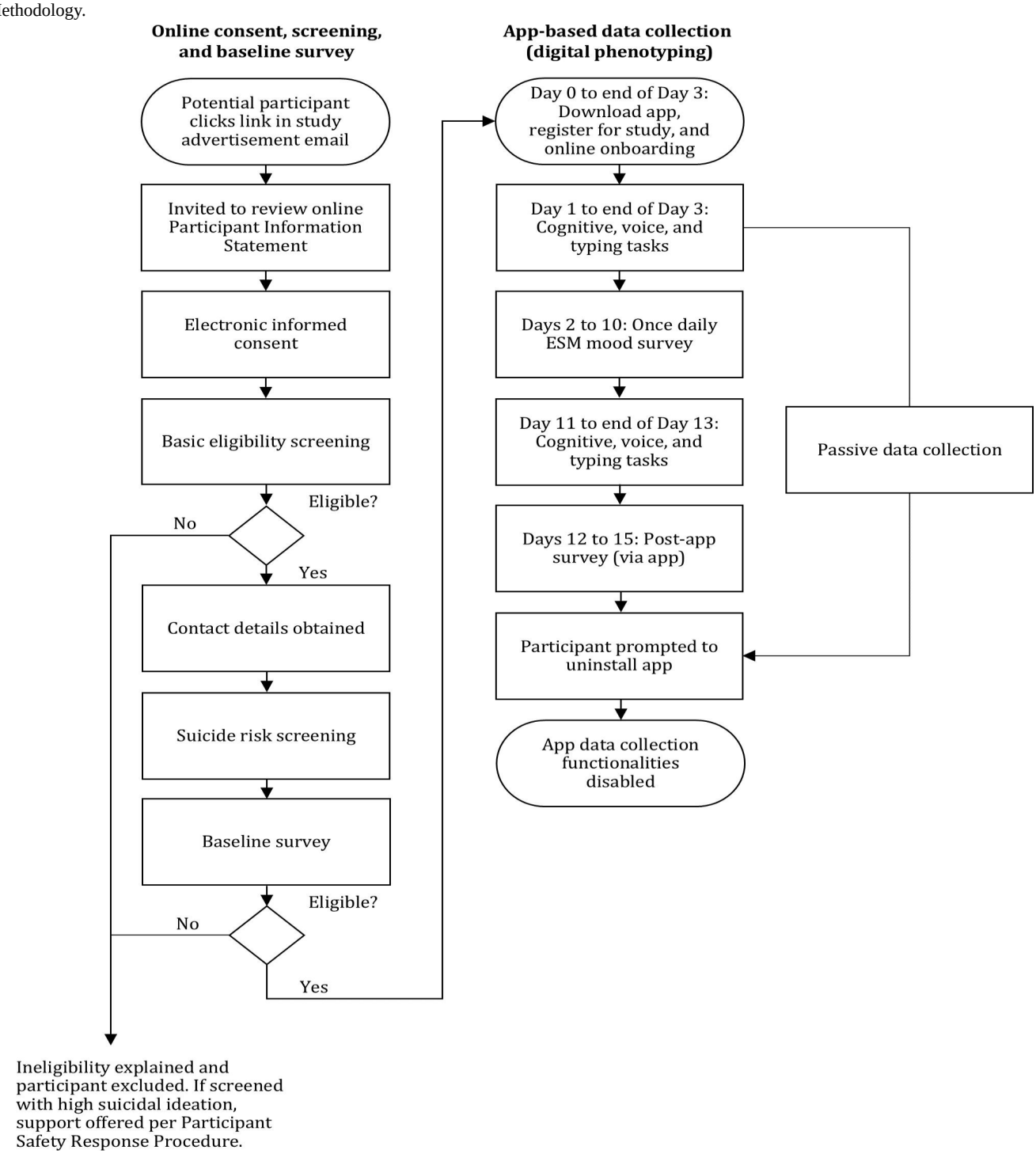

Ethical Considerations

Ethical approval was obtained from the University of New South Wales, Sydney Human Research Ethics Committee (HC220228) and QIMR Berghofer Human Research Ethics Committee (P3476). Participants provided online informed consent and all study data were deidentified. For compensation, participants were either entered into a draw for 10 Aus $100 (US $65) e-gift cards (cohort 1) or received an Aus $50 (US $33) e-gift card (cohort 2).

Data Linkage

Three unique identifiers were randomly assigned to each participant: one upon receipt of the study invitation; one during screening; and one upon confirmation of eligibility. These identifiers enabled linkage between baseline and digital phenotyping data for participants from AGDS. See Multimedia Appendix 1 for further details about linkage and privacy.

Study Uptake and Adherence

Study uptake was defined as the number of eligible participants who consented to participate, the number (%) of participants whose data could be successfully linked, and the number (%)

XSL•FO
RenderX

of duplicates. Study adherence was defined as the number (%) of participants that completed baseline and post self-report surveys, as well as the number (%) of participants that dropped out during the study.

Sample Characteristics

Baseline self-report data included age, education, relationship status, number of children, and current mental health symptoms and treatment. Standardized scales included the Suicide Ideation Attributes Scale [23], Patient Health Questionnaire 9-item [24], Generalized Anxiety Disorder 7-item Questionnaire [25], Snaith-Hamilton Pleasure Scale [26], and Short Health Anxiety Inventory [27]. Total scores and cut-off points for clinical severity were computed (See Table S1 in Multimedia Appendix 1 for internal consistency) [23-31].

Daily Diary Adherence and Engagement

MindGRID metadata were used to compute adherence to the daily diary protocol. We computed three indices of adherence: the average number (%) of completed evening assessments, the number (%) of participants that did not complete any diaries, and the number (%) of dropouts during the diary period. Engagement was defined as the average number of completed evening and event-based assessments.

Sustainability

Self-reported openness to take part in a similar future study was assessed postsurvey.

Results

Study Uptake and Adherence

A total of 1,292 participants from the AGDS study were invited to participate. Of the invited participants, 174 eligible participants consented to participate; 21 (12.1%) participants dropped out, leaving 153 (87.9%) who completed the baseline survey. Of these, 126 (72.4%) participants had successfully linked data. Linkage failures were attributed to technical errors: (1) an identifier was not appropriately generated during eligibility; or (2) the eligibility identifier was not recorded due to a bug in the cognitive tasks (ie, if tasks were not completed or timed out by the end of day 3, the identifier was not recorded). After removing 26 duplicates (all from cohort 2), 100 (57.5%) unique participants remained. Thirty-one (17.8%) participants dropped out during the app data collection phase; there were no significant differences in dropout rates between the recruitment cohorts (see Table S2 in the Multimedia Appendix 1). Overall, 69 (39.7%) unique participants with linked data completed postsurveys. The flow diagram is depicted in Figure 2.

XSL•FO
RenderX

Figure 2. Participant flow diagram illustrating the number of participants at each stage, from initial contact through to study completion. ESM: Experience Sampling Methodology; Generalized Anxiety Disorder 7-item Questionnaire; Patient Health Questionnaire 9-item; Short Health Anxiety Inventory; SIDAS: Suicide Ideation Attributes Scale.

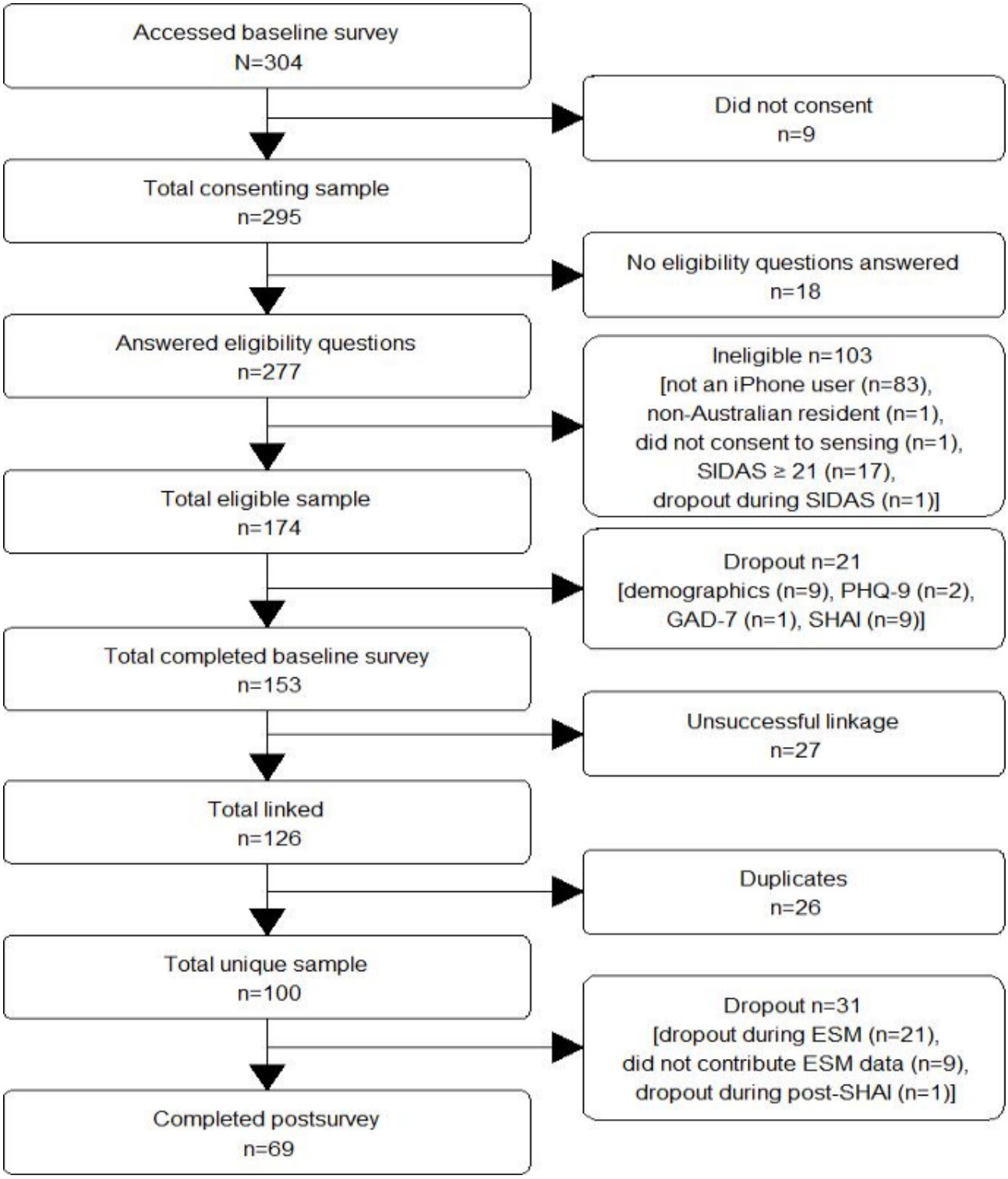

Sample Characteristics

Baseline sample characteristics are reported in Table 1.

XSL•FO
RenderX

Table . Baseline demographic and mental health characteristics of the total unique sample (n=100) and final postsurvey sample (n=69).

	Total unique sample[a] (n=100)	Final unique sample (n=69)
Age, mean (SD; range)	27.09 (2.05; 20 - 30)	27.13 (2.15; 20 - 30)
Education, n (%)		
High school	34 (34.0)	24 (34.8)
Undergraduate	40 (40.0)	27 (39.1)
Postgraduate	26 (26.0)	18 (26.1)
Relationship status, n *(%)*		
Single	39 (39.0)	26 (37.7)
In a relationship	26 (26.0)	19 (27.5)
De facto	18 (18.0)	14 (20.3)
Married	17 (17.0)	10 (14.5)
Number of children, mean (SD*;* range)[b]	0.21 (0.99; 0 - 9)	0.13 (0.42; 0 - 2)
Mental health symptoms (continuous), mean (SD*;* range)		
Suicidal ideation	5.42 (6.26); 0 - 20	4.96 (5.99); 0 - 20
Depression	10.99 (5.34); 0 - 22	11.07 (5.23); 2 - 22
Generalised anxiety	8.95 (5.09); 0 - 21	9.32 (5.21); 1 - 21
Anhedonia	2.93 (3.15); 0 - 13	2.84 (3.24); 0 - 13
Health anxiety[c]	20.18 (8.31); 3 - 43	20.15 (8.37); 3 - 39
Mental health symptoms (≥ clinical cut-off), n (%)[d]		
Suicidal ideation (≥1)	66 (66.0)	45 (65.2)
Depression (≥10)	61 (61.0)	42 (60.9)
Generalised anxiety (≥8)	52 (52.0)	36 (52.2)
Anhedonia (≥3)	45 (45.0)	29 (42.0)
Currently taking prescribed medications for mental health, n (%)		
Yes	68 (68.0)	47 (68.1)
No	32 (32.0)	22 (31.9)
Currently receiving psychological therapy for mental health, n (%)		
Yes	50 (50.0)	35 (50.7)
No	50 (50.0)	34 (49.3)

[a]Total unique sample refers to the total sample after removal of duplicate records.

[b]n=2 missing.

[c]n=1 missing.

[d]There is no established cut-off score for the Short Health Anxiety Inventory [32,33].

Daily Diary Adherence and Engagement

The average number of evening assessments completed (out of 9) was 5.30 (SD 2.76; range 0 - 9; equivalent to 58.9%). The average number of completed evening and event-based assessments was 7.21 (SD 3.79, range 0 - 19). Most app-phase dropouts occurred during the diary period (21/31, 67.7%); 9 (29.0%) participants did not complete any diaries. The average diary adherence and engagement was comparable across recruitment cohorts (see Tables S3 and S4 in Multimedia Appendix 1).

Sustainability

All participants who completed the postsurvey (69/69, 100.0%) reported that they would participate in a similar study in the future.

Current Mental Health Symptoms and Feasibility Indicators

There were no significant associations between symptoms and dropout (all $t<1.12$; all $P>.05$) or diary adherence and engagement (all $r<0.19$; all $P>.05$) (see Tables S5 and S6 in Multimedia Appendix 1 for statistics).

XSL•FO
RenderX

Discussion

Principal Findings

The Mobigene pilot study demonstrated that it was feasible to collect and link new data from an existing cohort that had already participated in extensive data collection procedures. Uptake and adherence to the study protocol were promising; most participants completed at least some daily diaries and there was evidence supporting sustainability. No associations were observed between current mental health symptoms and dropout, daily diary adherence, or engagement.

However, there were methodological and technical challenges during different study phases. Most (n=83) of the ineligible participants did not have an iOS smartphone to support the MindGRID app. Further, technical errors prevented linkage of app data to baseline data for some participants (n=27), and dropout was most common during the digital phenotyping phase (n=31). It is unclear whether dropout was due to issues with daily diaries (eg, timing of assessments), cognitive tasks, or other design-related factors, although it likely reflects the intensive nature of the multimodal data collection protocol. Gathering follow-up data on participants' subjective experiences of the study, and reasons for noncompletion or dropout is crucial for fully assessing feasibility. Uptake and adherence rates were broadly aligned with findings from other studies using remote monitoring technologies in individuals with a history of depression [34-36]. However, comparisons are challenging given the lack of uptake and adherence standards in intensive digital phenotyping protocols and differences in study design.

Limitations

First, small sample size limits generalizability of the findings. Second, our estimate of sustainability is likely to be inflated due to self-selection bias, as the 69 participants who completed the postsurvey may represent a particularly motivated subgroup. Third, a large participant pool was required to recruit the final sample.

Conclusions

Using multimodal data, including integration of existing datasets, is a novel approach to advance mental health prediction and minimize research waste. Our findings indicate that future studies should use a cross-platform data collection apps (eg, both iOS and Android), consolidate data collection via a single mobile platform, and implement adaptive incentives such as gamification to increase compliance [37]. This approach would improve data quality and quantity by broadening eligibility, minimizing technical linkage errors, and increasing participant engagement with data collection procedures.

Acknowledgments

This work was supported by Commonwealth of Australia Medical Research Future Fund grant (MRFAI000028)–Optimising treatments in mental health using AI. JRB received funding through a Marie Skłodowska-Curie fellowship (101063326). AW, JN, HC and BM were supported by NHMRC Investigator Grants (AW: 2017521; JN: 2008839; HC: GNT1155614). The funding bodies had no role in any aspect of the study design or this manuscript preparation.

Data Availability

The datasets generated or analyzed during this study are not publicly available given their sensitivity and concerns around privacy but are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary materials.
[DOCX File, 4145 KB - formative_v9i1e71377_app1.docx]

References

1. Kambeitz-Ilankovic L, Koutsouleris N, Upthegrove R. The potential of precision psychiatry: What is in reach? Br J Psychiatry 2022 Apr;220(4):175-178. [doi: 10.1192/bjp.2022.23] [Medline: 35354501]
2. Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ. Multimodal biomedical AI. Nat Med 2022 Sep;28(9):1773-1784. [doi: 10.1038/s41591-022-01981-2] [Medline: 36109635]
3. Montag C, Quintana DS. Digital phenotyping in molecular psychiatry-a missed opportunity? Mol Psychiatry 2023 Jan;28(1):6-9. [doi: 10.1038/s41380-022-01795-1] [Medline: 36171355]
4. Montag C, Dagum P, Hall BJ, Elhai JD. How the study of digital footprints can supplement research in behavioral genetics and molecular psychology. Mol Psychol 2022;1(2):2. [doi: 10.12688/molpsychol.17401.1]
5. Huckvale K, Venkatesh S, Christensen H. Toward clinical digital phenotyping: a timely opportunity to consider purpose, quality, and safety. NPJ Digit Med 2019;2(1):88. [doi: 10.1038/s41746-019-0166-1] [Medline: 31508498]
6. Oudin A, Maatoug R, Bourla A, et al. Digital phenotyping: data-driven psychiatry to redefine mental health. J Med Internet Res 2023 Oct 4;25:e44502. [doi: 10.2196/44502] [Medline: 37792430]

XSL•FO
RenderX

7. Torous J, Kiang MV, Lorme J, Onnela JP. New tools for new research in psychiatry: a scalable and customizable platform to empower data driven smartphone research. JMIR Ment Health 2016 May 5;3(2):e16. [doi: 10.2196/mental.5165] [Medline: 27150677]
8. Myin-Germeys I, Kuppens P, editors. The open handbook of experience sampling methodology: a step-by-step guide to designing, conducting, and analyzing ESM studies, 2nd edition: Leuven: Center for Research on Experience Sampling and Ambulatory Methods Leuven; 2022. URL: https://www.kuleuven.be/samenwerking/real/real-book [accessed 2025-06-11]
9. Myin-Germeys I, Kasanova Z, Vaessen T, et al. Experience sampling methodology in mental health research: new insights and technical developments. World Psychiatry 2018 Jun;17(2):123-132. [doi: 10.1002/wps.20513] [Medline: 29856567]
10. Marciano L, Vocaj E, Bekalu MA, La Tona A, Rocchi G, Viswanath K. The use of mobile assessments for monitoring mental health in youth: umbrella review. J Med Internet Res 2023 Sep 19;25:e45540. [doi: 10.2196/45540] [Medline: 37725422]
11. Moura I, Teles A, Viana D, Marques J, Coutinho L, Silva F. Digital phenotyping of mental health using multimodal sensing of multiple situations of interest: a systematic literature review. J Biomed Inform 2023 Feb;138:104278. [doi: 10.1016/j.jbi.2022.104278] [Medline: 36586498]
12. Bufano P, Laurino M, Said S, Tognetti A, Menicucci D. Digital phenotyping for monitoring mental disorders: systematic review. J Med Internet Res 2023 Dec 13;25:e46778. [doi: 10.2196/46778] [Medline: 38090800]
13. Beames JR, Han J, Shvetcov A, et al. Use of smartphone sensor data in detecting and predicting depression and anxiety in young people (12–25 years): a scoping review. Heliyon 2024 Aug;10(15):e35472. [doi: 10.1016/j.heliyon.2024.e35472]
14. Choi A, Ooi A, Lottridge D. Digital phenotyping for stress, anxiety, and mild depression: systematic literature review. JMIR Mhealth Uhealth 2024 May 23;12:e40689. [doi: 10.2196/40689] [Medline: 38780995]
15. dos Santos MP, Heckler WF, Bavaresco RS, Barbosa JLV. Machine learning applied to digital phenotyping: a systematic literature review and taxonomy. Comput Human Behav 2024 Dec;161:108422. [doi: 10.1016/j.chb.2024.108422]
16. Khoo LS, Lim MK, Chong CY, McNaney R. Machine learning for multimodal mental health detection: a systematic review of passive sensing approaches. Sensors (Basel) 2024 Jan 6;24(2):348. [doi: 10.3390/s24020348] [Medline: 38257440]
17. Leaning IE, Ikani N, Savage HS, et al. From smartphone data to clinically relevant predictions: a systematic review of digital phenotyping methods in depression. Neuroscience & Biobehavioral Reviews 2024 Mar;158:105541. [doi: 10.1016/j.neubiorev.2024.105541]
18. Murray GK, Lin T, Austin J, McGrath JJ, Hickie IB, Wray NR. Could polygenic risk scores be useful in psychiatry?: A review. JAMA Psychiatry 2021 Feb 1;78(2):210-219. [doi: 10.1001/jamapsychiatry.2020.3042] [Medline: 33052393]
19. Verhagen S, Os J, Delespaul P. Ecological momentary assessment and other digital technologies for capturing daily life in mental health. In: Stein DJ, Fineberg NA, editors. Mental Health in a Digital World: Academic Press; 2022:81-108. [doi: 10.1016/B978-0-12-822201-0.00017-4]
20. Byrne EM, Kirk KM, Medland SE, et al. Cohort profile: the Australian genetics of depression study. BMJ Open 2020 May 26;10(5):e032580. [doi: 10.1136/bmjopen-2019-032580] [Medline: 32461290]
21. Benning SD, Bachrach RL, Smith EA, Freeman AJ, Wright AGC. The registration continuum in clinical science: a guide toward transparent practices. J Abnorm Psychol 2019 Aug;128(6):528-540. [doi: 10.1037/abn0000451] [Medline: 31368732]
22. Beames JR. The feasibility of creating novel multimodal datsets in mental health research. The Open Science Framework. 2025. URL: https://osf.io/z5vnt/ [accessed 2025-06-11]
23. van Spijker BAJ, Batterham PJ, Calear AL, et al. The suicidal ideation attributes scale (SIDAS): community-based validation study of a new scale for the measurement of suicidal ideation. Suicide Life Threat Behav 2014 Aug;44(4):408-419. [doi: 10.1111/sltb.12084] [Medline: 24612048]
24. Kroenke K, Spitzer RL. The PHQ-9: a new depression diagnostic and severity measure. Psychiatr Ann 2002 Sep;32(9):509-515. [doi: 10.3928/0048-5713-20020901-06]
25. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
26. Snaith RP, Hamilton M, Morley S, Humayan A, Hargreaves D, Trigwell P. A scale for the assessment of hedonic tone the Snaith-Hamilton Pleasure Scale. Br J Psychiatry 1995 Jul;167(1):99-103. [doi: 10.1192/bjp.167.1.99] [Medline: 7551619]
27. Salkovskis PM, Rimes KA, Warwick HMC, Clark DM. The Health Anxiety Inventory: development and validation of scales for the measurement of health anxiety and hypochondriasis. Psychol Med 2002 Jul;32(5):843-853. [doi: 10.1017/s0033291702005822] [Medline: 12171378]
28. Manea L, Gilbody S, McMillan D. Optimal cut-off score for diagnosing depression with the Patient Health Questionnaire (PHQ-9): a meta-analysis. CMAJ 2012 Feb 21;184(3):E191-E196. [doi: 10.1503/cmaj.110829] [Medline: 22184363]
29. Plummer F, Manea L, Trepel D, McMillan D. Screening for anxiety disorders with the GAD-7 and GAD-2: a systematic review and diagnostic metaanalysis. Gen Hosp Psychiatry 2016;39:24-31. [doi: 10.1016/j.genhosppsych.2015.11.005] [Medline: 26719105]
30. Trøstheim M, Eikemo M, Meir R, et al. Assessment of anhedonia in adults with and without mental illness: a systematic review and meta-analysis. JAMA Netw Open 2020 Aug 3;3(8):e2013233. [doi: 10.1001/jamanetworkopen.2020.13233] [Medline: 32789515]

XSL•FO
RenderX

31. Abramowitz JS, Deacon BJ, Valentiner DP. The short health anxiety inventory: psychometric properties and construct validity in a non-clinical sample. Cognit Ther Res 2007;31(6):871-883. [doi: 10.1007/s10608-006-9058-1] [Medline: 32214558]
32. Alberts NM, Hadjistavropoulos HD, Jones SL, Sharpe D. The Short Health Anxiety Inventory: a systematic review and meta-analysis. J Anxiety Disord 2013 Jan;27(1):68-78. [doi: 10.1016/j.janxdis.2012.10.009] [Medline: 23247202]
33. Kikas K, Werner-Seidler A, Upton E, Newby J. Illness anxiety disorder: a review of the current research and future directions. Curr Psychiatry Rep 2024 Jul;26(7):331-339. [doi: 10.1007/s11920-024-01507-2] [Medline: 38748190]
34. de Angel V, Adeleye F, Zhang Y, et al. The feasibility of implementing remote measurement technologies in psychological treatment for depression: mixed methods study on engagement. JMIR Ment Health 2023 Jan 24;10:e42866. [doi: 10.2196/42866] [Medline: 36692937]
35. Matcham F, Leightley D, Siddi S, et al. Remote Assessment of Disease and Relapse in Major Depressive Disorder (RADAR-MDD): recruitment, retention, and data availability in a longitudinal remote measurement study. BMC Psychiatry 2022 Feb 21;22(1):136. [doi: 10.1186/s12888-022-03753-1] [Medline: 35189842]
36. Oetzmann C, White KM, Ivan A, et al. Lessons learned from recruiting into a longitudinal remote measurement study in major depressive disorder. NPJ Digit Med 2022 Sep 3;5(1):133. [doi: 10.1038/s41746-022-00680-z] [Medline: 36057688]
37. Dejonckheere E, Verdonck S, Andries J, et al. Real-time incentivizing survey completion with game-based rewards in experience sampling research may increase data quantity, but reduces data quality. Comput Human Behav 2024 Nov;160:108360. [doi: 10.1016/j.chb.2024.108360]

Abbreviations

AGDS: Australian Genetics of Depression Study
ESM: Experience Sampling Methodology

Edited by A Mavragani; submitted 26.01.25; peer-reviewed by AM Triana, R Chen; revised version received 14.04.25; accepted 11.05.25; published 23.06.25.

Please cite as:
Beames JR, Dabash O, Spoelma MJ, Shvetcov A, Zheng WY, Slade A, Han J, Hoon L, Kupper JF, Parker R, Mitchell B, Martin NG, Newby JM, Whitton AE, Christensen H
Feasibility of Collecting and Linking Digital Phenotyping, Clinical, and Genetics Data for Mental Health Research: Pilot Observational Study
JMIR Form Res 2025;9:e71377
URL: https://formative.jmir.org/2025/1/e71377
doi:10.2196/71377

© Joanne R Beames, Omar Dabash, Michael J Spoelma, Artur Shvetcov, Wu Yi Zheng, Aimy Slade, Jin Han, Leonard Hoon, Joost Funke Kupper, Richard Parker, Brittany Mitchell, Nicholas G Martin, Jill M Newby, Alexis E Whitton, Helen Christensen. Originally published in JMIR Formative Research (https://formative.jmir.org), 23.6.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Analyzing Trends in Suicidal Thoughts Among Patients With Psychosis in India: Exploratory Secondary Analysis of Smartphone Ecological Momentary Assessment Data

Ameya P Bondre[1*], MSPH; Aashish Ranjan[1*], MSc; Ritu Shrivastava[1*], MSW; Deepak Tugnawat[1*], PGDPHM; Nirmal Kumar Chaturvedi[1*], MTech; Anant Bhan[1*], MHSc; Snehil Gupta[2*], MD; Abhijit R Rozatkar[2*], MD; Srilakshmi Nagendra[3*], BE; Siddharth Dutt[3*], PhD PDF; Soumya Choudhary[3*], MSc; Preethi V Reddy[3*], MD PhD; Urvakhsh Meherwan Mehta[3*], MD PhD; John A Naslund[4*], PhD; John Torous[5*], MD

[1]Sangath, 106, Good Shepherd Colony, Kolar Rd, Bhopal, Madhya Pradesh, India

[2]Department of Psychiatry, All India Institute of Medical Sciences Bhopal, Bhopal, Madhya Pradesh, India

[3]Department of Psychiatry, National Institute of Mental Health and Neurosciences, Bengaluru, Karnataka, India

[4]Department of Global Health and Social Medicine, Harvard Medical School, Boston, MA, United States

[5]Division of Digital Psychiatry, Beth Israel Deaconess Medical Center, Boston, MA, United States

[*]all authors contributed equally

Corresponding Author:
Ameya P Bondre, MSPH
Sangath, 106, Good Shepherd Colony, Kolar Rd, Bhopal, Madhya Pradesh, India

Abstract

Background: India has the world's largest number of suicides, but there is little research on the trends in suicidal thoughts, especially for individuals with psychosis. More research is necessary to develop preventive interventions. Smartphone-based ecological momentary assessments (EMAs) can assess dynamic symptoms, but most EMA studies are conducted in higher income settings and have shorter (≤1 month) follow-up periods.

Objective: This study aimed to examine the duration of onset to offset of suicidal ideation (SI) in tertiary hospital outpatients with psychosis in India.

Methods: This study is an exploratory, secondary analysis of smartphone EMA data nested within the ongoing "Smartphone Health Assessment for Relapse Prevention (SHARP)" project. Tertiary hospital outpatients (n=50) with early course schizophrenia at 2 socioculturally different sites in India were recruited and given the "mindLAMP" app for monitoring mood through daily EMA surveys. The mood survey matched the 9-item Patient Health Questionnaire; the ninth item was used to define an instance of SI (score ≥1). A total of 14 patients with ≥1 SI instances who met the site-specific EMA survey use cutoff were included. We examined the between- and within-person variability in SI and computed the timescale of "episodic" SI (sequences of consecutive daily observations of SI score ≥1). Positive and Negative Syndrome Scale (PANSS) was used to assess changes in psychosis symptoms and its relationship with the temporality of SI.

Results: Over approximately 11 (SD 2.1) months of EMA reporting on average, 3253 mood surveys were filled by the 14 participants (median 213, IQR 147 - 256). A total of 521 instances of SI were reported. Monthly SI instances showed substantial within- and between-person variations. Timescale summary statistics revealed episodic SI patterns in 11 patients, with an average of 5.9 episodes (SD 4.4; range:1 - 14; n=65) with an episode lasting on average 2.5 days (SD 1.5; range:1 - 5.3; n=27). There was an average lag of approximately 59, 66, and 81 days between the time of the first drop in PANSS positive, negative, and general psychopathology scores, respectively, and the last reported SI instance. Results after imputation of missing data showed an average of 12.1 episodes and 228 days (average lag) between the first drop in PANSS scores and last reported SI. This indicated that SI was an enduring vulnerability subsequent to the beginning of clinical improvement in psychosis.

Conclusions: Our study adds to the much-needed evidence base in India to measure the dynamics of suicidal thinking within an individual, for more targeted preventive interventions. Further steps in EMA research are highlighted such as the use of higher frequency "burst" surveys to assess the duration of an SI episode in hours or minutes, and inclusion of both active and passive SI markers to measure the timescale of suicidal thinking.

International Registered Report Identifier (IRRID): RR2-10.1192/bjo.2020.142

(JMIR Form Res 2025;9:e67745) doi:10.2196/67745

XSL•FO
RenderX

KEYWORDS

schizophrenia; suicidality; ecological momentary assessments; India; trends; suicidal; suicidal ideation; suicides; suicidal thoughts; EMA; mobile apps; exploratory study; psychosis; smartphone; tertiary hospitals; outpatients

Introduction

India has the world's largest number of suicidal deaths [1], and for every death by suicide in India, there are more than 200 people with suicidal ideations (SIs) [2]. Suicide is a common cause of premature mortality among people living with schizophrenia [3,4] and a recent systematic review has reported a point prevalence of nearly 30% of SIs in people with schizophrenia [5].

Most of the empirical research on the risks of suicidality includes cross-sectional or retrospective studies that distinguish the characteristics between people who experience SIs or suicidal behavior, and those who do not [6-8]. This is true for Indian hospital-based studies on suicide in schizophrenia [9,10]. Longitudinal studies (without ecological momentary assessments [EMAs]) have also mostly examined the prevalence of, and factors associated with suicidal thoughts and behaviors among individuals with early course schizophrenia, as illustrated in a systematic review of 17 studies [11]. There is some evidence of trajectories of SI such as "low-decreasing," "frequent-stable," and "frequent-increasing" identified among patients with first-episode psychosis, but these studies do not involve frequent assessments (like in EMAs) and the long-duration prospective study designs involve wider-spaced follow-ups (eg, yearly) [12]. EMA studies involving participants with psychosis also often assess the relationships between SI and other relevant variables such as perceived burdensomeness and thwarted belongingness [13]; or social approach, functioning, or isolation in schizophrenia [14]. With respect to temporality of SI in longitudinal designs, a recent review of 26 studies has demonstrated that intensive time sampling approaches are highly capable in detecting SI at daily and hourly increments [15], but this emerging evidence is not specific to patients of psychosis [15,16]. Specifically, we know little about the temporal dynamics of SI in psychosis, in particular how SI changes or evolves over time (minutes, hours, or days) in an individual with psychosis, or when this individual is at risk [7,17], which is critical to understand how suicidal behavior develops, and design preventive interventions [18]. One key aspect is the duration of onset to offset of an episode of suicidal thinking [18]. Smartphone-based EMAs can assess dynamic symptoms such as SI, and help measure its duration, while enabling a safe disclosure of SI without face-to-face contact with the researcher or clinician [19,20]. However, most EMA studies on suicidality, as also EMA studies on day-to-day functioning of people with schizophrenia have been conducted in higher-income countries [21]. Furthermore, the available EMA studies have sampling periods typically lasting for ≤1 month [20], which is inadequate to inform the estimation of the evolution of SI during longer outpatient care periods in India, reducing the clinical use.

This study aimed to explore the feasibility of using EMA to track SI in tertiary hospital outpatients with psychosis in lower-resource settings in India, duration and temporal trends of SI, particularly the estimation of the duration of onset to offset of SI (primary objective) over a substantially long follow-up period, and association between temporal trends of SI and symptoms of schizophrenia.

Methods

Setting

The parent study was conducted at Beth Isreal Deaconess Medical Centre, Boston, United States; Sangath and All India Institute of Medical Sciences (AIIMS) in Bhopal, Madhya Pradesh, India; and National Institute of Mental Health and Neurosciences (NIMHANS) in Bengaluru, Karnataka. This study is a secondary analysis of data from the latter 2 sites in India. Sangath and the AIIMS Bhopal implemented the parent study in Bhopal. Sangath was founded in 1996 and is a leading mental health research nongovernmental organization in India. Since 2011, Sangath has been working closely with the health system in Madhya Pradesh, to advance research efforts aimed at implementing evidence-based mental health services in primary care settings. AIIMS Bhopal is a premier tertiary medical center and a national center of excellence in medical education, biomedical research, and service delivery. In Madhya Pradesh, AIIMS Bhopal represents a leading regional institution for provision of psychiatric services, including the treatment and management of schizophrenia spectrum disorders. NIMHANS is a tertiary care teaching facility that caters to about 400 daily outpatients with an approximately 600-bed psychiatric in-patient facility. Individuals with schizophrenia receiving treatment at NIMHANS comprise a mix of patients having first episode and chronic disease, coming from urban and surrounding rural localities, and those with difficult-to-treat symptoms, referred from across the country.

Procedures

This study is an exploratory, secondary analysis of smartphone EMA data nested within the ongoing "Smartphone Health Assessment for Relapse Prevention" (SHARP) project or the parent study [22-24], which is guiding the systematic development and adaptation of an open-source smartphone app and digital dashboard—Learn, Assess, Manage and Prevent (LAMP) across diverse cultures and contexts to promote patient-centered care [22-24]. This analysis focuses on the EMA reporting period from September 2021 to April 2023, of outpatients diagnosed with early course schizophrenia at AIIMS in Bhopal in collaboration with Sangath (central India), and NIMHANS in Bengaluru (southern India). The participant sample for the secondary analysis was drawn from the 25 hospital outpatients in each of these 2 parent study sites, recruited during routine outpatient services. Data from the parent study site in Boston, that is, Beth Israel Deaconess Medical Centre was not included for this analysis, as we focused on lower-resource settings in a low- and middle-income country.

Each participant had an initial intake visit and monthly follow-up visits at each site. Participants were recruited from among the

XSL•FO
RenderX

clinic outpatients at NIMHANS and AIIMS hospitals. Inclusion criteria of the parent study were the diagnosis of a psychotic spectrum disorder, or a report of symptom onset for psychosis, within the 5 years before the beginning of data collection in 2021, age 45 years or younger, residence in India, and sufficient proficiency in English or Kannada (in Bengaluru) or English or Hindi (in Bhopal). Further details of the parent study procedures have been published in previous studies [22-24].

During the intake visit in the parent study, the participants were provided a detailed study description followed by written informed consent, and then guided to download the smartphone app, mindLAMP 2, from the App store (iOS) or Play store (Android), and provided a walkthrough of the app and the necessary settings [22-24]. The study team provided a smartphone for patients who did not have it. After the app setup, clinical symptoms for psychosis were assessed using the Positive and Negative Syndrome Scale (PANSS; refer to Measures section), in addition to other parent study baseline assessments [22]. PANSS was readministered at each monthly clinical follow-up, in person or remotely, depending on the latest COVID-19 safety guidelines implemented at the sites. Over the postrecruitment period, participants received daily EMA notifications on their mindLAMP app, including EMA surveys, cognitive games, psychoeducation resources, and relaxation exercises. Daily scheduled activities appeared on the app's feed page, and participants were encouraged to complete the activities daily. In the parent study, mood, anxiety, psychosis, sleep, social functioning, and medication adherence (EMA) surveys were randomized such that 2 of the 6 surveys were sent to participants twice per day, to be completed at least once. Participants reported their mood at the time of the survey prompt.

EMA mood surveys (Patient Health Questionnaire-9 [PHQ-9]), the focus of the exploratory analysis, were sent as part of a "group," with other EMA surveys, during the morning and evening; they remained "open" for the 12-hour duration or till the next "group" of surveys was assigned. These surveys were used to reduce recall bias and increase the ecological validity of the findings [25], while complementing the monthly psychiatric follow-up assessments, for which the patient attended the hospital in person. Importantly, participants could also access the mood survey and submit the data at any time by directly filling in the survey in the mindLAMP's "ASSESS" section; especially, if they missed the survey prompt or could not fill the survey via the notification. We expected the participants to fill a minimum of 1 survey daily, to assess the survey "usage" (we have defined "usage" as discussed subsequently). The mood survey matched the PHQ-9 [26] used for screening depression.

Measures

An instance of SI was measured using the PHQ-9, ninth item score at the EMA prompt ("thoughts that you would be better off dead, or of hurting yourself," cutoff≥1, range: 0-3) [26]. To assess the relationship between the temporality of SI instances and changes in symptoms of psychosis, we used the PANSS positive, negative, and general psychopathology scales, each rated from 1=absent to 7 points (extreme symptoms). The score range is 7 - 49 for the positive and negative scales, and 16 - 112 for the general psychopathology scale [27]. PANSS was administered monthly (in person in Bhopal and in person or remotely in Bengaluru) by the site teams.

Erbe D et al [28] compared the interformat reliability of the paper-and-pen version with a computerized version of the PHQ-9 in a clinical sample and obtained comparable internal consistency between the computer (α=.88) and paper versions (α=.89), and highly significant correlations between the formats (r=0.92). The PANSS is a widely used and valid instrument for the assessment of symptom severity in schizophrenia; test-retest reliability for the total score and subscales is reported as 0.77 - 0.89 [27,27].

We want to clarify that mood, anxiety, psychosis, sleep, social functioning, and medication adherence (EMA) surveys were randomized such that 2 of the 6 surveys were sent to users twice per day (or total 2 EMA survey prompts), to be completed once (in the parent study). Therefore, the mood survey (PHQ-9) was not sent daily due to random selection. However, participants could access the mood survey from the "ASSESS" section of the mindLAMP app at any time. Therefore, we have considered survey "usage" as the measure of feasibility, instead of "compliance" to survey notifications. By defining a day's usage as a "minimum of 1 mood survey filled daily, and counting the first survey of the day (in case of multiple surveys filled)," we have accounted for the option that the participants had, that is, to directly access the mood survey in the app. This is consistent with the fact that while the participants may have missed survey notifications, the research team members tracked their app usage on the dashboard and made touch-base calls with participants to encourage them to use the app and fill the surveys (calls made every 2 - 3 days as a practice, or depending on observed usage on the dashboard).

Sampling and Participants

Out of the 50 outpatients (25 at each site) originally sampled for the parent study (convenience sampling), 10 patients in Bhopal and 12 in Bengaluru had more than 1 instance of SI, essential for conducting the intended analysis, and out of these, 6 patients in Bhopal showed a 40% or more daily usage rate, and 8 patients in Bengaluru showed 20% or more daily usage rate on EMA surveys (usage rate: percentage of mood surveys completed of the total surveys expected to be completed daily ie, minimum 1 survey per day over the reporting period). We defined different usage rate cut offs for the 2 sites after considering the average usage rates (56.3%, IQR 46.7%-64%, SD 13.1% for Bhopal and 34%, IQR 24.2%-43.3%, SD 11.8% for Bengaluru), and the substantial variance between the sites in the distribution of survey usage across participants. Missing data in the sampling process included the number of days when the survey was not filled for which we used an imputation model (discussed subsequently). Based on recent compliance rates of 44% for typical EMA surveys and 35% for high-frequency "burst" surveys over a 42-day monitoring period [18], we defined conservative EMA usage cutoffs, given our much longer data reporting period.

Data Analysis

We referred to analytical plans of similar studies that used EMA [18,22] and higher frequency "burst" surveys [18] to map the

XSL•FO
RenderX

duration or timescale of SI. All analyses were conducted in Stata (version 14; StataCorp) [29] and Microsoft Excel 2010. First, we examined the mean EMA reporting period and survey usage rates (mean, SD, and IQR) across the 14 patients. Second, we examined the descriptive properties such as mean number of instances of SI, duration (days) between first and last reported SI, percentages of modal responses ie, score of 0 on the ninth item of PHQ-9 or "resting state," average duration between 2 successive SI instances, and "no reported SI" period that is, days between last SI and last EMA date. Third, we used timescale summary statistics to quantify the rate at which self-reported SI changed over time within individuals. Through visual inspection of each individual time-series (Figure 1), we identified that most patients showed an "episodic" pattern of SI reporting. "Episodic" time-series include distinct periods of elevated (nonzero) responses interspersed by sequences of zeros, which we can interpret as episodes of heightened suicidal thinking. We quantified the rate of change in the time-series by calculating the frequency and duration of episodes of elevated suicidal thinking. We categorized sequences of consecutive daily observations that show SI score ≥1 as episodes; we calculated the duration of an episode by observing how much time elapses before the next occasion (day) when SI is back to its resting state (score=0). The estimated duration was, therefore, an upper bound of the true episode length, as participants may have returned to "normal" EMA response, before the next instance of SI. In addition, we assessed the duration between the date of the first "drop" in PANSS score (marker of the beginning of clinical improvement) and last reported SI instance, to check for a relationship between clinical changes in symptoms of psychosis and temporality of SI.

We acknowledge the challenge of missing data in EMA studies, where engaging with the app and providing symptom data through surveys can have substantial missing values, particularly in real-world settings such as clinic outpatients (as in this study), over a long follow-up period. Therefore, we have used the "Multiple Imputation for Categorical Time Series (MICT) algorithm," introduced by Halpin (2016), which handles missing data gaps, for the typical form of missing data in longitudinal datasets, by imputing them recursively from their edges [30,31]. MICT model for imputation was originally developed using a multinomial model, and to impute a gap of missing values, the algorithm includes past or future time points. Due to our small participant sample, we could not include covariates (eg, illness duration or gender) in the model, however, we have computed the average frequency of episodes of SI and the average episode duration after imputation, which is the primary aim of the study, and compared these metrics with the timescale summary statistics, which are unadjusted for missing values. Table S1 in Multimedia Appendix 1 includes the detailed data on frequency and duration of SI episodes after imputation for each model, per participant.

Figure 1. Time series plots of instances of suicidal ideations (PHQ-9 ninth item scores) of individual participants. This chart shows the time-series plots of the 14 total participants across the 2 sites. The blue line represents the total PHQ-9 mood scores (range: 0 - 27) plotted on the left-hand axis and the orange dots represent scores of suicidal ideation represented by the PHQ-9 ninth item score ranging from 0 (absent, or "resting") to 3, graded on the right-hand axis. The gray threshold line represents the PHQ-9 total score cutoff of 10 for moderate depression. The total surveys filled by each participant with the corresponding reporting period are mentioned at the top of the time-series plot. EMA: ecological momentary assessment; PHQ-9: Patient Health Questionnaire-9.

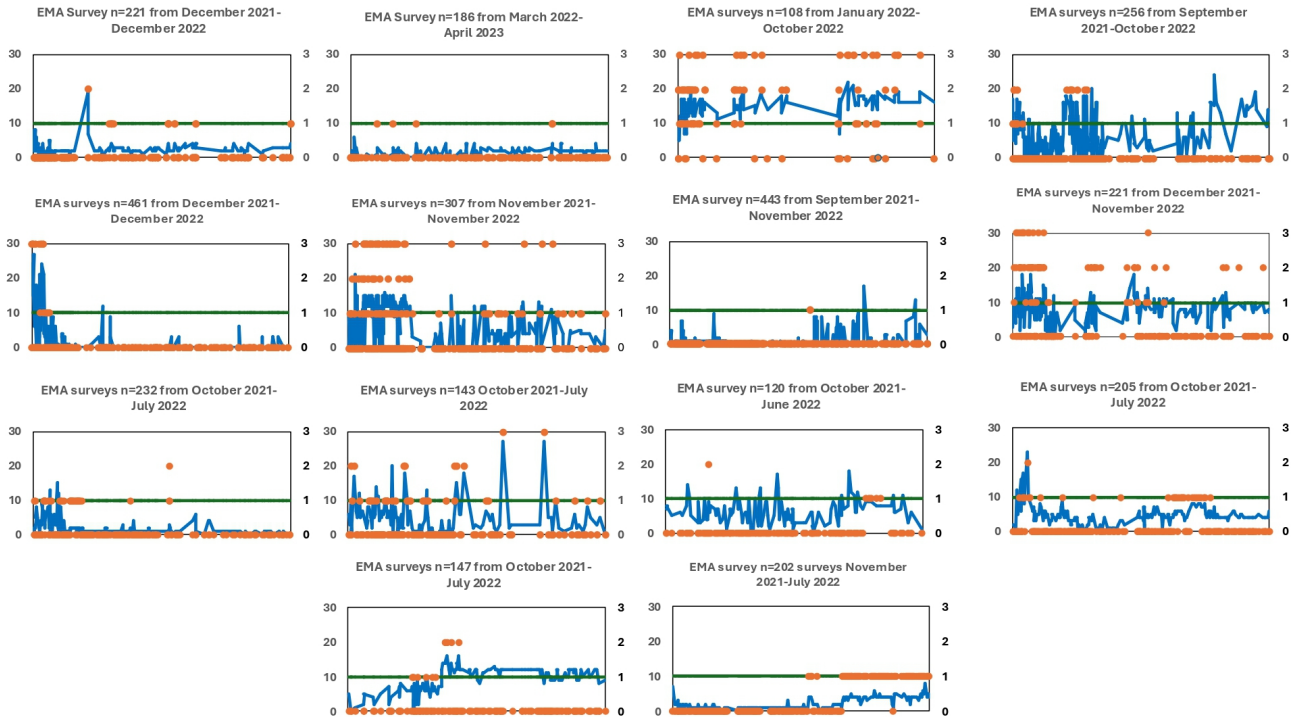

Ethical Considerations

Ethical approval for both study sites was granted by their respective institutional review boards (IRBs): Sangath IRB and AIIMS Bhopal Institutional Human Ethics Committee, and NIMHANS IRB, in accordance with relevant ethical guidelines for human research. All participants in the parent study provided written informed consent before participation, which allowed

XSL•FO
RenderX

the use of deidentified data in secondary analyses without additional consent. To protect participants' privacy and confidentiality, all data were deidentified (a study ID was assigned to each participant), no personally identifiable information was included in the final dataset, and all published data were presented in aggregate form. Participants' data (eg, case record forms, information sheets, and consent forms) were stored in a locked cabinet in the respective site's research office. During the study, only deidentified data were used, and the data were only accessible to the research team. There is no identification of individual participants or users in any figures of this paper or its supplementary material. Participants received compensation of INR 500 (approximately US $7) for each assessment session in which they participated.

Results

Participant Overview

Following is the summary of demographic characteristics of the subgroups of patients included in the analysis (n=14 as shown in Figure 2): Bhopal (n=6, average age: 30.8, SD 6.6, range: 23 - 42 years; n=3, 50% female; mean 11.8, SD 1.8 years of education; n=5, 83.3% were single); Bengaluru (n=8, average age: 33.9, SD 3.8, range: 31 - 42 years; n=3, 37.5% female; mean 13.6, SD 2.8, years of education; n=7, 87.5% were single). We also compared the demographic characteristics between the full (n=22) and the analytic sample (n=14) but found marginal differences (Tables S2 and S3 in).

XSL•FO
RenderX

Figure 2. Sample selection for the secondary analysis. The parent study included a cross-sectional sample of 50 patients from Bhopal and Bengaluru sites. A total of 22 patients reported more than 1 instance of suicidal ideation (SI), which was necessary for their inclusion in the secondary analysis, and furthermore, 6 patients from Bhopal and 8 patients from Bengaluru had daily mood survey usage rates, that is, ≥40% and ≥20%, respectively. These 14 patients were included for the final analysis. We defined different usage rate cutoffs for the 2 sites after considering the average usage rates (56.3%, IQR 46.7%-64% for Bhopal and 34%, IQR 24.2%-43.3% for Bengaluru), and the substantial variance between the sites in the distribution of survey usage across participants. EMA: ecological momentary assessment.

XSL•FO
RenderX

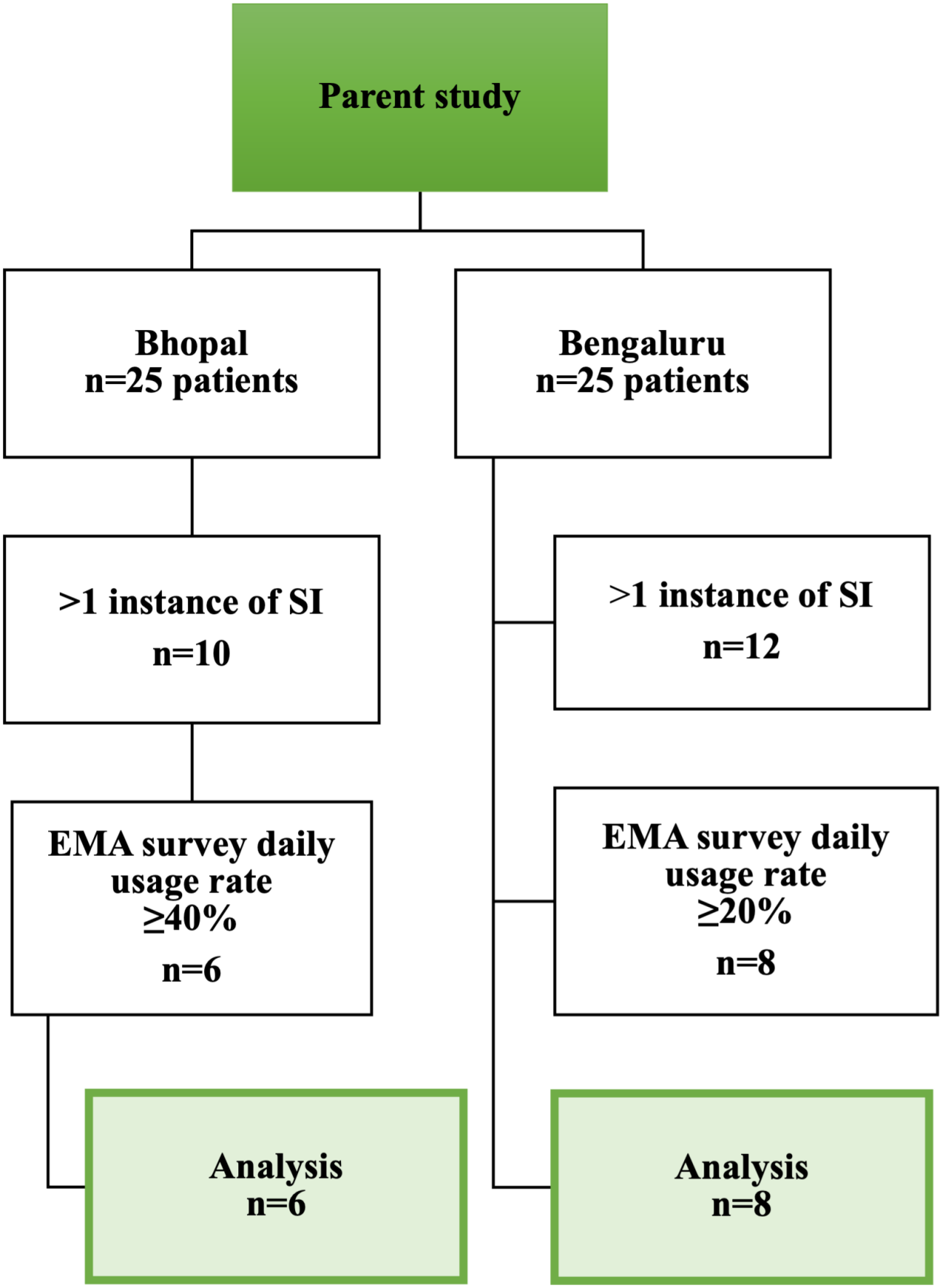

XSL•FO
RenderX

Timescale of Suicidal Ideation

For the 14 patients included in the secondary analysis, the average days of EMA were 326 or ~11 (IQR 9-13) months, with an average of 56.3% (IQR 46.7%-64%) of EMA survey usage rate for Bhopal and 34% (IQR 24.2%-43.3%) for Bengaluru. In total, 3253 mood surveys were filled by the 14 participants (average: 232 surveys , IQR 147-256) with a total 521 instances of SI, and 76 missing values on PHQ-9 ninth item across the filled surveys (2.3%). The number of SI instances in each month over the reporting period showed substantial within and between-person variations as shown in the grid in Figure 3.

Figure 3. Within and between-person variations in number of SI instances in each month of the reporting period. The grid below represents the participant-wise and month-wise instances of SI (score of 1 or more). The frequency of SI instances has been graded in the legend. The thicker black borders of the individual participant columns indicate the boundaries of the reporting period during which EMA data were collected. April 2023 was the last month of data collection; maximum reporting period was 15 months. EMA: ecological momentary assessment; PHQ-9: Patient Health Questionnaire-9; SI: suicidal ideation.

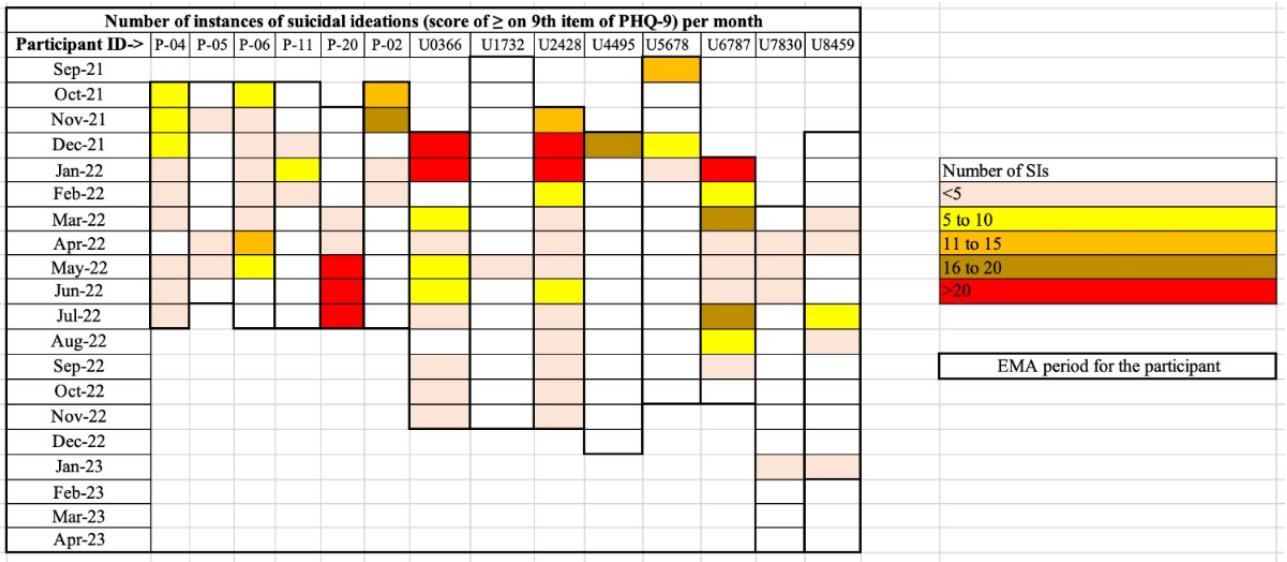

The descriptive analysis (Table 1) of raw data (n=14 patients), disregarding the episodic patterns of SI, revealed high SDs for the average number of SI instances (mean 37.2, SD 34.2), average days between first and last reported SI (mean 188, SD 118.3), percentage of modal ("resting," score=0) response (mean 78.9%, SD 23.4%), average days of "no reported SI" period after last reported SI (mean 95.5, SD 114.6) and average days between 2 successive SI instances (mean 12, SD 17.8). Table 2 presents these statistics after imputation of missing data.

Table . Descriptive statistics of instances of suicidal ideation (SI) reported during ecological momentary assessment (disregarding the episodic pattern of SI and before imputation). Instance of SI is defined as a score of 1 or higher on the ninth item of Patient Health Questionnaire-9; a score of 0 is the "modal" (resting) response. The results in this table do not account for missing values or days with no survey forms received or no ninth item filled in the mood survey.

Characteristics[a]	Statistical values		
	Mean (SD)	Median (IQR)	Range
Number of SI instances	37.2 (34.2)	24.5 (10-70)	2 - 106
Days between first and last reported SI	188 (118.3)	192.5 (112-284)	0 - 365
Percentage of modal ("resting," score=0) response	78.9 (23.4)	88.8 (65-95)	15 - 99.3
Duration (days) of "no reported SI" period, after last reported SI	95.5 (114.6)	53 (6-157)	0 - 359
Duration (days) between successive SI instances[b,c]	12 (17.8)	5.2 (2.8-7.9)	0 - 65

[a]Spearman correlation between Patient Health Questionnaire-9 total and SI score at ecological momentary assessment prompts was 0.47 (n=3176, *P*<.001); correlation between Patient Health Questionnaire-9 score>=10 and suicidal ideation score ≥1 was 0.32 (n=245, *P*<.001).

[b]Disregarding the episodic pattern.

[c]To understand the difference between "days between successive SI instances" and "episode duration" (refer to the episode data in Table 3), let's take an example of a patient who has 2 suicidal ideation over 2 consecutive days; the average duration between the successive suicidal ideations is 1 day and the episode duration is also 1 day. However, if there are 4 suicidal ideations over 4 consecutive days, the episode duration is 4 days but the average duration between suicidal ideations is 1 day.

XSL•FO
RenderX

Table . Descriptive statistics of instances of SI[a] reported during ecological momentary assessment (disregarding the episodic pattern of SI) after imputation. The results in this table have been calculated after imputation, and account for missing values, which explains the differences in the means, compared with Table 1.

Statistics	Mean (SD)	Median (IQR)	Range
Number of SI instances (n)	70.8 (67.4)	43.7 (23.1-87.5)	10.8 - 224.7
Days between first and last reported SI (days)	296.6 (54.8)	287.3 (266.85 - 344.45)	212.2 - 367.9
Percentage of modal ("resting," score=0) response (%)	76.8 (23.2)	84.5 (65 - 91.9)	17.1 - 97.2
Duration of "no reported SI" period, after last reported SI (days)	10.7 (11.1)	9.7 (1.2 - 19.4)	0 - 34.2
Duration between successive SI instances (days)	11.3 (13.5)	6.4 (2.9 - 12.2)	1.2 - 51.9

[a]SI: suicidal ideation.

As the SI patterns across patients were largely episodic (except for 3 patients, which showed highly variable distribution of SI), the mean duration quantities showed high SDs in the raw data (Tables 1 and 2). Therefore, conducting the descriptive analysis at the episode-level was important, which revealed less noise even with a smaller sample (n=11, refer to Tables 3 and 4). Timescale summary statistics revealed an average of 5.9 episodes (range: 1 - 14, SD 4.4, 65 total episodes) with an episode of heightened suicidal thinking lasting for 2.5 days on average (range: 1 - 5.3, SD 1.5, 27 total days). Please refer to Tables 3 and 4 for these results.

Table . Timescale summary statistics to calculate the frequency and duration of episodes of SI[a] (before imputation). Episodes are categorized sequences of consecutive daily observations that show SI score ≥1. We calculated the duration of an episode by observing how much time elapses before the next occasion (day) when SI is back to its resting state (score=0). The estimated duration was, therefore, an upper bound of the true episode length, as participants may have returned to "normal" ecological momentary assessment response, before the next instance of SI.

Statistics	Mean (SD)	Range	Sum	IQR
Frequency of SI episodes (episodes)	5.9 (4.4)	1 - 14	65	3 - 11
Average SI episode duration (days)	2.5 (1.5)	1 - 5.3	27	1.2 - 3.2

[a]SI: suicidal ideation.

Table . Timescale summary statistics to calculate the frequency and duration of episodes of suicidal ideation after imputation. We conducted the imputation modelling with 13 out of the 14 participants. We excluded one participant who had 2 suicidal ideation instances reported on the same day and no other instance reported throughout their participation, which precluded an estimation of episodes.

Statistics	Mean (SD)	Range	Sum	IQR
Frequency of SI[a] episodes (episodes)	12.1 (13.2)	1.1 - 41.5	157	2.3 - 16
Average SI episode duration (days)	2.1 (1.8)	1 - 6.3	28	1 - 2.1

[a]SI: suicidal ideation.

In addition, we have also analyzed the survey reporting rate and SI instance reporting rate of participants, site-wise, for a subsample of participants, for 4 different 6-hour time zones. For Bhopal participants, the maximum or 57.4% of the surveys (343 surveys filled) were filled during the 6 AM to <12 PM slot; and SI reporting rate was highest during this slot, ie, 66.7% (104 SI instances). For Bengaluru participants, 53.9% of the surveys (1147 surveys filled) were filled during the 12 PM to <6 PM slot; and SI reporting rate during this slot was 42.6% (141 SI instances).

Relationship Between Suicidal Ideation and Changes in Symptoms of Psychosis

We also conducted an analysis of the relationship between changes in symptoms of psychosis and the temporality of patterns of SI (Figures 4 and 5 refer to the same participant).

XSL•FO
RenderX

Figure 4. SI time series (ecological momentary assessment mood survey data) of a participant. Figures 4 and 5 aim to respectively show the temporal patterns in SI, and the changes in monthly Positive and Negative Syndrome Scale (clinical in person assessment) scores for the same (illustrated) participant, with available ecological momentary assessment and Positive and Negative Syndrome Scale (in person) assessment data. This figure shows an SI item-level response of 0 (no SI, or modal response), 1 and 2 (scaled on the right-hand axis); the horizontal line passing through 10 marks the SI cut-off of 1 used in this study. Each dot represents the score of SI instance graded from 0 to 3. The graph line indicates the total PHQ-9 score for the same timepoints, scaled on the left-hand axis. An episodic SI pattern is visible with clusters of SI instances (dots) separated by periods of modal response (before imputation). PHQ-9: Patient Health Questionnaire-9; SI: suicidal ideation.

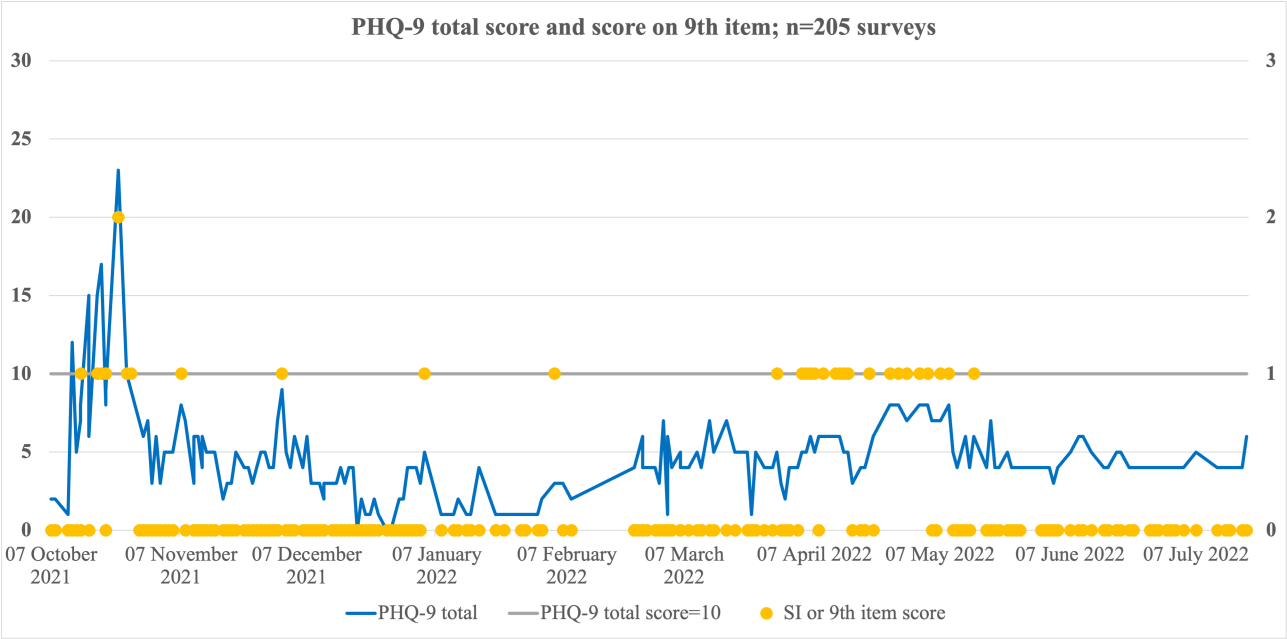

Figure 5. Monthly PANSS scores. This figure shows the monthly changes in positive, negative, and general psychopathology scores for the reporting period, in case of the same participant shown in Figure 4. First drop in PANSS scores was observed in November 2021, but enduring suicidal ideation vulnerability was observed till as late as May 2022 (Figure 4). PANSS: Positive and Negative Syndrome Scale.

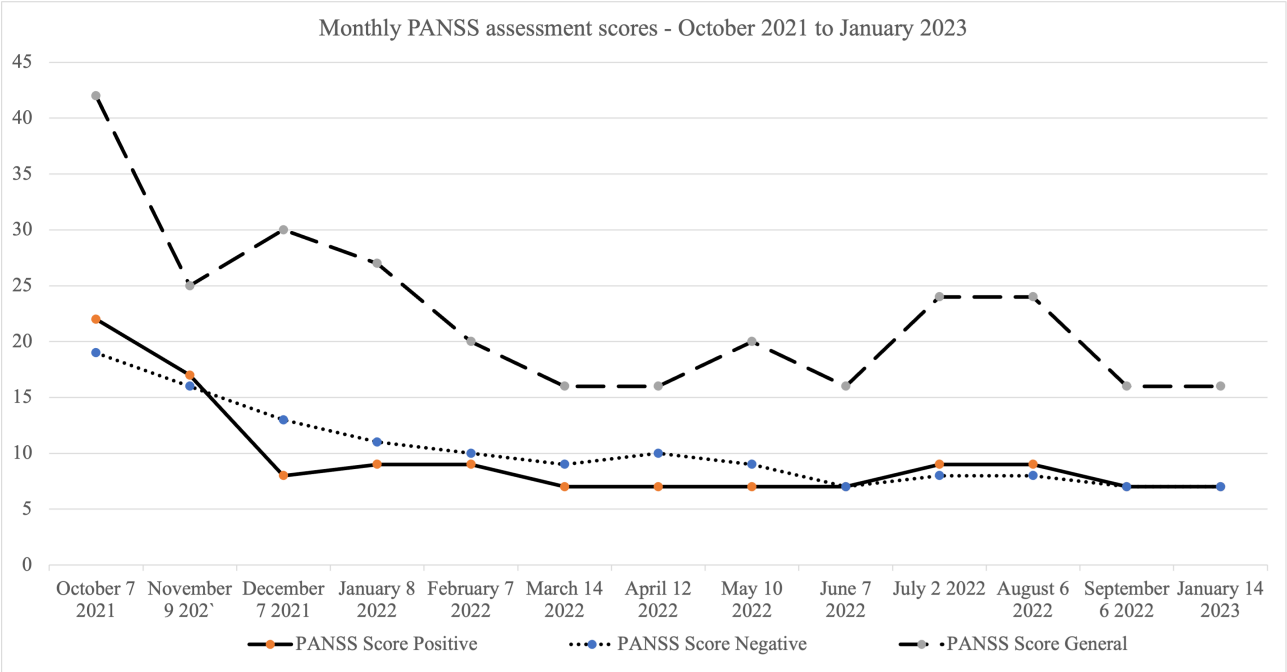

Table 5 shows the distribution of PANSS assessments from September 2021 (when EMA surveys also started) to April 2023 (n=14 patients, 154 total assessments, 11 assessments per patient on average [SD 2.1, range 6 - 13]).

XSL•FO
RenderX

Table . Distribution of Positive and Negative Syndrome Scale assessments (monthly), showing the number of assessments conducted, cumulative, and average per participant.

	Number of assessments
Mean (SD)	11 (2.1)
Range	6-13
Sum	154

We examined the number of days that passed between the first drop in PANSS score and the last reported SI, within the PANSS reporting period. As shown in Table 6, there was an average lag of ~59 days, 66 days and 81 days between the first drop in PANSS positive, negative, and general psychopathology score respectively, as a marker of the beginning of clinical improvement, and the last reported SI. Negative values for "days" implied that the drop in PANSS score occurred after the instances of SI, which partly inflated the SDs (n=3 patients), however, in most patients (n=11), the lag was positive indicating that the last reported SI was subsequent to the first signs of clinical improvement on PANSS. Please refer to Table 6 for these results. Table 7 shows these results after imputation of missing data.

Table . Duration (in days) between the timing of the first drop in Positive and Negative Syndrome Scale (PANSS) score and last reported suicidal ideation (SI), within the PANSS reporting period (before imputation). The results summarize the relationship between first signs of clinical improvement on PANSS and persistence of SI. These results are unadjusted for missing values.

Conditions	Duration (days)		
	Mean SD	Median (IQR)[a]	Range[a]
Between first drop in PANSS positive score and last SI	58.6 (66.7)	85 (–6.5 to 114)	–50 to 132
Between first drop in PANSS negative score and last SI	66 (82.9)	65 (9 to 107)	–57 to 207
Between first drop in PANSS general score and last SI	81 (74.4)	79 (28 to 121)	–40 to 207

[a]The negative value represents the suicidal ideation incidence before the first drop in PANSS score (for n=3 patients).

Table . Duration between the timing of the first drop in Positive and Negative Syndrome Scale (PANS) score (in days) and last reported suicidal ideation (SI), within the PANSS reporting period, after imputation. These results are adjusted for missing values, or after imputation.

Conditions	Duration (days)			
	Mean (SD)	Median (IQR)	Range[a]	Frequency, n
First drop in PANSS positive score and last SI	248.2 (47.7)	244.6 (210.4-285.5)	176.6 to 327.4	12
First drop in PANSS negative score and last SI	232.7 (89.4)	242.4 (197.5-293)	–14.55 to 327.4	13
First drop in PANSS general score and last SI	204.8 (112.5)	225.5 (197.5-254.3)	–37 to 337.9	13

[a]The negative value represents the suicidal ideation incidence before the first drop in PANSS score (for n=3 patients).

Discussion

Principal Findings

First, the findings of this study show that it is feasible to collect smartphone EMA data on SI from outpatients having schizophrenia in relatively lower-resource hospital settings in India. This builds on the feasibility and acceptability of EMAs as one of the digital phenotyping methods [16,32] to generate novel forms of data on suicide risk. The demonstrated user engagement in clinical settings in this study is notable (3253 mood surveys filled by 14 patients with schizophrenia over 15 months, with a low proportion [2%] of missing value on the ninth item of PHQ-9), especially as low engagement of health apps as a limiting factor in their regular clinical implementation has been cited in literature [33]. The research team had involved the patients, caregivers, and their attending clinicians in the development and refinements of mindLAMP through earlier extensive focus group discussions in the study sites [34,35], which could have contributed to the greater engagement toward this app to generate clinical data on SI, which was otherwise difficult to obtain in face-to-face models of care where patients usually meet with their attending clinician monthly or every 2 months. It is also important to note that the study could analyze SI data over more than a year of follow-up outpatient visits, representing the typical periods over which patients with psychosis usually follow-up at government hospitals in India, and potentially other similar contexts of the Global South, unlike the much shorter (≤1 month) EMA monitoring periods observed

XSL•FO
RenderX

in studies [19]. Related to feasibility, we also found (through reminder or touch-base phone calls to encourage the participants to use the app) that factors affecting EMA usage included low battery-life of the phone due to the passive data collection (eg, GPS data) on the mindLAMP app (beyond the scope of this analysis) and variations in clinical symptoms of schizophrenia potentially affecting app usage. For the latter, we computed Pearson correlations between mood survey usage rate, as defined by a minimum of 1 daily mood survey, and the PANSS score of the previous and the succeeding months. Overall, weak correlation coefficients were observed between survey usage and either the previous or succeeding month's PANSS score, with somewhat higher and positive correlations noted for the relationship between previous month's PANSS score and succeeding month's survey usage rate (succeeding month's usage and previous month PANSS positive score: 0.26, P=.003, n=127 observations; PANSS negative score: 0.01, P=.8, n=127 observations; and PANSS general score: 0.18, P=.0437, n=127 observations).

Second, the results show substantial within and between person variation in SI, and the attempt to map the timescale of SI among hospital outpatients with schizophrenia is an early effort in the Indian context, to the knowledge of the authors. The initial noise observed in the duration between first and last reported SI and between successive SI instances due to small sample size and episodic SI pattern was controlled by using timescale statistics to help assess the SI evolution better, which revealed that on average, there were 6 SI episodes per patient, each 2.5 days long. Furthermore, with imputation of missing values, 12 SI episodes per patient were reported on average, each 2.1 days long. Therefore, the timescale statistics provide a sense of the intensity of SI as experienced by the patient, and with more rigorous examinations (eg, by high-frequency burst EMA surveys), we can attempt to examine an "episode" in terms of hours.

Third, this secondary analysis and the parent study controlled to the extent possible, the biases that are found in EMA surveys; for example, participants were allowed to postpone an EMA prompt or reject a prompt, which helped reduce practice effects [36], although the statistical significance of the difference in practice effects for specific EMA items was not monitored in the parent study; also, the grid in Figure 3 shows possibly low "reactive" effects because clustering of SI counts of different degrees (color coded in the grid) was seen throughout the EMA period for most patients, and not necessarily in the beginning of the study [37].

Finally, in relation to first signs of clinical improvement in PANSS+, PANSS−, and PANSS general scores, SI instances were observed after approximately 60, 66, and 81 days (approximately 2 - 3 months) respectively (within the PANSS reporting period), indicating that SI was an enduring vulnerability even after the beginning of early improvement in positive, negative, and general psychopathology symptoms of psychosis, rather than a short-term crisis [38]. After imputation, the persistence of SI after early signs of clinical improvement was more pronounced; for example, after clinical improvement, SI instances were observed after approximately 248, 232, and 204 days, respectively (approximately 7 - 8 months) for PANSS+, PANSS–, and PANSS general scores.

Limitations

The study has several limitations. First, the PHQ-9 measures passive SI, and it has been shown through rigorous EMA and higher-frequency burst survey analyses that active and passive SI can operate on different timescales [18]. Furthermore, previous evidence states that item-9 of the PHQ-9 is not sufficient to assess suicide risk and suicide ideation, with limited use in certain demographic and clinical subgroups, which requires further research [39]. It was not possible in this study design of a secondary analysis nested in a completed primary study, to understand the frequency, intensity and duration of an instance of SI, or the exploration of active SI. Relatedly (aligned with the study aim), to explore the temporality of SI, we selected a lower cutoff of 1 on the PHQ-9 item-9; but this precluded an analysis of SI severity in relation to time. Second, while the EMA monitoring period in this study was long given the length of outpatient care for persons with psychosis, the observed mean mood survey usage rates of 56.3% (Bhopal) and 34% (Bengaluru) were to an extent, comparable with the mean EMA compliance (44%) observed in a recent assessment of the SI timescale [18], which was a reference for this analysis. There is a caveat though that we have little knowledge on the benchmarks or realistic expectations of what compliance or survey usage rates should be with EMA and higher-frequency sampling. Taking the usage rates together with a low sample size of patients in our study, and the fact that quantitative assessments of "fatigue" effects were not within the scope of the parent study (ie, the change in overall EMA survey usage over time spent in the study), it is challenging to make broad claims around the episodic nature of SI (frequency and duration of episodes) and the relationship of SI with improvement on PANSS. Third, the parent study could not use higher frequency burst surveys due to potential respondent burdens on hospital outpatients, which limited the scope of potentially identifying 19% more patients with SI [18], and rendered the durations of SI episodes in days, and not in hours. This also reduced the ecological validity, and data patterns resembled more of a daily dairy than conventional EMA; though it needs to be recognized that this is a real-world lower-resourced outpatient clinic sample in a long follow-up period. Higher-frequency (hours or minutes apart) sampling is necessary to accurately characterize within-person dynamics of suicidal thinking, although with the costs of respondent burden and potential reductions in overall survey usage. Fourth, "anonymous sampling," which increases the overall identification of endorsement of rates of SI [40,41] was not done in the parent study , that is , participants had to provide their name, address, or phone number in the process of recruitment in the outpatient clinic. Fifth, the type of population and setting for the above analysis to estimate the timescale of SI limits the generalizability of the findings in multiple ways, for example, (a) the findings may not be generalizable to people living with schizophrenia who are not engaged in outpatient care, or who may be unable or hesitant to seek formal psychiatric outpatient care; (b) the findings may not apply to older people living with the illness given our study age group, as also to younger people in the community (nonhospital) settings, and

XSL•FO
RenderX

adults with schizophrenia who are hospitalized; and (c) the data were collected during COVID-19, which may point to unique stressors and the findings therefore, may not generalize to people outside this period of time. Sixth, we want to highlight the small sample size for analysis, that is, 14 of 22 participants with SI had EMA data of sufficient quality that could be analyzed, which limits the feasibility of our approach. Finally, SI or suicidal thoughts (as captured by PHQ-9) are relatively common, and typically not serious unless the individual has a plan, intent, or makes a suicide attempt. This analysis does not have these details, and without knowing the complete suicidal intention, it is difficult to design or inform prevention efforts.

Conclusions

Even as identifying "when" a person is at risk of suicide is as important as identifying "who" is at risk, and it is recognized that mapping the evolution of suicidal thinking in high-risk individuals such as outpatients with psychosis is necessary to make significant progress in suicide prevention efforts, the national suicide prevention strategy in India [42] only mentions "digital interventions" as short-term prevention strategies. This is because the research on examining the patterns of suicidal thoughts in real time among mental health clinic outpatients in India is extremely scarce. Our study adds to the much-needed evidence base to further highlight the next steps in EMA research on measuring the changes in suicidal thinking, that is, the use of higher-frequency "burst" surveys to assess the duration of an episode of SI in hours or even minutes for more targeted interventions; inclusion of markers of active and passive SI and understanding the shift from passive to active SI, which presents a therapeutic window of opportunity [43], in addition to collecting data on attempt; addition of passive data (eg, mobility markers) to the analysis; and inclusion of other relevant measures (such as PANSS in this analysis) such as changes in anxiety level and sleep to arrive at more precise estimations of the intensity and duration of individual periods of heightened suicidal thinking experienced by an individual.

Acknowledgments

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article. This work was supported by the Wellcome Trust (grant no. 215843/Z/19/Z). The funder had no role in the study design; collection, analysis, and interpretation of data; writing of the paper; and decision to submit for publication. The authors wish to acknowledge Khushbu Verma, Junior Officer at AIIMS, Bhopal, and Ram Vishwakarma, Administration and Finance Officer at Sangath Bhopal for the coordination, administration, and financial management of the study activities, and Yogendra Sen for his contribution to data collection. The authors also wish to acknowledge the contribution of Rahul Singh, Shinto Thomas, and G Sai Teja for their assistance in the data collection and analysis.

Data Availability

The datasets generated and analyzed during this study are not publicly available and we do not have institutional review board approval to share the study data due to privacy and confidentiality of participants' mental health data.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplemental tables of imputation models for missing data and additional demographic characteristics of participants.
[DOCX File, 43 KB - formative_v9i1e67745_app1.docx]

References

1. Yadav S, K K A, Cunningham SA, et al. Changing pattern of suicide deaths in India. Lancet Reg Health Southeast Asia 2023 Sep;16:100265. [doi: 10.1016/j.lansea.2023.100265] [Medline: 37649643]
2. Amudhan S, Gururaj G, Varghese M, et al. A population-based analysis of suicidality and its correlates: findings from the National Mental Health Survey of India, 2015-16. Lancet Psychiatry 2020 Jan;7(1):41-51. [doi: 10.1016/S2215-0366(19)30404-3] [Medline: 31826821]
3. Piotrowski P, Gondek TM, Królicka-Deręgowska A, Misiak B, Adamowski T, Kiejna A. Causes of mortality in schizophrenia: an updated review of European studies. Psychiatr Danub 2017 Jun;29(2):108-120. [doi: 10.24869/psyd.2017.108] [Medline: 28636568]
4. Laursen TM. Causes of premature mortality in schizophrenia: a review of literature published in 2018. Curr Opin Psychiatry 2019 Sep;32(5):388-393. [doi: 10.1097/YCO.0000000000000530] [Medline: 31135491]
5. Bai W, Liu ZH, Jiang YY, et al. Worldwide prevalence of suicidal ideation and suicide plan among people with schizophrenia: a meta-analysis and systematic review of epidemiological surveys. Transl Psychiatry 2021 Oct 29;11(1):552. [doi: 10.1038/s41398-021-01671-6] [Medline: 34716297]
6. Fazel S, Runeson B. Suicide. N Engl J Med 2020 Jan 16;382(3):266-274. [doi: 10.1056/NEJMra1902944] [Medline: 31940700]

XSL•FO
RenderX

7. Millner AJ, Robinaugh DJ, Nock MK. Advancing the understanding of suicide: the need for formal theory and rigorous descriptive research. Trends Cogn Sci 2020 Sep;24(9):704-716. [doi: 10.1016/j.tics.2020.06.007] [Medline: 32680678]
8. Turecki G, Brent DA, Gunnell D, et al. Suicide and suicide risk. Nat Rev Dis Primers 2019 Oct;5(1):74. [doi: 10.1038/s41572-019-0121-0]
9. Mamtani H, Pathak H, Sakhardande K, et al. Suicide attempts in schizophrenia – a large retrospective cohort study from a tertiary care psychiatry centre in India. Asian J Psychiatr 2024 Jun;96:104053. [doi: 10.1016/j.ajp.2024.104053] [Medline: 38678797]
10. Nath S, Kalita KN, Baruah A, Saraf AS, Mukherjee D, Singh PK. Suicidal ideation in schizophrenia: a cross-sectional study in a tertiary mental hospital in North-East India. Indian J Psychiatry 2021;63(2):179-183. [doi: 10.4103/psychiatry.IndianJPsychiatry_130_19] [Medline: 34194063]
11. Sicotte R, Iyer SN, Kiepura B, Abdel-Baki A. A systematic review of longitudinal studies of suicidal thoughts and behaviors in first-episode psychosis: course and associated factors. Soc Psychiatry Psychiatr Epidemiol 2021 Dec;56(12):2117-2154. [doi: 10.1007/s00127-021-02153-2]
12. Madsen T, Karstoft KI, Secher RG, Austin SF, Nordentoft M. Trajectories of suicidal ideation in patients with first-episode psychosis: secondary analysis of data from the OPUS trial. Lancet Psychiatry 2016 May;3(5):443-450. [doi: 10.1016/S2215-0366(15)00518-0] [Medline: 26948484]
13. Parrish EM, Chalker SA, Cano M, et al. Ecological momentary assessment of interpersonal theory of suicide constructs in people experiencing psychotic symptoms. J Psychiatr Res 2021 Aug;140:496-503. [doi: 10.1016/j.jpsychires.2021.06.022]
14. Parrish EM, Chalker S, Cano M, et al. Ecological momentary assessment of social approach and avoidance motivations in serious mental illness: connections to suicidal ideation and symptoms. Arch Suicide Res 2024;28(1):123-140. [doi: 10.1080/13811118.2022.2137445] [Medline: 36377277]
15. Ammerman BA, Law KC. Using intensive time sampling methods to capture daily suicidal ideation: a systematic review. J Affect Disord 2022 Feb 15;299:108-117. [doi: 10.1016/j.jad.2021.10.121] [Medline: 34718039]
16. Butner JE, Bryan CJ, Tabares JV, et al. Temporal-dimensional examination of the Scale for Suicidal Ideation in a cohort of service members in treatment for PTSD. Psychol Trauma 2021 Oct;13(7):793-801. [doi: 10.1037/tra0001030] [Medline: 34723567]
17. Sewall CJR, Wright AGC. Personalizing suicidology. Crisis 2021 Nov;42(6):405-410. [doi: 10.1027/0227-5910/a000834] [Medline: 34783589]
18. Coppersmith DDL, Ryan O, Fortgang RG, Millner AJ, Kleiman EM, Nock MK. Mapping the timescale of suicidal thinking. Proc Natl Acad Sci U S A 2023 Apr 25;120(17):e2215434120. [doi: 10.1073/pnas.2215434120] [Medline: 37071683]
19. Kivelä L, van der Does WAJ, Riese H, Antypa N. Don't miss the moment: a systematic review of ecological momentary assessment in suicide research. Front Digit Health 2022;4:876595. [doi: 10.3389/fdgth.2022.876595] [Medline: 35601888]
20. Sedano-Capdevila A, Porras-Segovia A, Bello HJ, Baca-García E, Barrigon ML. Use of ecological momentary assessment to study suicidal thoughts and behavior: a systematic review. Curr Psychiatry Rep 2021 May 18;23(7):1-17. [doi: 10.1007/s11920-021-01255-7] [Medline: 34003405]
21. Mote J, Fulford D. Ecological momentary assessment of everyday social experiences of people with schizophrenia: a systematic review. Schizophr Res 2020 Feb;216:56-68. [doi: 10.1016/j.schres.2019.10.021] [Medline: 31874743]
22. Lakhtakia T, Bondre A, Chand PK, et al. Smartphone digital phenotyping, surveys, and cognitive assessments for global mental health: initial data and clinical correlations from an international first episode psychosis study. Digit Health 2022;8:20552076221133758. [doi: 10.1177/20552076221133758] [Medline: 36386246]
23. Cohen A, Naslund JA, Chang S, et al. Relapse prediction in schizophrenia with smartphone digital phenotyping during COVID-19: a prospective, three-site, two-country, longitudinal study. Schizophrenia (Heidelb) 2023 Jan 27;9(1):6. [doi: 10.1038/s41537-023-00332-5] [Medline: 36707524]
24. Rodriguez-Villa E, Mehta UM, Naslund J, et al. Smartphone Health Assessment for Relapse Prevention (SHARP): a digital solution toward global mental health. BJPsych Open 2021 Jan 7;7(1):e29. [doi: 10.1192/bjo.2020.142] [Medline: 33407986]
25. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol 2008;4(1):1-32. [doi: 10.1146/annurev.clinpsy.3.022806.091415] [Medline: 18509902]
26. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613. [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
27. Kay SR, Fiszbein A, Opler LA. The Positive and Negative Syndrome Scale (PANSS) for schizophrenia. Schizophr Bull 1987;13(2):261-276. [doi: 10.1093/schbul/13.2.261] [Medline: 3616518]
28. Erbe D, Eichert HC, Rietz C, Ebert D. Interformat reliability of the patient health questionnaire: validation of the computerized version of the PHQ-9. Internet Interv 2016 Sep;5:1-4. [doi: 10.1016/j.invent.2016.06.006] [Medline: 30135800]
29. Stata. URL: https://www.stata.com/company/ [accessed 2024-10-19]
30. Emery K, Studer M, Berchtold A. Comparison of imputation methods for univariate categorical longitudinal data. Qual Quant 2024 Dec;26:1-25. [doi: 10.1007/s11135-024-02028-z]
31. Halpin B. Multiple imputation for categorical time series. The Stata Journal: Promoting communications on statistics and Stata 2016 Sep;16(3):590-612. [doi: 10.1177/1536867X1601600303]

XSL•FO
RenderX

32. Pizzoli SFM, Monzani D, Conti L, Ferraris G, Grasso R, Pravettoni G. Issues and opportunities of digital phenotyping: ecological momentary assessment and behavioral sensing in protecting the young from suicide. Front Psychol 2023;14:1103703. [doi: 10.3389/fpsyg.2023.1103703] [Medline: 37441331]
33. Torous J, Nicholas J, Larsen ME, Firth J, Christensen H. Clinical review of user engagement with mental health smartphone apps: evidence, theory and improvements. Evid Based Ment Health 2018 Aug;21(3):116-119. [doi: 10.1136/eb-2018-102891] [Medline: 29871870]
34. Bondre AP, Shrivastava R, Raghuram H, et al. A qualitative exploration of perceived needs and barriers of individuals with schizophrenia, caregivers and clinicians in using mental health applications in Madhya Pradesh, India. SSM - Mental Health 2022 Dec;2:100063. [doi: 10.1016/j.ssmmh.2022.100063]
35. Rodriguez-Villa E, Rozatkar AR, Kumar M, et al. Cross cultural and global uses of a digital mental health app: results of focus groups with clinicians, patients and family members in India and the United States. Glob Ment Health 2021;8:1-10. [doi: 10.1017/gmh.2021.28]
36. Wrzus C, Neubauer AB. Ecological momentary assessment: a meta-analysis on designs, samples, and compliance across research fields. Assessment 2023 Apr;30(3):825-846. [doi: 10.1177/10731911211067538] [Medline: 35016567]
37. Husky M, Olié E, Guillaume S, Genty C, Swendsen J, Courtet P. Feasibility and validity of ecological momentary assessment in the investigation of suicide risk. Psychiatry Res 2014 Dec 15;220(1-2):564-570. [doi: 10.1016/j.psychres.2014.08.019] [Medline: 25155939]
38. Simon GE, Rutter CM, Peterson D, et al. Does response on the PHQ-9 Depression Questionnaire predict subsequent suicide attempt or suicide death? Psychiatr Serv 2013 Dec 1;64(12):1195-1202. [doi: 10.1176/appi.ps.201200587] [Medline: 24036589]
39. Na PJ, Yaramala SR, Kim JA, et al. The PHQ-9 Item 9 based screening for suicide risk: a validation study of the Patient Health Questionnaire (PHQ)−9 Item 9 with the Columbia Suicide Severity Rating Scale (C-SSRS). J Affect Disord 2018 May;232:34-40. [doi: 10.1016/j.jad.2018.02.045]
40. Deming CA, Harris JA, Castro-Ramirez F, et al. Inconsistencies in self-reports of suicidal ideation and attempts across assessment methods. Psychol Assess 2021;33(3):218-229. [doi: 10.1037/pas0000976]
41. Vannoy SD, Andrews BK, Atkins DC, et al. Under reporting of suicide ideation in US army population screening: an ongoing challenge. Suicide Life Threat Behav 2017 Dec;47(6):723-728. [doi: 10.1111/sltb.12318] [Medline: 27976807]
42. Vijayakumar L, Chandra PS, Kumar MS, et al. The national suicide prevention strategy in India: context and considerations for urgent action. Lancet Psychiatry 2022 Feb;9(2):160-168. [doi: 10.1016/S2215-0366(21)00152-8] [Medline: 34895477]
43. Nahum-Shani I, Smith SN, Spring BJ, et al. Just-in-time Adaptive Interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. Ann Behav Med 2018 May 18;52(6):446-462. [doi: 10.1007/s12160-016-9830-8] [Medline: 27663578]

Abbreviations

AIIMS: All India Institute of Medical Sciences
EMA: ecological momentary assessment
IRB: institutional review board
LAMP: Learn, Assess, Manage and Prevent
MICT: Multiple Imputation for Categorical Time Series
NIMHANS: National Institute of Mental Health and Neurosciences
PANSS: Positive and Negative Syndrome Scale
PHQ-9: Patient Health Questionnaire-9
SHARP: Smartphone Health Assessment for Relapse Prevention
SI: suicidal ideation

Edited by A Mavragani; submitted 21.10.24; peer-reviewed by GT Wallace, P Harvey; revised version received 15.03.25; accepted 02.04.25; published 29.05.25.

Please cite as:
Bondre AP, Ranjan A, Shrivastava R, Tugnawat D, Chaturvedi NK, Bhan A, Gupta S, Rozatkar AR, Nagendra S, Dutt S, Choudhary S, Reddy PV, Mehta UM, Naslund JA, Torous J
Analyzing Trends in Suicidal Thoughts Among Patients With Psychosis in India: Exploratory Secondary Analysis of Smartphone Ecological Momentary Assessment Data
JMIR Form Res 2025;9:e67745
URL: https://formative.jmir.org/2025/1/e67745
doi:10.2196/67745

XSL•FO
RenderX

© Ameya P Bondre, Aashish Ranjan, Ritu Shrivastava, Deepak Tugnawat, Nirmal Kumar Chaturvedi, Anant Bhan, Snehil Gupta, Abhijit R Rozatkar, Srilakshmi Nagendra, Siddharth Dutt, Soumya Choudhary, Preethi V Reddy, Urvakhsh Meherwan Mehta, John A Naslund, John Torous. Originally published in JMIR Formative Research (https://formative.jmir.org), 29.5.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Norwegian Version of the Self-Efficacy in Clinical Performance Scale (SECP): Psychometric Validation Study

Camilla Olaussen[1], RN, MSc, PhD; Marko Stojiljkovic[2], RN, MSc; Jaroslav Zlamal[1,3], RN, MSc; Tone Nygaard Flølo[2,4], RN, MSc, PhD; Andréa Aparecida Gonçalves Nes[1,5], RN, MSc, PhD

[1]Department of Nursing, Lovisenberg Diaconal University College, Lovisenberggt. 15B, Oslo, Norway

[2]Department of Nursing and Health Promotion, Faculty of Health Sciences, Oslo Metropolitan University, Oslo, Norway

[3]Clinical Simulation Center, Vestfold Hospital Trust, Tønsberg, Norway

[4]Voss Hospital, Haukeland University Hospital, Voss, Norway

[5]Department of Caring and Ethics, Faculty of Health Sciences, University of Stavanger, Stavanger, Norway

Corresponding Author:
Camilla Olaussen, RN, MSc, PhD
Department of Nursing, Lovisenberg Diaconal University College, Lovisenberggt. 15B, Oslo, Norway

Abstract

Background: Previous research has demonstrated a correlation between nursing students' self-efficacy and their clinical performance, competence, and behavior during clinical practice placements. Assessing students' self-efficacy in clinical performance could be a valuable method for identifying areas that need reinforcement and for recognizing students who may require additional support during clinical practice placements.

Objective: This study aimed to translate the Self-Efficacy in Clinical Performance Scale (SECP) from English into Norwegian and to evaluate the psychometric properties of the Norwegian version.

Methods: A cross-sectional study design was used. The SECP was translated into Norwegian following a 6-step process: forward translation, forward translation synthesis, backward translation, backward translation synthesis, cognitive debriefing, and psychometric testing. The validity and reliability of the translated version were assessed using confirmatory factor analysis (CFA), Cronbach α, McDonald ω, and composite reliability.

Results: A total of 399 nursing students completed the Norwegian version of the SECP. The CFA goodness-of-fit indices (χ^2/df ratio=1.578, comparative fit index=0.98, Tucker-Lewis index=0.98, standardized root mean square residual=0.056, root mean square error of approximation=0.038) indicated an acceptable model fit. Reliability measures, including Cronbach , McDonald ω, and composite reliability, were high, with factor-level values ranging from 0.94 to 0.98.

Conclusion: The Norwegian version of the SECP demonstrated strong potential as an instrument for assessing self-efficacy in both current and required competencies among nursing students in clinical practice within nursing education. Future research should aim to confirm the factor structure of the SECP and evaluate its test-retest reliability.

(JMIR Form Res 2025;9:e68173) doi:10.2196/68173

KEYWORDS

clinical performance; self-efficacy; instrument validation; nursing education; psychometric analysis; Norway; psychometric; validation study; competence; clinical practice; translate; translation; cross-sectional study; nursing student; reliability

Introduction

Background

Clinical practice placements in nursing education provide a platform for nursing students to apply theoretical knowledge to real patient care. Nursing knowledge encompasses both theoretical understanding ("knowing that") and practical skills ("knowing how") [1]. The nursing process, originally defined as "a systematic approach to care using the fundamental principles of critical thinking, client-centered approaches to treatment, goal-oriented tasks, evidence-based practice recommendations, and nursing intuition" [2], is recognized as an essential framework for applying nursing knowledge to enhance care quality [3]. The nursing process entails comprehending the rationale behind patients' treatment plans, understanding pathophysiology, identifying patient problems, conducting suitable assessments, prioritizing and implementing care, and evaluating the outcomes of provided care [3]. However, students report difficulties in integrating theoretical knowledge into clinical practice [4,5], as well as challenges in gathering, assessing, and using patient data to develop nursing care that caters to patient needs [6].

XSL•FO
RenderX

Previous studies have demonstrated that nursing students' self-efficacy is associated with their clinical performance, competence, and behavior during clinical practice placements [7-9]. Self-efficacy, a critical factor in academic success, refers to the confidence or belief in one's abilities to successfully accomplish specific tasks and persist despite challenges [10,11]. According to Bandura [10], mastery experiences have a strong effect on a student's self-efficacy development because they are the most authentic indicators of one's capabilities. In contrast, experienced failures may impair students' self-efficacy perceptions and result in avoidance behaviors [10].

Various factors, such as anxiety, stress, motivation, and the pedagogical atmosphere in the clinical setting, can potentially influence nursing students' self-efficacy [12] and thus their clinical achievements [13-15]. A vital resource in promoting nursing students' self-efficacy is emotional support from peers, academic educators, and clinical supervisors [10,12,16]. Hence, to enhance nursing students' clinical learning, performance, and competence during clinical practice placements, it is crucial not only to equip students with a conducive clinical learning environment but also to assist them in boosting their self-efficacy to manage various challenges they might face [8]. As self-efficacy may predict clinical performance and behavior among nursing students, assessing students' self-efficacy could be a valuable method to pinpoint those who may require additional support during clinical practice placements [17,18].

To assess students' self-efficacy in clinical practice, validated instruments are necessary. The Self-Efficacy in Clinical Performance Scale (SECP) was designed to collect empirical data on nursing students' self-efficacy in clinical performance [19]. The SECP explores students' self-efficacy perceptions in performing different facets of the nursing process. Students are queried about their confidence in patient assessments, diagnosis and planning, implementation of care, and evaluation of provided care. Such insights are valuable for educators and clinical supervisors to identify areas requiring reinforcement and to create educational strategies to promote students' self-efficacy, equip them for potential challenges, and ultimately enhance their clinical performance and competence. For these reasons, the SECP was selected for translation and validation into Norwegian in this study, given its established psychometric properties and relevance to clinical education settings [19]. As highlighted by Cheraghi et al [19], the SECP has demonstrated construct validity, internal consistency, and stability, ensuring consistent and reliable measurement of self-efficacy in clinical performance.

Objective

This study was initiated to address the need for culturally adapted and psychometrically validated tools to support educators in enhancing educational strategies in Norwegian nursing education. Therefore, the study aimed to translate the SECP from English into Norwegian and to evaluate its psychometric properties in a Norwegian academic context.

Methods

Design

This study used a cross - sectional survey design, including translating the SECP and testing its psychometric properties.

Translation Procedure

The original SECP, developed and validated by Cheraghi et al [19], is in Persian and consists of 37 items across 4 factors: assessment, diagnosis and planning, implementation, and evaluation. Each item is rated on an 11-point Likert scale, from 0 ("fully disagree") to 10 ("fully agree"), where higher scores indicate greater levels of agreement, and lower scores indicate disagreement. The hypothesized 4-factor model and corresponding items in the SECP are presented in Table 1.

XSL•FO
RenderX

Table . The hypothesized 4-factor model and corresponding items in SECP[a].

Factors and subscales	Items I am confident that in the clinical setting, I can:
Assessment	Q1. Collect significant data in a physical evaluation.
	Q2. Collect relevant data by obtaining the patient's history.
	Q3. Collect data efficiently, without burdening the patient unnecessarily.
	Q4. Collect data by organizing the available time.
	Q5. Collect objective data related to the patient's health status.
	Q6. Collect subjective data related to the patient's health status.
	Q7. See the relationship between data elements collected from different sources.
	Q8. Document the collected data based on the patient's health status.
	Q9. Analyze the data collected based on the patient's health status.
	Q10. Identify the patient's strengths in the care process.
	Q11. Identify the patient's health concerns in the care process.
	Q12. Prioritize the patient's needs based on the patient's health status.
Diagnosis and planning	Q13. Formulate a nursing diagnosis based on the collected data.
	Q14. Adjust the nursing diagnosis based on an assessment of the patient's data.
	Q15. Adjust the nursing diagnosis based on prioritizing the patient's needs.
	Q16. Formulate the overall goal of the patient's plan of care.
	Q17. Formulate short-term goals for the patient's plan of care
	Q18. Formulate long-term goals for the patient's plan of care.
	Q19. Establish measurable outcomes of care.
	Q20. Based on goals, set up the patient's daily plan of care.
	Q21. Establish a plan of care based on prioritizing the patient's needs.
Implementation	Q22. Implement the patient's established plan of care to attain the goals.
	Q23. Provide nursing care to the patient based on priorities in the plan of care.
	Q24. Implement the patient's plan of care with available resources.
	Q25. Explain each nursing intervention to the patient or family member before implementing it.
	Q26. Work together with the patient or family member in implementing the daily plan of care.
	Q27. Make decisions based on my prior experience in similar situations.
	Q28. Seek help from a mentor or nurse colleagues in difficult situations.
	Q29. Improve my skills based on feedback from a mentor and nursing colleagues.
	Q30. Develop teaching strategies for the patient's discharge.
	Q31. Document and report daily clinical work.
Evaluation	Q32. Evaluate the achievement of the patient's desired outcomes.
	Q33. Evaluate how nursing interventions were performed.
	Q34. Identify weaknesses in the structure of the care plan.
	Q35. Based on the patient's prognosis, determine if the plan of care should be followed as is or modified.
	Q36. Adjust the goals of the care plan in response to changes in the patient's condition.
	Q37. Reprioritize the care plan based on changes in the patient's condition.

[a]SECP: Self-Efficacy in Clinical Performance Scale.

Permission to translate, validate, and use the SECP developers' English and Persian versions of the instrument was obtained via email. The SECP was translated from their English version to Norwegian following six of a seven-step guideline suggested by Sousa and Rojjanasrirat [20] in the symmetrical translation approach: (1) forward translation, (2) synthesis of the forward translation, (3) backward translation, (4) synthesis of the backward translation, (5) cognitive debriefing, and (7) comprehensive psychometric testing. Step 6, preliminary

XSL•FO
RenderX

psychometric testing, was omitted due to the lack of a bilingual population.

Forward Translation and Synthesis

The forward translation was conducted independently by 2 translators, both registered nurses and researchers with expertise in the terminology of the area covered by the SECP. The translators were native Norwegian speakers and fluent in English. The 2 forward-translated versions were additionally compared with the original version of the instrument by a third independent translator, who was bilingual and bicultural. Any discrepancies in wording, sentences, and meanings were addressed and resolved through consensus among the translators and the last author (AAGN).

Backward Translation and Synthesis

The backward translation into English was carried out by 2 independent translators. One back translator was a registered nurse and researcher with expertise in the terminology of the area covered by the SECP, while the other was an English-language expert familiar with nursing terminology. Both back translators were native English speakers and were blinded to the original version of the instrument. The 2 back-translated versions were then compared by a third independent translator who was bilingual and bicultural. Any discrepancies in wording, sentences, and meanings were resolved through consensus among the translators and the last author (AAGN). The prefinal Norwegian version of the SECP was subsequently approved by the instrument developer.

Cognitive Debriefing

The prefinal Norwegian version of the SECP was tested with a pilot group of 10 nursing students who had completed clinical practice and represented the target population [20,21]. Each student was asked to evaluate the instructions, response format, and the 37 items of the SECP using a dichotomous scale (clear or unclear). As no unclear issues were identified by the students, no revisions to the instrument were done [22]. Additionally, an expert panel of 10 members assessed the conceptual equivalence (clarity) of the instrument [21]. The panel comprised experienced nurses who were educators holding positions as assistant professors, associate professors, or professors—all registered nurses familiar with the terminology covered by the SECP. Following the same procedure as the pilot group, the panel identified no issues with conceptual equivalence. The expert panel also evaluated each item for content equivalence (relevance). As no items were rated as irrelevant, difficult to assess, or needing minor alteration, no further revisions were necessary [23]. The instrument was then deemed ready for psychometric testing.

Psychometric Testing of the SECP Norwegian Version

Setting and Sample

The study was conducted at one of the biggest universities in Norway, which offers nursing education at the bachelor level. A convenience sampling method was used to recruit participants for psychometric testing of the SECP (Norwegian version). We aimed to recruit a sample of 400 nursing students to ensure at least 10 responses per item in the SECP, accounting for potential withdrawals from the study [20]. Nursing students from the first and second year of the Bachelor of Nursing Education at the university in the spring of 2022 and 2023 (approximately 800 students) were invited to participate by announcing the study through the university's learning platform. To ensure that all invited students had finished at least one of their clinical practice placements, the announcement of the study and a link to the Questback management system [24], which included written information about the study, web-based informed consent forms, and the SECP (Norwegian version), were distributed to the students through the university's learning platform in the second part of their spring semesters.

Data Collection

Data for the psychometric testing of the SECP (Norwegian version) were collected digitally between the spring of 2022 and the spring of 2023 using the Questback management system, which is a web-based survey system [24]. The web-based version of the SECP included the informed consent form. This combination provided data encompassing voluntary participants' names, email addresses, and SECP scores. No additional background data were collected. To avoid missing data, participants were required to complete all SECP items to finalize the survey.

Statistical Analyses

The data were analyzed using the R programming language [25]. The lavaan package was used to compute the goodness-of-fit indices [26], semTools to calculate internal consistency [27], and semPlot to generate the factor structure model of the SECP (Norwegian version) [28].

Internal Consistency

Internal consistency was assessed by Cronbach α, McDonald ω, and composite reliability coefficients, and values of ≥0.7 were classified as satisfactory [29,30].

Construct Validity

The SECP developers specified a 4-factor model of the SECP, as presented in Table 1 [19]. Discriminant validity was evaluated using the Fornell-Larcker criterion, where good discriminant validity is indicated when the square root of the average variance extracted (AVE) for each factor is greater than the correlations between factors [31].

Confirmatory factor analysis (CFA) using the weighted least squares mean and variance-adjusted estimator [32] was performed to evaluate whether the prehypothesized 4-factor model fit our observed data as evidence of construct validity [33]. The following goodness-of-fit indices were used: the χ^2/df ratio, the P value, the comparative fit index (CFI), the Tucker-Lewis index (TLI), the root mean square error of approximation (RMSEA), and the standardized root mean square residual (SRMR). A χ^2/df ratio of ≤2 was considered acceptable [34]. The P value was used to reject a null hypothesis representing a perfect fit [35]. Thus, a nonsignificant P value of >.05 was preferred. The acceptable range of the SRMR index was between 0 and 0.08 [36]. Following Hu and Bentler [36], a CFI and TLI of at least 0.95 was deemed acceptable. Lower RMSEA values indicate a better fit [33], and a value of ≤0.05

XSL•FO
RenderX

with a confidence interval ≤0.1 was considered to represent a close model fit [37].

Ethical Considerations

The Norwegian Centre for Research Data approved the study (reference number 891608). Participation in the psychometric testing of the SECP (Norwegian version) was based on written informed consent and performed in accordance with the 2013 revised version of the Declaration of Helsinki [38]. Participation was voluntary, and participants' rights were clearly outlined in the consent forms. Participants were informed about the nature and purpose of the study, as well as their right to withdraw at any time without any consequences. All data were deidentified and anonymized to ensure confidentiality. Participants were not compensated in any way for their involvement in the study.

Results

Mean Scores, Skewness, and Kurtosis

Of the estimated 800 active nursing students invited to participate, 399 (49.9%) completed and returned the instrument. Answers were skewed toward the "fully agree" end of the scale. The respondents' mean scores, skewness, and kurtosis are provided in Table 2.

Table . Mean score and internal consistency indicators by factor (n=399).

Factor or subscale	Mean (SD)	Skewness	Kurtosis	Cronbach α	McDonald ω	Composite reliability
Assessment	8.37 (1.56)	–1.02	2.48	0.96	0.97	0.96
Diagnosis and planning	8.07 (1.84)	–0.96	1.36	0.98	0.98	0.97
Implementation	8.69 (1.52)	–1.42	3.76	0.94	0.95	0.94
Evaluation	8.29 (1.85)	–1.02	1.50	0.96	0.97	0.96

Internal Consistency

The internal consistency indicators by factor level displayed values from 0.94 to 0.98, indicating high internal consistency. The mean score and internal consistency indicators for each factor are presented in Table 2.

Construct Validity

The factor structure model of the SECP Norwegian version is presented in Figure 1. The content of the items is presented in Table 1.

All factor loadings within each factor in the model presented statistical significance and had acceptable values ranging from 0.44 to 0.82, except for item Q28, which presented significant results, but with a low factor loading of 0.28 (Table 3).

The square roots of the AVE values were equal to or lower than the correlations between the factors, indicating potential issues with discriminant validity, as shown in Table 4.

However, overall, the goodness-of-fit indices from the CFA confirmed the prehypothesized factor structure model, indicating acceptable construct validity. The exception was the significant *P* value, which means that the model did not obtain a perfect fit for the data. The goodness-of-fit indices are shown in Table 5.

XSL•FO
RenderX

Figure 1. Factor structure model for the SECP Norwegian version. Ass: assessment; D_P: diagnosis and planning; Evl: evaluation; Imp: implementation; SECP: Self-Efficacy in Clinical Performance Scale.

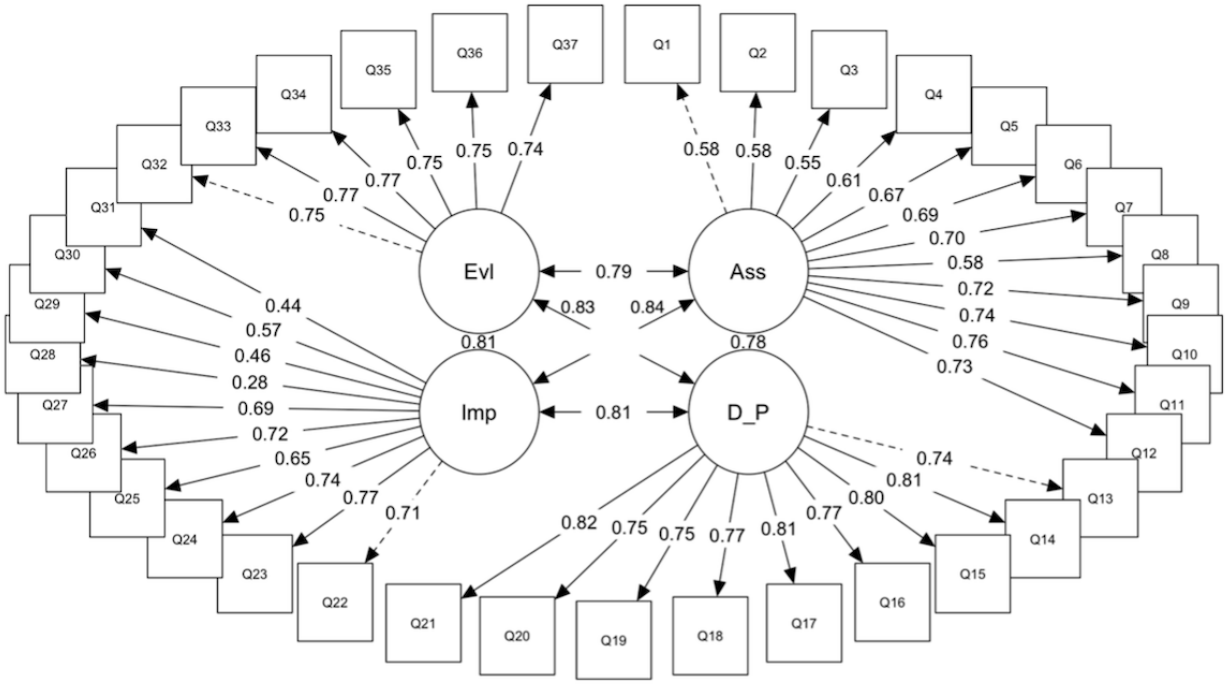

XSL•FO
RenderX

Table . Factor loadings for the factor structure model (n=399).

Factor or subscale and items		λ[a]	SE	*P* value
Assessment				
	Q1	0.58	0.051	<.001
	Q2	0.58	0.048	<.001
	Q3	0.55	0.055	<.001
	Q4	0.61	0.049	<.001
	Q5	0.67	0.042	<.001
	Q6	0.69	0.040	<.001
	Q7	0.70	0.044	<.001
	Q8	0.58	0.042	<.001
	Q9	0.73	0.041	<.001
	Q10	0.74	0.031	<.001
	Q11	0.76	0.036	<.001
	Q12	0.74	0.037	<.001
Diagnosis and planning				
	Q13	0.74	0.033	<.001
	Q14	0.82	0.026	<.001
	Q15	0.80	0.029	<.001
	Q16	0.77	0.035	<.001
	Q17	0.81	0.028	<.001
	Q18	0.77	0.036	<.001
	Q19	0.76	0.037	<.001
	Q20	0.75	0.039	<.001
	Q21	0.82	0.029	<.001
Implementation				
	Q22	0.71	0.034	<.001
	Q23	0.77	0.030	<.001
	Q24	0.74	0.030	<.001
	Q25	0.66	0.039	<.001
	Q26	0.72	0.032	<.001
	Q27	0.69	0.039	<.001
	Q28	0.28	0.043	<.001
	Q29	0.46	0.043	<.001
	Q30	0.57	0.047	<.001
	Q31	0.44	0.043	<.001
Evaluation				
	Q32	0.75	0.035	<.001
	Q33	0.77	0.031	<.001
	Q34	0.77	0.033	<.001
	Q35	0.75	0.036	<.001
	Q36	0.75	0.035	<.001
	Q37	0.74	0.038	<.001

XSL•FO
RenderX

[a]λ: factor loading.

Table . The square root of AVE[a] and correlations between factors (n=399).

	Assessment	Diagnosis and planning	Implementation	Evaluation
Assessment	0.66[b]	0.78	0.84	0.79
Diagnosis and planning	0.78	0.78[b]	0.81	0.83
Implementation	0.84	0.81	0.63[b]	0.81
Evaluation	0.79	0.83	0.81	0.75[b]

[a]AVE: average variance extracted.
[b]These values represent average variance extracted.

Table . Goodness-of-fit indices (n=399).

Goodness-of-fit indices	Value
χ^2	983.246
df	623
χ^2/df ratio	1.578
P value	<.001
CFI[a]	0.986
TLI[b]	0.985
SRMR[c]	0.056
RMSEA[d] (90% CI)	0.038 (0.034-0.043)

[a]CFI: comparative fit index.
[b]TLI: Tucker-Lewis index.
[c]SRMR: standardized root mean square residual.
[d]RMSEA: root mean squared error of approximation.

Discussion

Principal Results

The Norwegian version of the SECP demonstrated acceptability, internal consistency, and satisfactory construct validity.

Acceptability

To ensure that the conclusions drawn from the statistical analyses were based on genuine cultural differences and similarities rather than translation errors, we placed a strong emphasis on maintaining equivalence between the original and translated versions of the SECP throughout the translation process [39]. As the nursing process is accepted as a care standard with the stages of assessment, nursing diagnosis, planning, implementation, and evaluation worldwide [40], we found that identifying Norwegian words and expressions that captured the original meaning of the SECP was not difficult. The target population of nursing students who piloted and evaluated the SECP (Norwegian version), along with the panel of educators, confirmed the relevance of the wording and items, indicating the acceptability of the SECP within the Norwegian context.

Internal Consistency and Construct Validity

The internal consistency and construct validity tests were performed on a sample of 399 respondents. The suggested minimum size for conducting factor analysis differs in relative terms, from 3 to 20 times the number of variables [41]. Bryant and Yarnold [42] suggest that the subjects-to-variables ratio should be at least 5 times the number of variables. In this study, the subjects-to-variables ratio was above 10:1 and thus considered sufficient.

Cheraghi et al [19] used exploratory factor analysis to test and develop the original SECP compositions, resulting in the prehypothesized 4-factor structure model used in this study. In the original SECP, Cheraghi et al [19] presented Cronbach α scores at factor levels ranging from 0.90 to 0.92, which demonstrated high internal consistency. In this study, the internal consistency indicators for the hypothesized factors also indicated high internal consistency, with factor values ranging from 0.94 to 0.98. However, the high internal consistency values (>0.95), especially in the factor “diagnosis and planning,” can also indicate that some of the factor items in the SECP may be redundant [43]. However, although some of the SECP items are somewhat similarly worded, such as “Q14: Adjust the nursing diagnosis based on an assessment of the patient’s data”

XSL•FO
RenderX

and "Q15: Adjust the nursing diagnosis based on prioritizing the patient's needs," they seem to tap into slightly different aspects of the measured constructs within the factors, thus adding additional information.

Our CFA results revealed significant factor loadings for all 4 hypothesized factors. Ideally, factor loadings should be ≥0.7 [44], and this criterion was met by 24 of the 37 items in the SECP (Norwegian version). Tabachnick and Fidell [45] classify factor loadings of ≥0.55 as good, which applied to 34 of the 37 items. According to their criteria, item Q29 ("Improve my skills based on feedback from a mentor and nursing colleague") with a loading of 0.46 and item Q31 ("Document and report daily clinical work") with a loading of 0.44 may be considered "fair." In the original SECP, Cheraghi et al [19] reported that item Q28 ("Seek help from a mentor or nurse colleagues in difficult situations") demonstrated a factor loading of 0.75. In our study, item Q28 exhibited a factor loading of 0.28, indicating that the factor "Implementation" does not adequately account for the variance in this item [44]. Unlike the other items in "Implementation," item Q28 emphasizes seeking help from others. Consequently, it may have been poorly aligned with our respondents' perceptions of mastering nursing practice within "Implementation," potentially contributing to its low factor loading.

Generally, items with low loadings should be considered for removal as they contribute less to the construct and more to measurement error [44]. In this study, we retained all the original items, which may have contributed to our negative results when assessing discriminant validity using the Fornell-Larcker criterion [44]. We found that our estimated square root of AVE values was equal to or lower than the factor correlations, indicating that the SECP factors are not clearly distinct. Retaining items with low factor loadings will result in lower square roots of AVE, as AVE is calculated based on the squared loadings of items, thereby potentially compromising discriminant validity results [44]. However, several other factors beyond items with low loadings can contribute to negative results on the Fornell-Larcker criterion. These include items with similar content, which can lead to high correlations between factors, or a homogeneous sample, which can result in similar responses across factors and mask the true distinction between them [46].

In this study, we did not redefine the original SECP model, as we could not ensure that changes made would be stable and not influenced by the unique characteristics of our single sample. Furthermore, when assessing the goodness-of-fit indicators from the CFA, the construct validity of the SECP (Norwegian version) was considered acceptable. Our χ^2/df ratio was well below the recommended limit set by Byrne [34], indicating a satisfactory model fit [47]. For further evaluation, we assessed the CFI, TLI, SRMR, and RMSEA indices. The RMSEA and CFI are relatively robust in large samples [36], and the RMSEA is adjusted for model complexity [33]. Our RMSEA value indicated a close fit [36,37]. Additionally, the SRMR fell within the acceptable range established by Hu and Bentler [36], while both the CFI and the TLI exceeded the cutoff point of 0.95 suggested by Hu and Bentler [36], indicating a good fit. One goodness-of-fit indicator that opposed the hypothesized model was the chi-square P value, which was <.001. However, significant P values may arise, especially in large samples, even when the proposed model is only slightly inaccurate [47,48]. Furthermore, by using the χ^2/df, CFI, TLI, SRMR, and RMSEA fit indices, we evaluated various aspects of goodness of fit. Together, these fit indices confirmed the hypothesized SECP model, indicating that the overall model structure is sound.

An important step in improving nursing students' clinical performance is to pinpoint both students' strengths and areas requiring improvement. In this first Norwegian translation and testing of the SECP, the results show that the SECP (Norwegian version) has the potential as an instrument within Norwegian nursing education to assess students' self-efficacy in performing different facets of the nursing process. The SECP could be integrated into Norwegian nursing education for clinical course evaluations, and SECP results may be used to guide nurse educators in their work to better equip nursing students for the demands of the clinical setting. The SECP not only tracks nursing students' self-efficacy but also helps to identify learning gaps due to low self-efficacy, measure the impact of teaching interventions, guide curriculum development, and moreover tailor students' learning and subsequently reinforce students' well-being and self-efficacy in clinical performance. The SECP can be administered throughout the study program, for example, before clinical placements, midway through courses, or after targeted workshops. Aggregated results may reveal recurring weaknesses, the need for curriculum updates, and spotlight at-risk students in need of guidance, ultimately improving readiness, resilience, and overall performance in the nursing profession [49,50].

Limitations

The originally validated SECP is in Persian, and in this study, we translated the English version provided by the original SECP developers, Cheraghi et al [19], for which there was no prior validation. We found no other studies that had translated, adapted, or validated the SECP into other languages. Thus, we could not compare this study with studies from different linguistic and cultural contexts, apart from the original study by Cheragi et al [19]. This emphasizes the need for additional research to establish the SECP scale's validity and reliability across diverse populations worldwide.

Additionally, individual characteristics of our participants, such as age or gender, were not collected, which could have provided a more comprehensive overview of the sample. Although the CFA results in this study confirmed the prehypothesized SECP model, it should be noted that having an acceptable fitting model with data from a single sample does not necessarily confirm that the model is correctly specified [47]. Furthermore, we did not evaluate the stability of the SECP (Norwegian version) over time. Therefore, we propose that future psychometric validation studies also include evaluations of the instrument's test-retest reliability.

Conclusion

The SECP (Norwegian version) demonstrated its potential as a viable instrument for assessing self-efficacy in the current and

XSL•FO
RenderX

required competencies of nursing students in clinical practice within Norwegian nursing education. Collecting data in these areas may be crucial to evaluate, develop, and enhance nursing students' clinical performance. The SECP (Norwegian version) indicated internal consistency and acceptable construct validity in this study; however, future research should aim to confirm the factor structure of the SECP (Norwegian version) and evaluate its test-retest reliability.

Acknowledgments

We wish to thank Dr Fatemeh Cheraghi for permission to translate, validate, and use the Self-Efficacy in Clinical Performance (SECP) instrument. We also thank the translators, faculty members, pilot group, expert panel, and nursing students who contributed to the translation and testing of the SECP (Norwegian version).

Authors' Contributions

Conceptualization: JZ, AAGN
Data curation: CO
Formal analysis: CO
Investigation: MS, JZ, AAGN
Project administration: CO
Visualization: CO, AAGN
Writing – original draft: CO, JZ, AAGN
Writing – review & editing: MS, JZ, TNF, AAGN

Conflicts of Interest

None declared.

References

1. Benner P, Wrubel J. Skilled clinical knowledge: the value of perceptual awareness, part 2. J Nurs Adm 1982 Jun;12(6):28-33. [Medline: 6919573]
2. Toney-Butler TJ, Thayer JM. Nursing process. In: StatPearls [Internet]: StatPearls Publishing; 2023. URL: https://www.ncbi.nlm.nih.gov/books/NBK499937 [accessed 2025-04-02] [Medline: 29763112]
3. McCuistion LE, DiMaggio K, Winton MB, Yeager JJ. Pharmacology: A Patient-Centered Nursing Process Approach, 10th edition: Saunders; 2020.
4. Dinmohammadi M, Jalali A, Peyrovi H. Clinical learning experiences of Iranian student nurses: a qualitative study. Nurs Pract Today 2016;3(1):31-39 [FREE Full text]
5. Günay U, Kılınç G. The transfer of theoretical knowledge to clinical practice by nursing students and the difficulties they experience: a qualitative study. Nurs Educ Today 2018 Jun;65:81-86. [doi: 10.1016/j.nedt.2018.02.031]
6. Taskın Yilmaz F, Sabanciogullari S, Aldemir K. The opinions of nursing students regarding the nursing process and their levels of proficiency in Turkey. J Caring Sci 2015 Dec;4(4):265-275. [doi: 10.15171/jcs.2015.027] [Medline: 26744726]
7. Xiang D, Ge S, Zhang Z, Twumwaah Budu J, Mei Y. Relationship among clinical practice environment, creative self-efficacy, achievement motivation, and innovative behavior in nursing students: a cross-sectional study. Nurs Educ Today 2023 Jan;120:105656. [doi: 10.1016/j.nedt.2022.105656]
8. Yu M, Tong H, Li S, Wu XV, Hong J, Wang W. Clinical competence and its association with self-efficacy and clinical learning environments among Chinese undergraduate nursing students. Nurs Educ Pract 2021 May;53:103055. [doi: 10.1016/j.nepr.2021.103055]
9. Alosaimi D. Learning self-efficacy as predictor of nursing students' performance of clinical skills. Educ Sci Theor Pract 2021 Dec;21(3):120-131. [doi: 10.12738/jestp.2021.3.009] [Medline: 2637179889]
10. Bandura A. Self-Efficacy: The Exercise of Control: W. H. Freeman & Co Ltd; 1997.
11. Kang YN, Chang CH, Kao CC, Chen CY, Wu CC. Development of a short and universal learning self-efficacy scale for clinical skills. PLoS One 2019;14(1):e0209155. [doi: 10.1371/journal.pone.0209155] [Medline: 30615610]
12. Henderson A, Harrison P, Rowe J, et al. Students take the lead for learning in practice: a process for building self-efficacy into undergraduate nursing education. Nurs Educ Pract 2018 Jul;31:14-19. [doi: 10.1016/j.nepr.2018.04.003] [Medline: 29730303]
13. Arkan B, Ordin Y, Yılmaz D. Undergraduate nursing students' experience related to their clinical learning environment and factors affecting to their clinical learning process. Nurs Educ Pract 2018 Mar;29:127-132. [doi: 10.1016/j.nepr.2017.12.005] [Medline: 29274524]
14. Keeping-Burke L, McCloskey R, Donovan C, Yetman L, Goudreau A. Nursing students' experiences with clinical placement in residential aged care facilities: a systematic review of qualitative evidence. JBI Evid Synth 2020 May;18(5):986-1018. [doi: 10.11124/JBISRIR-D-19-00122] [Medline: 32813353]

XSL•FO
RenderX

15. Perry C, Henderson A, Grealish L. The behaviours of nurses that increase student accountability for learning in clinical practice: an integrative review. Nurs Educ Today 2018 Jun;65:177-186. [doi: 10.1016/j.nedt.2018.02.029]
16. Chesser-Smyth PA, Long T. Understanding the influences on self-confidence among first-year undergraduate nursing students in Ireland. J Adv Nurs 2013 Jan;69(1):145-157. [doi: 10.1111/j.1365-2648.2012.06001.x] [Medline: 22486537]
17. Yiin SJ, Shen KP, Lai CL, Liang JC. An evaluation of nursing students' learning self-efficacy: a multi-dimensional instrument development and structural validation. Nurs Educ Today 2024 Apr;135:106118. [doi: 10.1016/j.nedt.2024.106118]
18. Bulfone G, Iovino P, Mazzotta R, et al. Self-efficacy, burnout and academic success in nursing students: a counterfactual mediation analysis. J Adv Nurs 2022 Oct;78(10):3217-3224. [doi: 10.1111/jan.15231] [Medline: 35307859]
19. Cheraghi F, Hassani P, Yaghmaei F, Alavi-Majed H. Developing a valid and reliable self-efficacy in clinical performance scale. Int Nurs Rev 2009 Jun;56(2):214-221. [doi: 10.1111/j.1466-7657.2008.00685.x] [Medline: 19646171]
20. Sousa VD, Rojjanasrirat W. Translation, adaptation and validation of instruments or scales for use in cross-cultural health care research: a clear and user-friendly guideline. J Eval Clin Pract 2011 Apr;17(2):268-274. [doi: 10.1111/j.1365-2753.2010.01434.x] [Medline: 20874835]
21. Beaton DE, Bombardier C, Guillemin F, Ferraz MB. Guidelines for the process of cross-cultural adaptation of self-report measures. Spine (Phila Pa 1976) 2000 Dec 15;25(24):3186-3191. [doi: 10.1097/00007632-200012150-00014] [Medline: 11124735]
22. Topf M. Three estimates of interrater reliability for nominal data. Nurs Res 1986;35(4):253-255. [doi: 10.1097/00006199-198607000-00020] [Medline: 3636827]
23. Waltz C, Strickland O, Lenz E. Measurement in Nursing and Health Research, 3rd edition: Springer Publishing Company; 2005.
24. Questback managing software. Questback. 2022. URL: http://www.questback.com [accessed 2025-04-02]
25. R: a language and environment for statistical computing. R Foundation for Statistical Computing. 2021. URL: https://www.R-project.org [accessed 2025-04-02]
26. Rosseel Y. lavaan: an R package for structural equation modeling. J Stat Softw 2012;48(2):1-36. [doi: 10.18637/jss.v048.i02]
27. Jorgensen TD, Pornprasertmanit S, Schoemann AM, et al. semTools: useful tools for structural equation modeling [R package version 05-1]. R Foundation for Statistical Computing. 2018. URL: https://CRAN.R-project.org/package=semTools [accessed 2025-04-02]
28. Epskamp S. semPlot: path diagrams and visual analysis of various SEM packages. R Foundation for Statistical Computing. 2019. URL: https://CRAN.R-project.org/package=semPlot [accessed 2025-04-02]
29. Streiner DL. Starting at the beginning: an introduction to coefficient alpha and internal consistency. J Pers Assess 2003 Feb;80(1):99-103. [doi: 10.1207/S15327752JPA8001_18] [Medline: 12584072]
30. Hayes AF, Coutts JJ. Use omega rather than Cronbach's alpha for estimating reliability. But…. Commun Methods Meas 2020 Jan 2;14(1):1-24. [doi: 10.1080/19312458.2020.1718629]
31. Ab Hamid MR, Sami W, Mohmad Sidek MH. Discriminant validity assessment: use of Fornell & Larcker criterion versus HTMT criterion. J Phys Conf Ser 2017 Sep 1;890:012163. [doi: 10.1088/1742-6596/890/1/012163]
32. Li CH. Confirmatory factor analysis with ordinal data: comparing robust maximum likelihood and diagonally weighted least squares. Behav Res Methods 2016 Sep;48(3):936-949. [doi: 10.3758/s13428-015-0619-7] [Medline: 26174714]
33. Sun J. Assessing goodness of fit in confirmatory factor analysis. Meas Eval Couns Dev 2005 Jan;37(4):240-256. [doi: 10.1080/07481756.2005.11909764]
34. Byrne BM. A Primer of LISREL: Basic Applications and Programming for Confirmatory Factor Analysis Models: Springer-Verlag; 1989. [doi: 10.1007/978-1-4613-8885-2]
35. Browne MW, Cudeck R. Alternative ways of assessing model fit. In: Bollen KA, Long JS, editors. Testing Structural Equation Models: Sage; 1993:136-162.
36. Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Struct Equ Model 1999 Jan;6(1):1-55. [doi: 10.1080/10705519909540118]
37. Schermelleh-Engel K, Moosbrugger H, Müller H. Evaluating the fit of structural equation models: tests of significance and descriptive goodness-of-fit measures. Methods Psychol Res 2003;8(2):23-74. [doi: 10.23668/psycharchives.12784]
38. WMA declaration of Helsinki: ethical principles for medical research involving human subjects. World Medical Association. 2013. URL: https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/ [accessed 2025-04-02]
39. Wang WL, Lee HL, Fetzer SJ. Challenges and strategies of instrument translation. West J Nurs Res 2006 Apr;28(3):310-321. [doi: 10.1177/0193945905284712] [Medline: 16585807]
40. Potter PA, Perry AG, Stockert PA, Hall A. Fundamental of Nursing, 11th edition: Elsevier; 2023.
41. Mundfrom DJ, Shaw DG, Ke TL. Minimum sample size recommendations for conducting factor analyses. Int J Test 2005 Jun;5(2):159-168. [doi: 10.1207/s15327574ijt0502_4]
42. Bryant FB, Yarnold PR. Principal-components analysis and exploratory and confirmatory factor analysis. In: Grimm LG, Yarnold PR, editors. Reading and Understanding Multivariate Statistics: American Psychological Association; 1995:99-136.
43. Tavakol M, Dennick R. Making sense of Cronbach's alpha. Int J Med Educ 2011 Jun 27;2(53-5):53-55. [doi: 10.5116/ijme.4dfb.8dfd] [Medline: 28029643]

XSL•FO
RenderX

44. Hair JF, Black WC, Babin BJ, Anderson RE. Multivariate Data Analysis, 7th edition: Pearson Prentice Hall; 2010.
45. Tabachnick BG, Fidell LS. Using Multivariate Statistics, 6th edition: Allyn & Bacon; 2013.
46. Kline RB. Principles and Practice of Structural Equation Modeling, 4th edition: Guilford Press; 2015.
47. Alavi M, Visentin DC, Thapa DK, Hunt GE, Watson R, Cleary M. Chi-square for model fit in confirmatory factor analysis. J Adv Nurs 2020 Sep;76(9):2209-2211. [doi: 10.1111/jan.14399] [Medline: 32323338]
48. MacCallum RC. 2001 Presidential Address: working with imperfect models. Multivariate Behav Res 2003 Jan;38(1):113-139. [doi: 10.1207/S15327906MBR3801_5]
49. Zengin N, Pınar R, Akinci AC, Yildiz H. Psychometric properties of the self-efficacy for clinical evaluation scale in Turkish nursing students. J Clin Nurs 2014 Apr;23(7-8):976-984. [doi: 10.1111/jocn.12257] [Medline: 23876212]
50. George TP, DeCristofaro C, Murphy PF. Self-efficacy and concerns of nursing students regarding clinical experiences. Nurs Educ Today 2020 Jul;90:104401. [doi: 10.1016/j.nedt.2020.104401]

Abbreviations

AVE: average variance extracted
CFA: confirmatory factor analysis
CFI: comparative fit index
RMSEA: root mean square error of approximation
SECP: Self-Efficacy in Clinical Performance Scale
SRMR: standardized root mean square residual
TLI: Tucker-Lewis index

Edited by A Mavragani; submitted 30.10.24; peer-reviewed by BA Balgiu; revised version received 11.03.25; accepted 11.03.25; published 21.04.25.

Please cite as:
Olaussen C, Stojiljkovic M, Zlamal J, Flølo TN, Nes AAG
The Norwegian Version of the Self-Efficacy in Clinical Performance Scale (SECP): Psychometric Validation Study
JMIR Form Res 2025;9:e68173
URL: https://formative.jmir.org/2025/1/e68173
doi:10.2196/68173

© Camilla Olaussen, Marko Stojiljkovic, Jaroslav Zlamal, Tone Nygaard Flølo, Andrea Aparecida Gonçalves Nes. Originally published in JMIR Formative Research (https://formative.jmir.org), 21.4.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

CoVimmune COVID-19 Immunity Calculator: Web Application Development and Validation Study

Rebecca Slotkin[1], MD; Tassos C Kyriakides[2,3], PhD; Vinni Yu[4], BS; Xien Chen[4], BS; Anupam Kundu[5,6], PhD; Shaili Gupta[1,6], MBBS

[1]Department of Medicine, Yale School of Medicine, 950 Campbell Ave, Bldg 1, Floor 5, West Haven, CT, United States
[2]Cooperative Studies Program Coordinating Center, VA CT Healthcare System, West Haven, CT, United States
[3]Yale School of Public Health, Yale University, New Haven, CT, United States
[4]Department of Computer Science, Yale University, New Haven, CT, United States
[5]Yale Center for Analytical Sciences, Yale School of Public Health, New Haven, CT, United States
[6]Veterans Healthcare System of Connecticut, West Haven, CT, United States

Corresponding Author:
Shaili Gupta, MBBS
Department of Medicine, Yale School of Medicine, 950 Campbell Ave, Bldg 1, Floor 5, West Haven, CT, United States

Abstract

Background: This study illustrates the development of a simple web-based application, which demonstrates the relationship between serum anti-SARS-CoV-2 S1/receptor-binding domain immunoglobulin G (IgG) and anti-SARS-CoV-2 neutralizing antibody (nAb) half-maximal inhibitory concentration (IC50) titers in a vaccinated US adult population and compares them to prior data on nAb titers at different time points after vaccination.

Objective: The objective of this study is to create an easily accessible calculator that uses the results of commercially available anti-SARS-CoV-2 serum IgG to approximate the underlying ability to neutralize SARS-CoV-2.

Methods: Our web-based application leveraged two previously published datasets. One dataset demonstrated a robust correlation between nAb and serum IgG. The other dataset measured nAb titers at specific time periods over a year-long interval following a messenger RNA vaccination primary series and booster vaccine dose. Clinical factors that were statistically significant on a forward linear regression model examining the prediction of nAb from serum IgG were incorporated in the application tool.

Results: By combining the datasets described above, we developed a publicly available web-based application that allows users to enter a serum IgG value and determine their estimated nAb titer. The application contextualizes the estimated nAb titer with the theoretical distance from the corresponding vaccine-mediated antibody protection. Using the clinical variables that had a significant impact on how well IgG values predict nAb titers, this application allows for a patient-centered, nAb titer prediction.

Conclusions: This application offers an example of how we might bring the advances made in scientific research on protective antibodies post-SARS-CoV-2 vaccination into the clinical sphere with practical tools.

(JMIR Form Res 2025;9:e59467) doi:10.2196/59467

KEYWORDS

COVID-19; immunity; neutralizing antibody; immunoglobulin G; vaccine hesitancy; vaccine timing; patient-centered care; web application; vaccination; SARS-CoV-2

Introduction

High levels of anti-SARS-CoV-2 neutralizing antibodies (nAbs) have been shown to be protective against severe SARS-CoV-2 infection [1], but nAbs wane over time following vaccination or infection [2,3]. In addition, nAbs require significant funding and infrastructure to measure and are largely limited to research settings. Serum anti-SARS-CoV-2 immunoglobulin G (IgG) testing is available in commercial settings at relatively low cost. IgG testing has been sporadically ordered in primary care settings; although, there has been little to guide its clinical interpretation [4,5].

Methods

Ethical Considerations

Approval for the study from which both datasets were derived (SARS-CoV-2 Vaccine and You, protocol #1599488) was obtained from the Veterans Affairs Connecticut Healthcare System Institutional Review Board (IRB) in December 2020. Participants were eligible for inclusion if they were veterans or health care workers at the Veterans Affairs Connecticut Healthcare System and excluded if they were unable to provide informed consent. Written informed consent was obtained from

XSL•FO
RenderX

willing participants prior to study participation and they were aware they could opt out at any time. The ethics approval covered secondary analysis as done for this paper without additional consent. Samples were donated by participants without any compensation and we are immensely grateful for this. All identifiable subject data was secured in encrypted folders accessible only to IRB-approved investigators. All data was de-identified before serological analysis.

Source Data Sets

The first dataset was derived from a prospective cohort study which measured anti-SARS-COV-2 nAb half-maximal inhibitory concentration (IC50) titers against pseudotyped WA-1 (the original SARS-CoV-2 strain) at specific time periods over a year-long interval following the primary messenger RNA vaccination series and a subsequent booster vaccine dose (before dose 1, before dose 2, and after dose 2 at 1, 3, 6, and 12 months, with an additional collection 1 month after dose 3, which was approximately 10 months after dose 2). Clinical variables were collected by manual chart review as previously described. The nAb titers were found to fluctuate over time, peaking 1 month after the primary series, waning at 6 months, and rising again 1 month after the booster dose [2].

The second dataset leveraged the collected serum samples and measured IgG anti-S1/receptor-binding domain values in sera with predetermined nAb titers and clinical variables as previously described [6]. IgG and nAb IC50 titers had a robust linear correlation. Multimedia Appendix 1 illustrates this correlation for each clinical variable (COVID-19 infection status, estimated glomerular filtration rate (GFR), age, sex, history of malignancy, and race). Based on a forward stepwise linear regression, GFR (0 - 30, 31 - 59, and >60 mL/min/1.73 m^2), and history of COVID-19 infection within the past 12 months (yes or no) were found to impact this correlation (R^2=0.78) [6].

The model had undergone internal validation using repeated k-fold cross-validation. The model's performance was evaluated using root mean squared error (RMSE), R^2, and mean absolute error (MAE) across multiple versions of the model and different crossvalidation setups (ten 10-fold repeats, five 10-fold repeats, and three 5-fold repeats). For our model, which included IgG, COVID-19 infection status, and GFR, the RMSE was relatively low (around 0.27 - 0.28) across validation setups, suggesting a small average error between the predicted and actual values. The R^2 values were around 0.77 to 0.81 across validation setups, which indicated that the model explained about 77% to 81% of the variability in the data. The MAE values were around 0.2 for all validation setups, demonstrating a low average error magnitude (Multimedia Appendix 2).

Application Development

Backend (Software)

The application used the model-view-controller pattern. We initially used RStudio 2023.12.1+402 running R 4.3.2 (IBM Corp) and an integrated development environment with a set of tools to complement R. The original R script had initially used its own web application framework (Shiny) to create a local web-based application, but this framework was not compatible with the rest of the tools used in the development process. The R script was converted to Python 3.9.12 (Python Software Foundation) for ease of web development. The main challenge when developing the project was converting the original R script into Python and creating the website using an HTML and cascading style sheets (CSS). Python was chosen over R for several reasons. First, Python's versatility and ecosystem made integration with other tools like Flask, a popular web framework that simplifies the web application development process, and Docker, which simplifies the process of containerizing applications, much easier. Second, Python's versatility and rich ecosystem of libraries and web development tools allowed us to keep the backend and front end the same language. Third, Python could handle a high volume of web requests and was better suited for scalable applications, which was a point of consideration. Last, Python was also the better choice for long-term maintainability, as it is more readable and easier to update. The host server emphasized Docker as the primary tool for deployment, making it the easiest tool to use for the development process. The full source code is available on request.

Front End (Web Interface)

The front end was converted from a Shiny package in RStudio to an HTML and CSS. The Python Flask 3.0.0 (Armin Ronacher) module was used to build the CoVimmune website. The website was launched using a YaleInformation Technology Support secured self-service sandbox environment called Spinup. The Waitress Web Server Gateway Interface was used to deploy our application and Docker 4.27.1 (Docker, Inc) was used to package, ship, and run our application.

Results

This web-based application was based on 127 sera samples in 100 unique participants aged 20 - 93 years (mean 63.83, SD 15.63 years; 29% female; 67% White) from the second dataset. The user can input their IgG value, and the units can be from any of 4 commonly used platforms (Abbott-arbitrary U/ml, Siemens-U/ml, Roche-U/ml, and Euroimmune-relative U/ml) or the World Health Organization's standardized IgG value in binding antibody U/mL. The user can choose to select their COVID-19 infection status and GFR category from drop down lists. The application then generates an estimated nAb titer, which is compared to individuals with matching renal function and COVID-19 status from the first dataset, to provide a contextualized range of nAb titers at different time points after vaccination (Figure 1 and Figure 2). The application also visualizes the results in the context of time from the last vaccination among individuals with similar values in the application dataset. The application is free and available on the internet [7].

XSL•FO
RenderX

Figure 1. CoVimmune COVID-19 immunity calculator website interface for the application. GFR: glomerular filtration rate; IC50: half-maximal inhibitory concentration; IgG: immunoglobulin G.

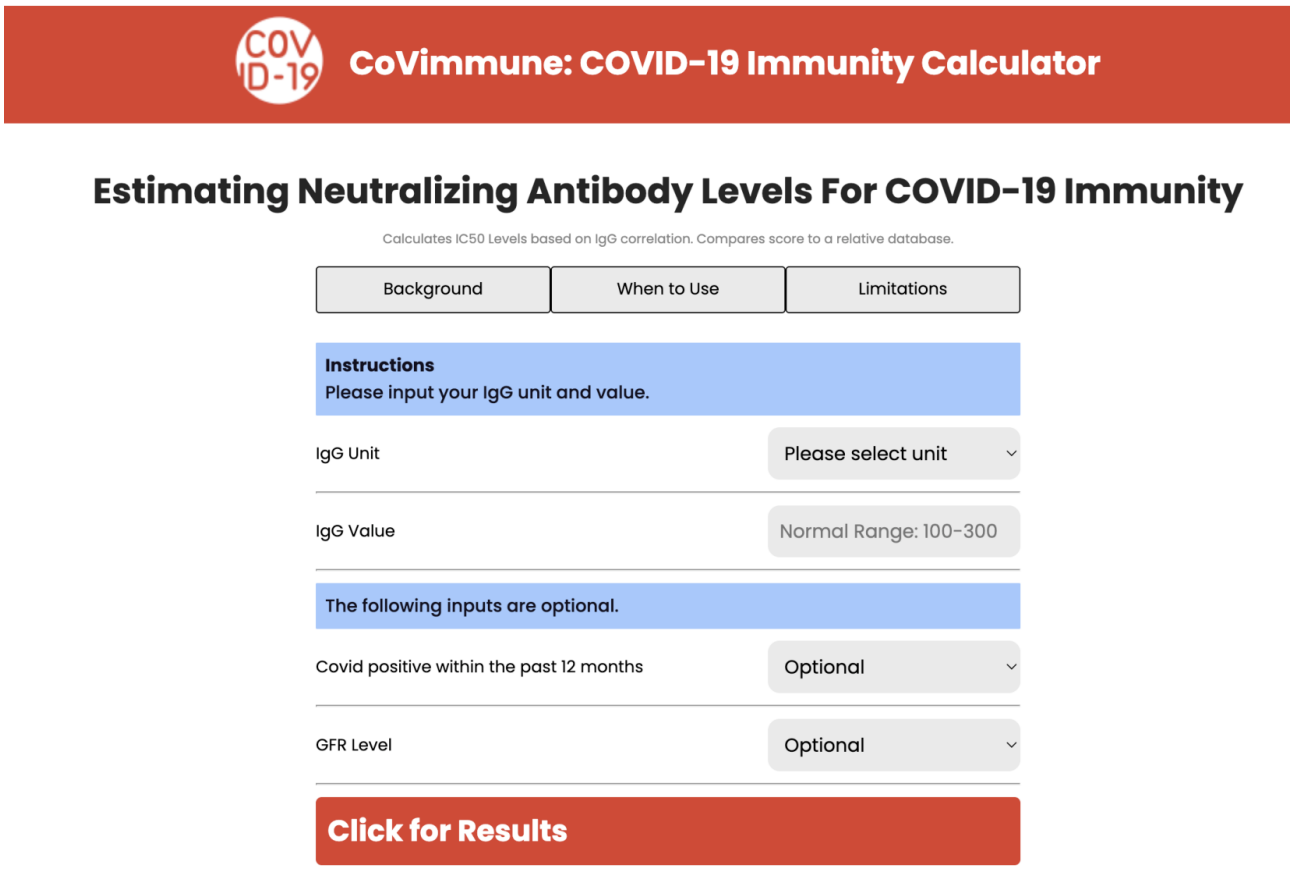

XSL•FO
RenderX

Figure 2. CoVimmune COVID-19 immunity calculator results with an example input. BAU: binding antibody unit; IC50: half-maximal inhibitory concentration; IgG: immunoglobulin G.

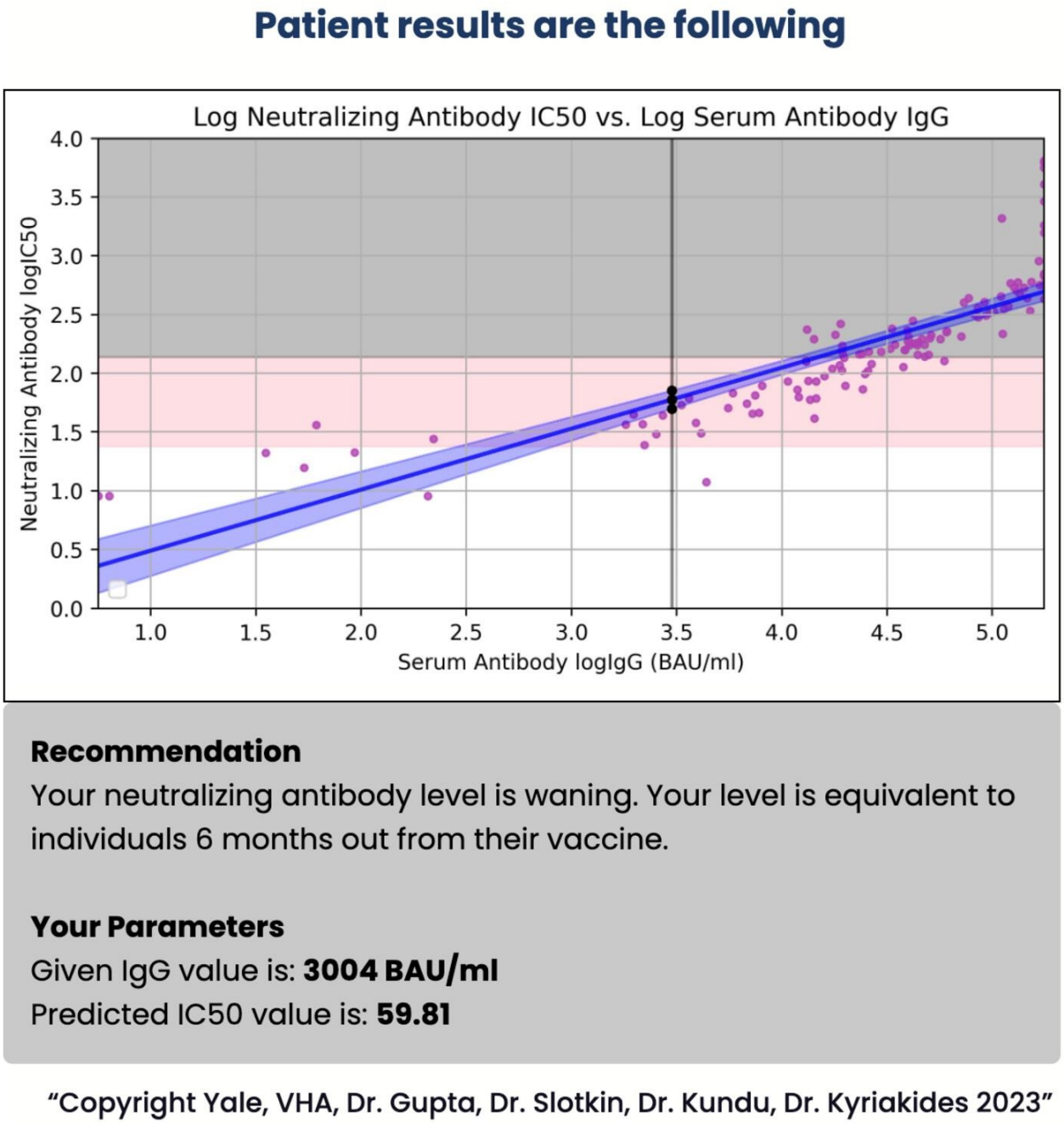

Discussion

The application illustrates the relationship between serum anti-SARS-CoV-2 IgG and anti-SARS-CoV-2 nAb IC50 titers in a vaccinated US adult population and allows for the adjustment of that relationship based on renal function and prior COVID-19 infection. It brings bench research into the clinical context and may provide a tool for physicians to personalize discussions about vaccine booster timing for patients where yearly vaccination may not be appropriate [8] or for those who are hesitant to get revaccinated.

This application is not intended to replace clinical judgment or offer conclusions about immune protection from COVID-19 or vaccine efficacy. The nAb titers are only one component of the immune response to infection. We cannot extrapolate cellular immunity from these data. This application is based on a small sample of patients and there were limited numbers of immunocompromised individuals within the sample, which may have limited generalizability to those individuals [9]. Ultimately, this application is a user-friendly example of how SARS-CoV-2 IgG nAb vaccine research, which has high barriers to replication and clinical application due to cost and infrastructure requirements, can be transformed into a practical tool.

XSL•FO
RenderX

Acknowledgments

We would like to thank Dr Gary Stack and Dr Richard Sutton for their contributions to the studies that form the basis of this application. A big thank you to Yale students VY and XC for their work on website creation. Funding for this study was provided by the National Institutes of Health (5T32AR007107-48 to RS) and (NCT04501978 to SG). The funders of the study had no role in the study design, data collection, data analysis, writing of the report, or in the decision to submit results for publication.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Correlation between the log immunoglobulin G values and the log half-maximal inhibitory concentration titers by clinical variable.
[PNG File, 828 KB - formative_v9i1e59467_app1.png]

Multimedia Appendix 2
Repeated k-fold cross validation results.
[PNG File, 29 KB - formative_v9i1e59467_app2.png]

References

1. Khoury DS, Cromer D, Reynaldi A, et al. Neutralizing antibody levels are highly predictive of immune protection from symptomatic SARS-CoV-2 infection. Nat Med 2021 Jul;27(7):1205-1211. [doi: 10.1038/s41591-021-01377-8] [Medline: 34002089]
2. Zhao M, Slotkin R, Sheth AH, et al. Serum neutralizing antibody titers 12 months after coronavirus disease 2019 messenger RNA vaccination: correlation to clinical variables in an adult, US population. Clin Infect Dis 2023 Feb 8;76(3):e391-e399. [doi: 10.1093/cid/ciac416] [Medline: 35639598]
3. Link-Gelles R, Weber ZA, Reese SE, et al. Estimates of bivalent mRNA vaccine durability in preventing COVID-19-associated hospitalization and critical illness among adults with and without immunocompromising conditions - VISION Network, September 2022-April 2023. MMWR Morb Mortal Wkly Rep 2023 May 26;72(21):579-588. [doi: 10.15585/mmwr.mm7221a3] [Medline: 37227984]
4. Wiencek JR, Head CL, Sifri CD, Parsons AS. Clinical ordering practices of the SARS-CoV-2 antibody test at a large academic medical center. Open Forum Infect Dis 2020 Oct;7(10):faa406. [doi: 10.1093/ofid/ofaa406] [Medline: 33072813]
5. Zarif A, McGagh D, Hoang U, de Lusignan S. The role of antibody tests for COVID-19 in primary care. Br J Gen Pract 2021;71(704):131-134. [doi: 10.3399/bjgp21X715169] [Medline: 33632693]
6. Slotkin R, Kyriakides TC, Kundu A, Stack G, Sutton RE, Gupta S. Correlation of a commercial platform's results with post-vaccination SARS-CoV-2 neutralizing antibody response and clinical host factors. PLoS ONE 2023;18(8):e0289713. [doi: 10.1371/journal.pone.0289713] [Medline: 37643190]
7. Estimating neutralizing antibody levels for COVID-19 immunity. CoVimmune: COVID-19 Immunity Calculator. URL: https://covimmune.yale.edu/ [accessed 2024-04-02]
8. Townsend JP, Hassler HB, Dornburg A. Infection by SARS-CoV-2 with alternate frequencies of mRNA vaccine boosting. J Med Virol 2023 Feb;95(2):e28461. [doi: 10.1002/jmv.28461] [Medline: 36602045]
9. Barnes E, Goodyear CS, Willicombe M, et al. SARS-CoV-2-specific immune responses and clinical outcomes after COVID-19 vaccination in patients with immune-suppressive disease. Nat Med 2023 Jul;29(7):1760-1774. [doi: 10.1038/s41591-023-02414-4] [Medline: 37414897]

Abbreviations

CSS: cascading style sheets
GFR: glomerular filtration rate
IC50: half-maximal inhibitory concentration
IgG: immunoglobulin G
MAE: mean absolute error
nAb: neutralizing antibody
RMSE: root mean squared error

XSL•FO
RenderX

Edited by A Mavragani; submitted 21.04.24; peer-reviewed by A Mohazzab, P Bare; revised version received 03.11.24; accepted 05.11.24; published 22.04.25.

Please cite as:
Slotkin R, Kyriakides TC, Yu V, Chen X, Kundu A, Gupta S
CoVimmune COVID-19 Immunity Calculator: Web Application Development and Validation Study
JMIR Form Res 2025;9:e59467
URL: https://formative.jmir.org/2025/1/e59467
doi:10.2196/59467

© Rebecca Slotkin, Tassos C Kyriakides, Vinni Yu, Xien Chen, Anupam Kundu, Shaili Gupta. Originally published in JMIR Formative Research (https://formative.jmir.org), 22.4.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Increasing the Uptake of Breast and Cervical Cancer Screening Via the MAwar Application: Stakeholder-Driven Web Application Development Study

Nurfarhana Nasrudin[1*], BEng, MEng; Shariff-Ghazali Sazlina[1*], MBBS, MMed, PhD; Ai Theng Cheong[1*], MBBS, MMed, PhD; Ping Yein Lee[2*], MBBS, MMed; Soo-Hwang Teo[3], MA, PhD; Abdul Rashid Aneesa[1], MBBCh BAO, MMed; Chin Hai Teo[2,4], BSc, PhD; Fakhrul Zaman Rokhani[5*], BEng, MSc, PhD; Nuzul Azam Haron[6*], BQS, MSc, PhD; Noor Harzana Harrun[7*], MD, MMed; Bee Kiau Ho[8*], MBBS, MMed; Salbiah Mohamed Isa[8*], MD, MMed

[1]Department of Family Medicine, Faculty of Medicine and Health Sciences, Universiti Putra Malaysia, Jalan Universiti 1, Serdang, Malaysia
[2]UM eHealth Unit, Faculty of Medicine, University of Malaya, Kuala Lumpur, Malaysia
[3]Cancer Research Malaysia, Subang Jaya, Malaysia
[4]Department of Primary Care Medicine, Faculty of Medicine, Universiti Malaya, Kuala Lumpur, Malaysia
[5]Department of Computer and Communication Systems Engineering, Faculty of Engineering, Universiti Putra Malaysia, Serdang, Malaysia
[6]Department of Civil Engineering, Faculty of Engineering, Universiti Putra Malaysia, Serdang, Malaysia
[7]Klinik Kesihatan Pandamaran, Ministry of Health Malaysia, Klang, Malaysia
[8]Klinik Kesihatan Bandar Botanic, Ministry of Health Malaysia, Klang, Malaysia
[*]these authors contributed equally

Corresponding Author:
Shariff-Ghazali Sazlina, MBBS, MMed, PhD
Department of Family Medicine, Faculty of Medicine and Health Sciences, Universiti Putra Malaysia, Jalan Universiti 1, Serdang, Malaysia

Abstract

Background: Digital health interventions such as web health applications significantly enhance screening accessibility and uptake, particularly for individuals with low literacy and income levels. By involving stakeholders—including health care professionals, patients, and technical experts—an intervention can be tailored to effectively meet the users' needs, ensuring contextual relevance for better acceptance and impact.

Objective: The aim of this study is to prioritize the content and user interface appropriate for developing a web health application, known as the MAwar app, to promote breast and cervical cancer screening.

Methods: A cross-sectional study for stakeholder engagement was conducted to develop a web-based application known as the MAwar app as part of a larger study entitled "The Effectiveness of an Interactive Web Application to Motivate and Raise Awareness on Early Detection of Breast and Cervical Cancers (The MAwar study)". The stakeholder engagement process was conducted in a public health district that oversees 12 public primary care clinics with existing cervical and breast cancer screening programs. We purposively selected the stakeholders for their relevant roles in breast and cervical cancer screening (health care staff, patients, and public representatives), as well as expertise in software and user interface design (technology experts). The Quality Function Deployment method was used to reflect the priorities of diverse stakeholders (health care, technology experts, patients, and public representatives) in its design. The Quality Function Deployment method facilitated the translation of stakeholder perspectives into app features. Stakeholders rated features on a scale from 1 (least important) to 5 (most important), ensuring the app's design resonated with user needs. The correlations between the "WHATs" (user requirements) and the "HOWs" (technical requirements) were scored using a 3-point ordinal scale, with 1 indicating weak correlation, 5 indicating medium correlation, and 9 indicating the strongest correlation.

Results: A total of 13 stakeholders participated in the study, including women who had either underwent or never had health screening, a health administrator, a primary care physician, medical officers, nurses, and software designers. Stakeholder evaluations highlighted cost-free access (mean 4.64, SD 0.81), comprehensive cancer information (mean 4.55, SD 0.69), detailed screening benefits (mean 4.45, SD 0.68), detailed screening facilities (mean 4.45, SD 0.68) and personalized risk calculator for breast and cervical cancers (mean 4.45, SD 0.68) as essential priorities of the app. The highest-ranked features include detailed information on screening procedures (weighted score [WS]=367.84), information on treatment options (WS=345.80), benefits of screening (WS=333.75), information about breast and cervical cancers (WS=332.15), and frequently asked questions about the concerns around screening (WS=312.00).

XSL•FO
RenderX

Conclusions: The MAwar app, conceived through a collaborative, stakeholder-driven process, represents a significant step in leveraging digital health solutions to tackle cancer screening disparities. By prioritizing accessibility, information quality, and clarity on benefits, the app promises to encourage early cancer detection and management for targeted communities.

Trial Registration: ISRCTN Registry ISRCTN10403163; https://www.isrctn.com/ISRCTN10403163

(JMIR Form Res 2025;9:e65542) doi:10.2196/65542

KEYWORDS

cancer screening; stakeholder engagement; Quality Function Deployment; web health app

Introduction

Breast and cervical cancers are the most common types of cancer in women in most countries worldwide [1]. In 2020, there were approximately 2.3 million new cases of breast cancer and 660,000 new cases of cervical cancer globally with up to 90% of breast and cervical cancer deaths occurred in low- and middle-income countries [2,3]. In Malaysia, breast and cervical cancer are the two most common cancers among women and result in high morbidity and mortality. In 2022, breast and cervical cancer accounted for about 16% and 4% of new cases and 11% and 3.2% of cancer related deaths, respectively [4].

Despite the availability of effective screening strategies, access to and uptake of screening remains suboptimal, particularly for minority populations and women with low income or low literacy [5]. This disparity often results in a significant gap in the stage of cancer detection, with individuals from lower-income backgrounds presenting at more advanced stages, ultimately affecting survival rates and exacerbating health inequities between high-income and low-income populations [6].

Digital health interventions, such as mobile apps, offer a promising avenue for improving the accessibility and uptake of screening, but its efficacy is highly dependent on the target population [7]. Previous literature has shown that increased involvement of the target population in the development of digital health tools leads to a greater acceptance of their use, but few have involved the target population consistently in the development of such interventions [8-13]. Stakeholder engagement, which encompasses actively involving those who have an interest in a particular decision, is critical to the successful development of digital health interventions [9]. Stakeholder engagement in the development of a mobile app is considered important when structural, cultural, and individual practices can affect its application. This underscores the necessity of addressing these potential barriers during the process of stakeholder engagement in the development of an app [10].

Stakeholder engagement in the development of digital health interventions in the Southeast Asia region is limited. As demonstrated by the MomConnect project conducted at primary care clinics in South Africa, stakeholder engagement is crucial for ensuring the successful implementation, scalability, and sustainability of such interventions [9]. The intended outcome is to achieve a holistic digital health solution that addresses the needs of marginalized communities, improves health services, and overcomes challenges like stakeholder mismanagement, lack of political support, and funding issues. This highlights the importance of obtaining the diverse perspectives of stakeholders that represent the Ministry of Health, clinical staff, and technical experts as well as the primary care service users who are the patients, as in the MomConnect project. This outcome is achieved through continuous collaboration among stakeholders, understanding the digital health ecosystem, and adopting a stakeholder-centered design approach [9-11,13].

Facilitating diverse stakeholder engagement effectively means creating an inclusive environment where all relevant parties, from patients to tech developers, are actively involved [13]. This engagement emphasizes open communication, regular feedback, and collaboration. By doing so, it can enhance information sharing and social interaction, which are crucial components of health app development. By aligning the development process with the needs and interests of the community it seeks to serve, the app's applicability and effectiveness can be enhanced [14]. Furthermore, stakeholder engagement strategies are applicable across a range of community-based health interventions, including the development of an app [15].

Existing digital health interventions have faced challenges in user engagement and content relevance, highlighting the need for apps that are both accessible and grounded in user-centered design [16]. In addition, according to the Malaysia National Health and Morbidity Survey in 2023, about 71% of women have never had a mammogram and 65% of women had not had a Papanicolaou test (Pap smear) in the past 3 years [17]. Further, the proportion of breast and cervical cancers diagnosed at late stages were 50.5% (stage 3) and 47.1% (stage 4) [18]. Hence, to address this gap, we conducted a research project entitled "The Effectiveness of an Interactive Web Application to Motivate and Raise Awareness on Early Detection of Breast and Cervical Cancers (The MAwar study)," a mixed methods feasibility study comprising a pre-post study and a qualitative study that were designed to be conducted in 3 urban primary care clinics in Malaysia. The MAwar study was preceded by a qualitative study that assessed the acceptability of an online web application to promote breast and cervical cancer screenings [19]. The qualitative study involved 15 health care professionals and 25 patients at primary care clinics in urban Malaysia. The findings of the qualitative study highlighted suggestions to incorporate easy-to-use features and visually comprehensible formats (such as more visuals and less text) and provide support for navigating the web app. The content recommended was including individual risk assessment and information on screening benefits, as well as addressing barriers like fear, embarrassment, logistics, and costs for patients. Utilizing the

XSL•FO
RenderX

foundational insights from this study, which focused on factors influencing women's engagement in cancer screenings, we incorporated these findings into the MAwar app's design.

This study aimed to describe the process of stakeholder engagement and prioritize a broad range of content and user interfaces appropriate for the development of a web-based application called MAwar. Through incorporating stakeholder engagement in the application development process, we aimed to create a tool that appropriately meets the needs and expectations of its users, thereby increasing the uptake of breast and cervical cancer screening.

Methods

Design and Setting

We conducted a cross-sectional study for stakeholder engagement between January and May 2022 at a public health district office in Selangor, Malaysia. This stakeholder engagement process allowed the researchers to develop a patient-centered intervention together with key stakeholders and community representatives. The health district office in this study oversees 12 public primary care clinics with existing cervical and breast cancer screening programs. In these clinics, Papanicolaou tests and clinical breast examination services are available. These clinics provide access to ultrasound and mammography services, as well as access to referral for early cancer detection and further evaluation and management in a public hospital in the district through the screening programs. This study was part of a larger research project entitled "The Effectiveness of an Interactive Web Application to Motivate and Raise Awareness on Early Detection of Breast and Cervical Cancers (The MAwar study)." The study protocol was registered with the ISRCTN Registry (ISRCTN10403163).

Participants

In our study, the stakeholders were identified through a discussion with the state Head of Service and with a research organization that was involved in cancer patient navigation programs in the state. The research assistant contacted the relevant individuals and invite them to participate after providing them with participant information. Both verbal and written consent were obtained from the participants prior to the stakeholder engagement.

The stakeholders were purposively selected for their relevant roles in breast and cervical cancer screening. The health care providers selected included primary care doctors (family physicians and medical officers) and nurses involved in cancer screening programs to ensure the app's development was aligned with clinical protocols and needs. We also included patients with varying experiences with screening to obtain insights into user requirements and app functionality needed to enhance the user-centric design of the app [20]. A representative from the Patient Advisory Board was invited to contribute perspectives on patient advocacy and community engagement [21,22]. Complementing this, technical experts were invited to provide their expertise in software and user interface design, ensuring the app was built with the latest technological advancements to improve efficiency and accessibility [23,24]. This purposive and balanced selection of stakeholders aimed to create a multidisciplinary team capable of addressing the multifaceted challenges of cancer screening, ensuring that the MAwar App meets the complex needs of all involved stakeholders [23,24].

Process of Stakeholder Engagement

The stakeholder engagement used a stakeholder-driven approach was based on the INVOLVE UK guidelines on patients and public involvement [21] and the Australian Government Department of Health's Stakeholder Engagement Framework [22]. The rationale for integrating these 2 distinct frameworks stems from their complementary strengths. The Australian framework offers a robust structure for systematically identifying and involving stakeholders in health-related projects, ensuring transparency and accountability [22]. On the other hand, the INVOLVE guidelines provide a specialized focus on patient and public involvement, emphasizing the importance of including the voices of end users in health research to enhance relevance and impact [21].

By merging these approaches, the engagement strategy was enriched, leveraging the structured engagement and accountability. This hybrid approach facilitated a more comprehensive understanding of stakeholder needs, enabling the development of an application that is not only technically sound but also deeply meaningful and aligned with the community's expectations and experiences.

To further validate this approach, an adapted version of the combined framework was reviewed by a panel of experts comprising 3 family physicians with significant experience in stakeholder engagement and patient and public involvement. This expert review ensured that the engagement strategy was not only theoretically sound but also viable and tailored to the specific context of the health app development. It comprised 5 strategic steps. These steps systematically capture the essence of our community-centric development of the MAwar application, each underscored by robust, participatory principles. Table 1 explains the 5 steps of the stakeholder engagement process for the MAwar application development.

XSL•FO
RenderX

Table . Five-step stakeholder engagement process for MAwar app development.

Step	Phase	Action	Study element	Timeline
1	Identify stakeholders	Engage diverse groups with a stake in app outcomes	Health care professionals, patients, public representatives, and tech experts identified	January 2022
2	Define methods of engagement	Determine stakeholder engagement depth using appropriate tools	QFD[a] framework and online survey tool were adopted for the group engagement	February 2022
3	Develop and implement engagement plan	Design detailed engagement plans tailored to stakeholder needs	Engagement plan was developed from the preliminary qualitative study insights [14]	March 2022
4	Evaluate the process	Gather and assess stakeholder feedback	Through QFD framework, prioritization of the top 5 user requirements and correlation of these with the technical requirements	April 2022
5	Provide feedback and follow-up	Share results for continuous improvement	Feedback loop established for app feature refinement	May 2022

[a]QFD: Quality Function Deployment.

Quality Function Deployment Framework for MAwar Application Design

The Quality Function Deployment (QFD) framework is a customer-driven planning process that guides the design of products and services, as it was originally designed for business [25]. This methodology was helpful in refining the MAwar application, translating insights from the previous study into specific design specifications to ensure the application effectively meets user needs. By applying QFD principles, our team was able to prioritize features that directly address user-reported barriers and motivators, enhancing the app's usability and impact on cancer screening uptake. The QFD process facilitated a structured approach to aligning the app's development with the expressed needs, challenges, and preferences of potential users, as identified in the preliminary qualitative study [19]. In addition, we used the term "user" to describe the "customer" as used in the QFD framework in order to address our study participants more appropriately.

The QFD process began with defining the user needs (the WHATs) and the technical requirements (the HOWs) based on the qualitative study [19] and incorporating these in the QFD framework. Table 2 shows the 25 user needs (the WHATs) identified, capturing essential requirements for the app based on barriers and motivators for cancer screening.

Table 3 presents the 28 technical requirements (the HOWs) from the QFD framework, detailing how each user's need is addressed through specific app functionalities and features.

XSL•FO
RenderX

Table . List of WHATs (user requirements) using the Quality Function Deployment framework.

Number	User requirements
1	Free to use (without cost)
2	Information about breast and cervical cancer (statistics, causes, signs and symptoms, risk factors, complications)
3	Information on the benefits of screening (survival rate with/without screening)
4	Information on screening facilities: who, where, recommendations, opening hours
5	A risk assessment function to know about my personal risk of breast and cervical cancers
6	Data are kept confidential and secure
7	Information about screening procedure for breast and cervical cancers
8	A list of frequently asked questions about the concerns about breast and cervical cancer screening
9	Information is presented in video format
10	Only information related to me is shown instead of all information (personalized)
11	Reminder function for appointment using, for example, WhatsApp
12	Information about cost of screening for breast and cervical cancers
13	Information is presented in picture format
14	Health screening appointment settings with health care center
15	Information on treatments for breast and cervical cancers
16	Available in 2 languages (Malay and English)
17	Testimonies from people who have undergone screening
18	Accessible using a QR code
19	Latest updates on ongoing health screening program locally
20	Access to a support group on social media via the web app
21	Motivational quotes to empower women to go for screening
22	Can interact with counsellors/health care providers via the web app
23	Information format is easy to understand and navigate
24	Reminder function for annual check-up
25	Use of a credible doctor as the avatar in the web app

XSL•FO
RenderX

Table . List of HOWs (technical requirements) using the Quality Function Deployment framework.

Number	Technical requirements
1	Funds (research grants, sponsors, etc)
2	Expert inputs (ie, Ministry of Health, institutes of higher learning)
3	Information (clinical evidence, literature review, clinical practice guide)
4	Ministry of Health website (list of locations of health clinics)
5	Expert input (encrypt data, access control, https)
6	SMS services (code and telco provider)
7	Video placement (web, video uploading platforms, ie, YouTube, Vimeo)
8	Testimonial video (patients, experience)
9	Video animation (Powtoon, Vyond)
10	Subtitles
11	Filtering information
12	Integration with WhatsApp
13	Screening center website (government and private)
14	Preliminary findings (barriers, motivators)
15	Image placement (web)
16	Image format (JPEG, GIF)
17	Related links (clinic appointment system, eg, QueueMed, Encoromed, BookDoc)
18	Content creator (translator)
19	Dual page
20	QR code generator
21	Login function (name, risk profile, age, history, recommendation, only needing to log in once, data are stored)
22	Access for medical practitioner to post screening campaign on the web app (post web app to social medias)
23	Related links in support group (on social media) and web app
24	Chatbox on web app
25	Voice app (VoIP, dial in)
26	Text format (eg, Arial)
27	Integration with email (for appointment reminder)
28	Celebrity face

Following this step, we integrated the QFD method with an online polling tool (VEVOX [26]) to gather feedback from the participants. They were asked to rate the importance of each feature on a scale from 1 to 5 (1 being the least important and 5 being the most important). Each rating was used to assign a “user importance value” (UIV) to each feature in the QFD, ensuring that the design of the MAwar application aligned with the needs and expectations of its users.

Then, the participants were asked to score the correlations between the WHATs (user requirements) and the HOWs (technical requirements). The correlations were scored using a 3-point ordinal scale, with 1 indicating weak correlation, 3 indicating medium correlation, and 9 indicating the strongest correlation [27]. This scoring system provided a quantitative measure of the alignment between the stakeholders’ needs and the application’s design features.

Ethical Considerations

This study obtained ethics approval from the Malaysia Ministry of Health Medical Research Ethics Committee (NMRR-21-951-58339 [IIR]) as well as permission from the State Health Department. Participants who voluntarily agreed to participate provided verbal and written informed consent after they had read the participant information sheet. We provided honoraria to participants as reimbursement for their travel to compensate for their time and participation in the study. Nonidentifiable identification codes were assigned to the participants for the purpose of data entry and data analysis. The consent forms and questionnaires will be stored in a locked

XSL•FO
RenderX

filing cabinet for 7 years and will only be accessible by the research team. After 7 years, these documents will be shredded and disposed of in secure bins. In the report writing or publication, the participants will not be identified.

Data Analysis

Data were analyzed descriptively using Microsoft Excel (Microsoft Corp). The QFD analysis ranked the technical requirements of the MAwar application by assessing the UIVs and correlation scores (CS). The UIVs were calculated based on the average of the participants' ratings of the importance of each feature on a scale from 1 (least important) to 5 (most important). The CS were rated using a 3-point ordinal scale (1=weak, 3=medium, and 9=strong correlations) to measure how each technical requirement supports user needs [21].

The technical elements were identified using the absolute importance value. These were calculated for the sum of UIV × CS for each of the "HOWs," showing the direct contribution of each technical requirement. These AI values were then summed across all relevant user needs to compute the weighted score (WS)=sum of absolute importance value, ranking the significance of each feature of the "WHATs" within the overall app structure. All these were ranked to highlight the most essential components of the MAwar application [21]. These analytical steps provided a detailed, quantified overview of the key priorities for the MAwar application's development. It identified the features most valued by stakeholders and pinpointed the technical requirements crucial to delivering these features, thus ensuring the application's design perfectly matched the needs and preferences of its intended users. Given the extensive data from the QFD (see the complete QFD analysis in Multimedia Appendix 1), we presented the top outcomes of the list of WHATs to represent the users' needs in the Results section to focus on the most significant findings, ensuring both clarity and relevance in our discussion of key impacts.

Results

Overview

We engaged 13 stakeholders, which was within the acceptable group size of 10 - 15 participants for achieving a balance of diverse perspectives and depth in discussions for stakeholder engagement [28]. The participants included 4 female patients, one of whom serves on the Patient Advisory Panel, who provided insights into patient experiences [23]. In the public health care sector, 6 participants included medical doctors and nurses, who brought specialized knowledge to the project based on their involvement in providing services to women related to health screening and education [24]. Additionally, 3 technical experts contributed their software and user interface design expertise, ensuring the application was developed with the latest technological advancements to improve efficiency and accessibility (see Table 4) [29].

Table . Stakeholders' classifications and roles.

Stakeholder group (subgroup)	Participants, n	Roles
Patient and public involvement (patients)	4	Advisory panel member for primary care clinic, women who had either underwent or never went for health screening
Health care professionals (medical doctors)	4	Administrator at Health District Office, a primary care physician, and 2 medical officers involved in the development of protocols and counseling on health screenings
Health care professionals (nurses)	2	Involved in health screening and education
Technical experts (software designer)	3	Involved in application development and user interface design

Users' Needs

Table 5 shows the top 20 user needs for the development of the MAwar application based on QFD analysis (refer to Multimedia Appendix 1 for the complete QFD analysis). The main 5 features were that the application should be free to use and provide comprehensive cancer information, including details on the benefits of screening, screening facilities, and personal risk assessment. Data security and detailed screening procedures were also deemed as major components. Educational features included Frequently Asked Questions, video-based facts, and personalized content. Although reminder functions for screening appointments were also rated highly, they were not applicable for our web application and are more valid for mobile apps.

Additional user needs include information on screening costs, cancer prevention strategies, and pictorial representations of key details. Users also highlighted the importance of appointment booking, cancer treatment information, and bilingual support in Malay and English. Other notable features include testimonials from screened individuals or cancer survivors and QR code access for convenience.

These findings provide crucial insights into user expectations, ensuring the MAwar application is developed with a user-centric approach, prioritizing accessibility, education, and engagement. However, since the MAwar application is a web application, features such as a reminder function for appointments using SMS text messaging or WhatsApp were not able to be incorporated into the application.

XSL•FO
RenderX

Table . The list of top 20 users' needs ranked by user importance value based on the Quality Function Deployment framework.

List of WHATs	User importance value, mean (SD)	Rank
Free to use (without cost)	4.64 (0.81)	1
Information about breast and cervical cancers (statistics, causes, signs and symptoms, risk factors, complications)	4.55 (0.69)	2
Information on the benefits of screening (survival rate with/without screening)	4.45 (0.68)	3
Information on screening facilities: who, where, recommendations, opening hours	4.45 (0.68)	4
A risk assessment function to know about my personal risk of breast and cervical cancers	4.45 (0.68)	5
Data are kept confidential and secure	4.45 (0.68)	6
Reminder function for appointment using, for example, SMS text messaging	4.36 (0.65)	7
Information about screening procedure for breast and cervical cancers	4.18 (0.60)	8
A list of frequently asked questions about the concerns about breast and cervical cancer screening	4.00 (0.89)	9
Information is presented in video format	4.00 (0.89)	10
Only information related to me is shown instead of all information (personalized learning)	4.00 (0.89)	11
Reminder function for appointment using, for example, WhatsApp	4.00 (0.89)	12
Information about cost of screening for breast and cervical cancers	3.91 (0.72)	13
Information on prevention for breast and cervical cancers (lifestyle factors like diet and exercise)	3.91 (0.72)	14
Information is presented in picture format	3.91 (0.72)	15
Function to set appointments with health screening center	3.91 (0.72)	16
Information on the treatments for breast and cervical cancers	3.64 (0.61)	17
Available in 2 languages (Malay and English)	3.64 (0.61)	18
Testimonies from people who have undergone screening or cancer survivors	3.36 (0.58)	19
Accessible using a QR code	3.36 (0.58)	20

The MAwar Application's Features

The QFD analysis identified the top 5 priority ranks of technical requirements relevant for the MAwar application's features for the health information platform, specifically targeting breast and cervical cancer awareness (see Table 6). These technical requirements ensure the MAwar application provides credible, accessible, and medically proven information, incorporating expert inputs, government resources, multimedia engagement, and seamless connectivity with health care services. Expert inputs from the Ministry of Health and institutions of higher learning form the foundation of medically reviewed content, ensuring accuracy and trust. The information component, including clinical evidence and literature reviews, strengthens the app's role as an educational tool. Ministry of Health integration facilitates access to official screening locations and health care services, while video content enhances user engagement through educational animations and testimonials. Lastly, integration features such as the WhatsApp application programming interface and screening center connections support direct interaction with health care providers. These prioritized technical elements enable the MAwar application to effectively promote breast and cervical cancer awareness, supporting early detection and informed decision-making among users.

XSL•FO
RenderX

Table . The top 10 technical requirements ranked by weighted score from the Quality Function Deployment framework for the MAwar application.

User requirements (list of WHATs)	Weighted score	Rank
Information about screening procedure for breast and cervical cancers	367.84	1
Information on the treatments for breast and cervical cancers	345.80	2
Information on the benefits of screening (survival rate with/without screening)	333.75	3
Information about breast and cervical cancer (statistics, causes, signs and symptoms, risk factors, complications)	332.15	4
A list of frequently asked questions about concerns about breast and cervical cancer screening	312.00	5
Information is presented in video format	284.00	6
Information on screening facilities: who, where, recommendations, opening hours	275.90	7
A risk assessment function to know about my personal risk of breast and cervical cancers	191.35	8
Information about the cost of screening for breast and cervical cancers	156.40	9
Function to set appointments with a health screening center[a]	148.58	10

[a]Feature evaluated but not applicable to the web application.

Discussion

Principal Findings

The findings from our stakeholder engagement underscored the crucial role of targeted, comprehensive information delivery in web health applications like the MAwar application. With the increasing accessibility of digital technologies, web-based and mobile health interventions have emerged as effective tools to overcome barriers to breast and cervical cancer screening. Several apps, such as BrAware, Amate, and iBreastCheck, have shown promising results to improve intention and uptake for cancer screenings [30-32]. The BrAware app has been effective in increasing knowledge of breast cancer risk factors and improving awareness of warning signs, as demonstrated in a study conducted in northeast Peninsular Malaysia [30]. Similarly, the Amate app in Colombia aims to improve access to early detection of both breast and cervical cancers by educating patients about and guiding them to national screening programs, resulting in the successful completion of recommended screening tests among some participants [31]. These interventions provided tailored health information, risk education, and reminders, which have been shown to improve health behaviors in various contexts. The prioritization of detailed cancer information, screening benefits, and interactive features like personal risk assessments align with previous literature suggesting that informed users are more likely to participate actively in their health management, potentially improving outcomes [33]. The emphasis on screening procedures and frequently asked questions further highlights the necessity for health apps to be both educational and reassuring, addressing common concerns and misconceptions that can hinder screening uptake [34].

User Needs

Providing comprehensive and accurate information about breast and cervical cancer through health applications is essential for enhancing user engagement and promoting better health outcomes. Detailed educational content helps users understand the importance of early detection and treatment, encouraging participation in screening programs. A mobile app designed for Colombian women has shown that tailored educational content and risk assessments can significantly raise awareness and participation rates in breast and cervical cancer screening programs [28]. Moreover, providing robust information on the benefits of screening empowers users to take proactive steps in cancer prevention and screening, leading to improved health outcomes [19,35].

Integrating personalized risk assessment functions into health applications enhances user engagement by providing tailored insights into specific health risks, thereby motivating preventive behaviors. Personalized health tools have been shown to significantly increase user interaction and satisfaction with health apps [36]. Additionally, providing detailed information about screening procedures for breast and cervical cancer helps demystify the process and encourages participation by alleviating fears and misconceptions. Clear, accessible information about screening can significantly increase the uptake of screening services [37]. Furthermore, comprehensive frequently asked questions sections addressing common concerns and misconceptions about breast and cervical cancer screenings can provide reassurance, reduce anxiety, and improve user confidence in the procedures. This approach has been proven to improve user satisfaction and participation in screening programs [35,38].

XSL•FO
RenderX

Features of the MAwar Application

The leading features identified for the MAwar application reflect a strategic emphasis on ensuring that content is not only trustworthy but also visually engaging and well-structured. By focusing on expert inputs, text formatting, direct contributions from content creators, strategic image placement, and image formatting, we aimed to enhance user comprehension and engagement. This approach facilitates a better understanding of critical health information, which is crucial for informed decision-making among users. Experts inputs ensure that the information is accurate and reliable, while strategic text formatting and visual content placement make the information accessible and appealing. Studies have shown that well-structured and visually appealing information presentation can greatly increase the understandability and usability of health apps [39].

By integrating these top technical requirements, the MAwar application systematically prioritizes development efforts to meet user needs effectively. Expert contributions ensure content accuracy, while strategic image placement and formatting enhance visual engagement, making complex medical information more digestible. For instance, health applications that included expertly curated content and optimized visual presentation have been found to improve users' satisfaction and interaction significantly [40]. Moreover, text formatting plays a critical role in making information more understandable, as demonstrated by studies on health care applications [41]. Strategic image placement further aids in improving user comprehension and retention of health information [42]. These features collectively help address the challenges of conveying complex health information through digital platforms, ultimately supporting users in making informed health decisions [35].

Stakeholder Engagement and Public and Patient Involvement

Incorporating a broad spectrum of stakeholders through stakeholder engagement and public and patient involvement has added significant value to the MAwar application development. By engaging medical professionals, patients, and technical experts, the development process benefited from a holistic view that encompassed diverse perspectives, enhancing the app's relevance and usability across different user groups. Previous studies have highlighted that stakeholder engagement and public and patient involvement contribute to higher product acceptance and greater adherence to health interventions, as these processes ensure that the end product aligns closely with the real-world needs and expectations of its users [43].

The inclusion of technical experts in the stakeholder engagement process not only ensures that the app is built using the latest technological advancements but also guarantees that the technical execution matches the high standards required for health applications. This collaboration between health professionals and technologists fosters innovation, leading to more effective and user-friendly solutions, ultimately improving uptake of and adherence to recommended health practices [44].

The MAwar application, conceived through a collaborative, stakeholder-driven process, represents a significant step in leveraging digital health solutions to tackle cancer screening disparities [44]. By prioritizing accessibility, information quality, and clarity on benefits, the app directly confronts the barriers impeding cancer screening in underserved populations [45]. The stakeholder engagement underscored the application's relevance and potential impact, showcasing the power of inclusive design in improving health outcomes. The MAwar application focused on critical features reflecting user needs to empower women to take part in early cancer detection and management for targeted communities.

Strengths and Limitations of the Study

The stakeholder engagement process played a crucial role in identifying key features that are critical for meeting the needs and expectations of the MAwar application users. This approach ensured that the application was designed with a user-centric focus, increasing its potential effectiveness and usability.

However, there were some limitations. The patient representatives in our stakeholder engagement session did not include older clients aged 50 years and above. This age group is a significant demographic for breast and cervical cancer screening, and their unique needs and perspectives may not have been fully captured in our study. However, their perspective was included in an earlier qualitative study, the results of which were used for the discussion during the stakeholder engagement in this study [14].

Additionally, family members of clients were not involved in the stakeholder engagement process. As potential co-users of the app, their input could have provided valuable insights into the design and functionality of the application. Future research should consider involving these important stakeholders to ensure a more comprehensive understanding of user needs and preferences.

Conclusion

By focusing on features such as cost-free access, comprehensive cancer information, and detailed screening benefits, the application directly targets the primary obstacles to screening adherence, thereby fostering improved health outcomes in targeted communities. To enhance the app's efficacy and reach, future development efforts must prioritize the inclusion of a broader demographic spectrum, especially older individuals and potential co-users like family members. Such inclusivity will ensure the application's design and functionalities resonate more effectively with the diverse needs of its user base, ultimately maximizing its public health impact.

In summary, the MAwar application represents a strategic innovation in digital health, aiming to reduce disparities in cancer screening uptake through a user-centered design. Its foundation, built on rigorous stakeholder engagement and evidence-based design principles, sets a precedent for future public health interventions aimed at promoting early cancer detection and management.

XSL•FO
RenderX

Acknowledgments

The authors would like to thank the Ministry of Higher Education and the Ministry of Health Malaysia for their support and permission to publish this article. This project was funded by the Ministry of Higher Education Fundamental Research Grant Scheme (FRGS/1/2020/SKK05/UPM/02/5).

Data Availability

All data have been presented in this manuscript and Multimedia Appendix 1.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Quality Function Deployment framework analysis.
[PDF File, 159 KB - formative_v9i1e65542_app1.pdf]

References

1. Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2021 May;71(3):209-249. [doi: 10.3322/caac.21660] [Medline: 33538338]
2. Arnold M, Morgan E, Rumgay H, et al. Current and future burden of breast cancer: global statistics for 2020 and 2040. Breast 2022 Dec;66:15-23. [doi: 10.1016/j.breast.2022.08.010] [Medline: 36084384]
3. Singh D, Vignat J, Lorenzoni V, et al. Global estimates of incidence and mortality of cervical cancer in 2020: a baseline analysis of the WHO Global Cervical Cancer Elimination Initiative. Lancet Glob Health 2023 Feb;11(2):e197-e206. [doi: 10.1016/S2214-109X(22)00501-0] [Medline: 36528031]
4. Ferlay J, Ervik M, Lam F, et al. Global cancer observatory: Malaysia 2022. International Agency for Research on Cancer, World Health Organization. 2024. URL: https://gco.iarc.who.int/today [accessed 2025-03-24]
5. Woldetsadik AB, Amhare AF, Bitew ST, Pei L, Lei J, Han J. Socio-demographic characteristics and associated factors influencing cervical cancer screening among women attending in St. Paul's Teaching and Referral Hospital, Ethiopia. BMC Womens Health 2020 Apr 6;20(1):70. [doi: 10.1186/s12905-020-00927-5] [Medline: 32252733]
6. Ji P, Gong Y, Jiang CC, Hu X, Di GH, Shao ZM. Association between socioeconomic factors at diagnosis and survival in breast cancer: a population-based study. Cancer Med 2020 Mar;9(5):1922-1936. [doi: 10.1002/cam4.2842] [Medline: 31960597]
7. Richardson-Parry A, Silva M, Valderas JM, Donde S, Woodruff S, van Vugt J. Interactive or tailored digital interventions to increase uptake in cervical, breast, and colorectal cancer screening to reduce health inequity: a systematic review. Eur J Cancer Prev 2023 Jul 1;32(4):396-409. [doi: 10.1097/CEJ.0000000000000796] [Medline: 37144585]
8. Radu I, Scheermesser M, Spiess MR, Schulze C, Händler-Schuster D, Pehlke-Milde J. Digital health for migrants, ethnic and cultural minorities and the role of participatory development: a scoping review. Int J Environ Res Public Health 2023 Oct 23;20(20):6962. [doi: 10.3390/ijerph20206962] [Medline: 37887700]
9. Sibuyi IN, de la Harpe R, Nyasulu P. A stakeholder-centered mHealth implementation inquiry within the digital health innovation ecosystem in South Africa: MomConnect as a demonstration case. JMIR mHealth uHealth 2022 Jun 16;10(6):e18188. [doi: 10.2196/18188] [Medline: 35708756]
10. Laird Y, Manner J, Baldwin L, et al. Stakeholders' experiences of the public health research process: time to change the system? Health Res Policy Syst 2020 Jul 18;18(1):83. [doi: 10.1186/s12961-020-00599-5] [Medline: 32682426]
11. Shrestha A, Tamrakar D, Shrestha B, et al. Stakeholder engagement in a hypertension and diabetes prevention research program: description and lessons learned. PLOS ONE 2022;17(10):e0276478. [doi: 10.1371/journal.pone.0276478] [Medline: 36264860]
12. Stelfox HT, Niven DJ, Clement FM, et al. Stakeholder engagement to identify priorities for improving the quality and value of critical care. PLOS One 2015;10(10):e0140141. [doi: 10.1371/journal.pone.0140141] [Medline: 26492196]
13. Chen SCI, Liu C, Wang Z, et al. How geographical isolation and aging in place can be accommodated through connected health stakeholder management: qualitative study with focus groups. J Med Internet Res 2020 May 27;22(5):e15976. [doi: 10.2196/15976] [Medline: 32459181]
14. Petkovic J, Magwood O, Lytvyn L, et al. Key issues for stakeholder engagement in the development of health and healthcare guidelines. Res Involv Engagem 2023 Apr 28;9(1):27. [doi: 10.1186/s40900-023-00433-6] [Medline: 37118762]
15. Holcomb J, Ferguson GM, Sun J, Walton GH, Highfield L. Stakeholder engagement in adoption, implementation, and sustainment of an evidence-based intervention to increase mammography adherence among low-income women. J Cancer Educ 2022 Oct;37(5):1486-1495. [doi: 10.1007/s13187-021-01988-2] [Medline: 33754327]

XSL•FO
RenderX

16. Dennison L, Morrison L, Conway G, Yardley L. Opportunities and challenges for smartphone applications in supporting health behavior change: qualitative study. J Med Internet Res 2013 Apr 18;15(4):e86. [doi: 10.2196/jmir.2583] [Medline: 23598614]
17. National Health and Morbidity Survey (NHMS) 2023: non-communicable diseases and healthcare demand. Institute for Public Health. URL: https://iku.gov.my/nhms-2023 [accessed 2025-03-24]
18. Summary of malaysia national cancer registry report 2017-2021. Ministry of Health Malaysia. URL: https://nci.moh.gov.my/images/pdf_folder/SUMMARY-OF-MALAYSIA-NATIONAL-CANCER-REGISTRY-REPORT-2017-2021.pdf [accessed 2025-03-24]
19. Cheong AT, Lee PY, Ghazali SS, et al. A qualitative exploration of the acceptability of an online web application to promote breast and cervical cancer screening in primary care settings in Malaysia. JCO Global Oncology 2022 May;8(Supplement_1):42-42. [doi: 10.1200/GO.22.48000]
20. Hennemann S, Beutel ME, Zwerenz R. Ready for eHealth? Health professionals' acceptance and adoption of eHealth interventions in inpatient routine care. J Health Commun 2017 Mar;22(3):274-284. [doi: 10.1080/10810730.2017.1284286] [Medline: 28248626]
21. Briefing notes for researchers: involving the public in NHS, public health and social care research. INVOLVE. URL: https://www.invo.org.uk/wp-content/uploads/2014/11/9938_INVOLVE_Briefing_Notes_WEB.pdf [accessed 2025-01-28]
22. Stakeholder engagement framework. Australian Government Department of Health and Aged Care. URL: https://www.health.gov.au/sites/default/files/stakeholder-engagement-framework.pdf [accessed 2025-01-28]
23. van Velsen L, Evers M, Bara CD, Op den Akker H, Boerema S, Hermens H. Understanding the acceptance of an eHealth technology in the early stages of development: an end-user walkthrough approach and two case studies. JMIR Form Res 2018 Jun 15;2(1):e10474. [doi: 10.2196/10474] [Medline: 30684434]
24. Kelders SM, van Zyl LE, Ludden GDS. The concept and components of engagement in different domains applied to eHealth: a systematic scoping review. Front Psychol 2020;11:926. [doi: 10.3389/fpsyg.2020.00926] [Medline: 32536888]
25. Akao Y. Quality Function Deployment: Integrating Customer Requirements into Product Design: Productivity Press; 1990. [doi: 10.4324/9781003578833]
26. Vevox Software Reviews. URL: https://www.softwareadvice.com/polling/vevox-profile/ [accessed 2025-01-28]
27. Franceschini F, Rupil A. Rating scales and prioritization in QFD. International Journal of Quality & Reliability Management 1999 Feb 1;16(1):85-97. [doi: 10.1108/02656719910250881]
28. Goodman MS, Ackermann N, Bowen DJ, Panel D, Thompson VS. Reaching consensus on principles of stakeholder engagement in research. Prog Community Health Partnersh 2020;14(1):117-127. [doi: 10.1353/cpr.2020.0014] [Medline: 32280129]
29. Concannon TW, Fuster M, Saunders T, et al. A systematic review of stakeholder engagement in comparative effectiveness and patient-centered outcomes research. J Gen Intern Med 2014 Dec;29(12):1692-1701. [doi: 10.1007/s11606-014-2878-x] [Medline: 24893581]
30. Yusuf A, P Iskandar YH, Ab Hadi IS, Nasution A, Lean Keng S. Breast awareness mobile apps for health education and promotion for breast cancer. Front Public Health 2022;10:951641. [doi: 10.3389/fpubh.2022.951641] [Medline: 36324460]
31. Munoz-Zuluaga CA, Gallo-Pérez JD, Pérez-Bustos A, et al. Mobile applications: breaking barriers to early breast and cervical cancer detection in underserved communities. JCO Oncol Pract 2021 Mar;17(3):e323-e335. [doi: 10.1200/OP.20.00665] [Medline: 33417491]
32. Mango VL, Olasehinde O, Omisore AD, et al. The iBreastExam versus clinical breast examination for breast evaluation in high risk and symptomatic Nigerian women: a prospective study. Lancet Glob Health 2022 Apr;10(4):e555-e563. [doi: 10.1016/S2214-109X(22)00030-4] [Medline: 35303464]
33. Beidas RS, Stewart RE, Adams DR, et al. A multi-level examination of stakeholder perspectives of implementation of evidence-based practices in a large urban publicly-funded mental health system. Adm Policy Ment Health 2016 Nov;43(6):893-908. [doi: 10.1007/s10488-015-0705-2] [Medline: 26658692]
34. Qudah B, Luetsch K. The influence of mobile health applications on patient - healthcare provider relationships: a systematic, narrative review. Patient Educ Couns 2019 Jun;102(6):1080-1089. [doi: 10.1016/j.pec.2019.01.021] [Medline: 30745178]
35. Haddad SM, Souza RT, Cecatti JG. Mobile technology in health (mHealth) and antenatal care-Searching for apps and available solutions: a systematic review. Int J Med Inform 2019 Jul;127:1-8. [doi: 10.1016/j.ijmedinf.2019.04.008] [Medline: 31128820]
36. Cai T, Huang Y, Zhang Y, Lu Z, Huang Q, Yuan C. Mobile health applications for the care of patients with breast cancer: a scoping review. Int J Nurs Sci 2021 Oct 10;8(4):470-476. [doi: 10.1016/j.ijnss.2021.07.003] [Medline: 34631997]
37. Cankaya BA, Yildirim Baş F, Cankaya İ. Using mobile applications in cancer scanning and evaluating the effectiveness. Turkish Journal of Family Medicine and Primary Care 2021;15(1):49-55. [doi: 10.21763/tjfmpc.803805]
38. Houghton LC, Howland RE, McDonald JA. Mobilizing breast cancer prevention research through smartphone apps: a systematic review of the literature. Front Public Health 2019;7:298. [doi: 10.3389/fpubh.2019.00298] [Medline: 31781525]
39. Olaniyi BY, Fernández del Río A, Periáñez Á, Bellhouse L. User engagement in mobile health applications. Presented at: KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining; Aug 14-18, 2022; Washington DC p. 4704-4712. [doi: 10.1145/3534678.3542681]

XSL•FO
RenderX

40. Boissin C, Blom L, Wallis L, Laflamme L. Image-based teleconsultation using smartphones or tablets: qualitative assessment of medical experts. Emerg Med J 2017 Feb;34(2):95-99. [doi: 10.1136/emermed-2015-205258] [Medline: 27707791]
41. Mosa ASM, Yoo I, Sheets L. A systematic review of healthcare applications for smartphones. BMC Med Inform Decis Mak 2012 Jul 10;12(1):67. [doi: 10.1186/1472-6947-12-67] [Medline: 22781312]
42. Wang Z, Zhang L, Ma L, Liu B. Modeling medical services with mobile health applications. J Healthc Eng 2018;2018(1):1385034. [doi: 10.1155/2018/1385034] [Medline: 29849991]
43. Lingg M, Lütschg V. Health system stakeholders' perspective on the role of mobile health and its adoption in the Swiss health system: qualitative study. JMIR mHealth uHealth 2020 May 11;8(5):e17315. [doi: 10.2196/17315] [Medline: 32391802]
44. Shippee ND, Domecq Garces JP, Prutsky Lopez GJ, et al. Patient and service user engagement in research: a systematic review and synthesized framework. Health Expect 2015 Oct;18(5):1151-1166. [doi: 10.1111/hex.12090] [Medline: 23731468]
45. Joosten YA, Israel TL, Williams NA, et al. Community engagement studios: a structured approach to obtaining meaningful input from stakeholders to inform research. Acad Med 2015 Dec;90(12):1646-1650. [doi: 10.1097/ACM.0000000000000794] [Medline: 26107879]

Abbreviations

CS: correlation scores
QFD: Quality Function Deployment
UIV: user importance values
WS: weighted score

Edited by A Mavragani; submitted 20.08.24; peer-reviewed by L Guo; revised version received 31.01.25; accepted 18.02.25; published 28.03.25.

Please cite as:
Nasrudin N, Sazlina SG, Cheong AT, Lee PY, Teo SH, Aneesa AR, Teo CH, Rokhani FZ, Haron NA, Harrun NH, Ho BK, Mohamed Isa S
Increasing the Uptake of Breast and Cervical Cancer Screening Via the MAwar Application: Stakeholder-Driven Web Application Development Study
JMIR Form Res 2025;9:e65542
URL: https://formative.jmir.org/2025/1/e65542
doi:10.2196/65542

© Nurfarhana Nasrudin, Shariff-Ghazali Sazlina, Ai Theng Cheong, Ping Yein Lee, Soo-Hwang Teo, Abdul Rashid Aneesa, Chin Hai Teo, Fakhrul Zaman Rokhani, Nuzul Azam Haron, Noor Harzana Harrun, Bee Kiau Ho, Salbiah Mohamed Isa. Originally published in JMIR Formative Research (https://formative.jmir.org), 28.3.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Using a Hybrid of AI and Template-Based Method in Automatic Item Generation to Create Multiple-Choice Questions in Medical Education: Hybrid AIG

Yavuz Selim Kıyak[1,2], MD, PhD; Andrzej A Kononowicz[2], PhD

[1]Department of Medical Education and Informatics, Faculty of Medicine, Gazi University, Ankara, Turkey

[2]Department of Bioinformatics and Telemedicine, Jagiellonian University Medical College, Medyczna Str 7, Kraków, Poland

Corresponding Author:
Andrzej A Kononowicz, PhD
Department of Bioinformatics and Telemedicine, Jagiellonian University Medical College, Medyczna Str 7, Kraków, Poland

Abstract

Background: Template-based automatic item generation (AIG) is more efficient than traditional item writing but it still heavily relies on expert effort in model development. While nontemplate-based AIG, leveraging artificial intelligence (AI), offers efficiency, it faces accuracy challenges. Medical education, a field that relies heavily on both formative and summative assessments with multiple choice questions, is in dire need of AI-based support for the efficient automatic generation of items.

Objective: We aimed to propose a hybrid AIG to demonstrate whether it is possible to generate item templates using AI in the field of medical education.

Methods: This is a mixed-methods methodological study with proof-of-concept elements. We propose the hybrid AIG method as a structured series of interactions between a human subject matter expert and AI, designed as a collaborative authoring effort. The method leverages AI to generate item models (templates) and cognitive models to combine the advantages of the two AIG approaches. To demonstrate how to create item models using hybrid AIG, we used 2 medical multiple-choice questions: one on respiratory infections in adults and another on acute allergic reactions in the pediatric population.

Results: The hybrid AIG method we propose consists of 7 steps. The first 5 steps are performed by an expert in a customized AI environment. These involve providing a parent item, identifying elements for manipulation, selecting options and assigning values to elements, and generating the cognitive model. After a final expert review (Step 6), the content in the template can be used for item generation through a traditional (non-AI) software (Step 7). We showed that AI is capable of generating item templates for AIG under the control of a human expert in only 10 minutes. Leveraging AI in template development made it less challenging.

Conclusions: The hybrid AIG method transcends the traditional template-based approach by marrying the “art” that comes from AI as a “black box” with the “science” of algorithmic generation under the oversight of expert as a “marriage registrar”. It does not only capitalize on the strengths of both approaches but also mitigates their weaknesses, offering a human-AI collaboration to increase efficiency in medical education.

(JMIR Form Res 2025;9:e65726) doi:10.2196/65726

KEYWORDS

automatic item generation; ChatGPT; artificial intelligence; large language models; medical education; AI; hybrid; template-based method; hybrid AIG; mixed-method; multiple-choice question; multiple-choice; human-AI collaboration; human-AI; medical education; algorithm; expert

Introduction

Multiple-choice items are integral to written assessment in medical education, serving as a useful tool for assessing a wide range of knowledge and skills. Their common use spans from evaluating factual knowledge to clinical reasoning and decision-making in various domains [1]. This assessment format supports high-volume testing with the added advantage of automated scoring to enhance the efficiency of examinations in medical education.

The traditional way of writing multiple-choice items, characterized by manual development processes, presents significant challenges in scalability. This challenge stems from the intensive time and effort required to create and review each question. This laborious process, which demands expertise and resources, faces a bottleneck in scaling up to meet the demand for a vast quantity of quality items. The challenge is particularly pronounced in medical education, where only a progress test administration in a year requires having 2400 multiple-choice

XSL•FO
RenderX

items [2], showing the inefficiency of traditional methods in satisfying the needs of question banks in medical schools.

Template-based automatic item generation (AIG) is a scalable method used in educational assessment that leverages predefined templates to systematically produce multiple-choice items with the help of software [3]. It has been implemented in 7 languages [3-6] and different health professions [3,7-9]. It consists of 3 sequential stages: development of a cognitive model, development of an item model (template), and using software for the rapid generation of hundreds of items [10]. Item models developed based on cognitive models are structured by subject matter experts to include variables and rules for item generation, allowing for a more efficient creation of consistent questions. This approach enhances the scalability of item development and review compared with traditional item writing [3], addressing the demand for high-quality assessment materials across various educational domains. Template-based AIG, while generating high-quality multiple-choice items efficiently [11-15], still heavily relies on expert effort in development of cognitive models and item models. Although it depends on the content area and the expert's skills, a study reported that development of a cognitive model and an item model can take hours for a subject matter expert [10]. Furthermore, this development process necessitates high levels of extraneous cognitive load due to high element interactivity [16], which requires experts to deal with several components simultaneously. Therefore, "creating cognitive models for AIG is challenging" [3] and it "requires a lot of practice" for experts [3].

On the other hand, nontemplate-based AIG, which "can be guided by the syntactic, semantic, or sequential structure of a text" [3], is an approach that leverages natural language processing to generate assessment items without relying on predefined templates. Unlike the template-based method, this method uses the ability of artificial intelligence (AI) to generate content dynamically, for example, using ChatGPT, which is an AI-based chatbot developed by OpenAI, for creating items based on specific topics or learning outcomes provided by users [17-21]. This approach allows for the generation of diverse and complex questions in seconds, offering flexibility and efficiency in item development. However, this AI-driven approach struggles with issues of inaccuracy and inconsistency [18], especially when good prompting strategies [22] are not used [23]. In AI-driven item generation, such as with ChatGPT, these issues often emerge due to the model's reliance on its training data, which may not always align perfectly with the specific objectives intended by educators. For example, an AI might generate content that includes incorrect information, such as asserting that "the human heart only has two chambers" [24], or misinterpret the complexity level required for a medical education context. Furthermore, the "black box" [25] nature of these AI models complicates diagnosing and correcting these errors within the AI mechanism, as it is challenging to trace back how the AI arrived at a particular output. Therefore, it raises concerns regarding validity, reliability, transparency, fairness, and equity [26], and the process still requires subject matter experts to review and revise each generated question [18,19,21,22,27]. Although it is more efficient than traditional item writing, necessity for reviewing each question is still inefficient.

As outlined above, recent advancements in AIG have offered efficiency, yet each method—template-based and nontemplate-based—brings its own set of limitations. The gap, therefore, lies in the need for a method that merges the structured efficiency of template-based AIG with the content generation capabilities of AI-driven, nontemplate-based, approaches. This convergence could potentially address the pressing need for tools augmenting capabilities of medical educators in test development. We are aware of the interdependence of social (human) and technical elements within an organization [28], advocating for the design of systems that concurrently optimize both human and technological components to achieve effective outcomes. In AIG, this can be interpreted as the need to harmonize the collaboration of a subject matter expert and AI tools working together on developing item and cognitive models.

In this paper, we propose a hybrid AIG method that uses AI to generate an item model (template) and a cognitive model for applying the item template in a template-based item generation process. This capitalizes on the strengths of both approaches but also mitigates their respective weaknesses, offering a novel human-AI collaboration to increase AIG efficiency in medical education.

Methods

Study Process

This is a proof-of-concept study. Drawing on existing guidance on prompt engineering [29,30] and our experience in building human-AI dialogues [20,31], we developed a series of flipped interactions through a series of iterative refinements. In this approach, the AI initiates the conversation and prompts the human expert to provide the necessary data, enabling a collaborative process for generating and refining item templates. The AI shoulders the significant cognitive load involved in template development, thereby reducing the cognitive burden on experts, allowing them to focus on deciding which elements of the question are essential for assessing students' knowledge.

We used GPT Builder (OpenAI), a platform for customizing ChatGPT to the user's needs [32], to train our Custom GPT. Since we conducted this study in February 2024, the Custom GPT worked based on the frontier model at that time, GPT-4.

We will present our approach in a manner analogous to how Gierl et al [10] described the template-based AIG. To illustrate the process, we used two items: one from Gierl et al's work [3], and another from a multiple-choice question (MCQ) similar to an item in the Turkish National Medical Specialty Exam, TUS (2021/1, clinical question number 58). The reason for choosing Gierl et al's [3] item is that it allows readers to compare it with the existing item model developed by a subject matter expert. The reason for choosing a TUS item is that Gierl et al's [3] item model has likely already been processed by ChatGPT, so we also aimed to focus on an item that had not been modelled before.

XSL•FO
RenderX

Ethical Considerations

This study did not involve human subjects, patient data, or personal identifiers, and therefore did not require ethical approval. No informed consent was necessary as no human participants were involved. The study is based on the demonstration of a methodological approach using AI for automatic item generation.

Results

Hybrid AIG and Prompts

The Hybrid AIG consists of 7 steps, with the last 2 steps carried out outside the AI environment. The AI environment requires a specialized GPT trained for generating item models and cognitive models. The Custom GPT we developed is titled "Item Model Maker for AIG" and is accessible at [33]..

In addition, the prompt we used is provided in Multimedia Appendix 1.

Step 1: Providing a Parent Item

The starting point for developing both the item model and the cognitive model in the hybrid AIG is a parent item, reflecting the approach used in the template-based method. It functions as a prototype for generating new questions that follow the template's structure.

In the first step of the hybrid AIG, the expert provides a parent item to AI. Although it is not a requirement, it would be better if the parent item has been chosen from well-performing items in an exam (eg, a published item from a past medical licensure examination). This could mitigate the possible problems that can stem from the lack of quality in the parent item. If an expert does not have a parent item, they can use AI tools, such as Case-Based MCQ Generator [20], to generate an MCQ as a parent item.

The purpose of generating MCQs in medical education is for a human expert to assess the students' skills and knowledge and to evaluate whether they have grasped the nuances of the information presented, aiming for outcomes that are consistent with expert reasoning. The parent item includes the necessary components and embedded possible cognitive models required for this purpose, as it is approved and provided by the expert. It serves as the well-established basis and acts as a seed for generating a new template. In the subsequent steps, AI will primarily handle the cognitive work in using one of the possible underlying cognitive models in the parent item to generate a template. The expert's role will be to monitor and ensure the AI maintains a consistent cognitive model in its output.

We present below the parent items provided to AI, adopting them from the sources described in the Methods section.

Gierl et al's [3] item:

"A 22-year-old female sees her doctor and reports that she's been experiencing a mild cough and slight body aches that have developed over a few days.

Upon examination, she presents with an oral temperature of 37°C. What is the most likely diagnosis?

A. Hay fever
B. Ear infection
C. Common cold
D. Acute sinusitis
E. Seasonal influenza"

TUS Item:

"A fifteen-year-old girl, who became ill during her physical education class after lunch, is brought to the infirmary complaining of coughing, shortness of breath, dizziness, facial swelling and redness, as well as itching in her hands and feet. During the physical examination, the patient appears anxious, with a blood pressure reading of 80/50 mmHg, periorbital edema, and wheezing upon auscultation. Which of the following is the most likely diagnosis for this patient?

A. Exercise-induced asthma
B. Cholinergic urticaria
C. Pulmonary embolism
D. Anaphylaxis
E. Vasovagal syncope"

Step 2: Identification of Elements for Manipulation

In the next step, the expert identifies words or phrases in the parent item for variable manipulation.

In the identification process, the expert can, preferably, consider the importance of the content in the case that is more influential in deciding what the correct answer is. Table 1 shows AI's response and what the expert chose as the elements for manipulation.

XSL•FO
RenderX

Table . Identification of elements for manipulation.

Who speaks	Gierl et al's [3] item	TUS item
AI[a]	• Thank you for sharing the parent item. Could you please identify words or phrases in the question that you would like to manipulate for variable creation? These can include demographic details, symptoms, examination findings, or any other elements that could vary across different versions of this question.	• Based on the parent item you've provided, let's proceed to the next step. Could you please identify words or phrases in the scenario and question that you would like to have as variable elements for manipulation? These could be specific symptoms, the patient's demographics, the setting, or any other details you think should be variable.
Expert	• Element 1: mild cough • Element 2: slight body aches • Element 3: over a few days • Element 4: 37°C	• Element 1: during her physical education class after lunch • Element 2: coughing, shortness of breath, dizziness, facial swelling and redness • Element 3: a blood pressure reading of 80/50 mm Hg, periorbital edema, and wheezing upon auscultation

[a]AI: artificial intelligence.

Step 3: Replacement of Elements and Choosing Options

In this step, AI replaces the words and phrases received from the expert with the elements in brackets, then the expert reviews the output, and prompts AI to make necessary changes, if any. Subsequently, the expert chooses 2 or more options from the options in the parent item. The AI will use them to define the possible values of elements. Table 2 shows the replaced version and the expert's choice among the options.

Table . Replacement of elements and choosing options.

Who speaks	Gierl et al's [3] item	TUS item
AI[a]	Based on the elements you've identified, here's how the parent item could be modified with these variables in brackets: "A 22-year-old female sees her doctor and reports that she's been experiencing a [Element 1] and [Element 2] that have developed [Element 3]. Upon examination, she presents with an oral temperature of [Element 4]. What is the most likely diagnosis? A. Hay fever B. Ear infection C. Common cold D. Acute sinusitis E. Seasonal influenza" Could you please choose two or more options from the parent item that the variables will be focused on? This will help in offering variables for each element that significantly impact the correct answer according to medical sources and guidelines.	Great, I'll incorporate these elements as variables into the parent item. Here's how it will look: "A fifteen-year-old girl, who became ill [Element 1], is brought to the infirmary complaining of [Element 2], as well as itching in her hands and feet. During the physical examination, the patient appears anxious, with [Element3]. Which of the following is the most likely diagnosis for this patient? A. Exercise-induced asthma B. Cholinergic urticaria C. Pulmonary embolism D. Anaphylaxis E. Vasovagal syncope" Now, could you please choose two or more options from the answer choices that you'd like the variables to significantly impact? This will help in designing the variables for each element to ensure they align closely with the medical scenarios corresponding to the chosen options.
Expert	• Common cold • Seasonal influenza	• Cholinergic urticaria • Anaphylaxis

[a]AI: artificial intelligence.

Step 4: Offering Values in Elements

At this stage, AI suggest possible values for each element that can impact the correct answer. Then, the expert reviews the output and makes necessary changes, if any, by providing prompts. Table 3 shows the response of AI in order for offering values. It is important to note that "values" were presented as "variables" by AI in the example because it did not work well when we call them "values" (eg, it was prone to offer numbers rather than words). It may not be a problem in newer versions.

XSL•FO
RenderX

Table . Artificial intelligence's (AI's) offer for values.

Who speaks	Gierl et al's [3] item	TUS item
AI	Thank you for choosing "Common cold" and "Seasonal influenza" as the focus. Based on these options, I'll offer variables for each element that significantly impact the correct answer, considering medical sources and guidelines: Element 1: Cough • Variable 1: Persistent dry cough • Variable 2: Mild cough • Variable 3: Severe hacking cough Element 2: Body aches • Variable 1: Slight body aches • Variable 2: Moderate body aches • Variable 3: Severe muscle pains Element 3: Duration of symptoms • Variable 1: Suddenly over a few hours • Variable 2: Over a few days • Variable 3: Gradually over more than a week Element 4: Oral temperature • Variable 1: 37°C (normal) • Variable 2: 37.5°C to 38°C (mild fever) • Variable 3: Over 38°C (fever) Do these variables seem appropriate for creating different versions of the MCQ[a] item focused on distinguishing between the common cold and seasonal influenza? If you need any adjustments or have additional suggestions, please let me know!	Thank you for choosing options B (Cholinergic urticaria) and D (Anaphylaxis). Based on these selections, I'll offer variables for each element that significantly impact the correct answer considering the chosen options. Here are the proposed variables: Element 1: Timing of symptoms • During her physical education class after lunch. • Immediately after taking a hot shower. • While sitting in a hot classroom. • After exposure to cold wind post-exercise. • During a stressful exam. Element 2: Symptoms • Coughing, shortness of breath, dizziness, facial swelling and redness. • Generalized hives, sweating, and palpitations • Localized itching and redness following sweating. • Throat tightness, difficulty swallowing, and a sense of impending doom. • Flushing, headache, and nausea after eating specific foods. Element 3: Clinical findings • A blood pressure reading of 80/50 mm Hg, periorbital edema, and wheezing upon auscultation. • Tachycardia with a blood pressure reading of 100/70 mmHg, and diffuse hives. • Normal blood pressure, with localized swelling and erythema. • Rapid drop in blood pressure upon exposure to an allergen, with angioedema. • Elevated blood pressure due to anxiety, with mild facial edema. Please review these variables and let me know if they are appropriate or if you would like any modifications.

[a]MCQ: multiple-choice question.

Step 5: Generating the Cognitive Model

In the following step, AI determines the constraints based on values (called "variables" by AI) approved by the expert, in order to create the cognitive model. Then, the expert reviews the output and prompts AI for necessary changes, if any. The presence of the expert is to validate the output, as in the previous steps.

Compared with the cognitive model in the traditional template-based method, our hybrid method adopts a simpler approach. By providing constraints, it allows to determine what items will be generated and what the correct option is in each item. Tables 4 and 5 present the cognitive models. Although the content needs small adjustments that can be easily carried out by the expert (eg, the expert may ask ChatGPT to remove "37.5°C to 38 °C" to make the correct answers more evident), we did not make any changes to present ChatGPT's original output in order for demonstrating that it is able to provide a useful cognitive model to begin with for further revisions. It is also possible due to the nature of GPT models that the output might defer in formatting as visible in the tables.

XSL•FO
RenderX

Table . Cognitive model generated by artificial intelligence (AI) for Gierl et al.'s [3] item.

Element (variable)	Constraints (supported options)
Cough	• Persistent dry cough (seasonal Influenza) • Mild cough (common cold) • Severe hacking cough (seasonal influenza)
Body aches	• Slight body aches (common cold) • Moderate body aches (seasonal influenza) • Severe muscle pains (seasonal influenza)
Duration of symptoms	• Suddenly over a few hours (seasonal influenza) • Over a few days (common cold) • Gradually over more than a week (common cold)
Oral temperature	• 37°C (normal) (common cold) • 37.5°C to 38°C (mild fever) (seasonal influenza) • Over 38°C (fever) (seasonal influenza)

Table . Cognitive model generated by artificial intelligence (AI) for TUS item.

Variable element	Constraints (supported options)
Timing of symptoms	B and D
During her physical education class after lunch.	D
Immediately after taking a hot shower.	B
While sitting in a hot classroom.	B
After exposure to cold wind postexercise.	D
During a stressful exam.	B
Symptoms	B and D
Coughing, shortness of breath, dizziness, and facial swelling and redness.	D
Generalized hives, sweating, and palpitations.	B and D
Localized itching and redness following sweating.	B
Throat tightness, difficulty swallowing, and a sense of impending doom.	D
Flushing, headache, and nausea after eating specific foods.	D
Clinical findings	B and D
A blood pressure reading of 80/50 mmHg, periorbital edema, and wheezing upon auscultation.	D
Tachycardia with a blood pressure reading of 100/70 mmHg, and diffuse hives.	B and D
Normal blood pressure, with localized swelling and erythema.	B
Rapid drop in blood pressure upon exposure to an allergen, with angioedema.	D
Elevated blood pressure due to anxiety, with mild facial edema.	B

These 5 steps have been completed in less than 10 minutes for each model. The whole process within the AI environment can be displayed by accessing the following public pages of the human-AI conversations: Gierl et al's [3] item [34] and TUS item [35].

Step 6: Final Review by the Expert(s)

During this phase, the expert carries out a final review of the item model and cognitive model provided by AI, preferably with other experts.

In the previous steps, in order to keep the demonstration simple and due to the fact that there was not a significant inaccuracy in the AI-generated content, the expert did not demand any additional changes during the process in AI environment. But in the hybrid method, the 5 steps within the AI environment should be actively monitored by the expert, and if necessary, the expert should input prompts to make changes because AI is always prone to provide inaccurate content and deviate from providing a consistent template. Expert oversight, and involvement if necessary, is a strong necessity for creating high-quality item models and cognitive models.

XSL•FO
RenderX

Following the first 5 steps, which can be completed in less than 10 minutes, the expert should carry out one more round of review for the item model and the cognitive model generated through human-AI collaboration. It would be better if the expert conducts this review together with other experts to make sure that there is no inaccuracy, inconsistency, or inappropriate way of presentation. The expert should keep in mind that content generated by AI, in this case ChatGPT, is generated through a large language model, so it could "hallucinate" [36] some inaccurate information and relationships that are harmful for the output quality. Apart from that, in this step, the expert may prefer to add more elements and variables, such as age and gender, in a way that does not change the correct answers, in order for increasing the number of the items.

Step 7: Item Generation Using a Non-AI Software

Finally, the expert inputs the final version of the item template and the constraints to a traditional template-based AIG tool (software without AI involvement), and then the software algorithmically produces multiple-choice items based on the input provided by the expert. It is crucial to emphasize that the expert must input the content accurately, as traditional software cannot handle inconsistent type of inputs, unlike AI in the previous steps. There is no difference between the traditional template-based method (stage 3) [10] and our hybrid method (step 7) in this regard. As is in the template-based method [10], Hybrid AIG also allows the software to generate hundreds of consistent items based on a single item model.

Discussion

In this study, we used AI to generate item models and cognitive models for generating multiple-choice items by using these models for template-based AIG. We demonstrated that AI is capable of providing AIG templates for this purpose under the control of human expert. Leveraging AI in template development has significantly reduced the time investment from hours [10] to less than 10 minutes, and provided a smoother experience for experts in this challenging task [3].

In our hybrid AIG method, cognitive work required to be carried out by experts in the past [10] is now shared with AI. It switches the role of experts from "the creators of item-cognitive models from scratch" to "the reviewer of AI-generated content," which brings an important efficiency to AIG without sacrificing consistency and accuracy. Our hybrid AIG method transcends the traditional template-based approach by marrying the "art" that comes from AI as a "black box" [25] with the "science" of algorithmic generation [10] under the oversight of expert as a "marriage registrar." Practically, this balanced fusion under human guidance reduces the extraneous cognitive load [16] on experts by allocating the burdensome tasks to AI in order for enhancing human efficiency and allowing them to concentrate on refining and validating the AI-generated content.

Similar to our approach, a recent study successfully incorporated a large language model into the process of developing reading comprehension items [37]. While addressing a critical issue in item development for a non–health care setting, its direct application to medical education is challenging due to the inherent complexities of health professions education. Furthermore, this approach integrates AI only into generating unique sentences based on rules imposed by experts, leaving the essential cognitive work dependent on expert input, which remains inefficient for medical education. In our hybrid method, we use AI not only for generating unique sentences but also for development of item models and cognitive models as a whole, hence transforming the role of experts from the main "cognitive workers" to reviewers. This shift reduces cognitive effort for experts while maintaining their essential contribution for accurate and consistent items. Considering the importance of clarity and constraints in the templates, we still can suggest that it is possible and desirable to create specifications and instructions using artificial intelligence. Our research demonstrated that even a minimal human oversight can be sufficient for using AI in the creation of specifications and instructions, particularly in challenging domains such as medical education, which suggests even greater possibility for less complex tasks like reading comprehension. By dismissing the potential of AI in this regard by labeling it as "impossible," humans might inadvertently limit AI's capacity to enhance efficiency in cognitive work needed to be done. Thus, we propose leveraging AI more effectively rather than relegating it to a lesser role.

While noting the improved efficiency of the proposed hybrid for cognitive tasks, we emphasize the importance of rigorous human oversight, and consequently, accountability for automatically generated content. As demonstrated in a study by Zack et al [38], even a state of the art large language model can still be prone to perpetuate racial and gender bias. Adding to the complexity, such bias may not be visible at the level of a single question item or template, but as an effect of prolonged use of generative AI tools. The role of human examiners is to be aware of such risks and to implement bias mitigation strategies at different steps and levels of the AI process to prevent the injection of harmful stereotypes into the assessment of students' skills.

Our study has some limitations. Although the templates generated by AI showed promising results, replicability depends on the consistency of the AI model, which is GPT-4 in this case. In addition, other AI models such as GPT-4o (Open AI), Claude (Anthropic), Gemini (Google), Llama (Meta), and Command *R*+ (Cohere) could lead to different outputs. While our study demonstrated that a hybrid AIG is possible, future research should explore this further by using different parent items across various settings to generate MCQs. As this is a proof-of-concept study, there is a lack of empirical evidence supporting the efficacy of the proposed hybrid AIG method, no qualitative reviews to assess the generated items' quality, and a lack of quantitative item analysis since the items were not tested on medical students. However, it is still valuable because it has shown for the first time that generating plausible, and possibly useful, item templates using AI is possible in medical education. A recent study has provided empirical evidence, demonstrating that experts correctly identified the answers in MCQs generated by using hybrid AIG [39]. In the future studies, we are planning to generate more items using these templates and investigate their effectiveness using qualitative and quantitative methods.

XSL•FO
RenderX

Moreover, a direct comparison of traditional template-based AIG and hybrid AIG could provide valuable evidence for effectiveness and efficiency. Another limitation is that we generated simple templates. There are multilayered templates for AIG [3], which require relatively complex structures, that might require from us to use different custom AIs for this purpose.

In conclusion, the hybrid AIG is a promising novel method that leverages AI in development of templates for template-based AIG that transforms the traditional role of experts from creators to reviewers. This shift can significantly reduce the cognitive burden on experts and streamline the item generation process while ensuring high-quality outcomes. We recommend piloting and improving the hybrid AIG in high-demand settings of increasing importance to investigate and improve its efficiency and quality benefits.

Acknowledgments

This work has been supported by TÜBİTAK (The Scientific and Technological Research Council of Turkiye) under the 2219 program for YSK and internal research funds at Jagiellonian University Medical College under the N41/DBS/000332 program for AK. Generative AI, as a research tool, was used in the preparation of this manuscript for multiple-choice question template generation, which has been transparently reported in the methods and results section of the article. In addition, generative AI was used for language correction. The authors have reviewed, edited, and verified all content. The authors take full responsibility for the integrity and validity of the final work.

Data Availability

All data generated or analyzed during this study are included in this published article.

Authors' Contributions

YSK contibuted to conceptualization, methodology, writing the original draft, data curation, writing-review and editing, and funding acquisition. AK contributed to conceptualization, methodology, writing-review and editing, and funding acquisition.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The prompt that has been used in the custom GPT.
[DOCX File, 16 KB - formative_v9i1e65726_app1.docx]

References

1. Pugh D, De Champlain A, Touchie C. Plus ça change, plus c'est pareil: Making a continued case for the use of MCQs in medical education. Med Teach 2019 May 4;41(5):569-577. [doi: 10.1080/0142159X.2018.1505035]
2. Wrigley W, van der Vleuten CPM, Freeman A, Muijtjens A. A systemic framework for the progress test: strengths, constraints and issues: AMEE Guide No. 71. Med Teach 2012;34(9):683-697. [doi: 10.3109/0142159X.2012.704437] [Medline: 22905655]
3. Gierl MJ, Lai H, Tanygin V. Advanced Methods in Automatic Item Generation, 1st edition: Routledge; 2021. [doi: 10.4324/9781003025634-1]
4. Kiyak YS, Budakoğlu I, Coşkun Ö, Koyun E. The first automatic item generation in Turkish for assessment of clinical reasoning in medical education. Tıp Eğitimi Dünyası 2023 Mar;22(66):72-90. [doi: 10.25282/ted.1225814]
5. Kiyak YS, Coşkun Ö, Budakoğlu I, Uluoğlu C. Psychometric analysis of the first Turkish multiple-choice questions generated using automatic item generation method in medical education. Tıp Eğitimi Dünyası 2023;22(68):154-161. [doi: 10.25282/ted.1376840]
6. Kıyak YS, Kononowicz AA, Górski S. Multilingual template-based automatic item generation for medical education supported by generative artificial intelligence models ChatGPT and Claude. Bio-Algorithms Med-Syst 2024 Dec 23;20(1):81-89. [doi: 10.5604/01.3001.0054.9192]
7. Leslie T, Gierl MJ. Using automatic item generation to create multiple-choice questions for pharmacy assessment. Am J Pharm Educ 2023 Oct;87(10):100081. [doi: 10.1016/j.ajpe.2023.100081] [Medline: 37852684]
8. Lai H, Gierl MJ, Byrne BE, Spielman AI, Waldschmidt DM. Three modeling applications to promote automatic item generation for examinations in dentistry. J Dent Educ 2016 Mar;80(3):339-347. [doi: 10.1002/j.0022-0337.2016.80.3.tb06090.x] [Medline: 26933110]
9. Falcão F, Costa P, Pêgo JM. Feasibility assurance: a review of automatic item generation in medical assessment. Adv in Health Sci Educ 2022 May;27(2):405-425. [doi: 10.1007/s10459-022-10092-z]

XSL•FO
RenderX

10. Gierl MJ, Lai H, Turner SR. Using automatic item generation to create multiple-choice test items. Med Educ 2012 Aug;46(8):757-765. [doi: 10.1111/j.1365-2923.2012.04289.x] [Medline: 22803753]
11. Kosh AE, Simpson MA, Bickel L, Kellogg M, Sanford - Moore E. A cost–benefit analysis of automatic item generation. Educational Measurement 2019 Mar;38(1):48-53 [FREE Full text] [doi: 10.1111/emip.12237]
12. Pugh D, De Champlain A, Gierl M, Lai H, Touchie C. Using cognitive models to develop quality multiple-choice questions. Med Teach 2016 Aug;38(8):838-843. [doi: 10.3109/0142159X.2016.1150989] [Medline: 26998566]
13. Gierl MJ, Lai H, Pugh D, Touchie C, Boulais AP, De Champlain A. Evaluating the psychometric characteristics of generated multiple-choice test items. Applied Measurement in Education 2016 Jul 2;29(3):196-210. [doi: 10.1080/08957347.2016.1171768]
14. Pugh D, De Champlain A, Gierl M, Lai H, Touchie C. Can automated item generation be used to develop high quality MCQs that assess application of knowledge? RPTEL 2020 Dec;15(1):12. [doi: 10.1186/s41039-020-00134-8]
15. Gierl MJ, Lai H. Evaluating the quality of medical multiple-choice items created with automated processes. Med Educ 2013 Jul;47(7):726-733. [doi: 10.1111/medu.12202] [Medline: 23746162]
16. van Merriënboer JJG, Sweller J. Cognitive load theory in health professional education: design principles and strategies. Med Educ 2010 Jan;44(1):85-93. [doi: 10.1111/j.1365-2923.2009.03498.x] [Medline: 20078759]
17. Kıyak YS. A ChatGPT prompt for writing case-based multiple-choice questions. Rev Esp Edu Med 2023;4(3):98-103. [doi: 10.6018/edumed.587451]
18. Zuckerman M, Flood R, Tan RJB, et al. ChatGPT for assessment writing. Med Teach 2023 Nov;45(11):1224-1227. [doi: 10.1080/0142159X.2023.2249239] [Medline: 37789636]
19. Laupichler MC, Rother JF, Grunwald Kadow IC, Ahmadi S, Raupach T. Large language models in medical education: comparing ChatGPT- to human-generated exam questions. Acad Med 2024 May 1;99(5):508-512. [doi: 10.1097/ACM.0000000000005626] [Medline: 38166323]
20. Kıyak YS, Kononowicz AA. Case-based MCQ generator: A custom ChatGPT based on published prompts in the literature for automatic item generation. Med Teach 2024 Aug;46(8):1018-1020. [doi: 10.1080/0142159X.2024.2314723] [Medline: 38340312]
21. Kıyak YS, Coşkun Ö, Budakoğlu I, Uluoğlu C. ChatGPT for generating multiple-choice questions: Evidence on the use of artificial intelligence in automatic item generation for a rational pharmacotherapy exam. Eur J Clin Pharmacol 2024 May;80(5):729-735. [doi: 10.1007/s00228-024-03649-x] [Medline: 38353690]
22. Indran IR, Paranthaman P, Gupta N, Mustafa N. Twelve tips to leverage AI for efficient and effective medical question generation: A guide for educators using Chat GPT. Med Teach 2024 Aug;46(8):1021-1026. [doi: 10.1080/0142159X.2023.2294703] [Medline: 38146711]
23. Ngo A, Gupta S, Perrine O, Reddy R, Ershadi S, Remick D. ChatGPT 3.5 fails to write appropriate multiple choice practice exam questions. Acad Pathol 2024;11(1):100099. [doi: 10.1016/j.acpath.2023.100099] [Medline: 38162414]
24. Lee H. The rise of ChatGPT: exploring its potential in medical education. Anatomical Sciences Ed 2024;17(5):926-931. [doi: 10.1002/ase.2270]
25. von Eschenbach WJ. Transparency and the black box problem: why we do not trust AI. Philos Technol 2021 Dec;34(4):1607-1622. [doi: 10.1007/s13347-021-00477-0]
26. Bulut O, Beiting-Parrish M, Casabianca JM. The rise of artificial intelligence in educational measurement: opportunities and ethical challenges. arXiv. Preprint posted online on Jun 27, 2024. [doi: 10.48550/arXiv.2406.18900]
27. Han Z, Battaglia F, Udaiyar A, Fooks A, Terlecky SR. An explorative assessment of ChatGPT as an aid in medical education: Use it with caution. Med Teach 2024 May;46(5):657-664. [doi: 10.1080/0142159X.2023.2271159] [Medline: 37862566]
28. Appelbaum SH. Socio - technical systems theory: an intervention strategy for organizational development. Manag Decis 1997 Aug;35(6):452-463. [doi: 10.1108/00251749710173823]
29. Mesko B. The ChatGPT (generative artificial intelligence) revolution has made artificial intelligence approachable for medical professionals. J Med Internet Res 2023 Jun 22;25:e48392. [doi: 10.2196/48392] [Medline: 37347508]
30. White J, Fu Q, Hays S, et al. A prompt pattern catalog to enhance prompt engineering with chatgpt. arXiv. Preprint posted online on Feb 21, 2023. [doi: 10.48550/arXiv.2302.11382]
31. Kıyak YS. Beginner-level tips for medical educators: guidance on selection, prompt engineering, and the use of artificial intelligence chatbots. Med Sci Educ 2024 Dec;34(6):1571-1576. [doi: 10.1007/s40670-024-02146-1] [Medline: 39758489]
32. Masters K, Benjamin J, Agrawal A, MacNeill H, Pillow MT, Mehta N. Twelve tips on creating and using custom GPTs to enhance health professions education. Med Teach 2024 Jun;46(6):752-756. [doi: 10.1080/0142159X.2024.2305365] [Medline: 38285894]
33. Item Model Maker for AIG. ChatGPT. URL: https://chatgpt.com/g/g-ISoiQOLyv-item-model-maker-for-aig [accessed 2025-04-01]
34. Item model 1. ChatGPT. URL: https://chatgpt.com/share/1c975655-6bc4-4a16-a082-efad2030a5e2 [accessed 2025-04-01]
35. Item model 2. ChatGPT. URL: https://chatgpt.com/share/7e093fef-8d43-4b8f-bd15-a00bdf4c25c5 [accessed 2025-04-01]
36. Masters K. Medical Teacher 's first ChatGPT's referencing hallucinations: lessons for editors, reviewers, and teachers. Med Teach 2023 Jul;45(7):673-675. [doi: 10.1080/0142159X.2023.2208731] [Medline: 37183932]

XSL•FO
RenderX

37. Sayin A, Gierl M. Using OpenAI GPT to generate reading comprehension items. Educational Measurement 2024 Mar;43(1):5-18 [FREE Full text] [doi: 10.1111/emip.12590]
38. Zack T, Lehman E, Suzgun M, et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. Lancet Digit Health 2024 Jan;6(1):e12-e22. [doi: 10.1016/S2589-7500(23)00225-X] [Medline: 38123252]
39. Kıyak YS, Emekli E, Coşkun Ö, Budakoğlu I. Keeping humans in the loop efficiently by generating question templates instead of questions using AI: Validity evidence on Hybrid AIG. Med Teach 2025 Apr;47(4):744-747. [doi: 10.1080/0142159X.2024.2430360] [Medline: 39604119]

Abbreviations

AI: artificial intelligence
AIG: automatic item generation
MCQ: multiple-choice question

Edited by A Mavragani; submitted 23.08.24; peer-reviewed by A Montazerolghaem, M Alshiekh; revised version received 16.03.25; accepted 18.03.25; published 04.04.25.

Please cite as:
Kıyak YS, Kononowicz AA
Using a Hybrid of AI and Template-Based Method in Automatic Item Generation to Create Multiple-Choice Questions in Medical Education: Hybrid AIG
JMIR Form Res 2025;9:e65726
URL: https://formative.jmir.org/2025/1/e65726
doi:10.2196/65726

© Yavuz Selim Kıyak, Andrzej A Kononowicz. Originally published in JMIR Formative Research (https://formative.jmir.org), 4.4.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Using Large Language Models to Automate Data Extraction From Surgical Pathology Reports: Retrospective Cohort Study

Denise Lee[1*], MD; Akhil Vaid[2*], MD; Kartikeya M Menon[1], BA; Robert Freeman[2], DNP, RN; David S Matteson[3], PhD; Michael L Marin[1], MD; Girish N Nadkarni[2], MD, MPH

[1]Department of Surgery, Icahn School of Medicine at Mount Sinai, 10 Union Square East, Suite 2L, New York, NY, United States

[2]Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, United States

[3]Department of Statistics and Data Science, Cornell University, Ithaca, NY, United States

[*]these authors contributed equally

Corresponding Author:
Denise Lee, MD
Department of Surgery, Icahn School of Medicine at Mount Sinai, 10 Union Square East, Suite 2L, New York, NY, United States

Abstract

Background: Popularized by ChatGPT, large language models (LLMs) are poised to transform the scalability of clinical natural language processing (NLP) downstream tasks such as medical question answering (MQA) and automated data extraction from clinical narrative reports. However, the use of LLMs in the health care setting is limited by cost, computing power, and patient privacy concerns. Specifically, as interest in LLM-based clinical applications grows, regulatory safeguards must be established to avoid exposure of patient data through the public domain. The use of open-source LLMs deployed behind institutional firewalls may ensure the protection of private patient data. In this study, we evaluated the extraction performance of a locally deployed LLM for automated MQA from surgical pathology reports.

Objective: We compared the performance of human reviewers and a locally deployed LLM tasked with extracting key histologic and staging information from surgical pathology reports.

Methods: A total of 84 thyroid cancer surgical pathology reports were assessed by two independent reviewers and the open-source FastChat-T5 3B-parameter LLM using institutional computing resources. Longer text reports were split into 1200-character-long segments, followed by conversion to embeddings. Three segments with the highest similarity scores were integrated to create the final context for the LLM. The context was then made part of the question it was directed to answer. Twelve medical questions for staging and thyroid cancer recurrence risk data extraction were formulated and answered for each report. The time to respond and concordance of answers were evaluated. The concordance rate for each pairwise comparison (human-LLM and human-human) was calculated as the total number of concordant answers divided by the total number of answers for each of the 12 questions. The average concordance rate and associated error of all questions were tabulated for each pairwise comparison and evaluated with two-sided *t* tests.

Results: Out of a total of 1008 questions answered, reviewers 1 and 2 had an average (SD) concordance rate of responses of 99% (1%; 999/1008 responses). The LLM was concordant with reviewers 1 and 2 at an overall average (SD) rate of 89% (7%; 896/1008 responses) and 89% (7.2%; 903/1008 responses). The overall time to review and answer questions for all reports was 170.7, 115, and 19.56 minutes for Reviewers 1, 2, and the LLM, respectively.

Conclusions: The locally deployed LLM can be used for MQA with considerable time-saving and acceptable accuracy in responses. Prompt engineering and fine-tuning may further augment automated data extraction from clinical narratives for the provision of real-time, essential clinical insights.

(JMIR Form Res 2025;9:e64544) doi:10.2196/64544

KEYWORDS

natural language processing; large language model; artificial intelligence; thyroid cancer; endocrine surgery; framework; privacy; medical; surgical pathology; report; NLP; medical question

Introduction

Surgical pathology reports contain narrative data essential for hospital-based comprehensive cancer surveillance databases and the real-time understanding of patient staging, recurrence risk, clinical trial eligibility, and personalized treatment options. Large-scale national databases such as the Surveillance, Epidemiology, and End Results (SEER) program and the National Cancer Database (NCDB) include population-level histologic data necessary to study rare cancers and

XSL•FO
RenderX

underrepresented patient subpopulations [1,2]. However, the large-scale extraction of key surgical oncologic insights contained within unstructured free pathology text is limited by the need for labor-intensive, manual review by human reviewers [3]. The maintenance of comprehensive cancer databases often requires skilled registrars who are trained to synthesize and extract information from patient narrative reports according to established data collection standards [4]. However, constantly updated guidelines, time-intensive labor for data extraction, and inevitable human error are challenges to coordinate cancer data collection.

Advances in natural language processing (NLP) techniques have sought to address challenges in the efficiency and accuracy of data abstraction. Efforts have included applying NLP methods to extract pain scores in patients with cancer undergoing radiation [5], classify metastatic phenotypes from radiology reports of patients with colorectal cancer [6], and identify recurrence status in patients with hepatocellular carcinoma [7]. Traditional NLP methods, however, require specific domain expertise and may be error-prone based on statistical or rule-based approaches [8]. Large language models (LLMs) power a new generation of natural language processing (NLP) whereby deep neural networks are trained on a massive corpus of human text that are then deconstructed into vectorized embeddings that depict linguistic relationships in a numerical format appropriate for easy analysis [9]. Popularized by ChatGPT and its user-friendly question-and-answer interface, LLMs are poised to transform the scalability of clinical NLP downstream tasks such as medical question answering (MQA) and data mining. LLMs may also enhance the ability to rapidly and accurately extract key information from surgical pathology reports [10].

However, as new LLM-based clinical applications are being increasingly explored, there is a substantial risk of protected health information (PHI) breaches without regulatory safeguards in place [11]. Ethical, privacy, and regulatory constraints preclude the transfer of PHI across the public domain through widely used proprietary LLM services (eg, ChatGPT, Gemini, and Claude) that can generate automated responses for MQA. However, significant health care resources (ie, utilization fees and institutional agreements) would be required to ensure the protection of PHI while accessing the largest state-of-the-art models such as GPT4 via a private cloud service provider (eg, Microsoft Azure OpenAI). Such considerable upfront costs and service agreements may preclude hospital systems without financial and IT resources from utilizing the largest and most state-of-the-art LLMs.

The rapid development in LLM technology has led to the development of many open-source models (eg, Llama 3.1 and Google Gemini) both comparable in performance to GPT4 and deployable on an institution's existing computing clusters. The use of a smaller, locally deployed LLM may help ensure that protected health information is maintained and democratize access to LLM technologies at medical centers where larger, private LLMs are inaccessible. In this study, we investigated the use of a locally deployed, open-source LLM to extract key staging and recurrence risk information from thyroid surgical pathology reports. We compared its performance to the gold standard of medically trained human reviewers. Although we utilize this for a single-use case, our work serves as a practical pathway to enable individual medical centers to utilize current LLM technology for clinical NLP tasks in a privacy-protecting manner (Figure 1).

XSL•FO
RenderX

Figure 1. Overview of medical data extraction workflows: (A) Pathology chart review: The traditional approach of manual data extraction from publicly available databases or private electronic health records (EHRs) to obtain predetermined oncologic insights. (B) Enterprise large language model (LLM, such as ChatGPT): Due to regulatory constraints, only publicly available data may be shared with enterprise LLMs. Prompt entry and question curation are used to gain oncologic insights. (C) Private LLMs: EHR data can be shared with a local hospital LLM, and prompt entry with question curation can be used to gain oncologic insights. EHR: electronic health record.

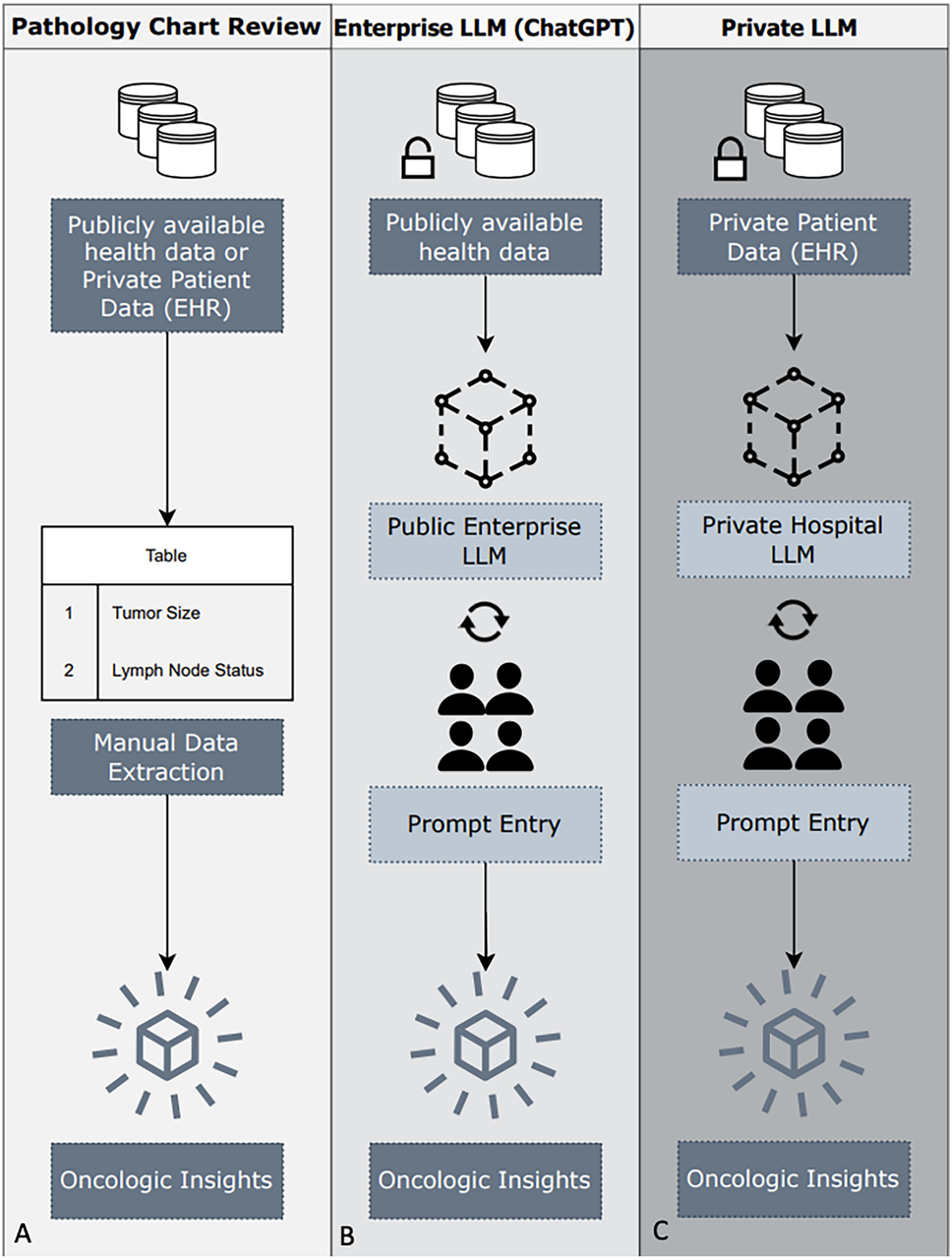

XSL•FO
RenderX

Methods

Ethical Considerations

This study was approved by the Institutional Review Board of the Icahn School of Medicine at Mount Sinai (no. 22 - 00347). Informed consent was waived given the retrospective nature of the study. To ensure confidentiality of patient data, only authorized research study personnel were permitted access to patient data. Study identification numbers were assigned to patients included in the study for deidentification.

Study Population

We queried the Mount Sinai Data Warehouse for a cohort of adult patients (>18 years old) with *ICD-9* (International Classification of Diseases, 9th revision) and *ICD-10* (International Classification of Diseases, 10th revision) diagnosis codes for thyroid cancer and who underwent thyroid surgery between 2010 and 2022. We reviewed 102 surgical pathology reports from 102 patients and excluded reports if they were from other organ sites (n=10), benign (n=2), cytopathology (n=5), or outside review (n=1). We included 84 reports for analysis. The study flowchart is shown in Figure 2.

Figure 2. Flowchart of the study design and analysis. *The concordance rate was calculated as the total number of concordant answers/total number of answers for each of the 12 medical question answering (MQA).

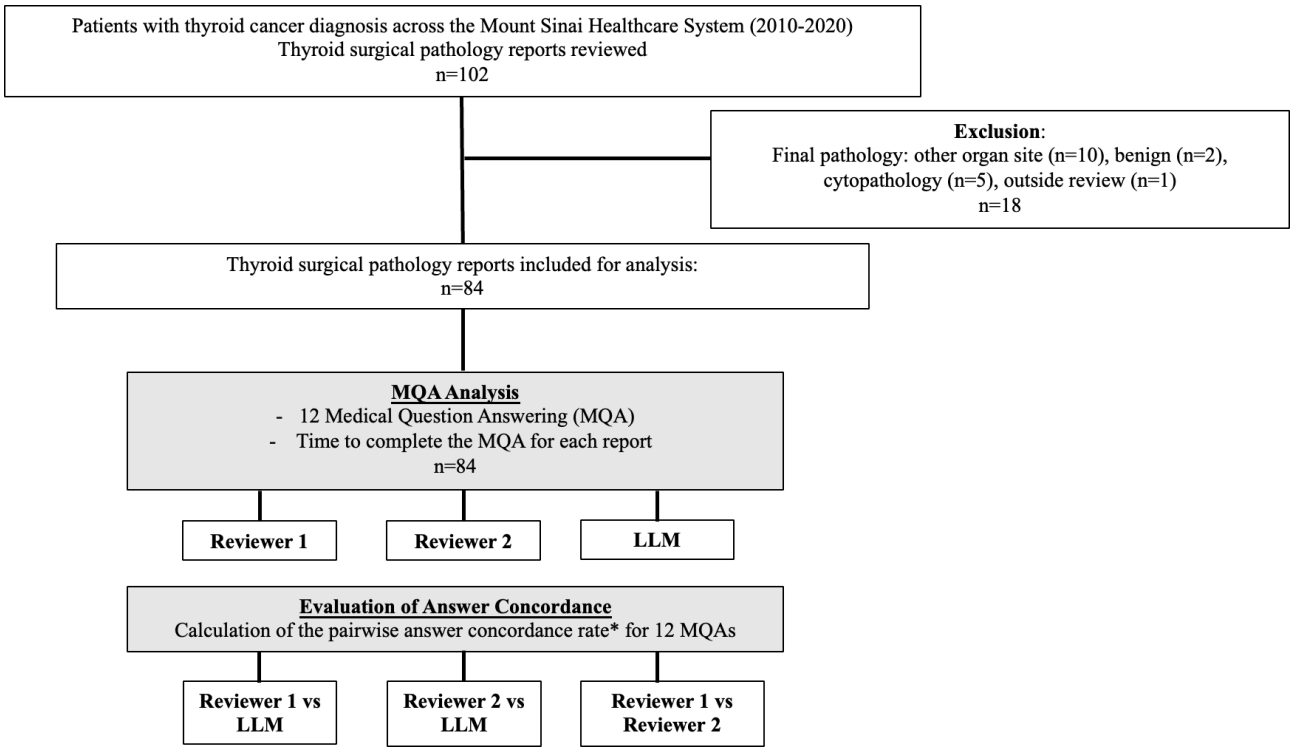

Large Language Model

We used the publicly available, open-source, and parameter-efficient FastChat-T5 3B-parameter LLM for our analysis. FastChat-T5 is a smaller model that requires significantly less computing, making it accessible in resource-constrained environments. This characteristic was essential for our study, as we wanted our findings to be easily replicable across institutions with varying levels of computational resources. In many real-world scenarios, particularly in health care or academic settings, access to high-end hardware or expensive cloud services is limited, and the need for locally deployable solutions behind firewalls is critical due to privacy concerns. FastChat-T5 provides a balance between performance and resource efficiency, making it a practical choice in these contexts. This LLM was deployed on existing hospital servers, which avoided the utilization of expensive cloud services and external transfer of patient data. The ready availability of such software has increased the use of open-source LLMs. The practice of deploying smaller models on local hardware aligns with the latest trends from leading LLM providers that are exploring the deployment of models on mobile devices.

In addition, our goal was to evaluate how a smaller, open-source model performs in comparison to human reviewers, who are generally considered the gold standard in tasks such as document review and content evaluation. We selected FastChat-T5 to explore whether such a parameter-efficient model could offer comparable performance to humans, without requiring highly specialized code or expensive infrastructure for running more recent and resource-intensive models, such as those in the LLaMA (large language Meta AI) family.

A limitation all LLMs have is the amount of context they may process at once, and for FastChat-T5, this limit is 2048 tokens. For reports of lengths greater than what the model could accommodate, we split the report text into 1200-character-long segments, followed by converting each of these segments into machine-readable numerical representations called embeddings. As embeddings encode meaning, calculating similarity scores using the cosine similarity metric between segment embeddings and posed questions allowed us to retrieve the pieces of text

XSL•FO
RenderX

most directly related to the content of the question. As such, three segments with the highest similarity scores were integrated to create the final context for the LLM. This context was made part of a plain language question for the LLM, alongside a question it was directed to answer.

We selected the top three segments with the highest similarity scores for several key reasons. First, using multiple segments allows us to enrich the context provided to the LLM, ensuring that it has access to a broader and more diverse set of relevant information. By incorporating three segments, we increase the likelihood that important details from different parts of the document are included, without overwhelming the model's context window.

Additionally, focusing on the top three ensures that we strike a balance between the relevance and the diversity of information. Using just one segment might omit crucial context, while using too many could dilute the focus on the most pertinent details. The number three was chosen empirically, as it allowed for a good compromise between feeding the LLM with enough useful content and avoiding the risk of both exceeding usable context and introducing unrelated or redundant information that could confound the model's operation. We did experiment with other segmentation strategies, including combining two or four closest segments, but the performance gains were inferior to those of the three-segment approach. In our preliminary trials, while combining more distinct segments could introduce variety, this approach led to a slight reduction in overall accuracy due to the introduction of less relevant or tangential information.

Finally, embedding models with cosine similarity offers an efficient and robust way to evaluate how closely each segment relates to the posed question, ensuring that the selected segments are not just textually similar but semantically aligned with the user's query. This method enhances the precision of the context provided to the LLM, resulting in more accurate and relevant answers. The source code can be found at [12] under a GPLv3 license.

The prompt template is available in Multimedia Appendix 1.

Development of MQA and the Evaluation of Concordance

We formulated 12 questions with expert clinical input that extracted key information for the assessment of the thyroid cancer staging and recurrence risk on the AJCC-TNM staging system (American Joint Committee on Cancer Tumor-Node-Metastasis, 8th edition) according to the American Thyroid Association Recurrence Risk Stratification System [13,14]. Two study authors (DL and KMM) reviewed 84 thyroid surgical pathology reports and recorded answers to each of the 12 questions and the time taken to complete the answers for each report. Then, we used the LLM to answer the same questions. For every question, we determined if the answers were concordant between Reviewer 1 and the LLM, between Reviewer 2 and the LLM, and between the two reviewers. The concordance rate for each pairwise comparison (human-LLM and human-human) was then calculated as the total number of concordant answers divided by the total number of answers for each of the 12 questions (Figure 2). The average concordance rate and associated error (calculated here as the SD) for all questions were tabulated for each pairwise comparison and evaluated with two-sided *t* tests.

Results

We report the sample LLM responses and the concordance rates between reviewers and the LLM for each question in Table 1. A total of 1008 questions were answered for the 84 thyroid surgical pathology reports. Reviewers 1 and 2 were concordant at an average (SD) overall rate of 99% (1%; 999 out of 1008 questions) with disagreement on 9 answers. Reviewers 1 and 2 took an average (SD) of 2 (0.6) minutes and 1.4 (0.4) minutes to respond to each pathology report and a total of 170.7 minutes and 115 minutes for all reports, respectively. The LLM was concordant with reviewers 1 and 2 at an average (SD) overall rate of 89% (7%; 896 out of 1008 questions) and 89 (7.2%; 903 out of 1008 questions), respectively. The LLM was significantly less accurate than both reviewer 1 ($P<.001$) and reviewer 2 ($P<.001$), but performed at the same level between both reviewers ($P=.82$). The average time to review each report for the LLM was 13.97 seconds/report and 19.56 minutes for all reports. The questions with the highest overall rates of concordant responses were questions requiring binary or categorical data extraction (eg, "Is lymphatic invasion present?", 100%; "Is vascular invasion present?", 98.81%; and "Where is the primary cancer located?", 98.1%). The question with the lowest overall concordance was, "Were cervical lymph nodes present?" at 75%.

XSL•FO
RenderX

Table . Twelve medical questions and sample large language model (LLM) answers with the rates of concordance between the reviewers and the LLM.

Medical question	Sample LLM answer	Concordance (%)			
		R1,[a] LLM	R2,[b] LLM	R1, R2	All 3
Staging tumor (T)					
What is the size of the primary tumor?	The size of the primary tumor was 5.5 cm.	82	84	98	82
Does the tumor extend beyond the capsule of the thyroid?	No, the tumor does not extend beyond the capsule of the thyroid.	90	93	97	90
Where is the primary cancer located (right thyroid lobe, left thyroid lobe, or isthmus)?	The primary cancer is located in the right thyroid lobe.	99	100	99	99
Node (N)					
Were any cervical lymph nodes present?	Yes, cervical lymph nodes were present.	76	100	99	75
If cervical lymph nodes are present, how many lymph nodes were positive for malignancy?	Three lymph nodes were positive for malignancy.	88	88	100	88
Stage (TNM)[c]					
What is the final pathology TNM stage of the thyroid cancer?	The final pathology TNM stage of the thyroid cancer is pT3a.	85	85	100	85
Recurrence risk factor					
Is vascular invasion present?	Yes, vascular invasion is present.	99	99	100	99
Is lymphatic invasion present?	No, lymphatic invasion is not present.	100	100	100	100
What was the histology of the primary cancer?	The histology of the primary thyroid cancer was papillary carcinoma.	87	87	100	87
Was there a variant of thyroid cancer such as tall cell, hobnail variant, or columnar cell variant present, indicative of a more aggressive histology?	No, a variant of thyroid cancer such as tall cell, hobnail variant, or columnar cell variant was not present indicative of a more aggressive histology.	89	92	97	89
Was there a second thyroid cancer present?	No, there was no second thyroid cancer present.	82	83	99	82
If there was a second thyroid cancer present, what was the histology?	The histology of the second thyroid cancer was classical variant papillary thyroid carcinoma.	88	88	99	88
Overall questions, % (SD)		89 (7)	89 (7)	99 (1)	89 (7.2)

[a]R1: Reviewer 1.

[b]R2: Reviewer 2.

[c]TNM: Tumor-Node-Metastasis.

XSL•FO
RenderX

Discussion

We demonstrated and evaluated the extraction performance of a locally deployed, open-source LLM for a specific clinical NLP task. To our knowledge, this is the first study to compare an LLM's ability to extract key thyroid cancer recurrence and staging information accurately and efficiently from surgical pathology reports using a conversational MQA format. The LLM took 19.56 minutes to evaluate and respond to all pathology reports, whereas it required an additional 187 minutes and 105 minutes for the reviewers to complete the same task—demonstrating a considerable reduction in time. Regarding the accuracy of the responses, we found that the rates of response concordance were higher among questions tasked with simpler binary or categorical responses. The increase in task complexity requiring textual interpretation and inconsistent word prompting, such as asking whether there was "cervical" lymph nodes present, resulted in the lowest rate of concordance. Furthermore, the question regarding the size of the primary tumor also seemed to be relatively straightforward but had an overall concordance rate of only 82%.

The augmentation of poorer-performing MQA may lie in the improvement of prompt engineering—an emerging subfield where domain-specific knowledge and linguistics are optimized to design questions that yield the best-performing response to a task. In addition, more expressive embeddings may help localize relevant text [15]. For example, "cervical" does not appear in most pathological reports verbatim, possibly limiting the model's ability to respond appropriately to the question regarding the presence of cervical lymph nodes. Moreover, the LLM often incorrectly identified the size of the "primary tumor" and would instead provide a dimension from another specimen in the report, such as the overall thyroid lobe. This response accuracy may also be improved by modifying the question prompt and will be the focus of future work.

Overall, the use of LLMs may be an advancement from early NLP methods that faced limitations such as restrictive data preprocessing and the inability to handle multiple positive diagnoses [16-18]. Studies exploring the ability of LLMs to handle real clinical data have been rapidly growing. Examples include the investigation of the use of LLMs for multiple clinical NLP tasks such as the summarization of magnetic resonance imaging and radiology reports of the knee [19], the simplification of anatomic pathology reports for patient interpretation [20], and data extraction from breast cancer pathology and radiology reports [21]. However, the protection of private patient information must be prioritized with the use of any emerging technology. Safeguards should be established to ensure data breaches can be avoided with the use of LLMs for clinical tasks. The approach of deploying smaller, open-source LLMs behind a health care institution's own computing resources ensures that researchers across all centers may use this emerging technology while maintaining patient privacy. Additionally, the increased language capacity of the latest generation of LLMs allows institutions to deploy their own data for in-context learning only while potentially achieving a reasonable performance.

The limitations of this study include the use of a single institutional dataset without comparison with an external dataset to evaluate the generalizability of the LLM performance. Differences in language, reporting style, and description length of surgical pathology reports of different cancer types may also affect the ability of a smaller LLM such as the open-source FastChat-T5 used in this study to accurately extract the requested data using prompts. Additionally, while the intention of the study was to compare the performance of a smaller, open-source LLM to the gold standard of human reviewers, a comparison of its performance with larger, private LLMs such as GPT should be explored in future studies, after data have been confirmed to be entirely deidentified.

This study demonstrated that a locally deployed, open-source LLM is capable of automating data extraction faster than human reviewers who need to perform a manual review of surgical pathology reports. However, although the responses between the LLM and expert reviewers reached concordance rates of 88 - 89%, the degree of error by the LLM precludes its exclusive use to populate cancer statistics databases. Until LLMs demonstrate comparable performance to humans, the use of LLMs for clinical NLP tasks still requires considerable oversight by human reviewers. As a human-assistance tool, the role of LLMs may lie in improving the time efficiency of populating clinical databases. As methods of prompt engineering and fine-tuning improve, we envision that LLMs will eventually allow medical institutions to harness cutting-edge advances in NLP for timely and privacy-preserving MQA data extraction from pathology reports and other clinical narratives for the provision of real-time, essential oncologic insights.

Acknowledgments

This work was supported in part through the computational and data resources and staff expertise provided by the Scientific Computing and Data Department at the Icahn School of Medicine at Mount Sinai and supported by the Clinical and Translational Science Awards (CTSA) grant UL1TR004419 from the National Center for Advancing Translational Sciences. No generative artificial intelligence was used in the preparation or writing of this manuscript.

Data Availability

Source code for the study may be found at [12] under a GPLv3 license.

XSL•FO
RenderX

Authors' Contributions

Conceptualization: DL, AV, RF, GNN, MLM
Data curation: DL, KMM, AV
Formal analysis: DL, KMM, AV, DSM
Methodology: DL, AV, DSM
Supervision: GNN, MLM
Validation: DL, AV
Visualization: DL, KMM
Writing – DL, KMM, AV
Writing – review & editing: DL, AV, KMM, RF, DSM, MM, GNN

Conflicts of Interest

None declared.

Multimedia Appendix 1
Transcript of prompts.
[DOCX File, 15 KB - formative_v9i1e64544_app1.docx]

References

1. Boffa DJ, Rosen JE, Mallin K, et al. Using the national cancer database for outcomes research: a review. JAMA Oncol 2017 Dec 1;3(12):1722-1728. [doi: 10.1001/jamaoncol.2016.6905] [Medline: 28241198]
2. Hankey BF, Ries LA, Edwards BK. The surveillance, epidemiology, and end results program: a national resource. Cancer Epidemiol Biomarkers Prev 1999 Dec;8(12):1117-1121. [Medline: 10613347]
3. Burger G, Abu-Hanna A, de Keizer N, Cornet R. Natural language processing in pathology: a scoping review. J Clin Pathol 2016 Jul 22:jclinpath-2016-203872. [doi: 10.1136/jclinpath-2016-203872] [Medline: 27451435]
4. Hochheiser H, Finan S, Yuan Z, et al. DeepPhe-CR: natural language processing software services for cancer registrar case abstraction. JCO Clin Cancer Inform 2023 Sep;7:e2300156. [doi: 10.1200/CCI.23.00156] [Medline: 38113411]
5. Naseri H, Kafi K, Skamene S, et al. Development of a generalizable natural language processing pipeline to extract physician-reported pain from clinical reports: Generated using publicly-available datasets and tested on institutional clinical reports for cancer patients with bone metastases. J Biomed Inform 2021 Aug;120:103864. [doi: 10.1016/j.jbi.2021.103864] [Medline: 34265451]
6. Causa Andrieu P, Golia Pernicka JS, Yaeger R, et al. Natural language processing of computed tomography reports to label metastatic phenotypes with prognostic significance in patients with colorectal cancer. JCO Clin Cancer Inform 2022 Sep;6:e2200014. [doi: 10.1200/CCI.22.00014] [Medline: 36103642]
7. Karimi YH, Blayney DW, Kurian AW, et al. Development and use of natural language processing for identification of distant cancer recurrence and sites of distant recurrence using unstructured electronic health record data. JCO Clin Cancer Inform 2021 Apr;5:469-478. [doi: 10.1200/CCI.20.00165] [Medline: 33929889]
8. Yim WW, Yetisgen M, Harris WP, Kwan SW. Natural language processing in oncology: a review. JAMA Oncol 2016 Jun 1;2(6):797-804. [doi: 10.1001/jamaoncol.2016.0213] [Medline: 27124593]
9. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023 Aug;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
10. Yang X, Chen A, PourNejatian N, et al. A large language model for electronic health records. NPJ Digit Med 2022 Dec 26;5(1):194. [doi: 10.1038/s41746-022-00742-2] [Medline: 36572766]
11. Rengers TA, Thiels CA, Salehinejad H. Academic surgery in the era of large language models: a review. JAMA Surg 2024 Apr 1;159(4):445-450. [doi: 10.1001/jamasurg.2023.6496] [Medline: 38353991]
12. LLMSurgicalPathologyNoteParserPublic. Github. URL: https://github.com/akhilvaid/LLMSurgicalPathologyNoteParser [accessed 2025-04-07]
13. Tuttle RM, Haugen B, Perrier ND. Updated American joint committee on cancer/tumor-node-metastasis staging system for differentiated and anaplastic thyroid cancer (eighth edition): what changed and why? Thyroid 2017 Jun;27(6):751-756. [doi: 10.1089/thy.2017.0102] [Medline: 28463585]
14. Haugen BR, Alexander EK, Bible KC, et al. 2015 American thyroid association management guidelines for adult patients with thyroid nodules and differentiated thyroid cancer: the American thyroid association guidelines task force on thyroid nodules and differentiated thyroid cancer. Thyroid 2016 Jan;26(1):1-133. [doi: 10.1089/thy.2015.0020] [Medline: 26462967]
15. Savova GK, Danciu I, Alamudun F, et al. Use of natural language processing to extract clinical cancer phenotypes from electronic medical records. Cancer Res 2019 Nov 1;79(21):5463-5470. [doi: 10.1158/0008-5472.CAN-19-0579] [Medline: 31395609]
16. Datta S, Bernstam EV, Roberts K. A frame semantic overview of NLP-based information extraction for cancer-related EHR notes. J Biomed Inform 2019 Dec;100:103301. [doi: 10.1016/j.jbi.2019.103301] [Medline: 31589927]

XSL•FO
RenderX

17. Schadow G, McDonald CJ. Extracting structured information from free text pathology reports. AMIA Annu Symp Proc 2003;2003:584-588. [Medline: 14728240]
18. Mitchell JR, Szepietowski P, Howard R, et al. A question-and-answer system to extract data from free-text oncological pathology reports (CancerBERT network): development study. J Med Internet Res 2022 Mar 23;24(3):e27210. [doi: 10.2196/27210] [Medline: 35319481]
19. López-Úbeda P, Martín-Noguerol T, Díaz-Angulo C, Luna A. Evaluation of large language models performance against humans for summarizing MRI knee radiology reports: a feasibility study. Int J Med Inform 2024 Jul;187:105443. [doi: 10.1016/j.ijmedinf.2024.105443] [Medline: 38615509]
20. Steimetz E, Minkowitz J, Gabutan EC, et al. Use of artificial intelligence chatbots in interpretation of pathology reports. JAMA Netw Open 2024 May 1;7(5):e2412767. [doi: 10.1001/jamanetworkopen.2024.12767] [Medline: 38776080]
21. Choi HS, Song JY, Shin KH, Chang JH, Jang BS. Developing prompts from large language model for extracting clinical information from pathology and ultrasound reports in breast cancer. Radiat Oncol J 2023 Sep;41(3):209-216. [doi: 10.3857/roj.2023.00633] [Medline: 37793630]

Abbreviations

ICD: International Classification of Diseases
LLaMA: large language Meta AI
LLM: large language model
MQA: medical question answering
NCDB: National Cancer Database
NLP: natural language processing
PHI: protected health information
SEER: Surveillance, Epidemiology, and End Results

Edited by A Mavragani; submitted 19.07.24; peer-reviewed by M Abdalla, ST Arasteh; revised version received 17.01.25; accepted 19.01.25; published 07.04.25.

Please cite as:
Lee D, Vaid A, Menon KM, Freeman R, Matteson DS, Marin ML, Nadkarni GN
Using Large Language Models to Automate Data Extraction From Surgical Pathology Reports: Retrospective Cohort Study
JMIR Form Res 2025;9:e64544
URL: https://formative.jmir.org/2025/1/e64544
doi:10.2196/64544

© Denise Lee, Akhil Vaid, Kartikeya M Menon, Robert Freeman, David S Matteson, Michael L Marin, Girish N Nadkarni. Originally published in JMIR Formative Research (https://formative.jmir.org), 7.4.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Moving Standard Deviation of Trunk Acceleration as a Quantification Index for Physical Activities: Validation Study

Takuya Suzuki[1], MHS; Yuji Kono[1], PhD; Takayuki Ogasawara[2], PhD; Masahiko Mukaino[3,4], MD, PhD; Yasushi Aoshima[1], BSc; Shotaro Furuzawa[1], MHS; Yurie Fujita[1], BSc; Hirotaka Matsuura[3], MD, PhD; Masumi Yamaguchi[2], PhD; Shingo Tsukada[2], MD, PhD; Yohei Otaka[3], MD, PhD

[1]Department of Rehabilitation, Fujita Health University Hospital, Toyoake, Japan

[2]NTT Basic Research Laboratories and Bio-medical Informatics Research Center, NTT Corporation, Atsugi, Japan

[3]Department of Rehabilitation Medicine, Fujita Health University School of Medicine, Toyoake, Japan

[4]Department of Rehabilitation Medicine, Hokkaido University Hospital, Kita14 Nishi5, Kita-Ku, Sapporo, Japan

Corresponding Author:
Masahiko Mukaino, MD, PhD
Department of Rehabilitation Medicine, Fujita Health University School of Medicine, Toyoake, Japan

Abstract

Background: Step count is used to quantify activity in individuals using accelerometers. However, challenges such as difficulty in detecting steps during slow or irregular gait patterns and the inability to apply this method to wheelchair (WC) users limit the broader utility of accelerometers. Alternative device-specific measures of physical activity exist, but their specificity limits cross-applicability between different device sensors. Moving standard deviation of acceleration (MSDA), obtained from truncal acceleration measurements, is proposed as another alternative variable to quantify physical activity in patients.

Objective: This study aimed to evaluate the validity and feasibility of MSDA for quantifying physical activity in patients with stroke-induced hemiparesis by comparing it with the traditional step count.

Methods: We enrolled 197 consecutive patients with stroke hemiparesis admitted to a convalescent rehabilitation ward. Using the hitoe system, a smart clothing–based physical activity measurement system, we measured the MSDA of trunk movement and step count. The correlation between MSDA and step count was examined in all participants. Based on their daily living mobility levels, measured using the Functional Independence Measure (FIM), participants were categorized into 6 subgroups: FIM1-4, FIM5 (WC), FIM5 (walking), FIM6 (WC), FIM6 (walking), and FIM7 (walking). Intersubgroup differences in MSDA were analyzed.

Results: A strong correlation was observed between MSDA and step count (r=0.78; P<.001), with a stronger correlation in the walking group (r=0.79; P<.001) compared with the WC group (r=0.55; P<.001). The Shapiro-Wilk test indicated no significant results for MSDA across all subgroups, supporting a normal distribution within these groups. In contrast, the step count data for the WC subgroups showed significant results, indicating a deviation from a normal distribution. Additionally, 10.2% (20/197) of participants recorded zero steps, demonstrating a floor effect in the step count data. The median MSDA values for the 6 subgroups (FIM1-4, FIM5 WC, FIM5 walking, FIM6 WC, FIM6 walking, and FIM7) were 0.006, 0.007, 0.010, 0.011, 0.011, and 0.014, respectively, reflecting their levels of independence based on the FIM mobility scores. The median step counts for these subgroups were 68, 233, 1386, 367, 2835, and 4462, respectively. FIM5 participants who walked had higher step counts than FIM6 participants using WCs, though the difference was marginally but not statistically significant (P=.07), highlighting the impact of mobility type (walking vs WC).

Conclusions: The results suggest the validity of MSDA as a variable for physical activity in patients with stroke, applicable to patients with stroke irrespective of their mobility measures. This finding highlights the potential of MSDA for use in individuals with motor impairments, including WC users, underscoring its broad utility in rehabilitation clinical practice.

(JMIR Form Res 2025;9:e63064) doi:10.2196/63064

KEYWORDS

smart clothing; step count; moving standard deviation of acceleration; MSDA; wheelchair; activity quantification; physical activities; validation study; accelerometer; regular gait patterns; older people; aging; motor impairments; step detection; stroke; hemiparesis; measurement system; walking; mobility; rehabilitation

XSL•FO
RenderX

Introduction

Poststroke rehabilitation is primarily undertaken to enhance patient independence in activities of daily living (ADL). To achieve this goal, improving movement ability and exercise tolerance is important, which is directly reflected in the daily amount of activity. Increases in physical activity and extended rehabilitation time are associated with improved physical function and shorter hospital stays [1,2]. Higher levels of physical activity can decrease the mortality risk [3]. Therefore, physical activity requirements should be actively assessed in patients in the poststroke stage. To quantify physical activity, the number of steps, measured using an accelerometer, is widely used and is a valid and reliable variable for assessing physical activity in clinical rehabilitation settings [4,5]. Numerous studies have established a close association between the number of steps taken and health-related quality of life [6], self-efficacy [7], and the risk of stroke recurrence [8]. Accelerometers are reliable and validated devices that can objectively measure physical activity in patients with stroke [9]. Exercise therapy, when supplemented with accelerometer feedback and a goal-centered step-activity monitoring program, can enhance the physical activity of hospitalized patients recovering from ischemic stroke [10,11].

While widely used to quantify activity, counting steps is only applicable to ambulatory patients. Most of the previous studies have focused on patients with stroke who could walk independently. However, a substantial number of these patients face challenges with walking in their daily lives. During discharge from acute hospitalization, approximately 30% - 50% of patients had a modified Rankin scale score of ≥4 [12], indicating significant mobility impairment. While step count, an acceleration threshold–based measure of activity, may partially capture movement in everyday activities, its validity in accurately measuring the activity of patients with walking difficulties remains questionable. These considerations highlight the need for acceleration-based measures that can be effectively used in patients with difficulty walking.

To address this issue, we proposed an alternative metric: the moving standard deviation of acceleration (MSDA), which is derived from the SD of the acceleration norm that is measured from the trunk using an accelerometer placed on the chest. The chest placement of the accelerometer enables movement measurement in a seated position, which is expected to have advantages in activity monitoring in wheelchair (WC) users. The utility of the MSDA as an activity-quantification measure has been validated previously, showing a strong correlation between the MSDA and percentage oxygen uptake reserve in healthy participants and between MSDA and percentage heart rate reserve in patients with motor dysfunction due to cerebrovascular disease [13]. However, its validity compared with traditional accelerometry-based measures, such as step count, remains unexplored. Additionally, the potential advantages of MSDA in quantifying physical activity for individuals with limited mobility, including WC users, have yet to be confirmed with real-world data.

Consequently, this study focused on investigating the validity and feasibility of the MSDA measurement with a hitoe system in the measurement of physical activity in patients with stroke, with or without a WC. Specifically, our objectives were (1) to examine the correlation between the MSDA and the number of steps taken and (2) to compare their data characteristics, which includes an understanding of the features of data distribution and investigating their respective relationships with the level of mobility limitation.

Methods

Participants

Between January 2018 and February 2020, we enrolled consecutive patients with stroke hemiparesis who were admitted to the recovery rehabilitation ward of Fujita Health University Hospital. The inclusion criteria were (1) a confirmed diagnosis of stroke resulting in hemiparesis, (2) aged 18 years or older, and (3) the provision of written informed consent. The exclusion criteria were (1) noncompliance with wearing the hitoe measurement wear and (2) the presence of severe comorbid conditions, such as advanced heart failure or respiratory diseases requiring oxygen support that could interfere with participation in the rehabilitation program or data collection.

Study Design and Protocol

This single-center cross-sectional study comprised participants who underwent physical activity assessment within 1 week of admission to the recovery rehabilitation ward.

Ethical Considerations

This study was conducted in accordance with the ethical principles outlined in the Declaration of Helsinki and was approved by the Institutional Review Board of Fujita Health University (approval number HM17-220). All eligible patients who were admitted were informed about the study by the clinicians and provided with detailed information about its objectives and procedures. Those who agreed to participate were enrolled in the study after providing written informed consent. To ensure privacy, all data were deidentified before analysis. Participants did not receive financial compensation for their involvement in this study.

Measurement

A hitoe system (NTT Corporation and Toray Industry, Inc) was used to measure the amount of physical activity [14]; this study specifically evaluated its accelerometry-based physical activity measurement function. The hitoe system is a clothing-based activity-monitoring system that consists of clothing with embedded electrodes for heart rate measurement, a transmitter equipped with an embedded accelerometer, a data receiver (which can be a smartphone or an Internet of Things gateway), and a data server. The wearer's activity is measured using an accelerometer to capture physical movement and a heart rate monitor to assess load. The transmitter was securely fastened to the chest with a tight-fitting strap to minimize measurement noise. The noninferiority of accuracy in accelerometry-based measurements using the hitoe system has been demonstrated previously through comparisons with various

XSL•FO
RenderX

accelerometry-based devices, including waist-worn accelerometers (pedometers) [15]. The sampling rate was set to 25 Hz. Monitoring started before noon on the first assessment day and ended in the afternoon on the third assessment day. As clothing was used for monitoring, monitoring was paused during bathing. After bathing, the hitoe wear was changed, and the monitoring resumed. To mitigate the impact of daily fluctuations and compensate for deficiencies [16], we used 48-hour data to calculate the 24-hour ensemble average. Ensemble averaging is a method in which multiple measurements recorded at the same time of the day are averaged together. For instance, the data recorded between 5 PM and 5:01 PM on the first day was averaged with the data recorded at the same time on the second day, producing the ensemble average for 5 PM-5:01 PM. This process was repeated for all time points throughout the 24-hour period. The data with ≥5% measurement error after ensemble averaging were excluded from the analysis. The acquired 24-hour average values of minute-by-minute data on the number of daily steps and the MSDA were used for the analysis. The step-detection capability of this system has been validated using data from various commercial pedometers [15]. The MSDA was derived from a 2-second window of 50 samples of the norm of truncal acceleration, as measured by the accelerometer. The acceleration norm was computed using the following equation:

$$\text{Norm}=(x2+y2+z2)$$

where x, y, and z represent the vertical, lateral, and anterior/posterior axes, respectively.

Mobility levels of the patients were assessed using the Functional Independence Measure (FIM) scores. The FIM consists of 18 (13 motor and 5 cognitive) items. Each item is rated on a 7-point scale from 1 (full assistance) to 7 (independence). The data on other clinical characteristics including age, sex, height, weight, and disease were collected from the patients' medical records.

Data Analysis

Continuous variables are presented as mean and SD and categorical variables as percentages. The Shapiro-Wilk test was used to examine the normality of the number of steps and MSDA. To further investigate the characteristics of these distributions, their skewness and kurtosis were examined using the Fisher-Pearson coefficient of skewness (g1) and Fisher coefficient of kurtosis (g2) [17]. The relationship between MSDA and number of steps was examined using Spearman rank correlation coefficient. This analysis was conducted for all participants and separately for the two groups: those in the walking group and those in the WC group. Finally, we categorized all participants into 6 subgroups based on the mobility scores from the FIM and their mobility patterns as follows: FIM1-4, FIM5 WC, FIM5 walking, FIM6 WC, FIM6 walking, and FIM7. The Kruskal-Wallis test was performed to analyze the association between MSDA and the number of steps. The Bonferroni method was used to control for multiple comparisons.

All analyses were performed using SPSS (version 26; IBM Corp). Statistical significance was set at $P<.05$.

Results

A total of 197 patients (129 men) were enrolled and included in the analysis (Table 1). The mean age was 66.1 (SD 12.5) years. Of these, 61 patients were ambulatory, and 136 patients used WCs. The median number of days after onset was 33 (IQR 24 - 45). The average MSDA and number of steps were 0.009 (SD 0.003) and 1229 (SD 1834), respectively.

XSL•FO
RenderX

Table . Participant characteristics in the study cohort. Baseline characteristics of 197 participants including participants' demographics, stroke etiology, mobility status, and Functional Independence Measure (FIM) scores (motor, cognitive, and total) are summarized.

Variables	Values
Age (years), mean (SD)	66.1 (12.5)
Sex, n (%)	
Male	129 (65.5)
Female	68 (34.5)
Height (cm), mean (SD)	161.7 (15.5)
Weight (kg), mean (SD)	58.6 (13.7)
Days after onset, median (IQR)	33 (24 - 45)
Etiology, n (%)	
Cerebral hemorrhage	71 (36)
Cerebral infarction	94 (47.7)
Subarachnoid hemorrhage	16 (8.1)
Others	16 (8.1)
Mobility, n (%)	
Walking	61 (31)
Wheelchair	136 (69)
FIM-motor, mean (SD)	48.5 (26.0)
FIM-cognitive, mean (SD)	22.5 (10.2)
FIM-total, mean (SD)	70.7 (34.6)

The relationship between the MSDA and the number of steps is shown in Figure 1. Significant positive correlations were found for all participants (r=0.784; P<.001), the walking group (r=0.787; P<.001), and the WC group (r=0.546; P<.001). When excluding the extreme data with 0 values, the correlation coefficient was raised, for all participants (r=823; P<.001), the walking group (r=0.798; P<.001), and the WC group (r=0.634; P<.001).

XSL•FO
RenderX

Figure 1. Scatter plot showing the correlations between MSDA and number of steps in the (A) wheelchair group (n=136) and (B) walking group (n=61). Open circles and open squares represent the wheelchair and walking groups, respectively. MSDA: moving standard deviation of acceleration.

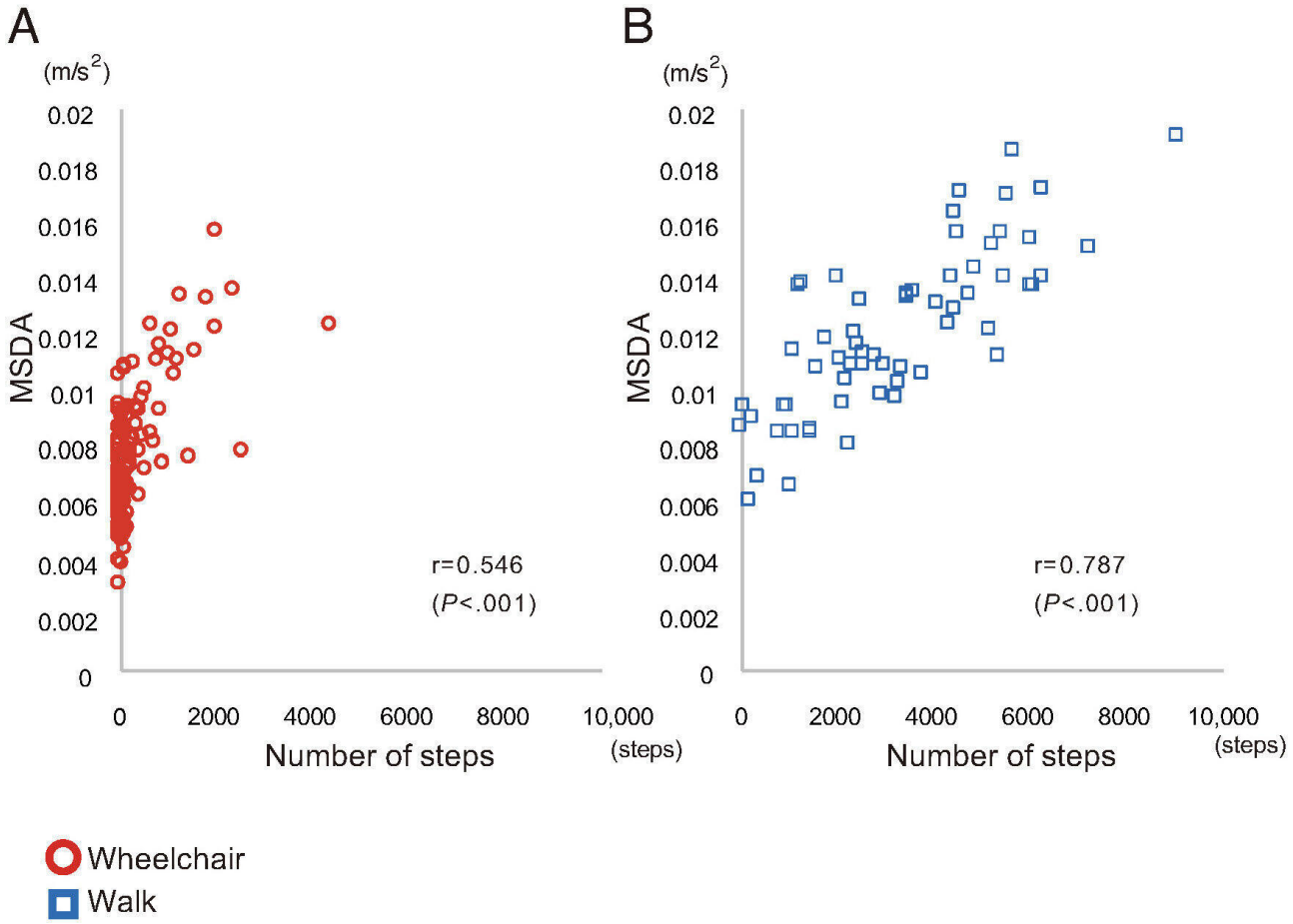

The detailed features of the MSDA distribution and the number of steps for all participants are shown in Figure 2. Neither the MSDA nor the number of steps exhibited a normal distribution. The distributions of MSDA and steps exhibited right-skewness, with skewness coefficients of 0.78 and 1.72, respectively (zero when normally distributed). The kurtosis coefficient for MSDA was 3.02, which is close to the typical value of 3 for a normal distribution. In contrast, the kurtosis coefficient for steps was 5.26, indicating a higher level of kurtosis. The normality of the distribution was further tested in each of the 6 subgroups (Table 2). The distribution of MSDA data within subgroups exhibited smaller deviations from a normal distribution in terms of kurtosis and skewness coefficients, suggesting that it was closer to a normal distribution compared with the steps data. The Shapiro-Wilk test did not yield significant results for MSDA in all the subgroups, whereas the step data showed significance, especially within the WC subgroups. In the step count, 10.2% (20/197) of participants exhibited zero count, showing the floor effect.

XSL•FO
RenderX

Figure 2. Histograms showing the distribution of (A) MSDA and (B) daily step count for all 197 participants, divided by mobility type. The wheelchair group is represented in red and the walking group in blue. MSDA: moving standard deviation of acceleration.

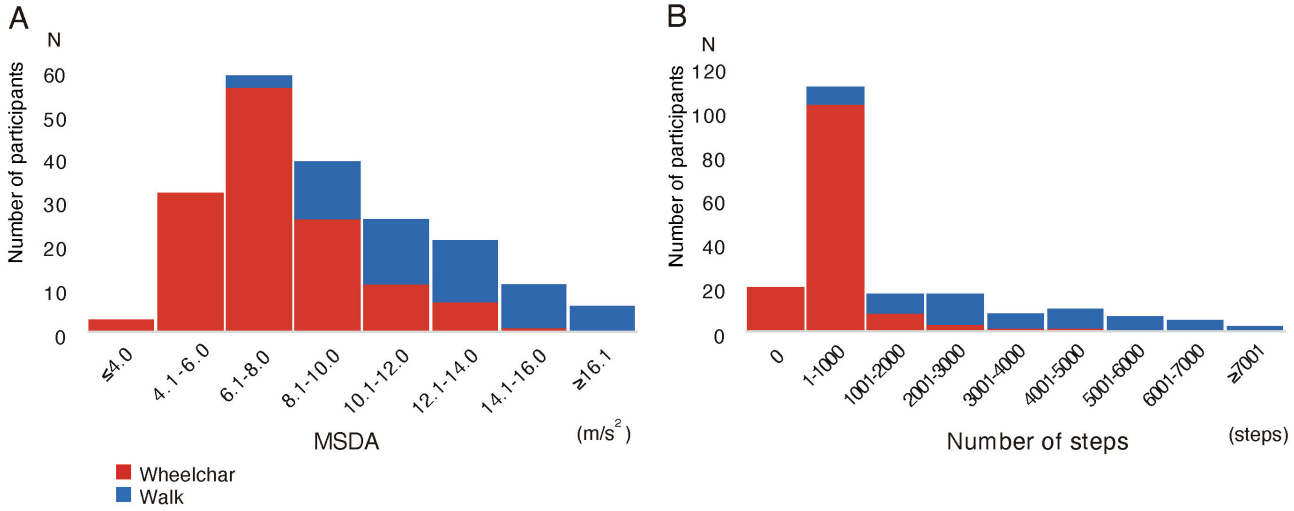

Table . Kurtosis and skewness values and results of the Shapiro-Wilk test. This table summarizes the normality test results for the moving standard deviation of acceleration (MSDA) and daily step count in participants classified by mobility type and functional levels. For normal distribution, kurtosis (g2)=3 and skewness (g1)=0.

	MSDA			Number of steps		
	Kurtosis (g2)	Skewness (g1)	*P* value (Shapiro-Wilk)	Kurtosis (g2)	Skewness (g1)	*P* value (Shapiro-Wilk)
WC[a] 1 - 4 (n=94)	3.33	0.51	.18	24.33	4.32	<.001
WC 5 (n=23)	3.15	0.94	.05	2.22	0.75	.005
WC 6 (n=19)	2.59	0.05	.92	4.37	1.38	.002
Walking 5 (n=15)	2.48	0.27	.79	1.64	0.01	.23
Walking 6 (n=18)	2.75	0.47	.67	1.95	0.09	.83
Walking 7 (n=28)	2.63	–0.38	.62	2.57	0.10	.42
WC (n=136)	3.91	0.93	<.001	19.48	3.61	<.001
Walking (n=61)	2.56	0.19	.72	2.61	0.40	.11
All participants (n=197)	3.02	0.78	<.001	5.26	1.72	<.001

[a]WC: wheelchair.

The relationships between the 6 subgroups according to the FIM scores of mobility items, mobility patterns, MSDA, and number of steps are shown in Figure 3. The median MSDA values in the 6 subgroups FIM1-4, FIM5 WC, FIM5 walking, FIM6 WC, FIM6 walking, and FIM7 were 0.006, 0.007, 0.010, 0.011, 0.011, and 0.014, respectively. Significant differences were found in MSDA between the low-FIM and high-FIM subgroups (Figure 3A). The median number of steps in the 6 subgroups were 68, 233, 1386, 367, 2835, and 4462, respectively. Significant differences in the steps between the low-FIM and high-FIM subgroups are shown in Figure 3B. The difference between FIM5 participants with walking and FIM6 WC was marginal (*P*=.07).

XSL•FO
RenderX

Figure 3. Relationship between Functional Independence Measure (FIM)-mobility scores and the MSDA for 6 groups divided by mobility type: FIM1-4, FIM5 (wheelchair), FIM5 (walking), FIM6 (wheelchair), FIM6 (walking), and FIM7 (walking). Physical activity was compared among the 6 subgroups according to the FIM scores of the mobility items and mobility patterns. (A) MSDA and (B) number of steps. MSDA: moving standard deviation of acceleration. *P<.05, **P<.01.

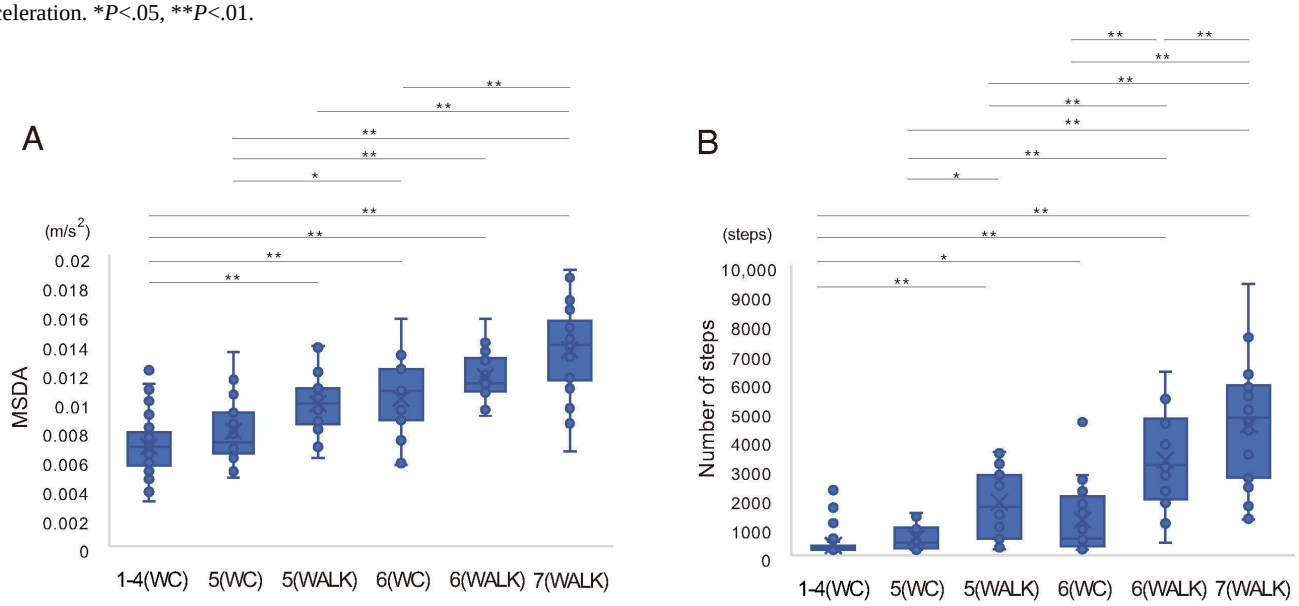

Discussion

Principal Findings

This study investigated the characteristics of the MSDA and step count in measuring physical activity in patients with stroke-induced hemiparesis. The MSDA not only had a positive correlation with the number of steps but was also associated with the FIM-mobility score, showing the validity as activity measurement index. Additionally, the MSDA demonstrated potential advantages in quantifying activity levels in patients with walking difficulties. First, in contrast to the step count, the distribution of the MSDA exhibited a closer alignment with a normal distribution, particularly within subgroups of patients using WCs and those with similar ADL levels. Second, the MSDA did not present a floor effect within the WC group, which was observed with the step count. Finally, the MSDA was consistent with varying levels of mobility independence, showing no significant differences across types of mobility measures (walking or WC use). In contrast, the step count was more susceptible to the mode of mobility; significant differences were observed in the step values between groups with the same level of independence but using different mobility measures.

The findings of this study demonstrated the strength of the MSDA as a measure of physical activity in patients after stroke, who often present with gait disturbance and relies on WC. The conducted analyses on the correlation between the step counts and MSDA and the distribution of those in the poststroke rehabilitation inpatients, of whom approximately 70% relied on a WC for mobility, support the usability of the MSDA in those with limited gait functionality. Our results indicated that the MSDA was positively correlated with the number of steps, which is an established measure of physical activity, supporting the validity of the MSDA. Further analysis revealed a weaker correlation in patients who used WCs. The weaker correlation between step counts and MSDA in WC users can be attributed to the skewed distribution and floor effects in their step counts. Notably, excluding participants with a step count of zero results in an improved correlation coefficient. Further stratified examination revealed that, in contrast to the number of steps, the MSDA was normally distributed within all levels of mobility independence. These findings suggest that the MSDA has broad applicability, including that for patients in poststroke stage who use WCs.

The strength of the MSDA in measuring physical activity among patients in poststroke stage, with fewer issues in distribution patterns, may stem from its nature as a continuous variable. The number of steps, which is a frequently used variable for activity monitoring, is a threshold-based variable that measures the number of movements with greater acceleration beyond a specific threshold [16]. However, this approach may induce challenges when applied to patients with neurological disorders who often manifest irregular step patterns and slower gait speeds. The use of accelerometers to count steps can produce misleading results in such cases [18]. In contrast, the MSDA quantifies truncal oscillations; as the trunk is the body's heaviest segment, this can, without imposing a specific threshold, be captured as a continuous value [13]. Thus, the MSDA could be more sensitive to variations in nonwalking activities. Consequently, the MSDA distribution encompasses both walking and nonwalking activities; this characteristic of the MSDA underscores its potential application as a comprehensive metric for evaluating activity levels irrespective of individual mobility methods, whether walking or using WCs.

Furthermore, our results illustrate that the MSDA effectively reflects the level of independence in daily life. Previous studies have demonstrated a strong correlation between physical activity and independence levels; for instance, the Life-Space Assessment, a physical activity measure, shows a positive correlation with Lawton instrumental ADL scores [19,20]. A similar positive correlation has been noted between accelerometer-measured physical activity and FIM-motor scores in patients with stroke [21].

XSL•FO
RenderX

The study's results further support previous observations and reveal significant differences in both the MSDA and step count between patients with and without independent mobility. Furthermore, in contrast to the step count, the MSDA appears to be less influenced by the mode of mobility. In the case of mobility levels FIM5 and FIM6, no significant differences were observed between patients using WCs and those walking, while significant disparities in step counts were observed across different mobility measures at both levels, leading to reduced coherence with respect to levels of independence—it tends to be higher for patients capable of walking with a level of FIM 5, as opposed to those who are using WC with a FIM score of 6. These findings suggest that the MSDA can serve as a universal measure of physical activity, regardless of an individual's gait capabilities. This objective measure could be beneficial for providing feedback and setting goals for rehabilitation, to encourage daily activities. Many community-dwelling patients with a stroke fail to achieve the recommended activity level [22]. Thus, using the MSDA to track daily activities could offer a valuable strategy for promoting more active daily living in the realm of rehabilitation.

Limitations

This study has some limitations. First, its applicability to nonhospitalized patients warrants further exploration. The study was conducted in a rehabilitation ward with a flat floor, which differs from the environment outside the hospital. With barriers such as steps or uneven surfaces, the MSDA values may be higher than those in hospital wards. The impact of environmental variables on MSDA merits investigation in future studies. Despite these considerations, this study is meaningful in demonstrating the advantage of MSDA in the measurement of physical activity in patients with low walking ability or WC users.

Second, the feasibility of using the MSDA with the hitoe system in a home setting should be further explored. In this study, therapist assistance was required for fitting and ensuring proper placement to maintain measurement quality. While the system has been designed to allow patients with hemiparesis to wear it independently, those with severe hemiparesis may still encounter difficulties. For this system to be effectively integrated into daily activity management for patients with hemiparesis, ease of fitting is essential. Further investigation is needed to adapt and implement this system in practical clinical settings.

Third, as the MSDA reflects truncal movement oscillations without identifying the specific movements involved, certain neurological symptoms may influence the measurements. For instance, patients with Parkinson disease experience tremors that can affect the MSDA values, which reflect movement variability. Further research is needed to clarify the movements or symptoms that may distort MSDA values. Finally, to enhance understanding, the clinical significance of MSDA values should be further investigated, specifically in terms of interpreting different values as corresponding levels of physical activity. While the step count serves as an easily understandable and intuitive metric for clinical feedback, the MSDA value lacks this straightforwardness, posing challenges for patients in correlating it with specific activities. Therefore, fostering a deeper comprehension and interpretation of this variable within clinical contexts is imperative. Efforts should be directed towards contextualizing the MSDA values and making them more intuitive and meaningful for patients.

Conclusion

This study indicates that the MSDA is a valid variable for measurement of physical activity, applicable to patients with stroke irrespective of their mobility measures. Future research is warranted to examine the factors influencing MSDA values and to establish the clinical utility of this variable.

Acknowledgments

The authors thank Mr Kenta Maruyama at NTT Basic Research Laboratories for their technical assistance with the data analysis.

Data Availability

The data collected and analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

The authors used leased wearable clothing and hardware devices from Toray Industries and NTT Corporation, respectively, which are the manufacturers of the tracking devices. TO, MY, and ST are employees of NTT Corporation. The other authors declare no conflicts of interest associated with this manuscript.

References

1. Schneider EJ, Lannin NA, Ada L, Schmidt J. Increasing the amount of usual rehabilitation improves activity after stroke: a systematic review. J Physiother 2016 Oct;62(4):182-187. [doi: 10.1016/j.jphys.2016.08.006] [Medline: 27637769]
2. English C, Shields N, Brusco NK, et al. Additional weekend therapy may reduce length of rehabilitation stay after stroke: a meta-analysis of individual patient data. J Physiother 2016 Jul;62(3):124-129. [doi: 10.1016/j.jphys.2016.05.015] [Medline: 27320831]
3. Inoue M, Iso H, Yamamoto S, et al. Daily total physical activity level and premature death in men and women: results from a large-scale population-based cohort study in Japan (JPHC study). Ann Epidemiol 2008 Jul;18(7):522-530. [doi: 10.1016/j.annepidem.2008.03.008] [Medline: 18504139]

XSL•FO
RenderX

4. Mathie MJ, Coster ACF, Lovell NH, Celler BG. Accelerometry: providing an integrated, practical method for long-term, ambulatory monitoring of human movement. Physiol Meas 2004 Apr;25(2):R1-20. [doi: 10.1088/0967-3334/25/2/r01] [Medline: 15132305]
5. Dowd KP, Szeklicki R, Minetto MA, et al. A systematic literature review of reviews on techniques for physical activity measurement in adults: a DEDIPAC study. Int J Behav Nutr Phys Act 2018 Feb 8;15(1):15. [doi: 10.1186/s12966-017-0636-2] [Medline: 29422051]
6. Kanai M, Izawa KP, Kubo H, Nozoe M, Mase K, Shimada S. Association of health utility score with physical activity outcomes in stroke survivors. Int J Environ Res Public Health 2020 Dec 31;18(1):251. [doi: 10.3390/ijerph18010251] [Medline: 33396274]
7. Kanai M, Nozoe M, Izawa KP, et al. Promoting physical activity in hospitalized patients with mild ischemic stroke: a pilot study. Top Stroke Rehabil 2017 May;24(4):256-261. [doi: 10.1080/10749357.2016.1259030] [Medline: 27881044]
8. Kono Y, Kawajiri H, Kamisaka K, et al. Predictive impact of daily physical activity on new vascular events in patients with mild ischemic stroke. Int J Stroke 2015 Feb;10(2):219-223. [doi: 10.1111/ijs.12392] [Medline: 25381835]
9. Rand D, Eng JJ, Tang PF, Jeng JS, Hung C. How active are people with stroke?: use of accelerometers to assess physical activity. Stroke 2009 Jan;40(1):163-168. [doi: 10.1161/STROKEAHA.108.523621] [Medline: 18948606]
10. Kanai M, Izawa KP, Kobayashi M, et al. Effect of accelerometer-based feedback on physical activity in hospitalized patients with ischemic stroke: a randomized controlled trial. Clin Rehabil 2018 Aug;32(8):1047-1056. [doi: 10.1177/0269215518755841] [Medline: 29400070]
11. Danks KA, Roos MA, McCoy D, Reisman DS. A step activity monitoring program improves real world walking activity post stroke. Disabil Rehabil 2014;36(26):2233-2236. [doi: 10.3109/09638288.2014.903303] [Medline: 24670193]
12. Toyoda K, Yoshimura S, Nakai M, et al. Twenty-year change in severity and outcome of ischemic and hemorrhagic strokes. JAMA Neurol 2022 Jan 1;79(1):61-69. [doi: 10.1001/jamaneurol.2021.4346] [Medline: 34870689]
13. Mukaino M, Ogasawara T, Matsuura H, et al. Validity of trunk acceleration measurement with a chest-worn monitor for assessment of physical activity intensity. BMC Sports Sci Med Rehabil 2022 Jun 10;14(1):104. [doi: 10.1186/s13102-022-00492-4] [Medline: 35689292]
14. Ogasawara T, Matsunaga K, Ito H, Mukaino M. Application for rehabilitation medicine using wearable textile “hitoe”. NTT Tech Rev 2018;16(9):6-12. [doi: 10.53829/ntr201809fa2]
15. Ogasawara T, Itoh Y, Kuwabara K, Kasahara R. Gait analysis using a wearable T-shirt type sensor. NTT Tech Rev 2016;14(4):47-53. [doi: 10.53829/ntr201604ra1]
16. Ogasawara T, Mukaino M, Otaka Y, et al. Validation of data imputation by ensemble averaging to quantify 24-h behavior using heart rate of stroke rehabilitation inpatients. J Med Biol Eng 2021 Jun;41(3):322-330. [doi: 10.1007/s40846-021-00622-2]
17. Pearson ES, Hartley HO. Biometrika Tables for Statisticians, 3rd edition: Cambridge University Press; 1970:207.
18. Elsworth C, Dawes H, Winward C, et al. Pedometer step counts in individuals with neurological conditions. Clin Rehabil 2009 Feb;23(2):171-175. [doi: 10.1177/0269215508098895] [Medline: 19164404]
19. Baker PS, Bodner EV, Allman RM. Measuring life-space mobility in community-dwelling older adults. J Am Geriatr Soc 2003 Nov;51(11):1610-1614. [doi: 10.1046/j.1532-5415.2003.51512.x] [Medline: 14687391]
20. Peel C, Sawyer Baker P, Roth DL, Brown CJ, Brodner EV, Allman RM. Assessing mobility in older adults: the UAB study of aging life-space assessment. Phys Ther 2005 Oct;85(10):1008-1119. [doi: 10.1093/ptj/85.10.1008] [Medline: 16180950]
21. Hobo K, Kurita H, Momose K. The relationship between energy expenditure and physical functions in patients hospitalised for stroke. Sci Rep 2021 Nov 4;11(1):21685. [doi: 10.1038/s41598-021-01135-3] [Medline: 34737355]
22. Michael KM, Allen JK, Macko RF. Reduced ambulatory activity after stroke: the role of balance, gait, and cardiovascular fitness. Arch Phys Med Rehabil 2005 Aug;86(8):1552-1556. [doi: 10.1016/j.apmr.2004.12.026] [Medline: 16084807]

Abbreviations

ADL: activities of daily living
FIM: Functional Independence Measure
MSDA: moving standard deviation of acceleration
WC: wheelchair

XSL•FO
RenderX

Edited by A Mavragani; submitted 21.06.24; peer-reviewed by A Opheim; revised version received 09.12.24; accepted 09.12.24; published 08.04.25.

Please cite as:
Suzuki T, Kono Y, Ogasawara T, Mukaino M, Aoshima Y, Furuzawa S, Fujita Y, Matsuura H, Yamaguchi M, Tsukada S, Otaka Y
Moving Standard Deviation of Trunk Acceleration as a Quantification Index for Physical Activities: Validation Study
JMIR Form Res 2025;9:e63064
URL: https://formative.jmir.org/2025/1/e63064
doi:10.2196/63064

© Takuya Suzuki, Yuji Kono, Takayuki Ogasawara, Masahiko Mukaino, Yasushi Aoshima, Shotaro Furuzawa, Yurie Fujita, Hirotaka Matsuura, Masumi Yamaguchi, Shingo Tsukada, Yohei Otaka. Originally published in JMIR Formative Research (https://formative.jmir.org), 8.4.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Provider Perspectives on Implementing an Enhanced Digital Screening for Adolescent Depression and Suicidality: Qualitative Study

Morgan A Coren[1], MD; Oliver Lindhiem[1], PhD; Abby R Angus[2], BS; Emma K Toevs[2], BA; Ana Radovic[1,3], MD, MSc

[1]School of Medicine, University of Pittsburgh, 100 N Bellefield, Office 537, Pittsburgh, PA, United States

[2]University of Pittsburgh Medical Center, Western Psychiatric Hospital, Pittsburgh, PA, United States

[3]Adolescent and Young Adult Medicine, University of Pittsburgh Medical Center, Children's Hospital of Pittsburgh, Pittsburgh, PA, United States

Corresponding Author:
Oliver Lindhiem, PhD
School of Medicine, University of Pittsburgh, 100 N Bellefield, Office 537, Pittsburgh, PA, United States

Abstract

Background: With a growing adolescent mental health crisis, pediatric societies are increasingly recommending that primary care providers (PCPs) engage in mental health screening. While symptom-level screens identify symptoms, novel technology interventions can assist PCPs with providing additional point-of-care guidance to increase uptake for behavioral health services.

Objective: In this study, we sought community PCP feedback on a web-based, digitally enhanced mental health screening tool for adolescents in primary care previously only evaluated in research studies to inform implementation in community settings.

Methods: A total of 10 adolescent providers were recruited to trial the new screening tool and participate in structured interviews based on the Consolidated Framework for Implementation Research domains. Interviews were audio recorded, transcribed, and coded according to a prespecified codebook using a template analysis approach.

Results: Providers identified improving mental health screening and treatment in pediatric primary care as a priority and agreed that a web-based digitally enhanced screening tool could help facilitate identification of and management of adolescent depression. Salient barriers identified were lack of electronic health record integration, time to administer screening, implications on clinic workflow, accessibility, and lack of transparency within health care organizations about the process of approving new technologies for clinical use. Providers made multiple suggestions to enhance implementation in community settings, such as incorporating customization options.

Conclusions: Technology interventions can help address the need for improved behavioral health support in primary care settings. However, numerous barriers exist, complicating implementation of new technologies in real-world settings.

(JMIR Form Res 2025;9:e67624) doi:10.2196/67624

KEYWORDS

depression; suicidality; adolescent mental health screening; primary care; digital tools

Introduction

Adolescent depression and suicidality are part of a growing public health crisis that has been further exacerbated by the COVID-19 pandemic [1]. In 2022, 19.5% of US adolescents aged 12 - 17 years experienced a major depressive episode and 13.4% reported having seriously considered suicide [2]. Pediatric societies recommend that pediatric primary care providers (PCPs) screen adolescents for both depression and suicidality [3-6]. Although screening adolescents for depression is recommended by the US Preventive Services Taskforce [7], there is no evidence to support that screening by itself improves health outcomes [7]. PCPs recognize the need for additional interventions to increase uptake of services and provide clinical benefit, but requirements for increased mental health staff and support obfuscate their implementation.

Technology interventions can aid providers to implement screening recommendations and augment initiatives to improve behavioral health support in primary care. In recent years, technology-based screeners have been developed for detecting conditions in addition to depression such as posttraumatic stress disorder (PTSD) [8], substance use [9], and interpersonal violence [10] in primary care settings, as well as to screen patients for social needs [11]. Such interventions have proven to be effective, accurate, well-received by patients, and helpful in streamlining clinic workflow and history-taking.

Screening Wizard (SW; see Figure 1 for a screenshot) is a web-based digitally enhanced screening tool for eliciting

XSL•FO
RenderX

symptoms of depression and suicidality in adolescents. In addition to surveying adolescents and their caregivers through separate unique weblinks with validated screening measures for comorbidities of anxiety and mania, the screening tool assesses additional considerations, such as treatment readiness, motivation, treatment preferences, and barriers. The surveys take approximately 5 - 7 minutes to complete. Once completed, a report succinctly synthesizes the information to provide PCPs with comparisons between adolescents and their caregiver responses that may guide management decisions and referrals. The initial qualitative study informing the design of SW [12], evaluation of iterative feedback from stakeholders, and pilot usability trial of the screening tool in a general sample of adolescents [13] are described in previously published papers.

Figure 1. Sample images of the Screening Wizard tool, as it appears to patients.

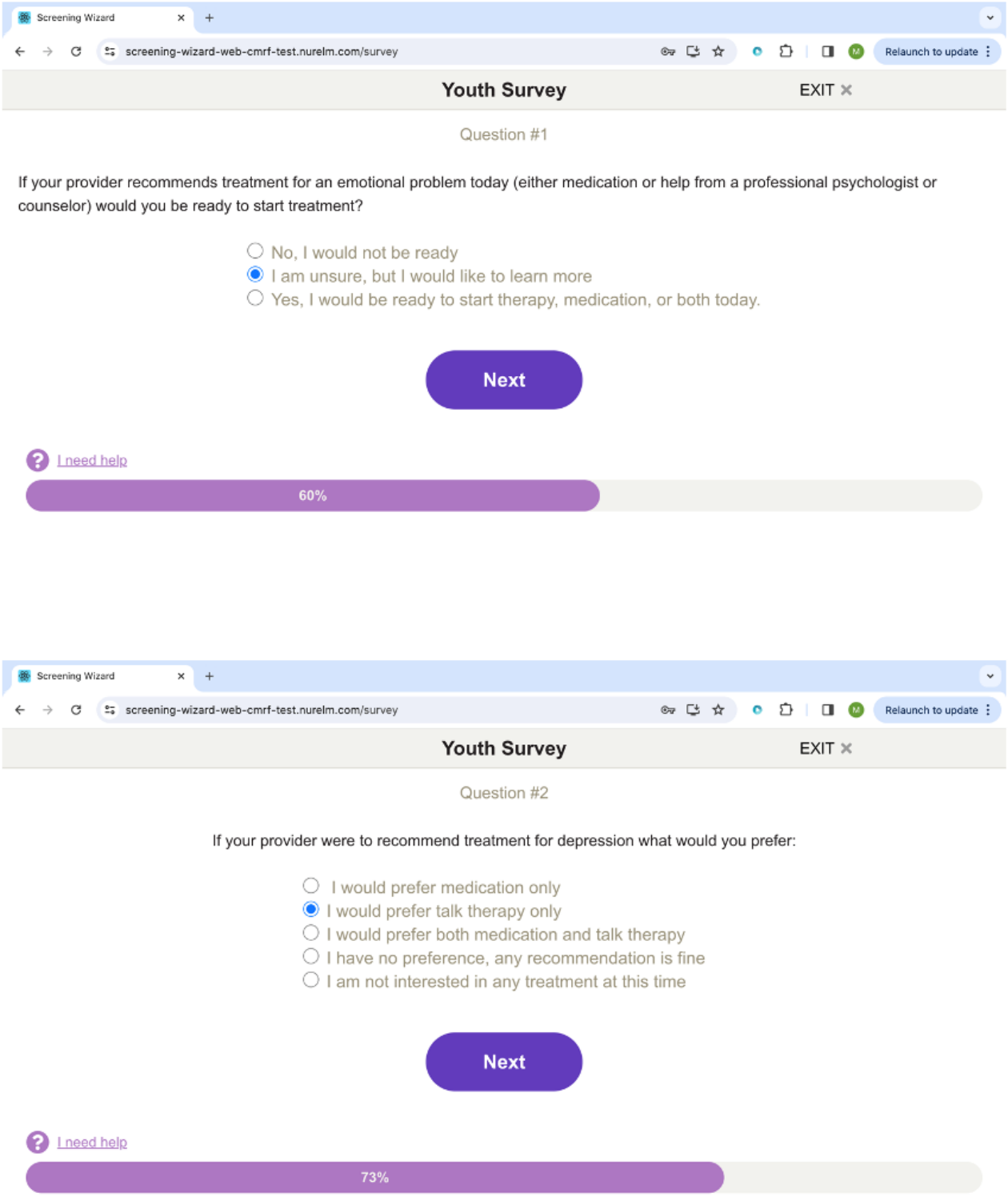

These previous studies and provider feedback informed several modifications to the enhanced screening tool so that it could be used clinically outside of a research context. The clinical use version was updated with functionality for a clinic-specific

XSL•FO
RenderX

Health Insurance Portability and Accountability Act (HIPAA)–secure log-in to the SW system to separate its administration and data collection from the SW research environment. This allowed clinical staff to assign adolescent patients and their caregiver unique weblinks to the enhanced screen and schedule when they would receive it through text message directly to their mobile phone; or alternatively print a list of QR code weblinks to distribute to individual patients. After completion, staff would be able to review results on a private clinician dashboard. In response to safety and liability concerns related to remote suicide screening, surveys were scheduled to expire at the end of the same clinic day to allow time for clinic staff to review all completed surveys. Planned changes also included condensing the number of items by presenting additional screening items only for participants with positive initial depression screens. Integration of the screening tool into the electronic health record (EHR) had been identified as a high priority initiative at this stage, though this required clinical system-wide permission, which was not accessible to the research team.

As the enhanced screening tool was initially developed within an academic setting with access to embedded behavioral health staff and research support, this study was intended to pilot the screening tool in more typical and less-resourced community health care settings where adolescents seek primary care or other services (eg, reproductive care) to elucidate barriers and facilitators to implementation from providers who are the intended users of the tool. This tool was offered to multiple clinical sites to pilot and provide feedback on the piloting process. Although health care providers at these sites were interested in using the intervention, many clinics (especially those within larger health care systems) had extensive IT audit and administrative approval processes to trial new technology tools, which presented substantial barriers to piloting. Only 3 out of 10 clinics were able to pilot the tool very briefly. It became apparent that barriers to implementation would differ in a real-world versus research setting and additional elicitation from providers about these barriers and facilitators would be needed to enable scalability. Preimplementation studies are most helpful when they allow participants to try out the tool [12,14]. However, due to the barriers to piloting, we developed an additional prototype of the tool for interested providers to test informally, using fabricated patient data.

In this study, we conducted structured interviews with primary care providers based on domains from the Consolidated Framework for Implementation Research [15,16], a widely used framework for systematically identifying barriers and facilitators to implementation of new interventions. The information learned will be used to inform future implementation efforts and make subsequent iterations of the enhanced screening tool better suited for use in a variety of practice settings. This paper describes our qualitative process and the feedback we received.

Methods

Participants

A total of 10 providers from 10 different clinical sites across the United States were recruited through purposive sampling to participate in structured interviews. Providers were included if they provided either primary care or specialty adolescent medicine care (eg, for reproductive, eating disorder, and gender concerns) to the adolescent age group. They were excluded if their clinic had participated in previous studies with the screening tool. Providers were recruited from primary care clinics within the University of Pittsburgh Medical Center network who had not participated in the previous study, and from a national organization adolescent medicine listserv where providers had previously expressed interest in trialing the clinical tool after Dr Radovic had replied to a member's question about existence of available computerized clinical tools.

Ethical Considerations

Participants provided verbal consent to participate in interviews and were compensated with a US $50 gift card. Participants and interview content were deidentified. This study was approved by the University of Pittsburgh institutional review board (STUDY21090136).

Procedure

Interested providers who were able to obtain organizational clearance were invited to trial the digitally enhanced mental health screening tool with actual patients, or if not, to informally explore a prototype of the tool on their own before the interview. All participants subsequently participated in a structured interview. Interview sessions were conducted by the first author through HIPAA-compliant Zoom (Zoom Communications) and comprised of a brief live demonstration of the screening tool followed by a 45-minute structured interview using a guide based on the Consolidated Framework for Implementation Research (CFIR) constructs [16] and informed by pilot work [13]. Table 1 shows the examples of interview questions.. The complete interview guide is available in Multimedia Appendix 1.

XSL•FO
RenderX

Table . Sample interview questions about a digitally enhanced mental health screening tool and their correspondence to the consolidated framework for implementation research.

CFIR[a] domain	Construct	Interview question
Inner setting	Compatibility	How well does the enhanced screening tool fit with existing screening practices in your clinic?
Process	Engaging	Who are the key influential individuals to get on board with the enhanced screening tool?
Intervention characteristics	Adaptability	What changes or alterations do you think would be needed to facilitate implementation and utility of the screening tool in your clinic?

[a]CFIR: Consolidated Framework for Implementation Research.

Coding and Qualitative Analysis

Interviews were audio-recorded, deidentified, and transcribed using Rev Human Transcription Services [17], removing identifiers and filler words (eg, "yeah"). Meeting notes were taken by the interviewer in real time for triangulation with eventual formal qualitative coding. Transcripts were coded by the first author following a template analysis approach [18] with NVivo software (version 14; Lumivero), using a prespecified codebook. The codebook was created according to a hybrid approach, using a deductive approach to generate high-level codes based on the CFIR Codebook Template [15], and an inductive approach to incorporate themes that arose from the data [19]. Subcodes were chosen based on topics recurring as key points across multiple interviews, and codes that did not fall within the CFIR framework were included on a case-by-case basis as relevant to future implementation of the screening tool. A research assistant coded data in parallel, and the 2 coders discussed and resolved discrepancies, with the availability of the senior author in case a resolution could not be found. The 2 coders had differences in organization of coding (for example, coding small sentence fragments versus coding complete blocks of text), but did not disagree on any content-level coding. All of these organizational discrepancies were resolved between the two coders and did not require escalation to the senior author. After consensus was reached between the two coders, the first author summarized and synthesized interview findings to present thematically. Participants were offered the opportunity to review findings presented in this manuscript and provide feedback for triangulation purposes. Participants did not suggest any corrections to the study's findings.

Results

Providers' feedback regarding the digitally enhanced mental health screening tool and its theoretical implementation in their respective practice settings focused on four major themes: (1) current screening practices, (2) perceived benefits of the tool, (3) barriers to implementation, and (4) structural or organizational considerations. We reached saturation around these major themes.

Description of Providers and Current Screening Practices

Of the providers contacted, 22 were invited by email to participate in interviews, 12 were interested in participating, and 10 ultimately completed interviews. All 10 providers viewed a prototype of the screening tool, and 3 of the providers attempted to pilot the tool in their respective clinics with patients but were unable to gain the required approval to do so. Of the 10 participants, 9 were physicians and 1 was a nurse practitioner. In addition, 8 of the providers specifically practiced adolescent medicine, while 2 providers worked in general pediatrics. In total, 7 providers practiced in hospital-based clinics associated with academic institutions, in a combination of primary care and adolescent specialty care settings. Furthermore, 1 provider worked in a college health clinic, and 2 providers practiced in community practices, with 1 working in a Federally Qualified Health Center.

In total, 6 of the 10 providers described using paper screeners for mental health, which clinic staff manually enter into the EHR, either by filling in a template with the patients' responses or scanning the screening result into the EHR as a media file. In addition, 2 providers used an electronic screening format, whereby patients answer questions electronically and these answers are automatically integrated into the EHR. Furthermore, 1 provider used a combination of paper and electronic screening modalities, and another reported no formalized mental health screening process.

Perceived Benefits of Digitally Enhanced Mental Health Screening Tool

Providers universally reported an uptick in adolescent depression and anxiety in recent years, particularly during the COVID-19 pandemic. Most providers supported mental health screening and treatment as high priority initiatives, with several reporting recent expansion of their practice's mental health services (eg, hiring new embedded mental health providers and partnering with telepsychiatry services). All providers perceived the digitally enhanced mental health screening tool would offer improvements on their current practices due to it being administered through mobile devices before appointments and eliciting more detailed information than currently used screeners. Providers generally had positive opinions about any new intervention that could increase efficiency of history-taking, allowing a larger proportion of appointment time to be devoted to management discussions. Some providers recognized potential benefits of the tool in assisting with diagnostic clarity due to its efficiency in eliciting additional information without relying on a longer appointment. Previewing the tool and the included measures inspired one provider to inquire about mania in an adolescent patient with depression:

XSL•FO
RenderX

> *I thought a lot about the mania [screener], because it's definitely not something that I traditionally screen for, unless I'm starting somebody on an SSRI…however, one of [my] patients screened positive for mania…I would not necessarily have thought to ask her those questions at that point, because we hadn't even really started talking about medication yet. So, it definitely did prompt me to [ask further questions. The patient then revealed a strong family history of bipolar disorder], and I was like, "Oh, okay. This is starting to make more sense." I think she actually does have bipolar type one.*

The tools' summarized results accompanied with additional symptom severity interpretation were recognized as efficiencies providers could use to identify mental health concerns at a glance and access direct links to current clinical care guidelines (ie, from the American Academy of Child and Adolescent Psychiatry and the American Academy of Pediatrics). One provider thought the tool would increase their comfort addressing mental health concerns in primary care, reflecting a lack of prior mental health training despite the need to address it regularly:

> *I wasn't trained on how to use SSRIs or even stimulant medication really, even though I knew I was going to primary care and I tried to get that training and now that's what I do a few times a day, which is kind of crazy.*

Providers proposed that if the tool is introduced as a standardized screening that all adolescents and caregivers within the practice are asked to complete, it could normalize mental health discussions while simultaneously providing psychoeducation. Adopting the screening tool into their clinical practice was mentioned by a provider as a way of "letting people know that this [answering questions about their mental health] is something okay to talk about even when they're not [currently symptomatic]," providing "a sense of safety and comfort for the patient." Another provider, who regularly precepts pediatric residents, appreciated that the tool automatically engages parents in conversation about mental health, stating that this is something trainees struggle to facilitate on their own. Another provider relayed the tool could help dismantle mental health stigma, as many parents show heightened concern navigating the lasting impacts of the COVID-19 pandemic on youth:

> *There's still a lot of stigma associated with mental health, but I think more and more people are getting more open, and I think a lot of parents will be very open to it right now, just given the current environment we're in, understanding that a lot of young people and teens, everybody needs help once in a while, and it's not something to be stigmatized.*

Providers thought the enhanced screening tool could be used in various patient visit types besides annual well child visits, including as a diagnostic aid for patients presenting with mental health chief complaints, given the tool's inclusion of several validated measures and specific questions regarding the patient's treatment preferences and perceived barriers to care; and as a college health clinic administered prescreening for transitional age youth upon entering college to facilitate parental involvement in students' transition to adult care.

Barriers to Implementation

The most frequently mentioned theme throughout provider interviews was that implementing a new digital screening, which is not embedded in the EHR would be exceedingly challenging, requiring clinical staff to take on burdensome tasks including scheduling and sending out survey links, monitoring responses, printing results for providers, and uploading results into the EHR. This concern was in the setting of clinics already struggling with staffing shortages, and the perspective that—while highly important—mental health screening is only one facet of well-child visits, which already involve multiple administrative tasks. For a new technology intervention to be adopted in their clinic, providers agreed it would need to show clear clinical benefit while simultaneously reducing burden on clinic staff. As such, EHR integration was seen as a necessity for successful implementation of the tool across clinical settings.

Providers expressed concern about the length of time needed to complete the screening survey, both from a workflow and user perspective. Providers suspected patients and their caregivers may be unwilling to complete the entire tool, noting the form could be particularly time-consuming for individuals with low literacy levels. Providers shared that if a patient or caregiver did not have a chance to complete the survey before their visit, it may negatively affect the duration of time patients spend in exam rooms, appointment length, room turnover, and patient wait time.

Nearly all providers endorsed safety and liability concerns related to screening for suicidality. Despite the screening tool being designed to limit availability of the survey to the clinic's hours of operation as described above, providers worried about setting up foolproof protocols for a notification system for completed screens, and assigning responsibility to staff for reviewing screens and identifying patients at risk in a timely manner to avoid positive screens being missed, or loss to follow-up:

> *The other thing that I think about too [is] getting a positive screen [from] someone who fills that out from home and then they don't come to the appointment. That could be a big deal because it's going to be our patients with mental health disorders. Generally, if a teenager has a mental health disorder, there's a reasonable chance that a parent has a mental health disorder. If they have a mental health disorder that's not treated, which is a lot of mental health disorders, it makes them less likely to follow through with things like appointments. So then you have a patient who's suicidal and is out there, or who's significantly affected and is out there, and then you can't get ahold of them because of that stuff. That makes me scared. Now is that better than just being dumb to it? Maybe it's better, but it makes me nervous.*

Similarly, providers worried about balancing measures and the unintended harms of screening broadly for mental health without adequate time allotted during the appointment to address

XSL•FO
RenderX

concerns that the screening tool may identify. A provider was concerned about the screening tool's potential to uncover otherwise undetected mental health concerns in patients without having the resources such as behavioral health staff or referral resources to respond appropriately.

Providers raised concerns about patient privacy when using a third-party product to obtain highly sensitive protected health information. In addition, providers worried about the potential for caregivers to access a patient's portal or survey or even to hover over their child while they complete the survey, which could in turn cause patients to answer questions inaccurately or not engage fully with the screening tool.

Accessibility was a major theme across all interviews, particularly for providers who worked with underserved populations. Providers stated it was important for the screening tool to be available in other languages, noting that if a child is fluent in English, their caregiver may be less fluent and require a version in their native language to obtain accurate collateral information. Some also wanted the screening tool to be available in other modalities, such that it would be accessible to those with poor literacy or learning disabilities. As with any new technology intervention, questions arose about health equity and inaccessibility for those without mobile devices or reliable internet access. There was general consensus that, while new technologies have potential to mitigate certain health care disparities, they also have the potential to exacerbate others by virtue of being inaccessible to many underserved populations, which is consistent with concerns in the existing literature [20].

Structural and Organizational Considerations

The most frequently mentioned requirements for getting leadership on board with implementing a new intervention were that the intervention is efficient, enhances quality of care, increases patient satisfaction, and has proof of clinical benefit. It was also important that the intervention be simple and easy for staff to use, so as not to burden staff or negatively impact clinic workflow.

Providers identified a major barrier to implementing new technologies as lack of transparency about the processes in their institution**,** with many providers unable to identify who they would need to obtain permission from within their organization to implement a new technology within their clinic. Providers described complicated networks of clinical leadership, business executives, and representatives from information technology and their respective EHRs that all must collaborate to approve new technologies. A provider reported that they previously attempted to implement a new technology intervention in their hospital-based clinic but grew frustrated and abandoned the initiative after numerous unfruitful meetings and emails with the administration.

Some providers also mentioned public insurance documentation requirements as a barrier to using a third-party tool, as public insurance plans often require routine screenings be completed using specific forms. Providers worried some patients would have to complete the same screening questions (eg, Patient Health Questionnaire-9) twice for the same office visit—once in the enhanced screening tool, and a second time on the form required for reimbursement.

Recommended Alterations to the Digitally Enhanced Mental Health Screening Tool

Integration of the enhanced screening tool into the EHR and automation of survey creation and distribution were thought to be the strongest facilitators necessary for implementing the tool in pediatric clinics. All 10 providers were in favor of redesigning the survey to offer all patients a subset of initial questions, and only the full screen to those patients who answered positively to those initial questions, as a means for increasing the screening tool's efficiency.

Providers agreed that customization options such as ability to choose which screening tools to include would make the intervention more appealing to a variety of practice settings. In addition to selecting or deselecting currently available measures, providers were interested in adding screening tools for the following (in order of frequency mentioned): disordered eating, sleep, adverse childhood events, social determinants of health, attention-deficit/hyperactivity disorder (ADHD), academic distress, and demographic questions about sexual orientation and gender identity.

Discussion

Principal Findings

This preimplementation study aimed to elicit provider perspectives regarding proposed implementation of a digitally enhanced mental health screening tool in their respective clinical settings to inform future implementation strategies and alterations to the screening tool. While all providers stated that improving mental health screening was a high priority in their clinic and believed that the screening tool had the potential to improve such practices, many providers expressed that the screening tool may not be feasible to implement in its current form, most notably due to the tool being a third-party application without EHR integration.

As new technologies for health care information gathering, evaluation, and intervention delivery become an increasingly popular area of interest, there is a simultaneous need for improved systems interoperability to allow for exchange of information between new technology tools and existing EHRs in order for such technologies to be practical for use in real clinical settings [21]. While EHRs show provider notes and metrics from previous medical encounters, they have limited ability to pull in data and services from third-party tools that have potential to enhance quality of health care, decrease health care costs, and improve patient outcomes. The lack of connectivity between different systems and software presents a nearly impermeable barrier to implementation of new technologies, highlighting the dire need for an application-based component to health information technology to allow for adaptation of new technologies and seamless integration with existing EHRs [22]. Several large research-based hospital systems have successfully deployed questionnaires for patient reported outcomes, the results of which are directly incorporated in the electronic medical record [23]. Such initiatives can serve

XSL•FO
RenderX

as a model for similar integration of screening tools such as ours and promote organizational buy-in by demonstrating that wide-scale implementation is feasible. The findings from this study further emphasize the need for enhanced systems interoperability in the health care space, and exemplify another way that improvement in this area can facilitate implementation of innovative health care technologies. Beyond the logistical barriers to implementation of the screening tool, providers identified administrative buy-in as another major barrier. Providers reported hesitancy at the administrative level due to new challenges that such technologies present, such as data privacy concerns and questions about liability in a relatively uncharted digital landscape. The findings in this study further reinforce the already identified need to address such issues at a structural or policy level [24] so that individual providers can more easily and independently trial and use new technology tools that have the potential to provide clinical benefit and improve patient outcomes without requiring system-wide changes.

Limitations

While homogeneity within our sample allowed thematic saturation to be reached, our findings may be less generalizable to all pediatric primary care settings due to our sample being small and self-selective, comprised of providers who may be more interested in improving adolescent behavioral health services as they expressed interest in learning how a digitally enhanced screening tool could augment their existing clinical practices. While not all primary care practices will have clinician leaders interested and willing to take on the burdensome task of pioneering new interventions, early adopters are critical to intervention development and implementation such that these new tools can be optimized, scalable and feasible to disseminate more widely [25,26]. Provider feedback was also limited by the inability to use the screening tool in vivo with real patients in real time, again underscoring the importance of pilot work and a need for greater consideration regarding the process of trialing new interventions in the community (eg, flexible method designs, quasi-experimental and nonrandomized approaches that account for the need to maintain efficient workflow in practice settings).

Conclusion

While technology interventions can help address the need for improved behavioral health support in primary care settings, numerous barriers to implementation of such tools exist. This study adds to the literature showing a need for improved health information systems interoperability for innovative technologies to be used to their greatest potential in real-world settings.

Acknowledgments

This study was funded by the University of Pittsburgh Medical Center Competitive Medical Research Fund, the ETUDES Center at the University of Pittsburgh (1P50MH115838-01), and the University of Pittsburgh School of Medicine Roth Fellow Award.

Data Availability

The data from this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Screening Wizard provider interview guide.
[DOCX File, 19 KB - formative_v9i1e67624_app1.docx]

References

1. Ros-DeMarize R, Chung P, Stewart R. Pediatric behavioral telehealth in the age of COVID-19: Brief evidence review and practice considerations. Curr Probl Pediatr Adolesc Health Care 2021 Jan;51(1):100949. [doi: 10.1016/j.cppeds.2021.100949] [Medline: 33436319]
2. 2022 NSDUH annual national report. : SAMHSA; 2022 URL: https://www.samhsa.gov/data/report/2022-nsduh-annual-national-report [accessed 2024-10-18]
3. Hua LL, Lee J, Rahmandar MH, Sigel EJ, COMMITTEE ON ADOLESCENCE, COUNCIL ON INJURY, VIOLENCE, AND POISON PREVENTION. Suicide and suicide risk in adolescents. Pediatrics 2024 Jan 1;153(1):e2023064800. [doi: 10.1542/peds.2023-064800] [Medline: 38073403]
4. Zuckerbrot RA, Cheung A, Jensen PS, Stein REK, Laraque D, GLAD-PC STEERING GROUP. Guidelines for Adolescent Depression in Primary Care (GLAD-PC): Part I. Practice preparation, identification, assessment, and initial management. Pediatrics 2018 Mar;141(3):e20174081. [doi: 10.1542/peds.2017-4081] [Medline: 29483200]
5. Cheung AH, Zuckerbrot RA, Jensen PS, Laraque D, Stein REK, GLAD-PC STEERING GROUP. Guidelines for Adolescent Depression in Primary Care (GLAD-PC): Part II. Treatment and ongoing management. Pediatrics 2018 Mar;141(3):e20174082. [doi: 10.1542/peds.2017-4082] [Medline: 29483201]

XSL•FO
RenderX

6. Suicide: blueprint for youth suicide prevention. American Academy of Pediatrics. 2021. URL: https://www.aap.org/en/patient-care/blueprint-for-youth-suicide-prevention/ [accessed 2024-10-18]
7. US Preventive Services Task Force, Mangione CM, Barry MJ, et al. Screening for depression and suicide risk in children and adolescents: US preventive services task force recommendation statement. JAMA 2022 Oct 18;328(15):1534-1542. [doi: 10.1001/jama.2022.16946] [Medline: 36219440]
8. Asnaani A, Narine K, Suzuki N, et al. An innovative mobile game for screening of pediatric PTSD: a study in primary care settings. J Child Adolesc Trauma 2021 Sep;14(3):357-366. [doi: 10.1007/s40653-020-00300-6] [Medline: 34471454]
9. McNeely J, Mazumdar M, Appleton N, et al. Leveraging technology to address unhealthy drug use in primary care: effectiveness of the substance use screening and intervention tool (SUSIT). Subst Abus 2022;43(1):564-572. [doi: 10.1080/08897077.2021.1975868] [Medline: 34586976]
10. Spence EE, Prabhakar P, Grace J, Fulda KG, Thompson EL, Ondersma SJ. Development and implementation of tablet-based screening for interpersonal violence in primary care settings. Health Care Women Int 2022 Dec;43(12):1503-1509. [doi: 10.1080/07399332.2022.2043861] [Medline: 35235498]
11. Peahl AF, Rubin-Miller L, Paterson V, et al. Understanding social needs in pregnancy: prospective validation of a digital short-form screening tool and patient survey. AJOG Glob Rep 2023 Feb;3(1):100158. [doi: 10.1016/j.xagr.2022.100158] [Medline: 36922957]
12. Radovic A, Anderson N, Hamm M, et al. Perceived facilitators of and barriers to implementation of a decision support tool for adolescent depression and suicidality screening: focus group and interview study (preprint). JMIR Mental Health. Preprint posted online on 2020. [doi: 10.2196/preprints.26035]
13. Radovic A, Coren MA, Wecht S, et al. Detecting and managing adolescent depression in primary care: a pilot randomized controlled trial. Procedia Comput Sci 2022;206:81-91. [doi: 10.1016/j.procs.2022.09.087]
14. Taylor SP, Short RT, Asher AM, Taylor B, Beidas RS. A rapid pre-implementation evaluation to inform a family engagement navigator program during COVID-19. Implement Sci Commun 2020 Dec 9;1(1):110. [doi: 10.1186/s43058-020-00098-2] [Medline: 33298192]
15. CFIR guide. Consolidated Framework for Implementation Research. 2024. URL: https://cfirguide.org/guide/app [accessed 2024-10-18]
16. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. Implement Sci 2009 Aug 7;4(1):50. [doi: 10.1186/1748-5908-4-50] [Medline: 19664226]
17. Rev. URL: https://www.rev.com/ [accessed 2025-03-28]
18. King N. Essential Guide to Qualitative Methods in Organizational Research: SAGE Publications Ltd; 2004. [doi: 10.4135/9781446280119]
19. Hamilton AB, Finley EP. Qualitative methods in implementation research: an introduction. Psychiatry Res 2019 Oct;280:112516. [doi: 10.1016/j.psychres.2019.112516] [Medline: 31437661]
20. Ralston AL, Andrews AR, Hope DA. Fulfilling the promise of mental health technology to reduce public health disparities: review and research agenda. Clinical Psychology: Science and Practice 2019;26(1). [doi: 10.1111/cpsp.12277]
21. O'Connor S. What is interoperability, and why is it important? Advanced Data Systems Corporation. 2017 May 30. URL: https://www.adsc.com/blog/what-is-interoperability-and-why-is-it-important [accessed 2024-10-18]
22. Mandl KD, Mandel JC, Kohane IS. Driving innovation in health systems through an apps-based information economy. Cell Syst 2015 Jul;1(1):8-13. [doi: 10.1016/j.cels.2015.05.001] [Medline: 26339683]
23. Papuga MO, Dasilva C, McIntyre A, Mitten D, Kates S, Baumhauer JF. Large-scale clinical implementation of PROMIS computer adaptive testing with direct incorporation into the electronic medical record. Health Syst (Basingstoke) 2018;7(1):1-12. [doi: 10.1057/s41306-016-0016-1] [Medline: 31214335]
24. Mohr DC, Azocar F, Bertagnolli A, et al. Banbury forum consensus statement on the path forward for digital mental health treatment. Psychiatr Serv 2021 Jun;72(6):677-683. [doi: 10.1176/appi.ps.202000561] [Medline: 33467872]
25. Green L, Gottlieb N, Parcel G. Diffusion theory extended and applied. Advances in Health Education and Promotion 1991.
26. Van Houtven CH, Drake C, Malo TL, et al. Ready, set, go! The role of organizational readiness to predict adoption of a family caregiver training program using the Rogers' diffusion of innovation theory. Implement Sci Commun 2023 Jun 19;4(1):69. [doi: 10.1186/s43058-023-00447-x] [Medline: 37337208]

Abbreviations

ADHD: attention-deficit/hyperactivity disorder
CFIR: Consolidated Framework for Implementation Research
EHR: electronic health record
HIPAA: Health Insurance Portability and Accountability Act
PCP: primary care provider
PTSD: posttraumatic stress disorder
SW: Screening Wizard

XSL•FO
RenderX

Edited by A Mavragani; submitted 23.10.24; peer-reviewed by H Kautz, R Pearson; revised version received 26.02.25; accepted 11.03.25; published 10.04.25.

Please cite as:
Coren MA, Lindhiem O, Angus AR, Toevs EK, Radovic A
Provider Perspectives on Implementing an Enhanced Digital Screening for Adolescent Depression and Suicidality: Qualitative Study
JMIR Form Res 2025;9:e67624
URL: https://formative.jmir.org/2025/1/e67624
doi:10.2196/67624

© Morgan A Coren, Oliver Lindhiem, Abby R Angus, Emma K Toevs, Ana Radovic. Originally published in JMIR Formative Research (https://formative.jmir.org), 10.4.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Reliability of Average Daily Steps Measured Through a Consumer Smartwatch in Parkinson Disease Phenotypes, Stages, and Severities: Cross-Sectional Study

Edoardo Bianchini[1,2], MD; Domiziana Rinaldi[1], MD; Lanfranco De Carolis[1], MD; Silvia Galli[1], MD; Marika Alborghetti[1], MD; Clint Hansen[2,3], PhD; Antonio Suppa[4,5], MD; Marco Salvetti[1,5], Prof Dr Med; Francesco Ernesto Pontieri[1], Prof Dr Med; Nicolas Vuillerme[2,6], PhD

[1]Department of Neuroscience, Mental Health and Sensory Organs (NESMOS), Sapienza University of Rome, Rome, Italy

[2]AGEIS, Université Grenoble Alpes, Grenoble, France

[3]Department of Neurology, Kiel University, Kiel, Germany

[4]Department of Human Neurosciences, Sapienza University of Rome, Rome, Italy

[5]IRCCS Neuromed Institute, Pozzilli, Italy

[6]Institut Universitaire de France, Paris, France

Corresponding Author:
Edoardo Bianchini, MD
Department of Neuroscience, Mental Health and Sensory Organs (NESMOS), Sapienza University of Rome, Rome, Italy

Abstract

Background: Average daily steps (avDS) could be a valuable indicator of real-world ambulation in people with Parkinson disease (PD), and previous studies have reported the validity and reliability of this measure. Nonetheless, no study has considered disease phenotype, stage, and severity when assessing the reliability of consumer wrist-worn devices to estimate daily step count in unsupervised, free-living conditions in PD.

Objective: This study aims to assess and compare the reliability of a consumer wrist-worn smartwatch (Garmin Vivosmart 4) in counting avDS in people with PD in unsupervised, free-living conditions among disease phenotypes, stages, and severity groups.

Methods: A total of 104 people with PD were monitored through Garmin Vivosmart 4 for 5 consecutive days. Total daily steps were recorded and avDS were calculated. Participants were dichotomized into tremor dominant (TD; n=39) or postural instability and gait disorder (PIGD; n=65), presence (n=57) or absence (n=47) of tremor, and mild (n=65) or moderate (n=39) disease severity. Based on the modified Hoehn and Yahr scale (mHY), participants were further dichotomized into earlier (mHY 1 - 2; n=68) or intermediate (mHY 2.5 - 3; n=36) disease stages. Intraclass correlation coefficient (ICC; 3,k), standard error of measurement (SEM), and minimal detectable change (MDC) were used to evaluate the reliability of avDS for each subgroup. The threshold for acceptability was set at an ICC ≥0.8 with a lower bound of 95% CI ≥0.75. The 2-tailed Student *t* tests for independent groups and analysis of 83.4% CI overlap were used to compare ICC between each group pair.

Results: Reliability of avDS measured through Garmin Vivosmart 4 for 5 consecutive days in unsupervised, free-living conditions was acceptable in the overall population with an ICC of 0.89 (95% CI 0.85 - 0.92), SEM below 10%, and an MDC of 1580 steps per day (27% of criterion). In all investigated subgroups, the reliability of avDS was also acceptable (ICC range 0.84 - 0.94). However, ICCs were significantly lower in participants with tremor (P=.03), with mild severity (P=.04), and earlier stage (P=.003). Moreover, SEM was below 10% in participants with PIGD phenotype, without tremor, moderate disease severity, and intermediate disease stage, with an MDC ranging from 1148 to 1687 steps per day (18% - 25% of criterion). Conversely, in participants with TD phenotype, tremor, mild disease severity, and earlier disease stage, SEM was >10% of the criterion and MDC values ranged from 1401 to 2263 steps per day (30% - 33% of the criterion).

Conclusions: In mild-to-moderate PD, avDS measured through a consumer smartwatch in unsupervised, free-living conditions for 5 consecutive days are reliable irrespective of disease phenotype, stage, and severity. However, in individuals with TD phenotype, tremor, mild disease severity, and earlier disease stages, reliability could be lower. These findings could facilitate a broader and informed implementation of avDS as an index of ambulatory activity in PD.

(JMIR Form Res 2025;9:e63153) doi:10.2196/63153

XSL•FO
RenderX

KEYWORDS

gait; Parkinson disease; phenotype; wearable sensors; smartwatch; step count; reliability; activity monitor; digital health technology; digital outcome measures; wearable; mHealth; motor; quality of life; fall; posture; mobile health

Introduction

Walking is a fundamental motor ability, and it is pivotal for functional independence and social well-being [1]. Gait disturbances are common in people with Parkinson disease (PD), and include shuffling gait, shortened step length, altered automaticity, decreased arm swing, and freezing of gait [2]. These represent a particularly disabling group of symptoms, significantly hampering the quality of life of people with PD [3] and increasing the risk of falls [4].

Daily steps are an easy-to-collect and useful measure of ambulatory activity and mobility [5]. Although this parameter could not provide details regarding subtle gait features, several evidence linked a reduced daily step count to overall mortality risk [6-9], as well as to a range of health conditions, such as dementia [10] and oncological and cardiovascular diseases [11,12]. Previous studies have also reported a negative correlation between daily steps and disease severity in PD [13] and proposed a minimum daily step goal of 4200 to match with physical activity recommendations in the early stages of the disease [14].

Wearable devices, including smartwatches, could represent a useful option to estimate daily steps in an unobtrusive, ecological way [15]. Moreover, wearables are widely available on the market, easily used by the general population, and enable unobtrusive continuous long-term data collection [16]. However, since these devices are usually tested in healthy populations, knowledge of the validity and reliability of collected data is generally limited when applied to different groups of patients. Therefore, a growing body of literature focused on the validity and reliability of consumer wearable devices for step counting, with generally positive results [17-19].

In PD, motor and gait manifestations could render step detection and step count challenging and hence significantly diminish the validity and reliability of device algorithms [20-22]. Nevertheless, a prior study by Ginis et al [23] demonstrated a good criterion validity in estimating average daily steps (avDS) of 2 Fitbit devices (Fitbit Alta and Fitbit Inspire 3) in 28 people with PD in a real-life setting, compared to a research-grade device (Dynaport Movemonitor, McRoberts, NL).

Similarly, a prior study from our group involving 47 individuals with PD demonstrated a good criterion validity in step counting using a consumer smartwatch (Garmin Vivosmart 4), when worn on the side least affected by the disease and under well-controlled pharmacological conditions in a supervised, in-clinic setting [20].

Besides criterion validity, we recently demonstrated that optimal reliability of avDS recorded by Garmin Vivosmart 4 in real-life conditions could be achieved if the smartwatch is worn for a minimum of 4 days [24].

However, the clinical heterogeneity of PD when evaluating metrological characteristics of step-counting devices has been neglected so far. Indeed, the clinical presentation of PD is highly variable among individuals, thus, significant efforts have been made to identify distinct clusters and subtypes [25]. Several classifications have been proposed over the years from clinical-based [26-28] to more recent biomarker-based classifications [29-31]. In this regard, one of the most used classifications distinguishes PD with predominant features of tremor (ie, tremor dominant [TD]) or gait, posture, and balance issues (ie, postural instability and gait disorder [PIGD]) [26-28] based on subitems scores of the Unified Parkinson's Disease Rating Scale (UPDRS) and its revision by the Movement Disorder Society (MDS-UPDRS) [32]. This classification could be relevant when measuring avDS in PD since tremors could increase the noise-to-signal ratio, making step detection more challenging [21]. Similarly, in PIGD individuals, the higher degree of gait alterations could alter the performance of step-detection algorithms [20,22].

Symptom severity and disease stage could also represent other relevant parameters to be considered when assessing the reliability of any wearable devices in counting avDS. Indeed, with disease progression, gait features increasingly deviate from normality [2,33], and tremor and bradykinesia could further alter the spatiotemporal and kinematic characteristics of walking, dampening, in turn, algorithm performance in step detection [20-22].

Nevertheless, no study to date considered disease phenotype, stage, and symptom severity when assessing the reliability of consumer wrist-worn devices for step counting in unsupervised, free-living conditions in mild-to-moderate people with PD. This study was hence specifically designed to address this issue. We hypothesized that reduced reliability might be observed in TD individuals and people with PD with more severe symptoms and in more advanced disease stages due to the aforementioned increased signal noise due to tremor, the higher degree of motor symptoms, and the more marked gait alterations. In addition, since no previous study specifically investigated the criterion validity of Garmin Vivosmart 4 in estimating avDS, we performed a proof-of-concept experiment in a subgroup of participants. Details of methods and results of this latter experiment are presented in Multimedia Appendix 1.

Methods

Population

Participants were consecutively screened and recruited during scheduled visits at the Movement Disorder Outpatient Service of the Sant'Andrea University Hospital (Rome, Italy) in the period between March 2023 and March 2024. Inclusion criteria were: (1) diagnosis of idiopathic PD according to the MDS criteria (Postuma et al [34]); (2) aged 18 years or older; (3) disease stage <4 according to the modified Hoehn and Yahr scale (mHY) [35]; (4) classification as TD or PIGD according to Stebbins et al [27]; and (5) stable medication in the 4 weeks before the experimental procedure. Exclusion criteria were: (1)

XSL•FO
RenderX

cognitive impairment, defined by Montreal Cognitive Assessment [36] score <21; and (2) orthopedic, rheumatologic, or systemic conditions affecting mobility as judged by the assessor.

Ethical Considerations

This cross-sectional study was performed in accordance with the ethical standards laid down in the 1964 Declaration of Helsinki and its later amendments. Approval was granted by the local Ethical Committee of Sapienza, University of Rome, Italy (0372/2022). Data collection and processing followed the current European regulations for data protection. All participants provided written informed consent before the beginning of measurements. All data were deidentified. Participants did not receive any form of compensation.

Demographic and Clinical Data

Participants were evaluated during scheduled visits. Demographics and anthropometric measures (including age, sex, weight, height, and BMI) were collected. Disease duration and disease stage according to mHY and levodopa equivalent daily dose [37] were also collected. MDS-UPDRS [32] part III was used to assess motor symptoms severity.

Participants were divided into 4 subgroup pairs. Based on MDS-UPDRS part II and III scores, participants were classified into TD or PIGD disease subtypes according to Stebbins et al [27]. To evaluate the effect of tremor presence on device reliability, participants were also classified as those with and without tremor based on a score of ≥1 at item 2.10 of MDS-UPDRS part II. Concerning disease severity, participants were grouped into those with mild or moderate disease severity based on the MDS-UPDRS score as proposed by Martínez-Martín et al [38]. Similarly, participants were dichotomized in earlier (mHY 1 - 2) or intermediate (mHY 2.5 - 3) stages, based on mHY score.

Experimental Procedure

Participants received the smartwatch Garmin Vivosmart 4 after the visit and were instructed to wear it at home for a minimum of 5 consecutive days, including at least 1 weekend day, on the wrist of the body side least affected by the disease [20]. No reminders or further instructions were provided to participants during registration. We chose a 5-day period since we demonstrated previously that a minimum of 4 days of monitoring is needed to reliably estimate daily step count in PD [24]. Each smartwatch was configured according to the producer's recommendations and participants were asked to perform daily activities as usual. After 5 days, participants returned the smartwatch. The total daily number of steps for each day was recorded and avDS were calculated [24]. Compliance was assessed based on the participants' dashboard data. The device recognizes that it is worn through heart rate and inertial motion unit signal. We considered all recording days with >80% wear time while awake and no interruption of device use greater than 3 hours to be valid. Details of methods of the proof-of-concept experiment regarding criterion validity of Garmin Vivosmart 4 are presented in Multimedia Appendix 2.

Data and Statistical Analysis

The statistical analyses were performed using JASP (version 0.18.3.0; JASP Team), R (version 4.0.3; R Core Team), and RStudio (version 2023.12.0+369; R Foundation for Statistical Computing) for Windows. Descriptive statistics were calculated for the examined variables. The normality of distributions was assessed by histogram and residual plots inspection.

To evaluate the relative reliability for the 5-day monitoring period in the overall population and each subgroup, a 2-way intraclass correlation coefficient (ICC) with a fixed set of raters and averaged ratings was used (ICC (3,k), where k was the number of days of measurement), together with a custom R script. The following reference cut-off values for ICC interpretation were used [39]: excellent: >0.90; good: 0.75 - 0.90; moderate: 0.50 - 0.75; and poor: <0.50. The a priori threshold for acceptable ICC was set at a point estimate ≥0.80 with a lower bound of 95% CI ≥0.75 in accordance with a previous study from our group [24].

To compare ICCs between the 4 subgroup pairs, 2 methods were applied. First, standard errors and point estimates of ICCs were used to compute *t* statistics and perform 2-tailed independent groups Student *t* tests. Second, the CI overlap between each group pair was graphically and numerically assessed. Nonoverlapping CIs were considered indicators of significantly different ICCs [40]. Previous evidence underscored that a 95% CI overlap assessment could inflate the risk of type II error and suggested that an 83.4% CI could be a more powerful option [40-42]. Therefore, we adopted this method for CI overlap evaluation.

Standard error of measurement (SEM) and minimal detectable change (MDC) with a CI of 95% were used to compute the absolute reliability for the 5-day recordings in the overall population and each subgroup [43]. SEM and MDC were reported as absolute value and percentage of criterion measure (SEM% and MDC%, respectively). The criterion was the avDS count derived from the 5 days. For all analyses, the significance threshold was set at $\alpha<.05$. All data were reported as mean (SD) or median (IQR) for numerical data and n (%) for categorical variables.

Results

Overview

A total of 104 people with PD were included in the study. All participants were monitored through Garmin Vivosmart 4 at home for a period of 5 consecutive days. No participants or days were excluded based on the prespecified compliance criteria. Participants took on average 5923 (SD 3014) daily steps, ranging from 357 to 12,620. Details of demographic, anthropometric, and clinical variables of the study population are shown in Table 1. The overlap between the subgroups is shown in Multimedia Appendix 1.

Overall, the results of the proof-of-concept experiment suggest that the smartwatch Garmin Vivosmart 4 is valid and decently accurate in estimating avDS in PD. Details are presented in the Multimedia Appendix 2.

XSL•FO
RenderX

Table . Demographic, anthropometric, and clinical characteristics of the study population.

	Overall (N=104)	TD[a] (n=39)	PIGD[b] (n=65)	Tremor (n=57)	No tremor (n=47)	Mild (n=65)	Moderate (n=39)	mHY[c] 1 - 2 (n=68)	mHY 2.5 - 3 (n=36)
Age (years), mean (SD)	68.0 (8.4)	66.4 (9.0)	68.9 (8.0)	69.5 (7.9)	66.7 (8.8)	66.8 (8.7)	69.9 (7.7)	65.3 (7.8)	73.0 (7.3)
Height (cm), mean (SD)	171 (9.0)	173 (7.6)	170 (9.6)	170 (9.6)	172 (8.4)	173 (9.3)	168 (7.6)	173 (8.1)	168 (9.8)
Weight (kg), mean (SD)	75.7 (13.1)	76.7 (13.0)	75.1 (13.2)	75.0 (12.1)	76.3 (13.9)	77.2 (13.7)	73.2 (11.8)	77.7 (13.8)	71.9 (10.8)
BMI (kg/m^2), mean (SD)	25.7 (3.4)	25.5 (3.6)	25.8 (3.4)	25.8 (3.0)	25.7 (3.8)	25.7 (3.5)	25.8 (3.4)	25.9 (3.7)	25.4 (3.0)
Sex (female), n (%)	34 (33)	11 (28)	23 (35)	19 (33)	15 (32)	22 (34)	12 (31)	22 (32)	12 (33)
Disease duration (years), mean (SD)	6.4 (4.4)	5.2 (4.4)	7.0 (4.4)	7.2 (3.9)	5.7 (4.8)	5.7 (4.5)	7.5 (4.2)	5.3 (4.3)	8.4 (4.1)
LEDD[d] (mg), mean (SD)	553 (302)	418 (247)	634 (304)	623 (271)	495 (316)	489 (289)	659 (296)	453 (248)	741 (307)
mHY, median (IQR)	2 (2-2.5)	2 (1-2)	2 (2-2.5)	2 (2-2.5)	2 (2-2)	2 (2-2)	2.5 (2-3)	2 (2-2)	2.5 (2.5 - 3)
MDS-UPDRS-III[e], median (IQR)	27 (21 - 32)	26 (18 - 31)	29 (22 - 33)	29 (22 - 33)	26 (21 - 32)	23 (19-29)	33 (29 - 37)	23 (18 - 29)	33 (29 - 36)
avDS[f], mean (SD)	5923 (3014)	6654 (2733)	5485 (3109)	4594 (2612)	7020 (2898)	6512 (2857)	4942 (3049)	6838 (2908)	4195 (2419)

[a]TD: tremor dominant.

[b]PIGD: postural instability and gait disorder.

[c]mHY: modified Hoehn and Yahr scale.

[d]LEDD: levodopa equivalent daily dose.

[e]MDS-UPDRS III: Movement Disorder Society Unified Parkinson's Disease Rating Scale part III.

[f]avDS: average daily steps.

Reliability of avDS in People With PD

Relative Reliability

AvDS collected during 5 consecutive days showed a level of relative reliability above the threshold of acceptability, as indicated by an ICC point estimate of ≥0.80 and a lower 95% CI limit of ≥0.75, in the overall population and all subgroups. Moreover, daily step count showed excellent reliability in PD in the intermediate disease stage. Details of ICC and CI limits are shown in Table 2.

XSL•FO
RenderX

Table . ICC[a] (3,k) values with 95% CI for the overall population and each subgroup[b].

	Overall (N=104)	TD[c] (n=39)	PIGD[d] (n=65)	Tremor (n=57)	No tremor (n=47)	Mild (n=65)	Moderate (n=39)	mHY[e] 1 - 2 (n=68)	mHY 2.5 - 3 (n=36)
ICC (3,k)	0.888	0.854	0.899	0.838	0.914	0.856	0.919	0.839	0.939
Lower 95% CI	0.850	0.767	0.855	0.760	0.868	0.793	0.871	0.769	0.900
Upper 95% CI	0.919	0.916	0.933	0.896	0.947	0.905	0.953	0.892	0.966
Lower 83.4% CI	N/A[f]	0.797	0.869	0.786	0.883	0.814	0.887	0.792	0.914
Upper 83.4% CI	N/A	0.901	0.925	0.881	0.939	0.893	0.945	0.878	0.959

[a]ICC: intraclass correlation coefficient.

[b]For subgroups, 83.4% CI to assess intervals overlap are also reported.

[c]TD: tremor dominant.

[d]PIGD: postural instability and gait disorder.

[e]mHY: modified Hoehn and Yahr scale.

[f]Not applicable.

Absolute Reliability

AvDS showed an SEM below 10% in the overall population with an MDC of 1580 (26.7% of the criterion). AvDS also showed an SEM below 10% in the PIGD disease subtype, in participants without tremor, with a moderate disease severity, and in an intermediate disease stage with an MDC ranging from 1148 to 1687 steps per day (18% to 25% of criterion). Conversely, in the TD disease subtype, in participants with tremor, with mild disease severity and in an early disease stage, SEM was >10% of criterion and MDC values ranged from 1401 to 2263 steps per day (30% to 33% of the criterion). Details of SEM and MDC are shown in Table 3.

Table . Absolute and percentage values of SEM[a] and MDC[b] for the overall population and each subgroup.

	Overall (N=104)	TD[c] (n=39)	PIGD[d] (n=65)	Tremor (n=57)	No tremor (n=47)	Mild (n=65)	Moderate (n=39)	mHY[e] 1 - 2 (n=68)	mHY 2.5 - 3 (n=36)
SEM	570	742	495	506	609	701	414	817	267
SEM%	9.6	11.1	9.0	11.0	8.7	10.8	8.4	11.9	6.4
MDC	1580	2056	1372	1401	1687	1944	1148	2263	741
MDC%	26.7	30.9	25.0	30.5	24.0	29.8	23.2	33.1	17.7

[a]SEM: standard error of measurement.

[b]MDC: minimal detectable change.

[c]TD: tremor dominant.

[d]PIGD: postural instability and gait disorder.

[e]mHY: modified Hoehn and Yahr scale.

Reliability Comparison Between Subgroups

When comparing ICCs between subgroups pairs, 2-tailed Student *t* test for independent groups showed a significant difference between participants with and without tremor (t_{102}=1.897; *P*=.03), between PD with mild and moderate disease severity (t_{102}=1.765; *P*=.04), and between individuals in early and intermediate disease stage (t_{102}=2.817; *P*=.003). Conversely, no significant difference was found between TD and PIGD participants (t_{102}=1.048; *P*=.15).

The analysis of 83.4% CI showed no overlap between interval limits between participants with and without tremor (Figure 1A), in early and intermediate disease stage (Figure 1C), and only a negligible overlap between individuals with mild and moderate disease severity (Figure 1B). Conversely, a degree of overlap of the two 83.4% CIs was observed between TD and PIGD participants (Figure 1D). Details of 83.4% CI limits are shown in Table 2.

XSL•FO
RenderX

Figure 1. ICC and 83.4% CI comparison for each subgroup pair. (A) Participants with reported presence or absence of tremor. (B) Participants with mild and moderate disease severity as indicated by MDS-UPDRS. (C) Participants with earlier (mHY 1 - 2) and intermediate (mHY 2.5 - 3) disease stage. (D) PIGD and TD phenotypes. ICC (3,k) point estimate is indicated by the black dot, CI limits are represented by the vertical bars. ICC: intraclass correlation coefficient; MDS-UPDRS: International Movement Disorders Society Unified Parkinson's Disease Rating Scale; mHY: modified Hoehn and Yahr stage; PIGD: postural instability and gait disorder; TD: tremor dominant.

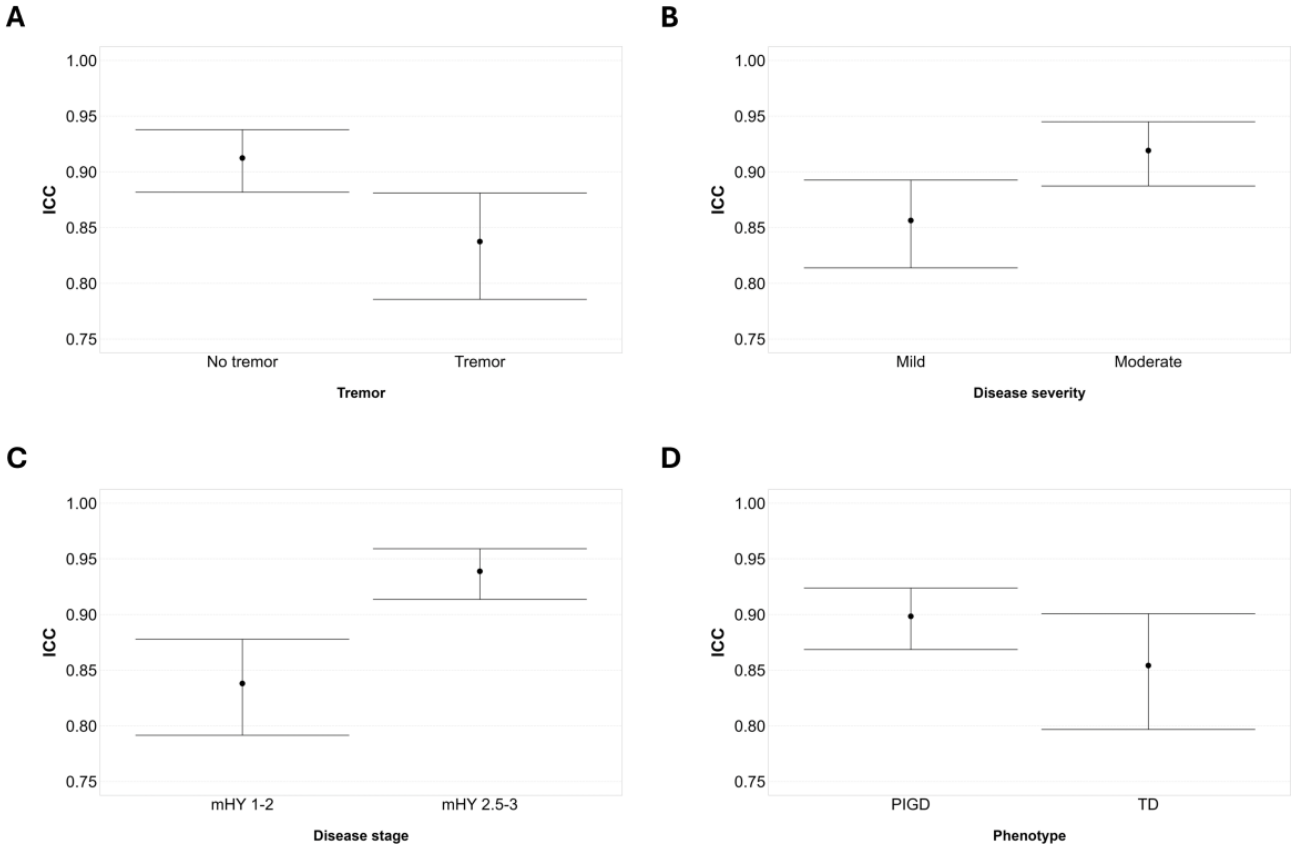

Discussion

Principal Findings

This cross-sectional study aimed at assessing and comparing the reliability of a consumer wrist-worn smartwatch (Garmin Vivosmart 4) in counting avDS in PD in unsupervised, free-living conditions for 5 consecutive days between disease phenotypes, stages, and severity groups.

Overall, our results showed, for the first time, that avDS were acceptably reliable in mild-to-moderate PD irrespective of disease severity, stage, or phenotype. Nevertheless, our results further indicated lower reliability in people with PD with TD phenotype, tremor, lower disease severity, and earlier stage.

Reliability of avDS in the Overall PD Population

We found that the relative reliability of avDS, measured in unsupervised, free-living conditions, by Garmin Vivosmart 4 for 5 consecutive days was within the a priori criteria for acceptability in the overall PD population. We identified only 2 studies investigating the reliability of wearable devices in measuring avDS in people with PD [24,44]. Paul et al [44] reported that 2 consecutive days of monitoring were sufficient to obtain an ICC of >0.9, using a research-grade, ankle-mounted, step counter (Step Activity Monitor) in 92 people with PD. One study from our group investigated the reliability of a wrist-worn consumer device in measuring avDS in PD before. In that study, we found an ICC (3,k) of 0.88 (0.82 - 0.93) for 4 days of monitoring in 56 mild-to-moderate people with PD using Garmin Vivosmart 4 [24]. The present results are consistent with our previous study, but also with other studies investigating the reliability of avDS through wrist-worn wearables in healthy older people [44-47], and in people with various neurological conditions, such as multiple sclerosis [48] and stroke [49]. Regarding absolute reliability, we found that SEM% was below 10% (9.6%) with an MDC of 1580 steps per day (27% of criterion). Only 1 study, from our group, investigated SEM and MDC for avDS in PD [24]. The present results are consistent with our previous study in which we reported an SEM% of 9.2% and an MDC of 1495 steps per day (or 26% of criterion) [24]. Taken together, these results confirm the reliability of avDS measured through Garmin Vivosmart 4 in unsupervised, free-living conditions in mild-to-moderate PD for 5 consecutive days.

Reliability of avDS in People With PD Subgroups

To our knowledge, this is the first study to investigate and compare the reliability of a consumer smartwatch in measuring avDS in different subgroups of PD based on disease phenotype, severity, and stage. We found that in all investigated subgroups (ie, TD vs PIGD, mild vs moderate, tremor vs nontremor, and earlier vs intermediate), ICC values were within the a priori criteria for acceptability (ICC range 0.84 - 0.94). However, a significantly lower ICC was observed in people with PD with tremor, mild disease severity, and earlier disease stage. Moreover, SEM% was below 10% in participants with PIGD

XSL•FO
RenderX

phenotype, moderate disease severity, intermediate disease stage, and without tremor, with an MDC ranging from 1148 to 1687 steps per day (18% - 25% of criterion). Conversely, in individuals with TD phenotype, tremor, mild disease severity, and earlier disease stage, SEM was >10% of criterion and MDC values ranged from 1401 to 2263 steps per day (30% - 33% of the criterion).

MDC, defined as the minimal change that falls outside the measurement error of an instrument, is extremely relevant in study design since it allows to calculate the sample size of studies aiming to assess the effectiveness of interventions [50]. MDC could be also crucial to define the appropriateness and feasibility of using a determinate device to measure a given construct. A prior work from Handlery et al [14] reported an increase in 1250 steps per day following a high-intensity physical activity intervention in PD measured through a research-grade wrist-worn device (Actigraph GT3X). In this study, we found that in individuals with TD phenotype, mild disease severity, and earlier disease stage, MDC was ~2000 steps per day. Although a direct comparison with the metrics reported in the work from Handlery et al [14] could not be performed due to the different devices, we could hypothesize that only large modifications in avDS could be reliably measured through Garmin Vivosmart 4 in the aforementioned PD subgroups. To this end, future studies will be needed to define the minimal clinically important difference for avDS measured through consumer wrist-worn devices and to assess the attainability of avDS modifications sufficiently large to be reliably detected by these devices.

The reduced reliability in participants with tremor is in line with our hypothesis that tremor could reduce the performance of the step-detection algorithm. In fact, tremor could increase the noise-to-signal ratio in the accelerometer signal, and in turn, render step detection more challenging [21]. Indeed, a previous study highlighted that tremor and dyskinesia together contributed to more than 19% of the variation in daily step counts when comparing measurements from waist-worn and wrist-worn devices in 46 people with PD with similar characteristics to those included in this study [21]. In this regard, our study further supports the assumption that tremor could reduce the step-detection performance of wrist-worn devices in PD.

On the other hand, our hypothesis that a reduced reliability might be observed in participants with higher symptom severity and more advanced stages was not supported by our results. Indeed, the reduced reliability of avDS observed in individuals with mild disease severity and earlier disease stage is somehow counterintuitive. In fact, previous evidence highlighted that step count was less accurate in people walking at slower gait speed and with shorter step length in several neurological and musculoskeletal conditions, including PD [20,22,51-54]. Reduced step length and slower walking speed are typical features of Parkinsonian gait, with a higher prevalence along the disease course [2,33]. Moreover, another typical characteristic of walking in PD is the reduced automaticity that leads to a more discontinuous and irregular gait pattern that can further reduce device accuracy in step detection [2,22].

Despite these considerations, our results showed that avDS estimation was more reliable in people with PD with moderate disease severity and intermediate disease stage, compared with individuals with mild disease severity and earlier disease stage. In this regard, it must be considered that reliability is a measure of consistency and reproducibility of measurement and not a measure of accuracy [55]. Therefore, a reduced accuracy could not directly translate into reduced reliability. We could hypothesize that in more advanced stages of PD, the variability of clinical presentation could be lower. In an earlier stage, indeed, symptom heterogeneity, both in terms of motor and nonmotor features, could be extremely high [56-58]. However, this variability could decrease with disease progression since motor symptoms tend to consistently worsen along the disease course and motor features such as gait and balance impairment become increasingly prevalent [59,60]. In addition, the phenotype is dynamic along the disease course and some researchers have proposed that the classification into PIGD/TD evolves over time [61,62]. One study, indeed, reported that over a period of 8 years, approximately 70% of TD individuals transitioned to PIGD, whereas only 4% of PIGD individuals transitioned to TD [61]. Another study reported that 45% of TD participants at baseline had a subtype shift along a 2-year follow-up while 85% of PIGD participants remained as PIGD [62]. This is mirrored in our study cohort, where 32 out of 65 (49%) individuals with mild disease severity were classified as TD, whereas only 7 out of 39 (18%) were in the moderate group. Similarly, 33 out of 68 (49%) participants with earlier disease stage were classified as TD, whereas only 6 out of 36 (17%) were in the intermediate group. Therefore, a regression toward a more uniform motor impairment along the disease course might be considered. Moreover, since we found that tremor could be a relevant factor in reducing avDS reliability, the different prevalence of TD phenotype could also contribute to explaining our results. However, it must be underlined that no study, to our knowledge, systematically compared the heterogeneity of PD features across early, intermediate, and advanced disease stages. Therefore, our hypothesis should be taken with caution and future studies are needed to confirm it.

In conclusion, our findings highlight that, although avDS were reliable across the examined subgroups, clinicians and researchers should consider disease phenotype, stage, and severity when implementing wrist-worn wearables and interpreting mobility data collected through these devices in PD.

Limitations

We acknowledge that this study has some limitations. First, the participants included in our study displayed relatively preserved cognitive functions, due to our exclusion of participants with a Montreal Cognitive Assessment of <21. Additionally, those with more advanced disease stages or requiring walking aids were not included. This potentially limits the generalizability of our findings. However, it should be considered that the sample in our study can be seen as representative of the typical target for interventions using consumer-grade wearable technology. Moreover, including individuals with a disease stage >3, using walking aids, or with more severe cognitive impairments poses significant challenges in the utilization of

XSL•FO
RenderX

consumer technology and is beyond our scope. Nevertheless, future research incorporating PD with lower functional scores, higher disease severity, and more impactful cognitive impairments would be valuable. Furthermore, we used only PD subtyping based on clinical features, yet other classification methods and clustering techniques have been proposed incorporating instrumental and biological data. Future studies are warranted to investigate the reliability of consumer smartwatches in PD subgroups defined using multimodal biomarkers. Finally, we did not account for antiparkinsonian treatment to control tremor. All included participants were under dopaminergic medical treatment and stable medical treatment in the 4 weeks before data collection, and we did not enroll drug-naïve individuals. In addition, we did not account for different drug classes since it would have required a much higher sample size to cover all the possible combinations. Future studies might investigate the metrological features of commercial smartwatches in drug-naïve PD to avoid the influencing factor of dopaminergic treatment. However, since almost all patients with a diagnosis of PD are medically treated, we considered it relevant to assess the measurement properties of commercial devices in a setting more representative of the real-world experience.

Conclusions

In mild-to-moderate PD, avDS measured through a consumer smartwatch in unsupervised, free-living conditions for 5 consecutive days are reliable irrespective of disease phenotype, stage, and severity. Researchers and clinicians who want to implement these instruments should consider that in individuals with TD phenotype, tremor, mild disease severity, and earlier disease stage, reliability could be lower and MDC could be higher. Future studies are needed to define the minimal clinically important difference for avDS measured through consumer wrist-worn devices and to assess the attainability of avDS modifications sufficiently large to be reliably detected by wrist-worn consumer devices. Taken together, we believe that these results could facilitate a broader implementation and an informed application of avDS as an index of ambulatory activity in PD and could be highly relevant to developing monitoring, preventive, educational, and rehabilitation strategies for PD.

Acknowledgments

The authors sincerely thank all people who were involved in the data collection, from study participants to students, who aided and assisted the assessments. This research was performed as part of EB's PhD (Sapienza University of Rome, Italy and Université Grenoble Alpes, France) and funded through a Doctoral Bursary from Sapienza University of Rome (student number 1472455). The research was supported by Sapienza University of Rome as part of the program Sapienza Research Call 2020 (grant AR120172B97AF626) and Research Call 2022 (grant AR222181683A15AA), by the French National Research Agency (ANR) in the framework of the Investissements d'avenir program (ANR-10- AIRT-05 and ANR-15-IDEX-02), and the MIAI Cluster (ANR-23-IACL-0006). This work also forms part of a broader translational and interdisciplinary GaitAlps research program (NV)

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Overlap between the groups of participants with and without self-reported tremor and TD or PIGD phenotypes as well as between the groups with mild or moderate disease severity and early or intermediate disease stage. mHY: modified Hoehn and Yahr scale; PIGD: postural instability and gait disorder; TD: tremor dominant.
[DOCX File, 22 KB - formative_v9i1e63153_app1.docx]

Multimedia Appendix 2
Proof-of-concept experiment to test criterion validity of Garmin Vivosmart 4 in estimating average daily steps in free-living conditions in people with Parkinson disease.
[DOCX File, 26 KB - formative_v9i1e63153_app2.docx]

References

1. MacKinnon CD. Sensorimotor anatomy of gait, balance, and falls. Handb Clin Neurol 2018;159(3–26):3-26. [doi: 10.1016/B978-0-444-63916-5.00001-X] [Medline: 30482322]
2. Mirelman A, Bonato P, Camicioli R, et al. Gait impairments in Parkinson's disease. Lancet Neurol 2019 Jul;18(7):697-708. [doi: 10.1016/S1474-4422(19)30044-4] [Medline: 30975519]
3. Lee M, Noh Y, Youm C, et al. Estimating health-related quality of life based on demographic characteristics, questionnaires, gait ability, and physical fitness in Korean elderly adults. IJERPH 2021 Nov 11;18(22):11816. [doi: 10.3390/ijerph182211816]

XSL•FO
RenderX

4. Creaby MW, Cole MH. Gait characteristics and falls in Parkinson's disease: a systematic review and meta-analysis. Parkinsonism Relat Disord 2018 Dec;57:1-8. [doi: 10.1016/j.parkreldis.2018.07.008] [Medline: 30041848]
5. Bassett DR, Toth LP, LaMunion SR, Crouter SE. Step counting: a review of measurement considerations and health-related applications. Sports Med 2017 Jul;47(7):1303-1315. [doi: 10.1007/s40279-016-0663-1] [Medline: 28005190]
6. Inoue K, Tsugawa Y, Mayeda ER, Ritz B. Association of daily step patterns with mortality in US adults. JAMA Netw Open 2023 Mar 1;6(3):e235174. [doi: 10.1001/jamanetworkopen.2023.5174] [Medline: 36976556]
7. Lee IM, Shiroma EJ, Kamada M, Bassett DR, Matthews CE, Buring JE. Association of step volume and intensity with all-cause mortality in older women. JAMA Intern Med 2019 Aug 1;179(8):1105-1112. [doi: 10.1001/jamainternmed.2019.0899] [Medline: 31141585]
8. Paluch AE, Bajpai S, Bassett DR, et al. Daily steps and all-cause mortality: a meta-analysis of 15 international cohorts. Lancet Public Health 2022 Mar;7(3):e219-e228. [doi: 10.1016/S2468-2667(21)00302-9] [Medline: 35247352]
9. Saint-Maurice PF, Troiano RP, Bassett DR Jr, et al. Association of daily step count and step intensity with mortality among US adults. JAMA 2020 Mar 24;323(12):1151-1160. [doi: 10.1001/jama.2020.1382] [Medline: 32207799]
10. Del Pozo Cruz B, Ahmadi M, Naismith SL, Stamatakis E. Association of daily step count and intensity with incident dementia in 78 430 adults living in the UK. JAMA Neurol 2022 Oct 1;79(10):1059-1063. [doi: 10.1001/jamaneurol.2022.2672] [Medline: 36066874]
11. Del Pozo Cruz B, Ahmadi MN, Lee IM, Stamatakis E. Prospective associations of daily step counts and intensity with cancer and cardiovascular disease incidence and mortality and all-cause mortality. JAMA Intern Med 2022 Nov 1;182(11):1139-1148. [doi: 10.1001/jamainternmed.2022.4000] [Medline: 36094529]
12. Banach M, Lewek J, Surma S, et al. The association between daily step count and all-cause and cardiovascular mortality: a meta-analysis. Eur J Prev Cardiol 2023 Dec 21;30(18):1975-1985. [doi: 10.1093/eurjpc/zwad229] [Medline: 37555441]
13. Skidmore FM, Mackman CA, Pav B, et al. Daily ambulatory activity levels in idiopathic Parkinson disease. J Rehabil Res Dev 2008;45(9):1343-1348. [doi: 10.1682/JRRD.2008.01.0002] [Medline: 19319758]
14. Handlery R, Stewart JC, Pellegrini C, et al. Physical activity in De Novo Parkinson disease: daily step recommendation and effects of treadmill exercise on physical activity. Phys Ther 2021 Oct 1;101(10):34244805. [doi: 10.1093/ptj/pzab174] [Medline: 34244805]
15. Mattison G, Canfell O, Forrester D, et al. The influence of wearables on health care outcomes in chronic disease: systematic review. J Med Internet Res 2022 Jul 1;24(7):e36690. [doi: 10.2196/36690] [Medline: 35776492]
16. Bianchini E, Maetzler W. Chapter five: wearable systems in movement disorders. In: International Review of Movement Disorders: Academic Press; 2023, Vol. 5. [doi: 10.1016/bs.irmvd.2023.04.002]
17. Chevance G, Golaszewski NM, Tipton E, et al. Accuracy and precision of energy expenditure, heart rate, and steps measured by combined-sensing Fitbits against reference measures: systematic review and meta-analysis. JMIR Mhealth Uhealth 2022 Apr 13;10(4):e35626. [doi: 10.2196/35626] [Medline: 35416777]
18. Fuller D, Colwell E, Low J, et al. Reliability and validity of commercially available wearable devices for measuring steps, energy expenditure, and heart rate: systematic review. JMIR Mhealth Uhealth 2020 Sep 8;8(9):e18694. [doi: 10.2196/18694] [Medline: 32897239]
19. Nakagata T, Murakami H, Kawakami R, et al. Step-count outcomes of 13 different activity trackers: results from laboratory and free-living experiments. Gait Posture 2022 Oct;98:24-33. [doi: 10.1016/j.gaitpost.2022.08.004] [Medline: 36030707]
20. Bianchini E, Caliò B, Alborghetti M, et al. Step-counting accuracy of a commercial smartwatch in mild-to-moderate PD patients and effect of spatiotemporal gait parameters, laterality of symptoms, pharmacological state, and clinical variables. Sensors (Basel) 2022 Dec 25;23(1):214. [doi: 10.3390/s23010214] [Medline: 36616812]
21. Kim DW, Hassett LM, Nguy V, Allen NE. A comparison of activity monitor data from devices worn on the wrist and the waist in people with Parkinson's disease. Mov Disord Clin Pract 2019 Nov;6(8):693-699. [doi: 10.1002/mdc3.12850] [Medline: 31745480]
22. Wendel N, Macpherson CE, Webber K, et al. Accuracy of activity trackers in Parkinson disease: should we prescribe them? Phys Ther 2018 Aug 1;98(8):705-714. [doi: 10.1093/ptj/pzy054] [Medline: 29718452]
23. Ginis P, Goris M, De Groef A, et al. Validation of commercial activity trackers in everyday life of people with Parkinson's disease. Sensors (Basel) 2023 Apr 21;23(8):4156. [doi: 10.3390/s23084156] [Medline: 37112496]
24. Bianchini E, Galli S, Alborghetti M, et al. Four days are enough to provide a reliable daily step count in mild to moderate Parkinson's disease through a commercial smartwatch. Sensors (Basel) 2023 Nov 4;23(21):8971. [doi: 10.3390/s23218971] [Medline: 37960670]
25. Outeiro TF, Alcalay RN, Antonini A, et al. Defining the riddle in order to solve it: there is more than one "Parkinson's disease". Mov Disord 2023 Jul;38(7):1127-1142. [doi: 10.1002/mds.29419] [Medline: 37156737]
26. Jankovic J, McDermott M, Carter J, et al. Variable expression of Parkinson's disease: a base-line analysis of the DATATOP cohort. The Parkinson Study Group. Neurology (ECronicon) 1990 Oct;40(10):1529-1534. [doi: 10.1212/wnl.40.10.1529] [Medline: 2215943]
27. Stebbins GT, Goetz CG, Burn DJ, Jankovic J, Khoo TK, Tilley BC. How to identify tremor dominant and postural instability/gait difficulty groups with the movement disorder society unified Parkinson's disease rating scale: comparison

XSL•FO
RenderX

with the unified Parkinson's disease rating scale. Mov Disord 2013 May;28(5):668-670. [doi: 10.1002/mds.25383] [Medline: 23408503]
28. Zetusky WJ, Jankovic J, Pirozzolo FJ. The heterogeneity of Parkinson's disease: clinical and prognostic implications. Neurology (ECronicon) 1985 Apr;35(4):522-526. [doi: 10.1212/wnl.35.4.522] [Medline: 3982637]
29. Cardoso F, Goetz CG, Mestre TA, et al. A statement of the MDS on biological definition, staging, and classification of Parkinson's disease. Mov Disord 2024 Feb;39(2):259-266. [doi: 10.1002/mds.29683] [Medline: 38093469]
30. Höglinger GU, Adler CH, Berg D, et al. A biological classification of Parkinson's disease: the SynNeurGe research diagnostic criteria. Lancet Neurol 2024 Feb;23(2):191-204. [doi: 10.1016/S1474-4422(23)00404-0] [Medline: 38267191]
31. Horsager J, Borghammer P. Brain-first vs. body-first Parkinson's disease: an update on recent evidence. Parkinsonism Relat Disord 2024 May;122:106101. [doi: 10.1016/j.parkreldis.2024.106101] [Medline: 38519273]
32. Goetz CG, Tilley BC, Shaftman SR, et al. Movement Disorder Society - sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS - UPDRS): scale presentation and clinimetric testing results. Mov Disord 2008 Nov 15;23(15):2129-2170. [doi: 10.1002/mds.22340]
33. Schlachetzki JCM, Barth J, Marxreiter F, et al. Wearable sensors objectively measure gait parameters in Parkinson's disease. PLoS ONE 2017;12(10):e0183989. [doi: 10.1371/journal.pone.0183989] [Medline: 29020012]
34. Postuma RB, Berg D, Stern M, et al. MDS clinical diagnostic criteria for Parkinson's disease. Mov Disord 2015 Oct;30(12):1591-1601. [doi: 10.1002/mds.26424]
35. Goetz CG, Poewe W, Rascol O, et al. Movement Disorder Society Task Force report on the Hoehn and Yahr staging scale: status and recommendations. Mov Disord 2004 Sep;19(9):1020-1028. [doi: 10.1002/mds.20213] [Medline: 15372591]
36. Nasreddine ZS, Phillips NA, Bédirian V, et al. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. J Am Geriatr Soc 2005 Apr;53(4):695-699. [doi: 10.1111/j.1532-5415.2005.53221.x] [Medline: 15817019]
37. Jost ST, Kaldenbach MA, Antonini A, et al. Levodopa dose equivalency in Parkinson's disease: updated systematic review and proposals. Mov Disord 2023 Jul;38(7):1236-1252. [doi: 10.1002/mds.29410] [Medline: 37147135]
38. Martínez-Martín P, Rodríguez-Blázquez C, et al. Parkinson's disease severity levels and MDS-Unified Parkinson's Disease Rating Scale. Parkinsonism Relat Disord 2015 Jan;21(1):50-54. [doi: 10.1016/j.parkreldis.2014.10.026] [Medline: 25466406]
39. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016 Jun;15(2):155-163. [doi: 10.1016/j.jcm.2016.02.012] [Medline: 27330520]
40. Knol MJ, Pestman WR, Grobbee DE. The (mis)use of overlap of confidence intervals to assess effect modification. Eur J Epidemiol 2011 Apr;26(4):253-254. [doi: 10.1007/s10654-011-9563-8] [Medline: 21424218]
41. Austin PC, Hux JE. A brief note on overlapping confidence intervals. J Vasc Surg 2002 Jul;36(1):194-195. [doi: 10.1067/mva.2002.125015] [Medline: 12096281]
42. Cho J, Seo DM, Uh Y. Clinical application of overlapping confidence intervals for monitoring changes in serial clinical chemistry test results. Ann Lab Med 2020 May;40(3):201-208. [doi: 10.3343/alm.2020.40.3.201] [Medline: 31858759]
43. Balaguier R, Madeleine P, Vuillerme N. Intra-session absolute and relative reliability of pressure pain thresholds in the low back region of vine-workers: effect of the number of trials. BMC Musculoskelet Disord 2016 Aug 18;17(1):350. [doi: 10.1186/s12891-016-1212-7] [Medline: 27538914]
44. Paul SS, Ellis TD, Dibble LE, et al. Obtaining reliable estimates of ambulatory physical activity in people with Parkinson's disease. J Parkinsons Dis 2016 May 5;6(2):301-305. [doi: 10.3233/JPD-160791] [Medline: 27164042]
45. Yao J, Tan CS, Lim N, Tan J, Chen C, Müller-Riemenschneider F. Number of daily measurements needed to estimate habitual step count levels using wrist-worn trackers and smartphones in 212,048 adults. Sci Rep 2021 May 5;11(1):9633. [doi: 10.1038/s41598-021-89141-3] [Medline: 33953288]
46. Clemes SA, Griffiths PL. How many days of pedometer monitoring predict monthly ambulatory activity in adults? Med Sci Sports Exerc 2008 Sep;40(9):1589-1595. [doi: 10.1249/MSS.0b013e318177eb96] [Medline: 18685533]
47. Sigmundová D, Vašíčková J, Stelzer J, Repka E. The influence of monitoring interval on data measurement: an analysis of step counts of university students. Int J Environ Res Public Health 2013 Jan 28;10(2):515-527. [doi: 10.3390/ijerph10020515] [Medline: 23358235]
48. Norris M, Anderson R, Motl RW, Hayes S, Coote S. Minimum number of days required for a reliable estimate of daily step count and energy expenditure, in people with MS who walk unaided. Gait Posture 2017 Mar;53:201-206. [doi: 10.1016/j.gaitpost.2017.02.005] [Medline: 28199925]
49. Fini NA, Burge AT, Bernhardt J, Holland AE. Two days of measurement provides reliable estimates of physical activity poststroke: an observational study. Arch Phys Med Rehabil 2019 May;100(5):883-890. [doi: 10.1016/j.apmr.2018.10.006] [Medline: 31030730]
50. Lexell JE, Downham DY. How to assess the reliability of measurements in rehabilitation. Am J Phys Med Rehabil 2005 Sep;84(9):719-723. [doi: 10.1097/01.phm.0000176452.17771.20] [Medline: 16141752]
51. Treacy D, Hassett L, Schurr K, Chagpar S, Paul SS, Sherrington C. Validity of different activity monitors to count steps in an inpatient rehabilitation setting. Phys Ther 2017 May 1;97(5):581-588. [doi: 10.1093/ptj/pzx010] [Medline: 28339904]

XSL•FO
RenderX

52. Fokkema T, Kooiman TJM, Krijnen WP, Van Der Schans CP, De Groot M. Reliability and validity of ten consumer activity trackers depend on walking speed. Med Sci Sports Exerc 2017 Apr;49(4):793-800. [doi: 10.1249/MSS.0000000000001146] [Medline: 28319983]
53. Chow JJ, Thom JM, Wewege MA, Ward RE, Parmenter BJ. Accuracy of step count measured by physical activity monitors: the effect of gait speed and anatomical placement site. Gait Posture 2017 Sep;57:199-203. [doi: 10.1016/j.gaitpost.2017.06.012] [Medline: 28666177]
54. Lamont RM, Daniel HL, Payne CL, Brauer SG. Accuracy of wearable physical activity trackers in people with Parkinson's disease. Gait Posture 2018 Jun;63:104-108. [doi: 10.1016/j.gaitpost.2018.04.034] [Medline: 29729611]
55. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Qual Life Res 2010 May;19(4):539-549. [doi: 10.1007/s11136-010-9606-8] [Medline: 20169472]
56. Erro R, Vitale C, Amboni M, et al. The heterogeneity of early Parkinson's disease: a cluster analysis on newly diagnosed untreated patients. PLoS ONE 2013;8(8):e70244. [doi: 10.1371/journal.pone.0070244] [Medline: 23936396]
57. Wüllner U, Borghammer P, Choe CU, et al. The heterogeneity of Parkinson's disease. J Neural Transm (Vienna) 2023 Jun;130(6):827-838. [doi: 10.1007/s00702-023-02635-4] [Medline: 37169935]
58. Berg D, Borghammer P, Fereshtehnejad SM, et al. Prodromal Parkinson disease subtypes—key to understanding heterogeneity. Nat Rev Neurol 2021 Jun;17(6):349-361. [doi: 10.1038/s41582-021-00486-9] [Medline: 33879872]
59. Fasano A, Fung VSC, Lopiano L, et al. Characterizing advanced Parkinson's disease: OBSERVE-PD observational study results of 2615 patients. BMC Neurol 2019 Apr 2;19(1):50. [doi: 10.1186/s12883-019-1276-8] [Medline: 30940119]
60. Stefani A, Tessitore A, Tambasco N, et al. Criteria for identification of advanced Parkinson's disease: the results of the Italian subgroup of OBSERVE-PD observational study. BMC Neurol 2022 Jan 28;22(1):41. [doi: 10.1186/s12883-022-02554-z] [Medline: 35090406]
61. Alves G, Larsen JP, Emre M, Wentzel-Larsen T, Aarsland D. Changes in motor subtype and risk for incident dementia in Parkinson's disease. Mov Disord 2006 Aug;21(8):1123-1130. [doi: 10.1002/mds.20897] [Medline: 16637023]
62. Che N, Ou R, Li C, et al. Plasma GFAP as a prognostic biomarker of motor subtype in early Parkinson's disease. NPJ Parkinsons Dis 2024 Mar 1;10(1):48. [doi: 10.1038/s41531-024-00664-8] [Medline: 38429295]

Abbreviations

avDS: average daily steps
ICC: intraclass correlation coefficient
MDC: minimal detectable change
MDS-UPDRS: International Movement Disorders Society Unified Parkinson's Disease Rating Scale
mHY: modified Hoehn and Yahr scale
PD: Parkinson disease
PIGD: postural instability and gait disorder
SEM: Standard error of measurement
TD: tremor dominant

Edited by A Mavragani; submitted 16.06.24; peer-reviewed by E Rainaldi, J Ruokolainen; revised version received 03.02.25; accepted 21.02.25; published 18.03.25.

Please cite as:
Bianchini E, Rinaldi D, De Carolis L, Galli S, Alborghetti M, Hansen C, Suppa A, Salvetti M, Pontieri FE, Vuillerme N
Reliability of Average Daily Steps Measured Through a Consumer Smartwatch in Parkinson Disease Phenotypes, Stages, and Severities: Cross-Sectional Study
JMIR Form Res 2025;9:e63153
URL: https://formative.jmir.org/2025/1/e63153
doi:10.2196/63153

© Edoardo Bianchini, Domiziana Rinaldi, Lanfranco De Carolis, Silvia Galli, Marika Alborghetti, Clint Hansen, Antonio Suppa, Marco Salvetti, Francesco Ernesto Pontieri, Nicolas Vuillerme. Originally published in JMIR Formative Research (https://formative.jmir.org), 18.3.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Patient and Provider Perspectives of a Web-Based Intervention to Support Symptom Management After Radioactive Iodine Treatment for Differentiated Thyroid Cancer: Qualitative Study

Alaina L Carr[1], PhD; Angela M Jenkins[1,2], BS; Jacqueline Jonklaas[3], MD, MPH, PhD; Kate Gabriel[4], BS; Kristen E Miller[5], DrPH, MSPH, MSL, CPPS; Kristi D Graves[1], PhD

[1]Department of Oncology, Lombardi Comprehensive Cancer Center, Georgetown University, Washington, DC, United States
[2]Department of Natural Sciences, Bowie State University College of Arts and Science, Bowie, MD, United States
[3]Division of Endocrinology, Georgetown University, Washington, DC, United States
[4]Department of Health Systems Administration, Georgetown University School of Health, Washington, DC, United States
[5]MedStar Health Research Institute, National Center for Human Factors in Healthcare, Washington, DC, United States

Corresponding Author:
Alaina L Carr, PhD
Department of Oncology, Lombardi Comprehensive Cancer Center, Georgetown University, Washington, DC, United States

Abstract

Background: Patients diagnosed with differentiated thyroid cancer (DTC) who receive radioactive iodine (RAI) treatment experience acute, medium, and late treatment effects. The timing and severity of these effects vary by individual; common posttreatment effects include dry mouth, salivary gland swelling, dry eyes, and nose bleeds. The nature of symptoms that patients experience after RAI treatment can significantly and negatively impact health-related quality of life. Adequate information during the postprimary treatment phase remains an unmet need among the population of patients diagnosed with DTC.

Objective: This qualitative study aimed to identify and understand self-management strategies for RAI-specific symptom burden from the perspectives of patients and stakeholders (cancer care providers and patient advocates). An additional aim included assessing features and functionalities desirable in the development of a web-based intervention to engage patients in their self-management and thyroid cancer survivorship care.

Methods: Following the Social Cognitive Theory framework and person-based principles, we conducted six focus groups with 22 patients diagnosed with DTC who completed RAI treatment and individual interviews with 12 stakeholders in DTC care. The interviews focused on participants' perspectives on current self-management strategies and mockups of a symptom management web-based intervention. Before focus groups and interviews, participants completed a demographics survey. Focus group discussions and interviews were transcribed and coded using content analysis. Interrater reliability was satisfactory (α=.88).

Results: A total of 34 individuals (patients and stakeholders) participated in the study; the mean age was 45 (SD 13.4) and 45.3 (SD 13) years, respectively. Three domains emerged from qualitative interviews: (1) difficult-to-manage RAI symptoms: short, medium, and late treatment effects; (2) key intervention structure and content feedback on mockups; and (3) intervention content to promote RAI symptom management and survivorship care. Focus group participants identified the most prevalent RAI symptoms that were difficult to manage as: dry mouth (11/22, 50%), salivary gland swelling (8/22, 36%), and changes in taste (12/22, 55%). Feedback elicited from both groups found education and symptom management mockup videos to be helpful in patient self-management of RAI symptoms, whereas patients and stakeholders provided mixed feedback on the benefits of a draft frequently asked questions page. Across focus groups and stakeholder interviews, nutrition-based symptom management strategies, communication with family members, and practical survivorship follow-up information emerged as helpful content to include in a future web-based supportive care intervention.

Conclusions: Results suggest education and symptom management videos can empower patients with DTC to self-manage mild to moderate RAI symptoms on a web-based platform. Findings emphasized the need for additional information for patients related to ongoing care following RAI treatment including social support and thyroid cancer surveillance. The findings provide insights for theoretically informed interventions and recommendations for refinements in thyroid cancer survivorship from patient and provider perspectives.

(JMIR Form Res 2025;9:e60588) doi:10.2196/60588

XSL•FO
RenderX

KEYWORDS

iodine radioisotopes; person-based approach; self-management; Social Cognitive Theory; survivorship; symptom burden; thyroid neoplasms; web-based intervention; radioactive iodine treatment; radiotherapy; thyroid cancer; qualitative; quality of life; survivorship care; supportive care; patient with cancer; QoL; cancer; carcinoma; malignancy; tumor; malignant; oncology; neoplasm; benign; neoplasia; thyroid

Introduction

Background

Thyroid cancer is often mislabeled as the "good cancer" [1-3] due to the excellent 5-year survival rates compared to other cancers [4]. Differentiated thyroid cancer (DTC) is the most common endocrine malignancy among the estimated 44,020 new cases of thyroid cancer in the United States annually [4]. DTC involves a spectrum of disease presentations characterized by different levels of risk of recurrence. For patients with a high risk of recurrence, the standard of care involves total thyroidectomy followed by radioactive iodine (RAI) treatment and thyroid hormone therapy [5-7]. For patients with intermediate-risk disease, clinical practice guidelines from the American Thyroid Association recommend thyroidectomy followed by consideration of RAI treatment in consultation with their medical team [6]. Approximately, 72% of these intermediate and high-risk patients receive RAI and thyroid hormone therapy as standard of care [5]. Despite the favorable 5-year survival rates, complications from RAI treatment are common, impacting health-related quality of life (HRQOL) [1,8-10]. The "good cancer" label given to thyroid cancer frustrates patients, particularly when they experience side effects as a result of RAI treatment [1].

The likelihood of acute and chronic side effects following RAI varies based on RAI dose administration; these treatment impacts are particularly evident in patients who receive higher doses of ≥150 mCi [7,11,12]. While trends toward lower dose activity are recommended, therapeutic dose recommendations vary across professional medical societies and practice patterns [6,11-14]. Short-, medium-, and long-term complications from RAI treatment include salivary dysfunction, dysphagia, dysphonia, sialadenitis, appetite changes, nausea, vomiting, xerostomia, epiphora, and secondary malignancies [1,15-17], with the onset of certain symptoms appearing as late as 1-year post-RAI treatment [12,16]. In a cross-sectional study of 69 patients diagnosed with DTC and treated with RAI, lower HRQOL was significantly associated with higher symptom burden, distress, and lower self-efficacy [18]. Furthermore, patients with thyroid cancer report lower HRQOL compared to patients with other types of cancer with worse survival [10]. Additional care challenges reported by patients with DTC are unmet information needs about follow-up care, psychosocial support, and coordination of care [17,19]. Research is needed to understand, address, and mitigate RAI symptom burden among patients with DTC.

Mobile health (mHealth) technologies offer an accessible approach to symptom management interventions [20] and are increasingly used in the delivery of cancer care. mHealth refers to mobile and wireless technologies such as smartphones, computers, and tablets to support patient care, education, and communication in health care [21,22]. mHealth technologies have been implemented in the prevention, promotion, treatment, and maintenance of health and health care for patients with breast, lung, leukemia, and prostate cancers [21-25]. mHealth interventions empower patients with cancer with direct access to educational and support tools to improve self-management of symptoms, including fatigue, pain, distress, and cognitive impairment [21,23,26,27]. Core elements of mHealth interventions for chronic disease management include evidence-based education, monitoring and tracking symptoms, tailored feedback on symptoms, self-management training to cope with the psychological and physical aspects of treatment, and communication with providers [21,28,29]. Applying evidence-based theories to inform the design characteristics of digital technology interventions during the development phase of the research is one way to engage patients, foster patient-provider communication, and improve patient-centered health outcomes [20-22,28].

Theoretically Driven mHealth Intervention Development

Social Cognitive Theory (SCT) [30,31], a longstanding theory-based approach to health behavior change, coupled with guiding principles from the person-based approach [20,32] to mHealth intervention development, offers a basis for core components of symptom management interventions. Behavioral interventions based on tenets of SCT are evaluated through patient knowledge of health risks and benefits, self-efficacy or confidence related to symptom management, outcome expectations about the costs and benefits of performing a particular behavior, and self-regulation impacting patient HRQOL [30,31]. Prior examples of the use of SCT in cancer-related health behavior change interventions include physical activity behavior change interventions [33-36] or interventions to promote HRQOL in breast cancer populations [37]. The application of SCT to promote self-management of RAI symptoms among patients with DTC is understudied, and further research is needed to identify core components of health-related interventions. The person-based approach to mHealth interventions yields insights into the behavioral components relevant to the populations intended to use the intervention [32]. As patients with DTC post-RAI treatment are commonly seen in ambulatory care settings, and their follow-up care is managed by specialists [17], a digital platform offers patients accessibility delivered at the point of need. Furthermore, involving target patient users and members of the health care team early in the intervention design process can identify key components [20,21,32] of symptom management interventions among this population.

We seek to develop and refine tailored mHealth symptom management interventions for patients with DTC who completed RAI. Thus, we conducted a qualitative study to identify and

XSL•FO
RenderX

understand potential self-management strategies for RAI-specific symptoms from the perspectives of patients with DTC who completed RAI treatment and health care providers. In addition, we wanted to elicit feedback on core content and features in the formative development of a web-based intervention to promote patient self-management of RAI symptoms.

Methods

Study Design

We followed the COREQ (Consolidated Criteria for Reporting Qualitative Research) to report the qualitative findings [38]. From August 2022 to January 2024, we conducted a qualitative study using focus groups with patient participants and individual stakeholder interviews at Georgetown University Medical Center.

Ethical Considerations

Ethics approval was performed in line with the principles of the Declaration of Helsinki. Approval was granted by the Georgetown University Institutional Review Board (#00005285). All focus group participants and stakeholders provided informed consent. Study data were deidentified prior to data analysis. Each participant received Amazon gift cards worth US $50 for their participation.

Participant Selection and Setting

Adult patients were recruited from MedStar Georgetown University Hospital to take part in the qualitative study. Treating endocrinologists referred patients to the study. Patients were invited to participate if they were diagnosed with DTC, completed RAI treatment within the past three years, and had access to an electronic device with an internet connection. We excluded non-English-speaking patients and those with psychiatric or cognitive impairments that would inhibit meaningful consent.

For the stakeholder interviews, health care specialists were recruited from four hospitals in the Washington metropolitan area. Health care specialists were referred to the study by colleagues or invited to participate directly by study staff through an introductory email. Stakeholder inclusion criteria were self-identification as a full-time specialist treating patients with DTC post-RAI treatment or allied health professionals and patient advocates affiliated with ThyCa: Thyroid Cancer Survivors' Association, Inc.

We used a purposive sampling strategy for patient age, gender, cancer type, and health subspecialists from academic and community hospital settings in the area. We recruited 22 patients diagnosed with DTC who completed RAI treatment and 12 stakeholders from local health care facilities and ThyCa groups in the DC area.

Procedures and Data Collection

Following Institutional Review Board approval, study staff approached potential participants with an introductory study email followed by a phone call. Study staff completed informed consent through a 15-minute telephone or Health Insurance Portability and Accountability Act-compliant video conference and obtained electronic informed consent from focus group participants and stakeholders. Focus group participants and stakeholders completed a brief survey through REDCap (Research Electronic Data Capture; Vanderbilt University) [39] to obtain sociodemographic, clinical characteristics, and occupation information. The interviews were conducted over a Health Insurance Portability and Accountability Act-compliant videoconference lasting approximately two hours for focus group participants and one hour for individual stakeholder interviews.

Two authors (ALC: psychologist, female; and KDG: behavioral scientist, female) trained in qualitative research methods developed moderator and semistructured interview guides (Multimedia Appendices 1 and 2), which were informed by tenets of SCT [30] and person-based design principles [32]. Both guides included items to elicit participants' perspectives about RAI symptoms that were challenging to manage, identify content that would be helpful in a symptom management website, and obtain feedback on the formatting and content of static images or mockups (Multimedia Appendix 3) of a prospective symptom management website. Participants were asked to provide feedback on the mockups after eliciting input on prospective content to minimize participant response bias. One author (ALC) conducted audio-recorded interviews for focus group and stakeholder interviews. A second member of the research team trained in qualitative research (KG: research assistant, female; Gautham Pillai: research assistant, male; and AMJ: research assistant, female) took field notes during the focus group and stakeholder interviews. All focus group and stakeholder interviews were transcribed verbatim and deidentified before data analysis. All focus group participants and stakeholders received a summary of the findings.

Data Analysis

We used thematic content analysis to analyze focus group and stakeholder interview transcripts using a well-established 6-step framework [40]. Three authors (ALC, AMJ, and KG) independently reviewed transcripts to familiarize themselves with the data and identify initial codes using ATLAS.ti, a computer-assisted qualitative data analysis software (ATLAS.ti Scientific Software Development GmbH). The authors (ALC, AMJ, and KG) discussed initial codes, shared definitions of each code, and grouped the codes into broad, meaningful categories. The coders further organized the codes into major domains and subthemes and examined discrepant data until resolution. Analysis was systematically and iteratively applied across all transcripts. Thematic saturation [40] occurred after no new information was identified from transcripts. We calculated Krippendorf Cu-α coefficients [41,42] to evaluate coding precision by the three team members for the three domains across transcripts. The final Krippendorf Cu-α=.88, indicating acceptable agreement [43]. Extracted quotations exemplified major domains and subthemes. Descriptive statistics were used to characterize focus group and stakeholder demographic, clinical, and occupation characteristics.

XSL•FO
RenderX

Results

Participants

We approached 93 patients and stakeholders to enroll in the qualitative study; 24 patients and 12 stakeholders consented, 2 patients were unable to participate in focus groups due to scheduling conflicts, and 34 interviews were analyzed (Figure 1). A total of 29 patients were excluded due to ineligibility (n=2), decline (n=6), and no response to the introductory study email (n=21). A total of 28 stakeholders were excluded from the study due to ineligibility (n=5), decline (n=5), and no response to the introductory study email (n=18).

Figure 1. Qualitative study flowchart for focus group participants and stakeholders. Participants, recruited from clinical settings, were patients diagnosed with thyroid cancer. Stakeholders, recruited through professional networks, were providers involved in the clinical care of patients diagnosed with thyroid cancer. Data were collected between August 2022 and January 2024.

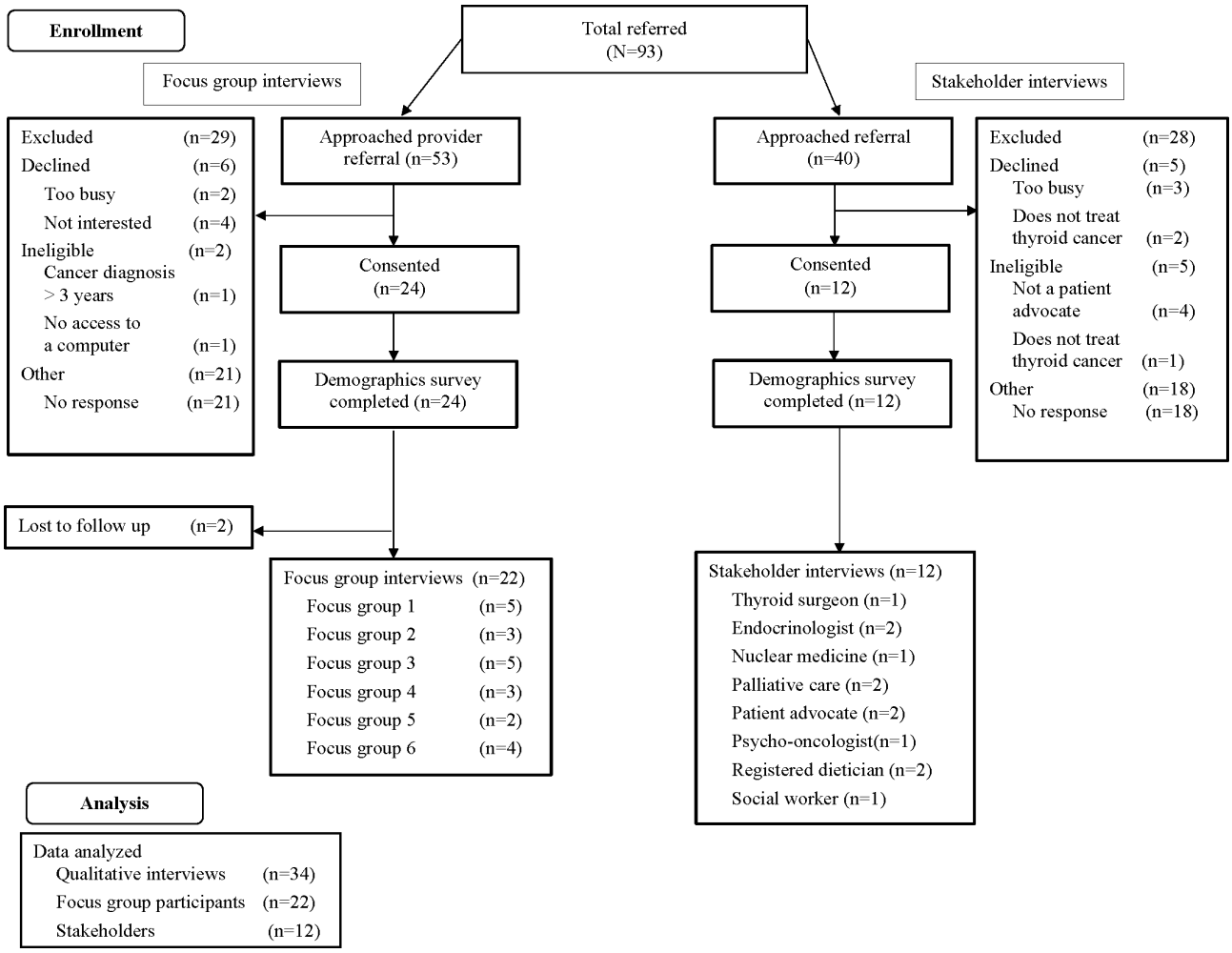

The age (in years) of focus group participants and stakeholders was 45.0 (SD 13.4; range 23.6 - 72.5) and 45.2 (13.0; range 32.3 - 75.7), respectively. The majority of focus group participants and stakeholders identified as female (15/22, 68%; 10/12, 83%), non-Hispanic or Latino (20/22, 91%; 11/12, 92%), and White (14/22, 64%; 8/12, 67%), respectively. In terms of clinical characteristics, the majority of focus group participants were diagnosed with papillary thyroid cancer (18/22, 82%). The average time since diagnosis (in years) was 1.5 (SD 1.0; range 0.03 - 2.9), and the time since treatment (in years) was 1.1 (SD 0.9; range 0.1 - 2.8). There was a total of six focus groups, with a range of 2 - 5 participants in each group. Stakeholders consisted of thyroid surgeons (n=1), endocrinologists (n=2), nuclear medicine specialists (n=1), palliative care providers (n=2), patient advocates (n=2), psycho-oncologists (n=1), registered oncology dietitians (n=2), and social workers (n=1). Table 1 presents sample characteristics of focus group participants and stakeholders and Table 2 presents clinical characteristics of focus group participants.

XSL•FO
RenderX

Table . Sample characteristics of focus group participants and stakeholders (N=34)[a].

Characteristics	Focus group participants (n=22)	Stakeholders (n=12)
Age (years)		
Mean (SD)	45.0 (13.4)	45.2 (13.0)
Range	23.6-72.5	32.3-75.7
Female, n (%)	15 (68)	10 (83)
Married or partnered, n (%)	14 (64)	—[b]
Hispanic or Latino, n (%)	2 (9)	1 (8)
Race, n (%)		
Asian	3 (14)	3 (25)
Black or African American	2 (9)	1 (8)
More than one race/ other	3 (14)	—
White	14 (64)	8 (67)
Education, n (%)		
High-school and below	3 (14)	—
College and above	19 (86)	—
Employment status, n (%)		
Full-time	17 (77)	—
Part-time	3 (14)	—
Student	2 (9)	—
Retired	2 (9)	—
Heath Insurance (Yes), n (%)	22 (100)	—

[a]Participants, recruited from clinical settings, were patients diagnosed with thyroid cancer. Stakeholders, recruited through professional networks, were providers involved in the clinical care of patients diagnosed with thyroid cancer. Data were collected between August 2022 and January 2024.
[b]Not applicable.

XSL•FO
RenderX

Table . Differentiated thyroid cancer clinical characteristics of focus group participants (n=22)[a].

Clinical characteristics	Focus group participants
Time since diagnosis (years)	
Mean (SD)	1.5 (0.9)
Range	0.03 - 2.9
Time since treatment (years)	
Mean (SD)	1.1 (0.9)
Range	0.10 - 2.8
Histology[b], n (%)	
Papillary	18 (82)
Follicular	2 (9)
Do not know	4 (18)
Stage[b], n (%)	
I	6 (27)
II	2 (9)
III	1 (5)
IVA	1 (5)
Do not know	11 (50)

[a]Participants recruited from clinical settings were patients diagnosed with thyroid cancer. Data was collected between August 2022 and January 2024.
[b]The value given are frequencies.

Qualitative Results

Qualitative results yielded three domains: (1) difficult-to-manage RAI symptoms: short, medium, and late treatment effects; (2) key intervention structure and content feedback on mockups; and (3) intervention content to promote RAI symptom management and survivorship care. Tables 3 and 4 display exemplar quotes with patient identifiers to depict domains and subthemes from focus group participant and stakeholder interviews.

XSL•FO
RenderX

Table . Summary and exemplar quotes from focus group participants and stakeholders related to domain 2 key intervention structure and content feedback on mockups with an SCT[a] framework.

Mockup	Exemplar quotes	SCT[a] component
Home page Description: benefits and consequences home page in its current format; recommended changes to enhance patient understanding of website component	• "In those websites you always see younger people. I think it's nice to include more diversity. If you are going to include more pictures have other ones where you see people that are older depicted." [Focus Group 5, PID 2] • "On the landing page ... You can have a tab on statistics ... Are you experiencing side effects? Side effects management tab here. Social resources here. Or supportive care resources. You could have a separate nutrition tab." [PID 10, Registered Oncology Dietitian]	Knowledge and outcome expectations
Education page RAI[b] salivary symptoms Description: benefits of education page on RAI symptom self-management; recommended changes to enhance engagement in website component	• "[It] Feels friendly discussing something that is very distressing. There is an empowerment element to this of, 'Hey, we get you. You're experiencing these things. Please watch out for these and call your doctor.' This page is magical. I wish this site existed already." [Focus Group 4, PID 3] • "It's a good general informational landing spot particularly things that they should definitely not try home remedies for. Like the bottom half [Redness over salivary glands, Pain doesn't go away with antiinflammatory medications, worsening swelling or salivary glands-signs of infection, Significant weight loss due to difficulty eating and swallowing]." [PID 8, Thyroid Surgeon]	Knowledge
Symptom management video Description: benefits of RAI symptom self-management video on self-efficacy; consequences of self-management videos on self-efficacy; recommended changes to enhance patient engagement in self-management symptom videos	• "Having this particular type of video on how to massage salivary glands, I would be interested in being able to go to the site to see precisely how I can do this and have a video to demonstrate just how it can be performed. So I'm interested to tap the play button. " [Focus Group 4, PID 1] • "There might be a little discussion of who might need this because if I just had RAI and I'm out of isolation, am I supposed to do this irrespective of symptoms ?" [PID 4, Patient Advocate]	Self-efficacy and self-regulation

XSL•FO
RenderX

Mockup	Exemplar quotes	SCT[a] component
FAQ[c] page Description: need for relevant and concise questions related to the post-RAI treatment timeframe. Stakeholders suggested providing answers to the questions within the web-based intervention to reduce the potential for increased demands on provider workflows.	• "Some of the things that I would not think about are dental appointments. I didn't think about that until recently, after I had my sialendoscopy. But other than that, all the other questions are relevant. You don't really think about it before or during the RAI treatment, because you didn't know that was a concern." [Focus Group 2, PID 2] • "Is there a value to having answers for these [questions]? I would wonder if I'm interested in the first question, and I want to ask my provider that, I also want to know the answer now. Or I want to get some guidance about that now. I don't want to wait until my clinic appointment in 2 weeks to get the answer. That was what, I wonder, is there a role for that? But with some answers. " [PID 6, Palliative Care Provider]	Self-regulation and outcome expectations

[a]SCT: Social Cognitive Theory.

[b]RAI: radioactive iodine.

[c]FAQ: frequently asked question.

XSL•FO
RenderX

Table . Subthemes and exemplar quotes related to domain 3 intervention content to promote RAI[a] symptom management and survivorship care needs within the SCT[b] framework.

Subthemes	Exemplar Quotes	SCT Construct
Nutrition description: information on nutrition strategies for RAI symptom management strategies for salivary symptoms. Topics included details about foods that stimulate salivary glands and help manage changes in taste.	• "It could be a video about what foods might be helpful [to] stimulate your salivary glands. And then, in case you lost taste, recommendations for you to still taste something, and being able to eat without affecting the health of your mouth." [Focus Group 5, PID 2] • "If there were tips presented earlier about managing dry mouth like ensuring adequate hydration, ensuring fluid or moisture, rich food intake, that is something that could be helpful." [PID 10 Registered oncology dietitian]	Self-efficacy and self-regulation
Communication with family description: information and strategies to help guide patient-led communication with families and children, about RAI symptoms. Topics included practical tips on how to talk about the social and emotional impact of DTC[c] diagnosis and RAI symptoms with children and family members.	• "I was thinking [what] would also be nice [is] to have some type of resource center for family and friends on how to best support someone with thyroid cancer ... [It is]a nice way of how best to support someone with thyroid cancer." [Focus Group 4, PID 1] • " How to speak to them, when they're young and explain the survival rate is very good. And when it comes to cancers, it's not [a] death sentence, because you hear the cancer part, but you know it is survivable." [Focus Group 6, PID 3] • "That's important, especially, if your children are young, [and] they hear the diagnosis of cancer. How you can communicate that to the children, if there is information on that, I think that would be helpful for patients." [PID 3, Nuclear Medicine Provider] • " I worry about broad strokes because what do you tell a 6-year-old versus the 10-year-old versus a 13-year-old and not all 13-year-olds are the same." [PID 8, Thyroid Surgeon] • " I think that's [communication] also intimidating for people. Some people have a better general sense of what this treatment is about, and how it can or can't impact others. And other people are less familiar and aware or able to articulate what they know about their treatment and how it can impact others. And I think that can just be hard and an additional burden on the patient." [PID 11, Psycho-oncologist]	Self-efficacy, and self-regulation

XSL•FO
RenderX

Subthemes	Exemplar Quotes	SCT Construct
Cancer survivorship follow-up care description: discussed the need for a general overview of lifelong follow-up with health care providers, including, which health care team members to consult with for RAI symptom management and follow-up tests (blood, ultrasound, or whole body radioiodine scans).	• "Are you going back to the nuclear medicine team? Are you [going to] your endo? Who's supposed to be quarterbacking all of your treatment? The right person? Is it whoever did your surgery? Who's the right person depending on what the symptoms are?" [Focus Group 1, PID 5] • "What would be helpful is how to explain to other providers that you've had this treatment, and these are some of the possible side effects ... I think having something like a quick reference sheet where they [patients] can talk to their doctor." [Focus Group 6, PID 2] • "Patients should know that they basically have this diagnosis of cancer, and that they need lifelong follow up. Follow up with their endocrinologists, follow up with their primary care, or whoever is managing their dose which they get for their thyroid hormone. Or to make sure that they that they do not get recurrence in the future." [PID 3, Nuclear Medicine Provider]	Knowledge, self-regulation, and outcome expectations

[a]RAI: radioactive iodine.

[b]SCT: Social Cognitive Theory.

[c]DTC: differentiated thyroid cancer.

Domain 1: Difficult-to-Manage RAI Symptoms: Short, Medium, and Late Treatment Effects

As listed in Table 5, we indicated the percentage of focus group participants and stakeholders who commented on difficult-to-manage RAI symptoms. In our sample, approximately half of our focus group participants (11/22) endorsed symptoms of dry mouth and changes in taste as challenging symptoms to manage, followed by salivary gland swelling (8/22, 36%), dry eyes (7/22, 31%), and dry nose (3/22, 13%). Stakeholders endorsed a similar prevalence of difficult-to-manage RAI symptoms experienced by patients in their clinical care, including dry mouth (6/12, 50%), changes in taste (5/12, 41%), salivary gland swelling (4/12, 33%), swallowing difficulties (3/12, 25%), and dry eyes (3/12, 25%). Focus groups highlighted that among patients with DTC post-RAI treatment, the effectiveness of their current self-management strategies varies across individuals. Participants emphasized the need for personalized approaches to manage the short and late effects of RAI treatment. Many stakeholders emphasized the need for patient referrals and collaboration with multidisciplinary care teams to enhance overall patient care and address severe RAI symptom management.

XSL•FO
RenderX

Table . RAI-specific[a] symptoms referenced as difficult to manage by focus group participants and stakeholders during qualitative interviews.

Difficult to manage RAI symptom	Focus group participants (n=22), n (%)	Stakeholders (n=12), n (%)
Salivary		
Dry mouth	11 (50)	6 (50)
Change in taste	12 (55)	5 (42)
Mouth ulcers	2 (9)	2 (17)
Salivary gland swelling	8 (36)	4 (33)
Swallowing difficulties	2 (9)	3 (25)
Dental issues	4 (18)	1 (8)
Lacrimal		
Dry eyes	7 (32)	3 (25)
Watery eyes	__[b]	1 (8)
Nasal		
Dry nose	3 (14)	
Loss of smell	1 (5)	—
Nose bleeds	2 (9)	—
Other		
Fatigue[c]	4 (18)	4 (33)
Headache	3 (14)	—
Gastrointestinal issues	2 (9)	—
Menstrual cycle changes[c]	1 (5)	—
Hair loss	5 (23)	—
Nausea	8 (36)	3 (25)

[a]RAI: radioactive iodine.

[b]Not endorsed by any focus group participants or stakeholders.

[c]Symptoms may occur from concurrent treatment such as thyroid stimulating hormone therapy.

Domain 2: Key Intervention Structure and Content Feedback on Mockups

The key message we derived from the focus group participant and stakeholder findings was that the web-based intervention requires a more streamlined home page narrative instructing patients on using the intervention in the context of experienced symptoms and personalized self-management strategies. Both focus group participants and stakeholders felt the home page mockup needed to provide relevant information about the website for a patient engaging in the intervention. Stakeholders emphasized the desirability of repeated language within the intervention content narrative to redirect patients to their health care team should emergent symptoms occur requiring medical oversight. Consequently, for the RAI support intervention to be perceived as credible, reliable, and trustworthy, focus group participants and stakeholders indicated it was very important that it include references to differentiate the intervention from nonacademic public-facing websites.

Overall, focus group participants and stakeholders found the education and self-management of symptoms mockup pages to be helpful components of a web-based intervention. Most focus group participants found that the education pages presented "distressing" symptoms in a simple and approachable visual format. Together, focus group participants and stakeholders found the RAI symptom management mockup to be useful and found the video format to allow for more engagement across a range of patient learning styles. Both patients and stakeholders emphasized the need for short video content (less than 3 minutes in length) and accessibility features such as closed captioning to maintain user video engagement. The frequently asked questions content page received mixed responses between focus group participants and stakeholders. Focus group participants highlighted the value of prepared questions to address patient RAI symptom concerns and increase patient-provider communication during appointments. In contrast, stakeholders shared concerns about the potential negative consequences on provider workload of "unanswered" questions that a patient may bring to a clinical encounter. Most stakeholders recommended answering the questions on the frequently asked questions page directly and including hyperlinks within the web-based intervention with answers to the questions. Both groups suggested the need to tailor questions to address RAI symptoms and aftercare and to use lay language to streamline information.

XSL•FO
RenderX

Domain 3: Intervention Content to Promote RAI Symptom Management and Survivorship Care

Focus group participants and stakeholders identified topics for intervention content and characteristics that had previously not been considered in the initial intervention design. Many focus group participants expressed uncertainty about which foods to eat after RAI treatment and suggested intervention content on foods that stimulate salivary gland secretion. Stakeholders reinforced the consideration of nutrition-based self-management strategies to reduce dry mouth and changes in taste, with an emphasis on the noncurative nature of these strategies. Many focus group participants indicated recommendations on how to communicate with family members about side effects, especially in light of the "good cancer" label, would be helpful after RAI treatment. A smaller number of focus group participants with children emphasized the need for intervention content on how to communicate with children about thyroid cancer, RAI treatment-related side effects, and survivorship. Most stakeholders supported the inclusion of communication strategies with family members but emphasized tailoring communication with adolescents and children at age-appropriate levels. Focus group participants expressed challenges in long-term monitoring for the management of RAI symptoms, monitoring for potential cancer recurrence, and not knowing which provider is appropriate for RAI symptom management.

Discussion

Principal Results

Following the SCT framework and the person-centered approach to inform the development of a theoretically driven mHealth symptom management intervention, we elicited feedback from focus group participants and stakeholders regarding which symptoms related to RAI treatment are the most challenging to manage. Findings in this study support evidence from prior research on RAI-related toxicities [11,12,16], including well-documented salivary gland swelling, xerostomia or dry mouth, and changes in taste [15]. Both patients and stakeholders emphasized the utility of a web-based intervention for patient self-management of mild to moderate symptoms with redirection to the patient health care team to treat severe symptoms.

Findings from this study build upon prior work investigating cancer survivor self-management interventions [44,45] with an emphasis on patient knowledge, self-efficacy, and skill development as intervention components from SCT [30,46]. Most self-management interventions are designed for a particular type of cancer or symptom [44,46], which does not capture the RAI treatment aftercare experiences among patients with DTC. RAI treatment-specific issues and aftercare needs reported by our focus group participants and stakeholders reinforce prior evidence of unmet informational needs related to RAI side effects and long-term symptom management in this population [1,17,19,47]. Both focus group participants and stakeholders found the content and digital format of the web-based mockups of salivary symptom education and a salivary gland massage video helpful in engaging and empowering patients in symptom self-management. These findings support the potential of web-based symptom management interventions in patient outcomes with accessible and digital information and support [20,28,44].

Additionally, our results may potentially address supportive care gaps through focus group patient and stakeholder identification of nutrition management strategies, communication with family, and follow-up care planning as tailored survivorship components to include in a web-based intervention. Survivorship care among this patient population involves unique coordination across specialists, with an endocrinologist or primary care provider at the center [17,47], and adjunct supportive programs such as a web-based intervention for patient self-management of RAI symptoms may offer assistance in reducing fragmented care gaps, improving self-efficacy in RAI symptom management, and reducing patient health care utilization.

Limitations

There are several limitations to this study. Our patients and stakeholders primarily identified as White and non-Hispanic, which limits the generalizability of the data. Furthermore, our patient sample comprised insured patients with thyroid cancer, and our qualitative findings did not capture the perspective of uninsured patients. Our qualitative results identified a gap in the current delivery of coordinated patient care across subspecialties. It should be noted that this care gap is reflective of one academic hospital center and is not generalizable across all settings. While our results reached saturation across stakeholder specialists, saturation within subspecialists did not occur due to the limited number of subspecialists enrolled in each group. In addition, we were not able to engage a dentist as a stakeholder. Future qualitative research should elicit endocrinologists' and primary care providers' perspectives on mHealth behavioral interventions for the self-management of RAI symptoms to enhance the implementation of coordinated care efforts.

A small subset of focus group participants attributed fatigue and menstrual cycle changes to RAI treatment; however, these are symptoms from concurrent treatments such as thyroidectomy and thyroid-stimulating hormones. Patients may have difficulty differentiating which symptoms are due to RAI or other concurrent treatment. In addition, symptoms such as menstrual cycle changes could be due to both RAI and adjustment of thyroid hormone dosage. Future studies can work toward disambiguating symptoms for patient education and content development of a web-based intervention for RAI treatment symptom management.

Conclusions

Through a SCT framework and person-based approach, results from this qualitative study highlight important intervention content considerations such as the inclusion of nutrition management strategies, communication skills, and follow-up care planning for the self-management of RAI symptoms among patients with DTC. These findings can inform the development of future self-management interventions tailored to patients with DTC after RAI treatment.

XSL•FO
RenderX

Acknowledgments

The authors express their gratitude to all patients who participated in the focus group interviews and medical specialists, other health care providers, and patient advocates for their participation in individual interviews. The authors would like to thank Gary Bloom, the executive director of ThyCa: Thyroid Cancer Survivors' Inc, for his time and effort in the recruitment of patient advocates for the study. We would like to thank Emily Lattie for her time and effort in codeveloping the focus group facilitator guide. Finally, the authors acknowledge the efforts of Gautham Pillai for his contributions to the interviews. This work was supported by the National Cancer Institute (K99CA277588 and CCSG: P30-CA051008). ALC received research support from the National Cancer Institute (K99CA277588, CCSG: P30-CA051008, and T32CA261787). AMJ received research support from the American Cancer Society Center for Diversity in Cancer Research Training (supplement to ACS IRG-17-177-23) and her efforts were supported in part through funding for trainees provided by the Lombardi Comprehensive Cancer Center.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to following ethical principles for the protection of confidentiality given the small sample size in this study, and following the guidance of the Health Insurance Portability and Accountability Act per our consent process but are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The open-ended qualitative focus group moderator guide is informed by Social Cognitive Theory and person-based design principles to elicit patient perspectives about RAI symptoms that were challenging to manage, identify content that would be helpful in a symptom management website, and obtain feedback on the formatting and content of static images or mockups.
[DOCX File, 37 KB - formative_v9i1e60588_app1.docx]

Multimedia Appendix 2
The open-ended qualitative semistructured interview guide is informed by Social Cognitive Theory and person-based design principles to elicit stakeholder perspectives about RAI symptoms that are challenging to manage, identify content that would be helpful in a symptom management website, and obtain feedback on the formatting and content of static images or mockups.
[DOCX File, 230 KB - formative_v9i1e60588_app2.docx]

Multimedia Appendix 3
Four mockups of a prospective web-based radioactive iodine treatment symptom management intervention reviewed by focus group participants and stakeholders: 1) home page, 2) salivary gland massage video, 3) salivary gland education page, and 4) frequently asked questions page.
[PDF File, 623 KB - formative_v9i1e60588_app3.pdf]

References

1. Diamond-Rossi SA, Jonklaas J, Jensen RE, et al. Looking under the hood of "the Cadillac of cancers:" radioactive iodine-related craniofacial side effects among patients with thyroid cancer. J Cancer Surviv 2020 Dec;14(6):847-857. [doi: 10.1007/s11764-020-00897-5]
2. Randle RW, Bushman NM, Orne J, et al. Papillary thyroid cancer: the good and bad of the "Good Cancer". Thyroid 2017 Jul;27(7):902-907. [doi: 10.1089/thy.2016.0632] [Medline: 28510505]
3. Cho S, Tutelman P, Chomistek T, et al. "No such thing as good cancer": a qualitative exploration of the experience of early-onset thyroid cancer in survivors. Support Care Cancer 2024 Sep 18;32(10):671. [doi: 10.1007/s00520-024-08850-3] [Medline: 39292348]
4. SEER cancer stat facts: thyroid cancer.: National Cancer Institute URL: https://seer.cancer.gov/statfacts/html/thyro.html [accessed 2024-11-25]
5. Guo H, Zhang N, Hu Y, Zhang F, Huang T, Shen N. Radioactive iodine therapy strategies for distinct types of differentiated thyroid cancer: a propensity score-matched analysis. Front Endocrinol (Lausanne) 2023;14:1158581. [doi: 10.3389/fendo.2023.1158581] [Medline: 37664843]
6. Haugen BR, Alexander EK, Bible KC, et al. 2015 American Thyroid Association Management guidelines for adult patients with thyroid nodules and differentiated thyroid cancer: the American Thyroid Association Guidelines task force on thyroid nodules and differentiated thyroid cancer. Thyroid 2016 Jan;26(1):1-133. [doi: 10.1089/thy.2015.0020] [Medline: 26462967]

XSL•FO
RenderX

7. Nguyen NC, Anigati EM, Desai NB, Öz OK. Radioactive iodine therapy in differentiated thyroid cancer: an update on dose recommendations and risk of secondary primary malignancies. Semin Nucl Med 2024 Jul;54(4):488-496. [doi: 10.1053/j.semnuclmed.2024.05.002] [Medline: 38772827]
8. Goswami S, Peipert BJ, Mongelli MN, et al. Clinical factors associated with worse quality-of-life scores in United States thyroid cancer survivors. Surgery 2019 Jul;166(1):69-74. [doi: 10.1016/j.surg.2019.01.034] [Medline: 30898373]
9. Goswami S, Mongelli M, Peipert BJ, Helenowski I, Yount SE, Sturgeon C. Benchmarking health-related quality of life in thyroid cancer versus other cancers and United States normative data. Surgery 2018 Nov;164(5):986-992. [doi: 10.1016/j.surg.2018.06.042] [Medline: 30149935]
10. Applewhite MK, James BC, Kaplan SP, et al. Quality of life in thyroid cancer is similar to that of other cancers with worse survival. World J Surg 2016 Mar;40(3):551-561. [doi: 10.1007/s00268-015-3300-5] [Medline: 26546191]
11. Avram AM, Giovanella L, Greenspan B, et al. SNMMI procedure standard/EANM practice guideline for nuclear medicine evaluation and therapy of differentiated thyroid cancer: abbreviated version. J Nucl Med 2022 Jun;63(6):15N-35N. [Medline: 35649660]
12. Singer MC, Marchal F, Angelos P, et al. Salivary and lacrimal dysfunction after radioactive iodine for differentiated thyroid cancer: American Head and Neck Society Endocrine Surgery Section and Salivary Gland Section joint multidisciplinary clinical consensus statement of otolaryngology, ophthalmology, nuclear medicine and endocrinology. Head Neck 2020 Nov;42(11):3446-3459. [doi: 10.1002/hed.26417] [Medline: 32812307]
13. Dedhia PH, Saucke MC, Long KL, Doherty GM, Pitt SC. Physician perspectives of overdiagnosis and overtreatment of low-risk papillary thyroid cancer in the US. JAMA Netw Open 2022 Apr 1;5(4):e228722. [doi: 10.1001/jamanetworkopen.2022.8722] [Medline: 35467735]
14. Thyroid carcinoma. NCCN clinical practice guidelines in oncology version 5. 2025. URL: https://www.nccn.org/professionals/physician_gls/pdf/thyroid_blocks.pdf [accessed 2025-03-04]
15. Lee SL. Complications of radioactive iodine treatment of thyroid carcinoma. J Natl Compr Cancer Netw 2010;8(11):1277-1287. [doi: 10.6004/jnccn.2010.0094]
16. Jonklaas J. Nasal symptoms after radioiodine therapy: a rarely described side effect with similar frequency to lacrimal dysfunction. Thyroid 2014 Dec;24(12):1806-1814. [doi: 10.1089/thy.2014.0162] [Medline: 25090584]
17. Banach R, Bartès B, Farnell K, et al. Results of the Thyroid Cancer Alliance international patient/survivor survey: psychosocial/informational support needs, treatment side effects and international differences in care. Hormones (Athens) 2013;12(3):428-438. [doi: 10.1007/BF03401308] [Medline: 24121384]
18. Carr AL, Pillai G, Jonklaas J, Bloom G. Health-related quality of life associations in patients with differentiated thyroid cancer after radioactive iodine treatment: a cross-sectional study. Thyroid® 2023;33(S1):A-74. [doi: 10.1089/thy.2023.29156.abstracts]
19. Hyun YG, Alhashemi A, Fazelzad R, Goldberg AS, Goldstein DP, Sawka AM. A systematic review of unmet information and psychosocial support needs of adults diagnosed with thyroid cancer. Thyroid 2016 Sep;26(9):1239-1250. [doi: 10.1089/thy.2016.0039] [Medline: 27350421]
20. Walsh JC, Groarke JM. Integrating behavioral science with mobile (mHealth) technology to optimize health behavior change interventions. Eur Psychol 2019 Jan;24(1):38-48. [doi: 10.1027/1016-9040/a000351]
21. Penedo FJ, Oswald LB, Kronenfeld JP, Garcia SF, Cella D, Yanez B. The increasing value of eHealth in the delivery of patient-centred cancer care. Lancet Oncol 2020 May;21(5):e240-e251. [doi: 10.1016/S1470-2045(20)30021-8] [Medline: 32359500]
22. Hong YA, Hossain MM, Chou WYS. Digital interventions to facilitate patient-provider communication in cancer care: a systematic review. Psychooncology 2020 Apr;29(4):591-603. [doi: 10.1002/pon.5310] [Medline: 31834650]
23. Hesse BW, Kwasnicka D, Ahern DK. Emerging digital technologies in cancer treatment, prevention, and control. Transl Behav Med 2021 Nov 30;11(11):2009-2017. [doi: 10.1093/tbm/ibab033]
24. Aapro M, Bossi P, Dasari A, et al. Digital health for optimal supportive care in oncology: benefits, limits, and future perspectives. Support Care Cancer 2020 Oct;28(10):4589-4612. [doi: 10.1007/s00520-020-05539-1]
25. Davis SW, Oakley-Girvan I. mHealth education applications along the cancer continuum. J Cancer Educ 2015 Jun;30(2):388-394. [doi: 10.1007/s13187-014-0761-4] [Medline: 25482319]
26. Richards R, Kinnersley P, Brain K, McCutchan G, Staffurth J, Wood F. Use of mobile devices to help cancer patients meet their information needs in non-inpatient settings: systematic review. JMIR Mhealth Uhealth 2018 Dec 14;6(12):e10026. [doi: 10.2196/10026] [Medline: 30552082]
27. O'Leary DP, Zaheer A, Redmond HP, Corrigan MA. Integration of advances in social media and mHealth technology are pivotal to successful cancer prevention and control. Mhealth 2016;2:38. [doi: 10.21037/mhealth.2016.09.02] [Medline: 28293611]
28. Fridriksdottir N, Gunnarsdottir S, Zoëga S, Ingadottir B, Hafsteinsdottir EJG. Effects of web-based interventions on cancer patients' symptoms: review of randomized trials. Support Care Cancer 2018 Feb;26(2):337-351. [doi: 10.1007/s00520-017-3882-6] [Medline: 28921391]

XSL•FO
RenderX

29. Moradian S, Voelker N, Brown C, Liu G, Howell D. Effectiveness of internet-based interventions in managing chemotherapy-related symptoms in patients with cancer: a systematic literature review. Support Care Cancer 2018 Feb;26(2):361-374. [doi: 10.1007/s00520-017-3900-8] [Medline: 28948360]
30. Bandura A. Health promotion by social cognitive means. Health Educ Behav Off Publ Soc Public Health Educ 2004 Apr;31(2):143-164. [doi: 10.1177/1090198104263660] [Medline: 15090118]
31. Graves KD. Social cognitive theory and cancer patients' quality of life: a meta-analysis of psychosocial intervention components. Health Psychol Off J Div Health Psychol Am Psychol Assoc 2003 Mar;22(2):210-219. [Medline: 12683741]
32. Yardley L, Morrison L, Bradbury K, Muller I. The person-based approach to intervention development: application to digital health-related behavior change interventions. J Med Internet Res 2015 Jan 30;17(1):e30. [doi: 10.2196/jmir.4055] [Medline: 25639757]
33. Stacey FG, James EL, Chapman K, Courneya KS, Lubans DR. A systematic review and meta-analysis of social cognitive theory-based physical activity and/or nutrition behavior change interventions for cancer survivors. J Cancer Surviv 2015 Jun;9(2):305-338. [doi: 10.1007/s11764-014-0413-z] [Medline: 25432633]
34. Rogers LQ, Matevey C, Hopkins-Price P, Shah P, Dunnington G, Courneya KS. Exploring social cognitive theory constructs for promoting exercise among breast cancer patients. Cancer Nurs 2004;27(6):462-473. [doi: 10.1097/00002820-200411000-00006] [Medline: 15632787]
35. Hatchett A, Hallam JS, Ford MA. Evaluation of a social cognitive theory-based email intervention designed to influence the physical activity of survivors of breast cancer. Psychooncology 2013 Apr;22(4):829-836. [doi: 10.1002/pon.3082] [Medline: 22573338]
36. Short CE, James EL, Plotnikoff RC. How social cognitive theory can help oncology-based health professionals promote physical activity among breast cancer survivors. Eur J Oncol Nurs 2013 Aug;17(4):482-489. [doi: 10.1016/j.ejon.2012.10.009] [Medline: 23177321]
37. Graves KD, Carter CL, Anderson ES, Winett RA. Quality of life pilot intervention for breast cancer patients: use of social cognitive theory. Palliat Support Care 2003 Jun;1(2):121-134. [doi: 10.1017/s1478951503030268] [Medline: 16594274]
38. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
39. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381. [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
40. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
41. Krippendorff K. Estimating the reliability, systematic error and random error of interval data. Educ Psychol Meas 1970 Apr;30(1):61-70. [doi: 10.1177/001316447003000105]
42. Krippendorff K. Content Analysis: An Introduction to Its Methodology, 2nd edition: SAGE Publications; 2004.
43. Marzi G, Balzano M, Marchiori D. K-alpha calculator-Krippendorff's alpha calculator: a user-friendly tool for computing Krippendorff's alpha inter-rater reliability coefficient. MethodsX 2024 Jun;12:102545. [doi: 10.1016/j.mex.2023.102545] [Medline: 39669968]
44. Leach CR, Hudson SV, Diefenbach MA, et al. Cancer health self-efficacy improvement in a randomized controlled trial. Cancer 2022 Feb 1;128(3):597-605. [doi: 10.1002/cncr.33947] [Medline: 34668569]
45. Boland L, Bennett K, Connolly D. Self-management interventions for cancer survivors: a systematic review. Support Care Cancer 2018 May;26(5):1585-1595. [doi: 10.1007/s00520-017-3999-7] [Medline: 29199362]
46. Rimmer B, Brown MC, Sotire T, et al. Characteristics and components of self-management interventions for improving quality of life in cancer survivors: a systematic review. Cancers (Basel) 2024;16(1):14. [doi: 10.3390/cancers16010014] [Medline: 39123377]
47. Kotwal A, Fingeret A, Knape A, Patel A, Bradford Bell E, Goldner W. Thyroid cancer survivorship: challenges and opportunities. Endocr Pract 2024 Nov;30(11):1097-1102. [doi: 10.1016/j.eprac.2024.08.003] [Medline: 39209023]

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Research
DTC: differentiated thyroid cancer
HRQOL: health-related quality of life
mHealth: mobile health
RAI: radioactive iodine
REDCap: Research Electronic Data Capture
SCT: Social Cognitive Theory

XSL•FO
RenderX

Edited by A Mavragani; submitted 16.05.24; peer-reviewed by M Kim; revised version received 11.02.25; accepted 12.02.25; published 19.03.25.

Please cite as:
Carr AL, Jenkins AM, Jonklaas J, Gabriel K, Miller KE, Graves KD
Patient and Provider Perspectives of a Web-Based Intervention to Support Symptom Management After Radioactive Iodine Treatment for Differentiated Thyroid Cancer: Qualitative Study
JMIR Form Res 2025;9:e60588
URL: https://formative.jmir.org/2025/1/e60588
doi:10.2196/60588

© Alaina L Carr, Angela M Jenkins, Jacqueline Jonklaas, Kate Gabriel, Kristen E Miller, Kristi D Graves. Originally published in JMIR Formative Research (https://formative.jmir.org), 19.3.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Customizing Computerized Adaptive Test Stopping Rules for Clinical Settings Using the Negative Affect Subdomain of the NIH Toolbox Emotion Battery: Simulation Study

Saki Amagai[1], BA; Aaron J Kaat[1], PhD; Rina S Fox[2], MPH, PhD; Emily H Ho[1], PhD; Sarah Pila[1], PhD; Michael A Kallen[1†], MPH, PhD; Benjamin D Schalet[3], PhD; Cindy J Nowinski[1], MD, PhD; Richard C Gershon[1], PhD

[1]Department of Medical Social Sciences, Northwestern University Feinberg School of Medicine, Chicago, IL, United States

[2]College of Nursing, University of Arizona, Tuscon, AZ, United States

[3]Department of Epidemiology and Data Science, Amsterdam University Medical Centers, Amsterdam, The Netherlands

[†]deceased

Corresponding Author:
Saki Amagai, BA
Department of Medical Social Sciences, Northwestern University Feinberg School of Medicine, Chicago, IL, United States

Abstract

Background: Patient-reported outcome measures are crucial for informed medical decisions and evaluating treatments. However, they can be burdensome for patients and sometimes lack the reliability clinicians need for clear clinical interpretations.

Objective: We aimed to assess the extent to which applying alternative stopping rules can increase reliability for clinical use while minimizing the burden of computerized adaptive tests (CATs).

Methods: CAT simulations were conducted on 3 adult item banks in the NIH Toolbox for Assessment of Neurological and Behavioral Function Emotion Battery; the item banks were in the Negative Affect subdomain (ie, Anger Affect, Fear Affect, and Sadness) and contained at least 8 items. In the originally applied NIH Toolbox CAT stopping rules, the CAT was stopped if the score SE reached <0.3 before 12 items were administered. We first contrasted this with a SE-change rule in a planned simulation analysis. We then contrasted the original rules with fixed-length CATs (4 - 12 items), a reduction of the maximum number of items to 8, and other modifications in post hoc analyses. Burden was measured by the number of items administered per simulation, precision by the percentage of assessments yielding reliability cutoffs (0.85, 0.90, and 0.95), and accurate score recovery by the root mean squared error between the generating θ and the CAT-estimated "expected a posteriori"–based θ.

Results: In general, relative to the original rules, the alternative stopping rules slightly decreased burden while also increasing the proportion of assessments achieving high reliability for the adult banks; however, the SE-change rule and fixed-length CATs with 8 or fewer items also notably increased assessments yielding reliability <0.85. Among the alternative rules explored, the reduced maximum stopping rule best balanced precision and parsimony, presenting another option beyond the original rules.

Conclusions: Our findings demonstrate the challenges in attempting to reduce test burden while also achieving score precision for clinical use. Stopping rules should be modified in accordance with the context of the study population and the purpose of the study.

(JMIR Form Res 2025;9:e60215) doi:10.2196/60215

KEYWORDS

computerized adaptive testing; CAT; stopping rules; NIH Toolbox; reliability; test burden; clinical setting; patient-reported outcome; clinician

Introduction

Background

The patient's perspective, typically collected using patient-reported outcome measures (PROMs), is an imperative component of effective medical decision-making and treatment evaluation [1-4]. However, PROMs can be time-consuming for respondents to complete and difficult for clinicians to interpret, thus limiting their practical utility. Increasing importance has therefore been placed on minimizing response burden. For example, newer PROM initiatives have incorporated item response theory (IRT) and computerized adaptive test (CAT) capabilities, which enable the derivation of efficient, reliable, and precise assessments from large banks of items [5,6].

The NIH Toolbox for Assessment of Neurological and Behavioral Function is one such measurement initiative [7]. The NIH Toolbox is a set of neurobehavioral measures that can be used to assess sensory, motor, emotional, and cognitive

XSL•FO
RenderX

functioning among individuals aged 3 to 85 years. While many measures included in the NIH Toolbox focus on performance-based assessments, the Emotion Battery consists exclusively of PROMs assessing emotions such as sorrow, joy, or fear [8]. The battery consists of 4 subdomains: Negative Affect (Anger, Fear, and Sadness), Psychological Well-Being (Life Satisfaction, Meaning and Purpose, and Positive Affect), Social Relationships (Companionship, Fear, Perceived Social Support, Positive Social Development, and Social Distress), and Stress and Self-Efficacy (Self-Efficacy and Stress). Across all item banks, higher scores indicate higher levels of the construct being assessed (eg, a higher score on the Sadness measure indicates higher levels of sorrowful feelings or emotions) [8]. The item banks were calibrated using the graded response IRT model [9]. This calibration lines up all the items along a common mathematical metric, which allows any subset of items to be used to track a given patient's performance longitudinally or compare performance scores across patients under a common metric [10-12]. The NIH Toolbox Emotion Battery calculates "expected a posteriori" (EAP) scores, which are a type of IRT-based score, on the θ metric (mean 0, SD 1). These scores are then transformed to a *t* score metric (mean 50, SD 10) for individual reporting. The Emotion Battery provides a psychometrically strong option for evaluating emotional health that is particularly well suited for use in combination with other performance-based measures included in the NIH Toolbox.

Measures in the NIH Toolbox Emotion Battery are administered as fixed forms or variable-length CATs. The fixed forms consist of a set of items that are all administered in the same order during every administration. Conversely, when variable-length CATs are administered, the first item typically reflects average functioning or symptom severity, as the maximum weighted information value for a starting item is at the population mean (θ=0; *t* score=50). The response to that item is then used to calculate the provisional score estimate for the measured construct and to identify the best "next" item, within the remaining items in the bank, that will maximize weighted information based on the provisional ability estimate for the patient [13-15]. A patient's estimated θ score is iteratively refined after each item until (1) a minimum number of items is administered and either (2a) a prespecified level of measurement precision is achieved or (2b) a prespecified maximum number of items is administered. These conditions are known as "stopping rules" [16,17]. After 4 items have been administered (the minimum), the default for the NIH Toolbox CAT administration is to stop the CAT after either 12 items have been administered (the maximum) or the SE of the EAP score estimate is less than 0.3, which corresponds to a minimum reliability of 0.91. Other CATs may implement these or additional stopping rules, or may include content balancing or item exposure constraints. There is also a fully Bayesian approach for estimating ability using the Markov chain Monte Carlo method, which can be used to incorporate empirical information about candidates [18]. The use of empirical prior information has been shown to reduce response burden for patients and clinical practitioners [19]. While these other stopping rules may appear in other contexts (including in the NIH Toolbox Cognitive Battery for exposure constraints), the NIH Toolbox Emotion CATs do not implement any of these other features [20].

Compared to legacy measures, these stopping rules substantially reduce response burden, yet NIH Toolbox CATs still impose a significant burden on respondents with no or minimal symptoms. Similar issues have been documented in educational measurement using precision rules as a primary criterion for stopping rules [21]. For example, if an item bank has numerous items reflecting extreme sadness but few items reflecting minimal sadness, the statistically most-informative items for someone experiencing little-to-no sadness will be exhausted quickly, and additional items will be increasingly less targeted and unable to notably improve the score precision (ie, reducing the SE).

In such a scenario, the CAT will continue to select the next best (but still mistargeted) item until the maximum CAT length is reached. While there are sufficient items in each item bank to stratify pathology, if a person is not impaired with respect to the domain being explored, it is difficult or impossible to accurately differentiate how healthy a person is (eg, how not sad a person may be). As a result, patients with the best health in a given domain are more likely to be administered the maximum 12 items in a CAT as compared to patients experiencing more severe symptoms [22]. Therefore, the first goal of this study was to develop and systematically assess alternative stopping rules that would minimize response burden and frustration for these patients with few-to-no symptoms, as patients with wide range of levels of functioning are administered NIH Toolbox measures.

The second goal of this study was to alter stopping rules to improve reliability and, more specifically, to achieve clinically relevant reliability (eg, reliability ≥0.95). This is especially important today as the NIH Toolbox is increasingly being used with clinical populations and in clinical settings, even though it was developed specifically for use in research. As stated, the current CAT stopping rules call for a test to end when the SE is <0.3, which corresponds to a minimum reliability of 0.91. Although this level of reliability is recommended for use in general populations [23], it may be inadequate to identify clinically relevant symptoms and individual patient changes across time [24]. Thus, as NIH Toolbox measures are implemented more frequently for clinical use in health care delivery settings [25,26], the ability to achieve this level of reliability appears increasingly valuable. Reaching clinical-level reliability will lead to fewer errors in patient assessments and increase confidence in the treatments or type of care to be recommended.

Recent work has evaluated alternative approaches to CAT administration in an effort to promote more efficient and effective testing. In their simulation study, Kallen and colleagues [22] shortened the minimum test length for individuals completely denying symptoms to only 2 items and added a minimum SE-change rule, while retaining the SE <0.3 rule and a maximum test length of 12 items. The minimum SE-change rule allowed the CAT to stop when the CAT algorithm detected that the administration of additional items would not reduce the SE. The impact of adding the SE-change rule was negligible.

XSL•FO
RenderX

This simulation, combined with the shortened test length for individuals reporting absolutely no symptoms, addressed the concern that high-functioning individuals often receive lengthier CATs under the originally applied stopping rules for the NIH Toolbox. Importantly, this approach yielded scores that varied minimally from those obtained using the original stopping rules, particularly for scores in the clinically relevant range. Very few administrations required more than 8 items under these new stopping rules. However, the application of a SE <0.3 stopping rule retained a target reliability of 0.91, below a clinically relevant range (eg, ≥0.95), thus questioning the appropriateness of these rules as optimal for application in clinical settings.

Purpose of This Study

The NIH Toolbox has been increasingly used in real-world clinical settings, and patients with a wide range of levels of functioning complete these measures. The purpose of this study was to propose and evaluate new stopping rules that could optimally balance two competing priorities and goals that are increasingly common in both research and clinical practice:

1. Increase reliability (≥0.95) for the clinical population (ie, adjust the SE threshold to require higher reliability) without unacceptably increasing the number of items
2. Decrease response burden for nonsymptomatic individuals (ie, reduce the number of items administered) without unacceptably reducing reliability (<0.85)

By using in-person, in-production item banks, this study provides real-world insights that go beyond the theoretical understanding of how stopping rules function.

Methods

Planned Simulation Study

Design

Firestar was used to simulate examinee responses and item bank CAT administration [27]. Three adult (age ≥18 years) NIH Toolbox Negative Affect item banks (Anger Affect, Fear Affect, and Sadness) were evaluated. All 3 of the item banks contain at least 8 items and are IRT-calibrated and EAP-scored (Table 1). More detailed descriptions of the item banks, along with examples, can be found in Table S1 in Multimedia Appendix 1. The simulation study manipulated the CAT administration to implement both the original stopping rules and a new set of rules, which we named the “SE-change” rules. The SE-change rules terminated a test when (1) a minimum of 4 items or maximum of 8 items had been administered, thus decreasing the maximum test length; (2) SE was <0.224, thus increasing target reliability to 0.95; or (3) the SE was reduced by less than 0.01, thus stopping the test when measurement precision would not increase sufficiently by administering an additional item (Table 2). We hypothesized that the SE-change stopping rules would increase the proportion of assessments achieving clinically relevant reliability without unacceptably increasing the number of items administered to nonsymptomatic individuals.

Table . NIH Toolbox Emotion Battery measures from the Negative Affect subdomain included in the simulation.

Measure	Typical computer adaptive test administration (number of items available in bank)	Slope parameter, mean (SD)	Average within-item thresholds, mean (SD)
Anger Affect	22	2.22 (0.43)	1.45 (0.44)
Fear Affect	29	2.72 (0.78)	1.44 (0.40)
Sadness	28	3.14 (0.67)	1.27 (0.24)

XSL•FO
RenderX

Table . Stopping rule models evaluated.

Stopping rule model	Minimum items	Maximum items	SE threshold for interim stopping rule	SE-change threshold for interim stopping
Original	4	12	0.3 (reliability 0.91)	__[a]
SE-change	4	8	0.224 (reliability 0.95)	<0.01
4-item fixed	4	4	—	—
5-item fixed	5	5	—	—
6-item fixed	6	6	—	—
7-item fixed	7	7	—	—
8-item fixed	8	8	—	—
9-item fixed	9	9	—	—
10-item fixed	10	10	—	—
11-item fixed	11	11	—	—
12-item fixed	12	12	—	—
Reduced maximum	4	8	0.224 (reliability 0.95)	—

[a]Not applicable.

Interim item selection (using maximum posterior weighted information), starting item, and scoring (using EAP estimates with a standard normal prior) were constrained to be the same across all simulated CATs. For each stopping rule, 1000 responses were simulated with uniform distribution (for θ values from –4 to 4) using the full item banks. The simulations used a uniform distribution to eliminate artificial biases that could be inadvertently introduced if the person ability distribution was simulated such that it mimicked the item difficulty distribution. The unit of analysis was individual CAT assessments. The CAT simulations proceeded until one of the stopping rules was met. All simulations and analyses were conducted using the *Firestar* package [27] and R (version 3.6.3; R Foundation for Statistical Computing).

Evaluation Criteria

We used 4 evaluation criteria—precision, accurate score recovery, test burden, and efficiency—to compare the impact of the different stopping rules. (1) Precision was evaluated by exploring the percentages of assessments yielding reliabilities less than 0.85, between 0.85 and 0.90, between 0.90 and 0.95, and greater than 0.95 across stopping rules. This approach allowed us to compare the reliability achieved by each set of stopping rules in a more nuanced way rather than assessing mean reliability, which can obscure the impact of stopping rules on conditional reliability across the trait range. Additionally, maximum attainable information curves were generated to determine theoretically attainable reliability for fixed-length CATs with up to 12 items for each of the 3 item banks. (2) Accurate score recovery refers to an evaluation of whether the CAT was able to recover the generating score and index the variability in score recovery; for this, we calculated the root mean squared error (RMSE) between the generating θ and the CAT-estimated, EAP-based θ from each version of the CAT. (3) Test burden, or response burden, was evaluated by comparing the number of items administered as a function of each set of stopping rules. (4) Efficiency was defined as the Fisher information per number of items administered, which is equivalent to $(1/SE^2)$/item count, such that higher values represent greater efficiency.

Post Hoc Simulation Study

Achieving reliability ≥0.85 is considered a critical issue for the priority of increasing reliability. However, the reduced number of items in the SE-change stopping rule led to an increase in the assessments achieving low empirical reliability (<0.85). That is, assessments that had higher reliability (>0.85) under the original stopping rules were no longer able to maintain the same level of precision under the SE-change rules. As one of our two goals was to reduce burden without unacceptably reducing reliability, we conducted post hoc analyses in which we evaluated two new sets of stopping rules to more optimally balance the two goals: increasing reliability for the clinical population while decreasing burden for nonsymptomatic individuals.

All methods used in the planned simulations, as outlined, were replicated for this post hoc analysis; the only adjustments made were the addition of two new sets of stopping rules: (1) fixed-length CATs ranging in length from 4 to 12 items, and (2) “reduced maximum” stopping rules in which tests were terminated when SE was <0.224 or a maximum of 8 items had been administered (Table 2). The same evaluation criteria were used to compare the different stopping rules.

Ethical Considerations

The research presented herein was a simulation study, which was determined to not be human subjects research. Therefore, ethical review and approval was not necessary. This study was conducted using data from the NIH Toolbox Negative Affect item banks. The analysis did not involve identified patient information or any interaction with human subjects. As such, it did not meet the criteria for human subjects research and did not require ethics board review.

XSL•FO
RenderX

Results

SE-Change Stopping Rules vs Original Stopping Rules

The average percentage of simulated assessments yielding reliability at cutoffs of 0.85, 0.90, and 0.95 across stopping rules can be found in Table 3. For Fear Affect and Sadness, a larger percentage of assessments yielded reliability >0.95 per the SE-change stopping rules compared to the original stopping rules (+33.9% Fear Affect;+20.4% Sadness). For Anger Affect, neither set of stopping rules were able to yield the clinically relevant target reliability of 0.95. This result is because Anger Affect is a "weaker" bank (ie, the maximum attainable test information function is lower) that requires more items to be administered in order to distinguish different levels of anger and to reach 0.95 reliability. Most importantly, the percentage of assessments yielding low reliability for all three item banks (ie, <0.85) substantially increased (+2.7% Anger Affect;+4.7% Fear Affect;+3.4% Sadness).

Table . Proportion of simulated assessments meeting various levels of reliability.

Domain and level of reliability	Proportion of assessments meeting specified level of reliability (%)											
	Original	SE-change	4-item, fixed	5-item, fixed	6-item, fixed	7-item, fixed	8-item, fixed	9-item, fixed	10-item, fixed	11-item, fixed	12-item, fixed	Reduced maximum
Anger Affect												
<0.85	27.4	30.1	43.4	34.2	30.9	30.8	30	29.6	28.9	28.8	28.6	28.6
0.85-0.90	5.5	12.8	56.6	52.7	17.7	11.3	8.7	6.7	6.4	5.1	4.9	6.9
0.90-0.95	67.1	57.1	0	13.1	51.4	57.9	61.3	63.7	64.7	66.1	47.4	64.5
≥0.95	0	0	0	0	0	0	0	0	0	0	19.1	0
Fear Affect												
<0.85	31.5	36.2	36.6	36.4	35.2	35.2	35	33.8	33	32.9	32.6	33.5
0.85-0.90	3.5	5.7	15.7	12.9	9.2	8.1	7.2	7	6.2	5.1	5	4.9
0.90-0.95	65	24.2	47.7	50.7	25.3	20	15.8	14.2	15.1	15.6	14.2	21.8
≥0.95	0	33.9	0	0	30.3	36.7	42	45	45.7	46.4	48.2	39.8
Sadness												
<0.85	32.5	35.9	37.8	49.8	36.4	36.1	35.6	35.6	35.1	34.9	34.7	36.1
0.85-0.90	5.3	18.3	19.9	4.4	16	12.8	10.2	8.9	8	7.6	6.2	5.8
0.90-0.95	49.1	12.4	41.3	15.6	13.9	12.7	13.8	13.6	13.6	13.2	14.1	17.9
≥0.95	13.1	33.4	1	30.2	33.7	38.4	40.4	41.9	43.3	44.3	45	40.2

Concerning accurate score recovery, the RMSE of θ estimates are presented in Figure S1 in Multimedia Appendix 2 and Figure 1, respectively. The RMSE increased for Fear Affect (+0.036), while it decreased slightly for the other two banks (Anger Affect –0.001; Sadness –0.001). Test burden decreased for all banks per the SE-change rules as opposed to the original stopping rules, as shown in Table 4 (Anger Affect −1.26 items; Fear Affect −1.39 items; Sadness −2.64 items). In the original stopping rules, the average number of items administered was 7.9, compared to 6.1 items under the SE-change stopping rules, a 22% savings in response burden across banks (Table 4). The largest decrease in response burden observed was 33% for the Sadness bank.

XSL•FO
RenderX

Figure 1. Root mean squared error (RMSE) of θ estimates.

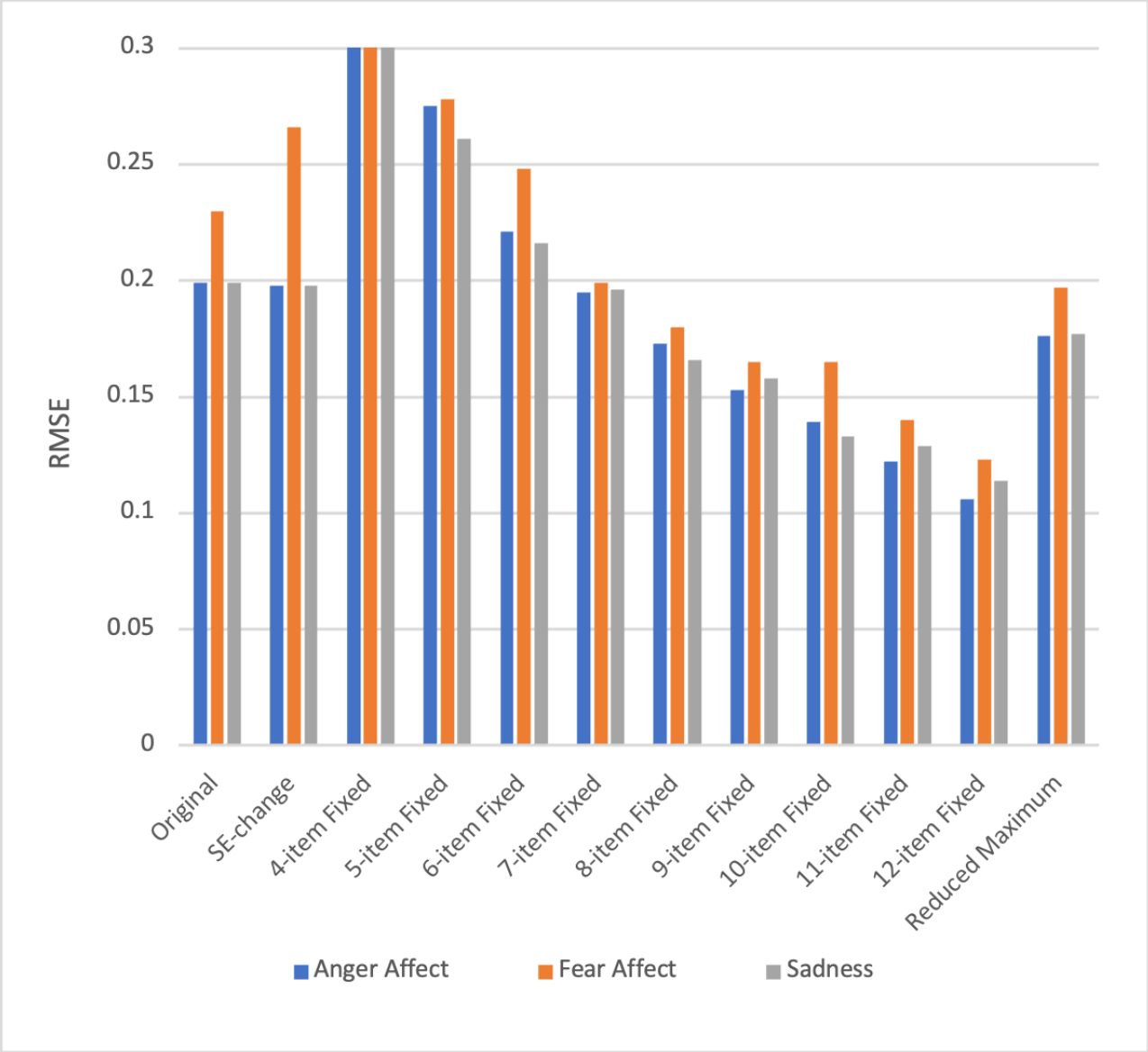

Table . Average number of items administered among 1000 simulated assessments.

Item bank	Number of items in item bank	Stopping rules		
		Original	SE-change	Reduced Maximum
Anger Affect	22	8.33	7.07	8.00
Fear Affect	29	7.40	6.01	7.36
Sadness	28	7.83	5.19	6.86

Fixed-Length CATs vs Original Stopping Rules

The fixed-length stopping rules led to an increased percentage of assessments reaching >0.85 reliability, as the test length is increased relative to the original stopping rules (Table 3). Per the fixed-length stopping rules, the target reliability level of 0.95 was achieved in the simulated assessments by only administering 6 items for Fear Affect and 4 items for Sadness, which is shorter than the average test length using the original stopping rules. However, more notably, the fixed-length CATs with 8 items or fewer also increased in simulated assessments with reliability <0.85 relative to the original stopping rules. Anger Affect was an exception and required 12 items to have any simulated assessments achieve the target reliability, and the assessments with reliability <0.85 did not improve relative to the original stopping rules even when administering 12 items.

Figure S2 in Multimedia Appendix 3 illustrates the potential inefficiencies for fixed-length CATs with up to 12 items. Consistent with the simulated results, Figure S2 in Multimedia Appendix 3 shows that there could be notable differences in the

XSL•FO
RenderX

number of items required to reach high reliability depending on the item bank. It is theoretically possible to achieve 0.95 reliability by administering at least 4 items for the Sadness and at least 5 items for the Fear Affect bank. Conversely, it can theoretically take at least 11 items for the Anger Affect bank to achieve 0.95 reliability, which suggests that the Anger Affect bank is a "weaker" bank that requires more items to be administered to distinguish levels of anger. Figure S2 in Multimedia Appendix 3 also suggests that individuals with little-to-no functioning of the trait (θ level=−1) will be subject to significant burden. For example, individuals with little-to-no sadness will never reach the 0.95 reliability threshold and will peak in maximum attainable information after 5 items, with diminishingly small gains in measurement precision. In this situation, high-functioning individuals will be administered more items until the maximum number of items is reached, thereby increasing burden.

The CAT was able to recover the generating score better when more items were administered in the fixed-length stopping rules as opposed to the original stopping rules. That is, RMSE of θ estimates decreased per the fixed-length stopping rules as opposed to the original stopping rules when administering more than 8 items for all banks. However, this improved accuracy score recovery came at a cost, as any fixed-length test greater than 8 items increases test burden relative to the original stopping rules (Table 4).

Reduced Maximum Stopping Rules vs Original Stopping Rules

A larger percentage of assessments yielded reliability >0.95 per the reduced maximum stopping rules as opposed to the original stopping rules (+0% Anger Affect, +39.8% Fear Affect, +40.2% Sadness). Importantly, this change did not excessively increase the percentage of assessments achieving low reliability (+1.2% Anger Affect; +2.0% Fear Affect; +3.6% Sadness).

The reduced maximum stopping rules improved the accuracy score recovery. The generating θ scores and score estimates were strongly correlated (>0.99) across banks for both stopping rules. RMSE also decreased for all banks (Anger Affect –0.023; Fear Affect –0.033; Sadness –0.022). Additionally, the average number of items administered was 7.9 compared to 7.4 items under the reduced maximum, a 6% savings in response burden across banks (Table 4). The largest decrease in response burden observed was 12% for the Sadness bank. Figure S3 in Multimedia Appendix 4 illustrates that the reduced maximum stopping rule may not be the most efficient choice across the θ range, but in general, all the rules are broadly comparable across banks.

Discussion

Principal Findings

Measures within the NIH Toolbox Emotion Battery that are administered as CATs are programmed to terminate when prespecified stopping rules are achieved. While the originally applied stopping rules are effective for many test takers, nonsymptomatic individuals often are burdened with many questions that are not relevant to their current health status. Additionally, many of the item banks have sufficient statistical information to achieve clinical-level reliability, but the original stopping rules terminate the assessment before that level of reliability is met. With increasing use of the NIH Toolbox in clinical settings, it is critical to simultaneously balance competing priorities: (1) maximizing score precision for clinical use and (2) minimizing burden for nonsymptomatic individuals. We evaluated alternative stopping rules to achieve these two goals.

The SE-change stopping rules were effective at reducing overall burden and were effective at ensuring adequate reliability and score recovery. Perhaps most importantly, although the SE-change stopping rules increased the percentage of assessments achieving clinically relevant reliability (≥0.95), the percentage of assessments achieving low empirical reliability (<0.85) also increased to an unacceptable level. Ultimately, we determined that the detrimental impact of the SE-change stopping rules on the percentage of assessments achieving low empirical reliability outweighed the modest benefits observed regarding test length and the percentage achieving very high reliability. This finding prompted us to explore additional stopping rule paradigms in post hoc simulations.

The fixed-length stopping rules were also not effective. Although it was possible to reduce burden relative to the original stopping rules by administering fewer than 8 items, it came at the cost of inflating the percentage of assessments achieving low empirical reliability (<0.85) while worsening accurate score recovery. By administering 8 or more items, it was also possible to maximize assessments achieving high reliability (>0.95) but at the cost of increasing burden. Given the benefits of longer fixed-length CATs, we opted to explore the reduced maximum stopping rules to determine if we could maintain this improvement in reliability without requiring 8 items across all administrations of all banks.

This subsequent evaluation identified a paradigm that more effectively balanced precision and parsimony. In general, the reduced maximum stopping rules demonstrated the same increase in reliability obtained with the 8-item fixed-length CATs; however, these stopping rules allowed for further decreased burden, as 8 items were not always required to reach a reliability level of 0.95.

One exception to the observed trends was the Anger Affect item bank. In simulations, this bank was unable to achieve reliability >0.95 for the reduced maximum stopping rule, as greater than 11 items are theoretically required to reach the reliability level of 0.95. As Figure S2 in Multimedia Appendix 3 showed, Anger Affect is an example of a "weak" bank, in which more items are needed to be administered to obtain clinically relevant reliability. The strength or weakness of a bank is related to the magnitude of the item slope parameters and the (statistically related) test information function. The Anger slope parameters are lower—therefore they are less-strongly related to the latent construct and contain more "error variance" or noise in their ratings. One possible hypothesis is that there are numerous and varied reasons someone might endorse or not endorse an Anger item, so the construct as a whole is not as cleanly defined as sadness or fear. Individuals may engage in more self-monitoring

XSL•FO
RenderX

about anger. Although Anger Affect will never achieve our clinically relevant target reliability with the reduced maximum stopping rule, it is not in our best interest to increase the test length to accommodate these "weaker" banks, as this will lead to increased response burden for all banks.

Given the competing priorities of improving reliability while minimizing burden, decisions for stopping rules should be adjusted depending on the context of the study population and the purpose of the study. The results of our study will be also relevant for other users who share a similar goal in achieving high reliability for use in clinical settings, as the NIH Toolbox has been increasingly used with clinical populations [28], while minimizing burden for nonsymptomatic individuals. When determining stopping rules, users should also consider the reliability achievable with the given item bank, the target reliability needed to answer their research or clinical questions, and the range of the domain in the sample population. Stopping rules should be modified after carefully considering these varied measurement contexts.

Limitations

Findings are limited to the 3 NIH Toolbox Emotion Battery adult item banks studied. As this study was designed to be largely a statistical investigation, our planned and post hoc stopping rules only incorporated psychometric considerations. These stopping rules were not applied to identical response data, although all simulations were based on the same uniform population distribution. To avoid artificial biases, we conducted the analyses in a way that prevented the person ability distribution from being simulated to match the item difficulty distribution. However, this may restrict the generalizability of our findings to real-world person distributions [29,30]. Indeed, using SE stopping rules in a multidimensional CAT setting can reduce test length by upward of 33% [31-33].

We did not explore multidimensional CATs for this study because those assessments are built on the assumption that all measures within a given composite (eg, Negative Affect subdomain) are important and should be assessed. While we expect clinicians and researchers would want to use all measures within a given subdomain, we chose not to impose that requirement. Future studies could extend the analyses to other subdomains and PROMs by considering multidimensional adaptive testing and theoretically driven content balancing techniques. As this analysis only examined adult responses, future extensions of this work could also examine the Emotion Battery for a pediatric population.

Conclusion

Taken together, these results demonstrate the challenges in attempting to reduce test burden while also achieving score precision for clinical use. The reduced maximum stopping rules provide one acceptable strategy to achieve these goals. The SE-change stopping rules and fixed-length stopping rules with less than 8 items are not recommended due to the sizable increase in simulations achieving empirical reliability <0.85. The reduced maximum stopping rules maximized the proportion of simulations achieving clinically relevant reliability (>0.95) without excessively inflating those achieving low reliability; they also decreased burden by allowing for fewer items to be administered when sufficient. That said, the reduced maximum stopping rules are not recommended for *all* occasions. Stopping rules should be modified in accordance with the context of the study population and the purpose of the study.

Acknowledgments

The authors thank the National Institutes of Health (NIH) for providing funding for this study. This study is funded in whole or in part with federal funds from the Blueprint for Neuroscience Research, NIH, under contract number HHS-N-260-2006-00007-C, by the Environmental Influences on Child Health Outcomes (ECHO) program, Office of the Director, NIH, under award number U24OD023319, and under contract number HHS-N-271-2018-0072-P.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Description of item banks.
[DOCX File, 15 KB - formative_v9i1e60215_app1.docx]

Multimedia Appendix 2
Correlation of the generating θ score with the estimated expected a posteriori score.
[DOCX File, 85 KB - formative_v9i1e60215_app2.docx]

Multimedia Appendix 3
Maximum attainable information curves for Anger Affect, Fear Affect, and Sadness.
[DOCX File, 361 KB - formative_v9i1e60215_app3.docx]

Multimedia Appendix 4

XSL•FO
RenderX

Comparisons of computerized adaptive test stopping rule efficiencies.
[DOCX File, 185 KB - formative_v9i1e60215_app4.docx]

References

1. Institute of Medicine (US) Committee on Quality of Health Care in America. Crossing the Quality Chasm: A New Health System for the 21st Century: National Academies Press (US); 2001. [Medline: 25057539]
2. Anker SD, Agewall S, Borggrefe M, et al. The importance of patient-reported outcomes: a call for their comprehensive integration in cardiovascular clinical trials. Eur Heart J 2014 Aug 7;35(30):2001-2009. [doi: 10.1093/eurheartj/ehu205] [Medline: 24904027]
3. Deshpande PR, Rajan S, Sudeepthi BL, Abdul Nazir CP. Patient-reported outcomes: a new era in clinical research. Perspect Clin Res 2011 Oct;2(4):137-144. [doi: 10.4103/2229-3485.86879] [Medline: 22145124]
4. Weldring T, Smith SMS. Patient-reported outcomes (PROs) and patient-reported outcome measures (PROMs). Health Serv Insights 2013;6:61-68. [doi: 10.4137/HSI.S11093] [Medline: 25114561]
5. Carlozzi NE, Kallen MA, Sander AM, et al. The development of a new computer adaptive test to evaluate anxiety in caregivers of individuals with traumatic brain injury: TBI-CareQOL Caregiver-Specific Anxiety. Arch Phys Med Rehabil 2019 Apr;100(4S):S22-S30. [doi: 10.1016/j.apmr.2018.05.027] [Medline: 29958902]
6. Smith AB, Hanbury A, Retzler J. Item banking and computer-adaptive testing in clinical trials: standing in sight of the PROMISed land. Contemp Clin Trials Commun 2018 Nov;13:100306. [doi: 10.1016/j.conctc.2018.11.005] [Medline: 30582066]
7. Gershon RC, Wagster MV, Hendrie HC, Fox NA, Cook KF, Nowinski CJ. NIH toolbox for assessment of neurological and behavioral function. Neurology (ECronicon) 2013 Mar 12;80(11 Suppl 3):S2-S6. [doi: 10.1212/WNL.0b013e3182872e5f] [Medline: 23479538]
8. Salsman JM, Butt Z, Pilkonis PA, et al. Emotion assessment using the NIH Toolbox. Neurology (ECronicon) 2013 Mar 12;80(11 Suppl 3):S76-S86. [doi: 10.1212/WNL.0b013e3182872e11] [Medline: 23479549]
9. Samejima F. Handbook of Moden Item Response Theory: Springer; 1997.
10. Chang CH, Reeve BB. Item response theory and its applications to patient-reported outcomes measurement. Eval Health Prof 2005 Sep;28(3):264-282. [doi: 10.1177/0163278705278275] [Medline: 16123257]
11. Nguyen TH, Han HR, Kim MT, Chan KS. An introduction to item response theory for patient-reported outcome measurement. Patient 2014;7(1):23-35. [doi: 10.1007/s40271-013-0041-0] [Medline: 24403095]
12. Stover AM, McLeod LD, Langer MM, Chen WH, Reeve BB. State of the psychometric methods: patient-reported outcome measure development and refinement using item response theory. J Patient Rep Outcomes 2019 Jul 30;3(1):50. [doi: 10.1186/s41687-019-0130-5] [Medline: 31359210]
13. Choi SW, Swartz RJ. Comparison of CAT item selection criteria for polytomous items. Appl Psychol Meas 2009 Sep 1;33(6):419-440. [doi: 10.1177/0146621608327801] [Medline: 20011456]
14. Lunz ME, Bergstrom BA, Gershon RC. Computer adaptive testing. Int J Educ Res 1994 Jan;21(6):623-634. [doi: 10.1016/0883-0355(94)90015-9]
15. van der Linden WJ. Bayesian item selection criteria for adaptive testing. Psychometrika 1998 Jun;63(2):201-216. [doi: 10.1007/BF02294775]
16. Wainer H. Computerized Adaptive Testing: A Primer, 2nd edition: Lawrence Erlbaum Associates; 2000.
17. Gershon R, Cella D, Dineen K, Rosenbloom S, Peterman A, Lai JS. Item response theory and health-related quality of life in cancer. Expert Rev Pharmacoecon Outcomes Res 2003 Dec;3(6):783-791. [doi: 10.1586/14737167.3.6.783] [Medline: 19807355]
18. Matteucci M, Veldkamp BP. On the use of MCMC computerized adaptive testing with empirical prior information to improve efficiency. Stat Methods Appl 2013 Jun;22(2):243-267. [doi: 10.1007/s10260-012-0216-1]
19. Frans N, Braeken J, Veldkamp BP, Paap MCS. Empirical priors in polytomous computerized adaptive tests: risks and rewards in clinical settings. Appl Psychol Meas 2023 Jan;47(1):48-63. [doi: 10.1177/01466216221124091] [Medline: 36425285]
20. Leung CK, Chang HH, Hau KT. Incorporation of content balancing requirements in stratification designs for computerized adaptive testing. Educ Psychol Meas 2003 Apr;63(2):257-270. [doi: 10.1177/0013164403251326]
21. Yao L. Comparing the performance of five multidimensional CAT selection procedures with different stopping rules. Appl Psychol Meas 2013 Jan;37(1):3-23. [doi: 10.1177/0146621612455687]
22. Kallen MA, Cook KF, Amtmann D, Knowlton E, Gershon RC. Grooming a CAT: customizing CAT administration rules to increase response efficiency in specific research and clinical settings. Qual Life Res 2018 Sep;27(9):2403-2413. [doi: 10.1007/s11136-018-1870-z] [Medline: 29730850]
23. Streiner DL, Norman GR, Cairney J. Health Measurement Scales: A Practical Guide to Their Development and Use: Oxford University Press; 2015.
24. Ho EH, Verkuilen J, Fischer F. Measuring individual true change with PROMIS using IRT-based plausible values. Qual Life Res 2023 May;32(5):1369-1379. [doi: 10.1007/s11136-022-03264-2] [Medline: 36282446]

XSL•FO
RenderX

25. Smith AW, Jensen RE. Beyond methods to applied research: realizing the vision of PROMIS®. Health Psychol 2019 May;38(5):347-350. [doi: 10.1037/hea0000752] [Medline: 31045416]
26. Yount SE, Cella D, Blozis S. PROMIS®: standardizing the patient voice in health psychology research and practice. Health Psychol 2019 May;38(5):343-346. [doi: 10.1037/hea0000741] [Medline: 31045415]
27. Choi SW. Handbook of Item Response Theory, 1st edition: Chapman and Hall/CRC; 2017.
28. Fox RS, Zhang M, Amagai S, et al. Uses of the NIH Toolbox® in clinical samples: a scoping review. Neurol Clin Pract 2022 Aug;12(4):307-319. [doi: 10.1212/CPJ.0000000000200060] [Medline: 36382124]
29. Paap MCS, Born S, Braeken J. Measurement efficiency for fixed-precision multidimensional computerized adaptive tests: comparing health measurement and educational testing using example banks. Appl Psychol Meas 2019 Jan;43(1):68-83. [doi: 10.1177/0146621618765719] [Medline: 30573935]
30. Yao L, Pommerich M, Segall DO. Using multidimensional CAT to administer a short, yet precise, screening test. Appl Psychol Meas 2014 Nov;38(8):614-631. [doi: 10.1177/0146621614541514]
31. Li YH, Schafer WD. Trait parameter recovery using multidimensional computerized adaptive testing in reading and mathematics. Appl Psychol Meas 2005 Jan;29(1):3-25. [doi: 10.1177/0146621604270667]
32. Luecht RM. Multidimensional computerized adaptive testing in a certification or licensure context. Appl Psychol Meas 1996 Dec;20(4):389-404. [doi: 10.1177/014662169602000406]
33. Paap MCS, Kroeze KA, Glas CAW, Terwee CB, van der Palen J, Veldkamp BP. Measuring patient-reported outcomes adaptively: multidimensionality matters!. Appl Psychol Meas 2018 Jul;42(5):327-342. [doi: 10.1177/0146621617733954] [Medline: 29962559]

Abbreviations

CAT: computerized adaptive test
EAP: expected a posteriori
IRT: item response theory
PROM: patient-reported outcome measure
RMSE: root mean squared error

Edited by A Mavragani; submitted 04.05.24; peer-reviewed by MAJ Luijten, N Guler, N Frans; revised version received 16.12.24; accepted 17.12.24; published 21.03.25.

Please cite as:
Amagai S, Kaat AJ, Fox RS, Ho EH, Pila S, Kallen MA, Schalet BD, Nowinski CJ, Gershon RC
Customizing Computerized Adaptive Test Stopping Rules for Clinical Settings Using the Negative Affect Subdomain of the NIH Toolbox Emotion Battery: Simulation Study
JMIR Form Res 2025;9:e60215
URL: https://formative.jmir.org/2025/1/e60215
doi:10.2196/60215

© Saki Amagai, Aaron J Kaat, Rina S Fox, Emily H Ho, Sarah Pila, Michael A Kallen, Benjamin D Schalet, Cindy J Nowinski, Richard C Gershon. Originally published in JMIR Formative Research (https://formative.jmir.org), 21.3.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

“Your Life, Your Health: Tips and Information for Health and Well-Being”: Development of a World Health Organization Digital Resource to Support Universal Access to Trustworthy Health Information

Danielle M Muscat[1], PhD; Rachael Hinton[2], PhD; Shyama Kuruvilla[2], PhD; Don Nutbeam[3], PhD

[1]Sydney Health Literacy Lab, Sydney School of Public Health, Faculty of Medicine and Health, University of Sydney, Sydney, Australia
[2]World Health Organization, Avenue Appia 20, Geneva, Switzerland
[3]Sydney School of Public Health, Faculty of Medicine and Health, University of Sydney, Sydney, Australia

Corresponding Author:
Shyama Kuruvilla, PhD
World Health Organization, Avenue Appia 20, Geneva, Switzerland

Abstract

Background: Access to trustworthy, understandable, and actionable health information is a key determinant of health and is an essential component of universal health coverage and primary health care. The World Health Organization has developed a new digital resource for the general public to improve health and well-being across different life phases and to support people in caring for themselves, their families, and their communities. The goal was to make trustworthy health information accessible, understandable, and actionable for the general public in a digital format and at the global scale.

Objective: The aim of this paper was to describe the multistage approach and methodology used to develop the resource *Your life, your health: Tips and information for health and well-being* (hereafter, *Your life, your health*).

Methods: A 5-step process was used to develop *Your life, your health*, including (1) reviewing and synthesizing existing World Health Organization technical guidance, member state health and health literacy plans, and international human rights frameworks to identify priority messages; (2) developing messages and graphics that are accessible, understandable, and actionable for the public using health literacy principles; (3) engaging with experts and stakeholders to refine messages and message delivery; (4) presenting priority content in an accessible digital format; and (5) adapting the resource based on feedback and new evidences.

Results: The *Your life, your health* online resource adopts a life-course approach to organize health information based on priority actions and rights that support peoples’ health and well-being across different life stages and specific health topics. The resource promotes health literacy by offering advice on asking questions to health workers, making informed decisions about personal and family health, and effectively using digital media to obtain reliable health information. Additionally, it reflects the ambitions of the Sustainable Development Goals by providing essential information on the social determinants of health and clarifies the distinct roles of individuals, frontline workers, governments, and the media in promoting and protecting health.

Conclusions: Making health information available—including to the public—is an essential step in strengthening the global health information system. The development process for the *Your life, your health* online resource outlined in this article offers a structured approach to translate technical health guidelines into accessible, understandable, and actionable health information for the general public.

(JMIR Form Res 2025;9:e57881) doi:10.2196/57881

KEYWORDS

health communication; health literacy; consumer health information; digital health; universal health care

Introduction

Background

Access to reliable, understandable, and actionable health information is a determinant of health and an essential component of universal health coverage (UHC) [1-3]. Similar to other social determinants of health, there is significant variability and inequity in access to health information and in peoples’ capacity to retrieve, appraise, and use it in making health-related decisions, also referred to as health literacy [4]. Differences in access to health information and levels of health literacy are observed both within and between countries [5-8].

Digital technology has significantly improved access to health information for many people. Several reviews of digital health

XSL•FO
RenderX

interventions have demonstrated the feasibility and potential effectiveness of eHealth and mHealth (mobile health) interventions for a range of health conditions in low-, middle-, and high-income countries [9-13]. However, the digital divide is real and significantly impacts peoples' access to health information [14] and perpetuates many existing health inequities. Additionally, health information accessed from digital platforms (ie, any electronic communication tool, including websites and social media) is not consistently reliable, understandable, and actionable.

Reliable health information originates from a trusted source and is based on an objective interpretation of available scientific evidence [1]. To be understandable and actionable, members of the public with diverse backgrounds and varying levels of health literacy must be able to understand key messages and identify appropriate actions based on the information presented [15]. However, these fundamental requirements of reliability, understandability, and actionability are not always met. This issue became particularly evident during the COVID-19 pandemic, when unintended misinformation and deliberately circulated disinformation were widely available and often amplified by social media [16-18]. A significant proportion of information sources are also biased, sometimes unconsciously but often deliberately, particularly in commercial communications designed to promote specific products or services [1]. Even when health information originates from a reliable, trustworthy source, it is rarely presented in a manner that most people can understand and act upon when making health-related decisions [19,20].

In response to the global challenge of misinformation and disinformation during the COVID-19 pandemic, the World Health Organization (WHO) developed a Framework for Managing the COVID-19 Infodemic. This framework proposed various actions that WHO member states and the civil society could implement to address the infodemic. This included clear guidance on ensuring that health information was provided in a manner that was "actionable" and "presented in ways that are understood by and accessible to all individuals in all parts of all societies" [21]. This framework recognized the important role of WHO, other United Nations (UN) partners, and civil society organizations in providing reliable health information that is adaptable for use in a wide range of contexts.

Long before this, the 1946 WHO Constitution specified that "informed opinion and active cooperation on the part of the public are of the utmost importance in the improvement of the health of the people" [22]. Despite this historical mandate, WHO has generally limited its health communication activities to providing technical advice and support to member states. These communications are not systematically designed for direct use by the general public [20]. Over the past 25 years, this has been gradually changing and WHO has developed global health information resources that can be accessed and used more directly by the general public, such as WHO's Facts in Pictures [23], Global Health Days [24], and public health advice on COVID-19 [25]. However, at present, much of WHO's publicly available information tends to be organized by different health topics or is limited to targeted communication campaigns or events. Therefore, it is not always easily accessible and is not available directly to the general public in a consistent form or integrated manner.

To address this gap, WHO has developed an integrated public health advice resource, *Your life, your health: Tips and information for health and well-being* [2] (hereafter referred to in short form as *Your life, your health*). This new digital resource, specifically developed for the general public, has been designed to make trustworthy health information accessible, understandable, and actionable, contributing toward achieving UHC. The aim of this paper is to describe the multistage process and methodology used to develop this resource.

Theoretical Framework

The theoretical underpinning for the resource was a UHC and life-course approach to health, based on the WHO Life Course Framework [26]. This framework emphasizes that strategies to achieve UHC must be explicitly people-centered and require integrated, evidence-based approaches across the life course. The WHO's Thirteen General Programme of Work included a dedicated platform on Improving Human Capital across the Life Course, within which this initiative was situated [27].

Additionally, we applied the principles of Universal Design for Learning (UDL) to the development of the resource. UDL is a theoretical paradigm that informs efforts to promote learning for the greatest possible proportion of a population [28]. It is an inclusive learning environment that is universally designed, considers all relevant dimensions of diversity, and proactively minimizes barriers to learning for all individuals [28]. This perspective recognizes that learning occurs in a dynamic interaction between the individual and learning environment. UDL leverages the flexibility of digital technology to design learning environments that, from the outset, offer options for diverse learner needs, ensuring that educational benefits are more equitably and effectively distributed [28]. In the context of health, health literacy universal precautions similarly assume that all patients and caregivers may have difficulty comprehending health information and recommend communicating in ways that ensure universal understanding [29,30].

Methods

Development of the "*Your Life, Your Health*" Resource

A 5-step process was used to develop *Your life, your health* [2]:

- Step 1 – A review and synthesis of existing WHO technical guidance, member state health and health literacy plans, and international human rights frameworks to identify evidence-based health guidance for the public;
- Step 2 – The development of messages and graphics to be accessible, understandable, and actionable for the public through the application of health literacy principles;
- Step 3 – Engagement with experts and other stakeholders to refine messages and message delivery;
- Step 4 – Presentation of priority content in an accessible digital format;
- Step 5 – Adaptation and updates based on feedback and new evidence.

XSL•FO
RenderX

Below, we outline the methods related to the first four steps in detail. Similar to WHO guidelines, the resource will evolve and will be updated as new evidence becomes available. It will also need to be adapted for different country contexts and for different audiences in Step 5.

Ethical Considerations

Ethical approval was not required as this study is based exclusively on published literature, including global and national strategies and plans, with no data collected from human participants.

Review of Evidence-Based Interventions and Guidance

Overview

The first stage of resource development was focused on identifying reliable health information to be included as the basis for *Your life, your health.* There is an overwhelming volume of health information accessible to most people. We sought to identify evidence–based public health actions that individuals or communities could undertake to care for themselves, their families, and their community—either with or without the support of the health system—to improve health and well-being across different life phases. We then prioritized this information by life phase to reduce the volume of information provided in the resource. While making decisions on priorities, we identified interventions or actions within current global and national health strategies that required or presumed an element of individual, caregiver, family, or community initiative or engagement. Interventions or actions that were exclusively health provider-oriented or related to specialized clinical and surgical interventions and care were mostly excluded from the review process.

To identify priority actions, we purposively searched 21 WHO global health strategies across the life course, and additional strategies, guidance, and evidence compendia suggested by WHO partners. This included strategies for specific target populations and life phases (eg, WHO Global Strategy for Women's, Children's and Adolescent's Health), communicable and noncommunicable diseases (eg, WHO Global Action Plan for the Prevention and Control of Noncommunicable Diseases) and cross-cutting strategies on systems, principles, and determinants (eg, Healthy Systems for Universal Health Coverage: A Joint Vision for Healthy Lives). A full list is available in Table S1 in Multimedia Appendix 1.

We also searched the WHO UHC Compendium of Interventions and the Disease Control Priorities 3rd Edition (DCP3) interventions (community worker and health post levels) [31,32]. The UHC Compendium is a database of health services and intersectoral interventions designed to assist countries in making progress towards UHC. The database includes over 3500 health actions across different health areas. The DCP3 provides an up-to-date comprehensive review of priority health interventions and intersectoral policies to reduce mortality and disability which are evidence-based, scalable, and adaptable in multiple settings. Information from the Compendium and DCP3 was cross-checked against the main global causes of mortality and morbidity at each life stage to ensure that the reviewed evidence-based guidance and interventions addressed these challenges.

Finally, we searched for contemporary national and regional health strategies, action plans, and road maps from all 194 WHO member state countries and 6 regions, with a focus on documents available in English and French. These documents were sourced primarily via WHO's Country Planning Cycle Database [33]. If a current document was unavailable in the Planning Cycle Database, a search was carried out on Google or PubMed using the following search string: *[((“National”) AND (“Health”)) AND ((“Strategy”) OR (“Road Map”) OR (“Action Plan”)) AND (Name of Country or Region)]*. We considered “contemporary” to mean any strategic document informing planning in a member state beyond 2019. We found relevant information from 84 countries across the 6 WHO regions, which is shown in Multimedia Appendix 2.

The evidence searches, review, and syntheses were led by SK, supported by RH and Global Health Insights, a specialist global health research and consulting collaborative, with experience in evidence synthesis and policy analysis.

Synthesis of Evidence

Using the documents identified in the evidence review, we listed the interventions and actions that required or presumed an element of individual, caregiver, family, or community initiative or engagement. Using the content analysis approach described by Mays, Pope, and Popay [34], we deductively classified these actions and interventions according to broad life phases adapted from the WHO Life Course Framework [26]: pregnancy, birth, and after childbirth; newborns and children under 5 years; children aged 5-9 years; adolescents and youth aged 10 - 24 years; early and middle adulthood (25 - 64 years); and later adulthood (≥65 years).

Within each life phase, interventions and actions were coded into topic areas derived from the broad health topics on the WHO website [35]. As we reviewed the content, we added further health topics until the saturation of categories was achieved. The final categories of health topics are shown in Textbox 1. Content analysis was then applied to identify recurrent themes arising from the tabulated content within each health topic.

XSL•FO
RenderX

Textbox 1. Final categories of health topics.

- Care and communication in families
- Nutrition
- Physical activity
- Mental health
- Immunization
- Sexual and reproductive health (including family planning)
- Healthy environments
- Social participation
- Violence and injuries
- Smoking, alcohol, and drugs
- Disabilities
- Connecting to services in health and other sectors
- Emergency illness
- Communicable disease
- Noncommunicable disease

Finally, we applied the international human rights framework and a categorization of human rights by the Sustainable Development Goals to develop a matrix of human rights categorized by life phase (eg, rights related to having children) and cross-cutting rights (eg, the right to health). This approach assured that we moved beyond providing only information that required individual action, to ensure that people are informed about their rights, obligations of the government, and personal responsibilities.

Using this methodology, we had reasonable confidence in the quality and consistency of the evidence and were able to use the findings as the basis for prioritization. To keep the volume of information manageable, we aimed to identify 5 essential health messages for each stage in the life course or specific health topic. A health message refers to actionable information related to health and well-being that individuals or communities can use to care for themselves, their families, and their communities (eg, helping children learn to swim). We identified the messages that had the highest potential for promoting health and those most frequently mentioned across the collated information. Inevitably, some issues of importance to individuals, communities, and countries were not included, and people's rights to health may vary between countries. We recognize the importance of these contextual considerations and have included them into our pilot testing and adaptation actions as indicated below.

Application of Health Literacy Principles to Make Content Accessible, Understandable, and Actionable

Following the prioritization of health messages, the next step was to draft the main messages and design the resource's information architecture in a manner that was optimally accessible, understandable, and actionable. We reviewed the national health and health literacy plans of 84 countries with publicly accessible documentation available in English or French to understand how countries developed public-facing messages that were aligned with WHO's evidence-based guidance and global strategies. We made selective use of a range of widely used tools and resources to inform decisions about the presentation of information once content was approved within WHO. These included the Health Literacy Universal Precautions Toolkit [36], the CDC Clear Communication Index, Health Literacy Online [37], and the Patient Education Materials Assessment Tool. All content was drafted by a WHO consultant with expertise in health communication, and technical and copy editing, who was trained in health literacy principles.

Review by Health Specialists and Communication and IT Experts

We developed a web-based prototype of the resource for consultation. Between October 2019 and March 2020, we conducted consultations regarding the draft resource with over 30 experts in health, health literacy, communication, and IT. This included consultation and review across WHO departments, including life course, health workforce, reproductive health and rights, environmental health, health promotion and health literacy, and communications, among others. We sought feedback via an online survey on the content (see Multimedia Appendix 3), particularly on the choice of prioritized actions and rights, and advice on the format and presentation of the messages and potential uses and users. We also sought feedback on various graphics related to the main messages to ascertain the most appropriate format for potential users. Additionally, content experts from WHO departments provided comments on both the simple and more detailed messages relevant to their health topic area to ensure alignment with WHO guidance and existing evidence. Finally, we consulted communication and IT experts within WHO on optimizing the digital platform to support the use and dissemination of the resource, depending on how people use the WHO website. This included the potential for translation of the content into other languages and graphics adaptation for various audiences and communication channels.

XSL•FO
RenderX

Feedback was collated and analyzed thematically by the first author and compiled into an action report for WHO to finalize priority content and graphics. In instances where experts had provided conflicting advice, all verbatim feedback provided was included, and final decisions were made by team consensus with all authors.

Presentation of Priority Content in an Accessible Digital Format

Following the consultation process, the content was revised to incorporate feedback on prioritized actions and rights, while also aiming to maintain a readability grade of 8 or lower (as assessed by the Hemingway Editor). The WHO communications and web team supported the development of the resource for the WHO website.

Results

Structure

Through the multistage process outlined above, *Your life, your health*: *Tips and information for health and well-being* online resource was designed to be accessible, understandable, and actionable. This resource has five key sections, accessible via the landing page:

1. Tips and information by life phase
2. How to find and use health information
3. A healthy world
4. Know your rights
5. Other health topics

This resource adopts WHO's life-course approach [26] to organize health-related tips and information (Section 1) and rights (Section 4) based on priority actions and rights that support peoples' health and well-being at different life phases. The consultation process also confirmed that health information should not be focused entirely on individual behaviors and risks, but should reflect how social determinants of health fundamentally influence peoples' choice of actions and their capacity, more accurately to engage successfully in responding to information. For people in most parts of the world, health decisions and actions are substantially restricted by social and environmental factors beyond their control; peoples' ability to respond to health information is not universal and equal. In addition, we recognized that people rarely make health decisions in isolation but rather through interaction and support from other people, institutions, organizations, services, and systems. Given this context, Section 3*: A Healthy World* was integrated to reflect the ambitions of the Sustainable Development Goals by providing information on the social determinants of health and clarifying the roles of individuals, frontline workers, governments, and the media in promoting and protecting health. Section 2: *How to find and use health information* was also included to more overtly support users in developing health literacy skills. It includes advice on asking questions to health workers, making informed decisions about personal and family health, and effectively using digital media to access health information (Figure 1).

Figure 1. Screenshots representing the structure of three sections of the "Your life, your health: Tips and information for health and well-being" online resource.

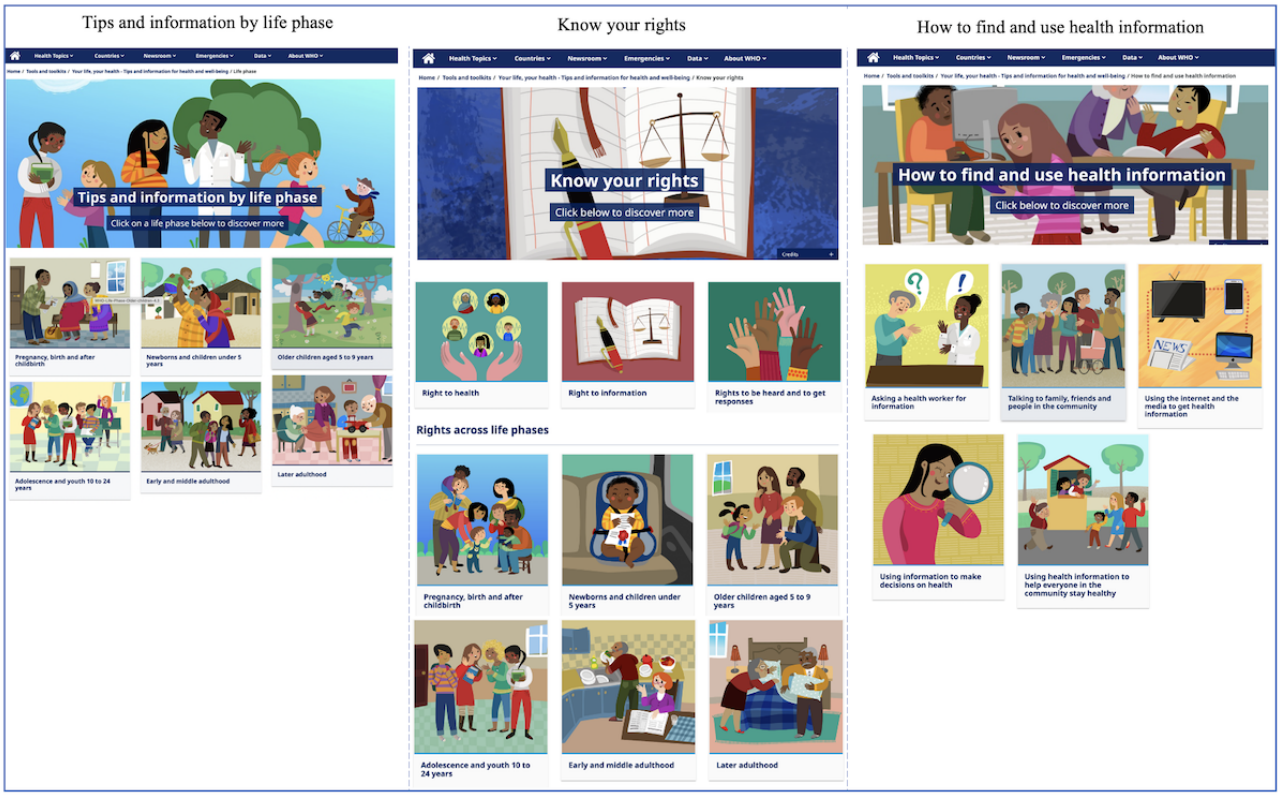

XSL•FO
RenderX

Content and Presentation

Within each section, we adopted a tiered structure, where information was sequenced from simple to more detailed messages at a level of Grade 8 readability or lower. This would allow the readers to move from the first level consisting of 5 main messages to a second level where each message is broken down into manageable and explicit steps. Further information and more detailed WHO technical content and products could then be made available via links to WHO infographics, videos, and fact sheets depending on needs and preferences. To enhance actionability, each message started with a verb and was worded in a direct way (eg, "Learn about..." "Find out…"). Content cues or markers (eg, bullet points) were also used to highlight or draw attention to each point and 3 styles of frequently used visuals for different life phases and messages (ie, photos, illustrations, and infographics) were developed. Figure 2 shows a screenshot of the *Your life, your health* online resource showcasing select design elements.

Figure 2. Screenshot from the "Your life, your health: Tips and information for health and well-being" online resource showcasing select design elements from Health Literacy Online and the Patient Education Materials Assessment Tool (PEMAT).

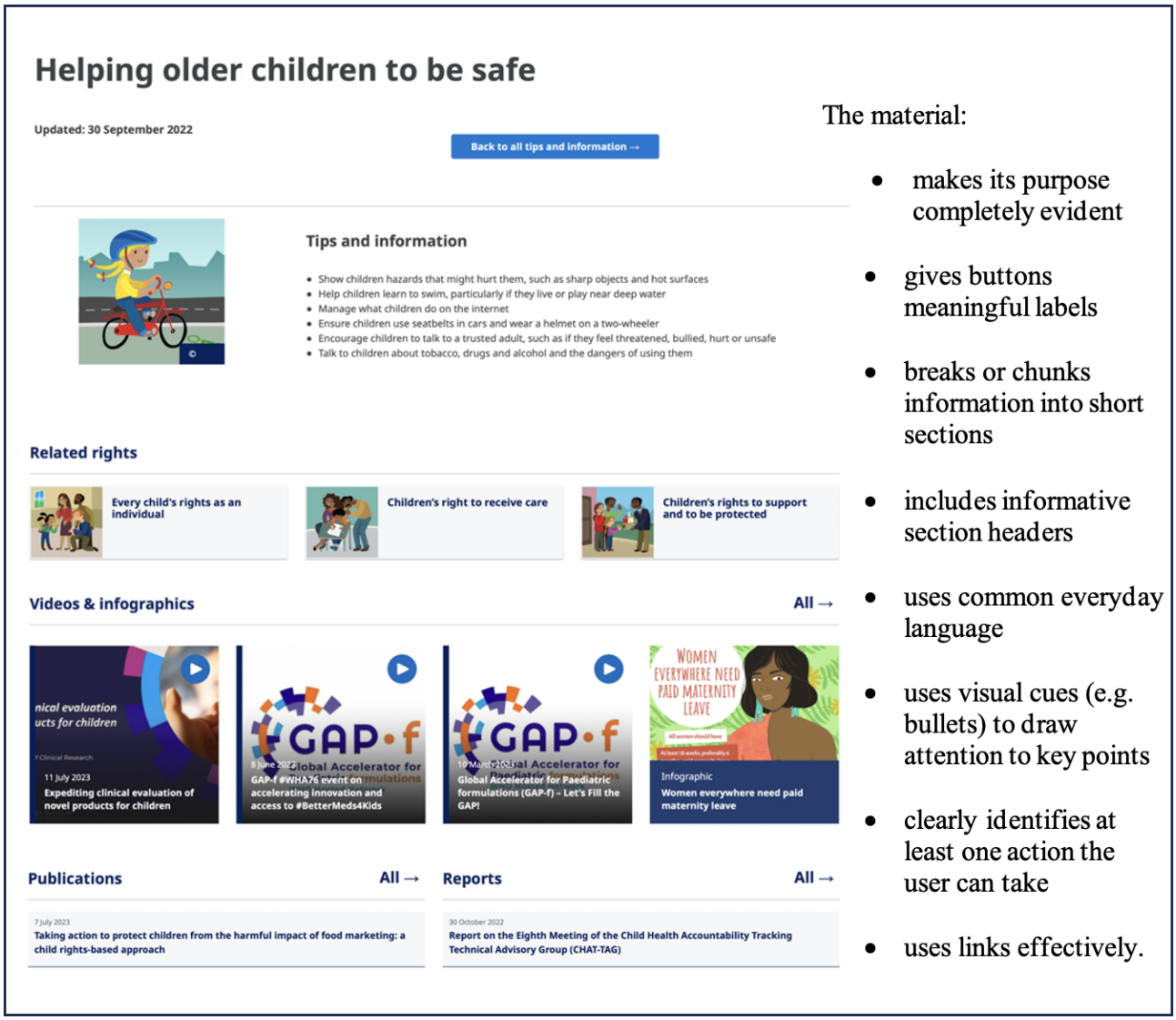

Hosting and Dissemination

As a WHO global good initiative, *Your life, your health* is publicly available to all. The resource is available on the WHO website and is linked to Health Topics [38], for easy access to public health information. To increase accessibility, health advice within the resource is available as downloadable and shareable digital cards with links to expanded messages and resources (See Multimedia Appendix 4). The planned next steps include engaging with the WHO academy to link the resource content to relevant courses and topics for learners from the global health workforce. Other adaptations include integrating *Your life, your health* in home-based records, such as integrated maternal and child health books, which often include health education messages, as well as using content on the WHO Facebook (chatbot) page [39]. Similar to WHO guidelines, the resource will evolve and be updated as new evidence becomes available. Given the synthesis approach adopted in the development of this resource, this will be aligned with updates from the WHO UHC Compendium of Interventions and the Disease Control Priorities. Standard practice steps and a template have been developed to ensure the health literacy principles described here (eg, tiered presentation of information, testing readability) and will continue to provide a guiding framework

XSL•FO
RenderX

as new information is integrated. A guidance document has also been developed for WHO departments for developing new content in *Your life, your health* in a standardized format for the WHO website.

Discussion

As a contribution to improving universal access to trustworthy health information, the WHO has developed an integrated resource of health advice for use by and with the general public. This resource is based on a synthesis of evidence-based guidance derived from the WHO and national-level documents. It is organized to reflect priority actions and rights across different life phases while concurrently seeking to build health literacy skills, and acknowledge and promote awareness of the social determinants of health. As a global resource, *Your life, your health* is intended to complement country-based initiatives such as the US government's MyHealthFinder or HealthDirect in Australia, which are also evidence-based, easy to read, and action oriented.

Making health information available—including to the public—is an essential step in the global health information system [1]. While developing consumer friendly health information is not a new concept [40,41], there is limited published guidance on how public-facing health information resources are developed, or how to achieve this on a global scale. The development process for *Your life, your health* as outlined in this article, includes reviews of technical guidance and health strategies, the application of health literacy principles, and an expert review and iterative consultation process. It offers a structure that can be adapted for related purposes within and beyond the WHO to translate technical health guidelines into accessible, understandable, and actionable health information for the general public. The development process was built on several well-established health literacy guidelines related to structuring, writing, and visually presenting health information. In addition to the resources referenced above (ie, Health Literacy Universal Precautions Toolkit [32], CDC Clear Communication Index, Health Literacy Online [33]; Patient Education Materials Assessment Tool), there are several other resources more broadly related to universal design. Future research and practice should focus on the selection of resources that are validated and accurate, given inconsistencies across some automated health literacy tools [42]. The selection of guiding resources will also necessarily be context-specific, as mandated guidelines exist across country contexts (eg, US Federal Plain Language Guidelines and Web Content Accessibility Guidelines).

There are strengths and limitations to the development process of WHO's *Your life, your health* resource. We conducted an evidence synthesis to identify health content, recommendations, and human rights across the life course, including both WHO and external sources. Given the breadth of the resource and scope of health topics included in *Your life, your health,* we reviewed existing global strategies and action plans that are publicly accessible and that already synthesize evidence-based recommendations. By ensuring transparency in the sourcing of evidence and identifying core, consistently advocated health information, we aimed to ensure that the information provided is recognizable, reliable, and trustworthy for those using the resource. This synthesis approach was more efficient than a review of primary studies and allowed for the identification of evidence and recommendations that were consistent across strategies and action plans. While we recognize that this methodology will result in the exclusion of some types of evidence—including newer or emerging evidence which is yet to be incorporated into global strategies—we determined that this approach offered the most secure method for ensuring trust in the information provided, and was supported by expert consultation and review. Additionally, while the exclusion criteria for language were a limitation in our evidence review, the review covered all WHO regions, many of which had national language and English versions available for their strategies.

Our expert review of the content and structure purposefully included diverse expertise, ranging from health literacy to IT; however, it has not yet included community members. Following the publication of this resource, we have initiated a consultative process to iteratively refine its content. This will be followed up by additional co-development and user testing, including with WHO representatives at the regional and country levels, UN agencies, and the general public. Partnering with the general public will result in incremental improvements to the understandability of health information materials [39,43]; the benefits of engaging stakeholders are well-documented and will likely include enhanced relevance, quality, and authenticity, the generation of alternative and innovative ideas and enhanced sustainability of the resource [44]. Adaptation for country-specific contexts will ensure that the resource is culturally acceptable, aligned with local health practices availability of health resources and services, and actionable in a variety of contexts [1]. User testing must include people with variations on literacy, health literacy, and cognitive and functional capacities. The consultations will need to consider feasibility, potential effectiveness, and necessary adaptations of this online digital resource in different settings. It will be important to balance digital engagement with appropriately resourced in-person engagement to ensure that vulnerable groups and those without access to digital channels are not excluded. The effectiveness of the multistage process outlined in this paper will be determined by future user testing to confirm that it can deliver clear and actionable information to those in need.

The current rising momentum and expansion of digital technologies in the UHC era offers a significant opportunity to provide more equitable access to information globally. The *Your life, your health* resource is the first of its kind to be developed by WHO. It aims to consolidate technical and often disparate health information available to the general public in an integrated, understandable, and actionable way. The next step will be to focus on user testing of the resource prior to implementation at the global, regional, and country levels.

XSL•FO
RenderX

Disclaimer

The named authors alone are responsible for the views expressed in this publication that does not necessarily represent the decisions or the policies of the World Health Organization.

Data Availability

Data sharing is not applicable to this article as no datasets were generated or analyzed during this study.

Authors' Contributions

Conceptualization: DMM, DN, RH, SK
Data curation: RH
Formal analysis: DMM, RH, SK
Funding acquisition: SK
Investigation: RH, SK
Methodology: DN, DMM, RH, SK
Project administration: RH
Resources: SK
Writing – original draft: DMM
Writing – review & editing: DMM, DN, RH, SK

Conflicts of Interest

DMM is a codirector of a health literacy consultancy company, Health Literacy Solutions PTY Ltd. DMM takes no personal income from Health Literacy Solutions PTY Ltd.

Multimedia Appendix 1
Global strategies, action plans, and guidance included in the evidence review.
[DOCX File, 14 KB - formative_v9i1e57881_app1.docx]

Multimedia Appendix 2
Countries for which national health strategies were included in the evidence review.
[DOCX File, 338 KB - formative_v9i1e57881_app2.docx]

Multimedia Appendix 3
“Actions and rights for healthy lives” consultation feedback form.
[DOCX File, 23 KB - formative_v9i1e57881_app3.docx]

Multimedia Appendix 4
Digital cards adapted from the “Your life, your health: Tips and information for health and well-being” online resource.
[DOCX File, 226 KB - formative_v9i1e57881_app4.docx]

References

1. Royston G, Pakenham-Walsh N, Zielinski C. Universal access to essential health information: accelerating progress towards universal health coverage and other SDG health targets. BMJ Glob Health 2020 May;5(5):e002475. [doi: 10.1136/bmjgh-2020-002475] [Medline: 32424012]
2. Muscat DM, Hinton R, Nutbeam D, Kenney E, Kuruvilla S, Jakab Z. Universal health information is essential for universal health coverage. Fam Med Com Health 2023 May;11(2):e002090. [doi: 10.1136/fmch-2022-002090]
3. National action plan to improve health literacy. Office of Disease Prevention and Health Promotion. 2010. URL: https://odphp.health.gov/our-work/national-health-initiatives/health-literacy/national-action-plan-improve-health-literacy [accessed 2023-09-07]
4. Nutbeam D, Muscat DM. Health Promotion Glossary 2021. Health Promot Int 2021 Dec 23;36(6):1578-1598. [doi: 10.1093/heapro/daaa157] [Medline: 33822939]
5. Aronson B. Improving online access to medical information for low-income countries. N Engl J Med 2004 Mar 4;350(10):966-968. [doi: 10.1056/NEJMp048009] [Medline: 14999107]
6. Chen X, Orom H, Hay JL, et al. Differences in rural and urban health information access and use. J Rural Health 2019 Jun;35(3):405-417. [doi: 10.1111/jrh.12335] [Medline: 30444935]
7. Pakenham-Walsh N, Bukachi F. Information needs of health care workers in developing countries: a literature review with a focus on Africa. Hum Resour Health 2009 Apr 8;7(1):30. [doi: 10.1186/1478-4491-7-30] [Medline: 19356239]

XSL•FO
RenderX

8. Sørensen K, Pelikan JM, Röthlin F, et al. Health literacy in Europe: comparative results of the European health literacy survey (HLS-EU). Eur J Public Health 2015 Dec;25(6):1053-1058. [doi: 10.1093/eurpub/ckv043] [Medline: 25843827]
9. Hall CS, Fottrell E, Wilkinson S, Byass P. Assessing the impact of mHealth interventions in low- and middle-income countries--what has been shown to work? Glob Health Action 2014;7(1):25606-25606. [doi: 10.3402/gha.v7.25606] [Medline: 25361730]
10. Lee SH, Nurmatov UB, Nwaru BI, Mukherjee M, Grant L, Pagliari C. Effectiveness of mHealth interventions for maternal, newborn and child health in low- and middle-income countries: systematic review and meta-analysis. J Glob Health 2016 Jun;6(1):010401. [doi: 10.7189/jogh.06.010401] [Medline: 26649177]
11. Sondaal SFV, Browne JL, Amoakoh-Coleman M, et al. Assessing the effect of mHealth interventions in improving maternal and neonatal care in low- and middle-income countries: a systematic review. PLoS One 2016;11(5):e0154664. [doi: 10.1371/journal.pone.0154664] [Medline: 27144393]
12. Stephani V, Opoku D, Quentin W. A systematic review of randomized controlled trials of mHealth interventions against non-communicable diseases in developing countries. BMC Public Health 2016 Jul 15;16(1):572. [doi: 10.1186/s12889-016-3226-3] [Medline: 27417513]
13. Müller AM, Alley S, Schoeppe S, Vandelanotte C. The effectiveness of e-& mHealth interventions to promote physical activity and healthy diets in developing countries: a systematic review. Int J Behav Nutr Phys Act 2016 Oct 10;13(1):109. [doi: 10.1186/s12966-016-0434-2] [Medline: 27724911]
14. Reddy H, Joshi S, Joshi A, Wagh V. A critical review of global digital divide and the role of technology in healthcare. Cureus 2022 Sep;14(9):e29739. [doi: 10.7759/cureus.29739] [Medline: 36340549]
15. Shoemaker SJ, Wolf MS, Brach C. Development of the Patient Education Materials Assessment Tool (PEMAT): a new measure of understandability and actionability for print and audiovisual patient information. Patient Educ Couns 2014 Sep;96(3):395-403. [doi: 10.1016/j.pec.2014.05.027] [Medline: 24973195]
16. Muhammed T S, Mathew SK. The disaster of misinformation: a review of research in social media. Int J Data Sci Anal 2022 May;13(4):271-285. [doi: 10.1007/s41060-022-00311-6]
17. Borges do Nascimento IJ, Pizarro AB, Almeida JM, et al. Infodemics and health misinformation: a systematic review of reviews. Bull World Health Organ 2022 Sep 1;100(9):544-561. [doi: 10.2471/BLT.21.287654] [Medline: 36062247]
18. Bin Naeem S, Kamel Boulos MN. COVID-19 misinformation online and health literacy: a brief overview. Int J Environ Res Public Health 2021 Jul 30;18(15):8091. [doi: 10.3390/ijerph18158091] [Medline: 34360384]
19. Mishra V, Dexter JP. Comparison of readability of official public health information about COVID-19 on websites of international agencies and the governments of 15 countries. JAMA Netw Open 2020 Aug 3;3(8):e2018033. [doi: 10.1001/jamanetworkopen.2020.18033] [Medline: 32809028]
20. Mac OA, Muscat DM, Ayre J, Patel P, McCaffery KJ. The readability of official public health information on COVID-19. Med J Aust 2021 Oct 18;215(8):373-375. [doi: 10.5694/mja2.51282] [Medline: 34580878]
21. Tangcharoensathien V, Calleja N, Nguyen T, et al. Framework for managing the COVID-19 infodemic: methods and results of an online, crowdsourced WHO technical consultation. J Med Internet Res 2020 Jun 26;22(6):e19659. [doi: 10.2196/19659] [Medline: 32558655]
22. Constitution of the World Health Organization. : World Health Organization; 1946 URL: https://www.who.int/about/governance/constitution [accessed 2025-02-25]
23. Facts in Pictures 2023. World Health Organization. URL: https://www.who.int/news-room/facts-in-pictures [accessed 2023-09-07]
24. Global health days 2023. World Health Organization. URL: https://www.who.int/campaigns [accessed 2023-09-07]
25. Advice for the public: coronavirus disease (COVID-19). World Health Organization. 2023. URL: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public [accessed 2023-09-07]
26. Kuruvilla S, Sadana R, Montesinos EV, et al. A life-course approach to health: synergy with sustainable development goals. Bull World Health Organ 2018 Jan 1;96(1):42-50. [doi: 10.2471/BLT.17.198358] [Medline: 29403099]
27. Thirteenth general programme of work 2019-2023. World Health Organization. 2019. URL: https://www.who.int/about/general-programme-of-work/thirteenth [accessed 2024-10-08]
28. Meyer A, Rose DH, Gordon D. Universal Design for Learning: Theory and Practice: CAST Professional Publishing; 2014. URL: https://publishing.cast.org/catalog/books-products/universal-design-for-learning-meyer-rose-gordon [accessed 2025-02-25]
29. DeWalt DA, Broucksou KA, Hawk V, et al. Developing and testing the health literacy universal precautions toolkit. Nurs Outlook 2011;59(2):85-94. [doi: 10.1016/j.outlook.2010.12.002] [Medline: 21402204]
30. Mabachi NM, Cifuentes M, Barnard J, et al. Demonstration of the health literacy universal precautions toolkit: lessons for quality improvement. J Ambul Care Manage 2016;39(3):199-208. [doi: 10.1097/JAC.0000000000000102] [Medline: 27232681]
31. UHC compendium: health interventions for universal health coverage. World Health Organization. 2023. URL: https://www.who.int/universal-health-coverage/compendium [accessed 2023-07-06]
32. DCP3 Disease Control Priorities. University of Washington Department of Global Health. 2021. URL: https://dcp-3.org [accessed 2023-07-06]

XSL•FO
RenderX

33. Country planning cycles database. World Health Organization. 2023. URL: https://extranet.who.int/countryplanningcycles [accessed 2023-07-06]
34. Mays N, Pope C, Popay J. Systematically reviewing qualitative and quantitative evidence to inform management and policy-making in the health field. J Health Serv Res Policy 2005 Jul;10 Suppl 1(1_suppl):6-20. [doi: 10.1258/1355819054308576] [Medline: 16053580]
35. Health Topics. World Health Organization. 2024. URL: https://www.who.int/health-topics [accessed 2024-02-27]
36. AHRQ health literacy universal precautions toolkit. Agency for Healthcare Research and Quality. 2020. URL: https://www.ahrq.gov/health-literacy/improve/precautions/index.html [accessed 2023-05-17]
37. Health literacy online. Office of Disease Prevention and Health Promotion. 2016. URL: https://health.gov/healthliteracyonline [accessed 2023-07-03]
38. Health topics. World Health Organization. URL: https://www.who.int/health-topics [accessed 2025-02-20]
39. Crawford MJ, Rutter D, Manley C, et al. Systematic review of involving patients in the planning and development of health care. BMJ 2002 Nov 30;325(7375):1263. [doi: 10.1136/bmj.325.7375.1263] [Medline: 12458240]
40. MyHealthfinder. Department of Health and Human Services Office of Disease Prevention and Health Promotion. URL: https://odphp.health.gov/myhealthfinder [accessed 2024-12-11]
41. Overman VP. Healthfinder. Int J Dent Hyg 2008 Nov;6(4):378-378. [doi: 10.1111/j.1601-5037.2008.00341.x] [Medline: 19138191]
42. Mac O, Ayre J, Bell K, McCaffery K, Muscat DM. Comparison of readability scores for written health information across formulas using automated vs manual measures. JAMA Netw Open 2022 Dec 1;5(12):e2246051. [doi: 10.1001/jamanetworkopen.2022.46051] [Medline: 36508219]
43. Mastroianni F, Chen YC, Vellar L, et al. Implementation of an organisation-wide health literacy approach to improve the understandability and actionability of patient information and education materials: a pre-post effectiveness study. Patient Educ Couns 2019 Sep;102(9):1656-1661. [doi: 10.1016/j.pec.2019.03.022] [Medline: 30962076]
44. Baines R, Bradwell H, Edwards K, et al. Meaningful patient and public involvement in digital health innovation, implementation and evaluation: a systematic review. Health Expect 2022 Aug;25(4):1232-1245. [doi: 10.1111/hex.13506] [Medline: 35526274]

Abbreviations

DCP3 : Disease Control Priorities 3rd Edition
mHealth: mobile health
UDL: Universal Design for Learning
UHC: universal health coverage
UN: United Nations
WHO: World Health Organization

Edited by A Mavragani; submitted 28.02.24; peer-reviewed by A Hassan, C Baur, G Berget; revised version received 20.01.25; accepted 22.01.25; published 06.03.25.

Please cite as:
Muscat DM, Hinton R, Kuruvilla S, Nutbeam D
"Your Life, Your Health: Tips and Information for Health and Well-Being": Development of a World Health Organization Digital Resource to Support Universal Access to Trustworthy Health Information
JMIR Form Res 2025;9:e57881
URL: https://formative.jmir.org/2025/1/e57881
doi:10.2196/57881

© Danielle M Muscat, Rachael Hinton, Shyama Kuruvilla, Don Nutbeam. Originally published in JMIR Formative Research (https://formative.jmir.org), 6.3.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

HUMAn, a Real-Time Evolutive Patient Model for Major Incident Simulation: Development and Validation Study

Maxence Laurent[1], Dipl.Ing.; Arnaud Jaccard[2], MSc; Laurent Suppan[3], MD; Elio Erriquez[3]; Xavier Good[1], MSc; Eric Golay[3], MAS; Dominique Jaccard[1], MSc; Mélanie Suppan[4], MD, MSc

[1]Media Engineering Institute, School of Management and Engineering Vaud, University of Applied Sciences and Arts of Western Switzerland, Avenue des Sports 20, Yverdon, Switzerland

[2]Faculty of Biology and Medicine, University of Lausanne, Lausanne, Switzerland

[3]Division of Emergency Medicine, Department of Anesthesiology, Clinical Pharmacology, Intensive Care and Emergency Medicine, University of Geneva Hospitals and Faculty of Medicine, Geneva, Switzerland

[4]Division of Anesthesiology, Department of Anesthesiology, Clinical Pharmacology, Intensive Care and Emergency Medicine, University of Geneva Hospitals and Faculty of Medicine, Geneva, Switzerland

Corresponding Author:
Dominique Jaccard, MSc
Media Engineering Institute, School of Management and Engineering Vaud, University of Applied Sciences and Arts of Western Switzerland, Avenue des Sports 20, Yverdon, Switzerland

Abstract

Background: Major incidents correspond to any situation where the location, number, severity, or type of casualties requires extraordinary resources. Major incident management must be efficient to save as many lives as possible. As any paramedic or emergency medical technician may unexpectedly have to respond to major incidents, regular training is mandatory. Those trainings usually include simulations. The vast majority of major incident simulations are limited by the fact that simulated patients do not evolve during the simulation, regardless of the time elapsed and treatment decisions. Therefore, most simulations fail to incorporate the critical temporal effect of decision-making.

Objective: This study aimed to develop and validate a simplified mathematical model of physiology, capable of plausibly simulating the real-time evolution of several injuries.

Methods: A modified version of the user-centered design framework, including a relevance, development, and validation phase, was used to define the development process of the physiological model. A 12-member design and development team was established, including prehospital physicians, paramedics, and computer scientists. To determine whether the developed model was clinically realistic, 15 experienced professionals working in the prehospital field participated in the validation phase. They were asked to rate clinical and physiological parameters according to a 5-point Likert scale ranging from 1 (impossible) to 5 (absolutely realistic).

Results: The design and development team led to the development of the HUMAn model (Human is an Uncomplicated Model of Anatomy). During the relevance phase, the team defined the needed features of the model: clinically realistic, able to compute the evolution of prehospital vital signs, yet simple enough to allow real-time computation for several simulated patients on regular computers or tablets. During the development phase, iterations led to the development of a heart-lung-brain interaction model coupled to functional blocks representing the main anatomical body parts. During the validation phase, the evolution of nine simulated patients presenting pathologies devised to test the different systems and their interactions was assessed. Overall, clinical parameters of all patients had a median rating of 5 (absolutely realistic; IQR 4-5). Most (n=52, 96%) individual clinical parameters had a median rating of 5, the remainder (n=2, 4%) being rated 4. Overall physiological parameters of all patients had a median rating of 5 (absolutely realistic; IQR 3-5). The majority of individual physiological parameters (n=43, 79%) had a median rating of 5, with (n=9, 17%) rated 4, and only (n=2 ,4%) rated 3.

Conclusions: A simplified model of trauma patient evolution was successfully created and deemed clinically realistic by experienced clinicians. This model should now be included in computer-based simulations and its impact on the teaching of major incident management assessed through randomized trials.

(JMIR Form Res 2025;9:e66201) doi:10.2196/66201

KEYWORDS

physiological model; mathematical model; computer simulation; major incident management; emergency medicine; mass casualties; healthcare professional education; professional education; continuing education

XSL•FO
RenderX

Introduction

Background and Importance

In the health sector, major incidents correspond to "any incidents where the location, number, severity or type of lives of casualties requires extraordinary resources" [1]. Major incidents are rare events that require efficient and effective management to save as many lives as possible [1,2]. The scarcity of these events and the impact they can have both on the victims and on society, as the impact such stressful situations may have on rescuers clinical performance [3] has led to the development of many different training modalities [4]. Thus, most prehospital providers are trained to manage major incidents through tabletop or computer-based simulations, practical exercises without casualties, or full-scale simulations [5].

These simulations all present limitations, some of which can often be addressed by using other forms of training. However, 1 important limitation is common to almost all major incident simulations and resides in the fact that simulated patients do not evolve over time [6]. Indeed, in most simulations, victims do not evolve at all regardless of the time elapsed and despite treatment and transport decisions. For instance, a simulated victim experiencing a massive internal hemorrhage (such as splenic rupture) will present identical vital signs throughout the simulation [7]. Consequently, most current simulations fail to incorporate the temporal effects of decision-making, which is a major characteristic of major incident management [8].

A simplified mathematical model of human physiology could be a solution for simulating victims' evolution over time [9]. However, most currently available mathematical models of human physiology present limitations preventing their use in real-time major incident simulations. In 2005, Sacco et al [10] proposed a model using respiratory rate, pulse rate, and motor response (RPM) to predict victims' survival probabilities. In this model, a single discrete modification of RPM parameters is made depending on whether the victims remains at the incident scene or is transported to a hospital. This change of RPM parameters is based on expert consensus. While insightful for predicting survival probabilities, this model does not enable to calculate the continuous time-based evolution of multiple physiological and clinical parameters, based on treatment decisions. Many others models are complex and ask for high processing power, making them incompatible with their use to simulate the synchronous evolution of several victims[11]. Furthermore, most models are not open source and cannot be embedded into third-party applications [12].

Objectives

Our objective was to develop and validate a simplified physiological mathematical model capable of simulating the evolution of several casualties, in real time, while remaining realistic in the context of major events training.

Methods

General Design

A modified version of the user-centered design (UCD) framework proposed by Farao et al [13] was used to define the development process of the physiological model. The UCD framework is based on the 5 modes of design thinking (empathize, define, ideate, prototype, and test) [14], with the addition of the knowledge base mode of the information systems research framework [15].

These 6 modes were sorted in 2 phases, the relevance phase and an iterative development phase. To validate the developed model, a validation phase was added. The overall resulting framework is summarized in Figure 1.

Figure 1. Framework used for the development and validation of the physiological model. The framework integrates the 6 modes of the user-centred design framework with the addition of a validation phase.

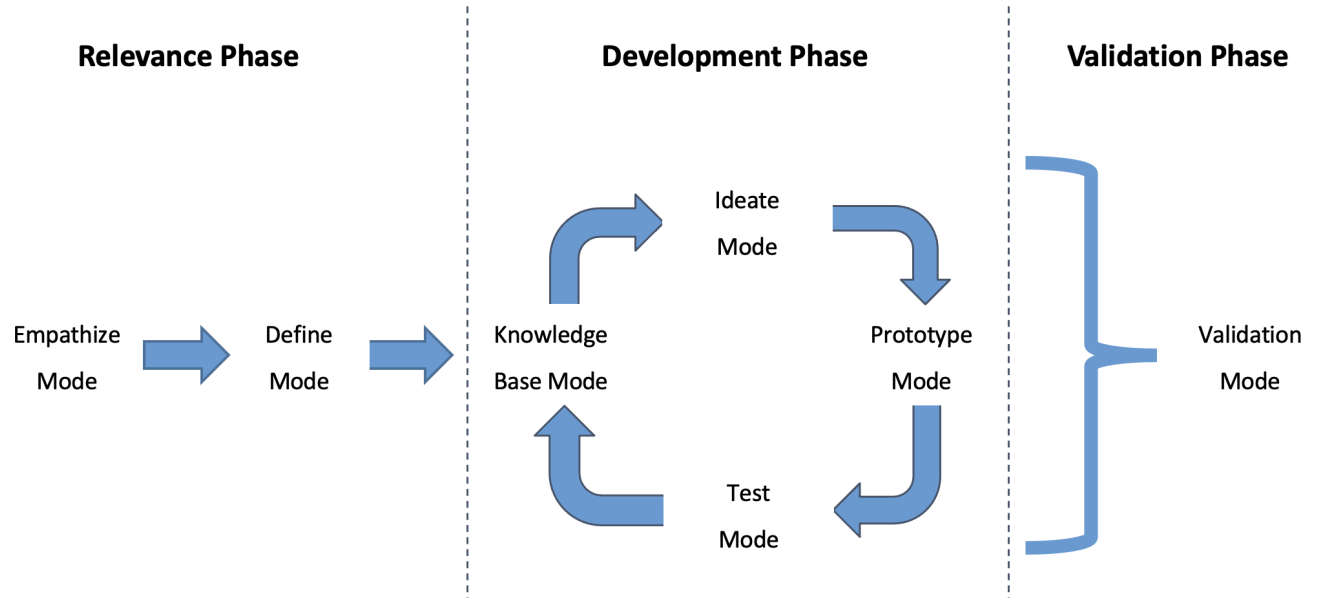

XSL•FO
RenderX

Relevance Phase

The objective of the relevance phase was to identify the problem to be solved and define the overall goal for the development phase.

Empathize Mode

The goal of the empathize mode was to identify and understand the end-users and the context of use of the physiological model. Relevant elements of the co.LAB framework, which has been shown to have a positive impact on team collaboration when designing serious games [16,17], were used to determine the learning contexts where the physiological model could be used. In line with the philosophy of the co.LAB framework, end-users were included in the design process early on. A 12-member team was established, including 3 prehospital senior physicians and 3 advanced paramedics (all with extensive emergency medicine experience), 1 medical student, 1 game designer, 3 computer scientists, and 1 instructional designer. A total of 8 team members are coauthors of this article.

Define Mode

Based on results of the empathize mode, the define mode aimed at identifying the needed characteristics of the physiological model as well as basic technical requirements. This was achieved through several collaborative sessions between medical doctors and computer scientists.

Development Phase

The objective of the development phase was to develop a conceptual physiological model, implement it into a software application, and test its ability to simulate the evolution of victims during a major incident.

Knowledge Base Mode

The knowledge base mode aimed at identifying existing solutions and reliable knowledge foundation for the development of the physiological model. First, Google Scholar and the regular Google search engine were used to identify already existing open-source computational models of physiology. Available models were assessed for their relevance against our objectives. Since no such model was found, a team of medical doctors and computer scientists collaborated to identify physiological and physiopathological formulas necessary to develop a model corresponding to the characteristics detailed through the define mode. Searches were done in research databases, textbooks and grey literature. After the first iteration, further searches were based on the results of the test mode and relevant formulas inserted accordingly.

Ideate Mode

The ideate mode was used to define the combination of formulas needed to develop a simple yet clinically plausible physiological model. It also included ideation for technological solutions for the development of the software simulation. Furthermore, this mode was used to investigate and define the algorithmic implementation of formulas in the software. This was carried out through brainswarming sessions involving medical doctors, computer scientists and paramedics [18].

Prototype Mode

Basic implementations of formulas were first done on spreadsheets and then implemented into a full software application. Computational models of independent physiological systems (respiratory, cardiovascular, and brain) were first developed. Once deemed realistic enough by the development team, interactions between these systems were implemented. After ensuring that interactions were appropriate under normal physiological conditions, elements required to simulate injuries were progressively added and tested.

Test Mode

Each prototype was tested by a panel of doctors and paramedics, all belonging to the development team. Tests were first conducted to assess the plausibility of the simulated independent physiological systems and, then, the simultaneous behavior of those systems. Situations involving single, then multiple injuries were then tested. Tests consisted of an evaluation of the plausibility of the evolution of clinical and physiological parameters.

Validation Phase

During the validation phase, relevant elements of Eysenbach's Checklist for Reporting Results of Internet E-Surveys (CHERRIES) [19] have been used to report the results.

Design and Study Sequence

To determine whether the developed model was clinically realistic, 15 experienced prehospital professionals (5 advanced paramedics and 10 senior registrars working in ground and helicopter emergency medical services, none of them being part of the development team) were invited to rate clinical and physiological parameters according to a 5-point Likert scale ranging from 1 (impossible) to 5 (absolutely realistic).

After a short briefing regarding the study's aims, these experts were asked to create a personal account on an web testing platform. The top of the page displayed the patient's characteristics (sex, age, height, and weight), their pathology, and their time of death if the model predicted it would happen within 4 hours if left untreated. All experts were clearly informed that they would have to conceptualize how patient parameters would evolve if they did not receive any medical treatment.

Clinical (oxygen saturation, respiratory rate, heart rate, mean arterial pressure, Glasgow coma scale, and ability to walk) and physiological (alveolar volume, PaO2, PaCO2, blood volume, stroke volume, and intracranial pressure) parameters were displayed at 5 different time points. The first time point (T0) corresponds to the initial state, just before the simulated patient was ascribed a specific pathology. The last point (T4) corresponds to the time of death of the simulated patient, with a maximum T4 equals to 4 hours if the survival time was more than 4 hours. To assess the plausibility of the evolution of the parameters, intermediate values between T0 and T4 were presented at times corresponding to 25% (T1), 50% (T2), and 75% (T3) of T4.

Although the use of relative times for T1 to T4 leads to an assessment of clinical and physiological parameters at different

XSL•FO
RenderX

absolute times for each patient, we chose this approach because it enable us to assess the plausibility of the evolution of the parameters at similar relative times to death. In order to give the complete information to the experts who assessed the model, the absolute values of T1, T2, T3, and T4 were displayed in the form of "hours : minutes : seconds" for each simulated patient.

Figure 2 shows a screenshot of the system used to display parameters and enable participants to rate them. In this particular case, the patient was expected to survive more than 4 hours. Thus the last time point is equal to 4 hours and intermediate time points correspond to 1, 2, and 3 hours.

Figure 2. Screenshot of the web assessment form. A summary of the simulated patients to be evaluated is displayed on the left side. The patient's characteristics and the Likert scale legend are shown on the top of the page. Each parameter is shown on a single row, and each column represents a specific timepoint. Once the expert has chosen a rating, the background turns to green.

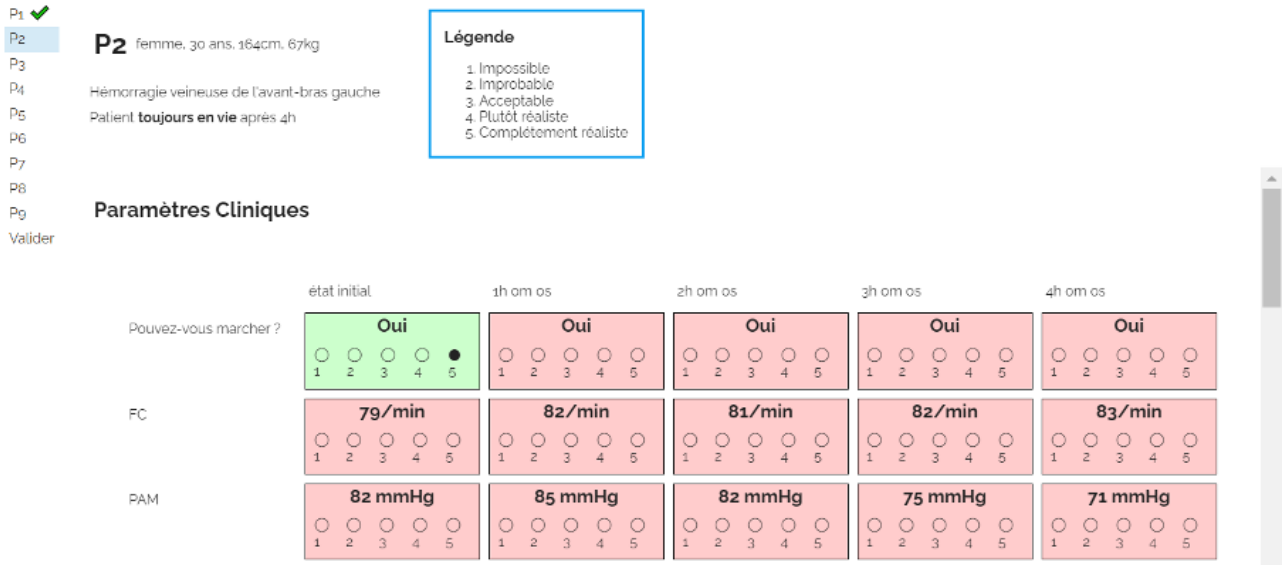

Statistical Analysis

Data were extracted to a .csv file and imported in Stata (version 17.0. StataCorp) for curation and statistical analysis. All clinical and physiological parameters were individually tested at each time point to determine their perceived reliability. Their overall perceived reliability was then assessed. Clinical parameters (oxygen saturation, respiratory rate, heart rate, mean arterial pressure, Glasgow coma scale, and ability to walk) were grouped and tested together at each timepoint and overall. The same procedure was applied to physiological parameters (alveolar volume, PaO_2, $PaCO_2$, blood volume, stroke volume, and intracranial pressure). Results are reported as median (IQR) to better reflect the definitions of the Likert scale.

All clinical and physiological parameters were individually tested at each time point to determine their perceived reliability. Their overall perceived reliability was then assessed. Clinical parameters (SpO_2, respiratory rate, heart rate, mean arterial pressure, Glasgow coma scale, ability to walk) were grouped and tested together at each timepoint and overall. The same procedure was applied to physiological parameters (alveolar volume, PaO_2, $PaCO_2$, blood volume, stroke volume, intracranial pressure).

Ethical Considerations

Since the participants who took part in the validation phase were all professionals working in the prehospital field, their participation did not fall within the scope of the Swiss Federal Act on Research involving Human Beings (HRA, RS 810.30). All the participants agreed to take part in this study without any form of compensation, and no incentive other than pure scientific interest was used to promote participation. The results of the evaluation phase were anonymized to ensure the confidentiality of participants' assessments.

All the participants agreed to take part in this study without any form of compensation, and no incentive other than pure scientific interest was used to promote participation. The results of the evaluation phase were anonymized to ensure the confidentiality of participants' assessments.

Results

Relevance Phase

Empathize Mode

Several collaborative design sessions were organized. All collaborative sessions were held in the presence of at least 3 members of the overall team, with the medical and game development professions systematically represented. At the end of each session, ideas were discussed and synthesized.

Define Mode

The team agreed on the characteristics of the physiological model that was to be developed. The most important feature was that the model was to be clinically realistic. This entailed that it should include all the elements required to compute vital signs and their evolution according to the presence of pathologies usually encountered in major incidents, such as hemorrhages, burns, pneumothoraxes, and head traumas. In addition, the model was to be simple enough to allow real-time computation of all relevant vital parameters for several simulated patients simultaneously. It was also decided that the system

XSL•FO
RenderX

should be able to randomly create simulated patients. Taking all these elements into account, it was determined that the model would be limited to the characteristics outlined in Table 1.

Table . Characteristics of simulated patients.

Characteristics	Values
Age (years), range	16-100
Sex (boolean)	Male or female
BMI (kg/m^2), mean (SD)	23 (3)
Height (cm), mean (SD)	
Man	178.4 (7.6)
Woman	164.7 (7.1)

Development Phase

Knowledge Base Analysis

When models were readily available, knowledge base analysis was a straightforward process and all relevant references were rapidly included. Several examples are given in Table 2.

Table . Examples of data and formulae obtained through knowledge base analysis.

Variable [study]	Data and formulae
Ideal weight (kg) [20]	
Man	50+0.9 (kg/cm)*(height [cm]-152.4 [cm])
Woman	45.5+0.9 (kg/cm)*(height [cm]-152.4 [cm])
Blood volume (mL) [21]	70*ideal weight (kg)
Dead space (mL) [22]	2.2 (mL/kg)*ideal weight (kg)
Respiratory quotient [23]	0.84
Cerebral blood flow (L/min) [24]	0.5 (L*min/kg)*1.4 (kg)
Cerebral perfusion pressure (mmHg) [25]	Mean arterial pressure – intracranial pressure

Ideate Mode

Since data regarding patient evolution according to specific injuries were often lacking, the mathematical formulae controlling the evolution of physiological parameters in the presence of such injuries had to be developed through a collaboration between clinicians and computer scientists. These formulae were progressively refined through multiple iterations.

Table 3 shows the steps used to compute systolic and diastolic blood pressures according to volemia, sympathetic nervous system activation, pericardic and pleural pressures, and cardiac contractility. While many parameters are taken into account, several were either simplified or used as constants. Nevertheless, the model enables the inclusion of future updates to simulate currently simplified elements such as vascular resistance.

XSL•FO
RenderX

Table . Steps used to determine diastolic and systolic blood pressure.

Step and parameters	Data and formulae	Comment
Initial data		
Blood volume (mL)	70*weight [21]	—[a]
Systemic vascular resistance (mmHg*min/L)	13 [26]	—
Initial EDV[b] $_{Initial}$ (mL)	120	—
Effect of blood volume on EDV ($\Delta EDV_{Volemia}$) (mL)		EDV is reduced in the event of blood loss (hypovolemia), increased in the event of hypervolemia. The model allows it to vary between +40ml and −70ml according to this parameter. (model-specific approximations rather than established formulas)
Effect of pericardial pressure on EDV ($\Delta EDV_{PPericardial}$) (mL)		A pericardial effusion causes fluid to accumulate in the pericardium. Consequently, EDV decreases along with the increase in pericardial fluid. (model-specific approximations rather than established formulas)
Effect of pleural pressure on EDV ($\Delta EDV_{PPleural}$) (mL)		This simulates a limitation in heart filling consecutive to an increase in pleural cavity pressure. (model-specific approximations rather than established formulas)
Effect of sympathetic nervous system activation on EDV ($\Delta EDV_{Sympathetic}$) (mL)		This is a very simplified approximation designed to simulate the increase in venous return consecutive to sympathetic activation.(model-specific approximations rather than established formulas)
Final EDV (EDV) (mL)	$EDV = EDV_{Initial} + \Delta EDV_{Volemia} + \Delta EDV_{PPericardial} + \Delta EDV_{PPleural} + \Delta EDV_{Sympathetic}$	—
ESV[c] (mL)		(model-specific approximations rather than established formulas)
SV[d] (ml)	SV = EDV – ESV [27]	—
CO[e] (L/min)	CO = SV * HR * 1000 [27]	—
Left ventricular pressure (mmHg)		(model-specific approximations rather than established formulas)
MAP[f] $_{th}$ (mmHg)	MAP_{th} = CO * SVR [27]	—
MAP[g] $_{eff}$ (mmHg)	• If MAP_{th}< VP: MAPeff = MAP_{th} • If $MAP_{th} \geq$ VP: MAPeff = VP	MAP cannot be higher than VP (model-specific approximations rather than established formulas)
Actual CO (CO_{eff}) (L/min)	• If MAP_{th}< VP: CO_{eff} = MAP_{eff} / SVR • Else: CO_{eff} = CO	Adjustment of CO if theoretical MAP cannot be achieved (model-specific approximations rather than established formulas)
Calculation of DBP[h] and SBP[i] (mmHg)	DBP = 6/7 * MAP_{eff} SBP = 3/2 * MAP_{eff}	(model-specific approximations rather than established formulas)

[a] —: not applicable.

[b] EDV: end-diastolic volume.

[c] ESV: end-systolic volume.

XSL•FO
RenderX

[d] SV: stroke volume.

[e] CO: cardiac output.

[f] MAP_{th}: theoretical mean arterial pressure.

[g] MAP_{eff}: actual MAP.

[h] DBP: diastolic blood pressure.

[i] SBP: systolic blood pressure.

Brainswarming processes led to retain two other main concepts for the development of the application: the heart-lung-brain interaction model and the concept of functional blocks. The heart-lung-brain interaction model was decided since these 3 organs are responsible for almost all the vital signs linked to the ABC (airways, breathing, and circulation) evaluation performed by prehospital providers. Although other organs, such as the liver and the kidney, are of course vital, their function does not significantly alter vital signs in the immediate aftermath of a major incident.

The concept of functional blocks (Figure 3) was elaborated to provide more information regarding the neurological status of simulated patients and to create hemorrhages at different levels, with different blood flows, thereby allowing specific treatment options.

XSL•FO
RenderX

Figure 3. Original drawing illustrating the concept of functional blocks. These blocks are used to divide the body into anatomical subsections, allowing traumas to be created at different locations thus enabling localized treatments.

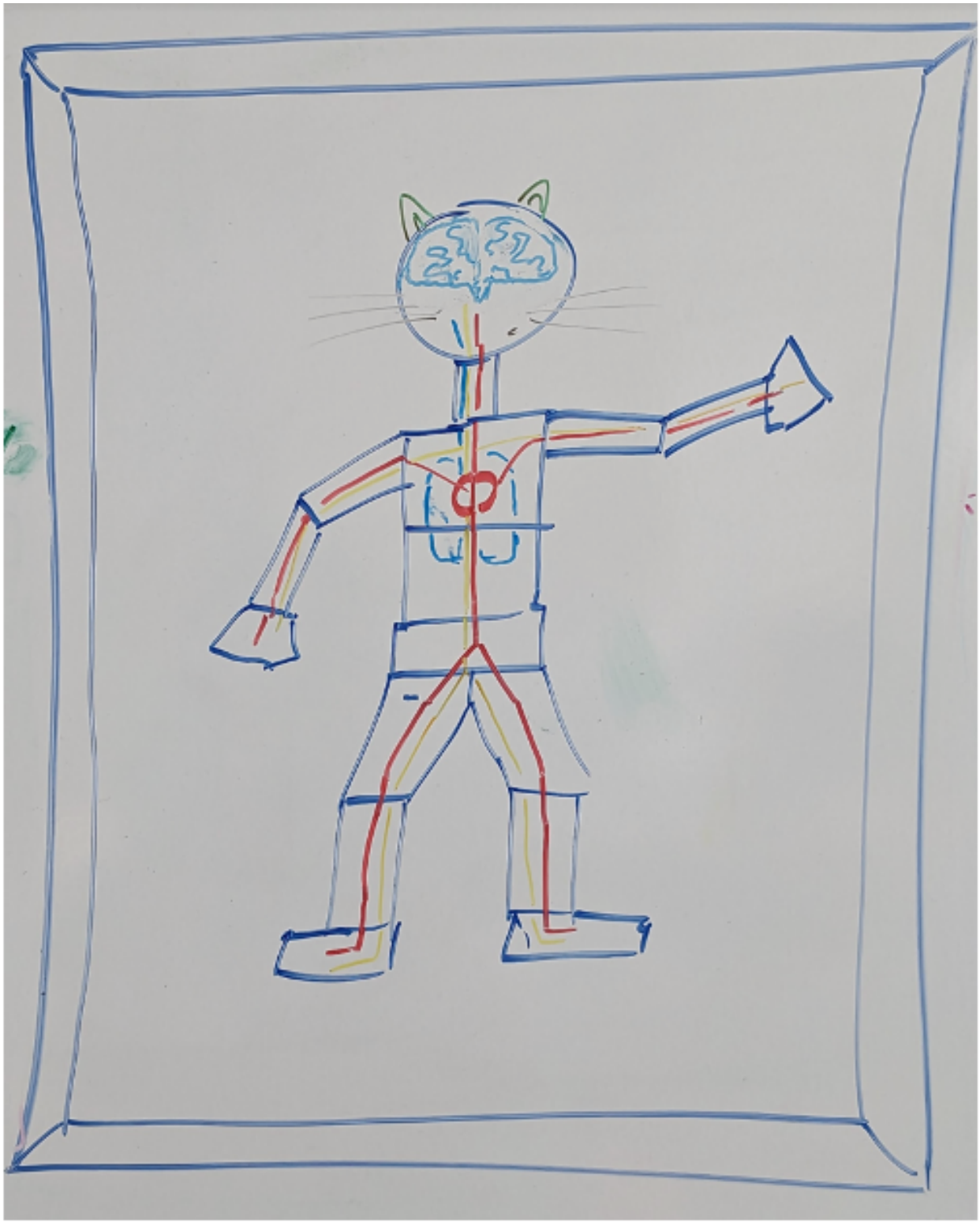

Prototype Mode

Prototypes were developed and tested on Wegas, an open-source web-based serious game authoring and execution platform [28]. Wegas source code can be downloaded on Github [29]. The model source code is freely available under a Creative Commons 4.0 CC-BY-NC licence. The source code can be downloaded on Github [30]. The model's source code is in English. At the implementation level, user interfaces are designed to be multilingual. The current version of the implementation is in French, but other languages may be added in the future.

The model source code is freely available under a Creative Commons 4.0 CC-BY-NC licence. The source code can be downloaded on Github [30]. The model's source code is in English. At the implementation level, user interfaces are designed to be multilingual. The current version of the

XSL•FO
RenderX

implementation is in French, but other languages may be added in the future.

Test Mode

All prototypes were tested by paramedics and physicians belonging to the development team. Their comments were used to progressively refine the model through new iterations of the knowledge base analysis and ideate modes.

Validation Phase

In total, 15 experienced prehospital professionals (5 advanced paramedics and 10 senior registrars working in ground and helicopter emergency medical services) were invited to participate in the validation study. All of them agreed to participate. Their characteristics are detailed in Table 4.

Table . Characteristics of the prehospital experts.

Characteristics	Overall (N=15)	Physicians (n=10)	Paramedics (n=5)
Gender, n (%)			
Man	11 (73)	6 (60)	5 (100)
Woman	4 (27)	4 (40)	0 (0%)
Other	0 (0)	0 (0)	0 (0%)
Age (years), mean (SD)	40 (6)	39 (5)	41 (7)
Postgraduate clinical experience (years), mean (SD)	11 (7)	10 (6)	13 (9)
Prehospital experience (years), mean (SD)	10 (8)	7 (7)	16 (5)

In order to test the different systems and their interactions, 9 simulated patients were automatically generated, with different personal characteristics. We then ascribed a specific pathology to each patient. No patient was ascribed more than 1 pathology. The characteristics of patients and ascribed pathologies are detailed in Table 5.

Table . Characteristics of the patients automatically generated by the simulator.

Patient	Sex (M/F)	Age (years)	Height (cm)	Weight (kg)	Ascribed pathology	Time of death (seconds)
Patient 1	M[a]	45	159	64	Massive internal hemorrhage (eg, ruptured spleen)	2700
Patient 2	F[b]	30	164	67	Veinous hemorrhage of the left forearm	>14,400
Patient 3	M	42	182	86	Traumatic cardiac tamponade	8850
Patient 4	F	63	161	53	Complete left pneumothorax	>14,400
Patient 5	F	46	168	68	Circular second-degree burns of the face and neck	5735
Patient 6	F	50	170	61	Complete circumferential third-degree burn of the chest	5670
Patient 7	F	32	161	62	Head trauma with cerebral hemorrhage	4745
Patient 8	F	61	164	69	Atlanto-occipital dislocation	345
Patient 9	M	44	168	57	Unstable, displaced cervical fracture between C5 and C7	>14,400

[a]M: male.

[b]F: female.

The validation team assessed each patient clinical and physiological parameters from T0 to T4. The initial and final states of all patients had a median rating of 5 (absolutely realistic) for both clinical (Table 6) and physiological (Table

XSL•FO
RenderX

7) parameters. The overall median rating was also 5 for each individual patient. Detailed assessments, by patient and time period, are displayed in Tables 6 and 7. Ratings as mean (SD) are available in Multimedia Appendix 1.

Table . Assessment of the simulated patients' overall clinical parameters.

Patient	T0, median (IQR)	T1, median (IQR)	T2, median (IQR)	T3, median (IQR)	T4, median (IQR)	Overall, median (IQR)
Patient 1	5 (5-5)	5 (4-5)	4 (4-5)	5 (4-5)	5 (5-5)	5 (4-5)
Patient 2	5 (5-5)	5 (5-5)	5 (5-5)	5 (4-5)	5 (4-5)	5 (5-5)
Patient 3	5 (5-5)	5 (4-5)	4 (3-5)	5 (4-5)	5 (5-5)	5 (4-5)
Patient 4	5 (5-5)	5 (4-5)	5 (4-5)	5 (4-5)	5 (4-5)	5 (4-5)
Patient 5	5 (5-5)	5 (4-5)	5 (3-5)	4 (3-5)	5 (5-5)	5 (4-5)
Patient 6	5 (5-5)	5 (4-5)	5 (3-5)	4 (3-5)	5 (5-5)	5 (4-5)
Patient 7	5 (5-5)	5 (4-5)	4 (3-5)	5 (4-5)	5 (5-5)	5 (4-5)
Patient 8	5 (5-5)	5 (3-5)	5 (3-5)	5 (3-5)	5 (5-5)	5 (4-5)
Patient 9	5 (5-5)	5 (4-5)	5 (4-5)	5 (4-5)	5 (4-5)	5 (4-5)

Table . Assessment of the simulated patients' overall physiological parameters.

Patient	T0, median (IQR)	T1, median (IQR)	T2, median (IQR)	T3, median (IQR)	T4, median (IQR)	Overall, median (IQR)
Patient 1	5 (4-5)	5 (4-5)	4 (3-5)	4 (3-5)	5 (4-5)	5 (3-5)
Patient 2	5 (5-5)	5 (5-5)	5 (5-5)	5 (4-5)	5 (4-5)	5 (5-5)
Patient 3	5 (5-5)	5 (4-5)	4 (3-5)	4 (3-5)	5 (5-5)	5 (4-5)
Patient 4	5 (5-5)	5 (4-5)	5 (4-5)	5 (4-5)	5 (4-5)	5 (4-5)
Patient 5	5 (5-5)	5 (4-5)	5 (4-5)	5 (4-5)	5 (5-5)	5 (4-5)
Patient 6	5 (5-5)	5 (4-5)	5 (4-5)	5 (3-5)	5 (5-5)	5 (4-5)
Patient 7	5 (5-5)	5 (5-5)	5 (4-5)	5 (4-5)	5 (5-5)	5 (5-5)
Patient 8	5 (5-5)	5 (4-5)	5 (4-5)	5 (4-5)	5 (5-5)	5 (4-5)
Patient 9	5 (5-5)	5 (3-5)	5 (3-5)	5 (3-5)	5 (3-5)	5 (3-5)

Overall clinical parameters include oxygen saturation, respiratory rate, heart rate, mean arterial pressure, Glasgow coma scale, and ability to walk. T4 corresponds to the time of death of the simulated patient (T4 equals 4 hours if the survival time was more than 4 hours). T1 to T4 correspond respectively to 25%, 50%, and 75% of T4. Results are presented as "median (IQR)."

Overall physiological parameters include alveolar volume, PaO2, PaCO2, blood volume, stroke volume, and intracranial pressure. T4 corresponds to the time of death of the simulated patient (T4 equals 4 hours if the survival time was more than 4 hours). T1 to T4 correspond respectively to 25%, 50%, and 75% of T4. Results are presented as "median (IQR)."

The overall median rating for each specific clinical (Table 8) and physiological (Table 9) parameter was 5 (absolutely realistic). Further results are available in Multimedia Appendix 1.

XSL•FO
RenderX

Table . Assessment of the simulated patients' clinical parameters.

Patient	SpO_2,[a] median (IQR)	RR,[b] median (IQR)	HR,[c] median (IQR)	MAP,[d] median (IQR)	GCS,[e] median (IQR)	Walks,[f] median (IQR)
Patient 1	5 (4-5)	5 (4-5)	5 (3-5)	5 (4-5)	5 (4-5)	5 (4-5)
Patient 2	5 (5-5)	5 (4-5)	5 (4-5)	5 (5-5)	5 (5-5)	5 (5-5)
Patient 3	5 (5-5)	5 (5-5)	5 (4-5)	5 (4-5)	5 (4-5)	5 (4-5)
Patient 4	4 (4-5)	5 (4-5)	5 (4-5)	5 (4-5)	5 (5-5)	5 (4-5)
Patient 5	5 (4-5)	5 (4-5)	5 (3-5)	5 (4-5)	5 (5-5)	5 (4-5)
Patient 6	5 (4-5)	5 (4-5)	5 (4-5)	5 (4-5)	5 (5-5)	5 (3-5)
Patient 7	5 (5-5)	5 (4-5)	5 (4-5)	5 (4-5)	5 (4-5)	5 (4-5)
Patient 8	5 (4-5)	5 (5-5)	5 (4-5)	5 (4-5)	4 (2-5)	5 (5-5)
Patient 9	5 (4-5)	5 (4-5)	5 (4-5)	5 (4-5)	5 (5-5)	5 (5-5)
Overall	5 (4-5)	5 (4-5)	5 (4-5)	5 (4-5)	5 (4-5)	5 (4-5)

[a]SpO_2: oxygen saturation.

[b]RR: respiratory rate.

[c]HR: heart rate.

[d]MAP: mean arterial pressure.

[e]GCS: Glasgow Coma Scale.

[f]Ability to walk.

Table . Assessment of the simulated patients' physiological parameters. Results are presented as "median (IQR)".

	PaO_2,[a] median (IQR)	$PaCO_2$,[b] median (IQR)	Alveolar volume, median (IQR)	Blood volume, median (IQR)	Stroke volume, median (IQR)	ICP,[c] median (IQR)
Patient 1	4 (3-5)	4 (2-5)	4 (3-5)	4 (3-5)	5 (4-5)	5 (5-5)
Patient 2	5 (5-5)	5 (4-5)	5 (5-5)	5 (4-5)	5 (4-5)	5 (5-5)
Patient 3	5 (4-5)	4 (3-5)	5 (4-5)	5 (4-5)	5 (4-5)	5 (5-5)
Patient 4	4 (4-5)	4 (4-5)	4 (4-5)	5 (5-5)	5 (4-5)	5 (5-5)
Patient 5	5 (4-5)	5 (4-5)	5 (4-5)	5 (5-5)	5 (5-5)	5 (5-5)
Patient 6	5 (4-5)	5 (4-5)	5 (4-5)	5 (4-5)	5 (5-5)	5 (5-5)
Patient 7	5 (5-5)	5 (4-5)	5 (5-5)	5 (5-5)	5 (5-5)	5 (4-5)
Patient 8	5 (4-5)	5 (4-5)	5 (5-5)	5 (5-5)	5 (3-5)	5 (5-5)
Patient 9	4 (3-5)	3 (2-5)	3 (3-5)	5 (5-5)	5 (4-5)	5 (5-5)
Overall	5 (4-5)	5 (3-5)	5 (4-5)	5 (4-5)	5 (4-5)	5 (5-5)

[a]PaO_2: arterial partial pressure of O_2.

[b]$PaCO_2$: arterial partial pressure of CO_2.

[c]ICP: intracranial pressure.

Discussion

Principal Findings

A simplified model of trauma patient evolution was successfully created using the modified version of the user-centered design framework and was deemed clinically realistic by experienced prehospital physicians and paramedics. The diversity of simulated clinical cases, which included nervous, cardiovascular and respiratory pathologies, and its suitability for real-time calculations makes this model a good candidate for inclusion in major incident simulations. Many kinds of simulations could be considered, computer-based table-tops and serious games, practical exercises without casualties, and even full-scale simulations. In this latter case, tablets could be used to display the vital parameters of simulated patients.

All clinical and physiological parameters had an overall median rating of 5 (absolutely realistic). Nevertheless, some parameters had lower median ratings when specific patients were analyzed. For instance, the oxygen saturation values of patient 4, who presented a complete left pneumothorax, had a median rating of 4 (rather realistic). Several hypotheses could explain these findings. First, even though an iterative process was followed to progressively refine the evolution of specific parameters, the current model is still an approximation of actual human

XSL•FO
RenderX

physiology. Thus, further iterations may enable further improvements. Second, prehospital clinicians quickly manage airway and breathing issues when responding to critical situations. Therefore, these professionals have never witnessed the evolution of patients left untreated and could find it harder to conceive how vital and physiological parameters would develop in such cases. Finally, given the inherent variability in injury presentations and the subjective nature of textual descriptions, there is potential for variation in interpretation among prehospital providers. This inter-rater variability may contribute to the differing assessments of injury realism.

Other models have previously been developed [31], none of which were however deemed suitable for our purpose, that is, real-time simulation of multiple trauma patients with realistic heart-lung-brain interactions. While models such as HumMod and Biogears provide realistic, scientifically grounded models of patient evolution [32,33], their complexity usually requires considerable computational power, making them hardly suitable for major incident simulations. In addition, and although many physiological processes are supported by clear scientific literature and rules, the lack of data regarding most prehospital injuries had to be compensated for by clinical experience. The inclusion of senior prehospital physicians and paramedics since inception has thus enabled the creation of a model realistic enough to simulate the evolution of trauma patients during a major incident.

The model developed through this study presents the advantage of being based on an anatomically realistic model made of several blocks simulating the different body compartments which could be involved in case of injury. Neural and vascular lesions are thus easy to simulate, and these blocks could be divided up further if needed. Furthermore, other organ systems could be added according to the end-user's requirements: for instance, adding clearance mechanisms could prove useful to the management of major incidents involving hazardous materials. Conversely, disabling or removing systems uninvolved in the simulated situation could save processing power.

Simulating the outcome of patients involved in major incidents could prove useful in several ways. Including an evolutive model in computer-based simulations could help ambulance and medical commanders better understand the impact of certain decisions, such as the use of human resources for treatment rather than for triage. The importance of accurate triage cannot be overstated and methods based on patient evolution should be promoted [9]. In this regard, this physiological model could be used to compare different triage algorithms among different populations. Indeed, some professional categories may be more efficient than others depending on the tool provided [34].

Limitations

This model is not without limitations. The most important and most obvious limitation is the lack of pediatric-specific formulae, which prevents the generation of patients aged 16 years or younger. Another significant limitation is the absence of treatment-specific effects, which would be necessary to assess the effect of human and material resources through computer-based simulations. Since physiological and clinical parameters of simulated patients were assessed at relative intermediate time points (25%, 50%, and 75%) of death time, the absolute times (hours: minutes: seconds) of those intermediate points varied across simulated patients. This may complicate assessing patients' conditions over time. However, to enhance clarity, the absolute value of each intermediate time point was displayed on the interface. In addition, experts were informed about the calculation method for intermediate time points before starting the evaluation of parameters. While the overall scores, both for clinical and physiological parameters, show the high realism of the validation study, overall scores could potentially hide specific inconsistencies. However, Q1 and Q3 values are mostly close to median values. The plausibility of simulated patients' parameters has been assessed by only 15 experts. Given the overall score used in the assessment, the relatively small number of experts who evaluated the model may not have highlighted individual inconsistencies. Finally, the current version of this simplified physiological model was only developed to simulate traumas. Thus, other causes of major incidents, such as those linked to hazardous materials, cannot be simulated without further developments.

Another significant limitation is the absence of treatment-specific effects, which would be necessary to assess the effect of human and material resources through computer-based simulations.

Since physiological and clinical parameters of simulated patients were assessed at relative intermediate time points (25%, 50%, 75%) of death time, the absolute times (hours: minutes: seconds) of those intermediate points varied across simulated patients. This may complicate assessing patients' conditions over time. However, to enhance clarity, the absolute value of each intermediate time point was displayed on the interface. Additionally, experts were informed about the calculation method for intermediate time points before starting the evaluation of parameters.

While the overall scores, both for clinical and physiological parameters, show the high realism of the validation study, overall scores could potentially hide specific inconsistencies. However, Q1 and Q3 values are mostly close to median values.

The plausibility of simulated patients' parameters has been assessed by only 15 experts. Given the overall score used in the assessment, the relatively small number of experts who evaluated the model may not have highlighted individual inconsistencies.

Finally, the current version of this simplified physiological model was only developed to simulate traumas. Thus, other causes of major incidents, such as those linked to hazardous materials, cannot be simulated without further developments.

Perspectives

Embedding this model in computer-based simulations and serious games designed to teach major incident management should now be considered. The inclusion of evolutive pathologies could help determine the most efficient ways of dealing with specific major incident situations and prepare health professionals to deal with the stress of managing such events

XSL•FO
RenderX

[3]. Further developments should also be contemplated. These should help address several limitations, including the unavailability of pediatric patients and the inclusion of treatment specific effects.

Conclusion

A simplified model of trauma patient evolution was successfully created and deemed clinically realistic by experienced clinicians. This model should now be included in computer-based simulations and its impact on the teaching of major incident management assessed through randomized trials.

Acknowledgments

Authors thank Innosuisse who supported this study with the grant 120.336 IP-SBM. This funding sources had no role in the design and execution of this study, nor on data analysis and interpretation. The authors warmly thank all the other members of the development team, Michèle Chan, Simon Regard, Audrey Huguenin, Ely Braun, and Pierre-Benjamin Monaco. We also thank the 15 emergency professionals who took part in the validation phase of the physiological model.

Data Availability

Overall results of the validation phase are available in Multimedia Appendix 1.

Authors' Contributions

ML, MS, and AJ contributed to conceptualization of the physiological model. ML, MS, AJ, LS, and DJ handled methodology. ML and XG managed software development. MS, EG, and LS handled validation. LS conducted formal analysis and data curation. DJ handled the resources. DJ and LS handled the project administration and funding acquisition. ML, MS, DJ, and LS wrote the original draft. All authors contributed to investigation, review and editing of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Overall results of the validation phase, with ratings as mean (SD).
[DOCX File, 61 KB - formative_v9i1e66201_app1.docx]

References

1. Wieteska S, Mackway-Jones K. Major Incident Medical Management and Support: The Practical Approach to Pre-Hospital Incident Command, 3rd edition: Wiley-Blackwell; 2012.
2. Hugelius K, Edelbring S, Blomberg K. Prehospital major incident management: how do training and real-life situations relate? A qualitative study. BMJ Open 2021 Sep 23;11(9):e048792. [doi: 10.1136/bmjopen-2021-048792] [Medline: 34556511]
3. Duncliffe TH, D'Angelo B, Brock M, et al. Examining the effects of stress on the driving abilities of paramedic students. International Paramedic Practice 2019 Dec 2;9(4):76-83. [doi: 10.12968/ippr.2019.9.4.76]
4. Hampson D. Major incident preparedness: a curriculum and workplace necessity. Scand J Trauma Resusc Emerg Med 2020 Dec 9;28(1):115. [doi: 10.1186/s13049-020-00812-y] [Medline: 33298122]
5. Porthouse A, Clancy H, Lax P. Training for major incidents. Surgery (Oxf) 2021 Jul;39(7):388-392. [doi: 10.1016/j.mpsur.2021.05.007] [Medline: 34092879]
6. Wolf P, Bigalke M, Graf BM, Birkholz T, Dittmar MS. Evaluation of a novel algorithm for primary mass casualty triage by paramedics in a physician manned EMS system: a dummy based trial. Scand J Trauma Resusc Emerg Med 2014 Aug 28;22:50. [doi: 10.1186/s13049-014-0050-6] [Medline: 25214310]
7. Lee CWC, McLeod SL, Van Aarsen K, Klingel M, Franc JM, Peddle MB. First responder accuracy using SALT during mass-casualty incident simulation. Prehosp Disaster Med 2016 Apr;31(2):150-154. [doi: 10.1017/S1049023X16000091] [Medline: 26857296]
8. Hugelius K, Rådestad M, Al-Dhahir H, Kurland L. Decision-making by medical officer in charge during major incidents: a qualitative study. Scand J Trauma Resusc Emerg Med 2021 Aug 21;29(1):120. [doi: 10.1186/s13049-021-00937-8] [Medline: 34419113]
9. Sigle M, Berliner L, Richter E, et al. Development of an anticipatory triage-ranking algorithm using dynamic simulation of the expected time course of patients with trauma: modeling and simulation study. J Med Internet Res 2023 Jun 15;25:e44042. [doi: 10.2196/44042] [Medline: 37318826]

XSL•FO
RenderX

10. Sacco WJ, Navin DM, Fiedler KE, Waddell RK, Long WB, Buckman RF. Precise formulation and evidence - based application of resource - constrained triage. Acad Emerg Med 2005 Aug;12(8):759-770 [FREE Full text] [doi: 10.1197/j.aem.2005.04.003] [Medline: 16079430]
11. Pruett WA, Hester RL. The creation of surrogate models for fast estimation of complex model outcomes. PLoS ONE 2016;11(6):e0156574. [doi: 10.1371/journal.pone.0156574] [Medline: 27258010]
12. Gilbart MK, Hutchison CR, Cusimano MD, Regehr G. A computer-based trauma simulator for teaching trauma management skills. Am J Surg 2000 Mar;179(3):223-228. [doi: 10.1016/s0002-9610(00)00302-0] [Medline: 10827325]
13. Farao J, Malila B, Conrad N, Mutsvangwa T, Rangaka MX, Douglas TS. A user-centred design framework for mHealth. PLoS ONE 2020;15(8):e0237910. [doi: 10.1371/journal.pone.0237910] [Medline: 32813711]
14. Plattner H, Meinel C, Leifer H, Leifer L, editors. Design Thinking Research: Making Distinctions: Collaboration versus Cooperation: Springer; 2018.
15. Hevner AR. A three cycle view of design science research. Association for Information Systems 2007;19(2).
16. Jaccard D, Suppan L, Sanchez E, Huguenin A, Laurent M. The co.LAB generic framework for collaborative design of serious games: development study. JMIR Serious Games 2021 Jul 2;9(3):e28674. [doi: 10.2196/28674] [Medline: 34255655]
17. Jaccard D, Suppan L, Bielser F. Contribution of the co.LAB framework to the collaborative design of serious games: mixed methods validation study. JMIR Serious Games 2021 Nov 24;9(4):e33144. [doi: 10.2196/33144] [Medline: 34822336]
18. BrainSwarming: because brainstorming doesn't work. HBR. URL: https://hbr.org/2014/06/brainswarming-because-brainstorming-doesnt-work [accessed 2024-06-28]
19. Eysenbach G. Improving the quality of web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34. [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
20. McCarron MM, Devine BJ. Clinical pharmacy: case studies: case number 25 gentamicin therapy. Drug Intell Clin Pharm 1974 Nov;8(11):650-655. [doi: 10.1177/106002807400801104]
21. Lasocki S, Loupec T, Parot-Schinkel E, et al. Study protocol for a multicentre, 2×2 factorial, randomised, controlled trial evaluating the interest of intravenous iron and tranexamic acid to reduce blood transfusion in hip fracture patients (the HiFIT study). BMJ Open 2021 Jan 17;11(1):e040273. [doi: 10.1136/bmjopen-2020-040273] [Medline: 33455926]
22. West JB, Luks A. West's Respiratory Physiology: The Essentials: Wolters Kluwer; 2016.
23. Singer P. Simple equations for complex physiology: can we use VCO2 for calculating energy expenditure? Crit Care 2016 Mar 21;20(1):72. [doi: 10.1186/s13054-016-1251-3] [Medline: 26997171]
24. Lassen NA. Normal average value of cerebral blood flow in younger adults is 50 ml/100 g/min. J Cereb Blood Flow Metab 1985 Sep;5(3):347-349. [doi: 10.1038/jcbfm.1985.48] [Medline: 4030914]
25. Mount CA, Das JM. Cerebral perfusion pressure. In: StatPearls: StatPearls Publishing; 2024. [Medline: 30725956]
26. Klingensmith ME. The Washington Manual of Surgery: Wolters Kluwer Health; 2016.
27. Klabunde RE. Cardiovascular Physiology Concepts: Wolters Kluwer; 2022.
28. Laurent M, Monnier S, Huguenin A, Monaco PB, Jaccard D. Design principles for serious games authoring tool. IJSG 2022 Nov 14;9(4):63-87. [doi: 10.17083/ijsg.v9i4.458]
29. ALBASIM Research Team, Jaccard D. Wegas. Github. URL: https://github.com/Heigvd/Wegas [accessed 2025-02-04]
30. Laurent M, Jaccard D, Good X, Santaniello F. HUMAn project. Github. 2023. URL: https://github.com/Heigvd/mimms/tree/main/gameModel/libs/ClientScript/HUMAn [accessed 2024-08-19]
31. Borri A, Panunzi S, Brancaleoni R, et al. Simulation of trauma incidents: modelling the evolution of patients and resources. J Med Syst 2016 Nov;40(11):234. [doi: 10.1007/s10916-016-0599-x] [Medline: 27653041]
32. Hester RL, Brown AJ, Husband L, et al. HumMod: a modeling environment for the simulation of integrative human physiology. Front Physiol 2011;2:12. [doi: 10.3389/fphys.2011.00012] [Medline: 21647209]
33. McDaniel M, Baird A. A full-body model of burn pathophysiology and treatment using the biogears engine. 2019 Presented at: 2019 41st Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); Jul 23-27, 2019; Berlin, Germany p. 261-264. [doi: 10.1109/EMBC.2019.8857686]
34. Marcussen CE, Bräuner KB, Alstrøm H, Møller AM. Accuracy of prehospital triage systems for mass casualty incidents in trauma register studies - a systematic review and meta-analysis of diagnostic test accuracy studies. Injury 2022 Aug;53(8):2725-2733. [doi: 10.1016/j.injury.2022.05.006] [Medline: 35660101]

Abbreviations

CHERRIES: Checklist for Reporting Results of Internet E-Surveys
RPM: respiratory rate, pulse rate, and motor response
UCD: user-centered design

XSL•FO
RenderX

Edited by A Mavragani, D Chartash; submitted 09.09.24; peer-reviewed by NH Mohamad Zainal, Y Bitan; revised version received 13.01.25; accepted 16.01.25; published 07.03.25.

Please cite as:
Laurent M, Jaccard A, Suppan L, Erriquez E, Good X, Golay E, Jaccard D, Suppan M
HUMAn, a Real-Time Evolutive Patient Model for Major Incident Simulation: Development and Validation Study
JMIR Form Res 2025;9:e66201
URL: https://formative.jmir.org/2025/1/e66201
doi:10.2196/66201

© Maxence Laurent, Arnaud Jaccard, Laurent Suppan, Elio Erriquez, Xavier Good, Eric Golay, Dominique Jaccard, Mélanie Suppan. Originally published in JMIR Formative Research (https://formative.jmir.org), 7.3.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Exploratory Co-Design on Electronic Health Record Nursing Summaries: Case Study

Suhyun Park[1], PhD, RN; Jenna L Marquard[2], PhD; Robin R Austin[2], PhD, DNP, DC, RN, NI-BC; Christie L Martin[2], PhD, MPH, RN-BC; David S Pieczkiewicz[3], PhD; Connie W Delaney[2], PhD, RN

[1]Cizik School of Nursing, The University of Texas Health Science Center at Houston, #557, 6901 Bertner Ave, Houston, TX, United States

[2]School of Nursing, University of Minnesota, Minneapolis, MN, United States

[3]Institute for Health Informatics, University of Minnesota, Minneapolis, MN, United States

Corresponding Author:
Suhyun Park, PhD, RN
Cizik School of Nursing, The University of Texas Health Science Center at Houston, #557, 6901 Bertner Ave, Houston, TX, United States

Abstract

Background: Although electronic health record nursing summaries aim to provide a concise overview of patient data, they often fall short of meeting nurses' information needs, leading to underutilization. This gap arises from a lack of involvement of nurses in the design of health information technologies.

Objective: The purpose of this exploratory co-design case study was to solicit insights from nurses regarding nursing summary design considerations, including key information types and the preferred design prototype.

Methods: We recruited clinical nurses (N=33) from 7 inpatient units at a university hospital in the Midwestern United States using a purposive sampling method. We used images from a simulated nursing summary to generate visual card versions of the 46 information types currently included in an electronic health record vendor–generated nursing summary. Participants selected which cards to include and arranged them in their designs based on their perceived relevance of the information types to the summary and their preferred reading layout. The nurses' perceived relevance of information types to the summary was analyzed by quantifying the frequency of included cards, while the nurses' preferred reading layout was analyzed by quantifying the occurrence of closely paired cards to identify common groupings. After participants evaluated the information type cards, debriefing interviews were conducted and analyzed thematically to explore their rationales for the desired content and its arrangement.

Results: The participants demonstrated a high level of engagement in the activities. On average, all 33 participants included 61% (n=28) of the total information types (n=46). The most frequently included cards were "unit specimen" (results of the analysis of body fluid, tissue, or urine), "activity," "diet," and "hospital problems," each included by 33 participants. Participants most frequently preferred adjacency of the following pairs: "activity" and "diet" (paired by 26 participants; 79%) and "notes to physicians" and "notes to treatment team" (paired by 25 participants; 76%). Participants preferred arranging the cards to improve information accessibility, focusing on key information types.

Conclusions: Involving nurses in the co-design process may result in more useful and usable designs, thereby reducing the time required to navigate nursing summaries. Future work should include refining and evaluating prototypes based on the designs created by the nurses.

(JMIR Form Res 2025;9:e68906) doi:10.2196/68906

KEYWORDS

electronic health records; interview; nurses; user-computer interface; co-design

Introduction

Background

Ineffective communication among clinicians is a leading cause of sentinel events, such as falls and delays in treatments [1]. Given time constraints, nurses review vast amounts of patient information in electronic health records (EHRs) at the beginning of each shift; thus, effective tools for summarizing patient information are crucial [2]. Nursing summaries in EHRs are intended to provide a concise overview of essential patient data (eg, vital signs, intake and output, weight) on a single page, but rigid and incomplete formats have led to underutilization [3], resulting in nurses developing highly variable methods, such as paper notes, for summarizing patient information.

End-users are involved in the EHR development phase to prevent low usability and support optimal EHR utilization [4]. Co-design (or participatory design) has gained attention as a strategy to integrate stakeholder viewpoints, bridging the gap

XSL•FO
RenderX

between tools and users' needs to ensure the long-term adoption of tools [5]. Co-design aims to actively involve potential user groups in designing tools and interventions, leveraging their lived experiences to tailor outputs to their expressed needs and preferences, while aligning with existing workflows and available resources [6]. In health care, co-design has been applied across various settings and user groups, including in the creation of predictive analytics dashboards and decision support systems to enhance care quality and client outcomes in aged care [6]. Co-design has also been used to examine clinical nurses' general ideas about information display design in a nursing summary [2] and to develop a mobile health application to assist patients with obesity in self-managing their preparation for elective surgery [7].

The use of information cards as a tool for stakeholder engagement is an approach adopted by designers to structure the co-design ideation process [8,9]. Card sorting, a prevalent user-centered design technique, actively involves participants categorizing labeled cards based on criteria that most resonate with them [10]. Card sorting offers numerous benefits, including facilitating the organization of complex information into intuitive groupings, promoting consensus building among diverse user groups, enhancing the user experience by ensuring content and functionality are logically structured, and providing valuable insights into users' preferences and priorities [11]. Although we did not fully follow the card sorting technique, we adapted key principles from the card sorting method, such as labeling and analyzing the arrangement of cards, to suit this study context. This exploratory research focused on nurses' perceived usability and their feedback to inform nursing summary design strategies. The design strategies informed by nurse users' input could streamline navigational pathways, thereby reducing cognitive load. This aligns with the goals of the National Burden Reduction Collaborative, which unites health care organizations, federal agencies, and standards organizations to reduce clinical burden nationwide [12].

Problem Statement

Nurses' participation in EHR design is often limited, despite their role as key users [2,13]. A deeper understanding of their unique requirements and suggestions for EHR design is needed. Recognizing the pivotal role of end-user involvement in design [5], we used a co-design activity to examine nurses' perspectives on the use and arrangement of information types (eg, vital signs, intake and output, weight) within an existing EHR vendor's nursing summary. Specifically, the co-design activity aimed to gain insights into nurses' perspectives on nursing summary design, including which information types to include and how to arrange the information types.

Methods

Study Design

This was an exploratory, single session of co-design activities that gathered nurses' feedback and identified their design preferences. Although conducted independently, this study was part of an eye-tracking simulation, which focused on examining nurses' visual attention to EHR nursing summary information types (submitted to another journal).

Participants and Recruitment

We recruited participants from 5 adult medical-surgical and 2 intermediate care units at a university hospital using a purposive sampling method. We aimed to recruit 30 or more participants, as this is recommended for eye-tracking simulations. These units used the same summary layout as the example summary employed in this study. To be included, participants must have had at least 12 months of experience as a registered nurse. We excluded nurses who were not in frontline patient care, such as nurse educators, nurse managers, and nurse research coordinators. Recruitment was facilitated by nurse managers, who distributed email flyers providing details about the study.

Data Collection

Participants were individually invited to a shared research space from August to September 2023. We used a nursing summary from a simulated patient case on a training platform that mirrored the EHR system used by the recruiting hospital. Since the hospital had used the Epic EHR system (Epic Systems) since 2011, we assumed all participants were familiar with its layout. The EHR vendor–generated nursing summary included 46 information types (eg, vital signs, intake and output, weight). We printed and cut out cards for each information type so the participants could move them around as they liked (Figure 1). See Multimedia Appendix 1 for the full original nursing summary layout and Multimedia Appendix 2 for details of the information types. A lead researcher (SP) conducted all co-design activities independently, while a co-author (JLM) who is a human factors expert provided guidance on the study design and monitored the co-design activities after each session.

XSL•FO
RenderX

Figure 1. Example of an information type card. Each information type card used a consistent design, including a colored boundary, heading, and optional content.

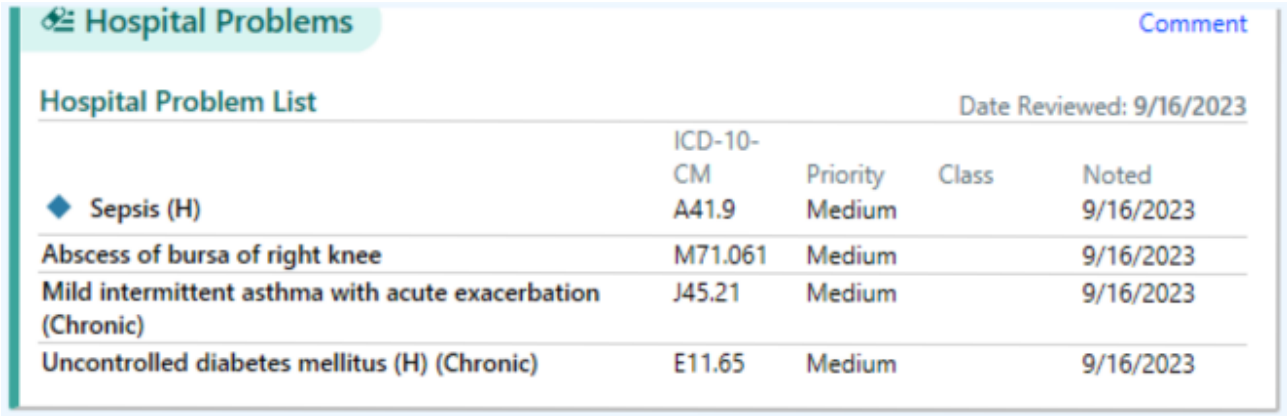

Hospital Problems

Comment

Hospital Problem List

Date Reviewed: 9/16/2023

	ICD-10-CM	Priority	Class	Noted
Sepsis (H)	A41.9	Medium		9/16/2023
Abscess of bursa of right knee	M71.061	Medium		9/16/2023
Mild intermittent asthma with acute exacerbation (Chronic)	J45.21	Medium		9/16/2023
Uncontrolled diabetes mellitus (H) (Chronic)	E11.65	Medium		9/16/2023

First, each participant reviewed each information type card to determine whether it was important enough to be included in the summary. Participants then arranged the included cards they wanted to present in the nursing summary on a desk to develop their preferred layout for the nursing summary screen. We clarified that the focus of the study was on determining the relevance of specific types of information (ie, social determinants of health) in the context of a nursing summary and arranging them, rather than the graphical design or the content of the information within the cards. We photographed each participant's design prototype for subsequent analysis. Figure 2 shows an example of a design prototype created by one participant. Following the activity, a debriefing interview was conducted to collect feedback, and each conversation was audio-recorded via Zoom for subsequent transcription and analysis. Participants were prompted with the following guided questions: (1) explain your reasoning process in relation to the design layout; (2) discuss any specific considerations you made regarding information arrangement; and (3) offer any additional comments on your design.

We collected demographic information through a survey administered after the debriefing interviews.

XSL•FO
RenderX

Figure 2. A sample summary design prototype by a participant. The layout shows the information types to include and their preferred arrangement.

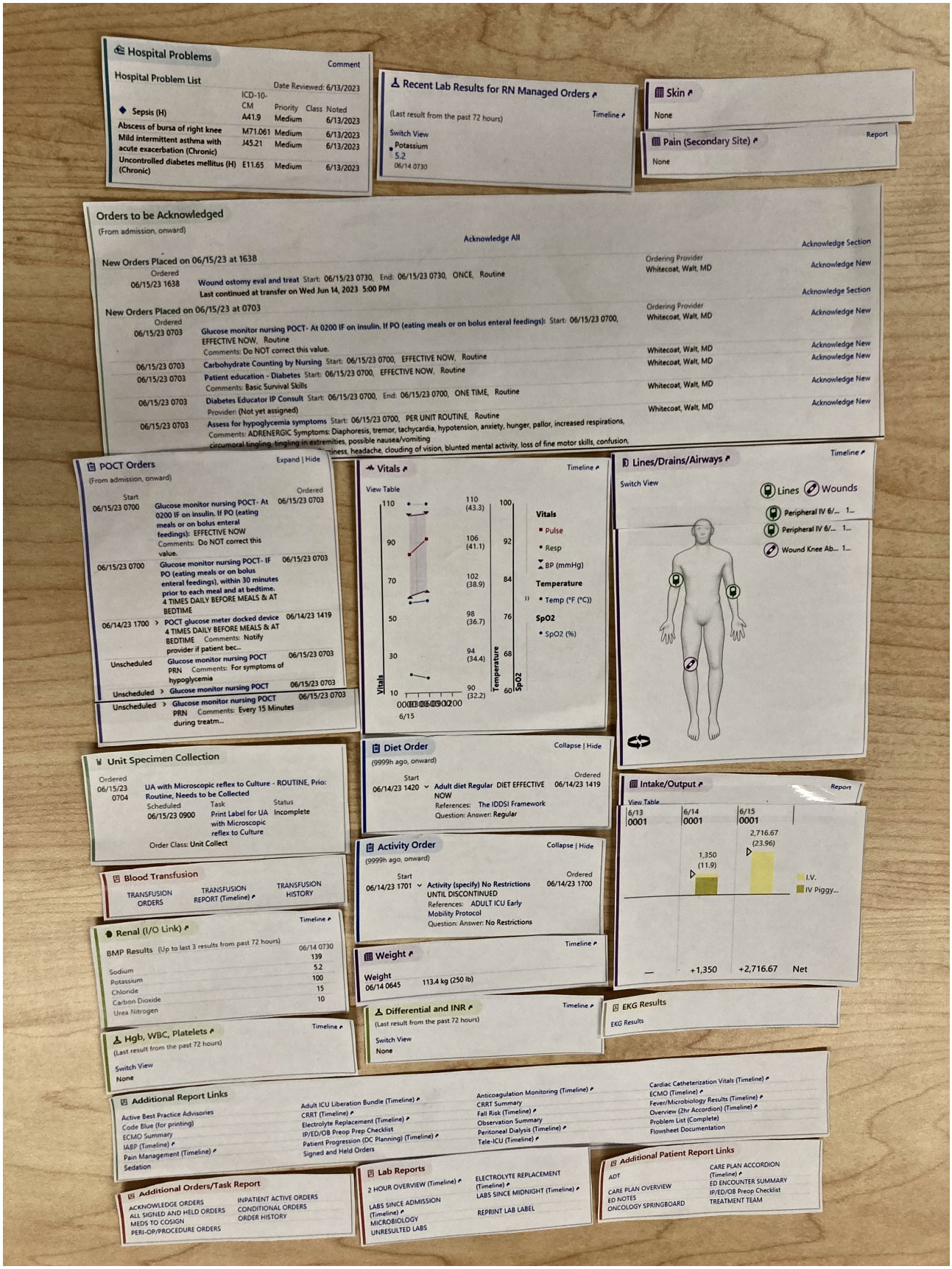

Ethical Considerations

The study protocol was approved by the University of Minnesota Institutional Review Board (STUDY00018047). The principal investigator (SP) provided verbal and written information about the study to the participants and obtained verbal consent, as the requirement for written consent was waived. All data collected during the study were deidentified and securely stored in a

XSL•FO
RenderX

university cloud-based storage system. The audio recordings and transcripts were deleted after the data analysis was completed. As this study was a part of an eye-tracking simulation, participants received a US $65 gift card for their time and effort upon completing all activities.

Data Analysis

We evaluated the relevance of each information type to the summary by quantifying the number of nurses who chose to include the information type card in the design output. We then quantified the frequency of card pairs to identify which information type cards nurses tended to place next to each other in their design prototype, providing insights into the cohesive grouping of information [11]. One researcher (SP) reviewed the deidentified transcripts of the debriefing interviews for accuracy and coded for thematic analysis. The codes and overarching themes were reviewed in weekly meetings with another researcher (JLM).

Results

Sample Characteristics

The characteristics of the 33 participants are presented in Table 1. More than half (n=19, 58%) of the participants were 25 - 44 years old, 85% (n=28) were female, 76% (n=25) were White, and none were Hispanic. Most nurses included in this analysis had at least a baccalaureate degree related to nursing (n=30, 91%). The participants generally had fewer years of nursing experience and worked fewer hours compared to a representative sample of national nurses (median=13 years and median=37 hours, respectively) [14].

XSL•FO
RenderX

Table . Baseline characteristics of the participants (N=33).

Characteristic	Value
Age (years), n (%)	
18 - 24	9 (27)
25 - 44	19 (58)
45 - 64	5 (15)
≥65	0 (0)
Sex, n (%)	
Male	5 (15)
Female	28 (85)
Other	0 (0)
Race, n (%)	
Black or African American	5 (15)
White	25 (76)
Asian	2 (6)
Native American or Alaskan	0 (0)
Other	0 (0)
Prefer not to say	1 (3)
Ethnicity, n (%)	
Hispanic or Latinx	0 (0)
Non-Hispanic	33 (100)
Other	0 (0)
Relationship status, n (%)	
Single	14 (42)
Married	14 (42)
Partnered	2 (6)
Divorced	2 (6)
Separated	1 (3)
Highest nursing degree, n (%)	
Diploma	3 (9)
BSN	28 (85)
MS or MSN	2 (6)
PhD or DNP	0 (0)
Hours worked per week, median (IQR)	36 (30-36)
Years of registered nurse experience, median (IQR)	4 (1.5-8)
Years of using electronic health records, median (IQR)	4 (1.5-8)

Card Frequency

On average, participants included 28 of the 46 information type cards in their design prototypes. Notably, all participants (N=33) included “unit specimen” (results of the analysis of body fluid, tissue, or urine), “hospital problems,” “diet,” and “activity” in their design prototype (Figure 3). However, certain information type cards were included less frequently: fewer than 50% of participants included 15 of the information type cards, and fewer than 10 participants included “key history and social determinants” (n=9), “about me” (individualized notes) (n=8), “intravenous pump settings” (n=7), “patient history notes” (n=3), and “scale and screen documentation” (n=2).

XSL•FO
RenderX

Figure 3. Frequency of included cards. Participants selected information type cards that they think are relevant to include in nursing summaries. EKG: electrocardiogram; Hgb: hemoglobin; INR: international normalized ratio; IO: intake and output; IV: intravenous; LDA: lines, drains, and airways; POCT: point-of-care testing; SDOH: social determinants of health; WBC: white blood cell.

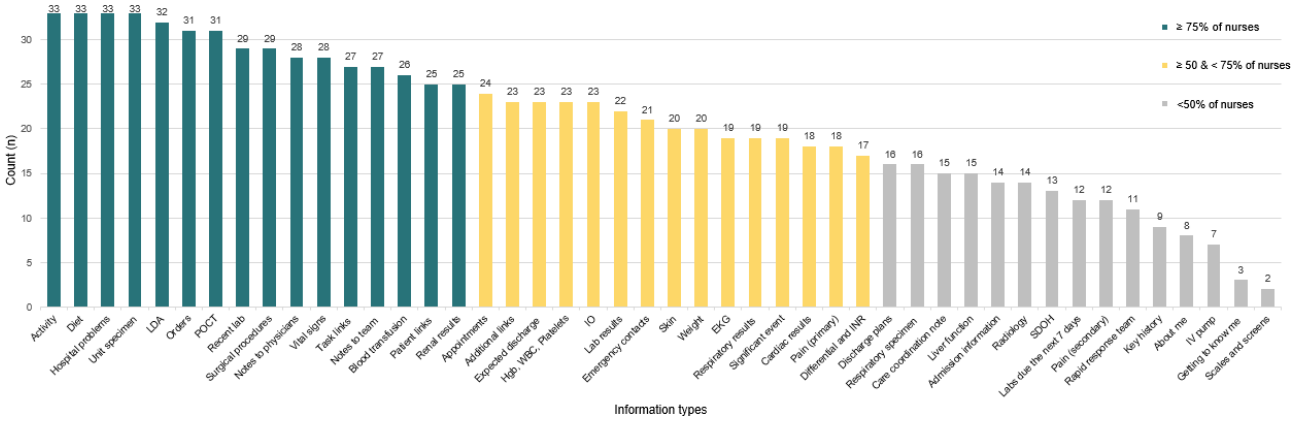

Card Adjacency

Figure 4 shows the number of participants who placed each pair of information type cards next to each other in their design. Darker shading was used to highlight cells where 2 information type cards were located next to each other more frequently. The pairs of information type cards located next to each other most frequently were "activity" and "diet" (n=26, 79%), followed by "notes to physicians" and "notes to treatment team" (n=25, 76%). Other pairs of cards that were placed next to each other included "orders" and "hospital problems" (n=16, 48%) and "orders" and "vital signs" (n=16, 48%). "Orders" and "hospital problems," and "orders" and "vital signs" are not located next to each other in the current EHR design.

Figure 4. Item-by-item matrix of card adjacency. The frequency of adjacency was manually counted when 2 cards were placed next to each other with no third card in between. EKG: electrocardiogram; Hgb: hemoglobin; INR: international normalized ratio; IO: intake and output; IV: intravenous; LDA: lines, drains, and airways; POCT: point-of-care testing; SDOH: social determinants of health; WBC: white blood cell.

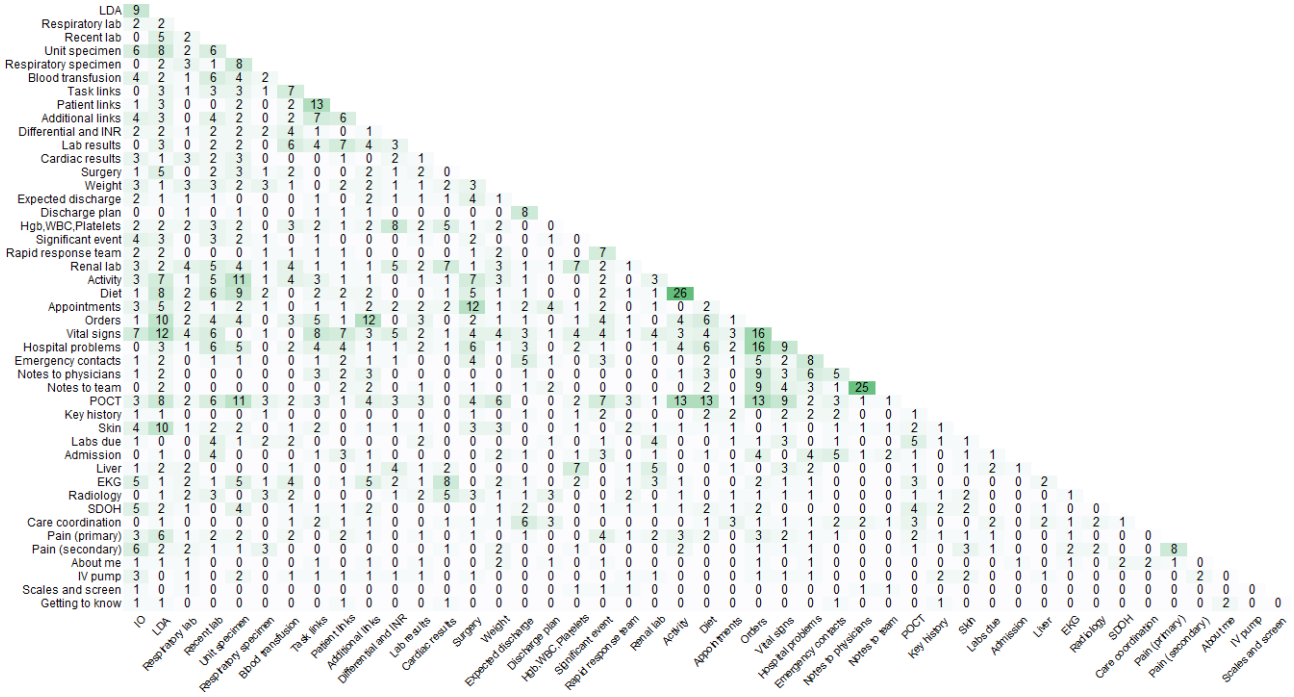

Interview Findings

Table 2 presents themes, subthemes, and exemplary quotes derived from the debriefing interviews. The participants emphasized the importance of readability in their design choices, opting for a layout that facilitates top-to-bottom reading of key information based on priority. They noted that grouping related information together would facilitate information scanning. They also prioritized identifying immediate action items (eg, physicians' orders) over delving into patient characteristics, such as social determinants of health. Owing to the typical time constraints needed to review patient information at the beginning of a shift, participants removed redundant information cards that hindered a swift and effective patient overview. Last, beyond the given design layout, participants provided design-related commentaries, pointing out that some information types required unnecessary clicks to get to the desired information, which limited usability.

XSL•FO
RenderX

Table . Common themes, subthemes, and supporting quotes from debriefing sessions based on their co-design activities, uncovering participants' reasoning process and design considerations.

Themes and subthemes	Quotes
Improving information accessibility	
Key information positioned upward	• "I have it from top to bottom. So it's the least important thing towards the bottom." [P11] • "Diet is not as important, so I kept it going to the bottom." [P24] • "I usually place it at the top of the page, not necessarily because it's important, but because it needs to be visible for everyone to see…Otherwise, no one else will notice it. Not everyone will scroll down." [P4] • "Most important would be at like the top, in a top-down order." [P1] • "I like that the orders are at the top because they're the most important; they're what the doctor wants us to be doing with the patient." [P9]
Proximity of similar information	• "I want lab results in the same spot." [P24] • "That'd be nice if they were together. Orders around orders. Orders can stick together." [P1]
Focusing on key information types	
Go-to action items over background details	• "I believe that social determinants of health are important. However, during my admission process, I do not check the social determinants of health. It is crucial for social workers and us to better understand the patient, but it is not my initial priority. When I first receive patients, I do not check the social determinants of health." [P26] • "I always want to know what orders are because these would be done right away." [P22]
Removing redundancy	• "The additional report things have a fall risk timeline, but then there's a fall risk injury (information card). I just feel like it's totally redundant." [P22] • "The card, 'Getting to know me,' I feel like this was redundant with 'About me.' I honestly never use this." [P5]
Easy-to-use design	
Unnecessary steps to the desired information	• "We cannot see the treatment team unless we do another click here and check the treating team." [P26] • "...a lot of people have a hard time with, like, everything's here but you have to click a lot to get out of here, right? If there's a transfusion order, this shouldn't be bold or an asterisk (on the order name). There should be an order you need to give blood or something that you could click on…Why do you need extra things to press?" [P22] • "I really don't use this summary page that much, because it's just so overwhelming. I guess it's just too many things there that I don't even know, like how to navigate it, which is why I made my system, which is more clicking." [P5]

Discussion

Principal Results

The co-design case study highlights the importance of aligning EHR nursing summary design with nurses' cognitive workflows, demonstrating that nurses prioritize and organize information in ways that make sense to them. Concise, relevant information allows nurses to quickly review a patient's status, while strategically grouping related data helps them navigate and interpret complex information more efficiently. The current nursing summary contains overwhelming amounts of information that is often presented ineffectively, necessitating extensive searching to consolidate scattered data. The findings indicate a significant need to remove less relevant information from the summary. Also, nurses emphasized the importance of "unit specimen" (results of the analysis of body fluid, tissue, or urine), "activity," "diet," and "hospital problems" in their design prototype, suggesting these are the key information types for patient overviews. Nurses tended to prioritize obtaining information on current problems, daily living status, or treatment plans. Medical history or miscellaneous notes can be a source of information but are not always necessary for the initial patient overview. Our findings are consistent with prior work in which clinicians used a limited number of clinical information concepts at the time of patient admission, suggesting better electronic data management strategies, including the priority display of frequently used clinical concepts [15].

XSL•FO
RenderX

Participants strategically arranged information cards to align with their cognitive workflow by placing them adjacent to other relevant cards. The adjacency of information often signifies the similarity or relative importance of the information [16]. Previous studies have also emphasized the benefits of placing paired EHR information adjacent to each other on the screen to facilitate user interface navigation [17-19]. Some information pairs nurses identified with high frequency did not align with the current nursing summary layout. For example, “orders” and “hospital problems,” “orders” and “vital signs,” “orders” and “point-of-care-testing order” (tests that are conducted at the patient’s bedside), and “point-of-care-testing order” and “activity” were frequently located next to each other in the participant designs, but not in the existing layout. In addition to the adjacent placement of relevant information, top-to-bottom organization of information was commonly found in the designs. The placement of items at the top of the screen naturally calls attention to those items [20]. Instead of scrolling down for details on the screen, nurses tended to place good-to-know but not essential information toward the bottom of their designs. The concise and cognitively aligned design supports nurses in narrowing their focus from the comprehensive patient record to a targeted subset of relevant information, thereby enhancing the efficiency of initial assessments and ensuring that pertinent details are readily accessible when required [21].

Strengths and Limitations

These findings show the value of co-design activities that involve nurses in the design process—to create an EHR nursing summary that better supports their needs. By prioritizing the most relevant patient information and streamlining the data presentation, the study offers design strategies that could alleviate the EHR burden of navigating from one piece of information to another. The identified design strategies could enhance efficiency and support the National Burden Reduction Collaborative’s mission to reduce the overall burdens on health care professionals [12].

While this case study successfully identified nurses’ information preferences aligning with their workflow, it has limitations. The use of a single simulated patient case for the co-design activities and the small sample size from a single institution may limit the generalizability of the findings. Due to the small sample size, we did not group participants by demographics (eg, unit-specific role, shift, work experience). While this study aimed to explore general information needs and design considerations for nursing summaries by including participants from multiple units, nurses’ specific needs may vary based on their characteristics. For example, nurses in surgical units may prioritize surgery-related information, while nurses in general medical units may focus on different aspects. Additionally, night shift nurses may prioritize information about upcoming examinations for the next day. Nurses’ characteristics may also influence their priorities and responsibilities, underscoring the importance of considering these variables in future research.

Conclusion

This co-design case study provides valuable insights into nurses’ preferences and strategies for organizing information in an EHR nursing summary, underscoring the need to redesign the layout based on its practical use. We emphasized the significance of reducing less relevant information and optimizing the layout according to nurses’ information priorities. These efforts have the potential to improve the usefulness of the current nursing summary. The direct involvement of nurses as co-designers holds promise for developing user-centered EHR tools that support nursing cognition, decision-making, and workflow.

Acknowledgments

We would like to express our gratitude to the participants for generously sharing their time and experiences. This study was supported by SP’s Doctoral Dissertation Fellowship, University of Minnesota and Midwest Nursing Research Society/Council for Advancement of Nursing Science Dissertation Grant (grant CON000000104586).

Data Availability

The data generated during this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Nursing summary layout.
[PDF File, 1437 KB - formative_v9i1e68906_app1.pdf]

Multimedia Appendix 2
Descriptions of information types in nursing summaries.
[DOCX File, 18 KB - formative_v9i1e68906_app2.docx]

References

XSL•FO
RenderX

1. Sentinel event data summary. The Joint Commission. 2024. URL: https://www.jointcommission.org/resources/sentinel-event/sentinel-event-data-summary/ [accessed 2024-11-19]
2. Park S, Marquard J. Nurse-centered co-design of an electronic health record nursing summary. Human Factors in Healthcare 2024 Jun;5:100065. [doi: 10.1016/j.hfh.2024.100065]
3. Staggers N, Clark L, Blaz JW, Kapsandoy S. Why patient summaries in electronic health records do not provide the cognitive support necessary for nurses' handoffs on medical and surgical units: insights from interviews and observations. Health Informatics J 2011 Sep;17(3):209-223. [doi: 10.1177/1460458211405809] [Medline: 21937463]
4. Busse TS, Jux C, Laser J, et al. Involving health care professionals in the development of electronic health records: scoping review. JMIR Hum Factors 2023 Jul 10;10:e45598. [doi: 10.2196/45598] [Medline: 37428524]
5. Voorend R, Derboven J, Slegers K. Distributed user-generated card based co-design: a case study. Presented at: CHI '19; May 4-9, 2019; Glasgow, Scotland, United Kingdom. [doi: 10.1145/3290607.3312815]
6. Ludlow K, Westbrook J, Jorgensen M, et al. Co-designing a dashboard of predictive analytics and decision support to drive care quality and client outcomes in aged care: a mixed-method study protocol. BMJ Open 2021 Aug 25;11(8):e048657. [doi: 10.1136/bmjopen-2021-048657] [Medline: 34433599]
7. Song T, Yu P, Bliokas V, et al. A clinician-led, experience-based co-design approach for developing mhealth services to support the patient self-management of chronic conditions: development study and design case. JMIR Mhealth Uhealth 2021 Jul 20;9(7):e20650. [doi: 10.2196/20650] [Medline: 34283030]
8. Alvarez CP, Martinez-Maldonado R, Buckingham Shum S. LA-DECK: a card-based learning analytics co-design tool. Presented at: LAK '20: 10th International Conference on Learning Analytics and Knowledge; Mar 23-27, 2020; Frankfurt, Germany. [doi: 10.1145/3375462.3375476]
9. Kwiatkowska J, Szóstek A, Lamas D. (Un)structured sources of inspiration: comparing the effects of game-like cards and design cards on creativity in co-design process. Presented at: PDC '14: Participatory Design Conference; Oct 6-10, 2014; Windhoek, Namibia. [doi: 10.1145/2661435.2661442]
10. Tankala S, Sherwin K. Card sorting: uncover users' mental models. : Nielsen Norman Group; 2024 Feb 2. URL: https://www.nngroup.com/articles/card-sorting-definition/ [accessed 2024-04-01]
11. Righi C, James J, Beasley M, et al. Card sort analysis best practices. J Usability Stud 2013;8(3):69-89 [FREE Full text]
12. New group hopes to focus efforts to trim clinician burdens. Health Data Management. 2023 Feb. URL: https://www.healthdatamanagement.com/articles/new-group-hopes-to-focus-efforts-to-trim-clinician-burdens [accessed 2024-11-16]
13. Martikainen S, Kaipio J, Lääveri T. End-user participation in health information systems (HIS) development: physicians' and nurses' experiences. Int J Med Inform 2020 May;137:104117. [doi: 10.1016/j.ijmedinf.2020.104117] [Medline: 32179254]
14. Melnick ER, West CP, Nath B, et al. The association between perceived electronic health record usability and professional burnout among US nurses. J Am Med Inform Assoc 2021 Jul 30;28(8):1632-1641. [doi: 10.1093/jamia/ocab059] [Medline: 33871018]
15. Pickering BW, Gajic O, Ahmed A, Herasevich V, Keegan MT. Data utilization for medical decision making at the time of patient admission to ICU. Crit Care Med 2013 Jun;41(6):1502-1510. [doi: 10.1097/CCM.0b013e318287f0c0] [Medline: 23528804]
16. Wagemans J, Elder JH, Kubovy M, et al. A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure-ground organization. Psychol Bull 2012 Nov;138(6):1172-1217. [doi: 10.1037/a0029333] [Medline: 22845751]
17. Marian AA, Dexter F, Tucker P, Todd MM. Comparison of alphabetical versus categorical display format for medication order entry in a simulated touch screen anesthesia information management system: an experiment in clinician-computer interaction in anesthesia. BMC Med Inform Decis Mak 2012 May 29;12(1):46. [doi: 10.1186/1472-6947-12-46] [Medline: 22643058]
18. Senathirajah Y. Safer design - composable EHRs and mechanisms for safety. Stud Health Technol Inform 2015;218:86-92. [Medline: 26262532]
19. Zheng K, Padman R, Johnson MP, Diamond HS. An interface-driven analysis of user interactions with an electronic health records system. J Am Med Inform Assoc 2009;16(2):228-237. [doi: 10.1197/jamia.M2852] [Medline: 19074301]
20. Senathirajah Y, Kaufman D, Bakken S. The clinician in the driver's seat: part 2 - intelligent uses of space in a drag/drop user-composable electronic health record. J Biomed Inform 2014 Dec;52:177-188. [doi: 10.1016/j.jbi.2014.09.008] [Medline: 25445921]
21. Horsky J, Ramelson HZ. Development of a cognitive framework of patient record summary review in the formative phase of user-centered design. J Biomed Inform 2016 Dec;64:147-157. [doi: 10.1016/j.jbi.2016.10.004] [Medline: 27725292]

Abbreviations

EHR: electronic health record

XSL•FO
RenderX

Edited by A Mavragani; submitted 19.11.24; peer-reviewed by A Kuusisto, M Alshammari; revised version received 15.02.25; accepted 19.02.25; published 11.03.25.

Please cite as:
Park S, Marquard JL, Austin RR, Martin CL, Pieczkiewicz DS, Delaney CW
Exploratory Co-Design on Electronic Health Record Nursing Summaries: Case Study
JMIR Form Res 2025;9:e68906
URL: https://formative.jmir.org/2025/1/e68906
doi:10.2196/68906

© Suhyun Park, Jenna L Marquard, Robin R Austin, Christie L Martin, David S Pieczkiewicz, Connie W Delaney. Originally published in JMIR Formative Research (https://formative.jmir.org), 11.3.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Improving the User Interface and Guiding the Development of Effective Training Material for a Clinical Research Recruitment and Retention Dashboard: Usability Testing Study

Leah Leslie Gardner[1], PhD; Pezhman Raeisian Parvari[1], MA; Mark Seidman[1], MA; Richard J Holden[1,2,3,4], PhD; Nicole R Fowler[1,2,4], MHSA, PhD; Ben L Zarzaur[5], MPH, MD; Diana Summanwar[6], MD; Cristina Barboi[7], MS, MD; Malaz Boustani[1,2,4,8], MPH, MD

[1]Center for Health Innovation and Implementation Science, School of Medicine, Indiana University, 410 W 10th St #1140, Indianapolis, IN, United States

[2]Department of Medicine, School of Medicine, Indiana University, Indianapolis, IN, United States

[3]School of Public Health, Indiana University Bloomington, Bloomington, IN, United States

[4]Center for Aging Research, Regenstrief Institute Inc, Indianapolis, IN, United States

[5]Department of Surgery, Division of Acute Care and Regional General Surgery, University of Wisconsin-Madison, Madison, WI, United States

[6]Department of Family Medicine, School of Medicine, Indiana University, Indianapolis, IN, United States

[7]Department of Anesthesiology, School of Medicine, Indiana University, Indianapolis, IN, United States

[8]Sandra Eskenazi Center for Brain Care Innovation, Eskenazi Health, Indianapolis, IN, United States

Corresponding Author:
Pezhman Raeisian Parvari, MA
Center for Health Innovation and Implementation Science, School of Medicine, Indiana University, 410 W 10th St #1140, Indianapolis, IN, United States

Abstract

Background: Participant recruitment and retention are critical to the success of clinical trials, yet challenges such as low enrollment rates and high attrition remain ongoing obstacles. RecruitGPS is a scalable dashboard with integrated control charts to address these issues by providing real-time data monitoring and analysis, enabling researchers to better track and improve recruitment and retention.

Objective: This study aims to identify the challenges and inefficiencies users encounter when interacting with the RecruitGPS dashboard. By identifying these issues, the study aims to inform strategies for improving the dashboard's user interface and create targeted, effective instructional materials that address user needs.

Methods: Twelve clinical researchers from the Midwest region of the United States provided feedback through a 10-minute, video-recorded usability test session, during which participants were instructed to explore the various tabs of the dashboard, identify challenges, and note features that worked well while thinking aloud. Following the video session, participants took a survey on which they answered System Usability Scale (SUS) questions, ease of navigation questions, and a Net Promoter Score (NPS) question.

Results: A quantitative analysis of survey responses revealed an average SUS score of 61.46 (SD 23.80; median 66.25) points, indicating a need for improvement in the user interface. The NPS was 8, with 4 of 12 (33%) respondents classified as promoters and 3 of 12 (25%) as detractors, indicating a slightly positive satisfaction. When participants compared RecruitGPS to other recruitment and study management tools they had used, 8 of 12 (67%) of participants rated RecruitGPS as better or much better. Only 1 of 12 (8%) participants rated RecruitGPS as worse but not much worse. A qualitative analysis of participants' interactions with the dashboard diagnosed a confusing part of the dashboard that could be eliminated or made optional and provided valuable insight for the development of instructional videos and documentation. Participants liked the dashboard's data visualization capabilities, including intuitive graphs and trend tracking; progress indicators, such as color-coded status indicators and comparison metrics; and the overall dashboard's layout and design, which consolidated relevant data on a single page. Users also valued the accuracy and real-time updates of data, especially the integration with external sources like Research Electronic Data Capture (REDCap).

Conclusions: RecruitGPS demonstrates significant potential to improve the efficiency of clinical trials by providing researchers with real-time insights into participant recruitment and retention. This study offers valuable recommendations for targeted refinements to enhance the user experience and maximize the dashboard's effectiveness. Additionally, it highlights navigation challenges that can be addressed through the development of clear and focused instructional videos.

XSL•FO
RenderX

(JMIR Form Res 2025;9:e66718) doi:10.2196/66718

KEYWORDS

recruitment strategies; clinical research; research subject recruitment; agile science; agile implementation; human-computer interaction

Introduction

Clinical studies are essential to advancing medical science and improving patient outcomes, and their successful completion relies on adequate participant recruitment [1]. Effective and efficient patient recruitment and retention ensure timely collection of data, while insufficient enrollment often leads to delays or even trial failure [2-5]. Studies indicate that 80% of trials face recruitment challenges [6], while 37% fail to meet their target sample size goals [7]. In fact, inadequate recruitment is the leading cause of trial discontinuation [8,9].

Clinical trial management systems (CTMSs), clinical trial recruitment support systems (CTRSSs), web-based platforms, patient registries, social media, and community engagement initiatives [10-14] are widely used to support recruitment and management of clinical trials. Electronic health records (EHRs) facilitate targeted recruitment by identifying potential participants based on health data, while CTMSs track participant enrollment, scheduling, and communication [15,16]. Online platforms like ClinicalTrials.gov provide global access to trial information, and social media platforms such as Facebook enable broader outreach through targeted advertisements [17,18]. While these tools offer valuable support for recruitment, they often face limitations such as high costs, limited reach, or sustainability challenges [19].

Dashboards are effective tools in clinical settings, particularly for tracking quality and safety metrics [20]. In clinical trials, dashboards are used to generate alerts when participant accrual is inadequate [3,21]. Commercially available CTMSs with dashboards to support large-scale clinical trials include Medidata Rave (Medidata Solutions) [22], Veeva Vault CDMS (Veeva Systems) [23], Ennov Clinical (Ennov) [24], Viedoc (Viedoc Technologies AB) [25], RealTime eClinical (RealTime Software Solutions) [26], and Clinevo (Clinevo Technologies) [27]. These systems have a variety of features for trial oversight, from patient recruitment to data monitoring, such as patient portals to streamline patient communication, consent capture, financial tracking, regulatory compliance, adverse event tracking, and study closeout.

Castor EDC (Castor) [28] and Ripple Science (Ripple Science Inc) [29] offer affordable packages for academic institutions and smaller companies. These packages have dashboards that help researchers track study progress, tasks, and participant data across different trials and that automate email or in-app notifications and reminders to both staff and participants.

OpenClinica (OpenClinica LLC) [30] and REDCap (Research Electronic Data Capture; Vanderbilt University) [31] are examples of open-source platforms for managing clinical trial data. OpenClinica Insight is an add-on module for OpenClinica 3 Enterprise and OpenClinica 4 that allows the user to create reports and dashboards. REDCap has limited dashboarding capabilities and requires the user to export data to a spreadsheet, such as Microsoft Excel, or statistical package, such as R and R Shiny, for detailed analysis and dashboard development.

Since REDCap requires the user to export data to a spreadsheet or other software for analysis, a research group from the Center for Health Innovation and Implementation Science (CHIIS) successfully developed and implemented an Excel-based dashboard in a single, multisite clinical study already in progress using REDCap [32]. This dashboard integrates control charts and enables seamless data updates from the REDCap database, allowing for effective monitoring of recruitment and participant progress. In this study, participants moved through various stages of the research workflow. The dashboard provided investigators with weekly updates on participant progress through these stages, facilitating timely interventions when participant counts or progress rates at any stage suggested that sample size targets might not be met. Additionally, it enabled the application of Agile science methods [33-37] to rapidly test and implement strategies for improving recruitment.

Building on the success of the dashboard used in the multisite clinical study, the research group developed a customizable, generic dashboard called RecruitGPS, which can be customized for other clinical research studies. This paper focuses on the detailed usability testing of the RecruitGPS user interface. A comprehensive description of RecruitGPS features is provided in Multimedia Appendix 1. The aim of this study is to identify the challenges and inefficiencies users experienced with the RecruitGPS interface so that these insights can be leveraged to improve usability and develop effective training materials.

Methods

Overview of the RecruitGPS Dashboard

The goal of the user interface is to provide clinical trial principal investigators and research coordinators with immediate access to weekly and cumulative counts at each stage of the study and weekly and cumulative transition rates from stage to stage. The interface uses compelling visual cues to bring the user's attention to potential problems in the study workflow. The first version of the user interface for RecruitGPS seen in Figure 1 was the version from the successful multisite study dashboard user interface [32] with the generic names step 1 through step 9 substituted for the names of the stages screened through 18-month assessment. The multisite study user interface was collaboratively designed by the stakeholders in the study and the dashboard developer.

XSL•FO
RenderX

Figure 1. The RecruitGPS dashboard interface evaluated during usability testing.

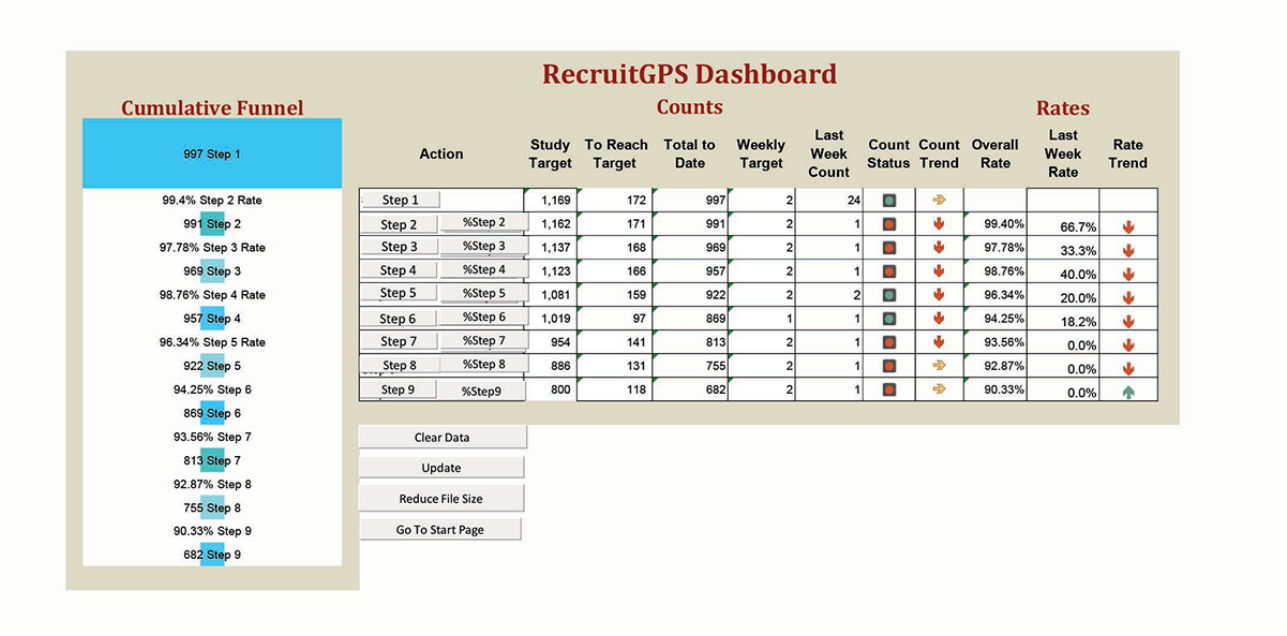

Think-aloud usability testing sessions were combined with a user survey to assess the effectiveness of, efficiency of, and satisfaction with the RecruitGPS dashboard. Effectiveness is the extent to which the user can achieve a goal with accuracy and completeness. Efficiency is the effort and resources necessary to achieve a complete and accurate goal. Satisfaction is a positive user experience and absence of discontent during task performance. Once the 3 are fulfilled adequately, the product can be considered to have attained an acceptable level of usability [38].

Participant Recruitment

Participants in the study were recruited to use the RecruitGPS user interface while being video recorded and to complete a survey about their experience. Most participants were affiliated with CHIIS or were clinical researchers primarily based in the Midwest region of the United States. Recruitment and usability testing were conducted in August 2024.

Participants were recruited using convenience and snowball sampling techniques, with the CHIIS team supporting the process through digital advertisements, emails, and word of mouth. To be eligible, participants needed experience with clinical research studies, fluency in English, and access to the internet and Zoom (Zoom Video Communications Inc.). Although the literature suggests that 5 to 8 participants are sufficient for think-aloud usability testing, 12 participants were recruited to ensure diverse and comprehensive insights [39,40].

Usability Testing Sessions and Surveys

A total of twelve 10-minute usability testing sessions of the RecruitGPS dashboard were conducted via Zoom. During each session, the facilitator shared their screen, displaying the dashboard, and participants were provided with remote control access to interact with the interface. Participants were instructed as follows: “consider[ing] a research project of your own, explore the various tabs of the dashboard, identify challenges, and note features that work well while thinking aloud.” To identify specific problems, the researchers recorded the 10-minute usability testing sessions for a qualitative analysis of the challenges users encountered and features that worked well.

Following each usability test, participants received a survey link to provide additional feedback. The survey gathered insights on users’ overall experience, including their System Usability Scale (SUS) (Textbox 1) [41] and Net Promoter Score (NPS) [42] ratings. The NPS is based on the question “On a scale of 0 to 10, how likely are you to recommend this dashboard to a colleague or friend?” It is reported as a number ranging from −100 to +100, where a higher score is desirable. Respondents are categorized as detractors (giving scores of 0 - 6), passives (giving scores of 7 - 8), and promoters (giving scores of 9 - 10) The percentages of promoters and detractors are calculated and then the percentage of detractors is subtracted from the percentage of promoters to obtain the NPS [42]. The survey also included open-ended questions about the most and least useful features, as well as overall impressions of the user experience and dashboard interface. To assess ease of navigation in the survey, users were asked the questions in Table 1.

XSL•FO
RenderX

Textbox 1. System Usability Scale (SUS) questions in the survey following the usability test.

SUS Questions (1=strongly disagree, 2=disagree, 3=neutral, 4=agree, 5=strongly agree)

1. I think that I would like to use this system frequently.
2. I found the system unnecessarily complex.
3. I thought the system was easy to use.
4. I think that I would need the support of a technical person to be able to use this system.
5. I found the various functions in this system were well integrated.
6. I thought there was too much inconsistency in this system.
7. I would imagine that most people would learn to use this system very quickly.
8. I found the system very cumbersome to use.
9. I felt very confident using the system.
10. I needed to learn a lot of things before I could get going with this system.

Table . Questions on ease of navigation in the survey following the usability test.

Question	Available responses
How easy was it to navigate through the dashboard?	Likert scale (1 - 5, with 1 being the most difficult and 5 being the easiest)
Overall, how does this recruitment dashboard compare to other recruitment tools you have used?	Scale: 1=much worse, 2=worse, 3=about the same, 4=better, 5=much better
How intuitive did you find the user interface (UI) of this recruitment dashboard?	Scale: 1=not intuitive at all, 2=slightly intuitive, 3=neutral, 4=intuitive, 5=very intuitive

Ethical Considerations

This study was approved by the Human Research Protection Program at Indiana University (23755) prior to participant recruitment. Informed consent was obtained from all participants. Throughout the study, strict confidentiality and privacy protocols were followed. All data collected were deidentified before analysis to maintain participant anonymity. To protect privacy, secure data storage and anonymization procedures were used. Furthermore, no identifiable images of participants were included in the data. Participation was voluntary and no financial compensation or incentives were provided.

Results

This analysis assesses the effectiveness and efficiency of the RecruitGPS dashboard and satisfaction with it.

Participants

Twelve medical professionals completed the usability testing and surveys. Their demographics are summarized in Table 2.

XSL•FO
RenderX

Table . Education levels, roles, and experience of study participants (N=12) completing the usability testing and surveys.

Characteristics	Participants, n (%)
Education	
Doctorate (PhD)	2 (17)
Master's degree	4 (33)
Professional degree (MD, JD)	6 (50)
Role in research	
Principal investigator or coprincipal investigator	5 (42)
Project manager	5 (42)
Research coordinator	2 (17)
Experience	
1 - 3 years	4 (33)
4 - 6 years	2 (17)
7 - 10 years	1 (8)
Less than 1 year	1 (8)
More than 10 years	4 (33)

Effectiveness

The 10 questions on the SUS together assess overall usability, computed on a scale from 0 to 100, as described in detail by Brooke [41]. Table 3 reports individual item scores and overall SUS scores for each participant. Higher ratings on odd-numbered questions and lower ratings on even-numbered questions indicate better usability.

Table . System Usability Scale (SUS) total scores and scores for each question (Q) by participant (P).

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12
Q1	2	3	2	4	3	3	4	1	4	4	3	4
Q2	1	4	1	1	2	3	3	1	2	1	3	4
Q3	1	4	1	3	3	3	3	0	3	3	3	4
Q4	1	4	0	2	3	0	3	0	1	3	3	4
Q5	2	3	2	3	3	3	2	1	2	3	3	4
Q6	4	4	2	2	3	3	3	1	4	3	3	4
Q7	1	3	2	4	2	3	3	0	3	3	3	4
Q8	2	4	1	4	2	3	3	1	1	3	3	4
Q9	0	2	0	3	2	3	2	0	1	3	3	4
Q10	1	1	0	4	2	1	2	1	4	2	3	4
SUS score	37.5	80.0	27.5	75.0	62.5	62.5	70.0	15.0	62.5	70.0	75.0	100.0

The mean SUS total score for all participants was 61.46 (SD 23.80; median 66.25) points. According to Bangor et al [43], a score of 85 and above is excellent and a score of 70 - 84 is good.

Efficiency—Assessing Ease of Navigation

Feedback on ease of navigation showed a mixed picture. While 25% (n=3) of the 12 respondents gave a rating of 5% and 58% (n=7) gave a rating of 4, 17% (n=2) rated ease of navigation a 1.

Satisfaction

The NPS [42] for RecruitGPS in this study was 8, with 33% (n=4) of respondents classified as promoters and 25% (n=3) as detractors. This indicates a slightly positive satisfaction.

A second question asked participants to compare RecruitGPS to other recruitment tools they had used on a scale of 1 - 5, with 1 meaning RecruitGPS is much worse and 5 meaning Recruit GPS is much better. The results were that 8 of 12 participants thought RecruitGPS was better. The results are summarized in Table 4.

XSL•FO
RenderX

Table . Survey responses (N=12) comparing RecruitGPS to other recruitment tools used by participants.

Score	Interpretation	Number of responses, n (%)
1	Much worse	0 (0)
2	Worse	1 (8)
3	About the same	3 (25)
4	Better	5 (42)
5	Much better	3 (25)

Qualitative Feedback From Participants

The framework developed by Clarke and Braun [44] was used for thematic analysis of the recorded 10-minute usability testing sessions to identify and analyze patterns within the data, including challenges users encountered and features that worked well.

Challenges

Several users reported difficulties with specific graphical elements, such as the cumulative funnel, which they found confusing or unintuitive. Additionally, some users experienced challenges with unclear instructions and labels, leading to difficulties in interpreting data and understanding the dashboard's functionality. These issues were reflected in the variability of ease of navigation ratings, with some users finding the dashboard extremely difficult to use.

Features That Worked Well

Positive feedback centered on the dashboard's data visualization capabilities, including intuitive graphs and trend tracking, which were highly valued by users. Progress indicators, such as color-coded status indicators and comparison metrics, were appreciated for helping users monitor their recruitment goals effectively. The dashboard's layout and design, which consolidated relevant data on a single page with clear button labeling, contributed to user satisfaction. Users also valued the accuracy and real-time updates of data, especially the integration with external sources like REDCap, which was seen as a critical feature for effective data management.

Discussion

Principal Findings

We found that 8 of 12 (67%) participants considered RecruitGPS better than other recruitment tools they had used, but the usability (SUS), satisfaction (NPS), and efficiency scores all indicated a need for improvement. Feedback on ease of navigation showed significant efficiency challenges faced by a minority of users, suggesting that some users struggled with navigation, despite overall positive feedback.

The recommendations for improvement based on the qualitative analysis are follows:

1. Remove the cumulative funnel or make it an optional feature for users who request it.
2. Create videos with accompanying documentation to address challenges users encountered with (a) instructions and labels and (b) navigation.

Although the quantitative analysis highlighted areas for improvement, the testing was conducted without any prior training for participants. The developers of RecruitGPS had not anticipated users attempting to operate the system without access to documentation or instructional materials, such as training videos. Despite this, RecruitGPS performed well enough. Additionally, the qualitative analysis offered valuable insights for enhancing the documentation and provided clear guidance for the development of instructional videos.

Limitations

The first limitation of this study was the relatively small sample size, consisting of only 12 participants in the usability testing sessions and survey responses. Conducting future studies with a larger and more diverse user group could yield additional insights and a broader understanding of user needs.

A related limitation was the participants' affiliations, as most were connected to 1 of 2 medical schools or their associated health care systems. While some participants had experience at other research institutions, their familiarity with alternative recruitment tools may have been limited to those used within these 2 medical schools, potentially restricting the study's generalizability.

Another limitation was that the study did not include testing of the existing documentation, and the instructional videos had not yet been developed. Future testing of these resources could provide valuable feedback to refine and improve their effectiveness.

Conclusions

RecruitGPS demonstrates promise and potential in enhancing recruitment and retention for clinical trials by providing real-time insights, supporting agile decision-making, and using scalable design with an accessible user interface. Removing the cumulative funnel and creating instructional videos and documentation focused on navigation challenges are likely to enhance its usability. With these enhancements, RecruitGPS could become a key tool in clinical trial management, supporting successful participant recruitment and better research outcomes.

XSL•FO
RenderX

Acknowledgments

The study was supported by a grant (R25AG078136) from the National Institute on Aging.

Data Availability

The datasets generated during and/or analyzed during this study are anonymized and restricted for confidentiality. Any further inquiries can be directed to the corresponding author.

Conflicts of Interest

MB serves as a chief scientific officer and cofounder of BlueAgilis and the chief health officer of DigiCare Realized. He has equity interest in Blue Agilis and DigiCare Realized; he has sold equity in Preferred Population Health Management and MyShift (previously known as RestUp). He serves as an advisory board member for Eli Lilly and Co, Eisai; Merck & Co; Biogen; and Genentech. These conflicts have been reviewed by Indiana University and have been appropriately managed to maintain objectivity.

Multimedia Appendix 1
Detailed features of RecruitGPS dashboard.
[DOC File, 251 KB - formative_v9i1e66718_app1.doc]

References

1. Campbell MK, Snowdon C, Francis D, et al. Recruitment to randomised trials: strategies for trial enrollment and participation study. The STEPS study. Health Technol Assess 2007 Nov;11(48):iii, ix-105. [doi: 10.3310/hta11480] [Medline: 17999843]
2. Visanji EC, Oldham JA. Patient recruitment in clinical trials: a review of literature. Phys Ther Rev 2001 Jun 1;6(2):141-150. [doi: 10.1179/108331901786166532]
3. Toddenroth D, Sivagnanasundaram J, Prokosch HU, Ganslandt T. Concept and implementation of a study dashboard module for a continuous monitoring of trial recruitment and documentation. J Biomed Inform 2016 Dec;64:222-231. [doi: 10.1016/j.jbi.2016.10.010] [Medline: 27769890]
4. Vadeboncoeur C, Foster C, Townsend N. Challenges of research recruitment in a university setting in England. Health Promot Int 2018 Oct 1;33(5):878-886. [doi: 10.1093/heapro/dax025] [Medline: 28531308]
5. Lai YS, Afseth JD. A review of the impact of utilising electronic medical records for clinical research recruitment. Clin Trials 2019 Apr;16(2):194-203. [doi: 10.1177/1740774519829709] [Medline: 30764659]
6. Brøgger-Mikkelsen M, Ali Z, Zibert JR, Andersen AD, Thomsen SF. Online patient recruitment in clinical trials: systematic review and meta-analysis. J Med Internet Res 2020 Nov 4;22(11):e22179. [doi: 10.2196/22179] [Medline: 33146627]
7. Jacques RM, Ahmed R, Harper J, et al. Recruitment, consent and retention of participants in randomised controlled trials: a review of trials published in the National Institute for Health Research (NIHR) Journals Library (1997-2020). BMJ Open 2022 Feb 14;12(2):e059230. [doi: 10.1136/bmjopen-2021-059230] [Medline: 35165116]
8. Pung J, Rienhoff O. Key components and IT assistance of participant management in clinical research: a scoping review. JAMIA Open 2020 Oct;3(3):449-458. [doi: 10.1093/jamiaopen/ooaa041] [Medline: 33215078]
9. Fitzer K, Haeuslschmid R, Blasini R, et al. Patient recruitment system for clinical trials: mixed methods study about requirements at ten university hospitals. JMIR Med Inform 2022 Apr 20;10(4):e28696. [doi: 10.2196/28696] [Medline: 35442203]
10. Cuggia M, Besana P, Glasspool D. Comparing semi-automatic systems for recruitment of patients to clinical trials. Int J Med Inform 2011 Jun;80(6):371-388. [doi: 10.1016/j.ijmedinf.2011.02.003] [Medline: 21459664]
11. Obeid JS, Beskow LM, Rape M, et al. A survey of practices for the use of electronic health records to support research recruitment. J Clin Transl Sci 2017 Aug;1(4):246-252. [doi: 10.1017/cts.2017.301] [Medline: 29657859]
12. Darmawan I, Bakker C, Brockman TA, Patten CA, Eder M. The role of social media in enhancing clinical trial recruitment: scoping review. J Med Internet Res 2020 Oct 26;22(10):e22810. [doi: 10.2196/22810] [Medline: 33104015]
13. Sampson R, Shapiro S, He W, et al. An integrated approach to improve clinical trial efficiency: linking a clinical trial management system into the Research Integrated Network of Systems. J Clin Transl Sci 2022;6(1):e63. [doi: 10.1017/cts.2022.382] [Medline: 35720964]
14. Darko EM, Kleib M, Olson J. Social media use for research participant recruitment: integrative literature review. J Med Internet Res 2022 Aug 4;24(8):e38015. [doi: 10.2196/38015] [Medline: 35925655]
15. Köpcke F, Prokosch HU. Employing computers for the recruitment into clinical trials: a comprehensive systematic review. J Med Internet Res 2014 Jul 1;16(7):e161. [doi: 10.2196/jmir.3446] [Medline: 24985568]
16. Gupta A, Calfas KJ, Marshall SJ, et al. Clinical trial management of participant recruitment, enrollment, engagement, and retention in the SMART study using a marketing and information technology (MARKIT) model. Contemp Clin Trials 2015 May;42:185-195. [doi: 10.1016/j.cct.2015.04.002] [Medline: 25866383]
17. Richesson RL, Krischer J. Data standards in clinical research: gaps, overlaps, challenges and future directions. J Am Med Inform Assoc 2007;14(6):687-696. [doi: 10.1197/jamia.M2470] [Medline: 17712081]

XSL•FO
RenderX

18. Akers L, Gordon JS. Using Facebook for large-scale online randomized clinical trial recruitment: effective advertising strategies. J Med Internet Res 2018 Nov 8;20(11):e290. [doi: 10.2196/jmir.9372] [Medline: 30409765]
19. Muldowney LA, Hynes SM, Oglesby M, Dwyer CP. A qualitative exploration of advantages and disadvantages to using technology in the process of randomised controlled trial recruitment. HRB Open Res 2023;6:56. [doi: 10.12688/hrbopenres.13776.2] [Medline: 39211398]
20. Rabiei R, Almasi S. Requirements and challenges of hospital dashboards: a systematic literature review. BMC Med Inform Decis Mak 2022 Nov 8;22(1):287. [doi: 10.1186/s12911-022-02037-8] [Medline: 36348339]
21. Afroz MA, Schwarber G, Bhuiyan MAN. Risk-based centralized data monitoring of clinical trials at the time of COVID-19 pandemic. Contemp Clin Trials 2021 May;104:106368. [doi: 10.1016/j.cct.2021.106368] [Medline: 33775899]
22. Rave EDC. Medidata. 2024. URL: https://www.medidata.com/en/clinical-trial-products/clinical-data-management/edc-systems/ [accessed 2024-09-25]
23. Veeva Systems. 2024. URL: https://www.veeva.com/products/clinical-data-management/ [accessed 2024-09-25]
24. Ennov. 2024. URL: https://en.ennov.com/solutions/clinical/ [accessed 2024-09-25]
25. Viedoc. 2024. URL: https://www.viedoc.com/campaign/eclinical-software/ [accessed 2024-09-25]
26. RealTime eClinical Solutions. 2024. URL: https://realtime-eclinical.com/ [accessed 2024-09-25]
27. Clinevo Technologies. 2024. URL: https://www.clinevotech.com/ [accessed 2024-09-25]
28. Castor. 2024. URL: https://www.castoredc.com/ [accessed 2024-09-25]
29. Ripple Science. 2024. URL: https://www.ripplescience.com/ [accessed 2024-09-25]
30. OpenClinica. 2024. URL: https://www.openclinica.com/ [accessed 2024-09-25]
31. REDCap. URL: https://www.project-redcap.org/ [accessed 2025-02-06]
32. Gardner L, Bylund P, Robbins S, et al. Agile monitoring dashboard for clinical research studies. Trials 2024 Nov 29;25(1):802. [doi: 10.1186/s13063-024-08646-0] [Medline: 39609928]
33. Boustani M, Alder CA, Solid CA. Agile implementation: a blueprint for implementing evidence-based healthcare solutions. J Am Geriatr Soc 2018 Jul;66(7):1372-1376. [doi: 10.1111/jgs.15283] [Medline: 29513360]
34. Boustani M, Azar J, Solid CA. Agile Implementation: A Model for Implementing Evidence-Based Healthcare Solutions into Real-World Practice to Achieve Sustainable Change: Morgan James Publishing; 2020.
35. Holden RJ, Boustani MA, Azar J. Agile Innovation to transform healthcare: innovating in complex adaptive systems is an everyday process, not a light bulb event. BMJ Innov 2021 Apr;7(2):499-505. [doi: 10.1136/bmjinnov-2020-000574]
36. Mehta J, Long E, Bynagari V, et al. Creating demand for unmet needs: agile storytelling. Front Health Serv 2024;4:1376695. [doi: 10.3389/frhs.2024.1376695] [Medline: 39544454]
37. Holden RJ, Boustani M. The value of an 'agile' mindset in times of crisis. Modern Healthcare. 2020 May 11. URL: https://www.modernhealthcare.com/opinion-editorial/value-agile-mindset-times-crisis [accessed 2025-02-06]
38. Bevan N, Carter J, Harker S. ISO 9241-11 revised: what have we learnt about usability since 1998? In: Kurosu M, editor. Human-Computer Interaction: Design and Evaluation. HCI 2015. Lecture Notes in Computer Science(), Vol 9169: Springer; 2015. [doi: 10.1007/978-3-319-20901-2_13]
39. Clemmensen T, Hertzum M, Hornbæk K, Shi Q, Yammiyavar P. Cultural cognition in the thinking-aloud method for usability evaluation. In: ICIS 2008 Proceedings: Association for Information Systems; 2008:189.
40. van Someren M, Barnard YF, Sandberg J. The Think Aloud Method: A Practical Approach to Modelling Cognitive Processes: Academic Press; 1994.
41. Brooke J. SUS: a 'quick and dirty' usability scale. In: Usability Evaluation In Industry: CRC Press; 1996.
42. Reichheld FF. The one number you need to grow. Harv Bus Rev 2003 Dec;81(12):46-54. [Medline: 14712543]
43. Bangor A, Kortum P, Miller J. Determining what individual SUS scores mean: adding an adjective rating scale. J Usability Stud 2009;4(3):114-123 [FREE Full text]
44. Clarke V, Braun V. Thematic analysis. J Posit Psychol 2017 May 4;12(3):297-298. [doi: 10.1080/17439760.2016.1262613]

Abbreviations

CHIIS: Center for Health Innovation and Implementation Science
CTMS: clinical trial management system
CTRSS: clinical trial recruitment support system
EHR: electronic health record
NPS: Net Promoter Score
REDCap: Research Electronic Data Capture; Vanderbilt University
SUS: System Usability Scale

XSL•FO
RenderX

Edited by A Mavragani; submitted 20.09.24; peer-reviewed by PV Asaro, S Jimenez-Jorge; revised version received 20.01.25; accepted 22.01.25; published 24.02.25.

Please cite as:
Gardner LL, Raeisian Parvari P, Seidman M, Holden RJ, Fowler NR, Zarzaur BL, Summanwar D, Barboi C, Boustani M
Improving the User Interface and Guiding the Development of Effective Training Material for a Clinical Research Recruitment and Retention Dashboard: Usability Testing Study
JMIR Form Res 2025;9:e66718
URL: https://formative.jmir.org/2025/1/e66718
doi:10.2196/66718

© Leah Leslie Gardner, Pezhman Raeisian Parvari, Mark Seidman, Richard J Holden, Nicole R Fowler, Ben L Zarzaur, Diana Summanwar, Cristina Barboi, Malaz Boustani. Originally published in JMIR Formative Research (https://formative.jmir.org), 24.2.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Exploring the Relationship Between Smartphone GPS Patterns and Quality of Life in Patients With Advanced Cancer and Their Family Caregivers: Longitudinal Study

Kyungmi Lee[1], PhD, RN; Andres Azuero[2], MBA, PhD; Sally Engler[2], MPH; Sidharth Kumar[3], PhD; Frank Puga[2], PhD; Alexi A Wright[4], MPH, MD; Arif Kamal[5], MHS, MBA, MD; Christine S Ritchie[6], MSPH, MD; George Demiris[7], FACMI, PhD; Marie A Bakitas[2,8,9], NP-C, DNSc; J Nicholas Odom[2,8,9], PhD, RN

[1]Frances Payne Bolton School of Nursing, Case Western Reserve University, Cleveland, OH, United States

[2]School of Nursing, University of Alabama at Birmingham, 1720 2nd Avenue South, Birmingham, AL, United States

[3]Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, United States

[4]Department of Medical Oncology, Harvard Medical School, Dana-Farber Cancer Institute, Boston, MA, United States

[5]Department of Medicine, Duke University, Durham, NC, United States

[6]Mongan Institute Center for Aging and Serious Illness, Division of Palliative Care and Geriatric Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, MA, United States

[7]School of Nursing & Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States

[8]Division of Geriatrics, Gerontology, and Palliative Care, Department of Medicine, University of Alabama at Birmingham, Birmingham, AL, United States

[9]Center for Palliative and Supportive Care, University of Alabama at Birmingham, Birmingham, AL, United States

Corresponding Author:
J Nicholas Odom, PhD, RN
School of Nursing, University of Alabama at Birmingham, 1720 2nd Avenue South, Birmingham, AL, United States

Abstract

Background: Patients with advanced cancer and their family caregivers often experience poor quality of life (QOL). Self-report measures are commonly used to quantify QOL of family caregivers but may have limitations such as recall bias and social desirability bias. Variables derived from passively obtained smartphone GPS data are a novel approach to measuring QOL that may overcome these limitations and enable detection of early signs of mental and physical health (PH) deterioration.

Objective: This study explored the feasibility of a digital phenotyping approach by assessing participant adherence and examining correlations between smartphone GPS data and QOL levels among family caregivers and patients with advanced cancer.

Methods: This was a secondary analysis involving 7 family caregivers and 4 patients with advanced cancer that assessed correlations between GPS sensor data captured by a personally owned smartphone and QOL self-report measures over 12 weeks through linear correlation coefficients. QOL as measured by the Patient-Reported Outcomes Measurement Information System (PROMIS) Global Health 10 was collected at baseline, 6, and 12 weeks. Using a Beiwe smartphone app, GPS data were collected and processed into variables including total distance, time spent at home, transition time, and number of significant locations.

Results: The study identified relevant temporal correlations between QOL and smartphone GPS data across specific time periods. For instance, in terms of PH, associations were observed with the total distance traveled (12 and 13 wk, with r ranging 0.37 to 0.38), time spent at home (−4 to −2 wk, with r ranging from −0.41 to −0.49), and transition time (−4 to −2 wk, with r ranging −0.38 to −0.47).

Conclusions: This research offers insights into using passively obtained smartphone GPS data as a novel approach for assessing and monitoring QOL among family caregivers and patients with advanced cancer, presenting potential advantages over traditional self-report measures. The observed correlations underscore the potential of this method to detect early signs of deteriorating mental health and PH, providing opportunities for timely intervention and support.

(JMIR Form Res 2025;9:e59161) doi:10.2196/59161

KEYWORDS

cancer; digital phenotyping; global positioning system; quality of life; smartphone; mobile phone; family caregiver

XSL•FO
RenderX

Introduction

Each year, an estimated 1.96 million individuals are expected to receive a cancer diagnosis in 2023 [1]. Projections indicate a near 25% increase in cancer diagnoses by 2032 when it is estimated that 22.5 million US individuals will have a history of cancer [2]. With cancer care transitioning increasingly to home settings, the impact of cancer will increasingly stretch beyond the individual diagnosed to those family and friend caregivers who provide support [3]. Consequently, family caregivers have become indispensable frontline clinicians for patients, especially when their cancer progresses to advanced stages [4]. Both patients and their family caregivers face an array of challenges, from logistical and practical challenges to managing treatments and symptoms to handling a range of strong emotions, including anxiety and depression [5]. Hence, it is not surprising that patients and their caregivers often experience diminished quality of life (QOL) [6-8], which cancer care clinicians aim to assess and optimize.

Traditionally, assessing QOL has primarily relied on patient-reported outcome measures (PROMs) using paper-and-pencil questionnaires, online surveys, or verbal self-report to a data collector reading from a list of items [9,10]. However, PROMs have several notable limitations [11,12]. They can be time consuming for both patients and clinicians and interrupt daily activities. Furthermore, they typically entail retrospective recall which can be subject to distorting biases [13,14] and may elicit socially desirable responses, such as over-reporting better health or greater well-being than actually experienced [15-17].

To address the limitations of a PROM approach to assessing patient and caregiver QOL states, digital phenotyping has emerged as a new and potentially innovative approach. Defined as the "moment-by-moment quantification of the individual-level human phenotype in situ using data from personal digital devices" [18], this method involves passive data collection from ubiquitously owned devices such as smartphones, now owned by approximately 85% of US individuals [19]. By using smartphone sensor data, digital phenotyping models changes in individuals' behaviors, including mobility (eg, GPS data), sociability (eg, text message and telephone logs), and sleep patterns (eg, accelerometer data and screen activity logs) [10].

Physical activity has been shown to improve both the physical and mental well-being of patients with cancer and their caregivers [20,21]. Recent studies have demonstrated that engaging in regular physical activity reduces the risk of cancer recurrence and alleviates symptoms of anxiety and depression, which in turn positively influence QOL [22,23]. These benefits are particularly important for caregivers, who often experience high levels of stress, as physical activity can improve their mental health (MH) and overall well-being [24].

GPS data has been shown to correlate with measures assessing mood, anxiety, and depression [25]. Better mood, lower anxiety, and reduced depression were associated with spending more time at locations of friends and family members. Conversely, more time spent at health care facilities was linked to poorer mood, higher anxiety, and increased depression [25]. Furthermore, GPS data were examined in comparing recovery metrics between mastectomy and breast-conserving surgery in patients with breast cancer [26]. The results indicated that patients undergoing mastectomy spent more time at home in the initial weeks postsurgery, whereas no significant differences were observed after 12 weeks. Despite a growing body of research that highlights the relationship between individual GPS-derived mobility behaviors and MH and well-being [27,28], a digital phenotyping approach to QOL assessment has yet to be explored in an advanced cancer context.

This study aimed to explore the temporal relationship between smartphone GPS sensor features and participants' physical health (PH) and MH over a 12-week period. We hypothesized that more consistent and stable GPS-derived mobility features would correlate with better patient and caregiver PROM reports of QOL.

Methods

Study Design

This was a secondary analysis of data collected from patient and caregiver participants of a larger, ongoing prospective longitudinal observational study [29]. This analysis used data collected over 12 weeks from individuals enrolled between September 2021 and August 2022 at an academic medical center in the southeastern United States as part of a larger research project investigating patterns in passively collected smartphone behavioral data, including GPS and accelerometer.

Ethical Considerations

This study is a secondary analysis of deidentified data collected from a prior study, which was approved by the University of Alabama at Birmingham Institutional Review Board (IRB; 300008890). The IRB determined that this secondary analysis does not constitute Human Subjects Research and is exempt from further IRB review. Details regarding the informed consent process and participant compensation in the original study are described in the Procedures section. For this secondary analysis, no direct participant interaction occurred, and the dataset was fully deidentified before access. Therefore, no additional informed consent or participant compensation was required.

Sample

We included 7 family caregivers and 4 patients with advanced cancer who were recruited as of August 4, 2022. Inclusion criteria for both caregivers and patients were (1) aged 18 years or older, (2) owned a personal smartphone (Android or iOS) and had the ability to download and run the study app, and (3) had proficiency in the English language (spoken and written). In addition, enrolled caregivers provided regular support to a patient with advanced cancer, while patients had to be diagnosed with metastatic, recurrent, or progressive stage III or IV cancer.

Procedures

Once potential participants were identified, they received an invitation letter containing the study description and a timeline of study activities through mail. Subsequently, participants were contacted by telephone to confirm their interest. Verbal consent

XSL•FO
RenderX

was obtained during these calls, and participants were provided with an informed consent document, baseline questionnaires, and a participant guide. After completing the baseline questionnaire, research coordinators assisted participants in downloading and installing the Beiwe app on their personally owned smartphone, which passively collected GPS data. Patient-reported questionnaires were completed through various modalities, including electronic completion through REDCap (Research Electronic Data Capture; Vanderbilt University), mailed paper-and-pencil surveys, or phone-based data collection. Both caregiver and patient participants received US $25 for phone usage and an additional US $25 for each completed questionnaire every 6 weeks.

Beiwe Research Platform

For this research, we used the Beiwe Research Platform, which has a primary focus on developing statistical and computational tools to extract biomedical and clinical insights from smartphone data [30]. The Beiwe app, developed by Torous and colleagues [30], passively collects smartphone GPS data and securely synchronizes it over Wi-Fi to a cloud-based server managed by Beiwe on Amazon Web Services. The platform includes a study portal, applications for both Android and iOS, secure data storage on Amazon Web Services, and data analysis tools and code [30]. To prioritize privacy, passive data undergo encryption both during transit and storage, and indirect identifiers, such as telephone numbers and IP addresses, are hashed, a process that converts the original data into a fixed-size string of characters to enhance the security and privacy of the data [30]. Data linkage is exclusively tied to an 8-character Beiwe participant ID to ensure confidentiality [30].

Data Collection

Demographics

Collected demographic information included age, gender, race or ethnicity, religion, marital status, household size, employment and education levels, relationship status, caregiving duration, and caregiving frequency.

QOL Data

Health-related QOL was assessed using the Patient-Reported Outcomes Measurement Information System (PROMIS) Global-10 questionnaire at baseline, 6 weeks, and 12 weeks. This questionnaire assesses respondents on their perception of overall health, pain, fatigue, social health, MH, and PH [9,31]. Two subscale scores are generated for PH (α=.86) and MH (α=.81), respectively [32]. Lower scores (MH <40, PH <42) indicated poor MH and PH [33], and are associated with higher risk of future health care use [34,35].

GPS Data

We established a data-quality dataset indicating the “valid” GPS hours for each participant per day, enabling the identification and exclusion of individuals with insufficient data during analyses [36]. Our threshold for valid GPS data was set at 15 hours per day, considering that participants might sleep for 9 hours during which GPS records may not be transmitted [37]. Missing location data resulting from issues such as phone battery depletion or power-off states were excluded cases. We determined the number of days with a minimum of 15 hours of recorded GPS data for each user in our dataset. Subsequently, days meeting this hourly cutoff were extrapolated to 24 hours. Additionally, for week-level analyses, participants with fewer than 3 valid days per week were excluded from our analysis [36].

Data Analysis and Feature Extraction From GPS Data

Overview of Feature Extraction

For data analysis, we used Python (version 3.9.0; Python Software Foundation) in Jupyter Notebooks and Google Collaboratory to preprocess the GPS data and extract mobility variables, including total distance, time spent at home, transition time, and number of significant locations.

Total Distance

This variable measured the total distance covered by each participant using the Haversine distance formula, which calculates the angular distance between 2 points on the Earth’s surface [38].

Time Spent at Home

The home cluster for each participant was identified based on two heuristics: (1) it was among the top 3 most visited clusters, and (2) it was the cluster most visited between 12 AM and 6 AM [39]. The area of the participant’s home was assumed to be in a circular area with a 2000 square feet area from the estimated latitude and longitude of the house location. The time spent within this 2000 square feet area was calculated for each day.

Transition Time

Location data points were categorized as either stationary (eg, working in an office) or transitional (eg, walking on the street) based on movement speed [39]. The time derivative of each data point was used to calculate the movement speed, and a threshold speed of 1 meter per second (m/s) was applied to distinguish between transitional (speed>1 m/s) and stationary states (speed<1 m/s) [39]. The time spent at speeds greater than 1 m/s was calculated for each day.

Number of Significant Locations

This feature represents the total number of clusters found using a machine learning clustering algorithm [40]. The algorithm, called adaptive k-means, classified data points into specific groups based on their locations during stationary states [41]. The goal was to identify places where participants spent most of their time, such as homes, workplaces, and parks. The algorithm used the elbow method to determine the optimal number of clusters, which helps minimize the overall distances of data points to the centers of their respective clusters [42].The number of significant locations visited was then calculated for each day.

Overview of Data Analysis

Statistical Methods Used

Data analysis was conducted using the R programming language (version 4.2.1, R Foundation), in RStudio (RStudio, PBC). Descriptive statistics were computed for variables of interest,

XSL•FO
RenderX

and patterns of missing data were explored. PROMIS Global-10 mental and PH T-scores were used as measures of QOL. The association between GPS location features and patient-reported QOL was assessed using linear correlation coefficients. Temporal association between QOL assessments and GPS data were calculated weekly, using repeated-measures correlations with 1000 cluster-bootstrap resamples [43] to estimate 95% CIs. Cluster bootstrapping [44] was applied to handle multiple observations per participant.

Temporal Correlations

As the QOL measurements did not necessarily align with weeks 0, 6, and 12 of smartphone data, we calculated the time difference between each patient-reported QOL assessment and daily smartphone GPS data on a weekly basis. Relative to the week of QOL assessment collection, time differences between smartphone data and QOL assessments ranged from −15 weeks to +14 weeks. The number of observations available for estimating a correlation at each temporal difference was computed. We considered data for temporal differences with more than 10 available observations per week, resulting in analysis for −6 to +13 weeks, excluding +10 weeks (Figure 1).

XSL•FO
RenderX

Figure 1. Time differences between patient-reported QOL assessments and daily smartphone GPS data were calculated on a weekly basis, as QOL measurements did not align with the smartphone data collection weeks of 0, 6, and 12. (A) Weekly time differences between QOL assessment and smartphone GPS data. (B) Example of weekly time differences in QOL and GPS data. (C) Data from temporal differences: –6 to +13 weeks (excluding +10 weeks) with >10 observations per week. QOL: quality of life.

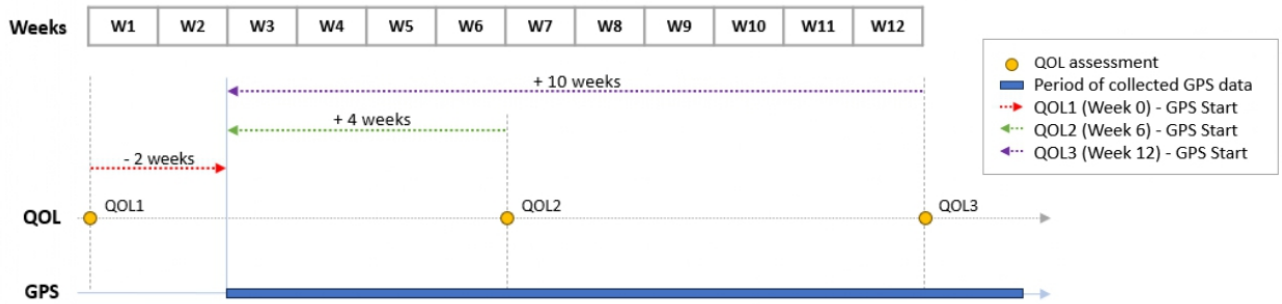

(A) Weekly time differences between QOL assessment and smartphone GPS data.

		QOL Assessment and Week of it		
		QOL1	QOL2	QOL3
		-2 weeks	+4 weeks	+10 weeks
Participant	GPS data week	Weeks from after QOL assessment		
Participant 1	1	3	-3	-9
Participant 1	2	4	-2	-8
Participant 1	3	5	-1	-7
Participant 1	4	6	0	-6
Participant 1	5	7	1	-5
Participant 1	6	8	2	-4
Participant 1	7	9	3	-3
Participant 1	8	10	4	-2
Participant 1	9	11	5	-1
Participant 1	10	12	6	0
Participant 1	11	13	7	1
Participant 1	12	14	8	2

(B) Example of weekly time differences in QOL and GPS data.

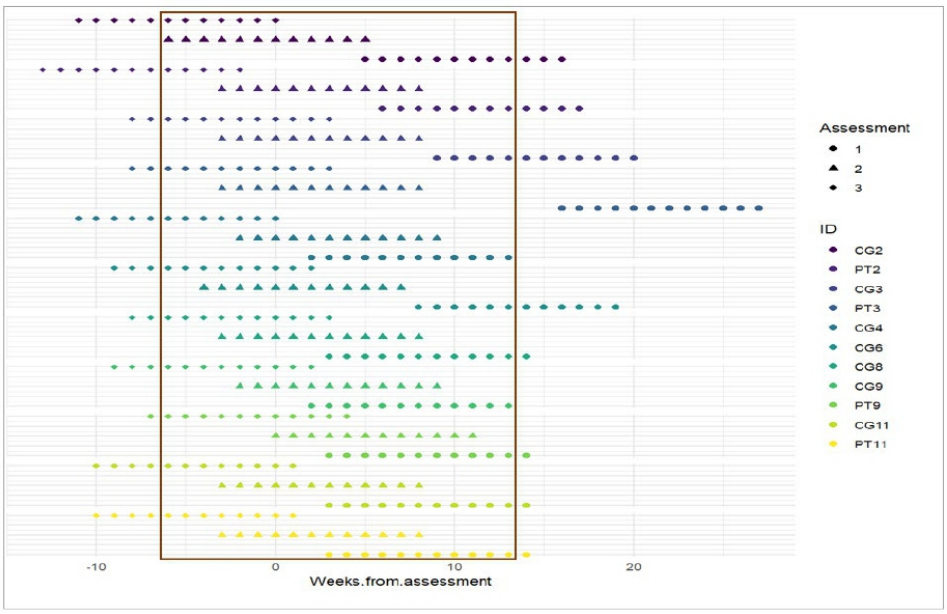

(C) Data from temporal differences: -6 to +13 weeks (excluding +10 weeks) with >10 observations per week.

Results

Demographics of Participants

The study included 11 participants (7 caregivers and 4 patients). Caregivers had a mean age of 47.7 (SD 15.9) years, while patients had a mean age of 55 (SD 3.7) years. Most participants identified as White (10/11, 90%) and non-Hispanic or non-Latino (11/11, 100%). Most participants (9/11, 82%) identified as Protestant, were married (10/11, 90%), employed (5/11, 45%), and had a high school diploma or General

XSL•FO
RenderX

Educational Development or higher (10/11, 90%). Among caregivers, 71% (5/7) cared for a spouse, and their average caregiving duration was 557 (SD 96) days. Most caregivers (6/7, 86%) lived with the patients. Regarding insurance, 75% (3/4) of patients had private or commercial medical insurance, and 25% (1/4) had military insurance. Furthermore, 75% (3/4) of patients resided in rural areas (Table 1).

XSL•FO
RenderX

Table . Demographics of participants.

Variables	Caregivers (n=7)	Patients (n=4)	All (N=11)
Age (years), mean (SD)	47.71 (15.86)	55 (3.65)	50 (13)
Gender, n (%)			
Male	3 (43)	2 (50)	5 (46)
Female	4 (57)	2 (50)	6 (55)
Race, n (%)			
Black or African American	1 (14)	0 (0)	1 (9)
White	6 (86)	4 (100)	10 (90)
Hispanic or Latino, n (%)			
Yes	0 (0)	0 (0)	0 (0)
No	7 (100)	4 (100)	11 (100)
Religion, n (%)			
Protestant	6 (86)	3 (75)	9 (82)
Catholic	0 (0)	1 (25)	1 (9)
None	1 (14)	0 (0)	1 (9)
Marital status, n (%)			
Never married	1 (14)	0 (0)	1 (9)
Married	6 (86)	4 (100)	10 (90)
Employment status, n (%)			
Full time	4 (57)	0 (0)	4 (36)
Part time	0 (0)	1 (25)	1 (9)
Unemployed due to disability or illness	1 (14)	1 (25)	2 (18)
Homemaker	1 (14)	1 (25)	2 (18)
Other	1 (14)	1 (25)	2 (18)
Education, n (%)			
≤8th grade	1 (14)	0 (0)	1 (9)
Some high school	0 (0)	0 (0)	0 (0)
High school graduate or General Educational Development	2 (29)	2 (50)	4 (36)
Some college or technical school	2 (29)	0 (0)	2 (18)
College graduate (bachelor's degree)	2 (29)	2 (50)	4 (36)
Relationship with patient, n (%)			
Spouse	5 (71)	—[a]	—
Parent	1 (14)	—	—
Other	1 (14)	—	—
Lives with patient, n (%)			
Yes	6 (86)	—	—
No	1 (14)	—	—
Length of caregiving (day), mean (SD)	556.57 (96)	—	—
The number of days of care per week, n (%)			

XSL•FO
RenderX

Variables	Caregivers (n=7)	Patients (n=4)	All (N=11)
Every day of the week (7 days)	3 (43)	—	—
2 to 3 days a week	2 (29)	—	—
1 day a week or less	2 (29)	—	—
Hours of care per day, n (%)			
5 to 6 hours a day	1 (14)	—	—
3 to 4 hours a day	1 (14)	—	—
1 to 2 hours a day	5 (71)	—	—
Medical insurance, n (%)			
Private or commercial	—	3 (75)	—
Military program	—	1 (25)	—
Residence, n (%)			
Urban	—	0 (0)	—
Suburban	—	1 (25)	—
Rural	—	3 (75)	—

[a]Not applicable.

QOL Assessment Results

PH Findings

At baseline, the mean PH score across all participants was 47.1 (SD 6.7). Caregivers had a slightly lower mean PH score (mean 45.4, SD 4.6) compared to patients (mean 50.1, SD 9.4). Around 18% (2/11) of participants reported baseline PH T-scores below 42, including 1 caregiver and 1 patient. By the 6-week mark, the average PH score had increased to 49 (SD 9.6), with caregivers reporting a mean score of 50.2 (SD 10.3) and patients a mean score of 47.1 (SD 9.2). Similarly, around 18% (2/11) of participants recorded 6-week PH T-scores below 42, including 1 caregiver and 1 patient. At the 12-week assessment, the overall mean PH score slightly decreased to 46.6 (SD 6.8), with caregivers and patients reporting mean scores of 46.8 (SD 6.1) and 46.3 (SD 9.1), respectively. This time, around 27% (3/11) of participants, including 1 caregiver and 2 patients, had 12-week PH T-scores below 42 (see Multimedia Appendix 1 for participant-reported QOL results).

MH Findings

At baseline, the mean MH score across all participants was 50.7 (SD 8.6). Caregivers had a slightly lower mean MH score (mean 47.9, SD 7.7) compared to patients (mean 55.5, SD 8.8). One caregiver (9%) reported a baseline MH T-score below 40. At the 6-week evaluation, the mean MH score across all participants slightly increased to 51.5 (SD 7.7), with caregivers and patients reporting mean scores of 51.1 (SD 9) and 52.2 (SD 6), respectively. By the 12-week mark, the overall mean MH score further improved to 52.8 (SD 7.2), with caregivers at a mean score of 52.2 (SD 7.3) and patients at 53.9 (SD 8.1). One caregiver (9%) reported a 12-week MH T-score below 40 (Multimedia Appendix 1).

Daily Smartphone GPS Data Weekly Averages

Over the 12-week period, the total distance variable ranged from a minimum of 102.2 (SD 66.5) km to a maximum of 186.7 (SD 153.2) km. The time spent at home variable ranged from a minimum of 437.2 (SD 388.5) minutes to a maximum of 573.3 (SD 271.1) minutes. The transition time variable ranged from a minimum of 800.4 (SD 225.3) minutes to a maximum of 903.4 (SD 109.4) minutes. The number of significant locations variable ranged from a minimum of 4.2 (SD 1.4) to a maximum of 5.1 (SD 1.6; see Multimedia Appendix 2 for details on daily smartphone GPS location features averaged by week).

Correlation Between PH and Daily Smartphone GPS Data Over a 12-Week Period

Table 2 illustrates the temporal correlation between PH and weekly GPS location features. Significant correlations, with a value of +0.3 or higher, were observed between total distance and PH at 3 weeks before and during weeks 12 and 13 after the QOL assessment. Positive correlations at 3 weeks before and during weeks 12 and 13 after the QOL assessment imply that increased distances correlate with higher PH. Correlations of −0.3 or higher were found between time spent at home and PH at 2, 3, and 4 weeks before the QOL assessment, as well as at the 3 weeks after the assessment. These negative correlations suggest a potential association between increased time spent at home and lower PH. Similarly, correlations of −0.3 or higher were found between transition time and PH at 2, 3, and 4 weeks before the QOL assessment. These negative correlations also suggest a potential association between increased transition time and lower PH. Moreover, correlations of ±0.3 or higher between the number of significant locations and PH were significant at 3 weeks before and at weeks 1, 3, and 4 after the QOL assessment. Negative correlations at week 3 after the QOL assessment suggest that as the number of places visited increases, PH tends to decline, while positive correlations at

XSL•FO
RenderX

week 3 before and at weeks 1 and 4 after the QOL assessment indicate a potential link to higher PH (Figure 2).

Table . Correlation between physical health and daily smartphone GPS features, including total distance, time spent at home, transition time, and number of significant locations, averaged by week and using 95% cluster-bootstrap CIs over a 12-week period.

From QOL[a], weeks[b]	Observations, n	Participants, n	Total distance mean, *r*[c] (95% CI)	Time spent at home mean, *r* (95% CI)	Transition time mean, *r* (95% CI)	Number of significant locations mean, *r* (95% CI)
−6	11	10	0.06 (−0.42 to 0.74)	−0.14 (−0.52 to 0.41)	0.14 (−0.31 to 0.63)	0.11 (−0.34 to 0.58)
−5	11	10	0.14 (−0.47 to 0.54)	0.02 (−0.26 to 0.38)	−0.05 (−0.56 to 0.35)	0.28 (−0.14 to 0.76)
−4	12	10	0.12 (−0.37 to 0.61)	−0.49 (−0.71 to −0.05)	−0.44 (−0.84 to 0.06)	−0.18 (−0.48 to 0.35)
−3	16	11	0.39 (0.01 to 0.65)	−0.41 (−0.69 to −0.02)	−0.38 (−0.74 to 0.04)	0.34 (−0.06 to 0.63)
−2	16	10	0.12 (−0.44 to 0.63)	−0.46 (−0.80 to 0.02)	−0.47 (−0.80 to 0.01)	0.21 (−0.23 to 0.66)
−1	19	11	0.17 (−0.40 to 0.76)	0.08 (−0.26 to 0.61)	0.20 (−0.15 to 0.57)	0.14 (−0.20 to 0.57)
0	19	11	0.08 (−0.54 to 0.53)	−0.10 (−0.47 to 0.24)	−0.06 (−0.41 to 0.18)	−0.02 (−0.28 to 0.29)
1	18	11	0.05 (−0.44 to 0.62)	−0.19 (−0.45 to 0.10)	−0.16 (−0.45 to 0.13)	0.39 (0.11 to 0.66)
2	19	11	0.26 (−0.54 to 0.80)	−0.12 (−0.47 to 0.33)	−0.23 (−0.46 to −0.03)	0.07 (−0.26 to 0.47)
3	20	11	0.20 (−0.21 to 0.60)	−0.41 (−0.65 to −0.20)	−0.16 (−0.66 to 0.31)	−0.31 (−0.55 to 0.43)
4	16	11	0.21 (−0.16 to 0.71)	−0.18 (−0.50 to 0.17)	0.14 (−0.31 to 0.63)	0.46 (0.13 to 0.65)
5	17	11	−0.09 (−0.66 to 0.36)	0.08 (−0.22 to 0.49)	−0.14 (−0.59 to 0.30)	0.17 (−0.41 to 0.63)
6	18	11	0.28 (−0.51 to 0.80)	0.08 (−0.27 to 0.51)	0.15 (−0.35 to 0.55)	0.09 (−0.14 to 0.60)
7	16	11	0.08 (−0.36 to 0.66)	−0.17 (−0.57 to 0.34)	−0.04 (−0.49 to 0.51)	0.19 (−0.25 to 0.57)
8	17	11	0.10 (−0.61 to 0.50)	−0.20 (−0.64 to 0.19)	−0.28 (−0.68 to 0.04)	0.23 (−0.23 to 0.54)
9	12	9	0.25 (−0.40 to 0.63)	−0.17 (−0.64 to 0.36)	0.12 (−0.37 to 0.51)	−0.20 (−0.49 to 0.29)
11	10	10	−0.01 (−0.88 to 0.52)	0.03 (−0.39 to 0.51)	−0.12 (−0.68 to 0.29)	0.18 (−0.17 to 0.73)
12	10	10	0.37 (−0.84 to 0.74)	0.21 (−0.18 to 0.65)	0.06 (−0.29 to 0.53)	0.00 (−0.44 to 0.55)
13	10	10	0.38 (−0.57 to 0.77)	0.01 (−0.33 to 0.39)	−0.01 (−0.51 to 0.57)	0.39 (−0.54 to 0.74)

[a]QOL: quality of life.

[b]Weeks from QOL: temporal correlations.

[c]Correlation (absolute value): small, r=0.1; medium, r=0.3; and large, r≥0.5.

XSL•FO
RenderX

Figure 2. Correlation between physical and mental health and daily smartphone GPS features, including total distance, time spent at home, transition time, and number of significant locations averaged by week and using 95% cluster-bootstrap CIs over a 12-week period. (A) Physical health. (B) Mental health. QOL: quality of life.

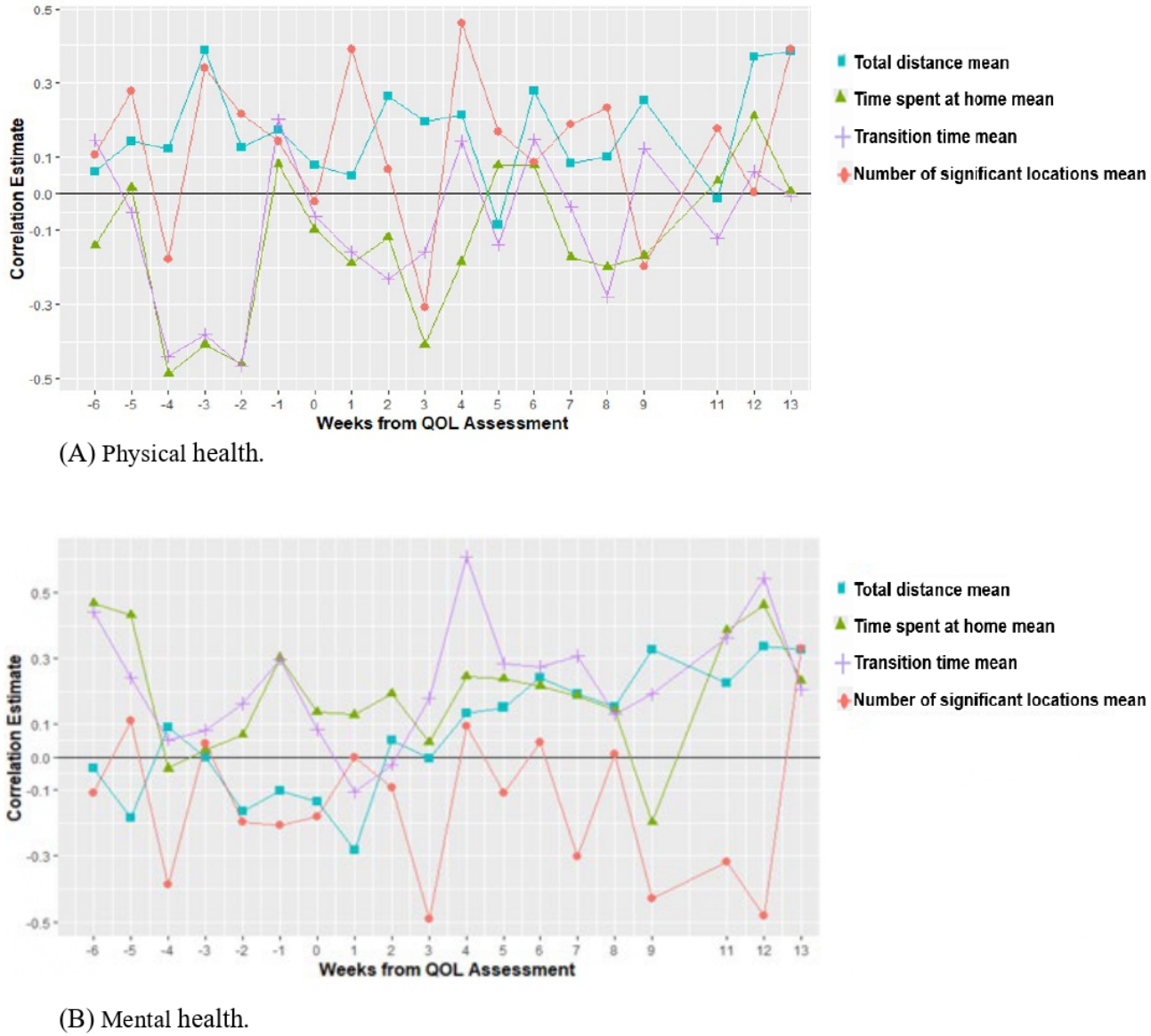

Correlation Between MH and Daily Smartphone GPS Data Over a 12-Week Period

Table 3 illustrates the temporal correlation between MH and weekly GPS location features. Correlations of −0.3 or higher were notable between time spent at home and MH at weeks 4 and 2 before the QOL assessment, as well as at week 9 after the QOL assessment, suggesting a potential link between increased time spent at home and lower MH. Correlations of ±0.3 or higher were significant between transition time and MH at weeks −4 and −2 before the QOL assessment, as well as at week 4 after the QOL assessment. Negative correlations at weeks −4 and −2 imply that increased transition time correlates with lower MH, while positive correlations at week 4 suggest a potential link between increased transition time and higher MH. Positive correlations of +0.3 or higher were significant between the number of significant locations and MH at 2 weeks before the QOL assessment and at week 13 after the QOL assessment, suggesting a potential link between an increased number of places visited and higher MH (Figure 2).

XSL•FO
RenderX

Table . Correlation between mental health and daily smartphone GPS features, including total distance, time spent at home, transition time, and number of significant locations, averaged by week and using 95% cluster-bootstrap CIs over a 12-week period.

From QOL[a], weeks[b]	Observations, n	Participants, n	Total distance mean, *r*[c] (95% CI)	Time spent at home mean, *r* (95% CI)	Transition time mean, *r* (95% CI)	Number of significant locations mean, *r* (95% CI)
−6	11	10	−0.03 (−0.50 to 0.38)	0.47 (0.07 to 0.77)	0.44 (0.07 to 0.87)	−0.11 (−0.35 to 0.29)
−5	11	10	−0.18 (−0.67 to 0.1)	0.43 (0.13 to 0.78)	0.24 (−0.01 to 0.8)	0.11 (−0.21 to 0.6)
−4	12	10	0.09 (−0.39 to 0.34)	−0.03 (−0.35 to 0.42)	0.05 (−0.48 to 0.59)	−0.39 (−0.75 to 0.02)
−3	16	11	0 (−0.33 to 0.41)	0.02 (−0.23 to 0.39)	0.08 (−0.31 to 0.51)	0.04 (−0.38 to 0.49)
−2	16	10	−0.16 (−0.51 to 0.27)	0.07 (−0.3 to 0.28)	0.16 (−0.15 to 0.37)	−0.2 (−0.56 to 0.16)
−1	19	11	−0.1 (−0.49 to 0.28)	0.3 (−0.08 to 0.62)	0.3 (−0.05 to 0.66)	−0.21 (−0.41 to 0.02)
0	19	11	−0.13 (−0.45 to 0.16)	0.14 (−0.15 to 0.38)	0.08 (−0.32 to 0.45)	−0.18 (−0.55 to 0.18)
1	18	11	−0.28 (−0.6 to 0.04)	0.13 (−0.25 to 0.47)	−0.11 (−0.45 to 0.22)	0 (−0.42 to 0.42)
2	19	11	0.05 (−0.24 to 0.35)	0.19 (−0.13 to 0.48)	−0.02 (−0.47 to 0.44)	−0.09 (−0.46 to 0.35)
3	20	11	0 (−0.4 to 0.41)	0.05 (−0.29 to 0.4)	0.18 (−0.3 to 0.62)	−0.49 (−0.65 to -0.09)
4	16	11	0.13 (−0.28 to 0.54)	0.24 (−0.09 to 0.51)	0.61 (0.26 to 0.84)	0.09 (−0.57 to 0.36)
5	17	11	0.15 (−0.3 to 0.65)	0.24 (−0.04 to 0.45)	0.28 (−0.09 to 0.52)	−0.11 (−0.62 to 0.25)
6	18	11	0.24 (−0.23 to 0.67)	0.22 (0.01 to 0.55)	0.27 (−0.01 to 0.53)	0.05 (−0.36 to 0.57)
7	16	11	0.19 (−0.29 to 0.68)	0.19 (−0.21 to 0.47)	0.31 (0.08 to 0.61)	−0.3 (−0.65 to 0.05)
8	17	11	0.15 (−0.51 to 0.62)	0.14 (−0.38 to 0.5)	0.13 (−0.41 to 0.44)	0.01 (−0.85 to 0.39)
9	12	9	0.33 (0 to 0.83)	−0.2 (−0.71 to 0.58)	0.19 (−0.36 to 0.69)	−0.43 (−0.77 to 0.07)
11	10	10	0.23 (−0.52 to 0.85)	0.39 (−0.23 to 0.82)	0.36 (−0.43 to 0.76)	−0.32 (−0.8 to 0.59)
12	10	10	0.34 (−0.3 to 0.77)	0.46 (−0.05 to 0.86)	0.54 (0.14 to 0.87)	−0.48 (−0.87 to 0.03)
13	10	10	0.33 (−0.15 to 0.94)	0.23 (−0.31 to 0.71)	0.2 (−0.31 to 0.7)	0.33 (−0.9 to 0.77)

[a]QOL: quality of life.

[b]Weeks from QOL: temporal correlations.

[c]Correlation (absolute value): small, r=0.1; medium, r=0.3; and large, r≥0.5.

Discussion

Principal Findings

The findings of this exploratory, secondary analysis demonstrate the potential association between smartphone-derived GPS-based movement features and QOL of patients with advanced cancer and their family caregivers. While not powered for statistical significance, correlations were observed between GPS movement variables and PH and MH during stretches of time before and after PROM assessment. These results warrant continued investigation of the potential of digital phenotyping approaches to passively assess patient and caregiver QOL in the context of cancer. Future analyses could investigate the interrelationship between the movement patterns and QOL of both patients and their caregivers, offering deeper insights into the dynamics between their experiences.

This study extends the existing body of research associating higher movement with higher QOL for patients with cancer and their family caregivers [45]. Notably, significant correlations were observed between PH and the total distance covered in GPS data collected 3 weeks before the QOL assessment. This indicates a potential connection between higher PH and longer distances traveled during this period. Additionally, our study findings align with previous research, indicating a correlation

XSL•FO
RenderX

between PH and time spent at home [45,46]. Our findings suggest that poor PH is associated with increased time spent at home at 2, 3, and 4 weeks before the QOL assessment. This finding is supported by earlier research, which has proposed that an extended duration at home might be indicative of fatigue or impaired motor function [45]. However, our study findings diverge from prior research on the correlations between PH and the transition time [47], or the number of significant locations [48]. While our results suggest that decreased mobility is associated with higher PH, earlier research has proposed that increased mobility is an indicator of recovered PH. These findings imply that the metrics of transition time and number of visited places may not adequately capture how these indices relate to PH. Given the different contexts of the study design, it is not surprising that our results were inconsistent with the findings of a previous study. Future research may, therefore, explore the context of the relationship between PH and GPS measures in further detail, such as clinical populations or specific environmental contexts.

Our study did not reveal significant correlations between MH and the time spent at home or the number of significant locations overall. However, specific weeks within the 12-week period showed noteworthy correlations. For example, a statistically significant correlation emerged at 1, 5, and 6 weeks when GPS data were collected before the QOL assessment, indicating an association between increased time spent at home and MH. Similarly, at 3, 7, 9, 10, 11, 12, and 13 weeks, when GPS data were collected after the QOL assessment, a significant correlation emerged, suggesting a link between the number of significant locations and MH. These findings indicate temporal variability in the relationship between individuals' daily activities, such as time spent at home and various locations, and their MH across distinct phases. While the precise significance of these periodic correlations remains uncertain, it underscores the necessity of accounting for temporal variations in understanding the connections between environmental exposure and MH outcomes. Further investigation is warranted to explore potential contributing factors to these temporal variations and to elucidate the nuanced dynamics of the interplay between daily routines and MH across different time frames.

Moreover, in our exploration of the relationship between MH and movement, we found that the total distance exhibited either no correlation or only a small correlation over time. Existing literature presents mixed findings; however, a significant portion indicates that traveling shorter distances is associated with worse outcomes in terms of emotional and cognitive well-being [49-51]. When exploring the association between MH and transition time, our results displayed mixed outcomes, aligning with previous studies where results on the correlation between MH and movement conflicted [39,52]. Notably, 1 study indicated that individuals with irregular movement patterns—described by irregularity features such as variations in circadian movement and routine index across days—tended to report fewer depressive symptoms and less loneliness compared to those with more regular patterns, hinting at a potential link between irregular movement and improved overall MH [53]. While our findings suggest that distance and movement behaviors may not be directly linked to specific MH aspects, they contribute to the broader understanding of general movement patterns, such as total distance and transition time.

Limitations

This study has several limitations. First, the small sample size introduces uncertainty in correlation estimates and limits the generalizability of the findings. Future research with a larger sample size is needed to validate and extend these findings. Second, the participant sample was recruited from a specific group of patients with advanced cancer and family caregivers attending a particular oncology clinic, raising the potential for selection bias. Future research endeavors should strive for a more geographically and racially diverse sample. Third, the pilot study encountered issues with missing sensor data, leading to the exclusion of weeks with less than 3 days of data and days with less than 10 hours of data per day from the analysis. This limitation narrows the generalizability of our findings for those periods. The exact cause of the missing data is unclear; it could be attributed to several factors, such as technical issues with the app or the participants' devices not being turned on, potentially due to reasons such as the device not being charged. Fourth, GPS data alone does not provide specific information about the type of activity performed outside the home, which limits the interpretation of the data and its correlation with quality-of-life questionnaires. Although we extracted several features from the GPS data to offer insights into participant behavior (eg, total distance traveled and time spent at home), the inability to classify activity types outside the home remains a limitation in the analysis. This limitation could impact the accuracy of the behavioral patterns examined in relation to QOL. Despite these acknowledged limitations, the study suggests the potential integration of digital phenotyping into health research, providing valuable insights into the challenges faced by individuals with complex health conditions and their caregivers in their daily lives.

Conclusion

This study investigated the correlation between the QOL of patients with advanced cancer and family caregivers and smartphone GPS data. The results revealed complex relationships between smartphone GPS features and participants' PH and MH outcomes. Specifically, total distance, time spent at home, transition time, and the number of significant locations showed various correlations with reported QOL. The study contributes to the field of digital phenotyping by shedding light on the intricate dynamic relationships between daily behaviors captured through smartphone GPS data and well-being among family caregivers and patients with advanced cancer. The findings emphasize the importance of considering personalized patterns and contextual factors when interpreting the impact of smartphone sensor data on health outcomes.

XSL•FO
RenderX

Acknowledgments

This study was funded by the Cambia Health Foundation (principal investigator: JNO; no grant number provided). We express our gratitude to the partnering oncologists, clinicians, and staff at the O'Neal Comprehensive Cancer Center at the University of Alabama at Birmingham for their support throughout this study. We also thank James Mapson, Brianna Brakefield, Stephanie Ford, and Charis Smith for their invaluable contributions as research coordinators.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

KL and JNO contributed to this study's concept and design. KL, AA, and SK performed the data analyses. All authors contributed to the interpretation of the data. KL and JNO drafted this paper, and all authors critically revised it for relevant intellectual content. All authors have read and approved the final version of this paper and take responsibility for their individual contributions.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Participant-reported quality of life.
[DOCX File, 19 KB - formative_v9i1e59161_app1.docx]

Multimedia Appendix 2
Daily smartphone GPS location features averaged by week.
[DOCX File, 19 KB - formative_v9i1e59161_app2.docx]

References

1. Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. CA Cancer J Clin 2023 Jan;73(1):17-48. [doi: 10.3322/caac.21763] [Medline: 36633525]
2. Cancer treatment & survivorship facts & figures 2022-2024. American Cancer Society. 2022. URL: https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/cancer-treatment-and-survivorship-facts-and-figures/2022-cancer-treatment-and-survivorship-fandf-acs.pdf [accessed 2025-01-28]
3. National Academies of Sciences, Engineering, and Medicine. Families Caring for an Aging America: National Academies Press; 2016.
4. Chua GP, Pang GSY, Yee ACP, et al. Supporting the patients with advanced cancer and their family caregivers: what are their palliative care needs? BMC Cancer 2020 Aug 15;20(1):768. [doi: 10.1186/s12885-020-07239-9] [Medline: 32799834]
5. Badr H. Psychosocial interventions for patients with advanced cancer and their families. Am J Lifestyle Med 2016;10(1):53-63. [doi: 10.1177/1559827614530966] [Medline: 30202258]
6. Given BA, Given CW, Sherwood P. The challenge of quality cancer care for family caregivers. Semin Oncol Nurs 2012 Nov;28(4):205-212. [doi: 10.1016/j.soncn.2012.09.002] [Medline: 23107177]
7. Kim MH, Dunkle RE, Lehning AJ, Shen HW, Feld S, Perone AK. Caregiver stressors and depressive symptoms among older husbands and wives in the United States. J Women Aging 2017;29(6):494-504. [doi: 10.1080/08952841.2016.1223962] [Medline: 27673406]
8. Ramasubbu SK, Pasricha RK, Nath UK, Rawat VS, Das B. Quality of life and factors affecting it in adult cancer patients undergoing cancer chemotherapy in a tertiary care hospital. Cancer Rep (Hoboken) 2021 Apr;4(2):e1312. [doi: 10.1002/cnr2.1312] [Medline: 33295136]
9. Cella D, Riley W, Stone A, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. J Clin Epidemiol 2010 Nov;63(11):1179-1194. [doi: 10.1016/j.jclinepi.2010.04.011] [Medline: 20685078]
10. Cote DJ, Barnett I, Onnela JP, Smith TR. Digital phenotyping in patients with spine disease: a novel approach to quantifying mobility and quality of life. World Neurosurg 2019 Jun;126:e241-e249. [doi: 10.1016/j.wneu.2019.01.297] [Medline: 30797933]
11. Faraj SSA, van Hooff ML, Holewijn RM, Polly DW Jr, Haanstra TM, de Kleuver M. Measuring outcomes in adult spinal deformity surgery: a systematic review to identify current strengths, weaknesses and gaps in patient-reported outcome measures. Eur Spine J 2017 Aug;26(8):2084-2093. [doi: 10.1007/s00586-017-5125-4] [Medline: 28534221]

XSL•FO
RenderX

12. Paulsen RT, Bouknaitir JB, Fruensgaard S, Carreon L, Andersen M. Prognostic factors for satisfaction after decompression surgery for lumbar spinal stenosis. Neurosurgery 2018 May 1;82(5):645-651. [doi: 10.1093/neuros/nyx298] [Medline: 28575297]
13. Ben-Zeev D, Young MA. Accuracy of hospitalized depressed patients' and healthy controls' retrospective symptom reports: an experience sampling study. J Nerv Ment Dis 2010 Apr;198(4):280-285. [doi: 10.1097/NMD.0b013e3181d6141f] [Medline: 20386257]
14. Stone AA, Shiffman S. Capturing momentary, self-report data: a proposal for reporting guidelines. Ann Behav Med 2002;24(3):236-243. [doi: 10.1207/S15324796ABM2403_09] [Medline: 12173681]
15. Gorzelitz J, Peppard PE, Malecki K, Gennuso K, Nieto FJ, Cadmus-Bertram L. Predictors of discordance in self-report versus device-measured physical activity measurement. Ann Epidemiol 2018 Jul;28(7):427-431. [doi: 10.1016/j.annepidem.2018.03.016] [Medline: 29681429]
16. Paulhus DL. Socially desirable responding on self-reports. In: Encyclopedia of Personality and Individual Differences: Springer International Publishing AG 2017; 2017:1-5. [doi: 10.1007/978-3-319-28099-8_1349-1]
17. Dagum P. Digital biomarkers of cognitive function. NPJ Digit Med 2018;1:10. [doi: 10.1038/s41746-018-0018-4] [Medline: 31304295]
18. Onnela JP, Rauch SL. Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. Neuropsychopharmacology 2016 Jun;41(7):1691-1696. [doi: 10.1038/npp.2016.7] [Medline: 26818126]
19. Mobile fact sheet. Pew Research Center. 2023. URL: https://www.pewresearch.org/internet/fact-sheet/mobile/ [accessed 2025-01-16]
20. Ng SM, Fung MHY, Chan JSM, Chan CHY, Chan CLW. Physical activity, confidence and quality of life among cancer patient-carer dyads. Sports Med Open 2021 Jul 1;7(1):46. [doi: 10.1186/s40798-021-00333-7] [Medline: 34196850]
21. Stout NL, Santa Mina D, Lyons KD, Robb K, Silver JK. A systematic review of rehabilitation and exercise recommendations in oncology guidelines. CA Cancer J Clin 2021 Mar;71(2):149-175. [doi: 10.3322/caac.21639] [Medline: 33107982]
22. Misiąg W, Piszczyk A, Szymańska-Chabowska A, Chabowski M. Physical activity and cancer care-a review. Cancers (Basel) 2022 Aug 27;14(17):4154. [doi: 10.3390/cancers14174154] [Medline: 36077690]
23. Kulchycki M, Halder HR, Askin N, et al. Aerobic physical activity and depression among patients with cancer: a systematic review and meta-analysis. JAMA Netw Open 2024 Oct 1;7(10):e2437964. [doi: 10.1001/jamanetworkopen.2024.37964] [Medline: 39378035]
24. Mahindru A, Patil P, Agrawal V. Role of physical activity on mental health and well-being: a review. Cureus 2023 Jan;15(1):e33475. [doi: 10.7759/cureus.33475] [Medline: 36756008]
25. Cai L, Boukhechba M, Gerber MS, et al. An integrated framework for using mobile sensing to understand response to mobile interventions among breast cancer patients. Smart Health (2014) 2020 Mar;15:100086. [doi: 10.1016/j.smhl.2019.100086]
26. Panda N, Solsky I, Hawrusik B, et al. Smartphone Global Positioning System (GPS) data enhances recovery assessment after breast cancer surgery. Ann Surg Oncol 2021 Feb;28(2):985-994. [doi: 10.1245/s10434-020-09004-5] [Medline: 32812109]
27. Saeb S, Lattie EG, Kording KP, Mohr DC. Mobile phone detection of semantic location and its relationship to depression and anxiety. JMIR Mhealth Uhealth 2017 Aug 10;5(8):e112. [doi: 10.2196/mhealth.7297] [Medline: 28798010]
28. Wang R, Wang W, Aung MH, et al. Predicting symptom trajectories of schizophrenia using mobile sensing. GetMobile Mobile Comp Comm 2018 Sep 5;22(2):32-37. [doi: 10.1145/3276145.3276157]
29. Odom JN, Lee K, Currie ER, et al. Feasibility and Acceptability of Collecting Passive Smartphone Data for Potential Use in Digital Phenotyping Among Family Caregivers and Patients With Advanced Cancer. JCO Clin Cancer Inform 2025 Jan;9:e2400201. [doi: 10.1200/CCI-24-00201] [Medline: 39746166]
30. Torous J, Kiang MV, Lorme J, Onnela JP. New tools for new research in psychiatry: a scalable and customizable platform to empower data driven smartphone research. JMIR Ment Health 2016 May 5;3(2):e16. [doi: 10.2196/mental.5165] [Medline: 27150677]
31. Hays RD, Bjorner JB, Revicki DA, Spritzer KL, Cella D. Development of physical and mental health summary scores from the Patient-Reported Outcomes Measurement Information System (PROMIS) global items. Qual Life Res 2009 Sep;18(7):873-880. [doi: 10.1007/s11136-009-9496-9] [Medline: 19543809]
32. Hays RD, Schalet BD, Spritzer KL, Cella D. Two-item PROMIS® global physical and mental health scales. J Patient Rep Outcomes 2017;1(1):2. [doi: 10.1186/s41687-017-0003-8] [Medline: 29757325]
33. PROMIS® score cut points. HealthMeasures. 2021. URL: https://www.healthmeasures.net/score-and-interpret/interpret-scores/promis/promis-score-cut-points [accessed 2025-01-16]
34. Blumenthal KJ, Chang Y, Ferris TG, et al. Using a self-reported global health measure to identify patients at high risk for future healthcare utilization. J Gen Intern Med 2017 Aug;32(8):877-882. [doi: 10.1007/s11606-017-4041-y] [Medline: 28341894]
35. Katzan IL, Thompson N, Schuster A, Wisco D, Lapin B. Patient-reported outcomes predict future emergency department visits and hospital admissions in patients with stroke. J Am Heart Assoc 2021 Mar 16;10(6):e018794. [doi: 10.1161/JAHA.120.018794] [Medline: 33666094]

XSL•FO
RenderX

36. Müller SR, Bayer JB, Ross MQ. Analyzing GPS data for psychological research: a tutorial. AMPPS 2022;5(2):25152459221082680. [doi: 10.1177/25152459221082680]
37. Harari GM, Müller SR, Stachl C, et al. Sensing sociability: individual differences in young adults' conversation, calling, texting, and app use behaviors in daily life. J Pers Soc Psychol 2020 Jul;119(1):204-228. [doi: 10.1037/pspp0000245] [Medline: 31107054]
38. Sinnott RW. Virtues of the Haversine. Sky Telescope 1984;68(2):158.
39. Saeb S, Zhang M, Karr CJ, et al. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. J Med Internet Res 2015 Jul 15;17(7):e175. [doi: 10.2196/jmir.4273] [Medline: 26180009]
40. Rasmussen EM. Clustering algorithms. In: Information Retrieval: Data Structures &Amp; Algorithms: Pearson Education; 1992:419-442.
41. Arthur D, Vassilvitskii S. K-means++: the advantages of careful seeding. Presented at: The Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2007; Jan 7-9, 2007; New Orleans, Louisiana URL: https://theory.stanford.edu/~sergei/papers/kMeansPP-soda.pdf [accessed 2025-01-16]
42. Magiya J. Clustering GPS coordinates and forming regions with python. Medium. 2019. URL: https://levelup.gitconnected.com/clustering-gps-co-ordinates-forming-regions-4f50caa7e4a1 [accessed 2025-01-16]
43. Efron B, Tibshirani RJ. An Introduction to the Bootstrap: CRC press; 1994.
44. Davison AC, Hinkley DV. Bootstrap Methods and Their Application: Cambridge University Press; 1997.
45. Yurkiewicz IR, Simon P, Liedtke M, Dahl G, Dunn T. Effect of Fitbit and iPad wearable technology in health-related quality of life in adolescent and young adult cancer patients. J Adolesc Young Adult Oncol 2018 Oct;7(5):579-583. [doi: 10.1089/jayao.2018.0022] [Medline: 29924668]
46. Petersen J, Austin D, Mattek N, Kaye J. Time out-of-home and cognitive, physical, and emotional wellbeing of older adults: a longitudinal mixed effects model. PLoS ONE 2015;10(10):e0139643. [doi: 10.1371/journal.pone.0139643] [Medline: 26437228]
47. Boaro A, Leung J, Reeder HT, et al. Smartphone GPS signatures of patients undergoing spine surgery correlate with mobility and current gold standard outcome measures. J Neurosurg Spine 2021 Dec 1;35(6):796-806. [doi: 10.3171/2021.2.SPINE202181] [Medline: 34450590]
48. Panda N, Rattner DW, Morse CR. Third-time ("redo-redo") anti-reflux surgery: patient-reported outcomes after a thoracoabdominal approach. Surg Endosc 2020 Jul;34(7):3092-3101. [doi: 10.1007/s00464-019-07059-4] [Medline: 31388809]
49. Ben-Zeev D, Scherer EA, Wang R, Xie H, Campbell AT. Next-generation psychiatric assessment: using smartphone sensors to monitor behavior and mental health. Psychiatr Rehabil J 2015 Sep;38(3):218-226. [doi: 10.1037/prj0000130] [Medline: 25844912]
50. Faherty LJ, Hantsoo L, Appleby D, Sammel MD, Bennett IM, Wiebe DJ. Movement patterns in women at risk for perinatal depression: use of a mood-monitoring mobile application in pregnancy. J Am Med Inform Assoc 2017 Jul 1;24(4):746-753. [doi: 10.1093/jamia/ocx005] [Medline: 28339686]
51. Wettstein M, Wahl HW, Shoval N, et al. Out-of-home behavior and cognitive impairment in older adults: findings of the SenTra Project. J Appl Gerontol 2015 Feb;34(1):3-25. [doi: 10.1177/0733464812459373] [Medline: 25548086]
52. Saeb S, Lattie EG, Schueller SM, Kording KP, Mohr DC. The relationship between mobile phone location sensor data and depressive symptom severity. PeerJ 2016;4:e2537. [doi: 10.7717/peerj.2537] [Medline: 28344895]
53. Müller SR, Peters H, Matz SC, Wang W, Harari GM. Investigating the relationships between mobility behaviours and indicators of subjective well–being using smartphone–based experience sampling and GPS tracking. Eur J Pers 2020 Sep;34(5):714-732. [doi: 10.1002/per.2262]

Abbreviations

IRB: institutional review board
MH: mental health
PH: physical health
PROM: patient-reported outcome measure
PROMIS: Patient-Reported Outcomes Measurement Information System
QOL: quality of life
REDCap: Research Electronic Data Capture

XSL•FO
RenderX

Edited by A Mavragani; submitted 03.04.24; peer-reviewed by A AL-Asadi, KL Du; revised version received 11.12.24; accepted 12.12.24; published 07.02.25.

Please cite as:
Lee K, Azuero A, Engler S, Kumar S, Puga F, Wright AA, Kamal A, Ritchie CS, Demiris G, Bakitas MA, Odom JN
Exploring the Relationship Between Smartphone GPS Patterns and Quality of Life in Patients With Advanced Cancer and Their Family Caregivers: Longitudinal Study
JMIR Form Res 2025;9:e59161
URL: https://formative.jmir.org/2025/1/e59161
doi:10.2196/59161

© Kyungmi Lee, Andres Azuero, Sally Engler, Sidharth Kumar, Frank Puga, Alexi A. Wright, Arif Kamal, Christine S. Ritchie, George Demiris, Marie A. Bakitas, J. Nicholas Odom. Originally published in JMIR Formative Research (https://formative.jmir.org), 7.2.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

eHealth Literacy and Cyberchondria Severity Among Undergraduate Students: Mixed Methods Study

Wan -Chen Hsu, PhD

Center for Teaching & Learning Development , National Kaohsiung University of Science and Technology, No.1, University Rd.,, Yanchao District, Kaohsiung, Taiwan

Corresponding Author:
Wan -Chen Hsu, PhD
Center for Teaching & Learning Development , National Kaohsiung University of Science and Technology, No.1, University Rd.,, Yanchao District, Kaohsiung, Taiwan

Abstract

Background: With the development of the internet, health care websites have become increasingly important by enabling easy access to health information, thereby influencing the attitudes and behaviors of individuals toward health issues. However, few studies have addressed public access to health information and self-diagnosis.

Objective: This study investigated the background factors and status of cyberchondria severity among college students by conducting a nationwide sample survey using the Cyberchondria Severity Scale. Further, we explored the perspective of eHealth literacy of those with scores higher than 1 SD from the mean by analyzing their recent experiences using web-based health information.

Methods: A nationally representative sample of college students was surveyed, and 802 valid responses were obtained (male: 435/802, 54.2%; female: 367/802, 45.8%; mean age 20.3, SD 1.4 years). The Cyberchondria Severity Scale was used, which consisted of 4 dimensions (increased anxiety, obsessive-compulsive hypochondria, perceived controllability, and web-based physician-patient interaction). Additionally, we recruited 9 volunteers who scored more than 1 SD above the mean for in-depth interviews on their web-based health information–seeking behaviors.

Results: Significant differences were found across the 4 dimensions of cyberchondria severity ($F_{3,2403}$=256.26; P<.001), with perceived controllability scoring the highest (mean 2.75, SD 0.87) and obsessive-compulsive hypochondria scoring the lowest (mean 2.19, SD 0.77). Positive correlations were observed between perceived controllability, web-based physician-patient interactions, increased anxiety, and obsessive-compulsive hypochondria (r=0.46-0.75, P<.001). Regression analysis indicated that health concern significantly predicted perceived controllability (β coefficient=0.12; P<.05) and web-based physician-patient interaction (β coefficient=0.16; P<.001). Interview data revealed that students often experienced heightened anxiety (8/9, 89%) and stress (7/9, 78%) after exposure to web-based health information, highlighting the need for improved health literacy and reliable information sources.

Conclusions: The study identified both benefits and risks in college students' use of web-based health information, emphasizing the importance of critical consciousness and eHealth literacy. Future research should examine how college students move from self-awareness to actionable change and the development of critical health literacy, which are essential for effective digital health engagement.

(JMIR Form Res 2025;9:e63449) doi:10.2196/63449

KEYWORDS

eHealth literacy; undergraduate student; cyberchondria; compucondria; web-based health information; health information seeking; college students

Introduction

eHealth information is widely accessible and essential for public health improvement. However, a digital divide persists, with individuals in rural areas facing greater barriers to internet access. The increasing demand for at-home self-diagnosis is challenged by the questionable quality of some health-related websites, where personal anecdotes and content from reputable organizations may lack proper verification, affecting users' judgment and decision-making. Self-diagnosis websites often provide convenience but may lack oversight and include biased information, such as covert marketing for specific treatments, leading to potential exploitation [1]. This highlights the risks of relying on unverified digital health information, including misinformation, privacy issues, and harmful health decisions. eHealth literacy generally refers to an individual's ability to obtain, read, select, critique, and use health information from the internet to solve health problems [2-4]. It comprises 3 levels:

XSL•FO
RenderX

functional literacy, which refers to basic reading skills for acquiring health information; interactive literacy, involving advanced cognitive and social skills for selecting, understanding, synthesizing, and using health information in supportive environments; and critical literacy, which includes analytical and critical skills for evaluating and responding to health information, thereby improving control over life situations [5,6].

The ability to discern and ensure the trustworthiness and usability of information in a saturated web-based environment is critical for demonstrating eHealth literacy. This ability correlates with the effectiveness of individuals' web-based health information-seeking behaviors [7]. Those with high eHealth literacy can use web-based search strategies to identify high-quality health information [8], reducing technological risks. Conversely, insufficient ability to judge web-based health information can lead to cyberchondria, which refers to excessive or repeated web searches for health-related information because of health anxiety, resulting in increased distress or anxiety [9].

Hsu [10] compared eHealth literacy between younger (18-22 years) and older (55-72 years) college students, finding significant differences in overall scores and functional literacy, with older students scoring higher. This highlights eHealth literacy's role in helping students navigate and evaluate web-based health information, a key focus of our study. While most participants found web-based health information accessible and understandable, they were skeptical about its quality, viewing it as a mix of subjective and objective content. This underscores the importance of understanding how individuals interpret health information in a complex web-based environment.

Individuals searching for health-related information on the internet seek reassurance. However, information found on the internet is not always accurate or reliable, potentially increasing uncertainty, worry, and anxiety and leading to incorrect self-diagnoses [11,12]. Chen and Lee [13] noted that people often use limited skills to search for and evaluate vast amounts of web-based information of varying quality. However, accessing health information websites can help with self-diagnosis, and excessive concern over health may lead to cyberchondria, causing anxiety and serious health risks that arise from misinterpreting symptoms based on unverified web-based information, potentially leading to undue worry or even dangerous health actions [14]. Myrick [15] conducted a naturalistic experiment with 380 Americans to examine how emotions influence attitudes and behaviors following an information search. The study identified several emotions—fear, hope, satisfaction, interest, and motivation—that impact individuals' reactions to web-based health information. In particular, Myrick highlighted that social cognitive factors, such as self-efficacy and outcome expectations, positively influence confidence and behavioral intentions when searching for health information. This finding emphasizes how emotions and cognitive assessments together encourage or deter health-related actions, demonstrating the significant role of social cognitive factors in web-based health information–seeking behaviors.

Further, Starcevic and Berle [9] highlighted that when individuals with high health anxiety search for health information on the internet, receiving timely reassurance from others (particularly doctors) can alleviate their anxiety and reinforce their behavior of searching for health information. Conversely, if the search process increases anxiety or panic, users may continue searching to alleviate their negative emotions. However, repeated disappointment can worsen their emotional burden, deepen cyberchondria severity, and potentially lead to the avoidance of web-based health information searches in the future.

Providing web-based health information and ensuring public access are significant technological issues in society today. However, the "prescriptions" found on the internet—advice or treatments from articles or user testimonials without professional oversight—often offer only temporary relief, potentially causing anxiety and worry. These web-based prescriptions can mislead individuals, resulting in inappropriate or ineffective self-treatment, underscoring the importance of reliable, professionally vetted health information to support safer health decision-making. Additionally, the difficulty of verifying the credibility of web-based health information often leads users to rely on self-diagnosis based on their symptoms, which can increase health-related anxiety. This is particularly concerning when individuals lack the tools to assess the reliability of web-based sources, potentially resulting in misinformation and heightened worry. Adolescents, for instance, frequently report feeling overwhelmed by ambiguous health data on the internet, sometimes exacerbating their fears about symptoms they may not fully understand [16].

Most people searching for web-based health information lack medical training. Specifically, being exposed to a vast amount of web-based health information often causes anxiety and emotional distress because of the inability to understand medical terms or judge the information, resulting in severe self-diagnosis reactions [17]. Shan et al [18] found that adults generally have low levels of critical health literacy. Developing this literacy through education can enhance the ability to critically assess health information, engage in critical dialogues, and make health-related decisions.

Further, the measurement methods for cyberchondria severity are in developmental stages. Early studies have measured cyberchondria using a single item; however, later studies have developed multidimensional and multi-item scales. These scales include dimensions such as compulsion (excessive searching for web-based health information), distress (negative emotions and physiological responses from searching), reassurance (seeking reassurance from medical professionals), and mistrust of medical professionals (conflict between trusting web-based health information and doctors) [19]. Starcevic [20] considered cyberchondria to be a psychological construct. Hsu et al [3] found that college students with high health standards and high health concerns develop better eHealth literacy and adopt healthy eating, exercising, and better sleeping behaviors. Thus, in addition to developing reliable assessment tools, specific contextual items for different populations should be considered. Therefore, this study includes health status and health concerns to explore their impact on cyberchondria.

XSL•FO
RenderX

Specifically, this study aims to conduct a national survey using the Cyberchondria Severity Scale to understand the background factors and current status of cyberchondria among college students; recruit participants whose Cyberchondria Severity Scale scores are more than 1 SD above the mean and explore their experiences using web-based health information from the perspective of eHealth literacy; and based on the phased research results, summarize the characteristics of college students' experiences with web-based health information using the eHealth literacy framework and provide specific recommendations for developing health-promotion competencies to enhance the application value for health education practices.

Methods

Overview

This study used a sequential mixed methods design, which used qualitative and quantitative research sequentially, depending on the purpose or problem of the study. The purpose was to attain complementarity, such as the rationality of quantitative data in additional sampling and to investigate further through qualitative research [21]. Accordingly, this study was conducted in two phases. The first phase, comprising the quantitative part of the study, aimed to screen the respondents. The second phase, comprising interviews, aimed to collect the respondents' eHealth literacy data, which were analyzed qualitatively.

Recruitment

In this study, we specifically focused on Taiwanese college students. For the purposes of this research, a college student was defined as an individual aged 18‑22 years currently enrolled in an undergraduate program at a Taiwanese higher education institution. This age range aligned with the typical age group of undergraduate students in Taiwan and provided a consistent basis for analysis within this demographic. The survey was conducted in two waves: at the end of the first (September 2021) and second (June 2022) semesters. We recruited 802 participants, including students from 14 universities in Taiwan, and we administered a questionnaire to them. Subsequently, we selected 9 volunteers for interviews based on their Cyberchondria Severity Scale scores, specifically those whose scores exceeded 1 SD above the mean, indicating a higher level of cyberchondria. Initial contact was made through the survey platform, and we followed strict ethical guidelines to protect participants' privacy throughout the recruitment and interview process.

Ethical Considerations

Ethical approval was granted by the Ethics Research Committee of National Cheng Kung University (No. 110-573-2), and informed consent was obtained from all participants. Participants had the right to withdraw from the study at any time. To preserve anonymity in the initial survey, all responses were collected without identifiable information. For follow-up interviews, we requested separate, voluntary consent from participants who met the selection criteria. All methods were carried out per relevant ethical standards, guidelines, and regulations. Participants who completed the initial survey received a US $3 gift voucher as a token of appreciation for their time and effort. Additionally, participants who took part in the follow-up interviews were compensated with a US $15 gift voucher to acknowledge their contribution to the study. These measures aimed to encourage participation while ensuring voluntary involvement without undue influence. By implementing these compensation protocols, we maintained ethical practices that aligned with institutional guidelines and ensured the rights and welfare of the participants were upheld throughout the research process.

Survey

The Cyberchondria Severity Scale was modified from those proposed by Durak Batigun et al [22] and White and Horvitz [23]. A 5-point Likert scale was used, with the following responses: (1) never, (2) rarely, (3) sometimes, (4) often, and (5) always. The exploratory factor analysis led to a Cyberchondria Severity Scale for college students with 4 constructs: increased anxiety (Cronbach α=0.91), obsessive-compulsive hypochondria (Cronbach α=0.87), perceived controllability (Cronbach α=0.88), and web-based physician-patient interaction (Cronbach α=0.86). The total scale Cronbach α was 0.92, with a variance of 66.81%. The confirmatory factor analysis indicated that item reliability ranged from 0.50 to 0.86, factor loadings ranged from 0.71 to 0.93, and the composite reliability for latent variables ranged from 0.83 to 0.90 (P<.001). The extracted average variance was from 0.46 to 0.60. Increased anxiety referred to the anxiety experienced by an individual due to apprehensions about web-based content-seeking to understand a disease or its symptoms (eg, "To understand a certain disease or symptom, I went on the internet to search for information. The large amount of information was too much for me to handle and increased my anxiety"). Obsessive-compulsive hypochondria referred to excessive worry or excessive internet surfing to understand a disease or its symptoms (eg, "I feel that I am someone who worries excessively about health"). Perceived controllability referred to the decreased doubts or psychological burdens caused by a web search for health information (eg, "Searching for health-related information on the internet allows me to understand my medical condition and decreases my anxiety"). Web-based physician-patient interaction refered to an individual's web-based information search that affected their judgment of their symptoms (eg, "After seeing a physician, I would use the internet to obtain detailed information on the diagnosis").

To understand the advantages, disadvantages, and risks of college students' access to web-based health information, this study used semistructured interviews. The interview outline was developed based on the research objectives and focused on collecting information about the background of frequent web-based health information users who engaged in self-diagnosis, as well as their experiences, including the advantages, disadvantages, and risks. This information was gathered from college students.

Data Analysis

In the first phase of the study, SPSS version 20 (IBM Corp) was used to calculate the mean, SD, and dependent‑sample one-way ANOVA to indicate the current status and differences

XSL•FO
RenderX

across various dimensions of college students' cyberchondria. In the second phase, the Strauss and Corbin constant comparative method was used to analyze the advantages, disadvantages, and risks of web-based health information [23]. Open coding was first conducted by assigning codes and noting data entry dates. For example, for Participant 8-20220920, the number 20220920 indicates that the interview with Participant 8 was conducted on September 20, 2022. Thereafter, the researcher read the transcripts, conceptualized the data, and grouped similar concepts into categories. Finally, axial coding was conducted to link the subcategories and main categories. This study followed Lincoln and Guba's [24] credibility, dependability, and confirmability criteria to ensure research quality. With the participants' consent, data were collected by recording and note-taking to ensure completeness and accuracy.

Results

Current Situation of Cyberchondria Severity Among College Students

The sociodemographic details of the sample are shown in Table 1. Regarding the different dimensions of cyberchondria, the average single-item scores ranged from 2.19 to 2.75 (between rarely and sometimes). The average scores for each dimension were below the median, indicating that college students did not frequently experience hypochondria during web-based health information searches. However, they did not often feel that their anxiety was controllable or engage in web-based physician-patient interactions (Table 2).

Table . Sociodemographic details of the sample (n=802).

Variable and group	Frequency, n (%)
Sex	
Male	435 (54.2)
Female	367 (45.8)
Grade	
Freshman year	565 (70.4)
Sophomore year	74 (9.2)
Junior year	87 (10.8)
Senior year	76 (9.5)
Health status	
Very unhealthy	4 (5)
Unhealthy	53 (6.6)
Neutral	252 (31.4)
Healthy	298 (37.2)
Very healthy	195 (24.3)
Degree of health concern	
Very unconcerned	12 (1.5)
Unconcerned	55 (6.9)
Neutral	266 (33.2)
Concerned	287 (35.8)
Very concerned	182 (22.7)

Table . Descriptive statistics of cyberchondria severity (n=802).

Single-item	Minimum score	Maximum score	Likert score, mean (SD)
Increased anxiety	1	5	2.39 (0.82)
Obsessive-compulsive hypochondria	1	5	2.19 (0.77)
Perceived controllability	1	5	2.75 (0.87)
Online physician–patient interaction	1	5	2.70 (0.80)

Owing to the different number of items in the 4 dimensions of cyberchondria among college students, a dependent - sample one-way ANOVA was conducted using the single-item mean scores of the 4 dimensions. The results showed significant differences across the dimensions ($F_{3,2403}$=256.26; P<.001), indicating significant variations in cyberchondria scores among college students across different dimensions. Further post hoc comparisons using the least significant difference test revealed

XSL•FO
RenderX

that perceived controllability had the highest score, with the highest single item being "When I search for health-related information on the internet, reading about my condition from reliable sources (eg, hospital websites) alleviates my anxiety" (mean 2.79, SD 0.75). Obsessive-compulsive hypochondria scored the lowest, indicating that college students did not consider themselves to be overly concerned about their health (Table 3).

Table . Dependent - sample one-way ANOVA of the different dimensions of cyberchondria severity among college students.

Analysis	Type III sum of squares (*df*)	Mean square
Between groups	513.11 (801)	0.64
Within groups	513.11 (801)	0.64
Each dimension[a]	513.61 (3)	171.20
Residual error	1605.41 (2403)	0.668
Total	2632.13 (3207)	—

[a]Post hoc test revealed perceived controllability>web-based physician–patient interaction>increased anxiety>obsessive-compulsive hypochondria ($F_{3,2403}$=256.26, P<.001).

The correlation matrix of different dimensions indicated significant positive correlations between perceived controllability, web-based physician-patient interaction, increased anxiety, and obsessive-compulsive hypochondria. This suggests that the more college students feel they can control their anxiety by searching for web-based health information and the more frequently they interact with web-based physicians, the higher their levels of anxiety and obsessive-compulsive hypochondria. Further, perceived controllability exhibited a significant positive correlation with web-based physician-patient interactions. The absolute values of the correlation coefficients were all above 0.40, indicating moderate-to-high correlations. Thus, the most significant dimensions of cyberchondria among college students showed moderate or high correlations (Table 4).

Table . Correlation analysis of different dimensions of cyberchondria severity among college students.

Variable	Increased anxiety	Obsessive-compulsive hypochondria	Perceived controllability	Online physician–patient interaction
Increased anxiety				
r	1.00	0.75	0.46	0.48
P value	__[a]	<.001	<.001	<.001
Obsessive-compulsive hypochondria				
r	0.75	1.00	0.50	0.54
P value	<.001	—	<.001	<.001
Perceived controllability				
r	0.46	0.50	1.00	0.73
P value	<.001	<.001	—	<.001
Online physician–patient interaction				
r	0.48	0.54	0.73	1.00
P value	<.001	<.001	<.001	—

[a]Statistical analysis was not done for reference values.

Table 5 indicates that all individual variables were significant predictors of the 4 dimensions of cyberchondria severity, with an overall explanatory power of 1% - 2%, corresponding with a low predictive power level. Notably, the degree of health concern emerged as a significant positive indicator of all dimensions of cyberchondria severity, except increased anxiety.

XSL•FO
RenderX

Table . Multiple regression analyses of individual factors predicting cyberchondria severity.

Variable	Increased anxiety[a]				Obsessive-compulsive hypochondria[b]				Perceived controllability[c]				Web-based physician-patient interaction[d]			
	B[e]	β[f]	*t* test (*df*)	*P* value	B	β	*t* test (*df*)	*P* value	B	β	*t* test (*df*)	*P* value	B	β	*t* test (*df*)	*P* value
Health status	−0.09	−0.10	−2.32 (790)	.02	−0.11	−0.13	−2.87 (790)	.006	0.00	0.00	0.07 (790)	.43	−0.01	−0.01	−0.29 (790)	.06
Degree of health concern	0.01	0.01	0.18 (790)	.08	0.11	0.13	2.93 (790)	.004	0.11	0.12	2.58 (790)	.02	0.14	0.16	3.69 (790)	<.001

[a]R=0.10, R^2=0.01 (P=.02); $F_{3,790}$=2.73 (P=.02).

[b]R=0.12, R^2=0.01 (P=.02); $F_{3,790}$=3.55 (P=.02).

[c]R=0.12, R^2=0.02 (P=.002); $F_{3,790}$=3.90 (P=.003).

[d]R=0.16, R^2=0.02 (P<.001); $F_{3,790}$=6.54 (P<.001).

[e]B: unstandardized coefficients.

[f]β: standardized coefficients.

Impact of Accessing Web-Based Health Information on the Physical and Mental Well-Being of College Students

The interviews with college students revealed that accessing web-based health information has both psychological and physical effects, often creating challenges due to the nature and quality of the information available on the internet. In terms of psychological impact, participants reported feeling overwhelmed by the plethora of opinions and conflicting advice, which often caused confusion and heightened anxiety. For example, one student shared the following:

> *I could not find any results through searching, so I mainly see a doctor because I have no other option. I don't know what to do anymore, so I must see a doctor. And since I've been paying more attention to health issues, I've noticed that the first things that pop up are usually some exaggerated news or bizarre articles from content farms. The headlines make me feel like, "Oh my god! I'm about to die."* [Participant 8-20220920]

This highlights the stress caused by exaggerated or unreliable health information, which drives the students to consult health care professionals as a last resort. As for strained doctor-patient relationships, the tendency to self-diagnose based on web-based information sometimes led students to question or dismiss professional medical advice, creating tension in doctor-patient interactions. One student remarked the following:

> *If you have doubts about the doctor's words, don't ask random questions. For instance, if I now Google some issues, like searching for stomach pain on Google, and then tell the doctor, "But Google says this is some specific disease, so you just need to prescribe me this medicine," it can lead to problems.* [Participant 1-20221101]

This underscores the potential of web-based health information to undermine the trust in medical expertise. Additionally, exposure to sensational or clickbait content exacerbated feelings of fear and helplessness. Students described how alarming headlines led to unnecessary worry:

> *And since I've been paying more attention to health issues, I've noticed that the first things that pop up are usually some exaggerated news… The headlines make me feel like, " Oh my god! I'm about to die."* [Participant 8-20220920]

The physical impact included misguided health practices and delayed medical treatment. Some students engaged in health practices based on unverified advice on the internet, which sometimes led to negative outcomes. For example, one participant shared a personal experience:

> *My chest is not very big, and I saw online that eating green papaya can make your chest bigger. I was just in middle school then, so I started eating one every day… Then I noticed that my hands were starting to turn yellow.* [Participant 9-20221024]

Moreover, in some cases, students delayed seeking professional care due to reliance on web-based advice. One participant shared a story about their friend:

> *Like my friend, he once sprained his finger while playing basketball… He looked it up online and found advice saying to ice it and not press it. He kept doing that, and now his finger is crooked.* [Participant 3-20221107]

This highlights the risk of self-experimentation based on health misinformation and demonstrates how misleading information can lead to improper self-care and long-term consequences. The findings emphasize the dual nature of web-based health information. Although such information is accessible and convenient, its unreliable or sensational nature can lead to considerable psychological and physical challenges for students.

XSL•FO
RenderX

This emphasizes the urgent need for improving students' health literacy on the internet and promoting access to credible and professionally vetted health information.

College Students' Experiences With Accessing Web-Based Health Information: The Advantages, Disadvantages, and Risks

Students' responses to the semistructured interviews revealed convenience, richness of information, privacy, and validation as the 4 main themes related to using web-based health information. The motivation for using web-based health information was to meet the needs of beautifying features, weight loss, and fitness. The two types of web-based health information access were fixed webpage browsing and nonfixed webpage retrieving. The criteria for checking the quality of web-based health information included subjective judgments and objective features. Regarding web-based health information reading experiences, problems were caused by technical terms, foreign languages, and numerical health information data. Typically, the participants searched for related information on the internet to clarify their doubts. Finally, the impacts of using web-based health information depend on assessing its quality, lifestyle changes, and the possibility of self-practice. Self-diagnosis can be a double-edged sword. It may empower individuals with preliminary health insights; however, it also poses risks, particularly if the information is misinterpreted or lacks professional verification. Individuals should be encouraged to evaluate web-based health information and approach self-diagnosis with caution. Additionally, we will consider the ethical implications of promoting self-diagnosis in health-related research, emphasizing the importance of fostering critical eHealth literacy among users as a suggestion.

The participants reported various risks and benefits while searching for health information on the internet. They tended to be wary of sensational or emphatically presented headlines, leading them to adopt a conservative approach to avoid harm. In assessing the quality of information, they paid attention to the source, crosschecked it with multiple sources, and judged its effectiveness based on their personal experiences and expert endorsements, ensuring content completeness. Their experiences included selective understanding, focusing on general ideas rather than technical terms, and further clarifying doubtful information. The impact of this information varied based on the closeness of relationships with others and the feasibility of implementing lifestyle changes. For instance, support from close friends and family could enhance the likelihood of adopting positive health behaviors and reduce anxiety related to conflicting information. Conversely, a lack of support or conflicting opinions from peers could heighten stress and uncertainty, making it more challenging to discern and apply accurate health information.

Summarizing the experiences of college students who engaged in self-diagnosis through extensive internet use for health information access, the benefits primarily included cross-referencing, synthesizing diverse information, and verifying questionable information. The harms mainly arose from confusion due to difficulty reading and understanding the diverse information and, in more severe cases, engaging in self-experimentation that affected physical health. The risks involved selective understanding, blind sharing of information, and extreme rejection of any information, which could undermine the positive intentions of effectively using web-based health information. Table 6 presents the interview analysis and coding framework.

XSL•FO
RenderX

Table . Interview analysis and coding framework.

Main and subcategory structure	Interviewee numbers
Background of using web-based health information	
Time of initial use	
I mostly started in high school	1, 3, 4, 6, 7, 8, 9
Initial motivation for using web-based health information	
Knowledge acquisition	1, 2, 4, 5, 6, 8, 9
Concern for one's condition	1, 2, 3, 4, 5, 6, 7, 8, 9
Searching for health and medical information for family	2, 5, 6, 8, 9
Searching for drug ingredients	2, 3, 5, 6, 7, 8
Looking for nearby medical facilities	1, 2, 4, 7, 8
Reason for choosing web-based health information	
Free of charge	1, 2, 3, 4, 6, 8
Richness of information	1, 2, 5, 7, 8, 9
Verifiability	1, 2, 3, 4, 5, 6, 7, 8
Convenience	1, 2, 4, 5, 7, 8, 9
Experience in accessing web-based health information	
Quality assessment of web-based health information: subjective judgment	
Sensational headlines	1, 4, 5, 6, 8
Headlines overly emphasizing efficacy	1, 4, 5, 6, 7, 8, 9
Conservative adoption to avoid harm	2, 3, 5, 7, 8
Judged by my common sense	1, 3, 5, 6, 8, 9
Quality assessment of web-based health information: objective standards	
Source of information	1, 2, 4, 5, 6, 8
Cross-verification	2, 3, 5, 6, 7, 8, 9
Judged by actual experiences of friends and family	1, 2, 3, 4, 7, 8
Expert evaluation of the efficacy of information	1, 3, 4, 5, 6, 7
Incomplete information	2, 3, 4, 5, 6, 9
Experience in reading web-based health information	
Summary understanding and further inquiry	
Handling doubts about web-based health information: continue searching or ask others interesting questions	1, 2, 3, 4, 5, 6, 7, 8, 9
Understanding of web-based health information: often ignoring technical terms and focusing on general meaning	1, 4, 5, 6, 8, 9
Impact of web-based health information	
Deliberation on object and risk outcome assessment	
Influence on others based on personal relationships: varying forms and degrees of influence	1, 2, 3, 4, 7, 9
Self-physical and mental impact: depending on the extent of lifestyle changes or feasibility	1, 2, 3, 4, 5, 6, 7, 8, 9

Discussion

Principal Findings

Digital platforms have revolutionized health and science by enabling collaboration, expanding evidence-based interventions, and advancing public health. Early successes in behavioral medicine demonstrated their potential to scale therapeutic solutions. With growing investments in health information technology, behavioral informatics has become essential, integrating medicine, psychology, and engineering to develop patient apps, clinical tools, communication programs, and population health strategies. The Society of Behavioral

XSL•FO
RenderX

Medicine, positioned at this interdisciplinary nexus, plays a pivotal role in shaping how science is applied to reduce health disparities, support patients and caregivers, enhance community well-being, and address challenges such as psychological pain and addiction [25]. This study examined eHealth literacy and cyberchondria severity in college students, highlighting their web-based health information-seeking behaviors. Key findings include links between perceived controllability, web-based interactions, anxiety, and obsessive-compulsive hypochondria, with health concern predicting controllability and interactions. Interviews showed that 89% (8/9) felt heightened anxiety, and 78% (7/9) reported stress from web-based health information.

These results highlight the dual nature of web-based health information. The high perceived controllability scores suggest that accessing reliable web-based health information—such as from trusted sources like hospital websites—can alleviate student anxiety by helping them manage health-related stress. This finding aligns with prior research showing the importance of credible sources in fostering a sense of control [10,23]. By contrast, the low obsessive-compulsive hypochondria scores suggest that students generally do not exhibit excessive worry or compulsive searching behaviors. This is contrary to the studies linking frequent web-based health searches to increased distress and cyberchondria severity [9]. This discrepancy may stem from differences in digital literacy or cultural attitudes toward web-based health information.

The positive association between health concern and cyberchondria dimensions underscores the dual impact of web-based health information. Although such information can promote awareness and self-management, it can also lead to anxiety and stress when students encounter conflicting or sensational content; this is consistent with the findings from Tan and Goonawardene's systematic review [26]. Moreover, the role of web-based physician-patient interactions in moderating anxiety highlights the importance of guiding students toward appropriate and reliable eHealth resources. This aligns with Hsu's [10] findings that enhancing eHealth literacy can empower students to critically evaluate web-based health information and reduce health-related anxiety.

Notably, the interviews revealed that excessive web-based health searches contribute to confusion and emotional distress, particularly when students encounter contradictory advice or exaggerated claims. These results align with those of Starcevic and Berle [9], who identified increased anxiety and obsessive-compulsive tendencies as key outcomes of frequent health information searches. Furthermore, exposure to sensational content, such as clickbait headlines, exacerbates fear and helplessness, increasing the likelihood of unnecessary medical consultations or misdiagnoses [11,23]. These findings emphasize the need to develop interventions that improve students' critical health literacy, which would enable them to discern credible information and mitigate the psychological impacts of misinformation.

This study has limitations. Self-reported data may introduce response bias, suggesting the need for objective health measures like medical records for greater reliability. The qualitative phase included only 9 participants, restricting the generalizability of the findings; a larger sample could offer deeper insights. Additionally, focusing solely on Taiwanese college students limits the applicability of the results to other cultures or demographics. Cross-cultural studies are recommended to examine variations in eHealth literacy and cyberchondria across populations.

This study highlights the complex interplay between eHealth literacy and cyberchondria severity among college students, revealing both benefits and risks associated with web-based health information-seeking behavior. Reliable web-based information can reduce anxiety and enhance self-management, whereas the prevalence of contradictory or sensational content underscores the importance of fostering critical eHealth literacy. Drawing on Nutbeam's framework of health literacy [18,27], this study advocates for integrating critical health literacy into health education programs to empower students to effectively navigate the digital health landscape.

Moreover, the findings contribute to the broader discourse on digital health education, emphasizing the need for educational interventions that promote self-awareness and actionable behavioral changes. As previous studies have shown, developing critical eHealth literacy not only improves individual decision-making but also strengthens patient-physician relationships by encouraging open and informed discussions [26,28]. Future efforts should focus on designing tailored health literacy initiatives that equip students with the skills to critically assess and apply web-based health information, fostering their ability to engage responsibly with digital health resources in a rapidly evolving technological landscape.

Conclusions

This study surveyed college students using the Cyberchondria Severity Scale and interviewed high scorers to explore their browsing behaviors and experiences with web-based health information. Significant score differences were observed across cyberchondria dimensions, with perceived controllability scoring highest and obsessive-compulsive hypochondria lowest. High scorers noted benefits such as cross-verification and clarification of information; drawbacks like confusion from complexity; and risks including selective understanding, reflexive sharing, and mistrust, which could undermine the advantages of web-based health resources.

The internet has revolutionized medical education and practice by enabling the rapid, large-scale exchange of information and ideas across geographically dispersed audiences [29]. To effectively use web-based health information, critical awareness and eHealth literacy are essential skills. Future research should investigate how college students move from self-awareness to a behavioral change that helps them develop critical health literacy as a crucial component of digital health education.

This study relied on self-reported questionnaires, which may introduce response bias and affect the accuracy of findings, particularly regarding health status. Future studies could improve data reliability by encouraging participants to reference recent health examination results, enhancing internal validity. Additionally, the small interview sample limits the

XSL•FO
RenderX

generalizability of conclusions. Expanding the sample in future research could reduce bias and support more robust findings.

Acknowledgments

The authors thank the National Science and Technology Council for funding this study.

Data Availability

The anonymized data from the survey are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

References

1. Ryan A, Wilson S. Internet healthcare: do self-diagnosis sites do more harm than good? Expert Opin Drug Saf 2008 May;7(3):227-229. [doi: 10.1517/14740338.7.3.227] [Medline: 18462181]
2. Chiang CH, Yang SC, Hsu WC. Development and validation of the e-health literacy scale and investigation of the relationships between e-health literacy and healthy behavior among undergraduate students in Taiwan. Formosa J Ment Health 2015;23(3):389-420. [doi: 10.30074/FJMH]
3. Hsu WC, Chiang CH, Yang SC. The effect of individual factors on health behaviors among college students: the mediating effects of eHealth literacy. J Med Internet Res 2014 Dec 12;16(12):e287. [doi: 10.2196/jmir.3542] [Medline: 25499086]
4. Norman CD, Skinner HA. eHEALS: the eHealth Literacy Scale. J Med Internet Res 2006 Nov 14;8(4):e27. [doi: 10.2196/jmir.8.4.e27] [Medline: 17213046]
5. Nutbeam D. The evolving concept of health literacy. Soc Sci Med 2008 Dec;67(12):2072-2078. [doi: 10.1016/j.socscimed.2008.09.050] [Medline: 18952344]
6. Nutbeam D, Kickbusch I. Advancing health literacy: a global challenge for the 21st century. Health Promot Int 2000 Sep 1;15(3):183-184. [doi: 10.1093/heapro/15.3.183]
7. Hsu WC, Chen SF, Ho CJ. Experience of using web health information among college students: an analysis from the health literacy perspective. J Health Promot Health Educ Contents 2011;35:1-22. [doi: 10.7022/JHPHE.201106.0001]
8. Quinn S, Bond R, Nugent C. Quantifying health literacy and eHealth literacy using existing instruments and browser-based software for tracking online health information seeking behavior. Comput Human Behav 2017 Apr;69:256-267. [doi: 10.1016/j.chb.2016.12.032]
9. Starcevic V, Berle D. Cyberchondria: towards a better understanding of excessive health-related internet use. Expert Rev Neurother 2013 Feb;13(2):205-213. [doi: 10.1586/ern.12.162] [Medline: 23368807]
10. Hsu WC. The effect of age on electronic health literacy: mixed-method study. JMIR Hum Factors 2019 Apr 21;6(2):e11480. [doi: 10.2196/11480] [Medline: 31066696]
11. Hsu WC. Undergraduate students' online health information-seeking behavior during the COVID-19 pandemic. Int J Environ Res Public Health 2021 Dec 16;18(24):13250. [doi: 10.3390/ijerph182413250] [Medline: 34948859]
12. Xu DY, Ding JL, Yang ZH. Pathogenesis of cyberchondria and its impact on individuals: a qualitative study. Chin Gen Pract 2018;21(22):2755-2759. [doi: 10.12114/j.issn.1007-9572.2018.22.022]
13. Chen W, Lee KH. More than search? Informational and participatory eHealth behaviors. Comput Human Behav 2014 Jan;30:103-109. [doi: 10.1016/j.chb.2013.07.028]
14. Smith PK, Fox AT, Davies P, Hamidi-Manesh L. Cyberchondriacs. Int J Adolesc Med Health 2006;18(2):209-213. [Medline: 16894859]
15. Myrick JG. The role of emotions and social cognitive variables in online health information seeking processes and effects. Comput Human Behav 2017 Mar;68:422-433. [doi: 10.1016/j.chb.2016.11.071]
16. Gray NJ, Klein JD, Noyce PR, Sesselberg TS, Cantrill JA. Health information-seeking behaviour in adolescence: the place of the internet. Soc Sci Med 2005 Apr;60(7):1467-1478. [doi: 10.1016/j.socscimed.2004.08.010] [Medline: 15652680]
17. Jokić-Begić N, Mikac U, Čuržik D, Sangster Jokić C. The development and validation of the Short Cyberchondria Scale (SCS). J Psychopathol Behav Assess 2019 Dec;41(4):662-676. [doi: 10.1007/s10862-019-09744-z]
18. Shan Y, Ji M, Dong Z, Xing Z, Xu X. Assessing patients' critical health literacy and identifying associated factors: cross-sectional study. J Med Internet Res 2023 Apr 5;25:e43342. [doi: 10.2196/43342] [Medline: 37018027]
19. McElroy E, Shevlin M. The development and initial validation of the Cyberchondria Severity Scale (CSS). J Anxiety Disord 2014 Mar;28(2):259-265. [doi: 10.1016/j.janxdis.2013.12.007] [Medline: 24508033]
20. Starcevic V. Cyberchondria: challenges of problematic online searches for health-related information. Psychother Psychosom 2017;86(3):129-133. [doi: 10.1159/000465525] [Medline: 28490037]
21. Venkatesh V, Brown SA, Bala H. Bridging the qualitative-quantitative divide: guidelines for conducting mixed methods research in information systems. MISQ 2013 Jan 1;37(1):21-54. [doi: 10.25300/MISQ/2013/37.1.02]

XSL•FO
RenderX

22. Durak Batigun A, Gor N, Komurcu B, Senkal Erturk I. Cyberchondria Scale (CS): development, validity and reliability study. Dusunen Adam 2018 Jun 29;31:148-162. [doi: 10.5350/DAJPN2018310203]
23. White RW, Horvitz E. Cyberchondria: studies of the escalation of medical concerns in web search. ACM Trans Inf Syst 2009;27(4):23. [doi: 10.1145/1629096.1629101]
24. Lincoln YS, Guba EG. Naturalistic Inquiry: Sage Publications, Inc; 1985.
25. Hesse BW. Riding the wave of digital transformation in behavioral medicine. Ann Behav Med 2020 Dec 1;54(12):960-967. [doi: 10.1093/abm/kaaa093] [Medline: 33416839]
26. Tan SSL, Goonawardene N. Internet health information seeking and the patient-physician relationship: a systematic review. J Med Internet Res 2017 Jan 19;19(1):e9. [doi: 10.2196/jmir.5729] [Medline: 28104579]
27. Abel T, Benkert R. Critical health literacy: reflection and action for health. Health Promot Int 2022 Aug 1;37(4):daac114. [doi: 10.1093/heapro/daac114] [Medline: 36047637]
28. Devan H, Perry MA, van Hattem A, et al. Do pain management websites foster self-management support for people with persistent pain? A scoping review. Patient Educ Couns 2019 Sep;102(9):1590-1601. [doi: 10.1016/j.pec.2019.04.009] [Medline: 30981410]
29. Levy JA, Strombeck R. Health benefits and risks of the internet. J Med Syst 2002 Dec;26(6):495-510. [doi: 10.1023/a:1020288508362] [Medline: 12385532]

Edited by A Mavragani; submitted 20.06.24; peer-reviewed by J Behre, R Peng; revised version received 12.01.25; accepted 15.01.25; published 03.02.25.

Please cite as:
Hsu C
eHealth Literacy and Cyberchondria Severity Among Undergraduate Students: Mixed Methods Study
JMIR Form Res 2025;9:e63449
URL: https://formative.jmir.org/2025/1/e63449
doi:10.2196/63449

© Wan-Chen Hsu. Originally published in JMIR Formative Research (https://formative.jmir.org), 3.2.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Estimating the Prevalence of Schizophrenia in the General Population of Japan Using an Artificial Neural Network–Based Schizophrenia Classifier: Web-Based Cross-Sectional Survey

Pichsinee Choomung[1], PharmD; Yupeng He[1], MD, PhD; Masaaki Matsunaga[1], MD, PhD; Kenji Sakuma[2], MD, PhD; Taro Kishi[2], MD, PhD; Yuanying Li[3], PhD; Shinichi Tanihara[4], PhD; Nakao Iwata[2], MD, PhD; Atsuhiko Ota[1], MD, PhD

[1]Department of Public Health, Fujita Health University School of Medicine, 1-98 Dengakugakubo, Kutsukake-cho, Toyoake, Japan

[2]Department of Psychiatry, Fujita Health University School of Medicine, Toyoake, Japan

[3]Department of Public Health and Health Systems, Nagoya University Graduate School of Medicine, Nagoya, Japan

[4]Department of Public Health, Kurume University School of Medicine, Kurume, Japan

Corresponding Author:
Yupeng He, MD, PhD
Department of Public Health, Fujita Health University School of Medicine, 1-98 Dengakugakubo, Kutsukake-cho, Toyoake, Japan

Abstract

Background: Estimating the prevalence of schizophrenia in the general population remains a challenge worldwide, as well as in Japan. Few studies have estimated schizophrenia prevalence in the Japanese population and have often relied on reports from hospitals and self-reported physician diagnoses or typical schizophrenia symptoms. These approaches are likely to underestimate the true prevalence owing to stigma, poor insight, or lack of access to health care among respondents. To address these issues, we previously developed an artificial neural network (ANN)–based schizophrenia classification model (SZ classifier) using data from a large-scale Japanese web-based survey to enhance the comprehensiveness of schizophrenia case identification in the general population. In addition, we also plan to introduce a population-based survey to collect general information and sample participants matching the population's demographic structure, thereby achieving a precise estimate of the prevalence of schizophrenia in Japan.

Objective: This study aimed to estimate the prevalence of schizophrenia by applying the SZ classifier to random samples from the Japanese population.

Methods: We randomly selected a sample of 750 participants where the age, sex, and regional distributions were similar to Japan's demographic structure from a large-scale Japanese web-based survey. Demographic data, health-related backgrounds, physical comorbidities, psychiatric comorbidities, and social comorbidities were collected and applied to the SZ classifier, as this information was also used for developing the SZ classifier. The crude prevalence of schizophrenia was calculated through the proportion of positive cases detected by the SZ classifier. The crude estimate was further refined by excluding false-positive cases and including false-negative cases to determine the actual prevalence of schizophrenia.

Results: Out of 750 participants, 62 were classified as schizophrenia cases by the SZ classifier, resulting in a crude prevalence of schizophrenia in the general population of Japan of 8.3% (95% CI 6.6%-10.1%). Among these 62 cases, 53 were presumed to be false positives, and 3 were presumed to be false negatives. After adjustment, the actual prevalence of schizophrenia in the general population was estimated to be 1.6% (95% CI 0.7%-2.5%).

Conclusions: This estimated prevalence was slightly higher than that reported in previous studies, possibly due to a more comprehensive disease classification methodology or, conversely, model limitations. This study demonstrates the capability of an ANN-based model to improve the estimation of schizophrenia prevalence in the general population, offering a novel approach to public health analysis.

(JMIR Form Res 2025;9:e66330) doi:10.2196/66330

KEYWORDS

schizophrenia; schizophrenic; prevalence; artificial neural network; neural network; neural networks; ANN; deep learning; machine learning; SZ classifier; web-based survey; epidemiology; epidemiological; Japan; classifiers; mental illness; mental disorder; mental health

XSL•FO
RenderX

Introduction

Schizophrenia is a complex psychiatric disorder characterized by significant impairments in the perception of reality and behavioral changes. It includes positive (eg, delusions, hallucinations, disorganized thinking, and notably disorganized behaviors) and negative symptoms (eg, limited speech, decreased emotional expression, inability to experience interest or pleasure, and social withdrawal). Affected individuals often experience cognitive impairments affecting memory, attention, and problem-solving abilities [1,2]. The Global Burden of Disease Study [3] reported that schizophrenia accounted for 23.6 million estimated cases in 2019, with an age-standardized prevalence rate of 287.4 per 100,000 people. Median estimates from internationally published studies indicate a 12-month prevalence of 0.33%, lifetime prevalence of 0.48%, and point prevalence of 0.32% [4]. However, there is a lack of published studies on the prevalence of schizophrenia in the general Japanese population. Only a few population-based studies have estimated the prevalence of schizophrenia in Japan. For instance, the point prevalence of schizophrenia stratified by sex and age ranges from 0.4% to 1.1% [5]. However, these figures were derived from the Vital Statistics and the Patient Survey in 2008, which probably do not represent the general population since the Patient Survey could not encompass patients who do not visit a doctor. The estimated lifetime prevalence of schizophrenia is 0.59%, according to a secondary analysis of the 2019 cross-sectional Japan National Health and Wellness Survey [6]. However, this study may produce underestimation since it only depended on physician diagnoses. Individuals with schizophrenia hardly self-identify as having schizophrenia because of stigma and lack of awareness of schizophrenia and its symptoms [7-13].

In our previous research, we developed an artificial neural network (ANN)–based schizophrenia classification model (SZ classifier) to classify schizophrenia cases in the Japanese population and verified its generalizability [14]. The model was trained by using data from a large-scale Japanese web-based survey; the presence of schizophrenia served as a response variable. The general information including demographics, health-related background, physical comorbidities, psychiatric comorbidities, and social comorbidities were collected and used as feature variables. This SZ classifier achieved high internal validity with a high area under the receiver operating characteristic curve (AUC) of 0.86 [14]. In another study, we also assessed the SZ classifier's external validity by applying it to psychiatric outpatients, achieving a sensitivity of 75% [15]. The SZ classifier suggests a potential for classifying schizophrenia cases in the general Japanese population.

In this study, we aimed to estimate the prevalence of schizophrenia in the general Japanese population using the SZ classifier through the following three approaches: (1) conducting a population-based survey to recruit participants and collect their general information, (2) sampling participants who match the distribution to the population demographic structure, and (3) using a pretrained machine learning model to detect schizophrenia cases. Study participants were randomly sampled from a pooled panel. The samples are matched to the demographic distribution of the Japanese population. General information, the same as that for the development of the SZ classifier, was collected through the web-based survey for each study participant. The estimation of the schizophrenia prevalence was implemented through the following two steps: (1) calculating the crude prevalence based on crude cases that are detected by the SZ classifier and (2) refining the crude prevalence by excluding false-positive estimates and including the false-negative estimates.

Methods

Development of the SZ Classifier

The SZ classifier was developed as part of a prior study [14] using a dataset obtained from an internet research agency's pooled panels (Rakuten Insight, Inc) [16]. The dataset included 223 individuals with schizophrenia and 1776 healthy controls. For each sample, schizophrenia status served as the response variable, while demographic details, health-related background information, and physical, psychiatric, and social comorbidities were considered feature variables. Evidence on the association between these feature variables and schizophrenia has been previously published [17,18]. For example, demographic and health-related characteristics included age, BMI, smoking status, restriction in functional capacity, and bowel frequency. Physical comorbidities included overweight, cancer, cardiovascular disease, heart failure, and hypertension. Psychiatric comorbidities included frequency of sleep medication use, bedtime, hours of sleep, perceived stress, and positive reasons for living. The social comorbidities included educational background, type of occupation, type of employment, household income, and marital status. The inclusion criteria of the study participants and definitions of the feature variables are detailed in previous publications [14,18] and provided in Multimedia Appendix 1.

The SZ classifier was constructed using an ANN concept. This multilayer perceptron neural network model comprised 5 hidden layers with 128, 64, 32, 16, and 8 neurons in each respective layer. It used the HeNormal weight initializer and the ReLU activation function in the hidden layers. A sigmoid activation function was used in the output layer. Comprehensive details can be found in a prior publication [14]. The model's performance was evaluated using the AUC score. During internal validation, the SZ classifier achieved an AUC of 0.86, accuracy of 0.93, specificity of 0.96, and sensitivity of 0.56. He et al [14] performed 10,000 bootstrapping iterations and a 10-fold cross-validation to further evaluate the algorithm's performance and achieved AUCs of 0.847 and 0.81, respectively.

External Validation of the SZ Classifier

External validation of the SZ classifier was conducted by applying it to a sample of patients who visited the psychiatric department of Fujita Health University Hospital between January 2022 and May 2023 [15]. Patients aged 20 - 75 years diagnosed with schizophrenia, major depressive disorder (MDD), bipolar disorder (BD), or obsessive-compulsive disorder (OCD) were selected. Therefore, 150 participants were included in the study: 61 with schizophrenia, 56 with MDD, 32 with BD, and one with

XSL•FO
RenderX

OCD. The participants were required to complete the same survey used to develop the SZ classifier to predict the likelihood of schizophrenia. The external validation process revealed that the SZ classifier achieved a sensitivity of 0.75 but misclassified 55.4% of patients with MDD and 59.4% of patients with BD as having schizophrenia [15].

Estimation of the Prevalence of Schizophrenia by Using the SZ Classifier

A cross-sectional study was conducted using data from a large-scale web-based survey in Japan. The number of participants was set to 750. The sample size was decided using the single population proportion formula ($n=[Z^2 \cdot p(1-p)]/e^2$) by assuming a 95% CI (Z=1.96), 1% margin of error (e=0.01), and 2% proportion of schizophrenia (p=2%) based on previous studies. Participants were randomly recruited from those registered as survey monitors with Rakuten Insight, a large-scale web-based survey incorporated approximately 2.2 million registered members as of September 2022 and provided reliable data with regular quality maintenance [16]. Specifically, the recruited criteria were set to the combination of 3 items of age intervals (20 - 75 years old, divided into 5 distinct age groups), gender (male and female), and regions (13 geographical administrative regions in Japan: Hokkaido, Tohoku, North Kanto, Greater Tokyo Area, Koshinetsu, Hokuriku, Tokai, Kinki excluding Keihanshin, Keihanshin, Chugoku, Shikoku, Kyushu excluding Okinawa, and Okinawa). For each combination category of these 3 items, the number of participants was set to match the composition of the national population as of October 1, 2021, as estimated by the Ministry of Internal Affairs and Communications, Japan. Participants were recruited until the target number for each combination category was reached. The recruitment period was from November 20 to 26, 2023. This was a web-based survey in which only authorized panelists were allowed to respond to the survey to prevent fraudulent registrants from participating. To administer the survey, respondents had to log in to their accounts to prevent multiple entries from the same individual. They were required to answer 38 questions (some questions had several subquestions) that are distributed into 38 pages. Respondents were asked to review their responses and ensure completeness before submission. Specifically, the web-based survey system notified respondents if there were any blank answers. The participation rate for this survey was estimated to be 0.30, and the completion rate was not calculated because if the respondents terminated during the answering process (ie, logged out, closed the web page, etc), they were not counted in the statistics. Incomplete questionnaires or questionnaires that were submitted too soon were not used.

Each participant's demographic information, health-related backgrounds, and physical, psychiatric, and social comorbidities were collected and applied to the SZ classifier, yielding the predicted crude number of schizophrenia cases and crude prevalence of schizophrenia. Textbox 1 shows a summary of the feature variable categories collected from study participants. These variables are the same as those used in developing the SZ classifier. For detailed information on these variables, please refer to Multimedia Appendix 1.

XSL•FO
RenderX

Textbox 1. Summary of categories of feature variables collected from study participants, including demographic, health-related backgrounds, physical comorbidities, psychiatric comorbidities, and social comorbidities.

Demographic, health-related backgrounds, and physical comorbidities:

- Age
- BMI
- Smoking
- Alcohol drinking
- Sports
- Eating behaviors
- Bowel movement
- Restriction in functional capacity
- Self-rated health status
- Physical comorbidities

Psychiatric comorbidities:

- Frequency of sleep medication use
- Bedtime
- Sleep hours and qualities
- Depressive symptoms
- Perceived stress
- Positive reason for living (ikigai)
- Internet use hours per week

Social comorbidities:

- Regularity of health checkups
- Educational background
- Household income
- Type of employment and occupation
- Marital status and family structure
- Social support and social capital

We further refined these crude estimates by excluding false positives and adding false negatives. This was based on the model's internal and external validation results and literature reviews. Details of the model's internal and external validation results are provided in the Multimedia Appendix 1.

False-positive cases included individuals without mental disorders who were misclassified as having schizophrenia and those with other mental disorders who were misclassified as having schizophrenia. The false-positive cases were calculated as follows:

$$Nfp=(Phc \cdot rhc+PMDD \cdot rMDD+PBD \cdot rBD) \cdot n$$

where N_{fp} represents the number of false-positive cases, P_{hc} represents the percentage of individuals who have not had any mental disorders in the population, r_{hc} represents the rate at which the SZ classifier falsely classified individuals without mental disorders as having schizophrenia, P_{MDD} represents the prevalence of MDD in the population, r_{MDD} represents the rate at which the SZ classifier falsely classified individuals with MDD as having schizophrenia, P_{BD} represents the prevalence of BD in the population, r_{BD} represents the rate at which the SZ classifier falsely classified individuals with BD as having schizophrenia, n represents the sample size, which included 750 participants.

The SZ classifier falsely classified 3.1% of the individuals without mental disorders as having schizophrenia [14]. The SZ classifier misclassified 55.4% of the patients with MDD and 59.4% of the patients with BD as having schizophrenia [15]. According to a literature review, 80% of the general population in Japan does not have mental disorders [19]. The point prevalence of MDD and BD was 7.9% [20] and 0.4% [21], respectively; namely, P_{hc}=80%, r_{hc}=3.1%, P_{MDD}=7.9%, r_{MDD}=55.4%, P_{BD}=0.4%, r_{BD}=59.4%, and n=750.

False-negative cases included individuals with schizophrenia but were not classified as schizophrenia cases. The proportion of true-positive to false-negative cases in the external validation

XSL•FO
RenderX

was 46:15 [15]. Hence, the false-negative cases can be calculated as follows:

$$Ntp=Ncrude-Nfp$$

$$Ntp/Nfn=p$$

where N_{tp} represents the cases predicted by the SZ classifier, N_{crude} represents the crude number of schizophrenia cases estimated by the SZ classifier, N_{fn} represents the false-negative cases, and p represents the proportion of true-positive cases to false-negative cases in the external validation, which was 46:15 [15].

Finally, the prevalence of schizophrenia in the population could be calculated by

$$Ppopulation=(Ncrude-Nfp+Nfn)/n$$

where $P_{population}$ represents the prevalence of schizophrenia in the population (Figure 1).

Figure 1. Flowchart illustrating the data analysis and refinement process for estimating the prevalence of schizophrenia in the general Japanese population. Data were collected from 750 participants during November 20 to 26, 2023. Participants' demographic information, health-related background, and physical, psychiatric, and social comorbidities were assessed using the schizophrenia classification model to initially estimate the crude number of schizophrenia cases and crude prevalence. The refinement process involved excluding false-positive cases and including false-negative cases, resulting in an actual estimation of schizophrenia prevalence in the general population of Japan.

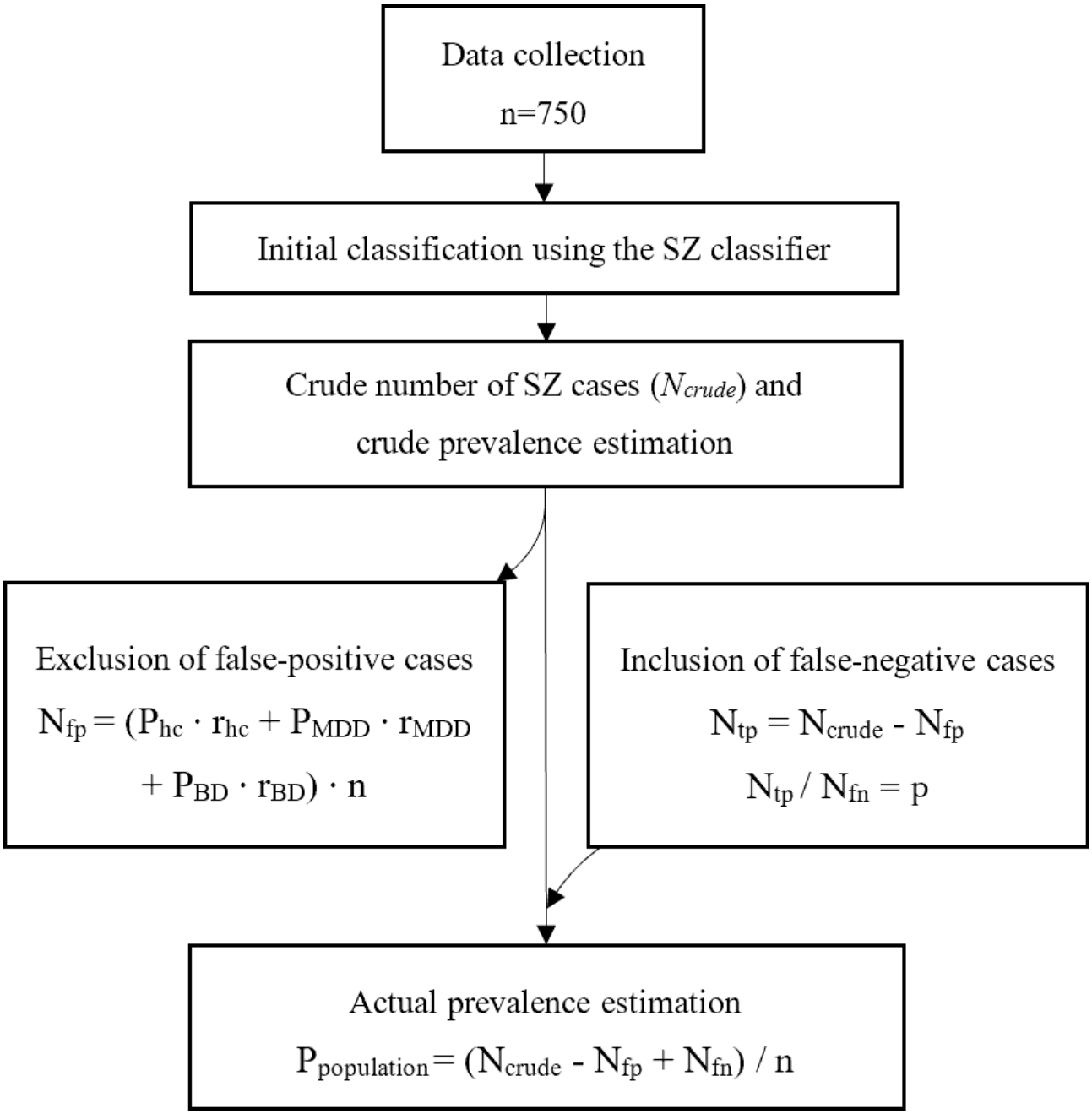

XSL•FO
RenderX

Statistical Analysis

The 95% CIs were calculated from 10,000 bootstrap replicates. Statistical analyses were performed using Python (version 3.8; Python Software Foundation), with a Jupyter Notebook (Project Jupyter) as a coding notebook.

Ethical Considerations

The study protocol was reviewed and approved by the institutional review board at Fujita Health University (HM23-253). All procedures were performed following the ethical standards of the Declaration of Helsinki. The internet survey agency respected the Act on the Protection of Personal Information in Japan. Informed consent was obtained at the initial web page of the survey, where the purpose of the survey and the instructions to the respondents were displayed. Only those who had read the survey's purpose and instructions and provided informed consent by clicking the agreement button could proceed with the survey. Participants for the external validation of the SZ classifier were diagnosed by experienced psychiatrists. The data were anonymous and deidentified before being provided to researchers. This was a voluntary survey. Participants received Rakuten Points, which can be used for Rakuten services such as web-based or offline shopping. No image data were collected so that there were no concerns about identifying individual participants.

Results

Of the 750 samples analyzed, 62 were classified as schizophrenia cases (N_{crude}). The crude prevalence of schizophrenia in the general population of Japan, as estimated using the SZ classifier, was 8.3% (95% CI 6.6%-10.1%; Table 1). Among these 62 cases, 53 were presumed to be false positives, and 3 were presumed to be false negatives. After adjustments, the actual prevalence of schizophrenia in the general population was estimated to be 1.6% (95% CI 0.7%-2.5%; Table 2).

Table . Results of crude estimates of schizophrenia cases by the schizophrenia classification model (SZ classifier), crude schizophrenia prevalence, false-positive estimates, and true-positive estimates.

	Cases, n	Prevalence, %
Crude estimates of predicted schizophrenia by the SZ classifier	62	8.3 (6.6-10.1)[a]
False positive: no mental disorders misclassified as schizophrenia	18	2.5
False positive: patients with MDD[b] misclassified as schizophrenia	33	4.4
False positive: patients with BD[c] misclassified as schizophrenia	2	0.2
True positive: patients with schizophrenia who were accurately classified as schizophrenia cases	9	1.2

[a]95% CI value.

[b]MDD: major depressive disorder.

[c]BD: bipolar disorder.

Table . Confusion matrix and adjusted prevalence of schizophrenia in the general population of Japan.

		Adjusted cases estimation	
		Schizophrenia	NonSchizophrenia
Predicted cases, n			
	Schizophrenia	9	53
	Nonschizophrenia	3	685
	Total	12	738
Prevalence, % (95%CI)		1.6 (0.7 - 2.5)	—[a]

[a]Not applicable.

Discussion

Principal Results

Our findings indicate that the estimated prevalence of schizophrenia in the general population of Japan using the SZ classifier was 1.6% (95% CI 0.7%-2.5%). This estimate is slightly higher than those in previously published figures [5,6]. This discrepancy could be attributed to several reasons, including the possibility that our SZ classifier estimates lifetime prevalence rather than point prevalence, a potentially more comprehensive disease classification methodology, and calculating the actual prevalence by referencing various prevalence data from diverse published sources.

XSL•FO
RenderX

To ensure the validity of the data used in our study, 3 main points were considered. The first point is the trustworthiness of the data source. The data were collected and managed by Rakuten Insight Panel, a large-scale web-based survey panel. The panel's quality has been maintained by company procedures and has been widely used by research institutions in Japan for population-based studies [16]. The second point is the appropriateness of the data used for model development. All feature variables were decided both through a literature review [17] and studies demonstrating that all these variables effectively differentiate between individuals with and those without schizophrenia [14,15,18]. Additionally, all the schizophrenia cases were defined by using a screening protocol instead of relying solely on the self-reported illness conditions. Schizophrenia cases had to meet the following four criteria: (1) currently having schizophrenia only, schizophrenia and migraine, schizophrenia and a sleep disorder, or schizophrenia, migraine, and a sleep disorder; (2) having experienced auditory hallucinations lasting more than 1 month; (3) not having used stimulants or other illegal drugs and not having been an alcoholic; and (4) having experienced the first auditory hallucination lasting more than 1 month at less than 60 years of age [18]. Third, the high performance of the developed model relies on high-quality data. The robustness and generalizability of the model were confirmed through good internal validation (AUC=0.86) [14] and external validation results (sensitivity=0.75) [15].

The SZ classifier predicts schizophrenia status using various feature variables. Each feature has been reported as being significantly associated with schizophrenia in previous studies [14,15,17,18]. By training the model with these features and applying data from the study participants, the probability of schizophrenia can be calculated. Although this probability may not be highly interpretable for diagnosis on an individual level, it reliably represents the proportion of specific cases within a target group. When the model demonstrates good generalizability, this proportion can be extended to estimate prevalence by considering the false positives and false negatives on a population level.

According to the feature importance ranking estimated by the model, social comorbidities, such as household income, type of employment, educational background, and type of occupation, play significant roles [14]. These factors may have been influenced by previous psychiatric conditions, which could have had long-term effects on the data used in this analysis. For instance, individuals diagnosed with schizophrenia may have had to stop attending school or leave their jobs, affecting their current income and occupation [22]. Consequently, the model might have predicted them as having schizophrenia even though they had recovered, which is consistent with the lifetime prevalence estimates. Several studies have shown that the prevalence of schizophrenia increases when it extends from one point to the lifetime of an individual [4,23].

To the best of our knowledge, this is the first prevalence study to apply an ANN model as a schizophrenia case classifier using web-based survey data from the general population of Japan. ANNs, which are computational models inspired by the inner processes of the human brain, are increasingly used in public health research [24,25]. Technically, ANNs are excellent at capturing complex and nonlinear relationships within large datasets. Hence, models constructed using ANNs perform better than classical methods based on linear equations [26,27]. Several studies have proposed ANN models or machine learning techniques to detect and classify schizophrenia [28,29]. These studies developed models using large amounts of clinical data (ie, magnetic resonance images, positron emission tomography scans, electroencephalography, and subjective patient posture interpretation). These studies have applied machine learning techniques to improve detection and diagnosis and aid in clinical decision-making. Bae et al [30] demonstrated the use of machine learning to detect signs of schizophrenia by analyzing social media texts, supporting the role of machine learning models in improving symptoms monitoring and potentially early detection of schizophrenia.

The use of ANN models to estimate prevalence has been reported in previous studies. For example, suicidal ideation risk prediction models were developed based on demographic, behavioral, and psychological data obtained from computer-based questionnaires. These models were then used to estimate the prevalence of suicidal ideation among Chinese college students. The models demonstrated a prediction accuracy of over 91.7%, and the AUC score increased to 0.96 [31]. Compared with previous studies, our SZ classifier has undergone external validation to demonstrate its generalizability [15]. This study supports the use of machine learning–based methods for prevalence estimation.

This study design may provide evidence to overcome biases encountered in traditional self-identifying surveys due to stigma, poor insight, and lack of access to health professionals experienced in conventional self-identifying surveys [32,33]. Stocking et al [34] reported that they first used a mathematical model to explore the correlation between several comorbidities and a specific illness in a small sample, and then applied this correlation to a larger population to identify previously underreported cases. They noted that this approach is effective for addressing the underreporting of various chronic conditions associated with social stigma, challenges in detection and treatment, or inconsistent treatment engagement (eg, mental disorders, obesity, hypertension, sleep disorders, etc). In this study, schizophrenia cases were estimated by incorporating general information, including demographics, health-related backgrounds, and physical, psychological, and social comorbidities, to calculate the likelihood of schizophrenia. This approach encompasses a broader range of patients in the general population than previous methodologies that relied on clinical records or typical schizophrenia symptoms. In contrast to a diagnosis that depends on the self-reporting of typical schizophrenia symptoms, which requires participants to openly acknowledge symptoms that may suggest schizophrenia, our methodology provides a more objective assessment. Patients with schizophrenia are often stigmatized and not widely accepted in society, leading to embarrassment and hesitation in disclosing their symptoms and avoiding negative perceptions. Stigmatizing attitudes related to mental illness have been reported to prevent individuals from seeking counseling [35] and harm the self-esteem of people with serious mental illnesses

XSL•FO
RenderX

[36]. Additionally, the wording of questionnaires that focus on negative aspects may cause respondents to feel negative and skip or conceal their actual symptoms, resulting in underreported results in self-identifying surveys [37,38]. By using general information, our SZ classifier mitigated these biases, potentially providing a more comprehensive estimate of patients with schizophrenia in Japan.

The SZ classifier has the advantage of potentially detecting patients with schizophrenia who do not perceive subjective symptoms. Hoffman et al [39] reported that a common reason why people with schizophrenia lack insight into their illness is that their ability to objectively assess their own pathological experiences is impaired. By using objective indicators such as demographics, health-related backgrounds, and physical, psychological, and social comorbidities, the SZ classifier showed a potential ability of detecting patients who do not perceive subjective symptoms and estimating the prevalence of schizophrenia in the population. This study may offer unique advantages over previous estimation methods based solely on medical data.

The higher prevalence estimated in our study suggests that schizophrenia might be more common than previously believed in the Japanese population. This finding confirmed our hypothesis regarding the underestimation of schizophrenia cases in the Japanese population. These findings have significant implications for public health management and resource allocation. Greater knowledge about the actual prevalence of schizophrenia can result in enhanced mental health services and increased support systems for managing this public health concern.

Limitations

Our study has several limitations. First, the actual prevalence of schizophrenia was calculated by referencing various prevalence data from diverse published sources, including the prevalence of individuals who did not have any mental disorder, the prevalence of those who had MDD, and the prevalence of those who had BD. Using data from different studies, which may have used different methodologies and population samples, causes uncertainty in the data, thereby affecting the estimated values. Second, although the data used in this study were collected from an extensive web-based survey that efficiently integrates large amounts of data, it may introduce a selection bias toward those with internet access. Those who are less likely to participate in web-based surveys, such as individuals without internet access, institutionalized patients, patients with severe comorbidities, older adults, and poor people, might be underrepresented. Third, the SZ classifier, designed to capture a broad range of patients, may include individuals with subclinical symptoms or comorbid conditions similar to schizophrenia. Despite efforts to rule out incorrectly classified cases, there is a chance that other mental disorders will not be sufficiently considered. This inclusion may lead to an overestimation of the true prevalence of schizophrenia. Fourth, for mental disorders comorbid with schizophrenia symptoms (eg, schizoaffective disorder, characterized by symptoms of both schizophrenia and mood disorders, etc), these cases were not included in the SZ classifier validation. Also, due to the complexity of mental disorders, the diagnosis of a disease without confirmation from a physician may seem arbitrary. Therefore, these results should be interpreted with caution.

Conclusions

This study applied a pretrained SZ classifier to estimate the prevalence of schizophrenia in a Japanese population. The estimated results are slightly higher than those of previous studies. Considering the shortcomings of focusing on a single type of mental disorder and the need to increase the sample size, this study shows promise for prevalence-estimation research. Potential future research directions may focus on developing models to classify other mental disorders, such as BD and MDD, as well as estimate prevalence. These findings are expected to support the versatility of pretrained models in epidemiological research.

Acknowledgments

This study was funded by the Ministry of Health, Labour and Welfare, Japan (JPMH21GC1018). The findings and conclusions of this article are the sole responsibility of the authors and do not represent the official views of the research funders. The authors attest that there was no use of generative artificial intelligence technology in the generation of text, figures, or other informational content of this manuscript.

Data Availability

The datasets analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: PC, YH, and AO
Data curation: MM and AO
Data interpretation: PC and YH
Data analysis: PC and YH
Investigation: MM and AO
Methodology: AO and YH
Supervision: AO

XSL•FO
RenderX

Writing—original draft: PC
Writing—review and editing: YH, KS, TK, YL, ST, NI, and AO

Conflicts of Interest

None declared.

Multimedia Appendix 1
Inclusion criteria, variable definition, and model validation results.
[DOCX File, 36 KB - formative_v9i1e66330_app1.docx]

References

1. Marder SR, Cannon TD. Schizophrenia. N Engl J Med 2019 Oct 31;381(18):1753-1761. [doi: 10.1056/NEJMra1808803] [Medline: 31665579]
2. Schizophrenia. World Health Organization. 2022 Jan 10. URL: https://www.who.int/news-room/fact-sheets/detail/schizophrenia [accessed 2024-05-17]
3. GBD 2019 Mental Disorders Collaborators. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet Psychiatry 2022 Feb;9(2):137-150. [doi: 10.1016/S2215-0366(21)00395-3] [Medline: 35026139]
4. Simeone JC, Ward AJ, Rotella P, Collins J, Windisch R. An evaluation of variation in published estimates of schizophrenia prevalence from 1990─2013: a systematic literature review. BMC Psychiatry 2015 Aug 12;15:193. [doi: 10.1186/s12888-015-0578-7] [Medline: 26263900]
5. Sado M, Inagaki A, Koreki A, et al. The cost of schizophrenia in Japan. Neuropsychiatr Dis Treat 2013;9:787-798. [doi: 10.2147/NDT.S41632] [Medline: 23785238]
6. Baba K, Guo W, Chen Y, Nosaka T, Kato T. Burden of schizophrenia among Japanese patients: a cross-sectional National Health and Wellness Survey. BMC Psychiatry 2022 Jun 18;22(1):410. [doi: 10.1186/s12888-022-04044-5] [Medline: 35717149]
7. Rodríguez-Rivas ME, Cangas AJ, Cariola LA, Varela JJ, Valdebenito S. Innovative technology-based interventions to reduce stigma toward people with mental illness: systematic review and meta-analysis. JMIR Serious Games 2022 May 30;10(2):e35099. [doi: 10.2196/35099] [Medline: 35635744]
8. Gray AJ. Stigma in psychiatry. J R Soc Med 2002 Feb;95(2):72-76. [doi: 10.1177/014107680209500205] [Medline: 11823548]
9. Kamaradova D, Latalova K, Prasko J, et al. Connection between self-stigma, adherence to treatment, and discontinuation of medication. Patient Prefer Adherence 2016;10:1289-1298. [doi: 10.2147/PPA.S99136] [Medline: 27524884]
10. Vrbova K, Prasko J, Holubova M, et al. Self-stigma and schizophrenia: a cross-sectional study. Neuropsychiatr Dis Treat 2016;12:3011-3020. [doi: 10.2147/NDT.S120298] [Medline: 27920538]
11. Konstantakopoulos G. Insight across mental disorders: a multifaceted metacognitive phenomenon. Psychiatriki 2019;30(1):13-16. [doi: 10.22365/jpsych.2019.301.13] [Medline: 31115349]
12. Ando S, Yamaguchi S, Aoki Y, Thornicroft G. Review of mental-health-related stigma in Japan. Psychiatry Clin Neurosci 2013 Nov;67(7):471-482. [doi: 10.1111/pcn.12086] [Medline: 24118217]
13. Goldner EM, Hsu L, Waraich P, Somers JM. Prevalence and incidence studies of schizophrenic disorders: a systematic review of the literature. Can J Psychiatry 2002 Nov;47(9):833-843. [doi: 10.1177/070674370204700904] [Medline: 12500753]
14. He Y, Matsunaga M, Li Y, et al. Classifying schizophrenia cases by artificial neural network using Japanese web-based survey data: case-control study. JMIR Form Res 2023 Nov 15;7:e50193. [doi: 10.2196/50193] [Medline: 37966882]
15. He Y, Sakuma K, Kishi T, et al. External validation of a machine learning model for schizophrenia classification. J Clin Med 2024 May 17;13(10):2970. [doi: 10.3390/jcm13102970] [Medline: 38792511]
16. Rakuten Insight. Rakuten Insight, Inc. URL: https://insight.rakuten.co.jp/en/aboutus.html [accessed 2024-05-30]
17. He Y, Tanaka A, Kishi T, et al. Recent findings on subjective well-being and physical, psychiatric, and social comorbidities in individuals with schizophrenia: a literature review. Neuropsychopharmacol Rep 2022 Dec;42(4):430-436. [doi: 10.1002/npr2.12286] [Medline: 35916310]
18. Matsunaga M, Li Y, He Y, et al. Physical, psychiatric, and social comorbidities of individuals with schizophrenia living in the community in Japan. Int J Environ Res Public Health 2023 Feb 28;20(5):4336. [doi: 10.3390/ijerph20054336] [Medline: 36901345]
19. Nishi D, Ishikawa H, Kawakami N. Prevalence of mental disorders and mental health service use in Japan. Psychiatry Clin Neurosci 2019 Aug;73(8):458-465. [doi: 10.1111/pcn.12894] [Medline: 31141260]
20. Hoshino E, Ohde S, Rahman M, Takahashi O, Fukui T, Deshpande GA. Variation in somatic symptoms by Patient Health Questionnaire-9 depression scores in a representative Japanese sample. BMC Public Health 2018 Dec 27;18(1):1406. [doi: 10.1186/s12889-018-6327-3] [Medline: 30587173]

XSL•FO
RenderX

21. Kato T. Bipolar disorder: from pathophysiology to treatment. Juntendo Iji Zasshi 2022;68(1):17-24. [doi: 10.14789/jmj.JMJ21-0026-R] [Medline: 38911011]
22. Hakulinen C, McGrath JJ, Timmerman A, et al. The association between early-onset schizophrenia with employment, income, education, and cohabitation status: nationwide study with 35 years of follow-up. Soc Psychiatry Psychiatr Epidemiol 2019 Nov;54(11):1343-1351. [doi: 10.1007/s00127-019-01756-0] [Medline: 31456027]
23. Bhugra D. The global prevalence of schizophrenia. PLoS Med 2005 May;2(5):e151. [doi: 10.1371/journal.pmed.0020151] [Medline: 15916460]
24. Lanillos P, Oliva D, Philippsen A, Yamashita Y, Nagai Y, Cheng G. A review on neural network models of schizophrenia and autism spectrum disorder. Neural Netw 2020 Feb;122:338-363. [doi: 10.1016/j.neunet.2019.10.014] [Medline: 31760370]
25. Razavi M, Ziyadidegan S, Mahmoudzadeh A, et al. Machine learning, deep learning, and data preprocessing techniques for detecting, predicting, and monitoring stress and stress-related mental disorders: scoping review. JMIR Ment Health 2024 Aug 21;11:e53714. [doi: 10.2196/53714] [Medline: 39167782]
26. Masegosa AR, Cabañas R, Langseth H, Nielsen TD, Salmerón A. Probabilistic models with deep neural networks. Entropy (Basel) 2021 Jan 18;23(1):117. [doi: 10.3390/e23010117] [Medline: 33477544]
27. He Y, Sun Q, Matsunaga M, Ota A. Can feature structure improve model's precision? A novel prediction method using artificial image and image identification. JAMIA Open 2024 Apr;7(1):ooae012. [doi: 10.1093/jamiaopen/ooae012] [Medline: 38348347]
28. Lai JW, Ang CKE, Acharya UR, Cheong KH. Schizophrenia: a survey of artificial intelligence techniques applied to detection and classification. Int J Environ Res Public Health 2021 Jun 5;18(11):6099. [doi: 10.3390/ijerph18116099] [Medline: 34198829]
29. Verma S, Goel T, Tanveer M, Ding W, Sharma R, Murugan R. Machine learning techniques for the schizophrenia diagnosis: a comprehensive review and future research directions. J Ambient Intell Human Comput 2023 May;14(5):4795-4807. [doi: 10.1007/s12652-023-04536-6]
30. Bae YJ, Shim M, Lee WH. Schizophrenia detection using machine learning approach from social media content. Sensors (Basel) 2021 Sep 3;21(17):5924. [doi: 10.3390/s21175924] [Medline: 34502815]
31. Liao S, Wang Y, Zhou X, et al. Prediction of suicidal ideation among Chinese college students based on radial basis function neural network. Front Public Health 2022;10:1042218. [doi: 10.3389/fpubh.2022.1042218] [Medline: 36530695]
32. Supina AL, Patten SB. Self-reported diagnoses of schizophrenia and psychotic disorders may be valuable for monitoring and surveillance. Can J Psychiatry 2006 Mar;51(4):256-259. [doi: 10.1177/070674370605100407] [Medline: 16629350]
33. Gould F, McGuire LS, Durand D, et al. Self-assessment in schizophrenia: accuracy of evaluation of cognition and everyday functioning. Neuropsychology 2015 Sep;29(5):675-682. [doi: 10.1037/neu0000175] [Medline: 25643212]
34. Stocking A, Duncan I, Huynh N. Estimating underdiagnosis of patients in chronically ill populations. N Am Actuar J 2023 Dec 18:1-10. [doi: 10.1080/10920277.2023.2281471]
35. Vogel DL, Wade NG, Hackler AH. Perceived public stigma and the willingness to seek counseling: the mediating roles of self-stigma and attitudes toward counseling. J Couns Psychol 2007;54(1):40-50. [doi: 10.1037/0022-0167.54.1.40]
36. Link BG, Struening EL, Neese-Todd S, Asmussen S, Phelan JC. Stigma as a barrier to recovery: the consequences of stigma for the self-esteem of people with mental illnesses. Psychiatr Serv 2001 Dec;52(12):1621-1626. [doi: 10.1176/appi.ps.52.12.1621] [Medline: 11726753]
37. Subramaniam M, Abdin E, Vaingankar JA, et al. Lifetime prevalence and correlates of schizophrenia and other psychotic disorders in Singapore. Front Psychiatry 2021;12:650674. [doi: 10.3389/fpsyt.2021.650674] [Medline: 33776823]
38. Karam EG, Sampson N, Itani L, et al. Under-reporting bipolar disorder in large-scale epidemiologic studies. J Affect Disord 2014 Apr;159:147-154. [doi: 10.1016/j.jad.2014.01.011] [Medline: 24679403]
39. Hoffman RE, McGlashan TH. Corticocortical connectivity, autonomous networks, and schizophrenia. Schizophr Bull 1994;20(2):257-261. [doi: 10.1093/schbul/20.2.257] [Medline: 8085129]

Abbreviations

ANN: artificial neural network
AUC: area under the receiver operating characteristic curve
BD: bipolar disorder
MDD: major depressive disorder
OCD: obsessive-compulsive disorder
SZ classifier: schizophrenia classification model

XSL•FO
RenderX

Edited by A Mavragani; submitted 10.09.24; peer-reviewed by D Abdel-Hady, SE Khamlichi; revised version received 05.01.25; accepted 07.01.25; published 29.01.25.

Please cite as:
Choomung P, He Y, Matsunaga M, Sakuma K, Kishi T, Li Y, Tanihara S, Iwata N, Ota A
Estimating the Prevalence of Schizophrenia in the General Population of Japan Using an Artificial Neural Network–Based Schizophrenia Classifier: Web-Based Cross-Sectional Survey
JMIR Form Res 2025;9:e66330
URL: https://formative.jmir.org/2025/1/e66330
doi:10.2196/66330

© Pichsinee Choomung, Yupeng He, Masaaki Matsunaga, Kenji Sakuma, Taro Kishi, Yuanying Li, Shinichi Tanihara, Nakao Iwata, Atsuhiko Ota. Originally published in JMIR Formative Research (https://formative.jmir.org), 29.1.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Measuring Self-Reported Well-Being of Physicians Using the Well-Being Thermometer: Cohort Study

Marios Adamou[1], MD, PhD; Sarah L Jones[2], PhD; Niki Kyriakidou[3], PhD; Andrew Mooney[4], PhD; Shriti Pattani[5]; Matthew Roycroft[6], MEd

[1]University of Huddersfield, Huddersfield, United Kingdom
[2]South West Yorkshire Partnership NHS Foundation Trust, Belle Isle Health Park, Wakefield, United Kingdom
[3]Leeds Beckett University, Leeds, United Kingdom
[4]Renal Unit, St James's Hospital, Leeds, United Kingdom
[5]London North West University Healthcare NHS Trust, London, United Kingdom
[6]Sheffield Teaching Hospitals NHS Foundation Trust, Sheffield, United Kingdom

Corresponding Author:
Sarah L Jones, PhD
South West Yorkshire Partnership NHS Foundation Trust, Belle Isle Health Park, Wakefield, United Kingdom

Abstract

Background: Advancements in medical science have focused largely on patient care, often overlooking the well-being of health care professionals (HCPs). This oversight has consequences; not only are HCPs prone to mental and physical health challenges, but the quality of patient care may also endure as a result. Such concerns are also exacerbated by unprecedented crises like the COVID-19 pandemic. Compared to other sectors, HCPs report high incidence of stress, depression, and suicide, among other challenging factors that have a significant negative impact on their well-being.

Objective: Given these substantial concerns, the development of a tool specifically designed to be used in clinical settings to measure the well-being of HCPs is essential.

Methods: A United Kingdom–based cross-sectional pilot study was carried out to measure self-reported well-being in a cohort of 148 physicians, using the newly developed well-being thermometer. The aim of the tool is to allow respondents to develop an individual sense of "well-being intelligence" thus supporting HCPs to have better insight and control over their well-being and allow insights into how to manage it. The tool consists of 5 well-being domains—health, thoughts, emotions, spiritual, and social. Each domain can be measured individually or combined to produce an overall well-being score.

Results: The tool demonstrated good internal consistency; the Cronbach α in this study was 0.84 for the total scale.

Conclusions: Results from this cohort demonstrated that the well-being thermometer can be used to gather intelligence of staff well-being. This is a promising new tool that will assist HCPs to recognize their own well-being needs and allow health care organizations to facilitate change in policies and practices to reflect a better understanding of staff well-being.

(JMIR Form Res 2025;9:e54158) doi:10.2196/54158

KEYWORDS

well-being; health care professionals; mental health; well-being thermometer; health care

Introduction

Medicine has developed greatly in relation to disease control and health interventions, yet it is questioned if health care professionals (HCPs) are fully aware of their own well-being, and the toll poor levels of well-being may have on their lives [1]. HCPs face unique challenges that could be damaging to their mental and physical health [2-5], and given the recent challenges of the COVID-19 pandemic [6,7], plus the current public health crisis, monitoring the well-being of our HCPs is vital for the mental health of the workforce.

Poor well-being is often reported in HCP cohorts [8-11], leading to serious health consequences and reduced quality of life for those affected [12]. In economic terms, higher staff turnover and sickness absence are the consequence [13], potentially resulting in poor quality of care for patients [14]. HCPs consistently report higher levels of sickness absence, job dissatisfaction, and stress compared to other work sectors [15-20]. A recent systematic review found moderate to high levels of stress, anxiety, depression, sleep disturbance, and burnout reported by frontline health care employees compared to workers in other sectors [21]. Moreover, a survey conducted in the United Kingdom (pre–COVID-19 pandemic) reported that, of public sector workers, those who work for the NHS

XSL•FO
RenderX

recorded the most stress [22]. There is also increasing morbidity in HCPs compared to the general population [23]. Reasons for these higher levels of stress are multifactorial, including lack of staff, increasing workload, patient expectation [23], emotional demands of the role [24], as well as others [25]. Research is supported by latest figures showing anxiety, stress, and depression are among the most reported reason for sickness absence in the NHS, accounting for 24.9% of absence in the year preceding September 2022 [26]. The issues pertaining to employee well-being have been acknowledged by the development of the "NHS workforce health and wellbeing framework" [27], although the outcomes of this are unclear at present. While there is extensive research around employee well-being [28], there are calls for deeper consideration of psychological needs of HCPs specifically [21,23,29]. The benefits of considering this not only help on an individual level, but also aid health care organizations with the economic cost of staff burnout [13].

Conceptualizing well-being is not straightforward. It is a complex construct, subject to much academic debate [30]. Current models of well-being tend to be grounded on concepts of mental illness or psychological functioning [31]. However, well-being is best described as multifaceted; measured by a range of subjective and objective concepts rather than one single notion [32]. Well-being is often used interchangeably with the term "mental health." For instance, according to the World Health Organization (WHO), "mental health" is "a state of well-being in which every individual realizes his or her own potential, can cope with the normal stresses of life, can work productively and fruitfully, and is able to make a contribution to her or his community" [33]. It is important to note that well-being is different from the terms "quality of life" or "health-related quality of life," terms that are primarily used to measure a person's perspective of their own life within a cultural context [34]. Furthermore "happiness" is not conducive to well-being. Happiness is often tied to external factors such as pursuit and fulfillment of life goals or life events, rather than a holistic concept such as well-being [35,36]. In terms of how we define well-being here, there is no commonly accepted definition. Subsequently there is no universally accepted approach to measure it. Instead, studies of well-being are often ambiguous in their approach and theoretical underpinnings [37]. A systematic review of self-report measures for assessing well-being found that while the 2 main theories referred to were Diener's model of subjective well-being and the WHO definition of health, authors were very rarely obvious about how theory had influenced the development of their tool or study. Further, argument suggests that the 2 most popular scales, The Warwick-Edinburgh Mental Wellbeing Scale and the WHO Five well-being index fail to capture the holistic nature of well-being [37].

Conceptualizing well-being is complex and often grounded in various subjective and objective factors. Despite various studies and tools aiming to measure well-being, none specifically target HCPs in a comprehensive manner. Adamou et al [31] developed a new theoretical framework of well-being specifically with HCPs in mind. The development of the well-being thermometer consisted of a three-step formation: (1) understanding the concept of well-being from existing literature and tools, (2) constructing a new framework of well-being, and (3) devising a tool to measure it. See Adamou et al [31] for further discussion of the conceptual development of the well-being thermometer. This study aims to pilot a new instrument, the well-being thermometer, specifically designed for HCPs. This tool aims to enhance individual "well-being intelligence," helping HCPs better manage their mental and physical health.

Methods

The Well-Being Thermometer

Developed by Adamou et al [31], this tool incorporates 5 well-being domains—health, thoughts, emotions, spiritual, and social. It aims to provide a comprehensive yet individualized snapshot of well-being, thereby allowing for targeted interventions.

Each of the domains comprises 5 items related to that domain. Each domain can be reviewed individually or collectively providing the individual with a score, allowing reflection of overall and domain-specific well-being. The health domain includes items which relate to the physical and eating health. The thoughts domain relates to mental health. Emotions domain measures the experience of emotions such as joy, satisfaction, and frustration. Spiritual health aligns with the connection with oneself and the meaning of life. The social domain relates to the experience of positive relationships and social networks. Total score can range from 0 to 25, with a score of 25 being the highest level of well-being. Example statements include "I tend to dwell on things more than I should" or "I feel there is a lot to enjoy in life." See Adamou et al [31] for further details.

Participants and Procedure

Doctors were recruited to participate in the well-being survey through verbal advertisements at 2 separate events—a diploma course in occupational medicine organized by the Royal Society of Public Health and a Regional (Yorkshire) Conference for Physicians organized by the Royal College of Physicians. At each event, information was provided about the survey's purpose and the automatic feedback participants would receive upon completion. The survey was administered using an online platform, and opportunity sampling was used to gather the cohort. Participants received an email containing a link to complete the survey online. By completing the survey, respondents were informed that they were consenting to have their data included in the research.

Ethical Considerations

This project was granted ethics approval in line with the Research Ethics Policy and Procedures at Leeds Beckett University (138110). Consent for participation was implicitly granted by the subjects through their active engagement with the survey, after being fully informed about the study's methodology and purpose via the online platform. Moreover, the study ensured the privacy and confidentiality of the participants by anonymizing or deidentifying all data used in the research. No compensation was provided to participants, aligning with the study's observational and noninterventional design.

XSL•FO
RenderX

Statistical Analysis

SPSS (version 29; IBM Corp) was used for statistical analyses. The Kolmogorov-Smirnov Test for Goodness of Fit determined data deviated significantly from normal distribution $P<.05$; therefore, nonparametric analysis was reported.

Results

Demographics

The cohort consisted of 148 physicians (without missing data), 68 (46%) recorded female sex, 78 (53%) reported male sex, and 2 (1%) respondents did not want to disclose gender. Age was recorded in category format, with age range of 40 - 44 years recorded most frequently. Ages ranged from 20 to 69 years. Two respondents chose not to disclose age (see Table 1). Respondents recorded level and speciality pertaining to their profession. See Table 1 for full demographic details.

Table 1 shows the demographic information of the 148 participants who participated in this cross-sectional United Kingdom–based study exploring the validity of the well-being thermometer using survey data. Details include self-reported gender, age, profession level, and speciality pertaining to their medical career.

XSL•FO
RenderX

Table . Respondent demographics (N=148).

Demographic information		Values, n (%)
Sex		
	Female	68 (45.9)
	Male	78 (52.7)
	Prefer not to say	2 (1.4)
	Missing	0 (0)
Age (years)		
	20 - 24	2 (1.4)
	25 - 29	4 (2.7)
	30 - 34	23 (15.5)
	35 - 39	16 (10.8)
	40 - 44	33 (22.3)
	45 - 49	21 (14.2)
	50 - 54	22 (14.9)
	55 - 59	13 (8.8)
	60 - 64	7 (4.7)
	65 - 69	5 (3.4)
	Prefer not to say	2 (1.4)
	Missing	0 (0)
Level		
	Foundation doctor	3 (2)
	Core trainee	5 (3.4)
	Higher specialty trainee	20 (13.5)
	SAS[b] or nontraining grade doctors	15 (10.1)
	Consultant or GP[a]	91 (61.5)
	Prefer not to say	5 (3.4)
	Other	9 (6.1)
	Missing	0 (0)
Specialty		
	Foundation programme	1 (7)
	Core medical training or internal medicine stage 1	3 (2)
	Acute internal medicine	12 (8.1)
	Cardiology	8 (5.4)
	Endocrinology and diabetes	4 (2.7)
	Gastroenterology	5 (3.4)
	General internal medicine	1 (0.7)
	Genitourinary medicine	11 (7.4)
	Geriatric medicine	33 (22.3)
	GP	22 (14.9)
	Infectious diseases	3 (2)
	Medical oncology	1 (0.7)
	Palliative medicine	5 (3.4)

XSL•FO
RenderX

Demographic information		Values, n (%)
	Rehabilitation medicine	1 (0.7)
	Renal medicine	2 (1.4)
	Respiratory medicine	14 (9.5)
	Rheumatology	3 (2)
	Sport and exercise medicine	1 (0.7)
	Stroke medicine	4 (2.7)
	Not applicable	3 (2)
	Prefer not to say	6 (4.1)
	Other	5 (3.4)
	Missing	0 (0)

[a]GP: general practitioner.

[b]SAS: specialty and specialist.

Well-Being Scores

A total of 148 participants in this cross-sectional United Kingdom–based study recorded an overall median score of 18 (IQR 14-22) on the well-being thermometer.

Table 2 shows the cross-sectional median and IQR score recorded by the 148 participants of the United Kingdom–based study of the well-being thermometer. Scores for individual well-being domains ranged between 3 and 4.

Table . Scores by domain.

Domain	Median (IQR)
Health	3 (2-4)
Social	3.5 (3-5)
Thoughts	4 (2-5)
Emotions	4 (3-5)
Spiritual	4 (3-5)
All domains	18 (14-22)

There was no significant effect of age, profession level, or speciality on well-being scores, both overall and domain specific.

Gender-Based Analysis

For the United Kingdom–based cross-sectional study exploring the well-being thermometer, organized by gender (N=148), the median score for men (n=78) was 20 (IQR 16-23) compared to 16 (IQR 12-20.5) for women (n=68).

Table 3 shows the median and IQR scores recorded by the sample (N=148) on individual domains (health, social, thoughts, emotions, and spiritual) of the well-being thermometer, organized by gender. The data are derived from a cross-sectional United Kingdom–based study exploring the validity of the well-being thermometer. Scores ranged between 4 and 5 for men and 3 to 4 for women.

Men reported significantly higher levels of well-being than women overall (U=1593; P=.002). In terms of specific well-being domains, men reported higher levels of well-being on health (U=1979.5; P=.015), thoughts (U=1948.5; P=.019), emotions (U=1971; P=.013), and spiritual (U=1914; P=.014) domains, compared to women. There was also a trend for men to score higher on levels of social well-being (U=2109.5; P=.05).

XSL•FO
RenderX

Table . Domain scores by gender.

Domain	Male (n=78) Median (IQR)	Female (n=68) Median (IQR)
Health	4 (3-5)	3 (1.5-4)
Social	4 (3-5)	3 (3-4)
Thoughts	4 (2-5)	3 (2-5)
Emotions	4 (3-5)	3 (2-4)
Spiritual	5 (4-5)	4 (3-5)
All domains	20 (7)	16 (8.5)

Correlation

Table 4 demonstrates that Spearman rho identified positive relationships between scores on individual domains on the well-being thermometer in this cross-sectional United Kingdom–based study of HCPs (N=148). Suggesting that higher scores on 1 domain of well-being was reflected in other domains of well-being.

Table . Spearman rho.

		Total for health domain, n	Total for social domain, n	Total for thoughts domain, n	Total for emotions domain, n	Total for spiritual domain, n
Total for health domain						
	Correlation coefficient	—[a]	0.399[b]	0.376[b]	0.468[b]	0.400[b]
	Significance (2-tailed)	—	.00	.00	.00	.00
	n	—	144	143	144	142
Total for social domain						
	Correlation coefficient	0.399[b]	—	0.492[b]	0.453[b]	0.453[b]
	Significance (2-tailed)	.000	—	.000	.195	.000
	n	144	—	142	144	141
Total for thoughts domain						
	Correlation coefficient	0.376[b]	0.492[b]	—	0.640[b]	0.537[b]
	Significance (2-tailed)	.000	.000	—	.000	.000
	n	143	142	—	143	142
Total for emotions domain						
	Correlation coefficient	0.468[b]	0.453[b]	0.640[b]	—	0.534[b]
	Significance (2-tailed)	.000	.000	.000	—	.000
	n	144	144	143	—	143
Total for spiritual domain						
	Correlation coefficient	0.400[b]	0.453[b]	0.537[b]	0.534[b]	—
	Significance (2-tailed)	.000	.000	.000	.000	—
	n	142	141	142	143	—

[a]Not applicable.

[b]Correlation is significant at the .01 level (2-tailed).

XSL•FO
RenderX

Cronbach α Analysis

Cronbach α values of 0.7 are considered high levels of internal consistency (DeVillis [38]; Kline [39]). Values above 0.5 are acceptable (Bowling [40]; Schmitt [41]). The scale had a high level of internal consistency, as determined by a Cronbach α of 0.872. Value would improve to α=0.878 if question 7 (I drink more alcohol than would be considered healthy) was removed.

Discussion

Principal Results

The aim of this study was to pilot a new instrument, the well-being thermometer. The tool was specifically designed for use with HCPs and was piloted here with a sample of physicians. In this cross-sectional study the well-being thermometer demonstrated a good level of internal consistency. It was evident that higher scores in 1 well-being domain correlated with higher scores on other well-being domains. The factors of age, speciality, and professional level had no significant effect on well-being. Results suggest that the well-being thermometer has the potential to be a useful and informative tool, both within clinical settings and on an individual level.

Interpretations

The purpose of the well-being thermometer is to enhance individual "well-being intelligence." To aid HCPs to better manage their mental and physical health, and to measure well-being in a more holistic framework than popular scales have allowed [31]. The well-being thermometer has shown to be a valid tool for measuring well-being in HCPs. Subsequently, with better understanding of well-being, services can offer better interventions, protection, and help to their workforce, and use this information to influence policy. The well-being thermometer also allows HCPs to identify elements of their individual well-being which may need attention.

Interestingly, results from this study demonstrated that there was a difference in well-being levels between the sexes, with men reporting higher levels of well-being than women overall, but also on the health, thoughts, emotions, and spiritual domains. This variable was considered important to explore here, as men have often reported higher levels of subjective well-being than women in numerous previous studies [21,29,42-44]. With best evidence derived from a large-scale study of 6397 HCPs, where men reported significantly greater level of overall well-being than women [45]. The findings from this study follow this trend.

Importantly, the well-being thermometer can be used to gather intelligence of staff well-being to facilitate change in policies and practices across health care organizations [31]. While better well-being is a valuable goal in its own right, HCPs should be a priority target because their roles require frontline responsibilities toward the general public and vulnerable populations. We have seen advancements in medical science largely focused on patient care, often overlooking the well-being of HCPs. This oversight has consequences—not only are HCPs prone to mental and physical health challenges, but the quality of patient care also endures as a result. Such concerns are exacerbated by unprecedented crises like the COVID-19 pandemic. HCPs are pressured with dealing with public health crisis in real time, in which difficult moral decision-making is associated with significant stress, lack of control, and feelings of fear [46]. Troublingly, the prevalence of burnout, depression, and suicide is high for this group [47-49]. Compromised well-being among HCPs leads to medical errors, reduced patient safety, high rates of staff turnover, increased absence due to sickness, and diminished patient care [12,14,50,51]. Thus, health care organizations have a responsibility to support and protect staff well-being for both patient and staff safety [52] and the well-being thermometer has shown it can be a useful tool in supporting this objective.

Limitations

Further work pertaining to the well-being thermometer should be conducted with additional demographic information such as race and ethnicity, as this information was not captured here, and is a limitation of the study. A focus on threshold analysis is also necessary to gain further insight. Also, future work with larger samples is recommended.

Conclusions

Conceptualizing well-being remains complex and often grounded in various subjective and objective factors, and despite various studies and tools aiming to measure well-being, none specifically target HCPs in a comprehensive manner. The aim of this study was to pilot a new instrument, the well-being thermometer, specifically designed for HCPs. This study demonstrated that the well-being thermometer can be used to gather intelligence of staff well-being. This information is required to facilitate much needed change in policies and practices across health care organizations. A specific focus on staff well-being will benefit not only HCPs, but also those who trust in our organizations to provide safe and efficient health care services.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

References

1. Shanafelt TD, Sloan JA, Habermann TM. The well-being of physicians. Am J Med 2003 Apr 15;114(6):513-519. [doi: 10.1016/s0002-9343(03)00117-7] [Medline: 12727590]

XSL•FO
RenderX

2. Lomas T, Medina JC, Ivtzan I, Rupprecht S, Eiroa-Orosa FJ. A systematic review of the impact of mindfulness on the well-being of healthcare professionals. J Clin Psychol 2018 Mar;74(3):319-355. [doi: 10.1002/jclp.22515] [Medline: 28752554]
3. Gao YQ, Pan BC, Sun W, Wu H, Wang JN, Wang L. Anxiety symptoms among Chinese nurses and the associated factors: a cross sectional study. BMC Psychiatry 2012 Sep 14;12:141. [doi: 10.1186/1471-244X-12-141] [Medline: 22978466]
4. Khamisa N, Oldenburg B, Peltzer K, Ilic D. Work related stress, burnout, job satisfaction and general health of nurses. Int J Environ Res Public Health 2015 Jan 12;12(1):652-666. [doi: 10.3390/ijerph120100652] [Medline: 25588157]
5. Givens JL, Tjia J. Depressed medical students' use of mental health services and barriers to use. Acad Med 2002 Sep;77(9):918-921. [doi: 10.1097/00001888-200209000-00024] [Medline: 12228091]
6. Serafini G, Parmigiani B, Amerio A, Aguglia A, Sher L, Amore M. The psychological impact of COVID-19 on the mental health in the general population. QJM 2020 Jun 22;113(8):531-537. [doi: 10.1093/qjmed/hcaa201] [Medline: 32569360]
7. Siddiqui I, Aurelio M, Gupta A, Blythe J, Khanji MY. COVID-19: causes of anxiety and wellbeing support needs of healthcare professionals in the UK: a cross-sectional survey. Clin Med (Lond) 2021 Jan;21(1):66-72. [doi: 10.7861/clinmed.2020-0502] [Medline: 33479070]
8. Johnson J, Hall LH, Berzins K, Baker J, Melling K, Thompson C. Mental healthcare staff well-being and burnout: a narrative review of trends, causes, implications, and recommendations for future interventions. Int J Ment Health Nurs 2018 Feb;27(1):20-32. [doi: 10.1111/inm.12416] [Medline: 29243348]
9. O'Connor DB, Hall LH, Johnson J. Job strain, burnout, wellbeing and patient safety in healthcare professionals. In: Montgomery A, van der Doef M, Panagopoulou E, Leiter MP, editors. Connecting Healthcare Worker Well-Being, Patient Safety and Organisational Change: The Triple Challenge: Springer International Publishing; 2020:11-23. [doi: 10.1007/978-3-030-60998-6_2]
10. Shanafelt TD, Balch CM, Bechamps GJ, et al. Burnout and career satisfaction among American surgeons. Ann Surg 2009 Sep;250(3):463-471. [doi: 10.1097/SLA.0b013e3181ac4dfd] [Medline: 19730177]
11. Elmore LC, Jeffe DB, Jin L, Awad MM, Turnbull IR. National Survey of Burnout among US general surgery residents. J Am Coll Surg 2016 Sep;223(3):440-451. [doi: 10.1016/j.jamcollsurg.2016.05.014] [Medline: 27238875]
12. Prudenzi A, D Graham C, Flaxman PE, O'Connor DB. Wellbeing, burnout, and safe practice among healthcare professionals: predictive influences of mindfulness, values, and self-compassion. Psychol Health Med 2022 Jun;27(5):1130-1143. [doi: 10.1080/13548506.2021.1898651] [Medline: 33856236]
13. Toppinen-Tanner S, Ojajärvi A, Väänänen A, Kalimo R, Jäppinen P. Burnout as a predictor of medically certified sick-leave absences and their diagnosed causes. Behav Med 2005;31(1):18-27. [doi: 10.3200/BMED.31.1.18-32] [Medline: 16078523]
14. Hall LH, Johnson J, Watt I, Tsipa A, O'Connor DB. Healthcare staff wellbeing, burnout, and patient safety: a systematic review. PLoS One 2016;11(7):e0159015. [doi: 10.1371/journal.pone.0159015] [Medline: 27391946]
15. Lee DJ, Fleming LE, LeBlanc WG, et al. Health status and risk indicator trends of the aging US health care workforce. J Occup Environ Med 2012 Apr;54(4):497-503. [doi: 10.1097/JOM.0b013e318247a379] [Medline: 22446575]
16. Edwards D, Burnard P. A systematic review of stress and stress management interventions for mental health nurses. J Adv Nurs 2003 Apr;42(2):169-200. [doi: 10.1046/j.1365-2648.2003.02600.x] [Medline: 12670386]
17. Raiger J. Applying a cultural lens to the concept of burnout. J Transcult Nurs 2005 Jan;16(1):71-76. [doi: 10.1177/1043659604270980] [Medline: 15608102]
18. Petterson IL, Arnetz BB. Psychosocial stressors and well-being in health care workers. The impact of an intervention program. Soc Sci Med 1998 Dec;47(11):1763-1772. [doi: 10.1016/s0277-9536(98)00245-7] [Medline: 9877346]
19. Ruotsalainen J, Serra C, Marine A, Verbeek J. Systematic review of interventions for reducing occupational stress in health care workers. Scand J Work Environ Health 2008 Jun;34(3):169-178. [doi: 10.5271/sjweh.1240] [Medline: 18728906]
20. Brooks SK, Gerada C, Chalder T. Review of literature on the mental health of doctors: are specialist services needed? J Ment Health 2011 Apr;20(2):146-156. [doi: 10.3109/09638237.2010.541300] [Medline: 21275504]
21. Danet Danet A. Psychological impact of COVID-19 pandemic in Western frontline healthcare professionals. A systematic review. Med Clin (Barc) 2021 May 7;156(9):449-458. [doi: 10.1016/j.medcli.2020.11.009] [Medline: 33478809]
22. Dudman J, Isaac A, Johnson S. Revealed: how the stress of working in public services is taking its toll on staff. The Guardian. 2015. URL: https://www.theguardian.com/society/2015/jun/10/stress-working-public-services-survey [accessed 2024-12-27]
23. Mohanty A, Kabi A, Mohanty AP. Health problems in healthcare workers: a review. J Family Med Prim Care 2019 Aug;8(8):2568-2572. [doi: 10.4103/jfmpc.jfmpc_431_19] [Medline: 31548933]
24. Tyssen R, Vaglum P, Gronvold NT, Ekeberg O. The impact of job stress and working conditions on mental health problems among junior house officers. A nationwide Norwegian prospective cohort study. Med Educ (Chicago Ill) 2000 May;34(5):374-384. [doi: 10.1046/j.1365-2923.2000.00540.x]
25. Greenberg N. Mental health of health-care workers in the COVID-19 era. Nat Rev Nephrol 2020 Aug;16(8):425-426. [doi: 10.1038/s41581-020-0314-5] [Medline: 32561871]
26. NHS sickness absence rates. NHS England Digital. 2022. URL: https://digital.nhs.uk/data-and-information/publications/statistical/nhs-sickness-absence-rates?__cf_chl_tk=GyQ_jyQo3bml1ydIfymQem8x2Tw06jHtOkp6RsAg9cQ-1734726614-1.0.1.1-Znz.bX9zDciYi0lG5RKljML9k7KYoFv1pCFRbJER63o [accessed 2024-12-21]

XSL•FO
RenderX

27. NHS health and wellbeing framework. NHS England. 2021. URL: https://www.england.nhs.uk/publication/nhs-health-and-wellbeing-framework/ [accessed 2024-12-21]
28. Adler A, Seligman MEP. Using wellbeing for public policy: theory, measurement, and recommendations. Intnl J Wellbeing 2016;6(1):1-35. [doi: 10.5502/ijw.v6i1.429]
29. Luo M, Guo L, Yu M, Jiang W, Wang H. The psychological and mental impact of coronavirus disease 2019 (COVID-19) on medical staff and general public - a systematic review and meta-analysis. Psychiatry Res 2020 Sep;291:113190. [doi: 10.1016/j.psychres.2020.113190] [Medline: 32563745]
30. Huppert FA. The state of wellbeing science: concepts, measures, interventions, and policies. In: Huppert FA, Cooper CL, editors. Interventions and Policies to Enhance Wellbeing: A Complete Reference Guide: John Wiley & Sons Inc; 2014:1-50. [doi: 10.1002/9781118539415]
31. Adamou M, Goddard A, Kyriakidou N, et al. The wellbeing thermometer: a novel framework for measuring wellbeing. PSYCH 2020;11(10):1471-1480. [doi: 10.4236/psych.2020.1110093]
32. Forgeard MJC, Jayawickreme E, Kern ML, Seligman MEP. Doing the right thing: measuring well-being for public policy. Intnl J Wellbeing 2011;1(1). [doi: 10.5502/ijw.v1i1.15]
33. Promoting Mental Health: Concepts, Emerging Evidence, Practice: World Health Organization; 2004.
34. Hewitt R. Measuring well-being in higher education. HEPI. 2019. URL: https://www.hepi.ac.uk/wp-content/uploads/2019/05/Policy-Note-13-Paper-May-2019-Measuring-well-being-in-higher-education-8-Pages-5.pdf [accessed 2024-12-27]
35. Easterlin RA. Explaining happiness. Proc Natl Acad Sci 2003;100(19):11176-11183. [doi: 10.1073/pnas.1633144100]
36. Jacobson B. What is happiness? J Soc Exist Anal 2007;18(1):39-50.
37. Linton MJ, Dieppe P, Medina-Lara A. Review of 99 self-report measures for assessing well-being in adults: exploring dimensions of well-being and developments over time. BMJ Open 2016 Jul 7;6(7):e010641. [doi: 10.1136/bmjopen-2015-010641] [Medline: 27388349]
38. DeVellis RF. Scale Development: Theory and Applications, 2nd edition: Sage Publications, Inc; 2003.
39. Kline T. Psychological Testing: A Practical Approach to Design and Evaluation: Sage Publications, Inc; 2005.
40. Bowling A. Research Methods in Health: Investigating Health and Health Services, 2nd edition: Open University Press, McGraw-Hill Education; 2002.
41. Schmitt N. Uses and abuses of coefficient alpha. Psychol Assess 1996;8(4):350-353. [doi: 10.1037/1040-3590.8.4.350]
42. Lombardo M, Aulisa G, Padua E, et al. Gender differences in taste and foods habits. NFS 2019 Jul 18;50(1):229-239. [doi: 10.1108/NFS-04-2019-0132]
43. Oncini F, Guetto R. Cultural capital and gender differences in health behaviours: a study on eating, smoking and drinking patterns. Health Sociol Rev 2018 Jan 2;27(1):15-30. [doi: 10.1080/14461242.2017.1321493]
44. Roothman B, Kirsten DK, Wissing MP. Gender differences in aspects of psychological well-being. S Afr J Psychol 2003 Nov;33(4):212-218. [doi: 10.1177/008124630303300403]
45. Dyer NL, Adan F, Barnett T, Dusek JA. Assessment of healthcare professionals' wellbeing during a peak of the COVID-19 pandemic in a healthcare system in Ohio. Glob Adv Health Med 2022;11:2164957X221089258. [doi: 10.1177/2164957X221089258] [Medline: 35634474]
46. Wald HS. Optimizing resilience and wellbeing for healthcare professions trainees and healthcare professionals during public health crises - practical tips for an "integrative resilience" approach. Med Teach 2020 Jul;42(7):744-755. [doi: 10.1080/0142159X.2020.1768230] [Medline: 32449867]
47. Dyrbye LN, Shanafelt TD. Physician burnout: a potential threat to successful health care reform. JAMA 2011 May 18;305(19):2009-2010. [doi: 10.1001/jama.2011.652] [Medline: 21586718]
48. Mata DA, Ramos MA, Bansal N, et al. Prevalence of depression and depressive symptoms among resident physicians: a systematic review and meta-analysis. JAMA 2015 Dec 8;314(22):2373-2383. [doi: 10.1001/jama.2015.15845] [Medline: 26647259]
49. Schernhammer E. Taking their own lives -- the high rate of physician suicide. N Engl J Med 2005 Jun 16;352(24):2473-2476. [doi: 10.1056/NEJMp058014] [Medline: 15958803]
50. Velez K. Relationship between moral distress, workplace bullying, staff burnout and turnover at an academic spinal cord injury/disorder (SCI/D) rehabilitation facility: a causality study. Arch Phys Med Rehabil 2020 Nov;101(11):e54. [doi: 10.1016/j.apmr.2020.09.161]
51. Wilson W, Raj JP, Narayan G, Ghiya M, Murty S, Joseph B. Quantifying burnout among emergency medicine professionals. J Emerg Trauma Shock 2017;10(4):199-204. [doi: 10.4103/JETS.JETS_36_17] [Medline: 29097859]
52. Ornell F, Halpern SC, Kessler FHP, Narvaez JCDM. The impact of the COVID-19 pandemic on the mental health of healthcare professionals. Cad Saude Publica 2020;36(4):e00063520. [doi: 10.1590/0102-311X00063520] [Medline: 32374807]

Abbreviations

HCP: health care professional
NHS: National Health Service

XSL•FO
RenderX

WHO: World Health Organization

Edited by A Mavragani; submitted 31.10.23; peer-reviewed by K Kamposioras, T Gladman; revised version received 04.11.24; accepted 06.11.24; published 09.01.25.

Please cite as:
Adamou M, Jones SL, Kyriakidou N, Mooney A, Pattani S, Roycroft M
Measuring Self-Reported Well-Being of Physicians Using the Well-Being Thermometer: Cohort Study
JMIR Form Res 2025;9:e54158
URL: https://formative.jmir.org/2025/1/e54158
doi:10.2196/54158

© Marios Adamou, Sarah L Jones, Niki Kyriakidou, Andrew Mooney, Shriti Pattani, Matthew Roycroft. Originally published in JMIR Formative Research (https://formative.jmir.org), 9.1.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Recruitment Challenges and Strategies in a Technology-Based Intervention for Dementia Caregivers: Descriptive Study

Eunjung Ko[1], RN, PhD; Ye Gao[2], PhD; Peng Wang[3], MS; Lahiru Wijayasingha[3], MS; Kathy D Wright[4], RN, PhD; Kristina C Gordon[5], PhD; Hongning Wang[3], PhD; John A Stankovic[3], PhD; Karen M Rose[4], RN, PhD

[1]Rory Meyers College of Nursing, New York University, New York, NY, United States
[2]Computer Science, William & Mary, Williamsburg, VA, United States
[3]School of Engineering and Applied Science, University of Virginia, Charlottesville, VA, United States
[4]College of Nursing, The Ohio State University, Columbus, OH, United States
[5]College of Education, Health, and Human Sciences, The University of Tennessee-Knoxville, Knoxville, TN, United States

Corresponding Author:
Eunjung Ko, RN, PhD
Rory Meyers College of Nursing, New York University, New York, NY, United States

Abstract

Background: Researchers have encountered challenges in recruiting unpaid caregivers of people living with Alzheimer disease and related dementias for intervention studies. However, little is known about the reasons for nonparticipation in in-home smart health interventions in community-based settings.

Objective: This study aimed to (1) assess recruitment rates in a smart health technology intervention for caregivers of people living with Alzheimer disease and related dementias and reasons for nonparticipation among them and (2) discuss lessons learned from recruitment challenges and strategies to improve recruitment.

Methods: The smart health intervention was a 4-month, single-arm trial designed to evaluate an in-home, technology-based intervention that monitors stressful moments for caregiving dyads through acoustic signals and to provide the caregivers with real-time stress management strategies. The recruitment involved two main methods: on-site engagement by a recruiter from a memory clinic and social media advertising. Caregivers were screened for eligibility by phone between January 2021 and September 2023. The recruitment rates, reasons for nonparticipation, and participant demographics were analyzed using descriptive statistics.

Results: Of 201 caregivers contacted, 11 were enrolled in this study. Eighty-two caregivers did not return the screening call, and others did not participate due to privacy concerns (n=30), lack of interest (n=29), and burdensome study procedures (n=26). Our recruitment strategies included addressing privacy concerns, visualizing collected data through a dashboard, boosting social media presence, increasing the recruitment budget, updating advertisements, and preparing and deploying additional study devices.

Conclusions: This study highlighted barriers to participation in the smart health intervention. Despite several recruitment strategies, enrollment rates remained below expectations. These findings underscore the need for future research to explore alternative methods for increasing the recruitment of informal dementia caregivers in technology-based intervention studies.

Trial Registration: ClinicalTrials.gov NCT04536701; https://clinicaltrials.gov/study/NCT04536701

International Registered Report Identifier (IRRID): RR2-10.1111/jan.14714

(JMIR Form Res 2025;9:e59291) doi:10.2196/59291

KEYWORDS

recruitment challenges and strategies; technology-based intervention; dementia caregivers; dementia; mobile phone; Alzheimer disease; smart health

Introduction

The number of people living with Alzheimer disease and related dementias (ADRD) in the United States is increasing, with the figure expected to reach 13.8 million by 2060 [1]. Alzheimer disease, the primary neurodegenerative condition contributing to ADRD, necessitates ongoing and escalating levels of support over time for those affected [1]. Most people living with ADRD receive unpaid assistance from family members, friends, or neighbors. In 2023, the number of unpaid caregivers surpassed 11 million [1]. Managing the wide range of caregiving responsibilities and coordinating dementia care can exacerbate caregivers' stress, burden, anxiety, and depression [1,2].

Various nonpharmacological interventions have emerged to mitigate the negative aspects of caregiving, aimed at supporting caregivers [3]. These interventions include acceptance and

XSL•FO
RenderX

commitment therapy, education, counseling, psychotherapy, and support groups. Although the effects of these interventions vary depending on their contexts, the interventions have been generally effective in alleviating caregivers' negative emotions and improving their quality of life [1,3].

Over the past few decades, technological advancements and the recent pandemic challenges prompted a transition from in-person to online delivery of interventions. This shift has made interventions more accessible by accommodating caregivers' limited time and geographical locations and alleviating care demand [1,4,5]. The delivery methods of interventions range from the basic tools, such as telephone calls, video conferences, and web page visits, to more advanced technologies such as wearable devices, video or image sensors, or intelligent audio systems [4,6]. These technologies can enhance home safety and well-being and reduce health care costs for people living with dementia and their caregivers [6].

Despite the potential of these novel interventions, their deployment still faces challenges among caregivers of people living with ADRD. Barriers include limited experience with technology, lack of awareness, technical issues, time constraints, privacy concerns, and mistrust [6-8]. Notably, Wen et al [7] found that online-based interventions are more likely to be accepted by younger family caregivers, contrasting with the predominantly middle-aged and older caregiver population [1,9]. Further, the American Association of Retired Persons report revealed that only half of the caregivers across all age groups and just 20% of those older than 65 years of age felt comfortable using in-home technologies (eg, Google Assistant or Alexa; Amazon.com, Inc.) [10]. Given the growing proportion of caregivers aged 60 years and older (28% in 2015 and 35.4% in 2022) [11], participation in technology-based interventions, specifically through those requiring advanced technological devices, remain limited.

Recent studies have explored recruitment challenges and strategies in interventions targeting caregivers of people living with ADRD [7,12-15]. However, there is still limited research specifically focused on these issues within the context of technology-based interventions. Understanding and addressing these challenges are pivotal for effectively engaging caregivers of people living with ADRD in such interventions, which is the primary focus of our study.

In this descriptive study, we aimed to explore the recruitment challenges encountered in an in-home, technology-based intervention designed to reduce stress among caregivers of people living with ADRD and to improve their dyadic relationships. We identified the overall recruitment rate and the reasons for nonparticipation from the caregivers' perspective and examined the demographic characteristics of the caregivers enrolled in this study. In addition, we discussed the recruitment challenges and strategies we used to address them during the intervention.

Methods

Study Design of the Smart Health Intervention

The smart health intervention was a 4-month, single-arm trial designed to develop an in-home, technology-based intervention and assess its feasibility for caregivers of people living with ADRD [16,17]. Caregivers received this study's devices (ie, a laptop, a microphone connected to the computer, a router, and a smartphone). They placed them in an area of their home where they frequently interacted with their care recipients. We captured acoustic signals from the caregiving dyads through the microphone and used a deep machine-learning approach to identify stressful situations [16]. The system then sent a real-time stress management tip to the caregivers' study smartphone via the ecological momentary assessment (EMA) system [16,17]. Examples of tips sent to caregivers included practicing body scan and deep-breathing exercises, taking a time out, or engaging in enjoyable activities. The messages encouraged caregivers to follow the tip and to report its usefulness. Caregivers were also asked to answer daily and weekly questions about their physical and emotional health status and the overall usefulness of the stress management tips [17]. We provided US $50 compensation to each caregiver and care recipient three times—1 month after starting the intervention, halfway through, and at the end of the intervention.

Recruitment

Recruitment took place between January 2021 and September 2023 using convenience sampling. Two main recruitment methods were used to recruit eligible participants: (1) on-site engagement at a memory clinic within a Midwestern academic medical center and (2) advertisements on social media platforms. At the memory clinic, a recruiter distributed study brochures to individuals who visited the clinic and might be eligible for this intervention. Once the caregivers expressed interest in this study, the recruiter shared their contact information with this study's personnel (EK and KMR). This study's personnel contacted the caregivers to explain this study's procedures in detail and confirm their eligibility.

We also advertised the smart health intervention via social media platforms, such as Facebook (Meta) and Instagram (Instagram from Meta), to recruit caregivers across the United States. We allocated between US $100 and US $300 to recruit participants through social media platforms for 1 week of recruitment postings. The uploaded posts included this study's title, primary text, target age and gender, study locations, and a link to a website containing this study's detailed information. Individuals who saw this study's advertisements and were interested in this intervention contacted the principal investigator (KMR) by phone or email. In addition to these two methods, individuals also contacted the principal investigator after finding the information about our study on websites listing clinical trials, such as ClinicalTrials.gov [18].

Participant Eligibility

We recruited participants who met the following inclusion criteria: (1) aged 21 years or older, (2) residing with and caring for older adults living with ADRD, (3) not receiving payment

XSL•FO
RenderX

for providing care, (4) having well-functioning home Wi-Fi, (5) receiving a score above 3 on the Revised Memory and Behavior Problems Checklist, (6) being fluent in English, and (7) agreeing to participate [17]. During the screening call, those who met the eligibility criteria provided verbal consent and signed the electronic consent form. Given that this intervention involved monitoring acoustic signals while caregivers communicated with their care recipients, care recipients were also required to sign the consent form. If care recipients could not consent, their caregivers provided proxy consent.

Data Collection

Overview

We manually recorded the participation status and the reasons for nonparticipation of each individual contacted in REDCap (Research Electronic Data Capture; Vanderbilt University), a Health Insurance Portability and Accountability Act compliant, web-based data collection platform that allows researchers to access and manage data [19]. We also gathered the reasons for nonparticipation during the screening call. After individuals agreed to participate, we emailed them a REDCap link to the consent form and the demographic questionnaire. Those who consented and completed the questionnaires received copies of the informed consent form. Participants were required to complete the survey only through this link, with responses automatically stored in REDCap.

Participation Status and Reasons for Nonparticipation

During the screening call, we asked individuals screening questions and documented their reasons for nonparticipation. The reasons were categorized into predeveloped options: (1) did not meet eligibility criteria, (2) not interested, (3) considered study protocol to be too burdensome, (4) worried about privacy issues, (5) others (eg, considering care replacement to facilities), (6) expressed interest but did not respond to the screening call, (7) noted invasion of privacy was undesirable, (8) harmful events occurred due to this study, and (9) no longer wanted to participate.

Demographic Characteristics of Participants

We collected demographic data from participants at baseline. The questionnaire included information on the age, gender, race (American Indian or Alaska Native, Asian Native Hawaiian or other Pacific Islander, Black or African American, White, unknown, or not reported), ethnicity (Hispanic or Latino, Not Hispanic or Latino, or not reported), relationship between caregivers and care recipients (spouse, partner, sibling, child, grandchild, or other), education level (never attended or kindergarten only, K grades from 1st to 12th [not diploma], high school graduate, general education diploma or equivalent, attended college but did not obtain a degree, associate degree, bachelor's degree, master's degree, professional school, doctoral degree, or unknown), employment status (working now, only temporarily laid off, sick leave, maternity leave, unemployed, retired, disabled permanently or temporarily, keeping house, student, other, or unknown), caregiving hours per day, caregiving duration in months, marital status, the number of household members, zip code, veteran status (yes or no), and any history of combat exposure for those with a military background.

Challenges and Strategies in Recruitment

To address recruitment challenges through on-site engagement, we conducted two 20-minute Zoom (Zoom Communications, Qumu Corporation) meetings with a memory clinic recruiter in June and November 2022. We also held weekly meetings with the research team to discuss challenges in recruiting through social media and brainstorm potential solutions.

Ethical Considerations

The smart health intervention was approved by the institutional review board at this study's institution (2019B0406) and was registered with ClinicalTrials.gov (NCT04536701) on September 3, 2020, and completed on December 31, 2023. All participants provided written, informed consent before the data collection. Participants were informed about this study's purpose, duration, location of data storage, potential benefits and risks of involvement in the intervention, and contact information for the principal investigators.

Data Analysis

We used SPSS (version 29; IBM Corp) for descriptive analyses, including frequencies, percentages, means, and SDs, to present the characteristics of recruitment and participant demographics. We also used a note-taking approach during discussions with a recruiter and team members to capture key points on the topic.

Results

Participation Status and Reasons for Nonparticipation

Figure 1 presents the recruitment flowchart and reasons for nonparticipation. We attempted to screen 201 caregivers through social media (125 individuals), memory clinic (63 individuals), and other sources, such as ClinicalTrials.gov or unknown routes (13 individuals). Of these, 22 caregivers and their care recipients agreed to consent; only half completed the baseline questionnaire and began the intervention. Six of the 11 participants were recruited through social media, and the remaining 5 were recruited through on-site engagement at the memory care clinic (Figure 1).

XSL•FO
RenderX

Figure 1. Flowchart illustrating participant recruitment and exclusion. *Some of 179 respondents addressed multiple reasons for their non-participation.

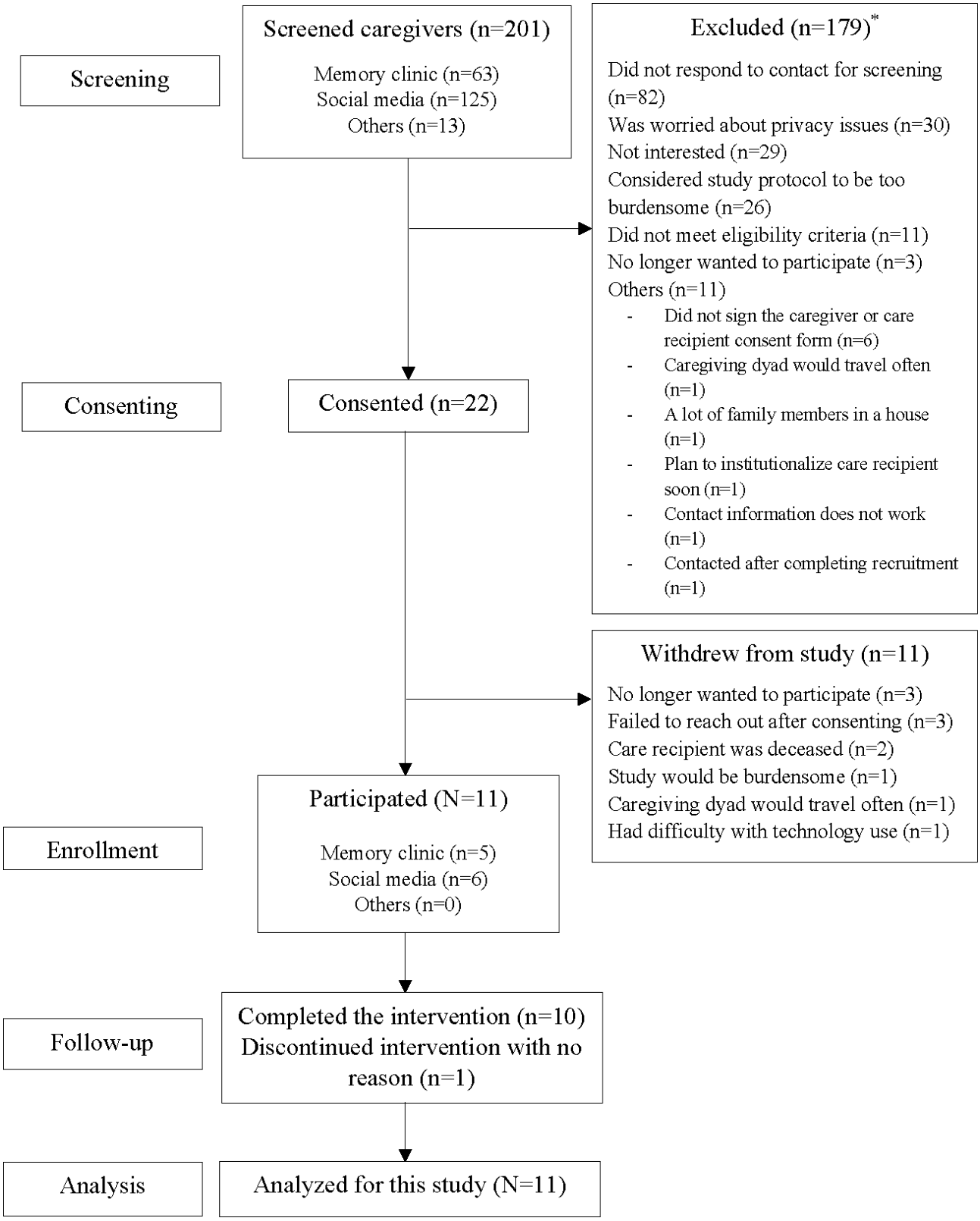

Among caregivers excluded, 82 did not respond to the screening contact, 30 declined to participate due to privacy concerns, 29 were not interested in the intervention, and 26 considered this study's protocol to be burdensome. Eleven respondents failed to meet the eligibility criteria, mainly because they were not primary or live-in caregivers, their care recipient was hospitalized or deceased, or they lacked functional Wi-Fi at home. Three caregivers expressed interest but later decided not to participate after further consideration. Other reasons for nonparticipation included (1) not providing written consent on a consent form after providing verbal agreement, (2) frequent travel plans, (3) having multiple household members besides the caregiving dyad, (4) planning to institutionalize their care recipient soon, and (5) providing incorrect contact information.

XSL•FO
RenderX

Further, 11 caregivers who consented ultimately withdrew from this study for reasons such as no longer wanting to participate, failure to reach out, and the death of the care recipient.

Demographic Characteristics of Participants

Table 1 presents the demographic data for the participating caregiving dyads. The average ages of caregivers and care recipients were 60.09 (SD 13.52) years and 76.73 (SD 8.04) years, respectively. Most caregivers were female (n=8, 73%), while most care recipients were male (n=7, 64%). Most caregiving dyads were non-Hispanic White (n=7, 82% of caregivers; n=9, 91% of care recipients). Education levels varied, but all caregivers had attained education beyond a high school degree. Regarding employment status, 8 (73%) caregivers and all care recipients were either unemployed or retired. Eight (73%) care recipients and 7 (64%) caregivers were married, with caregivers typically being spouses. The average amount of caregiving hours per day was 21.18 (SD 6.37) hours, and the average caregiving duration was 34.82 (SD 20.26) months.

XSL•FO
RenderX

Table . Demographics of actual participating dyads (N=11).

Characteristic	Caregivers	Care recipients
Age (years), mean (SD)	60.09 (13.52)	76.73 (8.04)
Gender, n (%)		
Female	8 (73)	4 (36)
Male	3 (27)	7 (64)
Race, n (%)		
African American or Black	1 (9)	1 (9)
Asian American	1 (9)	0 (0)
Non-Hispanic White	9 (82)	10 (91)
Ethnicity, n (%)		
Non-Hispanic	10 (91)	10 (91)
Not reported	1 (9)	1 (9)
Education level, n (%)		
General education diploma or equivalent	0 (0)	1 (9)
12th grade, not diploma	0 (0)	1 (9)
High school graduate	0 (0)	1 (9)
Attended college but did not obtain a degree	1 (9)	2 (18)
Associate degree	2 (18)	2 (18)
Bachelor's degree	4 (36)	1 (9)
Master's degree	3 (27)	1 (9)
Doctoral degree	1 (9)	2 (18)
Employment status, n (%)		
Employed	3 (27)	0 (0)
Not employed or retired	8 (73)	11 (100)
Marital status, n (%)		
Married	7 (64)	8 (73)
Never married	3 (27)	0 (0)
Widowed	0 (0)	2 (18)
Divorced	1 (9)	1 (9)
Veteran status and combat exposure, n (%)		
Yes, no combat exposure	2 (18)	1 (9)
No	9 (82)	10 (91)
Caregiving hours (h/d), mean (SD)	21.18 (6.37)	—[a]
Caregiving duration (mo), mean (SD)	34.82 (20.26)	—
Household member, mean (SD)	2.27 (0.65)	—

[a]The "—" indicates "not applicable"

Lessons Learned Through Recruitment Challenges

Challenges of On-Site Engagement

A recruiter responsible for informing caregivers about this study at the memory clinic provided some reasons for a low rate of face-to-face recruitment. First, staffing shortages at the clinic and the shift from in-person to remote medical appointments likely reduced recruitment opportunities. When patients transitioned to telehealth appointments, the recruiter could not meet with them and their caregivers to provide study information. Additionally, increased patient no-shows for in-person medical appointments further hindered recruitment efforts. Caregivers also expressed concerns about the time commitment and privacy issues associated with technology-based interventions, which echoed the feedback from the caregivers who refused participation in the smart health

XSL•FO
RenderX

intervention. The recruiter noted that caregivers showed more interest in pharmacological trials for their care recipients than in nonpharmacological trials for themselves.

To address these challenges, the recruiter suggested introducing this study virtually before or after telehealth appointments to raise awareness among caregivers and their care recipients. Increasing the clinic's staff who help study recruitment and having study personnel provide detailed explanations of this study—either in person or virtually—while caregivers wait for their care recipients could also help improve recruitment rates.

Challenges of Social Media Recruitment

We also encountered specific challenges with recruiting through social media platforms. A primary issue was the need to present study information briefly within social media posts, which often limited the detail we could provide. This brevity hindered potential participants from receiving detailed study information to decide, which is particularly critical in intervention studies. For example, of 601 caregivers who clicked the advertisements in April, only 7 reached out to the staff for further information. This suggests that the brief content in the advertisements may not have efficiently delivered this study's purpose or importance to potential participants.

Another issue was that the visibility of our study was directly tied to our advertising budget. As a result, we had to allocate more resources to increase our reach on social media platforms. Moreover, we faced difficulties maintaining the balance between demand and supply. When study advertisements were posted on social media, we received a flood of inquiries from interested participants, but limitations in the availability of study equipment caused delays. Some participants had to wait weeks or months to receive this study's equipment, likely contributing to decreased interest and higher withdrawal rates.

Strategies to Improve Recruitment

We discussed and attempted several strategies to alleviate the recruitment challenges, specifically targeting the individual concerns that led to the reluctance to participate. To mitigate privacy concerns in the smart health intervention, we reassured caregivers during recruitment that we would not record their conversation or behaviors but focus on the changes in acoustic characteristics, such as tone, intonation, and amplitude. We allowed participants to choose the microphone placement for data collection, provided it was in a room where frequent interactions with the care recipient occurred. Additionally, we clarified that we would only analyze acoustic data collected from the participating caregiving dyad, differentiating their voices from others in the household (ie, visitors).

We also developed a dashboard (Figure 2) and shared it with participants, showing the visualized data without capturing specific words or conversations. This approach helped ease privacy concerns, and as this study progressed, participants gained confidence in the system's utility. By the end of this study, they reported that privacy was no longer a concern.

To enhance recruitment through social media, we consulted with social media specialists on ways to improve the visibility of this study's advertisements. We increased our budget for social media advertising from US $100/wk in April to US $200/wk in September and revised the advertising text to provide more detailed information for eligible caregivers (Textbox 1). These modifications led to a slight rise in interest from potential study participants, as indicated by an increase in advertisement clicks from 601 in April to 1421 in September. The number of inquiries for detailed information also grew, from 7 in April to 17 in September. However, there was no change in terms of percentage (from 7/601, 1.16% in April to 16/1701, 1.20%), while the cost per click was higher (from US $0.33 in April to US $0.42 in September). Despite these efforts, we recruited only 5 out of 125 participants through social media, indicating that the strategies may not be effective and that further actions are needed to capture the attention of potential participants better.

XSL•FO
RenderX

Figure 2. Examples of dashboard data collection. These graphs were generated based on an actual participant's data. The x-axis represents the passage of time. (A) The cumulated number of daily mood states through acoustic monitoring. (B) Self-reported daily emotional status. (C) Followed recommendation tips per day (bars) and daily scores for helpfulness (line).

Textbox 1. Changes in the statements on advertising posts.

Statement on the advertisements *before* the revision

Researchers at the ### are seeking participants for a study to learn more about ways we can help family caregivers for persons with memory loss manage their stress better. The research study will help family caregivers recognize their stressful situations and find ways they may be able to feel less stress.

Statement on the advertisements *after* the revision

Researchers at the ### are seeking family or friends who care for people with memory loss for a study. *By identifying voice changes, this study* will help caregivers recognize their feelings in tense moments and *cope with their difficult conversations with persons with memory loss*. This may help caregivers find ways to i*ncrease self-care and satisfaction with their daily lives. You do not have to be good with technology to participate. Participant privacy will be carefully protected. Please do not hesitate to contact us with questions or for more information!*

We prepared and deployed additional study devices for participants on the waitlist. To enhance their motivation to participate in this study, we conducted regular follow-ups, approximately 2 to 3 times monthly. These actions enabled us to monitor and address study-related challenges and facilitate effective communication and rapport between study personnel and participants, helping ensure their continued involvement in this study. However, some potential participants experienced changes in their caregiving situations (eg, planning travel, experiencing a care recipient's death, or having second thoughts about engaging in this study) and withdrew from the waitlist.

Discussion

Principal Findings

This study aimed to identify recruitment challenges in a smart health technology intervention for caregivers of people living with dementia. We observed a low recruitment rate and identified various reasons for nonparticipation among caregivers. Beyond nonresponsiveness to screening, the most common reasons for nonparticipation included privacy concerns, a lack of interest, and perceptions of this study as being burdensome.

XSL•FO
RenderX

The demographic characteristics of participating caregivers showed that the majority were mainly female, White, spousal caregivers, and retired. Recruitment challenges were observed both on-site and through social media. On-site engagement challenges included the transition to remote medical appointments and staffing issues. At the same time, social media-based recruitment faced limitations related to the context of postings, budget considerations, and balancing demand and supply.

We observed a lower recruitment rate through social media (6/125 individuals) than on-site engagement (5/63 individuals). Previous literature has reported mixed results regarding the effectiveness of using social media for recruitment in intervention research [13,20-26]. While some studies have suggested that social media recruitment could be costly and ineffective [20-22,24], others have argued that social media could benefit the recruitment of potential target participants [13,23,25,26]. These conflicting findings highlight a need for further research to explore differences in recruitment rates across recruitment channels in technology-based interventions for people living with ADRD and their caregivers.

Caregivers of people living with ADRD may be anxious about sharing their data to use technology [27-29]. Additionally, they often prioritize dementia care and pay more attention to clinical trials focused on treating a specific disease than trials targeting their health [30]. While multiple pharmacological clinical trials may be ongoing in a clinic, caregivers of people living with ADRD may overlook nonpharmacological intervention studies. Moreover, they already spend considerable time and energy caring for their care recipients, and they may perceive participating in an intervention study, especially a technology-based trial, to be overly demanding [29,31,32]. Therefore, intervention research using technology should use strategies that reduce the perceived burden, such as providing training, using user-friendly platforms, or offering nondigital alternatives, including daily logs for those uncomfortable with technology in the caregiving population [8]. These tailored approaches have the potential to enhance the enrollment in dementia caregiving research.

Being female and non-Hispanic White are common characteristics of caregivers of people living with dementia in the United States [1], and spousal dementia caregivers are more likely to be retired [33]. Compared to male caregivers, females are more likely to sacrifice personal time, which can lead to a more significant burden and depressive symptoms [1]. Spousal dementia caregivers are also at a higher risk for depression, with rates being 2.5 times higher than that for nonspousal caregivers [1]. Caregivers with these characteristics may be more willing to participate in caregiving research on stress management. Still, the lack of demographic data on nonparticipants makes it difficult to confirm this. Therefore, future studies should gather information from participants and nonparticipants to better understand why some caregivers choose not to engage in research.

Limitations and Potential Implications

Despite our efforts, the strategies implemented in the smart health intervention needed to address recruitment barriers effectively. First, we needed more complete data on the socioeconomic status and race or ethnicity of potential participants who did not enroll in our study. Research shows that dementia caregiving dyads from lower socioeconomic backgrounds are less likely to participate in studies involving technology due to unfamiliarity and limited access to technological devices [31,34]. This gap in our data limited our ability to access the potential impact of the backgrounds of caregiving dyads on the recruitment rate.

Second, concerns about this study being too demanding persisted throughout recruitment, likely due to caregivers' unfamiliarity with the technology. This study's design, which included real-time messaging through the EMA system, may have contributed to this perception. However, an intervention study using the EMA system reported that the EMA approach did not bother participating caregivers [35]. Nonetheless, more research is needed to understand how the EMA system affects caregivers' daily lives while capturing data [36].

We could not manage the recruitment challenges, such as the transition to remote medical appointments and staffing issues, through on-site engagement. Recruitment might have improved with more advertising before or after virtual medical appointments, greater research personnel involvement, and additional resources; however, barriers to participating in telehealth visits made implementing these strategies unfeasible. The rise in health care since the COVID-19 pandemic—an approximately 150% increase in telehealth visits [37]—posed a barrier to recruiting caregivers in person. Future studies should explore ways to overcome these challenges when recruiting participants with limited in-person contact.

Despite these limitations, the findings of this study can inform strategies for addressing recruitment challenges in dementia caregiving research. Our findings offer insights into the reasons for nonparticipation and how to address participants' concerns regarding the smart health intervention. The findings may help researchers develop better strategies to improve recruitment rates in similar studies, offering caregivers opportunities to enhance their health and the quality of care they provide to their loved ones.

Acknowledgments

The authors thank Jenny Icenhour, BA, a research manager in the memory clinic, for providing opinions on the research topic. The data collection was funded by the National Science Foundation (grants 1838589 and 1838615). The National Center for Advancing Translational Sciences of the National Institutes of Health also supported this research (grant UL1TR002733). The material is the full responsibility of the authors and may not reflect the official views of the National Institutes of Health.

XSL•FO
RenderX

Conflicts of Interest

None declared.

References

1. 2024 Alzheimer's disease facts and figures. Alzheimers Dement 2024 May;20(5):3708-3821. [doi: 10.1002/alz.13809] [Medline: 38689398]
2. Pinyopornpanish K, Soontornpun A, Wongpakaran T, et al. Impact of behavioral and psychological symptoms of Alzheimer's disease on caregiver outcomes. Sci Rep 2022 Aug 19;12(1):14138. [doi: 10.1038/s41598-022-18470-8] [Medline: 35986203]
3. Sun Y, Ji M, Leng M, Li X, Zhang X, Wang Z. Comparative efficacy of 11 non-pharmacological interventions on depression, anxiety, quality of life, and caregiver burden for informal caregivers of people with dementia: a systematic review and network meta-analysis. Int J Nurs Stud 2022 May;129:104204. [doi: 10.1016/j.ijnurstu.2022.104204] [Medline: 35247788]
4. González-Fraile E, Ballesteros J, Rueda JR, Santos-Zorrozúa B, Solà I, McCleery J. Remotely delivered information, training and support for informal caregivers of people with dementia. Cochrane Database Syst Rev 2021;1(1):CD006440. [doi: 10.1002/14651858.CD006440.pub3]
5. Sharma RK, Teng A, Asirot MG, Taylor JO, Borson S, Turner AM. Challenges and opportunities in conducting research with older adults with dementia during COVID-19 and beyond. J Am Geriatr Soc 2022 May;70(5):1306-1313. [doi: 10.1111/jgs.17750] [Medline: 35285942]
6. Ali MT, Turetta C, Demrozi F, Pravadelli G. ICT-based solutions for Alzheimer's disease care: a systematic review. IEEE Access 2024;12:13944-13961. [doi: 10.1109/ACCESS.2024.3356348]
7. Wen Y, Xing Y, Ding Y, Xu W, Wang X. Challenges of conducting of online educational programs for family caregivers of people with dementia living at home: an integrative review. Int J Nurs Sci 2023 Jan;10(1):121-128. [doi: 10.1016/j.ijnss.2022.12.017] [Medline: 36860711]
8. Brookman R, Parker S, Hoon L, et al. Technology for dementia care: what would good technology look like and do, from carers' perspectives? BMC Geriatr 2023 Dec 16;23(1):867. [doi: 10.1186/s12877-023-04530-9] [Medline: 38104074]
9. Brodaty H, Donkin M. Family caregivers of people with dementia. Dialogues Clin Neurosci 2009;11(2):217-228. [doi: 10.31887/DCNS.2009.11.2/hbrodaty] [Medline: 19585957]
10. Keenan TA. U.S. caregivers' use of technology. : AARP Research; 2022. [doi: 10.26419/res.00566.001]
11. Kilmer G, Omura JD, Bouldin ED, et al. Changes in health indicators among caregivers - United States, 2015-2016 to 2021-2022. MMWR Morb Mortal Wkly Rep 2024 Aug 29;73(34):740-746. [doi: 10.15585/mmwr.mm7334a2] [Medline: 39207933]
12. Joshi S, Park T, Brody L, et al. Recruitment of family caregivers of persons with dementia: lessons learned from a pilot randomized controlled trial. Front Pain Res (Lausanne) 2023;4:1125914. [doi: 10.3389/fpain.2023.1125914] [Medline: 37051457]
13. Baker FA, Blauth L, Bloska J, et al. Recruitment approaches and profiles of consenting family caregivers and people living with dementia: a recruitment study within a trial. Contemp Clin Trials Commun 2023 Apr;32:101079. [doi: 10.1016/j.conctc.2023.101079] [Medline: 36949849]
14. Langbaum JB, Zissimopoulos J, Au R, et al. Recommendations to address key recruitment challenges of Alzheimer's disease clinical trials. Alzheimers Dement 2023 Feb;19(2):696-707. [doi: 10.1002/alz.12737] [Medline: 35946590]
15. Yang M, Samper-Ternent R, Volpi E, et al. The dementia care study (D-CARE): recruitment strategies and demographic characteristics of participants in a pragmatic randomized trial of dementia care. Alzheimers Dement 2024 Apr;20(4):2575-2588. [doi: 10.1002/alz.13698] [Medline: 38358084]
16. Gao Y, Jabbour J, Schlegel EC, et al. Out-of-the-box deployment to support research on in-home care of Alzheimer's patients. IEEE Pervasive Comput 2021;21(1):37-47. [doi: 10.1109/MPRV.2021.3106885]
17. Rose KM, Gordon KC, Schlegel EC, et al. Smarthealth technology study protocol to improve relationships between older adults with dementia and family caregivers. J Adv Nurs 2021 May;77(5):2519-2529. [doi: 10.1111/jan.14714] [Medline: 33576064]
18. Rose K. Collaborative research: learning and improving Alzheimer's patient-caregiver relationships via smart healthcare technology. Clinicaltrials.gov. 2024. URL: https://clinicaltrials.gov/study/NCT04536701 [accessed 2024-10-09]
19. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381. [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
20. Beattie E, Fielding E, O'Reilly M, Brooks D, MacAndrew M, McCrow J. Recruitment of individuals with dementia and their carers for social research: lessons learned from nine studies. Res Gerontol Nurs 2018 May 1;11(3):119-128. [doi: 10.3928/19404921-20180308-01] [Medline: 29634849]
21. Leach MJ, Ziaian T, Francis A, Agnew T. Recruiting dementia caregivers into clinical trials: lessons learnt from the Australian TRANSCENDENT trial. Alzheimer Dis Assoc Disord 2016;30(4):338-344. [doi: 10.1097/WAD.0000000000000149] [Medline: 27227995]

XSL•FO
RenderX

22. Leslie M, Khayatzadeh-Mahani A, MacKean G. Recruitment of caregivers into health services research: lessons from a user-centred design study. Res Involv Engagem 2019;5(1):17. [doi: 10.1186/s40900-019-0150-6] [Medline: 31139432]
23. Wasfi R, Stephens ZP, Sones M, et al. Recruiting participants for population health intervention research: effectiveness and costs of recruitment methods for a cohort study. J Med Internet Res 2021 Nov 12;23(11):e21142. [doi: 10.2196/21142] [Medline: 34587586]
24. Whitlatch CJ, Orsulic-Jeras S, Johnson J. Challenges to and strategies for recruiting chronic care dyads into intervention research. Chronic Illn 2021 Sep;17(3):232-241. [doi: 10.1177/1742395319869434] [Medline: 31426659]
25. Bremer W, Sarker A. Recruitment and retention in mobile application-based intervention studies: a critical synopsis of challenges and opportunities. Inform Health Soc Care 2023 Apr 3;48(2):139-152. [doi: 10.1080/17538157.2022.2082297] [Medline: 35656732]
26. Darko EM, Kleib M, Olson J. Social media use for research participant recruitment: integrative literature review. J Med Internet Res 2022 Aug 4;24(8):e38015. [doi: 10.2196/38015] [Medline: 35925655]
27. Block L, Gilmore-Bykovskyi A, Jolliff A, Mullen S, Werner NE. Exploring dementia family caregivers' everyday use and appraisal of technological supports. Geriatr Nurs 2020;41(6):909-915. [doi: 10.1016/j.gerinurse.2020.06.019] [Medline: 32665095]
28. Christie HL, Bartels SL, Boots LMM, Tange HJ, Verhey FRJ, de Vugt ME. A systematic review on the implementation of eHealth interventions for informal caregivers of people with dementia. Internet Interv 2018 Sep;13:51-59. [doi: 10.1016/j.invent.2018.07.002]
29. Hassan AYI. Challenges and recommendations for the deployment of information and communication technology solutions for informal caregivers: scoping review. JMIR Aging 2020 Jul 29;3(2):e20310. [doi: 10.2196/20310] [Medline: 32723720]
30. Clemmensen TH, Lauridsen HH, Andersen-Ranberg K, Kristensen HK. "I know his needs better than my own" - carers' support needs when caring for a person with dementia. Scand J Caring Sci 2021 Jun;35(2):586-599. [doi: 10.1111/scs.12875] [Medline: 32410395]
31. Brody AA, Convery KA, Kline DM, Fink RM, Fischer SM. Transitioning to remote recruitment and intervention: a tale of two palliative care research studies enrolling underserved populations during COVID-19. J Pain Symptom Manage 2022 Jan;63(1):151-159. [doi: 10.1016/j.jpainsymman.2021.06.017] [Medline: 34161811]
32. Trivedi RB, Szarka JG, Beaver K, et al. Recruitment and retention rates in behavioral trials involving patients and a support person: a systematic review. Contemp Clin Trials 2013 Sep;36(1):307-318. [doi: 10.1016/j.cct.2013.07.009] [Medline: 23916918]
33. Torres JM, Romero KRF, Kotwal AA, et al. Spouses of individuals living with mild cognitive impairment or dementia in the United States: a descriptive, population-based study. Alzheimers Dement 2024 Mar;20(3):1562-1572. [doi: 10.1002/alz.13555] [Medline: 38041823]
34. AboJabel H, Welsch J, Schicktanz S. Cross-cultural perspectives on intelligent assistive technology in dementia care: comparing Israeli and German experts' attitudes. BMC Med Ethics 2024 Feb 7;25(1):15. [doi: 10.1186/s12910-024-01010-6] [Medline: 38326778]
35. van Knippenberg RJM, de Vugt ME, Ponds RW, Myin-Germeys I, van Twillert B, Verhey FRJ. Dealing with daily challenges in dementia (deal-id study): an experience sampling study to assess caregiver functioning in the flow of daily life. Int J Geriatr Psychiatry 2017 Sep;32(9):949-958. [doi: 10.1002/gps.4552] [Medline: 27464472]
36. Han A, Malone LA, Lee HY, et al. The use of ecological momentary assessment for family caregivers of adults with chronic conditions: a systematic review. Health Psychol Res 2024;12:93907. [doi: 10.52965/001c.93907] [Medline: 38435338]
37. Koonin LM, Hoots B, Tsang CA, et al. Trends in the use of telehealth during the emergence of the COVID-19 pandemic - United States, January-March 2020. MMWR 2020 Oct 30;69(43):1595-1599. [doi: 10.15585/mmwr.mm6943a3] [Medline: 33119561]

Abbreviations

ADRD: Alzheimer disease and related dementias
EMA: ecological momentary assessment
REDCap: Research Electronic Data Capture

Edited by A Mavragani; submitted 24.05.24; peer-reviewed by B Davis, X Li; revised version received 22.11.24; accepted 05.12.24; published 17.01.25.

Please cite as:
Ko E, Gao Y, Wang P, Wijayasingha L, Wright KD, Gordon KC, Wang H, Stankovic JA, Rose KM
Recruitment Challenges and Strategies in a Technology-Based Intervention for Dementia Caregivers: Descriptive Study
JMIR Form Res 2025;9:e59291
URL: https://formative.jmir.org/2025/1/e59291
doi:10.2196/59291

XSL•FO
RenderX

© Eunjung Ko, Ye Gao, Peng Wang, Lahiru Wijayasingha, Kathy D Wright, Kristina C Gordon, Hongning Wang, John A Stankovic, Karen M Rose. Originally published in JMIR Formative Research (https://formative.jmir.org), 17.1.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Analysis of Social Performance and Action Units During Social Skills Training: Focus Group Study of Adults With Autism Spectrum Disorder and Schizophrenia

Hiroki Tanaka[1*], PhD; Kana Miyamoto[2*], PhD; Jennifer Hamet Bagnou[3], MS; Elise Prigent[3], PhD; Céline Clavel[3], PhD; Jean-Claude Martin[3], PhD; Satoshi Nakamura[2], PhD

[1]International Christian University, Mitaka, Tokyo, Japan
[2]Nara Institute of Science and Technology, Ikoma, Japan
[3]Université Paris-Saclay, Orsay, France
[*]these authors contributed equally

Corresponding Author:
Hiroki Tanaka, PhD
International Christian University, Mitaka, Tokyo, Japan

Abstract

Background: Social communication is a crucial factor influencing human social life. Quantifying the degree of difficulty faced in social communication is necessary for understanding developmental and neurological disorders and for creating systems used in automatic symptom screening and assistive methods such as social skills training (SST). SST by a human trainer is a well-established method. Previous SST used a modified roleplay test to evaluate human social communication skills. However, there are no widely accepted evaluation criteria or social behavioral markers to quantify social performance during SST.

Objective: This paper has 2 objectives. First, we propose applying the Social Performance Rating Scale (SPRS) to SST data to measure social communication skills. We constructed a Japanese version of the SPRS already developed in English and French. Second, we attempt to quantify action units during SST for people with autism spectrum disorder (ASD) or schizophrenia.

Methods: We used videos of interactions between trainers, adults with ASD (n=16) or schizophrenia (n=15), and control participants (n=19) during SST sessions. Two raters applied the proposed scale to annotate the collected data. We investigated the differences between roleplay tasks and participant groups (ASD, schizophrenia, and control). Furthermore, the intensity of action units on the OpenFace toolkit was measured in terms of mean and SD during SST roleplaying.

Results: We found significantly greater gaze scores in adults with ASD than in adults with schizophrenia. Differences were also found between the ratings of different tasks in the adults with schizophrenia and the control participants. Action units numbered AU06 and AU12 were significantly deactivated in people with schizophrenia compared with the control group. Moreover, AU02 was significantly activated in people with ASD compared with the other groups.

Conclusions: The results suggest that the SPRS can be a useful tool for assessing social communication skills in different cultures and different pathologies when used with the modified roleplay test. Furthermore, facial expressions could provide effective social and behavioral markers to characterize psychometric properties. Possible future directions include using the SPRS for assessing social behavior during interaction with a digital agent.

(JMIR Form Res 2025;9:e59261) doi:10.2196/59261

KEYWORDS

social performance rating scale; social skills training; autism spectrum disorder; schizophrenia; facial expressions; social; autism; training; communication; trainers; tool; neurological

Introduction

Social communication skills are essential in human life. Experienced psychiatrists identify people with social communication difficulties, such as autism spectrum disorder (ASD or autism spectrum condition) and schizophrenia, through interviews based on diagnostic criteria, responses, and various neuropsychological tests [1,2]. To improve the identification accuracy of neurodevelopmental and neurological disorders such as ASD and schizophrenia, it is crucial to discover symptom-specific behavioral markers and phenotyping. Both adults with ASD and and adults with schizophrenia are characterized by social difficulties [3]; in terms of deficits, there have been reports describing conversational gestures that are less closely synchronized with co-occurring speech [4], frequency of eye contact [5], speech prosody [6], and delayed

XSL•FO
RenderX

response time [7]. For schizophrenia, deficits in social cognition are a critical determinant of social functioning [8]. It has been reported that people with ASD and schizophrenia share several symptomatically similar characteristics such as flat or blunted affect (eg, reduced eye contact) and alogia (eg, impoverished speech) [9,10]. Thus, quantifying the degree of social communication difficulty is necessary to understand the nature of ASD and schizophrenia and to create systems for automatic symptom screening and early intervention methods such as social skills training (SST).

SST is a psychosocial treatment that has been widely applied and adapted to help people improve their social communication skills. It has been used in hospitals, employment support facilities, workplaces, and schools. A human trainer generally conducts SST to promote appropriate social communication skills, strengthen the individual's social self-efficacy, and reduce social anxiety. The Bellack method [11], or step-by-step SST, is a well-structured approach that includes defining target skills, modeling, roleplay, feedback, and homework. This method defines the SST framework and its 4 basic skills: asking for requests, declining requests, telling positive feelings, and listening to others. Some researchers have been conducting studies to build SST systems using socially interactive agents [12,13]. Personalizing is an essential element for such SST systems. For example, Bellack et al [11] recommended using shorter and more precise feedback for people with schizophrenia who experience hallucinations. Since they often hear voices distracting the SST, focusing is difficult. For ASD, feedback could be disengaging if trainers frequently mention exaggerated gestures or facial expressions, since these are major symptomatic characteristics. Toward more personalized SST systems, it is imperative to quantitatively measure the characteristics of ASD and schizophrenia.

However, there are no widely accepted evaluation criteria during SST. If a well-validated scale was available, it could be used as a ground truth value for machine learning models in the SST system [14]. Among social behavioral markers, individual roleplay performance can be rated with a modified roleplay test conducted [15,16], although there is no internationally validated scale. This type of test is further limited because it must be evaluated by experts such as psychiatrists and SST trainers. Moreover, there is no robust social behavioral marker to quantify social performance during SST. One of the important behavioral markers is facial expression. Previous works showed that action units AU06 and AU12 are less expressed in ASD participants [17,18], while AU01 is also less expressed in ASD participants [19]. For patients with schizophrenia, as far as the authors know, no previous study has provided an analysis of particular action units, although machine learning–based schizophrenia detection models have been proposed [20].

This paper proposes applying the Social Performance Rating Scale (SPRS) to measure social communication skills. The SPRS is a validated scale developed in English [21] and French [22]. The English version was applied to videos of a simulated dinner party for participants with a primary diagnosis of social phobia, those with a primary diagnosis of anxiety disorder, and control participants. The French version was applied to videos of collaborative games between neurotypical participants. This scale has been extensively used in a variety of research [23]. In this paper, we newly developed a Japanese version of the SPRS that can be used without special qualifications in SST. We collected data on 4 SST tasks performed by adults with ASD, with schizophrenia, and control groups. Two raters used the SPRS to annotate their performance in SST. The goals of our study are (1) to validate our Japanese version of the SPRS, (2) to apply the SPRS to SST and examine the correlations between the SPRS and the modified roleplay test, (3) to investigate whether the SPRS can be applied to participants with ASD or schizophrenia, and (4) to analyze the relations between the SPRS, facial expressions, and questionnaire results. Here, we aim to analyze the differences in action units during SST roleplay in people with ASD or schizophrenia. This is an extended version of a conference workshop's proceedings [24], in which we added an analysis of the facial expressions of people with ASD or schizophrenia.

In summary, we made contributions by developing and examining the following items: we developed a Japanese version of the SPRS; we analyzed social communication skills by focusing on the differences among adults with ASD or schizophrenia and a control group, on task differences of SST, and on questionnaire scores; and we analyzed action units by focusing on the differences among adults with ASD or schizophrenia and a control group, as well as questionnaire scores.

Methods

Ethical Considerations

This study was approved by the institutional review board of the Nara Medical University, Japan, and Nara Institute of Science and Technology, Japan (reference number: 2019-I-27 - 2). We explained the experiment's procedure and purpose to the participants and obtained informed consent. All data were anonymized. Participants received monetary rewards of up to 16,000 Japanese yen.

Development of Japanese Version of the SPRS

The items in the English version of the SPRS are provided in Table 1, where each of them is rated on a Likert scale of 1 to 5 [25].

XSL•FO
RenderX

Table . Social Performance Rating Scale (SPRS; modified from an earlier study [21]).

Item	Description
Gaze	• (1) Very poor: Participant completely avoids looking at the partner or stares continually. • (5) Very good: Participant keeps eye contact during the conversation, does not stare; shifts focus during pauses and conversation.
Voice quality	• (1) Very poor: (1) Participant speaks in a flat, monotonous voice; or (2) speaks at a low volume or mumbles; or (3) speaks overly loudly or has an intrusive tone (harsh or unpleasant voice quality). • (5) Very good: Participant is warm and enthusiastic in verbal expression without sounding condescending or gushy.
Length	• (1) Very poor: Monosyllabic ("hmmm," "yeah," "OK") speech turns; or responses so long that the partner must interrupt or cannot utter a reply. • (5) Very good: At most times, participant's utterances are 2 or more sentences long. Participant acknowledges partner's remarks without taking over and monopolizing the conversation.
Discomfort	• (1) Very high: Complete rigidity of arms, legs, or whole body. Constant leg movements or fidgeting with hands, hair, or clothing. Extremely stiff face or constant facial tics. Frequent nervous throat clearing, swallowing, or stuttering. Frequent inappropriate giggling or laughing. Look of extreme discomfort and desire to flee the situation shown by 2 or more breaks in the role. Participant does not pay attention to the roleplay tasks most of the time. • (5) Very low: Relaxed body posture and natural body movement. Participant laughs and smiles at appropriate times. She/he shows effective gesturing (to be distinguished from fidgeting). Participant focuses on the task all of the time, does not appear at all uncomfortable, but at ease in the situation.
Conversation flow	• (1) Very poor: Participant makes few attempts to initiate the conversation. Even when prompted by the partner, participant cannot maintain the conversation. Participant uses almost no open-ended questions or is intrusive in questions and shows no empathy. Participant does not pay attention to the information provided by partner. • (5) Very good: Participant easily maintains the conversation and responds smoothly to pauses in the conversation, often by following up on previous information provided by the partner or providing free information about themself on a related topic. Participant introduces new topics fluidly and frequently use open-ended questions. Participant shows genuine interest in the partner and follows up on the partner's remarks with warmth or enthusiasm.

We developed a Japanese version of the SPRS. As suggested by an earlier work [26], in its guide to good practices in translation and cultural adaptation, we first translated the scale's contents from English to Japanese and then from Japanese to English. This forward and backward translation method provides quality control to verify consistency between the translation and the original version. Inter Group Corp (Osaka, Japan), a company specializing in language and cross-cultural adaptation, translated the SPRS at our request. This Japanese version of the scale is available upon request to the authors.

SST Roleplay

We used a Japanese SST dataset collected in a previous study [14]. We collected SST data at Nara Institute of Science and Technology for the control group and at Nara Medical University for the clinical group. The dataset includes data from 50 adults with the following characteristics: 16 in the ASD group, 15 in the schizophrenia group, and 19 in the control group. We collected these psychiatric or neurodevelopmental clinical groups to investigate their social communication difficulties. We included these groups because the main SST clients in clinical facilities are those with ASD and schizophrenia. The control group participants were hired by a recruitment agency. Two psychiatrists with SST experience joined this study as trainers (roleplay partners). All participants performed SST using roleplay and feedback from the trainer. Four basic skills were performed as follows: asking for requests, declining requests, telling about positive feelings, and listening to others (ie, asking, declining, telling, and listening). Each roleplay lasts 1 to 3 minutes. Some participants performed SST multiple times, although our analysis was done on the initial roleplay data.

We excluded from the control group participants who had undergone eye surgery or had a history of psychiatric hospitalization. For the clinical group, we excluded participants

XSL•FO
RenderX

who scored less than 70 on the third edition of the Wechsler Adult Intelligence Scale [27]. We excluded participants with drug or alcohol addictions, personality disorders, or organic mental disorders. The data collection period was from January 2020 to January 2021. Due to COVID-19 concerns, a transparent partition was placed between the participants and trainers. Two video cameras were placed behind each conversationalist to record the other individual at chest level from the participants' faces. We also recorded a video and used 2 Kinect (Microsoft Corporation) sensors installed at the side of each conversationalist.

Questionnaire

We assessed each participant using the following questionnaires for correlation analysis with the SPRS and action units: the Facial Emotion Identification Test (FEIT) adopted in previous work [28], Kikuchi's scale of social skills with 18 items (Kiss-18) [29], Singelis independent-interdependent Self-Construal Scale (Self-Construal Scale) [30], the second edition of the Social Responsiveness Scale (SRS-2) [31] (Japanese version [32]), and the Japanese version of the Brief Assessment of Cognition in Schizophrenia (BACS-J) [33]. In addition, we used the Autism Diagnostic Observation Schedule, Second Edition (ADOS-2), for the ASD group and the Positive and Negative Syndrome Scale (PANSS) for the schizophrenia group [2,34]. Where possible, we obtained the total score and subscales of each questionnaire. Excluding the FEIT, the questionnaires were collected before the SST data collection.

The FEIT, which assesses the emotional perception of facial emotions, includes facial images of 19 different people classified into 1 of 6 emotions: happiness, sadness, anger, surprise, fear, or disgust. We included this assessment because people with ASD and schizophrenia struggle with social cognition in facial images [34]. Kiss-18, which measures social skill levels, is comprised of 18 questions based on 6 social skill categories. This metric comprehensively measures social skills. Self-Construal Scale, which consists of questions on a 7-point rating scale, was developed to measure how people view themselves in relation to others. The SRS-2, composed of 65 questions, is an evaluation metric of the severity of social impairment. Although the SRS-2 was initially designed to assess people with ASD, it can also differentiate among various social communication difficulties. Its effectiveness has been investigated with clinical subjects and members of the general population [31].

To evaluate the participants during SST roleplay, we also adapted a modified roleplay test [15]. The scale is available only in Japanese. It includes items more closely related to psychopathology than the SPRS because the main target of this test is adults with schizophrenia. For this test, third-party psychiatrists watch recorded videos from both oblique and side views. Our 2 raters used a Likert scale from 1 to 5 for eye contact, body direction and distance, facial expression, voice variation, clarity, fluency, and social appropriateness. Since the required skills are situation-dependent, the social appropriateness differs depending on each SST task. Let us explain examples of social appropriateness for Bellack's basic tasks. Listening to others, which determines whether the participants paid attention to the interlocutor, includes nodding, back-channels, and other nonverbal behaviors (eg, eye contact, smiling). In expressing positive feelings, social appropriateness involves expressing attention to the interlocutor's responses and the suitability of the participant's speech content. For making requests, social appropriateness assesses whether they explained the details of their request, including what kind of help they needed. It also includes whether they listened to the interlocutor. For declining, social appropriateness concerns whether they expressed remorse and appropriate reasons for their refusal. It also includes whether they proposed alternatives to the request (eg, "I'm sorry, but I propose doing it another way next time"), which is essential to this situation.

SPRS Rating

Two third-party raters performed the annotations; all are nonexperts in psychiatry, psychology, or SST. The raters watched the recorded videos from oblique and side views for this evaluation. They initially evaluated all participants' roleplay in the first of the 4 basic tasks during the SST. Here, they watched the first 3 videos and discussed the SPRS content for evaluation practice, and then they evaluated the scores of the other videos without discussion. They were not informed directly about the nature of the groups (clinical outpatients or control), but it is possible that they may have noticed the existence of these groups by the background in the video, since each group conducted SST in a different location. These raters were not given information about the other questionnaire scales, and they watched the videos in a shuffled order regardless of the group.

Action Unit Extraction

The intensity of action units on OpenFace toolkit (version 2.0) [35] is measured in terms of mean and SD during individual roleplaying. From the action units extracted by OpenFace, we selected specific action units: AU01 (inner brow raiser), AU02 (outer brow raiser), AU04 (brow lower), AU06 (cheek raiser), and AU12 (lip corner puller). These action units represent various facial muscle activations. For our analysis, we referred to a previous study that focused on action units less directly associated with lip movement [36]. The intensity of each action unit ranges from 0 to 5, indicating a spectrum from no activation to high activation. First, the intensity of frames with less than 0.7 confidence in the data was excluded as not a number. We then applied a moving average filter on 3-second frames. We then calculated the mean and SD of the cleaned intensity data for each action unit to assess the intensity and variability, respectively.

Statistical Analysis

The analyses were performed to investigate whether the SPRS could be used to evaluate social communication skills during SST. We analyzed the interrater agreement, items' characteristics, construct validity, and internal consistency. To check the interrater agreement, we calculated the intraclass correlation coefficient (ICC, 2 k) [37]. We averaged the SPRS scores for the 2 raters for our analysis. We calculated the Spearman correlation coefficients to examine the homogeneity of the SPRS and the revised roleplay test. We assessed the factor structure of the SPRS by exploratory factor analysis with

XSL•FO
RenderX

oblimin rotation to confirm the construct validity [38]. We examined the Bartlett test of sphericity (P<.001) and the Kaiser–Meyer–Olkin test for a sampling adequacy of 0.881 to apply the exploratory factorial analyses [39]. The internal consistency was assessed by calculating the Cronbach α coefficient [40] and McDonald ω [41].

We also analyzed the total of SPRS items in relation to the 3 groups, 4 tasks of SST, and the questionnaires. In comparing the control, ASD, and schizophrenia groups, we used the nonparametric Kruskal–Wallis test. Post-hoc analysis was conducted using the Mann-Whitney U test with Bonferroni correction. In comparing the asking, declining, telling, and listening tasks, we used the Wilcoxon signed rank test with Bonferroni correction. For action units, we applied the Mann-Whitney U test with Bonferroni correction to compare the control, ASD, and schizophrenia groups. Spearman correlation coefficients were calculated between the total SPRS items, action units, and questionnaire scores for all, control, ASD, and schizophrenia groups, although some questionnaires were not obtained from all participants. We used the JASP (Jeffreys's Amazing Statistics program) for the statistical analysis [42].

Results

SPRS Statistics

We measured the interrater agreement by calculating the ICC (2 k) for validation of the SPRS annotated by the 2 raters. We opted for the random 2-way ICC method with the mean as the unit of evaluation, as shown in Table 2. The ICC ranged from 0 to 1, and all items had reliability values between 0.509 and 0.804. The ICCs were above 0.5, which indicates more than moderate reliability [22]. The score for each SPRS item was averaged for the 2 raters' annotations in subsequent analyses. The Spearman correlation performed between the SPRS items and the modified roleplay test items shows significant results (all P<.001) and positive relationships. The correlation between all of them was greater than 0.6. Exploratory factor analyses were conducted on the 5 items of the SPRS using oblique rotation to investigate underlying constructs and assess dimensionality. Following a previous study [22], a 1-factor solution was adopted. Table 3 provides the factor loadings after rotation, illustrating how each of the 5 items aligns with the selected factor. The results show that the first factor explains 82.1% of the variance, indicating strong explanatory power. Factor loadings ranged from 0.874 for the "Discomfort" item to 0.932 for the "Conversation Flow" item, reflecting high values across all items. Additionally, Table 3 displays the uniqueness values for each variable, where uniqueness represents the proportion of variance specific to each variable that the factor does not explain. Higher uniqueness suggests lower relevance of that variable within the factorial model. For example, 23.6% of the variance in the "Discomfort" item remains unexplained by the factor, whereas the "Conversation Flow" item has a smaller unexplained variance (13.2%), showing greater alignment with the factor model. In sum, the factor analysis of the SPRS identifies a single factor composed of 5 items, capturing the verbal and nonverbal social skills essential for effective communication with others.

Table . Intraclass correlation coefficients for the Social Performance Rating Scale.

Item	Intraclass correlation coefficient
Gaze	0.509
Voice quality	0.795
Length	0.777
Discomfort	0.790
Conversation flow	0.804

Table . Factor loading for the Social Performance Rating Scale score.

Item	Factor 1	Uniqueness	Mean (SD)	Skewness	Kurtosis
Gaze	0.881	0.224	3.21 (1.00)	−0.351	−0.726
Voice quality	0.914	0.164	2.93 (1.18)	0.168	−0.975
Length	0.928	0.139	3.27 (1.10)	−0.247	−0.819
Discomfort	0.874	0.236	3.04 (1.02)	−0.074	−0.659
Conversation flow	0.932	0.132	3.15 (1.07)	−0.192	−0.585

We created an overall performance measure by summing the scores of the 5 SPRS items. This measure revealed that Cronbach α was 0.957 and McDonald ω was 0.958. These results indicate that the SPRS shows excellent internal consistency.

SPRS Differences Among Participant Groups

Box plots of the 5 items of the SPRS and the total of the items for the control, ASD, and schizophrenia groups are shown in Figure 1. There were significant differences in all items and in the total of the SPRS items (P<.001). Post-hoc analysis was conducted using the Mann-Whitney U test with Bonferroni correction. We found significant differences in gaze, voice

XSL•FO
RenderX

quality, length, discomfort, conversation flow, and the total of SPRS items for control-ASD and control-schizophrenia (P<.001). In ASD-schizophrenia, we found significantly greater gaze scores in the ASD group than in the schizophrenia group (P=.005).

Figure 1. Group differences of Social Performance Rating Scale scores. ASD: autism spectrum disorder; Sz: schizophrenia. *P=.02, **P<.001.

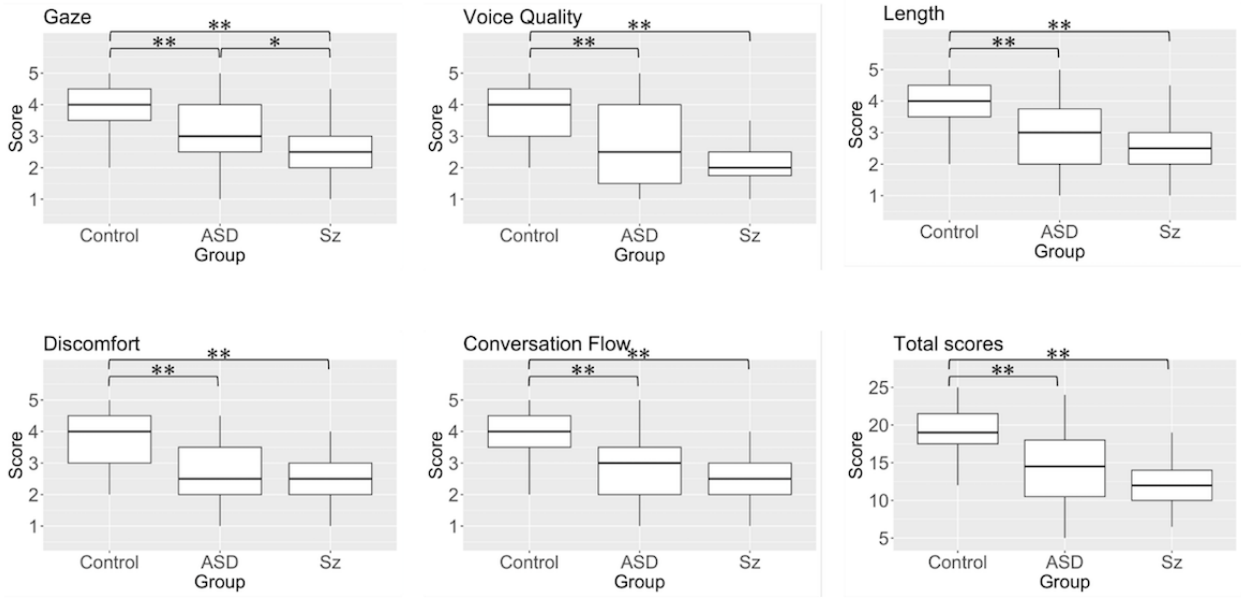

SPRS Differences Between 4 SST Tasks

SST task differences and a comparison are shown in Table 4. In the control group, we found significant differences in all items except gaze in asking-declining tasks and telling-declining tasks, and there were significant differences in the length of the asking-listening and telling-listening tasks. The declining task had lower scores than the asking and telling tasks for the control group. The length item of the listening task also had a low score compared with asking and telling tasks. We did not find significant differences between the 4 tasks in the ASD group. In the schizophrenia group, we found significant differences in the length of telling-listening tasks. For length, the listening task had a lower score than the telling task.

XSL•FO
RenderX

Table . Comparisons of Social Performance Rating Scale (SPRS) items for social skills training tasks. Parentheses represent significantly higher scored tasks. Effect size here is a rank-biserial correlation.

Item and combination			Control		Autism spectrum disorder		Schizophrenia	
			P value	Effect size	*P* value	Effect size	*P* value	Effect size
Gaze								
	Asking							
		Declining	.07	0.603	.08	−0.591	.58	−0.200
		Telling	.49	−0.289	.36	−0.295	.82	−0.091
		Listening	.48	0.219	.91	−0.044	.16	0.485
	Declining							
		Telling	.01	−0.743	.40	0.309	.97	0.026
		Listening	.15	−0.386	.17	0.491	.10	0.526
	Telling							
		Listening	.18	0.400	.24	0.500	.23	0.397
Voice quality								
	Asking							
		Declining	<.001 (Asking)[a]	0.963	.59	−0.250	.61	0.200
		Telling	.45	0.242	.45	−0.289	.22	−0.385
		Listening	.13	0.515	.81	0.091	.11	0.564
	Declining							
		Telling	.001 (Telling)[a]	−0.950	.97	−0.026	.12	−0.513
		Listening	.04	−0.600	.63	0.200	.49	0.255
	Telling							
		Listening	.28	0.341	.24	0.418	.03	0.758
Length								
	Asking							
		Declining	.001 (Asking)[a]	0.933	.52	0.198	.62	0.222
		Telling	.63	−0.167	.10	0.561	.43	−0.291
		Listening	.008 (Asking)[a]	0.824	.15	0.527	.04	0.727
	Declining							
		Telling	<.001 (Telling)[a]	−1.000	.30	0.364	.30	−0.364
		Listening	.56	−0.175	.24	0.350	.06	0.652
	Telling							
		Listening	.003 (Telling)[a]	0.912	.47	0.258	.005 (Telling)	0.923
Discomfort								
	Asking							
		Declining	.001 (Asking)[a]	0.933	.88	−0.073	.83	−0.111
		Telling	.90	−0.039	.81	0.111	.90	0.067
		Listening	.49	0.220	.60	−0.286	.29	0.382
	Declining							
		Telling	.001 (Telling)[a]	−0.912	.67	0.194	.67	0.238
		Listening	.02	−0.647	.62	−0.222	.25	0.385

XSL•FO
RenderX

Item and combination			Control		Autism spectrum disorder		Schizophrenia	
			P value	Effect size	*P* value	Effect size	*P* value	Effect size
	Telling							
		Listening	.41	0.248	.49	−0.267	.32	0.417
Conversation flow								
	Asking							
		Declining	.004 (Asking)[a]	0.797	.59	−0.179	.52	0.200
		Telling	.67	−0.164	.20	0.423	.15	−0.474
		Listening	.03	0.603	.50	0.242	.15	0.500
	Declining							
		Telling	<.001 (Telling)[a]	−0.949	.20	0.423	.15	−0.474
		Listening	.25	−0.342	.19	0.423	.37	0.318
	Telling							
		Listening	.02	0.700	>.99	0.010	.03	0.731
Total of SPRS items								
	Asking							
		Declining	<.001 (Asking)[a]	0.916	.41	−0.250	.96	0.036
		Telling	.63	−0.135	.61	0.162	.31	−0.330
		Listening	.07	0.510	.48	0.206	.03	0.681
	Declining							
		Telling	<.001 (Telling)[a]	−0.979	.38	0.276	.39	−0.267
		Listening	.09	−0.468	.27	0.333	.15	0.487
	Telling							
		Listening	.02	0.614	.53	0.192	.02	0.758

[a]Significant difference ($P<.008$).

Action Units Differences Between Participant Groups

Participant group differences and comparisons, as well as mean and SD, are shown in Figure 2. We present the cases where significant differences were observed. We found that AU02 is significantly activated in people with schizophrenia compared with other groups. For mean values of the action units, this implies that AU02 is significantly less for the control group than the ASD group. For AU06, schizophrenia is significantly less than the control group. For AU12, it is more evident that schizophrenia is significantly less than either the control or ASD group. Regarding SD, people with ASD have significantly activated AU02 compared with the control group. For AU12, schizophrenia has significantly less variance than the ASD group.

XSL•FO
RenderX

Figure 2. Action units' intensities of mean values. (**A**) ***P*<.001, **P*<.05 (AU02 control-ASD: *P*=.04, AU06 control-Sz *P*=.09, AU12 ASD-Sz *P*=.04); SDs. (**B**) **P*<.05 (AU02 control-ASD: *P*=.02, AU12 control-ASD: *P*=.04). ASD: autism spectrum disorder; Sz: schizophrenia.

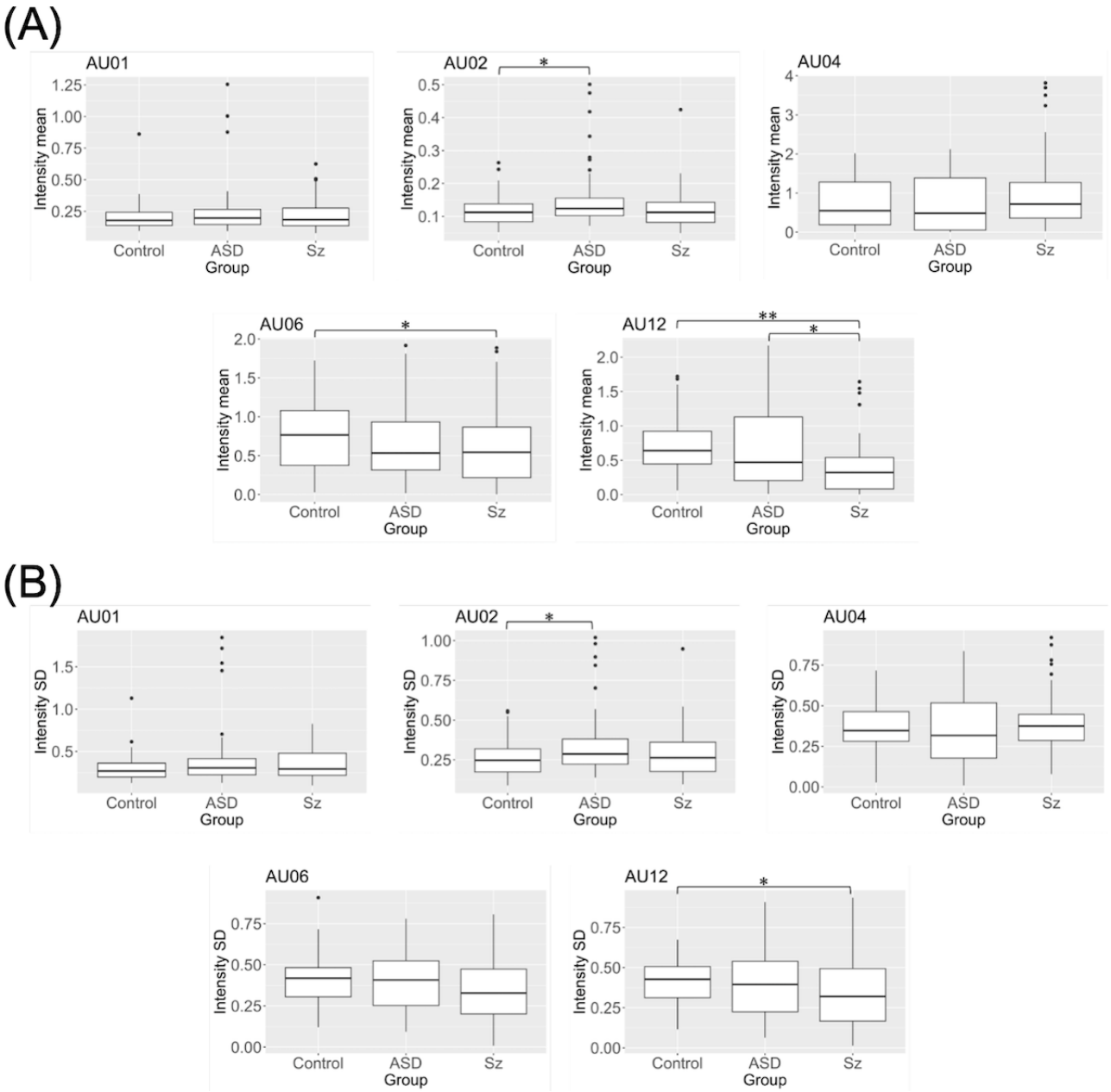

Correlation Between the SPRS, Action Units, and Questionnaire Scores

Spearman correlation coefficients were calculated between the total SPRS items and questionnaire scores for all the following groups: control, ASD, and schizophrenia. The correlation coefficients are displayed in Figure 3. Across all groups, significant correlations were found between the total SPRS items and various measures, including the FEIT, SRS-2, Self-Construal Scale, Kiss-18, ADOS-2, PANSS, as well as the mean and SD values of AU06 and AU12. These findings suggest a strong relationship between social performance and various action unit statistics, as well as questionnaire scores.

XSL•FO
RenderX

Figure 3. Correlation matrix ("?" indicates not applicable or missing data, cross mark represents nonsignificant correlations). (A) All groups, (B) control group, (C) ASD group, and (D) Sz group. ADOS: Autism Diagnostic Observation Schedule; FEIT: Facial Emotion Identification Test; Kiss-18; Kikuchi's scale of social skills with 18 items; PANSS: Positive and Negative Syndrome Scale; SPRS: Social Performance Rating Scale; SRS-2: Social Responsiveness Scale.

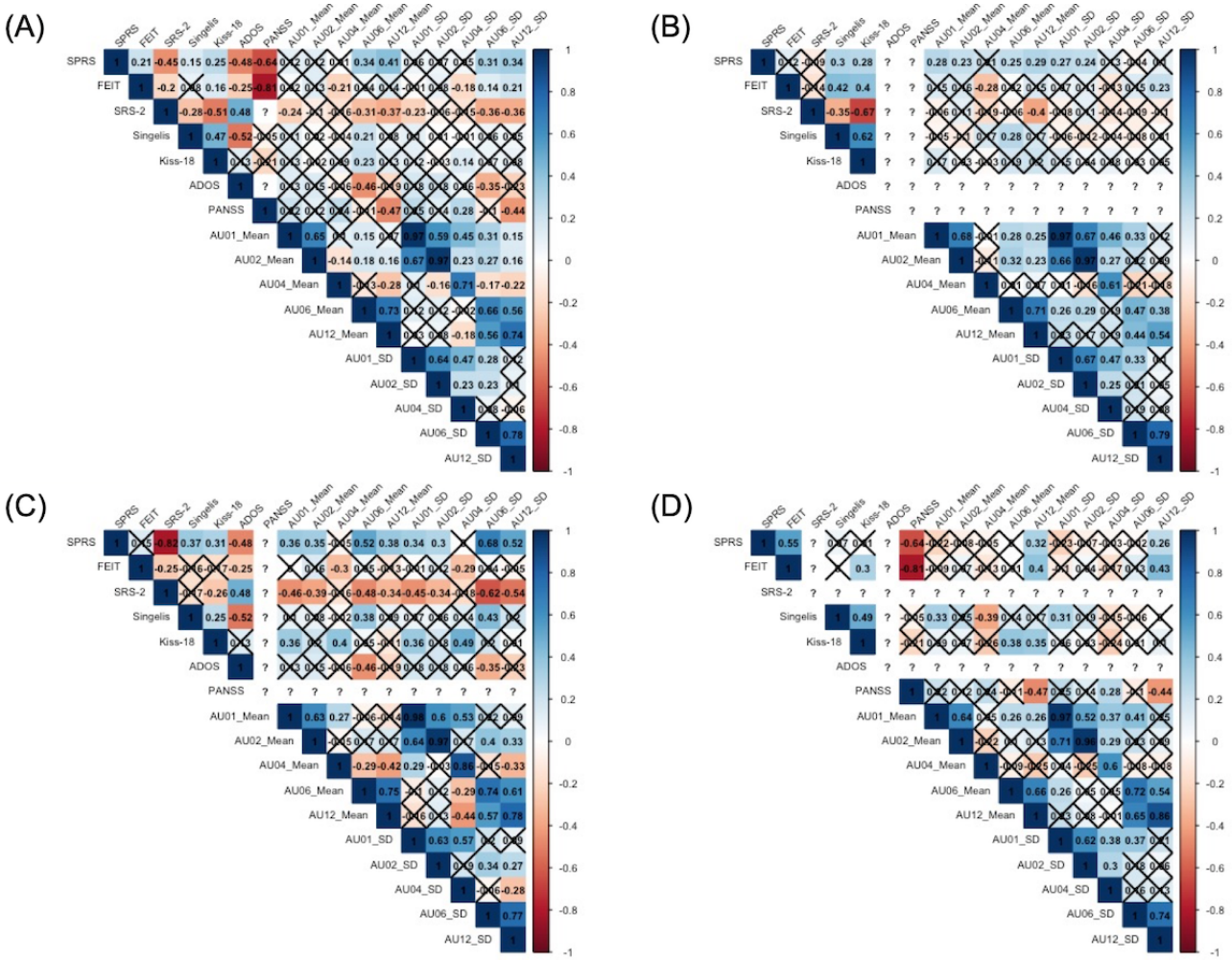

In the control group, significant correlations with the total SPRS items were noted for Kiss-18 and the Self-Construal Scale, reflecting the cultural emphasis on harmony over assertiveness, which may influence social communication skills in Japan. In the ASD group, significant correlations were found with Kiss-18, the Self-Construal Scale, SRS-2, BACS-J, and ADOS-2, suggesting a relationship between these questionnaires and the social performance of individuals with ASD. Notably, significant negative correlations were observed in the SRS-2, BACS-J, and ADOS-2, highlighting the complexities of social interactions within this group.

In the schizophrenia group, significant correlations with the total SPRS items were found for the FEIT, BACS-J, and PANSS. These results suggest that emotional awareness and social performance are interconnected.

Discussion

Principal Findings

This is the first study applying the SPRS and analysis of action units for SST roleplay data. In this paper, we propose evaluating roleplay performance using the SPRS. We calculated the interrater reliability between 2 raters' evaluated SPRS scores and found that the values were above 0.509 for any SPRS item. Our results show that the reliability was moderate to good [43]. Hamet Bagnou et al [22] calculated reliability among SPRS raters during collaborative tasks and showed that the reliability of gaze was the highest, but our results show that it has the lowest values. One reason could be the difference in conversation content. In our study, we applied 4 SST tasks, and the annotators needed to evaluate the SPRS based on the content of the tasks. The evaluation of SST tasks may have been more complex than that of collaborative tasks. Although there are slight differences between their study and ours, the results are nearly identical, with a reliability of 0.5 or higher in both studies.

The Spearman correlation between the SPRS items and revised roleplay test items was above 0.6. The factor analysis for the SPRS showed that the first factor explained 82.1% of the variance. The Cronbach α was 0.957, and the McDonald ω was 0.958. These results indicate that the SPRS shows excellent internal consistency and generally correlates to the modified roleplay test but is not identical.

We compared the SPRS scores by group and task. The control group differed significantly from the ASD and schizophrenia groups on every item of the SPRS. Several studies have shown that a control group shows higher social communication skills

XSL•FO
RenderX

than an ASD group and a schizophrenia group, and we showed similar results.

The ASD group had a higher gaze score than the schizophrenia group. We consider it relevant that the eye movement abnormalities were severe in the schizophrenia group compared with the ASD group [44].

We conducted analyses of the differences in the SPRS scores between tasks for each group. In the control group, the declining task had lower scores than the asking and telling tasks, except for gaze. Declining is the most difficult task of the basic SST tasks, which may have contributed to the lower scores [11]. The listening task had lower scores than the asking and telling tasks in length. In the schizophrenia group, the listening task also scored lower than the telling task in length. The control group met the trainer for the first time in the data recording. It is possible that some control participants were not interested in the topics provided by the trainer in the listening task. The ASD and schizophrenia groups had met with the trainer several times. However, the schizophrenia group had low scores. Some participants with schizophrenia might have seemed disinterested in the conversation [11], and the raters might have annotated the low scores.

The control group showed that the SPRS had significant positive correlations with Kiss-18 and with the Self-Construal Scale.

The independence of the Self-Construal Scale measures is characterized by autonomy and distinction from others [45], and it is embedded within the culture [46]. In cultures with strong interdependence, harmony is an important concern, while assertiveness tends to be discouraged. This is predominant in the East, including Japan, and it may be related to social communication skills in Japan.

In the ASD group, there were significant positive or negative correlations between the SPRS and Kiss-18 or the Self-Construal Scale; this was also the case for the control group. We found significant negative correlations in the SRS-2, BACS-J, and ADOS-2. Our results suggest a relationship between the questionnaire for ASD and the social communication skills of the ASD group. The schizophrenia group had significant positive or negative correlations between the SPRS and the FEIT, BACS-J, and PANSS. Emotional awareness and communication skills are also related [47], and we surmise that our results show a significant difference in the FEIT. Another study showed that negative symptoms related to social motivation may be primary determinants of social outcomes in schizophrenia, potentially having a greater impact than social competence [48].

The results on action units also resemble the previous findings, in which an ASD group showed less smiling as represented in AU06 and AU12 [15,17], but there was no significant difference in terms of AU01 [18]. We found that AU02 was significantly activated in people with ASD compared with other groups. Moreover, this was the first study to conduct a detailed analysis of action units for schizophrenia. We also found that the AU06 and AU12 action units were significantly deactivated in people with schizophrenia compared with the control group. Accordingly, our results demonstrated that action units could work as social behavioral markers of people with ASD or schizophrenia.

Limitations

This study is limited in terms of the population used and its generalizability. In particular, the SPRS was measured by 2 raters. Although we found high interrater correlations, a greater number of raters is needed to enhance the consistency and validity of the scale. In fact, previous work used an even larger number of raters, but a high interrater agreement was still maintained [21]. Our SST dataset was collected using outpatients with mild cases who were able to conduct SST verbally. Further study is needed to test whether the SPRS and action unit analysis could be applied to patients with more severe cases.

Conclusions

We applied and validated a Japanese version of the SPRS to SST roleplay data to measure individual social communication skills. We analyzed the SST data of 4 tasks carried out by adults with ASD or schizophrenia as well as control participants. The results found significant differences between the control and ASD as well as schizophrenia in all SPRS items. We found significantly greater gaze scores in the ASD group than in the schizophrenia group. There were also task-based differences in the schizophrenia and control groups. Our results suggest that the SPRS can be a useful tool for assessing social communication skills in different cultures and different pathologies and, in addition, can evaluate SST effectiveness as well as the modified roleplay test. We also found that the action units numbered AU06 and AU12 were significantly deactivated in people with schizophrenia compared with the control group. Possible future directions include using the SPRS for assessing social behavior during interaction with a digital agent.

Acknowledgments

This work was funded by the Japan Science and Technology Agency, Core Research for Evolutionary Science and Technology (grant JPMJCR19A5), ANR-19-JSTS-0001, and the Japan Society for the Promotion of Science (grant 22K12151).

Conflicts of Interest

None declared.

References

1. Guha M. Diagnostic and statistical manual of mental disorders: DSM-5 (5th edition). Ref Rev 2014;28(3):36-37. [doi: 10.1108/RR-10-2013-0256]

XSL•FO
RenderX

2. Bastiaansen JA, Meffert H, Hein S, et al. Diagnosing autism spectrum disorders in adults: the use of Autism Diagnostic Observation Schedule (ADOS) module 4. J Autism Dev Disord 2011 Sep;41(9):1256-1266. [doi: 10.1007/s10803-010-1157-x] [Medline: 21153873]
3. Marwaha S, Johnson S. Schizophrenia and employment - a review. Soc Psychiatry Psychiatr Epidemiol 2004 May;39(5):337-349. [doi: 10.1007/s00127-004-0762-4] [Medline: 15133589]
4. de Marchena A, Eigsti IM. Conversational gestures in autism spectrum disorders: asynchrony but not decreased frequency. Autism Res 2010 Dec;3(6):311-322. [doi: 10.1002/aur.159] [Medline: 21182208]
5. Senju A, Johnson MH. Atypical eye contact in autism: models, mechanisms and development. Neurosci Biobehav Rev 2009 Sep;33(8):1204-1214. [doi: 10.1016/j.neubiorev.2009.06.001]
6. Tanaka H, Sakti S, Neubig G, Toda T, Nakamura S. Linguistic and acoustic features for automatic identification of autism spectrum disorders in children's narrative. Presented at: Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: from Linguistic Signal to Clinical Reality; Jun 27, 2014; Baltimore, Maryland, USA p. 88-96. [doi: 10.3115/v1/W14-3211]
7. Heeman PA, Lunsford R, Selfridge E, et al. Autism and interactional aspects of dialogue. Presented at: Proceedings of the SIGDIAL 2010 Conference; Sep 24-25, 2010; Tokyo, Japan p. 249-252.
8. Dodell-Feder D, Tully LM, Hooker CI. Social impairment in schizophrenia: new approaches for treating a persistent problem. Curr Opin Psychiatry 2015 May;28(3):236-242. [doi: 10.1097/YCO.0000000000000154] [Medline: 25768085]
9. Konstantareas MM, Hewitt T. Autistic disorder and schizophrenia: diagnostic overlaps. J Autism Dev Disord 2001 Feb;31(1):19-28. [doi: 10.1023/a:1005605528309] [Medline: 11439750]
10. Foss-Feig JH, McPartland JC, Anticevic A, Wolf J. Re-conceptualizing ASD within a dimensional framework: positive, negative, and cognitive feature clusters. J Autism Dev Disord 2016 Jan;46(1):342-351. [doi: 10.1007/s10803-015-2539-x] [Medline: 26267330]
11. Bellack AS, Mueser KT, Gingerich S, et al. Social Skills Training for Schizophrenia: A Step-by-Step Guide: Guilford Publications; 2013.
12. Tanaka H, Iwasaka H, Matsuda Y, Okazaki K, Nakamura S. Analyzing self-efficacy and summary feedback in automated social skills training. IEEE Open J Eng Med Biol 2021;2:65-70. [doi: 10.1109/OJEMB.2021.3075567] [Medline: 35402987]
13. Nadel J, Grynszpan O, Martin JC. Autism and socially interactive agents. In: The Handbook on Socially Interactive Agents: 20 Years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 2: Interactivity, Platforms, Application: Association for Computing Machinery; 2022:437-462. [doi: 10.1145/3563659.3563673]
14. Saga T, Tanaka H, Matsuda Y, et al. Automatic evaluation-feedback system for automated social skills training. Sci Rep 2023 Apr 26;13(1):6856. [doi: 10.1038/s41598-023-33703-0] [Medline: 37100886]
15. Sasaki T. The reliability and validity of the revised role play test for assessment of social skills in schizophrenia. Clin Psychiatry 2006;48(11):1191-1198. [doi: 10.11477/mf.1405100353]
16. Tanaka H, Saga T, Iwauchi K, et al. The validation of automated social skills training in members of the general population over 4 weeks: comparative study. JMIR Form Res 2023 Apr 27;7:e44857. [doi: 10.2196/44857] [Medline: 37103996]
17. Bangerter A, Chatterjee M, Manfredonia J, et al. Automated recognition of spontaneous facial expression in individuals with autism spectrum disorder: parsing response variability. Mol Autism 2020 May 11;11(1):31. [doi: 10.1186/s13229-020-00327-4] [Medline: 32393350]
18. Alvari G, Furlanello C, Venuti P. Is smiling the key? machine learning analytics detect subtle patterns in micro-expressions of infants with ASD. J Clin Med 2021 Apr 19;10(8):1776. [doi: 10.3390/jcm10081776] [Medline: 33921756]
19. Legiša J, Messinger DS, Kermol E, Marlier L. Emotional responses to odors in children with high-functioning autism: autonomic arousal, facial behavior and self-report. J Autism Dev Disord 2013 Apr;43(4):869-879. [doi: 10.1007/s10803-012-1629-2] [Medline: 22918860]
20. Bishay M, Palasek P, Priebe S, Patras I. SchiNet: automatic estimation of symptoms of schizophrenia from facial behaviour analysis. IEEE Trans Affect Comput 2019;12(4):949-961. [doi: 10.1109/TAFFC.2019.2907628]
21. Fydrich T, Chambless DL, Perry KJ, Buergener F, Beazley MB. Behavioral assessment of social performance: a rating system for social phobia. Behav Res Ther 1998 Oct;36(10):995-1010. [doi: 10.1016/s0005-7967(98)00069-2] [Medline: 9714949]
22. Hamet Bagnou J, Prigent E, Martin JC, Clavel C. Adaptation and validation of two annotation scales for assessing social skills in a corpus of multimodal collaborative interactions. Front Psychol 2022;13:1039169. [doi: 10.3389/fpsyg.2022.1039169] [Medline: 36389487]
23. Stevens S, Cooper R, Bantin T, Hermann C, Gerlach AL. Feeling safe but appearing anxious: differential effects of alcohol on anxiety and social performance in individuals with social anxiety disorder. Behav Res Ther 2017 Jul;94:9-18. [doi: 10.1016/j.brat.2017.04.008] [Medline: 28441522]
24. Miyamoto K, Tanaka H, Bagnou JH, et al. Social performance rating during social skills training in adults with autism spectrum disorder and schizophrenia. Presented at: 2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW); Sep 10-13, 2023; Cambridge, MA, USA p. 1-8. [doi: 10.1109/ACIIW59127.2023.10388083]

XSL•FO
RenderX

25. Likert R. A Technique for the Measurement of Attitudes Archives of Psychology de Marchena A: Columbia University; 1932.
26. Wild D, Grove A, Martin M, et al. Principles of good practice for the translation and cultural adaptation process for patient-reported outcomes (PRO) measures: report of the ISPOR Task Force for Translation and Cultural Adaptation. Value Health 2005;8(2):94-104. [doi: 10.1111/j.1524-4733.2005.04054.x] [Medline: 15804318]
27. Tulsky DS, Chiaravalloti ND, Palmer BW, et al. The Wechsler Memory Scale: a new perspective. In: Clinical Interpretation of the WAIS-III and WMS-III: Elsevier; 2003:93-139. [doi: 10.1016/B978-012703570-3/50007-9]
28. Horan WP, Kern RS, Shokat-Fadai K, Sergi MJ, Wynn JK, Green MF. Social cognitive skills training in schizophrenia: an initial efficacy study of stabilized outpatients. Schizophr Res 2009 Jan;107(1):47-54. [doi: 10.1016/j.schres.2008.09.006] [Medline: 18930378]
29. Kikuchi A. Measure social skill: handbook of kiss-18 (110-115). : Toukyou: Sankou Digital Production Press; 2007.
30. Singelis TM, Sharkey WF. Culture, self-construal, and embarrassability. J Cross Cult Psychol 1995 Nov;26(6):622-644. [doi: 10.1177/002202219502600607]
31. Constantino JN, Gruber CP. Social Responsiveness Scale: SRS-2: Western Psychological Services; 2012.
32. Kamio Y, Inada N, Moriwaki A, et al. Quantitative autistic traits ascertained in a national survey of 22 529 Japanese schoolchildren. Acta Psychiatr Scand 2013 Jul;128(1):45-53. [doi: 10.1111/acps.12034] [Medline: 23171198]
33. Kaneda Y, Sumiyoshi T, Keefe R, Ishimoto Y, Numata S, Ohmori T. Brief assessment of cognition in schizophrenia: validation of the Japanese version. Psychiatry Clin Neurosci 2007 Dec;61(6):602-609. [doi: 10.1111/j.1440-1819.2007.01725.x] [Medline: 18081619]
34. Kay SR, Fiszbein A, Opler LA. The positive and negative syndrome scale (PANSS) for schizophrenia. Schizophr Bull 1987;13(2):261-276. [doi: 10.1093/schbul/13.2.261] [Medline: 3616518]
35. Baltrusaitis T, Zadeh A, Lim YC, Morency LP. OpenFace 2.0: facial behavior analysis toolkit. Presented at: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018); May 15-19, 2018; Xi'an, China p. 59-66. [doi: 10.1109/FG.2018.00019]
36. Yang L, Woo J, Achard C, Pelachaud C. Exchanging... watch out!. arXiv. Preprint posted online on Nov 8, 2023. [doi: 10.48550/arXiv.2311.04747]
37. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979 Mar;86(2):420-428. [doi: 10.1037//0033-2909.86.2.420] [Medline: 18839484]
38. Costello AB, Osborne J. Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. Pract Assess Res Eval 2005;10(1). [doi: 10.7275/jyj1-4868]
39. Beavers AS, Lounsbury JW, Richards JK, et al. Practical considerations for using exploratory factor analysis in educational research. Pract Assess Res Eval 2013;18(1):6. [doi: 10.7275/qv2q-rk76]
40. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951 Sep;16(3):297-334. [doi: 10.1007/BF02310555]
41. McDonald RP. Test Theory: A Unified Treatment: Psychology Press; 2013.
42. JASP (version 0.17). JASP. 2023. URL: https://jasp-stats.org/ [accessed 2024-12-23]
43. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016 Jun;15(2):155-163. [doi: 10.1016/j.jcm.2016.02.012] [Medline: 27330520]
44. Shiino T, Miura K, Fujimoto M, et al. Comparison of eye movements in schizophrenia and autism spectrum disorder. Neuropsychopharmacol Rep 2020 Mar;40(1):92-95. [doi: 10.1002/npr2.12085] [Medline: 31774635]
45. Kam C, Zhou X, Zhang X, Ho MY. Examining the dimensionality of self-construals and individualistic–collectivistic values with random intercept item factor analysis. Pers Individ Dif 2012 Oct;53(6):727-733. [doi: 10.1016/j.paid.2012.05.023]
46. Markus HR, Kitayama S. Culture and the self: implications for cognition, emotion, and motivation. Psychol Rev 1991;98(2):224-253. [doi: 10.1037//0033-295X.98.2.224]
47. Uzun G, Lok N. Effect of emotional awareness skills training on emotional awareness and communication skills in patients with schizophrenia: a randomized controlled trial. Arch Psychiatr Nurs 2022 Jun;38:14-20. [doi: 10.1016/j.apnu.2022.01.005] [Medline: 35461642]
48. Robertson BR, Prestia D, Twamley EW, Patterson TL, Bowie CR, Harvey PD. Social competence versus negative symptoms as predictors of real world social functioning in schizophrenia. Schizophr Res 2014 Dec;160(1-3):136-141. [doi: 10.1016/j.schres.2014.10.037] [Medline: 25468184]

Abbreviations

ADOS-2: Autism Diagnostic Observation Schedule, Second Edition
ASD: autism spectrum disorder
BACS-J: Brief Assessment of Cognition in Schizophrenia
FEIT: Facial Emotion Identification Test
ICC: intraclass correlation coefficient
JASP : Jeffreys's Amazing Statistics Program

XSL•FO
RenderX

Kiss-18: Kikuchi's scale of social skills with 18 items
PANSS: Positive and Negative Syndrome Scale
SPRS: Social Performance Rating Scale
SRS-2: Social Responsiveness Scale
SST: social skills training

Edited by A Mavragani; submitted 08.04.24; peer-reviewed by P Harvey; revised version received 04.11.24; accepted 06.11.24; published 10.01.25.

Please cite as:
Tanaka H, Miyamoto K, Hamet Bagnou J, Prigent E, Clavel C, Martin JC, Nakamura S
Analysis of Social Performance and Action Units During Social Skills Training: Focus Group Study of Adults With Autism Spectrum Disorder and Schizophrenia
JMIR Form Res 2025;9:e59261
URL: https://formative.jmir.org/2025/1/e59261
doi:10.2196/59261

© Hiroki Tanaka, Kana Miyamoto, Jennifer Hamet Bagnou, Elise Prigent, Céline Clavel, Jean-Claude Martin, Satoshi Nakamura. Originally published in JMIR Formative Research (https://formative.jmir.org), 10.1.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Unified Mobile App for Streamlining Verbal Autopsy and Cause of Death Assignment in India: Design and Development Study

Harleen Kaur[1*], MSc; Stuti Tripathi[1*], MSc; Manjeet Singh Chalga[2*], PhD; Sudhir K Benara[1], MD; Amit Dhiman[2], BTECH; Shefali Gupta[1], MPH; Saritha Nair[2], PhD; Geetha Menon[1], PhD; B K Gulati[1], PhD; Sandeep Sharma[3], MD; Saurabh Sharma[1], MD

[1]ICMR-National Institute for Research in Digital Health and Data Science, Ansari Nagar, New Delhi, India

[2]Indian Council of Medical Research, New Delhi, India

[3]ICMR-National JALMA Institute for Leprosy & Other Mycobacterial Diseases, Agra, India

[*]these authors contributed equally

Corresponding Author:
Saurabh Sharma, MD
ICMR-National Institute for Research in Digital Health and Data Science, Ansari Nagar, New Delhi, India

Abstract

Background: Verbal autopsy (VA) has been a crucial tool in ascertaining population-level cause of death (COD) estimates, specifically in countries where medical certification of COD is relatively limited. The World Health Organization has released an updated instrument (Verbal Autopsy Instrument 2022) that supports electronic data collection methods along with analytical software for assigning COD. This questionnaire encompasses the primary signs and symptoms associated with prevalent diseases across all age groups. Traditional methods have primarily involved paper-based questionnaires and physician-coded approaches for COD assignment, which is time-consuming and resource-intensive. Although computer-coded algorithms have advanced the COD assignment process, data collection in densely populated countries like India remains a logistical challenge.

Objective: This study aimed to develop an Android-based mobile app specifically tailored for streamlining VA data collection by leveraging the existing Indian public health workforce. The app has been designed to integrate real-time data collection by frontline health workers and seamless data transmission and digital reporting of COD by physicians. This process aimed to enhance the efficiency and accuracy of COD assignment through VA.

Methods: The app was developed using Android Studio, the primary integrated development environment for developing Android apps using Java. The front-end interface was developed using XML, while SQLite and MySQL were employed to streamline complete data storage on the local and server databases, respectively. The communication between the app and the server was facilitated through a PHP application programming interface to synchronize data from the local to the server database. The complete prototype was specifically built to reduce manual intervention and automate VA data collection.

Results: The app was developed to align with the current Indian public health system for district-level COD estimation. By leveraging this mobile app, the average duration required for VA data collection to ascertainment of COD, which typically ranges from 6 to 8 months, is expected to decrease by approximately 80%, reducing it to about 1 - 2 months. Based on annual caseload projections, the smallest administrative public health unit, health and wellness centers, is anticipated to handle 35 - 40 VA cases annually, while medical officers at primary health centers are projected to manage 150 - 200 physician-certified VAs each year. The app's data collection and transmission efficiency were further improved based on feedback from user and subject area experts.

Conclusions: The development of a unified mobile app could streamline the VA process, enabling the generation of accurate national and subnational COD estimates. This mobile app can be further piloted and scaled to different regions to integrate the automated VA model into the existing public health system for generating comprehensive mortality statistics in India.

(JMIR Form Res 2025;9:e59937) doi:10.2196/59937

KEYWORDS

verbal autopsy; cause of death; mortality; mHealth; public health; India; mobile health

Introduction

Effective and efficient public health surveillance systems rely heavily on communication, collaboration, and information sharing. Improving traditional data collection and reporting methods is essential for supporting surveillance and response efforts. Cause-specific mortality is an important component in planning and implementing public health interventions. Digital technologies, especially in India, have been instrumental in achieving these goals while advancing the vision of transforming

XSL•FO
RenderX

into a digitally empowered society. This revolution has accelerated the growth of health care services. These measures enable public health professionals across various regions to collaborate and communicate effectively while gathering crucial information efficiently, thereby facilitating rapid and targeted interventions by generating and disseminating public health information [1].

Cause of death (COD) information is essential for evaluating health interventions and programs at all levels. Ideally, civil registration systems provide relevant information related to births, deaths, and COD in any country [2-4]. However, in countries with a large population such as India, obtaining comprehensive and reliable information on COD has been a major challenge. The death registration rates in India have shown considerable progress with approximately 93% of estimated deaths officially registered. However, only 19% of these registered deaths have a medically certified cause of death (MCCD) [5]. To address this gap, research studies have suggested that detailed household interviews with family members of the deceased, known as verbal autopsy (VA), can improve the quality of information on COD [6-8]. The primary aim of VA is to ascertain population-level COD estimates. In a VA, the COD is ascertained based on an interview with next of kin or other caregivers using a standardized questionnaire that gathers information on signs, symptoms, medical history, and circumstances preceding death [9]. Data collected during these interviews is then processed and analyzed to assign a probable COD. Standard VA instruments include a VA questionnaire, a list of COD, and diagnostic criteria (based on either expert or data-driven algorithms) for assigning a COD. The World Health Organization's (WHO) VA instrument contains a concise list of COD that are of public health importance. These can be determined through a significant number of questions designed for use in VA interviews and with automated or physician-coded COD assignment using analytical software. The Office of the Registrar General of India has been successfully conducting VA since 1999, with the help of trained physicians to assign COD [4,10].

The existing WHO VA Instrument is compatible with the Open Data Kit (ODK) Android app and has been tested for feasibility in the Indian context [11]. The ODK app is designed to be used in various scenarios such as surveillance, surveys, research, and crisis response. The current ODK app provides diverse features such as reducing data entry costs and preventing missing or invalid data from VA questionnaires. However, the development of mobile apps often stems from specific requirements for national-level VA studies and surveillance programs. This study by the Indian Council of Medical Research–National Institute for Research in Digital Health and Data Science (ICMR-NIRDHDS) details the development of a unified mobile app to streamline the data collection process during VA by using the existing public health workforce. This app is designed to support the line listing of death records, with further segregation of data by user and field catchment area, which enhances scalability and efficiency in data management. As the volume of cases increases, integration of these capabilities directly into the app helps prevent bottlenecks in data collection and analysis, streamlining the entire process. Additionally, the app includes automated data transfer for physician-coded verbal autopsy (PCVA), reducing the need for manual data extraction and sharing, which accelerates data processing and minimizes delays. It could facilitate streamlining of the data collection process and enable physicians to assign COD within a single app, completely integrated with the public health system from the subcenters to district hospitals.

Methods

Study Design

The WHO's VA instrument was tested for its feasibility in the Indian context by using the computer-assisted personal interviewing device-based data collection method. The ICMR-NIRDHDS has thus taken the initiative to develop a mobile app to streamline VA data collection, which may be integrated into the existing public health system. A graphical example depicting the comparison of the existing ODK-based VA tool and the developed unified mobile app is shown in Figure 1.

XSL•FO
RenderX

Figure 1. Step-by-step workflow for cause of death assignment using the unified mobile app, highlighting data collection, verification, and physician review. ANM: auxiliary nursing midwife; CHC: community health care; CoD: cause of death; ODK: Open Data Kit; PCVA: physician-coded verbal autopsy; PHC: primary health care; VA: verbal autopsy.

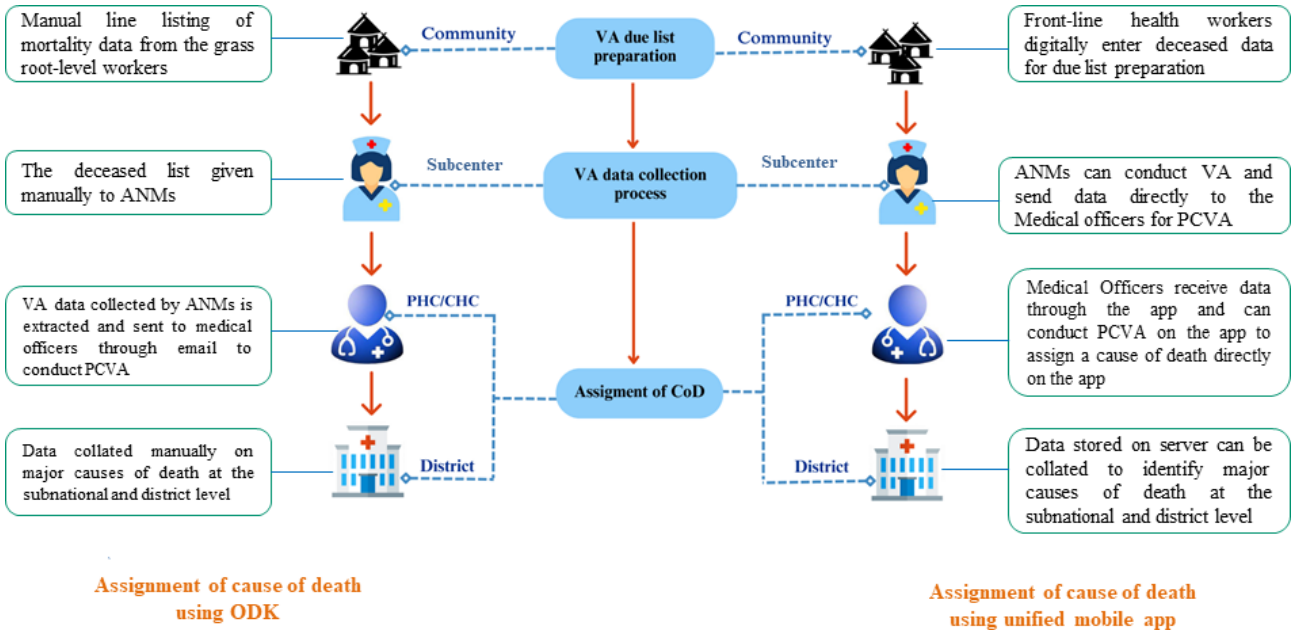

Phase 1: Ideation, Planning

Reliable cause-specific mortality statistics are required by administrators, policy planners, and researchers for evidence-based decision-making regularly. Assigning COD using PCVA and computer-coded VA methods has already been tested in urban as well as rural areas of India using the existing ODK app [11]. However, this approach involved manual intervention at various stages of the data collection process. Initial automated line listing of death records by accredited social health activist (ASHA) workers is not supported by the ODK and is instead maintained manually, followed by subsequent VAs conducted by auxiliary nursing midwives (ANMs), community health officers, and field investigators. Data collected during the VA process was then extracted from the ODK platform and transferred via email to medical officers for COD assignment. Manual interventions requiring an extensive workforce in large-scale studies are time-consuming and can potentially compromise the data quality and subsequent outcomes. Therefore, this study aimed to develop a unified app for eliminating the manual interventions at each stage of VA data collection and COD assignment, from death reporting to conducting PCVA in rural settings. Data collected using mobile or handheld devices such as computer-assisted personal interviewing tablets could also be transferred electronically to physicians, enabling them to assign COD efficiently. In addition to providing community-based perspectives on rural mortality, it also served as a basis for testing VA methods on an operational level. This app also offers offline data collection without an internet connection and limited resources. Hence, the primary goal of the study was to efficiently implement the WHO VA 2022 instrument in rural settings through the existing public health system, with plans for further expansion to integrate automated algorithms for assigning COD.

Phase 2: App Development

Having conceptualized the solution, the next step was the development of the VA app by the team of ICMR-NIRDHDS. It was developed using Android Studio, which is the primary IDE for developing Android apps. This app was based on the WHO 2022 VA questionnaire comprising three modules, for example, adult, children, and neonate (data elements' details are given in Multimedia Appendix 1). The entire app was developed using Java for programming and XML for designing the front-end interface. Data storage was streamlined using the MySQL database management system on a secure and authenticated server, while SQLite was used for local data storage within the app. The initial submission of data by each user is stored on the local database. Next, the user selects the Sync button to update data on the centralized server. Communication between the app and server was facilitated through a PHP application programming interface (API). The API acts as a bridge between the mobile app and the server to facilitate data synchronization. During this process, the app sends HTTP requests to the server using specific PHP API end points. These end points were designed to receive the submitted data, perform validation and sanitization, and subsequently insert or update the records in the MySQL server database. After data submission, the server response is sent back to the mobile app to update the corresponding local database records. Given the data complexity, we opted for a relational database model for comprehensive data analysis and to ensure structured data organization and efficient retrieval. Further, the developed app allowed the data collectors to partially complete the surveys and resume them at a later stage or edit responses after the initial submission of the survey. This facilitated standardized data collection for subsequent review by physicians involved in the study. The complete prototype was designed with the ability to capture data with minimal cognitive effort. The navigational structure depicting the complete network is shown in Figure 2. Snapshots of the user-based interface are shown in Figure 3.

XSL•FO
RenderX

Figure 2. Diagram depicting the step-by-step data flow from user input in the mobile app to the central database, highlighting the distinct pathways for different user roles. ANM: auxiliary nursing midwife; API: application programming interface; ASHA: accredited social health worker; CoD: cause of death; VA: verbal autopsy.

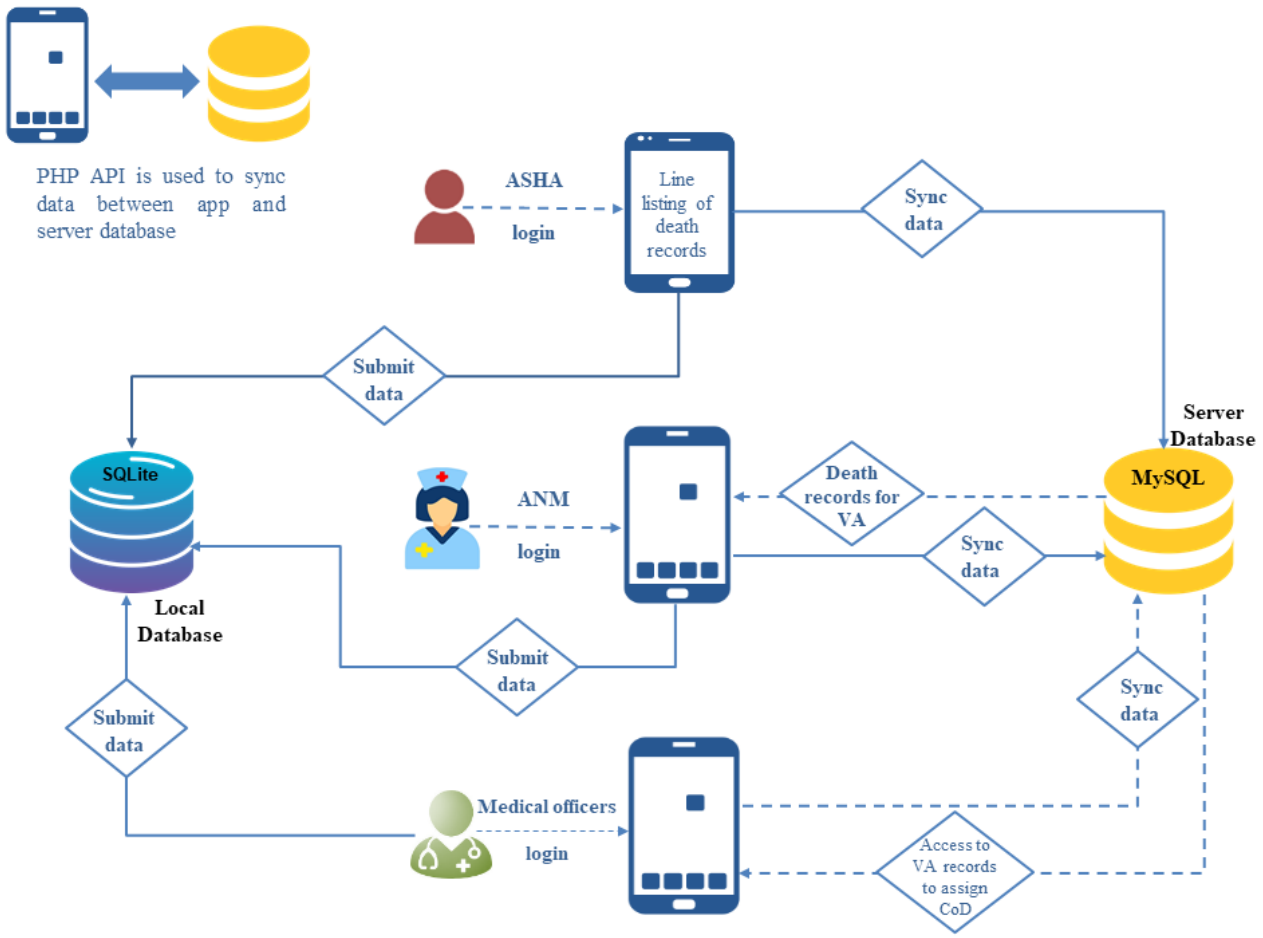

XSL•FO
RenderX

Figure 3. Customized app interface designed for specific user roles to streamline tasks for accredited social health workers, auxiliary nursing midwives, and medical officers.

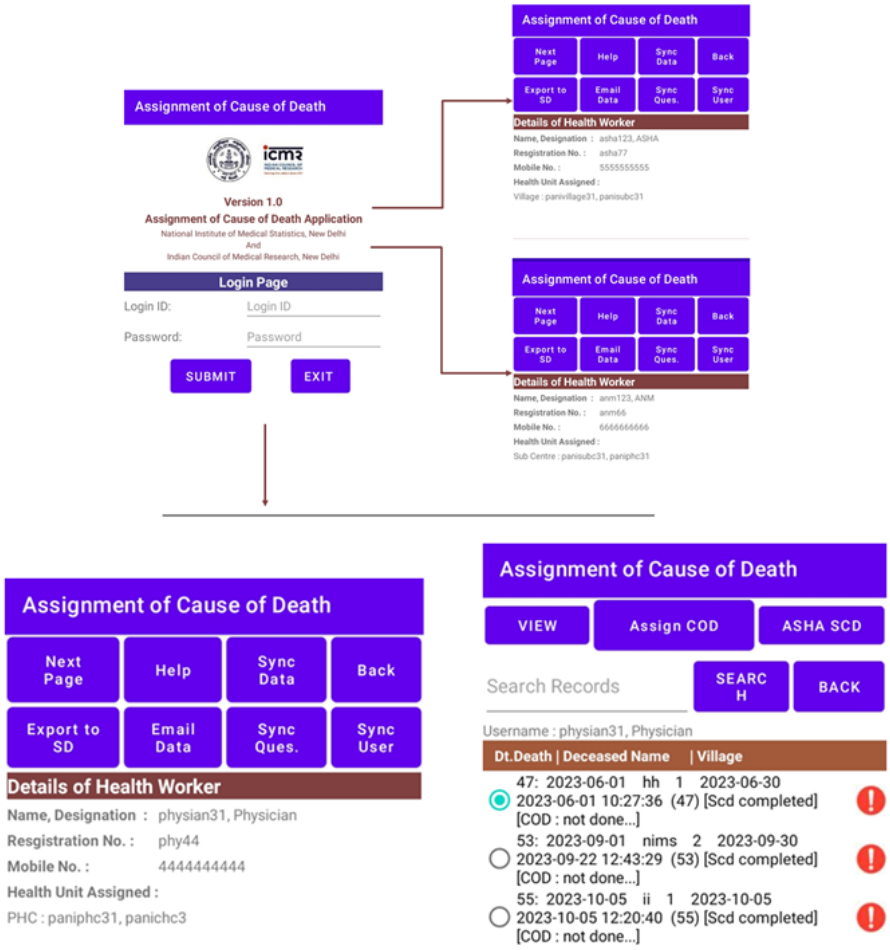

Phase 3: Data Management, Testing, and Quality Assurance

In this phase, the mobile app underwent field testing in both English and Hindi as an integral part of the methodology. Using APIs and servers for data communication ensured effective and secure data management, maintaining data integrity. Multiple data validation checks, skip patterns, and quality assurance measures were incorporated into the survey to maintain the accuracy and integrity of the data.

Ethical Considerations

The study was reviewed and approved by the ICMR-NIMS Ethics Committee on Health Research (ICMR_NIMS_2020). Informed consent was obtained from all subjects involved in the study. Data security and patient privacy were prioritized throughout the entire process. The app adheres to strict security standards, with access controls and user authentication. Patient identifiers are anonymized to protect their identities. Subject compensation is not applicable in the present study and hence was not provided.

Results

Overview

The mobile app was developed with a focus on the integration and implementation of VA within the existing Indian public health system. Health and wellness centers, which serve approximately 5000 individuals, were projected to manage around 35 - 40 VA cases annually, translating to about 3 - 4 cases per month per center. The ANMs and community health officers, who are the primary health care providers in these centers, are tasked with performing and recording these autopsies efficiently using the app. Similarly, primary health centers catering to larger populations of around 20,000 - 30,000 individuals are projected to handle a higher annual caseload of 150 - 200 VAs, equating to 12 - 16 autopsies per month per center for each medical officer. These projections were based on population densities and historical data on mortality rates in these regions, allowing for more accurate workload estimation. The app was specifically designed with these projections in mind, optimizing the workflow to manage these volumes with ease.

The traditionally lengthy process of VA data collection and COD determination, which generally spans 6 - 8 months, is anticipated to be significantly expedited with this app, reducing this duration by approximately 80%. This would shorten the

XSL•FO
RenderX

time from VA data collection to final COD determination to approximately 1 - 2 months.

To ensure the app's practicality and usability, extensive feedback was gathered from subject matter experts, including public health officials and field workers. This input resulted in several enhancements, such as simplifying data input processes and ensuring compatibility with existing health information systems.

Prototype Evaluation: Experts' Insights and Recommendations

The unified mobile app used in this study was developed according to the outlined objectives and methodologies. It was critically reviewed by the subject experts and its contents were refined according to their feedback and suggestions until final approval. Therefore, the developed app has automated the VA process and has been rated as efficient, user-friendly, and easy to navigate by the users and various stakeholders, strengthening the existing public health system. Textbox 1 lists the recommendations by the subject experts.

Textbox 1. Experts' recommendations implemented in the app.

Content

- Customization of the user interface for ease of use by frontline health workers
- Addition of detailed information of the health care workers along with the demarcated health care centers, enabling the generation of an area-wise due list for conducting the verbal autopsy (VA)
- Incorporation of the list of cause of death with VA codes, which can be assigned by medical officers

Interaction and workflow pathways

- Streamlining of workflow pathways for frontline workers (eg, accredited social health activist [ASHA], auxiliary nursing midwife [ANM]) and medical officers
- Transition from manual line listing to digital due list preparation of deceased data
- Addressing the language barrier by facilitating bilingual VA process in both English and Hindi
- Creation of a user-friendly interface with role-based access defined for each user group (eg, ANM, ASHA, medical officers)

App Overview

Figure 3 illustrates the app interface, beginning with the login module accessible only to authorized users. Role-based login credentials for specific roles such as ANM, ASHA, and medical officers will allow them to navigate to specific sections on the app. ASHA workers will be able to digitally enter the deceased data according to the demarcated health centers. This digitized data collection marks a significant shift from traditional manual methods, ensuring accuracy and minimizing errors.

This stored data will be further accessible to ANMs to conduct a VA using a bilingual interface (in Hindi and English), according to their feasibility. The VA process allows for partial saving of information, which can be completed at a later time—an important feature for real-world scenarios where time constraints are often encountered. Once ANMs sync the VA data, it becomes available for medical officers to review and assign a COD, facilitated by the integration of VA codes directly in the app.

The app's robust design, which is grounded in expert feedback, facilitates an integrated and real-time approach to VA and COD assignment.

Discussion

Overall death registration rates have shown significant improvement in India. However, the MCCD currently covers only a smaller percentage of deaths compared to the total registered deaths [12-14]. At present, the Office of the Registrar General of India conducts VA for a sample of deaths through the Sample Registration System survey [15]. VA holds significant potential to supplement MCCD and strengthen the mortality statistics at the population level [16-18]. In the conventional system, VA data collection involves a multistep, manual process where health workers fill out paper-based questionnaires and rely on physicians to review and assign causes of death, often leading to substantial delays.

The existing WHO VA instrument is compatible with the ODK software and has undergone feasibility testing in the Indian context. However, this app is primarily designed for conducting VA and requires manual intervention at each stage, which limits its flexibility and adaptability across different states. To scale up the existing VA survey, it could be integrated with the existing public health system through the grassroots-level workforces and digital health technology [9,19]. Over the past decade, the role of digital health technologies in public health has been increasing [20]. In India, several digital health initiatives, such as doctor consultation apps, e-pharmacies, and e-Sanjeevani, were developed to strengthen the existing public health system [21-24]. As an advancement, the institute has developed a unified mobile app to streamline the COD assignment process via VA. By utilizing the public health system, this mobile app for VA can act as a stepping stone toward addressing the challenge of timely and accessible mortality data collection in India. With the increasing penetration of mobile devices across urban and rural areas, this app has the potential to reach diverse populations, including those in remote regions. This study was able to demonstrate that a customized smartphone app for VA data collection, data

XSL•FO
RenderX

transfer, and COD identification can be used for disease burden estimation at subpopulation levels. The app is further scalable for use in larger populations, including urban regions and more tertiary hospitals. Additionally, India is characterized by its cultural and linguistic diversity, presenting a challenge in obtaining accurate VA data. The app can further accommodate multiple languages and aid in overcoming this barrier.

The mobile app developed comes with several key features and strengths: (1) it assists inline listing of the cases and due list preparation; (2) frontline workers can conduct VA through this mobile app; (3) VA cases can be automatically assigned based on the field catchment area; (4) physicians can retrieve information from the server and conduct PCVA; (5) it is bilingual, supporting both English and Hindi; (6) it has the potential to substantially improve the timeliness of mortality data reporting at both the local and national levels by reducing the overall time from 6 to 8 months to just 1 - 2 months; (7) the mobile app can be customized based on the variability of public health organizational structure across different states; (8) the app could accommodate multiple languages; (9) the app could be functional in remote areas with limited internet connectivity, as it allows users to transfer and retrieve data from the server offline; and (10) health care workers can retrieve data from the server at any point of time, enhancing the timely accessibility of the information.

However, mobile health implementation is often encountered with challenges [25,26]. Acquiring and maintaining hardware is difficult due to financial limitations and inadequate technical expertise. Developing the complete workflow from data collection to assigning the COD for each user was also challenging initially. However, it improved gradually through multiple discussions with various stakeholders. Additionally, Android apps may face compatibility issues across different devices and operating system versions, leading to inconsistent performance and user experience.

In conclusion, the development of a unified mobile app for streamlining VA and COD assignment through the public health system in India could act as a stepping stone for generating national and subnational level COD estimates. Feasibility studies based on this mobile app need to be undertaken across different regions to develop a VA model, which can be integrated into the public health system for generating mortality statistics in India.

Acknowledgments

The authors acknowledge the Indian Council of Medical Research, Delhi, for providing financial assistance for the study. This work was supported by the Indian Council of Medical Research, New Delhi (grant ID 5/4/8-32/CD/SS/2021-NCD-II).

Data Availability

All data generated or analyzed during this study are included in this published article.

Authors' Contributions

Conceptualization: Saurabh Sharma
App development: MSC, AD
App testing: AD
Graphics: HK, ST
Writing (original draft): HK, ST, Saurabh Sharma
Writing – review and editing: MSC, SKB, SG, SN, GM, BKG, Sandeep Sharma

Conflicts of Interest

None declared.

Multimedia Appendix 1
Data elements of World Health Organization verbal autopsy questionnaire for adults, children, and neonates.
[PDF File, 119 KB - formative_v9i1e59937_app1.pdf]

References

1. Mukhi S, Dhiravani K, Micholson B, et al. An innovative mobile data collection technology for public health in a field setting. Online J Public Health Inform 2018;10(2):e202. [doi: 10.5210/ojphi.v10i2.9114] [Medline: 30349620]
2. Vital statistics of India based on the civil registration system 2016. Census India. URL: https://censusindia.gov.in/nada/index.php/catalog/42538 [accessed 2024-04-24]
3. Medical certification of cause of death. Delhi Planning. URL: https://delhiplanning.delhi.gov.in/sites/default/files/report_on_medical_certification_of_cause_of_deaths_in_delhi_2016.pdf [accessed 2024-04-24]
4. Mahapatra P, Chalapati Rao PV. Cause of death reporting systems in India: a performance analysis. Natl Med J India 2001;14(3):154-162. [Medline: 11467144]

XSL•FO
RenderX

5. Kumar R. Mortality statistics in India: current status and future prospects. Indian J Community Med 2022;47(4):476-478. [doi: 10.4103/ijcm.ijcm_614_22] [Medline: 36742978]
6. Byass P, Huong DL, Minh HV. A probabilistic approach to interpreting verbal autopsies: methodology and preliminary validation in Vietnam. Scand J Public Health Suppl 2003;62:32-37. [doi: 10.1080/14034950310015086] [Medline: 14649636]
7. Rao C, Bradshaw D, Mathers CD. Improving death registration and statistics in developing countries: lessons from sub-Saharan Africa. S Afr J Demogr 2004;9(2):81-99. [doi: 10.2307/20853272]
8. Fottrell E, Byass P. Verbal autopsy: methods in transition. Epidemiol Rev 2010;32:38-55. [doi: 10.1093/epirev/mxq003] [Medline: 20203105]
9. Verbal autopsy standards: the 2022 WHO verbal autopsy instrument. World Health Organization. URL: https://cdn.who.int/media/docs/default-source/classification/other-classifications/autopsy/2022-va-instrument/verbal-autopsy-standards_2022-who-verbal-autopsy-instrument_v1_final.pdf [accessed 2024-04-18]
10. Jha P, Gajalakshmi V, Gupta PC, et al. RGI-CGHR Prospective study of one million deaths in India: rationale, design, and validation results. PLoS Med 2006 Feb;3(2):e18. [doi: 10.1371/journal.pmed.0030018] [Medline: 16354108]
11. Benara SK, Sharma S, Juneja A, et al. Evaluation of methods for assigning causes of death from verbal autopsies in India. Front Big Data 2023;6:1197471. [doi: 10.3389/fdata.2023.1197471] [Medline: 37693847]
12. Report on medical certification of cause of deaths in Delhi-2019. Government of the National Capital Territory of Delhi. URL: https://des.delhi.gov.in/sites/default/files/DES/generic_multiple_files/final_mccd_2019.pdf [accessed 2024-04-24]
13. Saikia N, Kumar K, Das B. Death registration coverage 2019-2021, India. Bull World Health Organ 2023 Feb 1;101(2):102-110. [doi: 10.2471/BLT.22.288889] [Medline: 36733620]
14. Ray S, Das DK. Verbal autopsy - concept and role in Indian scenario. J Comp Health 2019;7(2):12-15. [doi: 10.53553/JCH.v07i02.003]
15. Krishnan A, Gupta V, Nongkynrih B, et al. Mortality in India established through verbal autopsies (MINErVA): strengthening national mortality surveillance system in India. J Glob Health 2020 Dec;10(2):020431. [doi: 10.7189/jogh.10.020431] [Medline: 33282224]
16. Polprasert W, Rao C, Adair T, Pattaraarchachai J, Porapakkham Y, Lopez AD. Cause-of-death ascertainment for deaths that occur outside hospitals in Thailand: application of verbal autopsy methods. Popul Health Metr 2010 May 18;8:13. [doi: 10.1186/1478-7954-8-13] [Medline: 20482760]
17. Danso SO, Manu A, Fenty J, et al. Population cause of death estimation using verbal autopsy methods in large-scale field trials of maternal and child health: lessons learned from a 20-year research collaboration in Central Ghana. Emerg Themes Epidemiol 2023 Feb 16;20(1):1. [doi: 10.1186/s12982-023-00120-7] [Medline: 36797732]
18. Firth SM, Hart JD, Reeve M, et al. Integrating community-based verbal autopsy into civil registration and vital statistics: lessons learnt from five countries. BMJ Glob Health 2021 Nov;6(11):e006760. [doi: 10.1136/bmjgh-2021-006760] [Medline: 34728477]
19. de Savigny D, Riley I, Chandramohan D, et al. Integrating community-based verbal autopsy into civil registration and vital statistics (CRVS): system-level considerations. Glob Health Action 2017;10(1):1272882. [doi: 10.1080/16549716.2017.1272882] [Medline: 28137194]
20. Iyamu I, Xu AXT, Gómez-Ramírez O, et al. Defining digital public health and the role of digitization, digitalization, and digital transformation: scoping review. JMIR Public Health Surveill 2021 Nov 26;7(11):e30399. [doi: 10.2196/30399] [Medline: 34842555]
21. Wienert J, Jahnel T, Maaß L. What are digital public health interventions? First steps toward a definition and an intervention classification framework. J Med Internet Res 2022 Jun 28;24(6):e31921. [doi: 10.2196/31921] [Medline: 35763320]
22. Maroju RG, Choudhari SG, Shaikh MK, Borkar SK, Mendhe H. Role of telemedicine and digital technology in public health in India: a narrative review. Cureus 2023 Mar;15(3):e35986. [doi: 10.7759/cureus.35986] [Medline: 37050980]
23. Markan A, Kishore A, Agarwal A, et al. Demographic profile of patients seeking teleophthalmology consultations through e-Sanjeevani: retrospective analysis of 5138 patients from North India. Indian J Ophthalmol 2022 Dec;70(12):4238-4243. [doi: 10.4103/ijo.IJO_781_22] [Medline: 36453323]
24. Agarwal N, Biswas B. Doctor consultation through mobile applications in India: an overview, challenges and the way forward. Healthc Inform Res 2020 Apr;26(2):153-158. [doi: 10.4258/hir.2020.26.2.153] [Medline: 32547812]
25. Flaxman AD, Stewart A, Joseph JC, et al. Collecting verbal autopsies: improving and streamlining data collection processes using electronic tablets. Popul Health Metr 2018 Feb 1;16(1):3. [doi: 10.1186/s12963-018-0161-9] [Medline: 29391038]
26. University of Melbourne. Challenges associated with automated VA training and rollout. Bloomberg Philanthropies: Data for Health. 2018. URL: https://data4healthlibrary.org/sites/default/files/resources/104_UMelbourne_Challenges%20associated%20with%20automated%20VA%20training%20and%20rollout.pdf [accessed 2024-08-24]

Abbreviations

ANM: auxiliary nursing midwife
API: application programming interface

XSL•FO
RenderX

ASHA: accredited social health activist
COD: cause of death
ICMR: Indian Council of Medical Research
MCCD: medically certified cause of death
NIRDHDS: National Institute for Research in Digital Health and Data Science
ODK: Open Data Kit
PCVA: physician-coded verbal autopsy
VA: verbal autopsy
WHO: World Health Organization

Edited by A Mavragani; submitted 26.04.24; peer-reviewed by S Pal, VK Kamal; revised version received 09.10.24; accepted 15.10.24; published 10.01.25.

Please cite as:
Kaur H, Tripathi S, Chalga MS, Benara SK, Dhiman A, Gupta S, Nair S, Menon G, Gulati BK, Sharma S, Sharma S
Unified Mobile App for Streamlining Verbal Autopsy and Cause of Death Assignment in India: Design and Development Study
JMIR Form Res 2025;9:e59937
URL: https://formative.jmir.org/2025/1/e59937
doi:10.2196/59937

©Harleen Kaur, Stuti Tripathi, Manjeet Singh Chalga, Sudhir K Benara, Amit Dhiman, Shefali Gupta, Saritha Nair, Geetha Menon, B K Gulati, Sandeep Sharma, Saurabh Sharma. Originally published in JMIR Formative Research (https://formative.jmir.org), 10.1.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Public Understanding and Expectations of Digital Health Evidence Generation: Focus Group Study

Paulina Bondaronek[1,2], PhD; Jingfeng Li[2], MSc; Henry W W Potts[1], PhD

[1]Institute of Health Informatics, University College London, 222 Euston Road, London, United Kingdom
[2]School of Social Sciences, University of Westminster, London, United Kingdom

Corresponding Author:
Paulina Bondaronek, PhD
Institute of Health Informatics, University College London, 222 Euston Road, London, United Kingdom

Abstract

Background: The rapid proliferation of health apps has not been matched by a comparable growth in scientific evaluations of their effectiveness, particularly for apps available to the public. This gap has prompted ongoing debate about the types of evidence necessary to validate health apps, especially as the perceived risk level varies from wellness tools to diagnostic aids. The perspectives of the general public, who are direct stakeholders, are notably underrepresented in discussions on digital health evidence generation.

Objective: This study aimed to explore public understanding and expectations regarding the evidence required to demonstrate health apps' effectiveness, including at varying levels of health risk.

Methods: A total of 4 focus group discussions were held with UK residents aged 18 years and older, recruited through targeted advertisements to ensure demographic diversity. Participants discussed their views on evidence requirements for 5 hypothetical health apps, ranging from low-risk wellness apps to high-risk diagnostic tools. Focus groups were moderated using a structured guide, and data were analyzed using reflexive thematic analysis to extract common themes.

Results: A total of 5 key themes were established: personal needs, app functionality, social approval, expectations of testing, and authority. Participants relied on personal experiences and social endorsements when judging the effectiveness of low-risk digital health interventions, while making minimal reference to traditional scientific evidence. However, as the perceived risk of an app increased, there was a noticeable shift toward preferring evidence from authoritative sources, such as government or National Health Service endorsements.

Conclusions: The public have a preference for evidence that resonates on a personal level, but also show a heightened demand for authoritative guidance as the potential risk of digital health interventions increases. These perspectives should guide developers, regulators, and policy makers as they balance how to achieve innovation, safety, and public trust in the digital health landscape. Engaging the public in evidence-generation processes and ensuring transparency in app functionality and testing can bridge the gap between public expectations and regulatory standards, fostering trust in digital health technologies.

(JMIR Form Res 2025;9:e56523) doi:10.2196/56523

KEYWORDS

mobile apps; digital health; public expectations; evidence of effectiveness; health risk perception; effectiveness; health risk; health app; public health; well-being; public trust; diagnostic tools; safety; mobile phone

Introduction

The proliferation of health and wellness apps is substantial [1,2], with over 318,000 found in a 2018 study [3]. However, there is a wide gap between the availability and evaluation of digital health, with most publicly available apps not having been evaluated in the scientific literature [3]. They are typically rapidly developed and implemented with little formal evaluation to support their impact claims [4-6]. Many apps have proven not to be effective [7-10].

There are many reasons for this evaluation gap. The rapid advancement of digital health technologies has outpaced the development of comprehensive guidance. Governments have not wanted to hold back innovation around new technologies that promise to reduce health care costs and support economic growth. The dividing line between what should require effectiveness testing and what need not is unclear. This has all contributed to a gray area or "wild west" of regulation, a lack of clear guidelines, standards, and regulatory frameworks for evaluating digital health interventions. Stakeholders, including developers, health care providers, regulators, and users, face difficulties in navigating this uncertain terrain and ensuring that digital products and services meet appropriate standards of evidence.

What evaluation may be appropriate will vary depending on the digital technology under consideration. The National

XSL•FO
RenderX

Institute for Health and Care Excellence Evidence Standards Framework (NICE ESF) [11], for example, offers recommendations of different study methods for different technology types, broadly based on a risk stratification [11]. A distinction is sometimes made between "digital therapeutics," seen as akin to drug treatments and thus coming with rigorous evaluation [12], and "well-being apps," perhaps closer to a self-help book or going to a yoga class and not requiring the same oversight. Well-being apps are seen as less of a concern for various reasons. The direct risks associated with them are small compared with a diagnostic or therapeutic tool. However, we note there can be a significant opportunity cost to public health apps that do not work. If someone uses, for example, a smoking cessation app that is not effective, they may be less inclined to try another that would have worked.

While there have been initiatives to improve the situation (including in the United Kingdom with the NICE ESF and Evaluating Digital Health Products [11,13]; and elsewhere, for example, DiGA (Digitale Gesundheitsanwendungen) in Germany [14]), there is a live debate over what evaluation is appropriate for what digital health products. However, we have seen little in this literature asking what the populace may want. Do the public expect or want more or less evaluation of digital health products?

Public policy [15,16] and research [17] both profess the value of public participation. The technical nature of clinical research, however, presents challenges. Public understanding of clinical research can be poor, although people's perception of their own knowledge is higher [18]. Studies have focused on public understanding of randomized clinical trials, which is generally low, with ideas such as randomization and double blinding being challenging [19], as has been the case for many years [20]. This is significant as poor understanding undermines trial recruitment and public support for research, and may lead to people making poorer treatment choices [21,22].

Expectations of evidence generation will also relate to the public's broader understanding of and belief in digital health. Policy makers may have techno-utopian views, seeing digital interventions as a solution to challenges in health care [23] through a revolution in the collection and use of data [24]. However, the public can be more cautious, being fairly positive about the potential for digital health, but with some concerns, particularly around safety, privacy, and depersonalization through the loss of human contact [25-28]. Attitudes can vary with the novelty and form of the technology involved [26].

Given all this, we aimed to explore the public's understanding, expectations, and preferences regarding the evidence needed to demonstrate the effectiveness of digital health products and services. We also wanted to see how views vary by the nature of the digital health technology, particularly those that are seen as low-risk (well-being app) versus high-risk (diagnostic and therapeutic tools). We chose a focus group method as we thought the discursive nature of focus groups would help people consider a topic that may be relatively novel for them.

Methods

Sample

The research was carried out at the University of Westminster, and we sought to recruit a sample representative of the local population. We designed a recruitment poster outlining the content of the study and how to participate. The study was then advertised on social media (Twitter [subsequently rebranded X] and LinkedIn [Microsoft]) and through the University of Westminster's Black and minority ethnic community engagement, the Black and Minority Ethnic All Members Network.

The inclusion criteria were UK residents aged 18 years and older. A sample framework was used to monitor the recruitment to ensure a diverse sample of participants. Interested participants were asked to fill out screening questions including age, gender, and ethnicity. We assessed levels of digital literacy with 2 previously used questions [29], asking "How difficult is it to use your smartphone [AND] install mobile applications/apps on your smartphone without someone else helping you?", answered on a Likert scale. Participants could choose 1 of 4 meeting times for online focus groups. A total of 26 individuals contacted us, with 21 participants participating, divided into the 4 focus groups. Each participant received a participant information sheet and was told about the study goals and process and what they needed to do. If they had questions about the study, they were told they could ask them directly by email and get answers. The focus group interviews were conducted on July 11-July 12, 2023, through the University of Westminster's Microsoft Teams account. Each focus group lasted 90 minutes.

Ethical Considerations

The study was approved by the Psychology Ethics Committee at the University of Westminster (reference: ETH2223-3175). Participants received a Participant Information Sheet, provided written informed consent (Multimedia Appendix 1) through email, and could ask questions at any stage. Data were anonymized, securely stored, and handled in compliance with institutional data protection policies. Participants received vouchers valued at £20 (US $25.37) each to thank them for participating in the study. They were also provided with a summary of the interview results afterwards and given the opportunity to check and comment.

Data Collection

The focus groups used a moderator's guide. All groups were led by the same researcher (PB) for consistency. The other 2 authors attended all focus groups and took notes. The groups were recorded. We used the Microsoft Teams transcription as a starting point and then corrected this.

The goal throughout the focus groups was to explore participants' perspectives on the evidence or proof required to demonstrate the effectiveness and accuracy of different types of digital health interventions. The moderator used a Microsoft PowerPoint presentation to guide the focus group process. We started the groups by introducing our definition of digital health, providing examples of 3 different technologies—medical imaging in hospitals, treating mental health at home, and

XSL•FO
RenderX

improving fitness—to illustrate the broad spectrum of digital health. We then defined what we meant by evidence, which we loosely defined as "information or facts that support or prove something." We then introduced 5 hypothetical digital health scenarios (alternating the order of scenarios for different groups: 1 - 5; 5 - 1).

The 5 hypothetical digital health scenarios and how we described them are given in Textbox 1.

These scenarios were selected and presented in a hierarchy (reversed for half the groups) to cover the spectrum of potential safety implications. We were informed in our choice by the NICE ESF, but included more well-being apps as these are not covered by the NICE ESF.

Textbox 1. Five hypothetical digital health scenarios to guide the focus group.

Fitness app

- Focuses on promoting good health, specifically for running.
- Tracks running activities—maps the run, records your speed, and logs running session

Fitness app with artificial intelligence

- Same as the fitness app
- Uses sensors to track your heart rate
- Uses artificial intelligence to then make a personalized recommendation as to when or how much you should run based on your age, weight, and heart rate

Mindfulness app

- Helps you deal with everyday stress
- Guided relaxation sessions
- Tracks how you are feeling

Depression treatment app

- Designed for people who have been diagnosed by a doctor as having clinical depression
- Offers a treatment
- Includes thinking exercises for the user to practice

Skin cancer diagnosis app

- Allows users to take a picture of their skin, specifically moles, for a potential skin cancer diagnosis
- Provides a diagnosis of skin cancer from the picture

Data Analysis

We adopted an interpretivist epistemological framework and a social constructionism ontological stance. We used a reflexive thematic analysis [30,31] to interpret recurring themes based on the collected data and generate themes during coding. We used Reflexive Thematic Analysis Reporting Guidelines to guide the reporting of the analysis [32]. We anonymized the transcripts of the 4 focus groups, assigning a code to each participant. We also compared the transcriptions with our contemporary notes. We generated initial codes and applied them to the data, using Microsoft Excel sheets to identify common ground in participants' expectations, concerns, and views on the effectiveness of digital health products. The codes were reviewed through repeated reading and analysis of the overall data. We also paid attention to how the codes related to the different scenarios presented. A total of 5 themes were eventually developed, which we present on a continuum. The codes and themes were reviewed again. Finally, we drew a thematic map for each scenario. We compared the commonalities and differences encoded between the thematic maps for each scenario, with respect to the hierarchy of scenarios (low-risk to high-risk digital health interventions) and our continuum of themes (from individual to community to authority).

Results

Participants

We used a semi–open-ended questionnaire to screen participants. Gender, age, and ethnicity were asked as open-ended questions to provide participants the flexibility of answer (Table 1). In total, 21 people (8 women and 13 men), aged 21 - 50 years, participated in 4 focus groups. In the sample, 7 identified as Black British, 8 as Black other, 3 as White British, 2 as Asian, and 1 as Mixed. Everyone had access to smartphones, but 2 participants reported sometimes having difficulty installing software on their phones. Furthermore, 14 participants were in employment.

XSL•FO
RenderX

Table . Sample characteristics including gender, age, ethnicity, smartphone usage difficulty, and employment status (gender, age, and ethnicity were collected using open-ended questions).

Characteristic		Total (N=21), n
Self-reported gender		
	Women	8
	Men	13
Age group (years)		
	18 - 25	5
	26 - 35	13
	36 - 45	2
	46 - 55	1
Ethnicity		
	Black British	7
	Black other	8
	Asian	2
	White British	3
	Mixed	1
Use smartphone		
	Not difficult	21
Install mobile apps		
	Not difficult	19
	Somewhat difficult	2
Employment status		
	Full-time paid job (31+ hours)	9
	Part-time paid job (<31 hours)	5
	Doing paid work on a self-employed basis or within one's own business	4
	Out of work (more than 6 months)	1
	Unpaid work for a business, community or voluntary organization	1
	Prefer not to say	1

Overview

While we had sought to discuss evidence generation from a regulatory or research perspective, the participants often took a different perspective, discussing what factors influence their choice and use of digital health products and services. The 13 factors were eventually selected and grouped into 5 themes—personal needs, app functionality, social approval, testing expectations, and authority. Figure 1 shows the hierarchy of the 5 themes, from individual to community to authority. The color represents how often each factor was mentioned by the participants in the app scenario with different risk coefficients. The darker the color, the higher the frequency. White means not mentioned.

Across our hierarchy of 5 different digital health products, participants' views changed. As potential health risks increased, there was an increased call for greater testing, and the involvement of authorities and health care professionals. For example, with the skin cancer diagnosis apps, participants preferred digital health products to be recommended by professionals and to be checked with hospital diagnoses.

XSL•FO
RenderX

Figure 1. The expectations of evidence from different scenarios. The heatmap shows the approximate strength and frequency of participants' comments with respect to each scenario, but should not be interpreted as a quantitative analysis. This heatmap shows public expectations for evidence across 5 digital health app types, progressing from low to high risk (fitness to skin cancer diagnosis). Themes—personal needs, functionality, social approval, testing expectations, and authority—are shaded to indicate how frequently each was mentioned. Darker shades represent higher frequency, with expectations shifting from personal experience for low-risk apps to clinical validation for high-risk apps. AI: artificial intelligence.

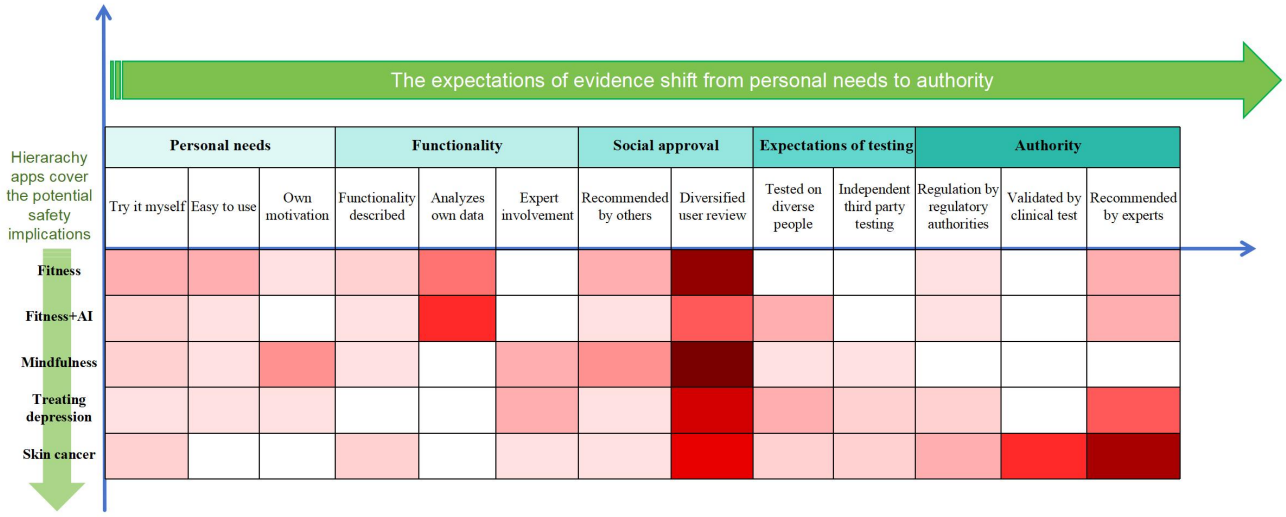

Personal Needs

Most participants first considered trying a digital health product themselves to see if it works for them. They generally reported believing they know their own physical and mental conditions best. Respondents explained that everyone is different, and what works for others may not work for them, so using an app yourself and checking effectiveness was reported to be one of the most reliable ways of evaluating it. Similarly, some participants thought they would only recommend the product to others after using it themselves successfully.

For example, with the fitness app, some participants said they would see the app as effective if they see positive changes in their body after using it:

> *If I were using the fitness app, I would want probably milestones probably after a week after two weeks to three, I should get changes and stuff like that.* [P18]

With the fitness app with artificial intelligence (AI), participants discussed whether the AI's recommendations could be trusted. As P11 said,

> *If I've used it a few times and it's working for me. Then you begin to build up trust, or this is what it can do, and this is what it has trouble doing.* [P11]

With the mindfulness app, as this is related to their daily mental state, participants felt more need to try it for themselves:

> *Like people every day, stress is different like stress is different for everybody, like what might be stressful for me might not be stressful for another individual.* [P18]

Similarly, with the depression treatment app:

> *I think I'll just go on and try it out on my own.* [P21]

Factors, such as ease of use and personal motivation, were highlighted, with many participants expressing that these aspects greatly influence their willingness to adopt and continue using digital health apps. Ease of use seemed to be a fundamental prerequisite for participants.

When talking about personal motivation to use an app, P3 explained,

> *There are many fitness apps. You gotta find the one you really like.* [P3]

When participants felt an app was what they wanted, they would be motivated to continue using it. For mindfulness and depression apps, some participants said that if the app made them feel better, that would mean the app was effective. They would then increase their motivation and become more committed to using the app:

> *I know it's different for everyone and depending, like the other person said, depending on how invested you are in using the app, that's how much of the result you see.* [P14]

Functionality

Participants saw a product's functionality as a determining factor of whether it is effective or not. Before using the fitness or fitness apps with AI, some participants said they would check online information about the product to understand its functions and effects. With the higher risk products, such as the skin cancer diagnosis app, some participants felt that developers demonstrating to the public how the app works could be evidence of the app's effectiveness, such as having videos or explaining within the app how it works. A product's diversity in functions was also seen to be related to its ability to work for diverse populations. For fitness apps, some participants talked about different levels of exercise being supported to meet the needs of different users, and how apps should also have the ability to collect physiological data and analyze data:

> *The application needs to identify the users you know to get data and then be able to process the data and give you a well-informed recommendation.* [P12]

XSL•FO
RenderX

For the other apps, participants wanted to have experts involved in the apps, and for the apps' features to allow users to connect with experts, which would increase their trust in using the app. For example, they wanted the depression treatment app to entail further follow-up by experts according to the degree of depression. Talking about the skin cancer diagnosis app, P19 explained:

> *It will be my only safety because it depends on the professionals attached, whatever the app wants to support.* [P19]

Social Approval

Most participants mentioned social approval as a strong influence on their belief in the effectiveness of an app. They described recommendations from friends and relatives as being critical when using an app. This is, in part, because they have a high degree of familiarity with relatives and friends, know their lifestyles and physical conditions, and this therefore allows them to put relatives' and friends' recommendations in context. As a result, across all the apps, participants trusted the recommendations of relatives and friends more:

> *For instance, if I'm into a particular type of job and the person is also into that type. I could read what everyone's stress goes through together with that, and I could greatly do the same. I could use that as evidence, and I trust such an app*
>
> *[P18]*

After recommendations from friends and family, user reviews were the main factor influencing participants to choose an app. For apps with varying health risks, the user reviews that people wanted to see differed. Participants said they would look at overall user reviews before downloading the fitness app to check its effectiveness. This would make it easier for them to understand what the app does and whether it is right for them. Participants believed that the more people who downloaded and positively reviewed an app, the more likely it was to be effective. Therefore, participants also mentioned the number of downloads as a source of evidence. While discussing the fitness app with AI, 1 participant questioned the authenticity of reviews. Some participants tended to ask questions in the app user's comments section and then use the subsequent interaction and answers from those users to help them decide whether to choose an app:

> *I need evidence cos it's already acting in place of my doctor and already giving me recommendations… I tend to stop and pause for reactions from people and get reviews.* [P17]

Discussing the mindfulness app, some participants felt they needed feedback from people close to them, such as observing how their friends used it to see how the app worked. As for the app to treat depression, 1 participant mentioned that the stories people share on social media are also a way for users to comment. Since they are real user experiences, this kind of comment has impact because

> *we buy because of what people have said, how it's done.* [P11]

Participants found the comments of those who suffer from depression more persuasive, as P9 explained:

> *Seeing that evidence from the person who used this app, the person was suffering from depression, and this kind of treatment covered. So I'll need such things to use this kind of app.* [P9]

Participants wanted to know if different people had used the app and it had worked for them. Different people could refer to differences in ethnicity, sex, age, and lifestyle. Some participants wanted to see physical evidence of results from fitness apps, such as outcomes like weight loss and faster running. We pushed participants to specify the number of people from whom they would like to have some feedback. Many participants considered the experience of 1‑5 people enough to demonstrate effectiveness.

For the AI function of the fitness app, participants were concerned about whether the AI could make accurate recommendations based on the age and gender of different people. Mindfulness apps, they said, need to have been used and tested for effectiveness by various different people, and the number of participants who brought this up was larger than with the other apps. Some participants believed that different lifestyles lead to varying stress levels:

> *I think they have different needs, especially for mindfulness apps. The option to choose a lifestyle would be helpful.*
>
> *[P3]*

Here, participants talked of how 5 or more people's reviews could be used as evidence, with feedback wanted from people with differing levels of depression:

> *If the treatment can cater for different people depending on their diagnosis… other people have used them for treatments, and it's a success for them. I think I'd be OK to use it.* [P14]

Expectations of Testing

Some participants mentioned that developers should conduct tests on different people before launching digital health products and services. What participants meant by different people again refered to differences in various aspects, such as different ethnicities, ages, genders, lifestyles, and health conditions. Depending on the app, participants felt that the focus of the test should vary. P1 believed that the fitness app with AI should also implement

> *Background checks and testing. For a different age group and probably put in some information.* [P1]

P20 thought mindfulness apps "*should cater to various individuals.*" The participants did not evince any familiarity with the details of scientific evaluations.

Being diagnosed with cancer was recognized as having a significant impact on people's lives. Thus, some participants mentioned that the testing for a skin cancer diagnosis app should be mainly to determine the app's accuracy and to improve the richness of the test data. In terms of testing being on a diverse group, for skin cancer diagnosis apps, participants were

XSL•FO
RenderX

concerned about differences in skin color and wanted to see more usage data than for other apps. Furthermore, 1 participant suggested testing on 20 people, 5 of whom had the same skin color as himself.

For apps to treat depression, it was suggested that testing should be conducted strictly following academic experiments to test actual effectiveness:

> *I'm expecting that this app has been used for more, more like a randomised set of people before it was rolled out, and then, you know, a study should be carried out on them before and after the intervention. Then you compare how well, how effective they there's been. So, for me, I think that would be the evidence that we're looking for.* [P15]

Some participants said that the testing of digital health products should be carried out by third-party institutions. This should, first, be an authority in a relevant professional field, such as the National Health Service (NHS) or other professional bodies. Second, it should be independent and unrelated to the developer's business operations. In addition, 1 participant noted for depression apps that such third-party tests will be more trustworthy than the developers' own explanations. Another participant said that skin cancer diagnosis app testing should be used in a real-life setting:

> *I also think the app should have been used to diagnose real-life diseases in some trials before the agency certified it for launch.* [P15]

Information on testing should be made available, for example, within the app. Discussing the skin cancer diagnosis app, a participant said,

> *I think the app should contain information about the testing and trials done to ascertain this app's effectiveness.* [P15]

Authority

Participants mentioned the need for government and NHS regulation for all the apps. Participants believed that government and NHS regulation would increase their trust in the apps' effectiveness. A digital health product vetted by the government or NHS was taken to mean that it has been tried by many people and has worked. Government regulation was seen as most important for skin cancer diagnosis apps because of the *"severity of the condition"* [P20] and the need to avoid misdiagnosis.

Participants felt it was necessary that the skin cancer app's diagnoses were compared with hospital diagnoses to determine the app's accuracy. It was the only app where participants talked about a comparison with a hospital diagnosis during the discussions. To avoid misdiagnosis, participants would not unilaterally believe the app diagnosis:

> *I also think any form of diagnosis that should be done with this app should also be related to the GP and others for clarification to get a better result.* [P13]

The use of the app, they said, should only be an intermediate process, and the effectiveness of the app would ultimately be proved by the hospital diagnosis.

Participants rated professional recommendations as being necessary for the other 4 apps. For fitness apps, recommending professionals referred to experts in this field, such as fitness coaches and sports athletes. For the skin cancer diagnosis app, most participants said only a doctor's approval would lead them to use it. One participant stated,

> *I would also be looking for certification from a cancer body.* [P15]

For depression treatment apps, people would consult their doctors for recommendations:

> *I would actually contact my doctor first and let him know about the app and if he was about to use it and can give me go ahead to use it.* [P1]

Most participants believed that the government or NHS should regulate all digital health products:

> *These apps must be censored and monitored because they involve our well-being. This application acts as a substitute for a visit to our GP. So, we need to be sure of how we administer the application and its reliability. About all the apps.* [P17]

Discussion

Principal Results

This study aimed to explore participants' trust in digital health technology and what evidence the public wants from digital health interventions for increasing health and well-being. Following the NICE ESF's risk hierarchy, different scenarios were designed, including wellness apps and treatment and diagnostic apps. We found that (1) the public's expectations for evidence of effectiveness are intertwined with the factors that influence their choice and use of digital health interventions; (2) the public's expectations for evidence of effectiveness are based more on individual experiences and social approval than scientific evidence; (3) as the perceived risks associated with different apps increase, so to do the public's expectations for the evidence of their effectiveness; and (4) there is a desire for recommendations or certifications from relevant bodies, particularly as the potential risk increases.

Comparison With Previous Work

This study supports the findings that the public understanding of what constitutes high-quality research evidence is low [18-20]. As shown previously [21,22], this can lead to challenges in research support and affect the decision-making related to the digital health selection, uptake and engagement, behavior and, ultimately, health outcomes.

We also saw people exhibiting cautious optimism, recognizing potential benefits while maintaining a healthy scepticism about new technologies, as seen in previous studies [25,26].

This study found that the public prefers to verify the effectiveness of digital interventions through personal experience. There are parallels with some research on earlier generations of digital health technology. Herian and colleagues [33] found that greater experience with health care was associated with lower trust in government to regulate electronic

XSL•FO
RenderX

health records, which suggests that personal experience was valued more. Several studies showed that people seeking health information online relied more on their own experience of websites than credibility assessments [34]. We know more generally that personal choice and personal recommendations are important in people's engagement with digital health interventions [35].

This focus on personal experience supports the view of Pagoto and Bennett [36] that interface design is crucial in the development process of the app. The "person-based" approach emphasizes using qualitative research to understand the psychosocial background of a wide range of target populations during the development process of digital interventions and getting feedback from them to improve applicability and effectiveness [37]. This is consistent with the findings of this study that the public expects greater transparency in the development and testing process, which is beneficial for improving the social acceptance of digital interventions by testing with different people and validating the results.

Implications

The study found that the public's concept of evidence diverges significantly from the scientific community's understanding. While researchers may refer to evidence as data collected from rigorous, controlled settings, such as randomized controlled trials, participants in our focus groups considered personal experiences and anecdotal successes as credible evidence for assessing digital health products.

Public policy [15,16] and research [17] both strongly endorse involving the public. This study calls for a systematic approach to involving the public in digital health evidence generation. To effectively engage the public, it is essential to grasp their level of understanding about the need for evidence generation in digital health and why it matters, that is, the potential for safety issues but also missed opportunities for prevention (for so-called lower risk digital health, such as apps for physical activity). This requires educational methods with a substantial instructive component as well as coproduction, such as citizen juries, and necessitates clear, accessible explanations of evidence generation, effectiveness, and the potential detrimental effects of digital health interventions that are widely used but not evaluated. Educating the public about the need for evidence generation in digital health could increase the pressure and motivation of developers to embed evidence generation practices in digital health. Equally, the pressure of the public may encourage regulators to require minimal evidence generation before publicly available digital health is introduced on the app distribution platforms.

Professional support and education can foster user engagement with digital health technology [38]. Social media platforms present opportunities for public communication and behavioral interventions [39]. Developers and relevant authorities can use diverse strategies, including leveraging social media, to enhance the efficacy of digital health apps for promoting public health behaviors. A detailed description of an app's functionalities, effects, and target population should be provided so that individuals can assess its effectiveness quickly. Using documentaries and press releases can promote transparency in digital health product development and testing processes, enhancing public trust. Using social media platforms to share user reviews, such as inviting long-term clients from diverse backgrounds and age groups to tell their stories or engaging professionals like renowned doctors, experts, and athletes to provide insights, fosters societal acceptance. It is crucial to increase public understanding of the impact of the lack of evidence generation in digital health, such as safety concerns and missed opportunities for health promotion when popularity rather than evidence for effectiveness is the primary factor driving downloads.

Strengths

To the best of our knowledge, this is the first study to explore public expectations and preferences regarding evidence generation for the effectiveness of digital health interventions—a stakeholder group that has been overlooked. The sampling strategy aimed to include those at risk of digital inequalities and included a range of ages, genders, socioeconomic statuses, and ethnicities, showcasing a broad spectrum of views. The various scenarios developed for this study were comprehensive.

Limitations

While the study was open to all UK residents, we recruited mostly from London. With 21 participants, the study may not capture the full range of public opinions. The materials used to guide the focus groups and the results could be used to facilitate the development of a survey to get larger samples, which could provide more robust data. We offered a £20 (US $25.37) voucher, which might have influenced motivations to participate in this study.

Conclusion

This study has explored public expectations regarding the evidence needed to demonstrate the effectiveness of digital health products and services. The public tends to favor evidence that is relatable and directly experienced, such as personal success stories, user reviews, and recommendations from significant others. The public made minimal reference to what researchers would perceive as rigorous evaluation methods. However, there was also support for authoritative recommendations and the independent testing of products. As health risks associated with digital health increased, there was a growing demand for more substantial evidence and authoritative recommendations from health care professionals.

Acknowledgments

Thanks to all the participants who took part in the research. ChatGPT was used to assist with grammar and style refinement.

XSL•FO
RenderX

Data Availability

The datasets generated during and analyzed during this study (consisting of qualitative focus group transcripts) are not publicly available due to the sensitive nature of the data and confidentiality commitments to participants. However, the deidentified data are available for research purposes from the corresponding author on reasonable request.

Conflicts of Interest

PB provides consultancy in digital health development and evaluation for ShiftPartner. HWWP provides or has provided consultancy on digital health evaluation for Flo Health Inc., Prova Health, and Thrive Therapeutic Software Ltd. He has PhD students in the field employed by or previously employed by, and with fees paid by or previously paid by AstraZeneca, Patients Know Best, and BetterPoints Ltd.

Multimedia Appendix 1
Informed consent.
[DOCX File, 42 KB - formative_v9i1e56523_app1.docx]

References

1. Istepanian RSH. Mobile health (m-Health) in retrospect: the known unknowns. Int J Environ Res Public Health 2022 Mar 22;19(7):3747. [doi: 10.3390/ijerph19073747] [Medline: 35409431]
2. Egermark M, Blasiak A, Remus A, Sapanel Y, Ho D. Overcoming pilotitis in digital medicine at the intersection of data, clinical evidence, and adoption. Adv Intell Syst 2022 Sep;4(9):2200056. [doi: 10.1002/aisy.202200056]
3. Byambasuren O, Sanders S, Beller E, Glasziou P. Prescribable mHealth apps identified from an overview of systematic reviews. NPJ Digit Med 2018;1(1):12. [doi: 10.1038/s41746-018-0021-9] [Medline: 31304297]
4. Gajarawala SN, Pelkowski JN. Telehealth benefits and barriers. J Nurse Pract 2021 Feb;17(2):218-221. [doi: 10.1016/j.nurpra.2020.09.013] [Medline: 33106751]
5. McKee M, van Schalkwyk MCI, Stuckler D. The second information revolution: digitalization brings opportunities and concerns for public health. Eur J Public Health 2019 Oct 1;29(Supplement_3):3-6. [doi: 10.1093/eurpub/ckz160] [Medline: 31738440]
6. Schwalbe N, Wahl B. Artificial intelligence and the future of global health. Lancet 2020 May 16;395(10236):1579-1586. [doi: 10.1016/S0140-6736(20)30226-9] [Medline: 32416782]
7. Freeman K, Dinnes J, Chuchu N, et al. Algorithm based smartphone apps to assess risk of skin cancer in adults: systematic review of diagnostic accuracy studies. BMJ 2020 Feb 10;368:m127. [doi: 10.1136/bmj.m127] [Medline: 32041693]
8. Crane D, Garnett C, Brown J, West R, Michie S. Behavior change techniques in popular alcohol reduction apps: content analysis. J Med Internet Res 2015 May 14;17(5):e4060. [doi: 10.2196/jmir.4060] [Medline: 25977135]
9. Fleming GA, Petrie JR, Bergenstal RM, Holl RW, Peters AL, Heinemann L. Diabetes digital app technology: benefits, challenges, and recommendations. A consensus report by the European Association for the Study of Diabetes (EASD) and the American Diabetes Association (ADA) Diabetes Technology Working Group. Diabetes Care 2020 Jan;43(1):250-260. [doi: 10.2337/dci19-0062] [Medline: 31806649]
10. Wicks P, Chiauzzi E. 'Trust but verify'--five approaches to ensure safe medical apps. BMC Med 2015 Sep 25;13(1):205. [doi: 10.1186/s12916-015-0451-z] [Medline: 26404791]
11. Evidence standards framework for digital health technologies. National Institute for Health and Care Excellence. 2018. URL: https://www.nice.org.uk/corporate/ecd7 [accessed 2025-1-3]
12. Dang A, Arora D, Rane P. Role of digital therapeutics and the changing future of healthcare. J Fam Med Prim Care 2020 May;9(5):2207-2213. [doi: 10.4103/jfmpc.jfmpc_105_20] [Medline: 32754475]
13. Evaluating digital health products. GOVUK. 2020. URL: https://www.gov.uk/government/collections/evaluating-digital-health-products [accessed 2025-01-03]
14. The Fast-Track Process for Digital Health Applications (DiGA) according to Section 139e SGB V. A Guide for Manufacturers, Service Providers and Users. Federal Institute for Drugs and Medical Devices. 2020. URL: https://www.bfarm.de/SharedDocs/Downloads/EN/MedicalDevices/DiGA_Guide.pdf [accessed 2024-02-12]
15. Accelerated Access Collaborative patient and public involvement strategy. NHS England. 2022. URL: https://www.england.nhs.uk/aac/publication/accelerated-access-collaborative-patient-and-public-involvement-strategy/ [accessed 2024-12-04]
16. Patient and public involvement policy. National Institute for Health and Care Excellence. 2013. URL: https://www.nice.org.uk/about/nice-communities/nice-and-the-public/public-involvement/public-involvement-programme/patient-public-involvement-policy [accessed 2025-01-03]
17. Gradinger F, Britten N, Wyatt K, et al. Values associated with public involvement in health and social care research: A narrative review. Health Expect 2015 Oct;18(5):661-675. [doi: 10.1111/hex.12158] [Medline: 24325553]

XSL•FO
RenderX

18. Report on the participation decision-making process. Center for Information and Study on Clinical Research Participation (CISCRP). 2017. URL: https://www.ciscrp.org/wp-content/uploads/2019/06/2017-CISCRP-Perceptions-and-Insights-Study-Decision-Making-Process.pdf [accessed 2025-01-03]
19. Stead M, Eadie D, Gordon D, Angus K. “Hello, hello--it’s English I speak!”: A qualitative exploration of patients’ understanding of the science of clinical trials. J Med Ethics 2005 Nov;31(11):664-669. [doi: 10.1136/jme.2004.011064] [Medline: 16269566]
20. Ganz PA. Clinical trials. Concerns of the patient and the public. Cancer 1990 May 15;65(10 Suppl):2394-2399. [doi: 10.1002/1097-0142(19900515)65:10+<2394::aid-cncr2820651509>3.0.co;2-#] [Medline: 2185874]
21. Enhancing the use of scientific evidence to judge the potential benefits and harms of medicines. The Academy of Medical Sciences. 2017. URL: https://acmedsci.ac.uk/file-download/44970096 [accessed 2025-1-3]
22. Finucane E, O’Brien A, Treweek S, et al. The People’s Trial: supporting the public’s understanding of randomised trials. Trials 2022 Mar 9;23(1):205. [doi: 10.1186/s13063-021-05984-1] [Medline: 35264220]
23. Lupton D. Beyond techno-utopia: critical approaches to digital health technologies. Societies (Basel) 2014;4(4):706-711. [doi: 10.3390/soc4040706]
24. Strange M. Three different types of AI hype in healthcare. AI Ethics 2024 Aug;4(3):833-840. [doi: 10.1007/s43681-024-00465-y]
25. Young AT, Amara D, Bhattacharya A, Wei ML. Patient and general public attitudes towards clinical artificial intelligence: a mixed methods systematic review. Lancet Digit Health 2021 Sep;3(9):e599-e611. [doi: 10.1016/S2589-7500(21)00132-1] [Medline: 34446266]
26. Sauchelli S, Pickles T, Voinescu A, et al. Public attitudes towards the use of novel technologies in their future healthcare: a UK survey. BMC Med Inform Decis Mak 2023 Feb 22;23(1):38. [doi: 10.1186/s12911-023-02118-2] [Medline: 36814262]
27. Lee L, Salami RK, Martin H, et al. “How I would like AI used for my imaging”: children and young persons’ perspectives. Eur Radiol 2024 Dec;34(12):7751-7764. [doi: 10.1007/s00330-024-10839-9] [Medline: 38900281]
28. Cowey AE, Potts HWW. What can we learn from second generation digital natives? A qualitative study of undergraduates’ views of digital health at one London university. Dig HEALTH 2018;4:2055207618788156. [doi: 10.1177/2055207618788156] [Medline: 30046453]
29. Hernandez-Ramos R, Aguilera A, Garcia F, et al. Conducting internet-based visits for onboarding populations with limited digital literacy to an mHealth intervention: development of a patient-centered approach. JMIR Form Res 2021 Apr 29;5(4):e25299. [doi: 10.2196/25299] [Medline: 33872184]
30. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
31. Braun V, Clarke V. One size fits all? What counts as quality practice in (reflexive) thematic analysis? Qual Res Psychol 2021 Jul 3;18(3):328-352. [doi: 10.1080/14780887.2020.1769238]
32. Braun V, Clarke V. Supporting best practice in reflexive thematic analysis reporting in palliative medicine: a review of published research and introduction to the Reflexive Thematic Analysis Reporting Guidelines (RTARG). Palliat Med 2024 Jun;38(6):608-616. [doi: 10.1177/02692163241234800] [Medline: 38469804]
33. Herian MN, Shank NC, Abdel-Monem TL. Trust in government and support for governmental regulation: the case of electronic health records. Health Expect 2014 Dec;17(6):784-794. [doi: 10.1111/j.1369-7625.2012.00803.x] [Medline: 22809222]
34. Eysenbach G. Credibility of Health Information and Digital Media: New Perspectives and Implications for Youth: MacArthur Foundation Digital Media and Learning Initiative; 2008:123-154. [doi: 10.1162/dmal.9780262562324.123]
35. O’Connor S, Hanlon P, O’Donnell CA, Garcia S, Glanville J, Mair FS. Understanding factors affecting patient and public engagement and recruitment to digital health interventions: a systematic review of qualitative studies. BMC Med Inform Decis Mak 2016 Sep 15;16(1):120. [doi: 10.1186/s12911-016-0359-3] [Medline: 27630020]
36. Pagoto S, Bennett GG. How behavioral science can advance digital health. Transl Behav Med 2013 Sep;3(3):271-276. [doi: 10.1007/s13142-013-0234-z] [Medline: 24073178]
37. Yardley L, Morrison L, Bradbury K, Muller I. The person-based approach to intervention development: application to digital health-related behavior change interventions. J Med Internet Res 2015 Jan 30;17(1):e30. [doi: 10.2196/jmir.4055] [Medline: 25639757]
38. Curtis K, Price K. Factors that influence older people’s engagement with digital health technology. Nurs Older People 2017 Nov 30;29(10):27-30. [doi: 10.7748/nop.2017.e986] [Medline: 29188928]
39. Pagoto S, Waring ME, May CN, et al. Adapting behavioral interventions for social media delivery. J Med Internet Res 2016 Jan 29;18(1):e24. [doi: 10.2196/jmir.5086] [Medline: 26825969]

Abbreviations

AI: artificial intelligence
DiGA: Digitale Gesundheitsanwendungen
NHS: National Health Service

XSL•FO
RenderX

NICE ESF: National Institute for Health and Care Excellence Evidence Standards Framework

Edited by A Mavragani; submitted 21.03.24; peer-reviewed by B Cliffe; revised version received 28.10.24; accepted 30.10.24; published 20.01.25.

Please cite as:
Bondaronek P, Li J, Potts HWW
Public Understanding and Expectations of Digital Health Evidence Generation: Focus Group Study
JMIR Form Res 2025;9:e56523
URL: https://formative.jmir.org/2025/1/e56523
doi:10.2196/56523

© Paulina Bondaronek, Jingfeng Li, Henry W W Potts. Originally published in JMIR Formative Research (https://formative.jmir.org), 20.1.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

International Expert Consensus on Relevant Health and Functioning Concepts to Assess in Users of Tobacco and Nicotine Products: Delphi Study

Vivienne Law[1*], BSc; Esther F Afolalu[2*], PhD; Linda Abetz-Webb[3], MA; Lee Andrew Wemyss[4], BSc; Andrew Turner[4], BSc; Christelle Chrea[2], PhD

[1]Zebra26 Ltd., Devon, United Kingdom
[2]PMI R&D, Philip Morris Product S.A., Neuchâtel, Switzerland
[3]Patient-Centered Outcomes Assessments Ltd., Cheshire, United Kingdom
[4]DaSH Global, Cheshire, United Kingdom
[*]these authors contributed equally

Corresponding Author:
Esther F Afolalu, PhD
PMI R&D, Philip Morris Product S.A.
Quai Jeanrenaud 5
Neuchâtel, 2000
Switzerland
Phone: 41 58 242 2218
Email: esther.afolalu@pmi.com

Abstract

Background: A Delphi study was conducted to reach a consensus among international clinical and health care experts on the most important health and functioning self-reported concepts when evaluating a switch from smoking cigarettes to using smoke-free tobacco and/or nicotine products (sf-TNPs).

Objective: The aim of this research was to identify concepts considered important to measure when assessing the health and functioning status of users of tobacco and/or nicotine products.

Methods: Experts (n=105), including health care professionals, researchers, and policy makers, from 26 countries with professional experience and knowledge of sf-TNPs completed a 3-round, adapted Delphi panel. Online surveys combining quantitative (MaxDiff best-worst scaling and latent class analysis) and qualitative assessments were used to rank and achieve alignment on the importance of 69 health and functioning concepts. All experts participating in round I completed round II, and 101 (95%) completed round III.

Results: The round I analysis identified 36 (52%) out of 69 concepts that were refined for the round II assessment. The highest-ranked concepts reflected health-related impacts, while the lowest-ranked ranked concepts were related to aesthetics and social impacts. Round II ranking reinforced the importance of concepts relating to health impacts, and the analysis resulted in 20 concepts retained for round III assessment. In round III, the 4 highest-ranked concepts were cardiovascular symptoms, shortness of breath, chest pain, and worry about smoking-related diseases and impact on general health, and they made up 50% of the total score in the MaxDiff analysis. Experts reported likelihood of seeing measurable levels of change in the final 20 concepts with a switch to an sf-TNP. The majority of experts felt it was “likely” or “extremely likely” to observe changes in concepts such as gum problems (74/101, 73%), phlegm or mucus while coughing or not coughing (72/101, 71%), general perception of well-being (72/101, 71%), and throat irritation or sore throat (72/101, 71%). Latent class analysis revealed subgroups of experts with different perceptions of the relative importance of the concepts, which varied depending on professional specialty and geographic region. For example, 74% (14/19) of oncologists aligned with the subgroup prioritizing physical health symptoms, while 71% (12/17) of experts from Asia aligned with the subgroup considering both physical health and psychosocial aspects.

Conclusions: This study identified key concepts to be considered in the development of a new measurement instrument to assess the self-reported health and functioning status of individuals using sf-TNPs. The findings contribute to the scientific evidence base for understanding and evaluating both the individual and public health impacts of sf-TNPs.

XSL•FO
RenderX

(JMIR Form Res 2025;9:e58614) doi:10.2196/58614

KEYWORDS

Delphi study; expert consensus; outcome measures; health and functioning; tobacco and/or nicotine products

Introduction

Reducing exposure to harmful and potentially harmful constituents in cigarette smoke by cessation of smoking or switching to reduced risk tobacco and/or nicotine products (TNPs) is a major public health focus worldwide [1]. Although nicotine-replacement therapy (NRT) and targeted counseling services can help individuals achieve their quitting goals, there can be hurdles to becoming and/or staying smoking free [2].

Smoke-free tobacco and/or nicotine products (sf-TNPs) refer to TNPs that do not undergo combustion and therefore do not produce the harmful smoke generated by cigarettes, which has been established as the main risk factor of tobacco-related diseases [3]. These sf-TNPs can include heated tobacco products, e-cigarettes (vapes), other e-vapor products (e-pipes, e-cigars, and so on), nicotine pouches, and smokeless tobacco products (snuff, snus, and chewing tobacco) [3]. Regulatory bodies, like the US Food and Drug Administration (FDA), require robust data to evaluate the health impacts of sf-TNPs and guide regulatory decisions on product authorization. For instance, the FDA's modified risk tobacco product pathway mandates comprehensive scientific evidence to support claims that a tobacco product reduces harm or the risk of tobacco-related disease in individual tobacco users and benefits the health of the population as a whole [4]. Some sf-TNPs have been recognized as modified-risk tobacco products under this pathway [4,5] and may help reduce the negative health effects associated with commercially marketed combustible tobacco products such as cigarettes [6-8].

The adverse effects of smoking on health have been well documented, whereas stopping smoking can improve health and reduce the risk of disease [6,9,10]. Evidence also suggests that switching away from cigarettes to using sf-TNPs can help reduce cigarette consumption and may lead to cessation in some cases [8]. While studies have shown that sf-TNPs reduce exposure to many of the harmful and potentially harmful constituents found in cigarette smoke, they may still pose health risks. For instance, e-cigarettes, heated tobacco products, and smokeless tobacco have been associated with some cardiovascular, respiratory, and oral health risks, albeit to a lesser extent than combustible cigarettes [11-14]. The long-term health effects of sf-TNPs are not yet fully understood, and further research is necessary to comprehensively assess their safety profile. In addition, little is known about the self-reported health impact of switching from cigarettes to sf-TNPs [6]. In this context, measuring the self-reported experience of health and functioning (including health status, functional status, and other health-related quality of life constructs) is crucial to understanding the impact of tobacco harm reduction strategies [15].

Generic health status measures, such as the Short Form Survey–12 [16] and Short Form Survey–36 instruments [17], have shown that those who smoke tend to report lower health status compared with those who never smoked, although the impact of cessation on health status seems to be more complex. Some smoking-specific measures have been developed but not yet widely standardized; therefore, they may not be directly applicable to sf-TNPs [9,18,19]. Crucially, current measures often lack the necessary sensitivity to detect longitudinal health changes in healthy populations due to high ceiling effects at baseline [20], hindering the assessment of changes in health and functioning following a switch from smoking cigarettes to using sf-TNPs.

To help address these challenges, the development of a new self-reported measure (ABOUT – Health and Functioning) was undertaken to assess the health and functioning status of individuals using sf-TNPs. This new measure is part of the portfolio of the ABOUT Toolbox [21], consisting of self-report measures to assess perceptions and behavioral outcomes related to the use of sf-TNPs. The initial preparatory phase of the instrument development was based on several research activities (systematic literature review, reanalysis of qualitative data, and expert insights [22]) and resulted in the identification of 69 health and functioning concepts relevant to TNP use. The qualitative research phase that followed consisted of (1) concept elicitation interviews of users to understand their perceptions of health and functioning after switching to sf-TNPs [23,24] and (2) the Delphi panel study reported in this paper. The aim of the Delphi study was to identify the health and functioning concepts considered most important by clinical and health care experts (ie, health care professionals, researchers, policy makers, and those involved in smoking cessation or tobacco harm reduction) when assessing the health and functioning status of individuals who stop smoking cigarettes or switch to using sf-TNPs or NRTs. Delphi panel methodology is a well-established process for determining consensus among relevant groups of individuals to aid issue prioritization, ranking, development, and obtaining agreement on guidelines, concept-framework development, and development of outcome measures [25-27], including issues relevant to smoking-related behaviors and sf-TNP use [25,28,29]. Delphi panels have previously been used to forecast trends and changes over time in health-related matters [30-32].

Methods

Objectives and Summary of Approach

This Delphi panel was organized for experts to select and rank health and functioning concepts considered important to assess when individuals stop smoking combustible TNPs (eg, cigarettes) or switch to using sf-TNPs (eg, e-cigarettes [vapes], heated tobacco, or smokeless tobacco products) or NRTs. The study was conducted and reported based on available guidance for conducting and reporting Delphi studies in health care research [33].

XSL•FO
RenderX

A preliminary set of 69 concepts was identified from earlier research activities including a scoping literature review, results from the reanalysis of previous TNP consumer qualitative research results, and expert opinion from a small group of key opinion leaders [22]. This list was presented for evaluation in this Delphi panel according to the process described in Figure 1. An adapted approach was used for the Delphi panel, whereby the results of round I were reviewed and refined in light of the ongoing qualitative concept elicitation interview findings from consumers of TNPs and ongoing review by key opinion leaders (Supplementary Information and Table S1 in Multimedia Appendix 1). The concept list to be evaluated in round II was adjusted accordingly and refined considering the round I findings. Round III was a detailed review of the final set of health and functioning concepts. This step also included an evaluation of likelihood of change that experts considered important for each concept for inclusion in a self-reported measure of clinical relevance.

Figure 1. Overview of the design and structure of the adapted Delphi panel study.

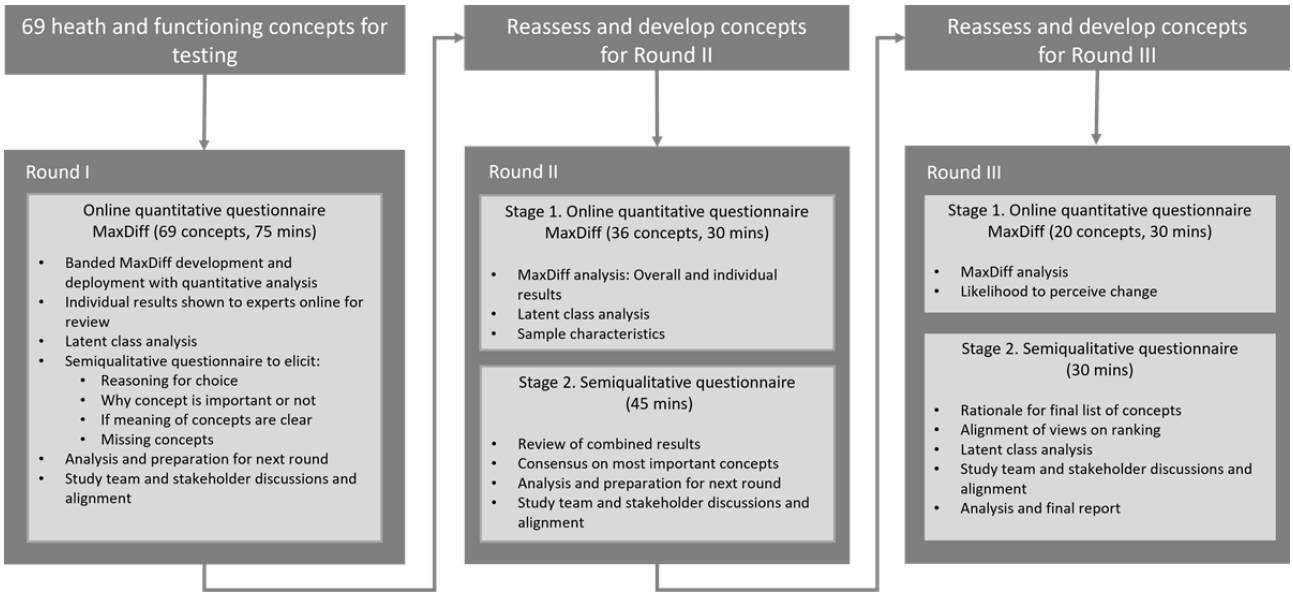

Participants and Recruitment

For this adapted Delphi panel, we recruited experts who routinely treat, communicate with, or advise individuals who smoke or wish to stop smoking or switch to sf-TNPs. The recruitment aim was to ensure that a minimum of 100 experts globally completed all 3 rounds of the Delphi panel. To anticipate attrition between rounds, the initial recruitment target was set at 120 experts.

Recruitment was undertaken by QualWorld, who identified experts through their survey panels, existing contacts, publications, word of mouth, and canvassing through appropriate organizations in individual countries. The process and route of contact to identify potential respondents are outlined in Figure 2.

To anticipate and account for potential sensitivities among health care practitioners about undertaking a project sponsored by a tobacco company, the recruitment company sent out an initial survey asking which industries experts would be interested in working on health-related topics. Tobacco companies were listed among other industries such as mining, pharmaceuticals, and oil and gas. Only experts stating unprompted that they would consider participating in studies by tobacco companies were chosen to receive further correspondence linked to the Delphi panel. Subsequent recruitment steps proceeded on an individual basis according to predefined inclusion and exclusion criteria (Textbox 1).

To ensure consistency in responses across different regions, the Delphi panel was conducted in English where possible and in local languages for experts in Japan, South Korea, Russia, Ukraine, Czech Republic, and Bulgaria. A professional translation service translated the survey materials and responses, and back-translation techniques were used to verify accuracy. The local bilingual recruitment team members also performed proofreading and quality checks from a language and contextual perspective to ensure the translated content accurately reflected the original meaning. For each step of the Delphi panel, the English and local language surveys were programmed and uploaded for a user acceptability test by the project team before launch.

XSL•FO
RenderX

Figure 2. Identification and recruitment of experts through specific channels for the adapted, 3-round Delphi panel.

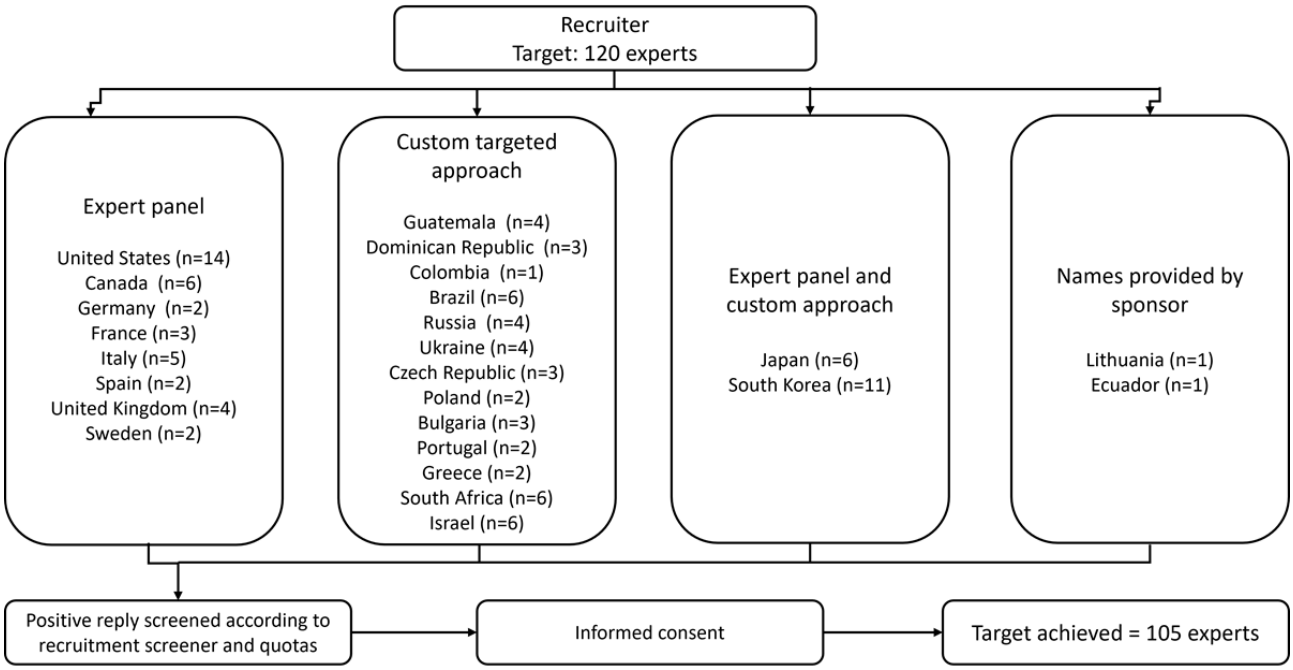

Textbox 1. Inclusion and exclusion criteria for the participants recruited for the Delphi panel.

Inclusion criteria

- Had current clinical, academic, or professional experience in the area of smoking-related diseases, smoking cessation, nicotine addiction, or other smoking-related conditions, as well as health policy or advocacy work related to tobacco control or tobacco harm reduction
- Currently working in research, prescribing, or a recommendation capacity
- Knowledge of smoke-free tobacco and/or nicotine products or nicotine replacement therapies, works with patients to some degree, and/or recommends smoke-free tobacco and/or nicotine products or nicotine replacement therapies
- Expertise in one of the following groups:
  - Specialist physicians: smoking-related oncology; smoking-related respiratory disease, and smoking-related cardiovascular disease (specifically treatment of patients)
  - General practitioners and internal medicine physicians: those who see patients with a smoking-related condition or disease and recommend smoking cessation advice
  - Dentists or oral hygienists: those who see patients with smoking-related dental issues
  - Smoking cessation, addiction, or dependence: those who help users quit smoking with less harmful products or nicotine replacement therapy or worked with nicotine addiction; nurses undertaking smoking cessation activity; researchers in nicotine addictions; psychologists, social workers, or counselors
  - Health policy, advocacy, and non-profit organizations: smoking-related health policy, including tobacco harm reduction; organizations or charities that encourage people to quit; health policy and advocacy related to tobacco control and tobacco harm reduction. This can include physicians who work alongside the government and local policy as well as nonprofit organizations
- Ability and willingness to participate across the whole study
- Ability to complete the research in the English language (except for these countries: South Korea, Japan, Ukraine, Czech Republic, Russia, and Bulgaria)
- At least 30% of experts to identify as female (although this was not a strict quota for recruitment)

Exclusion criteria

- More than 50% of experts should not have been a consultant with or worked for the tobacco industry
- Had less than 5 years of clinical, academic, or professional experience
- Did not fall into one of the categories for inclusion
- Not willing to complete all 3 rounds of research

XSL•FO
RenderX

Ethical Considerations

The New England Independent Review Board reviewed the study protocol and all study materials and any amendments and granted approval of the study (reference 1 - 9184 - 1).

Before participation, all participants were required to review and sign an informed consent form, which provided a detailed description of the study, outlined the procedures, and specified participant expectations. Participants had the opportunity to ask questions and receive answers before signing the consent form.

All study data were deidentified and participants were identified only by unique ID numbers throughout the research rounds.

All experts were compensated and received a cash honorarium according to local professional rates in their respective countries, ranging from £115 to £450 (approximately US $147 to US $576 at the time of the research), for their time after each round.

Procedure and Analytical Methods

MaxDiff Analysis

This study incorporated the use of MaxDiff—a type of best-worst (maximum difference) scaling analytical method [34-36]. This methodology was primarily chosen due to the large number of initial concepts to be evaluated. Ranking and rating exercises with numerous concepts can lead to a cognitive burden, so a best-worst scaling approach overcomes fatigue while generating additional information by making respondents choose the most and least important concepts. The result is a list of concepts in order of relative importance [36-38].

For each round, the MaxDiff Sawtooth software algorithm [39] generated a list of scaled scores from most to least important, as well as the relative importance of one concept compared with another. This was based on the number of times that concept was rated as the most or least important. In addition, cumulative scores were used to differentiate the number of concepts accounting for 20%, 50%, and 70% of the most important concepts; the 50th percentile was used as an initial guide for which concepts to retain.

Rounds I and II also featured an “anchor” value, which was a score marking the boundary between concepts considered important (to be retained in the next round) or unimportant (to be excluded from the next round). This anchor value was derived from an anchor question representing a threshold of “important or not important” included for each concept to be assessed as above (ie, important) or below (ie, unimportant) the anchor. The anchor question in this survey required one of the following 3 responses: “none of these [items] are important,” “some of these [items] are important,” or “all of these [items] are important.” The anchor value was statistically derived using the Sawtooth software [38]. An illustration of a typical MaxDiff question is shown in Figure 3.

Figure 3. Typical presentation of the Delphi panel MaxDiff ranking question, including the anchor evaluation. NRT: nicotine replacement therapy; TNP: tobacco and/or nicotine products.

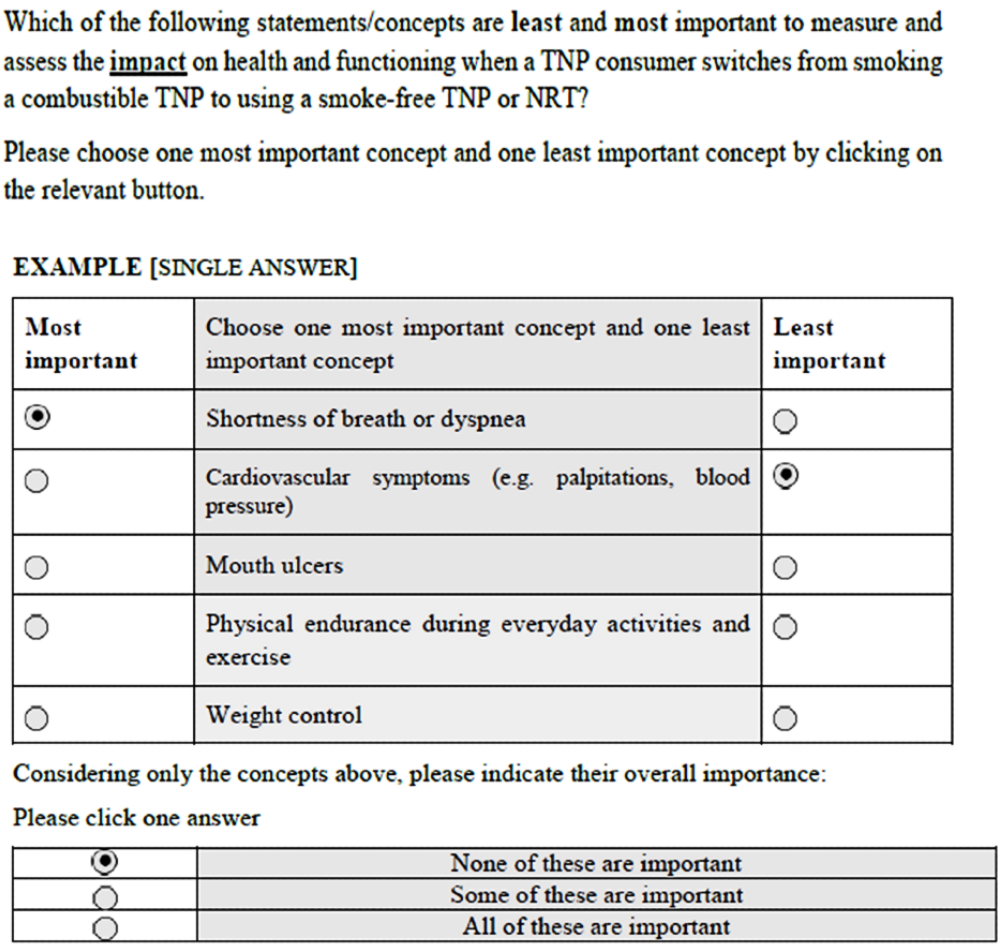

Which of the following statements/concepts are least and most important to measure and assess the impact on health and functioning when a TNP consumer switches from smoking a combustible TNP to using a smoke-free TNP or NRT?

Please choose one most important concept and one least important concept by clicking on the relevant button.

EXAMPLE [SINGLE ANSWER]

Most important	Choose one most important concept and one least important concept	Least important
◉	Shortness of breath or dyspnea	○
○	Cardiovascular symptoms (e.g. palpitations, blood pressure)	◉
○	Mouth ulcers	○
○	Physical endurance during everyday activities and exercise	○
○	Weight control	○

Considering only the concepts above, please indicate their overall importance:

Please click one answer

◉	None of these are important
○	Some of these are important
○	All of these are important

XSL•FO
RenderX

Latent Class Analysis

Latent class analysis [40] was also performed to determine whether some concepts were more important to some groups of experts over others, according to experts' demographic or other characteristics. This ensured that the final set of items was not biased toward one subset of health care professionals at the expense of others' preferences. Where there were concepts that were only important to subsets of the study sample but unimportant overall, these concepts could be retained.

Semiqualitative Assessment

Semiqualitative questions were incorporated into each Delphi panel round to provide an increased understanding of participants' responses to the results of the ranking and scoring exercise. These questions were asked to ensure no concepts were missing or needed rewording, and to understand the reasons behind experts' selections of importance or unimportance. They also sought to gauge agreement or disagreement with the overall ranking results and to gain insights into the reasons for these viewpoints. After each round, the semiqualitative comments were reviewed alongside the MaxDiff scores. This process involved coding and analyzing the qualitative feedback to identify common themes and insights that could provide context to the quantitative rankings. This feedback was then integrated with the quantitative MaxDiff results to refine and inform decisions about the list of concepts to be included for the next rounds. It helped explain the rationale behind the experts' choices and highlighted any discrepancies or areas of consensus, ensuring that both qualitative and quantitative perspectives were considered in the decision-making process.

Results

Descriptive Participant Demographics

Approximately 3100 experts were contacted, with a final sample of 105 experts recruited across 26 countries. Table 1 shows the demographics and participant characteristics, and Table 2 shows the participant quotas per region and per specialty. Of the initial 105 experts, all participated in exploratory round I and round II, and 101 (96%) completed round III (Table 2). Nearly all participants (90/105, 87%) reported frequently addressing smoking cessation with clients or patients. A total of 64% (67/105) of experts in the study identified as male, and most were located in the Americas, Europe, or Asia. Regarding place of work, three-quarters (79/105, 75%) of experts were based in hospitals or other clinical settings, mostly as medical specialists or general care practitioners.

XSL•FO
RenderX

Table 1. Demographics of Delphi panel participants (rounds I-III).

Demographic variable	Respondents	
	Rounds I and II (n=105), n (%)	Round III (n=101), n (%)
Sex		
Male	67 (64)	64 (63)
Female	38 (36)	37 (37)
Geographical region		
North America (United States and Canada)	20 (19)	17 (17)
Asia (Japan and South Korea)	17 (16)	17 (17)
Central, Caribbean, South America (Brazil, Colombia, Dominican Republic, and Guatemala)	15 (14)	15 (15)
Southern Europe (Spain, Portugal, Italy, Greece, and Bulgaria)	14 (13)	13 (13)
Western Europe (France, Germany, Sweden, United Kingdom, and the Netherlands)	13 (12)	13 (13)
Eastern Europe (Russia, Ukraine, and Lithuania)	9 (9)	9 (9)
Africa (South Africa)	6 (6)	5 (5)
Middle East (Israel)	6 (6)	6 (6)
Central Europe (Poland and Czech Republic)	5 (5)	5 (5)
Work setting		
Hospital	43 (41)	41 (41)
Clinical (other)	36 (34)	36 (36)
Professional	17 (16)	15 (15)
University hospital	5 (5)	5 (5)
University hospital; hospital setting	4 (4)	4 (4)
Main current role		
General practitioner or internal medicine	27 (26)	27 (27)
Oncology	19 (18)	19 (19)
Dentist or oral hygienist	16 (15)	15 (15)
Research	11 (10)	9 (9)
Cardiology	12 (11)	12 (12)
Respiratory	8 (8)	8 (8)
Counselor or psychologist	7 (7)	6 (6)
Advocacy or health policy or NGO[a]	5 (5)	5 (5)
Frequency of discussing smoking cessation with clients		
Frequent (at least weekly)	90 (86)	86 (85)
Not frequent (less than once a week)	11 (10)	11 (11)
Not applicable	4 (4)	4 (4)
Discussed the benefits of stopping smoking or provided information about sf-TNPs[b]or NRTs[c]to those who smoke and do not want or have difficulty quitting smoking		
Yes	103 (98)	99 (98)
No	0	0
Not available or prefer not to say	2 (2)	2 (2)[d,e]
Provided information on heat-not-burn or heated tobacco products (eg,*IQOS*, Glo) or tobacco vapor products (eg, Ploom Tech)		
Yes	50 (48)	46 (46)

XSL•FO
RenderX

Demographic variable	Respondents	
	Rounds I and II (n=105), n (%)	Round III (n=101), n (%)
No	53 (50)	53 (53)
Not available or prefer not to say	2 (2)	2 (2)
Provided information on smokeless tobacco (eg, Copenhagen Snuff, Swedish Match General Snus, Camel Snus, or any other local brands)		
Yes	39 (37)	35 (35)
No	64 (61)	64 (63)
Not available or prefer not to say	2 (2)	2 (2)
Provided information on e-cigarettes (eg, JUUL, Blu, Logic, or any other local brands)		
Yes	65 (62)	62 (61)
No	38 (36)	37 (37)
Not available or prefer not to say	2 (2)	2 (2)
Provided information on Nicotine replacement therapies (eg, nicotine gums, inhalers, nasal sprays, lozenges, patches)		
Yes	91 (87)	88 (87)
No	11 (10)	10 (10)
Not available or prefer not to say	3 (3)	3 (3)
Current or previous personal use of combustible TNPs[f]		
Never	65 (62)	62 (61)
Yes	35 (33)	34 (34)
Not available or prefer not to say	5 (5)	5 (5)
Current or previous personal use of sf-TNP or NRT of those that were currently or had used combustible TNPs		
Yes	12 (11)	12 (12)

[a]NGO: nongovernmental organization.
[b]sf-TNP: smoke-free tobacco or nicotine product.
[c]NRT: nicotine replacement therapy.
[d]Not available.
[e]Data missing for one respondent in round III, stage 2.
[f]TNP: tobacco or nicotine product.

Table 2. Participant quotas per region and per specialty for the start of round I of the Delphi panel (N=105).

Specialty	Asia	Africa	Middle East	North America	Central America	Southern Europe	Eastern Europe	Western Europe	Central Europe
Specialist physician	6	2	3	6	7	3	4	2	6
GP[a] or IM[b]	4	2	3	4	3	3	2	2	4
Dental specialist	3	1	0	2	2	2	3	1	2
SC[c] or DEP[d]	2	1	0	5	2	1	3	0	3
Health policy or NGO[e]	2	0	0	3	1	0	1	0	0
Total	17	6	6	20	14	9	13	5	15

[a]GP: general practitioner.
[b]IM: internal medicine specialist.
[c]SC: smoking cessation specialist.
[d]DEP: dependence or addiction specialist.
[e]NGO: nongovernmental organization.

XSL•FO
RenderX

MaxDiff Analysis and Semiqualitative Assessment for Reduction or Refinement of Health and Functioning Concepts

Round I

A cumulative score was calculated by adding all the scaled scores (in this case, the total score was 14,900.17) and dividing by the scaled score (concept 1: 425/14,900=3%) and adding to the scaled score of the next highest concept (concept 2: 422/14,900=3%+3% concept 1 and so on; Figure 4). Regarding the relative importance of concepts in round I, when switching from combustible to sf-TNPs, experts placed importance on withdrawal symptoms, general health impacts, and emotion-related impacts. Conversely, concepts that ranked low in priority were those that referred to aesthetic concerns that would have less of a health impact or were not considered exclusive to those who smoke. Experts considered withdrawal symptoms such as anxiety and irritability as potential barriers to switching to sf-TNPs and highlighted there were levels of satisfaction (eg, enjoyment and craving relief) that those who smoke cigarettes would not want to lose when switching to sf-TNPs

Figure 4. Scaling and ranking results for items considered in round I of the Delphi panel. TNP: tobacco and/or nicotine products.

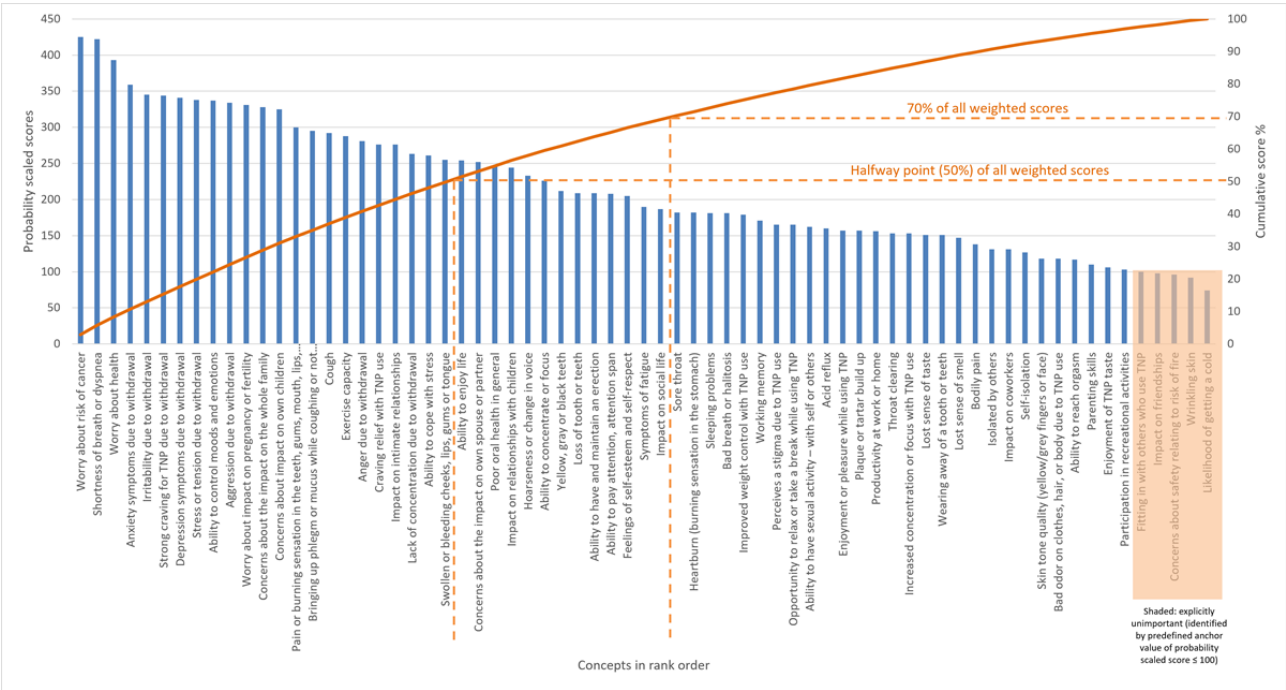

When determining which concepts to include in round II of the Delphi panel evaluation, the relative score (ie, importance) for each concept was considered, resulting in 36 out of 69 concepts (52% of all concepts in ranked order) accounting for 69% (10,285/14,900) of the total importance score. Based on the results and alignment with other activities ongoing within development of the health and functioning measurement instrument, the health and functioning concepts were further reviewed for relevance and clarity, and a different, refined, reduced, and synthesized list of 36 concepts was taken forward for evaluation in round II (refer to Supplementary Information and Table S1 in Multimedia Appendix 1 for further details).

Round II

Of the 36 concepts in this round, 14 made up 50% of the total scaled score (Figure 5). In general, physical health symptoms, and worries about the impact on health concepts remained the highest-ranked concepts and were considered by experts as more important than aesthetic considerations. Experts commented that concepts rated as unimportant and ranked lower in the logit analysis tended to lack specificity to smoking (ie, could be due to other causes), did not directly affect health, or were not reported frequently by TNP users. Of the 36 concepts evaluated in round II, a final selection of 20 highest-ranked concepts were identified for validation and discussion in round III.

XSL•FO
RenderX

Figure 5. Scaling and ranking results for items considered in round II of the Delphi panel. TNP: tobacco and/or nicotine products.

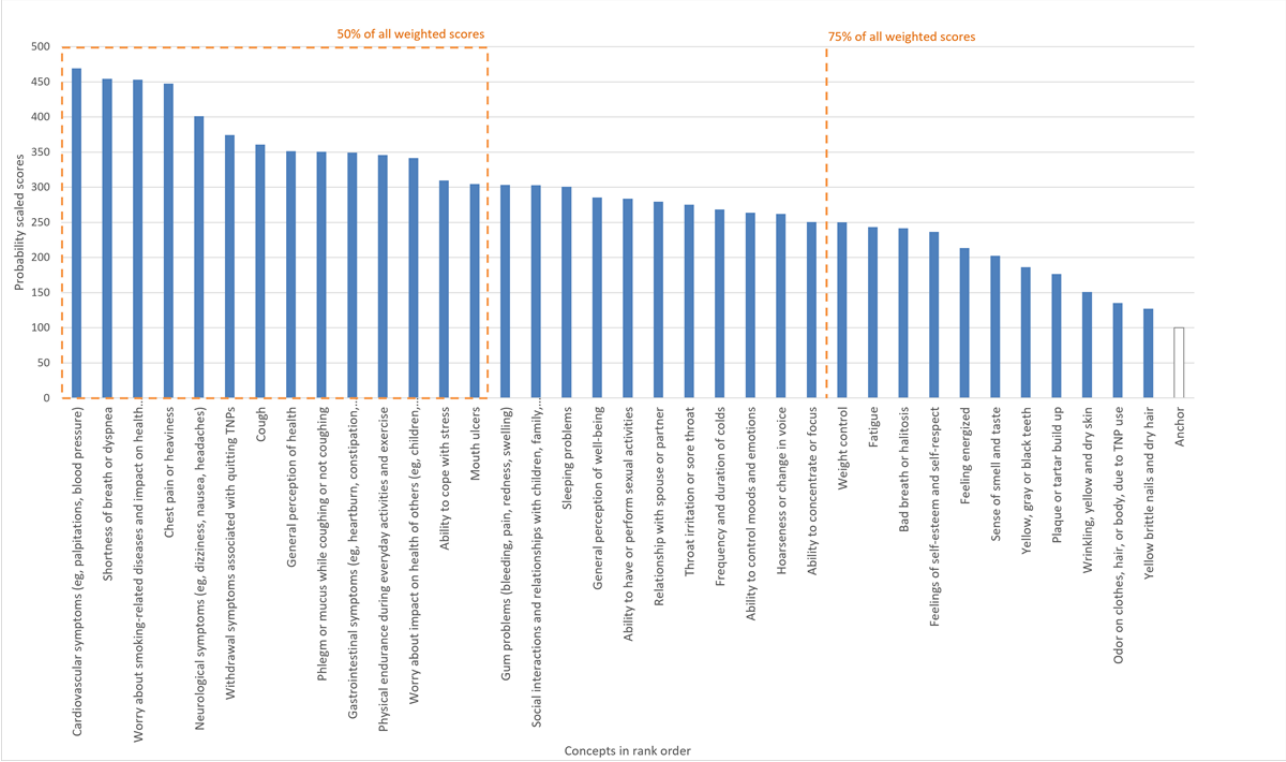

Round III

Of the final 20 concepts ranked in round III (Figure 6), the sum of the top 4 concept scores accounted for more 50% of the total score: cardiovascular symptoms, shortness of breath or dyspnea, chest pain heaviness, and worry about smoking-related diseases and impact on health in general. Physical health concepts remained the most important experts across all 3 rounds

Figure 6. Ranking and ordering of the final 20 items identified in round III of the Delphi panel.

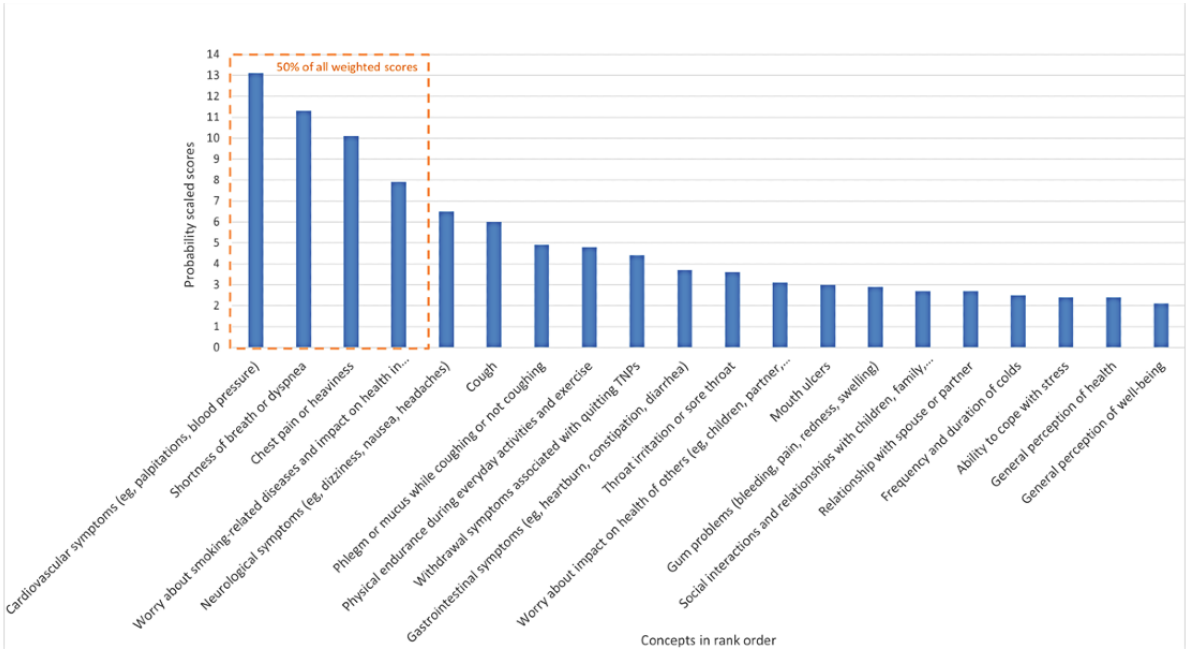

Latent Class Analysis

Latent class analysis of the results in round III revealed 2 identifiable subgroups, each with its own ranking and ordering of the final 20 concepts (Textbox 2). In both cases, concepts related to physical health, particularly cardiovascular and respiratory symptoms, ranked highest. In subgroup 1, the lowest-ranked concepts were related to psychosocial or general well-being and social interactions, and all scored equally low. For subgroup 2, there was minimal differentiation between concepts beyond the highest-ranking group.

XSL•FO
RenderX

Textbox 2. List of concepts in ranked order (highest to lowest) for the 2 subgroups identified by latent class analysis in round III of the Delphi panel.

Subgroup 1 in rank order

- Cardiovascular symptoms (eg, palpitations and blood pressure)
- Shortness of breath or dyspnea
- Chest pain or heaviness
- Neurological symptoms (eg, dizziness, nausea, and headaches)
- Cough
- Mouth ulcers
- Phlegm or mucus while coughing or not coughing
- Throat irritation or sore throat
- Gum problems (eg, bleeding, pain, redness, or swelling)
- Gastrointestinal symptoms (eg, heartburn, constipation, and diarrhea)
- Worry about smoking-related diseases and impact on health in general (eg, cancer and stroke)
- Withdrawal symptoms associated with quitting tobacco and/or nicotine products
- Frequency and duration of colds
- Physical endurance during everyday activities and exercise
- Ability to cope with stress
- Worry about impact on health of others (eg, children, partner, family, and friends)
- Relationship with spouse or partner
- General perception of health
- General perception of well-being
- Social interactions and relationships with children, family, friends, and colleagues

Subgroup 2 in rank order

- Cardiovascular symptoms (eg, palpitations and blood pressure)
- Shortness of breath or dyspnea
- Worry about smoking-related diseases and impact on health in general (eg, cancer and stroke)
- Chest pain or heaviness
- Physical endurance during everyday activities and exercise
- Neurological symptoms (eg, dizziness, nausea, and headaches)
- Cough
- Worry about impact on health of others (eg, children, partner, family, and friends)
- Social interactions and relationships with children, family, friends, and colleagues
- Withdrawal symptoms associated with quitting tobacco and/or nicotine products
- Relationship with spouse or partner
- Phlegm or mucus while coughing or not coughing
- General perception of health
- General perception of well-being
- Ability to cope with stress
- Gastrointestinal symptoms (eg, heartburn, constipation, and diarrhea)
- Throat irritation or sore throat
- Frequency and duration of colds
- Gum problems (eg, bleeding, pain, redness, and swelling)
- Mouth ulcers

XSL•FO
RenderX

In stage 2 of round III, following the MaxDiff analysis, the ranked concept lists for each of the 2 latent class subgroups were shown to the experts, who were asked to identify which group best reflected their own thinking and provide a rationale for their choice. Just over half of the experts (58/100, 58%) reported that subgroup 1 best reflected their thinking compared with (42/100, 42%) for subgroup 2. When asked about their rationale for aligning with subgroup 1, the main reason was a focus on the most important physical health symptoms (33/58, 57%). The main rationale for choosing subgroup 2 was a greater emphasis on the importance of the consumer and their perspective (10/42, 24%), including psychological elements, the well-being of consumers, and impact of smoking on the health of others (14/42, 33%).

Regarding the influence of professional specialty in self-alignment with subgroups (Table 3), preference for subgroup 1 tended to be among oncologists (14/19, 74%), dentists (11/15, 73%), counselors (4/6, 67%), and general practitioners or internal medicine specialists (17/27, 63%). Conversely, all policy advisors and charity workers (5/5) felt that subgroup 2 best reflected their views. When geographical location was considered (Table 3), there was a strong preference for subgroup 2 in participants from Asia (12/17, 71%) and a slight preference among those in North America (9/17, 53%). Experts from all other regions tended mostly to identify with subgroup 1 (Middle East: 5/6, 83%; Africa: 4/5, 80%; Central Europe: 4/5, 80%; Eastern Europe: 7/9, 78%; Western Europe: 9/13, 69%; Southern Europe: 8/13, 62%; and Central or South America: 8/15, 53%). Regarding the final selection of 20 concepts, both subgroups considered all concepts to be important and relevant for inclusion in any self-reported outcome measure.

Table 3. Region and specialty after self-identifying with a latent class subgroup in round III of the Delphi panel.

Experts' region and specialty	Total (n=100), n (%)[a]	Latent class subgroup 1, n (%)[b]	Latent class subgroup 2, n (%)[b]
Region	100 (100)	58 (58)	42 (42)
North America	17 (17)	8 (47)	9 (53)
Western Europe	13 (13)	9 (69)	4 (31)
Southern Europe	13 (13)	8 (62)	5 (38)
Africa	5 (5)	4 (80)	1 (20)
Central/South America	15 (15)	8 (53)	7 (47)
Central Europe	5 (5)	4 (80)	1 (20)
Asia	17 (17)	5 (29)	12 (71)
Eastern Europe	9 (9)	7 (78)	2 (22)
Middle East	6 (6)	5 (83)	1 (17)
Specialty	100 (100)	58 (58)	42 (42)
General practice or internal medicine	27 (27)	17 (63)	10 (37)
Dentist or oral hygienist	15 (15)	11 (73)	4 (27)
Oncologist	19 (19)	14 (74)	5 (26)
Cardiovascular specialist	11 (11)	5 (45)	6 (55)
Respiratory	8 (8)	3 (38)	5 (63)
Researcher	9 (9)	4 (44)	5 (56)
Counselor or psychologist	6 (6)	4 (67)	2 (33)
Policy advisor	4 (4)	0 (0)	4 (100)
Charity or advocacy	1 (1)	0 (0)	1 (100)

[a]Data for 100 experts were included, with one respondent (from South Africa) reporting not understanding the questions asked.

[b]Percentage use the n value in the "Total" column as the denominator.

Additional Semiqualitative Assessment—Likelihood to Perceive Change

In round III (in response to the question: "Now that you have seen the list of concepts collated from Rounds I & II, how likely do you think it is to see a change in this concept when a TNP user switches from a combustible TNP to an sf-TNP?"), most experts reported they would expect to find measurable levels of change in 13 out of 20 individual concepts. Furthermore, over half of the respondents stated it was "likely" or "extremely likely" that a measurable change would occur in these concepts when a TNP user switched from a combustible TNP to a sf-TNP (Figure 7). For example, majority of experts felt that it was "likely" or "extremely likely" to observe a change in gum problems (74/101, 73%), phlegm or mucus while coughing or not coughing (72/101, 71%), general perception of well-being (72/101, 71%), and throat irritation/sore throat (72/101, 71%). In addition, between 5% and 20% considered that switching

XSL•FO
RenderX

was "unlikely" or "very unlikely" to result in a change in the severity of any single concept, including cardiovascular symptoms, withdrawal symptoms associated with quitting, shortness of breath, and worry about smoking-related diseases. Reasons given by experts for considering a change in a concept to be unlikely are listed in Table 4 and were mostly related to lack of perceived substantial differences between combustible TNPs and sf-TNPs.

Figure 7. Experts' opinions (percentage of experts) regarding likelihood of observing change in the severity of the final 20 items in round III of the Delphi panel.

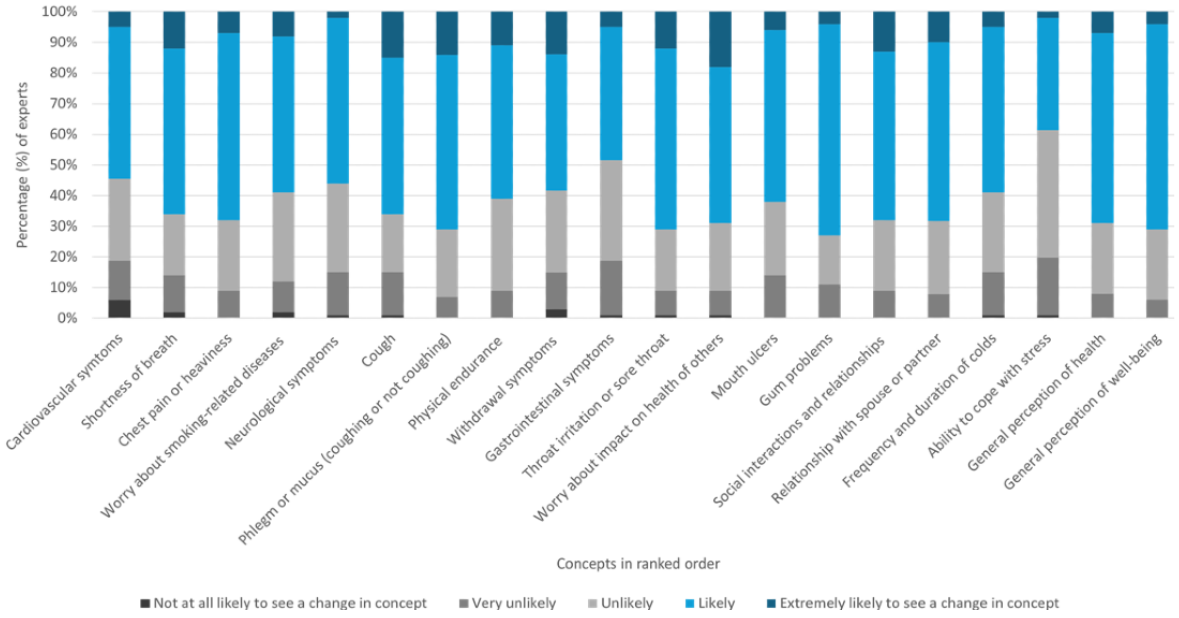

Table 4. Reasons given by experts for considering a change in a concept to be unlikely for the final 20 items in round III of the Delphi panel.

Concept (number of experts responding it was unlikely to see a change)	Rationale reported by experts
Cardiovascular symptoms (n=6)	• Unlikely to see substantial modifications between the combustible TNPs[a] and sf-TNPs[b] as tobacco or nicotine can cause cardiovascular disease • Active ingredients are the same and present in both combustible TNPs and sf-TNPs products
Shortness of breath or dyspnea (n=2)	• No change or no substantial modifications in switching to sf-TNPs
Worry about smoking-related diseases and impact on health in general (eg, cancer and stroke) (n=2)	• Symptoms will occur with using both combustible TNPs and sf-TNPs
Worry about impact on health of others (eg, children, partner, family, and friends) (n=1)	• Other dangerous chemicals in sf-TNPs
Frequency and duration of colds (n=1)	• Colds are unlikely to change as they are a nuisance and not a major health threat

[a]TNPs: tobacco and/or nicotine products.

[b]sf-TNPs: smoke-free tobacco and/or nicotine products.

Discussion

Principal Findings

This adapted Delphi panel study was conducted to identify the health and functioning concepts that experts consider to be most relevant and useful for incorporation into a new TNP-specific health and functioning measurement instrument to assess the self-reported impact of switching from combustible TNPs to sf-TNPs [21]. We recruited a geographically and professionally representative panel of experts who routinely treated those who smoke or communicated on smoking and TNP use. They were asked to reduce and refine a preliminary list of health and functioning concepts and to rank the final list in order of importance. After 3 rounds of investigation, the initial list of 69 concepts was refined to the most important 20 they considered could be included in a new self-reported measure (Table 5). The range of final concepts reflected health and psychosocial issues associated with smoking, including respiratory and cardiovascular health, withdrawal symptoms, and worries about the impact on health of self and others.

XSL•FO
RenderX

Table 5. Concepts recommended and agreed after each round of the Delphi panel.

Rank	Round I (n=69), ranked in order of importance (highest to lowest)	Round II (n=36), ranked in order of importance (highest to lowest)	Round III (n=20), ranked in order of importance (highest to lowest)
1	Worry about risk of cancer	Cardiovascular symptoms (eg, palpitations, blood pressure)	Cardiovascular symptoms (eg, palpitations, blood pressure)
2	Shortness of breath or dyspnea	Shortness of breath or dyspnea	Shortness of breath or dyspnea
3	Worry about health	Worry about smoking-related diseases and impact on health in general (eg, cancer, stroke)	Chest pain or heaviness
4	Anxiety symptoms due to withdrawal	Chest pain or heaviness	Worry about smoking-related diseases and impact on health in general (eg, cancer, stroke)
5	Irritability due to withdrawal	Neurological symptoms (eg, dizziness, nausea, headaches)	Neurological symptoms (eg, dizziness, nausea, headaches)
6	Strong craving for TNP[a] due to withdrawal	Withdrawal symptoms associated with quitting TNPs	Cough
7	Depression symptoms due to withdrawal	Cough	Phlegm or mucus while coughing or not coughing
8	Stress or tension due to withdrawal	General perception of health	Physical endurance during everyday activities and exercise
9	Ability to control moods and emotions	Phlegm or mucus while coughing or not coughing	Withdrawal symptoms associated with quitting TNPs
10	Aggression due to withdrawal	Gastrointestinal symptoms (eg, heartburn, constipation, diarrhea)	Gastrointestinal symptoms (eg, heartburn, constipation, diarrhea)
11	Worry about impact on pregnancy or fertility	Physical endurance during everyday activities and exercise	Throat irritation or sore throat
12	Concerns about the impact on the whole family	Worry about impact on health of others (eg, children, partner, family, friends)	Worry about impact on health of others (eg, children, partner, family, friends)
13	Concerns about impact on own children	Ability to cope with stress	Mouth ulcers
14	Pain or burning sensation in the teeth, gums, mouth, lips, throat, or tongue	Mouth ulcers	Gum problems (bleeding, pain, redness, swelling)
15	Bringing up phlegm or mucus while coughing or not coughing	Gum problems (bleeding, pain, redness, swelling)	Social interactions and relationships with children, family, friends, colleagues
16	Cough	Social interactions and relationships with children, family, friends, colleagues	Relationship with spouse or partner
17	Exercise capacity	Sleeping problems	Frequency and duration of colds
18	Anger due to withdrawal	General perception of well-being	Ability to cope with stress
19	Craving relief with TNP use	Ability to have or perform sexual activities	General perception of health
20	Impact on intimate relationships	Relationship with spouse or partner	General perception of well-being
21	Lack of concentration due to withdrawal	Throat irritation or sore throat	—
22	Ability to cope with stress	Frequency and duration of colds	—
23	Swollen or bleeding cheeks, lips, gums, or tongue	Ability to control moods and emotions	—
24	Ability to enjoy life	Hoarseness or change in voice	—
25	Concerns about the impact on own spouse or partner	Ability to concentrate or focus	—
26	Poor oral health in general	Weight control	—
27	Impact on relationships with children	Fatigue	—
28	Hoarseness or change in voice	Bad breath or halitosis	—
29	Ability to concentrate or focus	Feelings of self-esteem and self-respect	—
30	Yellow, gray, or black teeth	Feeling energized	—
31	Loss of tooth or teeth	Sense of smell and taste	—

XSL•FO
RenderX

Rank	Round I (n=69), ranked in order of importance (highest to lowest)	Round II (n=36), ranked in order of importance (highest to lowest)	Round III (n=20), ranked in order of importance (highest to lowest)
32	Ability to have and maintain an erection	Yellow, gray, or black teeth	—
33	Ability to pay attention, attention span	Plaque or tartar build up	—
34	Feelings of self-esteem and self-respect	Wrinkling, yellow, and dry skin	—
35	Symptoms of fatigue	Odor on clothes, hair, or body, due to TNP use	—
36	Impact on social life	Yellow brittle nails and dry hair	—
37	Sore throat	—	—
38	Heartburn (burning sensation in the stomach)	—	—
39	Sleeping problems	—	—
40	Bad breath or halitosis	—	—
41	Improved weight control with TNP use	—	—
42	Working memory	—	—
43	Perceives a stigma due to TNP use	—	—
44	Opportunity to relax or take a break while using TNP	—	—
45	Ability to have sexual activity—with self or others	—	—
46	Acid reflux	—	—
47	Enjoyment or pleasure while using TNP	—	—
48	Plaque or tartar build up	—	—
49	Productivity at work or home	—	—
50	Throat clearing	—	—
51	Increased concentration or focus with TNP use	—	—
52	Lost sense of taste	—	—
53	Wearing away of a tooth or teeth	—	—
54	Lost sense of smell	—	—
55	Bodily pain	—	—
56	Isolated by others	—	—
57	Impact on co-workers	—	—
58	Self-isolation	—	—
59	Skin tone quality (yellow or gray fingers or face)	—	—
60	Bad odor on clothes, hair, or body due to TNP use	—	—
61	Ability to reach orgasm	—	—
62	Parenting skills	—	—
63	Enjoyment of TNP taste	—	—
64	Participation in recreational activities	—	—
65	Fitting in with others who use TNP	—	—
66	Impact on friendships	—	—
67	Concerns about safety relating to risk of fire	—	—
68	Wrinkling skin	—	—
69	Likelihood of getting a cold	—	—

[a]TNP: tobacco and/or nicotine products.

Overall, the final 20 concepts were considered both clinically relevant and important to the experts in their evaluation of the possible impact of switching to sf-TNPs or stopping smoking cigarettes. We observed that the concepts consistently ranked highly among experts focused on objectively measurable health consequences such as respiratory and cardiovascular symptoms,

XSL•FO
RenderX

as well as overall physical functioning. In contrast, concepts such as the ability to cope with stress, risk of getting colds, oral health, physical appearance, social functioning, and sensory impacts (taste, sense of smell, and odor) were ranked as less important. Those concepts ranking highest reflect current evidence regarding smoking-related health outcomes and potential improvements attainable by ceasing cigarette consumption or switching to sf-TNPs [6,41-46]. The experts' knowledge of evidence for the effect of cessation or switching on pre-existing respiratory conditions [6,43,44,47,48] may also be reflected in the ranking.

It is essential to note that the relative ranking of these concepts is likely to be different for TNP consumers compared with health care professionals. Individuals trying to stop smoking or switch to sf-TNPs recognize the respiratory and other physical health and functional benefits and understand the potential for improved quality of life by reducing cigarette consumption or stopping smoking [49,50], as well as reduced exposure of others to smoking-related harm [6,51]. They also tend to report benefits and preferences for general hygiene and smell, better oral health, and sensory improvements [6,41,43]. However, no concepts related to general physical appearance or hygiene were in the final top-ranking concepts of the current study. These results explicitly reflect the participants' own experiences and opinions, which may differ from those of TNP users and be dependent on TNP user characteristics.

Although objective health outcomes such as cardiovascular and respiratory function were rated highly by experts throughout the study, some individuals rated psychosocial outcomes and effects on families and others (through secondhand smoke and so on) as equally important as objective physical measures. Latent class analysis in the final round of the study suggested that geographical location, cultural milieu, and professional specialism may each play a major role in these observations. Specifically, we identified a subgroup that focused mostly on physical, objectively measurable concepts (eg, cardiovascular symptoms, shortness of breath, chest pain, neurological symptoms, and cough), compared with the other subgroup, which considered both objective and subjective or emotional concepts (eg, worry about smoking-related diseases and impact on health both of self and others) to be of equal importance. These results may correspond with established patterns of perceptions and considerations among Western cultures, which may be broadly defined as individualistic (ie, driven by personal goals), whereas those from Asia and South America can be defined as collectivist (reflecting the primacy of mutual obligations among members of society) [51-53]. Furthermore, qualitative investigation of alignment with the subgroups revealed that for certain job roles (eg, counselors and policy advisors), the consumer-focused concepts were more important than for other health care professionals (eg, specialist clinicians), who were primarily concerned with objective, physical concepts. These findings have important implications for the development of a self-report measure that can be widely used and is adaptable to individual or local requirements [54]. When disseminating results based on the newly developed self-reported measure, it may be essential to tailor communications on sf-TNPs for specific audiences with different considerations of what is clinically meaningful to the experts themselves, as well as to the individual TNP user [26,33,54]. To validate the latent class findings in round III, we asked respondents to self-identify with the most relevant of the 2 subgroups. Overall, participants were able to do so readily, indicating that the 2 classes were culturally and professionally relevant, and that the new self-report measure will maximize content validity by including the range of concepts identified and ranked in round III to accommodate cultural and professional requirements and differences.

There are 3 main strengths of this study. First, we used the MaxDiff ranking to address the challenge of ranking an initial high number of concepts. Second, the incorporation of an anchor was essential in objectively identifying concepts for elimination. Third, a large panel of experts representing a broad range of professional expertise and geographies was recruited for this study. This allowed us to confirm validity of the concepts identified in the preparatory phase of the development of the new measure. It also enabled us to parse out cultural and professional subtleties that are important to consider when disseminating outcomes from the new measure. Development of an outcome measure involving both the target population and clinicians is a key component of creating a high-quality instrument [33,54]. Building this manner of collaboration and bridging between TNP users and relevant experts in this Delphi panel into the development of this new instrument would serve to enhance the potential quality and validity of the final measure by feeding into subsequent qualitative and psychometric evaluations [26].

There are also some limitations to our study. First, we did not use a strict Delphi panel process, and round I may be seen as a preliminary evaluation rather than a pure ranking and scoring exercise. Delphi panel methodology typically follows a highly structured process. In this particular case, it would have been ideal to perform additional preliminary qualitative insight work to review all the elements of a study, formally review and refine concept terminology, and test with external audiences. And, as previously mentioned, this Delphi panel focused on the experience and perceptions of experts and professionals who interact with TNP users. Consequently, the findings reported here do not have the vital context of TNP consumers opinions and perceptions. Unfortunately, it would have been impractical to address both angles in a single study with an initial list of 69 concepts. Instead, further research should consider evaluating TNP users' perceptions and rankings in a separate study using similar methodology. In addition, there is a potential for selection bias in the study due to the exclusion of experts who specified they would not consider participating in studies by tobacco companies. This exclusion could result in a sample that is not fully representative of the broader expert community, potentially impacting the generalizability of the findings. To mitigate this, we used a diverse recruitment strategy to include experts from various regions and specialties. However, the views of nonparticipating experts might differ from those who participated, and this limitation should be considered when interpreting the results. Finally, the survey was conducted in various languages, and despite the measures taken to ensure the consistency and validity of the translated surveys, linguistic and cultural differences may have led to variations in interpretation

XSL•FO
RenderX

of the survey. These variations could potentially affect the study's findings and should be considered when interpreting the results.

Conclusion

In conclusion, this 3-round, adapted Delphi panel identified a ranked list of 20 concepts to be considered when assessing the health and functioning status of individuals who stop smoking cigarettes or switch to using sf-TNPs or NRTs. It is to be expected that the scale of importance of each concept will vary based on the health status and concerns of an individual, whereas the ranking presented here represents a global, generalized view provided by the participating experts. In addition, the sensitivity of the concepts to accurately reflect changes in TNP use behavior will need to be determined. This would support the evaluation of the self-reported experience and impact of switching from conventional cigarettes to sf-TNPs on health risks and contribute to the regulatory and scientific evidence base for understanding both the individual and public health impacts of sf-TNPs.

Acknowledgments

We wish to thank David Floyd for providing editorial assistance. Philip Morris International is the sole source of funding and sponsor of this research.

Data Availability

The datasets generated during and/or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

The authors declare the following financial interests or personal relationships which may be considered as potential conflicting interests: EA and CC are employees of Philip Morris Products S.A. VL, LAW, LW, and AT were contracted and paid by Philip Morris Products S.A.

Multimedia Appendix 1
Amended/modified Delphi panel approach.
[DOCX File , 38 KB - formative_v9i1e58614_app1.docx]

References

1. WHO report on the global tobacco epidemic 2021: addressing new and emerging products. Geneva, Switzerland: World Health Organization (WHO); 2021.
2. Mersha AG, Gould GS, Bovill M, Eftekhari P. Barriers and facilitators of adherence to nicotine replacement therapy: a systematic review and analysis using the capability, opportunity, motivation, and behaviour (COM-B) model. Int J Environ Res Public Health 2020;17(23):8895 [FREE Full text] [doi: 10.3390/ijerph17238895] [Medline: 33265956]
3. Food and Drug Administration (FDA). 2020. URL: https://www.fda.gov/tobacco-products/products-guidance-regulations/products-ingredients-components [accessed 2021-12-21]
4. Guidance for Industry - Modified Risk Tobacco Product Applications (Draft Guidance). Silver Spring, MD: Food and Drug Administration (FDA) Center for Tobacco Products; 2012.
5. Modified risk orders. Food and Drug Administration (FDA). URL: https://www.fda.gov/tobacco-products/advertising-and-promotion/modified-risk-orders [accessed 2021-12-21]
6. Glasser AM, Collins L, Pearson JL, Abudayyeh H, Niaura RS, Abrams DB, et al. Overview of electronic nicotine delivery systems: a systematic review. Am J Prev Med 2017;52(2):e33-e66. [doi: 10.1016/j.amepre.2016.10.036] [Medline: 27914771]
7. Hajek P, Phillips-Waller A, Przulj D, Pesola F, Smith KM, Bisal N, et al. E-cigarettes compared with nicotine replacement therapy within the UK stop smoking services: the TEC RCT. Health Technol Assess 2019;23(43):1-82 [FREE Full text] [doi: 10.3310/hta23430] [Medline: 31434605]
8. McRobbie HJ, Phillips-Waller A, El Zerbi C, McNeill A, Hajek P, Pesola F, et al. Nicotine replacement treatment, e-cigarettes and an online behavioural intervention to reduce relapse in recent ex-smokers: a multinational four-arm RCT. Health Technol Assess 2020;24(68):1-82 [FREE Full text] [doi: 10.3310/hta24680] [Medline: 33270009]
9. Goldenberg M, Danovitch I, IsHak WW. Quality of life and smoking. Am J Addict 2014;23(6):540-562. [doi: 10.1111/j.1521-0391.2014.12148.x] [Medline: 25255868]
10. Livingston G, Huntley J, Sommerlad A, Ames D, Ballard C, Banerjee S, et al. Dementia prevention, intervention, and care: 2020 report of the lancet commission. Lancet 2020;396(10248):413-446 [FREE Full text] [doi: 10.1016/S0140-6736(20)30367-6] [Medline: 32738937]
11. Qasim H, Karim ZA, Rivera JO, Khasawneh FT, Alshbool FZ. Impact of electronic cigarettes on the cardiovascular system. J Am Heart Assoc 2017;6(9):e006353 [FREE Full text] [doi: 10.1161/JAHA.117.006353] [Medline: 28855171]
12. Critchley JA, Unal B. Health effects associated with smokeless tobacco: a systematic review. Thorax 2003;58(5):435-443 [FREE Full text] [doi: 10.1136/thorax.58.5.435] [Medline: 12728167]

XSL•FO
RenderX

13. Znyk M, Jurewicz J, Kaleta D. Exposure to heated tobacco products and adverse health effects, a systematic review. Int J Environ Res Public Health 2021;18(12):6651 [FREE Full text] [doi: 10.3390/ijerph18126651] [Medline: 34205612]
14. Marques P, Piqueras L, Sanz M. An updated overview of e-cigarette impact on human health. Respir Res 2021;22(1):151 [FREE Full text] [doi: 10.1186/s12931-021-01737-5] [Medline: 34006276]
15. Institute of Medicine, Board on Population Health and Public Health Practice, Committee on Scientific Standards for Studies on Modified Risk Tobacco Products. Scientific Standards for Studies on Modified Risk Tobacco Products. Washington, DC: National Academy of Science; 2012.
16. Gasperini B, Barbadoro P, Cherubini A, Pierri F, D'Errico MM, Di Stanislao F, et al. The earlier the better: health-related quality of life is inversely related to age of smoking cessation in older former Italian smokers. Aging Clin Exp Res 2017;29(4):655-663. [doi: 10.1007/s40520-016-0608-4] [Medline: 27456679]
17. Sarna L, Bialous SA, Cooley ME, Jun H, Feskanich D. Impact of smoking and smoking cessation on health-related quality of life in women in the nurses' health study. Qual Life Res 2008;17(10):1217-1227. [doi: 10.1007/s11136-008-9404-8] [Medline: 18931942]
18. Kulasekaran A, Proctor C, Papadopoulou E, Shepperd CJ, Guyer R, Gandek B, et al. Preliminary evaluation of a new German translated tobacco quality of life impact tool to discriminate between healthy current and former smokers and to explore the effect of switching smokers to a reduced toxicant prototype cigarette. Nicotine Tob Res 2015;17(12):1456-1464. [doi: 10.1093/ntr/ntv024] [Medline: 25914263]
19. Olufade AO, Shaw JW, Foster SA, Leischow SJ, Hays RD, Coons SJ. Development of the smoking cessation quality of life questionnaire. Clin Ther 1999;21(12):2113-2130. [doi: 10.1016/s0149-2918(00)87242-2] [Medline: 10645757]
20. Ware JE, Gandek B, Kulasekaran A, Guyer R. Evaluation of smoking-specific and generic quality of life measures in current and former smokers in Germany and the United States. Health Qual Life Outcomes 2015;13:128 [FREE Full text] [doi: 10.1186/s12955-015-0316-3] [Medline: 26276447]
21. Chrea C, Acquadro C, Afolalu EF, Spies E, Salzberger T, Abetz-Webb L, et al. Developing fit-for-purpose self-report instruments for assessing consumer responses to tobacco and nicotine products: the ABOUT™ toolbox initiative. F1000Res 2018;7:1878 [FREE Full text] [doi: 10.12688/f1000research.16810.1] [Medline: 30906527]
22. Afolalu EF, Spies E, Bacso A, Clerc E, Abetz-Webb L, Gallot S, et al. Impact of tobacco and/or nicotine products on health and functioning: a scoping review and findings from the preparatory phase of the development of a new self-report measure. Harm Reduct J 2021;18(1):79 [FREE Full text] [doi: 10.1186/s12954-021-00526-z] [Medline: 34330294]
23. Afolalu EF, Spies E, Bacso A, Clerc E, Abetz-Webb L, Gallot S, et al. The process of developing a new outcome measure to assess the impact of tobacco and/or nicotine-containing products on health and functioning. 2020 Presented at: 26th Annual Meeting Society for Research on Nicotine and Tobacco (SRNT); March 11-14, 2020; New Orleans, LA.
24. Afolalu EF, Lanar S, Abetz-Webb L, Bacso A, Blankenspoor M, Koochaki P, et al. Qualitative health and functioning concept elicitation interviews in Japanese users of heated tobacco products. 2021 Presented at: 27th Annual Meeting Society for Research on Nicotine and Tobacco (SRNT); February 24, 2021; Virtual event URL: https://www.pmiscience.com/content/dam/pmiscience/en/publications/presentations-and-posters/pdf/srnt-2021-poster-esther-afolalu.pdf
25. Asfar T, Schmidt M, Ebrahimi Kalan M, Wu W, Ward KD, Nakkash RT, et al. Delphi study among international expert panel to develop waterpipe-specific health warning labels. Tob Control 2020;29(2):159-167 [FREE Full text] [doi: 10.1136/tobaccocontrol-2018-054718] [Medline: 30696784]
26. Hepworth LR, Rowe FJ. Using delphi methodology in the development of a new patient-reported outcome measure for stroke survivors with visual impairment. Brain Behav 2018;8(2):e00898 [FREE Full text] [doi: 10.1002/brb3.898] [Medline: 29484260]
27. Santaguida PL, Oliver D, Gilsing A, Lamarche L, Griffith LE, Mangin D, et al. Delphi consensus on core criteria set selecting among health-related outcome measures (HROM) in primary health care. J Clin Epidemiol 2020;127:105-116. [doi: 10.1016/j.jclinepi.2020.04.028] [Medline: 32417261]
28. Blaser J, Cornuz J. Experts' consensus on use of electronic cigarettes: a Delphi survey from Switzerland. BMJ Open 2015;5(4):e007197 [FREE Full text] [doi: 10.1136/bmjopen-2014-007197] [Medline: 25877274]
29. O'Neill N, Dogar O, Jawad M, Kellar I, Kanaan M, Siddiqi K. Which behavior change techniques may help waterpipe smokers to quit? An expert consensus using a modified Delphi technique. Nicotine Tob Res 2018;20(2):154-160 [FREE Full text] [doi: 10.1093/ntr/ntw297] [Medline: 27798091]
30. Lintonen T, Konu A, Rönkä S, Kotovirta E. Drugs foresight 2020: a Delphi expert panel study. Subst Abuse Treat Prev Policy 2014;9:18 [FREE Full text] [doi: 10.1186/1747-597X-9-18] [Medline: 24885142]
31. Marten O, Koerber F, Bloom D, Bullinger M, Buysse C, Christensen H, et al. A Delphi study on aspects of study design to overcome knowledge gaps on the burden of disease caused by serogroup B invasive meningococcal disease. Health Qual Life Outcomes 2019;17(1):87 [FREE Full text] [doi: 10.1186/s12955-019-1159-0] [Medline: 31118091]
32. O'Donnell S, Doyle G, O'Malley G, Browne S, O'Connor J, Mars M, et al. Establishing consensus on key public health indicators for the monitoring and evaluating childhood obesity interventions: a Delphi panel study. BMC Public Health 2020;20(1):1733 [FREE Full text] [doi: 10.1186/s12889-020-09814-y] [Medline: 33203390]

XSL•FO
RenderX

33. Jünger S, Payne SA, Brine J, Radbruch L, Brearley SG. Guidance on conducting and reporting DElphi studies (CREDES) in palliative care: recommendations based on a methodological systematic review. Palliat Med 2017;31(8):684-706. [doi: 10.1177/0269216317690685] [Medline: 28190381]
34. Strasser A. Design and evaluation of ranking-type Delphi studies using best-worst-scaling. Technol Anal Strateg Manag 2018;31(4):492-501. [doi: 10.1080/09537325.2018.1521956]
35. Uy EJB, Bautista DC, Xin X, Cheung YB, Thio S, Thumboo J. Using best-worst scaling choice experiments to elicit the most important domains of health for health-related quality of life in Singapore. PLoS One 2018;13(2):e0189687 [FREE Full text] [doi: 10.1371/journal.pone.0189687] [Medline: 29420564]
36. Mühlbacher AC, Kaczynski A, Zweifel P, Johnson FR. Experimental measurement of preferences in health and healthcare using best-worst scaling: an overview. Health Econ Rev 2016;6(1):2. [doi: 10.1186/s13561-015-0079-x] [Medline: 26743636]
37. Cheung KL, Wijnen BFM, Hollin IL, Janssen EM, Bridges JF, Evers SMAA, et al. Using best-worst scaling to investigate preferences in health care. Pharmacoeconomics 2016;34(12):1195-1209. [doi: 10.1007/s40273-016-0429-5] [Medline: 27402349]
38. Sawtooth S. The MaxDiff System Technical Paper (Version 8). Orem, UT: Sawtooth Software, Inc; 2013.
39. Lighthouse Studio Manual. Orem, UT: Sawtooth Software, Inc; 2013.
40. Kongsted A, Nielsen AM. Latent class analysis in health research. J Physiother 2017;63(1):55-58. [doi: 10.1016/j.jphys.2016.05.018] [Medline: 27914733]
41. Abafalvi L, Pénzes M, Urbán R, Foley KL, Kaán R, Kispélyi B, et al. Perceived health effects of vaping among Hungarian adult e-cigarette-only and dual users: a cross-sectional internet survey. BMC Public Health 2019;19(1):302 [FREE Full text] [doi: 10.1186/s12889-019-6629-0] [Medline: 30866897]
42. Benowitz NL, Burbank AD. Cardiovascular toxicity of nicotine: implications for electronic cigarette use. Trends Cardiovasc Med 2016;26(6):515-523. [doi: 10.1016/j.tcm.2016.03.001] [Medline: 27079891]
43. Farsalinos KE, Romagna G, Tsiapras D, Kyrzopoulos S, Voudris V. Characteristics, perceived side effects and benefits of electronic cigarette use: a worldwide survey of more than 19,000 consumers. Int J Environ Res Public Health 2014;11(4):4356-4373 [FREE Full text] [doi: 10.3390/ijerph110404356] [Medline: 24758891]
44. Polosa R, Cibella F, Caponnetto P, Maglia M, Prosperini U, Russo C, et al. Health impact of E-cigarettes: a prospective 3.5-year study of regular daily users who have never smoked. Sci Rep 2017;7(1):13825 [FREE Full text] [doi: 10.1038/s41598-017-14043-2] [Medline: 29150612]
45. Choi S, Lee K, Park SM. Combined associations of changes in noncombustible nicotine or tobacco product and combustible cigarette use habits with subsequent short-term cardiovascular disease risk among South Korean men: a nationwide cohort study. Circulation 2021;144(19):1528-1538 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.121.054967] [Medline: 34601948]
46. Sharman A, Nurmagambetov T. Changes in respiratory function and physical capacity among smokers after switching to iqos: one year follow-up. Glob. J. Respir. Care 2020;6:21-27 [FREE Full text]
47. Polosa R, Morjaria JB, Prosperini U, Busà B, Pennisi A, Malerba M, et al. COPD smokers who switched to e-cigarettes: health outcomes at 5-year follow up. Ther Adv Chronic Dis 2020;11:2040622320961617 [FREE Full text] [doi: 10.1177/2040622320961617] [Medline: 33101622]
48. Lucchiari C, Masiero M, Mazzocco K, Veronesi G, Maisonneuve P, Jemos C, et al. Benefits of e-cigarettes in smoking reduction and in pulmonary health among chronic smokers undergoing a lung cancer screening program at 6 months. Addict Behav 2020;103:106222 [FREE Full text] [doi: 10.1016/j.addbeh.2019.106222] [Medline: 31838445]
49. Laverty AA, Vardavas CI, Filippidis FT. Prevalence and reasons for use of heated tobacco products (HTP) in Europe: an analysis of eurobarometer data in 28 countries. Lancet Reg Health Eur 2021;8:100159 [FREE Full text] [doi: 10.1016/j.lanepe.2021.100159] [Medline: 34557853]
50. Skelton E, Guillaumier A, Tzelepis F, Walsberger S, Paul CL, Dunlop AJ, et al. Alcohol and other drug health-care providers and their client's perceptions of e-cigarette use, safety and harm reduction. Drug Alcohol Rev 2021;40(6):998-1002. [doi: 10.1111/dar.13276] [Medline: 33774886]
51. Xu SS, Meng G, Yan M, Gravely S, Quah ACK, Ouimet J, et al. Reasons for regularly using heated tobacco products among adult current and former smokers in Japan: finding from 2018 ITC Japan survey. Int J Environ Res Public Health 2020;17(21):8030 [FREE Full text] [doi: 10.3390/ijerph17218030] [Medline: 33142757]
52. Betsch C, Böhm R, Airhihenbuwa CO, Butler R, Chapman GB, Haase N, Nurm, et al. Improving medical decision making and health promotion through culture-sensitive health communication: an agenda for science and practice. Med Decis Making 2016;36(7):811-833. [doi: 10.1177/0272989X15600434] [Medline: 26296619]
53. Brick C, McCully SN, Updegraff JA, Ehret PJ, Areguin MA, Sherman DK. Impact of cultural exposure and message framing on oral health behavior: exploring the role of message memory. Med Decis Making 2016;36(7):834-843. [doi: 10.1177/0272989X15570114] [Medline: 25654986]
54. Gerritsen A, Jacobs M, Henselmans I, van Hattum J, Efficace F, Creemers G, et al. Developing a core set of patient-reported outcomes in pancreatic cancer: a delphi survey. Eur J Cancer 2016;57:68-77. [doi: 10.1016/j.ejca.2016.01.001] [Medline: 26886181]

XSL•FO
RenderX

Abbreviations

sf-TNP: smoke-free tobacco and/or nicotine product
TNP: tobacco and/or nicotine product
NRT: nicotine replacement therapy
FDA: Food and Drug Administration

Edited by A Mavragani; submitted 20.03.24; peer-reviewed by N Peraza; comments to author 12.09.24; revised version received 31.10.24; accepted 07.11.24; published 02.01.25.

Please cite as:
Law V, Afolalu EF, Abetz-Webb L, Wemyss LA, Turner A, Chrea C
International Expert Consensus on Relevant Health and Functioning Concepts to Assess in Users of Tobacco and Nicotine Products: Delphi Study
JMIR Form Res 2025;9:e58614
URL: https://formative.jmir.org/2025/1/e58614
doi:10.2196/58614
PMID:

©Vivienne Law, Esther F Afolalu, Linda Abetz-Webb, Lee Andrew Wemyss, Andrew Turner, Christelle Chrea. Originally published in JMIR Formative Research (https://formative.jmir.org), 02.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Development of an eHealth Intervention Including Self-Management for Reducing Sedentary Time in the Transition to Retirement: Participatory Design Study

Lisa Hultman[1], MSc; Caroline Eklund[1], PhD; Petra von Heideken Wågert[1], Prof Dr; Anne Söderlund[1], Prof Dr; Maria Lindén[2], Prof Dr; Magnus L Elfström[3], PhD

[1]Division of Physiotherapy, School of Health, Care and Social Welfare, Mälardalen University, Västerås/Eskilstuna, Sweden
[2]Division of Intelligent Future Technologies, School of Innovation, Design and Engineering, Mälardalen University, Västerås/Eskilstuna, Sweden
[3]Division of Psychology, School of Health, Care and Social Welfare, Mälardalen University, Västerås/Eskilstuna, Sweden

Corresponding Author:
Lisa Hultman, MSc
Division of Physiotherapy
School of Health, Care and Social Welfare
Mälardalen University
Box 883
Västerås/Eskilstuna, 721 23
Sweden
Phone: 46 21 48 06 901
Email: lisa.hultman@mdu.se

Abstract

Background: Having a great amount of sedentary time is common among older adults and increases with age. There is a strong need for tools to reduce sedentary time and promote adherence to reduced sedentary time, for which eHealth interventions have the potential to be useful. Interventions for reducing sedentary time in older adults have been found to be more effective when elements of self-management are included. When creating new eHealth interventions, accessibility and effectiveness can be increased by including end users as co-designers in the development process.

Objective: The aim was to explore the desired features of an eHealth intervention including self-management for reducing sedentary time and promoting adherence to reduced sedentary time in older adults transitioning from working life to retirement. Further, the aim was to develop a digital prototype of such an eHealth intervention.

Methods: The study used the participatory design approach to include end users, researchers, and a web designer as equal partners. Three workshops were conducted with 6 older adults transitioning to retirement, 2 researchers, and 1 web designer. Thematic analysis was used to analyze the data from the workshops.

Results: Participants expressed a desire for an easy-to-use eHealth intervention, which could be accessed from mobile phones, tablets, and computers, and could be individualized to the user. The most important features for reducing sedentary time were those involving finding joyful activities, setting goals, and getting information regarding reduced sedentary time. Participants expressed that the eHealth intervention would need to first provide the user with knowledge regarding sedentary time, then offer features for measuring sedentary time and for setting goals, and lastly provide support in finding joyful activities to perform in order to avoid being sedentary. According to the participants, an eHealth intervention including self-management for reducing sedentary time in older adults in the transition to retirement should be concise, accessible, and enjoyable. A digital prototype of such an eHealth intervention was developed.

Conclusions: The developed eHealth intervention including self-management for reducing sedentary time in older adults transitioning to retirement is intended to facilitate behavior change by encouraging the user to participate in autonomously motivated activities. It uses several behavior change techniques, such as goal setting and action planning through mental contrasting and implementation intention, as well as shaping knowledge. Its active components for reducing sedentary time can be explained using the integrated behavior change model. Further research is needed to evaluate the feasibility and effectiveness of the eHealth intervention.

(JMIR Form Res 2025;9:e63567) doi:10.2196/63567

XSL•FO
RenderX

KEYWORDS

behavior change intervention; adherence; integrated behavior change model; autonomous motivation; affective determinants

Introduction

Background

Increased sedentary time is common among older adults [1], and it increases the risk of noncommunicable diseases and premature death [2-4]. Sedentary time is the time spent in sedentary behavior for any duration or in any context [5]. Sedentary behavior can be defined as any waking behavior in a seated, reclined, or lying position with an energy expenditure of <1.5 metabolic equivalents (METs) [5]. With an aging population globally [6], there is an urgent need to reduce sedentary time in older adults. This has been found to be possible [7,8] but difficult, as sedentary behavior is complex and multifaceted [9]. In a recent scoping review [8], researchers found that sedentary time can be reduced in community-dwelling older adults using multicomponent interventions. Further, their results showed that information, goal setting, health coaching, self-monitoring, and action planning were included in effective interventions. However, studies reporting effect size reported only a mild to moderate reduction in sedentary time. Consequently, there is a need to develop new interventions with new approaches to reduce sedentary time in older adults.

The transition to retirement could be a favorable time for the introduction of interventions for reducing sedentary time in older adults. This is because retirement leads to changes in health behaviors such as physical activity, sedentary time, and smoking [10,11]. A recent longitudinal cohort study found that 61% of participants made health behavior changes when retiring [12], indicating a window where health behavior interventions, such as interventions for reduced sedentary time, could be timely. Previous research has found that interventions for health behavior change are more effective when elements of self-management are included [8]. Self-management is a person's own day-to-day management of health promotion or disease, including management of their medicines, emotions, meaningful behaviors, and roles [13].

Previous research has found that older adults often equate sedentary behavior with physical inactivity, and are unaware of the health consequences of prolonged sedentary time, how much time they spend in sedentary behavior, and when they are sedentary [9,14]. Affect has an impact on sedentary time in older adults, as enjoyment can be a motive for taking part in nonsedentary behavior [15] and sedentary behavior [9]. Older adults transitioning to retirement experience sedentary behavior as something related to health and well-being, giving them an opportunity to engage in meaningful activities, such as self-reflection, but also as something unhealthy and related to loneliness and boredom [16].

A recent study found that older adults in the transition to retirement use and desire self-management strategies that influence both affective determinants (related to emotions and affects such as joyfulness) and cognitive determinants (related to cognitive processes and includes attitudes and self-efficacy) for reducing sedentary time [17]. Earlier interventions for reducing sedentary time have primarily targeted cognitive determinants, resulting in positive changes with only a mild to moderate effect size [8]. Phipps et al [18] found that sedentary time is influenced more by explicit affective attitudes than by explicit instrumental attitudes. In recent years, interventions addressing affective determinants have been used in health promotion interventions targeting health behaviors [19,20]. Lithopoulos et al [21] compared affective, instrumental, and self-regulatory information in office workers and found that affective information had the most influence on reducing sedentary time. However, there are no intervention studies targeting both cognitive and affective determinants to reduce sedentary time in older adults transitioning to retirement.

To be effective, interventions must also have the possibility of reaching the intended end users. eHealth, which involves health services and information delivery through the internet and internet-related technologies [22], might offer an opportunity to reach out to large populations in upper middle– and high-income countries, as the rates of internet use and access in these countries are 83% and 93%, respectively [23]. eHealth interventions include but are not limited to mobile device–delivered health interventions (mHealth), text messaging, use of stationary computers and the internet, and telephone calls. A recent systematic review and meta-analysis of the effect of mobile health app interventions on health behaviors in older adults stated that such eHealth interventions can be used to reduce sedentary time, but there is uncertainty regarding the results owing to small sample sizes [24]. Adherence to eHealth interventions is slightly higher than adherence to traditional interventions, which might be due to eHealth interventions being perceived as more fun [25]. Adherence can be further raised through increased acceptability of the intervention by involving older adults in the development process using participatory research design [26]. Involving end users in the development of eHealth interventions is a common approach [27,28] and can increase the quality of the knowledge created, effectiveness of the intervention developed [28], and accessibility among those who are unaccustomed to eHealth [29].

Aim

The aim was to explore desired features of an eHealth intervention including self-management for reducing sedentary time and promoting adherence to reduced sedentary time in older adults transitioning from working life to retirement. Further, the aim was to develop a digital prototype of such an eHealth intervention.

Methods

Design

The study applied participatory design (PD) [26] to allow for the knowledge of older adults transitioning to retirement, researchers, and a web designer to be utilized and valued equally in the eHealth intervention's development. PD is a research method in which end users are involved as equal partners in the

XSL•FO
RenderX

design process [26] and has previously been used to develop eHealth interventions targeting various health behaviors [30-32]. This study is part of a PD project aimed at health promotion through reduced sedentary time in the transition to retirement. The PD project consists of the following 4 phases: need assessment, idea generation, testing and retesting, and evaluation. This study involves the second phase, idea generation, in which needs and opinions are evaluated in order to generate ideas and develop a digital prototype [26].

Participants

Participants, who were older adults transitioning to retirement, were recruited by the first author (LH) at theme days and meetings held by retirement organizations and municipalities, and through an interview article in the local newspaper. The criteria for inclusion were as follows: age 60-75 years, ability to speak and understand Swedish, ability to read and comprehend the study instructions, and plan to retire within a maximum of 3 years or having been retired for a maximum of 5 years. The exclusion criteria were as follows: presence of a serious disease, severe loss of vision, and severe loss of communicative ability. All recruited participants were newly retired, and none were still working. For further background information regarding the participants, see Tables 1 and 2.

Table 1. Participant (older adult) demographics.

Participant	Age, years	Sex	Time retired, years	Mostly sedentary or nonsedentary at the current or last job	Living with other adults	Type of housing	Self-reported daily sedentary time, hours	Previous use of an eHealth intervention including self-management
Participant 1: Older adult	67	Male	4	Sedentary	Yes	House/townhouse	7-8	No
Participant 2: Older adult	68	Female	3	Nonsedentary	No	Apartment	2	Do not know
Participant 3: Older adult	67	Male	2	Sedentary	No	Apartment	8	No
Participant 4: Older adult	67	Female	2	Nonsedentary	Yes	House/townhouse	5	App-based arthritis treatment
Participant 5: Older adult	70	Female	4	Nonsedentary	No	Apartment	6-8	__[a]
Participant 6: Older adult	64	Female	2	Sedentary	Yes	House/townhouse	5	Activity watch

[a]Not applicable.

Table 2. Participant (designer and researcher) demographics.

Participant	Age, years	Sex	Description
Participant 7: Researcher	38	Female	Physiotherapist (PhD in physiotherapy) with a research focus on eHealth and participatory design
Participant 8: Researcher	37	Female	Physiotherapist (master's degree in clinical medical science) with experience in primary and specialist health care
Participant 9: Web designer	39	Female	Adequate education and several years of experience in web development

A workshop group consisting of 9 participants was convened. The group's size was selected to be large enough to capture different user perspectives and small enough to allow all participants' voices to be heard. Of the 9 participants, 6 were older adults in the transition from working life to retirement, 2 were researchers participating in the workshop (LH and CE), and 1 was a web designer (Tables 1 and 2). The workshops were led by an experienced workshop moderator to improve the balance of power among participants. The workshop moderator was a female researcher employed by the same university as the authors.

Ethical Considerations

Prior to recruitment, the study was reviewed and approved by the Swedish Ethical Review Authority (Dnr 2019-03836). Informed consent was obtained from the participants.

Workshops and Data Collection

Three workshops were conducted between November 2021 and January 2022 at time intervals of 2 and 8 weeks. All participants were invited to all 3 workshops (see Figure 1 for attendance). The workshops were held in a meeting room at a university. At the beginning of the first workshop, the participants were given written and verbal information about the study, and the study aim and the focus of each workshop were presented (Figure 1), as were the roles of the participants and the moderator. Further the participants were informed that their participation was

XSL•FO
RenderX

voluntary and that they could choose to end their participation at any time. Informed consent was signed. The first workshop was planned based on the results of a previous study [17]. In this workshop, focus group interviews were performed with older adults transitioning to retirement, and they involved perceptions of self-management strategies for reducing sedentary time and adhering to the reduced sedentary time. PD tool cards [33] were used at the first workshop, with participants being asked to list on Post-it notes which features in a digital solution they judged to be necessary to operationalize the self-management strategies for reducing sedentary time. This was followed by a discussion. Workshop 2 used 2D mapping [33], taking its starting point in a mind map of features developed from workshop 1. The third workshop used the PD tool user journey [33], with participants working based on a paper prototype developed after workshop 2. Participants were also asked to prioritize the importance of the features, voting for the first, second, and third most important features for reducing sedentary time. The voting was conducted using Post-it notes, with different colored notes used to represent the first, second, and third most important features. After the prioritization of the features, drafts of each feature of the paper prototype were presented, and suggestions for changes were discussed.

Figure 1. Workshop process map. BWS: between-workshop.

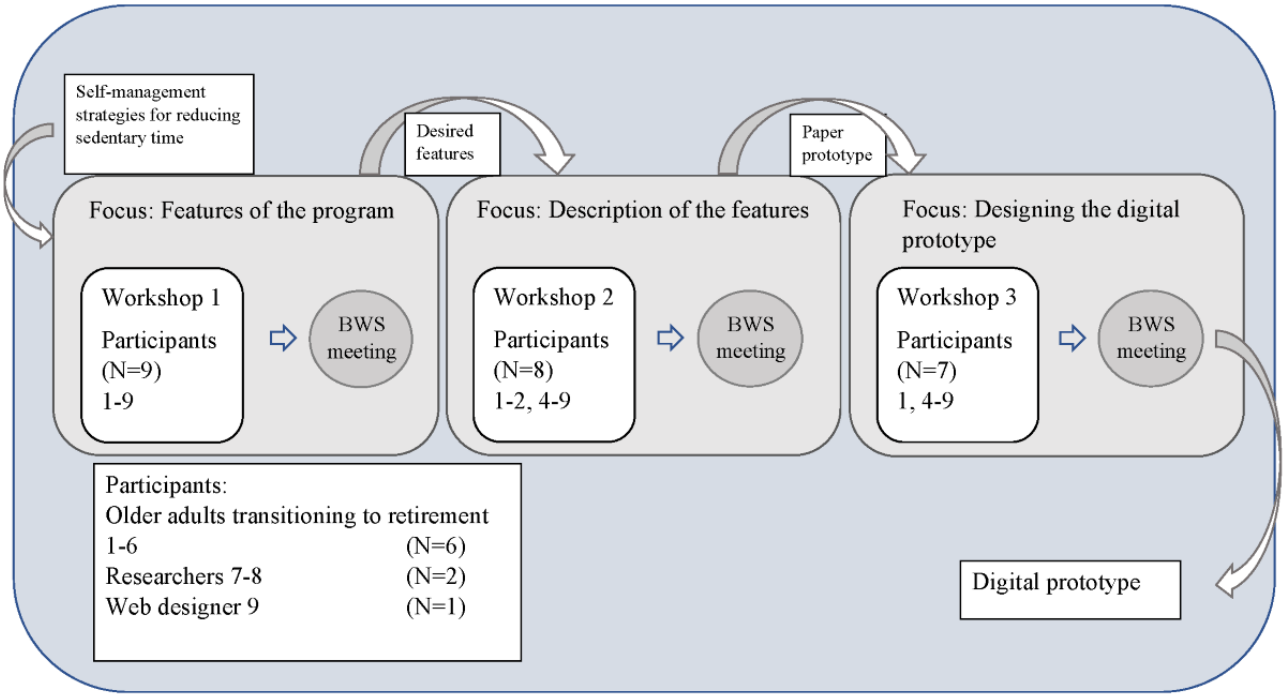

After each workshop, the first author (LH) went through the Post-it notes and photographs and listened to audio recordings to create mind maps and outline possible features of the eHealth intervention. Audio recordings were used at this stage to clarify and deepen the understanding of the text written on the Post-it notes. Between-workshop meetings were conducted with the 2 researchers participating in the workshop and 4 additional researchers (co-authors PHW, AS, ML, and MLE) with extensive experience in developing eHealth interventions, behavior change interventions, and gerontologic research. Preliminary results and proposed solutions were presented at the between-workshop meetings, and ideas and questions for the next workshop were discussed. After the last workshop, a digital prototype was developed.

Data Analysis

After the last workshop, the audio recordings from the workshops were transcribed verbatim. The transcripts were analyzed using inductive thematic analysis [34], with the analysis conducted in 2 steps based on the study aim and the focus of each workshop. The analyses of workshops 1 and 2 focused on desired features and their descriptions (feature analysis). This served as a verification of the paper prototype presented at workshop 3. The analysis of workshop 3 aimed to describe the necessary changes and the participants' desires regarding the final layout of the digital prototype (change analysis). The first author listened to the audio recordings and read the transcripts to get familiar with the dataset, after which relevant segments were coded at a semantic level. This was in accordance with phases 1 and 2 of the 6 phases of thematic analysis. In phase 3, initial themes were created. These were then checked against the whole data set, and meetings were held with all authors where the themes were discussed (phase 4). The themes were then defined and named (phase 5). Although writing had been carried out since phase 3, the analysis proceeded into the writing of the manuscript (phase 6) as themes were then clarified. The themes from the analysis can be found in Multimedia Appendix 1. The results from the prioritization of the features were summarized and used as a basis for discussion during workshop 3.

XSL•FO
RenderX

Results

Overview

The feature analysis resulted in the following 2 themes: general features and features for self-management. The latter consisted of 9 subthemes: finding activities that arouse joy, support in goal setting, information regarding sedentary behavior, measure sedentary time and receive feedback, schedule and planning, rewards, reminders, interact with other users, and timer. These subthemes corresponded to the features of the paper prototype presented at workshop 3. The change analysis resulted in 3 additional themes: evoke positive affect, less is more, and packaging.

The prioritization of the features revealed that the 3 most highly prioritized features in self-management for reducing sedentary time were finding activities that arouse joy, support for goal setting, and information regarding sedentary behavior. More information, such as the order in which participants wanted the features to be presented to the user, can be found in Figure 2.

Figure 2. Prioritization of features and time of introduction and inclusion of features in the paper prototype. The gray area represents functions not included in the paper prototype.

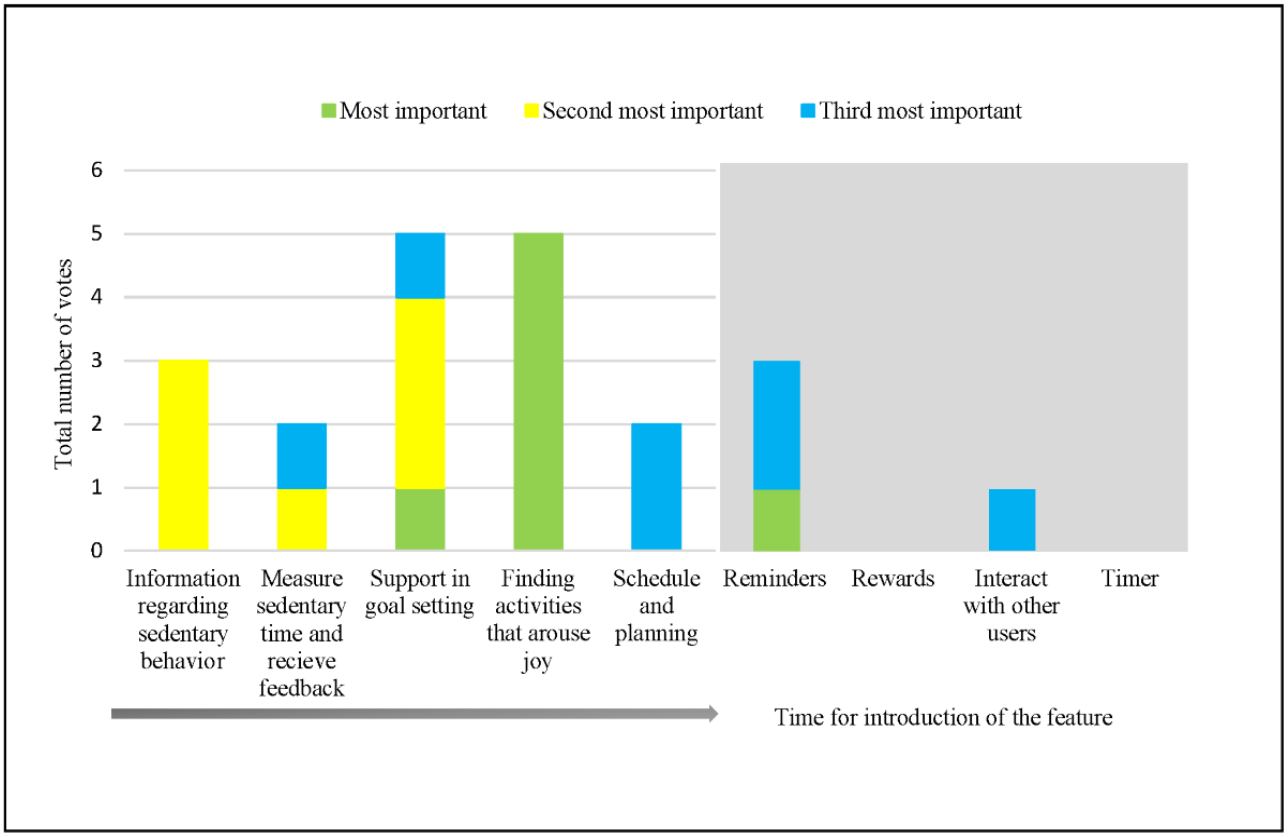

Feature Analysis

General Features

Participants wanted the eHealth intervention to include basic features, with the possibility to add more features for those who are more technically advanced.

> *I mean, there are so many people on their way into retirement and there are even more who are retired today who have a very low digitalization level... Yeah, it has to be incredibly simple. Has to be basal. Then you should be able to build on it, maybe, but... The foundation has to be based on something very, very simple.* [Participant in workshop 2]

To facilitate use, the user should be guided in the order in which the features should be used. To increase accessibility, the eHealth intervention should be available on mobile phones, tablets, and computers. There was concern regarding the inclusion of too many features, as this could lead to increased screen time rather than reduced sedentary time.

> *The more and more you build into it, it makes it so that, now I have to take a look.* [Participant in workshop 1]

To further facilitate use, an optional tutorial guiding the user through how the eHealth intervention worked was suggested.

Individualization according to the user's interests and affections, such as joy, was considered crucial for using the eHealth intervention, reducing sedentary time, and adhering to reduced sedentary time. At the same time, there was concern that individualization would result in the user having to answer too many questions at an early stage, which might exhaust them and prevent their further use of the eHealth intervention.

> *Yeah, it's... as soon as you have to go in and start making these kinds of choices that very many people stop.* [Participant in workshop 2]

XSL•FO
RenderX

Features for Self-Management

Finding Activities That Arouse Joy

To help them reduce their sedentary time, participants wanted the eHealth intervention to offer suggestions for joyful activities. Since what arouses feelings of joy is individual, participants saw a need to include different types of activities that could be performed in the home, outside the home, by oneself, or together with others, as part of the individualization. They expressed that the eHealth intervention should both allow the user to reflect on what activities they are interested in or could appreciate and provide suggestions for activities based on their areas of interest.

> *You get to answer questions about what you thought was interesting before and in that way be guided to... where you could look or what more you should do.* [Participant in workshop 2]

It should also provide information regarding what there is to do nearby when one is newly retired and facilitate contact with retirement organizations through links, information, and a contact person from the organization or activity. Participants expressed a wish for the activities offered by organizations to be presented in the eHealth intervention with their time and location, although they realized the difficulty that gathering this information would entail. Additionally, they wanted the eHealth intervention to include what activities other newly retired people do and enjoy, and provide the user with tips from other newly retired people for fun ways to reduce sedentary time. The difficulty in knowing what activities are available in the local area was expressed by 1 participant as follows:

> *I think a lot of people have problems with it... How in the world do I find things? What's here and where [is it], how do I get in contact with groups?* [Participant in workshop 1]

Support in Goal Setting

Participants considered goal setting to be important for reducing sedentary time, but described formulating goals as difficult. They suggested that it could be made easier by offering the user predetermined options or levels of sedentary time as goals, with these goals formulated based on either the maximum time to be spent in sedentary behavior or the number of activities to perform.

> *Can there be, like: Level 1, Level 2, Level 3, so you don't have to sit and make it up yourself, sort of?*
>
> *(You're thinking that there should be suggestions for goals?)*
>
> *Yeah, maybe if you're, you know, completely sedentary. What's Level 1. Yep. I finished that on the first day. Level 2, and so on.* [Participant in workshop 2]

Although goal setting was thought to be important for reducing sedentary time, a desire was expressed that this be made optional in the eHealth intervention. Participants expressed that, when formulating a goal regarding reduced sedentary time, they would like to be offered suggestions for what to do instead, and at follow-up, if the goal was not fully achieved, they would like the eHealth intervention to propose adapted suggestions to help them fulfill it in the future.

Information Regarding Sedentary Behavior

A need was expressed for information regarding sedentary behavior to be presented to the user at an early stage, so they would know why prolonged sedentary time should be avoided. In addition, participants wanted the information to be offered to the user both as optional short notifications and as static text to be found within the eHealth intervention. They expressed that they would like the information to include the benefits of not engaging in prolonged sedentary time, how long a person can stay sedentary before it is considered unhealthy, and tips on how sedentary time can be reduced. The participants stated that information must be formulated in a positive manner throughout the eHealth intervention. During one of the workshops, participants discussed information regarding sedentary behavior as follows:

> *That it's actually stated then that, where's this limit [for how long you can be sedentary]. That you get that information.* [Participant in workshop 2]

Measure Sedentary Time and Receive Feedback

To set realistic goals, it is necessary to first measure one's sedentary time. Further, participants saw a need to measure their sedentary time again at goal follow-up. This measurement could be done on specific days rather than daily to reduce the burden on the user. Different ways of taking the measurement were discussed, and while the participants saw the benefit of using sensors in an eHealth intervention, this would require users to always carry their smartphone, which was considered inconvenient. A watch or other device could be connected to the eHealth intervention, but it was thought that this would entail costs that would be unaffordable for some. Participants considered self-reporting one's sedentary time to be an appropriate solution. Suggested features for self-reporting sedentary time included tapping a stopwatch in the eHealth intervention when one sat down and tapping it again when one stood up, or estimating the time one spent in sedentary behavior. Using a stopwatch was considered unfavorable due to the risk of forgetting or tapping it in the wrong way, and participants did not want to have to carry their phone with them and tap it all day. The estimate of the time one spent in sedentary behavior could be given as an average for a typical day or as an estimate of the current day given in the evening. Participants expressed that after the user had measured their sedentary time, they would need feedback from the eHealth intervention to know whether they had spent too much time in sedentary behavior.

> *Then you get statistics on how you've been sitting. How long you've been sitting and how it's affected your body... Then you learn, okay, now I have to change my habits, if you're given information.* [Participant in workshop 2]

Schedule and Planning

Participants wanted a schedule or calendar to support their planning and facilitate new routines. It was important for the schedule to be easy to handle, with just a few tabs. It would need to be close at hand. For some, this meant being in their

XSL•FO
RenderX

phone, while for others, it meant a printed copy to put on their fridge. Participants said they would like a connection between the proposed activities and the schedule, so that activities could easily be placed in the schedule from the activity menu. For follow-up, the participants wanted the opportunity to check off those activities they had performed as planned.

> *I think this, that it's connected to the schedule, is quite good... because if you have to physically enter it in yourself. Then it's. It's a form of resistance and an extra step.* [Participant in workshop 2]

Rewards

Participants wanted the eHealth intervention to include rewards at goal fulfillment and as notifications in order to increase their motivation and adherence, and considered it important that the rewards lead to feelings of joy. Which rewards led to joy was said to be highly individual; for instance, some said they would like to have virtual stars and trophies in the eHealth intervention, while others suggested gift cards in line with the user's interests. To solve the problem of adapting to these differences, participants suggested that the eHealth intervention should urge users to treat themselves with a reward, with the possibility for users to specify in advance what this treat should be. While there was concern that this treat would be something that was unhealthy for the user, it was also discussed that engaging in a fun or amusing activity was rewarding and something to look forward to.

> *…also has to be something that's fun that you can look forward to. You can reward yourself.* [Participant in workshop 1]

Reminders

Participants considered reminders to be important for adherence to new routines and said they would like to have reminders about their use of the eHealth intervention, goals, and planned activities. It could be enough to get a reminder in the morning, but as the timing of when reminders are needed is individual, the participants wanted the user to be able to select when the eHealth intervention issued reminders. It was important for the reminders to be phrased positively and give joy, as it was expressed that positively phrased prompts lead to action and adherence more often than other prompts.

> *Support that's encouraging is very... We don't shut it off in the same way as support that's maybe not encouraging.*
>
> *That gives [us] a bad conscience instead.* [Participant in workshop 1]

As what gives joy is also individual, a reminder regarding a planned activity that brings joy could be a way to achieve individualization.

Interacting With Other Users

Participants expressed that they would like to have features that provide the opportunity to interact with others to reduce their sedentary time. They wanted the opportunity to share suggestions for activities and tips for reducing sedentary time and to see other users' suggestions. Moreover, they wanted users to be able to give each other positive feedback to increase their motivation.

> *...then you're a group, aren't you, who sort of ehh can cheer each other on and say good luck and well done and all that.* [Participant in workshop 2]

A chatroom was discussed as a possible way to facilitate interaction between users. There were concerns regarding the social climate in digital chatrooms. Several suggestions for how to improve it were discussed, and the participants preferred a solution in which users could only comment on each other's activity suggestions with predetermined answers or positive emojis. This interaction could also serve as an opportunity to initiate meeting up with other users with similar interests, but arranging meetings with unknown users also led to safety concerns.

> *Then there has to be some kind of block in it so you don't meet some strange person just to... I was going to say... but I'm just thinking that you can include the security aspect in it.* [Participant in workshop 1]

Timer

Participants wanted a timer with an alarm set to 30 minutes to avoid prolonged sedentary time, until the new behavior had become a habit. It was proposed that the timer could be started when the user sat down or could be set in advance to go off every 30 minutes during a selected period.

> *I might need ehh a pling, like: whoops. Now it's been 30 minutes.* [Participant in workshop 1]
>
> *...when you describe how you live. I don't live half as physically active... I live more "small". Thirty minutes, fine, and then I leave.* [Participant in workshop 2]

It was also suggested that, rather than simply sounding an alarm when the timer finished, the eHealth intervention could suggest joyful activities that were prespecified by the user.

Paper Prototype

At the third workshop, a paper prototype of an eHealth intervention including self-management for reducing sedentary time through joyful activities was presented. The prototype was planned as a website solution with important functions as well as information and links to websites and mobile apps. The prototype was designed to introduce the user to the features in the order chosen by the participants during the workshops. The user was first shown information regarding sedentary behavior, then encountered features for measuring sedentary time, and finally received support for goal setting, the selection of joyful activities, and the planning feature (Figure 2).

Brief paragraphs regarding sedentary behavior were presented, which could be shown to the user when they started using the eHealth intervention and could be found in the eHealth intervention at any time if the user wanted to read all the information.

Two tools for measuring sedentary time were presented. One was a self-assessment tool that asked the user to assess how many hours they had been performing different sedentary

XSL•FO
RenderX

behaviors during the day, such as eating, working, or commuting, and the other was a stopwatch for the user to start when they began being sedentary (Figure 3). Alongside the tools for measuring sedentary time, the user would receive adapted feedback based on different time intervals of sedentary time.

Figure 3. One of the tools presented at workshop 3 for measuring sedentary time.

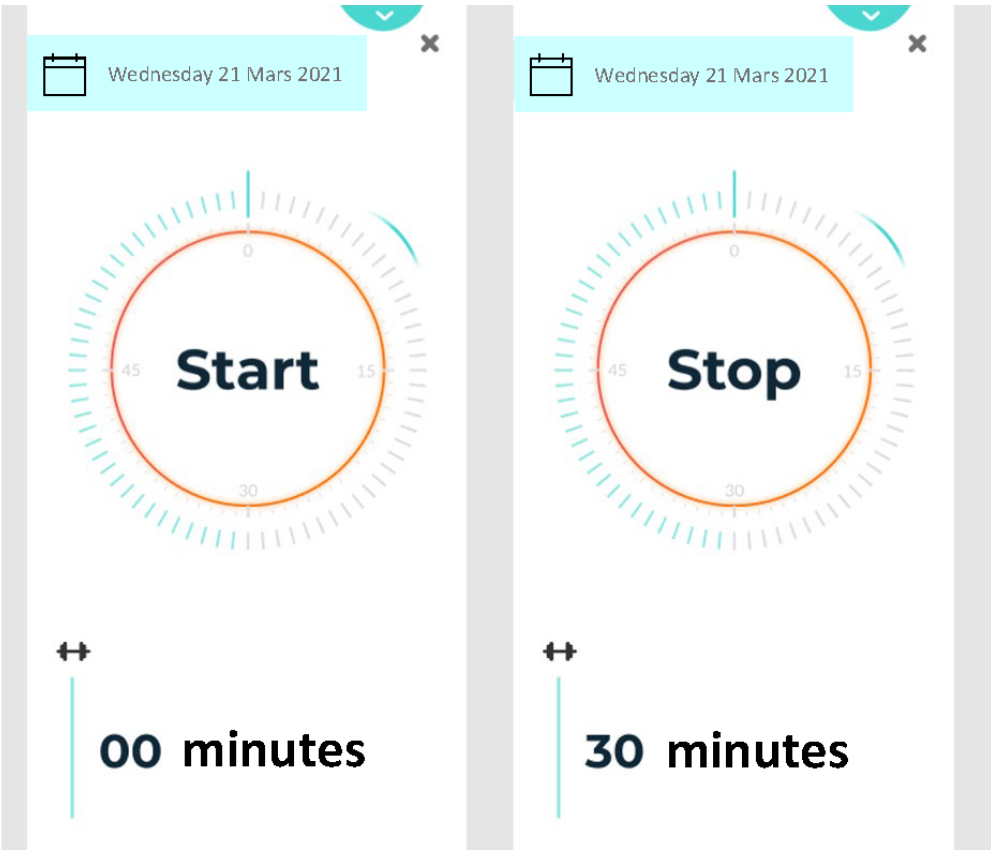

The prototype included the design of a support for goal setting. The user was given a written prompt asking them to formulate a goal regarding their sedentary time, and they received support for how this goal should be formulated, with examples of goals shown to them. They were then asked to focus on how they would feel if their sedentary time was reduced and what obstacles they might face, and to make a plan for how to overcome these obstacles. When the goal was then evaluated at follow-up, the user was asked to choose whether it was far from being achieved, nearly achieved, or fully achieved, and then received customized feedback.

The feature providing support for finding activities was presented to participants as a scroll list with activities shown along with a brief description and a photo (Figure 4). The scroll list could be sorted based on interests, local areas, and activities performed with others or by oneself. Tips on how to reduce sedentary time were offered as a bulleted list.

XSL•FO
RenderX

Figure 4. Feature for finding activities that arouse joy from the paper prototype presented at workshop 3.

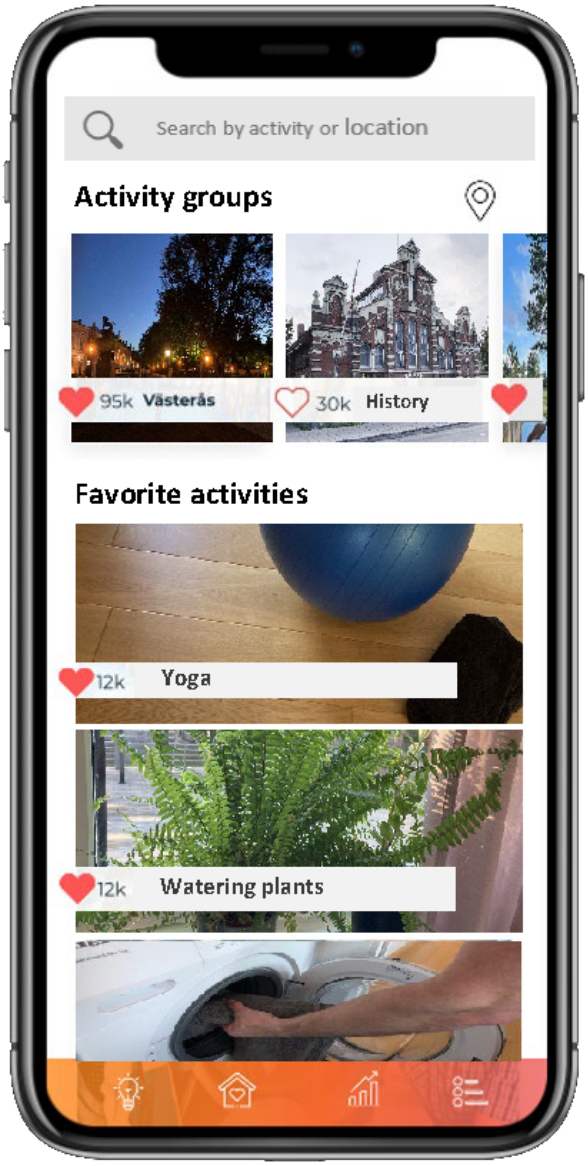

The schedule was presented both as a colorful printable schedule designed to evoke positive and cheerful emotions, and as a copy of the schedule from an iPhone and iPad. The user could write their planned activities digitally or manually on the printed schedule. The copy of the schedule from the iPhone and iPad was used to discuss whether the calendar found in a user's own phone could be used as a schedule.

In the prototype, rewards were delivered at goal evaluation as brief positively formulated phrases adapted to the user's goal achievement. As the prototype was designed as a website, the possibility to include reminders and interactions with other users was limited. It was suggested that the user make notes regarding planned activities and scheduled follow-ups, preferably in a calendar with the possibility to send notifications. No timer function was included in the prototype; instead, the suggestion to set an alarm when one was to be seated for a prolonged time was included among the tips for reducing sedentary time.

Change Analysis

Evoke Positive Affect

Participants stated that the eHealth intervention's first impression and introduction should arouse interest to facilitate further use.

> *If nobody opens this. Then nothing's worth anything. I mean, there has to be something that captures your interest.* [Participant in workshop 3]

They suggested that this could be done by having celebrities or cartoons convey the information or by having the user answer questions regarding sedentary time. While the participants considered it important to arouse interest in the eHealth intervention's introduction, when it came to the delivery of information, they expressed that credibility was important.

> *Isn't it two different things we're talking about... One's about capturing attention. Come look at our website. [mm] There you could have animated figures, I'm thinking... But when I open the website, then I'd*

XSL•FO
RenderX

> *really like to have, to know that there's evidence behind this in some way, I mean. Ehh am I thinking about it the wrong way?* [Participant in workshop 3]

They suggested that, to achieve credibility, the information could be presented by a physician or physiotherapist, or by people who were newly retired and could share their experiences. It was stated that all information in the eHealth intervention should be more positively phrased and that when information was evidence-based, this should be clearly stated.

Less is More

Overall, the participants wanted the text and information in the paper prototype to be shortened and simplified. The group agreed to choose the tool presented for the self-assessment of sedentary time, as it was considered easier to use than a stopwatch. However, there were concerns regarding whether people who spend a prolonged time in sedentary behavior would take the time to fill in such a form and the fact that it is difficult to make such an assessment. As a solution, it was suggested that the user could fill in an estimate of their total time in sedentary behavior or enter data from an activity watch if they used one. The participants pointed out the need to adapt the assessment form to better suit those who are retired, suggesting that the activities mentioned on the form be replaced with time intervals as every day is different for a person who is newly retired compared to when they had a job. They also suggested that the feedback be changed to give the user a clear overview of when prolonged sedentary behavior becomes unhealthy and whether there are different levels when it comes to harmful amounts of sedentary time, and if so, how close they are to moving on to the next level.

The content of the support for goal setting was considered helpful, but the layout was perceived as cumbersome and too extensive. It was suggested that the text be shortened and reformulated to use more accessible language.

> *I hadn't finished reading... There's too much text.* [Participant in workshop 3]

The participants pointed out the importance of letting the user know that they do not need to make massive changes to reduce their sedentary time and that small changes and activities in the home could suffice.

> *It's also so important that there's information about that the small things in daily life is enough.* [Participant in workshop 3]

They suggested that this could be done by sorting activities as "activities in the home" or "start small." The group was positive about simplifying the eHealth intervention by including links to organizations rather than giving information on each activity offered by the organizations.

The participants considered the rewards offered in the prototype to be sufficient. As the prototype was designed to encourage the user to engage in activities that bring positive affect and emotions, and to remind them of their feelings at goal fulfillment, the participants did not feel further rewards were needed.

The participants mentioned wanting the opportunity to interact with other users within the eHealth intervention. However, at this point, they agreed that the possibility to find activities with others in terms of links and information regarding retirement organizations and local events was sufficient for interacting with others.

> *Like we've talked about before, it was more this thing of being able to interact with other users. I mean, it's...* [not possible now]
>
> *We've dropped that a bit.*
>
> *But then you can find the organizations, you find them via this.* [Participant in workshop 3]

To keep the features included in the eHealth intervention to a minimum, a timer was not included; instead, among the tips provided, the user would be given a recommendation to use a timer in other mobile apps or use the timer feature on their phone. The participants discussed the notion that a timer might be useful for those who prefer not to enroll in planned activities with others and would rather perform activities in the home.

At the third workshop, participants did not express a need for a calendar or schedule in the eHealth intervention, as they preferred to write down their planned activities in the calendar on their phones rather than in the calendar suggested at the workshop. They felt that those who do not use the calendar on their phones would probably not use a digital eHealth intervention for planning but likely already use a paper calendar. Therefore, it was suggested that the eHealth intervention should simply suggest that the user write down their planning so that they could choose the type of planning tool they prefer. This was expressed as follows:

> *I mean, maybe it's a bit much for those who don't have it in their phone. I was thinking of this... on the website.*
>
> *You don't think they have a calendar on the fridge?*
>
> *I don't think they go to their computer and [fill it in].* [Participant in workshop 3]

Packaging

The participants agreed that the eHealth intervention should be delivered as a website rather than a mobile app, as it was important that the intervention be accessible from a computer so it would be easier to operate. However, choosing a website rather than a mobile app limited the possibility to include certain features. For instance, the participants would have liked the eHealth intervention to send reminders to promote adherence, and they suggested that the calendar function on the user's phone could be used to remind them of planned activities (although they still expressed a desire for the possibility to send a daily reminder or a reminder if the eHealth intervention was not being used as planned). They also wanted to be able to log in to the eHealth intervention in order to save their goals and interests digitally to facilitate the use of the eHealth intervention and for goal follow-up.

Participants expressed a wish for information to be delivered in different ways, allowing the user to listen to or read information, or sometimes get it as a video. It was suggested

XSL•FO
RenderX

that, during goal setting, the user should be able to listen to the information, as participants experienced that the spoken word was more appealing and evoked more emotions in this context.

> *Yeah, I thought about it... that you said it in a much more appealing way than the text does. When you said "Think about what you want to achieve. Imagine how...", then you get a feeling. I don't get that here [shows the paper].* [Participant in workshop 3]

Discussion

Principal Findings

To our knowledge, this is the first study to develop a self-management eHealth intervention for reducing sedentary time and promoting adherence to reduced sedentary time, which has been created for and with older adults in the transition to retirement. The study used PD [26] to include end users, researchers, and a web designer as mutual partners in the development process. The study has contributed new knowledge regarding features of this type of eHealth intervention and the timing of these features. The most important features were those involving finding activities that arouse joy, having support in goal setting, and getting information regarding sedentary behavior. The workshop participants regarded the order in which the included features are presented to the user as important: The user first receives knowledge regarding sedentary behavior, then encounters features for measuring sedentary time and for support in setting a goal, and lastly gets support in finding joyful activities to perform in order to avoid being sedentary. The results highlight the need for such an eHealth intervention to be concise, accessible, and enjoyable to suit its end users.

Comparison With Prior Work

This study differs from previous research [7,8] as it specifically targets older adults in the transition to retirement and involves them as co-designers in the development of an eHealth intervention including self-management for reducing sedentary time and promoting adherence to reduced sedentary time.

Participants perceived it as important for the eHealth intervention including self-management to use positive phrasing to increase both its use and users' adherence to reduced sedentary time, expressing that positively phrased prompts lead to adherence more than negatively phrased ones. Previous researchers have found positive affect to be connected to the use of eHealth interventions to increase adherence in older adults [25] and positively phrased prompts to be more effective in older adults [35]. Overall, this indicates the importance of an eHealth intervention that evokes positive affect to promote adherence to reduced sedentary time when transitioning to retirement.

The features considered to be the most important for reducing sedentary time in the transition to retirement were those involving finding joyful activities. The importance of joyfulness in reducing sedentary time and increasing physical activity in older adults has been found in previous research [15,36]. In this study, participants stated the importance of receiving suggestions for what to replace sedentary time with, rather than just getting a prompt to break their sedentary behavior. The results further highlight the need for awareness of the diversity in the activities that people consider joyful and for the inclusion of information on simple joyful activities that can be performed in one's own home to reduce sedentary time.

The eHealth intervention developed in this study includes features for information, measurement, goal setting, and planning, which have been used previously to reduce sedentary time in older adults [7,8] and have been found to be acceptable behavior change techniques for older adults to reduce their sedentary time [9]. The results of this study show that these behavior change techniques are also desired specifically when it comes to an eHealth intervention including self-management for reducing sedentary time, which targets older adults in the transition to retirement.

During the third workshop, self-assessment was chosen as the method for measuring one's sedentary time. As objectively measured sedentary time is more accurate than self-assessed sedentary time (total sedentary time is often underestimated [37]), the inclusion of objectively measured sedentary time might provide the user with more reliable results. However, doing so would mean adding a device, such as a smartwatch, and this would entail a cost to the user or require them to always carry their phone, which the participants considered inconvenient. The included feature for self-assessment also provides the user with more contextual information, which is an advantage of using self-assessed sedentary time [38].

Among the included features are behavior change techniques that target both affective determinants (eg, affective attitudes) [39] and cognitive determinants (eg, cognitive attitudes and self-efficacy) [40] for reduced sedentary time. The feature for goal setting and planning in the developed eHealth intervention is based on mental contrasting and implementation intention (MCII) [41], which has previously been found to be effective in health behavior change [42]. The inclusion of MCII also responds to the participants' wish to include goal setting and planning as well as an opportunity to consider activities to replace their sedentary time, which was discussed during the workshops. In the developed eHealth intervention, the user is asked to consider a positive affective outcome, which has been used with success in previous research targeting health behaviors [20]. The inclusion of a feature targeting affective determinants when reducing sedentary time in older adults transitioning to retirement can be an effective approach, as previous research has found that sedentary behavior is largely an affect-driven behavior [18]. Further, Lithopoulos et al [21] found that an intervention targeting affective determinants for reducing sedentary time in the office was effective in the short term.

Previous research has described sedentary behavior as a habitual behavior performed in unawareness [9]. This speaks for the inclusion of features, such as a timer, to increase users' awareness of when they are in prolonged sedentary behavior and thereby promote their adherence to reduced sedentary time. However, Suorsa et al [43] found that an activity tracker was not sufficient for reducing sedentary time in retirees. In this study, none of the participants rated a timer feature as important, and it was not included in either the paper or the digital prototype. However, among the tips the eHealth intervention

XSL•FO
RenderX

offered for reducing one's sedentary time was the suggestion to set a timer when engaged in sedentary behaviors.

The approach of the eHealth intervention for achieving behavior change, which was developed here, can be understood based on the integrated behavior change (IBC) model for health behavior [44]. The eHealth intervention is designed to guide the user in reducing their sedentary time by finding activities that they are autonomously motivated to perform, meaning that the intention to perform a health behavior is dependent on whether the behavior evokes positive affect and is in line with a person's interests and true self [45,46]. Further, the developed eHealth intervention includes information, self-monitoring, and feedback on behavior that can influence users' attitudes and perceived behavioral control regarding the behavior change [40,47]. Lastly, the feature for goal setting includes affective MCII, which can influence behavior through implicit pathways [48,49], as well as elements of action planning, which are included in the IBC model. The use of the IBC model can offer an understanding of how the developed eHealth intervention including self-management is intended to reduce sedentary time and promote adherence to reduced sedentary time, but further studies are needed in order to evaluate the eHealth intervention's feasibility and effectiveness.

Limitations and Strengths

Carefully considered choices were made to increase the study's trustworthiness (credibility, dependability, and transferability) [50]. The study's methodological approach is a strength, as it enabled the inclusion of older adults in the transition to retirement as co-designers. They are the only ones who hold the tacit knowledge that is needed to develop an eHealth intervention in line with their needs and preferences. The participants also included a web designer and 2 researchers, who contributed knowledge regarding web development and health behavior change. This combination of participants may contribute to the development of an effective and acceptable eHealth intervention as both tacit and explicit knowledge are included in the development process, which strengthens the study's credibility.

The credibility was further strengthened by a workshop leader guiding the workshops to ensure that everyone had the opportunity to speak and to avoid placing additional power with the researchers. Moreover, the workshops started with an icebreaker, and participants were offered coffee and fruits during the meetings to create a pleasant atmosphere. Three PD tool cards, 2D mapping, and a user journey tool were used to encourage active participation among the participants. These tools were chosen because they have been used successfully in PD research [33], and they ensured that all participants were encouraged and had the possibility to be active during the workshops. The recruitment of participants at theme days and retirement organization meetings might have led to the inclusion of participants with a greater interest in such activities than the population, possibly affecting the study's credibility and transferability. This was compensated by the recruitment of participants through a local newspaper. Additionally, the participants expressed their differences through participation and interest in activities held by organizations during the workshops. The older adults in the transition to retirement who participated in this study lived in different types of housing and reported different levels of daily sedentary time as well as levels of sedentary time at their latest jobs, which indicates a spread among them and thus increases credibility. The study's confirmability was strengthened as, during the third workshop, the participants were given the opportunity to discuss and further develop the paper prototype developed from workshops 1 and 2. Further, the feature analysis was used as a verification of the paper prototype to ensure that no important data had been excluded.

Conclusion

An eHealth intervention including self-management for reducing sedentary time and promoting adherence to reduced sedentary time in older adults transitioning to retirement was developed in a co-design approach involving representatives of end users, a web designer, and researchers. Participants desired features for providing information regarding sedentary behavior, measuring sedentary time, goal setting, finding joyful activities, and making plans and a schedule, and expressed that these should be presented to the user in this order. The most important features for reducing sedentary time were said to be those involving finding joyful activities. Further, participants expressed a desire for such an eHealth intervention to be concise, accessible, and enjoyable in order to be suitable for older adults transitioning to retirement. The developed eHealth intervention strives to support behavior change based on the IBC model by focusing on autonomously motivated activities and affective and cognitive determinants through implicit and explicit pathways, and includes action planning. It will be further tested for feasibility and effectiveness in upcoming studies.

Acknowledgments

This study was funded by the Swedish Research Council Dnr (2018-02928). We acknowledge Liv Nordström for being the workshop leader during the 3 workshops and Bluearth Fondation for sharing the sedentary time calculator for use during workshop 3.

Authors' Contributions

All authors (LH, CE, PHW, AS, ML, and MLE) contributed to the conception and design of the study. LH and CE participated during the workshops. All authors (LH, CE, PHW, AS, ML, and MLE) participated during the between-workshop meeting. LH analyzed the data and wrote the main content in the manuscript. CE, PHW, AS, ML, and MLE critically revised the manuscript and contributed intellectual content. All authors (LH, CE, PHW, AS, ML, and MLE) read and approved the final manuscript.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Theme summary.
[DOCX File , 18 KB - formative_v9i1e63567_app1.docx]

References

1. Harvey JA, Chastin SF, Skelton DA. How sedentary are older people? A systematic review of the amount of sedentary behavior. J Aging Phys Act 2015 Jul;23(3):471-487. [doi: 10.1123/japa.2014-0164] [Medline: 25387160]
2. Biswas A, Oh PI, Faulkner GE, Bajaj RR, Silver MA, Mitchell MS, et al. Sedentary time and its association with risk for disease incidence, mortality, and hospitalization in adults: a systematic review and meta-analysis. Ann Intern Med 2015 Jan 20;162(2):123-132. [doi: 10.7326/M14-1651] [Medline: 25599350]
3. Rossen J, Von Rosen P, Johansson UB, Brismar K, Hagströmer M. Associations of physical activity and sedentary behavior with cardiometabolic biomarkers in prediabetes and type 2 diabetes: a compositional data analysis. Phys Sportsmed 2020 May;48(2):222-228. [doi: 10.1080/00913847.2019.1684811] [Medline: 31663410]
4. Rojer AGM, Ramsey KA, Trappenburg MC, van Rijssen NM, Otten RHJ, Heymans MW, et al. Instrumented measures of sedentary behaviour and physical activity are associated with mortality in community-dwelling older adults: A systematic review, meta-analysis and meta-regression analysis. Ageing Res Rev 2020 Aug;61:101061 [FREE Full text] [doi: 10.1016/j.arr.2020.101061] [Medline: 32360669]
5. Tremblay MS, Aubert S, Barnes JD, Saunders TJ, Carson V, Latimer-Cheung AE, SBRN Terminology Consensus Project Participants. Sedentary Behavior Research Network (SBRN) - Terminology Consensus Project process and outcome. Int J Behav Nutr Phys Act 2017 Jun 10;14(1):75 [FREE Full text] [doi: 10.1186/s12966-017-0525-8] [Medline: 28599680]
6. Beard JR, Officer AM, Cassels AK. The world report on ageing and health. Gerontologist 2016 Apr;56 Suppl 2:S163-S166. [doi: 10.1093/geront/gnw037] [Medline: 26994257]
7. Chastin S, Gardiner PA, Harvey JA, Leask CF, Jerez-Roig J, Rosenberg D, et al. Interventions for reducing sedentary behaviour in community-dwelling older adults. Cochrane Database Syst Rev 2021 Jun 25;6(6):CD012784 [FREE Full text] [doi: 10.1002/14651858.CD012784.pub2] [Medline: 34169503]
8. Petrusevski C, Choo S, Wilson M, MacDermid J, Richardson J. Interventions to address sedentary behaviour for older adults: a scoping review. Disabil Rehabil 2021 Oct;43(21):3090-3101. [doi: 10.1080/09638288.2020.1725156] [Medline: 32058808]
9. Compernolle S, De Cocker K, Cardon G, De Bourdeaudhuij I, Van Dyck D. Older adults' perceptions of sedentary behavior: a systematic review and thematic synthesis of qualitative studies. Gerontologist 2020 Nov 23;60(8):572-582. [doi: 10.1093/geront/gnz127] [Medline: 31670766]
10. Ding D, Grunseit AC, Chau JY, Vo K, Byles J, Bauman AE. Retirement-a transition to a healthier lifestyle?: Evidence from a large Australian study. Am J Prev Med 2016 Aug;51(2):170-178. [doi: 10.1016/j.amepre.2016.01.019] [Medline: 26972491]
11. Vigezzi GP, Gaetti G, Gianfredi V, Frascella B, Gentile L, d'Errico A, Italian Working Group on RetirementHealth. Transition to retirement impact on health and lifestyle habits: analysis from a nationwide Italian cohort. BMC Public Health 2021 Sep 14;21(1):1670 [FREE Full text] [doi: 10.1186/s12889-021-11670-3] [Medline: 34521363]
12. Robinson SM, Westbury LD, Ward K, Syddall H, Cooper R, Cooper C, et al. Is lifestyle change around retirement associated with better physical performance in older age?: insights from a longitudinal cohort. Eur J Ageing 2021 Dec;18(4):513-521 [FREE Full text] [doi: 10.1007/s10433-021-00607-9] [Medline: 34786013]
13. Lorig KR, Holman H. Self-management education: history, definition, outcomes, and mechanisms. Ann Behav Med 2003 Aug;26(1):1-7. [doi: 10.1207/S15324796ABM2601_01] [Medline: 12867348]
14. Ramalho A, Petrica J. The quiet epidemic: An overview of emerging qualitative research trends on sedentary behavior in aging populations. Healthcare (Basel) 2023 Aug 06;11(15):2215 [FREE Full text] [doi: 10.3390/healthcare11152215] [Medline: 37570455]
15. McGowan LJ, Powell R, French DP. How acceptable is reducing sedentary behavior to older adults? perceptions and experiences across diverse socioeconomic areas. J Aging Phys Act 2019 Sep 01;27(5):642–652 [FREE Full text] [doi: 10.1123/japa.2018-0274] [Medline: 30676195]
16. Eklund C, Elfström ML, von Heideken Wågert P, Söderlund A, Gustavsson C, Cederbom S, et al. The meaning of sedentary behavior as experienced by people in the transition from working life to retirement: an empirical phenomenological study. Phys Ther 2021 Aug 01;101(8):pzab117 [FREE Full text] [doi: 10.1093/ptj/pzab117] [Medline: 33951141]
17. Hultman L, Eklund C, von Heideken Wågert P, Söderlund A, Elfström ML. It has to be my way-Reducing sedentary time in the transition to retirement. J Aging Phys Act 2024 Nov 20:1-10. [doi: 10.1123/japa.2023-0414] [Medline: 39566479]

XSL•FO
RenderX

18. Phipps DJ, Rhodes RE, Jenkins K, Hannan TE, Browning NG, Hamilton K. A dual process model of affective and instrumental implicit attitude, self-monitoring, and sedentary behavior. Psychology of Sport and Exercise 2022 Sep;62:102222. [doi: 10.1016/j.psychsport.2022.102222]
19. Forster AS, Buykx P, Martin N, Sadler S, Southgate B, Rockliffe L, et al. Using affective judgement to increase physical activity in British adults. Health Promot Int 2018 Aug 01;33(4):648-656 [FREE Full text] [doi: 10.1093/heapro/dax004] [Medline: 28334769]
20. Ruissen GR, Rhodes RE, Crocker PRE, Beauchamp MR. Affective mental contrasting to enhance physical activity: A randomized controlled trial. Health Psychol 2018 Jan;37(1):51-60. [doi: 10.1037/hea0000551] [Medline: 28981303]
21. Lithopoulos A, Kaushal N, Beauchamp MR, Conner M, de Bruijn GJ, Latimer-Cheung AE, et al. Changing sedentary behavior in the office: a randomised controlled trial comparing the effect of affective, instrumental, and self-regulatory messaging on sitting. Appl Psychol Health Well Being 2020 Nov;12(3):687-702 [FREE Full text] [doi: 10.1111/aphw.12202] [Medline: 32428341]
22. Eysenbach G. What is e-health? J Med Internet Res 2001;3(2):E20 [FREE Full text] [doi: 10.2196/jmir.3.2.e20] [Medline: 11720962]
23. Internet use. International Telecommunication Union. URL: https://www.itu.int/itu-d/reports/statistics/2023/10/10/ff23-internet-use/#footnote1 [accessed 2024-02-12]
24. Yerrakalva D, Yerrakalva D, Hajna S, Griffin S. Effects of mobile health app interventions on sedentary time, physical activity, and fitness in older adults: systematic review and meta-analysis. J Med Internet Res 2019 Nov 28;21(11):e14343 [FREE Full text] [doi: 10.2196/14343] [Medline: 31778121]
25. Valenzuela T, Okubo Y, Woodbury A, Lord SR, Delbaere K. Adherence to technology-based exercise programs in older adults: a systematic review. J Geriatr Phys Ther 2018;41(1):49-61. [doi: 10.1519/JPT.0000000000000095] [Medline: 27362526]
26. Clemensen J, Rothmann MJ, Smith AC, Caffery LJ, Danbjorg DB. Participatory design methods in telemedicine research. J Telemed Telecare 2017 Oct;23(9):780-785. [doi: 10.1177/1357633X16686747] [Medline: 28027678]
27. Wannheden C, Revenäs Å. How people with Parkinson's disease and health care professionals wish to partner in care using eHealth: co-design study. J Med Internet Res 2020 Sep 21;22(9):e19195 [FREE Full text] [doi: 10.2196/19195] [Medline: 32955448]
28. Moore G, Wilding H, Gray K, Castle D. Participatory methods to engage health service users in the development of electronic health resources: systematic review. J Particip Med 2019 Feb 22;11(1):e11474 [FREE Full text] [doi: 10.2196/11474] [Medline: 33055069]
29. Revenäs Å, Ström L, Cicchetti A, Ehn M. Toward digital inclusion of older adults in e-health: a case study on support for physical activity. Univ Access Inf Soc 2023:1049. [doi: 10.1007/s10209-023-01049-z]
30. Revenäs Å, Opava CH, Martin C, Demmelmaier I, Keller C, Åsenlöf P. Development of a web-based and mobile app to support physical activity in individuals with rheumatoid arthritis: results from the second step of a co-design process. JMIR Res Protoc 2015 Feb 09;4(1):e22 [FREE Full text] [doi: 10.2196/resprot.3795] [Medline: 25665589]
31. Bartlett YK, Selby DL, Newsham A, Keding A, Forman D, Brown J, et al. Developing a useful, user-friendly website for cancer patient follow-up: users' perspectives on ease of access and usefulness. Eur J Cancer Care (Engl) 2012 Nov;21(6):747-757. [doi: 10.1111/j.1365-2354.2012.01357.x] [Medline: 22533456]
32. Østervang C, Lassen A, Schmidt T, Coyne E, Dieperink K, Jensen C. Development of a health information system to promote emergency care pathways: A participatory design study. Digit Health 2022;8:20552076221145856 [FREE Full text] [doi: 10.1177/20552076221145856] [Medline: 36601282]
33. Vandekerckhove P, de Mul M, Bramer WM, de Bont AA. Generative participatory design methodology to develop electronic health interventions: systematic literature review. J Med Internet Res 2020 Apr 27;22(4):e13780 [FREE Full text] [doi: 10.2196/13780] [Medline: 32338617]
34. Braun V, Clarke V. Thematic Analysis: A Practical Guide. Thousand Oaks, CA: SAGE Publications Ltd; 2021.
35. Notthoff N, Carstensen LL. Positive messaging promotes walking in older adults. Psychol Aging 2014 Jun;29(2):329-341 [FREE Full text] [doi: 10.1037/a0036748] [Medline: 24956001]
36. Ioannou E, Chen HL, Bromley V, Fosker S, Ali K, Fernando A, et al. The key values and factors identified by older adults to promote physical activity and reduce sedentary behaviour using co-production approaches: a scoping review. BMC Geriatr 2023 Jun 16;23(1):371 [FREE Full text] [doi: 10.1186/s12877-023-04005-x] [Medline: 37328734]
37. Prince SA, Cardilli L, Reed JL, Saunders TJ, Kite C, Douillette K, et al. A comparison of self-reported and device measured sedentary behaviour in adults: a systematic review and meta-analysis. Int J Behav Nutr Phys Act 2020 Mar 04;17(1):31 [FREE Full text] [doi: 10.1186/s12966-020-00938-3] [Medline: 32131845]
38. Aunger J, Wagnild J. Objective and subjective measurement of sedentary behavior in human adults: A toolkit. Am J Hum Biol 2022 Jan;34(1):e23546 [FREE Full text] [doi: 10.1002/ajhb.23546] [Medline: 33277954]
39. Williams DM, Evans DR. Current emotion research in health behavior science. Emotion Review 2014 Jun 17;6(3):277-287. [doi: 10.1177/1754073914523052]
40. Ajzen I. The theory of planned behavior. Organizational Behavior and Human Decision Processes 1991 Dec;50(2):179-211. [doi: 10.1016/0749-5978(91)90020-T]

XSL•FO
RenderX

41. Oettingen G. Future thought and behaviour change. European Review of Social Psychology 2012 Mar;23(1):1-63. [doi: 10.1080/10463283.2011.643698]
42. Cross A, Sheffield D. Mental contrasting for health behaviour change: a systematic review and meta-analysis of effects and moderator variables. Health Psychol Rev 2019 Jun;13(2):209-225. [doi: 10.1080/17437199.2019.1594332] [Medline: 30879403]
43. Suorsa K, Leskinen T, Pulakka A, Pentti J, Löyttyniemi E, Heinonen I, et al. The effect of a consumer-based activity tracker intervention on accelerometer-measured sedentary time among retirees: a randomized controlled REACT trial. J Gerontol A Biol Sci Med Sci 2022 Mar 03;77(3):579-587 [FREE Full text] [doi: 10.1093/gerona/glab107] [Medline: 33839766]
44. Hagger MS, Chatzisarantis NL. An integrated behavior change model for physical activity. Exerc Sport Sci Rev 2014 Apr;42(2):62-69. [doi: 10.1249/JES.0000000000000008] [Medline: 24508739]
45. Deci EL, Ryan RM. Intrinsic Motivation and Self-Determination in Human Behavior. New York, NY: Springer; 1985.
46. Hagger MS, Protogerou C. Affect in the Context of Self-Determination Theory. In: Williams DM, Rhodes RE, Conner MT, editors. Affective Determinants of Health Behavior. Oxford, England: Oxford University Press; 2018.
47. Ajzen I. Perceived behavioral control, self‐efficacy, locus of control, and the theory of planned behavior. J Applied Social Pyschol 2006 Jul 31;32(4):665-683. [doi: 10.1111/j.1559-1816.2002.tb00236.x]
48. Klusmann V, Gow AJ, Robert P, Oettingen G. Using theories of behavior change to develop interventions for healthy aging. J Gerontol B Psychol Sci Soc Sci 2021 Sep 13;76(Suppl 2):S191-S205. [doi: 10.1093/geronb/gbab111] [Medline: 34515775]
49. Wittleder S, Kappes A, Krott N, Jay M, Oettingen G. Mental contrasting spurs energy by changing implicit evaluations of obstacles. Motivation Science 2020 Jun;6(2):133-155 [FREE Full text] [doi: 10.1037/mot0000140]
50. Lincoln YS, Guba EG, Pilotta JJ. Naturalistic inquiry. International Journal of Intercultural Relations 1985 Jan;9(4):438-439. [doi: 10.1016/0147-1767(85)90062-8]

Abbreviations

IBC: integrated behavior change
MCII: mental contrasting and implementation intention
PD: participatory design

Edited by A Mavragani; submitted 24.06.24; peer-reviewed by W Tumuhimbise, L McGowan; comments to author 26.11.24; revised version received 02.12.24; accepted 03.12.24; published 20.01.25.

Please cite as:
Hultman L, Eklund C, von Heideken Wågert P, Söderlund A, Lindén M, Elfström ML
Development of an eHealth Intervention Including Self-Management for Reducing Sedentary Time in the Transition to Retirement: Participatory Design Study
JMIR Form Res 2025;9:e63567
URL: https://formative.jmir.org/2025/1/e63567
doi:10.2196/63567
PMID:

©Lisa Hultman, Caroline Eklund, Petra von Heideken Wågert, Anne Söderlund, Maria Lindén, Magnus L Elfström. Originally published in JMIR Formative Research (https://formative.jmir.org), 20.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Methods to Adjust for Confounding in Test-Negative Design COVID-19 Effectiveness Studies: Simulation Study

Elizabeth AK Rowley[1], DrPH; Patrick K Mitchell[1], ScD; Duck-Hye Yang[1], PhD; Ned Lewis[2], MPH; Brian E Dixon[3,4], MPA, PhD; Gabriela Vazquez-Benitez[5], PhD; William F Fadel[3,4], PhD; Inih J Essien[5], MPH, OD; Allison L Naleway[6], PhD; Edward Stenehjem[7], MSc, MD; Toan C Ong[8], PhD; Manjusha Gaglani[9,10], MBBS; Karthik Natarajan[11,12], PhD; Peter Embi[3,13], MS, MD; Ryan E Wiegand[14], PhD; Ruth Link-Gelles[14], MPH, PhD; Mark W Tenforde[14], MD, PhD; Bruce Fireman[2], MA

[1]Westat, Rockville, MD, United States
[2]Vaccine Study Center, Northern California Division of Research, Kaiser Permanente, Oakland, CA, United States
[3]Center for Biomedical Informatics, Regenstrief Institute, Indianapolis, IN, United States
[4]Fairbanks School of Public Health, Indiana University, Indianapolis, IN, United States
[5]HealthPartners Institute, Minneapolis, MN, United States
[6]Center for Health Research, Kaiser Permanente, Portland, OR, United States
[7]Division of Infectious Diseases and Clinical Epidemiology, Intermountain Healthcare, Salt Lake City, UT, United States
[8]Department of Biomedical Informatics, University of Colorado Anschutz Medical Campus, Aurora, CO, United States
[9]Department of Pediatrics, Section of Pediatric Infectious Diseases, Baylor Scott & White Health, Temple, TX, United States
[10]Department of Medical Education, Texas A&M University College of Medicine, Killeen, TX, United States
[11]Department of Biomedical Informatics, Columbia University Irving Medical Center, New York, NY, United States
[12]New York Presbyterian Hospital, New York, NY, United States
[13]Vanderbilt University Medical Center, Nashville, TN, United States
[14]National Center for Immunization and Respiratory Diseases, Centers for Disease Control and Prevention, Atlanta, GA, United States

Corresponding Author:
Elizabeth AK Rowley, DrPH
Westat
1600 Research Blvd
Rockville, MD, 20850
United States
Phone: 1 301 251 1500
Email: ELIZABETHROWLEY@WESTAT.COM

Abstract

Background: Real-world COVID-19 vaccine effectiveness (VE) studies are investigating exposures of increasing complexity accounting for time since vaccination. These studies require methods that adjust for the confounding that arises when morbidities and demographics are associated with vaccination and the risk of outcome events. Methods based on propensity scores (PS) are well-suited to this when the exposure is dichotomous, but present challenges when the exposure is multinomial.

Objective: This simulation study aimed to investigate alternative methods to adjust for confounding in VE studies that have a test-negative design.

Methods: Adjustment for a disease risk score (DRS) is compared with multivariable logistic regression. Both stratification on the DRS and direct covariate adjustment of the DRS are examined. Multivariable logistic regression with all the covariates and with a limited subset of key covariates is considered. The performance of VE estimators is evaluated across a multinomial vaccination exposure in simulated datasets.

Results: Bias in VE estimates from multivariable models ranged from –5.3% to 6.1% across 4 levels of vaccination. Standard errors of VE estimates were unbiased, and 95% coverage probabilities were attained in most scenarios. The lowest coverage in the multivariable scenarios was 93.7% (95% CI 92.2%-95.2%) and occurred in the multivariable model with key covariates, while the highest coverage in the multivariable scenarios was 95.3% (95% CI 94.0%-96.6%) and occurred in the multivariable model with all covariates. Bias in VE estimates from DRS-adjusted models was low, ranging from –2.2% to 4.2%. However, the

XSL•FO
RenderX

DRS-adjusted models underestimated the standard errors of VE estimates, with coverage sometimes below the 95% level. The lowest coverage in the DRS scenarios was 87.8% (95% CI 85.8%-89.8%) and occurred in the direct adjustment for the DRS model. The highest coverage in the DRS scenarios was 94.8% (95% CI 93.4%-96.2%) and occurred in the model that stratified on DRS. Although variation in the performance of VE estimates occurred across modeling strategies, variation in performance was also present across exposure groups.

Conclusions: Overall, models using a DRS to adjust for confounding performed adequately but not as well as the multivariable models that adjusted for covariates individually.

(JMIR Form Res 2025;9:e58981) doi:10.2196/58981

KEYWORDS

disease risk score; propensity score; vaccine effectiveness; COVID-19; simulation study; usefulness; comorbidity; assessment

Introduction

Background

Real-world, observational vaccine effectiveness (VE) studies are critical for evaluating the post licensure performance of vaccines. These studies must address confounding from patients' demographic characteristics and underlying medical conditions because such factors may be associated with both disease outcomes and vaccination status. When vaccination status was dichotomous, inverse probability weighting based on propensity scores (PS) could be used to adjust for confounding. However, the usefulness of the PS is problematic when the exposure is multinomial rather than dichotomous—for example, if we are interested in comparing multiple levels of vaccination status distinguished by the number of doses received and by intervals of time after the last dose [1]. Multinomial PS approaches are feasible in some situations [2], however, multinomial PS poses computational challenges and is less intuitive.

An alternative to the PS in this context is the disease risk score (DRS), also called a confounder score, prognostic score, comorbidity score, or simply a risk score [1-9]. A DRS can combine covariates into a single score that reflects their associations with the outcome. However, if it is feasible to make a DRS that adjusts appropriately for the relevant covariates, it can be similarly feasible and appropriate to simply adjust for the covariates individually without first combining them into a DRS [1]. This simulation study compared logistic regression models that use a DRS to adjust VE estimates versus logistic regression models that adjust for covariates individually. We compared DRS adjustment with individual covariate adjustment in scenarios comprised of simulated data similar to real-world data used by the Virtual SARS-CoV-2, Influenza, and Other Respiratory Viruses Network (VISION) to report on COVID-19 VE.

VISION: Virtual SARS-CoV-2, Influenza, and Other Respiratory Viruses Network Studies of COVID-19 VE

The VISION network was established by the Centers for Disease Control and Prevention (CDC) in collaboration with 10 US health care systems with medical, laboratory, and vaccination records. VISION uses a case-control test-negative design (TND) to assess the effectiveness of COVID-19 vaccines in preventing laboratory-confirmed COVID-19–associated hospitalizations and visits to emergency departments or urgent care clinics [10]. Patients who received care in one of these settings for a COVID-19-like illness (CLI) are included in the study if they were tested for SARS-CoV-2 by molecular assay proximate to the encounter. Those who tested positive were considered cases; those who tested negative were considered controls. CLI diagnoses include acute respiratory diagnoses or related signs or symptoms, captured by diagnosis codes [10-13]. The case-control TND has been commonly used in studies of COVID-19 VE and VE studies against influenza, rotavirus, and other diseases [14-16]. This study was reviewed and approved by the institutional review board at Westat, Inc. This electronic health record-based study does not include factors necessitating patient consent. The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the CDC.

When VISION first began analyzing COVID-19 VE in early 2021, the comparison of interest was between fully vaccinated individuals and unvaccinated individuals. PS-based methods have been widely used in cohort studies with this kind of binary exposure [17-19]. VISION's case-control studies also used a PS, derived from the test-negative controls, to adjust VE estimates [10,20].

As the pandemic progressed, recommended vaccination schedules became more complex [21-23]. Research studies required consideration of more nuanced exposure categories, based on a combination of the number of vaccine doses received and time since the most recent vaccine dose (to examine the waning of vaccine-induced protection against COVID-19) [24-26]. VISION continued to rely on the PS for covariate adjustment by conducting a series of separate comparisons, each comparing one vaccine status with another. For example, 1 PS was derived and used in comparisons of vaccinees who received 3 doses versus the unvaccinated; then a separate PS was derived and used in comparisons of vaccinees who received 4 doses versus the unvaccinated. However, separate models do not allow for easy testing between different levels of exposure and are computationally challenging, with a single analysis often requiring over 1000 PS to account for different vaccination exposure groups and population subgroups being assessed. In addition, the positivity assumption of the PS requires that at no level of the PS is outcome probability close to 0 or 1 [27,28]. The widening difference in characteristics between those who closely follow the recommended vaccination schedule and those remaining unvaccinated put a strain on the positivity assumption

XSL•FO
RenderX

and decreased the appeal of collapsing vaccination categories into two.

Disease Risk Scores to Adjust for Confounding

An alternative to PS is the DRS. A DRS reflects the relationship between potential confounders and the outcome [1-4,7,9,29,30]. Conditioning on a DRS does not balance covariates across treatment groups as a PS would but results in a covariate balance where the potential outcome under the referent condition is independent of a set of covariates [29], which has been called prognostic balance by Hansen [30]. In short, the DRS is an observation's predicted risk of the outcome, assuming the observation was in the referent category of the exposure, reflecting the risk of the outcome relative to other observations. There are well-known generalizable risk scores, such as the Gail model score [31], Framingham score [32], APACHE score [33], and Charlson index [34] that can be used to control for baseline health status. Instead, a study may generate a DRS specific to the outcome and population. In the context of a TND, the outcome is not simply testing positive for the disease of interest but rather testing positive for the disease of interest conditional on receiving care and being tested. A key benefit of using the DRS is potentially only calculating it once for the entire dataset, or a few times for a few key subgroups, rather than generating up to thousands of PS for comparisons using different dichotomous comparisons between multiple vaccination exposure groups.

We compared the performance of DRS-adjusted estimators with multivariable-adjusted estimators that do not aggregate the covariates into a composite score. Several studies have implemented TNDs with this straightforward approach—that is, by fitting logistic regression models that include many covariates in addition to indicators of vaccination status [35-37]. While many analytic methods select variables or adjust for confounding, we are particularly interested in the possibility the DRS allows us to calculate a single score to apply to many analyses.

Methods

Monte Carlo Simulation

We conducted simulations to compare the performance of alternative approaches to the adjustment of VE estimates for potential confounding—either by DRS or by individual covariate adjustment—in VISION network-like scenarios. In order of generation, the simulated data consisted of a bootstrapped sample of individuals, each with (1) a profile of continuous and categorical covariates, (2) a 13-level vaccination status derived conditionally from the covariates, and (3) a 2-level outcome derived conditionally from vaccination status and the covariates. The simulation study used R=1000 replications, which ensured that an estimated coverage of 95% would have a Monte Carlo error of 0.7 [38]. Datasets were generated as follows:

1. A bootstrap sample of size N=1000 or 10,000 observations was drawn from observed emergency department and urgent care encounters in VISION data of adults aged ≥18 years from the Omicron predominance era (December 16, 2021-July 31, 2022). Each observation consisted of all covariates aside from vaccination status and test result for a sampled encounter. This maintained complex relationships between potential confounders in the real-world VISION study. The data used to initiate this simulation study were accessed beginning on August 8, 2022. The authors did not have access to information that could identify individual participants during or after the analysis.
2. Vaccination status was generated as a 13-level exposure variable incorporating the number of vaccine doses received and the time interval since receipt of the most recent dose from a multinomial distribution for each observation, V_i, based on the bootstrapped covariates, Xi (1).

, where k=1,2,…,13 and i=1,2,…, N (1)

β_k was derived from the dataset by multinomial logistic regression and shown in Table S1 in Multimedia Appendix 1. The generated 13-level vaccination status was collapsed into a 5-level vaccination status as shown in Table 1 with approximate frequencies. We focused on VE estimates for each of the 4 levels of vaccination status compared with a reference level of vaccination status.

Next, we used a Bernoulli distribution to generate an outcome, Y_i dependent on the covariates in the bootstrap sample, and generated 13-level vaccination status (2). γ are the coefficients describing the true VE and are shown in Table 1. α are defined in Table S2 in Multimedia Appendix 2. In the VISION data, from which the bootstrap sample was drawn, approximately 22% of encounters were cases. This percentage varied depending on covariates such as season or vaccination status. For example, cases comprise approximately 42% of encounters in January 2022 and only 13% of encounters in May. Among those unvaccinated, approximately 30% are cases, while among those recently vaccinated with their third dose, 11% are cases.

XSL•FO
RenderX

Table 1. Definition of vaccination status categories for data generation and analysis with the true VE used in generating simulated data.[a]

Thirteen-level vaccination status categories for data generation				Five-level vaccination status categories for VE analysis	
Definition[b]	Approximate frequency in simulated data[c,d], %	Coefficient for true VE, γ	Corresponding true VE	Definition	Approximate frequency in simulated data[c,d], %
Unvaccinated	40	__[e]	—	Unvaccinated	40
2 doses, 14-59 d	1	–1.05	65	2 doses, 14-179 d	6
2 doses, 60-119 d	2	–0.60	45	—	—
2 doses, 120-179 d	3	–0.43	35	—	—
2 doses, 180-239 d	5	–0.36	30	2 doses, ≥180 d	24
2 doses, 240-299 d	7	–0.29	25	—	—
2 doses, 300-359 d	6	–0.22	20	—	—
2 doses, ≥360 d	6	–0.16	15	—	—
3 doses, 7-59 d	6	–1.61	80	3 doses, 7-119 d	15
3 doses, 60-119 d	10	–1.39	75	—	—
3 doses, 120-179 d	8	–0.43	35	3 doses, ≥120 d	15
3 doses, 180-239 d	5	–0.16	15	—	—
3 doses, ≥240 d	1	–0.11	10	—	—

[a]VE: vaccine effectiveness.

[b]Vaccination categories are defined by the number of mRNA vaccines received at the index date of the medical encounter and the days between the most recent dose and the index date. The index date is the earlier of the date of specimen collection of a positive test within the allowable testing window (14 d before to 72 h after the encounter) and the time of the medical encounter.

[c]Approximate percent observed in the full dataset from which samples of 10,000 or 1000 are drawn.

[d]Data were generated using a 13-level vaccination status variable, with the approximate frequency of each category shown. This frequency varies across the 1000 replicates. The vector of coefficients, γ, is used as part of the data generation process, and its corresponding values on the VE scale are shown. In the analysis phase, the 13 categories are collapsed to the 5 shown in the column on the right, again with approximate frequencies.

[e]Not available.

Derivation of the DRS

Methods for deriving the DRS have been described elsewhere [1-5,7,29,39]. For this simulation study, we used Miettinen's approach, which models the outcome in relation to the covariates in the entire study population and includes the exposure of the vaccination status in the model with the covariates. The coefficients from the fitted model were used to calculate a predicted DRS for the entire dataset, predicting what risk of the outcome would be for each individual if they were unvaccinated (or if they were given the "reference" level of vaccination status in other scenarios) [5]. Multimedia Appendix 3 [40,41] contains more details about why this approach was selected.

Consideration around the use of a single DRS for the entire VISION study population for each analysis of VE was required. For example, VISION analyses were commonly stratified by age and immunocompromising status. Those with immunocompromising conditions had different vaccine recommendations and may have had reduced VE [42]. Accordingly, in each generated dataset, up to 3 DRS were estimated for each individual depending on their age and immunocompromised (IC) status. One used the full generated dataset, the second used only those with IC conditions according to VISION network criteria, and the third used only those 50 years and older (≥50).

In Miettinen's approach, incorrectly modeling the modification of exposure by baseline covariates can result in estimated scores that are influenced by the magnitude of the exposure effect [30]. Applied to the estimation of DRS, gradient-boosted regression trees look for multilevel interactions that are useful for describing the relationship between patient characteristics, including vaccination status, and the outcome [20]. DRS were estimated using gradient boosted regression and included the exposure and all the covariates. Table S1 in Multimedia Appendix 1 contains a listing of these covariates.

Final VE Models

The VE was calculated in a variety of logistic regression models, calculated as (1–adjusted odds ratio)×100%. Without a DRS, we estimated unadjusted VE, VE adjusted only for the key variables of patient age (as a spline), calendar date from January 1, 2021 (epi-day, as a spline), and VISION site and sub-region of the health care facility (site-region) (referred to as multivariable key), and VE adjusted for the key variables plus all other variables in the DRS (referred to as multivariable all). Note that the term spline refers to a natural cubic spline with 3 knots placed at the 25th, 50th, and 75th percentiles.

While a PS is typically used to balance comparisons of the exposed versus the unexposed by inverse probability weighting or by matching on the PS [43], there is less theory or experience

XSL•FO
RenderX

pertaining to the use of the DRS in the final model that yields the adjusted VE [2]. A model may include DRS as a covariate [44], or a model could be conditional on DRS [45], which under certain conditions, may yield an average of the individual treatment effects of those treated (ATT) [29]. Well-known generalizable risk scores, such as the Gail model score, Framingham score, APACHE score, and Charlson index, are sometimes included in models or are used for stratification. The risk score we investigated was not designed to be generalizable, however, the modeling strategies employed for including a DRS in a VE model are similar. We investigated DRS inclusion in VE models as strata in a conditional model as well as a continuous covariate fit with a flexible spline. Conditional models used DRS as centiles, or deciles with epi-week, and site as strata. Conditional models that are only conditioned on DRS and models that included DRS as a continuous covariate also adjusted for age (as a spline), epi-day (as a spline), and site region. Models that were conditioned on DRS epi-week and site were also adjusted for age (as a spline). Vaccination status in the final VE model was collapsed to the 5-level category. Table 2 summarizes the calculated models.

XSL•FO
RenderX

Table 2. Description of final vaccine effectiveness (VE) models assessed in the simulation study.

Model name	Subset of data used to estimate DRS[a,b]. Subset of data used for estimating VE[c].	Model stratification and adjustment	Dataset totals, n
Unadjusted	DRS dataset: no DRS. VE dataset: full simulated dataset.	Vaccination status only	10,000 and 1000
Multivariable all	DRS dataset: no DRS. VE dataset: full simulated dataset.	• Stratification: none • Adjustment: variables that would go into the DRS including age (spline), epi-day (spline), and site region	10,000 and 1000
Multivariable key	DRS dataset: no DRS. VE dataset: full simulated dataset.	• Stratification: none • Adjustment: age (spline), epi-day (spline), and site region	10,000 and 1000
Stratified week, site, DRS	DRS dataset: full simulated dataset. VE dataset: full simulated dataset.	• Stratification: DRS decile, epi-week, and site • Adjustment: age (spline)	10,000 and 1000
Stratified DRS	DRS dataset: full simulated dataset. VE dataset: full simulated dataset.	• Stratification: DRS centile • Adjustment: age (spline), epi-day (spline), and site region	10,000 and 1000
Spline DRS	DRS dataset: full simulated dataset. VE dataset: full simulated dataset.	• Stratification: none • Adjustment: DRS included with a natural spline with 4 degrees of freedom, age (spline), epi-day (spline), and site-region	10,000 and 1000
Multivariable key IC[d]	1. DRS datasets: no DRS. VE dataset: IC subset. 2. DRS datasets: IC subset. VE dataset: IC subset.	• Stratification: none • Adjustment: age (spline), epi-day (spline), and site	10,000
Stratified week, site, IC DRS	1. DRS datasets: full simulated dataset. VE dataset: IC subset. 2. DRS datasets: IC subset. VE dataset: IC subset.	• Stratification: DRS decile, epi-week, and site • Adjustment: age (spline)	10,000
Stratified IC DRS	1. DRS datasets: full simulated dataset. VE dataset: IC subset. 2. DRS datasets: IC subset. VE dataset: IC subset.	• Stratification: DRS centile • Adjustment: age (spline), epi-day (spline), and site	10,000
Spline IC DRS	1. DRS datasets: full simulated dataset. VE dataset: IC subset. 2. DRS datasets: IC subset. VE dataset: IC subset.	• Stratification: none • Adjustment: DRS included with a natural spline with 4 degrees of freedom, age (spline), epi-day (spline), and site	10,000
Multivariable key ≥50[e]	1. DRS datasets: no DRS. VE dataset: ≥50 subset. 2. DRS datasets: ≥50 subset. VE dataset: ≥50 subset.	• Stratification: none • Adjustment: only for age (spline), epi-day (spline), and site region	10,000
Stratified week, site, ≥50 DRS	1. DRS datasets: full simulated dataset. VE dataset: ≥50 subset. 2. DRS datasets: ≥50 subset. VE dataset: ≥50 subset.	• Stratification: DRS decile, epi-week, and site • Adjustment: age (spline)	10,000
Stratified ≥50 DRS	1. DRS datasets: full simulated dataset. VE dataset: ≥50 subset 2. DRS datasets: ≥50 subset. VE dataset: ≥50 subset.	• Stratification: DRS centile • Adjustment: age (spline), epi-day (spline), and site region	10,000

XSL•FO
RenderX

Model name	Subset of data used to estimate DRS[a,b]. Subset of data used for estimating VE[c].	Model stratification and adjustment	Dataset totals, n
Spline ≥50	1. DRS datasets: full simulated dataset. VE dataset: ≥50 subset. 2. DRS datasets: ≥50 subset. VE dataset: ≥50 subset.	• Stratification: none • Adjustment: DRS included with a natural spline with 4 degrees of freedom, age (spline), epi-day (spline), and site region	10,000

[a]DRS: disease risk score.

[b]One goal of this simulation study is to consider how the DRS might behave in smaller subsets of the full dataset, specifically a subset of patients with IC conditions and a subset of patients 50 years or older. For this reason, when analyzing these subsets, we consider a DRS built from the full dataset but applied to a smaller subset and a DRS built specifically from that smaller subset.

[c]VE: vaccine effectiveness.

[d]IC: patients with immunocompromising conditions.

[e]≥50: patients aged 50 years or older.

Assessing Model Performance

To evaluate performance, we examined bias, SE, the ratio of mean SE to empirical SE, and coverage. The true VE is built into the data generation mechanism, and these known values are shown in Table 1 for the 13 levels of vaccination status. However, VE was estimated for the collapsed 5-level vaccination status, $Vcollapsed_i$. Therefore, for each replication, r, the true effect for each 5-level vaccination status, $\gamma_{m,r}$, was approximated with a weighted average (weighted by sample size in each of the 13 vaccination status categories) as shown below (3) and summarized as a VE (4).

The estimated VE for each vaccine status in each replication, , was calculated as shown in (5).

Percent VE bias was calculated as the average percent bias across the replications on the VE scale (6).

The ratio of the observed SE of the OR compared with the empirical SE was calculated as shown in (7).

Presentation of Results

Simulation results are presented in figures depicting the complete distribution of observed percent VE bias and SE of $\gamma_{m,r}$, for each vaccination status. The average across all replications is shown as well as the 2.5th and 97.5th percentiles. The empirical SE, or SD of the coefficients, is presented as a vertical bar. In addition to low bias, an ideal model will provide estimates of SE that are also averaged across replications, close to the SD of the estimated coefficients. The figures include coverage. All results are in Tables S3 and S4 in Multimedia Appendices 4 and 5. Coverage is detailed in the results, with CIs for the coverage probabilities based on the binomial distribution of the R coverage indicators [38]. All simulations and analyses were conducted with R Statistical Software (R Foundation for Statistical Computing) [46].

Ethical Considerations

This study was reviewed and approved by the institutional review boards at participating sites and under a reliance agreement between the CDC and the Westat institutional review board (FWA# FWA00005551, expiry date 10/13/2027, IRB project number 6201.08). This activity was reviewed by the CDC and was conducted consistent with applicable federal law and CDC policy (eg, 45 CFR part 46.102(l)(2), 21 CFR part 56; 42 USC §241(d); 5 USC §552a; 44 USC §3501). This activity was reviewed and approved as a research activity by one VISION site. This study presented minimal risk to participants because there was no interaction or intervention with patients; therefore, a waiver of informed consent was granted.

Results

Comparing Strategies for Estimating VE

When N=10,000, the unadjusted models resulted in an average percent bias of VE ranging from –37% to 124% across the 4 vaccination categories (Figure 1, Table S3 in Multimedia Appendix 4). Average percent bias was dramatically lower in the models that accounted for the covariates: the 2 multivariable models without DRS (range for multivariable with all covariates, including key covariates from –2.2 to 3.8, multivariable with only key covariates from –5.3 to 6.1), the stratified model with DRS decile epi-week and site (range –2.6 to 2.3), the stratified model with DRS centile (range –1.5 to 4.2), and the model with DRS as a spline (range –1.6 to 4.1). Bias varied less across the 5 strategies for estimating VE than across the 4 levels of vaccination status for which VE was estimated. For example, the bias in estimates of 3-dose VE within 120 days ranged from –10% to 10% approximately, whereas the bias in estimates of 2-dose VE within 180 days ranged from –50% to 40% approximately. The difference in bias between vaccination status levels is related to the different sample sizes expected in each vaccine status group and the expected increasing instability of

XSL•FO
RenderX

the percent bias when VE is closer to 0 (ie, 5 units is a much greater percent of 25 than it is of 80).

Figure 1. Summary of simulation results in the overall sample of 10,000. DRS: disease risk score; VE: vaccine effectiveness.

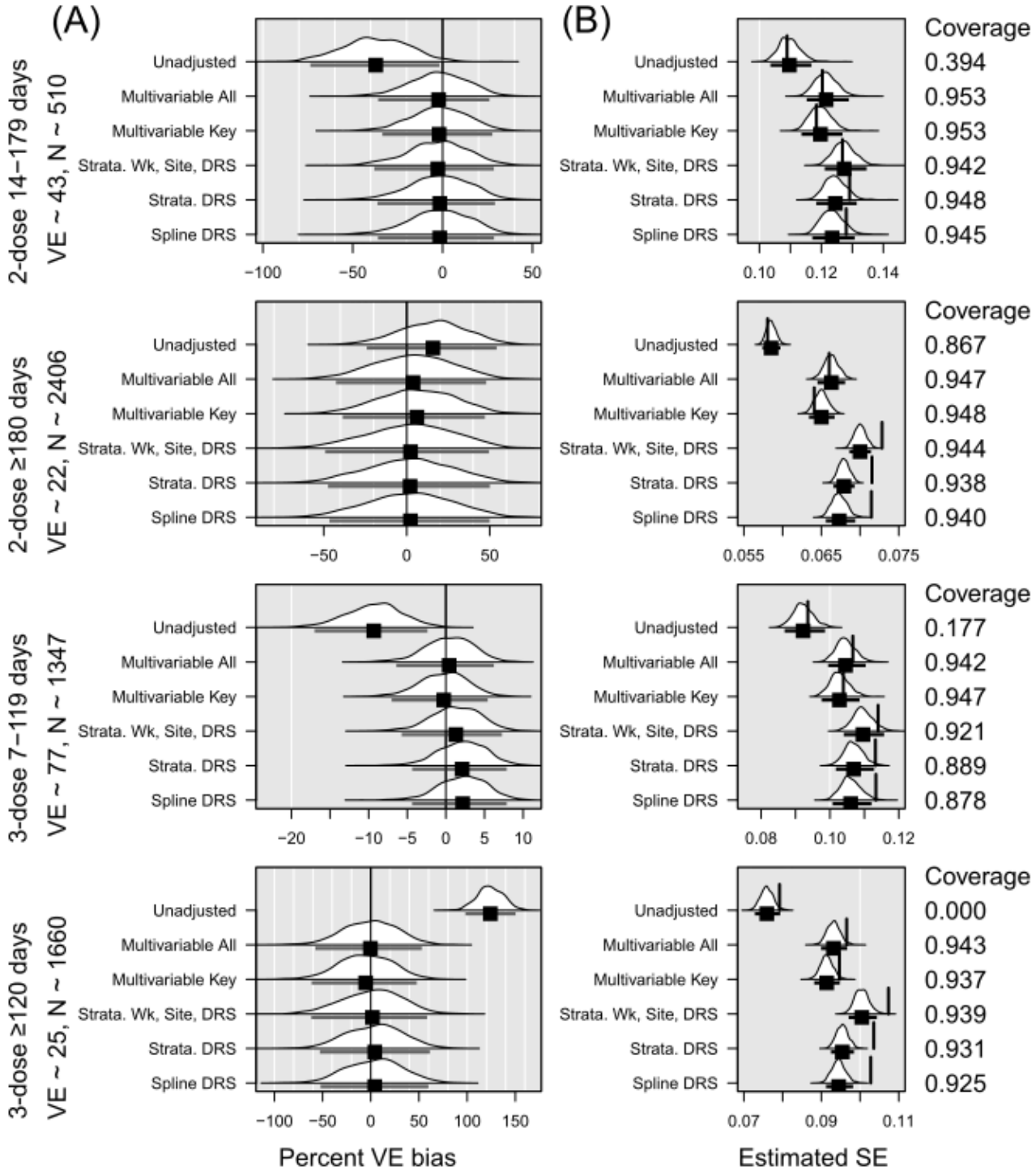

In all figures, panel A presents the percent VE bias for each vaccination level, with a vertical black line at 0. The distribution across 1000 replications is shown for each model strategy. The mean percent VE bias, represented with a black square, with an interval indicating the observed 2.5th and 97.5th percentiles, appears below each distribution. Panel B presents the distribution across 1000 replications of the estimated standard error for each vaccination level. The mean standard error, represented with a black square, with an interval indicating the observed 2.5th and 97.5th percentiles appears below the distribution. The vertical line shows the empirical standard error (SE), or SD of the coefficients. Ideally, the vertical line aligns with the mean standard error. Means to the left of the empirical SE indicate underestimation of the SE. Coverage is shown on the right of Panel B. Coverage is defined as the proportion of confidence intervals from the simulated analyses that include the true VE. Ideally, this would be close to 95%.

Across all 4 vaccination statuses, the multivariable models had the lowest SE. In most instances, the average SE came closest

XSL•FO
RenderX

to the SD of the coefficients in the multivariable models. In all but the 2-dose 14-179 days category, the models incorporating DRS as a spline and the models stratified on DRS centile resulted in empirical SE greater than the 97.5th percentile of observed standard errors. The model stratified on decile of DRS, epi-week, and site fared slightly better, with empirical SE greater than the 97.5th percentile in 2 of the 4 vaccination categories. However, the models stratified on decile of DRS, epi-week, and site had the highest mean SE in all vaccination categories. Models including DRS as a spline had the greatest underestimation of SE, as demonstrated with the lowest ratio of mean SE to empirical SE (Table S3 in Multimedia Appendix 4). With the exception of the multivariable all or multivariate key models, all models yielded CIs that covered the true VE in less than 95% of replicated analyses for at least one vaccination status, as shown in Table 3.

We also considered subgroups of the N=10,000 dataset defined by immunocompromised (IC) status and age, as these groups were routinely studied by VISION and others to guide vaccination policy decisions. Out of N=10,000, approximately 532 fell into the IC subgroup and 4782 into the 50 years or older subgroup. In the IC subgroup, the multivariable model with key covariates and without the DRS performed best in terms of bias—its VE estimates for each level of vaccination status were least biased (Figure 2 and Table S4 in Multimedia Appendix 5). This multivariable model also had the lowest SE while achieving 95% coverage across two of the four levels of vaccination status (Table 3). For estimates of 2-dose VE within 180 days, the coverage with the multivariable model was 96.8% (CI 95.1-97.5), and for 3-dose VE within 120 days, the coverage with the multivariable model was also 96.8% (CI 95.7-97.9). While having slightly higher bias, the spline model and the model stratified using DRS derived from the full 10,000 achieved 95% coverage in all 4 vaccination categories. The IC group was so small that several replications yielded very high SE estimates for at least one vaccination status indicating that a meaningful VE estimate was not obtained. This occurred in 21 of 1000 replications for the multivariate model, 11-132 replications for the models using the DRS from the full model, and 21-230 replications for the models using the DRS from the IC subset. In the older subgroup, the multivariable model with key covariates and without DRS again performed best in terms of bias across all 4 vaccination statuses (Figure 3 and Table S4 in Multimedia Appendix 5). Again, the multivariable model often had the lowest SE, which was also closest to the SD of the coefficients. Only the multivariable model achieved 95% coverage for its VE estimates for all four levels of vaccination status. Only the model stratified on DRS (from the full cohort), epi-week, and site-region achieved 95% coverage for 3 of the 4 levels of vaccination status; the other models only achieved this for 2 or fewer levels of vaccination status (Table 3). In both the IC subgroup and the 50 years or older subgroup, the models using a DRS derived from the subgroup tended to perform worse than models with a DRS derived from the full sample of 10,000.

When N=1000, the sizes in each vaccination category were dramatically lower, ranging from 50 in the 2-dose 14-179 days group to 242 in the 2-dose ≥180 days group. Again, the multivariable models, particularly the one with all the covariates, had the least bias in VE estimates for each level of vaccination status (Figure 4 and Table S3). The multivariable models, along with the model stratified by epi-week, site, and DRS decile all achieved 95% coverage across the four vaccination statuses (Table 3). The model stratified by DRS centile and the model with a DRS spline had coverages slightly below 90% for 3-dose VE within 120 days (89.6% coverage CI: 87.7%-91.5% and 89.1% coverage CI: 87.2%-91%, respectively).

XSL•FO
RenderX

Table 3. Estimates of coverage of the true vaccine effectiveness (VE) with 95% CIs.

Model strategy	Two-dose 14-179 d, coverage[a] (95% CI[b])	Two-dose ≥180 d, coverage[a] (95% CI[b])	Three-dose 7-119 d, coverage[a] (95% CI[b])	Three-dose ≥120 d, coverage[a] (95% CI[b])
N=10,000				
Multivariable all	95.3 (94.0-96.6)	94.7 (93.3-96.1)	94.2 (92.8-95.6)	94.3 (92.9-95.7)
Multivariable key[c]	95.3 (94.0-96.6)	94.8 (93.4-96.2)	94.7 (93.3-96.1)	93.7 (92.2-95.2)
Strata: week site, DRS[d]	94.2 (92.8-95.6)	94.4 (93.0-95.8)	92.1 (90.4-93.8)	93.9 (92.4-95.4)
Strata: DRS	94.8 (93.4-96.2)	93.8 (92.3-95.3)	88.9 (87.0-90.8)	93.1 (91.5-94.7)
Spline DRS	94.5 (93.1-95.9)	94.0 (92.5-95.5)	87.8 (85.8-89.8)	92.5 (90.9-94.1)
IC[e] subgroup				
Multivariable key[c]	96.3 (95.1-97.5)	94.9 (93.5-96.3)	96.8 (95.7-97.9)	94.7 (93.3-96.1)
Strata: week site, DRS	97.8 (96.9-98.7)	95.6 (94.3-96.9)	96.6 (95.5-97.7)	97.0 (95.9-98.1)
Strata: DRS	96.0 (94.8-97.2)	95.8 (94.6-97.0)	95.6 (94.3-96.9)	96.1 (94.9-97.3)
Spline DRS	96.0 (94.8-97.2)	94.5 (93.1-95.9)	94.7 (93.3-96.1)	94.7 (93.3-96.1)
Strata: week, site, IC DRS	96.6 (95.5-97.7)	98.3 (97.5-99.1)	98.4 (97.6-99.2)	98.2 (97.4-99.0)
Strata: IC DRS	95.4 (94.1-96.7)	97.6 (96.7-98.5)	95.8 (94.6-97.0)	94.2 (92.8-95.6)
Spline IC DRS	92.3 (90.6-94.0)	98.5 (97.7-99.3)	95.2 (93.9-96.5)	93.8 (92.3-95.3)
≥50[f] subgroup				
Multivariable key[c]	93.5 (92.0-95.0)	94.4 (93.0-95.8)	94.1 (92.6-95.6)	94.1 (92.6-95.6)
Strata: week site, DRS	94.9 (93.5-96.3)	93.8 (92.3-95.3)	93.0 (91.4-94.6)	93.5 (92.0-95.0)
Strata: DRS	93.7 (92.2-95.2)	94.2 (92.8-95.6)	90.6 (88.8-92.4)	93.0 (91.4-94.6)
Spline DRS	94.0 (92.5-95.5)	93.4 (91.9-94.9)	89.7 (87.8-91.6)	93.2 (91.6-94.8)
Strata: week, site, IC DRS	92.7 (91.1-94.3)	93.6 (92.1-95.1)	91.1 (89.3-92.9)	94.2 (92.8-95.6)
Strata: IC DRS	92.3 (90.6-94.0)	92.3 (90.6-94.0)	86.1 (84.0-88.2)	92.9 (91.3-94.5)
Spline IC DRS	91.5 (89.8-93.2)	92.2 (90.5-93.9)	84.8 (82.6-87.0)	92.3 (90.6-94.0)
N=1000				
Multivariable all	95.6 (94.3-96.9)	94.9 (93.5-96.3)	94.8 (93.4-96.2)	95.2 (93.9-96.5)
Multivariable key[c]	95.8 (94.6-97.0)	95.1 (93.8-96.4)	94.9 (93.5-96.3)	94.8 (93.4-96.2)
Strata: week site, DRS	95.6 (94.3-96.9)	94.1 (92.6-95.6)	93.8 (92.3-95.3)	95.2 (93.9-96.5)
Strata: DRS	94.2 (92.8-95.6)	92.8 (91.2-94.4)	89.6 (87.7-91.5)	93.3 (91.8-94.8)
Spline DRS	95.0 (93.6-96.4)	93.1 (91.5-94.7)	89.1 (87.2-91.0)	93.5 (92.0-95.0)

[a]Coverage is defined as the proportion of confidence intervals from the simulated analyses that include the true VE. Ideally, this would be close to 95%.
[b]Confidence intervals were calculated using the properties of the binomial distribution of the indicators that each replication covers the true VE.
[c]Key variables include simple adjustment only for age (spline), epi-day (spline), and site region.
[d]DRS: disease risk score.
[e]IC: immunocompromising conditions.
[f]≥50: patients 50 years old or older.

XSL•FO
RenderX

Figure 2. Summary of simulation results in the immunocompromised subset of 10,000. DRS: disease risk score; IC: immunocompromised; VE: vaccine effectiveness.

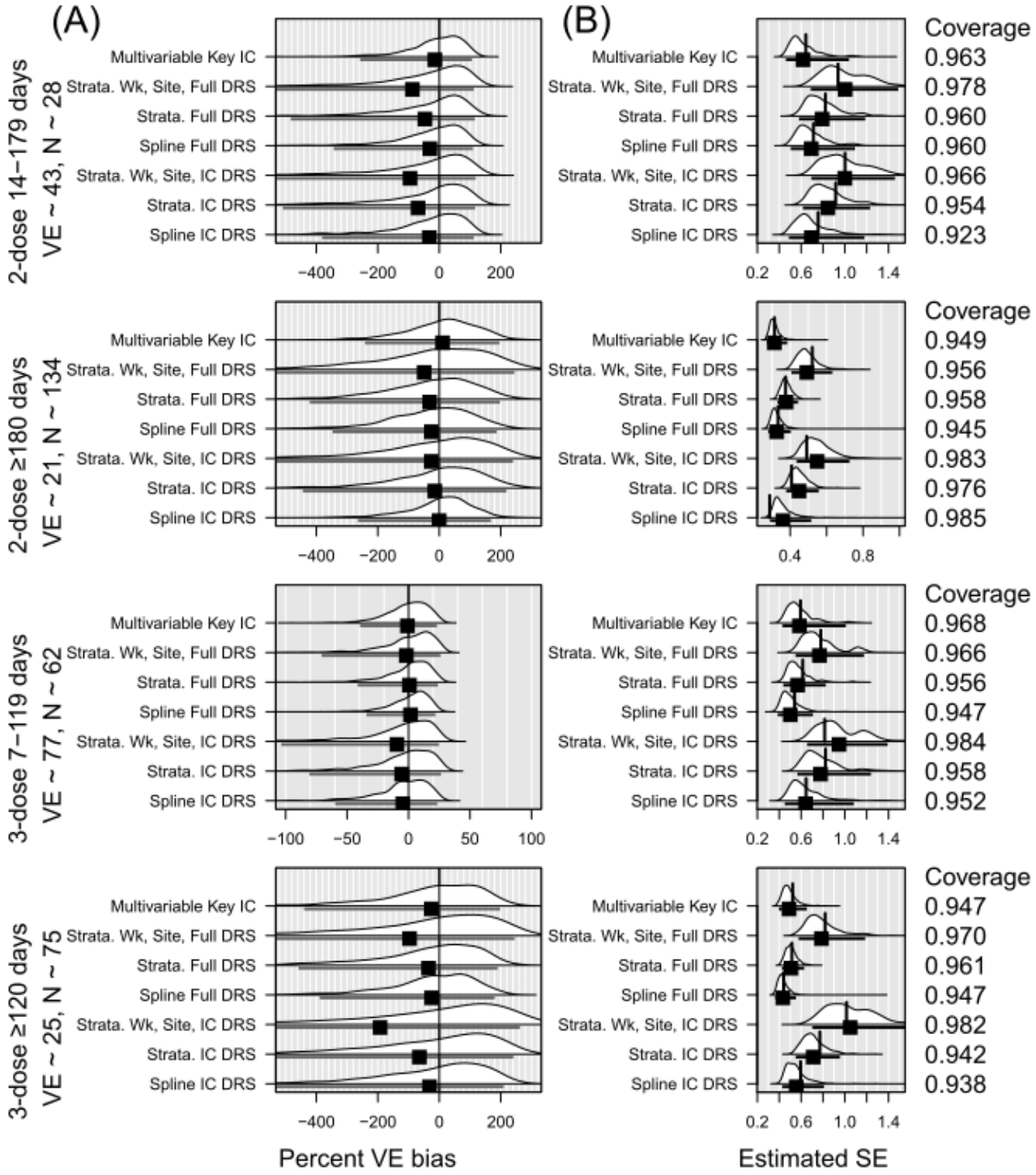

XSL•FO
RenderX

Figure 3. Summary of simulation results in the 50 years or older subset of 10,000. DRS: disease risk score; VE: vaccine effectiveness.

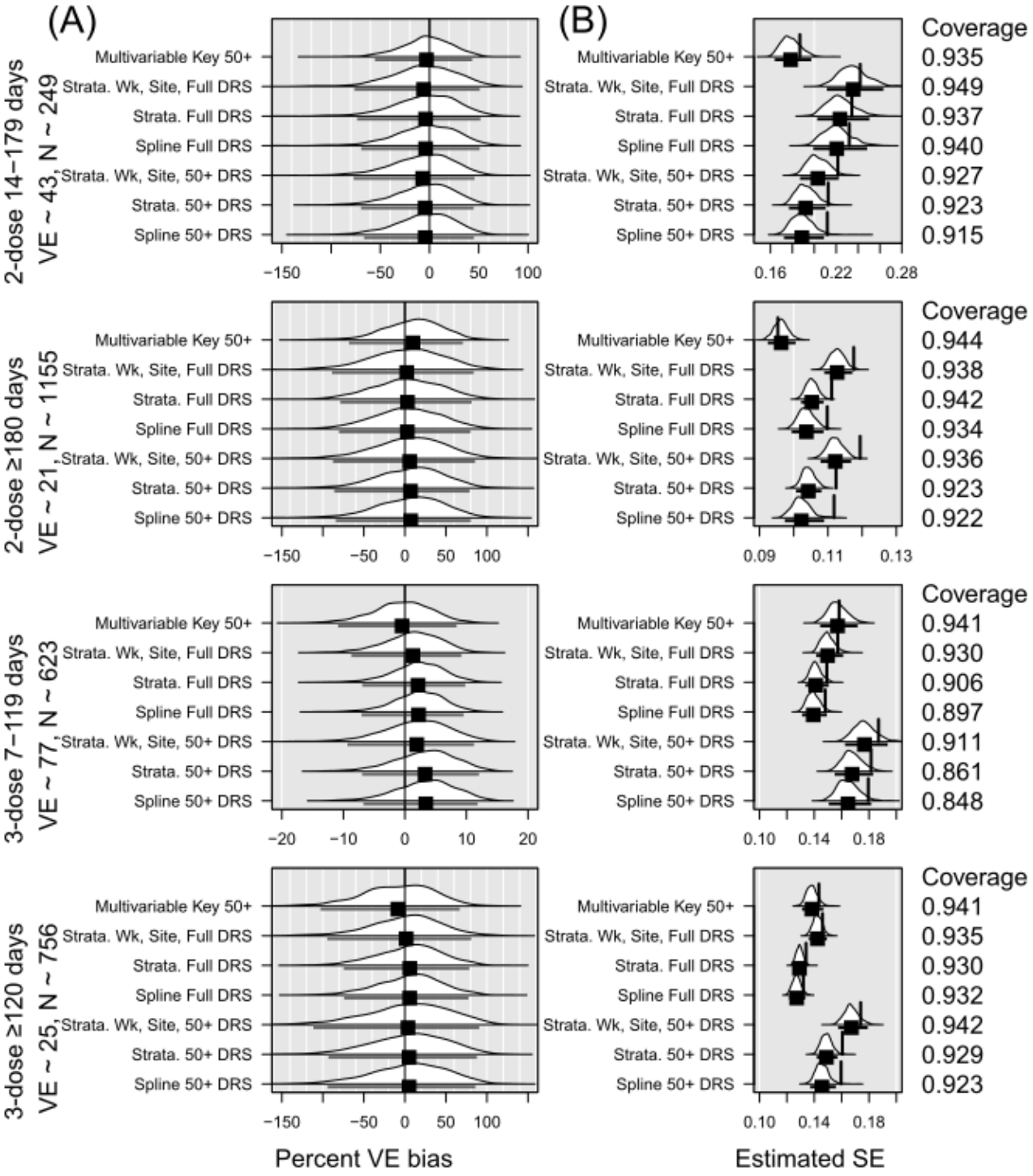

XSL•FO
RenderX

Figure 4. Summary of simulation results in a sample of size 1000. DRS: disease risk score; VE: vaccine effectiveness.

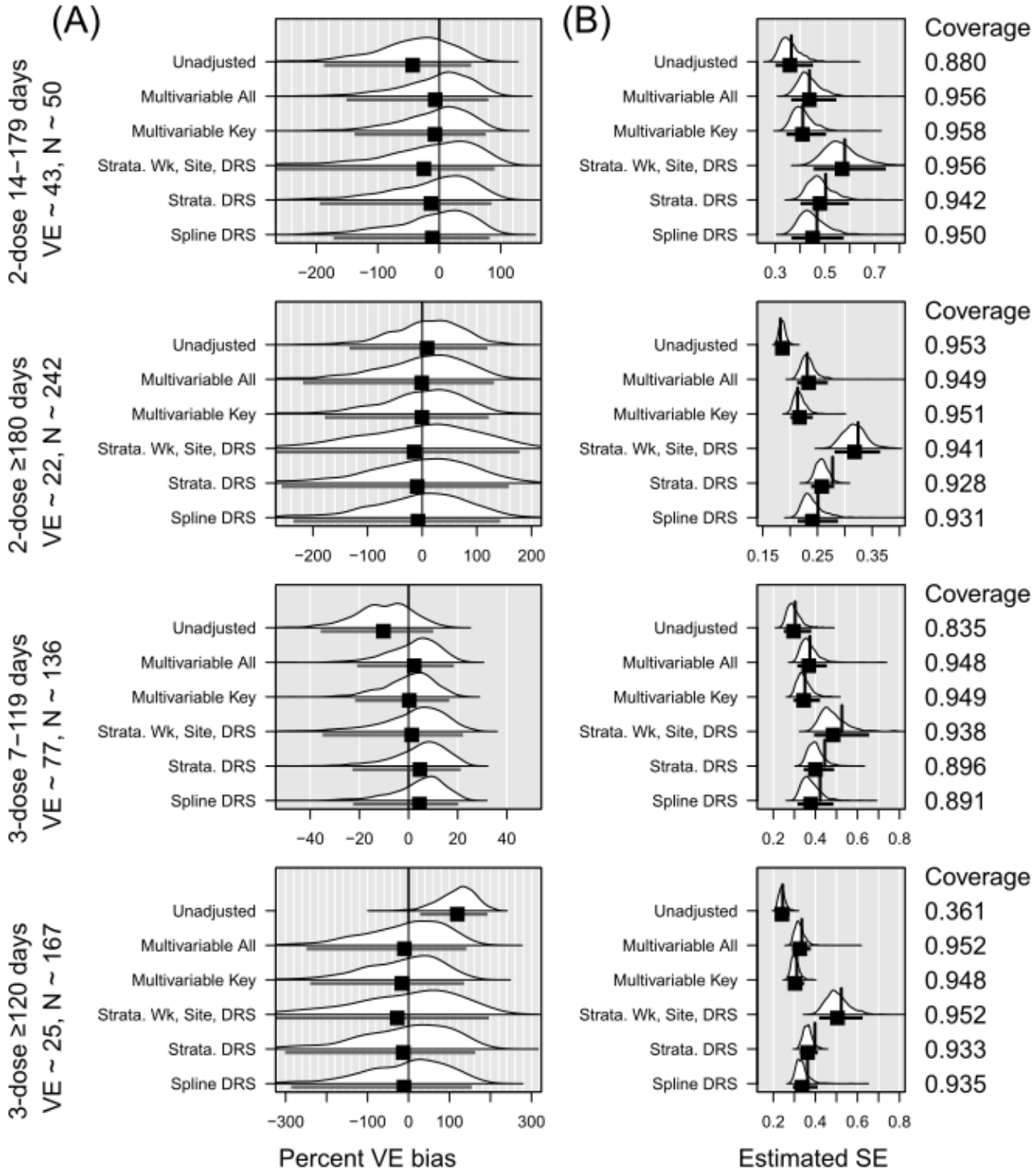

DRS-Related Decisions

When comparing the bias in analyses of subgroups defined by IC or age, we found little difference in the bias of VE between using a DRS built from the full dataset of 10,000 compared with using a DRS built exclusively from the subset of interest (Figures 2 and 3). When a difference was found, models using the DRS built from the full cohort tended to perform better than the models built from the smaller subset. Among the 12 VE estimates for the IC subgroup—a VE estimate for each of the 4 levels of vaccination status obtained by each of the 3 ways of using the DRS—9 of the 12 achieved 95% coverage when using a DRS from the full cohort compared with 5 of 12 when using the DRS from the IC subgroup. Among the 12 VE estimates for the subgroup aged ≥50 years, 6 of 12 achieved 95% coverage when using a DRS from the full cohort compared with 2 of 12 when using the DRS from the 50 years or older subgroup. Model-fitting challenges occurred more frequently when using the subgroup-specific DRS, particularly in the smaller IC subgroup.

XSL•FO
RenderX

Discussion

Principal Findings

This simulation study found that multivariable covariate adjustment, either with all the covariates or a subset of key covariates, performed well in the context of VISION's case-control test-negative studies of VE. Adjustment for a DRS comprised of the covariates performed adequately but tended to overestimate the precision of VE estimates.

It is possible the underestimate of the SE we have observed could be attributed to a high correlation between the confounders and vaccination status. Early studies of DRSs suggested they can help reduce the dimensionality of analyses that try to adjust for multiple covariates but that DRS adjustment tends to overestimate the precision of the effect estimate of interest [47]. Later studies suggested that the corresponding exaggeration of the statistical significance of a finding (eg, a finding about VE) would be trivial unless the covariates are very strongly associated with the exposure (eg, vaccination status) or with each other [1,48].

It should not be surprising that simple covariate adjustment performed well, given VISION's large samples of cases and controls. Other large VE studies with TNDs have also successfully employed individual covariate adjustment [35-37]. Furthermore, in previous VISION network studies, we observed that only a few of the available covariates (primarily age, calendar time, and geographic location) accounted for most of the confounding apparent in unadjusted VE estimates, which makes it feasible to adjust for the covariates individually even in subsets that might otherwise be too small to adjust for several dozen covariates or more. Even in our subgroup analyses, we did not find that use of the dimension-reducing DRS improved performance compared with straightforward adjustment for the individual covariates.

In the context of settings other than VISION, where sample sizes may be smaller, and the relevant covariates may be more numerous, the "curse of dimensionality" may render covariate adjustment more problematic. If an appropriate DRS has already been derived from large samples in similar settings, such a DRS could be helpful in several ways. First, an appropriate DRS can avoid overfitting. DRS adjustment outperformed individual covariate adjustment in some of our smaller subgroup analyses, especially when we used a DRS derived from the larger overall sample.

Second, confounding may arise from nonlinearities and interactions among covariate-outcome associations that could easily be overlooked unless previously captured by a DRS derived from large datasets using flexible machine learning methods. In this simulation study, we used boosted regression to derive the DRS, as described above, but our data-generating mechanism did not insert nonlinearities and interactions that would be undetectable in smaller datasets and would be challenging to specify in a logistic regression model with individual covariate adjustment.

Third, a DRS could be part of a hybrid approach to covariate adjustment. VE estimates could be derived from logistic regression models that include a DRS plus a few key covariates (which may also be components of the DRS). The key covariates would be those with high potential for confounding because of strong associations with the outcome (and exposure) that may differ in the current study population than in the population where the DRS was derived.

Fourth, a DRS can be an intuitive way to adjust for confounding and an intuitive tool to explore effect modification. Typically, effect modification examines one risk factor at a time—for example, examining whether the benefits of vaccination differ by age group. A DRS can be used to account for multiple risk factors as we examine whether the benefits of vaccination differ by level of risk. Furthermore, a DRS can facilitate the interpretation of findings. For example, if VE is found to be similar across levels of the DRS, yet the risk is 10-fold higher in the highest DRS decile as compared with the lowest, then we can infer that the vaccine benefits the highest decile 10-fold more (on the scale of cases prevented per 1000 persons vaccinated) than it benefits the lowest decile. However, a DRS may be less intuitive as a risk score in our test-negative VISION study than in a population-based cohort study to the extent that our test-negative controls are restricted to individuals with "a COVID-19-like illness" and may not be representative of the underlying population at risk. In the underlying population, risk factors for hospitalization for CLI may differ from risk factors for hospitalization for COVID-19.

This simulation study was motivated by challenges facing VISION's studies of VE. The scenarios we simulated emulate those examined by VISION, and this may limit the generalizability of our findings to other settings.

Conclusions

Our simulations found that logistic regression with individual covariate adjustment performed well in scenarios similar to those studied by the VISION network and is generally the current approach employed by the VISION network. DRS-adjusted models performed adequately but not as well as models that adjusted for the covariates individually.

Data Availability

The datasets generated during and/or analyzed during this study are not publicly available due to data sharing agreements between CDC and VISION network partner institutions prohibiting it.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
Definition of effects of covariates on 13-level exposure, β, for multinomial model determining simulated exposure with exposure status defined by number of mRNA doses and time since most recent dose.
[DOCX File , 33 KB - formative_v9i1e58981_app1.docx]

Multimedia Appendix 2
Definition of other coefficients, α, for the Bernoulli distribution defining the probability of the generated outcome.
[DOCX File , 17 KB - formative_v9i1e58981_app2.docx]

Multimedia Appendix 3
Who is included in the disease risk score model?
[DOCX File , 24 KB - formative_v9i1e58981_app3.docx]

Multimedia Appendix 4
Bias and standard error results, 5-level vaccine effectiveness, N=10,000 and 1000.
[DOCX File , 29 KB - formative_v9i1e58981_app4.docx]

Multimedia Appendix 5
Bias and SE results, subgroup effect 5-level vaccine effectiveness.
[DOCX File , 30 KB - formative_v9i1e58981_app5.docx]

References

1. Arbogast PG, Kaltenbach L, Ding H, Ray WA. Adjustment for multiple cardiovascular risk factors using a summary risk score. Epidemiology 2008;19(1):30-37. [doi: 10.1097/EDE.0b013e31815be000] [Medline: 18091000]
2. Arbogast P, Seeger J, DMCSVW Group. Summary variables in observational research: propensity scores and disease risk scores. Effective health care program research report.: Agency for Healthcare Research and Quality; 2012. URL: https://effectivehealthcare.ahrq.gov/products/observational-research-scores/research [accessed 2024-12-19]
3. Arbogast PG, Ray WA. Use of disease risk scores in pharmacoepidemiologic studies. Stat Methods Med Res 2009;18(1):67-80. [doi: 10.1177/0962280208092347] [Medline: 18562398]
4. Arbogast PG, Ray WA. Performance of disease risk scores, propensity scores, and traditional multivariable outcome regression in the presence of multiple confounders. Am J Epidemiol 2011;174(5):613-620. [doi: 10.1093/aje/kwr143] [Medline: 21749976]
5. Miettinen OS. Stratification by a multivariate confounder score. Am J Epidemiol 1976;104(6):609-620. [doi: 10.1093/oxfordjournals.aje.a112339] [Medline: 998608]
6. Cadarette SM, Gagne JJ, Solomon DH, Katz JN, Stürmer T. Confounder summary scores when comparing the effects of multiple drug exposures. Pharmacoepidemiol Drug Saf 2010;19(1):2-9 [FREE Full text] [doi: 10.1002/pds.1845] [Medline: 19757416]
7. Tadrous M, Gagne JJ, Stürmer T, Cadarette SM. Disease risk score as a confounder summary method: systematic review and recommendations. Pharmacoepidemiol Drug Saf 2013;22(2):122-129 [FREE Full text] [doi: 10.1002/pds.3377] [Medline: 23172692]
8. Stürmer T, Schneeweiss S, Brookhart MA, Rothman KJ, Avorn J, Glynn RJ. Analytic strategies to adjust confounding using exposure propensity scores and disease risk scores: nonsteroidal antiinflammatory drugs and short-term mortality in the elderly. Am J Epidemiol 2005;161(9):891-898 [FREE Full text] [doi: 10.1093/aje/kwi106] [Medline: 15840622]
9. Glynn RJ, Gagne JJ, Schneeweiss S. Role of disease risk scores in comparative effectiveness research with emerging therapies. Pharmacoepidemiol Drug Saf 2012;21 Suppl 2(Suppl 2):138-147 [FREE Full text] [doi: 10.1002/pds.3231] [Medline: 22552989]
10. Thompson MG, Stenehjem E, Grannis S, Ball SW, Naleway AL, Ong TC, et al. Effectiveness of Covid-19 vaccines in ambulatory and inpatient care settings. N Engl J Med 2021;385(15):1355-1371. [doi: 10.1056/nejmoa2110362] [Medline: 34496194]
11. Cates J, Lucero-Obusan C, Dahl RM, Schirmer P, Garg S, Oda G, et al. Risk for in-hospital complications associated with COVID-19 and influenza - veterans health administration, United States, October 1, 2018-May 31, 2020. MMWR Morb Mortal Wkly Rep 2020;69(42):1528-1534 [FREE Full text] [doi: 10.15585/mmwr.mm6942e3] [Medline: 33090987]
12. Garg S, Kim L, Whitaker M, O'Halloran A, Cummings C, Holstein R, et al. Hospitalization rates and characteristics of patients hospitalized with laboratory-confirmed coronavirus disease 2019 - COVID-NET, 14 States, March 1-30, 2020. MMWR Morb Mortal Wkly Rep 2020;69(15):458-464 [FREE Full text] [doi: 10.15585/mmwr.mm6915e3] [Medline: 32298251]
13. Kim L, Garg S, O'Halloran A, Whitaker M, Pham H, Anderson EJ, et al. Risk factors for intensive care unit admission and in-hospital mortality among hospitalized adults identified through the US coronavirus disease 2019 (COVID-19)-associated

XSL•FO
RenderX

hospitalization surveillance network (COVID-NET). Clin Infect Dis 2021;72(9):e206-e214 [FREE Full text] [doi: 10.1093/cid/ciaa1012] [Medline: 32674114]
14. Chua H, Feng S, Lewnard JA, Sullivan SG, Blyth CC, Lipsitch M, et al. The use of test-negative controls to monitor vaccine effectiveness: a systematic review of methodology. Epidemiology 2020;31(1):43-64 [FREE Full text] [doi: 10.1097/EDE.0000000000001116] [Medline: 31609860]
15. Foppa IM, Ferdinands JM, Chaves SS, Haber MJ, Reynolds SB, Flannery B, et al. The case test-negative design for studies of the effectiveness of influenza vaccine in inpatient settings. Int J Epidemiol 2016;45(6):2052-2059. [doi: 10.1093/ije/dyw022] [Medline: 26979985]
16. Jackson ML, Nelson JC. The test-negative design for estimating influenza vaccine effectiveness. Vaccine 2013;31(17):2165-2168. [doi: 10.1016/j.vaccine.2013.02.053] [Medline: 23499601]
17. Joffe MM, Rosenbaum PR. Invited commentary: propensity scores. Am J Epidemiol 1999;150(4):327-333. [doi: 10.1093/oxfordjournals.aje.a010011] [Medline: 10453808]
18. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983;70(1):41-55. [doi: 10.1093/biomet/70.1.41]
19. D'Agostino RB. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat Med 1998;17(19):2265-2281. [doi: 10.1002/(sici)1097-0258(19981015)17:19<2265::aid-sim918>3.0.co;2-b] [Medline: 9802183]
20. McCaffrey DF, Ridgeway G, Morral AR. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. Psychol Methods 2004;9(4):403-425. [doi: 10.1037/1082-989X.9.4.403] [Medline: 15598095]
21. Rosenblum HG, Wallace M, Godfrey M, Roper LE, Hall E, Fleming-Dutra KE, et al. Interim recommendations from the advisory committee on immunization practices for the use of bivalent booster doses of COVID-19 vaccines - United States, October 2022. MMWR Morb Mortal Wkly Rep 2022;71(45):1436-1441 [FREE Full text] [doi: 10.15585/mmwr.mm7145a2] [Medline: 36355612]
22. Mbaeyi S, Oliver SE, Collins JP, Godfrey M, Goswami ND, Hadler SC, et al. The advisory committee on immunization practices' interim recommendations for additional primary and booster doses of COVID-19 vaccines - United States, 2021. MMWR Morb Mortal Wkly Rep 2021;70(44):1545-1552 [FREE Full text] [doi: 10.15585/mmwr.mm7044e2] [Medline: 34735422]
23. Oliver SE, Gargano JW, Marin M, Wallace M, Curran KG, Chamberland M, et al. The advisory committee on immunization practices' interim recommendation for use of Pfizer-BioNTech COVID-19 vaccine - United States, December 2020. MMWR Morb Mortal Wkly Rep 2020;69(50):1922-1924 [FREE Full text] [doi: 10.15585/mmwr.mm6950e2] [Medline: 33332292]
24. Bozio CH, Grannis SJ, Naleway AL, Ong TC, Butterfield KA, DeSilva MB, et al. Laboratory-confirmed COVID-19 among adults hospitalized with COVID-19-like illness with infection-induced or mRNA vaccine-induced SARS-CoV-2 Immunity - Nine States, January-September 2021. MMWR Morb Mortal Wkly Rep 2021;70(44):1539-1544 [FREE Full text] [doi: 10.15585/mmwr.mm7044e1] [Medline: 34735425]
25. Tenforde MW, Weber ZA, Natarajan K, Klein NP, Kharbanda AB, Stenehjem E, et al. Early estimates of bivalent mRNA vaccine effectiveness in preventing COVID-19-associated emergency department or urgent care encounters and hospitalizations among immunocompetent adults - VISION network, nine states, September-November 2022. MMWR Morb Mortal Wkly Rep 2022;71(5152):1616-1624 [FREE Full text] [doi: 10.15585/mmwr.mm715152e1] [Medline: 36580430]
26. Ferdinands JM, Rao S, Dixon BE, Mitchell PK, DeSilva MB, Irving SA, et al. Waning of vaccine effectiveness against moderate and severe COVID-19 among adults in the US from the VISION network: test negative, case-control study. BMJ 2022;379:e072141 [FREE Full text] [doi: 10.1136/bmj-2022-072141] [Medline: 36191948]
27. Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. Am J Epidemiol 2008;168(6):656-664 [FREE Full text] [doi: 10.1093/aje/kwn164] [Medline: 18682488]
28. Austin PC, Stuart EA. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. Stat Med 2015;34(28):3661-3679 [FREE Full text] [doi: 10.1002/sim.6607] [Medline: 26238958]
29. Wyss R, Glynn RJ, Gagne JJ. A review of disease risk scores and their application in pharmacoepidemiology. Curr Epidemiol Rep 2016;3(4):277-284. [doi: 10.1007/s40471-016-0088-2]
30. Hansen BB. The prognostic analogue of the propensity score. Biometrika 2008;95(2):481-488 [FREE Full text] [doi: 10.1093/biomet/asn004]
31. Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. J Natl Cancer Inst 1989;81(24):1879-1886. [doi: 10.1093/jnci/81.24.1879] [Medline: 2593165]
32. Wilson PWF, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. Circulation 1998;97(18):1837-1847. [doi: 10.1161/01.cir.97.18.1837] [Medline: 9603539]
33. Zimmerman JE, Kramer AA, McNair DS, Malila FM. Acute physiology and chronic health evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients. Crit Care Med 2006;34(5):1297-1310. [doi: 10.1097/01.CCM.0000215112.84523.F0] [Medline: 16540951]

XSL•FO
RenderX

34. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J Chronic Dis 1987;40(5):373-383. [doi: 10.1016/0021-9681(87)90171-8] [Medline: 3558716]
35. Andrews N, Stowe J, Kirsebom F, Toffa S, Rickeard T, Gallagher E, et al. Covid-19 vaccine effectiveness against the Omicron (B.1.1.529) variant. N Engl J Med 2022;386(16):1532-1546 [FREE Full text] [doi: 10.1056/NEJMoa2119451] [Medline: 35249272]
36. Andrews N, Stowe J, Kirsebom F, Toffa S, Sachdeva R, Gower C, et al. Effectiveness of COVID-19 booster vaccines against COVID-19-related symptoms, hospitalization and death in England. Nat Med 2022;28(4):831-837 [FREE Full text] [doi: 10.1038/s41591-022-01699-1] [Medline: 35045566]
37. Andrews N, Tessier E, Stowe J, Gower C, Kirsebom F, Simmons R, et al. Duration of protection against mild and severe disease by Covid-19 vaccines. N Engl J Med 2022;386(4):340-350 [FREE Full text] [doi: 10.1056/NEJMoa2115481] [Medline: 35021002]
38. Koehler E, Brown E, Haneuse SJA. On the assessment of Monte Carlo error in simulation-based statistical analyses. Am Stat 2009;63(2):155-162 [FREE Full text] [doi: 10.1198/tast.2009.0030] [Medline: 22544972]
39. Wyss R, Hansen BB, Ellis AR, Gagne JJ, Desai RJ, Glynn RJ, et al. The "dry-run" analysis: a method for evaluating risk scores for confounding control. Am J Epidemiol 2017;185(9):842-852 [FREE Full text] [doi: 10.1093/aje/kwx032] [Medline: 28338910]
40. Peters CC. A method of matching groups for experiment with no loss of population. J Ed Res 1941;34(8):606-612 [FREE Full text]
41. Belson WA. A technique for studying the effects of a television broadcast. Applied Statistics 1956 Nov;5(3):195-202. [doi: 10.2307/2985420]
42. Britton A, Embi PJ, Levy ME, Gaglani M, DeSilva MB, Dixon BE, et al. Effectiveness of COVID-19 mRNA vaccines against COVID-19-associated hospitalizations among immunocompromised adults during SARS-CoV-2 Omicron predominance - VISION network, 10 States, December 2021-August 2022. MMWR Morb Mortal Wkly Rep 2022;71(42):1335-1342 [FREE Full text] [doi: 10.15585/mmwr.mm7142a4] [Medline: 36264840]
43. Brookhart MA, Wyss R, Layton JB, Stürmer T. Propensity score methods for confounding control in nonexperimental research. Circ Cardiovasc Qual Outcomes 2013;6(5):604-611 [FREE Full text] [doi: 10.1161/CIRCOUTCOMES.113.000359] [Medline: 24021692]
44. Shapiro AE, Bender Ignacio RA, Whitney BM, Delaney JA, Nance RM, Bamford L, CFAR Network of Integrated Clinical Systems. Factors associated with severity of COVID-19 disease in a multicenter cohort of people With HIV in the United States, March-December 2020. J Acquir Immune Defic Syndr 2022;90(4):369-376. [doi: 10.1097/QAI.0000000000002989] [Medline: 35364600]
45. Huh K, Ji W, Kang M, Hong J, Bae GH, Lee R, et al. Association of prescribed medications with the risk of COVID-19 infection and severity among adults in South Korea. Int J Infect Dis 2021;104:7-14 [FREE Full text] [doi: 10.1016/j.ijid.2020.12.041] [Medline: 33352326]
46. R: The R Project for Statistical Computing.: R Foundation for Statistical Computing URL: https://www.R-project.org/
47. Pike MC, Anderson J, Day N. Some insights into Miettinen's multivariate confounder score approach to case-control study analysis. Epidemiol Community Health 1979;33(1):104-106 [FREE Full text] [doi: 10.1136/jech.33.1.104] [Medline: 467396]
48. Cook EF, Goldman L. Performance of tests of significance based on stratification by a multivariate confounder score or by a propensity score. J Clin Epidemiol 1989;42(4):317-324. [doi: 10.1016/0895-4356(89)90036-x] [Medline: 2723692]

Abbreviations

CDC: Centers for Disease Control and Prevention
CLI: COVID-19–like illness
DRS: disease risk score
PS: propensity score
TND: test-negative design
VE: vaccine effectiveness
VISION: Virtual SARS-CoV-2, Influenza, and Other Respiratory Viruses Network

XSL•FO
RenderX

Edited by A Mavragani; submitted 08.04.24; peer-reviewed by N Andrews; comments to author 02.10.24; revised version received 22.10.24; accepted 10.11.24; published 27.01.25.

Please cite as:
Rowley EAK, Mitchell PK, Yang DH, Lewis N, Dixon BE, Vazquez-Benitez G, Fadel WF, Essien IJ, Naleway AL, Stenehjem E, Ong TC, Gaglani M, Natarajan K, Embi P, Wiegand RE, Link-Gelles R, Tenforde MW, Fireman B
Methods to Adjust for Confounding in Test-Negative Design COVID-19 Effectiveness Studies: Simulation Study
JMIR Form Res 2025;9:e58981
URL: https://formative.jmir.org/2025/1/e58981
doi:10.2196/58981
PMID:

©Elizabeth AK Rowley, Patrick K Mitchell, Duck-Hye Yang, Ned Lewis, Brian E Dixon, Gabriela Vazquez-Benitez, William F Fadel, Inih J Essien, Allison L Naleway, Edward Stenehjem, Toan C Ong, Manjusha Gaglani, Karthik Natarajan, Peter Embi, Ryan E Wiegand, Ruth Link-Gelles, Mark W Tenforde, Bruce Fireman. Originally published in JMIR Formative Research (https://formative.jmir.org), 27.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Multimodal Pain Recognition in Postoperative Patients: Machine Learning Approach

Ajan Subramanian[1], MS; Rui Cao[2], MS; Emad Kasaeyan Naeini[1], PhD; Seyed Amir Hossein Aqajari[2], PhD; Thomas D Hughes[3], RN; Michael-David Calderon[4], MS; Kai Zheng[5], PhD; Nikil Dutt[1], PhD; Pasi Liljeberg[6], PhD; Sanna Salanterä[7,8], RN, PhD; Ariana M Nelson[9], MD; Amir M Rahmani[1,2,3,10], MBA, PhD

[1]Department of Computer Science, University of California, Irvine, Irvine, CA, United States
[2]Department of Electrical Engineering and Computer Science, University of California, Irvine, Irvine, CA, United States
[3]School of Nursing, University of California, Irvine, Irvine, CA, United States
[4]College of Medicine, Kansas City University, Kansas City, MO, United States
[5]Department of Informatics, University of California, Irvine, Irvine, CA, United States
[6]Department of Computing, University of Turku, Turku, Finland
[7]Department of Nursing Science, University of Turku, Turku, Finland
[8]Turku University Hospital, University of Turku, Turku, Finland
[9]Department of Anesthesiology and Pain Medicine, University of California, Irvine, Irvine, CA, United States
[10]Institute for Future Health, University of California, Irvine, Irvine, CA, United States

Corresponding Author:
Ajan Subramanian, MS
Department of Computer Science
University of California, Irvine
3211 Donald Bren Hall
Irvine, CA, 92617
United States
Phone: 1 6506604994
Email: ajans1@uci.edu

Abstract

Background: Acute pain management is critical in postoperative care, especially in vulnerable patient populations that may be unable to self-report pain levels effectively. Current methods of pain assessment often rely on subjective patient reports or behavioral pain observation tools, which can lead to inconsistencies in pain management. Multimodal pain assessment, integrating physiological and behavioral data, presents an opportunity to create more objective and accurate pain measurement systems. However, most previous work has focused on healthy subjects in controlled environments, with limited attention to real-world postoperative pain scenarios. This gap necessitates the development of robust, multimodal approaches capable of addressing the unique challenges associated with assessing pain in clinical settings, where factors like motion artifacts, imbalanced label distribution, and sparse data further complicate pain monitoring.

Objective: This study aimed to develop and evaluate a multimodal machine learning–based framework for the objective assessment of pain in postoperative patients in real clinical settings using biosignals such as electrocardiogram, electromyogram, electrodermal activity, and respiration rate (RR) signals.

Methods: The iHurt study was conducted on 25 postoperative patients at the University of California, Irvine Medical Center. The study captured multimodal biosignals during light physical activities, with concurrent self-reported pain levels using the Numerical Rating Scale. Data preprocessing involved noise filtering, feature extraction, and combining handcrafted and automatic features through convolutional and long-short-term memory autoencoders. Machine learning classifiers, including support vector machine, random forest, adaptive boosting, and k-nearest neighbors, were trained using weak supervision and minority oversampling to handle sparse and imbalanced pain labels. Pain levels were categorized into baseline and 3 levels of pain intensity (1-3).

Results: The multimodal pain recognition models achieved an average balanced accuracy of over 80% across the different pain levels. RR models consistently outperformed other single modalities, particularly for lower pain intensities, while facial muscle activity (electromyogram) was most effective for distinguishing higher pain intensities. Although single-modality models,

XSL•FO
RenderX

especially RR, generally provided higher performance compared to multimodal approaches, our multimodal framework still delivered results that surpassed most previous works in terms of overall accuracy.

Conclusions: This study presents a novel, multimodal machine learning framework for objective pain recognition in postoperative patients. The results highlight the potential of integrating multiple biosignal modalities for more accurate pain assessment, with particular value in real-world clinical settings.

(JMIR Form Res 2025;9:e67969) doi:10.2196/67969

KEYWORDS

pain intensity recognition; multimodal information fusion; signal processing; weak supervision; health care; pain intensity; pain recognition; machine learning approach; acute pain; pain assessment; behavioral pain; pain measurement; pain monitoring; multimodal machine learning–based framework; machine learning–based framework; electrocardiogram; electromyogram; electrodermal activity; self-reported pain level; clinical pain management

Introduction

Pain is defined by the International Association for the Study of Pain as “an unpleasant sensory and emotional experience associated with actual or potential tissue damage or described in terms of such damage” [1]. Pain is a unique phenomenon that individuals experience and perceive independently. Younger et al [2] stated that pain is a subjective experience for which there is no current objective measure. Pain may be classified as either acute or chronic; Kent et al [3] described acute pain as encompassing the immediate, time-limited bodily response to a noxious stimulus that triggers actions to avoid or mitigate ongoing injury. Chronic pain was first defined loosely by Bonica [4] as pain that extends beyond an expected timeframe; currently, chronic pain is defined as “persistent or recurrent pain lasting longer than three months” [5]. The focus of this paper is on acute pain.

Acute pain is a common experience in the postanesthesia care unit in the immediate period following surgery. According to Chou et al [6], pain occurs in 80% of patients following surgery, and 75% of patients with pain report their pain as either moderate, severe, or extreme. Current guidelines for the assessment of pain in the postanesthesia care unit recommend using a Numerical Rating Scale (NRS) or Verbal Rating Scale for patients who are sufficiently awake and coherent to reliably report pain scores [7]. However, Herr et al [8] identified several patient populations who are at risk for being incapable of providing self-report scores of pain; specifically, these populations include the pediatric population who have yet to develop adequate cognition; older patients with dementia; individuals with intellectual disabilities; and those who are unconscious, critically ill, or terminally ill. In these patient populations, Small and Laycock [7] recommend the use of behavioral pain scales, such as the Pain Assessment in Advanced Dementia, Critical Care Pain Observation Tool (CPOT), or Behavioral Pain Scale. Despite the pain assessment measures of self-report and behavioral pain scales, each of these methods may be prone to biases. For example, Craig et al [9] discussed how self-reporting might be a means to obtain a particular goal that can be influenced by the individual reporting pain. In addition, Hadjistavropoulos and Craig [10] provided the Communications Model of Pain, which provided a basis for how expressive behaviors are decoded by observers of individuals in pain, which are influenced by the message clarity transmitted by the individual in pain as well as the unique biases (eg, knowledge level, assessment skills, and predisposing beliefs) of the individual assessing pain. The difficult nature of interpreting pain scores has resulted in disparities in pain management in minority populations, with research by Staton et al [11] showing that the Black race is a significant predictor of the underestimation of pain by physicians.

Multimodal pain assessment represents a potential method of circumventing the limitations of traditional self-report and behavioral pain assessment tools and an opportunity for enhancing pain assessment in vulnerable populations. Instead of having to rely on only one dimension of pain assessment, such as behaviors through the use of the CPOT or Behavioral Pain Scale, future multimodal pain assessment will incorporate physiological indicators, such as electrodermal activity (EDA), electrocardiogram (ECG), electroencephalogram, and electromyogram (EMG) as well as behaviors (eg, facial expression), and perhaps other as-yet undiscovered parameters to capture pain assessment in patient populations that might not be best represented by current assessment strategies. For example, a study by Gélinas et al [12] found that revisions to the CPOT were necessary because some brain-injured patients may not exhibit certain behaviors that are contained in the CPOT. Similarly, for individuals diagnosed with dementia, Achterberg et al [13] stated that there is a preponderance of observer-based pain assessment tools, however, these tools retain significant differences between them, as well as concerns for lack of reliability, validity, and sensitivity of change. Enhancing pain assessment through the combination of traditional pain assessment methods with novel multimodal approaches may serve to eventually enhance pain assessment in a greater majority of vulnerable patient populations.

With the advent of connected Internet of Things devices and wearable sensor technology, automated data collection may achieve continuous pain intensity measurement. A significant amount of research has been conducted in recent years, which has sought to develop methods of continuous, automatic, and multimodal pain assessment. For example, previous work conducted by Walter et al [14] and Werner et al [15] used skin conductance level, ECG, electroencephalogram, and EMG to monitor pain in response to thermal pain. Other works, such as Hammal and Cohn [16] and Werner et al [17], have incorporated facial expression monitoring as an indicator of pain. While these studies were immensely beneficial to the scientific community

XSL•FO
RenderX

in terms of their contributions to a better understanding of techniques to obtain continuous pain assessment, the setting of these experiments was in highly controlled laboratory environments with healthy participants. Collecting data in real-world situations as opposed to a laboratory setting would allow the researchers to assess a pain assessment technique's potential in relation to actual pain brought about through a surgical procedure instead of induced pain.

The aim of this study is to develop a robust and effective multimodal pain assessment framework for postoperative patients in real clinical settings. To the best of our knowledge, this is the first work proposing a multimodal pain assessment framework for postoperative patients. It should be noted that a pain assessment study on real patients is associated with several challenges (eg, imbalanced label distribution, missing data, motion artifacts, etc) since several parameters such as the intensity, distribution, frequency, and time of the pain as well as the environment cannot be controlled by researchers. Our main contributions are 4-fold:

1. We conducted a clinical study for multimodal signal acquisition from an acute pain unit of the University of California, Irvine Medical Center.
2. We propose a multimodal pain assessment framework using our database (iHurt Pain DB) collected from postoperative patients while obtaining a higher accuracy compared to existing works on healthy participants [17].
3. We use both handcrafted and automatically generated features outputted from deep learning networks to build our models.
4. We provide a novel method to mitigate the presence of sparse and imbalanced labels (due to the real clinical setting of the study) using weak supervision and minority oversampling.

Methods

Overview

Candidates were selected from the Acute Pain Service patient list at University of California Irvine Health in Orange, California. The Acute Pain Service unit at the medical center serves approximately 100 patients weekly, enabling the lead Doctor of Medicine to recruit patients. This is the first claimed study that collected biosignals from postoperative adult patients in hospitals. All participants (aged 23-89 years) were recruited to the study from July 2018 to October 2019.

iHurt Study Design

We conducted a biomedical data collection study on 25 postoperative patients reporting various degrees of pain symptoms. Multimodal biosignals (ECG, EMG, EDA, and photoplethysmography [PPG]) were collected from patients likely having mild to moderate pain who were asked to perform a few light physical activities while acquiring data. We also collected primary demographic information from each patient, including height, weight, sex, and BMI. All signals were collected using the iHurt system.

iHurt System

iHurt is a system that measures facial muscle activity (ie, changes in facial expression) in conjunction with physiological signals such as heart rate, heart rate variability, respiration rate (RR), and EDA for the purpose of developing an algorithm for pain assessment in hospitalized patients. The system uses the following 2 components to capture raw signals.

Eight-Channel Biopotential Acquisition Device

Our team at the University of Turku, Finland, developed a biopotential acquisition device to measure ECG and EMG signals. The device incorporates commercially available electrodes, electrode-to-device lead wires, an ADS1299-based portable device, and computer software (LabVIEW version 14.02f, National Instruments) to visualize data streaming from the portable device. Raw signals from the electrodes are sampled at 500 samples per second and are sent to the computer software through Bluetooth for visualization [18].

Empatica E4

We use the commercially available Empatica E4 wristband (Empatica Inc) [19] to measure EDA and PPG signals. The purpose of using a wristband was to allow our participants to move freely without any impediments. The Empatica E4 was connected to the participants' phones over Bluetooth for visualization.

We removed 3 participants' data from the final dataset due to the presence of excessive motion artifacts. We also excluded 2 additional patients since they were wearing the Empatica E4 watch on their arm that received intravenous medication. This resulted in unreliable EDA signals due to conditions like skin rash and itching. This left us with data from 20 patients to build our pain recognition system. The dataset also contains rich annotation with self-reported pain scores based on the 11-point NRS from 0 to 10. A detailed explanation of the dataset and the study design can be found in Kasaeyan Naeini et al [20]. We intend to make the deidentified dataset available to the research community for further analysis and applications.

Data Processing Pipeline

The first step in building our multimodal pain assessment system was to process the raw signals collected during trials. The data processing pipeline consisted of the following steps:

1. We filtered the signal to remove powerline interference, baseline wander, and motion artifact noise.
2. We performed feature extraction on the filtered signals to obtain amplitude and variability features in the time domain. The time domain features were extracted using 5.5-second and 10-second windows. The 5.5-second window size was extracted to be compared with previous work [17].
3. In addition to handcrafted features, we also used automatic features, which were outputted from a deep neural network.
4. Once the features were extracted, we tagged them with their corresponding labels based on the nearest timestamp of the label.
5. Each of these processing steps was applied individually to each of the 4 modalities. Processed data from each of the modalities were combined using either early fusion or late

XSL•FO
RenderX

fusion. The types of handcrafted features extracted from each modality and the deep learning pipeline for extracting automatic features are described in detail. An overview of our method is described as a flowchart in Figure 1.

Figure 1. Overview of the proposed method. AdaBoost: adaptive boosting; EDA: electrodermal activity; EMG: electromyogram; ECG: electrocardiogram; ML: machine learning; PPG: photoplethysmography; SMOTE: synthetic minority oversampling technique; SVM: support vector machine; RF: random forest; KNN: k-nearest neighbors.

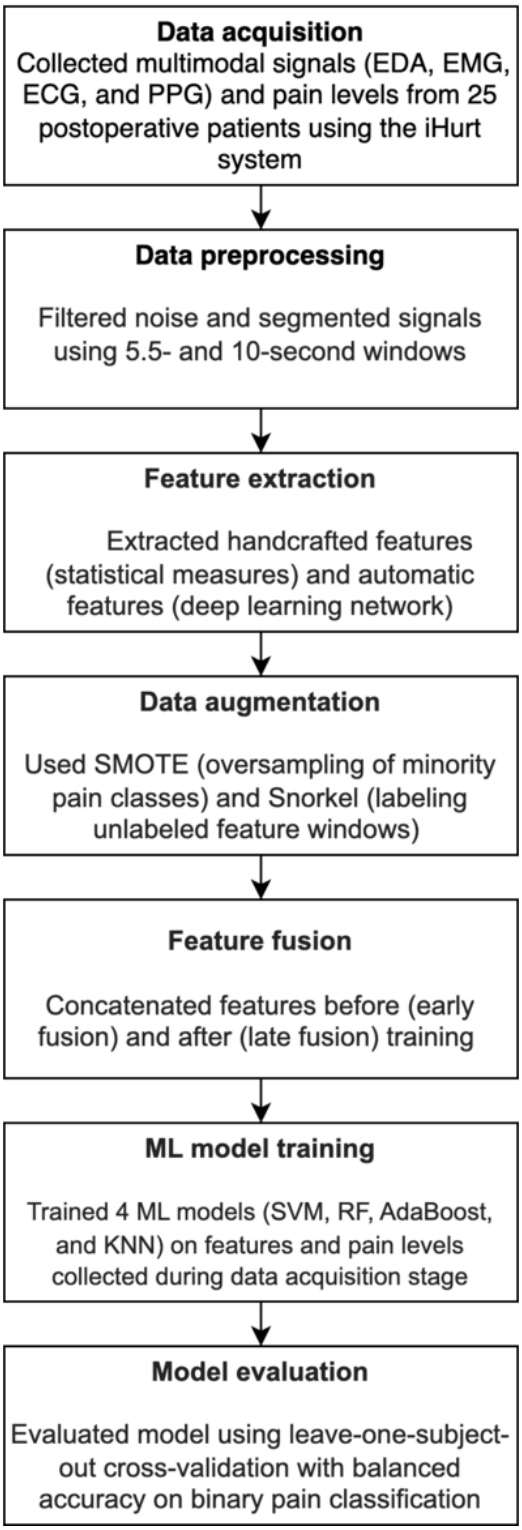

ECG Handcrafted Features

The ECG channel was filtered using a Butterworth band-pass filter with a frequency range of 0.1-250 Hz. The heart rate variability handcrafted features were extracted with *pyHRV*, an open-source Python (Python Software Foundation) toolbox [21], using the R-peaks extracted from the ECG signal through a bidirectional long short-term memory (LSTM) network [22]. These features were extracted from two window sizes, 5.5 and 10 seconds. There were 19 time-domain features. The

XSL•FO
RenderX

time-domain features extracted from NN intervals, or the time interval between successive R-peaks, comprised of the slope of these intervals, 5 statistical features (total count, mean, minimum, maximum, and SD), 9 difference features (mean difference, minimum difference, maximum difference, SD of successive interval differences, root mean square of successive interval differences, number of interval differences greater than 20 ms and 50 ms, and percentage of successive interval differences that differ by more than 20 ms and 50 ms), and 4 heart rate features (mean, minimum, maximum, and SD) [23].

EMG Handcrafted Features

The preprocessing phase of EMG channels comprised a 20 Hz high pass filter and two notch filters at 50 Hz and 100 Hz, all using a Butterworth filter. Like ECG features, we extracted EMG features from 5.5- and 10-second windows on 5 different channels for each major facial muscle. The ten amplitude features extracted were (1) peak, (2) peak-to-peak mean value, (3) root mean squared, (4) mean of the absolute values of the second differences, (5) mean of the absolute values of the first differences, (6) mean of the absolute values of the second differences of the normalized signal, (7) mean of the absolute values of the first differences of the normalized signal, (8) mean of local minima values, (9) mean of local maxima values, and (10) mean of absolute values. The four variability features were (1) variance, (2) SD, (3) range, and (4) IQR. All 14 features were calculated for 5 different EMG channels, resulting in 70 EMG features in total.

EDA Handcrafted Features

We used the *pyEDA* library [24] for preprocessing and feature extraction of EDA signals. In the preprocessing part, first, we used a moving average across a 1-second window to remove the motion artifacts and smooth the data [25]. Second, a low-pass Butterworth filter on the phasic data was applied to remove the line noise. Finally, preprocessed EDA signals corresponding to each different pain level were visualized to ensure the validity of the signals. In the feature extraction part, the *cvxEDA* algorithm [26] was used to extract the phasic component of EDA signals. The EDA signals' peaks or bursts are considered variations in the phasic component of the signal. Therefore, the clean signals and extracted phasic component of signals were fed to the statistical feature extraction module to extract the number of peaks, the average value, and the maximum and minimum value of the signals. Furthermore, these extracted features were further used in the post–feature extraction module to extract eight more features: (1) the difference between the maximum and the minimum value of the signal, (2) the SD, (3) the difference between the upper and lower quartiles (4) root mean square, (5) the mean value of local minima, (6) the mean value of local maxima, (7) the mean of the absolute values of the first differences, and (8) the mean of the absolute values of the second differences. This resulted in 12 EDA features in total.

PPG-Based RR Handcrafted Features

We preprocessed the PPG signal before extracting the RR from it. In total, 2 filters were used during the preprocessing [27]. We first used a Butterworth band-pass filter to remove noises, including motion artifacts. Then, a moving average filter was implemented to smooth the PPG signal. After that, we applied an empirical mode decomposition–based method proposed by Madhav et al [28] to derive respiration signals from filtered PPG signals. This method was proven to derive RR from a PPG signal with high accuracy (99.87%). A total of ten features were extracted from the respiratory signal, including (1) the number of inhale peaks, (2) the mean value of the signal, (3) the maximum value, (4) the minimum value, (5) the difference between the maximum and the minimum value, (6) SD, (7) the average value of the inhale peak intervals, (8) the SD of the inhale peak intervals, (9) the root mean square of successive differences between adjacent inhale peak intervals, (10) SD of inhale duration. A visualization of the handcrafted feature pipeline is shown in Figure 2.

XSL•FO
RenderX

Figure 2. Handcrafted feature extraction pipeline. dim.: dimension; ECG: electrocardiogram; EDA: electrodermal activity; EMG: electromyogram; PPG: photoplethysmography; RR: respiration rate.

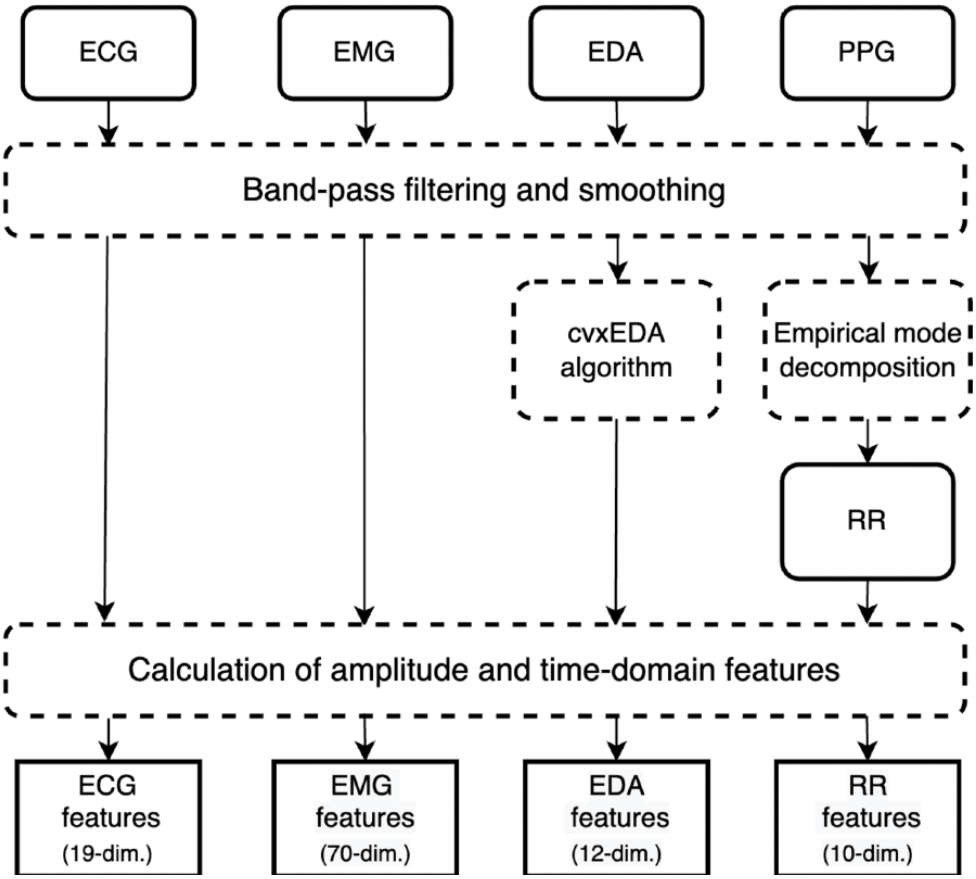

Automatic Feature Extraction Pipeline

As the dimensionality of biomedical data increases, it becomes increasingly difficult to train a machine learning algorithm on the entire uncompressed dataset. This often leads to a large training time and is computationally more expensive overall. A possible solution is to perform feature engineering to get a compressed and interpretable representation of the signal. Another alternative approach, however, is to use the compressed or latent representation of that data obtained from deep learning networks trained for that specific task. Using automatic features helps in dimensionality reduction and can provide us with a sophisticated yet succinct representation of the data that handcrafted features alone cannot provide. This automatic feature extraction is typically carried out by an autoencoder (AE) network, which is an unsupervised neural network that learns how to efficiently compress and encode the data into a lower-dimensional space [29,30]. AEs are composed of 2 separate networks: an encoder and a decoder. The encoder network acts as a bottleneck layer and maps the input into a lower-dimensional feature space. The decoder network tries to reconstruct this lower-dimensional feature vector into the original input size. The entire network is trained to minimize the reconstruction loss (ie, mean-squared error) by iteratively updating its weights and biases through backpropagation.

A convolutional AE from the *pyEDA* library was used to extract automatic features. Figure 3 shows the architecture of the AE. First, a linear layer (L1) is used to downsample the input signal with *Input_Shape* length to a length that is the closest power of 2 (CP2). This was done to make the model scalable to an arbitrary input size. The encoder half of the network consists of three 1D convolutional layers (C1, C2, and C3) and a linear layer (L2), which flattens and downsamples the input vector to a lower-dimensional latent vector. The number of dimensions of this latent vector (Feature Size) corresponds to the number of automatic features extracted and was set prior to training the network. A total of 32 features were extracted from ECG, EDA, and RR signals, whereas a total of 30 features were extracted from the EMG signal (6 features from each of the 5 channels). The decoder half of the network consists of three 1D deconvolutional layers (DeC1, DeC2, and DeC3) to reconstruct the input signal from the latent vector. A final linear layer (L3) is then used to flatten and reconstruct the signal to its original dimension. Both encoder and decoder networks have rectified linear unit activation between layers. Window sizes of both 5.5 and 10 seconds were applied to the filtered signals. This was done to compare the performance with handcrafted features. After signals from each of the modalities were normalized, they were trained on separate AE models for each modality. In addition to the convolutional AE, we also extracted features from an LSTM AE network. This resulted in two different feature extraction methods (convolutional and LSTM) that spanned two different window lengths (5.5 and 10 seconds).

XSL•FO
RenderX

Figure 3. The architecture of the pyEDA convolutional autoencoder.

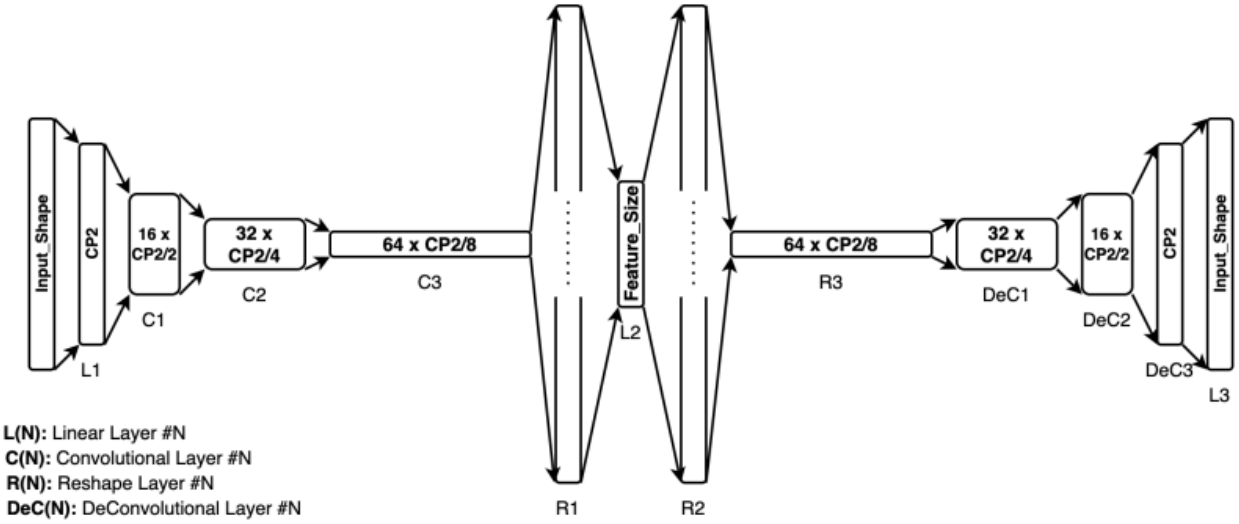

The batch size was set to 10, the number of training epochs was set to 100, and the ADAM optimizer [28] was used with a learning rate of 1×10^{-3}. A total of 126 feature vectors across all 4 modalities were extracted from each AE network. A visualization of our automatic feature extraction pipeline is shown in Figure 4.

Figure 4. Automatic feature extraction pipeline. dim.: dimension; ECG: electrocardiogram; EDA: electrodermal activity; EMG: electromyogram; RR: respiration rate; sec.: second.

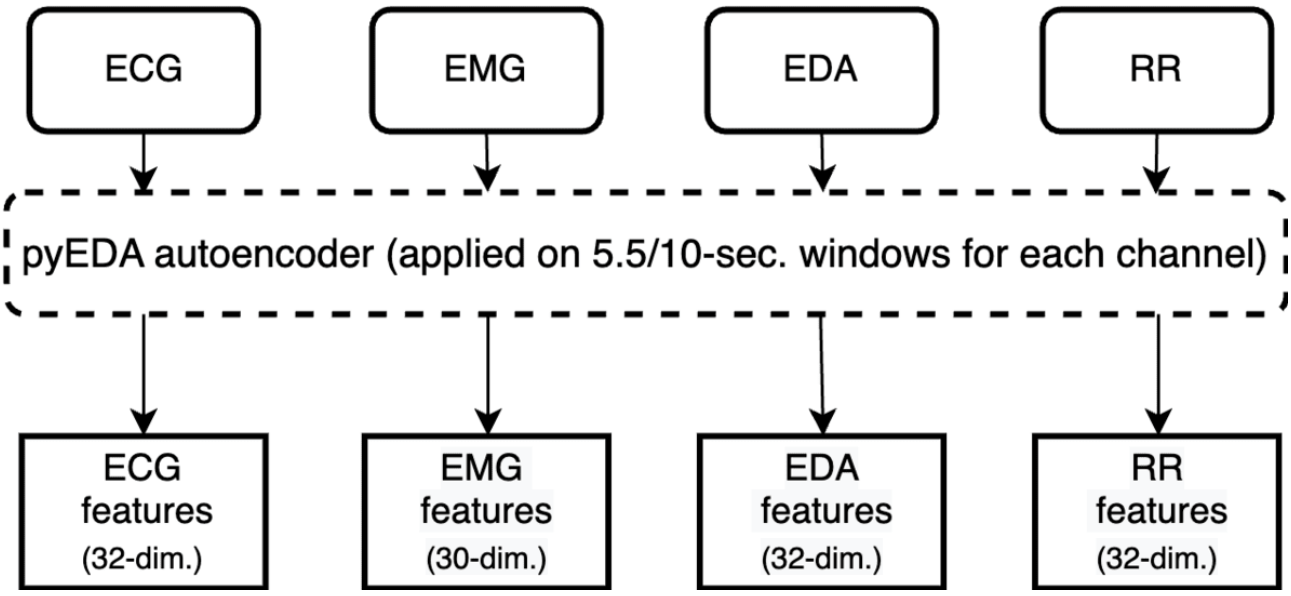

Data Augmentation

Overview

There were several inherent challenges in the distribution of labels, as NRS values recorded during the clinical trials of this study were collected from real postoperative patients. This problem bears less significance while studying healthy participants since the stimulated pain can be controlled during the experiments. Consequently, occurrences of some pain levels far exceeded those of others. For example, among all patients, there were only 4 reported occurrences of pain level 10, whereas there were more than 80 reported occurrences of pain level 4. This imbalanced distribution was inevitable due to the subjective nature and the different sources of pain among the participants. Therefore, while downsampling our pain labels to 4 classes, thresholds for each downsampled class were carefully chosen to ensure a more evenly distributed set of labels. The pain levels ranged from a baseline (BL) level of pain or no pain to 3 increasing intensities of pain (PL 1-3). The thresholds for the pain levels were as follows: (1) PL1 ranged from 0 to 3, (2) PL2 ranged from 4 to 6, and (3) PL3 ranged from 7 to 10. All the ranges here are inclusive.

Since we asked patients to report their pain levels only while they performed pain-inducing activities, the number of labels generated was sparse. Both handcrafted and automatic features were combined with the corresponding labels using timestamps that were within the nearest 5.5 or 10 seconds (labeling threshold) of the reported NRS value. This depended on the window size of the features extracted. Due to having sparse labels, many of the feature windows were not assigned a corresponding label. To mitigate the problem of having an

XSL•FO
RenderX

imbalanced and sparse label distribution, 2 techniques were exploited:

Minority Oversampling

The first technique, called synthetic minority oversampling technique (SMOTE), is a type of data augmentation that oversamples the minority class [31]. SMOTE works by first choosing a minority class instance at random and finding its k-nearest minority class neighbors. It then creates a synthetic example at a randomly selected point between two instances of the minority class in that feature space. The experiments involving SMOTE were implemented using the *imbalanced-learn* Python library [32].

Weak Supervision

The second technique we used is weak supervision using the Snorkel framework [33]. Rather than employing an expert to manually label the unlabeled instances, Snorkel allows its users to write labeling functions that can make use of heuristics, patterns, external knowledge bases, and third-party machine learning models. Weak supervision is typically employed to label large volumes of unlabeled data when there are noisy, limited, or imprecise sources. For our pain assessment algorithm, we decided to use third-party machine learning models to label the remaining unlabeled instances. All the data points that were within the labeling threshold were considered as "strong labels," or ground-truth values collected from patients during trials. The remaining unlabeled data points were kept aside for Snorkel to provide a weakly supervised label. The strong labels were fed into Snorkel's labeling function consisting of 3 off-the-shelf machine learning models: (1) a support-vector machine (SVM) with a radial basis function kernel, (2) a random forest (RF) classifier, and (3) a k-nearest neighbor (KNN) classifier with uniform weights. Once each model was trained on the strong labels, it was used to make predictions on the remaining unlabeled data. The predictions from these 3 models were collected and converted into a single confidence-weighted label per data point using Snorkel's *LabelModel* function. This function outputs the most confident prediction as the label for each data point. To perform a fair assessment of the reliability and accuracy of our algorithm, we used SMOTE and Snorkel only while training our machine learning models. The performance of these models was measured solely on ground-truth (strong) labels collected during trials. This way, there is no implicit bias introduced from mislabeling or upsampling certain data points to skew model predictions.

Multimodal Machine Learning Models

To compare the performance of our multimodal machine learning models with the previous work, we performed binary classification using a leave-one-subject-out cross-validation approach [34]. In this method, a model's performance is validated over multiple folds in such a way that data from each patient are either in the training set or in the testing set. The purpose of using this method is to provide generalizability to unseen patients and to avoid overfitting by averaging the results over multiple folds. The eventual goal of this study is to build personalized models that make predictions on a single patient but learn from data collected from a larger population of similar patients. The following machine learning models were used to evaluate the performance of our pain assessment algorithm: (1) KNN, (2) RF classifier, (3) adaptive boosting (AdaBoost), (4) and an SVM. The models were then evaluated using leave-one subject-out cross-validation. Four separate models were trained for each of the 3 pain intensities (eg, BL; no pain versus PL1, the lowest pain level; or BL vs PL3, the highest pain level).

Fusing Modalities

In total, 2 fusion approaches were used while combining features across different modalities. The first one is early or feature-level fusion, which concatenates feature vectors across different modalities based on their time stamps. The resulting data, which are now higher in dimension than any single modality, are then fed into our classifier to make predictions. While concatenating features across different modalities, a threshold of either 5.5 or 10 seconds was used to combine the modalities depending on the features extracted. The second approach was late or decision-level fusion, where each modality is fed to a separate classifier, and the final classification result is based on the fusion of outputs from the different modalities [35].

Feature Selection

Since there were a lot of features generated during the data processing phase, we had to select a subset of the most informative features to build our models with. Therefore, to reduce the complexity and training time of the resulting model, feature selection using Gini importance was performed. Gini is important as a lightweight method that is simple and fast to compute. Since we extracted a relatively large number of features in our method, it made sense to use a computationally low-cost algorithm for feature selection. We computed the Gini importance of the features from the data in the training fold with the help of a random forest classifier and selected the top 25 features. We then trained our model on these top 25 features and evaluated them in the validation fold. Our proposed multimodal pain recognition system is shown in Figure 5.

XSL•FO
RenderX

Figure 5. Proposed multimodal pain recognition system. BL: baseline; CNN: convolutional neural network; ECG: electrocardiogram; EDA: electrodermal activity; EMG: electromyogram; PL: pain level; RR: respiration rate; SMOTE: synthetic minority oversampling technique.

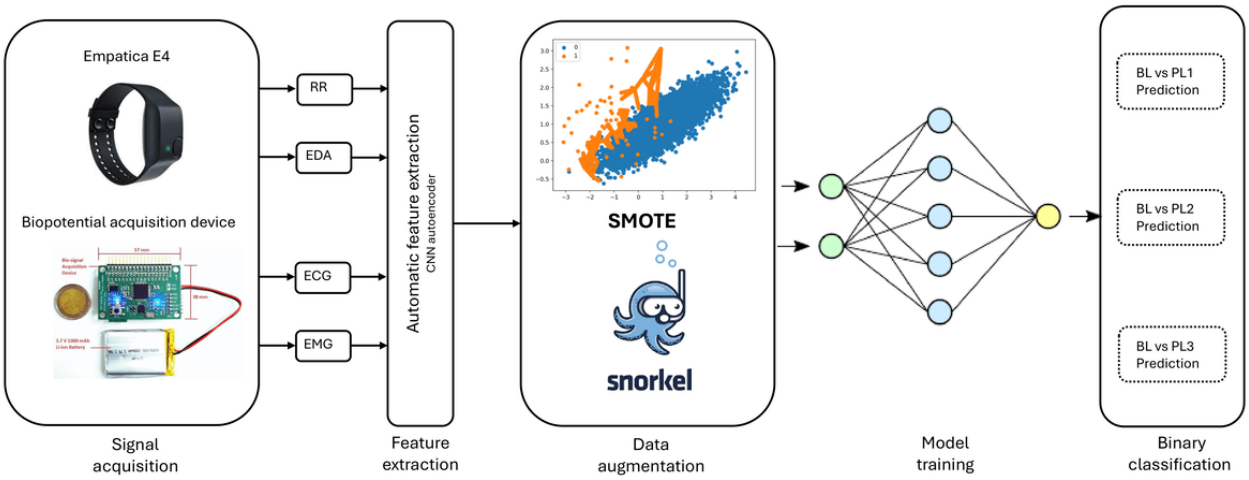

Ethical Considerations

The dataset used in this study was originally collected with approval from the Institutional Review Board (IRB) at the University of California, Irvine (Protocol HS# 2017-3747). Participants provided written informed consent after receiving detailed oral and written explanations of the study's objectives and procedures. They were encouraged to discuss participation with family and friends before consenting. Investigators ensured that all participants understood the study and had their questions answered prior to enrollment. Participants were informed of their right to withdraw at any time without impacting their care. For the secondary analysis conducted in this study, the IRB approval and original informed consent covered the reuse of the data, and no additional consent was required. All data utilized for this study were anonymized prior to analysis to protect participants' identities. Personal identifiers, such as names and contact information, were removed, and access to the data was restricted to authorized personnel only. The anonymized data were stored securely in compliance with institutional and regulatory guidelines to ensure confidentiality. Participation in the original study was entirely voluntary, and no compensation was provided. This ensured that participants' involvement was based solely on their willingness to contribute to the research.

Results

Experimental Settings

The goal of our experiments was to compare the performance of using only a single modality to build our models over using a combination of multiple modalities. We trained several different models for each of the pain intensities, which varied in the types of modalities, data augmentation techniques, machine learning models, and fusion techniques used. Figure 6 shows the general pipeline of the experiments we conducted. We first selected the type of modalities to train on, which varied from only using each of the single modalities separately to using a combination of all 4 modalities. Furthermore, these modalities varied depending on the type of features used, like handcrafted or automatic features. In the case of using multiple modalities, we had 2 choices of fusion: early (Figure 6, left) and late (Figure 6, right). These architectures varied in how the modalities were combined, either before training (early) or at the decision level (late) after training using majority voting. The data preparation process involved feature selection and data augmentation. These models could either be trained with no data augmentation, with just SMOTE or Snorkel, or a combination of both. The last step of the pipeline before making predictions involved choosing the type of machine learning algorithms, like SVM, RF, AdaBoost, or KNN. Due to the lack of space, only the best-performing single and multimodal model configurations are mentioned in the section below.

XSL•FO
RenderX

Figure 6. Our proposed general multimodal pipeline-based on early fusion (left) and late fusion (right). AdaBoost: adaptive boosting; KNN: k-nearest neighbors; ML: machine learning; RF: random forest; SVM: support vector machine; SMOTE: synthetic minority oversampling technique.

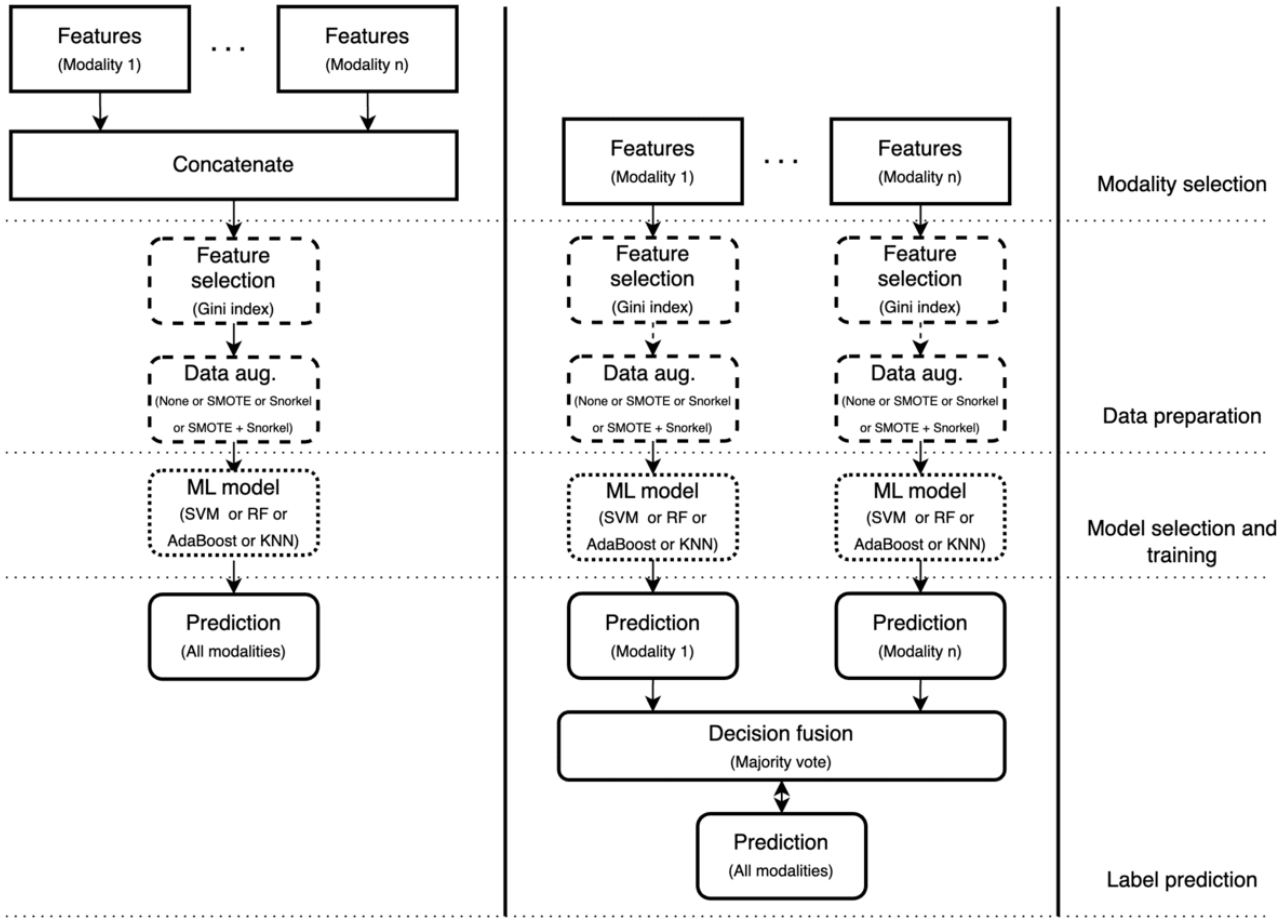

Experimental Results

Tables 1 and 2 present the best-performing single-modal and multimodal models for each of the 3 pain intensities. For comparison, the best multimodal results from Werner et al [17], Lopez-Martinez and Picard [36], Wang et al [37], and Subramaniam and Dass [38] are also mentioned. We use balanced accuracy as an evaluation criterion because our dataset had an imbalanced class distribution. Balanced accuracy is defined as the average of the true positive rate and the true negative rate.

XSL•FO
RenderX

Table 1. Best scores: single modality versus multiple modalities.

Pain levels	ECG[a] scores	EMG[b] scores	EDA[c] scores	RR[d] scores	Multiple modality
BL[e] vs PL[f]1	82.14	86	79.18	84.62	82.14
BL vs PL2	86.11	84.53	82.94	88.24	86.11
BL vs PL3	75	78.12	75	76.23	75
Mean (SD)	81.08 (5.03)	82.8 (5.03)	79.04 (5.03)	83.03 (5.03)	81.08 (4.59)
Classifier configuration	LSTM[g] AE[h] (10 s), Strong, SVM[j]	HC[i] (10 s), Snorkel, SVM	CNN[k] AE (10 s), Strong, SVM	HC (10 s), Strong, SVM	EF[l], LSTM AE (10 s), Strong, SVM

[a]ECG: electrocardiogram.
[b]EMG: electromyogram.
[c]EDA: electrodermal activity.
[d]RR: respiration rate.
[e]BL: baseline.
[f]PL: pain level.
[g]LSTM: long short-term memory.
[h]AE: autoencoder.
[i]HC: handcrafted.
[j]SVM: support vector machine.
[k]CNN: convoluted neural network.
[l]EF: early fusion.

Table 2. Multiple modalities: comparison with other methods.

Study	Value, mean (SD)	Modalities
Werner et al [17]	65.02 (8.72)	Video, ECG[a], EMG[b], and EDA[c]
Lopez-Martinez and Picard [36]	66.68 (10.87)	ECG and EDA
Wang et al [37]	70.4 (9.76)	ECG, EMG, and EDA
Subramaniam and Dass [38]	92.604 (3.49)	ECG and EDA
Our method	81.08 (4.59)	EDA, EMG, EDA, and RR[d]

[a]ECG: electrocardiogram.
[b]EMG: electromyogram.
[c]EDA: electrodermal activity.
[d]RR: respiration rate.

Discussion

Principal Findings

This study demonstrated that RR emerged as the strongest single-modality predictor of pain intensity, particularly for distinguishing between baseline and lower pain levels. EMG performed best for higher pain intensities, while EDA and ECG showed comparatively lower effectiveness as stand-alone modalities. Multimodal models, though offering potential advantages in robustness and complementary information, generally underperformed compared with the RR single-modality models, likely due to challenges related to noise and data alignment. The study highlights the importance of modality selection and data fusion strategies for pain recognition in postoperative settings.

Performance by Modality

Pain Recognition Using RR Alone

From the single-modality results (Table 1), it is evident that RR models outperform all other modalities, especially for the BL versus PL1 and BL versus PL2 models. This highlights the strong predictive power of RR in distinguishing between baseline and lower pain intensities. The best-performing model used RR alone. One justification for these results could be the dynamic nature of RR signals in response to pain stimuli. Since we effectively isolated and captured periods of higher pain intensity with smaller window sizes, this could have helped the models better distinguish between baseline and other pain levels.

Pain Recognition Using EMG Alone

For the highest pain category (BL vs PL3), the EMG model outperformed other single-modality models. This suggests that facial muscle activation captured by EMG signals is particularly effective for distinguishing higher pain intensities. The

XSL•FO
RenderX

comparatively lower performance of other modalities, such as EDA, could be attributed to the subtle variations in their responses to different pain levels.

Pain Recognition Using EDA Alone

EDA models exhibited comparatively lower performance across all pain categories. This may be due to the difficulty in capturing clear variations in EDA signal responses to different pain levels, as observed in our experiments.

Pain Recognition Using ECG Alone

While ECG features contributed strongly to the performance of multimodal models, their single-modality results were not as robust as those of RR or EMG. However, the best-performing multimodal models shared identical configurations with the best ECG models, suggesting that ECG features had a significant influence on the multimodal results.

Challenges With Extremes in Pain Levels

The BL versus PL1 and BL versus PL3 models had relatively poor performances across both single and multimodal approaches. BL versus PL1 struggled to distinguish the baseline from the lowest pain intensity due to the subtlety of the physiological responses collected while experiencing this pain level. The BL versus PL3 model, however, found it challenging to distinguish pain levels due to the scarcity of labels for the highest pain intensity. Although data augmentation can help mitigate these challenges, there is no substitute for real data. On the contrary, the BL versus PL2 models performed better due to the relative abundance of such labels reported during trials.

Multimodal Performance

The best-performing multimodal model was trained on automatic features outputted from an LSTM network with a 10-second window size. This model, which made use of strong labels without any data augmentation techniques, achieved comparable results to the best-performing ECG single-modality model. Early fusion outperformed late fusion, likely due to its ability to detect correlations across modalities during feature selection [39]. By treating each modality as independent, late fusion might lose correlations in the combined feature space.

However, single-modality models, particularly RR, generally outperformed multimodal models. This contrasts with previous studies on healthy participants, where multimodal approaches typically excelled. Our findings suggest that the unique challenges of real-world postoperative data, including noise and missing signals, may complicate the integration of multiple modalities.

Advantages and Trade-Offs

While multimodal models have the potential to add complementary information and robustness, they also introduce challenges related to data alignment and noise management. Single-modality models, by contrast, are simpler, easier to interpret, and computationally less expensive. These advantages make single modalities, such as RR and EMG, attractive for certain applications despite the overall potential of multimodal approaches. Multiple modalities certainly have the potential to add more useful information over a single modality and can be used to introduce complementary information and resiliency when any one modality fails or is too noisy [40].

While comparing our results to previous studies [17,36-38] in Table 2, it can be observed that our models outperform most of their models in mean pain assessment scores except Subramaniam and Dass [38]. However, this is not entirely a fair comparison because we use 3 pain levels instead of 4, and our patients are not healthy.

An additional consideration is the comfort and compliance of patients wearing multiple biosensors, especially in postoperative settings. While multimodal models rely on multiple sources of data, this could pose a burden to patients who may already be experiencing discomfort. Future iterations of the framework could focus on optimizing the number of biosensors by identifying the most informative modalities. This optimization could improve patient compliance while maintaining the accuracy and robustness of the system.

Limitations

The main limitation of our algorithm is the presence of noise in the form of motion artifacts produced while collecting physiological signals. Since we obtained data from real postoperative patients in a clinical setting, they were allowed to move more freely compared to experiments performed in controlled laboratory settings. The presence of these motion artifacts diminished the quality of our data, thus negatively impacting our machine learning algorithms.

In addition, our study was conducted in a setting with a limited and relatively homogeneous patient population. While this setting allowed us to focus on developing and testing the algorithm, it restricts the generalizability of our findings to broader and more diverse clinical environments. Testing the model in varied clinical settings and across a larger, more diverse patient population is essential for evaluating its scalability and effectiveness in real-world scenarios. This remains an important future research direction.

Furthermore, we must acknowledge the more complicated facets of pain that are not fully captured by our algorithm, such as the number of days after surgery, the amount of pain medication administered, and the location and type of pain experienced. Incorporating these factors in future studies could improve the accuracy and robustness of pain assessment systems.

Future Directions

One of the main research directions we would like to explore is the development of real-time multimodal pain assessment systems using deep learning architectures. In such scenarios, missing or incomplete data from one or more modalities are likely to be encountered. Real-time systems also face limitations related to computational complexity and power constraints. Building on the experiments conducted in this study, we aim to create models capable of dynamically determining which modalities to use in an energy-efficient manner without compromising performance given the clinical context.

In addition, a promising avenue for future work is to build personalized machine learning models. These models could

XSL•FO
RenderX

leverage data from groups of similar patients while being fine-tuned to make predictions for individual patients. This personalized approach accounts for the large interindividual variability in pain perception, which makes a monolithic model unsuitable. Previous research has demonstrated the feasibility of using multitask machine learning to address variability in mood prediction tasks [41]. This strategy could be extended to the domain of pain assessment, not only for acute postoperative pain but also for chronic pain scenarios. Personalized modeling will be a vital step toward creating clinically viable and effective pain assessment algorithms.

Conclusions

In this paper, we presented a multimodal machine learning framework for classifying pain in real postoperative patients using the iHurt Pain Database. Both traditional handcrafted features and deep learning–generated automatic features were extracted from physiological signals (ECG, EDA, EMG, and PPG). Several experiments were conducted to perform binary classification among 3 different pain intensities versus baseline levels of pain. Models were varied based on the modalities used, the data augmentation techniques applied (SMOTE, Snorkel, or both), the machine learning algorithms used, and the modality fusion methods implemented.

Our results showed that binary pain classification significantly benefits from the application of data augmentation techniques in conjunction with automatic features. The single-modality models based on RR and EMG outperformed the multimodal models. The BL versus PL3 model with the best results was trained on EMG data alone, highlighting the importance of facial muscle activation in distinguishing higher pain intensities from baseline levels. This finding is consistent from a clinical perspective, as higher pain intensities are commonly associated with acute pain.

Overall, this study highlights a novel approach to addressing the challenges of building a pain recognition system for real postoperative patients, particularly constraints such as label imbalances and missing data. By employing robust data preprocessing techniques, data augmentation strategies, and multimodal fusion approaches, our framework demonstrates the potential for accurate and objective pain classification in clinical settings. These findings lay the groundwork for advancing multimodal pain assessment methods tailored to real-world clinical scenarios.

Conflicts of Interest

None declared.

References

1. Merskey HA. Pain terms: a list with definitions and notes on usage. Recommended by the IASP subcommittee on taxonomy. Pain 1979;6(3):249. [Medline: 460932]
2. Younger J, McCue R, Mackey S. Pain outcomes: a brief review of instruments and techniques. Curr Pain Headache Rep 2009;13(1):39-43 [FREE Full text] [doi: 10.1007/s11916-009-0009-x] [Medline: 19126370]
3. Kent ML, Tighe PJ, Belfer I, Brennan TJ, Bruehl S, Brummett CM, et al. The ACTTION-APS-AAPM pain taxonomy (AAAPT) multidimensional approach to classifying acute pain conditions. Pain Med 2017;18(5):947-958 [FREE Full text] [doi: 10.1093/pm/pnx019] [Medline: 28482098]
4. Bonica JJ. Management of cancer pain. In: Pain in the Cancer Patient: Pathogenesis, Diagnosis and Therapy. Berlin, Germany: Springer; 1984.
5. Treede RD, Rief W, Barke A, Aziz Q, Bennett MI, Benoliel R, et al. A classification of chronic pain for ICD-11. Pain 2015;156(6):1003-1007 [FREE Full text] [doi: 10.1097/j.pain.0000000000000160] [Medline: 25844555]
6. Chou R, Gordon DB, de Leon-Casasola OA, Rosenberg JM, Bickler S, Brennan T, et al. Management of postoperative pain: a clinical practice guideline from the American pain society, the American society of regional anesthesia and pain medicine, and the American society of anesthesiologists' committee on regional anesthesia, executive committee, and administrative council. J Pain 2016;17(2):131-157 [FREE Full text] [doi: 10.1016/j.jpain.2015.12.008] [Medline: 26827847]
7. Small C, Laycock H. Acute postoperative pain management. Br J Surg 2020;107(2):e70-e80. [doi: 10.1002/bjs.11477] [Medline: 31903595]
8. Herr K, Coyne PJ, Ely E, Gélinas C, Manworren RCB. Pain assessment in the patient unable to self-report: clinical practice recommendations in support of the ASPMN 2019 position statement. Pain Manag Nurs 2019;20(5):404-417. [doi: 10.1016/j.pmn.2019.07.005] [Medline: 31610992]
9. Craig K. The facial expression of pain Better than a thousand words? APS Journal 1992 Sep;1(3):153-162. [doi: 10.1016/1058-9139(92)90001-S]
10. Hadjistavropoulos T, Craig KD. A theoretical framework for understanding self-report and observational measures of pain: a communications model. Behav Res Ther 2002;40(5):551-570. [doi: 10.1016/s0005-7967(01)00072-9] [Medline: 12038648]
11. Staton LJ, Panda M, Chen I, Genao I, Kurz J, Pasanen M, et al. When race matters: disagreement in pain perception between patients and their physicians in primary care. J Natl Med Assoc 2007;99(5):532-538. [Medline: 17534011]
12. Gélinas C, Boitor M, Puntillo KA, Arbour C, Topolovec-Vranic J, Cusimano MD, et al. Behaviors indicative of pain in brain-injured adult patients with different levels of consciousness in the intensive care unit. J Pain Symptom Manage 2019;57(4):761-773 [FREE Full text] [doi: 10.1016/j.jpainsymman.2018.12.333] [Medline: 30593909]

XSL•FO
RenderX

13. Achterberg W, Lautenbacher S, Husebo B, Erdal A, Herr K. Pain in dementia. Schmerz 2021;35(2):130-138. [doi: 10.1007/s00482-020-00501-w] [Medline: 32926239]
14. Walter S, Gruss S, Ehleiter H, Tan J, Traue HC, Werner P, et al. The biovid heat pain database data for the advancement and systematic validation of an automated pain recognition system. 2013 Presented at: Proceedings of the IEEE International Conference on Cybernetics (CYBCO); Jun 13-15, 2013; Lausanne, Switzerland p. 128-131. [doi: 10.1109/CYBConf.2013.6617456]
15. Werner P, Al-Hamadi A, Niese R, Walter S, Gruss S, Traue HC. Towards pain monitoring: Facial expression, head pose, a new database, an automatic system and remaining challenges. 2013 Presented at: Proceedings of the British Machine Vision Conference; September 01, 2013; London, UK p. 1-13 URL: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=03f075e95638bc66e687badd97a58c5de67e58e6 [doi: 10.5244/C.27.119]
16. Hammal Z, Cohn JF. Automatic detection of pain intensity. Proc ACM Int Conf Multimodal Interact 2012;2012:47-52 [FREE Full text] [doi: 10.1145/2388676.2388688] [Medline: 32724903]
17. Werner P, Al-Hamadi A, Niese R, Walter S, Gruss S, Traue H. Automatic pain recognition from video and biomedical signals. 2014 Presented at: Proceedings of the 22nd International Conference on Pattern Recognition; August 24-28, 2014; Stockholm, Sweden p. 4582-4587. [doi: 10.1109/ICPR.2014.784]
18. Sarker VK, Jiang M, Gia TN, Anzanpour A, Rahmani A, Liljeberg P. Portable multipurpose bio-signal acquisition and wireless streaming device for wearables. 2017 Presented at: Proceedings of the IEEE Sensors Applications Symposium (SAS); March 13, 2017; New York City, NY p. 1-6. [doi: 10.1109/SAS.2017.7894053]
19. Empatica E4 User Manual. 2015. URL: https://www.manualslib.com/manual/1318955/Empatica-E4.html [accessed 2024-12-13]
20. Kasaeyan Naeini E, Jiang M, Syrjälä E, Calderon MD, Mieronkoski R, Zheng K, et al. Prospective study evaluating a pain assessment tool in a postoperative environment: protocol for algorithm testing and enhancement. JMIR Res Protoc 2020;9(7):e17783 [FREE Full text] [doi: 10.2196/17783] [Medline: 32609091]
21. Gomes PM, Margaritoff P, Silva H. pyHRV: development and evaluation of an open-source Python toolbox for heart rate variability (HRV). 2019 Presented at: Proceedings of the International Conference on Electrical, Electronic and Computing Engineering (IcETRAN); June 01, 2019; Serbia URL: https://www.researchgate.net/publication/333611305_pyHRV_Development_and_Evaluation_of_an_Open-Source_Python_Toolbox_for_Heart_Rate_Variability_HRV
22. Laitala J, Jiang M, Syrjälä E, Naeini E, Airola A, Rahmani A, et al. Robust ECG R-peak detection using LSTM. 2020 Presented at: Proceedings of the 35th Annual ACM Symposium on Applied Computing; March 30, 2020; Brno, Czech Republic p. 1104-1111. [doi: 10.1145/3341105.3373945]
23. Kasaeyan Naeini E, Subramanian A, Calderon MD, Zheng K, Dutt N, Liljeberg P, et al. Pain recognition with electrocardiographic features in postoperative patients: method validation study. J Med Internet Res 2021;23(5):e25079 [FREE Full text] [doi: 10.2196/25079] [Medline: 34047710]
24. Aqajari SAH, Naeini EK, Mehrabadi MA, Labbaf S, Dutt N, Rahmani AM. pyEDA: an open-source Python toolkit for pre-processing and feature extraction of electrodermal activity. Procedia Comput. Sci 2021;184:99-106. [doi: 10.1016/j.procs.2021.03.021]
25. Aqajari SAH, Cao R, Kasaeyan Naeini E, Calderon M, Zheng K, Dutt N, et al. Pain assessment tool with electrodermal activity for postoperative patients: method validation study. JMIR Mhealth Uhealth 2021;9(5):e25258 [FREE Full text] [doi: 10.2196/25258] [Medline: 33949957]
26. Greco A, Valenza G, Lanata A, Scilingo EP, Citi L. cvxEDA: a convex optimization approach to electrodermal activity processing. IEEE Trans. Biomed. Eng 2016;63(4):797-804. [doi: 10.1109/tbme.2015.2474131]
27. Cao R, Aqajari SAH, Kasaeyan Naeini E, Rahmani AM. Objective pain assessment using wrist-based PPG signals: a respiratory rate based method. Annu Int Conf IEEE Eng Med Biol Soc 2021;2021:1164-1167. [doi: 10.1109/EMBC46164.2021.9630002] [Medline: 34891494]
28. Madhav K, Ram M, Krishna E, Komalla N, Reddy K. Estimation of respiration rate from ECG, BPPPG signals using empirical mode decomposition. 2011 Presented at: Proceedings of the IEEE International Instrumentation and Measurement Technology Conference; May 10-12, 2011; Hangzhou, China. [doi: 10.1109/imtc.2011.5944249]
29. Schmidhuber J. Deep learning in neural networks: an overview. Neural Netw 2015;61:85-117. [doi: 10.1016/j.neunet.2014.09.003] [Medline: 25462637]
30. Le QV. A tutorial on deep learning part 2: Autoencoders, convolutional neural networks and recurrent neural networks. Google Brain 2015:1-20 [FREE Full text]
31. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. jair 2002 Jun 01;16(1):321-357. [doi: 10.1613/jair.953]
32. Lemaître G, Nogueira F, Aridas C. Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. J Mach Learn Res 2017;18(1):559-563 [FREE Full text]
33. Ratner A, Bach SH, Ehrenberg H, Fries J, Wu S, Ré C. Snorkel: rapid training data creation with weak supervision. Proceedings VLDB Endowment 2017;11(3):269-282 [FREE Full text] [doi: 10.14778/3157794.3157797] [Medline: 29770249]

XSL•FO
RenderX

34. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. Proc. 14th IJCAI 1995:1137-1143 [FREE Full text]
35. Gunes H, Piccardi M. Affect recognition from face and body: early fusion vs. late fusion. 2005 Presented at: Proceedings of the IEEE International Conference on Systems, Man and Cybernetics; October 12, 2005; Waikoloa, HI, USA p. 3437-3443. [doi: 10.1109/ICSMC.2005.1571679]
36. Lopez-Martinez D, Picard R. Multi-task neural networks for personalized pain recognition from physiological signals. 2017 Presented at: Proceedings of the Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW); October 23-26, 2017; San Antonio, TX p. 181-184. [doi: 10.48550/arXiv.1708.08755]
37. Wang R, Xu K, Feng H, Chen W. Hybrid RNN-ANN based deep physiological network for pain recognition. 2020 Presented at: Proceedings of the 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); July 20-24, 2020; Montreal, QC p. 5584-5587. [doi: 10.1109/EMBC44109.2020.9175247]
38. Subramaniam SD, Dass B. Automated nociceptive pain assessment using physiological signals and a hybrid deep learning network. IEEE Sensors J 2021;21(3):3335-3343. [doi: 10.1109/jsen.2020.3023656]
39. Ross A. In: Li SZ, editor. Fusion, Feature-Level. Boston, MA: Springer; 2009.
40. Naeini EK, Shahhosseini S, Kanduri A, Liljeberg P, Rahmani A, Dutt N. AMSER: adaptive multimodal sensing for energy efficient and resilient eHealth systems. 2022 Presented at: Proceedings of the Design, Automation & Test in Europe Conference & Exhibition (DATE); March 14-23, 2022; Antwerp, Belgium p. 1455-1460. [doi: 10.48550/arXiv.2112.08176]
41. Taylor S, Jaques N, Nosakhare E, Sano A, Picard R. Personalized multitask learning for predicting tomorrow's mood, stress, and health. IEEE Trans Affect Comput 2020;11(2):200-213 [FREE Full text] [doi: 10.1109/TAFFC.2017.2784832] [Medline: 32489521]

Abbreviations

AdaBoost: adaptive boosting
AE: autoencoder
CNN: convolutional neural network
CPOT: Critical Care Pain Observation Tool
ECG: electrocardiogram
EDA: electrodermal activity
EMG: electromyogram
KNN: k-nearest neighbors
LSTM: long short-term memory
NRS: Numerical Rating Scale
PPG: photoplethysmography
RF: random forest
RR: respiration rate
SMOTE: synthetic minority oversampling technique
SVM: support vector machine

Edited by A Mavragani; submitted 24.10.24; peer-reviewed by G Chennekkattu Markose, X Gao; comments to author 25.11.24; revised version received 28.11.24; accepted 29.11.24; published 27.01.25.

Please cite as:
Subramanian A, Cao R, Naeini EK, Aqajari SAH, Hughes TD, Calderon MD, Zheng K, Dutt N, Liljeberg P, Salanterä S, Nelson AM, Rahmani AM
Multimodal Pain Recognition in Postoperative Patients: Machine Learning Approach
JMIR Form Res 2025;9:e67969
URL: https://formative.jmir.org/2025/1/e67969
doi:10.2196/67969
PMID:

©Ajan Subramanian, Rui Cao, Emad Kasaeyan Naeini, Seyed Amir Hossein Aqajari, Thomas D Hughes, Michael-David Calderon, Kai Zheng, Nikil Dutt, Pasi Liljeberg, Sanna Salanterä, Ariana M Nelson, Amir M Rahmani. Originally published in JMIR Formative Research (https://formative.jmir.org), 27.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The

XSL•FO
RenderX

complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Evaluating Reaction Videos of Young People Watching Edutainment Media (MTV Shuga): Qualitative Observational Study

Venetia Baker[1], MSc; Sarah Mulwa[1], MSc; David Khanyile[2], BA; Georgia Arnold[3], BA; Simon Cousens[1], PhD; Cherie Cawood[2], PGDip; Isolde Birdthistle[1], PhD

[1]Faculty of Epidemiology and Population Health, The London School of Hygiene and Tropical Medicine, London, United Kingdom

[2]Epicentre Health Research, Durban, South Africa

[3]GA-Agency, London, United Kingdom

Corresponding Author:
Venetia Baker, MSc
Faculty of Epidemiology and Population Health
The London School of Hygiene and Tropical Medicine
Keppel St
London, WC1E 7HT
United Kingdom
Phone: 44 020 7636 8636
Email: venetia.baker1@lshtm.ac.uk

Abstract

Background: Mass media campaigns, particularly edutainment, are critical in disseminating sexual health information to young people. However, there is limited understanding of the authentic viewing experience or how viewing contexts influence engagement with media campaigns. Reaction videos, a popular format in web-based culture in which users film themselves reacting to television shows, can be adapted as a research method for immediate and unfiltered insights into young people's engagement with edutainment media.

Objective: We explored how physical and social context influences young people's engagement with *MTV Shuga*, a dramatic television series based on sexual health and relationships among individuals aged 15 to 25 years. We trialed reaction videos as a novel research method to investigate how young people in South Africa experience the show, including sexual health themes and messages, in their viewing environments.

Methods: In Eastern Cape, in 2020, purposively selected participants aged 18 to 24 years of an evaluation study were invited to take part in further research to video record themselves watching *MTV Shuga* episodes with their *COVID-19 social bubble*. To guide the analysis of the visual and audio data, we created a framework to examine the physical setting, group composition, social dynamics, coinciding activities, and viewers' spoken and unspoken reactions to the show. We identified patterns within and across groups to generate themes about the nature and role of viewing contexts. We also reflected on the utility of the method and analytical framework.

Results: In total, 8 participants recorded themselves watching *MTV Shuga* episodes in family or friendship groups. Viewings occurred around a laptop in the home (living room or bedroom) and outside (garden or vehicle). In same-age groups, viewers appeared relaxed, engaging with the content through discussion, comments, empathy, and laughter. Intergenerational groups experienced discomfort, with older relatives' presence causing embarrassment and younger siblings' distractions interrupting the engagement. Scenes featuring physical intimacy prompted some viewers to hide their eyes or leave the room. While some would prefer watching *MTV Shuga* alone to avoid the self-consciousness experienced in group settings, others valued the social experience and the lively discussions it spurred. This illustrates varied preferences for consuming edutainment and the factors influencing these preferences.

Conclusions: The use of reaction videos for research captured real-time verbal and nonverbal reactions, physical environments, and social dynamics that other methods cannot easily measure. They revealed how group composition, dynamics, settings, and storylines can maximize engagement with *MTV Shuga* to enhance HIV prevention education. The presence of parents and the camera may alter young people's behavior, limiting the authenticity of their viewing experience. Still, reaction videos offer a unique opportunity to understand audience engagement with media interventions and promote participatory digital research with young people.

XSL•FO
RenderX

(JMIR Form Res 2025;9:e55275) doi:10.2196/55275

KEYWORDS

mass media; edutainment; adolescents; sexual health; HIV prevention; participatory research

Introduction

Background

Edutainment has become a crucial tool for disseminating sexual health information, especially to young people [1-5]. Singhal and Rogers [6] define *edutainment* as the deliberate use of media to entertain and educate, aiming to enhance knowledge, foster positive attitudes, and change behavior. This approach has traditionally used various media formats, including radio, television, comic books, and music, to influence audience awareness, attitudes, and behaviors in their external environment to help create the necessary conditions for social change [6]. The Sabido methodology is an approach to edutainment using soap operas to promote positive social change by integrating educational messages into compelling narratives. It relies on social learning theory, which posits that people learn behaviors by observing others, especially when they see those behaviors rewarded [6]. The Sabido methodology uses this theory to create positive, negative, and transitional role models that viewers can relate to, aiming to shift behaviors by showing the consequences of these characters' actions [6]. Audience engagement is key as the method emotionally invests viewers in the characters' journeys, encouraging reflection and discussion. The discussions that edutainment generates are central to its ability to diffuse new ideas into the community, shifting not just individual behavior but also social norms [7].

Edutainment plays a crucial role in the field of communication for development across various development challenges, such as HIV and AIDS, sexual and reproductive health, domestic violence, and sustainable livelihoods [3,4,8-10]. In South Africa, which has the largest population of people living with HIV (UNAIDS [11]), both governmental and nongovernmental organizations have made significant investments in communication campaigns aimed at correcting misinformation and reducing HIV-related stigma [1,12]. Edutainment was key in correcting misinformation and reducing stigma during South Africa's response to the HIV or AIDS epidemic [13-16].

Edutainment became a key component in raising awareness about HIV and AIDS in South Africa, particularly through the popular drama *Soul City* [10]. Created by the Soul City Institute, *Soul City* is a multilevel and culture-sensitive communication strategy centered on the television drama that sought change at the individual, community, and broader society levels to inspire social action and sociopolitical mobilization by incorporating community mobilization and action into the narrative [7,17]. It integrated HIV storylines and reached millions of viewers, showing positive effects on knowledge, attitudes, and behaviors related to HIV prevention [10].

Since *Soul City*'s success, HIV and sexual health edutainment have become a popular approach for reaching South African audiences, especially younger age groups. *MTV Shuga*, *Down South* (2017-present), which is the focus of this study, is a current edutainment series that aims to create storylines centered on HIV prevention, including pre-exposure prophylaxis (PrEP) and HIV self-testing, for young people, a critical population for addressing the high incidence and prevalence of HIV in South Africa [18]. *Down South* is part of the *MTV Shuga* mass media campaign created by the MTV Staying Alive Foundation, which aims to engage young people with positive sexual health messages using compelling and relatable storylines. *MTV Shuga* is part of a large media campaign and has reached millions of viewers worldwide through television and web streaming since 2009.

When evaluating health edutainment interventions, research has often relied heavily on surveys and structured interviews to determine effectiveness [4,5,10,19,20]. These types of studies have shown the mixed effects of edutainment on sexual health outcomes. While these studies can track measurable changes in attitudes, behaviors, and tangible outcomes—such as increased use of services—they do not explain why certain elements of edutainment work whereas others do not. To fully understand the impact of edutainment, other methods must be used to explore the nuanced ways in which audiences engage with the content in their everyday lives [8].

Many studies that use qualitative methods to evaluate edutainment dramas with HIV or sexual health storylines rely on self-reported, retrospective data obtained through qualitative interviews and focus groups [3,21-25]. These methods have proven valuable in providing deep insights into participants' personal experiences, perceptions, and emotional responses to edutainment content. Such approaches are essential for understanding the mechanisms through which edutainment influences behavior, attitudes, and beliefs as they allow participants to articulate their experiences in their own words.

Moreover, the retrospective nature of the data collected in these studies can capture longer effects of edutainment [3,4,26,27]. This includes how the content continues to resonate with participants over time, influencing their decisions and actions. Focus groups facilitate discussions that elicit diverse perspectives and highlight shared experiences among participants, contributing to a more comprehensive understanding of edutainment's effects [4,21,23,24,27-29]. However, several limitations are inherent in these qualitative methods. Social desirability bias, recall issues, and the power dynamics between participants and researchers are notable concerns that may compromise the reliability of the data. Furthermore, there is a tendency to sample participants who are already engaged with the content, potentially skewing the findings toward more favorable opinions [4,21,22,26].

While most studies have relied on retrospective researcher-led methods, alternative approaches such as participatory methods and observational evaluations have also been explored. Some studies have used participatory and observational methods such as storytelling and community screenings to evaluate audience

XSL•FO
RenderX

engagement [26,28,29]. Observations have provided valuable insights into physical settings and collective experiences similar to those seen in theaters or live performances [30-32]. However, these insights may not fully reflect the diverse influences and mediating factors present in more intimate home environments, where most edutainment is consumed.

MTV Shuga has been evaluated using both quantitative and qualitative retrospective methods. The results of our own research using a web-based survey in the Eastern Cape, South Africa, have shown a positive association, consistent with a causal impact of *MTV Shuga "Down South"* (based in South Africa) on young people's awareness and use of HIV prevention tools such as HIV self-screening and condoms and greater motivation among adolescents to use PrEP for HIV [4]. Evaluations of other *MTV Shuga* series produced in other contexts such as Nigeria and KwaZulu-Natal in South Africa also show positive effects of engaging with *MTV Shuga*, for example, on young people's adoption of chlamydia and HIV testing, condoms, and contraception, as well as significant attitude shifts and increased disclosure of sexual violence among youth [33-35]. Qualitative evaluations using in-depth interviews and focus group discussions of *MTV Shuga*, including our own [21], have found that *MTV Shuga*'s relatable, unjudgmental content helps young people think through scenarios that relate to their real lives. They also describe how the discussions and debates inspired by *MTV Shuga* created a safe place to discuss sexual health, reducing stigma and increasing support [21,27]. These methods contributed valuable knowledge to *MTV Shuga*'s effectiveness and the mechanisms in which *MTV Shuga* works. Participants self-reported how social context can modulate the viewing experience. However, we lacked data on how those interactions happen in real time and the way in which people consume *MTV Shuga* in their everyday lives.

Edutainment evaluations rarely use participatory approaches to capture viewers' natural, real-time interactions and reactions, including how mass media edutainment is experienced in its actual viewing environment. To gain a more comprehensive perspective, there is a need to capture the viewing experience and how the viewing context, including both social and physical elements, influences the effectiveness and impact of mass media edutainment. This holistic understanding can further enhance the effectiveness of sexual health messaging through edutainment.

Objectives

In this paper, we explore the role of reaction videos as a method of evaluating an edutainment campaign. Our objectives for using reaction videos as a research method were to (1) create a remote approach to observation without the presence of a researcher in which young people could record their experiences of watching the *MTV Shuga* show in viewing groups and (2) understand how physical context and social dynamics influence young people's experiences of watching *MTV Shuga* within their household or other social spaces, with family, relatives, and friends. We also sought to critique the successes and challenges of using self-filmed observation for research purposes to reflect on the utility of this technique more widely in qualitative investigations with young people.

We piloted and described the use of reaction videos self-filmed by viewers as part of a mixed methods evaluation of *MTV Shuga* in South Africa. Previously, young people's real-life engagement with the *MTV Shuga* series has not been documented at the time or place of private viewing, for example, at home with family and friends. We hoped that self-filmed reaction videos would complement the self-reported data we collected [4,21,36] by revealing the physical and social context in which young people view sexual health media content and how that experience may influence the show's effect.

Analyzing reaction videos can provide valuable insights into viewers' real-time emotional and physical reactions within their contextual environment. These videos, which record people's responses as they watch scenes from movies, television shows, music videos, or comedy sketches, have become a prominent feature in web-based and media culture. The rising popularity of reaction videos on YouTube, TikTok, and other social media platforms reflects a fascination with observing other people's natural, unscripted social interactions. As with other media formats, reaction videos have the potential to be both entertaining and educational. They offer a literal lens with which we can observe the influence of media in real-time and real-life situations [37]. While viewers are aware of the recording and potentially alter their behavior consciously or unconsciously, reaction videos offer an inside perspective that cannot be easily captured otherwise and which can complement other data sources for research and evaluation.

There is an emerging body of literature on the nature and influence of reaction videos, primarily focusing on reaction videos on YouTube. Analysis of YouTube reaction videos reveals that they vary from amateur—unscripted, spontaneous, and organic—to more professional, scripted, and edited formats. While primarily aimed at entertainment, these videos also create narratives that guide viewers in interpreting content through the creators' perspectives [38]. They help build communities and challenge dominant views, enabling viewers to share their own perspectives and influence how videos are consumed [38]. These videos foster a sense of community by uniting viewers worldwide to share and discuss in comment sections, with participatory elements such as content suggestions and creator interactions further strengthening these connections [39,40].

Adopting reaction videos as a research method presents valuable potential for data collection. Instead of analyzing existing reaction videos on the web, we asked participants to create their own reaction videos, which would not be posted on the web for entertainment but used exclusively for research purposes. This approach aimed to minimize the influence of performative, exaggerated reactions often seen in web-based content, where viewers may enhance their responses to build a fan base or generate income. We hoped that these reactions would be similar to what participants might naturally experience at home. This method can provide insights from individuals who are not web-based personalities who may be posting content for attention or commercial purposes, which can skew data toward extremes due to algorithmic pressures. By adopting this format, researchers can collect real-time responses and insights, offering a rich qualitative data source reflecting participants' immediate reactions and emotions. This approach deepens data collection

XSL•FO
RenderX

and provides a novel way to engage participants in the research process. The use of reaction videos as a research method can draw from forms of traditional data collection, including photovoice and observations. The participatory and visual principles of photovoice apply here as they empower participants to document and share their reactions and perspectives, fostering a more engaged and authentic form of data collection [41,42]. Similarly, the principles of observation can be used by capturing real-time interactions and spontaneous responses, providing rich, contextual insights. Video recordings can allow researchers to capture the participants' physical environment, social dynamics, and other contextual factors without physically being there [43,44]. Reaction videos introduce an additional participatory dimension to video observations as the participants are responsible for self-recording, allowing them to decide what to capture.

Methods

Recruitment

We recruited participants from a web-based survey that was live on the web in South Africa from September 2020 to December 2020. The survey was part of an evaluation to understand the impact of the mass media campaign *MTV Shuga Down South season 2 (DS2)*, and detailed methods and results to date have been published previously [4,21,36]. In short, 3431 young people aged 15 to 24 years completed the web-based survey. Participants were recruited through targeted advertisements to individuals aged 15 to 24 years in Mthatha through social media platforms, primarily Facebook. The questionnaire was also promoted in collaboration with local schools, universities, and clinics through their Facebook pages and WhatsApp groups.

This study was conducted in Mthatha, Eastern Cape, which was selected due to the recent distribution of HIV self-testing kits and the availability of oral PrEP in the area, ensuring that any demand generated by the *MTV Shuga DS2* campaign could be met. In addition, the offline components of the *MTV Shuga DS2* campaign were active in this region, providing a local context for examining the campaign's impact. Mthatha was also selected because of its relatively lower levels of HIV prevention research and testing compared to other provinces, making it a priority area for intervention. This decision was further supported by the South African Department of Health, which identified Mthatha as a strategic site for increasing HIV prevention efforts among youth.

Of those who reported ever having watched *MTV ShugaDS2* (N=238), 28.6% (n=68) within the study area of Mthatha in Eastern Cape opted in to further research to be conducted remotely due to COVID-19 restrictions at the time. Participants were first invited to individual interviews or focus groups, the findings of which are reported elsewhere [4,21], and participants of those activities were subsequently invited to create video recordings of themselves watching *MTV ShugaDS2* episodes. Initially, the eligibility criterion was that viewing groups should include a parent and their adolescent children (aged 15-19 years) to understand parent-child viewing dynamics. However, when it proved difficult to recruit adolescents willing to watch with their parents, eligibility was expanded to all 68 participants from Mthatha (aged 15-24 years) who opted in to further research to watch *MTV ShugaDS2* with any family members or housemates of their choice (within their COVID-19 "social bubble"). Of those, participants were called at random until 8 consented to and participated in the reaction videos.

The primary participant for each viewing group, recruited through focus groups and interviews, must have watched at least one episode of *MTV Shuga DS2*, but the other participants in the group did not require any previous exposure. We did not impose an age limit for this activity as the main participant was drawn from the evaluation study with individuals aged 15 to 24 years, within the target audience for *MTV Shuga*. As all participants responsible for organizing their viewing group had some familiarity with *MTV Shuga*, they were responsible for determining their viewing group and assessing the appropriateness of the situation.

A total of 8 groups was considered a suitable number to trial the method and analytical framework with a range of viewing groups. The decision to use a small sample size was justified by the study's aim to gain in-depth insights into family dynamics and viewing behaviors. We predicted that we would reach saturation with 8 groups ensured. By focusing our resources on a small number of groups watching multiple episodes of *MTV Shuga*, we prioritized depth over breadth, allowing for meaningful interactions and a thorough understanding of each family's viewing experience. Practical and budget constraints also influenced the sample size, ensuring a feasible research process.

Ethical Considerations

This study was approved by the Biomedical Research Ethics Committee at the University of KwaZulu-Natal (reference BREC/00000477/2019), the London School of Hygiene and Tropical Medicine (reference 17996), and the World Health Organization (reference ERC.0003283). Ethics approvals covered all phases of the research, including participant recruitment and data collection and analysis. Participants, who were recruited through the web-based survey, provided web-based consent before joining the study. For participants aged <18 years, consent was obtained from parents or guardians, and this was done via email.

Participants who opted into the individual interviews or focus groups and were later invited to the reaction videos were provided with detailed information sheets and had the opportunity to discuss the study with data collectors over the phone. During these discussions, participants could ask questions to ensure that they fully understood the nature of their participation. Parents of participants aged <18 years were required to give consent via a form sent over email.

This study took several measures to ensure privacy and confidentiality. All participant data were anonymized. The reaction videos themselves were not disseminated outside of the research team. To maintain secrecy while illustrating moments from these videos, the research team commissioned a young South African artist, Lunga Jonas, to create graphic art representations of the footage. These artworks altered

XSL•FO
RenderX

participants' facial features to prevent identification while retaining the social and emotional context of the group viewings.

In recognition of the time that the participants contributed to the study, they were compensated with 100 rand (US $5.54) for their participation. This compensation was provided to ensure that participants were fairly acknowledged for their involvement without coercion or undue influence.

This study followed the COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist to ensure the transparency and rigor of the research process, including participant selection and data collection and analysis procedures.

Data Collection

Researchers conducted informed consent procedures via a WhatsApp phone call with the entire viewing group, specifically to confirm that they were aware of the reaction videos and their purpose and consented to being filmed. The survey participants signed a consent form via email on behalf of their viewing group. For participants aged <18 years, the written permission of parents and guardians was also confirmed via email. We asked participants to use a smartphone to video record their group watching *MTV Shuga DS2* episodes 5 and 6 on YouTube, both 22 minutes long. It was not required for participants to have seen episodes 1 to 4 before participating in the reaction videos. These episodes include storylines about PrEP, HIV self-screening, and gender-based violence. They also feature 2 scenes with physical intimacy that include kissing, removing clothes, and opening condoms. All participants were able to choose where, when, and with whom they would watch the episodes. The participants transferred their video recordings via WhatsApp to the research team, who deleted them from the research phone once downloaded to a password-protected private server.

In total, 4 researchers familiar with the study setting and fluent in the local language reviewed each reaction video and took notes, describing the setting, interactions, and significant verbal and nonverbal communication in ways agreed upon during researcher training to aid consistency. Nonverbal behavior was defined to encompass body movements and postures, along with temporal speech markers such as gaps and silences. Verbal behavior was characterized by spoken words, verbal expressions such as gasps and laughter, and the strength or emotive inflection in vocal expressions [45,46].

Following their review, researchers called the main participant of each viewing group via a WhatsApp "debrief" call to ask about the experience of watching *MTV Shuga DS2* with this group of people and whether they had any advice to improve the show. Participants' responses were summarized and included in the notes. Research notes and debrief notes were recorded in English and translated from Xhosa into English as needed.

Analysis

To analyze reaction videos, we created a framework to capture the social and physical context of people consuming edutainment in their everyday lives [8,19]. We designed a framework with the following components: *group composition*, *physical setting*, *coinciding activities*, *social dynamics*, and *reactions to the DS2 show and this research activity* (Textbox 1).

Textbox 1. Reaction video analysis framework.

Components

1. Group composition: the characteristics of those involved in the viewing, including their age, gender, relationships with each other, and other relevant information provided by the participants.
2. Physical setting: this describes the physical space in which the participants viewed the *MTV Shuga DS2* episodes. It encompasses factors such as the location, the time of day, surrounding physical objects, lighting, and overall environment, as well as the mode or device through which they viewed *MTV Shuga DS2*.
3. Coinciding activities: this constitutes any concurrent activities happening during the viewing that might either enhance or distract from engagement with the show.
4. Social dynamics: this constitutes the interactions between participants during the viewings. It includes spoken dialogue or unspoken gestures and responses between members of the viewing group.
5. Reactions to the *MTV Shuga DS2* show: this encompasses any verbal or nonverbal reaction expressed by participants in response to the *MTV Shuga DS2* episodes and scenes they are watching. This includes body language, facial and bodily expressions (such as laughing or gasping), and their overall attitudes while watching (eg, bored, relaxed, awkward, or happy) as perceived by the researchers and from the debrief call with the main participant.
6. Reactions to the research activity: this refers to participants' awareness of the camera, including speaking directly to the camera, their chosen positioning for the filming, and any statements they made in the debrief call about their experience of filming.

To apply our analytical framework to the coding process, we used the framework's subheadings to code each video recording, observation note, and audio transcript. Short summaries were written for each video, organized by the framework's subheadings. These summaries were then combined to identify codes within and across the videos. The lead analyst collaborated with the data collection researchers and a translator to discuss and interpret the findings, ensuring that they accurately reflected the participants' experiences. These codes were synthesized into themes guided by the framework supported by specific data excerpts such as quotations from transcripts or descriptions of moments captured in the videos.

The researchers involved in conducting the follow-up call (Zulu and Xhosa, male and female, and aged <30 years) and the lead analyst (White, British woman, and aged 32 years) reflected on how their knowledge, attitudes, assumptions, and positions might affect the interpretation of the data. On the basis of these

XSL•FO
RenderX

reflections, we determined that the initial analysis by the lead analyst should be verified with the other researchers and transcribers to ensure that the data were accurately summarized and interpreted.

Results

Group Composition

A total of 8 self-filmed reaction videos were received from 8 distinct viewing groups (Multimedia Appendix 1). Only 1 group (group 1) included a parent who filmed her 3 sons (aged 15-26 years) while watching the show together. Other groups included viewers ranging in age from 7 to 32 years. These groups, confined to "COVID-19 social bubbles," primarily consisted of siblings and cousins. All groups included male individuals: 38% (3/8) included male individuals only and 62% (5/8) were mixed-gender groups. There were no exclusively female groups; all female viewers were accompanied by older male cousins or younger brothers who were not within the target age range for *MTV Shuga* (15-25 years).

Physical Setting

As illustrated in Multimedia Appendix 1, the environments in which the groups convened to watch *MTV Shuga DS2* were diverse, with a blend of communal and secluded spaces. Use of laptops by all groups may have allowed for flexibility in where they chose to watch the show. Groups 5, 6, and 7 viewed the episodes in spacious living rooms with plush sofas and chairs, suggesting an evening setting or at least an ambiance with the curtains closed. Group 1, on the other hand, opted for a more structured viewing setup, congregating around a dining table during daylight hours. Groups 4 and 8 chose the privacy of bedrooms, where they huddled around laptops with siblings or cousins. These indoor gatherings were marked by an informal, relaxed ambience, with participants donning loungewear and pajamas and sporting bonnets, all contributing to the gatherings' casual and comfortable essence.

In total, 67% (2/3) of the all-male groups, groups 2 and 3, opted for outdoor settings for their daytime viewings. Group 2 created a makeshift viewing area by arranging chairs around a crate situated beside an exterior wall. The 4 brothers in group 3 sought an enclosed environment, choosing to watch the episodes from the confines of a vehicle. Although they mostly had the windows and doors shut, they occasionally rolled them down to talk to passersby who approached the vehicle to greet the boys inside.

Coinciding Activities

Communal activities often accompanied the viewing of the show, particularly in groups watching at home. For instance, group 6 and 7 participants combined viewing with eating a meal, whereas group 5 engaged in doing their hair. These activities illustrated the intimacy or familiarity of these groups; however, they also introduced potential distractions. Engaging in activities unrelated to the show led to conversations and actions that diverted attention away from the content being watched. As a result, participants might have missed crucial dialogue or scenes, highlighting a trade-off between enhancing the communal viewing experience and maintaining focus on the show's narrative and messages.

Other distractions included people who were not involved coming to chat with the participants, which only happened in group 3, who met in the vehicle.

Phone use emerged as a significant distraction during the viewings. Although recording the viewings required smartphones, other phones were frequently present and easily accessible to participants. While most participants focused on the show, smartphones proved to be potential distractions. For instance, in group 2, a participant began scrolling through his phone for several minutes, diverting his attention away from the show. Similarly, in group 5, a viewer aged 13 years was less engaged with the show than her sister, who was aged 24 years; more absorbed in her phone; and visibly distracted by an ongoing text conversation, as indicated by her active notifications. Both the older sister and younger brother nudged her and told her to concentrate on the episodes. In addition, smartphones were sometimes deliberately used to "escape" the viewing. In group 7, during a scene that made him noticeably uncomfortable, one man turned to his phone to avoid watching the laptop screen.

Despite most groups experiencing some distractions, no group paused or rewound to avoid missing the show.

Social Dynamics

Awkwardness

While most of the time viewers appeared relaxed around each other, intimate scenes showing couples kissing, opening condoms, or discussing PrEP created awkwardness across many groups. Viewers aged ≤15 years often responded to the intimate scenes by closing their eyes, covering their faces, or laughing nervously. In group 1, the mother often teased and joked with her sons about their visible discomfort during some scenes, playfully remarking that "it's not that painful" when her children appeared awkward during an intimate scene about PrEP that ended with characters kissing and taking off clothing. Although the atmosphere appeared lighthearted, the boys, smiling shyly, laughing with flushed faces, shaking their head, and covering their eyes, were embarrassed watching these scenes with their mother. In group 6, the teenage girl looked up and smiled awkwardly, showing discomfort as she pushed her face onto a pillow to hide from the intimate scene. Her uncle remarked the following in the debrief:

> *It was awkward a little bit because my niece and I are not used to each other that much, so I was uncomfortable with some scenes.*

In the debrief calls, groups 5, 7, and 1 agreed that watching with family was awkward, but overall, they said that they enjoyed the experience.

Dialogue

The reaction videos highlighted the social aspect of watching a show as a group, especially how it elicits reactions and facilitates discussion among viewers. In groups of young men, the show's content related to HIV prevention, often intimate in nature, generated reactions or dialogue; for example, one participant in group 3 encouraged the characters on-screen to engage in sexual activity using the phrase "Shaya shaya [Hit

XSL•FO
RenderX

Hit].” Participants explained how watching the show created an environment where they could discuss topics that came up on the show once the viewing was complete. Group 2 shared the following:

> *I love watching shows like these with my brothers because we discuss and agree and disagree on some issues afterwards, but it helps because we grow each other.*

For group 3, the discussions were also a major component of enjoyment of the show; they stated that they “loved every moment of it because we discussed it afterwards and shed light, so it was very informative.” However, it was not universal that groups enjoyed watching the show in groups—mainly due to discomfort around intimate scenes and relationship dynamics with family members, as referenced in the aforementioned theme. As the group 6 uncle stated, he found it uncomfortable watching with his niece, but “I wouldn’t have a problem watching alone.”

Age Dynamics

This study revealed varying dynamics in viewing experiences across different age groups. Groups in which viewers were of a similar age, such as group 7, found the viewing experience agreeable and comfortable, noting the following:

> *It was okay [watching with family]. We are all adults.*

Similarly, group 4 said the following:

> *We enjoyed it. It was fine because we are almost the same age group.*

The presence of younger siblings aged 7 to 8 years in groups 5, 6, and 8 introduced unique dynamics. These younger viewers were notably active, often moving in and out of the room, yet engaged with the show by asking questions about the storyline, such as “What’s happening?” (group 8) or “Who is Lemo?” (group 5). These inquiries sometimes elicited impatience from their older siblings. For instance, during a scene in which a character discovers that his girlfriend has tested positive for HIV, the viewer aged 8 years in group 5 began narrating the event, only to be cut off by an older sister with a dismissive, “You don’t need to explain.” The video recording later captures the end of a discussion happening off camera, then shows a visibly upset younger brother remarking, “This is such a joke, this is supposed to be a family thing” before exiting the room.

In the group 8 recording, the younger brother playfully interacted with the camera, making faces and sticking out his tongue, prompting a reprimand from his older sister filled with a mix of discipline and humor:

> *If you do that one more time, I am going to make you rewatch this [laughs and pauses], but I’m serious.*

These interactions show that, although younger viewers were included, their older siblings frequently guided or managed their behavior. It led to distractions for the older viewers and irritation of the younger ones.

Reactions to the Videos

Entertainment Value

Overall, the participants seemed to enjoy watching the show, finding it engaging and fun. Many participants stated in the debrief that they enjoyed the music, and this was evident during the recordings, with some singing along and dancing in moments (groups 6 and 8) and recognizing famous actors on the show. Others enjoyed the other creative components integrated into the show; for instance, the older sister in group 5 hummed and snapped her fingers during a scene with a poetry slam.

Reactions to Storylines

Q and Dineo (Transactional Sex)

The storyline between Q and Dineo, which portrays a budding romantic relationship, elicited varied reactions from participants, particularly regarding the theme of transactional relationships. Dineo’s involvement in a transactional sexual relationship, where her school fees are paid by an older man (a “Blesser”), sparked comments on the motivations behind such relationships. In group 4, a viewer empathized saying, “Things we do for money” when Dineo explained to Q why she was sleeping with an older man.

Q’s initial response to Dineo’s situation was seen as understanding and supportive. The group 1 mother remarked, “That’s one understanding guy” as Q came to terms with Dineo’s arrangement with her Blesser.

However, Q’s attempts to win Dineo’s affection by spending money on her generated different reactions. In group 2, a male viewer aged 19 years noted Q’s financial expenditure, remarking, “u fit kanje,” which translates to “so much money.” This observation was echoed by other groups, with expressions of surprise and dismay watching Q withdraw his savings to spend on Dineo.

In group 3, participants expected physical intimacy in Q and Dineo’s relationship as a return for Q’s financial spending, encapsulated in the following comment—“Akamunike futhi (She must give it to him now)”—showcasing the perceived sexual obligation created by the financial exchange.

Reggie and Odirile (Same-Sex Relationship)

The responses of boys and men to scenes showing same-sex intimacy were characterized by notable discomfort. This was manifested by in-jokes and physically shielding their eyes during the scenes in which 2 male characters showed intimacy while discussing the use of PrEP. In groups 2 and 7, male viewers’ reactions to these scenes were marked by avoidance; a male individual from each group walked away when the 2 male characters started kissing. This strong aversion contrasted with a female viewer in group 7 who laughed at the boys’ uneasy reactions. In group 4, during a scene in which male characters Reggie and Odirile kissed, the group expressed their unease (shaking their head, covering their eyes, and groaning), with one stating, “They should move on to the next scene,” suggesting a collective discomfort or disinterest in engaging with same-sex romantic content. Young men’s reactions

XSL•FO
RenderX

potentially led them to miss the scene and the educational content about PrEP and HIV prevention options.

Daniel and Ipeleng (HIV Self-Testing)

The HIV self-testing scenes in *MTV Shuga DS2* were designed to be educational and showed the characters Daniel and Ipeleng using a mouth swab for self-testing at home. This scene elicited a subdued reaction compared to scenes involving PrEP and intimacy. Only groups 7 and 5 visibly reacted to these moments. In group 5, someone off camera expressed disbelief saying, "i-swab? [a swab?]," showing that she was surprised that HIV could be screened for using an oral swab. Group 7 participants showed skepticism; one doubted the method's reliability, preferring a blood test and saying, "Don't trust this one; I want the blood one." Another participant misconstrued the testing method's implications, mistakenly suggesting, "If they swab, that means we can get HIV through kissing."

Reaction to the Research Method

The reaction videos were, on average, 29.38 (SD 9.3) minutes long. The duration of the 2 episodes together is 44 minutes. Some of the videos were sent in 1 long recording, whereas other groups sent multiple shorter videos.

The awareness or presence of the camera played different roles in each video. In group 1, the mother, holding her phone to film the observation, stood over her sons, using the camera to zoom in and capture their expressions. Although her voice was audible, she remained invisible on-screen. Her sons were very aware of the camera and often hid their faces, smiling shyly as she brought the camera close. In contrast, the young female participants who watched with their younger siblings (groups 5 and 8) positioned themselves as the central characters in the viewing, ensuring the camera captured their reactions. They appeared to be aware of the camera, sometimes glancing at it. For instance, the viewer aged 24 years in group 5 set up the camera and positioned herself within the frame. She took charge of the viewing, reprimanding her sister for checking her phone and not paying attention to the show. Her glances at the camera showed an awareness that the video had captured the distraction.

Discussion

Principal Findings

This study offers valuable insights into how physical and social contexts shape engagement with edutainment content such as *MTV Shuga*. Reaction videos captured nuanced dynamics within diverse viewing groups, including siblings, cousins, parents, and friends, revealing varying levels of comfort and discussion. Participants' responses to themes such as same-sex intimacy, HIV prevention, and transactional sex were influenced by familial and peer relationships, as well as the settings in which they viewed the show—ranging from communal living rooms to private bedrooms and outdoor spaces. Concurrent activities such as eating, styling hair, and smartphone use further modulated the viewing experience. Our findings demonstrate that reaction videos effectively captured spontaneous engagement, providing unique insights into nonverbal cues, group dynamics, and emotional reactions often missed by retrospective methods. This real-time documentation deepened our understanding of how physical and social environments impact viewer engagement. While self-filmed observations proved valuable in capturing immediate reactions and discussions, challenges such as varying levels of camera awareness and engagement with the content could affect the authenticity of the viewing experience.

This approach leverages smartphone cameras to allow participants to show their physical and social context while enabling remote observation by researchers. This eliminates the need for a researcher's physical presence, potentially reducing response bias as the presence of the in-person researcher will not influence social dynamics. While the absence of a researcher can diminish direct influence, the visible awareness of being filmed and the anticipation of sharing these recordings for research could shape participants' behaviors.

However, while participants were aware of the camera, most did not directly engage with it. Unlike the highly entertainment-focused content on social media, the reaction videos in this research captured responses in natural viewing environments that appeared candid and unobtrusive [38,40]. The reactions recorded in this study likely reflect genuine, unguarded responses, similar to those occurring when no "audience" is watching the reaction. This contrasts with the more performative reactions often seen in content aimed at entertainment, persuasion, or monetization [38,39].

Granting participants the autonomy to decide their positioning in front of the camera and select the moments they wish to record transforms the research process into a participatory one. The dynamic of power and control shifts as participants handle the camera by self-focusing or navigating family interactions, such as older siblings monitoring their younger siblings' interaction with the camera or parents capturing playful dominance over their children's reactions. Participatory research is vital in adolescent research as it can examine power and voice while simultaneously being an empowering activity, building agency and giving voice to underrepresented or marginalized groups [47]. This shift in focus from researcher-controlled observation to participant-driven recordings highlights a key aspect of this study's participatory methodology. Drawing from the principles of visual methodologies and participatory research rooted in the traditions of ethnography within anthropology and sociology [43,46], this method empowered participants to engage with the activity on their own terms.

The recordings enabled participants to document their reactions to *MTV Shuga* within the familiarity of their physical environment. Filming reactions in the comfort of their own homes or chosen environments, participants provided insights closely reflecting how they naturally might consume and engage with *MTV Shuga*. While open living rooms with communal activities created potential distractions, more intimate settings such as bedrooms allowed participants to engage deeply with the content. Outdoor spaces offered another setting for engagement as participants viewed in peer groups away from home. The distractions from young siblings and the disruptions of coinciding activities such as smartphone use or passersby could have impeded the educational and entertainment value of watching *MTV Shuga*.

XSL•FO
RenderX

While these various physical environments shaped the nature of participants' engagement with the content, social dynamics within families also played a significant role in how the show was received. Even during the recruitment phase, participants resisted joining the qualitative reaction videos when we initially requested that adolescents (aged 15-19 years) watch the show with their parents. While we initially hypothesized that *MTV Shuga* could have a "family influence" by encouraging communication between parents and adolescents, we have since observed resistance to this in other research activities, including in-depth interviews and focus group discussions, in which adolescents cite discomfort in viewing sexually explicit content with parents for fear of awkwardness and judgment [21,27,36]. Some parents shared an interest in viewing and learning from the show, but encouraging "family" or intergenerational viewings can depend on the cultural context, for example, whether social norms about adolescents or sex can be challenged safely.

Participants noted that watching with peers of similar ages made the experience more engaging, sparking discussions and learning opportunities. The groups that reported engaging in postviewing discussions were mainly young men who viewed the show outside their homes. As observed in our previous study, creating youth-friendly spaces for media consumption could further enhance engagement with sexual health content [21].

Reaction videos capture the spontaneity of viewers' real-time responses, providing insights into their immediate emotional reactions to the show's content. Follow-up questions with a researcher further enrich these data, allowing participants to reflect on and articulate their experiences. This approach offers a dual layer of insight—combining initial reactions with more reflective postviewing feedback.

In particular, real-time reactions among viewing groups with mixed ages revealed that visual intimacy scenes in *MTV Shuga*, such as characters kissing, undressing, or reaching for condoms, caused discomfort for many participants. This led to distractions and awkwardness as participants covered their faces, walked out of the room, or picked up their phones often during critical and short HIV prevention scenes. It is not clear whether these reactions were due to the presence of others in the viewing group or the presence of the camera (spurred by social expectations) and may not have happened in the absence of either.

Despite the embarrassment caused by visual intimacy, participants approached discussions about sexual topics with more comfort when not accompanied by intimate scenes. For instance, several groups engaged in animated discussions about Q and Dineo's relationship and transactional sex. Groups of young men, for example, were content watching scenes of Reggie and Odirile discussing sexual health but reacted negatively when the couple kissed on-screen.

These findings suggest that *MTV Shuga* may be more effective in facilitating conversations around sexual health and relationships if it reduces the intimate scenes that make viewing uncomfortable. However, the show must balance its aim of creating destigmatized representative content (eg, to help normalize kissing between a same-sex couple) with creating content that is comfortable for its audience [48]. In addition, in some cases, generating a strong or discomfiting reaction can challenge viewers to think and learn and leave a more lasting impression as long as it is not harmful or gratuitous. Reducing intimate scenes could dilute the drama and sexual appeal that make the show compelling, potentially diminishing its appeal to younger viewers.

Beyond these broader social dynamics, participants' responses to specific content within *MTV Shuga* revealed key insights into the reception of public health messages. For example, the HIV screening scene, particularly the depiction of a swab for self-screening, generated limited dialogue and, concerningly, misconceptions about HIV transmission through saliva. Similar findings were documented in our earlier research [4]. To prevent misunderstandings and concerns among viewers, future series can embed clearer explanations of how oral swabs check for HIV antibodies in saliva and do not detect the HIV virus itself given the absence of real-time clarification from health professionals in media interventions.

Alongside responses to HIV prevention content, participants' reactions to the show's musical elements revealed how music shaped their engagement with the show. Reaction videos captured the immediate somatic responses of viewers to the music featured in *MTV Shuga* as they sang, danced, and snapped their fingers to the rhythm. Music holds a significant place in South African edutainment, particularly in programs that combine performance theatre and music to promote public health and HIV awareness [30-32]. These artistic methods leverage music's emotional and rhythmic power to create a sense of interconnectedness and euphoria among audiences, making them more receptive to new ideas and behaviors [49].

Young people find the integration of music into health interventions both engaging and refreshing, which is crucial given that they are often overwhelmed and fatigued by the constant stream of HIV messaging and promotions. This saturation can lead to disengagement, making innovative approaches such as music-based edutainment particularly valuable in capturing their attention and encouraging meaningful engagement [32].

Limitations and Adaptations

The findings of this study, involving a small sample of 8 reaction videos from a specific region of South Africa, may not be generalizable to other social and cultural contexts. The demographics of the participants do not represent *MTV Shuga* viewers in South Africa in general as no groups were female only. Female individuals constitute the largest viewing group of the MTV series [22]. Recruitment challenges, especially in engaging parents, impacted this study's original design, constraining insights into parent-child dynamics during *MTV Shuga* viewings.

This study was conducted during the COVID-19 pandemic, likely impacting participants' spaces and viewing habits. As a result, the viewing environment may not reflect the participants' natural choices. COVID-19 restrictions also limited participants' ability to select their preferred viewing partners, potentially affecting the authenticity of group composition and individual

XSL•FO
RenderX

responses. Their reactions may have been influenced by peers' or family members' collective atmosphere and expectations.

Not all participants adhered to instructions to ignore the camera, indicating that the instructions were not clear to all and may have caused variation in the participants' reactions to the activity. Nonetheless, their interactions with the camera provided valuable insights into social dynamics.

The success of this method hinges on participants' ease with technology and access to the necessary resources, such as the internet, smartphones, and a laptop, further restricting who would be able to participate in this study. Challenges arising from the reaction video method included participants using personal devices for filming, impacting the quality of audio and visuals in the recordings. The videos' narrow view makes it challenging to understand dynamics outside the camera's frame within the room or other spaces.

A final limitation of the reaction video method is that it primarily captures only what is externally visible, audible, or physical. It fails to delve into the internal processes, such as thoughts and emotions, that are crucial for understanding the deeper cognitive and emotional responses of participants. As a result, a significant layer of participant experience remains unexplored.

The reaction videos provided valuable insights into both the social context and content engagement, and these findings also point to opportunities for expanding the approach in future studies. By adopting a more interactive research design, reaction videos could capture different types of data, offering deeper insights into participants' thoughts and experiences. Researchers could capture more verbal data and gain insights into inner thoughts by asking participants to speak directly into the camera and express their thoughts more overtly. This method captures thoughts and feelings lost in a less interactive observational approach. Reaction videos can also be adapted to solo reactions, which could yield insights into more personalized viewpoints, again akin to those shared on social media platforms. As noted by some participants, some would be more comfortable viewing *MTV Shuga* alone.

Conclusions

This study's exploration of viewer engagement with *MTV Shuga* through reaction videos illuminated the nuanced impact of physical and social environments on the consumption and influence of edutainment. By capturing real-time reactions and discussions among various viewing groups in their chosen settings, this research method offered valuable insights into the immediate emotional responses and subsequent reflections on themes of intimacy, HIV prevention, and transactional sex. In addition, the innovative use of these reaction videos represents a significant opportunity for adolescent health research. This participatory approach respects young people's autonomy and perspectives and provides deeper insights into their interactions with health-related content. Furthermore, it showed reaction videos to be a valuable method to capture dimensions of context, the physical and social context in which edutainment is consumed, which has not been documented using other evaluation methods.

This study acknowledges the inherent limitations in capturing the authentic viewer dynamics due to the self-selective nature of the viewing groups, the challenges posed by the COVID-19 context, and the observer effect of the camera. The findings suggest that, while reaction videos can significantly enrich our understanding of audience reception, careful consideration must be given to the potential influences on participant behavior and the representativeness of the data collected.

Acknowledgments

The authors thank the young people who gave their time to this study. They especially thank Lunga Jonas for her beautiful artwork, which was used to capture the self-filmed observations. The authors are grateful to Luntu Fica, Bawelile Madlala, Samkele Khuma, and Lunga Mkrezo for their hard work collecting data and valuable ideas that strengthened this research and to Antonio Duran-Aparicio for his support managing the grant. The authors acknowledge the dedication of *MTV Shuga* peer educators and coordinators (Yvonne Diogo and Lesedi Thwala) and the commitment of the MTV Staying Alive Foundation, especially Alistair Chase and Sara Piot.

Data Availability

The datasets generated during and analyzed during this study are not publicly available due to the identifiable nature of the data but nonidentifiable data such as transcripts and summaries are available from the corresponding author on reasonable request.

Authors' Contributions

VB and IB conceived the study design and methods. VB, DK, CC, SC, and IB codeveloped the study protocol, and CC, DK, and VB managed the data collection. VB led the observation data analysis and wrote the first draft. IB and VB co-designed the analysis framework. All authors contributed to revisions and reviewed the final draft.

Conflicts of Interest

GA was the executive director of the MTV Staying Alive Foundation at the time this study was designed and conducted and has since left the foundation.

XSL•FO
RenderX

Multimedia Appendix 1
Illustrations and descriptions of the viewing groups partaking in the self-filmed reaction videos.
[PNG File , 4911 KB - formative_v9i1e55275_app1.png]

References

1. Kincaid D, Parker W, Johnson S, Schierhout G, Kelly K, Connolly C, et al. AIDS communication programmes, HIV prevention, and living with HIV and AIDS in South Africa, 2006. The United States Agency for International Development. 2008. URL: https://www.researchgate.net/publication/263234176_AIDS_Communication_Programmes_HIV_Prevention_and_Living_with_HIV_and_AIDS_in_South_Africa_2006 [accessed 2024-04-29]
2. LaCroix JM, Snyder LB, Huedo-Medina TB, Johnson BT. Effectiveness of mass media interventions for HIV prevention, 1986-2013: a meta-analysis. J Acquir Immune Defic Syndr 2014 Aug 15;66 Suppl 3:S329-S340. [doi: 10.1097/QAI.0000000000000230] [Medline: 25007204]
3. Pichon M, Carter DJ, Howard-Merrill L, Sono R, Gimunta V, Rutenge O, et al. A mixed-methods, exploratory, quasi-experimental evaluation of a radio drama intervention to prevent age-disparate transactional sex in Tanzania. Front Reprod Health 2022;4:1000853 [FREE Full text] [doi: 10.3389/frph.2022.1000853] [Medline: 36531443]
4. Birdthistle I, Mulwa S, Sarrassat S, Baker V, Khanyile D, O'Donnell D, et al. Effects of a multimedia campaign on HIV self-testing and PrEP outcomes among young people in South Africa: a mixed-methods impact evaluation of 'MTV Shuga Down South'. BMJ Glob Health 2022 Apr;7(4):e007641 [FREE Full text] [doi: 10.1136/bmjgh-2021-007641] [Medline: 35365480]
5. Orozco-Olvera V, Shen F, Cluver L. The effectiveness of using entertainment education narratives to promote safer sexual behaviors of youth: a meta-analysis, 1985-2017. PLoS One 2019;14(2):e0209969 [FREE Full text] [doi: 10.1371/journal.pone.0209969] [Medline: 30753185]
6. Singhal A, Rogers E. Entertainment-Education: A Communication Strategy for Social Change. New York, NY: Routledge; 2012.
7. Orr G. Diffusion of innovations, by Everett Rogers (1995). Stanford University. URL: https://web.stanford.edu/class/symbsys205/Diffusion%20of%20Innovations [accessed 2024-04-29]
8. Cardey S, Garforth C, Govender E, Dyll-Myklebust L. Entertainment education theory and practice in HIV/AIDS communication: a South Africa/United Kingdom comparison. Critical Arts 2013 Jun;27(3):288-310. [doi: 10.1080/02560046.2013.800665]
9. Usdin S, Scheepers E, Goldstein S, Japhet G. Achieving social change on gender-based violence: a report on the impact evaluation of Soul City's fourth series. Soc Sci Med 2005 Dec;61(11):2434-2445. [doi: 10.1016/j.socscimed.2005.04.035] [Medline: 16006028]
10. Goldstein S, Usdin S, Scheepers E, Japhet G. Communicating HIV and AIDS, what works? A report on the impact evaluation of Soul City's fourth series. J Health Commun 2005;10(5):465-483. [doi: 10.1080/10810730591009853] [Medline: 16199389]
11. The urgency of now: AIDS at a crossroads. Joint United Nations Programme on HIV/AIDS (UNAIDS). URL: https://www.unaids.org/sites/default/files/media_asset/2024-unaids-global-aids-update_en.pdf [accessed 2024-04-29]
12. Peltzer K, Parker W, Mabaso M, Makonko E, Zuma K, Ramlagan S. Impact of national HIV and AIDS communication campaigns in South Africa to reduce HIV risk behaviour. ScientificWorldJournal 2012;2012:384608-384606 [FREE Full text] [doi: 10.1100/2012/384608] [Medline: 23213285]
13. Bennet N. The AIDS conspiracy: science fights back. Lancet Infect Dis 2012 Aug;12(8):596. [doi: 10.1016/S1473-3099(12)70195-0]
14. Fassin D, Schneider H. The politics of AIDS in South Africa: beyond the controversies. BMJ 2003 Mar 01;326(7387):495-497 [FREE Full text] [doi: 10.1136/bmj.326.7387.495] [Medline: 12609950]
15. Fourie P, Meyer M. The Politics of AIDS Denialism: South Africa's Failure to Respond. New York, NY: Routledge; 2016.
16. Chigwedere P, Essex M. AIDS denialism and public health practice. AIDS Behav 2010 Apr;14(2):237-247. [doi: 10.1007/s10461-009-9654-7] [Medline: 20058063]
17. Tufte T. Entertainment-education and participation: assessing the communication strategy of Soul City. J Int Commun 2001 Dec;7(2):25-50. [doi: 10.1080/13216597.2001.9751908]
18. Francis SC, Mthiyane TN, Baisley K, Mchunu SL, Ferguson JB, Smit T, et al. Prevalence of sexually transmitted infections among young people in South Africa: a nested survey in a health and demographic surveillance site. PLoS Med 2018 Feb;15(2):e1002512 [FREE Full text] [doi: 10.1371/journal.pmed.1002512] [Medline: 29485985]
19. Tufte T. Fighting AIDS with edutainment: building on the Soul City experience in South Africa. In: Servaes J, editor. Communication for Development and Social Change. Thousand Oaks, CA: Sage Publications; 2008:327-344.
20. Bertrand JT, O'Reilly K, Denison J, Anhang R, Sweat M. Systematic review of the effectiveness of mass communication programs to change HIV/AIDS-related behaviors in developing countries. Health Educ Res 2006 Aug;21(4):567-597. [doi: 10.1093/her/cyl036] [Medline: 16847044]

XSL•FO
RenderX

21. Baker V, Mulwa S, Sarrassat S, Khanyile D, Cousens S, Cawood C, et al. 'It is guiding us to protect ourselves': a qualitative investigation into why young people engage with a mass-media HIV education campaign. Cult Health Sex 2023 Jul 28;25(7):803-817. [doi: 10.1080/13691058.2022.2100483] [Medline: 35900942]
22. Baker V, Arnold G, Piot S, Thwala L, Glynn J, Hargreaves J, et al. Young adults' responses to an African and US-Based COVID-19 edutainment miniseries: real-time qualitative analysis of online social media engagement. JMIR Form Res 2021 Oct 29;5(10):e30449 [FREE Full text] [doi: 10.2196/30449] [Medline: 34596568]
23. Jana M, Letsela L, Scheepers E, Weiner R. Understanding the role of the OneLove campaign in facilitating drivers of social and behavioral change in southern Africa: a qualitative evaluation. J Health Commun 2015 Dec 13;20(3):252-258. [doi: 10.1080/10810730.2014.925014] [Medline: 25495105]
24. Makwambeni B, Salawu A. Accounting for youth audiences' resistances to HIV and AIDS messages in the television drama Tsha Tsha in South Africa. SAHARA J 2018 Dec;15(1):20-30 [FREE Full text] [doi: 10.1080/17290376.2018.1444506] [Medline: 29508641]
25. Pichon M, Buller AM, Gimunta V, Rutenge O, Thiaw Y, Sono R, et al. Qualitative evaluation of an edutainment intervention to prevent age-disparate transactional sex in Tanzania: changes in educational aspirations and gender equitable attitudes towards work. PLOS Glob Public Health 2024;4(4):e0002527 [FREE Full text] [doi: 10.1371/journal.pgph.0002527] [Medline: 38568918]
26. Cameron M, Cockcroft A, Waichigo GW, Marokoane N, Laetsang D, Andersson N. From knowledge to action: participant stories of a population health intervention to reduce gender violence and HIV in three southern African countries. AIDS Care 2014 Jul 03;26(12):1534-1540. [doi: 10.1080/09540121.2014.931560] [Medline: 24991886]
27. Kyegombe N, Zuma T, Hlongwane S, Nhlenyama M, Chimbindi N, Birdthistle I, et al. A qualitative exploration of the salience of MTV-Shuga, an edutainment programme, and adolescents' engagement with sexual and reproductive health information in rural KwaZulu-Natal, South Africa. Sex Reprod Health Matters 2022 Dec 05;30(1):2083809 [FREE Full text] [doi: 10.1080/26410397.2022.2083809] [Medline: 35929964]
28. Le Port A, Seye M, Heckert J, Peterman A, Nganya Tchamwa A, Dione M, et al. A community edutainment intervention for gender-based violence, sexual and reproductive health, and maternal and child health in rural Senegal: a process evaluation. BMC Public Health 2022 Jun 10;22(1):1165 [FREE Full text] [doi: 10.1186/s12889-022-13570-6] [Medline: 35689180]
29. Pedrana A, Hellard M, Gold J, Ata N, Chang S, Howard S, et al. Queer as F**k: reaching and engaging gay men in sexual health promotion through social networking sites. J Med Internet Res 2013 Feb 07;15(2):e25 [FREE Full text] [doi: 10.2196/jmir.2334] [Medline: 23391459]
30. Walker GR. Musicking in applied theatre: exploring interdisciplinary approaches to drama-based health and social interventions. Res Drama Educ 2023 Aug 23;29(1):187-205. [doi: 10.1080/13569783.2023.2249409]
31. Dalrymple L, Jaffe A. Dramaide, a project in schools in Kwazulu-Natal, South Africa. In: Schenker II, Sabar-Friedman G, Sy FS, editors. AIDS Education: Interventions in Multi-Cultural Societies. Boston, MA: Springer; 1996:119-123.
32. Walker GR. Emotive media as a counterbalance to AIDS messaging fatigue in South Africa: responses to an HIV/AIDS awareness music video. AIDS Educ Prev 2022 Feb;34(1):17-32. [doi: 10.1521/aeap.2022.34.1.17] [Medline: 35192396]
33. Banerjee A, Ferrara EL, Orozco-Olvera V. The entertaining way to behavioral change: fighting HIV with MTV. National Bureau of Economic Research. URL: https://www.nber.org/papers/w26096 [accessed 2024-04-29]
34. Shahmanesh M, Mthiyane N, Chimbindi N, Zuma T, Dreyer J, Birdthistle I, et al. P407 'MTV shuga': mass media communication, HSV2 and sexual health in adolescent girls and young women in rural south africa. BMJ 2019;95(Issue Suppl 1):A196. [doi: 10.1136/sextrans-2019-sti.496]
35. Hutchinson P, Beaudoin CE, Meekers D, Omoluabi E, Akinyemi A. "You need to be able to stand up for what is right": MTV Shuga Naija's transformative impact on youth attitudes toward sexual violence in Nigeria. J Interpers Violence 2024 Aug 02:8862605241265408. [doi: 10.1177/08862605241265408] [Medline: 39092629]
36. Mulwa S, Baker V, Cawood C, Khanyile D, O'Donnell D, Sarrassat S, et al. Evaluating the intensity of exposure to MTV Shuga, an edutainment program for HIV prevention: cross-sectional study in eastern cape, South Africa. JMIR Form Res 2024 Feb 13;8:e44111 [FREE Full text] [doi: 10.2196/44111] [Medline: 38349729]
37. Baker L. Observation: a complex research method. Libr Trends 2006 Jun;55(1):171-189. [doi: 10.1353/lib.2006.0045]
38. McDaniel B. Popular music reaction videos: reactivity, creator labor, and the performance of listening online. New Media Soc 2020 Jun 02;23(6):1624-1641. [doi: 10.1177/1461444820918549]
39. Goddard M. Sincere, authentic, remediated: the affective labour and cross cultural remediations of music video reaction videos on YouTube. In: Rogers H, Freitas J, Porfirio JF, editors. Remediating Sound: Repeatable Culture, YouTube and Music. New York, NY: Bloomsbury Academic; 2023:73-92.
40. Swan AL. Transnational identities and feeling in fandom: place and embodiment in K-pop fan reaction videos. Commun Cult Crit 2018;11(4):548-565. [doi: 10.1093/ccc/tcy026]
41. Budig K, Diez J, Conde P, Sastre M, Hernán M, Franco M. Photovoice and empowerment: evaluating the transformative potential of a participatory action research project. BMC Public Health 2018 Apr 02;18(1):432 [FREE Full text] [doi: 10.1186/s12889-018-5335-7] [Medline: 29609576]

XSL•FO
RenderX

42. Sadati SM, Mitchell C, Nayriro M, Muthuri S, Njeri M. Critical perspectives on photovoice as a tool to explore the challenges of balancing work and childcare among mothers in Kenya. Institute for Studies in Industrial Development. 2019. URL: https://tinyurl.com/txwvem6z [accessed 2024-04-29]
43. Asan O, Montague E. Using video-based observation research methods in primary care health encounters to evaluate complex interactions. Inform Prim Care 2014;21(4):161-170 [FREE Full text] [doi: 10.14236/jhi.v21i4.72] [Medline: 25479346]
44. Sparrman A. Video recording as interaction: participant observation of children's everyday life. Qual Res Psychol 2005 Jan;2(3):241-255. [doi: 10.1191/1478088705qp041oa]
45. Gorden RL. Interviewing: Strategy, Techniques, and Tactics. Santa Barbara, CA: Dorsey Press; 1987.
46. Glaw X, Inder K, Kable A, Hazelton M. Visual methodologies in qualitative research: autophotography and photo elicitation applied to mental health research. Int J Qual Methods 2017 Dec 19;16(1):160940691774821. [doi: 10.1177/1609406917748215]
47. Bradbury-Jones C, Isham L, Taylor J. The complexities and contradictions in participatory research with vulnerable children and young people: a qualitative systematic review. Soc Sci Med 2018 Oct;215:80-91. [doi: 10.1016/j.socscimed.2018.08.038] [Medline: 30218806]
48. Booker NA, Miller AN, Ngure P. Heavy sexual content versus safer sex content: a content analysis of the entertainment education drama Shuga. Health Commun 2016 Dec 07;31(12):1437-1446. [doi: 10.1080/10410236.2015.1077691] [Medline: 27054689]
49. Robert Walker G. Pulsing bodies and embodying pulse: musical effervescence in a South African HIV/AIDS community outreach program. Anthropol Med 2022 Sep 16;29(3):289-304. [doi: 10.1080/13648470.2021.1994335] [Medline: 34913756]

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Research
DS2: Down South season 2
PrEP: pre-exposure prophylaxis

Edited by G Eysenbach; submitted 05.03.24; peer-reviewed by C Reagon, G Walker; comments to author 21.06.24; revised version received 30.07.24; accepted 01.10.24; published 31.01.25.

Please cite as:
Baker V, Mulwa S, Khanyile D, Arnold G, Cousens S, Cawood C, Birdthistle I
Evaluating Reaction Videos of Young People Watching Edutainment Media (MTV Shuga): Qualitative Observational Study
JMIR Form Res 2025;9:e55275
URL: https://formative.jmir.org/2025/1/e55275
doi:10.2196/55275
PMID:

©Venetia Baker, Sarah Mulwa, David Khanyile, Georgia Arnold, Simon Cousens, Cherie Cawood, Isolde Birdthistle. Originally published in JMIR Formative Research (https://formative.jmir.org), 31.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Proficiency, Clarity, and Objectivity of Large Language Models Versus Specialists' Knowledge on COVID-19's Impacts in Pregnancy: Cross-Sectional Pilot Study

Nicola Luigi Bragazzi[1,2], MPH, MD, PhD; Michèle Buchinger[3], MD; Hisham Atwan[4], MD; Ruba Tuma[3,5], MD; Francesco Chirico[6], MD; Lukasz Szarpak[7,8,9], MD; Raymond Farah[3], MD; Rola Khamisy-Farah[3], MD

[1]Laboratory for Industrial and Applied Mathematics, Department of Mathematics and Statistics, York University, Toronto, ON, Canada

[2]Department of Computer Science, Data Science, and Information Technology, Faculty of Natural and Applied Sciences, Sol Plaatje University, Kimberley, South Africa

[3]Azrieli Faculty of Medicine, Bar-Ilan University, Ramat Gan, Israel

[4]Kaplan Medical Centre, Department of Internal Medicine, Hebrew University, Rehovot, Israel

[5]Department of Obstetrics and Gynecology, Galilee Medical Center, Nahariya, Israel

[6]Post-Graduate School of Occupational Health, Università Cattolica del Sacro Cuore, Rome, Italy

[7]Henry JN Taub Department of Emergency Medicine, Baylor College of Medicine, Houston, TX, United States

[8]Department of Clinical Research and Development, LUXMED Group, Warsaw, Poland

[9]Institute of Medical Science, Collegium Medicum, The John Paul II Catholic University of Lublin, Lublin, Poland

Corresponding Author:
Nicola Luigi Bragazzi, MPH, MD, PhD
Laboratory for Industrial and Applied Mathematics
Department of Mathematics and Statistics
York University
4700 Keele Street
Toronto, ON, M3J 1P3
Canada
Phone: 1 416 736 2100
Email: robertobragazzi@gmail.com

Abstract

Background: The COVID-19 pandemic has significantly strained health care systems globally, leading to an overwhelming influx of patients and exacerbating resource limitations. Concurrently, an "infodemic" of misinformation, particularly prevalent in women's health, has emerged. This challenge has been pivotal for health care providers, especially gynecologists and obstetricians, in managing pregnant women's health. The pandemic heightened risks for pregnant women from COVID-19, necessitating balanced advice from specialists on vaccine safety versus known risks. In addition, the advent of generative artificial intelligence (AI), such as large language models (LLMs), offers promising support in health care. However, they necessitate rigorous testing.

Objective: This study aimed to assess LLMs' proficiency, clarity, and objectivity regarding COVID-19's impacts on pregnancy.

Methods: This study evaluates 4 major AI prototypes (ChatGPT-3.5, ChatGPT-4, Microsoft Copilot, and Google Bard) using zero-shot prompts in a questionnaire validated among 159 Israeli gynecologists and obstetricians. The questionnaire assesses proficiency in providing accurate information on COVID-19 in relation to pregnancy. Text-mining, sentiment analysis, and readability (Flesch-Kincaid grade level and Flesch Reading Ease Score) were also conducted.

Results: In terms of LLMs' knowledge, ChatGPT-4 and Microsoft Copilot each scored 97% (32/33), Google Bard 94% (31/33), and ChatGPT-3.5 82% (27/33). ChatGPT-4 incorrectly stated an increased risk of miscarriage due to COVID-19. Google Bard and Microsoft Copilot had minor inaccuracies concerning COVID-19 transmission and complications. In the sentiment analysis, Microsoft Copilot achieved the least negative score (–4), followed by ChatGPT-4 (–6) and Google Bard (–7), while ChatGPT-3.5 obtained the most negative score (–12). Finally, concerning the readability analysis, Flesch-Kincaid Grade Level and Flesch Reading Ease Score showed that Microsoft Copilot was the most accessible at 9.9 and 49, followed by ChatGPT-4 at 12.4 and 37.1, while ChatGPT-3.5 (12.9 and 35.6) and Google Bard (12.9 and 35.8) generated particularly complex responses.

XSL•FO
RenderX

Conclusions: The study highlights varying knowledge levels of LLMs in relation to COVID-19 and pregnancy. ChatGPT-3.5 showed the least knowledge and alignment with scientific evidence. Readability and complexity analyses suggest that each AI's approach was tailored to specific audiences, with ChatGPT versions being more suitable for specialized readers and Microsoft Copilot for the general public. Sentiment analysis revealed notable variations in the way LLMs communicated critical information, underscoring the essential role of neutral and objective health care communication in ensuring that pregnant women, particularly vulnerable during the COVID-19 pandemic, receive accurate and reassuring guidance. Overall, ChatGPT-4, Microsoft Copilot, and Google Bard generally provided accurate, updated information on COVID-19 and vaccines in maternal and fetal health, aligning with health guidelines. The study demonstrated the potential role of AI in supplementing health care knowledge, with a need for continuous updating and verification of AI knowledge bases. The choice of AI tool should consider the target audience and required information detail level.

(JMIR Form Res 2025;9:e56126) doi:10.2196/56126

KEYWORDS

COVID-19; vaccine; reproductive health; generative artificial intelligence; large language model; chatGPT; google bard; microsoft copilot; vaccination; natural language processing; obstetric; gynecology; women; text mining; sentiment; accuracy; zero shot; pregnancy; readability; infectious

Introduction

The COVID-19 pandemic has exerted unprecedented strain on health care systems worldwide, overwhelming hospitals and health care facilities with patients and stretching resources and staff to their limits, thus highlighting the critical need for efficient health care delivery, and underscoring the importance of robust, resilient public health infrastructure [1].

In addition, the pandemic gave rise to an "infodemic"—that is to say, a flood of information, often misleading or false, that spread through social media and other channels [2], which has significantly impacted maternal and child health [3]. This deluge of misinformation made the task of retrieving and providing reliable, science-based information more challenging yet crucial [4] and health care providers, including gynecologists and obstetricians, played a pivotal role during the pandemic, especially in advising pregnant women.

Pregnancy, by its nature, puts women at a higher risk for complications from viral infections like COVID-19 [5,6]. As vaccines became available, gynecologists and obstetricians were at the forefront, recommending COVID-19 vaccinations to pregnant women, and having to balance the lack of initial data on vaccine safety in pregnancy with the known risks of COVID-19, providing tailored advice to their patients. The role of these specialists, crucial in safeguarding the health of mothers and their unborn children, faced significant challenges during the pandemic. This period adversely affected their professional relationships with patients and colleagues, leading to anxiety, frustration, and physical strain. These stressors, stemming from both the workplace and pandemic conditions, have prompted some gynecologists and obstetricians to consider early retirement or leaving the profession, highlighting an increased risk of burnout among these specialists [7].

The advent of generative artificial intelligence (AI), such as large language models (LLMs) and related disruptive technologies, has shown immense potential in various fields, including health care [8]. These AI-based systems hold great potential in assisting and supporting biomedical researchers, health care professionals, and the general public. They can process vast amounts of data, provide medical insights, assist in diagnostic processes, and even support therapies. In the context of the COVID-19 pandemic, AI has been used not only to analyze virus structures, predict outbreaks, and sift through vast amounts of research to identify potential treatments [9] but also to respond to queries from people, providing accurate and timely information on symptoms, prevention measures, and vaccination.

However, with the rapid development and deployment of these AI-driven technologies, there arises an essential need to test the knowledge and proficiency of AI-based systems rigorously [10]. Ensuring the accuracy, reliability, and ethical use of AI is vital, especially in sensitive areas like health care, and involves continuous monitoring, evaluation, and updating of AI systems to align with the latest scientific knowledge and ethical standards. As AI becomes more integrated into human lives, the necessity for such oversight becomes increasingly important to prevent misuse and ensure that these powerful tools serve the greater good effectively and responsibly.

Recent studies have explored the potential use of AI and LLMs in public health initiatives, particularly in the context of immunization campaigns [11-20]. These investigations have demonstrated that AI can play a critical role in countering misinformation and dismissing conspiracy theories, thereby fostering trust and acceptance of vaccines among the public. By providing clear, evidence-based responses to common myths and misconceptions, AI-driven systems can effectively address vaccine hesitancy and promote informed decision-making. However, besides being limited in number, these studies are also constrained in scope, often concentrating on individual models without providing a comprehensive comparative analysis of their performance. This lack of comparative insight leaves gaps in understanding the relative strengths and weaknesses of different AI systems in supporting public health efforts, highlighting the need for broader investigations that include a variety of generative AI tools to fully assess their potential and limitations in these critical domains.

As of today, the landscape of AI prototypes is dominated by 3 leading models, that are ChatGPT-4 (which has surpassed its predecessor, ChatGPT-3.5), Microsoft Copilot, and Google

XSL•FO
RenderX

Bard (now rebranded as Google Gemini). ChatGPT-3.5, released by OpenAI and based on the GPT-3.5 model, is a conversational AI model designed primarily for text-based interactions, and skilled in a variety of tasks like answering questions, providing explanations, and generating text based on prompts, among others. The model's knowledge is up to the date it was last trained, which means it does not have information on events or developments that occurred after that date. ChatGPT-4 represents its further iteration. As such, this advanced version by OpenAI, succeeding ChatGPT-3.5, built on the GPT-4 architecture, offers enhanced capabilities in terms of understanding and generating more complex and nuanced text, and is more contextually aware, providing more relevant and accurate responses. Like its predecessor, it is trained up to a certain point in time and lacks real-time knowledge. Microsoft Copilot is primarily focused on assisting with coding, problem-solving, and programming tasks and can be integrated with development tools like GitHub and Visual Studio Code (Microsoft Inc). Finally, Google Bard, powered by Google's own language models, including BERT, is designed to integrate Google's search capabilities, potentially offering more up-to-date information, and is aimed at providing conversational AI experiences, including answering questions or generating text, with a strong emphasis on leveraging the web for current information.

Each of these models has unique strengths and applications, with ChatGPT models (3.5 and 4) focusing on general conversational tasks, Microsoft Copilot on programming assistance, and Google Bard on integrating search capabilities into conversational AI. As such, each of these LLMs was tested in the present investigation. More in detail, the research objective of this pilot study was to evaluate the applicability of major generative AI model prototypes in providing evidence-based responses to COVID-19–related queries, particularly in the context of maternal and child health. The evaluation focused on key aspects such as accuracy, reliability, and readability. By benchmarking these models against a range of factual statements, the study aimed to assess their potential not only to support health care professionals in decision-making and public health communication during pandemics but also to address the challenges faced by the public in accessing accurate and reliable information.

The findings will help determine the role of AI in enhancing health care delivery and improving public understanding of critical health issues during crises. By highlighting the dual challenge of supporting specialists and ensuring the public is correctly informed, this research underscores the complex interplay between information dissemination and health care communication. As a pilot study, it serves as a foundational step, emphasizing the need for further, more comprehensive investigations to fully explore and validate the capabilities and limitations of AI in health care and public health contexts.

Methods

Procedure

All 4 AI-based tools were queried about COVID-19 and related vaccines during pregnancy using zero-shot prompts, that is to say, without any previous examples or context specific to that task. They were queried all on the same day, on January 7, 2024.

In a comparative evaluation of ChatGPT-3.5, ChatGPT-4, Google Bard, and Microsoft Copilot, these AI-based tools were tested using a validated questionnaire. This had previously been administered to a sample of 159 Israeli obstetrician-gynecologists, with an average age of 44.9 years (SD 12.48), predominantly female (59.7%), and largely composed of attending physicians (60.4%) [21]. This assessment aimed to evaluate the AIs' ability to provide accurate information on COVID-19, particularly in the context of pregnancy, fertility, and related clinical issues, as well as to make evidence-based recommendations. The questionnaire was developed following a comprehensive review of the literature and in accordance with World Health Organization (WHO) guidelines. It consisted of 5 sections designed to capture a range of relevant information. The first section collected sociodemographic data, including participants' gender, age, marital status, number of children, religion, level of religious observance, professional role (resident or attending physician), years of practice, as well as previous COVID-19 infection and vaccination status. The second section assessed general knowledge about COVID-19 through 5 true or false questions that addressed basic preventive measures, transmission routes, and management strategies. The third section focused on specific knowledge related to pregnancy, childbirth, and breastfeeding during the COVID-19 pandemic, including 14 questions exploring the risks and complications of COVID-19 for pregnant women, the potential for vertical transmission, the safety of breastfeeding, and related clinical practices. The fourth section addressed knowledge of COVID-19 vaccination through 14 questions that covered vaccine types, safety, efficacy, administration during pregnancy, and potential side effects. Finally, the fifth section examined attitudes and practices regarding COVID-19 vaccination. This section explored physicians' recommendations about vaccination timing during pregnancy, vaccination for women planning pregnancy, and associated safety considerations, with responses recorded on a 5-point Likert scale. To ensure clarity and relevance, the questionnaire was piloted with 12 obstetrician-gynecologists before full deployment. The survey was then distributed using both digital platforms and in-person delivery during professional gatherings, achieving a high response rate. Logistic regression analysis of the data revealed that knowledge of COVID-19 vaccination was a significant predictor of whether obstetrician-gynecologists recommended vaccination for pregnant women.

In the present comparative assessment, items from the second through fifth sections of the questionnaire were used. Sections 2 to 4 were included in their original format to specifically evaluate the AI systems' proficiency in delivering accurate knowledge about COVID-19 prevention, risks during pregnancy, and vaccination safety and efficacy. For section 5, instead of using the original Likert scale, the AI tools were queried with direct recommendation prompts such as "Would you recommend...?" The responses from section 5 were subsequently analyzed for readability and sentiment (refer to next sections),

XSL•FO
RenderX

providing insights into the clarity and tone of the AI-generated recommendations.

All the queried items, along with the corresponding prompts, are detailed in Multimedia Appendix 1.

Scoring

The scoring process of the replies provided by the four AI models (reported in Multimedia Appendix 2) was conducted independently by 2 obstetrician-gynecologists (RT and RK-F) and 1 internal medicine doctor (RF). To ensure impartiality, the biostatistician (NLB) analyzing the data was blinded to the identities of the LLMs during the scoring process.

Sentiment Analysis

Textual data analysis of AI-generated responses to the items of the fifth section of the validated questionnaire (Multimedia Appendix 3) was carried out using XLSTAT (Data Analysis and Statistical Solution for Microsoft Excel, Addinsoft) for each LLM-generated text to ensure comprehensive and model-specific insights. Each text underwent preprocessing to prepare it for analysis. This included removing English stop words, punctuation, and numerical characters, as well as filtering out words with a frequency of less than two occurrences. A sparsity threshold of 0.975 was applied to exclude terms appearing in fewer than 2.5% of the dataset, resulting in a more refined and representative corpus. Following preprocessing, a Term-Document Matrix (TDM) was constructed, providing a structured representation of the frequency of each term across the documents generated by the LLMs. The TDM facilitated comparative and statistical analyses by organizing the data into a matrix format, with columns representing unique terms and rows labeling the specific LLM. A Bag-of-Words representation was derived to enable both word frequency visualization and sentiment evaluation. Word clouds were generated to visually represent the most frequently occurring words for each LLM-generated text. The size and prominence of the words in the cloud reflected their frequency, offering an intuitive summary of the dominant themes and linguistic patterns specific to each model. To assess the emotional tone and polarity of the texts, sentiment analysis was performed using the TDM. Tokenization was applied to isolate individual words, and the Bing sentiment lexicon was used to assign binary polarity scores (positive or negative) to each term. These scores were then aggregated to calculate the overall sentiment for each LLM-generated text.

Readability Analysis

The readability of the 4 AI's responses to the items of the fifth section of the validated questionnaire (Multimedia Appendix 3) was appraised by means of the Flesch-Kincaid Grade Level analysis, which is a test designed to indicate how difficult a passage in English is to understand. The formula was developed by Rudolf Flesch and J Peter Kincaid and was first used in the United States military to assess the readability of their technical manuals. It is widely used in the field of education, but also in other sectors that require clarity of written communication, such as legal and health care industries. The formula is based on 2 factors, that are the average number of words per sentence and the average number of syllables per word. The formula is,

Flesch-Kincaid Grade Level=0.39 × (average number of words per sentence) + 11.8 × (average number of syllables per word) − 15.59

The result of the Flesch-Kincaid Grade Level formula is a number that corresponds with a US school grade level. For example, a score of 8 means that the document can be understood by an eighth grader, with a higher score indicating more complex and difficult text.

In addition, the Flesch Reading Ease Score was also used to evaluate the readability of the responses. This score rates text on a 100-point scale, where higher scores indicate easier-to-read material. The formula for calculating the Flesch Reading Ease Score is Flesch Reading Ease Score = 206.835 − 1.015 × (average number of words per sentence) − 84.6 × (average number of syllables per word)

Higher scores on this scale indicate easier-to-read text, while lower scores indicate more complex material. For example, scores above 90 suggest very easy text suitable for an average 10- to 11-year-old (5th grade), while scores below 30 indicate very difficult text meant for college graduates.

Ethical Considerations

The primary study [21], which provided data from human participants, was exempted from formal approval by the Ethics Committee of the Medical Faculty at Bar-Ilan University in Tel Aviv, Israel, due to its survey-based design. Participants, consisting of obstetrician-gynecologists (both residents and attending physicians), voluntarily consented to participate in the study without receiving any compensation. All data were fully anonymized, ensuring the privacy and confidentiality of the participants. In addition, no ethical approval was required for the AI component of this study.

Results

LLMs' Knowledge of COVID-19 and Its Vaccines for Maternal and Fetal Health

ChatGPT-4 and Microsoft Copilot obtained a score of 97% (32/33), while Google Bard scored 94% (31/33), and ChatGPT-3.5 scored 82% (27/33; Multimedia Appendix 3).

ChatGPT-4 delivered correct responses across a range of topics, demonstrating a high level of proficiency in this specialized field. However, there was a notable exception in its performance: ChatGPT incorrectly stated that there is a heightened risk of miscarriage among pregnant women with COVID-19, even though, in reality, according to current literature, there is no evidence to support an increased risk of miscarriage due to COVID-19.

Similarly, Microsoft Copilot made only one mistake, related to the item concerning the increased risk of complications for pregnant women with COVID-19 compared to nonpregnant women of the same age. Google Bard's knowledge of COVID-19 and its vaccines for women's health was rather good, with the AI-based tool making only 2 mistakes, by ruling out COVID-19 intrauterine transmission, transmission during

XSL•FO
RenderX

delivery, or through contact and respiratory droplets during breastfeeding.

On the contrary, ChatGPT-3.5 exhibited poorer performance, with 6 wrong responses, demonstrating gaps in the knowledge of (1) risk, (2) complications of COVID-19 in pregnant women, (3) medication for pregnant and postpartum women with COVID-19, (4) risk of miscarriage in COVID-19–infected pregnant women, (5) mother-to-child transmission of COVID-19, and (6) COVID-19 transmission through breast milk.

LLMs' Attitudes and Practices Toward COVID-19 and Its Vaccines for Maternal and Fetal Health

Even if ChatGPT-3.5, ChatGPT-4, Microsoft Copilot, and Google Bard, as AI-based LLMs, cannot and do not provide personal opinions or medical recommendations, they can offer up-to-date and reliable information based on current guidelines and scientific understanding regarding COVID-19 vaccinations. In particular, concerning the recommendation of COVID-19 vaccine to all pregnant women (without contraindication), ChatGPT-4 stated that many health organizations, including the Centers for Disease Control and Prevention and the WHO, generally recommend COVID-19 vaccination for pregnant women due to the increased risk of severe illness from COVID-19 during pregnancy (including hospitalization, intensive care unit admission, and death). Google Bard concurred, emphasizing that the benefits of COVID-19 immunization far outweigh any potential risks to the mother or baby.

In terms of the proper timing for recommending the immunization of pregnant women against COVID-19, the AI-based chatbot clearly stated that the recommendation for COVID-19 vaccination during pregnancy often applies to all trimesters and, especially in the case of doubts, the timing should be discussed with a health care provider. Google Bard added that there is no evidence that vaccination at any particular trimester poses a higher risk to the mother or baby and Copilot stated that emphasis should be on vaccine receipt as soon as possible to maximize maternal and fetal health.

When queried about the recommendation of the COVID-19 vaccine only to pregnant women at high risk of contracting the virus, both ChatGPT-4 and Google Bard posited that the current guidance from many health organizations is to offer the vaccine to all pregnant women, not just those at high risk, due to the potentially severe impacts of COVID-19 during pregnancy.

Concerning the recommendation of COVID-19 vaccine to all women (without contraindication) of reproductive age who are not pregnant, ChatGPT-4 advised that COVID-19 vaccination is widely recommended for women of reproductive age to prevent severe illness, complications, and potential long-term effects of COVID-19. For those planning to undergo assisted reproduction, according to both the 2 AI-based tools, they are generally advised to get vaccinated against COVID-19, as there is no evidence that vaccines can affect fertility. When queried about the recommended interval between vaccination and pregnancy, the OpenAI chatbot stated that there is no recommended interval between receiving a COVID-19 vaccine and becoming pregnant according to most health organizations. Google Bard emphasized that there is no clear consensus on whether there is a need for an interval between vaccination and pregnancy, with some experts believing that there may be a slight increase in the risk of miscarriage in the first few weeks after vaccination, but this risk is very small, and other experts believing that there is no need to wait for any period of time before trying to conceive after vaccination. According to Google Bard, more research is warranted to determine whether there is a need for an interval between vaccination and pregnancy.

Finally, when tasked with recommending an interval between COVID-19 vaccination and the use of assisted reproductive technology (ART), ChatGPT-4 clarified that there is no specified interval recommended between COVID-19 vaccination and undergoing ART, while, according to Google Bard, there is no need for an interval between vaccination and ART, in that COVID-19 vaccination can safely be administered to women undergoing ART at any time.

On the contrary, ChatGPT-3.5 failed to offer up-to-date and reliable information, recommending immunization against COVID-19 in pregnant women at high risk, especially during the third trimester as it would be safer than during the first and second trimesters. Furthermore, ChatGPT-3.5 recommended an interval between vaccination and pregnancy as well as between immunization and ART based on vaccine type and individual risk factors.

LLMs' Text-Mining Analysis

The comparative text-mining analysis of Google Bard, Microsoft Copilot, ChatGPT-3.5, and ChatGPT-4 revealed both similarities and differences in the vocabulary used when discussing COVID-19 and immunization in relation to pregnancy. While all the models highlighted core themes such as "pregnant," "COVID-19," "vaccination," and "women," the nuances in their focus and vocabulary indicated variations in their approach to presenting the information. Google Bard and Microsoft Copilot both emphasized terms related to pregnancy and COVID-19, but their focal points diverged. Google Bard included more detailed terms associated with vaccination recommendations and the effects of COVID-19 on pregnancy outcomes, suggesting an emphasis on clinical guidance. In contrast, Microsoft Copilot featured broader terminology, incorporating aspects like "risk" and "breastfeeding," indicating a wider scope that touches on future planning and broader maternal and fetal health. When comparing Google Bard to ChatGPT-3.5 and ChatGPT-4, shared terms like "pregnant" and "vaccination" pointed to a shared understanding of key themes. However, Google Bard prioritized actionable guidance and recommendations, while ChatGPT variants included a more detailed vocabulary concerning medical and health-related risks. This distinction highlighted Google Bard's focus on practical advice, compared to ChatGPT's detailed explanations of risks and safety. Microsoft Copilot and the ChatGPT versions also shared several core terms but differed in their linguistic styles. Microsoft Copilot tended to lean toward broader generalizations, which makes it suitable for summarizing information succinctly. On the other hand, ChatGPT versions delved into specifics, incorporating medical and technical language to provide more

XSL•FO
RenderX

comprehensive information about COVID-19 and its impact on pregnancy. This distinction reflected the contrast between Copilot's generalist approach and ChatGPT's detailed focus. The comparison between ChatGPT-3.5 and ChatGPT-4 revealed a significant overlap in vocabulary, indicating consistency between the 2 versions. However, ChatGPT-4 introduced additional terms and nuances, suggesting an evolution in depth and breadth of content. These differences highlighted improvements in ChatGPT-4's ability to provide detailed and nuanced medical insights compared to its predecessor.

LLMs' Readability Analysis

The Flesch-Kincaid Grade Level and Flesch Reading Ease Score for Google Bard's response were 12.9 and 35.8, respectively, indicating that the text was highly complex and best understood by individuals with postsecondary education. Microsoft Copilot scored 9.9 and 49, suggesting a less complex response that was easier to read. Among the ChatGPT models, ChatGPT-3.5 exhibited a similar level of complexity to Google Bard, with scores of 12.9 and 35.6. ChatGPT-4, with scores of 12.4 and 37.1, was slightly less complex than both ChatGPT-3.5 and Google Bard but more complex than Microsoft Copilot. Overall, ChatGPT responses were best suited for readers with advanced reading skills, with ChatGPT-3.5 requiring graduate-level comprehension and ChatGPT-4 being more appropriate for advanced undergraduate-level readers (Table 1).

Table 1. Readability analysis for each large language model tested (Google Bard, Microsoft Copilot, ChatGPT-3.5, and ChatGPT-4).

AI Model	Flesch-Kincaid Grade Level	Flesch Reading Ease Score	Reading Complexity	Target Audience
Google Bard	12.9	35.8	Highly complex	Post-secondary education
Microsoft Copilot	9.9	49	Less complex	High school to early undergraduate level
ChatGPT-3.5	12.9	35.6	Similar complexity to Google Bard	Graduate-level comprehension
ChatGPT-4	12.4	37.1	Slightly less complex than ChatGPT-3.5 and Bard	Advanced undergraduate-level comprehension

LLMs' Sentiment Analysis

In the sentiment analysis, Microsoft Copilot had the least negative score (–4), reflecting a neutral and balanced tone. ChatGPT-4 scored (–6), slightly more negative due to its detailed discussion of risks. Google Bard followed with (–7), emphasizing practical recommendations that highlighted potential challenges. ChatGPT-3.5 had the most negative score (–12), reflecting a strong focus on risks and adverse outcomes. These differences, shown in Figure 1, highlight the tonal variations among the models, with Microsoft Copilot being the most neutral and ChatGPT-3.5 the most risk-oriented.

Figure 1. Sentiment analyses for Google Bard, Microsoft Copilot, ChatGPT-3.5, and ChatGPT-4.

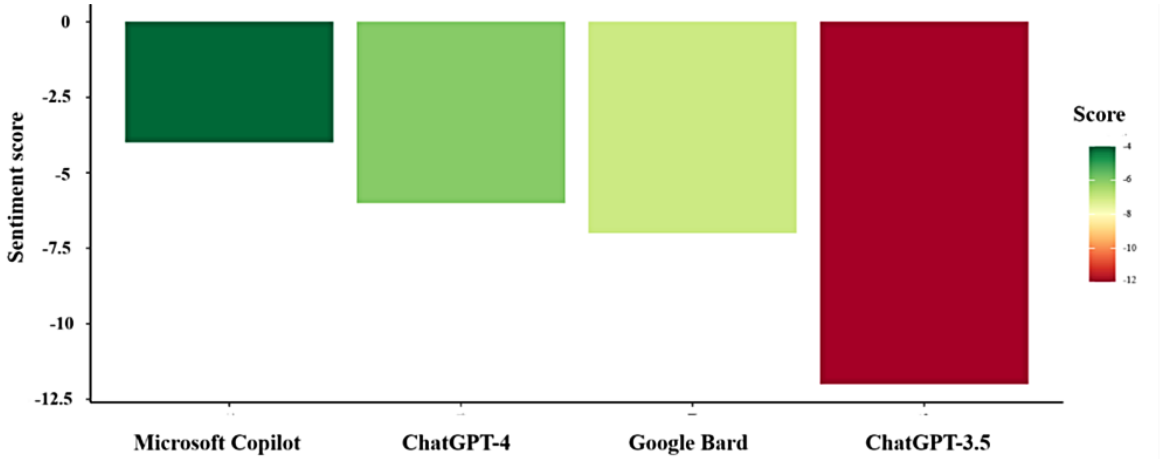

Discussion

Overview of Major Study Findings

This study comprehensively assessed the knowledge, clarity, and objectivity of 4 prominent LLMs (ChatGPT-3.5, ChatGPT-4, Microsoft Copilot, and Google Bard) regarding COVID-19 impacts on pregnancy. Aligning with the stated aims, the findings revealed distinct performance variations among the LLMs. ChatGPT-4 and Microsoft Copilot demonstrated the highest levels of accuracy (97%, 32/33), followed by Google Bard (94%, 31/33) and ChatGPT-3.5 (82%, 27/33). While all models provided generally reliable information on COVID-19 and vaccination in maternal health, they varied in readability, with Microsoft Copilot being the most accessible and ChatGPT-3.5 the least comprehensible. Sentiment analysis highlighted Microsoft Copilot as the most neutral, whereas ChatGPT-3.5 displayed a risk-heavy tone. These outcomes underscored the evolving potential of LLMs to support health care professionals and public health communication during global crises while emphasizing the necessity for ongoing

XSL•FO
RenderX

evaluation and model refinement to align with the latest scientific evidence.

The Four LLMs Differ in Terms of Knowledge, With ChatGPT-3.5 Being the Least Knowledgeable

The performance of AI models such as ChatGPT-4, Google Bard, and Microsoft Copilot proved exceptional, with only minor errors detected. Their capacity to deliver accurate and detailed information showcased a depth of knowledge that is particularly impressive when juxtaposed with the level of COVID-19-related expertise observed among obstetrics and gynecology specialists. This comparison is drawn from findings in a previous survey [21], which revealed an average knowledge score of 75.6 (SD 10.6) among healthcare professionals in the field. Notably, while 81.1% (129/159) of these specialists displayed adequate knowledge of general COVID-19 information, a much smaller percentage (19/159, 11.9%) demonstrated specific understanding related to pregnancy, childbirth, and breastfeeding in the context of the pandemic. Furthermore, only 40.3% (64/159) of respondents showed familiarity with COVID-19 vaccination guidelines for pregnant individuals. A particularly critical area of knowledge, the increased risk of preeclampsia in pregnant women infected with COVID-19, was correctly identified by only 27% (43/159) of participants. This highlights a significant gap in awareness that could impact patient care. Regarding immunization practices, about 65% (104/159) of respondents recommended COVID-19 vaccination during all trimesters, reflecting a broad acceptance of its safety and efficacy across pregnancy stages. However, preferences for specific trimesters varied: about 25% (39/159) recommended vaccination during the second trimester, 6% (9/159) in the third trimester, and a minority of 4% (6/159) in the first trimester.

These findings align with existing literature, which consistently highlights a trend where ChatGPT-3 underperformed compared to other LLMs in specific tasks or benchmarks. Studies have noted that while ChatGPT-3 demonstrated proficiency in generating coherent text and answering general questions, it often fell short in accuracy, contextual understanding, and handling of complex queries. This performance gap has been attributed to differences in training data, model architecture, and parameter tuning [22,23]. Conversely, studies have shown that more advanced LLMs, such as ChatGPT-4 and Microsoft Copilot, outperformed ChatGPT-3, with Google Bard following closely behind [24], consistent with our findings.

However, other research suggests that ChatGPT-3.5 performed comparably or even better than other LLMs in certain contexts, potentially indicating that performance may vary depending on the specific application or field of use [25].

The Four LLMs Differ in Terms of Recommendations of COVID-19 Vaccination During Pregnancy, With ChatGPT-3.5 Being the Least Aligned With Scientific Evidence

When comparing the information provided about COVID-19 in pregnant women and the vaccination during pregnancy, there were some notable variations, especially in alignment with scientific evidence. Concerning the risk of severe illness, Google Bard highlighted an increased risk of severe complications, including hospitalization, intensive care unit admission, and death for pregnant women. In addition, it pointed out a higher likelihood of preterm birth. Microsoft Copilot emphasized that, while pregnant women are not more likely to contract the virus, they face a greater risk of severe disease, especially if infected in the third trimester. ChatGPT-3.5 noted only a slightly higher risk of severe illness for pregnant women compared to nonpregnant individuals, including increased hospitalization and intensive care unit admission rates, while, more correctly, ChatGPT-4 stated that pregnant women with COVID-19 have a higher risk of severe illness and complications, such as preterm birth. Regarding COVID-19 vaccination during pregnancy, Google Bard strongly advocated for COVID-19 immunization during pregnancy, highlighting its safety and effectiveness in preventing serious illness and protecting both mother and baby. Microsoft Copilot stressed the importance of vaccination for those women who are pregnant, breastfeeding, or planning pregnancy, emphasizing its safety and efficacy. ChatGPT-3.5 discussed the safety of Messenger Ribonucleic Acid vaccines for pregnant individuals and suggested discussing the timing of vaccination with health care providers, while ChatGPT-4 recommended COVID-19 vaccines for pregnant individuals, noting their safety and efficacy and also advised discussing vaccination timing with health care providers. In terms of vaccine safety and side effects, Google Bard mentioned that the side effects of vaccination in pregnant women are similar to those in nonpregnant adults and are generally mild and short-lived. Microsoft Copilot, ChatGPT-3.5, and ChatGPT-4 all emphasized the safety of Messenger Ribonucleic Acid COVID-19 vaccines during pregnancy and did not report any serious adverse maternal or fetal effects. Finally, Google Bard advised pregnant women to get vaccinated as soon as possible after the first trimester, whereas ChatGPT-4 suggested postvaccination monitoring for pregnant women through programs like Centers for Disease Control and Prevention's V-safe and emphasized the importance of consulting healthcare providers. All sources agreed on the increased risk of severe COVID-19 in pregnant women and the benefits and safety of vaccination during pregnancy. However, they differed in their emphasis on specific risks and considerations.

The comparison of LLMs in providing information about COVID-19 in pregnant women and vaccination during pregnancy underscores their potential to disseminate health information effectively but also highlights significant variability in alignment with scientific evidence. This variability raises concerns about the accuracy and reliability of AI-generated health content, particularly when subtle differences in emphasis or framing could influence public perception and decision-making. The findings illustrate how LLMs like ChatGPT, Microsoft Copilot, and Google Bard can either reinforce or undermine public health messaging, depending on their training data and underlying algorithms. These differences emphasize the need for rigorous evaluation and validation of AI-generated content in healthcare contexts, as inconsistencies could amplify the risk of misinformation, contributing to the emerging "AI-driven infodemic" [26]. Establishing robust mechanisms for quality control and promoting transparency in

XSL•FO
RenderX

AI outputs is imperative to mitigate these risks and ensure that LLMs support, rather than hinder, public health objectives.

The Four LLMs Differ in Terms of Vocabulary, Clarity, and Readability, With ChatGPT-3.5 Being the Most Difficult to Understand

From the text-mining analysis, overall, while there was a significant overlap in key themes across all four sources, each one brought its unique focus and depth to the discussion. The unique terms in each source highlighted their different approaches to discussing the same topic, reflecting the diversity in how information can be presented and prioritized. Google Bard and Microsoft Copilot tended to have a more general approach, while ChatGPT versions provided a more detailed, medically-oriented perspective. In particular, ChatGPT-3.5, being excessively technical, is particularly difficult to understand, as also reflected by the Flesch-Kincaid grade level analysis.

The findings from our text-mining analysis align closely with a comparative study [27], which showed that ChatGPT-3.5 tended to generate overly technical content that may hinder readability.

The differences in readability and complexity across LLMs are important when considering the target audience for the information [28-30]. For a general audience, a lower Flesch-Kincaid Grade Level might be more appropriate to ensure wider comprehension. On the other hand, for professional or academic audiences, a higher grade level could be more suitable, as it allows for the conveyance of complex ideas using specialized vocabulary and detailed explanations. Each LLM's approach reflects its unique language processing capabilities and potentially the nature of the training data it was exposed to. It is of crucial importance for users to consider the context and audience when evaluating or choosing the most appropriate AI tool for generating text, with ChatGPT's and Microsoft Copilot's approach appearing more beneficial for specialist or academic audiences and for the general public, respectively.

The Four LLMs Differ in Terms of Sentiment and Objectivity, With ChatGPT-3.5 Being the Least Factual

In the sentiment analysis, the tools demonstrated varying degrees of negativity in their language when discussing COVID-19 and pregnancy-related topics. Microsoft Copilot achieved the least negative score (–4), indicating a relatively neutral or balanced tone compared to its counterparts. This suggests that Microsoft Copilot's responses tended to minimize emotionally charged or negative language, focusing instead on providing concise, objective, and factual information. ChatGPT-4 followed with a slightly more negative score of –6, reflecting a balanced but somewhat more cautious tone. This could be attributed to ChatGPT-4's detailed discussion of potential risks and complications, which, while comprehensive, likely included language that leaned toward highlighting challenges and considerations in the context of COVID-19 and pregnancy. Google Bard scored –7, placing it closer to ChatGPT-4 in terms of sentiment but slightly more negative. This may be indicative of its focus on delivering actionable recommendations and clinical insights, which, while practical, may also involve language emphasizing adverse effects or potential risks, thereby increasing its overall negativity score. ChatGPT-3.5 obtained the most negative score of –12, representing a significantly more cautious or risk-focused tone compared to the other models. This heightened negativity could reflect the model's tendency to elaborate extensively on risks and adverse outcomes, potentially overemphasizing negative aspects in its explanations.

These differences in sentiment scores highlight the varying tonal approaches of the AI models. Microsoft Copilot's minimal negativity suggests a focus on general, less emotionally charged language, while ChatGPT-3.5's highest negativity score underscores its more risk-heavy narrative. ChatGPT-4 and Google Bard occupy the middle ground, balancing detailed explanations with practical recommendations, albeit with some cautionary emphasis. This variation in sentiment reflects not only the linguistic choices of the models but also their inherent design priorities, ranging from objectivity and balance to detailed elaboration and actionable advice.

Strengths and Limitations

This study has several notable strengths. First, the use of a validated questionnaire ensures a robust and reliable assessment of LLMs' proficiency, clarity, and objectivity. The evaluation spans multiple dimensions, including knowledge, sentiment, and readability, offering a comprehensive analysis of the capabilities of ChatGPT-3.5, ChatGPT-4, Google Bard, and Microsoft Copilot. In addition, the comparison against specialist knowledge provides a meaningful benchmark, grounding the findings in real-world relevance. The inclusion of sentiment and readability analyses highlights the accessibility and emotional tone of AI-generated responses, further contextualizing their suitability for different audiences. Moreover, the study addresses a critical gap in assessing AI tools in the context of obstetrics, gynecology, and public health during a pandemic—a time when accurate information is essential. By incorporating text-mining and readability metrics, the research elucidates how LLMs can serve diverse user bases, ranging from healthcare professionals to general audiences. However, this study is not without its limitations. A key limitation is the static nature of the evaluation, as LLMs undergo continuous updates and learning processes. This dynamic evolution may render some of the current observations obsolete over time, necessitating periodic re-evaluation. Furthermore, the study focuses on a specific application of LLMs (COVID-19's impacts in pregnancy), which may not generalize to other medical or public health domains. The zero-shot prompting approach, while simulating real-world user queries, may not fully capture the potential of these models when used with tailored or fine-tuned inputs. Finally, the study is limited by its reliance on the English language and the specific design of the validated questionnaire. Future research could explore the multilingual capabilities of LLMs and their application to other health care domains, as well as integrate expert feedback during iterative model evaluation.

Future Directions

The present comprehensive analysis covering various aspects (knowledge, recommendations, vocabulary and readability,

XSL•FO
RenderX

sentiment, and objectivity) shows similarities and dissimilarities among four major LLMs.

The knowledge appraisal highlights a gap in awareness of COVID-19 and immunization during pregnancy among specialists, underscoring the potential role of AI like ChatGPT in supplementing healthcare knowledge and decision-making. Generative AI holds great potential in health care and global public health, serving as valuable tools for professionals, providing quick access to updated information, and filling knowledge gaps.

However, some discrepancies in ChatGPT-3.5's responses also serve as a reminder of the need for continuous updating and verification of AI knowledge bases. This ensures that the information provided is aligned with the latest research and clinical guidelines. As AI systems become more integrated into health care, their ability to adapt and learn from emerging data will be crucial in ensuring they remain reliable and beneficial adjuncts in medical practice [31,32].

Future research directions should consider the inclusion of emerging AI models to evaluate their comparative utility in clinical settings. Investigating the effectiveness of these tools across various medical specialties could shed light on their strengths and limitations. Studies assessing the real-world integration of LLMs into health care workflows, focusing on usability, clinician satisfaction, and patient outcomes, are also needed. In addition, exploring the ethical and legal implications of using AI in clinical practice (such as ensuring patient privacy, safeguarding data security, and clarifying liability in decision-making) remains a crucial area of inquiry. Longitudinal studies on the influence of AI tools on medical education, professional development, health care delivery, and communication would provide valuable insights into their sustained impact.

Overall, the analysis underscores the importance of choosing the right LLM for specific needs. While ChatGPT-4 and newer versions offer more detailed and updated information, earlier versions like ChatGPT-3.5 might present challenges in terms of comprehensibility. The choice between these LLMs should be influenced by the target audience and the level of detail and technicality required in the information. For general audiences, simpler language (as seen in Microsoft Copilot) is preferable, whereas more specialized audiences might benefit from the detailed approach of ChatGPT-4. For instance, in vaccine management, ChatGPT-4 can support specialized audiences by aiding clinical decision-making, providing detailed research insights, and tailoring vaccination guidelines for pregnant women. Microsoft Copilot excels in communicating simplified information to general audiences, making it ideal for educating patients, training health workers, or addressing vaccine hesitancy. Google Bard, with its real-time search integration, is suited for disseminating up-to-date policies and driving public health campaigns with clear, actionable messaging. ChatGPT-3.5, despite its limitations, can be used for simulating scenarios or gathering feedback to refine communication strategies. These applications demonstrate how aligning LLMs' capabilities with target audiences ensures effective vaccine management and can be potentially translated to other clinical settings and patient management contexts.

Conclusions

This study highlighted the strengths and limitations of LLMs such as ChatGPT-3.5, ChatGPT-4, Microsoft Copilot, and Google Bard in addressing critical healthcare topics, specifically the impacts of COVID-19 on pregnancy. While the findings underscored the potential of LLMs to provide accurate, up-to-date information, they also revealed gaps in their knowledge and variability in their communication styles, sentiment, and readability. Such variability underscores the importance of tailoring AI use to specific audiences and contexts. Beyond the immediate findings, the broader implications of this research emphasize the transformative potential of LLMs in health care and public health communication. These AI tools can bridge knowledge gaps among health care professionals, support evidence-based decision-making, and combat misinformation during health crises. Their ability to deliver clear, accessible, and accurate information positions them as valuable adjuncts in both clinical settings and public health campaigns. However, the study also underscores critical ethical and operational considerations. Ensuring the reliability of AI-generated content requires continuous updates to training datasets and rigorous validation against the latest scientific evidence. The dynamic nature of AI development means that its integration into health care systems must be accompanied by robust oversight mechanisms to mitigate risks such as the dissemination of misinformation or unintentional biases. Furthermore, the deployment of LLMs in health care highlights the need for interdisciplinary collaboration among AI developers, health care professionals, and policy makers. Such partnerships can ensure that these tools are designed and utilized in ways that prioritize patient safety, data privacy, and equitable access to information. By fostering these collaborations, the integration of AI into health care can move beyond theoretical potential to practical, impactful applications. In conclusion, while LLMs like ChatGPT-4 and Microsoft Copilot demonstrate remarkable capabilities in providing health care–related information, their limitations remind us of the essential role of human oversight. The findings of this study advocate for a future where AI complements, rather than replaces, human expertise in healthcare, enhancing both the quality and accessibility of medical knowledge. This balanced approach will be key to leveraging AI's full potential while safeguarding the ethical and practical dimensions of health care delivery.

Data Availability

All data are available in Multimedia Appendices 1-3.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Overview of queries and prompts used with Google Bard, Microsoft Copilot, ChatGPT-3.5, and ChatGPT-4.
[DOCX File , 16 KB - formative_v9i1e56126_app1.docx]

Multimedia Appendix 2
Responses provided by each large language model.
[DOCX File , 22 KB - formative_v9i1e56126_app2.docx]

Multimedia Appendix 3
Responses generated by each large language model (Google Bard, Microsoft Copilot, ChatGPT-3.5, and ChatGPT-4) when prompted to provide recommendations on COVID-19 vaccination for pregnant women.
[DOCX File , 16 KB - formative_v9i1e56126_app3.docx]

References

1. Filip R, Gheorghita Puscaselu R, Anchidin-Norocel L, Dimian M, Savage WK. Global challenges to public health care systems during the COVID-19 pandemic: a review of pandemic measures and problems. J Pers Med 2022;12(8):1295 [FREE Full text] [doi: 10.3390/jpm12081295] [Medline: 36013244]
2. Forati AM, Ghose R. Geospatial analysis of misinformation in COVID-19 related tweets. Appl Geogr 2021;133:102473 [FREE Full text] [doi: 10.1016/j.apgeog.2021.102473] [Medline: 34103772]
3. Marquini GV, Martins SB, Oliveira LM, Dias MM, Takano CC, Sartori MGF. Effects of the COVID-19 pandemic on gynecological health: an integrative review. Rev Bras Ginecol Obstet 2022;44(2):194-200 [FREE Full text] [doi: 10.1055/s-0042-1742294] [Medline: 35213918]
4. Alberca RW, Pereira NZ, Oliveira LMDS, Gozzi-Silva SC, Sato MN. Pregnancy, viral infection, and COVID-19. Front Immunol 2020;11:1672 [FREE Full text] [doi: 10.3389/fimmu.2020.01672] [Medline: 32733490]
5. Hanna N, Hanna M, Sharma S. Is pregnancy an immunological contributor to severe or controlled COVID-19 disease? Am J Reprod Immunol 2020;84(5):e13317 [FREE Full text] [doi: 10.1111/aji.13317] [Medline: 32757366]
6. Blakeway H, Prasad S, Kalafat E, Heath PT, Ladhani SN, Le Doare K, et al. COVID-19 vaccination during pregnancy: coverage and safety. Am J Obstet Gynecol 2022;226(2):236.e1-236.e14 [FREE Full text] [doi: 10.1016/j.ajog.2021.08.007] [Medline: 34389291]
7. Riggan KA, Reckhow J, Allyse MA, Long M, Torbenson V, Rivera-Chiauzzi EY. Impact of the COVID-19 pandemic on obstetricians/gynecologists. Mayo Clin Proc Innov Qual Outcomes 2021;5(6):1128-1137 [FREE Full text] [doi: 10.1016/j.mayocpiqo.2021.11.002] [Medline: 34849457]
8. Khamisy-Farah R, Furstenau LB, Kong JD, Wu J, Bragazzi NL. Gynecology meets big data in the disruptive innovation medical era: state-of-art and future prospects. Int J Environ Res Public Health 2021;18(10):5058 [FREE Full text] [doi: 10.3390/ijerph18105058] [Medline: 34064710]
9. Bragazzi NL, Dai H, Damiani G, Behzadifar M, Martini M, Wu J. How big data and artificial intelligence can help better manage the COVID-19 pandemic. Int J Environ Res Public Health 2020;17(9):3176 [FREE Full text] [doi: 10.3390/ijerph17093176] [Medline: 32370204]
10. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature 2023;620(7972):172-180 [FREE Full text] [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
11. Joshi S, Ha E, Amaya A, Mendoza M, Rivera Y, Singh VK. Ensuring accuracy and equity in vaccination information from ChatGPT and CDC: mixed-methods cross-language evaluation. JMIR Form Res 2024;8:e60939 [FREE Full text] [doi: 10.2196/60939] [Medline: 39476380]
12. Fiore M, Bianconi A, Acuti Martellucci C, Rosso A, Zauli E, Flacco ME, et al. Vaccination hesitancy: agreement between WHO and ChatGPT-4.0 or Gemini Advanced. Ann Ig 2024 [FREE Full text] [doi: 10.7416/ai.2024.2657] [Medline: 39373234]
13. Kim S, Kim K, Wonjeong Jo C. Accuracy of a large language model in distinguishing anti- and pro-vaccination messages on social media: the case of human papillomavirus vaccination. Prev Med Rep 2024;42:102723 [FREE Full text] [doi: 10.1016/j.pmedr.2024.102723] [Medline: 38659997]
14. Deiana G, Dettori M, Arghittu A, Azara A, Gabutti G, Castiglia P. Artificial intelligence and public health: evaluating ChatGPT responses to vaccination myths and misconceptions. Vaccines (Basel) 2023;11(7):1217 [FREE Full text] [doi: 10.3390/vaccines11071217] [Medline: 37515033]
15. Sohail SS, Madsen D, Farhat F, Alam MA. ChatGPT and vaccines: can AI chatbots boost awareness and uptake? Ann Biomed Eng 2024;52(3):446-450. [doi: 10.1007/s10439-023-03305-y] [Medline: 37428336]

XSL•FO
RenderX

16. Sallam M, Salim NA, Al-Tammemi AB, Barakat M, Fayyad D, Hallit S, et al. ChatGPT output regarding compulsory vaccination and COVID-19 vaccine conspiracy: a descriptive study at the outset of a paradigm shift in online search for information. Cureus 2023;15(2):e35029 [FREE Full text] [doi: 10.7759/cureus.35029] [Medline: 36819954]
17. Tekinay ON. Curious questions about Covid-19 pandemic with ChatGPT: answers and recommendations. Ann Biomed Eng 2023;51(7):1371-1373 [FREE Full text] [doi: 10.1007/s10439-023-03209-x] [Medline: 37071282]
18. Cheng K, Li Z, He Y, Guo Q, Lu Y, Gu S, et al. Potential use of artificial intelligence in infectious disease: take ChatGPT as an example. Ann Biomed Eng 2023;51(6):1130-1135 [FREE Full text] [doi: 10.1007/s10439-023-03203-3] [Medline: 37074486]
19. Alhasan K, Al-Tawfiq J, Aljamaan F, Jamal A, Al-Eyadhy A, Temsah M. Mitigating the burden of severe pediatric respiratory viruses in the post-COVID-19 era: ChatGPT insights and recommendations. Cureus 2023;15(3):e36263 [FREE Full text] [doi: 10.7759/cureus.36263] [Medline: 37073200]
20. Cheng K, Wu C, Gu S, Lu Y, Wu H, Li C. WHO declares the end of the COVID-19 global health emergency: lessons and recommendations from the perspective of ChatGPT/GPT-4. Int J Surg 2023;109(9):2859-2862 [FREE Full text] [doi: 10.1097/JS9.0000000000000521] [Medline: 37246993]
21. Bragazzi NL, Buchinger M, Szarpak L, Chmielewski J, Goździewska M, Gotlib J, et al. Knowledge, attitudes, and practices towards COVID-19 and COVID-19 vaccines among obstetrician-gynaecologists. Ann Agric Environ Med 2023;30(4):669-676 [FREE Full text] [doi: 10.26444/aaem/176959] [Medline: 38153070]
22. Zare S, Vafaeian S, Amini M, Farhadi K, Vali M, Golestani A. Comparing the performance of ChatGPT-3.5-Turbo, ChatGPT-4, and Google Bard with Iranian students in pre-internship comprehensive exams. Sci Rep 2024;14(1):28456 [FREE Full text] [doi: 10.1038/s41598-024-79335-w] [Medline: 39557926]
23. Gill GS, Blair J, Litinsky S. Evaluating the performance of ChatGPT 3.5 and 4.0 on statpearls oculoplastic surgery text- and image-based exam questions. Cureus 2024;16(11):e73812. [doi: 10.7759/cureus.73812] [Medline: 39691123]
24. Rossettini G, Rodeghiero L, Corradi F, Cook C, Pillastrini P, Turolla A, et al. Comparative accuracy of ChatGPT-4, microsoft copilot and Google Gemini in the Italian entrance test for healthcare sciences degrees: a cross-sectional study. BMC Med Educ 2024;24(1):694 [FREE Full text] [doi: 10.1186/s12909-024-05630-9] [Medline: 38926809]
25. Ponzo V, Rosato R, Scigliano MC, Onida M, Cossai S, De Vecchi M, et al. Comparison of the accuracy, completeness, reproducibility, and consistency of different AI chatbots in providing nutritional advice: an exploratory study. J Clin Med 2024;13(24):7810 [FREE Full text] [doi: 10.3390/jcm13247810] [Medline: 39768733]
26. De Angelis L, Baglivo F, Arzilli G, Privitera GP, Ferragina P, Tozzi AE, et al. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. Front Public Health 2023;11:1166120 [FREE Full text] [doi: 10.3389/fpubh.2023.1166120] [Medline: 37181697]
27. Marulli F, Campanile L, de Biase MS, Marrone S, Verde L, Bifulco M. Understanding readability of large language models output: an empirical analysis. Procedia Comput. Sci 2024;246:5273-5282. [doi: 10.1016/j.procs.2024.09.636]
28. Amin K, Doshi R, Forman HP. Large language models as a source of health information: are they patient-centered? A longitudinal analysis. Healthc (Amst) 2024;12(1):100731. [doi: 10.1016/j.hjdsi.2023.100731] [Medline: 38141269]
29. Tu W, Joe BN. The Era of ChatGPT and large language models: can we advance patient-centered communications appropriately and safely? Radiol Imaging Cancer 2024;6(3):e240038 [FREE Full text] [doi: 10.1148/rycan.240038] [Medline: 38668641]
30. Howell MD. Generative artificial intelligence, patient safety and healthcare quality: a review. BMJ Qual Saf 2024;33(11):748-754 [FREE Full text] [doi: 10.1136/bmjqs-2023-016690] [Medline: 39048292]
31. Lin Z, Guan S, Zhang W, Zhang H, Li Y, Zhang H. Towards trustworthy LLMs: a review on debiasing and dehallucinating in large language models. Artif Intell Rev 2024;57(9):243. [doi: 10.1007/s10462-024-10896-y]
32. Bragazzi NL, Garbarino S. Toward clinical generative AI: conceptual framework. JMIR AI 2024;3:e55957 [FREE Full text] [doi: 10.2196/55957] [Medline: 38875592]

Abbreviations

AI: artificial intelligence
ART: assisted reproductive technology
LLM: large language model
TDM: Term-Document Matrix
WHO: World Health Organization

XSL•FO
RenderX

Edited by A Mavragani; submitted 07.01.24; peer-reviewed by M Dagli, W Han, D Valdes; comments to author 16.05.24; revised version received 28.12.24; accepted 09.01.25; published 05.02.25.

Please cite as:
Bragazzi NL, Buchinger M, Atwan H, Tuma R, Chirico F, Szarpak L, Farah R, Khamisy-Farah R
Proficiency, Clarity, and Objectivity of Large Language Models Versus Specialists' Knowledge on COVID-19's Impacts in Pregnancy: Cross-Sectional Pilot Study
JMIR Form Res 2025;9:e56126
URL: https://formative.jmir.org/2025/1/e56126
doi:10.2196/56126
PMID:

©Nicola Luigi Bragazzi, Michèle Buchinger, Hisham Atwan, Ruba Tuma, Francesco Chirico, Lukasz Szarpak, Raymond Farah, Rola Khamisy-Farah. Originally published in JMIR Formative Research (https://formative.jmir.org), 05.02.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A 4-Site Public Deliberation Project on the Acceptability of Youth Self-Consent in Biomedical HIV Prevention Trials: Assessment of Facilitator Fidelity to Key Principles

Claire Burke Draucker[1], MSN, PhD; Andrés Carrión[1], MS; Mary A Ott[1,2], MA, MD; Ariel I Hicks[1], MPH; Amelia Knopf[1], MPH, PhD

[1]Indiana University, Indianapolis, IN, United States

[2]Icahn School of Medicine at Mount Sinai, New York, NY, United States

Corresponding Author:
Claire Burke Draucker, MSN, PhD
Indiana University
111 Middle Drive
Indianapolis, IN, 46202
United States
Phone: 1 317 274 4139
Email: cdraucke@iu.edu

Abstract

Background: Public deliberation is an approach used to engage persons with diverse perspectives in discussions and decision-making about issues affecting the public that are controversial or value laden. Because experts have identified the need to evaluate facilitator performance, our research team developed a framework to assess the fidelity of facilitator remarks to key principles of public deliberation.

Objective: This report describes how the framework was used to assess facilitator fidelity in a 4-site public deliberation project on the acceptability of minor self-consent in biomedical HIV prevention research.

Methods: A total of 88 individuals participated in 4 deliberation sessions held in 4 cities throughout the United States. The sessions, facilitated by 18 team members, were recorded and transcribed verbatim. Facilitator remarks were highlighted, and predetermined coding rules were used to code the remarks to 1 of 6 principles of quality deliberations. A variety of display tables were used to organize the codes and calculate the number of facilitator remarks that were consistent or inconsistent with each principle during each session across all sites. A content analysis was conducted on the remarks to describe how facilitator remarks aligned or failed to align with each principle.

Results: In total, 735 remarks were coded to one of the principles; 516 (70.2%) were coded as consistent with a principle, and 219 (29.8%) were coded as inconsistent. A total of 185 remarks were coded to the principle of equal participation (n=138, 74.6% as consistent; n=185, 25.4% as inconsistent), 158 were coded to expression of diverse opinions (n=110, 69.6% as consistent; n=48, 30.4% as inconsistent), 27 were coded to respect for others (n=27, 100% as consistent), 24 were coded to adoption of a societal perspective (n=11, 46% as consistent; n=13, 54% as inconsistent), 99 were coded to reasoned justification of ideas (n=81, 82% as consistent; n=18, 18% as inconsistent), and 242 were coded to compromise or movement toward consensus (n=149, 61.6% as consistent; n=93, 38.4% as inconsistent). Therefore, the counts provided affirmation that most of the facilitator remarks were aligned with the principles of deliberation, suggesting good facilitator fidelity. By considering how the remarks aligned or failed to align with the principles, areas where facilitator fidelity can be strengthened were identified. The results indicated that facilitators should focus more on encouraging quieter members to participate, refraining from expressing personal opinions, promoting the adoption of a societal perspective and reasoned justification of opinions, and inviting deliberants to articulate their areas of common ground.

Conclusions: The results provide an example of how a framework for assessing facilitator fidelity was used in a 4-site deliberation project. The framework will be refined to better address issues related to balancing personal and public perspectives, managing plurality, and mitigating social inequalities.

(JMIR Form Res 2025;9:e58451) doi:10.2196/58451

XSL•FO
RenderX

KEYWORDS

public deliberation; deliberative democracy; bioethics; ethical conflict; biomedical; HIV prevention; HIV research; group facilitation; fidelity assessment; content analysis

Introduction

Background

Public deliberations are community-focused activities in which persons with diverse perspectives are engaged in discussion and decision-making about a topic that is controversial or value laden or involves ethical uncertainty. Public deliberations typically include provision of factual and balanced information about the deliberation topic, facilitated small- and large-group discussions, and activities that yield collective solutions to policy questions [1-3].

Public deliberation has been used to explore a wide variety of issues that affect public health. The method has been used to elicit views on genetic research and testing in Alaska Native and American Indian communities [4], attitudes and recommendations to address antibiotic overuse [5], opinions on acceptable uses of deidentified health information in oncology settings [6], views on the allocation of limited health care resources [7], opinions on guidelines for the design and evaluation of decision aids [8], recommendations for opioid screening in the context of HIV care [2], preferences regarding health data sharing for patients with cancer [9], and views on precision medicine research in American Indian and Alaska Native communities [10].

With the increasing use of public deliberations to inform public health policy decisions, experts have called for increased attention to assessing and reporting the quality of deliberations [11,12]. To aid in this task, Scott et al [11] developed a deductive coding framework based on the operationalization of common deliberation goals and a synthesis of published frameworks designed to assess deliberation quality. Their framework includes 4 deliberation elements (ie, preferences and values are referenced, deliberants engage with each other, expert information is referenced, and information enriches deliberation) and 4 recommendation elements (ie, recommendations are clear and identifiable, recommendations are relevant, recommendations are justified, and recommendations reflect a societal perspective). Goold et al [12] also developed an analytic framework for the assessment of deliberation quality. This framework calls for an evaluation of structure (eg, demographic representativeness of the deliberants), process (eg, equality of participation), and outcomes (eg, deliberant knowledge of deliberation topics) using a variety of data sources and analytic approaches (eg, surveys, follow-up interviews, qualitative analysis of transcripts, word counts, and participation metrics).

Most evaluation frameworks do not provide for an assessment of the performance of facilitators. Deliberations typically rely on skilled facilitators who are responsible for promoting deliberative norms, fostering high-quality discussion, and promoting an outcome of shared solutions. Rubinelli and von Groote [13] outline common facilitator tasks, including conducting the dialogue, soliciting different opinions, ensuring that reasons are provided for points of view, and ensuring that reasons are grounded in evidence. The authors also discuss facilitator challenges such as identifying hidden goals of deliberants and managing sharp differences of opinion. Dillard [14] argues that facilitators are “the first and primary form of intervention” and provides an in-depth discussion of facilitator responsibilities. These responsibilities can include setting the discussion tone, establishing deliberative rules and norms, mitigating adversarial and polarizing dynamics, and ensuring that minority opinions are expressed. Moreover, the author identifies 3 facilitator types based on variations in their actions: passive facilitators (ie, they regulate deliberant speech time and direct participant turn taking), involved facilitators (ie, they ask elicitation questions and implement discursive strategies), and moderate facilitators (ie, they use elicitation questions and discursive strategies but less often than involved facilitators). Discursive strategies are facilitation techniques used to manage tensions and encourage fruitful deliberative discourse. These strategies include storytelling (ie, sharing personal experiences), playing devil’s advocate, framing (ie, highlighting interactions to shape discussion), focusing on local issues to consider specific implications, and eliciting responses.

Objectives

Despite the assumed importance of facilitators in ensuring effective deliberative processes, published reports of deliberation outcomes rarely include information regarding facilitator credentials, selection, training, and supervision and, with few exceptions (see the study by De Vries et al [15]), do not address the quality of facilitator involvement. Therefore, a framework is needed to guide researchers in systematically assessing public deliberation facilitator remarks. To address this gap, our research team developed such a framework. This framework is based on the assumption that the performance of facilitators can be assessed by determining the fidelity of their remarks to well-established principles of democratic deliberation. A blueprint to guide researchers in implementing the framework has been published in *JMIR Formative Research* [16] and includes several coding templates and a step-by-step analytic plan. The purpose of this report is to describe how the framework (hereafter referred to as the blueprint report) was used to assess facilitator fidelity in our 4-site public deliberation on the acceptability of youth self-consent in biomedical HIV prevention research. In this report, we (1) briefly describe the deliberation, (2) summarize the blueprint report, and (3) present our assessment of facilitator fidelity across the 4 sites.

Methods

Public Deliberation

The deliberation research project from which this work stemmed is described in depth elsewhere [17,18] and summarized in the blueprint report [16]. Therefore, we include only a brief recap of the main aspects of the deliberation project in this section.

XSL•FO
RenderX

The public health problem that was the focus of the deliberation was the underrepresentation of minors in biomedical HIV prevention trials due in part to barriers caused by standard regulatory requirements for parental consent and youth assent [19,20]. This consent model can result in parents becoming aware of some stigmatized behaviors or identities that their children may not wish to disclose and, thus, can lead to rejection or punishment of the youth or their exclusion from research [21]. Because youth from underrepresented populations assume a disproportionate burden of HIV disease and are the most likely to be harmed by existing consent processes, calls have been made for consideration of new consent procedures, including minor self-consent [22]. However, because parental consent has long been considered foundational to protecting youth from research-related harms, developing new consent models constitutes a public health issue that is both controversial and value laden. The purpose of our deliberation was to engage diverse stakeholders, including minor adolescents, parents, caregivers, and community members, in deliberations about the use of minor self-consent in biomedical HIV prevention research.

Ethical Considerations

This study was reviewed and approved by the Indiana University Institutional Review Board (IRB; protocol 2001909245), which served as the single IRB in accordance with reliance agreements from participating institutions, including the University of South Florida, Johns Hopkins University, the University of Colorado Denver, and the University of Chicago. An electronic copy of the IRB-approved study information sheet was provided to deliberants by study staff or site investigators. In web-based or face-to-face meetings, staff reviewed study information with interested deliberants and obtained verbal consent from those who agreed to participate. Although a waiver of parental permission was granted for this study, minors were asked whether they felt comfortable seeking parental permission to enroll in the study. Those who wished to engage a parent helped staff schedule a time to meet with both the minor and their parent or guardian to review the study information sheet and obtain parental permission. If a minor did not wish to engage a parent, the study staff contacted a site principal investigator, who determined whether the waiver of parental permission could be applied.

All participants were assigned a study ID number, and their files were labeled with that ID number. All data were stored on a dually authenticated, password-protected, cloud-based storage system. The servers are tightly controlled and inaccessible to anyone without appropriate credentials. Once all analyses are complete, audio and video files will be permanently deleted, and only deidentified data will be stored.

Participants received compensation for this study, which varied by activity and retention in the study. The maximum compensation for participants who completed all surveys and homework and attended all deliberative sessions was US $300 in the form of gift cards from a variety of web-based and physical stores; participants selected the retailer that worked best for their needs.

Deliberation Sessions

Overview

The deliberations were held via a web-based video platform in 4 cities: Tampa, Florida; Baltimore, Maryland; Denver, Colorado; and Chicago, Illinois. Educational materials related to the deliberation topic were developed in consultation with a youth advisory board and external advisers and displayed on a deliberant website. At each site, 4 deliberation sessions, each lasting 2 hours, were held. Each session included large-group meetings (plenary sessions) and small breakout group sessions. During the plenary sessions, research team members and expert stakeholders (eg, physicians and IRB administrators) presented information on topics pertinent to the deliberation (eg, research with adolescents, regulatory guidelines for biomedical research, and clinical trial principles). All deliberation sessions were video recorded and transcribed. Due to recording malfunctions in Baltimore, transcripts were not available for some of these sessions.

Deliberants

There were 88 deliberants in the deliberations. To provide context for our discussion on facilitator fidelity, a summary of the deliberants' demographic characteristics is presented in Table 1.

XSL•FO
RenderX

Table 1. Summary of deliberants' demographic characteristics (N=88).

Characteristics	Deliberants, n (%)
Age	
Minor	24 (27)
Young adult	20 (23)
Adult	44 (50)
Gender	
Women, including transgender women and transfeminine	62 (70)
Genderqueer or gender nonconforming	6 (7)
Men, including transgender men and transmasculine	17 (19)
Other	3 (3)
Sexual orientation	
Lesbian, bisexual, gay, or queer	26 (30)
Heterosexual or straight	53 (60)
Questioning	5 (6)
Other[a]	4 (5)
Educational level (n=60)[b]	
High school diploma or GED[c]	8 (13)
Some college or technical program	12 (20)
Finished vocational or technical program or 2-year degree	5 (8)
Bachelor's degree or higher	35 (58)
Racial identity	
American Indian or Alaska Native	1 (1)
Asian	6 (7)
Black or African American	39 (44)
Multiracial	6 (7)
White	28 (32)
Other[d]	8 (9)
Ethnic identity	
Hispanic or Latinx	19 (22)

[a]Other sexual orientations as entered by the respondents: pansexual (2/88, 2%), asexual (1/88, 1%), and other (1/88, 1%).

[b]Among deliberants aged >18 years.

[c]GED: General Educational Development.

[d]Other racial identities as entered by the respondents: Hispanic (mixed; 1/88, 1%), Afrolatina (1/88, 1%), Mestizo (1/88, 1%), "none of these races apply to me as I am Hispanic" (1/88, 1%), West Asian (1/88, 1%), and other (1/88, 1%).

Facilitators

A total of 18 persons facilitated the deliberations across the 4 sites. The number of facilitators at each site varied according to the size of the deliberant group. Of the 18 facilitators, 8 (44%) were graduate students, 4 (22%) were research staff members, and 6 (33%) were faculty. Their racial and ethnic identities included non-Hispanic White, non-Hispanic Black, Asian, Hispanic, and multiracial; 61% (11/18) were women. The research team and content experts had expertise in HIV prevention, adolescent health, research regulation, bioethics, and public deliberation. The site stakeholders included health care professionals and community advocates. Because the roles that the team members played in the deliberation sessions were fluid (eg, public deliberation experts lead plenary sessions and facilitated breakout group sessions, and scientists from the research team and site stakeholders presented expert information and helped facilitate sessions), for the purpose of this assessment, all team members were considered facilitators.

The facilitators had different training experiences. Of the graduate student facilitators, 75% (6/8) were employed and trained by the deliberation experts, and 25% (2/8) were employed and trained by the research team. Content experts did

XSL•FO
RenderX

not receive training in deliberative methods. A booster training session was provided between deliberations at sites 3 and 4; the session focused primarily on methods to support deliberative principles and goals but also included basic group facilitation skills.

Blueprint for Implementing a Framework for Assessing Facilitator Fidelity

The blueprint report [16] directs researchers to highlight pertinent facilitator remarks on verbatim transcripts of deliberation sessions and rate them as consistent or inconsistent with any of the 6 principles of public deliberations in the framework. This approach is based on the assumption that facilitator remarks consistent with one of the principles encourage deliberant discussions that reflect deliberation ideals, whereas facilitator remarks inconsistent with one of the principles can inhibit such discussions. The principles in the framework are equal participation [23], expression of diverse opinions [24], respect for others [23], adoption of a societal perspective [23], reasoned justification of ideas [23], and compromise or movement toward consensus [24]. We developed a coding rule table that displays the 6 principles, coding rules for each principle, and examples of text coded to each principle. This table can be accessed in the blueprint report [16].

The blueprint report also includes 2 tables to organize codes for analysis: a code display table and a code summary table. In the code display table, transcript line numbers that were coded to the principles are placed in cells defined by the principle, the role of the facilitator who made the remark, and the deliberation session in which the remark was made. In the code summary table, the number of remarks coded as consistent or inconsistent with each principle across all sessions is summarized and displayed. An example of the code display and code summary tables for site 1 (Tampa) can be found in the blueprint report [16].

The blueprint report also provides a step-by-step analytic plan to assess facilitator fidelity at each site [16]. In summary, 2 raters use the coding rule table to independently highlight facilitator remarks on deliberation transcripts that reflect 1 of the 6 principles and indicate whether each remark was consistent or inconsistent with the principle. They compare their coding and resolve discrepancies through discussion and review of transcript data. The transcript line numbers of the agreed-upon codes are placed in the appropriate cell in a code display table. Information in the code display tables for each site is tabulated to then complete a code summary table.

Results

Overview

Following the blueprint report [16] guidelines discussed previously, we coded all the transcripts and developed code display and code summary tables for each site. For the project described in this paper, we also developed 2 additional project summary tables to display counts across multiple sites. The first project summary table (Table 2) shows the number of facilitator remarks coded as consistent or inconsistent with the principles by site. The second project summary table (Table 3) shows the number of facilitator remarks coded as consistent or inconsistent with the principles by principle.

Facilitator fidelity to each principle across all sessions and sites can be described based on the project summary tables. For each principle, the total number of remarks coded to that principle, the number of remarks coded as consistent or inconsistent with the principle, and the number of remarks coded to the principle for each site are presented in the following sections. The types of remarks coded to the principles are then discussed, and verbatim examples of facilitator remarks and facilitator and deliberant dialogues are presented. The 3 principles related to facilitation of the discussion (ie, equal participation, expression of diverse opinions, and respect for others) are discussed first followed by the 3 principles related to facilitation of recommendations (ie, adoption of a societal perspective, reasoned justification of ideas, and compromise or movement toward consensus).

XSL•FO
RenderX

Table 2. First project summary table—number of facilitator remarks coded as consistent or inconsistent with principles of democratic deliberation by site.

Site and principle	Consistent, n (%)	Inconsistent, n (%)
Tampa, Florida		
Equal participation (n=11)	9 (81.8)	2 (18.1)
Expression of diverse opinions (n=41)	36 (87.8)	5 (12.2)
Respect for others (n=8)	8 (100)	0 (0)
Adoption of a societal perspective (n=6)	4 (66.7)	2 (33.3)
Reasoned justification of ideas (n=7)	7 (100)	0 (0)
Compromise or movement toward consensus (n=38)	29 (76.3)	9 (23.7)
Total (n=111)	93 (83.8)	18 (16.2)
Baltimore, Maryland		
Equal participation (n=6)	6 (100)	0 (0)
Expression of diverse opinions (n=7)	6 (85.7)	1 (14.3)
Respect for others (n=2)	2 (100)	0 (0)
Adoption of a societal perspective (n=0)	0 (0)	0 (0)
Reasoned justification of ideas (n=3)	3 (100)	0 (0)
Compromise or movement toward consensus (n=21)	18 (85.7)	3 (14.3)
Total (n=39)	35 (89.7)	4 (10.3)
Denver, Colorado		
Equal participation (n=94)	59 (62.8)	35 (37.2)
Expression of diverse opinions (n=41)	24 (58.5)	17 (41.5)
Respect for others (n=9)	9 (100)	0 (0)
Adoption of a societal perspective (n=14)	6 (42.9)	8 (57.1)
Reasoned justification of ideas (n=29)	27 (93.1)	2 (6.9)
Compromise or movement toward consensus (n=92)	58 (63)	34 (37)
Total (n=279)	183 (65.6)	96 (34.4)
Chicago, Illinois		
Equal participation (n=74)	64 (86.5)	10 (13.5)
Expression of diverse opinions (n=69)	44 (63.8)	25 (36.2)
Respect for others (n=8)	8 (100)	0 (0)
Adoption of a societal perspective (n=4)	1 (25)	3 (75)
Reasoned justification of ideas (n=60)	44 (73.3)	16 (26.7)
Compromise or movement toward consensus (n=91)	44 (48.4)	47 (51.6)
Total (n=306)	205 (67)	101 (33)
Total across sites (n=735)	516 (70.2)	219 (29.8)

XSL•FO
RenderX

Table 3. Second project summary table—number of facilitator remarks coded as consistent and inconsistent with principles of democratic deliberation by principle.

	Equal participation, n/N (%)		Expression of diverse opinions, n/N (%)		Respect for others, n/N (%)		Adoption of a societal perspective, n/N (%)		Reasoned justification of ideas, n/N (%)		Compromise or movement toward consensus, n/N (%)	
	Consistent	Inconsistent	Consistent	Inconsistent	Consistent	Inconsistent	Consistent	Inconsistent	Consistent	Inconsistent	Consistent	Inconsistent
Tampa, Florida	9/11 (81.8)	2/11 (18.2)	36/41 (87.8)	5/41 (12.2)	8/8 (100)	0/8 (0)	4/6 (66.7)	2/6 (33.3)	7/7 (100)	0/7 (0)	29/38 (76.3)	9/38 (23.7)
Baltimore, Maryland	6/6 (100)	0/6 (0)	6/7 (85.7)	1/7 (14.3)	2/2 (100)	0/2 (0)	0/0 (0)	0/0 (0)	3/3 (100)	0/3 (0)	18/21 (85.7)	3/21 (14.3)
Denver, Colorado	59/94 (62.8)	35/94 (37.2)	24/41 (58.5)	17/41 (41.5)	9/9 (100)	0/9 (0)	6/14 (42.9)	8/14 (57.1)	27/29 (93.1)	2/29 (6.9)	58/92 (63)	34/92 (37)
Chicago, Illinois	64/74 (86.5)	10/74 (13.5)	44/69 (63.8)	25/69 (36.2)	8/8 (100)	0/8 (0)	1/4 (25)	3/4 (75)	44/60 (73.3)	16/60 (26.7)	44/91 (48.4)	47/91 (51.6)
Total	138/185 (74.6)	47/185 (25.4)	110/158 (69.6)	48/158 (30.4)	27/27 (100)	0/27 (0)	11/24 (45.8)	13/24 (54.2)	81/99 (81.8)	18/99 (18.2)	149/242 (61.6)	93/242 (38.4)

Equal Participation

Overview

Across the 4 sites, facilitators made 185 remarks that were coded to the principle of equal participation. A total of 74.6% (138/185) of the remarks were coded as consistent with the principle (encouraged equal participation), and 25.4% (47/185) of the remarks were coded as inconsistent with the principle (discouraged equal participation). The proportions of consistent remarks made at each site were as follows: 6.5% (9/138) in Tampa, 4.3% (6/138) in Baltimore, 42.8% (59/138) in Denver, and 46.4% (64/138) in Chicago. The proportions of inconsistent remarks made at each site were as follows: 4% (2/47) in Tampa, 0% in Baltimore, 74% (35/47) in Denver, and 21% (10/47) in Chicago.

Consistent With the Principle

In total, 3 types of remarks were coded as consistent with the principle of equal participation. The remarks (1) conveyed the importance of equal participation, (2) invited individual deliberants or subgroups (eg, youth and parents) to speak, and (3) encouraged more talkative deliberants to allow quieter members time to speak.

Facilitators routinely promoted equal participation by making remarks that emphasized the importance of equal participation. These remarks often occurred when facilitators initially discussed deliberation goals and ground rules. The facilitators emphasized that they valued all perspectives and wished to hear from all deliberants. Facilitators also encouraged equal participation throughout the sessions by frequently asking whether anyone else had something to add. A facilitator in Tampa, for example, said, “Comments from others, anybody?” (session 1).

Facilitators also called on individual deliberants or subgroups of deliberants who had been quiet or who had not contributed to a particular discussion. Facilitators sometimes engaged quiet deliberants by directing questions specifically to them (eg, “[XXXX], what are your thoughts on...?”). Facilitators often encouraged youth deliberants to participate. For example, during a discussion about actions that could be taken to protect youth who self-consent for a study, a facilitator in Chicago, said the following:

> *Do we have any youth that would like to share an opinion or thoughts here?* [Session 4; breakout group 3]

In some instances, facilitators encouraged deliberants who had already contributed to a discussion to give quieter members a chance to speak. These requests respected the contributions of more talkative deliberants but provided opportunities for silent or disengaged deliberants to weigh in. A facilitator in Denver, for example, said the following:

> *So if you’re a person like me who has no problem speaking in front of others, maybe step back once before you make repeat additional comments to others. Whereas if you’re a person who tends to wait a little bit more, if you can, step out of your comfort zone and try to make a comment and just let people know where you’re coming from as early as you feel safe doing so and we would all appreciate that.* [Session 1]

Inconsistent With the Principle

In total, 2 types of remarks were coded as inconsistent with the principle of equal participation. The remarks (1) directed questions toward deliberants who had already spoken on an issue and (2) excused quieter deliberants from speaking.

Some facilitators called on deliberants who had already contributed multiple times instead of on those who had been quiet. A facilitator in Tampa, for example, said the following:

> *So, before we do that [move to another topic], I thought we’d just give one last comment to the one that—to the person that kicked this part of the*

XSL•FO
RenderX

> *discussion off. [XXXX], would you like to say anything else?* [Session 4]

In some instances, facilitators interacted with a few deliberants who were actively discussing an issue without attempting to draw in other deliberants.

Facilitators sometimes made remarks that served to excuse quieter deliberants from providing comments. In a few instances, facilitators suggested that not everyone needed to contribute to a particular discussion. A facilitator in Denver, for example, said the following:

> *And this case, it doesn't have to be everyone that shares...* [Session 1; breakout group 2]

Facilitators occasionally invited quieter deliberants to share their views but then excused them doing so. A facilitator in Denver, for example, asked deliberants who had not spoken to "move out of their comfort zone" but then said, "We would love it if you can do that verbally but if you can't, that is okay" (session 4).

Expression of Diverse Opinions

Overview

Across the 4 sites, facilitators made 158 remarks that were coded to the principle of expression of diverse opinions. A total of 69.6% (110/158) of the remarks were coded as consistent with the principle (encouraged expression of diverse opinions), and 30.3% (48/158) of the remarks were coded as inconsistent with the principle (discouraged expression of diverse opinions). The proportions of consistent remarks made at each site were as follows: 32.7% (36/110) in Tampa, 5.5% (6/110) in Baltimore, 21.8% (24/110) in Denver, and 40% (44/110) in Chicago. The proportions of inconsistent remarks made at each site were as follows: 10% (5/48) in Tampa, 2% (1/48) in Baltimore, 35% (17/48) in Denver, and 52% (25/48) in Chicago.

Consistent With the Principle

In total, 3 types of remarks were coded as consistent with the principle of expression of diverse opinions. The remarks (1) conveyed the importance of expression of diverse opinions, (2) invited diverse opinions during deliberative discussions, and (3) asked what others who were not represented in the deliberations might think about deliberative topics.

Facilitators routinely encouraged the expression of diverse opinions when initially discussing the purpose and the ground rules of the deliberations. They stressed that conflict and dissent, if done amicably, were important to the deliberative process. Facilitators also emphasized the importance of expressing diverse views throughout the sessions. A facilitator in Chicago, for example, said the following:

> *And I want to emphasize again that it is really important for us to mention that dissent is absolutely, completely fine here, and disagreement, as long as it's respectful, is quite helpful. So, please, if you do feel a little differently about any of the actions or any trade-offs than others, please don't feel like you need to self-censor in any way...* [Session 4]

Facilitators sometimes invited deliberants to share views that were at odds with views expressed by other deliberants. They often did this by inquiring about "other perspectives" among deliberants. In Baltimore, for example, as the group was reaching a consensus on a recommendation that community partners be involved in the consent process, the facilitator asked the following:

> *So, I don't think we want to jump to a show of hands yet, but can I just encourage people to weigh in on the notion of community advocates or community partners, especially if you've got concerns about them? If you have concerns about them, it'd be great if you mentioned that and articulated them.* [Session 4]

Facilitators regularly asked deliberants to surmise how the views of other persons or groups not represented in the deliberation might differ from the deliberants' views. The facilitators indicated that these questions were aimed at "bringing in as many perspectives as possible." In Denver, for example, a facilitator said the following:

> *Do we think that there were any perspectives that might be missing, you know, in our particular group?* [Session 2; breakout group 3]

The following excerpt from the Tampa deliberation demonstrates how questions about persons not represented in the group can draw out diverse perspectives:

> *Facilitator: So, I guess in other words, do we think—so I think a good amount of the discussion thus far is some of the group is really coming from similar places. So, for the folks who don't come from the same angle that members of this group are, are there perspectives out there that are different than what we're talking about? And what do you think they would say about the conversation we're having today?*
>
> *Deliberant: I would say that, of course, there's perspective that's very different from this group. I think we're all here for kind of a similar reason. And I think that there are a lot of people out there that do not want their children to have sex education. They do not want their children to have access to birth control. They do not want their children to have access to be able to be in medical research that has to do with HIV or STDs, period. I think there is a large group of parents that would be absolutely freaked out by everything that we've talked about today, you know. So yeah, there's a wide range of comfortableness with this topic.* [Session 1; breakout group 1]

Inconsistent With the Principle

In total, 2 types of remarks were coded as inconsistent with the principle of expression of diverse opinions. The remarks (1) revealed the personal opinions of the facilitators and (2) interrupted the expressions of diverse opinions by deliberants.

We chose to code remarks in which facilitators expressed their own personal opinions as inconsistent with the principle of

XSL•FO
RenderX

expression of diverse opinions. Deliberation facilitators are typically dissuaded from expressing their own views on topics germane to the deliberation as this can squelch dissenting or challenging views of deliberants due to power differences between facilitators and deliberants. Although facilitators in our deliberation did not routinely express their own opinions, they did so occasionally. In Tampa, for example, a facilitator said the following:

> *I don't know that we're very respectful of young people. I mean, just to be blunt, like I don't think we're respectful of their time, I don't think we're respectful of their interests. I think, you know, research—we do cater to adults, and I just don't think we often were thinking as much about, you know, like you said, interest in accessibility of young people, because we're—I don't know, maybe just not cognizant of what those were.* [Session 1; breakout group 2]

The following excerpt from the Tampa deliberation demonstrates how expression of personal opinions by facilitators can lead deliberants to agree rather than discuss their own views. In this example, a deliberant suggested that school personnel "flag" students who would benefit from or be eligible for a research study. A facilitator comment changed the focus from school personnel "flagging" potentially eligible students to school personnel being well positioned to weigh in on whether a student would be competent to consent:

> *Deliberant: So, maybe [teachers] just tapping into it in the beginning of the study for identifying people who would be eligible or benefit might be helpful.*
>
> *Facilitator: Or at least being able to give a comment in terms of their ability to consent and understand.*
>
> *Deliberant: Yes.*
>
> *Facilitator: So, I think the teachers really know the kids enough to know whether they will be able to understand the level of nuance that's required [to consent]. Maybe not with any responsibility related to that, but at least being able to address this question of maturity versus, you know, being able to—*
>
> *Deliberant: I completely agree. And we, I mean, oftentimes [we] know about the ability for their parents to consent, the ability for their parents—you know, their level of education, how much they understand about the study themselves. I mean, I know my kids pretty well. And I know their families pretty well. I know their living situations. I know a lot about them. So, at least tapping into that perspective, we definitely have a lot to offer in that point of view.* [Session 4]

In some instances, facilitators inquired about diverse opinions but did not wait for deliberants to comment, and instead, the facilitators answered their own inquiries. In Denver, for example, after a deliberant spoke in favor of community consultation for youth who had no support at home, the facilitator said the following:

> *What do others think about this? Maybe something perhaps how practical this might be because something like obtaining community input requires a community to show up...So do you think it would realistically work?* [Session 4; breakout group 1]

Respect for Others

Overview

Across the 4 sites, facilitators made 27 remarks that were coded to the principle of respect for others. All 27 remarks (100%) were coded as consistent with the principle (promoted respect for others). No remarks were coded as inconsistent with the principle (discouraged respect for others). The proportions of consistent remarks made at each site were as follows: 30% (8/27) in Tampa, 7% (2/27) in Baltimore, 33% (9/27) in Denver, and 30% (8/27) in Chicago.

Consistent With the Principle

In total, 2 types of remarks were coded as consistent with the principle of respect for others. The remarks (1) conveyed the importance of respect for others and (2) acknowledged instances in which deliberants showed respect for others.

The facilitators routinely stressed the importance of respect for others in their introductory remarks to deliberants. Facilitators emphasized that showing respect for others by actively listening and considering others' views was a ground rule of the deliberations. A facilitator in Denver, for example, said the following:

> *We were all experts in our own experience. And when people were speaking about that, the best thing we can do for each other is listen to that and learn from it.* [Session 1]

Facilitators at times acknowledged when deliberants showed respect for others by pointing out exchanges among deliberants that were particularly respectful. A facilitator in Baltimore, for example, reflected on a previous deliberation:

> *We have parents, and younger people, and people that have different relationships with their parents, and it was super respectful, and I think kind of collaborative.* [Session 3]

Adoption of a Societal Perspective

Overview

Across the 4 sites, facilitators made 24 remarks that were coded to the principle of adoption of a societal perspective. A total of 46% (11/24) of the remarks were coded as consistent with the principle (encouraged adoption of a societal perspective), and 54% (13/24) of the remarks were coded as inconsistent with the principle (discouraged adoption of a societal perspective). The proportions of consistent remarks made at each site were as follows: 36% (4/11) in Tampa, 0% in Baltimore, 55% (6/11) in Denver, and 9% (1/11) in Chicago. The proportions of inconsistent remarks made at each site were as follows: 15% (2/13) in Tampa, 0% in Baltimore, 62% (8/13) in Denver, and 23% (3/13) in Chicago.

XSL•FO
RenderX

Consistent With the Principle

One type of remark was coded as consistent with the principle of adoption of a societal perspective. This type of remark asked deliberants to consider issues of civic good in their consensus recommendations. Facilitators occasionally introduced the idea that the end goal of research was to benefit society as a whole rather than individual research participants. One facilitator in Chicago, for example, said the following:

> *Like clinical [practice] really focuses on the individual to improve that individual's health, whereas clinical research, while there's often benefit to the individual...the purpose of clinical research is to advance scientific knowledge so we know what's happening in the population.* [Session 1]

In a few instances, facilitators asked deliberants to consider what is "good for society" when forming their views. A facilitator in Denver, for example, said the following:

> *Think about these concerns...from a really broad perspective—that means not just from your specific stakeholder perspective but from a community and researcher perspective.* [Session 3; breakout group 5]

Inconsistent With the Principle

One type of remark was coded as inconsistent with the principle of adoption of a societal perspective. In these instances, facilitators encouraged deliberants to share personal experiences and consider how recommendations would affect them personally without inquiring about the societal perspective. For example, the following excerpt from the Denver deliberation demonstrates a missed opportunity to move a discussion from a deliberant's personal experience as a parent to a discussion of the greater good:

> *Deliberant: Me as an adolescent parent, I would be able to trust that if my adolescent made the choice to join the study or like a research that they were doing the right thing. And I think it's also good for them to know like what's going on today, versus them not knowing, like what's going on in the world.*
>
> *Facilitator: Yeah thanks. And what evidence helped you come to that conclusion? Like obviously that perspective as a parent?*
>
> *Deliberant: So I work at* [local hospital] and we do like birth control and you know, the adolescents are able to consent for themselves without their parents knowing. And I feel like some parents don't talk about like HIV and medications, and what the possibility can be. You know they want to keep their kids hidden from everything and protected from everything. But in our reality if we don't communicate with our adolescents then that's where us as parents get in trouble and our adolescents get in trouble. [Session 3; breakout group 5]

Reasoned Justification of Ideas

Overview

Across the 4 sites, facilitators made 99 remarks that were coded to the principle of reasoned justification of ideas. A total of 82% (81/99) of the remarks were coded as consistent with the principle (encouraged reasoned justification of ideas), and 18% (18/99) of the remarks were coded as inconsistent with the principle (discouraged reasoned justification of ideas). The proportions of consistent remarks made at each site were as follows: 9% (7/81) in Tampa, 4% (3/81) in Baltimore, 33% (27/81) in Denver, and 54% (44/81) in Chicago. The proportions of inconsistent remarks made at each site were as follows: 0% in Tampa, 0% in Baltimore, 11% (2/18) in Denver, and 89% (16/18) in Chicago.

Consistent With the Principle

In total, 2 types of remarks were coded as consistent with the principle of reasoned justification of ideas. The remarks asked deliberants to (1) provide a reason for their stated views or (2) reveal what information influenced their views.

Some facilitators encouraged deliberants to provide a rationale or explanation for why they took a particular stance on a deliberation topic. In some instances, facilitators discussed the importance of understanding the "why" or the "logic" behind opinions. A facilitator in Tampa, for example, said the following:

> *We want to give reasons for—in sentences we were writing [consensus recommendations], we want to also say why we think this, not only that, why we think it should be done.* [Session 4]

Facilitators at times asked deliberants to pinpoint information they had heard in the course of the deliberation or elsewhere that had influenced their opinions. Some facilitators asked deliberants to state what facts supported their views. A facilitator in Baltimore, for example, said the following:

> *So, I was going to try to nudge the discussion a little bit more towards explanations you've heard of current practice or the background reading you did coming in...* [Session 4]

Inconsistent With the Principle

One type of remark was coded as inconsistent with the principle of reasoned justification of ideas. This type of remark encouraged deliberants to justify their opinions but then interrupted them from doing so by supplying a rationale or by excusing the participant from having to explain their reasoning. We did not code remarks in which facilitators assured deliberants that they were free to share their views only if they were comfortable doing so, especially early in the deliberations, but rather remarks in which facilitators invited deliberants to discuss the reasoning for their views but did not give them the opportunity or space to do so.

The following excerpt from the Denver deliberation demonstrates how a facilitator remark can discourage discussion of reasoned justification of ideas. In this example, a deliberant opined that opening a study to cisgender or heterosexual youth

XSL•FO
RenderX

would not change their opinion about minor self-consent. The facilitator asked a deliberant to justify their opinion but then excused them from doing so:

> *Deliberant: And to what [deliberant] said, I think that opening it up to a larger population doesn't necessarily mean that you're going to be getting participants from that larger population if they don't think that it's an issue that impacts them. Yeah, so I'm not really sure if—how much of an impact it would have for me in terms of like the consent process.*
>
> *Facilitator: Okay. Thanks. And what would you say helped you come to that conclusion? What would you say is the information that helped you support that idea? And if you don't want to expand, you don't have to. So don't worry. All right everyone. I think this is a great time to move on to the next permutation.* [Session 3; breakout group 2]

Compromise or Movement Toward Consensus

Overview

Across the 4 sites, facilitators made 242 remarks that were coded to the principle of compromise or movement toward consensus. A total of 61.6% (149/242) of the remarks were coded as consistent with the principle (encouraged compromise and movement toward consensus), and 38.4% (93/242) of the remarks were coded as inconsistent with the principle (discouraged compromise and movement toward consensus). The proportions of consistent remarks made at each site were as follows: 19.5% (29/149) in Tampa, 12.1% (18/149) in Baltimore, 38.9% (58/149) in Denver, and 61.1% (91/149) in Chicago. The proportions of inconsistent remarks made at each site were as follows: 10% (9/93) in Tampa, 3% (3/93) in Baltimore, 37% (34/93) in Denver, and 51% (47/93) in Chicago.

Consistent With the Principle

In total, 3 types of remarks were coded as consistent with the principle of compromise or movement toward consensus. The remarks (1) introduced compromise and consensus as a goal of the deliberative process, (2) promoted discussion of common ground and encouraged deliberants to seek consensus, and (3) asked deliberants to summarize the consensus they had reached.

Facilitators discussed the goals of compromise and consensus when making introductory remarks about the deliberation. They informed deliberants that, while they did not need to agree, they were to explore common ground and, if possible, reach consensus on issues that were the focus of the deliberation. In Denver, for example, the facilitator stated the following:

> *And then in the final session, we'll review all the actions and tradeoffs that we've come up with, see where we have any common ground and collaborate to turn this into an actual list of recommendations or at least guiding principles...* [Session 1]

Often later in the deliberations, facilitators encouraged participants to consider common ground and move toward consensus. The facilitators asked whether deliberants were "on the same page" about an issue, and if they had not reached a consensus, the facilitators made remarks aimed at encouraging compromise. The facilitators often asked deliberants to consider what "tradeoffs" they would be willing to accept to agree on recommendations. In Baltimore, for example, a facilitator said the following at the start of the last session:

> *So, now in this phase, in addition to revising or adding or tweaking the things that we were just talking about, the main goal of this next phase is to come up with—there's different ways to put it, a set of principles or maybe a set of recommendations.* [Session 4]

While facilitators often summarized their perceptions of deliberant common ground, at times they asked deliberants to discuss the consensus they had reached and recommendations they wished to make. During deliberations in Chicago (the final deliberation site), for example, facilitators asked for volunteers from each breakout group to share a summary of their group's consensus and recommendations. The volunteers provided comprehensive summaries to the larger group. The following excerpt shows how a deliberant summarized the breakout group's conclusions, trade-offs, and recommendations related to the use of community town halls:

> *Facilitator: So, we're going to try to report out from all our three different groups and kind of keep track of where we had similarities and differences, and then we'll start to deliberate as a large group as best we can right after that. So, let's get to it. We'll start off with group one. I know that the note taker took some notes on the slides, and so—and we have a participant who will be walking us through those.*
>
> *Deliberant: OK. Yeah. So, as far as like actions and group conclusions, how we feel about a town hall with community before study recruitment begins...as well as like the paid community advisory board that can be engaged at every step along the way. We feel one—it's useful if the young people do not have support at home. We require honest discussion with the young person as well as like trained professionals to screen. So, not that this wouldn't be useful for people who have like a more stable and supportive home life but especially important to folks who don't have that or who don't have it consistently. A negative is in the community, people [who] aren't involved with other people that much would not—would like—prefer not to really know about these sorts of things unless like some sort of like major issue popped up. And of course, that varies from like neighborhood to neighborhood but it is kind of like a general thing. Like, at the end of the day, like the majority of the US [are] still like largely conservative, so, yeah. And then c, it could be good but it could be bad like if the broader community were invited. It could attract people who maybe have like more like ill-intent. So, like, you know, the fundamentalist types that come and like, you know, want to disrupt instead of really have a constructive conversation. So, we'd have to just be wary of people who want to show up and just cause dissent and not really like be a part of the*

XSL•FO
RenderX

conversation. Not—You know, and folks who aren't really centering the youth at the end of the day, or they may claim to center the youth but it's really like under the guise of like homophobia, transphobia, hodophobia, all those things. We decided it's better if we were going to have a community advisory board that it's a paid group of people who are screened. And we would want to be selective, including people already involved with the youth, such as, you know, leaders and folks working in schools, folks specifically working with like youth HIV prevention, people working within like the queer community who know how to advocate for folks who are—yeah, show up in the range of like queerness and gender and all of these things as well as like near peers. So, people who maybe like not too long ago were going through something similar to this. Yeah. OK... [Session 4]

The deliberant continued to discuss a number of other recommendations that the group had agreed on.

Inconsistent With the Principle

One type of remark was coded as inconsistent with the principle of compromise or movement toward consensus. In these remarks, facilitators conveyed their "take on" deliberant consensus, often without first asking deliberants what common ground they felt they had reached. Facilitators often provided somewhat lengthy summaries of conclusions reached by breakout groups or by deliberant groups as a whole. These reports eliminated the opportunity for deliberants to consider and articulate points of common ground. The following excerpt from Chicago occurred after deliberant volunteers had reported on the consensus of their breakout groups (as described previously). As the session wound down, the facilitator provided a summary of common ground among the 3 breakout groups based on notes taken by the research team:

Facilitator: Well, in the interest of time, I'm going to do something. I'm going to try to summarize the common ground that we've heard...I'm going to try to come up with a summary of what our common ground is that would involve the basic elements of what actions are acceptable under most circumstances and why?...You all see nuance that needs to be added that was in your small groups that we want to make sure we keep in our summarization of your discussion, please keep that in mind and we'll take your revisions right after this. So, I think that we agreed that common ground things that would be necessary or things that would be at least useful in most situations would be community supports such as a town hall and an advisory board. Although there may be some issues with some of these in certain implementation aspects. There was generally support for ongoing support for use throughout this with workshops, counseling during and counseling afterwards. Although people wanted to make sure that it was known while these things were positive, they were not a standalone. So generally, under most circumstances, both those things are good at least options to be part of this. For minor self-consent it sounds like what I was hearing from the group as common ground would be minor self-consent good after the age of 16. And/or in addition to that with teach back as part of it, or on research that does not necessarily have such intense side effects. Or, as one group mentioned, for behavioral, not medical research. General agreement that minor self-consent with adult help is useful, possibly especially for people under 16, but that both this and the trusted adult permission could introduce problems if it becomes in conflict with an adult—with a parent, and potentially for other reasons as well. And then, general agreement that under most situations, an ombudsperson being available as part of the project for all participants—all youth participants, but especially those without—where parental consent isn't feasible was positive. And then, parental permission for folks who were less than 16 in most areas... [Session 4]

The facilitator then reported on several other areas of common ground, but the session ended soon after this summary, and there was no opportunity for deliberants to discuss "nuances," as the facilitator had suggested.

Discussion

Principal Findings

The analysis presented in this paper provides an example of how the framework that our team developed to assess facilitator fidelity to principles of democratic deliberation was used in a 4-site public deliberation project focusing on minor self-consent for biomedical HIV prevention research. The assessment was implemented as outlined in our published blueprint report [16]. The examination of facilitator fidelity at all 4 sites allowed us to examine patterns of facilitator remarks across sessions and sites, provided sufficient data to delineate different types of remarks aligned with each principle, and yielded rich verbatim exemplars of facilitator remarks and facilitator and deliberant dialogue reflecting each principle. The counts provide affirmation of facilitator fidelity as most facilitator remarks made across all sessions at all sites were coded as consistent with a principle, and moreover, the number of coded facilitator remarks generally increased over time across the sites.

Implications

Although the code counts do not serve as definitive quantitative metrics of fidelity as the counts are influenced by variations in the deliberation sites and sessions (eg, number of deliberants attending, structure of sessions, and unique group characteristics), they do provide guideposts for improving deliberation quality. In Table 4, we summarize facilitator fidelity across the 4 sites, provide conclusions about facilitator fidelity across sites, and offer recommendations for future deliberations.

XSL•FO
RenderX

Table 4. Summary of facilitator fidelity across sites, conclusions about facilitator fidelity across the 4 sites, and recommendations for future deliberations.

Principle	Summary of facilitator fidelity across the sites	Conclusions about facilitator fidelity across the sites	Recommendations for future deliberations
Equal participation	Few remarks consistent with the principle were made at the first site, and many more were made at the last 2 sites (Tampa: n=9; Denver: n=59; Chicago: n=64). Remarks inconsistent with the principle varied across sites (Tampa: n=2; Denver: n=35; Chicago: n=10).	Remarks consistent with this principle increased after the first 2 deliberations because the research team noted during debriefings that some participants were not well engaged. The team then encouraged facilitators to actively invite quieter members, especially youth, to participate and ask talkative deliberants to pause to give quieter members time to formulate responses. Facilitators also continued on occasion throughout the sessions to encourage persons who had already spoken to comment again or invite quiet deliberants to comment but then immediately assure them that they did not "need" to if they were uncomfortable.	Facilitators should continue to encourage more reticent members to contribute to the discussion and, on occasion, respectfully encourage more talkative members to refrain until others have spoken. Facilitators should be discouraged from inviting deliberants to contribute and then "excuse" them from doing so.
Expression of diverse opinions	Many remarks consistent with the principle were made across the sites (Tampa: n=36; Denver: n=24; Chicago: n=44). Remarks inconsistent with the principle were made with some frequency in later deliberations (Tampa: n=5; Denver: n=17; Chicago: n=25).	Facilitators at all sites reassured participants that differences were valued and dissent was expected or asked them to consider views that might not be represented in their group. However, some facilitators also on occasion continued to give their own opinions rather provide facts or expert information about the issues being discussed.	Deliberants should be encouraged throughout the deliberations to express diverse opinions, and the sharing of personal opinions by facilitators should always be avoided.
Respect for others	Few remarks consistent with the principle were made across the sites (Tampa: n=8; Denver: n=9; Chicago: n=8). No remarks inconsistent with the principle were made across the sites.	There were very few instances in which deliberants did not interact with each other respectfully, and therefore, more facilitator remarks encouraging them to do so were unnecessary.	Once respect for others is established as a ground rule, deliberants may not need ongoing encouragement to act respectfully toward one another, but facilitators should remain alert to and intervene in response to any instances of disrespect.
Adoption of a societal perspective	Few remarks consistent with the principle (Tampa: n=4; Denver: n=6; Chicago: n=1) were made across the sites. Few remarks inconsistent with the principle (Tampa: n=2; Denver: n=8; Chicago: n=3) were made across the sites.	At times, facilitators diverted deliberants' attention toward their own concerns rather than toward consideration of what would be good for society. The concept of societal benefit, as opposed to self-benefit, might be a difficult concept for some facilitators to understand or promote.	Strategies are needed to help facilitators introduce the concept of societal good and encourage consideration of civic benefit as deliberants finalize their recommendations. While discussion of personal experiences early in deliberations can help members get to know one another, these discussions should not be encouraged later in deliberations as participants move toward consensus.
Reasoned justification of opinions	Remarks consistent with the principle were made often, especially during the last 2 sessions (Tampa: n=7; Denver: n=27; Chicago: n=44). Remarks inconsistent with the principle were made only infrequently (Tampa: n=0; Denver: n=2; Chicago: n=16).	The increase in consistent remarks at the last 2 sites might be attributed to facilitator training as facilitators were encouraged to ask deliberants to expand on how they came to the views they expressed rather than just state their opinions. However, facilitators were less likely to ask deliberants about information that influenced their views. Moreover, infrequently, some facilitators continued to supply reasons for deliberants' views rather than encouraging them to provide their own rationale.	Facilitators should routinely ask deliberants to discuss the rationale for their views as well as identify what information they considered in coming to these views. Facilitators should never supply their understanding of the reasons or the information underlying deliberant views if the deliberants have not articulated these justifications themselves.

XSL•FO
RenderX

Principle	Summary of facilitator fidelity across the sites	Conclusions about facilitator fidelity across the sites	Recommendations for future deliberations
Compromise or movement toward consensus	Many remarks consistent with the principle were made across the sites (Tampa: n=29; Denver: n=58; Chicago: n=44). Remarks inconsistent with the principle were fairly frequent in the last 2 deliberations (Tampa: n=9; Denver: n=34; Chicago: n=47).	Facilitators often encouraged deliberants to consider the views of the group, find common ground, and consider trade-offs but also frequently provided a summary of what the facilitators believed to be a consensus. The increase in consistent remarks in Chicago likely reflected a change in the deliberation plan (run of show) as facilitators were instructed to ask 1 deliberant from each breakout group to assume responsibility for reporting their groups' points of agreement and recommendations to the larger group. However, the frequency of the inconsistent remarks even late in the deliberations suggests that facilitators continued to discuss what they believed to be the consensus reached by deliberants.	Research teams should develop strategies that allow facilitators to guide deliberants toward compromise and consensus without taking over that responsibility.

Limitations

While the assessment presented in this paper can inform future deliberation planning and facilitator training, it also raises some issues for further consideration in refining the framework. For example, few facilitator remarks in our deliberation were coded as consistent with the principle of adoption of a societal perspective, and we surmised that this was because some facilitators did not fully understand the concept and, thus, often invited consideration of personal rather than public welfare in the deliberations. This conclusion is challenged by the work of Lehoux and Proulx [25], who suggest that some deliberants can serve as both public and patient representatives. In the former role, they bring a collective perspective, and in the latter role, they bring a personal perspective. In their deliberation on the value of technological health innovations, Lehoux and Proulx [25] found that deliberants alternated between these 2 modes of engagement. When arguing from a collective perspective, they addressed issues such as social costs, environmental protection, and individual freedoms, whereas when they argued from a personal perspective, they addressed issues such as the functionality of the technology, its dehumanizing effects on users, and its potential to increase individuals' autonomy. The researchers suggest that these 2 engagement modes are complementary and each is needed to advance deliberative discussions. They recommend that facilitators can first elicit one standpoint and then elicit counterpoints from the other standpoint. This work suggests that the facilitator fidelity framework may need to provide a way for researchers to evaluate the extent to which facilitators are able to elicit discussion of both perspectives and illuminate the tension between them rather than rate questions about personal experiences and needs as inconsistent with the principle of adoption of a societal perspective.

Another issue not sufficiently addressed in the framework is plurality. The framework includes the principles of expression of diverse opinions and compromise and movement toward consensus but does not call for an assessment of how facilitators unearth and account for multiple perspectives—despite the fact that deliberants are chosen for their diverse views. In our assessment, remarks coded as inconsistent with the principle of expression of diverse opinions were primarily personal opinions expressed by the facilitators, and remarks coded as inconsistent with the principle of compromise and movement toward consensus were primarily remarks that revealed the facilitators' "takes" on deliberant points of agreement. While these were important findings and led to recommendations for facilitator training, the nuances of effectively managing plurality are not captured completely in the framework. Baker et al [24] addressed the question of "what to do when people disagree" in deliberations and discussed how different perspectives should be elicited, deliberated, and integrated into recommendations. They developed a multilevel framework that allows for transparency of plurality and considers the integration of perspectives through counting (ie, strength of preferences), coherence (ie, the logical consistency of the argument), and consensus (ie, recommendations agreed upon through deliberation). They propose that agreements and disagreements occur at 3 levels: principles (ie, high-level normative statements), policies (ie, midlevel operational rules), and patients (ie, case-level judgments). Plurality can be disentangled by examining counting, consensus, and coherence at each level. For example, deliberants may reach consensus on general principles, but these principles may not be reflected in their policy recommendations, and thus, their decision-making lacks coherence. The authors suggest that facilitation methods can be developed to elicit values at all levels and uncover instances in which principles, policies, and patients are in tension, thereby promoting discussions of inconsistencies, identification of values that have the greatest force, and increased justification of collective outputs. This work could inform the refinement of our framework by introducing important nuances in our evaluation of how facilitators elicit diverse opinions and guide deliberants to integrate multiple perspectives into their final recommendations.

The issue of facilitator influence raised in our assessment is considered in a discussion by Kuhar et al [26]. We identified a few exchanges in which facilitators expressed their opinions, which in turn influenced deliberant responses. Kuhar et al [26] point out that neutrality of facilitators is a deliberation ideal but facilitator influence on outcomes occurs regularly in deliberative discussions. We recognize that, while our framework classified

XSL•FO
RenderX

a few overt instances of facilitator influence, it is possible that facilitators influenced discussions in more nuanced ways not addressed in the framework (eg, information presented in expert testimonies), suggesting that facilitator neutrality is an issue that warrants further development. For example, assessment of facilitators' performance might include the extent to which they have reflected on their own views and developed an awareness of how their views can influence deliberation outcomes, even if subtly.

In addition to the limitations described previously, a few other concerns need to be addressed in the refinement of the framework. Kuhar et al [26] pose that the privileging of logical argumentation, which is central to most deliberation theories but not valued by all cultures, can marginalize underrepresented groups. In future iterations of our framework, we intend to add a principle that extends beyond expression of diverse perspectives and addresses mitigating social inequalities more explicitly. In addition, for the assessment reported in this paper, all team members (eg, deliberation experts, research team members, and site stakeholders) were treated as one group of facilitators as there was fluidity in how they contributed to the sessions. However, in future assessments, we will examine role differences more closely when determining how facilitation processes overall affect deliberation quality.

Conclusions

Our assessment of facilitator fidelity to the principles of democratic deliberation in a 4-site deliberation project on the acceptability of youth self-consent for biomedical HIV prevention research affirmed that most facilitator remarks aligned with deliberative principles but also revealed areas in which facilitator performance could be strengthened. Facilitators play an important role in ensuring rich deliberative discussions, and a comprehensive assessment of deliberation quality should include an assessment of facilitator performance. The facilitator fidelity framework and blueprint report developed by our team offer a guide to conduct such an assessment. Systematically assessing facilitator fidelity can not only provide an indication of deliberation quality but can also inform future deliberation planning and facilitator training.

Acknowledgments

The research reported in this manuscript was funded through a Patient-Centered Outcomes Research Institute award (ME-2018C3-14766). The views and perspectives in this report are solely the responsibility of the authors and do not necessarily represent the views of the Patient-Centered Outcomes Research Institute or its board of governors or Methodology Committee. No generative artificial intelligence was used in any portion of manuscript writing.

Data Availability

The datasets generated and analyzed during this study are not publicly available, nor are they available from the corresponding author on reasonable request. The Patient-Centered Outcomes Research Institute determines which of its awards meet the criteria for inclusion in its data repository. The terms of the award that supported this study *do not* require participation in the data repository. Much of the information discussed in the deliberations was very sensitive in nature, and the small number of research participants would make it difficult to protect the privacy of their sensitive disclosures in the sessions. For these reasons, the full transcripts will not be made available.

Conflicts of Interest

MAO's spouse is an employee at Eli Lilly, Inc. Together, they are joint small stockholders.

References

1. Blacksher E, Hiratsuka VY, Blanchard JW, Lund JR, Reedy J, Beans JA, et al. Deliberations with American Indian and Alaska native people about the ethics of genomics: an adapted model of deliberation used with three tribal communities in the United States. AJOB Empir Bioeth 2021;12(3):164-178 [FREE Full text] [doi: 10.1080/23294515.2021.1925775] [Medline: 34125006]
2. Scherer M, Kamler A, Weiss L, Blacksher E, Jeavons J, Gold MR. Using public deliberation to set priorities: the case of COVID-19 vaccine access in New York City. J Public Health Manag Pract 2022;28(1):86-94. [doi: 10.1097/PHH.0000000000001471] [Medline: 34797242]
3. Slomp C, Edwards L, Burgess M, Sapir-Pichhadze R, Keown P, Bryan S. Public values and guiding principles for implementing epitope compatibility in kidney transplantation allocation criteria: results from a Canadian online public deliberation. BMC Public Health 2023 May 10;23(1):844 [FREE Full text] [doi: 10.1186/s12889-023-15790-w] [Medline: 37165330]
4. Hiratsuka VY, Beans JA, Blanchard JW, Reedy J, Blacksher E, Lund JR, et al. An Alaska Native community's views on genetic research, testing, and return of results: results from a public deliberation. PLoS One 2020 Mar 16;15(3):e0229540 [FREE Full text] [doi: 10.1371/journal.pone.0229540] [Medline: 32176704]
5. Richmond J, Mangrum R, Wang G, Maurer M, Sofaer S, Yang M, et al. An informed public's views on reducing antibiotic overuse. Health Serv Res 2019 Dec;54(6):1283-1294 [FREE Full text] [doi: 10.1111/1475-6773.13175] [Medline: 31172520]

XSL•FO
RenderX

6. Jagsi R, Griffith KA, Jones RD, Krenz C, Gornick M, Spence R, et al. Effect of public deliberation on patient attitudes regarding consent and data use in a learning health care system for oncology. J Clin Oncol 2019 Dec 01;37(34):3203-3211 [FREE Full text] [doi: 10.1200/JCO.19.01693] [Medline: 31577472]
7. Waljee AK, Ryan KA, Krenz CD, Ioannou GN, Beste LA, Tincopa MA, et al. Eliciting patient views on the allocation of limited healthcare resources: a deliberation on hepatitis C treatment in the Veterans Health Administration. BMC Health Serv Res 2020 May 01;20(1):369 [FREE Full text] [doi: 10.1186/s12913-020-05211-8] [Medline: 32357873]
8. Schwartz PH, O'Doherty KC, Bentley C, Schmidt KK, Burgess MM. Layperson views about the design and evaluation of decision aids: a public deliberation. Med Decis Making 2021 Jul;41(5):527-539 [FREE Full text] [doi: 10.1177/0272989X21998980] [Medline: 33813928]
9. Raj M, Ryan K, Nong P, Calhoun K, Trinidad MG, De Vries R, et al. Public deliberation process on patient perspectives on health information sharing: evaluative descriptive study. JMIR Cancer 2022 Sep 16;8(3):e37793 [FREE Full text] [doi: 10.2196/37793] [Medline: 36112409]
10. Trinidad SB, Blacksher E, Woodbury RB, Hopkins SE, Burke W, Woodahl EL, et al. Precision medicine research with American Indian and Alaska Native communities: results of a deliberative engagement with tribal leaders. Genet Med 2022 Mar;24(3):622-630 [FREE Full text] [doi: 10.1016/j.gim.2021.11.003] [Medline: 34906504]
11. Scott AM, Sims R, Degeling C, Carter S, Thomas R. Developing and applying a deductive coding framework to assess the goals of citizen/community jury deliberations. Health Expect 2019 Jun;22(3):537-546 [FREE Full text] [doi: 10.1111/hex.12880] [Medline: 30864216]
12. Goold SD, Biddle AK, Klipp G, Hall CN, Danis M. Choosing Healthplans All Together: a deliberative exercise for allocating limited health care resources. J Health Polit Policy Law 2005 Aug;30(4):563-601. [doi: 10.1215/03616878-30-4-563] [Medline: 16318163]
13. Rubinelli S, von Groote PM. Stakeholder dialogue as deliberation for decision making in health policy and systems: the approach from argumentation theory. Am J Phys Med Rehabil 2017 Feb;96(2 Suppl 1):S17-S22. [doi: 10.1097/PHM.0000000000000635] [Medline: 28059875]
14. Dillard KN. Envisioning the role of facilitation in public deliberation. J Appl Commun Res 2013 Aug 25;41(3):217-235. [doi: 10.1080/00909882.2013.826813]
15. De Vries R, Stanczyk AE, Ryan KA, Kim SY. A framework for assessing the quality of democratic deliberation: enhancing deliberation as a tool for bioethics. J Empir Res Hum Res Ethics 2011 Sep;6(3):3-17 [FREE Full text] [doi: 10.1525/jer.2011.6.3.3] [Medline: 21931233]
16. Draucker C, Carrión A, Ott MA, Knopf A. Assessing facilitator fidelity to principles of public deliberation: tutorial. JMIR Form Res 2023 Dec 13;7:e51202 [FREE Full text] [doi: 10.2196/51202] [Medline: 38090788]
17. Community solutions to adolescent research consent - minor consent for biomedical HIV research (C-START). National Institutes of Health National Library of Medicine. URL: https://clinicaltrials.gov/study/NCT05371327 [accessed 2023-11-20]
18. Do public deliberation methods help improve the informed consent process for research on sensitive topics? Patient-Centered Outcomes Research Institute. URL: https://www.pcori.org/research-results/2019/do-public-deliberation-methods-help-improve-informed-consent-process-research-sensitive-topics [accessed 2023-11-20]
19. Research involving children: report and recommendations. U.S. Department of Health, Education and Dhew Publication. 1977. URL: https://videocast.nih.gov/pdf/ohrp_research_involving_children.pdf [accessed 2023-11-20]
20. CFR - Code of federal regulations title 21. U.S. Food & Drug Administration. URL: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?C [accessed 2023-11-03]
21. Arrington-Sanders R, Wilson CM, Perumean-Chaney SE, Patki A, Hosek S. Brief report: role of sociobehavioral factors in subprotective TFV-DP levels among YMSM enrolled in 2 PrEP trials. J Acquir Immune Defic Syndr 2019 Feb 01;80(2):160-165 [FREE Full text] [doi: 10.1097/QAI.0000000000001901] [Medline: 30640203]
22. Ott MA, Knopf AS. Avoiding a tyranny of the majority: public deliberation as citizen science, sensitive issues, and vulnerable populations. Am J Bioeth 2019 Aug;19(8):28-31 [FREE Full text] [doi: 10.1080/15265161.2019.1619870] [Medline: 31544636]
23. De Vries R, Stanczyk A, Wall IF, Uhlmann R, Damschroder LJ, Kim SY. Assessing the quality of democratic deliberation: a case study of public deliberation on the ethics of surrogate consent for research. Soc Sci Med 2010 Jun;70(12):1896-1903 [FREE Full text] [doi: 10.1016/j.socscimed.2010.02.031] [Medline: 20378225]
24. Baker R, Mason H, McHugh N, Donaldson C. Public values and plurality in health priority setting: what to do when people disagree and why we should care about reasons as well as choices. Soc Sci Med 2021 May;277:113892 [FREE Full text] [doi: 10.1016/j.socscimed.2021.113892] [Medline: 33882440]
25. Lehoux P, Proulx S. Deliberating as a public representative or as a potential user? Two complementary perspectives that should inform health innovation policy. Healthc Policy 2019 May;14(4):28-38 [FREE Full text] [doi: 10.12927/hcpol.2019.25858] [Medline: 31322112]
26. Kuhar M, Krmelj M, Petrič G. The impact of facilitation on the quality of deliberation and attitude change. Small Group Res 2019 Aug 26;50(5):623-653. [doi: 10.1177/1046496419861439]

XSL•FO
RenderX

Abbreviations

IRB: institutional review board

Edited by A Mavragani; submitted 20.03.24; peer-reviewed by A Grosso; comments to author 13.11.24; revised version received 15.11.24; accepted 24.11.24; published 13.02.25.

Please cite as:
Draucker CB, Carrión A, Ott MA, Hicks AI, Knopf A
A 4-Site Public Deliberation Project on the Acceptability of Youth Self-Consent in Biomedical HIV Prevention Trials: Assessment of Facilitator Fidelity to Key Principles
JMIR Form Res 2025;9:e58451
URL: https://formative.jmir.org/2025/1/e58451
doi:10.2196/58451
PMID:

©Claire Burke Draucker, Andrés Carrión, Mary A Ott, Ariel I Hicks, Amelia Knopf. Originally published in JMIR Formative Research (https://formative.jmir.org), 13.02.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Designing a Smartphone-Based Virtual Reality App for Relaxation: Qualitative Crossover Study

Amandine Verstegen[1], MSc; Tom Van Daele[1], PhD; Bert Bonroy[2], ING, PhD; Glen Debard[2], ING, PhD; Romy Sels[2], ING, MSc; Marlon van Loo[3], MSc; Sylvie Bernaerts[1], PhD

[1]Psychology and technology, Centre of Expertise - Care and Well-being, Thomas More University of Applied Sciences, Antwerp, Belgium

[2]Mobilab & Care, Centre of Expertise - Care and Well-being, Thomas More University of Applied Sciences, Geel, Belgium

[3]Prevention and empowerment, Centre of Expertise - Care and Well-being, Thomas More University of Applied Sciences, Antwerp, Belgium

Corresponding Author:
Sylvie Bernaerts, PhD
Psychology and technology
Centre of Expertise - Care and Well-being
Thomas More University of Applied Sciences
Molenstraat 8
Antwerp, 2018
Belgium
Phone: 32 34324050
Email: sylvie.bernaerts@thomasmore.be

Abstract

Background: Accumulating evidence supports the use of virtual reality (VR) in mental health care, with one potential application being its use to assist individuals with relaxation exercises. Despite studies finding support for the potential of VR to effectively aid in relaxation, its implementation remains limited outside of specialized clinics. Known barriers are insufficient knowledge regarding VR operation, lack of availability of VR relaxation apps tailored to local health care systems, and cost concerns. Unfortunately, many VR relaxation apps are designed exclusively for stand-alone headsets, limiting accessibility for a broad audience.

Objective: We aimed to design an accessible, smartphone-based VR relaxation app based on user preferences. This paper describes the assessment of 2 stand-alone VR relaxation apps and the resulting smartphone-based VR relaxation app design.

Methods: Overall, 30 participants (n=23, 77% women; n=7, 23% men) took part in 2 separate VR sessions, assessing 1 of the 2 VR relaxation apps (Flowborne and Calm Place) in each session. After each session, participants were presented with open-ended questions to assess their experiences via a web-based survey tool. These questions explored positive and negative features, shortcomings, and suggestions for improvements while also allowing space for additional remarks concerning the 2 VR relaxation apps. Three of the authors analyzed the responses using inductive thematic analysis, a process comprising 6 phases.

Results: Across both the apps, 5 recurring themes and 13 recurring subthemes were identified in the participants' answers: audio (music and sounds, guidance), visuals (content, realism, variation and dynamics in the environment), features (language, options, feedback and instructions, duration, exercise), implementation (technical aspects, cybersickness, acceptability and usability), and experience. We analyzed the participants' findings and conducted a literature review, which served as the basis for developing the app. The resulting app is a Dutch-language, smartphone-based VR relaxation app, with customization options including 3 types of relaxation exercises, 2 guiding voices, and 3 different environments. Efforts have been made to ensure maximum variation and dynamism in the environments. Calming music and nature sounds accompany the exercises. The efficacy and effectiveness of the resulting app design were not assessed.

Conclusions: This study provides insights into key features of VR relaxation apps, which were subsequently used for the development of a novel smartphone-based VR relaxation app. Further research concerning the effectiveness of this app, along with a broader evaluation of the efficacy and user feedback for smartphone-based VR relaxation apps, is needed. More generally, there is a clear need for more research on the impact of interactivity, biofeedback, and type of environment in VR relaxation.

(JMIR Form Res 2025;9:e62663) doi:10.2196/62663

XSL•FO
RenderX

KEYWORDS

smartphone-based virtual reality; virtual reality; relaxation; stress; user experience; mobile phone

Introduction

Background

Immersive virtual reality (VR) has a long-standing history in the mental health domain. Using computer-generated environments or 360-degree videos, patients experience being in the virtual environment, rather than the real world [1-3]. In therapy, VR boasts a clear-cut advantage by affording the therapist absolute control over the emotional intensity of the environment, thereby facilitating customization in response to the unique needs of the patient [4]. As VR becomes more accessible, it is being explored for a wide range of applications in mental health care. These include exposure therapy for patients with anxiety disorders [5] or posttraumatic stress disorder [6], distraction techniques for children to manage pain during burn wound treatment [3,7], body swapping for addressing eating and weight disorders [8], and using avatars with customized voices to treat patients experiencing auditory hallucinations [9]. In addition, VR is being explored as a tool to deliver or facilitate psychoeducation, behavioral activation, treatment for obsessive-compulsive disorder, addiction treatment, rehabilitation following traumatic brain injuries, and more [10,11].

Another potential application of VR is to help facilitate relaxation in individuals experiencing stress. VR relaxation apps typically present a computer-generated or 360-degree captured natural environment, such as a beach or forest, sometimes with accompanying auditory breathing or meditation exercises. The prevalence of natural environments in VR relaxation stems from research confirming their stress-reducing qualities [12]. Pizzoli et al [13] described 2 efficacious approaches to VR for promoting relaxation, namely, relaxing VR and engaging VR, and proposed a new approach, namely, personalized VR. Relaxing VR is characterized by visual content illustrating generic, pleasant environments, such as nature-based scenes, and auditory content based on common relaxation techniques, such as meditation and progressive muscle relaxation, with the aim of inducing control over one's body. In contrast, engaging VR uses more specific scenarios and interactive sceneries with a focus on empowering and training the user to regulate one's emotions [13]. Personalized VR was explored as a third, theory-based approach, focusing on an individualized design and implementation. Numerous studies have affirmed the effectiveness of VR relaxation in promoting relaxation and alleviating stress [14-19]. Recent reviews by Riches et al [20-22] also demonstrate that VR relaxation is an acceptable, feasible, and effective tool for enhancing relaxation and reducing stress across diverse populations, including the general population [20], various clinical populations [21], and in workplace settings [22].

Despite an increasing body of evidence highlighting the effectiveness of VR relaxation, its routine care implementation remains limited outside of specialized clinics [23]. Technology adoption already proves challenging in mental health care [24], but several barriers hamper the uptake of VR relaxation specifically. A first barrier is that many therapists or individuals experiencing stress frequently lack the knowledge of how to operate VR [25]. Furthermore, numerous VR relaxation apps are not adapted to or available in patients' native language. Another challenge is the fact that VR remains relatively expensive [11,26,27]. However, concerning the latter barrier, it should be noted that VR headsets exist in different price ranges. One distinction that can be drawn among VR headsets that rely on smartphones and those that can be used stand-alone. Smartphone-based VR headsets, as implied by their name, require a smartphone to function, whereas stand-alone VR headsets operate independently. The latter are frequently more sophisticated, although considerably more expensive than smartphone-based VR headsets. A smartphone-based VR headset is readily available with a starting price as low as €15 (US $15). Unfortunately, most relaxation apps are designed exclusively for stand-alone headsets, limiting accessibility for a broad audience. Notably, there is a third category, of computer-connected (or tethered) headsets. However, these are even more expensive than stand-alone VR headsets, especially because they rely on an additional powerful computer.

Objectives

The purpose of this study is to develop a relaxation app specifically for smartphone-based VR headsets. Such VR headsets allow people to use their own smartphone, making them a potentially cheaper and more accessible alternative to stand-alone VR headsets. However, it is crucial to acknowledge that owning a smartphone is a prerequisite for using a smartphone-based VR headset. Fortunately, most of the population already owns a smartphone, making this cost less of a barrier. To overcome the previously highlighted barrier of inadequate adaptation to the local health care sector and language, the app is tailored to the local (ie, Flemish) context and developed in the Dutch language.

Methods

Overview

This study is 2-fold. In the first part, we assess the user's perspectives on existing VR relaxation apps using evidence-based techniques. In the second part, we describe the development of a smartphone VR app, considering previous user feedback and relevant literature. The validation of the resulting app design is beyond the scope of this study.

Study Design and Procedure

A crossover study to assess users' perceptions of 2 VR relaxation apps was conducted at Thomas More University of Applied Sciences, Belgium. Participants took part in 2 separate VR sessions, exploring 1 of the 2 VR relaxation apps per session. After each VR session, participants were presented with open-ended questions to assess their experiences via the web-based survey tool QuestionPro (QuestionPro, Inc). These questions explored positive and negative features by assessing

XSL•FO
RenderX

strengths (What did you like about the app?), shortcomings (What did you like less or dislike about the app?), and suggestions for improvements (What would you improve about the app?), while also allowing space for additional remarks concerning the 2 VR relaxation apps. Before and after each session, quantitative data concerning relaxation levels, positive and negative affect, attitude toward technology, and mood were also collected. However, these data are beyond the scope of this study. Each VR session lasted approximately 5 to 10 minutes and took place on 2 different days, with at least 1 week in between. The order in which the apps were presented was randomized. Each study session lasted approximately 40 minutes.

Ethical Considerations

The study design and consent forms were approved by the Ethical Committee of the Department of Applied Psychology at Thomas More University of Applied Sciences (ECTP2223_06). Before data collection, participants were informed about the study objectives, study design, and data processing both verbally and through an information letter. They also provided written informed consent. Participants were not compensated for their participation in the study. Data were collected, handled and stored according to the European Union General Data Protection Regulation (GDPR) and International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) Guideline for Good Clinical Practice (GCP). Data were deidentified and all analyses adhered to data privacy guidelines.

Participants

A total of 30 employees (n=23, 77% women; n=7, 23% men) from Thomas More University of Applied Sciences were recruited through a call in an internal newsletter between December 2022 and January 2023, across 2 campuses in Antwerp and Geel, Belgium. The call provided information about the study objectives and study procedures. Interested employees were asked to contact the researchers via email for additional information and to schedule the data collection. Their average age was 37.17 (SD 8.91) years. Individuals with neurological conditions, such as epilepsy, or severe neck problems were excluded. There were no dropouts.

Intervention

Overview

Participants were presented with 2 freely available VR relaxation apps via the Meta Quest 2 stand-alone VR headset (Meta Platforms, Inc), Flowborne (Flowborne) [28] and Calm Place (Mimerse) [23]. It is important to note that Calm Place is not publicly available, but the creators gave permission to use the app for research purposes.

Flowborne

Flowborne is a freely available respiratory biofeedback VR app designed to promote diaphragmatic breathing in a playful way. The app was developed in an academic context, and previous research has demonstrated its effectiveness in improving diaphragmatic breathing and breath awareness, promoting a sense of relaxation [29]. Flowborne consists of a low-polygon 3D environment, a digitally created space that uses fewer geometric shapes to ensure a smoother and faster VR experience, featuring natural elements such as grass, stones, flowers, and trees. The environments are composed of 4 different landscapes, each containing 6 different levels with an additional variation in duration (ie, short, medium, and long). The 4 different landscapes vary in terms of color themes and environmental characteristics. For instance, one of the landscapes resembles a hill, while another resembles an ocean. What makes this VR relaxation app stand out is the ability for users to navigate through the environment by breathing, accompanied by visual alterations in response to their breath (eg, grass growth and color shifts). During exercises, users are required to position the controller of the VR headset on their lower abdomen, allowing the controller to detect the expansion and contraction of the abdominal region and promote abdominal breathing. A slow, deep breathing pattern is necessary to progress fluently through the environment. There is no guiding voice throughout these exercises. At the outset, individuals have the opportunity to select a tutorial session that explains the exercise and the use of the controller on their abdomen. In addition, users experience ethereal and natural sounds during their journey in the environment. For this study, participants were instructed to complete a session of medium duration (ie, 9 minutes) with a landscape of their choice.

Calm Place

Calm Place is a VR relaxation app developed by Mimerse. Although this company is no longer operational, it has made its app available for research purposes upon request. In this VR relaxation app, users immerse themselves in a low-polygon natural environment, at a campfire surrounded by nature. During the exercise, one has the option to transition the environment from daytime to nighttime. The app uses a gaze-controlled interface, enabling users to navigate the environment by simply moving their heads. By directing their gaze toward specific objects, users can induce changes in the environment [23]. The exercises can be done without controllers; they are only needed to set up the exercises and customize the environment to individual preferences. This customization includes selecting whether it rains, choosing whether there are animals, setting a specific time to prevent a transition from day to night, adjusting sound volume, enabling or disabling environment sounds, choosing among 3 songs, and selecting which “relaxing object” you prefer. Exercises include guided relaxation, mindfulness exercises, or simply observing nature. In both the guided relaxation and mindfulness exercises, there is a voice that guides you step by step through the process. In the guided relaxation exercises, you can choose from different relaxation techniques: basic relaxation, muscle relaxation, and deep breathing. Similarly, in the mindfulness exercises, you can specify the session number you prefer to follow. There are also 4 different landscapes and 3 different duration options (ie, 9, 13, or 20 minutes). At the outset and conclusion, participants are asked to report on their emotional state (“How are you feeling now?”), using a scale featuring 5 emojis. For this study, participants were instructed to perform a 9-minute deep breathing exercise and were given the freedom to choose their preferred environment.

XSL•FO
RenderX

Data Analysis

Using inductive thematic analysis, qualitative data were analyzed by 3 team members (AV, MvL, and SB) for each app separately. The goal was to map out users' perspectives regarding strengths, weaknesses, and potential improvements in the 2 VR relaxation apps. Although there is no singular definitive method for conducting thematic analysis, Braun and Clarke [30] have outlined a potential procedural framework consisting of 6 phases: familiarization, coding, theme development, reviewing themes, defining and naming themes, and writing the report. In the familiarization stage, the responses to the open-ended questions underwent repeated readings by each of the 3 coders individually to attain a profound understanding. Subsequently, in the second phase, a brief summary or code was provided by each of the 3 coders individually for what was said in the responses. The data were presented in Excel (version 365; Microsoft Corp) and were open-coded through semantic coding, and no predetermined codebook was used. For instance, codes such as "more realistic" were assigned to responses such as the following:

> *The images are very "virtual." I think I would find it more pleasant if real images of mountains were used.*

Each coder individually went through multiple iterations of coding. However, these iterations and the corresponding changes and evolution of the codes were not documented. During the third stage, the 3 coders individually searched for overarching themes and subthemes and categorized the answers to the open-ended questions using colors associated with these initial themes and subthemes. Note that, in this respect, although an inductive thematic analysis approach was adopted, deductive analysis was necessary to some extent to ensure that the codes and themes contributed to answering the research questions, which is common in thematic analysis [31]. In the fourth stage, a comparison was made among the identified themes and the data to check whether these themes constitute an accurate representation of the data and aid in addressing the research question. In this phase, coders MvL, AV, and SB compared themes and codes to reach consensus about the codes and themes and subthemes. When similar codes and themes and subthemes were present, the clearest (ie, based on consensus) terminology was withheld. When codes and themes and subthemes differed, the researchers discussed until consensus was reached. These discussions were not documented. Official names were assigned to the themes in the fifth phase, which was led by AV. These tables were subsequently checked by SB, and the theme and subtheme structure was discussed until consensus was reached. MvL was not able to partake in this stage due to practical reasons. This thematic analysis revealed that participants' responses to the open-ended questions could be categorized according to the following themes across the 2 VR relaxation apps: audio, visuals, features, implementation, and experience. Each of the themes has been further subdivided into different subthemes. These themes and subthemes were reported in the sixth phase and translated into specific app features, with their respective findings elaborated upon in the Results section of this study.

Results

In the first part of this section, we discuss user perspectives on existing VR relaxation apps, categorized according to the aforementioned themes and subthemes. In the second part, we describe the translation from user preferences to app features and the resulting smartphone VR app, considering this user feedback and relevant literature.

Stage 1: Assessment of User Perspectives

Results of the thematic analyses for each separate app can be found in Tables 1 and 2. The corresponding codes and themes and translation to app features are discussed in stage 2.

XSL•FO
RenderX

Table 1. Summary of users' perspectives on the Flowborne app, organized according to the themes and subthemes from the thematic analysis.

Themes and subthemes	Strengths	Areas of improvement and negative features
Audio		
Music and sounds	• Music found to be calming (n=14, 47%) • Music mentioned spontaneously when asked for positive aspects of the app (n=9, 30%) • Nature sounds (n=2, 7%)	• Suggestion to add more nature sounds (n=1, 3%)
Guidance	• Lack of guiding voice (n=3, 10%)	• Preference for a guiding voice (n=1, 3%)
Visuals		
Content	• Beautiful setting (n=10, 33%) • Magical setting (n=1, 3%) • Animals (n=4, 13%)	• Desire for more animals (n=3, 10%) • World feels deserted, too little life (n=1, 3%)
Realism	• Realistic virtual setting (n=1, 3%) • Magical environment (n=1, 3%)	• Preference for more realistic environment (n=3, 10%) • Artificial environment (n=2, 7%) • Futuristic animals and elements (n=1, 3%)
Variation and dynamics in the environment	• Appreciation for existing variety (n=2, 7%) • High activity level in the environment (n=1, 3%)	• Desire for more variety (n=3, 10%) • Desire for more dynamism (n=2, 7%) • Distraction due to environmental activity (n=1, 3%)
Features		
Language	—[a]	—
Options	• Customizable duration (n=1, 3%)	• Desire for customizable route (n=1, 3%)
Feedback and instructions	—	• Unclear instructions and desire for example exercise (n=3, 10%) • Unclear feedback at the end (n=2, 7%) • Lack of reference for breathing pace (n=1, 3%) • Unclear end of exercise (n=1, 3%)
Duration	—	• Duration too short (n=2, 7%) • Negative feedback on exercise duration (n=1, 3%)
Exercise	• Movement through breathing (n=7, 23%) • Visual feedback linked to breathing (n=3, 10%)	• Lack of forward movement while inhaling (n=3, 10%) • Tendency to exhale more to move quickly through the environment (n=1, 3%)
Implementation		
Technical aspects	—	• Negative feedback on configuration of walking path (n=3, 10%) • Disappointment with the configuration of the environment boundaries (n=2, 7%) • Controller as a suboptimal breathing sensor (n=1, 3%) • Occasional image blurriness (n=1, 3%) • Occasional visual glitches (n=1, 3%)
Cybersickness	—	• Nausea (n=2, 7%) • Dizziness (n=2, 7%) • Nausea and headache (n=1, 3%)
Acceptability and usability	• Intention to use it again (n=1, 3%)	• Sense of an abandoned world (n=1, 3%)
Experience		
	• Pleasant experience (n=5, 17%) • Relaxing effect (n=5, 17%) • Energizing effect (n=1, 3%) • Interesting as a sleep aid (n=1, 3%) • Desire to keep wearing the headset (n=1, 3%)	• Competitive aspect hinders relaxation (n=1, 3%) • Uncertain about repeated enjoyment (n=1, 3%) • Difficulty staying awake (n=1, 3%)

[a]None.

XSL•FO
RenderX

Table 2. Summary of users' perspectives on the Calm Place app, organized according to the themes and subthemes from the thematic analysis.

Themes and subthemes	Strengths	Areas of improvement and negative features
Audio		
Music and sounds	• Positive feedback on nature sounds (n=9, 30%)	• Desire for more nature sounds (n=2, 7%) • Suggestion to add music (n=1, 3%)
Guidance	• Pleasant, relaxing voice (n=12, 40%)	• Desire for continuous guiding voice counting during breathing exercise (n=4, 13%) • Desire for less auditory guidance (n=3, 10%) • Desire for more interaction with guiding voice (n=1, 3%) • Guiding voice should not be too demanding (n=1, 3%)
Visuals		
Content	• Environment (n=12, 40%) • Campfire (n=2, 7%) • Northern lights (n=2, 7%) • Animals (n=1, 3%) • Beach and rocks (n=1, 3%) • Vacation feeling (n=1, 3%)	—[a]
Realism	• Realistic environment (n=1, 3%)	• Desire for greater realism in the environment (n=4, 13%)
Variation and dynamics in the environment	• Day-night transition (n=5, 17%) • Variation in the environment (n=4, 13%) • Dynamic elements (n=1, 3%) • Rain (n=1, 3%)	• Rain and thunder found to be unpleasant (n=5, 17%) • Static environment (n=1, 3%) • Monotonous environment (n=1, 3%) • Desire for more dynamic elements (n=1, 3%) • Desire for more variation (n=1, 3%)
Features		
Language	—	• Desire for the Dutch-language option to lower the threshold for certain target groups (n=3, 10%)
Options	• Ability to choose different environments (n=2, 7%) • Curiosity regarding the other environments (n=2, 7%)	—
Feedback and instructions	—	• Suggestion for better preparation or a test exercise (n=1, 3%) • Suggestion for instruction on appropriate posture (n=1, 3%) • Suggestion for option without indicating feelings at the end (n=1, 3%)
Duration	• Satisfaction with duration (n=2, 7%) • Positive feedback on customizable exercise duration (n=1, 3%)	• Exercise duration found to be short (n=1, 3%)
Exercise	—	• Desire for clearer indication of breathing rhythm (n=3, 10%) • Desire for visual cues for counting during breathing exercise (n=2, 7%) • Discomfort expressed regarding the fixed breathing pace (n=2, 7%) • Too many tasks at once (n=1, 3%)
Implementation		
Technical aspects	• Appreciation for minimal controller use (n=1, 3%)	• Controller visible at times (n=3, 10%) • Occasional visual glitches (n=3, 10%) • Difficulties with controller clicking (n=1, 3%)
Cybersickness	—	—

XSL•FO
RenderX

Themes and subthemes	Strengths	Areas of improvement and negative features
Acceptability and usability	—	• Distraction due to background noise (n=2, 7%) • Doubts about virtual reality relaxation's superiority over breathing exercises with only visual or video stimuli (n=1, 3%) • Suggestion to use headphones (n=1, 3%) • Suggestion to use the app while standing (n=1, 3%)
Experience		
	• Pleasant experience (n=10, 33%) • Relaxing effect (n=4, 13%) • Easy to imagine being in such an environment (n=1, 3%) • Minimal sensory stimuli (n=1, 3%)	• Boring experience (n=1, 3%) • Experience found to be tiring (n=1, 3%) • Fear due to the dark environment (n=1, 3%)

[a]None.

Stage 2: Translation of User Perspectives to App Features

Overview

A new VR relaxation app, Immersive Mental Health (IMH; Aeroplane XR), was designed for smartphones, incorporating insights from the user assessments of the 2 stand-alone VR apps and guided by a comprehensive literature review. The app was developed by the Flemish XR developer Aeroplane XR using the Unity platform. The app's conceptualization is presented in Figure 1. In this section, we describe the developed smartphone VR app, considering user feedback and relevant literature.

Figure 1. Personal configuration flow in the Immersive Mental Health app. The language in the screenshots has been translated from Dutch to English for publication purposes.

Audio

Music and Sounds

Numerous participants reported finding ethereal and nature sounds pleasant and conducive to relaxation. Studies have shown that the concurrent exposure to a nature-based VR setting and natural sounds is more effective in reducing stress compared to independent auditory or visual inputs [17,32]. Drawing upon participant evaluations and literature, we chose to integrate both natural and ethereal sounds into our new IMH app. Different nature sounds, such as chirping birds, the sound of the sea,

XSL•FO
RenderX

crackling fire, wind, and others, can be heard depending on the environment chosen.

Guidance

Opinions on the use of a guiding voice were divided, both regarding whether it should be included at all and whether it should provide continuous guidance during breathing exercises. Some participants preferred a continuous vocal count throughout the breathing exercise, stating the following:

> *I would rather hear the voice continuously guiding the breath exercise with simple words like "in and out." Counting to 10 myself kept me more focused on counting than on my breathing.*

In contrast, others preferred its absence:

> *Hearing the guiding voice count aloud the duration of breaths made me more anxious because my pace was slower than indicated, making me afraid of not completing on time.*

IMH does not require a counting voice during breathing exercises; consequently, such exercises are not included. We opted to include a guiding voice for 2 out of the 3 exercises to provide auditory explanations for the exercises, as participants have articulated a preference for clear instructions. In addition, participants were given the option to choose between a male or a female narrator. We included this customization option because the literature has consistently shown that providing personalization options leads to a higher level of relaxation [33].

Visuals

Content

A substantial number of participants found the environments to be pleasant:

> *Lovely environment. Feels like a vacation.*

Animals and natural features such as fire and water were especially well received. Literature has also shown that natural lighting and water elements are essential for alleviating stress in VR users [34]. Grounded in the findings of this study and the evaluative study, the IMH app features natural landscapes enriched with elements such as water, fire, and animals.

Realism

Participants valued a realistic environment, often suggesting that the environments could be even more realistic:

> *The images are very "virtual." I think it would be more pleasant if real images of mountains were used.*

According to a review study, VR natural scenarios have similar effects to real nature scenarios for stress reduction [34]. However, they highlight the importance of maximizing realism in these VR scenarios to improve the immersive experience. Furthermore, research indicates that a computer-generated natural environment provides a more emotionally restorative experience compared to a 360-degree video representation of a similar setting [35]. In addition, another study found that computer-generated VR enhances the sense of presence and elevates mood more effectively than 360-degree VR [36]. Therefore, based on the literature and participant feedback, we aimed to create as realistic a computer-generated environment as possible within the limits of the processing power of a smartphone (ie, Android operated).

Variation and Dynamics in the Environment

The participants placed great importance on variation. They found it pleasant when there were surprising elements, and the day-night transition in Calm Place was positively assessed:

> *It's nice that a day passes during these nine minutes. You see the sun first, then the stars, and then the sun again.*

In addition, more participants found rain and thunder unpleasant than those who found it enjoyable. Participants also expressed a preference for dynamic elements in the environment. Consequently, we integrated these considerations into our IMH app, featuring 3 distinct natural environments, each encompassing diverse fauna, natural and dynamic elements, and day-to-night transitions.

Features

Language

Some participants believed that using an app in a language other than the user's native language could pose a barrier. Therefore, the IMH app was developed in the Dutch language.

Options

Participants appreciated the option to select diverse environments and varied durations. Moreover, as noted earlier, the literature consistently demonstrates that offering customization choices results in increased levels of relaxation [33]. Therefore, we decided to incorporate a wide array of customization options. Within the IMH app, users could choose from 3 distinct environments: a tropical beach, a winter scene with the northern lights, or a mountainous landscape. Furthermore, users had the opportunity to select among 3 specific exercises: free exploration, a mindfulness exercise, or progressive muscle relaxation. In addition, participants had the option to choose between a male or female guiding voice during the mindfulness and progressive muscle relaxation exercises.

Feedback and Instructions

The participants expressed a preference for detailed instructions, a consideration we took into account in the development of our new IMH app. Before the exercises began, clear instructions were provided (Figure 1). First, participants were instructed to sit in a comfortable position. They were also prompted to horizontally rotate their smartphone in the direction indicated by an arrow. Subsequently, participants were asked to insert their smartphone into the VR headset. Clear instructions were also provided during the exercises: a guiding voice offered step-by-step guidance in 2 exercises. In addition, many participants indicated that they would have preferred to receive clear feedback at the end. However, the IMH app does not provide any feedback. Some participants also indicated wanting to try a sample exercise before starting the actual session:

> *At first, I was not sure if I was making progress. Perhaps add a small test before the meditation begins to check if you're doing it correctly.*

XSL•FO
RenderX

One of the 3 exercises in the IMH app involves simply exploring the environment without any specific tasks. This can serve as an introduction to familiarize users with the surroundings and the headset. A tutorial for the exercises is not provided in the IMH app, as the step-by-step instructions from the guiding voice eliminate the need for it.

Duration

During the evaluative study, certain participants expressed satisfaction with the presence of diverse duration options for the exercises. However, this feature was not integrated into the IMH app. Instead, users can exit the exercise whenever they desire. The participants also appreciated having a clear ending. Therefore, the conclusion is explicitly marked, as noted by the following statement: "your session is now complete."

Exercise

Many participants found navigating the environment through breathing enjoyable in Flowborne. However, there were some remarks regarding the limitation that forward movement only occurred during exhalation:

> *You should move forward during both inhalation and exhalation. Currently, you stand still while inhaling and move forward while exhaling. Right now, it feels more like a stop-and-go. It would be better in a continuous flow.*

The visual feedback based on breathing, such as the movement of dots and squares, was also appreciated. Certain participants found it unpleasant that a predetermined breathing pace was enforced, whereas others indicated a preference for clearer indications of the breathing pace. In the end, we made the decision to not include breathing exercises in the IMH app. We selected 2 exercises that have also been demonstrated to be effective in reducing stress: progressive muscle relaxation and mindfulness [36,37].

Implementation

Technical Aspects

Several issues were encountered with the controllers. Some participants observed the controllers persistently on the screen, while others struggled with their operation. Participants found it favorable when minimal controller use was required:

> *Convenient and pleasant that you don't have to use the controller system all the time, ensuring the focus is on relaxation.*

There were also occasional challenges with setting up the play area's boundaries. The IMH app uses smartphone-based VR headsets, which eliminate the need for a defined play area and do not require controllers. Furthermore, participants sporadically noted slight blurriness at the environment edges, an issue preventable by properly inserting the smartphone into the VR headset. In both the apps, it was also noted that errors in the environment (eg, occasional collisions with a rock or the campfire appearing almost on one's lap) are detrimental to the perceived realism of the environment, bringing individuals back to the realization that they are not in a real world:

> *Sometimes, the plants move through a stone or through the fire, and it brings you back to the awareness that you are not completely in a real environment, but in a virtual one.*

In response, the development of the IMH app focused on achieving the most realistic computer-generated environment feasible, while considering the technical constraints.

Cybersickness

In Flowborne, some participants experienced nausea due to a discrepancy between the movement in the VR headset and their actual physical movements. A participant expressed dissatisfaction in the following manner:

> *After the test, I felt sick and had a headache, which had an unpleasant impact on my body. To cope with the nausea and dizziness, I began to breathe less through my belly, aiming to move forward at a slower pace.*

This phenomenon, known as cybersickness in the literature, results from an inconsistency between the sensory inputs [11]. The design of the IMH app ensures that there is no forward motion within the VR environment while the user is seated in a chair in reality, thereby reducing the risk of cybersickness within the limits of the processing power of a smartphone (ie, Android operated).

Acceptability and Usability

Some participants found themselves distracted by background noises, leading 1 individual to propose the use of headphones. Moreover, 1 individual questioned whether the app offers any added value compared to non-VR relaxation apps:

> *I'm not sure if this app offers a significant advantage over regular apps that indicate the rhythm of breathing through figures or sounds.*

In addition, 1 person suggested to use the app in a standing position. Yet another participant described the environment as somewhat abandoned. Another individual expressed a willingness to use Flowborne again.

Experience

Certain participants reported feeling more relaxed after having used the VR relaxation apps. Overall, participants found the experience pleasant. However, responses did vary, with some participants reporting feelings of tiredness, others experiencing a sense of competitiveness, and some feeling energized, while 1 participant found the experience boring. Despite these diverse reactions, it is important to note that the primary aim of the IMH app is to facilitate relaxation for its users.

Discussion

Principal Findings

This study investigated users' opinions on 2 existing stand-alone VR relaxation apps and demonstrated how, based on those insights, a new smartphone-based VR relaxation app was developed. The aim of this new app is to overcome the barriers that currently limit the use of VR for relaxation, thereby making VR relaxation accessible to a wider audience. These barriers

XSL•FO
RenderX

include insufficient knowledge regarding VR operation, lack of availability of VR relaxation apps in the appropriate language for local health care systems, and cost concerns.

The evaluative study revealed 5 predominant themes that consistently recurred in the participants' evaluations: audio, visuals, features, implementation, and experience. We evaluated the main findings of the participants for each theme and incorporated them into the development process of our new IMH app. This app is a Dutch-language, smartphone-based VR relaxation tool and is freely available for both Android [38] and iOS [39]. It is designed to enable users to perform relaxation exercises within a computer-generated nature setting, accompanied by natural and ethereal auditory backgrounds. The app has been developed to incorporate a high degree of variability and dynamic features within these environments. Moreover, the design ensures minimal cybersickness by eliminating movement within the virtual environment while the user remains physically still. The IMH app has multiple customization options, as research has shown that they can enhance relaxation levels [33]. Primarily, users have the option to select from 3 exercises: 1 free exploration without auditory guidance and 2 auditory-guided exercises, involving mindfulness or progressive muscle relaxation. Our decision to offer "free exploration" as an option aims to introduce users to the VR environment, while the choice of progressive muscle relaxation and mindfulness is based on scientific evidence demonstrating their effectiveness in stress reduction [36,37]. Second, users can choose between a male or female guiding voice for the mindfulness and progressive muscle relaxation exercises. The decision to incorporate a guiding voice stemmed from participants' desire for clear instructions during the exercises. A third customization option allows users to select from 3 distinct environments: a tropical beach, a winter landscape featuring the northern lights, or a mountainous scenery.

Comparison With Other Smartphone-Based VR Relaxation Apps

To the best of our knowledge, there exist a few other smartphone-based VR relaxation apps, although they are relatively scarce. One such app is the VR-based relaxation self-training app, which focuses on self-training for relaxation using diaphragmatic breathing and progressive muscle relaxation [40]. Users initiate their training by learning relaxation techniques in a computer-generated clinic, subsequently reviewing these techniques in a comfortable environment (eg, on a beach) and finally practicing in various computer-generated settings (eg, in an elevator). Researchers have also conducted an effectiveness study comparing this app with traditional learning methods using text and illustrations, finding that the VR app is particularly effective for progressive muscle relaxation exercises, with significantly lower tension levels observed compared to a control group using worksheets. Serenity is another gamified smartphone-based VR app developed to reduce stress among university students [41]. It features a traveler role-playing game format where the user visits various vacation spots under different settings, including day and night. The game allows players to trigger special events by collecting items along paths, with these items being selectable through eye-tracking technology. The effectiveness of this immersive experience was monitored through physiological sensors, with positive outcomes particularly among students with lower stress levels. Third, Gaia VR merges neurofeedback with biofeedback in a meditation app, using feedback from a heart rate band and an electroencephalogram headband to tailor the VR experience based on the user's measured stress levels [42]. The user continuously wears the heart rate monitoring band throughout the day, enabling the app to suggest appropriate meditation techniques in response to the monitored stress levels. During the meditation sessions in VR, users can influence the virtual environment through their meditation-induced brainwave patterns, creating a dynamic and personalized meditation experience. This app has been tested thus far as a proof of concept; however, a more comprehensive evaluation is required to determine its long-term effectiveness and its applicability to a broader audience.

In summary, these 3 smartphone-based VR relaxation apps feature unique approaches and designs, differing from each other and from our IMH app, thereby demonstrating the extensive range of options available for inducing relaxation via smartphone-based VR. A key feature of the IMH app is its strong emphasis on both accessibility and customizability. As implied by its name, the Virtual Reality-Based Relaxation Self-Training app focuses mainly on the training of relaxation techniques. In contrast, the Serenity app targets students and uses gamification for stress reduction. The Gaia VR app primarily focuses on the integration of neurofeedback and biofeedback, providing real-time adjustments based on the user's physiological state. Unlike the other 3 apps, the IMH app was developed specifically using user feedback. Additional research is needed to determine the most effective design for relaxation and to identify user preferences in smartphone-based VR relaxation.

Comparison With Previous User Experience Studies

Although the effectiveness of VR for relaxation has been demonstrated in controlled research settings, insights into user experiences and preferred characteristics of VR relaxation apps remain sparse. It is important to acknowledge that, as far as we know, no user feedback studies specifically address smartphone-based VR relaxation. On the basis of 2 studies focusing on user experiences with VR for relaxation, although not specifically on smartphone-based VR, several aspects that users value can be highlighted [19,43]. Primarily, a study focusing explicitly on mindfulness VR demonstrated that sense of presence is a key factor [43]. Presence, which refers to the feeling of "being there" in a simulated environment, creates an "illusion of reality" where users behave as if the environment were real, despite being computer generated. Participants reported that this sense of presence within this virtual setting facilitated their mindfulness practices. Moreover, users noted that interruptions within the virtual environment could be distracting, leading to a reduction in the quality of the relaxation experience.

Second, users appreciated the integration of game elements into the VR relaxation experience. Features such as performance badges, use metrics, and leaderboards are appreciated and can therefore enhance the overall user experience and increase

XSL•FO
RenderX

engagement. Nonetheless, the extensive diversity in the design and implementation of these gamification elements in different apps complicates the process of determining which elements are most preferred.

Third, the potential for personalizing the experience is a key factor for users. This includes choosing the computer-generated setting, adjusting the relaxation exercises provided, and selecting the background music, all contributing to a personalized and potentially more efficacious user experience.

Fourth, participants indicated that technical aspects, such as video resolution and the weight of the VR headset, are sometimes perceived as disruptive. They specifically highlighted that issues with blurred images and areas of low resolution within the video frequently led to distractions and adversely affected their experience. In summary, the 2 studies showed that users generally have a positive opinion of VR relaxation apps [19,43]. The findings demonstrated that VR relaxation app users value an experience that seems real, has gamification elements, can be customized, and runs well without glitches. These findings are in line with what participants mentioned in our evaluative study.

In developing the IMH app, we aimed to align closely with these preferences, although gamification elements were not integrated into our app. Another interesting study developed a framework for designing a virtual environment for stress therapy [44], listing 5 factors (ie, attention, attraction, environment setting, environment exploration, and interaction) containing 14 design elements: fascination stimuli, nature - based theme, supporting audio, daylight lighting, realism, free exploration, predefined viewpoints, predefined path, passive interaction, active interaction, safety features, mind preparation, social support, and visual clarity. The IMH app meets 11 out of the 14 design elements, only lacking predefined paths, active interaction, or social support.

Limitations

Several limitations of this study warrant attention. First, obtaining insights from professionals highly experienced in conducting relaxation sessions would have added depth. Second, our participant group was relatively young, which could hinder the generalization of our results to the general population. Therefore, it would be valuable to examine the opinions of older adults concerning VR relaxation apps as well. Third, it would have been intriguing to involve the participants not only in the evaluative study but also in the development process of the new app [45]. A systematic review on cocreating digital mental health interventions indicated that involving various stakeholders, such as health care professionals and patients, in the development process contributes to the creation of culturally sensitive and widely accepted interventions [46]. Fourth, our evaluation study focused on 2 stand-alone VR relaxation apps, rather than smartphone-based VR relaxation apps. It would have been interesting to obtain user feedback specifically on smartphone-based VR relaxation apps. Finally, concerning the methodology, coding iterations and the corresponding changes and evolution of the codes were not recorded, and discussions among the 3 coders about codes and themes were also not documented. Therefore, no interrater reliability could be calculated to assess the consistency of the different researchers' coding. In this respect, it is not possible to quantitatively assess the validity and generalizability of the thematic analysis. However, we have described the process of thematic analysis in detail to ensure openness about the process.

Strengths

We developed a new smartphone-based app, grounded in both an evaluative study and a review of the literature. We believe this approach substantiates a robust methodology for app development. Furthermore, to the best of our knowledge, there are few smartphone-based VR relaxation apps available, suggesting that our development has addressed a significant gap in the field. Moreover, we hypothesize that our IMH app could also be used as a method or tool to teach relaxation techniques. For instance, individuals may, after a few sessions of VR relaxation involving progressive muscle relaxation, potentially practice these techniques without VR.

Future Directions

First, it is important to note that the presented app design, based on the assessed user preferences, has not yet been validated. This app design is an example of what a smartphone VR app for relaxation purposes could look like. In addition, future research should also focus on the efficacy and effectiveness of smartphone-based VR for relaxation. We have deliberately chosen to develop an app specifically for smartphone-based VR headsets to make VR relaxation accessible to a wider audience. Despite being more affordable and user-friendly than high-cost stand-alone VR systems, smartphone-based VR has its limitations, including a narrower field of view, lower resolution and frame rate, and less sophisticated optics and positional tracking. Nonetheless, a comparative study has not demonstrated significant differences in spatial presence, usability, cybersickness, satisfaction, workload, and learning outcomes between stand-alone VR headsets and smartphone-based VR in educational contexts [47]. Despite its limitations, smartphone-based VR might therefore still be able to deliver a satisfactory level of immersion [48]. To the best of our knowledge, no specific study has yet been conducted to explore the differences between smartphone-based VR and stand-alone VR specifically for relaxation purposes.

Second, additional investigation is required to examine other potentially engaging environments apart from natural settings, as well as to determine which particular aspects of nature are optimal for VR relaxation. In the design of the IMH app, the decision was made to include 3 different natural environments as options. This choice was driven by evidence from review studies showing that VR exposure to natural settings may lead to reductions in stress levels [34,49]. However, there is uncertainty in the literature regarding which type of VR environment is most effective for relaxation. Some researchers argue that the research has not adequately examined the diverse array of environments that can be used to promote relaxation in VR [50]. Besides urban computer-generated settings, only a handful of nonnatural alternatives have been extensively examined for VR relaxation. Consequently, a study was conducted to determine whether abstract environments could also serve as viable settings for VR relaxation. Results indicated

XSL•FO
RenderX

that the effects on perceived restoration, including sensations of being away, experiencing soft fascination, and sustaining interest, were comparable across both abstract and natural environments, pointing to a similar restorative capacity in the abstract computer-generated environment [50]. According to another review, there is a lack of understanding regarding the differentiation between green and blue virtual natural environments [51]. In addition, research has been conducted on the impact of biodiversity on stress recovery within VR settings [52]. Researchers found that stress recovery was most effective in VR environments characterized by low biodiversity. Overall, further research is needed to explore alternative engaging environments beyond natural settings and to identify the specific elements of nature that are most effective for VR relaxation.

Third, further research is needed to better understand the potential advantages of interactivity in VR relaxation, particularly within the context of smartphone-based VR. Within the literature, there is debate regarding the role of interactivity in VR relaxation environments. The underlying principle of such interactivity involves engaging the user in specific tasks within the computer-generated environment, such as gaming or art creation, which are aimed at providing distraction and subsequent stress relief [53]. It is also suggested that interactive VR serves as an interesting and motivating tool, enhancing engagement among participants [54]. A review paper concluded that the effectiveness of interactivity in VR relaxation depends on the type of game used; interactive may surpass noninteractive VR relaxation in certain contexts [51]. The IMH app was designed without interactive features, partly because the typical constraints of smartphone-based VR, such as the lack of physical controllers, make interactivity less straightforward. Notably, feedback from our participants during evaluations indicated a preference for the absence of controllers, suggesting that this feature increases the usability of VR relaxation, thereby reducing barriers to adoption. This reflects the necessity of balancing considerations such as the potential advantages of interactivity with the ease provided by nonuse of controllers, demonstrating the importance of tailoring VR relaxation to individual preferences.

Further research is needed to explore the advantages of interactive features in VR relaxation, especially within smartphone-based VR apps. Exploring other mechanisms for interaction in smartphone-based VR, aside from controllers, such as head-based movements and voice control, could also be considered. Currently, our IMH app does not incorporate biofeedback, but integrating such a feature in the future could offer substantial advantages. Interventions that include biofeedback sensors with real-time monitoring of, for example, respiration, neural activity, heart rate, and skin conductance represent a promising advancement. This biological information can be communicated to the user via dynamic environmental changes, including alterations in music, sound, color, lighting, the appearance of objects, and animations, to facilitate the detection and regulation of internal states [50]. The combination of VR with biofeedback presents multiple benefits. For instance, VR can improve the visual appeal and salience of the feedback stimuli, which may, in turn, foster greater motivation and engagement. In our evaluative study, some participants also expressed appreciation for the visual feedback based on their breathing, such as the movement of dots and squares. In addition, biofeedback may enhance trainees' belief in their capacity to relax both their body and mind, given that achievements in displaying healthy target behaviors or skills are instantly fed back and visible. Nevertheless, a significant obstacle is that most implementations of VR biofeedback demand considerable expertise and effort. Expensive physiological sensors must be purchased and integrated into the VR system [29]. Despite these challenges, it remains an intriguing avenue to explore. It would be interesting, for example, to develop a smartphone-based VR relaxation app that is linked to a smartwatch, which would provide feedback to the app.

Conclusions

In this study, a new smartphone-based VR relaxation app was designed to address the current challenges in implementing VR relaxation technologies. The design of this app draws from an evaluation study on 2 existing stand-alone VR relaxation apps, supplemented by an extensive literature review to determine key attributes of VR relaxation apps. The resulting IMH app is a Dutch-language, smartphone-based VR app featuring a variety of customization options, namely, 3 different relaxation exercises, options for male or female guiding voices, and 3 unique environments. The environments are designed to be dynamic and varied and are accompanied by soothing music and natural sounds. The app design itself was nevertheless not validated in this study. Therefore, future research should validate the resulting design and explore the effectiveness of this app, along with a broader evaluation of its efficacy and user experience, to further build on the potential of VR relaxation apps.

Acknowledgments

This study was conducted as a part of the Immersive Mental Health Project funded by the Flanders Innovation & Entrepreneurship TETRA program (HBC.2021.1023). We used generative AI (ChatGPT; OpenAI) to check sections of our writing and optimize spelling and grammar.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Conflicts of Interest

None declared.

References

1. Velana M, Sobieraj S, Digutsch J, Rinkenauer G. The advances of immersive virtual reality interventions for the enhancement of stress management and relaxation among healthy adults: a systematic review. Appl Sci 2022 Jul 20;12(14):7309. [doi: 10.3390/app12147309]
2. Bernaerts S, Bonroy B, Daems J, Sels R, Struyf D, Gies I, et al. Virtual reality for distraction and relaxation in a pediatric hospital setting: an interventional study with a mixed-methods design. Front Digit Health 2022 May 31;4:866119 [FREE Full text] [doi: 10.3389/fdgth.2022.866119] [Medline: 35712230]
3. Eijlers R, Utens EM, Staals LM, de Nijs PF, Berghmans JM, Wijnen RM, et al. Systematic review and meta-analysis of virtual reality in pediatrics: effects on pain and anxiety. Anesth Analg 2019 Nov;129(5):1344-1353 [FREE Full text] [doi: 10.1213/ANE.0000000000004165] [Medline: 31136330]
4. Rizzo A, Hartholt A, Grimani M, Leeds A, Liewer M. Virtual reality exposure therapy for combat-related posttraumatic stress disorder. Computer 2014 Jul;47(7):31-37. [doi: 10.1109/mc.2014.199]
5. Riva G. 10.06 - Virtual reality in clinical psychology. In: Asmundson GJ, editor. Comprehensive Clinical Psychology (Second Edition). Amsterdam, The Netherlands: Elsevier; 2022.
6. Rizzo A, Shilling R. Clinical virtual reality tools to advance the prevention, assessment, and treatment of PTSD. Eur J Psychotraumatol 2017 Jan 16;8(sup5):1414560 [FREE Full text] [doi: 10.1080/20008198.2017.1414560] [Medline: 29372007]
7. Hoffman HG, Meyer WJIII, Drever SA, Soltani M, Atzori B, Herrero R, et al. Virtual reality distraction to help control acute pain during medical procedures. In: Rizzo A, Bouchard S, editors. Virtual Reality for Psychological and Neurocognitive Interventions. New York, NY: Springer; 2019.
8. Serino S, Pedroli E, Keizer A, Triberti S, Dakanalis A, Pallavicini F, et al. Virtual reality body swapping: a tool for modifying the allocentric memory of the body. Cyberpsychol Behav Soc Netw 2016 Feb;19(2):127-133. [doi: 10.1089/cyber.2015.0229] [Medline: 26506136]
9. Vernal DL, Nordentoft M, Christensen MJ, Smith LC, Mariegaard L, Mainz J, et al. Status and clinical experiences from the challenge trial – a randomized controlled trial investigating virtual reality-based therapy for auditory hallucinations. World Soc Psychiatry 2023;5(1):71-76. [doi: 10.4103/wsp.wsp_4_23]
10. Rizzo A, Bouchard S. Virtual Reality for Psychological and Neurocognitive Interventions. New York, NY: Springer; 2019.
11. Baniasadi T, Ayyoubzadeh SM, Mohammadzadeh N. Challenges and practical considerations in applying virtual reality in medical education and treatment. Oman Med J 2020 May 19;35(3):e125 [FREE Full text] [doi: 10.5001/omj.2020.43] [Medline: 32489677]
12. Gaekwad JS, Sal Moslehian A, Roös PB. A meta-analysis of physiological stress responses to natural environments: Biophilia and Stress Recovery Theory perspectives. J Environ Psychol 2023 Sep;90:102085. [doi: 10.1016/j.jenvp.2023.102085]
13. Pizzoli SF, Mazzocco K, Triberti S, Monzani D, Alcañiz Raya ML, Pravettoni G. User-centered virtual reality for promoting relaxation: an innovative approach. Front Psychol 2019 Mar 12;10:479 [FREE Full text] [doi: 10.3389/fpsyg.2019.00479] [Medline: 30914996]
14. Valtchanov D, Barton KR, Ellard C. Restorative effects of virtual nature settings. Cyberpsychol Behav Soc Netw 2010 Oct;13(5):503-512. [doi: 10.1089/cyber.2009.0308] [Medline: 20950174]
15. Serrano B, Baños RM, Botella C. Virtual reality and stimulation of touch and smell for inducing relaxation: a randomized controlled trial. Comput Hum Behav 2016 Feb;55:1-8. [doi: 10.1016/j.chb.2015.08.007]
16. Hedblom M, Gunnarsson B, Iravani B, Knez I, Schaefer M, Thorsson P, et al. Reduction of physiological stress by urban green space in a multisensory virtual experiment. Sci Rep 2019 Jul 12;9(1):10113 [FREE Full text] [doi: 10.1038/s41598-019-46099-7] [Medline: 31300656]
17. Annerstedt M, Jönsson P, Wallergård M, Johansson G, Karlson B, Grahn P, et al. Inducing physiological stress recovery with sounds of nature in a virtual reality forest--results from a pilot study. Physiol Behav 2013 Jun 13;118:240-250. [doi: 10.1016/j.physbeh.2013.05.023] [Medline: 23688947]
18. Anderson AP, Mayer MD, Fellows AM, Cowan DR, Hegel MT, Buckey JC. Relaxation with immersive natural scenes presented using virtual reality. Aerosp Med Hum Perform 2017 Jun 01;88(6):520-526. [doi: 10.3357/amhp.4747.2017]
19. Fagernäs S, Hamilton W, Espinoza N, Miloff A, Carlbring P, Lindner P. What do users think about virtual reality relaxation applications? A mixed methods study of online user reviews using natural language processing. Internet Interv 2021 Apr;24:100370 [FREE Full text] [doi: 10.1016/j.invent.2021.100370] [Medline: 33665136]
20. Riches S, Azevedo L, Bird L, Pisani S, Valmaggia L. Virtual reality relaxation for the general population: a systematic review. Soc Psychiatry Psychiatr Epidemiol 2021 Oct 13;56(10):1707-1727 [FREE Full text] [doi: 10.1007/s00127-021-02110-z] [Medline: 34120220]

XSL•FO
RenderX

21. Riches S, Jeyarajaguru P, Taylor L, Fialho C, Little J, Ahmed L, et al. Virtual reality relaxation for people with mental health conditions: a systematic review. Soc Psychiatry Psychiatr Epidemiol 2023 Jul 20;58(7):989-1007 [FREE Full text] [doi: 10.1007/s00127-022-02417-5] [Medline: 36658261]
22. Riches S, Taylor L, Jeyarajaguru P, Veling W, Valmaggia L. Virtual reality and immersive technologies to promote workplace wellbeing: a systematic review. J Ment Health 2024 Apr 15;33(2):253-273 [FREE Full text] [doi: 10.1080/09638237.2023.2182428] [Medline: 36919828]
23. Lindner P, Miloff A, Hamilton W, Carlbring P. The potential of consumer-targeted virtual reality relaxation applications: descriptive usage, uptake and application performance statistics for a first-generation application. Front Psychol 2019 Feb 4;10:132 [FREE Full text] [doi: 10.3389/fpsyg.2019.00132] [Medline: 30778311]
24. Torous J, Bucci S, Bell IH, Kessing LV, Faurholt-Jepsen M, Whelan P, et al. The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. World Psychiatry 2021 Oct;20(3):318-335 [FREE Full text] [doi: 10.1002/wps.20883] [Medline: 34505369]
25. Garrett B, Taverner T, Gromala D, Tao G, Cordingley E, Sun C. Virtual reality clinical research: promises and challenges. JMIR Serious Games 2018 Oct 17;6(4):e10839 [FREE Full text] [doi: 10.2196/10839] [Medline: 30333096]
26. Segal R, Bhatia M, Drapeau M. Therapists' perception of benefits and costs of using virtual reality treatments. Cyberpsychol Behav Soc Netw 2011 Jan;14(1-2):29-34. [doi: 10.1089/cyber.2009.0398] [Medline: 21329440]
27. Schwartzman D, Segal R, Drapeau M. Perceptions of virtual reality among therapists who do not apply this technology in clinical practice. Psychol Serv 2012 Aug;9(3):310-315. [doi: 10.1037/a0026801] [Medline: 22867123]
28. Flowborne VR homepage. Flowborne VR. URL: https://flowborne.com/ [accessed 2024-02-01]
29. Rockstroh C, Blum J, Göritz AS. A mobile VR-based respiratory biofeedback game to foster diaphragmatic breathing. Virtual Real 2020 Oct 05;25(2):539-552. [doi: 10.1007/s10055-020-00471-5]
30. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
31. Byrne D. A worked example of Braun and Clarke's approach to reflexive thematic analysis. Qual Quant 2021 Jun 26;56(3):1391-1412. [doi: 10.1007/s11135-021-01182-y]
32. Naef AC, Jeitziner MM, Knobel SE, Exl MT, Müri RM, Jakob SM, et al. Investigating the role of auditory and visual sensory inputs for inducing relaxation during virtual reality stimulation. Sci Rep 2022 Oct 12;12(1):17073 [FREE Full text] [doi: 10.1038/s41598-022-21575-9] [Medline: 36224289]
33. Pardini S, Gabrielli S, Dianti M, Novara C, Zucco GM, Mich O, et al. The role of personalization in the user experience, preferences and engagement with virtual reality environments for relaxation. Int J Environ Res Public Health 2022 Jun 13;19(12):7237 [FREE Full text] [doi: 10.3390/ijerph19127237] [Medline: 35742483]
34. Gentile A, Ficarra S, Thomas E, Bianco A, Nordstrom A. Nature through virtual reality as a stress-reduction tool: a systematic review. Int J Stress Manag 2023 Nov;30(4):341-353. [doi: 10.1037/str0000300]
35. Nukarinen T, Istance HO, Rantala J, Mäkelä J, Korpela K, Ronkainen K, et al. Physiological and psychological restoration in matched real and virtual natural environments. In: Proceedings of the Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems. 2020 Presented at: CHI EA '20; April 25-30, 2020; Honolulu, HI. [doi: 10.1145/3334480.3382956]
36. Sanada K, Montero-Marin J, Alda Díez M, Salas-Valero M, Pérez-Yus MC, Morillo H, et al. Effects of mindfulness-based interventions on salivary cortisol in healthy adults: a meta-analytical review. Front Physiol 2016 Oct 19;7:471 [FREE Full text] [doi: 10.3389/fphys.2016.00471] [Medline: 27807420]
37. Muhammad Khir S, Wan Mohd Yunus WM, Mahmud N, Wang R, Panatik SA, Mohd Sukor MS, et al. Efficacy of progressive muscle relaxation in adults for stress, anxiety, and depression: a systematic review. Psychol Res Behav Manag 2024 Feb;Volume 17:345-365. [doi: 10.2147/prbm.s437277]
38. Immersive Mental Health. Google Play Store. URL: https://play.google.com/store/apps/details?id=com.Aeroplane.ImmersiveMentalHealth&hl=nl [accessed 2025-01-30]
39. Immersive Mental Health. Apple App Store. URL: https://apps.apple.com/be/app/immersive-mental-health/id6736897735 [accessed 2025-01-30]
40. Jeong HS, Oh J, Paik M, Kim H, Jang S, Kim BS, et al. Development and feasibility assessment of virtual reality-based relaxation self-training program. Front Virtual Real 2022 Jan 5;2. [doi: 10.3389/frvir.2021.722558]
41. De Asis KM, Guillem EJ, Reyes FA, Samonte MJ. Serenity: a stress-relieving virtual reality application based on Philippine environmental variables. In: Proceedings of the 6th International Conference on Frontiers of Educational Technologies. 2020 Presented at: ICFET '20; June 5-8, 2020; Tokyo, Japan. [doi: 10.1145/3404709.3404745]
42. Wang SY, Bang HS, Liu Y, Kim M, Raiti J. Gaia VR: a virtual reality-based neuro and biofeedback meditation application. In: Proceedings of the IEEE Global Humanitarian Technology Conference. 2022 Presented at: GHTC 2022; September 8-11, 2022; Santa Clara, CA. [doi: 10.1109/ghtc55712.2022.9910614]
43. Seabrook E, Kelly R, Foley F, Theiler S, Thomas N, Wadley G, et al. Understanding how virtual reality can support mindfulness practice: mixed methods study. J Med Internet Res 2020 Mar 18;22(3):e16106 [FREE Full text] [doi: 10.2196/16106] [Medline: 32186519]

XSL•FO
RenderX

44. Zaharuddin FA, Ibrahim N, Yusof AM. A conceptual framework for designing virtual environments for stress therapy. Appl Sci 2022 Oct 04;12(19):9973. [doi: 10.3390/app12199973]
45. van Bruinessen IR, van Weel-Baumgarten EM, Snippe HW, Gouw H, Zijlstra JM, van Dulmen S. Active patient participation in the development of an online intervention. JMIR Res Protoc 2014 Nov 06;3(4):e59 [FREE Full text] [doi: 10.2196/resprot.3695] [Medline: 25379679]
46. Brotherdale R, Berry K, Branitsky A, Bucci S. Co-producing digital mental health interventions: a systematic review. Digit Health 2024 Apr 25;10:20552076241239172 [FREE Full text] [doi: 10.1177/20552076241239172] [Medline: 38665886]
47. Papachristos NM, Vrellis I, Mikropoulos TA. A comparison between oculus rift and a low-cost smartphone VR headset: immersive user experience and learning. In: Proceedings of the IEEE 17th International Conference on Advanced Learning Technologies. 2017 Presented at: ICALT 2017; July 3-7, 2017; Timisoara, Romania. [doi: 10.1109/icalt.2017.145]
48. Amin A, Gromala D, Tong X, Shaw C. Immersion in cardboard VR compared to a traditional head-mounted display. In: Proceedings of the 8th International Conference on Virtual, Augmented and Mixed Reality. 2016 Presented at: VAMR 2016; July 17-22, 2016; Toronto, Canada. [doi: 10.1007/978-3-319-39907-2_25]
49. Spano G, Theodorou A, Reese G, Carrus G, Sanesi G, Panno A. Virtual nature, psychological and psychophysiological outcomes: a systematic review. J Environ Psychol 2023 Aug;89:102044. [doi: 10.1016/j.jenvp.2023.102044]
50. Barton AC, Do M, Sheen J, Byrne LK. The restorative and state enhancing potential of abstract fractal-like imagery and interactive mindfulness interventions in virtual reality. Virtual Real 2024 Feb 27;28:53. [doi: 10.1007/s10055-023-00916-7]
51. Li H, Zhang X, Wang H, Yang Z, Liu H, Cao Y, et al. Access to nature via virtual reality: a mini-review. Front Psychol 2021;12:725288 [FREE Full text] [doi: 10.3389/fpsyg.2021.725288] [Medline: 34675840]
52. Schebella MF, Weber D, Schultz L, Weinstein P. The nature of reality: human stress recovery during exposure to biodiverse, multisensory virtual environments. Int J Environ Res Public Health 2019 Dec 19;17(1):56 [FREE Full text] [doi: 10.3390/ijerph17010056] [Medline: 31861700]
53. Ladakis I, Filos D, Chouvarda I. Virtual reality environments for stress reduction and management: a scoping review. Virtual Real 2024 Feb 27;28:50. [doi: 10.1007/s10055-024-00943-y]
54. Choi YH, Paik NJ. Mobile game-based virtual reality program for upper extremity stroke rehabilitation. J Vis Exp 2018 Mar 08(133):56241 [FREE Full text] [doi: 10.3791/56241] [Medline: 29578520]

Abbreviations

IMH: Immersive Mental Health
VR: virtual reality

Edited by A Mavragani; submitted 31.05.24; peer-reviewed by C Bull; comments to author 30.10.24; revised version received 19.11.24; accepted 18.12.24; published 13.02.25.

Please cite as:
Verstegen A, Van Daele T, Bonroy B, Debard G, Sels R, van Loo M, Bernaerts S
Designing a Smartphone-Based Virtual Reality App for Relaxation: Qualitative Crossover Study
JMIR Form Res 2025;9:e62663
URL: https://formative.jmir.org/2025/1/e62663
doi:10.2196/62663
PMID:

©Amandine Verstegen, Tom Van Daele, Bert Bonroy, Glen Debard, Romy Sels, Marlon van Loo, Sylvie Bernaerts. Originally published in JMIR Formative Research (https://formative.jmir.org), 13.02.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Using an Interactive Voice Response Survey to Assess Patient Satisfaction in Ethiopia: Development and Feasibility Study

Dessalegn Shamebo[1*], PhD; Anagaw Derseh Mebratie[2*], PhD; Catherine Arsenault[3], PhD

[1]Department of Development Economics, Ethiopian Civil Service University, Addis Ababa, Ethiopia

[2]Department of Health Systems Management and Health Policy, School of Public Health, College of Health Sciences, Addis Ababa University, Addis Ababa, Ethiopia

[3]Department of Global Health, Milken Institute School of Public Health, The George Washington University, Washington, DC, United States

[*]these authors contributed equally

Corresponding Author:
Catherine Arsenault, PhD
Department of Global Health
Milken Institute School of Public Health
The George Washington University
950 New Hampshire Avenue
Washington, DC, 20052
United States
Phone: 1 2029941011
Email: catherine.arsenault@gwu.edu

Abstract

Background: Patient satisfaction surveys can offer crucial information on the quality of care but are rarely conducted in low-income settings. In contrast with in-person exit interviews, phone-based interactive voice response (IVR) surveys may offer benefits including standardization, patient privacy, reduced social desirability bias, and cost and time efficiency. IVR surveys have rarely been tested in low-income settings, particularly for patient satisfaction surveys.

Objective: In this study, we tested the feasibility of using an IVR system to assess patient satisfaction with primary care services in Addis Ababa, Ethiopia. We described the methodology, response rates, and survey costs and identified factors associated with survey participation, completion, and duration.

Methods: Patients were recruited in person from 18 public and private health facilities in Addis Ababa. Patients' sex, age, education, reasons for seeking care, and mobile phone numbers were collected. The survey included 15 questions that respondents answered using their phone keypad. We used a Heckman probit regression model to identify factors influencing the likelihood of IVR survey participation (picking up and answering at least 1 question) and completion (answering all survey questions) and a Weibull regression model to identify factors influencing the survey completion time.

Results: A total of 3403 individuals were approached across 18 health facilities. Nearly all eligible patients approached (2985/3167, 94.3%) had a functioning mobile phone, and 89.9% (2415/2685) of those eligible agreed to be enrolled in the study. Overall, 92.6% (2236/2415) picked up the call, 65.6% (1584/2415) answered at least 1 survey question, and 42.9% (1037/2415) completed the full survey. The average survey completion time was 8.1 (SD 1.7) minutes for 15 Likert-scale questions. We found that those aged 40-49 years and those aged 50+ years were substantially less likely to participate in (odds ratio 0.63, 95% CI 0.53-0.74) and complete the IVR survey (odds ratio 0.77, 95% CI 0.65-0.90) compared to those aged 18-30 years. Higher education levels were also strongly associated with survey participation and completion. In adjusted models, those enrolled in private facilities were less likely to participate and complete the survey compared to those in public health centers. Being male, younger, speaking Amharic, using a private hospital, and being called after 8 PM were associated with a shorter survey duration. The average survey costs were US $7.90 per completed survey.

Conclusions: Our findings reveal that an IVR survey is a feasible, low-cost, and rapid solution to assess patient satisfaction in an urban context in Ethiopia. However, survey implementation must be carefully planned and tailored to local challenges. Governments and health facilities should consider IVR to routinely collect patient satisfaction data to inform quality improvement strategies.

(JMIR Form Res 2025;9:e67452) doi:10.2196/67452

XSL•FO
RenderX

KEYWORDS

mobile phone surveys; patient satisfaction; interactive voice response; global health; surveys; Ethiopia; IVR; Africa

Introduction

Collecting information on patient's perception of health care quality is crucial to understand gaps in service provision. Patient satisfaction surveys have gained increasing attention as an essential source of information for quality improvement [1]. In high-income countries, these surveys are often paper-based and mailed to patients or delivered by email and completed on the web [2]. However, in low-income settings, these delivery methods can be problematic. Paper- and web-based surveys also assume that all recipients can read and write.

Other methods include phone surveys that have been increasingly popular, given the high rates of mobile phone ownership globally despite a relatively lower penetration rate in Sub-Saharan Africa (50%) [3]. Phone surveys with live interviewers can be effective but come with several challenges including the cost of hiring and training interviewers. They are also time-consuming for large samples and are prone to social desirability bias and to interviewer variability and bias [4,5]. Interactive voice response (IVR) surveys enable automated communications over the phone, where respondents engage with the survey using their phone's keypad. IVR surveys offer several benefits, particularly in the context of patient satisfaction surveys. IVR surveys can be accessed by patients who may not have internet access or are not comfortable using web-based platforms. They provide a sense of anonymity, encouraging patients to provide honest feedback [6]. The automated nature of IVR ensures that every patient is asked the same set of questions in the same manner. They are also cost-effective and highly scalable and can be delivered to thousands of patients at the same time and in multiple languages, leading to very short data collection periods for real-time feedback [7-9]. Furthermore, an IVR system can make multiple calls both during and outside of normal business hours, allowing respondents to participate in the survey at their convenience [10]. Given these benefits, they have the potential to be integrated into the routine operations of health systems.

Nonetheless, IVR surveys can face a unique set of limitations and challenges. First, they may not be suitable for individuals who are not comfortable with automated phone systems. They may lead to lower engagement, greater response fatigue, a higher rate of survey abandonment, and thus a lower overall response rate compared to other survey delivery methods. People who are not technology-savvy may struggle with navigating IVR systems. Despite these limitations, IVR surveys have been used for various purposes such as for patient satisfaction assessment after endoscopy [11], evaluation of child and adolescent psychiatric outpatient treatment [8], and decision-making about sun protection [12]. Yet, there is little published evidence on the acceptability and feasibility of using IVR to collect patient feedback in Ethiopia. We found only 2 peer-reviewed studies that discussed the feasibility of IVR in Ethiopia [13,14]. These studies assessed the feasibility of IVR messages for targeted client communication. One working paper assessed the feasibility of an IVR survey in Ethiopia to obtain nationally representative estimates at population levels [15]. To our knowledge, there has been no study assessing the use of IVR to measure patient satisfaction in Ethiopia.

In this paper, we assessed the feasibility of using an IVR survey to assess patient-perceived quality of primary care services in 18 health facilities in Addis Ababa, Ethiopia. We described the methodology, acceptability, cost, response rate, completion rate, and the time taken to complete the full survey. We also assessed the factors associated with survey participation, completion, and duration. These findings may be helpful for others conducting studies among patients or considering using IVR in a similar setting.

Methods

Study Design and Recruitment

This study was conducted in Addis Ababa, Ethiopia, in June to July 2024. Phone numbers and consent from primary care patients were obtained in person at health facilities. The study aimed to compare patient-perceived quality of primary care services across 4 types of facilities: public health centers, public general hospitals, private clinics, and private hospitals. Four facilities from each category were randomly selected from health management information system lists with the aim of enrolling a minimum of 100 patients per facility.

Enrollment was limited to patients receiving 11 common primary care services, including for chronic diseases, injuries, infectious diseases, gastrointestinal conditions, respiratory diseases, prenatal care, and family planning. Eligible participants were those who understood Amharic or Afaan Oromo, were aged 18 years and older, were at the facility for their own care, had received one of the selected services as an outpatient, and had a functioning mobile phone. Hospitalized individuals were excluded. Data collectors identified eligibility, obtained consent, and collected sex, educational level, mobile phone number, and preferred contact time for the IVR survey. No other personal identifiers were collected.

The recruitment period lasted 14 days. Each day, phone numbers were uploaded to the IVR platform. One day after enrollment, participants received an SMS text message at their preferred time to remind them of the study, followed by a call within 10 minutes to complete the IVR survey. During recruitment, respondents listened to a sample IVR question to familiarize themselves with the system. They were advised to respond honestly and to find a quiet location for the survey.

The IVR survey was programmed using the EngageSpark platform. The survey questions and response options were recorded in a recording studio in Amharic and Afaan Oromo and uploaded to the IVR platform. The IVR survey recordings were narrated by a female, given evidence that IVR surveys narrated by female individuals lead to higher response rates and can prevent the incidence of domestic violence [16]. The first call for the IVR survey took place 1 day after the primary care visit. If the participant did not pick up, the system automatically

XSL•FO
RenderX

called the respondent back after 30 minutes, 1 hour, 90 minutes, 1 day, 1 day and 1 hour, 2 days, and 2 days and 1 hour. For respondents who completed the full survey, the EngageSpark platform automatically sent the airtime top-up to the respondent's phone.

The patient satisfaction survey included 15 questions covering structural quality (eg, perceived quality and availability of equipment, medicine, and diagnostic test), competence of health providers, respectful care, patient-centered care, and user experience. The survey instrument is provided in Multimedia Appendix 1. Participants were also asked to report their monthly household income and how likely they were to recommend the health facility to a family member or friend. Most questions used a 5-point Likert scale. Among 15 questions asked, 13 items on the quality and cost of care were rated on a 5-point Likert scale (from 1=very poor to 5=very good). The question on endorsement of the health facility was answered on a 4-point Likert scale (1=not at all likely to 4=very likely). The household income question was categorized into 6 income groups.

Respondents were instructed to use their mobile phone touchpad to answer questions. Respondents were allowed to repeat or skip a question. A message was recorded to inform participants if they had pressed an invalid key. The survey question was repeated if respondents pressed an invalid key or did not answer for more than 20 seconds. The call was disconnected if they pressed an invalid key 3 times or failed to respond after a question was repeated twice.

Measures

The outcomes of interest in this study were participation in the IVR survey, completion of all survey questions, and time taken to complete all 15 survey questions (among those who completed the survey). Participation in the IVR survey was measured using a dummy variable, which was coded as 1 if the individual picked up the survey call and answered at least 1 of the survey questions and 0 otherwise. Similarly, the completion outcome was constructed as a dichotomous variable, coded as 1 if the respondent completed all 15 questions and 0 if they did not. Duration was based on the time taken in minutes to answer all 15 survey questions among those who completed the full survey. To assess the correlates of survey participation, completion, and duration, our analysis included a series of covariates collected during the in-person recruitment. These were the respondent's sex, age, educational achievement, type of primary health care service received, type of health facility used, preferred language (Amharic or Afaan Oromo), and preferred survey time.

Statistical Analysis

We used descriptive statistics to describe patient demographics, enrollment, survey response, completion rates, and survey duration. A Heckman probit regression model was used to identify factors influencing survey participation (answering at least 1 question) and completion (answering all questions), with the first step determining participation factors (selection model). We chose the Heckman probit model over the standard probit model due to its advantages in predicting outcomes when there is a dependency between the selection model and the outcome model as well as a correlation of error terms between the 2 models [17,18]. The participation model specification was:

$$S^*=ZB+\varepsilon$$

Where S^* is a latent variable representing the participation decision in the IVR survey, Z is a vector of covariates that include the respondents' characteristics (sex, age, education level, timing of the IVR survey, type of facility visited, type of care received, and the patient's preferred language), B represents the coefficients of interest, and ε is the error term. The second step of the model aimed to identify the factors associated with the completion of the IVR survey (outcome variable), conditional on participation in the survey. This equation used the inverse Mills ratio, derived from the participation equation, to overcome selection bias.

The outcome (completion) model was specified as:

$$Y^*=X\psi+\lambda+u$$

Where Y^* is a latent variable representing the outcome, X is a vector of covariates that includes all factors listed in the selection model plus number of reconnections made to complete the survey, ψ represents the coefficients to be estimated, λ is the inverse Mills ratio from the participation equation, and u is the normally distributed error term. Adding λ adjusts the outcome equation for more accurate estimates of the relationship between participation and covariates. Language was used as a selection variable for participation in the IVR survey, with Afaan Oromo speakers from rural areas around Addis Ababa being more likely to participate than those from urban areas.

For those who completed the full survey, a Weibull regression model was used to identify the factors associated with the time taken to complete the survey. The main cost driver of an IVR survey is the airtime cost while a respondent answers the survey. Identifying factors influencing survey duration is helpful due to cost implications. Duration data are often skewed, with many early events and few late ones. The Weibull model is versatile for providing insights about the effects of covariates on the timing of events (survey completion).

Ethical Considerations

The study was approved by the Addis Ababa University College of Health Sciences institutional review board (approval 057/23/SPH). Ethics approval was also obtained from the Addis Ababa City Health Bureau. Only those participants who provided verbal consent were interviewed. All data were deidentified to ensure participant privacy and confidentiality. Participants who completed the full survey received approximately US $0.86 mobile phone airtime top-up.

Results

Recruitment

Table 1 reports the number of individuals who were approached and enrolled in the study. Of 3403 individuals approached at the study health facilities, 3167 (93.1%) were receiving outpatient care for themselves, and 2985 (87.7%) had a functioning mobile phone with them at the time. The majority (n=2685, 78.9%) had received 1 of 11 eligible primary care

XSL•FO
RenderX

services. Of these, 2415 (86.9%) agreed to participate, provided a valid phone number, and were enrolled in the IVR survey.

Table 1. Recruitment results: eligibility and consent to participate in the interactive voice response survey.

	Values, n (%)
Number of people approached (n= 3403)	
Receiving care for themselves	3238 (95.2)
Did not spend the night at the facility	3167 (93.1)
Had a functioning mobile phone	2985 (87.7)
Received eligible primary care services[a]	2685 (78.9)
Eligible to participate in the study (n= 2685)	
Agreed to participate	2604 (97)
Agreed to provide their mobile phone number	2469 (92)
Provided a valid mobile phone number	2415 (89.9)

[a]Participants were eligible to participate in the study if they received at least 1 of the following services: care for hypertension or high blood pressure, care for diabetes or high blood sugar, care for kidney disease, care for an injury, prenatal care, family planning services, care for tuberculosis, HIV or AIDS or a sexually transmitted infection (including testing or treatment), care for diarrhea or gastrointestinal conditions, or care for a respiratory illness.

IVR Survey Completion and Duration

Of 2415 individuals enrolled in the IVR survey, 2236 (92.6%) participants picked up the call on their mobile phone a day later. However, 16.6% (n=400) of these respondents only listened to the introduction, and another 10.4% (n=252) listened to the first question but failed to answer it, leading to 1584 (65.6%) respondents answering at least 1 question. Among them, 8.6% (n=208) answered part of the survey and hung up, and 14% (n=339) answered part of the survey but were disconnected by the system (for failing to respond or pressing invalid keys more than twice). Finally, 1037 (42.9%) respondents completed the full survey (answered all 15 questions). Those who completed the survey took on average 8.5 (SD 1.68) minutes to answer all questions, with a median of 7.55 (IQR 6.54-8.58) minutes, a minimum of 4 minutes and 4 seconds, and a maximum of 14 minutes and 59 seconds (Table 2).

Table 2. Interactive voice response (IVR) survey response rate.

	Values, n (%)
Enrolled in the IVR survey (n=2415)	
Picked up the IVR survey call	2236 (92.6)
Answered at least 1 question	1584 (65.6)
Participated in the IVR survey (n=1584)	
Answered at least 1 question but the system ended the call[a]	339 (14)
Answered at least 1 question but hung up without completing the survey	208 (8.6)
Answered the full survey	1037 (42.9)

[a]The IVR system disconnected the call if the user did any of the following 3 times: failed to reply within 20 seconds or pressed an invalid key.

Characteristics of the Study Respondents

Among the 2415 enrolled individuals, 1584 (65.7%) participated in the survey by responding to at least 1 survey question. The characteristics of the eligible individuals by participation status are reported in Table 3. A little over half of the individuals were female (n=1398, 57.9%), and the majority were younger than 40 years of age. Educational attainment varied, with nearly half of the respondents having education above the secondary level, and only 5.6% (n=134) with no formal education. In the latest Demographic and Health Survey, the proportion of Addis Ababa respondents with higher education was 27.7% only [19]. Although this survey is more than 8 years old, it is possible that our sample is wealthier and more educated than the general population, as they represent health system users, and half of them were enrolled in private health facilities. This may overrepresent the private sector share of health care services delivered in the city. The participants were selected nearly equally from each of the 4 health facility types, including public hospitals (n=628, 26%), health centers (n=624, 25.8%), private clinics (n=621, 25.7%), and private hospitals (n=542, 22.4%).

XSL•FO
RenderX

Table 3. Characteristics of the interactive voice response (IVR) sample by participation status among those enrolled.

Variable	Participants[a] (n=1584), n (%)	Nonparticipants (n=831), n (%)	Total enrolled (n=2415), n (%)
Sex			
Female	908 (57.3)	490 (58.9)	1398 (57.9)
Male	676 (42.7)	341 (41)	1017 (42.1)
Age group (years)			
18-29	623 (39.3)	278 (33.4)	901 (37.3)
30-39	516 (32.6)	249 (30)	765 (31.7)
40-49	219 (13.8)	145 (17.5)	364 (15.1)
≥50	226 (14.3)	159 (19.1)	385 (15.9)
Education			
No education at all	67 (4.2)	67 (8.1)	134 (5.6)
Primary education	326 (20.6)	172 (20.8)	498 (20.6)
Secondary education	436 (27.5)	229 (27.6)	665 (27.6)
Higher education	755 (47.7)	361 (43.6)	1116 (46.3)
Care type			
Maternal or reproductive	374 (23.7)	222 (26.8)	596 (24.8)
Communicable disease	106 (6.7)	58 (7)	164 (6.8)
Noncommunicable disease	661 (41.9)	347 (42)	1008 (41.9)
Diarrhea and respiratory	436 (27.7)	200 (24.2)	636 (26.5)
Language			
Afaan Oromo	274 (17.3)	106 (12.8)	380 (15.7)
Amharic	1310 (82.7)	725 (87.2)	2035 (84.3)
Preferred survey time			
Morning (8-11:59 AM)	319 (20.1)	155 (18.7)	474 (19.6)
Mid-day (12-1:59 PM)	323 (20.4)	193 (23.2)	516 (21.4)
Afternoon (2-5:59 PM)	269 (17)	173 (20.8)	442 (18.3)
Evening (6-7:59 PM)	320 (20.2)	164 (19.7)	484 (20)
Night (8-10 PM)	353 (22.3)	146 (17.7)	499 (20.7)
Facility type			
Health centers	443 (28)	181 (21.8)	624 (25.8)
Public hospitals	411 (26)	217 (26.1)	628 (26)
Private hospitals	333 (21)	209 (25.2)	542 (22.4)
Private clinic	397 (25.1)	224 (27)	621 (25.7)

[a]This shows the characteristics of those enrolled based on their participation status in the interactive voice response survey (answered at least 1 question or not).

Regarding the type of care received, around 41.9% (n=1008) of the participants sought care for noncommunicable diseases (hypertension, diabetes, or kidney disease), 26.5% (n=637) were treated for diarrhea or respiratory conditions, 24.8% (n=596) received maternal or reproductive health care, and 6.8% (n=164) received care for communicable diseases (including tuberculosis, HIV/AIDS, or sexually transmitted infections). The majority of respondents (n=2035, 84.3%) spoke Amharic. Preferences for survey times were fairly evenly distributed between morning, mid-day, afternoon, evening, and night time.

Survey Costs and Data Quality

A total of US $13,512 was spent to implement the survey, but 39.3% of these costs (US $5316) were driven by in-person enrollment, which was necessary to obtain the patients' phone numbers (Table 4). Health facilities, governments, or researchers that can obtain patient phone numbers directly from health records could avoid this portion of the cost. Excluding the in-person expenses, the IVR survey costs were US $7.90 per respondent who completed the full survey (n=1037). Considering those who answered at least 1 question (n=1584),

XSL•FO
RenderX

the survey costs were estimated at US $5.17 per partial survey. Of 2415 patients called, only 1584 answered at least 1 question. A total of 831 patients answered no survey questions. The costs for these patients were negligible and included only the cost of the SMS primer and of any airtime used during which the patient may have listened to the introduction and the first question without answering.

Table 4. Survey costs.

	Costs (US $)
In-person enrollment	
Including enumerator payments, transport, training, and supplies	5316
Interactive voice response survey costs	
Survey translation and recording	300
SMS primer (2415 SMS text messages sent)	1017
Survey airtime	5439
Incentive (airtime top-up)	1440
Total	13,512

As expected, the number of respondents answering the survey questions declined as the survey progressed, and 34.5% (547/1584) of respondents who began the survey did not complete all questions. For 11 of 15 questions, fewer than 2% of respondents chose to skip the question by pressing “9.” However, there were 3 notable exceptions. First, 36.5% (n=578) of respondents opted to skip the first question, which asked them to rate the cleanliness of the facility. Despite this, most participants went on to answer the second question and continued the survey. This high skip rate is likely because respondents were testing the survey technology on the first question. Future implementers should consider adding a test question at the beginning of their survey, given the high skip rate for our first survey question. Second, for question 13, which informed participants that only 3 questions remained, 16.1% (n=184) chose to skip this question, which asked about the affordability of the services received. Finally, 6.9% (n=72) of respondents skipped the last question, which asked about their monthly household income.

Factors Affecting Participation in and Completion of the IVR Survey

A Heckman sample selection model was used to predict covariates associated with IVR survey participation (answering at least 1 question) and completion. Before executing this model, we assessed its appropriateness against the standard probit model. The results indicated a correlation between the error terms of the 2 equations, as the value of athrho was statistically significant (P<.001). Furthermore, the findings showed a dependence between the selection model and the outcome model, justifying the use of the Heckman probit model since ρ is significantly different from 0 (χ^2_1=9.2; P=.002). Additionally, the likelihood function of the Heckman probit model was significant (Wald χ^2_{19}=269.9; P<.001), indicating the better explanatory power of the Heckman probit model.

The model showed that as the age of the participants increased, the likelihood of both participation and completion declined (Table 5). For example, there was a significant difference in the odds of participating in (odds ratio [OR] 0.77, 95% CI 0.65-0.90) and completing the IVR survey (OR 0.63, 95% CI 0.53-0.74) between participants aged 50 years and older compared to those aged 18-30 years. Similarly, higher education increased the odds of participation and completion. For instance, for respondents with postsecondary education, the odds of participation were 1.54 times higher (95% CI 1.25-1.90) and the odds of survey completion were 1.74 times higher (95% CI 1.46-2.06) compared to those with no education.

XSL•FO
RenderX

Table 5. Results of the Heckman probit model.

Variables	Completion of the IVR[a] survey (2402 observations), OR[b] (95% CI)	Participation in the IVR survey (2402 observations), OR (95% CI)
Sex (reference=female)		
Male	1.010 (0.907-1.125)	0.983 (0.880-1.097)
Age category (years) (reference 18-30 years)		
30-39	0.853[c] (0.757-0.960)	0.958 (0.846-1.084)
40-49	0.666[c] (0.569-0.779)	0.785[c] (0.668-0.921)
≥50	0.628[c] (0.534-0.739)	0.765[c] (0.648-0.904)
Education level (reference=no education)		
Primary	0.993 (0.831-1.188)	1.339[c] (1.081-1.658)
Secondary	1.318[c] (1.111-1.563)	1.337[c] (1.085-1.648)
Above secondary	1.737[c] (1.462-2.063)	1.540[c] (1.248-1.901)
Time of the survey (reference=morning [8-11:59 AM])		
Mid-day (12-1:59 PM)	1.017 (0.873-1.184)	0.861[d] (0.735-1.007)
Afternoon (2-5:59 PM)	0.963 (0.819-1.133)	0.826[e] (0.699-0.975)
Evening (6-7:59 PM)	0.971 (0.831-1.135)	0.920 (0.783-1.081)
Night (8-10 PM)	1.163[d] (0.995-1.361)	1.047 (0.888-1.236)
Facility type (reference=public health center)		
Public hospital	0.866[d] (0.748-1.002)	0.920 (0.791-1.070)
Private hospital	0.797[c] (0.677-0.938)	0.724[c] (0.611-0.857)
Private clinic	0.840[e] (0.727-0.972)	0.819[c] (0.705-0.952)
Number of reconnection (reference=0)		
One	0.588[c] (0.514-0.673)	—[f]
Two or more	0.472[c] (0.365-0.611)	—
Type of care (reference=NCD[g])		
Maternal and reproductive	0.804[c] (0.699-0.925)	0.809[c] (0.700-0.934)
Communicable diseases	0.950 (0.775-1.165)	0.904 (0.735-1.110)
Diarrhea and respiratory	0.972 (0.854-1.107)	0.992 (0.867-1.135)
Language (reference=Afaan Oromo)		
Amharic	—	0.852[c] (0.761-0.954)
Athrho	— (10,193-1.090e+09)	—
Constant	—	1.810[c] (1.452-2.256)

[a]IVR: interactive voice response.
[b]OR: odds ratio.
[c]$P<.01$.
[d]$P<.1$.
[e]$P<.05$.
[f]Not applicable.
[g]NCD: noncommunicable disease.

We also found that participation in and completion of the IVR survey was also affected by time of day. Those who were called in the afternoon had lower odds of participating compared to those called in the morning (OR 0.84, 95% CI 0.70-0.98). There

XSL•FO
RenderX

was no difference in survey completion according to time of day.

Compared to patients who used services in health centers, those from private hospitals and private clinics were less likely to participate in and complete the survey. For example, primary care patients from private hospitals were substantially less likely to participate in (OR 0.72, 95% CI 0.61-0.86) and complete (OR 0.80, 95% CI 0.68-0.95) the survey compared to those in health centers.

Participants who had received maternal and reproductive health services were less likely to participate in and complete the IVR survey compared to those who received care for noncommunicable diseases. Another important factor identified in determining the completion of the IVR survey was the number of reconnection attempts made. The results revealed that as the number of reconnections increased, participants were less likely to complete the survey. For example, there was a significant difference in completion rates between participants reconnected twice and those not reconnected with an OR of 0.47 (95% CI 0.37-0.61).

Factors Affecting the Time Taken to Complete the IVR Survey

Results from the Weibull model revealed duration dependency, with a *P* value greater than 1 (*P*=5.037), implying that the hazard function increased over time. This suggests that IVR survey participants were more likely to complete the survey as time progressed. The results indicated that sex, age, time of day, language, and facility type were associated with the duration of IVR survey completion (Table 6).

XSL•FO
RenderX

Table 6. Results from Weibull regression (1034 observations).

Variables	Hazard rate (95% CI)
Sex (reference=female)	
Male	1.148[a] (1.005-1.311)
Age category (years) (reference<30 years)	
30-39	0.957 (0.828-1.106)
40-49	0.933 (0.759-1.146)
≥50	0.744[a] (0.589-0.941)
Education level (reference=no education)	
Primary	0.867 (0.565-1.329)
Secondary	0.972 (0.641-1.476)
Above secondary	1.243 (0.824-1.876)
Time of the IVR[b] (reference=morning [8-11:59 AM])	
Mid-day (12-1:59 PM)	1.139 (0.929-1.396)
Afternoon (2-5:59 PM)	1.034 (0.834-1.282)
Evening (6-7:59 PM)	1.078 (0.878-1.323)
Night (8-10 PM)	1.293[a] (1.063-1.574)
Language (reference=Afaan Oromo)	
Amharic	1.911[c] (1.607-2.273)
Facility type (reference=health center)	
Public hospital	1.003 (0.836-1.203)
Private hospital	1.210[a] (1.003-1.459)
Private clinic	1.164 (0.967-1.400)
Type of care (reference=NCD[d])	
Maternal and reproductive	1.140 (0.950-1.367)
Communicable diseases	0.983 (0.759-1.272)
Diarrhea and respiratory	1.154[e] (0.988-1.347)
P value	5.037[c] (4.826-5.258)
Constant	0[c] (0-0)

[a]*P*<.05.
[b]IVR: interactive voice response.
[c]*P*<.01.
[d]NCD: noncommunicable disease.
[e]*P*<.10.

Male participants were 1.15 times more likely to complete the survey early compared to female participants, controlling for other factors. As age increased, individuals were less likely to complete the survey early. For instance, the odds of completing the IVR survey quickly for participants aged 50 years and older was 0.74 (95% CI 0.59-0.94) compared to those participants aged 18 to 30 years. The odds of completing the IVR survey faster were higher at night (OR 1.29, 95% CI 1.06-1.57) compared to morning calls. The odds of completing the survey faster were higher for those enrolled in private hospitals (OR 1.21, 95% CI 1.03-1.46) compared to those from health centers. The regression model also controlled for survey language as a proxy for geographical area. Most people speaking Afaan Oromo were from rural areas surrounding Addis Ababa, while those responding in Amharic were from the urban area. The results showed that participants responding in Amharic were more likely to complete the survey early with an OR of 1.91 (95% CI 1.61-2.27).

XSL•FO
RenderX

Discussion

Principal findings

This study assessed the feasibility of using an IVR survey as an alternative approach to collect patient satisfaction data in Addis Ababa, Ethiopia. The findings demonstrate a high rate of mobile phone ownership and a willingness among patients to participate, indicating that IVR surveys can serve as a viable low-cost alternative to traditional survey methods to assess patient satisfaction in similar urban areas in a low-income country. Notably, 92.6% (n=2236) of those enrolled picked up the call, 65.6% (n=1584) answered at least 1 survey question, and 42.9% (n=1037) completed the full survey in an average time of 8.1 (SD 1.7) minutes. In adjusted regression models, being older, having less education, and using private facilities negatively affected survey participation and completion. In contrast, being male, being younger, speaking Amharic, using a private hospital, and being called after 8 PM were associated with a shorter survey duration.

The completion rate of 42.9% (n=1037) is encouraging, as almost all of the participants had limited experience with this type of technology. Previous IVR survey studies in Bangladesh, Burkina Faso, Tanzania, and Uganda have found varying response rates ranging from 1% to 41% [20-22]. Response rates were affected by the sampling methods (in-person household surveys or random digit dialing) and type of incentives. Our study, which sampled respondents in person in health facilities and used incentives, achieved a slightly higher response rate (n=1037, 42.9%) than previous studies. In the United States, IVR surveys have been used in various patient satisfaction studies including after endoscopies [11] and for child and adolescent psychiatric outpatient treatment [8]. According to EngageSpark, a global survey and engagement platform, IVR campaigns generally have higher acceptance rates than SMS campaigns. IVR surveys typically achieve response rates ranging from 25% to 75%, while the acceptance rate for SMS surveys ranges from 5% to 15% [23].

Our findings also highlighted the challenges of conducting an IVR survey. For instance, around 16.6% (n=400) of participants listened to the introduction but hung up before answering any questions, and the other 10.4% (n=252) failed to answer the first question despite staying on the call. This suggests that while the initial engagement was successful, there may be factors in the introduction or in the survey's initial questions that could be optimized to retain participants' attention or improve their understanding of the system.

We identified several factors influencing participation and completion of the IVR survey. Age emerged as a significant factor, with younger participants being more likely to participate in and complete the IVR survey compared to older participants. This finding aligns with previous studies suggesting that younger individuals are more comfortable with technology and value the privacy that IVR surveys can provide [24]. Education level also significantly impacts both participation and completion rates, with individuals with higher education levels being more likely to engage with IVR surveys. This supports the notion that higher education correlates with better use of technology and higher perceived importance of surveys [25,26].

The type of facility visited by participants influenced the likelihood of participating in and completing the survey. Participants from private hospitals and clinics were less likely to engage with the survey compared to those from health centers. This could be due to the perception that feedback is more likely to improve services in public facilities, which typically serve poorer segments of the population. Another potential reason could be the provision of incentives that attracted lower-income individuals who are more likely to use affordable health care from public health centers [27]. Our incentive consisted of approximately US $0.86 mobile top-up for those who completed the full survey.

The Weibull model for factors associated with the time taken to complete the survey suggested that the hazard function increases over time and that participants were more likely to complete the IVR survey as time progressed [28,29]. The regression results showed that younger individuals tend to complete the survey faster likely due to their greater familiarity with technology. This finding is consistent with literature indicating that older adults may face more challenges with technology, leading to longer completion times [20,30]. The time of day when the IVR survey was conducted also affected completion times. Surveys conducted at night were completed faster compared to those conducted in the morning. This could be due to participants having fewer distractions at night or that those completing the survey at night tend to be more educated. This finding aligns with the idea that identifying optimal times for survey calls can increase participation [24].

Although participants from private hospitals were less likely to participate in and complete the survey, those who completed it were more likely to have a shorter survey duration compared to those from health centers. This may be because higher-income individuals, who are more likely to visit private facilities, are more efficient in completing telephone-based surveys. Language, used as a proxy for area of residence, shows that participants responding in Amharic were more likely to complete the survey faster compared to those responding in Afaan Oromo. This suggests that urban participants, who are more likely to respond in Amharic, complete the survey faster than rural participants. Access to technology and familiarity with IVR systems in urban and rural areas are likely contributing factors.

To further inform our study, a debrief was organized with data collectors who enrolled participants in person and explained the IVR technology. During enrollment, patients revealed concerns with privacy, confusion about the technology, and a preference for in-person interviews. Some individuals were reluctant to share their mobile numbers due to fears of misuse, and face-to-face interactions were sometimes seen as more confidential. Language barriers also hindered participation, as some patients preferred languages that were not supported by our IVR survey. Technical challenges further complicate IVR use, as participants may struggle to navigate prompts without prior experience and with low educational backgrounds and poor literacy. Clear instructions are crucial for engagement. The study showed that older and less-educated individuals were

XSL•FO
RenderX

substantially less likely to participate in the phone-based study. This is consistent with evidence that older people and those with no education have limited digital competencies and thereby limited technological abilities [21,22,31].

Conclusions

IVR systems offer a promising tool for health system research in urban settings of a low-income country. Nonetheless, the implementation of an IVR survey among patients must be carefully planned and tailored to address local challenges. Addressing privacy concerns and building trust are essential. It is also necessary to ensure language inclusivity to improve participation rates and the overall effectiveness of IVR-based public health surveys. Others have found that providing incentives, such as mobile airtime top-ups, can encourage participation and improve response rates, especially for people of lower socioeconomic status [32,33]. Finally, using simple and common language, particularly for medical terms, may ensure better understanding among participants. In considering whether to implement an IVR survey as opposed to an in-person or a computer-assisted telephone interviewing survey with live interviewers, researchers should consider certain tradeoffs. First, we found that the IVR survey was cheap (US $7.90 per full survey); it was implemented quickly and is highly scalable to large samples as long as patient phone numbers can be obtained. However, although we did not find major data quality gaps, IVR may be more prone to data quality issues if users struggle to understand the questions or to answer correctly on their mobile phone keypads. In low-income country settings, in-person surveys may also be affordable if labor costs are low. However, in-person and computer-assisted telephone interviewing surveys take more time to implement and therefore are less scalable, as the number of interviewers and the time taken to conduct surveys must be multiplied when the sample size increases. With IVR, the survey is programmed once and can be answered simultaneously by thousands of respondents simultaneously. In addition, in-person surveys do not ensure privacy and are more prone to social desirability bias than IVR [6].

IVR is much cheaper than other survey methods. For example, health facility surveys such as the Service Provision Assessment surveys (conducted by the Demographic and Health Survey program) that include nationally representative samples of health facilities and patient exit interviews generally cost between US $800 and US $1500 per surveyed health facility [31]. Similarly, the World Bank estimated that a household survey in Sub-Saharan Africa will cost approximately US $300 per household. In addition, IVR surveys tend to be completed much faster. Our entire study was completed in 2 weeks (most of the time was taken for the in-person recruitment), and all data were immediately available through the IVR platform and ready for analysis [32].

There have been few studies in low- and middle-income countries that have tested the use of IVR to survey patients. To our knowledge, ours is the first study to use IVR to collect patient feedback in Ethiopia. IVR offers a low-cost and rapid solution to survey patients and obtain information on perceived quality and satisfaction. In high-income countries, patient satisfaction surveys are often conducted by mail on the web. These survey modes are largely not possible in Ethiopia. In Ethiopia, IVR surveys are likely to function well for younger and more educated respondents in an urban context. Future research should investigate feasibility and response rates in rural areas. Governments and health facilities should consider integrating IVR surveys into routine operations to gather patient satisfaction data and information on patient-perceived quality of care, which can inform quality improvement initiatives and strategies. By leveraging IVR technology, health systems can obtain timely and actionable feedback from patients, ultimately contributing to improved health care quality and better patient outcomes.

Acknowledgments

The authors thank colleagues from the Ethiopian Civil Service University and Addis Ababa University for their support in implementing the study as well as Alyssa Khan from the George Washington University for her support in the study design. This work was supported by a Research Innovation Award from the Office of Research Excellence, Milken Institute School of Public Health, The George Washington University, United States (grant 02RIA022023). Health facility directors were involved in the design and conduct of the study. Results were disseminated among health facility managers through policy briefs. Patients completed the survey and shared their opinions on the acceptability of the interactive voice response method.

Data Availability

The datasets generated and during this study are available from the corresponding author on reasonable request.

Authors' Contributions

DS and ADM contributed to designing the survey, data cleaning, data analysis, and report writing. CA contributed to designing the survey, data cleaning, data analysis, and editing and report writing.

Conflicts of Interest

None declared.

Multimedia Appendix 1

XSL•FO
RenderX

Interactive voice response survey instrument on the quality of primary health care in Addis Ababa health facilities. [DOCX File , 22 KB - formative_v9i1e67452_app1.docx]

References

1. Al-Abri R, Al-Balushi A. Patient satisfaction survey as a tool towards quality improvement. Oman Med J 2014;29(1):3-7 [FREE Full text] [doi: 10.5001/omj.2014.02] [Medline: 24501659]
2. Tiperneni R, Patel S, Heis F, Ghali S, Du D, Ghali W, et al. HCAHPS: having constant communication augments hospital and patient satisfaction. BMJ Open Qual 2022;11(4):e001972 [FREE Full text] [doi: 10.1136/bmjoq-2022-001972] [Medline: 36253015]
3. The mobile economy Sub-Saharan Africa. Global System for Mobile Communications. 2024. URL: https://www.gsma.com/mobileeconomy/sub-saharan-africa/ [accessed 2025-01-18]
4. Holbrook AL, Green MC, Krosnick JA. Telephone versus face-to-face interviewing of national probability samples with long questionnaires: comparisons of respondent satisficing and social desirability response bias. Public Opin Q 2003;67(1):79-125 [FREE Full text] [doi: 10.1086/346010]
5. Hughes S. The challenges and opportunities for collecting survey data in Sub-Saharan Africa. Mathematica. URL: https://www.mathematica.org/blogs/the-challenges-and-opportunities-for-collecting-survey-data-in-sub-saharan-africa [accessed 2019-02-04]
6. Gibson DG, Farrenkopf BA, Pereira A, Labrique AB, Pariyo GW. The development of an interactive voice response survey for noncommunicable disease risk factor estimation: technical assessment and cognitive testing. J Med Internet Res 2017;19(5):e112 [FREE Full text] [doi: 10.2196/jmir.7340] [Medline: 28476724]
7. Dillon B. Using mobile phones to collect panel data in developing countries. J Int Dev 2011;24(4):518-527 [FREE Full text] [doi: 10.1002/jid.1771]
8. Andersson C, Danielsson S, Silfverberg-Dymling G, Löndahl G, Johansson BA. Evaluation of interactive voice response (IVR) and postal survey in follow-up of children and adolescents discharged from psychiatric outpatient treatment: a randomized controlled trial. Springerplus 2014;3:77 [FREE Full text] [doi: 10.1186/2193-1801-3-77] [Medline: 24567883]
9. 5 Key benefits of interactive voice response (IVR). Primas. 2023. URL: https://primas.net/5-key-benefits-of-interactive-voice-response-ivr/ [accessed 2023-05-05]
10. Rigotti NA, Chang Y, Rosenfeld LC, Japuntich SJ, Park ER, Tindle HA, et al. Interactive voice response calls to promote smoking cessation after hospital discharge: pooled analysis of two randomized clinical trials. J Gen Intern Med 2017;32(9):1005-1013 [FREE Full text] [doi: 10.1007/s11606-017-4085-z] [Medline: 28616847]
11. Altman C, Birraux T, Lapuelle J, Letard JC, Tusseau F, Canard JM. Assessment of patient satisfaction with endoscopy using an interactive voice response system. Gastroenterol Clin Biol 2006 Mar;30(3):371-376 [FREE Full text] [doi: 10.1016/s0399-8320(06)73190-6] [Medline: 16633301]
12. Holland SM, Shuk E, Burkhalter J, Shouery M, Li Y, Hay JL. Feasibility and acceptability of using an IVRS to assess decision making about sun protection. Psychooncology 2020;29(1):156-163 [FREE Full text] [doi: 10.1002/pon.5231] [Medline: 31520426]
13. Daftary A, Hirsch-Moverman Y, Kassie GM, Melaku Z, Gadisa T, Saito S, et al. A qualitative evaluation of the acceptability of an interactive voice response system to enhance adherence to isoniazid preventive therapy among people living with HIV in Ethiopia. AIDS Behav 2017;21(11):3057-3067 [FREE Full text] [doi: 10.1007/s10461-016-1432-8] [Medline: 27221743]
14. Howard AA, Hirsch-Moverman Y, Saito S, Gadisa T, Daftary A, Melaku Z. The ENRICH Study to evaluate the effectiveness of a combination intervention package to improve isoniazid preventive therapy initiation, adherence and completion among people living with HIV in Ethiopia: rationale and design of a mixed methods cluster randomized trial. Contemp Clin Trials Commun 2017;6:46-54 [FREE Full text] [doi: 10.1016/j.conctc.2017.03.001] [Medline: 28626811]
15. Leo B, Morello R, Mellon J, Peixoto T, Davenport ST. Do mobile phone surveys work in poor countries? SSRN J 2015:398. [doi: 10.2139/ssrn.2597885]
16. Pegus CM. Crowdsourcing citizen feedback on district development in Ghana using interactive voice response surveys. The Institute of Development Studies. 2016. URL: https://opendocs.ids.ac.uk/articles/report/Crowdsourcing_citizen_feedback_on_district_development_in_Ghana_using_interactive_voice_response_surveys/26476822?file=48249943 [accessed 2025-01-18]
17. Wooldridge JM. Econometric Analysis of Cross Section and Panel Data. Cambridge: MIT press; 2010.
18. Koné S, Bonfoh B, Dao D, Koné I, Fink G. Heckman-type selection models to obtain unbiased estimates with missing measures outcome: theoretical considerations and an application to missing birth weight data. BMC Med Res Methodol 2019;19(1):231. [doi: 10.1186/s12874-019-0840-7] [Medline: 31815610]
19. Ethiopia: Demographic and Health Survey. Central Statistical Agency (CSA) [Ethiopia] and ICF. 2016. URL: https://www.dhsprogram.com/pubs/pdf/FR328/FR328.pdf [accessed 2025-01-18]
20. Chróinín DN, Hunter C, Shaw-Jones M, Duruchukwu E, Huang F, Ngov A, et al. Identifying the needs of our older community-based patients in the COVID era: a telephone survey. Aging Health Res 2022;2(2):100073 [FREE Full text] [doi: 10.1016/j.ahr.2022.100073] [Medline: 35434693]

XSL•FO
RenderX

21. Chen K, Chan AHS. Gerontechnology acceptance by elderly Hong Kong Chinese: a senior technology acceptance model (STAM). Ergonomics 2014;57(5):635-652 [FREE Full text] [doi: 10.1080/00140139.2014.895855] [Medline: 24655221]
22. Marasinghe KM, Lapitan JM, Ross A. Assistive technologies for ageing populations in six low-income and middle-income countries: a systematic review. BMJ Innov 2015;1(4):182-195 [FREE Full text] [doi: 10.1136/bmjinnov-2015-000065] [Medline: 26688747]
23. Ultimate guide to interactive voice response (IVR) for social impact programs. EngageSpark. URL: https://www.engagespark.com/guides/ultimate-guide-interactive-voice-response-ivr-social-impact-programs [accessed 2025-01-16]
24. Rutebemberwa E, Namutundu J, Gibson DG, Labrique AB, Ali J, Pariyo GW, et al. Perceptions on using interactive voice response surveys for non-communicable disease risk factors in Uganda: a qualitative exploration. Mhealth 2019;5:32 [FREE Full text] [doi: 10.21037/mhealth.2019.08.12] [Medline: 31620459]
25. Groves RM, Couper MP. Nonresponse in Household Interview Surveys. Hoboken, NJ: John Wiley & Sons; 2012.
26. Dillman DA, Smyth JD, Christian LM. Internet, phone, mail and mixed-mode surveys: the tailored design method. Reis 2016;154:161-176. [doi: 10.1002/9781394260645]
27. Smith MG, Witte M, Rocha S, Basner M. Effectiveness of incentives and follow-up on increasing survey response rates and participation in field studies. BMC Med Res Methodol 2019;19(1):230-213 [FREE Full text] [doi: 10.1186/s12874-019-0868-8] [Medline: 31805869]
28. Phadnis MA, Sharma P, Thewarapperuma N, Chalise P. Assessing accuracy of Weibull shape parameter estimate from historical studies for subsequent sample size calculation in clinical trials with time-to-event outcome. Contemp Clin Trials Commun 2020;17:100548 [FREE Full text] [doi: 10.1016/j.conctc.2020.100548] [Medline: 32154431]
29. Gómez YM, Gallardo DI, Marchant C, Sánchez L, Bourguignon M. An in-depth review of the Weibull model with a focus on various parameterizations. Mathematics 2023;12(1):56 [FREE Full text] [doi: 10.3390/math12010056]
30. Czaja SJ, Charness N, Fisk AD, Hertzog C, Nair SN, Rogers WA, et al. Factors predicting the use of technology: findings from the Center for Research and Education on Aging and Technology Enhancement (CREATE). Psychol Aging 2006;21(2):333-352 [FREE Full text] [doi: 10.1037/0882-7974.21.2.333] [Medline: 16768579]
31. Bertolazzi A, Quaglia V, Bongelli R. Barriers and facilitators to health technology adoption by older adults with chronic diseases: an integrative systematic review. BMC Public Health 2024;24(1):506 [FREE Full text] [doi: 10.1186/s12889-024-18036-5] [Medline: 38365698]
32. Abdelazeem B, Abbas KS, Amin MA, El-Shahat NA, Malik B, Kalantary A, et al. The effectiveness of incentives for research participation: a systematic review and meta-analysis of randomized controlled trials. PLoS One 2022;17(4):e0267534 [FREE Full text] [doi: 10.1371/journal.pone.0267534] [Medline: 35452488]
33. Ichimiya M, Muller-Tabanera H, Cantrell J, Bingenheimer JB, Gerard R, Hair EC, et al. Evaluation of response to incentive recruitment strategies in a social media-based survey. Digit Health 2023;9:20552076231178430 [FREE Full text] [doi: 10.1177/20552076231178430] [Medline: 37312941]

Abbreviations

IVR: interactive voice response
OR: odds ratio

Edited by A Mavragani; submitted 11.10.24; peer-reviewed by A Greenleaf; comments to author 18.11.24; revised version received 08.12.24; accepted 24.12.24; published 13.02.25.

Please cite as:
Shamebo D, Derseh Mebratie A, Arsenault C
Using an Interactive Voice Response Survey to Assess Patient Satisfaction in Ethiopia: Development and Feasibility Study
JMIR Form Res 2025;9:e67452
URL: https://formative.jmir.org/2025/1/e67452
doi:10.2196/67452
PMID:

©Dessalegn Shamebo, Anagaw Derseh Mebratie, Catherine Arsenault. Originally published in JMIR Formative Research (https://formative.jmir.org), 13.02.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Exploring Metadata Catalogs in Health Care Data Ecosystems: Taxonomy Development Study

Simon Scheider[1], PhD; Mostafa Kamal Mallick[1], MSc

Fraunhofer Institute for Software and Systems Engineering, Dortmund, Germany

Corresponding Author:
Simon Scheider, PhD
Fraunhofer Institute for Software and Systems Engineering
Speicherstraße 6
Dortmund, 44147
Germany
Phone: 49 231976774
Email: simon.scheider@isst.fraunhofer.de

Abstract

Background: In the European health care industry, recent years have seen increasing investments in data ecosystems to "FAIRify" and capitalize the ever-rising amount of health data. Within such networks, health metadata catalogs (HMDCs) assume a key function as they enable data allocation, sharing, and use practices. By design, HMDCs orchestrate health information for the purpose of findability, accessibility, interoperability, and reusability (FAIR). However, despite various European initiatives pushing health care data ecosystems forward, actionable design knowledge about HMDCs is scarce. This impedes both their effective development in practice and their scientific exploration, causing huge unused innovation potential of health data.

Objective: This study aims to explore the structural design elements of HMDCs, classifying them alongside empirically reasonable dimensions and characteristics. In doing so, the development of HMDCs in practice is facilitated while also closing a crucial gap in theory (ie, the literature about actionable HMDC design knowledge).

Methods: We applied a rigorous methodology for taxonomy building following well-known and established guidelines from the domain of information systems. Within this methodological framework, inductive and deductive research methods were applied to iteratively design and evaluate the evolving set of HMDC dimensions and characteristics. Specifically, a systematic literature review was conducted to identify and analyze 38 articles, while a multicase study was conducted to examine 17 HMDCs from practice. These findings were evaluated and refined in 2 extensive focus group sessions by 7 interdisciplinary experts with deep knowledge about HMDCs.

Results: The artifact generated by the study is an iteratively conceptualized and empirically grounded taxonomy with elaborate explanations. It proposes 20 dimensions encompassing 101 characteristics alongside which FAIR HMDCs can be structured and classified. The taxonomy describes basic design characteristics that need to be considered to implement FAIR HMDCs effectively. A major finding was that a particular focus in developing HMDCs is on the design of their published dataset offerings (ie, their metadata assets) as well as on data security and governance. The taxonomy is evaluated against the background of 4 use cases, which were cocreated with experts. These illustrative scenarios add depth and context to the taxonomy as they underline its relevance and applicability in real-world settings.

Conclusions: The findings contribute fundamental, yet actionable, design knowledge for building HMDCs in European health care data ecosystems. They provide guidance for health care practitioners, while allowing both scientists and policy makers to navigate through this evolving research field and anchor their work. Therefore, this study closes the research gap outlined earlier, which has prevailed in theory and practice.

(JMIR Form Res 2025;9:e63396) doi:10.2196/63396

KEYWORDS

data catalogs; data ecosystems; findability, accessibility, interoperability, and reusability; FAIR; health care; metadata; taxonomy

XSL•FO
RenderX

Introduction

Challenges of Health Care Systems

In the 21st century, health care systems worldwide are experiencing a tremendous increase in data, driven by advances in medical technology, digital health records, and wearable devices [1]. This flood of data holds immense potential for data-driven health innovations, building upon large-scale real-world data (RWD) and real-world evidence (RWE) [2]. However, health care systems face multiple challenges that hinder the effective use of RWD to generate RWE and thus data-driven health innovations. One primary issue is the integration of heterogeneous datasets [3]. RWD stem from diverse sources, such as electronic health records, imaging, and genomic data, frequently exhibiting incompatible or unknown data formats, which complicates harmonization, particularly across different entities [4,5]. Furthermore, finding and accessing suitable RWD represents a hurdle for medical research due to their origins from disparate patient populations, health care systems, and data collection methodologies [6]. This impairs the effective discovery of and access to a sufficient number of both available and adequate datasets. Moreover, even if enough RWD are discovered and access is established, another challenge lies in ensuring scientific rigor and reproducibility of generated RWE (ie, medical studies) that becomes increasingly difficult in today's data-intensive health research [7]. Medical studies constantly require larger, high-quality datasets to generate meaningful and reproducible RWE [5,6,8]. However, unknown RWD management practices threaten study reliability, while making the validation of results (ie, RWE) across studies difficult [2,7]. Besides that, the diversity of national health care systems increases the prevailing differences in health care data infrastructures across countries that, in turn, lead to additional barriers for organizations to share and use RWD at a large scale [9]. Moreover, health care systems must navigate complex legal requirements with regard to sharing and processing RWD [8,9]. For instance, the European jurisdiction mandates strict data governance and security standards that, while essential, impair data-driven health innovations in the absence of adequate data-sharing infrastructures [9]. As a result, RWD are fragmented and isolated within single organizations, whereby data sharing and use are limited. Because of all these challenges, the rapidly increasing amount of RWD cannot be harnessed to its full potential for producing health care innovations (ie, RWE).

Metadata Catalogs as a Promising Solution

Against this background, data ecosystems as technical and organizational infrastructures within the healthcare sector represent auspicious solutions, that is, *health care data ecosystems* (amplified in, eg, the studies by Lovestone and EMIF Consortium [10], Manogaran et al [11], and Sharon and Lucivero [12]). These evolving networks enable legally compliant use of RWD [13]. Their key function is market mechanisms instantiated by metadata catalogs [5,14] that define and describe the intricate web of RWD circulating between a potentially arbitrary number of actors in the ecosystem [15]. Hence, such *health metadata catalogs* (HMDCs) are crucial components of modern health care data ecosystems, for example, EHDEN (European Health Data and Evidence Network), EHDS2 (European Health Data Space 2), Elixir, EUCAIM (European Federation for Cancer Images), IDERHA (Integration of Heterogeneous Data and Evidence towards Regulatory and Health Technology Assessment Acceptance), and Gaia-X. Multimedia Appendix 1 provides a comprehensive overview about the most important European Union (EU) initiatives. For example, EHDS2 aims at creating a European infrastructure for the secure exchange and secondary use of health data across EU member states [16]. Therein, HMDCs belong to the core infrastructure services to enable standardized organization of and controlled access to RWD for research. The pilot infrastructure of the EHDS is implemented by the HealthData@EU initiative.

Since HMDCs provide an effective method for systematically sharing and using RWD within data ecosystems [3], they potentially allow harnessing crucial benefits corresponding to the challenges outlined earlier. First, HMDCs facilitate integrating heterogenous datasets across health care systems. They help to transcend diverse data types, which eases the integration, standardization, and harmonization of data within data ecosystems [3,5]. This is essential for medical research that requires huge pools of accessible RWD appropriate for their investigations [7]. Second, as finding and accessing adequate RWD effectively is vital for medical research [6], HMDCs entail added value by offering a governed data search and access framework embedded into the technical infrastructures of the underlying ecosystems [17]. They provide a tool for data discovery to precisely characterize, locate, and filter RWD on the basis of a myriad of factors [5,17]. Third, HMDCs support transparent and reproducible research processes by helping scientists in replicating studies and validating their results [18]. Such transparency is fundamental for building trust in the reliability and validity of RWE [2]. Finally, HMDCs facilitate data-intensive research, generally, as they establish unified health care data infrastructures for allocating, accessing, and using RWD of connected data providers [19]. In doing so, they reduce barriers for organizations to integrate their otherwise isolated RWD within data ecosystems. At the same time, HMDCs bridge prevailing differences between national health care systems and retain full control of data providers [5,19]. To this end, they establish robust data security and governance frameworks that are aligned to the applicable jurisdictions [5,15].

As a result, HMDCs represent an auspicious medium against the fragmentation and isolation of RWD [5]. However, since HMDCs are novel constructs, typically in premature phases [20], their ascribed benefits are primarily backed by the literature rather than evidence from practice. Nevertheless, HMDCs are likely to provide means for using RWD systematically within and across health care data ecosystems, potentially resulting in more efficient RWE generation.

In Europe, HMDCs are of particular importance because the EU health care sector exhibits a broad diversity across member states, all with their own health care systems and policies. Consequently, there is a need to focus heavily on standardizing and harmonizing both data and metadata across different countries for facilitating data sharing and legally compliant data

XSL•FO
RenderX

use [21]. More specifically, the diversity of national health care systems [22], the restrictiveness of data protection regulations [8,9], and the fragmentation and isolation of health data [23] make operative health care data ecosystems and HMDCs a paramount concern for the European health care industry. Therefore, this study adopts a European focus.

Theoretical Background

Originally, data catalogs are organized collections of datasets that provide descriptive information within an organization [24,25]. They act as centralized repositories, making it easier for data consumers to discover, understand, and access the information they need [26]. Enterprise data management platforms often comprise such centralized data catalogs implying storage of data within their peripheries [25,27]. If data are not encapsulated within the organization but integrated into decentral or federated networks [28], the literature commonly refers to such environments as data ecosystems with metadata catalogs as key function [5,14]. This study considers metadata catalogs as decentralized or federated constructs that are mutually exclusive to centralized ones. For simplification, only the term decentralized is used.

Metadata describe dataset attributes, such as source, format, structure, provenance, owner, access, or governance modalities [29]. Metadata catalogs act as "catalogues of data catalogues," dedicated to enhancing discoverability, usability, and management of distributed datasets [30]. Within data ecosystems, metadata catalogs are a mechanism that provides a standardized way for recording, disclosing, and making available information about all relevant kinds of phenotypes describing datasets, while ensuring legally compliant access and sharing practices [3,14]. If these datasets are health data, it is henceforth referred to such constructs as *HMDCs*. Consequently, HMDCs manage heterogenous health information integrated into health care data ecosystems [5]. They ensure that these diverse and highly sensitive datasets are effectively organized and understood [4], while facilitating their systematic use [3,5,14]. This requires dedicated, yet unknown, design elements to be unveiled by the study.

Research Gap, Objective, and Questions

After having clarified the added value of HMDCs, the research gap is demarcated by reviewing related work. Therefrom, the research problem is identified, which leads to the research objectives. These objectives then allow to derive the research questions, required to define a meaningful research methodology.

To begin with, Labadie et al [24] foster the understanding of data catalogs by classifying corresponding initiatives. The authors propose a taxonomy for data catalogs and present 3 case studies. However, similar to Ehrlinger et al [26] and Jahnke and Otto [25], Labadie et al [24] focus on intraorganizational data sharing using centralized catalogs. Moreover, they neglected health use cases. Remy et al [3] conducted a design science study to build an integrated catalog for health research metadata. The artifact enables medical scientists to analyze phenomena that require a view across several domains. The authors are among the first who provide design knowledge usable in HMDC contexts. Although, similar to the findings of the previously presented literature sources, Remy et al [3] accentuate centralized catalogs. Almeida et al [19] present a platform that provides a set of tools, compliant with the findability, accessibility, interoperability, and reusability (FAIR) principles, to help data holders sharing biomedical databases while allowing data consumers to discover and apply for them. However, the authors only consider a narrow use case instead of generating universally applicable design knowledge. Similarly, Oliveira et al [15] developed a holistic stakeholder agnostic catalog framework for biomedical datasets. Researchers can explore metadata held decentralized at federated nodes, with distinct levels of granularity being conceivable. Extending this initial design knowledge specific to biomedical data, Swertz et al [5] proposed a unified framework for sharing health data across catalogs. It encompasses multiple centralized and decentralized catalogs. The authors offer recommendations to establish an integrated community as an open catalog ecosystem. This theoretical basis for HMDCs builds upon and is enriched by similar research. Specifically, Bergeron et al [17] developed a catalog toolkit to support creating comprehensive as well as user- and study-friendly HMDCs. Almeida and Oliveira [30] produced a framework to simplify the process of building an HMDC for exposing metadata, while providing analysis capacities. Apparently, there is a tendency from centralized to decentralized data catalogs in health care. However, for HMDCs, a *research gap* prevails concerning (1) empirically grounded and actionable design knowledge that is (2) universally applicable to (3) the broad array of use cases and EU initiatives associated with health care data ecosystems.

In general, the generation of design knowledge about an artifact is crucial as it provides the intellectual foundation to advance the respective body of scientific knowledge, while facilitating development efforts in practice [31]. In particular, HMDC design knowledge can harmonize and sustain the multitude of different EU initiatives by following a systematic approach to problem-solving [31,32]. Therefore, its generation must adhere to a rigor design process [33,34]. This process must ensure empirical grounding for the sake of efficiency, effectiveness, and quality assurance, which inevitably favors quality, adaptiveness, and impact of the generated results [35]. Likewise, the prevailing lack of design knowledge causes difficulties concerning the adoption and use of HMDCs in practice and theory, revealing the *research problem*. To remedy this problem, the *research objective* is to provide actionable design knowledge that is universally applicable in real-world HMDC use cases, thus allowing to infer the following *research questions* (RQs):

- RQ1:
  What are taxonomy elements (ie, dimensions and characteristics) to structure HMDCs from a design science perspective?
- RQ2:
  How does the proposed taxonomy effect real-world use cases?

According to Hevner et al [33], a *design science* perspective means to examine and create information system (IS) artifacts to solve practical problems. A taxonomy is a suitable approach to address RQ1 because it provides a set of elementary building

XSL•FO
RenderX

blocks and prescriptions for effectively designing such artifacts [36,37]. It targets a broad and diverse audience, including health care IS engineers and architects, health data holders and scientists, health care economists and researchers, as well as legal and ethical regulatory bodies, while accentuating the European health care sector.

Methods

Overview

Taxonomies are common approaches in IS research to classify, understand, and examine complex issues [38]. For their development, the method of Nickerson et al [37] is applied to identify dimensions and characteristics of HMDCs. The authors propose generating knowledge conceptually (eg, from the literature) and empirically (eg, analyzing objects of interest). This approach is referred to as the gold standard to build taxonomies in IS research [36]. As refinement, the methodological update of Kundisch et al [36] is incorporated, adding an evaluation process by means of focus groups. The authors' refinement enhances the assessment of value created by the taxonomy [39]. Corresponding to these 2 methods, the research design is divided into the 7 steps shown in Figure 1 based on the studies by Nickerson et al [37] and Kundisch et al [36]. The numbers 1 to 7 represent methodological steps explained in the following sections.

Figure 1. Applied research method of the taxonomy development study based on Nickerson et al [37] and Kundisch et al [36].

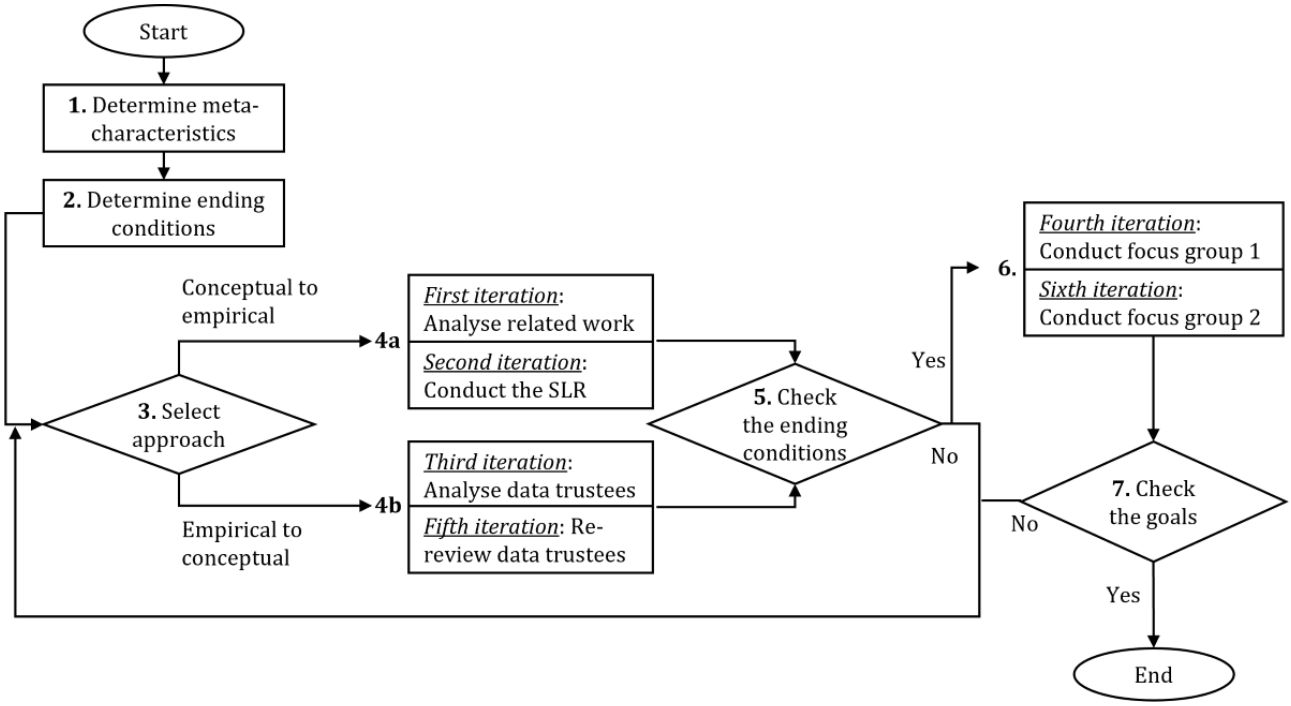

In step 1 in Figure 1, a *meta-characteristic* is specified in orientation towards the taxonomy's purpose so that each subordinated characteristic and dimension follows from it. On the basis of RQ1, the meta-characteristic was defined as "distinguishing key design elements of HMDCs." It facilitates selecting meta-dimensions as well as inferring characteristics and classifying them to dimensions. To define the meta-dimensions, the FAIR framework was used [7]. These well-known data principles postulate an accepted approach to the discoverability and usability of RWD [19]. While FAIR emphasizes making data interoperable and reusable, it inherently involves considerations related to data governance and harmonization [40].

In step 2 in Figure 1, *ending conditions* for the iterative part of the process are defined, determining its termination criteria. The ending conditions were chosen on the basis of Nickerson et al [37] and Scheider et al [41] in terms of subjective and objective criteria. Ultimately, 6 design iterations were required until all conditions listed in Table 1 were fulfilled.

XSL•FO
RenderX

Table 1. Ending conditions for taxonomy development adopted from Nickerson et al [37].

Ending conditions	Design iterations					
	1	2	3	4	5	6
Objective						
All papers were examined.			✓	✓	✓	✓
No object was merged with another or split.			✓		✓	✓
Each characteristic is classified by one object.	✓	✓	✓	✓	✓	✓
No new dimensions or characteristics were added.			✓		✓	✓
Dimensions or characteristics were neither merged nor split.			✓		✓	✓
Each dimension is unique and not duplicated.	✓	✓	✓	✓	✓	✓
Every characteristic is unique within its dimension.	✓	✓	✓	✓	✓	✓
Each cell is unique and not repeated.	✓	✓	✓	✓	✓	✓
Subjective						
Conciseness: no unnecessary dimensions and characteristics			✓	✓	✓	✓
Robust: dimensions and characteristics differentiate objects			✓	✓	✓	✓
Comprehensiveness: all objects can be classified			✓	✓	✓	✓
Extension: dimensions and characteristics can be added easily		✓	✓	✓	✓	✓
Explanatory: dimensions and characteristics describe all objects			✓		✓	✓

In steps 3 to 5 in Figure 1, we repeatedly chose between either an inductive or a deductive path. The former is a conceptual-to-empirical attempt (step 4a in Figure 1) to infer dimensions and characteristics from theory. The latter reflects an empirical-to-conceptual procedure (E2C; step 4b in Figure 1) to derive characteristics from real-world analysis objects and to classify them in dimensions. After each iteration (ie, steps 3 to 5 in Figure 1), the ending conditions are checked (ie, step 5 in Figure 1). If all ending conditions are fulfilled, an evaluation step (ie, step 6 in Figure 1) follows, integrated by focus groups [36]. In case the focus group iteration does not imply changes to the taxonomy (ie, step 7 in Figure 1), the artifact is finished and the methodological process terminates. After 5 design iterations (ie, 4 times of executing steps 4a and 4b and executing step 6 once in Figure 1), all ending conditions were fulfilled (ie, step 5 in Figure 1) and the subsequent focus group did not result in any major changes (ie, step 7 in Figure 1). Thus, the taxonomy was completed [36]. Because 6 design iterations were traversed and it was ensured that the focus group experts covered all dimensions relevant for HMDCs, the taxonomy achieved result saturation.

To ensure transparency of the taxonomy development process, Multimedia Appendix 2 shows intermediary taxonomies after certain iterations. Furthermore, it offers a table linking key references from the inductive (ie, the literature) and deductive iterations (ie, analysis objects) to the dimensions of the final taxonomy [5,10,14,16,17,19,24,28,30,41-78].

Inductive Design Iterations for Taxonomy Development

In the first iteration, an initial set of dimensions and characteristics was derived from former research (ie, step 4a in Figure 1), consolidating the related work addressed in the Introduction section.

In the second iteration, a structured literature review (SLR) was carried out (ie, step 4a in Figure 1) [36]. The method of Kitchenham et al [79] was applied (ie, 1-6), while orienting toward its application in the study by Scheider et al [41]. At the outset, RQ1 was adopted as the (1) research question guiding the SLR. The (2) search process comprised HMDC-related conference and journal papers. The search string was defined as (ALL (health AND data AND catalog) OR ALL (health AND metadata AND catalog) AND ALL (data AND catalog AND technologies)). Primarily, Scopus and IEEE Xplore were used, and the operands were deployed on documents' titles, abstracts, and authors' keywords. The 2 databanks were leveraged due to their multidisciplinary nature covering research in all fields relevant for HMDCs. Following Scheider et al [41], (3) inclusion and exclusion criteria were created to identify and filter papers. First, the literature not available in English was excluded. Second, inaccessible papers were removed. Third, each paper retrieved was reviewed by 2 researchers for whether it covers HMDCs in the broader sense. This means that papers had to address a design perspective, as defined by Hevner et al [33]. Articles were emphasized that dealt with "patient-related" data, while ones about aggregated health data (eg, regions and countries) were neglected. The same holds true for catalogs about health-oriented surveys and analysis results (eg, studies). Due to broadly formulated keywords in the search string, initially retrieved literature contained many papers outside the thematical scope. To this end, the third inclusion or exclusion criterion was examined by screening titles and abstracts before reviewing the entire content of the papers. Since 2 researchers constantly worked together in (3), one can argue for reliable objectivity in paper selection.

Building upon the inclusion and exclusion criteria, the initial (4) data collection resulted in 18 papers in the Scopus and IEEE

XSL•FO
RenderX

Xplore search (iteration 4 in Figure 2). Subsequently, backward (ie, referenced articles) and forward (ie, citing articles) stepping was conducted [80], which added 12 articles. Figure 2 shows the SLR statistics expressed by a PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart.

Figure 2. Structured literature review statistics presented in the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart.

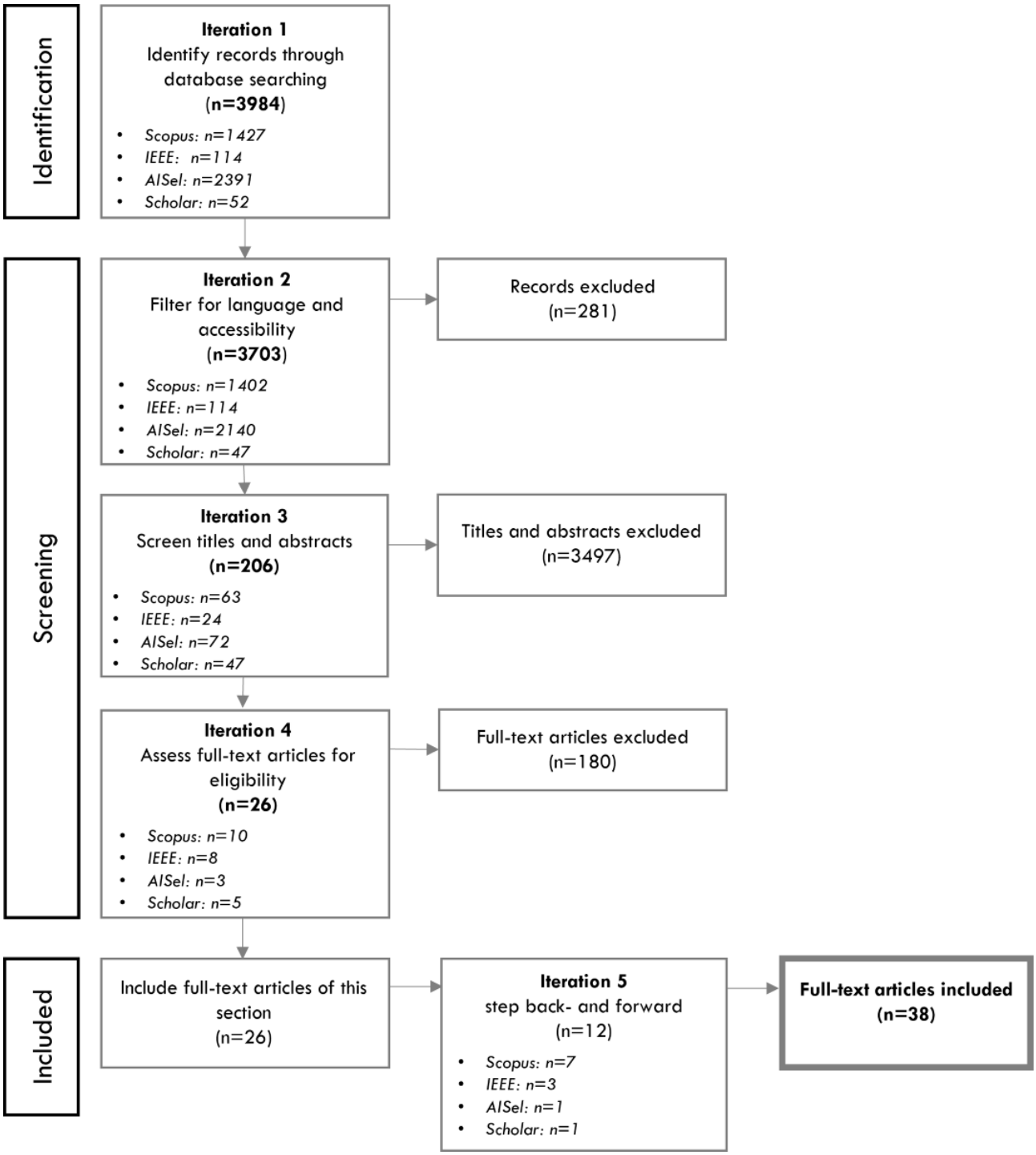

The SLR was expanded by a search via Google (Google Search) and the AISeL database for extension and verification. The Google search engine served to identify whitepapers using a consolidated search string compared to (2). For AISeL, the same steps executed in Scopus and IEEE Xplore were applied (ie, 2 and 3), except that the research team looked for the search terms in titles only to keep the number of results feasible. Once duplicates were removed, Google and AISeL added 8 papers to the literature collection. To test theoretical saturation [81], "quick searches" were carried out in other databases (eg, ACM) checking whether the top results, first, match the inclusion and exclusion criteria and, second, are not already in the collection. Since these quick searches did not add new papers, the literature collection was considered representative [41]. Excluding related work of the first iteration, the collection comprised 38

XSL•FO
RenderX

publications, of which the most important items are listed in Table 2.

Table 2. List of most important literature from the SLR[a] used for taxonomy development.

Study	Year	Title
Top 10 academic papers		
Alvarellos et al [42]	2023	Democratizing clinical-genomic data: How federated platforms can promote benefits sharing in genomics
Almeida and Oliveira [30]	2024	MONTRA2[b]: A web platform for profiling distributed databases in the health domain
Almeida et al [19]	2023	A FAIR[c] approach to real-world health data management and analysis
Bergeron et al [17]	2018	Fostering population-based cohort data discovery: The Maelstrom Research cataloguing toolkit
Scheider et al [41]	2023	Exploring design elements of personal data markets
Ehrlinger et al [26]	2021	Data catalogs: a systematic literature review and guidelines to implementation
Labadie et al [24]	2020	Fair enough? enhancing the usage of enterprise data with data catalogs
Lovestone and EMIF Consortium [10]	2020	The European medical information framework: a novel ecosystem for sharing health care data across Europe
Oliveira et al [15]	2019	EMIF[d] Catalogue: a collaborative platform for sharing and reusing biomedical data
Swertz et al [5]	2022	Towards an Interoperable Ecosystem of Research Cohort and Real-world Data Catalogues Enabling Multi-center Studies
Top 5 nonacademic papers		
European Medicines Agency [43]	2022	Good Practice Guide for the use of the Metadata Catalogue of Real-World Sources
European Medicines Agency [44]	2022	List of metadata for Real World Data catalogues
Directorate-General for Health and Food Safety [82]	2022	The European Health Data Space
Jahnke and Otto [25]	2022	Data Catalogs - Implementing Capabilities for Data Curation, Data Enablement and Regulatory Compliance
TEHDAS[e] [16]	2022	EHDS[f] Semantic interoperability framework

[a]SLR: structured literature review.

[b]MONTRA2: Modular Next-generation Research Analysis.

[c]FAIR: findability, accessibility, interoperability, and reusability.

[d]EMIF: European Medical Information Framework Catalogue.

[e]TEHDAS: Towards European Health Data Space.

[f]EHDS: European Health Data Space.

Throughout the steps (2) to (4), a (5) quality assessment step was integrated on the basis of the criteria suggested by Kitchenham et al [79], that is, inclusion or exclusion criteria, relevant article coverage, literature corpus assessment, and study descriptions.

For (6) data analysis, phrases ("quotes") from articles with useful content for HMDC designs were extracted. Following the approaches of Saldana [83] and Pratt [84], those phrases were coded, inserted in a tabular structure, and iteratively generalized. As in steps (3) and (9), two researchers analyzed the literature to reduce subjectivity biases. Figure 3 shows how quote extractions relating to the dimension of data linking are coded and design implications are derived. Particularly, whenever there was a direct connection to an HMDC context, quotes became design implications immediately, for example, linkage strategy (first quote in Figure 3). If a direct connection was missing (eg, linkage variable as a new characteristic; third quote in Figure 3), more evidence was required to transform codes into design implications (ie, second quote in Figure 3). Finally, considering the influential factors proposed by Mwita [81] (eg, study purpose, research design, sample size variability, and analysis approach), data saturation in the SLR is likely.

XSL•FO
RenderX

Figure 3. Examples for literature coding for inductive taxonomy development.

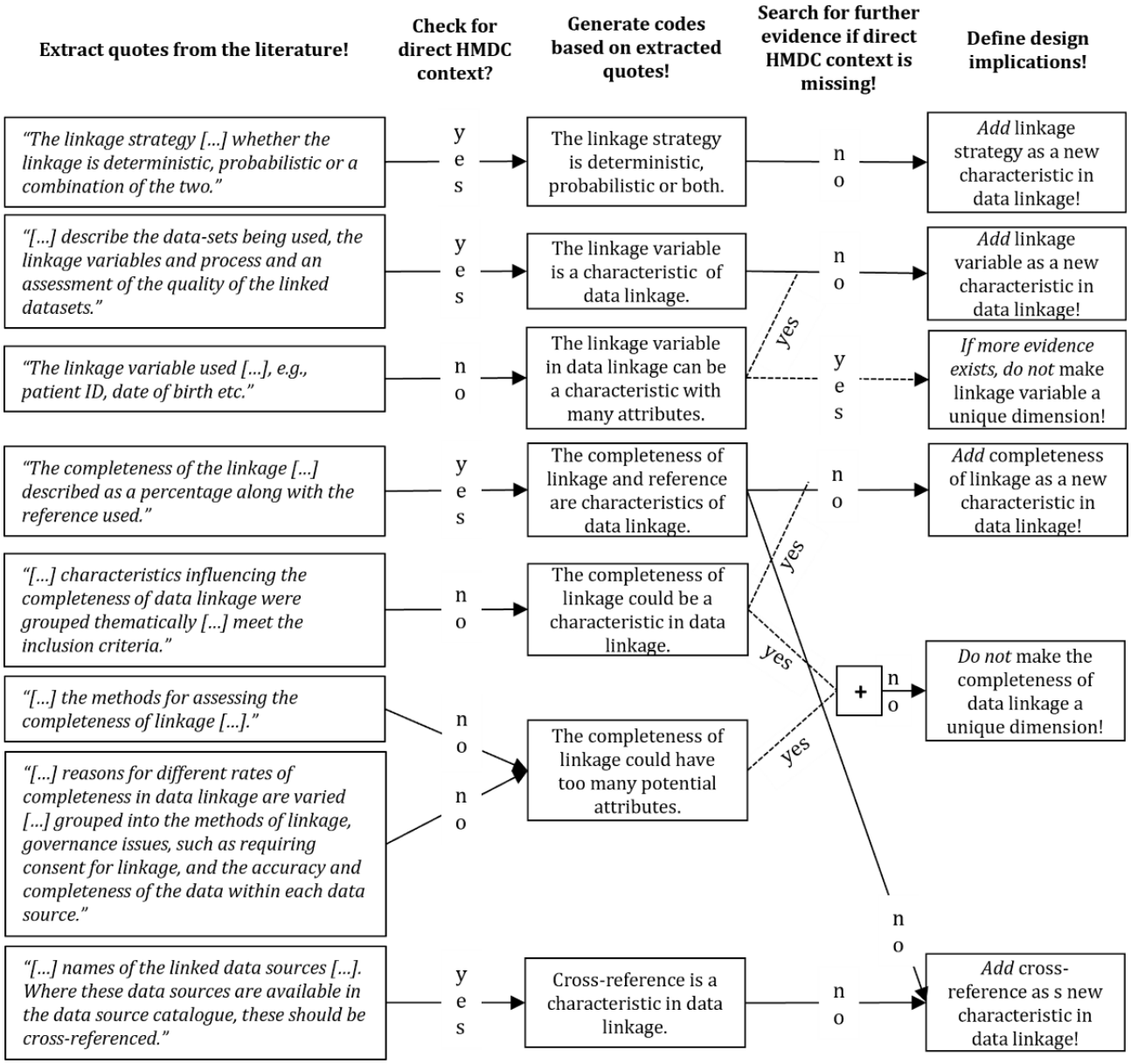

Deductive Design Iterations for Taxonomy Development

Applying the E2C approach (ie, step 4b Figure 1) in the third and fifth iterations, health data catalogs from practice were initially listed, as identified in the first 2 iterations. This set was extended by a Google search to identify analysis objects not encountered in inductive iterations. The research team searched for analysis objects using the browser's incognito mode to circumnavigate carryover effects from previous searches [85]. The keywords from the SLR were used as an orientation to avoid limiting the results unconsciously [41]. Analysis objects were excluded if meaningful information could not be obtained. This characteristic was defined as access to analyzable information describing the analysis object, that is, data retrievable either from websites or demo applications [85]. Analysis objects were also excluded if information was meaningful but unavailable in German or English. However, metadata catalogs under construction were not excluded per se [41]. The set of analysis objects was created in the first quarter of 2024. The final analysis objects are listed in Table 3.

XSL•FO
RenderX

Table 3. Health metadata catalogs from practice used for taxonomy development.

HMDCs[a]	Classification	Status
BBMRI-ERIC[b] Data Directory [45]	Decentral	Operative
Catalogue of Mental Health Measures [46]	Central	Operative
Compendium Data Catalog for Healthcare [47]	Central	Operative
EHDEN[c] Portal [48]	Decentral	Operative
Elixir (BioSamples [86] and FAIRsharing [87])	Decentral	Operative
EMIF[d] Data Catalogue [49]	Decentral	In progress
EUCAIM[e] Cancer Image Europe [50]	Decentral	Operative
European Health Information Portal [51]	Central	Operative
Fjelltopp Data Catalogues for Health [52]	Central	Operative
HealthRI[f] Data Catalogues [53]	Decentral	Operative
Helsedata Explore Data Sources [54]	Decentral	Operative
IDERHA[g] Metadata Catalogue (no public access)	Decentral	In progress
IHME[h] Global Health Data Exchange [88]	Central	Operative
IQVIA[i] Health Data Catalogue [55]	Central	Operative
Kraken Health Data Pilot [56]	Decentral	In progress
Lifebit Precision Medicine Data Catalogue [57]	Decentral	Operative
MACH[j] Clinical and Research Data Catalogue [89]	Decentral	Operative
Maelstrom Research Data Catalogue [58]	Decentral	Operative
Yoda Trials Data Catalogue [59]	Decentral	In progress

[a]HMDC: health metadata catalog.
[b]BBMRI-ERIC: Biobanking and Biomolecular Resources Research Infrastructure – European Research Infrastructure Consortium.
[c]EHDEN: European Health Data and Evidence Network.
[d]EMIF: European Medical Information Framework.
[e]EUCAIM: European Federation for Cancer Images.
[f]HealthRI: Health Research Infrastructure.
[g]IDERHA: Integration of Heterogeneous Data and Evidence towards Regulatory and Health Technology Assessment Acceptance.
[h]IHME: Institute for Health Metrics and Evaluation.
[i]IQVIA: Information, Quality, Value, Innovation, and Access.
[j]MACH: Melbourne Academic Centre for Health.

The catalogs were examined by classifying them alongside the design elements of the preliminary taxonomy (Multimedia Appendix 2). We tried to assign a catalog to a single characteristic in each dimension. This deductive activity was conducted by 2 researchers, whereby three cases could occur [41]: (1) on the basis of available information, the analysis object could be assigned to an existing characteristic in a dimension; (2) due to a lack of available information, the analysis object could not be assigned to any characteristic in a dimension; and (3) the analysis object contained information about a dimension but could not be associated with a characteristic defined therein. Evidently, the third case led to modifications of a current set of taxonomy design elements. During the third iteration, the occurrence of this case decreased continuously until only the first and second ones appeared. Because the ending conditions were fulfilled at that time, the taxonomy-building process was terminated and the first focus group session started. Alterations were proposed by focus group members in this session. This led to another E2C approach as the fifth iteration in which specific design elements were emphasized, as addressed by the focus group experts. The corresponding analysis procedure (ie, 1-3) remained the same as in the third iteration, while examining decentral analysis objects only (Table 3).

Evaluative Design Iterations for Taxonomy Development and Use Case Cocreation

Following Figure 1, an evaluation by focus groups is needed after the fulfillment of all ending conditions (ie, in step 6). Focus groups help gather more data than individual interviews, since experts respond to the input of others, triggering discussions and idea generation [23]. According to Szopinski et al [39], focus groups are particularly suitable to assess the comprehensiveness, robustness, understandability, and

XSL•FO
RenderX

extensibility of a taxonomy, as well as the shape of dimensions and characteristics. Members were recruited on the basis of the target audience of the taxonomy (see section about Research gap, objective, and questions). We ensured that they have substantial knowledge with regard to HMDCs, stemming from EU initiatives (Multimedia Appendix 1). Ultimately, the focus group consisted of 7 experts. However, scheduling conflicts affected the organization of meetings. The first focus group had to be split into 2 sessions, involving 5 experts. The second focus group iteration required 3 separate sessions to accommodate all 7 experts. Multimedia Appendix 3 shows the distribution of experts to sessions per iteration. Focus group sessions occurred in the fourth and sixth iteration. Their rough agenda was as follows: (1) the objective of the taxonomy (only fourth iteration) and current version (fourth and sixth iteration) were presented; (2) a dimension with its characteristics was explained; (3) experts shared their expectations about this dimension, especially with regard to real-world use cases; and (4) we triggered discussions by contrasting expectations. Because the focus groups served for evaluation purposes and concrete design elements were discussed in the plenum directly, we dispensed with detailed coding. This lightweight effort is inspired and justified by similar research studies published in high-ranking IS journals [14,41]. Therefore, we argue that our approach successfully mitigates any subjectivity biases because specific taxonomy design elements were addressed and discussed in the focus groups directly. This narrow focus led to the derivation of tangible design activities (ie, deletion, addition, alteration, and renaming of design elements and cocreation of use cases).

In the fourth iteration, the focus group resulted in substantial changes of dimensions and characteristics, particularly regarding taxonomy elements of data accessibility and findability (Multimedia Appendix 2). On the contrary, the second session in the sixth iteration caused minor adjustments only (eg, renaming a few characteristics), emphasizing the cocreation of use cases. Finally, the third session of the sixth iteration merely led to refined use case formulations and merging of content. Given these small adjustments, a sufficient saturation in results was considered [23] and the taxonomy development was terminated (ie, step 7). Multimedia Appendix 3 lists the focus group members who participated in the evaluation of the taxonomy and the cocreation of use cases during the fourth and sixth iterations. It further states the EU initiatives the experts were recruited from and whether these initiatives have an operative HMDC or one that is still under development.

Ethical Considerations

In the run-up to the sessions, the experts received an information sheet explaining the study context, the procedure, and the approach to gathering and processing interview data. It was assured that no personal information will be disclosed. Experts were informed about their right to opt out. During the focus group sessions, only anonymous interview data in the form of handwritten notes were collected. Neither a video or voice recording nor the transcription of interview material was used. Experts did not receive any financial compensation. Hence, conducting focus groups did not require the official approval of an ethics review committee.

Results

Taxonomy for FAIR HMDCs

Overview

Table 4 shows the taxonomy containing 20 dimensions (D_n) and 101 characteristics (C_{nm}) structured alongside the FAIR data principles as meta-dimensions [40]. For visualization, morphologies were considered (Multimedia Appendix 2) as they demonstrate the structure and arrangement of taxonomy elements [90]. Importantly, the meta-dimension of data accessibility has exclusive dimensions pertaining to the general design of HMDCs. The other meta-dimensions relate to design elements of metadata assets published *within* the HMDC and the data accessibility constraints. On the basis of the study by Nickerson et al [37], nonexclusiveness was chosen for some of the dimensions associated with metadata assets. The reason is that they propose structural characteristics, which might be accumulated for creating effective metadata assets in an HMDC. Exclusivity of dimensions allows to categorize design elements into mutually distinct characteristics, ensuring clarity and avoiding overlaps. Nonexclusivity can accommodate complex multidimensional relationships by enabling design elements belonging to multiple characteristics [37]. Correspondingly, the difference between OR and XOR in the middle column of Table 4 is that OR is used for nonexclusive characteristics, while XOR is used for exclusive ones.

XSL•FO
RenderX

Table 4. Taxonomy of FAIR[a] HMDCs[b] presented as a morphological box.

Dimension (D_n)	Characteristics (C_{nm})	Exclusive or nonexclusive
Findability		
D_1: data source	$C_{1.1}$: patients' health OR $C_{1.2}$: medical procedures OR $C_{1.3}$: medical products OR $C_{1.4}$: others	Nonexclusive
D_2: managerial. details	$C_{2.1}$: holder OR $C_{2.2}$: origin OR $C_{2.3}$: collection OR $C_{2.4}$: qualification OR $C_{2.5}$: financials OR $C_{2.6}$: others	Nonexclusive
D_3: data type	$C_{3.1}$: admin XOR $C_{3.2}$: primary care XOR $C_{3.3}$: secondary care XOR $C_{3.4}$: registries XOR $C_{3.5}$: others	Exclusive
D_4: population information	$C_{4.1}$: disease OR $C_{4.2}$: family linkages OR $C_{4.3}$: lifestyle factors OR $C_{4.4}$: population OR $C_{4.5}$: sociodemographic OR $C_{4.6}$: catchment area coverage	Nonexclusive
D_5: data sensitivity	$C_{5.1}$: synthetic data XOR $C_{5.2}$: anonymized data XOR $C_{5.3}$: pseudonymized data XOR $C_{5.4}$: personal data	Exclusive
Accessibility		
D_6: catalogue accessibility	$C_{6.1}$: public XOR $C_{6.2}$: hybrid XOR $C_{6.3}$: private	Exclusive
D_7: dataset accessibility	$C_{7.1}$: free XOR $C_{7.2}$: formal request XOR $C_{7.3}$: strictly limited XOR $C_{7.4}$: others	Exclusive
D_8: access control	$C_{8.1}$: catalog operator OR $C_{8.2}$: internal DAC[a] OR $C_{8.3}$: external DAC OR $C_{8.4}$: none OR $C_{8.5}$: others	Nonexclusive
Interoperability		
D_9: program discoverability	$C_{9.1}$: Beacon OR $C_{9.2}$: BBMRI-MIABIS[d] OR $C_{9.3}$: bioimage OR $C_{9.4}$: CESSDA[e] OR $C_{9.5}$: DCAT[f] OR $C_{9.6}$: ECRIN-CRMDR[g] OR $C_{9.7}$: FairShairing OR $C_{9.8}$: INSPIRE[h] OR $C_{9.9}$: PHIRI[i] OR $C_{9.10}$: others	Nonexclusive
D_{10}: semantic interoperability	$C_{10.1}$: CDISC-SDTM[j] OR $C_{10.2}$: LOINC[k] OR $C_{10.3}$: OMOP[l] OR $C_{10.4}$: Oorphanet standards OR $C_{10.5}$: SNOMED[m] OR $C_{10.6}$: others	Nonexclusive
D_{11}: interoperable communication	$C_{11.1}$: DICOM[n] OR $C_{11.2}$: HL7 FHIR[o] OR $C_{11.3}$: IDMP[p] OR $C_{11.4}$: ISO 800-110[q] OR $C_{11.5}$: others	Nonexclusive
D_{12}: CDM[r]	$C_{12.1}$: type OR $C_{12.2}$: reference OR $C_{12.3}$: release frequency	Nonexclusive
D_{13}: ETL[s] status	$C_{13.1}$: planned XOR $C_{13.2}$: in progress XOR $C_{13.3}$: completed	Exclusive
D_{14}: vocabularies	$C_{14.1}$: medicinal product OR $C_{14.2}$: cause of death OR $C_{14.3}$: quality of life measurement OR $C_{14.4}$: prescription OR $C_{14.5}$: dispensing OR $C_{14.6}$: indication OR $C_{14.7}$: procedures OR $C_{14.8}$: genetic data OR $C_{14.9}$: biomarker data OR $C_{14.10}$: medical event	Nonexclusive
Reusability		
D_{15}: collection methodology	$C_{15.1}$: collection governance OR $C_{15.2}$: collection process OR $C_{15.3}$: dataset updates OR $C_{15.4}$: others	Nonexclusive
D_{16}: collection events	$C_{16.1}$: patient encounter OR $C_{16.2}$: physical examination OR $C_{16.3}$: diagnostics OR $C_{16.4}$: treatment OR $C_{16.5}$: progress note OR $C_{16.6}$: communication OR $C_{16.7}$: regulatory OR $C_{16.8}$: others	Nonexclusive
D_{17}: data linkage	$C_{17.1}$: strategy OR $C_{17.2}$: variable OR $C_{17.3}$: completeness OR $C_{17.4}$: cross-reference OR $C_{17.5}$: none	Nonexclusive
D_{18}: data preservation	$C_{18.1}$: definite records XOR $C_{18.2}$: indefinite records	Exclusive
D_{19}: publish	$C_{19.1}$: approval needed XOR $C_{19.2}$: no approval needed	Exclusive
D_{20}: informed consent	$C_{20.1}$: not required XOR $C_{20.2}$: general use XOR $C_{20.3}$: all studies XOR $C_{20.4}$: specific studies XOR $C_{20.5}$: waiver XOR $C_{20.6}$: others	Exclusive

XSL•FO
RenderX

[a]FAIR: Findability, Accessibility, Interoperability, and Reusability.
[b]HMDC: health metadata catalog.
[c]DAC: Data Access Committee.
[d]BBMRI-MIABIS: Biobanking and Biomolecular Resources Research Infrastructure-Minimum Information About Biobank Data Sharing.
[e]CESSDA: Consortium of European Social Science Data Archives.
[f]DCAT: Data Catalog vocabulary.
[g]ECRIN-CRMDR: European Clinical Research Infrastructure Network – Clinical Research Metadata Repository.
[h]INSPIRE: Infrastructure for Spatial Information in Europe.
[i]PHIRI: Population Health Research Infrastructure.
[j]CDISC-SDTM: Clinical Data Interchange Standards Consortium - Study Data Tabulation Model.
[k]LOINC: Logical Observation Identifiers Names and Codes.
[l]OMOP: Observational Medical Outcomes Partnership.
[m]SNOMED: Systematized Nomenclature of Medicine.
[n]DICOM: Digital Imaging and Communications in Medicine.
[o]HL7 FHIR: Health Level 7 Fast Healthcare Interoperability Resources.
[p]IDMP: Identification for Medicinal Products.
[q]ISO800-110: International Organization for Standardization 800-110.
[r]CDM: common data model.
[s]ETL: extract, transform, load.

Data Findability

The meta-dimension prescribes that datasets orchestrated by an HMDC must be easily discoverable, requiring metadata assets to describe essential attributes of the decentral datasets [40]. The dimension *data source* (D_1) refers to abstract categories for data classification in the catalog system. Following European Medicines Agency guidelines [43], and implementations in practice [49,50], patients' health ($C_{1.1}$) comprises datasets attributable to conditions. Examples are diseases, causes of death, prescriptions and dispensing of medicines, clinical measurements, genetic data, units of health care use, and all other similar patient-generated data, for example, wearables. Medical procedures ($C_{1.2}$) encompass data describing hospital admission discharges, intensive care admissions, administration of vaccines or other injectables, medical operations, biomarker data, and diagnostic codes [43,44]. The latter includes, among others, the International Classification of Disease Code, the Major Comorbidity Code, and the Major Diagnostic Code. Medical products ($C_{1.3}$) span categories like prescribed medicinal products for human use, contraception, indication for use, and medical device data. Others ($C_{1.4}$) may refer to further data, for example, health care providers delivering diagnosis and treatment services [43]. The dimension *managerial details* (D_2) captures crucial organizational metadata to be disclosed by the HMDC as part of a dataset's metadata asset [24]. Above all, the data holder ($C_{2.1}$) must be publicized, including contact details (ie, data steward). This entity sustains the record collection in an underlying dataset [60]. The origin ($C_{2.2}$) of the data refers to the countries or geographical regions of their acquisition [60] and the language [44]. The characteristic of collection ($C_{2.3}$) details the acquisition dates as well as all data assemblage information, except collection methodology and events. If the dataset has received a formal qualification ($C_{2.4}$), this should also be disclosed in the metadata asset [43]. The same holds for sources of finance ($C_{2.5}$) having sponsored the dataset creation [23,61], for example, data holder, public, industry, research, or patient organizations [44]. Naturally, other ($C_{2.6}$) metadata attributes specifying managerial details are conceivable. The dimension *data type* (D_3) describes broader content related categories applicable to the dataset. It mainly distinguishes datasets containing administrative ($C_{3.1}$ [62]), primary ($C_{3.2}$) and secondary care ($C_{3.3}$ [63]), registry ($C_{3.4}$), and other ($C_{3.5}$) data types. Furthermore, *population information* (D_4) as metadata attribute refers to the specifics of the records within a dataset [17]. The taxonomy narrows the dimension down to collected disease information ($C_{4.1}$); particular population specifics ($C_{4.2}$; eg, age groups); family linkages ($C_{4.3}$; eg, household, parent-child, sibling, and not applicable) lifestyle factors ($C_{4.4}$; eg, tobacco use, physical exercises, and diet); sociodemographic data ($C_{4.5}$; eg, gender, ethnicity, education, and deprivation index); and ($C_{4.6}$) catchment area coverage [17,43,44]. *Data sensitivity* (D_5) addresses the identifiability of data subjects of whom records were collected. Records can be synthetic ($C_{5.1}$), anonymized ($C_{5.2}$), pseudonymized ($C_{5.3}$), or contain personal data ($C_{5.4}$) [15,41].

Data Accessibility

Once datasets are findable in the HMDC, their accessibility must be facilitated. By their design, HMDCs must ensure legal and ethical compliance of all data access and sharing processes in their health care ecosystems [19]. This particularly involves that datasets can be retrieved by authorized users only, implying rigor access control functions [5,41]. Accordingly, the dimension *catalogue accessibility* (D_6) refers to access modalities for data consumers to use the HMDC. HMDCs can be public ($C_{6.1}$) allowing anyone to browse metadata assets and discover datasets. Alternatively, HMDCs can be private ($C_{6.2}$), limited to a certain number of users who have been formally authorized by a dedicated authority. In addition, hybrid ($C_{6.3}$) forms exist. *Dataset accessibility* (D_7) describes the access modalities for data consumers to published datasets [30]. The access can be free ($C_{7.1}$), implying that data must at least be anonymized or

XSL•FO
RenderX

synthetic to comply with legal and ethical guidelines [41,64]. Datasets can also require a formal request ($C_{7.2}$) with a Data Access Committee (DAC) or a data steward deciding upon access requests [65,66]. HMDCs frequently require such requests of data consumers to be approved by their ethic committees before processing them within the ecosystem [66]. Moreover, data access can be strictly limited ($C_{7.3}$) to a demarcated group of data consumers. Although, members of such limited groups also need to make formal requests for data access [66]. This implies that other ($C_{7.4}$) data access modalities exist, especially combinations of $C_{7.1}$ to $C_{7.3}$. *Access control* (D_8) refers to the mechanisms implemented by HMDCs that facilitate the aforementioned decision-making by empowered entities [30]. The dimension specifies the entities who determine whether data consumers receive the requested datasets and are allowed to perform which kinds of processing activities [67]. It distinguishes the catalog operator ($C_{8.1}$); internal DACs at the sides of the data holders ($C_{8.2}$ [65,66]); external DACs ($C_{8.3}$) that are run centrally by an independent third party [23,66]; the absence of access control ($C_{8.4}$; ie, free data [$C_{7.1}$]); and any other forms ($C_{8.5}$). Generally, access control in HMDC designs is crucial for maintaining data security, confidentiality, and compliance with legal and ethical constraints [67].

Data Interoperability

A core objective of HMDCs is to enable data consumers accumulating datasets across organizations, effectively, to create meaningful connections and analyses [7]. This makes data interoperability crucial [91]. To that end, HMDCs leverage specific standards described in this meta-dimension. Thereof, *programmatic discoverability* (D_9) refers to the ability of data consumers to programmatically query, access, and retrieve metadata assets and search for their attributes. It is defined by the joint action Towards European Health Data Space (TEHDAS) [16] as the ability to identify, access, and understand health data by automated means. Associated approaches commonly involve application programming interfaces or similar programmatic methods to access and filter metadata in the HMDC [19]. Following the TEHDAS community [16], the dimension is narrowed down to the most frequent standards used by HMDCs. These are:

- Beacon ($C_{9.1}$),
- Biobanking and biomolecular resources research infrastructure-minimum information about biobank data sharing (BBMRI-MIABIS; $C_{9.2}$),
- Bio-image archive ($C_{9.3}$),
- Consortium of European Social Science Data Archives (CESSDA; C9.4),
- Data catalog vocabulary (DCAT; $C_{9.5}$),
- European Clinical Research Infrastructure Network – clinical research metadata repository (ECRIN-CRMDR; $C_{9.6}$),
- FairSharing ($C_{9.7}$),
- Infrastructure for Spatial Information in Europe (INSPIRE; C9.8),
- Population Health Research Infrastructure (PHIRI; C9.9), and
- Others ($C_{9.10}$).

HMDCs typically adhere to one of those standards to ensure programmatic discoverability of published data offerings (ie, the metadata assets). The dimension *semantic interoperability* (D_{10}) ensures that the precise format and meaning of datasets is preserved and understood, covering both semantic and syntactic aspects [68]. Similar to D_9, the characteristics of this dimension encompass standards commonly applied by HMDCs. These are

- Clinical Data Interchange Standards Consortium - Study Data Tabulation Model (CDISC-SDTM; C10.),
- Logical Observation Identifiers Names and Codes (LOINC; C10.2),
- Observational Medical Outcomes Partnership (OMOP; C10.3),
- Orphanet (C10.4),
- Systematized Nomenclature of Medicine (SNOMED; C10.5 [16,68,69] and
- Others (C10.6).

Subsequently, the dimension *interoperable communication* (D_{11}) comprises approaches implemented by HMDCs to facilitate seamless and effective data sharing between data holders and consumers [70]. Approaches typically used for interoperable communication are Digital Imaging and Communications in Medicine (DICOM; $C_{11.1}$), Health Level 7 Fast Healthcare Interoperability Resources (HL7 FHIR; $C_{11.2}$), Identification for Medicinal Products (IDMP; $C_{11.3}$), and International Organization for Standardization (ISO) 800-110 ($C_{11.4}$) [16,71]. As for the previous dimensions, other standards ($C_{11.5}$) are conceivable. For D_9 to D_{11}, detailed information is easily available in the web.

The following dimensions deal with "data harmonization" as a crucial aspect of data interoperability. They comprise HMDC design elements for standardizing disparate datasets. The purpose is to ensure consistency and coherence of all datasets classifiable to the same data type (D_3). Data harmonization aims to create a unified and cohesive view on datasets, enhancing their allocation, sharing, and use [7]. The common data model (CDM; D_{12}) describes the specifications relating to the structured representation of data records within datasets [5]. The CDM unfolds implications to the relationships between these records, as well as the rules and possibilities for data use. It defines how data consumers can access and process datasets, thus providing the foundation for data consistency, interoperability, and orchestration [10,72]. The taxonomy distinguishes the CDM type ($C_{12.1}$) [10,43,72], the CDM references ($C_{12.2}$), for example, websites or publications, and the release frequency of CDM specification updates ($C_{12.3}$) [43]. Furthermore, information about datasets on their transformation status (extract, transform, load [ETL]) to a CDM should be provided [44]. This *ETL status* (D_{13}) can be described as planned ($C_{13.1}$), in progress ($C_{13.2}$), or completed ($C_{13.3}$), indicating the readiness of the dataset for use. Finally, HMDCs leverage *vocabularies* (D_{14}) as sets of fixed terms, labels, or identifiers to describe and categorize the metadata assets [63]. Vocabularies facilitate understanding,

XSL•FO
RenderX

discovery, and allocation of metadata with a consistently applied language [63,73]. The dimension distinguishes 10 characteristics for classifying vocabularies on the basis of pertinent literature: medicinal product ($C_{14.1}$), cause of death ($C_{14.2}$), quality of life measuring ($C_{14.3}$), prescription ($C_{14.4}$), dispensing ($C_{14.5}$), indication ($C_{14.6}$), procedures ($C_{14.7}$), genetic data ($C_{14.8}$), biomarker data ($C_{14.9}$), and medical event ($C_{14.10}$) [43,44,73].

Data Reusability

FAIR datasets must be created and documented in a way that allows reusage for different purposes. For HMDCs, this implies providing contextual information beyond the metadata dimensions associated with data findability [40]. *Collection methodology* (D_{15}) encompasses characteristics that are associated with how data records were created [74]. Thereof, collection governance ($C_{15.1}$) addresses information about data capture, demonstrating legal and ethical compliance [43]. This includes data quality checks and validation activities [75]. The latter may also refer to the question of whether the dataset allows access to the actual records. Furthermore, the collection process ($C_{15.2}$) outlines how records in the dataset were created [74], for example, surveys, questionnaires, or data retrieval from hospital IS. Dataset updates ($C_{15.3}$) disclose refreshment dates of datasets, for instance, fixed dates around the year [43]. Naturally, the collection methodology can contain other ($C_{15.4}$) use case–specific characteristics as additional metadata attributes. Similar to D_{15}, *collection events* (D_{16}) narrows down the categories of incidents having triggered the creation of a record in the dataset [17]. The dimension comprises the characteristics of patient encounter ($C_{16.1}$; eg, interactions with health care providers); physical examination ($C_{16.2}$; eg, patient's health examined by a professional); diagnostics ($C_{16.3}$; eg, results of medical condition checks); treatment ($C_{16.4}$; eg, documentation of conditions and treatment plans); progress notes ($C_{16.5}$; eg, changes in patients' health status, responses to treatment, or modifications of care plans); communication ($C_{16.6}$; eg, information exchanged by health care providers); regulatory ($C_{16.7}$; eg, legally required documentation of patients' care); and others ($C_{16.8}$) [43,44].

The dimension *data linkage* (D_{17}) describes whether and how a dataset was created by linking others [43,44,76]. The metadata should disclose the linkage strategy ($C_{17.1}$) which could be deterministic, probabilistic, or both. In addition, the used linkage variable ($C_{17.2}$) should be published, along with the completeness of data linkage ($C_{17.3}$). Ideally, the linked datasets should be cross-referred ($C_{17.4}$) and, if applicable, their availability in the HMDC highlighted. In case no data linkage was applied, no corresponding metadata attribute is provided ($C_{17.5}$). Furthermore, *data preservation* (D_{18}) indicates whether records in the dataset are preserved indefinitely ($C_{18.1}$) or, if not ($C_{18.2}$), the time for which they are specified [77]. *Publishing constraints* (D_{19}) provides information to data consumers whether an approval of the data holder ($C_{19.1}$) is needed to publish results obtained from using the dataset or an approval is not needed ($C_{19.2}$) [43]. In the former case, the kind of approval and the approval process should be described. Finally, metadata assets of HMDCs should reveal whether *informed consent* (D_{20}) was obtained or needs to be obtained for data processing [78]. Generally, the characteristics not required ($C_{20.1}$), required for general use ($C_{20.2}$), required for all studies ($C_{20.3}$), required for specific studies ($C_{20.4}$), waiver ($C_{20.5}$), and other ($C_{20.6}$) are recommendable [43].

Cocreated HMDC Use Cases

Overview

The usability, effectiveness, and accuracy of the taxonomy are amplified by 4 "illustrative scenarios" for HMDCs that demonstrate how the FAIR dimensions and characteristics are reflected in real-world use cases [39]. These use cases facilitate the taxonomy's tangibility and the ascertainment of its practical implications, while triangulating the results. As such, they add depth and context to the taxonomy [39]. Originally, 6 abstract application scenarios were derived from recommendations of EMA [43]. Building upon this, we continuously developed and refined those scenarios on the basis of the insights gained during the taxonomy design iterations in general and the focus groups in particular. In the latter, we relied on the experts' reflections on what they expect from an HMDC and whether our dimensions and characteristics meet their expectations, contradict them, or miss out on certain aspects. In the section about the deductive design iterations, we have already described how the focus groups were conducted. We ensured that the experts possess extensive expertise relevant to HMDC designs, either from a development (ie, technology or legal) or a user perspective (ie, data consumer or provider; Multimedia Appendix 3). Hence, the use cases followed a cocreation approach that was contextualized to the methodological process of the study [39]. Moreover, by being refined within multiple design iterations, the cocreated use cases have been triangulated and their relevance for HMDCs ensured. Following, their connection to the taxonomy is highlighted by direct references to dimensions (see *Taxonomy for FAIR HMDCs*). Additionally, I to VII refer to statements of experts who are listed in Multimedia Appendix 3. Reference citations of the literature and analysis objects demonstrate further exemplary sources having contributed to the final versions of the cocreated use cases.

Study Planning

Use case 1 is as follows: *a data consumer wants to identify suitable datasets for a planned study*.

The HMDC must enable data consumers to effectively identify datasets for medical research studies [15] by implementing the following process: First, a data consumer who wants to access the HMDC, needs to be authorized as a qualified user (D_6; #I [50]). Second, this authorized user must be able to browse and filter published metadata assets to discover relevant datasets that fulfill specifications of an intended study (#V and VII [49,50]). For example, detailed data type [63] or population information of the metadata assets should be disclosed to enable verifying the relevance of datasets (D_3 and D_4; #I, II, V, and VII [17,58]). Third, the HMDC must allow to check managerial

XSL•FO
RenderX

details concerning information about the data holder, origin, collection, qualification, and financials (D_2 [43,50]), including the eligibility to receive synthetic, anonymized, pseudonymized, or personal data (D_5; #IV [15]). Subsequently, the data consumer must be facilitated to perform a preliminary assessment of datasets regarding their relevance for the planned study (#III and VII). At this stage, a first list of candidates should be possible to be established. Ideally, the data consumer can access links (ie, cross-references) within the metadata assets to identify former studies which were performed with the same dataset, addressing similar research questions ($C_{17.4}$; #II and VII [43,76]). Such studies are typically accessed outside the contextual boundaries of HMDCs (#IV). Finally, depending on the governance modalities of selected datasets (D_7), the HMDC must enable the data consumer to request data access (#I-VII [30]). Therefore, an official data order needs to be submitted by the HMDC on behalf of the data consumer, containing specifications about the planned study and required documents, for example, protocols and ethical assessments (#I-III). With respect to the datasets accessibility constraints (D_7) and their associated access control (D_8) characteristics, the HMDC must forward data orders to the data holders ($C_{8.2}$ [66]), external third parties ($C_{8.3}$ [23]) or determine request permission or denial decisions itself ($C_{8.1}$ and $C_{8.4}$; #I, III, and VI).

Study Assessment

Use case 2 is as follows: *a dataset is mentioned in a conducted study. The data consumer wants to evaluate this study based on the suitability of the datasets used therein.*

Given that the datasets used in a conducted study are available in the HMDC, data consumers must be enabled to verify, in retrospective, the suitability of these datasets (#I, II, and VII). To support such evaluations, the HMDC must provide different parts of the metadata asset, depending on the nature of the conducted study (#II and VII). For example, to assess the representativeness of the study population, the data consumer needs to examine qualitative metadata attributes (#II and VII), such as population information ($C_{4.1}$-$C_{4.5}$ [17]); data type (D_3 [43,63]); collection methodology (D_{15} [74]); and collection events (D_{16} [17,43]). In addition, quantitative metadata such as the percentage of the population covered in the catchment area ($C_{4.6}$) should be disclosed by the HMDC assets [17,43]. Furthermore, the data consumer might want to explore technical details to evaluate a study and its database, respectively (#I, II, and VII). Examples are the vocabularies used to define variables (D_{14} [43,63]), the CDM according to which the used datasets are structured (D_{12} [43]), the ETL status (D_{13} [44]) and, if applied, any data linkage strategies (D_{17} [76]). Moreover, cross-references should be listed in the metadata assets of the HMDC ($C_{17.4}$) to allow identifying and obtaining lessons learned from other studies, where the same dataset was leveraged (#VII [76]). In doing so, the HMDC facilitates data consumers to identify strengths and limitations of datasets used in conducted studies.

Study Creation and Data Benchmarking

Use case 3 is as follows: *a data consumer writes a study protocol that requires to describe the underlying data sets and compare their characteristics.*

An HMDC must enable data consumers to easily access standardized metadata information about datasets that need to be specified and compared in a study protocol to be written (#I-VII). For HMDCs, this requires making attribute values of metadata assets directly and easily retrievable for data consumers to facilitate an efficient description and comparison of datasets (particularly, D_1-D_5; #VII). As such, when writing a study protocol, the data consumer can simply provide links to the metadata assets available in the HMDC, alongside with all kinds of other information that is interesting in the protocol's context (#III and VII [43]). Providing such links is also beneficial because study readers could, in addition to basic metadata information (D_1-D_5), be interested in collection methodologies (D_{15} [74]) and events (D_{16} [17]), data preservation (D_{18} [77]), consent requirements (D_{20} [43,78]) as well as technical (D_9-D_{14}) and compliance specifics (D_6-D_8). Thus, HMDCs strengthen transparency and reproducibility of studies by facilitating an effective creation of study protocols (#III and V). At the same time, they support data benchmarking by providing detailed standardized metadata, potentially reaching beyond published datasets (#I-VII).

Data Analysis

Use case 4 is as follows: *a data consumer wants to benefit from the experience of others for the creation of a study's programming script or statistical analysis.*

If a study relies on a CDM, the HMDC should enable the data consumer to identify, for published datasets, the ETL procedure (D_{13}) from the dataset to the CDM (D_{12}) [43]. Irrespective of whether data holders have converted their entire datasets, or only an extraction thereof, this information can support the development of the study script (#VII [43]). If the HMDC publishes cross-references of datasets ($C_{17.4}$), the data consumer is also facilitated in finding further studies having investigated the same topic or used a comparable research design (#I, II, and VII [76]). These studies may disclose information on how to operationalize data variables, which offers the data consumer additional support in the development of a programming script (#I and VII). After data analysis, the HMDC may require the data consumer to record the developed script in a public repository and provide a link to the study protocol (see use case 3). This allows the HMDC to cross-reference it in the metadata assets of the datasets used (#V [43]). In doing so, transparency, reproducibility, and quality of studies are supported. Importantly, before publishing any study results, the HMDC should enable the data consumer to check whether approvals of data holders, of whom datasets were obtained, are required (D_{19}; #IV and V [43]).

XSL•FO
RenderX

Discussion

Principal Findings

The taxonomy detailed in this work provides 20 dimensions and 101 characteristics to develop FAIR HMDCs, representing initial, yet actionable design knowledge to answer RQ1. Comprehensive description is achieved when amplifying the taxonomy in the light of the cocreated use cases that entail real-world requirements (RQ2). The generated design knowledge provides added value because HMDCs facilitate effective and efficient use of RWD for generating RWE. Thereby, the obtained results accentuate the integration of FAIR principles into HMDCs to ensure "findability, accessibility, interoperability, and reusability" of the RWD circulating within the underlying health care data ecosystem [40]. The design knowledge can be classified into scientific and managerial contributions, as outlined in the following sections.

Interpretation of Findings

The taxonomy's *scientific contributions* intensify previous work on data catalogs, paying particular attention to, first, their implementation as decentralized components within data ecosystems and, second, their application in health care peripheries. Consequently, some of the design elements conceptualized in this study draw from prior research about metadata catalogs, as shown in the section Research gap, objective, and questions. Concurrently, they further spin the red paths of HMDC developments with respect to European initiatives pushing health care data ecosystems forward, for example, IDERHA, Elixir, HealthData@EU, EUCAIM, TEHDAS, or EHDEN (Multimedia Appendix 1). Therefore, the taxonomy closes the identified research gap.

On the one hand, the conceptualized dimensions and characteristics of the taxonomy describe and classify attributes of "HMDC metadata assets." First, they relate to attributes associated with data findability, such as data source, managerial details, the type of data records contained, data sensitivity, and population information. Second, they state crucial information pertaining to data reusability. This encompasses collection methodologies, data linkage, and preservation as well as consent and possible result publishing constraints. Third, the taxonomy emphasizes the need to specify data interoperability attributes in metadata assets. Among others, data standardization according to a CDM, the prescription of vocabularies, and the technological implementation of programmatic discoverability are important. On the other hand, the taxonomy contains design elements referring to more general data accessibility constraints pertaining to the "overall HMDC design." They accentuate basic security and governance considerations regarding dataset and catalog accessibility as well as the access control framework. Conclusively, from a scientific viewpoint, the artifact provides fundamental design knowledge [31] that unfolds broad implications and a solid starting point for future research.

Regarding *managerial contributions*, the taxonomy enables health care practitioners (see Introduction section for target audience) to navigate more effectively in the largely unexplored field of HMDCs, particularly focusing on their application in health care data ecosystems across Europe. It helps both researchers and practitioners to anchor and communicate their work [41]. The taxonomy also represents a support tool for developing HMDCs, where the illustrative scenarios assume an accentuated role. They showcase how the design elements are reflected in real-world use cases [43]. In essence, these use cases amplify that the taxonomy supports common activities for planning, assessing, and conducting medical research studies as well as benchmarking and analyzing the underlying RWD.

Subsequently, the contributions of the study are discussed in the light of 3 major issues. These are (1) the exclusiveness of taxonomy characteristics; (2) the difference between HMDCs and centralized data catalogs; and (3) the absence of data quality and data-sharing incentives as explicit taxonomy dimensions.

Depending on the meta-dimension, the taxonomy contains nonexclusive characteristics, which might be accumulated, to facilitate the design of metadata assets, that is, data findability, interoperability, and reusability. Alternatively, the taxonomy has mutual exclusive characteristics to classify and distinguish HMDC designs with respect to their data security and governance approaches, that is, data accessibility. This mixture of exclusive and nonexclusive dimensions can foster the understanding of health care practitioners, while allowing for an easy alteration of the taxonomy [37]. Convertibility is vital because HMDCs represent a rapidly evolving and changing field, where new solutions vanish and emerge constantly.

Furthermore, even though the objective of the taxonomy is not to differentiate between centralized and decentralized data catalogs, as distinguished in the Theoretical Background section, it pinpoints their fundamental design commonalities and differences. The meta-dimensions concerning the design of the metadata assets are conceivable for both approaches in health care contexts (ie, findability, interoperability, and reusability). The reason is that, despite datasets being stored centrally within intraorganizational data catalogs [26], meaningful metadata need to be disclosed to their data users by means of the catalog offerings. Naturally, the same holds for decentralized catalogs. However, the taxonomy also shows design differences with respect to its meta-dimension data accessibility. While decentralized catalogs can have various combinations of characteristics in the associated dimensions, centralized health data catalogs exhibit one specific pattern of characteristics. They usually are exclusively private systems ($C_{6.1}$) as their functionalities are only accessible to members of the operating organization. Similarly, dataset access is strictly limited to this specific group of predefined users ($C_{7.3}$). Finally, access control lies solely with the organization operating the centralized catalog ($C_{8.1}$).

As a last discussion point, the taxonomy contains neither dimensions associated with data quality nor incentives for data sharing. The reason is that these concepts, although important, are broad, multifaceted, and hardly explored, making a systematic categorization difficult. Generally, data quality involves subjective and context-dependent assessments [92], while incentives to share data are influenced by external, sociopolitical, and institutional factors [28]. Typically, HMDCs do not disclose any data quality metrics as those can barely be quantified and are subject to applied data types, formats, and

XSL•FO
RenderX

standards [41]. Rather, HMDCs publish test samples consisting of synthetic or fully anonymized data that do not justify a dimension in the taxonomy. Similarly, incentivizing data sharing, for example, via price tags for datasets or mandatory citations, represents an unsolved problem [93]. To circumnavigate this issue, HMDCs commonly rely on membership fees and public funding. The former restricts data access to members of the HMDC operating organization. The latter compensates data providers through public funds, usually applied in preliminary stages. Naturally, other business models exist. However, both data quality indicators and data-sharing incentives represent underdeveloped fields requiring future research beyond the study's design perspective [33]. Consequently, these concepts are not a part of the taxonomy, because they cannot be defined in universally applicable dimensions. Arguably, their inclusion would have overcomplicated the taxonomy and undermined its focus on actionable HMDC design knowledge.

Limitations

The taxonomy is mainly subject to the following limitations. In the inductive iterations, results were derived from a potentially limited number of publications because of the emphasis on 4 main databases. Similarly, in the deductive iterations, the examined analysis objects might merely cover a snapshot of what was available at the time (ie, many analysis objects have been in progress), be outdated quickly, and not be conclusive. As for the SLR, the conclusiveness of analysis objects is particularly questionable because of, first, the focus on European ecosystem initiatives and, second, a possible negligence of many centralized health data catalogs. In the evaluative iterations, the experts might not have captured the full range of relevant perspectives on HMDCs and are limited in number. Furthermore, the research design comprises certain limitations per se. As it is with qualitative research, taxonomy building requires substantial generalizations and simplifications of intricate and interdisciplinary content [83]. Although countermeasures were taken (see Methods section), these factors imply interpretative biases inevitably incorporated into the results [41], for example, extracting design elements from public data. Moreover, as shown in the Theoretical Background section, new HMDCs must be expected to arise constantly, while others are likely to disappear with a high frequency. Hence, the taxonomy must be altered swiftly. To conclude, the taxonomy provides first actionable design knowledge about HMDCs but requires continuous triangulation of design elements by future research.

Conclusions

Despite the limitations, the scientific and managerial contributions of this study unfold broad implications, which are formulated as recommendations for future research. Generally, HMDCs should be increasingly investigated in practice, for example, by more in-depth case studies. On the one hand, it is of utmost importance to keep track of the rapidly evolving HMDC-related initiatives in Europe. Their conceptual and technical advancements should be analyzed and evaluated constantly against the background of the taxonomy design elements, deriving the need for modifying dimensions and characteristics. On the other hand, by incorporating worldwide efforts toward health care data ecosystems and HMDCs, the scope of the taxonomy can be expanded and design knowledge beyond European jurisdictions can be created. In this regard, it is important to mention that the generated design knowledge about European HMDCs already entails such global implications. The FAIR dimensions of the taxonomy state fundamental characteristics of health data FAIR, making it universally relevant. In other words, the taxonomy conveys generally conceivable options for using catalog functionalities and underlying metadata assets. In addition, it outlines how to design those health metadata assets meaningfully. Therefore, despite the European focus, the taxonomy addresses global challenges with regards to health data sharing and metadata catalog designs, underlining its broad implications. Nevertheless, further research is essential, because HMDCs represent the fulcrum for allocating, exchanging, and using RWD to effectively generate RWE in emerging health care data ecosystems.

Acknowledgments

This research has emerged from the IDERHA (Integration of Heterogeneous Data and Evidence towards Regulatory and Health Technology Assessment Acceptance) project funded by the Innovative Health Initiative of the European Union (GAP-101112135) and the Digital Life Journey project funded by Fraunhofer-Gesellschaft (PN: 051-600000).

Data Availability

Data sharing is not applicable to this paper as no datasets were generated or analyzed during this study.

Authors' Contributions

SS conceptualized the study, conducted the investigation, developed the methodology, managed the project, supervised its execution, drafted the original manuscript, and contributed to reviewing and editing the manuscript. MKM contributed to the conceptualization and investigation, validated the findings, and participated in the review and editing of the manuscript.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
Leading European initiatives toward health care data ecosystems.
[DOCX File , 30 KB - formative_v9i1e63396_app1.docx]

Multimedia Appendix 2
Intermediary results of the iterative research process and key references.
[DOCX File , 70 KB - formative_v9i1e63396_app2.docx]

Multimedia Appendix 3
List of focus group members and distribution to sessions.
[DOCX File , 50 KB - formative_v9i1e63396_app3.docx]

References

1. Agarwal R, Gao GG, DesRoches C, Jha AK. The digital transformation of healthcare: current status and the road ahead. Inf Syst Res 2010;21(4):796-809. [doi: 10.1287/isre.1100.0327]
2. Singh G, Schulthess D, Hughes N, Vannieuwenhuyse B, Kalra D. Real world big data for clinical research and drug development. Drug Discov Today 2018;23(3):652-660 [FREE Full text] [doi: 10.1016/j.drudis.2017.12.002] [Medline: 29294362]
3. Remy L, Ivanović D, Theodoridou M, Kritsotaki A, Martin P, Bailo D, et al. Building an integrated enhanced virtual research environment metadata catalogue. Electron Libr 2019;37(6):929-951. [doi: 10.1108/el-09-2018-0183]
4. Feldman K, Johnson RA, Chawla NV. The state of data in healthcare: path towards standardization. J Healthc Inform Res 2018;2(3):248-271 [FREE Full text] [doi: 10.1007/s41666-018-0019-8] [Medline: 35415409]
5. Swertz M, van Enckevort E, Oliveira JL, Fortier I, Bergeron J, Thurin NH, et al. Towards an interoperable ecosystem of research cohort and real-world data catalogues enabling multi-center studies. Yearb Med Inform 2022;31(1):262-272 [FREE Full text] [doi: 10.1055/s-0042-1742522] [Medline: 36463884]
6. Bietz MJ, Bloss CS, Calvert S, Godino JG, Gregory J, Claffey MP, et al. Opportunities and challenges in the use of personal health data for health research. J Am Med Inform Assoc 2016;23(e1):42-48 [FREE Full text] [doi: 10.1093/jamia/ocv118] [Medline: 26335984]
7. Kush RD, Warzel D, Kush MA, Sherman A, Navarro EA, Fitzmartin R, et al. FAIR data sharing: the roles of common data elements and harmonization. J Biomed Inform 2020;107:103421 [FREE Full text] [doi: 10.1016/j.jbi.2020.103421] [Medline: 32407878]
8. Wiertz S, Boldt J. Ethical, legal, and practical concerns surrounding the implemention of new forms of consent for health data research: qualitative interview study. J Med Internet Res 2024;26:e52180 [FREE Full text] [doi: 10.2196/52180] [Medline: 39110970]
9. Molnár-Gábor F, Beauvais MJ, Bernier A, Jimenez MP, Recuero M, Knoppers BM. Bridging the European data sharing divide in genomic science. J Med Internet Res 2022;24(10):e37236 [FREE Full text] [doi: 10.2196/37236] [Medline: 36260387]
10. Lovestone S, EMIF Consortium. The European medical information framework: a novel ecosystem for sharing healthcare data across Europe. Learn Health Syst 2020;4(2):e10214 [FREE Full text] [doi: 10.1002/lrh2.10214] [Medline: 32313838]
11. Manogaran G, Varatharajan R, Lopez D, Kumar PM, Sundarasekar R, Thota C. A new architecture of internet of things and big data ecosystem for secured smart healthcare monitoring and alerting system. Future Gener Comput Syst 2018;82:375-387. [doi: 10.1016/j.future.2017.10.045]
12. Sharon T, Lucivero F. Introduction to the special theme: the expansion of the health data ecosystem – rethinking data ethics and governance. Big Data Soc 2019;6(2):205395171985296. [doi: 10.1177/2053951719852969]
13. Witte AK. A review on digital healthcare ecosystem structure: identifying elements and characteristics. In: Proceedings of the 23rd Pacific Asia Conference on Information System. 2020 Presented at: PACIS '20; July 9-12, 2020; Dubai, UAE p. 226-238 URL: https://aisel.aisnet.org/pacis2020/228 [doi: 10.1111/j.1467-8373.2010.01427.x]
14. Scheider S, Lauf F, Möller F, Otto B. A reference system architecture with data sovereignty for human-centric data ecosystems. Bus Inf Syst Eng 2023;65(5):577-595. [doi: 10.1007/S12599-023-00816-9]
15. Oliveira JL, Trifan A, Bastião Silva LA. EMIF catalogue: a collaborative platform for sharing and reusing biomedical data. Int J Med Inform 2019;126:35-45. [doi: 10.1016/j.ijmedinf.2019.02.006] [Medline: 31029262]
16. EHDS semantic interoperability framework 2022. TEHDAS. URL: https://tehdas.eu/app/uploads/2023/10/tehdas-recommendations-to-enhance-interoperability.pdf [accessed 2024-04-29]
17. Bergeron J, Doiron D, Marcon Y, Ferretti V, Fortier I. Fostering population-based cohort data discovery: the Maelstrom Research cataloguing toolkit. PLoS One 2018;13(7):e0200926 [FREE Full text] [doi: 10.1371/journal.pone.0200926] [Medline: 30040866]
18. Ulrich H, Kock-Schoppenhauer A, Deppenwiese N, Gött R, Kern J, Lablans M, et al. Understanding the nature of metadata: systematic review. J Med Internet Res 2022;24(1):e25440 [FREE Full text] [doi: 10.2196/25440] [Medline: 35014967]

XSL•FO
RenderX

19. Almeida JR, Silva JM, Oliveira JL. A FAIR approach to real-world health data management and analysis. In: Proceedings of the 36th International Symposium on Computer-Based Medical Systems. 2023 Presented at: CBMS '23; June 22-24, 2023; L'Aquila, Italy p. 892-897 URL: https://ieeexplore.ieee.org/document/10178764 [doi: 10.1109/cbms58004.2023.00338]
20. Derycke P, Kesisoglou I, Korsgaard T, Aage Huru A, Catsyne CA. Report on the landscape analysis of available metadata catalogues and the metadata standards in use. HaDEA & European Union. URL: https://ehds2pilot.eu/wp-content/uploads/2024/04/HealthData@EU-Pilot_MS6.1_FIN.pdf [accessed 2024-04-29]
21. Peng Y, Bathelt F, Gebler R, Gött R, Heidenreich A, Henke E, et al. Use of metadata-driven approaches for data harmonization in the medical domain: scoping review. JMIR Med Inform 2024;12:e52967 [FREE Full text] [doi: 10.2196/52967] [Medline: 38354027]
22. Lueschen G, van der Zee J. Health Systems in the European Union: Diversity, Convergence, and Integration: A Sociological and Comparative Analysis in Belgium, France, Germany, ... and Spain. München, Germany: Walter de Gruyter; 2016.
23. Lauf F, Scheider S, Friese J, Kilz S, Radic M, Burmann A. Exploring design characteristics of data trustees in healthcare - taxonomy and archetypes. In: Proceedings of the 31st European Conference on Information Systems. 2023 Presented at: ECIS '23; June 11-16, 2023; Kristiansand, Norway p. 33-57 URL: https://www.researchgate.net/publication/370060215_Exploring_Design_Characteristics_of_Data_Trustees_in_Healthcare_-_Taxonomy_and_Archetypes
24. Labadie C, Legner C, Eurich M, Fadler M. FAIR enough? Enhancing the usage of enterprise data with data catalogs. In: Proceedings of the 2020 IEEE 22nd Conference on Business Informatics. 2020 Presented at: CBI '20; June 22-24, 2020; Antwerp, Belgium URL: https://ieeexplore.ieee.org/document/9140254 [doi: 10.1109/cbi49978.2020.00029]
25. Jahnke N, Otto B. Data catalogs in the enterprise: applications and integration. Datenbank Spektrum 2023 Jun 21;23(2):89-96. [doi: 10.1007/S13222-023-00445-2]
26. Ehrlinger L, Schrott J, Melichar M, Kirchmayr N, Wöß W. Data catalogs: a systematic literature review and guidelines to implementation. In: Proceedings of the 2021 International Conference on Database and Expert Systems Applications. 2021 Presented at: DEXA '21; September 27-30, 2021; Virtual Event URL: https://link.springer.com/chapter/10.1007/978-3-030-87101-7_15 [doi: 10.1007/978-3-030-87101-7_15]
27. Gröger C. There is no AI without data. Commun ACM 2021 Oct 25;64(11):98-108. [doi: 10.1145/3448247]
28. Oliveira M, Barros Lima GD, Farias Lóscio B. Investigations into data ecosystems: a systematic mapping study. Knowl Inf Syst 2019;61(2):589-630. [doi: 10.1007/s10115-018-1323-6]
29. Sohail SA, Bukhsh FA, van Keulen M. Multilevel privacy assurance evaluation of healthcare metadata. Appl Sci 2021;11(22):10686. [doi: 10.3390/app112210686]
30. Almeida JR, Oliveira JL. MONTRA2: a web platform for profiling distributed databases in the health domain. Inform Med Unlocked 2024;45:101447. [doi: 10.1016/j.imu.2024.101447]
31. vom Brocke J, Winter R, Hevner A, Maedche A. Special issue editorial –accumulation and evolution of design knowledge in design science research: a journey through time and space. J Assoc Inf Syst 2020;21(3):520-544. [doi: 10.17705/1jais.00611]
32. van Aken JE. Valid knowledge for the professional design of large and complex design processes. Des Stud 2005;26(4):379-404. [doi: 10.1016/j.destud.2004.11.004]
33. Hevner AR, March ST, Park J, Ram S. Design science in information systems research. MIS Q 2004;28(1):75-105. [doi: 10.2307/25148625]
34. March ST, Storey VC. Design science in the information systems discipline: an introduction to the special issue on design science research. MIS Q 2008;32(4):725-730. [doi: 10.2307/25148869]
35. Cash PJ. Developing theory-driven design research. Des Stud 2018;56:84-119. [doi: 10.1016/j.destud.2018.03.002]
36. Kundisch D, Muntermann J, Oberländer AM, Rau D, Röglinger M, Schoormann T, et al. An update for taxonomy designers. Bus Inf Syst Eng 2021;64(4):421-439. [doi: 10.1007/s12599-021-00723-x]
37. Nickerson RC, Varshney U, Muntermann J. A method for taxonomy development and its application in information systems. Eur J Inf Syst 2017;22(3):336-359. [doi: 10.1057/ejis.2012.26]
38. Gregor S. The nature of theory in information systems. MIS Q 2006;30(3):611-642. [doi: 10.2307/25148742]
39. Szopinski D, Schoormann T, Kundisch D. Because your taxonomy is worth IT: towards a framework for taxonomy evaluation. In: Proceedings of the 27th European Conference on Information Systems. 2019 Presented at: ECIS '19; June 8-14, 2019; Stockholm-Uppsala, Sweden p. 25-44 URL: https://www.researchgate.net/publication/332711034_Because_your_taxonomy_is_worth_it_Towards_a_framework_for_taxonomy_evaluation
40. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data 2016;3(1):1-9 [FREE Full text] [doi: 10.1038/sdata.2016.18] [Medline: 26978244]
41. Scheider S, Lauf F, Geller S, Möller F, Otto B. Exploring design elements of personal data markets. Electron Markets 2023;33(1):1-16. [doi: 10.1007/S12525-023-00646-3]
42. Alvarellos M, Sheppard HE, Knarston I, Davison C, Raine N, Seeger T, et al. Democratizing clinical-genomic data: how federated platforms can promote benefits sharing in genomics. Front Genet 2022 Jan 10;13:1045450 [FREE Full text] [doi: 10.3389/fgene.2022.1045450] [Medline: 36704354]

XSL•FO
RenderX

43. Good practice guide for the use of the metadata catalogue of real-world data sources. European Medicines Agency. 2022. URL: https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/good-practice-guide-use-metadata-catalogue-real-world-data-sources_en.pdf [accessed 2024-04-29]
44. List of metadata for real world data catalogues. European Medicines Agency. 2022. URL: https://www.ema.europa.eu/en/documents/other/list-metadata-real-world-data-catalogues_en.pdf [accessed 2024-04-29]
45. Data directory. BBMRI-ERIC. URL: https://directory.bbmri-eric.eu/ERIC/directory/#/catalogue [accessed 2024-11-02]
46. An interactive catalogue of mental health and wellbeing measures in British cohort and longitudinal studies. Catalogue of Mental Health Measures. URL: https://www.cataloguementalhealth.ac.uk/ [accessed 2024-10-27]
47. Data catalogue for healthcare. Compendium. URL: https://compendiumdatacatalog.com/data-catalog/ [accessed 2024-10-28]
48. Findable, standardised data at scale through the EHDEN database catalogue. EHDEN Portal. URL: https://www.ehden.eu/ehden-portal/ [accessed 2024-10-22]
49. EMIF catalogue. EMIF. URL: https://www.emif.eu/emif-catalogue/ [accessed 2024-10-19]
50. Cancer image Europe. EUCAIM Catalogue. URL: https://catalogue.eucaim.cancerimage.eu/#/ [accessed 2024-10-25]
51. Welcome to the one-stop shop that facilitates access to population health and health care data, information and expertise across Europe. European Health Information Portal. URL: https://www.healthinformationportal.eu/ [accessed 2024-10-18]
52. Data catalogues for health. Fjelltopp. URL: https://www.fjelltopp.org/service/data-catalogues-for-health/ [accessed 2024-10-28]
53. Data catalogues. Health RI. URL: https://catalogus.healthdata.nl/datasets [accessed 2024-10-29]
54. Explore data sources. Helsedata. URL: https://helsedata.no/en/ [accessed 2024-10-29]
55. Health data catalog. IQVIA. URL: https://www.iqvia.com/library/fact-sheets/iqvia-health-data-catalog [accessed 2024-09-25]
56. Health pilot. Kraken. URL: https://www.krakenh2020.eu/pilots/health [accessed 2024-11-01]
57. Precision medicine data catalog. Lifebit. URL: https://www.lifebit.ai/federated-data-catalogue [accessed 2024-11-02]
58. Maelstrom catalogue. Maelstrom Research. URL: https://www.maelstrom-research.org/page/catalogue [accessed 2024-11-26]
59. Trials data catalogue. Yoda. URL: https://yoda.yale.edu/trials-search/ [accessed 2024-10-19]
60. Shi J, Zheng M, Yao L, Ge Y. DIR — a semantic information resource for healthcare datasets. In: Proceedings of the 2017 IEEE International Conference on Bioinformatics and Biomedicine. 2017 Presented at: BIBM '17; November 13-16, 2017; Kansas City, MO p. 805-810 URL: https://ieeexplore.ieee.org/abstract/document/8217758 [doi: 10.1109/bibm.2017.8217758]
61. McCoy D, Chand S, Sridhar D. Global health funding: how much, where it comes from and where it goes. Health Policy Plan 2009;24(6):407-417. [doi: 10.1093/heapol/czp026] [Medline: 19570773]
62. Yurkovich M, Avina-Zubieta JA, Thomas J, Gorenchtein M, Lacaille D. A systematic review identifies valid comorbidity indices derived from administrative health data. J Clin Epidemiol 2015;68(1):3-14. [doi: 10.1016/j.jclinepi.2014.09.010] [Medline: 25441702]
63. Pereira A, Almeida JR, Lopes RP, Oliveira JL. Querying semantic catalogues of biomedical databases. J Biomed Inform 2023;137:104272 [FREE Full text] [doi: 10.1016/j.jbi.2022.104272] [Medline: 36563828]
64. Brunswick D. Data privacy, data protection, and the importance of integration for GDPR compliance. Isaca J 2019;1:14-27.
65. Yang K, Jia X, Ren K, Zhang B, Xie R. DAC-MACS: effective data access control for multiauthority cloud storage systems. In: Proceedings of the 2013 IEEE Conference on Computer Communications Workshops. 2013 Presented at: INFOCOM '13; April 14-19, 2013; Turin, Italy p. 2895-2903 URL: https://ieeexplore.ieee.org/document/6567100 [doi: 10.1109/tifs.2013.2279531]
66. Shabani M, Borry P. "You want the right amount of oversight": interviews with data access committee members and experts on genomic data access. Genet Med 2016;18(9):892-897 [FREE Full text] [doi: 10.1038/gim.2015.189] [Medline: 26795589]
67. Munoz-Arcentales A, López-Pernas S, Pozo A, Alonso Á, Salvachúa J, Huecas G. An architecture for providing data usage and access control in data sharing ecosystems. Procedia Comput Sci 2019;160:590-597. [doi: 10.1016/j.procs.2019.11.042]
68. Ngouongo SM, Löbe M, Stausberg J. The ISO/IEC 11179 norm for metadata registries: does it cover healthcare standards in empirical research? J Biomed Inform 2013;46(2):318-327 [FREE Full text] [doi: 10.1016/j.jbi.2012.11.008] [Medline: 23246614]
69. de Mello BH, Rigo SJ, da Costa CA, da Rosa Righi R, Donida B, Bez MR, et al. Semantic interoperability in health records standards: a systematic literature review. Health Technol (Berl) 2022;12(2):255-272 [FREE Full text] [doi: 10.1007/s12553-022-00639-w] [Medline: 35103230]
70. Iroju O, Soriyan A, Gambo I, Olaleke J. Interoperability in healthcare: benefits, challenges and resolutions. Int J Innov Appl Stud 2013;3(1):262-270.
71. Roehrs A, da Costa CA, da Rosa Righi R, Rigo SJ, Wichman MH. Toward a model for personal health record interoperability. IEEE J Biomed Health Inform 2019;23(2):867-873. [doi: 10.1109/jbhi.2018.2836138]
72. Kent S, Burn E, Dawoud D, Jonsson P, Østby JT, Hughes N, et al. Common problems, common data model solutions: evidence generation for health technology assessment. Pharmacoeconomics 2021;39(3):275-285 [FREE Full text] [doi: 10.1007/s40273-020-00981-9] [Medline: 33336320]
73. Ivanović M, Budimac Z. An overview of ontologies and data resources in medical domains. Expert Syst Appl 2014;41(11):5158-5166. [doi: 10.1016/j.eswa.2014.02.045]

XSL•FO
RenderX

74. Hentschel J. Contextuality and data collection methods: a framework and application to health service utilisation. J Dev Stud 1999;35(4):64-94. [doi: 10.1080/00220389908422581]
75. Tijhuis M, Finger JD, Slobbe L, Sund R, Tolonen H. Data collection. In: Marieke V, van Oers H, editors. Population Health Monitoring: Climbing the Information Pyramid. Cham, Switzerland: Springer; 2019:59-81.
76. Bohensky MA, Jolley D, Sundararajan V, Evans S, Pilcher DV, Scott I, et al. Data linkage: a powerful research tool with potential problems. BMC Health Serv Res 2010;10(1):346-354 [FREE Full text] [doi: 10.1186/1472-6963-10-346] [Medline: 21176171]
77. Rasmussen KB, Blank G. The data documentation initiative: a preservation standard for research. Arch Sci 2007;7(1):55-71. [doi: 10.1007/S10502-006-9036-0]
78. Angrist M. Eyes wide open: the personal genome project, citizen science and veracity in informed consent. Per Med 2009;6(6):691-699 [FREE Full text] [doi: 10.2217/pme.09.48] [Medline: 22328898]
79. Kitchenham B, Pearl Brereton O, Budgen D, Turner M, Bailey J, Linkman S. Systematic literature reviews in software engineering – a systematic literature review. Inf Softw Technol 2009;51(1):7-15. [doi: 10.1016/j.infsof.2008.09.009]
80. Webster J, Watson RT. Analyzing the past to prepare for the future: writing a literature review. MIS Q 2002;26(2):13-23 [FREE Full text]
81. Mwita K. Factors influencing data saturation in qualitative studies. Int J Bus Soc 2022 Jun 05;11(4):414-420. [doi: 10.20525/ijrbs.v11i4.1776]
82. Proposal for a regulation - the European Health Data Space. Directorate-General for Health and Food Safety. 2022. URL: https://health.ec.europa.eu/publications/proposal-regulation-european-health-data-space_en#details [accessed 2024-11-07]
83. Saldana J. The Coding Manual for Qualitative Researchers. Thousand Oaks, CA: SAGE Publications; 2021.
84. Pratt MG. Fitting oval pegs into round holes. Organ Res Methods 2007;11(3):481-509. [doi: 10.1177/1094428107303349]
85. Fruhwirth M, Rachinger M, Prlja E. Discovering business models of data marketplaces. In: Proceedings of the 53rd Hawaii International Conference on System Sciences. 2020 Presented at: HICSS '20; January 7-10, 2020; Maui, HI p. 5738-5747 URL: https://scholarspace.manoa.hawaii.edu/server/api/core/bitstreams/cf7bab54-478b-412a-8742-6b02f10dd7ca/content [doi: 10.24251/hicss.2020.704]
86. Database (part of Elixir infrastructure). Elixir BioSamples. URL: https://www.ebi.ac.uk/biosamples/ [accessed 2024-12-03]
87. A registry of knowledgebases and repositories of data. Elixir FAIRsharing. URL: https://fairsharing.org/search?fairsharingRegistry=Database [accessed 2024-12-04]
88. Global health data exchange. Institute for Health Metrics and Evaluation. URL: https://ghdx.healthdata.org/ [accessed 2024-10-29]
89. MACH clinical and research dataset catalogue. Melbourne Academic Centre for Health. URL: https://figshare.unimelb.edu.au/MACH-catalogue?searchMode=1 [accessed 2024-11-24]
90. Möller F, Stachon M, Azkan C, Schoormann T, Otto B. Designing business model taxonomies – synthesis and guidance from information systems research. Electron Mark 2021;32(2):701-726. [doi: 10.1007/s12525-021-00507-x]
91. Pine KH. The qualculative dimension of healthcare data interoperability. Health Informatics J 2019;25(3):536-548 [FREE Full text] [doi: 10.1177/1460458219833095] [Medline: 31002277]
92. Wang RY, Strong DM. Beyond accuracy: what data quality means to data consumers. J Manag Inf Syst 2015;12(4):5-33. [doi: 10.1080/07421222.1996.11518099]
93. Spiekermann S, Acquisti A, Böhme R, Hui KL. The challenges of personal data markets and privacy. Electron Markets 2015;25(2):161-167. [doi: 10.1007/S12525-015-0191-0]

Abbreviations

CDM: common data model
DAC: Data Access Committee
E2C: empirical-to-conceptual
EHDEN: European Health Data and Evidence Network
EHDS: European Health Data Space
EHDS2: European Health Data Space 2
ETL: extract, transform, load
EU: European Union
EUCAIM: European Federation for Cancer Images
FAIR: findability, accessibility, interoperability, and reusability
HMDC: health metadata catalog
IDERHA: Integration of Heterogeneous Data and Evidence towards Regulatory and Health Technology Assessment Acceptance
IS: information systems
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RQ: research question

XSL•FO
RenderX

RWD: real-world data
RWE: real-world evidence
SLR: structured literature review
TEHDAS: Towards European Health Data Space

Edited by A Mavragani; submitted 19.06.24; peer-reviewed by Y Arayici, M Gersch; comments to author 06.08.24; revised version received 20.08.24; accepted 12.12.24; published 18.02.25.

Please cite as:
Scheider S, Mallick MK
Exploring Metadata Catalogs in Health Care Data Ecosystems: Taxonomy Development Study
JMIR Form Res 2025;9:e63396
URL: https://formative.jmir.org/2025/1/e63396
doi:10.2196/63396
PMID:39964739

©Simon Scheider, Mostafa Kamal Mallick. Originally published in JMIR Formative Research (https://formative.jmir.org), 18.02.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Design and Evaluation of a Simulation Tool for Audiology Screening Education: Design Science Approach

John Gerdes[1], BSc, MSc, MEng, MBA, PhD; Benjamin Schooley[2], MBA, PhD; Dakota Sharp[3], AuD; Juliana Miller[4], BSc, MSc, CCC-SLP

[1]Department of Integrated Information Technology, University of South Carolina, Columbia, SC, United States
[2]College of Engineering and Computing, Brigham Young University, Provo, UT, United States
[3]School of Medicine, University of North Carolina, Durham, NC, United States
[4]Department of Communication Sciences and Disorders, University of South Carolina, Columbia, SC, United States

Corresponding Author:
John Gerdes, BSc, MSc, MEng, MBA, PhD
Department of Integrated Information Technology
University of South Carolina
Bert Storey Innovation Center, Suite 1300
550 Assembly Street
Columbia, SC, 29208
United States
Phone: 1 803 777 8529
Email: jgerdes@sc.edu

Abstract

Background: The early identification of hearing loss and ear disorders is important. Regular screening is recommended for all age groups to determine whether a full hearing assessment is necessary and allow for timely treatment of hearing problems. Procedural training is needed for new speech-language pathology students as well as continuing education for those trained to perform this screening procedure. Limited availability and access to physical training locations can make it difficult to receive the needed training.

Objective: The aims of this study were to (1) develop a new hearing screening simulation software platform and (2) assess its effectiveness in training a group of graduate-level speech-language pathology students in hearing screening procedures.

Methods: An audiology simulator modeled after the commercial Grason-Stadler GSI39 combination audiometer and tympanometer device was developed to serve as a precursor to traditional face-to-face clinical instruction. A description of the simulator development process, guided by a design science approach, is presented. The initiation phase established the initial criteria for the simulator design. This was followed by an iterative process involving prototype development, review, and critique by the clinical faculty. This feedback served as input for the subsequent iteration. The evaluation of the final prototype involved 33 speech-language pathology graduate students as part of an introductory audiology class. These students were randomly assigned to control (receiving in-person instruction) and test (in-person instruction and simulation tool use) groups. Students in both groups were subsequently evaluated as they performed audiology screenings on human participants and completed a 25-item pretest and posttest survey. Nonparametric Mann-Whitney *U* tests were conducted on the mean differences between pretest and posttest ordinal survey response data to compare the control and intervention groups.

Results: The results indicated that the students who used the simulation tool demonstrated greater confidence in their ability to (1) explain hearing screening procedures to a child (P=.02), (2) determine whether otoscopy results are normal (P=.02), and (3) determine whether otoscopy results are abnormal (P=.03). Open-ended responses indicated that the students found that the hands-on experience provided by the simulator resulted in an easy-to-use and useful learning experience with the audiometer, which increased their confidence in their ability to perform hearing screenings.

Conclusions: Software-based education simulation tools for audiology screening may provide a beneficial approach to educating students and professionals in hearing screening training. The tool tested in this study supports individualized, self-paced learning with context-sensitive feedback and performance assessment, incorporating an extensible approach to supporting simulated subjects.

(JMIR Form Res 2025;9:e47150) doi:10.2196/47150

XSL•FO
RenderX

KEYWORDS

design science; audiology; simulation; hearing screening; framework; speech pathology; training

Introduction

Background

A national study in the United States found that 14.9% of children ranging in age from 6 to 19 years had low- or high-frequency hearing loss in one or both ears [1]. Early identification, diagnosis, and intervention are beneficial, improving the likelihood that children will develop effective communication and language skills and achieve successful learning outcomes [2]. For this reason, hearing screenings should be carried out regularly to identify problems early. To accomplish this, health care students and personnel must receive adequate training on how to perform these screenings and appropriately refer patients for evaluation [3].

The American Speech-Language-Hearing Association guidelines for audiological screening [4] specify procedures for completing hearing loss screenings, along with recommendations for screening the outer and middle ears, collecting an optional case history, performing a visual examination (otoscopy), and conducting acoustic immittance testing (tympanometry). Those who do not pass a hearing screening (the term “fail” is not used) are typically referred to a licensed audiologist for more detailed hearing testing or to a physician for any medical concerns. The screener can also refer the individual to both an audiologist and physician when the results warrant such action.

The American Speech-Language-Hearing Association recognizes the use of standardized patients and simulation technologies as alternatives to traditional clinical education methods [5]. The use of clinical simulation increased significantly during the COVID-19 pandemic because many practicum sites became inaccessible to students. Simulation is increasingly being used in clinical training and has been found to improve clinical knowledge and skills, self-confidence, communication skills, empathy, critical thinking abilities, leadership, and situation management among students [6-11]. Simulation has been found to improve the confidence in both knowledge and skills of audiology and speech-language pathology graduate students during training [12,13]. A survey of clinical educators found that simulation was a useful learning tool, provided a quality control measure, was an aid in professional development, and exposed students to clinical pathologies that were not routinely encountered [13]. Barriers to the use of simulation include faculty lacking training with simulators, limited funding to purchase simulator technology, and a lack of resources such as time and space [13].

Training speech-language pathology students to perform hearing screenings is an important undertaking. Students must learn how to interact with children and adults, properly work through the assessment protocol, and learn how to operate the equipment.

Studies have shown simulation to be effective in audiology education and training [14-16]. Simulation can help transfer theory to practice in an integrated teaching and learning model [17]. Simulators can present a full range of patient cases that represent real-world clinical diagnoses. These simulated cases present patients with symptoms that must be detected and diagnosed by the student. However, the use of simulation in audiology education is still in its infancy compared to some other fields, such as nursing and medicine [3].

It can be challenging for speech-language pathology students to obtain the necessary audiology training due to limited practicum opportunities. One alternative is to use a commercial simulator application, such as Otis–the virtual patient [18], which is a simulator designed for use in the education of audiologists. Many of these commercial simulators are designed to train students in conducting full audiological evaluations, rather than teaching hearing screening skills that follow an abbreviated protocol. However, these simulators can be cost prohibitive for training speech-language pathology students. Our goal was to deliver a realistic simulation tool providing a realistic representation of the controls found on a commercial testing instrument that students would work with when in the field. This design allows students to practice and develop testing skills without having to travel to a testing center. It also exposes them to a broad array of clinical cases and provides feedback on their performance. These goals became particularly relevant during the COVID-19 pandemic, when participating in live practicum experiences became difficult.

Prior research has explored the use of pure-tone audiometry simulators [19-23]. One system incorporated multilingual capability, improving accessibility for non-English speakers [21]. Audiology training using virtual reality was found to yield positive results compared to traditional training [11]. There are also commercially available simulation tools that are useful for training audiology skills. AudSim Flex [24] offers basic and advanced simulators for audiology students to learn pure-tone and bone conduction testing, both masked and unmasked. Otis–the virtual patient [18] realistically simulates the behavior of the patient and detects possible user errors immediately. It supports both pure-tone and screening training and guides the learner through exercises at several levels of difficulty, while providing useful help.

There are similarities between the hearing screening simulator (HSS) we developed and existing simulation tools. All incorporate a set of simulated subjects that present a range of hearing disorders, provide background narratives of the subjects to provide context, and model human behavior or actual subjects to improve realism. Most tools allow students to perform a simulated assessment and collect data that are then plotted on a standard audiogram, with the simulator assessing and critiquing the students’ analysis and conclusions.

The HSS differs from other simulators in 2 important ways. First, the HSS provides a digital twin—a realistic representation of the Grason-Stadler GSI39 combination audiometer and tympanometer device [25]—whereas other simulators use a generic design. By modeling the actual instrument interface, the HSS allows the student to become familiar with the instrument layout and practice working with the modeled

XSL•FO
RenderX

device's controls. Second, the HSS is designed for the simpler hearing screening protocol rather than full audiological testing. The protocols are different; therefore, the feedback and assessment verification must be tailored accordingly. Full audiological testing is outside the scope of practice for speech-language pathology students.

A design science approach was adopted for the development of the HSS. Initial design specifications were first established, followed by an iterative process involving prototype development, review, and critique by the clinical faculty, with the feedback used to refine the next iteration. After several iterations, the final prototype was tested to assess whether students felt that the simulator was useful in learning how to administer hearing screenings and to determine their confidence in their skills after using the simulator.

Objectives

The aims of this study were to (1) develop a new hearing screening simulation software platform and (2) assess its effectiveness in training a group of speech-language pathology students in hearing screening procedures. This work's contributions include a framework for developing a simulation of an existing process (hearing screening testing) and the description of a unique artifact with a user interface that closely emulates an existing commercial tool used to perform hearing screenings. This artifact supports an individualized, self-paced learning environment incorporating context-sensitive feedback and performance assessment. While further testing is needed, initial testing indicates that the HSS shows promise as a training aid for speech-language pathology students.

Methods

Overview

This research follows a design science methodology consisting of (1) problem identification and motivation, (2) the definition of the objectives for a solution, (3) design and development, (4) demonstration, (5) evaluation, and (6) communication [26,27]. We identified and translated design requirements for an audiology simulator into design components. We then classified these functional components in the application into design principles and features for designing a health care simulation app for clinical education and training. An abbreviated description of the design phase is provided in the next subsection, offering context for the evaluation phase of the study. For a more detailed description of the development process, the interested reader is directed to the conference paper [28]. For the evaluation of the simulation tool, participants were randomly separated into control (did not use the simulator) and test (used the simulator) groups, with both groups completing a pretest and posttest survey. Survey responses were analyzed using the nonparametric Mann-Whitney *U* test, which is appropriate for ordinal data analysis when a normal distribution cannot be assumed. Four qualitative survey question responses were also collected and analyzed. Moreover, audiology instructors assessed the skill levels of the student participants. Additional details on the analysis can be found in the Evaluation Study subsection.

The Simulation Tool's Conceptual Framework, Requirements, and Design

The simulation tool was designed to train speech-language pathology students on the proper use of audiometer equipment, how to administer a hearing screening test, and how to interpret the results. The tool's clinical and administrative functions, which are summarized in Textbox 1, served as the design criteria for the HSS. The clinical requirements reflect the data typically collected and presented during a traditional training session using commercial equipment. The operational requirements include additional features that enable a web-based digital learning experience, such as calculating student grades, and therefore represent value-added aspects of the simulator.

The HSS is designed to simulate 3 clinical procedures used in hearing screening and audiological evaluations. The otoscopic examination is a visual procedure used to assess the condition of the external auditory canal and tympanic membrane. Tympanometry tests how well the subject's eardrum moves and determines the volume of the ear canal and the pressure in the middle ear space. The audiometer presents a series of tones to the subject based on the appropriate protocol, and the subject indicates whether they hear the tone by raising their hand. For screening tests, tones are presented at a fixed level considered "normal hearing" (20 dB hearing level), in contrast to a typical diagnostic hearing test where multiple levels are tested to determine the subject's hearing threshold at each frequency (ie, the lowest level the subject can hear). The HSS uses standardized subjects with known auditory hearing profiles, along with sample otoscopic images and tympanometry results for both ears.

The simulator user interface consists of 3 panels. Figure 1A shows the simulated audiometer device, which was modeled after the Grason-Stadler GSI39 combination audiometer and tympanometer device [25]. The upper portion of the unit contains a graphic display that shows either the tympanogram or audiogram depending on the testing mode selected. The lower portion of the unit provides the controls used to perform the screening. These controls and their purpose are described herein. Figure 1B portrays the panel that provides information about the subject, as well as the command buttons to show and print a larger version of the audiogram chart, and displays the debriefing report that summarizes the students' performance on the simulation. Figure 1C shows the technician notes panel. It allows students to record their findings and include their observations, as is typical in a traditional audiological screening.

XSL•FO
RenderX

Textbox 1. Required clinical and operational functionalities of the hearing screening simulator.

Clinical requirements

- Authenticate and authorize users
- Provide a graphical user interface similar to that of standard audiology screening hardware-based systems
- Simulate 3 audiometry screens: video otoscopy, tympanometry, and audiometry
- Configure audiometer settings for a specific subset of frequencies used in hearing screening
- Display output similar to that obtained with commercial devices, including otoscopic images, tympanograms, and audiograms
- Include simulated patients representative of a range of real-world clinical cases

Operational requirements

- Provide a web-based platform
- Track and report time spent on the simulation
- Document and print a record of the simulation session
- Include automatic grading and feedback on students' performance
- Provide a full-featured audiometer with additional features beyond those needed for the hearing screening learning experience but present in commercial units
- Log all simulation screenings to permit reporting of an individual student's simulation history and enable cross-sectional data analysis across multiple students' results

Figure 1. The hearing screening simulator user interface. (A) The simulated audiometer device. (B) The panel that provides the simulated subject's details, with the avatar responding to tone stimulus. (C) Area for the technician to enter results and comments.

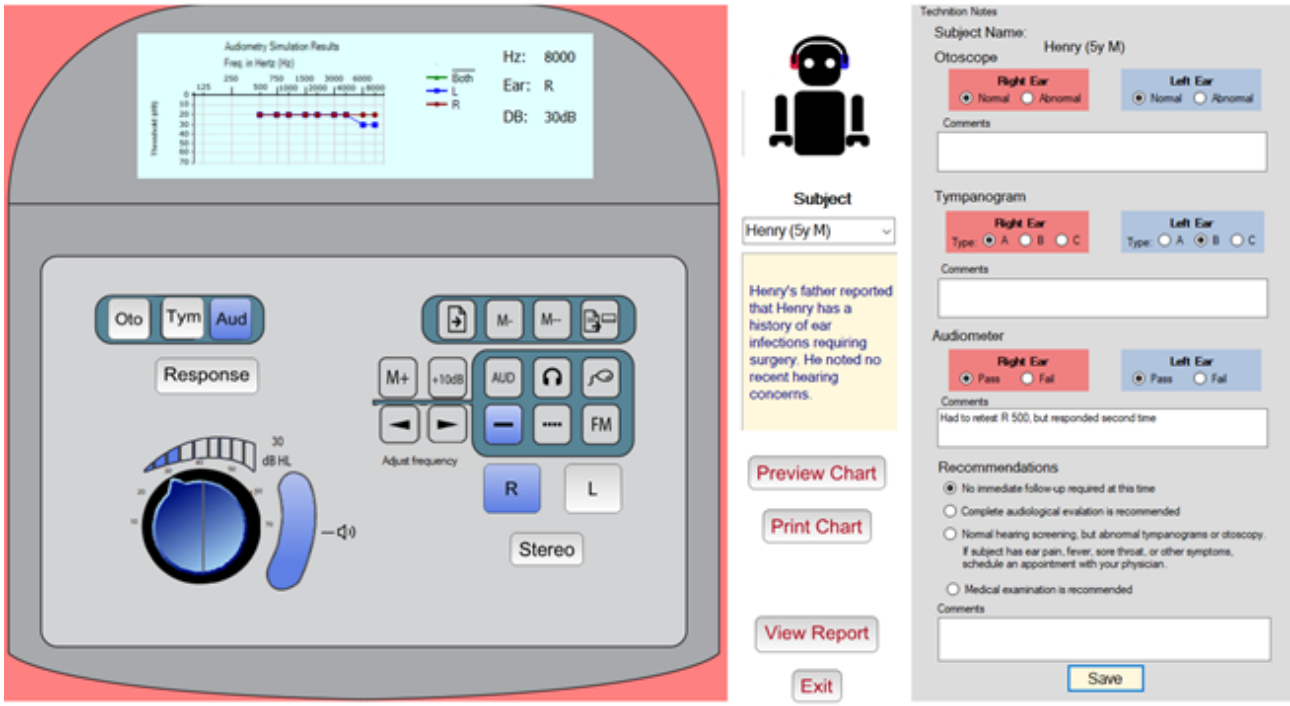

The HSS is equipped with 10 simulated patients, with the capability to expand this list. Each patient has a short case history, which is presented to the student when the subject is selected, providing the subject's age and sex and a brief biographical background related to any hearing concerns. The 10 artificial subjects represent different audiological conditions and pathologies, with audiometry, tympanometry, and otoscopy results consistent with their condition. One child and 1 adult standard patient with normal results were included. To obfuscate the identities of these simulated patients and create the illusion of having a much larger subject base, the profile information is randomized each time the program is run. Subject's sex is randomly assigned, age is adjusted slightly, the subject's name is randomly selected from a list of 100 diverse names, and the backstory is modified to match the new subject profile. This feature expands the number of possible simulated patient profiles to >1000 unique profiles. This prevents students from circumventing the training by using results from earlier simulations and also enables the assessment of a student's

XSL•FO
RenderX

precision when presented with multiple subjects derived from the same base simulated patient.

To perform the training, the student selects a standardized subject and reads the brief biographical sketch provided. The simulated screening test is then performed. The student must correctly set up the audiometer and work through the testing procedure. The otoscopic examination is simulated by selecting either the right or left ear and pressing the “Oto” button, which displays the subject’s right or left otoscopic image, depending on which ear is specified. Similarly, the student simulates the tympanometry test by selecting the ear and pressing the “Tymp” button, displaying the subject’s tympanometry result. To simulate the auditory test, the student selects the ear being tested, the tone level (20 dB hearing level) for the test, and then the series of frequencies specified by the testing protocol.

As the student proceeds through the simulation, they are asked to record their observations and patient results in the technician notes panel. There are note sections for each of the 3 screening procedures (otoscopy, tympanometry, and audiometry). The student is asked to provide the result for each ear and to add comments and observations related to the screening. Recording this information allows the simulator to assess the student’s performance. The system knows which result is correct based on the known profile of the subject being screened. Taking into consideration the 3 individual screening procedures, the student must also make a final recommendation regarding the patient’s subsequent follow-up and add a related comment. The student’s final recommendation is automatically graded and reported by the simulator, immediately providing appropriate feedback as needed. The patient cases, associated correct results, and correct recommendations were defined by the study clinical team and implemented in the application database.

Once the student completes the screening and enters their session notes, the data are saved to a database, and a session report is generated (Figure 2). This report serves as a starting point for a debriefing that would occur between a facilitator and the student. It provides metadata related to the simulation, including the technician’s name (ie, the student’s name), the subject’s name, the duration of the testing portion of the simulation, the screening procedures run by the student during the evaluation, and the total time spent on the simulation session. This information allows the simulator to perform validity checking of the training session, including whether the proper procedure was followed and the proper tests were run. The session report provides information to the simulation facilitator about the steps completed by the student during the session. Any shortcuts taken by the student during the session would be noted in the report. The report also displays the student’s results along with the correct results, with an explanation justifying the results. This provides the student with immediate feedback on their performance as well as their performance over time. Reports can be printed or saved as a PDF file to serve as required documentation of their clinical training.

XSL•FO
RenderX

Figure 2. Simulation report summary displaying student performance in administering a hearing screening to a simulated client. It includes session metadata and a simulation debriefing report that provides the students' responses, notes, and correct results.

Technician: John Smith

Date : Wednesday, June 2, 2021 9:34:46 AM

Testing duration:	36.6 Min		
Number Of Subjects tested:	1		
Auditory tests run:	21 right	4 left	0 both
Tympanometry tests run:	1 right	1 left	
Otoscope images viewed:	1 right	1 left	

Subject:	Henry (5y M)			
	Henry's father reported that Henry has a history of ear infections requiring surgery. He noted no recent hearing concerns.			
Otoscope	Right		Left	
Status:	Normal	[✗ - Incorrect]	Normal	[✗ - Incorrect]
Comment:				
Correct Response:	Abnormal, pressure equalization tube present in eardrum		Abnormal, PE tube	
Tympanogram	Right		Left	
Type:	Type A	[✗ - Incorrect]	Type B	[✓ - Correct]
Comment:				
Correct Response:	Type B, no eardrum movement and large ear canal volume		Type B, no eardrum movement and large ear canal volume	
Audiometer	Right		Left	
Screening:	Pass	[✓ - Correct]	Pass	[✓ - Correct]
Comment:	Had to retest R 500, but responded second time			
Correct Response:	Pass		Pass	
Recommendation: No immediate follow-up required at this time				[✗ - Incorrect]
Comment:				
Correct Response:	Normal hearing screening, but abnormal tympanograms or otoscopy. If you have ear pain, fever, sore throat, or other symptoms, schedule an appointment with your physician			

Evaluation Study

Student Participant Sampling, Intervention Procedure, and Data Collection

All 33 first-year speech-language pathology graduate students enrolled in an Introduction to Audiology class participated in this pilot study. The study started with 34 students, but since one student in the text group failed to submit the posttest survey they were dropped from the analysis resulting in 33 subjects. This is a required course in the graduate Speech-Language Pathology curriculum.

All participants received the same lecture on hearing loss concepts and performing hearing screening assessments. The 33 students were randomly assigned to either the control (n=16, 48%; all female) or test (n=17, 52%; all female) group by listing all participants in Microsoft Excel and using the built-in

XSL•FO
RenderX

randomization function. The control group received brief in-person training on equipment operation before being evaluated when they conducted their live, in-person audiology screening. The test group received the same in-person equipment operation introduction but used the simulation tool before being evaluated doing their live audiology screening. Students in each group were divided into 2-member teams, except for 1 group of 3 in the test group. Each student team alternated between roles, acting as a test subject (ie, patient) and as the test evaluator. The practice of students acting as both subject and evaluator is common in small class settings for audiology training. Approximately half of the students in the control group (16/33, 48%) and those in the test group (17/33, 52%) had previous hands-on experience with a hearing screening audiometer. Randomization included sorting students into an approximately equal number of pairings of "no previous experience," "1 partner with previous experience," and "both partners with previous experience."

Participants in the test group received supplementary training with the simulation tool. They were introduced to the simulator and given approximately 15 minutes to work through approximately 5 cases with their partner. A 25-item questionnaire was administered to all students to assess their confidence in their skills both before and after performing the live hearing screening. This means that the control and test groups completed the questionnaire both before and after the intervention. The questionnaire was created by the audiology faculty to gauge students' knowledge and confidence in administering and interpreting hearing screenings. A 5-point Likert scale was used. The response options and corresponding coding values were *strongly disagree* (1), *disagree* (2), *neither disagree nor agree* (3), *agree* (4), and *strongly agree* (5). A set of 7 to 9 questions were asked about each assessment mode (otoscopy, tympanometry, and audiometry). The test group was also asked 4 additional free-response questions in the same questionnaire to assess their perceptions of the simulation tool. A total of 16 surveys were completed by the control group and 17 completed by the test group. One test group survey was excluded from analysis because it represented the second of 2 surveys completed by the same participant, each providing identical feedback. Thus the analysis use 16 surveys in each group.

Student Participant Survey Data Analysis

A nonparametric analysis was conducted using the Mann-Whitney *U* test to compare differences between the groups at baseline (ie, before the test) and after the intervention (ie, after the test). The goal was to assess the change in the test and control groups' pretest versus posttest survey responses. The Mann-Whitney *U* test is a nonparametric method suitable for analyzing ordinal data when a normal distribution cannot be assumed. In this case, the ordinal variables were responses to survey questions measured on a Likert scale. This method is also appropriate when the groups being analyzed represent 2 independent samples (test and control groups created via randomization). The null hypothesis is that the means of each group are identical, meaning that all groups come from the same distribution. The alternative hypothesis is that at least 1 of the groups has a different mean, meaning that at least 1 group comes from a different distribution than the others. The test was corrected for tied ranks.

It was expected that using the simulation training tool together with in-person training would be perceived by the students to be at least as effective as, and possibly an improvement over, in-person–only training using commercial audiology equipment. Although some indication of improvement was expected, statistically significant differences between the groups were not expected for most questionnaire items between the test and control groups. All statistical analyses were conducted using a standard, commercially available statistical software tool (NCSS 2021; NCSS Statistical Software). A thematic analysis of qualitative responses was conducted to produce categorical findings, which were then discussed.

Student Qualitative Data Collection and Analysis

As part of the posttest survey administered to the test group, 4 free-response questions were included regarding their impressions of using the simulator tool. Qualitative responses were typed, and we conducted a thematic analysis across the responses, which were then grouped, summarized, and counted.

Instructor Evaluation of Student Hearing Screening Assessment Technique

A live assessment of the students' clinical techniques was completed using commercial audiometry units, with peers role-playing as patients. Certified instructors supervised the assessment, observing and grading the students' performance on otoscopy, tympanometry, and hearing screening techniques on a scale ranging from 1 to 5. Scores were compiled, averaged, and compared across the control and test groups.

Ethical Considerations

The institutional review board of the University of South Carolina determined that this research did not involve human subjects. All participants received a verbal and written description of the study and consented to participate. Participants' identities were kept confidential and anonymized before data analysis. No compensation was provided for participation in the study.

Results

The Simulation Tool's Conceptual Framework

Overview

In a design science study, the objective is to assess how well the created artifact, in this case, the HSS, meets the target operational requirements. In this study, we focused on assessing how well the tool performed in delivering the necessary training on the audiometer equipment. The results are presented in the following subsections.

Evaluation Study

Likert Scale Survey Results

The test group showed higher posttest mean scores and greater percentage change than the control group across all 3 test component groupings: otoscope, tympanometer, and hearing screening (Table 1).

XSL•FO
RenderX

Table 1. Multifactor assessment of student confidence change resulting from the use of the simulation tool. Pretest and posttest values represent the averages for both control and test groups.

	Control group (n=16)			Test group (n=17)		
	Pretest score, mean (SD)	Posttest score, mean (SD)	Change (% change)	Pretest score, mean (SD)	Posttest score, mean (SD)	Change (% change)
Otoscope (7 factors)	3.59 (0.41)	3.90 (0.63)	0.31 (8.64)	3.73 (0.58)	4.28 (0.55)	0.55 (14.75)
Tympanometer (9 factors)	3.63 (0.32)	4.28 (0.50)	0.63 (17.35)	3.79 (0.70)	4.62 (0.44)	0.83 (21.90)
Hearing screening (9 factors)	4.01 (0.56)	4.39 (0.45)	0.38 (9.48)	3.84 (0.59)	4.43 (0.45)	0.59 (15.36)

As shown in Table 2, the responses to the following 3 survey questions showed statistically significant differences between the groups with significance set at $P<.05$: “I am confident in my ability to explain hearing screening procedures to a child” (Mann-Whitney $U=71.0$, n1=n2=16; $P=.02$, 2-tailed t test), “I am confident in my ability to determine if otoscopy is normal” (Mann-Whitney n1=n2=16; $P=.02$, 2-tailed), and “I am confident in my ability to determine if otoscopy is abnormal” (Mann-Whitney $U=74.5$, n1=n2=16 $P=.02$, 2-tailed t test). For technology pilot studies, findings at the $P<.10$ significance level can also demonstrate important results. In this sense, the responses to the following 2 survey questions showed statistically significant differences between the groups with significance set at $P<.10$: “I am confident in my ability to locate landmarks while completing otoscopy” (Mann-Whitney $U=86.5$, n1=n2=16; $P=.08$, 2-tailed t test) and “I am confident which dB level should be tested during a hearing screening” (Mann-Whitney $U=84.0$, n1=n2=16; $P=.07$). The results were not significant across 19 survey questions at the $P<.10$ significance level, underscoring the specific impact of the simulator on the aforementioned confidence measures.

The survey asked whether students had previously conducted an audiology screening test on a human subject before the class audiology training exercise and study. A total of 8 (24%) of the 33 participants in the test group and 6 (18%) of the 33 participants in the control group reported in the affirmative. For this smaller subset of participants, mixed results showed higher posttest mean scores and greater percentage change for the otoscope and tympanometer control groups than the test groups. There was a negative change and percentage change for the combined hearing screening scores for the control group, while the test group had a positive change and percentage change (Table 3). Statistically significant differences were not calculated due to the small sample sizes across the groups.

XSL•FO
RenderX

Table 2. Analysis of survey responses grouped by otoscopy, tympanometry, and hearing screening questions, comparing the control and test group mean scores for pre- and postassessment sentiment. The Mann-Whitney U test was used to assess the statistical significance of the change in sentiment between the 2 groups.

Survey questions	Control group		Test group		*P* value
	Preassessment score, mean (SD)	Postassessment score, mean (SD)	Preassessment score, mean (SD)	Postassessment score, mean (SD)	
Otoscopy					
I am confident in my ability to interpret otoscopy	3.63 (0.62)	3.93 (0.80)	3.65 (1.00)	4.06 (0.75)	.74
I am confident in my ability to complete otoscopy on a patient	3.19 (0.75)	3.80 (0.77)	3.24 (1.15)	4.18 (0.64)	.39
I am confident in my ability to locate landmarks while completing otoscopy	3.44 (0.63)	3.73 (0.70)	3.53 (0.94)	4.35 (0.70)	.08[a]
I am confident in my ability to determine if otoscopy is abnormal	3.81 (0.54)	3.60 (1.06)	3.82 (0.73)	4.18 (0.81)	.03[b]
I am confident in my ability to determine if otoscopy is normal	3.88 (0.50)	4.00 (0.85)	3.94 (0.66)	4.41 (0.62)	.02[b]
I am confident in my ability to determine if otoscopy warrants a referral	3.75 (0.68)	3.93 (0.59)	3.88 (0.49)	4.24 (0.75)	.43
I am confident in my understanding of the impact of abnormal otoscopy on hearing screening results	3.38 (0.89)	4.27 (0.46)	4.06 (0.66)	4.56 (0.51)	.32
Tympanometry					
I am confident in my ability to interpret a tympanogram	3.53 (0.50)	4.40 (0.51)	3.76 (0.97)	4.65 (0.49)	.43
I can confidently identify a normal ear canal volume (ECV) on a tympanogram	3.81 (0.54)	4.33 (0.49)	4.06 (0.56)	4.65 (0.49)	.45
I can confidently differentiate between tympanogram types	3.75 (0.45)	4.33 (0.62)	3.88 (0.86)	4.65 (0.49)	.76
I can confidently identify an abnormal tympanogram	3.81 (0.40)	4.33 (0.62)	4.00 (0.94)	4.71 (0.47)	.66
I can confidently identify a normal tympanogram	3.88 (0.34)	4.40 (0.51)	4.24 (0.56)	4.76 (0.44)	.86
I can confidently complete a tympanogram on a patient	3.13 (0.81)	4.29 (0.83)	3.18 (1.19)	4.59 (0.51)	.51
I am confident in my ability to determine if a tympanometry result warrants a referral	3.69 (0.70)	4.13 (0.52)	3.65 (0.86)	4.53 (0.62)	.12
I am confident in my understanding of the impact of abnormal tympanometry on hearing screening results	3.38 (0.81)	4.20	3.76 (0.75)	4.59 (0.51)	.85
I am confident in my ability to connect abnormal otoscopy results to abnormal tympanometry results	3.56 (0.51)	3.93 (0.80)	3.59 (1.12)	4.47 (0.62)	.76
Hearing screening					
I am confident in which frequencies should be tested during a hearing screening	4.31 (0.60)	4.79 (0.43)	4.12 (0.78)	4.65 (0.61)	.24
I am confident which dB level should be tested during a hearing screening	4.13 (0.72)	4.53 (0.52)	3.76 (0.90)	4.53 (0.72)	.07[a]
I am confident in my ability to explain hearing screening procedures to an adult	3.63 (0.96)	4.27 (0.70)	3.65 (0.86)	4.65 (0.49)	.42

XSL•FO
RenderX

Survey questions	Control group		Test group		*P* value
	Preassessment score, mean (SD)	Postassessment score, mean (SD)	Preassessment score, mean (SD)	Postassessment score, mean (SD)	
I am confident in my ability to explain hearing screening procedures to a child (e.g., conditioned play audiometry)	3.75 (1.00)	4.00 (0.85)	3.38 (0.81)	4.29 (0.85)	.02[b]
I am confident in how to present stimuli/tones while completing a hearing screening	3.81 (0.98)	4.36 (0.50)	3.47 (1.07)	4.65 (0.49)	.11
I am confident in how to verify a response while completing a hearing screening	4.00 (0.82)	4.27 (0.70)	4.00 (0.87)	4.53 (0.51)	.22
I am confident in knowing when a referral is necessary based on hearing screening results	4.06 (0.57)	4.50 (0.52)	3.94 (0.75)	4.41 (0.62)	.60
I am confident in knowing which professional to refer to based on hearing screening results	4.38 (0.50)	4.47 (0.52)	4.35 (0.49)	4.47 (0.62)	.75

[a]Significance set at $P<.10$.

[b]Significance set at $P<.05$.

Table 3. Multifactor assessment of student confidence change for students reporting having conducted an audiology screening test on live subjects before the audiology test. Pretest and posttest values represent the average values.

	Control group (n=6)			Test group (n=8)		
	Pretest scores, mean (SD)	Posttest scores, mean (SD)	Change (% change)	Pretest scores, mean (SD)	Posttest scores, mean (SD)	Change (% change)
Otoscope (7 factors)	3.83 (0.40)	4.36 (0.40)	0.53 (13.8)	3.91 (0.49)	4.30 (0.51)	0.39 (10.0)
Tympanometer (9 factors)	3.83 (0.21)	4.67 (0.42)	0.84 (21.9)	4.24 (0.47)	4.75 (0.38)	0.51 (12.0)
Hearing screening (9 factors)	4.56 (0.30)	4.66 (0.41)	−0.10 (−2.2)	4.21 (0.61)	4.67 (0.45)	0.46 (10.9)

Qualitative Questionnaire Results

The results of the 4 free-response questions that were administered to the test group regarding their impression of the simulator tool are presented in Table 4. We categorized the responses based on their context similarity (type) and frequency of being mentioned.

The responses to the first question indicate that the students found that the hands-on experience gained by using the simulator provided an easy-to-use and useful learning experience with the audiometer that increased their confidence in their ability to perform hearing screening.

With regard to the second question, the simulator was designed to display an image or chart for a simulated subject, rather than demonstrate how to perform an otoscopic or tympanometric examination. Students wanted more explanation on how to perform these examinations. Others felt that the instructions, in general, could be improved, possibly incorporating pictures or a video. One participant expressed a concern that the simulator design did not resemble an audiometer machine. Given that the simulator was modeled to closely resemble the Grason-Stadler GSI39 combination audiometer and tympanometer device, it is assumed that the student was familiar with a different model.

The most frequent response to the third question was that the simulator should improve training on how to perform otoscopy and tympanometry screenings. It was also suggested that the simulator model the instructions provided by the technician to the subject. Future software iterations will address these suggestions.

The responses to the final question suggested positivity toward the simulation tool, indicating its potential benefit in terms of providing a range of realistic learning experiences. The responses do suggest that some features of the simulator need to be explained further in the system documentation. The final 3 comments emphasized that the simulator was easy to use and was perceived as a useful addition to the class.

XSL•FO
RenderX

Table 4. Poststudy survey of impressions of the hearing screening simulator (n=17).

	Frequency of being mentioned, n (%)
"What aspects of the simulation did you like?"	
Provides hands-on experience, helped me become more familiar with an audiometer, and increased my confidence	5 (29)
Ease of use	4 (23)
Practice with multiple virtual patients	4 (23)
Provided useful feedback on the simulation debriefing report	1 (6)
Thought it was helpful	1 (6)
"What aspects of the simulator did you not like?"	
Did not show how to do otoscopic or tympanometry exam	4 (23)
Having better instructions would be helpful	4 (23)
Nothing, this is a useful simulation	2 (12)
It was repetitive	1 (6)
"Do you have any recommendations to improve the simulator?"	
The simulator did not show how to perform otoscopy and tympanometry. Clearer instructions are needed in general	5 (29)
Nothing, no improvement needed	3 (18)
Model the technician giving required directions to the subject	1 (6)
Have the design of the simulator match that of the audiometer	1 (6)
"Would you like to share anything else?"	
A great tool to add to any audiology class	1 (6)
Really liked that sample participants showed inconsistencies in hearing thresholds. This realistic touch prepares one to be not thrown off when performing a live, in-person screening	1 (6)
The simulation was easy to use	1 (6)
The virtual simulation was very helpful	1 (6)
I loved the activity!	1 (6)

Instructor Evaluation of Student Hearing Screening Assessment Technique

Certified instructors supervised the assessment, observing and grading the students' performance and technique. The summary data from the instructor's assessment is given in Table 5. Assessment data for two students in the control group were not available which reduced the size of the control group to 14. Data for one student in the test group was dropped because of incomplete data resulting in a group size of 16. All students in both the control and test groups performed the 3 screenings satisfactorily (ie, otoscopy, tympanometry, and hearing screening). Some students, primarily in the control group, did not follow the prescribed protocol exactly, as reflected in the technique scores shown in Table 5. Students performed similarly across test and control groups, except in the case of providing instructions and screening according to protocol.

Table 5. Certified technicians' assessment of study participants' techniques in performing otoscopic, tympanometric, and audiometric testing on live subjects using commercial audiometer equipment.

	Control group (n=14), mean (SD)	Test group (n=16), mean (SD)
Otoscope	4.71 (0.47)	4.50 (0.52)
Tympanometer	4.64 (0.74)	4.38 (0.62)
Audiometer		
Provide instructions	3.64 (1.39)	4.63 (0.81)
Test according to protocol	4.64 (0.50)	4.81 (0.40)

XSL•FO
RenderX

Discussion

Principal Findings

The HSS is designed as a training tool for speech-language pathologists, nurses, and other professionals to learn the skills needed to perform hearing screening. It can provide remote training when it is not practical to use a traditional practicum setting with commercial audiology equipment, and thus represents an improvement on the state of the art [26]. The design science framework described in this study could be used to replicate and extend the existing simulator and develop other simulation training tools; for example, this implementation is modeled on a specific commercial audiometer. It would be useful to expand this to an array of available audiometers and allow students to receive a flexible range of training on instruments of their choosing. This would be similar to how Microsoft Flight Simulator [29] allows users to choose from 30 different cockpits. During our testing, a respondent commented that the simulation interface did not represent the instrument they were using, highlighting the need to expand the tool to model other instruments. The user training interaction model created for this study incorporated artificial patients in a software simulation within a theoretical framework, operating in real-time and real-world environments. The tool uses external configuration files that enable the addition of new simulated subjects. Each simulated subject has a corresponding biographical sketch, auditory and tympanometric profiles, and otoscopic images. This design feature allows the addition of new simulated patients, covering a range of possible real patient conditions. The software simulation tool provides a design artifact that was tested and evaluated across 33 participants and thus provides a complete design science research cycle, inclusive of artifact evaluation.

This study focused on the user experience that influenced the design parameters of the HSS. In the evaluation, we surveyed users to assess usability, as well as efficacy as a teaching tool. While the mean scores of the test group were largely higher than those of the control group and demonstrated a larger percentage change from before to after the intervention, there were few statistically significant differences between the groups. This was not surprising, considering the very early phase of research in this pilot program and the known difficulties of measuring educational interventions, especially with small sample sizes. Furthermore, the percentage changes from before to after the intervention were mixed for those participants who had previously conducted a screening examination on a real person before participating in this training exercise and study. While the small sample size of this subgroup makes it difficult to draw firm conclusions in this regard, it is important to note that prior screening experience may impact hearing screening learning outcomes and the effect of the simulator on learning. Taken together, the quantitative and qualitative results from this endeavor show significant promise for audiology simulator tools in enhancing the education of speech-language pathology students. The results indicate that the working prototype of the simulator may provide a strong supplement to in-person training using commercially available equipment. Participants noted that the simulator was intuitive and useful for learning the skills needed to perform hearing screenings, while also indicating areas for improvement.

Limitations

The number of participants in this study was low (n=33), yet appropriate for an initial pilot test. Future studies should include a more diverse set of participants, possibly from multiple geographic locations. This represents the first test of the HSS simulation software. Future iterations of the software may demonstrate different results. A common issue in the control group was a failure of student evaluators to communicate instructions to the subjects as prescribed in the protocol. This issue was not as prevalent in the test group, which may indicate that the added simulator practice impacted the students' performance in following the procedure, although the simulator did not explicitly indicate or model this behavior. The observed issues were not modeled in the simulation and thus represent potential areas for future improvement.

The study design could be improved. In this iteration, the students in the test group received traditional lectures on the material and then completed the simulation before their performance and technique were evaluated. The control group received only the traditional lectures before evaluation. The difference in the observed results could therefore be due to the test group's greater exposure to the relevant content. In future testing, the control group could be given an alternative, related activity with a similar duration as the simulation. This could include additional reading or researching audiometer test equipment on the internet.

Future Directions

Our testing of the HSS found that it has the potential to be an effective audiometry teaching tool, but some areas for improvement and future research directions were identified to validate these findings. The most significant improvement would be to convert it to a web-based platform for broader availability and accessibility. While remote students can currently use the simulator, it requires them to download and install the tool locally. A web-based option would make using the simulator easier, with the only requirement being a web browser.

The functionality of the simulator could be extended by allowing the configuration to support customizable pass and refer cutoffs to address different environments or countries. Similarly, allowing the configuration to support multiple languages could be valuable. This could be implemented by modifying and displaying text in different user-specified languages or incorporating "tool tips" to offer second-language support in pop-up windows.

Another important extension would be the inclusion of different models of commercial audiometer test equipment. A key benefit of the tool is that it provides training using a digital twin, providing the learning experience with a test unit user interface. The value of this feature diminishes if the technician is not using the same equipment modeled in the tool. This issue was mentioned by a participant in the exit survey. Although the tool is designed for hearing screening test training, it could be extended to handle diagnostic hearing test training as well.

XSL•FO
RenderX

While the physical equipment remains the same, the testing protocols differ and would have to be modified.

Recommendations from students using the system included that a video tutorial should be created to either replace or supplement the current text-based instructions. The students also suggested adding a learning mode, where context-sensitive feedback could be displayed on command during the training phase but hidden during the evaluation phase. As a result of the testing, it was determined that the simulator did not verify the students' compliance with the required auditory screening protocols. While the system counts the number of frequency screens completed, it does not track the testing of specified frequencies (1000, 2000, and 4000 Hz tones), repeated tests, the number of positive responses, or the sound levels used. Therefore, while the simulator did provide an assessment of the students' ability to determine the correct result, it did not fully capture the performance as specified in the audiology screening guidelines [4]. These factors were considered by the evaluator when grading the students during the evaluation phase of the exercise. These features are planned for future versions of the simulator.

To further access the effectiveness of the simulator tool, a series of studies would need to be conducted with a broader set of participants across a broader set of training programs (eg, universities) dispersed geographically. These studies would need to assess the educational outcomes associated with the use of the tool.

Conclusions

This work's contributions include a design framework for developing a tool that simulates an existing process, the description of a unique artifact that supports an individualized and self-paced learning environment incorporating context-sensitive feedback and performance assessment, and an extensible approach to supporting simulated subjects in clinical training testing artificial patients. The software development process, developed artifact, and user evaluation together provide a comprehensive, descriptive framework for designing audiology simulation training tools as well as validation that such tools may provide a cost-effective mechanism for training a wide range of providers for audiology screening.

Acknowledgments

The authors wish to acknowledge the assistance of Danielle Varnedoe, the director of the Montgomery Speech-Language-Hearing Clinic at the University of South Carolina, for her support in the development of the hearing screening simulator. The authors also thank Dr Gordon Schooley, PhD, for his consultation and assistance with the statistical analysis. The authors also acknowledge the work by the student capstone teams that worked on early versions of the tool: Austin Wicker; Elizabeth Saiz; Saniya Jumani; Davon Davis; Kristin Jackson; Oneil Clemenson, Jr; Brian Morrison; Christopher Bomar; Kasara Tyer; and Mary Holliday.

Disclaimer

An earlier version of this paper was presented at the 55th Hawaii International Conference on System Sciences [28]. This paper differs from the earlier conference paper. The discussion of the iterative design of the simulator tool was removed, with interested readers directed to the conference paper. This study changed statistical methodology from the Kruskal-Wallis test to the Mann-Whitney U test to more appropriately address the study's pretest-posttest design with ordinal data. Statistical reporting was expanded to include $P<.05$ and $P<.10$, both of which are often used for reporting in pilot studies. Additional mean change and percentage change analyses were conducted and reported on regarding the prior audiology testing experience of participants. A new table was added to present the additional analysis. Corrections were made throughout the paper, with significant reduction and rewriting of the introduction and literature review. The discussion section was expanded based on the expanded statistical analysis.

Data Availability

The datasets generated and analyzed during this study are available in the OSF repository "Hearing Screening Simulator Data" [30] under the Creative Commons Attribution 4.0 International Public License.

Authors' Contributions

JG was the lead application developer of the hearing screening simulator discussed in the paper and is the primary author. BS processed the statistics and analysis and assisted in the preparation of the manuscript. DS and JM tested and provided feedback on each iteration as it was developed. DS and JM provided domain-specific knowledge that aided in the development of the tool and contributed to the paper.

Conflicts of Interest

DS and JM work in the Montgomery Speech-Language-Hearing Clinic in the Department of Communication Sciences and Disorders at the University of South Carolina, which was a partner in the development of the hearing screening simulator. All other authors declare no other conflicts of interest.

XSL•FO
RenderX

References

1. Niskar AS, Kieszak SM, Holmes A, Esteban E, Rubin C, Brody DJ. Prevalence of hearing loss among children 6 to 19 years of age: the Third National Health and Nutrition Examination Survey. JAMA 1998 Apr 08;279(14):1071-1075. [doi: 10.1001/jama.279.14.1071] [Medline: 9546565]
2. Guralnick MJ. Why early intervention works: a systems perspective. Infants Young Child 2011 Jan 01;24(1):6-28 [FREE Full text] [doi: 10.1097/IYC.0b013e3182002cfe] [Medline: 21532932]
3. Alanazi AA. The use of simulation in audiology education to improve students' professional competency. ED.gov. 2017. URL: https://eric.ed.gov/?id=ED577397 [accessed 2024-04-29]
4. American Speech-Language-Hearing Association. Guidelines for Audiologic Screening. New York, NY: American Speech-Language-Hearing Association; 1997.
5. 2014 standards for the certificate of clinical competence in speech-language pathology. Council for Clinical Certification in Audiology and Speech-Language Pathology of the American Speech-Language-Hearing Association (ASHA). 2020. URL: https://www.asha.org/certification/2020-slp-certification-standards/ [accessed 2025-01-19]
6. Cant RP, Cooper SJ. Simulation-based learning in nurse education: systematic review. J Adv Nurs 2010 Jan;66(1):3-15. [doi: 10.1111/j.1365-2648.2009.05240.x] [Medline: 20423432]
7. Guise V, Chambers M, Välimäki M. What can virtual patient simulation offer mental health nursing education? J Psychiatr Ment Health Nurs 2012 Jun;19(5):410-418. [doi: 10.1111/j.1365-2850.2011.01797.x] [Medline: 22070549]
8. La Cerra C, Dante A, Caponnetto V, Franconi I, Gaxhja E, Petrucci C, et al. Effects of high-fidelity simulation based on life-threatening clinical condition scenarios on learning outcomes of undergraduate and postgraduate nursing students: a systematic review and meta-analysis. BMJ Open 2019 Feb 22;9(2):e025306 [FREE Full text] [doi: 10.1136/bmjopen-2018-025306] [Medline: 30798316]
9. Sherwood RJ, Francis G. The effect of mannequin fidelity on the achievement of learning outcomes for nursing, midwifery and allied healthcare practitioners: systematic review and meta-analysis. Nurse Educ Today 2018 Oct;69:81-94. [doi: 10.1016/j.nedt.2018.06.025] [Medline: 30015220]
10. Brown DK. Simulation before clinical practice: the educational advantages: audiology today. American Academy of Audiology. 2017. URL: https://www.audiology.org/news-and-publications/audiology-today/articles/simulation-before-clinical-practice-the-educational-advantages/ [accessed 2024-04-29]
11. Bakhos D, Galvin J, Aoustin JM, Robier M, Kerneis S, Bechet G, et al. Training outcomes for audiology students using virtual reality or traditional training methods. PLoS One 2020;15(12):e0243380 [FREE Full text] [doi: 10.1371/journal.pone.0243380] [Medline: 33270806]
12. Alanazi AA, Nicholson N. Audiology and speech-language pathology simulation training on the 1-3-6 early hearing detection and intervention timeline. Am J Audiol 2019 Jun 10;28(2):348-361. [doi: 10.1044/2019_AJA-18-0185] [Medline: 31095403]
13. Andre A, Reed A, Ananthakrishnan S, Korczak P. An evaluation of simulation techniques in audiology and allied health professions. Am J Audiol 2021 Jun 14;30(2):295-308. [doi: 10.1044/2021_AJA-20-00190] [Medline: 33872515]
14. Alanazi AA, Mohamud MS, AlSuwailem SS. The effect of simulation learning on audiology and speech-language pathology students' self-confidence related to early hearing detection and intervention: a randomized experiment. Speech Lang Hear 2020 Nov 17;25(2):211-224. [doi: 10.1080/2050571X.2020.1846839]
15. Ozelie R, Panfil P, Swiderski N, Walz E. Hearing voices simulation: impact on occupational therapy students. Open J Occup Ther 2018 Oct 01;6(4):10. [doi: 10.15453/2168-6408.1452]
16. Toader E. Clinical simulations for learning medical skills: a work-based approach to simulators. Procedia Soc Behav Sci 2015 Jul;197:2443-2448. [doi: 10.1016/j.sbspro.2015.07.408]
17. Halm BM, Lee MT, Franke AA. Improving toxicology knowledge in preclinical medical students using high-fidelity patient simulators. Hawaii Med J 2011 Jun;70(6):112-115 [FREE Full text] [Medline: 22162607]
18. Otis – the virtual patient. Innoforce. URL: https://www.innoforce.com/en/products/otis-the-virtual-patient/ [accessed 2024-04-29]
19. Kocian A, Cattani G, Chessa S, Grolman W. An artificial patient for pure-tone audiometry. J Audio Speech Music Proc 2018 Jul 27;2018(1):1. [doi: 10.1186/S13636-018-0131-Y]
20. Calandruccio L, Weidman D. Online simulation education for audiometry training. Am J Audiol 2022 Mar 03;31(1):1-10. [doi: 10.1044/2021_AJA-21-00121] [Medline: 34958736]
21. Kompis M, Steffen P, Caversaccio M, Brugger U, Oesch I. A multilingual audiometer simulator software for training purposes. Acta Otolaryngol 2012 Apr;132(4):428-433. [doi: 10.3109/00016489.2011.636379] [Medline: 22235845]
22. Lieberth AK, Martin DR. The instructional effectiveness of a web-based audiometry simulator. J Am Acad Audiol 2005 Feb;16(2):79-84. [doi: 10.3766/jaaa.16.2.3] [Medline: 15807047]
23. Wilding T. Pure tone audiometry simulator. The University of Manchester. URL: https://personalpages.manchester.ac.uk/staff/tim.wilding/PTA_Sim/index.html [accessed 2024-04-29]
24. Audiometer simulator for pure-tone audiometry including masking. AudSim Flex. URL: https://audsim.com/) [accessed 2024-04-29]
25. Combination audiometer and tympanometry. Grason-Stadler Inc. URL: https://www.grason-stadler.com/products/tympanometers/gsi-39 [accessed 2024-04-29]

XSL•FO
RenderX

26. Peffers K, Tuunanen T, Rothenberger MA, Chatterjee S. A design science research methodology for information systems research. J Manag Inf Syst 2014 Dec 08;24(3):45-77. [doi: 10.2753/MIS0742-1222240302]
27. Drechsler A, Hevner AR. Utilizing, producing, and contributing design knowledge in DSR projects. In: Proceedings of the 13th International Conference on Designing for a Digital and Globalized World. 2018 Presented at: DESRIST '18; June 3-6, 2018; Chennai, India p. 82-97 URL: https://link.springer.com/chapter/10.1007/978-3-319-91800-6_6 [doi: 10.26686/wgtn.13088369]
28. Gerdes J, Schooley B, Sharp D. Design of a simulation tool for audiology education to provide hearing screening training. In: Proceedings of the 55th Hawaii International Conference on System Sciences. 2022 Presented at: HICSS '22; January 4-7, 2022; Virtual Event p. 3793-3802 URL: https://scholarspace.manoa.hawaii.edu/server/api/core/bitstreams/9c77643b-9269-4cc4-ad3c-7b085a841fcd/content [doi: 10.24251/hicss.2022.462]
29. Microsoft flight simulator, 40th anniversary edition. Flight Simulator. URL: https://www.flightsimulator.com/ [accessed 2024-04-29]
30. Gerdes J, Schooley B. Hearing screening simulator data. Open Science Framework. URL: https://osf.io/8s6rj/ [accessed 2024-04-29]

Abbreviations

HSS: hearing screening simulator

Edited by A Mavragani; submitted 10.03.23; peer-reviewed by L Coco, N Mungoli; comments to author 15.05.23; revised version received 28.07.23; accepted 03.09.24; published 20.02.25.

Please cite as:
Gerdes J, Schooley B, Sharp D, Miller J
The Design and Evaluation of a Simulation Tool for Audiology Screening Education: Design Science Approach
JMIR Form Res 2025;9:e47150
URL: https://formative.jmir.org/2025/1/e47150
doi:10.2196/47150
PMID:39977027

©John Gerdes, Benjamin Schooley, Dakota Sharp, Juliana Miller. Originally published in JMIR Formative Research (https://formative.jmir.org), 20.02.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effects of Missing Data on Heart Rate Variability Measured From A Smartwatch: Exploratory Observational Study

Hope Davis-Wilson[1], PhD; Meghan Hegarty-Craver[1], PhD; Pooja Gaur[1], PhD; Matthew Boyce[1], BS; Jonathan R Holt[1], PhD; Edward Preble[1], PhD; Randall Eckhoff[1], BS; Lei Li[1], PhD; Howard Walls[1], PhD; David Dausch[1], PhD; Dorota Temple[1], PhD

RTI International, Morrisville, NC, United States

Corresponding Author:
Hope Davis-Wilson, PhD
RTI International
3040 E Cornwallis Rd
Morrisville, NC, 27709
United States
Phone: 1 7043404816
Email: hdaviswilson@rti.org

Abstract

Background: Measuring heart rate variability (HRV) through wearable photoplethysmography sensors from smartwatches is gaining popularity for monitoring many health conditions. However, missing data caused by insufficient wear compliance or signal quality can degrade the performance of health metrics or algorithm calculations. Research is needed on how to best account for missing data and to assess the accuracy of metrics derived from photoplethysmography sensors.

Objective: This study aimed to evaluate the influence of missing data on HRV metrics collected from smartwatches both at rest and during activity in real-world settings and to evaluate HRV agreement and consistency between wearable photoplethysmography and gold-standard wearable electrocardiogram (ECG) sensors in real-world settings.

Methods: Healthy participants were outfitted with a smartwatch with a photoplethysmography sensor that collected high-resolution interbeat interval (IBI) data to wear continuously (day and night) for up to 6 months. New datasets were created with various amounts of missing data and then compared with the original (reference) datasets. 5-minute windows of each HRV metric (median IBI, SD of IBI values [STDRR], root-mean-square of the difference in successive IBI values [RMSDRR], low-frequency [LF] power, high-frequency [HF] power, and the ratio of LF to HF power [LF/HF]) were compared between the reference and the missing datasets (10%, 20%, 35%, and 60% missing data). HRV metrics calculated from the photoplethysmography sensor were compared with HRV metrics calculated from a chest-worn ECG sensor.

Results: At rest, median IBI remained stable until at least 60% of data degradation (P=.24), STDRR remained stable until at least 35% of data degradation (P=.02), and RMSDRR remained stable until at least 35% data degradation (P=.001). During the activity, STDRR remained stable until 20% data degradation (P=.02) while median IBI (P=.01) and RMSDRR P<.001) were unstable at 10% data degradation. LF (rest: P<.001; activity: P<.001), HF (rest: P<.001, activity: P<.001), and LF/HF (rest: P<.001, activity: P<.001) were unstable at 10% data degradation during rest and activity. Median IBI values calculated from photoplethysmography sensors had a moderate agreement (intraclass correlation coefficient [ICC]=0.585) and consistency (ICC=0.589) and LF had moderate consistency (ICC=0.545) with ECG sensors. Other HRV metrics demonstrated poor agreement (ICC=0.071-0.472).

Conclusions: This study describes a methodology for the extraction of HRV metrics from photoplethysmography sensor data that resulted in stable and valid metrics while using the least amount of available data. While smartwatches containing photoplethysmography sensors are valuable for remote monitoring of patients, future work is needed to identify best practices for using these sensors to evaluate HRV in medical settings.

(JMIR Form Res 2025;9:e53645) doi:10.2196/53645

XSL•FO
RenderX

KEYWORDS

plethysmography; electrocardiogram; missing data; smartwatch; wearable; ECG; photoplethysmography; PPG; mobile phone; heart rate; pilot study; photoplethysmography; detection; sensor; monitoring; health metric; measure; electrocardiogram; real-world settings; rest; physical activity; remote monitoring; medical setting; youth; adolescent; teen; teenager

Introduction

Remote patient monitoring using digitally transmitted health data through wearable physiological and activity sensors offers many benefits and is gaining acceptance in the medical community [1]. For example, applications of remote monitoring include acute conditions such as COVID-19 recovery [2] along with chronic conditions such as heart failure [3], chronic obstructive pulmonary disease [4], and diabetes [5]. Remote patient monitoring via wearable sensors can improve clinical outcomes, improve self-management of diseases, and increase patient engagement and satisfaction [6]. Many commercial off-the-shelf (COTS) wearable sensors provide data on important physiological information, particularly wrist-worn devices used to measure heart rate [7,8]. Wrist-worn devices most often use embedded photoplethysmography sensors that detect changes in light intensity on the skin surface due to changes in blood volume during the cardiac cycle to estimate heart rate [9]. While not all smartwatches are certified as medical devices currently, they offer a way of assessing the viability of photoplethysmography sensors, which are used in some cleared medical devices [9-11].

Heart rate variability (HRV) is an indicator of autonomic nervous system functioning and can be used to identify health conditions such as respiratory illness [12], multiple sclerosis [13], Parkinson disease [14], and traumatic brain injury [15]. While collecting wearable sensor data such as HRV can provide multiple benefits to patients and create a more holistic picture of health status, real-world collection of HRV can be difficult to assess due to missing data, differences in data processing techniques, and the quality of data collected [16]. Missing data due to compliance and low-quality signals is an issue when using wearable sensors to monitor HRV in patients in clinical or research settings [16]. Missing data is common in longitudinal research studies, and it is problematic because it can reduce statistical power and cause bias in the estimation of parameters [17]. This is of concern when attempting to identify health anomalies because identifying a change in health relies on consistent and accurate baseline HRV data [18]. As demand for remote patient monitoring grows, an important next step is to identify best practices for data acquisition, processing, and analysis of HRV data from wearable sensors.

In addition to the need for best practices regarding data missingness, more understanding is still needed for accuracy in wrist-worn photoplethysmography sensors compared with more accurate wearable devices. Wrist-worn sensors typically result in greater wear time compared with chest-worn devices; however, wrist-worn sensors are more prone to movement-related artifacts that can negatively influence HRV signal quality [6]. Overwhelmingly, real-world studies have opted to use wrist-worn photoplethysmography sensors to evaluate HRV over other more accurate methods such as chest-worn electrocardiogram (ECG) sensors, likely improving patient compliance in these studies. As a result, we must determine how to best assess data from wrist-worn photoplethysmography sensors, including accounting for missing data and assessing data validity compared with ECG HRV data. This study conducted long-term (months) continuous monitoring of high-resolution HRV data during daily life activities using a COTS smartwatch with a photoplethysmography sensor and activity (step count) data. The primary aim of this study was to evaluate the influence of missing data on HRV metrics collected from a photoplethysmography-based smartwatch both at rest and during physical activity in real-world settings. The secondary aim of this study was to evaluate HRV agreement and consistency between the wrist-worn photoplethysmography sensor and a chest-worn ECG heart rate monitor both at rest and during periods of physical activity and sleep.

Methods

Ethical Considerations

For this prospective observational study, wearable sensor data was collected from local community members. The protocol was approved by the Research Triangle Institute (RTI) international institutional review board (MOD00001413). Written consent was received from all participants for participation and data use. The consent process emphasized that participation was voluntary and that individuals were allowed to withdraw their consent for participation at any time. Personal identifying information was stored in a secured file and kept separate from data collected once enrolled in the study. All smartwatch and survey data were deidentified and labeled by a participant identifier for analysis. Participants were offered US $50 every 3 months as study compensation, and participants could earn up to US $100 if they were enrolled for the entire 6 months of the study duration.

Participants

Healthy volunteers aged 18 years and older were included in this study. Individuals who were pregnant or who had a pre-existing health condition were not excluded from participation if they were able to use the wearable sensor. Participants were enrolled for 3 months with the opportunity to re-enroll for up to 6 months. Demographic data were obtained and stored in Research Electronic Data Capture (REDCap). Recruitment and data collection took place between May 2022 and June 2023.

Device Protocol

Each participant was outfitted with a Garmin Fenix 6/6s smartwatch, which was worn on the nondominant wrist. We chose to use Garmin over other wrist-worn photoplethysmography sensors because its Health Companion Software Development Kit provides access to high-resolution interbeat interval (IBI) data collected by photoplethysmography sensors that allow for monitoring of HRV. Other COTS sensors

XSL•FO
RenderX

only provide summary and other processed data. The average heart rate derived from Garmin watches strongly correlates with clinical ECG heart rate data (r=0.96) [19]; however, the accuracy of HRV metrics derived from Garmin has yet to be studied. Garmin watches integrate 2 photoplethysmography sensors for monitoring HR and a 3-axis accelerometer for detecting movement in the x, y, and z directions. The watch firmware uses accelerometer signals to calculate step counts in 1-minute windows. We used the Companion Software Development Kit to access minimally processed IBI data and step counts from the participants. The SIGMA+ Health app (S+ app) was installed onto the participant's personal mobile device or a study-issued Samsung Galaxy 12 smartphone if participants preferred not to use their personal mobile device for the study. The Garmin smartwatch was paired through a Bluetooth low energy connection to the mobile device and data from the watch was ingested through the S+ app onto an RTI secure Amazon Web Services server in JavaScript Object Notation format for offline data cleaning, metric extraction and standardization, and analysis. The initial setup included giving each user the study Personal identification number and a unique participant ID that did not contain personally identifiable information, which allowed the S+ app to track the user's data from one or more devices through the duration of the study. Within the app, the participant could view the connectivity status of their Garmin Fenix watch and manually trigger a data sync if required. Participants completed a daily health survey administered within the S+ app, which asked questions related to illness symptoms, sleep quality, sleep duration, and any unusual events that may have occurred that day (ie, psychological stress, excessive exercise, etc). Results were stored in a REDCap database.

A subset of participants (n=5) wore a Polar H10 ECG heart rate monitor with a sampling frequency of 1000 Hz over their chest simultaneously with the Garmin smartwatch to assess agreement and consistency between HRV metrics. IBI values derived from the Polar H10 chest strap show strong agreement compared with clinical ECG recordings at rest (intraclass correlation coefficient [ICC]=0.95) and with light exercise (ICC=0.93) [20]. Participants were encouraged to wear the Polar H10 continuously (day and night) for 5 days and were instructed to only remove it during showers, baths, or swimming. Polar H10 data was retrieved through the Elite HRV mobile app, which was downloaded to the participant's personal smartphone before data collection [21]. Excel files containing raw IBI data were exported from the participant's phone and sent to an RTI email account. We did not get measures of step counts from the Polar H10 device because Elite HRV only collects HRV data.

Data Processing for IBI Data From Garmin and Polar Devices

Each IBI data point acquired through the Garmin smartwatch and S+ app had an associated timestamp with millisecond precision. Steps from the Garmin were reported at 60-second intervals and IBI values were reported with each detected beat. Each IBI data point acquired through the Polar H10 monitor and Elite HRV app was recorded with millisecond precision. IBI artifacts (IBI>1500 ms or IBI <300 ms) were flagged and removed. For Garmin data quality control, we evaluated the expected data fraction, calculated as a ratio of the number of incoming data points to the maximum number of data points expected during the monitoring period, to assess data missingness. For simplicity, we assumed the expected number of IBI data points would be 60 IBI values/minute. This would correlate with an average heart rate of 60 beats per minute, which is a typical heart rate for someone who is sitting or resting. The exact number of data points expected varied due to within- and between-person differences in average heart rate. The expected data fraction was multiplied by 100 to be represented as a percentage between 0 and 100. For analysis, we chose to evaluate Garmin datasets with at least 70% of expected data to ensure we were using high-quality datasets with minimal missing data points. The first approximately 2 weeks of data were selected for each participant with at least 70% expected data fraction. Although there is some variation in the number of IBI data points in each 5 min window (a higher heart rate results in more IBI data points within a window), the datasets with at least 70% of expected data that were used for analysis contained a mean of 250 (SD 126) IBI data points within each 5 min window.

Garmin and Polar data were processed into 5-minute windows with each point in the window formatted as (timestamp, value) and aligned based on real-time so that the windows start at 5-minute marks (eg, 12:00 AM, 12:05 AM). The following metrics were extracted in 5 min windows: time-domain metrics (median IBI, SD of IBI values [STDRR], root-mean-square of the difference in successive IBI values [RMSDRR[) and frequency-domain metrics (low-frequency [LF] power, high-frequency [HF] power, and ratio of LF to HF power [LF/HF]) [18,22]. To calculate HRV metrics in the frequency domain, we resampled IBI data to 5 Hz using the Piecewise Cubic Hermite Interpolating Polynomial method [23]. If there was a gap in the signal greater than 15s, the longest continuous segment in the 5-minute window was used. If there were more than 3 gaps of this length, we did not use this window. After interpolation, the signal was filtered to the low-frequency (LF: 004 to 0.15 Hz) and high-frequency regions (HF: 0.15 to 1.0 Hz; we increased the upper limit to 1.0 Hz vs the typically used 0.40 Hz to account for physical activity when breathing rates and heart rates increase [24,25]). We computed a value representing the LF and HF power for 60-second and 30-second windows, respectively, using the formula: value = ln (variance [data window]). Values less than 2.5 and more than 9.0 were eliminated, and the median values were calculated to represent the 5-minute window. For Garmin smartwatch step data, reported every minute, we calculated the average of values in the 5-minute window; we required at least 2 valid data points for the calculation.

Semisimulation of Garmin Data Degradation

A semisimulation design based on previous work [26] was used to create new datasets with various amounts of Garmin IBI data from original Garmin datasets. The premise of this semisimulation approach was to create new datasets with different amounts of missing data based on real wear patterns. Using this method, we ensured that we removed data in a realistic manner rather than randomly removing data. In this approach, we considered data missingness patterns from datasets with low expected data fractions and applied these patterns to

XSL•FO
RenderX

the datasets with high expected data fractions. These semisimulated datasets were then compared to the reference datasets to understand how varying amounts of data missingness influence HRV metrics. In total, 4 additional 5-minute window samples were randomly selected that contained fewer IBI data points. These 4 samples were inspected to identify data gap patterns and then used to model data missingness patterns in the selected datasets.

To recreate data missingness, we matched the data missingness pattern from the sample 5-minute windows to each 5-minute window in the reference datasets used for analysis. For example, one sample dataset with 120 IBI data points (missing 60% of data) was missing data in the first 3.15 min of the window. The first 3.15 min of each 5-min window was removed from the reference dataset to simulate data missingness. The simulated datasets were then compared with the reference datasets to identify differences in HRV metrics. Missing data simulations were performed using Python (Python Software Foundation) [27]. The various datasets were labeled for analysis, that consists of (1) a reference dataset with no missing data, (2) missing 10% of data, (3) missing 20% of data, (4) missing 35% of data, and (5) missing 60% of data. No simulated data degradation was performed for Polar H10 IBI data.

Statistical Analysis

Mean and SD were calculated for all descriptive variables and HRV metrics included in the analysis. For the primary aim, due to the large sample size of Garmin 5-minute windows for rest (48,442 windows) and light activity (19,976 windows), Shapiro-Wilk normality tests were not valid; therefore, we visually inspected each Garmin HRV metric for normality using Quantile-Quantile (Q-Q) plots. Q-Q plots that appeared as a roughly straight 45-degree line were considered normally distributed. Variables that were not normally distributed were transformed using a Box-Cox transformation before analysis. Q-Q plots of each HRV variable are shown at rest in Figure 1 and at light activity in Figure 2.

XSL•FO
RenderX

Figure 1. Quantile-Quantile plots for the median interbeat interval (IBI), SD of IBI values (STDRR), root-mean-square error of IBI values (RMSDRR), low frequency (LF), high frequency (HF), and the ratio of low to high frequency power at rest (steps/min=0). Box-Cox transformations were performed for STDRR and RMSDRR metrics. Smartwatch IBI data was collected and reported from 16 individuals recruited from a community setting between May 2022 and June 2023.

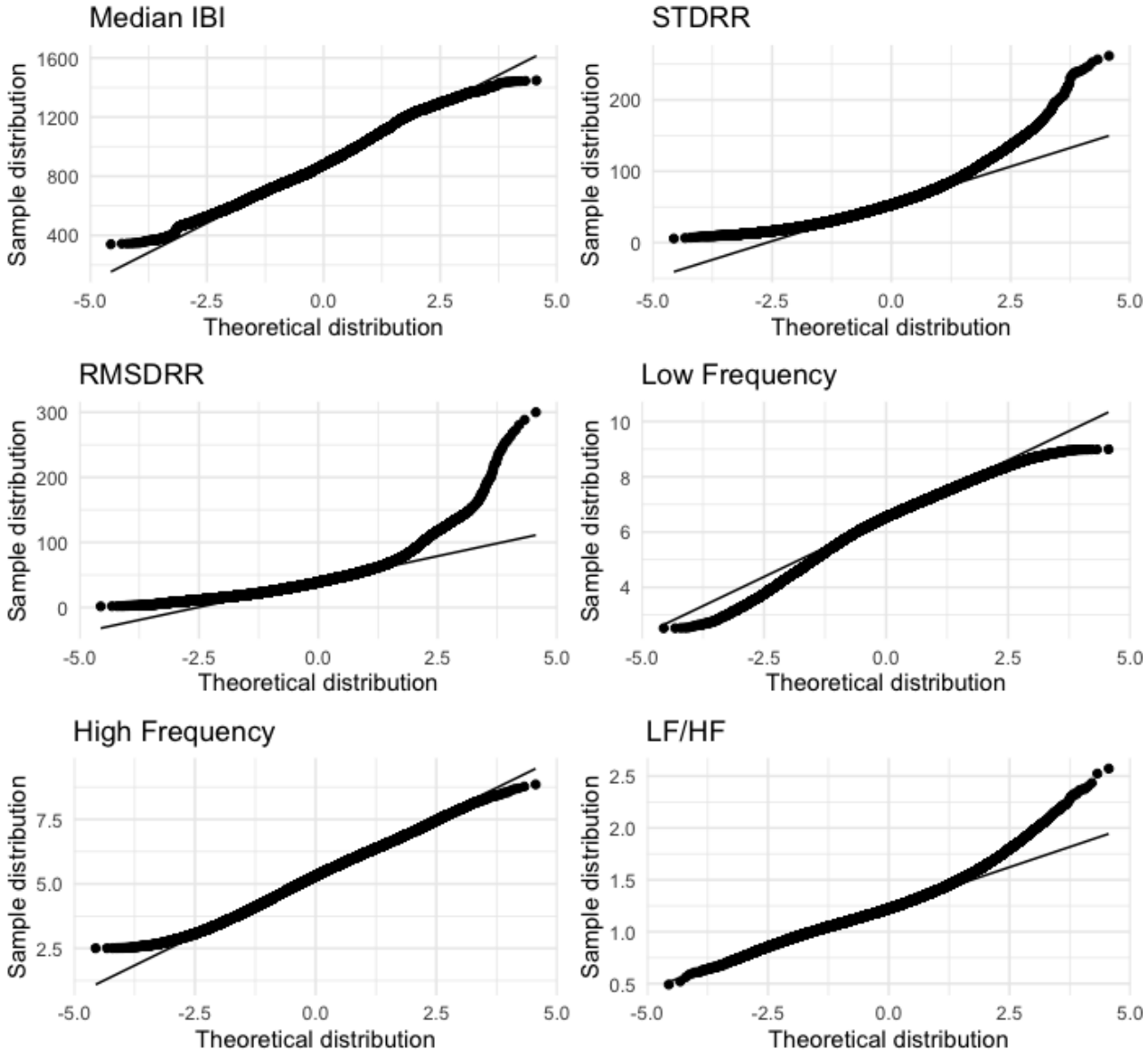

XSL•FO
RenderX

Figure 2. Quantile-Quantile plots for the median interbeat interval (IBI), SD of IBI values (STDRR), root-mean-square error of IBI values (RMSDRR), low frequency (LF), high frequency (HF), and the ratio of low to high frequency power at light activity (100>steps/min>0). Box-Cox transformations were performed for RMSDRR and LF/HF metrics. Smartwatch IBI data were collected and reported from 16 individuals recruited from a community setting between May 2022 and June 2023.

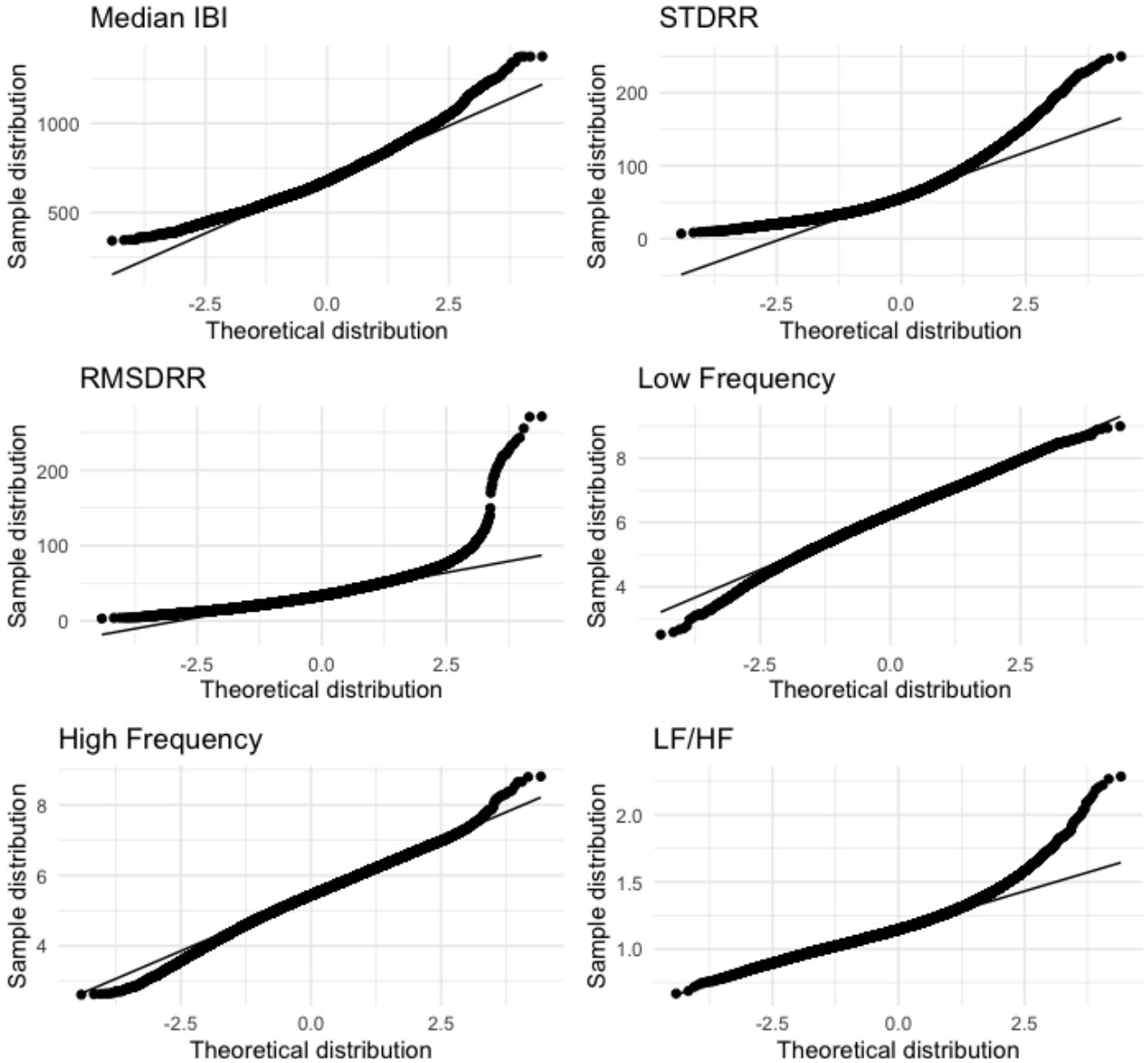

We then used linear mixed-effect models to estimate the mean differences in each Garmin HRV metric between the datasets with missing samples by level of missingness and the reference dataset. A participant-specific random intercept was included in the models to account for correlations of our sample data among multiple 5-minute windows within each participant [28]. An indicator of the level of data missingness (data without missing samples as reference, data with 10%, 20%, 35%, or 60% samples removed) was included in the models as a fixed-effect covariate. The lme4 package in the statistical environment R (R Core Team) was used for the model estimations [29]. The statistical significance of the estimated differences was tested using Satterthwaite's method with a 2-tailed significance level of 0.05, and the 95% CI were obtained using the bootstrap method through the bootMer function of the lme4 package in R (1000 simulations). Separate linear mixed models were estimated using sample data under 2 different conditions: at rest (steps/min=0) and during light physical activity (steps/min of >0 and <100).

For the secondary aim, we assessed the reliability of Garmin-derived HRV metrics within each participant and their validity compared to Polar-derived HRV metrics. The intraclass correlation coefficients (class 2, mean rating; $ICC_{2,k}$) for consistency and absolute agreement with 95% CI were obtained using the irr package in the statistical environment R [30]. The magnitude of the ICC was generally interpreted as $ICC_{2,k}$ of 0.5-0.75 as "moderate", $ICC_{2,k}$ >0.75 as "good", and $ICC_{2,k}$ >0.9 as "excellent" [31]. In addition, we visually assessed agreement between Garmin-derived HRV metrics and polar-derived HRV metrics using Bland-Altman analyses to evaluate the mean differences and 95% limits of agreement [32]. R version 4.2.2 [33] was used for statistical analyses.

XSL•FO
RenderX

Results

Overview

Garmin data were collected from 31 participants for up to 180 days. The analyzed sample in aim 1 used a mean of 16 (SD 3) consecutive days of data from 16 participants resulting in 48,442 five-minute windows of data that represented rest (steps/min=0) and 19,976 windows of data that represented light activity (100 >steps/min >0). Although there is not a clear precedence for what percentage of missing data is acceptable, we wanted the reference datasets to contain a minimal amount of missing data. Therefore, each participant included in the analysis had an expected data fraction of at least 70% (including daytime and nighttime), and participants who did not meet this requirement (n=15) were not included in the analysis. Differences in days of data occurred because of between-person variance in unprocessed IBI data and differences in wear patterns. For aim 2, 5 participants wore both the Garmin Fenix watch and Polar heart rate monitor for 5 consecutive days. Demographic data is shown in Table 1. The averages of each HRV metric for aims 1 and 2 are shown in Table 2 and Table 3. Box-Cox transformations were performed for STDRR and RMSDRR metrics at rest (steps/min=0; Figure 1) and RMSDRR and LF/HF at light activity (100>steps/min>0; Figure 2). Tables 4 and 5 show results from the linear mixed models, specifically the estimated error between the reference and missing data, the SE, and the 95% CIs of the estimated difference.

Table 1. Participant characteristics of a community sample of individuals who provided smartwatch data between May 2022 and June 2023. In total, 31 individuals were recruited, data from 16 individuals were used in aim 1, and data from 5 individuals were used in aim 2.

	Total sample (n=31), mean (SD)	Aim 1 (n=16), mean (SD)	Aim 2 (n=5), mean (SD)
Age (years)	38.6 (10.7)	38.6 (13.3)	35.2 (4.1)
Sex (females), n (%)	16 (53)	13 (81)	3 (60)
Height (m)	1.73 (0.12)	1.77 (0.12)	1.77 (0.08)
Mass (kg)	83.4 (17.2)	87.1 (16.5)	95.4 (24.7)
BMI (kg/m^2)	27.9 (4.7)	27.7 (4.1)	30.2 (5.6)

Table 2. Heart rate variability metrics for aim 1 datasets in a community sample of 16 individuals who provided smartwatch data between May 2022 and June 2023.

	Reference Garmin data	10% Missing Garmin data	20% Missing Garmin data	35% Missing Garmin data	60% Missing Garmin data
Resting (steps/min=0)					
Median interbeat interval	888.86	888.17	887.88	888.45	887.85
SD of interbeat interval values	57.41	57.44	57.44	57.14	52.87
Root-mean-square of the difference in successive interbeat interval values	41.70	41.83	41.86	41.53	40.98
Low-frequency power	6.52	6.38	6.31	6.50	6.50
High-frequency power	5.51	5.01	4.83	5.50	5.51
Ratio of low-frequency to high-frequency power	1.19	1.29	1.32	1.20	1.20
Light activity (0<steps/min<100)					
Median interbeat interval	690.70	688.02	687.19	692.65	695.30
SD of interbeat interval values	63.90	63.55	63.39	63.96	50.31
Root-mean-square of the difference in successive interbeat interval values	35.20	35.72	35.87	35.11	34.57
Low-frequency power	6.28	6.22	6.19	6.26	6.26
High-frequency power	5.59	5.26	5.13	5.55	5.55
Ratio of low-frequency to high-frequency power	1.13	1.19	1.21	1.13	1.14

XSL•FO
RenderX

Table 3. Heart rate variability metrics for aim 2 datasets in a community sample of 5 individuals who provided smartwatch data between May 2022 and June 2023.

	Reference Garmin data	10% Missing Garmin data	20% Missing Garmin data	35% Missing Garmin data	60% Missing Garmin data
Resting (steps/min=0)					
Median interbeat interval	888.86	888.17	887.88	888.45	887.85
SD of interbeat interval values	57.41	57.44	57.44	57.14	52.87
Root-mean-square error	41.70	41.83	41.86	41.53	40.98
Low-frequency power	6.52	6.38	6.31	6.50	6.50
High-frequency power	5.51	5.01	4.83	5.50	5.51
Ratio of low-frequency to high-frequency power	1.19	1.29	1.32	1.20	1.20
Light activity (0<steps/min<100)					
Median interbeat interval	690.70	688.02	687.19	692.65	695.30
SD of interbeat interval values	63.90	63.55	63.39	63.96	50.31
Root-mean-square error	35.20	35.72	35.87	35.11	34.57
Low-frequency power	6.28	6.22	6.19	6.26	6.26
High-frequency power	5.59	5.26	5.13	5.55	5.55
Ratio of low-frequency to high-frequency power	1.13	1.19	1.21	1.13	1.14

XSL•FO
RenderX

Table 4. Linear mixed-effects model results of the effects of missing raw data on heart rate variability metrics collected at rest (steps/min=0) in a community sample of 16 individuals who provided smartwatch data between May 2022 and June 2023.

Heart rate variability metric and level of missing data compared with reference	Estimated difference between reference data and missing data	SE	Lower limit of the 95% CI	Upper limit of the 95% CI	*P* value
Median interbeat interval					
10%	–0.739	0.939	–2.643	1.139	.43
20%	–1.04	0.939	–3.019	0.978	.27
35%	–0.468	0.939	–2.341	1.374	.62
60%	–1.095	0.939	–2.880	0.928	.24
SD of interbeat interval values					
10%	–0.002	0.006	–0.014	0.010	.69
20%	–0.003	0.006	–0.015	0.008	.59
35%	–0.020	0.006	–0.031	–0.008	.001[a]
60%	–0.215	0.006	–0.228	–0.204	<.001[a]
Root-mean-square					
10%	0.005	0.004	–0.003	0.013	.27
20%	0.005	0.004	–0.003	0.014	.20
35%	-0.014	0.004	–0.021	–0.005	.001[a]
60%	-0.053	0.004	–0.061	–0.044	<.001[a]
Low-frequency power					
10%	-0.134	0.006	–0.145	–0.122	<.001[a]
20%	-0.204	0.006	–0.215	–0.193	<.001[a]
35%	-0.016	0.006	–0.028	–0.005	.005[a]
60%	-0.016	0.006	–0.028	–0.005	.004[a]
High-frequency power					
10%	-0.501	0.005	–0.512	–0.490	<.001[a]
20%	-0.678	0.005	–0.689	–0.667	<.001[a]
35%	-0.013	0.005	–0.024	–0.002	.02[a]
60%	-0.006	0.005	–0.016	0.005	.27
Ratio of low frequency to high frequency					
10%	0.096	0.001	0.094	0.098	<.001[a]
20%	0.130	0.001	0.128	0.132	<.001[a]
35%	0.005	0.001	0.003	0.007	<.001[a]
60%	0.004	0.001	0.002	0.006	<.001[a]

[a]*P* value <.05 indicates a significant difference compared with the reference dataset.

XSL•FO
RenderX

Table 5. Linear mixed-effects model results of the effects of missing raw data on heart rate variability metrics collected at light activity (100>steps/min>0) in a community sample of 16 individuals who provided smartwatch data between May 2022 and June 2023.

Heart rate variability metric and level of missing data compared with reference	Estimated difference between reference data and missing data	SE	Lower limit of the 95% CI	Upper limit of the 95% CI	*P* value
Median interbeat interval					
10%	–2.645	1.049	–4.638	–0.627	.01[a]
20%	–3.473	1.049	–5.454	–1.481	<.001[a]
35%	1.998	1.049	–0.051	3.958	.06
60%	4.631	1.050	2.513	6.577	<.001[a]
SD of interbeat interval values					
10%	–0.005	0.003	–0.011	0.002	.11
20%	–0.008	0.003	–0.014	–0.001	.02[a]
35%	–0.006	0.003	–0.012	0.001	.08
60%	–0.217	0.003	–0.223	–0.210	<.001[a]
Root-mean-square					
10%	0.039	0.007	0.026	0.052	<.001[a]
20%	0.050	0.007	0.037	0.064	<.001[a]
35%	–0.013	0.007	–0.025	–.001	.05
60%	–0.069	0.007	–0.081	–0.055	<.001[a]
Low-frequency power					
10%	–0.054	0.006	–0.067	–0.042	<.001[a]
20%	–0.087	0.006	–0.100	–0.075	<.001[a]
35%	–0.027	0.006	–0.040	–0.015	<.001[a]
60%	–0.019	0.006	–0.031	–0.006	.003[a]
High-frequency power					
10%	–0.324	0.006	–0.336	–0.313	<.001[a]
20%	–0.452	0.006	–0.464	–0.441	<.001[a]
35%	–0.033	0.006	–0.045	–0.021	<.001[a]
60%	–0.038	0.006	–0.050	–0.027	<.001[a]
Ratio of low frequency to high frequency					
10%	0.043	0.004	0.042	0.045	<.001[a]
20%	0.059	0.001	0.058	0.061	<.001[a]
35%	–0.001	0.001	–0.003	0.0004	0.12
60%	–0.001	0.001	–0.001	0.002	0.33

[a]*P* value <0.05 indicates a significant difference compared to the reference dataset.

Aim 1: Heart Rate Variability Metrics at Rest

At rest (steps/min=0), there was no difference in median IBI values between reference and simulated data (*P*>.05; Table 4 and Figure 3). There were significant differences in STDRR and RMSDRR values between the reference and 35% and 60% missing data (*P*<.05; Table 4, Figure 3). LF and LF/HF metrics demonstrated significant differences between the reference and all simulated datasets (*P*<.05; Table 4, Figure 3). HF in the reference dataset was significantly higher compared to the 10%, 20%, and 35% missing datasets (*P*<.05; Table 4, Figure 3). There was no significant difference between the reference dataset and the 60% missing dataset (*P*>.05; Table 4 and Figure 3).

XSL•FO
RenderX

Figure 3. Heart rate variability metrics at rest (steps/min=0). HRV: heart rate variability; IBI: interbeat interval; LF/HF: ratio of low frequency power to high frequency power; RMSDRR: root-mean-square of the difference in successive IBI values; STDRR: standard deviation of interbeat interval values.

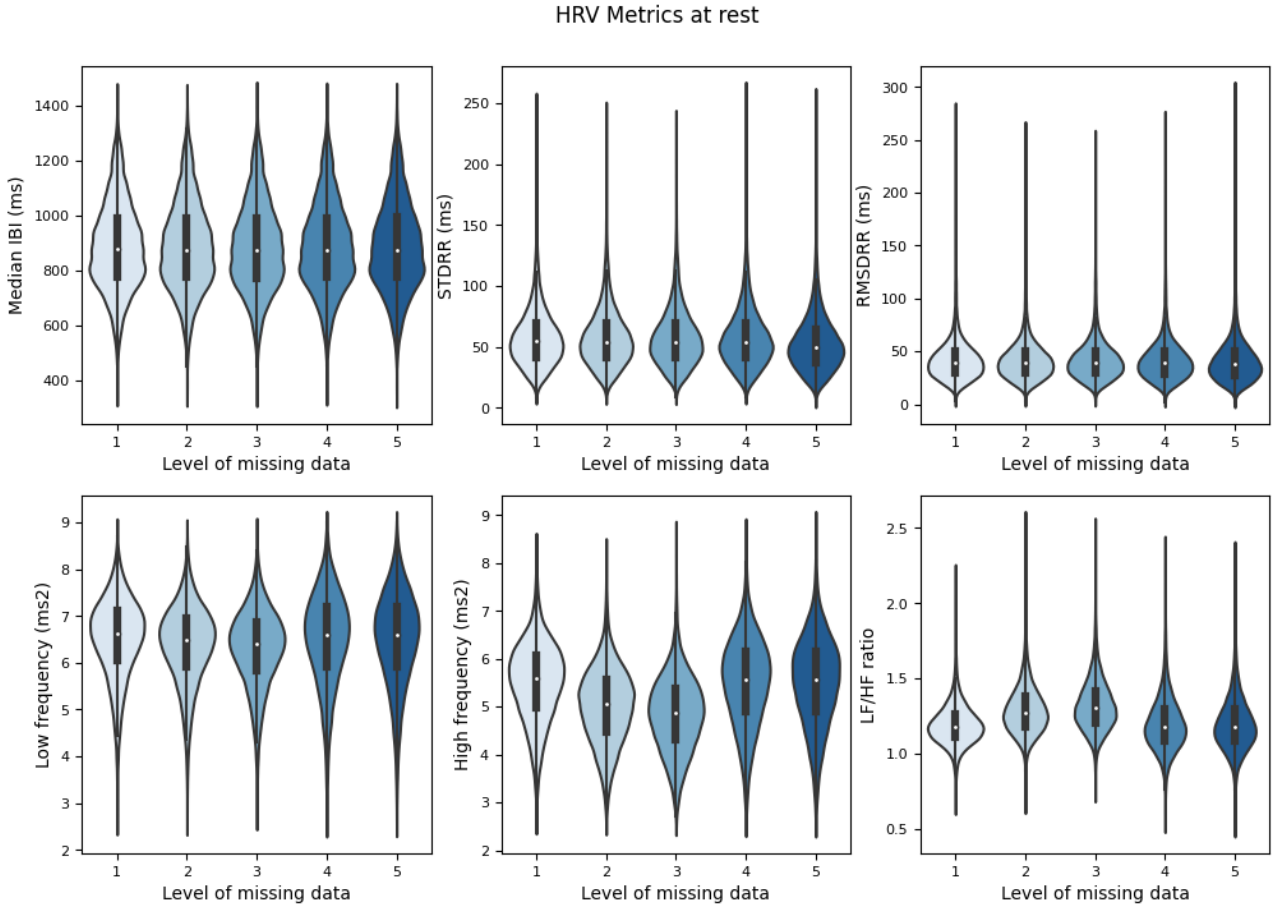

In Figure 3, reference smartwatch data were compared to (1) 10% missing data (2), 20% missing data (3), 35% missing data (4), and 60% missing data (5) per 5-minute window. Smartwatch IBI data were collected and reported from 16 individuals recruited from a community setting between May 2022 and June 2023.

Aim 1: Heart Rate Variability Metrics at Light Activity

At light activity (100>steps/min>0), there were significant differences in median IBI between the reference and all simulated datasets (P<.05; Table 5 and Figure 4). STDRR and RMSDRR values were different between the reference and simulated datasets with 10%, 20%, and 60% missing data (P<.05; Table 5 and Figure 4). Differences between the reference dataset and 35% missing data were not statistically significant, although results trended toward statistical significance (P=.06). LF and HF metrics demonstrated significant differences between the reference dataset and all simulated datasets (P<.05; Table 5 and Figure 4). LF/HF in the reference dataset was significantly higher compared with the 10% and 20% missing datasets (P<.05; Table 5 and Figure 4). There was no significant difference between the reference dataset and the 35% and 60% missing datasets (P>.05; Table 5 and Figure 4).

XSL•FO
RenderX

Figure 4. Heart rate variability metrics at light activity (100>steps/min>0). HRV: heart rate variability; IBI: interbeat interval; LF/HF: ratio of low frequency power to high frequency power; RMSDRR: root-mean-square of the difference in successive IBI values; STDRR: standard deviation of interbeat interval values.

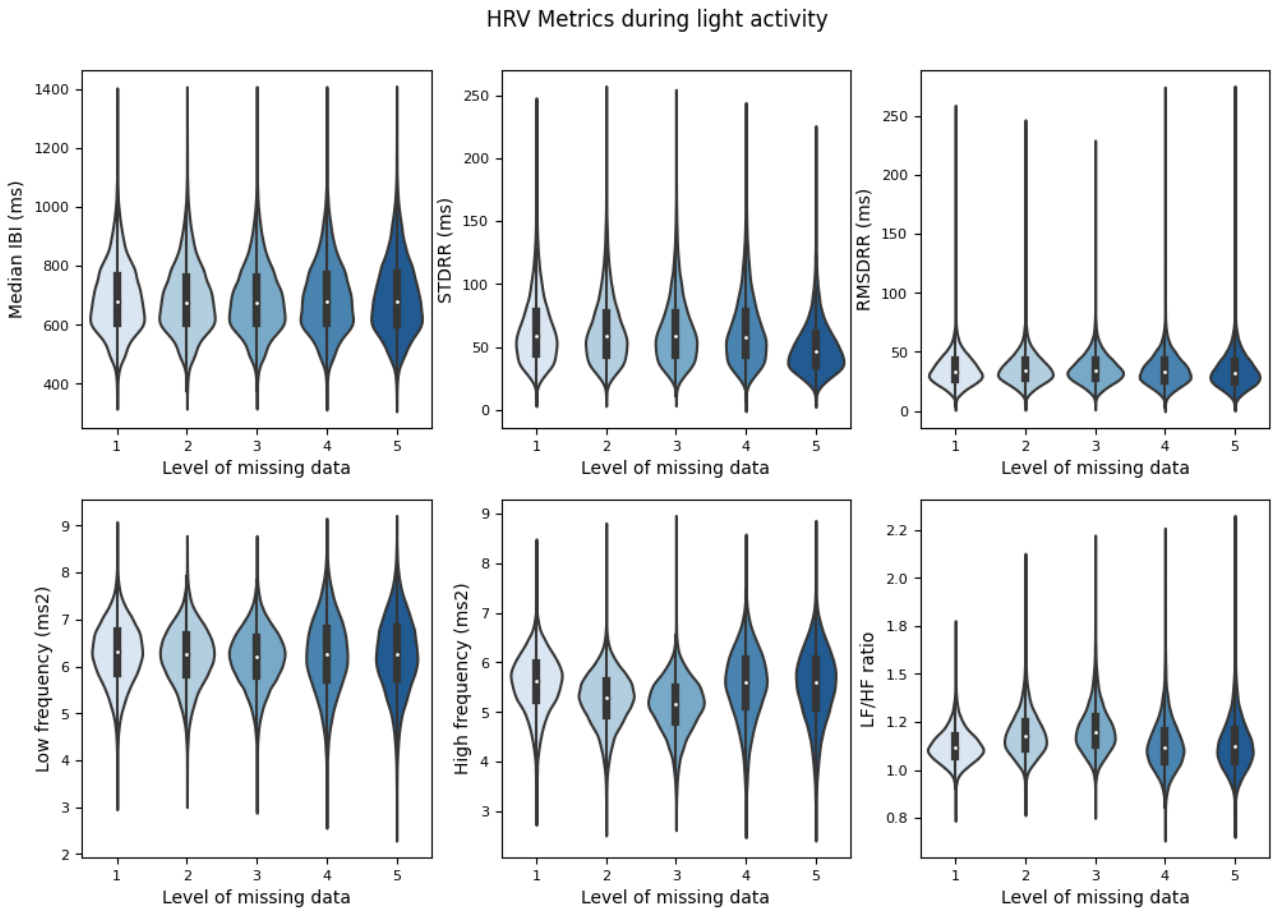

In Figure 4, Reference smartwatch data were compared to (1) 10% missing data (2), 20% missing data (3), 35% missing data (4), and 60% missing data (5) per 5 min window. Smartwatch IBI data was collected and reported from 16 individuals recruited from a community setting between May 2022 and June 2023.

Aim 2: Validity and Agreement of Photoplethysmography-Derived Heart Rate Variability Compared With ECG-Derived Heart Rate Variability

Median IBI values calculated from Garmin photoplethysmography sensors had moderate agreement (ICC=0.585) and consistency (ICC=0.589) with Polar ECG data, and LF calculated from Garmin photoplethysmography sensors had moderate consistency (ICC=0.545) with Polar ECG data. All other HRV metrics (STDRR, RMSDRR, LF, HF, and LF/HF) demonstrated poor agreement between Garmin photoplethysmography and Polar ECG data (ICC=0.071-0.470, Table 6). Bland-Altman visual analysis indicated systematic bias present in all HRV metrics, with particularly notable bias in RMSDRR, HF, and LF/HF (Figure 5). The root-mean-square error between Garmin and Polar devices for each HRV metric are listed in Table 6.

XSL•FO
RenderX

Figure 5. Bland-Altman plots for differences in heart rate variability metrics measured with a Garmin smartwatch and Polar heart rate monitor. HRV data was collected and reported from 5 individuals recruited from a community setting between May 2022 and June 2023. HRV: heart rate variability; IBI: interbeat interval; LF/HF: ratio of low frequency power to high frequency power.

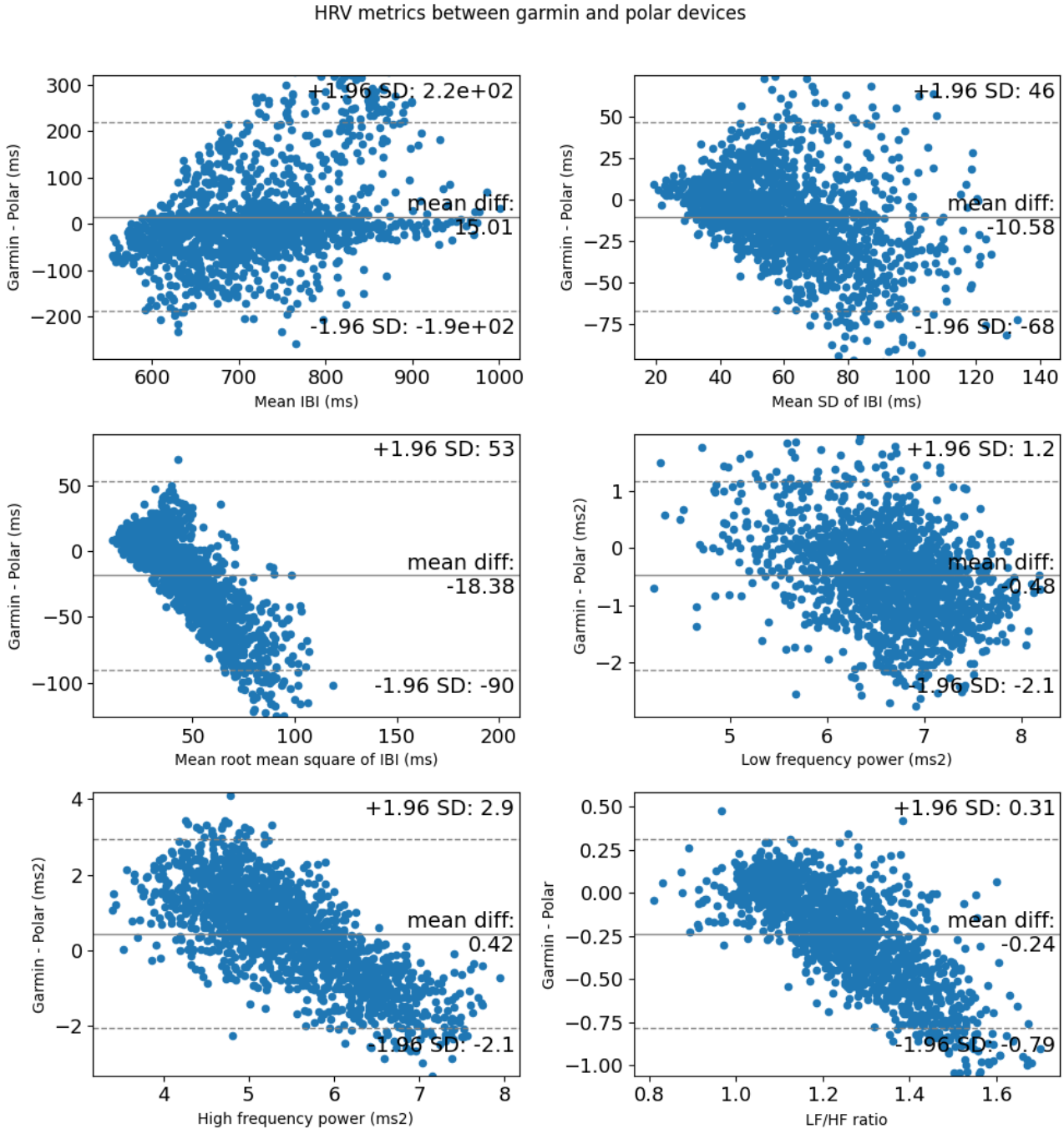

XSL•FO
RenderX

Table 6. Agreement and consistency between heart rate variability metrics were measured with a Garmin smartwatch and Polar heart rate monitor in a community sample of 5 individuals who provided data between May 2022 and June 2023.

Heart rate variability metric	ICC[a] agreement, mean (95% CI)	ICC consistency, mean (95% CI)	Root-mean-square error
Median interbeat interval	0.585 (0.538 to 0.626)	0.589 (0.547 to 0.628)	105.23
SD of interbeat interval values	0.441 (0.302 to 0.545)	0.472 (0.417 to 0.521)	30.92
Root mean square of the difference in successive interbeat interval values	0.214 (0.028 to 0.357)	0.255 (0.178 to 0.324)	40.98
Low-frequency power	0.473 (0.163 to 0.645)	0.545 (0.497 to 0.588)	0.97
High-frequency power	0.404 (0.283 to 0.499)	0.429 (0.370 to 0.483)	1.35
Low-frequency to high-frequency power ratio	0.071 (–0.09 to 0.207)	0.117 (0.024 to 0.2)	0.37

[a]ICC: intraclass correlation coefficient.

Discussion

Principal Findings

To our knowledge, this paper shows for the first time the results of long-term (months) continuous monitoring of high-resolution HRV data during daily life activities using a COTS smartwatch with photoplethysmography sensor and activity (step count) data. This study describes a methodology for the extraction of HRV metrics from IBI time-series data that resulted in stable and valid metrics while using the least amount of available data. By understanding patterns of missing data, we can maximize the amount of usable data and minimize the impact of data gaps due to suboptimal wear compliance or any issues in the data synchronization between the watch and the smartphone. We found that time-domain HRV metrics (median IBI, STDRR, and RMSDRR) are more resilient to missing data when the participant is in a resting state; however, during light activity missing data influence time-domain HRV metrics. Median IBI measurements remain stable at rest even at 60% data degradation. STDRR and RMSDRR measurements remain stable at rest until 35% data degradation. Frequency-domain HRV metrics (LF, HF, and LF/HF) are less resilient to missing data both at rest and during light activity and are unstable even at 10% data degradation. It is unclear why HF (during rest) was not significantly different at 60% data degradation and LF/HF (during light activity) was not significantly at 35% and 60% data degradation compared to reference data. We speculate that there may not be enough data at these levels of data degradation for accurate comparison. Results from this study suggest that analyses and algorithms that use primarily frequency-domain HRV metrics are more sensitive to sparse data collection. Researchers and engineers should carefully evaluate HRV metrics in situations where data are sparse. If missing data are an issue in a research study, it may be beneficial to rely on metrics and algorithms using time-domain HRV metrics.

A few previous studies have simulated missing data patterns in HRV data by evaluating differences between scattered missing beats (one or two missing data points dispersed throughout the dataset) and bursts of missing beats (longer periods of missing data likely due to not wearing the device) [34,35]. Our study considered data characteristics from real HRV data patterns to remove data instead of using a random data removal approach. We found that datasets with very sparse data (missing 35% and 60% of data points per 5-minute window) typically had one burst of missing data. For example, we identified a 3.15 min data gap (60%) during the beginning of one 5-minute window. We also identified missing data between 1 minute and 2.75 minutes (35%) of another 5-minute window. The 20% and 10% missing datasets were missing 0.80 minutes and 0.63 minutes, respectively, of data scattered randomly throughout the trial, most often in 1-2 second increments. In general, 5 min windows that had data gaps that were smaller and more randomly dispersed had a greater number of IBI datapoints, whereas 5-minute windows that had one larger burst of missing data, often at the beginning of the 5-minute window, had fewer IBI datapoints. We hypothesize that these larger bursts of missing data may be the result of noncompliance (ie, not wearing the Garmin watch), and the smaller data gaps may be due to movement artifacts or device malfunction. Scattered missing data due to movement artifacts is typically considered data missing at random or missing completely at random [16]. Data missing at random or missing completely at random do not necessarily produce a bias in outcomes [17]. Bursts of missing data that occur at similar times each day or routinely for a particular purpose (ie, bathing) may be considered data missing not at random, which could lead to biases in HRV outcomes [16]. Further understanding of the type of missing data within a 5-minute window is needed in future work.

Our secondary objective was to report validity and agreement between the Garmin photoplethysmography wrist-worn sensor compared with a chest-worn ECG sensor (Polar), which is closer to a gold standard that would be used in a laboratory setting to evaluate HRV [36,37] and has been used in the field for HRV biofeedback [38]. ICCs for agreement and consistency were mostly poor between Garmin and Polar devices. ICCs for consistency were moderate for median IBI and LF, suggesting these HRV metrics may be more reliable to measure in a real-world setting. Bland-Altman plots showed systematic bias in RMSDRR, HF, and LF/HF metrics. Specifically, Garmin tended to overreport lower values and underreport higher values, indicating that Garmin generally produced a smaller range of RMSDRR, HF, and LF/HF values compared to Polar. Our study compared processed HRV metrics (beyond a simple averaged heart rate calculation) between devices under free-living conditions. Previous studies have shown higher validity and agreement between wrist-worn photoplethysmography sensors

XSL•FO
RenderX

and Polar ECG average heart rate [36,37]; however, these studies performed protocols in controlled, laboratory settings. Results from this study highlight that agreement and consistency between photoplethysmography wrist-worn sensors and ECG sensors are lower in free-living conditions and should be considered when evaluating HRV metrics. Commercially available wearable technology continues to grow at a rapid pace, and sensors that provide more accurate data are needed. Future studies should evaluate the accuracy of HRV data in emerging sensors.

Our study is the first to our knowledge to address the effects of missing data using real examples of HRV data patterns collected with photoplethysmography sensors; however, there are some limitations to our current work. In our current sample, we have limited labeling of HRV data. Future studies should include daily self-reported activity logs to determine when participants are sleeping, exercising, or not wearing their watch. We were unable to evaluate data missingness during levels of high activity (steps/min>100) because we did not have enough 5-minute windows with high step counts. It is best to evaluate high levels of activity in smaller time windows since many people engage in high activity for very short durations; however, a minimum window of 5 minutes is needed to accurately calculate frequency-domain HRV metrics [39]. In future work, we will aim to recruit individuals who are engaging in high levels of activity. In addition, we recruited healthy individuals for this study, but we did not exclude individuals with potential confounders that could have affected HRV data. For example, some participants may have had infrequent irregular heart rhythms (ie, heart skipping a beat) or may have taken medications that could have affected HRV values. Finally, our small sample size when comparing Garmin to Polar HRV data limits the generalizations we can make regarding Garmin's validity in real-world settings.

Conclusion

Findings from this study have important implications for best practices for photoplethysmography wearables use in clinical and research settings. Time-domain HRV metrics (median IBI, STDRR, and RMSDRR) collected in a resting state remain stable until at least 35% data degradation. Frequency-domain HRV metrics (LF, HF, and LF/HF) are less resilient to missing data both at rest and during light activity and are unstable even at 10% data degradation. Studies that allow for greater than 10% data degradation for frequency-domain HRV metrics may introduce bias into their estimates, specifically the underestimation of LF and HF values. Correction methods such as gap-filling are possible to replace missing data; however, differences in HRV outcomes will still exist compared with more complete datasets without data loss [34]. In general, most derived HRV metrics from the photoplethysmography-based sensors used demonstrated moderate to poor agreement and consistency with ECG-based sensors in a real-world setting, and only median IBI values had reasonable agreement and consistency between the 2 sensing modalities. Despite the differences between photoplethysmography and ECG, we may conclude that given the stability of time-domain HRV metrics for up to 35% data degradation, these metrics could result in more reliable calculations of health metrics compared with frequency-domain metrics when monitoring patients using wrist-worn photoplethysmography sensors. In conclusion, while photoplethysmography sensors are valuable for remote monitoring of patients, future work is needed to identify best practices and the most accurate HRV metrics when using photoplethysmography sensors to evaluate HRV in medical settings.

Acknowledgments

This research was developed with funding from the Defense Advanced Research Projects Agency (contract 140D63-19-C-0023). The views, opinions, and findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the US Government. This paper is approved for public release, and the distribution is unlimited. Artificial intelligence was not used in any portion of writing this manuscript.

Data Availability

The data generated during and/or analyzed during this study are not publicly available due to restrictions of the Defense Advanced Research Projects Agency (contract 140D63-19-C-0023) but are available from the corresponding author on reasonable request.

Authors' Contributions

DST, DED, and HJW conceptualized the study. DST and MHC developed the study methodology. EAP, JRH, MDB, and RPE developed the software and mobile apps. DED performed project administration. MHC, PG, and HDW managed the study cohort. HDW and LL performed the statistical analyses. HDW wrote the original manuscript. All authors edited and reviewed the manuscript and approved the final version.

Conflicts of Interest

None declared.

References

XSL•FO
RenderX

1. de Farias FAC, Dagostini CM, Bicca YDA, Falavigna VF, Falavigna A. Remote patient monitoring: a systematic review. Telemed J E Health 2020;26(5):576-583. [doi: 10.1089/tmj.2019.0066] [Medline: 31314689]
2. Aalam AA, Hood C, Donelan C, Rutenberg A, Kane EM, Sikka N. Remote patient monitoring for ED discharges in the COVID-19 pandemic. Emerg Med J 2021;38(3):229-231. [doi: 10.1136/emermed-2020-210022] [Medline: 33472870]
3. Emani S. Remote monitoring to reduce heart failure readmissions. Curr Heart Fail Rep 2017;14(1):40-47. [doi: 10.1007/s11897-017-0315-2] [Medline: 28215030]
4. Cruz J, Brooks D, Marques A. Home telemonitoring effectiveness in COPD: a systematic review. Int J Clin Pract 2014;68(3):369-378. [doi: 10.1111/ijcp.12345] [Medline: 24472009]
5. Kirkland EB, Marsden J, Zhang J, Schumann SO, Bian J, Mauldin P, et al. Remote patient monitoring sustains reductions of hemoglobin A1c in underserved patients to 12 months. Prim Care Diabetes 2021;15(3):459-463 [FREE Full text] [doi: 10.1016/j.pcd.2021.01.005] [Medline: 33509728]
6. Kruklitis R, Miller M, Valeriano L, Shine S, Opstbaum N, Chestnut V. Applications of remote patient monitoring. Prim Care 2022;49(4):543-555. [doi: 10.1016/j.pop.2022.05.005] [Medline: 36357060]
7. Lubitz SA, Faranesh AZ, Selvaggi C, Atlas SJ, McManus DD, Singer DE, et al. Detection of atrial fibrillation in a large population using wearable devices: the fitbit heart study. Circulation 2022;146(19):1415-1424. [doi: 10.1161/CIRCULATIONAHA.122.060291] [Medline: 36148649]
8. Ray D, Collins T, Woolley S, Ponnapalli P. A review of wearable multi-wavelength photoplethysmography. IEEE Rev Biomed Eng 2023;16:136-151. [doi: 10.1109/RBME.2021.3121476] [Medline: 34669577]
9. Baman JR, Mathew DT, Jiang M, Passman RS. Mobile health for arrhythmia diagnosis and management. J Gen Intern Med 2022;37(1):188-197 [FREE Full text] [doi: 10.1007/s11606-021-07007-w] [Medline: 34282532]
10. Tison GH, Sanchez JM, Ballinger B, Singh A, Olgin JE, Pletcher MJ, et al. Passive detection of atrial fibrillation using a commercially available smartwatch. JAMA Cardiol 2018;3(5):409-416 [FREE Full text] [doi: 10.1001/jamacardio.2018.0136] [Medline: 29562087]
11. Ding EY, Han D, Whitcomb C, Bashar SK, Adaramola O, Soni A, et al. Accuracy and usability of a novel algorithm for detection of irregular pulse using a smartwatch among older adults: observational study. JMIR Cardio 2019;3(1):e13850 [FREE Full text] [doi: 10.2196/13850] [Medline: 31758787]
12. Alqahtani JS, Aldhahir AM, Alghamdi SM, Al Ghamdi SS, AlDraiwiesh IA, Alsulayyim AS, et al. A systematic review and meta-analysis of heart rate variability in COPD. Front Cardiovasc Med 2023;10:1070327 [FREE Full text] [doi: 10.3389/fcvm.2023.1070327] [Medline: 36873414]
13. Damla O, Altug C, Pinar KK, Alper K, Dilek IG, Kadriye A. Heart rate variability analysis in patients with multiple sclerosis. Mult Scler Relat Disord 2018;24:64-68. [doi: 10.1016/j.msard.2018.06.012] [Medline: 29957350]
14. Heimrich KG, Lehmann T, Schlattmann P, Prell T. Heart rate variability analyses in Parkinson's disease: a systematic review and meta-analysis. Brain Sci 2021;11(8):959 [FREE Full text] [doi: 10.3390/brainsci11080959] [Medline: 34439578]
15. Esterov D, Greenwald BD. Autonomic dysfunction after mild traumatic brain injury. Brain Sci 2017;7(8):1-8. [doi: 10.3390/brainsci7080100] [Medline: 28800081]
16. Di J, Demanuele C, Kettermann A, Karahanoglu FI, Cappelleri JC, Potter A, et al. Considerations to address missing data when deriving clinical trial endpoints from digital health technologies. Contemp Clin Trials 2022;113:106661 [FREE Full text] [doi: 10.1016/j.cct.2021.106661] [Medline: 34954098]
17. Kang H. The prevention and handling of the missing data. Korean J Anesthesiol 2013;64(5):402-406 [FREE Full text] [doi: 10.4097/kjae.2013.64.5.402] [Medline: 23741561]
18. Temple DS, Hegarty-Craver M, Furberg RD, Preble EA, Bergstrom E, Gardener Z, et al. Wearable sensor-based detection of influenza in presymptomatic and asymptomatic individuals. J Infect Dis 2023;227(7):864-872 [FREE Full text] [doi: 10.1093/infdis/jiac262] [Medline: 35759279]
19. Helmer P, Hottenrott S, Rodemers P, Leppich R, Helwich M, Pryss R, et al. Accuracy and systematic biases of heart rate measurements by consumer-grade fitness trackers in postoperative patients: prospective clinical trial. J Med Internet Res 2022;24(12):e42359 [FREE Full text] [doi: 10.2196/42359] [Medline: 36583938]
20. Schaffarczyk M, Rogers B, Reer R, Gronwald T. Validity of the polar H10 sensor for heart rate variability analysis during resting state and incremental exercise in recreational men and women. Sensors (Basel) 2022;22(17):6536 [FREE Full text] [doi: 10.3390/s22176536] [Medline: 36081005]
21. Perrotta AS, Jeklin AT, Hives BA, Meanwell LE, Warburton DER. Validity of the elite HRV smartphone application for examining heart rate variability in a field-based setting. J Strength Cond Res 2017;31(8):2296-2302. [doi: 10.1519/JSC.0000000000001841] [Medline: 28195974]
22. Malik M, Bigger JT, Camm AJ, Kleiger RE, Malliani A, Moss AJ, et al. Heart rate variability: standards of measurement, physiological interpretation, and clinical use. European Heart Journal 1996;17(3):354-381. [doi: 10.1093/oxfordjournals.eurheartj.a014868]
23. Lewis GF, Furman SA, McCool MF, Porges SW. Statistical strategies to quantify respiratory sinus arrhythmia: are commonly used metrics equivalent? Biol Psychol 2012;89(2):349-364 [FREE Full text] [doi: 10.1016/j.biopsycho.2011.11.009] [Medline: 22138367]

XSL•FO
RenderX

24. Shader TM, Gatzke-Kopp LM, Crowell SE, Jamila Reid M, Thayer JF, Vasey MW, et al. Quantifying respiratory sinus arrhythmia: effects of misspecifying breathing frequencies across development. Dev Psychopathol 2018;30(1):351-366. [doi: 10.1017/S0954579417000669] [Medline: 28554343]
25. Hayano J, Yuda E. Pitfalls of assessment of autonomic function by heart rate variability. J Physiol Anthropol 2019;38(1):3 [FREE Full text] [doi: 10.1186/s40101-019-0193-2] [Medline: 30867063]
26. Herrmann SD, Barreira TV, Kang M, Ainsworth BE. Impact of accelerometer wear time on physical activity data: a NHANES semisimulation data approach. Br J Sports Med 2014;48(3):278-282. [doi: 10.1136/bjsports-2012-091410] [Medline: 22936409]
27. van Rossum G, Drake FL. Python 3 Reference Manual. Scotts Valley, CA: Scotts Valley; 2009.
28. Giboin LS, Tokuno C, Kramer A, Henry M, Gruber M. Motor learning induces time-dependent plasticity that is observable at the spinal cord level. J Physiol 2020;598(10):1943-1963 [FREE Full text] [doi: 10.1113/JP278890] [Medline: 32115702]
29. Bates D, Maechler M, Bolker B, Walker S. lme4: Linear mixed-effects models using Eigen and S4. R package version 1. URL: http://CRAN.R-project.org/package=lme4 [accessed 2024-12-14]
30. Gamer M, Lemon J, Fellows I, Singh P. irr: Various coefficients of interrater reliability and agreement. R package version 0.84.1. Cran 2019:1-32.
31. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016;15(2):155-163 [FREE Full text] [doi: 10.1016/j.jcm.2016.02.012] [Medline: 27330520]
32. Gerke O. Reporting standards for a bland-altman agreement analysis: a review of methodological reviews. Diagnostics (Basel) 2020;10(5):334 [FREE Full text] [doi: 10.3390/diagnostics10050334] [Medline: 32456091]
33. Core Team R. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2019. URL: https://www.r-project.org/ [accessed 2024-12-14]
34. Cajal D, Hernando D, Lázaro J, Laguna P, Gil E, Bailón R. Effects of missing data on heart rate variability metrics. Sensors (Basel) 2022;22(15):5774 [FREE Full text] [doi: 10.3390/s22155774] [Medline: 35957328]
35. Baek HJ, Shin J. Effect of missing inter-beat interval data on heart rate variability analysis using wrist-worn wearables. J Med Syst 2017;41(10):147. [doi: 10.1007/s10916-017-0796-2] [Medline: 28812280]
36. Montalvo S, Martinez A, Arias S, Lozano A, Gonzalez MP, Dietze-Hermosa MS, et al. Commercial smart watches and heart rate monitors: a concurrent validity analysis. J Strength Cond Res 2023;37(9):1802-1808. [doi: 10.1519/JSC.0000000000004482] [Medline: 36862131]
37. Khushhal A, Nichols S, Evans W, Gleadall-Siddall DO, Page R, O'Doherty AF, et al. Validity and reliability of the apple watch for measuring heart rate during exercise. Sports Med Int Open 2017;1(6):E206-E211 [FREE Full text] [doi: 10.1055/s-0043-120195] [Medline: 30539109]
38. Davila MI, Kizakevich PN, Eckhoff R, Morgan J, Meleth S, Ramirez D, et al. Use of mobile technology paired with heart rate monitor to remotely quantify behavioral health markers among military reservists and first responders. Mil Med 2021;186(Suppl 1):17-24. [doi: 10.1093/milmed/usaa395] [Medline: 33499533]
39. Bourdillon N, Schmitt L, Yazdani S, Vesin JM, Millet GP. Minimal window duration for accurate HRV recording in athletes. Front Neurosci 2017;11:456 [FREE Full text] [doi: 10.3389/fnins.2017.00456] [Medline: 28848382]

Abbreviations

COTS: commercial off the shelf
ECG: electrocardiogram
HF: high-frequency
HRV: heart rate variability
IBI: interbeat interval
ICC: intraclass correlation coefficient
LF: low-frequency
REDCap: Research Electronic Data Capture
RMSDRR: root-mean-square of the difference in successive IBI values
RTI: Research Triangle Institute
STDRR: standard deviation of interbeat interval values
Q-Q: Quantile-Quantile

XSL•FO
RenderX

Edited by A Mavragani; submitted 16.10.23; peer-reviewed by N Peterson, N Shafagati; comments to author 28.04.24; revised version received 28.06.24; accepted 24.11.24; published 24.02.25.

Please cite as:
Davis-Wilson H, Hegarty-Craver M, Gaur P, Boyce M, Holt JR, Preble E, Eckhoff R, Li L, Walls H, Dausch D, Temple D
Effects of Missing Data on Heart Rate Variability Measured From A Smartwatch: Exploratory Observational Study
JMIR Form Res 2025;9:e53645
URL: https://formative.jmir.org/2025/1/e53645
doi:10.2196/53645
PMID:

©Hope Davis-Wilson, Meghan Hegarty-Craver, Pooja Gaur, Matthew Boyce, Jonathan R Holt, Edward Preble, Randall Eckhoff, Lei Li, Howard Walls, David Dausch, Dorota Temple. Originally published in JMIR Formative Research (https://formative.jmir.org), 24.02.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Development and Implementation of MyPainHub, a Web-Based Resource for People With Musculoskeletal Conditions and Their Health Care Professionals: Mixed Methods Study

Kerrie Evans[1,2], PhD; Jonathan Ko[1], MPhysio; Dragana Ceprnja[3], PhD; Katherine Maka[3], MBA; Darren Beales[4], PhD; Michele Sterling[5], PhD; Kim L Bennell[6], PhD; Gwendolen Jull[5], PhD; Paul W Hodges[5], PhD; Marnee J McKay[1], PhD; Trudy J Rebbeck[1], PhD

[1]The University of Sydney, Sydney, Australia
[2]Healthia Limited, Bowen Hills, Australia
[3]Western Sydney Local Health District, Westmead, Australia
[4]Curtin University, Perth, Australia
[5]The University of Queensland, Brisbane, Australia
[6]The University of Melbourne, Melbourne, Australia

Corresponding Author:
Kerrie Evans, PhD
The University of Sydney
City Road, Camperdown
Sydney, 2050
Australia
Phone: 61 0731804900
Email: kerrie.evans@healthia.com.au

Abstract

Background: Musculoskeletal conditions, including low back pain (LBP), neck pain, and knee osteoarthritis, are the greatest contributors to years lived with disability worldwide. Resources aiming to aid both patients and health care professionals (HCPs) exist but are poorly implemented and adopted.

Objective: We aimed to develop and implement MyPainHub, an evidence-based web-based resource designed to provide comprehensive, credible and accessible information for people with, and HCPs who manage, common musculoskeletal conditions.

Methods: This mixed methods study adhered to the New South Wales Translational Research Framework and was evaluated against the Reach, Effectiveness, Adoption, Implementation, and Maintenance (RE-AIM) framework. Consultation with key stakeholders (patients, HCPs, researchers, industry, consumer groups, and website developers) informed content, design, features, and functionality. Development then aimed to meet the identified need for a “one-stop shop”—a central location for information about common musculoskeletal conditions tailored to a person’s condition and risk of poor outcomes. MyPainHub was then developed through an iterative process and implementation strategies were tailored to different health care settings. Quantitative and qualitative evaluation occurred with patients and HCPs.

Results: In total, 127 stakeholders participated in the development phase; initial consultation with them led to embedding 2 validated screening tools (the Short Form Örebro Musculoskeletal Pain Screening Questionnaire and the Keele STarT MSK tool) in MyPainHub to guide information tailoring for patients based on risk of poor outcomes. Development occurred in parallel and feedback from stakeholders informed design and content including structure, functionality, and phrasing and images to use to emphasize key points. Consultation resulted in information for patients being categorized using key guideline-based messages (general information, your pathway, exercise, and imaging) while information for clinicians was categorized into assessment, management, and prognosis. Implementation occurred in different health care settings with the most effective strategies being interactive education via webinars and workshops. The evaluation phase involved web-based questionnaires (patients: n=44; HCPs: n=29) and focus groups (patients: n=6; HCPs: n=6). Patients and HCPs found MyPainHub user-friendly, acceptable, credible, and potentially able to support self-management. Patient participants identified areas for improvement such as including more specific information on preventative measures and pain relief options. Despite positive feedback, only 35% (10/29) of HCPs used MyPainHub with their patients. HCP participants identified challenges including insufficient training and lack of familiarity

XSL•FO
RenderX

with using web-based resources in existing clinical workflows. Following implementation, the information contained on MyPainHub changed knowledge and practice for some patients and HCPs.

Conclusions: Following extensive and iterative stakeholder engagement, MyPainHub was developed as an evidence-based web-based resource and perceived by patients and HCPs as user-friendly, credible, and acceptable. Active implementation strategies are required for adoption and implementation and greater training focusing on strategies to implement MyPainHub into clinical practice may be necessary.

Trial Registration: Australian New Zealand Clinical Trials Registry ACTRN12619000871145; https://tinyurl.com/438kkyt3

(JMIR Form Res 2025;9:e63780) doi:10.2196/63780

KEYWORDS

clinical pathways; allied health; self-management; health information; ehealth; co-design

Introduction

Musculoskeletal conditions are characterized by pain and reduced physical function, often resulting in deterioration of mental well-being, increased susceptibility to developing other chronic health conditions, and loss of productive life years in the workforce [1]. Globally, musculoskeletal conditions affect 1.7 billion people and are ranked as the highest contributors to years lost to disease, above mental, respiratory, and cardiovascular diseases [1-3]. Among people presenting for care, low back pain (LBP), neck pain, whiplash associated disorders (WAD), and knee osteoarthritis are the most prevalent and disabling musculoskeletal conditions [1,4]. In Australia, the health care system operates as a mixed public-private model, with primary funding through Medicare, a universal public insurance system providing subsidized access to health care services, supplemented by private insurance options. The federal and state governments manage health care collaboratively, with federal oversight on Medicare and Pharmaceutical Benefits Scheme funding, while states are responsible for public hospital management and service delivery. Despite targeted government initiatives to mitigate musculoskeletal disease burden, such as the national strategic action plan for arthritis and funding under the chronic disease management plan, musculoskeletal conditions in Australia still account for 20% of the total burden of disease [5].

One factor contributing to this burden is poor implementation of guideline-based care [6,7]. Broadly speaking, clinical guidelines for LBP [8,9], neck pain [10,11], WAD [12], and knee osteoarthritis [13-16], contain similar recommendations for the provision of evidence-based care—screening for red flags and risk factors for poor outcome, providing patient-centered education (including options for care, advice to stay active), and avoiding unnecessary investigations and interventions. However, health care professionals (HCPs) often fail to screen for prognostic factors and continue to overuse nonrecommended treatments, imaging, pharmacological interventions, and surgery when managing musculoskeletal conditions [17-22]. Non–guideline-based care occurs across health care settings (ie, primary and tertiary care settings) despite evidence that these low-value practices do not improve health outcomes, waste health care resources and can be harmful [17,23-25].

Barriers to implementation and adoption of clinical guidelines include system-related factors (eg, funding models [26,27]) but also practitioner and patient-related factors. For example, lack of time, limited access to evidence-based resources, and discordant expectations are cited by both HCPs and patients as reasons why evidence-based recommendations are not followed [28-31]. In response to some of these challenges and the increasing trend for people to obtain health information from the internet, websites and other digital resources about musculoskeletal conditions have been developed [32]. The sheer volume of resources now available is overwhelming and information from the internet can be inaccurate, perceived to have low credibility, lack cultural diversity and not meet the needs of consumers [33-38]. In addition, often there is little training or education provided to HCPs as to how to best use internet-based resources to help implement, adopt and support evidence-based care [39-41].

To empower and engage patients and HCPs to make informed decisions about evidence-based recommendations, websites should be developed and co-designed by end users, include input from experts and opinion leaders, and be practical and scalable, that is, follow a translational research framework [42]. Examples of websites that have adopted a translational research framework include MyBackPain [33], MyWhiplashNavigator [43], and MyJointPain [44]. The development of these websites began with a needs assessment to identify research priorities and consumer needs before creating and curating guidelines and evidence-based health care information. Stakeholder engagement was essential in content creation and testing. A dissemination strategy was developed to ensure that the website was reached by the target audience. Monitoring and evaluation mechanisms tracked website usage and user feedback for quality improvement and their impact has been investigated in clinical trials [45,46]. Although these websites have undergone a robust and rigorous process of development, the information contained on these websites is condition specific. Given that many people experience multiple coexisting musculoskeletal conditions [47], and that clinical guidelines for common musculoskeletal conditions contain similar key messages [48], the aim of this study was to develop a website that could serve as a hub or a "one-stop shop" for both patients with, and HCPs managing, musculoskeletal conditions. The website was also initially developed to support patients and HCPs enrolled in a randomized control trial that was due to commence—the PAthway of CarE (PACE) musculoskeletal (MSK) clinical trial

XSL•FO
RenderX

[49]. PACE-MSK, funded by Australia's National Health and Medical Research Council (GNT1141377), was designed to evaluate the effectiveness of a clinical PACE for people with musculoskeletal disorders where care is provided based on people's risk of poor outcome [49]. A website that could serve as a one-stop shop for both patients and clinicians in PACE-MSK, as well as support patients and HCPs beyond the trial, was an important part of the design. This study reports on the development, implementation, and evaluation of MyPainHub, a website designed to provide support to patients and HCPs in the management of LBP, neck pain, WAD, and knee osteoarthritis.

Methods

Overview

This mixed methods study involved 3 stages (development, implementation, and evaluation), and followed the New South Wales Translational Research Framework [42] with outcomes assessed according to the Reach, Effectiveness, Adoption, Implementation, and Maintenance (RE-AIM) framework [50]. Both frameworks provide structured approaches to ensure that research is relevant and scalable and are widely used for research in the Australian health care context.

Ethical Considerations

Ethics approval was obtained from the Human Research Ethics Committees at The University of Sydney (2018/926, 2021/951), The University of Melbourne (1954239), The University of Queensland (2019000700), Curtin University (HRE2019-0263, HRE2020-0562) and Western Sydney Local Health District (2021/ETH11872). Participants were informed about the study before providing their written consent (either on paper or via a web-based form). Participation was voluntary and participants were informed about their right to withdraw from the study at any point without any consequences. Collected data were coded and deidentified.

Recruitment

Each stage involved different recruitment procedures. During the development stage, researchers, consumer, industry and clinical partners were identified by members of the research team and invited to participate via direct contact. These stakeholders were those with recognized expertise in musculoskeletal conditions and who had previously collaborated with members of the research team or been involved in previous studies. HCPs were involved in the PACE-MSK trial or had been involved in previous trials and had consented to be contacted for future trials. HCPs were working in either primary or tertiary (hospital) settings. People with MSK conditions had been involved in previous trials and had consented to be contacted for future trials. Contact was made with these groups via email or direct phone call. Students allied to HCPs were enrolled in a unit of study at a participating university and had consented to participate.

In the implementation and evaluation phases, HCPs and patients were involved in the PACE-MSK trial or had been involved in previous trials and had consented to be contacted for future trials. In addition, HCPs working in occupational health settings known to the research team or who had been involved in previous studies were invited to participate either by email or direct phone call. Those HCPs working in our clinical partners' tertiary care setting (Western Sydney Local Health District, [WSLHD]) were directly invited by their peers and managers. Patients with MSK conditions were also invited to participate by their treating clinician.

Stage 1: Development

The aim of the development stage was to design a website where patients and HCPs could access relevant, up-to-date, high-quality evidence-based information on common musculoskeletal conditions (LBP, neck pain, WAD and knee osteoarthritis). Given the expected commencement date of the PACE-MSK trial, the development stage incorporated comprehensive stakeholder engagement rather than a formal co-design process. The problem identified by the research team was that multiple resources exist, with neither a central nor coordinated location for people to access credible evidence-based information [51].

The first step of the New South Wales Research Translation Framework is Idea Generation, where key stakeholders are engaged to design a solution to the problem. Our core team comprised clinician- researchers from 4 Australian Universities (The University of Sydney, The University of Queensland, The University of Melbourne and Curtin University) with research expertise in musculoskeletal conditions. The team determined that key stakeholders likely to benefit from this resource were allied HCPs working in both primary and tertiary care, researchers involved in creating evidence for musculoskeletal conditions, specialist musculoskeletal clinicians, specialist medical professionals, people with musculoskeletal disorders, consumer organizations and industry partners. We then determined the level of influence and interest of each stakeholder using a stakeholder engagement matrix and considered appropriate strategies for managing their involvement and expectations [52]. Engagement methods ranged from being involved in co-design and implementation (manage) to piloting of resources (meet needs) and meetings (keep informed; Table 1).

XSL•FO
RenderX

Table 1. Classification of key stakeholders involved in the development stage of MyPainHub according to influence and interest, appropriate strategies and engagement methods.

Key stakeholder	Influence and interest	Appropriate strategies	Engagement methods
Allied HCPs[a], Specialist musculoskeletal allied HCPs	Higher influence, higher interest	Involve	• Consultation on key content • Facilitated meetings to review and refine content • Focus groups to pilot and identify gaps • Involvement in implementation process
Student allied HCPs	Lower influence, higher interest	Involve	• Involve in implementation and evaluation process
Researchers involved in musculoskeletal health	Higher influence, higher interest	Involve	• Consultation on key content and existing resources
Specialist medical professionals	Lower influence, higher interest	Meet needs	• Semistructured interview on resources and pathway
People with musculoskeletal conditions	Lower influence, higher interest	Meet needs	• Review and piloting of resources with feedback
Consumer partners: Arthritis Australia, Australian Pain Management Association, Musculoskeletal Australia	Higher influence, lower interest	Keep informed	• In-person meetings
Industry and clinical partners: Australian Physiotherapy Association, Chiropractic Australia, State Insurance Regulatory Authority of New South Wales, Primary Healthcare Networks, Knowledge Translation Australia	Higher influence, lower interest	Keep informed	• In-person meetings

[a]HCPs: health care professionals.

We held work-in-progress meetings with key stakeholders between July 2018 and October 2022. Discussions included design concepts, content, functionality, website maintenance and support and how this could be implemented. To ensure optimal functionality and ease of use, a digital agency (Vivo Group, Brisbane, Queensland, Australia) with experience in development of user-friendly websites was selected to design the website following a robust tendering process. As the content was developed, each stakeholder group shared their views and perspectives and provided feedback about additional key messages and resources that should be included or omitted. Before implementation, the website went through final checks by the research team and the digital agency to test features, functionality and resolve any issues.

Stage 2: Implementation

The aim of the implementation stage was to determine strategies that would help ensure effective and scalable implementation of the website before its widespread public release. Groups thought likely to source information about musculoskeletal conditions were identified and included allied health students and graduate allied HCPs working in health care settings where people with musculoskeletal conditions present. Settings included primary care, occupational health, and tertiary care settings. Finally, allied HCPs working in a specialist capacity were engaged. This process ensured implementation occurred with clinicians with a range of clinical experiences and in different care settings. We chose implementation strategies ranging from passive (dissemination) to more active strategies (facilitated and discussion-based education with opinion leaders) based on strategies that had been successful in our previous work (Table 1) [33,43,44]. During this stage, the website was not publicly accessible.

Patients with musculoskeletal conditions were invited by their treating clinician to use the resources on the website and then provided consent to participate in qualitative studies associated with both our primary and tertiary care implementation initiatives. After using the website, patients participated in a semistructured interview (via a web platform) or focus group (in person), depending on what was feasible in the particular health care setting and on the patient's preference and location, to provide their opinions on the resource.

Allied HCPs Working in Primary Health Care Settings

They were engaged via a passive (dissemination) strategy to use the website. The primary HCPs providing care for people with musculoskeletal conditions in the PACE-MSK trial [49] were informed of the trial and the website via a phone call. Both the HCP and their patients were emailed a link to the website and encouraged to refer to the resources during the episode of care. Over an 18-month period, the researchers held fourteen 1-hour video-conferenced noncompulsory educational meetings inviting HCPs to attend with the aim of keeping them engaged in the trial and website.

XSL•FO
RenderX

Allied HCPs Working in Occupational and Tertiary Care Settings

They were invited to participate in an active (interactive education using opinion leaders) implementation strategy. HCPs working in occupational health settings were invited to participate either by email or direct phone call. Those working in our clinical partners' tertiary care setting (WSLHD) were directly invited by their peers and managers. Those accepting the invitation then attended a web-based (occupational health) or in-person (tertiary care) discussion-based educational workshop explaining the purpose of the website and its key features. These HCPs were then encouraged to use the resources on MyPainHub to assist in assessment and management of their patients. After completion of the episode of care (or after 3 months), the HCPs were invited to complete a questionnaire to evaluate acceptability, feasibility, credibility, and actual practice of using the website and then participated in a focus group to explore their opinions in more depth (Multimedia Appendix 1).

Specialist Allied HCPs

They were invited to participate in a 2-day facilitated discussion-based educational workshop described elsewhere [53]. The workshop was designed to inform specialist HCPs about how to implement risk-stratified care (supported by MyPainHub) for people at risk of poor outcome. Key features that may facilitate the provision of this care were discussed during the workshop, including how MyPainHub could be used to support management. Two follow-up web-based meetings were held to encourage use of the website and implementation of a new pathway of care.

Student HCPs

They were engaged through classroom education. The website was integrated as a learning resource in 2 units of study for physiotherapy students in their final year of study at The University of Sydney. Key components of the education model included (1) risk assessment using the web-based risk-assessment tools resources available on the website, (2) using the resources to assist the provision of evidence-based treatment (guideline-based exercises and risk-based advice) for people with musculoskeletal disorders, and (3) timely and appropriate referral to specialist musculoskeletal clinicians for people at high risk of poor outcomes.

Outcome Assessed

The main outcome assessed for the implementation stage was the reach of the strategy. Reach was primarily assessed by the number registered on the resource or those invited to use it. Website traffic and trends (eg, total visits, total page views, and top landing pages) were collected through Google Analytics and built-in website reports.

Stage 3: Evaluation

The aim of the evaluation stage was to investigate both effectiveness, acceptability, and adoption of the website (RE-AIM framework) by allied HCPs (students, allied HCPs, and specialist HCPs) and people with musculoskeletal conditions. Mixed methods involving quantitative (web-based survey) and qualitative (focus groups or semistructured interviews) were used.

Allied HCPs (Students to Specialists)

Participants were invited to share their perspectives about MyPainHub via a web-based survey. Participants answered questions relating to change in knowledge and practice (eg, "The website has changed the way I use prognostic tools"), acceptability and credibility of the website (eg, "I trusted the information on the website"), structure and features of the website (eg, "It was easy to navigate through the website"), and overall satisfaction with the website. HCPs were also asked open-ended questions which explored opinions on the best features of the website, barriers to use and suggestions for improvement. Student HCPs were asked additional questions relating to whether the information on the website aligned with what they had been taught and helped them achieve the learning outcomes of their course (eg, "The messages on the website about imaging made sense to me"; "Viewing the resources in MyPainHub helped me achieve the learning outcomes of my course").

Qualitative data were collected from a proportion of HCPs who had been directly invited to participate by their peers and managers and who had consented to participate in a focus group. Participants were provided with the option of completing only the web-based survey or participating in both the web-based survey and the focus group. The focus groups were conducted in person and were designed to explore HCPs' opinions in greater depth.

People With Musculoskeletal Conditions

As occurred with HCPs, people with musculoskeletal conditions were invited to share their perspectives about MyPainHub via a web-based survey. Patient participants answered questions designed to assess (change in) knowledge (eg, "People with my condition should not exercise until they have had scans done"), acceptability and credibility of the website (eg, "I trusted the information on the website"), structure and features of the website (eg, "It was easy to navigate through the website"), and overall satisfaction with the website. Patient participants were also asked open-ended questions which explored opinions on the best features of the website, barriers to use and suggestions for improvement.

Qualitative data were collected from a proportion of people with musculoskeletal conditions who consented to participate in a focus group (in person) to explore their perspectives in greater depth. Participants in the PACE-MSK trial participated in semistructured interviews (via a web platform) 3 months after entering the trial but these findings will not be reported in this paper.

Data Analysis

With respect to the web-based surveys, closed-ended questions were rated using a 5-point Likert scale ranging from 1 (strongly disagree or extremely unsatisfied) to 5 (strongly agree or extremely satisfied). The open-ended questions underwent a descriptive analysis.

XSL•FO
RenderX

With respect to the qualitative data collected during the semistructured interviews and focus groups, the discussions were recorded and transcribed. Analysis was conducted first by familiarization, then preliminary coding, searching, reviewing and naming and then reporting [54]. The transcriptions were manually checked by a member of the research team and corrected against the verbatim account. The transcriptions were read several times to make sense of the data. Open coding was conducted by naming sections of the participant responses in the text. Codes were then grouped to form meaningful categories in discussion with all authors. The next step was to construct broad themes from the categories with authors working collaboratively in the interpretation of qualitative data to draw on the combined insights of all members of the research team.

Results

Overview

The total number of participants involved in each stage is provided in Table 2.

Table 2. Stakeholders involved in development, implementation and evaluation of MyPainHub.

Stages and stakeholders	Participants involved, n
Development	
HCPs[a]	67
Researchers	28
Specialist medical professionals	6
People with musculoskeletal conditions	18
Consumer partners	3
Industry and clinical partners	5
Implementation	
HCPs	136
People with musculoskeletal conditions	778
Evaluation	
HCPs (web-based surveys)	29
HCPs—focus group (tertiary care, in person)	6
People with musculoskeletal conditions (web-based surveys)	11
People with musculoskeletal conditions—focus groups (tertiary care, in person)	6
People with musculoskeletal conditions—semistructured interviews (primary care, via a web platform)[b]	33

[a]HCPs: health care professionals.

[b]Findings not reported in the present study.

Stage 1: Development

The stakeholder consultation generated ideas that informed the website design and implementation. This process informed the development of the website both for “Patients” and “Clinicians.” Key features and content were identified and then modified iteratively as a result of the consultation process. Modifications did not follow a strict, linear order occurring instead in parallel.

MyPainHub for Patients

The core clinician-researcher team initially identified that common key guideline-based messages for implementation among people with musculoskeletal conditions were to provide accurate advice, appropriate messaging regarding imaging, to promote exercise, and provide differential care based on risk of poor outcome. The stakeholder consultation process resulted in agreement to structure the website around prognostic risk and musculoskeletal condition, with resources automated thereafter. Two risk tools, the Short Form Örebro Musculoskeletal Pain Screening Questionnaire (SF-ÖMPSQ) [55] and the Keele STarT MSK tool [56] were identified by the research team as being validated tools to risk-stratify people with musculoskeletal conditions. Initial surveys of the specialist HCPs (n=50) suggested that 44% (22/50) either frequently or always used the SF-ÖMPSQ, 14% (7/50) the WhipPredict tool [57,58], 8% (4/50) the Keele STarT MSK tool [56], and 6% (3/50) the STarT Back Tool [59,60]. However, given WhipPredict is specific for WAD and the STarT Back Tool is specific for LBP, it was decided to embed the SF-ÖMPSQ and the Keele STarT MSK tool (both designed for broader musculoskeletal conditions) on the website for patients to complete with scores being automatically calculated. After completing the tools and nominating their primary condition (LBP, neck pain, or knee osteoarthritis), patients access resources that are designed specifically for both their risk subgroup (according to their score on the SF-ÖMPSQ) and condition. Each time a patient logs into the website, they are taken to their condition and risk-specific home page but can access information for the other conditions

XSL•FO
RenderX

through simple navigation links and drop-down menus where appropriate.

The stakeholder consultation process identified that people with musculoskeletal conditions (n=18) wanted general information about their condition, ideas on how they could adapt to their lifestyle, their work and what they could do to help themselves. Consultation with industry (Table 1) resulted in adding links to other evidence-based websites and resources (eg, the Royal Australian College of General Practitioner's guidelines for the management of knee osteoarthritis) and highlighted that options for care should be clearly explained (eg, people at low risk of poor outcome should be encouraged to avoid imaging and surgery and be provided with evidence-based advice and education whereas those at high risk should be provided with more comprehensive education related to the multifactorial, biopsychosocial nature of pain disorders). Consultation with patients with musculoskeletal conditions resulted in information being written in a friendly and nonjudgemental way and links to other websites and resources were selected so that they were a "simple click away" to avoid users being sent to another website where they would have to create an account. The consultation process resulted in the resources being structured under the categories of "General Information," "Your Pathway," "Exercise," and "Imaging."

MyPainHub for HCPs

Initial stakeholder consultation with 67 HCPs resulted in >200 individual pieces of feedback about the content, structure, features, and resources to use. Feedback included, for example, modifications to phrasing used or emphasis of key points (eg, "More emphasis on 'not one posture is best' would be good and give permission to sit in a comfortable, relaxed position to reduce or prevent fear of movement"), additional information to add (eg, "You could add some info about basic exercise such as walking, range of movement, balance and light upper limb weights"), structure (eg, "It was very easy to navigate"), and appearance (eg, "I think the pages that have banner images should have those images reduced in size – the banners are little dominating"). This feedback, and changes made to the website, was iterative. Overall, the feedback suggested that the website should contain key guideline-based messages, include "bite-sized" videos of information delivered by credible HCPs, include key practice points, open-access contemporary articles, clinical guidelines and outcome measures, and links to other credible resources. Subsequently, the information on MyPainHub for clinicians was structured under the categories of "Assessment," "Prognosis," and "Management."

Structure and Appearance

The website designers provided expertise on the color scheme, structure and functionality of the website to achieve the aims of the project. The content was written in a professional and friendly tone, and images were selected that avoided emphasizing, for example, a pathoanatomical focus where not appropriate and that avoided stereotyping.

The sitemap is shown in Figure 1. Pages were kept simple to optimize ease of navigation (Figure 2).

Figure 1. Sitemap for the development of MyPainHub.

Sitemap	Page Layout
Home Page	Home Layout
About the Trial	Standard Layout
FAQs	Standard Layout
Contact	Contact Layout
Create Account	Create Account Layout
Pain Questionnaire	Pain Questionnaire Layout
Resource Category Selection	Resource Categories Layout
Resource Dashboard	Dashboard Layout
Individual Resource	Resource Layout
ORPHAN PAGES	
Privacy Policy	Standard Layout

XSL•FO
RenderX

Figure 2. Wireframes used in the development of MyPainHub were designed to be clear and simple to facilitate navigation and clarity of information for the user.

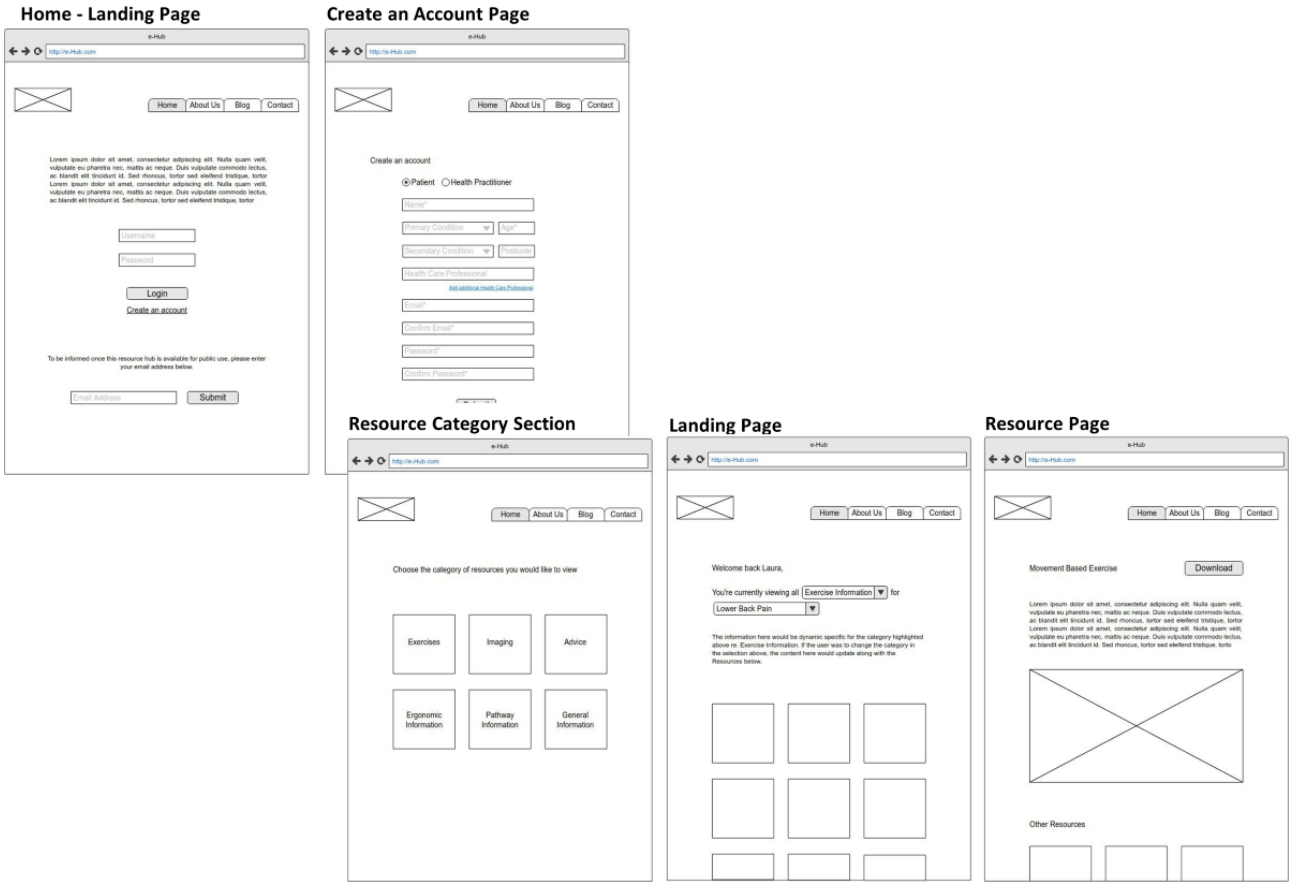

Permissions

The website was built on the website designer's content management system which was developed in a software platform in PHP (using the Symfony framework) that is based on standard customizable modules. A tiered permission system was used for MyPainHub. Administrators (ie, researchers) were able to access and amend all information on the website. HCPs were able to review all content on the website. Patients were able to view content related to their risk profile (ie, low risk or high risk) but they could see information for all conditions. However, once they identified their primary complaint during their registration process, their landing page was set to their primary condition. Only the researchers had access to patients' scores on the SF-ÖMPSQ and the Keele STarT MSK tool (during the PACE-MSK trial, the researchers would contact the patient's treating HCP to inform them of their patient's score [49]).

Stage 2: Implementation

Implementation of the website took place in 4 Australian states (New South Wales, Queensland, Western Australia, and Victoria) from February 2018 to December 2023 as a component of PACE-MSK, the clinical trial for which the website was built [49]. As part of PACE-MSK, 767 people with musculoskeletal conditions had consented to be part of the trial and were therefore required to register on MyPainHub. A total of 693 HCPs were involved in the PACE-MSK trial but registering on MyPainHub, whilst possible, was not a requirement of being involved in the trial. In addition to the PACE-MSK trial, the website was implemented within WSLHD (both patients and HCPs), with allied HCPs in occupational health settings and with students (Table 3).

The implementation strategy adopted influenced the number of participants who, despite consenting to participate, registered on the website (Table 3). As a result of the education via webinars and workshops, 100% of allied HCPs working in occupational settings, tertiary care settings and specialist HCPs who consented to participate registered on the website. However, despite being embedded in a unit of study in a university physiotherapy program, only 16% of students registered on the site. Passive strategies, such as phone calls or reminders, were ineffective with only 29% of HCPs working in primary care registering (Table 3).

Website data for a 12-month period (June 2021-June 2022) are presented in Table 4. The total page views of the website during the period were 23025 of which 12149 were unique page views. The pages with the highest unique views were general information about knee osteoarthritis, neck pain assessment (clinicians), frequently asked questions, LBP assessment (clinicians), general information about neck pain (general information), and exercise for knee osteoarthritis (Table 4). Users viewed on average 4 pages during a visit and stayed approximately 4 minutes on each page.

XSL•FO
RenderX

Table 3. Target markets of website users, implementation strategies, the numbers invited in each target market and percentage successfully registered.

Target market	Implementation strategy	N=invited	N=consented, (participants)	N=registered, n (%)
HCPs[a] in primary care	Reminders via phone calls and emails	693	196	57 (29)
Allied HCPs in occupational health settings	Discussion-based education (webinars) and reminders	48	6	6 (100)
Allied HCPs in tertiary care (WSLHD[b])	Interactive education (in-person workshop) and reminders	9	6	6 (100)
Specialist allied HCPs	Interactive education (in-person workshop and 2 web-based workshop reminders)	128	50	50 (100)
Student HCPs	Embedded classroom education	102	102	17 (17)
People with musculoskeletal conditions in primary care	Invitation via treating HCPs and research team (PACE-MSK trial)	767	767	767 (100)
People with musculoskeletal conditions in tertiary care (WSLHD)	Invitation via treating HCPs	18	11	11 (100)

[a]HCPs: health care professionals.

[b]WSLHD: Western Sydney Local Health District.

Table 4. Summary of the number and duration of page views for MyPainHub.

Page title	Unique page views, N	Average time (seconds)
Knee osteoarthritis: general information	657	42
Neck pain: assessment (clinicians)	408	41
Frequently asked questions	157	107
Lower back pain: assessment (clinicians)	128	88
Neck pain: general information	121	47
Knee osteoarthritis: exercise	99	29

Stage 3: Evaluation

Allied HCPs (Students to Specialists)

Of the 102 allied HCPs invited, 29 (28%) HCPs completed the evaluation questionnaire. Of these, 93% (27/29) responded that MyPainHub had the potential to change knowledge as the information contained on the website was credible, trustworthy, clearly presented and easy to navigate. In terms of acceptability of the website, 93% (27/29) of HCP participants were satisfied with the website, were likely to revisit, and were likely to refer other colleagues to the site. However, less than 35% (10/29) of HCPs reported using the website with their patients. Suggestions for improvement included considering changing some of the information to enhance understanding for people with low health literacy and to include other musculoskeletal conditions (HCPs' opinions are shown in Figure 3).

From the focus group (6 HCP participants), four key themes emerged: (1) MyPainHub has utility as a web-based resource; (2) MyPainHub contains credible information, which reinforces best practice; (3) challenges in implementation; and (4) potential opportunities for enhancement.

XSL•FO
RenderX

Figure 3. Health care professionals' (HCP) opinions about MyPainHub. Opinions on knowledge and acceptability were positive, but this contrasted to how often HCPs used MyPainHub with patients.

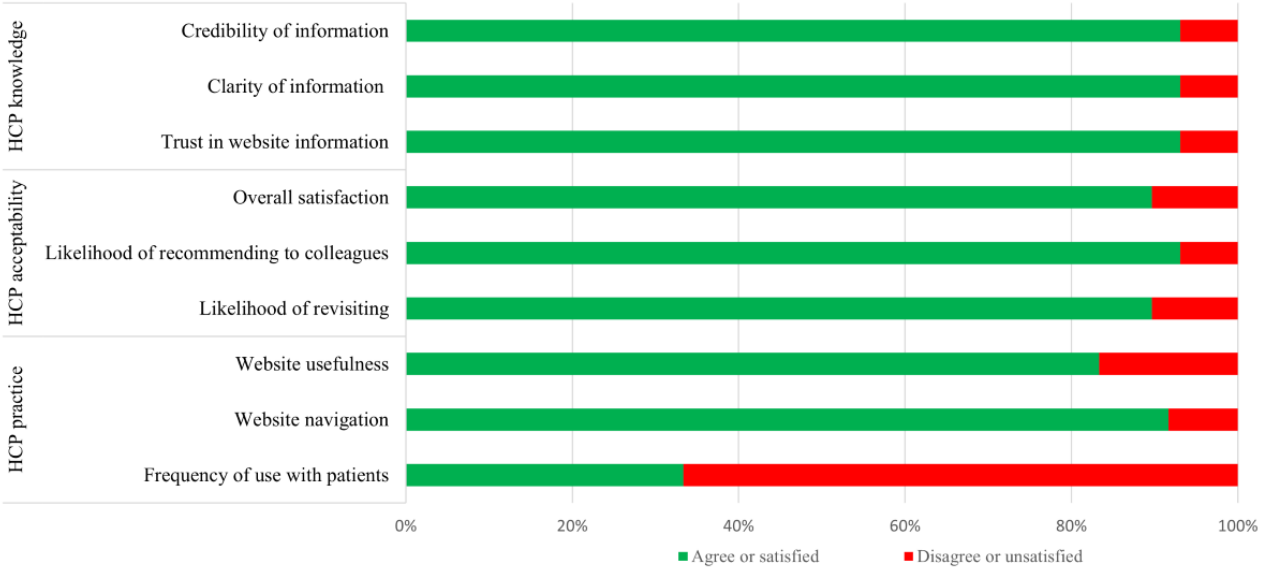

Theme 1: MyPainHub Has Utility as a Web-Based Resource

HCP participants reported MyPainHub contained relevant information, was easy to navigate and improved professional confidence. HCPs supported the relevance and breadth of topics on the site and its utility in enhancing quality of care, "especially for junior staff or those with lesser experience." Despite challenges with implementation, participants had felt that it was useful for "encouraging patient engagement with HCPs," fostering an educational continuum, and "...builds on education and use of resources between sessions based on patient learning preferences." The "easy to navigate" resource had increased professional confidence as it "could help through processes, guide treatment and reduce fears when unsure or confused." This theme highlighted the website's potential in bridging knowledge gaps and facilitating a more informed patient-provider interaction.

Theme 2: MyPainHub Contains Credible Information, Which Reinforces Best Practice

The HCPs reported information on MyPainHub was credible and aligned with best practice, particularly because of the inclusion of video content and articles from expert clinicians and researchers. Comments regarding the information on MyPainHub "reinforced stuff previously learnt at university" and "reinforced what we already say (to patients) in a different format." Some remarked that the resource helped their own learning around best practice which in turn would influence the education they provided to their patients: "I don't have too much outpatient experience nor am I very confident. So I really like the clinician end of it...using that to understand things for myself and then being able to explain that to someone."

Theme 3: Challenges With Implementation

HCP participants reported implementation of MyPainHub was harder than expected and identified several challenges including internet availability in the clinic, time constraints, and a lack of familiarity with using web-based resources to support patient interactions. Acknowledging "change in practice is difficult to achieve in busy clinical settings," participants identified that "more training and preparation to support new workflows is needed." HCPs perceived those patients "may expect a more hands-on approach" which made them "hesitant" to adopt MyPainHub within sessions. As some HCPs were "not accustomed to using online resources" to aid their management of patients, they acknowledged they "often forgot about MyPainHub as an option," particularly in the context of "competing demands and heavy clinical workloads."

Theme 4: Potential Opportunities for Enhancement

Opportunities for improvement were noted, including the simplification of how information was presented, reduction of hyperlinks, and expansion of the number of conditions included on the site. HCPs voiced the need for the key messages and information to be "reduced" and "summarized" so that users would not have to "scroll as much." Some participants suggested that "more downloadable resources may be more helpful" and proposed more "ready-made printouts and resources of information."

People With Musculoskeletal Conditions

Eleven people with musculoskeletal conditions in the tertiary setting (WSLHD) completed the evaluation questionnaire. All responded that MyPainHub had the potential to change knowledge as the information contained on the website was credible, trustworthy and clearly presented. In terms of acceptability, all participants were satisfied with MyPainHub, likely to revisit and recommend the website to others. In addition, all participants indicated that the web-based resources helped them understand the best ways to manage their condition, the website was easy to navigate, and that MyPainHub would be a useful tool for other people with musculoskeletal conditions.

Four key themes emerged from the focus group (n=6 participants) from participants at WSLHD and these aligned

XSL•FO
RenderX

with the findings from the questionnaire: (1) MyPainHub adds to the understanding of the condition, (2) MyPainHub is credible as it aligns with HCP advice and information, (3) MyPainHub supports self-management, and (4) potential areas for improvement.

Theme 1: MyPainHub Adds to the Understanding of the Condition

Participants reported information on MyPainHub added to the understanding of their musculoskeletal condition. It was reported that the website offered clear explanations that were "definitely helpful," "adding knowledge and understanding," and "changing their perspectives on their condition." This aspect was further supported by the ability to "revisit information" via the "links" as needed, allowing for engagement with the content and facilitating ongoing learning.

Theme 2: MyPainHub Is Credible as it Aligns With HCPs' Advice and Information

Participants reported information on MyPainHub as credible, particularly when it aligned with their HCP's advice and information. This alignment reassured patients and "built trust" in both the website content and the advice provided by their HCPs. Trust was further reinforced when the website was recommended by their HCP or "experts."

Theme 3: MyPainHub Has the Potential to Support Self-Management

Participants reported the website increased their confidence in managing their condition. This was achieved by supporting "self-help and lifestyle changes" and "promoting exercise, and providing a holistic approach to physical activity and mental health." The ability to "support more activity and distract from negative thoughts" was reported, underlining the website's role in supporting both physical and psychological aspects of self-efficacy.

Theme 4: Potential Areas for Improvement

Participants identified areas for improvement. This included the need for more specific information tailored to "work or employment challenges," "detailed do's and don'ts for condition management," "preventative measures," "pain relief options," and "information about surgery and when it may be needed." "Regular updates" and the creation of "shared chat groups" were also suggested as potential areas for improvement.

Discussion

Principal Findings

This paper describes the process undertaken to develop, implement and evaluate a website designed to support patients with, and HCPs who manage, musculoskeletal conditions, specifically LBP, neck pain, WAD, and knee osteoarthritis. The findings suggest that following a research translation framework and incorporating extensive and iterative consultation with multiple key stakeholders resulted in MyPainHub being acceptable, credible, and easy to navigate with participants reporting they are likely to revisit and recommend the website to others. Embedding prognostic tools (the SF-ÖMPSQ and the Keele STarT MSK tool) allowed information to be tailored to a person's risk of poor outcome which may help both patients and HCPs better adopt guideline-based management. However, training for HCPs that includes more specific strategies on how to incorporate MyPainHub into clinical practice would be beneficial for more successful implementation.

A key component of the development of MyPainHub was the extensive involvement of a wide range of key stakeholders with different levels of experience and from different health care settings over a 4-year period. The initial development phase focused on gathering perspectives from people with musculoskeletal conditions and HCPs from primary care settings, researchers, key opinion leaders, professional associations, consumer groups and website developers. The latter development phase involved seeking perspectives of HCPs and patients from tertiary care settings, allied professionals from occupational settings and from allied health students. This broad stakeholder engagement reflects contemporary user-centered design and was undertaken to facilitate future wider implementation of MyPainHub across a range of settings and populations [61]. Whilst active implementation strategies (eg, interactive education and workshops) were generally more effective than passive strategies (eg, phone calls and emails), implementing MyPainHub in a unit of study was not successful amongst allied health students. In our study, students were introduced to MyPainHub during a lecture and were then responsible for designing a treatment program using MyPainHub but no specific tutorial time was allocated for this task. A previous study [43] that had implemented an innovative, guideline-based website, My Whiplash Navigator, found student HCPs had a higher uptake of the website than primary HCPs but the students were allocated time during both a lecture and a tutorial to using the resource which may explain the difference in the level of engagement.

The main purpose of this project was to develop a resource that would support guideline-based care for patients at low and high risk of poor outcome. Tailoring information to a person's risk of poor outcome afforded the opportunity to provide different information on likely recovery pathways and interventions that may be required. In this way, MyPainHub reinforces key guideline-based messages, that people at low risk should do well with minimal care but that people at high risk may require more comprehensive assessment of multiple domains and a multidisciplinary approach to optimize health outcomes. Greater adoption and wider implementation of MyPainHub may improve the efficiency of health service delivery by helping to educate people about musculoskeletal conditions and influence care decisions but requires further research.

Despite the opinions of stakeholders that MyPainHub is acceptable, credible, and easy to navigate, few used the tools with patients, and additional training that focuses on how best to integrate the website into clinical practice in different health care settings is likely to be required for greater adoption and usage. Active engagement strategies were more effective than passive strategies and taking a multifaceted approach, for example embedding MyPainHub in existing educational opportunities (eg, workshops, curriculum, and training programs) and using key opinion leaders to reinforce the credibility of information, is likely to be important in future

XSL•FO
RenderX

implementation [62,63]. The poor response rate and level of engagement of students could be overcome by integrating MyPainHub as a supplementary resource into the physiotherapy program curriculum and ensuring the information provided aligns with the learning objectives of the specific unit of study with which it is associated. Effective incorporation of MyPainHub for students may be achieved through thoughtful planning and implementation, integrating into specific lesson plans including activities such as group discussions followed by monitoring of student engagement via Learning Management Systems or linked to assessable content [64].

As with any web-based resource, updates to MyPainHub are ongoing with the maintenance of the website being the responsibility of the research team and the digital agency. Addition of information about other conditions (eg, hip pain) is underway and incorporating additional features such as more dynamic elements (eg, carousels and expandable sections) and gamification to increase engagement [65] are likely to be important features of future updates. Limitations of this study include that we did not control the time between participants viewing the website and asking for their feedback. Despite efforts to engage general practitioners and other medical professionals, the majority of those involved in the development of MyPainHub were allied HCPs. Further work is needed to determine whether MyPainHub, or specific sections of MyPainHub, would be useful for general practitioners. Further work is also required to implement MyPainHub in other primary (eg, community-based care) and tertiary care settings.

Conclusion

This study details the development, implementation, and evaluation of MyPainHub, a web-based resource designed to support guideline-based care for musculoskeletal conditions. Stakeholder engagement throughout the process contributed to its acceptability, credibility, and ease of use. While MyPainHub effectively tailors information based on prognostic tools to support both patients and health care providers, further efforts are needed to facilitate its integration into clinical practice. Training for health care professionals, particularly in how to apply MyPainHub in diverse settings, is likely essential for greater adoption.

Acknowledgments

We would like to thank the study participants for contributing their time and knowledge to this study.

This work was supported by The National Health and Medical Research Council (NHMRC) and Sydney Health Partners Implementation Science Academy. The funder had no role in the design or conduct of the study. TR was supported by an NHMRC Career Development Fellowship. KLB and PH are supported by NHMRC Investigator grants. MS receives unrestricted grant funding from the Motor Accident Insurance Commission of Queensland.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Healthcare professional survey.
[PNG File , 63 KB - formative_v9i1e63780_app1.png]

References

1. Briggs AM, Woolf AD, Dreinhöfer K, Homb N, Hoy DG, Kopansky-Giles D, et al. Reducing the global burden of musculoskeletal conditions. Bull World Health Organ 2018;96(5):366-368 [FREE Full text] [doi: 10.2471/BLT.17.204891] [Medline: 29875522]
2. Hoy W. Australian burden of disease study: impact and causes of illness and death in Australia 2011. In: Australian burden of disease study series. Canberra: Australian Institute of Health and Welfare; 2011.
3. Cieza A, Causey K, Kamenov K, Hanson SW, Chatterji S, Vos T. Global estimates of the need for rehabilitation based on the global burden of disease study 2019: a systematic analysis for the global burden of disease study 2019. Lancet 2021;396(10267):2006-2017 [FREE Full text] [doi: 10.1016/S0140-6736(20)32340-0] [Medline: 33275908]
4. El-Tallawy SN, Nalamasu R, Salem GI, LeQuang JAK, Pergolizzi JV, Christo PJ. Management of musculoskeletal pain: an update with emphasis on chronic musculoskeletal pain. Pain Ther 2021;10(1):181-209 [FREE Full text] [doi: 10.1007/s40122-021-00235-2] [Medline: 33575952]
5. de Luca K, Briggs AM, French SD, Ferreira ML, Cross M, Blyth F, et al. Disability burden due to musculoskeletal conditions and low back pain in Australia: findings from GBD 2019. Chiropr Man Therap 2022;30(1):22 [FREE Full text] [doi: 10.1186/s12998-022-00434-4] [Medline: 35505334]

XSL•FO
RenderX

6. Wallace J, Nwosu B, Clarke M. Barriers to the uptake of evidence from systematic reviews and meta-analyses: a systematic review of decision makers' perceptions. BMJ Open 2012;2(5):e001220 [FREE Full text] [doi: 10.1136/bmjopen-2012-001220] [Medline: 22942232]
7. Paci M, Faedda G, Ugolini A, Pellicciari L. Barriers to evidence-based practice implementation in physiotherapy: a systematic review and meta-analysis. Int J Qual Health Care 2021;33(2). [doi: 10.1093/intqhc/mzab093] [Medline: 34110410]
8. Oliveira CB, Maher CG, Pinto RZ, Traeger AC, Lin CC, Chenot J, et al. Clinical practice guidelines for the management of non-specific low back pain in primary care: an updated overview. Eur Spine J 2018;27(11):2791-2803. [doi: 10.1007/s00586-018-5673-2] [Medline: 29971708]
9. Malfliet A, Ickmans K, Huysmans E, Coppieters I, Willaert W, Bogaert WV, et al. Best evidence rehabilitation for chronic pain part 3: low back pain. J Clin Med 2019;8(7):1063 [FREE Full text] [doi: 10.3390/jcm8071063] [Medline: 31331087]
10. Blanpied PR, Gross AR, Elliott JM, Devaney LL, Clewley D, Walton DM, et al. Neck pain: revision 2017. J Orthop Sports Phys Ther 2017;47(7):A1-A83. [doi: 10.2519/jospt.2017.0302] [Medline: 28666405]
11. Sterling M, de Zoete RMJ, Coppieters I, Farrell SF. Best evidence rehabilitation for chronic pain part 4: neck pain. J Clin Med 2019;8(8):1219 [FREE Full text] [doi: 10.3390/jcm8081219] [Medline: 31443149]
12. Guidelines for the management of acute whiplash associated disorders for health professionals. 3rd edition. State Insurance Regulatory Authority (SIRA). 2014. URL: https://www.sira.nsw.gov.au/resources-library/motor-accident-resources/publications/for-professionals/whiplash-resources/SIRA08104-Whiplash-Guidelines-1117-396479.pdf [accessed 2023-06-23]
13. Kolasinski SL, Neogi T, Hochberg MC, Oatis C, Guyatt G, Block J, et al. 2019 American college of rheumatology/arthritis foundation guideline for the management of osteoarthritis of the hand, hip, and knee. Arthritis Rheumatol 2020;72(2):220-233 [FREE Full text] [doi: 10.1002/art.41142] [Medline: 31908163]
14. van Doormaal MCM, Meerhoff GA, Vliet Vlieland TPM, Peter WF. A clinical practice guideline for physical therapy in patients with hip or knee osteoarthritis. Musculoskeletal Care 2020;18(4):575-595. [doi: 10.1002/msc.1492] [Medline: 32643252]
15. Guideline for the management of knee and hip osteoarthritis. 2nd ed. East Melbourne, Vic: RACGP. The Royal Australian College of General Practitioners. 2018. URL: https://www.racgp.org.au/getattachment/71ab5b77-afdf-4b01-90c3-04f61a910be6/Guideline-for-the-management-of-knee-and-hip-osteoarthritis.aspx [accessed 2023-06-23]
16. Rice D, McNair P, Huysmans E, Letzen J, Finan P. Best evidence rehabilitation for chronic pain part 5: osteoarthritis. J Clin Med 2019;8(11):1769 [FREE Full text] [doi: 10.3390/jcm8111769] [Medline: 31652929]
17. Zadro J, O'Keeffe M, Maher C. Do physical therapists follow evidence-based guidelines when managing musculoskeletal conditions? Systematic review. BMJ Open 2019;9(10):e032329 [FREE Full text] [doi: 10.1136/bmjopen-2019-032329] [Medline: 31591090]
18. Bahns C, Happe L, Thiel C, Kopkow C. Physical therapy for patients with low back pain in Germany: a survey of current practice. BMC Musculoskelet Disord 2021;22(1):563 [FREE Full text] [doi: 10.1186/s12891-021-04422-2] [Medline: 34147077]
19. Logan GS, Dawe RE, Aubrey-Bassler K, Coombs D, Parfrey P, Maher C, et al. Are general practitioners referring patients with low back pain for CTs appropriately according to the guidelines: a retrospective review of 3609 medical records in Newfoundland using routinely collected data. BMC Fam Pract 2020;21(1):236 [FREE Full text] [doi: 10.1186/s12875-020-01308-5] [Medline: 33208086]
20. Epstein-Sher S, Jaffe D, Lahad A. Are they complying? Physicians' knowledge, attitudes, and readiness to change regarding low back pain treatment guideline adherence. Spine (Phila Pa 1976) 2017;42(4):247-252. [doi: 10.1097/BRS.0000000000001714] [Medline: 28207666]
21. Avoundjian T, Gidwani R, Yao D, Lo J, Sinnott P, Thakur N, et al. Evaluating two measures of lumbar spine MRI overuse: administrative data versus chart review. J Am Coll Radiol 2016;13(9):1057-1066. [doi: 10.1016/j.jacr.2016.04.013] [Medline: 27344246]
22. Williams CM, Maher CG, Hancock MJ, McAuley JH, McLachlan AJ, Britt H, et al. Low back pain and best practice care: a survey of general practice physicians. Arch Intern Med 2010;170(3):271-277. [doi: 10.1001/archinternmed.2009.507] [Medline: 20142573]
23. Buchbinder R, Underwood M, Hartvigsen J, Maher CG. The Lancet series call to action to reduce low value care for low back pain: an update. Pain 2020;161 Suppl 1(1):S57-S64 [FREE Full text] [doi: 10.1097/j.pain.0000000000001869] [Medline: 33090740]
24. Kharel P, Zadro JR, Maher CG. Physiotherapists can reduce overuse by Choosing Wisely. J Physiother 2021;67(3):151-155 [FREE Full text] [doi: 10.1016/j.jphys.2021.06.006] [Medline: 34154950]
25. Rutten GM, Degen S, Hendriks EJ, Braspenning J, Harting J, Oostendorp R. Adherence to clinical practice guidelines for low back pain in physical therapy: do patients benefit? Phys Ther 2010;90(8):1111-1122. [doi: 10.2522/ptj.20090173] [Medline: 20488978]
26. van Staalduinen DJ, van den Bekerom P, Groeneveld S, Kidanemariam M, Stiggelbout AM, van den Akker-van Marle ME. The implementation of value-based healthcare: a scoping review. BMC Health Serv Res 2022;22(1):270 [FREE Full text] [doi: 10.1186/s12913-022-07489-2] [Medline: 35227279]

XSL•FO
RenderX

27. Macpherson S, Rebbeck TR, Coates S, Evans K. Referral practices of recent graduate and experienced physiotherapists working in Australian primary care for people with musculoskeletal conditions. Musculoskelet Sci Pract 2023;64:102745 [FREE Full text] [doi: 10.1016/j.msksp.2023.102745] [Medline: 36924580]
28. Gorgon E, Maka K, Kam A, Nisbet G, Sullivan J, Regan G, et al. Needs assessment for health service design for people with back pain in a hospital setting: a qualitative study. Health Expect 2022;25(2):721-731 [FREE Full text] [doi: 10.1111/hex.13419] [Medline: 35150036]
29. Al Zoubi FM, Menon A, Mayo NE, Bussières AE. The effectiveness of interventions designed to increase the uptake of clinical practice guidelines and best practices among musculoskeletal professionals: a systematic review. BMC Health Serv Res 2018;18(1):435 [FREE Full text] [doi: 10.1186/s12913-018-3253-0] [Medline: 29884165]
30. Maher C, Sherrington C, Elkins M, Herbert R, Moseley A. Challenges for evidence-based physical therapy: accessing and interpreting high-quality evidence on therapy. Phys Ther 2004;84(7):644-654. [doi: 10.1093/ptj/84.7.644]
31. da Silva TM, Costa LDCM, Garcia AN, Costa LOP. What do physical therapists think about evidence-based practice? A systematic review. Man Ther 2015;20(3):388-401. [doi: 10.1016/j.math.2014.10.009] [Medline: 25458142]
32. Kummervold PE, Chronaki CE, Lausen B, Prokosch H, Rasmussen J, Santana S, et al. eHealth trends in Europe 2005-2007: a population-based survey. J Med Internet Res 2008;10(4):e42 [FREE Full text] [doi: 10.2196/jmir.1023] [Medline: 19017584]
33. Hodges PW, Setchell J, Nielsen M. An internet-based consumer resource for people with low back pain (MyBackPain): development and evaluation. JMIR Rehabil Assist Technol 2020;7(1):e16101 [FREE Full text] [doi: 10.2196/16101] [Medline: 32229467]
34. Scurlock-Evans L, Upton P, Upton D. Evidence-based practice in physiotherapy: a systematic review of barriers, enablers and interventions. Physiotherapy 2014;100(3):208-219. [doi: 10.1016/j.physio.2014.03.001] [Medline: 24780633]
35. Jette DU, Bacon K, Batty C, Carlson M, Ferland A, Hemingway R, et al. Evidence-based practice: beliefs, attitudes, knowledge, and behaviors of physical therapists. Phys Ther 2003;83(9):786-805. [doi: 10.1093/ptj/83.9.786]
36. Harding K, Porter J, Horne-Thompson A, Donley E, Taylor NF. Not enough time or a low priority? Barriers to evidence-based practice for allied health clinicians. J Contin Educ Health Prof 2014;34(4):224-231. [doi: 10.1002/chp.21255] [Medline: 25530292]
37. Reis FJJ, Nijs J, Parker R, Sharma S, Wideman TH. Culture and musculoskeletal pain: strategies, challenges, and future directions to develop culturally sensitive physical therapy care. Braz J Phys Ther 2022;26(5):100442 [FREE Full text] [doi: 10.1016/j.bjpt.2022.100442] [Medline: 36209626]
38. Costa N, Nielsen M, Jull G, Claus AP, Hodges PW. Low back pain websites do not meet the needs of consumers: a study of online resources at three time points. Health Inf Manag 2020;49(2-3):137-149. [doi: 10.1177/1833358319857354] [Medline: 31266366]
39. Bourne JA, Dziedzic K, Morris SJ, Jones PW, Sim J. Survey of the perceived professional, educational and personal needs of physiotherapists in primary care and community settings. Health Soc Care Community 2007;15(3):231-237. [doi: 10.1111/j.1365-2524.2007.00677.x] [Medline: 17444986]
40. Peterson S, Rainey N, Weible K. Who writes this stuff? Musculoskeletal information quality and authorship of popular health websites: a systematic review. Musculoskelet Sci Pract 2022;60:102563. [doi: 10.1016/j.msksp.2022.102563] [Medline: 35453015]
41. Wanless B, Bermingham C, Yeates N, O'Rahelly J. Creating a valuable patient information resource for MSK conditions: By not re-inventing the wheel. Physiotherapy 2021;113:e120-e121. [doi: 10.1016/j.physio.2021.10.106]
42. Translational research framework. Sax Institute. 2016. URL: https://www.medicalresearch.nsw.gov.au/app/uploads/2019/02/Translational-Research-Grants-Scheme-translation-research-framework.pdf [accessed 2023-07-20]
43. Bandong AN, Mackey M, Leaver A, Ingram R, Sterling M, Ritchie C, et al. An interactive website for whiplash management (My Whiplash Navigator): process evaluation of design and implementation. JMIR Form Res 2019;3(3):e12216 [FREE Full text] [doi: 10.2196/12216] [Medline: 31452515]
44. Clarkson P, Vassilev I, Rogers A, Brooks C, Wilson N, Lawson J, et al. Integrating a web-based self-management tool (Managing Joint Pain on the Web and Through Resources) for people with osteoarthritis-related joint pain with a web-based social network support tool (Generating Engagement in Network Involvement): design, development, and early evaluation. JMIR Form Res 2020;4(11):e18565 [FREE Full text] [doi: 10.2196/18565] [Medline: 33242011]
45. Hodges PW, Hall L, Setchell J, French S, Kasza J, Bennell K, et al. Effect of a consumer-focused website for low back pain on health literacy, treatment choices, and clinical outcomes: Randomized controlled trial. J Med Internet Res 2021;23(6):e27860 [FREE Full text] [doi: 10.2196/27860] [Medline: 34128822]
46. Wang X, Urban H, Bennell KL, Dickson C, Dobson F, Fransen M, et al. My Joint Pain, a web-based resource, effects on education and quality of care at 24 months. BMC Musculoskelet Disord 2020;21(1):79 [FREE Full text] [doi: 10.1186/s12891-020-3074-2] [Medline: 32028927]
47. van der Zee-Neuen A, Putrik P, Ramiro S, Keszei A, de Bie R, Chorus A, et al. Impact of chronic diseases and multimorbidity on health and health care costs: the additional role of musculoskeletal disorders. Arthritis Care Res (Hoboken) 2016;68(12):1823-1831 [FREE Full text] [doi: 10.1002/acr.22913] [Medline: 27111195]

XSL•FO
RenderX

48. Lin I, Wiles L, Waller R, Goucke R, Nagree Y, Gibberd M, et al. What does best practice care for musculoskeletal pain look like? Eleven consistent recommendations from high-quality clinical practice guidelines: systematic review. Br J Sports Med 2020;54(2):79-86. [doi: 10.1136/bjsports-2018-099878] [Medline: 30826805]
49. Rebbeck T, Evans K, Ferreira P, Beales D, Sterling M, Bennell KL, et al. Implementation of a novel stratified PAthway of CarE for common musculoskeletal (MSK) conditions in primary care: protocol for a multicentre pragmatic randomised controlled trial (the PACE MSK trial). BMJ Open 2021;11(12):e057705 [FREE Full text] [doi: 10.1136/bmjopen-2021-057705] [Medline: 37039086]
50. Glasgow RE, Vogt TM, Boles SM. Evaluating the public health impact of health promotion interventions: the RE-AIM framework. Am J Public Health 1999;89(9):1322-1327. [doi: 10.2105/ajph.89.9.1322] [Medline: 10474547]
51. Baiardini I, Braido F, Bonini M, Compalati E, Canonica GW. Why do doctors and patients not follow guidelines? Curr Opin Allergy Clin Immunol 2009;9(3):228-233. [doi: 10.1097/ACI.0b013e32832b4651] [Medline: 19390434]
52. Spangenberg JH, Heong KL, Klotzbücher A, Klotzbücher T, Nguyen QA, Tekken V, et al. Doing what with whom? Stakeholder analysis in a large transdisciplinary research project in South-East Asia. Paddy Water Environ 2018;16(2):321-337. [doi: 10.1007/s10333-018-0634-2]
53. Kang K, Evans K, Simic M, Ferreira P, Bandong AN, Coates S, et al. Impact of an interactive workshop on specialist physiotherapists' practice when implementing a new clinical care pathway for people with musculoskeletal conditions. Musculoskelet Sci Pract 2022;57:102466. [doi: 10.1016/j.msksp.2021.102466] [Medline: 34742051]
54. Braun V, Clarke V. Using thematic analysis in psychology. Qualitative Research in Psychology 2006;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
55. Linton SJ, Boersma K. Early identification of patients at risk of developing a persistent back problem: the predictive validity of the Örebro Musculoskeletal Pain Questionnaire. Clin J Pain 2003;19(2):80-86. [doi: 10.1097/00002508-200303000-00002] [Medline: 12616177]
56. Campbell P, Hill J, Protheroe J, Afolabi E, Lewis M, Beardmore R, et al. Keele aches and pains study protocol: validity, acceptability, and feasibility of the Keele STarT MSK tool for subgrouping musculoskeletal patients in primary care. J Pain Res 2016;9:807-818 [FREE Full text] [doi: 10.2147/JPR.S116614] [Medline: 27789972]
57. Ritchie C, Hendrikz J, Jull G, Elliott J, Sterling M. External validation of a clinical prediction rule to predict full recovery and ongoing moderate/severe disability following acute whiplash injury. J Orthop Sports Phys Ther 2015;45(4):242-250. [doi: 10.2519/jospt.2015.5642] [Medline: 25827122]
58. Sterling M, Ritchie C, Rebbeck T, Cameron ID, Griffin A, Jagnoor J, et al. Comparison of the accuracy of WhipPredict to that of a modified version of the Short-Form Örebro Musculoskeletal Pain Screening Questionnaire to predict poor recovery after whiplash injury. J Orthop Sports Phys Ther 2021;51(5):207-215. [doi: 10.2519/jospt.2021.9987] [Medline: 33607916]
59. Suri P, Delaney K, Rundell SD, Cherkin DC. Predictive validity of the STarT Back Tool for risk of persistent disabling back pain in a U.S. primary care setting. Arch Phys Med Rehabil 2018;99(8):1533-1539.e2 [FREE Full text] [doi: 10.1016/j.apmr.2018.02.016] [Medline: 29625095]
60. Fuhro F, Fagundes F, Manzoni A, Costa L, Cabral C. Örebro Musculoskeletal Pain Screening Questionnaire Short-Form and STarT Back Screening Tool: Correlation and agreement analysis. Spine (Phila Pa 1976) 2016;41(15):E931-E936. [doi: 10.1097/BRS.0000000000001415] [Medline: 26720177]
61. Moore G, Wilding H, Gray K, Castle D. Participatory methods to engage health service users in the development of electronic health resources: systematic review. J Particip Med 2019;11(1):e11474 [FREE Full text] [doi: 10.2196/11474] [Medline: 33055069]
62. Rebbeck T, Maher CG, Refshauge KM. Evaluating two implementation strategies for whiplash guidelines in physiotherapy: a cluster randomised trial. Aust J Physiother 2006;52(3):165-174. [doi: 10.1016/s0004-9514(06)70025-3] [Medline: 16942451]
63. Goorts K, Dizon J, Milanese S. The effectiveness of implementation strategies for promoting evidence informed interventions in allied healthcare: a systematic review. BMC Health Serv Res 2021;21(1):241 [FREE Full text] [doi: 10.1186/s12913-021-06190-0] [Medline: 33736631]
64. Kenwright D, Dai W, Osborne E, Gladman T, Gallagher P, Grainger R. “Just tell me what I need to know to pass the exam!” Can active flipped learning overcome passivity? TAPS 2017;2(1):1-6. [doi: 10.29060/TAPS.2017-2-1/OA1007]
65. Alfieri FM, da Silva Dias C, de Oliveira NC, Battistella LR. Gamification in musculoskeletal rehabilitation. Curr Rev Musculoskelet Med 2022;15(6):629-636 [FREE Full text] [doi: 10.1007/s12178-022-09797-w] [Medline: 36301514]

Abbreviations

HCP: health care professional
LBP: low back pain
MSK: muscoluskeletal
PACE: PAthway of CarE
RE-AIM: Reach, Effectiveness, Adoption, Implementation, and Maintenance
SF-ÖMPSQ: Short Form Örebro Musculoskeletal Pain Screening Questionnaire

XSL•FO
RenderX

WAD: whiplash associated disorders
WSLHD: Western Sydney Local Health District

Edited by A Mavragani; submitted 05.07.24; peer-reviewed by N Khan, J Pearson; comments to author 09.10.24; revised version received 18.11.24; accepted 28.12.24; published 24.02.25.

Please cite as:
Evans K, Ko J, Ceprnja D, Maka K, Beales D, Sterling M, Bennell KL, Jull G, Hodges PW, McKay MJ, Rebbeck TJ
Development and Implementation of MyPainHub, a Web-Based Resource for People With Musculoskeletal Conditions and Their Health Care Professionals: Mixed Methods Study
JMIR Form Res 2025;9:e63780
URL: https://formative.jmir.org/2025/1/e63780
doi:10.2196/63780
PMID:

©Kerrie Evans, Jonathan Ko, Dragana Ceprnja, Katherine Maka, Darren Beales, Michele Sterling, Kim L Bennell, Gwendolen Jull, Paul W Hodges, Marnee J McKay, Trudy J Rebbeck. Originally published in JMIR Formative Research (https://formative.jmir.org), 24.02.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Feasibility of Large Language Models in Verbal Comprehension Assessment: Mixed Methods Feasibility Study

Dorit Hadar-Shoval[1], BA, MA, PhD; Maya Lvovsky[1], BA; Kfir Asraf[1], BA, MA; Yoav Shimoni[2,3], BA, MA, PHD; Zohar Elyoseph[4,5], BA, MA, PHD

[1]Department of Psychology, Max Stern Academic College of Emek Yezreel, Afula, Israel
[2]School of Psychology, University of Haifa, Haifa, Israel
[3]Department of Education and Psychology, Open University of Israel, Raanana, Israel
[4]Department of Brain Sciences, Faculty of Medicine, Imperial College London, London, United Kingdom
[5]School of Counseling and Human Development, Faculty of Education, University of Haifa, Haifa, Israel

Corresponding Author:
Zohar Elyoseph, BA, MA, PHD
Department of Brain Sciences
Faculty of Medicine
Imperial College London
Reynolds Building, St Dunstan's Road
Charing Cross Hospital, Hammersmith
London, W6 8RP
United Kingdom
Phone: 972 547836088
Email: Zohar.j.a@gmail.com

Abstract

Background: Cognitive assessment is an important component of applied psychology, but limited access and high costs make these evaluations challenging.

Objective: This study aimed to examine the feasibility of using large language models (LLMs) to create personalized artificial intelligence–based verbal comprehension tests (AI-BVCTs) for assessing verbal intelligence, in contrast with traditional assessment methods based on standardized norms.

Methods: We used a within-participants design, comparing scores obtained from AI-BVCTs with those from the Wechsler Adult Intelligence Scale (WAIS-III) verbal comprehension index (VCI). In total, 8 Hebrew-speaking participants completed both the VCI and AI-BVCT, the latter being generated using the LLM Claude.

Results: The concordance correlation coefficient (CCC) demonstrated strong agreement between AI-BVCT and VCI scores (Claude: CCC=.75, 90% CI 0.266-0.933; GPT-4: CCC=.73, 90% CI 0.170-0.935). Pearson correlations further supported these findings, showing strong associations between VCI and AI-BVCT scores (Claude: r=.84, P<.001; GPT-4: r=.77, P=.02). No statistically significant differences were found between AI-BVCT and VCI scores (P>.05).

Conclusions: These findings support the potential of LLMs to assess verbal intelligence. The study attests to the promise of AI-based cognitive tests in increasing the accessibility and affordability of assessment processes, enabling personalized testing. The research also raises ethical concerns regarding privacy and overreliance on AI in clinical work. Further research with larger and more diverse samples is needed to establish the validity and reliability of this approach and develop more accurate scoring procedures.

(JMIR Form Res 2025;9:e68347) doi:10.2196/68347

KEYWORDS

large language models; verbal comprehension assessment; artificial intelligence; AI in psychodiagnostics; personalized intelligence tests; verbal comprehension index; Wechsler Adult Intelligence Scale; WAIS-III; psychological test validity; ethics in computerized cognitive assessment

XSL•FO
RenderX

Introduction

Cognitive assessment is one of the core skills of applied psychologists, and it is considered one of the most important contributions of scientific psychology to clinical practice [1]. Intelligence tests are widely used to measure individuals' cognitive abilities. They provide a standardized measure of a person's intellectual capacities, aiding in educational placement and career selection, and identifying potential cognitive impairments. Intellectual assessment measures serve as valuable tools in psychology, education, and research, helping identify gifted individuals, track cognitive development, and evaluate the effectiveness of interventions [1]. Intelligence tests have their limitations [2] but their consistent use and refinement over time have contributed to our understanding of human intelligence and its variations. Traditionally, the development of an intelligence test involves an intricate process and the construction of age- and gender-based norms based on extensive data collected from a sizable sample representative of the population [3].

Intelligence tests are designed to assess various cognitive abilities, such as logical reasoning, problem-solving skills, spatial awareness, and verbal comprehension [4]. The verbal comprehension index (VCI) is known for its strong correlation with the g factor (the hypothesized construct of general intelligence) [5], and it measures the capability to use language for reasoning, comprehension, and understanding social norms. The VCI includes three subtests, that are, similarities, vocabulary, and information [5].

The significance of a comprehensive psychological assessment including measures of verbal intelligence cannot be overstated. [6-10]. The assessment of verbal intelligence also extends to personality traits crucial for interpersonal interactions, such as the recognition of emotions and the ability to express oneself effectively [11]. The comprehensive psychological evaluation of various populations under different conditions requires a nuanced understanding of verbal intelligence.

The administration of intelligence tests and the subsequent interpretation of the results are typically undertaken by psychologists. These professionals usually possess at least a master's degree in psychology or in a related field, which qualifies them to conduct the tests and analyze the results accurately and ethically. Intellectual testing is usually administered in a standardized form, where the individuals being tested are presented with similar tasks under similar conditions. The standardized administration process helps ensure an ethical process and produces valid results, leading to accurate interpretations [1,10].

The field of intelligence testing still heavily relies on conventional paper-and-pencil tests, despite recent efforts to embrace technological advancements. Some current intelligence testing measures, such as the fifth edition of the Wechsler Intelligence Scale for Children [12], presuppose the use of tablets by clinicians to improve the administration and scoring procedures. Globally, however, the use of technology in psychological assessment is still developing and to date has had relatively modest influence.

At the end of 2022, Open AI launched ChatGPT-3.5, a generative artificial intelligence (GenAI) product based on a large language model (LLM) trained on abundant text data and capable of generating human-like responses to text-based inputs [13]. This open access application set a record as the fastest-growing consumer application in history [14], making significant contributions to academia [13], psychology [14-27], medicine [28], and programming [29]. In March 2023, Anthropic launched Claude, another GenAI product based on LLM, with a similar interface to that of ChatGPT but with Hebrew language capabilities that are considered to be more nuanced and accurate than those of ChatGPT.

While these technological advances offer promising opportunities for psychological assessment, it is essential to acknowledge their current limitations. LLMs cannot replace comprehensive clinical evaluations, as they lack the ability to observe crucial behavioral cues, interpret complex clinical contexts, or make nuanced professional judgments. Therefore, any application of AI in cognitive testing must be viewed as a potential supportive tool requiring professional oversight.

This pilot study represents an innovative endeavor in the field of GenAI and its application in psychological assessment, serving as a proof-of-concept investigation into the feasibility of using LLMs for verbal comprehension testing. Rather than aiming to validate these tools across a large population, our primary objective was to demonstrate whether LLMs can, in principle, generate personalized verbal comprehension assessments that produce results comparable to standardized tests. Central to this study is the development of a personalized cognitive test comprising 30 questions without human norms (artificial intelligence [AI]–based verbal comprehension test [AI-BVCT]). Unlike standard testing procedures, this approach does not rely on pre-established diagnostic tools but uses the algorithmic capabilities of the LLM to encode, assess, and calculate verbal intelligence scores.

A critical aspect of this study is the comparative analysis conducted between the AI-BVCT results and those obtained from the VCI, under the supervision of an expert psychologist. To establish the basic viability of this approach, we conducted a focused comparison between AI-generated tests and traditional assessments on a small sample. We hypothesized that the differences between the AI-BVCT and traditional VCI scores would be within half an (SD 7.5), indicating a high level of accuracy and that the correlation between the AI-BVCT and the VCI would be high, demonstrating accurate cognitive ranking across participants. By focusing exclusively on verbal comprehension abilities, the study aims to establish proof of concept for how LLMs can contribute to psychological testing, emphasizing the potential for new personalized and dynamic testing methodologies.

Methods

Overview

This study used 2 LLMs: GPT-4 (OpenAI) and Claude (Anthropic). Data collection occurred in September 2023. The selection of these models was based on several key

XSL•FO
RenderX

considerations. At the time of the study, these models represented the most advanced capabilities in natural language processing, particularly for the Hebrew language. Claude demonstrated superior proficiency in Hebrew language processing and a nuanced understanding of cultural context, making it our primary choice for generating the AI-BVCT. Both models were used to evaluate participants' responses, allowing for a comparison between different approaches to language model development. The use of models from different leading companies (OpenAI and Anthropic) enabled us to examine various approaches to language model development and training, providing a broader context for our findings. It is important to note that this selection limits the study to Western commercial models and does not include open-source models or those developed in other cultural contexts. This limitation will be discussed further in the paper. We did not alter their default settings when using these models. We used the standard chat interface provided by the companies, without special adjustments to hyperparameters such as temperature or top-k. An additional limitation of these tools is their ownership by commercial entities, which restricts full transparency of their training and development processes. It is also worth noting that the field of LLMs is rapidly evolving, and more advanced models have become available since the study was conducted.

Participants

This pilot study included 8 native Hebrew-speaking volunteer participants (4 females), aged 28.5 (SD 2.13) years. The participants were recruited through random sampling, facilitated by a research assistant who distributed targeted recruitment notices to undergraduate and graduate psychology students at Max Stern Yezreel Valley College. The recruitment process extended over a 2-month period, during which participants volunteered without compensation. The recruitment notices specifically outlined the study's objectives and requirements, ensuring informed participation. All participants who initially enrolled completed the study in its entirety, with no withdrawals recorded during the research period. The participants were selected from the student population due to their accessibility and their presumed average to above-average cognitive abilities, which were deemed appropriate for this pilot study. However, we acknowledge the limitations of this small, nonrepresentative sample, which are addressed in detail in the limitations section of this paper.

Research Design and Procedure

The study used a within-participants design, with each participant undergoing 2 cognitive tests: the VCI from the WAIS-III, and a personalized AI-BVCT generated by the LLM Claude system.

After each participant underwent the VCI test, the participant's age and gender were entered into the LLM Claude system, together with a prompt to generate a personalized AI-BVCT to measure the participant's verbal intelligence. In the next stage, the participant was required to respond to the AI-BVCT. The participant's responses were then fed into 2 separate LLMs, Claude and GPT-4, with a prompt containing scoring instructions. To ensure the reliability of the scoring across different runs, each participant's responses were entered into each of the 2 models 10 times, using a separate tab for each iteration. As a result, each participant received a total of 20 scores (10 from each model) on the AI-BVCT.

Finally, we compared the participant's results on the VCI test and their performance on the AI-BVCT to examine the degree of concordance and accuracy of the test developed by the LLM system. A schematic representation of the study procedure is provided in Figure 1.

XSL•FO
RenderX

Figure 1. Schematic representation of the study procedure for assessing verbal comprehension using AI-based tests in a proof-of-concept study of Hebrew-speaking university students (N=8) in Israel, September 2023. AI-BVCT: artificial intelligence-based verbal comprehension tests; LLM: large language models; VCI: verbal comprehension index.

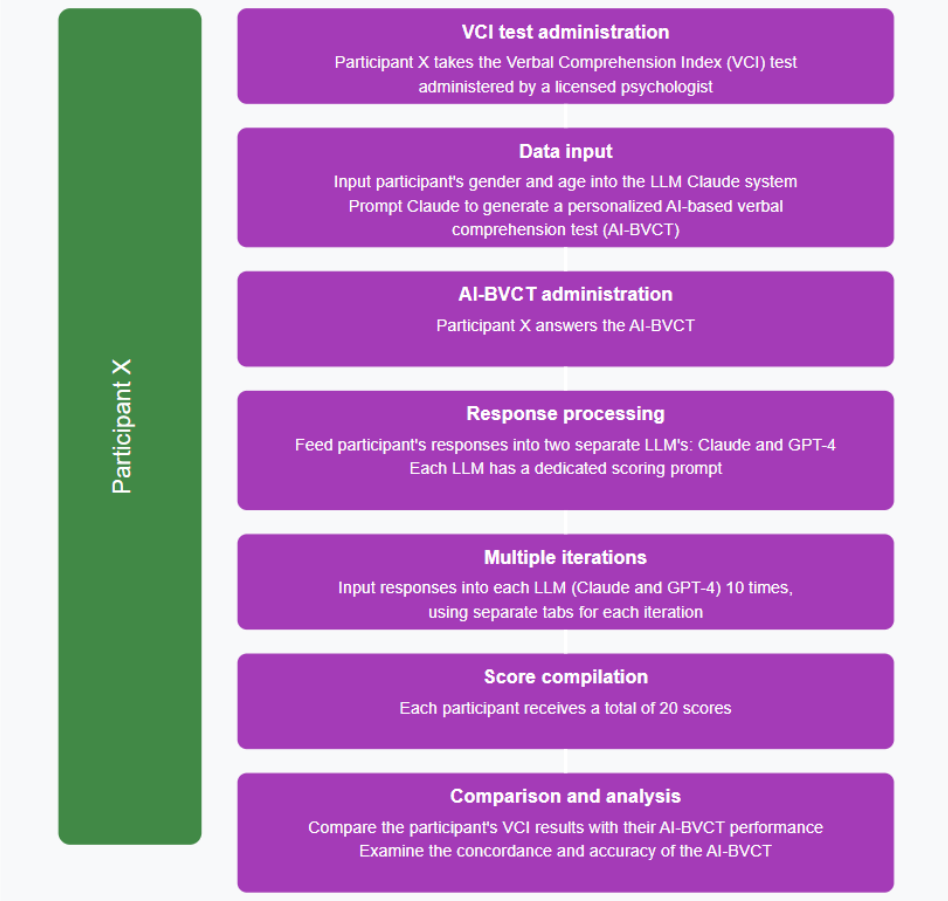

Materials

The VCI

Overview

The VCI is an index designed to assess verbal abilities, such as verbal concept formation and reasoning skills. This index is part of the Wechsler Adult Intelligence Scale-3 (WAIS-III; [30]), which measures various cognitive abilities in adults. VCI scores are normalized with a mean of 100 and an SD of 15 [31]. The WAIS-III was used in this study as it represents the most recent version of the WAIS-III that has been officially translated, adapted, and normalized for Hebrew-speaking populations in Israel. While newer versions exist, they have not yet been standardized for Hebrew-speaking populations.

The index consists of three subtests: (1) similarities, which measures the ability to identify commonalities between terms; (2) vocabulary, which measures the understanding and use of words; and (3) information, which measures general knowledge and the ability to provide factual information [5]. The VCI score is calculated by combining the 3 subtests.

VCI Scoring

Scoring was performed according to the accepted scoring manual. Each question was assigned a score of 0-2 based on the acceptable responses provided in the scoring guide. The calculation of standard scores and results was conducted in accordance with the guidelines established in the WAIS manual.

Personalized AI-Based Verbal Comprehension Test

Generating the Personalized AI-BVCT

The LLM Claude (Anthropic, Ltd) was used for generating a personalized cognitive test for each participant. We used Claude because at the time the study was conducted its performance in Hebrew was considered superior to those of other LLMs, such as GPT-4 or Bard. To generate the AI-BVCT, we composed a request for Claude to develop a 30-item verbal intelligence test, having provided the participant's age and gender. The inclusion of age information was intended to ensure age-appropriate content and difficulty level, similar to standardized tests. Gender information was provided to account for the gendered nature of the Hebrew language in question phrasing, although the content and difficulty of questions were not intended to differ based on gender. To enhance the quality of the test, we also provided examples of questions to guide Claude in shaping questions with the appropriate structure and difficulty level. Specifically, we provided 3 sample items for each of the 3 intelligence levels, which consist of low (<70), medium (90-110), and high (>125) for each of the subtests.

Personalized AI-BVCT Prompt

The prompt that was entered into Claude are shown in Textbox 1.

XSL•FO
RenderX

Textbox 1. Prompt for constructing a verbal intelligence test with artificial intelligence.

Cognitive Assessment

- Act as a professor of psychology specializing in psychodiagnostics and psychometrics. Please create a 30-question test that can assess the verbal intelligence score of an X-year-old Hebrew-speaking man or woman. Try to create questions whose answers will be as similar as possible to the results of the verbal intelligence WAIS 3 test in the verbal intelligence index, which includes a vocabulary test (defining words at increasingly difficult levels), common denominator (finding a semantic common denominator between words at increasingly difficult levels), and general knowledge (general knowledge at increasingly difficult levels). Construct the questions so that they can give as accurate a score as possible in the range of 50-150. This means that some questions should be very easy and some very difficult. Build questions that can be answered via the chat box, without any human intervention, and without questions requiring additional tools other than the chat itself.

To refine your ability to build questions at different difficulty levels, you are given examples.

- Vocabulary: (1) Here are some words for which the level of difficulty to define for a 30-year-old adult is hard and reflects a higher than average verbal intelligence (verbal intelligence level 125 and above): obelisk, orthography, ether, allegory, atom. (2) Here are some words for which the level of difficulty to define for a 30-year-old adult is medium and reflects average verbal intelligence (verbal intelligence level 90-110): fortification, settlement, takeover, degeneration, outsmarting. (3) Here are some words for which the level of difficulty to define for a 30-year-old adult is easy and reflects much lower-than-average verbal intelligence (verbal intelligence level 70 and below: table, chair, cupboard, sofa, television
- SimList Paragraphilarities: (1) Here are some words for which the level of difficulty to find a common denominator for a 30-year-old adult is hard and reflects higher than average verbal intelligence (verbal intelligence level 125 and above): clothes and food, dream and reality, life and death, light and darkness, sky and earth. (2) Here are some words for which the level of difficulty to find a common denominator for a 30-year-old adult is medium and reflects average verbal intelligence (verbal intelligence level 90-110): wave and wind, freedom and imprisonment, land and sea, inside and outside, number and letter. (3) Here are some words for which the level of difficulty to find a common denominator for a 30-year-old adult is easy and reflects much lower than average verbal intelligence (verbal intelligence level 70 and below). Fork and knife, ball and football, rain and snow, dog and cat, pants and shirt.
- General Knowledge: (1) Here are some questions that reflect high general knowledge of a 30-year-old adult reflecting higher than average verbal intelligence (verbal intelligence level 125 and above): Who painted “The Scream”? Who wrote the Odyssey? In which century did Beethoven live? In which sea are the Maldives Islands located? What is the biological component that converts RNA to protein? (2) Here are some questions that reflect the medium general knowledge of a 30-year-old adult reflecting average verbal intelligence (verbal intelligence level 90-110): Who was the first king of Israel according to the Bible? Who invented the telephone? In which country is the famous Himalayan Mountain, Everest, located? In which year did World War II end? Who was Nelson Mandela? (3) Here are some questions that reflect the low general knowledge of a 30-year-old adult reflecting much lower than average verbal intelligence (verbal intelligence level 70 and below): What is the vehicle that moves over water and carries people? On which Israeli holiday is it customary to light candles? What animal is the king of animals? What is the name of the first month of the Gregorian calendar? What shape do most balls have?
- Now build the test to assess verbal intelligence. The examples provided to you are meant to reflect different levels of difficulty to help you build a test sensitive to a score range of 50-150. You don't need to build the questions in the same structure as the examples (common denominator, vocabulary, and knowledge).

The AI-BVCT Scoring

After the participants completed the AI-BVCT, their responses to the test questions were entered into 2 LLMs, Claude and GPT-4, for evaluation and scoring. The LLMs were instructed to review and grade the participants’ responses to each of the 30 questions, assigning a score ranging from 0 to 2 for each answer (0=incorrect, 1=partial answer, and 2=full answer), making the total possible score range 0-60. Given the assumption that the test has a normal distribution of difficulty levels, correctly answering half the test questions (15 correct answers score of 30) is considered equivalent to a verbal IQ score of 100, correctly answering all test questions (30 correct answers, a raw score of 60) indicates a verbal IQ score of 150. Consequently, each point above or below 30 in the AI-BVCT corresponds to an IQ change of 1.67 points. For example, the IQ score of participants whose AI-BVCT was determined to be 31 would be 101.67. To increase reliability and avoid random errors, each participant’s responses were entered into each of the LLM 10 times.

The scoring process sought to address 2 main challenges. The first is that the LLMs struggle to perform numeric summation consistently. The second concerns certain cases where the LLMs incorrectly classify a correct response as incorrect or an incorrect response as correct. To address these challenges, we took two steps: (1) an MA student in psychology manually scored each participant’s responses to the AI-BVCT and (2) the automatic scoring process was supervised by the researchers regarding two key criteria in Textbox 2.

XSL•FO
RenderX

Textbox 2. Correction criteria for artificial intelligence–based scoring errors.

- In cases where there was an error in summing the number of correct responses into a total score [For example, in a case where a participant correctly answered 10 questions (2 points each) and partially answered 10 questions (1 point each), the large language model (LLM) erroneously calculated a total score of 29 instead of the correct 30], the LLMs were instructed: "You have summed the cases incorrectly, please re-sum."
- In extreme cases where the LLMs assigned a score of 2 when the human scorer assigned a score of 0, or vice versa [For example, when asked to define "ambivalence," a participant responded: "Ambivalence is the simultaneous existence of conflicting attitudes or emotions towards an object or situation." The human rater scored this as fully correct (2 points), but the LLM incorrectly assigned 0 points.], we instructed the LLMs: "You have made an error on question number X, classifying it as incorrect when it is correct. Please adjust the score accordingly." While this correction process was necessary for this proof-of-concept study, we acknowledge that such iterative adjustments could potentially introduce bias. Future implementations should consider automated validation methods and more structured scoring protocols to minimize human intervention.

Prompt for Reviewing the Personalized AI-BVCT

The prompt used is outlined in Textbox 3.

Textbox 3. Prompt for reviewing and scoring artificial intelligence–based verbal comprehension test (AI-BVCT) responses.

(Insert the participant's responses to the test questions here)

Your task now is to create a table in the following order

- In column 1, write the question and its number.
- In column 2, write the participant's answer.
- In column 3, carefully review each question based on your knowledge. Cross-reference different information sources and score according to the following key: a correct answer receives a score of 2. A partial or not entirely accurate answer receives a score of 1. Also, give a score of 1 for a slight inaccuracy. An entirely incorrect answer or a statement from the participant that they do not know the answer receives a score of 0. Ignore spelling errors in the scoring.
- In column 4, indicate the cumulative score accumulated throughout the questions.
- Below the table, write how many correct and partially correct answers there were out of 30 and calculate the intermediate score by accurately summing the scores obtained in step c. Caution: Please ensure the sum is correct before submitting the answer.
- Perform an additional calculation, this time counting the number of errors separately and noting it.
- Verify that the number of correct, partially correct, and incorrect answers adds up correctly to a total of 30 questions.
- After verification, indicate the final intermediate score at the end-this should be the score calculated in section e.

Prompt for Scoring the Personalized AI-BVCT

The following prompts were used:

> *(Insert the intermediate score obtained from the AI-BVCT here)*
>
> *Now you need to convert the obtained score to a new score. First, create a heading: "Final Score Calculation"*
>
> *Calculation method: If the participant received an intermediate score of 30 by answering 15 questions correctly with 2 points each, they will receive a score of 100. For each point above the intermediate score of 30, add 1.67 points (a participant who received an intermediate score of 31 will receive a final score of 101.67). The same applies in reverse for scores below 15 correct answers; they will receive below 100 in a similar ratio.*

Ethical Considerations

This study was approved by the institutional review board of Max Stern Yezreel Valley College (approval number: YVC 2023-66). All participants provided written informed consent after receiving a comprehensive explanation of the research procedures and objectives. Data were collected anonymously and stored securely with access restricted to the research team. All identifying information was removed from responses before analysis. Participants received course credit as compensation for their participation. No personally identifiable information or images were included in the study materials or findings.

Statistical Analysis

To assess the agreement between the AI-BVCTs and the VCI scores, we computed the intraclass correlation coefficient (ICC) and Bland-Altman limits of agreement (LoA; a method to assess agreement between 2 different measurement techniques by plotting the differences against the averages of the 2 measures). We also calculated the concordance correlation coefficient (CCC) to evaluate the agreement between the AI-BVCTs and the VCI scores (CCC measures the degree to which pairs of observations fall on the 45° line through the origin, indicating perfect agreement). We used Pearson correlations to examine the association between the VCI score, AI-BVCT Claude 2 scores, and AI-BVCT GPT-4 scores. The reliability of the AI-BVCTs was assessed using internal consistency (Cronbach α).

To test the hypothesis that the differences between the AI-BVCTs and VCI scores would be bounded within equivalence bounds of 15 points, we conducted Two One-Sided Tests (TOST) and paired-samples *t* test (TOST, used to establish

XSL•FO
RenderX

statistical equivalence between 2 measures). We compared the proportion of AI-BVCT Claude 2 and GPT-4 scores falling within the 95% WAIS CI using a *z* test.

Finally, we used a generalized mixed model to investigate the differences between the AI-BVCT Claude 2 scores and GPT-4 scores and their interaction with the VCI scores. We conducted post hoc comparisons to examine the differences between the LLMs at various levels of VCI (mean, SD 1) using Jamovi (version 2.3.28).

Results

Preliminary Analysis

First, we examined the reliability of the 10 AI-BVCTs generated by each LLM. Both LLMs had high internal consistency (AI-BVCT Claude 2: α=.988, AI-BVCT GPT-4: α=.986). Next, we computed the ICC to assess the agreement between the AI-BVCTs and VCI scores. The ICC was 0.961 (95% CI 0.872-0.991; 2-way mixed, average measures, absolute agreement), which is considered excellent reliability [32].

We then proceed to calculate Bland-Altman LoA (Figure 2) and the CCC.

Figure 2. Bland-Altman plots demonstrating agreement between VCI scores and AI-generated verbal comprehension test scores in Hebrew-speaking university students (N=8): A comparison of traditional and AI-based assessment methods. Study conducted in Israel, September 2023. LoA: limits of agreement.

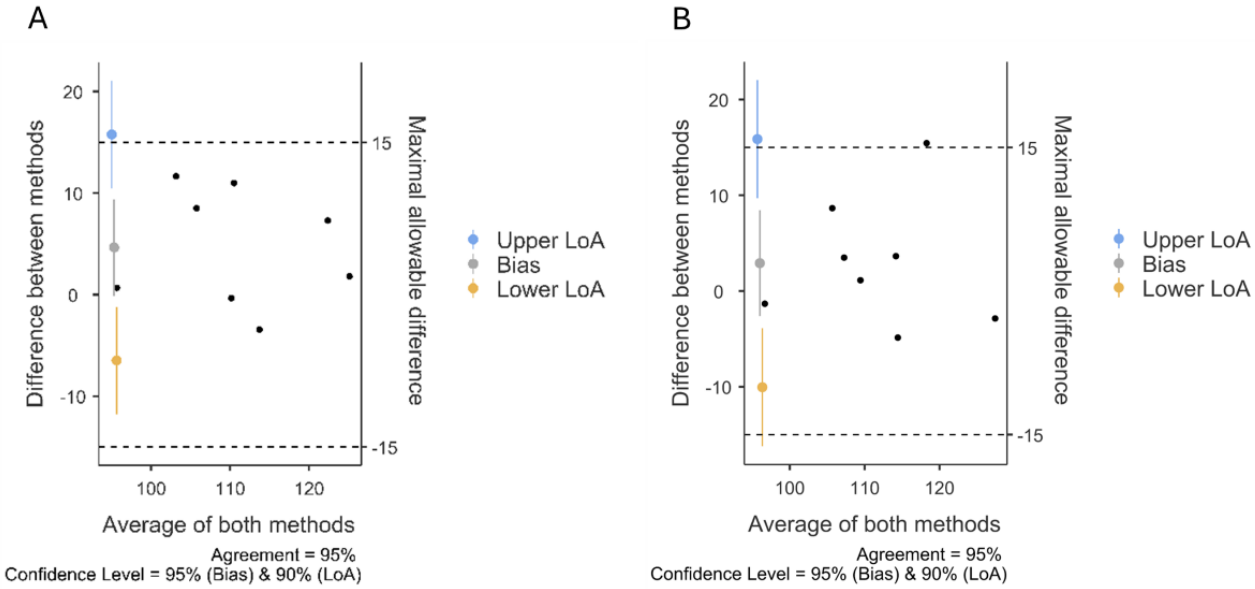

For AI-BVCT Claude 2, the Bland-Altman LoA indicated that the mean bias (4.64) was not significantly different from 0 (95% CI –0.10 to 9.38), the lower LoA was –6.48 (90% CI –11.78 to 1.18), and the upper LoA was 15.76 (90% CI 10.47-21.06). The CCC was 0.752 (90% CI 0.266-0.933), which is considered an almost excellent agreement.

For AI-BVCT GPT-4, the Bland-Altman LoA indicated that the mean bias (2.91) was not significantly different from 0 (95% CI –2.61 to 8.43), the lower LoA was –10.04 (90% CI –16.21 to –3.87), and the upper LoA was 15.86 (90% CI 9.70-22.03). The CCC was 0.733 (90% CI 0.170-0.935), which is considered an almost excellent agreement [33].

Following these results, we found high Pearson correlations between the VCI score and AI-BVCT Claude 2 scores (r=0.844, P<.001), the VCI score and AI-BVCT GPT-4 scores (r=0.771, P=.02), and AI-BVCT Claude 2 and AI-BVCT GPT-4 scores (r=0.866, P=.01).

In sum, the verbal IQ-like scoring of the AI-BVCTs generated by the LLMs was found to be reliable.

Hypothesis Testing

We hypothesized that the differences between the AI-BVCTs and VCI scores, if any, would be bounded within equivalence bounds of 15 points (half an SD of an IQ score), and examined it using a TOST paired-samples *t* test (Figure 3).

No difference was found between Claude 2 AI-BVCT scores (108.48, SD 10.43) and the VCI score (113.12, SD 9.78; t_7=–2.31, P=.05). The mean difference between the VCI and AI-BVCT scores was –4.64, with a 90% CI of –10.65 to 1.37. The 2 one-sided tests were statistically significant, with the score difference significantly higher than the lower bound of –15 (t_7=5.16, P<.001), and the score difference significantly lower than the upper bound of 15 (t_7=–9.78, P<.001).

No difference was found between the GPT-4 AI-BVCT scores (110.21, SD 9.75) and the VCI score (113.12, SD 9.78; t_7=–1.24, P=.25). The mean difference between the VCI and AI-BVCT scores was –2.91, with 90% CI of –9.91 to 4.09. The 2 one-sided tests were statistically significant, with the score difference significantly higher than the lower bound of –15 (t_7= 5.17,

XSL•FO
RenderX

P<.001), and the score difference significantly lower than the upper bound of 15 (t_7=–7.66, P<.001).

In conclusion, the IQ scores derived from the AI-BVCTs can be seen as equivalent on average to the VCI IQ scores.

Next, we examined whether the scores of the AI-BVCTs were within the range of the 95% WAIS CIs (Table 1). 5 out of 8 AI-BVCT Claude 2 scores (62.5%) and 75% of AI-BVCT GPT-4 scores were within the confidence intervals. The 2 proportions were significantly different (z=–0.53, P=.59).

Figure 3. Comparison of VCI scores with AI-BVCT scores from two language models (Claude and GPT-4) in Hebrew-speaking university students (N=8), Israel, September 2023, demonstrating score equivalence within predetermined bounds (SD 15 points). AI-BVCT: artificial intelligence-based verbal comprehension tests.

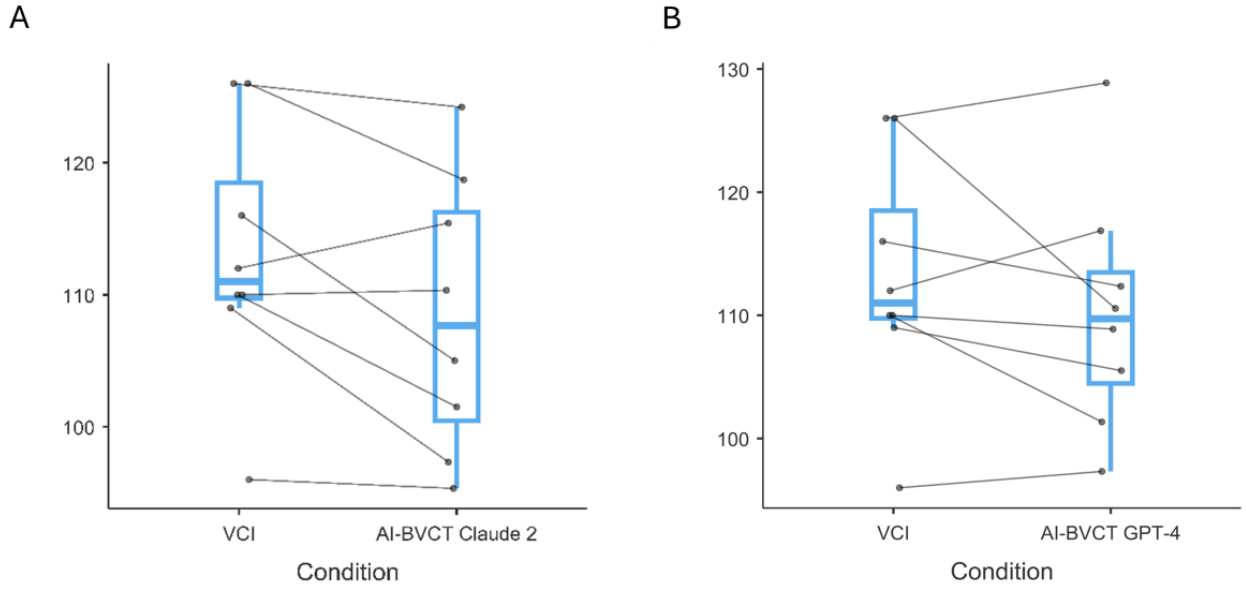

Table 1. Comparison of individual VCIs[a] with artificial intelligence–generated test scores: analysis of score agreement within WAIS[b] CIs in Hebrew-Speaking University Students (N=8) from Northern Israel, September 2023.

	VCI score	WAIS (95% CI)	AI-BVCT[c] GPT-4 scores		AI-BVCT Claude 2 scores	
			Scores	X[d] or V[e]	Scores	X or V
1	126	119-131	110.55	X	118.70	X
2	109	103-114	105.51	V	97.32	X
3	112	106-117	116.86	V	115.43	V
4	100	94-106	101.34	V	101.50	V
5	116	110-121	112.35	X	105.00	X
6	110	104-115	108.87	V	110.34	V
7	126	119-131	128.86	V	124.21	V
8	96	91-102	97.32	V	95.33	V
Total	__[f]	—	—	6/8 (75%)	—	5/8 (62.5%)

[a]VCI: verbal comprehension index.

[b]WAIS: Wechsler Adult Intelligence Scale.

[c]AI-BVCT: artificial intelligence–based verbal comprehension tests.

[d]X: Score outside the WAIS 95% CI.

[e]V: Score within the WAIS 95% CI.

[f]Not applicable.

Differences in AI-BVCT Scores Between the LLMs

Based on these findings, we examined the differences between the 2 LLMs and their interaction with the VCI score using a generalized mixed model (Figure 4). The 2 LLM scores were significantly different (χ^2_1=7.2, P<.001), with AI-BVCT GPT-4 having a higher mean score (109.86, SD 1.73) than AI-BVCT Claude 2 (108.08, SD 1.71). The difference between the LLMs interacted with the VCI score (χ^2_1=3.9, P=.04), so at 1 SD below

XSL•FO
RenderX

the mean (103.90), GPT-4 (103.18, SD 2.28) estimated the AI-BVCT score significantly higher than Claude 2 (100.31, SD 2.21; z=–3.08, P=.01). At the mean (113.08), GPT-4 (109.86, SD 1.74) estimated the AI-BVCT score significantly higher than Claude 2 (108.07, SD 1.72; z=–2.68, P=.01) but at 1 SD above the mean (122.27), GPT-4 (116.97, SD 2.58) was not significantly different than Claude 2 (116.44, SD 2.57; z=–0.53, P=.59).

These results suggest that Claude 2 underestimated and GPT-4 more closely estimated the true VCI scores, except for the higher VCI scores, where no difference was found between the LLMs. Both LLMs underestimated the true VCI scores at the mean and 1 SD above the mean levels. These results should be interpreted cautiously because the sample size was small and individuals with lower IQ than the mean were underrepresented.

Figure 4. Differential performance of language models (Claude vs GPT-4) in VCI estimation across intelligence levels in Hebrew-speaking university students (N=8), Israel, September 2023. GPT-4 demonstrated superior estimation accuracy at both -1 SD below mean and mean levels. AI-BVCT: artificial intelligence-based verbal comprehension tests. **p<0.001.

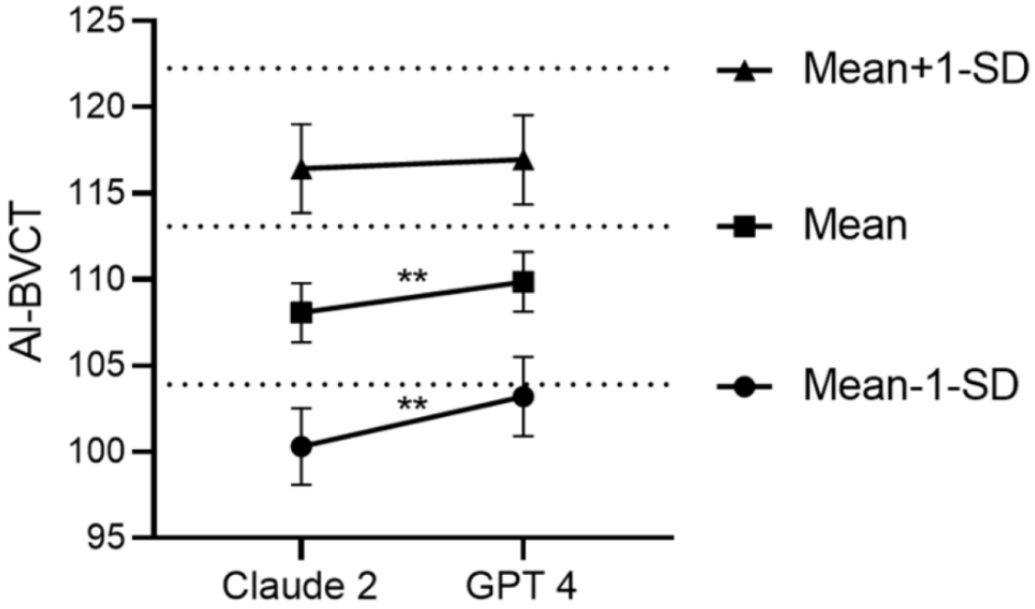

Discussion

Principal Findings

The present proof of concept study successfully demonstrated the potential feasibility of using LLMs to generate AI-based verbal comprehension tests. While the sample size was small, as appropriate for an initial feasibility investigation, the strong agreement between AI-BVCT and VCI scores provides compelling preliminary evidence that LLMs can generate psychologically meaningful assessments. However, it is important to note that these findings are based on a student population with above-average intelligence, and the application of this approach to broader populations requires careful consideration. The cognitive profile of university students differs systematically from the general population, particularly from clinical populations or individuals with below-average intelligence, and these differences might affect how LLMs interpret and score responses.

The findings showed a high ICC between the AI-BVCT and the VCI scores, indicating that the AI-generated test effectively discriminated between individuals with different verbal intelligence levels, ranking them similarly to the VCI. A paired-sample t test found no significant differences between the AI-BVCT scores and the VCI scores, demonstrating that the AI-generated test was properly calibrated to yield scores comparable to the traditional VCI. Finally, AI-BVCT scores fell within the 95% CI of the VCI scores, providing additional evidence that the AI-generated scores were consistent with what would be expected based on the well-established VCI measure. These preliminary results suggest that LLMs show promise in differentiating between individuals with varying verbal intelligence levels and produce scores that appear to correlate with the traditional VCI test. Further research with larger and more diverse samples is needed to fully establish the reliability and validity of this approach.

The findings of the present proof of concept study provide preliminary evidence that LLMs have the potential to contribute to the field of psychodiagnostics. The ability to accurately assess verbal skills is associated with the diagnosis of significant clinical states [6-10,27]. The results suggest that we may be on the cusp of a transformative moment in psychological assessments, a field that has not been substantially affected by technological advancements. The demonstrated capacity of LLMs to “understand” and process language at a high level holds immense promise for the diagnostic process.

Nevertheless, the process of coding and scoring responses still requires significant human involvement. We found that at times LLMs have difficulty with determining whether a particular answer is correct or incorrect and with accurately adding up scores. Specifically, we identified 2 main types of errors in using LLMs for scoring verbal intelligence tests. The primary issue was with summing up the scores accurately, which occurred frequently. In addition, there were a few instances where the LLMs did not provide correct scoring, although these

XSL•FO
RenderX

cases were not numerous. These challenges highlight the current limitations of LLMs in performing precise numerical calculations and occasional misinterpretations in scoring criteria. This situation requires a "human-in-the-loop" for supervision and correction. The need for human oversight is crucial in verifying the accuracy of score summations and ensuring the correct application of scoring criteria. In the future, with improvements in prompting techniques or the development of more powerful models, these issues may diminish. Potential improvements could include enhancing the LLMs' numerical computation abilities, refining the prompts to emphasize careful score tallying, and developing more robust error-checking mechanisms within the AI system. These advancements could significantly reduce the frequency of calculation errors and improve the overall accuracy of automated scoring. At present, human involvement remains essential to ensure the accuracy of the process.

In addition, we observed differences in scoring patterns between Claude 2 and GPT-4, with GPT-4 providing estimates closer to true VCI scores, especially for lower VCI scores. However, both LLMs underestimated scores at the mean and higher levels. While these differences may stem from varying capabilities in handling linguistic nuances and contextual understanding, particularly in Hebrew, the exact reasons remain unclear and may be temporary due to ongoing model updates. This finding highlights the importance of comparative analyses between LLM platforms in AI-based assessments. Future research should investigate these differences with larger, more diverse samples to ensure reliability across platforms. This underscores the need for ongoing monitoring and evaluation in the development of AI-assisted psychological assessment tools.

The preliminary findings of this study have several potential implications for the field of psychodiagnostics. One of the most significant is the potential of AI-based cognitive tests to make assessment processes more widely accessible and affordable. Traditional cognitive assessments are often expensive and require administration by trained psychologists, limiting their availability to a relatively small subset of the population, primarily those with suspected clinical difficulties. In the future AI-based cognitive tests may be easily generated and administered without the need for a psychologist's involvement, significantly reducing the associated costs. This could make cognitive assessments more readily available to a broader range of individuals, beyond those with clinical indications. The adaptability of AI-based tests could facilitate their translation and adaptation for use in different cultural and linguistic settings, further expanding their reach [15,22,27].

LLMs enable personalized testing tailored to each individual's unique characteristics and needs. This adaptive approach allows for the dynamic selection of item difficulty based on the examinee's performance, ensuring that the test is optimally challenging and informative. Such personalized testing can be particularly beneficial in cases that require repeated assessments over time, such as monitoring cognitive decline [34]. By generating novel items for each administration, AI-based tests can minimize practice effects and provide a more accurate measure of an individual's cognitive abilities.

The flexibility of AI-generated tests opens up new possibilities for expanding our understanding of cognitive functioning across diverse populations. Large-scale modeling of population distributions for various cognitive attributes may eventually enable the creation of customized norms that consider a wide range of factors, beyond age and gender. For example, norms could be tailored to specific learning disabilities, socioeconomic backgrounds, cultural contexts, and more [35]. This level of customization would allow for a more nuanced and comprehensive assessment of an individual's strengths and weaknesses. As LLMs continue to advance, we will likely see an increasing integration of these technologies into psychological assessment practices, leading to a new era of personalized and data-driven cognitive evaluation.

Despite the many positive implications demonstrated in this preliminary pilot study, consideration must also be given to the potential risks involved. Most significantly, the present pilot demonstrates the capability of LLMs to perform the technical aspects of test administration, such as test generation, coding, and scoring. Other aspects, like evaluating the context created during the assessment, identifying additional factors through clinical observation such as body language, or discerning problem-solving patterns, were not demonstrated in the current study. It remains uncertain whether LLMs will be able to perform these functions in the near future. In other words, the current research showed the ability to arrive at a similar score as a professional assessment but did not seek to identify or predict factors that may resolve or exacerbate difficulties.

Ethical Implications and Safeguards

The potential integration of AI-based cognitive assessment tools into psychological practice, even at this preliminary stage, raises important ethical considerations. Drawing from established professional guidelines, including those of major psychological associations, several key ethical domains require attention as this technology develops: protection of individual rights, professional oversight, prevention of misuse, concerns regarding epistemic authority, and current technological limitations [36].

First, privacy and data protection must be paramount [37]. While AI-based assessments could increase accessibility, they also create risks of unauthorized data collection. Professional psychological testing committees emphasize the need for (1) secure data handling protocols, (2) clear consent procedures specifically addressing AI involvement, and (3) strict limitations on data access and retention. These foundational requirements would need further development and validation in larger-scale studies.

Second, the risk of commercial misuse must be addressed through appropriate safeguards [38]. Based on our preliminary findings, key protections should include (1) explicit prohibition of covert cognitive profiling, (2) requirements for professional oversight, and (3) clear limitations on acceptable use cases. While these initial recommendations provide a framework, comprehensive protocols would need to be developed through extensive consultation with professional bodies and validation in clinical settings.

XSL•FO
RenderX

Third, professional standards and oversight are essential. Initial recommendations include maintaining clear boundaries between commercial applications and clinical assessment, requiring professional interpretation of results, and establishing guidelines for identifying and mitigating potential harm [39]. Regular consultation with senior clinical and neuropsychological professionals would be crucial in developing these standards as both the technology and its applications evolve.

Fourth, there is significant concern about overreliance on AI systems, which may be regarded as "all-knowing" with their recommendations taken as absolute [22,27]. This relates to the concept of epistemic authority, which is particularly concerning when working with vulnerable populations such as those with mental health challenges [40]. Professional guidelines emphasize that AI systems should serve only as supportive tools, with ultimate judgment residing with qualified clinicians. This requires (1) clear documentation of AI limitations, (2) training for professionals on appropriate AI integration, and (3) maintenance of critical evaluation practices.

Fifth, our findings highlight important limitations in current AI capabilities. While this study demonstrated the potential for automating technical aspects like test administration and scoring, critical clinical functions, such as observing nonverbal behavior, evaluating problem-solving patterns, and understanding broader context — remain beyond AI capabilities. In addition, the use of public AI models raises specific privacy concerns, as these systems are often trained on user data. Any clinical implementation would require dedicated, secure systems with proper privacy safeguards. Professional guidelines emphasize that validation must go beyond simple score comparison to include a comprehensive assessment of clinical utility and potential risks [27].

The implementation of these safeguards requires ongoing collaboration between AI developers, mental health professionals, and regulatory bodies. Regular consultation with senior clinical, neuropsychological, and educational psychologists will be essential for updating and refining these protocols as both AI capabilities and potential risks continue to evolve. Further research and broader professional consultation will be crucial to establish comprehensive ethical guidelines that protect individual rights while allowing beneficial applications of this emerging technology.

Study Limitations and Further Studies

While this pilot study included only 8 participants, this sample size was appropriate for our primary aim of demonstrating proof of concept. The strong correlation between AI-generated and traditional test scores, even in this small sample, provides compelling initial evidence for the feasibility of this approach. However, as this study was designed to demonstrate basic viability rather than population-level validation, further research with larger and more diverse samples is required to fully establish the viability of AI-generated testing. Moreover, the small sample was relatively homogeneous in education level and did not represent the full variation in the population. The intelligence examined was high, therefore the study did not take into account lower levels of intelligence. This represents range restriction, and it is possible that at lower levels the results would be different. Further research with larger and more diverse samples is required to fully establish the viability of AI-generated testing. Moreover, the study included a normative population, and future research should also examine populations with diverse cognitive disabilities (eg, learning disability). Another limitation is that the study was conducted in Hebrew; future research needs to examine the capabilities of LLMs in other languages as well.

Overall, these findings provide preliminary support for the use of LLMs in generating valid verbal intelligence assessments without reliance on normative data. The personalized approach allowed the creation of an individualized test tailored to the participant's age and gender background. This study was the first to examine the capabilities of AI in cognitive assessment and intelligence testing as a proof of concept. Given the pilot nature of the research, many additional studies are needed to further validate these initial findings in both the general population and in diverse clinical groups. As a preliminary investigation in this domain, the study helps lay the foundation for psychological testing by AI and serves as a proof of concept**.** Yet, considerably more research is required to fully characterize the use of AI for psychological testing across populations. This pilot study alone is insufficient to draw definitive conclusions, underscoring the need for rigorous continued inquiry in this nascent area. The findings expand the existing knowledge regarding what we know about AI in several aspects, including the ability to create a test and diagnose medical and psychological aspects. It represents an important first step in harnessing LLMs for innovative advances in psychological assessment. The ability to automate test development and scoring shows great promise in making testing more efficient, accessible, and personalized. As LLMs continue to evolve in sophistication, their integration can meaningfully augment and potentially transform standard practices in the field.

Conclusion

This preliminary study explores the potential of LLMs in generating personalized verbal intelligence tests. While our initial findings are promising, they highlight both opportunities and challenges in using AI for psychological assessments. Further research with larger, more diverse samples is necessary to validate these results and address the ethical considerations of integrating AI into psychological testing practices.

Acknowledgments

During the preparation of this manuscript, the authors used Claude 3.5 Sonnet to assist with English language refinement and sentence clarity. All output was carefully reviewed and edited by the authors, who take full responsibility for the content of this manuscript. This use of AI was limited to language enhancement and did not contribute to the scientific analysis or conclusions.

XSL•FO
RenderX

Data Availability

The datasets generated and analyzed during the current study are available from the corresponding author upon reasonable request. Access is restricted due to privacy considerations and ethical requirements regarding participant confidentiality. The institutional ethics committee will review requests for access to anonymized data which may be granted for research purposes that align with the informed consent provided by participants.

Conflicts of Interest

None declared.

References

1. Groth-Marnat G. Handbook of Psychological Assessment. Hoboken, NJ: Wiley & Sons; 2009.
2. Reynolds CR, Altmann RA, Allen DN. The problem of bias in psychological assessment. In: Mastering Modern Psychological Testing. Heidelberger Platz: Springer; 2021:573-613.
3. Norfolk PA, Farmer RL, Floyd RG, Woods IL, Hawkins HK, Irby SM. Norm block sample sizes: a review of 17 individually administered intelligence tests. J Psychoeduc Assess 2015;33(6):544-554. [doi: 10.1177/0734282914562385]
4. Climie EA, Rostad K. Test review: wechsler adult intelligence scale. J Psychoeduc Assess 2011;29(6):581-586. [doi: 10.1177/0734282911408707]
5. Lange RT, Lippa SM. Verbal comprehension index. In: Encyclopedia of Clinical Neuropsychology. Cham, Switzerland: Springer International Publishing; 2017:1-3.
6. Boyle R, Knight SP, De Looze C, Carey D, Scarlett S, Stern Y, et al. Verbal intelligence is a more robust cross-sectional measure of cognitive reserve than level of education in healthy older adults. Alzheimers Res Ther 2021;13(1):128 [FREE Full text] [doi: 10.1186/s13195-021-00870-z] [Medline: 34253231]
7. Henry JD, Crawford JR, Phillips LH. Verbal fluency performance in dementia of the alzheimer's type: a meta-analysis. Neuropsychologia 2004;42(9):1212-1222. [doi: 10.1016/j.neuropsychologia.2004.02.001] [Medline: 15178173]
8. Pauli L, Daseking M, Petermann F, Stemmler M. [Associations between cognitive performance in a dementia screening test (SKT) and an intelligence test (WAIS IV) : which deficits in cognitive performance in old age indicate a possible pathological deterioration process?]. Z Gerontol Geriatr 2018;51(3):266-274. [doi: 10.1007/s00391-017-1263-x] [Medline: 28600610]
9. Cornoldi C, Giofrè D, Orsini A, Pezzuti L. Differences in the intellectual profile of children with intellectual vs. learning disability. Res Dev Disabil 2014;35(9):2224-2230. [doi: 10.1016/j.ridd.2014.05.013] [Medline: 24927516]
10. Maseda A, Lodeiro-Fernández L, Lorenzo-López L, Núñez-Naveira L, Balo A, Millán-Calenti JC. Verbal fluency, naming and verbal comprehension: three aspects of language as predictors of cognitive impairment. Aging Ment Health 2014;18(8):1037-1045. [doi: 10.1080/13607863.2014.908457] [Medline: 24797556]
11. Bertie BL, Huang HW. Cognition and beyond: intersections of personality traits and language. Psychol Learn Motiv 2024;80:105-148. [doi: 10.1016/bs.plm.2024.03.003]
12. Na SD, Burns TG. Wechsler intelligence scale for children-v: test review. Appl Neuropsychol Child 2016;5(2):156-160. [doi: 10.1080/21622965.2015.1015337] [Medline: 25923224]
13. Rudolph J, Tan S, Tan S. ChatGPT: bullshit spewer or the end of traditional assessments in higher education? JALT 2023;6(1):342-363. [doi: 10.37074/jalt.2023.6.1.9]
14. Hu K. ChatGPT Sets Record for Fastest-Growing User Base - Analyst Note. New York: Reuters; 2023.
15. Elyoseph Z, Gur T, Haber Y, Simon T, Angert T, Navon Y, et al. An ethical perspective on the democratization of mental health with generative AI. JMIR Ment Health 2024;11:e58011 [FREE Full text] [doi: 10.2196/58011] [Medline: 39417792]
16. Elyoseph Z, Hadar Shoval DH, Levkovich I. Beyond personhood: ethical paradigms in the generative artificial intelligence era. Am J Bioeth 2024;24(1):57-59. [doi: 10.1080/15265161.2023.2278546] [Medline: 38236857]
17. Elyoseph Z, Hadar-Shoval D, Asraf K, Lvovsky M. ChatGPT outperforms humans in emotional awareness evaluations. Front Psychol 2023;14:1199058 [FREE Full text] [doi: 10.3389/fpsyg.2023.1199058] [Medline: 37303897]
18. Elyoseph Z, Levkovich I. Beyond human expertise: the promise and limitations of ChatGPT in suicide risk assessment. Front Psychiatry 2023;14:1213141 [FREE Full text] [doi: 10.3389/fpsyt.2023.1213141] [Medline: 37593450]
19. Elyoseph Z, Levkovich I. Comparing the perspectives of generative AI, mental health experts, and the general public on schizophrenia recovery: case vignette study. JMIR Ment Health 2024;11:e53043 [FREE Full text] [doi: 10.2196/53043] [Medline: 38533615]
20. Elyoseph Z, Levkovich I, Shinan-Altman S. Assessing prognosis in depression: comparing perspectives of AI models, mental health professionals and the general public. Fam Med Community Health 2024;12(Suppl 1):e002583 [FREE Full text] [doi: 10.1136/fmch-2023-002583] [Medline: 38199604]
21. Elyoseph Z, Refoua E, Asraf K, Lvovsky M, Shimoni Y, Hadar-Shoval D. Capacity of generative AI to interpret human emotions from visual and textual data: pilot evaluation study. JMIR Ment Health 2024;11:e54369 [FREE Full text] [doi: 10.2196/54369] [Medline: 38319707]

XSL•FO
RenderX

22. Haber Y, Levkovich I, Hadar-Shoval D, Elyoseph Z. The artificial third: a broad view of the effects of introducing generative artificial intelligence on psychotherapy. JMIR Ment Health 2024;11:e54781 [FREE Full text] [doi: 10.2196/54781] [Medline: 38787297]
23. Hadar-Shoval D, Asraf K, Mizrachi Y, Haber Y, Elyoseph Z. Assessing the alignment of large language models with human values for mental health integration: cross-sectional study using schwartz's theory of basic values. JMIR Ment Health 2024;11:e55988 [FREE Full text] [doi: 10.2196/55988] [Medline: 38593424]
24. Hadar-Shoval D, Elyoseph Z, Lvovsky M. The plasticity of ChatGPT's mentalizing abilities: personalization for personality structures. Front Psychiatry 2023;14:1234397 [FREE Full text] [doi: 10.3389/fpsyt.2023.1234397] [Medline: 37720897]
25. Levkovich I, Elyoseph Z. Suicide risk assessments through the eyes of ChatGPT-3.5 versus ChatGPT-4: vignette study. JMIR Ment Health 2023;10:e51232 [FREE Full text] [doi: 10.2196/51232] [Medline: 37728984]
26. Levkovich I, Shinan-Altman S, Elyoseph Z. Can large language models be sensitive to culture suicide risk assessment? Journal of Cultural Cognitive Science 2024;8(3):257-287. [doi: 10.21203/rs.3.rs-4066705/v1]
27. Tal A, Elyoseph Z, Haber Y, Angert T, Gur T, Simon T, et al. The artificial third: utilizing ChatGPT in mental health. Am J Bioeth 2023;23(10):74-77. [doi: 10.1080/15265161.2023.2250297] [Medline: 37812102]
28. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
29. Denny P, Leinonen J, Prather J, Luxton-Reilly A, Amarouche T, Becker BA, et al. Promptly: using prompt problems to teach learners how to effectively utilize AI code generators. 2023 Presented at: SIGCSE 2024: Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1; 2024 March 20 - 23; Portland OR USA.
30. Ryan JJ, Lopez SJ. Wechsler adult intelligence scale-III. In: Understanding Psychological Assessment. Heidelberger Platz: Springer; 2001:19-42.
31. Kaufman AS, Lichtenberger EO. Assessing Adolescent and Adult Intelligence. Hoboken, NJ: Wiley; 2006.
32. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol Assess 1994;6(4):284-290. [doi: 10.1037//1040-3590.6.4.284]
33. Altman DG. Practical Statistics for Medical Research. London: Chapman and Hall; 1991.
34. Gibbons RD, Lauderdale DS, Wilson RS, Bennett DA, Arar T, Gallo DA. Adaptive measurement of cognitive function based on multidimensional item response theory. Alzheimers Dement (N Y) 2024;10(4):e70018. [doi: 10.1002/trc2.70018] [Medline: 39748843]
35. Waheed W, Mirza N, Waheed MW, Malik A, Panagioti M. Developing and implementing guidelines on culturally adapting the addenbrooke's cognitive examination version III (ACE-III): a qualitative illustration. BMC Psychiatry 2020;20(1):492 [FREE Full text] [doi: 10.1186/s12888-020-02893-6] [Medline: 33023520]
36. Wright AJ, Chávez L, Edelstein BA, Grus CL, Krishnamurthy R, Lieb R, et al. Education and training guidelines for psychological assessment in health service psychology. Am Psychol 2021;76(5):794-801. [doi: 10.1037/amp0000742] [Medline: 33180536]
37. Boone KB, Sweet JJ, Byrd DA, Denney RL, Hanks RA, Kaufmann PM, et al. Official position of the American academy of clinical neuropsychology on test security. Clin Neuropsychol 2022;36(3):523-545. [doi: 10.1080/13854046.2021.2022214] [Medline: 35043752]
38. Braunack-Mayer A, Carolan L, Street J, Ha T, Fabrianesi B, Carter S. Ethical issues in big data: a qualitative study comparing responses in the health and higher education sectors. PLoS One 2023;18(4):e0282285 [FREE Full text] [doi: 10.1371/journal.pone.0282285] [Medline: 37097990]
39. Ogugua JO, Okongwu CC, Akomolafe OO, Anyanwu EC, Daraojimba OD. Mental health and digital technology: a public health review of current trends and responses. Med Sci Res 2024;4(2):108-125. [doi: 10.51594/imsrj.v4i2.754]
40. Hadar SD, Haber Y, Tal A, Simon T, Elyoseph T, Elyoseph Z. Transforming perceptionsxploring the multifaceted potential of generative AI for people with cognitive disabilities. JMIR Preprints 2024:64182. [doi: 10.2196/64182]

Abbreviations

AI: artificial intelligence
AI-BVCT: artificial intelligence-based verbal comprehension test
CCC: concordance correlation coefficient
GenAI: generative artificial intelligence
ICC: intraclass correlation coefficient
LLM: large language model
LoA: limits of agreement
TOST: Two One-Sided Tests
VCI: verbal comprehension index
WAIS-III: Wechsler Adult Intelligence Scale

XSL•FO
RenderX

Edited by A Mavragani; submitted 11.11.24; peer-reviewed by DJ Samuel, M Thomas; comments to author 18.12.24; revised version received 04.01.25; accepted 05.01.25; published 24.02.25.

Please cite as:
Hadar-Shoval D, Lvovsky M, Asraf K, Shimoni Y, Elyoseph Z
The Feasibility of Large Language Models in Verbal Comprehension Assessment: Mixed Methods Feasibility Study
JMIR Form Res 2025;9:e68347
URL: https://formative.jmir.org/2025/1/e68347
doi:10.2196/68347
PMID:

©Dorit Hadar-Shoval, Maya Lvovsky, Kfir Asraf, Yoav Shimoni, Zohar Elyoseph. Originally published in JMIR Formative Research (https://formative.jmir.org), 24.02.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

An Integrated Platform Combining Immersive Virtual Reality and Physiological Sensors for Systematic and Individualized Assessment of Stress Response (bWell): Design and Implementation Study

Budhachandra Khundrakpam[1], PhD; Melanie Segado[1], PhD; Jesse Pazdera[1], PhD; Vincent Gagnon Shaigetz[1], MSc; Joshua A Granek[2], PhD; Nusrat Choudhury[1], MEng

[1]National Research Council Canada, Boucherville, QC, Canada
[2]Defence Research and Development Canada, Toronto, ON, Canada

Corresponding Author:
Nusrat Choudhury, MEng
National Research Council Canada
75 de Mortagne
Boucherville, QC, J4B 6Y4
Canada
Phone: 1 450 978 6198
Email: nusrat.choudhury@nrc-cnrc.gc.ca

Abstract

Background: Stress is a pervasive issue in modern society, manifesting in various forms such as emotional, physical, and work-related stress, each with distinct impacts on individuals and society. Traditional stress studies often rely on psychological, performance, or social tests; however, recently, immersive virtual reality (VR), which provides a sense of presence and natural interaction, offers the opportunity to simulate real-world tasks and stressors in controlled environments. Despite its potential, the use of VR to investigate the multifaceted manifestations of stress has not been thoroughly explored.

Objective: This study aimed to explore the feasibility of using a VR-based platform, *bWell*, to elicit multifaceted stress responses and measure the resulting behavioral and physiological changes. Specifically, we aimed to design various VR stress exercises based on neurocardiac models to systematically test cardiac functioning within specific contexts of self-regulation (executive functioning, physical efforts, and emotional regulation).

Methods: The development process adhered to guidelines for VR clinical trials and complex health interventions, encompassing 3 phases: preparation, development, and verification. The preparation phase involved a comprehensive literature review to establish links between stress, the heart, and the brain, leading to the formulation of a conceptual model based on the Neurovisceral Integration Model (NVIM) and Vagal Tank Theory (VTT). The development phase involved designing VR exercises targeting specific stressors and integrating physiological sensors such as photoplethysmography (PPG) and electromyography (EMG) to capture heart rate variability (HRV) and facial expressions. The verification phase, conducted with a small number of trials, aimed to design a study and implement a workflow for testing the feasibility, acceptability, and tolerability of the VR exercises. In addition, the potential for capturing physiological measures along with subjective ratings of stress for specific dimensions was assessed.

Results: Verification trials demonstrated that the VR exercises were well tolerated, with negligible cybersickness and high user engagement. The different VR exercises successfully elicited the intended stress demands, along with the physiological responses.

Conclusions: The study presents a novel VR-based experimental setup that allows a systematic and individualized assessment of stress responses, paving the way for future research to identify features that confer stress resilience and help individuals manage stress effectively. While our conceptual model highlights the role of HRV in providing valuable insights into stress responses, future research will involve multivariate and machine learning analyses to predict individual stress responses based on comprehensive sensor data, including EMG and the VR-based behavioral data, ultimately guiding personalized stress management interventions.

(JMIR Form Res 2025;9:e64492) doi:10.2196/64492

XSL•FO
RenderX

KEYWORDS

virtual reality; stress; physiological response; NASA-Task Load Index; cognitive demand; physical demand; vagal tone; heart rate variability

Introduction

Background

A rising problem in our society is *stress*, with serious burdens imposed on individuals and society by different types of stressors, such as emotional stress, physical stress, and work stress [1,2]. The multifaceted nature of stress makes it even more problematic, with different stressors yielding different responses and consequences. For example, increased workload and time pressure has led to increased mental stress [3]. Similarly, in recent years, there has been an increase in emotional, behavioral, and psychological stress due to the COVID-19 pandemic [4,5].

Cumulative evidence from studies of stress neurobiology indicates individual variability in how people respond to stress. While for some individuals, stress can lead to negative outcomes (eg, deterioration of physical and mental health), other individuals may not be affected to that extent [6,7]. Identifying the risk factors that make a person vulnerable to negative consequences of stress would allow timely detection and intervention. By contrast, identifying and delineating the individual features that potentially confer stress resilience (defined as the ability of most people to maintain normal functioning even under extreme stress) is important in helping individuals reach their full potential. Factors that confer resilience include genetic predisposition, age, sex, prenatal environment, sleep, and exercise [8]. In terms of biological mechanisms, stress resilience has been attributed to a coordinated activity of the autonomic nervous system (ANS); the hypothalamic-pituitary-adrenal axis; the immune system; and the brain regions that control their activity, including the hypothalamus, amygdala, hippocampus, medial prefrontal cortex, and ventral tegmental area [9-11]. In particular, stress-induced activation of ANS has been linked to changes in heart rate and vasoconstriction [9]. In addition, stress responses via activation of sympathetic-adrenomedullary and hypothalamic-pituitary-adrenal axes have been observed in response to psychosocial and physical demands during intense exercise [12].

Given the complex coordinated involvement of several bodily systems (both central and peripheral), studies on stress resilience require measurement of attributes that can reflect these interactions. One such attribute is heart rate variability (HRV), which captures heart-brain interactions and ANS dynamics [13]. HRV is defined as the variation in the time intervals between consecutive heartbeats and is an emergent property of complex interactions among several interdependent physiological systems. Several studies have shown the usefulness of HRV in the analysis of stress that the body experiences during physical training and to increase insight into physiological recovery after training [14,15]. Measurement of HRV (using electrocardiogram [ECG] or photoplethysmography [PPG]) is an easy noninvasive technique, and therefore, it has been used to investigate stress-related changes in cardiovascular health in different conditions, including exercise [16] and sleep [17]. More importantly, studies on individuals with stress-related disorders (such as anxiety and depression) revealed altered HRV metrics, indicating impairment of neurocardiac function mediated by changes to the ANS [18]. In addition, integrating HRV with other modalities such as electroencephalography have shown the impact of stress on cardiac and brain health [19,20].

Traditionally, the impact of stress on biological measures such as HRV and brain activity is investigated after inducing a specific stressor (mainly in live scenarios) using psychological, performance, or social tests [21,22]. However, in recent years, the advancement of virtual reality (VR) and head-mounted displays has allowed the investigation of different stressors that are otherwise difficult to reproduce in real life. Using head-mounted displays, immersive VR offers a sense of presence and a more natural interaction, allowing multisensorial (visual, audio, and touch) simulation of real-world tasks with tight experimental control, enabling the study of specific stressors. In a study, VR was used to elicit stress in high-altitude simulations, and stress-related responses in the heart and brain were observed using ECG and electroencephalography, respectively [23]. In a VR study using ECG signals, a multilevel stress assessment (3 levels of stress) was performed using spectrograms and 1D ECG signals that performed better than traditional machine learning (ML) models [24]. In another example, playing 3D games (a combination of the first-person shooter and horror genres) in immersive VR was shown as an effective stressor for practicing stress management skills [25]. In alignment with these findings, a meta-analysis of 52 studies revealed that stress tasks in VR elicit a variety of physiological stress reactivity, highlighting the scope of VR in the study of stress [26].

Objectives

Despite the advantages offered by VR (multisensorial—visual, audio, and touch simulation of real-world tasks), previous studies have mostly used VR to induce particular stressor at a time, for example, cognitive stress [27] and emotional stress [28]. In this work, we explored the feasibility of implementing a VR-based experimental setup to (1) elicit multifaceted stress response and (2) measure the resulting behavioral and physiological response. Using our validated VR platform *bWell* [29], we aimed to leverage specific exercises among its battery for eliciting different stressor responses, specifically for physical, cognitive, and emotional stress. Particular emphasis was given on designing VR stress exercises to systematically test cardiac functioning within specific contexts of self-regulation (executive functioning and emotional regulation) based on neurocardiac models [30,31]. In addition, by integrating VR with PPG and electromyography (EMG), we captured various physiological measures, including HRV and facial expressions (collected via EMG), along with subjective ratings of stress (namely, the Perceived Stress Scale scores for

XSL•FO
RenderX

overall stress and NASA-Task Load Index [NASA-TLX] for specific stress dimensions) and mood.

The overarching goal is that the study design or paradigm tailor to different stress domains and measurable behavioral and physiological outputs. The long-term goal is to perform multivariate and ML-based analysis for predicting the subjective perceived stress of an individual. In addition, it should be adaptable to the specific needs of each individual, allowing a systematic and individualized assessment of stress response.

Methods

Development Process

Because the work involved immersive VR in a health context, the "Recommendations for Methodology of Virtual Reality Clinical Trials in Health Care by an International Working Group" were followed during the development process [32]. Following the recommendations, we focused on the first two phases: (1) VR1 phase with an emphasis on content development of different VR stress exercises and (2) VR2 phase with an emphasis on verification trials testing feasibility, acceptability, and tolerability. In addition, the study development process also followed guidelines from the Medical Research Council, namely "Guidance on how to develop complex interventions to improve health and healthcare," [33] which recommends combining the opinions or input elicited during content development with existing or available theories (eg, the rationale for intervention, expected changes, and how changes may be achieved). On the basis of the guidelines, our development study included a precursor step to identify the prevailing theories and frameworks in the domain to formulate hypotheses for the study.

Consequently, our development process comprised three phases: (1) a *preparation phase*, during which an extensive literature review was performed to establish the links between heart, brain, and stress, leading to the formulation of a conceptual model and formation of hypotheses (precursor phase); (2) a *development phase*, which involved iterative design of VR exercises for targeted stressors and hardware integration (VR1 phase); and (3) a *verification phase* (VR2 phase), focusing on implementing a workflow able to test the outcome measures associated with VR2 trials—tolerability, acceptability, and feasibility (Figure 1).

Figure 1. Schema of the development process. The development process comprised three phases: preparation phase, virtual reality (VR)-1 development phase and VR-2 verification phase.

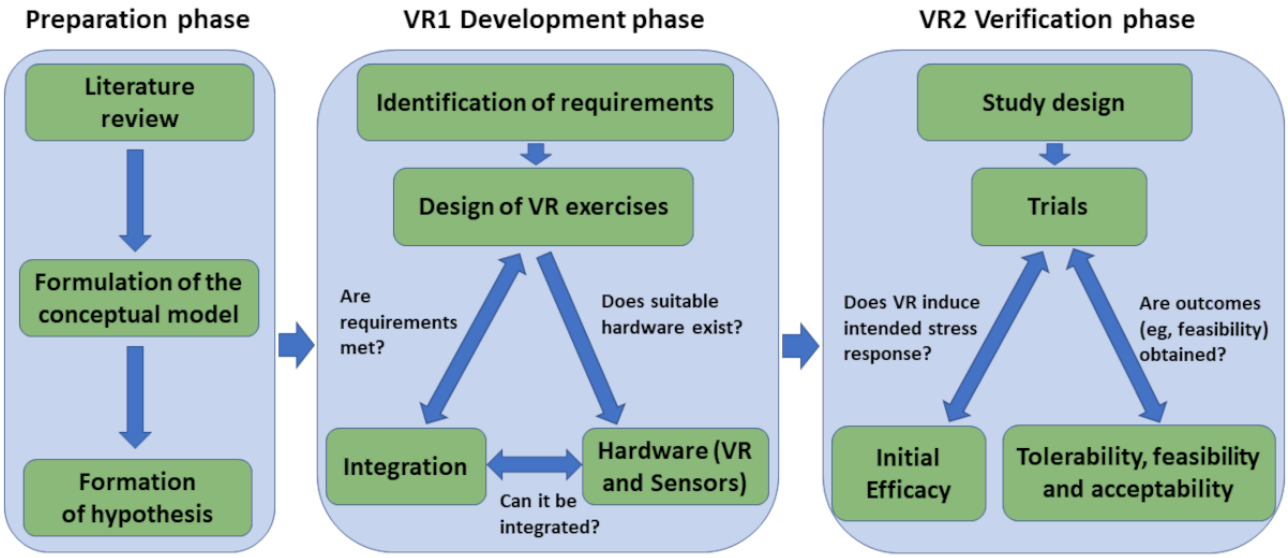

Preparation Phase

Literature Review and Survey

A literature survey was performed to establish links between stress, the heart, the brain, and cognition. Given the involvement of several bodily systems in stress, we observed diverse findings: on the one hand, associations with mental load [34] and emotional stress [35], on the other hand, associations with cognitive performance [36,37]. With the goal of linking these diverse findings, we next focused on publications that have reported integrated models of the heart, the brain, and cognition (especially self-regulation, a cognitive process implicated in stress). The first such model is the *Neurovisceral Integration Model(NVIM)* [38,39]. According to the NVIM, the heart and the brain are connected by direct and indirect pathways that influence the various processes (physiological, behavioral, emotional, and cognitive) related to self-regulation and adaptability. More importantly, the NVIM postulates vagal tone as the common index that is linked to physiological, affective, and cognitive regulation. The second model is based on the *Vagal Tank Theory (VTT)* [31]. The VTT suggests that a higher vagal tone is associated with better self-regulation and adaptive functioning.

It is worth noting that the theories of interest focus on vagal tone, which refers to the activity of the vagus nerve. The vagus nerve is the largest autonomic nerve, innervating nearly every organ in the body. More importantly, the vagus nerve regulates metabolic homeostasis by controlling the heart rate [40] as well as the interval between 2 heartbeats (HRV) [41]. While direct measurement of vagal activity (ie, physiological studies of the vagus nerve) has been limited due to difficulties in recording chronic signals in the vagus [42,43], HRV provides an indirect

XSL•FO
RenderX

measure of vagal tone (a clinical metric to specify overall levels of vagal activity) because it is largely regulated by cardiac vagal control [44,45]. HRV has been shown to be a sensitive metric for detecting abnormalities in vagal tone, as observed in clinical conditions such as diabetes, heart failure, and hypertension [46-48]. More broadly, HRV, as an indirect measure of vagal tone, has shown to be useful in many aspects of psychophysiology, such as self-regulation mechanisms linked to cognitive, affective, social, and health phenomena [13,30,49]. In summary, HRV is commonly used in research as a practical index for vagal tone, and therefore, we will be using HRV in our study to link vagal tone with stress.

Formulation of the Conceptual Model

To formulate our model, we first examined the hypotheses of the NVIM and VTT. Specifically, the NVIM suggests that greater parasympathetic nervous system activity (ie, vagal tone) increases regulatory control in the brain and body [38,39]. This control supports certain biological processes such as digestion, while facilitating brain functioning related to cognitive control and emotional regulation. In contrast, withdrawal of vagal tone supports physical action, such as a fight-or-flight response. The parasympathetic nervous system is thought to monitor and balance feedback from numerous subsystems to determine the optimal level of vagal tone to deploy, given one's current context, active goal, and unmet needs [50]. HRV reflects vagal tone and, as such, should correlate with the integrity of cognitive and emotional control processes [30]. By contrast, VTT suggests that there are *3* states of vagal control, the "3 Rs": *resting*, *reactivity*, and *recovery* [31]. Optimal vagal tone is high at rest, so a healthy system should settle to a high baseline level of parasympathetic activity. In response to mental demands, vagal control systems should react by maintaining a high vagal tone to promote executive control. In response to more active demands, these systems should react by withdrawing vagal tone to promote physical performance. In practice, vagal tone often drops below resting levels even during cognitive tasks, possibly due to the strain on and gradual exhaustion of vagal control systems. Once demands are met, vagal control should transition into a recovery state, where vagal tone returns to resting levels.

The NASA-TLX provides a formal framework for categorizing different aspects of workload and task demands [51]. Therefore, we identified this index as a useful tool for understanding the demands that vagal control systems must continuously monitor and react to, as per the *VTT* and the *NVIM*. Specifically, the NASA-TLX defines 3 types of task demands—*mental*, *physical*, and *temporal*—and 3 additional factors—*performance*, *effort*, and *frustration*—that contribute to perceived workload. On the basis of the *NVIM* model of vagal control, *mental*, *physical*, and *temporal* task demands should be associated with different optimal levels of vagal tone and, therefore, different states of reactivity. *Mental* demands are best solved by promoting executive control through high levels of parasympathetic nervous system activity; in contrast, active *physical* and *temporal* demands are best addressed by a withdrawal of vagal tone.

Formation of Hypothesis

On the basis of previous publications, we formed specific hypotheses for self-regulation for each level of cardiac vagal control (Table 1). At *resting (baseline)*, we hypothesized that the higher the vagal tone, the better the self-regulation [49,52-54]. For *reactivity*, different hypotheses were formed for emotional, physical, and cognitive demands. It may be noted that for consistency with terminology in neuroscience, we have used cognitive demand as representing mental demand and emotional demand as representing frustration in the NASA-TLX. For emotional demand, (1) a complete vagal withdrawal would elicit a fight-or-flight response [52,55], (2) a small increase in vagal tone would indicate successful emotion regulation [53,54], and (3) a small decrease would indicate that the task did not elicit a sufficiently large stress response [49,56]. For cognitive demand, (1) we hypothesized an initial increase in vagal tone under low cognitive load (self-regulatory effort) and a decrease in vagal tone under high cognitive load [31,57], and (2) in the context of vagal withdrawal (ie, high cognitive load), we hypothesized that a small decrease in vagal tone would predict better performance [31,58]. For physical demand, a small vagal withdrawal would indicate quick adaptation to meet the demands of the situation [59,60]. Finally, for *recovery*, we hypothesized that the faster the vagal tone returns to the baseline level, the better the self-regulation [61].

XSL•FO
RenderX

Table 1. Description of the hypotheses corresponding to our conceptual model based on the Neurovisceral Integration Model (NVIM) and Vagal Tank Theory (VTT). A distinct hypothesis was made for each phase—resting (baseline), reactivity, and recovery.

Cardiac vagal control level	Hypothesis of self-regulation, as indicated by cardiac vagal control
1. Resting (baseline)	• A higher vagal tone indicates better self-regulation
2. Reactivity	
2a. Emotional demand	• Complete vagal withdrawal indicates that a fight-or-flight response was elicited • A small increase indicates successful emotional regulation • A small decrease indicates that the task did not elicit a sufficiently large stress response
2b. Cognitive demand	• An initial increase in vagal tone under low cognitive load (self-regulatory effort) and a decrease under high cognitive load • A smaller decrease (under high cognitive load) is associated with better performance
2c. Physical demand	• Small vagal withdrawal indicates quick adaptation to meet demands of the situation
2d. Physical and cognitive demand	• Combination of hypotheses 2c (vagal withdrawal) and 2b (initial increase under low cognitive load with a more pronounced effect)
3. Recovery (from demand)	• The faster the vagal tone returns to the baseline level, the better the coping mechanisms

VR1: Development Phase

Identification of Requirements

The fundamental objective of this work is to harness the affordances of immersive VR to provide new opportunities for practicing stress management skills, particularly the synthetic recreation of scenes that are otherwise too difficult or resource intensive to replicate in real life. The latter is of particular interest in the application toward training stress resilience, a complex and dynamic process. Combined with physiological monitoring, which can help to capture additional information on the user state, the use of VR-based technologies can enable research in support of better understanding the impacts of well-being and individual characteristics, such as mental and psychological resilience.

To study and validate the different parts of our conceptual model, we identified four main requirements:

1. *Requirement 1:* recreation of each level of cardiac vagal control (3 “Rs”: resting, reactivity, and recovery)
2. *Requirement 2:* ability to elicit different types of mental demands
3. *Requirement 3:* ability to elicit physical demand
4. *Requirement 4:* integration of physiological sensing with immersive VR, essential for the integration of HRV measures as an index of vagal tone

Design of VR Exercises

We have previously developed *bWell*, an interactive and immersive VR platform designed as a broadly applicable toolkit delivering multisensory tasks, targeting general aspects of cognition and everyday functioning [29,62-64]. For this study, we made use of *3* bWell exercises, namely *Tent*, *Mole*, and *Stroll*, which were adapted to systematically test adaptive aspects of HRV within specific contexts of self-regulation (executive functioning, emotional regulation, and physical demand) and for different levels of cardiac vagal control, as identified in our conceptual model and requirements 1 and 2. The bWell platform and exercise battery are described in detail in our previous publication [29]. The following is a brief summary of the three exercises:

1. *Tent*—an exercise designed for self-guided stress management in which the user is immersed in scenic views and asked to look around while focusing on their breathing
2. *Mole*—an exercise targeting response inhibition and cognitive control in which the user has a hammer in each hand and has to hit cylinders that pop up in front of them (only when the cylinder and hammer are of the same color)
3. *Stroll*—an exercise targeting sustained attention with a continuous performance test (CPT) in which the user has to press a button for each new shape that appears, except when it is a green diamond. The user is represented by a self-avatar and is passively displaced for a stroll in a natural scene.

To test our specific hypotheses for different stress demands or stressors (Table 1), we adapted the aforementioned bWell exercises (except for the *Mole* exercise for cognitive demand) as follows:

1. *Tent*
   - Nature scene to collect data on user
     - *Resting:* As per recommended guidelines for experimental settings, baseline HRV should be recorded in a “nontask situation that best controls for the presence of task comparison” [65]. Consequently, the administration of the exercise

XSL•FO
RenderX

was modified such that the user was asked to simply immerse themselves in the scene, limit physical movement, and breathe spontaneously for 5 minutes.

- *Recovery:* this scene was also selected to be administered after a given task to allow a period of recovery after stress demand. As a measure of recovery, a Likert scale implemented in VR was added to collect a report on mood right after the exposure to a stressor and after the recovery period in a relaxing nature environment.

- *City*—in addition to the nature scenes, to elicit emotional demand, we adapted the *Tent* exercise to include an environment such that the user is immersed in a busy city with noise pollution—roads with ongoing traffic, including audio honking cars and sirens.

2. *Stroll*—to exert physical demand, we adapted the Stroll exercise such that the user must physically displace themselves to advance in the scene. The user was asked to jog, keeping their heart rate at a moderate level of exertion (the predefined threshold was displayed as a gauge based on individual heart rate). The exercise could include a CPT, as described earlier, to exert both physical and cognitive demand (*Stroll+CPT*) or could be administered without the test as a purely physical task (*Stroll*).

The formulated hypotheses shown in Table 1 can now be systematically matched with our designed VR exercises (Table 2).

Table 2. Description of the hypotheses corresponding to our conceptual model with the designed virtual reality (VR) exercises. A distinct hypothesis was made for each phase—resting (baseline), reactivity, and recovery—and these hypotheses were tested with different VR exercises.

Cardiac vagal control level	Requirement	bWell exercise
1. Resting (baseline): Tent—nature scene, no specific task given	User to immerse themselves, limit physical movement, and breathe spontaneously as a baseline state	✗
2. Reactivity		
2a. Emotional demand: city—(passive) city scene	User to immerse in a busy city with excessive noise pollution to elicit emotional demand	✗
2b. Cognitive demand: mole—(active) targeting response inhibition	User to engage in a dynamic and bimanual go or no-go task to elicit cognitive demand	✗
2c. Physical demand: stroll—(active) physical running task	User to jog to create a physical demand, with heart rate biofeedback guiding them to maintain a consistent moderate level of activity	✗
2d. Physical and cognitive demand: stroll+CPT[a]−dual task with running task as 2c along with a cognitive task (response inhibition)	User to jog and engage in the task to elicit both physical and cognitive demands	✗
3. Recovery from demand: Tent—nature scene	User to immerse themselves and limit physical movement to recover	✗

[a]CPT: continuous performance test.

Hardware Identification (VR+Sensors)

Based on the identified requirements, we aimed to identify and integrate hardware for use with the bWell exercises. We focused on evaluating the suitability of the use of immersive VR technologies. The choice of hardware considered specialized equipment aiming to maximize the feeling of "*presence*" in virtual environments as well as the potential to democratize the use of VR for training more broadly. Consequently, we evaluated both a high-end system—*HTC Vive Pro Eye*—or use within dedicated facilities as well as a more accessible option—*Pico Neo 3 Eye*. The latter targets the opportunity for individuals to practice at home or in settings outside the laboratory (eg, on the field, classroom, and clinical).

We aimed to reproduce the required factors in the simulated environment with the integration of specialized hardware. A VR treadmill (KatVR) was identified to recreate the physical demands. We also aimed to capture stress via the inclusion of physiological sensing, with a focus on integrating heart-based measures. The *EmteqPRO* facemask was identified for use in this project because it is a commercially available "plug-and-play" system for use with immersive VR, provides researchers with access to raw data, and is equipped with automatic classification algorithms for processing raw heart rate data into HRV as well as classification of emotion [66]. The system can accommodate multimodal measures of EMG and HRV via the use of a facemask that can be directly integrated onto the VR headset. The Emteq device does not add any additional steps to the workflow for carrying out trials with immersive VR; instead, the user simply must put on the headset as they normally would.

XSL•FO
RenderX

Hardware Integration

A first evaluation was conducted to determine the suitability of the identified hardware.

Immersive VR Headset

The *HTC Vive Pro Eye* and the *Pico Neo 3 Eye* headsets were identified because they both offer eye tracking. In addition, the Emteq solution is available for both. The HTC Vive Pro Eye is a tethered device, meaning it requires a high-performance, VR-ready PC to operate. A tethered VR headset operates only when connected to a PC through a cable connection and offers superior graphics and fidelity in comparison to stand-alone headsets. The *Pico Neo 3 Eye* is a stand-alone headset, which is wireless, and the VR is run entirely within the headset. The benefits of a stand-alone headset include increased comfort and mobility, a simple setup, and a lower cost compared to tethered devices. However additional efforts to optimize the VR experience in terms of graphical rendering and design are required to minimize lags related to hardware performance.

It was determined that both headsets are suitable for use in the proposed study with certain adaptations. A wireless adapter was integrated into the HTC Vive Pro Eye, enabling untethered use with the treadmill. For the Pico Neo 3 Eye, an external dongle was required to enable screen casting to an external monitor, allowing the test administrator to view the participant's perspective within the VR during trials.

Emteq Integration With bWell in Unity

The Emteq face mask was integrated within *bWell* through the use of the “EmteqLabs Unity” plugin. Using the plugin, the EMG amplitude as well as the heart rate (beats per minute) of each participant was extracted. Using these extracted values, several widgets (Figure 2) were implemented in the 2 existing layers of the *bWell* user interface (UI): test administrator UI and participant UI. The first widget was part of the EmteqLabs plugin package and was used to display EMG amplitude (Figure 2A). The same widget also provided the administrator with an indication of sensor contact. Good contacts were represented by green dots on the widget, while bad contacts were represented by gray dots. The second widget was composed of 2 simple text boxes that were used to inform the test facilitator of the current connection status of the Emteq (Figure 2B). The third widget was a speedometer that was used to inform the participant whether they were in a heart rate zone of moderate effort (Figure 2C). The widget was implemented as a gauge to prompt participants to raise or lower their effort level, with the zone in the widget turning green when participants are at the correct range. Specifically, (zone turns green at correct range). A heart icon that grew and shrank based on a speed that matched the current beats per minute was also included in the speedometer.

Figure 2. Integration of physiological sensor with virtual reality (VR). Implementation of widgets in the user interface for integration of EmteqPRO (physiological sensor for heart rate variability) into VR platform (bWell): (A) widget for displaying electromyography (EMG) amplitude in VR, (B) widget for connection status of EmteqPRO, and (C) widget for displaying heart rate.

Treadmill Integration

The usability of donning a wireless immersive VR headset while running on the treadmill was tested. Running speed (vigorous vs moderate) was evaluated, and it was found that both vigorous and moderate running was acceptable in terms of usability for both headsets. However, it was seen that running on a treadmill was not intuitive and that a warm-up time was required. A test run was added to the protocol, such that the user could have some time to familiarize themselves with running on the treadmill before putting on the VR headset.

It was also seen that vigorous running may prove to be difficult for participants running on the treadmill for the first time. In addition, running on the treadmill while performing a cognitive exercise (*Mole* exercise) occasionally resulted in inadvertent errors from the user. As a result, the protocol was adjusted to have users run at a moderate pace (light jog) during the study. For the dual task (Stroll+CPT), a cognitive exercise with simpler user interaction was use. The exercise was changed from the initial choice to use *Mole* (requiring users to swing with their arms) to the *Stroll* (requiring users to pull a trigger on the controller).

Finally, the physical demand exercise (*Stroll*) was originally designed for passive displacement. In this task, it is the self-avatar taking a nature stroll while the actual user is not moving. To promote a higher sense of *presence* for the user, we added advancement in the VR scene based on the user’s own steps while advancing on the treadmill. The running pace was added as an output measure in the VR logs generated automatically after a session with bWell.

The conclusion from the hardware integration tests was that both the *HTC Vive Pro Eye* and *Pico Neo 3 Eye* were suitable. However, because the primary use at this stage was for study purposes, it was decided that it would be better to use the hardware that is designed for external administration of VR

XSL•FO
RenderX

versus self-administration with a headset. In the workflow for administration, the bWell platform provides a UI that allows the test facilitator to dynamically change the exercise settings while a user is immersed in VR [29]. Consequently, it would be better suited for the iterative approach in finding the optimal exercise settings for the study. Thus, the decision was to use the *HTC Vive Pro Eye* for the study and the ensuing verification trials.

VR2: Verification Phase

In this phase, we focused on conducting early trials for evaluation and verification of feasibility, acceptability, tolerability, and initial efficacy (VR2 phase).

Study Design

We designed a study to systematically test our hypotheses (Figure 3). The study protocol consisted of a 5-minute initial baseline consistent with what is used in the HRV literature (Table S1 in Multimedia Appendix 1). During baseline, participants were in a standing position with their eyes open to optimally match the body position used during task conditions. Participants experienced a neutral nature scene. The baseline HRV was collected to establish a reference point. Participants then experienced alternating blocks of stressor tasks interspersed with recovery (Tent nature scenes). Stressor task presentation order was pseudorandomized across participants. Each participant performed all 4 exercises (stress conditions). Each stressor condition was followed by a 2-minute recovery period in the *Tent* (Table S1 in Multimedia Appendix 1).

Figure 3. Description of the study design. A within-participants repeated measures study design was implemented. It comprised an initial baseline (5 minutes), followed by 4 blocks of stressor (2 minutes each) and recovery (2 minutes each) periods. The stressors were 4 virtual reality (VR) exercises targeting emotional demand, cognitive demand, physical demand, and cognitive+physical demand. The sequence of stressors was pseudorandomized. CPT: continuous performance test.

The measures of VR2 trials were captured as follows:

- *Acceptability:* user experience, as assessed by the game user experience satisfaction scale (GUESS)-18 gamer user experience satisfaction scale [67] and the System Usability Scale [68]
- *Tolerability:* user comfort and any side effects from the use of technology, as assessed by the Simulator Sickness Questionnaire (SSQ; captured both before and after VR exposure)
- *Feasibility:* ability to elicit the intended stress response, as measured via physiological response as well as self-report

For the self-report measure, the NASA-TLX [51] and single-item rating for stress questionnaires were used.

Mood Likert scales were also automatically administered in immersive VR before and after the recovery exercise. Participant and VR views were also recorded. After completing the bWell session, participants reviewed a screen recording of the completed virtual task, provided qualitative feedback, and completed the set of questionnaires.

It may be noted that while the NVIM and VTT predict the dependence of vagal tone on task conditions, the NASA-TLX measures 6 aspects of workload. The study design involved 5 conditions, 1 without a stressor and 4 others in which each condition features a different exercise representing specific stressors based on the NASA-TLX (summarized in Table 3). To effectively assess the effects of each condition on stress responses, we proposed a within-*participants* design, where each participant completed all conditions of the experiment. The order of the stressor conditions was pseudorandomized across participants. The approach taken to experimentally isolate the contributions of each type of stress required the removal of

XSL•FO
RenderX

performance criteria (eg, score) from the tasks, as they may create confounds in the stressor type not related to the experimental manipulation of the condition. However, the cumulative effects of stressors were also investigated using a physical+cognitive stressor dual task, as the inclusion of multiple simultaneous stressors is more representative of everyday stressful situations.

Table 3. Description of the different types of stressor demands for different virtual reality (VR) exercises. bWell VR exercises elicited different types of demands, as defined by the NASA-Task Load Index workloads. The main stress or demand captured for each VR exercise is shown in italics.

	Mental demand	Physical demand	Temporal demand	Performance	Effort	Frustration
Relaxing nature environment (*tent*: nature scene)	No	No	No	No	No	No
Emotional stressor (*city*: busy city scene)	No	No	No	No	No	*Yes* [a]
Cognitive stressor (*mole*: dynamic bimanual go or no-go task)	*Yes* [a]	No	Yes[a]	Yes[a]	Yes[a]	No
Physical stressor (*stroll*: running task)	No	*Yes* [a]	No	Yes[a]	Yes[a]	No
Physical+cognitive stressor (*stroll+CPT:* running with continuous performance test)	*Yes* [a]	*Yes* [a]	No	Yes[a]	Yes[a]	No

[a]Demands elicited by the bWell virtual reality exercises.

We set thresholds to the NASA-TLX as a self-reported measure to ensure that the designed VR exercises targeted the intended type of demand, that is, *cognitive*, *emotional*, *physical*, and *dual (ie, physical+cognitive)*. We defined the following threshold in terms of NASA-TLX score (score range=0-10): a score>5 indicates that the demand is being elicited (eg, *City* for emotional demand, *Mole* for cognitive demand, *Stroll* for physical demand, and *Stroll+CPT* for physical and cognitive demand; Table 4). An additional aim was to not elicit demand related to poor performance because this may lead to a potential confound. In the NASA-TLX, performance is reverse coded (ie, 0=perfect performance). As such the threshold was set to<4, so that that participant should feel as if they performed well in all exercises. Finally, since all exercises invoked a type of demand except the one involving emotional demand, they were expected to involve an element of effort.

Table 4. Description of the thresholds for NASA-Task Load Index (NASA-TLX) scores for eliciting different stress demands. We set the NASA-TLX subscale scores to above the threshold, for the different types of stress demands that were elicited by each virtual reality exercise in Table 3.

	Frustration	Physical demand	Mental demand	Performance
City	>5[a]	—[b]	—	<4
Mole	<4	<4	>5[a]	<4
Stroll	—	>5[a]	—	<4
Stroll+CPT[c]	—	>5[a]	>5[a]	—

[a]NASA-TLX subscale scores that were above the threshold.

[b]Not available.

[c]CPT: continuous performance test.

More specifically, we targeted the following scores:

- *City*—a NASA-TLX score of >5 for *frustration*, which would indicate an emotional demand. We also intended for all other NASA-TLX scores to remain <4 to solely evoke an emotional response. Given the passive nature of this exercise, other demands should not be induced.
- *Mole*—a NASA-TLX score of >5 for *mental*, which would indicate a cognitive demand. A score of >5 would indicate that these demands were elicited. We also targeted *frustration* and *physical* NASA-TLX scores of <4 to solely target demands related to cognitive performance.
- *Stroll*—a NASA-TLX score of >5 for *physical*, which would indicate a physical demand. A demand of >5 would indicate that this was the case. We also targeted *mental* and *frustration* NASA-TLX scores of <4 to solely target demands related to physical performance.
- *Stroll+CPT*−a NASA-TLX score of >5 for *mental* and *physical*, which would indicate a dual cognitive and physical demand. We also intended to maintain a *frustration* score of <4 because we wanted to target solely cognitive and physical demands.

In addition, we sought to verify that *mental* demands are comparable between the *Mole* and *Stroll+CPT* tasks.

Verification Trials

Verification trials were conducted using HTC Vive Pro Eye on 3 healthy participants according to the protocol defined during the study design.

XSL•FO
RenderX

Ethical Considerations

Ethics approval for the study was obtained from the Research Ethics Board (REB) of the National Research Council (NRC 2020-2023 and NRC 2023-46) and from Defence Research and Development Canada Human Research Ethics Committee (HREC 2022-032). Informed consent was obtained before the start of the study. Data were deidentified before any analysis. There was no compensation paid to the participants.

Results

The results of the verification trials are reported below in terms of tolerability, acceptability, and feasibility.

Tolerability: Cybersickness During the VR Sessions

We first checked whether the VR exercises designed in our study induced cybersickness using SSQ [69]. Table S2 in Multimedia Appendix 1 shows the scores for SSQ for pre- and post-VR sessions for 3 participants. Total SSQ scores were obtained from the scores of the 16 items of the questionnaire [70]. Results showed negligible score (total SSQ score: 0-5), which are within the acceptable range in the literature [71]. Thus, our results show that the participants did not experience adverse effects, especially cybersickness during the VR exercises.

Acceptability: User Experience

Using the GUESS-18 scale [67], we checked whether the VR exercises designed in our study were immersive. GUESS-18 has 9 subscales with 2 statements associated with each subscale. Of the 9 subscales, we did not include "social connectivity," as there was no interaction between participants for the VR exercises. Table S3 in Multimedia Appendix 1 shows the scores of the 8 subscales of GUESS-18 for the 3 participants. The subscale scores of the GUESS were computed by averaging the scores within each subscale. Then, the overall GUESS score was computed by summing the subscale scores [67]. We observed an overall GUESS score of 43.17 (Table S3 in Multimedia Appendix 1), which aligns well with previous findings [72] and indicates the usability and immersiveness of the VR exercises designed in our study. Regarding the individual subscales, we observed greater average scores for *usability*, *play engrossment*, and *personal gratification* and lower average scores for *creative freedom* and *enjoyment* (Figure 4).

Figure 4. Findings of user acceptability. User ratings on the game user experience satisfaction scale (GUESS)-18 questionnaires showing greater average scores for the subscales usability, play engrossment, and personal gratification and lower average scores for creative freedom and enjoyment. An overall GUESS score of 43.17 (Table S3 was observed, which aligns well with previous findings, indicating the usability and immersiveness of the virtual reality exercises designed in our study.

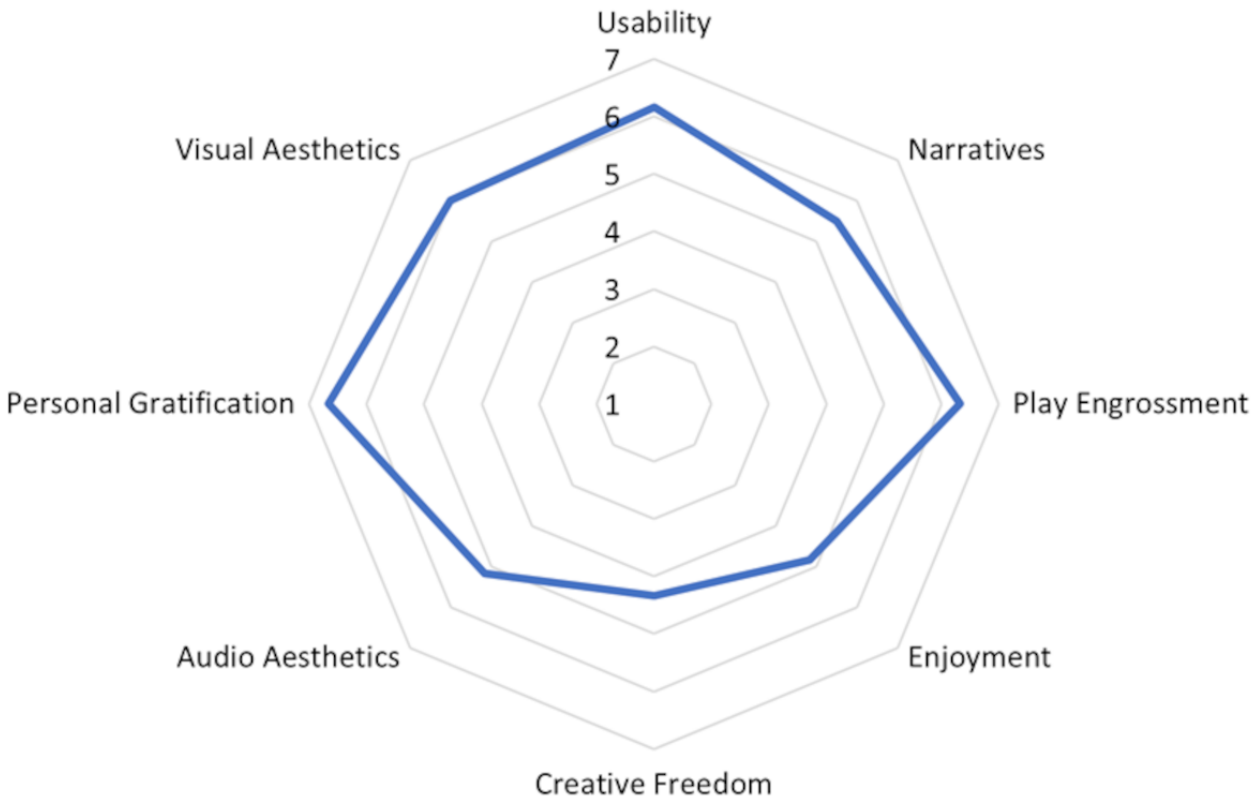

Feasibility: Do the VR Exercises Induce Genuine Stress Responses (Perceived)?

We evaluated the NASA-TLX scores to investigate the types of stressors and their associated demands. Figure 5 summarizes the data for NASA-TLX from verification trials of 3 participants.

XSL•FO
RenderX

Figure 5. Comparative analysis of different types of demands for different stressors. User ratings on the NASA-Task Load Index (NASA-TLX) revealed that participants perceived different types of demands (eg, mental, physical, and temporal) for the different VR exercises. Here, P1, P2, and P3 denote participant 1, participant 2, and participant 3, respectively, and the dashed line represents the threshold set for different demands associated with the different stressors (exercises). CPT: continuous performance test.

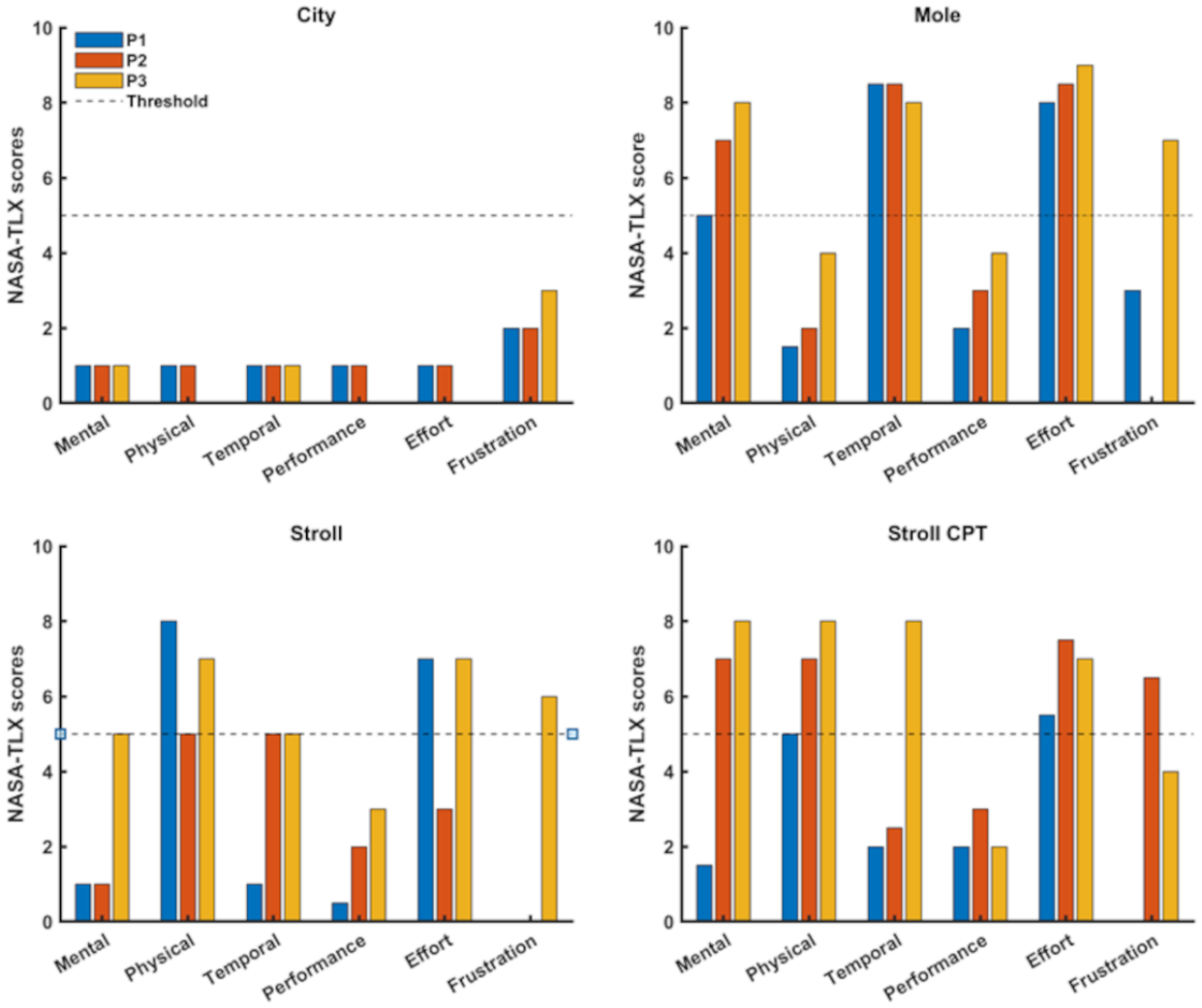

On the basis of the designed thresholds for the different stressors and their associated demands (refer to the Methods section), we observed the following for each of the bWell exercises (Figure 5), which led to further iterative development:

- *City*: a NASA-TLX rating of >5 for *frustration* was not achieved, indicating a need for adapting the exercise settings and creating new functionalities. All other demands were also not elicited (<4). In particular, users felt that they performed well (<4), and as such, it was verified that there would not be any confounds in *frustration* being elicited due to emotional response related to feelings of poor performance rather than VR stimuli.
- *Mole*: a cognitive demand was elicited with a *mental* demand of >5. We also observed a high level of *temporal* demand and associated *effort*.
- *Stroll*: a *physical* demand and associated *effort* was elicited (>5).
- *Stroll+CPT*: a dual cognitive (*mental*) and *physical* demand was elicited (>5), along with associated *effort*.

Because cognitive demand was elicited in both tasks, similar to the Mole task, the 2 can be compared in the analysis of cognitive demand, even though they involve different cognitive tasks. It appears that *temporal* demand may also be elicited but was not conclusive at this stage because of high variability in the ratings. This may be clarified with results from additional participant trials from the full study. From Table 4, we observed that this demand was not consistently present. This may be attributed to the fact that this was a dual task, where the *temporal* demand may be felt for one aspect (eg, running) but not for the other (eg, cognitive task), and when prompted by the NASA-TLX, the participant may respond with only 1 task in mind.

Overall, the results showed that we achieved eliciting different stress demands as intended in our study design: for *Stroll* to exert a *physical* demand, *Mole* to exert a *mental* demand, and a combination of exercises (*Stroll+CPT*) to exert both *mental* and *physical* demands. At the same time, we observed that participants felt they performed well in all exercises, minimizing confounds due to feelings of poor performance. These verification trials allowed us to finalize the exercise configurations to be used for the full study trials.

XSL•FO
RenderX

Feasibility: Do the VR Exercises Induce Genuine Stress Responses (Physiological)?

To evaluate the physiological stress responses, we examined the HRV measures obtained, specifically the root mean square of successive differences (RMSSD) between normal heartbeats. We adopted RMSSD as the HRV measure because previous studies have shown that RMSSD reflects vagal tone [38,73] and is relatively free of respiratory influences. It was seen that the data were quite noisy, artifacts were present, and the RMSSD values from the raw PPG data were too high. For example, values in the range of 150 to 300 were obtained, which is much higher than normal. Discussion with the Emteq team revealed that this was likely because the measures were sampled in a dynamic VR environment, particularly where in some instances, users were physically running during PPG data acquisition. Emteq advised using their signal quality metric and using only the data where good quality was indicated.

To have better control of the data processing, the team decided instead to implement an in-house data processing pipeline. For this, the raw PPG signals were processed using the Python library *HeartPy* [74]. On the basis of the *HeartPy* documentation, which states that RMSSD is extremely sensitive to incorrectly detected heartbeat locations, any RMSSD value artifacts (RMSSD does not typically exceed 130) were removed and peak interpolation was used to minimize the number of incorrectly placed peaks due to low sampling rate [74]. After these steps, an improved signal-to-noise ratio was observed and the resultant RMSSD values fell well within the normal range expected (between 30 and 70).

We then compared HRV across each of the different conditions for 2 individuals. In the example shown in Figure 6, it can be seen that a given task appears to elicit a change in response. It is also evident that baseline, reactivity, and recovery HRV patterns were distinct for the 2 sample individuals. As can be seen in the lower panel of Figure 6, some instances of residual noise remain in the Mole exercise after processing. Given that the data are being collected during interactive tasks with a good degree of movement, a high noise-to-signal ratio was expected. The next steps will involve exploring other techniques, particularly for motion artifact removal, with the goal of further reducing noise in the signal.

Figure 6. Characterization of distinct patterns of heart rate variability (HRV) for different virtual reality (VR) exercises. HRV data (root mean square of successive differences [RMSSD]) for the different VR (bWell) exercises are shown for 2 participants (row 1 for participant 1 and row 2 for participant 2) along y-axis and time along x-axis. The plots illustrate distinct HRV responses for the different exercises as well as individual variability in the HRV responses. CPT: continuous performance test.

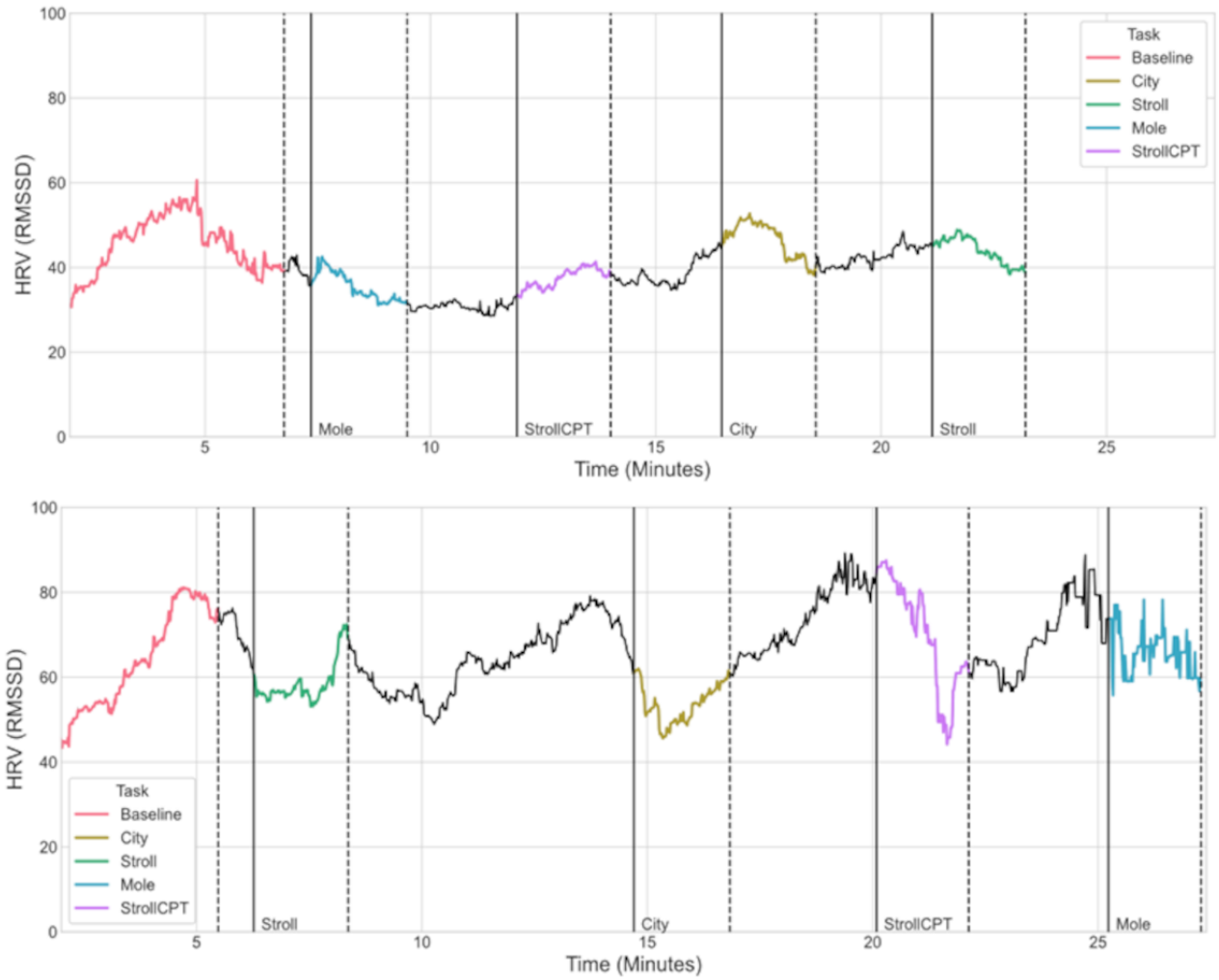

XSL•FO
RenderX

Discussion

Principal Findings

In this paper, we described the development process of a study design for testing the ability of immersive VR exercises to elicit genuine stress responses. The study development followed recommendations for use of VR in clinical trials [32] as well as guidance in developing complex interventions for improving health care [33]. On the basis of these guidelines, our study comprised a combination of *prevailing theories and frameworks* (eg, NVIM and the VTT), *experiences of users*, and *iterative actions during development*, which resulted in concrete outcomes [33]. More specifically, we built on our validated VR platform *bWell* [29] to integrate physiological sensors (PPG) and assessed how different types of VR exercises target different types of stressors (characterized by the NASA-TLX). The results of verification trials showed that the designed VR exercises induced negligible cybersickness, were well tolerated, and demonstrated high user engagement. In addition, in alignment with our hypotheses, the different VR exercises successfully elicited the intended stress demands (as measured with NASA-TLX subscales, eg, mental, physical, and temporal demands), along with distinct physiological responses.

In line with the recommended guidelines for VR trial design in the health domain [44], our study design considered both tolerability and acceptability. Keeping in mind user comfort during VR, the exercises were designed in such a way that the duration for the entire session lasted only 25 minutes. In addition, during the VR session, every exercise was followed by a 2-minute recovery state. By allowing the participants to experience alternating blocks of stressor tasks interspersed with recovery, this study design was used as a guide to evaluate the technical feasibility and scope for the phase of verification trials. Indeed, the results of the verification trials showed user comfort and tolerability, especially VR as an acceptable framework with *negligible* experience of cybersickness (assessed with SSQ scores; Table S2 in Multimedia Appendix 1). As a measure of acceptability, results of verification trials using the GUESS-18 scores showed a high degree of "*presence*" during our VR sessions (Table S3 in Multimedia Appendix 1 and Figure 4), consistent with earlier findings [72]. Because the sense of "presence" (or the subjective feeling of "being there") characterizes the experience of users, results showing a high degree of presence indicate that the VR exercises designed for the study created the sensation of physical presence and resulted in users behaving and reacting as in the real world, highlighting the importance of a good user experience and engagement. Best practices for HRV data collection were also considered for the study design. In experimental settings, it is recommended that the baseline HRV should be recorded in a "nontask situation that best controls for the presence of task comparisons" [65]. Consequently, in this study, baseline was recorded in VR and in a similar position to how the participants would complete the task. Specifically, baseline was recorded with the user on the treadmill and passively experiencing the VR environments. The session to collect baseline HRV was designed as follows: eyes open, standing, spontaneous breathing, and free to look around and immerse themselves in the VR environment. In addition and according to guidance on HRV reporting, RMSSD was selected as the HRV measure because it is appropriate for short sampling times (ie, <5 minutes) [41].

While the results of verification trials determined the feasibility and validity of our study protocol, some results indicated that modifications were also required. For example, the NASA-TLX scores showed high levels of *Frustration*, which captured participants' physical demands or frustration related to performance and not necessarily due to an emotional stressor. Because our original goal was to investigate the effects of emotion, we made certain modifications in our bWell exercises to elicit emotional responses in the participants, such as amplifying the noise pollution in the virtual environment. The results of the verification trials also indicated that we need to modify the administration of the NASA-TLX, which currently probes *Frustration* related to both performance and emotion. Instead of asking "How insecure, discouraged, irritated, stressed and annoyed were you," the question will be modified to include only the elements related to emotion "How irritated and/or annoyed were you?"

Implications

It may be mentioned that the overarching aim for the integration of behavioral and physiological outputs was to eventually enable systematic and individualized assessment of stress response. Responses to stress are often multifaceted with a profile specific to the individual. A given stressor may cause a response in one person but not necessarily in another, which may be attributed to the influence of personal experiences. The result is that everyone does not experience stress in the same way. In an effort to standardize the demands placed on individuals, we relied on several mechanisms to adapt the VR exercises in real time and customized them to the participant. For cognitive demand, we aimed to elicit a uniform cognitive demand among varied individuals. To do so, we leveraged the adaptive algorithms in *bWell* to control the difficulty level based on performance (for *Mole* exercise). The adaptive mechanism is designed to maintain an 80% success rate, ensuring the exercise is neither too easy (boring) nor too difficult (demotivating). For physical demand, we made use of the heart rate biofeedback to ensure consistent physical demand because running at a specified speed may be more or less exerting for one individual versus another. To achieve this, a heart rate gauge was displayed to the participant to encourage them to maintain a level of physical activity that corresponds to moderate intensity. This measure is additionally tailored to the individual because it considers the age of the participant, where moderate physical activity is calculated at 50% to 70% of maximum heart rate (maximum heart rate=220 – age).

Study Limitations

There were certain limitations in our study. First, participants reported issues with disengagement and a lack of frustration during the VR exercise (city). The reason for this could be the low resolution of the visual assets in the scene, a concern that will be addressed in future work by improving the resolution. Second, there could be a learning effect in terms of interactions in VR for participants who had no prior VR experience. In the full study (planned next steps), participants will be given

XSL•FO
RenderX

tutorials for immersion in VR. Finally, the verification trials were done with few participants (N=3), limiting our ability to generalize findings. However, we would like to emphasize that our study was designed to identify, create, and implement an experimental setup for testing the ability of immersive VR exercises to elicit genuine responses to different types of stressors. In addition, it aimed to assess the feasibility, tolerability, and acceptability of these exercises through verification trials. This is consistent with previous studies that focused on study development [75]. Because the purpose of the verification trials was to show the feasibility, tolerability, and user acceptability, we collected data from only 3 participants, with the intention to run full trials in the future. It is worth noting that, despite the limited number of participants, we were able to demonstrate the feasibility, tolerability, and user acceptability of the study.

Conclusions

In conclusion, we identified, designed, and implemented an experimental setup for testing the ability of immersive VR exercises to elicit genuine responses corresponding to different types of stressors and tested its feasibility, tolerability, and acceptability using verification trials. While only a small number of trials were carried out with participants, the verification trials provided the support required to carry out a full study (planned next steps).

A strong point of our study design was the collection of several physiological measures that permitted the collection of quantitative measures complementing that of self-report with the NASA-TLX. With the capture of measures such as HRV and facial expressions, it will be possible to compare how the bWell exercises may physiologically elicit response versus what users perceive to be elicited. Such evaluations may serve to obtain a more complete picture of the in-situ user state and may permit gaining insights on the use of traditional self-report, which has been critiqued as an unreliable index of functional outcomes [76]. In this regard, future work will explore multivariate analysis such as partial least squares [77] and ML techniques such as random forest [78] for predicting individual user responses to different stressors based on multiple input data, including sensor data (eg, HRV and facial expressions) and VR-based behavioral data (eg, response time). Such work will help in identifying and delineating the individual features that confer stress resilience and will guide us in helping individuals reach their full potential.

Acknowledgments

The authors would like to thank the participants who agreed to complete the verification trials. The authors also thank Alejandro Olivares Hernandez for help in integrating KatVR to the *bWell* platform.

Data Availability

The datasets generated and analyzed during this study are not publicly available due to legislation regarding privacy and security. None of the data for the experiments reported here are available, and none of the experiments were preregistered.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Description of the timeline for different task and recovery, Simulator Sickness Questionnaire scores for pre– and post–virtual reality session, and subjective ratings of game user experience satisfaction scale-18 questionnaires.
[DOCX File , 390 KB - formative_v9i1e64492_app1.docx]

References

1. Yaribeygi H, Panahi Y, Sahraei H, Johnston TP, Sahebkar A. The impact of stress on body function: a review. EXCLI J 2017;16:1057-1072 [FREE Full text] [doi: 10.17179/excli2017-480] [Medline: 28900385]
2. Chen B, Wang L, Li B, Liu W. Work stress, mental health, and employee performance. Front Psychol 2022 Nov 8;13:1006580 [FREE Full text] [doi: 10.3389/fpsyg.2022.1006580] [Medline: 36425815]
3. Alberdi A, Aztiria A, Basarab A. Towards an automatic early stress recognition system for office environments based on multimodal measurements: a review. J Biomed Inform 2016 Mar;59:49-75 [FREE Full text] [doi: 10.1016/j.jbi.2015.11.007] [Medline: 26621099]
4. Pedrosa AL, Bitencourt L, Fróes AC, Cazumbá ML, Campos RG, de Brito SB, et al. Emotional, behavioral, and psychological impact of the COVID-19 pandemic. Front Psychol 2020 Oct 2;11:566212 [FREE Full text] [doi: 10.3389/fpsyg.2020.566212] [Medline: 33117234]
5. Saladino V, Algeri D, Auriemma V. The psychological and social impact of COVID-19: new perspectives of well-being. Front Psychol 2020 Oct 2;11:577684 [FREE Full text] [doi: 10.3389/fpsyg.2020.577684] [Medline: 33132986]
6. Ebner K, Singewald N. Individual differences in stress susceptibility and stress inhibitory mechanisms. Curr Opin Behav Sci 2017 Apr;14:54-64. [doi: 10.1016/j.cobeha.2016.11.016]

XSL•FO
RenderX

7. Sapolsky RM. Stress and the brain: individual variability and the inverted-U. Nat Neurosci 2015 Oct 25;18(10):1344-1346. [doi: 10.1038/nn.4109] [Medline: 26404708]
8. Bangasser DA, Valentino RJ. Sex differences in stress-related psychiatric disorders: neurobiological perspectives. Front Neuroendocrinol 2014 Aug;35(3):303-319 [FREE Full text] [doi: 10.1016/j.yfrne.2014.03.008] [Medline: 24726661]
9. Russo SJ, Murrough JW, Han MH, Charney DS, Nestler EJ. Neurobiology of resilience. Nat Neurosci 2012 Nov 14;15(11):1475-1484 [FREE Full text] [doi: 10.1038/nn.3234] [Medline: 23064380]
10. McEwen BS. Sex, stress and the hippocampus: allostasis, allostatic load and the aging process. Neurobiol Aging 2002;23(5):921-939. [doi: 10.1016/s0197-4580(02)00027-1] [Medline: 12392796]
11. Ulrich-Lai YM, Herman JP. Neural regulation of endocrine and autonomic stress responses. Nat Rev Neurosci 2009 Jun 13;10(6):397-409 [FREE Full text] [doi: 10.1038/nrn2647] [Medline: 19469025]
12. Clark A, Mach N. Exercise-induced stress behavior, gut-microbiota-brain axis and diet: a systematic review for athletes. J Int Soc Sports Nutr 2016 Apr 01;13(1):43 [FREE Full text] [doi: 10.1186/s12970-016-0155-6] [Medline: 27924137]
13. McCraty R, Shaffer F. Heart rate variability: new perspectives on physiological mechanisms, assessment of self-regulatory capacity, and health risk. Glob Adv Health Med 2015 Jan;4(1):46-61 [FREE Full text] [doi: 10.7453/gahmj.2014.073] [Medline: 25694852]
14. Amano M, Kanda T, Ue H, Moritani T. Exercise training and autonomic nervous system activity in obese individuals. Med Sci Sports Exerc 2001 Aug;33(8):1287-1291 [FREE Full text] [doi: 10.1097/00005768-200108000-00007] [Medline: 11474328]
15. Dong JG. The role of heart rate variability in sports physiology. Exp Ther Med 2016 May;11(5):1531-1536 [FREE Full text] [doi: 10.3892/etm.2016.3104] [Medline: 27168768]
16. Mongin D, Chabert C, Extremera MG, Hue O, Courvoisier DS, Carpena P, et al. Decrease of heart rate variability during exercise: an index of cardiorespiratory fitness. PLoS One 2022 Sep 2;17(9):e0273981 [FREE Full text] [doi: 10.1371/journal.pone.0273981] [Medline: 36054204]
17. Tobaldini E, Nobili L, Strada S, Casali KR, Braghiroli A, Montano N. Heart rate variability in normal and pathological sleep. Front Physiol 2013 Oct 16;4:294 [FREE Full text] [doi: 10.3389/fphys.2013.00294] [Medline: 24137133]
18. Siepmann M, Weidner K, Petrowski K, Siepmann T. Heart rate variability: a measure of cardiovascular health and possible therapeutic target in dysautonomic mental and neurological disorders. Appl Psychophysiol Biofeedback 2022 Dec 22;47(4):273-287 [FREE Full text] [doi: 10.1007/s10484-022-09572-0] [Medline: 36417141]
19. Ahn JW, Ku Y, Kim HC. A novel wearable EEG and ECG recording system for stress assessment. Sensors (Basel) 2019 Apr 28;19(9):23 [FREE Full text] [doi: 10.3390/s19091991] [Medline: 31035399]
20. Attar ET, Balasubramanian V, Subasi E, Kaya M. Stress analysis based on simultaneous heart rate variability and EEG monitoring. IEEE J Transl Eng Health Med 2021;9:1-7. [doi: 10.1109/jtehm.2021.3106803]
21. Shilton AL, Laycock R, Crewther SG. The Maastricht acute stress test (MAST): physiological and subjective responses in anticipation, and post-stress. Front Psychol 2017 Apr 19;8:567 [FREE Full text] [doi: 10.3389/fpsyg.2017.00567] [Medline: 28469586]
22. Frisch JU, Häusser JA, Mojzisch A. The trier social stress test as a paradigm to study how people respond to threat in social interactions. Front Psychol 2015 Feb 02;6:14 [FREE Full text] [doi: 10.3389/fpsyg.2015.00014] [Medline: 25698987]
23. Aspiotis V, Miltiadous A, Kalafatakis K, Tzimourta KD, Giannakeas N, Tsipouras MG, et al. Assessing electroencephalography as a stress indicator: a VR high-altitude scenario monitored through EEG and ECG. Sensors (Basel) 2022 Aug 03;22(15):5792 [FREE Full text] [doi: 10.3390/s22155792] [Medline: 35957348]
24. Ahmad Z, Rabbani S, Zafar MR, Ishaque S, Krishnan S, Khan N. Multilevel stress assessment from ECG in a virtual reality environment using multimodal fusion. IEEE Sensors J 2023 Dec 1;23(23):29559-29570. [doi: 10.1109/jsen.2023.3323290]
25. Bouchard S, Bernier F, Boivin É, Guitard T, Laforest M, Dumoulin S, et al. Modes of immersion and stress induced by commercial (off-the-shelf) 3D games. J Def Model Simul 2012 May 04;11(4):339-352. [doi: 10.1177/1548512912446359]
26. van Dammen L, Finseth TT, McCurdy BH, Barnett NP, Conrady RA, Leach AG, et al. Evoking stress reactivity in virtual reality: a systematic review and meta-analysis. Neurosci Biobehav Rev 2022 Jul;138:104709. [doi: 10.1016/j.neubiorev.2022.104709] [Medline: 35644278]
27. Hemmeter U, Stormer R, Mager R, Kuntze M, Mueller-Spahn F, Hennig J, et al. Modification of virtual reality combined with a mental task stimulates cortisol in volunteers. Neuropsychobiology 2005 May 2;51(3):165-172. [doi: 10.1159/000085208] [Medline: 15838187]
28. Breuninger C, Sláma DM, Krämer M, Schmitz J, Tuschen-Caffier B. Psychophysiological reactivity, interoception and emotion regulation in patients with agoraphobia during virtual reality anxiety induction. Cogn Ther Res 2016 Oct 15;41(2):193-205. [doi: 10.1007/S10608-016-9814-9]
29. Shaigetz VG, Proulx C, Cabral A, Choudhury N, Hewko M, Kohlenberg E, et al. An immersive and interactive platform for cognitive assessment and rehabilitation (bWell): design and iterative development process. JMIR Rehabil Assist Technol 2021 Nov 03;8(4):e26629 [FREE Full text] [doi: 10.2196/26629] [Medline: 34730536]
30. Thayer JF, Hansen AL, Saus-Rose E, Johnsen BH. Heart rate variability, prefrontal neural function, and cognitive performance: the neurovisceral integration perspective on self-regulation, adaptation, and health. Ann Behav Med 2009 Apr 8;37(2):141-153. [doi: 10.1007/s12160-009-9101-z] [Medline: 19424767]

XSL•FO
RenderX

31. Laborde S, Mosley E, Mertgen A. Vagal tank theory: the three Rs of cardiac vagal control functioning - resting, reactivity, and recovery. Front Neurosci 2018 Jul 10;12:458 [FREE Full text] [doi: 10.3389/fnins.2018.00458] [Medline: 30042653]
32. Birckhead B, Khalil C, Liu X, Conovitz S, Rizzo A, Danovitch I, et al. Recommendations for methodology of virtual reality clinical trials in health care by an international working group: iterative study. JMIR Ment Health 2019 Jan 31;6(1):e11973 [FREE Full text] [doi: 10.2196/11973] [Medline: 30702436]
33. O'Cathain A, Croot L, Duncan E, Rousseau N, Sworn K, Turner KM, et al. Guidance on how to develop complex interventions to improve health and healthcare. BMJ Open 2019 Aug 15;9(8):e029954 [FREE Full text] [doi: 10.1136/bmjopen-2019-029954] [Medline: 31420394]
34. Delliaux S, Delaforge A, Deharo JC, Chaumet G. Mental workload alters heart rate variability, lowering non-linear dynamics. Front Physiol 2019 May 14;10:565 [FREE Full text] [doi: 10.3389/fphys.2019.00565] [Medline: 31156454]
35. Low A, McCraty R. Heart rate variability: new perspectives on assessment of stress and health risk at the workplace. Heart Mind 2018;2(1):16-27. [doi: 10.4103/hm.hm_11_18]
36. Eggenberger P, Annaheim S, Kündig KA, Rossi RM, Münzer T, de Bruin ED. Heart rate variability mainly relates to cognitive executive functions and improves through exergame training in older adults: a secondary analysis of a 6-month randomized controlled trial. Front Aging Neurosci 2020 Jul 15;12:197 [FREE Full text] [doi: 10.3389/fnagi.2020.00197] [Medline: 32760267]
37. Grässler B, Hökelmann A, Cabral RH. Resting heart rate variability as a possible marker of cognitive decline. Kinesiol 2020;52(1):72-84. [doi: 10.26582/k.52.1.9]
38. Thayer JF, Lane RD. A model of neurovisceral integration in emotion regulation and dysregulation. J Affect Disord 2000 Dec;61(3):201-216. [doi: 10.1016/s0165-0327(00)00338-4] [Medline: 11163422]
39. Thayer JF, Lane RD. Claude Bernard and the heart-brain connection: further elaboration of a model of neurovisceral integration. Neurosci Biobehav Rev 2009 Mar;33(2):81-88. [doi: 10.1016/j.neubiorev.2008.08.004] [Medline: 18771686]
40. Pavlov VA, Tracey KJ. The vagus nerve and the inflammatory reflex--linking immunity and metabolism. Nat Rev Endocrinol 2012 Dec 21;8(12):743-754 [FREE Full text] [doi: 10.1038/nrendo.2012.189] [Medline: 23169440]
41. Task Force of the European Society of Cardiology the North American Society of Pacing Electrophysiology. Heart rate variability: standards of measurement, physiological interpretation and clinical use. Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology. Circulation 1996 Mar 01;93(5):1043-1065. [Medline: 8598068]
42. Liu CY, Mueller MH, Grundy D, Kreis ME. Vagal modulation of intestinal afferent sensitivity to systemic LPS in the rat. Am J Physiol Gastrointest Liver Physiol 2007 May;292(5):G1213-G1220 [FREE Full text] [doi: 10.1152/ajpgi.00267.2006] [Medline: 17204546]
43. Ozaki N, Sengupta JN, Gebhart GF. Mechanosensitive properties of gastric vagal afferent fibers in the rat. J Neurophysiol 1999 Nov 01;82(5):2210-2220 [FREE Full text] [doi: 10.1152/jn.1999.82.5.2210] [Medline: 10561400]
44. Laborde S, Mosley E, Thayer JF. Heart rate variability and cardiac vagal tone in psychophysiological research - recommendations for experiment planning, data analysis, and data reporting. Front Psychol 2017;8:213 [FREE Full text] [doi: 10.3389/fpsyg.2017.00213] [Medline: 28265249]
45. Shaffer F, Ginsberg JP. An overview of heart rate variability metrics and norms. Front Public Health 2017 Sep 28;5:258 [FREE Full text] [doi: 10.3389/fpubh.2017.00258] [Medline: 29034226]
46. Dekker JM, Crow RS, Folsom AR, Hannan PJ, Liao D, Swenne CA, et al. Low heart rate variability in a 2-minute rhythm strip predicts risk of coronary heart disease and mortality from several causes: the ARIC Study. Atherosclerosis risk in communities. Circulation 2000 Sep 12;102(11):1239-1244. [doi: 10.1161/01.cir.102.11.1239] [Medline: 10982537]
47. Grassi G. Assessment of sympathetic cardiovascular drive in human hypertension: achievements and perspectives. Hypertension 2009 Oct;54(4):690-697. [doi: 10.1161/hypertensionaha.108.119883]
48. Ronkainen E, Korpelainen JT, Heikkinen E, Myllylä VV, Huikuri HV, Isojärvi JI. Cardiac autonomic control in patients with refractory epilepsy before and during vagus nerve stimulation treatment: a one-year follow-up study. Epilepsia 2006 Mar 06;47(3):556-562 [FREE Full text] [doi: 10.1111/j.1528-1167.2006.00467.x] [Medline: 16529621]
49. Porges SW. The polyvagal perspective. Biol Psychol 2007 Mar;74(2):116-143 [FREE Full text] [doi: 10.1016/j.biopsycho.2006.06.009] [Medline: 17049418]
50. Smith R, Thayer JF, Khalsa SS, Lane RD. The hierarchical basis of neurovisceral integration. Neurosci Biobehav Rev 2017 May;75:274-296. [doi: 10.1016/j.neubiorev.2017.02.003] [Medline: 28188890]
51. Hart SG, Staveland LE. Development of NASA-TLX (task load index): results of empirical and theoretical research. Adv Psychol 1988;52:139-183. [doi: 10.1016/S0166-4115(08)62386-9]
52. Porges SW. Cardiac vagal tone: a physiological index of stress. Neurosci Biobehav Rev 1995;19(2):225-233. [doi: 10.1016/0149-7634(94)00066-a] [Medline: 7630578]
53. Thayer JF, Åhs F, Fredrikson M, Sollers JJ, Wager TD. A meta-analysis of heart rate variability and neuroimaging studies: implications for heart rate variability as a marker of stress and health. Neurosci Biobehav Rev 2012 Mar;36(2):747-756. [doi: 10.1016/j.neubiorev.2011.11.009] [Medline: 22178086]

XSL•FO
RenderX

54. Park G, Vasey MW, Van Bavel JJ, Thayer JF. When tonic cardiac vagal tone predicts changes in phasic vagal tone: the role of fear and perceptual load. Psychophysiology 2014 May 27;51(5):419-426. [doi: 10.1111/psyp.12186] [Medline: 24571084]
55. Porges SW, Doussard-Roosevelt JA, Portales AL, Greenspan SI. Infant regulation of the vagal "brake" predicts child behavior problems: a psychobiological model of social behavior. Dev Psychobiol 1996 Dec;29(8):697-712. [doi: 10.1002/(SICI)1098-2302(199612)29:8<697::AID-DEV5>3.0.CO;2-O] [Medline: 8958482]
56. Porges SW. A phylogenetic journey through the vague and ambiguous Xth cranial nerve: a commentary on contemporary heart rate variability research. Biol Psychol 2007 Mar;74(2):301-307 [FREE Full text] [doi: 10.1016/j.biopsycho.2006.08.007] [Medline: 17055142]
57. Crameri L, Hettiarachchi IT, Hanoun S. Effects of dynamic resilience on the reactivity of vagally mediated heart rate variability. Front Psychol 2020 Jan 20;11:579210 [FREE Full text] [doi: 10.3389/fpsyg.2020.579210] [Medline: 33551903]
58. Beauchaine TP, Gatzke-Kopp L, Mead HK. Polyvagal Theory and developmental psychopathology: emotion dysregulation and conduct problems from preschool to adolescence. Biol Psychol 2007 Mar;74(2):174-184 [FREE Full text] [doi: 10.1016/j.biopsycho.2005.08.008] [Medline: 17045726]
59. Fisher JP. Autonomic control of the heart during exercise in humans: role of skeletal muscle afferents. Exp Physiol 2014 Mar 02;99(2):300-305 [FREE Full text] [doi: 10.1113/expphysiol.2013.074377] [Medline: 23995102]
60. Fu Q, Levine BD. Exercise and the autonomic nervous system. Handb Clin Neurol 2013;117:147-160. [doi: 10.1016/B978-0-444-53491-0.00013-4] [Medline: 24095123]
61. Kimhy D, Crowley O, McKinley PS, Burg M, Lachman M, Tun P, et al. The association of cardiac vagal control and executive functioning--findings from the MIDUS study. J Psychiatr Res 2013 May;47(5):628-635 [FREE Full text] [doi: 10.1016/j.jpsychires.2013.01.018] [Medline: 23434176]
62. Proulx C, Cabral A, Choudhury N, Debergue P. Acceptability study of a novel immersive cognitive care platform for remediation of cognitive deficits. Annu Rev Cyber Therapy Telemed 2018;16:43-49.
63. Hernandez Hernandez ME, Michalak E, Choudhury N, Hewko M, Torres I, Menon M, et al. Co-design of a virtual reality cognitive remediation program for depression (bWell-D) with patient end users and clinicians: qualitative interview study among patients and clinicians. JMIR Serious Games 2023 May 07;11:e43904 [FREE Full text] [doi: 10.2196/43904] [Medline: 37027183]
64. Cabral A, Choudhury N, Proulx C, Harmouche R, Kohlenberg E, Deberguevaluation OC. Evaluation of cybersickness in a passive walking virtual reality cognitive exercise. Annu Rev CyberTherapy Telemed 2019;17:59-64 [FREE Full text]
65. Quintana DS, Heathers JA. Considerations in the assessment of heart rate variability in biobehavioral research. Front Psychol 2014 Jul 22;5:805 [FREE Full text] [doi: 10.3389/fpsyg.2014.00805] [Medline: 25101047]
66. Gnacek M, Broulidakis J, Mavridou I, Fatoorechi M, Seiss E, Kostoulas T, et al. emteqPRO—fully integrated biometric sensing array for non-invasive biomedical research in virtual reality. Front Virtual Real 2022 Mar 11;3:35. [doi: 10.3389/frvir.2022.781218]
67. Keebler Assoc JR, Shelstad WJ, Google DC, Chaparro BS, Phan Google MH. Validation of the GUESS-18. J Usability Stud 2020;16(1):49-62. [doi: 10.5555/3532747.3532752]
68. Brooke J. SUS: a 'quick and dirty' usability scale. In: Jordan PW, Thomas B, McClelland IL, Weerdmeester B, editors. Usability Evaluation In Industry. London, UK: CRC Press; 1996.
69. Kennedy RS, Lane NE, Berbaum KS, Lilienthal MG. Simulator sickness questionnaire: an enhanced method for quantifying simulator sickness. Int J Aerosp Psychol 1993 Jul;3(3):203-220. [doi: 10.1207/s15327108ijap0303_3]
70. Bouchard S, Robillard G, Renaud P. Revising the factor structure of the simulator sickness questionnaire. Annu Rev CyberTherapy Telemed 2007;5:128-137.
71. Bouchard S, St-Jacques J, Renaud P, Wiederhold B. Side effects of immersions in virtual reality for people suffering from anxiety disorders. J Cyber Ther Rehabil 2009;2(2):127-137 [FREE Full text]
72. Carman WA. Validation of the GUESS-18 for the usability of a virtual reality racing game. LSU Master's Theses. URL: https://digitalcommons.lsu.edu/gradschool_theses/5606 [accessed 2023-06-14]
73. Kleiger RE, Stein PK, Bigger JT. Heart rate variability: measurement and clinical utility. Ann Noninvasive Electrocardiol 2005 Jan;10(1):88-101 [FREE Full text] [doi: 10.1111/j.1542-474X.2005.10101.x] [Medline: 15649244]
74. van Gent P, Farah H, van Nes N, van Arem B. HeartPy: a novel heart rate algorithm for the analysis of noisy signals. Transp Res Part F Traffic Psychol Behav 2019 Oct;66:368-378. [doi: 10.1016/j.trf.2019.09.015]
75. Watland S, Nes LS, Hanson E, Ekstedt M, Stenberg U, Børøsund E. The caregiver pathway, a model for the systematic and individualized follow-up of family caregivers at intensive care units: development study. JMIR Form Res 2023 May 25;7:e46299 [FREE Full text] [doi: 10.2196/46299] [Medline: 37097744]
76. Huckvale K, Venkatesh S, Christensen H. Toward clinical digital phenotyping: a timely opportunity to consider purpose, quality, and safety. NPJ Digit Med 2019 Sep 6;2(1):88 [FREE Full text] [doi: 10.1038/s41746-019-0166-1] [Medline: 31508498]
77. McIntosh AR, Lobaugh NJ. Partial least squares analysis of neuroimaging data: applications and advances. Neuroimage 2004;23 Suppl 1:S250-S263. [doi: 10.1016/j.neuroimage.2004.07.020] [Medline: 15501095]

XSL•FO
RenderX

78. Khundrakpam B, Booij L, Jeon S, Karama S, Tohka J, Evans AC. Individualized prediction of future cognition based on developmental changes in cortical anatomy. NeuroImage Rep 2022 Dec;2(4):100127. [doi: 10.1016/j.ynirp.2022.100127]

Abbreviations

ANS: autonomic nervous system
CPT: continuous performance test
ECG: electrocardiogram
EMG: electromyography
GUESS: game user experience satisfaction scale
HRV: heart rate variability
ML: machine learning
NASA-TLX: NASA-Task Load Index
NVIM: Neurovisceral Integration Model
PPG: photoplethysmography
RMSSD: root mean square of successive differences
SSQ: Simulator Sickness Questionnaire
UI: user interface
VR: virtual reality
VTT: Vagal Tank Theory

Edited by A Mavragani; submitted 18.07.24; peer-reviewed by N Khan; comments to author 16.10.24; revised version received 06.11.24; accepted 27.11.24; published 04.03.25.

Please cite as:
Khundrakpam B, Segado M, Pazdera J, Gagnon Shaigetz V, Granek JA, Choudhury N
An Integrated Platform Combining Immersive Virtual Reality and Physiological Sensors for Systematic and Individualized Assessment of Stress Response (bWell): Design and Implementation Study
JMIR Form Res 2025;9:e64492
URL: https://formative.jmir.org/2025/1/e64492
doi:10.2196/64492
PMID:40053709

©Budhachandra Khundrakpam, Melanie Segado, Jesse Pazdera, Vincent Gagnon Shaigetz, Joshua A Granek, Nusrat Choudhury. Originally published in JMIR Formative Research (https://formative.jmir.org), 04.03.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Long-Term Experiences of Health Care Providers Using Iris Scanning as an Identification Tool in a Vaccine Trial in the Democratic Republic of the Congo: Qualitative Study

Trésor Zola Matuvanga[1,2,3], PhD; Antea Paviotti[2,3], PhD; Freddy Bikioli Bolombo[1,2,3], MD; Gwen Lemey[2,3], MA; Ynke Larivière[2,3], PhD; Maha Salloum[2,3], MSc; Bernard Isekah Osang'ir[2,3], MSc; Emmanuel Esanga Longomo[4], MD; Solange Milolo[1], MD; Junior Matangila[1], PhD; Vivi Maketa[1], PhD; Patrick Mitashi[1], PhD; Pierre Van Damme[2], PhD; Hypolite Muhindo-Mavoko[1], PhD; Jean-Pierre Van geertruyden[3], PhD

[1]Department of Tropical Medicine, Faculty of Medicine, University of Kinshasa, Kinshasa, the Democratic Republic of the Congo
[2]Centre for the Evaluation of Vaccination, Vaccine and Infectious Disease Institute, University of Antwerp, Wilrijk, Belgium
[3]Global Health Institute, Department of Family Medicine and Population Health, University of Antwerp, Wilrijk, Belgium
[4]Division Provinciale de la Santé, Ministry of Health, Tshuapa Province, the Democratic Republic of the Congo

Corresponding Author:
Trésor Zola Matuvanga, PhD
Department of Tropical Medicine
Faculty of Medicine
University of Kinshasa
Avenue Université, N° 1, Commune de Lemba
Kinshasa, BP 834 KINSHASA XI
the Democratic Republic of the Congo
Phone: 243 810046306
Email: zola.matuvanga@unikin.ac.cd

Abstract

Background: Iris scanning has increasingly been used for biometric identification over the past decade, with continuous advancements and expanding applications. To better understand the acceptability of this technology, we report the long-term experiences of health care providers and frontline worker participants with iris scanning as an identification tool in the EBL2007 Ebola vaccine trial conducted in the Democratic Republic of the Congo.

Objective: This study aims to document the long-term experiences of using iris scanning for identity verification throughout the vaccine trial.

Methods: Two years after the start of the EBL2007 vaccine trial (February to March 2022), 69 trial participants—including nurses, first aid workers, midwives, and community health workers—were interviewed through focus group discussions. Additionally, 13 in-depth individual interviews were conducted with physicians involved in the trial, iris scan operators, trial staff physicians, and trial participants who declined iris scanning. Qualitative content analysis was used to identify key themes.

Results: Initially, interviewees widely accepted the iris scan and viewed it as a distinctive tool for identifying participants in the EBL2007 vaccine trial. However, over time, perceptions became less favorable. Some participants expressed concerns that their vision had diminished shortly after using the tool and continued to decline until the end of the study. Others reported experiencing perceived vision loss long after the trial had concluded. However, no vision impairment was reported as an adverse event or assessed in the trial as being linked to the iris scan, which uses a previously certified safe infrared light for scanning.

Conclusions: Our findings highlight the sustained acceptability and perceived high accuracy of the iris scan tool for uniquely identifying adult participants in a vaccine trial over time. Continued efforts to systematically disseminate and reinforce information about the function and safety of this technology are essential. Clearly presenting iris scanning as a safe procedure could help dispel misconceptions, concerns, and perceived risks among potential users in vaccine trials.

(JMIR Form Res 2025;9:e54921) doi:10.2196/54921

XSL•FO
RenderX

KEYWORDS

iris scan; vaccine trial; iris; perception; experience; views; biometric identification; Democratic Republic of the Congo

Introduction

In many low- and middle-income countries (LMICs), the lack of reliable patient identification can hinder the effectiveness of routine medical care [1-3]. In rural settings, the digitization of personal information and its integration into databases remain limited. Additionally, challenges in identifying patients enrolled over long periods (eg, in longitudinal cohort studies) can lead to misclassification and pose a significant barrier to maintaining study data integrity. A recurring issue during (un)scheduled study visits is that participants often lack a consistent form of identification—whether due to misplacement, forgetfulness, or deterioration. Recently, biometric recognition has gained importance as a tool for accurately identifying individuals in routine health information processes [4-6]. In clinical trials, the use of human biometric recognition has also expanded over the past decade, with adoption becoming increasingly widespread due to its advantages over traditional identification methods [4,7].

Biometric recognition involves a process that relies on the technical processing of data related to the physical, physiological, or behavioral characteristics of the human body (including movement) for authentication purposes [8]. Among the biometric recognition methods used to date—such as fingerprints, facial recognition, iris scans, ear biometrics, and voice recognition—fingerprints are the most widely utilized and have become the predominant form of biometric data [9]. However, unlike fingerprints, which lose clarity over time due to the widening of ridge gaps, the distinct patterns within the iris remain unchanged throughout a person's lifetime, making iris scan identification exceptionally stable [10-12]. Even among identical twins or between an individual's own eyes, iris patterns are unique, ensuring their suitability as a reliable proxy for identification [13]. Additionally, iris scanning offers the advantage of being a noncontact process, eliminating the risk of disease transmission through direct contact. As a result, it has the potential to become an affordable, fast, and reliable identification tool across various contexts, including electoral voting, access control, and vaccine trials [4,5,14,15]. In clinical trials, the use of iris scans could help prevent errors and fraudulent entries while ensuring unique participant identification across (un)scheduled visits, thereby safeguarding the integrity of trial results [5]. However, experience with iris scanning as a method of identity verification in clinical trials remains limited.

Recently, the implementation of the iris scan tool demonstrated high precision and strong acceptance among health care providers and frontline worker participants (Multimedia Appendix 1), who were interviewed both before and at enrollment in an Ebola vaccine trial (hereafter referred to as the "EBL2007 vaccine trial") conducted in a remote area of the Democratic Republic of the Congo (DRC) [4]. This new qualitative research aimed to document long-term experiences with iris scanning for identity verification throughout the trial.

Methods

Study Design

This study used a qualitative methodology, specifically a phenomenological approach, to explore experiences related to the long-term use of iris scanning as an identification tool during (un)scheduled visits throughout the EBL2007 vaccine trial.

Research Team and Reflexivity

The research team comprised TZM and FBB, both male medical doctors and PhD candidates at the University of Kinshasa (DRC), under the coordination of AP, a female social scientist (PhD) from the University of Antwerp (Belgium). TZM and AP developed the interview and topic guides. TZM and FBB conducted the interviews and focus group discussions (FGDs), also facilitating translation between French and Lingala and taking notes as needed. EEL, a male MD from the Provincial Health Division in Boende (Ministry of Health), provided logistical support and assisted with recruitment. Previously, TZM and EEL also served as subinvestigators and trial site coordinators for the EBL2007 vaccine trial.

Study Setting and Time

This qualitative study on iris scan perception was nested within the EBL2007 vaccine trial (the main study) and conducted in the Health District of Boende, located 1200 km from Kinshasa, the capital of the DRC. Conversations took place on the premises of the General Referral Hospital of Boende (Hôpital Général de Référence de Boende, HGR), which served as the main trial site.

Participant recruitment for the main study began in December 2019, with the final patient visit occurring in October 2022 (Figure 1). During the screening process, potential vaccine trial participants were invited to opt into an innovative identification method—an iris scan that captures biometric data from the iris for personal identification. They were informed that this tool was safe and would enable more accurate recognition at both scheduled and unscheduled trial visits.

XSL•FO
RenderX

Figure 1. Chronology of events in the EBL2007 vaccine trial. HCPs: health care providers.

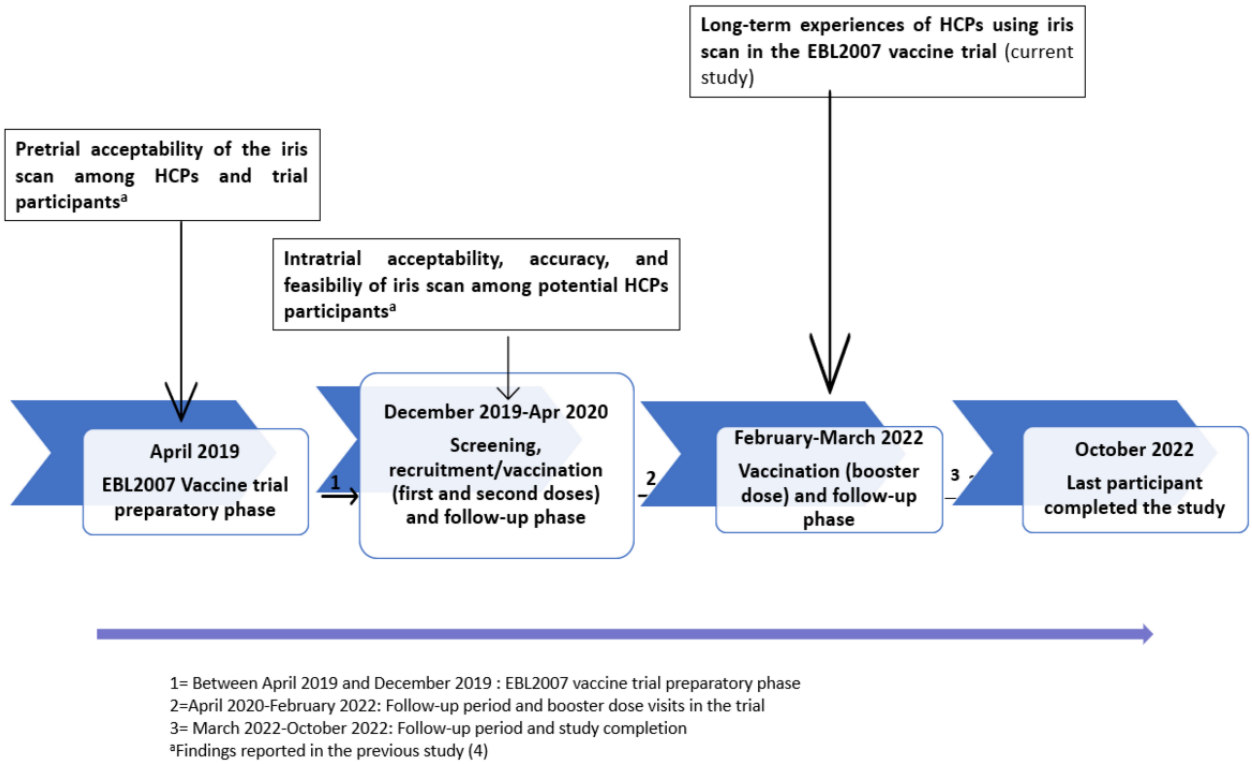

This qualitative study was conducted between February and March 2022, during the second follow-up period of the main study, when most participants had already completed their last scheduled visit.

Long-term experiences refer to users' perceptions of the operational viability of the iris scan after 2 years of systematic use within the EBL2007 vaccine trial. Further details on the vaccine trial are reported elsewhere [16].

Participants' Recruitment and Sampling

In this qualitative study, purposive sampling was used to include a diverse group of participants and staff from the EBL2007 vaccine trial as it neared completion. The study included the following groups: participants who used iris scanning for identification, participants who initially opted out of iris scanning (including various health care providers involved in the main study, such as nurses, doctors, midwives, first aid workers, community health workers, and health facility cleaners), physician staff involved in the trial, and iris scan operators. Potential study participants were purposively selected through referrals based on their professional category, their consent or refusal to use the iris scanning system, and their level of involvement in its implementation during the main study.

Data Collection Methods

A combination of FGDs and in-depth individual interviews (IDIs) was applied to collect data from these participants.

Trial participants who had used iris scanning as their identification method from the outset were invited to participate in FGDs. These participants represented a diverse range of professional categories, including community health care workers, who serve as intermediaries between health services and the local population by disseminating information, providing health education, identifying early disease cases, promoting positive health practices, and encouraging individuals to seek professional health care. First aid workers, trained to provide immediate care before professional medical services become available, were also included, along with midwives, nurses, and health facility cleaners. To facilitate open dialogue within each FGD, considerations were given to both the sex and professional categories of the participants.

Furthermore, IDIs were conducted to gather insights from trial participants who had initially opted against using iris scanning for identification. Physicians involved as trial participants and iris scanning operators were also invited to participate in these IDIs.

The selection of trial participants who declined iris scanning was based on documented refusals, as established in prior research [4]. The primary reason for refusing iris scanning among these participants was concern about ocular safety.

Procedure

For both IDIs and FGDs, a semistructured questionnaire was used (Multimedia Appendix 2). FGDs were conducted in groups of 6-10 participants. Before commencing the IDIs or FGDs, interviewees were reminded of the study's purpose. Each interview lasted between 60 and 90 minutes, and the conversations were recorded after obtaining the interviewees' consent.

Voice recordings in languages other than French were translated into French and then into English before being transcribed by 2 independent individuals. The accuracy and coherence of the transcriptions were thoroughly verified by AP and TZM. TZM coded all transcripts derived from both FGDs and IDIs, and AP

XSL•FO
RenderX

and TZM reached a consensus on the coding system and categories used for thematic analysis.

Analysis

This study follows the Consolidated Criteria for Reporting Qualitative Research (COREQ) guidelines to ensure rigor and transparency in the design, data collection, analysis, and reporting of findings (Multimedia Appendix 3). The audio recordings of all conversations were transcribed, translated into French if necessary, and imported into NVivo software (QSR International) for analysis. Each transcript was coded and assigned a unique identifier to ensure confidentiality. AP and TZM held regular collaborative meetings to ensure consistency in the coding process.

The data underwent a thematic analysis approach, beginning with an initial phase of familiarization with the data set. The coding process combined both inductive and deductive approaches [17,18]. Initially, a deductive approach was applied, with codes derived directly from the 2 main themes predetermined in the interview guide: (1) the acceptability of the iris scan and (2) knowledge, perception, and use of the iris scan. The initial codes addressing similar themes were grouped into these predetermined themes using a start list [19]. As the analysis progressed, an inductive approach was incorporated, allowing for the identification of new, unanticipated codes. The start list was subsequently updated to integrate these emergent codes. An iterative process was used to refine and develop subthemes and themes, ensuring alignment with the study objectives. The final themes were substantiated by significant quotations from the transcripts. The analysis was conducted in French to preserve the authenticity and nuance of participants' responses.

Trustworthiness

To ensure trustworthiness in our qualitative analysis, we implemented multiple strategies aligned with the criteria outlined by Creswell and Poth [20], as well as the principles proposed by Braun and Clarke [21]. Specifically, credibility was strengthened through source triangulation by integrating data from IDIs, FGDs, and observations across different professional and gender roles within the trial. This methodological approach facilitated the comparison of diverse perspectives, ensuring a more comprehensive understanding of participants' experiences.

Reflexivity was explicitly addressed in a dedicated section that examined the roles of the investigators throughout the research process. Additionally, verbatim quotes from participants were incorporated into the text to preserve the authenticity of their perspectives and enhance the depth of the analysis.

To enhance transferability, we explicitly detailed our working process and methodology, enabling other researchers to understand and potentially replicate the conditions and results of our study.

Confirmability was strengthened by using multiple data collection methods and engaging independent transcribers and translators to ensure neutrality. Each transcript was meticulously reviewed by the researchers to confirm the accuracy and minimize researcher bias.

Finally, dependability was ensured by maintaining a consistent methodological approach throughout the study, with exhaustive documentation to allow verification of the research process.

We believe that the aforementioned measures provide a solid framework to ensure the reproducibility of our findings.

Iris Scan Equipment and Procedures

The operator responsible for the iris scanning procedure was a trained and authorized EBL2007 vaccine trial staff member. The iris scan was conducted using an iris camera (Iritech; Irishield Monocular) in conjunction with a tablet (Samsung Tab Active 2) connected to a local ruggedized server via Wi-Fi. Further details regarding the procedures are described elsewhere (Multimedia Appendix 4) [4].

The iris scanner used in the EBL2007 vaccine trial was certified as safe for use with infrared light under all operating conditions, in accordance with the international standard IEC (International Electrotechnical Commission) 62471:2006-07 [22]. Additionally, its irradiance was less than 2% of the Eye Safety Standard Regulation and was tested for photobiological safety.

At enrollment, a list of trial participants who consented to the use of the iris scanning tool for identification during trial visits was compiled. Based on this list, the iris scan operator was instructed not to perform iris scans on participants who had not provided consent. For these individuals, an alternative identification method was used. Their unique participant identification number, assigned at trial enrollment, was recorded, and only their demographic data and identification photo were captured using the iris scanning tool's tablet. During subsequent (un)scheduled visits, entering their identification number into the tablet allowed for participant recognition based on the stored photograph. For consenting individuals, the operator collected demographic data, took an identification photo, and scanned both irises (left and right eye) as the primary method of identification.

Ethics Approval

This research was conducted in accordance with ethical principles designed to protect the rights and welfare of all participants. Approval to conduct the study was granted by the National Ethics Committee of Health of the DRC (approval reference number 368/CNES/BN/PMMF/2022). Explicit oral informed consent was obtained before any data collection, following a clearly defined protocol. Interviewees were fully informed about the study's objectives, the use of their personal data, potential risks and benefits, and their rights to confidentiality and voluntary participation. Consent was recorded, and discussions proceeded only after confirmation. This study had its own protocol and was not a secondary analysis of existing data; therefore, the informed consent process was specific to this research. To protect participants' privacy and confidentiality, all data were anonymized during transcription and analysis. Unique identifiers were used in place of personal information, and no identifying details were included in the reported findings. Interviewees did not receive monetary compensation for their participation. However, refreshments and transportation allowances were provided to facilitate their involvement in FGDs and IDIs. This was disclosed to

XSL•FO
RenderX

participants before they provided consent and was approved by the National Ethics Committee.

Results

Participant Recruitment and Data Saturation

Interviews were concluded once data saturation was achieved, with a total of 82 trial participants and staff included in the study. Data saturation was determined when no new themes emerged from subsequent discussions. A total of 69 trial participants took part in FGDs, representing community health care workers, first aid workers, midwives, nurses, and health facility cleaners (Table 1). Additionally, IDIs were conducted with 6 trial participants who refused iris scanning, 5 medical doctors responsible for safety monitoring in the trial, and 2 iris scanning operators.

Table 1. Data collection activities for the long-term experiences of health care providers using iris scanning in the EBL2007 vaccine trial.

Interviewees and data collection method	Number of activities, n	Interviewees occupation	Male (n=42), n	Female (n=40), n	Total (N=82), n
Trial participants					
Focus group discussion	2	Nurses	10	8	18
Focus group discussion	2	First aid workers	9	8	17
Focus group discussion	2	Community health workers	8	7	15
Focus group discussion	1	Midwife	0	9	9
Focus group discussion	1	Health facility cleaner	2	8	10
In-depth individual interview	1	Physician	3	0	3
In-depth individual interview	6	Entered the trial but refused iris scan (participants A, B, C, D, E, and F)[a]	6	0	6
Staff in the trial					
In-depth individual interview	2	Physician monitoring safety worker	2	0	2
In-depth individual interview	2	Iris scan operator	2	0	2

[a]Participants referred to as A, B, C, D, E, and F represent those who chose not to use the iris scanner for identification in the EBL2007 vaccine trial.

Participants had an average age of 51 (SD 11) years. At least one representative from each main professional category enrolled in the EBL2007 vaccine trial participated in either the IDIs or FGDs. The study included 40 (49%) female EBL2007 vaccine trial participants and staff (N=82; Table 1). Three key themes emerged from the collected data: (1) long-term experiences of using iris scanning as an identification tool in the EBL2007 vaccine trial, (2) the potential use of iris scanning in future vaccine trials, and (3) comparisons between iris scanning and previously known identification tools. Within each theme, additional subthemes and categories were identified.

Long-Term Experiences of Health Care Providers Using Iris Scan as an Identification Tool in the EBL2007 Vaccine Trial

Purpose-of-Use Understanding of Trial Participants Who Opted to Use Iris Scan

Some respondents demonstrated a clear understanding of the rationale behind using iris scanning, as reflected in the following statements:

> *It helped us because it brought out the whole face.* [FGD, midwife, trial participant, woman]
>
> *It was for identification purposes.* [FGD, midwife, trial participant, woman]

However, a few respondents had a different understanding of the iris scanning tool, perceiving it as a method for vaccine trial investigators to detect disease or assess the impact of the experimental vaccine on the eyes.

> *Me, I believed they're going to find the disease in the eyes, and they're going to tell us, but they haven't told us, and we haven't asked.* [FGD, midwife, trial participant, woman]

Acceptability of the Iris Scan

In general, health care providers and frontline workers volunteering in the EBL2007 vaccine trial widely accepted the use of iris scan technology, consistent with findings from the start of the trial [4]. One community health worker mentioned:

> *[...] I haven't come across anyone who would tell me that they didn't accept it. You see, I haven't encountered any group of people or any individual who would refuse to be examined by this device.* [FGD, community health worker, trial participant, man]

Reasons for Accepting the Iris Scan

The primary reason most interviewees accepted the iris scan was their willingness to receive the study vaccine. Their confidence in the trial staff further motivated them to trust the procedures proposed in the trial.

> *We accepted it because we were looking at our study vaccine, we were looking at the advantage of being enrolled in the Ebola vaccine trial.* [FGD, nurse, trial participant, man]

XSL•FO
RenderX

However, some respondents stated that they adopted the iris scan because they understood it as the appropriate identification tool for accurately recognizing participants enrolled in the trial.

> *We have observed that it is a valuable tool for identification, which is why it was accepted.* [FGD, first aid worker, trial participant, woman]

Reasons for Not Accepting the Iris Scan

A handful of interviewees indicated that they did not feel comfortable with the tool due to concerns about the retention of their demographic data recorded by it.

> *Yes, for a psychological reason, for example, we might take this photo and put it in a book, in a documentary, so that people can see you. So that's why I refused. But as I didn't have any information, I couldn't accept.* [FGD, community health worker, trial participant, man]

Some interviewees, due to the condition of their eyes before the use of the iris scan, expressed a fear of vision loss associated with the tool, compounded by their apprehension upon seeing it for the first time.

> *I refused because I'm sick. My eye hurts, especially my left eye, which has been bothering me since 2012. I've been suffering for four years. It's in this sense that I refused because the way it was flashing, it may burn my eye. Especially when she filmed it, it scared me. That's why I refused.* [Interview, trial participant B who refused iris scan]

Purpose-of-Use Understanding of Trial Participants Who Declined the Use of Iris Scan

Some interviewees who declined iris scan identification reported that they were unaware of the rationale for using this tool in the vaccine trial.

> *Well, I don't know the importance of this.* [Interview, trial participant A who refused iris scan]

> *I still have doubts, but I've seen the people who have had iris scans, and I don't see the point in continuing to doubt, given that there are friends who have accepted, and they're still here.* [Interview, trial participant E who refused iris scan]

> *In my opinion, regarding the importance of iris scan, when they captured us, I explained that my eye is diseased, but they told me they would capture the other eye, and I refused. But I don't know the significance of it.* [Interview, trial participant F who refused iris scan]

Some interviewees reported understanding that the scanning tool was intended for identification purposes during visits to prevent possible fraud.

> *[...] We mentioned that we would use it to confirm the person's identity. If someone else tries to pretend to be me, when they put the iris scan in their eye, it will show that it is not them. That is what I remember about the iris scan.* [Interview, trial participant E who refused iris scan]

The Perceived Accuracy of Iris Scan in Identifying Participants

Most interviewees reported that they found the iris scan tool in the trial to be highly accurate, as it helped prevent cases of fraud among participants. They provided examples of fraud attempts to illustrate the tool's effectiveness and accuracy. One such example is the following statement:

> *[...] There was an incident where I arrived for a scheduled visit and the person sitting close to me presented his father's identification card, claiming to be him. Upon verification, the operator discovered that the person in front of him was not who he claimed to be. The photo in the system of iris scan did not match his appearance. He eventually admitted that his father had passed away. The individual conducting the check informed him that we had not been informed of the father's passing, and upon further investigation, it was revealed that the father was traveling and had authorized his son to represent him. This incident revealed potentially fraudulent activity, as the person attempted to manipulate the results of the vaccine trial by assuming someone else's identity.* [FGD, community health care worker, trial participant, woman]

Perceived Risk of Iris Scan Identification

Some participants perceived a link between the use of the iris scan and vision loss. Additionally, most participants expressed concerns that vision loss might occur later due to the amount of light emitted by the device.

Certain interviewees compared the light emitted by the iris scan to the eclipse phenomenon or sunlight, suggesting that this may have contributed to vision loss.

> *We were scared, there was the light and there was uncertainty as to whether the eye would crack or not [...]* [FGD, community health care worker, trial participant, man]

> *Thank you, now after the iris scan, I have noticed that there is a reduction of vision, especially for reading, so we have to use glasses now.* [FGD, first aid worker, trial participant, man]

Certain interviewees mentioned that they did not observe any unusual occurrences during the scanning process and expressed no apprehension about the future safety of their eyes, as they did not perceive the iris scan as hazardous.

> *There was no reaction. They just tell you to stare like this and then they tell you, it's okay. There wasn't really any direct reaction like that.* [FGD, nurse, trial participant, woman]

> *I was just afraid for my eyes, but it is not dangerous. Even if it comes back to my village, we will make people aware of this device. It is for identifying people.* [Interview, trial participant E who refused iris scan]

XSL•FO
RenderX

Rumors and Reactions From the Surroundings

Certain interviewees and FDG expressed concerns about the well-being of their eyes based on conversations. Nevertheless, most emphasized that they had not heard any rumors related to the iris scan. Instead, the rumors primarily focused on the experimental vaccine and other study procedures, such as blood sampling.

> *They said your eye is sick if you use this device, your eye will be completely damaged, which is why I was afraid. Otherwise, there wasn't much to it.* [Interview, trial participant B who refused iris scan]

> *No, in the neighbourhoods there hadn't been any rumours, but it was about the vaccine and blood sampling that people were talking nonsense about, not about the iris scan.* [FGD, first aid worker, trial participant, man]

> *The people around us didn't know that we were having the iris scan in the study, they only knew that we were selling our blood and getting vaccinated, period, but concerning the Iris scan, nothing was said, it was only us, trial participants, who knew about the Iris scan, but not the community, they didn't know anything about it. In the neighbourhood, we were nicknamed blood sellers.* [FGD, health facility cleaner, trial participant, woman]

Use of Iris Scans in Future Vaccine Trials or Other Public Health Activities

Acceptability of Iris Scan in the Wider Community

As the iris scan tool is part of a vaccination monitoring system that collects additional data (eg, demographics, a photo of the face) beyond those recorded for routine vaccinations (eg, demographics, previous vaccinations, the name of the vaccine administered, its lot number, and expiry date), certain participants believed that the iris scan tool would neither be accepted nor feasible in broader vaccination efforts involving the general population. Participants referenced a yellow fever vaccination campaign that had taken place in Boende. During this campaign, people agreed to receive the vaccine, but many were unwilling to provide full identities and demographic data. Therefore, some interviewees and FGD participants suggested that a different tool, such as fingerprint scanning, would be preferable instead.

> *I wanted to say that for the population, it's going to be a bit difficult, because we've noticed here that with yellow fever, we only recorded the name on the card and then gave the vaccine directly. It was also difficult to get someone to agree to give their full identity so that they could be vaccinated, so it would also be very difficult with the iris scan. It's better even with the fingerprint, maybe it will be all right. With the iris scan, it will be a bit difficult with this population.* [FGD, nurse, trial participant, woman]

Some interviewees emphasized that iris scanning could be an acceptable, effective, and reliable tool for uniquely identifying individuals who might volunteer in future clinical trials. However, they stressed that this should be accompanied by a robust awareness campaign to repeatedly communicate sufficient information about the tool's safety.

> *At first, people will refuse, but after awareness-raising and testimonials from those who have experienced the tool, they will accept.* [Interview, trial participant A who refused iris scan]

Recommendations From Interviewees and FGD Participants

When implementing iris scanning in vaccine trials, particularly in remote areas such as Boende, some interviewees and FGD participants recommended considering the availability of ophthalmic specialists. According to their statements, the iris scan operator in the EBL2007 vaccine trial was not fully aware of the risks involved in scanning the eyes and was unable to provide clear explanations regarding the safety of trial participants' eyes.

> *[...] when you come to scan people's eyes, come with the eye specialist. [...]. All those who have handled our eyes are not specialists. [...] They are photographers, so you should to come with eye specialists. An ophthalmologist because it's a sensitive organ.* [FGD, first aid worker trial participant, trial participant, man]

Few of those who participated in the FGDs voiced concerns about not seeing the vaccine trial investigators undergo iris scanning.

> *[...] Until now we haven't seen the staff being vaccinated, or scanned the eyes with iris scan. We haven't seen; they haven't scanned themselves.* [FGD, first aid worker, trial participant, man]

Some participants in this research recommended reducing the amount of light used during scanning, increasing the distance between the eye and the scanner, and conducting demonstrations during the screening/consent process.

> *I think that, as my colleague the Community health worker just said, the distance from the iris scan is too close. Isn't there some way of finding ways of making it even bigger? [...].* [FGD, community health worker, trial participant, woman]

Comparison of Iris Scan With Previous Known Identification Tools

Referring to previously used and well-known identification methods, most interviewees stated that iris scanning would be the best way to uniquely identify volunteers in the trial.

> *With the experience that I have, with the age that I have...I believe that the only method of escaping fraud is scanning [...]* [FGD, community health worker, trial participant, woman]

> *[...] So, with today's technology, we may easily modify the photo by taking someone's face and putting it on another body to make it look like it's me, but it's not. But with the iris scan, it's easy to see that it's not me, it's just someone else's face. So, with the iris scan,*

XSL•FO
RenderX

it's hard to commit fraud. [FGD, community health worker, trial participant, woman]

Some interviewees suggested using the traditional fingerprint biometric tool to minimize the risk of compromising their eyes with iris scanning or deploying other methods, such as recording names, dates of birth, and identification numbers in a computer. Others proposed more innovative identification tools, such as collecting identity data through the laser thermometer used to measure temperature at entry points or utilizing blood samples already collected during the first visit.

I'm going to recommend the fingerprint because the signature may be imitated, but your fingerprint, your own blood, will reveal all your data. [FGD, community health worker, trial participant, man]

[...] instead of using this device, there was no way of using our fingerprints. Because...Isn't it possible to use a fingerprint? [FGD, community health worker, trial participant, man]

[...] As we've received all the doses of vaccine as well as blood samples were taken, it's the computer that will indicate that for such and such a participant, he's finished his doses, these appointments are over, so that's it. The computer is a method. [FGD, health facility cleaner, trial participant, woman]

[...] We may have a thermometer that records all the identity as well as the blood pressure and everything, so as not to have any problems with the use of that iris scan laser. [FGD, community health worker, trial participant, man]

Discussion

Principal Findings

This qualitative study aimed to document the long-term experiences of EBL2007 vaccine trial participants and staff regarding the use of an innovative iris scan biometric tool. Overall, the tool was found to be acceptable, accurate, and effective in verifying participants' identities throughout the trial, preventing fraud and errors.

Although it was clearly explained during the consent procedure that the iris scan was noncompulsory, some participants may have feared that declining it would prevent their enrollment. Similarly, the safety of the iris scan was thoroughly explained to trial personnel at the start of the trial. Despite these efforts, some interviewees and FGD participants still felt that scanning their eyes posed safety risks or that it might cause problems in the future.

It is important to highlight the expected motivational benefits of participating in the EBL2007 vaccine trial, as some interviewees may have accepted the iris scan primarily to receive the Ebola vaccine regimen, given that they reside in an area at risk of an outbreak [23], or to obtain reimbursement for travel costs and time associated with the trial [24]. Additionally, concerns about eye safety after using this tool should not be overlooked. Various vision problems were perceived to be associated with the iris scan.

In studies conducted elsewhere, similar reasons for hesitancy—such as general safety concerns and anxiety about the physical effects of biometric scanning—have been reported [14,25,26]. However, it is important to note that vision impairment was neither reported as an adverse event nor assessed as being associated with iris scanning during the trial. Additionally, concerns were raised regarding the collection and safeguarding of additional personal data following the iris scan. Notably, collecting such information was new to trial participants in the remote area of Boende (the trial site), where the most common biometric practice—used for payroll verification, passport applications, and voter registration—remains fingerprinting [27]. This familiarity with fingerprinting may have contributed to greater comfort with it compared with iris scanning and may have also raised security concerns about iris scanning, despite its general acceptance and perceived accuracy.

It is likely that the vision disorders perceived to be associated with iris scanning in the EBL2007 vaccine trial had other causes. Some participants may have had preexisting eye conditions. For instance, while the mean age of trial participants was 45 years [27], it is well known that the incidence of vision impairment increases from middle age onward [28,29]. Additionally, at the time of the EBL2007 vaccine trial, no ophthalmological care was available in Boende. As a result, participants' vision or ophthalmological complaints may not have been addressed at the time of enrollment. Promotional and preventive activities aimed at improving eye health may also be necessary, as studies have shown that visual impairment is prevalent in populations living in remote, resource-constrained areas due to limited access to quality health care services [30]. However, implementing such activities may introduce additional costs for researchers. Given the increasing spread of digital legislation and democracy in Africa [31], these concerns may gradually diminish with the wider use of iris scanning technology. This suggests that even if research participants develop vision problems—likely due to aging or other factors—they may not attribute them to iris scanning as digital technology becomes more commonplace.

Some interviewees noted that the iris scanner operator appeared more like a photographer than someone capable of properly explaining the tool and its safety information. This suggests that the purpose and function of the iris scan tool were not sufficiently explained at inclusion or during follow-up visits in the trial. As a lesson learned, it is crucial to provide more detailed training to the iris scan operator to ensure they can answer specific questions from research participants.

It is also worth emphasizing that the timing of this qualitative study—conducted 2 years after the trial began—may explain why some interviewees and FGD participants gradually forgot the information about the iris scan tool provided during the initial consent process. Previous studies have highlighted that trial volunteers often forget or retain less of the information given at the time of enrollment [32,33]. This suggests that in long-term studies, the contents of the informed consent form should be periodically re-explained to participants.

XSL•FO
RenderX

This study demonstrated that iris scanning in vaccine trials conducted in resource-poor settings has valuable potential and is generally accepted for participant identification. The acceptability of iris scanning aligns with the key properties required of a high-performance biometric tool, such as universality, uniqueness, permanence, collectability, and resistance to circumvention [31,34]. However, during the EBL2007 vaccine trial, the iris scanning process became cumbersome when identification was not possible due to a participant's failure to follow the operator's instructions. To facilitate rapid identification through iris scanning, it is essential that individuals remain attentive and adhere strictly to the operator's guidance. This challenge may be particularly pronounced in vaccine trials involving younger infants (under 1 year old). Although infants are among the populations most in need of vaccines, they are unable to follow detailed instructions—such as looking into a camera—required for iris recognition [15]. Alternative approaches, such as scanning the iris of the accompanying adult (proxy identification number) or using ear-based or palm-based automatic recognition, may be more suitable for infants and other dependent populations [35,36].

Limitations

This study has some limitations. First, our findings are based on long-term experiences—spanning multiple iris scanning moments over a 2-year study period—of a population of health care providers and frontline workers who likely have a higher level of understanding of health-related phenomena. As a result, these findings may not be generalizable to the broader population. Second, some of the researchers who conducted FGDs and interviews, although not directly involved in the medical aspects of the vaccine trial, may have been perceived by interviewees as representing the trial team. This perception could have introduced desirability bias in the IDIs and FGDs. Finally, some participants working at the General Referral Boende Hospital—also the study site of the trial—may have hesitated to express negative concerns about the iris scan due to the location of the interviews (ie, at the hospital) or because some interviews were conducted by a trial investigator. Nevertheless, the findings of this qualitative research reflect participants' and staff's retrospective experiences with iris scanning over time and complement previous qualitative research on its acceptability. To our knowledge, this study, combined with the earlier research on the acceptability of biometric identity verification tools, provides the only comprehensive analysis of both the initial acceptability and the actual experiences of iris scanning within the same trial population [4]. The insights gained can inform the broader implementation of iris scanning in vaccine trials or other long-term longitudinal research.

Conclusions

The findings of this qualitative research highlight the sustained acceptability and perceived high accuracy of the iris scan tool for uniquely identifying adult participants in a vaccine trial over time. However, when the functionality of the iris scan is not well understood or remembered by users, certain concerns may arise, including perceived risks to long-term vision, the use of retained data, and the tool's ability to rapidly verify information regardless of age, education level, or health condition. To support broader implementation in vaccine trials or other research, further efforts should be made to provide clear information to users and dispel misconceptions about the fears and perceived risks associated with the iris scan tool.

Acknowledgments

This work was supported by the EBOVAC3. This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement number 800176 (IMI-EU). This Joint Undertaking receives support from the European Union's Horizon 2020 Research and Innovation Programme, the European Federation of Pharmaceutical Industries and Associations (EFPIA), and the Coalition for Epidemic Preparedness Innovations (CEPI). All vaccines were provided by Janssen Vaccines & Prevention B.V. The funders had no role in the data collection, analysis, and interpretation; writing of the paper; or the decision to submit for publication. We extend our sincere gratitude to the health care providers and frontline workers of the Boende Health District for their valuable participation in this study. The authors deeply appreciate the hard work and dedication of the local trial staff. Additionally, we acknowledge with gratitude the supportive role of all partners within the EBOVAC3 Consortium.

Data Availability

The data sets generated during this study are available and included as multimedia appendices (Multimedia Appendices 1-5).

Disclosure of Artificial Intelligence Use

This manuscript was prepared without the assistance of generative artificial intelligence tools or algorithms for data analysis, research design, or writing substantial content.

Authors' Contributions

TZM was responsible for the conceptualization of the study protocol, data curation, formal analysis, investigation, methodology, software, and writing of the original draft. AP contributed to the conceptualization of the study protocol, data curation, formal analysis, investigation, methodology, software, supervision, validation, visualization, and writing—review and editing. FBB participated in the investigation and writing—review and editing. GL, YL, EEL, SM, and PM were involved in project administration

XSL•FO
RenderX

and writing—review and editing. MS, BIO, JM, and VM contributed to writing—review and editing. PVD and JPVG provided funding acquisition, supervision, validation, and writing—review and editing. HMM contributed to supervision, validation, and writing—review and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The profile of health care providers included in our previous study [4], which evaluated the acceptability of iris scanning in the EBL2007 trial before and during the recruitment process.
[DOCX File , 15 KB - formative_v9i1e54921_app1.docx]

Multimedia Appendix 2
The semistructured questionnaire used to interview participants.
[DOCX File , 18 KB - formative_v9i1e54921_app2.docx]

Multimedia Appendix 3
Reporting follows the COREQ (Consolidated Criteria for Reporting Qualitative Research) guidelines, which best match our study design to ensure completeness and detail.
[DOCX File , 24 KB - formative_v9i1e54921_app3.docx]

Multimedia Appendix 4
The process and the iris scanning tool used as an identification method in the EBL2007 trial, with the user experience—including clinical trial participants, staff, and operators—reported in our qualitative study.
[PPTX File , 641 KB - formative_v9i1e54921_app4.pptx]

Multimedia Appendix 5
It refers to the collection of transcripts (French) from all conversations (interviews and focus group discussions).
[DOCX File , 571 KB - formative_v9i1e54921_app5.docx]

References

1. Etter LP, Ragan EJ, Campion R, Martinez D, Gill CJ. Ear biometrics for patient identification in global health: a field study to test the effectiveness of an image stabilization device in improving identification accuracy. BMC Med Inform Decis Mak 2019 Jun 18;19(1):114 [FREE Full text] [doi: 10.1186/s12911-019-0833-9] [Medline: 31215427]
2. Gatwa T. "Rwanda and Burundi". In: Christianity in Sub-Saharan Africa. Edinburgh, Scotland: Edinburgh University Press; May 15, 2017:119-131.
3. ID4D data: global identification challenge by the numbers. World Bank. URL: https://id4d.worldbank.org/global-dataset/visualization [accessed 2022-05-10]
4. Zola Matuvanga T, Johnson G, Larivière Y, Esanga Longomo E, Matangila J, Maketa V, et al. Use of iris scanning for biometric recognition of healthy adults participating in an Ebola vaccine trial in the Democratic Republic of the Congo: mixed methods study. J Med Internet Res 2021 Aug 09;23(8):e28573 [FREE Full text] [doi: 10.2196/28573] [Medline: 34378545]
5. Singh T, Zaka-Ur-Rab S. Effect of pupil dilation on biometric iris recognition systems for personal authentication. Indian J Ophthalmol 2022 Dec 30;71(1):57-61 [FREE Full text] [doi: 10.4103/ijo.ijo_1417_22]
6. Dhir L, Habib NE, Monro DM, Rakshit S. Effect of cataract surgery and pupil dilation on iris pattern recognition for personal authentication. Eye (Lond) 2010 Jun 13;24(6):1006-1010. [doi: 10.1038/eye.2009.275] [Medline: 19911017]
7. SonLa Study Group. Using a fingerprint recognition system in a vaccine trial to avoid misclassification. Bull World Health Organ 2007 Jan 01;85(1):64-67 [FREE Full text] [doi: 10.2471/blt.06.031070] [Medline: 17242760]
8. Wendehorst C, Yannic D. Biometric recognition and behavioural detection: study commissioned by European parliament. European Parliament 2021 Aug 01:1-96 [FREE Full text]
9. Roberts CH, Stott C, Shawe-Taylor M, Chaudhry Z, Lal S, Marks M. Biometric linkage of longitudinally collected electronic case report forms and confirmation of subject identity: an open framework for ODK and related tools. Front Digit Health 2023 Aug 4;5:1072331 [FREE Full text] [doi: 10.3389/fdgth.2023.1072331] [Medline: 37600479]
10. Daugman JG. High confidence visual recognition of persons by a test of statistical independence. IEEE Trans Pattern Anal Machine Intell 2002;15(11):1148-1161 ISSN: 0162-8828. [doi: 10.1109/34.244676]
11. Jain AK, Arora SS, Cao K, Best-Rowden L, Bhatnagar A. Fingerprint recognition of young children. IEEE Trans Inform Forensic Secur 2017 Jul;12(7):1501-1514. [doi: 10.1109/tifs.2016.2639346]

XSL•FO
RenderX

12. Moolla Y, De Kock A, Mabuza-Hocquet G, Ntshangase CS, Nelufule N, Khanyile P. Biometric recognition of infants using fingerprint, iris, and ear biometrics. IEEE Access 2021;9:38269-38286. [doi: 10.1109/access.2021.3062282]
13. Hollingsworth K, Bowyer KW, Lagree S, Fenker SP, Flynn PJ. Genetically identical irises have texture similarity that is not detected by iris biometrics. Computer Vision and Image Understanding 2011 Nov;115(11):1493-1502. [doi: 10.1016/j.cviu.2011.06.010]
14. Okello E, Ayieko P, Kwena Z, Nanyonjo G, Bahemuka U, Price M, et al. Acceptability and applicability of biometric iris scanning for the identification and follow up of highly mobile research participants living in fishing communities along the shores of Lake Victoria in Kenya, Tanzania, and Uganda. Int J Med Inform 2023 Apr;172:105018 [FREE Full text] [doi: 10.1016/j.ijmedinf.2023.105018] [Medline: 36774907]
15. Masyn S, Vuchelen A, Santermans E, Rasschaert F, Bangura A, Parys W, et al. Overcoming the challenges of iris scanning to identify minors (1-4 years) in the real-world setting. BMC Res Notes 2019 Jul 22;12(1):448 [FREE Full text] [doi: 10.1186/s13104-019-4485-8] [Medline: 31331369]
16. Larivière Y, Zola T, Stoppie E, Maketa V, Matangila J, Mitashi P, et al. Open-label, randomised, clinical trial to evaluate the immunogenicity and safety of a prophylactic vaccination of healthcare providers by administration of a heterologous vaccine regimen against Ebola in the Democratic Republic of the Congo: the study protocol. BMJ Open 2021 Sep 28;11(9):e046835 [FREE Full text] [doi: 10.1136/bmjopen-2020-046835] [Medline: 34588237]
17. Braun V, Clarke V. Using thematic analysis in psychology. Qualitative Research in Psychology 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
18. Fereday J, Muir-Cochrane E. Demonstrating rigor using thematic analysis: a hybrid approach of inductive and deductive coding and theme development. International Journal of Qualitative Methods 2006 Mar 01;5(1):80-92. [doi: 10.1177/160940690600500107]
19. Azungah T. Qualitative research: deductive and inductive approaches to data analysis. QRJ 2018 Oct 31;18(4):383-400. [doi: 10.1108/qrj-d-18-00035]
20. Creswell JW, Poth CN. Qualitative Inquiry and Research Design: Choosing Among Five Approaches (4th Edition). Los Angeles, CA: Sage Publications; 2016.
21. Braun V, Clarke V. Toward good practice in thematic analysis: avoiding common problems and be(com)ing a researcher. Int J Transgend Health 2023 Oct 25;24(1):1-6 [FREE Full text] [doi: 10.1080/26895269.2022.2129597] [Medline: 36713144]
22. Reidenbach HD. Laser safety. In: Träger F, editor. Springer Handbook of Lasers and Optics. New York, NY: Springer; 2007:1251-1276.
23. Maganga GD, Kapetshi J, Berthet N, Kebela Ilunga B, Kabange F, Mbala Kingebeni P, et al. Ebola virus disease in the Democratic Republic of Congo. N Engl J Med 2014 Nov 27;371(22):2083-2091. [doi: 10.1056/nejmoa1411099]
24. Zola Matuvanga T, Larivière Y, Lemey G, De Bie J, Milolo S, Meta R, et al. Setting-up an Ebola vaccine trial in a remote area of the Democratic Republic of the Congo: challenges, mitigations, and lessons learned. Vaccine 2022 May 31;40(25):3470-3480 [FREE Full text] [doi: 10.1016/j.vaccine.2022.04.094] [Medline: 35550847]
25. Kalka N, Zuo J, Schmid N. Image quality assessment for iris biometric. In: Biometric Technology for Human Identification III. Bellingham, WA, USA: SPIE; Apr 01, 2006:124-136.
26. Daugman J, Downing C. Effect of severe image compression on iris recognition performance. IEEE Trans Inform Forensic Secur 2008 Mar;3(1):52-61. [doi: 10.1109/tifs.2007.916009]
27. Larivière Y, Garcia-Fogeda I, Zola Matuvanga T, Isekah Osang'ir O, Milolo S, Meta R, et al. Safety and immunogenicity of the heterologous 2-dose Ad26.ZEBOV, MVA-BN-Filo vaccine regimen in health care providers and frontliners of the Democratic Republic of the Congo. J Infect Dis 2024 Apr 12;229(4):1068-1076 [FREE Full text] [doi: 10.1093/infdis/jiad350] [Medline: 37673423]
28. Barber SJ. An examination of age-based stereotype threat about cognitive decline. Perspect Psychol Sci 2017 Jan 11;12(1):62-90 [FREE Full text] [doi: 10.1177/1745691616656345] [Medline: 28073332]
29. Dawes P, Dickinson C, Emsley R, Bishop PN, Cruickshanks KJ, Edmondson-Jones M, et al. Vision impairment and dual sensory problems in middle age. Ophthalmic Physiol Opt 2014 Jul 29;34(4):479-488 [FREE Full text] [doi: 10.1111/opo.12138] [Medline: 24888710]
30. Williamson S, Seewoodhary R, Dampies L. Effect of poverty on eye health and implications for nursing practice. Nurs Stand 2016 Aug 10;30(50):42-51. [doi: 10.7748/ns.2016.e10014] [Medline: 27507393]
31. Idowu HA. Biometric technologies and the prospect of sustainable democracy in Africa. JAE 2021 Jun 1;20(1):23-43. [doi: 10.20940/jae/2021/v20i1a2]
32. Bergler JH, Pennington AC, Metcalfe M, Freis ED. Informed consent: how much does the patient understand? Clin Pharmacol Ther 1980 Apr;27(4):435-440. [doi: 10.1038/clpt.1980.60] [Medline: 6987027]
33. Howard JM, DeMets D. How informed is informed consent? The BHAT experience. Control Clin Trials 1981 Dec;2(4):287-303. [doi: 10.1016/0197-2456(81)90019-2] [Medline: 6120794]
34. Bolle RM, Connell JH, Pankanti S, Ratha NK. Guide to Biometrics. New York, NY: Springer Science & Business Media; 2013.

XSL•FO
RenderX

35. Khatun F, Distler R, Rahman M, O'Donnell B, Gachuhi N, Alwani M, et al. Comparison of a palm-based biometric solution with a name-based identification system in rural Bangladesh. Glob Health Action 2022 Dec 31;15(1):2045769 [FREE Full text] [doi: 10.1080/16549716.2022.2045769] [Medline: 35343885]
36. Lemes R, Bellon O, Silva L. Biometric recognition of newborns: identification using palmprints. New York, NY: IEEE; 2011 Presented at: International Joint Conference on Biometrics (IJCB); October 11-13, 2011; Washington, DC p. 1-6. [doi: 10.1109/IJCB.2011.6117475]

Abbreviations

DRC: The Democratic Republic of the Congo
FGD: focus group discussion
IDIs: in-depth individual interview
IEC: International Electrotechnical Commission
LMIC: low- and middle-income country

Edited by A Mavragani; submitted 27.11.23; peer-reviewed by S Rajput, M Shabdiz, N Manshor, H Campos; comments to author 06.09.24; revised version received 02.10.24; accepted 05.12.24; published 06.03.25.

Please cite as:
Zola Matuvanga T, Paviotti A, Bikioli Bolombo F, Lemey G, Larivière Y, Salloum M, Isekah Osang'ir B, Esanga Longomo E, Milolo S, Matangila J, Maketa V, Mitashi P, Van Damme P, Muhindo-Mavoko H, Van geertruyden JP
Long-Term Experiences of Health Care Providers Using Iris Scanning as an Identification Tool in a Vaccine Trial in the Democratic Republic of the Congo: Qualitative Study
JMIR Form Res 2025;9:e54921
URL: https://formative.jmir.org/2025/1/e54921
doi:10.2196/54921
PMID:40053756

©Trésor Zola Matuvanga, Antea Paviotti, Freddy Bikioli Bolombo, Gwen Lemey, Ynke Larivière, Maha Salloum, Bernard Isekah Osang'ir, Emmanuel Esanga Longomo, Solange Milolo, Junior Matangila, Vivi Maketa, Patrick Mitashi, Pierre Van Damme, Hypolite Muhindo-Mavoko, Jean-Pierre Van geertruyden. Originally published in JMIR Formative Research (https://formative.jmir.org), 06.03.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Assessment of Fraud Deterrence and Detection Procedures Used in a Web-Based Survey Study With Adult Black Cisgender Women: Description of Lessons Learned and Recommendations

Amber I Sophus[1*], MPH, PhD; Jason W Mitchell[1*], MPH, PhD

Department of Health Promotion and Disease Prevention, Robert Stempel College of Public Health & Social Work, Florida International University, Miami, FL, United States

[*]all authors contributed equally

Corresponding Author:
Amber I Sophus, MPH, PhD
Department of Health Promotion and Disease Prevention
Robert Stempel College of Public Health & Social Work
Florida International University
11200 S.W. 8th Street
Miami, FL, 33199
United States
Email: asophus@fiu.edu

Abstract

Background: Online research studies enable engagement with more Black cisgender women in health-related research. However, fraudulent data collection responses in online studies raise important concerns about data integrity, particularly when incentives are involved.

Objective: The purpose of this study was to assess the strengths and limitations of fraud deterrence and detection procedures implemented in an incentivized, cross-sectional, online study about HIV prevention and sexual health with Black cisgender women living in Texas.

Methods: Data for this study came from a cross-sectional web-based survey that examined factors associated with potential pre-exposure prophylaxis use among a convenience sample of adult Black cisgender women from 3 metropolitan areas in Texas. Each eligibility screener and associated survey entry was evaluated using 4 fraud deterrence features and 7 fraud detection benchmarks with corresponding decision rules.

Results: A total of 5862 respondents provided consent and initiated the eligibility screener, of whom 2150 (36.68%) were ineligible for not meeting the inclusion criteria, and 131 (2.23%) completed less than 80% of the survey and were removed from further consideration. Other entries were removed for not passing level 1 fraud deterrent safeguards: duplicate entries with the same IP address (388/5862, 6.62%), same telephone number (69/5862, 1.18%), same email address (114/5862, 1.94%), and same telephone number and email address (17/5862, 0.29%). Of the remaining 2993 entries, 1652 entries were removed for not passing the first 2 items of the level 2 fraud detection benchmarks: screeners and surveys with latitude and longitude coordinates outside of the United States (347/2993, 11.59%) and survey completion time of less than 10 minutes (1305/2993, 43.6%). Of the remaining 1341 entries, 130 (9.69%) passed all 5 of the remaining level 2 data validation benchmarks, and 763 (56.89%) entries were removed due to passing less than 3. An additional 33.4% (423/1341) entries were removed after passing 4 of the 5 remaining validation benchmarks, being contacted to verify survey information, and not providing legitimate contact information or being unable to confirm personal information. The final enrolled sample in this online study consisted of 155 respondents who provided consent, were deemed eligible, and passed fraud deterrence features and fraud detection benchmarks. In this paper, we discuss the lessons learned and provide recommendations for leveraging available features in survey software programs to help deter bots and enhance fraud detection procedures beyond relying on survey software options.

Conclusions: Effectively identifying fraudulent responses in online surveys is an ongoing challenge. The data validation approach used in this study establishes a robust protocol for identifying genuine participants, thereby contributing to the removal of false data from study findings. By sharing experiences and implementing thorough fraud deterrence and detection protocols, researchers can maintain data validity and contribute to best practices in web-based research.

XSL•FO
RenderX

(JMIR Form Res 2025;9:e59955) doi:10.2196/59955

KEYWORDS

Black women; HIV; fraud deterrence; fraud detection; web-based research; online research; data integrity; data collection; survey

Introduction

Background

In the past 2 decades, there has been a rise in online health-related research studies. Online research provides a practical approach to recruitment and data collection, including electronic dexterity, participant anonymity, speed, reduced error, easy and remote participation [1], cost efficiency, and the potential to recruit larger study samples [2,3]. While online research has expanded recruitment opportunities, Black/African Americans remain underrepresented in health-related research studies [4-9], particularly Black women [7-9] who are considered a "hard-to-reach" population [9,10].

Online research studies offer one way to engage more Black cisgender women in health-related research. However, conducting research online, particularly when incentives are involved, warrants careful attention. For instance, respondents may misrepresent themselves to improve their likelihood of meeting the study eligibility criteria to obtain the incentive [11,12]. The lack of face-to-face interaction (eg, video) restricts researchers' ability to verify whether the data come from real individuals [13,14]. Other challenges associated with conducting online research include the presence of bots (ie, computer program software designed to automatically fill out survey questions with preprogrammed responses), the ability to bypass IP address restrictions, and respondent misrepresentation (ie, ineligible participants providing inaccurate information to gain entry into the study) [2,13,14]. Although online recruitment (ie, through social media) has been successful in attracting underrepresented populations in health-related research studies involving incentives [15-18], the use of a standard web link for online recruitment purposes may exacerbate the potential for fraudulent responses and threats to data integrity [19-22]. In response to these challenges, researchers have provided recommendations for how best to screen for fraudulent survey entries and handle potentially invalid responses [11,12,22-26].

Teitcher et al [12] recommends checking for inconsistent survey responses, using a CAPTCHA, collecting paradata to examine how individuals are responding to survey questions, tracking personal information (eg, email, home address, or telephone number), checking for encrypted IP addresses or multiple survey entries from the same IP address, enabling cookies to prevent multiple survey completion attempts, and including an interview to determine whether an individual has already participated or is being dishonest on responses. Other researchers have published similar recommendations [2,23,27,28]. In addition to these recommendations, researchers are encouraged to develop a system of decision rules to detect and handle invalid and fraudulent research data [11-13,25]. Pozzar et al [21] investigated threats to their sample validity and data integrity after online social media recruitment of health research participants to complete an online survey. Although the authors used a data collection platform with fraud prevention features, collected verifiable information, and included open-ended items to identify those who provided false information regarding their eligibility criteria, many respondents still bypassed the validity and data integrity measures [21]. Dewitt et al [29] applied similar participant validity procedures in a small internet health survey among a sample of gay and bisexual survivors of prostate cancer. However, the authors discovered that some invalid entries bypassed their validation protocol, although they passed checks for nonduplicate IP addresses, valid zip codes, and reCAPTCHA verification [29]. These findings underscore the importance of ongoing vigilance and periodic reassessment of validation strategies throughout the recruitment phase of a study to enhance the overall effectiveness of fraud prevention strategies. Furthermore, the tactics used to deceive or manipulate online study eligibility and enrollment systems are continually evolving and, as such, call for new insights and lessons learned to help improve rigor and data integrity.

Purpose

This analysis builds on extant literature by assessing the strengths and limitations of fraud deterrence and detection procedures that were implemented in an incentivized, cross-sectional, online study about HIV prevention and sexual health with a convenience sample of Black cisgender women living in Texas, United States. We share the lessons learned from implementing this study and discuss different strategies that are available to help improve the likelihood of collecting valid data from legitimate research participants.

Methods

This study adhered to the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) reporting guidelines [30].

Recruitment

Between December 2020 and January 2022, we recruited Black cisgender women from Houston, San Antonio, and Dallas in Texas to complete a 1-time online study survey about HIV prevention and women's sexual health, with a primary focus on pre-exposure prophylaxis [31]. Participants were recruited using online advertisements and print flyers that contained information about the study and a web link to access the online consent form and eligibility screener. Online advertisements were placed on Facebook and Instagram (clickable web link embedded), whereas printed flyers (with a QR code and web link) were placed in locations (eg, coffee shops, restaurants, grocery stores, and gyms) in Houston, Dallas, and San Antonio frequented by Black women. Snowball sampling was also used to recruit participants. All recruitment materials contained information about the study and a method (eg, QR code or direct web link) to access the online consent form and eligibility screener [32].

XSL•FO
RenderX

Procedures

Irrespective of the recruitment method, all interested individuals who clicked the provided web link were taken to the study landing page housed in Qualtrics (Qualtrics International Inc), a Health Insurance Portability and Accountability Act (HIPAA) compliant survey program. The landing page provided information about the study, the electronic consent form, and the online eligibility screener. The electronic consent form provided detailed information about the survey study, such as its purpose and participation process (including time commitment), incentives, privacy and confidentiality, risks and benefits, and contact details for the primary investigator and institutional review board. Respondents who consented to participate were immediately routed to the eligibility screener. Once consented and eligible, individuals were prompted to provide their email address for incentive distribution before being automatically granted access to complete the 30-minute online study survey. Individuals were also instructed to provide their telephone number if they consented to participate in a possible 1-time, online interview (if purposively selected). In addition, individuals could provide their contact information to be contacted for future research opportunities.

Consented and eligible individuals who completed at least 80% of the study survey were enrolled in the parent study once their data were evaluated and had passed the fraud deterrence and detection protocols (described later in the Methods section). Individuals who were ineligible were thanked for their interest in the study and provided with relevant resources related to HIV prevention.

Eligibility Criteria

For study inclusion, participants had to meet all of the following eligibility criteria through self-report: (1) be at least 18 years of age; (2) self-identify as a cisgender woman; (3) self-identify as Black (ie, African American or Caribbean American [eg, Haitian American]); (4) have an HIV-negative or unknown serostatus; (5) live in or near (within 25 miles [approximately 40 km]) Houston, San Antonio, or Dallas in Texas; (6) have engaged in at least 1 HIV risk behavior within the past 6 months (ie, unprotected vaginal or anal sex with a male partner, injection drug use, or sex exchange) or have been diagnosed with a sexually transmitted infection (STI); and (7) be fluent in English. Individuals who did not meet all 7 eligibility criteria were excluded.

Application of the First Level of Fraud Deterrent Protocols: Lesson 1

To help deter the collection of invalid data [11,23,27,33-37], 4 fraud prevention safeguards—built-in features available in Qualtrics—were selected for use in the online eligibility screener and study survey before recruitment began. The 4 fraud prevention safeguards are CAPTCHA verification, prevent indexing, prevent ballot box stuffing (ie, duplicate entries), and bot detection. Descriptions of the 4 fraud prevention safeguards are provided in Textbox 1.

Textbox 1. Descriptions of the level 1 fraud deterrent protocol, which included 4 fraud prevention safeguards used in the Qualtrics eligibility screener for all respondents.

CAPTCHA

CAPTCHA requires a respondent to successfully pass a task or challenge (eg, select all squares containing fire hydrants in the image) to gain access to the next web page (eg, survey)

Prevent indexing

A tool that prevents bots or software from finding the survey within web search engines

Prevent ballot box stuffing (currently called "Prevent Multiple Submissions")

A tool that places a cookie on the individual's browser (cookies are small data files generated by a web server and sent to a web browser; they track user activity, help websites recognize returning users, and improve the browsing experience); if the same respondent returns using the same browser on the same device, without having cleared their cookies, they are flagged as a duplicate

Bot detection

A tool that assesses the likelihood of a response being from a bot or a human by assigning a probability score based on interactions with invisible Google reCAPTCHA technology embedded in the survey

Soon after recruitment began, there was a huge influx of data entries—1498 within the first 3 days. This unusual activity prompted us to pause the study (ie, recruitment as well as screener and survey) to evaluate whether the screener and study survey entries were valid. During this review, the safeguard options presented in Textbox 1 were thoroughly examined to determine whether there were limitations to using these 4 features.

We learned that the prevent ballot box stuffing option in Qualtrics does not fully prevent duplicate entries. Respondents could still access the survey by switching browsers or devices, even when this option was enabled. In addition, Qualtrics does not prevent duplicate entries based on additional criteria such as name, telephone number, email address, or IP address. Of the 1498 entries, approximately half (n=769, 51.3%) were linked to an entry that shared an IP address with at least 1 other survey entry. Several other patterns were noted while evaluating the data entries: (1) some (53/1498, 3.54%) came from latitude and longitude coordinates outside of the United States and its territories; (2) a little more than one-fifth (350/1498, 23.36%) were completed in record time (eg, 9 min vs an estimated completion time of 30 min); (3) some (355/1498, 23.69%) contained names or email addresses with unusual handles, unconventional characters, excessive numbers, or uncommon symbols (eg, a987quaiigi@yahoo.com and Â2xylzggf@gmail.com); and (4) a few (18/1498, 1.2%) included

XSL•FO
RenderX

a US telephone number with an incorrect digit count (>10 digits or <10 digits). Upon further investigation, we noticed that responses to certain survey items were either nonsensical, highly improbable, or were incongruent with a response to a related item. For example, in an open-text item asking, "How did you hear about PrEP?" some respondents provided nonsensical answers, such as "Jsjsjd," or an illogical response, such as "Yes" (27/1498, 1.8%). A highly improbable response pattern was observed when some entries (17/1498, 1.13%) selected all options for HIV behavioral risk factors (eg, "I had unprotected vaginal sex with a male partner [vaginal sex without using a condom] in the past 6 months," "I had unprotected anal sex with a male partner [anal sex without using a condom] in the past 6 months," "I had sex in exchange for something of value [such as food, shelter, money, or drugs] in the past 6 months," "I have taken drugs by injection with a needle [such as heroin, cocaine, amphetamines, or steroids; not including anything taken under a doctor's order] in the past 6 months," and "I have been diagnosed with an STI in the past 6 months [STIs include chlamydia, gonorrhea, syphilis, human papilloma virus (HPV and warts) and herpes simplex virus]"). Selecting all options was considered highly unlikely.

After evaluating the influx of responses and noting the aforementioned patterns, we decided to call the first 20 people who met the eligibility criteria, consented to participate, and completed the study survey and whose responses showed no signs of fraudulent activity. The goal was to verify whether the provided telephone numbers were valid and belonged to the respondents. The majority of telephone numbers called (16/20, 80%) were linked to a person, company, or business not associated with the individual listed in the survey entry or were disconnected or no longer in service. Only 4 (20%) of the 20 telephone numbers were valid and belonged to the individual who completed the survey. In summary, our evaluation of data entries in relation to the limitations of the survey-based fraud prevention safeguard options highlighted the need to implement additional fraud deterrence and detection procedures to help ensure that valid data were being collected from legitimate research participants.

Application of the Second Level of Fraud Detection: Lesson 2

We made a number of changes before resuming the study. First, we retained and implemented the original 4 safeguard options provided by Qualtrics. In addition to the existing CAPTCHA, which was placed before the first question of the eligibility screener, we added a CAPTCHA before the first question of the study survey to further deter bots. We also decided to manually evaluate all completed eligibility screeners (with the associated study survey) for duplicate records (ie, >1 entry) [38] by assessing whether the same IP address, telephone number, or email address appeared across multiple entries. If >3 entries were found to have the same IP address, we retained the first 3 and removed the rest. If multiple entries were found to have the same telephone number or email address, we kept only the first entry and removed all others. On the basis of prior research and our initial findings, we created and implemented 7 additional data validity benchmarks for fraud detection. Each benchmark included a predetermined decision rule that was created a priori. These benchmarks, along with their corresponding decision rules, are described in Table 1 and were implemented alongside the 4 original Qualtrics safeguard options.

After duplicates were removed, each screener and study survey entry was evaluated by applying the 7 fraud detection benchmarks presented in Table 1. First, any entry with latitude and longitude coordinates outside of the United States was labeled as fraudulent and excluded (ie, ineligible and not enrolled). Any entry that had a completion time of ≤10 minutes was also labeled as fraudulent and not enrolled. All remaining entries with latitude and longitude coordinates within the United States and a completion time of >10 minutes were further evaluated using the 5 remaining benchmarks (items 3-7 in Table 1). These 5 benchmarks evaluated whether the self-reported telephone number contained 10 digits in the United States format (ie, xxx-xxx-xxxx); the self-reported email address was correctly formatted and deemed valid using a third-party email validation tool or was associated with a Facebook profile; the responses to certain survey questions (eg, HIV behavioral risk factors or open-ended survey questions) were nonsensical or highly improbable (ie, selecting having engaged in all 5 options for HIV behavioral risk factors in the 6 months before study participation: unprotected vaginal sex with a male partner, unprotected anal sex with a male partner, injection drug use, sex exchange, and diagnosed with an STI); the self-reported responses to the items about zip code and residential city matched; and name, email address, or open-ended text responses contained nonstandard characters or symbols not commonly used in the United States (eg, Â). To be considered a highly probable respondent with valid data, individuals had to pass all 5 fraud detection benchmarks (items 3-7 in Table 1). Respondents who passed ≤3 were deemed fraudulent (ie, ineligible and not enrolled). Those who passed at least 4 benchmarks were marked for further review. A team member then contacted these individuals via telephone or email to ask 2 questions about the personal information provided in the screener or study survey (eg, current age, birth month, zip code, city of residence, or relationship status). Individuals who could not be reached or provided incorrect answers were categorized as "fraudulent" and not enrolled. The goal of these procedures was to identify additional fraudulent entries to improve the integrity of the data collected.

XSL•FO
RenderX

Table 1. Description of the level 2 fraud detection protocol, which included 7 data validation benchmarks and corresponding decision rules used to identify additional fraudulent survey respondent entries[a].

Data used for benchmark	Benchmark rule	Data source	Decision rule
1. Latitude and longitude coordinates of IP address	Location must be within the United States	Eligibility screener	If individual does not pass, labeled as fraudulent; no longer reviewed
2. Start and finish time stamps	Completion time must exceed 10 minutes	Eligibility screener and study survey	If individual does not pass, labeled as fraudulent; no longer reviewed
3. Telephone number	Digits must be provided in the correct United States format (xxx-xxx-xxxx)	Eligibility screener	If not in United States format, labeled for further review
4. Email address	Email address must be valid when verified online through a third party (eg, s223456h@yahoo.com or have altering letters and numbers – 12s2n4n5P3@gmail.com [28,29])	Eligibility screener	If not in correct format or fake email address, labeled for further review
5. Nonsensical or highly improbable responses to survey questions	Selected all 5 options for HIV behavioral risk factors in the screener or provided irregular responses to open-ended survey questions	Eligibility screener and study survey	If all 5 options for HIV behavioral risk factors are selected, labeled for further review; if nonsensical or highly improbable responses have been provided to open-ended questions, labeled for further review
6. Matched responses: zip code and city	Provided a zip code that did not match their self-reported residential city	Eligibility screener	If self-reported zip code and city do not match, labeled for further review
7. Nonstandard characters or symbols	Name, email address, or open-ended text responses containing nonstandard characters or symbols not commonly used in the United States (eg, Â and Didnâ€™t)	Eligibility screener and study survey	If nonstandard characters or symbols identified, labeled for further review

[a]Respondents were first evaluated against items 1 and 2 of the 7 data validation benchmarks. Those who did not meet these criteria were labeled as "fraudulent." Respondents who passed were then evaluated against the remaining 5 benchmark items (items 3-7) and labeled based on the number of criteria met: "pass" (met all 5 criteria), "further review" (met 4 criteria), and "fraudulent" (met ≤3 criteria).

Ethical Considerations

The procedures associated with the primary study, including ethics approval and oversight, were approved by the institutional review board at the University of Hawai i at Mānoa (IRB 2020-00030). A certificate of confidentiality for human participant research was obtained from the National Institutes of Health to help keep participant data private. Informed electronic consent was obtained from all participants included in the study; each respondent was required to individually provide consent electronically before taking the eligibility screener. Upon study completion, all documents containing identifying information were deidentified and coded with a unique study number. As approved by the ethics committee, enrolled participants (ie, individuals who provided consent, met the eligibility criteria, completed at least 80% of the study survey, and passed all fraud deterrence and detection protocols) were emailed a US $25 electronic gift card for taking part in the study.

Results

Consent, Eligibility Criteria, and Level 1 Fraud Deterrent Protocols

Figure 1 presents the number of entries removed by implementing the 4 fraud deterrent safeguards provided by Qualtrics and the 7 fraud detection benchmarks. There were 5862 entries representing those who clicked the study link and provided consent. First, entries were removed if they did not pass the study eligibility criteria (ie, ineligible; 2150/5862, 36.68%). Entries were also removed if the respondents provided consent and were eligible but completed <80% of the study survey (131/5862, 2.23%). Second, duplicate entries (ie, entries with the same IP address and same telephone number or email address; 588/5862, 10.03%) were removed. With respect to duplicate IP address, there were 1839 entries that had the same IP address as at least 1 other screener and survey. Among these 1839 entries, 121 unique IP addresses were linked to an entry that shared the same IP address with ≥4 screeners and surveys. Following the data validation protocol for entries with the same IP address (ie, retaining the first 3 entries with the same IP address and discarding additional ones), 388 (21.1%) of the 1839 entries were removed. In sum, there were 2993 respondents who provided consent, met the eligibility criteria, and passed level 1 fraud deterrent safeguards.

XSL•FO
RenderX

Figure 1. Final participant enrollment after implementing the 4 fraud deterrent safeguards (provided by Qualtrics) and the 7 fraud detection benchmarks.

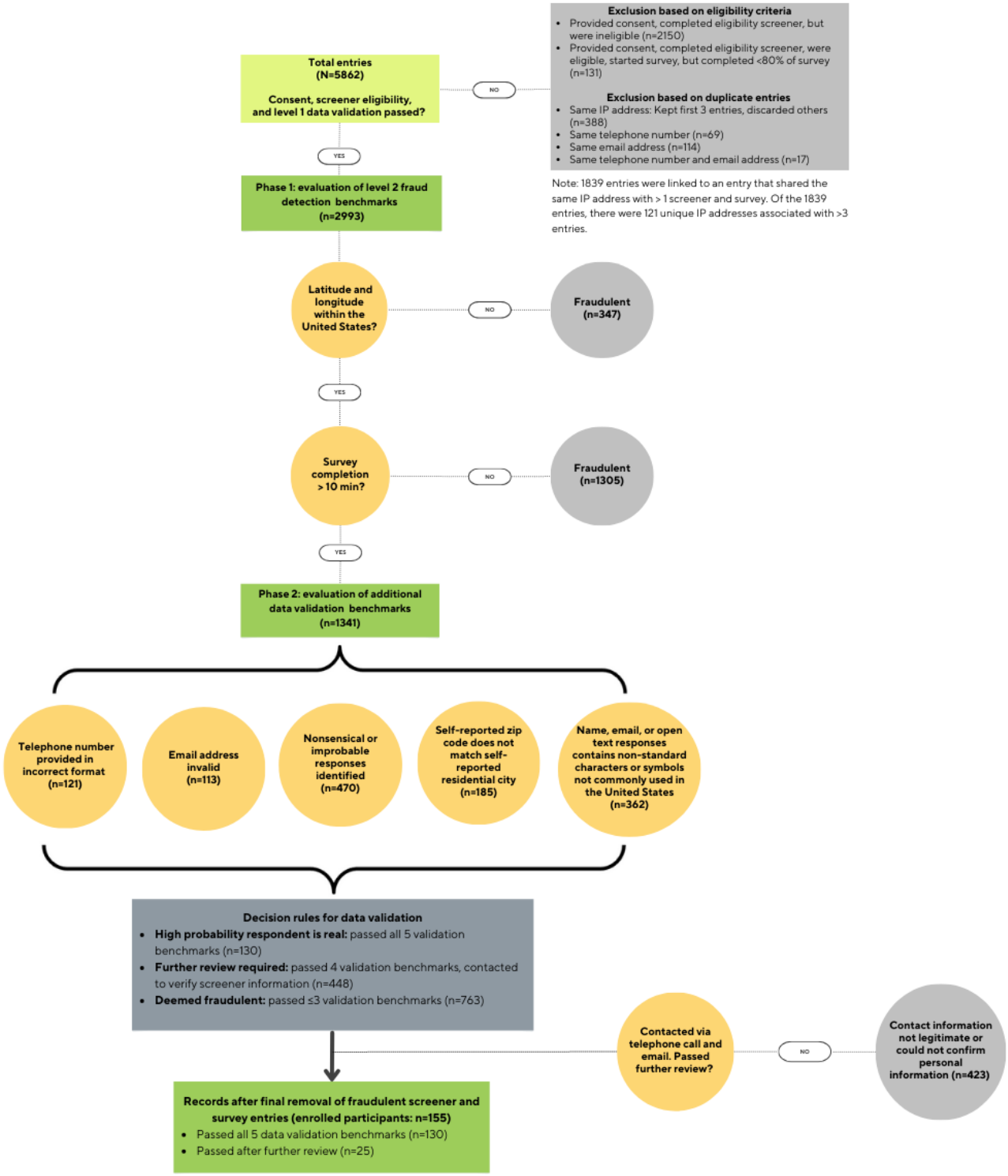

Level 2 Fraud Detection

Seven additional fraud detection benchmarks were implemented, resulting in the removal of additional entries. Of the 2993 entries that remained, 347 (11.59%) with latitude and longitude coordinates outside of the United States were removed, while 1305 (43.6%) were removed for having a survey completion time of <10 minutes. Among the remaining 1341 entries, implementing the 5 remaining fraud detection benchmarks led to the exclusion of 56.9% (763/1341) of the entries that met <4 of the benchmarks. Only 130 (9.7%) of the 1341 respondents passed all 5 data validation benchmarks and were enrolled in the study. A total of 448 respondents were contacted via telephone or email for further legitimacy review, of whom 25 (5.6%) were successfully reached and verified as legitimate.

Enrolled Participants

After removing entries based on the fraud deterrence and detection protocols, 155 Black cisgender women met all study criteria, including providing consent, meeting eligibility requirements, completing at least 80% of the study survey, and

XSL•FO
RenderX

passing all fraud deterrence and detection protocols, and were enrolled in the study.

Discussion

Principal Findings

This study, which focused on Black cisgender women, emphasizes the importance of implementing effective fraud deterrence and detection procedures in incentivized online research. Our findings, stemming from the lessons learned, highlight that solely relying on fraud deterrence features offered by online survey programs (eg, Qualtrics) is insufficient to ensure the collection of valid data from legitimate research participants. We demonstrate and argue that when conducting incentivized online research, additional fraud deterrence and detection protocols are needed to uphold data integrity and research rigor. These protocols should be implemented before the start of recruitment and enrollment, and should be adjusted as needed throughout the process. Of the 5862 entries received for the study, only 155 (2.64%) provided consent, were eligible, and met fraud detection and deterrent criteria. A large proportion of the entries (2838/5862, 48.41%) were deemed fraudulent. Through reflection on the protocols used in this online study, we offer suggestions and ideas for researchers to consider using to help deter and detect fraudulent data entries, with the objective of helping to increase the likelihood of collecting valid data from genuine participants.

Leveraging Available Features in Survey Software Programs to Deter Bots

Effectively identifying and managing bots is important to ensure the accuracy, reliability, and integrity of collected data. We recommend using all available features provided by the survey software while also being aware of their limitations to deter fraud. In this online study, we used the CAPTCHA verification, prevent indexing, prevent ballot box stuffing (ie, duplicate entries), and bot detection features offered by Qualtrics. We recommend leveraging the CAPTCHA feature because it can be used more than once within an online survey. Although 1 prior study [24] found the CAPTCHA feature to be insufficient in dissuading fraudulent responses, the authors did not provide details about how often it was actually used in their survey. In this study, we strategically placed a CAPTCHA before the first question in the eligibility screener as well as before the first question in the study survey. Strategically placing a CAPTCHA before the initial question in both the eligibility screener and the study survey potentially aided in the identification and elimination of additional fraudulent entries that might have bypassed the initial CAPTCHA. This assertion is supported by the discovery of participants (n=131) who provided consent and met the eligibility criteria but did not start the survey (Figure 1). The strategic placement of CAPTCHAs provides an additional layer of security against automated bot submissions and may help reduce the likelihood of receiving invalid responses.

The “prevent ballot box stuffing” option uses cookies to prevent multiple survey entries but does not fully prevent duplicate entries based on IP address or self-reported information, such as name, email address, and telephone number. Although prior studies have used this fraud detection feature, its limitations were not acknowledged or discussed [28,37]. Our evaluation suggests that respondents, in an attempt to identify which responses would qualify them for the study, may have repeatedly taken the eligibility screener using different internet browsers, as evidenced by the high number of duplicate entries from the same IP addresses (n=1839). When we discovered this, we manually evaluated screener entries to label and assess which ones came from the same IP address. We then applied the same process to evaluate screener entries that contained the same email addresses and telephone numbers, given the limitations of the prevent ballot box stuffing option. The implementation of these evaluation methods was valuable in discerning whether entries originated from the same entity. Our experience and the relevant lessons learned underscore the importance of understanding the limitations of fraud deterrence features in online survey software.

It is important to acknowledge that the 4 fraud deterrence features (CAPTCHA verification, prevent indexing, prevent ballot box stuffing, and bot detection) were implemented in the Qualtrics-hosted study survey between December 2020 and January 2022. Since then, Qualtrics may have updated or enhanced these features. Furthermore, Qualtrics offers 2 other fraud deterrence features that were not used in this study: Security Scan Monitor and RelevantID. Security Scan Monitor prevents email scanning software from inadvertently starting a survey session when a survey link is included in the email. This feature applies to all links, regardless of whether they were distributed via Qualtrics or a third-party system (ie, any software or platform not directly associated with Qualtrics, such as marketing automation software). RelevantID analyzes respondent metadata to determine the likelihood of multiple survey attempts by the same respondent by examining browser, operating system, and location details. Future surveys that are hosted on Qualtrics ought to consider using Security Scan Monitor and RelevantID, depending on the needs of the study. We also recommend that researchers refrain from exclusively relying on the fraud deterrence features provided by Qualtrics (or other survey software). Our lessons learned indicate that using a combination of fraud deterrent and fraud detection procedures will help provide a more robust defense against the collection of invalid data in online survey studies.

Enhancing Fraud Detection

Every online survey study is vulnerable to fraudulent entries. As such, the best approach is to identify potential vulnerabilities and implement strategies to prevent their exploitation before starting recruitment and data collection [25]. Initially, researchers should create and implement a protocol that incorporates both fraud deterrent and fraud detection strategies. Drawing from our experiences with this online survey study, we found that a combined approach of fraud deterrent and fraud detection worked best to identify fraudulent entries. Specifically, we—as well as prior research—evaluated the time taken to complete the eligibility screener and study survey [21,24,28,29,36], contacted respondents via telephone or email to verify their identity [36,39], and used matched survey

XSL•FO
RenderX

questions to identify nonsensical or highly improbable responses [21,24,29].

Predetermining an anticipated time range for survey completion adds an additional option to help detect fraud. For instance, data collected about survey completion time can be used to decipher whether a "participant" was likely to have completed it. Establishing a priori criteria for anticipated eligibility screener and study completion time can aid in fraud detection. Before launching this study, we used the automated survey completion time estimate provided by Qualtrics, which was 53 minutes. However, while this estimate accounts for survey flow, it does not account for skip logic or display logic. To address this, we manually tested survey completion times across different response paths. Our completion times ranged between 18 and 30 minutes. On the basis of these findings, we determined that it would be highly unlikely for an individual to complete the eligibility screener and associated study survey in <10 minutes. Therefore, we flagged all entries that completed the screener and survey in <10 minutes and considered them to represent automated, nonhuman involvement or rushed responses. We recommend that researchers consider implementing a similar strategy by predetermining a time range for survey completion to help detect fraudulent or rushed survey entries.

Similar to prior research [24,28], we verified respondents' telephone numbers and email addresses. The verification of contact information is important because respondents can easily create multiple email addresses or use Google Voice numbers to complete the survey multiple times. Third-party services that validate email addresses and Facebook were used to verify email addresses, while telephone numbers were validated through direct telephone calls to confirm legitimacy and ownership. Respondents labeled for "further review" were contacted via telephone or email and asked 2 questions about the personal information provided in the screener or study survey (eg, current age, birth month, zip code, city of residence, and relationship status). Although resource intensive, verifying phone numbers and emails helped to differentiate between legitimate and fraudulent entries, offering confidence in the authenticity of participant details and their survey data.

A notable concern with manual checks and using third-party verification services is the associated cost and time commitment. These expenses can be significant, especially for research projects with limited budgets. The time-consuming nature of manual verification of telephone numbers becomes apparent when handling a high volume of entries within a short time frame. Calling respondents individually can lead to delays and logistical challenges, impacting the efficiency of the research process. To deter fraudulent behavior, Ballard et al [28] sent an SMS text message to suspicious respondents stating, "You recently completed a survey for a health study online. However, we detected that your survey entry was fraudulent. If you think this is a mistake, please contact us." If participants did not respond, the authors considered the survey entry invalid [28]. The authors noted that they did not receive any responses categorized as "fraudulent" [28]. In summary, researchers must weigh the benefits of enhanced participant validation against practical constraints, such as cost, time, and personnel involvement, when considering the adoption of email and telephone verification protocols.

We also incorporated a matched response approach using 2 survey questions to enhance our fraud detection protocol. Respondents' self-reported zip code had to match their self-reported residential city; otherwise, they were flagged for possible fraud. Matching responses to different survey questions can help identify potential fraudulent entries. We recommend that researchers consider using at least 1, if not 2, matched responses to preselected survey questions. For example, asking for a respondent's current age at the beginning of an eligibility screener and then requesting their birth month and year—either toward the end of the screener or in the study survey—may help reveal discrepancies [34,36,37]. Alternatively, a survey could ask for a respondent's age range at the beginning and their specific age toward the end.

Additional Strategies and Recommendations for Fraud Deterrence and Detection

There may be other options that work best for online survey studies that can be used to discourage fraudulent behavior. Pratt-Chapman et al [24] recommend including a check box for participants to acknowledge that responses from ineligible respondents or multiple entries from the same respondent will disqualify them from receiving financial incentives. The authors also suggest indicating the investigators' right to confirm eligibility by telephone (or other means) to aid in identifying bots and eliminating duplicate entries [24]. In addition, when requests for payment are received from respondents whose entries are identified as fraudulent, Dewitt et al [29] recommend informing these respondents that there was a concern about their survey entry and asking them to call a 1-800 study line and leave a callback number for verification. The authors applied this approach to all suspicious entries and found that none resulted in return calls [29]. Another option is to contact participants via telephone or email to set up a Zoom meeting to determine their legitimacy [39] before inviting them to complete the study survey. During this meeting, participants could be asked to upload or show evidence to prove their identity or to verify they had met some of the inclusion criteria; for example, a study recruiting active US military service members could require potential participants to show their military ID as proof of service. The methods described here offer additional strategies for deterring and detecting fraudulent entries in online research studies.

While manual checks can be a viable option for online research studies with limited resources, more financially robust studies can deploy advanced fraud deterrence and detection strategies. One such option is incorporating email and telephone verification features that require respondents to validate their contact information by receiving and verifying a code. Guest et al [36] required respondents to submit their mobile phone number during the eligibility screener to receive a 3-digit verification code via SMS text message. After receiving the 3-digit code, respondents were required to enter the code in the eligibility screener survey to validate their mobile phone number for the study [36]. Those who failed to input the code were unable to continue with the eligibility screener [36].

XSL•FO
RenderX

Another recommendation for online eligibility screening is the use of automated electronic algorithms that allow researchers to create parameters and decision rules for inclusion criteria. Automated electronic algorithms also enable online eligibility screeners to restrict entries based on IP address (ie, blocking multiple entries from the same IP address or requiring a US-based IP address), completion time, and allowable responses that align with the inclusion criteria (eg, current age must be at least 18 years) [37]. These algorithms enhance fraud deterrent strategies while streamlining the eligibility screening process. However, implementing such techniques requires adequate financial and computational resources (eg, a coder or a web developer) tailored to the study's specific needs. It should be noted that implementing these procedures results in a major reduction in the study sample size. This is often the case when researchers opt to prioritize data integrity to ensure that the findings are representative of real people.

Overall, ongoing efforts are needed to refine and optimize fraud deterrence and detection protocols to maintain research rigor and improve the collection of valid data from legitimate research participants. This study uniquely contributes to existing related literature because it is one of the first to evaluate, describe lessons learned, and offer insights into fraud deterrence and detection protocols used for an incentivized online survey study with adult Black cisgender women. Most online research studies that have evaluated the use of fraud deterrence and detection methods have focused on samples of sexual minority men and male couples [11,12,23,34,37]. Further research is needed to evaluate the application and effectiveness of fraud deterrence and detection methods in online studies with diverse populations. It remains unknown whether distinct demographic groups require unique fraud deterrence and detection procedures. We encourage researchers to evaluate and publish findings stemming from their use of fraud deterrence and detection methods to help advance the rigor of online research studies.

Limitations

The limitations of this study are important to consider in light of the lessons learned and the recommendations provided. Despite the changes we made to enhance the fraud deterrence and detection methods used in this online survey study, it is possible that some participants provided false information during the verification attempts. In addition, reliance on telephone calls and emails for participant validation may introduce biases or errors due to participant nonresponsiveness or communication challenges. Participants may forget to respond to telephone calls or emails. As a result, there may be inaccuracies in the data. Although we used a third-party service for email validation, it is possible that some of the verified email addresses belonged to individuals other than the persons or entities completing the eligibility screener. Moreover, the evolving nature of fraud detection algorithms in online survey platforms such as Qualtrics may limit the long-term generalizability of recommendations based on research conducted within specific time frames. In addition, this study's fraud detection and deterrent protocols were designed to recruit Black cisgender women, potentially limiting the generalizability of these protocols to other populations. Future research should aim to use these protocols and provide feedback regarding whether these protocols improved their data integrity by improving their ability to identify bots and imposter participants during recruitment and enrollment. The findings would help to refine and strengthen fraud detection and deterrent protocols used to recruit diverse populations for online studies. It is important to acknowledge that the use of fraud deterrence and detection methods does not guarantee the complete elimination of all fraudulent entries; however, their use and evaluation help to enhance the confidence that valid data are being collected from verified participants, thereby contributing to the rigor and integrity of online research studies.

Conclusions

Effectively identifying fraudulent responses in web-based surveys is an ongoing challenge. With the increasing shift toward web-based research and online recruitment, the threat of fraudulent participation poses a real challenge to data validity. Protocols for identifying fraudulent survey entries and verifying and validating potential study participants should be considered for all internet-based studies. Researchers conducting online studies with Black cisgender women must actively share their experiences in deterring and detecting fraud to help contribute to the rigor of best practices and maintaining the validity of data and associated findings. The lessons learned and recommendations offered from the experiences of conducting this online study, which recruited and enrolled a study sample of Black cisgender women, highlight two important take-home points: (1) develop a thorough fraud deterrent and fraud detection plan to implement before study launch; and (2) monitor and evaluate how well these methods are working while data are being collected, as well as once data collection has ended. We encourage researchers to leverage all resources they may have at their disposal, given the number of different fraud deterrence and detection options that exist. This study—which emphasizes the importance of the aforementioned take-home points—used a combination of fraud deterrence and detection methods to identify a large number of fraudulent entries that would have otherwise been included in the data.

Acknowledgments

This study was funded, in part, by the Surgeon General C. Everett Koop HIV/AIDS Research Grant from the Rural Center for AIDS/STD Prevention, Indiana University School of Public Health-Bloomington. This research was supported in part by the National Institute on Minority Health and Health Disparities of the National Institutes of Health (U54MD012393), Florida International University Research Center in Minority Institutions. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

XSL•FO
RenderX

Data Availability

The datasets generated and analyzed during this study are not currently publicly available because the primary data are still being analyzed. Once primary analyses are complete, the study data will be available from the corresponding author on reasonable request.

Authors' Contributions

AIS contributed to the study design, data collection procedures, analysis, interpretation of the results, and drafting and critically reviewing the paper. JWM contributed to the concept and study design, interpretation of the results, and drafting and critically reviewing the paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
STROBE Checklist.
[PDF File (Adobe PDF File), 84 KB - formative_v9i1e59955_app1.pdf]

References

1. Rhodes SD, Bowie DA, Hergenrather KC. Collecting behavioural data using the world wide web: considerations for researchers. J Epidemiol Community Health 2003 Jan;57(1):68-73 [FREE Full text] [doi: 10.1136/jech.57.1.68] [Medline: 12490652]
2. Pequegnat W, Rosser BR, Bowen AM, Bull SS, DiClemente RJ, Bockting WO, et al. Conducting internet-based HIV/STD prevention survey research: considerations in design and evaluation. AIDS Behav 2007 Jul 20;11(4):505-521. [doi: 10.1007/s10461-006-9172-9] [Medline: 17053853]
3. Evans JR, Mathur A. The value of online surveys: a look back and a look ahead. Internet Res 2018 Aug 06;28(4):854-887. [doi: 10.1108/intr-03-2018-0089]
4. Zahm SH, Pottern LM, Lewis DR, Ward MH, White DW. Inclusion of women and minorities in occupational cancer epidemiologic research. J Occup Med 1994 Aug;36(8):842-847. [Medline: 7807263]
5. Svensson CK. Representation of American Blacks in clinical trials of new drugs. JAMA 1989 Jan 13;261(2):263-265. [doi: 10.1001/jama.1989.03420020117041]
6. Hall WD. Representation of Blacks, women, and the very elderly (aged > or = 80) in 28 major randomized clinical trials. Ethn Dis 1999;9(3):333-340. [Medline: 10600055]
7. Mak WW, Law RW, Alvidrez J, Pérez-Stable EJ. Gender and ethnic diversity in NIMH-funded clinical trials: review of a decade of published research. Adm Policy Ment Health 2007 Nov 10;34(6):497-503. [doi: 10.1007/s10488-007-0133-z] [Medline: 17690976]
8. Isler MR, Brown AL, Eley N, Mathews A, Batten K, Rogers R, et al. Curriculum development to increase minority research literacy for HIV prevention research: a CBPR approach. Prog Community Health Partnersh 2014 Dec;8(4):511-521 [FREE Full text] [doi: 10.1353/cpr.2014.0059] [Medline: 25727984]
9. George S, Duran N, Norris K. A systematic review of barriers and facilitators to minority research participation among African Americans, Latinos, Asian Americans, and Pacific Islanders. Am J Public Health 2014 Feb;104(2):e16-e31. [doi: 10.2105/AJPH.2013.301706] [Medline: 24328648]
10. Demark-Wahnefried W, Schmitz KH, Alfano CM, Bail JR, Goodwin PJ, Thomson CA, et al. Weight management and physical activity throughout the cancer care continuum. CA Cancer J Clin 2018 Jan;68(1):64-89 [FREE Full text] [doi: 10.3322/caac.21441] [Medline: 29165798]
11. Bauermeister JA, Pingel E, Zimmerman M, Couper M, Carballo-Diéguez A, Strecher VJ. Data Quality in web-based HIV/AIDS research: Handling Invalid and Suspicious Data. Field methods 2012 Aug 01;24(3):272-291 [FREE Full text] [doi: 10.1177/1525822X12443097] [Medline: 23180978]
12. Teitcher JE, Bockting WO, Bauermeister JA, Hoefer CJ, Miner MH, Klitzman RL. Detecting, preventing, and responding to "fraudsters" in internet research: ethics and tradeoffs. J Law Med Ethics 2015 Jan 01;43(1):116-133 [FREE Full text] [doi: 10.1111/jlme.12200] [Medline: 25846043]
13. Kramer J, Rubin A, Coster W, Helmuth E, Hermos J, Rosenbloom D, et al. Strategies to address participant misrepresentation for eligibility in web-based research. Int J Methods Psychiatr Res 2014 Mar 16;23(1):120-129 [FREE Full text] [doi: 10.1002/mpr.1415] [Medline: 24431134]
14. Quach S, Pereira JA, Russell ML, Wormsbecker AE, Ramsay H, Crowe L, et al. The good, bad, and ugly of online recruitment of parents for health-related focus groups: lessons learned. J Med Internet Res 2013 Nov 14;15(11):e250 [FREE Full text] [doi: 10.2196/jmir.2829] [Medline: 24231040]

XSL•FO
RenderX

15. Chen J, Wang Y. Social media use for health purposes: systematic review. J Med Internet Res 2021 May 12;23(5):e17917 [FREE Full text] [doi: 10.2196/17917] [Medline: 33978589]
16. Sanchez C, Grzenda A, Varias A, Widge AS, Carpenter LL, McDonald WM, et al. Social media recruitment for mental health research: a systematic review. Compr Psychiatry 2020 Nov;103:152197 [FREE Full text] [doi: 10.1016/j.comppsych.2020.152197] [Medline: 32992073]
17. Farr DE, Battle DA, Hall MB. Using Facebook advertisements for women's health research: methodology and outcomes of an observational study. JMIR Form Res 2022 Jan 12;6(1):e31759 [FREE Full text] [doi: 10.2196/31759] [Medline: 35019843]
18. Ellington M, Connelly J, Clayton P, Lorenzo CY, Collazo-Velazquez C, Trak-Fellermeier MA, et al. Use of Facebook, Instagram, and Twitter for recruiting healthy participants in nutrition-, physical activity-, or obesity-related studies: a systematic review. Am J Clin Nutr 2022 Feb 09;115(2):514-533 [FREE Full text] [doi: 10.1093/ajcn/nqab352] [Medline: 34669955]
19. Willis TA, Wright-Hughes A, Skinner C, Farrin AJ, Hartley S, Walwyn R, et al. The detection and management of attempted fraud during an online randomised trial. Trials 2023 Aug 04;24(1):494 [FREE Full text] [doi: 10.1186/s13063-023-07517-4] [Medline: 37537678]
20. Kayrouz R, Dear BF, Karin E, Titov N. Facebook as an effective recruitment strategy for mental health research of hard to reach populations. Internet Interv 2016 May;4:1-10 [FREE Full text] [doi: 10.1016/j.invent.2016.01.001] [Medline: 30135786]
21. Pozzar R, Hammer MJ, Underhill-Blazey M, Wright AA, Tulsky JA, Hong F, et al. Threats of bots and other bad actors to data quality following research participant recruitment through social media: cross-sectional questionnaire. J Med Internet Res 2020 Oct 07;22(10):e23021 [FREE Full text] [doi: 10.2196/23021] [Medline: 33026360]
22. Wang J, Calderon G, Hager ER, Edwards LV, Berry AA, Liu Y, et al. Identifying and preventing fraudulent responses in online public health surveys: lessons learned during the COVID-19 pandemic. PLOS Glob Public Health 2023 Aug 23;3(8):e0001452 [FREE Full text] [doi: 10.1371/journal.pgph.0001452] [Medline: 37610999]
23. Birnbaum MH. Human research and data collection via the internet. Annu Rev Psychol 2004 Feb 01;55(1):803-832. [doi: 10.1146/annurev.psych.55.090902.141601] [Medline: 14744235]
24. Pratt-Chapman M, Moses J, Arem H. Strategies for the identification and prevention of survey fraud: data analysis of a web-based survey. JMIR Cancer 2021 Jul 16;7(3):e30730 [FREE Full text] [doi: 10.2196/30730] [Medline: 34269685]
25. Lawlor J, Thomas C, Guhin AT, Kenyon K, Lerner MD, UCAS Consortium, et al. Suspicious and fraudulent online survey participation: introducing the REAL framework. Methodol Innov 2021 Oct 28;14(3). [doi: 10.1177/20597991211050467]
26. Yarrish C, Groshon L, Mitchell JK, Appelbaum AI, Klock S, Winternitz T, et al. Finding the signal in the noise: minimizing responses from bots and inattentive humans in online research. The Behavior Therapist 2019;42(7):235-242 [FREE Full text]
27. Bowen AM, Daniel CM, Williams ML, Baird GL. Identifying multiple submissions in internet research: preserving data integrity. AIDS Behav 2008 Nov 1;12(6):964-973 [FREE Full text] [doi: 10.1007/s10461-007-9352-2] [Medline: 18240015]
28. Ballard AM, Cardwell T, Young AM. Fraud detection protocol for web-based research among men who have sex with men: development and descriptive evaluation. JMIR Public Health Surveill 2019 Feb 04;5(1):e12344 [FREE Full text] [doi: 10.2196/12344] [Medline: 30714944]
29. Dewitt J, Capistrant B, Kohli N, Rosser BR, Mitteldorf D, Merengwa E, et al. Addressing participant validity in a small internet health survey (the restore study): protocol and recommendations for survey response validation. JMIR Res Protoc 2018 Apr 24;7(4):e96 [FREE Full text] [doi: 10.2196/resprot.7655] [Medline: 29691203]
30. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. The Lancet 2007 Oct;370(9596):1453-1457. [doi: 10.1016/s0140-6736(07)61602-x]
31. Sophus AI, Mitchell JW, Barroso J, Sales JM. Factors associated with planned future use of PrEP in the next 3 months and likelihood to use PrEP among Black cisgender HIV-negative women in Texas. AIDS Behav 2024 Jan 28;28(1):72-92. [doi: 10.1007/s10461-023-04188-9] [Medline: 37768428]
32. Sophus AI, Mitchell JW, Sales JM, Braun K. "Our community comes first": investigating recruitment ads that represent and appeal to Black women for online, HIV-related research studies. J Racial Ethn Health Disparities 2024 Dec 18;11(6):3478-3488. [doi: 10.1007/s40615-023-01800-5] [Medline: 37723375]
33. Konstan JA, Simon RB, Ross MW, Stanton J, Edwards WM. The story of subject naught: a cautionary but optimistic tale of internet survey research. J Comput Mediat Commun 2005;10(2). [doi: 10.1111/j.1083-6101.2005.tb00248.x]
34. Mitchell J, Lee JY, Stephenson R. How best to obtain valid, verifiable data online from male couples? Lessons learned from an eHealth HIV prevention intervention for HIV-negative male couples. JMIR Public Health Surveill 2016 Sep 20;2(2):e152 [FREE Full text] [doi: 10.2196/publichealth.6392] [Medline: 27649587]
35. Grey JA, Konstan J, Iantaffi A, Wilkerson JM, Galos D, Rosser BR. An updated protocol to detect invalid entries in an online survey of men who have sex with men (MSM): how do valid and invalid submissions compare? AIDS Behav 2015 Oct 25;19(10):1928-1937 [FREE Full text] [doi: 10.1007/s10461-015-1033-y] [Medline: 25805443]

XSL•FO
RenderX

36. Guest JL, Adam E, Lucas IL, Chandler CJ, Filipowicz R, Luisi N, et al. Methods for authenticating participants in fully web-based mobile app trials from the iReach project: cross-sectional study. JMIR Mhealth Uhealth 2021 Aug 31;9(8):e28232 [FREE Full text] [doi: 10.2196/28232] [Medline: 34463631]
37. Mitchell JW, Chavanduka TM, Sullivan S, Stephenson R. Recommendations from a descriptive evaluation to improve screening procedures for web-based studies with couples: cross-sectional study. JMIR Public Health Surveill 2020 May 12;6(2):e15079 [FREE Full text] [doi: 10.2196/15079] [Medline: 32396133]
38. Griffin M, Martino RJ, LoSchiavo C, Comer-Carruthers C, Krause KD, Stults CB, et al. Ensuring survey research data integrity in the era of internet bots. Qual Quant 2022 Oct 05;56(4):2841-2852 [FREE Full text] [doi: 10.1007/s11135-021-01252-1] [Medline: 34629553]
39. Saberi P. Research in the time of coronavirus: continuing ongoing studies in the midst of the COVID-19 pandemic. AIDS Behav 2020 Aug 18;24(8):2232-2235 [FREE Full text] [doi: 10.1007/s10461-020-02868-4] [Medline: 32303924]

Abbreviations

STI: sexually transmitted infection
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by A Mavragani; submitted 26.04.24; peer-reviewed by V Cofini, V Astha, M Hill; comments to author 20.11.24; revised version received 19.12.24; accepted 25.01.25; published 12.03.25.

Please cite as:
Sophus AI, Mitchell JW
Assessment of Fraud Deterrence and Detection Procedures Used in a Web-Based Survey Study With Adult Black Cisgender Women: Description of Lessons Learned and Recommendations
JMIR Form Res 2025;9:e59955
URL: https://formative.jmir.org/2025/1/e59955
doi:10.2196/59955
PMID:

©Amber I Sophus, Jason W Mitchell. Originally published in JMIR Formative Research (https://formative.jmir.org), 12.03.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Trauma Support App for Young People: Co-design and Usability Study

Maria Thell[1], MSc; Kerstin Edvardsson[1], MSc; Reem Aljeshy[1]; Kalid Ibrahim[1]; Georgina Warner[1], MA, PhD

Child Health and Parenting, Department of Public Health and Caring Science, Uppsala University, Uppsala, Sweden

Corresponding Author:
Georgina Warner, MA, PhD
Child Health and Parenting, Department of Public Health and Caring Science
Uppsala University
BMC, Husargatan 3
Uppsala, 751 23
Sweden
Phone: 46 18 471 6632
Email: georgina.warner@uu.se

Abstract

Background: One of the most common reasons young people with mental health issues, such as posttraumatic stress disorder, do not seek help is stigma, which digital support tools could help address. However, there is a lack of trauma support apps specifically designed for young people. Involving the target group in such projects has been shown to produce more engaging and effective results.

Objective: This study aimed to apply a child rights–based participatory approach to develop a trauma support app with young people.

Methods: Seven young people (aged 14-19 years; 3 males and 4 females) with experiences of trauma were recruited as coresearchers. A child rights–based framework guided the working process. The app was developed through a series of Design Studio workshops and home assignments, using the manualized intervention Teaching Recovery Techniques as the foundation for its content. The coresearchers were trained in research methodology and conducted usability testing with other young people (n=11) using the think-aloud method, the System Usability Scale (SUS), and qualitative follow-up questions.

Results: A functional app prototype was developed using a no-code platform, incorporating various trauma symptom management techniques. These techniques covered psychoeducation, normalization, relaxation, and cognitive shifting, presented in multiple formats, including text, audio, and video. The contributions of the coresearchers to the design can be categorized into 3 areas: mechanics (rules and interactions shaping the app's structure), dynamics (user-visible elements, such as the outcome when pressing a button), and aesthetics (the emotional responses the app aimed to evoke in users during interaction). Beyond influencing basic aesthetics, the coresearchers placed significant emphasis on user experience and the emotional responses the app could evoke. SUS scores ranged from 67.5 to 97.5, with the vast majority exceeding 77.5, indicating good usability. However, usability testing revealed several issues, generally of lower severity. For instance, video content required improvements, such as reducing light flickering in some recordings and adding rewind and subtitle selection options. Notably, the feature for listening to others' stories was removed to minimize emotional burden, shifting the focus to text formats with more context.

Conclusions: Young people who have experienced trauma can actively participate in the cocreation of a mental health intervention, offering valuable insights into the needs and preferences of their peers. Applying a child rights–based framework to their involvement in a research project supported the fulfillment of the Convention on the Rights of the Child Article 12.

(JMIR Form Res 2025;9:e57789) doi:10.2196/57789

KEYWORDS

co-design; young people; trauma; app development; usability testing

XSL•FO
RenderX

Introduction

Background

It is well recognized that one of the main barriers preventing young people from seeking face-to-face help for mental health problems, such as symptoms of posttraumatic stress disorder (PTSD), is stigmatization and embarrassment [1-3]. This includes perceived stigma, fear of stigma, societal stigma, and self-stigma [2], and may result in many young people experiencing PTSD symptoms not receiving the help and support they need. However, these well-known barriers can potentially be overcome through app-based support [3], as mental health apps can engage young people who might not otherwise seek help through traditional routes. Particularly, in the case of PTSD, which differs from other mental illnesses due to its clear cause, young people could receive scalable trauma interventions without delay through digital technology [4]. Despite this, there is a lack of apps specifically developed for young people [5,6].

Teaching Recovery Techniques (TRT) is a manualized group intervention for children aged 8 and older with trauma symptoms [7]. The intervention consists of 5 sessions for children and 2 sessions for parents and caregivers. The content for children and adolescents includes components aligns with trauma-focused cognitive behavioral therapy: psychoeducation, trauma narrative, affective modulation, cognitive coping and processing, in vivo mastery of trauma memories, and future development. TRT helps normalize responses to trauma and provides strategies to manage intrusive thoughts and memories as well as regulate arousal. Participants in a TRT group also work on gradually exposing themselves to thoughts and situations they avoid [7]. TRT has been evaluated in randomized controlled trials with children and young people in several international settings, with an overall positive impact on symptoms of PTSD and depression reported [8]. The initial idea for the current project emerged during a previous participatory study in which TRT was adapted to an online format in response to the COVID-19 pandemic [9]. During the adaptation phase, which involved TRT group leaders, parents, and young people with experience of TRT, the young people suggested consolidating all the techniques into a separate digital format, that is, a mobile app [9].

There has been a growing use of participatory methods with children and young people in the development of health interventions [10]. Previous research indicates that involving the target group in such projects can lead to more engaging and useful outcomes [11,12]. However, when conducting participatory research in the context of PTSD, it is important to recognize that traumatic events create a power imbalance, where one entity holds power over another. An individual's experience may be shaped by feelings of powerlessness, humiliation, guilt, shame, betrayal, or silencing. In interpersonal interactions, especially in public health research, it is crucial to ensure that these feelings of powerlessness are not replicated or reinforced [13].

It has been suggested that research involving young people as coresearchers should be conducted within a rights-based framework [14]. One argument is that this approach can help address the power imbalance between adults and children. The Convention on the Rights of the Child (CRC) [15] acknowledges the position and disempowerment of children in matters affecting them. Adopting a child rights–based approach as a researcher involves implementing and being accountable to the rights set out in the CRC [16]. For this study, we applied the Lundy model [17] of child and youth participation, including the associated planning checklist for researchers [18]. The Lundy model conceptualizes the two components of Article 12 of the CRC: (1) the right to express views and (2) the right to have those views duly considered. By organizing the 4 elements—space, voice, audience, and influence—in a logical, chronological order, the model helps researchers plan, implement, and monitor projects. Space refers to providing a safe and inclusive environment for children to express their views and ensuring that all children can participate. Voice involves offering appropriate information and facilitating the expression of children's views. Audience ensures that children's views are communicated to someone responsible for listening. Influence involves ensuring that children's views are taken seriously and acted upon where appropriate [17].

Objective

This study aimed to (1) apply a child rights–based participatory approach to developing a trauma support app with young people who have personal experiences of trauma and (2) conduct usability testing of the app with the young people as coresearchers.

Methods

Phase 1: App Development

Coresearcher Recruitment

Seven young people were recruited for the project through 3 different channels. Two of them had previously collaborated with the research team on a related TRT research project [9] and were invited to participate again. Additionally, contact was made with a youth-run organization for young people with experience in out-of-home care, which agreed to distribute the recruitment advertisement to its members. Furthermore, a youth center in an area categorized as "vulnerable" by the Swedish Police Authority due to high rates of crime and social exclusion [19] posted the advertisement on its physical noticeboard and social media accounts. Examples of traumatic events and situations, as well as PTSD symptoms, were included in the recruitment materials. Young people who identified with these experiences and enjoyed being creative were encouraged to contact the researchers via email, SMS text message, or WhatsApp (Meta Platforms, Inc.). The advertisement emphasized that participants were not expected to share their personal experiences while working on the project.

Following an initial individual dialogue with all interested young people, 7 chose to participate as coresearchers in the project. The youth coresearcher team consisted of 3 boys and 4 girls, aged 14-19 at the start of the project. They lived in different neighborhoods in 1 large city and 1 medium-sized city in southern Sweden. The older youth attended both theoretical and practical secondary school programs. In accordance with the regulations of the Swedish Work Environment Authority [20],

XSL•FO
RenderX

information about the project was provided to the guardians of participants under 18 years of age. Additionally, permission to participate was obtained from the guardians of those under 16 years of age [20]. The coresearchers were remunerated for their time spent on project activities in line with existing recommendations [21].

Before starting the app development process, a "start-up" session was conducted (Figure 1) to help the team get acquainted through games and other activities. During this session, decisions were made about how to collaborate, including setting common approaches and ground rules to guide the work throughout the project. At the end of the session, an experienced TRT group leader visited to introduce the intervention. As part of this introduction, they guided the group through a popular visual imagery technique called "Safe Inner Place."

Figure 1. Overview of the Teaching Recovery Techniques app development process with youth co-researchers conducted over 10 months.

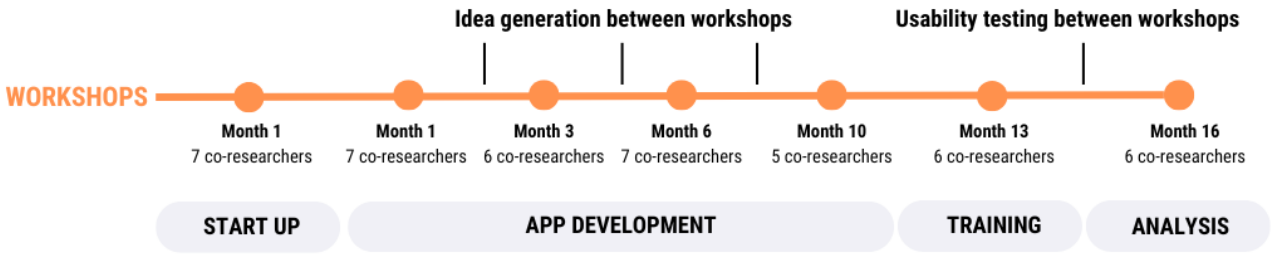

App Development Workshops

The app was developed over 4 workshops spanning 10 months (Figure 1), using an adapted version of the Design Studio Workshop method [22]. In addition, the coresearchers individually generated ideas between workshops. To support their independent work, app design prototypes, along with questions and topics for the next workshop, were shared with them. The researcher leading the workshops (MT) collected the individual contributions before each session and presented them to the group for collaborative refinement.

The first app development workshop aimed to cocreate a logo for the app and pilot the Design Studio Workshop method [22]. The coresearchers were given tablets with internet access and individually explored ideas, first by analyzing their favorite app logos and then by considering a design for a trauma support app. They submitted their ideas to the researcher leading the workshops (MT), who compiled them and projected them onto a screen using a digital noticeboard software tool [23]. Applying the Design Studio method, the team explored colors, fonts, and other design elements. The final logo incorporated the strongest elements from the generated ideas, featuring "soft," "encouraging," and "happy" colors while avoiding excessive or overused colors in app logos. The coresearchers selected a font that contrasted with the background and stood out from other popular apps. Figure 2 illustrates the logo development process.

XSL•FO
RenderX

Figure 2. Overview of the Teaching Recovery Techniques (TRT) app logo development process using the Design Studio method.

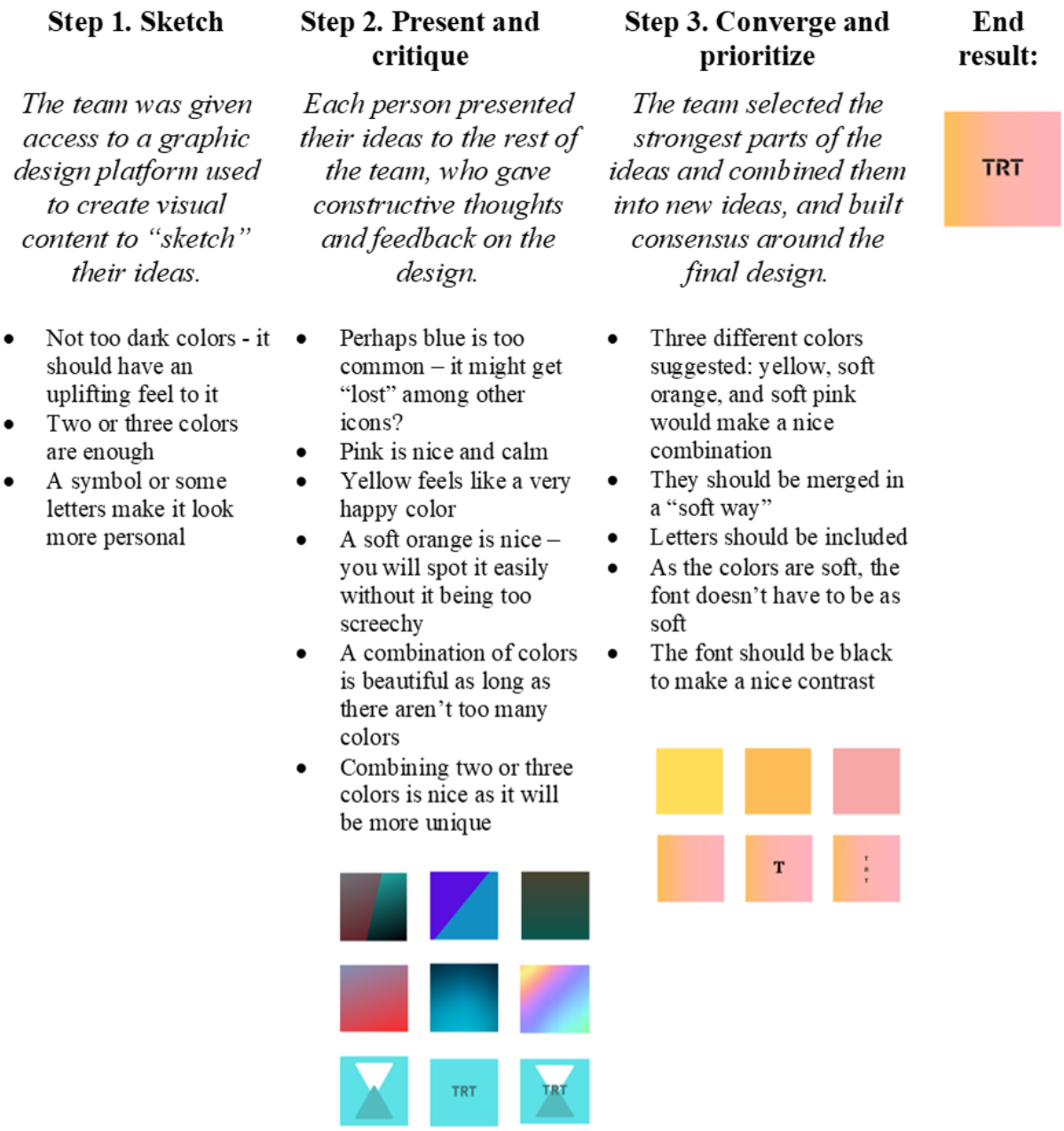

The second workshop focused on prioritizing content and layout for the app using the TRT manual. In the preparatory phase, coresearchers were provided with concise written descriptions of the techniques outlined in the TRT manual. While simplified, these descriptions carefully preserved the original phrasing to maintain the integrity of the intervention logic. Coresearchers were tasked with considering how each technique could be integrated into an app format. To accommodate diverse preferences and abilities, they were encouraged to express their ideas in various ways—whether orally, in writing, or through visual representations such as images or vision boards. Consequently, the initial "sketch" step of the Design Studio method was conducted independently. Upon convening for the workshop, participants engaged in the second step, during which the researcher leading the workshops (MT) presented the collective ideas and facilitated a critique of the proposed concepts. The ideas were then synthesized and prioritized. To discuss app content related to traumatic experiences, case vignettes were used in the workshops, allowing coresearchers to explore different traumatic experiences in a general rather than personal context. Following the workshop, the selected concepts were developed into a prototype of the TRT app using a no-code app-building program [24].

During the third workshop, the coresearchers were presented with an app prototype developed by a research assistant (KE), incorporating their ideas. They provided feedback on both the content and design. For certain techniques, the prototype successfully met their expectations. In other cases, minor

XSL•FO
RenderX

adjustments were made in real time during the workshop using the app-building program [24]. For techniques requiring more substantial revisions, the Design Studio cycle was reinstated. This process involved coresearchers initially "sketching" the necessary amendments, followed by presentations and critiques, ultimately converging and prioritizing the changes. Having access to the app-building program during the workshops facilitated this process, as coresearchers could explore the range of functionalities available within the program. Ideas that the coresearchers found valuable but were not feasible to implement at the time were placed in what the team collectively called the "ice box." This meant that the ideas would be revisited in any future development of the app. One such example was the app's color scheme, which had been jointly agreed upon. Some coresearchers had also expressed a preference for an alternative, slightly darker color scheme for nighttime use. While the first prototype featured only a single color scheme, the idea of selectable color schemes was considered both valuable and important. Rather than being discarded, it was added to the ice box for future consideration.

Further revisions to the app were made after workshop 3 and presented to the coresearchers for feedback in workshop 4. Some changes, primarily concerning the wording of written content, were informed by discussions among the academic researchers to ensure the integrity of the TRT intervention logic. Any modifications resulting from these discussions were communicated to the coresearchers during workshop 4. As in the previous workshop, coresearchers were invited to provide feedback on the prototype. After a series of minor revisions, they expressed their satisfaction with the product, and a collective decision was made to transition from app development to the usability testing phase of the project.

Overall, attendance at the workshops remained consistently high (Figure 1). One of the 7 coresearchers participated in the entire development process but did not take part in the training and usability testing described below. Other absences were due to illness and were compensated for through individual meetings or conversations with the researcher leading the workshops (MT). Those who were ill also completed the home assignments related to the missed session.

Phase 2: Usability Testing

Coresearcher Training

The usability testing was conducted by the coresearchers, with no academic researchers present in the room during data collection. Given the youth-led nature of the process, training sessions were held before the usability testing began (Figure 1). The training covered both theoretical and practical aspects of research ethics and methodology, enabling the coresearchers to actively lead and carry out the research tasks [25-27].

Careful consideration was given to the safety of study participants. As part of the training, the coresearchers and the academic researcher reflected together on how to recognize when a person might be negatively affected and practiced the necessary actions to take if needed. While training is considered essential for young people conducting research, it is also important to balance the extent of training with maintaining their ability to participate meaningfully [14]. Although the training in this study was based on existing guidelines for research with children [25], it was adapted for young people by condensing the content into 3 sessions held on the same day, with breaks in between. Plans were in place to provide additional sessions if needed, but both the coresearchers and researchers agreed that the initial sessions provided sufficient preparation.

Participant Recruitment

The target group for the usability testing consisted of young people aged 15-21 years who spoke and understood Swedish. As the purpose of the usability testing was to assess the intuitiveness of the design and evaluate it with users who had no prior exposure to the app—rather than to determine its efficacy as a trauma support tool—a population sample was recruited instead of a clinical sample with trauma symptomatology. The goal was to recruit at least five and up to 15 participants, ensuring age and gender representation. Research has shown that gender plays a role in usability testing with children and young people, with girls tending to report more issues than boys [28]. Studies indicate that testing with 5 participants per target group is sufficient to identify approximately 80% of usability problems [29].

The recruitment process consisted of 2 steps. First, the coresearchers identified other young people within their networks, provided initial information, and inquired whether they might be interested in participating in the study. These young people were primarily from the coresearchers' school networks and, like the coresearchers, resided in 1 large city and 1 medium-sized city in southern Sweden. If interested, they were then contacted by the academic researchers, who provided additional oral and written information and obtained consent for participation. The date and time of each usability test were mutually decided, with most sessions taking place in the evenings, as the coresearcher, participant, and researcher all needed to be physically present on-site. Five of the coresearchers conducted individual usability tests with participants (n=11; 6 male and 5 female; 15-20 years) over a 6-week period in January and February 2022.

Usability Testing Methods

To assess the app's usability, the think-aloud method was applied for data collection. This method involves participants verbalizing their thoughts while completing specific or nonspecific tasks. Participants are encouraged to articulate what they are looking at, thinking, doing, and feeling as they navigate the app. During the usability test, the observer records participants' verbalized thoughts and actions without interpreting them, paying particular attention to moments or areas where users experience difficulties [30].

All usability tests were conducted on university premises, with the coresearcher and participant in the same room and an academic researcher in an adjacent room. The participant was provided with a mobile phone and instructed to navigate through the entire app using the main menu while thinking aloud. The coresearcher documented the participant's verbalized thoughts and observed behaviors as described above. After completing the app test, participants were asked by the coresearcher to

XSL•FO
RenderX

complete the System Usability Scale (SUS) [31], which had been translated into Swedish. Following this, they participated in a structured interview led by the coresearcher to capture their general impressions of the app [32]. All notes from the test sessions were documented in Research Electronic Data Capture (REDCap), a secure online survey platform, during the session. No voice or video recordings were made.

Usability Testing Analysis

It was agreed in advance that immediately after each session, the coresearcher would report any major or serious usability problems identified during testing to the academic researcher. Following the procedure described by Hertzum [32], 2 academic researchers independently reviewed all notes from the test sessions, each compiling a list of positive and negative usability issues. These lists were then compared, discussed, and merged into a single final list. This consolidated list of usability issues was presented to the coresearchers in a dedicated meeting (Figure 2). Using a live interaction polling tool [33], the usability issues were displayed, and coresearchers individually rated their severity and suggested a priority order. The final prioritization was determined through group discussion.

Ethical Considerations

The coresearchers participated in the project as experts rather than as participants of the study; therefore, standard ethical requirements did not apply. However, their safety was carefully considered throughout the process. From the outset, it was made clear that coresearchers were not expected to share personal experiences of trauma, as this was not necessary for the project. Discussing one's own traumatic experiences could lead to retraumatization or secondary traumatization of others in the group [34]. To mitigate this risk, case vignettes were used in the workshops when developing app content related to trauma. This approach enabled coresearchers to engage in discussions about different traumatic experiences in a general manner rather than on a personal level.

Another key safety consideration was the nature of the tasks assigned to coresearchers between workshops. To minimize the risk of any negative impact, tasks conducted outside the workshop setting focused on app design and layout rather than potentially sensitive content.

The usability testing phase, involving young people as study participants, was submitted to and approved by the Swedish Ethical Review Authority (reference number 2022-04217-01). Study participants were fully informed about the research, and their consent was documented. As participants had access to the app's full content during testing, a safety protocol was implemented. Before each test session, the academic researcher checked in with participants to assess their well-being. If a participant appeared unwell, the test session was not conducted. After each session, the academic researcher conducted separate follow-up conversations with both the participant and the coresearcher to ensure that no one had been adversely affected by the testing experience. If any issues arose during a session, such as a participant showing signs of distress, the coresearcher conducting the test would have immediately alerted the researcher, as practiced during the training session. As part of the safety protocol, the researcher documented the contact details of guardians for participants under 18 years of age before each test session.

Results

Phase 1: App Development

Overview

The TRT app was developed as a progressive web app using a no-code, cross-platform application development program [24]. During the initial discussions about compiling TRT into an app, the young people highlighted potential issues with native apps, particularly regarding data storage requirements. By building the app as a progressive web app, these concerns were mitigated while still providing a user experience comparable to that of a native app.

The coresearchers' contributions to the app's design can be categorized into 3 areas using the Mechanics, Dynamics, and Aesthetics (MDA) framework [35]. Mechanics refers to the fundamental rules and interactions that define the app's structure. Dynamics encompasses what the user experiences in real time, such as the response when a button is pressed, which directly supports the app's aesthetics. Aesthetics pertains to the intended emotional responses elicited in users when interacting with the app and includes its purpose and narrative. Beyond influencing visual elements such as colors, fonts, and symbols, the coresearchers placed significant emphasis on the overall user experience. A selection of app screenshots is presented in Figure 3.

XSL•FO
RenderX

Figure 3. A selection of screenshots from the Teaching Recovery Techniques (TRT) app.

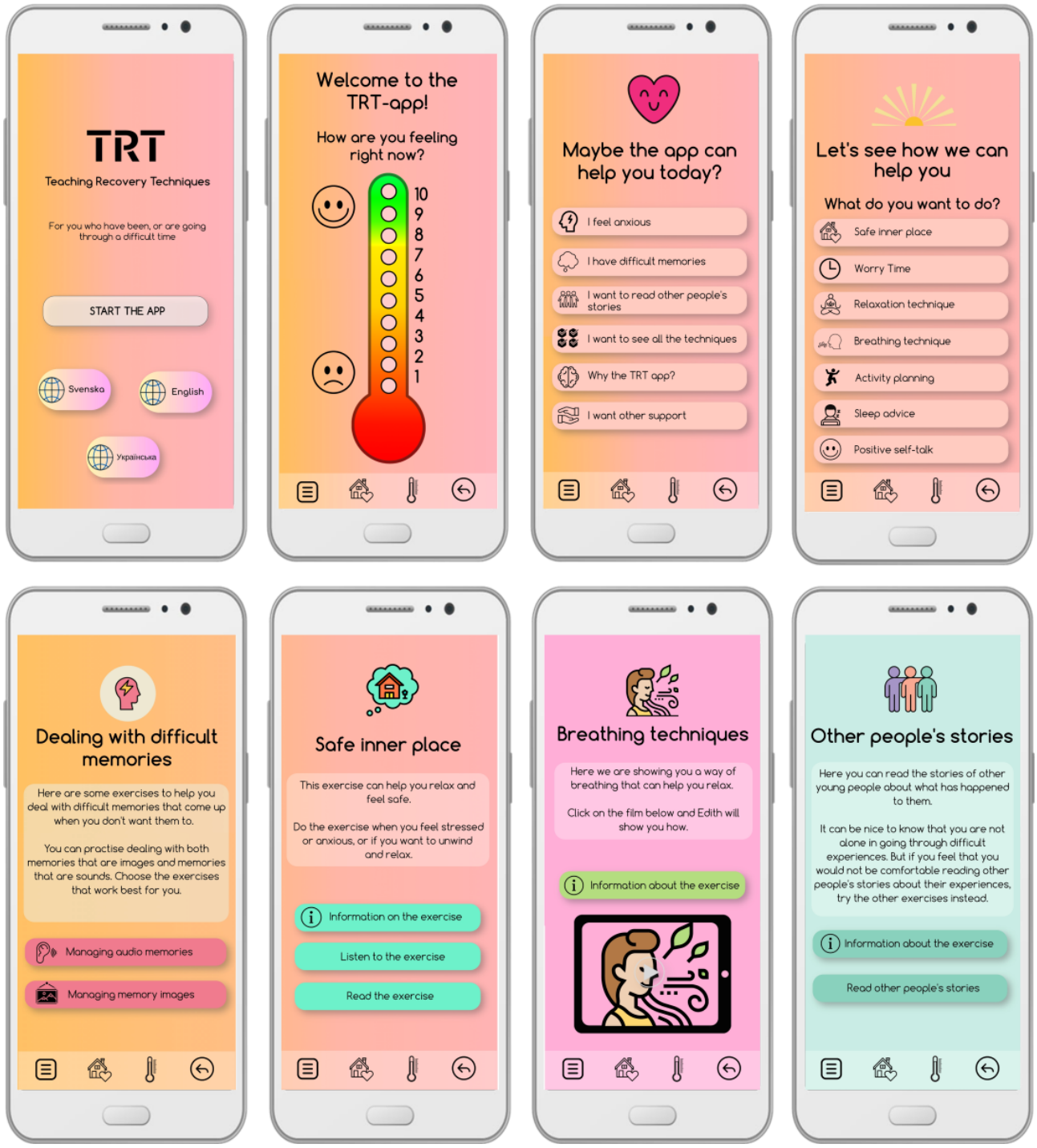

Mechanics

The coresearchers' contribution to the app's mechanics—its basic rules and structure—included selecting techniques from the TRT manual, determining the format of the exercises, and structuring the app's content. They emphasized the importance of the app's structure, particularly the role of the main menu, which is the first thing users see upon entering the app. A clear and intuitive menu ensures easy and efficient navigation, allowing users to quickly find what they need. This not only saves time but also minimizes frustration when searching for specific features or content. The menu provides an overview of the app's content, enhancing the user experience and improving usability. By featuring a front page with a well-organized menu, users can easily select their area of interest—whether it be reading texts, watching films, or completing exercises.

Both coresearchers with prior experience of TRT and those without agreed that the techniques in the app should follow the same order as in the manual—that is, as they would be introduced in a TRT group. There was consensus that the technique groupings should be labeled according to how young people who have had difficult experiences might feel (eg, "I feel anxious" or "I have intrusive memories"; Figure 3), making navigation more intuitive. Throughout the development process, coresearchers and academic researchers discussed the differences

XSL•FO
RenderX

between being introduced to the techniques via an app versus in a facilitated group session. In the latter, group leaders present the exercises and provide support in response to participants' reactions. Given this distinction, the coresearchers emphasized the importance of clear presentation within the app, ensuring that each exercise included information on its purpose and how to use it effectively.

Dynamics

The coresearchers emphasized the importance of the app being adaptable to the preferences of the young person using it, which is central to the app's dynamics—the visible elements and user interactions. The content primarily consists of simple illustrative icons, instructional and demonstrational films, audio recordings, and short texts. Long texts were seen as detrimental to the user experience; however, careful consideration was required to shorten the content while maintaining fidelity to the TRT manual and the therapeutic intent of each technique. The descriptive texts introducing the techniques were kept short and simple. For young people seeking more information about the purpose of each technique, an "Information about the exercise" option is available (Figure 3). To enhance customization, most exercises offer the choice of either reading or listening to the instructions. Additionally, exercises presented as videos include optional subtitles.

Aesthetics

The app's visual aesthetics were intentionally kept simple. The screen layout was designed to be uniform, featuring a representative icon and header at the top, followed by text or buttons framed by softly rounded boxes. An "easy-to-read" font with rounded edges was selected. The orange-pink color of the logo was chosen for its soft and friendly appearance—described as being "like the sun coming in from the left." The app's color theme was primarily based on the logo's colors, with additional shades introduced throughout development to enhance variation and create clearer visual distinctions between different techniques (Figure 3).

The app's aesthetics also relate to the emotional responses it aimed to evoke in users. The coresearchers emphasized that in-app films and voice recordings can be particularly helpful in promoting a sense of calm. These elements were seen as valuable for creating a relaxed atmosphere, providing guidance and instructions, and adding variety and creativity to the app experience. At the same time, the coresearchers stressed the importance of offering different exercise formats, recognizing that individuals have diverse preferences and needs when it comes to relaxation. While some may prefer breathing exercises, others might find meditation or guided visualization more effective. By providing a wide range of exercises, the app can better accommodate different users. In the videos, exercises are led by young people, as the coresearchers underscored the importance of featuring individuals with whom they can identify or relate.

Overview of App Content and Navigation

Basis of Support Techniques

The trauma support techniques featured in the app are based on the TRT manual and incorporate several common approaches to trauma intervention, including psychoeducation, normalization, relaxation, and cognitive shifting. Psychoeducation provides individuals with information to enhance their understanding of psychological conditions—PTSD in this context. Normalization helps users recognize that their symptoms are not signs of weakness or character flaws but rather a natural response to their psychological condition. Relaxation techniques aim to alleviate stress-related symptoms, while cognitive shifting involves deliberately redirecting one's attention from a singular fixation to alternative focal points. For safety reasons, the TRT app intentionally excludes in vivo exposure—a technique that involves directly confronting a feared object, situation, or activity in real life. A comprehensive overview of the app's content is provided below.

Main Menu

To facilitate navigation for both new and experienced users, the app offers 2 ways to explore its content. In the main menu, the "Why the TRT app?" option provides first-time users with an introduction to TRT, trauma, trauma reactions, and how the app is intended to be used. Additionally, the main menu is structured around the user's immediate needs, suggesting groups of exercises based on whether they are feeling anxious, experiencing intrusive memories, or seeking to learn more about trauma through psychoeducation and normalization of trauma reactions. For young people who are more experienced with TRT or the app and want to find a specific exercise, the main menu includes the option "I want to see all techniques." This section lists all exercises in the order they appear in the TRT manual, allowing users to either select an individual exercise or follow a sequence similar to that of a facilitated TRT group. Additionally, the main menu features the option "I want other support," which provides a comprehensive list of care and support organizations, including phone numbers and links to relevant web pages.

Thermometer

The Thermometer is the first screen displayed in the app, prompting the young person to rate how they are feeling on a scale from 1 (low mood) to 10 (high mood) before accessing the app's content. If they report a score of 3 or lower, they are directed to a screen with instructions to call the national emergency number or contact a suicide prevention line if they are experiencing thoughts of self-harm. From there, they can choose to either proceed directly to the main menu or visit the "Other support" screen, which provides a list of care and support organizations. If a score above 3 is reported, the young person is encouraged to return to the Thermometer if their mood changes while using the app, before being directed to the main menu. The Thermometer remains accessible from every screen in the app via a tab bar (described below).

Safety Features

Techniques from the TRT manual that involve in vivo exposure or pose a risk of causing high levels of arousal or negative

XSL•FO
RenderX

emotions have been excluded from the app or modified to minimize these risks. Techniques in the app that still involve recalling traumatic memories include shortcuts to safety features. Young people who feel they need additional support beyond what the app provides can access the "Other support" screen from the main menu or the "Thermometer" if they score low. This screen offers an extensive list of links and contact details for various digital and in-person care and support services for young people, providing more personalized help. The app only lists these services; no referrals are made.

The "Safe Inner Place" technique helps young people take control of intrusive memories and relax their bodies during periods of high arousal. This visual imagery technique allows users to create positive mental images, enabling them to influence their imagination and emotions. It offers 2 options: reading or listening to the guide. The Safe Inner Place serves as a secure base that users can return to when experiencing high arousal. Easy access to this feature from the tab bar (described below) was seen as empowering, allowing young people to explore more challenging exercises with greater confidence.

Tab Bar

A tab bar with 4 icons is located at the bottom of each screen in the app. These icons provide quick access to the Main Menu, the Safe Inner Place technique, the Thermometer, and the previous screen. The tab bar is designed to facilitate navigation while also serving as a safety feature. If a young person experiences high physiological arousal or negative emotions while using the app, they can quickly access the Safe Inner Place relaxation technique or the Thermometer. If they indicate a low level of well-being, the Thermometer will present them with options for additional care and support.

Anxiety Reducing Exercises

From the main menu, selecting "I feel anxious" directs the young person to a submenu of techniques designed to help manage negative emotions related to trauma. In addition to the Safe Inner Place technique, 6 other techniques support emotional regulation and coping with arousal and intrusive thoughts. Among them, 2 video-based exercises—*Breathing Technique* and *Relaxation Exercise*—are specifically designed to promote relaxation and help the young person regulate physiological arousal.

Activity Scheduling is a text-based technique designed to encourage social and physical activity while providing a distraction from anxious thoughts and feelings. The exercise offers 3 options: receiving suggestions for meaningful activities, creating a personalized list of activities, or scheduling planned activities. The suggested activities are provided by youth advisors to ensure they are relevant to young people. Another text-based technique, *Worry Time*, encourages setting aside 10 minutes each day for intrusive thoughts and worries, helping to limit worry time and regain control over intrusive thoughts. *Sleep Advice* provides a list of tips to improve sleep hygiene for young people experiencing sleep difficulties or nightmares due to trauma. Finally, *Positive Self-Talk* helps young people recognize the connection between thoughts and feelings and learn to replace negative self-perceptions with positive ones. The exercise offers a choice between receiving suggested positive affirmations or creating personalized ones.

Techniques for Intrusion Control

Selecting the option "I have intrusive memories" from the main menu provides a choice of techniques to help manage auditory or visual intrusive memories. *Coping with Sound Memories* is a video-based exercise where the young person practices controlling intrusive auditory memories by mentally manipulating the sound in different ways. Coping with Visual Memories includes 2 exercises: *Screen Techniques* and *Clapping Technique.* Screen Techniques is a video-based exercise that, similar to the Auditory Intrusion Exercise, guides users in manipulating visual memories as if they were displayed on a TV screen or in the palm of their hand to regain control over them. The Clapping Technique, known as bilateral stimulation in the TRT manual—a term the coresearchers found too complex for the app—demonstrates how rhythmically tapping the knees while recalling intrusive images can help modify and regulate those memories.

Other People's Stories

The "Other People's Stories" page is accessible via a button in the main menu. It features a collection of fictional stories in which young people describe their personal experiences of trauma in writing. The purpose of these stories is to provide psychoeducation and normalize emotional and psychological reactions to trauma. The traumatic events depicted were suggested by the coresearchers, with plans to expand the content to include a more diverse range of experiences. At the time of usability testing, the available stories covered themes such as losing a parent, being bullied, and experiencing war. Each story is clearly labeled with a heading that indicates its subject matter.

Phase 2: Usability Testing

For the SUS, complete data were available from 10 participants, with scores ranging from 67.5 to 97.5 [32]. Based on existing reference values, a product must achieve an SUS score of at least 71 to be in the top half [32,36]. Nine out of the 10 completed SUS recordings had scores of 77.5 or higher.

Table 1 summarizes the identified usability issues along with the corresponding modifications. Some issues were related to technical preferences, such as the ability to rewind video and audio files to revisit missed content or replay important segments. Additionally, the option to enable subtitles was considered important for those who wished to use them. As a result, rewind and subtitle features were added to all audio and video recordings. While the video format was well received, some recordings were noted to have low quality, with slight flickering from background lighting. Some participants reported that this negatively impacted the overall user experience. To address this, all relevant videos were rerecorded in a professional studio, with the same young person demonstrating the techniques.

XSL•FO
RenderX

Table 1. Usability issues identified with the prototype of the Teaching Recovery Techniques app and the corresponding modifications.

Usability issues	Modifications to the app
Some of the video and audio recordings lacked the ability to rewind as well as the option for subtitles.	Rewind function and option for subtitles were added to all videos and recordings.
Users found the quality of some video and audio recordings to be low, affecting the overall user experience negatively.	The videos were rerecorded in a studio. The audios were rerecorded by the same narrator, while ensuring no interferences in the recording.
Users were bothered by difficulties related to the device, such as the screen going black during long audio and video recordings, and the lower tab bar being difficult to access with an iPhone.	Instructional text was added, directing the user to the device settings to enable the screen to stay on for an extended period. Adjustments were made so access to the tab bar is not affected by the type of device.
The *Other People's Stories* exercise, aimed at identification and normalization, was generally perceived as important and helpful, but the audio recordings were also perceived as stressful to listen to. It could also be perceived as pressurizing to hear from others who have managed to overcome their difficulties.	A decision was made to keep *Other People's Stories* in text format but exclude the audio recordings from the app. Additional information about the contents of the stories, as well as the purpose of the exercise, was added. A reference to other exercises in the app in case the user feels they would not benefit from reading other people's stories was also added.
Users reacted to the advice to "reach out to someone close to you" and expressed that some people don't have that option, which may be why they use the app.	A reference to the *Other Support* screen was added.
There was too little variety in the tips on sleep, activities, and positive self-talk.	More tips were added for each exercise.
The app only being available in Swedish.	The app was translated into more languages.

Usability issues related to device functionality were also identified. Some video and audio recordings lasted several minutes, and if participants remained passive during playback—without touching the screen—the display would time out, disrupting their ability to focus on the exercise. To address this, a reference was added to each technique's information, guiding users on how to adjust screen time-out settings on their devices. Another device-related issue involved the tab bar at the bottom of each screen, which was unnecessarily difficult to access on iPhones. Adjustments were made to ensure a consistent user experience of the tab bar across different devices.

Some aspects were identified as potentially incompatible with the reality of the target group. Participants responded to the recurring advice to "talk to a trusted person," noting that some users might not have this option—which may, in fact, be why they turn to the app. To address this, a clarification was added, stating that if the user cannot or does not want to talk to someone close to them, the "Other support" section in the app provides several recommendations for both in-person and digital support services.

Additionally, the option to listen to other people's stories, narrated by young people for identification purposes, was found to carry the risk of causing undesirable feelings. Hearing someone speak about their personal difficulties could be burdensome and might also create pressure for the listener to cope with their own situation. As a result, the audio option was removed, leaving only the text format. Additional information was added to clarify the content and purpose of the stories. A reference to other exercises in the app was also included, along with the following message: "If you feel that reading other people's stories may not be helpful for you, you can try these exercises instead."

Discussion

Principal Findings

This project was conducted with young people who had personal experiences of trauma, actively involving them as experts from an early stage. In line with the Lundy model, creating a safe and inclusive space for children to express their views proved essential from the outset and remained a priority throughout [17]. All activities, except for the usability tests, were held in a conference room at a centrally located hotel. The coresearchers expressed early on that they appreciated the setting, describing the space as cozy, the food and refreshments as enjoyable, and the environment as one where they felt comfortable expressing themselves.

Careful consideration was given to recruitment strategies to ensure that young people from diverse backgrounds and experiences were informed about the project. While previous participatory research projects have highlighted the challenges of recruiting young people [37], our experience was different, with several young individuals quickly expressing interest. The use of multiple recruitment arenas proved valuable, and further exploration of effective strategies for engaging young coresearchers is warranted.

With regard to voice [17], coresearchers were encouraged to express themselves in various formats. Some felt comfortable speaking freely in workshops, while others preferred responding to direct questions. One coresearcher expressed early on that writing down their thoughts during workshops would be easier than speaking in a group. This option was made available and encouraged for all participants. Even those who preferred verbal communication occasionally utilized written expression. A standing agreement ensured that coresearchers could share their thoughts through writing, drawings, or digital images, sometimes compiled into vision boards. Several coresearchers contributed their insights this way at least once during or after a workshop.

XSL•FO
RenderX

Likewise, both written and verbal information was consistently provided to accommodate different preferences.

Research has shown that participatory approaches with young people must be flexible and pragmatic, aligning with their preferences and opportunities. Participation should be offered at different levels and in various formats, allowing young people to decide how they engage [38]. In this project, coresearchers' preferences guided practical aspects such as venue selection, scheduling, and breaks, including refreshments (see Sarkadi et al [39] for more details). All workshops were held in the same easily accessible venue, ensuring convenience for participants using any mode of transport. To foster a safe and inclusive environment, coresearchers collaboratively developed guidelines outlining how they wished to work together, including common rules for respect and understanding. A key expectation, as expressed in these discussions, was that open conversations would be easier if participants were assured that no one would interrupt or dismiss their opinions.

Sufficient time was allocated to allow researchers to work cocreatively with the coresearchers, ensuring the flexibility required for this type of collaboration. Joint workshops were scheduled in the evenings and on weekends, allowing coresearchers to participate without missing school—an aspect they highlighted as important. Additionally, researchers remained accessible throughout the project via multiple communication channels, accommodating individual preferences. Early in the project, a discussion was held to establish mutually agreed-upon times for communication. The coresearchers generally preferred communicating in the evenings but understood that responses might sometimes take until the next day. This flexibility contributed to consistently high workshop attendance. When a coresearcher was unable to attend, they were offered an individual meeting or conversation with a researcher, as they expressed a strong interest in staying engaged with the project's progress.

While enabling expression is important, so is providing the option not to participate. Early in the project, the coresearchers emphasized that they sometimes have a lot going on in their lives, particularly with school. To minimize absences, proposed dates and times were communicated in advance, allowing for joint decisions on scheduling. Additionally, the youth advisors were reminded that they could choose which parts of the process they wanted—and were able—to be involved in.

Audience refers to whether and how young people's views are being listened to [17]. Each workshop was attended by a researcher (MT) and the research assistant responsible for app development (KE), ensuring a direct communication pathway from coresearcher idea generation to app implementation. Having 2 adults present also allowed for more effective engagement—while one (KE) took notes, the other (MT) could focus on facilitating discussions and actively listening. To demonstrate that their input was valued, ideas from previous workshops and independent work tasks were shared back with the coresearchers at each session. This process was further reinforced by sharing app prototypes developed based on decisions made in previous workshops. This aligns with the Influence component of the Lundy model [17], as it allows youth advisors to see their input being implemented in real time. Additionally, the project was presented at multiple scientific conferences and regularly reported to the funding organization. Preparations for these events, as well as feedback received, were always shared with the coresearchers to ensure transparency about how and where their contributions were being presented. Those who wished to be recognized were also listed as coauthors on conference poster presentations.

The influence of the coresearchers extended beyond app development to the usability testing process. When reflecting on their experience, they described it as "fun and easy to do." They emphasized the value of peer-led research, noting that having young people conduct the testing and recruit from their own friendship networks made participants more comfortable expressing their true opinions—both positive and negative. By contrast, they felt that if adult researchers had conducted the testing, participants might not have responded as openly. The coresearchers emphasized that their close age and shared experiences with the participants allowed them to relate more directly, fostering a sense of understanding and connection. As peers, they created a more relaxed and informal atmosphere, making it easier for participants to communicate openly and feel comfortable. These reflections align with existing literature, which suggests that having observers and users with similar backgrounds can reduce misinterpretation of data in usability testing [32]. Arguments have been made in favor of involving children and young people as coresearchers, as their firsthand experience of childhood provides an insider perspective and expertise that adults cannot fully replicate [16,40]. While adult researchers may be considered outsiders, young people are embedded in peer culture, allowing them to be closer to the subject of research [41]. In line with the coresearchers' reflections, peer-led data collection has been shown to generate nuanced and valuable insights that differ from those typically produced by adult researchers [41].

Methodological Considerations

This study took steps to ensure diversity among coresearchers, representing different genders, ages, and backgrounds. However, it cannot be assumed that they fully represent all young people who have experienced trauma. The recruitment of study participants by the coresearchers is largely a strength, as their diversity was reflected in the sample. One limitation of the study was the decision not to video record the usability testing sessions, which could have provided additional insights into participant behavior. However, combining a standardized instrument with interview questions after usability tests was a strength, as was the multistep analysis process, which involved 2 researchers when applicable.

Conclusions

This study demonstrates that young people with personal experiences of trauma can actively contribute as coresearchers in the development and usability testing of a trauma support app. A child rights framework can facilitate this process, ensuring alignment with Article 12 of the CRC. The young people's contributions to the app's design can be categorized into 3 key areas: mechanics, dynamics, and aesthetics. The design studio method proved to be an effective approach for

XSL•FO
RenderX

cocreation, allowing for both the generation and refinement of their ideas. No formal training was required for the young people to contribute as experts in the development process. However, relevant training was essential to equip them for independent data collection. In parallel with this project, work has been conducted on potential negative consequences of the app—referred to as dark logic—which has been reported elsewhere [42]. The next phase involves further testing the app in a larger population. Ongoing efforts are focused on piloting the app both with young people participating in a TRT group and as a stand-alone intervention.

Acknowledgments

The copyright of Teaching Recovery Techniques is held by the Children and War Foundation. The research team thanks the coresearchers who have been part of the project for all their great efforts. We also thank Antónia Tökés for assisting with the analysis and for the further work on the app. This research was funded by a grant from the Skandia Ideas for Life Foundation (principal investigator: GW).

Data Availability

Data generated for this study are not publicly available to protect the privacy of those involved. Requests to access the data should be directed to GW.

Disclaimer

No generative AI toold were used when writing this manuscript.

Authors' Contributions

MT led the workshops, performed the coresearcher training, analyzed usability findings, and wrote sections of the manuscript. KI and RA participated in the workshops, performed the usability testing, participated in analyzing the usability findings, and wrote sections of the manuscript. KE coled the workshops, built the app prototype, and wrote sections of the manuscript. GW supervised all project activities, discussed usability analysis findings, and wrote parts of the manuscript. All authors reviewed the manuscript.

Conflicts of Interest

None declared.

References

1. Radez J, Reardon T, Creswell C, Lawrence P, Evdoka-Burton G, Waite P. Why do children and adolescents (not) seek and access professional help for their mental health problems? A systematic review of quantitative and qualitative studies. Eur Child Adolesc Psychiatry 2021 Feb;30(2):183-211 [FREE Full text] [doi: 10.1007/s00787-019-01469-4] [Medline: 31965309]
2. Aguirre Velasco A, Cruz I, Billings J, Jimenez M, Rowe S. What are the barriers, facilitators and interventions targeting help-seeking behaviours for common mental health problems in adolescents? A systematic review. BMC Psychiatry 2020 Jun 11;20(1):293-292 [FREE Full text] [doi: 10.1186/s12888-020-02659-0] [Medline: 32527236]
3. Gulliver A, Griffiths K, Christensen H. Perceived barriers and facilitators to mental health help-seeking in young people: a systematic review. BMC Psychiatry 2010 Dec 30;10(1):a [FREE Full text] [doi: 10.1186/1471-244x-10-113]
4. Kuhn E, Owen J. Advances in PTSD treatment delivery: the role of digital technology in PTSD treatment. Curr Treat Options Psych 2020 Mar 27;7(2):88-102 [FREE Full text] [doi: 10.1007/s40501-020-00207-x]
5. Grist R, Porter J, Stallard P. Mental health mobile apps for preadolescents and adolescents: a systematic review. J Med Internet Res 2017 May 25;19(5):e176 [FREE Full text] [doi: 10.2196/jmir.7332] [Medline: 28546138]
6. Sander L, Schorndanner J, Terhorst Y, Spanhel K, Pryss R, Baumeister H, et al. 'Help for trauma from the app stores?' A systematic review and standardised rating of apps for post-traumatic stress disorder (PTSD). Eur J Psychotraumatol 2020;11(1):1701788 [FREE Full text] [doi: 10.1080/20008198.2019.1701788] [Medline: 32002136]
7. Yule W, Dyregrov A, Raundalen M, Smith P. Children and war: the work of the Children and War Foundation. Eur J Psychotraumatol 2013;4:18424 [FREE Full text] [doi: 10.3402/ejpt.v4i0.18424] [Medline: 23330058]
8. Davis R, Meiser-Stedman R, Afzal N, Devaney J, Halligan S, Lofthouse K, et al. Systematic review and meta-analysis: group-based interventions for treating posttraumatic stress symptoms in children and adolescents. J Am Acad Child Adolesc Psychiatry 2023 Nov;62(11):1217-1232 [FREE Full text] [doi: 10.1016/j.jaac.2023.02.013] [Medline: 36948393]
9. Pérez-Aronsson A, Thell M, Lampa E, Löfving SG, Tökés A, Torakai N, et al. Adaptation of the trauma group intervention 'Teaching Recovery Techniques' for online delivery: a participatory design and usability study. Internet Interv 2022 Dec;30:100589 [FREE Full text] [doi: 10.1016/j.invent.2022.100589] [Medline: 36411786]

XSL•FO
RenderX

10. Freire K, Pope R, Jeffrey K, Andrews K, Nott M, Bowman T. Engaging with children and adolescents: a systematic review of participatory methods and approaches in research informing the development of health resources and interventions. Adolescent Res Rev 2022 Feb 03;7(3):335-354 [FREE Full text] [doi: 10.1007/s40894-022-00181-w]
11. Liabo K, Roberts H. Coproduction and coproducing research with children and their parents. Arch Dis Child 2019 Dec 08;104(12):1134-1137. [doi: 10.1136/archdischild-2018-316387] [Medline: 31068317]
12. Thabrew H, Fleming T, Hetrick S, Merry S. Co-design of eHealth interventions with children and young people. Front Psychiatry 2018;9:481 [FREE Full text] [doi: 10.3389/fpsyt.2018.00481] [Medline: 30405450]
13. Shimmin C, Wittmeier KDM, Lavoie JG, Wicklund ED, Sibley KM. Moving towards a more inclusive patient and public involvement in health research paradigm: the incorporation of a trauma-informed intersectional analysis. BMC Health Serv Res 2017 Aug 07;17(1):539 [FREE Full text] [doi: 10.1186/s12913-017-2463-1] [Medline: 28784138]
14. Fløtten KJØ, Guerreiro A, Simonelli I, Solevåg AL, Aujoulat I. Adolescent and young adult patients as co-researchers: a scoping review. Health Expect 2021 Aug;24(4):1044-1055 [FREE Full text] [doi: 10.1111/hex.13266] [Medline: 33991369]
15. Convention on the Rights of the Child. UNICEF. 1989. URL: https://www.unicef.org/child-rights-convention/convention-text [accessed 2024-02-01]
16. Lundy L, McEvoy L. Children's rights and research processes: assisting children to (in)formed views. In: Children's Rights. London, UK: Routledge; 2017:525-540.
17. Lundy L. 'Voice' is not enough: conceptualising Article 12 of the United Nations Convention on the Rights of the Child. British Educational Res J 2013 Jan 02;33(6):927-942 [FREE Full text] [doi: 10.1080/01411920701657033]
18. Participation framework. Hub na nÓg - young voices in decision-making. 2021. URL: https://hubnanog.ie/participation-framework/ [accessed 2024-02-01]
19. Utsatta områden - polisens arbete. Swedish Police Authority. 2023. URL: https://polisen.se/om-polisen/polisens-arbete/utsatta-omraden/ [accessed 2024-02-01]
20. Så får barn och unga arbeta. Swedish Work Environment Authority. 2015. URL: https://www.av.se/globalassets/filer/publikationer/broschyrer/sa-far-barn-och-ungdomar-arbeta-broschyr-adi043.pdf [accessed 2024-02-01]
21. INVOLVE. Involving children and young people in research: top tips and essential key issues for researchers. INVOLVE. Southampton, UK: INVOLVE; 2019. URL: https://www.invo.org.uk/wp-content/uploads/2016/01/involvingcyp-tips-keyissues-January2016.pdf [accessed 2024-02-01]
22. Kaplan K. Facilitating an effective design studio workshop. Nielsen Norman Group. 2017. URL: https://www.nngroup.com/articles/facilitating-design-studio-workshop/ [accessed 2024-02-01]
23. Padlet. URL: https://padlet.com [accessed 2024-02-01]
24. Pandasuite. URL: https://pandasuite.com [accessed 2024-02-01]
25. Kellett M. How to Develop Children as Researchers. New York, NY: Sage Publications; 2005.
26. Kim C, Sheehy K, Kerawalla L. Developing Children as Researchers: A Practical Guide to Help Children Conduct Social Research. London, UK: Taylor & Francis; 2017.
27. Shier H. Children as researchers in Nicaragua: children's consultancy to transformative research. Global Studies of Childhood 2015 May 25;5(2):206-219 [FREE Full text] [doi: 10.1177/2043610615587798]
28. Baauw E, Markopoulous P. A comparison of think-aloud and post-task interview for usability testing with children. In: IDC '04: Proceedings of the 2004 Conference on Interaction Design and Children: Building a Community. 2004 Presented at: IDC '04: The 2004 Conference on Interaction Design and Children: Building a Community; June 1-3, 2004; Baltimore, MD p. 115-116. [doi: 10.1145/1017833.1017848]
29. Virzi R. Refining the test phase of usability evaluation: how many subjects is enough? Hum Factors 1992 Aug 01;34(4):457-468 [FREE Full text] [doi: 10.1177/001872089203400407]
30. Eccles D, Arsal G. The think aloud method: what is it and how do I use it? Qualitative Research in Sport, Exercise and Health 2017 May 26;9(4):514-531. [doi: 10.1080/2159676x.2017.1331501]
31. Brooke J. SUS: a "quick and dirty'usability scale. Usability evaluation in industry 1996;189(3):1-7.
32. Hertzum M. Usability Testing: A Practitioner's Guide to Evaluating the User Experience. San Rafael, CA: Morgan and Claypool; 2020:1-15.
33. Mentimeter. 2023. URL: https://www.mentimeter.com [accessed 2024-02-01]
34. Motta RW. Secondary trauma. Int J Emerg Ment Health 2008;10(4):291-298. [Medline: 19278145]
35. Browning H. Guidelines for Designing Effective Games as Clinical Interventions: Mechanics, Dynamics, Aesthetics, and Outcomes (MDAO) Framework. New York, NY: IGI Global; 2016:105-130.
36. Bangor A, Kortum P, Miller J. An empirical evaluation of the System Usability Scale. International Journal of Human-Computer Interaction 2008 Jul 30;24(6):574-594 [FREE Full text] [doi: 10.1080/10447310802205776]
37. Beresford P. Beyond the usual suspects: Research report. Shaping Our Lives. 2013. URL: https://shapingourlives.org.uk/wp-content/uploads/2021/08/BTUS-Report.pdf [accessed 2024-02-01]
38. Rouncefield-Swales A, Harris J, Carter B, Bray L, Bewley T, Martin R. Children and young people's contributions to public involvement and engagement activities in health-related research: a scoping review. PLoS One 2021;16(6):e0252774 [FREE Full text] [doi: 10.1371/journal.pone.0252774] [Medline: 34106978]

XSL•FO
RenderX

39. Sarkadi A, Thell M, Fängström K, Dahlberg A, Fäldt A, Pérez-Aronsson A, et al. Are we ready to really hear the voices of those concerned? Lessons learned from listening to and involving children in child and family psychology research. Clin Child Fam Psychol Rev 2023 Dec;26(4):994-1007 [FREE Full text] [doi: 10.1007/s10567-023-00453-4] [Medline: 37700107]
40. Bradbury-Jones C, Taylor J. Engaging with children as co-researchers: challenges, counter-challenges and solutions. International Journal of Social Research Methodology 2013 Dec 03;18(2):161-173 [FREE Full text] [doi: 10.1080/13645579.2013.864589]
41. Kellett M. Rethinking Children and Research: Attitudes in Contemporary Society. New York, NY: Bloomsbury Publishing; Apr 15, 2010.
42. Warner G, Edvardsson K, Thell M. Look before you leap: theorising harmful consequences ahead of a pilot trial, the case of app-based trauma support for youth. 2022 Presented at: The 13th EUSPR Conference; September 29 and 30, 2022; Tallinn, Estonia.

Abbreviations

CRC: the Convention on the Rights of the Child
MDA: Mechanics, Dynamics, and Aesthetics
PTSD: posttraumatic stress disorder
REDCap: Research Electronic Data Capture
SUS: System Usability Scale
TRT: Teaching Recovery Techniques

Edited by A Mavragani; submitted 26.02.24; peer-reviewed by C Tsai; comments to author 20.01.25; revised version received 24.01.25; accepted 13.02.25; published 18.03.25.

Please cite as:
Thell M, Edvardsson K, Aljeshy R, Ibrahim K, Warner G
A Trauma Support App for Young People: Co-design and Usability Study
JMIR Form Res 2025;9:e57789
URL: https://formative.jmir.org/2025/1/e57789
doi:10.2196/57789
PMID:

©Maria Thell, Kerstin Edvardsson, Reem Aljeshy, Kalid Ibrahim, Georgina Warner. Originally published in JMIR Formative Research (https://formative.jmir.org), 18.03.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Kinetic Pattern Recognition in Home-Based Knee Rehabilitation Using Machine Learning Clustering Methods on the Slider Digital Physiotherapy Device: Prospective Observational Study

Clement Twumasi[1], BSc, MPhil, PhD; Mikail Aktas[2*], MBBS, MRES; Nicholas Santoni[3*], BSc

[1]Nuffield Department of Medicine, Experimental Medicine Division, University of Oxford, Oxford, United Kingdom

[2]Department of Bioengineering, Imperial College London, South Kensington, London, United Kingdom

[3]Department of Biology, University of North Carolina, Greensboro, NC, United States

[*]these authors contributed equally

Corresponding Author:
Clement Twumasi, BSc, MPhil, PhD
Nuffield Department of Medicine
Experimental Medicine Division
University of Oxford
Peter Medawar Building for Pathogen Research
South Parks Road
Oxford, OX1 3SY
United Kingdom
Phone: 44 7411484454
Email: clement.twumasi@ndm.ox.ac.uk

Abstract

Background: Recent advancements in rehabilitation sciences have progressively used computational techniques to improve diagnostic and treatment approaches. However, the analysis of high-dimensional, time-dependent data continues to pose a significant problem. Prior research has used clustering techniques on rehabilitation data to identify movement patterns and forecast recovery outcomes. Nonetheless, these initiatives have not yet used force or motion datasets obtained outside a clinical setting, thereby limiting the capacity for therapeutic decisions. Biomechanical data analysis has demonstrated considerable potential in bridging these gaps and improving clinical decision-making in rehabilitation settings.

Objective: This study presents a comprehensive clustering analysis of multidimensional movement datasets captured using a novel home exercise device, the “Slider”. The aim is to identify clinically relevant movement patterns and provide answers to open research questions for the first time to inform personalized rehabilitation protocols, predict individual recovery trajectories, and assess the risks of potential postoperative complications.

Methods: High-dimensional, time-dependent, bilateral knee kinetic datasets were independently analyzed from 32 participants using four unsupervised clustering techniques: k-means, hierarchical clustering, partition around medoids, and CLARA (Clustering Large Applications). The data comprised force, laser-measured distance, and optical tracker coordinates from lower limb activities. The optimal clusters identified through the unsupervised clustering methods were further evaluated and compared using silhouette analysis to quantify their performance. Key determinants of cluster membership were assessed, including demographic factors (eg, gender, BMI, and age) and pain levels, by using a logistic regression model with analysis of covariance adjustment.

Results: Three distinct, time-varying movement patterns or clusters were identified for each knee. Hierarchical clustering performed best for the right knee datasets (with an average silhouette score of 0.637), while CLARA was the most effective for the left knee datasets (with an average silhouette score of 0.598). Key predictors of the movement cluster membership were discovered for both knees. BMI was the most influential determinant of cluster membership for the right knee, where higher BMI decreased the odds of cluster-2 membership (odds ratio [OR] 0.95, 95% CI 0.94-0.96; $P<.001$) but increased the odds for cluster-3 assignment relative to cluster 1 (OR 1.05, 95% CI 1.03-1.06; $P<.001$). For the left knee, all predictors of cluster-2 membership were significant ($.001\leq P\leq .008$), whereas only BMI ($P=.81$) could not predict the likelihood of an individual belonging to cluster 3 compared to cluster 1. Gender was the strongest determinant for the left knee, with male participants significantly likely to belong to cluster 3 (OR 3.52, 95% CI 2.91-4.27; $P<.001$).

XSL•FO
RenderX

Conclusions: These kinetic patterns offer significant insights for creating personalized rehabilitation procedures, potentially improving patient outcomes. These findings underscore the efficacy of unsupervised clustering techniques in the analysis of biomechanical data for clinical rehabilitation applications.

(JMIR Form Res 2025;9:e69150) doi:10.2196/69150

KEYWORDS

machine learning; cluster analysis; force measurement; knee replacement; musculoskeletal; physical therapy; Slider device; knee osteoarthritis; digital health; telerehabilitation

Introduction

Background of the Study

The challenge of optimizing knee rehabilitation exercises following arthroplasty or injury is critical, given the high prevalence of knee surgeries and the need for effective recovery strategies. According to the *American Joint Replacement Registry 2022 Annual Report*, over 2.5 million knee and hip arthroplasty cases were documented, constituting a significant proportion of knee replacements. The quality and effectiveness of the ensuing rehabilitation, which has historically relied on patient self-reporting, subjective assessments, and poor compliance, is crucial to the success of these procedures. Such methods can lead to suboptimal outcomes, as they fail to capture biomechanical data that could inform treatment planning and postoperative complications [1,2].

Recent developments in sensor technology have made it possible to collect detailed kinetic data during physical therapy exercises. This presents a chance to transition from subjective assessments to more objective, data-driven methods. This change is especially significant in home-based rehabilitation, where it is crucial to collect valuable data consistently. The newly developed Slider device [3] exemplifies this technological advancement by capturing multidimensional movement data, including force and spatial coordinates, directly from patients' lower limbs during exercise. The Slider device is the first rehabilitation device capable of remotely collecting force and motion data in a patient's home. This capability enhances the ability to monitor patients in real time and to produce more informative datasets that capture the dynamic interaction of force and movement in a real-world, home-based setting [4].

A recent study has extensively examined the efficacy of the Slider device for prehabilitation prior to total knee replacement surgery [3]. The study revealed that the device can assist patients in performing autonomous physiotherapy prior to surgery in nonclinical environments. The research involved 17 patients awaiting knee replacement surgery at a UK National Health Service hospital. Their findings confirmed the device's practicality and user-friendliness for patient-initiated exercises, with no recorded safety issues or adverse occurrences [3]. Principal findings revealed improvements in user involvement and physical preparedness. The study advised that subsequent research should involve a bigger and more diverse participant cohort to comprehensively validate its findings regarding the device. In order to enhance our understanding of this new device, additional studies may be required to investigate the multidimensional kinetic and kinematic data from the Slider using unsupervised clustering methods for the first time across a comparatively larger and diverse population. This exploration is necessary to identify potential movement patterns that could offer further insights into workout performance and recovery. The insights gained from these computational methods may improve comprehension of patient heterogeneity, enabling customized rehabilitation programs that more effectively address individual needs and improve outcomes for diverse groups.

The user datasets from this new data-collection device represent a significant contribution to the field, as they provide a new level of detail in understanding how kinetic patterns evolve during rehabilitation outside of a clinical environment. Despite these advancements, analyzing such high-dimensional, time-dependent data remains a complex challenge. Previous studies have applied clustering methods to other rehabilitation data to identify movement patterns and predict recovery outcomes. However, these efforts have yet to use force or motion data captured outside of a clinical environment, limiting their ability to inform clinical decisions [5,6].

Data were collected from healthy volunteers performing lying flexions and extensions of their knees, including relevant demographic characteristics and pain levels. K-means, hierarchical clustering, partition around medoids (PAM), and CLARA (Clustering Large Applications) were applied to identify distinct movement patterns (based on force and displacement measurements as primary outcomes). Performance evaluations were conducted using the classical silhouette analysis to determine the most appropriate clustering method. Then, post hoc analyses were performed with multinomial logistic regression to explore potential demographic predictors of cluster membership based on the best models for the right and left knee datasets, respectively. This study contributes to the understanding of rehabilitation dynamics beyond the clinical setting. The insights gained from the uniquely identified data clusters may therefore enable clinicians to make more informed, data-driven decisions, ultimately enhancing the precision and effectiveness of rehabilitation programs. The primary aim of this study is to advance the understanding of kinetic patterns during knee rehabilitation by applying and comparing four unsupervised clustering algorithms to new datasets collected for the first time using the Slider device.

Paper Structure and Research Questions

This study is structured into five main sections. The *Introduction* section provides an overview of the study's background, outlines the paper's organization, and highlights its primary contributions. The *Methods* section outlines the empirical kinetic data used in the study, participant recruitment, study design, and ethical considerations, followed by a concise summary of

XSL•FO
RenderX

the clustering methodologies and additional modeling considerations. The *Results* section presents the findings derived from the statistical analysis of the experimental data. Finally, the final section (*Discussion*) offers a discussion of the key findings, draws conclusions, and suggests directions for future research. With the help of experimental bilateral knee kinetic datasets obtained from the newly developed home-based rehabilitation device (Slider), this study seeks to address the following open research questions:

- What distinct movement patterns can be identified from multidimensional kinetic data during home physiotherapy?
- Can real-time clustering of movement data provide meaningful feedback for personalized physiotherapy?
- How effective is cluster analysis in distinguishing between varying levels of movement efficiency and recovery states?
- How do kinetic movement patterns evolve over the course of physiotherapy, as detected through the new Slider device?
- Can the multidimensional movement data collected remotely provide reliable insights for monitoring physiotherapy outcomes?

Methods

Data Collection, Recruitment, and Study Design

Continuously observed, time-dependent, and high-dimensional datasets were collected from both knees of 32 participants (totaling 64 time-varying datasets with large data points) using Slider during lying flexion and extension exercises. Participants were not required to have undergone any prior treatment. Demographic characteristics such as gender (21 male and 11 female participants), ethnicity (30 White participants and 2 mixed White and Asian participants), laterality (29 right and 3 left), age (mean 47.3, SD 2.11; range 21-77 years), height (cm), and weight (kg) were recorded. Each participant's BMI was calculated from their height and weight measurements (mean 26.2, SD 4.99; range 18.3-40.6 kg/m^2). Pain levels or scores were quantified using a standard analog pain scale, ranging from a score of 0=no pain to 10=worst imaginable pain.

Participants were not excluded from the study due to deformities or complications from prior surgeries or therapy treatments. The study was conducted without supervision, as it was intended to be performed remotely in the participants' homes due to the inherently mobile nature of the Slider device. Consequently, the study was deemed low risk and noninvasive. Each participant completed 20 cycles, at a self-selected speed with the device recording force (in Newtons) and X-Y displacement (in meters) measured with an optical tracker. The data output was time stamped (in seconds). Figure 1 is a time-series visualization of observed data of the main unadjusted outcome variables for clustering of a study participant. All statistical tests, with a significance level set at $P<.05$, and cluster analyses were performed in R statistical software (version 4.3.1; R Core Team) [7].

The participants comprised the entire existing pool of volunteers registered with the Slider device manufacturer. These individuals were invited to participate based on their prior engagement with the device, ensuring familiarity with its use. Recruitment procedures followed ethical guidelines, with all participants providing informed consent before data collection. This study follows a machine learning (ML) predictive modeling design in biomedical research, using unsupervised clustering techniques to analyze high-dimensional kinetic data and identify distinct movement patterns in home-based knee rehabilitation.

XSL•FO
RenderX

Figure 1. Time-series plots illustrating the unadjusted observed kinetic data, including force and X-Y displacements, recorded at the (A-C) right and (D-F) left knees of a study participant (ID: 1 MA). These plots provide an overview of the raw movement patterns captured by the Slider device, highlighting temporal variations in force exertion and spatial displacement during rehabilitation exercises.

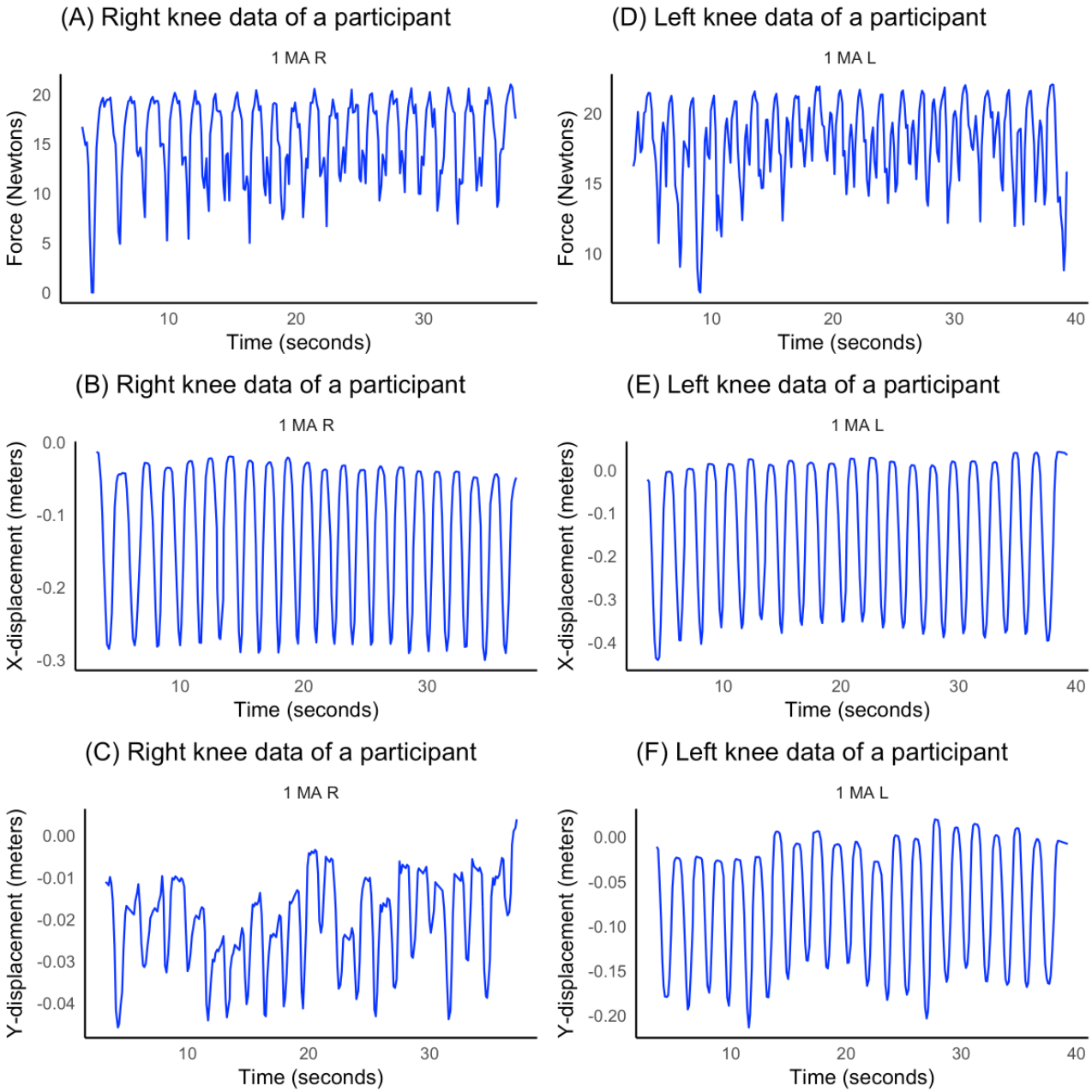

Ethical Considerations

This low-risk and noninvasive study was classified as a service evaluation in accordance with the guidelines from the UK Health Research Authority. Consultation with the UK Research and Innovation Medical Research Council using Integrated Research Application System (348336) confirmed that the study adhered to all relevant ethical standards, and therefore, formal ethical committee approval was not required. All participants provided informed consent before participating, with explicit consent obtained for both data collection and subsequent analysis. To protect privacy and confidentiality, all study data were fully anonymized before analysis, ensuring no personally identifiable information was included. No financial compensation or other incentives were provided for participation, as the study was conducted on a voluntary basis. Finally, no identifiable images or personal data from participants are included in the study dataset or any supplementary materials.

Data Preprocessing

Data preprocessing is a crucial step in medical data analysis, especially for high-dimensional datasets with missing or incomplete information. It ensures data integrity, accuracy, and usability by handling missing values (where applicable), normalizing distributions, and reducing dimensionality, ultimately improving predictive models and clustering techniques performance [8]. Consequently, a two-stage data preprocessing method was used on the time-dependent datasets before conducting cluster analysis. In the first stage, potential outliers in the time-series data for each participant were automatically detected and corrected using the *tsoutliers* R package [9], based on the force and X-Y displacement data. In

XSL•FO
RenderX

the second stage, the cleaned time-series data underwent decomposition using locally estimated scatterplot smoothing (decomposition). This process aided in detrending the series for each participant, making them more stationary, and removing cyclical patterns to mitigate cyclical effects, thereby enhancing the accuracy of the clustering analysis results. Because outlier correction and locally estimated scatterplot smoothing decomposition effectively standardize the time-series data, decreasing noise and maintaining comparability between participants by addressing both extreme values and nonstationary trends, this two-stage approach is sufficient without additional data normalization. Reliable clustering outcomes are therefore possible since any residual variation in the data represents significant variations in participant movement patterns rather than artifacts from data scale or trend discrepancies.

Unsupervised Clustering Methods

Overview

In this study, the conventional hierarchical clustering and three widely used partitional clustering algorithms, namely: k-means [10,11], PAM [10,12], and CLARA [10], were used and compared to respectively divide the preprocessed bilateral knee datasets into an optimal number of clusters (represented by $k>1$). This partitioning ensures that data points representing kinetic features within the same cluster exhibit greater similarity to each other than to those in other clusters. These clustering methods have their respective algorithmic processes for distinctly grouping data, each with advantages and disadvantages regarding its robustness, scalability, and applicability related to the data structure, dataset size, and noise sensitivity. These justify the choice of models for this study and highlight the need to identify the best clustering algorithms for the right and left knee kinetic datasets, which were continuously observed over time for each volunteer. A summary of the underlying clustering methods used in this study is presented below.

K-Means Clustering

K-means is a centroid-based clustering algorithm that partitions a set of n data points $\{x_1, x_2, ..., x_n\} \in R^d$ into k clusters $\{C_1, C_2, ..., C_k\}$ by minimizing the within-cluster sum of squares objective function [10]:

where μ_i is the mean of the points in cluster C_i. The algorithm starts with k initial centroids, assigns each point to the nearest centroid or cluster center, and updates centroids by computing the mean of assigned points. This process repeats until convergence, typically when cluster assignments no longer change. K-means is computationally efficient, with a time complexity of $O(n \cdot k \cdot d \cdot t)$, where t is the number of iterations; but is sensitive to the initial choice of centroids.

Hierarchical Clustering

Hierarchical clustering constructs a nested tree of clusters by either agglomerative (bottom-up) or divisive (top-down) methods. The agglomerative approach starts with each data point as a single cluster and iteratively merges the two closest clusters based on a chosen linkage criterion. For example, in single linkage, the distance between clusters A and B is defined as [13]:

The agglomerative algorithm first computes the distance matrix for all pairs of points, then merges the closest clusters based on the chosen linkage criterion, and finally, updates the distance matrix, repeating this process until all points form a single cluster. The divisive algorithm starts with all data points in a single cluster and recursively splits them into smaller clusters until each data point forms its own cluster. At each step, it maximizes a splitting criterion such as the between-cluster variance denoted by:

where A and B are the resulting clusters from the split, and their respective centroids are denoted by μ_A and μ_B. Hierarchical clustering is effective for detecting nested structures, but it has a time complexity of either $O(n^2\log_n)$ for the agglomerative method (making it effective for small to medium-sized datasets) or $O(2^n)$ for the divisive approach (making it less effective for large datasets).

PAM

PAM is a medoid-based clustering algorithm that extends k-means by selecting actual data points (medoids) as cluster centers, minimizing the sum of dissimilarities between points and their nearest medoid. The objective is to find a set of k medoids $\{m_1, m_2, ..., m_k\}$ that minimizes [10]:

where $d(x, m_i)$ is the dissimilarity (eg, Euclidean distance) between a data point x and the medoid m_i of cluster C_i. PAM is robust to noise and outliers but has a higher computational cost of $O(k[n-k]^2)$, which can limit its scalability for large datasets. Different variants and extensions of PAM also exist [10].

CLARA

CLARA is an extension of PAM designed for large datasets. It draws multiple samples of the data and applies PAM to each sample, selecting the set of medoids that minimizes the clustering cost over the entire dataset [12]. CLARA reduces the computational burden of PAM by using random sampling, maintaining a complexity of $O(k|S|^2 + k[n - |S|])$, where $|S|$ is the sample size. CLARA scales better to large datasets, but its performance depends on the quality and representativeness of the random samples [10].

Other Statistical Modeling Considerations and Performance Evaluation

To respectively determine the optimal number of unique clusters in the data for both knees, the *NbClust* R Package [14] was used. This package evaluates over 30 different cluster identification schemes or indices, with the optimal cluster size determined automatically using a majority rule. After clustering, the performance of the four ML approaches for the right and left

XSL•FO
RenderX

knee datasets was evaluated and compared using silhouette analyses, which validate clustering methods [15]. Silhouette coefficients or scores, which range from –1 (indicating incorrect clustering) to +1 (indicating highly dense or correct clustering), were calculated for each cluster identified across the four data partitioning methods. Based on their distributions, appropriate pooled estimates of these scores were used to compare the models' performance for the right and left knee data clustering, respectively. Principal component analysis (PCA) was adopted to visualize the clustered high-dimensional data $\{x_1, x_2, ..., x_n\} \in R^d$ in a reduced dimensional space.

While the primary cluster analysis using force and displacement outcome data can potentially categorize movement patterns into distinct groups in the right and left knees, it remains essential to identify additional demographic factors that may influence these patterns and their clinical relevance [16]. Thus, a multinomial logistic regression model with analysis of covariance adjustment was used to investigate whether demographic variables such as age, gender, BMI, and pain scores could serve as significant predictors of cluster membership. This approach allows for a more comprehensive understanding of the determinants of cluster assignment by examining how these demographic factors, when adjusted for kinetic measurements, contribute to the likelihood of a participant belonging to a specific cluster at a particular time (since cluster membership is time-varying given the time-dependent kinetic data structure). Incorporating these variables enhances the ability to tailor personalized rehabilitation protocols by considering both biomechanical and individual patient characteristics, ultimately supporting more precise and effective clinical decision-making.

Using an analysis of covariance–based logistic model with adjustment for the actual observed force and X-Y displacement measurements, as well as temporal effects, the model is mathematically defined as follows:

for each cluster category $1<k\leq K$ relative to a baseline category (cluster 1), where K denotes the total number of uniquely identified clusters. $P(C_i=k)$ represents the probability of the ith individual being in cluster k at time $T\geq 0$, and the regression coefficients: β_{1k}, β_{2k}, β_{3k}, and β_{4k} represent the respective effects of the main demographic predictors (age, gender, BMI, and pain score) on the log-odds of being in cluster $k\geq 2$ versus the baseline cluster 1, adjusted for force (F), X-Y displacements (denoted by X and Y) and time of measurement (T).

The coefficients γ_{1k}, γ_{2k}, γ_{3k}, and γ_{4k} control for the influence of force, X-Y displacements, and time of measurement, respectively. The model was fitted with the help of the *nnet* R package [17]. Finally, the odds ratios (ORs) were estimated, which quantify the magnitude of the discriminatory effect of each of the main demographic predictors by exponentiating the estimated log-odds regression coefficients from $\exp(\beta_{gk})$ for $1\leq g\leq 4$ and $1<k\leq K$. The right and left knee datasets were respectively split into 80% training sets (for model fitting) and 20% testing sets for cross-validation to estimate the prediction accuracy from the resulting confusion matrix and the area under the receiver operating characteristic curve statistic.

Results

Cluster Analysis at Both Knees

Summary of Key Findings

Three clusters were optimally identified for both knees, each representing a unique kinetic pattern. K-means, hierarchical, PAM, and CLARA clustering were subsequently performed based on these predetermined clusters. Figures 2 and 3 show a PCA visualization of the clustered datasets in a lower-dimensional space for both knees. The silhouette analysis (Figure 4) revealed that hierarchical clustering was relatively found to be the most effective for the right knee kinetic datasets (with an average silhouette score of 0.637) followed by k-means (with an average silhouette score of 0.612), while CLARA proved to be the best-performing method for the left knee datasets (with an average silhouette score of 0.598) followed by k-means (with an average silhouette score of 0.588). It can be inferred from these results that the right knee datasets resulted in a better clustering result than the left knee datasets. Figure 5 presents a scatterplot of the adjusted kinetic features (which have been detrended and cycle-adjusted) in a 3D space for the right and left knees, stratified by the identified clusters, based on the best-performing, knee-specific clustering algorithms.

XSL•FO
RenderX

Figure 2. Principal component analysis visualizations comparing the three identified movement clusters at the right knee across the four clustering methods: (A) k-means, (B) hierarchical, (C) PAM, and (D) CLARA. These visualizations are based on adjusted kinetic features (force and X-Y displacements) and illustrate how different clustering techniques distinguish movement patterns in the multidimensional dataset. CLARA: Clustering Large Applications; PAM: partition around medoids.

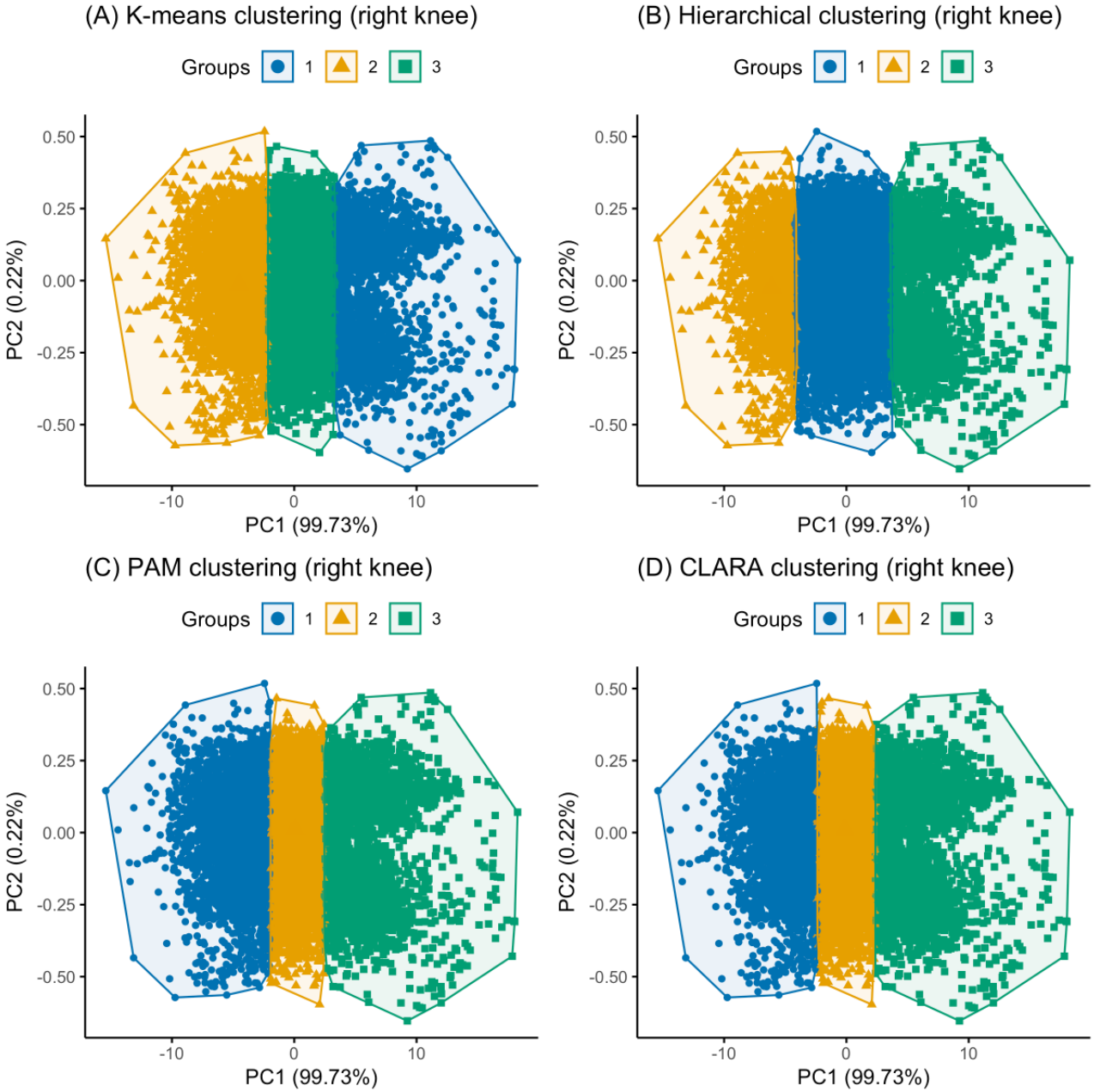

XSL•FO
RenderX

Figure 3. Principal component analysis visualizations comparing the three identified movement clusters at the left knee across the four clustering methods: (A) k-means, (B) hierarchical, (C) PAM, and (D) CLARA. These plots provide insights into the separability of kinetic movement patterns based on adjusted force and spatial displacement data. CLARA: Clustering Large Applications; PAM: partition around medoids.

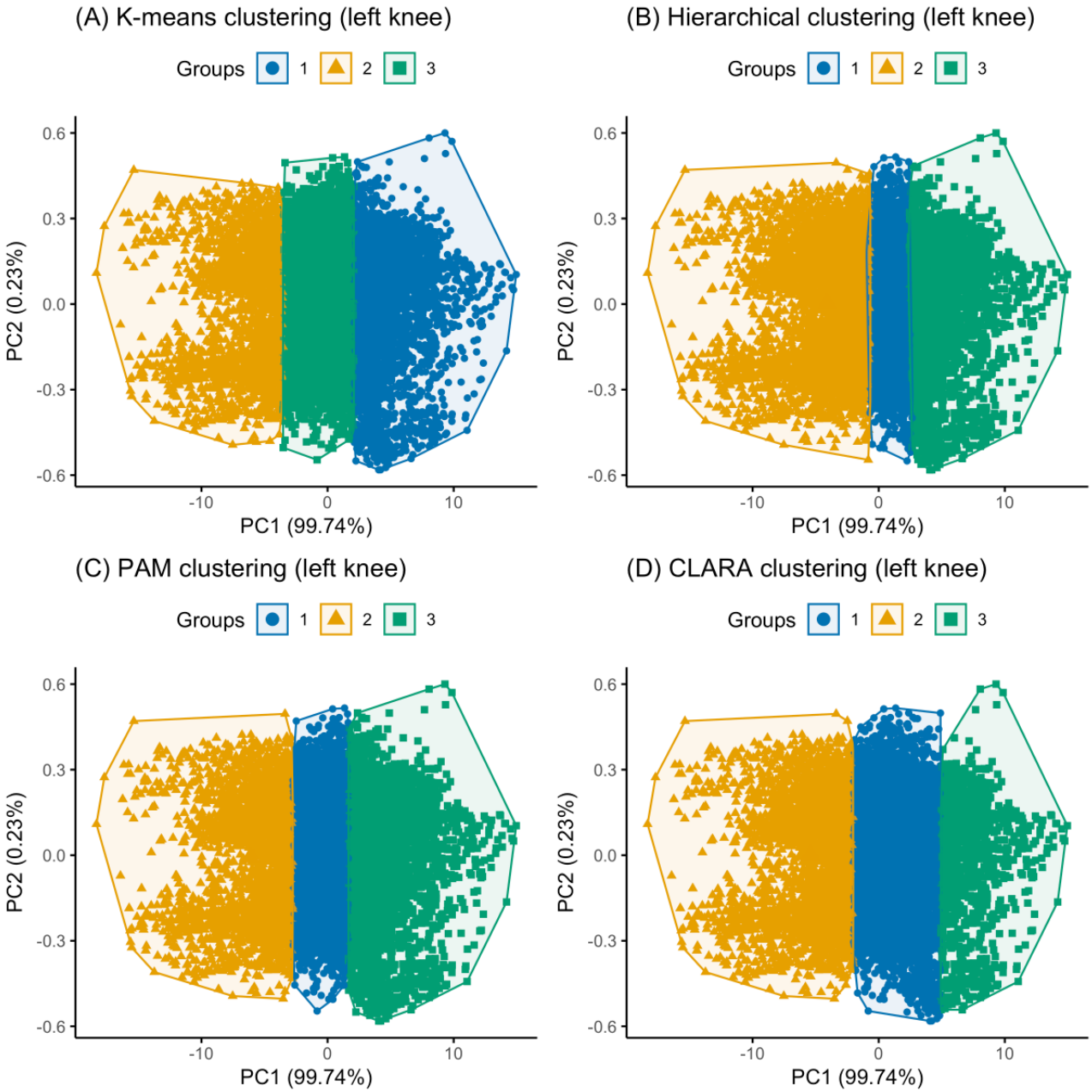

XSL•FO
RenderX

Figure 4. Distributional plot comparing estimated silhouette coefficients across the four clustering methods (k-means, hierarchical, PAM, and CLARA) at the (A) right and (B) left knees. The plot presents both cluster-specific and pooled estimates, highlighting the relative performance of each clustering method in distinguishing movement patterns from kinetic data at both knees. CLARA: Clustering Large Applications; PAM: partition around medoids.

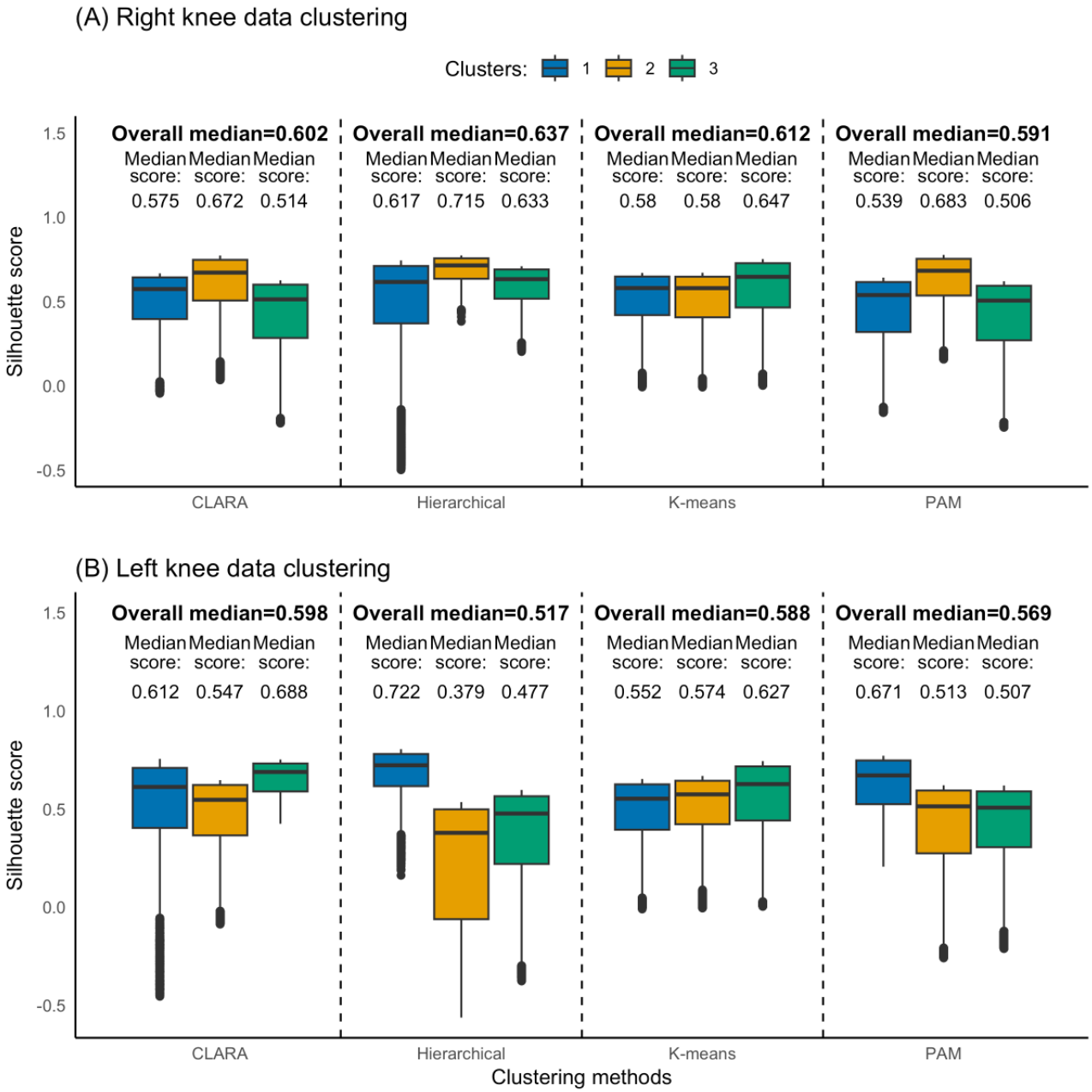

XSL•FO
RenderX

Figure 5. 3D scatterplot depicting the clustering of adjusted force measurements (in Newtons) against adjusted X-Y displacements (in meters) for the (A) right and (B) left knees. The clustering is based on the most optimal knee-specific methods identified through silhouette analysis—hierarchical clustering for the right knee and CLARA for the left knee—demonstrating the spatial distribution of movement patterns within the dataset. CLARA: Clustering Large Applications.

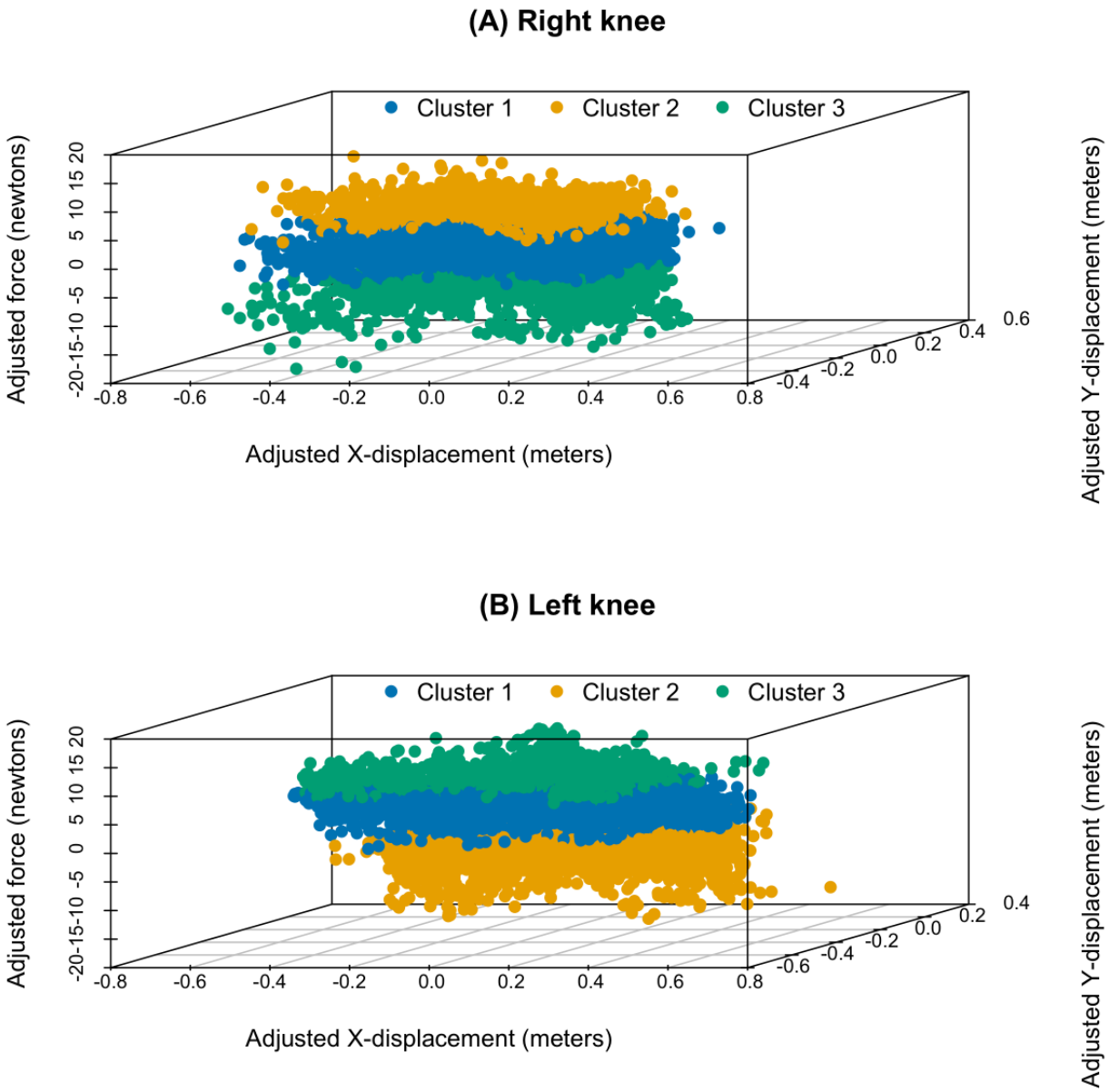

To assess whether significant differences existed in the distribution of multivariate kinetic features (force and X-Y displacements) across the three clusters for both knees, a multivariate Kruskal-Wallis (MKW) test [18] was used. This test was chosen due to the multivariate nonnormality of the bilateral knee datasets (Henze-Zirkler multivariate normality test: E=246.25, P<.001 for the right-knee datasets; E=219.34, P<.001 for left-knee datasets). The results indicated that at least one of the clusters significantly differed in distribution for both the right (MKW: W_6=9204.35; P<.001) and left knees (MKW: W_6=9973.54; P<.001). Subsequent univariate Kruskal-Wallis tests were performed for each kinetic feature at both knees, coupled with Dunn multiple comparison tests (Figure 6). There were significant differences in the distribution of the kinetic features between the three clusters at both knees, suggesting the identified knee-specific clusters were uniquely distinct. Figure 7 shows the frequency of times cluster membership varied temporally across the 32 study participants. The proportion of times participants were assigned to cluster 1, cluster 2, and cluster 3 at the right knee was 72.7%, 12.3%, and 15%, respectively. For the left knee, the proportion of times participants were assigned to cluster 1, cluster 2, and cluster 3 were 65.1%, 25.4%, and 9.5%, respectively.

XSL•FO
RenderX

Figure 6. Comparative distribution plots of adjusted kinetic features (force and X-Y displacements) across the three k-means–identified clusters at the (A-C) right and (D-F) left knees. Statistical differences between clusters were assessed using Kruskal-Wallis tests followed by Dunn multiple comparison tests, with significance levels indicated (***P<.001; **P<.01; *P<.05). These results highlight variations in movement characteristics among identified rehabilitation patterns.

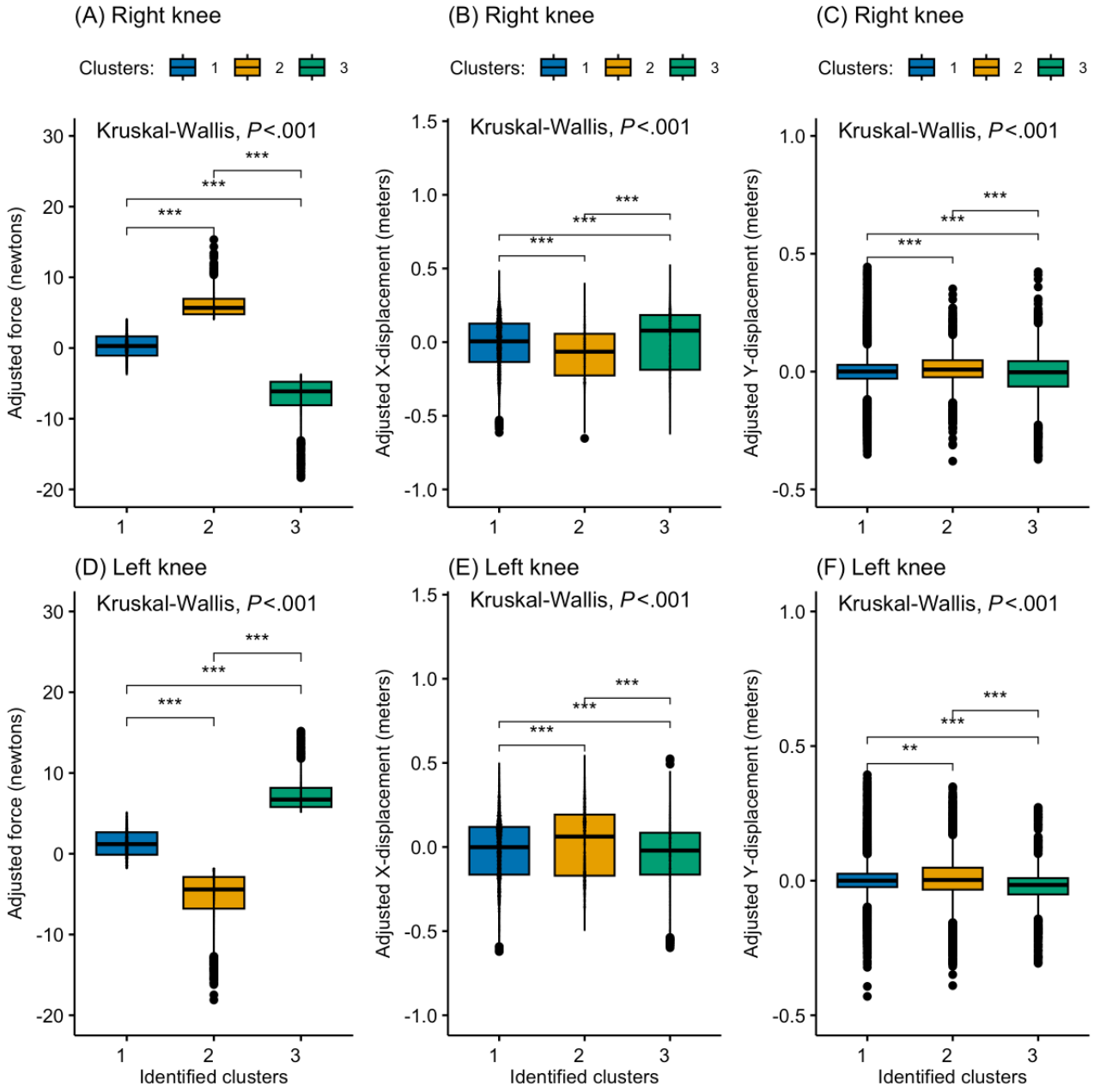

XSL•FO
RenderX

Figure 7. Slope graph visualizing temporal changes in cluster membership across the 32 study participants at the (A) right and (B) left knees. This plot captures how individuals transitioned between movement clusters over time due to the time-dependent structure of kinetic data collected from both knees, reflecting the dynamic nature of rehabilitation progress.

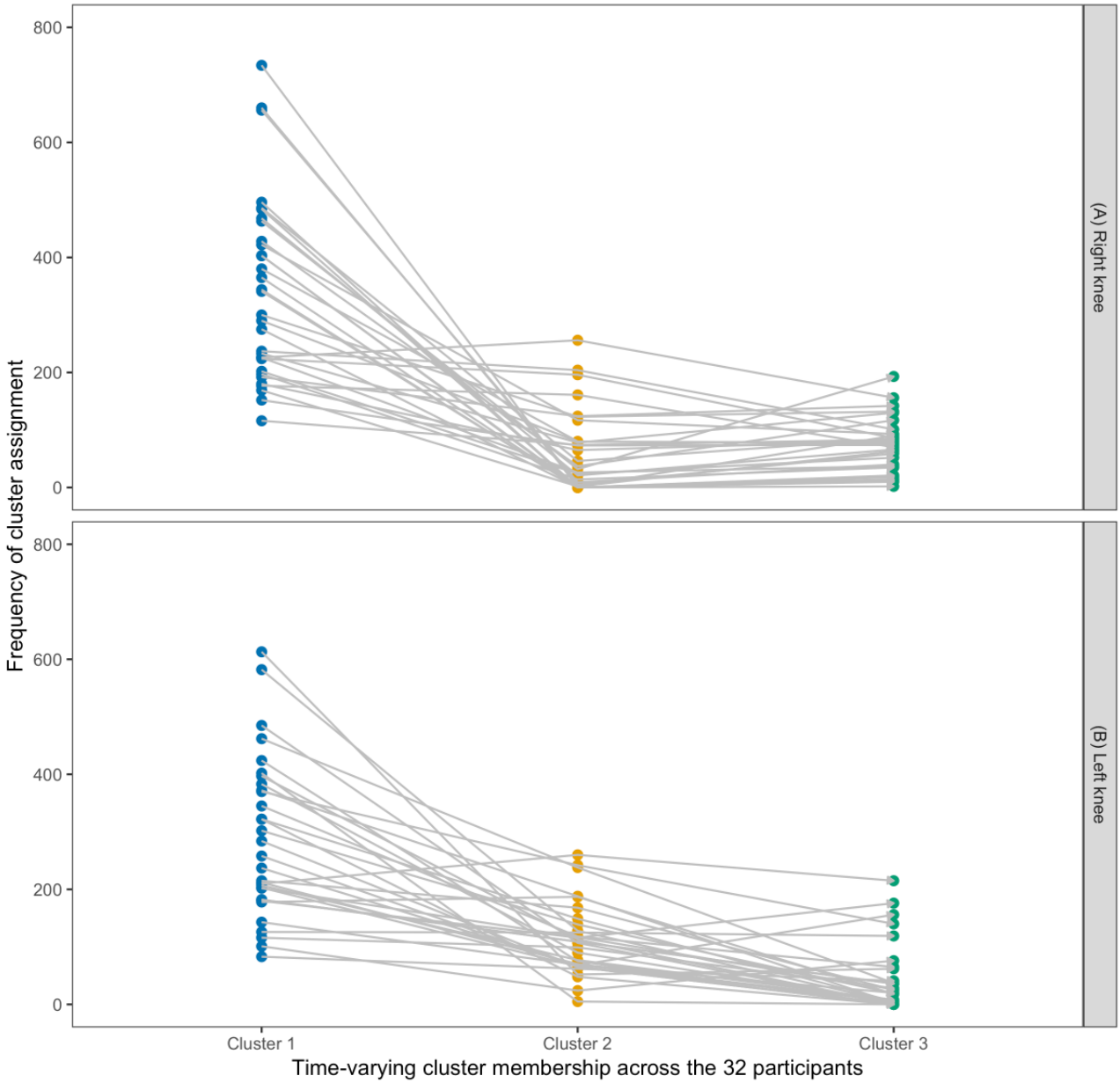

There was a reverse observation of the relative positions of clusters 2 and 3 in the right and left knees of the study participants (Figure 5). Possible practical kinetic interpretations of the clustered data with respect to force measurements across X-Y displacements, given this reverse kinetic pattern in right and left knees, could be understood as follows.

Cluster 1 (Middle Position)

This cluster suggests intermediate force levels and displacements. In both knees, movements in this cluster might involve moderate force exertion and a moderate range of motion in the X-Y plane. These could correspond to exercises aimed at strengthening muscles and improving joint mobility without excessive stress.

Cluster 2 (Bottom Position—Left Knee; Top Position—Right Knee)

This cluster likely represents movements with lower force exertion and smaller displacements in both X and Y directions. In the left knee, these movements may represent exercises performed with less force and a limited range of motion. It could represent muscle weakening or joint pathology. In the right knee, however, these movements may involve higher force exertion but still within a constrained range of motion. The lower range of motion could represent joint stiffness due to arthritis. This cluster may represent the least physically fit group of individuals. In the right knee, it could represent moderate-intensity exercises, while in the left knee, it might involve active movements with moderate intensity.

XSL•FO
RenderX

Cluster 3 (Top Position—Left Knee; Bottom Position—Right Knee)

This cluster indicates higher force levels and larger displacements. In the left knee, movements in this cluster likely involve high-force exertion combined with extensive movements in the X-Y plane. These could correspond to activities requiring significant strength and flexibility, such as high-intensity exercises. The right knee, however, shows lower force exertion but still with extensive movements, suggesting exercises focused more on flexibility, agility, and possibly lower muscle activities. This cluster may represent the most physically fit group of individuals.

Determinants of the Identified Clusters at Both Knees

The multinomial logistic regression analysis identified several significant demographic predictors of cluster membership for both right and left knee datasets (Table 1). For the right knee, being male was associated with significantly lower odds of belonging to cluster 2 compared to cluster 1 (OR 0.30, 95% CI 0.26-0.33) and to cluster 3 (OR 0.55, 95% CI 0.47-0.63). BMI also emerged as a significant predictor, with higher BMI decreasing the odds of being in cluster 2 (OR 0.95, 95% CI 0.94-0.96) but increasing the odds for cluster-3 membership (OR 1.05, 95% CI 1.03-1.06). Age was a significant predictor for cluster 3 (OR 1.02, 95% CI 1.01-1.03) but not for cluster 2 (P=.95). For the left knee, we found that all the demographic determinants of belonging to cluster 2 relative to cluster 1 were significant (.001≤P≤.008), whereas only BMI (P=.81) could not predict the likelihood of an individual belonging to cluster 3 compared to cluster 1. Gender emerged as the strongest determinant for the left knee data, with male participants being significantly more likely to belong to cluster 3 (OR 3.52, 95% CI 2.91-4.27; P<.001).

XSL•FO
RenderX

Table 1. Estimated OR[a] and their 95% CIs for the demographic determinants of cluster membership, derived from a fitted multinomial regression model (N=32 participants)[b,c].

Clusters and variables	OR (95% CI)	SE	*P* value
Right knee[d]			
Cluster 2 (reference: cluster 1)			
Male (reference: female)	0.30 (0.26-0.33)	0.061	<.001
BMI (kg/m^2)	0.95 (0.94-0.96)	0.006	<.001
Age	0.99 (0.98-1.01)	0.002	.95
Pain score	1.03 (0.99-1.06)	0.015	.05
Cluster 3 (reference: cluster 1)			
Male (reference: female)	0.55 (0.47-0.63)	0.077	<.001
BMI (kg/m^2)	1.05 (1.03-1.06)	0.007	<.001
Age	1.02 (1.01-1.03)	0.002	<.001
Pain score	1.03 (0.99-1.06)	0.016	.11
Left knee[e]			
Cluster 2 (reference: cluster 1)			
Male (reference: female)	0.63 (0.55-0.72)	0.069	<.001
BMI (kg/m^2)	1.10 (1.09-1.12)	0.007	<.001
Age	1.04 (1.03-1.05)	0.002	<.001
Pain score	0.95 (0.91-0.99)	0.020	.008
Cluster 3 (reference: cluster 1)			
Male (reference: female)	3.52 (2.91-4.27)	0.100	<.001
BMI (kg/m^2)	1.01 (0.98-1.02)	0.009	.81
Age	0.98 (0.97-0.99)	0.002	<.001
Pain score	1.05 (1.01-1.09)	0.020	.01

[a]OR: odds ratio.

[b]These results quantify the influence of demographic factors such as BMI, age, gender, and pain levels on the likelihood of belonging to specific movement clusters.

[c]Each of the multinomial logistic regression models for both knees' data was adjusted for the primary kinetic features (force and X-Y displacements) and temporal effect.

[d]For right knee measurement data: percentage accuracy=77.82% (95% CI 76.29%-79.3%); area under the receiver operating characteristic curve=0.778.

[e]For the left knee measurement data: percentage accuracy=76.24% (95% CI 74.62%-77.81%); area under the receiver operating characteristic curve=0.762.

Discussion

Results Key Insights Enabled by the Slider Device for Home-Based Rehabilitation Monitoring

This study explored further an innovative method for knee rehabilitation using the Slider device, a novel tool that enables the remote acquisition of multidimensional kinetic data, including force and spatial coordinates, within a home setting. By facilitating objective real-time monitoring of rehabilitation progress, Slider represents a significant advancement in personalized physiotherapy. A previous study has already validated its feasibility and safety for empowering patients to perform presurgery physiotherapy exercises independently, eliminating the need for a clinical environment [3]. However, no study has yet examined and analyzed the multidimensional bilateral kinetic and kinematic datasets generated by users of this new device. This analysis aimed to uncover potential patterns that could transform raw data into actionable insights, enabling therapists to deliver data-driven, patient-specific interventions that may optimize rehabilitation outcomes. This study used and compared four unsupervised clustering methods to find three unique time-varying movement patterns or clusters based on the observed time-dependent, multidimensional datasets from the right and left knees of the study participants.

The application of multiple clustering methods provides a comparative perspective on clustering performance, reinforcing the robustness of the findings. We have identified these movement patterns outside of a clinical setting for the first time. Significant differences in the distributions of force and displacement between clusters show how differently participants move. Demographic factors like gender, BMI, age, and pain

XSL•FO
RenderX

score also predicted membership in a cluster. These findings provide significant potential for individualized rehabilitation programs that enhance therapist accuracy and results. By leveraging ML and kinetic data collection, this study further demonstrates how at-home rehabilitation can be made more autonomous, reducing patient dependency on in-clinic physiotherapy sessions. The ability to remotely track movement progress offers a scalable solution for long-term rehabilitation, minimizing the need for frequent clinical visits.

Consequently, this study has effectively addressed the five open research questions by examining multidimensional kinetic data obtained during home-based physiotherapy. Using four clustering algorithms, the study found three different, time-varying, kinetic movement patterns for both knees. These patterns may represent different levels of fitness and recovery, answering the first question about how to find movement patterns in multidimensional kinetic data. The study answered the second question by showing that real-time clustering of movement data could help physiotherapists better understand how to meet the needs of each patient by showing how the clusters changed in response to force and displacement data. To answer the third question, the study confirmed that clustering methods, especially hierarchical clustering for the right knee and CLARA for the left knee, could tell the difference between different levels of movement efficiency and recovery states.

The mathematical modeling revealed that over time, participants progressed from erratic to more coordinated movement patterns, reflecting the evolution of the natural rhythm of motion during exercise, thus answering the fourth question on how movement patterns evolve. The insights gained from movement clustering suggest potential future advancements in physiotherapy technology, particularly in enhancing real-time feedback mechanisms of rehabilitation devices. These findings could contribute to the development of more adaptive and intelligent physiotherapy tools, improving their application in both research and clinical settings. Finally, the Slider device provided reliable, multidimensional kinetic data, enabling accurate cluster identification and analysis, thereby confirming its potential for remotely monitoring physiotherapy outcomes and answering the fifth question.

Practical Implications of the Study

Identification of Movement Patterns

Clusters that differentiate based on movement patterns (such as the range and direction of motion) can help clinicians identify abnormal movement behaviors in participants. This is particularly useful in postoperative rehabilitation or in monitoring the progress of participants with musculoskeletal disorders. For a theoretical application in a clinical setting, cluster characterized by limited range and lower force output might indicate participants who are either in the early stages of recovery from injury or surgery or who are experiencing complications or poor recovery. Hitherto, such kinetic evaluations have only been possible if participants attended sessions in a gait laboratory or clinic [5]. Slider allows these evaluations to be done in a participant's home using data from exercises done by the participants over the entire duration of their rehabilitation program. A possible weakness of this study would be that it is intended to account for movement of the knee but does not account for limitations in movement due to the flexibility of the ankle or hip. Such variables could have impacted the data collected and changed which cluster the participant fell into, however, no participants that participated reported having such issues before conducting the exercises, so such variables should not have been considered significant during data collection.

Tailored Rehabilitation Programs

Understanding which cluster a participant's movement data fall into allows clinicians to tailor rehabilitation programs more effectively. By customizing programs based on the typical movement characteristics of each cluster, clinicians can optimize recovery times and outcomes. By understanding which cluster a participant's movement data fall into, therapists can customize exercises that target specific movement deficiencies [6]. For a practical application, participants in a cluster showing high force and extreme movement ranges might benefit from a different set of exercises compared to those in a cluster with conservative movement patterns, possibly due to differences in pain tolerance, injury severity, or recovery stage. Remote data collection can enable clinicians to monitor progress in real time, offering greater flexibility in treatment planning and providing objective data for faster, more efficient feedback. This benefits both participants and care providers.

Prediction of Recovery Trajectories

The data clusters could help predict different recovery trajectories by identifying which participants will likely recover faster based on similarities in movement patterns. Participants showing gradual movement range and force improvements over time by progressing through clusters may indicate a positive recovery trajectory. Clusters may serve as predictive markers for rehabilitation outcomes, allowing for proactive adjustments in treatment plans [19].

Customization of Assistive Devices

Insights from clustering can inform the customization of assistive devices to better support individual participant needs based on the commonalities in movement patterns, unique movement limitations, or capabilities observed within each cluster. Participants in a cluster that shows significantly altered gait patterns might benefit from specific types of orthotic supports designed to compensate for those specific deviations [20].

Enhanced Understanding of Pathophysiology

Exploring cluster formation further can improve understanding of the pathophysiology of conditions, aiding diagnosis and treatment. Differentiating clusters by movement patterns can help identify causes of pain or dysfunction, such as muscular versus joint issues, and analyzing force and movement data can reveal significant correlations. This is significant as it allows therapists to obtain data regarding multiple variables at once during one exercise instead of having to test the functionality of the knee with the participant in the same room using multiple tests to achieve a similar outcome [21].

XSL•FO
RenderX

Limitations of the Study

This study has some limitations that must be acknowledged from both statistical and clinical viewpoints. From a statistical perspective, the sample size of 32 participants, although adequate for preliminary temporal exploratory analysis, is not sufficient for generalizing the findings to larger and more diverse groups. A larger cohort would provide a more robust assessment of the observed kinetic patterns and enhance the reliability of the used clustering techniques. Furthermore, whereas unsupervised clustering effectively identified distinct movement patterns, the study design was deficient in longitudinal follow-up to assess the direct correlation between these patterns and long-term rehabilitation outcomes. Clinically, while the Slider device offers an option for home-based rehabilitation, the absence of real-time clinician supervision may lead to variability in movement execution and data quality, thereby impacting cluster accuracy. The predictive efficiency of the identified clusters in guiding treatment strategies remains uncertain without rigorous validation through randomized controlled trials or extended clinical monitoring.

Conclusions

Three distinct, kinetic, time-varying clusters were identified for both the right and left knees, representing unique movement patterns, based on measurement data from the Slider exercise device. Data clustering performance was found to be knee specific, with hierarchical and CLARA clustering methods producing better results at the right and left knees, respectively. The kinetic patterns at the right and left knees were significantly different. Gender, BMI, age, and pain score were identified as significant demographic predictors of cluster membership, with varying impacts on cluster assignment for the right and left knees. Future research should also explore other clustering algorithms, such as density-based clustering (DBSCAN) and Gaussian mixture models. Additionally, the impact of various preprocessing techniques should be investigated, especially for time-dependent and cyclical data. Techniques such as Fourier transforms, wavelet transforms, and PCA for dimensionality reduction may also enhance the analysis of movement patterns.

Furthermore, future studies should develop additional kinetic features, such as velocity, acceleration, and jerk, to provide a more comprehensive understanding of movement patterns and improve clustering accuracy. Expanding this study to include a more diverse participant population would help assess the generalizability of the findings, refine personalized rehabilitation protocols, and address the limitations of this study. Finally, further longitudinal and trial studies could be conducted to correlate identified clusters with clinical outcomes, such as recovery trajectories and therapy efficacy, to validate the practical significance of the clustering results. Further case studies are currently being planned, which will expand the scope of data collected and study participants. It is expected that with further data collection, it will be easier to identify more accurately what subgroup knee replacement participants fall into, as well as what range of movement and force recorded can be considered optimal based on certain demographics. This method appears to be effective at remotely measuring movement and force for knee joint movement. Future projections for Slider device development will include artificial intelligence–assisted projections and therapy progress analysis to streamline the process for physicians.

Acknowledgments

The authors express their heartfelt gratitude to the volunteers for their valuable time and consent, which made this study possible. This study was conducted independently and did not receive external funding.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to requirements of intellectual property protection but are available from the corresponding author upon reasonable request. The well-documented R scripts used for data analysis are publicly available for reproducibility of results [22].

Authors' Contributions

Each author contributed significantly to the study's conception and design and provided approval for the final manuscript. MA and NJS played a key role in the collection of empirical data, with support from AI Rehab LLC, Atlanta, Georgia, who provided the devices. CT contributed by developing the mathematical methodologies and performing all statistical analyses. All authors actively participated in drafting and revising the manuscript.

Conflicts of Interest

The Slider Device and Slider digital app were generously provided at no cost by AI Rehab LLC, Atlanta, Georgia. The authors did not receive any financial support from AI Rehab LLC and have no financial ties to the company. AI Rehab LLC also provided complimentary training on the device's use but had no role in the design, execution, data analysis, or interpretation of the results of this study. The authors serve as unpaid advisors to AI Rehab LLC.

References

XSL•FO
RenderX

1. Shan L, Shan B, Suzuki A, Nouh F, Saxena A. Intermediate and long-term quality of life after total knee replacement: a systematic review and meta-analysis. J Bone Joint Surg Am 2015 Jan 21;97(2):156-168. [doi: 10.2106/JBJS.M.00372] [Medline: 25609443]
2. Bloomfield RA, Broberg JS, Williams HA, Lanting BA, McIsaac KA, Teeter MG. Machine learning and wearable sensors at preoperative assessments: functional recovery prediction to set realistic expectations for knee replacements. Med Eng Phys 2021 Mar;89:14-21. [doi: 10.1016/j.medengphy.2020.12.007] [Medline: 33608121]
3. Islam R, Gooch D, Karlakki S, Price B. A device for prehabilitation of total knee replacement surgery (Slider): usability study. JMIR Form Res 2023 Dec 18;7:e48055 [FREE Full text] [doi: 10.2196/48055] [Medline: 38109191]
4. Whatling GM, Biggs PR, Wilson C, Holt CA. Assessing functional recovery following total knee replacement surgery using objective classification of level gait data and patient-reported outcome measures. Clin Biomech (Bristol) 2022 May;95:105625 [FREE Full text] [doi: 10.1016/j.clinbiomech.2022.105625] [Medline: 35429691]
5. Young-Shand KL, Roy PC, Dunbar MJ, Abidi SSR, Astephen Wilson JL. Gait biomechanics phenotypes among total knee arthroplasty candidates by machine learning cluster analysis. J Orthop Res 2023 Feb;41(2):335-344 [FREE Full text] [doi: 10.1002/jor.25363] [Medline: 35538599]
6. Lange JK, DiSegna ST, Yang W, Li W, Franklin PD. Using cluster analysis to identify patient factors linked to differential functional gains after total knee arthroplasty. J Arthroplasty 2020 Jan;35(1):121-126.e6. [doi: 10.1016/j.arth.2019.08.039] [Medline: 31530461]
7. R: a language and environment for statistical computing. R Core Team. 2023. URL: http://www.R-project.org/ [accessed 2024-09-28]
8. Hassler AP, Menasalvas E, García-García FJ, Rodríguez-Mañas L, Holzinger A. Importance of medical data preprocessing in predictive modeling and risk factor discovery for the frailty syndrome. BMC Med Inform Decis Making 2019 Feb 18;19(1):33 [FREE Full text] [doi: 10.1186/s12911-019-0747-6] [Medline: 30777059]
9. López-De-Lacalle J. R package Tsoutliers. The Comprehensive R Archive Network. 2019. URL: https://cran.hafro.is/web/packages/tsoutliers/vignettes/tsoutliers-intro.pdf [accessed 2024-06-19]
10. Schubert E, Rousseeuw PJ. Fast and eager k-medoids clustering: O(k) runtime improvement of the PAM, CLARA, and CLARANS algorithms. Inf Syst 2021 Nov;101:101804 [FREE Full text] [doi: 10.1016/J.IS.2021.101804]
11. Shi N, Liu X, Guan Y. Research on k-means clustering algorithm: an improved k-means clustering algorithm. New York: IEEE; 2010 Presented at: Third International Symposium on Intelligent Information Technology and Security Informatics; April 2-4, 2010; Jian, China p. 63-67 URL: https://ieeexplore.ieee.org/document/5453745 [doi: 10.1109/IITSI.2010.74]
12. Park H, Jun C. A simple and fast algorithm for K-medoids clustering. In: Expert Systems with Applications. Cham, Switzerland: Springer; 2009:3336-3341.
13. Murtagh F, Contreras P. Algorithms for hierarchical clustering: an overview. WIREs Data Min Knowl 2011 Dec 07;2(1):86-97 [FREE Full text] [doi: 10.1002/widm.53]
14. Charrad M, Ghazzali N, Boiteau V, Niknafs A. NbClust: an R package for determining the relevant number of clusters in a data set. J Stat Software 2014;61(6):1-36 [FREE Full text] [doi: 10.18637/jss.v061.i06]
15. Shutaywi M, Kachouie NN. Silhouette analysis for performance evaluation in machine learning with applications to clustering. Entropy (Basel) 2021 Jun 16;23(6):759 [FREE Full text] [doi: 10.3390/e23060759] [Medline: 34208552]
16. Bucci CM, Legnani WE, Armentano RL. Clustering of cardiovascular risk factors highlighted the coronary artery calcium as a strong clinical discriminator. Health Technol 2016 Sep 30;6(3):159-165 [FREE Full text] [doi: 10.1007/s12553-016-0139-1]
17. Ripley B, Venables W. Package nnet. nginx. 2016. URL: https://staff.fmi.uvt.ro/~daniela.zaharie/dm2019/RO/lab/lab3/biblio/nnet.pdf [accessed 2024-06-20]
18. He F, Mazumdar S, Tang G, Bhatia T, Anderson SJ, Dew MA, et al. Non-parametric MANOVA approaches for non-normal multivariate outcomes with missing values. Commun Stat Theory Methods 2017;46(14):7188-7200 [FREE Full text] [doi: 10.1080/03610926.2016.1146767] [Medline: 29416225]
19. Giacomozzi C, Martelli F, Nagel A, Schmiegel A, Rosenbaum D. Cluster analysis to classify gait alterations in rheumatoid arthritis using peak pressure curves. Gait Posture 2009 Feb;29(2):220-224 [FREE Full text] [doi: 10.1016/j.gaitpost.2008.08.004] [Medline: 18815041]
20. Scivoletto G, Galli G, Torre M, Molinari M, Pazzaglia M. The overlooked outcome measure for spinal cord injury: use of assistive devices. Front Neurol 2019;10:272 [FREE Full text] [doi: 10.3389/fneur.2019.00272]
21. Loftus TJ, Shickel B, Balch JA, Tighe PJ, Abbott KL, Fazzone B, et al. Phenotype clustering in health care: a narrative review for clinicians. Front Artif Intell 2022;5:842306 [FREE Full text] [doi: 10.3389/frai.2022.842306] [Medline: 36034597]
22. Twumasi C. Kinetic pattern recognition in novel home-based physiotherapy: cluster analysis of multidimensional movement data from the Slider digital rehabilitation device (GitHub R codes). GitHub. 2024. URL: https://github.com/twumasiclement/Cluster-Analysis-of-Multidimensional-Movement-Data-from-the-Slider- [accessed 2025-02-17]

Abbreviations

CLARA: Clustering Large Applications

XSL•FO
RenderX

ML: machine learning
MKW: multivariate Kruskal-Wallis
OR: odds ratio
PAM: partition around medoids
PCA: principal component analysis

Edited by A Mavragani; submitted 23.11.24; peer-reviewed by A Holzinger; comments to author 15.02.25; revised version received 17.02.25; accepted 25.02.25; published 18.03.25.

Please cite as:
Twumasi C, Aktas M, Santoni N
Kinetic Pattern Recognition in Home-Based Knee Rehabilitation Using Machine Learning Clustering Methods on the Slider Digital Physiotherapy Device: Prospective Observational Study
JMIR Form Res 2025;9:e69150
URL: https://formative.jmir.org/2025/1/e69150
doi:10.2196/69150
PMID:

©Clement Twumasi, Mikail Aktas, Nicholas Santoni. Originally published in JMIR Formative Research (https://formative.jmir.org), 18.03.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Smartwatch-Based Ecological Momentary Assessment for High-Temporal-Density, Longitudinal Measurement of Alcohol Use (AlcoWatch): Feasibility Evaluation

Chris Stone[1,2], BSc, RSci; Sally Adams[3], PhD; Robyn E Wootton[2,4,5,6], PhD; Andy Skinner[1], PhD

[1]Integrative Cancer Epidemiology Programme, Bristol Medical School, University of Bristol, Bristol, United Kingdom
[2]School of Psychological Science, University of Bristol, Bristol, United Kingdom
[3]School of Psychology, University of Birmingham, Birmingham, United Kingdom
[4]Nic Waals Institute, Lovisenberg Diaconal Hospital, Oslo, Norway
[5]Medical Research Council Integrative Epidemiology Unit, Bristol Medical School, University of Bristol, Bristol, United Kingdom
[6]PsychGen Centre of Genetic Epidemiology and Mental Health, Norwegian Institute of Public Health, Oslo, Norway

Corresponding Author:
Chris Stone, BSc, RSci
School of Psychological Science
University of Bristol
12a Priory Road
Bristol, BS8 1TU
United Kingdom
Phone: 44 07983 317748
Fax: 44 0117 928 8588
Email: chris.stone@bristol.ac.uk

Related Article:

This is a corrected version. See correction statement: https://formative.jmir.org/2025/1/e82194

Abstract

Background: Ecological momentary assessment methods have recently been adapted for use on smartwatches. One particular class of these methods, developed to minimize participant burden and maximize engagement and compliance, is referred to as microinteraction-based ecological momentary assessment (µEMA).

Objective: This study explores the feasibility of using these smartwatch-based µEMA methods to capture longitudinal, high-temporal-density self-report data about alcohol consumption in a nonclinical population selected to represent high- and low-socioeconomic position (SEP) groups.

Methods: A total of 32 participants from the Avon Longitudinal Study of Parents and Children (13 high and 19 low SEP) wore a smartwatch running a custom-developed µEMA app for 3 months between October 2019 and June 2020. Every day over a 12-week period, participants were asked 5 times a day about any alcoholic drinks they had consumed in the previous 2 hours, and the context in which they were consumed. They were also asked if they had missed recording any alcoholic drinks the day before. As a comparison, participants also completed fortnightly online diaries of alcohol consumed using the Timeline Followback (TLFB) method. At the end of the study, participants completed a semistructured interview about their experiences.

Results: The compliance rate for all participants who started the study for the smartwatch µEMA method decreased from around 70% in week 1 to 45% in week 12, compared with the online TLFB method which was flatter at around 50% over the 12 weeks. The compliance for all participants still active for the smartwatch µEMA method was much flatter, around 70% for the whole 12 weeks, while for the online TLFB method, it varied between 50% and 80% over the same period. The completion rate for the smartwatch µEMA method varied around 80% across the 12 weeks. Within high- and low-SEP groups there was considerable variation in compliance and completion at each week of the study for both methods. However, almost all point estimates for both smartwatch µEMA and online TLFB indicated lower levels of engagement for low-SEP participants. All participants scored “experiences of using” the 2 methods equally highly, with “willingness to use again” slightly higher for smartwatch µEMA.

XSL•FO
RenderX

Conclusions: Our findings demonstrate the acceptability and potential utility of smartwatch µEMA methods for capturing data on alcohol consumption. These methods have the benefits of capturing higher-temporal-density longitudinal data on alcohol consumption, promoting greater participant engagement with less missing data, and potentially being less susceptible to recall errors than established methods such as TLFB. Future studies should explore the factors impacting participant attrition (the biggest reason for reduced engagement), latency issues, and the validity of alcohol data captured with these methods. The consistent pattern of lower engagement among low-SEP participants than high-SEP participants indicates that further work is warranted to explore the impact and causes of these differences.

(JMIR Form Res 2025;9:e63184) doi:10.2196/63184

KEYWORDS

smartwatch; ecological momentary assessment; µEMA; alcohol; ALSPAC

Introduction

Ecological momentary assessment (EMA) refers to a class of methods in which people are asked to answer questions about specific aspects of their health, feelings, and behaviors as they go about their normal lives. As individuals typically complete these assessments throughout the day, rather than completing a survey and having to remember details from days or weeks ago, recall errors are significantly reduced [1]. One domain in which EMA methods have been used extensively is the study of modifiable health behaviors, including physical activity, diet, tobacco smoking, and alcohol consumption [2]. The benefits of EMA methods are of particular interest for the measurement of alcohol consumption. Commonly used questionnaires and diary-based methods are known to suffer from systematic biases toward underreporting, capturing somewhere between 20% and 60% of the true level of consumption [3,4], with retrospective recall bias being a major factor [5]. Furthermore, recent studies have highlighted the importance of both longitudinal measures of alcohol consumption to help us explore relationships between alcohol consumption and conditions such as depression [6], and high-temporal-resolution measures of alcohol consumption to enable exploration of factors such as subjective responses (eg, feeling of stimulation, feeling of sedation, liking the effects of alcohol, wanting more alcohol) in heavy drinkers in natural environments [7]. There is, therefore, a need for EMA methods that can be used to capture high-temporal-density measures of alcohol consumption over extended periods.

In response to the increased adoption of mobile and wearable technologies, EMA methods have been adapted for use on smartphones, and more recently, smartwatches. For smartwatch-based EMA methods, a particular area of focus has been reducing participant burden, so that even more frequent sampling over longer periods can be tolerated. Smartwatches are worn on the wrist so they are never beyond reach, which reduces the time to access the device and respond to prompts. Further, because they are worn against the body, they enable the use of potentially more discrete haptic prompts for responses. One smartwatch EMA method aiming to minimize participant burden is microinteraction-based EMA (µEMA), which uses simple, brief questions with a limited set of answers that can be responded to with a single tap multiple times a day [8]. When compared with smartphone-based EMA over a 4-week period, smartwatch-based µEMA had better performance in terms of compliance rates (defined as the percentage of all scheduled prompts to which participants responded, regardless of the success of prompt delivery: 82% vs 64%) and completion rates (defined as the percentage of scheduled prompts actually delivered to the participant to which they responded: 92% vs 67%) [8].

Having demonstrated high levels of engagement in these initial 4-week explorations of smartwatch-based µEMA, the approach was subsequently used to capture self-report measures over longer periods in various settings. Beukenhorst and colleagues [9] explored the use of smartwatch-based µEMA methods in participants with osteoarthritis of the knee to capture self-report measures of pain and quality of life 4 or 5 times a day over 3 months. The compliance rate across all participants ranged from around 85% at the beginning of the study to around 30% at the end of the 3 months. The authors suggested the reduction over time was largely a result of participant attrition, possibly driven by technical factors including poor watch battery life. Considering just active participants, compliance remained above 75% throughout the 3 months. Ponnada and colleagues [10] explored using the same smartwatch-based µEMA methods for capturing self-report data of physical activity and its effect in young adults over an even longer time frame of 12 months. Preliminary findings at 6 months (when the authors published interim results of their study) indicated an overall compliance of rate 67% and a completion rate of 79% [10]. These findings suggest that while it is necessary to be mindful of technical issues that can lead to attrition, smartwatch-based µEMA appears to be a viable method for collecting high-temporal-resolution self-report data over extended time frames.

In terms of the effect of sample characteristics on EMA study compliance, several recent studies have reported mixed findings for age and sex [11-13]. One sample characteristic not explored as extensively is socioeconomic position (SEP). For studies of alcohol use, SEP is of particular importance as there are known to be complex relationships between SEP, alcohol consumption, and health outcomes (the “alcohol harm paradox”), in which individuals are more likely to have drunk in the past week and drank 6 units of alcohol or more in 1 drinking episode if they are a high-income earning managerial/professional worker [14,15], but individuals with lower individual or neighborhood SEP have an increased susceptibility to the negative and harmful health effects of alcohol consumption [16,17]. Understanding if SEP has a systematic effect on engagement with any new methods measuring alcohol use is therefore clearly important.

XSL•FO
RenderX

The objective of this study was to explore the feasibility of using smartwatch-based µEMA methods for the capture of high-resolution self-reported data about alcohol use over extended periods in high- and low-SEP groups. We asked participants to use a smartwatch-based µEMA system to record self-reported data about alcohol use and the context in which alcohol was consumed 5 times a day over a 3-month period. For comparison, we also asked all participants to record their alcohol use every 2 weeks using an online version of one of the currently most established methods for capturing data on alcohol consumption, the Timeline Followback (TLFB) method. We report engagement metrics (compliance and completion) for both smartwatch µEMA and TLFB methods, qualitative experiences of using both methods captured in interviews at the end of the study, and some initial findings comparing the levels of alcohol recorded by the 2 methods at the level of individual participants. Finally, in line with recommendations from a recent evaluation of pressing issues in EMA studies [18], we also report the characteristics of missing data for the smartwatch µEMA for all participants and high- and low-SEP groups, to explore whether SEP impacts the nature of missing data in a systematic manner.

Methods

Study Overview

The design and analyses of the study followed those detailed in the preregistered study protocol [19]. Any deviations from the study protocol are detailed in the following sections. An overview of the study elements, with their timings throughout the study, is shown in Figure 1. Data collection happened between September 2019 and June 2020.

Figure 1. Overview of study elements with timings.

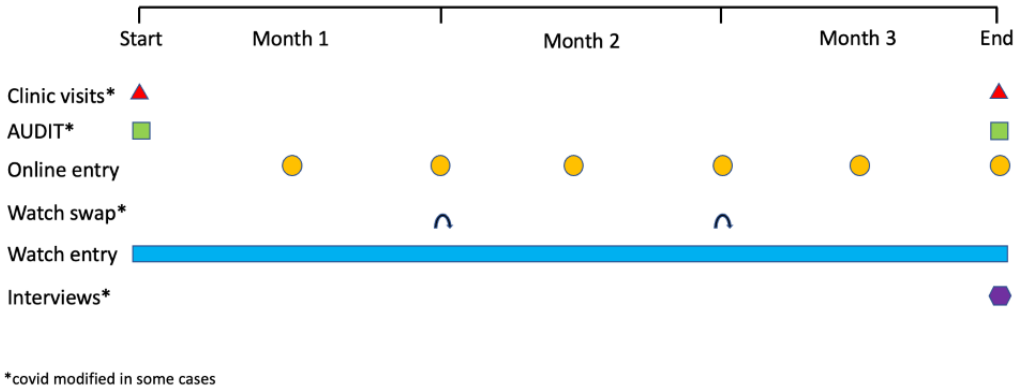

Study/Clinical Setting

This study was embedded within the Avon Longitudinal Study of Parents and Children (ALSPAC), a longitudinal birth-cohort study. During phase I of ALSPAC enrollment, 14,541 pregnancies in the former Avon Health Authority in the southwest of England with expected dates of delivery between April 1, 1991, and December 31, 1992, were recruited. Of these initial pregnancies, there was a total of 14,676 fetuses, resulting in 14,062 live births and 13,988 children who were alive at 1 year of age. A further 906 pregnancies were recruited during phases II, III, and IV, respectively, resulting in an additional 913 children being enrolled. The total sample size is 15,447 pregnancies, of which 14,901 were alive at 1 year of age.

Further details on the cohort profile, representativeness, and phases of recruitment are described in 3 cohort-profile papers [20-22].

The ALSPAC study website provides details on all available data through a fully searchable data dictionary and variable search tool [23]. Informed consent for data collected via questionnaires and clinics was obtained from participants in accordance with the recommendations of the ALSPAC Ethics and Law Committee at the time.

In this study, data were captured using smartwatch-based µEMA techniques as participants went about their normal daily lives and through online TLFB methods periodically throughout the study. An initial session was held at the ALSPAC premises in the Bristol Medical School, during which participants were briefed, provided consent, and completed questionnaires. A final session was also held there, where participants were debriefed, completed a qualitative interview about their experiences during the study, and returned equipment.

Study data for the TLFB method were collected and managed using REDCap, a secure, web-based electronic data capture platform hosted at the University of Bristol. REDCap (Research Electronic Data Capture) is designed to support data capture for research studies [24].

Ethical Considerations

Ethics Review

As this research involved human participants, ethics approval was obtained from the ALSPAC Ethics and Law Committee, which reviewed the study and granted formal approval (reference code 83643) in accordance with the University of Bristol institutional guidelines.

Informed Consent

Prospective participants received an information sheet before the initial study session, explaining the study's nature, purpose, and risks, and were given the opportunity to ask investigators any questions before deciding to participate. Upon arrival at the initial session, participants could review the information sheet again and ask further questions. Informed consent for data collection was then obtained through a written consent form,

XSL•FO
RenderX

following the recommendations of the ALSPAC Ethics and Law Committee. Participants were also informed that they could withdraw from the study at any time without providing a reason.

Privacy and Confidentiality

The data obtained from participants have been anonymized to ensure they are not identifiable. No individual participant can be identified from this manuscript or its multimedia appendices.

Compensation

Participants were reimbursed £15 (US $19.41) per month for their time and expenses, with a bonus of £15 (US $19.41) for completing all 3 months. Thus, those who completed the full study received a total of £60 (US $78 at the time of the study).

Participants

Participants were drawn from the first generation of children born in the ALSPAC study (ALSPAC-G1) and were selectively recruited by SEP based on maternal highest educational attainment. Those whose mothers had a degree or higher were assigned to the high-SEP group, while those whose mothers had less than a degree were assigned to the low-SEP group.

The inclusion criteria were as follows:

- Consume at least 6 units of alcohol per week.
- Be able to wear and use the smartwatch from 12 PM to 10 PM (the timings of questions in the current version of the smartwatch system cannot be altered to accommodate shift work, etc)
- Have access to the internet to complete the TLFB assessments.

Exclusion criteria (standard exclusion criteria used in alcohol studies to prevent harm during pregnancy and to those at risk of addiction) were as follows:

- Pregnant or breastfeeding (self-reported). Note that if a participant becomes pregnant during the study, they must withdraw from the study at this point.
- History of alcohol or drug addiction (self-reported).
- Strong familial history of alcoholism defined as 1 or more immediate relatives (parent, sibling) or more than 1 other relative (eg, cousin, grandparent; self-reported).

Clinic Visits

All participants began the study with a visit to an ALSPAC clinic. In this session (approximately 1 hour in duration) participants

- were given details of the study and provided written consent;
- were given the smartwatch and shown how to use the µEMA application and maintain the smartwatch (including instructions on keeping it charged, not getting it wet, and removing it before contact sports);
- were given instructions on how to use the online TLFB diary, provided with log-on details, and asked about any events (eg, birthdays, nights out) that should be entered into the diary as reminders;
- were provided with contact details of whom to call if they encountered any issues during the study.

Participants completed a second clinic visit at the end of the study. In this session (approximately 30 minutes) participants

- returned the smartwatch;
- completed a brief semistructured interview about their experiences using the smartwatch and online TLFB diary led by the researcher;
- provided written final consent; and
- were given their cash reimbursement.

A total of 13 participants (8 low SEP and 5 high SEP) completed the second clinic visit in person.

As a result of COVID-19 restrictions, the remaining 19 participants (11 low SEP and 8 high SEP) completed their second clinic visit remotely. In these remote sessions, ALSPAC researchers conducted telephone interviews, and smartwatch collection was arranged via courier.

Smartwatch Data Collection

We developed a bespoke µEMA smartwatch application incorporating elements proposed by Intille et al [8] to reduce participant burden. These included using haptic rather than audible prompts to minimize disruption and designing brief questions with a limited set of single-tap responses to reduce effort.

To optimize our app design, we collaborated with the ALSPAC Original Cohort Advisory Panel, a committee representing the diverse backgrounds and perspectives of ALSPAC participants. Using a think-aloud approach, panel members walked through the µEMA application, identifying issues related to visual design, wording, and navigation. Several common themes emerged, leading to the following changes in the final µEMA application design:

- The introduction of the “My usual” shortcut to speed up the recording of alcoholic drinks that are the same type, quantity, and context.
- The introduction of the “Back” option so participants could go back and correct previous entries in the current epoch.

Participants received a TicWatch C2 (Mobvoi Information Technology Company Limited) smartwatch running Android Wear 2.6 with our µEMA application preloaded.

At 5 time points each day (2 PM, 4 PM, 6 PM, 8 PM, and 10 PM), participants were asked about any alcoholic drinks consumed in the past 2 hours and the context of these drinking events. At each time point, they were presented with the set of questions and response options listed in Textbox 1.

Examples of these questions as rendered in our µEMA application are shown in Figure 2, with the full question sequence illustrated in Multimedia Appendix 1.

XSL•FO
RenderX

Textbox 1. Questions and options presented to participants.

- Question 1: Did you drink alcohol in the last 2 hours? Response options [Yes | No | My Usual].
- Question 2: What were you drinking? Response options [Beer/cider | Wine | Spirits].
- Question 3: What size was your drink? Response options: Beer/cider: [Half a pint | 330 bottle | Pint], Wine: [125 mL | 175 mL |250 mL], Spirits: [Single | Double | Free pour].
- Question 4: Who were you with? Response options [Alone | In company].
- Question 5: Where were you? Response options [At home | Elsewhere].
- Question 6 Any more alcoholic drinks to be recorded? Response options [Yes | No | My Usual].

"Usual": On the first day of the study, participants were given the option to set up a "My Usual" shortcut to simplify data entry for their most frequently consumed alcoholic drink. If they opted in, they answered questions 2-5 to specify their usual drink. Thereafter, selecting "My Usual" automatically recorded those drink details for that time point.

Figure 2. Example smartwatch microinteraction-based ecological momentary assessment (µEMA) application questions.

Each day at noon, participants were asked if they had missed recording any drinks the previous day. If they responded "Yes," they were prompted to report those drinks using the same set of questions.

Captured data were not visible to participants and were stored locally on the smartwatch rather than uploaded to a cloud-based platform to minimize the risk of personal data loss. As a result of the lack of remote monitoring, smartwatches were swapped at the end of months 1 and 2 to mitigate the risk of undetected technical issues disrupting data collection. However, COVID-19 restrictions prevented this for 15 participants (7 low SEP and 8 high SEP), who were instead instructed to continue using their current smartwatch until the study's conclusion.

Online Data Collection

The TLFB method uses a calendar prepopulated with participant-defined diary entries (eg, birthdays, theatre visits, nights out) to aid in recall of alcohol consumption. Online self-administration of TLFB has been validated [25], and previous studies suggest that a 14-day recall period optimally balances data quality and participant burden [26]. Therefore, participants completed an online TLFB every 14 days, recording their alcohol consumption over the previous 2 weeks.

We implemented TLFB using REDCap [24], hosted on the ALSPAC application platform. Participants had a 5-day window to complete each 2-week TLFB entry and received SMS text message reminders at the start of this period. In each entry, participants recorded the number of alcohol units consumed each day. As data captured in the smartwatch µEMA application were not visible to participants, they could not use it to complete TLFB entries. In line with standard TLFB practice, no context questions were asked. The format of the TLFB diary entry screen is shown in Multimedia Appendix 2.

Interviews

In the second clinic session at the end of the study (conducted either in person or remotely by telephone), each participant completed a semistructured interview exploring their experiences during the study.

For both smartwatch µEMA and online TLFB methods, participants were asked:

Question 1/6: Overall, how would you rate your experience of using the smartwatch/online system during the study, on a scale from 1 (I did not like it at all) to 10 (I really liked it)?

Question 2/7: If you were asked to use the smartwatch/online system again in another study, how likely would you be to say yes, on a scale from 1 (I would not want to use it again) to 10 (I would really like to use it again)?

Question 3/8: If you agreed to use the smartwatch/online system again in another study, what is the maximum length of time you would be willing to use it for?

Question 4/9: What things did you like about using the smartwatch/online system?

Question 5/10: What things did you dislike about using the smartwatch/online system?

XSL•FO
RenderX

Participants were asked additional questions specifically about their experiences using the smartwatch µEMA system:

Question 11: How did you feel about having to charge the smartwatch every night?

Question 12: Did you use any of the other functions on the smartwatch during the study (eg, step counter)?

Question 13: Did you pair the smartwatch with any smartphones during the study?

Participants were also asked about their attitudes toward future developments of the smartwatch µEMA system:

Question 14: How would you feel if the smartwatch system used data from its motion sensors to work out when the best time is to ask you the questions (eg, so it does not ask you questions when you are driving)?

Question 15: How would you feel if the smartwatch system used data from its global positioning system (GPS) location sensor (which tells the watch where the person is) to work out when the best time is to ask you the questions (eg, so it does not ask you questions when you are at work)?

Engagement Measures

In reporting engagement, we follow the approach used by Intille and colleagues [8]. We report compliance rates, defined as the proportion of scheduled questions answered by participants, both for all participants who started the study and for those still active in the study. Participants are considered active until they make their final response and drop out.

In addition, for smartwatch µEMA, we report the completion rate, defined as the proportion of questions delivered to active participants by the µEMA application that they respond to. This helps distinguish whether nonresponses result from factors preventing question delivery (eg, the smartwatch being switched off or the battery being flat) or from participants' propensity and ability to respond to delivered questions. The completion rate is not reported for online TLFB, as questions were not actively presented to participants, meaning there were no prompts that could be delivered or missed.

For all engagement measures, we report data for all participants as well as separately for the low- and high-SEP groups.

Missing Data

We examined whether prompts were more likely to be missed at certain times of the day by comparing missingness rates across 2-hour epochs. Similarly, we assessed whether prompts were more likely to be missed on specific days of the week using the same approach. These analyses were conducted for all participants as well as separately for the low- and high-SEP groups.

Comparison of Levels of Alcohol Consumption Recorded Using the Two Methods

To provide an initial indication of whether µEMA might capture higher levels of alcohol consumption than TLFB, we calculated the total number of units consumed over the 12-week study period for each participant, using only weeks where data were available from both methods. Within the high- and low-SEP groups, we compared the number of participants whose recorded consumption was higher in one method than the other. Note that this was an exploratory analysis and not preplanned.

Results

Participants

Our original aim was to recruit 40 participants (20 low SEP and 20 high SEP). However, recruitment had to be halted prematurely in March 2020 due to COVID-19 restrictions, by which point 32 participants had been recruited. Of these, 13 were high SEP (mean age 27 years, range 26-28 years; 7, 54%, females) and 19 were low SEP (mean age 28 years, range 26-28 years; 13, 68%, females).

Engagement

Figure 3 presents the compliance rate for all participants who commenced the study using the smartwatch µEMA and online TLFB methods. Figure 4 illustrates the compliance rate for participants who remained active in the study (ie, those who had not yet dropped out) for both methods. Figure 5 displays the completion rate for all active participants using the smartwatch µEMA method (not applicable to TLFB, as explained above). Figure 6 depicts the participant attrition rate, representing the percentage of participants still actively engaged in the study, for both smartwatch µEMA and online TLFB methods.

XSL•FO
RenderX

Figure 3. Smartwatch microinteraction-based ecological momentary assessment (top) and timeline followback (TLFB; bottom) compliance rates (mean and 95%CI) across all participants who started the study, for all (black solid line), high-SEP (red dotted dashed line), and low-SEP (blue dashed line) participants. SEP: socioeconomic position.

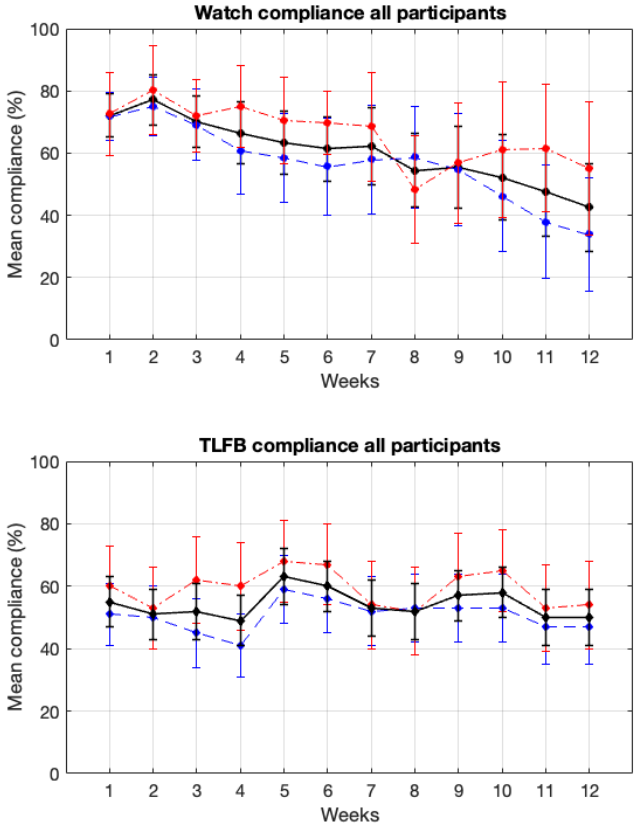

Figure 4. Smartwatch microinteraction-based ecological momentary assessment (top) and timeline followback (TLFB; bottom) compliance rates (mean and 95%CI) across all participants still active in the study for all (black solid line), high-SEP (red dotted dashed line), and low-SEP (blue dashed line) participants. SEP: socioeconomic position.

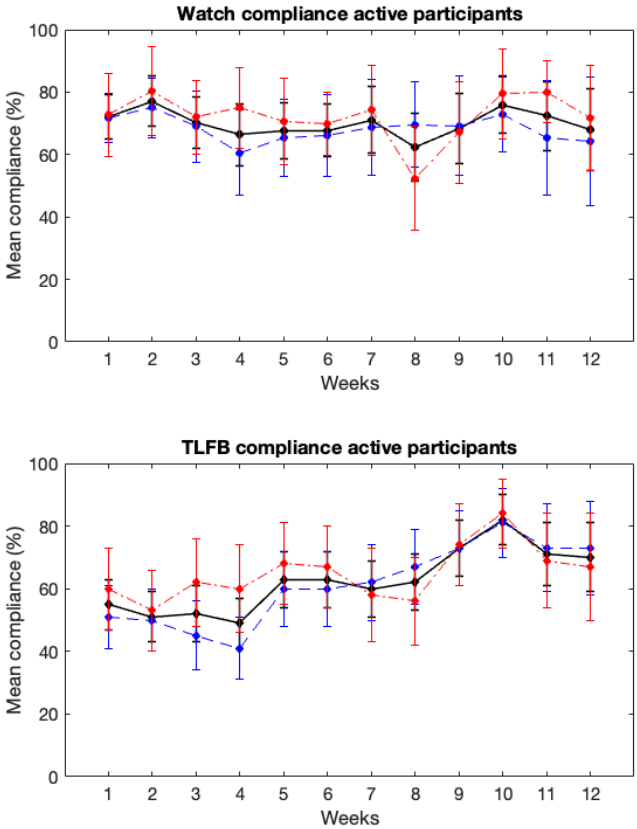

XSL•FO
RenderX

Figure 5. Smartwatch microinteraction-based ecological momentary assessment completion rate (mean and 95%CI) for all (black solid line), high-SEP (red dotted dashed line), and low-SEP (blue dashed line) participants. SEP: socioeconomic position.

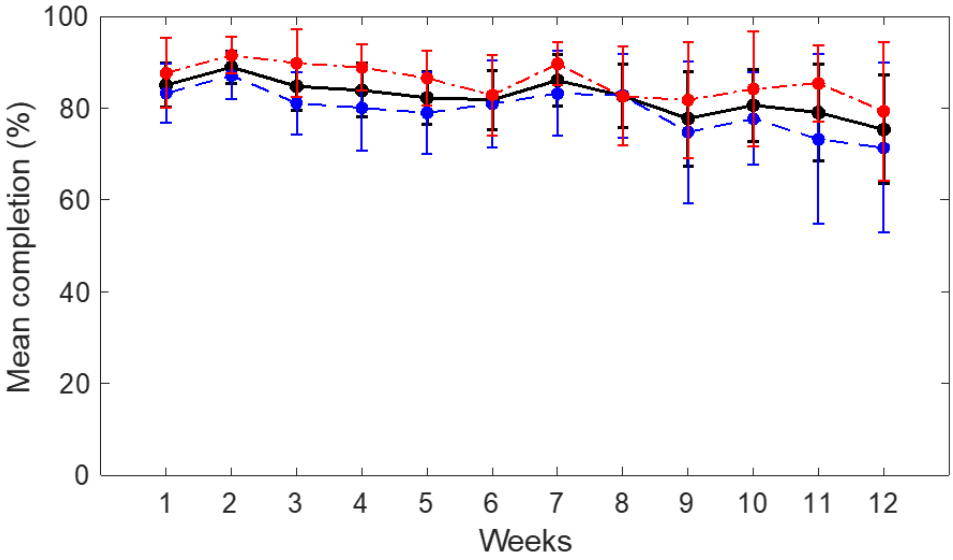

Figure 6. Participant attrition for smartwatch microinteraction-based ecological momentary assessment (circles) and online timeline followback (diamonds) for all (black solid line), high-SEP (red dotted dashed line), and low-SEP (blue dashed line) participants. SEP: socioeconomic position.

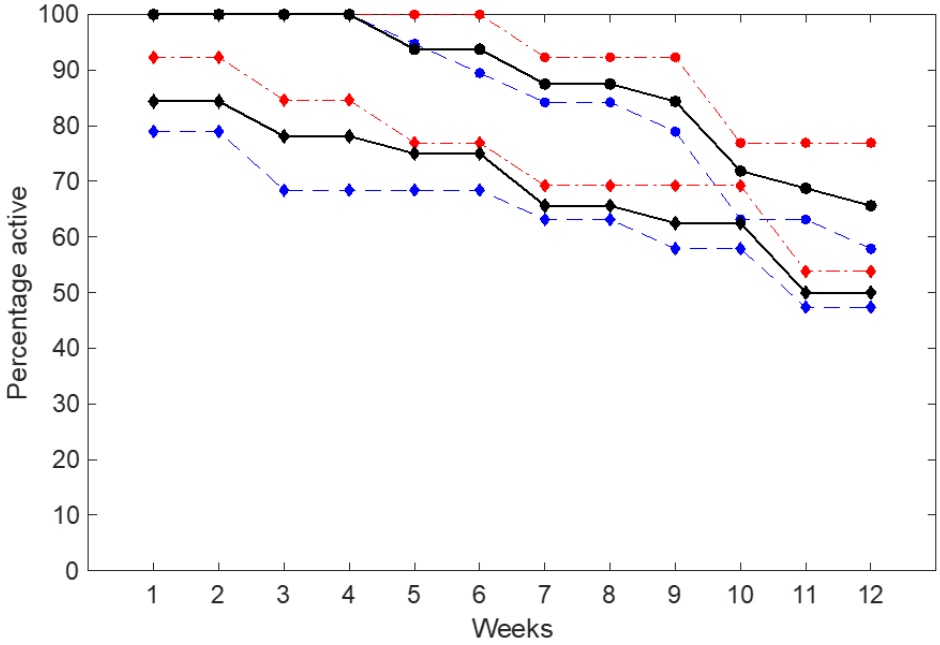

Qualitative Feedback

All participants completed interview questions regarding both the online and smartwatch methods for data collection. Participants' responses to the questions "How would you rate your experience of using the smartwatch µEMA/online TLFB from 1 (did not like) to 10 (really liked)?," "How willing would you be to use the smartwatch µEMA/online TLFB again in another study from 1 (would not) to 10 (really like to)?," and "How long would you be willing to use it for (in months)?" are presented in Table 1.

Themes identified from participants' responses to the question "What things did you like and dislike about using the smartwatch system?" are presented in Table 2, grouped by SEP. The equivalent themes for the online TLFB system are shown in Multimedia Appendix 3. Table 2 also includes participants' responses to the questions "How did you feel about having to charge the smartwatch every night?," "Did you pair the smartwatch with any smartphones during the study?," and "Did you use any of the other functions on the smartwatch during the study (eg, step counter)?"

Distinct themes identified from participants' responses to the questions "How would you feel if the smartwatch system used data from its GPS location sensor to determine the best time to ask you questions (eg, to avoid prompting during work)?" and "How would you feel if the smartwatch system used data from its motion sensors to determine the best time to ask you questions (eg, to avoid prompting while driving)?" are presented in Table 3.

Because of a technical issue with the questionnaire delivery, 6 participants (2 high SEP [1 female] and 4 low SEP [2 female])

XSL•FO
RenderX

were not asked the questions about smartwatch charging, pairing and other functions, or about the potential use of GPS and motion sensors.

Table 1. Participants' ratings of the extent to which they liked using, would be willing to use it again, and for how long, for the smartwatch µEMA[a] (Watch) and online (TLFB[b]) methods.

Participants' ratings of smartwatch µEMA/TLFB	All SEP[c], median (IQR)		Low SEP, median (IQR)		High SEP, median (IQR)	
	Watch	TLFB	Watch	TLFB	Watch	TLFB
Experience of using	8 (6-8)	7 (6-8)	8 (6-8)	7 (5-9)	8 (7-8)	8 (5-8)
Willing to use again	10 (7-10)	8 (7-10)	10 (7-10)	9 (7-10)	9 (6.5-10)	8 (6-10)
How long for (months)?	3 (3-12)	4 (3-12)	6 (3-12)	6 (3-12)	3 (2-5)	3 (2-5)

[a]µEMA: microinteraction-based ecological momentary assessment.
[b]TLFB: timeline followback.
[c]SEP: socioeconomic position.

Table 2. Participants' likes and dislikes of the smartwatch µEMA[a] method, attitudes to daily charging, attempts to pair with phone, and other watch features used, for low- and high-SEP[b] participants.

Participants' likes/dislikes of smartwatch µEMA	Low SEP, n/N (%)	High SEP, n/N (%)
Watch liked		
Quick and easy	9/19 (47)	6/13 (46)
Being prompted	1/19 (5)	4/13 (31)
Having a watch	0/19 (0)	2/13 (15)
Liked the smartwatch	2/19 (10)	2/13 (15)
Additional smartwatch features	2/19 (10)	0/13 (0)
No need for extra device	0/19 (0)	2/13 (15)
Always to hand	4/19 (21)	2/13 (15)
Watch disliked		
Timing/frequency of prompts	4/19 (21)	3/13 (23)
Technical issues	3/19 (16)	4/13 (31)
Battery life	4/19 (21)	2/13 (15)
Watch too big	3/19 (16)	2/13 (15)
Unsure how to handle missing drinks	1/19 (5)	2/13 (15)
Already have a watch/smartwatch	2/19 (10)	1/13 (8)
Attitude to needing to charge every night		
No issues	11/15 (73)	8/11 (73)
Frustrating	4/15 (27)	2/11 (18)
Attempted to pair with a smartphone		
Did not attempt	13/15 (87)	9/11 (82)
Attempted but failed	2/15 (13)	0/11 (0)
Attempted and succeeded	0/15 (0)	2/11 (18)
Other watch features used		
None	11/15 (73)	5/11 (45)
Step counter	3/15 (20)	3/11 (27)
Timer	1/15 (7)	2/11 (18)

[a]µEMA: microinteraction-based ecological momentary assessment.
[b]SEP: socioeconomic position.

XSL•FO
RenderX

Table 3. Participants' views on using GPS[a] data and motion sensor data to optimize smartwatch µEMA[b], for low- and high-SEP[c] participants.

Participants' attitudes toward the use of sensor data	Low SEP (n=15), n (%)	High SEP (n=11), n (%)
Attitudes to using GPS in the future		
No issues	13 (87)	7 (64)
Depends on context	3 (20)	1 (9)
OK if consented	0 (0)	1 (9)
Concerned about battery life	0 (0)	1 (9)
Scary	2 (13)	0 (0)
Intrusive	2 (13)	0 (0)
Attitudes to using motion sensors in the future		
No issues	15 (100)	9 (82)
Concerns about reliability	0 (0)	1 (9)
Uncomfortable with idea	0 (0)	1 (9)

[a]GPS: global positioning system.

[b]µEMA: microinteraction-based ecological momentary assessment.

[c]SEP: socioeconomic position.

Missing Data

The rates of missingness for the smartwatch µEMA method among participants still active in the study, stratified by all, low-, and high-SEP groups, are shown for each day of the week in Table 4. Considerable variability was observed across the 3 groups, with overlapping CIs for all days.

The rates of missingness for the smartwatch µEMA method among participants still active in the study, stratified by all, low-, and high-SEP groups, are shown for epochs within days in Table 5. Considerable variability was observed within each group, with overlapping CIs across epochs. However, there was notably less overlap when comparing the first four epochs (12:00–14:00, 14:00–16:00, 16:00–18:00, 18:00–20:00) with the final epoch (20:00–22:00), which had the highest level of missing data.

Table 4. Percentage missingness by day of the week for all, high-, and low-SEP[a] groups (mean and 95% CI).

Weekday	All participants (n=32), mean (95% CI)	High SEP (n=13), mean (95% CI)	Low SEP (n=19), mean (95% CI)
Monday	28.3 (21.2-35.5)	23.8 (14.4-33.3)	31.4(20.8-42.0)
Tuesday	25.6 (18.8-32.4)	24.9 (15.9-33.9)	26.1 (15.8-36.5)
Wednesday	25.0 (19.4-30.6)	23.6 (14.0-33.2)	26.0 (18.5-33.4)
Thursday	26.5 (21.1-31.2)	24.2 (15.3-33.1)	28.1 (20.6-35.5)
Friday	25.2 (19.1-31.2)	24.9 (14.6-35.1)	25.3 (17.0-33.7)
Saturday	28.0 (21.3-34.6)	25.8 (15.0-36.6)	29.5 (20.2-38.8)
Sunday	28.2 (21.0-35.4)	29.6 (18.2-41.0)	27.2 (17.0-37.4)

[a]SEP: socioeconomic position.

Table 5. Percentage missingness by 2-hour epoch for all, high-, and low-SEP[a] groups.

Epoch	All participants (n=32), mean (95% CI)	High SEP (n=13), mean (95% CI)	Low SEP (n=19), mean (95% CI)
12:00-14:00	26.7 (20.4-32.9)	26.1 (17.3-34.9)	27.0 (17.6-36.4)
14:00-16:00	23.6 (17.7-29.5)	22.2 (13.7-30.8)	24.5 (15.9-33.2)
16:00-18:00	24.7 (18.5-30.8)	23.1 (14.1-32.0)	25.8 (16.8-34.7)
18:00-20:00	26.0 (20.4-31.5)	24.2 (15.5-32.8)	27.2 (19.3-35.1)
20:00-22:00	32.5 (25.5-40.0)	30.7 (20.1-41.2)	33.8 (23.5-44.0)

[a]SEP: socioeconomic position.

XSL•FO
RenderX

Comparison of Levels of Alcohol Consumed Recorded Using the Two Methods

In the high-SEP group:

- 8 of the 13 (62%) participants had higher levels of alcohol units recorded using smartwatch µEMA than TLFB (percentage increase from TLFB: +44, +63, +27, +34, +76, +19, +2, and +71).
- 4 of the 13 (31%) had lower levels of alcohol recorded using smartwatch µEMA than TLFB (percentage decrease from TLFB: –11, –42, –10, and –50).
- 1 of the 13 (8%) did not have sufficient data for a comparison.

In the low-SEP group:

- 10 of the 19 (53%) participants had higher levels of alcohol units recorded using smartwatch µEMA than TLFB (percentage increase from TLFB: +42, +47, +18, +217, +10, +126, +33, +45, +12, and +15).
- 5 of the 19 (26%) had lower levels of alcohol recorded using smartwatch µEMA than TLFB (percentage decrease from TLFB: –6, –15, –61, –4, and –18).
- 4 of the 19 (21%) did not have sufficient data for a comparison.

Discussion

Principal Findings

Comparing Smartwatch µEMA and Online TLFB Engagement Metrics

Using the most stringent measure of engagement (compliance rate for all participants who started the study), compliance with the smartwatch µEMA method gradually declined over the 12 weeks, from approximately 70% in week 1 to around 45% in week 12. By contrast, compliance with the online TLFB method remained relatively stable at approximately 50% throughout the study. Based on this metric, the smartwatch µEMA method outperformed the online TLFB method until around week 9, after which their performance was broadly similar.

Comparing the compliance rates for all participants with those for participants still active in the study, and considering the attrition rate, it is evident that attrition significantly contributed to the observed decline in compliance among all participants over time. When focusing solely on participants still active in the study (ie, removing the effects of attrition), compliance with the smartwatch µEMA method remained relatively stable at around 70% throughout the 12 weeks. By contrast, compliance with the online TLFB method increased from approximately 50% in week 1 to around 80% in week 10, before declining to 70% in week 12. As before, the performance of the smartwatch µEMA method surpassed that of the online TLFB method until approximately week 9, after which only minor differences were observed between them. Notably, compliance with the online TLFB method increased during the first 10 weeks of the study. While no definitive explanation can be provided, one possible contributing factor is that COVID-19 lockdowns may have resulted in participants spending more time at home during certain periods, potentially improving ease of access to the online TLFB system. By contrast, access to the smartwatch µEMA method would not have been affected to the same extent, as it was specifically designed to be accessible regardless of location. Consequently, a similar increase in compliance would not be expected for this method.

Turning to the completion rate of the smartwatch µEMA method (not applicable to the online TLFB method, as previously discussed), this remained broadly stable at approximately 85% for the first 8 weeks, before declining to around 75% over the final 4 weeks. This suggests that, when the µEMA application functions correctly and the smartwatch device remains charged to ensure prompt delivery, participants exhibit a strong propensity to respond. The slight decline over the 12-week period indicates that longer data collection periods may still be feasible while maintaining acceptable completion rates using this method.

Comparing Engagement Metrics for Low- and High-SEP Groups

Although there is considerable variation within groups, and the differences between groups are generally modest (ranging from approximately 5% to 20%), nearly all point estimates for participant compliance, active participant compliance, completion rate, and attrition suggest lower engagement among low-SEP participants compared with high-SEP participants for both the smartwatch µEMA and online TLFB methods.

Patterns of Missing Data

There was considerable variability in all 3 groups (all SEP, low SEP, and high SEP), with overlapping CIs when comparing levels of missing data by days of the week and by 2-hour epochs within each group. The most notable pattern was the lower overlap in CIs observed in the all-participant group when comparing the first 4 epochs (12:00-14:00, 14:00-16:00, 16:00-18:00, and 18:00-20:00) with the final epoch (20:00-22:00), which had the highest level of missing data. Modifications to the µEMA application, such as allowing participants to configure their bedtime individually, may enhance engagement and reduce missing data during this period.

Regarding differences between high- and low-SEP groups, there was considerable overlap in CIs across all days and epochs. However, the point estimates for the proportion of missing data were consistently higher for low-SEP participants than for high-SEP participants across all epochs and on all days except Friday and Sunday.

Comparison of Levels of Alcohol Consumed Recorded Using the Two Methods

In both high- and low-SEP groups, approximately twice as many participants recorded higher alcohol consumption using the smartwatch µEMA method compared with the TLFB method over the study period. As this is a feasibility study, it was not powered to detect definitive differences in such comparisons, and these findings should be interpreted with caution. However, given the well-documented issue of underreporting in existing alcohol consumption recording methods [3,4], these preliminary results suggest that further research is warranted to assess whether smartwatch µEMA offers improvements in this area.

XSL•FO
RenderX

Qualitative Feedback

Across all participants, both methods were rated highly, with a slight preference for the smartwatch µEMA (8/10) over the online TLFB method (7/10). In the low-SEP group, the smartwatch µEMA received a slightly higher rating (8/10) than the online TLFB method (7/10), while the high-SEP group rated both methods equally (8/10).

For willingness to use the methods again, both scored highly, with all participants showing a preference for the smartwatch µEMA (10/10) over the online TLFB method (8/10). This pattern was consistent across SEP groups, with the high-SEP group rating the smartwatch µEMA slightly higher (9/10 vs 8/10) and the low-SEP group showing a stronger preference (10/10 vs 9/10).

Regarding the duration participants would be willing to use the methods again, online TLFB was rated slightly higher overall (4 months) compared with the smartwatch µEMA (3 months). However, within SEP groups, no differences were observed, with the low SEP group indicating a willingness to use both methods for 6 months and the high-SEP group for 3 months.

For the smartwatch µEMA method, the most appreciated factor was its ease and speed of use. However, participants reported a range of disliked aspects, including the timing and frequency of prompts, technical issues, battery life, and smartwatch size, all of which may have contributed to attrition. Regarding daily charging, 19 out of 26 (73%) participants in both SEP groups had no issues, while 4 out of 15 (27%) in the low-SEP group and 2 out of 11 (18%) in the high-SEP group found it frustrating. Additionally, 22 out of 26 (85%) participants in both groups adhered to study instructions by not pairing the smartwatch with their own smartphone. In the low-SEP group, 2 of the 15 (13%) who attempted to pair the smartwatch with their smartphone were unsuccessful, whereas in the high-SEP group, 2 of the 11 (18%) who attempted were successful. High-SEP participants (6/11, 55%) explored additional smartwatch features more frequently than low-SEP participants (4/15, 27%), with the step counter and timer being the most commonly used features.

Regarding attitudes toward using additional smartwatch sensors for data capture, the low-SEP group was more accepting of both GPS (13/15, 87%, had no concerns vs 7/11, 64%, in the high-SEP group) and motion sensors (15/15, 100% vs 9/11, 82%, in the high-SEP group).

Comparison With Prior Work

Alcohol EMA Studies

Perski et al's [2] systematic review and meta-analysis of EMA studies on health behaviors reported a median adherence of 84% (IQR 77%-91%) across 175 alcohol EMA studies, with smartphone-based EMA methods showing higher adherence than other approaches (eg, online and paper-and-pen). Notably, no studies included wrist-worn wearable EMA methods. Among the engagement metrics we assessed, the completion rate appeared most comparable to their adherence metric, averaging around 80% over the 12 weeks of our study—placing it within the reported range.

In his review of alcohol EMA methods, Piasecki [27] identified several studies employing high-resolution EMA approaches—primarily smartphone based—that prompted participants for responses throughout the day, with study durations ranging from a few days to 8 weeks. Reported response rates varied between 63% and 90%. While direct comparisons with our engagement metrics are not straightforward, our completion rate (~80%) and active participant compliance rate (~70%) fall within this range. Our all-participant compliance rate started at ~70% but dropped below 60% by week 8, remaining broadly consistent with these studies. However, as our study extended beyond 8 weeks, all-participant compliance declined further to 45% by week 12, suggesting that engagement may decrease more substantially in studies of longer duration.

Howard and Lamb [28] recently employed SMS text message–initiated smartphone-based EMA surveys to study alcohol consumption among undergraduates, reporting high compliance levels for active participants—starting at 89% in week 1 and gradually declining to 70% by week 14. By contrast, our active participant compliance rates were lower, fluctuating around 70% over the 12-week period. While no single factor can be identified as the cause of this difference, potential contributors include the nature of the prompts—the study by Howard and Lamb [28] used audible phone notifications, which may have been more attention grabbing than the smartwatch's haptic prompts—and differences in reward structures, with their incentives tied to individual survey completion, whereas ours were based on months of study participation. However, prior meta-analyses suggest that variations in reward strategies do not consistently impact engagement levels [11,12].

Smartwatch µEMA Studies

To the best of our knowledge, there are currently no broadly agreed benchmarks for compliance, completion, and attrition rates in smartwatch µEMA studies. In a study with a comparable assessment period, Beukenhorst et al [9] used smartwatch µEMA methods to collect self-reported pain data from patients with knee osteoarthritis, administering 4-5 questions per day over 13 weeks. Their attrition rate closely mirrored that of our study, with active participant numbers steadily declining to approximately 70% by week 10. However, by the end of the 13-week period, their attrition rate had dropped further to 42%, whereas ours remained at 65% by week 12. The compliance rate for all participants in the study by Beukenhorst et al [9] was also similar to that observed in our study, starting at approximately 75% before gradually declining to around 50% by week 10 and 45% by week 12. Completion rates showed a comparable pattern, beginning at approximately 85% and remaining around 80% throughout the study duration. Additionally, participants in both studies identified short battery life as a barrier to smartwatch use.

In what may be the longest smartwatch µEMA study to date, Ponnada et al [10] collected self-reported data on physical activity and its effect in young adults over a 12-month period, using smartwatch-based µEMA methods across 24 four-day burst periods. In their preliminary findings at the 6-month mark, the mean overall compliance of µEMA for active participants

XSL•FO
RenderX

during burst periods was 67%, comparable to the 70% observed over the 12 weeks of this study. Similarly, their reported mean completion rate of 79% at 6 months aligns with the completion rate in this study, which began at 85% and declined to 75% by week 12.

Effects of High and Low SEP

Regarding the impact of SEP on engagement in alcohol EMA studies, Howard and Lamb [28] found no significant differences in the number of students with college-educated parents (a measure similar to the SEP indicator used in this study) across their 4 responder categories ("poor," "adequate," "good," and "super"). However, they noted that the proportion of students with college-educated parents was higher in the "good" and "super" responder categories than in the "poor" category, suggesting that higher SEP may have contributed to increased engagement—consistent with the findings of this study.

Considering the broader impact of SEP on compliance in alcohol studies, Thern and Landberg [29] reported that low-SEP participants had approximately twice as much missing data as high-SEP participants for self-reported factors such as self-rated health, health-related quality of life, and level of social support received. While the current feasibility study is not powered to provide definitive evidence of higher levels of missing data in low-SEP participants, the point estimates observed here align with this pattern.

Strengths and Limitations

This is the first study to assess the feasibility of using smartwatch-based µEMA methods to collect high-temporal-density data on alcohol consumption over extended periods.

While our smartwatch-based EMA app was largely based on the µEMA methods originally developed by Intille and colleagues [8], a key aspect of true µEMA implementation is the presentation of only 1 question at a time. In our approach, we chained questions to capture details on drink type, size, and consumption context, making it more accurately described as a modified µEMA. However, this modification appeared to have little impact on engagement.

Unfortunately, due to COVID-19, participant recruitment had to be halted early, leading to a reduced sample size and an unequal mix of high- and low-SEP participants. As a result, the generalizability of our findings may be limited compared with studies conducted outside the context of a global pandemic.

Our participants were recruited from a UK birth cohort study [20]. A key advantage of this approach is that future research can explore new methods for integrating high-temporal-density, longitudinal data on alcohol consumption from our smartwatch-based µEMA methods with the extensive phenotypic and genetic data available for these participants. This could enable novel investigations into the impact of alcohol consumption on health. However, a potential limitation is that participants in cohort studies are accustomed to regular assessments, which may have increased their motivation to engage with the study compared to individuals not involved in such cohorts [20].

There is growing recognition that the diverse adherence metrics reported in EMA studies can hinder meaningful comparisons across the literature [10,11]. In this study, we selected adherence metrics primarily to facilitate comparisons with existing smartwatch-based µEMA studies. However, we acknowledge that these metrics do not always allow for direct comparisons with other EMA studies using different methodologies.

Recommendation for Future Studies

The primary factor affecting adherence in this study appeared to be participant attrition. Feedback from the end-of-study interviews provided insights into potential causes, with some issues—such as smartwatch size and battery life—likely to improve as technology advances. However, other factors, including the timing and frequency of prompts, warrant further investigation to optimize compliance and minimize attrition. Additionally, future studies should consider whether specific contextual factors (eg, study location) influence engagement with µEMA methods. A recent work by Ponnada and colleagues [30] offers a valuable overview of this topic.

The consistency of the differences observed between high- and low-SEP groups in compliance, completion, and attrition suggests the need for further research to quantify these disparities and identify their underlying drivers. This also highlights the importance of incorporating user input from diverse SEP backgrounds during the early stages of system design. Beyond SEP, future research should also consider other demographic factors, such as age, that may influence engagement and the quality of data collected using these methods.

This study has focused on the feasibility of using new smartwatch-based µEMA methods to capture data on alcohol consumption. Future research should also investigate response latency (the time between prompt delivery and response), the validity of the data collected compared with existing methods such as TLFB, and how both feasibility and validity vary across other health-related behaviors beyond alcohol consumption.

As a community developing new EMA methods, it is crucial to adopt metrics that facilitate the widest and most meaningful comparisons of findings. Adherence measurement has become increasingly complex, as new EMA methods leveraging the latest technologies benefit from specific metrics (eg, completion and compliance) that help distinguish between technical limitations and participant engagement. Given these advancements, now may be an opportune time to revisit previous recommendations for standardized reporting in EMA studies [31,32], considering the capabilities of mobile and wearable technology–enabled methods.

Conclusions

In this study, participants demonstrated higher engagement—reflected in higher compliance and completion rates and lower attrition—when using smartwatch-based µEMA methods compared with online TLFB methods for recording alcohol consumption. These engagement levels were consistent with previous studies employing smartwatch µEMA methods. While participants reported similar experiences with both methods and expressed comparable willingness to use either in

XSL•FO
RenderX

the future, they showed a slight preference for the smartwatch µEMA approach over the online TLFB method.

While certain challenges remain—particularly for studies extending over months rather than weeks—our findings highlight the considerable potential of smartwatch µEMA methods in alcohol use research. These methods offer the advantage of capturing high-temporal-density, longitudinal data on alcohol consumption, along with contextual information about drinking events. As with EMA methods more broadly, they are less susceptible to recall and bias errors that affect traditional alcohol consumption reporting methods. However, it is important to acknowledge that social desirability bias remains a significant factor in alcohol reporting, and some level of underreporting may persist despite the use of EMA techniques.

Future studies should investigate factors contributing to participant attrition—the primary driver of reduced adherence—as well as latency issues and the validity of alcohol data captured using these methods. While our study does not provide definitive evidence of engagement differences between high- and low-SEP participants, the consistent pattern of lower engagement observed among low-SEP participants suggests that further research is warranted to explore the existence, magnitude, and underlying causes of such disparities.

Acknowledgments

We are extremely grateful to all the families who took part in this study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists, and nurses. This work was funded by a UK Research and Innovation Fellowship to AS (MR/S003894/1), and further supported by the Integrative Cancer Epidemiology Programme, funded by Cancer Research UK (grant C18281/A29019). The Medical Research Council Integrative Epidemiology Unit is supported by the Medical Research Council and the University of Bristol (grants MC_UU_00032/05 and MC_UU_00032/07). REW is funded by a postdoctoral fellowship from the South-Eastern Norway Regional Health Authority (2020024). The UK Medical Research Council and Wellcome (grant 217065/Z/19/Z) and the University of Bristol provide core support for ALSPAC.

Data Availability

The informed consent obtained from ALSPAC participants does not allow the data to be made freely available through any third-party–maintained public repository. However, data used for this submission can be made available on request to the ALSPAC Executive. The ALSPAC data management plan describes in detail the policy regarding data sharing, which is through a system of managed open access. Full instructions for applying for data access can be found on the ALSPAC researchers' access web page [33]. The ALSPAC study website contains details of all the data that are available [23].

Disclaimer

No generative AI tools were used in any portion of this manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
AlcoWatch microinteraction-based ecological momentary assessment smartwatch question formatting and sequencing.
[PNG File , 173 KB - formative_v9i1e63184_app1.png]

Multimedia Appendix 2
Timeline followback diary format.
[PNG File , 69 KB - formative_v9i1e63184_app2.png]

Multimedia Appendix 3
High- and low-socioeconomic position participants' likes and dislikes of the online timeline followback method.
[PNG File , 46 KB - formative_v9i1e63184_app3.png]

References

1. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol 2008 Apr 01;4(1):1-32. [doi: 10.1146/annurev.clinpsy.3.022806.091415] [Medline: 18509902]
2. Perski O, Keller J, Kale D, Asare BY, Schneider V, Powell D, et al. Understanding health behaviours in context: a systematic review and meta-analysis of ecological momentary assessment studies of five key health behaviours. Health Psychol Rev 2022 Dec;16(4):576-601 [FREE Full text] [doi: 10.1080/17437199.2022.2112258] [Medline: 35975950]

XSL•FO
RenderX

3. Embree BG, Whitehead PC. Validity and reliability of self-reported drinking behavior: dealing with the problem of response bias. J Stud Alcohol 1993 May;54(3):334-344. [doi: 10.15288/jsa.1993.54.334] [Medline: 8487543]
4. Midanik L. The validity of self - reported alcohol consumption and alcohol problems: a literature review. British Journal of Addiction 2010 Feb 12;77(4):357-382. [doi: 10.1111/j.1360-0443.1982.tb02469.x] [Medline: 6762224]
5. Gmel G, Daeppen J. Recall bias for seven-day recall measurement of alcohol consumption among emergency department patients: implications for case-crossover designs. J Stud Alcohol Drugs 2007 Mar;68(2):303-310. [doi: 10.15288/jsad.2007.68.303] [Medline: 17286350]
6. Hammerton G, Lewis G, Heron J, Fernandes G, Hickman M, Lewis G. The association of alcohol dependence and consumption during adolescence with depression in young adulthood, in England: a prospective cohort study. Lancet Psychiatry 2023 Jul;10(7):490-498 [FREE Full text] [doi: 10.1016/S2215-0366(23)00138-4] [Medline: 37271164]
7. Fridberg DJ, Lee Z, Fischer AM, Cursio JF, King AC. Assessing real-time positive subjective effects of alcohol using high-resolution ecological momentary assessment in risky versus light drinkers. Alcohol Clin Exp Res (Hoboken) 2023 Aug;47(8):1570-1581. [doi: 10.1111/acer.15130] [Medline: 37578682]
8. Intille S, Haynes C, Maniar D, Ponnada A, Manjourides J. µEMA: microinteraction-based ecological momentary assessment (EMA) using a smartwatch. Proc ACM Int Conf Ubiquitous Comput 2016 Sep;2016:1124-1128 [FREE Full text] [doi: 10.1145/2971648.2971717] [Medline: 30238088]
9. Beukenhorst AL, Howells K, Cook L, McBeth J, O'Neill TW, Parkes MJ, et al. Engagement and participant experiences with consumer smartwatches for health research: longitudinal, observational feasibility study. JMIR Mhealth Uhealth 2020 Jan 29;8(1):e14368 [FREE Full text] [doi: 10.2196/14368] [Medline: 32012078]
10. Ponnada A, Wang S, Chu D, Do B, Dunton G, Intille S. Intensive longitudinal data collection using microinteraction ecological momentary assessment: pilot and preliminary results. JMIR Form Res 2022 Feb 09;6(2):e32772 [FREE Full text] [doi: 10.2196/32772] [Medline: 35138253]
11. Ottenstein C, Werner L. Compliance in ambulatory assessment studies: investigating study and sample characteristics as predictors. Assessment 2021 Jul 20;29(8):1765-1776. [doi: 10.1177/10731911211032718] [Medline: 34282659]
12. Jones A, Remmerswaal D, Verveer I, Robinson E, Franken IHA, Wen CKF, et al. Compliance with ecological momentary assessment protocols in substance users: a meta-analysis. Addiction 2019 Apr;114(4):609-619 [FREE Full text] [doi: 10.1111/add.14503] [Medline: 30461120]
13. Rintala A, Wampers M, Myin-Germeys I, Viechtbauer W. Momentary predictors of compliance in studies using the experience sampling method. Psychiatry Res 2020 Feb 29;286:112896. [doi: 10.1016/j.psychres.2020.112896] [Medline: 32146247]
14. Office for National Statistics, UK Government. Alcohol-specific deaths in the UK registered in 2018. UK Government. URL: https://www.gov.uk/government/statistics/alcohol-specific-deaths-in-uk-registered-in-2018 [accessed 2025-03-10]
15. Grittner U, Kuntsche S, Gmel G, Bloomfield K. Alcohol consumption and social inequality at the individual and country levels--results from an international study. Eur J Public Health 2013 Apr;23(2):332-339 [FREE Full text] [doi: 10.1093/eurpub/cks044] [Medline: 22562712]
16. Mäkelä P, Paljärvi T. Do consequences of a given pattern of drinking vary by socioeconomic status? A mortality and hospitalisation follow-up for alcohol-related causes of the Finnish Drinking Habits Surveys. J Epidemiol Community Health 2008 Aug;62(8):728-733. [doi: 10.1136/jech.2007.065672] [Medline: 18621959]
17. Mackenbach JP, Kulhánová I, Bopp M, Borrell C, Deboosere P, Kovács K, et al. Inequalities in alcohol-related mortality in 17 European countries: a retrospective analysis of mortality registers. PLoS Med 2015 Dec;12(12):e1001909 [FREE Full text] [doi: 10.1371/journal.pmed.1001909] [Medline: 26625134]
18. Stone A, Schneider S, Smyth JM. Evaluation of pressing issues in ecological momentary assessment. Annu Rev Clin Psychol 2023 May 09;19:107-131 [FREE Full text] [doi: 10.1146/annurev-clinpsy-080921-083128] [Medline: 36475718]
19. Skinner A, Adams S, Stone CJ, Munafo MR, Timpson NJ, Tilling K, et al. Assessing the feasibility of using smartwatch-based ecological momentary assessment techniques for determining longitudinal patterns and contexts of alcohol use in UK adults. Open Science Framework. 2020 Mar 10. URL: https://doi.org/10.17605/OSF.IO/4ZA26 [accessed 2025-03-10]
20. Boyd A, Golding J, Macleod J, Lawlor DA, Fraser A, Henderson J, et al. Cohort Profile: the 'children of the 90s'--the index offspring of the Avon Longitudinal Study of Parents and Children. Int J Epidemiol 2013 Feb;42(1):111-127 [FREE Full text] [doi: 10.1093/ije/dys064] [Medline: 22507743]
21. Fraser A, Macdonald-Wallis C, Tilling K, Boyd A, Golding J, Davey Smith G, et al. Cohort profile: the Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. Int J Epidemiol 2013 Feb;42(1):97-110 [FREE Full text] [doi: 10.1093/ije/dys066] [Medline: 22507742]
22. Northstone K, Lewcock M, Groom A, Boyd A, Macleod J, Timpson N, et al. The Avon Longitudinal Study of Parents and Children (ALSPAC): an update on the enrolled sample of index children in 2019. Wellcome Open Res 2019;4:51 [FREE Full text] [doi: 10.12688/wellcomeopenres.15132.1] [Medline: 31020050]
23. Avon Longitudinal Study of Parents and Children. University of Bristol. URL: http://www.bris.ac.uk/alspac/researchers/our-data/ [accessed 2025-03-10]

XSL•FO
RenderX

24. Fiellin DA, McGinnis KA, Maisto SA, Justice AC, Bryant K. Measuring alcohol consumption using timeline followback in non-treatment-seeking medical clinic patients with and without HIV infection: 7-, 14-, or 30-day recall. J Stud Alcohol Drugs 2013 May;74(3):500-504 [FREE Full text] [doi: 10.15288/jsad.2013.74.500] [Medline: 23490581]
25. Rueger SY, Trela CJ, Palmeri M, King AC. Self-administered web-based timeline followback procedure for drinking and smoking behaviors in young adults. J Stud Alcohol Drugs 2012 Sep;73(5):829-833 [FREE Full text] [doi: 10.15288/jsad.2012.73.829] [Medline: 22846247]
26. Harris P, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381 [FREE Full text] [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
27. Piasecki TM. Assessment of alcohol use in the natural environment. Alcohol Clin Exp Res 2019 Apr;43(4):564-577 [FREE Full text] [doi: 10.1111/acer.13975] [Medline: 30748019]
28. Howard AL, Lamb M. Compliance trends in a 14-week ecological momentary assessment study of undergraduate alcohol drinkers. Assessment 2024 Mar;31(2):277-290 [FREE Full text] [doi: 10.1177/10731911231159937] [Medline: 36914966]
29. Thern E, Landberg J. Understanding the differential effect of alcohol consumption on the relation between socio-economic position and alcohol-related health problems: results from the Stockholm Public Health Cohort. Addiction 2021 Apr;116(4):799-808 [FREE Full text] [doi: 10.1111/add.15213] [Medline: 32738005]
30. Ponnada A, Li J, Wang S, Wang W, Dunton GF, Intille SS. Contextual biases in microinteraction ecological momentary assessment (µEMA) non-response. Proc ACM Interact Mob Wearable Ubiquitous Technol 2022 Mar;6(1):1-24. [doi: 10.1145/3517259] [Medline: 39866712]
31. Stone AA, Shiffman S. Capturing momentary, self-report data: a proposal for reporting guidelines. Ann Behav Med 2002;24(3):236-243. [doi: 10.1207/S15324796ABM2403_09] [Medline: 12173681]
32. Liao Y, Skelton K, Dunton G, Bruening M. A systematic review of methods and procedures used in ecological momentary assessments of diet and physical activity research in youth: an adapted STROBE Checklist for Reporting EMA Studies (CREMAS). J Med Internet Res 2016 Jun 21;18(6):e151 [FREE Full text] [doi: 10.2196/jmir.4954] [Medline: 27328833]
33. Process for applying for data access. Avon Longitudinal Study of Parents and Children, University of Bristol. URL: http://www.bris.ac.uk/alspac/researchers/access/ [accessed 2025-03-10]

Abbreviations

ALSPAC: Avon Longitudinal Study of Parents and Children
EMA: ecological momentary assessment
GPS: global positioning system
SEP: socioeconomic position
TLFB: timeline followback
µEMA: microinteraction-based ecological momentary assessment

Edited by A Mavragani; submitted 12.06.24; peer-reviewed by K Kurisu; comments to author 30.10.24; revised version received 12.02.25; accepted 02.03.25; published 25.03.25.

Please cite as:
Stone C, Adams S, Wootton RE, Skinner A
Smartwatch-Based Ecological Momentary Assessment for High-Temporal-Density, Longitudinal Measurement of Alcohol Use (AlcoWatch): Feasibility Evaluation
JMIR Form Res 2025;9:e63184
URL: https://formative.jmir.org/2025/1/e63184
doi:10.2196/63184
PMID:40131326

©Chris Stone, Sally Adams, Robyn E Wootton, Andy Skinner. Originally published in JMIR Formative Research (https://formative.jmir.org), 25.03.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Using Social Media to Engage and Enroll Underrepresented Populations: Longitudinal Digital Health Research

Christiana Harry[1], MPH; Sarah Goodday[1,2], MSc, PhD; Carol Chapman[3], MPH; Emma Karlin[1], MPH; April Joy Damian[4], MSc, PhD; Alexa Brooks[1], MS, RD; Adrien Boch[5], MA; Nelly Lugo[6], BS; Rebecca McMillan[7], MHIHIM; Jonell Tempero[8], MS; Ella Swanson[1], BS; Shannon Peabody[9], BA; Diane McKenzie[1], MHA; Stephen Friend[1,2], MD, PhD

[1]4YouandMe, Great Neck, NY, United States
[2]Department of Psychiatry, University of Oxford, Oxford, United Kingdom
[3]Crohn's and Colitis Foundation, New York, NY, United States
[4]Moses/Weitzman Health System, Middletown, CT, United States
[5]Evidation Health, San Mateo, CA, United States
[6]Washington University, St. Louis, MO, United States
[7]University of California San Diego Health, San Diego, CA, United States
[8]RespirAI US Inc, Omaha, NE, United States
[9]Dana-Farber Cancer Instituite, Boston, MA, United States

Corresponding Author:
Christiana Harry, MPH
4YouandMe
185 Great Neck Rd Ste 447
Great Neck, NY, 11021
United States
Phone: 1 206 861 3655
Email: christianacharry@gmail.com

Abstract

Background: Emerging digital health research poses roadblocks to the inclusion of historically marginalized populations in research. Exclusion of underresourced communities in digital health research is a result of multiple factors (eg, limited technology access, decreased digital literacy, language barriers, and historical mistrust of research and research institutions). Alternative methods of access and engagement may aid in achieving long-term sustainability of diversified participation in digital health research, ensuring that developed technologies and research outcomes are effective and equitable.

Objective: This study aims to (1) characterize socioeconomic and demographic differences in individuals who enrolled and engaged with different remote, digital, and traditional recruitment methods in a digital health pregnancy study and (2) determine whether social media outreach is an efficient way of recruiting and retaining specific underrepresented populations (URPs) in digital health research.

Methods: The Better Understanding the Metamorphosis of Pregnancy (BUMP) study was used as a case example. This is a prospective, observational, cohort study using digital health technology to increase understanding of pregnancy among 524 women, aged 18-40 years, in the United States. The study used different recruitment strategies: patient portal for genetic testing results, paid/unpaid social media ads, and a community health organization providing care to pregnant women (Moses/Weitzman Health System).

Results: Social media as a recruitment tool to engage URPs in a digital health study was overall effective, with a 23.6% (140/594) enrollment rate of those completing study interest forms across 25 weeks. Community-based partnerships were less successful, however, resulting in 53.3% (57/107) engagement with recruitment material and only 8.8% (5/57) ultimately enrolling in the study. Paid social media ads provided access to and enrollment of a diverse potential participant pool of race- or ethnicity-based URPs in comparison to other digital recruitment channels. Of those that engaged with study materials, paid recruitment had the highest percentage of non-White (non-Hispanic) respondents (85/321, 26.5%), in comparison to unpaid ads (Facebook and Reddit; 37/167, 22.2%). Of the enrolled participants, paid ads also had the highest percentage of non-White (non-Hispanic) participants (14/70, 20%), compared to unpaid ads (8/52, 15.4%) and genetic testing service subscribers (72/384, 18.8%). Recruitment

XSL•FO
RenderX

completed via paid ads (Instagram) had the highest study retention rate (52/70, 74.3%) across outreach methods, whereas recruitment via community-based partnerships had the lowest (2/5, 40%). Retention of non-White (non-Hispanic) participants was low across recruitment methods: paid (8/52, 15.4%), unpaid (3/35, 14.3%), and genetic testing service subscribers (50/281, 17.8%).

Conclusions: Social media recruitment (paid/unpaid) provides access to URPs and facilitates sustained retention similar to other methods, but with varying strengths and weaknesses. URPs showed lower retention rates than their White counterparts across outreach methods. Community-based recruitment showed lower engagement, enrollment, and retention. These findings highlight social media's potential for URP engagement and enrollment, illuminate potential roadblocks of traditional methods, and underscore the need for tailored research to improve URP enrollment and retention.

(JMIR Form Res 2025;9:e68093) doi:10.2196/68093

KEYWORDS

digital health research; digital health technology; recruitment; research subject; participant; pregnancy; maternal health; underrepresented populations; health equity; diversity; marginalized; advertisement; social media; retention; attrition; dropout

Introduction

Background

Participation in research studies by historically racial and ethnic marginalized populations is significantly lower compared to their White counterparts. This has resulted in the underrepresentation of these groups across research study disciplines [1-3]. The National Institutes of Health (NIH) defines underrepresented populations (URPs) as those with disproportionately low representation relative to their overall or disease-specific population [4]. Despite mandates to include these groups in federally funded research, factors such as general attitudes toward research, sociocultural barriers, and accessibility continue to hinder diversity in research populations [1,5-11].

Certain URPs, such as African American populations, face additional historical and structural barriers to research participation including low health literacy, lack of access to care, and mistrust of institutions [2]. For URPs, particularly African American populations, mistrust of academic and research institutions is the largest reported participation barrier due to historical events, such as the Tuskegee syphilis study, and worsened by present inequalities in the health care system [1]. As a direct result of this severe institutional mistrust among URPs generally, lack of engagement and subsequent enrollment of URPs in research studies persists and continues to contribute to health disparities within these specific population subgroups, as they tend to reject participatory outreach from more traditional recruitment methods associated with these mistrusted institutions and present health inequities [1].

These health inequities are especially seen in pregnant populations, as illustrated by present maternal health outcome statistics [12]. Since 2019, the maternal mortality rate in the United States has continued to rise steadily, jumping to 32.9 deaths per 100,000 live births in 2021, compared with a rate of 20.1 in 2019 [12]. For pregnant women of ethnic or racial URPs, this discrepancy in maternal health outcomes is hypervisible. The historical exclusion of pregnant women from research studies has been present due to a perceived concern of lack of direct participant benefit or risk of significant harm to offspring. This exclusion has resulted in gaps in scientific knowledge regarding key women's health issues that further exacerbate women's health outcome disparities [13].

As these disparities and roadblocks to participation continue to be observed, additional efforts are also necessary to identify successful evidence-based enrollment strategies for reaching and engaging URPs, especially in digital research studies. Current engagement and recruitment strategies of URPs often emphasize outreach via community-forward channels (eg, partnerships with community leaders, churches, community-based clinic sites, and other organizations) to alleviate some of the barriers of mistrust present in these communities [10]. However, conflicting evidence exists regarding the effectiveness of community outreach as an isolated strategy for increasing the enrollment of URPs [10].

Another suggested recruitment strategy to bolster the engagement and involvement of URPs in studies is social media advertising. More recently, social media has been used by researchers as a new source of participant recruitment and may be an especially useful tool in locating specific URPs due to the widespread reach of social media to potential participants who otherwise might have been not visible to researchers, the anonymity that social media provides users, and general mistrust of health care institutions and more traditional recruitment methods by URPs [2,14,15].

Over the past decade, social media use in the United States has steadily increased. As of 2022, over 300 million people in the United States use social media, making the United States home to the third largest social media audience globally [16,17]. The social media platforms Facebook (Meta), Twitter (X Corp), and Instagram (Meta) account for a combined 83% of all social media site visits in the United States [16]. As nearly 3 quarters of Americans report social media use, social media platforms now provide researchers a new, accessible alternative to traditional recruitment methods (eg, printed flyers and in-person outreach) to meet individuals where they are and help bridge the digital divide [18].

The digital divide refers to the patterns of difference in the utilization of technology, the internet, and social media by race, ethnicity, or socioeconomic status. The underlying drivers of the digital divide are manifold but include 2 key factors: differential access to and engagement (motivation) in the use

XSL•FO
RenderX

of digital health technologies (DHTs; smartphones, wearable devices or other digital products, electronic health records, and telehealth options). The use of social media or DHTs in digital research studies requires ownership of devices (eg, smartphones and wearable devices), adequate internet access, and a level of digital and health literacy, indicating the presence of additional cost and other barriers to digital health engagement and use of DHTs [19].

The other key driver of the digital divide is the disconnect in engagement, especially among URPs. Although social media usage for racially based URPs continues to surpass that of their White counterparts, these populations are significantly less likely to engage with health information on the internet and social media or use social networking channels to learn more about their health [20-22]. However, studies indicate expressed interest by specific URPs (eg, African American women) in using DHTs and social media to engage with health information [23-25].

As the use of DHTs in digital research studies continues to rapidly expand, new and innovative ways in which these technologies may be used to support individual and population health research also evolve [26,27]. DHTs and digital research studies present the potential to mitigate existing gaps in health care access, quality of care, and health care outcomes, further suggesting that social media may be a powerful, untapped resource in reaching and engaging URPs in DHT research studies and digital health [22].

Previous Work

Existing research studies have primarily used Facebook as a means of recruiting URPs, especially for stigmatized groups, but few research studies explore the feasibility and effectiveness of using other social media platforms and recruitment methods specifically to engage URPs in digital health studies [28]. The use of more nuanced social media outreach methods—such as advertisements on Instagram (a platform that has nearly 127.2 million active users in the United States as of 2023)—to recruit and engage URPs in research studies is still in its infancy [17].

Goal of This Paper

This paper aims to (1) characterize the socioeconomic and demographic differences among individuals who enrolled and participated in a US-based pregnancy study—the Better Understanding the Metamorphosis of Pregnancy (BUMP) study—through various remote, digital, and traditional recruitment methods and (2) to determine whether social media outreach is an effective means of recruiting and retaining historically URPs to participate in pregnancy-related digital health research.

Methods

Overview

The BUMP study is a prospective, observational, cohort study conducted by 4YouandMe, a nonprofit organization that pilots open-source, digital health research. The BUMP study aims to leverage DHTs in the collection of objective and subjective measurements of health to increase understanding of pregnancy and subsequent complications in a sample of women, aged 18-40 years, in the United States (n=524) [29]. The BUMP study sample was primarily White non-Hispanic (381/524, 72.7%) and between 26-35 years of age (363/524, 69.3%). The first participant was enrolled in the BUMP study on February 23, 2021, and the last participant completed the study on July 1, 2023. Participants were enrolled over a 6- to 12-month period and followed for up to 16 months.

The BUMP study used various recruitment methods, including a patient portal for genetic testing results during pregnancy, social media platforms like Facebook, Reddit, and Instagram through both paid and unpaid advertisements; in-person clinics at a trusted community health center offering prenatal and postnatal care; as well as other outreach strategies such as word of mouth and participant referrals. Data were collected through surveys and active tasks on a study smartphone app, wearable DHTs (eg, a Garmin smartwatch, an Oura smart ring, and a Bodyport smart scale), and patients' electronic medical records.

Recruitment

Recruitment for the BUMP study was primarily conducted via 2 methods: genetic testing service subscriber recruitment (Sema4) and 4YouandMe direct recruitment. Genetic testing service subscriber recruitment includes outreach to subscribers to genetic testing via Sema4's patient portal. 4YouandMe direct recruitment consists of 3 recruitment suboutlets: social media (including both paid and unpaid outreach), community partnerships, and others (eg, word of mouth, press, and ClinicalTrials.gov; Figure 1).

XSL•FO
RenderX

Figure 1. Flowchart of recruitment channels for the Better Understanding the Metamorphosis of Pregnancy (BUMP) study.

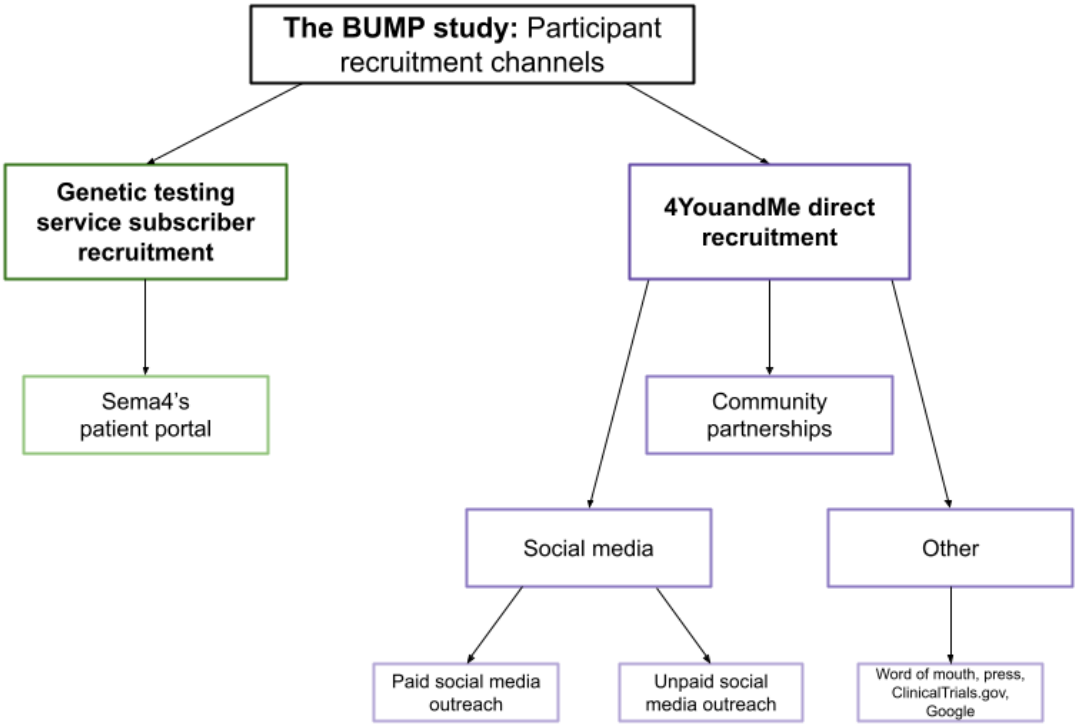

Genetic Testing Service Subscriber Recruitment (Sema4)

Primary recruitment for the BUMP study was completed via Sema4's patient portal, specifically focusing on those who signed up to receive results of genetic testing results during pregnancy. Sema4's patient portal combines mobile health and eHealth as it provides a mechanism for Sema4 to engage with its patients via communication of test results and management of electronic consents [29]. Patients who consented to having their electronic medical record data accessed by Sema4, and those requesting results of genetic testing or other tests commonly associated with pregnancy, received an email from Sema4 with information about the BUMP study, including contact information for study staff if further interested in participation [29].

4YouandMe Direct Recruitment (Social Media Recruitment, Community Partnerships, and Others)

All recruitment conducted by methods other than outreach to genetic testing service subscribers via Sema4's patient portal of genetic testing subscribers are referred to as 4YouandMe direct recruitment. 4YouandMe direct recruitment was conducted primarily through social media platforms (eg, Facebook, Reddit, and Instagram).

Unpaid Social Media Outreach (Social Media Recruitment)

Social media sites with moderated group forums related to pregnancy (eg, Facebook and Reddit) were the focus of unpaid social media recruitment efforts to reach the wider population of all study-eligible pregnant women for the BUMP study. The BUMP study staff conducted outreach efforts with forum and group moderators to gain permission to post fliers on these forums. The BUMP study staff monitored public response to the live group and forum posts and interacted with moderators and interested individuals by answering questions and engaging with posted content. The goal was to build trust and credibility within these online communities while mitigating some of the noted common challenges faced in social media–based recruitment.

Text study descriptions and image-heavy digital recruitment fliers were used to reach potential participants and posted as both standalone and response posts in various pregnancy-related online forums (eg, "Pregnant Women's Group" and "Mom's Talk: Pregnancy to Toddler" on Facebook; and "Science Based Parenting," "Pregnant," "Mommit," and monthly "BUMP Groups," ie, "January 2023 BUMP Group" on Reddit; Multimedia Appendices 1-4).

Paid Social Media Outreach (Social Media Recruitment)

A digital recruitment flier was posted on 4YouandMe's personal Instagram profile, and its visibility was boosted via Instagram's paid promotion feature, which allows you to select a "goal" (ie, driving traffic to a specific website). The total spend for paid social media recruitment was US $1939.30 over several paid Instagram advertisements.

Interested participants in the BUMP study were able to click on the advertisement and were then guided directly to complete the study interest form, hosted on the secure electronic data capture platform, REDCap (Research Electronic Data Capture; Vanderbilt University). The advertisement parameters for the paid social media outreach method included women, aged 18-42 years, currently living in the United States; and women, aged

XSL•FO
RenderX

18-36 years, living in the lowest 10% of United States zip codes by per capita income.

Community-Based Recruitment or Outreach (Community Partnerships)

Community-based partnerships were formed with the Moses/Weitzman Health System, a large federally qualified health center that works to deliver comprehensive health care services to those in need throughout the state of Connecticut. Partnership costs were covered by an agreed-upon fee for collaboration between Moses/Weitzman Health System and 4YouandMe. Medical assistants at three locations (Meriden, Middletown, and Clinton) that provide prenatal and postnatal care were educated by the BUMP study staff on the study overview, goals, eligibility requirements, wearable DHT use, and participant expectations. Moses/Weitzman Health System staff members were also provided with printed recruitment materials for the BUMP study.

Eligible potential participants were informed about the BUMP study by Moses/Weitzman Health System staff during one of their initial prenatal visits. Potential participants who expressed interest in the BUMP study were given a printed flier, tailored to the outreach of this specific community, with a scannable QR code linking interested participants to the study interest form, as well as contact information for study staff available to answer any questions before providing any personal information (Multimedia Appendices 5-6).

Other Recruitment Sources (Other)

Additional recruitment for the 4YouandMe direct outreach method was completed by other avenues such as referrals from current BUMP study participants to friends, family, and colleagues; general word of mouth about the BUMP study; articles published about 4YouandMe and the BUMP study on the Oura Ring blog and in STAT Magazine; and other sources such as Google and ClinicalTrials.gov.

Statistical Analysis

Engagement, for the purpose of this analysis, is defined as any interaction by a potential participant with the BUMP study material such as clicking on an advertisement or filling out a study interest form. Engagement with study advertisements from social media sites was tracked by Google Analytics, and REDCap was used to monitor completed study interest forms and collect additional information regarding recruitment channels. Due to the nature of recruitment conducted of genetic testing service subscribers via Sema4's patient portal, engagement was unable to be calculated for this group. Enrollment is defined as any participant who signed the informed consent and officially enrolled into the study, for any duration of time. In contrast, retention refers to participants who completed the full study period through birth or at least 9 months.

Engagement, enrollment, and retention rates were stratified by different recruitment channels and race. Sociodemographic characteristics including age, race or ethnicity, and income per capita (US $) by zip code were compared across different recruitment channels. Enrolled participants were also asked to share their home zip code upon enrollment. Participant's zip codes were used to calculate the distribution of per capita income of enrolled study participants, stratified by their recruitment method. Income data were provided by the United States Census Bureau's 2020 American Community Survey [30]. All analyses were conducted using Microsoft Excel (version 16.0).

Ethical Considerations

Human subjects research for the BUMP study was conducted in accordance with all applicable ethical guidelines and received full approval from the institutional review board, Advarra (Pro00047893). Participant confidentiality was maintained via deidentified data, secure data storage, limiting access of identifiable information to authorized members of the research team, and following clearly outlined privacy measures in the electronic consent (eConsent) document.

The BUMP study app served as the participants' hub for all data collection and study-related communication. Participants were instructed to download the app from the App Store (Apple Inc) or Google Play Store (Android). After downloading the study app, participants were walked through the study onboarding process, beginning with completing an eConsent form. ResearchKit's eConsent framework was used (for iOS) that includes brief summary screens with a link out to a "learn more" section to convey the consent narrative in a more digestible way. Participants were then provided with the eConsent in its entirety for their review and given a brief comprehension quiz before digitally signing the eConsent form. Participants immediately receive a digital, signed copy of the eConsent to their provided email address [29].

BUMP study participants were financially compensated in accordance with their engagement in daily surveys and active task completions on the BUMP study app. Participants received daily reward points that were tallied at the end of the study, and financial compensation was awarded proportionate to their accumulated study points for task completion. Participants were able to earn up to US $25 per month of participation in the BUMP study, for up to 12 months (including the 9-month study period and 3-month postpartum observation period). Participants who completed the study were also offered to keep their wearable devices—the Oura smart ring and Garmin smartwatch.

Further details of the BUMP study are described in external publications [29].

Results

Engagement, Enrollment, and Retention

Across all recruitment channels, 524 pregnant participants were enrolled in the BUMP study: 379 (72.3%) completed the full study period, 384 (73.3%) participants were recruited via Sema4's portal for genetic testing service subscribers, and 140 (26.7%) participants were recruited to 4YouandMe directly from the other recruitment approaches. Participants recruited via Sema4's patient portal (genetic testing service subscribers) had similar rates of study retention as those recruited via unpaid social media (Reddit and Facebook) and paid social media

XSL•FO
RenderX

(Instagram). The Moses/Weitzman Health System had the lowest percentage of study retention.

Genetic testing service subscriber recruitment was conducted via Sema4's patient portal for genetic testing service subscribers. Engagement data for this group is unknown. Of those who responded to an initial email with information about the BUMP study, 654 participants completed a screening call with a study coordinator. Of this group, 64.1% (419/654) met eligibility criteria, and 91.7% (384/654) of eligible participants enrolled in the study, resulting in 73.2% (281/654) of participants retained for the full study. (Figure 2).

Figure 2. Engagement, enrollment, and retention for the Better Understanding the Metamorphosis of Pregnancy (BUMP) study from outreach via genetic testing service subscribers (Sema4).

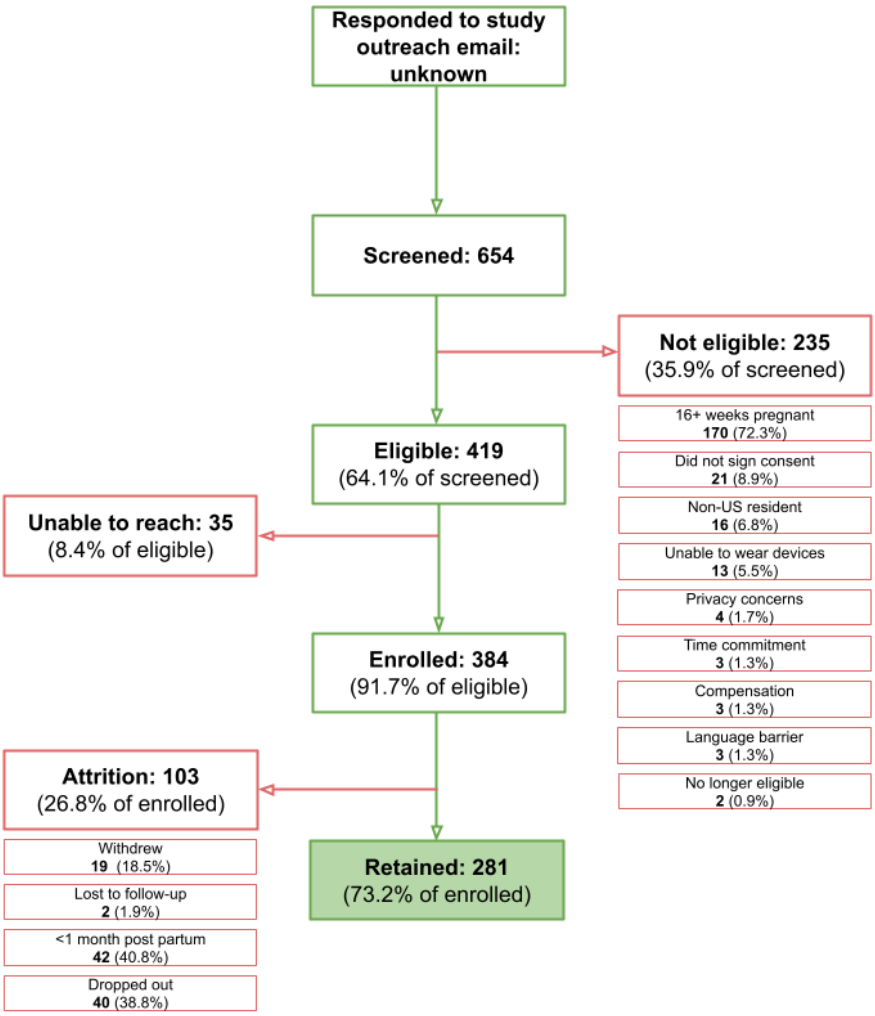

Social media recruitment of participants was conducted via unpaid advertisements (Reddit and Facebook) and paid advertisements (Instagram). Of those who engaged with an unpaid ad posted on Reddit and Facebook, 74.9% (167/223) completed a study interest form. Of this group, 31.1% (52/167) met eligibility criteria, and 100% (52/52) of eligible participants enrolled in the study, resulting in 67.3% (35/52) retained by the study's end (Figure 3).

XSL•FO
RenderX

Figure 3. Engagement, enrollment, and retention for the Better Understanding the Metamorphosis of Pregnancy (BUMP) study from outreach via unpaid social media recruitment (Reddit and Facebook).

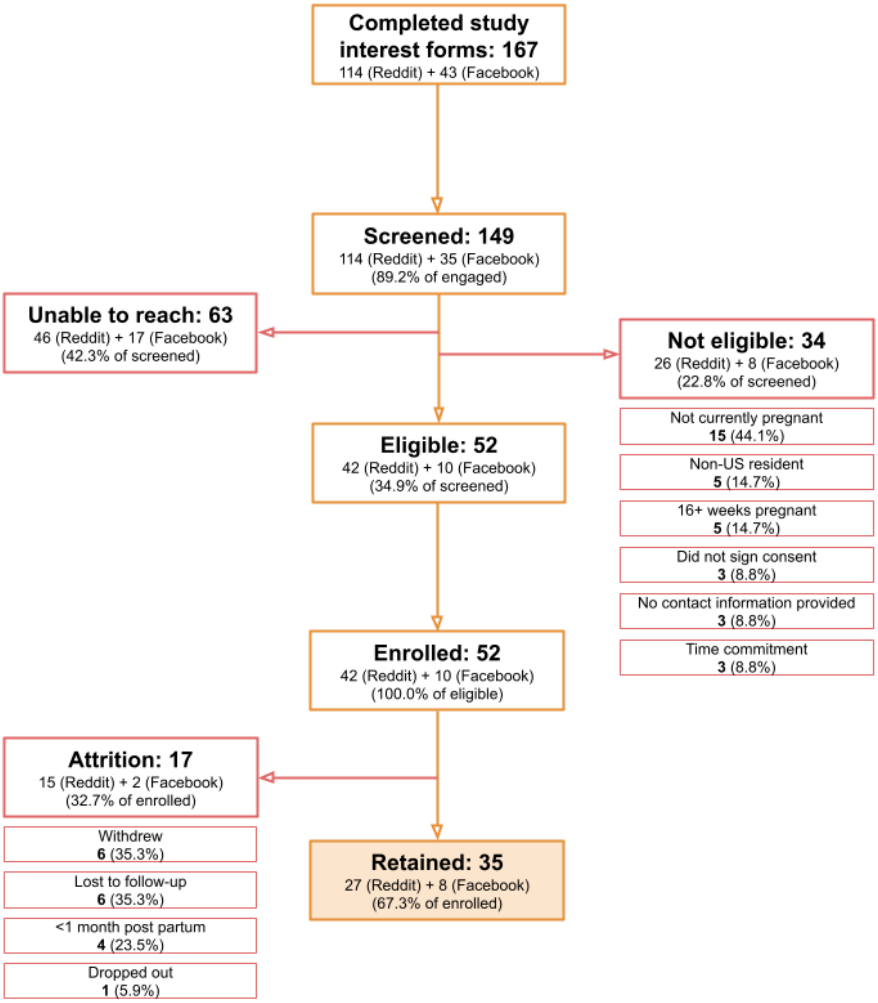

For individuals recruited from paid social media advertisements on Instagram, 321 individuals completed study interest forms; of those, 85 (26.5%) were not eligible for enrollment as a result of a studywide enrollment pause due to high recruitment volume. These individuals were notified of the enrollment pause but were not recontacted for further enrollment. Of the remaining 71 eligible participants, 70 (98.6%) enrolled in the study, and 74.3% (52/70) of those enrolled were retained (Figure 4).

XSL•FO
RenderX

Figure 4. Engagement, enrollment, and retention for the Better Understanding the Metamorphosis of Pregnancy (BUMP) study from outreach via paid social media recruitment (Instagram).

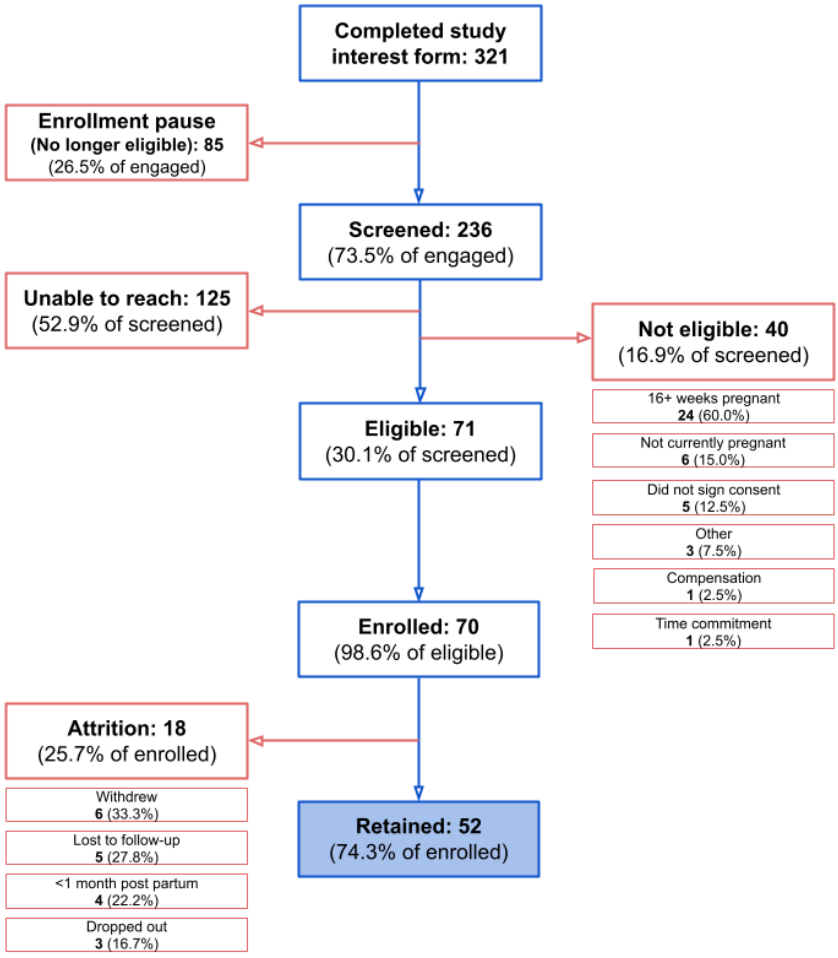

Of participants recruited via in-person visits with their clinician at a community health center providing affordable prenatal care (Moses/Weitzman Health System), 107 met eligibility criteria and were able to be approached on-site by staff. Of those approached, 24.3% (26/107) engaged with study material by completing the study interest form. A total of 76.9% (20/26) of participants did not enroll for a variety of reasons, such as being unable to contact (n=10), having more than 16 weeks pregnant (n=4), language barrier (n=3: n=2 Spanish and n=1 Haitian Creole), having no personal smartphone (n=1), not currently pregnant (n=1), or did not sign the consent (n=1). In total, 1 participant was no longer eligible due to the previously mentioned studywide enrollment pause, 5 participants ultimately enrolled in the study, and 40% (2/5) were retained—the lowest retention rate of any used recruitment method (Figure 5).

Over a 25-week period of active advertisements on both paid and unpaid social media channels (excluding community partnerships and other methods), 491 study interest forms were completed (the total from all 4YouandMe direct recruitment outlets was 594). This resulted in 122 enrolled participants (the total from 4YouandMe direct recruitment was 140) via paid and unpaid social media recruitment into the BUMP study (Table 1).

XSL•FO
RenderX

Figure 5. Engagement, enrollment, and retention for the Better Understanding the Metamorphosis of Pregnancy (BUMP) study from outreach via the Moses/Weitzman Health System.

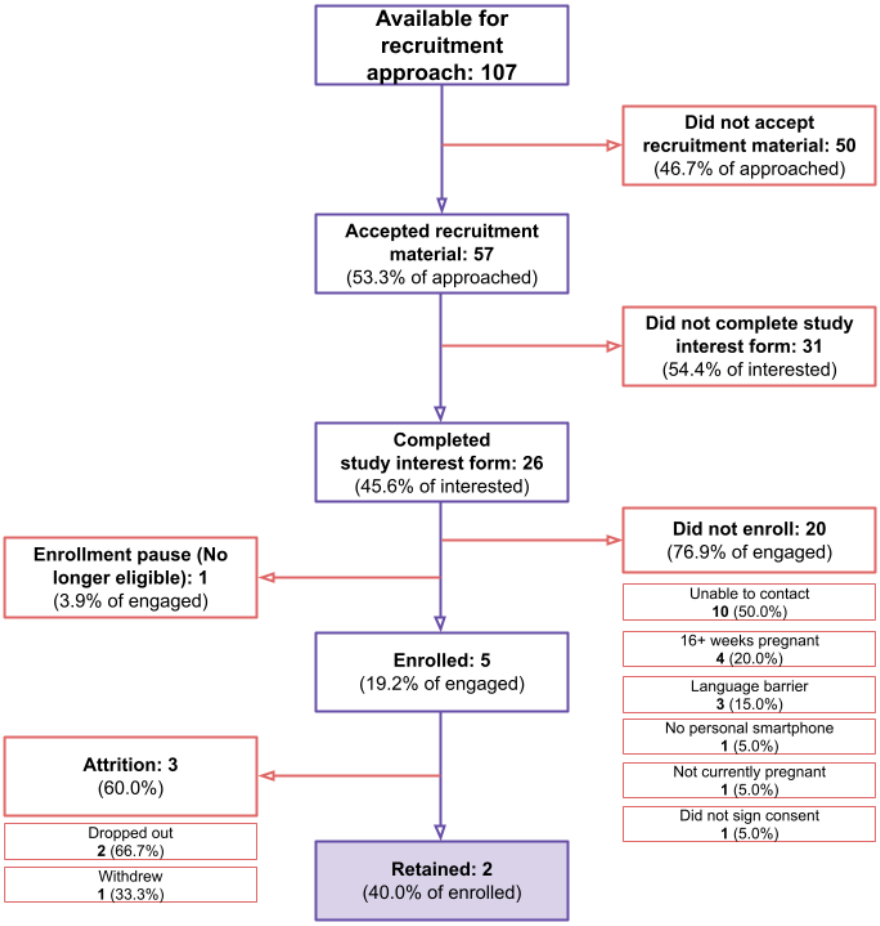

Table 1. Engagement and enrollment characteristics across different social media channels.

	Unpaid advertisements (Reddit)	Unpaid advertisements (Facebook)	Paid advertisements (Instagram)
Ad reach, n	__[a]	—	97,475
Website or advertisement taps, n	197	26	1534
Completed study interest forms, n	124	43	324
Participants enrolled, n	42	10	70
Start and end dates	April 5, 2022, to September 8, 2022	May 18, 2022, to September 8, 2022	September 2, 2022, to October 19, 2022

[a]Not applicable.

Characteristics of the Cohort

The majority of enrolled participants in the BUMP study were between the ages of 26-35 years (363/524, 69.3%) and White non-Hispanic (381/524, 72.7%; Table 2). For those that engaged with materials via 4YouandMe direct recruitment, participants were asked “Where did you hear about the BUMP study?” and were also given the option to self-identify both their race and ethnicity. Of participants that responded via each recruitment outlet (Reddit: n=124, Moses/Weitzman Health System: n=26, and Instagram: n=321), Reddit had the highest percentage of White non-Hispanic potential participants (98/124, 79%), while the Moses/Weitzman Health System had the lowest percentage of White non-Hispanic potential participants (7/26, 26.9%) and the highest relative percentage of Black potential participants (5/26, 19.2%), followed by Instagram (32/321, 9.9%; Table 3).

XSL•FO
RenderX

Table 2. Characteristics of the overall cohort among those who enrolled and were retained in the BUMP[a] study.

Characteristics	Genetic testing service subscriber recruitment (Sema4)		4YouandMe direct recruitment (social media, community partnerships, and other)		The BUMP study (all recruitment methods)	
	Enrolled (n=384)	Retained (n=281)	Enrolled (n=140)	Retained (n=98)	Enrolled (n=524)	Retained (n=379)
Age group (years), n (%)						
18-25	18 (4.7)	10 (3.6)	3 (2.1)	3 (3.1)	21 (4)	13 (3.4)
26-35	251 (65.4)	187 (66.6)	112 (80)	77 (78.6)	363 (69.3)	264 (69.7)
36-45	115 (29.9)	84 (29.9)	25 (21.7)	18 (18.4)	140 (26.7)	102 (26.9)
Race, n (%)						
Black	15 (3.9)	10 (3.6)	6 (4.3)	4 (4.1)	21 (4)	14 (3.7)
White (non-Hispanic)	274 (71.4)	213 (75.8)	107 (76.4)	77 (78.6)	381 (72.7)	290 (76.5)
White (Hispanic)	4 (1)	1 (0.4)	5 (3.6)	3 (3.1)	9 (1.7)	4 (1.1)
Asian	35 (9.1)	26 (9.3)	8 (5.7)	5 (5.1)	43 (8.21)	31 (8.2)
More than one race	18 (4.7)	13 (4.6)	7 (5)	4 (4.1)	25 (4.8)	17 (4.5)
Unknown or not reported	38 (9.9)	18 (6.4)	7 (5)	5 (5.1)	45 (8.6)	23 (6.1)
Ethnicity, n (%)						
Hispanic or Latino	12 (3.1)	6 (2.1)	10 (7.1)	6 (6.1)	22 (4.2)	12 (3.2)
Not Hispanic or Latino	372 (96.9)	275 (97.9)	128 (91.4)	91 (92.9)	500 (95.4)	366 (96.6)
Unknown or not reported	0 (0)	0 (0)	2 (1.4)	1 (1)	2 (0.4)	1 (0.3)

[a]BUMP: Better Understanding the Metamorphosis of Pregnancy.

XSL•FO
RenderX

Table 3. Characteristics of the cohort among those who engaged with recruitment materials, enrolled, and were retained in the BUMP[a] study by recruitment method[b].

Characteristics	Genetic testing service subscriber recruitment		4YouandMe direct recruitment											
	Sema4		Instagram			Facebook			Reddit			Moses/Weitzman Health System		
	Enrolled (n=384)	Retained (n=281)	Engaged (n=321)	Enrolled (n=70)	Retained (n=52)	Engaged (n=43)	Enrolled (n=10)	Retained (n=8)	Engaged (n=124)	Enrolled (n=42)	Retained (n=27)	Engaged (n=26)	Enrolled (n=5)	Retained (n=2)
Age group (years), n (%)														
18-25	18 (4.7)	10 (3.6)	40 (12.5)	1 (1.4)	1 (1.9)	4 (9.3)	1 (10)	1 (12.5)	5 (4)	1 (2.4)	1 (3.7)	8 (30.8)	0 (0)	0 (0)
26-35	251 (65.4)	187 (66.6)	224 (69.8)	56 (80)	42 (80.8)	35 (81.4)	7 (70)	6 (75)	102 (82.3)	34 (80.9)	20 (74)	16 (61.5)	4 (80)	2 (100)
36-45	115 (29.9)	84 (29.9)	57 (17.8)	13 (18.6)	9 (17.3)	4 (9.3)	2 (20)	1 (12.5)	17 (13.7)	7 (16.7)	6 (22.2)	2 (7.7)	1 (20)	0 (0)
Race, n (%)														
Black	15 (3.9)	10 (66.7)	32 (9.9)	2 (2.9)	1 (50)	4 (9.3)	0 (0)	0 (0)	4 (3.2)	1 (2.4)	1 (3.7)	5 (19.2)	1 (20)	1 (50)
White (non-Hispanic)	274 (71.4)	213 (77.7)	215 (66.9)	52 (74.3)	40 (76.9)	28 (65.1)	9 (90)	7 (87.5)	98 (79)	34 (80.9)	22 (81.5)	7 (26.9)	2 (40)	1 (50)
White (Hispanic)	4 (1)	1 (25)	18 (5.6)	3 (4.3)	2 (66.7)	2 (4.7)	0 (0)	0 (0)	5 (4)	0 (0)	0 (0)	5 (19.2)	1 (20)	0 (0)
Asian	35 (9.1)	26 (74.3)	16 (4.9)	4 (4.29)	2 (66.7)	1 (2.3)	0 (0)	0 (0)	9 (7.3)	5 (11.9)	3 (11.1)	0 (0)	0 (0)	0 (0)
American Indian or Alaska Native	0 (0)	0 (0)	1 (0.3)	0 (0)	0 (0)	1 (2.3)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (3.9)	0 (0)	0 (0)
Native Hawaiian or Other Pacific Islander	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (2.3)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
More than one race	18 (4.7)	13 (72.2)	18 (5.6)	5 (7.1)	3 (60)	4 (9.3)	0 (0)	0 (0)	6 (4.8)	2 (4.8)	1 (3.7)	2 (7.7)	0 (0)	0 (0)
Unknown or not reported	38 (9.9)	18 (47.4)	21 (6.5)	5 (7.1)	4 (80)	2 (4.7)	1 (10)	1 (12.5)	2 (1.6)	0 (0)	0 (0)	6 (23)	1 (20)	0 (0)
Ethnicity, n (%)														
Hispanic or Latino	12 (3.1)	6 (2.1)	36 (11.2)	5 (7.1)	4 (7.7)	4 (9.3)	0 (0)	0 (0)	7 (5.7)	0 (0)	0 (0)	13 (50)	3 (60)	1 (50)
Not Hispanic or Latino	372 (96.9)	275 (97.9)	276 (85.9)	63 (90)	47 (90.8)	39 (90.7)	10 (100)	8 (100)	114 (91.9)	42 (100)	27 (100)	13 (50)	2 (40)	1 (50)
Unknown or not reported	0 (0)	0 (0)	9 (2.8)	2 (2.9)	1 (1.9)	0 (0)	0 (0)	0 (0)	3 (2.4)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)

[a]BUMP: Better Understanding the Metamorphosis of Pregnancy.

[b]Values for participants recruited by other methods are included in Multimedia Appendix 7.

Those recruited via Sema4 (genetic testing service subscribers) had the highest participant income per capita by zip code at nearly every income percentile (Figure 6). Participants recruited via community outreach partnerships (Moses/Weitzman Health System) had the lowest number of participants in the 50-99th percentiles of per capita income by zip code, studywide (Figure 6).

XSL•FO
RenderX

Figure 6. Distribution of income per capita (US $) by zip code of enrolled participants in the Better Understanding the Metamorphosis of Pregnancy (BUMP) study by recruitment method (genetic testing subscriber recruitment: n=384 and 4YouandMe direct recruitment: n=140).

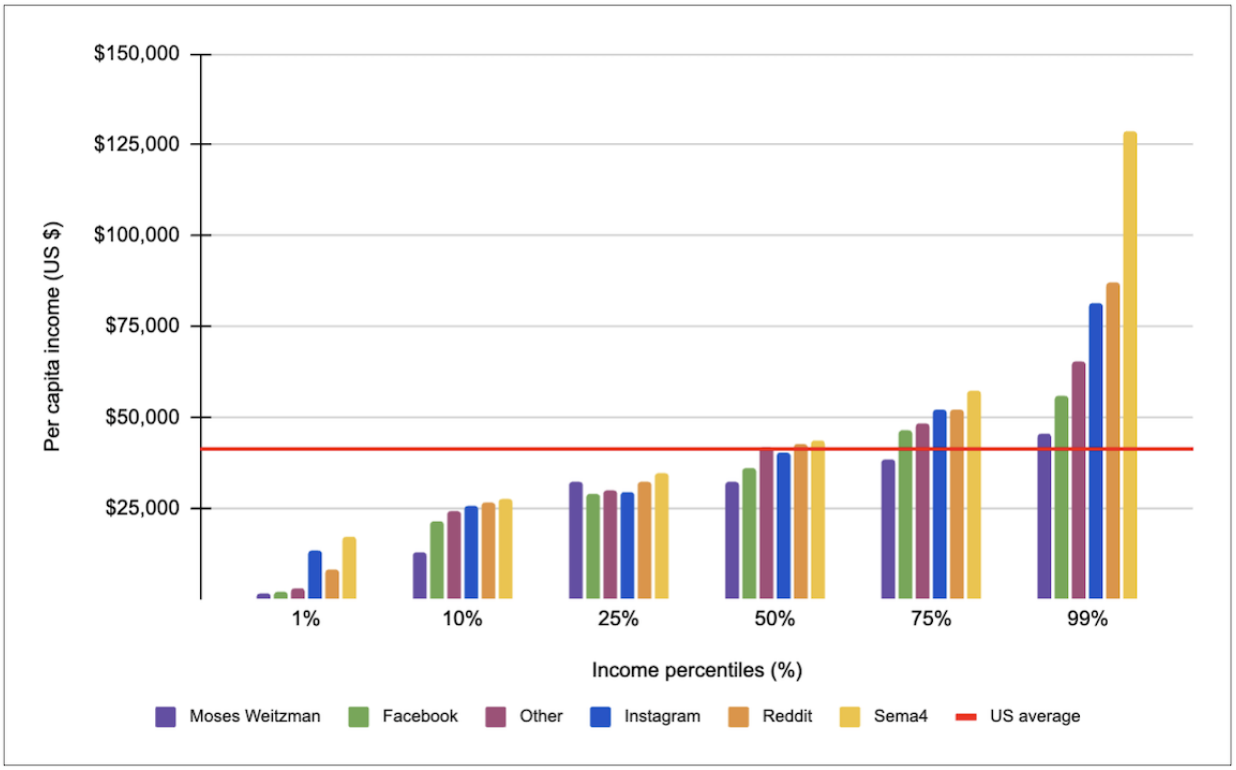

Discussion

Principal Results

The results of this analysis suggest that the use of social media (paid and unpaid advertisements) as a tool to engage, enroll, and retain URPs in digital health pregnancy research can be effective. Unpaid social media proved to be successful and low-cost, and achieved the highest engagement-to-enrollment rate among all approaches used by the 4YouandMe direct recruitment channel. However, maintaining successful recruitment on these platforms required significant sustained effort and time from the study team.

Paid social media ads on Instagram also effectively reached and engaged pregnant women as URP in digital health research. Running paid advertisements was a low burden on the study team, in comparison to unpaid advertisements, which required continued monitoring and engagement on behalf of the study team. Paid advertisements incurred a cost over 25 weeks; US $1939.30 was spent on Instagram advertisements, resulting in 70 enrollments, or US $27.70 per enrolled participant. This cost per enrolled participant is generally favorable, as previous systematic reviews of the costs of social media–based participant recruitment ranged from US $0.00 to US $517.00 per participant, compared to US $19.00 to US $777.00 for traditional recruitment methods [31]. Additional studies found a mean cost per participant of US $16.22 for Google ads and between US $13.12 and US $250.00 for other traditional methods. Data on the cost-effectiveness of paid social media ads for recruiting participants to research studies are severely limited [32].

Participants recruited via paid social media (Instagram) showed the highest overall retention rate as 74.3% (52/70) of participants enrolled completed the full study period (Table 3). This suggests that paid social media outreach was more effective in engaging and retaining study participants than other recruitment methods. This outreach method was even more effective than recruiting through genetic testing service subscribers (Sema4), where participants may have had a higher intrinsic motivation to participate and stay in the study due to potential previous interest in genetic testing and related research.

Interactions with participants recruited via Instagram during study onboarding revealed a distinct motivation to engage, enroll, and complete the full study period. Although not statistically quantified, this trend may be linked to specific user behavior within the Instagram-recruited cohort, or the platform's age demographics aligning closely with the study's target population. Among participants retained in the BUMP study recruited from Instagram, 80.8% (42/52) were aged 26-35 years, compared to 74% (20/27) from Reddit (Table 3). This aligns with broader trends, as 28.3% of the US Instagram users in 2024 were aged 25-34 years, while 44% of Reddit users in 2023 were aged 18-29 years [33,34]. Given that the mean age of first-time mothers in the United States is 27.4 years, Instagram's higher retention may reflect its unique demographic fit with the specific study population and subsequent increased study interest and engagement [35]. These observations remain speculative, as the available data cannot fully confirm the underlying reasons.

In addition, paid social media recruitment facilitated the widespread engagement and enrollment of URPs. Paid recruitment had a higher proportion of non-White, non-Hispanic

XSL•FO
RenderX

respondents who engaged with and enrolled in the BUMP study, in comparison to unpaid social media. Retention of non-White, non-Hispanic participants was low across all recruitment methods.

Recruitment via social media (including paid and unpaid recruitment) also provided access to and enrollment of a different socioeconomic population than those enrolled via genetic testing service subscriber (Sema4) recruitment. Participants recruited via genetic testing service subscribers (Sema4), which is associated with Mount Sinai Health System, reflect a unique group of individuals with means and access to genetic testing. This was shown in the enrolled participants from this group who had significantly higher per capita incomes, exceeding US $75,000 far surpassing the US national average income per capita [36].

In contrast, enrolled participants from social media recruitment and community-based outreach's highest average per capita incomes were US $74,771.15 and US $45,401.00, respectively (Figure 6). While data are limited, one study suggests that 46% of Instagram users and 23% of Reddit users have a reported household income under US $70,000.00 [37]. From this, we may infer that recruitment via social media (and community partnerships) also facilitated the increased enrollment of socioeconomic-based URPs.

Community-based partnerships for BUMP study recruitment revealed unique and informative data trends. Over a 24-week period at three Moses/Weitzman Health System clinics that serve a large URP, only 53.3% (57/107) accepted recruitment materials (Figure 5). The overall number of potential in-clinic approaches was limited by multiple factors. Time constraints with in-clinic visits and frequent no-shows hindered clinic staff from fully introducing the study to potential participants, thoroughly answering questions, or handing out recruitment material—further illustrating additional downsides of more traditional, in-person recruitment.

In addition, a significant number of potential participants were immediately excluded before the recruitment approach due to a Spanish language barrier, curtailing the overall number of eligible approaches. Of those remaining participants eligible for approach, some displayed immediate apprehension to study engagement upon simply hearing the word "study," and one participant explicitly reported "not wanting to feel like a guinea pig" to clinic staff." These responses show an intrinsic hesitancy to engage in research studies and a deep mistrust of research overall, even when the information is conveyed by familiar health care providers or their staff.

Community health center–recruited participants who indicated study interest also showcased population-specific barriers to study enrollment or participation. Despite accepting recruitment material, 15% (3/20) of participants did not express sufficient English language fluency to complete surveys and tasks entirely in English on the study app. A total of 50% (10/20) of participants who engaged with study materials were difficult to reach for a further onboarding call (Figure 5). While many of these participants were responsive via text message, BUMP study staff reported difficulty completing study onboarding via phone with participants, citing their extensive and limiting work schedules. Other significant, population-specific barriers to study enrollment included those who did not own a personal smartphone.

Despite dedicated efforts from the BUMP study team and Moses/Weitzman Health System staff, recruiting, engaging, and enrolling participants from community health centers proved uniquely challenging compared to other methods. Of the potential participants who received recruitment materials, only 8.8% (5/57) ultimately enrolled in the study (Figure 5). This low enrollment underscores many of the well-documented challenges of maintaining consistent enrollment among URPs when relying solely on community outreach channels. It also highlights opportunities for overcoming these challenges in future studies.

Strategies to address logistical and participant-voiced barriers may include conducting digital recruitment follow-ups with no-show potential participants, enhancing language inclusivity in recruitment and study app materials, incorporating text messaging for primary communication, providing smartphones for study participant use, and embedding opportunities for recruitment outreach into clinical care. Seamless communication between researchers, providers, and medical staff can also aid in addressing participants' needs in real time as researchers work with clinic staff, using feedback from potential participants, to address population-specific concerns and develop tailored language and approaches; combating mistrust head-on by leading with transparency; and creating a culture of embedding options for research participation into clinical care.

Limitations

Due to the methodology of recruitment and outreach for the BUMP study via genetic testing service subscribers (Sema4), overall study interest and engagement were unable to be followed and analyzed in the same manner as 4YouandMe direct recruitment avenues. Similarly, demographic information (eg, race or ethnicity, socioeconomic status) was not provided for interested or potential participants from this group, limiting the comparison and further analysis of engagement with study materials by various URPs between recruitment of genetic testing service subscribers and social media and other methods.

Additional limitations were present in the analysis of the recruitment potential for the general pregnant population via social media platforms. We were unable to discern how many of the total reached population was actually pregnant or study eligible; this was particularly relevant for paid social media outreach and recruitment. For paid social media outreach, we relied on inputting the lowest 10% of US zip codes by per capita income to increase our socioeconomic diversity; however, it is unknown if zip codes are too narrow of a parameter for focused Instagram advertisements, ultimately impacting an in-depth review of engagement statistics.

True analysis of study engagement to enrollment statistics was further limited by the presence of a study enrollment pause, due to the high volume of completed study interest forms and potential participants. This enrollment pause rendered 132 participants no longer eligible for study participation. Without this pause, the engagement-to-enrollment rate may have reached

XSL•FO
RenderX

45.8% (272/594). Recruitment for the BUMP study via the various 4YouandMe direct approaches was not conducted for the entire study duration, impacting the overall ability to assess engagement to enrollment effectiveness. Finally, the small size of this study sample (n=524) indicates the need for larger, similar studies to further validate the trends observed.

Conclusions

Social media and virtual outreach are both emerging and exciting tools that researchers can use to engage and recruit various URPs to participate in digital health research studies, specifically about pregnancy. Paid social media advertisements provide unique, innovative, and low-time-burden opportunities for researchers to engage and interact with a larger volume of URPs in comparison to traditional recruitment methods.

Though commonly perceived as the gold standard for outreach to URPs, community-based partnerships, even when combined with other outreach methods, did not provide a smooth, consistent path to outreach, engagement, or retention of URPs in digital health research. This lower engagement and retention underscore the need for further tailored strategies of addressing additional barriers to sustained participation in digital health studies for URPs to better bridge the digital divide and engage URPs.

By embracing novel recruitment outreach methods such as via social media platforms, researchers have a unique opportunity to refine often stagnant, traditional recruitment strategies and adapt to an evolving digital landscape. The integration of community-driven recruitment approaches with tailored digital interventions that uniquely address the needs of URPs will create more inclusive and equitable opportunities for participation in digital health studies and foster more representative and impactful research that may ultimately improve health outcomes for URPs and others alike.

Acknowledgments

The Better Understanding the Metamorphosis of Pregnancy (BUMP) study was funded by 4YouandMe and Sema4 along with supplemental in-kind contributions from coalition partners (Evidation Health, Vector Institute, Cambridge Cognition, and Bodyport).

4YouandMe primarily conducted the BUMP study and carried out all 4YouandMe direct recruitment (including paid and social media, community partnerships with the Moses/Weitzman Health System, and others [eg, word of mouth, press, and ClinicalTrials.gov]). Sema4 facilitated genetic testing subscriber recruitment. Evidation Health was responsible for data collection and maintenance. Cambridge cognition provided active and cognitive tasks used in the BUMP study app. The Vector Institute aided in poststudy data analysis.

In addition, we would like to express our gratitude to the Moses/Weitzman Health System for their partnership in helping to make this research possible—particularly, Amy Gagliardi and staff at the Middletown, Clinton, and Meriden sites for their on-the-ground efforts. Additional thanks to Ridley McKenzie and Lesley Harris from the 4YouandMe team for their study support, as well as all of the BUMP study participants for their time, effort, and contributions to this research.

Data Availability

The datasets generated or analyzed during this study will be hosted in the public domain and available on the Synapse platform at Sage Bionetworks [29]. Please check the 4YouandMe website or contact the corresponding author for an update on when study data will be available and how to access it.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The BUMP study recruitment flier for outreach via 4YouandMe direct recruitment approaches designed for long-form social media sites (eg, Reddit). BUMP: Better Understanding the Metamorphosis of Pregnancy.
[PNG File , 170 KB - formative_v9i1e68093_app1.png]

Multimedia Appendix 2
The BUMP study recruitment flier for outreach via 4YouandMe direct recruitment approaches designed for long-form social media sites (eg, Reddit). BUMP: Better Understanding the Metamorphosis of Pregnancy.
[PNG File , 641 KB - formative_v9i1e68093_app2.png]

Multimedia Appendix 3
The BUMP study recruitment flier for outreach via 4YouandMe direct recruitment approaches designed for long-form social media sites (text version), designed for long-form social media sites with specific forums that did not allow images (eg, Reddit). BUMP: Better Understanding the Metamorphosis of Pregnancy.

XSL•FO
RenderX

[PNG File , 271 KB - formative_v9i1e68093_app3.png]

Multimedia Appendix 4
The BUMP study recruitment flier for outreach via 4YouandMe direct recruitment approaches designed for long-form social media sites, tailored for image-dominant social media sites (eg, Instagram). BUMP: Better Understanding the Metamorphosis of Pregnancy.
[PNG File , 163 KB - formative_v9i1e68093_app4.png]

Multimedia Appendix 5
The BUMP study recruitment flier for outreach via 4YouandMe direct recruitment approach, tailored to community health partnerships (Moses/Weitzman Health System). BUMP: Better Understanding the Metamorphosis of Pregnancy.
[PNG File , 2377 KB - formative_v9i1e68093_app5.png]

Multimedia Appendix 6
The BUMP study recruitment flier for outreach via 4YouandMe direct recruitment approach (alternate flier), tailored to community health partnerships (Moses/Weitzman Health System). BUMP: Better Understanding the Metamorphosis of Pregnancy.
[PNG File , 528 KB - formative_v9i1e68093_app6.png]

Multimedia Appendix 7
Characteristics of the cohort among those who engaged (ENG) with recruitment materials, enrolled (ENR), and were retained (RTD) in the BUMP study for participants recruited via other methods. BUMP: Better Understanding the Metamorphosis of Pregnancy.
[PNG File , 285 KB - formative_v9i1e68093_app7.png]

References

1. Scharff DP, Mathews KJ, Jackson P, Hoffsuemmer J, Martin E, Edwards D. More than Tuskegee: understanding mistrust about research participation. J Health Care Poor Underserved 2010 Aug;21(3):879-897 [FREE Full text] [doi: 10.1353/hpu.0.0323] [Medline: 20693733]
2. Staccini P, Lau AYS. Consuming health information and vulnerable populations: factors of engagement and ongoing usage. Yearb Med Inform 2022 Aug;31(1):173-180 [FREE Full text] [doi: 10.1055/s-0042-1742549] [Medline: 36463876]
3. Swartz T, Palermo A, Masur S, Aberg J. The science and value of diversity: closing the gaps in our understanding of inclusion and diversity. J Infect Dis 2019 Aug 20;220(220 Suppl 2):S33-S41 [FREE Full text] [doi: 10.1093/infdis/jiz174] [Medline: 31430380]
4. NIH grants and funding acronym and glossary list. National Institutes of Health. 2022. URL: https://toolkit.ncats.nih.gov/glossary/underrepresented-population/ [accessed 2023-11-10]
5. Dauphin C, Clark N, Cadzow R, Saad-Harfouche F, Rodriguez E, Glaser K, et al. #BlackBreastsMatter: process evaluation of recruitment and engagement of pregnant African American women for a social media intervention study to increase breastfeeding. J Med Internet Res 2020 Aug 10;22(8):e16239 [FREE Full text] [doi: 10.2196/16239] [Medline: 32773377]
6. Farmer DF, Jackson SA, Camacho F, Hall MA. Attitudes of African American and low socioeconomic status White women toward medical research. J Health Care Poor Underserved 2007 Feb;18(1):85-99. [doi: 10.1353/hpu.2007.0008] [Medline: 17337800]
7. Washington V, Franklin JB, Huang ES, Mega JL, Abernethy AP. Diversity, equity, and inclusion in clinical research: a path toward precision health for everyone. Clin Pharmacol Ther 2023 Mar;113(3):575-584. [doi: 10.1002/cpt.2804] [Medline: 36423203]
8. UyBico SJ, Pavel S, Gross CP. Recruiting vulnerable populations into research: a systematic review of recruitment interventions. J Gen Intern Med 2007 Jun;22(6):852-863 [FREE Full text] [doi: 10.1007/s11606-007-0126-3] [Medline: 17375358]
9. Inclusion of women and minorities as participants in research involving human subjects. National Institutes of Health. 2022. URL: https://grants.nih.gov/policy/inclusion/women-and-minorities.htm [accessed 2023-07-30]
10. Yancey AK, Ortega AN, Kumanyika SK. Effective recruitment and retention of minority research participants. Annu Rev Public Health 2006;27:1-28. [doi: 10.1146/annurev.publhealth.27.021405.102113] [Medline: 16533107]
11. Shavers VL, Lynch CF, Burmeister LF. Racial differences in factors that influence the willingness to participate in medical research studies. Ann Epidemiol 2002 May;12(4):248-256. [doi: 10.1016/s1047-2797(01)00265-4] [Medline: 11988413]
12. Hoyert DL. Maternal mortality rates in the United States, 2021. NCHS Health E-Stats. 2023. URL: https://stacks.cdc.gov/view/cdc/124678 [accessed 2023-03-16]
13. Blehar MC, Spong C, Grady C, Goldkind SF, Sahin L, Clayton JA. Enrolling pregnant women: issues in clinical research. Womens Health Issues 2013 Jan;23(1):e39-e45 [FREE Full text] [doi: 10.1016/j.whi.2012.10.003] [Medline: 23312713]

XSL•FO
RenderX

14. Russomanno J, Patterson JG, Jabson Tree JM. Social media recruitment of marginalized, hard-to-reach populations: development of recruitment and monitoring guidelines. JMIR Public Health Surveill 2019 Dec 02;5(4):e14886 [FREE Full text] [doi: 10.2196/14886] [Medline: 31789598]
15. Ellington M, Connelly J, Clayton P, Lorenzo CY, Collazo-Velazquez C, Trak-Fellermeier MA, et al. Use of Facebook, Instagram, and Twitter for recruiting healthy participants in nutrition-, physical activity-, or obesity-related studies: a systematic review. Am J Clin Nutr 2022 Feb 09;115(2):514-533 [FREE Full text] [doi: 10.1093/ajcn/nqab352] [Medline: 34669955]
16. Leading social media websites in the United States as of August 2023, based on share of visits. Statista. 2023. URL: https://www.statista.com/statistics/265773/market-share-of-the-most-popular-social-media-websites-in-the-us/ [accessed 2023-11-15]
17. Number of social network users in selected countries in 2022 and 2027. Statista. 2023. URL: https://www.statista.com/statistics/278341/number-of-social-network-users-in-selected-countries/ [accessed 2023-12-30]
18. Social media fact sheet. Pew Research Center. 2022. URL: https://www.pewresearch.org/internet/fact-sheet/social-media/#panel-2fc5fff9-9899-4317-b786-9e0b60934bcf [accessed 2023-07-30]
19. Ginsburg GS, Picard RW, Friend SH. Key issues as wearable digital health technologies enter clinical care. N Engl J Med 2024 Mar 21;390(12):1118-1127. [doi: 10.1056/NEJMra2307160] [Medline: 38507754]
20. Chesser A, Burke A, Reyes J, Rohrberg T. Navigating the digital divide: a systematic review of eHealth literacy in underserved populations in the United States. Inform Health Soc Care 2016;41(1):1-19. [doi: 10.3109/17538157.2014.948171] [Medline: 25710808]
21. Mitchell SJ, Godoy L, Shabazz K, Horn IB. Internet and mobile technology use among urban African American parents: survey study of a clinical population. J Med Internet Res 2014 Jan 13;16(1):e9 [FREE Full text] [doi: 10.2196/jmir.2673] [Medline: 24418967]
22. Mitchell UA, Chebli PG, Ruggiero L, Muramatsu N. The digital divide in health-related technology use: the significance of race/ethnicity. Gerontologist 2019 Jan 09;59(1):6-14. [doi: 10.1093/geront/gny138] [Medline: 30452660]
23. James DCS, Harville C. eHealth literacy, online help-seeking behavior, and willingness to participate in mHealth chronic disease research among African Americans, Florida, 2014-2015. Prev Chronic Dis 2016 Nov 17;13:E156 [FREE Full text] [doi: 10.5888/pcd13.160210] [Medline: 27854421]
24. James DCS, Harville C, Whitehead N, Stellefson M, Dodani S, Sears C. Willingness of African American women to participate in e-Health/m-Health research. Telemed J E Health 2016 Mar;22(3):191-197 [FREE Full text] [doi: 10.1089/tmj.2015.0071] [Medline: 26313323]
25. George N, Reynolds S, de Long R, Kacica M, Ahmed R, Manganello J. Social media and Black maternal health: the role of health literacy and ehealth literacy. Health Lit Res Pract 2023 Jul;7(3):e119-e129 [FREE Full text] [doi: 10.3928/24748307-20230614-01] [Medline: 37463292]
26. Kasoju N, Remya NS, Sasi R, Sujesh S, Soman B, Kesavadas C, et al. Digital health: trends, opportunities and challenges in medical devices, pharma and bio-technology. CSIT 2023 Apr 11;11(1):11-30. [doi: 10.1007/s40012-023-00380-3]
27. Friend SH, Ginsburg GS, Picard RW. Wearable digital health technology. N Engl J Med 2023 Nov 30;389(22):2100-2101. [doi: 10.1056/NEJMe2303219] [Medline: 38048193]
28. Saeed SA, Masters RM. Disparities in health care and the digital divide. Curr Psychiatry Rep 2021 Jul 23;23(9):61 [FREE Full text] [doi: 10.1007/s11920-021-01274-4] [Medline: 34297202]
29. Goodday SM, Karlin E, Brooks A, Chapman C, Karlin DR, Foschini L, et al. Better Understanding of the Metamorphosis of Pregnancy (BUMP): protocol for a digital feasibility study in women from preconception to postpartum. NPJ Digit Med 2022 Mar 30;5(1):40 [FREE Full text] [doi: 10.1038/s41746-022-00579-9] [Medline: 35354895]
30. 2022 American Community Survey (ACS). United States Census Bureau. 2022. URL: https://www.census.gov/programs-surveys/acs/news/data-releases.html [accessed 2023-12-04]
31. Oudat Q, Bakas T. Merits and pitfalls of social media as a platform for recruitment of study participants. J Med Internet Res 2023 Oct 11;25:e47705 [FREE Full text] [doi: 10.2196/47705] [Medline: 37819692]
32. Tsaltskan V, Sanchez Baez R, Firestein GS. Cost-effectiveness of social media advertising as a recruitment tool: a systematic review and meta-analysis. J Clin Transl Sci 2023;7(1):e180 [FREE Full text] [doi: 10.1017/cts.2023.596] [Medline: 37745929]
33. Distribution of instagram users in the United States as of October 2024, by age group. Statista. URL: https://www.statista.com/statistics/398166/us-instagram-user-age-distribution/ [accessed 2024-12-04]
34. Percentage of U.S. adults who use Reddit as of September 2023, by age group. Statista. 2024. URL: https://www.statista.com/statistics/261766/share-of-us-internet-users-who-use-reddit-by-age-group/ [accessed 2024-12-04]
35. Osterman MJK, Hamilton BE, Martin JA. Births: final data for 2022. National Vital Statistics Reports 2024;73(2):1-39. [doi: 10.15620/cdc:145588]
36. 2023 American Community Survey (ACS). United States Census Bureau. 2023. URL: https://www.census.gov/programs-surveys/acs/news/data-releases.html [accessed 2023-12-04]
37. How use of online platforms – such as Facebook, Instagram or TikTok – differs among some U.S. demographic groups. Pew Research Center. 2024. URL: https://www.pewresearch.org/internet/2024/01/31/americans-social-media-use/pi_2024-01-31_social-media-use_00_04-png [accessed 2024-12-04]

XSL•FO
RenderX

Abbreviations

BUMP: Better Understanding the Metamorphosis of Pregnancy
DHTs: digital health technologies
eConsent: electronic consent
NIH: National Institutes of Health
REDCap: Research Electronic Data Capture
URPs: Underrepresented populations

Edited by A Mavragani; submitted 29.10.24; peer-reviewed by M Gasmi , C Banko, C Nnajiofor; comments to author 17.11.24; revised version received 11.12.24; accepted 25.02.25; published 15.04.25.

Please cite as:
Harry C, Goodday S, Chapman C, Karlin E, Damian AJ, Brooks A, Boch A, Lugo N, McMillan R, Tempero J, Swanson E, Peabody S, McKenzie D, Friend S
Using Social Media to Engage and Enroll Underrepresented Populations: Longitudinal Digital Health Research
JMIR Form Res 2025;9:e68093
URL: https://formative.jmir.org/2025/1/e68093
doi:10.2196/68093
PMID:

©Christiana Harry, Sarah Goodday, Carol Chapman, Emma Karlin, April Joy Damian, Alexa Brooks, Adrien Boch, Nelly Lugo, Rebecca McMillan, Jonell Tempero, Ella Swanson, Shannon Peabody, Diane McKenzie, Stephen Friend. Originally published in JMIR Formative Research (https://formative.jmir.org), 15.04.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Influence of Medical Expertise and Information Search Skills on Medical Information Searching: Comparative Analysis From a Free Data Set

Aline Chevalier[1*], PhD, Prof Dr; Cheyenne Dosso[2*], PhD

[1]Laboratoire Cognition, Langues, Langage, Ergonomie (Centre National de la Recherche Scientifique UMR5263) Maison de la Recherche, Université de Toulouse, Toulouse, France

[2]Laboratoire d'anthropologie et de psychologie cliniques, cognitives et sociales (UPR7278), Université Côte d'Azur, Nice, France

[*]all authors contributed equally

Corresponding Author:
Aline Chevalier, PhD, Prof Dr
Laboratoire Cognition, Langues, Langage, Ergonomie (Centre National de la Recherche Scientifique UMR5263) Maison de la Recherche
Université de Toulouse
5 allées Machado
Toulouse, 31058
France
Phone: 33 561503531
Email: aline.chevalier@univ-tlse2.fr

Abstract

Background: Nowadays, the internet has become the primary source of information for physicians seeking answers to medical questions about their patients before consulting colleagues. However, many websites provide low-quality, unreliable information that lacks scientific validation. Therefore, physicians must develop strong information search skills to locate relevant, accurate, and evidence-based content. However, previous studies have shown that physicians often have poor search skills and struggle to find information on the web, which may have detrimental consequences for patient care.

Objective: This study aims to determine how medical students and residents searched for medical information on the internet, the quality of the web resources they used (including their nature and credibility), and how they evaluated the reliability of these resources and the answers they provided. Given the importance of domain knowledge (in this case, medicine) and information search skills in the search process, we compared the search behaviors of medical students and residents with those of computer science students. While medical students and residents possess greater medical-related knowledge, computer science students have stronger information search skills.

Methods: A total of 20 students participated in this study: 10 medical students and residents, and 10 computer science students. Data were extracted from a freely accessible data set in accordance with FAIR (Findable, Accessible, Interoperable, and Reusable) principles. All participants searched for medical information online to make a diagnosis, select a treatment, and enhance their knowledge of a medical condition—3 primary activities they commonly perform. We analyzed search performance metrics, including search time, the use of medical-related keywords, and the accuracy of the information found, as well as the nature and credibility of web resources used by medical students and residents compared with computer science students.

Results: Medical students and residents provided more accurate answers than computer science students without requiring additional time. Their medical expertise also enabled them to better assess the reliability of resources and select high-quality web sources, primarily from hospital websites. However, it is noteworthy that they made limited use of evidence-based tools such as PubMed.

Conclusions: Although medical students and residents generally outperformed computer science students, they did not frequently use evidence-based tools. As previously observed, they may avoid databases due to the risk of encountering too many irrelevant articles and difficulties in applying appropriate filters to locate relevant information. Nevertheless, clinical and practical evidence-based medicine plays a crucial role in updating physicians' knowledge, improving patient care, and enhancing physician-patient relationships. Therefore, information search skills should be an integral part of medical education and continuing professional development for physicians.

XSL•FO
RenderX

(JMIR Form Res 2025;9:e62754) doi:10.2196/62754

KEYWORDS

information searching; credibility; internet; medicine; information search skills

Introduction

Internet as a Tool for Medical Information Searching by Physicians

Nowadays, searching for information on the web is an integral part of a physician's professional activity. While colleagues were the primary source of information in 2006 [1], more recently, the internet has become the first source before consulting colleagues when physicians have medical questions about patients [2,3].

Efficient information searching has various positive outcomes for physicians and, by extension, their patients [2]. It may improve patient care [4,5], clinical decision-making [5], and physicians' knowledge [5,6]. It also reduces delays in diagnosis, treatment initiation, and modification [5], as well as potential medical errors and unnecessary medical examinations [6].

More precisely, physicians, especially younger ones, use the internet to search for information for various purposes [1]: (1) finding information and solutions to specific patient diagnostic issues (reported by 33.7%-57% of respondents in the previous study [1]); (2) obtaining new information on diseases (20.4%); and (3) staying informed about the latest research on specific topics and general matters, such as diseases (27%), new therapies and treatments (8.8%), and drug dosing (5.6%), as also shown by [7-9].

Concerning information related to treatment and diagnosis, physicians take between 2 and 32 minutes to find answers on the internet, depending on the nature of the information sought (eg, side effects, drug dosing, new therapies) [2]. Nevertheless, most studies indicate that physicians spend less than 10 minutes searching for answers and less than 2 minutes during patient consultations [2]. During consultations, they primarily search for information on drug dosing and its potential side effects or to show images to patients [10].

Difficulties in Finding Relevant Web Resources and Its Consequences

Although physicians reported finding information in most cases (>50% of searches for 70% of physicians) and over 60% were confident in their ability to find the information they needed on the internet [2], they still face various difficulties and barriers in information searching. For instance, a survey of general practitioners identified the main obstacles as a lack of information search knowledge or skills, clinical practice conditions (time pressure or patient-related concerns), information-related issues (information overload, quality concerns, or low relevance), and language barriers [11]. It is noteworthy that, occasionally, reliance on internet-supplied information can lead to clinical errors, as physicians may respond to and process such information—even when it contradicts their preexisting knowledge [12]—regardless of its accuracy. While the use of web resources can enhance physicians' clinical decisions [6], occasional discrepancies in the quality and reliability of the information found may introduce errors in the decision-making process. Outpatient diagnostic error rates have been estimated at 5%, affecting more than 12 million individuals per year, while inpatient diagnostic errors range from 6% to 7%, with 20% of initial diagnoses being modified [13]. Efficient use of information access systems (eg, search engines) can positively impact patient care, but misuse of the internet may have the opposite effect. Therefore, physicians and medical students must develop accurate information search skills (ie, procedural knowledge of tools such as Google [Alphabet Inc] or PubMed [National Institutes of Health]) to use these systems effectively within the specificities of the medical domain.

Previous studies have reported a lack of information search skills, requiring technical support for information retrieval [1], unawareness of accessible sources [2], difficulties in accessing reliable and up-to-date medical information, and challenges in developing efficient search strategies [4]. Additionally, on the internet, physicians face the challenge of processing large amounts of information in a very short time [14].

Previous studies have shown that well-designed, easy-to-navigate websites are generally considered more reliable than those with a confusing or complex interface [15]. The first websites listed by search engine tools are often perceived as more credible than those appearing later [16]. However, neither a well-designed website nor a high search engine ranking guarantees that the information provided is credible and reliable. Although physicians require reliable and relevant documents to integrate into clinical practice, the internet presents a vast array of medical-related information, not all of which is verified. For instance, a study found that only 58% of eHealth websites in the United States met the criteria for content accuracy and credibility [17]. Consequently, physicians must develop information search skills to effectively find and assess web resources.

However, an institutionalized perception persists that the information found does not significantly enhance medical practice [6]—for a review, see [2]. This perception may be partly because not all physicians are aware of the availability of evidence-based medical resources on the internet (eg, PubMed) or they may lack the search skills to access them [18].

Main Web Resources Used

To effectively improve physicians' information search skills, it is essential to understand how medical students search for medical information and which web resources they use during their university program. A study on medical students found that they searched for online medical and clinical information daily (2-3 times per day) [19]. They accessed numerous websites (>50), including some recommended by the NHS (National Health Service), such as PubMed, which was used by only 30% of physicians and accounted for just 8% of all their searches.

XSL•FO
RenderX

However, they also relied on many general websites, particularly Wikipedia (Wikimedia Foundation; used by 71% of students and accounting for 26% of searches), patient forums, medical-specific wikis, and Google (used by 79% of students and comprising 32% of all searches). Indeed, Google is the most widely used general web search engine and has become an important tool in physicians' searches. Nevertheless, it is not specialized in medicine, and numerous websites present medical-related information without being reviewed by scientists or physicians [20,21], often leading to low-quality information [22]. Consequently, physicians must carefully evaluate the quality of information obtained from websites. Furthermore, physicians and medical students often prefer using Google over medicine-specific databases to answer clinical questions [23]. For instance, Krause et al [23] provided clinical questions to emergency department residents and asked them to search for answers either using a search engine (Google) or without it (by searching in a real environment). The results showed that while the number of incorrect answers increased with Google, the number of unsure responses decreased compared with searches conducted without Google. The residents developed a false sense of confidence in their answers and demonstrated poor efficiency in searching for and identifying correct clinical information using Google. While students and physicians may perceive Google as an evidence-based medical tool, it is not inherently an evidence-based medical database like PubMed.

Based on these prior studies, we can assert that medical-related information available on the internet can support clinical decision-making in patient health care, provided that physicians and medical students are trained to evaluate the quality of sources and content. The quality of medical-related information can be assessed according to 2 dimensions: (1) The credibility of sources, which refers to the degree of trustworthiness associated with the information and its origin [24-27]. This includes evaluations of accuracy, authority, objectivity, currency, and coverage [26,28]. Credibility is closely related to trustworthiness, reliability, accuracy, authority, and quality [24,25,29], although not always inseparable (eg, reliable information can sometimes be found on sites with low credibility). Consequently, credibility may influence the perceived reliability of information. (2) Reliability, in turn, refers to the extent to which information remains consistent and trustworthy under similar circumstances [30,31].

In the medical domain, Mikalef et al [6] classified credible documents as follows: (1) Authoritative web documents include scientific digital databases (eg, PubMed), websites of medical organizations (eg, the European Society of Cardiology), online peer-reviewed scientific journals (eg, The Lancet, JAMA), specialized medical journals (eg, *Circulation*, *Heart*), e-books, and government agency websites (eg, the Public Health Administration, the Food and Drug Administration). (2) Nonauthoritative web documents include medical portals, personal websites of physicians, pharmaceutical company websites, social media, and medical equipment websites. While authoritative sources can generally be considered credible and reliable, nonauthoritative sources require in-depth evaluation to determine their credibility and reliability. Therefore, when physicians and medical students use general websites to find medical-related information, they must:

- Translate a clinical or medical question into a search strategy to locate reliable information. This requires using specific medical terms or selecting credible web resources (eg, scientific databases).
- Evaluate credibility, reliability, relevance, and accessibility to determine which web resources to consult and use (see the review by [2]).

These 2 activities can be highly complex, especially when using general-purpose search engines (eg, Google), as individuals must select appropriate keywords and identify credible websites to obtain reliable information relevant to their professional objectives.

To this end, medical students and residents can rely on their prior domain knowledge to navigate the web more effectively. However, they often lack information search skills and familiarity with scientific tools and search systems, which are critical for efficient information retrieval. Users with strong information search skills are better guided by the components of the search space, regardless of the domain in which they seek information [32], a phenomenon known as the cross-domain transfer of information search skills [33]. Specifically, such users demonstrate more adaptive search behaviors tailored to their tasks and employ advanced search strategies [34]. For instance, they are more likely to incorporate Boolean operators into their queries to construct complex search equations rather than relying solely on simple keyword searches [35,36]. Additionally, they effectively utilize indexes and thesauri provided by search systems to identify relevant vocabulary, allowing them to refine their queries with greater specificity and precision as their search sessions progress [32].

However, not all academic programs place equal emphasis on developing information search skills. For example, computer science students acquire more advanced operational search skills compared with students in other fields [36,37]. Consequently, computer scientists have a deeper understanding of search systems and use them more effectively—leveraging features such as recommendations, thesauri, and domain-specific scientific databases—to conduct advanced searches, expand their vocabulary, and achieve their research objectives [38]. These search skills enable them to construct more structured and precise queries as their searches evolve [39].

In summary, prior domain knowledge facilitates attentional guidance as well as the reading and comprehension of web-based information [40]. By contrast, prior information search skills enhance the effective use and understanding of information systems.

Research Objectives and Questions

This study aimed to examine the effects of domain-specific prior knowledge versus information search skills on information search behaviors and performance. While medical students and residents benefit from their medical expertise when conducting medical searches, they lack the information search skills of their computer science counterparts, and vice versa. Identifying performance indicators and the challenges faced by these distinct

XSL•FO
RenderX

user profiles will help clarify the limitations of relying solely on domain knowledge in medical information searches.

More precisely, most prior studies examining how physicians search for information have relied on interviews or questionnaires to assess their search behaviors [2]. However, very few have analyzed the actual search activities of physicians or medical students and residents, despite the importance of understanding how they navigate the web when conducting necessary searches. To address this gap, we conducted an experimental study in which medical students and residents were compared with computer science students in their web-based information searches. More specifically, this study aimed to determine whether prior medical knowledge enabled medical students and residents to achieve higher search performance and make better choices regarding reliable and credible resources compared with computer science students, who had greater experience in information search and higher search skills.

More specifically, this study aimed to provide insights into the following research questions (RQs):

- RQ1: Do the most common searches conducted by physicians—treatment, diagnosis, and learning about a specific medical topic—result in different search performances between computer science students (who have higher information search skills but less medical knowledge) and medical students and residents (who have greater medical knowledge but lower information search skills)?
- RQ2: What types of resources do they consult during medical search tasks, and how do they assess the credibility of these resources based on their domain expertise (medicine vs information search)?
- RQ3: Do these 2 groups of students evaluate the reliability of the resources they use and the quality of the answers they provide differently when performing a medical search task?

Methods

Recruitment

The freely accessible data set used in this study was extracted from a user study conducted by Dosso et al [41] during the COVID-19 lockdown. The data set is available online [42].

A total of 20 students from the University of Toulouse participated in this study:

- 10 medical students and residents in medicine: 5 females and 5 males (mean age 25.4 years, SD 3.1 years); 2 were pursuing a master's degree and 8 were in an MD (Doctor of Medicine) program.
- 10 computer science students: 4 females and 6 males (mean age 23.8 years, SD 1.5 years); 5 were pursuing a master's degree and 5 were PhD (Doctor of Philosophy) students.

To evaluate their level of knowledge in medicine and their familiarity with the internet, all the participants had:

- To self-evaluate their knowledge in medicine, on a 5-point Likert scale ranging from 1 (very low) to 5 (very high). Medical students self-evaluated their level of medical knowledge higher than computer science students ($mean_{med}$ 4.2, SD 0.63; $mean_{cs}$ 1.3, SD 0.48; t_{18}=–11.5, P<.001, d=–5.15).
- To complete a multiple-choice questionnaire in the Medicine domain, elaborated by a physician, which consisted of 10 questions, each with 5 response options (1 correct, 3 incorrect, and 1 "I don't know"). Only correct answers earned 1 point, so the score per participant could range from 0 to 10. Results indicated that student-residents in medicine scored significantly higher (mean 5, SD 0.66) than students in computer science (mean 0.5, SD 0.97; t_{18}=–12.1, P<.001, d=–5.4).
- To complete the Information Search Self-Efficacy Scale developed by Rodon et al [43] and indicate the number of years of internet use and the contexts in which it is used. Results showed that students in computer science (mean 4.4, SD 0.51) had higher information search self-efficacy scores than those in medicine (mean 3.1, SD 0.73; t_{18}=4.56, P<.001, d=2.04). However, there was no significant difference in the number of years of internet use between the 2 groups ($mean_{CS}$ 12.3, SD_{CS} 1.83 and $mean_{med}$ 14, SD_{med} 2.79; t_{18}=–1.61, P=.12, d=–0.721). All the participants reported using the internet for both university and personal purposes.

Ethical Considerations

Before starting the search tasks, all participants received a detailed explanation of the study's objectives and procedures. Informed consent was obtained from each participant through a written agreement, which outlined the study's purpose, potential risks and benefits, data usage protocols, and participant rights, including the right to withdraw at any time without consequence. This study was reviewed and approved by the Ethics and Research Committee of the University of Toulouse (approval/project number 2021-354).

Procedure and Material

First, all participants completed an online prequestionnaire that included a free and informed consent form, questions on sociodemographic variables (eg, age, gender, education level, native language), internet information search scales, and multiple-choice questions assessing their level of medical knowledge. Additionally, participants described their general internet habits (ie, browsing device used, preferred browser, and search engine) and completed the Information Search Self-Efficacy Scale [43].

Then, the participants conducted the experiment at home during the COVID-19 lockdown in 2020-21. The experimental materials included the following:

- A printed booklet with general instructions for the experiment, a training exercise on using the logging browser, statements for the search tasks, and designated spaces for writing answers. As the written production task was conducted with pen and paper, the computer was used only for the search sessions.

XSL•FO
RenderX

- A memory stick containing the browser was provided for the search sessions, enabling participants to directly search for information while reviewing the experimental material.

Participants were instructed to complete each search task independently, without taking notes during the search session. This ensured that their final answers primarily reflected the information retained in memory after the search.

Both groups of participants (medical students/residents and computer science students) were presented with 3 distinct search tasks in general medicine (see Table 1 for detailed task descriptions). These tasks reflected real-world scenarios commonly encountered in medical practice, such as (1) enhancing knowledge in a specific disease area, (2) searching for information related to a particular patient problem, and (3) identifying and selecting the most appropriate treatment options for patients.

These scenarios are well-documented in medical research, as supported by previous studies [1,2].

The instructions specified that participants had to write their answers for each task only after completing the search session (ie, after closing the browser). There was no time limit. Note-taking was not permitted during the search session to ensure that responses closely reflected the knowledge retained in participants' memory.

Three different search tasks in general medicine were assigned to the participants (medical students/residents and computer science students; see Table 1 for the search tasks). These tasks closely resembled real-world scenarios encountered by physicians and medical students, such as improving their knowledge in a specific disease area, searching for information on a particular patient problem, and determining the most appropriate treatment for patients [1,2].

Table 1. The 3 different tasks provided to the participants and correct answers.

Three different tasks and answers	Tasks' instructions	Subtasks required
Learning task		
• Purpose: Improving knowledge of a specific area/disease • Correct answer: No correct answer was defined; instead, participants' responses were analyzed to compute the number of relevant elements they provided.	• You want to learn more about endometriosis.	• Understanding endometriosis • Understand the issues related to endometriosis • Discover/learn more about medical solutions to treat endometriosis
Treatment decision task		
• Purpose: Choosing the most appropriate treatment • Correct answer: Anticoagulant treatment (because the old woman did not fall too many times for her age)	• An 83-year-old woman had a nonsequelae stroke 5 months ago. • At the stroke assessment, atrial fibrillation was discovered. • She had fallen 3 times in the last 2 months. • Should anticoagulant therapy be initiated? • After evaluating the benefit-risk ratio of initiating or not initiating anticoagulant treatment, select the management option that you find most appropriate and justify your choice.	• Understanding nonsequential stroke • Understanding atrial fibrillation • Understanding the risk of falls • Identify the advantages and disadvantages of anticoagulant therapy • Judge this information according to the criteria to be established (benefit/risk assessment)
Diagnosis task		
• Purpose: Determining the pathology • Correct answer: Splenic infarction or spleen infarction	• A 47-year-old man presented to the emergency department with left hypochondrium pain that had been evolving for 24 hours and was not relieved by level 1 analgesics. • His history included cutaneous lupus and polycythemia. • The biological workup was normal. • The abdominopelvic scanner found 2 splenic hypodensities. With the information collected on the internet, propose your diagnosis and etiological hypotheses	• Understanding left hypochondrium • Understanding of level 1 analgesics • Understanding cutaneous lupus and polycythemia • Understanding splenic hypodensity • Analyze the information previously collected and propose a diagnosis with etiological hypotheses

Variables

Two independent variables (IVs) were manipulated:

- IV1: Domain expertise of participants (medical students/residents vs computer science students) as a between-participants variable.
- IV2: Search tasks (learning, choosing treatment, diagnosis) as a within-participants variable.

A total of 6 dependent variables (DVs) were measured to bring some answers to the 3 RQs. For RQ1, we measured 3 DVs:

- DV1: Total time (seconds) spent on each search task, measured from the first query entered to the completion of the multiquery process.

XSL•FO
RenderX

- DV2: Number of medical-related keywords generated by participants, excluding those explicitly provided in the search task instructions.
- DV3: Accuracy and relevance of retrieved information:
  - For diagnostic and treatment tasks, responses were scored as 0 (incorrect or abandoned) or 1 (correct).
  - For the learning task, responses were scored from 0 to 4 based on the amount of relevant information provided and interconnected. Relevant information was defined as content directly related to the assigned topic (endometriosis) in accordance with the information context [44]. This is rated as follows: 0 (incorrect information), 0.5 (correct information), 1 (correct and relevant information), 1.5 (correct and relevant information linked to general vocabulary), 2 (correct and relevant information linked to medical vocabulary), 3 (correct and relevant information linked to general vocabulary and interconnected), and 4 (correct and relevant information linked to medical vocabulary and interconnected).

For RQ2, we measured:

- DV4: Evaluation of the credibility and type of web resources consulted:
  - DV4a: Credibility score of the websites visited for each task.
  - DV4b: Distribution of the types of web resources consulted, categorized as websites or as documents.

For DV4a, the credibility of web resources consulted per search task was scored as follows, based on Mikalef et al [6]:

- Peer-reviewed scientific journals or scientific databases (eg, PubMed, ScienceDirect, The Lancet)=2 points per URL
- Documents from hospitals, health authorities (eg, World Health Organization), and government agencies=1.5 points per URL
- Health-related websites verified by HONcode (Health On the Net code; certification by the nonprofit Health On the Net Foundation [45])=1 point per URL
- Other websites (not scientifically verified)=0 points per URL

For DV4b, we analyzed the distribution of resources consulted across different categories (peer-reviewed scientific journals, hospital/government documents, HONcode-certified websites, and other websites) during each search task. This distribution was based on the number of resources consulted per category (as defined in DV4a) and per search task.

For RQ3, 2 DVs were measured:

- DV5: Self-assessment of reliability of web documents visited while achieving the search task [46]: “Overall, the search engine results, websites, or documents (results from Google, websites, pdf, etc) consulted were: (1) absolutely not trustworthy, (2) very untrustworthy, (3) not very trustworthy, (4) moderately trustworthy, (5) somewhat trustworthy, (6) trustworthy, or (7) extremely trustworthy. The higher the score was, the more participants perceived web documents as reliable.”
- DV6: Self-assessment of the quality of their own answers on a 4-point scale for the 3 tasks. Participants rated their perceived answer quality as follows: (1) very bad, (2) bad, (3) good, and (4) very good. A higher score indicated a more favorable self-evaluation of their answers.

Statistical Analysis

For all DVs, we compared student-residents in medicine with students in computer science for each task using either parametric or nonparametric tests. Specifically, for all DVs except DV4b, we performed Student *t* tests. For DV4b, we conducted a repeated-measures ANOVA. Tests for homogeneity of variance (Levene test) and normality (Shapiro-Wilk test) were performed before conducting ANOVA and unpaired (1-tailed) *t* tests. Partial η^2 was used as an index of effect size. When homogeneity of variance was violated, we applied the Mann-Whitney test. Additionally, some quantitative results are illustrated with qualitative examples.

Results

Overview

The results are presented according to the 3 RQs.

- *RQ1: Do the most popular searches developed by physicians (ie, treatment, diagnosis, and learning on a specific medical topic) lead to different search performances between computer science students (with higher information search skills but lower medical knowledge) and medical student-residents (with higher medical knowledge but lower information search skills)?*

Table 2 presents the dependent variables for RQ1.

XSL•FO
RenderX

Table 2. Dependent variables for RQ1 with regard to the search task and expertise domain.

Group of students	Learning task, mean (SD)	Diagnostic task, mean (SD)	Treatment task, mean (SD)
DV1: Search time (seconds)			
Medicine	548 (474)	705 (444)	634 (407)
Computer science	520 (373)	1099 (1085)	936 (699)
DV2: New keywords related to medicine (mean number)			
Medicine	1.80 (0.91)[a]	4 (4.19)	4.8 (4.13)
Computer science	0.3 (0.48)[a]	2 (2.35)	2.7 (3.97)
DV3: Correct answers or relevant information provided (mean number)			
Medicine	3.9 (0.31)[a]	0.9 (0.31)[a]	0.85 (0.15)[a]
Computer science	3.5 (0.52)[a]	0 (0)[a]	0.1 (0.31)[a]

[a]Significant differences between medical students/residents and computer science students.

Time Spent to Find Information

Regarding the time required to find information, no significant differences were observed between student-residents in medicine and those in computer science for any task (all *t* tests were nonsignificant: t_{18}=-.146, *P*=.88 for the learning task; t_{18}=1.06, *P*=.30 for the diagnostic task; and t_{18}=1.18, *P*=.25 for the treatment task).

Keywords Related to Medicine Added by the Participants

Student-residents in medicine added more medicine-related keywords than students in computer science, but only for the learning task (t_{18}=–4.57, *P*<.001, *d*=2.04). On average, student-residents in medicine added 1.6 keywords (SD 0.91), compared with 0.3 (SD 0.48) for students in computer science. The keywords added by student-residents in medicine were primarily related to specific websites they were familiar with (eg, associations of doctors in gynecology or government health agency websites, where they could find medical information).

For the treatment and diagnostic tasks, statistical analyses revealed no significant differences between student-residents in medicine and students in computer science (t_{18}=–1.32, *P*=.20, and t_{18}=–1.16, *P*=.26, respectively).

Correct information retrieval for the diagnostic and treatment tasks, as well as the number of relevant information elements provided by participants for the learning task, was analyzed. The results are as follows:

For the treatment and diagnostic tasks, medical student-residents provided more correct answers than those in computer science.

- For the treatment task (t_{18}=–6.77, *P*<.001, *d*=–3.03), the mean score was 0.85 (SD 0.15) for medical student-residents, compared with 0.1 (SD 0.31) for computer science students.
- For the diagnostic task (t_{18}=–9, *P*<.001, *d*=–4.02), the mean score was 0.9 (SD 0.31) for medical student-residents, compared with (SD 0) for computer science students.

For the learning task, even after applying the square root correction ($\sqrt{2}$), the homogeneity of variance was violated. Therefore, we applied the nonparametric Mann-Whitney test. Medical student-residents provided more relevant information elements than computer science students (U_{18}=30, *P*=.03, rank biserial correlation=0.4), with a mean of 3.9 (SD 0.31) for medical students and 3.5 (SD 0.52) for computer science students.

As illustrated in Multimedia Appendix 1 with 4 examples, the nature of the information elements provided for the learning task varied depending on the participants' domain of expertise. This task was exploratory and was approached differently based on the participants' expertise. Medical student-residents generally provided comprehensive responses specific to medicine, including definition and prevalence, symptoms, causes and consequences, diagnosis, and possible treatments. By contrast, computer science students primarily focused on definition, prevalence, and symptoms, with very few mentioning possible treatments. Their answers were more superficial compared with those of medical student-residents, who applied a medical framework to obtain information about the pathology.

- *RQ2: What types of resources do participants consult, and what is their credibility level when performing medical search tasks, considering their domain of expertise (medicine vs information search)?*

Credibility Resources Consulted, Scores, and Websites Used

First, regarding the credibility scores obtained (DV4a), medical student-residents scored higher than computer science students for the learning task, with a mean of 5.85 (SD 4.37) compared with 2.2 (SD 2; t_{18}=–2.40, *P*=.046, Cohen *d*=1.08). For the diagnostic task, only a marginal effect was observed (t_{18}=–1.48, *P*=.07), with a mean of 13.45 (SD 9.63) for medical student-residents versus 8 (SD 6.53) for computer science students.

For the treatment task, no significant difference was found between the 2 groups, with a mean of 6.9 (SD 4.67) for medical

XSL•FO
RenderX

student-residents and 5.4 (SD 4.09) for computer science students (t_{18}=–0.76, P=.45).

Next, we analyzed the distribution of internet resources used (DV4b; Figure 1). The resources did not hold the same prominence during the information search ($F_{3,54}$=28.12, P<.001, η^2p=0.61); hospital websites (P<.001) and documents (P<.001) were used more frequently than other sources.

We conducted a more detailed analysis of resource usage based on expertise and tasks. We observed:

- For evidence-based resources, surprisingly, no significant difference was found between medical student-residents and computer science students ($F_{1,18}$=0.31, P=.58). These resources were rarely visited by any participants.
- For hospital websites, medical student-residents used them more frequently than computer science students ($F_{1,18}$=13.96, P<.001, η^2p=0.44).
- For HONcode health websites, computer science students consulted these resources more frequently than medical student-residents ($F_{1,18}$=6, P=.02, η^2p=0.25).
- For other websites, computer science students used them more frequently than medical student-residents ($F_{1,18}$=5.21, P=.001, η^2p=0.34).

Next, qualitative analyses examined the main websites accessed through opened URLs (see Table 3).

Figure 1. Means (and SDs) of web resources consulted with regard to the expertise domain of students.

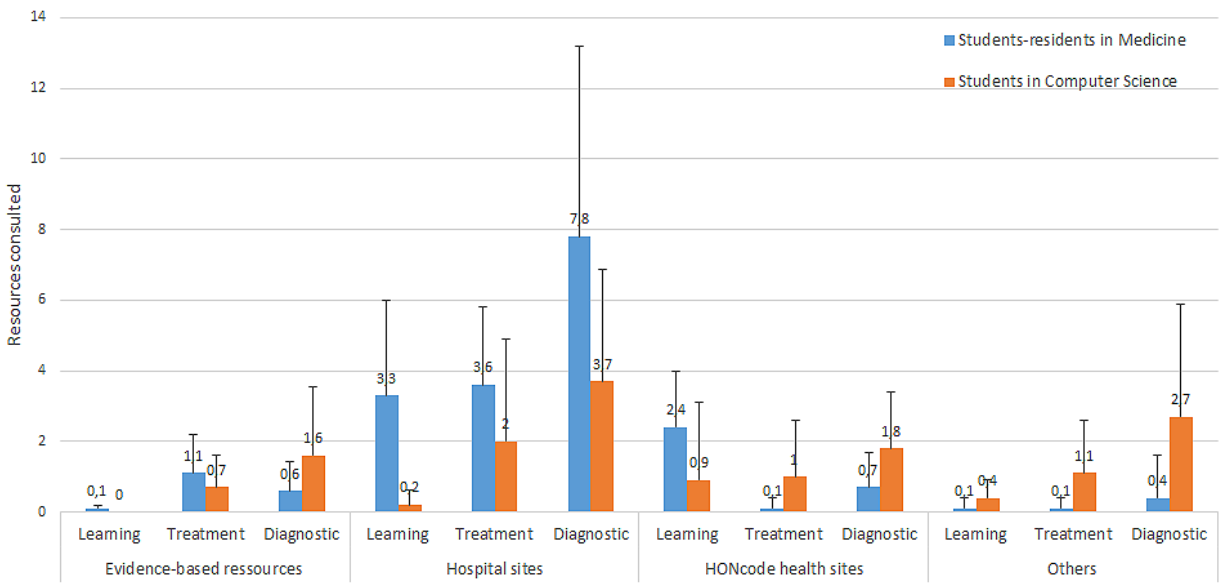

XSL•FO
RenderX

Table 3. The main websites visited by participants and their score with regard to the search task to be performed and expertise.

Tasks and websites (resources)		Authority versus nonauthority websites and scores obtained
Learning task		
Student-residents in medicine	• A government health agency [47] • Association of Physicians in Gynecology [48]	• Authority; score=1.5 points • Authority; score=1.5 points
Students in computer science	• Association of Patients With Endometriosis [49] • Nonofficial health websites (eg, [50,51])	• Nonauthority; score=1 points (verified by HONcode) • Nonauthority; score=0 points
Diagnostic task		
Student-residents in medicine	• Digital medical campus (not updated since 2016) [52] • A university hospital [53]	• Authority; score=1.5 points • Authority; score=1.5 points
Students in computer science	[54-56]	• Nonauthority; score=1 point (verified by HONcode) • For [55,56]: nonauthority; score=0 points
Treatment task		
Student-residents in medicine	• A Swiss medical journal [57] • A government health agency [47]	• Authority; score=2 points • Authority; score=1.5 points
Students in computer science	• A Swiss medical journal [57] • The other websites consulted are very different from one participant to another	• Authority; score=2 points

- *RQ3: Do they assess differently the reliability of resources they choose to use and the quality of answers they provided for the search task to be performed?*

Results of the self-assessment of the reliability of web documents consulted when answering this question at the end of each session (maximum of 7 points): a significant difference was found between medical student-residents and computer science students for the learning task (t_{18}=–3.13, P=.003, d=–1.39), with a mean of 6.8 (SD 0.42) for medical student-residents and 5.8 (SD 0.91) for computer science students. For the treatment and diagnostic tasks, statistical analyses did not reveal any significant differences between the 2 groups (t_{18}=–1.24, P=.12, and t_{18}=–1, P=.16; Table 4).

Table 4. Dependent variables for research question 3 with regard to the search task and expertise domain.

Research question 3 dependent variables and group of students	Learning task, mean (SD)	Diagnostic task, mean (SD)	Treatment task, mean (SD)
DV5: Self-assessment of resource reliability consulted (maximum score=7 points)			
Medicine	6.8 (0.42)[a]	6 (0.66)	5.4 (0.51)
Computer Science	5.8 (0.91)[a]	5.6 (1.07)	5.1 (0.56)
DV6: Self-assessment of quality of answer (maximum score=4 points)			
Medicine	1.84 (0.14)[a]	1.73 (0.13)[a]	1.79 (0.11)[a]
Computer Science	1.76 (0.08)[a]	1.31 (0.29)[a]	1.70 (0.10)[a]

[a]Significant differences between medical students/residents and computer science students.

Results of the self-assessment of the quality of answers provided on a 4-point scale for the 3 tasks; participants rated their perceived answer quality on a scale from 1 (very bad) to 4 (very good).

For all search tasks, medical student-residents rated the quality of their answers higher than computer science students:

- For the learning task, medical student-residents had a mean self-assessed answer quality of 3.5 (SD 0.52), compared with 3.1 (SD 0.31) for computer science students (t_{18}=–2.05, P=.03, d=–0.92).
- For the diagnostic task, medical student-residents had a mean self-assessed answer quality of 3.2 (SD 0.42), compared with 2.9 (SD 0.31) for computer science students (t_{18}=–1.8, P=.04, d=–0.80).
- For the treatment task, medical student-residents had a mean self-assessed answer quality of 2.9 (SD 0.31), compared with 1.8 (SD 0.78) for computer science students (t_{18}=–4.09, P<.001, d=–1.83).

Discussion

Principal Findings

Prior studies have suggested that physicians' primary challenge when searching for information online is their lack of information search skills [2,23], which may negatively impact

XSL•FO
RenderX

the quality of information found and, consequently, patient care. This study aimed to examine the strategies used by medical students and residents when searching for medical information online, the web resources they consulted (including the nature and credibility of these resources), how they assessed the reliability of these resources, and the quality of the answers they provided. The search behaviors of medical students and residents were compared with those of computer science students. While medical students and residents had greater medical knowledge, computer science students exhibited stronger information search skills but lower medical-related knowledge. The main findings indicated that medical students and residents outperformed computer science students. Overall, they found more relevant information, used more credible resources, rated these resources as more reliable, and had greater confidence in their answers than computer science students. Two particularly surprising and interesting results emerged: (1) medical students and residents generated more medical keywords than computer science students. These keywords, at least in part, helped them locate familiar websites, making them indirectly relevant to information searching; and (2) medical students and residents rarely used evidence-based resources, instead relying on familiar medical websites, such as hospital websites.

Interpretation of the Results Obtained

First, regarding information search performance, the results confirmed expectations: medical students and residents outperformed computer science students. They provided more correct answers for treatment and diagnostic tasks and included more relevant and structured elements in learning tasks—all without requiring additional time. Their deeper understanding of the search topics enabled them to deliver more structured and comprehensive responses. For example, medical students and residents frequently included details on symptoms, causes, diagnosis, and possible treatments. By contrast, computer science students' responses tended to be more superficial, focusing primarily on general statistics and potential consequences. These findings align with prior research on the positive effects of domain knowledge in generating more accurate and higher-quality search outcomes, regardless of the field of expertise [58-60]. Indeed, when processing information, individuals with domain expertise engage in top-down processing, which enhances reading comprehension and supports a deeper understanding of the content [61-63]. As a result, at the search outcome stage, domain experts can more effectively reuse extracted information to achieve their goals and produce higher-quality responses [60]. The advantages of domain expertise in medical information searches were also evident in keyword selection. Notably, medical students and residents used more medical-related keywords, particularly when researching a specific disease. These findings align with previous studies showing that medical experts employ a more specialized medical vocabulary than nonexperts [63-65]. The medical knowledge applied during query formulation allowed physicians to locate more precise information at a search speed comparable to that of computer science students. However, despite their speed, computer science students' searches tended to be far less precise.

Regarding these initial findings—the quality of final responses, keyword formulation, and search precision—it is worth noting that medical students accurately self-assessed their responses. Their higher confidence in the reliability of their answers, compared with computer science students, likely stemmed from their reliance on familiar medical websites. This underscores their ability to effectively locate and utilize credible information from online and hospital-based resources, as previously demonstrated [2].

When comparing the objective credibility scores of consulted web resources with participants' self-assessed reliability of those resources, the results showed that medical students and residents accessed more credible websites than computer science students for learning and diagnostic tasks. A significant difference was observed for learning tasks, while scores were comparable for diagnostic tasks. Learning and diagnostic tasks may require greater medical expertise to identify credible and reliable sources compared with treatment-related tasks. Accurate diagnosis often involves a careful differential diagnosis, requiring the exclusion of other conditions with similar symptoms. This process requires access to credible medical resources, as general websites are often insufficient. Medical students and residents effectively relied on reputable medical websites to access relevant case studies and other pertinent information. Likewise, learning about specific topics required current and reliable information, underscoring the need for high-quality medical sources.

A detailed analysis of the consulted websites revealed task-dependent differences between medical students/residents and computer science students. Notably, evidence-based resources were rarely used by any participants and were primarily consulted for diagnostic and treatment tasks. This finding was unexpected, as medical students and residents, given their rigorous training, were expected to prioritize evidence-based resources. While this result aligns with previous research highlighting the limited use of evidence-based resources among medical practitioners [66,67], it raises concerns about the effectiveness of current training programs in fostering evidence-based practice skills. As previously noted, physicians often struggle to formulate effective search queries in databases such as PubMed, leading to the retrieval of numerous irrelevant articles [68]. When using general search engines, they face 2 main challenges [11]: (1) managing and processing a vast amount of information and (2) assessing the reliability of content based on scientific evidence, all within a limited time frame. On average, physicians spend between 9 and 11 minutes searching the internet for medical information [2], a trend also observed among the medical students and residents in this study. To navigate these challenges, physicians tend to rely on familiar websites rather than exploring new content or consulting databases like PubMed [1,6,19].

The preference for familiar hospital websites observed in our study may offer a more streamlined and efficient—though potentially less comprehensive—pathway to relevant information. Our findings further support this trend, as medical students and residents consistently relied on familiar hospital resources across all search tasks, favoring established sources over broader exploration. In our experimental study, medical students and residents primarily consulted hospital-based resources for all tasks and did so more frequently than computer science students. This observation aligns with previous research

XSL•FO
RenderX

showing that health care professionals frequently rely on a limited set of preferred online resources [1,6,19]. One study found that physicians visit familiar websites to reduce search time [18], often going directly to trusted medical sources [69-71]. They apply various criteria to assess information quality (usefulness and accuracy) and authority (trustworthiness and credibility) [19], which may be particularly relevant when evaluating hospital websites and documents. Additionally, physicians tend to search for methodical and empirically grounded medical solutions to minimize errors in decision-making while maximizing diagnostic and therapeutic effectiveness, as previously demonstrated [6]. By contrast, our study found that computer science students relied more on HONcode-certified health websites and other unverified sources. Despite their advanced search skills, they may struggle to assess the credibility of medical websites or fully comprehend the information presented on platforms such as PubMed and hospital websites. Their limited medical knowledge likely hindered their ability to identify relevant and accurate medical information.

Consequently, the medical expertise of students and residents allowed them to achieve higher scores, evaluate resources more effectively, and select more relevant web sources compared with computer science students.

Implications

The results we obtained may have several implications.

First, a surprising finding was the limited use of evidence-based resources such as PubMed by medical students and residents. This may be due to the challenge of filtering relevant information from a vast number of irrelevant results, as previously documented [68]. Additionally, difficulties in effectively using search filters within these databases may have further hindered their ability to retrieve pertinent information. This finding underscores the need for enhanced training programs to equip health care professionals with the skills necessary to effectively utilize evidence-based resources. Integrating clinical and practical evidence-based medicine principles is crucial for updating medical knowledge, improving patient care, and fostering strong physician-patient relationships [72]. A study by the Institute for Healthcare Informatics [73] highlighted that physicians frequently rely on Wikipedia as a primary source of health care information, further emphasizing the need for targeted training initiatives. These programs should focus on strengthening information literacy skills, including the effective use of evidence-based databases such as PubMed and the critical evaluation of online health information. This could include instruction on HONcode principles, which help distinguish reliable from unreliable sources based on factors such as authorship, funding sources, and URL domains (.org, .gov, .net). Beyond their clinical practice, physicians who are well-versed in evaluating online information can also better support their patients. Many patients seek medical information on the internet, often without the ability to assess its accuracy [74], which can have negative consequences for their health care. Equipping physicians with the skills to identify credible and effective online resources would enable them to guide patients in interpreting the information they find online, ultimately improving patient education and decision-making. Physicians at all stages of their careers must stay updated on advancements in research, treatments, and best practices to ensure safe and effective patient care. Online scientific databases are essential tools in this process, particularly as the use of generative artificial intelligence (AI), such as ChatGPT (OpenAI, Inc.), continues to grow. Unlike Google, AI-generated responses often lack proper sourcing and reliability [75]. Given the evolving role of AI in knowledge acquisition, structured training programs are crucial to helping medical students and physicians navigate these resources effectively. Additionally, integrating verified sources—such as Google Scholar or PubMed—directly into general search engine results for medical queries could further support evidence-based learning and decision-making. General search engines could further support specialized medical queries by suggesting relevant medical keywords based on the task at hand. This approach would help medical students and physicians—who are already familiar with general search engine interfaces—find the required information more quickly than on PubMed, which they often find difficult to navigate efficiently. Physicians' preference for traditional search engines over PubMed may stem from their perception that PubMed is not well-suited to the practical demands of clinical medicine. General practitioners, in particular, require rapid access to high-quality content and do not have the time to sift through an overwhelming volume of information or conduct an extensive evaluation process [11]. Ideally, search engines should enhance usability by providing clear and precise indicators about the credibility and nature of retrieved results directly on the search results page. This includes specifying whether the content has undergone peer review and clarifying the level of empirical evidence presented in the source. Future research in design ergonomics and medical informatics should focus on developing and testing the impact of such enhanced search interfaces on physicians' search performance.

Limitations

This study had some limitations. The small sample size suggests that replicating the study with a larger and more diverse group of participants—including medical students, residents, and both junior and senior physicians—would be valuable. Given that physicians must continuously update their knowledge throughout their careers, future research should explore how search behaviors (eg, keyword formulation, websites visited) evolve at different professional stages, particularly in the context of emerging AI tools.

In addition to better distinguishing the impact of domain knowledge versus information search skills, future research could compare the performance of students in library and information science with that of medical students and residents. As library and information science students have specialized expertise in information retrieval and have outperformed computer science students in prior studies [33], such a comparison could provide further insights into the relationship between search expertise and domain-specific knowledge.

Although this study was conducted remotely—potentially raising concerns about data quality due to limited experimental control—the findings align with previous research. Notably, the search times recorded for medical students and residents in

XSL•FO
RenderX

this study were consistent with those reported in prior studies [2]. Moreover, the analysis of search times did not reveal any significant outliers, further supporting the reliability of the collected data.

Conclusions

The findings of this experimental study have significant implications for both current and future physicians.

A surprising finding was the limited use of evidence-based resources by medical students and residents. This may be due to the challenge of extracting relevant information from a high volume of irrelevant results, as previously documented [64]. This finding highlights the need for enhanced training programs, both at the university and at the postgraduate levels, to equip health care professionals with the skills necessary to effectively utilize evidence-based resources. These programs should focus on improving information search skills, including the efficient use of evidence-based databases and the critical evaluation of online health information. This could include instruction on HONcode principles, which help distinguish between reliable and unreliable sources based on factors such as authorship, funding sources, and URL domains. Beyond their clinical practice, knowledge of source evaluation criteria could also assist physicians in guiding patients who seek medical information online. This is especially important with the rise of AI tools, which are fundamentally reshaping how information is accessed and interpreted.

Training physicians and future physicians to effectively use these tools will also encourage them to expand the range of websites they consult. Rather than limiting themselves to familiar sources, such as hospital websites, they will be better equipped to explore a broader array of credible medical resources.

Acknowledgments

This work was supported by the French National Research Agency (ANR) CoST project (ANR-18-CE23-0016).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Examples of answers provided by students-residents in medicine and students in computer science with their description.
[DOCX File , 24 KB - formative_v9i1e62754_app1.docx]

References

1. Bennett NL, Casebeer LL, Zheng S, Kristofco R. Information-seeking behaviors and reflective practice. J Contin Educ Health Prof 2006;26(2):120-127. [doi: 10.1002/chp.60] [Medline: 16802313]
2. Daei A, Soleymani MR, Ashrafi-Rizi H, Zargham-Boroujeni A, Kelishadi R. Clinical information seeking behavior of physicians: a systematic review. Int J Med Inform 2020 Jul;139:104144. [doi: 10.1016/j.ijmedinf.2020.104144] [Medline: 32334400]
3. Sarbaz M, Kimiafar K, Sheikhtaheri A, Taherzadeh Z, Eslami S. Nurses' information seeking behavior for clinical practice: a case study in a developing country. Stud Health Technol Inform 2016;225:23-27. [Medline: 27332155]
4. Magrabi F, Coiera EW, Westbrook JI, Gosling AS, Vickland V. General practitioners' use of online evidence during consultations. Int J Med Inform 2005 Jan;74(1):1-12. [doi: 10.1016/j.ijmedinf.2004.10.003] [Medline: 15626631]
5. Schilling LM, Steiner JF, Lundahl K, Anderson RJ. Residents' patient-specific clinical questions: opportunities for evidence-based learning. Acad Med 2005 Jan;80(1):51-56. [doi: 10.1097/00001888-200501000-00013] [Medline: 15618093]
6. Mikalef P, Kourouthanassis PE, Pateli AG. Online information search behaviour of physicians. Health Info Libr J 2017 Mar 07;34(1):58-73 [FREE Full text] [doi: 10.1111/hir.12170] [Medline: 28168831]
7. Heale BSE, Khalifa A, Stone BL, Nelson S, Del Fiol G. Physicians' pharmacogenomics information needs and seeking behavior: a study with case vignettes. BMC Med Inform Decis Mak 2017 Aug 01;17(1):113 [FREE Full text] [doi: 10.1186/s12911-017-0510-9] [Medline: 28764766]
8. Devi P, George J. Drug information needs of physicians treating diabetic nephropathy in a tertiary care hospital. Kathmandu Univ Med J (KUMJ) 2008;6(1):23-27. [Medline: 18604110]
9. Casebeer L, Bennett N, Kristofco R, Carillo A, Centor R. Physician internet medical information seeking and on-line continuing education use patterns. J Contin Educ Health Prof 2002;22(1):33-42. [doi: 10.1002/chp.1340220105] [Medline: 12004639]
10. Vallo Hult H, Master Östlund C, Pålsson P. Why and when physicians Google: resident physicians' information-seeking strategies during patient consultations. Stud Health Technol Inform 2023 Jun 29;305:580-583. [doi: 10.3233/SHTI230563] [Medline: 37387097]
11. Bernard E, Arnould M, Saint-Lary O, Duhot D, Hebbrecht G. Internet use for information seeking in clinical practice: a cross-sectional survey among French general practitioners. Int J Med Inform 2012 Jul;81(7):493-499. [doi: 10.1016/j.ijmedinf.2012.02.001] [Medline: 22425281]

XSL•FO
RenderX

12. Westbrook JI. Do online information retrieval systems help experienced clinicians answer clinical questions? Journal of the American Medical Informatics Association 2005 Jan 31;12(3):315-321. [doi: 10.1197/jamia.m1717]
13. Muse E, Godino J, Netting JAJ, Alexander JF, Moran HJ, Topol EJ. From second to hundredth opinion in medicine: a global consultation platform for physicians. npj Digital Med 2018 Oct 09;1(1):55 [FREE Full text] [doi: 10.1038/s41746-018-0064-y] [Medline: 31304334]
14. Del Fiol G, Mostafa J, Pu D, Medlin R, Slager S, Jonnalagadda SR, et al. Formative evaluation of a patient-specific clinical knowledge summarization tool. Int J Med Inform 2016 Feb;86:126-134 [FREE Full text] [doi: 10.1016/j.ijmedinf.2015.11.006] [Medline: 26612774]
15. Sun W, Ma R, Zhao X, Li Z, Lindqvist J, El AA, et al. Trusting the search: unraveling human trust in health information from Google and ChatGPT. arXiv :e1 Preprint posted online on March 15, 2024 [FREE Full text]
16. Häußler H. Do users trust search engines? And if so, why?: Developing a trust measure and applying it in an experiment. In: CHIIR '23: Proceedings of the 2023 Conference on Human Information Interaction and Retrieval. New York, NY: Association for Computing Machinery; 2023 Presented at: CHIIR '23: ACM SIGIR Conference on Human Information Interaction and Retrieval; March 19-23, 2023; Austin, TX p. 471-474. [doi: 10.1145/3576840.3578280]
17. Devine T, Broderick J, Harris LM, Wu H, Hilfiker SW. Making quality health websites a national public health priority: toward quality standards. J Med Internet Res 2016 Aug 02;18(8):e211 [FREE Full text] [doi: 10.2196/jmir.5999] [Medline: 27485512]
18. Schuers M, Griffon N, Kerdelhue G, Foubert Q, Mercier A, Darmoni SJ. Behavior and attitudes of residents and general practitioners in searching for health information: from intention to practice. Int J Med Inform 2016 May;89:9-14. [doi: 10.1016/j.ijmedinf.2016.02.003] [Medline: 26980354]
19. Hughes B, Joshi I, Lemonde H, Wareham J. Junior physician's use of Web 2.0 for information seeking and medical education: a qualitative study. Int J Med Inform 2009 Oct;78(10):645-655. [doi: 10.1016/j.ijmedinf.2009.04.008] [Medline: 19501017]
20. Sirohi D, Ng CHM, Bidargaddi N, Slater H, Parker MA, Hull ML, et al. High-quality eHealth websites for information on endometriosis: systematic search. J Med Internet Res 2024 Jan 10;26:e48243. [doi: 10.2196/48243] [Medline: 38198205]
21. Yuan X, White R. Building the trail best traveled: effects of domain knowledge on web search trailblazing. In: CHI '12: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. New York, NY: Association for Computing Machinery; 2012 Presented at: CHI '12: CHI Conference on Human Factors in Computing Systems; May 5-10, 2012; Austin, TX p. 1795-1804. [doi: 10.1145/2207676.2208312]
22. Bergmo TS, Sandsdalen V, Manskow US, Småbrekke L, Waaseth M. Internet use for obtaining medicine information: cross-sectional survey. JMIR Form Res 2023 Feb 02;7:e40466. [doi: 10.2196/40466] [Medline: 36729577]
23. Krause R, Moscati R, Halpern S, Schwartz D, Abbas J. Can emergency medicine residents reliably use the internet to answer clinical questions? West J Emerg Med 2011 Nov;12(4):442-447 [FREE Full text] [doi: 10.5811/westjem.2010.9.1895] [Medline: 22224135]
24. Szegőfi A, Heintz C. Institutions of epistemic vigilance: the case of the newspaper press. Social Epistemology 2022 Aug 23;36(5):613-628 [FREE Full text] [doi: 10.1080/02691728.2022.2109532]
25. Sperber D, Clément F, Heintz C, Mascaro O, Mercier H, Origgi G, et al. Epistemic vigilance. Mind and Language 2010 Sep;25(4):359-393. [doi: 10.1111/j.1468-0017.2010.01394.x]
26. Metzger MJ. Making sense of credibility on the web: models for evaluating online information and recommendations for future research. J Am Soc Inf Sci 2007 Sep 21;58(13):2078-2091. [doi: 10.1002/asi.20672]
27. Fogg BJ, Tseng H. The elements of computer credibility. In: CHI '99: Proceedings of the SIGCHI conference on Human Factors in Computing Systems. New York, NY: Association for Computing Machinery; 1999 Presented at: CHI99: Conference on Human Factors in Computing Systems; May 15-20, 1999; Pittsburgh, PA p. 80-87. [doi: 10.1145/302979.303001]
28. Metzger MJ, Flanagin AJ. Credibility and trust of information in online environments: the use of cognitive heuristics. Journal of Pragmatics 2013 Dec;59:210-220. [doi: 10.1016/j.pragma.2013.07.012]
29. Rieh SY. Judgment of information quality and cognitive authority in the Web. J Am Soc Inf Sci 2001 Dec 14;53(2):145-161 [FREE Full text] [doi: 10.1002/asi.10017]
30. Babbie ER. The Basics of Social Research (7 edition). Belmont, CA: Wadsworth Publishing; 2016.
31. Trochim W, Donnelly J. The Research Methods Knowledge Base (3rd Edition). Mason, OH: Atomic Dog Publishing Inc; Dec 1, 2006:56-65.
32. Smith C. Domain‐independent search expertise: a description of procedural knowledge gained during guided instruction. Asso for Info Science & Tech 2014 May 19;66(7):1388-1405 [FREE Full text] [doi: 10.1002/asi.23272]
33. Aduba DE, Obot IN, Baro EE. A comparative study of information literacy skills among students of library and information science and computer science in Nigerian universities. Niger Delta J Libr Inf Sci 2022 Jun 30;3(1):1-15 [FREE Full text]
34. Aula A, Nordhausen K. Modeling successful performance in Web searching. Journal of the American Society for Information Science Technology 2006;57(12):1678-1693.
35. Hölscher C, Strube G. Web search behavior of internet experts and newbies. Computer Networks 2000 Jun;33(1-6):337-346. [doi: 10.1016/s1389-1286(00)00031-1]

XSL•FO
RenderX

36. White RW, Morris D. Investigating the querying and browsing behavior of advanced search engine users. In: SIGIR '07: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, NY: Association for Computing Machinery; 2007 Presented at: SIGIR07: The 30th Annual International SIGIR Conference; July 23-27, 2007; Amsterdam The Netherlands p. 255-262. [doi: 10.1145/1277741.1277787]
37. Ebiefung R, Adetimirin A, Oyewole O. Discipline variation in information retrieval skills and use of electronic information resources by undergraduates in University Of Uyo, Akwa Ibom State, Nigeria. Library Philosophy and Practice 2020:1-20 [FREE Full text] [doi: 10.1108/dlp-08-2020-0080]
38. Marchionini G. Information Seeking in Electronic Environments. Cambridge, UK: Cambridge University Press; 1995.
39. Fidel R, Pejtersen AM. From information behaviour research to the design of information systems: the Cognitive Work Analysis framework. Information Research. 2004. URL: http://InformationR.net/ir/10-1/paper210.html [accessed 2025-03-25]
40. Dosso C, Tamine L, Paubel P, Chevalier A. Navigational and thematic exploration–exploitation trade-offs during web search: effects of prior domain knowledge, search contexts and strategies on search outcome. Behaviour & Information Technology 2023 Aug 10;43(10):2232-2258. [doi: 10.1080/0144929x.2023.2242514]
41. Dosso C, Moreno JG, Chevalier A, Tamine L. CoST: an annotated data collection for complex search. In: CIKM '21: Proceedings of the 30th ACM International Conference on Information & Knowledge Management.: The 30th ACM International Conference on Information and Knowledge Management. ACM; 2021 Presented at: CIKM '21: The 30th ACM International Conference on Information and Knowledge Management; November 1-5, 2021; Virtual p. 4455-4464. [doi: 10.1145/3459637.3481998]
42. Figshare. URL: https://figshare.com/articles/Data [accessed 2025-03-24]
43. Rodon C, Meyer T. Self-efficacy about information retrieval on the web across all domains: a short measure in French and English. Behaviour & Information Technology 2018 Mar 16;37(5):430-444 [FREE Full text] [doi: 10.1080/0144929x.2018.1449252]
44. Sun Y, Zhang Y, Gwizdka J, Trace CB. Consumer evaluation of the quality of online health information: systematic literature review of relevant criteria and indicators. J Med Internet Res 2019 May 02;21(5):e12522. [doi: 10.2196/12522] [Medline: 31045507]
45. Health On the Net Foundation. URL: https://www.hon.ch [accessed 2025-03-25]
46. Jiang J, He D, Kelly D, Allan J. Understanding ephemeral state of relevance. In: CHIIR '17: Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval. New York, NY: Association for Computing Machinery; 2017 Presented at: CHIIR '17: Conference on Human Information Interaction and Retrieval; March 7-11, 2017; Oslo, Norway p. 137-146. [doi: 10.1145/3020165.3020176]
47. Haute Autorité de Santé. URL: https://www.has-sante.fr [accessed 2025-03-25]
48. CNGOF. URL: https://www.cngof.fr [accessed 2025-03-25]
49. Endofrance. URL: https://www.endofrance.org/ [accessed 2025-03-25]
50. PasseportSanté. URL: https://www.passeportsante.net/ [accessed 2025-03-25]
51. www. Pourquoi Docteur. URL: https://www.pourquoidocteur.fr [accessed 2025-03-25]
52. Le Centre de ressources et d'information sur les multimédias pour l'enseignement supérieur (CERIMES). URL: https://www.campus.cerimes.fr [accessed 2025-03-25]
53. CHU Toulouse. URL: http://www.medecine.ups-tlse.fr/ [accessed 2025-03-25]
54. Santé. URL: https://www.e-sante.fr/ [accessed 2025-03-25]
55. Journal des Femmes. URL: https://www.journaldesfemmes.fr/ [accessed 2025-03-25]
56. Wikipedia. URL: https://www.wikipedia.org/ [accessed 2025-03-25]
57. Revue Médicale Suisse. URL: https://www.revmed.ch/ [accessed 2025-03-25]
58. Mao J, Liu Y, Kando N, Zhang M, Ma S. How does domain expertise affect users' search interaction and outcome in exploratory search? ACM Trans Inf Syst 2018 Jul 17;36(4):1-30. [doi: 10.1145/3223045]
59. Wilson MJ, Wilson ML. A comparison of techniques for measuring sensemaking and learning within participant - generated summaries. J Am Soc Inf Sci Tec 2013 Jan 04;64(2):291-306. [doi: 10.1002/asi.22758]
60. Wood E, De Pasquale D, Mueller JL, Archer K, Zivcakova L, Walkey K, et al. Exploration of the relative contributions of domain knowledge and search expertise for conducting internet searches. The Reference Librarian 2016 Feb 05;57(3):182-204. [doi: 10.1080/02763877.2015.1122559]
61. Amadieu F, Salmerón L, Cegarra J, Paubel P, Lemarié J, Chevalier A. Learning from concept mapping and hypertext: an eye tracking study. Educ Technol Soc 2015 Oct;18(4):100-112 [FREE Full text]
62. Bhattacharya N, Gwizdka J. Relating eye-tracking measures with changes in knowledge on search tasks. In: ETRA '18: Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications. New York, NY: Association for Computing Machinery; 2018 Presented at: ETRA '18: 2018 Symposium on Eye Tracking Research and Applications; June 14-17, 2018; Warsaw, Poland p. 1-5. [doi: 10.1145/3204493.3204579]
63. White RW, Dumais ST, Teevan J. Characterizing the influence of domain expertise on web search behavior. In: WSDM '09: Proceedings of the Second ACM International Conference on Web Search and Data Mining. New York, NY: Association for Computing Machinery; 2009 Feb 9 Presented at: WSDM'09: Second ACM International Conference on Web Search and Web Data Mining; February 9-12, 2009; Barcelona, Spain p. 132-141. [doi: 10.1145/1498759.1498819]

XSL•FO
RenderX

64. Tamine L, Chouquet C. On the impact of domain expertise on query formulation, relevance assessment and retrieval performance in clinical settings. Information Processing & Management 2017 Mar;53(2):332-350. [doi: 10.1016/j.ipm.2016.11.004]
65. White R, Dumais S, Teevan J. How medical expertise influences web search interaction. In: SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. New York, NY: Association for Computing Machinery; 2008 Presented at: SIGIR '08: The 31st Annual International ACM SIGIR Conference; July 20-24, 2008; Singapore, Singapore p. 791-792 URL: http://susandumais.com/SIGIR2008-Expertise-pp879-white.pdf [doi: 10.1145/1390334.1390506]
66. Palotti J, Hanbury A, Muller H. Exploiting health related features to infer user expertise in the medical domain. In: Proceedings of WSCD Workshop on Web Search and Data Mining. Hoboken, NJ: John Wiley & Sons; 2014 Presented at: WSCD Workshop on Web Search and Data Mining; February 24-28, 2014; New York, NY p. e1 URL: https://tinyurl.com/4uj47xz4
67. Kritz M, Gschwandtner M, Stefanov V, Hanbury A, Samwald M. Utilization and perceived problems of online medical resources and search tools among different groups of European physicians. J Med Internet Res 2013 Jun 26;15(6):e122. [doi: 10.2196/jmir.2436] [Medline: 23803299]
68. Shariff SZ, Sontrop JM, Haynes RB, Iansavichus AV, McKibbon KA, Wilczynski NL, et al. Impact of PubMed search filters on the retrieval of evidence by physicians. CMAJ 2012 Feb 21;184(3):E184-E190. [doi: 10.1503/cmaj.101661] [Medline: 22249990]
69. Falagas ME, Pitsouni EI, Malietzis GA, Pappas G. Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. FASEB J 2008 Feb;22(2):338-342. [doi: 10.1096/fj.07-9492LSF] [Medline: 17884971]
70. De Leo G, LeRouge C, Ceriani C, Niederman F. Websites most frequently used by physician for gathering medical information. AMIA Annu Symp Proc 2006;2006:902 [FREE Full text] [Medline: 17238521]
71. Taubert M. Use of Google as a diagnostic aid. BMJ 2006 Dec 14;333(7581):1270.1-121270. [doi: 10.1136/bmj.39058.703194.3a]
72. Heighes PT, Doig GS. Intensive care specialists' knowledge, attitudes, and professional use of published research evidence: a mail-out questionnaire survey of appropriate use of research evidence in clinical practice. J Crit Care 2014 Feb;29(1):116-122. [doi: 10.1016/j.jcrc.2013.10.014] [Medline: 24331946]
73. Evans R. IMS Institute for Healthcare Informatics. Nursing Standard 2014 Apr 09;28(32):30-30. [doi: 10.7748/ns2014.04.28.32.30.s36]
74. Van Heel KL, Nelson A, Handysides D, Shah H. The factors associated with confidence in using the internet to access health information: cross-sectional data analysis. JMIR Form Res 2023 Apr 11;7:e39891 [FREE Full text] [doi: 10.2196/39891] [Medline: 37040161]
75. Hristidis V, Ruggiano N, Brown EL, Ganta SRR, Stewart S. ChatGPT vs Google for queries related to dementia and other cognitive decline: comparison of results. J Med Internet Res 2023 Jul 25;25:e48966 [FREE Full text] [doi: 10.2196/48966] [Medline: 37490317]

Abbreviations

AI: artificial intelligence
DV: dependent variable
FAIR: Findable, Accessible, Interoperable, and Reusable
HON: Health On the Net
IV: independent variable
MD: Doctor of Medicine
NHS: National Health Service
PhD: Doctor of Philosophy
RQ: research question

Edited by A Mavragani; submitted 30.05.24; peer-reviewed by M Tajafari, JL Claggett, S Ganesh; comments to author 07.10.24; revised version received 26.01.25; accepted 16.02.25; published 17.04.25.

Please cite as:
Chevalier A, Dosso C
The Influence of Medical Expertise and Information Search Skills on Medical Information Searching: Comparative Analysis From a Free Data Set
JMIR Form Res 2025;9:e62754
URL: https://formative.jmir.org/2025/1/e62754
doi:10.2196/62754
PMID:

XSL•FO
RenderX

©Aline Chevalier, Cheyenne Dosso. Originally published in JMIR Formative Research (https://formative.jmir.org), 17.04.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Application of an Innovative Methodology to Build Infrastructure for Digital Transformation of Health Systems: Developmental Program Evaluation

M Claire Buchan[1], BSc, MSc; Tarun Reddy Katapally[2,3,4], MS, PhD; Jasmin Bhawra[5], MSc, PhD

[1]School of Public Health Sciences, University of Waterloo, Waterloo, ON, Canada
[2]DEPtH Lab, Faculty of Health Sciences, Western University, London, ON, Canada
[3]Department of Epidemiology and Biostatistics, Schulich School of Medicine and Dentistry, Western University, London, ON, Canada
[4]Lawson Health Research Institute, London, ON, Canada
[5]CHANGE Research Lab, School of Occupational and Public Health, Toronto Metropolitan University, Toronto, ON, Canada

Corresponding Author:
Jasmin Bhawra, MSc, PhD
CHANGE Research Lab, School of Occupational and Public Health
Toronto Metropolitan University
350 Victoria Street
Toronto, ON, M5B2K3
Canada
Phone: 1 416 979 5000 ext 553466
Email: jasmin.bhawra@torontomu.ca

Abstract

Background: The current public health crises we face, including communicable disease pandemics such as COVID-19, require cohesive societal efforts to address decision-making gaps in our health systems. Digital health platforms that leverage big data ethically from citizens can transform health systems by enabling real-time data collection, communication, and rapid responses. However, the lack of standardized and evidence-based methods to develop and implement digital health platforms currently limits their application.

Objective: This study aims to apply mixed evaluation methods to assess the development of a rapid response COVID-19 digital health platform before public launch by engaging with the development and research team, which consists of interdisciplinary researchers (ie, key stakeholders).

Methods: Using a developmental evaluation approach, this study conducted (1) a qualitative survey assessing digital health platform objectives, modifications, and challenges administered to 5 key members of the software development team and (2) a role-play pilot with 7 key stakeholders who simulated 8 real-world users, followed by a self-report survey, to evaluate the utility of the digital health platform for each of its objectives. Survey data were analyzed using an inductive thematic analysis approach. Postpilot test survey data were aggregated and synthesized by participant role.

Results: The digital health platform met original objectives and was expanded to accommodate the evolving needs of potential users and COVID-19 pandemic regulations. Key challenges noted by the development team included navigating changing government policies and supporting the data sovereignty of platform users. Strong team cohesion and problem-solving were essential in the overall success of program development. During the pilot test, participants reported positive experiences interacting with the platform and found its features relatively easy to use. Users in the community member role felt that the platform accurately reflected their risk of contracting COVID-19, but reported some challenges interacting with the interface. Those in the decision maker role found the data visualizations helpful for understanding complex information. Both participant groups highlighted the utility of a tutorial for future users.

Conclusions: Evaluation of the digital health platform development process informed our decisions to integrate the research team more cohesively with the development team, a practice that is currently uncommon given the use of external technology vendors in health research. In the short term, the developmental evaluation resulted in shorter sprints, and the role-play exercise enabled improvements to the log-in process and user interface ahead of public deployment. In the long term, this exercise informed the decision to include a data scientist as part of both teams going forward to liaise with researchers throughout the development

XSL•FO
RenderX

process. More interdisciplinarity was also integrated into the research process by providing health system training to computer programmers, a key factor in human-centered artificial intelligence development.

(JMIR Form Res 2025;9:e53339) doi:10.2196/53339

KEYWORDS

digital health platform; citizen science; evaluation; health systems; digital health; app; innovative; digital transformation; public health; crises; communicable disease; coronavirus; chronic diseases; decision-making; assessment; thematic analysis; self-report survey; risk; artificial intelligence; AI

Introduction

Background

The current public health crises we face are global in nature, ranging from communicable disease pandemics, such as COVID-19, to endemic noncommunicable diseases and long-term care burdens [1,2]. The presence of multiple overlapping health issues, termed a *syndemic*, is further complicated in the age of polycrises. Crisis responses increasingly require coordination from multiple sectors, including, but not limited to, health care, environment, and social services [3,4]. In order to address the gaps in our current systems of care, deliberate and cohesive societal efforts are required to understand and respond to existing inefficiencies in health systems.

Digital health platforms, which range from mobile health apps and virtual care products to digital health dashboards [5,6], have immense potential to transform our health systems by increasing access to care, predicting symptoms and outcomes, and enabling rapid responses to health crises [4,6-9]. For instance, digital health platforms can enable patients to connect with their health care providers remotely [10,11], and, in turn, health care providers are able to predict risks by ethically leveraging big data from patients [12,13] and provide support by engaging with patients remotely [10,14]. More importantly, the application of digital health platforms is not limited to patients; that is, they can be used by apparently healthy individuals to self-monitor and track their health behaviors and outcomes [15,16] as well as share their data ethically with health care providers and scientists [13,17]. Given the ubiquity of digital devices [18-20], their adaptability, and their reach across geographic regions and sociodemographic groups, digital health platforms are capable of bridging existing gaps in health information and care access [6,21,22]. More importantly, such platforms, particularly with the incorporation of artificial intelligence (AI) and machine learning, can enable precision prediction of health outcomes [23,24] as well as rapid responses to help monitor, mitigate, and manage existing and emerging health crises [6,25]. While the development of digital health platforms has increased significantly in the past decade, and in particular during the COVID-19 pandemic [26], there are no standardized processes for development and evaluation to ensure that evidence-based approaches are applied using interdisciplinary expertise.

The role of digital health platforms in managing public health and promoting health care access is predicted to grow exponentially [27]. Thus, it is critical that the development of these platforms is evaluated using rigorous methods to ensure effectiveness and efficacy. The World Health Organization recently released a guide for evaluating digital health interventions [28], and many other groups have adapted methods to assess digital platform technology [29-31]. However, these guidelines do not provide directions to evaluate digital platform *development processes*, including prototype development sprints and troubleshooting [28], which are key steps that come well ahead of the actual implementation of digital interventions.

Another key challenge is ensuring data privacy and data sovereignty of digital health platforms, particularly when serving communities that have been historically disenfranchised or discriminated against. Data sovereignty refers to meaningful control or ownership of one's own data and is a critical aspect of self-governance and self-determination, particularly among Indigenous and other colonized communities [32,33]. In creating digital health platforms that serve these communities, the development process thus requires the integration of rigid privacy and data protocols, in addition to consideration of culturally appropriate features tailored to communities' specific needs [4,7,32,34]. Moreover, as technology continues to evolve, it is crucial to evaluate the development and use of AI, especially to ensure that AI is designed in a way that is truly human centered [35,36]. Evaluating development processes helps to identify and mitigate the risks associated with AI, which could include neglecting user needs, overlooking the social impact, and failing to address biases and accountability issues [37]. By systematically evaluating development processes, we can foster AI systems that are better aligned with human values and needs (ie, health services) [36].

Study Objective

Overall, the evaluation of digital health platform development has enormous consequences for eventual citizen health and well-being, as well as data safety, security, and data sovereignty [38]. The potential for causing harm to populations, the sensitivity of personally identifiable big data that are collected via digital health platforms, as well as the need to develop human-centered AI require a rigorous evaluation process, including internal pilot testing before public testing. Most importantly, a significant gap exists in the peer-reviewed literature regarding evaluation approaches for the development of evidence-based digital health platforms, particularly in academic settings, where there is a critical need for co-design [39] not only from a citizen or patient perspective, but also by cohesive engagement between research and development teams. To address this critical gap in digital health platform development, this study aimed to apply mixed evaluation methods to assess the development of a digital health platform that was exclusively developed by a research and development

XSL•FO
RenderX

team working remotely during the COVID-19 pandemic to manage, monitor, and mitigate household risk of COVID-19.

Methods

CO-Away Digital Health Platform

In an effort to address the imminent public health crisis of the COVID-19 pandemic, the CO-Away digital health platform was developed to track, manage, and mitigate household risk of COVID-19. During the pandemic, rural, remote, and Northern communities in Canada were disproportionately impacted and experienced challenges with health information, data sovereignty, and care access [40]. In the Canadian context, Indigenous communities reported gaps in access to health information and care [34,41]; thus, after conducting a comprehensive needs assessment [34], the CO-Away platform was developed to enable near real-time monitoring, rapid response, and health care data access. CO-Away comprises a progressive web application for users to manage household COVID-19 risk as well as a backend digital decision-making dashboard that visualizes aggregated and anonymized big data relayed in real time from the progressive web application [6]. Progressive web applications are a type of web application that leverage web technologies to provide a more app-like user experience, combining the best features of web applications and mobile apps [42]. Figure 1 shows screenshots of the CO-Away home page and dashboard. The CO-Away platform provides local jurisdictional decision makers access to aggregate-level data to track and respond to emerging risk patterns and trends in near real time.

Figure 1. Screenshots of the CO-Away platform interface for community members, including the platform landing page, avatar creation process, COVID-19 symptom assessment, food security request, and citizen reporter feature.

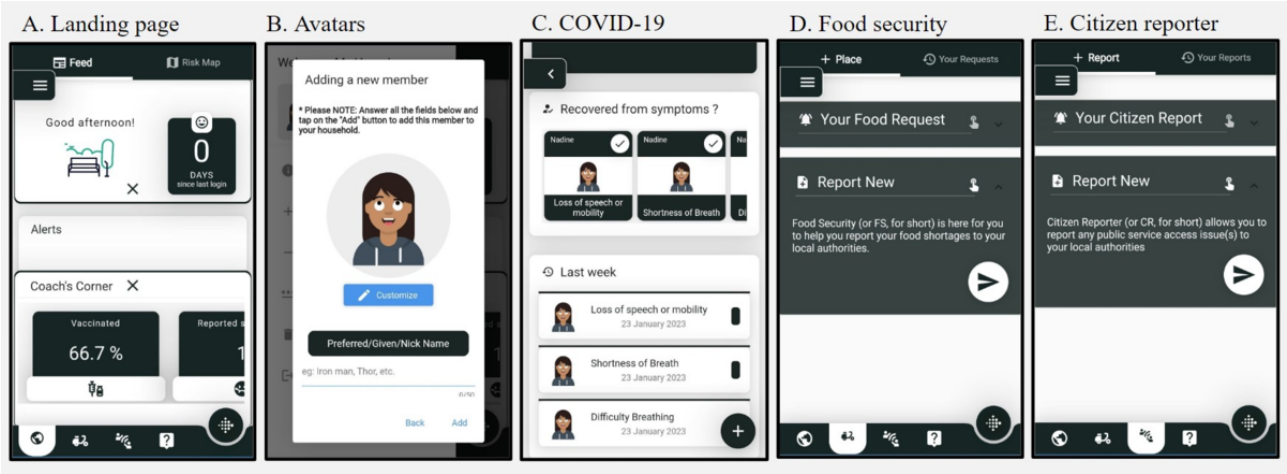

Given the wide-ranging impacts of the COVID-19 pandemic on other aspects of health and wellness, CO-Away takes a holistic approach that extends beyond COVID-19 symptom assessment to include other key digital platform features, such as (1) food security request, which helps monitor and manage food shortages within a jurisdiction and (2) citizen reporter, which reports any public service or access issues experienced by community members [6]. CO-Away serves as a link to connect citizens with decision makers and offers value to both households and decision makers. This digital health platform, driven by citizen-generated big data collected from ubiquitous tools, not only supports citizens by providing them with real-time support and valuable insights to enhance their decision-making abilities, but also aggregates and anonymizes citizen data. These aggregated, anonymized data are then delivered to decision makers, enabling real-time exchanges of information and alerts through direct bidirectional engagement between citizens and decision makers. This innovative approach represents a paradigm shift in the way community health is approached, placing priority on addressing the immediate needs of citizens. This study evaluated the development processes of both the app that citizens or community members use and the digital health dashboard that decision makers use. Additional details on the CO-Away platform can be found in another publication [43].

Evaluation Approach

To examine the evolution, challenges, successes, and utility of the CO-Away platform, a mixed methods developmental evaluation was performed after the completion of the first prototype to inform subsequent iterations of the platform. Developmental evaluations generate learnings to inform the development of an initiative; thus, are often used in complex and unpredictable scenarios, such as the COVID-19 pandemic [44,45]. A developmental evaluation is utilization focused and should be designed and implemented in ways that maximize utility for the primary intended users [46]. This evaluation approach generates data and findings in near real time, thereby facilitating developmental decision-making and course corrections throughout the development process [44,47].

Evaluation Questions

This evaluation was guided by the following overarching evaluation questions:

- What factors influenced the CO-Away digital health platform development?
- How has the CO-Away digital platform development evolved over the course of the project?
- To what extent does the CO-Away digital health platform achieve its goals and objectives from the perspective of community members or users and decision makers?

XSL•FO
RenderX

- How do these findings translate to the development process?
- What direction will the CO-Away project take going forward?

Evaluation Design

The evaluation included the following two key components: (1) a qualitative survey assessing program objectives, modifications, and challenges administered to 5 key members of the software development team and (2) a role-playing pilot test conducted with key stakeholders who were part of the research team, followed by a self-report survey, to evaluate the utility of the program for each of its goals or objectives.

Data Collection

Data collection for the CO-Away developmental evaluation included two primary components: (1) a development team survey (refer to Multimedia Appendix 1 for the survey) and (2) a research team pilot test.

Part A: Development Team Survey

A total of 5 members of the CO-Away software development team were asked to complete a web-based self-report survey. The survey was designed to assess the evolution of program objectives, target audience, and key modifications made throughout the development process. A guiding framework developed for conducting needs assessments to conceptualize and implement digital infrastructure was used to guide the development of this evaluation survey [34]. These key team members were identified by the principal investigator and the evaluation team and included software developers and data scientists. Participants were all involved in the development of CO-Away and project organization or management. Due to the potential for bias that could arise from having key members of the development team involved in the data collection process, participants were not involved in the preparation of the evaluation survey and were blinded to the responses of their fellow interviewees.

Participants were asked to complete the survey within a 1-week period from July 18, 2022, to July 25, 2022. The survey was designed to be completed in <60 minutes and was administered through Qualtrics (Qualtrics International Inc) [48]. Participants were asked to describe their role in the digital platform design, any major deviations from the initial project proposal, any changes in the target audience, and the digital platform's potential for impact using open text fields. Participants were also asked to reflect on the barriers to and facilitators of developing the CO-Away platform within the context of the COVID-19 pandemic.

Part B: Pilot Test

To evaluate the utility of the CO-Away platform for each of its objectives, a role-playing pilot test was conducted with 7 key stakeholders who were part of the interdisciplinary research team. All project stakeholders had expertise in digital health, epidemiology, and public health. Project stakeholders assumed the role of 1 of the 8 character roles developed by the evaluation team. These character roles consisted of 6 (75%) community members and 2 (25%) decision makers. One project stakeholder was asked to assume 2 roles within the pilot test, both of a community member and a decision maker, to reflect a real-world scenario where decision makers can be community members requiring access to both digital dashboards, that is, dual roles. The evaluation team developed the characters to simulate a range of potential user backgrounds and provided each participant with a brief overview of their character detailing their age, household composition, occupation, remote work ability, vaccination status, and food security status (Figure 2). Participating stakeholders reviewed the list of character descriptions and submitted their preferred roles to the evaluation team.

XSL•FO
RenderX

Figure 2. Overview of the 8 pilot test character descriptions.

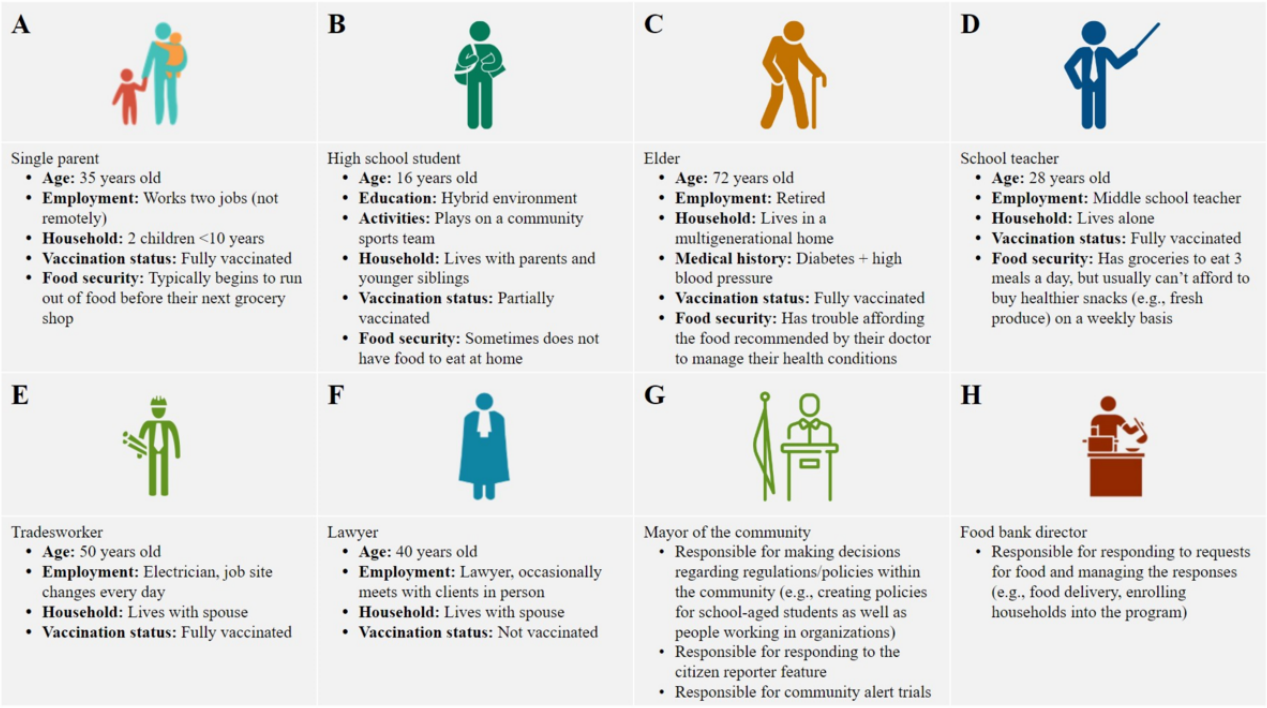

Participants were assigned their character role for the duration of the pilot test and were asked to use their provided character descriptions to create avatars within the CO-Away app. Participants were advised to use their assigned character description as a starting point and to interact with the platform as they saw fit to simulate real-world conditions. Participants were instructed to improvise their engagement with specific features of the platform (eg, COVID-19 diagnoses and symptoms). The participant who assumed both a decision maker and community member role set up 2 distinct profiles and was asked to engage with the platform (to the best of their ability) under each role independently. The pilot test was conducted over a 2-week period from July 8, 2022, to July 21, 2022. Over the course of the activity, participants were sent daily reminders via email to input their data into the app using the integrated features.

Following the pilot test, participants were asked to complete a brief self-report survey detailing their experience using the app by giving them a variety of response options, including yes or no and Likert scale questions. The survey assessed the general usability of the app and specific app features, including COVID-19 risk, food security, and citizen reporter features. Two surveys were developed by the evaluation team, one for community member participants (Multimedia Appendix 2) and the other for the decision maker participants (Multimedia Appendix 3). The survey administered was dependent on the participants' assumed roles in the study (ie, community members or decision makers), while some general questions were asked of both groups, and most questions differed between groups. The one participant who assumed both a community member and a decision maker role was asked to complete both surveys. The survey was designed to be completed in 30 minutes and was distributed electronically using Qualtrics [48].

By accessing the cloud-based database [6], daily user engagement reports were generated over the course of the pilot test to examine user engagement across the features of the platform. Reports included data on user engagement (eg, the number of new sign-ups and total log-ins), platform usability (eg, the number of households reporting their behaviors, the number of food requests, and the number of food requests processed and delivered), and reported issues (eg, the number of reported issues using the platform).

Ethical Considerations

The research and development team who participated in this project either provided simulated data as part of a role play scenario, or were subject matter experts, respectively. This developmental evaluation enabled testing of the functionality, usability, and areas of improvement for the digital platform, thus standard research ethics requirements did not apply [48]. However, survey response data remained anonymous to protect participant privacy, and survey responses were not associated with individual software development team members. Moreover, all data collected in the role-play exercise were simulated hypothetical scenarios, did not reflect true participant behaviors, and thus did not require strict anonymity and privacy protocols. Participants were not compensated for their participation in this evaluation.

Analysis

Data from the development team survey were downloaded from Qualtrics and synthesized in Microsoft Word. Data were analyzed using an inductive thematic analysis approach. To generate an initial coding manual, 1 evaluation team member (MCB) independently reviewed the first survey and organized the data into themes and meaningful codes. A second evaluation team member (JB) subsequently reviewed the coded transcript to ensure consistency and agreement. Once all transcripts were

XSL•FO
RenderX

coded by 1 evaluation team member (MCB), the evaluation team (MCB and JB) reconvened to discuss relationships between codes that could be grouped into themes.

Data from the postpilot test survey were downloaded from Qualtrics and aggregated in Microsoft Excel. Data were analyzed and summarized for each question and results were grouped by community members and decision makers. Result tables were sorted according to question topic areas (eg, COVID-19, food security, and citizen reporter). Quotations in the result tables indicated participant responses to open text fields on the survey. For questions where participants could select more than 1 option, responses were totaled across categories, and, as such, percentages may add to more than 100.

Results

Part A: Development Team Survey

Overview

All 5 members of the software development team completed the survey. The roles of the participating team members included the principal investigator (n=1, 20%), 3 (60%) software developers—one of whom was involved in project organization and management—and 1 (20%) project coordinator. The 5 subsequent subsections describe the key themes that arose through analysis of the development survey.

Theme 1: Alignment With Original Objectives

Over the course of the development process, several modifications to the app structure were made, many of which were made to accommodate the evolving nature of the COVID-19 pandemic. These changes included modifying existing features of the app (ie. the COVID-19 feature) and the addition of new features (eg, the food security feature and citizen reporter). Despite these modifications, the overall objective of CO-Away remained consistent throughout development. One participant noted the following:

> *The prototype satisfies the originally planned platform's mandate. The changes that took place were more logistical, which were beyond the scope of the development team. The critical takeaway is that the concept and the ultimate impact remain the same.* [DT5]

In addition, the team noted improvements in the visual appearance of the app. Overall, the development team indicated that the final app not only met the original objectives but also the plan expanded to accommodate the evolving needs of the community and COVID-19 regulations.

Theme 2: Project Challenges

There were a series of challenges identified by the development team, including tailoring the platform to a diverse audience (ie, both young and older community members), learning new software development tools (eg, programming languages), and the strict regulations surrounding COVID-19–related platforms on mobile app stores. The CO-Away app required integrating various disciplines to address community health needs, which was described as a unique challenge:

> *As this project's goal is to evaluate a COVID-19 mitigating digital health platform for Indigenous self-governance, determination, and data sovereignty in a remote [Indigenous] community...the initiative from its inception has been complex not only due to the integration of multiple disciplines [epidemiology, data science, Indigenous health, public health, among others], but also due to the ultimate purpose of the digital platform: rapidly responding to community needs.* [DT5]

A digital platform developed for a remote Indigenous community app would have a substantially greater chance for success if it took a holistic approach, reiterating Traditional Indigenous Knowledge of holistic health. This approach required continuous collaboration with the citizen scientist advisory council [6,34], and a part of this process required developing strategies and features within the app that manage and mitigate other population health crises that were worsened by the COVID-19 pandemic (eg, food security, mental health and substance misuse, and negative interactions of youth with law enforcement).

One of the key challenges that was pertinent for COVID-19 app development, in particular, was constantly evolving government policies. Participants reported that several previously existing features required repeated modifications, and features not initially outlined in the project plan required development on the go. For example, one participant commented that calculating their risk of COVID-19 required incorporating evolving guidelines on vaccination dosage:

> *...changes with increase in the number of doses of vaccination.* [DT4]

The nature of the COVID-19 pandemic not only required the software development team to adapt to changing government policies but also to the evolution of the virus variants themselves. Another participant highlighted the addition of the vaccine passport feature to the app, which was not within the original project scope.

While the ability of team members to work remotely was noted as a facilitator of project execution, participants underscored the delays that resulted from the constant need to update program features to align with government policies. Of high importance were the challenges associated with supporting the data sovereignty of the platform users. One participant commented the following:

> *We have had to go above and beyond existing data safety regulations in the world to build a digital health platform that provides Indigenous citizens control and ownership over their data.* [DT5]

Despite the plethora of challenges posed by the COVID-19 pandemic, the software development team was able to adapt their approach and practices to build a successful app.

Theme 3: Factors Influencing App Development

There was a range of factors that influenced app development, particularly for an app focused on addressing the rapidly changing virus during the COVID-19 pandemic. In particular,

XSL•FO
RenderX

participants noted that working during the COVID-19 pandemic posed specific barriers to the recruitment of developers. Moreover, navigating the development process dominated by, and dependent on big technology companies (Big Tech), as well as a shift to a new academic institution for the principal investigator were key external factors that influenced overall app development. Recruitment challenges and a change in academic institution created logistical challenges that added to the complexity of this large-scale project:

> *Due to the restriction of movement during the pandemic, we faced significant difficulties to recruit high-quality personnel. This issue is especially challenging in the disciplines of computer science and data science, where most of the personnel who fit our criteria are recent international graduates, who face IRCC [Immigration, Refugees and Citizenship Canada] backlogs for work permits.* [DT5]

One challenge noted by several participants was launching a platform in Big Tech app stores, which ultimately resulted in the creation of a progressive web application. For example, one participant stated the following:

> *To truly ensure self-determination and data sovereignty, we had to abandon launching our digital health platform on Apple Store and Google PlayStore to eliminate the control of these stores over the development process. This decision meant the development of a progressive web-based platform that does not need these stores to be launched and the platform would only be provided to the community members via a password protected process.* [DT5]

The 2 primary app hosting Big Tech companies, Apple Store and Google Play Store, implemented strict regulations around launching COVID-19–related platforms on their mobile app stores, which hindered the development process. While this transition created delays and roadblocks that the development team needed to overcome, it was essential in honoring the team's commitment to data sovereignty and self-determination.

Overall, the development team indicated that strong communication, workload distribution and delegation, and continuous evaluation were key internal factors that positively contributed to the platform development:

> *The communication was really strong between the team despite working from home and not working physically in the same place.* [DT2]

Participants also noted that having an agile development mindset as a team was helpful, particularly when collaborating with a citizen scientist advisory council consisting of project stakeholders. Digital literacy was also identified as an important component of app usability by the development team. Therefore, it was important for the development team to increase the accessibility of the app by modifying traditional surveys (ie, sociodemographic questions) to include more visuals and graphics while minimizing text where possible (Figure 1).

Theme 4: Project Successes and Areas of Improvement

Participants noted three primary moments of success: (1) the development of the household risk management feature, (2) the launch of the progressive web application, and (3) the clearance of the final software tests. One participant noted, in particular, the following:

> *The constant variation of policies and virus strains resulted in the creation of household risk management, rather than individual risk management, which I think is going to change the way we develop risk management tools for infectious diseases in the future.* [DT5]

When asked which factors stood out in making the development process successful, participants underscored the importance of the team dynamic. In particular, participants noted the strong leadership, cohesion, ability to problem solve, and their commitment and dedication.

Participants indicated that improving the efficiency of the team overall could be beneficial moving forward, "work on developing a more efficient workload distribution model" (DT4), particularly through simplifying platform features or implementing agile methodology [49] (ie, managing a project by breaking it up into several phases).

One participant noted that the team changed their internal sprint processes and incorporated internal evaluations of ongoing progress:

> *We had to...rewrite digital application development sprint protocols.* [DT5]

As described in theme 3, the development team experienced challenges launching health-focused apps, particularly restrictions for COVID-19 apps; thus, the team had to shift to a progressive web application. One participant noted that if they were to restart the project today, they would start with a progressive web application as it would have streamlined project development quite substantially.

Theme 5: App Launch Requirements

The participants identified a range of factors that would enable a successful launch. Most of these factors focused on successful uptake; hence, the responses described the importance of awareness and education. When asked to describe some of the factors that would contribute to a successful app launch, participants noted the platform's adaptability (ie, "rapid adaptation to changing scenarios"), consistent team engagement, and project planning. Strong community leadership and participation was highlighted as an important contributor to the overall success of the platform. One participant noted the following:

> *The most important factor that I think will contribute to a successful app launch is education and awareness [of the community in which the app is launched].* [DT3]

Taken together, these results suggested that despite the many challenges faced by the development team, their strong team cohesion, communication, and problem-solving were all essential in the overall success of program development.

XSL•FO
RenderX

Part B: Pilot Test

A total of 6 stakeholders of the interdisciplinary research team participated in the pilot test, playing the role of citizens or community members and decision makers to simulate real-world conditions. Participant character profiles included a diverse range of ages, occupations, household compositions, COVID-19 risk factors, and food security statuses (Figure 2). Over the course of the 2-week pilot test, there were an average of 7 (SD 3) log-ins per day across an average of 4 users per day (SD 1). The largest proportion (average: 40%, SD 18%) of the daily log-ins occurred in the evening.

In total, 5 (63%) of the 8 participants reported COVID-19 diagnoses, with 21 hospital visits reported over the duration of the pilot test. Among the 8 participants, 2 (25%) were fully vaccinated, 3 (38%) were fully vaccinated with boosters, and 2 (25%) were unvaccinated. Over the duration of the 2-week pilot test, participants reported inconsistently wearing face masks (n=38), making essential shopping trips (n=48), having outdoor visits with other people (n=49), having social gatherings (n=38), and travelling (n=11).

A total of 4 food security requests were placed over the 2-week period, all of which were processed by decision makers using the digital health dashboard. One participant submitted a citizen report, which was processed by a decision maker, and 5 community alerts were issued, all of which were categorized as high urgency.

Results from the postpilot survey are presented in Tables S1 and S2 in Multimedia Appendix 4 for community members and decision makers, respectively.

Community Members

Usability

All (6/6, 100%) participants felt that their avatars accurately reflected their COVID-19 risk, found the consent process clear, and appreciated the anonymity it provided (Figure 1B). All (6/6, 100%) participants reported that the app was easy to navigate and that they could find each feature when needed. Participants found it straightforward to set up their avatars, taking ≤5 minutes. In total, 33% (2/6) of the participants created avatars for multiple household members, while the other 67% (4/6) created avatars for themselves only. According to participants, the ideal frequency of notifications from the app was once per day. No (0/6, 0%) participants reported any issues in using the app.

Feature-Specific Feedback

Most (4/6, 67%) of the participants reported that they were comfortable interacting with the platform's COVID-19 feature, with 33% (2/6) of the participants reporting that they were neither comfortable nor uncomfortable. All (6/6, 100%) participants felt that the recommendations for their COVID-19 risk were clear and easy to understand, though some (2/6, 33%) provided feedback to increase convenience and subsequent usability (eg, readily accessible links or saved log-in information). Nearly all participants (5/6, 83%) were comfortable sharing their vaccination status. Issues such as random logouts and confusion with report creation were raised, though the feature was still found to be easy and clear to use.

Most (5/6, 83%) of the participants used the food security feature, all of which found it "easy" to use. A total of 60% (3/5) of the participants expressed comfort with using this feature, with the 40% (2/5) remaining participants reporting that they were "neither comfortable nor uncomfortable." Moreover, 60% (3/5) of the participants reported that the response of the decision makers was "neither appropriate nor inappropriate," while the remaining 40% (2/5) of the participants reported "appropriate" responses. Similarly, half (3/5, 60%) of the participants were "comfortable" revealing their identity when support was needed, while the remainder (2/5, 40%) felt "neither comfortable nor uncomfortable." When asked about room for improvement, 40% (2/5) of the participants suggested the inclusion of more detailed options for food security requests, 40% (2/5) of the participants reported issues viewing responses to their submitted food requests, and 20% (1/5) of the participants desired additional instructions for interacting with the feature.

Of the 6 participants, 3 (50%) used the citizen reporter feature and felt comfortable doing so. A total of 2 (67%) of the 3 participants found it easy to use, and 1 (33%) participant experienced difficulty during the report submission process, expressing uncertainty regarding the completeness of their submission due to the lack of response. Moreover, 1 (33%) participant reported an appropriate response from the decision maker, while the remaining 2 (67%) participants reported neither appropriate nor inappropriate responses. In total, 2 (67%) of the 3 participants felt comfortable revealing their identity when seeking support, while the third (33%) participant felt neither comfortable nor uncomfortable.

Overall Feedback

Participants generally found the app useful, providing clear information and serving as a good resource. The straightforwardness, accessibility, and organization of the app were highlighted along with the easy navigation and relevant tips. However, there were mentions of difficulty in finding where to input information initially and occasional log-in issues requiring password resets. Some (2/6, 33%) participants also mentioned challenges when using the app on a phone, particularly related to limited screen visibility.

Decision Makers

Usability

Decision makers reported on the usability of the digital dashboard, including ease of finding information, navigation, and overall use. Both decision makers found the information used to create their avatars accurate in reflecting their COVID-19 risk. They were satisfied with the clarity of the consent process and believed the app was clear and easy to navigate, ensuring anonymity and accessibility to all features. No issues or improvement suggestions were reported regarding navigability, and participants felt that they knew who to contact with questions regarding the app, their data, or their rights. Both decision makers had no difficulty interacting with notifications, finding them organized and user-friendly.

XSL•FO
RenderX

Feature-Specific Feedback

Figure 3 shows examples of how community member data were aggregated and visualized for decision makers to review. Participants had a positive experience with the dashboard data visualizations of the dashboard, finding them easy to understand and appreciating the level of control they had over them. Decision makers felt confident in using the presented data to make informed decisions regarding the community's response to the COVID-19 pandemic. Response rates to individual incidents were relatively short, typically taking less than 5 minutes to complete. No suggestions for improvement or increasing participants' confidence were provided.

Figure 3. Screenshots of the CO-Away platform interface for decision makers, including the general dashboard, the community behavior dashboard, and the community exposure level dashboard.

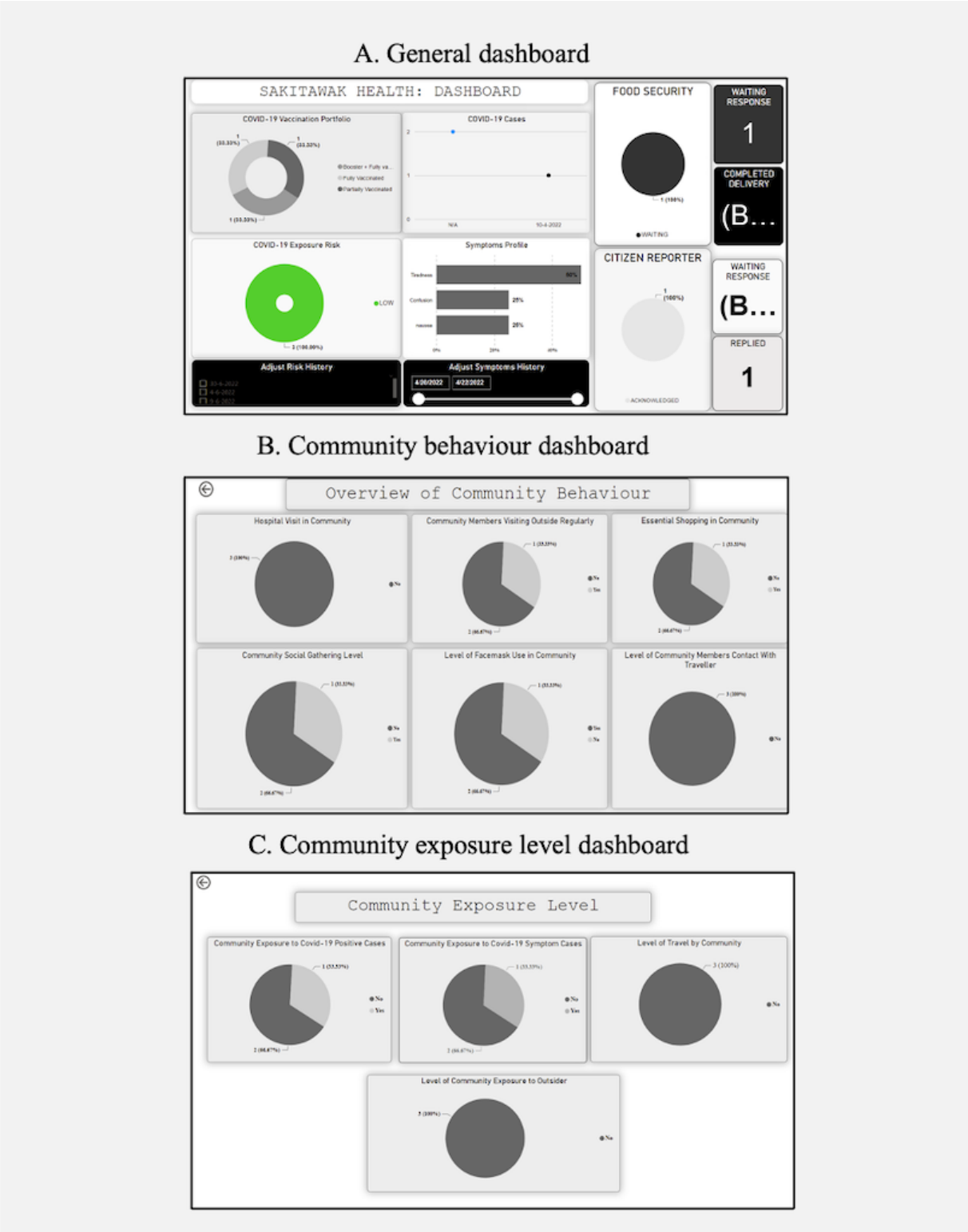

XSL•FO
RenderX

Similar to the COVID-19 feature, both decision makers found the data visualizations easy to understand and appreciated the appearance and level of control over them. Decision makers expressed confidence in using the presented data for informed decision-making regarding the COVID-19 pandemic response. Response rates for individual incidents were deemed prompt, with completion times less than 5 minutes. No suggestions for improving confidence or the visualizations were given.

Only 1 of the 2 decision makers interacted with the citizen reporter feature, reported that the data visualizations were easy to understand, and appreciated their appearance and level of control. The decision maker expressed confidence in using the presented data for informed decision-making, and the participant considered the response times for individual incidents, completed in less than 5 minutes, to be appropriate. No suggestions for improving the visualizations or participant confidence were given.

Overall Feedback

Participants had a positive overall experience with the platform, with a participant suggesting the addition of reference material, such as a brief presentation or a short video, for future use. The food security section was found to be easy to navigate and straightforward, providing users with simplified access. However, there were uncertainties regarding the ability to delete requests and whether citizens could contact local decision makers and/or food services for questions or concerns, as a participant did not receive a notification when attempting to do so.

Discussion

Principal Findings

Evaluation of the CO-Away digital health platform enabled systematic capture of the development team's perspective on the overall development process as well as the testing of platform usability from multiple stakeholders simulating real-world scenarios before finalizing platform features. This evaluation demonstrated that while the final digital health platform met its original objectives, it had to be expanded to accommodate the evolving needs of potential users and the COVID-19 pandemic regulations. Strong team cohesion, communication, and problem-solving were essential to the overall success of platform development. The pilot test found that participants reported positive experiences interacting with the platform and found its features relatively easy to use, with a few community members reporting challenges interacting with the interface. Users in the decision maker role found the data visualizations intuitive in helping them to understand the information.

Relevance and Application of Findings

The developmental evaluation findings will not only influence the decisions of both research and development teams going forward, but the process described for conducting an assessment of the CO-Away development phase also importantly advances a replicable methodology that can inform empirical development, implementation, and evaluation of digital health platforms [6]. The field of digital health evaluation is rapidly evolving; therefore, emerging research on innovations in digital technology and evaluation practices to assess impact and utility [28,38,50-52] have emerged over the past 5 years. The World Health Organization [28] has released guidelines for monitoring and evaluating digital health interventions, with many existing mixed methods (ie, qualitative focus groups and quantitative surveys) being applied to digital health platform evaluations. However, there is little guidance on how and when to apply these evaluation methods in the scheme of a digital health project [53,54]. In particular, there is a lack of peer-reviewed literature describing the evaluation of the development of evidence-based digital health platforms.

When working with marginalized populations, such as Indigenous or global south communities, evaluation of the development process is a critical step in ensuring that a digital platform is safe, accessible, and not widening existing inequities. Specific considerations related to data ownership, cultural competency, and privacy must be built into the digital platform development protocol and rigorously tested [34,52,53]. Thus, this developmental evaluation was designed by initially referring to existing evaluation literature to inform the key evaluation questions and study approach [38,44-47]. Given the dearth of evidence on *how* to practically adapt and apply digital evaluation methods [53], our study aimed to address this gap by implementing an innovative mixed methods approach informed by previous evaluation work in Indigenous communities [34] and evaluation literature [29,44-47,55-59].

The use of role-play has been demonstrated in other contexts [55-59] where the application of a digital intervention or program can be simulated before public deployment. For example, during the evaluation of a digital platform designed to improve the reproductive health of adolescents in Rwanda, a role-play exercise was used during the prototyping stage of development to evaluate the end-to-end user experience [29]. During the role-play exercise, participants were provided with mock scenarios and roles to help the design team understand how users interacted with the interface during the improvised scenario-specific performances. Similar to the approach used in this developmental evaluation, feedback was requested from participants following the role-play exercise to gain an understanding of user experience and areas of improvement. This step is essential not only when working with marginalized communities but also for launching digital platforms focused on sensitive health issues, that is, when potential risks are heightened, thereby requiring more rigorous testing protocols. Identifying qualified participants for role-play can be a challenge in research studies; therefore, this evaluation uniquely used an interdisciplinary team to conduct the exercise during the development phase to minimize commonly reported challenges with role-play and simulate a real-world scenario before the public launch [56].

It was through this evaluation that our internal testing of the platform generated key learnings that informed specific project decisions, including integrating the research team more cohesively with the development team. In academic research, a team of external technology vendors is commonly used to develop digital platforms. As a result, the external development team is often not embedded in the research and evaluation

XSL•FO
RenderX

process. In order to effectively evaluate the digital development process, our team identified the need to involve a data scientist as part of both the development and research teams going forward (whereas previously, only a researcher was liaising with the development team). Including a common computer programmer or developer in both teams improved communication and coordination during the development process. By bridging these traditionally distinct teams, we have been able to maximize cross-disciplinary collaboration, with research and practice informing development decisions. This integration also proved instrumental in aligning the platform's features with users' needs and developmental feasibility, that is, integrating interdisciplinarity into the development process.

In recognizing the potential of concurrent health domain expertise and software development skills, we initiated a health system training for all computer programmers [60,61]. This training contributed to increasing interdisciplinary understanding and communication and was deemed necessary for the creation of user-centred CO-Away feature development, which considered distinct risks that could be introduced with AI systems [36,37]. This training not only supports the programmers' comprehension of the health context but also provides them with the ability to envision the platform's functionalities from the user's perspective. This is a key step in both co-designing digital infrastructure as well as designing human-centered AI [35,62], which is a gap in the current literature [63-66]. In general, this approach resulted in digital health platform features that were user-friendly, that is, an improved user experience, which is critical to the success of platform use and implementation [64,67].

The developmental evaluation also changed development sprint processes [49,68], which paved the way for shorter sprints with a quick internal evaluation of ongoing progress. Software development sprints, which are part of the agile coding process, are implemented as part of the agile methodology, an iterative project management framework that breaks projects down into several dynamic phases, commonly known as "sprints" [49]. When the team started the CO-Away development process, sprints were approximately 4 to 5 months, but the developmental evaluation findings informed our decision to shorten the sprints to 3 to 4 months to enable faster internal testing and modification of the digital platform. This decision has allowed the software development team to pivot quickly when necessary, responding to stakeholder feedback and incorporating changes to the digital platform iteratively.

The evaluation provided noteworthy insights into the dynamics of our team, particularly the role that strong communication played throughout the development process. The COVID-19 pandemic presented a unique set of challenges that required adaptive strategies [69-71]. Despite the physical distancing mandates, our team managed to function cohesively, leveraging digital tools, regular virtual meetings, and transparent channels of communication, that is, open discussion of weekly updates and key decisions accessible to all team members on a virtual portal. The development process not only involved remote hiring of our team [72] but also remote functioning [73] and retention of our team even after the pandemic restrictions ended, lessons which we implemented into the continuous functioning of our development processes. For instance, we currently work in a hybrid setting [74], which leverages the flexibility of virtual work as well as in-person brainstorming, a system that has further improved communication and collaboration with research and development teams.

Perhaps the defining element that shaped the development process of the CO-Away platform was the complexities that arose with its COVID-19–specific focus. The evolving nature of the coronavirus [75-77], coupled with the rapidly changing landscape of pandemic management policies [78-80], presented a constant need for flexibility and adaptability. The virus's mutations necessitated continuous updates to the platform's algorithms to ensure that users would have access to the most up-to-date information. Similarly, the ever-changing policies surrounding lockdowns [81,82], vaccine requirements [83,84], and mask mandates [85,86] required the platform's features to be responsive to shifting guidelines [6]. Navigating these complexities highlighted the importance of further adapting the agile methodology [49,68] throughout the platform design and development process to accommodate the changing COVID-19 landscape.

The role-play pilot test highlighted the effectiveness of adhering to user-centered design principles throughout the development of the CO-Away digital health platform. The stakeholders who participated in the pilot reported overall positive reactions to the platform avatars, transparent consent procedures, and intuitive navigation—aspects that reiterated the importance of tailoring the platform design to users' needs and ensuring a user-friendly experience [64,67]. While participants generally found the app user-friendly, their feedback on minor obstacles, such as entering initial information and sporadic log-in difficulties, highlighted the importance of iterative improvement during the development process. This constructive feedback provided an opportunity for the development team to adapt CO-Away to meet the needs of users.

The mixed feedback regarding participants' comfort levels when interacting with the platform features reveals the importance of balancing interactivity with user comfort. For instance, while some participants were comfortable engaging with various features and sharing personal information for support, others expressed hesitancy. This finding highlights the significance of providing users with control over their data, a critical aspect of data sovereignty [32] that has been built into platform development after this vital input.

While we did not observe marked differences in the way participants engaged with the platform based on the assigned participant roles, it is worth noting that there were differences in platform use between decision makers and community members. These variations align with expectations stemming from the differing ways these distinct user groups engage with the platform's features. For instance, decision makers are less focused on their own privacy as they are not sharing data, while the community members are sharing data. We anticipate that various community members will interact with and use the app according to their specific circumstances and needs. As such, the potential differences in app use based on community member

XSL•FO
RenderX

roles will be further explored in the future during the external community pilot phase.

Finally, the digital health dashboard feature of CO-Away offered decision makers with valuable tools for informed decision-making, especially within the context of COVID-19 response and community management. The pilot decision makers appreciated the usability of the platform and expressed confidence in using the data visualizations for informed decision-making. However, it is important to recognize that although some real-world scenarios were simulated in the pilot, a larger external community pilot will require more complex decision-making due to the increased scale of implementation. However, the key goal of the pilot was to test the functionality of a cloud-based digital health dashboard to engage with community members and send real-time alerts, features that were successfully tested and functionality that was confirmed in the evaluation of the pilot. Real-world policy changes using real-time big data would require stakeholder buy-in and addressing nuanced matters, such as food security and citizen reporter reports, demanding sensitivity and community-wide consideration. Further exploration of these dynamics will be carried out during our upcoming external community pilot.

Key Guidelines for Digital Health Platform Development

On the basis of the findings of this developmental evaluation of a digital health platform development, we suggest the following guidelines:

1. Integration of research and development teams is key to the success of digital health platforms, where the software developers are provided an opportunity to understand scientific goals, and the research stakeholders are made privy to cloud computing challenges.
2. Conducting a developmental evaluation after the completion of a prototype facilitates the incorporation of research and development team perspectives empirically into the iterative digital development process before a product is tested widely in the community.
3. Conducting an internal role-play pilot simulating real-world conditions is critical to test not only the functionality of digital health platforms but also the nuanced perceptions of potential users, which will be essential for successful community implementation.

Strengths and Limitations

This developmental evaluation has several notable strengths. First, we deployed a mixed methods approach to data collection and analysis; open-ended survey questions to capture participant perceptions of the platform and its usability and CO-Away use metrics data provided an objective assessment of platform engagement. The integration of both quantitative and qualitative data collection methods enriches the findings generated by the evaluation. Second, the inclusion of development team surveys allowed for an in-depth understanding not only of CO-Away's developmental process but also offered valuable insights that have and will continue to inform improvements in the development processes. Finally, the use of "character roles" was used to simulate a community pilot test. Although not a substitute for a real-world pilot, this approach served as a valuable first step for exploring real-world simulations using varied user interactions and experiences, an approach that will shape a future external community pilot. Providing fictional role-play characters through this remote evaluation also avoided potential issues, such as staying in character [56]. There were also several study limitations. The participation of the stakeholders in the pilot test may have affected the nature of the issues reported, that is, more focus on app function than readability. In order to minimize potential bias, the evaluation questions were blinded, and the digital platform development team did not participate in the pilot test. Moreover, the stakeholders who participated in the pilot did not analyze the evaluation data. The use of "character roles," while insightful, does not capture all user scenarios that would arise in a real-world setting. As such, it is possible that certain user interactions and challenges were not captured in this evaluation and will be explored during our external community pilot. Notably, the sample size for each population was small, so caution should be taken when interpreting or comparing percentages. Another limitation pertains to the digital literacy of participants. While the pilot participants did not have previous experience using the platform, it is possible that few issues were reported due to the inherent high digital literacy of the interdisciplinary stakeholders. In future digital evaluations, it would be imperative to not only capture the digital literacy of platform users but also to enable the increase of digital literacy through innovative digital literacy programs [87].

Conclusions

The evaluation of digital health platform development is critical for the success of not only platform functionality but also eventual implementation and scale-up. The innovative mixed methods approach applied in this evaluation combines both perspectives of the development team as well as a real-world simulation of platform users to extend key guidelines for digital health platform development. For sensitive community health issues, role-play simulation and testing before public launch can avoid potential risks and improve user experience. The evaluation informed several key decisions of the digital health platform development, including integrating the research team more cohesively with the development team, which resulted in a data scientist being part of both teams going forward. Another key development process decision was to integrate more interdisciplinarity into the development process by providing health system training to computer programmers and shortening development sprints. Combined with more efficient sprints, the role-play exercise importantly enabled improvements to the log-in process and the user interface ahead of public deployment.

XSL•FO
RenderX

Acknowledgments

The authors would like to acknowledge Nadine Elsahli for her support with this manuscript, including the creation of figures to visually depict the digital health platform design. The authors would also like to thank all members who participated in the developmental evaluation process and provided valuable feedback throughout.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Development survey.
[DOCX File , 14 KB - formative_v9i1e53339_app1.docx]

Multimedia Appendix 2
Postpilot survey—community member.
[DOCX File , 17 KB - formative_v9i1e53339_app2.docx]

Multimedia Appendix 3
Postpilot survey—decision maker.
[DOCX File , 18 KB - formative_v9i1e53339_app3.docx]

Multimedia Appendix 4
Evaluation survey results.
[DOCX File , 152 KB - formative_v9i1e53339_app4.docx]

References

1. Alokaily F. COVID-19 global health crisis. Saudi Med J 2021 Jan 25;42(1):3-4 [FREE Full text] [doi: 10.15537/smj.2021.1.25663] [Medline: 33399164]
2. Noncommunicable diseases. World Health Organization. 2024 Dec 23. URL: https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases [accessed 2025-03-22]
3. Lawrence M, Homer-Dixon T, Janzwood S, Rockstrom J, Renn O, Donges JF. Global polycrisis: the causal mechanisms of crisis entanglement. SSRN J 2023 Jun 27 [FREE Full text] [doi: 10.2139/ssrn.4483556]
4. Katapally TR. A global digital citizen science policy to tackle pandemics like COVID-19. J Med Internet Res 2020 May 26;22(5):e19357 [FREE Full text] [doi: 10.2196/19357] [Medline: 32408267]
5. Sheikh A, Anderson M, Albala S, Casadei B, Franklin BD, Richards M, et al. Health information technology and digital innovation for national learning health and care systems. Lancet Digit Health 2021 Jun;3(6):e383-e396. [doi: 10.1016/s2589-7500(21)00005-4]
6. Katapally TR, Ibrahim ST. Digital health dashboards for decision-making to enable rapid responses during public health crises: replicable and scalable methodology. JMIR Res Protoc 2023 Jun 30;12:e46810 [FREE Full text] [doi: 10.2196/46810] [Medline: 37389905]
7. Katapally TR. The SMART framework: integration of citizen science, community-based participatory research, and systems science for population health science in the digital age. JMIR Mhealth Uhealth 2019 Aug 30;7(8):e14056 [FREE Full text] [doi: 10.2196/14056] [Medline: 31471963]
8. Stoumpos AI, Kitsios F, Talias MA. Digital transformation in healthcare: technology acceptance and its applications. Int J Environ Res Public Health 2023 Feb 15;20(4):3407 [FREE Full text] [doi: 10.3390/ijerph20043407] [Medline: 36834105]
9. Budd J, Miller BS, Manning EM, Lampos V, Zhuang M, Edelstein M, et al. Digital technologies in the public-health response to COVID-19. Nat Med 2020 Aug 07;26(8):1183-1192. [doi: 10.1038/s41591-020-1011-4] [Medline: 32770165]
10. Rosenlund M, Kinnunen UM, Saranto K. The use of digital health services among patients and citizens living at home: scoping review. J Med Internet Res 2023 Mar 27;25:e44711 [FREE Full text] [doi: 10.2196/44711] [Medline: 36972122]
11. Jnr BA. Use of telemedicine and virtual care for remote treatment in response to COVID-19 pandemic. J Med Syst 2020 Jun 15;44(7):132 [FREE Full text] [doi: 10.1007/s10916-020-01596-5] [Medline: 32542571]
12. Batko K, Ślęzak A. The use of big data analytics in healthcare. J Big Data 2022 Jan 06;9(1):3 [FREE Full text] [doi: 10.1186/s40537-021-00553-4] [Medline: 35013701]
13. Dash S, Shakyawar SK, Sharma M, Kaushik S. Big data in healthcare: management, analysis and future prospects. J Big Data 2019 Jun 19;6(1):54. [doi: 10.1186/s40537-019-0217-0]
14. Burton L, Rush KL, Smith MA, Davis S, Rodriguez Echeverria P, Suazo Hidalgo L, et al. Empowering patients through virtual care delivery: qualitative study with micropractice clinic patients and health care providers. JMIR Form Res 2022 Apr 27;6(4):e32528 [FREE Full text] [doi: 10.2196/32528] [Medline: 35413002]

XSL•FO
RenderX

15. Kuwabara A, Su S, Krauss J. Utilizing digital health technologies for patient education in lifestyle medicine. Am J Lifestyle Med 2020 Dec 13;14(2):137-142 [FREE Full text] [doi: 10.1177/1559827619892547] [Medline: 32231478]
16. Mair JL, Salamanca-Sanabria A, Augsburger M, Frese BF, Abend S, Jakob R, et al. Effective behavior change techniques in digital health interventions for the prevention or management of noncommunicable diseases: an umbrella review. Ann Behav Med 2023 Sep 13;57(10):817-835 [FREE Full text] [doi: 10.1093/abm/kaad041] [Medline: 37625030]
17. Riso B, Tupasela A, Vears DF, Felzmann H, Cockbain J, Loi M, et al. Ethical sharing of health data in online platforms - which values should be considered? Life Sci Soc Policy 2017 Aug 21;13(1):12 [FREE Full text] [doi: 10.1186/s40504-017-0060-z] [Medline: 28825221]
18. Mosa AS, Yoo I, Sheets L. A systematic review of healthcare applications for smartphones. BMC Med Inform Decis Mak 2012 Jul 10;12:67 [FREE Full text] [doi: 10.1186/1472-6947-12-67] [Medline: 22781312]
19. Langford AT, Solid CA, Scott E, Lad M, Maayan E, Williams SK, et al. Mobile phone ownership, health apps, and tablet use in US adults with a self-reported history of hypertension: cross-sectional study. JMIR Mhealth Uhealth 2019 Jan 14;7(1):e12228 [FREE Full text] [doi: 10.2196/12228] [Medline: 31344667]
20. Kaplan WA. Can the ubiquitous power of mobile phones be used to improve health outcomes in developing countries? Global Health 2006 May 23;2:9 [FREE Full text] [doi: 10.1186/1744-8603-2-9] [Medline: 16719925]
21. Willis VC, Thomas Craig KJ, Jabbarpour Y, Scheufele EL, Arriaga YE, Ajinkya M, et al. Digital health interventions to enhance prevention in primary care: scoping review. JMIR Med Inform 2022 Jan 21;10(1):e33518 [FREE Full text] [doi: 10.2196/33518] [Medline: 35060909]
22. Haleem A, Javaid M, Singh RP, Suman R. Telemedicine for healthcare: capabilities, features, barriers, and applications. Sens Int 2021;2:100117 [FREE Full text] [doi: 10.1016/j.sintl.2021.100117] [Medline: 34806053]
23. Shahid N, Rappon T, Berta W. Applications of artificial neural networks in health care organizational decision-making: a scoping review. PLoS One 2019;14(2):e0212356 [FREE Full text] [doi: 10.1371/journal.pone.0212356] [Medline: 30779785]
24. Vazquez-Ingelmo A, Garcia-Penalvo FJ, Theron R. Information dashboards and tailoring capabilities - a systematic literature review. IEEE Access 2019;7:109673-109688. [doi: 10.1109/access.2019.2933472]
25. Kostkova P. Disease surveillance data sharing for public health: the next ethical frontiers. Life Sci Soc Policy 2018 Jul 04;14(1):16 [FREE Full text] [doi: 10.1186/s40504-018-0078-x] [Medline: 29971516]
26. Bouabida K, Lebouché B, Pomey MP. Telehealth and COVID-19 pandemic: an overview of the telehealth use, advantages, challenges, and opportunities during COVID-19 pandemic. Healthcare (Basel) 2022 Nov 16;10(11):2293 [FREE Full text] [doi: 10.3390/healthcare10112293] [Medline: 36421617]
27. Abernethy A, Adams L, Barrett M, Bechtel C, Brennan P, Butte A, et al. The promise of digital health: then, now, and the future. NAM Perspect 2022 Jun 27;2022(22):10.31478/202206e [FREE Full text] [doi: 10.31478/202206e] [Medline: 36177208]
28. Monitoring and evaluating digital health interventions: a practical guide to conducting research and assessment. World Health Organization. 2016 Dec 12. URL: https://www.who.int/publications/i/item/9789241511766 [accessed 2023-10-31]
29. Ippoliti N, Sekamana M, Baringer L, Hope R. Using human-centered design to develop, launch, and evaluate a national digital health platform to improve reproductive health for Rwandan youth. Glob Health Sci Pract 2021 Nov 29;9(Supplement 2):S244-S260. [doi: 10.9745/ghsp-d-21-00220]
30. Beynon F, Guérin F, Lampariello R, Schmitz T, Tan R, Ratanaprayul N, et al. Digitalizing clinical guidelines: experiences in the development of clinical decision support algorithms for management of childhood illness in resource-constrained settings. Glob Health Sci Pract 2023 Jul 27;11(4):e2200439. [doi: 10.9745/ghsp-d-22-00439]
31. Tamrat T, Chandir S, Alland K, Pedrana A, Taighoon Shah M, Footitt C, et al. Digitalization of routine health information systems: Bangladesh, Indonesia, Pakistan. Bull World Health Org 2022 Oct 01;100(10):590-600. [doi: 10.2471/blt.22.287816]
32. Bhawra J. Decolonizing digital citizen science: applying the bridge framework for climate change preparedness and adaptation. Societies 2022 Apr 17;12(2):71. [doi: 10.3390/soc12020071]
33. Hummel P, Braun M, Tretter M, Dabrock P. Data sovereignty: a review. Big Data Soc 2021 Jan 22;8(1). [doi: 10.1177/2053951720982012]
34. Bhawra J, Buchan MC, Green B, Skinner K, Katapally TR. A guiding framework for needs assessment evaluations to embed digital platforms in partnership with Indigenous communities. PLoS One 2022 Dec 22;17(12):e0279282 [FREE Full text] [doi: 10.1371/journal.pone.0279282] [Medline: 36548382]
35. Bingley WJ, Curtis C, Lockey S, Bialkowski A, Gillespie N, Haslam SA, et al. Where is the human in human-centered AI? Insights from developer priorities and user experiences. Comput Hum Behav 2023 Apr;141:107617. [doi: 10.1016/j.chb.2022.107617]
36. Ozmen Garibay O, Winslow B, Andolina S, Antona M, Bodenschatz A, Coursaris C, et al. Six human-centered artificial intelligence grand challenges. Int J Hum Comput Interact 2023 Jan 02;39(3):391-437. [doi: 10.1080/10447318.2022.2153320]
37. Chen Y, Clayton EW, Novak LL, Anders S, Malin B. Human-centered design to address biases in artificial intelligence. J Med Internet Res 2023 Mar 24;25:e43251 [FREE Full text] [doi: 10.2196/43251] [Medline: 36961506]
38. Murray E, Hekler EB, Andersson G, Collins LM, Doherty A, Hollis C, et al. Evaluating digital health interventions: key questions and approaches. Am J Prev Med 2016 Nov;51(5):843-851. [doi: 10.1016/j.amepre.2016.06.008] [Medline: 27745684]

XSL•FO
RenderX

39. Sanz MF, Acha BV, García MF. Co-design for people-centred care digital solutions: a literature review. Int J Integr Care 2021 Apr 30;21(2):16 [FREE Full text] [doi: 10.5334/ijic.5573] [Medline: 33981193]
40. The impact of COVID-19 on rural and remote mental health and substance use. Mental Health Commission of Canada. 2021. URL: https://mentalhealthcommission.ca/wp-content/uploads/2021/10/The-Impact-of-COVID-19-on-Rural-and-Remote-Mental-Health-and-Substance-Use.pdf.pdf [accessed 2023-10-31]
41. Huyser KR, Yellow Horse AJ, Collins KA, Fischer J, Jessome MG, Ronayne ET, et al. Understanding the associations among social vulnerabilities, indigenous peoples, and COVID-19 cases within Canadian health regions. Int J Environ Res Public Health 2022 Sep 29;19(19):12409 [FREE Full text] [doi: 10.3390/ijerph191912409] [Medline: 36231708]
42. Rêgo F, Portela F, Santos MF. Towards PWA in healthcare. Procedia Comput Sci 2019;160:678-683. [doi: 10.1016/j.procs.2019.11.028]
43. Bhawra J, Elsahli N, Patel J. Applying digital technology to understand human experiences of climate change impacts on food security and mental health: scoping review. JMIR Public Health Surveill 2024 Jul 23;10:e54064 [FREE Full text] [doi: 10.2196/54064] [Medline: 39042453]
44. Dozois E, Langlois M, Blanchet-Cohen N. DE 201: a practitioner's guide to developmental evaluation. The J.W. McConnell Family Foundation and the International Institute for Child Rights and Development. 2010. URL: https://mcconnellfoundation.ca/wp-content/uploads/2017/07/DE-201-EN.pdf [accessed 2025-03-22]
45. Patton MQ. Emergent developmental evaluation developments. J Multidiscip Eval 2021 Jun 01;17(41):23-34 [FREE Full text] [doi: 10.56645/jmde.v17i41.699]
46. Patton MQ. Utilization-Focused Evaluation. Thousand Oaks, CA: SAGE Publications; 2008.
47. Patton MQ. Developmental Evaluation: Applying Complexity Concepts to Enhance Innovation and Use. New York, NY: Guilford Publications; 2010.
48. Tri-council policy statement: ethical conduct for research involving humans. Canadian Institutes of Health Research, Natural Sciences and Engineering Research Council of Canada, and Social Sciences and Humanities Research Council. 2018. URL: https://ethics.gc.ca/eng/documents/tcps2-2018-en-interactive-final.pdf [accessed 2025-04-02]
49. Nazir S, Price B, Surendra NC, Kopp K. Adapting agile development practices for hyper-agile environments: lessons learned from a COVID-19 emergency response research project. Inf Technol Manag 2022 Jul 30;23(3):193-211 [FREE Full text] [doi: 10.1007/s10799-022-00370-y] [Medline: 36188730]
50. Kolasa K, Kozinski G. How to value digital health interventions? A systematic literature review. Int J Environ Res Public Health 2020 Mar 23;17(6):2119 [FREE Full text] [doi: 10.3390/ijerph17062119] [Medline: 32209988]
51. Bashi N, Fatehi F, Mosadeghi-Nik M, Askari MS, Karunanithi M. Digital health interventions for chronic diseases: a scoping review of evaluation frameworks. BMJ Health Care Inform 2020 Mar;27(1):e100066 [FREE Full text] [doi: 10.1136/bmjhci-2019-100066] [Medline: 32156751]
52. Labrique AB, Wadhwani C, Williams KA, Lamptey P, Hesp C, Luk R, et al. Best practices in scaling digital health in low and middle income countries. Global Health 2018 Nov 03;14(1):103 [FREE Full text] [doi: 10.1186/s12992-018-0424-z] [Medline: 30390686]
53. Guo C, Ashrafian H, Ghafur S, Fontana G, Gardner C, Prime M. Challenges for the evaluation of digital health solutions-a call for innovative evidence generation approaches. NPJ Digit Med 2020 Aug 27;3(1):110 [FREE Full text] [doi: 10.1038/s41746-020-00314-2] [Medline: 32904379]
54. Wang T, Giunti G, Goossens R, Melles M. Timing, indicators, and approaches to digital patient experience evaluation: umbrella systematic review. J Med Internet Res 2024 Feb 05;26:e46308 [FREE Full text] [doi: 10.2196/46308] [Medline: 38315545]
55. Skoura-Kirk E, Brown S, Mikelyte R. Playing its part: an evaluation of professional skill development through service user-led role-plays for social work students. Soc Work Educ 2020 May 21;40(8):977-993. [doi: 10.1080/02615479.2020.1764521]
56. Seland G. Empowering end users in design of mobile technology using role play as a method: reflections on the role-play conduction. In: Kurosu M, editor. Human Centered Design. Berlin, Germany: Springer; 2009.
57. Elhilu AH, El-Setouhy M, Mobarki AS, Abualgasem MM, Ahmed MA. Peer role-play simulation: a valuable alternative to bedside teaching during the COVID-19 pandemic. Adv Med Educ Pract 2023 Mar;Volume 14:257-264. [doi: 10.2147/amep.s399531]
58. Xu L, Liu W, Jiang X, Li Y. Impact of using peer role-playing on the clinical skills performance of pediatric trainees. BMC Med Educ 2023 Sep 08;23(1):654 [FREE Full text] [doi: 10.1186/s12909-023-04554-0] [Medline: 37684627]
59. Reiff M, Spiegel A, Williams E, Ramesh B, Madabhushi S, Bvunzawabaya B. From role-play to real life: using gatekeeper skills in real-world situations. J Coll Stud Ment Health 2022 Nov 22;38(1):119-148. [doi: 10.1080/87568225.2022.2144789]
60. Moreb M, Mohammed TA, Bayat O. A novel software engineering approach toward using machine learning for improving the efficiency of health systems. IEEE Access 2020;8:23169-23178. [doi: 10.1109/access.2020.2970178]
61. Ibeneme S, Karamagi H, Muneene D, Goswami K, Chisaka N, Okeibunor J. Strengthening health systems using innovative digital health technologies in Africa. Front Digit Health 2022 Mar 31;4:854339 [FREE Full text] [doi: 10.3389/fdgth.2022.854339] [Medline: 35434700]

XSL•FO
RenderX

62. Mhlanga D. Human-centered artificial intelligence: the superlative approach to achieve sustainable development goals in the fourth industrial revolution. Sustainability 2022 Jun 27;14(13):7804. [doi: 10.3390/su14137804]
63. Grindell C, Coates E, Croot L, O'Cathain A. The use of co-production, co-design and co-creation to mobilise knowledge in the management of health conditions: a systematic review. BMC Health Serv Res 2022 Jul 07;22(1):877 [FREE Full text] [doi: 10.1186/s12913-022-08079-y] [Medline: 35799251]
64. Nusir M, Rekik M. Systematic review of co-design in digital health for COVID-19 research. Univers Access Inf Soc 2022 Dec 31:1-15 [FREE Full text] [doi: 10.1007/s10209-022-00964-x] [Medline: 36618758]
65. Slattery P, Saeri AK, Bragge P. Research co-design in health: a rapid overview of reviews. Health Res Policy Syst 2020 Feb 11;18(1):17 [FREE Full text] [doi: 10.1186/s12961-020-0528-9] [Medline: 32046728]
66. Tremblay M, Hamel C, Viau-Guay A, Giroux D. User experience of the co-design research approach in eHealth: activity analysis with the course-of-action framework. JMIR Hum Factors 2022 Aug 09;9(3):e35577 [FREE Full text] [doi: 10.2196/35577] [Medline: 35943783]
67. Madanian S, Nakarada-Kordic I, Reay S, Chetty T. Patients' perspectives on digital health tools. PEC Innov 2023 Dec;2:100171 [FREE Full text] [doi: 10.1016/j.pecinn.2023.100171] [Medline: 37384154]
68. Kortum F, Klünder J, Brunotte W, Schneider K. Sprint performance forecasts in agile software development - the effect of futurespectives on team-driven dynamics. In: Proceedings of the 31st International Conference on Software Engineering and Knowledge Engineering. 2019 Presented at: SEKE 2019; July 10-12, 2019; Lisbon, Portugal. [doi: 10.18293/seke2019-224]
69. Graziani AR, Botindari L, Menegatti M, Moscatelli S. Adaptive coping strategies at the time of COVID-19: the role of social and general trust. Int J Environ Res Public Health 2023 Aug 03;20(15):6512 [FREE Full text] [doi: 10.3390/ijerph20156512] [Medline: 37569052]
70. Maison D, Jaworska D, Adamczyk D, Affeltowicz D. The challenges arising from the COVID-19 pandemic and the way people deal with them. A qualitative longitudinal study. PLoS One 2021 Oct 11;16(10):e0258133 [FREE Full text] [doi: 10.1371/journal.pone.0258133] [Medline: 34634054]
71. Fegert JM, Vitiello B, Plener PL, Clemens V. Challenges and burden of the Coronavirus 2019 (COVID-19) pandemic for child and adolescent mental health: a narrative review to highlight clinical and research needs in the acute phase and the long return to normality. Child Adolesc Psychiatry Ment Health 2020;14:20 [FREE Full text] [doi: 10.1186/s13034-020-00329-3] [Medline: 32419840]
72. Kowalski G, Ślebarska K. Remote working and work effectiveness: a leader perspective. Int J Environ Res Public Health 2022 Nov 20;19(22):15326 [FREE Full text] [doi: 10.3390/ijerph192215326] [Medline: 36430045]
73. Henke JB, Jones SK, O'Neill TA. Skills and abilities to thrive in remote work: what have we learned. Front Psychol 2022 Dec 19;13:893895 [FREE Full text] [doi: 10.3389/fpsyg.2022.893895] [Medline: 36600705]
74. Krajčík M, Schmidt DA, Baráth M. Hybrid work model: an approach to work–life flexibility in a changing environment. Adm Sci 2023 Jun 06;13(6):150. [doi: 10.3390/admsci13060150]
75. Sharma A, Ahmad Farouk I, Lal SK. COVID-19: a review on the novel coronavirus disease evolution, transmission, detection, control and prevention. Viruses 2021 Jan 29;13(2):202 [FREE Full text] [doi: 10.3390/v13020202] [Medline: 33572857]
76. Sun J, He WT, Wang L, Lai A, Ji X, Zhai X, et al. COVID-19: epidemiology, evolution, and cross-disciplinary perspectives. Trends Mol Med 2020 May;26(5):483-495 [FREE Full text] [doi: 10.1016/j.molmed.2020.02.008] [Medline: 32359479]
77. -. Effectiveness of public health measures in reducing the incidence of covid-19, SARS-CoV-2 transmission, and covid-19 mortality: systematic review and meta-analysis. BMJ 2021 Dec 03;375:n2997 [FREE Full text] [doi: 10.1136/bmj.n2997] [Medline: 34862160]
78. Cavalcante de Oliveira AP, Galante ML, Maia LS, Craveiro I, da Silva AP, Fronteira I, et al. Implementation of policy and management interventions to improve health and care workforce capacity to address the COVID-19 pandemic response: a systematic review. Hum Resour Health 2023 Oct 10;21(1):80 [FREE Full text] [doi: 10.1186/s12960-023-00856-y] [Medline: 37817165]
79. First lessons from government evaluations of COVID-19 responses: a synthesis. Organisation for Economic Co-operation and Development. 2022 Jan 21. URL: https://www.oecd.org/coronavirus/policy-responses/first-lessons-from-government-evaluations-of-covid-19-responses-a-synthesis-483507d6/ [accessed 2023-11-03]
80. Sigahi TF, Kawasaki BC, Bolis I, Morioka SN. A systematic review on the impacts of Covid-19 on work: contributions and a path forward from the perspectives of ergonomics and psychodynamics of work. Hum Factors Ergon Manuf 2021 Jul 12;31(4):375-388 [FREE Full text] [doi: 10.1002/hfm.20889] [Medline: 33821126]
81. Yanovskiy M, Socol Y. Are lockdowns effective in managing pandemics? Int J Environ Res Public Health 2022 Jul 29;19(15):9295 [FREE Full text] [doi: 10.3390/ijerph19159295] [Medline: 35954650]
82. Panchal U, Salazar de Pablo G, Franco M, Moreno C, Parellada M, Arango C, et al. The impact of COVID-19 lockdown on child and adolescent mental health: systematic review. Eur Child Adolesc Psychiatry 2023 Jul 18;32(7):1151-1177 [FREE Full text] [doi: 10.1007/s00787-021-01856-w] [Medline: 34406494]

XSL•FO
RenderX

83. Bardosh K, de Figueiredo A, Gur-Arie R, Jamrozik E, Doidge J, Lemmens T, et al. The unintended consequences of COVID-19 vaccine policy: why mandates, passports and restrictions may cause more harm than good. BMJ Glob Health 2022 May 26;7(5):e008684 [FREE Full text] [doi: 10.1136/bmjgh-2022-008684] [Medline: 35618306]
84. Maquiling A, Jeevakanthan A, Ho Mi Fane B. The effect of vaccine mandate announcements on vaccine uptake in Canada: an interrupted time series analysis. Vaccine 2023 May 02;41(18):2932-2940 [FREE Full text] [doi: 10.1016/j.vaccine.2023.03.040] [Medline: 37019696]
85. Zhang YS, Young Leslie H, Sharafaddin-Zadeh Y, Noels K, Lou NM. Public health messages about face masks early in the COVID-19 pandemic: perceptions of and impacts on Canadians. J Community Health 2021 Oct 20;46(5):903-912 [FREE Full text] [doi: 10.1007/s10900-021-00971-8] [Medline: 33611755]
86. He L, He C, Reynolds TL, Bai Q, Huang Y, Li C, et al. Why do people oppose mask wearing? A comprehensive analysis of U.S. tweets during the COVID-19 pandemic. J Am Med Inform Assoc 2021 Jul 14;28(7):1564-1573 [FREE Full text] [doi: 10.1093/jamia/ocab047] [Medline: 33690794]
87. Buchan MC, Bhawra J, Katapally TR. Navigating the digital world: development of an evidence-based digital literacy program and assessment tool for youth. Smart Learn Environ 2023;11:8. [doi: 10.31219/osf.io/aqvcu]

Abbreviations

AI: artificial intelligence

Edited by K Bond; submitted 24.01.24; peer-reviewed by J Knight, MR Tanniru; comments to author 07.07.24; revised version received 23.08.24; accepted 02.03.25; published 17.04.25.

Please cite as:
Buchan MC, Katapally TR, Bhawra J
Application of an Innovative Methodology to Build Infrastructure for Digital Transformation of Health Systems: Developmental Program Evaluation
JMIR Form Res 2025;9:e53339
URL: https://formative.jmir.org/2025/1/e53339
doi:10.2196/53339
PMID:

©M Claire Buchan, Tarun Reddy Katapally, Jasmin Bhawra. Originally published in JMIR Formative Research (https://formative.jmir.org), 17.04.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Ecological Momentary Assessment of Parental Well-Being and Time Use: Mixed Methods Compliance and Feasibility Study

Laura Altweck[1,2], PhD; Samuel Tomczyk[1,2], PhD

[1]Department Health and Prevention, Institute of Psychology, Universität Greifswald, Greifswald, Germany

[2]German Center for Child and Adolescent Health (DZKJ), partner site Greifswald/Rostock, Greifswald, Germany

Corresponding Author:
Laura Altweck, PhD
Department Health and Prevention
Institute of Psychology
Universität Greifswald
Robert-Blum-Str. 13
Greifswald, 17487
Germany
Phone: 49 3834 420 ext 3813
Email: laura.altweck@uni-greifswald.de

Abstract

Background: Parents often juggle multiple conflicting responsibilities, including work, childcare, and the household, making them a particularly burdened group. However, the impact of daily routines and associated (poor) well-being among parents has received relatively little attention. Ecological Momentary Assessment (EMA) is increasingly being used to capture real-time data and can help address this research gap.

Objective: This study aims to examine compliance rates and the feasibility of EMA for measuring daily well-being and time use among parents.

Methods: An exploratory mixed-methods study was conducted with 74 German parents (57/74, 77% women, (age: mean 37.6, SD 5.9 years). Participants completed a baseline questionnaire, followed by 4 daily EMA surveys (at 7:30 AM, 12 PM, 16:30 PM, and 21:30 PM) over a 1-week period, and a follow-up questionnaire. A subset of parents was also subsequently interviewed. Sociodemographic background and expected feasibility (open-ended questions) were surveyed at baseline, and feasibility was assessed at follow-up (closed- and open-ended questions) and in the interviews. State well-being (affective and cognitive), state stress, state as well as retrospective time-use were measured in the EMA surveys. Compliance and feasibility were examined using a combination of quantitative (descriptive analyses) and qualitative methodologies.

Results: Participants completed an average of 83% (SD 13%) of the daily surveys. Compliance varied by gender and age, where men (90% vs 80%) and older parents showed higher rates. Participants generally found the survey frequency and length manageable, though some suggested adjustments to the study period depending on their individual routines. The 7:30 AM survey was reported as the most challenging due to childcare drop-offs (40%-49%), followed by the 16:30 PM survey for similar reasons (7%-17%). The qualitative analysis further revealed additional points for improvement, for instance, the need for personalization (eg, individual adjustment of the survey timings and intervals), technical support, and the incorporation of gamification elements. Most interviewees (46% vs 23%) found the used measurement of well-being and stress to be appropriate. Regarding time use, they felt that the predefined activity groups (eg, personal care, working) were suitable (46%) but noted challenges assigning less frequent activities (eg, medical appointments) (5%-54%). Reporting the timings of time-use via consecutive questions (ie, specifying the duration or start and end times of an activity) was perceived as confusing (9%-69%), with participants expressing a preference for a visual overview, such as a Gantt chart.

Conclusions: The study demonstrates that, when accounting for certain sociodemographic and study design factors, EMA can be a feasible method for data collection regarding daily well-being and time use, even in highly time-constrained populations like parents. This shows great potential for future research, such as exploring work-family conflict or performative gender roles and complementing established methods (eg, retrospective daily diaries).

Trial Registration: OSF Registries osf.io/8qj3d; https://osf.io/8qj3d

International Registered Report Identifier (IRRID): RR2-10.2196/54728

XSL•FO
RenderX

(JMIR Form Res 2025;9:e67451) doi:10.2196/67451

KEYWORDS

ecological momentary assessment; parents; feasibility; compliance; time use; well-being; stress; EMA; mixed methods; daily life; parents; online questionnaires; questionnaires; surveys; quantitative; qualitative; sociodemographic; mobile phone

Introduction

Background

As parents often have to fulfill professional, domestic, and parental responsibilities, they can be considered a particularly burdened group [1]. These demands of balancing work, family, and personal life can lead to worse physical and mental health outcomes, underscoring the need for targeted support and interventions in daily life [2]. However, daily routines and associated (poor) well-being among parents have received relatively little attention, for instance, using real-time assessment methods, such as ecological momentary assessment (EMA) or experience sampling methods. Given that parents' well-being is not only important for their own health but also for their children's [3,4], further research into the feasibility and implementation of EMA in parents is important.

Parents as a Target Group

Parents today face significant challenges by balancing childcare responsibilities, such as managing school drop-offs and pick-ups, with their careers. Although time use between men and women has become more similar in recent decades [5,6], substantial differences remain, with men generally spending more time on paid work and less on unpaid work [7-9]. As men continue to prioritize work, many women face a double burden, struggling to balance family and professional demands [7,9,10]. This discrepancy extends to care work and childcare. Women often reduce their paid work hours after giving birth, whereas men do not but may instead increase theirs [8]. Women's share of housework is typically double that of men's, with the largest gender differences observed among those with preschool-aged children [11,12]. Mothers experience more variation in how they use their time, often engaging in part-time work and childcare simultaneously, and activities frequently overlapping with others (eg, doing housework while taking care of the children) [13]. They generally also have less free time than fathers and tend to combine their leisure time with other activities, which can exacerbate stress and reduce opportunities for recovery [13].

Understanding how individuals behave in their everyday lives and how their actions affect mental and physical health is essential [2]. According to the Conservation of Resources Theory [14], individuals manage their time to preserve or enhance their resources (eg, energy, money, or social support), while the Work-Family Conflict Model [15] suggests that time is divided between work and family based on competing demands and conflicts and this conflict may lead to stress and poor health outcomes. The gender perspective complements these theories, examining how gender roles shape time use, with theories such as Doing Gender [16] and Gender Performativity [17] emphasizing the construction of gender roles through actions and behaviors. Previous research has established gender differences in daily activities and their impact on general health and well-being [7,9,10,18,19]. Certain activities—such as household cleaning and unpaid care work—are consistently related to worse health outcomes [20,21], and these are in turn more likely to be undertaken by women [9]. Therefore, examining the daily well-being and time use of parents, particularly mothers, is important for understanding and addressing gender disparities in health using a real-time lens.

Ecological Momentary Assessment

The widespread use of smartphones provides an opportunity to collect real-time data and access a wider population. EMA involves "repeated collection of real-time data on participants' momentary states in the natural environment" [19]. EMA can, for example, be implemented via smartphone apps, which can collect data both internet-based and offline and use various methods such as open-ended questions, text- or voice-based responses, Likert scales, multiple-choice options, number wheels, or sliders [22]. Data can be gathered at deliberately selected time points (ie, experience-based sampling, eg, after the morning coffee), randomly selected time points (ie, time-based sampling), or fixed intervals throughout the day (ie, interval-based sampling, eg, each morning at 9:00 AM).

While EMA has been widely used in clinical populations to monitor changes in state well-being during illness or substance withdrawal [23], its application among the population, for example, in parent samples, is limited. State well-being is relatively stable over time compared with mood [24], but there is a greater variation in positive and negative state well-being (eg, affect) than in cognitive state well-being (eg, life satisfaction) [25]. Overall, daily state well-being tends to decline throughout the day and toward the end of the week [26,27]. Understanding these patterns among parents and their association with daily routines and activities could inform interventions aimed at improving their well-being.

Although the research community has acknowledged the importance of daily activities for well-being and health, EMA is rarely implemented. So far, time use is generally measured through paper-pencil diary methods [28], studies that examine time-use patterns using EMA are rare [22,29]. EMA as opposed to paper-pencil diaries is better suited for capturing short and infrequent activities like errands and housework [29], which are precisely the areas where gender differences in time use are most pronounced, making EMA a valuable method for measuring these disparities. A comprehensive comparison of different methodological approaches is beyond the scope of this study, as we focus on the compliance and feasibility of the measurement of well-being and time use using EMA in a parent sample, without a control condition that uses daily diaries.

XSL•FO
RenderX

Compliance and Feasibility of Ecological Momentary Assessment

Compliance is a key factor in the success of EMA studies, especially given the unpredictability of daily life. High compliance rates are essential for the validity of the data collected. Electronic EMA surveys have been shown to have higher compliance rates than paper-based methods (96% vs 70%), due to advantages such as automatic prompts, immediate data transfer, secure servers, and greater accessibility and convenience for participants [30]. Electronic assessment ensures that the responses are provided at the intended times, reducing issues with delayed or retrospective entries that can occur with paper-based methods. Higher compliance with electronic assessments has been found in both adults and children [31]. However, compliance rates can vary. De Vries and Bartels [32] reported an average compliance rate of 72%, ranging from 43% to 95%, while Passini et al [33] reported an average compliance rate of 83%. Factors such as the frequency of prompts and asking repeating questions can affect compliance; studies with higher prompt frequencies tend to have lower compliance rates [30]. There were no significant differences in compliance when using personal smartphones versus research-provided devices, nor did incentives influence compliance in app-based EMA applications [32].

Compliance and Feasibility of Ecological Momentary Assessment in Parents

Parents, due to their high levels of stress and busy schedules, present unique challenges for EMA studies. Stress has been shown to impair episodic memory [34], suggesting that real-time data collection methods like EMA could be particularly advantageous for this group by reducing reliance on memory and recall.

Studies have shown that EMA is a feasible method for researching parenting behaviors. For instance, Passini et al [33] found that using EMA to record parenting behaviors 3 times a day was effective in capturing different dimensions of parenthood. In their study, 90% of mothers reported low difficulties using EMA, and 79% did not perceive it as burdensome. In addition, 60% would recommend EMA studies to other mothers. However, 74% indicated that the EMA study was at least partly challenging, suggesting a need to explore ways to reduce these challenges for parents.

Compliance rates in EMA studies involving parents vary. Kerr et al [35] reported that mothers responded to an average of 34 out of 50 prompts (ie, 68%) while El Dahr et al [36] reported a compliance rate of 82% in parents. Compliance was shown to be higher in the evenings [36,37].

Technical issues can also affect compliance; about 18% of EMA prompts are not answered due to problems such as data loss, battery difficulties, software bugs, or prompts occurring outside participants' wake times [37]. Socioeconomic factors play a role as well, for example, Dzubur et al [37] found that mothers with higher annual incomes responded to prompts more quickly than those with lower incomes. This suggests that financial stability may afford parents more flexibility or resources to engage with EMA studies.

These findings suggest that while EMA is a promising method for studying parental well-being and time use, there are challenges that need to be addressed to enhance compliance and feasibility. Understanding the specific needs and constraints of parents is essential for designing EMA studies that are both effective and participant-friendly.

Research Aims and Research Questions

The following research questions (RQ) were posed:

RQ1: What is the overall feasibility of parents participating in an EMA study measuring daily well-being and time use?

RQ2: What are the overall compliance rates for parents participating in an EMA study measuring daily well-being and time use?

RQ3: Which aspects of the study design (eg, time point) and sociodemographic factors are associated with noncompliance and feasibility?

Methods

Study Design

This work is part of a quasi-experimental pilot study using EMA [38]. The main study consisted of internet-based questionnaires (before and after a week of EMA), daily EMA surveys over the course of 7 days, and subsequent structured interviews for the intervention group. The control group only took part in the internet-based questionnaires and did not partake in the interviews or the EMA surveys. This study focuses on the intervention group; for the full study design see Altweck et al [38].

The minimum sample size (ie, 64 participants) for the intervention group was calculated using G*Power (version 3.1.9.6; Erdfelder, Faul, and Buchner; ANOVA, repeated measures, within a group, α=.05, power=0.8, correlation among repeated measures=0.5, and effect size [well-being]=0.25 [39]). For the interviews, to achieve data saturation [40,41], we aimed to recruit 6-17 participants.

The mixed methods approach facilitated the examination of compliance rates and their associations with various factors (eg, sociodemographic characteristics and study design features), while also providing insights into the feasibility of the study design through short, open-ended questionnaire responses and in-depth interview answers. First, participants in the intervention group filled in an internet-based questionnaire. On the next day, they started filling in the EMA surveys (each approximately 5-7 min), which were completed on participants' personal Android smartphones using the app movisensXS (2024; movisens GmbH). The surveys took place 4 times per day (7:30 AM, 12 PM, 4:30 PM, 9 PM) with a buffer of 30 minutes over a period of 7 days. These time points were selected to also test feasibility in high-stress situations, for example, at 7:30 AM when parents get their children ready for nursery or school before work. The first 2 days were classed as a familiarization period and were not included in the analyses. After one week, participants filled in another internet-based questionnaire. The first participants to complete the daily EMA surveys were invited to take part in a structured interview to comprehensively

XSL•FO
RenderX

understand the feasibility of the study design and to be able to make adjustments if necessary, which was not the case.

Sample and Recruitment

Parents with at least one child under 18 years and living in the household were included. Only parents over the age of 18 years, without foster children, and without children with a chronic illness or disability were included, as these factors have been shown to further increase the burden of parenting [42]. Participants also needed access to an Android smartphone, in order to complete the daily surveys, for technical reasons.

Participants were recruited primarily from the northeast region of Germany using convenience sampling by distributing leaflets to relevant groups (eg, parent groups, nurseries, and schools) and digital channels (eg, the university's website and social media). In an initial phone conversation, participants were informed about the study goals and methods, informed consent was collected, and, finally, the inclusion and exclusion criteria were checked. The first participants who completed the follow-up questionnaire were invited to participate in an interview.

Procedure and Instruments

Baseline Questionnaires

Participants first filled in an internet-based questionnaire, which took approximately 20-30 minutes to complete.

The following sociodemographic variables were assessed: gender (woman, man), age (in years), education (International Standard Classification of Education-97 [43]; low and medium: level 0-4 [reference], high: level 5-6), employment status (full-time work, part-time work, not working), partnership status (in partnership, single, other), number of children under 14 years (numeric), number of children between 14 and 18 years (numeric), and family time (average h/d spent on family commitments [eg, childcare, family time, and housework] on both a regular workday and a free day). In addition, the expected feasibility of the study design was measured using the 4 open-ended questions (see Multimedia Appendix 1 for full item wording). For instance, participants were asked about the integration of EMA into their daily life and the survey's length and time points.

Daily Ecological Momentary Assessment Surveys

On the next day, participants began with the daily EMA surveys. Participants started the EMA surveys either on Tuesday or Thursday. The daily EMA surveys included assessments of cognitive (life satisfaction item from the WHO-QoL [World Health Organization Quality of Life]-8) [44], positive (happy, relaxed), and negative (nervous, downcast) affective well-being [45], and stress (a global measure of perceived stress) [46]. Also, current activity and the sequence of activities since the last survey time point were queried (activity categories according to Tomczyk et al [9]; eg, working, housework, childcare). In the final surveys of the day, received and desired social support [3] as well as state mindfulness (Multidimensional State Mindfulness Questionnaire) [47] was assessed. Participants answered between 12 and 27 items per survey point, depending on (1) the variation in time use between time points (ie, the number of different activities reported) and (2) the survey time point, with more items presented in the evening. For details, see Altweck et al [38] and for the full EMA survey, see Multimedia Appendix 2.

Follow-Up Questionnaires

A total of 8 days after the baseline questionnaire, participants filled out another internet-based questionnaire.

The feasibility of the study design was surveyed with a mix of closed- and open-ended questions. The questions were inspired and adapted from the existing literature that examined the feasibility of EMA methods [22,48-50] (see Multimedia Appendix 1 for full item wording and scales). An example item included "In my opinion...I was able to accommodate the survey time point at 7:30am", which was rated on a scale from 1 (strongly disagree) to 5 (strongly agree). Participants were able to explain their choices in an open textbox.

Interviews

Finally, a subgroup of the intervention group was invited to participate in an internet-based, structured interview that lasted 20-30 minutes. The interview guide was developed based on previous qualitative research on usability and feasibility evaluations [51,52] and was tailored to the current procedure, including focusing on the specific survey time points. The interviews focused on the feasibility of the EMA survey method and study design in greater detail. The interview followed the following structure: (1) general assessment of the study design, (2) time expenditure and integration into everyday life, (3) contents of the EMA surveys, (4) effects of the study, and (5) general and conclusion; see Multimedia Appendix 3 for the full interview guide.

Ethical Considerations

The pilot study [38] was approved by the Ethics Committee of the University Medicine Greifswald in July 2022 (BB 113/23) and was conducted following the ethical standards set by the 1964 Declaration of Helsinki, its later amendments, or similar standards. Informed verbal and written consent was obtained from each participant, informing them about how their data would be used and stored, and that their participation was pseudonymized. They were also informed that they could withdraw from the study at any time without repercussions. Ethical guidelines and general data protection regulations were adhered to concerning data usage and storage. Participants in the intervention group were offered a monetary incentive of €35 (US $36.63) for the completion of the baseline and follow-up questionnaires as well as at least 50% of the daily surveys, an additional €15 (US $15.70) if they completed at least 80% of the daily surveys, and an additional €10 (US $10.47) if they took part in the interview.

Analytical Approach

Analysis Software

The statistical analyses were conducted using jamovi (version 3.5) [53], while the qualitative analyses were conducted using MAXQDA (version 24; VERBI Software GmbH) [54].

XSL•FO
RenderX

Data and Sample Description

For the sample description, we calculated mean values and SD for continuous variables and relative frequencies for categorical variables. Comparisons across gender were made using *t* tests, chi-square tests, or, where relevant, nonparametric equivalents (eg, Fisher exact test, Mann-Whitney *U* test, Welch *t* test).

Compliance

We analyzed compliance following guidelines by Liao et al [30] and Dao et al [55], reporting overall compliance (number of overall completed prompts), attrition (completed prompts by survey day and time point), and latency (time from prompt to survey completion and time to complete individual surveys). Associations of compliance with relevant sociodemographic factors were examined.

Feasibility

Closed questions regarding feasibility were analyzed using descriptive statistics (see the Data and Sample Description section). For the evaluation of the interviews and open-ended questions in the questionnaires, qualitative content analysis according to Kuckartz and Rädiker [56] was used. The qualitative data from the baseline questionnaires (ie, expected feasibility), follow-up questionnaires (feasibility), and interviews (feasibility) were coded and categorized separately.

The audio or video data from the interviews were transcribed for further analysis. The research team reflected on and noted their previous assumptions regarding the sample, research questions, and expected results. Summaries for each participant were compiled to gain an initial overview of the qualitative data. Text segments or sense units were coded and categorized either into existing categories or new ones. The first version of the main categories was created deductively based on the questionnaire and interview questions. Once coding began, an inductive approach was adopted, allowing the categories to be refined or extended with subcategories according to the data. In the first stage of the coding process, the qualitative data was once fully coded by a single person using the main deductive categories. Throughout this process, the research team noted overlapping categories, possible amendments (ie, new categories or redundant ones), and potential subcategories. Following Kuckartz and Rädiker [56], interrater agreement was achieved through discussions within the research team. In the second stage of the coding process, one researcher coded the data using the final categories twice, 2 weeks apart.

Results

Data and Sample Description

In total, 193 parents expressed interest in taking part in the study by leaving their personal information (Figure 1). We were able to reach 168 parents to complete the telephonic screening, of these, 88 met the inclusion criteria and were invited to the intervention group. In total, 74 parents completed the baseline questionnaires, 68 parents also completed the daily EMA surveys, 67 parents also completed the follow-up questionnaires, and 13 parents took part in the interviews. With 4 surveys per day over a 7-day period, we collected 1352 data points. To gain the most complete overview of the compliance and feasibility of this methodology in a parent sample, we included all possible cases in the analyses and excluded missing values case-wise (N=74). There was no missing data in the questionnaire and interview data, except average daily family time (2/74, 3%). The samples showed no significant differences in any sociodemographic variables or survey compliance ($P>.05$).

Figure 1. Sample overview including sample size and retention.

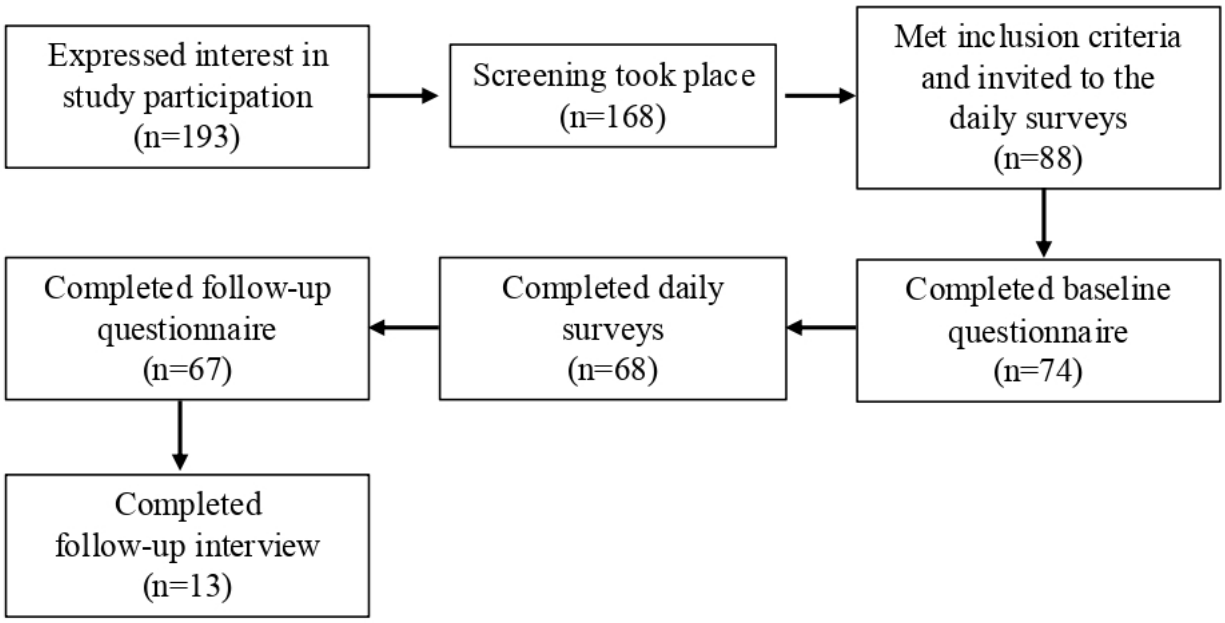

The final sample description is given in Table 1. The complete sample was made up of 77% (57/74) women and 23% (17/74) men. The average age was 37.6 (SD 5.9) years. The sample was highly educated (44/74, 60%), mostly in a relationship (53/74, 72%), and nearly equally divided across full-time, part-time, and no employment. On average, participants spent 9.6 (SD 4)

XSL•FO
RenderX

hours per day on family commitments. There were few significant differences across genders, men were merely significantly older than women. Also, there were no significant differences across subsamples (baseline only, baseline and follow-up, interview group; *P*>.05).

Table 1. Descriptive statistics of the sample at baseline (total and by gender; N=74).

Sample variable	Total sample	Women	Men
Age (years)			
Participants, n (%)	74 (100)	57 (77)	17 (23)
Mean (SD)	37.6 (5.9)	36.9 (5.1)	40.3 (7.6)[b]
Range	—[a]	28-53[b]	26-59[b]
Partnership status, n (%)			
In a relationship	53 (72)	39 (68)	14 (82)
Single	6 (8)	6 (11)	0 (0)
Other	15 (20)	12 (21)	3 (18)
Number of children			
Younger than 14 years			
Mean (SD)	1.7 (0.8)	1.7 (0.8)	1.8 (0.8)
Range	1-4	1-4	1-3
Between 14 and 18 years			
Mean (SD)	0.3 (0.5)	0.4 (0.6)	0.1 (0.4)
Range	0-2	0-2	0-1
Educational background, n (%)			
Low or medium	30 (41)	26 (46)	4 (24)
High	44 (60)	31 (54)	13 (77)
Employment status, n (%)			
Full-time	28 (38)	19 (34)	9 (53)
Part-time	29 (40)	24 (43)	5 (29)
None	16 (22)	13 (23)	3 (18)
Family time (h/d)			
Mean (SD)	9.6 (4)	10 (4.7)	8.2 (2.3)
Range	2-24	2-24	3-12.5

[a]Not applicable.
[b]*P*<.05.

Compliance

From the 28 maximum possible daily surveys to complete (4 times/d for 7 days), the average individual number of completed surveys was 23.1 (SD 3.57) (23.1/28, 82.5%; range 9-28) for results on compliance, refer to Table 2 and for grouped by time point or survey day, refer to Multimedia Appendix 4). Men (25/28, 90%; mean 25.1, SD 1.9) completed significantly more daily surveys than women (22.4/28, 80%; mean 2.4, SD 3.75) and the latter also showed greater variance ($t_{54.9}$=3.87, *P*<.001). Neither the survey time point nor the survey day significantly affected compliance (*P*>.05).

XSL•FO
RenderX

Table 2. Descriptive statistics of compliance at follow-up (total and by gender; N=68).

Variable	Total sample		Women		Men	
	Mean (SD)	Range	Mean (SD)	Range	Mean (SD)	Range
Average completed surveys (app data)	23.1 (3.6)	9-28	22.4[a] (3.8)	—[b]	25.1[a] (1.9)	—
Time to respond to trigger (app data)						
Across time points	2.6 (17.4)	0-633	2.3 (8.9)	0-270	3.3 (30.9)	0-633
7:30 AM	2.6 (13.9)	0-270.2	3 (16.5)	0-270.2	1.6 (3.1)	0-22.2
12 PM	1.7[c] (3.1)	0-26.1	1.7 (3.1)	0-26.1	1.5 (3.1)	0-21.7
4:30 PM	2.7[c] (5.3)	0-34.7	3 (5.5)	0-28.5	2 (5)	0-34.7
9 PM	3.4 (31.6)	0-632.5	1.6 (3.5)	0-25.6	8 (61.3)	0-632.6
Time to complete the survey						
Reported in the interviews	4.1 (1.8)	1.8-5.8	3.5 (2)	1.8-5.8	4.7 (1.5)	3.3-5.6
App data						
Across time points	3 (6.6)	0.4-137	3.2 (6.8)	0.5-137	2.8 (6.7)	0.4-134
7:30 AM	2.3[d] (3.4)	0.4-60.2	2.4[e] (4)	0.7-60.2	2.2[e,f] (1.5)	0.4-11.3
12 PM	2.8[g] (7.7)	0.4-134.4	2.8 (5)	0.6-60.4	3[h] (13)	0.4-134.4
4:30 PM	2.4[i] (3)	0.5-42.7	2.4[j] (4.4)	0.5-42.7	2.2[j,k] (1.8)	0.5-9.3
9 PM	4.7[d,g,i] (9.6)	1.1-136.8	5.1[e,j] (11.4)	1.2-136.8	3.8[e,f,h,j,k] (2.7)	1.1-20.5

[a]Significant differences across gender (*P*<.05).
[b]Not applicable.
[c]Significant differences across time points (*P*<.05).
[d]Significant differences across time points (*P*<.05).
[e]Significant differences across time points (*P*<.05).
[f]Significant differences across time points (*P*<.05).
[g]Significant differences across time points (*P*<.05).
[h]Significant differences across time points (*P*<.05).
[i]Significant differences across time points (*P*<.05).
[j]Significant differences across time points (*P*<.05).
[k]Significant differences across time points (*P*<.05).

Age significantly predicted the average individual number of completed surveys, where older participants showed higher compliance (r=0.33, P=.006). Other sociodemographic factors were not significantly related to compliance (P>.05).

On average, participants took 2.6 (SD 17.4) minutes to respond to a survey after receiving a signal; there was no significant gender difference (P>.05). Both men and women took longer to respond to the survey at 4:30 PM compared with the 12 PM one.

In the qualitative interviews, participants reported taking 4.1 (SD 1.8) minutes to complete the survey per time point, yet the app data showed that, on average, it took participants 3 (SD 6.6) minutes (see Multimedia Appendix 5 for examples and descriptions); there were no gender differences (P>.05). Both men and women took significantly longer to fill out the evening surveys (mean 4.7, SD 9.6) than the other time points (mean 2.3-2.8, SD 3.4-7.7); this discrepancy was also reported in the qualitative interviews.

Feasibility

Technical Aspects

Interviewees found the app easy to use and user-friendly, which was confirmed in the follow-up questionnaires. Participants highlighted the reminder function of the surveys and appreciated the real-time support provided during the study by the research team. However, some technical issues were raised, including compatibility problems with the operating system, display issues, unreliable reminders, and the nonautomatic snooze feature. For more details regarding technical aspects see Multimedia Appendix 6.

Study Design

In terms of the overall feasibility of the study design, at baseline, 74% (53/74) did not report any expected concerns, and 13% (9/74) expressed excitement about taking part in the study questionnaires (see Multimedia Appendix 7 for qualitative results from the baseline questionnaires). However, most participants felt that the study design was generally "doable" (high feasibility: 20/74, 28%, moderate feasibility: 51/74, 71%)

XSL•FO
RenderX

but expected certain difficulties in integrating the study into their daily lives (eg, some parts of the study design like a certain time point) and reported concerns, mainly around time management. The study duration (ie, the length of the overall study period and the individual surveys) appeared acceptable.

Participants found answering the questions every day adequately easy (mean 3.37, SD 1.04; scale: 1=difficult to 5=easy; refer to Tables 3 and 4 for quantitative results from the follow-up questionnaires). Some participants in the follow-up questionnaires (3/57, 5%) and 92% (12/13) interviewees reported a habituation effect because after a while "you already knew what you had to answer" (see Multimedia Appendix 8 for qualitative results from the follow-up questionnaires). Most interviewees reported the length and frequency of the individual survey time points to be adequate (12/13, 92% and 10/13, 77%, respectively). In the follow-up questionnaires, participants felt that the length of the individual time points was adequate, leaning slightly toward too much (mean 2.63, SD 0.8; scale: 1=too much to 5=far too little) but did not feel that the number of prompts was "annoying" (mean 2.48, SD 1.20; scale: 1=strongly disagree to 5=strongly agree). Most participants also did not feel that the study period was too long (mean 1.71, SD 0.94; scale: 1=strongly disagree to 5=strongly agree) and expressed that the survey period was adequate. However, 54% (7/13) of interviewees expressed a desire for the EMA study period to be longer than one week. Some expressed a desire for a longer survey period between 10 days and 3 weeks in order to better reflect fluctuations in everyday life and well-being. In contrast, one interviewee preferred a shorter survey period of 3 or 4 days, stating that the participation was stressful. In the follow-up questionnaires, 19% (11/57) of participants also requested a longer survey period, with the caveat that the number of daily surveys is reduced.

Table 3. Descriptive results from the follow-up questionnaires regarding feasibility (N=67).

Compliance	Mean (SD)
I found answering the questions everyday... (1=difficult to 5=easy)	3.37 (1.04)
I found the length of the daily surveys to be... (1=too much to 5=far too little)	2.63 (0.8)
In my opinion... (1=strongly disagree to 5=strongly agree)	
...the duration of the test (7 d) was too long.	1.71 (0.94)
...the number of daily prompts that were sent (4 times/d) was annoying.	2.48 (1.2)
...I ignored or postponed the prompts.	1.75 (1.01)
...the questions were understandable.	3.33 (1.04)
...I was able to remember the activities since the last measurement time.	4.17 (0.94)

Table 4. Descriptive results from the follow-up questionnaires regarding the feasibility of the survey time points (N=67).

Feasibility	Total sample, mean (SD)	Men, mean (SD)	Women, mean (SD)
In my opinion, I was able to accommodate the survey time at...well. (1=strongly disagree to 5=strongly agree)			
...7:30 AM...	3.42[a,b] (1.33)	3.76 (1.39)	3.31 (1.29)
...12 PM...	4.06[a,c] (1.16)	4.18 (1.07)	4.02 (1.2)
...4:30 PM...	3.62[c] (1.24)	3.35 (1.32)	3.71 (1.21)
...9:30 PM...	3.88[b] (1.07)	4.24[d] (0.66)	3.76[d] (1.16)

[a]Significant differences across time points (P<.05).
[b]Significant differences across time points (P<.05).
[c]Significant differences across time points (P<.05).
[d]Significant differences across gender (P<.05).

Survey Time Points

In the baseline questionnaires, 8% (6/74) of participants anticipated that it would be difficult to integrate the survey time points in general into their daily life. About half of the interviewees (7/13, 54%) found it difficult to fill out the surveys at the predetermined time points. Most interviewees (9/13, 69%) found the 30-minute response time sufficient, while 15% (2/13) preferred a longer response window of one hour. In the follow-up questionnaires, participants generally reported that they did not ignore or postpone survey prompts (mean 1.75, SD 1.01; scale: 1=strongly disagree to 5=strongly agree).

In the baseline (29/74, 40%), follow-up questionnaires (28/57, 49%), and the interviews (6/13, 46%), the 7:30 AM time point was reported to be the single most difficult survey to complete "because most of the time it's really the children's drop-off time. So if you have school-age children (laughs) or have to take them to the nursery, it was always a bit difficult to fit that in somehow". Instead, they suggested moving the early morning survey either a bit earlier (eg, 6:30 or 7:00 AM) or later (eg, 8:00 or 9:00 AM). The survey time point at 7:30 AM was said

XSL•FO
RenderX

to be the hardest to accommodate, especially compared with 12 PM and 9:00 PM (*P*<.05).

At baseline, the 12 PM time point was expected to be feasible (15% vs moderate-low feasibility: 10%), but at follow-up, it showed mixed feasibility (15% vs moderate-low feasibility: 10%). It was feasible if it took place during the lunch break at work but bad if it coincided with work duties or the children's lunchtime nap. For this time point, participants also suggested earlier (eg, 11:30 AM) or later times (eg, 2:00 PM).

In the baseline questionnaires, participants anticipated the time point at 4:30 PM to be the second most difficult to complete (17% vs moderate-low feasibility: 7%), which was confirmed at follow-up (high feasibility: 2%, low: 7%). The survey time point at 4:30 PM was also significantly easier to accommodate compared with 12 PM (*P*<.05). This time point was reported to be difficult if it coincided with the end of work and picking up the children from nursery or school or if it was after work and so the parents were in the middle of childcare and running errands or taking children to afternoon classes. Participants generally suggested a time a bit later, such as 5:00 PM or 5:30 PM.

In the interviews, the 9:00 PM time point was reported to be the second most feasible, whereas, in the follow-up questionnaires, 9:00 PM was reported to have been difficult to complete (high feasibility: 4% vs low feasibility: 19%). For some, this time point was "too late." If the children were already asleep, this time point appeared to be the most feasible; however, if the parents were still getting the children ready for bed, as the children sleep late in general or just on that day, then this time point appears to be similarly difficult as the one at 7:30 AM. The only significant gender difference in the feasibility of time points was regarding 9:00 PM, which men found easier to accommodate than women (*P*<.05).

Participants mentioned a difference in feasibility between weekdays and weekends; for instance, the 7:30 AM time point was too early on weekends. They also expressed a desire for more flexibility in tailoring the survey points to individual routines.

Measurement of Well-Being and Time Use Using Ecological Momentary Assessment

Generally, participants reported that the questions in the daily surveys were adequately understandable (mean 3.33, SD 1.04; scale: 1=strongly disagree to 5=strongly agree).

Most interviewees (46% vs 23%) found the used measurement of well-being to be appropriate, describing the questions as "multifaceted," while some reported the opposite and for example, felt that the items were too "superficial." Instead, 23% (3/13) reported that the questions did not reflect fluctuations over the day or the week.

Most interviewees (10/13, 77%) agreed that the query about the current activity 4 times per day was not a full reflection of the full day and that the questions about the activities between assessments were necessary to complete the picture: "I think the combination really makes the most sense, because then you really get everything packed into it."

There was some uncertainty regarding the predefined time-use categories. About half of the interviewees expressed uncertainty about assigning rare activities, for example, a doctor's appointment, to a certain category, which was also mentioned in the follow-up questionnaires. Others were unsure how to classify leisure and physical activity, that is, whether to assign this to hygiene, eating, etc., or childcare, when, for example, watching TV while the children were around. Some participants also reported that they would have preferred logging the individual activities (eg, watching TV) rather than an activity group (eg, free or leisure time).

At baseline, 83% (60/74) expected to be able to remember the activities since the last survey time point, this was confirmed in the follow-up questionnaires (mean 4.17, SD 0.94; scale: 1=strongly disagree to 5=strongly agree) and interviews (12/13, 92%). Participants were asked to report what they were currently doing and then the sequence of what they had been doing since the last survey. Both in the interviews and the questionnaires, participants mentioned that this was confusing and they would have preferred reporting activities in chronological order. Also, some felt the sequential report of activities was redundant, for example, when they were at work at 12 PM and still at work at 4:30 PM. Participants were also able to click on multiple activities at the same time, while some found this useful, others found it confusing to report the sequence in this way. Participants also reported being unsure whether to report the time (eg, 7:05 AM) or the time passed in hours and minutes (eg, since 1 h 30 min); there was no clear preference. In both the follow-up questionnaire and interviews, participants suggested including a visual overview of the activity sequence, similar to a "Gantt chart."

Miscellaneous

All interviewees also deemed the monetary incentive as appropriate; however, some also noted that as parents are short on time, a slightly higher incentive might be useful (between €50 [US $52.33] and €100 [US $104.65]). Some interviewees also mentioned that because the incentive was proportional to completed daily surveys, they would have liked a personal overview of their compliance.

One interviewee also mentioned that an offer of psychological support would be useful for participants who notice constant poor well-being or low levels of social support while completing the study.

Overall, the results offer a few key recommendations for future EMA studies with similar aims (Textbox 1). Ideally, these aspects should be discussed with the target group via participatory co-design workshops when preparing the study to ensure high commitment and compliance from the beginning.

XSL•FO
RenderX

Textbox 1. Key recommendations for designing ecological momentary assessments of time use and well-being in parents based on participants' perspectives.

- Technical support: Incorporate reminder functions and automatic snooze functions to facilitate completion and provide swift and adequate technical support throughout the study.
- Assessment schedule: Choose intervals that are acceptable for participants (eg, 4 times/d for 7 days or longer time periods with reduced frequency, eg, 2-3 times/d for 10-14 days).
- Customization: If possible: allow flexibility (eg, participants set their own time points within prespecified time windows).
- Visualization: Use visual aids, such as Gantt charts to illustrate sequential time use patterns instead of long (dropdown) lists'. Discuss whether to include open-ended or self-labelled responses and allow the logging of multiple overlapping activities.
- Incentives: Prepare for sufficient incentives as an appreciation for participants (eg, financial incentives of about €50 [US $52.33] per person).
- Gamification: If feasible, include elements of gamification (eg, live updates on compliance rates, positive feedback after completing a survey).

Discussion

Compliance and Feasibility

This study examined compliance and feasibility of an EMA of parental time use and well-being. By using qualitative (interviews, open-ended questions) and quantitative (EMA use data, surveys) methods, the study found that compliance was high (with an average of 83% [23.1/28] completed prompts) and feasibility was good for most parents that participated in the study. Older and male participants showed higher compliance. Participants also discussed several challenges to implementation regarding technical issues (eg, system compatibility and missing notifications), integration into daily life (eg, time points and frequency of assessments), and personal preferences (eg, how to log and classify daily activities).

The findings largely correspond to previous research that found similar compliance rates [30,33,36] and also reported a substantial amount of nonresponders due to technical issues [37]. However, in contrast to previous studies, we observed differences between fathers and mothers, with higher compliance rates and acceptance of evening assessments in fathers. Such differences were not found in other reviews that analyzed gender differences in EMA studies [57]. It is possible that mothers experienced excessive burden compared with fathers due to them more often having to balance family and professional demands [7,9,10] and having more informal care duties than fathers, which would give them less time to complete the assessments. Furthermore, fathers primarily having work duties would explain their frustration with repeated time use assessments during working hours (ie, monotonous responses at 12 PM and 4:30 PM), and would give them more time to complete evening assessments compared with mothers. However, this is merely speculative, and it is of interest to further examine time management and work-family conflict [15] in association with compliance and feasibility in the future. If work-family conflict affects well-being and time management, this could also be visible in daily life. Our sample was also highly educated, and previous research [37] found that mothers with higher annual incomes responded to prompts more quickly. This could indicate a bias and more research is needed in more diverse samples along different socioeconomic gradients.

Another important issue raised by participants referred to the customization of the assessment (eg, how to log and classify daily activities) and possible unintended consequences (eg, stress due to continued exposure to low well-being or visualization of work-family conflict in the time-use assessment). Since customization is in line with increased app engagement [58], this could be a good way to maintain compliance over time. However, this also requires more resources for the development, implementation, and data analysis. Therefore, more research is needed to explore the trade-off between complexity, compliance, and feasibility of the EMA design.

In addition, so far unintended consequences of such assessments have rarely been addressed in this target group, although research suggests that reactivity due to EMA monitoring is low [59,60]. Nevertheless, it is recommended to provide psychosocial support or resources in case of continuous reports of low levels of well-being. Since this can also affect compliance, additional incentives or reminders might be useful if multiple consecutive surveys are skipped or not completed. Other recommendations include extended training of participants to ensure a common understanding of data collection procedures. Fluctuations in well-being and negative affect can also influence app engagement. For example, parents may be less able to comply with EMA protocols on days when stress levels are higher, when their mood is lower or their children's behavior is more difficult to manage. There is also a possibility of potential reporting bias due to measurement reactivity as well as the risk that parents might skip surveys.

This has implications for ecological momentary interventions and just-in-time adaptive interventions: For instance, participants experiencing high levels of stress might benefit from a brief stress management intervention (eg, breathing techniques and mindfulness minute) to support their momentary emotion regulation before completing a more complex survey. This could not only improve their well-being but also their compliance. On the other hand, participants reporting positive mood might benefit from further reinforcement (in the sense of an upward spiral of positive emotions, following the broaden-and-build theory [61]) to sustain their positive mood and also their compliance. Therefore, analyzing stress, well-being, and app compliance concurrently is useful to account for days with higher or lower levels of engagement and well-being. Future studies may also want to use different types of operationalization of well-being or stress [31] to identify key moments to increase engagement and ensure compliance. For instance, participants

XSL•FO
RenderX

with low compliance could receive compliance-facilitating interventions, such as higher incentives or gamification elements, which could then lead to self-efficacy and a positive mood. In summary, we argue that looking more closely at the association between compliance, well-being, and stress might be beneficial to study quality (ie, lower attrition, higher retention rates) as well as participant well-being (eg, via EMIs or just-in-time adaptive interventions).

Finally, we present key recommendations for EMA designs assessing time use and well-being in parents (Textbox 1). These include ensuring robust technical support, tailoring assessment schedules to participants' needs, and offering flexibility in survey timing. We also highlight the importance of effective visualizations, sufficient incentives, and incorporating gamification elements to enhance engagement and usability.

Strengths and Limitations

The study is based on a convenience sample of parents, and therefore, it was not representative of the population. Particularly the female-dominant sample limits the generalizability of findings to fathers. In particular, parents with increased vulnerability (eg, few socioeconomic resources or children with a medical condition) were not part of the study, which could have biased the results. Furthermore, the assessment was limited to a few key questions for each outcome, which limits the scope of the examined constructs (eg, well-being). Furthermore, we did not compare different variations of the study design (eg, more or less frequent assessments, active vs passive data collection) and cannot generalize our findings to other methods or assessment schedules. Compliance and feasibility may differ between interval-contingent and random-signal-contingent sampling regimes, as parents in the former can mentally prepare and plan accordingly. Using brief scales and single-item measures is common and recommended in EMA research to limit the potential burden on participants, avoid dropouts, and achieve valid results [32]. Since we assumed parents to be a high-burdened group, because of demands from both family and work environments, it was important to us to use a brief survey for each EMA prompt. We also used psychometrically validated and previously tested items whenever possible to increase the comparability and methodological quality of the assessment. To explore other potential weaknesses and strengths of the study design, we also asked open-ended questions and conducted qualitative interviews to explore participants' perspectives and gain a better understanding of the challenges they face in this type of study.

Conclusion

In summary, our pilot study demonstrates that EMA methodologies are valuable for investigating parental time use and well-being in everyday life. They present unique opportunities for future research, such as exploring work-family conflict or performative gender roles, complementing established methods like daily diaries. In addition, they enable advanced big data analyses (eg, sequential cross-lagged panel models, sequence-based grouping algorithms) and support data-driven interventions that address the complex relationship between voluntary and involuntary daily activities, gender, and health.

Acknowledgments

We would like to thank Fredericke Schmitz for a preliminary literature review, Fredericke Schmitz and Pia Dietrich for the transcription of the interview material, and Anna Stoinski, Jette Giese, Heidi Schurig, Finja Knibbe, and Annalena Krupskin for supporting the preliminary coding of the qualitative material. Funding was received from the University of Greifswald for, for example, incentives and software licenses (F1704100).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

LA and ST contributed to conceptualization. LA performed data curation and formal analysis. LA and ST were involved in funding acquisition. LA participated in writing – original draft. LA and ST were responsible for writing – reviewing and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Questions regarding feasibility in the baseline and follow-up questionnaires.
[DOCX File , 17 KB - formative_v9i1e67451_app1.docx]

Multimedia Appendix 2
Full ecological momentary assessment questionnaire.
[DOCX File , 25 KB - formative_v9i1e67451_app2.docx]

Multimedia Appendix 3

XSL•FO
RenderX

Full interview guide.
[DOCX File , 18 KB - formative_v9i1e67451_app3.docx]

Multimedia Appendix 4
Compliance by survey day and time point.
[DOCX File , 18 KB - formative_v9i1e67451_app4.docx]

Multimedia Appendix 5
Frequency and description of the themes from the interviews (N=13).
[DOCX File , 35 KB - formative_v9i1e67451_app5.docx]

Multimedia Appendix 6
Technical aspects.
[DOCX File , 23 KB - formative_v9i1e67451_app6.docx]

Multimedia Appendix 7
Frequency and description of the themes from the baseline questionnaires (N=74).
[DOCX File , 23 KB - formative_v9i1e67451_app7.docx]

Multimedia Appendix 8
Frequency and description of the themes from the follow-up questionnaires (N=57).
[DOCX File , 26 KB - formative_v9i1e67451_app8.docx]

References

1. Matias M, Fontaine AM. Coping with work and family: how do dual-earners interact? Scand J Psychol 2015;56(2):212-222. [doi: 10.1111/sjop.12195] [Medline: 25600426]
2. Bauman A, Bittman M, Gershuny J. A short history of time use research; implications for public health. BMC Public Health 2019;19(Suppl 2):607 [FREE Full text] [doi: 10.1186/s12889-019-6760-y] [Medline: 31159790]
3. Bischoff M, Schmidt S, Muehlan H, Ulbricht S, Heckmann M, Berg NVD, et al. Ecological momentary assessment of parent-child attachment via technological devices: a systematic methodological review. Infant Behav Dev 2023;73:101882. [doi: 10.1016/j.infbeh.2023.101882] [Medline: 37633249]
4. Keijsers L, Boele S, Bülow A. Measuring parent-adolescent interactions in natural habitats. The potential, status, and challenges of ecological momentary assessment. Curr Opin Psychol 2022;44:264-269 [FREE Full text] [doi: 10.1016/j.copsyc.2021.10.002] [Medline: 34788708]
5. J Gershuny. Gender symmetry, gender convergence and historical work-time invariance in 24 countries. Centre for Time Use Research. 2018. URL: https://www.timeuse.org/sites/default/files/2018-02/CTUR%20WP%202%202018_0.pdf [accessed 2025-02-26]
6. Sayer LC. Trends in women's and men's time use, 1965–2012: back to the future? In: Gender and Couple Relationships. Cham: Springer; 2016:43-77.
7. Altweck L, Hahm S, Schmidt S, Ulke C, Fleischer T, Helmert C, et al. Even now women focus on family, men on work: an analysis of employment, marital, and reproductive life-course typologies in relation to change in health-related quality of life. App Res Qual Life 2022 Dec 02;18(3):1205-1223. [doi: 10.1007/s11482-022-10130-3]
8. Rubiano-Matulevich E, Viollaz M. Gender Differences in Time Use: Allocating Time between the Market and the Household. Washington, DC: World Bank; 2019.
9. Tomczyk S, Altweck L, Schmidt S. How is the way we spend our time related to psychological wellbeing? A cross-sectional analysis of time-use patterns in the general population and their associations with wellbeing and life satisfaction. BMC Public Health 2021;21(1):1858 [FREE Full text] [doi: 10.1186/s12889-021-11712-w] [Medline: 34649542]
10. Sullivan O. Gender inequality in work-family balance. Nat Hum Behav 2019;3(3):201-203 [FREE Full text] [doi: 10.1038/s41562-019-0536-3] [Medline: 30953017]
11. Anxo D, Mencarini L, Pailhé A, Solaz A, Tanturri ML, Flood L. Gender differences in time use over the life course in France, Italy, Sweden, and the US. Fem Econ 2011 Jul;17(3):159-195. [doi: 10.1080/13545701.2011.582822]
12. Coltrane S. Research on household labor: modeling and measuring the social embeddedness of routine family work. J Marriage Fam 2004;62(4):1208-1233. [doi: 10.1111/j.1741-3737.2000.01208.x]
13. Offer S. Free time and emotional well-being. Gender & Society 2015 Aug 03;30(2):213-239. [doi: 10.1177/0891243215596422]
14. Hobfoll SE. Conservation of resources. A new attempt at conceptualizing stress. Am Psychol 1989;44(3):513-524. [doi: 10.1037//0003-066x.44.3.513] [Medline: 2648906]

XSL•FO
RenderX

15. Greenhaus JH, Beutell NJ. Sources of conflict between work and family roles. Acad Manage Rev 1985;10(1):76-88. [doi: 10.2307/258214]
16. West C, Zimmerman DH. Doing gender. Gend Soc 1987;1(2):125-151. [doi: 10.1177/089124328700100200] [Medline: 22489612]
17. Butler J. Gender Trouble: Feminism and The Subversion of Identity. New York, NY: Routledge; 1990.
18. Altweck L, Tomczyk S, Schmidt S. Under contract and in good health: a multigroup cross-lagged panel model of time use and health-related quality of life in working-age men and women. Health Qual Life Outcomes 2022;20(1):151 [FREE Full text] [doi: 10.1186/s12955-022-02052-3] [Medline: 36376955]
19. Gálvez-Muñoz L, Rodríguez-Modroño P, Domínguez-Serrano M. Work and time use by gender: a new clustering of European welfare systems. Feminist Economics 2011 Oct;17(4):125-157. [doi: 10.1080/13545701.2011.620975]
20. Biddle S, Whitehead S, O'Donovan T, Nevill M. Correlates of participation in physical activity for adolescent girls: a systematic review of recent literature. J Phys Act Health 2005;2(4):423-434. [doi: 10.1123/jpah.2.4.423]
21. Dumas O, Le Moual N. Damaging effects of household cleaning products on the lungs. Expert Rev Respir Med 2020;14(1):1-4 [FREE Full text] [doi: 10.1080/17476348.2020.1689123] [Medline: 31682770]
22. Runyan JD, Steenbergh TA, Bainbridge C, Daugherty DA, Oke L, Fry BN. A smartphone ecological momentary assessment/intervention "app" for collecting real-time data and promoting self-awareness. PLoS One 2013;8(8):e71325 [FREE Full text] [doi: 10.1371/journal.pone.0071325] [Medline: 23977016]
23. Burgess-Hull A, Epstein DH. Ambulatory assessment methods to examine momentary state-based predictors of opioid use behaviors. Curr Addict Rep 2021;8(1):122-135 [FREE Full text] [doi: 10.1007/s40429-020-00351-7] [Medline: 33425652]
24. Mengelkoch S, Moriarity DP, Novak AM, Snyder MP, Slavich GM, Lev-Ari S. Using ecological momentary assessments to study how daily fluctuations in psychological states impact stress, well-being, and health. J Clin Med 2023;13(1):24 [FREE Full text] [doi: 10.3390/jcm13010024] [Medline: 38202031]
25. Gadermann AM, Zumbo BD. Investigating the intra-individual variability and trajectories of subjective well-being. Soc Indic Res 2006;81:1-33. [doi: 10.1007/s11205-006-0015-x]
26. Akay A, Martinsson P. Sundays are blue: aren't they? The day-of-the-week effect on subjective well-being and socio-economic status. SSRN Journal 2009:1-47. [doi: 10.2139/ssrn.1506315]
27. Stieger S, Reips U. Well-being, smartphone sensors, and data from open-access databases: a mobile experience sampling study. Field Methods 2019;31(3):277-291. [doi: 10.1177/1525822x18824281]
28. Eurostat. Harmonised European Time Use Surveys — 2018 Guidelines. Luxembourg: Eurostat; 2009.
29. Sonnenberg B, Riediger M, Wrzus C, Wagner GG. Measuring time use in surveys – how valid are time use questions in surveys? Concordance of survey and experience sampling measures. SSRN Journal 2012;41(5):1037-1052. [doi: 10.2139/ssrn.1895307]
30. Liao Y, Skelton K, Dunton G, Bruening M. A systematic review of methods and procedures used in ecological momentary assessments of diet and physical activity research in youth: an adapted STROBE checklist for reporting EMA studies (CREMAS). J Med Internet Res 2016;18(6):e151 [FREE Full text] [doi: 10.2196/jmir.4954] [Medline: 27328833]
31. Heron KE, Everhart RS, McHale SM, Smyth JM. Using mobile-technology-based ecological momentary assessment (EMA) methods with youth: a systematic review and recommendations. J Pediatr Psychol 2017;42(10):1087-1107. [doi: 10.1093/jpepsy/jsx078] [Medline: 28475765]
32. de Vries L, Bartels M. Individual differences in wellbeing dynamics: a genetically-informed comparison of Ecological Momentary Assessment and longitudinal survey data. Pers Individ Differ 2025 Apr;236:112996. [doi: 10.1016/j.paid.2024.112996]
33. Passini CM, Pihet S, Favez N, Schoebi D. Assessment of parental discipline in daily life. J Fam Psychol 2013;27(2):324-329. [doi: 10.1037/a0031504] [Medline: 23458698]
34. Sherman B, Huang I, Wijaya E, Turk-Browne N, Goldfarb EV. Acute stress effects on statistical learning and episodic memory. J Cogn Neurosci 2024;36(8):1741-1759. [doi: 10.31234/osf.io/546nv]
35. Kerr ML, Buttitta KV, Smiley PA, Rasmussen HF, Borelli JL. Mothers' real-time emotion as a function of attachment and proximity to their children. J Fam Psychol 2019;33(5):575-585. [doi: 10.1037/fam0000515] [Medline: 30742463]
36. El Dahr Y, Perquier F, Moloney M, Woo G, Dobrin-De Grace R, Carvalho D, et al. Feasibility of using research electronic data capture (REDCap) to collect daily experiences of parent-child dyads: ecological momentary assessment study. JMIR Form Res 2023;7:e42916 [FREE Full text] [doi: 10.2196/42916] [Medline: 37943593]
37. Dzubur E, Huh J, Maher JP, Intille SS, Dunton GF. Response patterns and intra-dyadic factors related to compliance with ecological momentary assessment among mothers and children. Transl Behav Med 2018;8(2):233-242 [FREE Full text] [doi: 10.1093/tbm/ibx002] [Medline: 29381785]
38. Altweck L, Schmidt S, Tomczyk S. Daily time-use patterns and quality of life in parents: protocol for a pilot quasi-experimental, nonrandomized controlled trial using ecological momentary assessment. JMIR Res Protoc 2024;13:e54728 [FREE Full text] [doi: 10.2196/54728] [Medline: 38820576]
39. Boemo T, Nieto I, Vazquez C, Sanchez-Lopez A. Relations between emotion regulation strategies and affect in daily life: a systematic review and meta-analysis of studies using ecological momentary assessments. Neurosci Biobehav Rev 2022;139:104747 [FREE Full text] [doi: 10.1016/j.neubiorev.2022.104747] [Medline: 35716875]

XSL•FO
RenderX

40. Guest G, Bunce A, Johnson L. How many interviews are enough?: An experiment with data saturation and variability. Field Methods 2006;18(1):59-82. [doi: 10.1177/1525822X05279903]
41. Hennink M, Kaiser BN. Sample sizes for saturation in qualitative research: a systematic review of empirical tests. Soc Sci Med 2022;292:114523 [FREE Full text] [doi: 10.1016/j.socscimed.2021.114523] [Medline: 34785096]
42. Crespo C, Santos S, Canavarro MC, Kielpikowski M, Pryor J, Féres-Carneiro T. Family routines and rituals in the context of chronic conditions: a review. Int J Psychol 2013;48(5):729-746. [doi: 10.1080/00207594.2013.806811] [Medline: 23848452]
43. Classifying educational programmes: manual for ISCED-97 implementation in OECD countries. Organisation for Economic Co-operation and Development. 1999. URL: https://www.voced.edu.au/content/ngv:11701 [accessed 2025-02-26]
44. The Whoqol Group. Development of the world health organization WHOQOL-BREF quality of life assessment. The WHOQOL group. Psychol Med 1998;28(3):551-558. [doi: 10.1017/s0033291798006667] [Medline: 9626712]
45. Blanke ES, Riediger M, Brose A. Pathways to happiness are multidirectional: associations between state mindfulness and everyday affective experience. Emotion 2018;18(2):202-211. [doi: 10.1037/emo0000323] [Medline: 28481572]
46. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. J Health Soc Behav 1983;24(4):385-396. [doi: 10.2307/2136404]
47. Blanke ES, Brose A. Multidimensional state mindfulness questionnaire (MSMQ). In: Medvedev ON, Krägeloh CU, Siegert RJ, Singh NN, editors. Handbook of Assessment in Mindfulness Research. Cham: Springer; 2022:1-13.
48. Businelle MS, Ma P, Kendzor DE, Frank SG, Vidrine DJ, Wetter DW. An ecological momentary intervention for smoking cessation: evaluation of feasibility and effectiveness. J Med Internet Res 2016;18(12):e321 [FREE Full text] [doi: 10.2196/jmir.6058] [Medline: 27956375]
49. Messner EM, Terhorst Y, Barke A, Baumeister H, Stoyanov S, Hides L, et al. The German version of the Mobile App Rating Scale (MARS-G): development and validation study. JMIR mHealth uHealth 2020;8(3):e14479 [FREE Full text] [doi: 10.2196/14479] [Medline: 32217504]
50. Spook JE, Paulussen T, Kok G, Van Empelen P. Monitoring dietary intake and physical activity electronically: feasibility, usability, and ecological validity of a mobile-based ecological momentary assessment tool. J Med Internet Res 2013;15(9):e214 [FREE Full text] [doi: 10.2196/jmir.2617] [Medline: 24067298]
51. Alqahtani F, Winn A, Orji R. Co-designing a mobile app to improve mental health and well-being: focus group study. JMIR Form Res 2021;5(2):e18172 [FREE Full text] [doi: 10.2196/18172] [Medline: 33635281]
52. Olsen SH, Saperstein SL, Gold RS. Content and feature preferences for a physical activity app for adults with physical disabilities: focus group study. JMIR mHealth uHealth 2019;7(10):e15019 [FREE Full text] [doi: 10.2196/15019] [Medline: 31605518]
53. Jamovi. 2024. URL: https://www.jamovi.org [accessed 2025-02-26]
54. MAXQDA 2022. Berlin, Germany: VERBI Software; 2021. URL: https://www.maxqda.com/about [accessed 2025-02-26]
55. Dao KP, De Cocker K, Tong HL, Kocaballi AB, Chow C, Laranjo L. Smartphone-delivered ecological momentary interventions based on ecological momentary assessments to promote health behaviors: systematic review and adapted checklist for reporting ecological momentary assessment and intervention studies. JMIR mHealth uHealth 2021;9(11):e22890 [FREE Full text] [doi: 10.2196/22890] [Medline: 34806995]
56. Kuckartz U, Rädiker S. Qualitative Content Analysis. Methods, Practice. Computer Support [Book in German]. Weinheim and Basel: Beltz Juventa; 2024.
57. Wrzus C, Neubauer AB. Ecological momentary assessment: a meta-analysis on designs, samples, and compliance across research fields. Assessment 2023;30(3):825-846 [FREE Full text] [doi: 10.1177/10731911211067538] [Medline: 35016567]
58. Oakley-Girvan I, Yunis R, Longmire M, Ouillon JS. What works best to engage participants in mobile app interventions and e-health: a scoping review. Telemed J E Health 2022;28(6):768-780 [FREE Full text] [doi: 10.1089/tmj.2021.0176] [Medline: 34637651]
59. Hufford MR, Shields AL, Shiffman S, Paty J, Balabanis M. Reactivity to ecological momentary assessment: an example using undergraduate problem drinkers. Psychol Addict Behav 2002;16(3):205-211. [doi: 10.1037/0893-164X.16.3.205]
60. Toth R, Trifonova T. Somebody's watching me: smartphone use tracking and reactivity. Comput Hum Behav Rep 2021;4:100142. [doi: 10.1016/j.chbr.2021.100142]
61. Fredrickson BL. The broaden-and-build theory of positive emotions. Philos Trans R Soc Lond B Biol Sci 2004;359(1449):1367-1378 [FREE Full text] [doi: 10.1098/rstb.2004.1512] [Medline: 15347528]

Abbreviations

EMA: ecological momentary assessment
WHO-QoL: World Health Organization Quality of Life

XSL•FO
RenderX

Edited by A Mavragani; submitted 18.10.24; peer-reviewed by L de Vries, AKC Chau; comments to author 19.12.24; revised version received 17.01.25; accepted 24.01.25; published 23.04.25.

Please cite as:
Altweck L, Tomczyk S
Ecological Momentary Assessment of Parental Well-Being and Time Use: Mixed Methods Compliance and Feasibility Study
JMIR Form Res 2025;9:e67451
URL: https://formative.jmir.org/2025/1/e67451
doi:10.2196/67451
PMID:

©Laura Altweck, Samuel Tomczyk. Originally published in JMIR Formative Research (https://formative.jmir.org), 23.04.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Protocol

Incorporating Patient-Reported Outcome Measures and Patient-Reported Experience Measures in Addiction Treatment Services in Belgium: Naturalistic, Longitudinal, Multicenter Cohort Study

Amine Zerrouk[1], MSc; Charlotte Migchels[2], MD, MSc; Clara De Ruysscher[1], PhD; Kim Fernandez[3], MSc; Jerome Antoine[3], MSE; Florian De Meyer[1], MSc; Frieda Matthys[2], MD, PhD; Wim van den Brink[4], MD, PhD; Cleo Lina Crunelle[2], MD, PhD; Wouter Vanderplasschen[1], PhD

[1]Department of Special Needs Education, Ghent University (UGent), Ghent, Belgium

[2]Department of Psychiatry, Universitair Ziekenhuis Brussel (UZ Brussel), Vrije Universiteit Brussel (VUB), Brussels, Belgium

[3]Department of Epidemiology and Public Health, Sciensano, Brussels, Belgium

[4]Department of Psychiatry, Amsterdam University Medical Centers, University of Amsterdam, Amsterdam, The Netherlands

Corresponding Author:
Amine Zerrouk, MSc
Department of Special Needs Education
Ghent University (UGent)
Henri Dunantlaan 1
Ghent, 9000
Belgium
Phone: 32 93310310
Email: ElAmine.Zerrouk@Ugent.be

Abstract

Background: Traditionally, treatment outcomes of service users with a substance use disorder (SUD) are measured using objective and provider-reported indicators. In recent years, there has been a shift toward incorporating patient-reported outcome measures (PROMs) and patient-reported experience measures (PREMs) to capture service users' perspectives on treatment outcomes and experiences.

Objective: The OMER-BE (Outcome Measurement and Evaluation as a Routine Practice in Alcohol and Other Drug Services in Belgium) study evaluates the acceptability and feasibility of PROMs and PREMs in different SUD treatment services, using the recently developed International Consortium for Health Outcomes Measurement Standard Set for Addictions. This paper presents the design and baseline characteristics of the study, indicators of attrition at 45-day follow-up, and the feasibility of the implementation of PROMs and PREMs in residential and outpatient services.

Methods: A convenience sample of 189 treatment-seeking individuals with SUD from different inpatient (therapeutic communities and psychiatric centers) and outpatient treatment services was followed for six months. Sociodemographic characteristics; clinical factors; and PROMs including recovery strengths, quality of life, and global health were assessed at baseline and within 3 weeks after starting treatment. Additionally, PROMs and PREMs were measured 45, 90, and 180 days later. Comparisons were made between treatment modalities, and indicators of attrition at the 45-day follow-up were assessed using ANOVA and chi-square tests.

Results: Baseline differences were observed between the three treatment modalities regarding education, SUD treatment history, primary substance, and Attention-Deficit/Hyperactivity Disorder Self-Report scores. Overall, patients in psychiatric treatment centers had a higher education level and less polysubstance use, while outpatients had fewer previous SUD treatments but received relatively more often opioid agonist treatment. Inpatients reported more attention-deficit/hyperactivity disorder symptoms and higher SUD severity than outpatients. Additionally, recovery strength scores were significantly lower in the outpatient group compared to the other groups, particularly in the subdomains of "Substance Use," "Self-care," and "Outlook on Life." At the 45-day follow-up assessment, the attrition rate was 36.6%. Comparisons between participants who completed the 45-day follow-up and those who dropped out revealed that completers were significantly older, had a higher level of education, were more likely to live alone, and were more likely to have a mother born in Belgium. They also had higher average scores on the "Material

XSL•FO
RenderX

Resources" domain of the Substance Use Recovery Evaluator, which includes questions about stable housing, a steady income, and effective financial management.

Conclusions: Evaluating PROMs and PREMs appears to be feasible in a diverse group of treatment-seeking patients with SUD in Belgium. However, challenges remain for structural implementation in practice, especially in outpatient services. Routine monitoring of PROMs and PREMs has the potential to empower patients, service providers, and policy makers by providing a comprehensive understanding of service users' needs and treatment effectiveness.

(JMIR Form Res 2025;9:e65686) doi:10.2196/65686

KEYWORDS

patient-reported outcome measures; patient-reported experience measures; substance use disorder; recovery; ICHOM; International Consortium for Health Outcomes Measurement; addiction; PROMs; PREMs; SUD; treatment; protocol; substance use; inpatient; services; perspectives; treatment outcome

Introduction

Alcohol and other substance use disorders (SUDs) are linked to a range of adverse psychological, physical, and social consequences. Their chronic, relapsing nature and related judicial, housing, and relational problems impact individuals with SUDs, as well as their environment and the broader community [1-4]. SUDs have a significant and growing impact on global morbidity and mortality [5-7]. Worldwide, harmful alcohol use causes 3 million deaths annually, representing 5.3% of all deaths. Alcohol use accounts for 5.1% of the global burden of disease [8]. Furthermore, an estimated 60,000 years of life lost were attributed to drug use in Europe in 2019 [5].

Treatment cohort studies conducted in the United States, Australia, and various European countries have shown the benefits of engaging in SUD treatment, generally resulting in increased abstinence rates, improved social integration, and reduced psychopathology [9-12]. Various SUD treatment modalities, however, seem to impact treatment outcomes in different ways. Stahler et al [13], for example, found that people with problematic opioid use benefitted more from residential treatment than individuals with alcohol as their primary substance. Conversely, people with problematic cannabis use were less likely to benefit from residential treatment compared to people with alcohol as their primary substance. Several studies found higher treatment completion rates in residential programs compared to outpatient settings [13,14]. Treatment engagement and retention have been consistently associated with positive outcomes, irrespective of the treatment modality [15-17]. Individuals who stay in treatment for a longer period of time are more likely to remain abstinent, experience fewer relapses and readmissions, engage less in criminal activity, and show greater improvements in general health measures [13,14,18]. Despite evidence for the effectiveness of SUD treatment, significant challenges such as high dropout and relapse rates persist and warrant further research and innovative approaches [16].

SUD treatment outcomes are typically measured using objective (eg, abstinence and rearrest or reincarceration) or provider-reported indicators (eg, treatment completion or compliance and absence of symptoms). Increasing emphasis on service users' perspectives in measuring treatment outcomes and experiences has recently led to the introduction of patient-reported outcome measures (PROMs) and patient-reported experience measures (PREMs) in SUD and other mental health services [19-22]. PROMs refer to individual, subjective treatment outcomes, including information about psychological well-being, quality of life, symptomatology, and physical functioning. PREMs provide information on how individuals experience health care and measure practical aspects of care, such as coordination, continuity and accessibility of care, and quality of patient-provider relationships [23]. International organizations, such as the International Consortium for Health Outcomes Measurement (ICHOM) [24], promote the implementation of patient-reported measures for routine assessment of health outcomes and experiences in all health care areas, including primary care, psychotherapy, and SUD services [21]. The use of PROMs and PREMs is advised to systematically monitor and improve the delivery of effective, patient-centered care and shared decision-making, which is not a standard practice in SUD and other mental health services [20,21,25-30]. Additionally, adopting standardized outcome measures opens opportunities to compare performances between treatment services within regions, countries, and even globally. This evolution will facilitate knowledge sharing among practitioners and provide policy makers with the tools and evidence needed to improve the quality and effectiveness of care [24]. The systematic use of PREMs is also likely to advance the field toward more personalized and effective support since it provides a direct evaluation of the accessibility, continuity, and coordination of care by service users [21,23]. Yet, although PROMs and PREMs show clear promise in improving the quality of SUD treatment, there is a lack of research on (the implementation of) these measures in clinical practice within SUD treatment settings [21]. The OMER-BE (Outcome Measurement and Evaluation as a Routine Practice in Alcohol and Other Drug Services in Belgium) study (2022-2025) aims to introduce and integrate the use of PROMs and PREMs in different SUD treatment modalities and to evaluate and enhance the quality of care provided in SUD treatment services in Belgium. The primary objective of the OMER-BE study is to assess the routine measurement of PROMs and PREMs in SUD treatment services, based on the recently developed and internationally validated ICHOM Standard Set for Addictions (ICHOM SSA) [24]. Therefore, we translated and validated the ICHOM SSA into Dutch and French, enhancing its applicability in Belgium and other Dutch- and French-speaking countries. By incorporating PROMs and PREMs in routine assessments, SUD treatment services are provided with information to better

XSL•FO
RenderX

align treatment approaches with service user needs. The secondary objective of the OMER-BE study is to measure and compare various recovery indicators and treatment experiences between different treatment modalities and assess their evolution during and after treatment. In this paper, we describe the design and baseline characteristics of the study. In order to enhance study adherence and routine implementation of PROMs and PREMs, we also examined indicators of attrition.

Methods

Study Setting

The OMER-BE study was a naturalistic, longitudinal, multicenter cohort study, in which 189 individuals with SUDs were followed up over a 6-month period in various residential and outpatient treatment modalities in the Dutch- and French-speaking regions of Belgium. Data collection for the study started in July 2022. Participants were recruited from 8 residential services (4 psychiatric treatment centers [PCs] and 4 therapeutic communities [TCs]) and 9 outpatient treatment services.

Specialized wards of PCs offer long-term (3 to 6 months) residential care that provides intensive medical and psychological support, addressing SUDs and in some cases co-occurring mental health disorders. Treatment consists of group counseling, psychoeducation, individual psychotherapy, and occupational activities.

TCs for addictions have a long history and were set up for individuals with SUDs, complementing traditional mental health care services that were traditionally not open to persons with drug problems [31]. In a TC, individuals with SUDs live together in a structured environment, typically for a period of 6 to 12 months, aiming for positive changes that lead to a drug-free life in society. The TC approach centers around the concept of “community as a method,” thus highlighting the influential role of peers and the power of mutual support in fostering recovery [32,33].

Outpatient treatment services provide more autonomy to service users and offer various nonresidential care options consisting of drug-free counseling interventions and harm-reduction approaches like opioid agonist therapy (OAT) and needle exchange programs. OAT refers to the use of opioid replacement medication such as methadone or buprenorphine to help manage withdrawal symptoms and reduce craving, often combined with some form of counseling and social support. Drug-free counseling focuses on psychosocial interventions, such as motivational interviewing and cognitive behavioral therapy, aimed at helping individuals develop coping strategies and support systems to become or remain abstinent.

Study Sample

The OMER-BE study aims to follow up a naturalistic cohort of service users with SUDs as they start a new treatment episode in a selected number of SUD treatment services, focusing on individuals with a primary alcohol or primary (illicit) drug problem. Eligibility criteria were (1) having a documented SUD (eg, a *Diagnostic and Statistical Manual of Mental Disorders* diagnosis of a SUD or previous treatment for a SUD), (2) being at least 18 years old, (3) being able to communicate in Dutch or French, and (4) having started treatment no longer than 21 days ago. First-time users, as well as returning service users, were considered eligible for this cohort study, as long as they started a new treatment episode during the recruitment period [34].

Upon treatment entry, service users were informed about the aims and design of the OMER-BE study through posters, leaflets, and staff members of the selected treatment facilities. During the initial meeting with the researcher, participants were informed extensively about the (follow-up) study and the implications of study participation and were asked for written informed consent to participate [34].

Study Procedure

Baseline Assessment

Sociodemographic and clinical factors (Textbox 1) were assessed at baseline, followed by the assessment of PROMs (Textbox 1). Data were collected through self-report using a tablet provided by one of the researchers, who were available throughout the assessment to address any questions participants may have. Administering the set of baseline variables and PROMs took between 20 and 45 minutes. After completion of the baseline assessment, participants received a voucher of €10 (US $10.46) as remuneration.

XSL•FO
RenderX

Textbox 1. Overview of included instruments in the OMER-BE (Outcome Measurement and Evaluation as a Routine Practice in Alcohol and Other Drug Services in Belgium) outcome measurement tool.

Sociodemographic factors

- Year of birth, sex, highest level of education completed, housing status, ethnicity

Clinical factors

- Treatment history for substance use disorders
- Primary Care Posttraumatic Stress Disorder Screen for Diagnostic and Statistical Manual of Mental Disorders (PC-PTSD-5)
- Depression, Anxiety, and Stress Scale (DASS-21)
- Adult Attention-Deficit/Hyperactivity Disorder Self-Report Scale (ASRS-v1.1)

Patient-reported outcome measures

- Patient-Reported Outcomes Measurement Information System (PROMIS) Short Form v1.0—Alcohol Use 7a (PROMIS-Alcohol)
- PROMIS SF v1.0—Severity of Substance Use 7a (PROMIS-Substance)
- Heaviness of Smoking Index (HSI)
- National Health Service Treatment Outcomes Profile for Substance Misuse—section 1 (TOP-S1)
- PROMIS Scale v1.2—Global Health (PROMIS-GH-10)
- Substance Use Recovery Evaluator (SURE)
- Brief version of the World Health Organization Quality of Life Scale (WHOQoL-BREF)

Patient-reported experience measures

- Patient-Reported Experience Measure for Addiction Treatment (PREMAT)

Follow-up Assessments

Study participants are contacted again 45, 90, and 180 days after the baseline assessment (Figure 1). We use a time window of four weeks for the 45- and 90-day follow-up and five weeks for the 180-day follow-up assessment to collect and complete the questionnaires. Researchers contacted all study participants either via email, SMS text messages, directly by phone, or through the treatment setting where they were enrolled in the study. Participants can complete the online survey via a personalized link provided by email or during a face-to-face or telephone interview, depending on their preference. The follow-up assessments include a measurement of both PROM and PREM variables (Textbox 1), which are supposed to take 20 to 30 minutes to complete. Participants receive a €10 (US $10.46) voucher for each completed follow-up assessment.

Figure 1. Overview of the data collection process. PREM: patient-reported experience measure; PROM: patient-reported outcome measure.

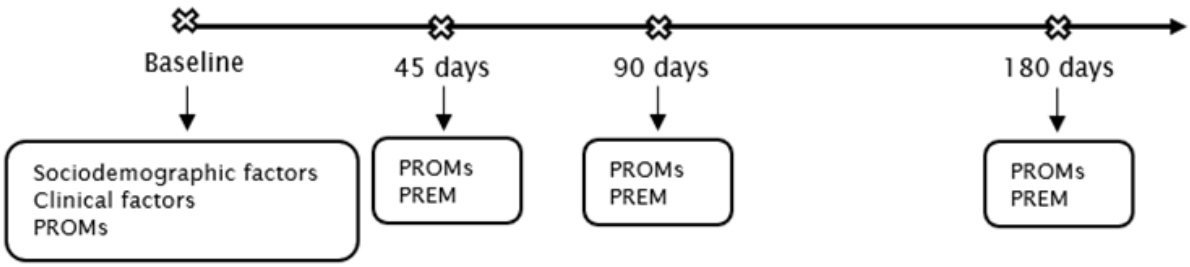

As the OMER-BE project is a naturalistic study, some participants will no longer be in treatment at the 45-, 90- or 180-day follow-up. Consequently, the PREM instrument is only administered when service users are still in treatment and at the first follow-up moment after leaving treatment. For example, if a participant stops treatment after 50 days, the PREM questionnaire will be administered at the 45- and 90-day follow-up moments.

XSL•FO
RenderX

Instruments

Overview

The baseline and follow-up assessments (Textbox 1) are largely based on the ICHOM SSA [24], a set of brief, existing, validated questionnaires to measure and monitor treatment outcomes routinely in SUD services that were developed by an international panel of SUD specialists. The ICHOM SSA focuses on patient-centered outcome indicators and provides an internationally agreed-upon method for measuring a variety of outcome domains. The tool offers potential for routine use since it is relatively short and has been specifically developed for and validated in the population of SUD service users. It can be easily administered and is applicable in a wide range of treatment settings [24]. To facilitate its application in Belgium, nontranslated questionnaires were translated into French and Dutch using forward or backward translation, following guidelines provided by Tsang et al [35], and subsequently validated. Compared to the ICHOM SSA procedure, we added a 45-day follow-up assessment, which allowed us to have an additional measurement point, keep participants more engaged in the study, and reduce attrition.

The OMER-BE measurement tool consists of three sections (Textbox 1): (1) sociodemographic and clinical factors, (2) PROMs, and (3) PREM.

Sociodemographic and Clinical Factors

Overview

The first section of the tool includes sociodemographic and clinical factors that may influence treatment outcomes.

The following sociodemographic variables were assessed: age, sex, education level, current living situation, country of birth of the participants, and country of birth of their parents.

Clinical factors include questions regarding SUD treatment history and three validated and widely used screening instruments to assess common comorbid psychiatric disorders (trauma, depression or anxiety, and attention-deficit/hyperactivity disorder [ADHD]) that are likely to affect treatment outcomes. These clinical factors were included to ensure a comprehensive understanding of how these variables influence recovery trajectories.

Primary Care Posttraumatic Stress Disorder Screen for Diagnostic and Statistical Manual of Mental Disorders

The Primary Care Posttraumatic Stress Disorder (PTSD) Screen [36] was developed to assess the occurrence of symptoms of PTSD over the last month. This 5-item screening tool is rated on a binary scale (No=0 and Yes=1). Higher total sum scores suggest the presence of more PTSD symptoms. The reliability of the English version has been found satisfactory in a sample of people with SUDs, as measured by a Cronbach α value of 0.73 [37].

Depression, Anxiety, Stress Scale

The Depression, Anxiety, and Stress Scale [38] is a commonly used instrument consisting of 21 items, divided into three subscales (depression, anxiety, and stress), each containing seven items. Each question is rated on a scale from 0=did not apply to me at all to 3=applied to me very much. Subscales and total scores are calculated by adding up the scores on the items and multiplying these by a factor of 2. In a population of Dutch-speaking SUD service users in the Netherlands, Beaufort et al [39] found the total score to be highly reliable (Cronbach α=0.91).

Adult ADHD Self-Report Scale

The Adult ADHD-Self-Report Scale v1.1 consists of several questions encompassing each of the 18 symptoms cited in the *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition* [40]. This study uses the short version, consisting of six items that have been found the most predictive of an ADHD diagnosis. These items are rated on a scale from 0=never to 4=very often. Scores equal to or higher than 2 for items 1-3 and equal to or higher than 3 on items 4-6 are indicative of a diagnosis of ADHD. A Spanish study by Blanco et al [41] showed a sensitivity of 0.88 and a specificity of 0.69 for a diagnosis of ADHD in a sample of outpatients with SUDs. In an international study among treatment-seeking in- and outpatients with an SUD, very similar results were obtained with a sensitivity of 0.83 and a specificity of 0.68 for *Diagnostic and Statistical Manual of Mental Disorders* ADHD [42].

PROMs

Overview

The second section of the tool is focused on PROMs and is based on the ICHOM SSA [24]. To extend the focus to subjective well-being beyond substance use and health outcomes, we added a widely used instrument to assess the quality of life and overall well-being (ie, brief version of the World Health Organization Quality of Life questionnaire [WHOQoL-BREF]).

Patient-Reported Outcomes Measurement Information System Alcohol Use Short Form 7a

The Patient-Reported Outcomes Measurement Information System (PROMIS) Alcohol Use Short Form [43] is a 7-item self-report questionnaire, derived from the 37-item PROMIS Alcohol Use item bank. This tool is designed to assess alcohol use in the last 30 days. Respondents were instructed to complete the seven questions only if they consumed alcohol in the last 30 days. Answers are rated on a 5-point scale ranging from 0=never to 4=almost always. The internal consistency of the total score has proven to be excellent in participants from the general population and a clinical sample of service users in treatment for SUDs (Cronbach α=0.95) [43].

PROMIS SF v1.0—Severity of Substance Use 7a

The PROMIS Severity of Substance Use Short Form [44] consists of seven items and is a shorter version of the 37-item PROMIS Severity of Substance Use item bank. This tool is designed to assess the severity of substance use in the last 30 days. Answers are rated on a 5-point scale ranging from 0=never to 4=almost always. The internal consistency of the total score was found to be excellent in participants from the general population and a clinical sample of service users in treatment for SUDs (Cronbach α=0.94) [44].

XSL•FO
RenderX

Heaviness of Smoking Index

The Heaviness of Smoking Index consists of two items: "How soon after waking up do you usually have your first smoke?" and a question assessing the number of cigarettes smoked each day. The first question was rated on a 4-point scale ranging from more than 60 minutes (0) to less than 5 minutes (3). The second question was replaced by an assessment of the number of cigarettes smoked in the last 30 days to maintain consistency with the assessment of the other substances. Internal consistency for the total Heaviness of Smoking Index score was found to be relatively low in a sample of male individuals with SUDs and nicotine dependence (Cronbach α=0.49) [45].

National Health Service Treatment Outcomes Profile for Substance Misuse—Section 1

The Treatment Outcomes Profile [46] is a multidimensional assessment instrument for monitoring outcomes in SUD treatment. This questionnaire measures four key life domains (substance use, crime, health, and social functioning). For this study, we only assessed the substance use domain and made modifications to the time frame in line with the other questionnaires. Specifically, participants were asked about their primary substances of use and the number of days and quantity consumed over the past 30 days.

PROMIS Scale v1.2—Global Health

The PROMIS Global Health consists of 10 items, scored on a 5-point scale. Due to overlap with other questionnaires, and in accordance with the ICHOM SSA, we only included two items of this scale relating to physical and mental health. The response options of these items are "excellent," "very good," "good," "fair," and "poor." Its psychometric properties were evaluated using a sample of 4370 individuals from the Dutch general population. Results indicated a 2-factor structure with good internal consistency. The subscales "Global Mental Health" and "Global Physical Health" had Cronbach α of 0.83 and 0.78, respectively [47].

Substance Use Recovery Evaluator

The Substance Use Recovery Evaluator (SURE) [48] is a self-report questionnaire consisting of 21 items, which is completed on a 5-point scale but scored on a 3-point scale (1-3). The first two response options correspond to a score of 3, the third response option to a score of 2, and the final 2 response options to a score of 1. Response options for the first three questions are "never," "on 1 or 2 days," "on 3 or 4 days," "on 5 or 6 days," and "every day," while for the remaining questions, response options are "all of the time," "most of the time," "a fair amount of the time," "a little of the time," and "none of the time." Higher total sum scores suggest greater recovery strengths. The items are categorized into five subscales: "substance use," "relationships," "self-care," "outlook on life," and "material resources." Internal consistency of the SURE total score was found to be high in a sample of current and former SUD service users (Cronbach α=0.92) [48]. Psychometric properties of the translated Dutch version of the SURE showed good internal consistency (Cronbach α=0.83) [49].

World Health Organization Quality of Life Scale

The WHOQoL-BREF [50] consists of 26 items and is a short version of the WHOQoL-100 questionnaire. Questions are rated on a 5-point scale (1-5) and response options, from the lowest to highest score, are "very poor/very dissatisfied/not at all/never," "poor/dissatisfied/a little/seldom," "neither poor nor good/neither satisfied nor dissatisfied/a moderate amount/moderately/quite often," "good/satisfied/very much/mostly/very often," and "very good/very satisfied/an extreme amount/extremely/completely/always." Higher sum scores are an indication of a better quality of life. The items are grouped into four domains: "psychological health," "physical health," "environment" and "social relationships." The WHOQoL-BREF has demonstrated good internal consistency with Cronbach α values for each of the domains ranging from 0.66 to 0.84 [51].

PREMs

Overview

As the objectives of the OMER-BE study also included the measurement of PREMs, the third section of the OMER-BE outcome measurement tool is not part of the ICHOM SSA. We use a newly validated PREM, the Patient-Reported Experience Measure for Addiction Treatment (PREMAT), to assess service users' experiences regarding the treatment they received.

PREMAT

The PREMAT is a recently developed relatively brief (23-item) questionnaire that aims to capture the experiences of people in residential SUD treatment services [52-54]. More precisely, the following topics are addressed in the PREMAT: "Individualized support," "Self-determination and Empowerment," "Program structure," "Treatment environment," "Coordination of care," and "Personal responsibility" [53,54]. The instrument consists of 23 statements rated on a 5-point Likert scale: 1=strongly disagree, 2=disagree, 3=neither agree nor disagree, 4=agree, and 5=strongly agree. The total score ranges from 23 to 115, with higher scores reflecting a more positive experience. Additionally, the PREMAT includes 2 open-ended questions ("How could your experience at this service have been improved?" and "What have been the best things about your experience here?"), which allows respondents to elaborate on certain aspects of the questionnaire or discuss topics that are not covered by the PREMAT items [52]. Internal consistency for the total score was found excellent in a sample of participants from specialist residential SUD treatment services in Australia (Cronbach α=0.91) [53].

Ethical Considerations

This study was granted ethical approval from the Medical Ethics Committee of the University Hospital Brussels on May 11, 2022 (UZ Brussel; BUN: 1432022000071). The participants were informed about the confidentiality of the data they provide, and written informed consent was acquired from all participants before being included in the study. Participants received a €10 (US $10.46) voucher for each completed assessment. All data are treated confidentially and reported anonymously. This study was conducted in accordance with the Declaration of Helsinki.

XSL•FO
RenderX

Data Analysis of Baseline Characteristics

This paper presents baseline characteristics of the study cohort and comparative analyses between participants from different treatment modalities, as well as between participants who completed the 45-day follow-up and those who did not participate in the 45-day assessment. For categorical demographic and patient characteristics, significant differences between treatment modalities were assessed using either the chi-square test or Fisher exact test when the data did not meet the assumptions required for the chi-square test. For continuous variables, we used ANOVA to assess group differences. Welch 2-tailed *t* test was used when the assumption of equal variances was not met. Post hoc comparisons between groups were made using the Tukey test when variances were equal or the Games-Howell test when the assumption of equal variances was violated. Initial power analyses indicated that a minimum of 159 participants were needed to detect moderate effects (with α=0.05 and 80% power). Statistical analyses were conducted using SPSS Statistics (version 29; IBM Corp).

Future analyses will center on longitudinal changes in PROMs and PREMs, using repeated measures and mixed model analyses. The role of mediating and moderating variables (eg, socioeconomic status, comorbidity, and recovery strength) will be assessed, along with differences between treatment modalities. All statistical analyses will be conducted using R studio (Posit PBC) and SPSS Statistics (IBM Corp).

Results

In total, 189 individuals participated in the OMER-BE study at baseline, of which 161 (85.2%) participants underwent residential treatment (81 participants treated in a SUD treatment ward in a psychiatric facility and 80 participants in a drug-free TC). Additionally, 28 (14.8%) participants were recruited in outpatient services. At the 45-day follow-up assessment, 120 participants completed the questionnaire, resulting in an overall attrition rate of 36.5% (69/189). The proportion of participants who did not complete the follow-up at 45 days was 38.8% (31/80) in drug-free TCs, 34.6% (28/81) in SUD treatment wards within psychiatric facilities, and 35.7% (10/28) in outpatient services.

Details regarding the baseline sociodemographic and clinical characteristics of the participant sample are presented in Table 1. The average age is 35.5 (SD 9.9) years. The majority of participants were male (n=156, 82.5%), completed secondary education as their highest level of education (n=114, 60.3%), and lived alone (n=90, 47.6%). Most participants were born in Belgium (n=178, 94.2%) and resided in psychiatric facilities (n=81, 42.9%) and TCs (n=80, 42.3%) during the baseline assessment. A total of 153 (81%) participants received previous treatment for SUDs. The most frequently reported main substances were alcohol (n=100, 53.8%), cocaine (n=81, 43.5%), and cannabis (n=64, 34.4%), indicating the presence of many problematic poly-substance users.

XSL•FO
RenderX

Table 1. Overview of the sociodemographic and clinical characteristics of the participant sample.

Variables	Total (n=189)	PC[a] (n=81)	TC[b] (n=80)	Outpatient (n=28)	*P* value
Age (years), mean (SD)[c]	35.5 (9.9)	36.6 (11.7)	34.1 (7.9)	36.5 (9.3)	.20[d]
Sex (male), n (%)	156 (82.5)	63 (77.8)	71 (88.8)	22 (78.6)	.16
Education level, n (%)					.04
Primary	45 (23.8)	14 (17.3)	23 (28.7)	8 (28.6)	
Secondary	114 (60.3)	47 (58)	51 (63.7)	16 (57.1)	
Higher	30 (15.9)	20 (24.7)	6 (7.5)	4 (14.3)	
Current living situation, n (%)					.37[e]
Alone	90 (47.6)	36 (44.4)	39 (48.8)	15 (53.6%)	
Alone with children	9 (4.8)	4 (4.9)	5 (6.3)	0 (0)	
Living together with partner and children	18 (9.5)	8 (9.9)	6 (7.5)	4 (14.3)	
Living together with partner without children	13 (6.9)	8 (9.9)	2 (2.5)	3 (10.7)	
Living together with others	59 (31.2)	25 (30.9)	28 (35)	6 (21.4)	
Country of birth, n (%)					
Belgium	178 (94.2)	74 (91.4)	76 (95)	28 (100)	.27[e]
Belgium (father)	156 (82.5)	64 (79)	65 (81.3)	27 (96.4)	.10
Belgium (mother)	161 (85.2)	68 (84)	66 (82.5)	27 (96.4)	.19
Previous treatment for SUDs[f] (yes), n (%)	153 (81)	67 (82.7)	68 (85)	18 (64.3)	.048
Opioid agonist therapy	37 (19.6)	9 (11.1)	15 (18.8)	13 (46.4)	<.001
Main substances (n=186)[g,h], n (%)					
Alcohol	100 (53.8)	52 (64.2)	38 (48.7)	10 (37)	.03
Amphetamines	41 (22)	9 (11.1)	23 (29.5)	9 (33.3)	.006
Benzodiazepines	20 (10.8)	8 (9.9)	12 (15.4)	0 (0)	.08
Cannabis	64 (34.4)	23 (28.4)	31 (39.7)	10 (37)	.31
Crack	33 (17.7)	9 (11.1)	19 (24.4)	5 (18.5)	.09
Codeine + Promethazine	2 (1.1)	2 (2.5)	0 (0)	0 (0)	.63[e]
Cocaine	81 (43.5)	21 (25.9)	50 (64.1)	10 (37)	<.001
GHB[i]	14 (7.5)	1 (1.2)	10 (12.8)	3 (11.1)	.02
Hallucinogens	3 (1.6)	0 (0)	2 (2.6)	1 (3.7)	.21[e]
Ketamine	20 (10.8)	13 (16)	5 (6.4)	2 (7.4)	.12
New psychoactive substances	6 (3.2)	2 (2.5)	3 (3.8)	1 (3.7)	.87[e]
Opioids	35 (18.8)	10 (12.3)	14 (17.9)	11 (40.7)	.005
Number of main substances (n=186)[c,g,h], mean (SD)	2.3 (1.4)	1.9 (1.3)	2.7 (1.4)	2.3 (1.4)	.001
More than one main substance (n=186)[g,h], n (%)	111 (59.7)	33 (40.7)	60 (76.9)	18 (66.7)	<.001

[a]PC: psychiatric treatment center.
[b]TC: therapeutic community.
[c]ANOVA.
[d]Welch test.
[e]Fisher exact test.
[f]SUD: substance use disorder.
[g]Some service users reported more than one main substance used.
[h]Data missing for three participants.

XSL•FO
RenderX

[i]GHB: gamma-hydroxybutyric acid.

Initial comparisons were made between the three treatment modalities (Table 1). When considering sociodemographic and clinical characteristics, no significant differences were found in terms of age, sex, living situation, and country of birth (all *P*>.05). However, significant differences were observed regarding education level, history of SUD treatment, OAT, and the primary substances reported. Post hoc analyses revealed that participants in the PC group had the highest level of education, followed by those in the outpatient group, and finally, the individuals from the TC group. On average, 83% (67/81) of participants from the PC group and 85% (68/80) of the participants from the TC group had a previous history of SUD treatment, with no statistically significant difference between the two groups. Moreover, a significantly higher percentage of participants in the outpatient group (13/28, 46%) were engaged in some form of OAT (*P*<.001). This was followed by participants in the TC group (15/80, 19%), and finally, the PC group (9/81, 11%). In terms of substance use, alcohol was more frequently reported as the primary substance in the PC group, followed by TC and outpatient groups. In contrast, opioids were most frequently reported in the outpatient group, followed by TC and PC groups. Amphetamine (*P*=.006), cocaine (*P*<.001), and gamma-hydroxybutyric acid (*P*=.02) were significantly more reported in the TC group, followed by the outpatient, and finally, the PC group. A significantly higher percentage of participants in the TC group reported more than one primary substance, followed by the outpatient and PC group (*P*<.001).

Further comparisons were made regarding comorbid psychiatric conditions and PROMs at baseline (Table 2). No significant differences were found in levels of PTSD, depression, anxiety, and stress scores between the treatment modalities (all *P*>.05). However, significant differences were observed in scores on the ADHD Self-Report scale, with participants in the outpatient group scoring significantly lower compared to the TC group (*P*=.03). No significant differences were found in general health and quality of life scores (all *P*>.05). In contrast, the total score on the SURE questionnaire revealed significant differences at baseline between the groups, with participants in the outpatient group scoring lower overall than those in the other groups (*P*<.001). This trend was significant in the subdomains of “Substance Use” (*P*<.001), “Self-care” (*P*<.001), and “Outlook on Life” (*P*=.008).

XSL•FO
RenderX

Table 2. Comparison of comorbidity and PROMs[a] at baseline.

Questionnaires	Total (n=189), mean (SD)	PC[b] (n=81), mean (SD)	TC[c] (n=80), mean (SD)	Outpatient (n=28), mean (SD)	*P* value
PC-PTSD-5[d]	2.24 (2.00)	2.09 (1.92)	2.51 (2.07)	1.89 (1.95)	.25
DASS 21[e]					
Depression	18.94 (10.92)	17.58 (10.80)	20.23 (10.80)	19.21 (11.54)	.31
Anxiety	14.03 (9.50)	12.94 (9.12)	15.53 (9.96)	12.93 (8.92)	.18
Stress	19.31 (10.15)	18.00 (10.47)	21.00 (9.72)	18.29 (10.09)	.15
ASRS[f]	3.44 (1.68)	3.36 (1.73)	3.75 (1.50)	2.79 (1.89)	.03
PROMIS GH[g]					
Physical	2.77 (.86)	2.73 (.92)	2.85 (.83)	2.77 (.86)	.55
Mental	2.59 (.92)	2.62 (.94)	2.61 (.88)	2.46 (.96)	.73
SURE[h] total	52.05 (7.70)	52.19 (7.99)	54.03 (5.46)	46.00 (9.31)	<.001[i]
Substance use	14.97 (2.89)	15.07 (2.98)	15.60 (2.42)	12.89 (3.01)	<.001
Self-care	12.02 (2.78)	11.81 (3.00)	12.91 (1.97)	10.07 (3.07)	<.001[i]
Relationships	10.90 (1.61)	10.98 (1.60)	11.13 (1.26)	10.04 (2.20)	.05[i]
Material resources	7.56 (1.76)	7.72 (1.73)	7.48 (1.80)	7.36 (1.77)	.55
Outlook on life	6.59 (1.88)	6.60 (1.92)	6.91 (1.66)	5.64 (2.08)	.008
WHOQoL-BREF[j]					
Perception quality of life	2.94 (.82)	2.93 (.77)	2.95 (.87)	2.96 (.84)	.97
Perception health	2.85 (.96)	2.91 (.91)	2.81 (.98)	2.79 (1.03)	.74
Physical health	13.52 (2.63)	13.33 (2.77)	13.85 (2.41)	13.14 (2.80)	.33
Psychological health	11.43 (2.81)	11.68 (2.67)	11.33 (2.88)	10.98 (3.01)	.48
Social relationships	12.14 (3.70)	12.59 (3.72)	11.65 (3.55)	12.24 (3.98)	.27
Environment	13.24 (2.87)	13.74 (2.84)	12.98 (2.85)	12.55 (2.84)	.09

[a]PROM: patient-reported outcome measure.

[b]PC: psychiatric treatment center.

[c]TC: therapeutic community.

[d]PC-PTSD-5: Primary Care Posttraumatic Stress Disorder Screen.

[e]DASS 21: Depression, Anxiety, and Stress Scale.

[f]ASRS: Adult Attention-Deficit/Hyperactivity Disorder-Self-Report Scale.

[g]PROMIS GH: Patient-Reported Outcomes Measurement Information System Global Health.

[h]SURE: Substance Use Recovery Evaluator.

[i]Welch test.

[j]WHOQoL-BREF: brief version of the World Health Organization Quality of Life questionnaire.

Additionally, comparisons were conducted of baseline sociodemographic, clinical, and PROM scores between participants who completed the 45-day follow-up and those who did not participate in the 45-day assessment (Table 3). Participants who completed the 45-day follow-up were significantly older (mean age 37.1, SD 9.4 years) compared to noncompleters (mean age 32.8, SD 10.3 years; *P*=.005). In general, noncompleters had a lower level of education (*P*=.009). Significant differences were also found regarding living situations, with a higher proportion of completers living alone compared to noncompleters (*P*=.03). The country of birth of participants' mothers differed significantly between groups, as a larger proportion of completers had a mother born in Belgium (109/120, 91%) compared to noncompleters (52/69, 75%; *P*=.004). A similar trend was observed for paternal country of birth, but the difference was not statistically significant (*P*=.44). Regarding substance use characteristics, alcohol was more frequently reported as the primary substance among completers (72/120, 61%) compared to noncompleters (28/69, 42%; *P*=.01). However, no significant differences were observed between groups concerning other primary substances of use or the number of primary substances reported (all *P*>.05).

XSL•FO
RenderX

Table 3. Comparison of sociodemographic and clinical characteristics of 45-day follow-up completers and noncompleters.

Variables	Total (n=189)	45-day completers (n=120)	Noncompleters (n=69)	*P* value
Age (years), mean (SD)[a]	35.5 (9.9)	37.1 (9.4)	32.8 (10.3)	.005
Sex (male), n (%)	156 (82.5)	100 (83.3)	56 (81.2)	.71
Education level, n (%)				.009
Primary	45 (23.8)	20 (16.7)	25 (36.2)	
Secondary	114 (60.3)	78 (65)	36 (52.2)	
Higher	30 (15.9)	22 (18.3)	8 (11.6)	
Current living situation, n (%)				.03[b]
Alone	90 (47.6)	67 (55.8)	23 (33.3)	
Alone with children	9 (4.8)	6 (5)	3 (4.3)	
Living together with partner and children	18 (9.5)	11 (9.2)	7 (10.1)	
Living together with partner without children	13 (6.9)	7 (5.8)	6 (8.7)	
Living together with others	59 (31.2)	29 (24.2)	30 (43.5)	
Country of birth, n (%)				
Belgium	178 (94.2)	113 (94.2)	65 (94.2)	.99[b]
Belgium (father)	156 (82.5)	101 (84.2)	55 (79.7)	.44
Belgium (mother)	161 (85.2)	109 (90.8)	52 (75.4)	.004
Previous treatment for SUDs[c] (yes), n (%)	153 (81)	98 (81.7)	55 (79.7)	.74
Opioid agonist therapy, n (%)	37 (19.6)	19 (15.8)	18 (26.1)	.09
Main substances (n=186)[d,e], n (%)				
Alcohol	100 (53.8)	72 (60.5)	28 (41.8)	.01
Amphetamines	41 (22)	27 (22.7)	14 (20.9)	.78
Benzodiazepines	20 (10.8)	15 (12.6)	5 (7.5)	.28
Cannabis	64 (34.4)	37 (31.1)	27 (40.3)	.21
Crack	33 (17.7)	22 (18.5)	11 (16.4)	.72
Codeine + Promethazine	2 (1.1)	1 (0.8)	1 (1.5)	.99[b]
Cocaine	81 (43.5)	52 (43.7)	29 (43.3)	.96
GHB[f]	14 (7.5)	12 (10.1)	2 (3)	.08
Hallucinogens	3 (1.6)	3 (2.5)	0 (0)	.55[b]
Ketamine	20 (10.8)	9 (7.6)	11 (16.4)	.06
New psychoactive substances	6 (3.2)	2 (1.7)	4 (6)	.19[b]
Opioids	35 (18.8)	19 (16)	16 (23.9)	.19
Number of main substances (n=186)[a,d,e], mean (SD)	2.3 (1.4)	2.3 (1.3)	2.2 (1.5)	.75
More than one main substance (n=186)[d,e], n (%)	111 (59.7)	75 (63)	36 (53.7)	.22
Treatment modality, n (%)				.86
TC[g]	80 (42.3)	49 (40.8)	31 (44.9)	
PC[h]	81 (42.9)	53 (44.2)	28 (40.6)	
Outpatient	28 (14.8)	18 (15)	10 (14.5)	

[a]ANOVA.

[b]Fisher exact test.

[c]SUD: substance use disorder.

XSL•FO
RenderX

[d]Some service users reported more than one main substance used.

[e]Data missing for three participants.

[f]GHB: gamma-hydroxybutyric acid.

[g]TC: therapeutic community.

[h]PC: psychiatric treatment center.

In terms of PROMs, a significant difference was observed in the SURE domain "material resources," where completers had significantly higher scores (mean 7.78, SD 1.62) than noncompleters (mean 7.19, SD 1.94; *P*=.04; Table 4). Additionally, noncompleters scored significantly higher (mean 2.99, SD 0.87) on the physical health domain of the PROMIS Global Health compared to people who completed the 45-day follow-up (mean 2.65, SD 0.84; *P*=.01), indicating better perceived physical health. No significant differences were found between completers and noncompleters across other PROM domains (all *P*>.05).

Table 4. Comparison of comorbidity and PROMs[a] at baseline between 45-day follow-up completers and noncompleters.

Questionnaires	Total (n=189), mean (SD)	45-day completers (n=120), mean (SD)	Noncompleters (n=69), mean (SD)	*P* value
PC-PTSD-5[b]	2.24 (2.00)	2.30 (2.00)	2.13 (2.00)	.58
DASS 21[c]				
Depression	18.94 (10.92)	18.43 (10.77)	19.83 (11.21)	.40
Anxiety	14.03 (9.50)	14.30 (9.66)	13.57 (9.25)	.61
Stress	19.31 (10.15)	18.70 (10.38)	20.38 (9.72)	.28
ASRS[d]	3.44 (1.68)	3.35 (1.74)	3.59 (1.58)	.34
PROMIS GH[e]				
Physical	2.77 (.86)	2.65 (.84)	2.99 (.87)	.01
Mental	2.59 (.92)	2.56 (.87)	2.65 (1.00)	.49
SURE[f] total	52.05 (7.70)	52.73 (7.33)	50.87 (8.22)	.11
Substance use	14.97 (2.89)	15.21 (2.75)	14.57 (3.10)	.14
Self-care	12.02 (2.78)	12.15 (2.71)	11.80 (2.90)	.40
Relationships	10.90 (1.61)	10.99 (1.52)	10.74 (1.75)	.29
Material resources	7.56 (1.76)	7.78 (1.62)	7.19 (1.94)	.04[g]
Outlook on life	6.59 (1.88)	6.60 (1.84)	6.58 (1.94)	.94
WHOQoL-BREF[h]				
Perception quality of life	2.94 (.82)	2.94 (.80)	2.94 (.86)	.99
Perception health	2.85 (.96)	2.83 (.96)	2.88 (.95)	.73
Physical health	13.52 (2.63)	13.51 (2.70)	13.55 (2.52)	.92
Psychological health	11.43 (2.81)	11.26 (2.68)	11.72 (3.01)	.28
Social relationships	12.14 (3.70)	11.83 (3.83)	12.68 (3.40)	.13
Environment	13.24 (2.87)	13.40 (2.87)	12.96 (2.86)	.30

[a]PROM: patient-reported outcome measure.

[b]PC-PTSD-5: Primary Care Posttraumatic Stress Disorder Screen.

[c]DASS 21: Depression, Anxiety, and Stress Scale.

[d]ASRS: Adult Attention-Deficit/Hyperactivity Disorder Self-Report Scale.

[e]PROMIS GH: Patient-Reported Outcomes Measurement Information System Global Health.

[f]SURE: Substance Use Recovery Evaluator.

[g]Welch test.

[h]WHOQoL-BREF: brief version of the World Health Organization Quality of Life questionnaire.

XSL•FO
RenderX

Discussion

Principal Findings

The use of PROMs and PREMs in SUD treatment services is limited to date, and no systematic monitoring system of patient-reported outcomes in Belgium is currently available [34]. Therefore, the OMER-BE study was set up to assess PROMs and PREMs systematically and to improve the quality of SUD services through the routine measurement and monitoring of patient-reported outcomes and experiences at regular times during and after treatment.

Comparisons across the three treatment modalities at baseline revealed several noteworthy differences and similarities. While no significant differences were observed in age, sex, living situation, or country of birth, significant variations in education levels, history of SUD treatment, engagement in OAT, and primary substance of problematic use were observed. Regarding co-occurring mental health disorders, no significant differences were found in PTSD, depression, anxiety, and stress scores between the treatment modalities. However, significant differences were noted in ADHD scores, with the TC group scoring significantly higher than the outpatient group. These differences can be attributed to several factors. TCs are often more suitable for individuals with more severe ADHD symptoms, as these environments provide the structured support and comprehensive care needed to manage both SUDs and comorbid conditions. These settings allow for intensive monitoring and support, which may be crucial for individuals struggling with ADHD symptoms, such as impulsivity [55].

Analysis of PROM scores showed no significant differences between treatment modalities in general health and quality of life. However, significant differences were found in recovery strengths, with participants in the outpatient group scoring lower overall than those in the other groups. These differences, particularly noted in the subdomains of “Substance Use,” “Self-care,” and “Outlook on Life,” suggest that the structure of the treatment environment plays a role in shaping recovery trajectories. For example, we observed that participants in residential treatment score significantly higher on the “Substance Use” subscale compared to those in outpatient treatment, suggesting lower levels of substance use among those in residential settings. This difference can be attributed to the structured and controlled environment of residential treatment settings, where strict measures are often in place to encourage abstinence and limit access to substances [56]. The supervision and supportive community in residential facilities may also play a crucial role in reducing substance use, which is less enforceable in outpatient settings where individuals have more autonomy and access to substances. The significant difference in the “Self-care” subscale, with residential participants scoring higher, may reflect the comprehensive care and support provided in residential settings. These environments typically offer structured daily routines and more intensive psychological and physical care [56]. In contrast, outpatient treatment often places a greater emphasis on self-management [57], which can be challenging for individuals with limited resources or support networks. Participants in residential treatment also scored significantly higher on the “Outlook on Life” subscale, which may be attributed to the supportive environment of residential treatment settings, where participants are provided with continuous care and peer support, all of which contribute to a more positive and optimistic outlook on life. Notably, the group difference in the “Relationship” subscale was close to significance. As participants were assessed in the first weeks of treatment, it is possible that participants in residential facilities did not have enough time to build meaningful relationships with other service users and providers.

In addition to differences between treatment modalities, comparisons between participants who completed the 45-day follow-up and those who did not participate in this assessment revealed several significant differences. Participants who completed the 45-day follow-up were significantly older, had a higher education level, were more likely to live alone, and were more likely to have a mother born in Belgium. Additionally, they scored higher on the “Material Resources” domain of the SURE, which includes questions about stable housing, a steady income, and effective financial management. These findings are in line with studies that suggest that factors such as lower education level, younger age, unemployment, and financial instability are associated with higher attrition in follow-up assessments [58,59]. Furthermore, some studies have found that factors such as low socioeconomic status are associated with limitations in people’s ability to use electronic devices [21,60,61], which may contribute to reduced participation in follow-up assessments. The higher completion rates among individuals living alone could indicate that they experience fewer external obligations that interfere with follow-up participation. The differences observed in maternal country of birth suggest that cultural or socioeconomic factors may also play a role in study retention, such as language barriers or differing levels of trust in research participation. Furthermore, alcohol was more frequently reported as the primary substance of use among completers and individuals with alcohol as primary substance of use had a lower attrition rate in comparison to the total sample (28% vs 36.6%). The higher prevalence of alcohol as a primary substance among completers may reflect differences in treatment pathways and more stable living situations that can be linked to socioeconomic factors and higher education levels.

Limitations

Despite being innovative and pioneering, the proposed study also has some limitations. First, recruiting participants in outpatient facilities proved to be extremely challenging, leading to a limited number of participants in these settings. One significant obstacle was the presence of long waiting lists, resulting in a limited number of new treatment episodes and lower turnover rates compared to the residential services. The frequency of contact between potential participants and service providers was limited in many outpatient settings, as many service users have only weekly or biweekly appointments. This poses significant scheduling constraints, especially since participants were expected to complete the baseline questionnaires within three weeks of treatment initiation. Additionally, there was a high occurrence of no-shows for scheduled appointments with the researchers, which resulted in

XSL•FO
RenderX

missed opportunities to engage potential participants. To address these challenges, several strategies were implemented. Flyers and posters were provided in waiting areas to inform potential participants about the study and proactive follow-up calls were made to confirm appointments. Moreover, regular contact was maintained with service providers (both through email and phone) to encourage participation. Despite these efforts, maintaining consistent engagement in outpatient facilities has been proven particularly challenging, resulting in some outpatient centers withdrawing from the study. Consequently, future comparative analyses will primarily focus on residential treatment services due to the limited number of outpatient participants. These challenges raise important considerations for the broader implementation of PROMs and PREMs in outpatient addiction services. While comprehensive assessments seem feasible in residential settings, these findings underscore the need for effective strategies for the structural implementation of PROMs and PREMs in outpatient services. Potential strategies to increase participation and retention may include broadening recruitment periods, simplifying assessment batteries, and using short-form PROMs and PREMs that could be completed during brief waiting periods before appointments. Additionally, codeveloping implementation strategies with service users and providers could further enhance feasibility and engagement in outpatient settings.

Second, recruiting facilities and participants in the French-speaking regions of Belgium proved to be more challenging than in the Dutch-speaking parts. This might be caused by differing attitudes toward the importance of outcome evaluation and measurement. Hence, beyond providing feasible measurement instruments and implementation strategies, there is a need to build service cultures that value the implementation of measurements to improve the quality of support. Organizing policies and funding that support the routine use of PROMs and PREMs might be an important step toward realizing this. Yet, this also includes overcoming practical challenges such as staffing shortages that hinder broader participation.

Finally, the study is limited by the lack of standardized diagnostic assessment of participants entering the study. Inclusion was based on proxy indicators (previous treatment for SUD, clinical assessment, or history of substance misuse), rather than on clinical diagnoses. This lack of structural diagnosis may have introduced variability in our findings and may limit generalization to other populations.

Conclusions

Despite these challenges, this study is one of the first studies to explore the assessment of PROMs and PREMs in SUD treatment services using the ICHOM SSA set of instruments. By systematically and routinely monitoring PROMs and PREMs, the project aims to empower service providers and give them tools to evaluate subjective treatment outcomes and experiences. This approach can provide service providers and policy makers with benchmarks for assessing outcomes and experiences during and after treatment across various treatment modalities. By applying an internationally validated tool, the study allows international comparison with similar interventions and treatment modalities globally. Future analyses will explore longitudinal changes in PROMs and PREMs, with a particular focus on the influence of baseline mediating and moderating variables (eg, socioeconomic status, comorbidity, and recovery strength) and differences between treatment modalities. The findings from this study will offer new perspectives and insights into the effectiveness of different SUD treatment modalities and their impact on diverse service user populations. This shift toward person-centered care not only supports better recovery outcomes but also fosters continuous improvement and innovation in SUD treatment services. Ultimately, systematically monitoring PROMs and PREMs has the potential to enhance the quality of care by offering a comprehensive understanding of service users' needs and treatment effectiveness, extending beyond abstinence.

Acknowledgments

The research was funded by a grant from the Belgian Science Policy Office (BELSPO), research project DR/90. BELSPO has no role in the design, data collection, data analysis, and reporting of this study.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

WB discloses relationships with Clearmind, Camurus AB, and Takeda Pharmaceutical Company Limited, involving consultancy or advisory roles.

References

1. Deak JD, Johnson EC. Genetics of substance use disorders: a review. Psychol Med 2021;51(13):2189-2200. [doi: 10.1017/s0033291721000969]
2. Dennis M, Scott CK. Managing addiction as a chronic condition. Addict Sci Clin Pract 2007;4(1):45-55 [FREE Full text] [doi: 10.1151/ascp074145] [Medline: 18292710]

XSL•FO
RenderX

3. Kelly JF, Greene MC, Bergman BG, White WL, Hoeppner BB. How many recovery attempts does it take to successfully resolve an alcohol or drug problem? Estimates and correlates from a national study of recovering U.S. adults. Alcohol Clin Exp Res 2019;43(7):1533-1544 [FREE Full text] [doi: 10.1111/acer.14067] [Medline: 31090945]
4. McKay JR, Hiller-Sturmhofel S. Treating alcoholism as a chronic disease: approaches to long-term continuing care. Alcohol Res Health 2011;33(4):356-370 [FREE Full text] [Medline: 23580020]
5. Castelpietra G, Knudsen AKS, Agardh EE, Armocida B, Beghi M, Iburg KM, et al. The burden of mental disorders, substance use disorders and self-harm among young people in Europe, 1990-2019: findings from the global burden of disease study 2019. Lancet Reg Health Eur 2022;16:100341 [FREE Full text] [doi: 10.1016/j.lanepe.2022.100341] [Medline: 35392452]
6. Degenhardt L, Charlson F, Ferrari A, Santomauro D, Erskine H, Mantilla-Herrara A, et al. The global burden of disease attributable to alcohol and drug use in 195 countries and territories, 1990–2016: a systematic analysis for the global burden of disease study 2016. Lancet Psychiatry 2018;5(12):987-1012. [doi: 10.1016/s2215-0366(18)30337-7]
7. Whiteford HA, Degenhardt L, Rehm J, Baxter AJ, Ferrari AJ, Erskine HE, et al. Global burden of disease attributable to mental and substance use disorders: findings from the global burden of disease study 2010. Lancet 2013;382(9904):1575-1586. [doi: 10.1016/s0140-6736(13)61611-6]
8. Alcohol. World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/alcohol [accessed 2022-05-09]
9. Cox G, Comiskey C. Characteristics of opiate users presenting for a new treatment episode: baseline data from the national drug treatment outcome study in Ireland (ROSIE). Drugs Educ Prev Policy 2009;14(3):217-230. [doi: 10.1080/09687630601032977]
10. Fletcher BW, Tims FM, Brown BS. Drug abuse treatment outcome study (DATOS): treatment evaluation research in the United States. Psychol Addict Behav 1997;11(4):216-229. [doi: 10.1037//0893-164x.11.4.216]
11. McKeganey N, Bloor M, McIntosh J, Neale J. Key findings from the Drug Outcome Research in Scotland (DORIS) study. ResearchGate. 2008. URL: https://www.researchgate.net/profile/Neil-Mckeganey/publication/255672321_Key_findings_from_the_Drug_Outcome_Research_in_Scotland_DORIS_study/links/0c96053b46da8b4f42000000/Key-findings-from-the-Drug-Outcome-Research-in-Scotland-DORIS-study.pdf [accessed 2025-02-13]
12. Teesson M, Mills K, Ross J, Darke S, Williamson A, Havard A. The impact of treatment on 3 years' outcome for heroin dependence: findings from the Australian treatment outcome study (ATOS). Addiction 2008;103(1):80-88. [doi: 10.1111/j.1360-0443.2007.02029.x] [Medline: 17995994]
13. Stahler GJ, Mennis J, DuCette JP. Residential and outpatient treatment completion for substance use disorders in the U.S.: moderation analysis by demographics and drug of choice. Addict Behav 2016;58:129-135. [doi: 10.1016/j.addbeh.2016.02.030] [Medline: 26925821]
14. Brorson HH, Arnevik EAA, Rand-Hendriksen K, Duckert F. Drop-out from addiction treatment: a systematic review of risk factors. Clin Psychol Rev 2013;33(8):1010-1024. [doi: 10.1016/j.cpr.2013.07.007] [Medline: 24029221]
15. Katz EC, Brown BS, Schwartz RP, Weintraub E, Barksdale W, Robinson R. Role induction: a method for enhancing early retention in outpatient drug-free treatment. J Consult Clin Psychol 2004;72(2):227-234. [doi: 10.1037/0022-006x.72.2.227]
16. Lappan SN, Brown AW, Hendricks PS. Dropout rates of in-person psychosocial substance use disorder treatments: a systematic review and meta-analysis. Addiction 2020;115(2):201-217. [doi: 10.1111/add.14793] [Medline: 31454123]
17. Simpson DD, Joe GW. A longitudinal evaluation of treatment engagement and recovery stages. J Subst Abuse Treat 2004;27(2):89-97. [doi: 10.1016/j.jsat.2004.03.001] [Medline: 15450643]
18. Arria AM, TOPPS-II Interstate Cooperative Study Group. Drug treatment completion and post-discharge employment in the TOPPS-II interstate cooperative study. J Subst Abuse Treat 2003;25(1):9-18. [doi: 10.1016/s0740-5472(03)00050-3] [Medline: 14512103]
19. Churruca K, Pomare C, Ellis LA, Long JC, Henderson SB, Murphy LED, et al. Patient-reported outcome measures (PROMs): a review of generic and condition-specific measures and a discussion of trends and issues. Health Expect 2021;24(4):1015-1024 [FREE Full text] [doi: 10.1111/hex.13254] [Medline: 33949755]
20. Davis EL, Kelly PJ, Deane FP, Baker AL, Buckingham M, Degan T, et al. The relationship between patient-centered care and outcomes in specialist drug and alcohol treatment: a systematic literature review. Subst Abus 2020;41(2):216-231. [doi: 10.1080/08897077.2019.1671940] [Medline: 31638870]
21. Migchels C, Zerrouk A, Crunelle CL, Matthys F, Gremeaux L, Fernandez K, et al. Patient reported outcome and experience measures (PROMs and PREMs) in substance use disorder treatment services: a scoping review. Drug Alcohol Depend 2023;253:111017. [doi: 10.1016/j.drugalcdep.2023.111017] [Medline: 37995391]
22. Trujols J, Portella MJ, Iraurgi I, Campins MJ, Siñol N, de Los Cobos JP. Patient-reported outcome measures: are they patient-generated, patient-centred or patient-valued? J Ment Health 2013;22(6):555-562. [doi: 10.3109/09638237.2012.734653] [Medline: 23323928]
23. Kingsley C, Patel S. Patient-reported outcome measures and patient-reported experience measures. BJA Educ 2017 Apr;17(4):137-144. [doi: 10.1093/bjaed/mkw060]

XSL•FO
RenderX

24. Black N, Chung S, Tisdale C, Fialho LS, Aramrattana A, Assanangkornchai S, et al. An international, multidisciplinary consensus set of patient-centered outcome measures for substance-related and addictive disorders. J Clin Med 2024;13(7):2154 [FREE Full text] [doi: 10.3390/jcm13072154] [Medline: 38610919]
25. Alves P, Sales C, Ashworth M. Does outcome measurement of treatment for substance use disorder reflect the personal concerns of patients? A scoping review of measures recommended in Europe. Drug Alcohol Depend 2017;179:299-308. [doi: 10.1016/j.drugalcdep.2017.05.049] [Medline: 28830035]
26. Friedrichs A, Spies M, Härter M, Buchholz A. Patient preferences and shared decision making in the treatment of substance use disorders: a systematic review of the lliterature. PLoS One 2016;11(1):e0145817 [FREE Full text] [doi: 10.1371/journal.pone.0145817] [Medline: 26731679]
27. Garnick DW, Horgan CM, Acevedo A, McCorry F, Weisner C. Performance measures for substance use disorders—what research is needed? Addict Sci Clin Pract 2012;7(1):18 [FREE Full text] [doi: 10.1186/1940-0640-7-18] [Medline: 23186374]
28. Kendrick T, El-Gohary M, Stuart B, Gilbody S, Churchill R, Aiken L, et al. Routine use of patient reported outcome measures (PROMs) for improving treatment of common mental health disorders in adults. Cochrane Database Syst Rev 2016;7(7):CD011119 [FREE Full text] [doi: 10.1002/14651858.CD011119.pub2] [Medline: 27409972]
29. Kolind T, Hesse M. Patient-centred care-perhaps the future of substance abuse treatment. Addiction 2017;112(3):465-466. [doi: 10.1111/add.13673] [Medline: 28168786]
30. Roe D, Mazor Y, Gelkopf M. Patient-reported outcome measurements (PROMs) and provider assessment in mental health: a systematic review of the context of implementation. Int J Qual Health Care 2021;34(Suppl 1):ii28–ii39. [doi: 10.1093/intqhc/mzz084] [Medline: 31725153]
31. Vanderplasschen W, Colpaert K, Autrique M, Rapp RC, Pearce S, Broekaert E, et al. Therapeutic communities for addictions: a review of their effectiveness from a recovery-oriented perspective. Sci World J 2013;2013(1):427817 [FREE Full text] [doi: 10.1155/2013/427817] [Medline: 23401669]
32. de Leon G. Community as Method: Therapeutic Communities for Special Populations and Special Settings. Westport, CT: Greenwood Publishing Group, Inc; 1997.
33. Vanderplasschen W, Vandevelde S, Broekaert E. Therapeutic Communities for Treating Addictions in Europe: Evidence, Current Practices and Future Challenges. Luxembourg: Publications Office of the European Union; 2014.
34. Zerrouk A. Outcome measurement and evaluation as a routine practice in alcohol and other drug services in Belgium (OMER-BE). Lisbon Addictions. URL: https://www.lisbonaddictions.eu/lisbon-addictions-2022/presentations/outcome-measurement-and-evaluation-routine-practice-alcohol-and-other-drug-services [accessed 2022-11-25]
35. Tsang S, Royse C, Terkawi A. Guidelines for developing, translating, and validating a questionnaire in perioperative and pain medicine. Saudi J Anaesth 2017;11(5):80. [doi: 10.4103/sja.sja_203_17]
36. Prins A, Bovin MJ, Smolenski DJ, Marx BP, Kimerling R, Jenkins-Guarnieri MA, et al. The primary care PTSD screen for DSM-5 (PC-PTSD-5): development and evaluation within a veteran primary care sample. J Gen Intern Med 2016;31(10):1206-1211 [FREE Full text] [doi: 10.1007/s11606-016-3703-5] [Medline: 27170304]
37. Patton SC, Hinojosa CA, Lathan EC, Welsh JW, Powers A. Validating the primary care posttraumatic stress disorder screen for DSM-5 (PC-PTSD-5) in a substance misusing, trauma-exposed, socioeconomically vulnerable population. Addict Behav 2023;139:107592 [FREE Full text] [doi: 10.1016/j.addbeh.2022.107592] [Medline: 36584543]
38. Lovibond SH, Lovibond PF. Manual for the Depression Anxiety Stress Scales. Sydney, Australia: Psychology Foundation of Australia; 1995.
39. Beaufort IN, de Weert-van Oene GH, Buwalda VA, de Leeuw JRJ, Goudriaan AE. The Depression, Anxiety and Stress Scale (DASS-21) as a screener for depression in substance use disorder inpatients: a pilot study. Eur Addict Res 2017;23(5):260-268 [FREE Full text] [doi: 10.1159/000485182] [Medline: 29224000]
40. Adler LA, Kessler RC, Spencer T. Adult ADHD Self-Report Scale-v1. 1 (ASRS-v1. 1) Symptom Checklist. New York, NY: World Health Organization; 2003.
41. Blanco CD, Ramos-Quiroga JA, Valero S, Bosch R, Roncero C, Gonzalvo B, et al. Adult ADHD self-report scale (ASRS-v1.1) symptom checklist in patients with substance use disorders. Actas Esp Psiquiatr 2009;37(6):299-305 [FREE Full text] [Medline: 20066581]
42. van de Glind G, van den Brink W, Koeter MW, Carpentier PJ, van Emmerik-van Oortmerssen K, Kaye S, et al. Validity of the adult ADHD self-report scale (ASRS) as a screener for adult ADHD in treatment seeking substance use disorder patients. Drug Alcohol Depend 2013;132(3):587-596 [FREE Full text] [doi: 10.1016/j.drugalcdep.2013.04.010] [Medline: 23660242]
43. Pilkonis PA, Yu L, Colditz J, Dodds N, Johnston KL, Maihoefer C, et al. Item banks for alcohol use from the patient-reported outcomes measurement information system (PROMIS): use, consequences, and expectancies. Drug Alcohol Depend 2013;130(1-3):167-177 [FREE Full text] [doi: 10.1016/j.drugalcdep.2012.11.002] [Medline: 23206377]
44. Pilkonis PA, Yu L, Dodds NE, Johnston KL, Lawrence SM, Hilton TF, et al. Item banks for substance use from the patient-reported outcomes measurement information system (PROMIS(®)): severity of use and positive appeal of use. Drug Alcohol Depend 2015;156:184-192 [FREE Full text] [doi: 10.1016/j.drugalcdep.2015.09.008] [Medline: 26423364]
45. Burling AS, Burling T. A comparison of self-report measures of nicotine dependence among male drug/alcohol-dependent cigarette smokers. Nicotine Tob Res 2003;5(5):625-633. [doi: 10.1080/1462220031000158708] [Medline: 14577979]

XSL•FO
RenderX

46. Marsden J, Farrell M, Bradbury C, Dale-Perera A, Eastwood B, Roxburgh M, et al. Development of the treatment outcomes profile. Addiction 2008;103(9):1450-1460. [doi: 10.1111/j.1360-0443.2008.02284.x] [Medline: 18783500]
47. Pellicciari L, Chiarotto A, Giusti E, Crins MHP, Roorda LD, Terwee CB. Psychometric properties of the patient-reported outcomes measurement information system scale v1.2: global health (PROMIS-GH) in a Dutch general population. Health Qual Life Outcomes 2021;19(1):226 [FREE Full text] [doi: 10.1186/s12955-021-01855-0] [Medline: 34579721]
48. Neale J, Vitoratou S, Finch E, Lennon P, Mitcheson L, Panebianco D, et al. Development and validation of 'SURE': a patient reported outcome measure (PROM) for recovery from drug and alcohol dependence. Drug Alcohol Depend 2016;165:159-167 [FREE Full text] [doi: 10.1016/j.drugalcdep.2016.06.006] [Medline: 27344196]
49. Migchels C, Brink WVD, Zerrouk A, Matthys FI, de Ruysscher C, Debeer D, et al. Psychometric evaluation of the dutch version of the substance use recovery evaluator (SURE-NL). Eur Addict Res 2024:1-10 [FREE Full text] [doi: 10.1159/000541584] [Medline: 39522511]
50. Skevington SM, Lotfy M, O'Connell KA. The World Health Organization's WHOQOL-BREF quality of life assessment: psychometric properties and results of the international field trial. a report from the WHOQOL Group. Qual Life Res 2004;13(2):299-310. [doi: 10.1023/b:qure.0000018486.91360.00]
51. WHOQOL Group. Development of the World Health Organization WHOQOL-BREF quality of life assessment. Psychol Med 1998;28(3):551-558. [doi: 10.1017/s0033291798006667] [Medline: 9626712]
52. Hinsley K, Kelly PJ, Davis E. Experiences of patient-centred care in alcohol and other drug treatment settings: a qualitative study to inform design of a patient-reported experience measure. Drug Alcohol Rev 2019;38(6):664-673. [doi: 10.1111/dar.12965] [Medline: 31281991]
53. Kelly PJ, Hatton EL, Hinsley K, Davis E, Larance B. Preliminary psychometric evaluation of the patient reported experience measure for addiction treatment (PREMAT). Addict Behav 2021;123:107048. [doi: 10.1016/j.addbeh.2021.107048] [Medline: 34348223]
54. Migchels C, van den Brink W, Zerrouk A, Matthys F, de Ruysscher C, Vanderplasschen W, et al. Psychometric evaluation of the dutch version of the patient-reported experience measure for addiction treatment (PREMAT-NL). Drug Alcohol Rev 2024;43(7):2021-2034. [doi: 10.1111/dar.13939] [Medline: 39256987]
55. Coghill DR. Organisation of services for managing ADHD. Epidemiol Psychiatr Sci 2016;26(5):453-458. [doi: 10.1017/s2045796016000937]
56. de Andrade D, Elphinston RA, Quinn C, Allan J, Hides L. The effectiveness of residential treatment services for individuals with substance use disorders: a systematic review. Drug Alcohol Depend 2019;201:227-235. [doi: 10.1016/j.drugalcdep.2019.03.031] [Medline: 31254749]
57. Mohamed SM, Abdelmonem RM, Hassan SS. Relationship between self-efficacy, social support and treatment motivation among addict patients. Minia Sci Nurs J 2022;012(1):106-113. [doi: 10.21608/msnj.2022.169551.1039]
58. de Graaf R, van Dorsselaer S, Tuithof M, ten Have M. Sociodemographic and psychiatric predictors of attrition in a prospective psychiatric epidemiological study among the general population. Result of the Netherlands mental health survey and incidence study-2. Compr Psychiatry 2013;54(8):1131-1139. [doi: 10.1016/j.comppsych.2013.05.012] [Medline: 23810078]
59. Katsuno N, Li PZ, Bourbeau J, Aaron S, Maltais F, Hernandez P, et al. Factors associated with attrition in a longitudinal cohort of older adults in the community. Chronic Obstr Pulm Dis 2023;10(2):178-189. [doi: 10.15326/jcopdf.2022.0380] [Medline: 37099700]
60. Myers B, van der Westhuizen C, Pool M, Hornsby N, Sorsdahl KR. Responding to COVID-19 threats to trial conduct: lessons learned from a feasibility trial of a psychological intervention for South African adolescents. Trials 2021;22(1):440 [FREE Full text] [doi: 10.1186/s13063-021-05400-8] [Medline: 34243806]
61. Zhai Y. A call for addressing barriers to telemedicine: health disparities during the COVID-19 pandemic. Psychother Psychosom 2021;90(1):64-66 [FREE Full text] [doi: 10.1159/000509000] [Medline: 32498070]

Abbreviations

ADHD: attention-deficit/hyperactivity disorder
ICHOM: International Consortium for Health Outcomes Measurement
OAT: opioid agonist therapy
OMER-BE: Outcome Measurement and Evaluation as a Routine Practice in Alcohol and Other Drug Services in Belgium
PC: psychiatric treatment center
PREM: patient-reported experience measure
PREMAT: Patient-Reported Experience Measure for Addiction Treatment
PROM: patient-reported outcome measure
PROMIS: Patient-Reported Outcomes Measurement Information System
PTSD: posttraumatic stress disorder
SSA: Standard Set for Addictions

XSL•FO
RenderX

SUD: substance use disorder
SURE: Substance Use Recovery Evaluator
TC: therapeutic community
WHOQoL-BREF: brief version of the World Health Organization Quality of Life questionnaire

Edited by A Mavragani; submitted 22.08.24; peer-reviewed by M Farrell, M Ferri; comments to author 10.11.24; revised version received 05.02.25; accepted 06.02.25; published 30.04.25.

Please cite as:
Zerrouk A, Migchels C, De Ruysscher C, Fernandez K, Antoine J, De Meyer F, Matthys F, van den Brink W, Crunelle CL, Vanderplasschen W
Incorporating Patient-Reported Outcome Measures and Patient-Reported Experience Measures in Addiction Treatment Services in Belgium: Naturalistic, Longitudinal, Multicenter Cohort Study
JMIR Form Res 2025;9:e65686
URL: https://formative.jmir.org/2025/1/e65686
doi:10.2196/65686
PMID:

©Amine Zerrouk, Charlotte Migchels, Clara De Ruysscher, Kim Fernandez, Jerome Antoine, Florian De Meyer, Frieda Matthys, Wim van den Brink, Cleo Lina Crunelle, Wouter Vanderplasschen. Originally published in JMIR Formative Research (https://formative.jmir.org), 30.04.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Mixed Method Survey of Characteristics of HIV Care Facilities: Medical Monitoring Project Facility Survey Project

Dustin Williams[1*], MA; John Weiser[2*], MPH, MD; Timothy McManus[2*], MSc; Hanna B Demeke[2*], RN, MSN, PhD; Darryl Creel[1], MSc; Jason Craw[2*], MPH; Milton Cahoon[1*], MA; Linda Beer[2*], PhD

[1]RTI International, Durham, NC, United States

[2]Centers for Disease Control and Prevention, Atlanta, GA, United States

[*]these authors contributed equally

Corresponding Author:
Dustin Williams, MA
RTI International
3040 East Cornwallis Road
Durham, NC, 27709
United States
Phone: 1 8432241561
Email: dwilliams@rti.org

Abstract

Background: Measuring the capacity of HIV medical facilities to deliver quality treatment and prevention care to people with HIV is essential to the over 1 million Americans living with HIV and supports federal efforts to end the HIV epidemic. To fill this gap and complement the ongoing Medical Monitoring Project—which conducts annual surveys of people with HIV and periodic surveys of HIV care providers—the US Centers for Disease Control and Prevention (CDC) and RTI (Research Triangle Institute) International conducted the Medical Monitoring Project Facility Survey (MMPFS).

Objective: We aimed to describe the survey methods designed to achieve a high response rate from the 1022 facilities providing care to people with HIV as part of the Medical Monitoring Project—including frame development, survey instrument development, facility recruitment, and postsurvey data processing.

Methods: For the MMPFS, the CDC and RTI developed a sequential multimode data collection approach (paper, web, and phone), including an abbreviated nonresponse follow-up (NRFU) instrument and the collection of administrative data for all facilities. Data were then processed to produce raw, imputed, and weighted datasets. Analyses included comparisons of responses to the full survey and NRFU survey.

Results: The full MMPFS survey yielded 455 complete survey respondents and the NRFU survey yielded 59 complete survey responses, a combined response rate of 50.3% (514/1022). A nonresponse bias analysis comparing the 2 surveys found a significant difference in the raw datasets for 4 (12%) of the 34 categorical variables that were identical between the 2 surveys (all $P>.0014$). Weighted and imputed datasets were then generated and compared. There was no significant difference between the 2 datasets for any variable (all $P>.05$).

Conclusions: The CDC and RTI's MMPFS methodology proved to be a valuable means of collecting data from HIV care providers and providing estimates for facility characteristics related to the provision of health care for people with HIV. The combined response rate allowed the CDC and RTI to generate facility-level estimates and an imputed dataset that can be linked to MMPFS patient data. The methods may be applied to other facility survey studies.

(JMIR Form Res 2025;9:e52123) doi:10.2196/52123

KEYWORDS

multimode surveys; establishment surveys; imputation; weighting; nonresponse medical monitoring; HIV care; people with HIV; data collection; United States; Medical Monitoring Project; physicians; clinician; web surveys method; methods; HIV medical facilities; Americans; support; HIV epidemic; centers; mobile phone

XSL•FO
RenderX

Introduction

Over 1.1 million Americans currently live with HIV, and many more are at risk of infection [1]. In 2019, the federal government launched the Ending the HIV Epidemic (EHE): A Plan for America initiative to end the US HIV epidemic by 2030. EHE focuses on early diagnosis, rapid and effective treatment for sustained viral suppression, preventing new transmissions using pre-exposure prophylaxis and syringe services programs, and responding quickly to outbreaks with prevention and treatment services for people who need them [2].

To guide the EHE, the National HIV/AIDS Strategy, and other national HIV prevention and care efforts, the US Centers for Disease Control and Prevention (CDC) conducts the Medical Monitoring Project (MMP). The MMP is a nationwide representative behavioral and clinical surveillance system of adults diagnosed with HIV in the United States, sponsored by the CDC [3]. The MMP is an annual, cross-sectional, complex sample survey with a 2-stage sampling design. In the first stage, 16 states (including 6 separately funded cities) and Puerto Rico were sampled, that is, from all US states, the District of Columbia, and Puerto Rico. In the second stage, simple random samples of persons aged ≥18 years with diagnosed HIV were drawn for each jurisdiction from the National HIV Surveillance System. Data are collected through phone or in-person interviews and medical records are abstracted at the most frequent source of HIV care during the past 24 months.

To complement the MMP and provide information to inform EHE activities, the CDC and RTI (Research Triangle Institute) International conducted the Medical Monitoring Project Facility Survey (MMPFS) to collect data on care delivery, patient engagement and retention practices, and workforce characteristics from facilities where MMP participants received HIV care. Specifically, the MMPFS surveyed all HIV care facilities at which a medical record abstraction took place during the MMP's 2019 data collection cycle. The MMPFS collected data on the capacity of care facilities to deliver prevention care and services; provide HIV prevention messaging; partner with public health programs; offer services for HIV negative partners of people with HIV; engage and retain patients; and offer HIV pre-exposure prophylaxis, medication-assisted therapy, and other ancillary care and support services to those who need them. Information on facility location and workforce capacity was also collected to identify areas in need of expanded support to deliver these services. MMPFS data were also linked to MMP person-level data to enhance the understanding of facility-level facilitators and barriers and health care access and outcomes among people with HIV. Numerous studies have shown that HIV care facility characteristics—such as those related to funding, service provision, and workforce—play a role in access to services and shaping patient outcomes [4,5].

In this paper, we describe the MMPFS methods designed to achieve a high response rate and minimal nonresponse bias—including the creation of the frame of facilities surveyed; survey instrument development; data collection methods including the development of multiple survey modes and phases of data collection and facility recruitment; response rates; and postsurvey data processing, including nonresponse bias analysis, weighting, and imputation.

Methods

Instrumentation

The CDC and RTI attempted to collect data from 1022 HIV care facilities that provided HIV care during the MMP 2019 data collection cycle (June 2019 to May 2020) to a probability sample of adults with diagnosed HIV located in 23 MMP project areas: California (including the separately funded jurisdictions of Los Angeles County and San Francisco); Delaware; Florida; Georgia; Illinois (including Chicago); Indiana; Michigan; Mississippi; New Jersey; New York (including New York City); North Carolina; Oregon; Pennsylvania (including Philadelphia); Puerto Rico; Texas (including Houston); Virginia; and Washington. MMP participants identified these facilities as their most frequent source of HIV care during the past 24 months and were the place at which MMP medical record abstraction took place [3,6]. Human participants' approvals were not obtained for the MMFS because respondents were staff of the facilities and no information on individuals was collected.

For the MMPFS, the CDC and RTI's primary consideration was building survey instruments that offered busy physicians and facility administrators (targeted respondents for the MMPFS) a variety of options to reduce their response burden while maximizing response. To achieve this goal, the MMPFS full survey was programmed in web, mail, and phone modes, and a nonresponse follow-up survey (NRFU) was programmed in a web mode. The web surveys were programmed using Voxco (Voxco Group Inc) online to be easily viewed on a desktop or laptop computer, tablet, and smartphone to provide flexibility and reduce the burden for care providers and facility administrators with limited time to complete surveys. The phone survey was then programmed in Voxco's computer-assisted telephone interviewing mode using the web survey as its base to minimize programming costs. The mail survey was programmed using TeleForm (Open Text Corporation) to minimize the time needed to capture mail survey data while improving accuracy.

Any of the 1022 sampled HIV care facilities from the 23 MMPS project areas that did not respond to the full survey were asked to complete the abbreviated MMPFS NRFU survey, which consisted of a subset of critical survey items and was estimated to take 5 minutes to complete. Additionally, we collected publicly available administrative data for all MMPFS facilities.

The full survey instrument was designed to document the provision of services that support the 4 pillars of the EHE: prevent, diagnose, treat, and respond. Survey domains included general characteristics (eg, facility type, sources of revenue, patient volume, and provider staffing levels), HIV testing and pre- and postexposure prophylaxis provision, clinical and supportive services provided onsite and through established referral relationships, capacity for supporting rapid enrollment and initiation of antiretroviral therapy and retention in care, collaboration with health departments on public health activities, availability of telehealth visits, and policies to prevent patient

XSL•FO
RenderX

exposure to COVID-19. The survey instrument was developed in collaboration with MMP provider and community advisory board members, selected HIV care facility leaders, and the CDC and other federal agency subject matter experts.

Ethical Considerations

Because respondents were staff of the facilities and no information was collected on individuals, human participants' approval was exempted by the RTI's Institutional Review Board (ethics exemption approval: STUDY00021321). All responding facility data were maintained in a secure environment that could only be accessed by project staff. All publications of data were deidentified. Responding facilities were offered no compensation for their participation.

Frame Development

Concurrent with MMPFS instrumentation, the CDC and RTI worked with MMP project areas to develop and refine the MMPFS frame. MMPFS frame development began in October 2020 with the development of a MMPFS mailing list upload site for MMP project areas to upload a workbook containing data for all facilities providing care to MMP participants during the 2019 MMP data collection cycle. From October 2020 through April 2021, MMP project areas uploaded workbooks containing facility information (name, geographic location, and mailing address), primary contact information (name, title, phone number, and email address), and secondary contact information (name, title, phone number, and email address) for each facility.

Due to high levels of item missingness for primary contact name and email address information, the CDC and RTI conducted web-based tracing on 544 facilities with missing frame data from May to June 2021. During web-based tracing, the RTI telephone interviewers conducted web searches to identify all missing information for each facility. After locating missing information, telephone interviewers then conducted for up to 5 call attempts to verify the accuracy of information identified via web search. Using web-based tracing, the CDC and RTI were able to locate sufficient contact information for mail and phone recruitment for nearly all MMPFS facilities. After web-based tracing, of the 1022 MMPFS facilities, 960 had no missing contact information, 58 had a partial facility record (missing some, but not all contact information), and 4 were missing all contact information. All 1022 cases were loaded into the MMPFS case management system with a unique case ID after web-based tracing—results from web-based tracing are presented in Figure 1.

Figure 1. Medical Monitoring Project Facility Survey interactive tracing results (N=1022). CDC: Centers for Disease Control and Prevention; MMPFS: Medical Monitoring Project Facility Survey; RTI: Research Triangle Institute.

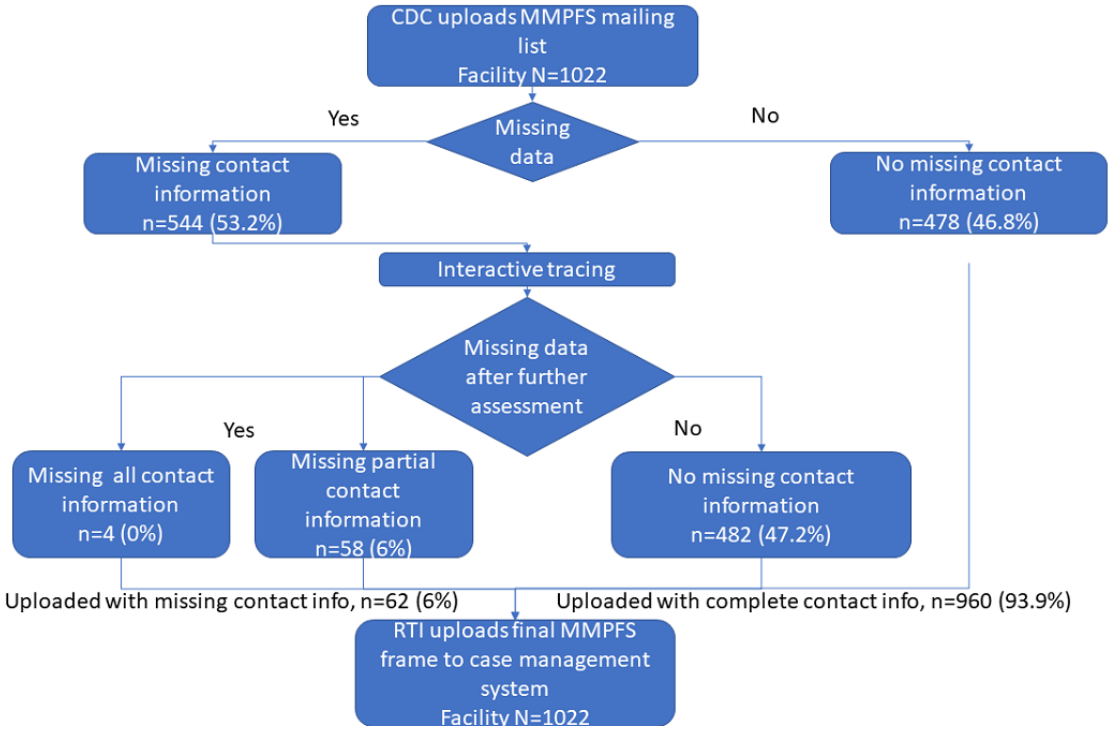

Survey Recruitment

For the MMPFS full survey, the CDC and RTI used a modified version of Don Dillman's Tailored Design Method to recruit facilities to participate using a variety of recruitment methods and varied mailing types [7].

MMPFS recruitment began with mail and email invitations followed by a reminder postcard and 3-week mail and email reminders—the 4 facilities missing all contact information were excluded from mail and email recruitment activities but were included in response rate calculations. Given the slower-than-expected response after the 3-week reminders, the CDC and RTI worked with the Health Resources and Services Administration, HIV Medicine Association, and Association of Nurses in AIDS Care to develop an email and letter endorsing the MMPFS that leaders of these 3 agencies cosigned. The 3-agency letter of support email was sent 2 weeks after the 3-week reminders. In the 2 weeks following the letter of support email, the number of completed full surveys increased from

XSL•FO
RenderX

176 to 278, which suggests this email had a positive impact on recruitment.

The 3-agency letter of support was included in the 7-week mail and email reminders. After all mail and email outreach was completed for the full survey, phone outreach was initiated to nonresponding agencies.

MMPFS phone recruitment began on September 14, 2021, lasting through October 22, 2021. During phone survey prompting, a facility received up to 5 call attempts—unless they completed the survey during a prior call attempt, completed a survey via full web or mail survey between call attempts, or refused to participate during a call attempt. Overall, a total of 707 facilities received call attempts. While only 1 phone survey was completed during phone prompting, an additional 219 cases were completed via the web or mail modes following phone prompting, which suggests this was an effective recruitment tactic for the web and mail modes.

After completion of the full survey recruitment, the CDC and RTI launched the MMPFS NRFU survey. NRFU surveys are surveys where the nonrespondents (refusals or noncontacts) are asked to complete a shorter questionnaire [8]. The MMPFS NRFU survey was critical for ensuring the MMPFS represented the full population of HIV facilities at which MMP participants received care. For example, if certain types of MMP facilities responded at lower rates than other facilities, it could introduce bias. The MMPFS NRFU survey was designed to determine whether nonresponding facilities differed systematically from responding facilities on key survey data elements and mitigate any potential biases.

From October 18, 2021, to October 25, 2021, depending on the date that phone recruitment ended for the facility, the CDC and RTI sent a NRFU web survey invitation email followed by a reminder email to all 490 facilities that had not responded or refused to participate.

In addition, on October 21, 2021, a list of all nonresponding agencies in each project area's jurisdiction was made available for each project area and the CDC asked project areas to encourage nonresponding facilities to complete the MMPFS NRFU survey. Before this outreach, there were 674 pending facilities and there were an additional 33 completed full surveys at the end of the full survey field period without additional outreach. The CDC and RTI found the number of additional completed full surveys to be greater than anticipated at this point in data collection (and with NRFU data collection ongoing) and believe the advocacy of MMP project areas helped boost response, with 111 respondents completing after the start of this recruitment activity. Additional experimentation would be needed to determine the efficacy of such interventions for other surveys, but this suggests that engaging entities that have established relationships with facilities in recruitment activities (as was the case for the MMP project areas) increases response rates.

Additional details on MMPFS full and NRFU survey recruitment activities and completion rates before each activity are shown in Tables 1 and 2—completion rates use a denominator of 1022 and include the 4 facilities that did not receive recruitment correspondence due to missing contact information.

XSL•FO
RenderX

Table 1. MMPFS[a] recruitment milestones by date, type, facilities receiving outreach, and cumulative response and completion percentage on the date of recruitment activity: full survey.

Event	Date	Facilities receiving recruitment outreach	Cumulative responses on the date of recruitment activity (N=1022), n (%)	Notes
MMPFS email invitation	July 23, 2021	1018	0 (0)	__[b]
MMPFS mail invitation	July 23, 2021	1018	0 (0)	Invitation letter, paper survey, and business reply envelope
MMPFS reminder postcards	July 30, 2021	975	43 (4.2)	—
MMPFS 3-week reminder email	August 13, 2021	914	104 (10.2)	—
MMPFS 3-week reminder letter	August 13, 2021	914	104 (10.2)	Reminder letter, paper survey, and business reply envelope
HRSA[c]/HIVMA[d]/ANAC[e] support email	August 25, 2021	842	176 (17.2)	—
MMPFS 7-week reminder email	September 10, 2021	740	278 (27.2)	—
MMPFS 7-week reminder letter	September 10, 2021	740	278 (27.2)	Reminder letter, HRSA/HIVMA/ANAC letter of support, paper survey, and business reply envelope
MMPFS phone nonresponse prompting	September 13, 2021, to October 25, 2021	707	311 (30.4)	Up to 5 call attempts
MMPFS project area outreach	October 21, 2021, to November 15, 2021	674	344 (33.7)	—
MMPFS data collection closes	November 15, 2021	—	455 (44.5)	—

[a]MMPFS: Medical Monitoring Project Facility Survey.

[b]Not applicable.

[c]HRSA: Health Resources and Services Administration.

[d]HIVMA: HIV Medicine Association.

[e]ANAC: Association of Nurses in AIDS Care.

Table 2. MMPFS[a] recruitment milestones by date, type, facilities receiving outreach, and cumulative response and completion percentage on the date of recruitment activity: NRFU[b] survey.

Event	Date	Facilities receiving recruitment outreach	Cumulative responses (N=1022), n (%)
MMPFS NRFU survey email invitations	October 18, 2021, to October 25, 2021	490	0 (0)
MMPFS NRFU survey reminder email	November 8, 2021	452	38 (3.7)
MMPFS data collection closes	November 15, 2021	0	59 (5.8)

[a]MMPFS: Medical Monitoring Project Facility Survey.

[b]NRFU: nonresponse follow-up.

Data Processing

In November 2021, the CDC and RTI conducted web searches to collect publicly available administrative data for all 1022 facilities, including Federal Information Processing System code, primary care health professional shortage area designation, medically underserved area or population designation, Rural-Urban Continuum Code, and Ryan White HIV/AIDS Program funding type. In addition, the CDC provided the number of abstractions performed at each facility during the 2019 MMP data collection cycle to be included in the minimal dataset. The goal of the minimal dataset was to fill the remaining gaps in the survey data on critical items (eg, Ryan White Program funding), improve survey estimates, and provide valuable information for nonresponse adjustments.

After completing the minimal dataset, a dataset for the full MMPFS survey was produced. The full survey dataset contained data on 155 analytic variables for all 1022 responding

XSL•FO
RenderX

facilities—only frame and minimal dataset data were included for full survey nonresponders.

A raw dataset was then produced for the MMPFS NRFU survey. The NRFU dataset only included the 567 facilities that did not complete a MMPFS full survey. The MMPFS NRFU dataset did not include full survey respondents to allow for comparisons between the responders in the 2 datasets. The MMPFS NRFU dataset had 36 analytic variables—only frame and minimal dataset data were included for NRFU survey nonresponders.

After producing the raw datasets, data quality checks and editing procedures were applied to both datasets, including checks for valid values, correct codes assigned for missing and unknown data, checks for skip pattern consistency and appropriate coding of skipped variables, and appropriate coding of missing variables from broken off or partial surveys.

All cases were then assigned a final status code based on response type (full complete, full partial, NRFU complete, NRFU partial, noncontact, and noncooperation) and the number of valid responses for a facility—the number of valid responses excludes the address verification question as a possible response variable. Facilities in the eligible, noninterview noncontact status group did not attempt to complete a web, mail, or phone survey for the full or NRFU surveys. Facilities in the eligible, noninterview noncooperation status groups accessed a full or NRFU survey but did not provide any valid responses beyond the address verification question. The Full Survey Respondent status group had at least one valid response and a full survey complete response—excluding address confirmation. The NRFU Survey Respondent status group had at least one valid response and a NRFU survey complete response. The MMPFS full and NRFU surveys did not have any ineligible facilities. That is, all facilities were deemed eligible at the start of this study because they provided care to a 2019 MMP participant.

Nonresponse Bias Analysis

The CDC and RTI then conducted a nonresponse bias analysis on the MMPFS data. The goal of the nonresponse bias analysis was to identify the best variables to use for postdata collection weight adjustments, which use variables that are associated with response propensity and key outcome variables [9]. For the MMPFS, the RTI and CDC wanted to assess whether certain HIV care facilities' characteristics (funding, types of health coverage or insurance accepted, HIV caseload, and staff training); onsite clinical and support services or referral agreements for off-site services; rapid linkage to HIV care and antiretroviral therapy initiation practices and barriers; use of HIV telemedicine; use of data to systematically monitor retention of care; and technology to increase retention of care were associated with response propensity.

First, full survey and NRFU survey data were compared for the frame and minimal dataset variables to compare differences in facility types or operations. Second, the CDC and RTI compared the responses for the full survey respondents to the NRFU survey respondents for the 35 analytic variables on both datasets to compare responses. Further, 34 of the variables were categorical and 1 variable was continuous. Categorical variables were analyzed by comparing the proportions with a value of 1, usually the "yes" category. Instead of viewing the 35 tests as individual tests at α equal to .05, we viewed the 35 tests as a single "family" of tests and controlled the familywise error rate to minimize the probability of making false discoveries. To account for the multiple testing conducted, the α levels were adjusted for significance using the Bonferroni adjustment, which is α divided by the number of tests. For MMPFS, it was assumed the α was equal to .05 and the Bonferroni adjusted α is .05/35 (approximately .0014). Therefore, for the MMPFS, observed P values had to be less than .0014 to be statistically significant. Bonferroni is a conservative approach to the multiple comparison problem, but, if we were to err, we wanted to err on the side of not making a false discovery (ie, decrease the chance of a type I error, with a possible increased chance of a type II error). In addition, the RTI ran a Holm-Bonferroni adjustment, which yielded the same results.

Postsurvey Weight Adjustments

For the MMPFS postsurvey weight adjustments, we first compared the number of abstractions completed at each facility during the 2019 data collection cycle with the patient with HIV load size reported on the full survey question. Based on this comparison, no clear association was identified between the reported patient load and number of MMP abstractions completed, indicating that the number of MMP abstractions completed was not an accurate measure of facility size.

Next, postsurvey weight adjustments were calculated so that the 455 full survey respondents would represent the 1022 facilities. The postsurvey weight evaluation started with all 9 frame and minimal dataset variables in the weight adjustment model (not counting the number of abstractions that had been previously eliminated). Variables were then removed using backward elimination, with the variable with the largest P value greater than .05 being removed from the model. The backward elimination modelling process was repeated until all variables in the weight adjustment model had P values less than or equal to .05. Table 3 shows the variable description and type for 3 Ryan White HIV/AIDS Program funding variables, frame data on MMP abstractions (other frame data was respondent contact information, which was not of interest), and all minimal dataset variables included in the weight adjustment model—this includes all minimal dataset variables. Finally, the CDC and RTI used the cross-classification of the 3 Ryan White variables to create the weight adjustments and final weights.

XSL•FO
RenderX

Table 3. Medical Monitoring Project Facility Survey weight adjustment types by variable types and inclusion in the weight adjustment model

Description	Variable type	Included in weight adjustment model
Number of Medical Monitoring Project abstractions at this facility (a proxy for the size of the facility)	Continuous	No
Primary health care professional shortage area	Binary	No
Medically underserved area or population	Binary	No
Rural-urban continuum code	Ordinal (categories 4-9 into a single nonmetro county category)	No
Does the facility receive Ryan White HIV/AIDS program funding?	Binary	Yes
Does the facility receive Ryan White HIV/AIDS Part A program funding?	Binary	Yes
Does the facility receive Ryan White HIV/AIDS Part B program funding?	Binary	No
Does the facility receive Ryan White HIV/AIDS Part C program funding?	Binary	No
Does the facility receive Ryan White HIV/AIDS Part D program funding?	Binary	No
Does the facility receive Ryan White HIV/AIDS Part F program funding?	Binary	Yes

Imputation

The CDC and RTI then performed imputation for nonresponding MMPFS facilities. Imputation allowed the CDC to link facility characteristics to MMP patient-level data for all MMP patients. The imputation methodology consisted of 2 steps: recursive partitioning (trees) to create imputation classes and then a weighted sequential hot deck to produce the imputed values. For each variable or vector of variables imputed, a prediction tree was constructed from the respondents. The terminal nodes of the tree are the imputation classes and every respondent was in an imputation class. Once the tree was constructed, we ran the nonrespondents through the tree so they could be placed in an imputation class. After the nonrespondents were placed in an imputation class there was a respondent and nonrespondent imputation class. Next, the weighted sequential hot deck was implemented within each imputation class. The final weight for the full survey respondents was used for their imputations, and a weight of 1 was used for all facility imputations. For imputation, Ryan White Funding and the node were used as class variables; and Ryan White Part A Funding, Ryan White Part F Funding, and the number of patients for whom the facility provided HIV care during the past year were used as sorting variables. If the cross-classification of Ryan White Funding and node created an imputation class without any donors, the minimum size of the nodes was increased until every node had donors.

The imputation methodology was applied in three steps:

1. Used data from donors (completed survey responses from facilities that had similar frame and minimal dataset characteristics) to impute missing items on the full survey respondent dataset for nonrespondents.
2. Imputed missing items on the NRFU respondent dataset. For items on the full and NRFU surveys, missing NRFU values were imputed using donors (respondents) from the full and NRFU survey data. For items not on the NRFU survey, answers were imputed using data from full survey respondents.
3. Imputed all items for all nonrespondents using the full survey and NRFU imputed dataset respondents.

The imputed and weighted survey estimates were then compared using the weighted survey estimate and imputed survey estimate to develop a contrast (difference) survey count, contrast survey estimates, contrast SE, contrast 95% CI lower limit, contrast 95% CI upper limit, and contrast *P* value.

Results

Instrumentation

For the MMPFS, 455 respondents completed full surveys (253 web surveys, 201 mail surveys, and 1 phone survey) and 59 completed a NRFU web survey of the entire sample of 1022 facilities. The final combined final response rate was 50.3% (514/1022)—the full survey response rate was 44.5% (455 full survey responses out of 1022 sampled facilities) and the NRFU survey response rate was 5.8% (59 NRFU survey responses out of 1022 sampled facilities) of the total sample. While NRFU response rates are expected to be low, the low NRFU response rate suggests that the length of the full survey may not have been the reason for not responding to the full survey for many facilities.

Survey Recruitment

The full MMPFS survey was open from July 23, 2021, to November 15, 2021, and allowed facilities to complete the survey across 3 modes: web, mail, or phone. Overall, a total of 260 full web surveys were submitted, including duplicate and partial web surveys; 253 were conducted on desktop computers and 7 on smartphones. The average completion time for the full

XSL•FO
RenderX

web survey was 14 (SD 11) minutes. The mail survey was conducted simultaneously and a total of 201 MMPFS mail surveys were completed. The phone survey was open from September 13, 2021, to November 15, 2021, and a total of 1 phone survey was completed.

The MMPFS NRFU survey was open from October 18, 2021, to November 15, 2021, and allowed facilities to complete an abbreviated web survey. Overall, a total of 59 surveys were submitted via the web NRFU survey. For the NRFU web survey, 54 were conducted on desktop computers and 5 on smartphones. The average completion time for NRFU web surveys was 6 (SD 4) minutes. Details on survey completion by mode are included in Tables 4 and 5. In addition, 111 responses were received after project area outreach started, which suggests that the personal and professional relationships the project areas have with facility administrators might have been a powerful tool for encouraging response.

Table 6 contains a summary of MMPFS facilities by status group.

Table 4. Medical Monitoring Project Facility Survey responses and completion rate by mode: full survey.

Event	Responses by mode (N=1022), n (%)
Web survey	260 (25.4)
Mail survey	194 (19)
Phone survey	1 (0.1)
Full survey total	455 (44.5)

Table 5. Medical Monitoring Project Facility Survey responses and completion rate by mode: NRFU[a] survey.

Mode	Cumulative responses (N=1022), n (%)
Web survey	59 (5.8)
NRFU survey total	59 (5.8)

[a]NRFU: nonresponse follow-up.

Table 6. Medical Monitoring Project Facility Survey response count, percent, cumulative count, and cumulative percentage by final contact status groups (N=1022).

Status group	Value, n (%)	Cumulative value, n (%)
Eligible, noninterview noncontact (never accessed either survey)	474 (46.38)	474 (46.38)
Eligible, noninterview noncooperation (accessed a survey, but did not provide any valid responses)	34 (3.33)	508 (49.71)
Full survey respondent	455 (44.52)	963 (94.23)
Nonresponse follow-up survey respondent	59 (5.77)	1022 (100)

Next, MMPFS contact, cooperation, and response rates were calculated. For the MMPFS, contact rates were calculated by dividing the number of respondents who made any attempt to complete a full or NRFU survey (n=548), including break offs, by the total sample (N=1022). The cooperation rate is then calculated by dividing the number of completed surveys by the number of contacted respondents who provided 1 valid response beyond the address verification question for each survey, using the number of contacted respondents (n=548) for the denominator on both surveys. For the MMPFS surveys, the response rate is the product of the contact and cooperation rates. Table 7 shows 3 different rates (contact, cooperation, and response rates) for 3 different groups of respondents (combined full and NRFU survey respondents together, only full survey respondents, and only NRFU survey respondents).

Table 7. Medical Monitoring Project Facility Survey contact counts, contact rates, cooperation counts, cooperation rates, and response rates by respondent type (N=1022).

Group	Contact (N=1022), n (%)	Cooperation (n=548), n (%)	Response rate (contact × cooperation; %)
Full and NRFU[a] survey respondents	548 (53.6)	514 (93.8)	51
Full survey respondents	548 (53.6)	455 (83)	45
NRFU survey respondents	548 (53.6)	59 (10.8)	6

[a]NRFU: nonresponse follow-up.

XSL•FO
RenderX

Nonresponse Bias Analysis

There was no statistically significant difference for 30 (88%) of the 34 categorical variables tested (all $P>.0014$), and 2 of the differences were where the NRFU survey estimate was 0, meaning the MMPFS full survey data provides essentially the same results as analyzing the MMPFS full and NRFU survey combined data. Table 8 shows the variable description, full survey estimate, NRFU survey estimates, contrast (difference) estimate, contrast SE estimate, contrast 95% CI lower limit, contrast 95% CI upper limit, and contrast *P* value for the 4 statistically significant MMPFS analytic variables—those with observed contrast *P* values smaller than .0014 (.0005, .0009, .0005, and .0004), the Bonferroni adjusted α. Given that the vast majority of the variables were not statistically significantly different between the datasets—meaning analysis of the MMPFS full survey data alone provides essentially the same results as analyzing the MMPFS full and NRFU survey combined data—the CDC and RTI decided not to use the NRFU data for the weighted facility estimates.

Table 8. Medical Monitoring Project Facility Survey full survey and NRFU[a] survey variable estimates, contrast survey counts, contrast estimates, contrast SE estimates, contrast 95% CI lower and upper limits, and contrast *P* values for variables with a significant difference (.0005, .0009, .0005, and .0004) between the full and NRFU survey estimates.

Variable description	Full survey estimate	NRFU survey estimate	Contrast survey count[b]	Contrast estimate	Contrast SE estimate	Contrast 95% CI limit	Contrast *P* value
Is the facility a federally qualified health center look-alike?	2.65	0	511	–2.65	0.76	–4.14, –1.17	.0005
Does the facility accept no common types of health coverage?	2.44	0	509	–2.44	0.73	–3.88, –1.01	.0009
Is there normally a physician at the facility at least 5 days per week who can provide HIV care?	77.03	91.53	503	14.5	4.14	6.36, 22.64	.0005
Does the facility provide onsite substance use disorders treatment?	31.54	55.93	506	24.39	6.83	10.96, 37.81	.0004

[a]NRFU: nonresponse follow-up.

[b]Number of combined responses from full and nonresponse follow-up surveys for each variable.

Postsurvey Weight Adjustments

Table 9 shows the final weights for all 1022 facilities, including 567 facilities with a weight of 0 (all nonrespondents to the full survey, which includes MMPFS NRFU survey respondents). The remaining 455 facilities are the full survey respondents with weights ranging from 1 to approximately 3 with an unequal weighting of ~1.1153. This results in an estimate of the effective sample size—the sample size divided by the design effect (for this study, the unequal weighting effect)—of approximately 408 facilities (455 facilities/1.1153).

Table 9. Final survey weights for Medical Monitoring Project Facility Survey (N=1022).

Final weight	Value, n (%)	Cumulative value, n (%)
0.0000	567 (55.48)	567 (55.48)
1.0000	94 (9.2)	661 (64.68)
1.8229	96 (9.39)	757 (74.07)
2.5522	67 (6.56)	824 (80.63)
2.9394	198 (19.37)	1022 (100)

Imputation

The comparison of the imputed and weighted data revealed that none of the contrast *P* values were statistically significant (all $P>.05$) Therefore, using either the imputed or weighted dataset would generate essentially the same point estimates.

The critical difference between the MMPFS imputed and weighted datasets is that the SEs for the imputed data are roughly 60% of the SEs of the weighted data, which could lower the *P* values for analyses performed using the imputed dataset. This is because the imputed dataset has about twice as many facilities as the weighted dataset, has no variability in the weights, and is treated as observed respondent data. That is, the imputed data do not account for the uncertainty in the imputed values. Due to the weighted data having generally larger SEs, a conservative approach to facility-level data analysis would be to use the weighted dataset. An alternative would be to conduct multiple imputation, which creates better estimates of SEs by accounting for differences in the SEs from the multiple imputations. However, multiple imputation was not used because it would require creating multiple datasets for the facility analysis and complicate patient-level analysis that would use the multiple imputed facility datasets. That is, each of the *M* multiple imputed facility datasets would have to be merged to the patient-level dataset to create *M* merged facility or patient

XSL•FO
RenderX

datasets to conduct the analysis, which could introduce difficulties for users trying to conduct patient-level analysis using facility characteristics.

Despite the smaller SEs for facility-level estimates using the single imputation dataset, we concluded that imputed facility values could be validly appended onto the patient-level MMP dataset without impacting patient-level variance estimates because those are based on the patient-level dataset and its structure. That is, the imputed facility data used for patient-level analysis are used as characteristics of patients and not as facility-level analytic variables.

Discussion

The CDC and RTI's MMPFS methodology proved to be a valuable means of collecting data from HIV care facilities and providing estimates for critical factors related to the provision of health care for people with HIV. The methodology yielded a 50.3% response rate—which is relatively high for establishment surveys—with minimal nonresponse bias by using the described frame development, instrument development, data collection methods, facility recruitment, and postsurvey data processing methods. The combined response rate allowed the CDC and RTI to generate facility-level estimates and an imputed dataset that can be linked to MMPFS patient data.

These methods may be applied to other surveys of HIV care facilities, medical establishments, and health care providers, as this methodology used innovative methods (eg, the NRFU survey) and demonstrated ways to maximize response. This is significant given the lack of publications on establishment surveys, broadly, and medical establishment surveys, specifically. These types of surveys are critical for public health but often use methods designed and tested for household surveys.

The 50.3% combined response rate was notable given the potential negative impacts on facility response due to the ongoing COVID-19 pandemic during MMPFS data collection and might be partially attributed to the multimode design. The MMPSFS also allowed the CDC and RTI to generate facility-level estimates and an imputed dataset that can be linked to MMPFS patient data. Additionally, MMPFS results suggest that there may be a benefit to using letters of support from federal agencies and professional organizations and the endorsement of local health departments to boost participation in surveys of medical facilities, which is worth further study.

While the minimal dataset for the MMPFS only yielded 3 weight adjustment variables, these variables were valuable in producing datasets that could be linked to patient data. Future surveys of medical establishment and health care providers should consider the variables used by the MMPFS and explore additional sources of administrative data to identify other significant variables that could be used to improve future weighting and imputation efforts.

In addition, if there are future iterations of the MMPFS (or similar studies of health care providers and medical establishments), the costs and benefits of the NRFU survey should be reconsidered, given the costs of the NRFU survey and that there were few significant differences (P=.0005, .0009, .0005, and .0004) between the full survey and NRFU survey data for this iteration. If there are significant changes in the facility characteristics and provision of care for people with HIV (eg, more MMP facilities are part of large health networks at the time of survey administration), it may be worth conducting another NRFU survey. However, if there are few changes in the population and a similar response is elicited from the initial survey, it may be prudent to forego a NRFU survey and use those resources to boost response (eg, additional recruitment activities).

A limitation of this assessment is that we did not perform any formal experiments to test the MMPFS methodology. In the future, it would be useful to design formal experiments on medical establishment and health care provider surveys to better understand the effects of time between survey outreach events and response rates. For busy medical facilities that may take longer to process noncritical mail than households or other establishments, it could be useful to test whether more time between mailings and a longer data collection period yield a higher response rate. It may also be useful to test the effect of the use of "special mailing" types (eg, Federal Express or Priority Mail) on medical facility response. Such special mailings may be more likely to stand out to overburdened medical facilities and generate an improved response rate. In addition, we did not conduct a formal assessment of survey costs, which could be helpful for the implementation of MMPFS methods in other settings.

The MMPFS methodology proved valuable in collecting data on care delivery, patient engagement and retention practices, and workforce characteristics from facilities where MMP participants received HIV care. This data will provide information that will help inform future EHE activities. In addition, MMPFS methods can be used to inform and improve response on future surveys of HIV care facilities, medical establishments, and medical providers. Specifically, the findings illuminate the benefits of using letters of support, multimode and multiphase data collection approaches, and the use of administrative data for data processing and analysis activities.

Acknowledgments

We thank MMP facility survey participants, project area staff, and provider and community advisory board members. We also acknowledge the contributions of the Clinical Outcomes Team and Behavioral and Clinical Surveillance Branch at the CDC and RTI International. The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention. The funding for the Medical Monitoring Project is provided by the Centers for Disease Control and Prevention.

XSL•FO
RenderX

Data Availability

Data are not publicly available due to security and confidentiality guidelines governing the collection of public health surveillance data. Restricted access may be granted on request from the CDC.

Authors' Contributions

DW did the conceptualization, methodology, project administration, supervision, software, writing of the original draft, and review and editing of the writing. LB and JW handled the conceptualization, methodology, supervision, and review and editing of the writing. TM worked on the data curation, formal analysis, methodology, software, validation, and review and editing of the writing. HBD and MC assisted with the conceptualization, data curation, investigation, methodology, and review and editing of the writing. JC carried out the conceptualization, methodology, supervision, and review and editing of the writing. DC aided with the conceptualization, data curation, validation, formal analysis, methodology, and review and editing of the writing.

Conflicts of Interest

None declared.

References

1. HIV surveillance supplemental report: estimated HIV incidence and prevalence in the United States, 2015-2019. Centers for Disease Control and Prevention. 2021. URL: https://www.cdc.gov/hiv/pdf/library/reports/surveillance/cdc-hiv-surveillance-supplemental-report-vol-26-1.pdf [accessed 2024-10-25]
2. Office of Infectious Disease and HIV/AIDS Policy. About ending the HIV epidemic: plan for America. HIV.gov. URL: https://www.hiv.gov/ending-hiv-epidemic#:~:text=among%20healthcare%20providers.-,Goal,250%2C000%20total%20HIV%20infections%20averted.&text=to%20end%20the%20HIV%20epidemic.%E2%80%9D [accessed 2024-10-11]
3. Beer L, Johnson CH, Fagan JL, Frazier EL, Nyaku M, Craw JA, et al. A national behavioral and clinical surveillance system of adults with diagnosed HIV (the Medical Monitoring Project): protocol for an annual cross-sectional interview and medical record abstraction survey. JMIR Res Protoc 2019;8(11):e15453 [FREE Full text] [doi: 10.2196/15453] [Medline: 31738178]
4. Weiser J, Beer L, Frazier EL, Patel R, Dempsey A, Hauck H, et al. Service delivery and patient outcomes in Ryan White HIV/AIDS Program-funded and -nonfunded health care facilities in the United States. JAMA Intern Med 2015;175(10):1650-1659 [FREE Full text] [doi: 10.1001/jamainternmed.2015.4095] [Medline: 26322677]
5. Weiser J, Dempsey A, Mandsager P, Shouse RL. Documenting successes 30 years after passage of the Ryan White CARE Act: to the editor. J Assoc Nurses AIDS Care 2021;32(2):138-139 [FREE Full text] [doi: 10.1097/JNC.0000000000000224] [Medline: 33427768]
6. HIV surveillance special report 28: behavioral and clinical characteristics of persons with diagnosed HIV infection-medical monitoring project, United States, 2019 cycle (June 2019' May 2020). Centers for Disease Control and Prevention. 2021. URL: https://stacks.cdc.gov/view/cdc/149117/cdc_149117_DS1.pdf [accessed 2024-05-01]
7. Dillman DA, Smyth JD, Christian LM. Internet, Phone, Mail, and Mixed Mode Surveys: The Tailored Design Method. 4th ed. Hoboken, NJ: John Wiley & Sons Inc; 2014.
8. Bethlehem JG, Kersten HMP. On the treatment of nonresponse in social survey. J Off Stat 1985;1(3):287-300 [FREE Full text]
9. Little RJ, Vartivarian S. Does weighting for nonresponse increase the variance of survey means? Surv Methodol 2005;31(2):161-168 [FREE Full text]

Abbreviations

CDC: Centers for Disease Control and Prevention
EHE: Ending the HIV Epidemic
MMP: Medical Monitoring Project
MMPFS: Medical Monitoring Project Facility Survey
NRFU: nonresponse follow-up
RTI: Research Triangle Institute

XSL•FO
RenderX

Edited by A Mavragani; submitted 23.08.23; peer-reviewed by KK Kota, A Fullem, P Wang, NS Wong; comments to author 17.04.24; revised version received 31.05.24; accepted 04.09.24; published 29.05.25.

Please cite as:
Williams D, Weiser J, McManus T, Demeke HB, Creel D, Craw J, Cahoon M, Beer L
A Mixed Method Survey of Characteristics of HIV Care Facilities: Medical Monitoring Project Facility Survey Project
JMIR Form Res 2025;9:e52123
URL: https://formative.jmir.org/2025/1/e52123
doi:10.2196/52123
PMID:

©Dustin Williams, John Weiser, Timothy McManus, Hanna B Demeke, Darryl Creel, Jason Craw, Milton Cahoon, Linda Beer. Originally published in JMIR Formative Research (https://formative.jmir.org), 29.05.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Co-Designing an Infant Early Childhood Mental Health Mobile App for Early Childhood Education Teachers' Professional Development: Community-Based Participatory Research Approach

Ruby Natale[1*], PsyD, PhD; Elizabeth Howe[1*], PhD; Carolina Velasquez[1*], PhD; Emperatiz Guzman Garcia[1*], LMHC, IMH-E; Karen Granja[1*], MS, RMHC; Bianca Caceres[2*], LCSW, CWLC; Elizabeth Erban[3*], LMFT, IMH-E; Tania Ramirez[4*], LMFT, IMH-E; Jason Jent[1*], PhD

[1]School of Medicine/Mailman Center for Child Development, University of Miami, Miami, FL, United States

[2]Community Health of South Florida, Miami, FL, United States

[3]Family Central, Miami, FL, United States

[4]Jewish Community Services, Miami, FL, United States

[*]all authors contributed equally

Corresponding Author:
Ruby Natale, PsyD, PhD
School of Medicine/Mailman Center for Child Development
University of Miami
1601 NW 12th Avenue
Miami, FL, 33137
United States
Phone: 1 3052434617
Email: rnatale@med.miami.edu

Abstract

Background: Many young children spend at least some time in early care and education programs, where they develop social-emotional skills that prepare them for future success. However, young children may exhibit behavioral challenges in these settings, negatively impacting their social-emotional development. It is critical that the early childhood workforce is prepared to support young children's burgeoning social-emotional skills to address challenging behaviors in early care and education classrooms. Infant and early childhood mental health consultation is an evidence-informed approach for increasing teachers' skills for managing young children's emotions and behaviors. One mechanism to increase teachers' access and use of the infant and early childhood mental health consultation programs is through on-demand mobile apps.

Objective: This study aims to investigate 2 primary objectives: to document the development of the Jump Start on the Go (JS Go) app through community-based participatory research (CBPR) methodologies, and to evaluate and refine the app based on early childhood education (ECE) teacher feedback using a mixed methods assessment approach.

Methods: This study used a community-based participatory research approach to design and evaluate the effectiveness of the JS Go app across 3 phases. In phase 1, a description of how the JS Go app was developed using CBPR principles is provided. In phase 2, teachers (n=12) were interviewed after reviewing mockups of the JS Go app to gather feedback about the interface and usefulness of the app to current and new teachers. Rapid qualitative analysis generated themes to inform phase 3 (n=31) of the study.

Results: Phase 2 findings suggested that teachers viewed the app as aesthetically pleasing with concise information, but there were design and content features that needed to be refined to improve ease of use for accessing content. Teachers also described the app as beneficial and useful to both current and new ECE teachers and identified it as a tool to support sustainability for the use of JS practices. In phase 3, teachers rated the JS Go app favorably across all mHealth (mobile health) App Usability Questionnaire dimensions, including interface satisfaction (mean 6.12 on a 7-point scale), ease of use (mean 5.56), and usefulness (mean 5.37). Despite positive usability ratings, teachers expressed less certain intentions to adopt the app, scoring near the midpoint on the Technology Acceptance Model Instrument-Fast Form's predicted future use scale (mean 1.60, –4 to +4-point scale). Implications for how the findings were used to make adaptions to the app are discussed. The next steps for testing the efficacy of the app in a randomized control trial are described.

XSL•FO
RenderX

Conclusions: ECE teachers have overall positive perceptions about the value of the JS Go app. Future research will need to test the efficacy of the app for increasing and sustaining teacher's use of JS practices.

(JMIR Form Res 2025;9:e66714) doi:10.2196/66714

KEYWORDS

early care and education; child care; social-emotional development; young children; professional development; infant early childhood mental health consultation; mobile apps; community-based participatory research; mixed methods; mobile phones

Introduction

Background

Behavioral and mental health issues represent some of the greatest unaddressed challenges for young children, impacting their school readiness [1]. Preschool children who exhibit behavioral challenges are suspended and expelled from early care and education (ECE) programs at 3 times the rate of children in kindergarten-12th grade schools [2]. Moreover, racially and ethnically diverse children living in low-income communities are more likely to exhibit behavioral challenges leading to suspension or expulsion [1,2]. ECE programs play a vital role in promoting young children's mental health by providing environments where children can develop social-emotional skills, including emotional regulation and resilience [3]. Research indicates that young children who attend high-quality ECE programs exhibit lower levels of behavioral problems and higher social competence compared to those who do not attend ECE programs [1,4].

ECE teachers in high-quality ECE programs can act as a protective factor by supporting young children's burgeoning social-emotional development, thus preventing challenging behaviors [1,5]. Moreover, enhancing young children's social-emotional development is linked with their school readiness, making a focus on this aspect of children's development a mechanism for eliminating disparities [1]. However, there are limited professional development supports available to ECE teachers for implementing evidence-based social-emotional practices, particularly to support young children from low-income communities [6].

Infant early childhood mental health consultation (IECMHC) is an evidence-based model for supporting teachers in developing competencies for implementing social-emotional practices to support young children's social competence and prevent challenging behaviors [7-9]. In an IECMHC model, teachers receive support from trained mental health consultants (MHC) to learn and practice new skills related to building social-emotional skills. Less than 10% of ECE teachers across the nation report access to MHC, yet over half report feeling that they struggle to manage children's behavior [10]. IECMHC models offer accessible continuous professional development (CPD) for ECE teachers to guide the implementation of evidence-based social-emotional teaching practices. In a CPD model, teachers are responsible for creating their learning journey [11] which can increase their knowledge, sense of agency and confidence, perceptions of their competence, and ability to reflect on their practices leading to sustained use of new skills [12].

Mobile App Technology

Mobile technology is a CPD mechanism that may increase the uptake of IECMHC practices. The prevalence of mobile technology is increasing, suggesting that mobile apps may be a critical mechanism for eliminating disparities in service receipt. For example, 258 billion apps were downloaded in 2022 [13], with the World Development Report indicating that the poorest households are more likely to have a smartphone than electricity [14], and Hispanic adults are one of the leading users of smartphones [15].

Mobile apps are critical CPD tools for just-in-time, on-demand learning [16]. Mobile apps address the personal agency of the learner to decide when, where, and how they want to learn [16]. Mobile apps are becoming an increasingly common mechanism for providing CPD support in ECE programs [17]. ECE teachers have made progress in integrating digital tools and technology into their everyday practice [18] and these tools can support the sustainability of new skills by teachers [18]. In the ECE space, mobile apps are used for a variety of pedagogical purposes, including teaching strategies to increase young children's physical activity, implement educational activities (eg, science, technology, engineering, and mathematics), and increase teachers' creative competence [19-21].

There is emerging evidence that mobile health (mHealth) apps can support the development of a therapeutic alliance between an MHC and teacher [22] which is a key component of IECMHC models. However, there is limited research on how mobile apps support teachers in building relationships with children and collaborating with families [23], which are also essential elements of an IECMHC model. To our knowledge, there is yet to be a mobile app dedicated to providing IECMHC content to ECE teachers.

The Study

To address a gap in IECMHC mobile technology, our team engaged in a community-based participatory research (CBPR) model to co-develop a mental health app to provide the Jump Start (JS) program, an IECMHC model [7]. Jump Start on the Go (JS Go) is a mechanism for providing CPD to ECE teachers. The first objective of this study was to document the development and features of the JS Go mobile app, which emerged through CBPR methodologies The second objective of this investigation was to systematically evaluate and refine the JS Go app through comprehensive feedback from ECE teachers, using a mixed methods approach to assess both implementation successes and opportunities for app enhancement. The findings were used to evaluate teachers' perceptions of the perceived ease of use, usefulness, user interface, and the predicted future use of the JS Go mobile app

XSL•FO
RenderX

as a CPD tool for the delivery of IECMHC practices to ECE teachers in Miami-Dade and Broward Counties, Florida. Children and teachers in ECE programs in Miami-Dade and Broward Counties are racially and ethnically diverse, and many live in lower-income communities with multiple neighborhood risk factors contributing to poorer school outcomes. Given the population of children, teachers, and ECE programs in these counties, it is a priority that the community voice and perspective contribute to the development and evaluation of JS Go.

Methods

Study Design

The study was implemented in three phases using exploratory sequential mixed methods. Phase 1 focused on the use of CBPR principles to develop the app. Phase 2 was dedicated to preliminary user testing using qualitative interviews with a small number of ECE teachers who reviewed mock-ups of JS Go to answer the following research questions in order to refine the app: (1) What feedback do ECE teachers have about the interface and aesthetics of the JS Go app for ease of use? (2) How do ECE teachers describe the usefulness of the JS Go app for both current and new teachers? Phase 3 used formal real-world testing of the app by ECE teachers currently receiving JS IECMHC services to answer the following questions: (1) What are ECE teachers ratings for the ease of use, their satisfaction, the user interface, and the usefulness of JS Go for accessing IECMHC content using JS Go? (2) What are teachers' ratings for the usefulness, ease of use, and predicted future use of the JS Go app as a mechanism for consultation?

Ethical Considerations

This study was approved by the University of Miami School of Medicine, Human Research Subjects Internal Review Board (approval number 20231026). Written informed consent was obtained from all participants prior to data collection. Study data are anonymous, and all participant responses were deidentified using unique identification numbers. Data were stored in cloud-based electronic systems for security. Participants in phase 2 received a US $20 gift card for the interview and participants in phase 3 received a US $20 gift card for beta testing the app and a certificate for 5 hours of in-service training for their participation in the study.

Study Phases

Phase 1: CBPR for Mobile App Development

CBPR is a collaborative approach to research that actively involves community members in the research process to address issues of local concern and promote social change [24]. We applied a CBPR framework developed by Israel et al [25,26] that includes 9 principles (Textbox 1). Unlike traditional research methods that often distance researchers from the communities they study, CBPR seeks to equalize the research process by engaging community members as equal partners in the design, implementation, and evaluation of research projects [25,26]. The CBPR approach emphasizes mutual respect, shared decision-making, and a focus on actionable outcomes that benefit the community [27]. CBPR has been shown to enhance the relevance and quality of research by ensuring that it reflects the community's needs and perspectives, thus improving the likelihood of successful interventions and sustainable solutions [28].

Textbox 1. Description of how Israel et al's 9 principles of community-based participatory research were used to develop and test Jump Start on the Go (JS Go).

Through our long history of effective community partnerships, we address the following 9 principles of CBPR.

- Recognizing the community as a unit of identity, we acknowledge our partners' unique and individual contributions to ensure the cultural appropriateness and accessibility of JS on the Go.
- Building on community strengths and resources by including the community voice in the design, development, and refinement of JS on the Go.
- Facilitating collaborative equitable partnerships in all phases of the research by including our partners in all stages of JS on the Go for co-development, refinement, testing, and dissemination.
- Involving an empowering and power-sharing process that attends to social inequities by including our community partners who are trusted sources as co-investigators to foster co-learning and ensure reach to early childhood education (ECE) programs in neighborhoods' historic disparities.
- Integrating and achieving a balance between knowledge generation and intervention for the mutual benefit of all partners through quarterly co-creation meetings with our community partners to design, develop, and refine JS on the Go.
- Focusing on the local relevance of public health problems and ecological perspectives that recognize and attend to multiple determinants of health by including partners from communities with historic disparities in access to ECE programs resulting in disproportionate outcomes to obtain their perspectives and gain an understanding of their backgrounds.
- Involving systems development using a cyclical iterative process through our exploratory sequential mixed methods research design implemented in multiple phases to gain an understanding of the acceptability, feasibility, usefulness, and usability of JS on the Go.
- Disseminating findings and knowledge gained to all partners and involving them in the dissemination process by co-developing publications and presentations about JS on the Go.
- Involving a long-term process and commitment to sustainability with pledged support from our community partners to support the development of future grants to bring JS to Go to scale across Miami-Dade County and nationwide.

Using a CBPR approach for the development and testing of a mobile app can enhance its relevance, usability, and effectiveness by incorporating the perspectives of community members throughout the process [15,29,30]. This approach can

XSL•FO
RenderX

ensure that the mobile app meets the specific needs of the intended users and fosters community engagement and ownership [30]. A CBPR approach for mobile app development and testing is imperative for use in low-income, racially, and ethnically diverse communities [15,29] because it allows community members to tailor technology to meet their unique and individualized needs using a strengths-based approach, increasing the likelihood of their sustained use of the tool [30].

The app was developed using the 9 principles of CBPR by creating a community advisory committee. Our partners in the community advisory committee are composed of 3 community-based organizations: Jewish Community Services, Community Health of South Florida, and Family Central. All of these community-based organizations provide JS MHC in ECE programs in Miami-Dade County. The community advisory committee also included members of the Women's Fund of Miami-Dade County and their grantees. The Women's Fund is a nonprofit advocacy organization investing in initiatives for girls and women, including the early childhood workforce, in Miami-Dade County.

The community advisory committee oversaw the development of the app using the 8 dimensions of the Ecological Validity Model (Textbox 2 [31]) to ensure that the app had the necessary cultural and linguistic adaptations to reduce health disparities [31,32]. The Ecological Validity Model is the most widely used approach for culturally adapting interventions to ensure they are usable for the intended audience and feasible to use within the desired settings [33]. Our community advisory committee assisted with co-developing scripts for the videos and modifying content to ensure their relevance to ECE programs within their communities. Given that the app was multilingual, the community advisory committee ensured that the dialect of Spanish and Haitian Creole that was used in videos and text within the app aligned with local community linguistic patterns. Agendas during the initial meetings included explanations of the purpose and goals of the study, describing the JS IECMHC intervention, sharing prototypes, and eliciting feedback. We revised the multilevel intervention based on community advisory committee feedback, presenting the revised content of JS Go at subsequent meetings. If there were extensive changes based on feedback during the final meetings, we requested their permission to have additional meetings to ensure that all feedback was appropriately incorporated.

Textbox 2. The 8 dimensions of the Ecological Validity Model used in the development of Jump Start on the Go (JS Go) and the methods for testing.

- Language: Culturally appropriate language
- Persons: Role of ethnic or racial similarities and differences between client and therapist in shaping the relationship
- Metaphors: Symbols and concepts shared by populations
- Content: Cultural knowledge, values, traditions—the uniqueness of groups
- Concepts: Concepts consonant with culture and context
- Goals: Transmission of positive and adaptive cultural values from the culture of origin
- Methods: Development and cultural adaptation of treatment methods
- Context: Consideration of changing contexts in assessment during treatment or intervention

Phase 2: Preliminary Qualitative End User Testing

Participants and Procedures

After the content of the JS Go app was developed, we interviewed 12 teachers between April 2024 and May 2024 on the usability and feasibility of JS Go. We purposefully recruited participants (n=12) from 12 ECE centers in low-income tracts within Miami-Dade and Broward Counties currently participating in the JS program. Participants were selected based on their role as program champions in the larger randomized control trial for JS and their participation in supplemental professional development to obtain their infant mental health endorsement. There was a total of 32 program champions who completed the professional development. Participants were notified about the study by their MHC. Once participants were notified, they were contacted via email by the interviewer to schedule the interview.

Measures

The data from phase 1 was collected as part of a larger qualitative study. Interviewers used a semistructured interview guide developed by the research team for data collection. The questions in the interview guide were developed using the Reach, Effectiveness, Adoption, Implementation, and Maintenance framework [34,35]. The Reach, Effectiveness, Adoption, Implementation, and Maintenance framework uses an implementation science lens for evaluating interventions to determine their applicability to real-world settings (Textbox 3 and Multimedia Appendix 1). The questions posed to participants in phase one focused on the maintenance of JS Go.

XSL•FO
RenderX

Textbox 3. Preliminary end user testing interview questions for phase 2.

Phase 2: Interview questions

- **What feedback do you have regarding the 5 screenshots?**
  - Home page
  - Journey
  - Providers
  - Services
  - Videos
- What additional features or changes would you suggest for the app?
- How helpful can this app be for helping teachers sustain their use of Jump Start (JS) practices?
- How helpful can this app be for helping new teachers learn to use JS practices?
- Are there any other comments, questions, or concerns you would like to share about the app?

Interviews with teachers were conducted by a Master level MHC unfamiliar to the teacher to minimize bias. Interviews were conducted one-on-one with participants using Zoom videoconferencing software (Zoom Video Communications). Interviews were conducted in either English or Spanish depending on the preference of the participant. The average length of the interviews was 45 minutes 32 seconds with a range of 37 minutes 11 seconds to 65 minutes 29 seconds.

Formal user testing was captured via think-aloud walkthrough interviews. Think-aloud walkthroughs are a common method to evaluate the usability of apps and involve users verbalizing experiences and perceptions as they navigate through the app [36-39]. Participants are able to provide both positive and negative reactions to their experiences during the walkthrough to minimize bias. Mock-ups of the app were shared with teachers via screenshots of each page by the MHC (Multimedia Appendix 1). This allowed participants to react to the features of the app including the aesthetics and content. Transcripts of the recorded interviews were downloaded from Zoom to a secure cloud storage platform and deidentified. Transcripts were cleaned by the lead interviewer and translated, then back-translated by certified translators prior to analysis.

Phase 3: Formal Quantitative Real-World Testing

Participants and Procedures

Following the completion of phase 2, and consistent with CBPR, we then engaged in a 3-month feasibility field study (June 2024-August 2024) in which MHC from our 3 community-based organizations co-facilitated formal real-world, or beta testing, of the JS Go app with teachers. Teachers in phase 3 did not participate in the interviews during phase 2. Purposeful sampling was used to have MHC identify ECE teachers they were currently or had recently worked with to participate in phase 3 of the study (n=31). ECE teachers came from 18 centers in low-income tracts in Miami-Dade and Broward Counties in South Florida participating in the JS program. Participants were eligible to participate if they were currently or recently enrolled (within 3 months) in the JS program. Participants were notified by their consultant to determine their interest and informed consent was obtained prior to beta testing.

Formal real-world testing occurred over a 3-week period during regularly scheduled consultation sessions between the teacher and their MHC. During consultations, MHC used the app in tandem with the teachers. Teachers completed surveys following beta testing with the app. Surveys were completed by teachers through the JS Go app. MHC did not have access to teachers' responses to minimize bias. Teachers' responses to the surveys in the app were linked to a secure cloud-based data management platform, REDCap (Research Electronic Data Capture, Vanderbilt University) [40,41].

Formal real-world testing was completed via rapid-cycle prototyping in 3 cycles with each cycle lasting 3 weeks. In cycle 1, seven centers and 16 teachers participated; in cycle 2, seven centers and 13 teachers participated; and in cycle 3, two centers and 4 teachers participated. During this time period, teachers used the app as part of their consultation sessions with their MHC. Teachers completed surveys after using the app during their consultations to gauge their perceptions of the acceptability, feasibility, and usability of the app's features and the functionality of the app for telehealth consultation. To aid in using the survey data open-ended questions were used during weekly meetings between the app developer and MHC to address poorly scoring items. Examples of issues addressed between cycles included how teachers create their profile and log in to the app, ensuring resources are downloaded to their smartphone rather than opening in a new browser window, and ensuring infographics are available in all languages.

Following the completion of each cycle, we made iterative changes to the app based on the survey data and feedback from MHC.

Measures

During phase 3, teachers completed the mHealth App Usability Questionnaire (MAUQ) in English and Spanish [42]. The mHealth App Usability Questionnaire is a short reliable and customizable questionnaire to assess the usability of mobile apps to gauge teachers' perceptions of the usefulness of JS Go and its acceptability as a tool for delivering IECMHC. The mHealth App Usability Questionnaire has 18 items organized into 3 subscales; ease of use and satisfaction, system information

XSL•FO
RenderX

arrangement, and usefulness. Participants rated JS Go using a 7-point Likert scale (1=strongly disagree to 7=strongly agree). The internal consistency for all scales within the mHealth App Usability Questionnaire measure was high ranging from α=.876 (ease of use and satisfaction) and α=.908 (system information arrangement) to α=.843 (usefulness) [42].

In addition to the MAUQ, the Technology Acceptance Model Instrument-Fast Form (FF-TAM) in English and Spanish was used to measure teachers' beliefs about the functionality of the JS Go app as a mechanism for teleconsultation [43]. The FF-TAM measures teachers' acceptance of technology which predicts their perceptions of the ease of use of the app as a telehealth consultation mechanism. The FF-TAM has 16 items organized into 3 subscales; usefulness, ease of use, and predicted future use. The FF_TAM evaluates attitudes toward technology use and can be modified to include the technology of interest. In this study, the form was adapted to measure participants' beliefs about the JS Go app for teleconsultation. Participants rate their beliefs about the JS Go app using an 8-point semantic differential scale ranging from –4=most negative rating of an item to +4=most positive rating of an item. The Cronbach α for the scales ranged from 0.935 to 0.957 [43].

Data Analysis

During phase 2, rapid qualitative analysis was used to identify pertinent themes from the interviews with teachers [44-48]. The rapid qualitative analysis allowed the software developer to make real-time changes to the app based on feedback from teachers in phase 2 to optimize beta testing in phase 3. Rapid qualitative analysis has standards for rigor [47] and the use of this type of analysis is supported by implementation science research. Rapid qualitative analysis is a method of expediting qualitative analysis to address pressing issues, such as the timely adaptation of the app to meet the needs of the ECE workforce. The process involves summarizing interview transcripts using a standard template and subsequently transferring key points into a matrix to explore relevant themes [48]. The MHC who conducted the interview completed the initial summaries. A second member of our research team transferred the key points from the interview summary into the matrix. A third member of the team used the matrix to identify themes, which were then refined through discussion with study investigators and other members of the team.

Quantitative data collected during phase 3 were analyzed using SPSS (IBM Corp) [49]. The Cronbach α was calculated for both measures, including each subscale. Descriptive statistics were used to calculate the means and SDs for each of the subscales for both the MAUQ and FF-TAM as well as the total score and item range.

Results

JS Go

The content of JS Go is based on the evidence-based JS program [50]. JS includes 4 pillars based on Caring for Our Children–National Health and Safety Standards [51]. The 4 pillars are safety and planning, trauma-informed behavior support, adult self-care, and effective communication. JS Go encourages independent learning by allowing users to drive their access to the app by navigating content based on their personal needs. The app includes a self-assessment for using JS practices, infographics describing JS content, videos illustrating JS practices in action, and access to other supplemental materials and resources. JS Go also includes an artificial intelligence–driven chatbot that is specifically trained on JS content. Teacher users can ask questions and receive responses based on machine learning. An app, such as JS Go, developed with community partners via CBPR, could serve as a scalable paradigm shift by bringing MHC to teachers who need it most in real time.

Participant Characteristics

All of the teachers in phase 2 were female. Most of the teachers reported they were White and Hispanic, and their primary language was Spanish. The majority of the teachers were between the ages of 40 and 49 years and most have a bachelor's degree (Table 1).

All of the teachers in phase 3 were female. Most of the teachers reported they were White and Hispanic, and their primary language was Spanish. The majority of the teachers were between the ages of 30 and 39 years and most reported having some college, which includes technical training and holding an associate's degree (Table 1).

XSL•FO
RenderX

Table 1. Participant characteristics for preliminary end user testing (phase 2) and formal real-world testing (phase 3) of JS Go.

Characteristics	Phase 2 participants	Phase 3 participants
Sex, n (%)[a]		
Female	12 (100)	31 (100)
Male	0 (0)	0 (0)
Age (years), n (%)[b]		
20-29	1 (9)	1 (3)
30-39	2 (18)	10 (32)
40-49	6 (55)	7 (23)
50-59	1 (9)	8 (26)
More than 60	1 (9)	5 (16)
Education level, n (%)[c]		
High school or GED[d]	2 (18)	4 (15)
Some college[e]	3 (27)	13 (48)
Bachelors	4 (36)	7 (23)
Graduate	2 (18)	4 (15)
Primary language, n (%)[a]		
English	3 (25)	7 (23)
Spanish	9 (75)	24 (77)
Race, n (%)[a]		
Black non-Hispanic	1 (8)	5 (16)
White	11 (92)	26 (84)
Ethnicity, n (%)[a]		
Hispanic	11 (92)	27 (87)
Non-Hispanic	1 (8)	4 (13)

[a]12 participants in phase 2 and 31 in phase 3.
[b]11 participants in phase 2 and 31 in phase 3.
[c]11 participants in phase 2 and 27 in phase 3.
[d]GED: General Educational Development.
[e]Some college includes technical training and associate's degree programs.

Phase 2: Preliminary Qualitative End User Testing

Research Questions 1. JS Go Interface

Two themes were generated regarding teachers' perception of the JS Go interface for ease of use. Figure 1 shows the JS Go home page The first theme was the high perceived ease of use with 3 subthemes: concise information, aesthetically pleasing, and language accessibility. Most of the participants reported they were satisfied with the JS Go app's appearance as it allowed for ease of use. One participant noted, "It's, it's been very good, I like that. You know, the first thing we get is a little introduction; it's quite good." The participants also described the JS Go app's interface as clear and concise, with the training section as straightforward and direct. For example, 1 respondent stated "The good thing about the page is that I see it with little information. Pages loaded with information, what they tend to do is confuse us." High perceived ease of use also included the appearance of the content for different sections of the app and for resources such as videos. Participants described the aesthetics as "eye-catching" with a "minimalistic approach" that is "straightforward." Participants also appreciated the translation feature and language change option. One respondent stated the following.

> *For me, it's fine. And you even have it translated. Because what it says up there that says SPA (Spanish), it's very important that it's translated. There are many teachers who don't speak English* [Teacher #3, female]

XSL•FO
RenderX

Figure 1. JS Go (Jump Start on the Go) home page.

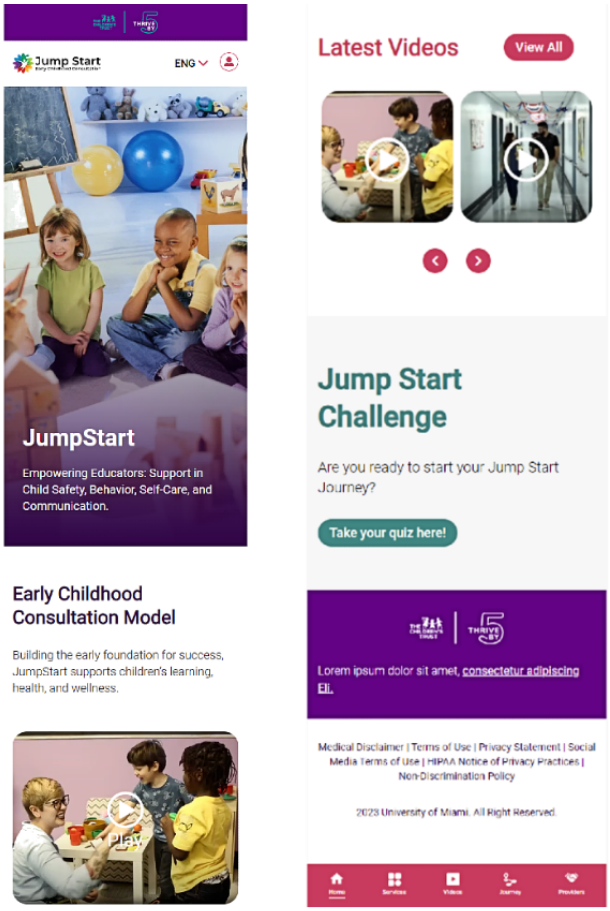

Another participant noted the following.

> *It's a well-thought-out app, but the only thing I see is that it has an option to change languages. If it has an option for Spanish, it would be good to create one for Creole as well because, at least in the area where we work, there are many teachers who are Haitian. They do speak English, but obviously, their first language is Creole* [Teacher #2, female]

The second theme generated included suggestions related to the app's visual attractiveness, functionality, and app content (Figure 1). Suggestions related to aesthetics included color choices, and reducing white space to optimize visibility on smartphones. Suggestions related to functionality include adding a search bar to the home page, changing the wording on the home page to "empowering educators and families," and improving the navigation of the home page to make it clearer. Other suggestions related to functionality include having a space for specific child cases to obtain follow-up strategies akin to an individual behavior plan and how the app is structured, such as including a start button. Suggestions about the content included requests for additional education videos that are concise and directly related to the JS program and videos demonstrating how to resolve various classroom situations. While yet another suggestion from a participant regarding content was to integrate songs as supplementary tools for classroom use.

Research Question 2. Helpfulness of JS Go

There were 4 themes generated describing how ECE teachers and directors perceive the helpfulness of the JS Go app; benefits, usefulness for current teachers, usefulness for new teachers, and sustainability (Figure 1). Participants describe the JS Go app as beneficial for accessing resources such as handouts, videos, courses, and infographics to facilitate teachers' jobs for timely receipt of information. Benefits also included the use of the app for planning and access to additional support when an IECMH consultant is not available. For example, 1 respondent noted the following.

> *I see it as an extra resource, right at their fingertips. That they can even like on their break, can look something up. I mean, I know I rather look up things on the phone or like an app, then go to like a website like sometimes I will be doing something on my phone. And I'm in front of the computer at work. But just because it was so much easier to just click, click, and just keep doing it on the phone* [Teacher #3, female]

Participants perceived the JS Go app as useful for current teachers for reviewing, refreshing, or reinforcing the use of JS practices for on-demand learning. For example, 1 respondent stated the following.

> *I think so because when you have doubts about something, you can go there and search the application. I think so. There are times when you have*

XSL•FO
RenderX

> *doubts about something or you don't remember something and you say 'you have to do this, I have the doubt' and you go in and you can reflect and look and do...* [Teacher #1, female]

Participants perceived the JS Go app as especially important for new teachers, especially following the workforce shortage resulting from the pandemic [52]. Participants described the app as useful for introducing JS content and having access to direct resources. The benefits included increasing new teachers' autonomy, bridging the gap between theory and practice, and having direct access to resources when things are not working in the classroom. For example, 1 participant stated, "New teachers have two ways of learning, either through theory when they go to school or during training, but often it doesn't apply in practice." Another respondent noted the following.

> *For her, it would be like a learning method, telling her, 'Look, here you have this tool that will help you when you feel things aren't working, when you feel frustrated, or when you have any problem. Open the app and, just like you study the courses or anything else, start studying the app* [Teacher #7, female]

Participants described the JS Go app as important for promoting the use of JS practices by providing supportive resources such as videos and classroom management strategies. The potential of the app for sustainability can be attributed to the practicality and clarity of the videos, and by providing an immediate resource to address unexpected learning situations, which is especially important when the director or consultant is not available. For example, 1 respondent noted the following.

> *The videos, the videos then stay a lot because that is like a practical example that was given and sometimes one looks for videos and sometimes they are difficult to find because they are like inside a link and the link that I send you something else and then that it has this direct and that you can see it. With the titles and names I think it's excellent* [Teacher #4, female]

The themes generated using rapid qualitative analysis are enlisted in Textbox 4.

Textbox 4. Relevant themes generated through rapid qualitative analysis.

Research question 1: Teachers' perceptions for the interface of Jump Start on the Go (JS Go)

1. High perceived use
   - Concise information
   - Aesthetically pleasing
   - Language accessibility
2. Suggestions
   - Aesthetics
   - Functionality
   - Content

Research question 2: Teachers' perceptions for the helpfulness of JS Go

1. Benefits
   - Access to resources
   - Planning
   - Access to support
2. Usefulness for current teachers
   - On-demand learning
3. Usefulness for new teachers
   - Content introduction
   - Access to resources
   - Autonomy
4. Sustainability
   - Supportive resources
   - Practicality
   - Immediate resource

XSL•FO
RenderX

Phase 3: Formal Quantitative Real-World Testing

mHealth App Usability Questionnaire

The total mean for the measure was 5.64 with an SD of 1.71, indicating participants were satisfied with the JS Go app and believed it is easy to use and useful (Table 2). The mean subscale scores across all 3 cycles of testing indicated that teachers had a positive perception of the ease of use, satisfaction with the app interface, and usefulness of the JS Go app.

Table 2. Teacher perceptions of the ease of use, satisfaction, usefulness, and predicted future use of the JS Go app (N=31).

Scale	Mean (SD)	Item response range (lowest to highest)
mHealth App Usability Questionnaire[a]		
Ease of use	5.56 (1.48)	1.80-7.00
Interface and satisfaction	6.12 (1.02)	3.29-7.00
Usefulness	5.37 (1.38)	1.83-7.00
Technology Acceptance Model Instrument-Fast Form[b]		
Usefulness	2.97 (1.56)	–2.00 to 4.00
Ease of use	2.76 (1.77)	–2.67 to 4.00
Predicted future use	1.60 (2.94)	–4.00 to 4.00

[a]mHealth: mobile health; mHealth App Usability Questionnaire scores range from 1=most negative rating to 7=most positive rating.

[b]Technology Acceptance Model Instrument-Fast Form scores range from –4=most negative rating to +4=most positive rating.

Technology Acceptance Model Instrument-Fast Form

The total mean for the measure was 2.44 with an SD of 0.52. The mean subscale scores across all 3 cycles of testing indicate that the teachers perceived JS Go to be easy to use and valuable but they had less favorable beliefs about using the app in the future (Table 2).

Discussion

Principal Findings

This study provides preliminary evidence regarding the benefits and usefulness of the JS Go mobile app as a CPD tool for teachers in accessing IECMHC practices in a timely, on-demand manner. The ECE teachers in this sample were all from centers in low-income tracts in Miami-Dade and Broward Counties, Florida. Focusing on the delivery of CPD in these centers is critical for minimizing disparities typically encountered by young children in these underserved neighborhoods [53]. To further address the diversity of the communities and where these centers reside to ensure the app's social validity, we used a CBPR approach based on 9 principles to develop and test JS Go [54].

By using a CBPR approach, we recognized and valued our community partners and the neighborhoods they represent, as important units of identity. Recognizing each community as a unit of analysis allows us to build on their strengths and resources by facilitating collaborative and equitable partnerships in all phases of our research. Through our ongoing and regular meetings with our community partners, we achieved a balance between knowledge generation and intervention about JS Go, mutually benefitting all partners. Including our community partners as co-investigators affirmed their value as trusted knowledgeable sources assuring our reach to ECE teachers and programs in their neighborhoods to understand their experiences with multiple determinants of health.

Similar to other CBPR research, we used an iterative approach to develop, refine, and test the JS Go app across multiple phases [15,29,30]. One aspect of the JS Go that appears to be particularly appealing to ECE teachers' is the access the app provides to IECMHC resources, which is in line with prior research on mobile technology using a CBPR approach [30]. One reason for this may be due to support for not only implementing IECMHC practices but also for successfully navigating the app itself as a CPD tool [30]. A strength of our research design and approach is the use of ECE teachers in the communities we strive to reach as end users to obtain valid and culturally relevant information about the usability of the app [30].

This study used an exploratory sequential mixed methods design to gather information from ECE teachers working in existing centers participating in the JS program to inform app development and evaluation across 3 phases. The iterative design allowed us to include our community partners' voices and perspectives during each phase to ensure the acceptability, feasibility, and usefulness of JS Go for ECE teachers and programs in underserved neighborhoods. Finally, we are ensuring a long-term commitment to sustainability, with pledged support from our community partners, who will support ongoing dissemination about JS Go as a germane CPD tool. This long-term commitment can lead to widespread use of JS Go across South Florida and nationwide.

Notably, the ECE teachers who participated in this study were from ECE programs in typically underresourced neighborhoods serving diverse children. To address the diversity of teachers and ECE programs in these neighborhoods, we used the Ecological Validity Model as a framework to guide the development and evaluation of JS Go, ensuring it is culturally relevant and generalizable to the communities in Miami-Dade and Broward Counties [31,32]. Given our attention to the cultural contexts of the neighborhoods in which JS Go will be implemented, it is culturally sensitive. For example, the teachers

XSL•FO
RenderX

can use the app in any of three languages, English, Spanish, and Haitian Creole, with teacher participants indicating their appreciation for the availability of app content in multiple languages.

Phases 2 and 3

ECE teachers had overall favorable perceptions of the JS Go mobile app, including their satisfaction and high perceived use of the app's interface. Feedback garnered from teachers about the app's interface during the qualitative interviews in phase 2 and quantitative real-world testing was shared with the app developer to aid in making adjustments to improve functionality. Users found the app to be easy to use, functional, beneficial, and useful for both current and new ECE teachers. Yet, they were less likely to rate it as a tool they would use in the future for telehealth consultation, suggesting some hesitancy with mobile apps as useful for this form of professional development. This was in contrast to their favorable report about using the app where they indicated they would use the app again in the future. One explanation for this finding could be that the teachers in this study were older, and research suggests that the age of end user influences their willingness to use mobile technology as a professional development tool while still valuing access to IECMHC content via an app [55].

A CPD tool that delivers individualized IECMHC content in a timely, on-demand manner can be a critical mechanism for training new ECE teachers while addressing their personal agency, confidence, and competence in their new role [12]. Addressing new teacher training, agency, confidence, and competence is vital, given the ongoing ECE workforce crisis exacerbated by the COVID-19 pandemic [52]. Thus, JS Go has the potential to be a viable tool for onboarding new teachers by giving them immediate access to the app with essential information on how to improve classroom quality. Additionally, teachers felt JS Go is an important mechanism to support the sustainability of JS practices by ECE teachers following receipt of IECMH consultation. Implementation science cites the sustainability of evidence-based practices as critical for scaling interventions [56]. The app would allow us to sustain and quadruple our reach across South Florida. As such, the JS Go app has the potential to increase equitable access for teachers to evidence-based IECMHC content helping to eliminate disparities typically experienced by young children in these communities. The JS Go app is a tool that could offer a paradigm shift in how teachers receive timely on-demand CPD tailored to their individual learning needs, increasing their autonomy and potentially leading to sustained use of JS practices.

This study contributes to a growing body of research regarding the viability of mobile apps for the delivery of CPD [19-21,57]. Research indicates ECE teachers integrate digital tools into their everyday practice to construct their learning using informal resources [17,18]. Mobile apps offer a tool for timely, on-demand learning, which can be a critical feature for ECE teachers working with diverse children in dynamic classrooms [16,18]. Additionally, mobile apps are a viable option for access to help and resources when in-person services are unavailable [58]. Mobile apps are also a CPD tool with the potential to support teachers' sustained use of new practices [57], critical for scaling evidence-based interventions [58]. Prior research shows that mobile apps are rated highly for their usability [59], which is in line with the findings in this study. However, there is a paucity of research on the use of mobile apps to support young children's social-emotional development, such as building relationships and communicating with families [22,23].

Next Steps

Future research by our team will pilot the efficacy of the JS Go app through a randomized controlled trial in six centers in South Florida. Each of the centers will be randomly assigned (via a random number table) to 1 of 2 arms, the intervention arm or the attention control arm. Both arms will be followed for 1 school year and then a 1-month follow-up. Centers in the JS Go arm will receive their IECMHC through the app in which they will participate in 12-24 mental health consultation sessions. Centers in the attention control arm will receive business-as-usual JS mental health consultation. We will implement the randomized controlled trial using CBPR using the 9 principles of Israel et al [54]. Our community advisory committee will continue to meet quarterly to review findings and recommendations and to support dissemination and sustainability. Meeting with partners will allow us to achieve knowledge generation that will mutually benefit the community and the research team by focusing on the local relevant public health issue regarding implementing IECMHC to support young children's social-emotional development and school readiness. Involving community partners in the research process will also allow us to build on each community's unique strengths and resources to attend to the social inequities ECE teachers and young children experience in these neighborhoods.

Limitations

There are several limitations to this study that should be noted. First, participants in qualitative interviews and quantitative testing were not identical, although both came from the same population of teachers currently participating in the JS program. Different populations pose a threat to the legitimation of making inferences from both qualitative and quantitative data [60,61]. However, using this sampling strategy is in line with the purpose of this research by allowing us to compare perceptions about the app from 2 separate groups of teachers from the same population. Additionally, the sample for the quantitative testing was small, limiting generalizability about the satisfaction and usefulness of the app with teachers who are not participating in the JS program, thus it is not possible to know the perceptions of these teachers [60]. Moreover, purposeful sampling of teachers may have introduced bias as MHC selected teachers based on their perceptions about teachers' ability to engage in the research. We attempted to minimize bias throughout the study in our methodology. However, it is possible that the response of teachers participating in phase 3 may not be entirely independent given their close relationship with their MHC. Finally given the different sample sizes of teachers in each cycle, with decreasing sample sizes, we elected to present total scores for the overall sample. The changes in scores across the 3 cycles are difficult to interpret given the differences in the sample sizes of teachers.

XSL•FO
RenderX

Conclusions

This paper describes the use of CBPR to develop the JS Go mobile app, a tool for delivering IECMHC content to ECE teachers in real time. The communities in this study face ongoing disparities in access to high-quality ECE service. Community stakeholders provided input at all stages of the app's development, including participation in data collection, allowing for contextualization of the data. This study used sequential exploratory mixed methods to evaluate the usability, acceptability, and feasibility of JS Go. Exploratory sequential mixed methods research is particularly salient for evaluating new approaches, as it allows for nuanced feedback via qualitative methods, and measurable input through quantitative data [60,61]. Specific to our population, the JS Go pilot data shows teacher satisfaction, given quick, easy, anytime-anywhere access to IECMHC information. To our knowledge, the JS Go app is the first mobile app that addresses the CPD needs of ECE teachers in implementing IECMHC content.

Acknowledgments

The authors wish to acknowledge the following for their contributions to this study: Jana Raphael and Maria Nieto for help with the analysis of the interviews in phase 2; Marya Meyer and Veronica Korbenfield of the Women's Fund of Miami Dade County for staunchly supporting JS Go, including providing valuable contributions to the app development during phase 1. The research was made available by funding from the University of Miami's Micah Batchelor Award for Excellence in Children's Health Research (grant BG005747 FY23), the Eunice Kennedy Shriver National Institute of Child Health and Human Development (grant R01HD105474), and the Children's Trust (grant 2212-7574). Generative artificial intelligence was not used in writing this manuscript.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

RN conceived of the study design and assisted with manuscript development. EH assisted with data collection, interpretation, and manuscript writing. JJ and CV assisted with statistical analyses and editing the final manuscript. EGG and KG assisted with data collection and interpretation. BC, EE, and TR assisted with app development and provided contextual knowledge of the interpretation of data. All authors reviewed and approved the manuscript for publication.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Mockups of JS Go (Jump Start on the Go) app.
[DOCX File , 979 KB - formative_v9i1e66714_app1.docx]

References

1. Hartman S, Winsler A, Manfra L. Behavior concerns among low-income, ethnically and linguistically diverse children in child care: importance for school readiness and kindergarten achievement. Early Educ Dev 2016;28(3):255-273. [doi: 10.1080/10409289.2016.1222121]
2. Gilliam WS, Shahar G. Preschool and child care expulsion and suspension: rates and predictors in one state. Infants Young Child 2006;19(3):228-245. [doi: 10.1097/00001163-200607000-00007]
3. Yoshikawa H, Weiland C, Brooks-Gunn J, Burchinal M, Espinosa L, Gormley W, et al. Investing in our Future: The Evidence Base on Preschool Education. Washington, DC: Society for Research in Child Development; 2013.
4. Schindler HS, Kholoptseva J, Oh SS, Yoshikawa H, Duncan GJ, Magnuson KA, et al. Maximizing the potential of early childhood education to prevent externalizing behavior problems: a meta-analysis. J Sch Psychol 2015;53(3):243-263 [FREE Full text] [doi: 10.1016/j.jsp.2015.04.001] [Medline: 26054817]
5. Harding J, Paulsell D. Improving access to early care and education: an equity-focused policy research agenda. Mathematica Policy Research. 2018. URL: https://www.mathematica.org/-/media/publications/pdfs/earlychildhood/2018/rwjf-ece-ib.pdf [accessed 2025-03-23]
6. Saitadze I, Lalayants M. Mechanisms that mitigate the effects of child poverty and improve children's cognitive and social-emotional development: a systematic review. Child Fam Soc Work 2020;26(3):289-308. [doi: 10.1111/cfs.12809]
7. Natale R, Kolomeyer E, Futterer J, Mahmoud FD, Schenker M, Robleto A, et al. Infant and early childhood mental health consultation in a diverse metropolitan area. Infant Ment Health J 2022;43(3):440-454. [doi: 10.1002/imhj.21983] [Medline: 35613369]

XSL•FO
RenderX

8. Trevedi P, deMonsabert J, Horen N. Infant and early childhood mental health consultation: Overview of research, best practices, and examples. SRI Education. 2021. URL: https://tinyurl.com/3mpnvkz7 [accessed 2025-03-20]
9. Gilliam WS, Maupin AN, Reyes CR. Early childhood mental health consultation: results of a statewide random-controlled evaluation. J Am Acad Child Adolesc Psychiatry 2016;55(9):754-761. [doi: 10.1016/j.jaac.2016.06.006] [Medline: 27566116]
10. Pears KC, Miao AJ, Burton M, Green BL, Hatfield B, Scheidt D. Survey on the Effects of COVID-19 on Oregon's Early Care & Education Workforce and Programs. Salem, OR: Oregon Early Learning Division and Early Learning Council; 2021.
11. Solomon J, Tresman S. A model for continued professional development: knowledge, belief and action. J Serv Educ 1999;25(2):307-319. [doi: 10.1080/13674589900200081]
12. Peleman B, Lazzari A, Budginaitė I, Siarova H, Hauari H, Peeters J, et al. Continuous professional development and ECEC quality: findings from a European systematic literature review. Euro J Educ 2018;53(1):9-22. [doi: 10.1111/ejed.12257]
13. Marinova I. 25+ Insightful App Usage Statistics to Know in 2024. Review 42. URL: https://review42.com/resources/app-usage-statistics/ [accessed 2025-03-20]
14. Barne D, Wadhwa D. Year in review: 2019 in 14 charts. World Bank Group. 2019. URL: https://www.worldbank.org/en/news/feature/2019/12/20/year-in-review-2019-in-charts [accessed 2025-03-20]
15. Doty JL, Brady SS, Monardez Popelka J, Rietveld L, Garcia-Huidobro D, Doty MJ, et al. Designing a mobile app to enhance parenting skills of latinx parents: a community-based participatory approach. JMIR Form Res 2020;4(1):e12618 [FREE Full text] [doi: 10.2196/12618] [Medline: 32012034]
16. Khaddage F, Müller W, Flintoff K. Advancing mobile learning in formal and informal settings via mobile app technology: Where to from here, and how? Educ Technol Soc 2016;19(3):16-26 [FREE Full text]
17. Hatzigianni M, Stephenson T, Harrison LJ, Waniganayake M, Li P, Barblett L, et al. The role of digital technologies in supporting quality improvement in Australian early childhood education and care settings. Int J Child Care Educ Policy 2023;17(1):5 [FREE Full text] [doi: 10.1186/s40723-023-00107-6] [Medline: 36778763]
18. Dwyer A, Jones C, Rosas L. What digital technology do early childhood educators use and what digital resources do they seek? Australas J Early Child 2019;44(1):91-105. [doi: 10.1177/1836939119841459]
19. Ek A, Sandborg J, Delisle Nyström C, Lindqvist A, Rutberg S, Löf M. Physical activity and mobile phone apps in the preschool age: perceptions of teachers and parents. JMIR Mhealth Uhealth 2019;7(4):e12512 [FREE Full text] [doi: 10.2196/12512] [Medline: 30994465]
20. Ekanayake T, Wishart J. Developing teachers' pedagogical practice in teaching science lessons with mobile phones. Technol Pedagogy Educ 2013;23(2):131-150. [doi: 10.1080/1475939x.2013.810366]
21. KyoungEun K. A pilot study on developing a mobile application to enhance teachers' creative teaching competence. Educ Adm Theory Pract 2024;30(7):58-63 [FREE Full text]
22. Berry K, Salter A, Morris R, James S, Bucci S. Assessing therapeutic alliance in the context of mHealth interventions for mental health problems: development of the mobile Agnew Relationship Measure (mARM) questionnaire. J Med Internet Res 2018;20(4):e90 [FREE Full text] [doi: 10.2196/jmir.8252] [Medline: 29674307]
23. Higgins A, Cherrington S. What's the story?: Exploring parent–teacher communication through ePortfolios. Australas J Early Child 2017;42(4):13-21. [doi: 10.23965/ajec.42.4.02]
24. Collins SE, Clifasefi SL, Stanton J, The Leap Advisory Board, Straits KJE, Gil-Kashiwabara E, et al. Community-based participatory research (CBPR): Towards equitable involvement of community in psychology research. Am Psychol 2018;73(7):884-898 [FREE Full text] [doi: 10.1037/amp0000167] [Medline: 29355352]
25. Israel B, Lantz P, McGranaghan R, Guzman J, Lichtenstein R, Rowe Z. Documentation and evaluation of CBPR partnerships: The use of in-depth interviews and closed-ended questionnaires. In: Methods in Community-Based Participatory Research for Health. West Sussex, UK: Wiley-Blackwell; 2012.
26. Israel BA, Coombe CM, Cheezum RR, Schulz AJ, McGranaghan RJ, Lichtenstein R, et al. Community-based participatory research: a capacity-building approach for policy advocacy aimed at eliminating health disparities. Am J Public Health 2010;100(11):2094-2102. [doi: 10.2105/AJPH.2009.170506] [Medline: 20864728]
27. Wallerstein N, Duran B. Community-based participatory research contributions to intervention research: the intersection of science and practice to improve health equity. Am J Public Health 2010;100:S40-546. [doi: 10.2105/AJPH.2009.184036] [Medline: 20147663]
28. Minkler M, Wallerstein N. Community-Based Participatory Research for Health: From Process to Outcomes. Hoboken, NJ: John Wiley & Sons; 2011.
29. Meherali S, Munro S, Puinean G, Salami B, Wong JP, Vandermorris A, et al. Co-designing a sexual health app with immigrant adolescents: protocol for a qualitative community-based participatory action research study. JMIR Res Protoc 2023;12:e45389 [FREE Full text] [doi: 10.2196/45389] [Medline: 36947124]
30. Merid B, Whitfield CO, Skolarus LE. Reflections on the values of community-based participatory research in supporting mobile health technology use. Circ Cardiovasc Qual Outcomes 2020;13(9):e007131 [FREE Full text] [doi: 10.1161/CIRCOUTCOMES.120.007131] [Medline: 32885680]
31. Bernal G, Jiménez-Chafey MI, Domenech Rodríguez MM. Cultural adaptation of treatments: a resource for considering culture in evidence-based practice. Prof Psychol Res Pract 2009;40(4):361-368. [doi: 10.1037/a0016401]

XSL•FO
RenderX

32. Bernal G, Bonilla J, Bellido C. Ecological validity and cultural sensitivity for outcome research: issues for the cultural adaptation and development of psychosocial treatments with Hispanics. J Abnorm Child Psychol 1995;23(1):67-82. [doi: 10.1007/BF01447045] [Medline: 7759675]
33. Pina AA, Polo AJ, Huey SJ. Evidence-based psychosocial interventions for ethnic minority youth: the 10-year update. J Clin Child Adolesc Psychol 2019;48(2):179-202. [doi: 10.1080/15374416.2019.1567350] [Medline: 30746965]
34. Glasgow RE, Battaglia C, McCreight M, Ayele R, Maw AM, Fort MP, et al. Use of the reach, effectiveness, adoption, implementation, and maintenance (RE-AIM) framework to guide iterative adaptations: Applications, lessons learned, and future directions. Front Health Serv 2022;2:959565 [FREE Full text] [doi: 10.3389/frhs.2022.959565] [Medline: 36925843]
35. Natale RA, Kolomeyer E, Robleto A, Jaffery Z, Spector R. Utilizing the RE-AIM framework to determine effectiveness of a preschool intervention program on social-emotional outcomes. Eval Program Plann 2020;79:101773. [doi: 10.1016/j.evalprogplan.2019.101773] [Medline: 31877485]
36. Nicholas JC, Ntoumanis N, Smith BJ, Quested E, Stamatakis E, Thøgersen-Ntoumani C. Development and feasibility of a mobile phone application designed to support physically inactive employees to increase walking. BMC Med Inform Decis Mak 2021;21(1):23 [FREE Full text] [doi: 10.1186/s12911-021-01391-3] [Medline: 33478495]
37. White BK, Martin A, White JA, Burns SK, Maycock BR, Giglia RC, et al. Theory-based design and development of a socially connected, gamified mobile app for men about breastfeeding (Milk Man). JMIR Mhealth Uhealth 2016;4(2):e81 [FREE Full text] [doi: 10.2196/mhealth.5652] [Medline: 27349756]
38. White BK, Burns SK, Giglia RC, Scott JA. Designing evaluation plans for health promotion mHealth interventions: a case study of the Milk Man mobile app. Health Promot J Austr 2016;27(3):198-203 [FREE Full text] [doi: 10.1071/HE16041] [Medline: 27784506]
39. Jaspers MW. A comparison of usability methods for testing interactive health technologies: methodological aspects and empirical evidence. Int J Med Inform 2009;78(5):340-353. [doi: 10.1016/j.ijmedinf.2008.10.002] [Medline: 19046928]
40. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009;42(2):377-381 [FREE Full text] [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
41. Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O'Neal L, REDCap Consortium. The REDCap consortium: Building an international community of software platform partners. J Biomed Inform 2019;95:103208 [FREE Full text] [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]
42. Zhou L, Bao J, Setiawan IMA, Saptono A, Parmanto B. The mHealth App Usability Questionnaire (MAUQ): Development and validation study. JMIR Mhealth Uhealth 2019;7(4):e11500 [FREE Full text] [doi: 10.2196/11500] [Medline: 30973342]
43. Chin WW, Johnson NA, Schwarz A. A fast form approach to measuring technology acceptance and other constructs. MIS Q 2008;32(4):687-703. [doi: 10.2307/25148867]
44. Hamilton AB. Qualitative Methods in Rapid Turn-Around Health Services Research. 2013. URL: http://www.hsrd.research.va.gov/for_researchers/cyber_seminars/archives/video_archive.cfm?SessionID=780 [accessed 2025-03-20]
45. Hamilton AB, Finley EP. Qualitative methods in implementation research: an introduction. Psychiatry Res 2019;280:112516 [FREE Full text] [doi: 10.1016/j.psychres.2019.112516] [Medline: 31437661]
46. Hamilton A. Rapid qualitative analysis: Updates/developments. US Department of Veteran Affairs. 2020. URL: https://www.hsrd.research.va.gov/for_researchers/cyber_seminars/archives/video_archive.cfm?SessionID=3846 [accessed 2025-03-20]
47. Kowalski CP, Nevedal AL, Finley EP, Young JP, Lewinski AA, Midboe AM, et al. Planning for and Assessing Rigor in Rapid Qualitative Analysis (PARRQA): a consensus-based framework for designing, conducting, and reporting. Implement Sci 2024;19(1):71 [FREE Full text] [doi: 10.1186/s13012-024-01397-1] [Medline: 39394597]
48. St George SM, Harkness AR, Rodriguez-Diaz CE, Weinstein ER, Pavia V, Hamilton AB. Applying rapid qualitative analysis for health equity: Lessons learned using "EARS" with Latino communities. Int J Qual Methods 2023;22 [FREE Full text] [doi: 10.1177/16094069231164938] [Medline: 38463016]
49. IBM SPSS Statistics for Windows, Version 26.0. Armonk, NY: IBM Corp; 2019.
50. Natale R, Agosto Y, Bulotsky Shearer RJ, St George SM, Jent J. Designing a virtual mental health consultation program to support and strengthen childcare centers impacted by COVID-19: a randomized controlled trial protocol. Contemp Clin Trials 2023;124:107022 [FREE Full text] [doi: 10.1016/j.cct.2022.107022] [Medline: 36442806]
51. American Public Health Association, American Academy of Pediatrics, National Resource Center for Health and Safety in Child Care and Early Education, United States Maternal and Child Health Bureau. Caring for Our Children: National Health and Safety Performance Standards: Guidelines for Early Care and Education Programs. Washington, DC: APHA Press; 2019.
52. Lachlan L, Kimmel L, Mizrav E, Holdheide L. Examining the Impact of COVID-19 on the Teaching Workforce. Arlington, VA: American Institutes for Research; 2020.
53. Conners-Burrow NA, Patrick T, Kyzer A, McKelvey L. A preliminary evaluation of REACH: Training early childhood teachers to support children‘s social and emotional development. Early Childhood Educ J 2016;45(2):187-199. [doi: 10.1007/s10643-016-0781-2]

XSL•FO
RenderX

54. Israel BA, Schulz AJ, Parker EA, Becker AB. Review of community-based research: assessing partnership approaches to improve public health. Annu Rev Public Health 1998;19:173-202. [doi: 10.1146/annurev.publhealth.19.1.173] [Medline: 9611617]
55. Ateş H, Garzón J. Drivers of teachers' intentions to use mobile applications to teach science. Educ Inf Technol 2022;27(2):2521-2542 [FREE Full text] [doi: 10.1007/s10639-021-10671-4] [Medline: 34462627]
56. Chambers DA, Glasgow RE, Stange KC. The dynamic sustainability framework: Addressing the paradox of sustainment amid ongoing change. Implement Sci 2013;8:117 [FREE Full text] [doi: 10.1186/1748-5908-8-117] [Medline: 24088228]
57. Dahri NA, Vighio MS, Bather JD, Arain AA. Factors influencing the acceptance of mobile collaborative learning for the continuous professional development of teachers. Sustainability 2021;13(23):13222. [doi: 10.3390/su132313222]
58. Yu W, He J, Gong Y. A systematic review of information technology in workplace learning. Psychol Sch 2023;60(12):4904-4929. [doi: 10.1002/pits.22901]
59. Høgsdal H, Kyrrestad H, Rye M, Kaiser S. Exploring adolescents' attitudes toward mental health apps: concurrent mixed methods study. JMIR Form Res 2024;8:e50222 [FREE Full text] [doi: 10.2196/50222] [Medline: 38224474]
60. Onwuegbuzie AJ, Collins KMT. A typology of mixed methods sampling designs in social science research. Qual Rep 2007;12(2):281-316. [doi: 10.46743/2160-3715/2007.1638]
61. Palinkas LA, Mendon SJ, Hamilton AB. Innovations in Mixed Methods Evaluations. Annu Rev Public Health 2019 Apr 01;40:423-442 [FREE Full text] [doi: 10.1146/annurev-publhealth-040218-044215] [Medline: 30633710]

Abbreviations

CBPR: community-based participatory research
CPD: continuous professional development
ECE: early care and education
FF-TAM: Technology Acceptance Model Instrument-Fast Form
IECMHC: infant and early childhood mental health consultation
JS: Jump Start
JS Go: Jump Start on the Go
MAUQ: mHealth App Usability Questionnaire
MHC: mental health consultant
mHealth: mobile health
REDCap: Research Electronic Data Capture

Edited by A Mavragani; submitted 20.09.24; peer-reviewed by JM Smith, J Arcand; comments to author 30.11.24; revised version received 20.12.24; accepted 02.03.25; published 02.06.25.

Please cite as:
Natale R, Howe E, Velasquez C, Guzman Garcia E, Granja K, Caceres B, Erban E, Ramirez T, Jent J
Co-Designing an Infant Early Childhood Mental Health Mobile App for Early Childhood Education Teachers' Professional Development: Community-Based Participatory Research Approach
JMIR Form Res 2025;9:e66714
URL: https://formative.jmir.org/2025/1/e66714
doi:10.2196/66714
PMID:

©Ruby Natale, Elizabeth Howe, Carolina Velasquez, Emperatiz Guzman Garcia, Karen Granja, Bianca Caceres, Elizabeth Erban, Tania Ramirez, Jason Jent. Originally published in JMIR Formative Research (https://formative.jmir.org), 02.06.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Idiographic Lapse Prediction With State Space Modeling: Algorithm Development and Validation Study

Eric Pulick[1], MS; John Curtin[2], PhD; Yonatan Mintz[1], PhD

[1]Department of Industrial and Systems Engineering, College of Engineering, University of Wisconsin–Madison, Madison, WI, United States

[2]Department of Psychology, College of Letters & Science, University of Wisconsin–Madison, Madison, WI, United States

Corresponding Author:
Eric Pulick, MS
Department of Industrial and Systems Engineering
College of Engineering
University of Wisconsin–Madison
1513 University Avenue
Madison, WI, 53706
United States
Phone: 1 6082622686
Email: pulick@wisc.edu

Abstract

Background: Many mental health conditions (eg, substance use or panic disorders) involve long-term patient assessment and treatment. Growing evidence suggests that the progression and presentation of these conditions may be highly individualized. Digital sensing and predictive modeling can augment scarce clinician resources to expand and personalize patient care. We discuss techniques to process patient data into risk predictions, for instance, the lapse risk for a patient with alcohol use disorder (AUD). Of particular interest are idiographic approaches that fit personalized models to each patient.

Objective: This study bridges 2 active research areas in mental health: risk prediction and time-series idiographic modeling. Existing work in risk prediction has focused on machine learning (ML) classifier approaches, typically trained at the population level. In contrast, psychological explanatory modeling has relied on idiographic time-series techniques. We propose state space modeling, an idiographic time-series modeling framework, as an alternative to ML classifiers for patient risk prediction.

Methods: We used a 3-month observational study of participants (N=148) in early recovery from AUD. Using once-daily ecological momentary assessment (EMA), we trained idiographic state space models (SSMs) and compared their predictive performance to logistic regression and gradient-boosted ML classifiers. Performance was evaluated using the area under the receiver operating characteristic curve (AUROC) for 3 prediction tasks: same-day lapse, lapse within 3 days, and lapse within 7 days. To mimic real-world use, we evaluated changes in AUROC when models were given access to increasing amounts of a participant's EMA data (15, 30, 45, 60, and 75 days). We used Bayesian hierarchical modeling to compare SSMs to the benchmark ML techniques, specifically analyzing posterior estimates of mean model AUROC.

Results: Posterior estimates strongly suggested that SSMs had the best mean AUROC performance in all 3 prediction tasks with ≥30 days of participant EMA data. With 15 days of data, results varied by task. Median posterior probabilities that SSMs had the best performance with ≥30 days of participant data for same-day lapse, lapse within 3 days, and lapse within 7 days were 0.997 (IQR 0.877-0.999), 0.999 (IQR 0.992-0.999), and 0.998 (IQR 0.955-0.999), respectively. With 15 days of data, these median posterior probabilities were 0.732, <0.001, and <0.001, respectively.

Conclusions: The study findings suggest that SSMs may be a compelling alternative to traditional ML approaches for risk prediction. SSMs support idiographic model fitting, even for rare outcomes, and can offer better predictive performance than existing ML approaches. Further, SSMs estimate a model for a patient's time-series behavior, making them ideal for stepping beyond risk prediction to frameworks for optimal treatment selection (eg, administered using a digital therapeutic platform). Although AUD was used as a case study, this SSM framework can be readily applied to risk prediction tasks for other mental health conditions.

(JMIR Form Res 2025;9:e73265) doi:10.2196/73265

XSL•FO
RenderX

KEYWORDS

mental health; digital health; alcohol use disorder; substance use disorder; mobile health; mHealth; personalized medicine; digital therapeutics

Introduction

Data-driven psychological modeling methods can help researchers leverage increasingly available patient data to better understand and treat pressing mental health conditions [1-3]. In parallel, digital interventions introduce new, low-burden treatment pathways to patients [4,5]. Together, these advancements give providers exciting new tools for low-barrier, personalized risk monitoring and treatment, which are core elements of continuing care for chronic mental health conditions [6]. Continuing care is an important recovery support for chronic conditions, such as alcohol use disorder (AUD) and other substance use disorders (SUDs) [7,8], but is underavailable and underutilized [9]. Novel approaches are needed to close this gap.

Advances in personal sensing have spurred new modeling approaches by empowering researchers to collect rich longitudinal patient datasets [10]. Ubiquitous mobile devices, such as smartphones and wearables, increase both the types and quantities of available measurements. Diary-type methods, such as ecological momentary assessment (EMA) [11], have long been an important way to gather *in situ* measurements of a patient's experience or mental state, and this type of active data collection has been made easier by the prevalence of mobile devices [12]. In addition to actively queried measurements like EMA, researchers can also use mobile devices to gather patient data passively, such as physiological measurements or a patient's GPS location. The explosion of data availability has led researchers to examine mental health conditions in more quantitative ways through 2 (often parallel) perspectives: explanation-focused modeling and prediction-focused modeling.

Explanation-focused modeling has largely turned to time-series methods to better understand the dynamics of different mental health conditions [13], sometimes explicitly focused on identifying causal mechanisms [14]. This line of research has generated a rich set of modeling approaches, typically using various forms of network analysis or structural equation modeling [15-23]. These approaches create mathematical models that explain variations in individual patient behavior over time, often implemented as multilevel models that separate group-level and person-specific model effects. Implicit in this separation is the notion of an idiographic-nomothetic spectrum [24-27]. While some behaviors may be well-described by nomothetic principles that hold *on average* across a population, there is no guarantee that those principles will apply well to a particular individual. Importantly, there is evidence suggesting that the presentation and progression of many mental health conditions are meaningfully individualized (eg, in mood dynamics [28,29], suicidal thoughts [30], depression [31], and panic disorder [32]). These examples emphasize the importance of personalized modeling approaches that can capture person-specific heterogeneity. Note that here we use "idiographic" synonymously with "personalized" and "person-specific," describing when separate models are fit to each individual, even if group-level information is incorporated into the patient's model in some way. Under a strict definition of idiographic, such a model might be better described as "idiothetic" [33]. For our purposes, the important distinction is whether each participant has a separate model.

In contrast to explanatory modeling, prediction-focused modeling emphasizes building models that predict behavior well, even if doing so does not necessarily give insights into the underlying behavior-generating process [34]. For instance, many mental health conditions exhibit high-acuity binary outcomes (eg, lapse in individuals with SUDs, suicidal ideation, panic attacks, binge eating, etc) that fit naturally into a machine learning (ML) prediction framework. ML classifier methods like logistic regression (LR) or gradient boosting models have used digital sensing data to successfully predict, for example, patient risk for AUD lapse or binge eating [6,35-38]. This type of risk assessment is an important component of successful patient intervention. In situations where treatments may be costly or scarce, such as appointments with clinicians, accurate patient risk assessment could help direct treatment resources to high-risk individuals. In settings like digital therapeutics, where intervention costs and resource scarcity are less relevant, accurate risk assessment could inform the urgency, intensity, or type of a just-in-time adaptive intervention [4].

However, ML classification approaches have 2 important drawbacks with respect to idiographic risk prediction. First, it is difficult to fit classifier models idiographically when outcomes are rare, so they are often fit at the population level instead. In the most extreme case, an exclusively person-specific model cannot be fit at all until the positive outcome label (eg, lapse) is observed. In practice, a person-specific classifier model might not show adequate performance until it can train on a sufficient number of positive outcomes, at which point it may no longer be relevant (eg, a person with SUD would likely have already relapsed). These concerns can sometimes be alleviated by modifications to incorporate prior or group-level information, such as Bayesian estimation or multilevel modeling, depending on the classifier approach being used.

The second drawback of classifier methods, however, is entirely structural: they are not time-series methods. Classifier models are trained only to make a prediction for a particular outcome, such as the probability of a patient lapsing today, and thus, they do not give insights into how a patient's behavior will evolve over time. If a clinician is also interested in a patient's lapse risk for other time windows (eg, tomorrow or within the next week), a separate classifier model must be trained for every such circumstance. While this type of predictive capability is clearly important, our broader goal is to use quantitative modeling to inform personalized treatment recommendations. This involves understanding "*what* treatment, by *whom*, is most effective for *this* individual with *that* specific problem under *which* set of circumstances" [39]. Models that output a single predictive risk score provide limited insights into these

XSL•FO
RenderX

questions. A dynamic model describing how relevant aspects of a patient's state of mind evolve over time would be much more helpful. Time-series models directly capture these behavioral dynamics as part of model fitting *and* can be used to produce risk scores (eg, by simulating trajectories of patient behavior, including responses to different treatment approaches). Time-series models are thus better integrated into intervention planning frameworks to optimize long-run patient outcomes.

We advocate combining these explanation- and prediction-focused research areas by using idiographic, time-series models for prediction. Specifically, we propose the use of an approach called state space modeling that supports fitting person-specific time-series models of patient behavior. State space models (SSMs) originate from the field of control engineering and share much in common with structural equation modeling and network analysis. We believe that SSMs are particularly well-suited for idiographic modeling as they focus specifically on modeling the behavior of a single subject (for a deeper discussion of commonalities and differences between structural equation models and SSMs, see [40]).

To demonstrate how SSMs can be applied to personalized mental health prediction, we use the setting of predicting lapse for individuals with AUD as a case study. We provide an overview of state space modeling and present a framework for fitting person-specific SSMs for AUD lapse prediction with longitudinal EMA observations. We compare the predictive performance of SSMs to that of commonly used ML classifiers and demonstrate that this idiographic, time-series approach can have comparable or better predictive performance than ML approaches used in the literature. As part of this analysis, we explicitly characterize how predictive performance changes when more data are available for each study participant. This is an important practical consideration for the real-world use of personalized models. Further, we discuss how the SSMs we propose are readily integrated into advanced personalized intervention frameworks. Lastly, we discuss how the SSM frameworks presented here can be similarly applied to other mental health conditions beyond AUD, such as anxiety and panic disorders, depression, and suicidal ideation.

Methods

Ethical Considerations

This study conducted analyses to address the second aim of a larger grant-funded parent project (R01 AA024391). This parent effort collected the study data between 2017 and 2019. The University of Wisconsin-Madison Institutional Review Board approved all procedures (IRB #2015-0780). All study participants provided written informed consent prior to their participation. Study consent documentation is provided in a persistent repository (see [41]) on Open Science Framework (OSF). To protect participant data privacy, the parent study that collected these data received a Certificate of Confidentiality from the National Institutes of Health. Further, no identification of individual data is possible from this manuscript or supplementary materials. We do not present individual patient data; all data are presented for the complete population using a deidentified version of the original dataset. Participants were compensated up to approximately US $800, depending on measure adherence (see Section S2 in Multimedia Appendix 1).

Transparency and Openness

We adhere to research transparency principles that are crucial for robust and replicable science. Data, analysis scripts, annotated results, measures, preregistration of the hypotheses and analysis plan (see [42]), and other study materials are publicly available at our OSF repository (see [41]). We report how the sample size, all data exclusions, all manipulations or procedures, and all study measures were determined. Measures and procedures that were collected as part of the parent project but were not relevant to this study have been described in [6,43]. We also provide a transparency report [44] in Multimedia Appendix 2 and a biomedical ML checklist [45] in Multimedia Appendix 3.

Recruitment

We analyzed 148 study participants in early remission from AUD (1-8 weeks of abstinence). Participants were recruited from Madison, Wisconsin and surrounding communities through print and targeted digital advertisements and partnerships with treatment centers. The requirements for inclusion are provided in Textbox 1. For additional information regarding recruitment, please refer to the study design documentation available on OSF (see [41]) and in the parent study [6].

Textbox 1. Criteria for inclusion in the parent project.

Inclusion criteria

1. Age ≥18 years
2. Ability to write and read English
3. Presence of at least moderate alcohol use disorder (≥4 symptoms from the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, measured with a self-report survey during a screening visit)
4. Abstinence from alcohol for at least 1 week but no longer than 2 months
5. Willingness to use a single smartphone (their personal phone or one provided to them) while enrolled in the study

Alcohol abstinence was used as a behavioral indicator of a commitment to recovery. Although recovery may be possible without complete abstinence, clinicians typically recommend abstinence for patients who present with moderate or more

XSL•FO
RenderX

severe AUD. Individuals were excluded if they exhibited severe symptoms of psychosis or paranoia, defined as scores greater than 2.2 or 2.8, respectively, on the psychosis or paranoia scales in the Symptom Checklist-90 [46]. The sample size for the parent project was determined based on power analyses for its broader set of aims. The authors of the parent project determined preliminary eligibility by phone and completed screening and obtained written informed consent at a subsequent in-person screening visit. The study period for each participant lasted up to 3 months, beginning with an instructional intake visit, which was followed by in-person visits every 30 days.

Measures

The study collected baseline demographics and participant characteristics relevant to alcohol use at the screening and intake visits. Participants completed brief EMAs (7-10 questions) 4 times each day following prompts sent by text messages. These text messages included a link to a Qualtrics survey that was optimized for completion on their smartphones. All EMAs contained a common set of 7 questions asking about any alcohol use that had not yet been reported, current affective state (pleasantness and arousal), greatest urge to drink alcohol since the last EMA, any pleasant or positive events and any hassles or stressful events that occurred since the last EMA, and any exposure to risky situations (ie, people, places, or things) since the last EMA. The first EMA each day (the "morning EMA") asked 3 additional questions about how likely participants were to encounter a risky situation, encounter a stressful event, and drink alcohol in the upcoming week. A copy of the morning EMA is available on OSF [41].

Although the parent study collected up to 4 EMAs per participant each day over the 3-month study period, we believe that 1 EMA per day represents a more realistic sensing burden for real-world implementation, especially for situations where patients might be followed for longer time periods [43,47]. Accordingly, we chose to compare models built using only 1 EMA response from participants each day, specifically the morning EMA. We chose the morning EMA as it contained the most informative set of questions and occurred at a similar time near the beginning of each participant's day. The first EMA question was used to report past alcohol consumption, leaving 9 Likert-style questions for analysis. Note that all 4 daily EMAs were used, along with in-person follow-up visits, to collect self-reported lapse (ie, outcome label) information, but only the Likert-style questions from the morning EMA were used as inputs in our modeling. We do not believe this has a practical impact on our findings, as a future implementation could equivalently collect the same lapse information using only 1 EMA per day. The original study [6] had 151 participants, but we excluded 3 participants who responded to fewer than 25% of the morning EMAs during their study period, leaving 148 participants for analysis.

Labels and Prediction Tasks

For each participant, we constructed binary labels indicating if the participant drank alcohol each day. Days began and ended at 4 AM. If the participant self-reported alcohol use during this interval, the day was labeled as 1; otherwise, it was labeled as 0. As all study participants stated a goal of abstinence, any use of alcohol was inconsistent with their goals. For brevity, we refer to any such goal-inconsistent use as a lapse. We established the 4 AM start and end time for each day to align better with the waking hours between periods of sleep. For instance, if a participant's study began on January 1 and they experienced a lapse at 2 AM on January 2, this lapse would be included as part of day 1 in their study. For 2 participants who typically woke up before 4 AM, study days were defined to begin and end at 3:30 AM.

To provide a richer view of how well each method assesses participant risk, we compared the predictive performance of the methods on 3 lapse prediction tasks ranging from same-day to week-long windows, as given in Textbox 2. Window-style predictions were only generated for days where the window was contained entirely within the participant's study period (ie, for a participant with a 90-day study, 7-day window predictions were only made up to day 83).

Textbox 2. Prediction tasks.

Given a participant's ecological momentary assessment data from day 1 to day t and lapse data from day 1 to day t–1, our models made lapse predictions for 3 intervals of increasing length:

1. Predict the risk of lapse on day t (same-day lapse prediction)
2. Predict the risk of lapse occurring on any day between day t and day t+3, inclusive (lapse within 3 days)
3. Predict the risk of lapse occurring on any day between day t and day t+7, inclusive (lapse within 7 days)

State Space Modeling

SSMs are equations that describe how a system changes over time. This framework comes from the field of control engineering and was critical, for instance, in the navigation and control of rockets in the Apollo space program [48,49]. Since then, the same modeling framework has proven valuable in many domains, such as ecology (eg, modeling wildlife population dynamics [50]) and psychology (eg, modeling affective mood dynamics [51]). State space modeling has many commonalities with structural equation modeling [40], often used for explanatory psychological modeling, but typically focuses on modeling the dynamics of a single subject.

The "state" of a system is something that usually cannot be observed directly. For a rocket, this might be its exact position or speed. In ecology, it could refer to the true size of an animal population. For more abstract systems, such as those in psychology, it might describe the intensity of a particular emotion or some other mental construct. The core insight of SSMs is the separation of a process model from an observation model. The process model governs how that system state changes over time, while the observation model describes how

XSL•FO
RenderX

we can gather, often noisy, measurements of the system state. For a physical system like a rocket, the process model would be the relevant laws of physics, while the observation model would come from the noisy sensors we use to measure the system. For psychological systems, the correct, causal process model may not be known, and discovering or defining such a model is not the focus of this study. Instead, we seek to demonstrate that simple models can be good approximations for complex psychological processes and can thus be useful for patient risk assessment.

In our state space modeling, the unobserved state (also called the hidden or latent state) can be thought of as representing a set of mental constructs, relevant to lapse, that evolve from one day to the next. We cannot observe these constructs directly and instead rely on passively or actively sensed data from the participants. In this study, we observed the participants' responses to 9 EMA questions each morning and their lapse behavior each day. For each study participant, we defined 2 key equations: the observation equation and the transition (or process) equation.

The observation equation mathematically defines the relationship between the quantities we cannot observe (the participant's mental states) and those that we can observe (how they respond to EMAs and whether they lapse). The transition equation describes how the hidden state evolves over time. In our modeling, time steps are discretized as days of the study period. Figure 1 provides a visualization of the modeling framework. Note that this representation is a more general formulation of other autoregressive models, such as vector autoregression, that model observed quantities as functions of previously observed quantities. SSMs readily handle missing data without the need for deletion or imputation, as measured quantities are used for inference about the hidden state and not directly to predict future observations.

Figure 1. Visual representation of state space modeling for this alcohol use disorder study, following the time series for a single participant from day 1 to the end of their study period (day *T*). This modeling framework explains observable quantities (ie, a participant's lapse decision and ecological momentary assessment [EMA] responses each day) in terms of hidden states (ie, the participant's state of mind) that change over time. The observation equation explains how the participant's hidden state produces EMA responses and lapses each day, while the transition equation explains how the hidden state evolves from one day to the next.

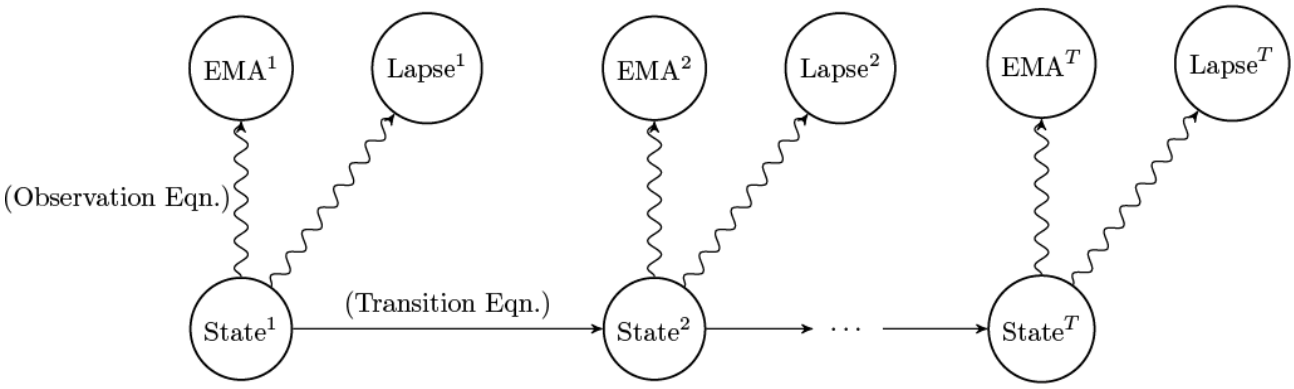

Many different equation structures are possible under the broad framework of SSMs. More complex equation structures may better describe the underlying processes but can become more difficult or even intractable to fit. We use a simple structure for the observation and transition equations: linear dynamics with b noise. This formulation is a common approximation and has convenient mathematical properties. Concretely, let y_t be the response to a particular EMA question on day t and x_t be the hidden state on that day. This gives the following equations:

$$y_t = ax_t + c + v_t \text{ (observation equation)} \quad \textbf{(1)}$$

$$x_{t+1} = bx_t + d + w_t \text{ (transition equation)} \quad \textbf{(2)}$$

where v_t and w_t represent independent, zero-mean Gaussian noise at each time step. Concretely, this structure states that on day t, the response to an EMA question (or the chance of lapse) is a linear function of the participant's current hidden mental state plus random noise. The participant's mental state the following day, t+1, is a separate linear function of their current mental state plus random noise. More generally, the observable quantities (EMA responses and lapse behavior) are expressions of the hidden state (mental constructs relevant to lapse), which evolve noisily over time. Model fitting identifies the joint "story" (ie, the observation and transition equations) that best explains the individual's observed behavior over time. Predictions are straightforward to obtain by simulation, for instance, by "playing forward" the transition equation over the desired horizon and passing the state quantities through the observation equation. For additional details, see Section S3.5 in Multimedia Appendix 1.

Note that this approach is inherently personalized, as the parameters *a*, *b*, *c*, *d*, and noise variance are estimated separately for each participant rather than for the overall population. All participants have the same linear-Gaussian model structure, but the parameters are estimated individually to create a personalized model that best explains their observed behavior. Importantly, since these fitted equations directly use the observed EMA responses and lapse values, this approach does not require the manual creation of features. We detail the vector forms of these equations used to model all 9 EMA questions and lapse in Section S3 in Multimedia Appendix 1.

Fitting SSMs of this form is computationally inexpensive, so it is feasible to improve parameter estimates over time by refitting the model for each participant with each new day of available data. A common approach for fitting such models is maximum likelihood estimation (MLE) [52]. MLE identifies the parameters that best explain the participant's behavior, but this approach

XSL•FO
RenderX

is problematic for model fitting with rare outcomes. If a participant has never been observed to lapse, the best fitting parameters will be those that suggest they will not lapse in the future either. To address this, our modeling uses a Bayesian fitting approach called maximum *a posteriori* estimation [53], which incorporates prior distributions for the model parameters. The resultant fit is one that best explains both the participant's observed data and the parameter priors, leading to models that better anticipate lapse risk even when the participant has never before lapsed. When limited data for the participant are available, the fit relies heavily on the prior distributions for the model parameters. The influence of these prior distributions diminishes as more data become available and the model fit primarily reflects the participant's data. The prior distributions used in this paper were found by fitting models using MLE on the full trajectories of other participants within this studied cohort, gathering the set of estimates for each parameter, and fitting well-behaved distributions to those parameter sets. For a detailed explanation, see Section S3.6 in Multimedia Appendix 1.

Note that this approach combines idiographic and nomothetic principles. For instance, fitting a model to each participant is fundamentally idiographic, but this process incorporates nomothetic principles as well. First, this approach assumes that the behavior of all participants can be well-approximated by a common linear-Gaussian model structure built from observations of the same set of EMA questions. Further, we incorporate information from other participants through the parameter prior distributions. These priors are built from the results of MLE model fits for other participants. This means that we begin by assuming that participants behave similarly (ie, have similar parameter values) but personalize these model fits as data for each participant become available. We provide thorough descriptions of SSM fitting, implementation, prediction, creation of priors, and limitations in Section S3 in Multimedia Appendix 1.

Benchmark ML Methods

We compared SSM performance to that of LR and gradient boosted classifiers. LR was implemented using Scikit-learn [54], and gradient boosting was implemented with extreme gradient boosting (XGB) [55]. Both LR and XGB used a common feature set given in Textbox 3 and the lapse labels described previously.

Textbox 3. Features used in the logistic regression and extreme gradient boosting models.

Features

- Day of the week: Expressed as one-hot encoding for each day of the week. One of the 7 Booleans was dropped for logistic regression to avoid collinearity issues.
- Most recent ecological momentary assessment (EMA) responses: The ordinal responses to the most recent available morning EMA, scaled to [0,1]. As noted previously, the first morning EMA question dealt with reporting previous lapses, so these features correspond to the 9 ordinal responses provided from EMA questions 2-10.
- Short-run mean of EMA responses: The average of the participant's responses in the 3 most recent available morning EMAs (9 questions), scaled to [0,1]. Fewer EMAs were used to calculate these features when 3 EMAs were not available.
- Long-run mean of EMA responses: The average of the participant's responses to all available morning EMAs in their study period up to the current day (9 questions), scaled to [0,1]. For the data availability analysis, this average was calculated using smaller intervals (eg, the most recent 15 days of data).
- Recent lapses: Three Booleans indicating whether the individual had lapsed on any day in the last 1, 3, and 5 days.

Given the difficulty of fitting classification models for rare outcomes, these models were fit at the population level. On one hand, this is a fundamentally nomothetic approach as it creates 1 model mapping features to risk for the entire population. However, this approach allows for some measure of personalization through the model's features. For instance, by including calculations of short- and long-run mean values for each EMA question as features, the model can capture risk associated with deviations from an individual's baseline response level. For additional technical details, see Section S4 in Multimedia Appendix 1.

Fitting Procedures

The SSMs in this study are time-series models trained for each participant, meaning they have inherently different fitting procedures than the population-trained benchmark ML classifiers. This section describes the idea of data availability used to assess the impact of model personalization and explains the fitting procedures used for each method.

Data Availability

A model's ability to personalize to an individual becomes relevant as more of that individual's data become available. The SSMs are explicitly personalized, with a separate model fit to each participant. While the benchmark ML classifiers are fit at the population level, they allow for limited personalization through feature construction. We evaluated the impact of model personalization on performance by varying what we call "data availability," which is the number of days of EMA data available to a model from the test participant. For the benchmark ML models, this meant varying the number of days of EMA data available for feature construction. For the SSMs, which were trained separately for each individual, this meant varying the number of days of data used to train the model. To make this notion more concrete, we provide examples in the training procedures below for each model type.

LR and XGB

LR and XGB were trained and evaluated using 15 repeats of 5-fold nested cross-validation. In each repeat, the dataset was

XSL•FO
RenderX

divided into 5 participant-stratified folds. We relied on scikit-learn's StratifiedGroupKFold function [54]. This function forms folds such that the overall proportion of positive labels is similar across folds, and all data for each participant are contained in exactly 1 fold. Each fold was used once for testing, with the remaining 4 folds used for hyperparameter tuning and model training. We tuned model hyperparameters by grid search, using 5-fold participant-stratified cross-validation nested within the training folds. Additional details regarding hyperparameter tuning can be found in Section S4.2 in Multimedia Appendix 1. The model was refit on the training folds using the best performing hyperparameters and tested on the held-out fold. To assess how the performance of LR and XGB changed with respect to data availability, we created separate featurizations for the test fold using 15, 30, 45, 60, and 75 days of data availability. For example, the 15-day test set included only study days from day 15 onward, and its features were constructed with only the most recent 15 days of data (ie, the long-run average EMA features for day 70 included only data from study days 56-70). In contrast, the 45-day featurization for day 70 used data from days 26-70 to calculate long-run average values. These alternative featurizations were used only at test time and not for model training. Models were trained using the full study periods of all participants in the training folds.

Approach Involving SSMs

The SSMs were fitted using the same 15 repeated 5-fold cross-validation splits as the LR and XGB models. However, SSMs differ from population-trained models in that they are idiographic, with a different model trained for each participant. For each iteration within a cross-validation repeat, fitting proceeded as follows. Data from the participants in the 4 training folds were used to create prior distributions for the SSM parameters. We then fit an idiographic model to each participant in the test fold using their EMA and lapse data. To assess how the SSMs performed with respect to data availability and provide a direct comparison to LR or XGB predictions, we used rolling training periods of fixed length (15, 30, 45, 60, and 75 days). Figure 2 provides a diagram illustrating this fitting procedure. This fitting process created a set of predictions from the SSMs that was directly comparable to the LR or XGB predictions. While the fitting procedures differed between models, they had access to the same data and made predictions for the same set of tasks each day. For example, consider the 15-day case. Each model type produced predictions for each participant's study day from day 15 onward. For the SSMs, predictions were made with models trained on the most recent 15 days of EMA and lapse data. For LR and XGB, the models were trained on participants from the 4 training folds and tested using the 15-day featurizations of EMA and lapse data.

Figure 2. Illustration of the data availability fitting procedure used for state space models in this study. Data availability refers to the amount of the participant's data that is available to use when training the model. As an example, this graphic considers 15 days of data availability, leading to models trained with data from days 1-15, 2-16, 3-17, and so on. The red bracketed case describes the scenario where a model is trained on ecological momentary assessment data from days 4-18 and lapse data from days 4-17. The trained model is used to make a same-day lapse prediction for day 18 and 2 window-style predictions for day 18 (ie, a lapse between days 18-21 and 18-25). This fitting procedure is repeated for each participant for different data availabilities (15, 30, 45, 60, and 75 days).

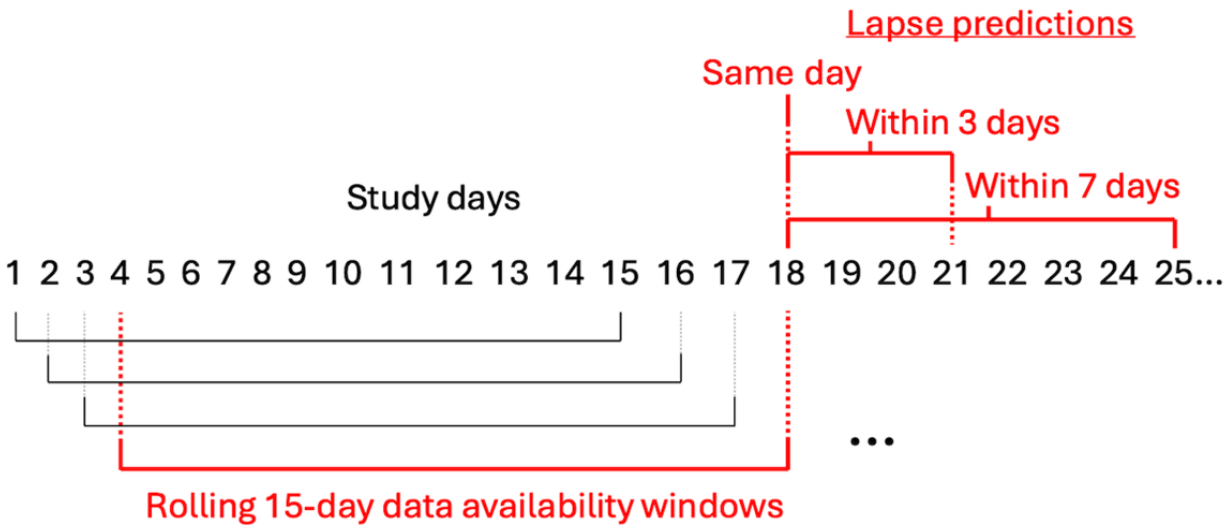

Statistical Analysis of Model Comparisons

Although fitting procedures differed between the models, they generated predictions that could be used to directly compare model performance. We gathered all predictions made for each test fold to create a receiver operating characteristic curve and summarized performance using the area under the receiver operating characteristic curve (AUROC) [56]. We focused our analysis on AUROC because we plan to directly use the continuous probability scores output by the models to inform treatment strategies and not to make a classification decision involving a particular decision threshold. However, other performance metrics, for instance, the area under the precision-recall curve (AUPRC), provide additional value in interpreting model performance [57]. For all assessments presented in terms of AUROC, we provide comparable assessments by AUPRC in Section S6.3 in Multimedia Appendix 1.

XSL•FO
RenderX

We compared the models using Bayesian hierarchical modeling [58,59] to quantify uncertainty in different aspects of the experiments, namely separating uncertainty in mean model AUROC performance from uncertainty in the cross-validation process. In this approach, the AUROC performance of a model on cross-validation repeat i and test fold j was modeled using a generalized linear model with a logit link function as follows:

where b_i is a cross-validation repeat intercept, c_{ij} is a fold-within-repeat intercept, β is the model type's mean AUROC performance, and ε_{ij} is zero-mean Gaussian noise. This structure implies that each method has logit-transformed Gaussian mean AUROC performance, but that the mean is shifted based on the specific repeat and fold of the cross-validation process. The repeat intercept implies that, across all models, the test folds within a given repeat are expected to show similar AUROC performance since the training folds share large amounts of training data. The fold-within-repeat intercept reflects that common AUROC trends are expected for each test fold (ie, easier or harder across all methods). Analysis was performed using the TidyPosterior package [60]. Additional details are provided in Section S5 in Multimedia Appendix 1.

Results

Study Cohort Demographics

The collected demographic information for the study cohort is provided in Table 1.

Table 1. Demographic information of the alcohol use disorder study participants considered in our modeling (N=148).

Characteristic	Value
Age (years), mean (SD)	40.8 (11.9)
Sex, n (%)	
Male	76 (51.4)
Female	72 (48.6)
Race, n (%)	
American Indian/Alaska Native	3 (2.0)
Asian	2 (1.4)
Black/African American	7 (4.7)
White/Caucasian	129 (87.2)
Other/Multiracial	7 (4.7)
Hispanic, Latino, or Spanish origin, n (%)	
Yes	4 (2.7)
No	144 (97.3)
Personal income (US$), mean (SD)	34,249 (31,517)
Number of quit attempts, mean (SD)	8.6 (31.0)
DSM-5[a] alcohol use disorder symptom count, mean (SD)	8.8 (1.8)

[a]DSM-5: Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition.

EMA Adherence and Lapses

We measured EMA adherence as the proportion of study days that each participant completed a morning EMA. The distribution of adherence across the study population is shown in Figure 3. Lapses were relatively rare, with approximately 7.5% of study days containing lapses across the study population. The distribution of total lapses by participants in the study period is shown in Figure 4.

XSL•FO
RenderX

Figure 3. Histogram describing the distribution of ecological momentary assessment (EMA) adherence for the complete study cohort. We characterize adherence as the proportion of study days that each participant responded to the EMA after being prompted each morning. The median adherence was approximately 0.86.

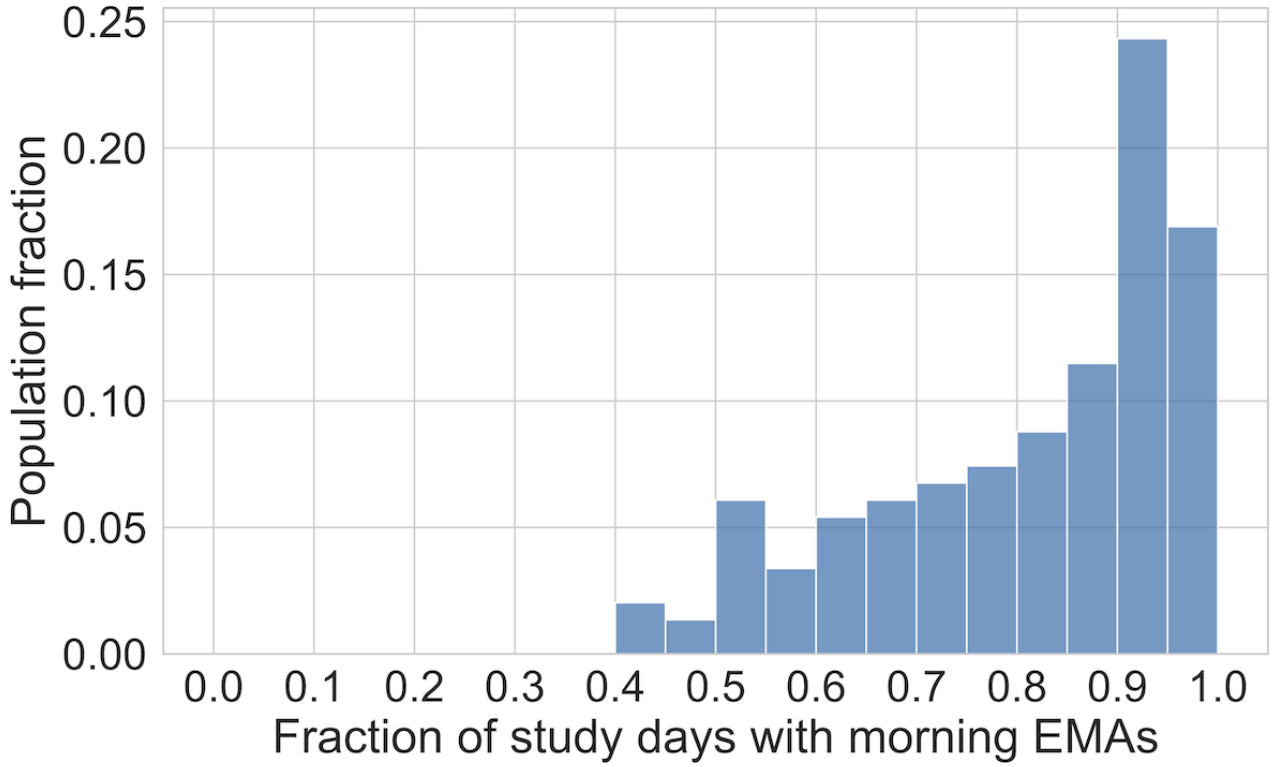

Figure 4. Histogram describing the distribution of total lapses during the study for the study cohort. Specifically, this histogram summarizes the fraction of the study cohort that reported each number of total lapses for the complete study period. Note that nearly 45% of participants reported no lapses during the study. The median lapse count for a participant was 1.

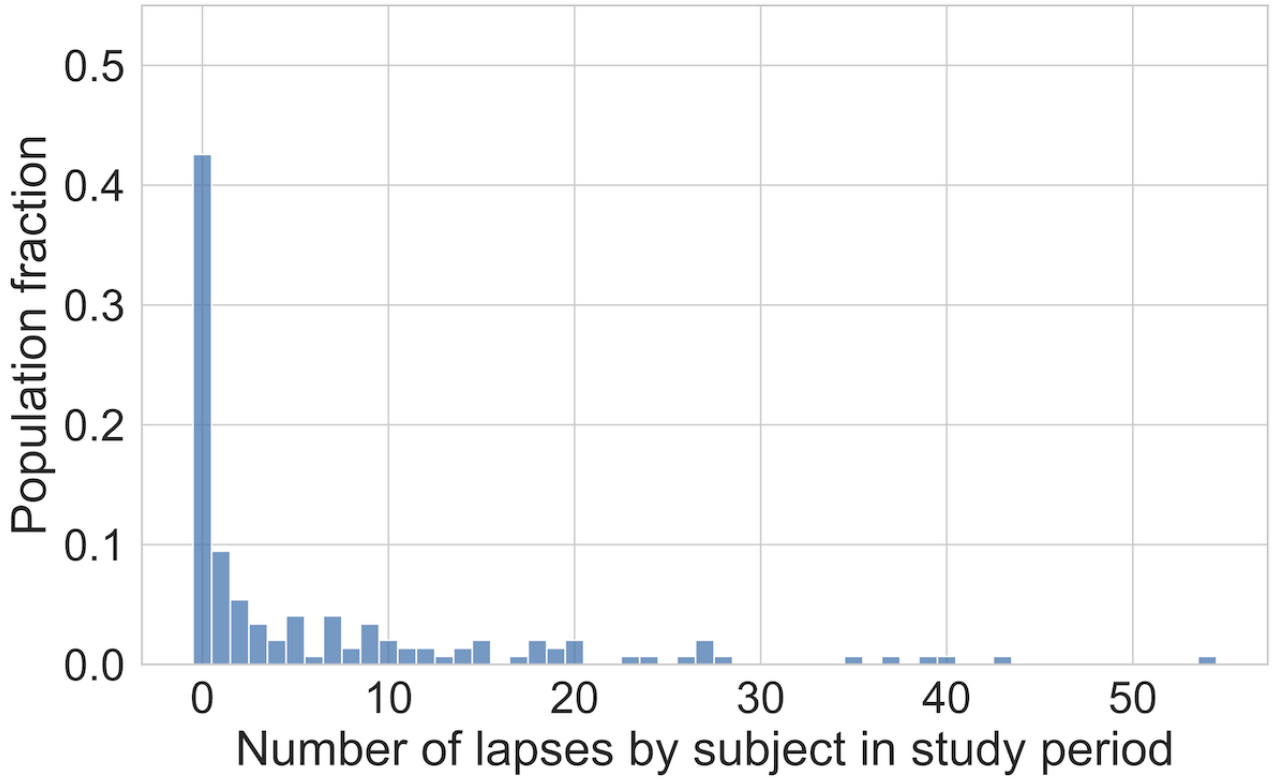

XSL•FO
RenderX

Lapse Prediction Performance

Figure 5 shows the posterior mean AUROC performance obtained from Bayesian hierarchical modeling across prediction tasks and data availability.

Plots of the raw AUROC values from repeated cross-validation are provided in Section S6.1 in Multimedia Appendix 1. Table 2 provides a particular summary of the posterior distribution, specifically the posterior probability that SSMs had the best mean AUROC across all methods. These probabilities were calculated by drawing posterior samples of model performance from the Bayesian hierarchical model and identifying the fraction of samples with mean AUROC performance for the SSMs greater than that for LR and XGB. Note that additional posterior summaries discussed in the OSF preregistration are available in Section S5.3 in Multimedia Appendix 1.

Section S6.3 in Multimedia Appendix 1 contains plots and tables for AUPRC, which are analogous to those presented here for AUROC.

Figure 5. Plots of posterior mean area under the receiver operating characteristic curve (AUROC) performance for 3 prediction tasks and 3 model types (state space model [SSM], logistic regression [LR], and extreme gradient boosting [XGB]). Medians are marked with dots, and 95% credible intervals (CrI) are provided as error bars. Each panel describes performance on a different prediction task (ie, same-day lapse, lapse within 3 days, and lapse within 7 days). The values reported for the 3 models at each x-value (number of days of data availability) are offset slightly by method for easier viewing of the median and CrIs. Note that there is substantial correlation between the performance of the different models in the posterior samples, making visual ranking comparisons using the CrIs incomplete. Model ranking is instead assessed by a separate summary calculation of posterior samples that accounts for this correlation.

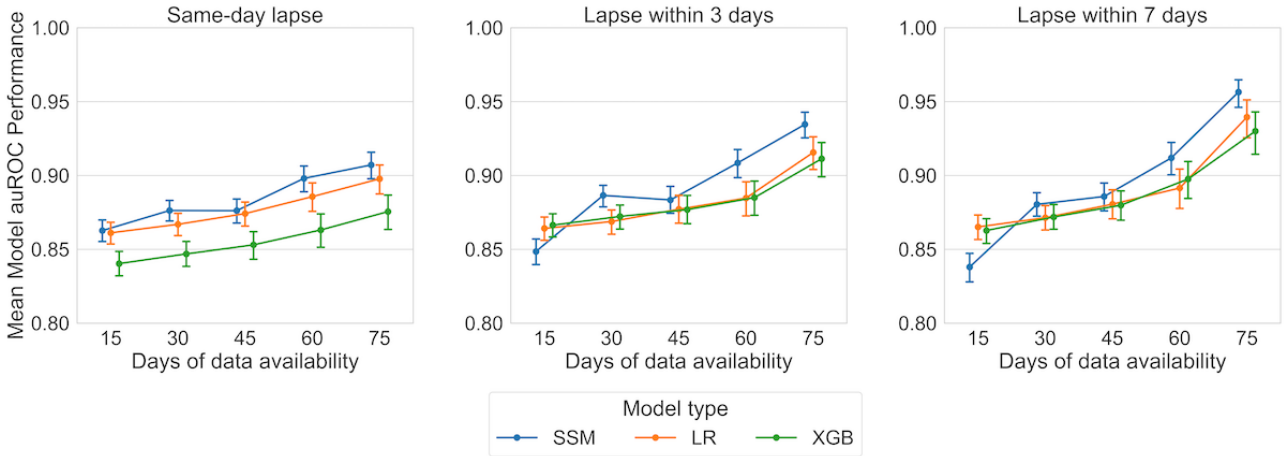

Table 2. A summary of state space model performance (posterior probability that the state space models had the highest mean area under the receiver operating characteristic curve across all methods) relative to the benchmark methods for each prediction task and amount of data availability.

Prediction task	Data availability (days)				
	15	30	45	60	75
Same-day lapse	0.732	0.999	0.760	0.999	0.995
Lapse within 3 days	<0.001	0.999	0.985	0.999	0.999
Lapse within 7 days	<0.001	0.997	0.913	0.999	0.999

Discussion

Principal Findings

We noted excellent performance across the tested methods, and our findings echo the sentiment of previously cited works [6,35-38]. The use of predictive modeling has the potential to meaningfully improve clinical treatment for AUD through accurate and automated risk assessment. Specifically, the experiments provide strong posterior evidence that SSMs can outperform existing ML approaches for AUD lapse prediction.

Figure 5 provides a coarse visual summary of mean AUROC performance, showing that the SSMs generally displayed higher mean performance than the benchmark methods. More formally, we rank-ordered the methods for each level of data availability using the posterior samples from our Bayesian hierarchical modeling. Table 2 provides the posterior probabilities that the SSMs had the best mean AUROC among all 3 assessed model types. We found strong evidence that, given 30 or more days of data availability, the SSMs outperformed the LR and XGB models. This held across all 3 prediction types, with the difference between the SSMs and LR or XGB models typically growing as data availability increased. There was significant posterior correlation between the mean AUROC performance of the different models. This explains why we obtained strong posterior evidence for the SSMs outperforming the other methods, even when there was overlap in the credible intervals for mean model AUROCs (Figure 5). A deeper discussion of this posterior correlation is provided in Section S6.2 in Multimedia Appendix 1.

More broadly, these results suggest a threshold of data availability occurring between 15 and 30 days of data availability, which divides 2 relevant performance regimes. Above this threshold, we noted uniformly better performance

XSL•FO
RenderX

across all tasks using the SSMs as compared to the benchmarks. Below this threshold, the performance results were mixed, with posterior estimates suggesting the SSMs had better same-day prediction performance but worse performance than the benchmark classifiers on longer prediction window tasks. We believe that this performance trend reflects the fact that SSMs are idiographic (training a unique model for each participant).

In the regime of low data availability (ie, only 15 days of data for a given participant), SSMs have very little patient-specific information to personalize a model beyond the initial parameter priors. SSMs estimate a model of each participant's time-series behavior, and this single model generates predictions for all tasks by simulating trajectories of the behavior over 1, 3, or 7 days. This means that errors in the participant's estimated model can be magnified when making predictions for longer time windows, likely explaining the poorer performance of SSMs on the latter prediction tasks.

As data availability increases, however, idiographic SSMs have more flexibility than population-trained ML benchmarks to capture the nuances of each participant's behavior. Since a unique model is trained for each participant, it can learn personalized patterns of behavior without impacting predictive performance for the rest of the population. Once a critical amount of participant data is available, the benefits of personalization dominate and SSMs outperform benchmark approaches across all tasks.

These performance differences highlight the importance of idiographic modeling for AUD lapse risk. If there were no meaningful heterogeneity in the presentation of AUD symptoms and lapse behavior, we would not expect to see a gap between idiographic and population-trained models. Further, AUD is a chronic, relapsing condition calling for continuing care for months or years. While population-trained models may have superior performance in the initial weeks of observation (depending on the prediction task), the vast majority of treatment would occur in a regime of data availability favoring idiographic SSMs.

Extensions to Intervention Frameworks

Beyond improved predictive performance, the use of SSMs enables exciting new approaches to treatment selection. SSMs build an approximate dynamic system model for how an individual's state of mind evolves over time. Using this, we can draw on powerful mathematical tools like control theory [61-65] to guide treatment decision-making. Control problems have 3 core components: a system model, a set of possible actions to take, and a defined objective. For example, in the case of automobile cruise control, the control task is to choose the amount of acceleration or braking to comfortably maintain a particular speed or distance from a neighboring vehicle. Critical to calculating the optimal action is the dynamic system model, which describes how the vehicle will respond to a given amount of acceleration or braking [13].

With an approximate system model for a patient's state of mind, we can apply a similar approach to treatment decisions. Using a digital therapeutics platform, a patient submits an EMA each morning. In response, the platform could automatically provide a risk estimate and a treatment recommendation to best support the patient's recovery (eg, performing a mindfulness or treatment-related exercise, recalling a particular strategy to control craving, meeting a supportive friend for lunch, providing a reminder to avoid high-risk locations, etc). This type of automated support, which considers the personalized factors driving each patient's risk, is a promising way to expand and improve care.

In principle, one could construct this type of treatment-recommendation platform using the benchmark classifier models, but SSMs offer a few specific advantages. First, as time-series models, SSMs allow for treatment recommendations that take the individual's behavioral dynamics into account. In control theory, this notion is called *planning*. Choosing the optimal action considers both the state of the system today and how the system might evolve in the future (including its response to today's chosen action). Interventions that myopically address factors causing elevated lapse risk *today* might be suboptimal choices to support the patient's long-term recovery. As classifier models do not model the time series dynamics of a patient, they do not support this type of principled treatment selection.

Second, since SSMs are trained separately for each patient, they can capture the effectiveness of a treatment for an individual patient rather than for the overall population. To make this concrete, we recall the SSM transition equation. This equation describes how an individual's hidden mental state (x) evolves over time, for instance, from day t to day $t+1$. Treatments would be included as additive terms in the transition equation:

A γ term is added to the right-hand side and signifies the treatment effect of the chosen intervention. This treatment effect may differ between individuals and is thus an important consideration for how to optimally recommend treatments. Just as the parameters related to the individual's behavioral dynamics, such as b and d here, are better estimated over time, we can similarly learn the personalized value of each treatment (γ) as additional patient data are collected. Note that this approach extends to cases where multiple possible interventions are considered. This would be expressed as a summation of distinct treatment effect terms (γ_i), each multiplied by a binary decision variable reflecting the choice to recommend that treatment, such that only the selected treatments impact the patient's system dynamics.

Further, a critical consideration in implementing effective digital therapeutic interventions is patient engagement [66]. The SSMs presented here can be augmented to include a state representing the participant's capacity for engagement. This would allow the model to consider how engagement evolves over time and to incorporate this information in the treatment selection problem. For instance, a higher burden treatment exercise might only be effective when the patient has the mental capacity to engage fully with it, compared to when they are stressed or overwhelmed. By including engagement dynamics in the patient model, the digital therapeutic platform could better recommend treatments to support patient recovery.

XSL•FO
RenderX

Clinical Interpretability of Modeling Approaches

Broadly, we believe that SSMs have the potential to provide improved clinical interpretability over the tested benchmark approaches, specifically through their personalization and the modeling of patient dynamics. To make these notions more concrete, we consider interpretability approaches for each model type.

LR models offer significant interpretability, as a clinician can directly evaluate fitted model coefficients to understand how input features impact risk assessment, and further examining the patient's specific covariates surrounding lapses can provide insights into what factors are driving lapse risks for a particular patient. Importantly, however, these interpretations are limited when models are fit at the population level, as the model coefficients are not necessarily representative of an individual patient.

Boosted tree-based models like XGB typically offer improved predictive performance but are less directly interpretable than an LR model. To understand which features are driving a patient's risk, a clinician might employ feature explainability tools like Shapley values [67]. These provide a way to assess the contribution of a particular feature (or groups of features) to predictions. Similar to LR, these values are somewhat limited when models are fit at the population level rather than to individuals.

SSMs are similar to LR models in that a clinician can directly examine coefficient values within the observation equations to see how particular mental constructs affect lapse risk. Analyzing specific latent state values can further inform which constructs are driving a patient's lapse risk. Importantly, since SSMs are fit to each individual, these coefficients reflect the behavior of a specific patient rather than the overall population. Additionally, SSMs capture the dynamics of these mental constructs through the transition equation. For instance, a patient's dynamics might indicate that one mental construct tends to decay from elevated levels toward baseline levels very slowly compared to another more transient construct (these quantities are directly visible as fitted parameters within the transition equations). Even if the current values of these example constructs contribute similarly to immediate risk, a clinician might choose to address the slower-decaying construct to better mitigate both immediate and longer-term lapse risk. This type of insight is not possible without a model that captures an individual's time-varying dynamics.

Limitations and Opportunities

The SSMs and study data presented in this manuscript have several limitations and considerations relevant for future work. We provide a discussion of the broader limitations here and point technically interested readers to additional modeling discussions in Section S3 in Multimedia Appendix 1.

Limited Demographic Diversity in the Study Population

This study is limited by its consideration of a narrow geographical area with correspondingly limited demographic diversity. This is typically a concern because trained models may show poor performance in populations that are underrepresented in the training sample [68,69]. We believe that such fairness concerns may be partially alleviated by the type of idiographic modeling framework presented here. Rather than applying a single, population-trained model to everyone, we advocate fitting personalized models to each patient that capture their unique patterns of behavior. However, whether this approach truly alleviates such concerns will require validation in separate studies. Evaluating the potential algorithmic fairness benefits of personalized models will be the focus of future work.

Limited Interpretability of SSMs due to Latent State Definition

The SSMs presented here have specific interpretability limitations, but these are not necessarily reflective of SSM approaches in general. In principle, SSMs can be constructed such that each latent state represents a specific mental construct. In this case, it is helpful to carefully structure EMA questions to provide targeted measurements of each relevant construct. Our modeling was performed after the parent study was designed and completed, and thus, we could not influence the existing EMA format. As such, we chose to use 2 latent states with no specific interpretations. This demonstrates that it is possible to approximate behavior well even with a low-dimensional linear model but indicates that this work cannot highlight the additional explanatory benefits of SSMs relative to classifier methods. We hope that future work will explore interpretable latent structures to bring SSMs closer to the existing literature of explanation-focused psychological modeling. Doing so would allow clinicians to examine fitted models to improve personalized treatment planning with patients [70].

Proposed SSMs Assume Time Invariance for Patient Models

The SSMs we presented assume that participants' dynamic models are time invariant. Over the 3-month study period, this may be a plausible assumption but is a clear limitation for longer time horizons. If individuals are trending toward recovery, this change should be represented through time-varying model parameters. For example, while craving might strongly influence lapse early in a participant's recovery, these dynamics may change over time as the participant learns to better address cravings, which can diminish their influence. SSMs can be structured to model these types of time-varying parameters [71], but this structure would require modifications to the model estimation strategies presented in this manuscript.

Performance Results Should be Considered in the Context of the Study Cohort

The predictive performance values achieved by our SSMs and the benchmark methods were linked to the size and composition of the study cohort as well as our chosen cross-validation structure. In the case of SSMs, the study cohort size and cross-validation fold count affected the number of other participants included in the fitting of prior distributions for individuals in the test fold. As our prior distributions were entirely data-driven, this effective cohort (ie, the count and composition of the individuals in the training folds) determined the strength of the priors constructed for the test fold. For the

XSL•FO
RenderX

benchmark methods, the cohort size and cross-validation fold count determined the size of the model training set. In future work, we plan to examine larger participant cohorts to rigorously characterize the impact of these factors on model performance.

Actively Collected Patient Data Have Important Limitations and Potential Biases

EMA data collection naturally imposes a burden on participants, although EMA data collection once per day, as used in this study, appears to be well-tolerated by patients for monitoring over months or years [43,47]. We thus believe that the performance reported in this study sets a plausible benchmark for the real-world use of such methods. However, we note that we excluded 3 of 151 participants from the original study based on poor EMA compliance (responding to fewer than 25% of prompted EMAs), as EMA noncompliance can introduce problems for all tested methods. Without reasonably up-to-date patient data, the prediction problem is not well-posed. While this decision represents a relatively small exclusion, it does mean that our results are more representative of compliant individuals. The use of passively collected data, such as GPS or sleep quality data, will be the focus of future work, as this may provide a means of gathering sufficient patient data for prediction, even with limited EMA compliance.

Further, while EMA represents an important step forward in patient data collection, it is still important to consider its structural limitations. For instance, as noted above, the primary use of EMA data introduces the possibility of participant selection bias [72], participant responses may be influenced by social desirability factors [73], and data quality may suffer from inattentive or careless responding [74]. Despite these potential limitations, we are encouraged to find that all tested models still performed well in our testing. We believe that the integration of passively collected data can help address the limitations of EMA data and make predictive systems more robust.

None of the modeling approaches (SSM, LR, or XGB) in this study incorporated passively sensed information, such as GPS, but all could, in principle, be modified to do so. For the LR or XGB models, integrating passively collected data would require the creation of new model features. For SSMs, integrating passive data would involve adding elements to the observation equations that relate passively observed quantities, such as the fraction of time spent in low-risk locations versus high-risk locations, to new functions of the latent states. We plan to incorporate passively sensed data in future work to decrease the sensing burden on participants, potentially improve prediction quality, and mitigate sources of bias.

Extensions to Other Mental Health Conditions

This work presents a case study for the use of SSMs in the context of AUD lapse prediction. The promising performance of SSMs in this domain suggests that they merit examination for risk assessment tasks in other mental health conditions (eg, predicting panic attacks or suicidal ideation). Of course, whether SSMs similarly outperform the benchmark ML approaches for other conditions will require validation in separate studies. It may be that, for instance, the benefits of personalization are more pronounced in some conditions versus others. For conditions with minimal patient heterogeneity, we would expect all approaches (SSM, LR, and XGB) to have similar predictive performance. However, even in these cases, we believe the structural benefits of SSMs (eg, easier integration into intervention frameworks for treatment selection) may make them a compelling choice for clinicians.

From an implementation standpoint, both SSMs and the benchmark ML approaches (LR and XGB) would require similar modifications to adapt them to new risk assessment tasks. SSMs model observable quantities (eg, EMA responses or clinically relevant events like lapse for AUD) as functions of dynamic processes of unobserved quantities (eg, a patient's evolving state of mind). Applying this framework to new conditions would involve determining which observable quantities are relevant to the condition under study. For instance, this might involve a different set of EMA questions (which could be smaller or larger than the 9-question EMA in this study) and a different primary prediction target (eg, panic attacks rather than AUD lapses). This is similar to the changes needed to apply LR or XGB approaches to a new condition. These models would also require potentially different EMA data sources, redefinition of appropriate features, and redefinition of the prediction target.

It is possible that different mental health conditions would require further structural modifications to the tested models. For the benchmark ML methods, this could entail different strategies in feature construction or the exploration of different model hyperparameters. Structural modifications to SSMs could involve changes, such as a larger or smaller number of latent states, a different granularity in time discretization, or modifications to the structure of the transition equations. These changes may drive modifications to the fitting approaches proposed in this manuscript but would not alter the types of predictions that can be made or the possible time series and personalization benefits of SSMs over the benchmark approaches. To begin exploring these questions more deeply, we plan to validate this approach in other forms of SUD prediction, such as lapses in opioid use.

Conclusion

This work presents a framework for the use of SSMs to predict lapses in individuals with AUD. Using a 3-month study of 148 individuals in early recovery from AUD, we demonstrated that SSMs can offer better lapse prediction performance compared to benchmark ML methods used in the literature. We have discussed the additional benefits of state SSMs as time-series methods. In particular, time-series approaches bring prediction-focused work closer to the methodologies used in explanation-focused psychological modeling and enable exciting new continuing care approaches using digital therapeutics. While our work deals specifically with AUD, we believe that the same modeling principles are suitable for evaluation in a variety of pressing mental health conditions. New approaches for automated risk monitoring and treatment suggestion can both improve care and help extend care to undertreated populations.

XSL•FO
RenderX

Acknowledgments

Support for this research was provided by the University of Wisconsin-Madison Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation, the National Science Foundation (grant number: 2041428), the National Institute on Drug Abuse (grant number: R01 DA047315), and the National Institute on Alcohol Abuse and Alcoholism (grant number: R01 AA024391). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors. The authors wish to thank Susan E Wanta for her role as the project administrator and for her help with data curation. The authors also wish to thank Kendra Wyant and Sarah J Sant'Ana for their help with data curation; Candace Lightheart, Jill Nagler, Kerry Keiser, and Megan Schultz for their contributions to data collection; and Chris Gioia for the clinical supervision he provided to graduate students. The authors attest that there was no use of generative artificial intelligence technology in the generation of text, figures, or other informational content of this manuscript.

Data Availability

We have provided all data necessary to reproduce the reported lapse prediction performance results at our publicly available, persistent project repository on Open Science Framework (see [41]). This repository also contains the code used for model fitting and the Bayesian analysis of predictive performance. Participant-level data are not included in this repository in order to safeguard protected health information. To discuss access to these data, please contact JC.

Authors' Contributions

Conceptualization: JC, EP, YM
Data curation: JC, EP
Methodology: EP, YM
Formal analysis: EP
Funding acquisition: JC
Investigation: JC
Supervision: JC, YM
Writing – original draft: EP
Writing – review and editing: JC, EP, YM

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional technical information.
[PDF File (Adobe PDF File), 929 KB - formative_v9i1e73265_app1.pdf]

Multimedia Appendix 2
Transparency report.
[DOCX File , 17 KB - formative_v9i1e73265_app2.docx]

Multimedia Appendix 3
Checklist for machine learning predictive modeling in biomedical applications.
[PDF File (Adobe PDF File), 118 KB - formative_v9i1e73265_app3.pdf]

References

1. Borsboom D, Deserno MK, Rhemtulla M, Epskamp S, Fried EI, McNally RJ, et al. Network analysis of multivariate data in psychological science. Nat Rev Methods Primers 2021 Aug 19;1(1):58. [doi: 10.1038/s43586-021-00055-w]
2. Chekroud AM, Bondar J, Delgadillo J, Doherty G, Wasil A, Fokkema M, et al. The promise of machine learning in predicting treatment outcomes in psychiatry. World Psychiatry 2021 Jun 18;20(2):154-170 [FREE Full text] [doi: 10.1002/wps.20882] [Medline: 34002503]
3. Deisenhofer A, Barkham M, Beierl ET, Schwartz B, Aafjes-van Doorn K, Beevers CG, et al. Implementing precision methods in personalizing psychological therapies: Barriers and possible ways forward. Behav Res Ther 2024 Jan;172:104443 [FREE Full text] [doi: 10.1016/j.brat.2023.104443] [Medline: 38086157]
4. Nahum-Shani I, Smith SN, Spring BJ, Collins LM, Witkiewitz K, Tewari A, et al. Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. Ann Behav Med 2018 May 18;52(6):446-462 [FREE Full text] [doi: 10.1007/s12160-016-9830-8] [Medline: 27663578]

XSL•FO
RenderX

5. Andersson G, Titov N, Dear BF, Rozental A, Carlbring P. Internet-delivered psychological treatments: from innovation to implementation. World Psychiatry 2019 Feb;18(1):20-28 [FREE Full text] [doi: 10.1002/wps.20610] [Medline: 30600624]
6. Wyant K, Sant'Ana SJ, Fronk G, Curtin J. Machine learning models for temporally precise lapse prediction in alcohol use disorder. J Psychopathol Clin Sci 2024 Oct;133(7):527-540. [doi: 10.1037/abn0000901] [Medline: 39172368]
7. McLellan AT, Starrels JL, Tai B, Gordon AJ, Brown R, Ghitza U, et al. Can substance use disorders be managed using the chronic care model? review and recommendations from a NIDA consensus group. Public Health Rev 2013 Dec 8;35(2):8. [doi: 10.1007/bf03391707]
8. Proctor S, Herschman P. The continuing care model of substance use treatment: what works, and when is "enough," "enough?". Psychiatry J 2014;2014:692423 [FREE Full text] [doi: 10.1155/2014/692423] [Medline: 24839597]
9. Socías ME, Volkow N, Wood E. Adopting the 'cascade of care' framework: an opportunity to close the implementation gap in addiction care? Addiction 2016 Dec 13;111(12):2079-2081 [FREE Full text] [doi: 10.1111/add.13479] [Medline: 27412876]
10. Mohr DC, Zhang M, Schueller SM. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. Annu Rev Clin Psychol 2017 May 08;13:23-47 [FREE Full text] [doi: 10.1146/annurev-clinpsy-032816-044949] [Medline: 28375728]
11. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol 2008 Apr;4(1):1-32. [doi: 10.1146/annurev.clinpsy.3.022806.091415] [Medline: 18509902]
12. Bolger N, Davis A, Rafaeli E. Diary methods: capturing life as it is lived. Annu Rev Psychol 2003;54:579-616. [doi: 10.1146/annurev.psych.54.101601.145030] [Medline: 12499517]
13. Gates KM, Chow SM, Molenaar PCM. Intensive Longitudinal Analysis of Human Processes. Boca Raton, FL: Chapman and Hall/CRC; 2023.
14. Molenaar PCM, Lo LL. Alternative Forms of Granger Causality, Heterogeneity, and Nonstationarity. In: Wiedermann W, von Eye A, editors. Statistics and Causality: Methods for Applied Empirical Research. Hoboken, NJ: John Wiley & Sons; 2016.
15. Molenaar PCM. A dynamic factor model for the analysis of multivariate time series. Psychometrika 2025 Jan 01;50(2):181-202. [doi: 10.1007/bf02294246]
16. Wood P, Brown D. The study of intraindividual differences by means of dynamic factor models: Rationale, implementation, and interpretation. Psychological Bulletin 1994;116(1):166-186. [doi: 10.1037//0033-2909.116.1.166]
17. Kim J, Zhu W, Chang L, Bentler PM, Ernst T. Unified structural equation modeling approach for the analysis of multisubject, multivariate functional MRI data. Hum Brain Mapp 2007 Feb 22;28(2):85-93 [FREE Full text] [doi: 10.1002/hbm.20259] [Medline: 16718669]
18. Nesselroade JR, Gerstorf D, Hardy SA, Ram N. Focus article: idiographic filters for psychological constructs. Measurement: Interdisciplinary Research & Perspective 2007 Dec 04;5(4):217-235. [doi: 10.1080/15366360701741807]
19. Borsboom D, Cramer AO. Network analysis: an integrative approach to the structure of psychopathology. Annu Rev Clin Psychol 2013 Mar 28;9(1):91-121. [doi: 10.1146/annurev-clinpsy-050212-185608] [Medline: 23537483]
20. Beltz AM, Gates KM. Network mapping with GIMME. Multivariate Behav Res 2017 Nov 21;52(6):789-804 [FREE Full text] [doi: 10.1080/00273171.2017.1373014] [Medline: 29161187]
21. Epskamp S, van Borkulo CD, van der Veen DC, Servaas MN, Isvoranu A, Riese H, et al. Personalized network modeling in psychopathology: the importance of contemporaneous and temporal connections. Clin Psychol Sci 2018 May 19;6(3):416-427 [FREE Full text] [doi: 10.1177/2167702617744325] [Medline: 29805918]
22. Asparouhov T, Hamaker EL, Muthén B. Dynamic structural equation models. Structural Equation Modeling: A Multidisciplinary Journal 2017 Dec 27;25(3):359-388. [doi: 10.1080/10705511.2017.1406803]
23. Briganti G, Scutari M, Epskamp S, Borsboom D, Hoekstra RHA, Golino HF, et al. Network analysis: An overview for mental health research. Int J Methods Psychiatr Res 2024 Dec 14;33(4):e2034. [doi: 10.1002/mpr.2034] [Medline: 39543824]
24. Levine FM, Sandeen E, Murphy CM. The therapist's dilemma: Using nomothetic information to answer idiographic questions. Psychotherapy: Theory, Research, Practice, Training 1992;29(3):410-415. [doi: 10.1037/h0088544]
25. Molenaar PCM. A manifesto on psychology as idiographic science: bringing the person back into scientific psychology, this time forever. Measurement: Interdisciplinary Research & Perspective 2004 Oct;2(4):201-218. [doi: 10.1207/s15366359mea0204_1]
26. Piccirillo ML, Rodebaugh TL. Foundations of idiographic methods in psychology and applications for psychotherapy. Clin Psychol Rev 2019 Jul;71:90-100 [FREE Full text] [doi: 10.1016/j.cpr.2019.01.002] [Medline: 30665765]
27. Wright AGC, Woods WC. Personalized models of psychopathology. Annu Rev Clin Psychol 2020 May 07;16:49-74. [doi: 10.1146/annurev-clinpsy-102419-125032] [Medline: 32070120]
28. Fisher AJ, Reeves JW, Lawyer G, Medaglia JD, Rubel JA. Exploring the idiographic dynamics of mood and anxiety via network analysis. J Abnorm Psychol 2017 Nov;126(8):1044-1056. [doi: 10.1037/abn0000311] [Medline: 29154565]
29. Piccirillo ML, Frumkin MR, Spink KM, Tonge NA, Foster KT. Examining time-varying dynamics of co-occurring depressed mood and anxiety. J Affect Disord 2024 Oct 01;362:24-35. [doi: 10.1016/j.jad.2024.06.064] [Medline: 38906224]

XSL•FO
RenderX

30. Kuehn KS, Piccirillo ML, Kuczynski AM, King KM, Depp CA, Foster KT. Person-specific dynamics between negative emotions and suicidal thoughts. Compr Psychiatry 2024 Aug;133:152495 [FREE Full text] [doi: 10.1016/j.comppsych.2024.152495] [Medline: 38728844]
31. Bringmann LF, Vissers N, Wichers M, Geschwind N, Kuppens P, Peeters F, et al. A network approach to psychopathology: new insights into clinical longitudinal data. PLoS One 2013 Apr 4;8(4):e60188 [FREE Full text] [doi: 10.1371/journal.pone.0060188] [Medline: 23593171]
32. Robinaugh DJ, Brown ML, Losiewicz OM, Jones PJ, Marques L, Baker AW. Towards a precision psychiatry approach to anxiety disorders with ecological momentary assessment: the example of panic disorder. Gen Psychiatr 2020 Feb 09;33(1):e100161 [FREE Full text] [doi: 10.1136/gpsych-2019-100161] [Medline: 32175524]
33. Park JJ, Chow S, Epskamp S, Molenaar PCM. Subgrouping with chain graphical VAR models. Multivariate Behav Res 2024 Feb 13;59(3):543-565 [FREE Full text] [doi: 10.1080/00273171.2023.2289058] [Medline: 38351547]
34. Yarkoni T, Westfall J. Choosing prediction over explanation in psychology: lessons from machine learning. Perspect Psychol Sci 2017 Nov;12(6):1100-1122 [FREE Full text] [doi: 10.1177/1745691617693393] [Medline: 28841086]
35. Bae S, Chung T, Ferreira D, Dey AK, Suffoletto B. Mobile phone sensors and supervised machine learning to identify alcohol use events in young adults: Implications for just-in-time adaptive interventions. Addict Behav 2018 Aug;83:42-47 [FREE Full text] [doi: 10.1016/j.addbeh.2017.11.039] [Medline: 29217132]
36. Soyster PD, Ashlock L, Fisher AJ. Pooled and person-specific machine learning models for predicting future alcohol consumption, craving, and wanting to drink: A demonstration of parallel utility. Psychol Addict Behav 2022 May;36(3):296-306. [doi: 10.1037/adb0000666] [Medline: 35041441]
37. Bae SW, Suffoletto B, Zhang T, Chung T, Ozolcer M, Islam MR, et al. Leveraging mobile phone sensors, machine learning, and explainable artificial intelligence to predict imminent same-day binge-drinking events to support just-in-time adaptive interventions: algorithm development and validation study. JMIR Form Res 2023 May 04;7:e39862 [FREE Full text] [doi: 10.2196/39862] [Medline: 36809294]
38. Leenaerts N, Soyster P, Ceccarini J, Sunaert S, Fisher A, Vrieze E. Person-specific and pooled prediction models for binge eating, alcohol use and binge drinking in bulimia nervosa and alcohol use disorder. Psychol. Med 2024 May 22;54(10):2758-2773. [doi: 10.1017/s0033291724000862]
39. Paul GL. Strategy of outcome research in psychotherapy. J Consult Psychol 1967 Apr;31(2):109-118. [doi: 10.1037/h0024436] [Medline: 5342732]
40. Chow S, Ho MR, Hamaker EL, Dolan CV. Equivalence and differences between structural equation modeling and state-space modeling techniques. Structural Equation Modeling: A Multidisciplinary Journal 2010 Apr 13;17(2):303-332. [doi: 10.1080/10705511003661553]
41. Pulick E, Curtin J, Mintz Y. Idiographic lapse prediction with state space modeling. OSF. URL: https://osf.io/8xbk2/ [accessed 2025-02-28]
42. Pulick E, Mintz Y, Curtin JJ. Idiographic models for lapse prediction. OSF. URL: https://osf.io/wbxfp [accessed 2025-02-28]
43. Wyant K, Moshontz H, Ward SB, Fronk GE, Curtin JJ. Acceptability of personal sensing among people with alcohol use disorder: observational study. JMIR Mhealth Uhealth 2023 Aug 28;11:e41833 [FREE Full text] [doi: 10.2196/41833] [Medline: 37639300]
44. Aczel B, Szaszi B, Sarafoglou A, Kekecs Z, Kucharský Š, Benjamin D, et al. A consensus-based transparency checklist. Nat Hum Behav 2019;4(1):4-6 [FREE Full text] [doi: 10.1038/s41562-019-0772-6] [Medline: 31792401]
45. Klement W, El Emam K. Consolidated reporting guidelines for prognostic and diagnostic machine learning modeling studies: development and validation. J Med Internet Res 2023 Aug 31;25:e48763 [FREE Full text] [doi: 10.2196/48763] [Medline: 37651179]
46. Derogatis LR, Lipman RS, Covi L. SCL-90: an outpatient psychiatric rating scale--preliminary report. Psychopharmacol Bull 1973 Jan;9(1):13-28. [Medline: 4682398]
47. Moshontz H, Colmenares AJ, Fronk GE, Sant'Ana SJ, Wyant K, Wanta SE, et al. Prospective prediction of lapses in opioid use disorder: protocol for a personal sensing study. JMIR Res Protoc 2021 Dec 07;10(12):e29563 [FREE Full text] [doi: 10.2196/29563] [Medline: 34559061]
48. Kalman RE. A new approach to linear filtering and prediction problems. J Basic Eng 1960;82(1):35-45. [doi: 10.1115/1.3662552]
49. McGee LA, Schmidt SF. Discovery of the Kalman filter as a practical tool for aerospace and industry. NASA. 1985. URL: https://ntrs.nasa.gov/citations/19860003843 [accessed 2025-05-21]
50. Auger‐Méthé M, Newman K, Cole D, Empacher F, Gryba R, King AA, et al. A guide to state–space modeling of ecological time series. Ecological Monographs 2021 Aug 25;91(4):e01470. [doi: 10.1002/ecm.1470]
51. Lodewyckx T, Tuerlinckx F, Kuppens P, Allen NB, Sheeber L. A hierarchical state space approach to affective dynamics. J Math Psychol 2011 Feb 01;55(1):68-83 [FREE Full text] [doi: 10.1016/j.jmp.2010.08.004] [Medline: 21516216]
52. Shumway RH, Stoffer DS. Time Series Analysis and Its Applications: With R Examples. Cham, Switzerland: Springer International Publishing; 2017.
53. McElreath R. Statistical Rethinking: A Bayesian Course with Examples in R and Stan. Boca Raton, FL: CRC Press/Taylor & Francis Group; 2016.

XSL•FO
RenderX

54. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. The Journal of Machine Learning Research 2011;12:2825-2830. [doi: 10.5555/1953048.2078195]
55. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016 Presented at: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13-17, 2016; San Francisco, CA. [doi: 10.1145/2939672.2939785]
56. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982 Apr;143(1):29-36. [doi: 10.1148/radiology.143.1.7063747] [Medline: 7063747]
57. Boyd K, Eng KH, Page CD. Area under the Precision-Recall Curve: Point Estimates and Confidence Intervals. In: Blockeel H, Kersting K, Nijssen S, Železný F, editors. Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2013. Lecture Notes in Computer Science, vol 8190. Berlin, Heidelberg: Springer; 2013:451-466.
58. Kruschke JK, Liddell TM. The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. Psychon Bull Rev 2018 Feb 7;25(1):178-206. [doi: 10.3758/s13423-016-1221-4] [Medline: 28176294]
59. Benavoli A, Corani G, Demšar J, Zaffalon M. Time for a change: a tutorial for comparing multiple classifiers through Bayesian analysis. J Mach Learn Res 2017;18(77):1-36 [FREE Full text]
60. Kuhn M. tidyposterior: Bayesian Analysis to Compare Models using Resampling Statistics. tidyposterior. 2023. URL: https://tidyposterior.tidymodels.org/ [accessed 2025-05-21]
61. Bertsekas D. Dynamic programming and optimal control: Volume I. Nashua, NH: Athena Scientific; 2012.
62. Leigh JR. Control Theory: A Guided Tour. London, England: Institution of Engineering and Technology; 2012.
63. Aswani A, Kaminsky P, Mintz Y, Flowers E, Fukuoka Y. Behavioral modeling in weight loss interventions. Eur J Oper Res 2019 Feb 01;272(3):1058-1072 [FREE Full text] [doi: 10.1016/j.ejor.2018.07.011] [Medline: 30778275]
64. Li Q, Gavin KL, Voils CI, Mintz Y. An Adaptive Optimization Approach to Personalized Financial Incentives in Mobile Behavioral Weight Loss Interventions. arXiv. 2023. URL: https://arxiv.org/abs/2307.00444 [accessed 2025-05-21]
65. Adams KB, Boutilier JJ, Deo S, Mintz Y. Planning a Community Approach to Diabetes Care in Low- and Middle-Income Countries Using Optimization. arXiv. 2023. URL: https://arxiv.org/abs/2305.06426 [accessed 2025-05-21]
66. Nahum-Shani I, Yoon C. Towards the science of engagement with digital interventions. Curr Dir Psychol Sci 2024 Aug 03;33(4):239-246. [doi: 10.1177/09637214241254328] [Medline: 39552747]
67. Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions. 2017 Presented at: 31st Conference on Neural Information Processing System; December 4-9, 2017; Long Beach, CA URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf
68. Olfat M, Mintz Y. Flexible Regularization Approaches for Fairness in Deep Learning. 2020 Presented at: 59th IEEE Conference on Decision and Control (CDC); December 14-18, 2020; Jeju, Korea (South). [doi: 10.1109/CDC42340.2020.9303736]
69. Mitchell S, Potash E, Barocas S, D'Amour A, Lum K. Algorithmic fairness: choices, assumptions, and definitions. Annu Rev Stat Appl 2021 Mar 07;8(1):141-163. [doi: 10.1146/annurev-statistics-042720-125902]
70. Frumkin MR, Piccirillo ML, Beck ED, Grossman JT, Rodebaugh TL. Feasibility and utility of idiographic models in the clinic: A pilot study. Psychother Res 2021 Apr 24;31(4):520-534 [FREE Full text] [doi: 10.1080/10503307.2020.1805133] [Medline: 32838671]
71. Holmes EE. Derivation of an EM algorithm for constrained and unconstrained multivariate autoregressive state-space (MARSS) models. arXiv. 2013. URL: https://arxiv.org/abs/1302.3919 [accessed 2025-05-21]
72. Stone AA, Schneider S, Smyth JM. Evaluation of pressing issues in ecological momentary assessment. Annu Rev Clin Psychol 2023 May 09;19(1):107-131 [FREE Full text] [doi: 10.1146/annurev-clinpsy-080921-083128] [Medline: 36475718]
73. Schwarz N. Self-reports: How the questions shape the answers. American Psychologist 1999 Feb;54(2):93-105. [doi: 10.1037/0003-066X.54.2.93]
74. Eisele G, Vachon H, Lafit G, Kuppens P, Houben M, Myin-Germeys I, et al. The effects of sampling frequency and questionnaire length on perceived burden, compliance, and careless responding in experience sampling data in a student population. Assessment 2022 Mar;29(2):136-151. [doi: 10.1177/1073191120957102] [Medline: 32909448]

Abbreviations

AUD: alcohol use disorder
AUPRC: area under the precision-recall curve
AUROC: area under the receiver operating characteristic curve
EMA: ecological momentary assessment
LR: logistic regression
ML: machine learning
MLE: maximum likelihood estimation
OSF: Open Science Framework

XSL•FO
RenderX

ROC: receiver operating characteristic
SSM: state space model
SUD: substance use disorder
XGB: extreme gradient boosting

Edited by A Mavragani; submitted 06.03.25; peer-reviewed by X Zheng; comments to author 08.04.25; revised version received 25.04.25; accepted 13.05.25; published 03.06.25.

Please cite as:
Pulick E, Curtin J, Mintz Y
Idiographic Lapse Prediction With State Space Modeling: Algorithm Development and Validation Study
JMIR Form Res 2025;9:e73265
URL: https://formative.jmir.org/2025/1/e73265
doi:10.2196/73265
PMID:

©Eric Pulick, John Curtin, Yonatan Mintz. Originally published in JMIR Formative Research (https://formative.jmir.org), 03.06.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Mobile App (Joint Effort) to Support Cannabis Use Self-Management and Reinforce the Use of Protective Behavioral Strategies: Development Process and Usability Testing

José Côté[1,2,3], PhD; Patricia Auger[2,3], MSc; Gabrielle Chicoine[3,4], PhD; Jinghui Cheng[5], PhD; Sylvie Cossette[1,6], PhD; Guillaume Fontaine[7,8,9,10,11], PhD; Christine Genest[1,12], PhD; Shalini Lal[2,13,14], PhD; Judith Lapierre[15,16], PhD; M Gabrielle Pagé[2,17], PhD; Marc-André Maheu-Cadotte[1], PhD; Geneviève Rouleau[3,18,19,20], PhD; Billy Vinette[1,3], PhD; Didier Jutras-Aswad[2,21], MD

[1]Faculty of Nursing, Université de Montréal, Montreal, QC, Canada
[2]Research Centre, Centre Hospitalier de l'Université de Montréal, Montreal, QC, Canada
[3]Research Chair in Innovative Nursing Practices, Montreal, QC, Canada
[4]Knowledge Translation Program, Li Ka Shing Knowledge Institute, St. Michael's Hospital, Toronto, ON, Canada
[5]Department of Computer Engineering and Software Engineering, Polytechnique Montréal, Montreal, QC, Canada
[6]Montreal Heart Institute, Montreal, QC, Canada
[7]Ingram School of Nursing, Faculty of Medicine and Health Sciences, McGill University, Montreal, QC, Canada
[8]Centre for Clinical Epidemiology, Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, QC, Canada
[9]Centre for Nursing Research, Jewish General Hospital, Montreal, QC, Canada
[10]Viral Hepatitis Clinical Research Program, Kirby Institute, UNSW Sydney, Sydney, Australia
[11]Centre for Implementation Research, Methodological and Implementation Research Program, Ottawa Hospital Research Institute, Ottawa, ON, Canada
[12]Centre d'étude sur le trauma, Research Centre, Hôpital Louis-H Lafontaine, Montreal, QC, Canada
[13]School of Rehabilitation, Faculty of Medicine, Université de Montréal, Montreal, QC, Canada
[14]Douglas Research Centre, Douglas Mental Health University Institute, Montreal, QC, Canada
[15]Faculty of Nursing, Université Laval, Québec, QC, Canada
[16]Center for Research on Social Innovations, Université Laval, Quebec, QC, Canada
[17]Departement of Anesthesioloy and Pain Medecine, Faculty of Medecine, Université de Montréal, Montreal, QC, Canada
[18]Department of Nursing, Université du Québec en Outaouais, Gatineau, QC, Canada
[19]Institute for Health System Solutions and Virtual Care, Women's College Hospital, Toronto, ON, Canada
[20]Institut du Savoir Montfort, Ottawa, ON, Canada
[21]Department of Psychiatry and Addictology, Faculty of Medecine, Université de Montréal, Montreal, QC, Canada

Corresponding Author:
José Côté, PhD
Faculty of Nursing
Université de Montréal
2375 Chem. de la Côte-Sainte-Catherine
Montreal, QC, H3S 2N4
Canada
Phone: 1 514 343 7156
Email: jose.cote@umontreal.ca

Abstract

Background: Canada's legalization of recreational cannabis use (CU) has further highlighted the need for innovative interventions that promote lower-risk CU. Young adults aged 18-25 years represent the age group with the highest prevalence of CU. Protective behavioral strategies (PBSs) have been shown to help manage CU and reduce its negative consequences. To date, only a few interventions have focused on PBSs. To address this gap, a mobile app prototype using PBSs to influence CU was developed with and for young adults.

XSL•FO
RenderX

Objective: This study aims to describe the development process and usability testing of Joint Effort, a CU self-management mobile app prototype centered on promoting the use of PBSs among young adults with any past 30-day CU.

Methods: Intervention mapping (IM) and a co-design approach were used. Six steps were followed: (1) focus groups were conducted to identify needs and preferences regarding CU interventions; (2) a matrix of change objectives was used to select target behaviors and determinants; (3) theory-based intervention methods and practical applications were selected; (4) focus groups were held to validate the intervention structure and examples of tailored messages; (5) preliminary intervention content was created; and (6) the intervention content was transposed into a mobile app prototype. Usability was assessed through qualitative semistructured interviews and the User Version of the Mobile Application Rating Scale (uMARS), completed by a sample of 20 university students with a mean age of 21.8 (median 22) years, 14 (70%) of whom were women and 15 (75%) were undergraduates. Qualitative data were analyzed using thematic analysis.

Results: Four themes were identified from the interviews: Joint Effort was visually pleasing and easy to use; the content was well-adapted to the target audience and nonjudgmental; customization functions were appreciated; and the app was perceived as helpful and relevant for initiating behavior change. The prototype received a mean quality score of 4.43/5.0 (SD 0.53) per item on the uMARS. The mean scores on the 5 subscales were as follows: engagement (4.14, SD 0.53), functionality (4.60, SD 0.47), aesthetics (4.53, SD 0.52), information quality (4.44, SD 0.61), and subjective quality (3.36, SD 0.53).

Conclusions: Our findings highlight the added value of IM and a co-design approach, underscoring the importance of incorporating user feedback in the development of mobile apps. Building on the strong usability results, the Joint Effort prototype has since been developed into an iOS mobile app, and larger-scale evaluations are currently underway to assess its acceptability, feasibility, and efficacy.

(JMIR Form Res 2025;9:e71924) doi:10.2196/71924

KEYWORDS

cannabis; intervention; mobile app; mHealth; self-management; young adults; students; development; usability; prototype

Introduction

Background

Previously conducted prospective cohort studies have suggested that substance use begins in adolescence and peaks in young adulthood [1], and that regular cannabis use (CU) typically starts between the ages of 18 and 19 years [2]. In Canada, the 18-24-year age group has been found to have the highest percentage of cannabis users [3].

Recreational CU was legalized nationwide in Canada in 2018 [4]. Since then, the use of cannabis products has increased considerably, particularly among the 18-25-year age group [1,5]. It has been shown that increased availability of legal cannabis can lead to higher CU and associated riskier use behaviors [6]. Legalization has made it all the more imperative to develop and offer preventive interventions aimed at reducing CU-related harm and encouraging safer behaviors [7]. Young adults, in particular, stand to benefit from such initiatives, given the high prevalence of CU in the 18-25-year age group and the limited number of existing interventions designed specifically for them [8].

A new generation of interventions targeting CU through digital technologies and mobile health (mHealth) has drawn researchers' attention over the past decade. Several systematic reviews and meta-analyses [9-14] have demonstrated the relative effectiveness of various digital interventions in reducing CU frequency, although the effect sizes have been relatively small. However, these reviews have primarily focused on web-based interventions.

Smartphone usage and social media are increasingly popular among young people [15], and young adults are open to digital health services, especially when delivered via mobile apps [16]. However, our own systematic review of the literature found that none of the digital CU-related interventions available through mobile apps had been evaluated using a randomized controlled trial (RCT) [17]. A scoping review based on 5 studies, which aimed to explore the technical and functional characteristics of mHealth apps, found that this intervention modality demonstrated adequate feasibility and acceptability [18]. The following CU-related mobile apps were covered in the review: APPT (Assess, Plan, Track, and Tips; 18-50-year age group; Australia) [19]; MiSARA, a substance abuse research assistant (16-24-year age group; United States) [20]; VoltEgySzer (Once Upon a High; 14-18-year age group; Hungary) [21]; Stop-cannabis (14-59-year age group; Switzerland) [22]; and MApp (Marijuana Smartphone App; 18-25-year age group; United States) [23]. Since this review, 2 additional mobile apps have been mentioned in the scientific literature: HAP-app, a mobile app intended to help individuals reduce or cease CU (Norway) [24]; and LooseLeaf, a mobile app for young people at clinically high risk for psychosis, designed to help monitor CU and cannabis-related experiences (14-30-year age group; Canada) [25]. In short, the current research on CU mobile apps remains at an early stage, and most published studies have targeted either specific clinical populations or broad segments of the general population.

An innovative and positive way of promoting lower-risk CU is through the use of protective behavioral strategies (PBSs). PBSs are approaches that can help mediate CU and reduce negative CU consequences [26-28]. They involve behaviors—enacted immediately before, during, or after using cannabis—that support better CU self-management [26]. The use of PBSs has been associated with reduced negative effects of CU and lower CU frequency and quantity in a population of young university

XSL•FO
RenderX

students [26-28]. In a previous study involving Canadian university students who reported any CU in the past 30 days (n=211), we found that greater use of PBSs was related to lower CU frequency, and that daily cannabis users employed fewer marijuana PBSs on average than lower-frequency users did [28]. A recent scoping review showed that PBSs may be associated with reduced CU-related negative consequences [29].

While the incorporation of PBSs is linked to CU self-management and holds great promise for lower-risk CU interventions, there has been very little research on PBS-centered interventions beyond the realm of alcohol use [9], and only a few interventions to date have targeted CU specifically [30]. Preliminary testing of MApp, a smartphone app used as part of a brief in-person intervention, indicated that PBS use reduced CU and that young adults were inclined to use a mobile app to help moderate their CU [23]. More recently, Lewis et al [31] published an RCT study protocol to evaluate a brief PBS-focused web-based and SMS intervention, although the results have yet to be published.

In sum, although the evaluation of PBS use as an intervention target in the context of safe CU is still at an early stage, findings to date suggest that PBS use is a positive and innovative approach to promoting lower-risk CU [31]. Against this background, we used intervention mapping (IM) to develop a mobile app prototype designed to support CU self-management and reinforce PBS use, with the goal of promoting lower-risk CU among young adults in the context of legalized recreational CU in Canada.

Objective

This study aimed to develop Joint Effort, a mobile app prototype designed to support CU self-management and reinforce PBS use among young adults, and to assess its usability.

Methods

Structured Intervention Development Through IM

IM is a rigorous, theory- and evidence-based approach that emphasizes the role of research and theory in the intervention development process [32]. Our process was based on IM and a co-design approach. It included 6 steps: (1) conducting focus groups to identify the needs and preferences of young adults regarding CU interventions; (2) creating matrices of change objectives to select target behaviors and determinants; (3) selecting theory-based intervention methods and practical applications; (4) developing preliminary intervention content and structure; (5) conducting focus groups to validate the intervention structure and examples of tailored messages; and (6) transposing the intervention content into a mobile app prototype. The completion of one step guided the development of the next [32]. Various research phases and activities were embedded in the IM steps. These are described below.

Usability testing aimed to document the experience of using the mobile app prototype. This was conducted through semistructured interviews and the User Version of the Mobile Application Rating Scale (uMARS) [33]. A multimethod approach combining quantitative and qualitative data is generally recommended in usability testing to provide a more comprehensive understanding of user experiences [34]. The methods and results are reported in accordance with the Consolidated Criteria for Reporting Qualitative Research (COREQ) [35] (see Multimedia Appendix 1).

The target population for this study was selected based on 3 elements. First, substance use peaks in young adulthood [1], and the 18-24-year age group has the highest percentage of cannabis users [4]. Second, the legal age to purchase cannabis in Canada varies from 18 to 21 years, depending on provincial laws [36]. In Québec, where this study was conducted, it is illegal for anyone under 21 years to purchase cannabis. The lower age threshold for the study population was therefore set at 21 years. Third, most research on PBSs has been conducted with young adults and college or university students [26,37,38]. Given that the proof of concept for interventions based on PBSs has been demonstrated with this clientele, an academic setting was favored. This is why university students aged 21-24 years were chosen as a convenience sample to represent the young adult demographic. Recruitment was conducted at the Université de Montréal, a predominantly French-language university located in a metropolitan region (Montreal, Québec, Canada) in Eastern Canada.

Development Process (IM Steps)

Step 1: Focus Groups to Identify Needs and Preferences

Overview

Focus groups were conducted to identify needs and preferences related to CU interventions.

Participant and Procedures

Participants were recruited through posters and social media advertisements on campus and in various Facebook student groups. We aimed to recruit at least 15 students aged 21-24 years from the Université de Montréal (Québec, Canada).

The 3 focus groups were conducted in person and led in French by 2 members of the research team. Two major topics were addressed: cannabis-related needs and interests, and preferences for technology-based interventions. These were explored through questions such as “What aspects of CU would you like to know more about?” “In your opinion, what constitutes safe and informed CU behaviors?” “How would you like to receive information or support (eg, device type, medium, form)?” “What might motivate you to use a digital intervention to receive information and support regarding your CU?”

All focus groups were audio-recorded and transcribed with participants’ consent. Each participant received CAD $50 (US $36.4) as compensation for their participation.

Analysis

Thematic analysis principles were used to analyze the data collected through the focus groups [39]. The transcripts were read multiple times to develop a deductive thematic coding tree. Three coders pilot-tested the coding tree by independently coding 1 of the focus groups. Differences among the coders were resolved through discussion, and the coding was then merged. The coding tree was finalized, and the definition of each code was discussed in depth before being applied to the

XSL•FO
RenderX

remaining 2 transcripts. Descriptive codes were grouped into higher-order thematic categories, and the relationships between themes were detailed in a narrative summary. The summary was reviewed by the first author (JC). NVivo version 12 (QSR International Pty Ltd) was used for data management.

Findings

A total of 13 participants met the eligibility criteria (ie, university students aged 21-24 years who reported CU) and were divided into 3 groups of 4 or 5, based on their availability. The focus groups were conducted in December 2019 and January 2020. Sessions lasted 61, 62, and 76 minutes, respectively. Most participants self-identified as women (11/13, 85%). The mean age of participants was 22.4 (median 22) years.

Participants expressed that safe and responsible CU behaviors were characterized primarily by knowledge of the "facts" (eg, risks associated with CU, side effects of cannabis) or of "themselves" (eg, knowing one's limits). Safe and responsible behavior involved ensuring that CU did not interfere with daily functioning or with personal, professional, and relational responsibilities. Participants emphasized the importance of having access to information and resources regarding CU.

In terms of technology preferences, participants emphasized that CU self-management interventions should be personalized and customizable. For example, they indicated that the amount and depth of information provided, notification frequency, and message format (eg, video, audio, or text) should be tailored to individual preferences. Additionally, participants suggested that the proposed intervention should be discreet, user-friendly, and engaging. They also expressed interest in having access to a CU monitoring feature (eg, where, when, quantity, with whom) to support self-monitoring and personal goal setting. The themes identified through thematic analysis, along with associated quotations, are presented in Multimedia Appendix 2.

Key Implications for Design

The data generated during the focus groups contributed to a deeper understanding of the needs and preferences of potential intervention users and laid the groundwork for the subsequent steps of intervention development. According to the target population, the intervention tool should be both educational and fun. It should provide information that is reliable, easily accessible and understandable, confidential, personalized, and encouraging. Additionally, it should include a CU logbook, reminders, and positive feedback.

Step 2: Matrix of Change Objectives

The process of creating a matrix of change objectives involved 3 steps: (1) specifying performance objectives (POs); (2) selecting important and changeable determinants of behavior; and (3) determining specific change objectives. The goal of the proposed intervention was to promote CU self-management and reinforce PBS use among young adult cannabis users. The target behaviors (ie, CU self-management and PBS use) encompassed various subbehaviors, which were translated into 4 POs described in Table 1.

Table 1. Matrix of change objectives.

Performance objectives	Determinants			
	Attitude	Self-efficacy	Social norms	Intention
• Gain awareness of your CU[a] (frequency and consequences).	• Assess the reasons for wanting to change your CU.	• Identify risky situations related to your CU.	• Determine how you measure up against other users.	• N/A[b]
• Set a goal for yourself.	• Appreciate the importance of setting this goal.	• Reflect upon your capacity to pursue this goal.	• N/A	• Reflect upon and assess your motivation and intention for setting this goal.
• Commit to a change process and strive to achieve your goal.	• Identify the benefits of change.	• Identify possible obstacles and facilitators. • Focus on means and resources.	• N/A	• Formulate an action plan based on PBS[c].
	• Focus on the benefits of and motivation for change.	• Identify ways of overcoming obstacles.	• N/A	• Implement your action plan.
• Overcome possible obstacles and consolidate your goal.	• Remind yourself of the benefits identified.	• Call on factors capable of facilitating change and on external resources.	• N/A	• Revise or adjust your action plan.

[a]CU: cannabis use.

[b]N/A: not applicable.

[c]PBS: protective behavioral strategy.

XSL•FO
RenderX

Considering the explanatory power of the Ajzen Theory of Planned Behavior (TPB) in behavior change and adoption [40], 4 TPB determinants were targeted to promote CU self-management and PBS use, namely, attitude, perceived behavioral control (self-efficacy), social norms, and intention [40]. In the context of CU, a few studies have shown that higher levels of self-efficacy are significantly associated with increased PBS use [41,42]. Attitude and perceived behavioral control/self-efficacy have also been found to predict the behavioral intention to use PBSs [43].

The intersection of POs and specific behavioral determinants to be targeted is presented in Table 1. The specific change objectives outline what an individual needs to do to optimize CU self-management and PBS use.

Step 3: Selection of Theory-Based Intervention Methods and Practical Applications

In IM, it is crucial to select theory-based intervention methods and practical applications—referred to as behavior change techniques (BCTs) in other approaches [44]—that appropriately align with the targeted determinants [32].

To support this process, a thorough BCT analysis was conducted to identify explicit behavior change mechanisms reported in digital interventions for recreational CU among young adults [17]. The most frequent BCT clusters identified were "Feedback and monitoring," "Goals and planning," "Natural consequences," and "Comparison of outcomes." Feedback on behavior emerged as a core component in nearly all of the CU behavioral interventions analyzed. These findings provided valuable insights for identifying the key active ingredients necessary to develop an effective intervention.

Specific theory-based intervention methods were selected to guide practical applications that effectively address the targeted behavioral determinants and support behavior adoption (ie, PBS use). These methods included personalized feedback, modeling, verbal persuasion, self-monitoring, and positive reinforcement. Examples of these methods and their corresponding practical applications are presented in Table 2.

Table 2. Examples of theory-based intervention methods and their practical applications used to target determinants.

Targeted determinants and theory-based intervention method	Examples of practical applications
Intention	
Goal setting	SMART[a] action plan
Activation of intention	If/then technique
Self-efficacy	
Modeling	Lived experience: CU[b] consequences
Coping planning	Identify obstacles and ways to overcome them
Attitude	
Belief selection	Reflective questions regarding motivation to change
Anticipated outcome	Benefits of adopting a new behavior
Social norms	
Personalized feedback	Feedback regarding CU frequency

[a]SMART: Specific, Measurable, Achievable, Relevant, and Time-bound.
[b]CU: cannabis use.

Integrating results from the initial steps of the IM process led to the development of a logic model of change that illustrates the underlying mechanism. Based on the target behavior and subbehaviors, the determinants and theory-based intervention methods form the active ingredients of the mobile app designed to increase the use of PBSs (principal outcome). Increased PBS use is expected to be associated with a decrease in the frequency of CU (secondary outcome). The determinants (intention and self-efficacy) serve as mediating variables in this model, as illustrated in Multimedia Appendix 3.

Step 4: Creation of Preliminary Intervention Content and Structure

The preliminary structure and content of the intervention were developed based on the matrix of change objectives established in step 2 and the theory-based intervention methods selected in step 3. Subgroups of the research team, along with a working group of potential young adult users, were formed and met multiple times to cocreate the intervention content.

The title of the intervention—Joint Effort—relies on a playful pun (which also works in French) and suggests that users will need to put in some work to change, though they will be supported in doing so.

As detailed in Table 3, the content was presented in 5 sections: (1) Assess—gain awareness of your CU; (2) Mobilize—support your decision to take action; (3) Act—support the establishment of your action plan; (4) Strengthen—consolidate change (booster session); and (5) Observe—monitor your CU.

The intervention was designed to be self-directed and, as such, focused on simple strategies that users could apply independently without relying on external resources.

XSL•FO
RenderX

Table 3. Summary of intervention content.

Section	Aim	Format/components/features	Topics/key content
Assess	Enable users to gain a better awareness of their CU[a]	Questions and personalized feedback	CU frequency, CU motivation, CU consequences, and motivation to change
Mobilize	Support decision-making process	Generic/general messages and questions to reflect upon	PBSs[b], benefits of change, possible difficult situations, and strategies and resources
Act	Support the establishment of an action plan	Generic/general messages, questions (for the SMART[c] action plan), personalized feedback, and questions to reflect upon	SMART action plan, anticipate difficulties and barriers (if/then), and how to overcome obstacles
Strengthen	Support the sustainability of the action plan	Personalized feedback	The SMART action plan (booster)
Observe	Personalized monitoring	Questions and personalized feedback	CU logbook (eg, daily CU frequency, products, open text log) and SMART action plan follow-up

[a]CU: cannabis use.

[b]PBS: protective behavioral strategy.

[c]SMART: Specific, Measurable, Achievable, Relevant, and Time-bound.

Step 5: Focus Groups to Validate Intervention Structure and Examples of Tailored Messages

Overview

Focus groups were conducted to validate the mobile app plan and examples of personalized messages used to develop the prototype.

Participants and Procedures

All participants from the previous focus groups (step 1; n=13) were invited to take part in this validation phase, and 8 agreed (4 were unavailable and 1 did not respond). Three new participants were recruited through ads posted in various Facebook groups targeting students, as well as via snowball sampling. All of these new focus groups were conducted online via Zoom (Zoom Communications), in French, by the same team as in step 1.

Participants discovered the intervention content during the focus group through a PowerPoint (Microsoft Corporation) presentation that summarized the structure and included examples of messages. Questions such as “What do you think of the topics covered?” “What do you think of the sample message?” and “How could it be improved?” were used to lead the discussion.

All meetings were recorded and transcribed, with field notes taken. Participants received CAD $50 (US $36.4) for taking part in the focus groups.

Analysis

The transcripts of the meetings and field notes were summarized. Comments and suggestions were organized thematically. The research team then discussed the results to agree on any changes to be made.

Findings

In December 2020, 11 participants were divided into 4 groups of 2 or 3 based on their availability, and 1 participant was interviewed individually. The mean age was 22.2 (median 22) years, with 7 out of 11 (64%) participants being women. The focus groups lasted between 49 and 115 minutes (mean 83.3 minutes; median 84.5 minutes), while the individual interview lasted 28 minutes. Duration was mainly influenced by the number of participants.

Regarding the main objective, participants emphasized that not all cannabis users may be interested in reducing or modifying their use; some might only want to track it. In this regard, they highlighted the importance of having a logbook option available to all users at all times (ie, not dependent on the completion of previous sections).

Overall, the proposed structure and sequence (ie, 5 sections: assess, mobilize, act, strengthen, and observe) were received positively. Participants appreciated the topics and progression of content, the personalized feedback, and the intervention’s name—Joint Effort.

Moreover, the examples of messages were considered interesting and appropriate. Participants indicated that personalized feedback could help reassure them and normalize their behavior. For example, personalized feedback in the Assess section provided information about CU frequency and motivation based on the results of a population-based survey [45]. For instance, to the question “In the past month, how often have you used cannabis?,” users could receive the following feedback “Frequency: regular. 38% of Quebecers aged 18-24 report having used cannabis in the past year. Among those who use it, just like you, 1 out of 5 did it regularly”. To the question “Thinking back over the past month, for what reason(s) have you used cannabis?,” they could receive the following message “Like you, 9 out of 10 people who regularly use cannabis take it for fun” or “Like you, 96% of people who regularly use cannabis take it to relax”.

The proposed strategies (ie, PBSs) were considered relevant, and the list of examples was appreciated. However, the vocabulary and phrasing received mixed reviews. In some cases, the text felt overly formal and disconnected from the reality experienced by young people.

XSL•FO
RenderX

The themes that emerged from the analysis, along with associated quotations, are presented in Multimedia Appendix 4.

Key Implications for Design

Validating the content and message algorithms was an essential step in the design process before prototyping the mobile app. Participants appreciated the proposed intervention structure, its conciseness and comprehensiveness, as well as the personalized messages, visual feedback, and graphic representations.

Minor changes were made to the intervention structure. For example, the "Observe" section (ie, CU logbook) was initially accessible only after completing the "Assess" section. Based on feedback from focus groups—highlighting that some users might be interested solely in tracking their CU—it was decided to make this section accessible from the start, allowing users to access their logbook without completing any prerequisite sections. Additionally, users could complete the first 3 sections (Assess, Mobilize, and Act) at their convenience; an order of completion was suggested but not imposed.

All intervention content was reviewed to better tailor the examples and language to the target population. A committee representing potential end users, composed of 4 focus group participants, was formed to help fine-tune the intervention. They revised and approved all updated texts and messages in the final validation loop.

Step 6: Transposing the Intervention Content to a Mobile App Prototype

Overview

The intervention content was transposed into a mobile app format. At this stage of the iterative prototype development process, members of the potential end users committee (formed in step 4) were actively involved in validating the prototype.

Intervention Content Scripting and Wireframes

The intervention content validated in step 4 was scripted into short messages organized to create a navigation algorithm (user flow). This content included general information messages, personalized feedback, and questionnaires (checkbox answers and reflection questions) covering various topics. It also featured interactive components, such as a self-monitoring function to track CU. More than 100 wireframes were created using InVision (Dribbble Holdings Ltd) and validated.

Graphic and Visual Design

The graphic identity was designed to appeal to the target audience. Several iterations of the chosen logo and screen mock-ups were produced and validated by the potential end user committee to finalize the graphic identity. Over 50 illustrations were created to support the messages conveyed in the theoretical content of the intervention. Examples of screenshots from the Joint Effort prototype are presented in Figure 1.

Figure 1. Selected screenshots of the Joint Effort mobile app prototype.

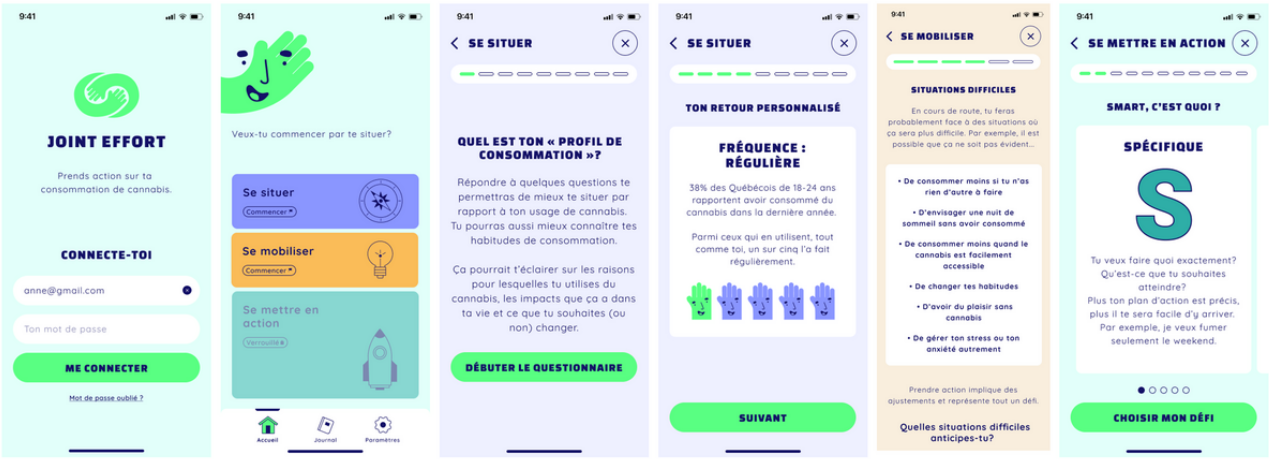

Elements of gamification were integrated into the prototype. A virtual character was created to be at the heart of the user experience, guiding users through various app steps (tutorial, transitions, home page, and logbook). The menu featured locked/unlocked sections to create a playful visual dynamic, and a personalized dashboard allowed users to visualize their progress and see how much of their goal had been achieved.

The user interface was designed with a focus on the user journey. The prototype was created using Figma (Figma, Inc), a collaborative design and prototyping tool [46].

Planned Exposure

In terms of dosage, the intervention was intended to be visited multiple times over a 30-day period. Users could view the content at their own pace (eg, 1 session per day or more than 1 per day). However, they had to wait 7 days before accessing the Booster section.

In terms of duration, the estimated time to complete all the content (ie, Assess, Mobilize, Act, and Strengthen) was about 15 minutes. The time devoted to the Observe section (ie, CU logbook) varied depending on the amount of data input (ie, up to each individual).

Usability Testing

Participant Selection and Sample Size

The inclusion criteria for participants were being a university student at Université de Montréal, being 21-24 years old, and having engaged in any CU in the past 30 days. Recruitment ads

XSL•FO
RenderX

were posted in various Facebook groups targeting students. Interested individuals contacted the study team and had their eligibility confirmed. A sample size of 20-30 participants was targeted, as proposed by Thabane et al [47].

Data Collection

Think-Aloud Interviews

As suggested by Noushad et al [48], we purposefully reflected on the context and research questions before choosing to use the think-aloud method. This method was selected to capture real-time cognitive processes and insights of participants as they performed a specific task (ie, using the mobile prototype). The methodology focuses on asking participants to verbalize their thoughts while completing a task or to recall what was going through their minds after performing it [49]. At that phase, the app was available only as a Figma prototype (ie, accessed via a private URL, with limited features and navigation that needed to be explained). These interviews aimed to understand users' spontaneous reactions to the intervention content, to document how they interacted with the components, and ultimately to inform usability testing.

The interviews were conducted online (via Zoom), in French, by 2 members of the research team (same as in step 1 and 5) with prior experience in qualitative research. Participants were invited to use a prototype of the intervention (via Figma) while thinking out loud about various aspects, including aesthetics, navigability, and reactions to the content. Using a shared link in a web browser, they viewed mock-ups of an iPhone screen featuring visual elements (images and text) and interactive buttons simulating the mobile app.

No interview guide was used. Instead, interviewers encouraged participants to verbalize what they saw, thought, and did while interacting with the prototype. Reformulation and clarification questions (eg, "You mentioned that..., can you tell me more about it?") were asked for explanatory purposes. All interviews were audio-recorded and transcribed.

After each interview, participants were invited to complete a short online sociodemographic questionnaire and the uMARS [33]. Unique links to the questionnaires were sent via LimeSurvey (LimeSurvey Team; Université de Montréal license).

The short sociodemographic questionnaire aimed to collect information to describe the study sample (eg, gender, age, educational level, CU frequency).

The uMARS is a reliable tool for assessing the quality of mHealth apps. It consists of 20 items grouped into 4 quality subscales—engagement, functionality, aesthetics, and information—as well as 1 subjective quality subscale. Each item is rated on a 5-point scale ranging from 1 (inadequate) to 5 (excellent). According to Stoyanov et al [33], the uMARS has demonstrated excellent internal consistency for the full scale (Cronbach α=0.90). To determine an app's quality score on the uMARS, Stoyanov et al [33] suggested calculating the mean score for each of the 4 quality subscales and then averaging these 4 mean scores. The uMARS was translated from English to French following a rigorous process similar to that used for the 17-item version of the Protective Behavioral Strategies for Marijuana Scale [28].

Participants were offered CAD $30 (US $21.8) for taking part in the usability testing.

Data Analysis

The data analysis process for usability testing was the same as that presented in step 1. Descriptive statistics were used to analyze the quantitative data, with Excel software (Microsoft Corporation) used for this purpose.

Ethical Considerations

All steps of the study were approved by the Centre Hospitalier de l'Université de Montréal Research Ethics Board (approval number 20.172) and the Comité d'éthique de la recherche en sciences et en santé (CERSES) of the Université de Montréal (approval number CERSES-20-114-D). All participants provided informed consent. All participants provided informed consent by signing a copy of the informed consent form. Compensation to participants was disclosed in previous sections. Privacy and confidentiality of participants' data or identity were maintained during the study process and publication.

Results

Sociodemographic Characteristics

In May 2021, usability testing was conducted with 20 participants whose mean age was 21.8 (median 22) years. Most were women (14/20, 70%), born in Canada (15/20, 75%), and undergraduates (15/20, 75%). Additionally, most reported engaging in weekly CU in the past 30 days (14/20, 70%).

Think-Aloud Interviews

Overview

A total of 20 think-aloud interviews were conducted, lasting between 20 and 51 minutes. Four themes emerged from the interviews: (1) the mobile app prototype was visually pleasing and easy to use; (2) the content was well adapted to the target audience and nonjudgmental; (3) the customization possibilities were an appreciated feature; and (4) the mobile app was helpful and relevant for initiating behavior change. Examples of participant verbatim quotations are presented in Table 4.

XSL•FO
RenderX

Table 4. Themes and associated quotes (think-aloud analysis).

Themes	Examples of quotes[a]
Visually pleasing and easy to use	• "The colors are nice, soft, pastel." [ID01] • "I personally like the little images, the little drawings, a lot because I find that it brings in a playful side that's not too serious." [ID14] • "For real, I think the app is easy to use, and the information on it is clear and precise." [ID04]
Content well adapted to target audience and nonjudgmental	• "What I liked best was that when I used this app I didn't feel like I was being judged...." [ID10] • "I think it's a good thing that right from the start the intentions are announced, we see that it's really meant to help to take action." [ID11] • "Yes, yes, there are lots of choices, I think that it could represent a fair portion of users." [ID20] • "Yeah, I like that it gives me little facts as well, I find that it normalizes consumption, I don't feel like stigmatized." [ID10]
Customization possibilities, an appreciated feature	• "I like it because so far it really leaves it up to the consumer....You go at it how you feel, the app is there but you go at it at your pace." [ID14] • "The option of being able to go through it all at once or in several parts is good." [ID19]
Relevant to initiate behavior change	• "It allows you to question your own consumption, put it back into perspective, understand why you use, if ever you wanted to change something." [ID01] • "OK, so, strategies before, you see I've been smoking for a really long time, and I've never seen any strategies before....It's really more informative and educational." [ID08] • "The mobile app is already a good step forward and it's a good compromise and it's easy to access." [ID15] • "I think that it can be just as useful to someone who's just trying to stop though it's more to gain some insight on the bad effects of your consumption...even if you don't necessarily stop completely, just understanding and knowing your use, it's all good." [ID3]

[a]Quotes were translated from French to English by a specialized translator and validated for accuracy by the research team.

Visually Pleasing and Easy to Use

Participants expressed positive feedback regarding the app's visuals. The visual layout was considered uncluttered, and the neutral, soft, and soothing color scheme was appreciated. The images and virtual characters were seen as playful and amusing without being childish. Participants found the app easy to use, clear, and accessible. Navigation was described as fluid and intuitive, with participants finding it easy to find their way around.

Well-Adapted to Target Audience and Nonjudgmental

The neutral tone and nonjudgmental wording contributed to a nonmoralizing aspect of the app. As the app's goals were clearly stated in the introduction (ie, "aims to help you take action on your cannabis use, not to convince you to stop or change at any cost"), participants felt reassured about its purpose. In the personalized feedback, the comprehensive statistics helped normalize CU without stigmatizing it. When selecting answers from a set list, participants felt the choices reflected their reality and were well adapted to the target audience.

Customization Possibilities, an Appreciated Feature

The various customization aspects of the app were highly appreciated. For instance, the option to choose or write personalized goals was particularly appealing to some participants. They also liked the flexibility of going through the content at their own pace, as they could either complete most sections in a single sitting or pause and return to them later if they wished.

Relevant to Initiate Behavior Change

Participants highlighted various benefits of the app for all types of users. It offered support and monitoring for those wishing to initiate a process of behavior change. The information and strategies presented in the app were considered relevant and educational, and were seen as encouraging users to reflect on their CU. Participants also noted that the app could help facilitate access to other services.

uMARS

After the think-aloud portion of the interview, participants (n=20) were sent a link to complete the uMARS. Results per item and per subscale are presented in Table 5. The Joint Effort prototype received the following mean scores on the 4 subscales: 4.14/5.0 (SD 0.53) for Engagement, 4.60/5.0 (SD 0.47) for Functionality, 4.53/5.0 (SD 0.52) for Aesthetics, and 4.44/5.0 (SD 0.61) for Information. This yielded an overall app quality mean score of 4.43/5.0 (SD 0.53). The mean scores on the subjective items varied widely. The item "Would you recommend this app to people who might benefit from it?" received the highest score at 4.35/5.0 (SD 0.67), while "Would you pay for this app?" received the lowest at 1.65/5.0 (SD 1.09).

XSL•FO
RenderX

Table 5. Results on the User Version of the Mobile Application Rating Scale (N=20).

Subscale and item	Mean (SD)	Median (range[a])
Engagement	4.14 (0.53)	N/A[b]
1. Entertainment: Is the app fun/entertaining to use? Does it have components that make it more fun than other similar apps?	4.05 (0.51)	4 (3-5)
2. Interest: Is the app interesting to use? Does it present its information in an interesting way compared to other similar apps?	4.60 (0.50)	5 (4-5)
3. Customization: Does it allow you to customize the settings and preferences that you would like to (eg, sound, content, and notifications)?	3.80 (0.89)	4 (2-5)
4. Interactivity: Does it allow user input, provide feedback, and contain prompts (reminders, sharing options, notifications, etc)?	3.65 (1.09)	4 (1-5)
5. Target group: Is the app content (visuals, language, and design) appropriate for the target audience?	4.60 (0.68)	5 (3-5)
Functionality	4.60 (0.47)	N/A
6. Performance[c]: How accurately/fast do the app features (functions) and components (buttons/menus) work?	N/A	N/A
7. Ease of use: How easy is it to learn how to use the app? How clear are the menu labels, icons, and instructions?	4.75 (0.55)	5 (3-5)
8. Navigation: Does moving between screens make sense? Does the app have all the necessary links between screens?	4.60 (0.50)	5 (4-5)
9. Gestural design: Do taps/swipes/pinches/scrolls make sense? Are they consistent across all components/screens?	4.45 (0.69)	5 (3-5)
Aesthetics	4.53 (0.52)	N/A
10. Layout: Is the arrangement and size of buttons, icons, menus, and content on the screen appropriate?	4.65 (0.67)	5 (3-5)
11. Graphics: How high is the quality/resolution of graphics used for buttons, icons, menus, and content?	4.70 (0.57)	5 (3-5)
12. Visual appeal: How good does the app look?	4.25 (0.64)	4 (3-5)
Information	4.44 (0.61)	N/A
13. Quality of information: Is app content correct, well written, and relevant to the goal/topic of the app?	4.50 (0.69)	5 (3-5)
14. Quantity of information: Is the information within the app comprehensive but concise?	4.20 (0.95)	4 (1-5)
15. Visual information: Is visual explanation of concepts—through charts/graphs/images/videos, etc—clear, logical, and correct?	4.70 (0.47)	5 (4-5)
16. Credibility of source: Does the information within the app seem to come from a credible source?	4.35 (0.93)	5 (2-5)
Subjective items	3.36 (0.53)	N/A
17. Would you recommend this app to people who might benefit from it?[d]	4.35 (0.67)	4 (3-5)
18. How many times do you think you would use this app in the next 12 months if it was relevant to you?[e]	3.25 (1.02)	3 (1-5)
19. Would you pay for this app?[f]	1.65 (1.09)	1 (1-4)
20. What is your overall (star) rating of the app?[g]	4.20 (0.41)	4 (4-5)

[a]Possible score range: 1-5.

[b]N/A: not applicable.

[c]The item “Performance” (How accurately/fast do the app features [functions] and components [buttons/menus] work?) could not be evaluated on the Figma prototype.

[d]Choices of answer ranging from 1 “not at all” to 5 “definitely.”

[e]Choices of answer: 1=“none”; 2=“1-2”; 3=“3-10”; 4=“10-50”; 5=“>50.”

[f]Choices of answer ranging from 1=“definitely not” to 5=“definitely yes.”

[g]Choices of answer ranging from 1=“1 star” to 5=“5 stars.”

XSL•FO
RenderX

Discussion

Principal Findings

The objective of this study was to codevelop and conduct usability testing of a mobile app prototype, aimed at promoting CU self-management and reinforcing PBS use among university students.

Focus groups were conducted to identify the needs and preferences of young adults regarding CU intervention. The data collected during this phase informed the development of the intervention content. A second round of focus groups was then held to validate the structure of the intervention and examples of tailored messages it might convey. Subsequently, the intervention content was transposed into digital form through iterative software development, and the mobile app prototype was created. A third validation phase was carried out using think-aloud interviews and a short questionnaire to document the user experience.

Usability findings suggest that participants appreciated the app. The prototype received an overall app quality score of 4.43/5. The mean scores on the uMARS subscales were relatively high for Engagement (4.14/5), Functionality (4.60/5), Information (4.44/5), and Aesthetics (4.53/5). Interestingly, the item that received the lowest rating was “Would you pay for this app?” (1.65/5.0, SD 1.09). This suggests that, despite the app being perceived positively, participants were not willing to pay a subscription fee. A systematic review has shown that offering health-related apps for free or at a low cost can positively influence their uptake and engagement [50].

The results of the think-aloud interviews (qualitative component) aligned with the uMARS findings. Participants found navigation smooth and intuitive, which made it easy for them to find their way around, and the content conducive to prompting behavior change. A systematic review and thematic synthesis (n=35 studies) of mHealth interventions identified reinforcement, communication, navigation, credibility, message presentation, and interface aesthetics as key design features to consider for improving user engagement [51]. Given the strong uMARS scores obtained for these components, it is reasonable to expect that the Joint Effort app may successfully elicit user engagement.

However, comparing our uMARS results with those of other studies is challenging. Although this tool has been used elsewhere, most mobile apps evaluated with the uMARS differ substantially from Joint Effort. To date, in the field of cannabis and other substance use research, the use of uMARS has primarily been mentioned in protocols [52-54]. Only Santesteban-Echarri et al [55] reported using the uMARS to evaluate a cannabis-related mobile app for youth at high risk for psychosis. They found a good overall score (3.75/5) and high subscale ratings for esthetics (4.48/5), information (4.32/5), and functionality (4.29/5) [55]. In comparison, the Joint Effort prototype received a higher overall app quality score (4.43/5). The slightly higher scores obtained in our study may be attributed to the IM and co-design approach used, along with multiple validation rounds. The target population was involved from the outset (ie, needs analysis), and potential end users were engaged throughout the process. However, these comparisons should be interpreted with caution, as subjectivity is inherent when using a scale such as the uMARS. For instance, aesthetic preferences may vary across cultural groups. Moreover, the uMARS was completed after a single use of the app prototype, rather than after extended use of its final version.

The aim of the mobile app prototype developed is to promote PBS use, which involves setting a goal, engaging in a change process, and implementing an action plan (SMART, take action). Our qualitative results suggest that participants appreciated the opportunity to engage at their own pace, set their own goals, and initiate the behavior change they desired, if any. In a rapid review of the literature (n=43 studies), Monarque et al [56] concluded that the uptake of digital interventions among youth depended on the incorporation of harm reduction principles and skills training. These approaches were favored in the development of Joint Effort, as participants were supported in self-managing their CU by reinforcing PBS use, rather than being encouraged to quit altogether.

Different approaches were used to support CU self-management and reinforce PBS use in developing the app, including personalized feedback, anticipated outcomes, positive reinforcement, goal setting, self-observation, and activation of intention. In a systematic review examining the efficacy of behavior change smartphone apps (n=27 studies), Schoeppe et al [57] emphasized common strategies frequently used in successful app-based interventions, such as goal setting, self-monitoring, and performance feedback. These findings were later reinforced in the systematic review by Milne-Ives et al [58], which identified 6 common behavior change techniques associated with user engagement in mHealth apps: goal setting, self-monitoring of behavior, feedback on behavior, prompts/cues, rewards, and social support. In light of the above, the Joint Effort mobile app prototype was designed using the most effective strategies to promote engagement and intervention success. These strategies will be evaluated in subsequent research phases, with a focus on both experiential and behavioral engagement [59].

An unexpected element emerged from the qualitative interviews regarding the nonjudgmental approach used. Participants reported not feeling judged or stigmatized and expressed that their behavior felt normalized. This is particularly important given the ongoing stigmatization of CU, both in jurisdictions where cannabis is legal and where it remains illegal [4,60]. In terms of intervention access, several barriers have been documented, including limited availability, confidentiality concerns, and stigma [61]. The intervention was developed with the intention of serving as a tool for self-directed prevention, offering simple and achievable strategies that require no reliance on external resources. This could address a specific need among young adults who use cannabis and are interested in changing their consumption habits. The advantages of interventions delivered via mobile apps are well documented in the literature. For example, mobile apps provide an economical and easily accessible means of delivering low-intensity interventions for mental health–related issues [62]. Anonymity, asynchrony, and easy, immediate access (without the need to leave home) were the advantages most appreciated by participants in this study.

XSL•FO
RenderX

Strengths and Limitations

The principal strength of this study lies in the rigorous, theory-based process used to cocreate the mobile app prototype. IM is a robust framework for developing theory- and evidence-based interventions [32]. Multiple methods were used to conduct an extensive needs analysis, including a systematic review [17], an online CU study [28,63], and focus groups. The overall IM process involved the continuous integration and analysis of 3 types of knowledge: empirical (previous studies and needs analysis), experiential (input from key informants), and theoretical (behavioral determinants). The iterative codevelopment process included multiple phases of consultation with potential end users of the mobile app and several validation loops to ensure the intervention aligned with user preferences and needs. Finally, the study was conducted by a well-rounded, thoughtfully assembled interdisciplinary team of researchers with diverse expertise, including cannabis, information and communication technologies, health promotion, youth mental health, knowledge transfer, and mixed methods research (both quantitative and qualitative), working in collaboration with potential end users.

A key limitation of the study is the overwhelmingly positive user reactions and reviews received during the think-aloud interviews. Participants did not mention any aspects they disliked or suggested elements for improvement at that time. This may be attributed to the methodology used: given the stage of the design process, participants interacted with the prototype for the first and only time during the interview. As such, the results reflected their initial general impressions. A more extended and in-depth use over time may have elicited more nuanced feedback.

Another limitation is the predominance of female participants across the focus groups conducted to identify user needs (11/13, 85%), validate the intervention content (7/11, 64%), and during the usability testing phase (14/20, 70%). Given that negative CU consequences are disproportionately concentrated in the young adult male population [7,29], it is reasonable to assume that a higher proportion of males with different viewpoints might have led to a slightly different intervention. In sum, these findings highlight the challenges of developing effective mobile app interventions for CU that account for gender differences across various clinical correlates (eg, readiness to change, sources of motivation for behavior change) that may influence young adults in their decision to take steps toward better managing their CU.

Conclusions

In the context of the recent legalization of cannabis and its widespread consumption, strategies to promote safer use and reduce harm need to extend beyond traditional abstinence-based approaches. mHealth interventions appear promising in this regard, particularly in addressing challenges related to health care access and stigmatization. Following this study, and given the promising results from usability testing, we proceeded with the further development of the Joint Effort mobile app. While the work described in this manuscript yielded positive signals from potential end users, further evaluation is needed to confirm its acceptability, feasibility, and ultimately, its efficacy in terms of CU outcomes. A pilot randomized trial [64] was conducted to assess the intervention's acceptability (user uptake, user engagement, user-participant profiles, and intervention appreciation) and to document the feasibility of the study process (online recruitment rate, adherence to online data collection methods, and attrition rate). Based on the logic model of change (Multimedia Appendix 3), an RCT is ongoing to evaluate the intervention's efficacy [65] on PBSs as the primary outcome, with frequency of CU as a secondary outcome, and intention to take action on CU as a mediator outcome. If proven efficacious, Joint Effort could help diversify the available tools to improve CU outcomes.

Acknowledgments

This project was conducted in collaboration with the Government of Québec and the Québec Ministry of Health and Social Services. The views and opinions expressed in this manuscript do not necessarily reflect those of the funding entities. J. Côté is the Chairholder of the Research Chair in Innovative Nursing Practices. J. Cheng is the Chairholder of the Canada Research Chair in User Experience Design for Data-Driven Systems. GC is the recipient of a Postdoctoral Fellowship Award from the Canadian Institutes of Health Research (CIHR). GF is the recipient of a Junior 1 Research Scholar Award from the Fonds de recherche du Québec – Santé (FRQS). SL is supported by the Canada Research Chairs program. MGP is a Junior 2 Research Scholar supported by the Fonds de la recherche du Québec – Santé. DJA is a Senior Clinical Research Scholar supported by the Fonds de la recherche du Québec – Santé. We thank Emma Février for her support with data collection and analysis, and Amel Baghdadli for her contribution to data analysis. We also thank our corporate partners, Akufen and Osedea, and, last but not least, the students who participated in the research.

Authors' Contributions

J Côté and PA contributed to conceptualization, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, supervision, visualization, and writing—original draft. J Cheng, GC, SC, GF, CG, SL, JL, MGP, MAMC, GR, and BV contributed to conceptualization, funding acquisition, methodology, and writing—review and editing. DJA contributed to conceptualization, funding acquisition, methodology, supervision, and writing—review and editing.

XSL•FO
RenderX

Conflicts of Interest

DJA received study materials from Cardiol Therapeutics for a clinical trial funded by the Québec Ministry of Health and Social Services. All other authors declare no conflicts of interest.

Multimedia Appendix 1
COREQ checklist.
[PDF File (Adobe PDF File), 1107 KB - formative_v9i1e71924_app1.pdf]

Multimedia Appendix 2
Results from focus groups to identify needs and preferences (step 1).
[PDF File (Adobe PDF File), 131 KB - formative_v9i1e71924_app2.pdf]

Multimedia Appendix 3
The logic model of change (step 3).
[PDF File (Adobe PDF File), 46 KB - formative_v9i1e71924_app3.pdf]

Multimedia Appendix 4
Results from focus groups to validate intervention structure and examples of tailored messages (step 5).
[PDF File (Adobe PDF File), 145 KB - formative_v9i1e71924_app4.pdf]

References

1. Degenhardt L, Stockings E, Patton G, Hall WD, Lynskey M. The increasing global health priority of substance use in young people. Lancet Psychiatry 2016 Mar;3(3):251-264. [doi: 10.1016/S2215-0366(15)00508-8] [Medline: 26905480]
2. Chan GCK, Becker D, Butterworth P, Hines L, Coffey C, Hall W, et al. Young-adult compared to adolescent onset of regular cannabis use: a 20-year prospective cohort study of later consequences. Drug Alcohol Rev 2021 May;40(4):627-636. [doi: 10.1111/dar.13239] [Medline: 33497516]
3. Rotermann M, Macdonald R. Analysis of trends in the prevalence of cannabis use in Canada, 1985 to 2015. Health Rep 2018 Feb 21;29(2):10-20 [FREE Full text] [Medline: 29465739]
4. Kourgiantakis T, Lee E, Kosar AKT, Tait C, Lau CKY, McNeil S, et al. Youth cannabis use in Canada post-legalization: service providers' perceptions, practices, and recommendations. Subst Abuse Treat Prev Policy 2023 Jun 22;18(1):36. [doi: 10.1186/s13011-023-00550-1] [Medline: 37349741]
5. Hall W, Stjepanović D, Caulkins J, Lynskey M, Leung J, Campbell G, et al. Public health implications of legalising the production and sale of cannabis for medicinal and recreational use. Lancet 2019 Oct 26;394(10208):1580-1590. [doi: 10.1016/S0140-6736(19)31789-1] [Medline: 31657733]
6. Manthey J, Jacobsen B, Hayer T, Kalke J, López-Pelayo H, Pons-Cabrera MT, et al. The impact of legal cannabis availability on cannabis use and health outcomes: a systematic review. Int J Drug Policy 2023 Jun;116:104039. [doi: 10.1016/j.drugpo.2023.104039] [Medline: 37126997]
7. Fischer B, Robinson T, Bullen C, Curran V, Jutras-Aswad D, Medina-Mora ME, et al. Lower-Risk Cannabis Use Guidelines (LRCUG) for reducing health harms from non-medical cannabis use: a comprehensive evidence and recommendations update. Int J Drug Policy 2022 Jan;99:103381 [FREE Full text] [doi: 10.1016/j.drugpo.2021.103381] [Medline: 34465496]
8. Bonar EE, Goldstick JE, Tan CY, Bourque C, Carter PM, Duval ER, et al. A remote brief intervention plus social media messaging for cannabis use among emerging adults: a pilot randomized controlled trial in emergency department patients. Addict Behav 2023 Dec;147:107829. [doi: 10.1016/j.addbeh.2023.107829] [Medline: 37598642]
9. Boumparis N, Loheide-Niesmann L, Blankers M, Ebert DD, Korf D, Schaub MP, et al. Short- and long-term effects of digital prevention and treatment interventions for cannabis use reduction: a systematic review and meta-analysis. Drug Alcohol Depend 2019 Jul 01;200:82-94. [doi: 10.1016/j.drugalcdep.2019.03.016] [Medline: 31112834]
10. Halladay J, Scherer J, MacKillop J, Woock R, Petker T, Linton V, et al. Brief interventions for cannabis use in emerging adults: a systematic review, meta-analysis, and evidence map. Drug Alcohol Depend 2019 Nov 01;204:107565. [doi: 10.1016/j.drugalcdep.2019.107565] [Medline: 31751868]
11. Hoch E, Preuss UW, Ferri M, Simon R. Digital interventions for problematic cannabis users in non-clinical settings: findings from a systematic review and meta-analysis. Eur Addict Res 2016;22(5):233-242. [doi: 10.1159/000445716] [Medline: 27160333]
12. Imtiaz S, Roerecke M, Kurdyak P, Samokhvalov AV, Hasan OSM, Rehm J. Brief interventions for cannabis use in healthcare settings: systematic review and meta-analyses of randomized trials. J Addict Med 2020;14(1):78-88. [doi: 10.1097/ADM.0000000000000527] [Medline: 32012140]

XSL•FO
RenderX

13. Olmos A, Tirado-Muñoz J, Farré M, Torrens M. The efficacy of computerized interventions to reduce cannabis use: a systematic review and meta-analysis. Addict Behav 2018 Apr;79:52-60. [doi: 10.1016/j.addbeh.2017.11.045] [Medline: 29248863]
14. Tait RJ, Spijkerman R, Riper H. Internet and computer based interventions for cannabis use: a meta-analysis. Drug Alcohol Depend 2013 Dec 01;133(2):295-304. [doi: 10.1016/j.drugalcdep.2013.05.012] [Medline: 23747236]
15. Massarat ERG. Teens, Social Media and Technology. Pew Research Center. 2022. URL: https://www.pewresearch.org/internet/2022/08/10/teens-social-media-and-technology-2022/ [accessed 2024-10-29]
16. Ferretti A, Hubbs S, Vayena E. Global youth perspectives on digital health promotion: a scoping review. BMC Digit Health 2023 Jul 25;1(1):25. [doi: 10.1186/s44247-023-00025-0]
17. Côté J, Chicoine G, Vinette B, Auger P, Rouleau G, Fontaine G, et al. Digital interventions for recreational cannabis use among young adults: systematic review, meta-analysis, and behavior change technique analysis of randomized controlled studies. J Med Internet Res 2024 Apr 17;26:e55031 [FREE Full text] [doi: 10.2196/55031] [Medline: 38630515]
18. Sedrati H, Belrhiti Z, Nejjari C, Ghazal H. Evaluation of mobile health apps for non-medical cannabis use: a scoping review. Procedia Computer Science 2022;196:581-589. [doi: 10.1016/j.procs.2021.12.052]
19. Albertella L, Gibson L, Rooke S, Norberg MM, Copeland J. A smartphone app intervention for adult cannabis users wanting to quit or reduce their use: a pilot evaluation. J Cannabis Res 2019 Aug 16;1(1):9 [FREE Full text] [doi: 10.1186/s42238-019-0009-6] [Medline: 33526112]
20. Coughlin LN, Nahum-Shani I, Philyaw-Kotov ML, Bonar EE, Rabbi M, Klasnja P, et al. Developing an adaptive mobile intervention to address risky substance use among adolescents and emerging adults: usability study. JMIR Mhealth Uhealth 2021 Jan 15;9(1):e24424. [doi: 10.2196/24424] [Medline: 33448931]
21. Kapitány-Fövény M, Vagdalt E, Ruttkay Z, Urbán R, Richman MJ, Demetrovics Z. Potential of an interactive drug prevention mobile phone app (Once Upon a High): questionnaire study among students. JMIR Serious Games 2018 Dec 04;6(4):e19 [FREE Full text] [doi: 10.2196/games.9944] [Medline: 30514697]
22. Monney G, Penzenstadler L, Dupraz O, Etter J, Khazaal Y. mHealth app for cannabis users: satisfaction and perceived usefulness. Front Psychiatry 2015;6:120 [FREE Full text] [doi: 10.3389/fpsyt.2015.00120] [Medline: 26379561]
23. Prince MA, Collins RL, Wilson SD, Vincent PC. A preliminary test of a brief intervention to lessen young adults' cannabis use: episode-level smartphone data highlights the role of protective behavioral strategies and exercise. Exp Clin Psychopharmacol 2020 Apr;28(2):150-156 [FREE Full text] [doi: 10.1037/pha0000301] [Medline: 31144836]
24. Vederhus J, Rørendal M, Bjelland C, Skar AKS, Kristensen. Can a smartphone app for cannabis cessation gain a broader user group than traditional treatment services? Subst Abuse 2020;14:1178221820902237 [FREE Full text] [doi: 10.1177/1178221820902237] [Medline: 32071540]
25. Santesteban Echarri O, Kim G, Haffey P, Tang J, Addington J. LooseLeaf, a mobile-based application to monitor cannabis use and cannabis-related experiences for youth at clinical high-risk for psychosis: development and user acceptance testing. International Journal of Human–Computer Interaction 2020 Oct 20;37(6):501-511. [doi: 10.1080/10447318.2020.1832744]
26. Bravo AJ, Anthenien AM, Prince MA, Pearson MR, Marijuana Outcomes Study Team. Marijuana protective behavioral strategies as a moderator of the effects of risk/protective factors on marijuana-related outcomes. Addict Behav 2017 Jun;69:14-21. [doi: 10.1016/j.addbeh.2017.01.007] [Medline: 28110137]
27. Bravo AJ, Prince MA, Pearson MR, Marijuana Outcomes Study Team. Can I use marijuana safely? An examination of distal antecedents, marijuana protective behavioral strategies, and marijuana outcomes. J Stud Alcohol Drugs 2017 Mar;78(2):203-212. [doi: 10.15288/jsad.2017.78.203] [Medline: 28317500]
28. Côté J, Cossette S, Auger P, Page G, Coronado-Montoya S, Fontaine G, et al. Psychometric properties of the French and English short form of the Protective Behavioural Strategies for Marijuana Scale in Canadian university students. BMJ Open 2022 Apr 06;12(4):e053715 [FREE Full text] [doi: 10.1136/bmjopen-2021-053715] [Medline: 35387810]
29. Grigsby TJ, Lopez A, Albers L, Rogers CJ, Forster M. A scoping review of risk and protective factors for negative cannabis use consequences. Subst Abuse 2023;17:11782218231166622. [doi: 10.1177/11782218231166622] [Medline: 37056398]
30. Mian MN, Altman BR, Luba R, Ueno LF, Dalal D, Earleywine M. Factor analysis of a short form of the Protective Behavioral Strategies for Marijuana scale. Addict Behav 2021 Jun;117:106852 [FREE Full text] [doi: 10.1016/j.addbeh.2021.106852] [Medline: 33581678]
31. Lewis MA, Litt DM, Fairlie AM, Kilmer JR, Kannard E, Resendiz R, et al. Investigating why and how young adults use protective behavioral strategies for alcohol and marijuana use: protocol for developing a randomized controlled trial. JMIR Res Protoc 2022 Apr 19;11(4):e37106 [FREE Full text] [doi: 10.2196/37106] [Medline: 35438642]
32. Bartholomew L, Markham C, Ruiter R, Fernández M, Kok G, Parcel G. Planning Health Promotion Programs: An Intervention Mapping Approach. Hoboken, NJ: Jossey-Bass; 2016.
33. Stoyanov SR, Hides L, Kavanagh DJ, Wilson H. Development and validation of the user version of the Mobile Application Rating Scale (uMARS). JMIR Mhealth Uhealth 2016 Jun 10;4(2):e72 [FREE Full text] [doi: 10.2196/mhealth.5849] [Medline: 27287964]
34. Bradbury K, Watts S, Arden-Close E, Yardley L, Lewith G. Developing digital interventions: a methodological guide. Evid Based Complement Alternat Med 2014;2014:561320 [FREE Full text] [doi: 10.1155/2014/561320] [Medline: 24648848]

XSL•FO
RenderX

35. Tong A, Sainsbury P, Craig J. Consolidated Criteria for Reporting Qualitative Research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
36. Hall W, Stjepanović D, Dawson D, Leung J. The implementation and public health impacts of cannabis legalization in Canada: a systematic review. Addiction 2023 Nov;118(11):2062-2072 [FREE Full text] [doi: 10.1111/add.16274] [Medline: 37380613]
37. Peterson R, Kramer MP, Pinto D, De Leon AN, Leary AV, Marin AA, et al. A comprehensive review of measures of protective behavioral strategies across various risk factors and associated PBS-related interventions. Exp Clin Psychopharmacol 2021 Jun;29(3):236-250. [doi: 10.1037/pha0000498] [Medline: 34264735]
38. Pedersen ER, Hummer JF, Rinker DV, Traylor ZK, Neighbors C. Measuring protective behavioral strategies for marijuana use among young adults. J Stud Alcohol Drugs 2016 May;77(3):441-450. [doi: 10.15288/jsad.2016.77.441] [Medline: 27172576]
39. Paillé P, Mucchielli A. L'analyse Qualitative en Sciences Humaines et Sociales. Paris, France: Armand Colin; 2016.
40. Godin G, Kok G. The theory of planned behavior: a review of its applications to health-related behaviors. Am J Health Promot 1996;11(2):87-98. [doi: 10.4278/0890-1171-11.2.87] [Medline: 10163601]
41. Hetelekides EM, Tyskiewicz AJ, Prince MA, Collins RL. The indirect effects of self-efficacy on cannabis use via cannabis protective strategies: a preliminary study. Cannabis 2018;1(2):66-76 [FREE Full text] [doi: 10.26828/cannabis.2018.02.007] [Medline: 34327312]
42. Pearson MR, Bravo AJ, Protective Strategies Study Team. Marijuana protective behavioral strategies and marijuana refusal self-efficacy: independent and interactive effects on marijuana-related outcomes. Psychol Addict Behav 2019 Jun;33(4):412-419 [FREE Full text] [doi: 10.1037/adb0000445] [Medline: 30869920]
43. Ciolino J. The role of attitude, subjective norms, behavioral control, and personality in predicting intention and use of protective strategies for cannabis. MS Thesis. Illinois State University. Illinois State University. 2024. URL: https://ir.library.illinoisstate.edu/etd/2012 [accessed 2025-04-28]
44. Michie S, Carey RN, Johnston M, Rothman AJ, de Bruin M, Kelly MP, et al. From theory-inspired to theory-based interventions: a protocol for developing and testing a methodology for linking behaviour change techniques to theoretical mechanisms of action. Ann Behav Med 2018 May 18;52(6):501-512. [doi: 10.1007/s12160-016-9816-6] [Medline: 27401001]
45. Conus F, Street M. Enquête québécoise sur le cannabis 2019. La consommation de cannabis et les perceptions des Québécois. Portrait et comparaison avec l'édition de 2018. Institut de la statistique du Québe. 2020. URL: https://statistique.quebec.ca/fr/document/enquete-quebecoise-sur-le-cannabis-2019 [accessed 2025-06-16]
46. Figma: the collaborative interface design tool. Figma, Inc. 2016. URL: https://www.figma.com/ [accessed 2021-03-02]
47. Thabane L, Ma J, Chu R, Cheng J, Ismaila A, Rios LP, et al. A tutorial on pilot studies: the what, why and how. BMC Med Res Methodol 2010 Jan 06;10:1. [doi: 10.1186/1471-2288-10-1] [Medline: 20053272]
48. Noushad B, Van Gerven PWM, de Bruin ABH. Twelve tips for applying the think-aloud method to capture cognitive processes. Med Teach 2024 Jul;46(7):892-897 [FREE Full text] [doi: 10.1080/0142159X.2023.2289847] [Medline: 38071621]
49. Charters E. The use of think-aloud methods in qualitative research an introduction to think-aloud methods. Brock Education Journal 2003 Jul 01;12(2):68-82. [doi: 10.26522/brocked.v12i2.38]
50. Szinay D, Jones A, Chadborn T, Brown J, Naughton F. Influences on the uptake of and engagement with health and well-being smartphone apps: systematic review. J Med Internet Res 2020 May 29;22(5):e17572 [FREE Full text] [doi: 10.2196/17572] [Medline: 32348255]
51. Wei Y, Zheng P, Deng H, Wang X, Li X, Fu H. Design features for improving mobile health intervention user engagement: systematic review and thematic analysis. J Med Internet Res 2020 Dec 09;22(12):e21687 [FREE Full text] [doi: 10.2196/21687] [Medline: 33295292]
52. Manning V, Piercy H, Garfield JBB, Lubman DI. Personalized approach bias modification smartphone app ("SWIPE") to reduce alcohol use among people drinking at hazardous or harmful levels: protocol for an open-label feasibility study. JMIR Res Protoc 2020 Aug 14;9(8):e21278 [FREE Full text] [doi: 10.2196/21278] [Medline: 32795989]
53. Hides L, Baker A, Norberg M, Copeland J, Quinn C, Walter Z, et al. A web-based program for cannabis use and psychotic experiences in young people (Keep It Real): protocol for a randomized controlled trial. JMIR Res Protoc 2020 Jul 29;9(7):e15803 [FREE Full text] [doi: 10.2196/15803] [Medline: 32723727]
54. Serre F, Moriceau S, Donnadieu L, Forcier C, Garnier H, Alexandre J, for Craving-Manager RCT investigator group. The Craving-Manager smartphone app designed to diagnose substance use/addictive disorders, and manage craving and individual predictors of relapse: a study protocol for a multicenter randomized controlled trial. Front Psychiatry 2023;14:1143167 [FREE Full text] [doi: 10.3389/fpsyt.2023.1143167] [Medline: 37255691]
55. Santesteban-Echarri O, Haffey P, Kim G, Tang J, Addington J. A mobile-based app to monitor cannabis use among youth at clinical high risk (CHR) for psychosis: feasibility and acceptability of LooseLeaf. Schizophr Res 2020 Aug;222:505-506. [doi: 10.1016/j.schres.2020.05.031] [Medline: 32522468]
56. Monarque M, Sabetti J, Ferrari M. Digital interventions for substance use disorders in young people: rapid review. Subst Abuse Treat Prev Policy 2023 Feb 17;18(1):13. [doi: 10.1186/s13011-023-00518-1] [Medline: 36805783]

XSL•FO
RenderX

57. Schoeppe S, Alley S, Van Lippevelde W, Bray NA, Williams SL, Duncan MJ, et al. Efficacy of interventions that use apps to improve diet, physical activity and sedentary behaviour: a systematic review. Int J Behav Nutr Phys Act 2016 Dec 07;13(1):127. [doi: 10.1186/s12966-016-0454-y] [Medline: 27927218]
58. Milne-Ives M, Homer SR, Andrade J, Meinert E. Potential associations between behavior change techniques and engagement with mobile health apps: a systematic review. Front Psychol 2023;14:1227443 [FREE Full text] [doi: 10.3389/fpsyg.2023.1227443] [Medline: 37794916]
59. Perski O, Blandford A, West R, Michie S. Conceptualising engagement with digital behaviour change interventions: a systematic review using principles from critical interpretive synthesis. Transl Behav Med 2017 Jun;7(2):254-267 [FREE Full text] [doi: 10.1007/s13142-016-0453-1] [Medline: 27966189]
60. King DD, Gill CJ, Cadieux CS, Singh N. The role of stigma in cannabis use disclosure: an exploratory study. Harm Reduct J 2024 Jan 26;21(1):21 [FREE Full text] [doi: 10.1186/s12954-024-00929-8] [Medline: 38273362]
61. Copeland J, Rooke S, Rodriquez D, Norberg MM, Gibson L. Comparison of brief versus extended personalised feedback in an online intervention for cannabis users: short-term findings of a randomised trial. J Subst Abuse Treat 2017 May;76:43-48. [doi: 10.1016/j.jsat.2017.01.009] [Medline: 28162849]
62. Linardon J, Cuijpers P, Carlbring P, Messer M, Fuller-Tyszkiewicz M. The efficacy of app-supported smartphone interventions for mental health problems: a meta-analysis of randomized controlled trials. World Psychiatry 2019 Oct;18(3):325-336 [FREE Full text] [doi: 10.1002/wps.20673] [Medline: 31496095]
63. Côté J, Auger P, Pagé GM, Cossette S, Coronado-Montoya S, Fontaine G, et al. Traduction française, adaptation culturelle et évaluation des propriétés psychométriques préliminaires de l'échelle des stratégies de protection comportementale liées à la consommation de cannabis. Can J Psychiatry 2022 Aug;67(8):608-615 [FREE Full text] [doi: 10.1177/07067437211025216] [Medline: 34160302]
64. Joint Effort Study (JointEffort). ClinicalTrials.org. URL: https://clinicaltrials.gov/ct2/show/NCT05099016 [accessed 2025-06-16]
65. Joint Effort 2.0: a mobile application with university students (JOINTEFFORT). ClinicalTrials.gov. URL: https://clinicaltrials.gov/ct2/show/NCT05620433 [accessed 2025-06-12]

Abbreviations

APPT: Assess, Plan, Track, and Tips
CERSES: Comité d'éthique de la recherche en sciences et en santé
COREQ: Consolidated Criteria for Reporting Qualitative Research
CU: cannabis use
IM: intervention mapping
MApp: Marijuana Smartphone App
mHealth: mobile health
PBS: protective behavioral strategy
PO: performance objective
RCT: randomized controlled trial
TPB: Theory of Planned Behavior
uMARS: User Version of the Mobile Application Rating Scale

Edited by A Bucher; submitted 31.01.25; peer-reviewed by Y Cui; comments to author 23.04.25; revised version received 13.05.25; accepted 24.05.25; published 23.06.25.

Please cite as:
Côté J, Auger P, Chicoine G, Cheng J, Cossette S, Fontaine G, Genest C, Lal S, Lapierre J, Pagé MG, Maheu-Cadotte MA, Rouleau G, Vinette B, Jutras-Aswad D
A Mobile App (Joint Effort) to Support Cannabis Use Self-Management and Reinforce the Use of Protective Behavioral Strategies: Development Process and Usability Testing
JMIR Form Res 2025;9:e71924
URL: https://formative.jmir.org/2025/1/e71924
doi:10.2196/71924
PMID:40550117

©José Côté, Patricia Auger, Gabrielle Chicoine, Jinghui Cheng, Sylvie Cossette, Guillaume Fontaine, Christine Genest, Shalini Lal, Judith Lapierre, M Gabrielle Pagé, Marc-André Maheu-Cadotte, Geneviève Rouleau, Billy Vinette, Didier Jutras-Aswad. Originally published in JMIR Formative Research (https://formative.jmir.org), 23.06.2025. This is an open-access article distributed

XSL•FO
RenderX

under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Therapy Companion Mobile App for Acceptance and Commitment Therapy Exercises (ACTaide): Therapist and Client Co-Design Study

Serena Thapar[1], HBSc; Daniela Quesada[1], BA (Hons); Bärbel Knäuper[1], PhD

Department of Psychology, McGill University, Montreal, QC, Canada

Corresponding Author:
Serena Thapar, HBSc
Department of Psychology
McGill University
2001 McGill College Avenue
Montreal, QC
Canada
Phone: 1 514 398 8186
Email: serena.thapar@mail.mcgill.ca

Abstract

Background: Acceptance and commitment therapy (ACT) relies heavily on the between-session practice of therapeutic exercises to promote skill acquisition and improve psychological flexibility. However, adherence to this between-session practice remains a challenge. Mobile apps offer a promising solution to bridge this gap. However, few ACT apps focus exclusively on supporting clients in their between-session practice, and fewer apps involve stakeholders in their design. ACTaide, a therapy companion mobile app co-designed with stakeholders, addresses these barriers by guiding clients through ACT exercises and metaphors using annotated image sequences, supporting their between-session practice.

Objective: This study aimed to co-design ACTaide with therapists and clients, incorporating their feedback to ensure the app aligns with clinical goals and the needs of end users. The research explored stakeholder preferences and feedback on app functionality, design, and features to guide iterative design improvements.

Methods: Using a qualitative, user-centered design framework, we conducted 4 consecutive focus groups: 2 with 10 licensed ACT therapists and 2 with 14 psychotherapy clients. Each focus group included semistructured discussions and co-design activities. Data were collected through audio recordings and design artifacts (eg, sketches), which were analyzed using thematic content analysis.

Results: A total of 9 themes were identified, reflecting areas of convergence and divergence between therapists and clients. The therapists and clients expressed enthusiasm for ACTaide as a tool to support between-session practice. Both groups emphasized the importance of a user-friendly, intuitive, and aesthetically appealing interface, with a preference for high-quality visuals over text-heavy features. Personalization and customization were viewed as essential for enhancing app engagement. The therapists prioritized accessibility and clinical appropriateness, voicing concerns about features that may be inconsistent with ACT principles, such as symptom rating scales, and clarified their role in app delivery. By contrast, the clients emphasized wanting greater interactivity and elements of gamification to improve engagement. Slight discrepancies were noted between therapists' preferences for minimal designs and clients' preferences for more vibrant and engaging aesthetics. Overall, both groups recognized the app's potential to address barriers to homework adherence and to extend the benefits of therapy into clients' daily lives.

Conclusions: The study illustrates the value of using a user-centered, co-design approach in the development of ACTaide, an adjunctive mental health app for the between-session practice of ACT exercises and metaphors tailored to therapist and client preferences. Through the integration of stakeholder feedback, the findings provide actionable insights for designing psychotherapy tools that balance clinical goals with user preferences. Future research will focus on testing high-fidelity prototypes to evaluate acceptability, usability, and engagement.

(JMIR Form Res 2025;9:e69532) doi:10.2196/69532

KEYWORDS

mental health; mobile app; focus group; co-design; user-centered design; acceptance and commitment therapy; psychotherapy

XSL•FO
RenderX

Introduction

Background

Mental health problems are a global concern, representing a leading cause of disability [1]. Among evidence-based psychotherapies, acceptance and commitment therapy (ACT) has been found to be efficacious across a wide range of conditions, including anxiety and depression [2], the most common mental health problems worldwide [1]. For those who have access to psychotherapy, it has been reported that clients do not fully engage in homework, with surveys indicating that approximately 20% to 50% of clients do not engage in any between-session practice [3]. Homework (ie, between-session practice of skills taught in therapy) has long been considered central to the effectiveness of psychotherapy across a wide variety of clinical conditions [3-5], with meta-analyses showing large effect sizes of homework quantity and quality on therapy effectiveness [6,7]. Through engaging in skill acquisition exercises, homework extends the therapeutic benefits of treatment beyond therapy sessions, thereby promoting enhanced symptom improvement [8,9]. Unfortunately, there is a lack of therapeutic tools that focus specifically on encouraging clients to practice exercises between sessions [10], including for ACT [11]. This study addresses this issue by developing a low-fidelity prototype of ACTaide, a therapy companion mobile app aimed at helping clients practice ACT exercises between sessions via visually illustrated and annotated ACT exercise and metaphor sequences.

ACT Model

As part of the third wave cognitive behavioral therapies, ACT combines mindfulness, acceptance, and behavior change processes to enhance psychological flexibility [12]. At the core of healthy emotional functioning, psychological flexibility refers to the ability to live in the present moment and be open to experiences for the purpose of engaging in values-guided actions [13]. It uses psychoeducation, experiential exercises, and metaphor exercises to positively impact clinical outcomes.

The efficacy of ACT has been shown across a wealth of evidence. ACT has been extensively studied in >1000 randomized trials [14,15] that have been summarized in several meta-analyses [16,17] and a review of 20 meta-analyses [2], demonstrating medium-to-large effect sizes across a broad range of mental health conditions. The World Health Organization has adapted ACT in >25 languages to be used with and by people experiencing stressful circumstances [18]. ACT relies heavily on between-session practice of experiential and metaphorical exercises for psychological improvement [15,19]. However, ACT has not been exempted from the issue of insufficient homework engagement [11].

Psychological flexibility is the overarching target of ACT and is meant to be cultivated by developing the 6 skills or therapeutic processes specified in the ACT hexaflex theoretical model [4,20,21]. These 6 processes are *acceptance* (ie, learning to accept difficult thoughts and feelings rather than trying to fight or avoid them), *cognitive defusion* (ie, learning to gain a psychological distance from thoughts and feelings), *present-moment awareness* (ie, learning to maintain awareness of current thoughts, emotions, sounds, and sights), *self-as-context* (ie, learning to observe internal experiences from a distance), *values* (ie, learning to identify and live by deeply held personal values), and *committed action* (ie, learning to act or not act in line with these values).

To foster the development of psychological flexibility, ACT uses metaphors and experiential exercises that target the 6 processes [12,13,22]. Metaphors have been long-standing therapeutic devices in clinical settings and across therapeutic interventions [23,24]. Their easy, story-like properties enable clients' understanding of their personal circumstances while directing them to solutions that could promote positive behavior change [23,25]. Finally, metaphors allow clients to get some distance from their own difficult thoughts, thus preventing the occurrence of cognitive fusion (ie, the opposite of defusion) and enhancing psychological flexibility [23]. ACT's emphasis on experiential learning is also demonstrated through its use of experiential exercises. These exercises assist clients in facing psychological processes, such as thoughts, feelings, memories, and sensations that they have previously avoided, by encouraging experiential engagement with them [26]. Experiential exercises also serve as a vehicle for skill acquisition [4]. What remains a challenge is clients' engagement with such exercises as part of homework.

Homework has long been considered central to the effectiveness of psychotherapy, with meta-analyses showing large effect sizes of homework quantity and quality on therapy effectiveness [5,6,27]. However, it is not uncommon for psychotherapy patients to engage minimally in between-session practice of exercises learned during therapy sessions [3,28]. Reasons for low engagement include low motivation, not remembering the details of the exercises, not fully understanding them, or not having access to the exercises when needed [29-31].

Existing ACT Mobile Apps

Mental health apps have prospered in the past decade, with >15,000 on the market [32]. They have gained attention for their potential to enhance skills training in ACT [33,34]. More than 20 ACT apps exist, displaying considerable heterogeneity in their purpose and features [35-56]. Many of these apps rely primarily on lengthy psychoeducational content delivered via text and video, which generally has been found to hinder user engagement within mental health apps [57]. Furthermore, to the best of our knowledge, no existing app focuses exclusively on the between-session practice of ACT metaphors and exercises.

Low engagement is a pervasive issue within mental health apps [58], including those focused on ACT [11,49,59,60]. While reasons for this poor engagement remain unclear, researchers posit that app design often fails to adequately consider users' experiences and needs [58,61,62]. Among the existing ACT apps, few have incorporated participatory approaches during development [41-43,52,54], such as co-design [52]. While many ACT apps show promise in research settings [43,52,54], their access is limited, given that many of them were created for research endeavors and are not publicly available [37-39,41-44,47-49,52-54]. This limitation is particularly significant given the interest expressed by practitioners in

XSL•FO
RenderX

integrating ACT apps into practice. For example, results of a survey of 356 members of the Association for Contextual Behavioral Science revealed that ACT practitioners expressed strong interest in ACT apps and would find them beneficial for practicing ACT skills between sessions [63]. Indeed, participants in an ACT workshop reported wanting a mobile app to access learned skills as a support for implementation in their daily lives [64].

ACTaide

This study will address this issue by developing and evaluating a smartphone app, ACTaide, that clients can use to practice ACT exercises between sessions. The app will include the most commonly used ACT exercises and metaphors, such as those identified in *The Big Book of ACT Metaphors* [13] and *ACT Made Simple* [65]. Unlike existing apps, ACTaide offers a novel delivery of ACT exercises and metaphors to be used exclusively for between-session practice of homework. The proposed smartphone app will visually illustrate the processes as annotated image sequences that guide patients step-by-step through exercises and metaphors that reflect the therapeutic processes of ACT. Each image will be accompanied by a brief text-based annotation.

The use of images and visuals has been documented to improve retention and comprehension [66]. For example, a meta-analysis of the effects of graphics on reading comprehension found that the combined use of both image and text was particularly beneficial for understanding text compared to the use of words alone [67]. Empirical evidence from medical education shows that images are more easily processed than words, are easier to recall, and have better retention than text alone [68-72]. For example, annotated image sequences on how to take medication have been shown to lead to enhanced recall, comprehension, and medication adherence [70,73].

User-Centered Design

Clients' current low engagement with mental health apps has prompted researchers to move away from developing mental health apps in isolation from end users. Instead, the user-centered design framework, an evidence-based approach that emphasizes the involvement of end users in the design process of mobile apps, has gained attention [74]. This approach has been demonstrated to be effective for the development of adjunctive mobile apps, in addition to being recommended by the World Health Organization to maximize outcomes such as engagement [75,76]. Therefore, the development of ACTaide is guided by the user-centered design framework that is built upon 2 frameworks: the information system framework (ie, user-centered design process) and the design thinking framework (ie, user-centered solution-focused approach) [75]. The information system framework emphasizes an iterative approach where understanding end users' needs, researching, learning from feedback, and integrating their insight into the design are central to the app development process. Design thinking, on the other hand, underscores the importance of empathy and creativity to further understand end users' needs. Both frameworks enable the creation of the app through a collaborative development process informed by relevant stakeholders, a process that has been successfully used for mental health app development [77-79].

Aims and Research Question

This qualitative research study will inform the first aim of this research project, which is to incorporate the feedback of end users into the smartphone app design process. The study aims to answer the following research question: "What is therapists' and clients' feedback on a low-fidelity wireframe prototype of ACTaide, a therapy companion smartphone app for visually delivering ACT exercises and metaphors?" Specifically, *feedback* refers to opinions, preferences, and perceptions regarding the prototype's interface, features, content, and overall usability. In this context, a *low-fidelity wireframe* refers to an initial, basic, and noninteractive visual representation of the ACTaide smartphone app (Figure 1). This wireframe is a simplified, static draft of the app's layout that outlines its visual structure and content, serving as a starting point for design and development without full functionality or graphics.

Figure 1. Example low-fidelity wireframes of ACTaide, depicting the basic layout and content of the home page, start exercise page, and reminders page. These were presented to the final focus group of the clients after incorporating feedback from previous focus groups as part of an iterative design process.

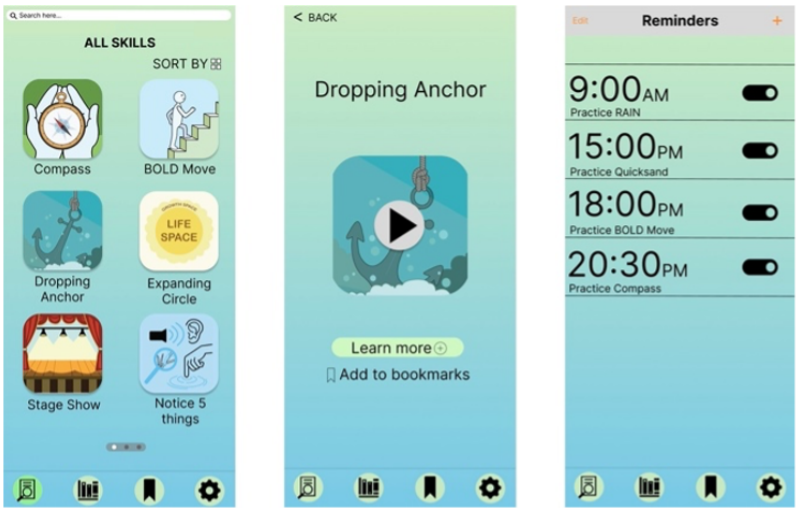

XSL•FO
RenderX

Methods

Study Design

To engage relevant stakeholders in a user-centered design process, data for this research were collected through focus groups, a widely recognized method that encourages discourse among participants on clearly defined topics [80]. Focus groups were selected for this study because of their effectiveness in eliciting diverse opinions and perceptions from therapists and patients involved in the co-design of the ACTaide smartphone app. This methodology has been successfully applied in previous mental health app development research [78]. Focus groups allow the exploration of both converging and divergent views, offering rich insights into participants' feedback and enhancing our understanding of their experiences. A total of 4 focus groups were conducted between January 2024 and October 2024: the first 2 with therapists and the subsequent 2 with clients. The focus groups were sequential, meaning that feedback from each session informed adjustments to the ACTaide wireframes before subsequent focus groups (Multimedia Appendix 1), allowing an iterative development process. The sequential order ensured that client input directly built upon a framework already vetted by clinical professionals, resulting in a well-integrated and clinically accurate tool. A summary of design changes made based on feedback during each focus group can be found in Multimedia Appendix 2.

Participants

Recruitment for therapist participants was conducted using purposive sampling to align with the research project's needs for clinical input [81]. To create a pool of potential participants to recruit from, we randomly selected and contacted 200 psychotherapists listed on *Psychology Today* (100 in Quebec and 100 in Ontario) [82] who identified as offering mindfulness-based therapy or ACT. Of these, 33.5% (67/200) expressed interest in contributing to the app development process. Therapists who worked in Montreal or the Greater Montreal Area were prioritized for recruitment to allow in-person attendance. Additional therapists in Montreal were identified through web searches and snowball sampling. For this pool, eligibility criteria included a minimum of 5 years of ACT experience, being proficient in English, and being a licensed practitioner. Therapists were invited to participate in the study via email. Using purposive sampling, 10 therapists were selected to represent a diversity of age, race, and gender (refer to Table 1 for demographic details). A total of 2 focus groups (5 in each group) were conducted.

Client participants, who represented the other stakeholder group, were recruited through flyers distributed on social media and web forums. All participants were adults (aged ≥18 years) with previous psychotherapy experience that included assigned between-session exercises. All participants were proficient in English and able to attend an in-person focus group in Montreal. The participants completed a screening questionnaire to determine eligibility and were then contacted via email with an invitation to participate in the study. Quota sampling was used to ensure representation across age ranges and gender, consistent with psychotherapy client demographics [83]. A total of 2 focus groups (8 attendees in the first group and 6 in the second group) were conducted (refer to Table 2 for demographic details).

XSL•FO
RenderX

Table 1. Demographic data of the therapists (N=10).

Demographic variables and categories	Participants, n (%)
Gender	
Men	3 (30)
Women	7 (70)
Age (y)	
29-45	4 (40)
≥46	6 (60)
Years of experience with ACT[a]	
5-10	7 (70)
≥10	3 (30)
Years of clinical experience	
<15	5 (50)
≥15	5 (50)
Primary practice location	
Hospital (outpatient)	4 (40)
Private practice	4 (40)
Other	2 (20)
Clinical position	
Doctoral-level psychologist	8 (80)
Master's-level social worker	2 (20)
Ethnicity	
European	8 (80)
Hispanic or Latinx	2 (10)
Southeast Asian	1 (10)

[a]ACT: acceptance and commitment therapy.

XSL•FO
RenderX

Table 2. Demographic data of the psychotherapy clients (N=14).

Demographic variables and categories	Participants, n (%)
Gender	
Men	6 (43)
Women	8 (57)
Age (y)	
18-29	5 (36)
30-39	2 (14)
40-49	4 (29)
50-59	2 (14)
≥60	1 (7)
Education	
Some college, CEGEP[a] (Quebec), or university	6 (43)
Bachelor's degree	5 (36)
Graduate degree (master's or PhD)	2 (14)
High school diploma	1 (7)
Employment status	
Unemployed	3 (21)
Employed or self-employed	5 (36)
Undergraduate student	3 (21)
Graduate student	3 (21)
Ethnicity	
African	1 (7)
European	9 (64)
Hispanic or Latinx	1 (7)
Middle Eastern	1 (7)
South Asian	2 (14)
Exposure to ACT[b]	
Yes	4 (29)
Unsure	6 (43)
No	4 (29)
Therapy format	
Individual (one-on-one)	10 (71)
Individual and group	4 (29)

[a]CEGEP: Collège d'enseignement général et professionnel (General and Professional College).

[b]ACT: acceptance and commitment therapy.

Procedure

Overview

We first conducted 2 focus groups with 10 therapists, followed by 2 with 14 clients, ensuring sufficient group size to facilitate interactive discussions without being too large and disorganized. This focus group size aligns with the recommendations of 5 to 8 participants per group [84] and reflects the average sizes found in similar research [85]. Focus group sessions were held at the McGill Health Psychology Lab, with the exception of the second therapist group, which took place via videoconferencing. Each session consisted of three parts: (1) a structured background presentation, (2) a focus group discussion, and (3) a co-design session for clients. Each session lasted approximately 120 minutes.

Phase 1: Exploring Users' Perceptions of ACTaide

The moderator presented an overview of ACTaide's rationale and purpose. Participants shared feedback on the concept of a

XSL•FO
RenderX

therapy companion mobile app for ACT exercises and metaphors to be practiced between sessions.

Phase 2: Understanding Participants' Perspectives, Opinions, Preferences, and Concerns Related to ACTaide Wireframes

Participants reviewed all app wireframes, receiving a high-level description of each wireframe. While low-fidelity wireframes typically do not contain graphics, preliminary sketches of 3 annotated image sequences were presented to provide context for the exercise delivery format. Therefore, feedback on the preliminary sketches themselves was not solicited. Feedback on the design, features, and functionalities of wireframes in their entirety was solicited. After reviewing all wireframes, key questions prompted deeper insights into participants' initial reactions, likes, dislikes, and suggestions for modifications.

Phase 3: Engaging Participants in a Co-Design Workshop

In the final segment, participants participated in a co-design session in which they sketched their ideal versions of the app.

The focus groups used semistructured interviewing techniques facilitated by a moderator (ST), who used predetermined opening, introductory, transition, key, and closing questions to stimulate conversation (Multimedia Appendix 3). The moderator-guided discussions related to wireframes while also adapting to emerging ideas. In total, 2 undergraduate researchers documented nonverbal behavior and group dynamics to support data analysis and capture emergent phenomena. All sessions were audio recorded and transcribed for analysis, with photographs taken of physical artifacts (eg, sketches from the co-design session).

Data Analysis

Data for this study were collected through audio recordings from each focus group session, which resulted in 4 transcripts. These transcripts were meticulously reviewed by the research team and cross-referenced with the audio recordings to ensure accuracy. To uphold methodological rigor, we used thematic analysis as the primary analytical approach [86]. This method is particularly suited for systematically identifying collective patterns within a large dataset through an inductive lens, while also considering the contextual nuances of participants' contributions [87]. Thematic analysis enables researchers to remain open to the complexities that may arise during discussions, free from the constraints of a rigid conceptual coding framework. In addition, recorded observations of nonverbal behavior, group dynamics, and emerging themes were reviewed to further enrich the analysis. Physical artifacts, such as sketches created during client co-design sessions, were also coded and analyzed.

The six-phase framework for conducting thematic analysis guided our approach: (1) becoming familiar with the data, (2) generating initial codes, (3) searching for themes, (4) defining themes, (5) iteratively reviewing themes, and (6) writing up the results [86]. ST and DQ independently read and reread all transcripts, applying open coding to identify initial codes. Subsequent meetings facilitated the expansion and refinement of codes as new insights emerged, leading to the identification of themes through iterative reviews. To enhance the validity and reliability of our analysis, we discussed the emerging themes with a group of researchers from the McGill Health Psychology team who were familiar with this research area, incorporating their feedback into the final themes. Quotes were extracted from transcripts for therapists and clients to illustrate the main themes that emerged from the data.

We then used a case study methodology to compare feedback from distinct end-user groups—therapists and clients. This comparative analysis aimed to uncover converging and diverging opinions [88,89], revealing how the app could be optimally tailored to meet the diverse needs of all end users. This comparative perspective was intended to provide a nuanced understanding of the differing attitudes toward the low-fidelity wireframes among therapists and patients. Ultimately, this analysis will inform decisions regarding the app's redesign based on shared and contrasting viewpoints.

Ethical Considerations

This study received approval from the McGill Research Ethics Office (institutional review board; 22-11-068). All participants provided written consent before participating in the study. To ensure privacy and confidentiality, all transcripts were deidentified before analysis. Furthermore, demographic information was aggregated in the reporting of study data. All participants were compensated for their participation in the study. Those who participated in the therapist focus groups were compensated CAD $200 (US $146.15). Those who participated in the client focus groups were compensated CAD $80 (US $58.46).

Results

Overview

A total of 9 themes emerged from the focus group discussions. In total, 6 themes were common across both therapist and client focus groups: the value of ACTaide, aesthetic design preferences, ease of use, customization and personalization, app guidance, and app engagement and gamification. A total of 3 additional themes were unique to therapist focus groups: therapist involvement, accessibility, and clinical concerns.

The Value of ACTaide

Both the therapists and the clients consistently acknowledged the inherent value of ACTaide as an adjunctive tool to facilitate between-session practice. The therapists particularly endorsed the app's potential for supporting the continuity of psychotherapy beyond the therapy room, which they deemed necessary for meaningful change:

> *Therapy isn't really about what happens in the office...[it's] about what our clients are actually doing different outside of sessions...something that can support out-of-session practice really moves people forward toward their goals.* [Therapist 1]

Many therapists also expressed enthusiasm for the therapy companion mobile app and its use as an adjunct to their clinical practice:

XSL•FO
RenderX

> *I think the idea of a companion app versus a standalone thing is a kind of nice idea...We are not expecting the app to do the therapy.* [Therapist 4]

The clients similarly valued the concept of the app, with several expressing that they often struggle to engage in between-session practice of psychotherapy skills, citing the app's potential to help them stay accountable and consistent. For example, 1 client reflected on their experience in therapy, stating the following:

> *I think it [referring to the app] is a really good idea...it's easy to forget the exercises that you're supposed to implement.* [Client 3]

In addition to the overall value of the app itself, both the clients and the therapists acknowledged the value of various app features. For example, both the therapists and the clients saw value in the app having a reminder feature, believing that it would promote greater app use:

> *I just forget these apps are on my phone...but if it sends me prompts during the day, maybe I would be more likely to do whatever it tells me to do.* [Client 13]

Similarly, both the therapists and the clients expressed their appreciation for a *favorites* feature that would provide the option to shortlist exercises for easy access.

A unique perspective that emerged from the therapists' focus group was the recognition of the value of the app's format (ie, using annotated image sequences) for delivering exercises and metaphors:

> *The fact that it's image-based is really appealing...that's the big thing that stood out.* [Therapist 7]

While the clients did not express contrasting opinions, they did not comment positively or negatively on the format of delivery.

Aesthetic Design Preferences

Both the therapists and the clients offered critical feedback on the visual appeal of the app and offered suggestions for improvement. Most expressed the importance of an appealing aesthetic design to encourage app use:

> *People are going to want to look at something that looks kind of like nice and fancy...there is an importance to the aesthetic appeal of it.* [Therapist 9]

> *For me personally, I wouldn't use this app if it didn't look good to me.* [Client 12]

While they acknowledged that the app was a low-fidelity wireframe, they expressed that the visual design of the app was not on par with current market standards. The clients expressed a similar sentiment, expressing the need for the design to be adapted for the popular market:

> *It doesn't look modern. It doesn't look like today's apps. It looks like...the beginnings of the internet.* [Client 13]

Some therapists appreciated what they described as a low-attention and boring design:

> *I see the value in a break from, you know being bombarded with color and moving stuff.* [Therapist 3]

By contrast, the clients expressed a desire for a more attention-grabbing, colorful design (client 8).

A recommendation to enhance the app's aesthetic focused on reducing the amount of text. For instance, therapists suggested simplifying the exercise titles to make them less wordy and using images to complement the list of exercises and metaphors (therapists 1 and 8). The clients echoed this idea, preferring symbols to represent words (eg, bookmarks and learn more):

> *You don't need all those words like we have symbols...And you know that pressing that [symbol] will add the exercise to the bookmarks. And you know that an eye in a circle means you can get more information for example.* [Client 10]

Both the therapists and the clients shared suggestions to improve ACTaide's color palette. Drawing inspiration from other mental health apps they have experience with, some therapists recommended incorporating background colors, such as green and blue, into the app (therapist 1). The clients specifically generally requested greater contrast in the colors, feeling that the app was "plain" (client 1). The therapists leaned toward wanting nature-inspired colors in the color palette, a suggestion that echoed client feedback favoring earthy tones for a calming effect (client 13):

> *I would like to put nature-like colors...the browns, the greens, mainly blue sky.* [Therapist 4]

Similarly, clients provided suggestions to improve the appearance of the slider feature by incorporating colors to visually represent changes in their feelings.

Other specific suggestions made by both the therapists and the clients were discussed to rectify certain elements of the app, such as the size and location of the icons, with the goal of enhancing the app's appeal:

> *The icons at the bottom...feel like they're a little too big for the circles.* [Client 2]

The clients further found some pages to be fragmented and preferred simplicity. For example, they recommended merging the play button with the cover image of the exercise to create a more streamlined visual (Figure 2):

XSL•FO
RenderX

Figure 2. Implementation ideas for an exercise start page, generated by the clients in the fourth focus group. Suggestions focused on enhancing visual appeal and improving clarity.

> *Having the play button underneath the actual icon, it's just, it all looks very, very fragmented and kinda clunky and kinda busy.* [Client 10]

Finally, some clients noted that our aesthetic seemed to more closely resemble a web page and suggested ensuring it remained consistent with typical app designs*:*

> *The landing page looks like an Internet website, whereas the other pages look like an app.* [Client 13]

Ease of Use

Both the clients and the therapists offered their feedback on the intuitiveness and simplicity of the interface. Consensus was reached about the app being user-friendly, with most agreeing that the app seemed simple and easy to use. Many therapists made comparisons to their favorite mental health apps, noting that their simplicity and ease of use were what stood out:

> *It looks fairly simple and user-friendly.* [Therapist 1]
>
> *It's easy...it's not complicated.* [Client 3]

The clients and the therapists also agreed that there were often too many steps or clicks involved, which could overcomplicate things. For example, the therapists felt that there were too many steps (eg, a distress scale and breathing prompt) to complete before one could start practicing an exercise:

> *There seemed to be a lot of pages in between saying that you wanted to do the exercise and actually getting into the exercise.* [Therapist 1]

Similarly, the clients felt that having a home screen or a pop-up screen to learn more about a given exercise was unnecessary, with 1 noting that "it's a lot of clicking" (client 10).

Both the therapists and the clients emphasized the importance of intuitive labeling across the app to ensure ease of use. For example, the therapists expressed that an overly minimalist design could pose challenges for some clients with various aptitudes in technology. For example, an arrow was not considered explicit enough for one to understand how to exit a given wireframe. The clients had similar concerns with the language of various labels, such as *all skills*, *snooze*, *see more*, and *bookmarks*, which they reported were not intuitive. For example, the term *bookmarks* was considered confusing, as it is commonly associated with internet browsers rather than mobile apps (client 13). Alternative suggestions were agreed upon, such as *exercises*, *learn more*, and *favorites*:

> *It's not clear to me what would happen if I click on that [referring to the see more button]. Maybe learn more.* [Client 6]

Similarly, the clients also reported that certain icons were not intuitive, such as the play button to start an exercise. They believed that this was not intuitive, as it primed them to expect a video (client 4). The bookmark icon and the library icon were also not rated as intuitive, with suggestions made to align them with their new labels (eg, a heart to represent favorites and a dumbbell to represent exercises).

While the clients acknowledged the value of various app features, such as reminders and a journal, they reported that such features should be more visible in the app. For example, they proposed moving such features into the navigation bar for easy access, as the previous ways of accessing such functions were not intuitive.

Finally, clients uniquely expressed a desire for increased app guidance generally. For example, they requested an optional

XSL•FO
RenderX

tutorial and a help button to provide technological support to ensure that the app was user-friendly:

> *[When you] download a new app, you have...a tutorial that [says] welcome, [and] then takes us here [referring to the list of exercises].* [Client 6]

Customization and Personalization

Both the therapists and the clients underscored the need to incorporate more options for customization and personalization of the app and its features to ensure that it accurately reflects the clients' realities and needs. For instance, the therapists suggested that the facial expressions in the graphics could be customized to the user's felt emotions through an input where one could enter their primary emotion. Such a sentiment was echoed by the clients, who felt that such an option would allow the exercise to resonate better with their experience. A therapist stated the following:

> *The actual graphic could be changed based on the emotion they identified...[such as] anger, shame, sadness, fear, which is going to more or less encompass the majority of negative experience.* [Therapist 9]

The therapists also proposed that an optional journal feature, where a client could reflect upon a given exercise, may be useful for the clients:

> *There could be an additional page where the person is asked to reflect...have them provide feedback on how it was for them.* [Therapist 6]

The clients expressed a strong desire for this feature as well, believing that it would aid in the app feeling more personalized:

> *I like the idea of being able to take notes of my thoughts during this process somewhere in the app.* [Client 7]

Finally, the therapists and the clients alike appreciated the option for customized reminders for exercise practice. The therapists emphasized the importance of avoiding the default reminders, suggesting instead that the clients have control over the frequency of such notifications.

The clients noted that the app had minimal opportunities for personalization, reporting that changes could be made to humanize the experience. They reported that small changes, such as being welcomed by the app with their name or with a personalized avatar, would make it feel less generic (Figure 3).

Figure 3. Implementation ideas for a home screen, generated by the clients in the fourth focus group. Suggestions emphasized personalization, gamification, and intuitive design.

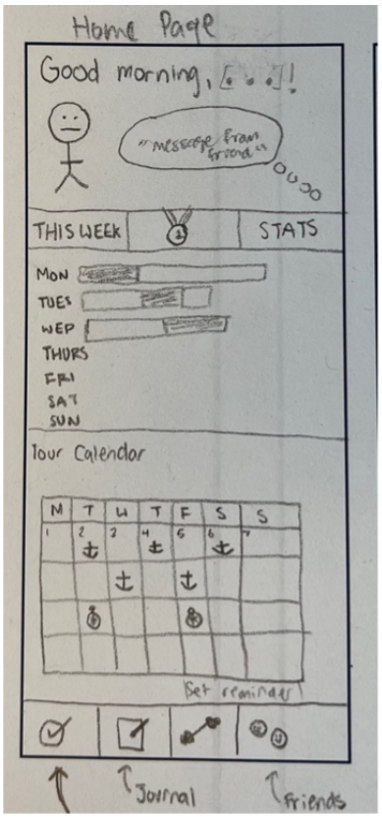

The clients and the therapists also requested further opportunities for customization throughout the app. For example, suggestions were made for the clients to customize the background color, stating that it would increase one's desire to use the app (client 12). Furthermore, the therapists uniquely acknowledged that the clients may remember exercises by terms different from the formal labels used by their therapists. To rectify such discrepancies, they suggested an option for the clients to customize the label of the exercise based on their personal recollection of it (therapist 9).

App Guidance

Both the therapists and the clients expressed a desire for the app to offer guidance regarding exercise selection. The therapists

XSL•FO
RenderX

recommended incorporating more guidance in the form of assessments that can help identify the clients' needs in a given moment. One would use the information from the assessment to guide them toward a specific exercise that targets the specific process or processes they are having the most difficulty with:

> *I wonder about maybe [adding] a question instead.... Something that is a bit more clear and is asking the client what they need in that moment? And then their answer to that will guide to the next part of the app.* [Therapist 3]

When such a feature was presented to the clients, they unanimously expressed appreciation for it, validating the therapists' suggestions for assessments to guide exercise selection.

When the clients were shown the "get-recommendation" feature, a brief assessment to obtain a recommendation for an exercise to practice, they indicated a strong preference for completing the 6 items aligned with the hexaflex, as opposed to the 3 items aligned with the triflex. The clients reported favoring a 6-item questionnaire because it allowed greater specificity of recommendations and promoted self-reflection:

> *I wouldn't mind answering six questions too if it gives me a better recommendation.* [Client 12]

In addition to desired guidance on exercise selection, the clients expressed consensus in valuing the "see-more" feature, wherein they could obtain a brief description of any given exercise and its purpose:

> *That's very useful. I would definitely like to know what the exercise is about...before actually doing it.* [Client 1]

Some therapists similarly echoed the need for this feature, suggesting providing information on why a given exercise could be helpful or when one might want to try it (therapist 7). Most felt that very brief and optional preparatory psychoeducation could be helpful for clients:

> *Having some psychoeducation worked in...to remind people...why try this...would be really helpful.* [Therapist 9]

App Engagement and Gamification

The therapists and the clients alike expressed a need to enhance app engagement. The therapists expressed mixed views about the use of gamification. Some felt that gamification could undermine intrinsic motivation:

> *[Gamification] removes kind of deliberate, kind of willful agency.* [Therapist 3]

Others appreciated the use of gamification features, such as progress tracking, that could reinforce app use, comparing it to other apps on the market that integrate such features. For example, 1 therapist commented that integrating a progress tracker would allow people to think, "hey, I'm building up this habit and feeling good about it" (participant 9).

By contrast, the clients reached a consensus that gamification elements would be their preferred way to increase app engagement. The clients emphasized their desire for similar progress tracking features to promote a sense of reward and achievement:

> *There's no sense of achievement in the app.* [Client 4]
>
> *Something that help us to see, maybe a graphic,...what is the progress made by this app.* [Client 5]

Some suggested integrating goal-setting features for exercise practice:

> *You set a goal...right after your last therapy session, [such as] "I want to do my exercise like two times a week"..., then it shows your progress on your goals.* [Client 9]

Furthermore, both the clients and the therapists expressed concerns that the clients may bypass the exercise content by mindlessly skipping through the steps in the annotated image sequences without fully engaging with them. To address this concern, both the clients and the therapists suggested embedding questions and forced-choice prompts within the steps to require user input, such that the exercises are more interactive:

> *If you just have the app as is with images, it's so easy to just press next, next, next without actually doing the steps. There has to be some sort of input because then otherwise I wouldn't even want to go on the app to do it because...[I can do it] on my own.* [Client 9]

The clients proposed that greater interactivity could be achieved by incorporating sound and animations into the image sequences. They also expressed that an option for text to be read aloud might enhance their desire to engage with the exercises.

Finally, the clients considered the utility of a social component to encourage app engagement, such as the ability to connect with others by sharing their progress or sending words of encouragement:

> *I wouldn't want people that I'm friends with necessarily knowing which exercises are completed or not.... But just some sort of indication that they have also been working on their exercises.* [Client 9]

The participants drew comparisons to other apps they were familiar with that use similar strategies. While some expressed interest, others noted that psychotherapy is a personal activity and were hesitant about such a feature (participant 9).

Therapist Involvement

The potential for therapist involvement through a designated interface that would allow the therapists to assign exercises and monitor progress was a significant focus across the therapist focus groups. Such interaction was queried in the client focus groups but was not elaborated upon as a topic of discussion. Some therapists were initially interested in an optional shared data feature that would allow them to access patient data to monitor user engagement. Through further discussion, the therapists identified that such a feature was not likely to be useful, given the additional time necessary to examine the data:

> *It might be useful to know how much time they spent on the app during the week and...where they spent*

XSL•FO
RenderX

> *the most time. But...I don't know that I would be pulling up the data.* [Therapist 1]

Further clinical concerns were echoed by most therapists, who felt that such a mechanism may negatively impact patient autonomy and motivation:

> *I want them using it for their own reasons and not because they feel they're being watched or that they're accountable to someone.* [Therapist 1]

However, 1 therapist observed that encouraging a sense of accountability might be clinically useful for some clients who lack intrinsic motivation:

> *Being accountable to your therapist for actually doing the exercises, might be tolerated and actually clinically useful for these patients..... I think the ability for them to feel that they are accountable to a therapist might actually promote usage and psychological flexibility.* [Therapist 3]

Similarly, the therapists initially expressed interest in taking an active role by assigning exercises to clients using a therapist interface, rather than having all exercises available to the patients:

> *I really like the idea of having things sort of locked or unlocked...Therapists will have some sort of control over the order in which things are...opened up to the patient.* [Therapist 1]

Others did not see the need for restricting access to exercises and metaphors:

> *I am okay with my patients going further and attempting things without me having said we're working on this thing and then coming back.* [Therapist 2]

Most therapists once again acknowledged the burden of such a task, sensing that it would further add to their workload.

Accessibility

The therapists emphasized the importance of accessibility throughout the app design, a topic that was not a focus of the client focus groups. The therapists appreciated the use of annotated image sequences with minimal, nonacademic text for promoting accessibility across people with differing education levels:

> *It is accessible...for people who might have difficulties with reading or benefit from something that is more image-based.* [Therapist 8]

More specifically, 1 therapist mentioned that images could facilitate the comprehension of content for clients who are not fluent in English. Some therapists expressed a desire to adopt a more universally accessible app design. For instance, they were keen on including an auditory component for clients with visual impairments:

> *An accessibility feature that has an audio option [would be necessary]....if it's someone who has any type of visual impairment, this app will be inaccessible to them.* [Therapist 8]

All the therapists recommended further improving accessibility by reducing academic jargon, such as removing hexaflex terminology from exercise categories visible to the clients.

They also suggested enhancing visual and tactile accessibility by incorporating options to adjust the font and button sizes.

Clinical Concerns

Various clinically relevant concerns were raised across the therapist focus groups. More specifically, the therapists expressed their doubts regarding the clinical utility of certain ACTaide features and their alignment with ACT principles. For example, the distress rating scale garnered the most concern, as it was deemed to be inconsistent with clinical use and geared toward research purposes. Specifically, the distress rating scale was poorly received, given that the therapists felt that focusing on reducing distress conflicted with ACT's goal of enhancing functionality rather than targeting symptom reduction:

> *But the goal is not to reduce distress...it's to tolerate it...Reduced distress is sometimes a by-product, but it's not the emphasis.* [Therapist 1]

One participant noted that it is illogical to expect a change in distress after a brief exercise, while another raised concerns that clients might feel disappointed if they do not perceive immediate changes:

> *Is it realistic to expect the change after a 3 to 5 minute exercise?* [Therapist 5]

By setting up expectations of change among clients, the therapists expressed that the distress scale might compel some of them to feign their levels of distress, rendering the assessment meaningless. Finally, 1 therapist noted a preference for clients to engage with the exercises beyond moments of distress, which the distress scale seems to discourage by implying that users are distressed before using the app (therapist 3).

Most therapists felt that the "breathe feature," which prompted clients to take a deep breath before doing an exercise, was not clinically appropriate, as the anchor of attention on the breath is not universally clinically indicated (therapist 5). They suggested using the phrase "take a moment" instead, to make it more accessible for all clients (therapist 1).

Finally, while the therapists appreciated the functionality of the favorites feature, they expressed hesitancy toward the language, proposing an alternative, such as "bookmarks." The term "favorites" was believed to encourage the practice of only easy exercises, thereby creating an environment for clients that promotes avoidance of more challenging exercises:

> *Some of the exercises that might be most helpful to people actually might be the ones that are their least favorites...They might really dislike them, and yet it's still going to be good for them to do it. So, I think it almost sets up avoidance.* [Therapist 9]

Discussion

Principal Findings

This study explored therapists' and clients' feedback on a low-fidelity prototype of ACTaide, a mobile app designed to

XSL•FO
RenderX

enhance between-session practice of ACT exercises. The participants provided valuable insights into their preferences and perceptions, which guided iterative refinements of the app's wireframes. A total of 9 themes emerged, reflecting both shared and differing perspectives between the therapists and the clients. Both groups emphasized the value of an adjunctive app for between-session practice and highlighted the need for a user-friendly and intuitive design. They preferred a streamlined interface with high-quality visuals rather than text-heavy features, while also advocating for greater interactivity to encourage consistent use. Personalization and customization were seen as essential for meeting diverse user needs. The therapists offered additional professional insights, including suggestions for maintaining alignment with ACT principles, increasing accessibility, and clarifying the therapist's role in app delivery. These findings provide actionable guidance for designing mobile apps that balance clinical goals with user preferences.

Usability emerged as a critical theme, with participants emphasizing the need for a streamlined, intuitive interface. According to the International Organization for Standardization, usability is commonly defined as "the degree to which a product or system can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use" [90]. The factors that contribute to the usability of an app include the amount of effort required to learn and use the app [91]. This finding is consistent with perceptions of both consumers and mental health professionals, who highlight ease of use as one of the most valuable app features [92,93], with ACT therapists specifically identifying it as a key criterion they use for evaluating ACT apps [63]. Specific suggestions, such as reducing the number of clicks, simplifying navigation, and providing onboarding tutorials, mirror well-established usability heuristics [94,95]. These recommendations were particularly salient among the clients, who noted that overly complex designs could detract from their willingness to engage with the app. Changes based on client feedback, such as incorporating intuitive icons, were iteratively made to further simplify the wireframes. Icons have been found to enhance app navigation and improve comprehension [96], provided they are familiar to users and unambiguous [91]. Similarly, suggestions were made to increase the visibility and location of certain app features, such as reminders and the journal. Older adults have identified such enhancements as crucial for improving the ease of use of apps [97]. Finally, the clients desired greater app guidance and requested an optional tutorial. This aligns with app development recommendations emphasizing that onboarding instructions should be provided for first-time users, regardless of how intuitive the app is perceived to be [61,98].

Consistent with prior research emphasizing the role of aesthetics in fostering digital engagement [99,100], both the therapists and the clients emphasized that a visually appealing app would be necessary to encourage app use. Research has shown that an app's aesthetic design correlates with an app's perceived usability [91]. The focus on aesthetic design is unsurprising, given that users tend to focus closely on aesthetics and visual interface elements during their initial experiences with apps [101]. Both the therapists and the clients focused on providing suggestions to enhance the app's visual appeal. For example, the color palette was a focal concern, with participants requesting calming colors, such as blues and greens. Research has shown that blue and green tones with medium saturation and high brightness, a color scheme common to medical apps, evoke a sense of calmness, trust, and professionalism [102]. Given the influence of colors on emotions and visual appeal [100], such feedback was implemented into iterative wireframes.

The participants further highlighted the desire to customize many of the app features, such as the labels, reminders, images, and colors. Designing mental health apps for a wide audience can be achieved by incorporating opportunities for the app to be personalized to align with individual needs and preferences [61]. Customization has been identified in mental health apps as a factor positively affecting both app effectiveness and engagement [103,104]. However, it is not commonly implemented [105]. For example, the therapists acknowledged the value of reminders but emphasized the need for end users to customize their timing and frequency. This iterative change was appreciated by the clients in the subsequent focus groups, echoing the feedback given by users in other studies regarding customizable reminders [61,77,106]. Such customization strategies in mental health apps have been shown to foster end users' sense of self-agency and self-determination, ultimately enhancing their engagement with apps [91].

Concerns and suggestions to improve overall app engagement emerged across focus groups, reflecting the widespread issue of low engagement faced by many mental health apps [58]. Gamification to promote engagement emerged as a particularly contentious topic, with the therapists expressing concern that gamified elements might undermine intrinsic motivation, while the clients viewed such features as necessary to maintain interest. Research on gamification in mental health apps has yielded mixed results, with some studies demonstrating its effectiveness in promoting app use, while others have found the opposite [107,108]. For example, a meta-analysis on apps for depression found no difference in engagement between gamified and nongamified apps [109]. Progress levels, a feature that shows how much progress a user has made toward a goal or task, have been found in >40 mental health apps [110]. The common use of this feature may explain why clients emphasized the inclusion of a similar feature, often making comparisons to apps they have used before. The therapists expressed concerns about incorporating gamification, noting that it may hinder intrinsic motivation. This aligns with critiques that argue gamifying technology relies on extrinsic motivators and positive reinforcement, such as overt rewards, to motivate user engagement [111]. Research has shown that gamification in health technologies is frequently implemented without accounting for its effects on intrinsic motivation, without clear objectives, and without understanding its mechanisms [110,112]. To reconcile these contrasting views on gamification given by the therapists and the clients, a compromise might involve subtle gamification strategies, such as progress tracking and badges for completed exercises, that support engagement without detracting from the therapeutic focus.

XSL•FO
RenderX

Both the therapists and the clients highlighted a need to incorporate interactive elements into ACTaide to promote engagement with the exercises. According to the taxonomy of user interactivity [113], which refers to the degree to which users can input and modify system outputs, the wireframes presented place ACTaide at a static level of interactivity, with limited to no user input of personal data. Incorporating the therapists' and clients' recommendations, such as adding questions, forced-choice prompts, and animations, would elevate ACTaide from a static to a basic level of interactivity within the taxonomy. These changes are important, given that interactivity features are known to enhance user engagement in digital interventions [113].

The therapists were particularly concerned with their level of involvement in the delivery of ACTaide, reporting mixed feelings about the potential inclusion of a therapist interface that would enable them to monitor user engagement and assign exercises to the clients. The discussion was characterized by an immediate interest in these features, followed by concerns about the feasibility of implementation in clinical practice. Most therapists acknowledged that it would undesirably add to their workload. Such opinions have been expressed among other health professionals, who emphasized that any new technology they adopt in their practice should clearly enhance efficiency, noting that their auxiliary tasks were already time consuming [93]. This finding aligns with the therapists' reluctance to take on additional responsibilities. Similarly, a qualitative exploratory pilot study exploring therapists' perspectives on blended therapy (eg, the combination of psychotherapy with digital components) highlighted similar concerns over workload as a barrier to implementation [114]. Therefore, the therapists noted that their involvement in the interface would be unnecessary for the app to achieve its intended purpose for the clients.

The therapists also emphasized the importance of accessibility, expressing appreciation for the app's strengths while noting areas for improvement. ACTaide's minimal use of language and emphasis on imagery were considered suitable for clients with lower literacy levels, who often encounter language barriers when accessing mental health technologies [115]. The therapists also suggested reducing academic language, which aligns with findings from a systematic review that found that participants engage better with interventions that are tailored to them, which could be achieved by limiting the use of jargon and technical language [116]. This reduction in academic jargon not only enhances user engagement but also simplifies language, a factor that end users have identified as important in mental health apps [101]. Other suggestions to improve accessibility were made, such as changing font size and button size and adding optional features to accommodate end users with disabilities. A cross-sectional study exploring 578 mental health apps in the marketplace found that accessibility features, such as adjustable text size and text-to-speech or speech-to-text abilities, were present in less than half of these apps [117]. Thus, the findings highlight the importance of prioritizing accessibility features in iterative app design.

An important contribution of this study is the therapists' identification of features that may inadvertently conflict with ACT principles. For instance, the therapists raised concerns about the inclusion of a distress rating scale, arguing that it risked reinforcing symptom reduction as a primary goal, contrary to ACT's emphasis on psychological flexibility [118]. The results from an Association for Contextual Behavioral Science survey revealed that practitioners rated consistency with ACT principles as the most important criterion for evaluating apps [63]. This feedback underscores the value of incorporating professional expertise in the co-design process to ensure that digital tools align with the theoretical foundations of their respective interventions. Following this feedback, the distress rating scale was removed from the app, demonstrating how iterative design can safeguard intervention fidelity. Furthermore, the therapists disliked the "favorites" label, as it was thought to implicitly promote experiential avoidance of more challenging exercises. The term "bookmarks" was thought to be more inclusive. However, the clients disliked the "bookmarks" label, as it was not considered to be intuitive, and preferred the term "favorites." This is an example of how a user-centered iterative design process allows one to refine and adjust the app based on the evolving preferences and needs of various stakeholders, which may differ by group [119].

Implications

The findings from this study provide actionable insights for refining ACTaide and underscore the value of user-centered, iterative design in mental health app development. The feedback obtained from this study will be used in the iterative design of ACTaide, ensuring alignment with end users'' needs and preferences. New wireframes based on the identified themes will be translated into a high-fidelity functional prototype. To refine the intervention, the next step will involve beta testing a high-fidelity prototype to assess acceptability, engagement, and usability in real-world settings. Aligned with the Obesity-Related Behavioral Intervention Trials model for designing behavioral interventions, further data will need to be obtained through a proof-of-concept study, pilot testing, and then a feasibility pilot in preparation for an efficacy trial [120]. A longitudinal efficacy trial could then evaluate the app's impact on homework adherence and psychological flexibility, addressing a critical gap in the literature. Experimental studies might also benefit from testing the format of exercise delivery (ie, annotated image sequences) to assess the impact on comprehension and retention. The facilitators and barriers to widespread adoption of ACTaide may also be explored during implementation, alongside effectiveness in real-world settings.

The incorporation of the perspectives of multiple stakeholders can potentially promote the acceptability of the tool while enhancing uptake and retention [103,121]. Importantly, involving health professionals alongside end users in the design of mental health apps may boost confidence in the final product [61,122]. Research has found that endorsement of patient-facing technologies by mental health providers is associated with greater interest from end users [123], emphasizing the credibility of an adjunctive tool that has been co-designed with mental health professionals.

Strengths and Limitations

This study has several strengths, notably its use of focus groups with a user-centered co-design framework, which facilitated

XSL•FO
RenderX

in-depth discussions and provided a platform for eliciting feedback on a low-fidelity wireframe prototype of ACTaide from distinct stakeholder groups—the therapists and clients. Another strength lies in the diversity achieved within the client focus groups, supported by quota sampling across ethnicities, which largely reflected the demographics of Montreal [124]. This sampling approach allowed representation of a wide spectrum of client experiences, though only 1 older adult client participated, which likely mirrors the age distribution in psychotherapy clientele [125]. While the therapist focus groups showed less diversity, this likely reflects workforce demographics; in Canada, for example, <25% of psychologists are men [126] and >75% are aged between 30 and 59 years [127]. The iterative nature of focus group discussions also allowed the research team to refine wireframes based on real-time participant feedback, promoting a user-centered approach to app development. The geographical restriction to participants who resided in or were able to commute to Montreal, combined with the requirement for English proficiency, limits the generalization of findings to other regions or non–English-speaking populations.

Conclusions

This study illustrates the feasibility and value of involving both users and professionals in the design of an adjunctive therapy app at the first stage of the process, aligned with the user-centered design approach [75]. The key themes identified were the value of ACTaide, aesthetic design preferences, ease of use, customization and personalization, app guidance, app engagement and gamification, therapist involvement, accessibility, and clinical concerns. By integrating feedback from both the therapists and clients, the app may address long-standing challenges of homework engagement through a user-friendly solution, the value of which was recognized by all stakeholders consulted. Through the integration of feedback from end users, ACTaide has the potential to address barriers to between-session practice, enhancing therapeutic outcomes and bridging the gap between therapy sessions in a creative, innovative, and scalable way.

Acknowledgments

This work was supported by a Canadian Institutes of Health Research doctoral award (193289) awarded to ST. The authors gratefully acknowledge the financial support provided by the James McGill Research Fund. These funding agencies were not involved in the study design, analysis, writing of the manuscript, or decision to submit the manuscript for publication. The authors would also like to thank the therapists and clients who participated in this study.

Data Availability

Data sharing is not applicable to this article. The datasets generated or analyzed during this study (ie, the audio recordings of the focus group sessions and the corresponding transcripts) are not publicly available because of confidentiality concerns and ethical restrictions imposed by our institutional review board. In addition, deidentified transcripts cannot be shared because not all participants provided consent for their deidentified data to be used in future, unspecified research. Moreover, given the inherently collective nature of focus group discussions, it is not feasible to disaggregate participant contributions.

Authors' Contributions

ST led the conception of the study design; coordinated the study; collected, analyzed, and interpreted the data; and drafted the manuscript. DQ supported the coordination of the study; collected, analyzed, and interpreted the data; and supported the writing of the manuscript. BK oversaw the design and coordination of the study and revised the manuscript. All authors read and approved the final manuscript submitted for publication.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Iteratively designed low-fidelity wireframes.
[DOCX File , 6986 KB - formative_v9i1e69532_app1.docx]

Multimedia Appendix 2
Summary of the design changes.
[PDF File (Adobe PDF File), 183 KB - formative_v9i1e69532_app2.pdf]

Multimedia Appendix 3
Semistructured interview guide.
[DOCX File , 15 KB - formative_v9i1e69532_app3.docx]

XSL•FO
RenderX

References

1. GBD 2019 Diseases and Injuries Collaborators. Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet 2020 Oct 17;396(10258):1204-1222 [FREE Full text] [doi: 10.1016/S0140-6736(20)30925-9] [Medline: 33069326]
2. Gloster AT, Walder N, Levin ME, Twohig MP, Karekla M. The empirical status of acceptance and commitment therapy: a review of meta-analyses. J Contextual Behav Sci 2020 Oct;18:181-192. [doi: 10.1016/j.jcbs.2020.09.009]
3. Helbig S, Fehm L. Problems with homework in CBT: rare exception or rather frequent? Behav Cogn Psychother 2004 Jun 14;32(3):291-301. [doi: 10.1017/S1352465804001365]
4. Hayes SC, Strosahl KD, Wilson KG, Strosahl K. Acceptance and Commitment Therapy: The Process and Practice of Mindful Change. New York, NY: Guilford Publications; 2012.
5. Strunk DR. Homework. Cogn Behav Pract 2022 Aug;29(3):560-563. [doi: 10.1016/j.cbpra.2022.02.009]
6. Kazantzis N, Whittington C, Zelencich L, Kyrios M, Norton PJ, Hofmann SG. Quantity and quality of homework compliance: a meta-analysis of relations with outcome in cognitive behavior therapy. Behav Ther 2016 Sep;47(5):755-772. [doi: 10.1016/j.beth.2016.05.002] [Medline: 27816086]
7. Mausbach BT, Moore R, Roesch S, Cardenas V, Patterson TL. The relationship between homework compliance and therapy outcomes: an updated meta-analysis. Cognit Ther Res 2010 Oct;34(5):429-438 [FREE Full text] [doi: 10.1007/s10608-010-9297-z] [Medline: 20930925]
8. Haller E, Watzke B. The role of homework engagement, homework-related therapist behaviors, and their association with depressive symptoms in telephone-based CBT for depression. Cogn Ther Res 2020 Jul 22;45:224-235. [doi: 10.1007/s10608-020-10136-x]
9. Kazantzis N, Lampropoulos GK. Reflecting on homework in psychotherapy: what can we conclude from research and experience? J Clin Psychol 2002 May;58(5):577-585. [doi: 10.1002/jclp.10034] [Medline: 11967882]
10. Tang W, Kreindler D. Supporting homework compliance in cognitive behavioural therapy: essential features of mobile apps. JMIR Ment Health 2017 Jun 08;4(2):e20 [FREE Full text] [doi: 10.2196/mental.5283] [Medline: 28596145]
11. van Aubel E, Bakker JM, Batink T, Michielse S, Goossens L, Lange I, et al. Blended care in the treatment of subthreshold symptoms of depression and psychosis in emerging adults: a randomised controlled trial of Acceptance and Commitment Therapy in Daily-Life (ACT-DL). Behav Res Ther 2020 May;128:103592. [doi: 10.1016/j.brat.2020.103592] [Medline: 32146218]
12. Hayes SC, Pistorello J, Levin ME. Acceptance and commitment therapy as a unified model of behavior change. Couns Psychol 2012 Sep 26;40(7):976-1002. [doi: 10.1177/0011000012460836]
13. Stoddard JA, Afari N. The Big Book of ACT Metaphors: A Practitioner's Guide to Experiential Exercises and Metaphors in Acceptance and Commitment Therapy. Oakland, CA: New Harbinger Publications; 2014.
14. ACT Randomized Controlled Trials (1986 to present). Association for Contextual Behavioral Science. URL: https://contextualscience.org/act_randomized_controlled_trials_1986_to_present [accessed 2025-07-07]
15. Hayes SC, Ciarrochi J, Hofmann SG, Chin F, Sahdra B. Evolving an idionomic approach to processes of change: towards a unified personalized science of human improvement. Behav Res Ther 2022 Sep;156:104155. [doi: 10.1016/j.brat.2022.104155] [Medline: 35863243]
16. A-Tjak JG, Davis ML, Morina N, Powers MB, Smits JA, Emmelkamp PM. A meta-analysis of the efficacy of acceptance and commitment therapy for clinically relevant mental and physical health problems. Psychother Psychosom 2015;84(1):30-36. [doi: 10.1159/000365764] [Medline: 25547522]
17. Han A, Yuen HK, Lee HY, Zhou X. Effects of acceptance and commitment therapy on process measures of family caregivers: a systematic review and meta-analysis. J Contextual Behav Sci 2020 Oct;18:201-213. [doi: 10.1016/j.jcbs.2020.10.004]
18. Doing what matters in times of stress. World Health Organization. 2020 Apr 29. URL: https://www.who.int/publications/i/item/9789240003927 [accessed 2023-10-15]
19. Twohig MP, Pierson HM, Hayes SC. Acceptance and commitment therapy. In: Kazantzis N, LǏAbate L, editors. Handbook of Homework Assignments in Psychotherapy. Boston, MA: Springer; 2007.
20. Hayes SC, Strosahl K, Wilson KG. Acceptance and Commitment Therapy: An Experiential Approach to Behavior Change. New York, NY: Guilford Publications; 1999.
21. Rolffs JL, Rogge RD, Wilson KG. Disentangling components of flexibility via the Hexaflex model: development and validation of the multidimensional psychological flexibility inventory (MPFI). Assessment 2018 Jun;25(4):458-482. [doi: 10.1177/1073191116645905] [Medline: 27152011]
22. Twohig MP. Acceptance and commitment therapy: introduction. Cogn Behav Pract 2012 Nov;19(4):499-507. [doi: 10.1016/j.cbpra.2012.04.003]
23. Foody M, Barnes-Holmes Y, Barnes-Holmes D, Törneke N, Luciano C, Stewart I, et al. RFT for clinical use: the example of metaphor. J Contextual Behav Sci 2014 Oct;3(4):305-313. [doi: 10.1016/j.jcbs.2014.08.001]
24. Varra AA, Drossel C, Hayes SC. The use of metaphor to establish acceptance and mindfulness. In: Didonna F, editor. Clinical Handbook of Mindfulness. New York, NY: Springer; 2009.
25. Webster M. Introduction to acceptance and commitment therapy. Adv Psychiatr Treat 2018 Jan 02;17(4):309-316. [doi: 10.1192/apt.bp.107.005256]

XSL•FO
RenderX

26. Eifert GH, Forsyth JP, Arch J, Espejo E, Keller M, Langer D. Acceptance and commitment therapy for anxiety disorders: three case studies exemplifying a unified treatment protocol. Cogn Behav Pract 2009 Nov;16(4):368-385. [doi: 10.1016/j.cbpra.2009.06.001]
27. Kazantzis N, Lampropoulos GK, Deane FP. A national survey of practicing psychologists' use and attitudes toward homework in psychotherapy. J Consult Clin Psychol 2005;73(4):742-748. [doi: 10.1037/0022-006x.73.4.742]
28. Marteau TM, Hollands GJ, Fletcher PC. Changing human behavior to prevent disease: the importance of targeting automatic processes. Science 2012 Sep 21;337(6101):1492-1495. [doi: 10.1126/science.1226918] [Medline: 22997327]
29. Bryant MJ, Simons AD, Thase ME. Therapist skill and patient variables in homework compliance: controlling an uncontrolled variable in cognitive therapy outcome research. Cogn Ther Res 1999;23:381-399. [doi: 10.1023/A:1018703901116]
30. Detweiler-Bedell JB, Whisman MA. A lesson in assigning homework: therapist, client, and task characteristics in cognitive therapy for depression. Prof Psychol Res Pract 2005;36(2):219-223. [doi: 10.1037/0735-7028.36.2.219]
31. Thase ME, Callan JA. The role of homework in cognitive behavior therapy of depression. J Psychother Integr 2006;16(2):162-177. [doi: 10.1037/1053-0479.16.2.162]
32. Chandrashekar P. Do mental health mobile apps work: evidence and recommendations for designing high-efficacy mental health mobile apps. Mhealth 2018 Mar 23;4:6 [FREE Full text] [doi: 10.21037/mhealth.2018.03.02] [Medline: 29682510]
33. Lindhiem O, Bennett CB, Rosen D, Silk J. Mobile technology boosts the effectiveness of psychotherapy and behavioral interventions: a meta-analysis. Behav Modif 2015 Nov;39(6):785-804 [FREE Full text] [doi: 10.1177/0145445515595198] [Medline: 26187164]
34. Torous J, Levin ME, Ahern DK, Oser ML. Cognitive behavioral mobile applications: clinical studies, marketplace overview, and research agenda. Cogn Behav Pract 2017 May;24(2):215-225. [doi: 10.1016/j.cbpra.2016.05.007]
35. ACT on it. Google Play. URL: https://play.google.com/store/apps/details?id=com.subtlelabs.mindfulness.actonit&hl=en_CA [accessed 2024-11-18]
36. PsychFlex. Google Play. URL: https://play.google.com/store/search?q=psychflex&c=apps&hl=en_CA [accessed 2024-11-18]
37. Ahtinen A, Mattila E, Välkkynen P, Kaipainen K, Vanhala T, Ermes M, et al. Mobile mental wellness training for stress management: feasibility and design implications based on a one-month field study. JMIR Mhealth Uhealth 2013 Jul 10;1(2):e11 [FREE Full text] [doi: 10.2196/mhealth.2596] [Medline: 25100683]
38. Bricker JB, Mull KE, Kientz JA, Vilardaga R, Mercer LD, Akioka KJ, et al. Randomized, controlled pilot trial of a smartphone app for smoking cessation using acceptance and commitment therapy. Drug Alcohol Depend 2014 Oct 01;143:87-94 [FREE Full text] [doi: 10.1016/j.drugalcdep.2014.07.006] [Medline: 25085225]
39. Bricker JB, Watson NL, Mull KE, Sullivan BM, Heffner JL. Efficacy of smartphone applications for smoking cessation: a randomized clinical trial. JAMA Intern Med 2020 Nov 01;180(11):1472-1480 [FREE Full text] [doi: 10.1001/jamainternmed.2020.4055] [Medline: 32955554]
40. Catella S, Gendreau RM, Kraus AC, Vega N, Rosenbluth MJ, Soefje S, et al. Self-guided digital acceptance and commitment therapy for fibromyalgia management: results of a randomized, active-controlled, phase II pilot clinical trial. J Behav Med 2024 Feb;47(1):27-42 [FREE Full text] [doi: 10.1007/s10865-023-00429-3] [Medline: 37382794]
41. Fledderus M, Schreurs KM, Bohlmeijer ET, Vollenbroek-Hutten MM. Development and pilot evaluation of an online relapse-prevention program based on acceptance and commitment therapy for chronic pain patients. JMIR Hum Factors 2015 Jan 05;2(1):e1 [FREE Full text] [doi: 10.2196/humanfactors.3302] [Medline: 27025550]
42. García-Torres F, García-Carmona M, Rubio S, Luque B, Gómez-Solís Á, Aranda E. Feasibility study of a mobile application (ACT-ON) to complement acceptance and commitment therapy (ACT) intervention in cancer patients. Int J Med Inform 2023 Sep;177:105131. [doi: 10.1016/j.ijmedinf.2023.105131] [Medline: 37354875]
43. Gentili C, Zetterqvist V, Rickardsson J, Holmström L, Simons LE, Wicksell RK. ACTsmart - development and feasibility of digital acceptance and commitment therapy for adults with chronic pain. NPJ Digit Med 2020 Feb 13;3:20 [FREE Full text] [doi: 10.1038/s41746-020-0228-4] [Medline: 32128450]
44. Haeger JA. Utilizing act daily as a self-guided mobile app intervention for depression and anxiety in a college counseling center. Utah State University. 2016. URL: https://www.proquest.com/docview/1864629015/abstract/D3B28FF50CCB455DPQ/1 [accessed 2024-11-10]
45. Hemmings NR, Kawadler JM, Whatmough R, Ponzo S, Rossi A, Morelli D, et al. Development and feasibility of a digital acceptance and commitment therapy-based intervention for generalized anxiety disorder: pilot acceptability study. JMIR Form Res 2021 Feb 09;5(2):e21737 [FREE Full text] [doi: 10.2196/21737] [Medline: 33560232]
46. Hitch C, Leightley D, Murphy D, Dymond S. ACT Vet: a mixed-methods pilot study of efficacy, usability and cost-effectiveness of an Acceptance and Commitment Therapy (ACT) based app for veterans with harmful gambling and PTSD. JMIR Preprints. Preprint posted online on October 24, 2023 2025 [FREE Full text] [doi: 10.2196/preprints.53924]
47. Kroska EB, Hoel S, Victory A, Murphy SA, McInnis MG, Stowe ZN, et al. Optimizing an acceptance and commitment therapy microintervention via a mobile app with two cohorts: protocol for micro-randomized trials. JMIR Res Protoc 2020 Sep 23;9(9):e17086 [FREE Full text] [doi: 10.2196/17086] [Medline: 32965227]
48. Levin ME, Pierce B, Schoendorff B. The acceptance and commitment therapy matrix mobile app: a pilot randomized trial on health behaviors. J Contextual Behav Sci 2017 Jul;6(3):268-275. [doi: 10.1016/j.jcbs.2017.05.003]

XSL•FO
RenderX

49. Ly KH, Asplund K, Andersson G. Stress management for middle managers via an acceptance and commitment-based smartphone application: a randomized controlled trial. Internet Interv 2014 Jul;1(3):95-101. [doi: 10.1016/j.invent.2014.06.003]
50. Sargisson RJ, Li F, Lobo D, Roche M. Using an acceptance and commitment therapy app to reduce anxiety for students and employees. Int J Psychol Behav Anal 2019 Dec 26;5(2):164. [doi: 10.15344/2455-3867/2019/164]
51. Short B. The effectiveness of an ACT app in promoting wellbeing and psychological flexibility. Murray State University. 2021. URL: https://digitalcommons.murraystate.edu/honorstheses/101 [accessed 2024-07-07]
52. Tighe J, Shand F, McKay K, Mcalister TJ, Mackinnon A, Christensen H. Usage and acceptability of the iBobbly app: pilot trial for suicide prevention in Aboriginal and Torres Strait Islander youth. JMIR Ment Health 2020 Dec 01;7(12):e14296 [FREE Full text] [doi: 10.2196/14296] [Medline: 33258782]
53. Vaessen T, Steinhart H, Batink T, Klippel A, Van Nierop M, Reininghaus U, et al. ACT in daily life in early psychosis: an ecological momentary intervention approach. Psychosis 2019 Mar 19;11(2):93-104. [doi: 10.1080/17522439.2019.1578401]
54. Zucchelli F, Donnelly O, Rush E, White P, Gwyther H, Williamson H. An acceptance and commitment therapy prototype mobile program for individuals with a visible difference: mixed methods feasibility study. JMIR Form Res 2022 Jan 21;6(1):e33449 [FREE Full text] [doi: 10.2196/33449] [Medline: 35060908]
55. ACT iCoach: Acceptance Commitm. Google Play. URL: https://play.google.com/store/apps/details?id=co.swasth.actcoach&hl=en_CA [accessed 2024-11-28]
56. ACT Companion: Happiness Trap. Apple Inc. URL: https://apps.apple.com/ca/app/act-companion-happiness-trap/id668468577 [accessed 2024-11-28]
57. Baumel A, Muench F, Edan S, Kane JM. Objective user engagement with mental health apps: systematic search and panel-based usage analysis. J Med Internet Res 2019 Sep 25;21(9):e14567 [FREE Full text] [doi: 10.2196/14567] [Medline: 31573916]
58. Torous J, Nicholas J, Larsen ME, Firth J, Christensen H. Clinical review of user engagement with mental health smartphone apps: evidence, theory and improvements. Evid Based Ment Health 2018 Aug;21(3):116-119 [FREE Full text] [doi: 10.1136/eb-2018-102891] [Medline: 29871870]
59. Krafft J, Ong CW, Davis CH, Petersen JM, Levin ME, Twohig MP. An open trial of group acceptance and commitment therapy with an adjunctive mobile app for generalized anxiety disorder. Cogn Behav Pract 2022 Nov;29(4):846-859. [doi: 10.1016/j.cbpra.2021.05.008]
60. Mattila E, Lappalainen R, Välkkynen P, Sairanen E, Lappalainen P, Karhunen L, et al. Usage and dose response of a mobile acceptance and commitment therapy app: secondary analysis of the intervention arm of a randomized controlled trial. JMIR Mhealth Uhealth 2016 Jul 28;4(3):e90 [FREE Full text] [doi: 10.2196/mhealth.5241] [Medline: 27468653]
61. Alqahtani F, Orji R. Insights from user reviews to improve mental health apps. Health Informatics J 2020 Sep 10;26(3):2042-2066 [FREE Full text] [doi: 10.1177/1460458219896492] [Medline: 31920160]
62. Vial S, Boudhraâ S, Dumont M. Human-centered design approaches in digital mental health interventions: exploratory mapping review. JMIR Ment Health 2022 Jun 07;9(6):e35591 [FREE Full text] [doi: 10.2196/35591] [Medline: 35671081]
63. Pierce B, Twohig MP, Levin ME. Perspectives on the use of acceptance and commitment therapy related mobile apps: results from a survey of students and professionals. J Context Behav Sci 2016 Oct;5(4):215-224. [doi: 10.1016/j.jcbs.2016.08.001]
64. Di Sante J, Frayn M, Angelescu A, Knäuper B. Proof-of-concept testing of a brief virtual ACT workshop for emotional eating. Appetite 2024 Aug 01;199:107386 [FREE Full text] [doi: 10.1016/j.appet.2024.107386] [Medline: 38692511]
65. Harris R. ACT Made Simple: An Easy-To-Read Primer on Acceptance and Commitment Therapy. Oakland, CA: New Harbinger Publications; 2019.
66. Kordjazi Z. The effect of visual mnemonic support practice on the reading comprehension of psychology texts. Eur Online J Nat Soc Sci 2014;3(4) [FREE Full text]
67. Guo D, Zhang S, Wright KL, McTigue EM. Do you get the picture? A meta-analysis of the effect of graphics on reading comprehension. AERA Open 2020 Jan 24;6(1). [doi: 10.1177/2332858420901696]
68. Bobek E, Tversky B. Creating visual explanations improves learning. Cogn Res Princ Implic 2016;1(1):27 [FREE Full text] [doi: 10.1186/s41235-016-0031-6] [Medline: 28180178]
69. Fletcher JD, Tobias S. The multimedia principle. In: Mayer R, editor. The Cambridge Handbook of Multimedia Learning. Cambridge, UK: Cambridge University Press; 2005.
70. Katz MG, Kripalani S, Weiss BD. Use of pictorial aids in medication instructions: a review of the literature. Am J Health Syst Pharm 2006 Dec 01;63(23):2391-2397. [doi: 10.2146/ajhp060162] [Medline: 17106013]
71. Levie WH, Lentz R. Effects of text illustrations: a review of research. ECTJ 1982 Dec;30(4):195-232. [doi: 10.1007/BF02765184]
72. Norris EM. The constructive use of images in medical teaching: a literature review. JRSM Short Rep 2012 May 23;3(5):33-38 [FREE Full text] [doi: 10.1258/shorts.2012.011158] [Medline: 22666530]
73. Mansoor LE, Dowse R. Effect of pictograms on readability of patient information materials. Ann Pharmacother 2003;37(7-8):1003-1009. [doi: 10.1345/aph.1C449] [Medline: 12841808]

XSL•FO
RenderX

74. McCurdie T, Taneva S, Casselman M, Yeung M, McDaniel C, Ho W, et al. mHealth consumer apps: the case for user-centered design. Biomed Instrum Technol 2012;Suppl:49-56. [doi: 10.2345/0899-8205-46.s2.49] [Medline: 23039777]
75. Farao J, Malila B, Conrad N, Mutsvangwa T, Rangaka MX, Douglas TS. A user-centred design framework for mHealth. PLoS One 2020;15(8):e0237910 [FREE Full text] [doi: 10.1371/journal.pone.0237910] [Medline: 32813711]
76. Saparamadu AA, Fernando P, Zeng P, Teo H, Goh A, Lee JM, et al. User-centered design process of an mHealth app for health professionals: case study. JMIR Mhealth Uhealth 2021 Mar 26;9(3):e18079 [FREE Full text] [doi: 10.2196/18079] [Medline: 33769297]
77. Alqahtani F, Winn A, Orji R. Co-designing a mobile app to improve mental health and well-being: focus group study. JMIR Form Res 2021 Feb 26;5(2):e18172 [FREE Full text] [doi: 10.2196/18172] [Medline: 33635281]
78. Brotherdale R, Berry K, Branitsky A, Bucci S. Co-producing digital mental health interventions: a systematic review. Digit Health 2024 Apr 25;10:20552076241239172 [FREE Full text] [doi: 10.1177/20552076241239172] [Medline: 38665886]
79. Vial S, Boudhraâ S, Dumont M, Tremblay M, Riendeau S. Developing a mobile app with a human-centered design lens to improve access to mental health care (Mentallys Project): protocol for an initial co-design process. JMIR Res Protoc 2023 Aug 22;12:e47220 [FREE Full text] [doi: 10.2196/47220] [Medline: 37606978]
80. Krueger RA, Casey MA. Focus Groups: A Practical Guide for Applied Research. Thousand Oaks, CA: SAGE Publications; 2014.
81. Morgan DL. Basic and Advanced Focus Groups. Thousand Oaks, CA: SAGE Publications; 2018.
82. Psychology Today homepage. Psychology Today. URL: https://www.psychologytoday.com/us [accessed 2025-07-11]
83. Hunsley J, Lee CM, Aubry T. Who uses psychological services in Canada? Can Psychol 1999 Aug;40(3):232-240. [doi: 10.1037/h0086839]
84. Breen RL. A practical guide to focus-group research. J Geogr High Educ 2006;30(3):463-475. [doi: 10.1080/03098260600927575]
85. Carlsen B, Glenton C. What about N? A methodological study of sample-size reporting in focus group studies. BMC Med Res Methodol 2011 Mar 11;11:26 [FREE Full text] [doi: 10.1186/1471-2288-11-26] [Medline: 21396104]
86. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
87. Green J, Thorogood N. Qualitative Methods for Health Research. Thousand Oaks, CA: SAGE Publications; 2018.
88. Stake R. The Art of Case Study Research. Thousand Oaks, CA: SAGE Publications; 1995.
89. Yin RK. Case Study Research and Applications: Design and Methods. Thousand Oaks, CA: SAGE Publications; 2017.
90. ISO 9241-11:2018 Ergonomics of human-system interaction part 11: usability: definitions and concepts. International Organization for Standardization. 2018. URL: https://www.iso.org/standard/63500.html [accessed 2024-11-27]
91. Oyebode O, Alqahtani F, Orji R. Using machine learning and thematic analysis methods to evaluate mental health apps based on user reviews. IEEE Access 2020 Jun 12;8:111141-111158. [doi: 10.1109/access.2020.3002176]
92. Chan AH, Honey ML. User perceptions of mobile digital apps for mental health: acceptability and usability - an integrative review. J Psychiatr Ment Health Nurs 2022 Feb;29(1):147-168. [doi: 10.1111/jpm.12744] [Medline: 33604946]
93. Schueller SM, Washburn JJ, Price M. Exploring mental health providers' interest in using web and mobile-based tools in their practices. Internet Interv 2016 May;4(2):145-151 [FREE Full text] [doi: 10.1016/j.invent.2016.06.004] [Medline: 28090438]
94. Nielsen J. 10 usability heuristics for user interface design. Nielsen Norman Group. URL: https://www.nngroup.com/articles/ten-usability-heuristics/ [accessed 2024-11-25]
95. Parente Da Costa R, Canedo ED, De Sousa RT, De Oliveira Albuquerque R, Garcia Villalba LJ. Set of usability heuristics for quality assessment of mobile applications on smartphones. IEEE Access 2019 Apr 11;7:116145-116161. [doi: 10.1109/access.2019.2910778]
96. Wei Z, Kuang Y, Gao Z. Measurement and optimization of navigation bar icon design in mental health apps. In: Design Studies and Intelligence Engineering. Amsterdam, The Netherlands: IOS Press; Feb 2024.
97. Morey SA, Stuck RE, Chong AW, Barg-Walkow LH, Mitzner TL, Rogers WA. Mobile health apps: improving usability for older adult users. Ergon Des Q Hum Factors Appl 2019 Apr 01;27(4):4-13. [doi: 10.1177/1064804619840731]
98. Strahm B, Gray CM, Vorvoreanu M. Generating mobile application onboarding insights through minimalist instruction. In: Proceedings of the 2018 Designing Interactive Systems Conference. 2018 Presented at: DIS '18; June 9-13, 2018; Hong Kong, Hong Kong. [doi: 10.1145/3196709.3196727]
99. Chan G, Alslaity A, Wilson R, Orji R. Feeling moodie: insights from a usability evaluation to improve the design of mHealth apps. Int J Hum Comput Interact 2023 Aug 09;40(19):5857-5875. [doi: 10.1080/10447318.2023.2241613]
100. Garrido S, Doran B, Oliver E, Boydell K. Desirable design: what aesthetics are important to young people when designing a mental health app? Health Informatics J 2024;30(4):14604582241295948 [FREE Full text] [doi: 10.1177/14604582241295948] [Medline: 39504119]
101. Schueller SM, Neary M, O'Loughlin K, Adkins EC. Discovery of and interest in health apps among those with mental health needs: survey and focus group study. J Med Internet Res 2018 Jun 11;20(6):e10141 [FREE Full text] [doi: 10.2196/10141] [Medline: 29891468]

XSL•FO
RenderX

102. Zhou C, Jiang Z, Wang Z. Research on color image of interactive interface of medical and health app. In: Proceedings of the 2nd International Conference on Information Science and Education. 2021 Presented at: ICISE-IE 2021; November 26-28, 2021; Chongqing, China. [doi: 10.1109/icise-ie53922.2021.00354]
103. Burn AM, Ford TJ, Stochl J, Jones PB, Perez J, Anderson JK. Developing a web-based app to assess mental health difficulties in secondary school pupils: qualitative user-centered design study. JMIR Form Res 2022 Jan 10;6(1):e30565 [FREE Full text] [doi: 10.2196/30565] [Medline: 35006079]
104. Jakob R, Harperink S, Rudolf AM, Fleisch E, Haug S, Mair JL, et al. Factors influencing adherence to mHealth apps for prevention or management of noncommunicable diseases: systematic review. J Med Internet Res 2022 May 25;24(5):e35371 [FREE Full text] [doi: 10.2196/35371] [Medline: 35612886]
105. Vo V, Auroy L, Sarradon-Eck A. Patients' perceptions of mHealth apps: meta-ethnographic review of qualitative studies. JMIR Mhealth Uhealth 2019 Jul 10;7(7):e13817 [FREE Full text] [doi: 10.2196/13817] [Medline: 31293246]
106. Wei Y, Zheng P, Deng H, Wang X, Li X, Fu H. Design features for improving mobile health intervention user engagement: systematic review and thematic analysis. J Med Internet Res 2020 Dec 09;22(12):e21687 [FREE Full text] [doi: 10.2196/21687] [Medline: 33295292]
107. Cheng VW. Recommendations for implementing gamification for mental health and wellbeing. Front Psychol 2020 Dec 7;11:586379 [FREE Full text] [doi: 10.3389/fpsyg.2020.586379] [Medline: 33365001]
108. Fleming T, Poppelaars M, Thabrew H. The role of gamification in digital mental health. World Psychiatry 2023 Feb 14;22(1):46-47 [FREE Full text] [doi: 10.1002/wps.21041] [Medline: 36640378]
109. Six SG, Byrne KA, Tibbett TP, Pericot-Valverde I. Examining the effectiveness of gamification in mental health apps for depression: systematic review and meta-analysis. JMIR Ment Health 2021 Nov 29;8(11):e32199 [FREE Full text] [doi: 10.2196/32199] [Medline: 34847058]
110. Cheng VW, Davenport T, Johnson D, Vella K, Hickie IB. Gamification in apps and technologies for improving mental health and well-being: systematic review. JMIR Ment Health 2019 Jun 26;6(6):e13717 [FREE Full text] [doi: 10.2196/13717] [Medline: 31244479]
111. Johnson D, Deterding S, Kuhn KA, Staneva A, Stoyanov S, Hides L. Gamification for health and wellbeing: a systematic review of the literature. Internet Interv 2016 Nov;6:89-106 [FREE Full text] [doi: 10.1016/j.invent.2016.10.002] [Medline: 30135818]
112. Auf H, Dagman J, Renström S, Chaplin J. Gamification and nudging techniques for improving user engagement in mental health and well-being apps. Proc Des Soc 2021 Aug;1:1647-1656 [FREE Full text] [doi: 10.1017/pds.2021.426]
113. Hashim MJ. User interactivity in eHealth applications: a novel taxonomy. In: Proceedings of the 12th International Conference on Innovations in Information Technology. 2016 Presented at: IIT 2016; November 28-30, 2016; Al Ain, United Arab Emirates. [doi: 10.1109/innovations.2016.7880023]
114. Titzler I, Saruhanjan K, Berking M, Riper H, Ebert DD. Barriers and facilitators for the implementation of blended psychotherapy for depression: a qualitative pilot study of therapists' perspective. Internet Interv 2018 Jun;12:150-164 [FREE Full text] [doi: 10.1016/j.invent.2018.01.002] [Medline: 30135779]
115. Ralston AL, Andrews ARIII, Hope DA. Fulfilling the promise of mental health technology to reduce public health disparities: review and research agenda. Clin Psychol Sci Pract 2019 Mar;26(1). [doi: 10.1111/cpsp.12277]
116. Borghouts J, Eikey E, Mark G, De Leon C, Schueller SM, Schneider M, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. J Med Internet Res 2021 Mar 24;23(3):e24387 [FREE Full text] [doi: 10.2196/24387] [Medline: 33759801]
117. Camacho E, Cohen A, Torous J. Assessment of mental health services available through smartphone apps. JAMA Netw Open 2022 Dec 01;5(12):e2248784 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.48784] [Medline: 36576737]
118. Gaudiano BA, Herbert JD, Hayes SC. Is it the symptom or the relation to it? Investigating potential mediators of change in acceptance and commitment therapy for psychosis. Behav Ther 2010 Dec;41(4):543-554 [FREE Full text] [doi: 10.1016/j.beth.2010.03.001] [Medline: 21035617]
119. Schnall R, Rojas M, Bakken S, Brown W, Carballo-Dieguez A, Carry M, et al. A user-centered model for designing consumer mobile health (mHealth) applications (apps). J Biomed Inform 2016 Apr;60:243-251 [FREE Full text] [doi: 10.1016/j.jbi.2016.02.002] [Medline: 26903153]
120. Czajkowski SM, Powell LH, Adler N, Naar-King S, Reynolds KD, Hunter CM, et al. From ideas to efficacy: the ORBIT model for developing behavioral treatments for chronic diseases. Health Psychol 2015 Oct;34(10):971-982 [FREE Full text] [doi: 10.1037/hea0000161] [Medline: 25642841]
121. Li SH, Achilles MR, Spanos S, Habak S, Werner-Seidler A, O'Dea B. A cognitive behavioural therapy smartphone app for adolescent depression and anxiety: co-design of ClearlyMe. Cogn Behav Ther 2022 Mar 16;15:e13. [doi: 10.1017/s1754470x22000095]
122. Osma J, Martínez-García L, Prado-Abril J, Peris-Baquero Ó, González-Pérez A. Developing a smartphone app based on the unified protocol for the transdiagnostic treatment of emotional disorders: a qualitative analysis of users and professionals' perspectives. Internet Interv 2022 Dec;30:100577 [FREE Full text] [doi: 10.1016/j.invent.2022.100577] [Medline: 36213084]

XSL•FO
RenderX

123. Lipschitz J, Miller CJ, Hogan TP, Burdick KE, Lippin-Foster R, Simon SR, et al. Adoption of mobile apps for depression and anxiety: cross-sectional survey study on patient interest and barriers to engagement. JMIR Ment Health 2019 Jan 25;6(1):e11334 [FREE Full text] [doi: 10.2196/11334] [Medline: 30681968]
124. Focus on geography series, 2016 census. Statistics Canada. 2016. URL: https://www12.statcan.gc.ca/census-recensement/2016/as-sa/fogs-spg/Facts-cma-eng.cfm?LANG=Eng&GK=CMA&GC=462&TOPIC=7 [accessed 2024-10-28]
125. Hunsley J, Aubry T, Lee C. A profile of Canadian consumers of psychological services. Canadian Psychological Association. 1997. URL: https://www.cpa.ca/documents/Hunsley_report.html [accessed 2024-10-27]
126. A profile of selected mental health and substance use health care providers in Canada, 2021. Canadian Institute for Health Information. URL: https://www.cihi.ca/en/a-profile-of-selected-mental-health-and-substance-use-health-care-providers-in-canada-2021 [accessed 2024-10-28]
127. Mental health in Ontario: insights from askforhelptoday.ca. Ontario Psychological Association. 2024 Apr 02. URL: https://tinyurl.com/y7e6vzh5 [accessed 2024-07-07]

Abbreviations

ACT: acceptance and commitment therapy

Edited by A Mavragani; submitted 02.12.24; peer-reviewed by R Nieto, C Abbatantuono; comments to author 06.06.25; revised version received 10.06.25; accepted 27.06.25; published 24.07.25.

Please cite as:
Thapar S, Quesada D, Knäuper B
Therapy Companion Mobile App for Acceptance and Commitment Therapy Exercises (ACTaide): Therapist and Client Co-Design Study
JMIR Form Res 2025;9:e69532
URL: https://formative.jmir.org/2025/1/e69532
doi:10.2196/69532
PMID:40705405

©Serena Thapar, Daniela Quesada, Bärbel Knäuper. Originally published in JMIR Formative Research (https://formative.jmir.org), 24.07.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Service Quality Assessment of Digital Health Solutions in Outpatient Care: Qualitative Item Repository Development Study

Dominik Rigo[1], MSc; Leonard Fehring[1,2], Priv-Doz, Dr Med; Achim Mortsiefer[3], Prof Dr Med; Sven Meister[1,4], Prof Dr

[1]Health Care Informatics, Faculty of Health, School of Medicine, Witten/Herdecke University, Witten, Germany

[2]Department of Gastroenterology, Helios University Hospital Wuppertal, Wuppertal, Germany

[3]General Practice II and Patient-Centredness in Primary Care, Institute of General Practice and Primary Care, Faculty of Health, School of Medicine, Witten/Herdecke University, Witten, Germany

[4]Department Healthcare, Fraunhofer Institute for Software and Systems Engineering ISST, Dortmund, Germany

Corresponding Author:
Sven Meister, Prof Dr
Health Care Informatics, Faculty of Health
School of Medicine
Witten/Herdecke University
Pferdebachstraße 11
Witten, 58455
Germany
Phone: 49 2302926 ext 78629
Email: sven.meister@uni-wh.de

Abstract

Background: The integration of digital health solutions (DHSs) into health care systems has the potential to significantly enhance service delivery and health outcomes. Despite their benefits, the adoption remains slow, especially in outpatient care, and is hindered by various barriers, such as unclear effectiveness and high costs.

Objective: This study aimed to address the uncertainties regarding the cost-benefit ratio of DHSs by developing a comprehensive instrument to evaluate their impact on health care service quality across diverse settings (eg, across different diseases or types of DHSs).

Methods: We conducted a multistaged rapid review and semistructured, qualitative interviews to identify and adapt existing instruments evaluating the effects of DHSs. The first rapid review screened 4957 records and included 40 relevant papers to identify instruments currently used for DHS assessment after their deployment, yielding a total of 126 reported outcomes. Subsequently, we conducted interviews with 19 health care practitioners across 4 countries to validate and refine the 7 health care service quality dimensions derived from merging the Outpatient Experience Questionnaire (OPEQ), selected after the first rapid review, and Health Care Service Quality (HEALTHQUAL), an established instrument for measuring health care service quality derived from gray literature. On the basis of the results of the interviews, a second rapid review with 35 papers out of 493 screened records was conducted to identify instruments used to measure patient satisfaction, yielding a total of 29 patient satisfaction instruments.

Results: From the first rapid review, OPEQ was selected out of 18 relevant instruments identified among the 126 reported outcomes and combined with HEALTHQUAL. The interviews with health care professionals confirmed the relevance of all 7 health care service quality dimensions derived from OPEQ and HEALTHQUAL. In addition, 4 interviewees mentioned patient satisfaction as a further dimension missing in the framework presented during the interviews. From the subsequent rapid review, the Patient Satisfaction Questionnaire-Short Form was selected out of 6 relevant instruments identified among the 29 identified patient satisfaction instruments. By combining HEALTHQUAL, OPEQ, and Patient Satisfaction Questionnaire-Short Form, we derived the Digital Healthcare Service Quality (DigiHEALTHQUAL) questionnaire, which consists of 51 items across 8 dimensions, including accessibility, efficiency, empathy, general satisfaction, degree of improvements of care services, information, safety, and tangibles.

Conclusions: The DigiHEALTHQUAL questionnaire aims to provide a standardized approach for assessing the impact of DHSs on health care service quality across various use cases, therapeutic areas, and perspectives, facilitating comparison between DHSs and supporting decision makers in resource allocation and implementation decisions. Future research will focus on validating

XSL•FO
RenderX

the DigiHEALTHQUAL in real-life settings and further refining it to comprehensively encompass both patient and health care practitioner perspectives.

(JMIR Form Res 2025;9:e68276) doi:10.2196/68276

KEYWORDS

digital health solutions; health care service quality; patient satisfaction; health service outcomes; implementation outcomes; health technology assessment

Introduction

Theoretical Background

The integration of digital health solutions (DHSs) into health care has become an area of extensive research [1,2]. For our study, we adopted the European Commission's definition of DHSs, which describes them as "tools and services that use information and communication technologies to improve prevention, diagnosis, treatment, monitoring and management of health-related issues and to monitor and manage lifestyle-habits that impact health" [3]. The many ways in which DHSs can transform health care have been shown in various studies, such as achieving cost and time savings [4], improving medical safety and treatment adherence [5], enhancing service quality [6], and ameliorating health outcomes [7].

Despite these benefits, the adoption of DHSs in health care is slow [8]. This is even more pronounced in outpatient settings compared to inpatient settings [9,10]. The key obstacles include required workflow adjustments, high costs for installation or maintenance, and unclear or limited effectiveness [11,12].

These challenges, especially the considerations between effectiveness and costs, underscore the importance of reliably measuring the impact of DHSs. Reliable measurement can support the further adoption of DHSs, that is, by helping decision makers compare the effects of different DHSs, select the most suitable DHS for their organization, and monitor the progress of deployments [12-14].

Moreover, as the landscape of DHSs grows increasingly complex with more and more digital use cases, tools, and providers [15], it is becoming more difficult for decision makers to navigate this landscape. Therefore, robust evaluation frameworks become more important for their support [16]. Recent literature reveals that the instruments currently used to assess the implementation of DHSs often lack robust psychometric properties, leading to a lack of consensus on preferred assessment methodologies [17-19]. Existing validated instruments tend to focus on specific use cases or diseases, which limits their utility for comparison across multiple DHSs and health care settings [20-22]. This variation in quality and scope of reported effects complicates the task for decision makers attempting to compare, select, and implement DHSs.

Objectives

To address the limitations of currently applied instruments, it is essential to develop and establish comprehensive instruments. Existing studies have used a range of broadly applicable instruments, for example, measuring costs, resource use, patient satisfaction, and quality of life [20]. For example, the Mobile Healthcare App Database uses the Mobile Application Rating Scale to increase transparency regarding the quality of content and data security of >1300 DHSs [23,24].

Although service quality is a well-established and extensively studied outcome in various sectors, including health care—with numerous validated instruments available [25]—to our knowledge, established instruments for health care service quality (HCSQ) evaluation have not yet been used to assess the effects of DHSs. Recognizing the necessity for a versatile outcome measure, we considered whether HCSQ might be an effective metric to assess the effects of DHSs in a variety of health care settings. Therefore, this study aimed to address the lack of broadly applicable and reliable instruments by identifying, adapting, or developing a questionnaire that can measure the service quality of DHSs across different settings leveraging existing instruments. The broad adoption of such a tool would be particularly valuable for decision makers when considering the selection and implementation of DHSs.

Methods

Overview

This study used a sequential approach (Figure 1) following predefined construction criteria to create a comprehensive questionnaire that can measure HCSQ in settings with different DHSs, across therapeutic areas, and from both patient and health care practitioner (HCP) perspectives for the assessment of the effects of DHS deployment on HCSQ.

XSL•FO
RenderX

Figure 1. Instrument consolidation and adaptation strategy for the development of Digital Healthcare Service Quality (DigiHEALTHQUAL). The stepwise approach included the development of the theoretical background and identification of Health Care Service Quality (HEALTHQUAL), a first rapid review identifying Outpatient Experience Questionnaire (OPEQ), interviews with health care practitioners (HCPs), and a second rapid review identifying Patient Satisfaction Questionnaire-Short Form (PSQ-18). Consolidation steps ensured alignment with 5 construction criteria: postdeployment applicability, cross-therapeutic relevance, multistakeholder perspectives, outpatient care focus, and applicability across digital use cases. DHS: digital health solution.

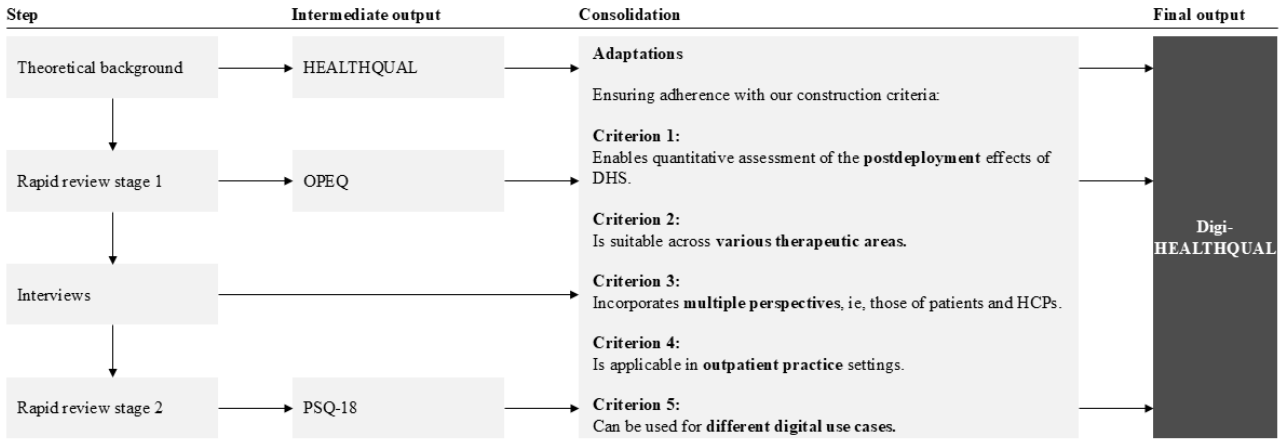

We selected Health Care Service Quality (HEALTHQUAL) [26]—a validated, health care–specific service quality instrument based on the well-established service quality model Service Quality (SERVQUAL) [27]—for further adaptation to the context of this study. As HEALTHQUAL has not previously been used in digital health settings, we conducted a rapid review adhering to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [28], as well as recommendations from Tricco et al [29] and King et al [30]. This review helped identify existing instruments used to assess the postdeployment effects of DHSs. The Outpatient Experience Questionnaire (OPEQ) was selected for the subsequent steps [31,32].

Neither the original HEALTHQUAL nor OPEQ have been validated for use with HCPs. To address this limitation, we conducted semistructured qualitative interviews with 19 HCPs, following the COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist [33]. This process validated the relevance of the included dimensions for HCPs and led to the addition of a dimension for patient satisfaction, which was derived from the Patient Satisfaction Questionnaire-Short Form (PSQ-18) after an additional literature review on patient satisfaction measurement instruments [34].

Consequently, we derived our suggested set of dimensions and items, hereafter referred to as Digital Healthcare Service Quality (DigiHEALTHQUAL).

Construction Criteria for the Questionnaire

To identify, adapt, or develop a questionnaire, we defined the following five criteria (Figure 1):

1. Enables quantitative assessment of the postdeployment effects of DHSs following the definition of Proctor et al [35] for implementation outcomes
2. Is broadly applicable and suitable across therapeutic areas
3. Incorporates multiple perspectives, that is, those of patients and HCPs, including health professionals, health associate professionals, and health management and support personnel, as defined by the World Health Organization (WHO) [36]
4. Is applicable in outpatient practice settings, following the definition provided by the National Cancer Institute [37]
5. Can be used for different digital use cases, that is, those already prioritized for the next research stage following this work (online appointment booking platforms, digital anamnesis tools, video consultations, and electronic patient records)

Selection of Research Framework

One well-established tool in the area of health service is SERVQUAL [27], which is widely used to measure service quality across different sectors [38]. SERVQUAL has often been adapted to measure HCSQ in various health care settings [25,39]. This led to the development of several health care–specific variations of SERVQUAL. The 3 frequently used variations are called HEALTHQUAL and were developed by Miranda et al [40], Lee [26], and Mosadeghrad and Sokhanvar [41]. Despite evidence that DHSs impact HCSQ [1,6], none of these HEALTHQUAL instruments have been used to assess these effects. Therefore, this study aimed to fill this gap by adapting HEALTHQUAL by Lee [26] to evaluate the impact of DHSs on HCSQ.

For this study, we used HEALTHQUAL by Lee [26] because of its comprehensive design approach, which incorporated input from experts and patients as well as international health care service accreditation systems. HEALTHQUAL by Lee [26] closely aligns with the original SERVQUAL framework and adds a dimension for health outcomes, making it a robust tool for assessing HCSQ [26]. In addition, HEALTHQUAL by Lee [26] has been widely applied internationally in both inpatient and outpatient settings (eg, South Korea, India, Syria, and Ghana) [42-47]. The instrument consists of 5 dimensions with a total of 32 items. These dimensions are empathy, tangibles, safety, efficiency, and the degree of improvements of care services.

XSL•FO
RenderX

Multistaged Rapid Review

Overview

We conducted 2 rapid literature reviews following the PRISMA-RR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Rapid Reviews) guidelines and recommendations for rapid reviews outlined by Tricco et al [29] and King et al [30]. As PRISMA-RR is still under development, we adapted PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines (especially PRISMA-ScR [Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews]) to align with the scope of this project. The search strategy and selection criteria are detailed below. More details can be found in Multimedia Appendices 1 and 2. Owing to the larger number of abstracts retrieved after title screening in the rapid review stage 1, we used ASReview LAB (Zenodo), an artificial intelligence–assisted screening tool, to expedite the abstract screening process [48]. This step was terminated upon meeting predefined stopping criteria in line with recent literature [49,50], specifically after screening at least 50% (422/843) of the abstracts and encountering at least 50 consecutive irrelevant abstracts. Data from the included studies were synthesized by extracting details from the reported findings and the instruments applied. This information was used to select suitable instruments for the following instrument development.

Rapid Review Stage 1

Overview

We conducted our search across 3 databases: PubMed, Scopus, and APA PsycINFO. Our search used a structured string comprising 4 conceptual blocks: *digital health*, *implementation*, *effects*, and *assessment*. The search terms were adapted from recent reviews within this field [1,7,17,19,51-60]. Articles were selected according to the criteria described in the following sections.

Inclusion and Exclusion Criterion 1: Type of Article

The articles were included if they reported the outcomes of a comprehensive data collection and analysis process. The articles were excluded if they (1) focused on describing the study design or methodology, (2) were statements, (3) reported single case studies, or (4) were conference papers.

Inclusion and Exclusion Criterion 2: Reported Outcomes

The articles were included if they (1) reported quantitative measures to assess postdeployment effects of DHSs and (2) included outcomes beyond health-related outcomes. The articles were excluded if they (1) focused on qualitative methods, (2) focused on intervention development and implementation, (3) focused on health-related outcomes, or (4) focused on describing the current status and trends of the health care sector in this area.

Inclusion and Exclusion Criterion 3: Type of Use Cases

Articles were included if they described the postdeployment outcome of a DHS that (1) belonged to 1 of the 4 selected digital use cases, (2) was implemented in an outpatient setting, and (3) did not require specific hardware. Articles were excluded if (1) the articles did not describe a DHS within the defined scope, (2) the DHS was deployed in an inpatient setting, or (3) the DHS required specific hardware.

Rapid Review Stage 2

Overview

We limited our search to the PubMed database after observing in rapid review stage 1 that it yielded the most relevant results for our research. Our search used a structured string comprising 3 conceptual blocks: *patient satisfaction*, *assessment*, and *primary health care*. The articles were selected according to the criteria described in the following sections.

Inclusion and Exclusion Criterion 1: Type of Article

The same criteria applied in rapid review stage 1 were used here to determine eligible article types.

Inclusion and Exclusion Criterion 2: Reported Outcomes

The articles were included if they contained patient satisfaction measures as at least 1 of the reported findings. The articles were excluded if they did not include a patient satisfaction measure.

Inclusion and Exclusion Criterion 3: Reported Health Care Setting

The articles were included if the study was conducted in a primary health care setting. The articles were excluded if the study was conducted in a different setting.

Inclusion and Exclusion Criterion 4: Reported Instruments

The articles were included if patient satisfaction was assessed using a validated instrument or if they referenced another publication as the source for the instrument. The articles were excluded if the applied instrument was (1) self-constructed, (2) not validated, or (3) did not reference previous scientific work.

HCP Interviews

We conducted qualitative, semistructured interviews with 19 HCPs. These interviews followed the COREQ checklist [33]. More details can be found in Multimedia Appendix 3.

We designed an interview guide to validate findings from rapid review stage 1 and to explore additional HCSQ dimensions relevant to HCPs. More details are provided in Multimedia Appendix 4. The interviews focused on two domains: (1) implementation and current experience with DHSs and (2) evaluation and further exploration of HCSQ dimensions.

Interviewees were recruited through targeted outreach to both contacts within the researchers' network as well as to HCPs with relevant profiles and publicly available contact information. The following profiles were included: (1) working as an HCP [36], (2) working in outpatient care [37], and (3) considering the implementation or already using at least 1 of the selected DHS use cases. We ensured a diverse sample across specialties, roles, demographics, and countries.

The interviews were conducted with each interviewee individually over the phone. Each session was audio recorded and transcribed verbatim. Recruitment concluded after 19 interviews, upon reaching data saturation. The average duration of each interview was 29 (SD 8) minutes.

XSL•FO
RenderX

We performed coding and qualitative analysis using MAXQDA software (Verbi software) [61]. Our content analysis approach combined deductive and inductive methods [62]:

- Deductive coding established codes based on our rapid review stage 1 results to facilitate comparison.
- Inductive theme derivation derived additional themes inductively from the interview content if mentioned by >1 interviewee.

On the basis of these analyses, we quantitatively scored the relevance of each literature-derived dimension.

Ethical Considerations

The study protocol was reviewed and approved by the Ethics Committee of Witten/Herdecke University (S-113/2023). All interviewees received written information outlining the study's objectives, procedures, and data handling practices and provided written informed consent before participation. Participation was entirely voluntary, and no incentives were offered. Audio recordings were deleted following verbatim transcription. During transcription and subsequent analysis, all personally identifiable information was removed. The interview data were factually anonymized, and results were presented only in aggregated form, making individual identification possible only with disproportionate effort. All data were stored on secure, password-protected systems.

Results

Rapid Review Stage 1

Our objective was to identify which instruments are currently used to measure the effects of DHSs. Our query yielded 9043 publications. To focus on the most current publications, we limited our inclusion to articles in English or German and published within the last 5 years, up to 2023. After removing duplicates, 4957 articles remained for screening, which resulted in 98 potentially relevant articles. Of these, 62 articles were accessible and underwent full-text screening, ultimately leading to the inclusion of 40 publications (Figure 2).

Figure 2. Flowchart of literature screening for rapid review stage 1 following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines to identify validated instruments used to assess the postdeployment effects of digital health solutions in outpatient care. Records were retrieved from PubMed, Scopus, and APA PsycINFO on December 20, 2023 (n=9043), filtered by language (English and German) and publication date (2018-2023). A total of 40 studies met the eligibility criteria and were included in the analysis.

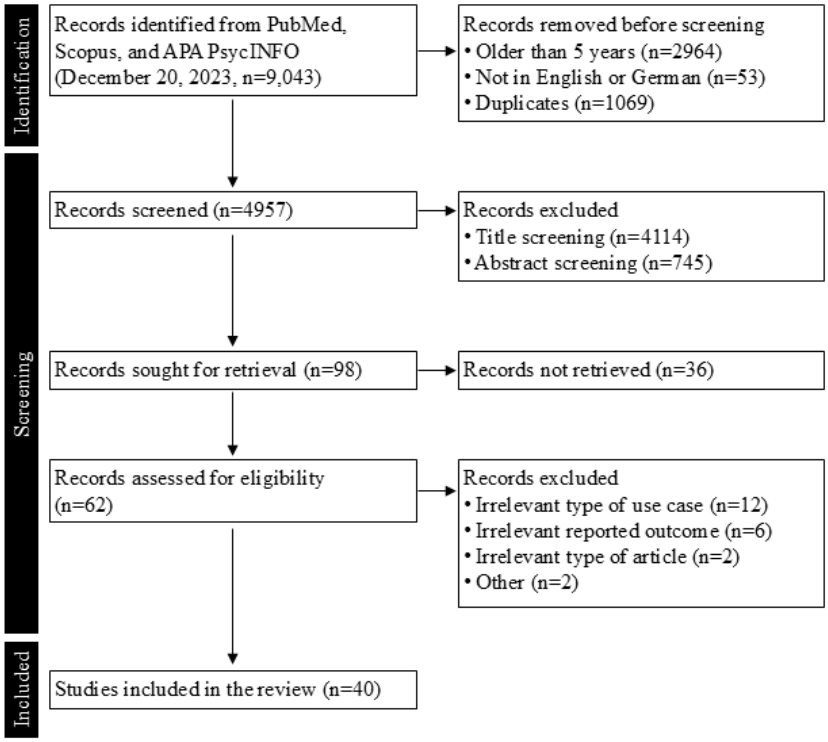

In the 40 eligible studies, we identified a total of 126 reported outcomes. We applied the eligibility criteria used during literature screening to these outcomes analogously, resulting in 85 quantitative, nonhealth-related outcomes applied in outpatient settings measuring the postdeployment effects of selected digital use cases. Many of these were based on self-constructed instruments without reporting validity or reliability characteristics. Only 23 were based on instruments that have either been validated or have been used in previous publications [32,63-76].

In total, we identified 18 distinct instruments (Table 1). We assessed the instruments to ensure adherence to our predefined construction criteria, especially for applicability across digital use cases and perspectives. None of the instruments fulfilled both criteria; instead, 3 instruments assessed both the patient's and HCPs' perspectives but were use case–specific [63,67,69,72], and 9 instruments were not use case–specific but did not assess different perspectives [32,64-68,70,73,74,76].

XSL•FO
RenderX

Table 1. Instruments identified in rapid review stage 1 for evaluating the postdeployment effects of DHSs[a] in outpatient care, selected based on represented perspectives, use case specificity, and occurrence[b].

Instrument	Applied to multiple perspectives	Not specific for a DHS use case	Mentions, n	Identified publications
ACES[c]	No	No	1	[74]
CAHPS[d,e]	No	Yes	2	[64,66]
Deloitte US health care consumer survey	No	Yes	1	[73]
GMC[f] Patient Questionnaire	No	Yes	1	[67]
HCCQ[g]	No	Yes	1	[66]
JSPPPE[h]	No	Yes	1	[64]
MISS-21[i]	No	No	1	[76]
OPEQ[e,j]	No	Yes	2	[32,65]
PEPPI[k]	No	Yes	1	[74]
STOHFLA[l]	No	Yes	1	[66]
Smartphone Use Score	Yes	No	1	[69]
SUS[m]	Yes	No	1	[69]
TAM[n]	No	Yes	1	[68]
TESS[o]	No	No	1	[67]
TUQ[p]	Yes	No	3	[63,67,72]
TMPQ[q]	No	No	1	[66]
TSUQ[r]	No	No	1	[67]
Trust in Physician Scale	No	No	1	[70]

[a]DHS: digital health solution.

[b]A total of 18 instruments were identified and reviewed for adherence to the predefined criteria; compliance with the following criteria was ensured through the previous selection steps: (1) postdeployment evaluation, (2) non–health related, and (3) application in an outpatient setting; open criteria were checked: applicability across multiple perspectives and digital use cases; and the number of mentions is an additional criterion for further selection.

[c]ACES: Ambulatory Care Experience Survey.

[d]CAHPS: Consumer Assessment of Health Care Providers and Systems.

[e]Instruments that were selected for further analysis.

[f]GMC: General Medical Council.

[g]HCCQ: Health Care Climate Questionnaire.

[h]JSPPPE: Jefferson Scale of Patient Perception of Physician Empathy.

[i]MISS-21: Medical Interview Satisfaction Scale.

[j]OPEQ: Outpatient Experience Questionnaire.

[k]PEPPI: Perceived Efficacy in Patient-Physician Interactions Scale.

[l]STOHFLA: Short test of functional health literacy in adults.

[m]SUS: System Usability Score.

[n]TAM: Technology Acceptance Model.

[o]TESS: Telehealth Satisfaction Scale.

[p]TUQ: Telemedicine Satisfaction and Usefulness Questionnaire.

[q]TMPQ: Telemedicine Perception Questionnaire Score.

[r]TSUQ: Telehealth Usability Questionnaire.

For further instrument development, we hypothesized that adjusting an existing instrument from the patient perspective to the HCP perspective would require fewer changes to the original instrument than making a use case–specific instrument more generalizable.

XSL•FO
RenderX

To select an instrument, we focused on those that had been applied multiple times, which is the case for the Consumer Assessment of Health Care Providers and Systems (CAHPS) [64,66] and OPEQ [32,65]. We mapped their dimensions against the HCSQ dimensions outlined by Lee [26]. The comparison revealed that both CAHPS and OPEQ cover dimensions similar to those of HEALTHQUAL (Figure 3 [26]). Notably, OPEQ covers a slightly broader range of HCSQ dimensions compared to HEALTHQUAL. Therefore, we selected OPEQ for further use, merged mapping dimensions with HEALTHQUAL, and added accessibility (merging clinic access and previsit communication) and information as additional HCSQ dimensions.

Figure 3. Mapping of instrument dimensions against health care service quality dimensions. Three instruments (Health Care Service Quality [HEALTHQUAL], Consumer Assessment of Health Care Providers and Systems [CAHPS], and Outpatient Experience Questionnaire [OPEQ]) align with common health care service quality dimensions according to Lee [26]. Shaded cells indicate conceptual overlaps between instrument-specific domains and health care service quality dimensions.

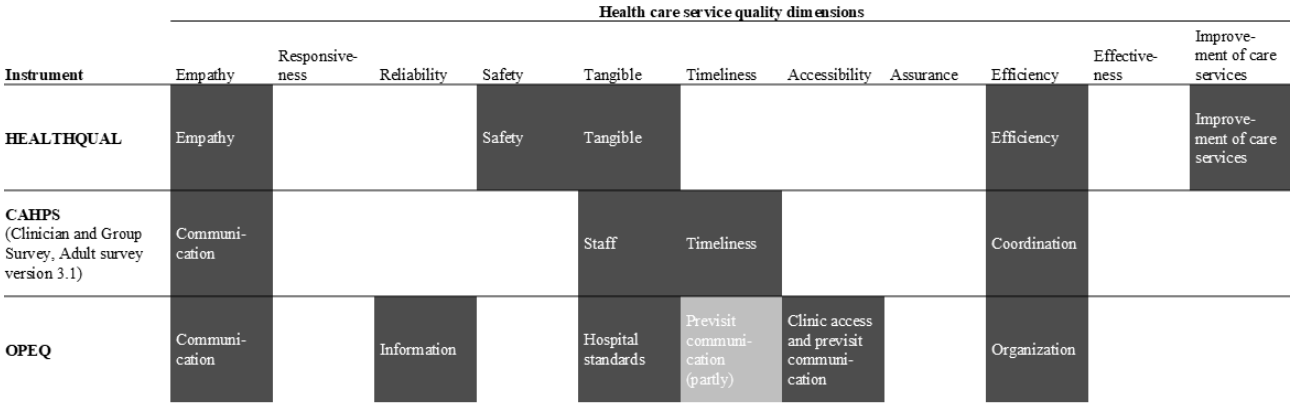

HCP Interviews

Given that HEALTHQUAL and OPEQ have not been previously used with HCPs, we conducted interviews with HCPs with two objectives: (1) to confirm existing and derive new dimensions for the evaluation of HCSQ from the perspective of HCPs and (2) to understand the importance of each dimension for the overall HCSQ.

We interviewed 19 HCPs with diverse backgrounds (Figure 4). We derived a total of 11 HCSQ dimensions from the interviews, with a broad consensus for four of them (mentioned by at least 12/19, 60% of interviewees):

Figure 4. Demographic and professional characteristics of interviewed health care practitioners (n=19). Descriptive statistics of health care practitioners participating in semistructured, qualitative interviews conducted between 2023 and 2025 across Germany, the United States, Switzerland, and Sweden. Variables include gender, age, profession, professional experience, position type, working hours, country, practice type, team size, patient population, size of the location, type of practice, and specialty.

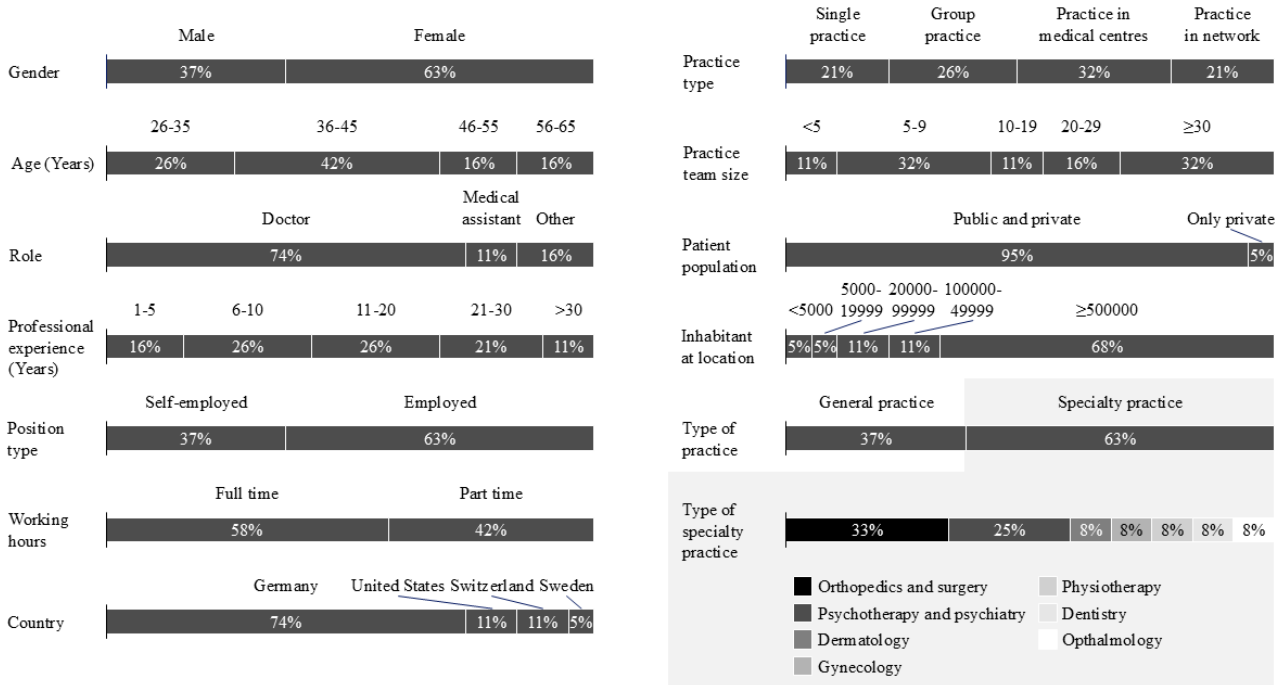

XSL•FO
RenderX

1. Accessibility (17/19, 89%); HCPs agreed that being accessible via phone, email, and at the front desk is crucial.
2. Efficient administration (16/19, 16%); HCPs highlighted the importance of efficient administrative processes to keep patient waiting times low.
3. Availability of appointments (13/19, 68%); HCPs pointed out the significance of having appointments available within an acceptable timeframe.
4. Patient interaction (12/19, 63%); HCPs emphasized the importance of spending enough time with patients and maintaining a close physician-patient relationship.

Other less frequently mentioned dimensions included providing information, the atmosphere in the practice, available medical offerings, and improved health conditions. These dimensions match those in HEALTHQUAL and OPEQ. Notably, 4 (21%) out of 19 interviewees explicitly mentioned patient satisfaction as a separate dimension of HCSQ not represented by the dimensions of HEALTHQUAL and OPEQ. The 2 dimensions not represented by the dimensions of HEALTHQUAL and OPEQ were mentioned but with low frequency. Two interviewees mentioned integrated care and the continuous effort of the practice to improve the service.

Furthermore, interviewees rated the relevance of the HEALTHQUAL and OPEQ dimensions (Figure 5). We found that there was a consensus regarding the most relevant HCSQ dimension (safety) but more variation regarding the remaining dimensions. However, all dimensions received a rating of at least 7 on a scale of 10, so we considered all of them as relevant.

Figure 5. Dimensions of health care service quality (HCSQ) identified and rated by health care practitioners (n=19). Left panel: the frequency with which HCSQ dimensions were mentioned across 19 interviews with outpatient care providers. Right panel: relevance ratings (scale: 1=not at all relevant and 10=extremely relevant) of predefined dimensions from the Health Care Service Quality (HEALTHQUAL) and Outpatient Experience Questionnaire (OPEQ). Bars indicate mean ratings with SD. HCP: health care practitioner.

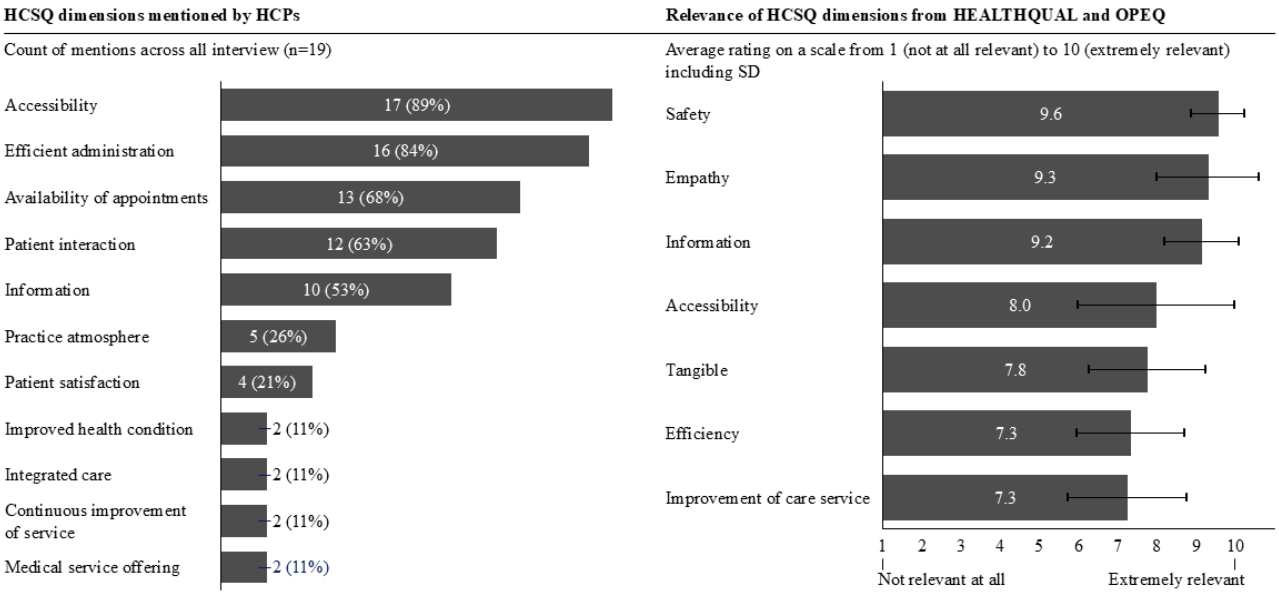

These findings confirmed the relevance of the selected HCSQ dimensions for HCPs. However, patient satisfaction was mentioned by >20% (4/19) of interviewees as an additional relevant dimension. Consequently, we conducted another rapid review to identify instruments measuring patient satisfaction in the literature to be added to our questionnaire.

Rapid Review Stage 2

The objective was to identify the instruments currently used to measure patient satisfaction. We decided to narrow down the scope of this review further to primary health care settings because the 4 mentions of patient satisfaction were made by HCPs working in primary health care. Our initial search resulted in 1573 publications. To focus on recent research, we included only articles written in English or German that were published within the last 5 years, up to 2023, reducing the dataset to 493 articles. Further screening narrowed the selection to 136 potentially relevant articles. Of these, 105 were accessible and subjected to full-text review, leading to the selection of 35 publications (Figure 6).

XSL•FO
RenderX

Figure 6. Flowchart of literature screening for rapid review stage 2 following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines to identify validated patient satisfaction instruments used in primary health care. Records were retrieved from PubMed on November 30, 2023 (n=1573), filtered by language (English and German) and publication date (2018-2023). A total of 35 studies met the eligibility criteria and were included in the analysis.

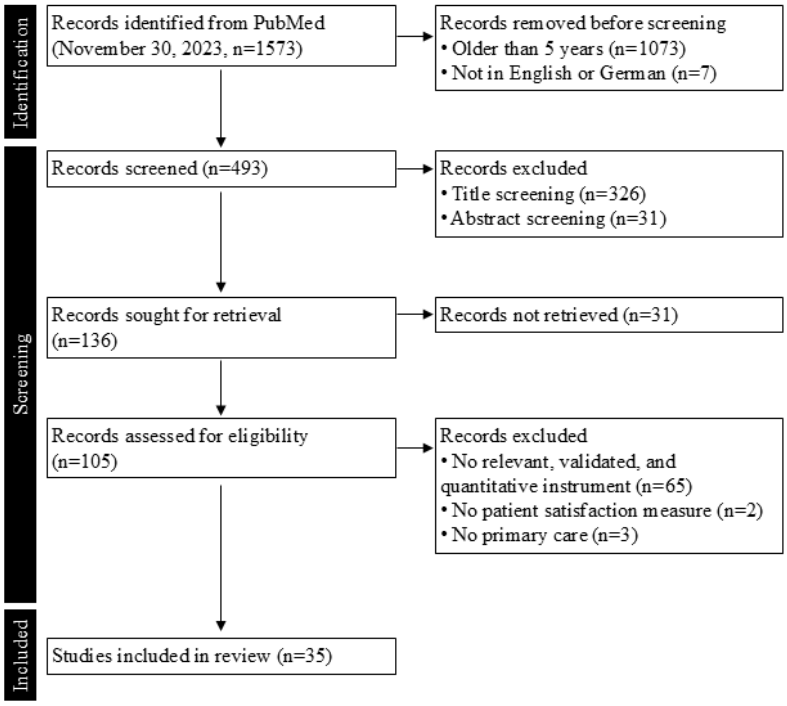

In the 35 eligible studies, 29 different instruments related to patient satisfaction were used or mentioned (Tables 2 and 3). To select the most suitable instrument to add to our questionnaire, we evaluated the following criteria: (1) the frequency of use, (2) the internal consistency (ie, Cronbach α), and (3) the feasibility for integration (ie, number of items).

XSL•FO
RenderX

Table 2. Instruments identified in rapid review stage 2 for measuring patient satisfaction in primary health care settings, selected based on occurrence[a].

Instrument (cluster)	Clustered instruments (if applicable)	Applied, n	Mentioned, n	References
PSQ[b,c]	PSQ II, PSQ III, and PSQ-18[d]	6	2	[77-82]
CAHPS[c,e]	Clinician and group version, emergency department version, hospital version, and patient-centered medical home version	6	0	[83-88]
PCAT[c,f]	—[g]	3	1	[89-92]
EUROPEP[c,h]	—	2	2	[90,93-95]
GPAQ[c,i]	GPAQ and GPAQ-R2[j]	2	1	[89,90,96]
AAFP[k]-MGMA[l] Patient Satisfaction Survey[c]	—	2	0	[97,98]
CQ[m] Index	—	1	0	[99]
Client Satisfaction Questionnaire	—	1	0	[100]
Tool to Improve Quality of Health Care	—	1	0	[101]
EORTC[n] Satisfaction with Care Questionnaire	—	1	0	[102]
General Practitioner Patient Survey	—	1	0	[103]
Medical Interview Satisfaction Scale	—	1	0	[104]
Primary Care Consultation Questionnaire	—	1	0	[105]
Paediatric Otolaryngology Telemedicine Satisfaction Survey	—	1	0	[67]
PROMIS[o]	—	1	0	[86]
Patient Satisfaction Survey Questionnaire	—	1	0	[106]
Patient Survey for Quality of Care Scale	—	1	0	[107]
Satisfaction with Stroke Care Questionnaire	—	1	0	[108]
Survey of Healthcare Experiences of Patients	—	1	0	[87]
Telemedicine Satisfaction Questionnaire	—	1	0	[81]
WHO SARA[p]	—	1	0	[101]
Components of Primary Care Index	—	0	2	[89,90]
GMC[q] Patient Questionnaire	—	0	1	[67]
General Practice Assessment Survey	—	0	1	[90]
General Satisfaction Questionnaire	—	0	1	[108]
SERVQUAL[r]	—	0	1	[79]
TESS[s]	—	0	1	[67]
TSUQ[t]	—	0	1	[67]
TUQ[u]	—	0	1	[67]

[a]A total of 29 instruments were identified in 35 papers; the identified instruments were reviewed in 2 waves; in wave 1, the selection was based on the frequency of use.

[b]PSQ: Patient Satisfaction Questionnaire.

XSL•FO
RenderX

[c]Instruments that were selected for further evaluation.
[d]PSQ-18: Patient Satisfaction Questionnaire-Short Form.
[e]CAHPS: Consumer Assessment of Health Care Providers and Systems.
[f]PCAT: Primary Care Assessment Tool.
[g]Not applicable.
[h]EUROPEP: European Task Force on Patient Evaluations of General Practice Care.
[i]GPAQ: General Practice Assessment Questionnaire.
[j]GPAQ-R2: General Practice Assessment Questionnaire Revised Version 2.
[k]AAFP: American Academy of Family Physicians.
[l]MGMA: Medical Group Management Association.
[m]CQ Index: Consumer Quality Index.
[n]EORTC: European Organization for Research and Treatment of Cancer.
[o]PROMIS: Patient-Reported Outcomes Measurement Information System.
[p]WHO SARA: World Health Organization Service Availability and Readiness Assessment.
[q]GMC: General Medical Council.
[r]SERVQUAL: Service Quality.
[s]TESS: Telehealth Satisfaction Scale.
[t]TSUQ: Telemedicine Satisfaction and Usefulness Questionnaire.
[u]TUQ: Telehealth Usability Questionnaire.

Table 3. Comparison of patient satisfaction instruments selected from results of the rapid review stage 2 based on internal consistency and feasibility of integration in future research (approximated through the number of items)[a].

Instrument (cluster)	Internal consistency (Cronbach α)	Items, n	Additional references
PSQ[b]	0.91[c]	≥18	[34,109]
CAHPS[d]	0.61-0.91[e]	32	[110]
PCAT[f]	0.59-0.94	≥81	[111,112]
EUROPEP[g]	0.81-0.95	23	[113,114]
GPAQ[h]	0.86-0.97[i]	19	[115]
AAFP[j]-MGMA[k] Patient Satisfaction Survey	>0.7	23	[116]

[a]A total of 29 instruments were identified in 35 papers; the identified instruments were reviewed in 2 waves; in wave 2, selection was based on internal consistency and feasibility for integration.
[b]PSQ: Patient Satisfaction Questionnaire.
[c]PSQ-18. Patient Satisfaction Questionnaire-Short Form.
[d]CAHPS: Consumer Assessment of Health Care Providers and Systems.
[e]Clinician group and patient-centered medical home version.
[f]PCAT: Primary Care Assessment Tool.
[g]EUROPEP: European Task Force on Patient Evaluations of General Practice Care.
[h]GPAQ: General Practice Assessment Questionnaire.
[i]GPAQ-R2: General Practice Assessment Questionnaire Revised Version 2.
[j]AAFP: American Academy of Family Physicians.
[k]MGMA: Medical Group Management Association.

A total of 6 instruments (accounting for different variations or versions of an instrument as 1) were used multiple times. All of them demonstrated overall acceptable Cronbach α values. However, CAHPS and Primary Care Assessment Tool were reported to have Cronbach α values <0.7. The Patient Satisfaction Questionnaires and the General Practice Assessment Questionnaire contained the shortest instruments, with 18 and 19 items, respectively. On the basis of these criteria, we selected the most recent version of the Patient Satisfaction Questionnaires, the PSQ-18, for inclusion in our questionnaire based on its widespread use in the literature, reliability, and feasibility.

DigiHEALTHQUAL Development

We developed DigiHEALTHQUAL by synthesizing 3 validated instruments—HEALTHQUAL, OPEQ, and PSQ-18—guided by insights from our rapid reviews and interviews in the previous steps. HEALTHQUAL measures HCSQ via 32 items across 5 dimensions (empathy, efficiency, safety, tangible, and degree of improvements of care services), with responses rated on a 5-point scale ranging from “very bad” to “very good” [26]. The

XSL•FO
RenderX

OPEQ assesses outpatient experiences through 24 items in 6 dimensions (clinic access, communication, hospital standards, organization, previsit communication, and quality of information), using a 10-point scale with anchoring phrases for each end of the scale [31,32]. The PSQ-18 evaluates patient satisfaction via 18 items across 7 domains (accessibility and convenience, communication, financial aspects, general satisfaction, interpersonal manner, technical quality, and time spent with doctors) on a 5-point Likert-like scale ranging from "strongly agree" to "strongly disagree" [34].

The 3 instruments were compared item by item to remove duplicates and ensure contextual applicability by adapting or removing items as needed (Multimedia Appendix 5). Of the initially longlisted 74 items, 13 were adjusted to better fit the focus of our study, 24 were removed because of redundancy or to account for differences between national health care systems, and 1 new item was added because adjusting the existing question to fit the focus of this study was not feasible (ie, data security considerations).

Subsequently, the remaining dimensions of the 3 original instruments were merged based on content overlap where feasible. The resulting DigiHEALTHQUAL questionnaire consists of 51 items across 8 dimensions: accessibility, efficiency, empathy, general satisfaction, degree of improvements of care services, information, safety, and tangibles.

This current questionnaire draft is designed to use a response scale similar to that of the PSQ-18, a 6-point Likert scale ranging from "completely disagree" to "completely agree," allowing calculation of a satisfaction score for each subdimension. This scale has proven effective in practice [12]. Consequently, all items were converted accordingly.

Discussion

Principal Findings

This study reveals key findings regarding the measurement of the effects of DHSs on HCSQ. First, there is a notable lack of consensus on standardized instruments to measure the effects, especially for nonhealth-related outcomes. We hypothesize that this lack of agreement is a reflection of the novelty and wide variety of available DHSs, which consequently makes it challenging for decision makers to compare the effects of different DHSs across diverse contexts.

In contrast, there is greater agreement on the instruments used to measure patient satisfaction, indicating that these tools are well established and widely accepted. This consensus provides a solid foundation for evaluating patient satisfaction in the context of DHSs, offering a reliable starting point for further research.

Finally, instruments typically used to measure HCSQ have not yet been used to assess the impact of DHSs. This gap underscores the need to adapt existing HCSQ instruments for digital use cases. Interestingly, some of the instruments identified in this study include dimensions relevant to HCSQ. This indicates that some level of assessment of HCSQ elements is already taking place, but it is not being explicitly reported as such.

Comparison to Prior Work

Previous research has highlighted a lack of consensus on instruments for measuring implementation research outcomes in health care, including DHS deployments, primarily because of insufficient psychometric properties [17-19]. Our study confirms this general observation and provides additional insights specific to DHS implementation research outcome measurement. We found that the lack of standardized instruments is particularly pronounced for nonhealth-related outcomes in DHS implementation studies. Many of these studies used self-constructed instruments that neither referenced previous work nor reported validity or reliability properties. This variability and lack of rigor make it challenging to compare findings across different studies and contexts. Consequently, our findings underscore the need to validate new or existing instruments for nonhealth-related outcomes to enhance the reliability and comparability of DHS implementation studies and generate a consensus within the research community related to the applicable instruments.

In our study, we identified instruments to measure the effects of DHSs and several instruments to measure HCSQ. However, none of the currently used instruments explicitly assess the intersection of these 2 areas: how HCSQ is influenced by DHS deployment. Recognizing this gap, we followed the recommendations of Boateng et al [117] by choosing to adapt and combine existing instruments rather than create a questionnaire de novo. This approach also aligns with the experimental practice of modifying established instruments such as SERVQUAL to fit specific health care contexts, although it has not yet been widely applied to DHSs [118]. By following this adaptation strategy, we ensure the relevance and applicability of our instruments while maintaining the validity, reliability, and comparability of the results by leveraging well-established and previously validated instruments.

Our HCP interviews revealed the recommendation to add patient satisfaction as an additional dimension to the set of HCSQ dimensions derived from HEALTHQUAL and OPEQ. This recommendation is anecdotally already supported by our rapid review stage 1, for example, where the CAHPS instrument included a question on overall satisfaction. The literature consistently described that there is a correlation between HCSQ and patient satisfaction or happiness in health care settings [119-121]. Various studies identify factors influencing patient satisfaction. For example, Duc Thanh et al [120] identified key factors affecting patient satisfaction, including satisfaction with facilities, service provision, information transparency and administrative procedures, accessibility, and staff interaction and communication. Similarly, Ferreira et al [119] highlighted factors such as waiting times, information provided, cleanliness, communication with patients, and doctors' characteristics. These factors affecting patient satisfaction align closely with the HCSQ dimensions outlined by Lee [26], underscoring the interconnectedness of the metrics patient satisfaction and HCSQ and their collective impact on health care

XSL•FO
RenderX

evaluation—supporting the inclusion of patient satisfaction in DigiHEALTHQUAL.

Strengths and Limitations

By leveraging existing, validated instruments rather than creating new ones, we ensure comparability with previous studies. This approach not only strengthens the validity and reliability of DigiHEALTHQUAL but also facilitates meaningful comparisons and meta-analyses for future research. In addition, our study aimed to add the HCP perspective to patient-focused instruments, providing a holistic view of HCSQ. This comprehensive approach ensures that the nuances and needs of both key stakeholder groups are considered in future research, enhancing the overall applicability and relevance of DigiHEALTHQUAL. Moreover, our approach aligns with the best practices for developing and validating scales for health research, as described by Boateng et al [117], who describe 9 steps across 3 phases for the development of rigorous scales. Following their work, we completed the first phase, "item development," comprising the 2 steps "identification of domain and item generation" and "content validity" in this study [117].

To enable broad application, we excluded dimensions from the original instruments, for example, items related to financial aspects, such as "appropriate costs" or "reasonable medical expenses," because of varying reimbursement across national health care systems. This may result in a lack of insights into selected aspects of HCSQ. Furthermore, the interviews informing our questionnaire design were conducted exclusively with HCPs from high-income countries, which may limit the applicability of the resulting questionnaire to countries with different health care systems. However, the interviews only informed the decision to keep the HCSQ dimensions derived from the validated instruments HEALTHQUAL and OPEQ and to add patient satisfaction as an additional dimension—also derived from a validated instrument used in international studies. Therefore, we estimate that the geographic bias of the interviewees is negligible at this stage of our research. It is important to note that our resulting country-unspecific instrument may have limitations in practice, where national regulatory frameworks influence the adoption of DHSs and data use [15,122].

Generalization issues also arise, as some selected digital use cases may not impact all included HCSQ dimensions. For example, we theorize that online appointment booking platforms are rather unlikely to affect health outcomes. Furthermore, nonprioritized DHSs may require other evaluation criteria. For instance, we hypothesize that the effects of Digital Health Applications will materialize mainly in health outcomes rather than HCSQ and, therefore, necessitate specific, targeted assessment tools. This underscores the need for researchers to balance between using broadly applicable instruments that facilitate comparison across studies and specific instruments that provide detailed insights.

Within the broader context of health care, it is important to note the following: merely assessing the effects of DHSs is insufficient to accelerate their adoption. Comprehensive implementation strategies are necessary to facilitate widespread use.

Moreover, while scaling the digitalization of health care can improve health care delivery, it may also reinforce or exacerbate existing disparities. Health care access and quality are already unevenly distributed, and DHSs risk exacerbating these existing inequalities [123].

On the one hand, digital technologies can help mitigate disparities, for example, by improving access in underserved areas or personalizing care. On the other hand, they can also create new barriers for populations with limited digital literacy, lower socioeconomic status, or restricted access to technology. Studies show that the use of digital health tools correlates with sociodemographic factors such as age; income; education; and, critically, health and digital literacy [124-126].

This introduces 2 challenges. First, in the context of validating DigiHEALTHQUAL, it is important that diverse user groups are appropriately represented. Digitally literate individuals are easier to recruit through online channels, whereas individuals with limited digital access or skills may be systematically underrepresented, leading to bias in the findings. Second, beyond the scope of our work, there is a need for comprehensive planning that ensures equity of access. DHSs must be designed and implemented in a way that safeguards against creating additional barriers and addresses common user concerns such as data privacy, usability, and accessibility.

Practical Implication and Further Research

The implications for further research based on our study are threefold.

Scale Development and Evaluation

Following the best practices for developing and validating scales as outlined by Boateng et al [117], it is necessary to complete the next phases of scale development and evaluation to ensure the validity and reliability of the DigiHEALTHQUAL questionnaire. This involves conducting validation studies with patients to test the resulting questionnaire and subsequently performing validation studies with HCPs to adjust DigiHEALTHQUAL to reflect their perspectives as well. The authors have already initiated this process by conducting preliminary face validity testing with 10 representatives from each target group—patients and HCPs. On the basis of their feedback, minor adjustments were made to improve clarity and contextual relevance. Full-scale quantitative validation studies are currently in preparation, aiming to collect statistically adequate samples for each group. These studies are expected to conclude by the end of 2025 and will include psychometric analyses to support refinement, including item reduction and dimensional validation. To ensure methodological rigor, the planned validation will apply exploratory and confirmatory factor analysis to assess the dimensional structure, internal consistency testing (eg, Cronbach α), and construct validity evaluation. Each quantitative study will aim to include a minimum of 200 participants per target group (patients and HCPs).

Data Collection From DHS Deployment Studies

Collecting data from DHS deployment studies will enable decision makers to retrospectively assess whether the impact

XSL•FO
RenderX

they aimed for has been realized after deployment. This data collection is crucial for understanding the real-world effects of DHSs and refining the evaluation tools based on practical insights.

Meta-Analysis of Postdeployment Studies

Conducting meta-analyses across postdeployment studies using DigiHEALTHQUAL will allow decision makers to prospectively consider the effects of DHSs and make informed decisions before deployment. This is especially important in resource-constrained environments, such as those commonly encountered in health care, where efficient allocation of resources is critical.

To optimize the impact of this research, we suggest initially focusing on use cases with the highest current implementation rates. Our prioritization of 4 use cases can serve as an initial starting point. In addition, there should be a strong emphasis on the outpatient setting, as digital transformation significantly lags behind in outpatient care compared to inpatient care [9,10]. A robust evaluation tool would be particularly valuable in outpatient settings, where clear and actionable guidance is critical for choosing among the many available DHSs that cater to patients with diverse health needs. In these settings, decision makers are often physicians who must juggle multiple roles. We hypothesize that these physicians have limited capacity to make business decisions or compare DHS offerings, given the workforce shortages and growing demand for health care services. Therefore, the decision makers need a robust evaluation framework even more than larger health care organizations with dedicated functions in place.

Conclusions

Our study underscores the importance of developing and validating instruments to measure the effects of DHS deployment on HCSQ in real-life settings. By adapting and combining validated instruments, we provide a robust framework for assessing HCSQ, extending their application to the new and less-explored context of digital health.

Currently, decision makers face significant challenges in comparing outcomes from digital health implementation research because of the lack of consensus on evaluation tools, the use of many different instruments tailored to specific digital use cases, and the variance in the quality of the applied instruments. Although existing tools for HCSQ have not yet been used to assess the effects of DHSs, instruments for digital tools often evaluate relevant HCSQ factors but lack broad applicability.

We propose a comprehensive set of dimensions to measure HCSQ based on the validated instruments HEALTHQUAL, OPEQ, and PSQ-18. This set can be used by decision makers and practitioners to assess ongoing or completed DHS deployments. However, future research is essential to validate the reliability and validity of this scale further and to collect a robust database that can support pre-deployment decision-making in the future.

In summary, DigiHEALTHQUAL aims to fill the gap in current digital health evaluation practices, providing a standardized approach that enhances the comparability of research outcomes and ultimately supports better decision-making in the implementation of DHSs.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

DR, LF, and SM developed the overarching research question and study design. DR performed the literature reviews; conducted, transcribed, and qualitatively analyzed the interviews; and developed the DigiHEALTHQUAL questionnaire. DR drafted the manuscript. SM, LF, and AM provided critical review. SM coordinated the project. All authors reviewed the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA-ScR checklist—rapid review stage 1.
[PDF File (Adobe PDF File), 223 KB - formative_v9i1e68276_app1.pdf]

Multimedia Appendix 2
PRISMA-ScR checklist—rapid review stage 2.
[PDF File (Adobe PDF File), 202 KB - formative_v9i1e68276_app2.pdf]

Multimedia Appendix 3
COREQ checklist—health care practitioner interviews.
[PDF File (Adobe PDF File), 206 KB - formative_v9i1e68276_app3.pdf]

Multimedia Appendix 4

XSL•FO
RenderX

Translated semistructured interview guide.
[PDF File (Adobe PDF File), 198 KB - formative_v9i1e68276_app4.pdf]

Multimedia Appendix 5
Development of the Digital Healthcare Service Quality Questionnaire.
[PDF File (Adobe PDF File), 211 KB - formative_v9i1e68276_app5.pdf]

References

1. Black AD, Car J, Pagliari C, Anandan C, Cresswell K, Bokun T, et al. The impact of eHealth on the quality and safety of health care: a systematic overview. PLoS Med 2011 Jan 18;8(1):e1000387 [FREE Full text] [doi: 10.1371/journal.pmed.1000387] [Medline: 21267058]
2. Eden R, Burton-Jones A, Scott I, Staib A, Sullivan C. Effects of eHealth on hospital practice: synthesis of the current literature. Aust Health Rev 2018 Sep;42(5):568-578. [doi: 10.1071/AH17255] [Medline: 29986809]
3. eHealth: digital health and care. Directorate-General for Health and Food Safety. URL: https://health.ec.europa.eu/ehealth-digital-health-and-care/overview_en [accessed 2024-05-24]
4. Eze ND, Mateus C, Cravo Oliveira Hashiguchi T. Telemedicine in the OECD: an umbrella review of clinical and cost-effectiveness, patient experience and implementation. PLoS One 2020;15(8):e0237585 [FREE Full text] [doi: 10.1371/journal.pone.0237585] [Medline: 32790752]
5. Keasberry J, Scott IA, Sullivan C, Staib A, Ashby R. Going digital: a narrative overview of the clinical and organisational impacts of eHealth technologies in hospital practice. Aust Health Rev 2017 Dec;41(6):646-664. [doi: 10.1071/AH16233] [Medline: 28063462]
6. Brönneke JB, Debatin JF. [Digitalization of healthcare and its effects on quality of care]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2022 Mar;65(3):342-347 [FREE Full text] [doi: 10.1007/s00103-022-03493-3] [Medline: 35181795]
7. Brands MR, Gouw SC, Beestrum M, Cronin RM, Fijnvandraat K, Badawy SM. Patient-centered digital health records and their effects on health outcomes: systematic review. J Med Internet Res 2022 Dec 22;24(12):e43086 [FREE Full text] [doi: 10.2196/43086] [Medline: 36548034]
8. Iyanna S, Kaur P, Ractham P, Talwar S, Najmul Islam A. Digital transformation of healthcare sector. What is impeding adoption and continued usage of technology-driven innovations by end-users? J Bus Res 2022 Dec;153:150-161. [doi: 10.1016/j.jbusres.2022.08.007]
9. Rahurkar S, Vest JR, Finnell JT, Dixon BE. Trends in user-initiated health information exchange in the inpatient, outpatient, and emergency settings. J Am Med Inform Assoc 2021 Mar 01;28(3):622-627 [FREE Full text] [doi: 10.1093/jamia/ocaa226] [Medline: 33067617]
10. Neunaber T, Meister S. Digital maturity and its measurement of general practitioners: a scoping review. Int J Environ Res Public Health 2023 Feb 28;20(5):e309 [FREE Full text] [doi: 10.3390/ijerph20054377] [Medline: 36901387]
11. Scott Kruse C, Karem P, Shifflett K, Vegi L, Ravi K, Brooks M. Evaluating barriers to adopting telemedicine worldwide: a systematic review. J Telemed Telecare 2018 Jan;24(1):4-12 [FREE Full text] [doi: 10.1177/1357633X16674087] [Medline: 29320966]
12. Weik L, Fehring L, Mortsiefer A, Meister S. Understanding inherent influencing factors to digital health adoption in general practices through a mixed-methods analysis. NPJ Digit Med 2024 Feb 27;7(1):47 [FREE Full text] [doi: 10.1038/s41746-024-01049-0] [Medline: 38413767]
13. Brantnell A, Temiz S, Baraldi E, Woodford J, von Essen L. Barriers to and facilitators of the implementation of digital mental health interventions as perceived by primary care decision makers: content analysis of structured open-ended survey data. JMIR Hum Factors 2023 Jun 26;10:e44688 [FREE Full text] [doi: 10.2196/44688] [Medline: 37358902]
14. Ross J, Stevenson F, Dack C, Pal K, May C, Michie S, et al. Developing an implementation strategy for a digital health intervention: an example in routine healthcare. BMC Health Serv Res 2018 Oct 19;18(1):794 [FREE Full text] [doi: 10.1186/s12913-018-3615-7] [Medline: 30340639]
15. Essén A, Stern AD, Haase CB, Car J, Greaves F, Paparova D, et al. Health app policy: international comparison of nine countries' approaches. NPJ Digit Med 2022 Mar 18;5(1):31 [FREE Full text] [doi: 10.1038/s41746-022-00573-1] [Medline: 35304561]
16. Mathews SC, McShea MJ, Hanley CL, Ravitz A, Labrique AB, Cohen AB. Digital health: a path to validation. NPJ Digit Med 2019;2:38 [FREE Full text] [doi: 10.1038/s41746-019-0111-3] [Medline: 31304384]
17. Khadjesari Z, Boufkhed S, Vitoratou S, Schatte L, Ziemann A, Daskalopoulou C, et al. Implementation outcome instruments for use in physical healthcare settings: a systematic review. Implement Sci 2020 Aug 18;15(1):66 [FREE Full text] [doi: 10.1186/s13012-020-01027-6] [Medline: 32811517]
18. Godinho MA, Ansari S, Guo GN, Liaw ST. Toolkits for implementing and evaluating digital health: a systematic review of rigor and reporting. J Am Med Inform Assoc 2021 Jun 12;28(6):1298-1307 [FREE Full text] [doi: 10.1093/jamia/ocab010] [Medline: 33619519]

XSL•FO
RenderX

19. Clinton-McHarg T, Yoong SL, Tzelepis F, Regan T, Fielding A, Skelton E, et al. Psychometric properties of implementation measures for public health and community settings and mapping of constructs against the Consolidated Framework for Implementation Research: a systematic review. Implement Sci 2016 Nov 08;11(1):148 [FREE Full text] [doi: 10.1186/s13012-016-0512-5] [Medline: 27821146]
20. Enam A, Torres-Bonilla J, Eriksson H. Evidence-based evaluation of eHealth interventions: systematic literature review. J Med Internet Res 2018 Nov 23;20(11):e10971 [FREE Full text] [doi: 10.2196/10971] [Medline: 30470678]
21. Sakamoto JL, Carandang RR, Kharel M, Shibanuma A, Yarotskaya E, Basargina M, et al. Effects of mHealth on the psychosocial health of pregnant women and mothers: a systematic review. BMJ Open 2022 Feb 15;12(2):e056807 [FREE Full text] [doi: 10.1136/bmjopen-2021-056807] [Medline: 35168981]
22. Wang X, Hunter DJ, Vesentini G, Pozzobon D, Ferreira ML. Technology-assisted rehabilitation following total knee or hip replacement for people with osteoarthritis: a systematic review and meta-analysis. BMC Musculoskelet Disord 2019 Nov 03;20(1):506 [FREE Full text] [doi: 10.1186/s12891-019-2900-x] [Medline: 31679511]
23. Messner EM, Terhorst Y, Barke A, Baumeister H, Stoyanov S, Hides L, et al. The German version of the Mobile App Rating Scale (MARS-G): development and validation study. JMIR Mhealth Uhealth 2020 Mar 27;8(3):e14479 [FREE Full text] [doi: 10.2196/14479] [Medline: 32217504]
24. Stach M, Kraft R, Probst T, Messner EM, Terhorst Y, Baumeister H, et al. Mobile health app database - a repository for quality ratings of mHealth apps. In: Proceedings of the 33rd International Symposium on Computer-Based Medical Systems. 2020 Presented at: CBMS '20; July 28-30, 2020; Rochester, MN p. 427-432 URL: https://ieeexplore.ieee.org/document/9183243 [doi: 10.1109/cbms49503.2020.00087]
25. Endeshaw B. Healthcare service quality-measurement models: a review. J Health Res 2020 Apr 07;35(2):106-117. [doi: 10.1108/jhr-07-2019-0152]
26. Lee D. HEALTHQUAL: a multi-item scale for assessing healthcare service quality. Serv Bus 2016 Jun 13;11(3):491-516. [doi: 10.1007/s11628-016-0317-2]
27. Parasuraman A, Zeithaml V, Berry L. SERVQUAL: a multiple- item scale for measuring consumer perceptions of service quality. J Retail 2024:1988 [FREE Full text]
28. Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ 2021 Mar 29;372:n160 [FREE Full text] [doi: 10.1136/bmj.n160] [Medline: 33781993]
29. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
30. King VJ, Stevens A, Nussbaumer-Streit B, Kamel C, Garritty C. Paper 2: performing rapid reviews. Syst Rev 2022 Jul 30;11(1):151 [FREE Full text] [doi: 10.1186/s13643-022-02011-5] [Medline: 35906677]
31. Garratt AM, Bjaertnes Ø, Krogstad U, Gulbrandsen P. The OutPatient Experiences Questionnaire (OPEQ): data quality, reliability, and validity in patients attending 52 Norwegian hospitals. Qual Saf Health Care 2005 Dec;14(6):433-437 [FREE Full text] [doi: 10.1136/qshc.2005.014423] [Medline: 16326790]
32. Augestad KM, Sneve AM, Lindsetmo RO. Telemedicine in postoperative follow-up of STOMa PAtients: a randomized clinical trial (the STOMPA trial). Br J Surg 2020 Apr;107(5):509-518. [doi: 10.1002/bjs.11491] [Medline: 32100297]
33. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
34. Marshall GN, Hays RD. The patient satisfaction questionnaire short-form (PSQ-18). RAND Corporation. 1994. URL: https://www.rand.org/content/dam/rand/pubs/papers/2006/P7865.pdf [accessed 2025-05-29]
35. Proctor E, Silmere H, Raghavan R, Hovmand P, Aarons G, Bunger A, et al. Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda. Adm Policy Ment Health 2011 Mar;38(2):65-76 [FREE Full text] [doi: 10.1007/s10488-010-0319-7] [Medline: 20957426]
36. Classifying health workers: mapping occupations to the international standard classification. World Health Organization. 2019. URL: https://cdn.who.int/media/docs/default-source/health-workforce/dek/classifying-health-workers.pdf [accessed 2024-05-24]
37. NCI dictionary of cancer terms. National Cancer Institute. URL: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/outpatient [accessed 2024-05-24]
38. Souca L. SERVQUAL- thirty years of research on service quality with implications for customer satisfaction. In: Proceedings of the 4th Edition of the International Conference on Marketing - From Information to Decision. 2011 Presented at: MID '11; October 28-29, 2011; Cluj-Napoca, Romania p. 420-429 URL: https://www.researchgate.net/publication/346470601_SERVQUAL_-Thirty_years_of_research_on_service_quality_with_implications_for_customer_satisfaction
39. Darzi MA, Islam SB, Khursheed SO, Bhat SA. Service quality in the healthcare sector: a systematic review and meta-analysis. LBS J Manag Res 2023 Jan 16;21(1):13-29. [doi: 10.1108/lbsjmr-06-2022-0025]
40. Miranda FJ, Chamorro A, Murillo LR, Vega J. Assessing primary healthcare services quality in Spain: managers vs. patients perceptions. Serv Ind J 2010 Nov;30(13):2137-2149. [doi: 10.1080/02642060903215055]

XSL•FO
RenderX

41. Mosadeghrad A, Sokhanvar M. Measuring quality of services in Tehran teaching hospitals using HEALTHQUAL instrument. Razi J Med Sci 2018;25:1 [FREE Full text]
42. Lee D, Kim KK. Assessing healthcare service quality: a comparative study of patient treatment types. Int J Qual Innov 2017 Jan 25;3(1):93. [doi: 10.1186/S40887-016-0010-5]
43. Barrios-Ipenza F, Calvo-Mora A, Velicia-Martín F, Criado-García F, Leal-Millán A. Patient satisfaction in the Peruvian Health Services: validation and application of the HEALTHQUAL scale. Int J Environ Res Public Health 2020 Jul 15;17(14):943 [FREE Full text] [doi: 10.3390/ijerph17145111] [Medline: 32679844]
44. Asiamah N, Frimpong Opuni F, Aggrey M, Adu-Gyamfi K. ADAPTED SERVQUAL: a health service quality scale incorporating indicators of sanitation and hygiene. Qual Manag Health Care 2021;30(3):184-193. [doi: 10.1097/QMH.0000000000000269] [Medline: 33306656]
45. Allahham L, Mouselli S, Jakovljevic M. The quality of Syrian healthcare services during COVID-19: a HEALTHQUAL approach. Front Public Health 2022;10:970922 [FREE Full text] [doi: 10.3389/fpubh.2022.970922] [Medline: 36033782]
46. Chauhan VA, Limbad SJ. Service quality assessment of government and private hospitals using extended HEALTHQUAL dimensions in Surat region. Int J Curr Res 2023;11(8):889-898 [FREE Full text]
47. Shaikh MI, Sarkar A. A comparison of patients' satisfaction treated at public and private healthcare institutions under the health scheme using the HEALTHQUAL model. J Integr Med Public Health 2024 Jan 17;2(2):75-81. [doi: 10.4103/jimph.jimph_22_23]
48. ASReview LAB - a tool for AI-assisted systematic reviews. zenodo. URL: https://zenodo.org/records/15466337 [accessed 2025-05-29]
49. Campos DG, Fütterer T, Gfrörer T, Lavelle-Hill R, Murayama K, König L, et al. Screening smarter, not harder: a comparative analysis of machine learning screening algorithms and heuristic stopping criteria for systematic reviews in educational research. Educ Psychol Rev 2024 Feb 08;36(1):e369. [doi: 10.1007/S10648-024-09862-5]
50. König L, Zitzmann S, Fütterer T, Campos D, Scherer R, Hecht M. An evaluation of the performance of stopping rules in AI-aided screening for psychological meta-analytical research. Res Synth Methods 2024 Nov;15(6):1120-1146. [doi: 10.1002/jrsm.1762] [Medline: 39412090]
51. Abdolkhani R, Petersen S, Walter R, Zhao L, Butler-Henderson K, Livesay K. The impact of digital health transformation driven by COVID-19 on nursing practice: systematic literature review. JMIR Nurs 2022 Aug 30;5(1):e40348. [doi: 10.2196/40348] [Medline: 35867838]
52. Allen P, Pilar M, Walsh-Bailey C, Hooley C, Mazzucca S, Lewis CC, et al. Quantitative measures of health policy implementation determinants and outcomes: a systematic review. Implement Sci 2020 Jun 19;15(1):47 [FREE Full text] [doi: 10.1186/s13012-020-01007-w] [Medline: 32560661]
53. Chatterjee A, Prinz A, Gerdes M, Martinez S. Digital interventions on healthy lifestyle management: systematic review. J Med Internet Res 2021 Nov 17;23(11):e26931 [FREE Full text] [doi: 10.2196/26931] [Medline: 34787575]
54. Granja C, Janssen W, Johansen MA. Factors determining the success and failure of eHealth interventions: systematic review of the literature. J Med Internet Res 2018 May 01;20(5):e10235 [FREE Full text] [doi: 10.2196/10235] [Medline: 29716883]
55. Krick T, Huter K, Domhoff D, Schmidt A, Rothgang H, Wolf-Ostermann K. Digital technology and nursing care: a scoping review on acceptance, effectiveness and efficiency studies of informal and formal care technologies. BMC Health Serv Res 2019 Jun 20;19(1):400 [FREE Full text] [doi: 10.1186/s12913-019-4238-3] [Medline: 31221133]
56. Li E, Clarke J, Neves AL, Ashrafian H, Darzi A. Electronic health records, interoperability and patient safety in health systems of high-income countries: a systematic review protocol. BMJ Open 2021 Jul 14;11(7):e044941 [FREE Full text] [doi: 10.1136/bmjopen-2020-044941] [Medline: 34261679]
57. Odendaal WA, Anstey Watkins J, Leon N, Goudge J, Griffiths F, Tomlinson M, et al. Health workers' perceptions and experiences of using mHealth technologies to deliver primary healthcare services: a qualitative evidence synthesis. Cochrane Database Syst Rev 2020 Mar 26;3(3):CD011942 [FREE Full text] [doi: 10.1002/14651858.CD011942.pub2] [Medline: 32216074]
58. Peters GM, Kooij L, Lenferink A, van Harten WH, Doggen CJ. The effect of telehealth on hospital services use: systematic review and meta-analysis. J Med Internet Res 2021 Sep 01;23(9):e25195 [FREE Full text] [doi: 10.2196/25195] [Medline: 34468324]
59. Ross J, Stevenson F, Lau R, Murray E. Factors that influence the implementation of e-health: a systematic review of systematic reviews (an update). Implement Sci 2016 Oct 26;11(1):146 [FREE Full text] [doi: 10.1186/s13012-016-0510-7] [Medline: 27782832]
60. Rouleau G, Gagnon MP, Côté J, Payne-Gagnon J, Hudson E, Dubois C. Impact of information and communication technologies on nursing care: results of an overview of systematic reviews. J Med Internet Res 2017 Apr 25;19(4):e122 [FREE Full text] [doi: 10.2196/jmir.6686] [Medline: 28442454]
61. MAXQDA 2020. VERBI Software. Berlin; 2020. URL: https://www.maxqda.com/blogpost/how-to-cite-maxqda
62. Kuckartz U, Rädiker S. Qualitative Inhaltsanalyse. Methoden, Praxis, Computerunterstützung (Grundlagentexte Methoden). Weinheim, Germany: Beltz Juventa; 2022.

XSL•FO
RenderX

63. Bennell KL, Lawford BJ, Metcalf B, Mackenzie D, Russell T, van den Berg M, et al. Physiotherapists and patients report positive experiences overall with telehealth during the COVID-19 pandemic: a mixed-methods study. J Physiother 2021 Jul;67(3):201-209 [FREE Full text] [doi: 10.1016/j.jphys.2021.06.009] [Medline: 34147399]
64. Bruno B, Mercer MB, Hizlan S, Peskin J, Ford PJ, Farrell RM, et al. Virtual prenatal visits associated with high measures of patient experience and satisfaction among average-risk patients: a prospective cohort study. BMC Pregnancy Childbirth 2023 Apr 06;23(1):234. [doi: 10.1186/s12884-023-05421-y] [Medline: 37024808]
65. Buvik A, Bugge E, Knutsen G, Småbrekke A, Wilsgaard T. Patient reported outcomes with remote orthopaedic consultations by telemedicine: a randomised controlled trial. J Telemed Telecare 2019 Sep;25(8):451-459. [doi: 10.1177/1357633X18783921] [Medline: 29973130]
66. Danila MI, Sun D, Jackson LE, Cutter G, Jackson EA, Ford EW, et al. Satisfaction with modes of telemedicine delivery during COVID-19: a randomized, single-blind, parallel group, noninferiority trial. Am J Med Sci 2022 Nov;364(5):538-546 [FREE Full text] [doi: 10.1016/j.amjms.2022.06.021] [Medline: 35793732]
67. Darr A, Senior A, Argyriou K, Limbrick J, Nie H, Kantczak A, et al. The impact of the coronavirus (COVID-19) pandemic on elective paediatric otolaryngology outpatient services - An analysis of virtual outpatient clinics in a tertiary referral centre using the modified paediatric otolaryngology telemedicine satisfaction survey (POTSS). Int J Pediatr Otorhinolaryngol 2020 Nov;138:110383 [FREE Full text] [doi: 10.1016/j.ijporl.2020.110383] [Medline: 33152974]
68. Horsham C, Snoswell C, Vagenas D, Loescher LJ, Gillespie N, Soyer HP, et al. Is teledermoscopy ready to replace face-to-face examinations for the early detection of skin cancer? Consumer views, technology acceptance, and satisfaction with care. Dermatology 2020;236(2):90-96. [doi: 10.1159/000506154] [Medline: 32114570]
69. Kim SM, Kim T, Cha WC, Lee JH, Kwon IH, Choi Y, et al. User experience of mobile personal health records for the emergency department: mixed methods study. JMIR Mhealth Uhealth 2020 Dec 15;8(12):e24326 [FREE Full text] [doi: 10.2196/24326] [Medline: 33320102]
70. Orrange S, Patel A, Mack WJ, Cassetta J. Patient satisfaction and trust in telemedicine during the COVID-19 pandemic: retrospective observational study. JMIR Hum Factors 2021 Apr 22;8(2):e28589. [doi: 10.2196/28589] [Medline: 33822736]
71. Rickwood D, Wallace A, Kennedy V, O'Sullivan S, Telford N, Leicester S. Young people's satisfaction with the online mental health service eheadspace: development and implementation of a service satisfaction measure. JMIR Ment Health 2019 Apr 17;6(4):e12169 [FREE Full text] [doi: 10.2196/12169] [Medline: 30994470]
72. Tan NG, Yang LW, Tan MZ, Chng J, Tan MH, Tan C. Virtual care to increase military medical centre capacity in the primary health care setting: a prospective self-controlled pilot study of symptoms collection and telemedicine. J Telemed Telecare 2022 Sep;28(8):603-612. [doi: 10.1177/1357633X20959579] [Medline: 33016187]
73. Thomas E, Lee CM, Norman R, Wells L, Shaw T, Nesbitt J, et al. Patient use, experience, and satisfaction with telehealth in an Australian population (reimagining health care): web-based survey study. J Med Internet Res 2023 Aug 17;25:e45016 [FREE Full text] [doi: 10.2196/45016] [Medline: 37590037]
74. Walker J, Leveille S, Bell S, Chimowitz H, Dong Z, Elmore JG, et al. OpenNotes after 7 years: patient experiences with ongoing access to their clinicians' outpatient visit notes. J Med Internet Res 2019 May 06;21(5):e13876 [FREE Full text] [doi: 10.2196/13876] [Medline: 31066717]
75. Watson L, Qi S, Delure A, Link C, Photitai E, Chmielewski L, et al. Virtual cancer care during the COVID-19 pandemic in Alberta: evidence from a mixed methods evaluation and key learnings. JCO Oncol Pract 2021 Sep;17(9):e1354-e1361 [FREE Full text] [doi: 10.1200/OP.21.00144] [Medline: 34351822]
76. Zammit M, Siau R, Williams C, Hussein A. Patient satisfaction from ENT telephone consultations during the coronavirus disease 2019 pandemic. J Laryngol Otol (Forthcoming) 2020 Nov 17:1-6 [FREE Full text] [doi: 10.1017/S0022215120002480] [Medline: 33198822]
77. Alsayali MM, Al-Sahafi A, Mandoura N, Usman Shah HB, Abdul Rashid OA, AlSharif K, et al. Patients' satisfaction after primary health care centers' integration with Ministry of Health Hospitals, Jeddah. J Epidemiol Glob Health 2019 Jun;9(2):135-142 [FREE Full text] [doi: 10.2991/jegh.k.190522.001] [Medline: 31241872]
78. Hayek S, Derhy S, Smith ML, Towne SD, Zelber-Sagi S. Patient satisfaction with primary care physician performance in a multicultural population. Isr J Health Policy Res 2020 Mar 25;9(1):13 [FREE Full text] [doi: 10.1186/s13584-020-00372-7] [Medline: 32213194]
79. Leow HT, Liew SM. A cross sectional study on patient satisfaction and its association with length of consultation at the University Malaya Medical Centre Primary Care Clinic. Malays Fam Physician 2022 Jul 24;17(2):71-80 [FREE Full text] [doi: 10.51866/oa1339] [Medline: 35950009]
80. Li H, Swe KM, Jabbar MA, Ching SM. Patients' satisfaction on primary healthcare services in China and its associated factors during COVID-19 pandemic: a cross-sectional study. Malays Fam Physician 2023;18:4 [FREE Full text] [doi: 10.51866/oa.194] [Medline: 36992956]
81. Mauro E, Marciano S, Torres MC, Roca JD, Novillo AL, Gadano A. Telemedicine improves access to hepatology consultation with high patient satisfaction. J Clin Exp Hepatol 2020;10(6):555-562 [FREE Full text] [doi: 10.1016/j.jceh.2020.04.017] [Medline: 33311892]

XSL•FO
RenderX

82. Staub MB, Pellegrino R, Gettler E, Johnson MC, Roumie CL, Grijalva CG, et al. Association of antibiotics with veteran visit satisfaction and antibiotic expectations for upper respiratory tract infections. Antimicrob Steward Healthc Epidemiol 2022;2(1):e100 [FREE Full text] [doi: 10.1017/ash.2022.233] [Medline: 36483414]
83. Abass G, Asery A, Al Badr A, AlMaghlouth A, AlOtaiby S, Heena H. Patient satisfaction with the emergency department services at an academic teaching hospital. J Family Med Prim Care 2021 Apr;10(4):1718-1725 [FREE Full text] [doi: 10.4103/jfmpc.jfmpc_8_20] [Medline: 34123918]
84. Alkahtani AS, Abbas AH, Bin Rsheed AM, Alabood AF, Alqahtani AA. Assessing the impact of a service excellence program on improving patient experience at primary health care centers. Cureus 2023 Aug;15(8):e43749 [FREE Full text] [doi: 10.7759/cureus.43749] [Medline: 37727190]
85. Hung DY, Mujal G, Jin A, Liang SY. Patient experiences after implementing lean primary care redesigns. Health Serv Res 2021 Jun;56(3):363-370 [FREE Full text] [doi: 10.1111/1475-6773.13605] [Medline: 33305379]
86. Jordan JN, Wadsworth TG, Robinson R, Hruza H, Paul A, O'Connor SK. Patient satisfaction with pharmacist-provided health-related services in a primary care clinic. Pharmacy (Basel) 2021 Nov 21;9(4):187 [FREE Full text] [doi: 10.3390/pharmacy9040187] [Medline: 34842798]
87. Legler A, Price M, Parikh M, Nebeker JR, Ward MC, Wedemeyer L, et al. Effect on VA patient satisfaction of provider's use of an integrated viewer of multiple electronic health records. J Gen Intern Med 2019 Jan;34(1):132-136 [FREE Full text] [doi: 10.1007/s11606-018-4708-z] [Medline: 30338474]
88. Zakare-Fagbamila RT, Park C, Dickson W, Cheng TZ, Gottfried ON. The true penalty of the waiting room: the role of wait time in patient satisfaction in a busy spine practice. J Neurosurg Spine 2020 Jul 01;33(1):95-105. [doi: 10.3171/2019.12.SPINE191257] [Medline: 32084633]
89. Albahrani S, Albidy H, Alomar N, Bin Mutreb L, Alkhofi A, Alsaleh Z, et al. Patient satisfaction with primary healthcare services in Al-Ahsa, Saudi Arabia. Cureus 2022 Nov;14(11):e31478 [FREE Full text] [doi: 10.7759/cureus.31478] [Medline: 36408313]
90. Kijima T, Matsushita A, Akai K, Hamano T, Takahashi S, Fujiwara K, et al. Patient satisfaction and loyalty in Japanese primary care: a cross-sectional study. BMC Health Serv Res 2021 Mar 25;21(1):274 [FREE Full text] [doi: 10.1186/s12913-021-06276-9] [Medline: 33766027]
91. Pitrou I, Berbiche D, Vasiliadis HM. Mental health and satisfaction with primary care services in older adults: a study from the patient perspective on four dimensions of care. Fam Pract 2020 Sep 05;37(4):459-464. [doi: 10.1093/fampra/cmaa019] [Medline: 32201895]
92. Wang W, Haggerty J. Development of primary care assessment tool-adult version in Tibet: implication for low- and middle-income countries. Prim Health Care Res Dev 2019 Jul 01;20:e94 [FREE Full text] [doi: 10.1017/S1463423619000239] [Medline: 32800017]
93. Chen W, Feng Y, Fang J, Wu J, Huang X, Wang X, et al. Effect of trust in primary care physicians on patient satisfaction: a cross-sectional study among patients with hypertension in rural China. BMC Fam Pract 2020 Sep 21;21(1):196 [FREE Full text] [doi: 10.1186/s12875-020-01268-w] [Medline: 32957936]
94. Ferreira PL, Raposo V, Tavares AI. Primary health care patient satisfaction: explanatory factors and geographic characteristics. Int J Qual Health Care 2020 Apr 27;32(2):93-98. [doi: 10.1093/intqhc/mzz134] [Medline: 32047931]
95. Sparkes SP, Atun R, B rnighausen T. The impact of the family medicine model on patient satisfaction in Turkey: panel analysis with province fixed effects. PLoS One 2019;14(1):e0210563 [FREE Full text] [doi: 10.1371/journal.pone.0210563] [Medline: 30699131]
96. Mohamoud G, Mash R. Evaluation of the quality of service delivery in private sector, primary care clinics in Kenya: a descriptive patient survey. S Afr Fam Pract (2004) 2020 Oct 22;62(1):e1-12 [FREE Full text] [doi: 10.4102/safp.v62i1.5148] [Medline: 33179953]
97. Alshowair A, Altamimi S, Alruhaimi F, Tolba A, Almeshari A, Almubrick R, et al. Assessment of primary health care specialized reference clinics in Riyadh first health cluster: outcome, cost-effectiveness and patient satisfaction. Clinicoecon Outcomes Res 2022;14:371-381 [FREE Full text] [doi: 10.2147/CEOR.S355507] [Medline: 35547100]
98. Nowaskie DZ, Fogel RS, Fogel JM. Impact on patient satisfaction and importance of medical intake and office staff in a multidisciplinary, one-stop shop transgender program in Indianapolis, Indiana. J Multidiscip Healthc 2019;12:665-673 [FREE Full text] [doi: 10.2147/JMDH.S216969] [Medline: 31616155]
99. Rutten MH, Giesen PH, Assendelft WJ, Westert G, Smits M. Effects of access to radiology in out-of-hours primary care on patient satisfaction and length of stay. Eur J Gen Pract 2021 Dec;27(1):221-227 [FREE Full text] [doi: 10.1080/13814788.2021.1959911] [Medline: 34369252]
100. Valero-Cantero I, Casals C, Espinar-Toledo M, Barón-López FJ, Martínez-Valero FJ, Vázquez-Sánchez M. Cancer patients' satisfaction with in-home palliative care and its impact on disease symptoms. Healthcare (Basel) 2023 Apr 29;11(9):1272 [FREE Full text] [doi: 10.3390/healthcare11091272] [Medline: 37174814]
101. Saric J, Kiefer S, Peshkatari A, Wyss K. Assessing the quality of care at primary health care level in two pilot regions of Albania. Front Public Health 2021;9:747689 [FREE Full text] [doi: 10.3389/fpubh.2021.747689] [Medline: 35004572]
102. Perfors IA, Helsper CW, Noteboom EA, Visserman EA, van Dorst EB, van Dalen T, et al. Effects of structured involvement of the primary care team versus standard care after a cancer diagnosis on patient satisfaction and healthcare use: the GRIP

XSL•FO
RenderX

randomised controlled trial. BMC Prim Care 2022 Jun 04;23(1):145 [FREE Full text] [doi: 10.1186/s12875-022-01746-3] [Medline: 35659264]
103. Feng Y, Gravelle H. Patient self-reported health, clinical quality, and patient satisfaction in English primary care: practice-level longitudinal observational study. Value Health 2021 Nov;24(11):1660-1666 [FREE Full text] [doi: 10.1016/j.jval.2021.05.019] [Medline: 34711367]
104. Eksteen LB, Mash RJ. Evaluating the validity and reliability of the Medical Interview Satisfaction Scale in South African primary care consultations. Fam Pract 2019 May 23;36(3):310-316. [doi: 10.1093/fampra/cmy076] [Medline: 30192946]
105. Stark S, Worm L, Kluge M, Roos M, Burggraf L. The patient satisfaction in primary care consultation-Questionnaire (PiC): an instrument to assess the impact of patient-centred communication on patient satisfaction. PLoS One 2021;16(7):e0254644 [FREE Full text] [doi: 10.1371/journal.pone.0254644] [Medline: 34270615]
106. Lee CF, Wang XF, Wong PN, Koh YL, Ngoh SH, Mohtar ZM, et al. Psychometric evaluation of a patient satisfaction survey questionnaire to assess advanced practice nurse ambulatory services in primary care. J Nurs Manag 2020 Oct;28(7):1481-1488 [FREE Full text] [doi: 10.1111/jonm.13072] [Medline: 32594581]
107. Steyl T. Satisfaction with quality of healthcare at primary healthcare settings: perspectives of patients with type 2 diabetes mellitus. S Afr J Physiother 2020;76(1):1321 [FREE Full text] [doi: 10.4102/sajp.v76i1.1321] [Medline: 32285017]
108. Abdul Aziz AF, Tan CE, Ali MF, Aljunid SM. The adaptation and validation of the satisfaction with stroke care questionnaire (Homesat) (SASC10-My™) for use in public primary healthcare facilities caring for long- term stroke survivors residing at home in the community. Health Qual Life Outcomes 2020 Jun 20;18(1):193 [FREE Full text] [doi: 10.1186/s12955-020-01450-9] [Medline: 32563246]
109. Satpathy S, Wundaville LT, Satapathy S, Malik A, Singh S, Singh AR, et al. A systematic review of patient satisfaction scales and their applicability to COVID-19 hospitalized patients: gaps and emerging needs. J Patient Exp 2022;9:23743735221079132 [FREE Full text] [doi: 10.1177/23743735221079132] [Medline: 35237716]
110. Care coordination measures atlas update: chapter 6. Measure maps and profiles (continued, 3). Agency for Healthcare Research and Quality. URL: https://www.ahrq.gov/ncepcr/care/coordination/atlas/chapter6b.html [accessed 2025-05-29]
111. Pasarín MI, Berra S, González A, Segura A, Tebé C, García-Altés A, et al. Evaluation of primary care: the "primary care assessment tools - facility version" for the Spanish health system. Gac Sanit 2013 Jan;27(1):12-18 [FREE Full text] [doi: 10.1016/j.gaceta.2012.03.009] [Medline: 22921943]
112. Shi L, Starfield B, Xu J. Validating the adult primary care assessment tool. J Fam Pract 2001;50(2):161 [FREE Full text]
113. Grol R, Wensing M, Mainz J, Jung HP, Ferreira P, Hearnshaw H, European Task Force on Patient Evaluations of General Practice Care (EUROPEP). Patients in Europe evaluate general practice care: an international comparison. Br J Gen Pract 2000 Nov;50(460):882-887 [FREE Full text] [Medline: 11141874]
114. Dimova R, Stoyanova R, Keskinova D. The EUROPEP questionnaire for patient's evaluation of general practice care: Bulgarian experience. Croat Med J 2017 Feb 28;58(1):63-74 [FREE Full text] [doi: 10.3325/cmj.2017.58.63] [Medline: 28252877]
115. Mead N, Bower P, Roland M. The General Practice Assessment Questionnaire (GPAQ) - development and psychometric characteristics. BMC Fam Pract 2008 Feb 20;9:13 [FREE Full text] [doi: 10.1186/1471-2296-9-13] [Medline: 18289385]
116. White B. Measuring patient satisfaction: how to do it and why to bother. Fam Pract Manag 1999 Jan;6(1):40-44 [FREE Full text] [Medline: 10344931]
117. Boateng GO, Neilands TB, Frongillo EA, Melgar-Quiñonez HR, Young SL. Best practices for developing and validating scales for health, social, and behavioral research: a primer. Front Public Health 2018;6:149 [FREE Full text] [doi: 10.3389/fpubh.2018.00149] [Medline: 29942800]
118. Fatima I, Humayun A, Iqbal U, Shafiq M. Dimensions of service quality in healthcare: a systematic review of literature. Int J Qual Health Care 2019 Feb 01;31(1):11-29. [doi: 10.1093/intqhc/mzy125] [Medline: 29901718]
119. Ferreira DC, Vieira I, Pedro MI, Caldas P, Varela M. Patient satisfaction with healthcare services and the techniques used for its assessment: a systematic literature review and a bibliometric analysis. Healthcare (Basel) 2023 Feb 21;11(5):378 [FREE Full text] [doi: 10.3390/healthcare11050639] [Medline: 36900644]
120. Duc Thanh N, My Anh BT, Xiem CH, Quynh Anh P, Tien PH, Thi Phuong Thanh N, et al. Patient satisfaction with healthcare service quality and its associated factors at one polyclinic in Hanoi, Vietnam. Int J Public Health 2022;67:1605055 [FREE Full text] [doi: 10.3389/ijph.2022.1605055] [Medline: 36439278]
121. Munawarah S, Arifin S, Febriana SK. Examining the associations between service quality, educational level, occupational background, and patient satisfaction in healthcare facilities. River Stud 2023 Aug 10;1(2):104-118. [doi: 10.61848/rst.v1i2.6]
122. Al Meslamani AZ. Technical and regulatory challenges of digital health implementation in developing countries. J Med Econ 2023;26(1):1057-1060 [FREE Full text] [doi: 10.1080/13696998.2023.2249757] [Medline: 37594521]
123. Lawrence K. Digital Health Equity. Brisbane, Australia: Exon Publication; 2022.
124. Li S, Cui G, Yin Y, Xu H. Associations between health literacy, digital skill, and eHealth literacy among older Chinese adults: a cross-sectional study. Digit Health 2023;9:20552076231178431 [FREE Full text] [doi: 10.1177/20552076231178431] [Medline: 37256010]

XSL•FO
RenderX

125. Nakayama LF, Binotti WW, Link Woite N, Fernandes CO, Alfonso PG, Celi LA, et al. The digital divide in Brazil and barriers to telehealth and equal digital health care: analysis of internet access using publicly available data. J Med Internet Res 2023 Jul 21;25:e42483 [FREE Full text] [doi: 10.2196/42483] [Medline: 37477958]
126. Cornejo Müller A, Wachtler B, Lampert T. [Digital divide-social inequalities in the utilisation of digital healthcare]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2020 Feb;63(2):185-191 [FREE Full text] [doi: 10.1007/s00103-019-03081-y] [Medline: 31915863]

Abbreviations

CAHPS: Consumer Assessment of Health Care Providers and Systems
COREQ: Consolidated Criteria for Reporting Qualitative Research
DHS: digital health solution
DigiHEALTHQUAL: Digital Healthcare Service Quality
HCP: health care practitioner
HCSQ: health care service quality
HEALTHQUAL: Health Care Service Quality
OPEQ: Outpatient Experience Questionnaire
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-RR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Rapid Reviews
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
PSQ-18: Patient Satisfaction Questionnaire-Short Form
SERVQUAL: Service Quality

Edited by A Mavragani; submitted 01.11.24; peer-reviewed by O Allela; comments to author 06.02.25; revised version received 25.05.25; accepted 25.05.25; published 24.07.25.

Please cite as:
Rigo D, Fehring L, Mortsiefer A, Meister S
Service Quality Assessment of Digital Health Solutions in Outpatient Care: Qualitative Item Repository Development Study
JMIR Form Res 2025;9:e68276
URL: https://formative.jmir.org/2025/1/e68276
doi:10.2196/68276
PMID:40705408

©Dominik Rigo, Leonard Fehring, Achim Mortsiefer, Sven Meister. Originally published in JMIR Formative Research (https://formative.jmir.org), 24.07.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Predicting Social Frailty in Older Adults Using Fitbit-Derived Circadian and Heart Rate Biomarkers: Cross-Sectional Study

Hiroki Maekawa[1,2*], MSRS; Yu Kume[3*], PhD

[1]Department of Rehabilitation, Faculty of Health Sciences, Tohoku Fukushi University, Sendai, Japan
[2]Doctoral Course of Health Sciences, Graduate School of Medicine, Akita University, Akita, Japan
[3]Department of Health Sciences, Graduate School of Medicine, Akita University, Akita, Japan
[*]all authors contributed equally

Corresponding Author:
Hiroki Maekawa, MSRS
Department of Rehabilitation
Faculty of Health Sciences
Tohoku Fukushi University
6-149-1, Kunimigaoka
Aoba-ku
Sendai, 989-3201
Japan
Phone: 81 22 728 6030
Email: h-maekawa@tfu.ac.jp

Abstract

Background: Social frailty poses a potential risk even for relatively healthy older adults, necessitating development of early detection and prevention strategies. Recently, consumer-grade wearable devices have attracted attention due to their ability to continuously collect physiological and activity-related data. These data can potentially be used to calculate digital biomarkers for screening social frailty in older adults.

Objective: The objective of this study was to explore digital biomarkers associated with social frailty using sensor data recorded via Fitbit devices and evaluate their relationship with health outcomes in older adults.

Methods: This cross-sectional study was conducted in 102 community-dwelling older adults. Participants attending frailty prevention programs wore devices from the Fitbit Inspire series on their nondominant wrist for at least 7 consecutive days, during which step count and heart rate data were collected. Standardized questionnaires were used to assess physical functions, cognitive functions, and social frailty, and based on the scores, the participants were categorized into 3 groups: robust, social prefrailty, and social frailty. The sensor data were analyzed to calculate nonparametric and extended cosinor rhythm metrics, along with heart rate–related metrics.

Results: The final sample included 86 participants who were categorized as robust (n=28, 33%), social prefrailty (n=39, 45%), and social frailty (n=19, 22%). The mean age of the participants was 77.14 (SD 5.70) years, and 91% (78/86) were women. Multinomial logistic regression analysis revealed that a step-based rhythm metric (intradaily coefficient of variation) was significantly associated with social frailty (odds ratio 1.05, 95% CI 1.01-1.11; *P*=.01). The heart rate metrics, including the delta resting heart rate and time of transition from rest to activity, showed significant associations with both social prefrailty (odds ratio 0.82, 95% CI 0.68-0.99; *P*=.04) and social frailty (odds ratio 0.69, 95% CI 0.50-0.95; *P*=.01). Specifically, delta resting heart rate, defined as the difference between the overall average heart rate and resting heart rate, exhibited significant negative associations with social prefrailty (odds ratio 0.82, 95% CI 0.68-0.97; *P*=.02) and social frailty (odds ratio 0.74, 95% CI 0.58-0.94; *P*=.02). Furthermore, analysis using a linear regression model revealed a significant association between the intradaily coefficient of variation and the word list memory score, a measure of cognitive decline (β=−0.04; *P*=.02).

Conclusions: This study identified associations between novel rhythm and heart rate metrics calculated from the step count and heart rate recorded by Fitbit devices and social frailty. These findings suggest that consumer-grade wearable devices, which are low cost and accessible, hold promise as tools for evaluating social frailty and its risk factors through enabling the calculation of digital biomarkers. Future research should include larger sample sizes and focus on the clinical applications of these findings.

(JMIR Form Res 2025;9:e71393) doi:10.2196/71393

XSL•FO
RenderX

KEYWORDS

social frailty; digital biomarkers; Fitbit; rest-activity rhythm; heart rate metrics; nonparametric indexes; extended cosinor model; community-dwelling older adults

Introduction

Background

In an aging society, early detection and prevention of frailty are urgent priorities. Frailty is a multifaceted concept encompassing physical, cognitive, and social aspects, all of which play critical roles in comprehensively assessing the health status of older adults [1,2]. Among these, social frailty precedes physical frailty and has been reported to promote unhealthy aging while increasing the risks of all-cause mortality and functional impairment [3,4]. Bunt et al [5] proposed that, in addition to general resource–related factors such as living alone and economic difficulties, social resource factors such as connections with neighbors and friends, as well as basic social needs such as social behavior, social isolation, and social support, are critical components of social frailty. Social frailty profoundly affects the lifestyles of older adults and has been associated with a poor nutritional status, depression, reduced physical capability, and decreased activity levels [6]. Furthermore, it has been linked to declines in activities of daily living [7] and instrumental activities of daily living [8,9]. Rest-activity rhythm (RAR) has long garnered attention as an established indicator for evaluating health conditions related to such lifestyle factors.

Previous studies using actigraphy have used RAR parameters derived from activity trackers to predict various health outcomes, including physical frailty, cognitive decline, and depression [10-14]. These parameters quantify 24-hour cycles of rest and activity based on activity counts and have been widely used in health research. However, their application to social frailty remains underexplored [15], and studies leveraging wearable sensors for this purpose are particularly scarce.

Recently, wearable devices such as Fitbit smartwatches have enabled the collection of detailed physiological and behavioral data, including heart rate, respiratory rate, and sleep data in addition to traditional activity data. These advancements have spurred the development of innovative digital biomarkers such as heart rate–based rhythm metrics, which complement conventional activity count–based indicators [16,17]. Fitbit devices, known for their affordability and accessibility, have demonstrated accuracy comparable to research-grade devices for measuring step count and heart rate [18-20]. Despite their potential, the use of wearable devices to evaluate social frailty and its associated risk factors remains limited, highlighting the need for further investigation, such as studies similar to the one reported in this paper.

Objectives

Fitbit and other wearable devices provide step count, heart rate, and other data that allow for the calculation of digital biomarkers to predict health outcomes. However, studies using raw data from these devices to compute novel metrics and evaluate their associations with health outcomes remain insufficient.

The primary aim of this study was to identify digital biomarkers associated with social frailty. Social frailty has been linked to reduced daytime physical activity, as demonstrated in previous research [7,9]. Although there is limited evidence directly connecting social frailty to heart rate, it is plausible that decreased physical activity in individuals with social frailty leads to lower overall heart rate variability (HRV) and altered heart rate patterns. Therefore, we hypothesized that the frailty group would exhibit significant associations with RAR parameters (nonparametric and extended cosinor analysis) and heart rate–based parameters (eg, nighttime heart rate and resting heart rate [RHR]), assessed by leveraging Fitbit data. These parameters have the potential to shed light on the physiological and behavioral characteristics of social frailty.

The secondary aim of this study was to explore the relationship between RAR metrics derived from Fitbit data and cognitive function. While previous research has linked nonparametric and cosinor-based RAR metrics to cognitive decline, most studies have relied on research-grade actigraphy devices. This study introduces novel RAR metrics derived from Fitbit data to exploratorily assess their associations with cognitive function.

Methods

Study Design and Participants

This study was designed as a cross-sectional analysis to investigate the relationship between digital biomarkers and social frailty in community-dwelling older adults in Akita Prefecture, Japan. To determine the appropriate sample size for the logistic regression analysis, we used G*Power (version 3.1.9.6) with the likelihood-ratio test as the test statistic. The estimation indicated that a minimum of 82 participants would be required to detect a clinically significant effect based on the following parameters: 3 groups, a significance level (α) of .05, a statistical power of 80%, and an odds ratio (OR) of 2.0 [21]. Considering a potential dropout rate of 20% to 30%, we recruited a total of 102 older adults between October 2023 and December 2024. Recruitment sites included municipally organized programs conducted at care management centers in the Yabase, Kawabe, Ushijima, Goshono, Iijima, and Shimoshinjo districts of the city of Akita, as well as in the city of Yurihonjō and the town of Mitane. Additional participants were recruited from a community-based exercise group regularly followed by the principal investigator held at a temple in the Katsuhira area. The inclusion criteria were (1) age of ≥65 years, (2) independent walking ability in daily life, and (3) residence at home. The exclusion criteria were (1) a central nervous system disorder, (2) requiring assistance in daily living, (3) a cardiac condition, or (4) requiring support or care as certified under Japan's public long-term care insurance system.

Data collection used consumer-grade fitness trackers, specifically the Fitbit Inspire 2 and 3 (Google). The accuracy of Fitbit data has been evaluated in several studies. According to systematic reviews [22,23], Fitbit devices consistently

XSL•FO
RenderX

demonstrate acceptable levels of accuracy for daily step counts and heart rate measurements despite a tendency to underestimate these metrics under certain conditions. Generally, Fitbit wearables provide activity assessment accuracy comparable to research-grade devices but tend to overestimate moderate to vigorous physical activity under free-living conditions. One of the major challenges faced by medical researchers when interpreting physical activity data is related to participants forgetting to wear the activity tracker [23]. However, the Inspire series of Fitbit devices, being waterproof and offering a battery life of up to 10 days, minimizes interruptions in data collection. These features contribute to ensuring high-quality data with fewer missing values, making them particularly advantageous for calculating rhythm metrics.

Participants were instructed to wear the device for at least 7 consecutive days. The research staff placed the device on the participants' nondominant wrist, approximately 1 inch below the wrist bone, ensuring that the back of the device was in contact with the skin. Participants were instructed to wear the tracker at all times, including during bathing, whenever possible. However, they were advised not to use the device in situations prohibited by the manufacturer, such as diving under high water pressure or in high-temperature environments such as saunas.

Ethical Considerations

This study was conducted in accordance with the ethical principles outlined in the Declaration of Helsinki. The research protocol and informed consent forms were approved by the institutional review board of Akita University School of Medicine (approval 3063). All participants provided written informed consent before taking part. This study involved both the collection of wearable device data and in-person assessments, including evaluations of social frailty and cognitive performance. To ensure participant confidentiality and data protection, all collected data were anonymized before analysis. Participants did not receive monetary compensation, nor did they incur any costs by taking part in the study. As a token of appreciation, each participant received a booklet summarizing their individual results and interpretations.

Classification of Social Frailty Index and Assessment

Demographic data, including age (years), sex (female or male), and educational level (years), were collected from all participants. Social frailty was classified using the social frailty index developed by Makizako et al [24], which reflects the accumulation of social risk factors and indicates reduced social participation or roles. The index consists of five components: (1) living alone (yes), (2) talking with someone every day (no), (3) feeling helpful to friends or family (no), (4) going out less frequently compared with the previous year (yes), and (5) visiting friends sometimes (no). Participants with a total score of 0 were classified as robust, those with a score of 1 were classified as social prefrailty, and those with scores of 2 to 5 were classified as social frailty. Physical performance was assessed by measuring grip strength (kg) and usual walking speed (m/s) over a 5-m course. Cognitive performance was evaluated using 4 subtests from the National Center for Geriatrics and Gerontology–Functional Assessment Tool [25]. The word list memory (WLM) test included immediate recognition and delayed recall tasks, where the mean correct answers across 3 trials (range 0-10) and the number of correct answers in delayed recall (range 0-10) were summed to calculate a total score (range 0-20). The trail making test versions A and B was used to measure cognitive flexibility, with completion times recorded in seconds. Finally, the digit symbol substitution task was conducted to assess information processing speed, where the number of correct responses in 90 seconds was recorded.

Wearable Data Preprocessing

After the monitoring period, devices were collected, and their stored data were synchronized via Bluetooth to a smartphone before being transferred to the Fitbit database. Researchers then accessed the Fitbit developer platform linked to a Google account [26], generated an access token (refreshed every 8 hours), and executed a Python script (Python Software Foundation) in Google Colaboratory to retrieve the raw step count and heart rate data at 1-minute intervals. To ensure data integrity, periods without heart rate data were excluded, and corresponding step count data were marked as invalid; only time points with both valid heart rate and step count data were retained as clean data. Consistent with the work by Rykov et al [17], a valid day was defined as one with at least 20 hours of concurrent heart rate and step count data. On the basis of evidence that 5 days of wear yield reliable RAR parameters in older adults [27], only participants with at least 5 valid days were included in the study.

Extraction of Digital Biomarkers

Raw data from Fitbit were used to extract digital biomarkers related to circadian rhythms and heart rate variation. Step count and heart rate data were processed using a modified version of the script developed by Rykov et al [17] to calculate both extended cosine-based and nonparametric circadian rhythm indexes.

The Extended Cosinor Metrics

Extended cosine-based indexes included the midline estimating statistic of rhythm (MESOR), acrophase, amplitude, minimum, pseudo–*F* statistic, time of transition from rest to activity (UpMesor), time of transition from activity to rest (DownMesor), and the α and β parameters. These indexes were derived using an extended cosinor model, which incorporates nonlinear transformations to refine the classic cosine function $c(t)$:

$$r(t) = \text{mes} + \text{amp} \times c(t)$$

In these equations, *mes* is the rhythm-adjusted mean level, *amp* is the amplitude, and φ (acrophase) is the time of the mathematically defined peak.

In contrast, the extended cosinor model applies nonlinear transformations to $c(t)$, enabling it to represent more complex waveform shapes. Three transformation methods are commonly used: (1) hill transformation, (2) antilogistic transformation, and (3) arctangent transformation.

XSL•FO
RenderX

These transformations allow for better modeling of circadian rhythms that deviate from simple cosine waveforms [28,29]. This study used the antilogistic transformation, which is defined as follows:

The extended cosinor model, incorporating this transformation, is expressed as $r(t) = \min + \text{amp} \times l(c[t])$, where *min* is the minimum value of the function, *amp* is the difference between the minimum and maximum values, $c(t)$ represents the classic cosine wave, and $l(c[t])$ is the antilogistically transformed function.

Interpretation-wise, *min* is the minimum value of the function, *amp* is the difference between the minimum and maximum of the function, φ is the time at which $r(t)$ has its mathematically well-defined "peak," α controls the "width" of the function, and β controls the "steepness" of the function.

In this study, we used the extended cosinor model to analyze rhythmic patterns derived from Fitbit data. *Minimum* represents the lowest value of the fitted function, indicating the period of least activity or lowest heart rate within a 24-hour cycle. *Amplitude* measures half the extent of predictable variation within a cycle, serving as an indicator of the maximum achievable activity level. The α parameter determines whether the peaks of the curve are wider or narrower than the troughs; specifically, a small α value signifies narrow troughs and wide peaks, whereas a large α value indicates wide troughs and narrow peaks. The β parameter assesses the steepness of the rise and fall of the curve relative to a standard cosine wave, with large β values producing nearly square waveforms that represent abrupt transitions between high and low activity levels. *Acrophase* refers to the time of day at which the peak of the rhythmic pattern occurs, initially measured in radians and subsequently converted into time units (hours) and referred to as acrotime when necessary. *F_pseudo* measures the improvement in the fit obtained through the nonlinear estimation of the transformed cosine model, serving as an indicator of the robustness of the rhythm. *UpMesor* denotes the time of day when activity transitions from low to high, reflecting the timing of the RAR; lower (earlier) values suggest an earlier onset of daytime activity and a more advanced circadian phase. *DownMesor* indicates the time of day when activity transitions from high to low, representing the cessation of peak activity within the RAR; lower (earlier) values imply an earlier decline in daily activities, also suggesting a more advanced circadian phase.

Finally, MESOR is an estimated 24-hour average activity level calculated as the sum of the minimum plus half the amplitude (MESOR = [minimum + amplitude]/2) and provides a measure of the central tendency in the daily activity cycle. It should be noted that the MESOR in the extended cosinor model differs from the MESOR in the standard cosinor model [29]. A visual representation of the extended cosinor variables used in this study is presented in Figure 1.

XSL•FO
RenderX

Figure 1. Illustration of the extended cosinor model applied to 24-hour heart rate data from a representative participant. The green curve represents the fitted extended cosinor function used to estimate rhythm-related parameters: minimum (lowest value), amplitude (half the predictable variation), midline estimating statistic of rhythm (MESOR; 24-hour mean), acrotime (time of peak), transition from rest to activity (UpMesor), transition from activity to rest (DownMesor), α (width of peak), and β (steepness of transition). Data were collected as part of a cross-sectional study conducted between October 2023 and December 2024 using Fitbit Inspire 2 and 3 devices worn by community-dwelling older adults in Akita Prefecture, Japan.

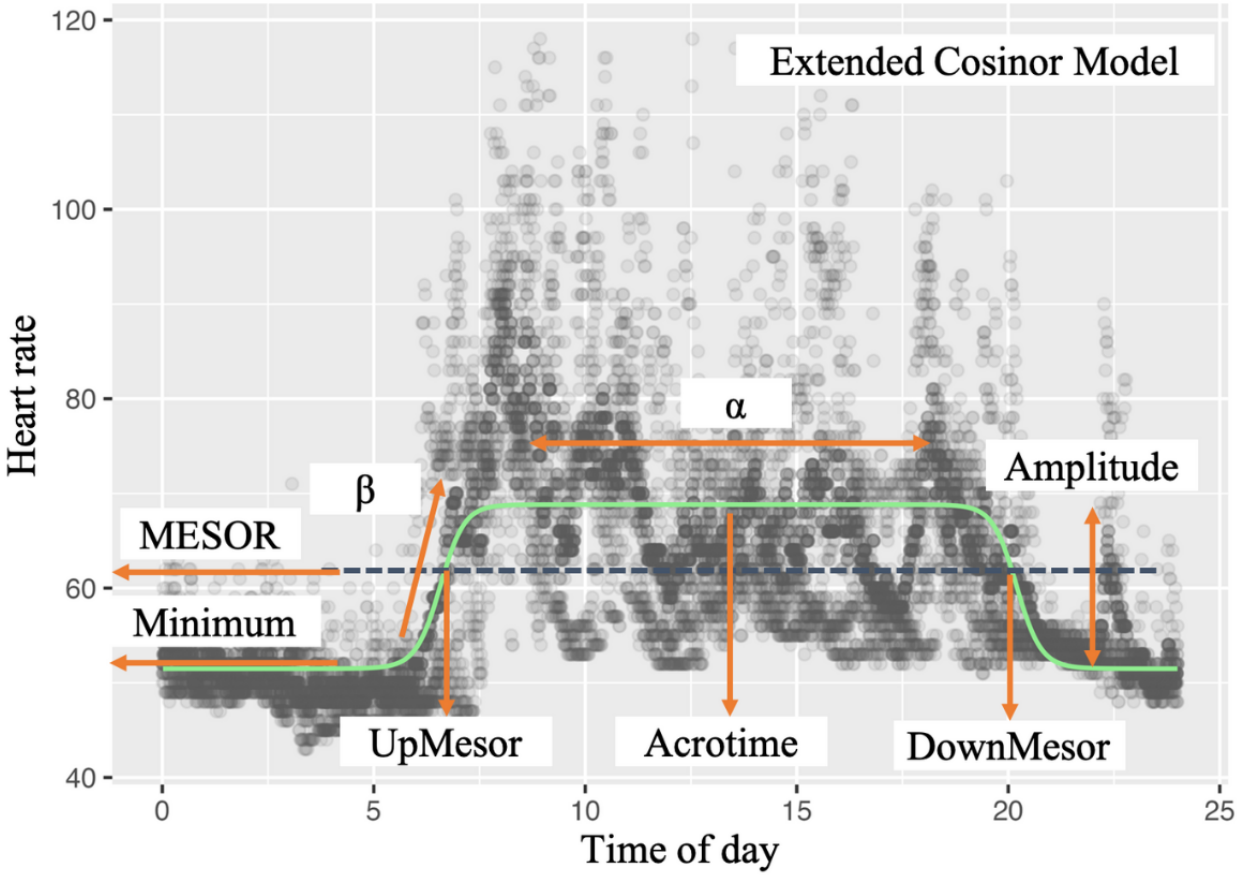

Compared to the classic cosine model, the extended cosinor model provides greater flexibility to capture complex waveform shapes. The extended cosinor model—through the application of an antilogistic transformation—can accommodate, for example, rapid activity onset in the morning, sustained daytime activity, and gradual transitions into nighttime rest. This approach enables a more realistic fit of circadian patterns while preserving interpretability [29].

Nonparametric Metrics

Nonparametric indexes included interdaily stability (IS), intradaily variability (IV), daytime activity level (most active 10-hour activity level [M10]), nighttime activity level (least active 5-hour activity level [L5]), relative amplitude (RA), intradaily coefficient of variation (ICV), autocorrelation, and peaks. Among these, IS, IV, M10, L5, and RA were originally proposed in studies by Van Someren et al [30] and Witting et al [31]. ICV, autocorrelation, and peaks have been introduced in more recent studies to provide complementary insights into circadian rhythm variability and stability.

IS measures the stability and regularity of activity patterns across 24-hour cycles. It is calculated as the ratio of the variance of the average 24-hour activity profile to the total variance of data aggregated by the hour across all days. A higher IS value indicates a more stable and regular circadian rhythm.

IV quantifies the fragmentation of rest and activity periods within a 24-hour cycle. It is calculated as the mean square of the differences between successive hourly aggregated data normalized by the total variance across all days. A higher IV value indicates a more fragmented rhythm, characterized by shorter alternating periods of rest and activity rather than 1 extended active period and 1 extended rest period.

RA reflects the difference between the most and the least active periods during the day. It is calculated as the difference between M10 (mean activity level during the most active 10 hours of the day) and L5 (mean activity level during the least active 5 hours of the day) divided by the sum of M10 and L5. Higher RA values indicate a greater amplitude in activity levels.

ICV is a novel measure of rhythm stability proposed by Rykov et al [17]. Unlike IS, which evaluates the stability and regularity of daily rhythms across days by calculating the variance of the mean 24-hour activity profile relative to the total variance, ICV focuses on the variation within a single day. ICV is calculated as the 24-hour mean of the coefficients of variation (CVs), where the CV is the ratio of the SD of the mean for each hour across days. Higher ICV values indicate greater variation and less stable rhythms within a day. In general, individuals with more regular and healthy lifestyles tend to exhibit higher IS values, reflecting their stable and consistent daily rhythms, and lower

XSL•FO
RenderX

ICV values, indicating more stable activity patterns within each hour of a day. Conversely, higher ICV values may suggest fragmented or inconsistent activity patterns, often characterized by intermittent bursts of activity and prolonged periods of inactivity.

Autocorrelation is another measure of rhythm stability calculated as the lagged autocorrelation of time-series data. Autocorrelation was computed for time series aggregated into 15-, 30-, and 60-minute intervals, with a lag corresponding to 1 day.

In addition, a robust peak detection algorithm based on *z* scores was applied to the time-series data to identify peaks in steps and heart rate data [32]. For this, the daily mean number of peaks and its SD were calculated.

Heart Rate Metrics

Extracted heart rate–based metrics included the overall average heart rate, RHR, delta RHR (dRHR), daytime and nighttime heart rate, and the variability of these measures calculated using the SD and CV. In addition, the root mean square of successive differences (RMSSD) of the heart rate was calculated.

RHR was defined as the average heart rate during 15-minute intervals with 0 recorded steps, representing resting periods. Daytime heart rate was calculated as the mean heart rate between 2 PM and 4 PM, whereas nighttime heart rate was calculated by averaging values from 3 consecutive 2-hour intervals: midnight to 2 AM, 2 AM to 4 AM, and 4 AM to 6 AM. These intervals and metrics were selected following the approach used by Rykov et al [17].

The SD and CV were computed for each heart rate measure to assess variability. dRHR, reflecting the difference between activity and rest, was calculated as the difference between the overall average heart rate and RHR. RMSSD, an established metric for assessing HRV, was calculated using both raw data and hourly aggregated data.

Custom Modifications to Digital Biomarker Calculation Programs

The calculation of digital biomarkers in this study was conducted using step and heart rate data by executing the R script provided by Rykov et al in R (version 4.4.1; R Foundation for Statistical Computing). However, for extended cosinor–based metrics, the *ActCR* package available on the Comprehensive R Archive Network was used instead of the actigraphy package used by Rykov et al [17]. The *ActCR* package offers functionality to calculate a greater number of extended cosinor–based metrics than the actigraphy package and was adopted to comprehensively capture the participants' circadian rhythms [33].

Although only days with >20 hours of data were included in the analysis, missing values were still present in some datasets. Specifically, for the calculation of peaks, the original program by Rykov et al [17] did not include a mechanism to skip missing values, which could result in errors and prevent the metric from being calculated correctly. To address this limitation, we modified the program by adding a process to skip missing values during the calculation of peaks, ensuring that the metric could be computed accurately using the available data points. In total, we extracted 69 digital biomarkers. The R scripts used for feature extraction are available in Multimedia Appendix 1.

Statistical Analysis

The normality of the variables was assessed using the Shapiro-Wilk test. For variables following a normal distribution, multiple regression analysis was conducted with the WLM score as the dependent variable, and explanatory variables were selected based on Pearson correlation coefficients. For nonnormally distributed variables, the Spearman correlation analysis was conducted. The Spearman analysis was applied to the entire dataset (all participants) as well as separately to the robust, social prefrailty, and social frailty groups. Kruskal-Wallis tests, followed by Dunn post hoc tests, were used to evaluate differences across the 3 frailty groups. Finally, multinomial logistic regression models were constructed to identify predictors of social frailty status (robust, social prefrailty, and social frailty). On the basis of the results of the Kruskal-Wallis tests, models I and III included age- and sex-adjusted factors as independent variables, whereas models II and IV were unadjusted for these factors. All analyses were conducted using RStudio (version 4.4.1; Posit PBC).

Results

Characteristics of the Data and Participants

The average device wearing time per participant was 186 (SD 26) hours, of which an average of 164 (SD 30) hours per participant consisted of complete data recorded on days with at least 20 hours of valid data. Of the 102 participants, only 86 (84.3%) had sufficient data for at least 5 days and were included in the analysis.

The participants had a mean age of 77.14 (SD 5.7; range 65-88) years, and most were women (78/86, 91%). The average daily step count was 7824 (SD 3794), and the mean heart rate was 72.3 (SD 5.9) beats per minute. According to the social frailty criteria defined by Makizako et al [24], the sample was categorized into robust (28/86, 33%), social prefrailty (39/86, 45%), and social frailty (19/86, 22%) groups.

Figure 2 presents the averaged 24-hour profiles of hourly heart rate and step count for all participants (n=86). Figure 3 shows the same 24-hour profiles stratified by social frailty status (robust, social prefrailty, and social frailty).

XSL•FO
RenderX

Figure 2. Hourly profiles of heart rate and step count across a 24-hour period among community-dwelling older adults (N=86) in Akita Prefecture, Japan. Data were collected between October 2023 and December 2024 as part of a cross-sectional study examining social frailty and circadian biomarkers. Participants wore Fitbit Inspire 2 and 3 devices for at least 7 consecutive days. The hourly averages (eg, midnight-12:59 AM and 1 AM-1:59 AM) of heart rate and step count were computed using only time windows with at least 30 minutes of valid data per participant. Red and blue lines indicate mean values, and gray shaded areas represent SDs. bpm: beats per minute.

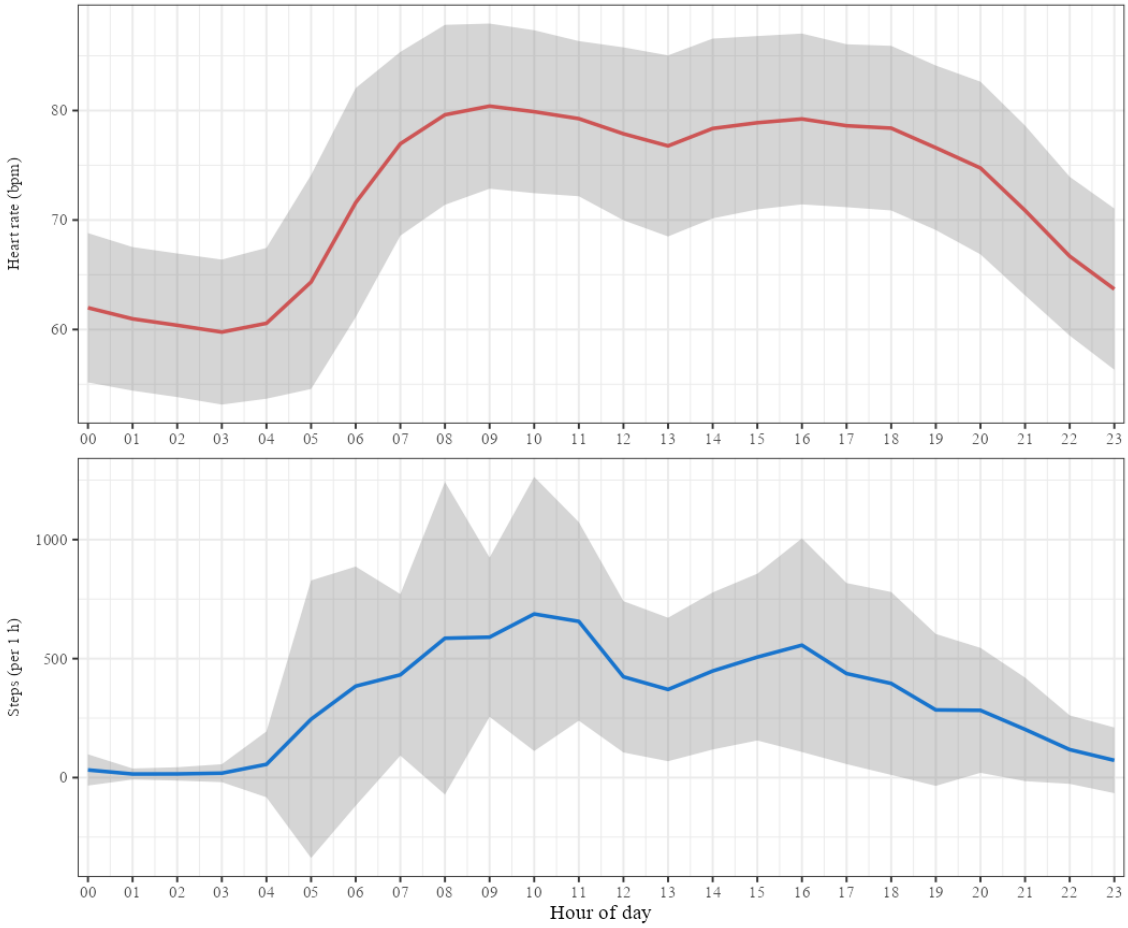

XSL•FO
RenderX

Figure 3. Average 24-hour profiles of heart rate and step count measured using Fitbit Inspire 2 and 3 devices among 3 groups of community-dwelling older adults in Akita Prefecture, Japan: robust (n=28), social prefrailty (n=29), and social frailty (n=19). Data were collected between October 2023 and December 2024 as part of a cross-sectional study on digital biomarkers of social frailty. The hourly means (eg, midnight-12:59 AM and 1 AM-1:59 AM) and SDs were calculated for heart rate and step count. Only time windows with at least 30 minutes of valid data per participant were included in the analysis. The horizontal axis represents the hour of the day. bpm: beats per minute.

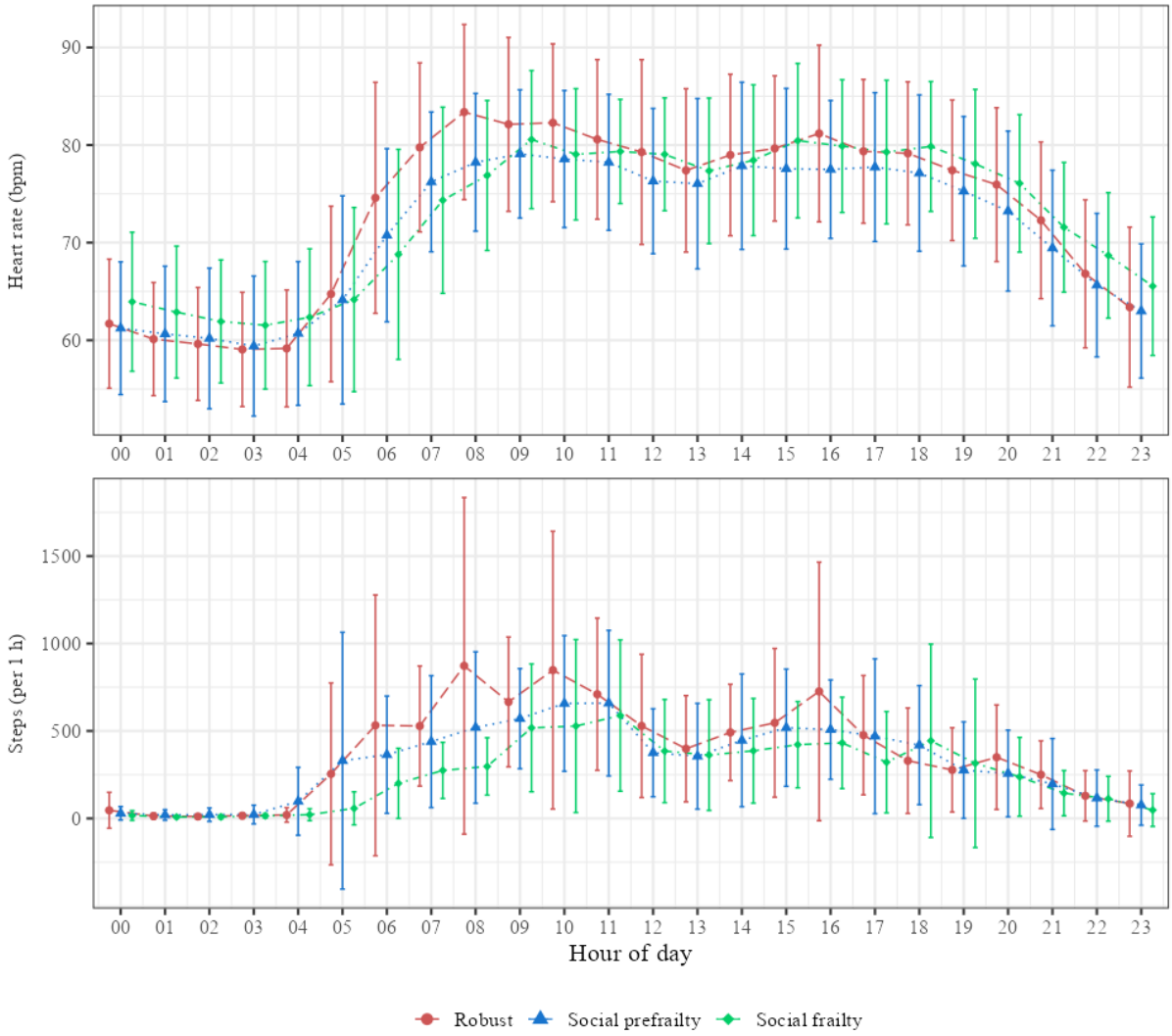

The demographic data for all participants and each social frailty group are summarized in Table 1, including key variables such as age, sex, and variables from physical and cognitive domains, as well as digital biomarkers. Only the key variables are shown in this table, with the full version available in Table S1 in Multimedia Appendix 2.

XSL•FO
RenderX

Table 1. Demographic information for all participants and each group (robust, social prefrailty, and social frailty; N=86)[a].

Variable	All participants	Robust (n=28)	Social prefrailty (n=39)	Social frailty (n=19)
Demographics				
Age (y), mean (SD)	77.14 (5.70)	76.29 (5.38)	77.18 (5.88)	78.32 (6.01)
Sex, n (%)				
Female	78 (91)	26 (93)	34 (87)	18 (95)
Male	8 (9)	2 (7)	5 (13)	1 (5)
Step count, mean (SD)	7824 (3794)	9080 (4529)	7809 (3104)	6005 (3424)
Physical domain variables, mean (SD)				
Grip strength (kg)	23.81 (5.38)	23.84 (4.37)	24.61 (5.91)	22.11 (5.61)
Usual walking speed (m/s)	1.36 (0.29)	1.34 (0.29)	1.38 (0.30)	1.33 (0.30)
Cognitive domain variables, mean (SD)				
WLM[b] (score)	12.26 (3.14)	12.62 (3.07)	12.49 (3.26)	11.26 (3.02)
TMT-A[c] (s)	1.33 (0.34)	1.31 (0.39)	1.30 (0.29)	1.40 (0.37)
TMT-B[d] (s)	3.22 (2.80)	3.04 (2.61)	3.42 (3.51)	3.09 (1.13)
DSST[e] (score)	44.70 (13.50)	46.96 (15.81)	45.56 (12.88)	39.58 (10.40)
Circadian rhythm metrics—nonparametric analysis, mean (SD)				
IS.st[f]	0.46 (0.17)	0.51 (0.19)	0.45 (0.16)	0.38 (0.13)
IV.st[g]	1.28 (0.27)	1.20 (0.23)	1.29 (0.27)	1.39 (0.28)
L5.st[h]	22 (27)	20 (24)	28 (34)	12 (10)
M10.st[i]	575 (287)	655 (339)	570 (230)	465 (292)
RA.st[j]	0.92 (0.09)	0.94 (0.06)	0.90 (0.11)	0.93 (0.07)
ICV.st[k]	1.12 (0.20)	1.08 (0.25)	1.09 (0.16)	1.25 (0.16)
Circadian rhythm metrics—extended cosinor analysis, mean (SD)				
Minimum.hr	58.71 (10.04)	58.79 (7.10)	57.19 (12.78)	61.72 (6.82)
Amplitude.hr	21.68 (14.24)	21.88 (8.49)	23.57 (19.70)	17.52 (3.97)
α parameter.hr	−0.41 (0.24)	−0.48 (0.15)	−0.35 (0.32)	−0.41 (0.13)
β parameter.hr	17.59 (20.99)	20.98 (21.19)	13.51 (15.58)	20.96 (30.25)
Acrotime.hr (h)	14.01 (1.09)	14.12 (0.85)	13.7 (1.21)	14.50 (1.01)
F_pseudo.hr[l]	1562.73 (1361.06)	2011.37 (1314.32)	1483.74 (1553.21)	1063.72 (759.72)
UpMesor.hr[m]	6.30 (1.20)	6.15 (0.92)	6.14 (1.41)	6.85 (0.98)
DownMesor.hr[n]	21.73 (1.97)	22.08 (1.25)	21.26 (2.55)	22.15 (1.30)
MESOR.hr[o]	69.55 (6.07)	69.73 (5.74)	68.97 (6.50)	70.48 (5.98)
HR[p] metrics, mean (SD)				
HR (bpm)	72.30 (5.91)	73.10 (5.98)	71.52 (6.00)	72.72 (5.89)
SD of the HR (bpm)	12.72 (2.48)	13.71 (2.74)	12.26 (2.25)	12.20 (2.30)
RHR[q] (bpm)	64.04 (6.52)	63.39 (6.57)	63.54 (6.46)	66.01 (6.70)
dRHR[r] (bpm)	8.26 (3.03)	9.71 (3.55)	7.98 (2.69)	6.70 (1.96)
RMSSD[s] of the HR	3.66 (1.03)	3.85 (1.04)	3.59 (1.06)	3.51 (0.99)
RMSSD of the HR based on hourly means	7.93 (1.86)	8.53 (2.21)	7.64 (1.62)	7.63 (1.71)

XSL•FO
RenderX

[a]This table presents demographic characteristics, cognitive performance scores, and rhythm-based digital biomarkers for community-dwelling older adults (N=102) participating in a cross-sectional study conducted in Akita Prefecture, Japan, between October 2023 and December 2024. Participants were classified into 3 groups: robust, social prefrailty, and social frailty. Cognitive tests included the word list memory test, the trail making test versions A and B, and the digit symbol substitution task. Wearable-derived metrics obtained from Fitbit Inspire 2 and 3 devices included average daily step counts and rhythm indicators derived from steps or heart rate data: interdaily stability, intradaily variability, least active 5-hour period, most active 10-hour period, relative amplitude, and interdaily coefficient of variation. The suffix ".st" denotes rhythm indicators based on steps; ".hr" denotes those based on heart rate. Additional physiological measures included resting heart rate, delta resting heart rate, and heart rate variability indexes such as root mean square of successive differences and root mean square of successive differences based on hourly means. Mean 186 (SD 26) sampling hours; mean 164 (SD 30) analyzed hours.

[b]WLM: word list memory.

[c]TMT-A: trail making test version A.

[d]TMT-B: trail making test version B.

[e]DSST: digit symbol substitution task.

[f]IS: interdaily stability.

[g]IV: intradaily variability.

[h]L5: least active 5-hour activity level.

[i]M10: most active 10-hour activity level.

[j]RA: relative amplitude.

[k]ICV: interdaily coefficient of variation.

[l]F_pseudo: the improvement in the fit obtained through the nonlinear estimation of the transformed cosine model.

[m]UpMesor: time of transition from rest to activity.

[n]DownMesor: time of transition from activity to rest.

[o]MESOR: midline estimating statistic of rhythm.

[p]HR: heart rate.

[q]RHR: resting heart rate.

[r]dRHR: delta resting heart rate.

[s]RMSSD: root mean square of successive differences.

Statistical Analysis Results

Figure 4 illustrates the results of the Spearman correlation analysis conducted across the entire dataset (all participants) and separately for the robust, social prefrailty, and social frailty groups. Age, physical domain, and cognitive domain variables are included in the analysis, whereas other variables are reported only if their correlation coefficients showed significant associations in at least 3 of the panels. Tables S2-S5 in Multimedia Appendix 2 provide detailed correlation coefficients.

XSL•FO
RenderX

Figure 4. Correlation matrices of rhythm-based digital biomarkers derived from Fitbit Inspire 2 and 3 devices among community-dwelling older adults (N=86) in Akita Prefecture, Japan. This cross-sectional study was conducted between October 2023 and December 2024. Participants were post hoc classified into 4 groups: all participants (A), robust (B; n=28), social prefrailty (C; n=29), and social frailty (D; n=19). The strength and direction of correlations are visualized using ellipses—deep red indicates strong negative correlations, deep blue indicates strong positive correlations, and lighter shades indicate weaker correlations. Variables other than age, physical function, and cognitive function are displayed only if their correlation coefficients showed significant differences across at least 3 of the 4 panels. α parameter.hr: the width of troughs and peaks in the rhythm; Amplitude.hr: half the extent of predictable variation within a cycle; DHR: daytime heart rate (average heart rate between 2 PM and 4 PM); dRHR: delta resting heart rate (HR); DSST: digit symbol substitution task; F_pseudo.hr: the improvement in the fit obtained through the nonlinear estimation of the transformed cosine model derived from HR; ICV.hr: intradaily coefficient of variation; M10.hr: most active 10-hour activity level; Peaks.hr: number of heart rate peaks per day detected in time-series data; RHR: resting HR; TMT-A: Trail Making Test version A; TMT-B: Trail Making Test version B; UpMesor.hr: time of transition from rest to activity; WLM: word list memory.

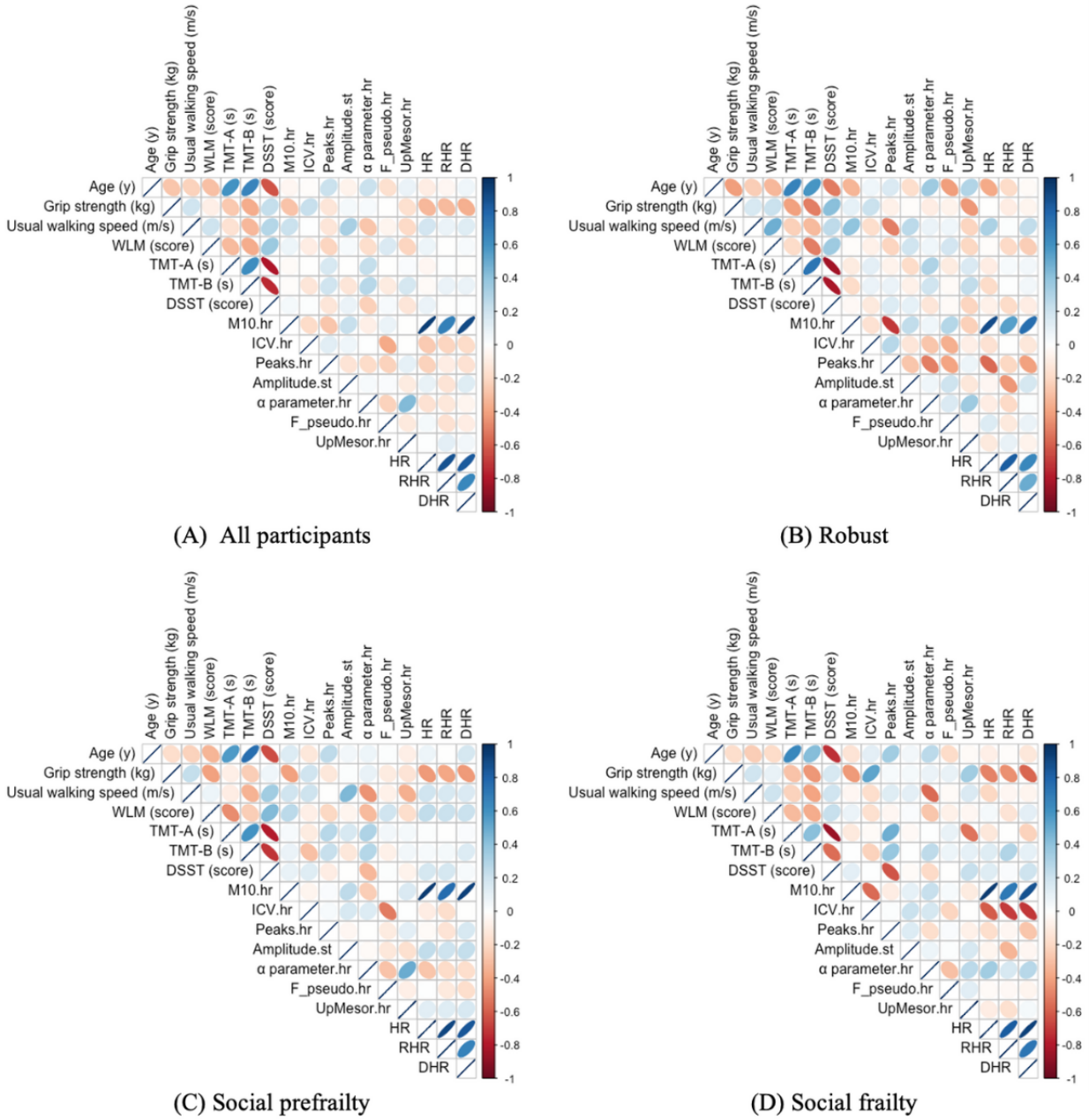

Table 2 presents the results of the multiple linear regression analyses with WLM score as the dependent variable. In total, 2 models were analyzed. Model I included age and ICV × 100 as independent variables, whereas model II added one more variable to the predictors. ICV × 100 represents the step-based intrahourly CV scaled by multiplying the original values by 100 to enhance interpretability.

XSL•FO
RenderX

Table 2. Models according to multiple logistic regression analysis[a].

	β coefficient (SE; 95% CI)	*P* value
Model I[b]		
Age	−0.15 (0.06; –0.26 to –0.03)	.01
ICV.st[c] × 100	−0.04 (0.02; –0.07 to –0.01)	.02
Model II[d]		
Age	−0.13 (0.06; –0.25 to –0.02)	.03
ICV.st × 100	−0.04 (0.02; –0.07 to 0.00)	.03
Peaks.hr[e]	−0.39 (0.29; –0.97 to 0.20)	.19

[a]The dependent variable for the analysis was word list memory score, derived from cognitive assessments. Independent variables included age, the step-based intrahourly coefficient of variation standardized and scaled by 100, and the number of heart rate peaks per day detected in time-series data. This cross-sectional study was conducted among community-dwelling older adults in Akita Prefecture, Japan, with data collected between October 2023 and December 2024. Model fit was evaluated using R^2 (proportion of variance explained), adjusted R^2 (adjusted for the number of predictors), and the *F* statistic (testing the overall model significance).

[b]R^2=0.150; adjusted R^2=0.129; *F* statistic *P*=.001.

[c]ICV.st: step-based intradaily coefficient of variation.

[d]R^2=0.167; adjusted R^2=0.137; *F* statistic *P*=.002.

[e]Peaks.hr: number of heart rate peaks per day detected in time-series data.

In model I, both age (β=−0.15, 95% CI −0.26 to −0.03; *P*=.01) and ICV × 100 (β=−0.04, 95% CI −0.07 to −0.01; *P*=.02) were significant predictors of WLM score. In model II, age (β=−0.13, 95% CI −0.25 to −0.02; *P*=.03) and ICV × 100 (β=−0.04, 95% CI −0.07 to 0.00; *P*=.03) remained significant.

Table 3 summarizes the results of the Kruskal-Wallis and Dunn post hoc tests for digital biomarkers across the robust, social prefrailty, and social frailty groups. The Dunn post hoc tests were conducted following the Kruskal-Wallis test to identify pairwise differences, and the *P* values from these tests were adjusted using the Bonferroni method. The table reports only the 17 variables with significant associations identified through the Dunn test (*P*<.05). Notably, comparisons between the robust and social frailty groups revealed significant differences in 13 biomarkers.

XSL•FO
RenderX

Table 3. Analysis of digital biomarkers by social frailty categories using the Kruskal-Wallis and Dunn tests[a].

Variable	Robust (n=28), median (IQR)	Social prefrailty (n=39), median (IQR)	Social frailty (n=19), median (IQR)	Overall *P* value	Robust vs social prefrailty *P* value	Robust vs social frailty *P* value	Social prefrailty vs social frailty *P* value
Steps	7685 (6562)	7488 (4832)	5114 (2782)	.02	.58	.008	.04
IS.st[b]	0.52 (0.21)	0.43 (0.18)	0.38 (0.19)	.03	.15	.01	.28
IV.st[c]	1.22 (0.23)	1.25 (0.41)	1.33 (0.42)	.10	.50	.045	.22
M10.st[d]	577 (406)	527 (342)	383 (261)	.06	.78	.03	.09
L5.st[e]	10 (16)	16 (19)	10 (10)	.07	.20	.63	.04
ICV.st[f]	1.04 (0.24)	1.08 (0.22)	1.24 (0.14)	.001	.61	<.001	.003
AC.st.60m[g]	0.30 (0.17)	0.30 (0.16)	0.18 (0.21)	.07	>.99	.046	.08
AC.st.30m[h]	0.24 (0.17)	0.25 (0.16)	0.14 (0.13)	.05	.66	.02	.09
AC.st.15m[i]	0.18 (0.15)	0.18 (0.16)	0.11 (0.10)	.08	.63	.04	.16
F_pseudo.st[j]	111.30 (169.36)	65.50 (150.43)	45.74 (74.92)	.03	.30	.01	.14
SD of the HR[k] (bpm)	13.28 (2.74)	11.95 (2.79)	11.74 (2.32)	.03	.03	.047	>.99
dRHR[l] (bpm)	10.23 (5.02)	7.31 (3.27)	6.85 (2.86)	.008	.07	.003	.22
α parameter.hr	−0.49 (0.17)	−0.36 (0.19)	−0.36 (0.16)	.04	.02	.11	>.99
F_pseudo.hr[m]	1586.53 (1436.61)	1036.27 (1448.31)	665.80 (1064.46)	.02	.04	.02	.81
UpMesor.hr[n]	6.39 (1.13)	6.39 (1.15)	6.87 (0.95)	.046	.87	.03	.06

[a]Comparison of digital biomarkers related to circadian rhythm and heart rate among the robust, social prefrailty, and social frailty groups using the Kruskal-Wallis and post hoc Dunn tests. Data were collected in a cross-sectional study conducted between October 2023 and December 2024 among community-dwelling older adults in Akita Prefecture, Japan. The table reports the median and IQR for each variable by group. *P* values reflect overall group differences and pairwise comparisons between groups, with significance corrected using the Bonferroni method. Only variables showing at least one significant pairwise difference (*P*<.05) are included in the table. The suffix ".st" denotes rhythm indicators based on steps; ".hr" denotes those based on heart rate.

[b]IS: interdaily stability.

[c]IV: intradaily variability.

[d]M10: most active 10-hour activity level.

[e]L5: least active 5-hour activity level.

[f]ICV: intradaily coefficient of variation.

[g]AC.st.60m: autocorrelation of step counts aggregated in 60-minute intervals.

[h]AC.st.30m: autocorrelation of step counts aggregated in 30-minute intervals.

[i]AC.st.15m: autocorrelation of step counts aggregated in 15-minute intervals.

[j]F_pseudo.st: the improvement in the fit obtained through the nonlinear estimation of the transformed cosine model derived from steps.

[k]HR: heart rate.

[l]dRHR: delta resting heart rate.

[m]F_pseudo.hr: the improvement in the fit obtained through the nonlinear estimation of the transformed cosine model derived from heart rate.

[n]UpMesor: time of transition from rest to activity.

Table 4 summarizes the results of the multiple logistic regression analysis, with models I and III adjusted for age and sex and models II and IV excluding these adjustments. The robust group was used as the reference category.

XSL•FO
RenderX

Table 4. Models according to multiple logistic regression analysis[a].

	β coefficient (SE)	Odds ratio (95% CI)	*P* value
Model I[b]			
Social prefrailty group			
Age (y)	0.01 (0.03)	1.01 (0.95-1.08)	.76
Sex (male=1; female=0)	−0.53 (0.76)	0.59 (0.13-2.58)	.49
dRHR[c]	−0.21 (0.10)	0.82 (0.68-0.99)	.04
α parameter × 100	0.05 (0.02)	1.05 (1.01-1.08)	.006
UpMesor.hr[d]	−0.55 (0.25)	0.58 (0.36-0.93)	.03
F_pseudo.hr[e]	0.00 (0.00)	1.00 (0.99-1.00)	.93
Social frailty group			
Age (y)	0.06 (0.03)	1.06 (0.99-1.13)	.08
Sex (male=1; female=0)	−1.98 (0.38)	0.14 (0.07-0.28)	<.001
dRHR	−0.37 (0.15)	0.69 (0.50-0.95)	.01
α parameter × 100	0.01 (0.02)	1.01 (0.97-1.05)	.61
UpMesor.hr	0.33 (0.30)	1.39 (0.77-2.50)	.27
F_pseudo.hr	0.00 (0.00)	1.00 (0.99-1.00)	.17
Model II[f]			
Social prefrailty group			
dRHR	−0.20 (0.09)	0.82 (0.68-0.97)	.02
α parameter.hr × 100	0.05 (0.01)	1.05 (1.02-1.07)	<.001
UpMesor.hr	−0.56 (0.11)	0.57 (0.46-0.71)	<.001
F_pseudo.hr	−0 (0.00)	1.00 (1.00-1.00)	.97
Social frailty group			
dRHR	−0.31 (0.13)	0.74 (0.58-0.94)	.02
α parameter.hr × 100	0.01 (0.02)	1.01 (0.98-1.04)	.68
UpMesor.hr	0.37 (0.15)	1.45 (1.09-1.94)	.01
F_pseudo.hr	0.00 (0.00)	1.00 (1.00-1.00)	.19
Model III[g]			
Social prefrailty group			
Age (y)	0.01 (0.05)	1.01 (0.92-1.11)	.79
Sex (male=1; female=0)	0.02 (0.05)	1.02 (0.92-1.13)	.76
ICV.st[h] × 100	0.00 (0.02)	1.00 (0.97-1.03)	.96
α parameter.hr × 100	0.05 (0.02)	1.05 (1.01-1.12)	.02
UpMesor.hr	−0.59 (0.31)	0.55 (0.30-1.01)	.06
SD of the HR[i]	−0.29 (0.12)	0.75 (0.59-0.96)	.02
Social frailty group			
Age (y)	0.04 (0.06)	1.04 (0.93-1.16)	.47
Sex (male=1; female=0)	−2.22 (1.65)	0.11 (0.01-1.69)	.18
ICV.st × 100	0.05 (0.02)	1.05 (1.01-1.11)	.01
α parameter.hr × 100	0.01 (0.02)	1.01 (0.96-1.07)	.68
UpMesor.hr	0.44 (0.39)	1.55 (0.72-3.35)	.26

XSL•FO
RenderX

	β coefficient (SE)	Odds ratio (95% CI)	*P* value
SD of the HR	−0.18 (0.15)	0.83 (0.63-1.11)	.20
Model IV[j]			
Social prefrailty group			
ICV.st × 100	0.00 (0.01)	1.00 (0.97-1.03)	.99
α parameter.hr × 100	0.05 (0.02)	1.05 (1.01-1.09)	.02
UpMesor.hr	−0.60 (0.31)	0.55 (0.30-1.01)	.06
SD of the HR	−0.29 (0.12)	0.75 (0.59-0.94)	.02
Social frailty group			
ICV.st × 100	0.04 (0.02)	1.04 (1.01-1.08)	.01
α parameter.hr × 100	0.01 (0.02)	1.01 (0.96-1.06)	.70
UpMesor.hr	0.46 (0.38)	1.58 (0.74-3.32)	.24
SD of the HR	−0.17 (0.14)	0.85 (0.64-1.12)	.25

[a]Multiple logistic regression analyses predicting social prefrailty and social frailty status relative to the robust group among community-dwelling older adults in Akita Prefecture, Japan. Data were collected between October 2023 and December 2024 in a cross-sectional study. This table summarizes the results of 4 models (models I-IV), each using a different combination of predictors related to heart rate variability and rhythm metrics derived from wearable Fitbit devices. The dependent variable was group classification (robust, social prefrailty, or social frailty), with the robust group used as the reference category. Independent variables included delta resting heart rate, α parameter (scaled by 100), rest-activity transition timing, step-based intradaily variability, model fit from the extended cosinor analysis, and SD of the heart rate. The suffix ".st" denotes rhythm indicators based on steps; ".hr" denotes those based on heart rate. Model fit was evaluated using the likelihood-ratio test, the Nagelkerke R^2, and the Akaike information criterion (AIC). Model IV demonstrated the best fit, indicated by the lowest AIC (170.1) and a relatively high R^2 (0.351).

[b]Likelihood-ratio test *P* value=.002; Nagelkerke R^2=0.343; AIC=179.0.

[c]dRHR: delta resting heart rate.

[d]UpMesor: time of transition from rest to activity.

[e]F_pseudo.hr: the improvement in the fit obtained through the nonlinear estimation of the transformed cosine model derived from heart rate.

[f]Likelihood-ratio test *P*<.001; Nagelkerke R^2=0.315; Akaike information criterion=173.9.

[g]Likelihood-ratio test *P*=.001; Nagelkerke R^2=0.378; Akaike information criterion=175.2.

[h]ICV: intradaily coefficient of variation.

[i]HR: heart rate.

[j]Likelihood-ratio test *P*<.001; Nagelkerke R^2=0.351; Akaike information criterion=170.1.

In model II, the dRHR showed significant negative associations with both social prefrailty (OR 0.82, 95% CI 0.68-0.97; *P*=.02) and social frailty (OR 0.74, 95% CI 0.58-0.94; *P*=.02). Similarly, UpMesor.hr demonstrated significant associations with both social prefrailty (OR 0.57, 95% CI 0.46-0.71; *P*<.001) and social frailty (OR 1.45, 95% CI 1.09-1.94; *P*=.01).

In model IV, the α parameter.hr × 100 was significantly associated with social prefrailty (OR 1.05, 95% CI 1.01-1.09; *P*=.02). The SD of the heart rate showed a significant negative association with social prefrailty (OR 0.75, 95% CI 0.59-0.94; *P*=.02). For social frailty, ICV.st × 100 demonstrated a significant positive association (OR 1.04, 95% CI 1.01-1.08; *P*=.01).

The results of models I and III, which were adjusted for age and sex, were generally consistent with their corresponding models (model I corresponds to models II, and model III corresponds to model IV). The primary difference was observed in UpMesor.hr, which was significantly associated with both social prefrailty and social frailty in model II but only with social prefrailty in model I. Regarding other variables, dRHR, ICV.st × 100, the α parameter.hr × 100, and the SD of the heart rate showed consistent associations across corresponding models.

Discussion

Principal Findings

Our results lent support to the hypothesis that the frailty group would exhibit significant associations with RAR parameters (nonparametric and extended cosinor analysis) and heart rate–based parameters. The observed significant relationships with dRHR, UpMesor, and other heart rate–derived metrics validate their relevance in understanding social frailty. In addition, Fitbit-derived RAR metrics were found to be associated with cognitive function, particularly memory decline, in partial alignment with the proposed hypothesis.

The results of the analyses using the multinomial logistic regression models partially supported the primary hypothesis. Among the step-based rhythm metrics, only the ICV was significantly associated with social prefrailty. On the other hand, the dRHR and UpMesor showed significant associations with both social frailty and prefrailty. Furthermore, the α parameter and the SD of the heart rate were significant predictors of social prefrailty. Notably, except for the ICV, all the significant predictors of social frailty were derived from heart rate data.

XSL•FO
RenderX

First, the dRHR showed significant negative associations with both social prefrailty (OR 0.82, 95% CI 0.68-0.97; *P*=.02) and social frailty (OR 0.74, 95% CI 0.58-0.94; *P*=.02) in model II. The dRHR reflects the difference between the daily average heart rate and the RHR, which may provide insights into the activity levels and physiological responses to rest and activity. In this study, we found that the dRHR values were smaller in the social frailty and prefrailty groups, potentially indicating reduced daytime activity levels and a closer approximation of the average heart rate to the RHR.

Figure 2 illustrates our finding of lower daytime heart rates and higher nocturnal heart rates in the social frailty group. This pattern corresponds to smaller dRHR values, suggesting a possible disruption in the typical fluctuations between rest and activity. Notably, insufficient nocturnal heart rate reduction in persons with social frailty might contribute to an elevated RHR. In addition, increased physical activity has been associated with reduced RHRs, likely mediated by autonomic nervous system regulation [34]. Therefore, lower dRHR values might reflect both reduced daytime activity and limited heart rate reduction during rest. In this context, dRHR can be interpreted as a metric associated with variations in heart rate across periods of activity and rest.

Studies on the association between the severity of depression and heart rate variation have suggested that reduced activity levels can lead to altered heart rate fluctuation patterns [35]. Specifically, a higher nocturnal resting mean heart rate has been associated with more severe depressive symptoms (β=0.09; *P*<.001), whereas greater daytime heart rate variation (β=−0.34; *P*<.001) and a higher SD of daytime heart rate (β=−0.18; *P*<.001) have been linked to a lower severity of depression [35].

Similarly, previous research has identified a high RHR as a risk factor for dementia [36], worsening of depression [35], noncardiac disability [37], and physical frailty [38]. Social frailty has also been reported to predict disability [24], cognitive decline [39], depression [40], and physical frailty [6,39,41]. On the basis of these findings, dRHR may function as a comprehensive indicator reflecting these interconnected factors.

However, research on the dRHR is limited. Only the studies by Rykov et al [16,17] have examined dRHR in patients with heart disease and depression. Our study is the first to report on dRHR in the context of social frailty and has highlighted its relevance and potential value for understanding social frailty.

In addition, another heart rate indicator, the SD of the heart rate, was identified as a predictor of social prefrailty (OR 0.75, 95% CI 0.59-0.94; *P*=.02; model IV). Similarly to dRHR, the SD of the heart rate can be interpreted as an indicator of the heart rate fluctuations associated with daily physical activity levels. It is also classified as a measure of the HRV, a widely recognized method for assessing autonomic nervous system function. The SD of the heart rate has been classified as a time-domain HRV metric in previous research [42]. While the SD of the heart rate has been less frequently studied compared with other HRV metrics such as the RMSSD and SD of the NN interval, its potential relevance in the context of social frailty is noteworthy. In this study, we also calculated the RMSSD, but no significant differences were observed between groups. While HRV is a well-established metric for assessing autonomic nervous system function, its specific significance in social frailty remains underexplored. Future research should incorporate not only time-domain measures such as RMSSD, which primarily reflect parasympathetic nervous system activity, but also frequency-domain metrics, including low- and high-frequency components and their ratio (LF/HF), which reflect the balance between sympathetic and parasympathetic activity [43]. This may lead to a deeper understanding of the role of HRV in the context of social frailty.

Second, this study found that the UpMesor was delayed in the social frailty group (OR 1.45, 95% CI 1.09-1.94; *P*=.01; model II). Although previous studies have not used heart rate–based UpMesor, our results align with reports from previous studies that delayed activity onset is associated with depressive symptoms and functional decline [44]. Given the association between social frailty and depression [45-47], delayed activity rhythms may represent a characteristic feature of social frailty.

Conversely, the UpMesor was advanced in the social prefrailty group (OR 0.57, 95% CI 0.46-0.71; *P*<.001; model II). Typically, social frailty progresses linearly from robust to prefrailty and then to frailty, suggesting that UpMesor would also be progressively delayed across these stages. However, our findings present an apparent contradiction.

According to the report by Ji et al [48], UpMesor, DownMesor, and acrophase are parameters reflecting temporal rhythms, including the onset and cessation of nighttime sleep and peak daytime activity levels. The observed opposing patterns of UpMesor suggest the importance of understanding how specific temporal rhythm elements impact health across the stages of social frailty.

Previous research has reported on acrophase, revealing complex associations. For example, delayed acrophase has been linked to developing dementia or mild cognitive impairment [49], whereas early acrophase has been associated with cognitive decline [50,51]. These studies indicate that temporal rhythms do not exhibit simple linear relationships with health; both excessively delayed and advanced rhythms may negatively impact health.

The observed delay in UpMesor in the social frailty group and its advancement in the social prefrailty group may support this nonlinear association. Specifically, early activity onset in the prefrailty stage and delayed activity onset in the frailty stage may each contribute to health risks. This finding underscores the need for further research to examine how factors such as sex, age, and physiological characteristics might influence these associations.

Third, the α parameter × 100 was significantly associated with social prefrailty (OR 1.05, 95% CI 1.01-1.09; *P*=.02; model IV), suggesting that an increase in α may heighten the risk of social prefrailty. High α values reflect shorter activity periods and longer rest periods, corresponding to insufficient daytime activity. This aligns with the findings of Smagula et al [44], which linked high α values to increased health risks such as depression.

XSL•FO
RenderX

However, α may also indicate rhythm clarity, as suggested by Ji et al [48] using the sleep regularity, satisfaction, alertness, timing, efficiency, and duration model [48]. Therefore, α must be interpreted cautiously as it may reflect both the quantity and quality of daytime activity and nighttime rest.

Fourth, the ICV × 100 was significantly associated with social frailty (OR 1.04, 95% CI 1.01-1.08; *P*=.01; model IV). ICV reflects the intrahourly CV, with higher values indicating greater variability in activity levels. In this study, individuals with social frailty exhibited more frequent alternations between short activity bursts and rest, consistent with the report that reduced physical activity and limited social engagement are associated with the risk of frailty [6]. These results suggest that intradaily rhythm instability may be a key feature of social frailty.

In addition, we found a significant association between ICV and WLM score (β=−0.04; *P*=.02), suggesting that higher rhythm variability may be linked to lower cognitive performance. This finding aligns with those of previous research indicating that increased IV and decreased IS are associated with cognitive decline [52,53]. These results may imply that the ICV metric, which reflects short-term rhythm fragmentation, is related to cognitive function. However, as Fitbit does not record activity counts, relying solely on step-based data may lead to underestimation of activity levels in individuals with impaired mobility.

Strengths and Limitations

One of the significant strengths of this study is its use of Fitbit data to elucidate new rhythm-related characteristics associated with social frailty. Fitbit devices have demonstrated acceptable precision comparable to that of research-grade devices and are also affordable and accessible, making them promising tools for the development of digital biomarkers. In this study, step count and heart rate data recorded by Fitbit were used to replicate rhythm metrics based on the scripts by Rykov et al [17], and new rhythm metrics were also calculated. We demonstrated in this study that this comprehensive approach offers the potential for evaluating rhythm-related characteristics in social frailty.

Furthermore, the combined use of nonparametric metrics and extended cosinor metrics has shown that multifaceted evaluation of rhythm characteristics is achievable. Nonparametric metrics are well suited for capturing the daily regularity of rhythms, whereas extended cosinor metrics are useful for analyzing phase changes along the time axis. This complementary use uncovered new aspects of social frailty that were not previously identifiable through conventional methods.

In addition, extended cosinor metrics, which have traditionally been calculated based on activity counts, were applied to heart rate data in this study, revealing new possibilities for rhythm analysis. This development suggests that rhythm evaluations can be conducted even for individuals with reduced walking ability or limited daytime movement. The use of heart rate–based data expands the scope of analysis to groups that have been challenging to evaluate previously, providing a more comprehensive understanding of the rhythm characteristics.

Despite its strengths, this study had several limitations. First, the sample size posed a constraint. Due to the limited sample size, the findings should be interpreted as observational results. For example, determining the “optimal range” for UpMesor requires validation in larger sample sizes. Furthermore, the small sample size restricted our ability to conduct comprehensive analyses integrating nonlinear models or multiple health outcomes.

Second, sex differences may have influenced the results. The potential impact of sex on the outcomes necessitates cautious interpretation of the regression model estimates. Future research should incorporate sex considerations to provide more nuanced insights.

Third, there is the issue of uncertainty surrounding the standardization of social frailty indexes. According to a meta-analysis conducted by Jia et al [54], lack of a standardized social frailty index and reliance on operational choices by individual researchers remain significant challenges. This calls for further discussion and validation of the reliability of social frailty indexes.

Fourth, although a significant association between rhythm metrics and cognitive function (WLM score) was observed, this analysis was exploratory and secondary in nature. As the primary objective of this study was to model social frailty, these findings should be interpreted as preliminary. Advanced modeling techniques such as machine learning were not used in this study due to the limited sample size and the descriptive aim of the cognitive analysis. Future studies with larger samples should incorporate machine learning to develop predictive models for cognitive outcomes and validate the utility of rhythm-based digital biomarkers.

In addition, the interpretation of extended cosinor metrics presents a challenge. The current body of research on extended cosinor metrics remains limited, and careful consideration tailored to specific objectives and populations is required. Unlike nonparametric metrics, cosinor metrics do not allow for simple binary evaluations such as “stable” or “unstable.” Both excessive stability and instability of rhythms may be associated with health risks, which necessitates a comprehensive approach to rhythm evaluation.

Finally, there are limitations related to Fitbit data. While Fitbit and similar smartwatches provide minute-by-minute step data through their application programming interfaces, they do not provide activity count data. This lack of activity count data raises questions about the validity of rhythm metrics derived solely from step data, highlighting the need for future research to address this issue.

Conclusions

This study revealed associations between novel rhythm metrics derived from step count and heart rate data recorded via Fitbit devices and social frailty. These findings suggest that digital biomarkers for evaluating social frailty and its risk factors can be derived from data provided by consumer-grade wearable devices, which are affordable and accessible.

XSL•FO
RenderX

In particular, it was shown that heart rate–based metrics such as dRHR, UpMesor, and the α parameter may play a crucial role in the assessment of social frailty and prefrailty. These metrics have the potential to complement existing activity count–based rhythm analyses by capturing new dimensions of health that were previously overlooked.

Furthermore, the findings of our study suggest that social frailty is associated with specific rhythm characteristics such as decreased daytime activity levels and increased nighttime heart rates. These findings indicate that rhythm metrics may be valuable for early detection and assessment of social frailty.

This research provides preliminary insights into the evaluation of social frailty using Fitbit data. However, limitations such as the small sample size and uncertainty surrounding standardized indexes for social frailty must be addressed. Future studies should focus on validating these findings in larger samples and conducting comparative research using other devices. In addition, further exploration is needed to determine whether heart rate–based metrics can be applied to assess other health conditions and diseases.

Acknowledgments

The authors wish to thank all the participants, as well as the staff and volunteers from each care management center and temple. They also thank the Advanced Research Center for Geriatric and Gerontology, Akita University, for providing the Fitbit devices. Portions of this manuscript were edited for grammar and clarity using ChatGPT (OpenAI) [55]. The final content was reviewed and approved by all authors, who take full responsibility for its accuracy and integrity. This study was also carried out as part of the doctoral degree work of HM. This work was supported by the Japan Society for the Promotion of Science Grant-in-Aid for Scientific Research JP23K12632.

Data Availability

The datasets generated during the current study are not publicly available due to ethical restrictions but may be available from the corresponding author upon reasonable request and approval by the institutional review board.

Conflicts of Interest

None declared.

Multimedia Appendix 1
R-code scripts for data processing.
[ZIP File (Zip Archive), 7 KB - formative_v9i1e71393_app1.zip]

Multimedia Appendix 2
Detailed statistical analysis tables.
[XLSX File (Microsoft Excel File), 141 KB - formative_v9i1e71393_app2.xlsx]

References

1. Gobbens RJ, van Assen MA, Luijkx KG, Wijnen-Sponselee MT, Schols JM. Determinants of frailty. J Am Med Dir Assoc 2010 Jun;11(5):356-364. [doi: 10.1016/j.jamda.2009.11.008] [Medline: 20511103]
2. Pilotto A, Custodero C, Maggi S, Polidori MC, Veronese N, Ferrucci L. A multidimensional approach to frailty in older people. Ageing Res Rev 2020 Jul;60:101047 [FREE Full text] [doi: 10.1016/j.arr.2020.101047] [Medline: 32171786]
3. Zhang XM, Cao S, Gao M, Xiao S, Xie X, Wu X. The prevalence of social frailty among older adults: a systematic review and meta-analysis. J Am Med Dir Assoc 2023 Jan;24(1):29-37.e9 [FREE Full text] [doi: 10.1016/j.jamda.2022.10.007] [Medline: 36402197]
4. Goto T, Kishimoto T, Fujiwara S, Shirayama Y, Ichikawa T. Social frailty as a predictor of all-cause mortality and functional disability: a systematic review and meta-analysis. Sci Rep 2024 Feb 10;14(1):3410 [FREE Full text] [doi: 10.1038/s41598-024-53984-3] [Medline: 38341512]
5. Bunt S, Steverink N, Olthof J, van der Schans CP, Hobbelen JS. Social frailty in older adults: a scoping review. Eur J Ageing 2017 Sep 31;14(3):323-334 [FREE Full text] [doi: 10.1007/s10433-017-0414-7] [Medline: 28936141]
6. Pek K, Chew J, Lim JP, Yew S, Tan CN, Yeo A, et al. Social frailty is independently associated with mood, nutrition, physical performance, and physical activity: insights from a theory-guided approach. Int J Environ Res Public Health 2020 Jun 14;17(12):4239 [FREE Full text] [doi: 10.3390/ijerph17124239] [Medline: 32545853]
7. Park H, Jang IY, Lee HY, Jung HW, Lee E, Kim DH. Screening value of social frailty and its association with physical frailty and disability in community-dwelling older Koreans: aging study of Pyeongchang rural area. Int J Environ Res Public Health 2019 Aug 07;16(16):2809 [FREE Full text] [doi: 10.3390/ijerph16162809] [Medline: 31394719]

XSL•FO
RenderX

8. Teo N, Gao Q, Nyunt MS, Wee SL, Ng T. Social frailty and functional disability: findings from the Singapore longitudinal ageing studies. J Am Med Dir Assoc 2017 Jul 01;18(7):637.e13-637.e19. [doi: 10.1016/j.jamda.2017.04.015] [Medline: 28648903]
9. Ong M, Pek K, Tan CN, Chew J, Lim JP, Yew S, et al. Social frailty and executive function: association with geriatric syndromes, life space and quality of life in healthy community-dwelling older adults. J Frailty Aging 2022 Mar;11(2):206-213 [FREE Full text] [doi: 10.14283/jfa.2021.43] [Medline: 35441199]
10. Cai R, Gao L, Gao C, Yu L, Zheng X, Bennett DA, et al. Circadian disturbances and frailty risk in older adults. Nat Commun 2023 Nov 16;14(1):7219 [FREE Full text] [doi: 10.1038/s41467-023-42727-z] [Medline: 37973796]
11. Maekawa H, Kume Y. Imbalance of nonparametric rest-activity rhythm and the evening-type of chronotype according to frailty indicators in elderly community dwellers. Chronobiol Int 2019 Sep;36(9):1208-1216. [doi: 10.1080/07420528.2019.1626416] [Medline: 31234663]
12. Musiek ES, Bhimasani M, Zangrilli MA, Morris JC, Holtzman DM, Ju YS. Circadian rest-activity pattern changes in aging and preclinical Alzheimer disease. JAMA Neurol 2018 May 01;75(5):582-590 [FREE Full text] [doi: 10.1001/jamaneurol.2017.4719] [Medline: 29379963]
13. Targa AD, Benítez ID, Dakterzada F, Fontenele-Araujo J, Minguez O, Zetterberg H, et al. The circadian rest-activity pattern predicts cognitive decline among mild-moderate Alzheimer's disease patients. Alzheimers Res Ther 2021 Sep 25;13(1):161 [FREE Full text] [doi: 10.1186/s13195-021-00903-7] [Medline: 34563258]
14. Hayashi M, Takeshima M, Hosoya T, Kume Y. 24-hour rest-activity rhythm in middle-aged and older persons with depression. Int J Environ Res Public Health 2023 Mar 27;20(7):5275 [FREE Full text] [doi: 10.3390/ijerph20075275] [Medline: 37047891]
15. Kume Y, Kodama A, Arai S, Nagaoka M, Sato A, Saito A, et al. Improvement of social frailty is associated with stability of nonparametric characteristics of the rest-activity rhythm and improvement of the usual walking ability in the elderly. Chronobiol Int 2024 Sep;41(9):1239-1247. [doi: 10.1080/07420528.2024.2393877] [Medline: 39164845]
16. Rykov Y, Thach TQ, Dunleavy G, Roberts AC, Christopoulos G, Soh C, et al. Activity tracker-based metrics as digital markers of cardiometabolic health in working adults: cross-sectional study. JMIR Mhealth Uhealth 2020 Jan 31;8(1):e16409 [FREE Full text] [doi: 10.2196/16409] [Medline: 32012098]
17. Rykov Y, Thach TQ, Bojic I, Christopoulos G, Car J. Digital biomarkers for depression screening with wearable devices: cross-sectional study with machine learning modeling. JMIR Mhealth Uhealth 2021 Oct 25;9(10):e24872 [FREE Full text] [doi: 10.2196/24872] [Medline: 34694233]
18. Henriksen A, Haugen Mikalsen M, Woldaregay AZ, Muzny M, Hartvigsen G, Hopstock LA, et al. Using fitness trackers and smartwatches to measure physical activity in research: analysis of consumer wrist-worn wearables. J Med Internet Res 2018 Mar 22;20(3):e110 [FREE Full text] [doi: 10.2196/jmir.9157] [Medline: 29567635]
19. Fuller D, Colwell E, Low J, Orychock K, Tobin MA, Simango B, et al. Reliability and validity of commercially available wearable devices for measuring steps, energy expenditure, and heart rate: systematic review. JMIR Mhealth Uhealth 2020 Sep 08;8(9):e18694 [FREE Full text] [doi: 10.2196/18694] [Medline: 32897239]
20. Chevance G, Golaszewski NM, Tipton E, Hekler EB, Buman M, Welk GJ, et al. Accuracy and precision of energy expenditure, heart rate, and steps measured by combined-sensing Fitbit against reference measures: systematic review and meta-analysis. JMIR Mhealth Uhealth 2022 Apr 13;10(4):e35626 [FREE Full text] [doi: 10.2196/35626] [Medline: 35416777]
21. Faul F, Erdfelder E, Buchner A, Lang AG. Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. Behav Res Methods 2009 Nov;41(4):1149-1160. [doi: 10.3758/BRM.41.4.1149] [Medline: 19897823]
22. Feehan LM, Geldman J, Sayre EC, Park C, Ezzat AM, Yoo JY, et al. Accuracy of Fitbit devices: systematic review and narrative syntheses of quantitative data. JMIR Mhealth Uhealth 2018 Aug 09;6(8):e10527 [FREE Full text] [doi: 10.2196/10527] [Medline: 30093371]
23. Puri A, Kim B, Nguyen O, Stolee P, Tung J, Lee J. User acceptance of wrist-worn activity trackers among community-dwelling older adults: mixed method study. JMIR Mhealth Uhealth 2017 Nov 15;5(11):e173 [FREE Full text] [doi: 10.2196/mhealth.8211] [Medline: 29141837]
24. Makizako H, Shimada H, Tsutsumimoto K, Lee S, Doi T, Nakakubo S, et al. Social frailty in community-dwelling older adults as a risk factor for disability. J Am Med Dir Assoc 2015 Nov 01;16(11):1003.e7-1003.11 [FREE Full text] [doi: 10.1016/j.jamda.2015.08.023] [Medline: 26482055]
25. Makizako H, Shimada H, Park H, Doi T, Yoshida D, Uemura K, et al. Evaluation of multidimensional neurocognitive function using a tablet personal computer: test-retest reliability and validity in community-dwelling older adults. Geriatr Gerontol Int 2013 Oct 11;13(4):860-866 [FREE Full text] [doi: 10.1111/ggi.12014] [Medline: 23230988]
26. Fitbit developer. Fitbit. URL: https://dev.fitbit.com/ [accessed 2025-01-17]
27. Danilevicz IM, Vidil S, Landré B, Dugravot A, van Hees VT, Sabia S. Reliable measures of rest-activity rhythm fragmentation: how many days are needed? Eur Rev Aging Phys Act 2024 Oct 19;21(1):29 [FREE Full text] [doi: 10.1186/s11556-024-00364-5] [Medline: 39427121]
28. Di J, Spira A, Bai J, Urbanek J, Leroux A, Wu M, et al. Joint and individual representation of domains of physical activity, sleep, and circadian rhythmicity. Stat Biosci 2019 Jul 15;11(2):371-402 [FREE Full text] [doi: 10.1007/s12561-019-09236-4] [Medline: 32440309]

XSL•FO
RenderX

29. Marler MR, Gehrman P, Martin JL, Ancoli-Israel S. The sigmoidally transformed cosine curve: a mathematical model for circadian rhythms with symmetric non-sinusoidal shapes. Stat Med 2006 Nov 30;25(22):3893-3904. [doi: 10.1002/sim.2466] [Medline: 16381069]
30. Van Someren EJ, Swaab DF, Colenda CC, Cohen W, McCall WV, Rosenquist PB. Bright light therapy: improved sensitivity to its effects on rest-activity rhythms in Alzheimer patients by application of nonparametric methods. Chronobiol Int 1999 Jul 07;16(4):505-518. [doi: 10.3109/07420529908998724] [Medline: 10442243]
31. Witting W, Kwa IH, Eikelenboom P, Mirmiran M, Swaab DF. Alterations in the circadian rest-activity rhythm in aging and Alzheimer's disease. Biol Psychiatry 1990 Mar 15;27(6):563-572. [doi: 10.1016/0006-3223(90)90523-5] [Medline: 2322616]
32. van Brakel JP. Peak signal detection in realtime timeseries data: robust peak detection algorithm using z-scores. Stack Overflow. URL: https://stackoverflow.com/questions/22583391/peak-signal-detection-in-realtime-timeseries-data [accessed 2025-01-15]
33. Package ActCR. CRAN. URL: https://cran.r-project.org/web/packages/ActCR/index.html [accessed 2025-01-16]
34. Saxena A, Minton D, Lee D, Sui X, Fayad R, Lavie CJ, et al. Protective role of resting heart rate on all-cause and cardiovascular disease mortality. Mayo Clin Proc 2013 Dec;88(12):1420-1426 [FREE Full text] [doi: 10.1016/j.mayocp.2013.09.011] [Medline: 24290115]
35. Siddi S, Bailon R, Giné-Vázquez I, Matcham F, Lamers F, Kontaxis S, RADAR-CNS consortium. The usability of daytime and night-time heart rate dynamics as digital biomarkers of depression severity. Psychol Med 2023 Jun;53(8):3249-3260. [doi: 10.1017/S0033291723001034] [Medline: 37184076]
36. Imahori Y, Vetrano DL, Xia X, Grande G, Ljungman P, Fratiglioni L, et al. Association of resting heart rate with cognitive decline and dementia in older adults: a population-based cohort study. Alzheimers Dement 2022 Oct;18(10):1779-1787. [doi: 10.1002/alz.12495] [Medline: 34859936]
37. Ogliari G, Mahinrad S, Stott DJ, Jukema JW, Mooijaart SP, Macfarlane PW, et al. Resting heart rate, heart rate variability and functional decline in old age. CMAJ 2015 Oct 20;187(15):E442-E449 [FREE Full text] [doi: 10.1503/cmaj.150462] [Medline: 26323697]
38. Weinstein G, Lutski M, Goldbourt U, Tanne D. Midlife resting heart rate, but not its visit-to-visit variability, is associated with late-life frailty status in men with coronary heart disease. Aging Male 2020 Dec;23(5):1052-1058 [FREE Full text] [doi: 10.1080/13685538.2019.1664456] [Medline: 31524042]
39. Tsutsumimoto K, Doi T, Makizako H, Hotta R, Nakakubo S, Makino K, et al. Association of social frailty with both cognitive and physical deficits among older people. J Am Med Dir Assoc 2017 Jul 01;18(7):603-607 [FREE Full text] [doi: 10.1016/j.jamda.2017.02.004] [Medline: 28411094]
40. Ma L, Sun F, Tang Z. Social frailty is associated with physical functioning, cognition, and depression, and predicts mortality. J Nutr Health Aging 2018;22(8):989-995 [FREE Full text] [doi: 10.1007/s12603-018-1054-0] [Medline: 30272104]
41. Makizako H, Shimada H, Doi T, Tsutsumimoto K, Hotta R, Nakakubo S, et al. Social frailty leads to the development of physical frailty among physically non-frail adults: a four-year follow-up longitudinal cohort study. Int J Environ Res Public Health 2018 Mar 10;15(3):3002-3013 [FREE Full text] [doi: 10.3390/ijerph15030490] [Medline: 29534470]
42. Nagaraj SB, Ramaswamy SM, Biswal S, Boyle EJ, Zhou DW, Mcclain LM, et al. Heart rate variability as a biomarker for sedation depth estimation in ICU patients. Annu Int Conf IEEE Eng Med Biol Soc 2016 Aug;2016:6397-6400 [FREE Full text] [doi: 10.1109/EMBC.2016.7592192] [Medline: 28269712]
43. Akselrod S, Gordon D, Ubel FA, Shannon DC, Berger AC, Cohen RJ. Power spectrum analysis of heart rate fluctuation: a quantitative probe of beat-to-beat cardiovascular control. Science 1981 Jul 10;213(4504):220-222. [doi: 10.1126/science.6166045] [Medline: 6166045]
44. Smagula SF, Krafty RT, Taylor BJ, Martire LM, Schulz R, Hall MH. Rest-activity rhythm and sleep characteristics associated with depression symptom severity in strained dementia caregivers. J Sleep Res 2017 Dec;26(6):718-725 [FREE Full text] [doi: 10.1111/jsr.12549] [Medline: 28488270]
45. Kume Y, Kodama A, Takahashi T, Lee S, Makizako H, Ono T, et al. Social frailty is independently associated with geriatric depression among older adults living in northern Japan: a cross-sectional study of ORANGE registry. Geriatr Gerontol Int 2022 Feb;22(2):145-151. [doi: 10.1111/ggi.14330] [Medline: 34935256]
46. Wada A, Makizako H, Nakai Y, Tomioka K, Taniguchi Y, Sato N, et al. Association between cognitive frailty and higher-level competence among community-dwelling older adults. Arch Gerontol Geriatr 2022 Mar;99:104589. [doi: 10.1016/j.archger.2021.104589] [Medline: 34896796]
47. Demichelis OP, Grainger SA, Hubbard RE, Henry JD. Emotion regulation mediates the relationship between social frailty and stress, anxiety, and depression. Sci Rep 2023 Apr 20;13(1):6430 [FREE Full text] [doi: 10.1038/s41598-023-33749-0] [Medline: 37081078]
48. Ji L, Wallace ML, Master L, Schade MM, Shen Y, Derby CA, et al. Six multidimensional sleep health facets in older adults identified with factor analysis of actigraphy: results from the Einstein Aging Study. Sleep Health 2023 Oct;9(5):758-766 [FREE Full text] [doi: 10.1016/j.sleh.2023.03.002] [Medline: 37246064]

XSL•FO
RenderX

49. Tranah GJ, Blackwell T, Stone KL, Ancoli-Israel S, Paudel ML, Ensrud KE, et al. Circadian activity rhythms and risk of incident dementia and mild cognitive impairment in older women. Ann Neurol 2011 Nov;70(5):722-732 [FREE Full text] [doi: 10.1002/ana.22468] [Medline: 22162057]
50. Rogers-Soeder TS, Blackwell T, Yaffe K, Ancoli-Israel S, Redline S, Cauley JA, Osteoporotic Fractures in Men Study Research Group. Rest-activity rhythms and cognitive decline in older men: the osteoporotic fractures in men sleep study. J Am Geriatr Soc 2018 Nov;66(11):2136-2143 [FREE Full text] [doi: 10.1111/jgs.15555] [Medline: 30136716]
51. Jeon SY, Byun MS, Yi D, Jung G, Lee J, Kim YK, KBASE Research Group. Circadian rest-activity rhythm and longitudinal brain changes underlying late-life cognitive decline. Psychiatry Clin Neurosci 2023 Apr;77(4):205-212 [FREE Full text] [doi: 10.1111/pcn.13521] [Medline: 36527292]
52. Winer JR, Lok R, Weed L, He Z, Poston KL, Mormino EC, et al. Impaired 24-h activity patterns are associated with an increased risk of Alzheimer's disease, Parkinson's disease, and cognitive decline. Alzheimers Res Ther 2024 Feb 14;16(1):35 [FREE Full text] [doi: 10.1186/s13195-024-01411-0] [Medline: 38355598]
53. Xiao Q, Sampson JN, LaCroix AZ, Shadyab AH, Zeitzer JM, Ancoli-Israel S, Osteoporotic Fractures in Men (MrOS) Study Group. Nonparametric parameters of 24-hour rest-activity rhythms and long-term cognitive decline and incident cognitive impairment in older men. J Gerontol A Biol Sci Med Sci 2022 Feb 03;77(2):250-258 [FREE Full text] [doi: 10.1093/gerona/glab275] [Medline: 34558603]
54. Jia B, Wang Z, Zhang T, Yue X, Zhang S. Prevalence of social frailty and risk factors among community-dwelling older adults: a systematic review and meta-analysis. Arch Gerontol Geriatr 2024 Aug;123:105419. [doi: 10.1016/j.archger.2024.105419] [Medline: 38522381]
55. GPT-4. OpenAI. URL: https://openai.com/index/gpt-4-research/ [accessed 2025-05-21]

Abbreviations

CV: coefficient of variation
DownMesor: time of transition from activity to rest
dRHR: delta resting heart rate
HRV: heart rate variability
ICV: intradaily coefficient of variation
IS: interdaily stability
IV: intradaily variability
L5: least active 5-hour activity level
M10: most active 10-hour activity level
MESOR: midline estimating statistic of rhythm
OR: odds ratio
RA: relative amplitude
RAR: rest-activity rhythm
RHR: resting heart rate
RMSSD: root mean square of successive differences
UpMesor: time of transition from rest to activity
WLM: word list memory

Edited by A Mavragani; submitted 08.02.25; peer-reviewed by V Gupta, K Esquivel; comments to author 02.05.25; revised version received 05.05.25; accepted 25.05.25; published 24.07.25.

Please cite as:
Maekawa H, Kume Y
Predicting Social Frailty in Older Adults Using Fitbit-Derived Circadian and Heart Rate Biomarkers: Cross-Sectional Study
JMIR Form Res 2025;9:e71393
URL: https://formative.jmir.org/2025/1/e71393
doi:10.2196/71393
PMID:

©Hiroki Maekawa, Yu Kume. Originally published in JMIR Formative Research (https://formative.jmir.org), 24.07.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Digital Tool for Assessing Well-Being at the Workplace and in Personal Life: Development and Validation of the Quan Well-Being Index

Georgia A Floridou[1], BSc, MSc, PhD; Freya Katre[2], BA, MSc; Emile Jeuken[1], BSc, MSc, PhD

[1]Quan, The Hague, The Netherlands

[2]King's Business School, King's College London, London, United Kingdom

Corresponding Author:
Georgia A Floridou, BSc, MSc, PhD
Quan
Wilhelmina van Pruisenweg 35, 2595 AN
The Hague
The Netherlands
Email: georgina.floridou@quanwellbeing.com

Abstract

Background: Quan is a workplace well-being digital platform that supports employees, teams, and organizations in measuring, understanding, and improving their well-being. It is important to develop a validated measurement instrument that enables users to assess and track their well-being over time. Currently, no digital measurement instrument comprehensively evaluates well-being across both personal and professional domains.

Objective: We detail the development and validation of the Quan Well-being Index, a new digital self-report measure for assessing well-being in personal life and at the workplace.

Methods: We performed 3 studies. The first study involved the conceptualization of 6 initial factors, the generation of 51 items, and the steps of face and content validity. In the second study, revised items were presented to a UK sample. In the third study, an independent UK sample completed the final assessment along with a battery of well-being and personality questionnaires. A subsample of participants from the third study retook the assessment approximately 2 weeks after initial completion.

Results: In the first study, after face and content validity processes, the number of items was reduced to 45. In the second study, exploratory factor analysis on data from 1020 participants (age: mean 43.06, SD 12.98 years; 525 female participants) identified a 4-factor solution with 35 items (Kaiser-Meyer-Olkin value=0.98; Bartlett test: χ^2_{990}=37063.54; P<.001), accounting for 64% of variance. The 4 factors were *thrive and connect in personal life*, *thrive and connect at work*, *mental health*, and *physical health*. In the third study, confirmatory factor analysis on data from 966 participants (age: mean 44.4, SD 12.52 years; 480 female participants) tested 4 structural models. A hierarchical model (model 1) where the general factor influenced the 4 group factors demonstrated the best fit (χ^2_{521}=3467.00; Bentler comparative fit index=0.906; Tucker-Lewis index=0.892; root mean square error of approximation=0.077; standardized root mean square residual=0.048; ΔAkaike information criterion=0.0; ΔBayesian information criterion=0.0). Internal reliability was high across subscales (Cronbach α=.88-.93; McDonald ω total=0.89-0.94; Guttman λ6=0.86-0.92). Convergent validity was demonstrated by strong correlations with the Warwick-Edinburgh Mental Well-being Scale (r=0.45-0.85; P<.001) and Flourishing-at-Work Scale (r=0.80-0.84; P<.001). Divergent validity was supported through weak or negative correlations with Big Five Personality Inventory traits (eg, neuroticism: r=–0.29; P<.001). Test-retest reliability assessed in a subset of 275 participants (age: mean 52.12, SD 9.56 years; 170 female participants) over a 2-week interval was strong to very strong across factors (r=0.74-0.81; P<.001).

Conclusions: The Quan Well-being Index provides a comprehensive assessment of well-being at the workplace and in personal life, and is anticipated to be a valuable digital tool, enabling individuals, teams, and organizations to gain insights, monitor progress, and implement appropriate interventions for a healthier workforce.

(JMIR Form Res 2025;9:e73713) doi:10.2196/73713

KEYWORDS

workplace well-being; personal life well-being; index development; digital health

XSL•FO
RenderX

Introduction

Background

Well-being has become a pressing concern in the modern workplace, with stress, burnout, and other issues on the rise. Evidence suggests that well-being issues can result in reduced productivity, increased absenteeism, and medical expenses for organizations [1,2], as well as impact teamwork, decision-making, and innovation [3]. By addressing well-being in the workplace, organizations can create a healthier culture and work environment, which can result in increased employee engagement, performance, and productivity, as well as better financial outcomes and retention [4]. Moreover, fostering workplace well-being can have broader societal benefits, such as reducing health care burdens, promoting mental health awareness, and contributing to overall community well-being [5-7]. Therefore, it is crucial to prioritize well-being in the workplace, and digital health tools offer promising solutions for achieving this goal.

In recent years, there have been calls for digital health implementation in workplace settings [8]. Although the digital health market is thriving, with over 350,000 products available and a growth rate of 25% per year [9], digital health technologies for the workplace are minimal, and existing ones focus on employees at the individual level and do not address team and organizational workplace issues [10]. Recognizing the need for digital health interventions, the World Health Organization [11] and the European Commission [12] have published guidelines for digital health tools, with the main objective of evidence-based practice. A key problem with existing digital health products is that the regulation around them is minimal [13]. The majority of these products do not always reach the basic standards of scientific rigor, such as peer-reviewed publications, with only 2.08% being evidence-based [14]. Since traditional approaches, such as employee assistance programs, are challenging to reach and have low uptake [15] and remote and hybrid work is rising, there is a need for evidence-based digital health solutions in the workplace.

According to Stevenson and Farmer [8], well-being measurement is a crucial workplace core standard for employees and teams to monitor well-being and take positive action to improve it. It also enables team leads and organizations to use the results to establish appropriate well-being strategies and enhance organizational well-being in general. Therefore, there is a crucial need for a robust well-being measurement instrument in the workplace for employees, teams, and organizations. One platform that aims to fulfill the need for such solutions is Quan, a digital well-being platform for employees, teams, and organizations, which has a 3-fold process: (1) quantify well-being at employee, team, and organizational levels; (2) provide comprehensive and personalized well-being insights as well as information on the underlying reasons; and (3) deliver tailored interventions for individual employees and teams. Our research is the first step toward this approach and concerns the development of a novel, valid, and reliable measurement instrument that assesses well-being at the workplace and in personal life. This instrument is called the Quan Well-being Index (QWI).

Assessment of Well-Being at the Workplace and in Personal Life

The development of a novel, valid, and reliable measurement instrument for the well-being of employees, teams, and organizations, which will be used in a digital setting, is of pivotal importance. Currently, no measurement instrument assesses well-being in both personal and work life. A plethora of measures exist for assessing various well-being aspects of personal life. Despite the growth of self-report well-being questionnaires, there is significant variation in their conceptualization of well-being (eg, eudemonic and hedonic well-being, and quality of life) and their methodological aspects such as length and psychometric properties [16]. Furthermore, the majority of available measures address well-being in general or in personal life (eg, the Warwick-Edinburgh Mental Well-being Scale [17]) or focus on only 1 aspect, such as quality of life (eg, Quality of Life Inventory [18]). Moreover, the majority of well-being measures prioritize mental well-being over aspects related to physical health, such as sleep, nutrition, and exercise. On the other hand, the assessment of well-being in domain-specific settings, such as work, is still in its infancy, with only a few existing validated measures [19-21] that capture only workplace-related well-being such as job satisfaction, burnout, and stress [22-24]. To establish a comprehensive model of occupational well-being, work-related and personal facets need to be considered [25].

Frameworks for Well-Being in the Workplace

Various theoretical models have been developed to explain the factors that contribute to well-being in the workplace, with particular emphasis on the interaction between job characteristics and psychological needs. Among these, the Job Demands-Resources (JD-R) model and Self-Determination Theory (SDT) have emerged as 2 of the most influential frameworks, offering complementary perspectives on how workplace conditions and intrinsic motivation shape well-being. The JD-R model [26] stands out as the primary framework for conceptualizing the drivers of employee well-being, and its effectiveness is, in part, attributed to its suitability in interpreting interventions within the workplace [27-29]. According to the JD-R model, the mental well-being of workers results from the interplay of 2 distinct categories: job demands and job resources. Job demands encompass the various facets of work, such as physical, psychological, social, or organizational, that necessitate sustained effort, whether it be physical, cognitive, or emotional [27]. Job resources refer to the physical, psychological, social, or organizational elements of work that have the potential to motivate, assist in achieving work-related goals, mitigate the effects of job demands, and foster personal development and learning [27]. Beyond workplace factors, the JD-R model also acknowledges the role of personal demands and resources in shaping well-being. Personal demands refer to individual characteristics or life circumstances, such as caregiving responsibilities, financial stress, and chronic health conditions, that require effort and may contribute to strain [30]. Conversely, personal resources, including resilience, self-efficacy, optimism,

XSL•FO
RenderX

and emotional intelligence, can help employees cope with challenges, enhance motivation, and buffer the negative effects of high job demands [30]. The interaction between job-related factors and personal factors is crucial, as personal resources can amplify the benefits of job resources, while personal demands may exacerbate workplace stressors, further influencing overall well-being.

The SDT has emerged as a crucial framework for understanding and promoting well-being in the workplace. Grounded in the work of Ryan and Deci [31], the SDT highlights the fundamental psychological needs of autonomy, competence, and relatedness, asserting that meeting these needs is essential for ideal human functioning and well-being [31]. In the workplace, the SDT has been instrumental in explaining the impact of autonomy on employee motivation and satisfaction [32]. When individuals perceive a high degree of autonomy in their work, they could have increased job satisfaction and engagement [33]. Furthermore, the fulfillment of competence needs, achieved through opportunities for skill development and feedback, is linked to enhanced job performance and personal well-being [34]. Finally, creating a workplace environment that supports relatedness and fosters positive social connections and interpersonal relationships has been linked with increased levels of well-being and job satisfaction [35]. Thus, the application of the SDT in the workplace not only contributes to individual well-being but also holds promise for enhancing organizational outcomes. However, so far, no measurement instrument of workplace and personal life well-being has considered the JD-R model and the SDT. Therefore, there is a need for a novel, holistic, and robust measurement instrument that takes into consideration these frameworks and simultaneously assesses well-being at the workplace and in personal life, as these settings are intricately associated.

Objectives of the QWI

The development of a digital workplace and personal life well-being index will support employees at the individual level to track, reflect, act on, and ultimately improve their well-being. Existing evidence suggests that digital mental health platforms that provide measurement scores to their users may improve their mental health by supporting them to be aware of and reflect on it [36,37]. At the team level, the index will act as a catalyst for establishing bidirectional well-being communication between employees and their team leads, which has been shown to contribute to improved well-being and increased trust between them [38]. Furthermore, such a measurement instrument will support users in assessing and monitoring their well-being across time by identifying areas of strength and risk and subsequently the impact of tailored interventions. Finally, at the organizational level, the index will provide organizations with metrics that will help them identify which well-being initiatives they should invest in and offer to their employees. This, in turn, will lead to improved employee and team well-being and better functioning organizations.

Overview of Studies

The primary objective of this project is to create and validate a comprehensive self-report tool called the QWI. This instrument aims to evaluate the complete range of well-being aspects in both professional and personal domains for individuals, teams, and organizations. Building upon existing well-being theories and models, this research encompasses 3 studies focused on developing and refining the QWI using a British sample. Study 1 details the creation of the initial item pool and the process to ensure face and content validity. Study 2 describes the development of the QWI, using data collected from online panels to explore the factor structure. Study 3 employs a separate sample to verify the structure of the QWI and examines various models using a confirmatory approach. Additionally, study 3 reports on the internal reliability, convergent and divergent validity, and test-retest reliability of the QWI through psychometric indicators.

Aim

The aim is to develop and validate a novel, psychometrically robust self-report instrument called the QWI, which is designed to assess well-being across both the workplace and personal life domains. The QWI seeks to capture individual, team, and organizational well-being to support digital health applications in workplace settings.

Study 1

Methods

Item Generation

The initial item pool was generated by an experienced organizational psychologist (EJ), who has over 15 years of practice in academia and occupational health psychology, leadership development, and employee well-being consultation across diverse industries. The item development was informed by a combination of applied practice and an integrative review of the literature on personal life and workplace well-being, including key constructs from the JD-R model [26] and SDT [31]. The JD-R model posits that well-being at work is shaped by a balance between job demands (eg, workload and emotional strain) and job resources (eg, autonomy and social support). This model provided the foundation for generating items throughout the initial model, ensuring that both protective and risk-related aspects of the work environment were represented. The SDT highlights the importance of fulfilling basic psychological needs, such as autonomy, competence, and relatedness, that are essential for well-being. Additional well-being models, such as PERMA (positive emotion, engagement, relationships, meaning, and accomplishment) [39] and Keyes' model of flourishing [40], also informed item generation, particularly for capturing positive functioning beyond symptom reduction. These frameworks collectively informed the conceptual structure of the measure.

Based on this theoretical foundation, a preliminary set of 51 items was generated, each intended to capture an aspect of individual well-being relevant in the personal life and workplace contexts. The items were grouped into emergent and coherent themes based on conceptual similarity, resulting in 6 factors, which formed the foundation of our well-being definition (ie, the experience of mental and physical health and the sense of meaning, self-fulfillment, and social connectedness). These were labeled as: *mind* (11 items), *body* (8 items), *social*

XSL•FO
RenderX

connectedness (8 items), *meaning* (10 items), *self-fulfillment* (11 items), and *general well-being* (3 items). These factors collectively formed the initial conceptual framework for the QWI. This process ensured that the QWI is grounded in theory, responsive to practical workplace realities, and reflective of multidimensional well-being that spans both personal and professional settings.

We acknowledge that relying on a single expert for the item generation and thematic grouping is a limitation, particularly regarding potential bias and limited perspective. However, subsequent phases of the project, including expert review and empirical validation (described in the face and content validity subsections and in studies 2 and 3), were used to mitigate this limitation and refine the structure and content of the QWI.

Face Validity

We presented the 51 items to 4 English-speaking nontechnical staff members at Quan (2 males, 1 female, and 1 nonbinary) with a mean work experience of 16 (range 1-25) years and individually asked them to assess the items by rating them on how clear and understandable they are (1=not clear and understandable, 2=somewhat clear and understandable, 3=clear and understandable, and 4=very clear and understandable) and to provide qualitative feedback on what is unclear and not easily understandable for items rated as 1 or 2 and provide suggestions on rewording.

Content Validity

We asked 4 experts in the field of well-being (2 academic psychologists and 2 practitioners; 2 females and 2 males) with a mean professional experience of 15.6 (range 10-20) years to rate the 51 items on how well they assessed each construct (1=not relevant, 2=somewhat relevant, 3=quite relevant, and 4=highly relevant) and to provide qualitative feedback and suggestions for rewording or removing items.

Ethical Considerations

Human Subject Ethics Review Approvals or Exemptions

All studies were reviewed and approved by the Ethics Board of King's Business School, King's College London, under reference number HR/DP-22/23-34006. The research involved human participants, and all procedures were conducted in accordance with the ethical standards outlined in the Declaration of Helsinki and institutional guidelines for research involving human subjects.

Informed Consent

All participants provided informed consent electronically prior to participation. The consent form outlined the purpose of the study, the voluntary nature of participation, the procedures for data handling and anonymity, and the right of participants to withdraw at any time without penalty. Consent was obtained via the Qualtrics platform in the form of ticking a checkbox and was recorded following Ethics Board–approved procedures.

Privacy and Confidentiality

Participant data were collected anonymously, with no identifying information retained. IP addresses were not stored, and all responses were deidentified at the point of collection. Data were stored securely on encrypted institutional servers, accessible only to authorized research team members. All analyses were conducted on anonymized datasets, and all reported findings were based on aggregated results.

Compensation Details

Participants received monetary compensation for their time, consistent with fair market rates for online panel participation. Compensation was provided by Qualtrics in line with their participant remuneration policies, and participants were informed of compensation prior to consenting to take part in the study.

Results

Face and Content Validity

The face validity of all items was high (mean 3.6). Based on the qualitative feedback, we reworded 2 items to further improve face validity. To assess content validity, we calculated interrater agreement using the Item Content Validity Index, following the threshold of 0.80 recommended by Wynd et al [41]. Items with an Item Content Validity Index value below this threshold were removed (n=6), and 3 additional items were reworded based on expert feedback.

Study 2

Methods

Participants

A sample of UK residents (n=1037) was recruited through the Qualtrics [42] online research panel, using quota-based convenience sampling. Eligibility criteria included age (18-65 years), location (United Kingdom), and fluency in English. Quotas were applied to align the sample with UK population distributions regarding age, gender, ethnicity, and region, based on 2020 census data and state statistics [43]. This approach ensured demographic diversity reflective of national benchmarks, although the sample was not probabilistically representative in a strict statistical sense. Recruitment was stratified by UK regions (England, Scotland, Wales, and Northern Ireland). Based on quality and eligibility criteria (total survey completion duration less than one-third of the median survey duration, incorrect responses to 1 attention filter question, nonfluency in English, location other than the United Kingdom, and age younger than 18 years or older than 65 years), we removed 17 participants. The final sample included 1020 participants.

Measures

The QWI includes 45 items and asks individuals to evaluate their experiences over the past 2 weeks. Participants are presented with the cue "During the past two weeks..." and then with the item (eg, "I have been experiencing stress in my work life"). We opted for a 2-week recall period as it helps minimize recall bias while still capturing meaningful patterns in recent experiences [44,45]. Furthermore, previous research has suggested that a 2-week recall period offers a practical balance between reducing recall bias and capturing a representative summary of well-being, rather than a momentary state [46,47].

XSL•FO
RenderX

This timeframe has been adopted in several validated and widely used mental health questionnaires, including the Warwick-Edinburgh Mental Well-being Scale [17], the World Health Organization-5 Well-being Index [48], the Patient Health Questionnaire-9 [49], and the Generalized Anxiety Disorder-7 [50]. Next, participants rate each item on how frequently each applies to them from "Never" (0 points) to "Always" (6 points). We chose a 7-point Likert scale as it provides better reliability and differentiation than shorter scales [51], reduces acquiescence biases and measurement errors while maintaining cognitive ease [52], and is commonly used in validated well-being scales (World Health Organization-5 [53,54]), allowing for comparability.

Furthermore, frequency-based measures have been increasingly favored in well-being research for their ecological validity and greater clarity in capturing behavioral patterns, compared to agreement-based formats [55,56]. The choice of a 7-point scale balances granularity with usability, providing sufficient sensitivity to detect meaningful differences while maintaining simplicity for workplace administration [51]. The presence of a clear midpoint further supports respondent clarity and reduces cognitive load, particularly in diverse occupational settings [57].

Procedure

The survey was built online on Qualtrics. Participants reviewed the research information and gave their consent to take part in the study. Next, participants completed demographic questions and then the QWI items. The consent form and demographic questions were presented in a fixed order, while the items from the QWI and the attention filter were shown in a random order. The median completion duration of the survey was 6 minutes.

Statistical Analysis

For exploratory factor analysis (EFA), we used the psych package for R [58]. We used oblimin rotation as we expected the factors to be correlated. We verified the sampling adequacy with the Kaiser-Meyer-Olkin measure and Bartlett test of sphericity. We used different factor extraction criteria such as the scree plot, Kaiser criterion of eigenvalues >1, VSS (very simple structure) criterion [59], and parallel analysis [60]. We removed items with loadings <0.50, uniqueness >0.50, and commonality <0.50.

Results

Sample Characteristics

The final sample consisted of 1020 UK residents (525 female, 51.5%; 1 person did not disclose their gender), ranging in age from 18 to 65 years (mean 43.06, SD 12.98 years). The sample was spread around the United Kingdom (England: 886/1020, 86.8%; Scotland: 77/1020, 7.5%; Wales: 38/1020, 3.8%; Northern Ireland: 19/1020, 1.9%) and was highly educated (tertiary education: 740/1020, 72.5%; secondary education: 276/1020, 27.1%; primary education: 4/1020, 0.4%). In terms of ethnicity, 84.4% (861/1020) identified as White; 5.6% (57/1020) as Asian or Asian British; 7.1% (72/1020) as Black, African, Caribbean, or Black British; 2.5% (25/1020) as mixed; and 0.5% (5/1020) as other ethnic groups. The majority of the sample was in full-time employment (760/1020, 74.5%) and in nonmanagerial positions (627/1020, 61.5%).

EFA Results

The Kaiser-Meyer-Olkin value was 0.98, and the Bartlett test of sphericity (χ^2_{990}=37063.54; P<.001) indicated that correlations between the 45 items were adequately large for the EFA. The scree plot indicated a 4-factor solution, the Kaiser and VSS criteria recommended a 5-factor model, and the parallel analysis suggested a 6-factor solution. We opted for a 5-factor solution as recommended by 2 criteria (Kaiser and VSS criteria) and ran the EFA. When visually inspecting the loadings of the items, we observed that all items loaded on 4 factors. Therefore, we reran the EFA with 4 factors, and the model explained 0.60 of the variance.

We excluded 10 items considering their low loadings, high uniqueness, and low commonality levels. Subsequently, the model underwent refitting with the remaining 35 items. This final iteration, comprising 4 factors, accounted for 64% of the variance, with each factor demonstrating an eigenvalue of 1.

Next, we defined appropriate labels for the 4 factors. The items that loaded on the first factor suggested aspects, such as positive affect, social connectedness, life satisfaction, and mindfulness, related to personal life well-being, and thus, the factor was labeled "thrive and connect in personal life." The second factor was related to positive affect, social connectedness, and life satisfaction of workplace well-being and was labeled "thrive and connect at work." The third and fourth factors corresponded to our hypothesized constructs of mind and body and were labeled "mental health" and "physical health," respectively. Table 1 presents the 4 factors, their items, and their loadings.

XSL•FO
RenderX

Table 1. Quan Well-being Index factors, their items, and their loadings.

Quan Well-being Index factors and items	Loading
Thrive and connect in personal life	
1. I have been in a positive mood in my personal life.	0.61
2. I have been able to focus on the present moment and accept my feelings and thoughts in my personal life.	0.54
3. I have been satisfied with my relationships in my personal life.	0.79
4. Support from others has been available to me in my personal life.	0.69
5. I have been able to support others in my personal life.	0.61
6. I have been feeling appreciated in my personal life.	0.79
7. I have been experiencing a sense of belonging in my personal life.	0.86
8. I have been feeling satisfied with my personal life.	0.84
9. I have been feeling fully engaged in my personal life activities.	0.60
10. I have been feeling that I am able to continuously learn, change, and thrive in my personal life.	0.57
11. I have been feeling a sense of purpose in my life.	0.58
12. I have been feeling that my purpose aligns with my personal actions.	0.50
13. I have been feeling that what I do in my personal life is meaningful.	0.74
14. I have been feeling optimistic about the future of my personal life.	0.77
Thrive and connect at work	
15. I have been in a positive mood in my work life.	0.63
16. I have been satisfied with my relationships in my work life.	0.65
17. Support from others has been available to me in my work life.	0.69
18. I have been able to support others in my work life.	0.55
19. I have been feeling appreciated in my work life.	0.84
20. I have been experiencing a sense of belonging in my work life.	0.84
21. I have been feeling satisfied with my work life.	0.86
22. I have been feeling fully engaged in my job-related activities.	0.71
23. I have been feeling that I am able to continuously learn, change, and thrive in my job.	0.75
24. I have been feeling that my purpose aligns with my work activities.	0.66
25. I have been feeling that what I do in my organization is meaningful.	0.78
26. I have been feeling optimistic about the future of my organization.	0.70
Physical health	
27. I have been eating and drinking healthily.	0.76
28. I have been satisfied with the amount of physical activity I have had.	0.70
29. I have been able to manage my cravings without depending on them.	0.73
30. I have been feeling good about my body.	0.63
Mental health	
31. I have been experiencing stress in my personal life.	0.62
32. I have been experiencing stress in my work life.	0.69
33. I have been feeling nervous, anxious, or on edge in my personal life.	0.67
34. I have been feeling nervous, anxious, or on edge in my work life.	0.70
35. I have been feeling burned out.	0.55

XSL•FO
RenderX

Study 3

Methods

Participants

An independent sample of participants (n=966) was recruited to conduct the confirmatory factor analysis (CFA) via the Qualtrics platform, with the same inclusion criteria as in study 2. To assess test-retest reliability, a subset of participants (n=275) completed the QWI 2 weeks after the CFA stage.

Measures and Procedure

CFA and Convergent and Divergent Validity

A section of the questionnaire included a demographics survey that was identical to the one used in study 2. Participants completed the 35-item QWI developed in study 2. For convergent validity, the Flourishing-at-Work Scale–Short Form [61] and the Warwick-Edinburgh Mental Well-being Scale [17] were used, and for divergent validity, the Big Five Personality Inventory short version [62] was used. All questionnaires and items were presented in a randomized order. The median completion duration of the survey was 13 minutes.

Test-Retest Reliability

For the test-retest analysis, participants completed the 35-item QWI a second time (items presented in a randomized order) approximately 2 weeks after the CFA and convergent and divergent validity stages. The median completion duration of the survey was 3 minutes.

Statistical Analysis

The lavaan package for R [63] was used for the CFA. To evaluate the factor structure of the QWI, we tested 4 different factor models. In model 1, we tested a hierarchical model with a general factor (general well-being) that influenced the 4 group factors (thrive and connect in personal life, thrive and connect at work, mental health, and physical health), which in turn affected the items that belong in each of them. In model 2, we explored a transformed variant of the hierarchical model (Schmid-Leiman) in which both the general factor and the 4 group factors directly influenced the items. In model 3, we assessed a simple CFA model, where only the 4 group factors influenced the items, and there was no general factor. In model 4, we tested a similar CFA model as model 3, but allowed intercorrelations between the group factors. All 35 items were loaded on the same factors as suggested by the EFA, and only the factor relation structure differed.

Results

Sample Characteristics

The final sample consisted of 966 participants, with a mean reported age of 44.4 (SD 12.52) years. Of the 966 participants, 486 (50.3%) were male and 480 (49.7%) were female. Regarding ethnicity, 87.6% (846/966) of participants identified as White; 6.9% (67/966) as Black, African, Caribbean, or Black British; 3.3% (32/966) as Asian or Asian British; 1.9% (18/966) as mixed; and 0.3% (3/966) as other ethnic groups. Regarding region, 88.9% (859/966) of participants were based in England, 6.2% (60/966) in Scotland, 3.6% (35/966) in Wales, and 1.2% (12/966) in Northern Ireland. The majority of the sample was highly educated (tertiary education: 508/966, 52.6%; secondary education: 455/966, 47.1%; primary education: 3/966, 0.3%). The majority of participants were in permanent employment (full-time or part-time work: 786/966, 81.4%), while 14.9% (144/966) were self-employed and 3.7% (36/966) were employed in temporary or contractual work.

In the test-retest reliability subset (n=275), the mean age was 52.12 (SD 9.56) years. Of the 275 participants in this subset, 105 (38.2%) were male and 170 (61.8%) were female. Regarding ethnicity, 93.1% (256/275) of participants identified as White; 3.3% (9/275) as Asian or Asian British; 1.5% (4/275) as Black, African, Caribbean, or Black British; and 2.2% (6/275) as mixed.

CFA: Model Testing and Evaluation

The fit indices, model comparison fit indices, and model comparison statistics for all 4 models are presented in Table 2. Model 1 demonstrated the best fit to the data (χ^2_{521}=3467.00; Bentler comparative fit index=0.906; Tucker-Lewis index=0.892; root mean square error of approximation=0.077; standardized root mean square residual=0.048) and was superior to all other models in terms of both Akaike information criterion (AIC) and Bayesian information criterion (BIC) values (ΔAIC=0; ΔBIC=0). Models 2, 3, and 4 displayed poorer fit, with substantially higher AIC and BIC values (ΔAIC ≥590.08; ΔBIC ≥531.17). Interestingly, models 3 and 4 yielded identical fit indices (χ^2_{554}=4657.33; Bentler comparative fit index=0.869; Tucker-Lewis index=0.859; root mean square error of approximation=0.088; standardized root mean square residual=0.064), suggesting that the addition of interfactor correlations in model 4 did not improve model fit. This finding was likely due to weak or negligible correlations among the group factors.

XSL•FO
RenderX

Table 2. Goodness-of-fit statistics comparison results for the 4 structural equation models.

CFA[a] models	Chi-square (*df*)	CFI[b]	TLI[c]	RMSEA[d]	SRMR[e]	ΔAIC[f]	ΔBIC[g]
Model 1 (hierarchical)	3467.00 (521)	0.906	0.892	0.077	0.048	0.00	0.00
Model 2 (Schmid-Leiman)	4658.26 (556)	0.869	0.859	0.087	0.064	590.08	531.17
Model 3 (simple factor)	4657.33 (554)	0.869	0.859	0.088	0.064	1124.32	1494.68
Model 4 (factor intercorrelation)	4657.33 (554)	0.869	0.859	0.088	0.064	593.15	963.51

[a]CFA: confirmatory factor analysis.

[b]CFI: Bentler comparative fit index.

[c]TLI: Tucker-Lewis index.

[d]RMSEA: root mean square error of approximation.

[e]SRMR: standardized root mean square residual.

[f]ΔAIC: relative difference in the Akaike information criterion between the model and the best model.

[g]ΔBIC: relative difference in the Bayesian information criterion between the model and the best model.

We assessed the internal reliability of the group factors with Cronbach α, McDonald ω total [64], and Guttman λ6 [65]. As presented in Table 3, estimates of internal reliability were good or very good for all the factors.

Table 3. Summary statistics of the Quan Well-being Index factors, and indicators of internal reliability (n=966) and test-retest reliability (n=275).

Reliability indicator	Thrive and connect in personal life	Thrive and connect at work	Mental health	Physical health	General well-being
Mean (SD)	51.83 (19.21)	40.57 (18.56)	13.42 (5.83)	18.10 (6.82)	123.92 (41.51)
Maximum	84	72	24	30	219
Minimum	0	0	0	0	9
Cronbach α	.96	.97	.86	.85	.97
McDonald ω total	0.96	0.97	0.86	0.85	0.97
Guttman λ6	0.96	0.97	0.86	0.82	0.98
Test-retest correlation	0.81	0.74	0.74	0.80	0.81

Convergent and Divergent Validity

Table 4 presents the correlations of the 5 QWI factors with all the questionnaires that were implemented for convergent and divergent validity. The convergent validity of the QWI can be seen in the strong correlations of all 5 factors with the Warwick-Edinburgh Mental Well-being Scale (r=0.45 to 0.85), with the highest value obtained for the general well-being factor (r=0.85). Further contributing to the convergent validity of the QWI are the very strong correlations between the "thrive and connect at work" factor of the QWI and all the Flourishing-at-Work Scale factors (r=0.80 to 0.84). Divergent validity was established by the weak and negative correlations between all QWI factors and all Big Five Personality Inventory factors.

XSL•FO
RenderX

Table 4. Pearson correlation coefficients between Quan Well-being Index factors and existing measures of well-being in personal life and at the workplace, and personality traits.

Existing measures	Thrive and connect in personal life	Thrive and connect at work	Mental health	Physical health	General well-being
Warwick-Edinburgh Mental Well-being Scale	0.82	0.67	0.45	0.68	0.85
Flourishing-at-Work Scale					
Emotional well-being	0.57	0.80	0.27	0.48	0.73
Psychological well-being	0.54	0.82	0.16	0.46	0.71
Social well-being	0.54	0.84	0.21	0.47	0.72
Big Five Personality Inventory					
Extraversion	0.10	0.11	0.04	0.02	0.10
Agreeableness	0.30	0.33	0.17	0.26	0.35
Conscientiousness	0.27	0.26	0.19	0.27	0.31
Neuroticism	–0.46	–0.38	–0.48	–0.44	–0.52
Openness to experience	0.09	0.04	–0.02	0.06	0.06

Test-Retest Reliability

Table 3 presents the test-retest correlations for the 5 factors, which were determined to be strong to very strong (ranging from 0.74 to 0.81) and statistically significant ($P<.001$).

Overall, the results from study 3 indicated that model 1 (Figure 1) was the best-fitting model, supporting the hypothesized hierarchical structure where the "general well-being" factor influenced the 4 group factors, which in turn impacted their associated individual items. The Bentler comparative fit index, the observed reliability coefficients, and the correlations with the additional questionnaires that were used indicate that the 35-item QWI has the required level of reliability and validity.

XSL•FO
RenderX

Figure 1. The factor structure of the Quan Well-being Index as demonstrated by model 1 (hierarchical model) in the confirmatory factor analysis.

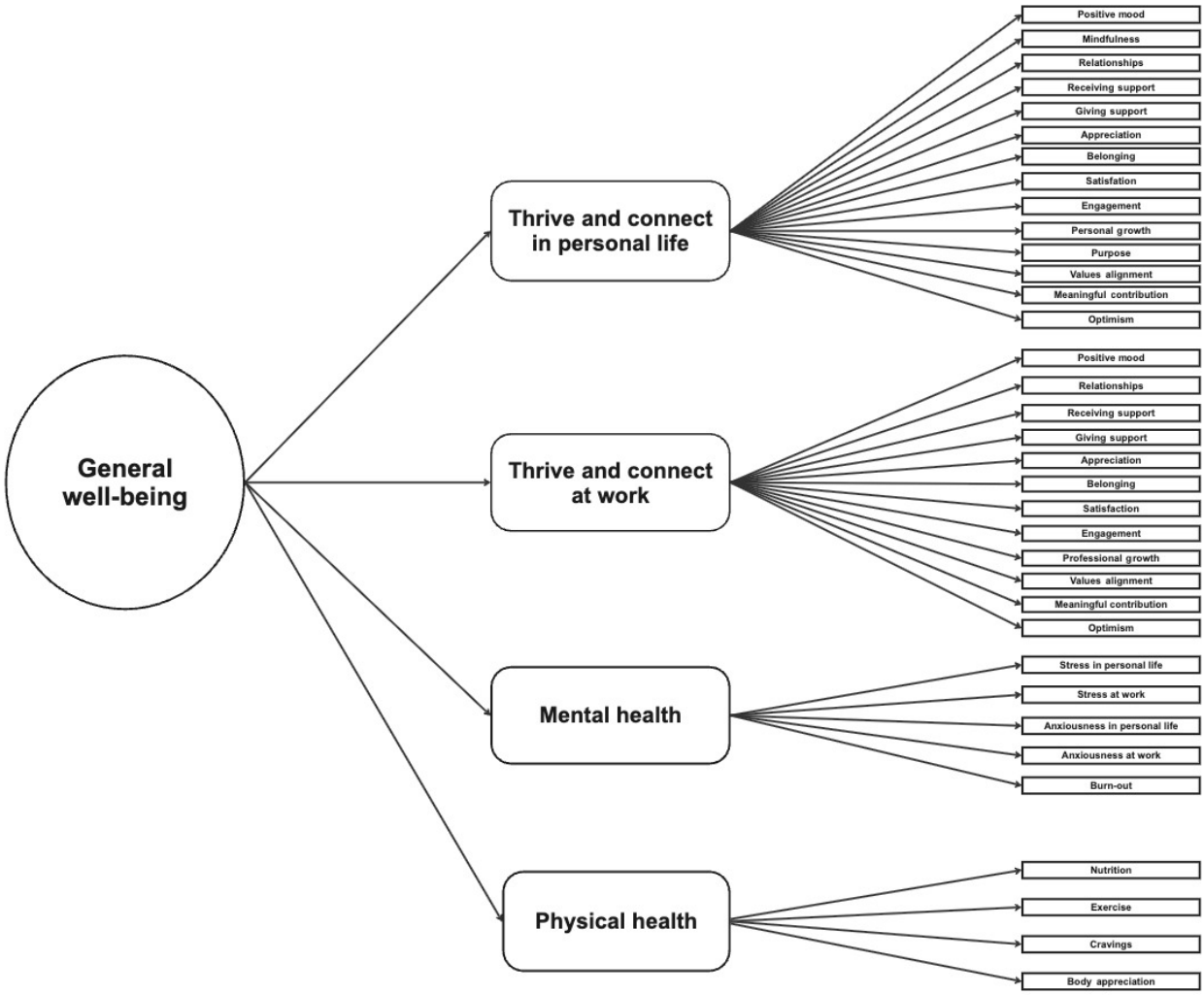

Discussion

Overview

We report on the development of the QWI, a novel self-report index that measures well-being in personal life and at the workplace. The factors identified in this index encompass, corroborate, and quantify aspects of well-being that have previously only been examined in isolation, integrating them into a single measurement instrument for the first time [39,55]. The development of QWI factors has also generated new insights into well-being in personal life and at the workplace, which are presented below.

Summary of Key Findings

The development and validation of the QWI as a novel and holistic measure of well-being in personal life and at the workplace have been demonstrated through a series of studies, showing good internal validity, convergent and divergent validity, and test-retest reliability. This comprehensive tool encompasses 5 key factors that collectively provide a holistic well-being assessment.

The first factor "thrive and connect in personal life" covers various aspects of personal life, including emotional state, mindfulness, relationships, social support, engagement, and sense of purpose. These results indicate that well-being in personal life is a multifaceted concept, encompassing various interconnected elements that contribute to an individual's overall life satisfaction and psychological health [39,66]. The second factor "thrive and connect at work" covers various aspects of work life, such as emotions, relationships, personal growth, and feeling valued, aligning with prior research emphasizing the importance of workplace well-being [26,31]. The third factor "mental health" emphasizes the need to consider mental health issues in both personal and professional contexts, recognizing that individuals may experience different levels of feelings (eg, stress) in personal life and at the workplace [66]. The fourth factor "physical health" is an essential aspect of holistic well-being that covers characteristics related to physical activity and nutrition, consistent with research highlighting the interdependence of physical and psychological health [40,67].

Theoretical and Practical Implications

The presence of a general well-being factor above the 4 group factors (*thrive and connect in personal life*, *thrive and connect at work*, *physical health*, and *mental health*) suggests that well-being is best conceptualized as a higher-order, integrative construct that unites experiences across both personal and professional life domains. This supports the growing body of

XSL•FO
RenderX

research emphasizing the systemic and interrelated nature of well-being across life domains [68,69].

Theoretically, this structure resonates with recent multidimensional and ecological models of well-being [70,71], which move beyond binary hedonic or eudaimonic frameworks and instead frame well-being as a dynamic model influenced by physical, psychological, social, and contextual factors. The QWI's higher-order structure reflects this complexity, recognizing that individual well-being outcomes do not occur in isolation but are shaped by conditions both at the workplace and in personal life.

Furthermore, the finding that *thrive and connect in personal life* and *thrive and connect at work* load onto a common higher-order factor challenges the often rigid boundaries between occupational and personal well-being that dominate traditional models. This aligns with the "whole person" perspective emerging in organizational psychology [72,73], which posits that work and nonwork experiences are mutually influential. For instance, positive interpersonal dynamics at work can buffer stress in personal life, while strong personal relationships may enhance workplace engagement and mood [67].

QWI's hierarchical structure is broadly consistent with prominent multidimensional models of well-being, including the PERMA model [39], which conceptualizes well-being as a composite of positive emotion, engagement, relationships, meaning, and accomplishment. However, the QWI makes a distinctive contribution by operationalizing both domain-specific well-being (*thrive and connect at work* vs *thrive and connect in personal life*) and foundational health factors (*physical health* and *mental health*) within a single integrated measurement framework. This dual focus on "where" well-being is experienced and "how" it is sustained may extend existing models by offering a more pragmatic approach for organizational interventions.

The presence of the "thrive and connect" concept as a factor both in personal life and at the workplace indicates that there is a need for distinction between aspects that take place in different settings. As such, individuals are provided with a tool to track and take action on their well-being in separate contexts, which might require different approaches. For example, issues related to work could be brought up in one-to-one meetings between employees and their managers to facilitate bidirectional communication about aspects that might not be apparent in one's work but can be adjusted to promote healthier organizations [74]. Moreover, the ability to differentiate between personal and professional well-being factors could lead to more targeted interventions and support systems both at the individual level and within organizational structures, ultimately contributing to a more holistic approach to well-being management.

From a practical standpoint, the presence of a general well-being factor allows for both macro- and micro-level analysis. At the macro level, the general factor serves as a useful indicator of an individual's or team's holistic well-being, enabling organizations to track broad shifts over time or following interventions. At the micro level, the 4 factors facilitate precise diagnostics, allowing leaders or practitioners to identify specific areas (eg, workplace social connectedness vs mental health) that require tailored support.

Scoring is structured to align with the hierarchical factor model validated in study 3. Each item contributes to 1 of 4 factors: *thrive and connect in personal life*, *thrive and connect at work*, *physical health*, and *mental health*. Factor scores are calculated by averaging responses to the items that belong to each factor, and the overall QWI score is the average of the 4 factors. For interpretative purposes in applied settings, we offer the following score ranges: scores of 0.0-1.9 suggest low well-being and may indicate risk, scores of 2.0-3.9 reflect moderate well-being, and scores of 4.0-6.0 indicate high well-being or "thriving." These guidelines are intended to facilitate meaningful interpretation and intervention planning by researchers, practitioners, and organizational stakeholders.

Comparison With Existing Measures of Well-Being

A key contribution of the QWI lies in its integrated assessment of both the personal and professional domains of well-being within a single, psychometrically validated instrument. While several established measures of well-being exist, such as the Warwick-Edinburgh Mental Well-being Scale [17], World Health Organization-5 Well-being Index [54], and PERMA Profiler [75], these tend to focus on general mental well-being and positive functioning, often without capturing the complex interplay between work and nonwork settings.

The QWI addresses this gap by explicitly incorporating 4 interrelated dimensions: *thrive and connect in personal life*, *thrive and connect at work*, *physical health*, and *mental health*. This structure allows for the simultaneous assessment of cross-domain experiences and provides a more ecologically valid representation of how individuals experience well-being across the life course. In contrast, measures like the Job Satisfaction Survey [76] or the Utrecht Work Engagement Scale [77] focus narrowly on occupational well-being, whereas personal well-being tools, such as the Ryff Psychological Well-being Scales [78], do not account for work-related dynamics.

From a practical standpoint, the QWI is designed for ease of use in organizational settings. It comprises 35 items, takes approximately 5 to 7 minutes to complete, and uses a uniform 7-point Likert response scale ranging from "Never" (0 points) to "Always" (6 points). This balance between comprehensiveness and brevity makes it well suited for routine monitoring in time-constrained environments, a feature less common in longer instruments like Ryff scales (84 items) or multidimensional surveys requiring separate administration for work and nonwork domains.

Moreover, by capturing personal and professional well-being within 1 coherent framework, the QWI reduces the burden on respondents and provides practitioners with an integrated profile that enhances intervention planning. This is particularly advantageous in applied settings where holistic well-being strategies (eg, action planning, leadership feedback, and human resource initiatives) benefit from a unified data source. However, we acknowledge that combining domains may reduce granularity

XSL•FO
RenderX

in highly specialized research contexts, which remains a consideration when choosing between integrated and domain-specific tools.

Taken together, the QWI's content coverage, time-efficient administration, and psychometric strength position it as a novel and practical addition to the well-being assessment landscape, particularly for contexts requiring cross-domain insights and targeted, team-level interventions.

Limitations and Future Directions

Certain limitations and future directions that stemmed from the outcomes of this research need to be addressed. The United Kingdom–based sample we recruited may limit the generalizability of the findings to other cultures, for example, non-Western cultures [79]. Validation in other countries might require the inclusion or exclusion of certain items to support different conceptualizations of well-being [80]. For example, in countries that have indigenous populations, such as New Zealand and Australia, items that represent family, land, and rituals might be necessary to encompass their well-being experiences [81]. We are planning to translate and validate the QWI for other continents as well as countries across Europe to explore how well it maps in other languages and what further improvements need to be made. In doing so, particular care will be taken to ensure conceptual, linguistic, and metric equivalence across cultural contexts, which may include exploratory and confirmatory factor analyses in each new language group.

Additionally, although the QWI demonstrated good convergent validity with established well-being measures, all data were derived from self-report instruments. This mono-method approach may introduce shared method variance and raise concerns about response biases, including social desirability and acquiescence effects [82]. To strengthen construct validity and provide a more holistic view of well-being, future studies should incorporate a multimethod approach. This may include behavioral data (eg, absenteeism rates or job performance metrics) and objective physiological indicators, such as cortisol levels, heart rate variability, and sleep quality measures, which can serve as indirect proxies for stress and overall health.

Lastly, while the 7-point Likert response format (ranging from "Never" to "Always") was chosen based on evidence suggesting that it improves reliability and reduces response bias, it may not be optimal in all cultural or linguistic contexts. Adaptations to the scale's format may be necessary when validating the QWI globally, particularly in regions where numeric or verbal anchors are interpreted differently due to linguistic or cultural norms.

By addressing these limitations, future research will be able to further assess the utility, adaptability, and impact of the QWI as a cross-culturally valid and practically useful tool for measuring employee well-being across personal and professional life domains.

Conclusion

In summary, the QWI is a novel, reliable, and validated instrument that systematically measures multiple and distinct aspects of well-being in personal life and at the workplace. It is anticipated that the QWI will be valuable to digitally assess well-being, enabling individuals, teams, and organizations to gain insights and monitor their progress over time, implement appropriate interventions, and ultimately enhance well-being areas that require improvement.

Acknowledgments

The authors would like to thank Professor Ute Stephan for her input throughout the research process.

Data Availability

The Quan Well-being Index (QWI) is freely available for academic research purposes. Researchers interested in using the QWI in scholarly studies are encouraged to contact the corresponding author to request access to the full instrument and accompanying scoring guidelines. For commercial or applied use in organizational settings, the QWI is proprietary and forms part of the well-being assessment platform developed by Quan BV. Commercial entities or practitioners seeking to implement the QWI in workplace interventions should contact the authors or Quan BV to discuss licensing or access terms.

Authors' Contributions

Conceptualization: EJ (lead), GAF (equal)
Data curation: GAF (lead), EJ (supporting), FK (supporting)
Formal analysis: GAF (lead), EJ (supporting), FK (supporting)
Investigation: GAF (lead), EJ (supporting), FK (supporting)
Methodology: GAF (lead), EJ (equal)
Project administration: GAF (lead), EJ (equal)
Resources: GAF (lead), FK (supporting)
Supervision: GAF (lead), EJ (equal)
Validation: GAF
Visualization: GAF
Writing – original draft: GAF
Writing – review & editing:GAF (lead), EJ (equal), FK (supporting)

XSL•FO
RenderX

Conflicts of Interest

GAF and EJ are employees at Quan BV, and EJ owns shares in Quan BV. They both created and validated the Quan Well-being Index reported in this article. EJ was involved in the conceptual development of the Quan Well-being Index. However, the study design, data collection, statistical analyses, and interpretation of results were conducted collaboratively by the full author team, with individual authors independently verifying key aspects of the work. To ensure objectivity, all analyses were independently conducted and verified by co-authors with no financial interest in the platform. All procedures were designed to minimize bias and uphold ethical research standards. This potential conflict of interest has been transparently disclosed, and appropriate safeguards were implemented to protect the integrity of the research. FK is affiliated with King's Business School, which has an in-kind contribution academic partnership agreement with Quan for research projects.

References

1. Hassard J, Thomson L, Blake H. Understanding and Exploring the Cost of Poor Mental Health at Work for Organizations and Society. In: Day A, Cooper CL, editors. The Routledge Companion to Mental Health at Work. New York, NY: Routledge; 2023:77-96.
2. Pinheiro M, Ivandic I, Razzouk D. The Economic Impact of Mental Disorders and Mental Health Problems in the Workplace. In: Razzouk D, editor. Mental Health Economics. Cham: Springer; 2017:415-430.
3. Lu L. Balancing Performance and Well-Being: Motivational and Resource Regulation in the Context of Excessive Availability for Work. In: Day A, Cooper CL, editors. The Routledge Companion to Mental Health at Work. New York, NY: Routledge; 2023:248-266.
4. Harshita L, Senthil A. Impact of employee well-being on organizational performance in workplace. Wesleyan J Res 2021;14(30):27-36 [FREE Full text]
5. Cooper CL, Cartwright S. Healthy mind; healthy organization: a proactive approach to occupational stress. Human Relations 1994 Apr 01;47(4):455-471. [doi: 10.1177/001872679404700405]
6. Goetzel R, Roemer E, Holingue C, Fallin M, McCleary K, Eaton W, et al. Mental health in the workplace: a call to action proceedings from the mental health in the workplace-public health summit. J Occup Environ Med 2018 Apr;60(4):322-330 [FREE Full text] [doi: 10.1097/JOM.0000000000001271] [Medline: 29280775]
7. Guidelines on mental health at work. World Health Organization. 2022. URL: https://www.who.int/publications/i/item/9789240053052 [accessed 2024-03-08]
8. Stevenson D, Farmer P. Thriving at Work: a review of mental health and employers. GOV.UK. 2017. URL: https://www.gov.uk/government/publications/thriving-at-work-a-review-of-mental-health-and-employers [accessed 2024-01-10]
9. Pohl M. 325,000 mobile health apps available in 2017 – Android now the leading mHealth platform. Research2Guidance. 2017. URL: https://research2guidance.com/325000-mobile-health-apps-available-in-2017/ [accessed 2022-10-04]
10. Fleming WJ. Employee well‐being outcomes from individual‐level mental health interventions: Cross‐sectional evidence from the United Kingdom. Industrial Relations Journal 2024 Jan 10;55(2):162-182. [doi: 10.1111/irj.12418]
11. WHO guideline: recommendations on digital interventions for health system strengthening. World Health Organization. 2019. URL: https://apps.who.int/iris/handle/10665/311941 [accessed 2024-06-08]
12. Report of the Working Group on mHealth assessment guidelines. European Commission. 2017. URL: https://digital-strategy.ec.europa.eu/en/library/report-working-group-mhealth-assessment-guidelines [accessed 2024-06-08]
13. Magrabi F, Habli I, Sujan M, Wong D, Thimbleby H, Baker M, et al. Why is it so difficult to govern mobile apps in healthcare? BMJ Health Care Inform 2019 Nov 18;26(1):e100006 [FREE Full text] [doi: 10.1136/bmjhci-2019-100006] [Medline: 31744843]
14. Lau N, O'Daffer A, Colt S, Yi-Frazier JP, Palermo TM, McCauley E, et al. Android and iPhone mobile apps for psychosocial wellness and stress management: systematic search in app stores and literature review. JMIR Mhealth Uhealth 2020 May 22;8(5):e17798 [FREE Full text] [doi: 10.2196/17798] [Medline: 32357125]
15. Azzone V, McCann B, Merrick EL, Hiatt D, Hodgkin D, Horgan C. Workplace stress, organizational factors and EAP utilization. J Workplace Behav Health 2009 Jul;24(3):344-356 [FREE Full text] [doi: 10.1080/15555240903188380] [Medline: 24058322]
16. Cooke P, Melchert T, Connor K. Measuring well-being: A review of instruments. The Counseling Psychologist 2016 Aug 05;44(5):730-757. [doi: 10.1177/0011000016633507]
17. Tennant R, Hiller L, Fishwick R, Platt S, Joseph S, Weich S, et al. The Warwick-Edinburgh Mental Well-being Scale (WEMWBS): development and UK validation. Health Qual Life Outcomes 2007 Nov 27;5:63 [FREE Full text] [doi: 10.1186/1477-7525-5-63] [Medline: 18042300]
18. Frisch MB, Cornell J, Villanueva M, Retzlaff PJ. Clinical validation of the Quality of Life Inventory. A measure of life satisfaction for use in treatment planning and outcome assessment. Psychological Assessment 1992;4(1):92-101. [doi: 10.1037//1040-3590.4.1.92]
19. Kun, Balogh P, Krasz KG. Development of the work-related well-being questionnaire based on Seligman's PERMA model. Period Polytech Soc Man Sci 2016 Dec 05;25(1):56. [doi: 10.3311/ppso.9326]

XSL•FO
RenderX

20. Orsila R, Luukkaala T, Manka M, Nygard C. A new approach to measuring work-related well-being. Int J Occup Saf Ergon 2011 Jan 08;17(4):341-359. [doi: 10.1080/10803548.2011.11076900] [Medline: 22152501]
21. Parker G, Hyett M. Measurement of well-being in the workplace: the development of the work well-being questionnaire. J Nerv Ment Dis 2011 Jun;199(6):394-397. [doi: 10.1097/NMD.0b013e31821cd3b9] [Medline: 21629018]
22. Leiter MP, Maslach C. Job burnout. In: Tetrick LE, Fisher GG, Ford MT, Quick JC, editors. Handbook of Occupational Health Psychology. Washington, DC: American Psychological Association; 2024:291-307.
23. Stanton JM, Balzer WK, Smith PC, Parra LF, Ironson G. A general measure of work stress: the stress in general scale. Educational and Psychological Measurement 2001 Oct 01;61(5):866-888. [doi: 10.1177/00131640121971455]
24. van Saane N, Sluiter J, Verbeek J, Frings-Dresen M. Reliability and validity of instruments measuring job satisfaction--a systematic review. Occup Med (Lond) 2003 May;53(3):191-200. [doi: 10.1093/occmed/kqg038] [Medline: 12724553]
25. Taris TW, Schaufeli WB. Individual Well-Being and Performance at Work: A conceptual and theoretical overview. In: Cooper C, editor. Current Issues in Work and Organizational Psychology. London, UK: Routledge; 2018:189-204.
26. Bakker AB, Demerouti E. The Job Demands-Resources model: State of the art. Journal of Managerial Psychology 2007;22(3):309-328. [doi: 10.1108/02683940710733115]
27. Bakker AB, Demerouti E, Sanz-Vergel A. Job demands–resources theory: ten years later. Annu Rev Organ Psychol Organ Behav 2023 Jan 23;10(1):25-53. [doi: 10.1146/annurev-orgpsych-120920-053933]
28. Hartmann S, Weiss M, Newman A, Hoegl M. Resilience in the workplace: a multilevel review and synthesis. Applied Psychology 2019 Apr 04;69(3):913-959. [doi: 10.1111/apps.12191]
29. Tetrick LE, Winslow CJ. Workplace stress management interventions and health promotion. Annu Rev Organ Psychol Organ Behav 2015 Apr 10;2(1):583-603. [doi: 10.1146/annurev-orgpsych-032414-111341]
30. Xanthopoulou D, Bakker AB, Demerouti E, Schaufeli WB. The role of personal resources in the job demands-resources model. International Journal of Stress Management 2007 May;14(2):121-141. [doi: 10.1037/1072-5245.14.2.121]
31. Ryan RM, Deci EL. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. Am Psychol 2000 Jan;55(1):68-78. [doi: 10.1037//0003-066x.55.1.68] [Medline: 11392867]
32. Gagné M, Deci EL. Self‐determination theory and work motivation. J Organ Behavior 2005 Apr 14;26(4):331-362. [doi: 10.1002/job.322]
33. Deci EL, Connell JP, Ryan RM. Self-determination in a work organization. Journal of Applied Psychology 1989;74(4):580-590. [doi: 10.1037//0021-9010.74.4.580]
34. Van den Broeck A, Ferris DL, Chang C, Rosen CC. A review of self-determination theory's basic psychological needs at work. Journal of Management 2016 Mar 09;42(5):1195-1229. [doi: 10.1177/0149206316632058]
35. Baard PP, Deci EL, Ryan RM. Intrinsic need satisfaction: a motivational basis of performance and well-being in two work settings. J Applied Social Pyschol 2006 Jul 31;34(10):2045-2068. [doi: 10.1111/j.1559-1816.2004.tb02690.x]
36. Kauer SD, Reid SC, Crooke AHD, Khor A, Hearps SJC, Jorm AF, et al. Self-monitoring using mobile phones in the early stages of adolescent depression: randomized controlled trial. J Med Internet Res 2012 Jun 25;14(3):e67 [FREE Full text] [doi: 10.2196/jmir.1858] [Medline: 22732135]
37. Wichers M, Simons C, Kramer I, Hartmann J, Lothmann C, Myin-Germeys I, et al. Momentary assessment technology as a tool to help patients with depression help themselves. Acta Psychiatr Scand 2011 Oct;124(4):262-272. [doi: 10.1111/j.1600-0447.2011.01749.x] [Medline: 21838742]
38. Truong H, McLachlan CS. Analysis of start-up digital mental health platforms for enterprise: opportunities for enhancing communication between managers and employees. Sustainability 2022 Mar 26;14(7):3929. [doi: 10.3390/su14073929]
39. Seligman MEP. Flourish: A Visionary New Understanding of Happiness and Well-being. New York, NY: Simon Element/Simon Acumen; 2011.
40. Keyes CLM. Promoting and protecting mental health as flourishing: a complementary strategy for improving national mental health. Am Psychol 2007;62(2):95-108. [doi: 10.1037/0003-066X.62.2.95] [Medline: 17324035]
41. Wynd CA, Schmidt B, Schaefer MA. Two quantitative approaches for estimating content validity. West J Nurs Res 2003 Aug 01;25(5):508-518. [doi: 10.1177/0193945903252998] [Medline: 12955968]
42. Qualtrics. URL: https://www.qualtrics.com [accessed 2025-06-30]
43. Eurostat. URL: https://ec.europa.eu/eurostat [accessed 2024-05-15]
44. Bradburn N, Rips L, Shevell S. Answering autobiographical questions: the impact of memory and inference on surveys. Science 1987 Apr 10;236(4798):157-161. [doi: 10.1126/science.3563494] [Medline: 3563494]
45. Sudman S, Bradburn NM. Effects of time and memory factors on response in surveys. Journal of the American Statistical Association 1973 Dec;68(344):805-815. [doi: 10.1080/01621459.1973.10481428]
46. Broderick JE, Schwartz JE, Shiffman S, Hufford MR, Stone AA. Recall bias in reporting mental health symptoms: an experience sampling study. J Psychiatr Res 2009;43(4):312-320.
47. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol 2008;4:1-32. [doi: 10.1146/annurev.clinpsy.3.022806.091415] [Medline: 18509902]
48. Bech P, Gudex C, Johansen KS. The WHO (Ten) Well-Being Index: validation in diabetes. Psychother Psychosom 1996;65(4):183-190. [doi: 10.1159/000289073] [Medline: 8843498]

XSL•FO
RenderX

49. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
50. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
51. Preston CC, Colman AM. Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. Acta Psychol (Amst) 2000 Mar;104(1):1-15 [FREE Full text] [doi: 10.1016/s0001-6918(99)00050-5] [Medline: 10769936]
52. Finstad K. The usability metric for user experience. Interacting with Computers 2010 Sep;22(5):323-327. [doi: 10.1016/j.intcom.2010.04.004]
53. Wellbeing measures in primary health care/the DepCare Project. World Health Organization. 1998. URL: https://iris.who.int/handle/10665/349766 [accessed 2025-06-30]
54. Topp CW, Østergaard SD, Søndergaard S, Bech P. The WHO-5 Well-Being Index: a systematic review of the literature. Psychother Psychosom 2015 Mar 28;84(3):167-176 [FREE Full text] [doi: 10.1159/000376585] [Medline: 25831962]
55. Diener E, Lucas RE, Oishi S. Subjective Well-Being The Science of Happiness and Life Satisfaction. In: Snyder CR, Lopez SJ, editors. Handbook of Positive Psychology. NY: New York; 2018.
56. Goodman FR, Disabato DJ, Kashdan TB, Kauffman SB. Measuring well-being: A comparison of subjective well-being and PERMA. The Journal of Positive Psychology 2017 Oct 10;13(4):321-332. [doi: 10.1080/17439760.2017.1388434]
57. Weijters B, Baumgartner H, Schillewaert N. Reversed item bias: an integrative model. Psychol Methods 2013 Sep;18(3):320-334. [doi: 10.1037/a0032121] [Medline: 23646990]
58. Package 'psych'. R Project. URL: https://cran.r-project.org/web/packages/psych/psych.pdf [accessed 2024-07-10]
59. Revelle W, Rocklin T. Very simple structure: an alternative procedure for estimating the optimal number of interpretable factors. Multivariate Behav Res 1979 Oct 01;14(4):403-414. [doi: 10.1207/s15327906mbr1404_2] [Medline: 26804437]
60. Horn JL. A rationale and test for the number of factors in factor analysis. Psychometrika 2025 Jan 01;30(2):179-185. [doi: 10.1007/bf02289447]
61. Rautenbach C, Rothmann S. Psychometric validation of the Flourishing-at-Work Scale – Short Form (FWS-SF): Results and implications of a South African study. Journal of Psychology in Africa 2017 Aug 29;27(4):303-309. [doi: 10.1080/14330237.2017.1347748]
62. Rammstedt B, John OP. Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. Journal of Research in Personality 2007 Feb;41(1):203-212. [doi: 10.1016/j.jrp.2006.02.001]
63. Rosseel Y. lavaan: An R package for structural equation modeling. J Stat Soft 2012;48(2):1-36. [doi: 10.18637/jss.v048.i02]
64. Zinbarg RE, Revelle W, Yovel I, Li W. Cronbach's α, Revelle's β, and Mcdonald's ωH: their relations with each other and two alternative conceptualizations of reliability. Psychometrika 2025 Jan 01;70(1):123-133. [doi: 10.1007/s11336-003-0974-7]
65. Guttman L. A basis for analyzing test-retest reliability. Psychometrika 2025 Jan 01;10(4):255-282. [doi: 10.1007/bf02288892]
66. Diener E. Subjective well-being: The science of happiness and a proposal for a national index. American Psychologist 2000;55(1):34-43. [doi: 10.1037//0003-066x.55.1.34]
67. Sonnentag S. The recovery paradox: Portraying the complex interplay between job stressors, lack of recovery, and poor well-being. Research in Organizational Behavior 2018;38:169-185. [doi: 10.1016/j.riob.2018.11.002]
68. How's Life? Organisation for Economic Co-operation and Development. URL: https://www.oecd.org/en/publications/serials/how-s-life_g1g317ee.html [accessed 2025-06-30]
69. VanderWeele TJ. On the promotion of human flourishing. Proc Natl Acad Sci USA 2017 Aug 01;114(31):8148-8156 [FREE Full text] [doi: 10.1073/pnas.1702996114] [Medline: 28705870]
70. Dodge R, Daly A, Huyton J, Sanders L. The challenge of defining wellbeing. Intnl J Wellbeing 2012 Aug 28;2(3):222-235. [doi: 10.5502/ijw.v2i3.4]
71. Huppert FA, So TTC. Flourishing across Europe: application of a new conceptual framework for defining well-being. Soc Indic Res 2013 Mar 15;110(3):837-861 [FREE Full text] [doi: 10.1007/s11205-011-9966-7] [Medline: 23329863]
72. Grant AM. Relational job design and the motivation to make a prosocial difference. AMR 2007 Apr;32(2):393-417. [doi: 10.5465/amr.2007.24351328]
73. Nielsen K, Nielsen MB, Ogbonnaya C, Känsälä M, Saari E, Isaksson K. Workplace resources to improve both employee well-being and performance: A systematic review and meta-analysis. Work & Stress 2017 Mar 24;31(2):101-120. [doi: 10.1080/02678373.2017.1304463]
74. Penedo FJ, Dahn JR. Exercise and well-being: a review of mental and physical health benefits associated with physical activity. Curr Opin Psychiatry 2005 Mar;18(2):189-193. [doi: 10.1097/00001504-200503000-00013] [Medline: 16639173]
75. Butler J, Kern ML. The PERMA-Profiler: A brief multidimensional measure of flourishing. Intnl. J. Wellbeing 2016 Oct 13;6(3):1-48. [doi: 10.5502/ijw.v6i3.526]
76. Spector PE. Measurement of human service staff satisfaction: Development of the Job Satisfaction Survey. American J of Comm Psychol 1985 Dec;13(6):693-713. [doi: 10.1007/bf00929796]
77. Schaufeli WB, Salanova M, González-Romá V, Bakker AB. The measurement of engagement and burnout: A two sample confirmatory factor analytic approach. J Happiness Stud 2002;3(1):71-92. [doi: 10.1023/A:1015630930326]

XSL•FO
RenderX

78. Ryff CD. Happiness is everything, or is it? Explorations on the meaning of psychological well-being. Journal of Personality and Social Psychology 1989;57(6):1069-1081. [doi: 10.1037//0022-3514.57.6.1069]
79. Scully D, Kremer J, Meade MM, Graham R, Dudgeon K. Physical exercise and psychological well being: a critical review. Br J Sports Med 1998 Jun 01;32(2):111-120 [FREE Full text] [doi: 10.1136/bjsm.32.2.111] [Medline: 9631216]
80. Bakker AB, Demerouti E. Job demands-resources theory: Taking stock and looking forward. J Occup Health Psychol 2017 Jul;22(3):273-285. [doi: 10.1037/ocp0000056] [Medline: 27732008]
81. Henrich J, Heine SJ, Norenzayan A. The weirdest people in the world? Behav Brain Sci 2010 Jun 15;33(2-3):61-83. [doi: 10.1017/s0140525x0999152x]
82. Joshanloo M, Van de Vliert E, Jose PE. Four Fundamental Distinctions in Conceptions of Wellbeing Across Cultures. In: Kern ML, Wehmeyer ML, editors. The Palgrave Handbook of Positive Education. Cham: Palgrave Macmillan; 2021:675-703.

Abbreviations

AIC: Akaike information criterion
BIC: Bayesian information criterion
CFA: confirmatory factor analysis
EFA: exploratory factor analysis
JD-R: Job Demands-Resources
PERMA: positive emotion, engagement, relationships, meaning, and accomplishment
QWI: Quan Well-being Index
SDT: Self-Determination Theory
VSS: very simple structure

Edited by A Mavragani; submitted 10.03.25; peer-reviewed by C Udensi; comments to author 09.04.25; revised version received 30.04.25; accepted 21.05.25; published 08.08.25.

Please cite as:
Floridou GA, Katre F, Jeuken E
A Digital Tool for Assessing Well-Being at the Workplace and in Personal Life: Development and Validation of the Quan Well-Being Index
JMIR Form Res 2025;9:e73713
URL: https://formative.jmir.org/2025/1/e73713
doi:10.2196/73713
PMID:

©Georgia A Floridou, Freya Katre, Emile Jeuken. Originally published in JMIR Formative Research (https://formative.jmir.org), 08.08.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Development of a Web-Based Intervention to Support Primary Health Care Professionals in Digital Health Measurement: User-Centered Participatory Approach

Kristel Meijers[1,2], MA; Esther Bols[1,3], PhD; Emmylou Beekman[1,3,4], PhD; Rachel Slangen[1,5], PhD; Peter Alexander van de Hoef[6], PhD; Darcy Ummels[1,3,7], PhD; Rik Crutzen[2], PhD

[1]Research Center for Autonomy and Participation for Persons with a Chronic Illness, Academy for Speech and Language Therapy, Zuyd University of Applied Sciences, Heerlen, The Netherlands

[2]Department of Health Promotion, Care and Public Health Research Institute, Maastricht University, Maastricht, The Netherlands

[3]Academy for Physiotherapy, Zuyd University of Applied Sciences, Heerlen, The Netherlands

[4]Department of Family Medicine, Care and Public Health Research Institute, Maastricht University, Maastricht, The Netherlands

[5]Research Center for Personalized Lifestyle and Movement-based Care, Academy for Physiotherapy, Zuyd University of Applied Sciences, Heerlen, The Netherlands

[6]Research Group Innovation of Human Movement Care, Research Center for Healthy and Sustainable Living, HU University of Applied Sciences Utrecht, Utrecht, The Netherlands

[7]Department of Rehabilitation Medicine, Care and Public Health Research Institute, Maastricht University, Maastricht, The Netherlands

Corresponding Author:
Kristel Meijers, MA
Research Center for Autonomy and Participation for Persons with a Chronic Illness
Academy for Speech and Language Therapy
Zuyd University of Applied Sciences
Postbus 550
Heerlen, 6400 AN
The Netherlands
Phone: 31 (0) 45 400 60 60 ext 6382
Email: kristel.meijers@zuyd.nl

Abstract

Background: Digital health measurement offers opportunities to address several primary health care challenges, but health care professionals encounter significant implementation barriers. Therefore, resources need to be developed to facilitate the integration of digital health measurement into daily practice.

Objective: We aim to identify the most appropriate format and content for an intervention to support primary health care professionals in adopting digital health measurement. In addition, we describe and reflect on the development process.

Methods: We used a participatory action research approach as well as user-centered design principles. A total of 19 primary health care professionals from 4 disciplines—physical therapy, occupational therapy, speech and language therapy, and general practitioner practice assistance—participated in intervention development as end users. External experts were consulted to broaden perspectives. Data were collected across 3 iterative stages (concept, design, and testing and trials) between January 2022 and December 2023 during cocreative meetings, individual interviews, focus group discussions, usability testing, and prototype use in daily practice. Data were analyzed using content analysis and descriptive statistics.

Results: A web-based, stepwise intervention combining theoretical information, practical aids, examples, and experiences proved most suitable. Key features were concise content, intuitive and attractive graphic design, and flexible navigation and functionalities. Iterative improvements led to an increase in usability ratings from “okay” to “good to excellent.”

Conclusions: Different health care disciplines benefit from similar support strategies; yet, this requires a careful balancing of intervention design and content. Combining participatory action research and user-centered design principles was useful to tailor the intervention to end users’ daily routines. The described development process offers a replicable framework for creating support strategies for digital health measurement in various health care settings.

(JMIR Form Res 2025;9:e72331) doi:10.2196/72331

XSL•FO
RenderX

KEYWORDS

cocreation; digital health; eHealth; health measurement; general practitioners; implementation; language therapy; occupational therapy; participatory action research; physical therapy; primary health care; support; telemedicine; telemonitoring; user-centered design

Introduction

Background

Health care in the Netherlands faces multifaceted challenges that mirror global trends. These challenges include maintaining quality, accessibility, and affordability amid rising demand for care and the scarcity of health care professionals [1-3]. These system-wide pressures are particularly acute in primary care, partly due to the concurrent movement toward personalized health care provided in the home environment [2,4-6]. To address these challenges, national and international governmental bodies, alongside professional associations and research institutes, advocate for the implementation of digital health as a partial remedy [6-11].

Digital health encompasses the use of digital technologies for a diverse array of health-related functionalities, such as patient education, health status monitoring, and the management of patient records [7,12,13]. Preconditions for the effective use of digital health vary across functionalities and health care settings [7,11]. Consequently, this study focuses on one of these functionalities in a specific setting: the use of digital health for measurement purposes within primary health care.

In the Netherlands, primary health care comprises various health care disciplines and operates within a gatekeeping model, where general practitioners (GPs) coordinate care and manage referrals to hospital-based or specialist services. Primary care is largely delivered by independent, privately practicing GPs and other frontline professionals—such as physical therapists and occupational therapists—and is generally accessible to patients without prior GP referral.

Measurement outcomes guide the treatment process by providing diagnostic, prognostic, and evaluative information and by supporting shared decision-making [14,15]. However, to be effective, measurements must be goal oriented and meaningful [14,15]. This requires alignment with patients' needs and objectives and an understandable presentation of measurement outcomes to health care professionals as well as patients and their significant others. Although assessing the effects of digital health use proves to be complicated [4,16], empirical evidence—such as studies on symptom monitoring demonstrating improvements in patients' awareness, self-management, and treatment satisfaction in chronic conditions [6,17,18]—suggests objective benefits, while health care professionals themselves report perceived positive effects of digital health measurement on quality of care, efficiency, patients' self-management, and prevention [3,11,19].

Despite generally positive attitudes of both health care professionals and patients toward digital health [3,11,19-22] and the widespread availability of digital health measurement instruments [6,7,23,24], the integration of these instruments into daily practice lags behind due to implementation barriers in various domains [6,11,21,22]. These barriers include technical, organizational, legal, and financial challenges as well as deficits in knowledge, skills, training, and time [6,20,25,26]. Furthermore, the sheer number of available digital health measurement instruments can itself pose an implementation barrier. The literature indicates that health care professionals, regardless of discipline or type of digital health technology, feel overwhelmed by the growing number of available instruments [16,24,27]. An informal survey among primary health care professionals in the Netherlands—comprising physical therapists, occupational therapists, speech and language therapists, and GP practice assistants—confirmed this sentiment for digital health measurement. Consequently, the clinical decision-making process, which involves goal-oriented identification, evaluation, and the selection of appropriate digital health measurements tailored to specific purposes and contexts, becomes challenging [27]. Hence, there is an urgent need to develop support strategies to address these barriers and promote the sustainable implementation and meaningful use of digital health measurement in daily practice [22,28,29].

Objectives

While some resources are available to support primary health care professionals in the Netherlands with aspects of digital health use, no overarching intervention currently exists that (1) enables the implementation of digital health measurement across primary health care disciplines and regardless of use context, (2) integrates relevant strategies and resources to guide professionals in assessing whether and how to use digital health measurement to inform and enhance patient management in real-world practice, and (3) is freely accessible to all intended users. Furthermore, the appropriate format and content for such an intervention—tailored to seamlessly integrate into daily clinical routines and address key barriers faced by primary health care professionals—remain undefined. Therefore, we aim to determine the optimal format and content for an intervention to support sustainable and goal-oriented digital health measurement in primary practice and to describe and reflect on the intervention's development process. The study is reported in accordance with the SRQR (Standards for Reporting Qualitative Research) checklist [30].

Methods

Design

To align with the needs of primary health care professionals, the intervention was developed using a bottom-up approach rooted in participatory action research (PAR) [31]. PAR entails equitable collaboration between researchers and stakeholders [32]. It involves iterative, overlapping cycles of planning and implementing change, followed by evaluation, reflection, and adjustment. PAR aims to both generate knowledge and drive meaningful practice change [31,33-36]. We used PAR to identify preferred strategies and resources for supporting the

XSL•FO
RenderX

implementation of digital health measurement while enhancing its sustainable and goal-oriented use in daily practice. In this study, digital health measurement refers to (1) the use of digital questionnaires, electronic patient-reported outcome measures, measurement tools, apps, wearables, and websites in practice settings or at patients' homes (2) by primary health care professionals alone or with their patients (3) to measure health conditions, body functions, body structures, activities, participation, and environmental or personal factors [37] (4) for diagnostic, treatment, and evaluation purposes [12,38].

To ensure that the intervention is user-friendly and tailored to real-world contexts, user-centered design (UCD) principles were incorporated into the PAR approach [39,40]. UCD emphasizes considering implementation conditions from the outset and testing easily understandable prototypes in their intended context. User involvement spans 4 UCD stages: concept, design, testing and trials, and deployment [39,41]. This paper focuses on the first 3 stages, as they form the foundation of the intervention's development. Deployment will be addressed in a subsequent evaluation.

The study is part of a broader PAR initiative that also investigates the intervention's implementation, effectiveness, and impact in real-world primary care settings. Within this broader framework, end users were involved actively not only in designing the intervention but also in co-defining the research problem, aims, and questions as well as in participating in the study's evaluation. Their level of participation encompassed advisory input (eg, sharing adjustment proposals), partnership (eg, codeciding on intervention format), and control (eg, defining relevant support needs and content) [42,43].

Ethical Considerations

All procedures involving human participants were reviewed and approved by the local medical research ethics committee, Medisch Ethische Toetsings Commissie Zuyderland (METCZ20220049 and METCZ20220112). All participants provided written informed consent after receiving detailed written information about the study (in Dutch), including their right to withdraw at any time without consequence. All data were pseudonymized by replacing identifiable information—such as the names and locations of health care professionals and practices—with unique codes in both the transcripts and contextual datasets. Participants received partial financial compensation only for the time invested in study activities that extended beyond their regular professional responsibilities. No images or other media containing identifiable individual participants are included in this manuscript or supplementary materials.

Participants

PAR Team

Typically, in PAR, researchers and those participating in the research cooperate as a team [32,33,35]. In this study, the initial PAR team consisted of primary health care professionals, participating as end users, and researchers from Zuyd University of Applied Sciences, Utrecht University of Applied Sciences, and Maastricht University. All researchers brought extensive and diverse expertise relevant to the study. As the intervention evolved, additional specialists, including a communication expert and website developers, joined the team to meet emerging needs.

Recruitment of End Users

To promote the applicability of the intervention across Dutch primary health care, recruitment aimed to ensure diversity in both health care disciplines and the characteristics of practices and end users. Three quartets were to be formed, each consisting of 1 practice from 4 distinct health care disciplines: physical therapy, occupational therapy, speech and language therapy, and GP practice assistance. These disciplines were selected because they commonly use patient measurements to guide treatment, have comparable roles and responsibilities in delivering individualized care, and collaborate interprofessionally within Dutch primary care. Collectively, they represent a subset of disciplines with the highest volume of care delivery, the most frequent patient contact, and some of the fastest professional growth within the Dutch primary care system [44]. While their roles are comparable and span both prevention and treatment, their clinical foci differ. GP practice assistants—a discipline introduced in the Netherlands in 1999 and less commonly known internationally—independently perform basic clinical tasks previously carried out by GPs, referring patients back to the GP when more complex care is required [45,46]. With 77% to 80% of their time dedicated to patient-specific activities, their role is primarily clinical. Practices were eligible for inclusion if they were located near the main research facility in the Dutch province of Limburg and willing to actively engage in intervention development and testing. Quartets were to be enrolled sequentially, allowing each to build on and validate insights from earlier enrollees.

Before the study, quartets 1 and 2 were purposively recruited [47] through the researchers' professional networks based on health care discipline and variation in practice and professional characteristics. During the study, quartet 3 was also purposively recruited through the networks of both researchers and end users. Practices were assigned to quartets based on discipline and availability. The practices independently determined which and how many of their health care professionals would participate.

External Experts

An advisory board provided feedback throughout all stages of intervention development. The board consisted of representatives from key national professional associations [48-51], a human movement scientist, and 2 professors specialized in digital health care innovations and the implementation of clinical guidelines in general practice. In addition, human factors engineers, lecturers, students, researchers, and primary health care professionals from Zuyd University of Applied Sciences, Utrecht University of Applied Sciences, Windesheim University of Applied Sciences, and Maastricht University provided feedback based on their areas of expertise at various points during the development process. This collaborative effort aimed to broaden perspectives and address end users' time constraints.

XSL•FO
RenderX

Data Collection and Analysis

Overview

Data were collected and analyzed between January 2022 and December 2023 during multiple sessions at various sites in the Netherlands. The following participant characteristics were documented: end users' sex, age, work experience, and technology user personas [52] as well as practices' size (number of staff and locations), socioeconomic status (SES) score, and expertise.

Throughout all UCD stages, participants' perspectives on the planned intervention were collected and analyzed using various approaches tailored to the specific objectives of each stage, the type of data required, the depth of analysis warranted, the emerging group dynamics, and time constraints; for instance, during the concept and design stages, sessions were intentionally loosely structured, with data primarily collected through forms, minutes, and photographs. This approach aimed to foster creativity, minimize inhibition among participants, and support an egalitarian researcher-participant relationship. During the testing and trials stage, data collection focused on testing instead of cocreating and thus primarily relied on verbatim transcripts to allow for more rigorous analysis. In a few cases, the need to ensure inclusive input and accommodate participant availability led to the unanticipated addition of individual sessions.

Data analysis was performed through deductive-inductive content analysis, descriptive statistics, or predefined criteria. The initial deductive coding drew on theoretical models, commonly used categories in relevant literature, team expertise, and study priorities to define broad, non–topic-specific categories, leading to categories such as skeleton plane [53,54], barriers, facilitators, support questions, and professional journey. Inductive coding was then used to refine or expand these with content-specific categories or subcategories based on participant input—such as data presentation, website title, and user-input features—ensuring both theoretical grounding and contextual relevance. The aim was to collect and analyze data until saturation was reached. Saturation was determined by reflexive discussions among the researchers during data collection and analysis, based on data saturation (repetition of data) as well as meaning saturation (adequacy and completeness of themes as well as their meaning and coherence with other themes) [55,56]. Detailed descriptions of data collection and analysis per stage are provided in the following subsections.

Concept Stage

The concept stage aimed to generate an initial outline for intervention development. Two iterations were conducted (Table 1). In the first iteration, the "six thinking hats" method [57-59] was used by end users and the advisory board during a cocreative session to share experiences on digital health measurement as well as intervention format and content needs. This method invites participants to adopt different perspectives (eg, benefits, caution, and creativity)—symbolized by 6 colored hats—to encourage structured and multifaceted input. Participants were prompted to metaphorically switch hats to reflect on the topics from different angles. This session was augmented by individual conversations to deepen insights and ensure input from all end users.

XSL•FO
RenderX

Table 1. Aims, methods, data collection methods, participants, and analysis approaches for each iteration across the concept, design, and testing and trials stages.

Stages and iterations and aims	Methods	Data collection methods	Participants	Analysis
Concept stage				
Iteration 1				
Gaining insight into digital health measurement experiences and intervention format and content needs	"Six thinking hats" [57,59] method, sensitizing preparatory questions	Minutes and forms	End users (quartet 1) and advisory board	Deductive-inductive content analysis [60]
Gaining insight into digital health measurement experiences and intervention format and content needs	Semistructured individual conversations	Minutes	End users (quartet 1)	Deductive-inductive content analysis [60]
Iteration 2				
Refining intervention content needs to create low-fidelity prototype 1	MoSCoW[a] method [59,61], brainstorm session, and plenary discussion	Photographs, forms, and minutes	End users (quartet 1)	Predefined MoSCoW criteria
Refining intervention format needs to create low-fidelity prototype 1	Plenary discussion	Minutes	End users (quartet 1)	Deductive-inductive content analysis [60]
Refining intervention content needs to create low-fidelity prototype 1	MoSCoW method and plenary discussion	Minutes	Advisory board	Predefined MoSCoW criteria
Refining intervention format needs to create low-fidelity prototype 1	Plenary discussion	Minutes	Advisory board	Deductive-inductive content analysis [60]
Design stage				
Iteration 1				
Enhancing visual aesthetics, structure, and content of medium-fidelity prototype 1	Independent website visit, content review[b], and semistructured individual conversations	Reviewed documents and minutes	End users (quartet 1)	Deductive-inductive content analysis [60]
Enhancing visual aesthetics, structure, and content of medium-fidelity prototype 1	(1) Independent website visit and plenary discussion or, based on availability of participants, (2) independent website visit, sensitizing preparatory questions, and semistructured individual conversations	(1) Minutes or (2) forms and minutes	Advisory board	Deductive-inductive content analysis [60]
Enhancing visual aesthetics, structure, and content of medium-fidelity prototype 1	Website demonstration and plenary discussion	Minutes	End users (quartet 2)	Deductive-inductive content analysis [60]
Iteration 2				
Enhancing visual aesthetics, structure, and content of medium-fidelity prototype 2	Codiscovery[c] [62,63], SUS[d] [64], and plenary discussion	Forms, questionnaires, and minutes	Session 1: external experts; session 2: end users (quartets 1 and 2); session 3: external experts	Forms and minutes: inductive content analysis [60]; SUS: descriptive statistics

XSL•FO
RenderX

Stages and iterations and aims	Methods	Data collection methods	Participants	Analysis
Enhancing visual aesthetics, structure, and content of medium-fidelity prototype 2	Codiscovery and plenary discussion	Forms, logs and minutes	Advisory board and external experts	Inductive content analysis [60]
Enhancing visual aesthetics, structure, and content of medium-fidelity prototype 2	Plenary discussion	Minutes	External experts	Inductive content analysis [60]
Enhancing content of medium-fidelity prototype 2	Content review[e]	Reviewed documents	End users (quartet 1)	Inductive content analysis [60]
Testing and trials stage				
Iteration 1				
Optimizing usability and persuasive design principles of high-fidelity prototype 1	Heuristic evaluation	Minutes and quotes	External experts	Summary for each heuristic, illustrated by quotes, and matched with recommendations for improvement
Optimizing surface and skeleton of high-fidelity prototype 1 [53]	Individual usability testing and SUS	Verbatim transcripts of audio recordings and questionnaires	End users (quartets 1 and 2) and external experts	Verbatim transcripts: deductive-inductive content analysis [60]; SUS: descriptive statistics
Optimizing content of high-fidelity prototype 1	Website[f] use in daily practice supported by coaching on the job and focus group discussion	Minutes (coaching on the job) and verbatim transcripts of audio recordings (focus group discussion)	End users (quartets 1, 2, and 3)	Directed content analysis [65]
Iteration 2				
Optimizing structure and scope of high-fidelity prototype 2 [53]	Individual usability testing and SUS	Verbatim transcripts of audio recordings and questionnaires	End users (quartets 1, 2, and 3)	Verbatim transcripts: deductive-inductive content analysis [60]; SUS: descriptive statistics
Optimizing content of high-fidelity prototype 2	Independent prototype use and focus group discussion	Verbatim transcripts of audio recordings	End users (quartets 1, 2, and 3)	Directed content analysis [65]

[a]MoSCoW: must have, should have, could have, would have.

[b]Content pages in Microsoft Word or Microsoft Excel format, part 1.

[c]During the codiscovery sessions, pairs of participants were asked to write down their first impressions, actions, and experiences while first navigating freely through the website and then solving a practice-based measurement case using the website.

[d]SUS: System Usability Scale.

[e]Content pages in Microsoft Word or Microsoft Excel format, part 2.

[f]Quartet 1 used medium-fidelity prototype 3, quartet 2 used medium-fidelity prototype 3 and high-fidelity prototype 1, and quartet 3 used high-fidelity prototype 1.

In the second iteration, group meetings were conducted to refine content and format needs. The MoSCoW (must have, should have, could have, would have) method [58,59,61] was used by end users and the advisory board to prioritize content needs. Needs rated as “must have” by at least 1 participant or “should have” by at least 2 participants were considered crucial for the intervention. Subsequently, end users were asked to identify content categories relevant to their needs during brainstorming sessions, followed by a plenary discussion. Finally, end users and advisory board members were invited to deliberate on further preferences for the intervention format. As an outline for intervention development, content needs and matching content categories were organized in low-fidelity prototype 1.

To support end users to take implementation conditions into account during the design stage, the researchers also created the initial version of a professional journey [59] that linked intervention use to clinical practice.

Design Stage

To foster conceptual understanding and consensus on the intervention’s scope and content while following UCD principles, the researchers developed realistic medium-fidelity prototypes early in the design stage. As end users preferred a web-based intervention format, a communication specialist was integrated into the PAR team to support prototype development. From this stage onward, all prototypes consisted of a website with a structured plan for digital health measurement at its core, accompanied by content pages and progressively added functionalities. Each prototype underwent initial evaluation and refinement by the researchers before presentation to end users or external experts.

XSL•FO
RenderX

In 2 iterations (Table 1), the design stage centered on enhancing the visual aesthetics, structure, and content of 2 subsequent medium-fidelity prototypes to fit the intervention's objectives. In the first iteration, end users and members of the advisory board participated in individual conversations (to accommodate availability), plenary discussions, and content reviews to comment on medium-fidelity prototype 1. Furthermore, the initial professional journey was reviewed, and revised into its final version (Multimedia Appendix 1). In the second iteration, medium-fidelity prototype 2 was assessed by end users, advisory board members, and external experts at the 3 collaborating universities of applied sciences using codiscovery [62,63], plenary discussion, and content reviews. In 3 sessions, participants also completed the System Usability Scale (SUS) [64], which had been translated into Dutch for this purpose. These quantitative data were collected to support qualitative data, and both provided input for a final plenary discussion debating overall experiences.

Two researchers analyzed data from each iteration and deliberated to formulate initial proposals for prototype adjustment based on the analyzed data. These findings were member checked and then discussed among all researchers and the communication specialist until consensus was reached on the changes to implement in the next prototype. Agreed-upon changes were then, where possible, integrated into medium-fidelity prototype 3; for example, when analysis revealed diverging preferences regarding the level of content detail in medium-fidelity prototype 2, the researchers and communication specialist reviewed the corresponding adjustment proposals. Suggestions that were considered to reconcile these differences were selected by consensus and incorporated into the subsequent prototype for evaluation in the next iteration.

Testing and Trials Stage

Overview

In light of the developmental stage and the complexity of some of the required alterations, professional website developers were added to the PAR team. They incorporated all required modifications from the design stage into the initial high-fidelity version of the website. Simultaneously, testing commenced using medium-fidelity prototype 3 to allow for a timely progression of the testing and trials stage.

The testing and trials stage involved 2 iterations (Table 1) primarily aimed at refining the prototype's visual aesthetics, structure, and content and functionalities for practical use and eventual dissemination. This stage included heuristic evaluation, usability testing, and website use in daily practice. Throughout this stage, minor adjustments were continually implemented to the prototype layout.

Heuristic Evaluation

Heuristic evaluation sought to identify potential issues related to both usability and persuasive design principles. Lecturers and second-year students in communication and multimedia design scrutinized the navigation structure, forms, layout, and use of visual elements in high-fidelity prototype 1. They used the 10 usability heuristics formulated by Nielsen [66] as well as the 6 persuasive heuristics formulated by Cialdini [67] to assess the prototype (Multimedia Appendix 2 [66,67]). Subsequently, they correlated the usability and design problems that had been identified with recommendations for improvement.

Usability Testing

The framework for website design developed by Garrett [53] and the study by Crutzen et al [54] were used as guides for usability testing. The framework consists of 5 planes, broadly referring to the following elements: from the bottom up, the strategy plane pertains to the website's objectives and the needs it addresses. This informs the scope plane, which defines the content and functionalities to be incorporated. The structure plane then determines how these elements are linked to work together. The skeleton plane translates this structure into layout by placing elements at specific locations and enabling end users to navigate the website. Finally, the surface plane pertains to the website's visual appearance. User experiences occur at each of these planes but are typically perceived from the top down, starting with the surface plane.

Usability testing involved individual sessions with end users as well as lecturers and researchers. Each session consisted of free navigation through the website, problem-solving while synchronously thinking aloud [59], completion of the translated SUS [64], and an interview. During the first iteration, the interview covered the surface and skeleton planes of the framework developed by Garrett [53]; during the second iteration, it focused on the structure and scope planes.

Website Use in Daily Practice

In the first iteration, end users used either medium-fidelity prototype 3 or high-fidelity prototype 1 in their daily practice to tackle self-selected digital health measurement challenges. Simultaneously, they received coaching-on-the-job support from 2 researchers. In this early stage of intervention development [68], coaching on the job was applied as a temporary implementation strategy to ensure the timely progression of prototype testing and facilitate comprehensive data collection on development issues. The 3 quartets of end users initiated prototype testing sequentially, with each quartet commencing 2 months apart. During the coaching-on-the-job sessions and subsequent focus group discussions per quartet, data were collected on website use experiences, focusing on the content requirements delineated in the scope plane of the framework developed by Garrett [53].

In the second iteration, all end users used high-fidelity prototype 2 independently for 2 months to address any remaining or new digital health measurement challenges within their daily practice. Their experiences were discussed during 2 focus group discussions in which all quartets were mixed.

On the basis of the analyzed results from each iteration and evaluation, the researchers and website developers discussed which prototype adjustments would be suitable and feasible within the study context. These adjustments were incorporated into high-fidelity prototype 2 (adjustments from iteration 1) and the final website version (adjustments from iteration 2).

XSL•FO
RenderX

Trustworthiness

Efforts were made to safeguard the credibility, confirmability, transferability, and dependability of the research [69-71]. Researchers and end users worked together on a regular basis during a 2-year period. Several researchers had prior professional or academic experience in the health care disciplines under study and had investigated the implementation of health measurement in primary care over multiple years. This insider perspective enriched the understanding of the context in which the intervention was to be used. Credibility was also bolstered by using various data sources, methods of data collection, and investigator triangulation. All data were collected, analyzed, and interpreted by a minimum of 2 investigators. Throughout the research process, findings were continually shared with end users and external experts for validation through real-time summaries and member checking of minutes and transcripts [69]. In addition, interpretations and conclusions derived from the data underwent member checking in follow-up sessions that involved oral presentations and the presentation of subsequent prototypes. Peer debriefing and reflexive notes, for instance on insider and outsider positionality, were used to augment confirmability. To mitigate potential limitations associated with positionality, researchers with more outsider perspectives—unfamiliar with the specific health care disciplines or settings—were intentionally included. This diversity also extended to differing levels of experience with and attitudes toward digital health among all PAR team members, ensuring a broader range of insights throughout the study. To facilitate transferability, a “rich description” of the characteristics of the end users and their practices was provided. To ensure the dependability of the research findings, an emergent research design was adopted, with decisions on the next steps based on earlier results. The sequential inclusion of quartets allowed for iterative data collection and analysis.

Results

Participant Characteristics, Data Collection, and Data Analysis

A total of 19 health care professionals across 11 practices were recruited as end users (Tables 2 and 3). These participants engaged their colleagues by sharing experiences with intervention use, encouraging them to use the intervention, and incorporating peer feedback into its further development. In total, 89 health care professionals were involved through this process of peer engagement. Quartets 1 and 2 each included 4 practices representing all 4 health care disciplines under examination. Quartet 3 incorporated 3 practices and health care disciplines. Despite extensive recruitment efforts, a GP practice assistant could not be enlisted in quartet 3. The reasons given for inability to participate included workload constraints and participation in concurrent research projects. Table 2 demonstrates that the recruited end users displayed a variety of ages, work experiences, and experiences and attitudes toward digital health. In addition, the sample represented a range of practice sizes, expertise, and SES scores based on the practice’s postcode areas (Table 3). Two end users (the speech and language therapist in quartet 1 and the GP practice assistant in quartet 2) withdrew from the study—one at the beginning of the testing and trials stage and one midway through the testing and trials stage—due to organizational challenges (significant staff turnover) and health-related reasons. These individuals represented 2 of the 11 participating practices (18%) and 2 of the 19 end users (11%).

XSL•FO
RenderX

Table 2. Demographic, occupational, and technology-related characteristics of the end users (N=19).

Quartets[a] and end users (technology user persona [52])	Sex	Age (y)	Work experience in health care discipline (y)
Quartet[b] 1			
PT[c]1.1 (digitally proficient professional)	Male	66	43
PT1.2 (digitally proficient professional)	Male	28	9
PT1.3 (digitally proficient professional)	Female	24	5
PT1.4 (digitally proficient professional)	Female	38	17
PT1.5 (hesitant technology user)	Female	31	10
OT[d]1 (digital enthusiast)	Female	43	20
SLT[e]1 (digital enthusiast)	Male	63	9[f]
GP[g] practice assistant 1 (digitally proficient professional)	Female	51	16
Quartet 2[h]			
PT2.1 (digital enthusiast)	Female	24	2
PT2.2 (hesitant technology user)	Female	37	16
OT2 (digital enthusiast)	Female	36	14
SLT2 (hesitant technology user; digitally proficient professional)	Female	54	30
GP practice assistant 2 (digitally proficient professional)	Female	64	16
Quartet 3[i]			
PT3.1 (digitally proficient professional; digital enthusiast)	Male	29	7
PT3.2 (hesitant technology user)	Female	45	25
OT3.1 (digitally proficient professional)	Male	26	5
OT3.2 (digitally proficient professional)	Male	27	5
SLT3.1 (digital enthusiast)	Female	42	21
SLT3.2 (digitally proficient professional)	Female	49	26

[a]The term *quartet* refers to the number of primary health care practices and health care disciplines that were intended to be included in each group. It does not reflect the number of end users in each group.

[b]Quartet 1's median age and work experience in health care discipline (years) are 40.5 (range 24-66) and 13 (range 5-43), respectively.

[c]PT: physical therapist.

[d]OT: occupational therapist.

[e]SLT: speech and language therapist.

[f]SLT1's work experience refers to their role as a practice manager in an SLT practice.

[g]GP: general practitioner.

[h]Quartet 2's median age and work experience in health care discipline (years) are 37 (range 24-64) and 16 (range 2-30), respectively.

[i]Quartet 3's median age and work experience in health care discipline (years) are 35.5 (range 26-49) and 14 (range 5-26), respectively.

XSL•FO
RenderX

Table 3. Organizational and contextual characteristics of the participating primary health care practices (N=11).

Quartets and practices	Health care professionals employed, n	Practice locations, n	SES[a] scores[b]	Expertise
Quartet 1				
Physical therapy	Physical therapists: 10	5	Range[c] −0.543 to 0.209	Pelvic health, manual therapy, pediatrics, geriatrics, neurology, COVID-19, edema, medical taping, intermittent claudication, preoperative care, rehabilitation, vocational rehabilitation, and lifestyle intervention
Occupational therapy	Occupational therapists: 3	2	0.122 and 0.135	Workplace ergonomics, pediatrics, neurology, compression stockings, mental health, chronic pain, and COVID-19
Speech and language therapy	Speech and language therapists: 7	20	Range −0.625 to 0.135	Voice, speech, language, swallowing, hearing, orofacial myology, dyslexia, and dyscalculia
General practitioner practice assistance	General practitioners: 3; general practitioner in training: 1; general practitioner's assistants: 6; and general practitioner practice assistants: 5	1	0.028	General care and chronic care (diabetes mellitus type 2, cardiovascular diseases, geriatrics, mental health, and pulmonary conditions)
Quartet 2				
Physical therapy	Physical therapists: 8; occupational therapists: 3; speech and language therapists: 4; osteopath: 1; and exercise therapist: 1	4	Range −0.439 to 0.215	Oncology, pulmonary conditions, neurology, Parkinson disease, chronic pain, edema, orthopedics, headache, hand therapy, intermittent claudication, and holistic coaching
Occupational therapy	Occupational therapists: 12	1	−0.248	Sensory processing, dementia, neurology, pediatrics, osteoarthritis, and rheumatism
Speech and language therapy	Speech and language therapist: 1	1	−0.469	Speech, language, and dyslexia
General practitioner practice assistance	General practitioners: 3; general practitioner's assistants: 4; and general practitioner practice assistants: 3	1	0.061	General care and chronic care (diabetes mellitus type 2, cardiovascular diseases, geriatrics, mental health, and pulmonary conditions)
Quartet 3				
Physical therapy	Physical therapists: 7	1	−0.310	Geriatrics, oncology, edema, hand therapy, pulmonary conditions, manual therapy, pelvic health, headache, (chronic) pain, orthopedics, osteoarthritis, neurology, cardiovascular diseases, postoperative care, and lifestyle intervention
Occupational therapy	Occupational therapists: 2	2	−0.258 and 0.036	Mental health, occupational health and prevention, pediatrics, oncology, neurology, dementia, COVID-19, osteoarthritis, and rheumatism
Speech and language therapy	Speech and language therapists: 5	3	Range −0.354 to 0.045	Voice, speech, language, breathing, swallowing, globus pharyngeus, neurology, facial palsy, and dyslexia

[a]SES: socioeconomic status.

[b]The average scores for the practice locations' postcode areas [72] are shown. The score for a postcode area is the average of all household scores within that area. The average SES-WOA (Wijk- en Omgevingsanalyse; Dutch for Neighborhood and Environmental Analysis) score for the Netherlands is 0. Scores typically range from −2 to 1. Higher scores indicate higher wealth, education, and employment duration.

[c]The range represents the lowest and highest average SES scores of the practice's locations if it has >2 locations.

The duration of data collection sessions varied: 30 to 75 minutes for usability testing and individual interviews, 30 to 120 minutes for group discussions, 60 to 120 minutes for cocreative sessions and codiscovery, and 240 minutes for heuristic evaluation. Data were collected and analyzed from all participants until responses no longer provided new insights related to the stage's objectives. At the end of each stage, categories resulting from data analysis reflected general agreement on which aspects of the prototype

XSL•FO
RenderX

needed improvement and why, although, in a few instances, there was no consensus on how to execute these improvements. Nonetheless, depth of understanding was deemed sufficient to progress with iterative prototype development [55,56]. Iteration continued until only minor issues were reported, and usability scores were "good to excellent." Multimedia Appendices 3-5 visualize website development by providing illustrative screenshots of key components from 2 iterative prototypes as well as the final website version.

Concept Stage

The data collected through the "six thinking hats" method [57-59] and individual conversations in the first iteration provided insight into barriers, facilitators, content and format needs, and good practices regarding the use of digital health for measurement purposes.

Table 4 summarizes the barriers and facilitators identified as relevant to our bottom-up approach to enhancing digital health measurement. Throughout all developmental stages, prototype format and content were tailored to mitigate these barriers and leverage these facilitators. For instance, adjustments were made to accommodate time constraints at the practice level, ensuring the intervention's efficiency. Furthermore, the intervention content was customized to promote goal-oriented use of digital health measurement, which the end users considered a facilitator for digital health measurement.

Table 4. Barriers and facilitators to digital health measurement in primary care, as identified by end users and advisory board members.

Barriers	Facilitators
Health care professional	
Lack of knowledge and skills concerning several topics (general unfamiliarity; lack of experience; digital skills; patient selection; finding and selecting measurement instruments; goal-oriented use; recording, selecting, and presenting measurement data; and implementation) and attitude	Innovation and oversight generate positive feelings; more targeted coaching and therapy
Patient	
General unfamiliarity; applicability for target audience; attitude; and readiness for digital health measurement (understanding of what is expected)	Added value; quality of care; registration and presentation of measurement data; wide applicability (goals and target audiences); more goal-oriented therapy; and positive experiences
Practice	
Time constraints and excessive transitions in primary care	Efficiency; reputational benefit; and consistency with digital skills within team
Characteristics of digital health measurement instruments	
Supply and clinimetric quality	Possible uses; quality of measurements; functionalities for registration, analysis, communication, and feedback; and efficiency
Policy and outer context	
Legal aspects (privacy and General Data Protection Regulations); little scientific evidence; and lack of resources providing oversight	Consistency with policies of external parties

Content needs shared by end users and members of the advisory board were translated into 23 support questions (Multimedia Appendix 6). Concerning intervention format, end users expressed the need for an interdisciplinary stepwise plan, flowchart, or checklist. These content and format needs were elaborated on during the second iteration of the concept stage and guided prototype development during the design stage. The identified good practices encompassed various aspects of digital health use, namely examples of digital health measurement instruments, implementation processes, benefits for patients, listing sites, and financing options. These good practices were used and expanded on during the testing and trials stage to inspire intervention application in daily practice.

In the second iteration, MoSCoW ratings [59,61] indicated that all 23 support questions were essential for the intervention (Multimedia Appendix 6). The "10-step plan for integrated use of measurement instruments in practice" [14,15] was used to organize these support questions, as the plan aligned closely with both the content and format needs from the first iteration. Together, end users and researchers identified content categories to address the support questions. This process yielded the low-fidelity prototype of the intervention, which consists of a matrix that organizes the support questions according to the 10-step measurement plan and aligns them with content categories (Multimedia Appendix 6).

End users deliberated on diverse formats for the intervention. Ultimately, they agreed on a web-based intervention:

> *You need to choose a format for the intervention that has appeal: a website or app to generate enthusiasm among employees. It should be systematic and scientific, but wrapped in an attractive package.* [Physical therapist, quartet 1]

> *A hard copy I will put on a shelf, and I won't look at it again, a clickable PDF I will save once and won't come back to, a website I can access anytime.* [Occupational therapist, quartet 1]

XSL•FO
RenderX

Design Stage

Overview

The first medium-fidelity prototype consisted of (1) a website home page with a matrix representing the 10-step plan for digital health measurement, linking to layers of basic and in-depth content pages for steps 1 and 2 as well as background information; and (2) Microsoft Word and Microsoft Excel documents containing proposed content for all steps. In the second medium-fidelity prototype, the content of the Word and Excel documents was added to the website's basic or in-depth content layers. The prototypes integrated various types of content, including text, icons, pictures, and hyperlinks to authoritative websites and resources.

The data collected from end users and external experts during the design stage primarily focused on the visual aesthetics, structure, and content of medium-fidelity prototypes 1 and 2. Given the conceptual similarity of the reviewed prototypes across both iterations and the consistent participant feedback on the same topics, the qualitative results from iterations 1 and 2 are clustered per topic. Quantitative data were collected only for medium-fidelity prototype 2 during codiscovery sessions 1 to 3, using the SUS. Of the 14 participants of these sessions, 12 (86%) completed the SUS, resulting in a mean score of 58.3 (SD 11.4), which corresponds to an "okay" usability rating (percentile range 15-34) [73,74].

Visual Aesthetics

Participants widely regarded the overall aesthetics of the website as attractive, clear, and inviting. However, certain icons and images were considered incongruent with digital health and the primary health care context:

> *The first moving image is not illustrative, but it is attractive. The health care professional in that image is not using digital health, but papers. Something along these lines is nice.* [Physical therapist, quartet 1]

The use of bold font to signify key words led to confusion among participants.

Structure

Participants found the website structure clear, effective, and pleasant. They often used the website intuitively, without resorting to the provided instructions. However, they suggested making the stepwise plan more prominent because of its central role in the website's goals and functioning:

> *Personally, I would like the 10 steps to be more explicitly visible without having to scroll. Now, attention is primarily directed to the title and a dynamic picture.* [Professional association member]

In addition, the flexibility for end users to either follow all steps or select specific ones was not immediately apparent. Furthermore, participants indicated a need for functionalities that allow experienced users to bypass steps and quickly locate relevant resources.

While participants appreciated the use of layers to differentiate between basic and in-depth information, they experienced disorientation within the website's structure:

> *Lots of opportunities everywhere to continue clicking through the site, with examples and explanations, but you sometimes get lost as a user of the site.* [Researcher]

Participants requested improvements to the text structure as well as for quicker retrieval of relevant information, including revisiting the classification of basic and in-depth content. They also critiqued the use of full URLs for hyperlinks, suggesting keywords, tiles, or explanations for clarity.

Content

Overall, participants considered the website content relevant, comprehensive, and educational and appreciated the usefulness of the resources and their integration within the intervention. However, they suggested improvements to the website title and background information to better align with the intervention's objectives and more clearly delineate digital health measurement.

End users reported that the stepwise plan addressed their support questions from the concept stage. While they found all steps useful, they identified additional relevant information and resources to include. Personal preferences varied regarding the balance of visual and textual information, detail level, and applicability to specific disciplines. Overall, participants felt that the content was too theoretical and necessitated translation into concrete, practical use:

> *The website now comes across as an educational tool to understand where to find what, what digital health is, and what is involved. It's more of a study than a practical real-world tool.* [Speech and language therapist, quartet 2]

> *All very well reasoned, though maybe a little too well. Above all, you want to get started on your own question and it takes a long time to get there.* [Researcher]

Participants recommended using more accessible language, accompanied by practical, interdisciplinary examples. Moreover, they observed an overemphasis on generic measuring concepts, rather than highlighting focus areas for digital health measurement.

In terms of implementation conditions, participants raised concerns about the volume of information being incompatible with end users' time constraints. Finally, they recommended minimizing links to paid resources to enhance user-friendliness.

Integrating all findings, the prototype adjustments aimed to clarify the website's goals and target audiences. They also aimed to balance content depth with user-friendly structuring of information to accommodate flexible and efficient use within the identified implementation conditions. Visual aesthetics, structure, and content of the subsequent medium-fidelity prototypes were revised, as illustrated in Figure 1. The final medium-fidelity prototype streamlined the website into a 5-step plan with overarching themes: formulating a measurement

XSL•FO
RenderX

challenge, finding an appropriate digital health measurement instrument, assessing its quality, using the instrument and using its outcomes for shared decision-making, and implementing the instrument. Each step contained limited and practical content. Other requested adjustments were postponed for further exploration and integration during the testing and trials stage.

Figure 1. Adjustments made to iteratively advance the low-fidelity prototype of the intervention designed to support primary health care professionals in digital health measurement into medium-fidelity prototypes.

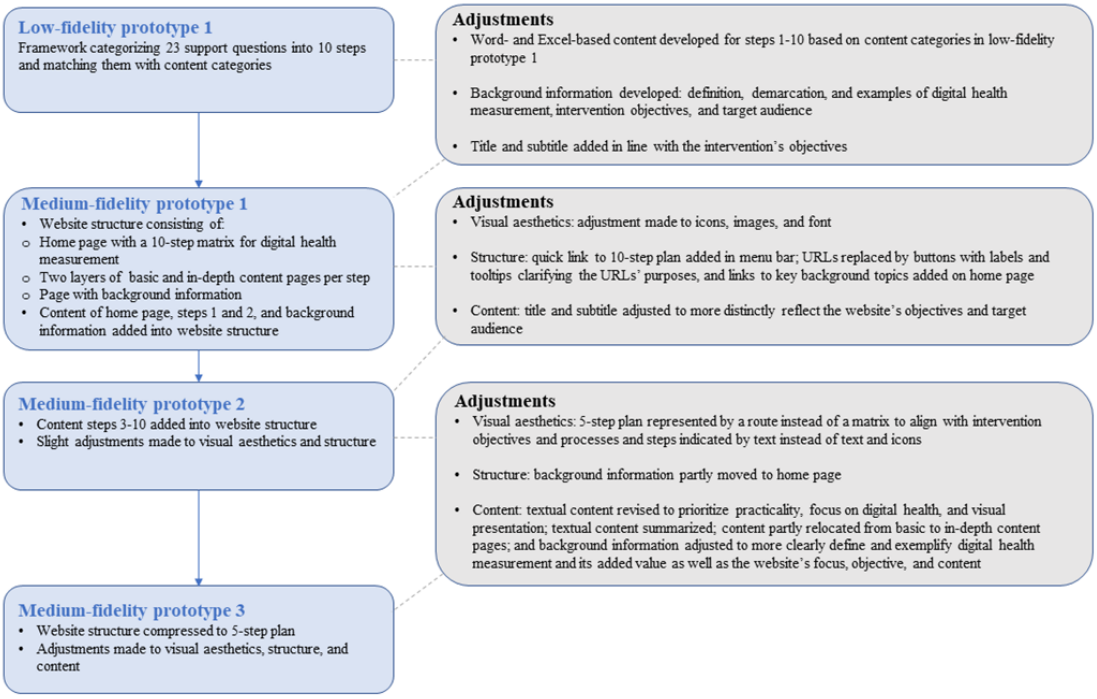

Testing and Trials Stage

In the first iteration, the usability testing of high-fidelity prototype 1 yielded a mean SUS score of 76.6 (SD 5.8; 8/8, 100%), which indicates a "good" rating (percentile range 70-79) [73,74]. Qualitative data pertained to the categories of graphic design, navigation, and content and functionalities.

Visual Aesthetics

During the transition from medium-fidelity prototype 3 to high-fidelity prototype 1, professional website developers refined the graphic design. Participants in heuristic evaluation and usability testing praised the readability and ease of finding information, attributing this to user interface elements such as tooltips, colors, and headings. Tooltips and contrasting colors effectively highlighted supplementary information, thereby enhancing usability. However, participants noted inconsistencies in design across website layers, which disrupted the sense of calm and clarity, and they found it unclear at times whether graphic elements were intended for navigation or information purposes:

> *The design is calm, and the imagery reinforces the text and fits the context.* [Human factors engineer 1]
>
> *Imagery of the home page is calm and functional, but not thereafter.* [Human factors engineer 2]

While participants agreed that images should add value and reflect the website's content and target audience, opinions diverged on which images fulfilled these criteria.

Structure

Heuristic evaluation and usability testing revealed that improving the website's navigation structure—by enabling end users to track progress within the 5-step plan and offering multiple pathways to access information—would significantly enhance usability:

> *And here it would be convenient if I could already click through to the next one here. Because here I have to go back first, right? And then I have to go all the way back again. So, if clicking through from step one to step two could actually be on the same page would also be nice.* [Lecturer]

Participants suggested adding a menu bar to establish a standardized website structure and prominently feature the 5-step plan, emphasizing the website's objectives. In addition, half of the participants considered the inclusion of a search function crucial for improving user-friendliness, usability, and efficiency.

Content and Functionalities

The potential to use social proof and authority in persuasive design was initially overlooked. Human factors engineers recommended integrating end users' personal stories to enhance engagement and emphasizing the involvement of expert health care practices and research institutions in the intervention's development to build trust:

> *Share reviews from others who have used this [intervention] and experienced in practice that they got better at using it [digital health measurement] or*

XSL•FO
RenderX

that it [digital health measurement] is clearer to them now. [Human factors engineer 1]

There is no connection displayed to an organization that offers trust...Is there a link to that kind of party? [Human factors engineer 2]

Usability testing and use in daily practice also revealed further content needs and a desire for increased user input. Participants proposed adding functionalities for sharing reviews of digital health measurement instruments, suggesting website content, tracking progress within the stepped plan, and recording results per step.

The findings from the first iteration were incorporated into high-fidelity prototype 2 (Figure 2). However, 3 adjustment proposals were deemed inapplicable or infeasible for this study. First, the researchers and website developers considered the website content too limited to justify a search function in addition to a menu bar. Instead, screening questions were introduced for identifying relevant content. Second, user-input was achieved by the aforementioned screening questions and by personal stories, instead of functionalities for directly sharing reviews and content. Third and last, recording individual results per step was excluded to avoid log-in requirements, maintaining accessibility and privacy.

Figure 2. Adjustments made to iteratively advance medium-fidelity prototype 3 and the high-fidelity prototypes of the intervention designed to support primary health care professionals in digital health measurement into the final website version.

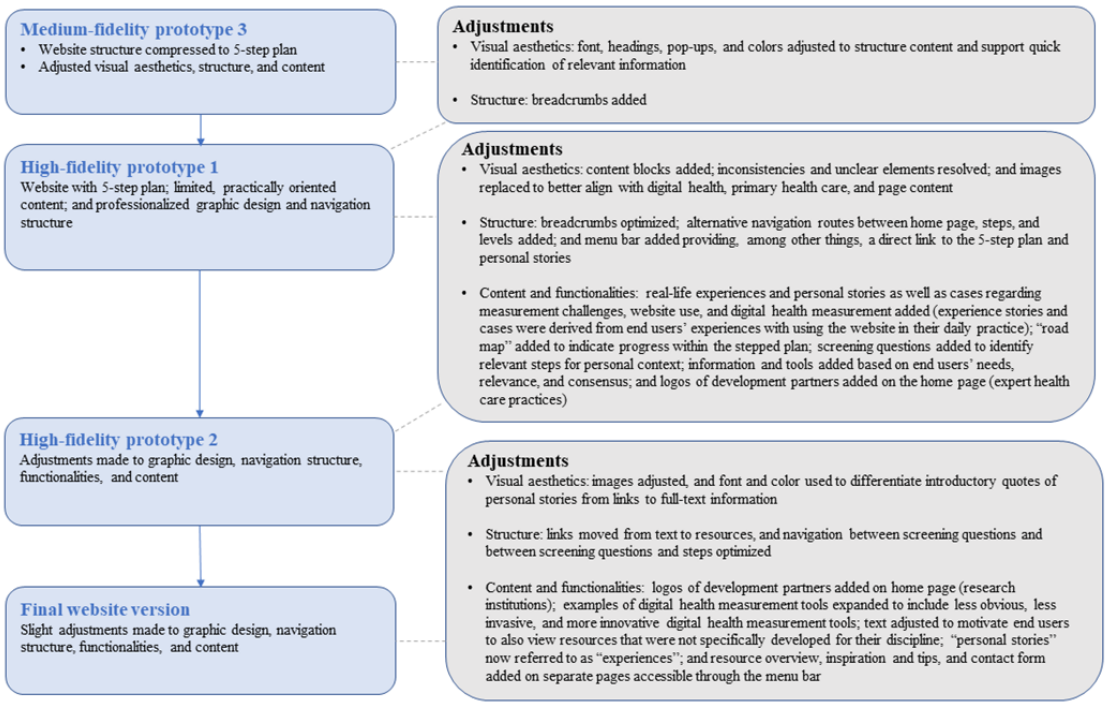

Final Experiences and Adjustment Proposals

During the final iteration, 11 (100%) of the 11 participants to the usability testing completed the SUS for high-fidelity prototype 2, yielding a mean rating of 78.6 (SD 8.8), which indicates a "good to excellent" usability rating (percentile range 80-84) [73,74]. Apart from some minor issues, end users found the prototype easy to navigate, with consistent graphic design that enhanced predictability. The 5-step plan was considered logical and effective to guide clinical decision-making for the implementation of digital health measurement. Most end users described the content as practical, complete, concise, and well balanced between text and visuals:

You do get taken by the hand and the steps are logical and especially if you have a slightly less concrete measurement challenge, then I do find that you can really be helped in what possibilities there are. [Physical therapist, quartet 1]

Newly added functionalities, such as screening questions and personal stories, were valued for directing users to relevant steps; enhancing recognition, motivation, and inspiration; and adding some "lightness" to the website:

Actually, you just can't go wrong if you use this [screening questions]. If you end up in a strange place now, I would find that very odd. [Physical therapist, quartet 2]

It should be an enjoyable website. Having "light" things in between makes it less clinical. It's pleasant. [Speech and language therapist, quartet 2]

If you land on a website, and you are looking for something, then, indeed, to read someone else's stories, that gives you something to hold on to, and yes, maybe it feels more familiar, I don't know, but I think it's a good thing. [Occupational therapist, quartet 3]

That's something that a colleague may have experienced as well, because, we are people from practice...if you see that other people have tried something and they are enthusiastic about it. [General practitioner practice assistant, quartet 1]

XSL•FO
RenderX

While the intervention's generic approach was considered beneficial for interprofessional use, some end users suggested stronger recognition of specific health care disciplines through images or explicit mentions:

> *I would appreciate it if there would be a physical therapist or, you know, more of my profession in it.* [Physical therapist, quartet 2]

Other adjustment proposals concerned adding a separate resource overview and including a blog or forum to share reviews more actively.

The final website version included the requested resource overview. Images were fine-tuned to increase recognition for specific health care disciplines while preserving broad applicability. To avoid content monitoring challenges, the proposed blog or forum was replaced with a page featuring broadly scoped inspiration and tips from end users and a contact form for inquiries and suggestions. End users agreed with these changes.

Figure 2 shows all major and minor adjustments integrated into the final website version, which was published in December 2023 [75] and is depicted in Figure 3.

Figure 3. Screenshots of the final website version designed to support primary health care professionals in digital health measurement. (A) Title, subtitle, and logo. (B) Five-step plan for digital health measurement. (C) Basic content layer for step 2.

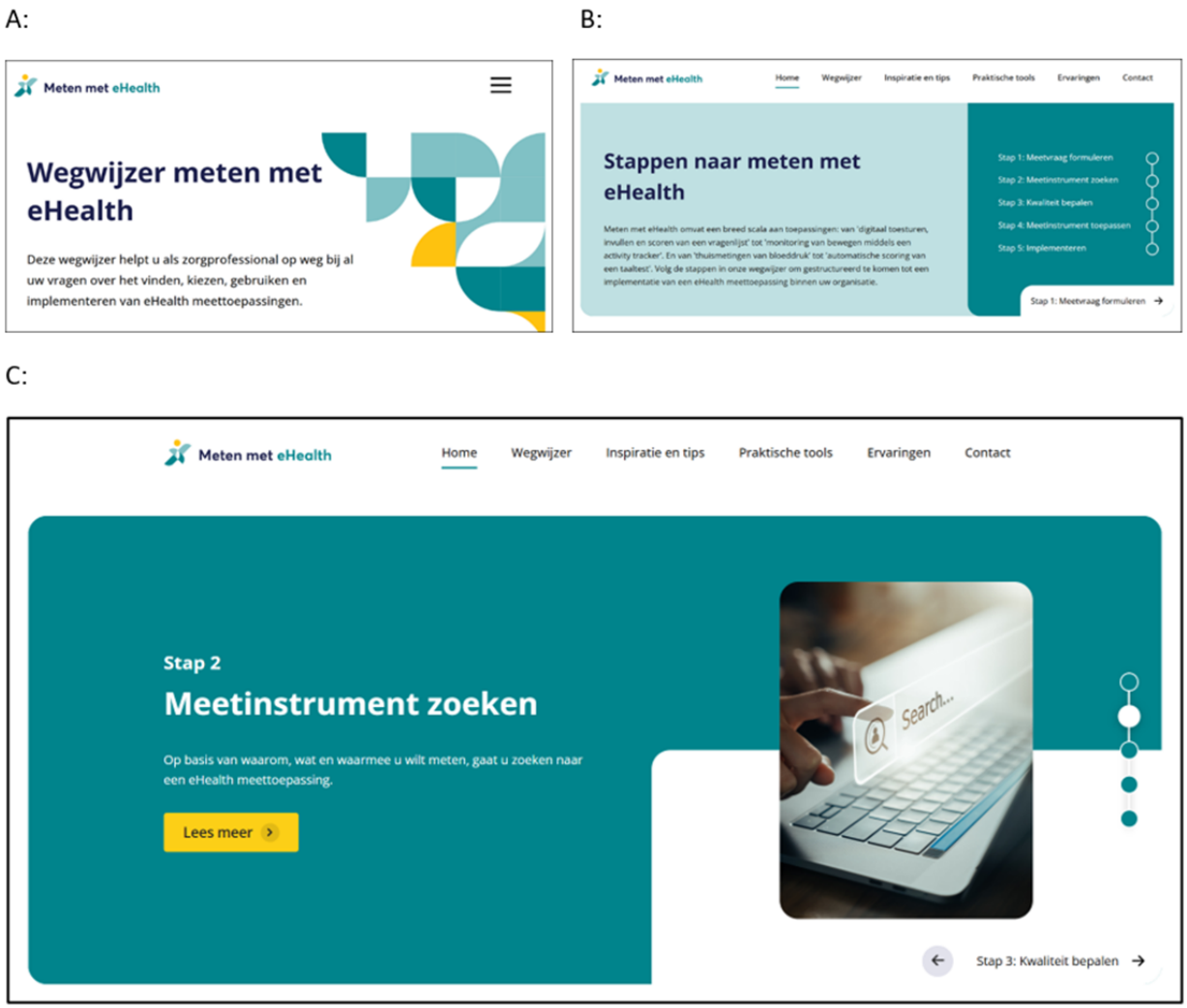

Discussion

Principal Findings

This study used a PAR approach to develop an intervention to support sustainable, goal-oriented digital health measurement in primary practice while also stimulating change within participating practices. PAR prioritizes the integration of lived experiences and practical expertise as key knowledge sources [31,33]. In alignment with this approach, the study foregrounded end users' experiences in identifying barriers, facilitators, and preferred strategies and resources for digital health implementation. Furthermore, researchers' expertise and input from external experts were incorporated to enrich and triangulate these insights with theoretical frameworks.

Major barriers to digital health implementation identified in this study included limited knowledge and skills, time constraints within primary practice, and negative attitudes of health care professionals and patients. Facilitators included compatibility with team skills and external policies, positive attitudes of innovation enthusiasts, and the perceived advantages of digital

XSL•FO
RenderX

health measurement for health care professionals and patients. These barriers and facilitators reflect the capability, opportunity, and motivation required for digital health performance, as outlined in the COM-B model, a scientifically validated behavior change framework that posits that for any behavior (B) to occur, a person must have the capability (C), opportunity (O), and motivation (M) to perform it [29,76,77]. Furthermore, they are consistent with barriers and facilitators reported for digital health implementation in general [6,20,25,26]. However, prior studies have also identified additional factors, including the influence of decision-making processes, improved patient-provider communication, and patients' access to the internet [6,20,25,78]. These differences may be attributed to this study's specific focus on barriers and facilitators experienced by health care professionals when implementing digital health for measurement purposes in Dutch primary care.

End users expressed support needs throughout the entire clinical decision-making process outlined in the "10-step plan for integrated use of measurement instruments in practice" [14,15]. However, to ensure the intervention's feasibility, these 10 steps were consolidated into 5 overarching steps with concise and practical content. Engaging health care professionals in digital health implementation and enabling them to break the implementation process into smaller, manageable steps are shown to be promising strategies to improve capability and motivation for digital health performance [29]. End users identified a web-based, open-access format as the most suitable for delivering the intervention content. To meet their needs, graphic design, navigation and functionalities needed to be designed to ensure flexibility, personalized use, and quick identification of relevant information. This preference aligns with existing research showing that health care professionals favor information resources that are efficient, easy to use, and accessible without requiring a password or account, provided that the information is accurate [79,80]. To enhance trust in the accuracy of the website content, the website prominently displayed the practices and experts involved in its development—a strategy that may also improve digital health competency [29].

The intervention combined the website, which featured theoretical information, practical aids, examples, personal stories, and experiences, with coaching on the job. Together, these intervention components encompassed multiple strategies—education, persuasion, skills training, modeling, enablement, and empathic support—that are recognized as appropriate for enhancing capability, opportunity, and motivation for digital health performance [29,76,77]. However, while there is support for the effectiveness of both the website and coaching on the job stimulating digital health performance [29,76], in this study focusing on website development, coaching served as a temporary implementation strategy for the website and facilitated data collection for prototype development. Given its time-consuming nature, coaching may pose greater challenges for dissemination compared to the website and may need to be replaced by alternative implementation strategies that require less time.

The development process combined PAR and UCD principles to ensure that the intervention met the needs and routines of primary health care professionals. End-user involvement and iterative processes were essential throughout the reported UCD stages to enable the development of a user-friendly and suitable intervention. At the concept stage, comprehensive content needs were identified, aligning in part with core competencies and skills reported in previous research [81]. During the design stage, the necessity to focus on key content became evident, while the testing and trials stage highlighted the content that was essential. The emergent nature of the research design also proved invaluable. It allowed the PAR team to adapt to end users' responses, such as addressing conceptual consensus issues by crafting a professional journey and strategically integrating disciplines into the team to facilitate the creation of early web-based prototypes.

Strengths and Limitations

This study has several strengths. The study design facilitated flexible methods, active involvement of relevant stakeholders at all stages, and direct impact in participating practices [31,33-36,39,40,58,59]. At the same time, the fluid design posed challenges for maintaining consistency across data collection, analysis, and interpretation with respect to the research objectives and priorities. To ensure qualitative rigor, the study implemented triangulation, prolonged engagement, member checking, peer debriefing, and reflection, all of which are notable strengths. The intervention's broad applicability across health care disciplines, digital health measurement technologies, and the entire clinical decision-making process is unique and aligns with evolving professional competencies and future health care developments [8,82,83]. To enhance transferability of the intervention within the Dutch primary health care context—particularly to disciplines that regularly conduct patient measurements to guide treatment, such as midwifery care and community nursing—the study intentionally included practices from 4 distinct health care disciplines, emphasized the interprofessional aspects of clinical decision-making, and used a generic framework [14,15]. To support scalability from the regional to the national level, the study incorporated a broad range of practice settings and engaged key national professional associations affiliated with the participating disciplines.

A limitation of this study was its reliance on photographs, forms, logs, reviewed documents, and minutes—rather than verbatim transcripts—for data collection and analysis during cocreative sessions in the concept and design stages [31,84,85]. This approach was adopted to encourage open expression and support a collaborative, nonhierarchical atmosphere between researchers and other participants. To enhance credibility, the minutes, interpretations, and conclusions underwent member checking through document reviews, real-time summaries, oral presentations, and iterative prototype feedback. Moreover, during the testing and trials stage, verbatim transcripts and quantitative questionnaires were used, contributing to extensive method triangulation. In addition, the sample did not include the "analog idealist" persona [52] because end users were recruited based on their willingness to enhance digital health measurement in daily practice. Although this limits the intervention's applicability for this group, it aligns with the intervention's objective of supporting primary health care professionals who are open to using digital health measurement.

XSL•FO
RenderX

Persuading more reluctant health care professionals might require alternative strategies, such as incentivization [76,77], which were beyond the study's scope. Another limitation was the underrepresentation of GP practice assistants compared to other primary health care disciplines. This was partially offset by the significant involvement of nursing lecturers and researchers among the external experts.

Future Work

Although future research remains essential to guide broader adoption and scaling, the effectiveness of the intervention in fostering sustainable and goal-oriented digital health measurement by primary health care professionals—as well as the contribution of both the website and coaching on the job during this phase of intervention development—is already under study [68]. As successful intervention use in daily practice is influenced by both implementation processes and contextual factors [86], the ongoing combination of process and outcome evaluations is expected to yield meaningful insights into these dynamics and their impact on the use of goal-oriented digital health measurement. Moreover, as coaching may be difficult to disseminate due to constraints in time and personnel, the evaluation also examines its mechanisms of impact to inform the development of more scalable implementation strategies. The 4 included health care disciplines were selected to reflect a broader group of primary health care professionals with similar roles and responsibilities in measurement-based care. Future studies could explore whether adaptations are needed for specific professional groups. In addition, several practical needs and suggestions raised by end users were beyond this study's scope and merit further investigation. These include developing an interdisciplinary database of digital health measurement instruments, facilitating the interprofessional use of digital health measurement, and exploring the intervention's relevance for pre- and postlicensure education as well as secondary care.

Conclusions

This study provides detailed insights into primary health care professionals' format and content needs for an intervention to overcome barriers and leverage the potential benefits of sustainable and goal-oriented digital health measurement. Stepwise, easily accessible, visually appealing, and concise delivery of practical content—spanning the entire clinical decision-making process—proved crucial to facilitating the intervention's integration into daily practice.

The development process described offers a structured approach to creating support strategies for digital health measurement in other contexts. Its PAR approach and UCD principles effectively engaged primary health care professionals in the cocreation of these strategies. Comparing the study's findings with existing research and theoretical perspectives highlights significant alignment. This reinforces the relevance of these theories in practice while underscoring the value of end-user participation to tailor interventions to specific contexts.

Moreover, the study demonstrates that different health care disciplines and technology user personas can benefit from similar support strategies. Achieving this broad applicability requires a careful balancing of design and content.

Acknowledgments

The authors sincerely thank all participating primary health care professionals from ICL Fysio, Ergotherapiepraktijk Martens, Logopedica, Huisartsenpraktijk Reinoud, ParaMedisch Centrum Zuid, Ergotherapie Midden-Limburg, BLIJ Logopedie, Huisartsenpraktijk Herkenbosch, Fysio Annadal, Mijn Ergo, Logopedie SamenSpraak, communication specialists Veronique Tumini and Hello Sunshine, and colleagues Laura Hochstenbach, Gerrie Bours, Corelien Kloek, Albine Moser, Renée Verwey, and Marissa Gerards. The authors are also grateful to the advisory board members and external experts: Erik van Rossum, Cindy Veenhof, Marike Hettinga, Trudy van der Weijden, Sander Holterman, Loek van der Heide, Joris van der Steen and Anouk Claessens from Koninklijk Nederlands Genootschap voor Fysiotherapie, Lucelle van de Ven from Ergotherapie Nederland, Cindy Koolhaas and Yvonne Brookman from Nederlandse Vereniging voor Logopedie en Foniatrie, and Talisha de Rouw and Jolanda Schaliboom from Nederlandse Vereniging van Praktijkondersteuners en Praktijkverpleegkundigen. This study was funded by the Brightlands innovation program Limburg Meet (DOE MEE project [Duurzaam implementeren van doelgericht meten met eHealth in de eerstelijns zorgpraktijk; Dutch for "Sustainable implementation of targeted measurement with eHealth in primary care practice"]). The funders had no role in the study design, data collection, data analysis, or writing of this manuscript. The generative AI tool ChatGPT [87] was used only for language editing.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

E Bols, E Beekman, RC, KM, and DU conceptualized the study. KM, RS, DU, E Bols, and Laura Hochstenbach curated the data. KM, DU, RS, E Bols, Gerrie Bours, and Laura Hochstenbach were responsible for the formal analysis. E Bols and E Beekman were responsible for funding acquisition. KM, E Bols, E Beekman, RS, PAvdH, DU, Gerrie Bours, Albine Moser, Laura Hochstenbach, and Marissa Gerards were responsible for investigation. KM, DU, PAvdDH, RS, Corelien Kloek, Veronique Tumini, and Hello Sunshine were responsible for intervention development. KM, E Bols, E Beekman, RS, DU, RC, Gerrie Bours, Renée Verwey, Marissa Gerards, and Laura Hochstenbach were responsible for methodology. E Bols, RS, and KM were responsible

XSL•FO
RenderX

for project administration. KM, E Bols, E Beekman, RS, PAvdH, DU, and RC were responsible for resources. E Bols, E Beekman, RS, and RC were responsible for supervision. RC, Erik van Rossum, Cindy Veenhof, Marike Hettinga, and Trudy van der Weijden were responsible for validation. KM and PAvdH were responsible for visualization. KM wrote the original draft. KM, PAvdH, E Bols, RC, E Beekman, RS, and DU reviewed and edited the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Final version of a professional journey that linked intervention use to clinical practice.
[PDF File (Adobe PDF File), 335 KB - formative_v9i1e72331_app1.pdf]

Multimedia Appendix 2
The 10 usability heuristics formulated by Nielsen [66] and the 6 persuasive heuristics formulated by Cialdini [67] used to assess high-fidelity prototype 1.
[PDF File (Adobe PDF File), 202 KB - formative_v9i1e72331_app2.pdf]

Multimedia Appendix 3
Prototype screenshots: medium-fidelity prototypes 1 and 2.
[PNG File , 354 KB - formative_v9i1e72331_app3.png]

Multimedia Appendix 4
Prototype screenshots: medium-fidelity prototype 3.
[PNG File , 415 KB - formative_v9i1e72331_app4.png]

Multimedia Appendix 5
Website screenshots: final version.
[PNG File , 331 KB - formative_v9i1e72331_app5.png]

Multimedia Appendix 6
Low-fidelity prototype 1 (columns 1 and 3), supplemented by MoSCoW (must have, should have, could have, would have) prioritization ratings (column 2).
[PDF File (Adobe PDF File), 713 KB - formative_v9i1e72331_app6.pdf]

References

1. den Broeder L, Couwenberg C, Hilderink H, Polder J. Public health and health services missions in the period to 2050: a look ahead to the Public Health Foresight Study 2024. Rijksinstituut voor Volksgezondheid en Milieu. 2023. URL: https://www.rivm.nl/bibliotheek/rapporten/2023-0408.pdf [accessed 2024-02-05]
2. Universal health coverage (UHC). World Health Organization. 2025 Mar 26. URL: https://www.who.int/news-room/fact-sheets/detail/universal-health-coverage-(uhc) [accessed 2025-05-25]
3. van der Vaart R, van Tuyl LHD, Versluis A, Wouters MJM, van Deursen L, Standaar L, et al. E-healthmonitor 2022. Stand van zaken digitale zorg. Rijksinstituut voor Volksgezondheid en Milieu. 2023. URL: https://www.rivm.nl/bibliotheek/rapporten/2022-0153.pdf [accessed 2024-02-05]
4. van der Kleij RMJJ, Kasteleyn MJ, Meijer E, Bonten TN, Houwink EJF, Teichert M, et al. SERIES: eHealth in primary care. Part 1: concepts, conditions and challenges. Eur J Gen Pract 2019 Oct 10;25(4):179-189 [FREE Full text] [doi: 10.1080/13814788.2019.1658190] [Medline: 31597502]
5. Ministerie van Volksgezondheid Welzijn en Sport. Integraal zorgakkoord: 'samen werken aan gezonde zorg'. Rijksoverheid. 2022. URL: https://www.rijksoverheid.nl/onderwerpen/kwaliteit-van-de-zorg/documenten/rapporten/2022/09/16/integraal-zorgakkoord-samen-werken-aan-gezonde-zorg [accessed 2024-03-04]
6. Stoumpos AI, Kitsios F, Talias MA. Digital transformation in healthcare: technology acceptance and its applications. Int J Environ Res Public Health 2023 Feb 15;20(4):3407 [FREE Full text] [doi: 10.3390/ijerph20043407] [Medline: 36834105]
7. WHO guideline: recommendations on digital interventions for health system strengthening. World Health Organization. 2019. URL: https://iris.who.int/bitstream/handle/10665/311941/9789241550505-eng.pdf?ua=1 [accessed 2024-04-25]
8. Visie eerstelijnszorg 2030. Rijksoverheid. 2024. URL: https://open.overheid.nl/documenten/5a92918b-0fab-4998-bd42-9488e0911de0/file [accessed 2024-03-04]

XSL•FO
RenderX

9. Ross J, Stevenson F, Lau R, Murray E. Factors that influence the implementation of e-health: a systematic review of systematic reviews (an update). Implement Sci 2016 Oct 26;11(1):146 [FREE Full text] [doi: 10.1186/s13012-016-0510-7] [Medline: 27782832]
10. van Gemert-Pijnen JE, Wynchank S, Covvey HD, Ossebaard HC. Improving the credibility of electronic health technologies. Bull World Health Organ 2012 May 01;90(5):323-33A [FREE Full text] [doi: 10.2471/BLT.11.099804] [Medline: 22589561]
11. Keij B, Versluis A, Alblas EE, Keuper JJ, van Tuyl LHD, van der Vaart R. 2023 eHealth Monitor: state of digital care. Rijksinstituut voor Volksgezondheid en Milieu. 2024 Apr 11. URL: https://www.rivm.nl/publicaties/e-healthmonitor-2023-stand-van-zaken-digitale-zorg [accessed 2024-10-10]
12. van Lettow B, Wouters M, Sinnige J. E-Health, wat is dat? Nictiz. 2019 Mar 27. URL: https://nictiz.nl/app/uploads/2022/08/E-health-Wat-is-dat.pdf [accessed 2024-02-05]
13. Digital health. World Health Organization. URL: https://www.who.int/europe/health-topics/digital-health#tab=tab_1 [accessed 2024-02-22]
14. Bols E, Beurskens A, Berghmans B. Selection and goal-oriented application of measurement instruments in pelvic physical therapy. In: Bø K, Berghmans B, Mørkved S, Van Kampen M, editors. Evidence-Based Physical Therapy for the Pelvic Floor, Third Edition. Amsterdam, The Netherlands: Elsevier; 2024.
15. Beurskens S, van Peppen R, Swinkels R, Vermeulen H, Wittink H. Meten in de Praktijk: Stappenplan Voor het Gebruik van Meetinstrumenten in de Gezondheidszorg, Third Edition. Houten, The Netherlands: Bohn Stafleu van Loghum; 2020.
16. Mathews SC, McShea MJ, Hanley CL, Ravitz A, Labrique AB, Cohen AB. Digital health: a path to validation. NPJ Digit Med 2019;2:38 [FREE Full text] [doi: 10.1038/s41746-019-0111-3] [Medline: 31304384]
17. María Gómez A, Cristina Henao D, León Vargas F, Mauricio Muñoz O, David Lucero O, García Jaramillo M, et al. Efficacy of the mHealth application in patients with type 2 diabetes transitioning from inpatient to outpatient care: a randomized controlled clinical trial. Diabetes Res Clin Pract 2022 Jul;189:109948 [FREE Full text] [doi: 10.1016/j.diabres.2022.109948] [Medline: 35700926]
18. Michaud TL, Ern J, Scoggins D, Su D. Assessing the impact of telemonitoring-facilitated lifestyle modifications on diabetes outcomes: a systematic review and meta-analysis. Telemed J E Health 2021 Feb;27(2):124-136. [doi: 10.1089/tmj.2019.0319] [Medline: 32397845]
19. van der Vaart R, van Deursen L, Standaar L, Wouters M, Suijkerbuijk A, van Tuyl L, et al. E-healthmonitor 2021: stand van zaken digitale zorg. Rijksinstituut voor Volksgezondheid en Milieu. 2022 Jan. URL: https://open.overheid.nl/documenten/ronl-1bbc17ad438a6b3361eebc806435e4d3f73fb6af/pdf [accessed 2024-02-05]
20. Peeters JM, Krijgsman JW, Brabers AE, Jong JD, Friele RD. Use and uptake of eHealth in general practice: a cross-sectional survey and focus group study among health care users and general practitioners. JMIR Med Inform 2016 Apr 06;4(2):e11 [FREE Full text] [doi: 10.2196/medinform.4515] [Medline: 27052805]
21. Keuper J, Meurs M, Victoor A, Huygens M, de Groot J, de Jong J. E-health in de huisartsenzorg. Nivel. 2020. URL: https://nivel.nl/sites/default/files/bestanden/1003811.pdf [accessed 2024-03-04]
22. Wouters M, Huygens M, Voogdt H, Meurs M, de Groot J, Lamain A, et al. Samen aan zet! eHealth-monitor 2019. Nictiz en het Nivel. 2019. URL: https://www.nivel.nl/sites/default/files/bestanden/1003604.pdf [accessed 2024-02-05]
23. Aitken M, Clancy B, Nass D. The growing value of digital health: evidence and impact on human health and the healthcare system. IQVIA Institute for Human Data Science. 2017 Nov 7. URL: https://www.iqvia.com/insights/the-iqvia-institute/reports-and-publications/reports/the-growing-value-of-digital-health [accessed 2024-07-02]
24. Larson RS. A path to better-quality mHealth apps. JMIR Mhealth Uhealth 2018 Jul 30;6(7):e10414 [FREE Full text] [doi: 10.2196/10414] [Medline: 30061091]
25. Schreiweis B, Pobiruchin M, Strotbaum V, Suleder J, Wiesner M, Bergh B. Barriers and facilitators to the implementation of eHealth services: systematic literature analysis. J Med Internet Res 2019 Nov 22;21(11):e14197 [FREE Full text] [doi: 10.2196/14197] [Medline: 31755869]
26. Versluis A, van Luenen S, Meijer E, Honkoop PJ, Pinnock H, Mohr DC, et al. SERIES: eHealth in primary care. Part 4: addressing the challenges of implementation. Eur J Gen Pract 2020 Dec 07;26(1):140-145 [FREE Full text] [doi: 10.1080/13814788.2020.1826431] [Medline: 33025820]
27. Jacob C, Lindeque J, Klein A, Ivory C, Heuss S, Peter MK. Assessing the quality and impact of eHealth tools: systematic literature review and narrative synthesis. JMIR Hum Factors 2023 Mar 23;10:e45143 [FREE Full text] [doi: 10.2196/45143] [Medline: 36843321]
28. Elbert NJ, van Os-Medendorp H, van Renselaar W, Ekeland AG, Hakkaart-van Roijen L, Raat H, et al. Effectiveness and cost-effectiveness of eHealth interventions in somatic diseases: a systematic review of systematic reviews and meta-analyses. J Med Internet Res 2014 Apr 16;16(4):e110 [FREE Full text] [doi: 10.2196/jmir.2790] [Medline: 24739471]
29. Virtanen L, Kaihlanen AM, Laukka E, Gluschkoff K, Heponiemi T. Behavior change techniques to promote healthcare professionals' eHealth competency: a systematic review of interventions. Int J Med Inform 2021 May;149:104432 [FREE Full text] [doi: 10.1016/j.ijmedinf.2021.104432] [Medline: 33684712]
30. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014 Sep;89(9):1245-1251 [FREE Full text] [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]

XSL•FO
RenderX

31. Koshy E, Koshy V, Waterman H. Action Research in Healthcare. London, UK: SAGE Publications; 2011.
32. Cargo M, Mercer SL. The value and challenges of participatory research: strengthening its practice. Annu Rev Public Health 2008;29:325-350. [doi: 10.1146/annurev.publhealth.29.091307.083824] [Medline: 18173388]
33. Baum F, MacDougall C, Smith D. Participatory action research. J Epidemiol Community Health 2006 Oct;60(10):854-857 [FREE Full text] [doi: 10.1136/jech.2004.028662] [Medline: 16973531]
34. Lingard L, Albert M, Levinson W. Grounded theory, mixed methods, and action research. BMJ 2008 Aug 07;337:a567. [doi: 10.1136/bmj.39602.690162.47] [Medline: 18687728]
35. Whitehead D, Taket A, Smith P. Action research in health promotion. Health Educ J 2003 Mar 01;62(1):5-22. [doi: 10.1177/001789690306200102]
36. Waterman H, Tillen D, Dickson R, de Koning K. Action research: a systematic review and guidance for assessment. Health Technol Assess 2001;5(23):iii-157 [FREE Full text] [Medline: 11785749]
37. How to use the ICF: a practical manual for using the International Classification of Functioning, Disability and Health (ICF). World Health Organization. 2013. URL: https://cdn.who.int/media/docs/default-source/classification/icf/drafticfpracticalmanual2.pdf?sfvrsn=8a214b01_4&download=true [accessed 2024-04-25]
38. Krijgsman J, Klein Wolterink G. Ordening in de wereld van eHealth. Nictiz. 2012 Aug 24. URL: https://nictiz.nl/publicaties/ordening-in-de-wereld-van-ehealth/ [accessed 2024-01-18]
39. Moser A, Korstjens I. Series: practical guidance to qualitative research. Part 5: co-creative qualitative approaches for emerging themes in primary care research: experience-based co-design, user-centred design and community-based participatory research. Eur J Gen Pract 2022 Dec;28(1):1-12 [FREE Full text] [doi: 10.1080/13814788.2021.2010700] [Medline: 35037811]
40. Gulliksen J, Göransson B, Boivie I, Blomkvist S, Persson J, Cajander Å. Key principles for user-centred systems design. Behav Inf Technol 2003 Nov;22(6):397-409. [doi: 10.1080/01449290310001624329]
41. Shah SG, Robinson I. User involvement in healthcare technology development and assessment: structured literature review. Int J Health Care Qual Assur Inc Leadersh Health Serv 2006;19(6-7):500-515. [doi: 10.1108/09526860610687619] [Medline: 17100220]
42. de Wit M, Beurskens A, Piškur B, Stoffers E, Moser A. Preparing researchers for patient and public involvement in scientific research: development of a hands-on learning approach through action research. Health Expect 2018 Aug 08;21(4):752-763 [FREE Full text] [doi: 10.1111/hex.12671] [Medline: 29418053]
43. Arnstein SR. A ladder of citizen participation. J Am Inst Planners 1969 Jul;35(4):216-224. [doi: 10.1080/01944366908977225]
44. Aantallen Paramedische zorg. Zorginstituut Nederland. 2023. URL: https://www.zorgcijfersdatabank.nl/databank?infotype=zvw&label=00-totaal&tabel=B_volu&geg=jjverdiepnew22&item=207 [accessed 2025-05-09]
45. van Hassel D, Batenburg R, van der Velden L. Praktijkondersteuners (POH's) in beeld: aantallen, kenmerken en geografische spreiding in Nederland. Nivel. 2016. URL: https://www.nivel.nl/sites/default/files/bestanden/Rapport_analyse_arbeidsmarkt_POH.pdf [accessed 2025-05-19]
46. Flinterman L, Vis E, Noordman J, Batenburg R. De praktijkondersteuner en praktijkverpleegkundige (POH) in beeld: inzicht in de omvang en samenstelling van de POH-beroepsgroep in Nederland. Nivel. 2023. URL: https://www.nivel.nl/sites/default/files/bestanden/1004466_0.pdf [accessed 2025-05-19]
47. Moser A, Korstjens I. Series: practical guidance to qualitative research. Part 3: sampling, data collection and analysis. Eur J Gen Pract 2018 Dec;24(1):9-18 [FREE Full text] [doi: 10.1080/13814788.2017.1375091] [Medline: 29199486]
48. De beroepsvereniging van en voor fysiotherapeuten in Nederland. Koninklijk Nederlands Genootschap voor Fysiotherapie. URL: https://www.kngf.nl/ [accessed 2024-04-12]
49. Ergotherapie - beroepsvereniging voor ergotherapeuten. Ergotherapie Nederland. URL: https://ergotherapie.nl/ [accessed 2024-04-12]
50. Nederlandse Vereniging voor Logopedie en Foniatrie homepage. Nederlandse Vereniging voor Logopedie en Foniatrie. URL: https://www.nvlf.nl/ [accessed 2024-04-12]
51. Nederlandse Vereniging van Praktijkondersteuners en Praktijkverpleegkundigen homepage. Nederlandse Vereniging van Praktijkondersteuners en Praktijkverpleegkundigen. URL: https://www.nvvpo.nl/ [accessed 2024-04-12]
52. Onderzoek naar digitale vaardigheden in de zorg. Vilans. 2020. URL: https://www.vilans.nl/wat-doen-we/projecten/digivaardig-in-de-zorg [accessed 2024-01-18]
53. Garrett JJ. The Elements of User Experience: User-Centered Design for the Web. Berkeley, CA: New Riders Press; 2002.
54. Crutzen R, de Nooijer J, Brouwer W, Oenema A, Brug J, de Vries NK. A conceptual framework for understanding and improving adolescents' exposure to internet-delivered interventions. Health Promot Int 2009 Sep;24(3):277-284. [doi: 10.1093/heapro/dap018] [Medline: 19515716]
55. Saunders B, Sim J, Kingstone T, Baker S, Waterfield J, Bartlam B, et al. Saturation in qualitative research: exploring its conceptualization and operationalization. Qual Quant 2018;52(4):1893-1907 [FREE Full text] [doi: 10.1007/s11135-017-0574-8] [Medline: 29937585]
56. Rahimi S, Khatooni M. Saturation in qualitative research: an evolutionary concept analysis. Int J Nurs Stud Adv 2024 Jun;6:100174 [FREE Full text] [doi: 10.1016/j.ijnsa.2024.100174] [Medline: 38746797]
57. de Bono E. Six Thinking Hats: An Essential Approach to Business Management. Boston, MA: Back Bay Books; 1999.

XSL•FO
RenderX

58. van Dijk-de Vries A, Stevens A, van der Weijden T, Beurskens AJHM. How to support a co-creative research approach in order to foster impact. The development of a Co-creation Impact Compass for healthcare researchers. PLoS One 2020 Oct 12;15(10):e0240543 [FREE Full text] [doi: 10.1371/journal.pone.0240543] [Medline: 33045033]
59. Stevens A, van Dijk A, Jamin G, Beurskens A, Jorissen R. Co-creation impact compass: working together to foster impact. LIME. 2021. URL: https://www.limeconnect.nl/wp-content/uploads/2021/07/Co-creation-Impact-Compass.pdf [accessed 2024-04-25]
60. Elo S, Kyngäs H. The qualitative content analysis process. J Adv Nurs 2008 Apr;62(1):107-115. [doi: 10.1111/j.1365-2648.2007.04569.x] [Medline: 18352969]
61. Bradbury K, Watts S, Arden-Close E, Yardley L, Lewith G. Developing digital interventions: a methodological guide. Evid Based Complement Alternat Med 2014;2014:561320 [FREE Full text] [doi: 10.1155/2014/561320] [Medline: 24648848]
62. Siau K. Evaluating the usability of a group support system using co-discovery. J Comput Inf Syst 2004;44(2):17-28. [doi: 10.1080/08874417.2004.11647563]
63. Kemp JAM, van Gelderen T. Co-discovery exploration: an informal method for the iterative design of consumer products. In: Jordan PW, Thomas B, McClelland IL, Weerdmeester B, editors. Usability Evaluation in Industry. London, UK: CRC Press; 2014.
64. Brooke J. SUS: a 'quick and dirty' usability scale. In: Jordan PW, Thomas B, McClelland IL, Weerdmeester B, editors. Usability Evaluation in Industry. London, UK: CRC Press; 2014.
65. Hsieh H, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005 Nov;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
66. Nielsen J. 10 usability heuristics for user interface design. Nielsen Norman Group. 2024 Jan 30. URL: https://www.nngroup.com/articles/ten-usability-heuristics/ [accessed 2025-02-02]
67. Schenker M. How to use Cialdini's 7 principles of persuasion to boost conversions. CXL. 2024 Jul 30. URL: https://cxl.com/blog/cialdinis-principles-persuasion/ [accessed 2025-02-02]
68. Skivington K, Matthews L, Simpson SA, Craig P, Baird J, Blazeby JM, et al. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. BMJ 2021 Sep 30;374:n2061 [FREE Full text] [doi: 10.1136/bmj.n2061] [Medline: 34593508]
69. Lincoln YS, Guba EG. Naturalistic Inquiry. Thousand Oaks, CA: SAGE Publications; 1985.
70. Korstjens I, Moser A. Series: practical guidance to qualitative research. Part 4: trustworthiness and publishing. Eur J Gen Pract 2018 Dec;24(1):120-124 [FREE Full text] [doi: 10.1080/13814788.2017.1375092] [Medline: 29202616]
71. Frambach JM, van der Vleuten CP, Durning SJ. AM last page. Quality criteria in qualitative and quantitative research. Acad Med 2013 Apr;88(4):552. [Medline: 23531762]
72. Sociaal-economische status per postcode, 2019. Centraal Bureau voor de Statistiek. 2022. URL: https://tinyurl.com/yeefkz76 [accessed 2024-01-18]
73. Bangor A, Kortum PT, Miller JT. An empirical evaluation of the system usability scale. Int J Hum Comput Interact 2008 Jul 30;24(6):574-594. [doi: 10.1080/10447310802205776]
74. Sauro J. 5 ways to interpret a SUS score. Measuring U. 2018 Sep 19. URL: https://measuringu.com/interpret-sus-score/ [accessed 2024-06-22]
75. Zuyd University of Applied Science, Utrecht University of Applied Science, Maastricht University, Windesheim University of Applied Science. Guide to measuring with eHealth. Measuring with eHealth. 2023 Dec. URL: https://metenmetehealth.nl/ [accessed 2025-07-22]
76. Michie S, van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. Implement Sci 2011 Apr 23;6:42 [FREE Full text] [doi: 10.1186/1748-5908-6-42] [Medline: 21513547]
77. Lagarde M, Huicho L, Papanicolas I. Motivating provision of high quality care: it is not all about the money. BMJ 2019 Sep 23;366:l5210 [FREE Full text] [doi: 10.1136/bmj.l5210] [Medline: 31548200]
78. Jacob C, Sanchez-Vazquez A, Ivory C. Social, organizational, and technological factors impacting clinicians' adoption of mobile health tools: systematic literature review. JMIR Mhealth Uhealth 2020 Feb 20;8(2):e15935 [FREE Full text] [doi: 10.2196/15935] [Medline: 32130167]
79. Hermes-DeSantis ER, Hunter RT, Welch J, Bhavsar R, Boulos D, Noue MA. Preferences for accessing medical information in the digital age: health care professional survey. J Med Internet Res 2021 Jun 19;23(6):e25868 [FREE Full text] [doi: 10.2196/25868] [Medline: 36260374]
80. Brennan N, Edwards S, Kelly N, Miller A, Harrower L, Mattick K. Qualified doctor and medical students' use of resources for accessing information: what is used and why? Health Info Libr J 2014 Sep;31(3):204-214 [FREE Full text] [doi: 10.1111/hir.12072] [Medline: 25041386]
81. Cavanagh R, Gerson SM, Gleason A, Mackey R, Ciulla R. Competencies needed for behavioral health professionals to integrate digital health technologies into clinical care: a rapid review. J Technol Behav Sci 2022 Feb 17;8(4):1-14 [FREE Full text] [doi: 10.1007/s41347-022-00242-w] [Medline: 35194561]
82. Huber C. [Interprofessional Collaboration in Health Care]. Praxis (Bern 1994) 2022 Jan;110(1):3-4 [FREE Full text] [doi: 10.1024/1661-8157/a003808] [Medline: 34983202]

XSL•FO
RenderX

83. Spaulding EM, Marvel FA, Jacob E, Rahman A, Hansen BR, Hanyok LA, et al. Interprofessional education and collaboration among healthcare students and professionals: a systematic review and call for action. J Interprof Care 2021;35(4):612-621 [FREE Full text] [doi: 10.1080/13561820.2019.1697214] [Medline: 31865823]
84. Lindsey L, Rathbone AP. Beyond the numbers: utilising existing textual data for qualitative research in pharmacy and health services research. Res Social Adm Pharm 2022 Jan;18(1):2193-2199 [FREE Full text] [doi: 10.1016/j.sapharm.2021.04.010] [Medline: 33903065]
85. van Tol LS, Smaling HJA, Groothuijse JM, Doornebosch AJ, Janus SIM, Zuidema SU, et al. COVID-19 management in nursing homes by outbreak teams (MINUTES) - study description and data characteristics: a qualitative study. BMJ Open 2021 Nov 30;11(11):e053235 [FREE Full text] [doi: 10.1136/bmjopen-2021-053235] [Medline: 34848521]
86. Moore GF, Audrey S, Barker M, Bond L, Bonell C, Hardeman W, et al. Process evaluation of complex interventions: Medical Research Council guidance. BMJ 2015 Mar 19;350:h1258. [doi: 10.1136/bmj.h1258] [Medline: 25791983]
87. ChatGPT homepage. ChatGPT. URL: https://chatgpt.com/ [accessed 2024-01-18]

Abbreviations

COM-B: capability, opportunity, and motivation for any behavior to occur
GP: general practitioner
MoSCoW: must have, should have, could have, would have
PAR: participatory action research
SES: socioeconomic status
SRQR: Standards for Reporting Qualitative Research
SUS: System Usability Scale
UCD: user-centered design

Edited by A Mavragani; submitted 10.02.25; peer-reviewed by P Tremoulet, C Kunze; comments to author 13.04.25; revised version received 23.05.25; accepted 23.06.25; published 16.09.25.

Please cite as:
Meijers K, Bols E, Beekman E, Slangen R, van de Hoef PA, Ummels D, Crutzen R
Development of a Web-Based Intervention to Support Primary Health Care Professionals in Digital Health Measurement: User-Centered Participatory Approach
JMIR Form Res 2025;9:e72331
URL: https://formative.jmir.org/2025/1/e72331
doi:10.2196/72331
PMID:

©Kristel Meijers, Esther Bols, Emmylou Beekman, Rachel Slangen, Peter Alexander van de Hoef, Darcy Ummels, Rik Crutzen. Originally published in JMIR Formative Research (https://formative.jmir.org), 16.09.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Optimizing the Color Shapes Task for Ambulatory Assessment and Drift Diffusion Modeling: A Factorial Experiment

Sharon Haeun Kim[1,2*], MA; Jonathan G Hakun[3,4,5,6*], PhD; Yanling Li[1*], PhD; Karra D Harrington[6*], PhD; Daniel B Elbich[3*], PhD; Martin J Sliwinski[1,6*], PhD; Joachim Vandekerckhove[7*], PhD; Zita Oravecz[1,6,8*], PhD

[1]Department of Human Development and Family Studies, The Pennsylvania State University, University Park, PA, United States
[2]Department of Social Data Analytics, Pennsylvania State University, University Park, United States
[3]Department of Neurology, College of Medicine, The Pennsylvania State University, Hershey, PA, United States
[4]Department of Psychology, The Pennslyvania State University, University Park, PA, United States
[5]Department of Public Health Services, College of Medicine, The Pennsylvania State University, Hershey, PA, United States
[6]Center for Healthy Aging, The Pennsylvania State University, University Park, PA, United States
[7]Department of Cognitive Sciences, University of California, Irvine, Irvine, CA, United States
[8]Institute for Computational and Data Sciences, Pennsylvania State University, University Park, United States
[*]all authors contributed equally

Corresponding Author:
Sharon Haeun Kim, MA
Department of Human Development and Family Studies
The Pennsylvania State University
235 Health and Human Development Building
University Park, PA, 16802
United States
Phone: 1 814 863 8000
Email: skim@psu.edu

Abstract

Background: Recent advances in cognitive digital assessment methodology, including high-frequency, ambulatory assessments, promise to improve the detection of subtle cognitive changes. Computational modeling approaches may further improve the sensitivity of digital cognitive assessments to detect subtle cognitive changes by capturing features that map onto core cognitive processes.

Objective: We explored the validity of a brief smartphone-based adaptation of a visual working memory task that has shown sensitivity for detecting preclinical Alzheimer disease risk. We aimed to optimize properties of the task for computational cognitive feature extraction with drift diffusion modeling.

Methods: We analyzed data from 68 participants (n=47, 69% women; n=55, 81% White; mean age 49, SD 14; range 24-80 years) who completed 60 trials for each of 16 variations of a visual working memory binding task (the Color Shapes task) on smartphones, over an 8-day period. A drift diffusion model was fit to the response time and accuracy data from the task. We experimentally manipulated 3 properties of the Color Shapes task (study time, probability of change, and choice urgency) to test how they yielded differences in key drift diffusion model parameters (drift rate, initial bias toward a response option, and caution in decision-making). We also evaluated how an additional task property, the test array size, impacted responses across all conditions. For array size, we tested a *whole display* of 3 shapes against a *single probe* of 1 shape only.

Results: The 3 task property manipulations yielded the following results: (1) increasing the ratio of *different* responses was credibly associated with higher initial bias toward the *different* response (mean 0.06, SD 0.02 for the whole display; mean 0.15, SD 0.02, for the single probe condition); (2) increasing the choice urgency during the test phase was credibly associated with decreased caution in decision-making in the single probe condition (mean −0.04, SD 0.02) but not in the whole display (mean −0.01, SD 0.02); and (3) contrary to expectation, longer study times did not yield a credibly faster drift rate but produced credibly slower ones for the whole display condition (mean −0.28, SD 0.05) and a null effect for the single probe condition (mean 0.01, SD 0.05). In addition, as expected, we found that individual differences in drift rate were associated with age in both array sizes (r=−0.45 with Bayes factor=191), with older participants having a slower drift rate. Older participants also showed higher caution (r=0.42 with Bayes factor=80.76) in the single probe condition.

XSL•FO
RenderX

Conclusions: We identified a version of the Color Shapes task optimized for smartphone-based cognitive assessments in real-world settings, with data designed for analysis through computational cognitive modeling. Our proposed approach can advance the development of tools for efficient and effective early detection and monitoring of risk for Alzheimer disease.

(JMIR Form Res 2025;9:e66300) doi:10.2196/66300

KEYWORDS

subtle cognitive decline; computational cognitive markers; smartphone-based cognitive testing; drift diffusion model; Bayesian multilevel modeling; mobile phone

Introduction

Background

Conventional cognitive testing is widely used to diagnose neurocognitive impairments from various clinical conditions. However, many of these conditions have a long preclinical phase, characterized by cognitive changes that are more subtle than what most conventional cognitive tests are designed to detect [1,2]. In other words, while cognitive performance summary scores (typically obtained from 1-time assessment) may be sufficient for detecting overt impairments in neurodegenerative diseases, they often lack the sensitivity to capture the early-stage changes in subtle cognitive subprocesses. Ratcliff and McKoon [3] have suggested that cognitive and neuropsychological testing could capitalize on advances in cognitive modeling to move beyond static performance summary scores and better assess the latent cognitive dynamics. With advances in digital health tools, smartphone-based cognitive assessments have shown promise for capturing subtle processes in cognitive performance daily contexts (ie, ambulatory assessment), facilitating more representative and richer measurements [4,5]. Here, we introduce a smartphone-based implementation of a working memory task that evaluates a cognitive function that has been linked with early-stage Alzheimer disease (AD) risk [6,7].

Recent evidence suggests that sensitive measures of subtle cognitive processes may be captured through repeated ambulatory cognitive assessments with smartphones [8], particularly when combined with computational cognitive modeling [9]. High-frequency ecological momentary assessments [10] of cognitive functioning across days produces a rich data stream across time and daily life contexts [11], providing a representative sample of daily cognition. These repeated assessments in natural contexts also mitigate reliability concerns [8], while preserving ecological validity [12], and permit the implementation of computational methods to effectively explore subtle within-person cognitive processes [13]. However, many cognitive tasks are designed for laboratory administration, and the resulting performance scores are often summarized via simple summary statistics. Here, we show how subtle cognitive processes underlying ambulatory reaction time and accuracy data can be captured with cognitive modeling. Our approach emphasizes the advantages of identifying individual differences in cognitive features via state-of-the-art computational cognitive psychometric modeling from repeated ambulatory assessments of cognitive performance.

A consistent theme in the cognitive modeling literature highlights the benefits associated with cognitive model–based approaches, which support the widely used scaling transformations [14,15], assist in identifying potential confounds [6,16], and improve parameter estimation stability [17]. When it comes to evaluating processing speed for example, cognitive modeling can jointly consider both response time (RT) and choice behavior and contribute to greater analytical sensitivity than mean-level insights alone [3]. Cognitive psychometric models [18] provide mathematically disciplined approaches for simulating data-generating mechanisms, distilling observed behavioral scores, and extracting individual differences in their latent cognitive features. In summary, this approach can identify subtle latent processes underlying manifest cognitive performance data and explore individual differences therein to capture latent mechanisms and dynamic processes [19].

There has been a growing shift toward adopting cognitive models to diagnostic challenges posed by early AD, such as a need to distinguish AD-specific markers from other neuropsychological conditions, normative age-related declines, and practice effects [3,7,16]. Cognitive modeling approaches have shown promise in identifying early cognitive markers of dementia risk, such as reduced episodic memory [20], higher task-switching costs [21], declines in attentional control [22], and higher performance inconsistency [23]. Moreover, cognitive modeling can parse out individual differences in these cognitive features linked to clinically relevant person-level characteristics, such as age [24], genetic and familial risks [6,20], and cognitive impairment status [3,9]. The drift diffusion model (DDM) framework has also been increasingly applied to characterize latent cognitive processes underlying neurological conditions [25-28] and age-related cognitive changes [29,30].

Parra et al [6] identified visual working memory binding as one of the earliest underlying functions impacted in preclinical AD, particularly deficits in “binding” visual features. This feature-binding deficit has demonstrated sensitivity and specificity across a range of at-risk populations, validated as a reliable early cognitive marker in symptomatic and asymptomatic mutation carriers [6,25], individuals with mild cognitive impairment and AD [26,30], and individuals across the clinical continuum with elevated AD-related biomarkers [27,31]. The original task was designed to isolate visual feature binding by using novel, abstract shapes and randomized color-shape pairings [6]. The abstract shapes were selected to make it difficult to attach a label to them (eg, triangle and square), thus preventing undesirable verbal encoding instead of relying on visual working memory. Building on this paradigm, we tested a mobile adaptation of this task, which we will refer to as the “Color Shapes” task [5]. Previous implementations of the Color Shapes task for mobile administration have been

XSL•FO
RenderX

largely web based [28,29] and focused on analyzing aggregated mean performance data across all administrations and did not explore the cognitive processes through computational modeling.

To address this gap, we propose that cognitive modeling of repeated administrations of Color Shapes task can further enhance our ability to capture subtle cognitive change processes. This approach can formally assess the validity of the model by evaluating fit with observed data [19]. One such model, the DDM [32], characterizes decision-making as a noisy process of evidence accumulation toward response options and can disentangle into hypothesized generative cognitive processes. The DDM captures the latent processes underlying observed performance of RT and accuracy [3,14,33]. By fitting the DDM to RT and accuracy data, we can disentangle subtle underlying cognitive processes, such as rate of evidence accumulation, caution in decision-making, and response biases [32,34]. For instance, individual differences in the speed of evidence accumulation and caution can elucidate important idiosyncrasies in decision-making so that we can better identify those at risk for AD [20,35].

Objectives

The aim of this study was twofold: (1) to conduct an experimental validation of the DDM analytic approach in the context of RT and accuracy data from the Color Shapes task and (2) to identify an ambulatory Color Shapes task version that is most optimal for DDM-based data analysis. We propose that the DDM-based computational modeling approach can extract useful signals that can serve as novel digital cognitive markers of risk for subtle cognitive decline. Consistent with the literature on cognitive testing and aging, we hypothesized the following effects on cognitive features in response to our experimental manipulations: (1) allowing for longer study time would increase the evidence accumulation rate, (2) showing trials with a higher number of *different* trials (48/60, 80% *different* trials) would bias the starting point of the decision process toward *different*, and (3) imposing higher decision urgency (3000 ms limit) would lower the decision boundary, corresponding to lower response cautiousness. Overall, our study aimed to lay the groundwork for this approach and illustrate the corresponding statistical tools for data analysis. As a corollary to our optimization effort, we also considered which test array size produces data in ambulatory setting that are in line with the assumptions of the DDM (ie, single shot decision-making) and measuring visual working memory (as opposed to memory relying on verbal encoding). For this, we varied 2 test phase formats: simultaneous presentation of 3 visual elements phase (labeled as *whole display*) versus simply showing a single item in the test phase (*single probe*).

Methods

Ethical Considerations

This study was approved by the Pennsylvania State University Institutional Review Board (STUDY00018787). All procedures were conducted in accordance with the ethical principles outlined in the Declaration of Helsinki and all applicable guidelines and regulatory standards. All participants provided written informed consent before participation in the study. All data were fully deidentified before analysis and handled in anonymized form. Participants received compensation of up to US $100 for completing the study.

Participant Recruitment

We recruited 69 adults from Pennsylvania (n=47, 69% women; n=21, 31% men) on the web via ResearchMatch, a nonprofit, National Institutes of Health–funded national registry established through the Clinical Translational Science Award program to support medical study participation [36,37]. Participants were screened via telephone to assess eligibility. Participant inclusion criteria are presented in Textbox 1.

Textbox 1. Participant inclusion criteria.

- Aged >18 years
- Fluent in English
- Has access to a reliable internet connection
- Does not have a motor or visual impairment that would interfere with operating a smartphone
- Does not have a history of neurological injury or disease (eg, stroke and seizures)

After an initial remote onboarding session (conducted via videoconference), participants were asked to complete cognitive tasks (Color Shapes trials; details in the measures section) for 8 days for approximately 20 minutes each day. More information is provided in the design and procedure section, which follows the brief description of the task and the computational modeling framework chosen to analyze the data.

Measures

Demographics

Participants responded to gender, age, race and ethnicity, ethnic status, and level of education. Participants were asked whether they had ever tested positive for COVID-19 (data were not analyzed). Basic health information was provided on ResearchMatch upon sign-up (eg, BMI, medical conditions, and medication).

The Color Shapes Task

The Color Shapes task is a visual array change detection task where participants were asked to determine whether the combination of features (color and shape) among visual objects distributed throughout the array change between study and test arrays. In our study, participants were asked to study a set of 3 abstract shapes, each with a unique color. After a brief 900-ms delay, they were presented with a test array and asked to

XSL•FO
RenderX

determine if the stimuli contained in the test array contained the same combination of shape and color as the study array. To identify an optimal combination of task parameters for mobile administration and drift diffusion modeling, we experimentally manipulated 4 features of the task in a within-persons 2 (study duration)×2 (probability of change)×2 (response duration)×2 (probe type) full factorial experimental design. Specifically, each trial of the Color Shapes task began with a fixation cross displayed for 500 ms after which 3 colored abstract shapes appeared for either a short (500 ms) or longer (2000 ms) period (study duration). Participants were asked to study the specific combinations of shape and color presented in the study array objects.

The abstract shape set consisted of 8 distinct shapes and 6 distinct colors. All stimuli used in the experiment are presented in Multimedia Appendix 1 [4]. For each trial, 3 shapes and 3 colors were randomly selected without replacement to form 3 unique shape-color pairings. In the study array, these 3-colored shapes were presented within a 3×3 grid layout (9 possible positions), and each shape was randomly assigned to one of the selected locations. After a brief 900-ms delay period, either a 1-colored shape or 3-colored shapes reappeared (probe type) at different locations throughout the test array, as illustrated in Figure 1. The test arrays either maintained the same combinations of color and shape (a *same* trial) or two of the shapes swapped colors (a *different* trial). The 3 selected color and shape elements remained constant within each trial, only their pairings were changed. Participants indicated whether they believed the combination of shapes and colors presented in the test array match the combinations presented in the study array by pressing the *same* or *different* buttons located at the bottom of the screen. Test array stimuli remained on the screen for either a short (3000 ms) or long (10,000 ms) period (response duration). Each trial has either a 50% to 50% chance or 20% to 80% chance (probability of change) of being a *same* or *different* trial.

Figure 1. Study and test arrays of the Color Shapes task. An illustration of the Color Shapes task showing the study array and test array with four configurations from left to right: (1) different trial with whole display (3-colored shapes) shown in the test array, (2) same trial with whole display (3-colored shapes) shown in the test array, (3) different trial with single probe (1-colored shape) shown in the test array, and (4) same trial with single probe (1-colored shape) shown in the test array.

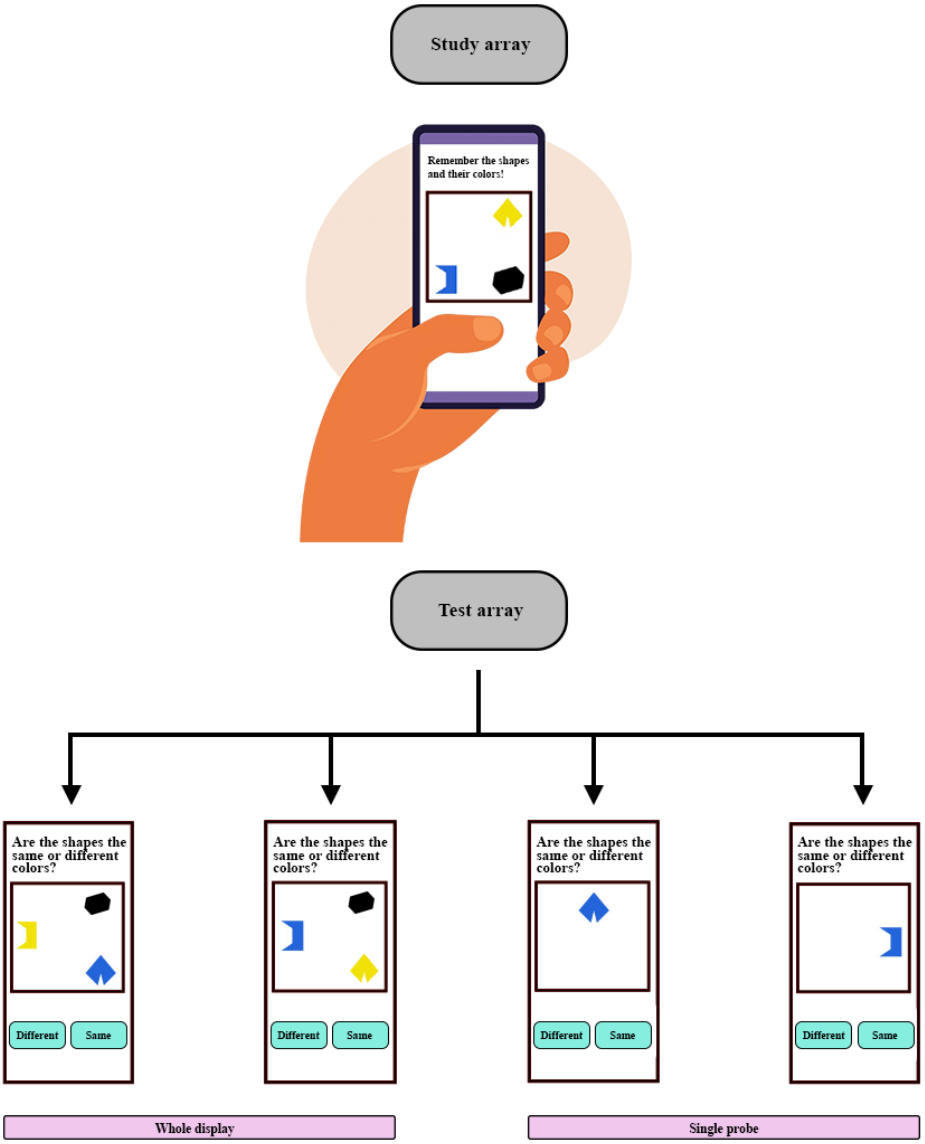

XSL•FO
RenderX

All manipulations were implemented within-persons across 8 consecutive days of ambulatory testing. Each participant completed 2 versions of the task per day, with 60 trials per version, culminating in up to 960 total trials across the 8 days. The 2 task versions were administered according to a randomized schedule. To evaluate testing experience and assess potential strategy use, participants completed a free response debrief item at the end of the study. The type of probes used at testing (all 3 items presented simultaneously in the whole-display format or a single item presented in the single-probe format) may have enabled the use of cognitive strategies such as focusing on a subset of items or using self-generated mnemonic heuristics. Instructions did not explicitly prohibit strategy use (eg, verbal encoding) to avoid inadvertently increasing its salience, but we assessed for potential strategy use during the poststudy debrief. Overall, our goal was to translate the Color Shapes tasks for unsupervised, ambulatory use while evaluating task designs suitable for extracting latent cognitive parameters through drift diffusion modeling, described in the next section.

Debrief Survey

Upon completing the final session, participants completed the exit survey consisting of 3 items. First, participants were asked how typical their routine, activities, and experiences were during the 8 study days on a 5-point scale, where 1=very unusual, 2=unusual, 3=neutral, 4=typical and 5=very typical, with a prefer not to respond option. Second, participants were asked whether they were impacted by any unusual circumstances or stressful events over the study duration. There were 13 response options ranging from nothing unusual to various stressful events (eg, negative social interaction, health issues, financial issues, and world events). Finally, participants were asked, “Over time, did you develop any strategies for better performance on the brain games?” (yes or no). If they responded yes, they were asked to provide more details about the strategies they used. This was an unstructured free response question. We only analyzed responses from the third item on cognitive strategy use in this study.

Computational Modeling With the DDM

The DDM falls under the family of sequential sampling models with continuous time and continuous evidence [32]. A graphical illustration of the model is shown in Figure 2. This model assumes that each decision-making task starts a stochastic information accumulation process of incoming evidence toward one of the two response options**.** The observed decision is determined by the process terminating at one of the decision boundaries, represented in Figure 2 as the upper (*respond Different*) and lower (*respond Same*) bounds of the Color Shapes task. The distance between these bounds (the boundary separation) captures the threshold amount of evidence needed for a response to be chosen. Mathematically, the information accumulation process is modeled as a 1D or Wiener diffusion process that terminates at the absorbing response boundary. With evidence toward the upper boundary represented by positive values and lower boundary represented by negative values, decisions are based on this Wiener diffusion process in continuous time as a single total that integrates both antagonistic types of evidence.

Figure 2. An illustration of the drift diffusion model. In a decision-making process, evidence is accumulated over time at an average drift rate of δ. The decision process terminates if the cumulative evidence value reaches 0 (lower boundary for "respond Same") or α (upper boundary for "respond Different"). At the onset of a trial, the decision process can be biased due to the amount of preliminary evidence given by αβ. The nondecision time τ reflects the duration of perceptual processes such as stimulus encoding and response execution, which occur outside the actual decision-making process. Equation 1 in Multimedia Appendix 2 describes the response time (RT) distributions that follow from these model assumptions. Figure generated via [39].

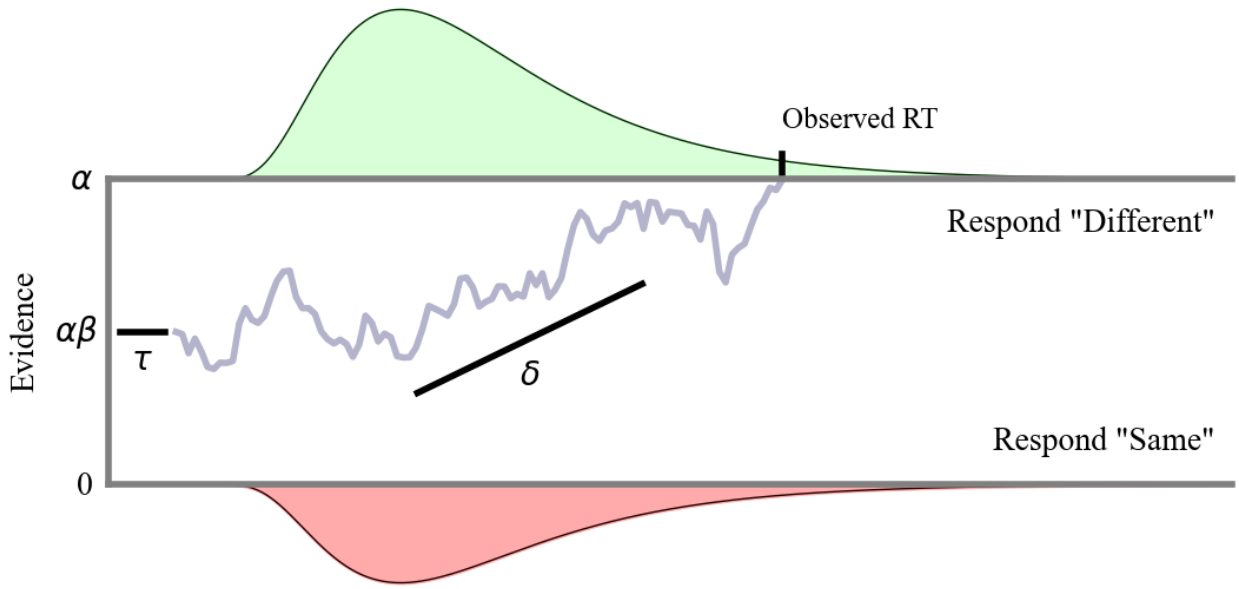

XSL•FO
RenderX

The 4 core parameters of the DDM are presented in Table 1 (adapted from the studies by Ratcliff and Vandekerckhove [32,38]). Formally, the drift rate (δ) describes the speed of information accumulation by capturing the average amount of evidence that is accumulated from the stimulus over a small period, the boundary separation (α) describes the level of evidence required to make a decision (capturing speed-accuracy trade-off), the initial bias (β) describes the starting status toward one or the other boundary before the evidence accumulation process, and the nondecision time (τ) represents the component of RT that is not related to the decision-making process. When an individual is presented with 3 colored shapes in the study array across repeated trials, the DDM models the evidence accumulation process toward a *different* or *same* decision using the individual's initial bias, nondecision time, boundary separation, and evidence accumulation rate. More details on the mathematical formulation of the model are provided in Multimedia Appendix 2 [17,39-46]. Given a participant's choice RT data, we can estimate these parameters and interpret them in the context of the Color Shapes task [38].

Table 1. Summary of the parameters of the drift diffusion model.

Symbol	Parameter	Definition	Interpretation
δ	Drift rate	Average quality of the stimulus and information accumulation process	Higher δ indicates faster accumulation of evidence; lower δ indicates slower accumulation
β	Initial bias	Starting point bias for either response	β>.5 indicates bias toward *different* response; β<.5 toward *same*
α	Boundary separation	Distance quantifying evidence required to make a decision; speed-accuracy trade-off	Higher α indicates more caution and higher accuracy; lower α indicates reduced caution and lower accuracy
τ	Nondecision time	Motor response time, encoding time	High τ indicates slow encoding or motor response; low τ toward faster encoding

In a decision-making process, evidence is accumulated over time at an average drift rate of δ. The decision process terminates if the cumulative evidence value reaches 0 (lower boundary for *respond Same*) or α (upper boundary for *respond Different*). At the onset of a trial, the decision process can be biased due to the amount of preliminary evidence given by αβ. The nondecision time τ reflects the duration of perceptual processes, such as stimulus encoding and response execution, which occur outside the actual decision-making process. Equation 1 in Multimedia Appendix 2 describes the response-time distributions that follow from these model assumptions. Figure generated via the study by Chavez et al [47].

Design and Procedure

After providing informed consent and enrolling in the study, participants were sent smartphones configured with the Color Shapes cognitive assessments. Next, a remote meeting with a study administrator was scheduled where they were onboarded onto the study and trained on the study protocol (eg, smartphone instructions). Over the next 8 consecutive days, participants were instructed to complete a self-initiated daily session consisting of overall 16 versions of the Color Shapes task (see description below), with each daily session taking <20 minutes. Two versions of the task were completed once per day over 8 days, and the order of the presented 16 versions were drawn randomly from 3 different schedules. Each version consisted of 60 ultrabrief trials per task, with a maximum of 960 total possible trials. The participants could only complete the testing once per day, the testing had to be completed in a single setting, which timed out after 30 minutes. Make-up opportunities were not provided for missed or incomplete sessions, and participants were presented with the next scheduled versions. Reminder alerts were provided every hour until 9 PM. Upon completing the final session, participants completed a final exit survey on the smartphone. Afterward they were debriefed by a study coordinator via videoconference and were provided with instructions to return the smartphone.

We experimentally manipulated features of the Color Shapes task in line with prior parameter selection experiment in cognitive modeling. These included four within-person variables in the Color Shapes task: (1) study time, (2) probability of change, (3) choice urgency, and (4) probe types.

We had 2 objectives in mind: to test the sensitivity for capturing changes in terms of the latent parameters of the DDM and to compare the 3-colored shapes test array and 1-colored shape test array versions of the task shown in Figure 1. Specifically, we manipulated across four task features, each with 2 levels:

1. Short (500 ms) versus long (2000 ms) study array time
2. Low (50% *different* trials and 50% *same* trials) versus high (80% *different* trials and 20% *same* trials) proportion of *different* trials, capturing different probabilities of change in the properties of the shapes from the study array to the test array
3. Minimal choice urgency (10,000 ms to respond) versus high choice urgency (3000 ms to respond)
4. Whole display (3 colored shapes simultaneously presented in the test array) versus single probe (1 colored shape simultaneously presented in the test array)

Previous cognitive modeling research has demonstrated that each of these manipulations map onto distinct parameters of the DDM [48,49]. Drift rate, the speed of information accumulation, is enhanced by longer study times as it enables more effective encoding and speed [50,51]. Adjusting the probability of change shapes the initial bias point of decision-making, shifting it toward the more frequently occurring response [34]. Manipulating response urgency

XSL•FO
RenderX

influences boundary separation, mapping onto the speed accuracy trade-off where increased urgency leads to more impulsive responses and has been found to be more prevalence with age [50]. Finally, single-probe designs have been shown to effectively isolate feature-binding processes in visual working memory targeted by the Color Shapes task that has been identified as an early cognitive marker of preclinical AD [6], alongside broader utility in assessing individuals with diminished memory capacity [52]. Accordingly, we designed each experimental manipulation to target a distinct DDM parameter and evaluate its sensitivity in an ambulatory, unsupervised testing context (Table 2).

Manipulations (1-3) aimed at influencing parameters of the DDM: (1) drift rate (δ) with the study array, (2) initial bias (β) with the different probabilities of change, and (3) boundary separation (α) with the different time urgency to respond. The nondecision time (τ) was estimated across participants and conditions, but we did not make any predictions about possible differences in it across conditions. Experimental manipulations were not communicated to participants. An illustration of these 3 cognitive task manipulations is shown in Figure 3 for the single probe trials.

We had the following hypotheses about the effects of the manipulations on the parameters:

1. We expected faster drift rates with longer study times, that is, more time for encoding leading to faster accumulation of evidence
2. Initial bias toward *different* responses with higher probability of *different* trials than the 50% to 50% probability
3. Lower (closer) boundary separation, that is, less caution in decision-making with increased urgency to choose a response
4. We studied how the first 3 manipulations would work with 2 visually different versions of the task, where the test array with single probe version has the additional advantage of reducing unwanted strategy use.

The experimental manipulations are summarized in Table 2.

Table 2. Drift diffusion model experimental manipulations and hypothesized effects.

Condition, parameter, and levels	Hypotheses
Study time	
Drift rate (δ)	
500 ms study time (short study time)	Slower drift
2000 ms study time (long study time)	Faster drift
Probability of change	
Initial bias (β)	
50% *different* or 50% *same* (low change)	No bias
80% *different* or 20% *same* (high change)	Bias toward *different*
Choice urgency	
Boundary separation (α)	
10,000 ms response-time window (minimal urgency)	Farther boundaries
3000 ms response-time window (high urgency)	Closer boundaries
Probe type	
Test array with 3 colored shapes (whole display)	No expected effects
Test array with 1 colored shape (single probe)	No expected effects

XSL•FO
RenderX

Figure 3. An illustration of 8 out of 16 Color Shapes task versions (all with test arrays with 1 colored shapes) with examples of different trial types shown here. Manipulations for task arrays are emphasized in green font and baseline conditions in black font. Versions have been ordered to flow from least manipulated to most manipulated versions.

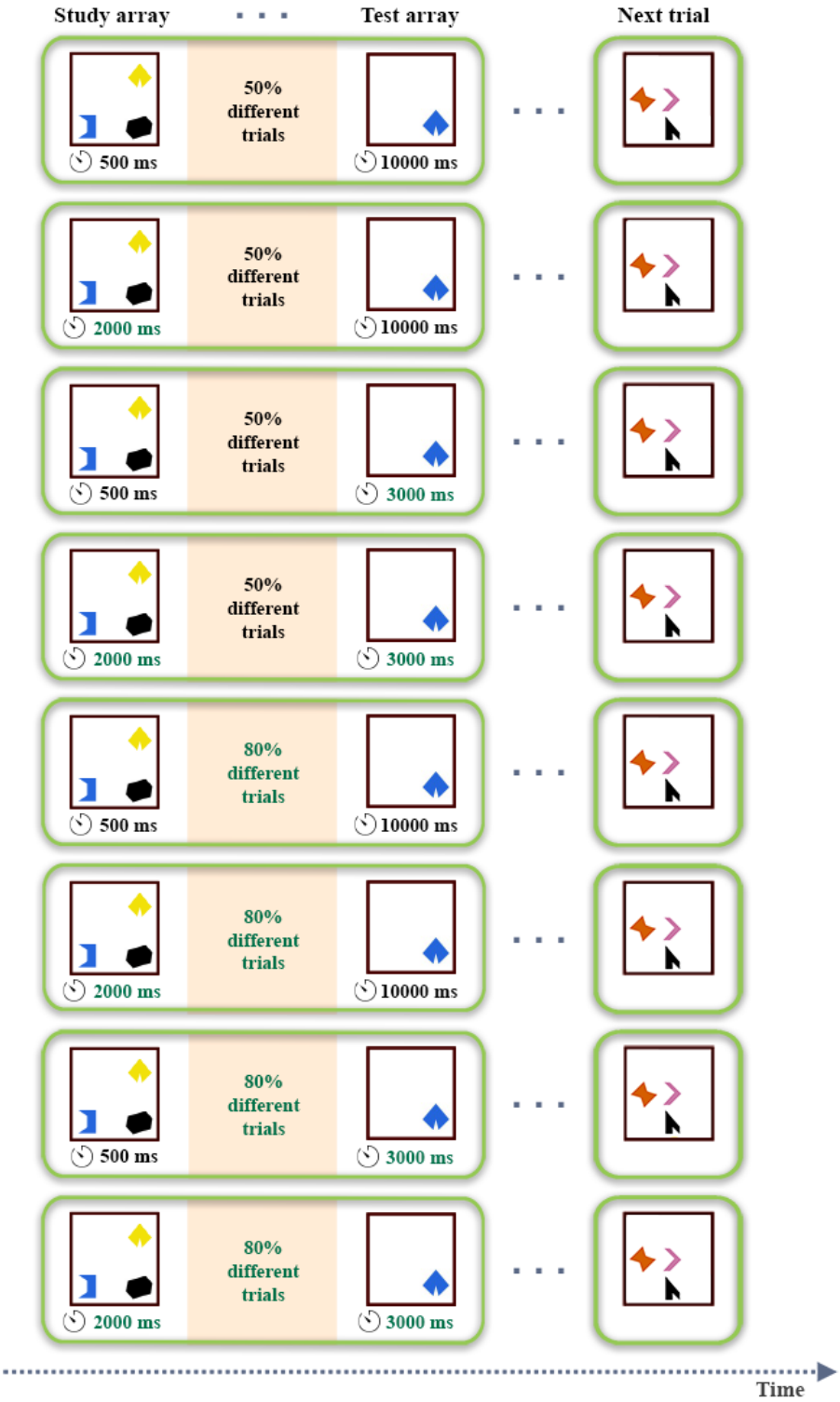

Data Processing

We set the RT-based outliers attending to the real-world constraints of the Color Shapes task. The DDM was intended to assess decision-making for fast 1-shot processes. Trials with RTs faster than 200 ms and slower than 7000 ms were excluded. Ultrafast RTs suggested technical errors, and long RTs were assessed as not following speeded task instructions or not paying attention to the task.

Data Analysis

We fit a tailored, multilevel DDM to RT and accuracy to derive latent cognitive parameters of interest. The multilevel DDM was cast in a Bayesian framework [41]. We estimated drift rates, initial bias, boundary separation, and the non–decision time

XSL•FO
RenderX

parameter for each condition and individual, allowing these parameters to vary between all 16 conditions with no constraints. Finally, the RT data were scaled from milliseconds to seconds for the analysis. A detailed description of our multilevel DDM is provided in Multimedia Appendix 2. All analysis scripts and data are provided on the Open Science Framework [39].

Results

Participant Characteristics

We recruited 69 participants for the study but excluded 1 participant due to missing 99.1% of data (9/960 trials completed). Therefore, our final sample size was 68, and demographics for these participants are displayed in Table 3. The average age of participants was 49 years; 69% (47/68) were women, 31% (21/68) were men, and none identified differently. Participants' ages spanned across younger to older adulthood (mean 49, SD 14; range 24-80 years), as visualized in Multimedia Appendix 3. Regarding race, 81% (55/68) identified as White, 4% (3/68) as Black or African American, 6% (4/68) as Asian, 1.5% (1/68) as American Native or Alaska Native, 1.5% (1/68) as Native Hawaiian or Pacific Islander, 3% (2/68) as multiethnic, and 3% (2/68) as other. As for ethnicity, 96% (65/68) identified as non-Hispanic and 4% (3/68) identified as Hispanic. In terms of highest education received, 5% (3/68) reported high school, 14% (9/68) reported vocational or some college, 12% (8/68) received an associate's degree, and 68% (46/68) reported at least a bachelor's degree or higher.

Table 3. Demographic characteristics of the sample (N=68).

	Values
Gender, n (%)	
Women	47 (69)
Men	21 (31)
Race, n (%)	
Asian	4 (6)
American Native or Alaska Native	1 (1)
Black or African American	3 (4)
Multiethnic	2 (3)
Native Hawaiian or Pacific Islander	1 (1)
Other	2 (3)
White	55 (81)
Ethnicity, n (%)	
Non-Hispanic	65 (96)
Hispanic	3 (4)
Education, n (%)	
High school	3 (5)
Vocational or some college	9 (14)
Associate's degree	8 (12)
Bachelor's degree	19 (29)
Post–bachelor's degree	2 (3)
Master's degree	19 (29)
Doctoral degree	6 (9)
N/A[a]	2 (3)
Age (y), mean (SD); range	49 (14); 24-80

[a]N/A: not applicable.

Descriptive Statistics

In total, 41 participants completed all possible 960 trials (16 versions by 60 trials); 22 participants completed at least 75% (720/960) of the trials, and 4 participants completed 62.5% (600/960), and only 1 participant completed only 25% (240/960) of the trials. A summary of participant-level trial engagement is shown in Table 4. The overall rate of unrecorded or missing trials was 8.45% (5,518/65,280). See Multimedia Appendix 4 for a detailed breakdown of trial exclusion percentages. In short, from the 65,280 total possible trials, 59,762 (91.55%) trials were recorded, and a total of 58,978 (90.35%) trials were retained for analysis. In total, 784 trials (1.20%) were excluded based on predefined criteria. Among these, 720 (1.10%) were

XSL•FO
RenderX

removed because they exceeded the allowed response window in the choice urgency condition, and 54 trials (0.09%) were removed due to unrealistic values (eg, unrealistically fast RTs <200 ms or slow RTs >7000 ms).

Table 5 shows that accuracy rates by test array condition (single probe vs whole display). The mean overall accuracy rate of correct trials among all trials was 85.83% (50,623/58,978). Accuracy rate was higher in the whole display (26,945/29,427, 91.57%) than in the single probe (23,678/29,551, 80.13%) condition.

Table 4. Summary on participant engagement during the study (N=68).

Completion status, %	Trials completed, n	Versions completed, n	Participants, n (%)
100.0	960	16	41 (60)
93.8	900	15	2 (3)
87.5	840	14	13 (19)
75.0	720	12	7 (10)
62.5	600	10	4 (6)
25.0	240	4	1 (2)

Table 5. Accuracy rates and response times across all trials and between test array probe types.

Probe type	Correct trials, n	Total trials, n	Accuracy, %	Response time (ms)
Overall	50,623	58,978	85.83	1290.95
Whole display	26,945	29,427	91.57	1314.97
Single probe	23,678	29,551	80.13	1270.35

Model Fit

To evaluate model fit, we simulated posterior predictive data (n=6000) based on the DDM. In Figure 4, we plotted observed (data-based) and model-predicted RT quantiles for all 16 conditions. Empty circles represent quantiles of incorrect response RT distributions, and filled circles are correct responses. We can see that the observed and model predicted values line up well for RTs below 2 seconds, with more variance occurring after 2 seconds. However, even for these larger RTs, the fit is reasonable. In Figure 5 we only show these for a single probe version, with 500-ms study time, no choice urgency, and half *different* or half *same* probes (our recommended constellation of features for future use in combination with drift diffusion modeling, see the discussion section). This plot also shows great fit with some flaring for slow RTs, which are more variable, and some underestimation of error RTs, which are only a small proportion of the data.

Figure 4. Posterior predictive check for all 16 conditions. RT: response time.

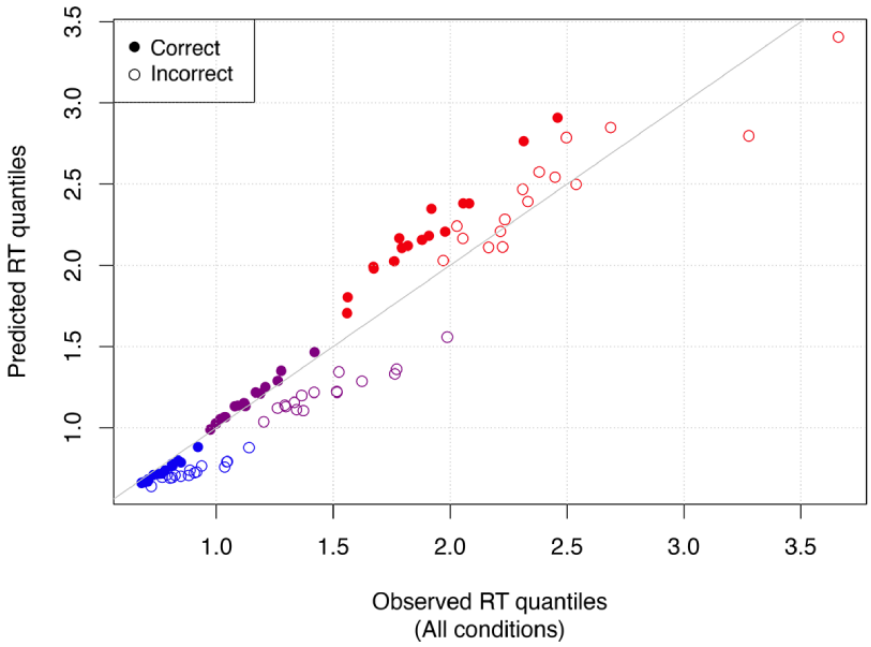

XSL•FO
RenderX

Figure 5. Posterior predictive check for conditions (short study time, 50%-50% distribution of trials, high urgency, single probe). RT: response time.

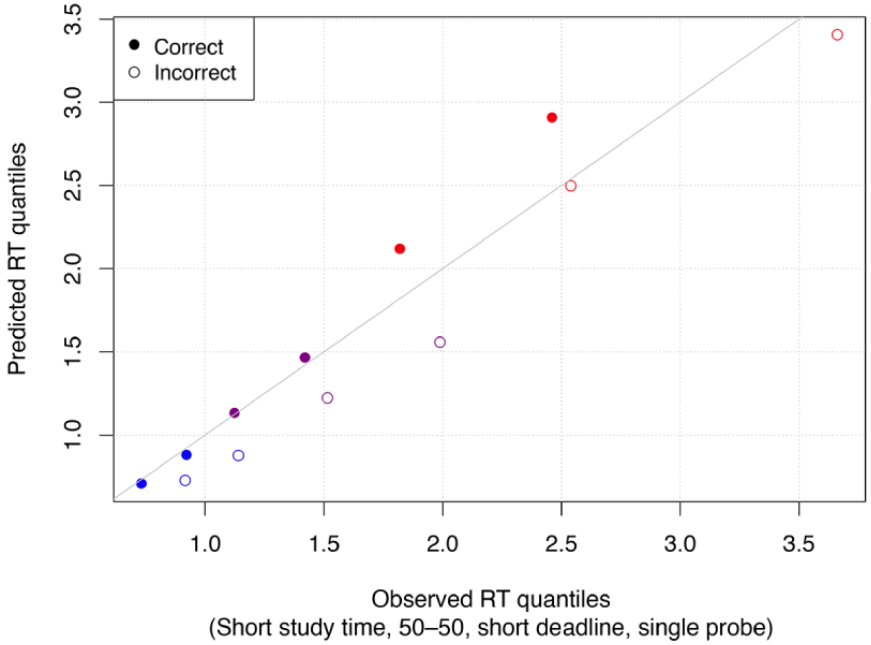

Group-Level Summaries of DDM Parameters

Summaries of DDM parameter estimates are shown for the condition that was aimed for their manipulations in Table 6. Results are separated for the different probe type test array conditions: whole display and single probe. Mean and 95% credible interval estimates are based on the posterior distribution of the corresponding parameters. The credible interval is a Bayesian statistic that can quantify the uncertainty that the true parameter value lies within the interval. A 95% credible interval indicates that the interval has a 95% probability of containing the true value. If 0 is included in the 95% credible interval, we conclude that there is no credible evidence that the parameter is different from 0.

For clarity, findings are presented similarly in structure across the three DDM manipulations: (1) drift rate is displayed with the 2 conditions related to study time, (2) initial bias is shown with the probability of change conditions, and (3) boundary separation is shown with choice urgency conditions.

For drift rate, higher absolute δ values represent faster accumulation of evidence. By definition, the initial bias ranges with β values between 0 and 1, with a β=.5 indicating no bias. For the ease of regression model (initial bias estimates were regressed on task conditions), we worked with logit transformed initial bias estimates (see details in Multimedia Appendix 2), which meant that positive estimated β values expressed bias toward *different* response, while negative β values for *same* response. When the corresponding 95% credible interval contained 0, we concluded that there was no credible bias in either way (which was expected for the condition with equal proportion of *same* and *different* trials). Finally, as the boundary separation parameter by definition could only take positive values, we applied a log-transformation to this parameter to ease regression modeling. Higher α values simply relate to farther response boundaries in the log scale as well, just like in the original scale.

While we can see clear differences in Table 6 between the conditions in both probe types, we need to test whether these differences are credible. Therefore, for the manipulated features listed in (1-3), we estimated contrast parameters by taking the difference between each condition (manipulation—baseline) to examine differences on expected cognitive processes. Specifically, we estimated 3 contrast parameters: study-time-contrast (difference in long study time—short study time), probability-of-change-contrast (high change—low change), and choice-urgency-contrast (high urgency—minimal urgency). The Bayesian framework allowed us to derive these contrasts as regular model parameters with posterior distributions, which allows for principled testing of our hypothesis. Table 7 summarizes the estimates for these below. All parameters were estimated simultaneously within a single-step model estimation process. The strength of this approach is the correct carryover of uncertainty and avoiding bias in sequential estimations [53]. We discuss results for each DDM parameter manipulation next.

XSL•FO
RenderX

Table 6. Posterior summaries of the drift diffusion model parameters per conditions[a].

Parameter, condition, and probe type	Posterior mean (SD)	95% credible interval
Drift rate (δ)		
Long study time[a]		
Whole display	1.37 (0.12)	1.13 to 1.58
Single probe	0.90 (0.13)	0.66 to 1.12
Short study time		
Whole display	1.65 (0.18)	1.37 to 2.00
Single probe	0.88 (0.09)	0.70 to 1.06
Initial bias (β)		
High change[a]		
Whole display	0.13 (0.03)	0.07 to 0.20
Single probe	0.10 (0.04)	0.02 to 0.18
Low change		
Whole display	0.07 (0.04)	−0.02 to 0.14
Single probe	−0.05 (0.09)	−0.19 to 0.10
Boundary separation (α)		
High urgency[a]		
Whole display	0.91 (0.09)	0.78 to 1.05
Single probe	0.66 (0.07)	0.56 to 0.77
Minimal urgency		
Whole display	0.92 (0.13)	0.73 to 1.12
Single probe	0.71 (0.08)	0.59 to 0.87

[a]Manipulation condition.

Table 7. Posterior differences of manipulated condition on drift diffusion model parameters.

Parameters, condition, and probe type	Posterior mean (SD)	95% credible interval
Drift rate (δ)		
Long study–short study		
Whole display	−0.28 (0.05)	−0.39 to −0.18
Single probe	0.01 (0.05)	−0.09 to 0.11
Initial bias (β)		
High change–low change		
Whole display	0.06 (0.02)	0.02 to 0.10
Single probe	0.15 (0.02)	0.11 to 0.20
Boundary separation (α)		
High urgency–minimal urgency		
Whole display	−0.01 (0.02)	−0.04 to 0.03
Single probe	−0.04 (0.02)	−0.08 to −0.01

Study Time on Drift Rates

First, we examined the differences in drift rate estimates (δ) between the study time conditions, broken down by probe type conditions (upper part of Table 7). For the posterior means of the study-time-contrast parameter (difference in long study time–short study time), positive difference values indicate faster drift rates in the manipulated long study time (2000 ms) than in the control short study time (500 ms), while negative difference values indicate slower drift rates in the manipulated

XSL•FO
RenderX

condition compared with the baseline condition. As can be seen from Table 7, the drift rate was credibly higher in the control condition in the whole display probe version of the task. While the manipulation had no credible effect in the single probe version of the task.

Our drift rate manipulation did not show the hypothesized effect. We expected faster drift rates when participants were allowed more time to study the visual features; however, we observed no credible change for single probe and slower draft rates as well. Figure 6A illustrates the posterior mean summaries of the estimated average drift rate between study time conditions for the different probe types (see the values in Table 6). Drift rates were credibly faster on whole display trials with shorter time to inspect the study array (baseline condition: mean 1.65, SD 0.18; 95% credible interval 1.37-2.00) than when given more time to study (manipulated condition: mean 1.37, SD 0.12; 95% credible interval 1.13-1.58). This is also illustrated in Figure 6B, which shows that the light gray posterior distribution of the drift rate differences (study-time-contrast parameter) between study time and probe type conditions (mean –0.28, SD 0.05; 95% credible interval –0.39 to –0.18) does not overlap with 0. This finding was unexpected, as we expected that longer time to study the test array would result in faster drift rates. Drift rates on single probe trials did not differ credibly between shorter study times (mean 0.88, SD 0.09; 95% credible interval 0.70-1.06) and longer study times (mean 0.90, SD 0.13; 95% credible interval 0.66-1.12). As shown in Figure 6B, the dark gray posterior distribution of the study-time-contrast parameter overlaps with 0 (mean 0.01, SD 0.05; 95% credible interval –0.09 to 0.11); that is, our drift rate manipulation simply did not influence single probe trials.

Figure 6. Posterior means and posterior differences of drift rate parameters for study time conditions. (A) Drift rate posterior group means; (B) Distribution of drift rate posterior differences.

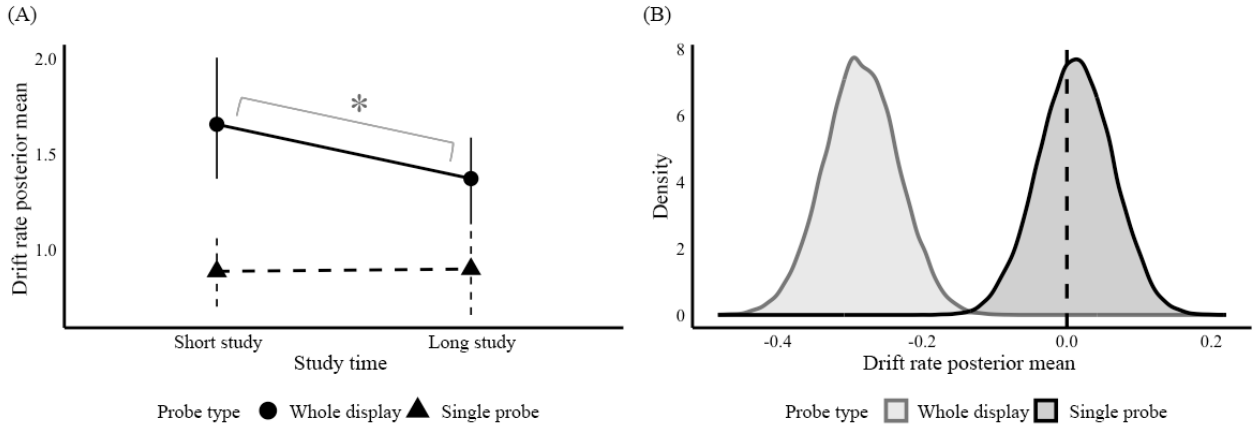

Probability of Change Manipulation on Initial Bias

Second, we examined the estimated probability-of-change-contrast parameter between probe types. This manipulation was intended to influence the initial bias parameter (β) that captures the degree of a priori bias toward *different* or *same* response options. As mentioned earlier, this parameter was logit-scaled, with values at 0 representing no systematic bias toward both binary responses, positive values indicating greater bias toward the *different* response, and negative values indicating bias toward the lower *same* response. Therefore, in Table 7, for the calculated contrast parameter taking the difference in initial bias estimates between the probability of change conditions (high change—low change), the positive values indicated greater a priori biases toward the *different* response boundary in the manipulated high change condition.

We observed positive changes in initial bias with higher probability of change manipulation in both probe types—all contrast values were credibly positive. Figure 7A shows the posterior mean summaries of the initial bias parameter between probability of change conditions and probe types. When tasks were manipulated to have higher probabilities of *different* trials (48/60, 80% of trials) than *same* trials (12/68, 20% of trials), estimates of the initial bias were sensitive to the manipulation and took positive values overall, reflecting a tendency toward selecting the *different* boundary, both on whole display trials (mean 0.13, SD 0.03; 95% credible interval 0.07-0.20), and in single probe trials (mean 0.10, SD 0.04; 95% credible interval 0.02-0.18). As expected, trials with even presentation of *different* and *same* stimuli indicated negligible response bias in both whole display (mean 0.07, SD 0.04; 95% credible interval –0.02 to 0.14), and in single probe trials (mean –0.05, SD 0.09; 95% credible interval –0.19 to 0.10), as shown in Table 5.

XSL•FO
RenderX

Figure 7. Posterior means and posterior differences of initial bias parameter for probability of change conditions. (A) Initial bias posterior group means; (B) Distribution of initial bias posterior differences.

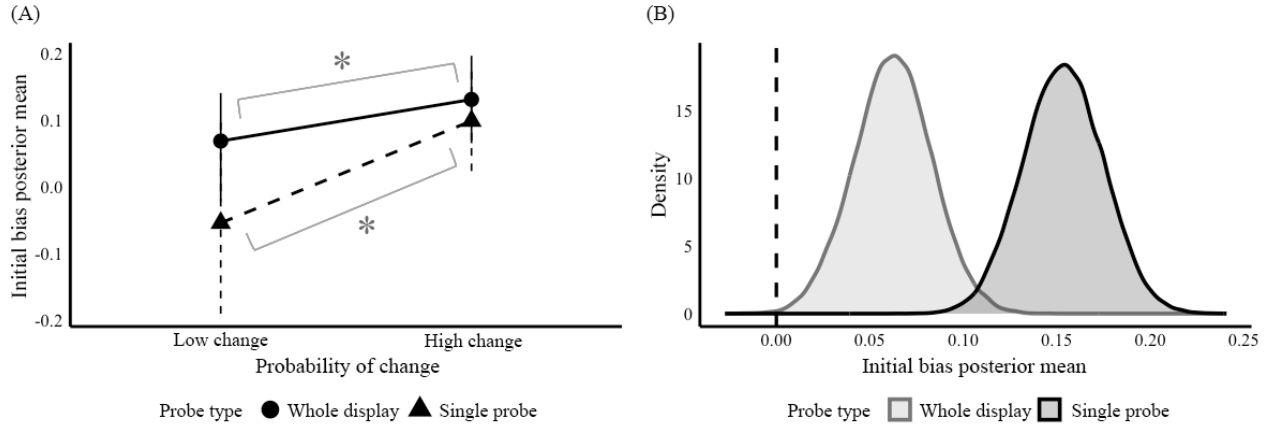

The initial bias manipulation effect is further illustrated in Figure 7B, where the distribution of posterior differences in β per probability of change conditions are shown. The range of the contrast parameter's distributions for both probe types do not include 0, suggesting credible effects of this manipulation on the initial bias parameter across probe types. This effect is more pronounced in the single probe condition (dark gray distribution: mean 0.15, SD 0.02; 95% credible interval 0.11-0.20) than the whole display condition (light gray distribution: mean 0.06, SD 0.02; 95% credible interval 0.02-0.10).

Choice Urgency Manipulation on Boundary Separation

Third, we examined differences in the boundary separation parameter by choice urgency and probe type conditions. This manipulation was intended to influence the response boundary separation: the amount of evidence required to make a decision. Larger boundary separation corresponds to higher threshold values (α). We expected the boundary separations to be lower in the manipulated condition with higher choice urgency to respond (3000 ms) than in the baseline condition with minimal urgency and more time to respond (10,000 ms). The boundary separation parameter was log-transformed to ease the regression modeling, but this scale remained easy to interpret as higher boundary separation values still indicated more information needed before making a decision (with values closer to 0 indicating less information acquired before making a decision). In Table 7, negative choice-urgency-contrast values (high urgency-minimal urgency) indicated lower boundary separation in the manipulated condition (high urgency). Positive differences represented higher response boundary separation in the manipulated high urgency condition than the control minimal urgency condition.

Our boundary separation manipulation worked as expected for the single probe version of the task. The posterior mean summaries of the boundary separation parameter between choice urgency conditions and probe types are shown in Figure 8A, based on values from Table 6. In the single probe trials, this manipulation was successful as boundary separation parameters were sensitive to the different choice urgency conditions (mean –0.04, SD 0.02; 95% credible interval –0.08 to –0.01), with participants exhibiting lower boundary separation with heightened urgency to respond to trials. This contrast parameter for single probe trials is shown in Figure 8B, with almost all the dark gray posterior distribution being <0. Comparatively, looking at the light gray distribution corresponding to the whole display trials, we saw that 0 is almost in the middle of that distribution (mean –0.01, SD 0.02; 95% credible interval –0.04 to 0.03), signifying no credible effect for this manipulation in the whole display condition.

Figure 8. Posterior means and posterior differences of the boundary separation parameter for choice urgency conditions. (A) Boundary separation posterior group mean; (B) Distribution of boundary separation posterior differences.

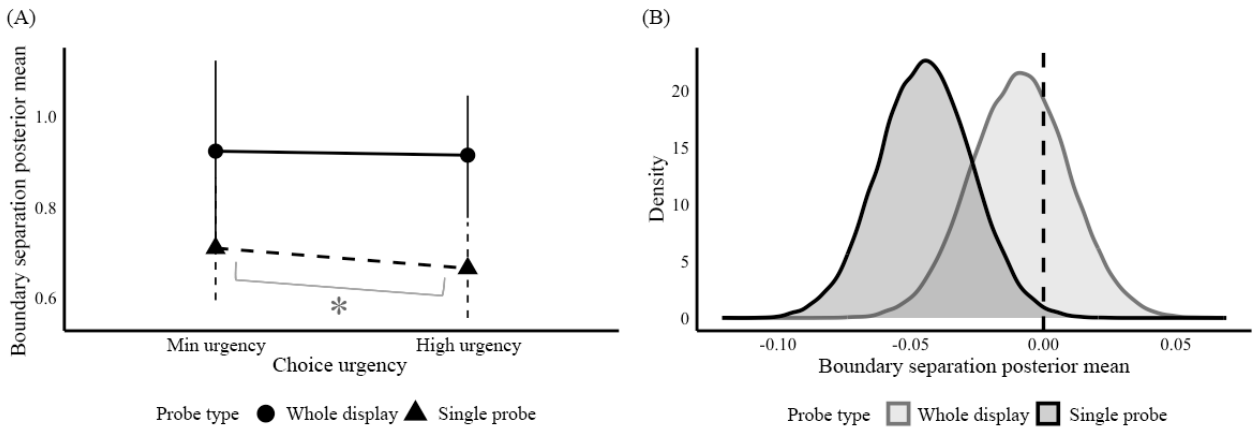

XSL•FO
RenderX

Bayesian Correlations

We explored associations between age and the 3 person-specific parameters of the DDM (drift rate, boundary separation, and initial bias) by running Bayesian correlation analysis in JASP (version 0.17.3) [54]. This analysis yields a regular Pearson correlation coefficient for each tested link and a corresponding Bayes factor that in our case summarizes the amount of evidence for the coefficient to be different from 0 [15]. For instance, Bayes factors of (BF_{10}) 3 to 10 indicate moderate evidence for the correlation, 10 to 30 indicate strong evidence, 30 to 100 indicate very strong evidence and beyond. On the basis of a previous study [24], we predicted that older participants would have slower drift rates and higher boundary separation in decision-making, while we did not expect age to be credibly linked to initial bias. We tested associations across all trials then stratified by probe types. Correlation coefficients and their corresponding Bayes factors are shown in Table 8.

Table 8. Bayesian correlations of drift diffusion model parameters and age.

Parameters and probe type	Pearson *r*	BF_{10} [a]	95% credible interval
Drift rate (δ)			
Overall	−0.45[b]	191.11	−0.61 to −0.23
Whole display	−0.39[c]	32.33	−0.57 to −0.17
Single probe	−0.42	76.04	−0.59 to −0.20
Initial bias (β)			
Overall	0.21	0.62	−0.03 to 0.42
Whole display	0.31	4.26	0.08 to 0.51
Single probe	−0.04	0.16	−0.27 to 0.19
Boundary separation (α)			
Overall	0.32	4.88	0.09 to 0.51
Whole display	0.22	0.74	−0.02 to 0.43
Single probe	0.42	80.76	0.20 to 0.59

[a]Represents ratio of likelihood of evidence in favor of alternative model (1) than null model (0).

[b]BF_{10}>30.

[c]BF_{10}>10.

As expected, the Pearson *r* coefficients capturing association between age and overall drift rate indicated a credibly negative association with age (*r*=–0.45; 95% credible interval –0.61 to –0.23; BF_{10}=191.11), suggesting that older participants have slower drift rates. Results indicated extreme evidence for this association and that the data were at least 191 times more likely under the alternative hypothesis of correlation than the null hypothesis of no correlation. We found similar links when we broke down this association by probe type. The results indicated very strong evidence for moderate negative association between drift rate and age in single probe trials (*r*=–0.42; 95% credible interval –0.59 to –0.20; BF_{10}=76.04) and whole display trials (*r*=–0.39; 95% credible interval –0.57 to –0.17; BF_{10}32.33). As expected, we did not find evidence for association between initial bias and age across both probe type trials (*r*=0.21; 95% credible interval −0.03 to 0.42; BF_{10}0.62). People's age was not expected to be linked with whether they are more likely to choose *different* or *same* on a trial. However, when we separated the probe types, we found some weak evidence of association in the whole display condition (*r*=0.31; 95% credible interval 0.08-0.51; BF_{10}=4.26). Importantly, the single probe condition showed evidence for the lack of correlation (*r*=−0.04; 95% credible interval −0.27 to 0.19; BF_{10}=0.16)—this can be seen by taking the reciprocal of the corresponding Bayes factor (BF_{10}=1/0.16=6.25), which represented 6 times more evidence for the lack of correlation.

The Pearson *r* coefficients showed a positive association between age and boundary separation (*r*=0.32; 95% credible interval 0.09-0.51; BF_{10}=4.88). This suggested that older participants had higher boundary separation. Across all probe type trials, and the Bayes factor indicated moderate evidence and that the data were at least 4 times more likely under the alternative hypothesis of correlation than the null hypothesis of no correlation. Importantly, this boundary separation association was driven by the single probe trials. The correlation between boundary separation and age in single probe trials was *r*=0.42 (95% credible interval 0.20-0.59; BF_{10}=80.76), with very strong evidence. For the link between age and boundary separation in whole display, we did not find a credible association and the Bayes factor provided no evidence (BF_{10}≈1) for the alternative hypothesis (*r*=0.22; 95% credible interval –0.02 to 0.43; BF_{10}=0.74).

Strategy Use

Subsequently, we conducted a post hoc qualitative analysis of cognitive strategy use in light of the unexpected effect of study time on drift rates. As part of the debrief and exit survey, participants were presented with a brief series of unstructured

XSL•FO
RenderX

questions. Summary of strategy use is presented in Table 9, and all the raw responses are reported in Multimedia Appendix 5. From the 68 participants, 81% (n=55) reported using some strategy, whereas 13% (n=9) reported no strategy used. Individuals who responded "Yes" were presented with an open-ended response to elaborate on specific strategies. One common theme emerged upon reviewing the responses. Many strategies involved using some tactic involving ≥2 shapes. Responses were categorized based on whether individuals reported any mention of 2 shapes or not. Approximately 60% (n=33) reported some strategy using 2 shapes, whereas the other 40% (n=22) individuals reported a different strategy that did not involve 2 shapes. For instance, participants reported using mnemonic strategies, such as organizing objects by size, creating heuristics with objects, or integrating the array into a single object.

Table 9. Summary of responses related to cognitive strategy use (N=68) and post hoc analysis of reported cognitive strategy use.

Cognitive strategy reported	Responses, n (%)
Used some strategy	55 (81)
Description includes 2 shapes	33 (60)
Any strategy that does not include 2 shapes	22 (40)
Did not report any strategy	9 (13)
No response	4 (6)

Discussion

Overview

We explored the feasibility of fitting a computational model (ie, the DDM) to ambulatory Color Shapes data by experimentally manipulating features of the Color Shapes task. Our proposed approach can integrate theory-based working memory measures with computational cognitive methodology to capture computational model–based metrics of cognitive ability in daily life settings. This approach was powered by the Bayesian statistical engine, which allowed for simultaneous estimation of all latent features, as well as person- and condition-specific variations in it.

Overall, the results of this study revealed several patterns that emerged consistently across conditions. The hierarchical Bayesian model provided a good fit to the data, confirmed via posterior predictive checks. Of the 3 experimental manipulations, the initial bias and response boundary conditions effectively shifted the corresponding cognitive parameters in the single probe condition. However, the drift rate manipulation effect did not work for the single probe condition and had the opposite effect in the whole display condition. In terms of individual differences in DDM parameters, age was meaningfully associated with drift rate and boundary separation, as was expected based on previous studies on aging [55,56].

Principal Findings

We presented a novel ambulatory adaptation of the Color Shapes task. The Color Shapes task has previously been established in the literature as a sensitive and specific measure of symptomatic and asymptomatic preclinical AD [6,7,25]. RT and accuracy data from this task have not been analyzed previously in a drift diffusion modeling framework. Our approach focused on disentangling subtle cognitive processes underlying manifest Color Shapes task performance with the DDM.

Specifically, we aimed to optimize the Color Shapes task for drift diffusion modeling in ambulatory assessment settings. Models, such as the DDM can incorporate speed and accuracy characteristics in RT data that may offer novel insights into cognitive changes that account for the speed-accuracy trade-off [14]. For this, we tested whether we are able to experimentally manipulate 3 parameters of the DDM by changing features of the Color Shapes task. We compared whole display and single probe versions across these manipulations. We also tested whether individual-level DDM parameters were meaningfully associated with participant's age, given previous studies showed that older participants tended to have declines in speed of cognitive processing [6] and monitor responses to prioritize accuracy over swift responses [57]. Finally, we looked at whether participants would be likely to use strategies for performing better on the Color Shapes task.

Our results showed that 2 Color Shapes property manipulations affected cognitive processes as we expected, as captured by the DDM parameters. First, when we changed the proportion of *different* versus *same* trials toward *different*, the initial bias shifted, indicating subtle preference toward the *different* response in both single probe and whole display versions of the task. Second, when we manipulated the boundary separation feature by triggering choice urgency, we observed the expected effect only in the single probe version, with practically no effect in the whole display version. In the single probe condition, imposing a higher choice urgency (ie, 3000 ms response window before autoadvancement) resulted in a lower response caution, as indicated by a closer boundary separation, relative to the minimal urgency condition with a longer response window (10,000 ms). However, this manipulation did not show this similar effect in the whole display condition.

Our results showed that the presentation duration manipulation did not yield the hypothesized effect on drift rate. Specifically, increasing the amount of time to study the array did not increase the drift rate in the single probe version but produced slower drift rates in whole display trials. While our decision to include this manipulation was based on previous studies using similar methods [48], and making allowances for the difference between laboratory studies and our ambulatory setting, it is likely that our short 500 ms duration already put most subjects at ceiling level for presentation duration (see the study by Ratcliff [58] for the equivalent finding in a laboratory setting). Successfully

XSL•FO
RenderX

manipulating the drift rate for the Color Shapes task in question might require a different approach, possibly making use of visual masking. Indeed, Smith [59] argues that the critical factor distinguishing whether presentation duration manipulation works is whether or not backward masks are used to limit the information extracted from the display. Note, however, that this failure to find the hypothesized effect does not otherwise speak to the validity of the study's premise of testing the DDM, whose parameter estimates appropriately reflect the pattern in the data.

Altogether, we conclude that the single probe version of the Color Shapes task is better optimized for capturing DDM-based cognitive process dissociations in ambulatory settings than the whole display version. The single probe version might also more effectively activates a single decision-making process, in line with DDM assumptions [34]. In contrast, the whole display version may engage multi-strategy decision processes, which could obscure the attribution of cognitive shifts and diminish the validity of model parameter estimates [60]. In addition, the correlations analysis between individual-level DDM parameters and age was more consistent with the single probe version: (1) there was evidence for links between age and drift rate, which is in line with previous findings and (2) evidence for the lack of correlation between age and initial bias, as expected [61,62]. The whole display version did not show evidence for correlation between age and boundary separation and had some evidence for unexpected weak correlation between age and initial bias. Post hoc qualitative data analysis showed that most of the unwanted strategy use was related to encoding only 2 shapes out of the 3 in the study array. This strategy would not be feasible for participants in the single probe version of the task.

Our results indicated that this ambulatory task version (single probe version of the Color Shapes task with short study time [500 ms], low probability of change [50%-50% *different* or *same*], and minimal choice urgency) responded well to the intended manipulations and was associated with convergent and divergent constructs as theorized. Giving 500 ms to study has been commonly used in clinical settings and might also prevent verbal encoding of the stimuli. We found that drift rates were higher in the 500-ms condition, which might support a more 1-shot decision-making style assumed by the DDM. The other two features of the task, that is, even presentation of stimuli and no strict limit on time to respond are also typical in the literature. In terms of raw data, this condition had an average reaction time of 1252.87 ms (95% central quantile range 800.43-1799.48 ms) with a 76.82% average accuracy rate (95% central quantile range 57.63%-96.63%), which are also in range of typical data modeling with DDM. These results align with broader research demonstrating that DDM parameters are sensitive to subtle cognitive variations associated with aging, AD, and a range of neuropsychiatric conditions, including obsessive compulsive disorder, attention-deficit/hyperactivity disorder, and Parkinson disease [63-66].

This optimization of an ambulatory version of the Color Shapes task for accessibility and computational cognitive modeling advances the development of digital tools that extract novel digital markers to empower individuals to initiate health services earlier. By implementing the DDM to disentangle ambulatory cognitive performance data into more nuanced processes underlying performance, we can open opportunities to integrate cognitive aging theoretical frameworks with early detection advances ushered in by technological advances in computational tools [67] and multi-model data [68]. Our study integrated the Color Shapes task linked with preclinical AD-related cognitive changes with computational cognitive methodology and focused on their respective methodological strengths for capturing cognitive processes in daily life settings. This approach is powered by the Bayesian statistical engine that allows a person-centered approach for quantifying individual-specific risk probabilities, for example, in terms of DDM parameter estimates.

Taken together with drift diffusion modeling, the results found that this approach was effective in measuring subtle differences in latent cognitive processes from observed performance. This approach has the potential to extract sensitive digital-based cognitive indicators that can reveal subtle cognitive decline at earlier stages of AD and related dementias. The longitudinal implementation of this approach holds promise for being able to monitor these digital cognitive features over short and long periods. On a broader scale, this approach can be effectively tailored for designing person-centered treatment plans for monitoring cognitive health status over adulthood.

Limitations and Future Directions

Finally, this study had some limitations. First, there may be underlying bias related to participants and study selection. For instance, the population of individuals that register for research participation may be qualitatively compared with individuals that do not sign-up. Second, cognitive health status was assessed via self-reported information and may be bolstered with biological or clinical validation. Third, there may be relevant reasons for missing data on cognitive performance that can be important to address in future computerized assessment designs. Fourth, we applied fixed RT cutoffs informed by domain expertise, and recommend that future work explore alternative, more flexible approaches to managing RT outliers. Fifth, participants were not formally screened for color vision deficiencies, which may have affected performance despite the task being designed using color palettes and shape elements informed by visual accessibility guidelines [69].

As this study was conducted entirely remotely, it may require additional consideration to control for unintended cognitive strategy use. We opted to ask participants at the end regarding any cognitive strategy use, rather than providing explicit instructions that may potentially bias performance data. As the Color Shapes task is intended to assess cognitive processes underlying visual working memory [5,7], using unwanted cognitive strategies using different cognitive domains (eg, verbal encoding), the Color Shapes task could fail to measure changes in visual short-term memory system and bias results. Designs to assess subtle changes associated with cognitive impairment may benefit from conducting a priori technical optimization to identify best strategies to control for unwanted cognitive strategy use. By optimizing the task for unsupervised remote testing, this aids in designing valid and reliable ambulatory cognitive instruments.

XSL•FO
RenderX

AD has a long preclinical phase marked by neuropathological changes [70]. While early stages of preclinical AD can be detected using biomarkers of amyloidosis and neurodegeneration (eg, cerebrospinal fluid, brain imaging, and blood [71]), mobile cognitive testing offers an easily accessible early monitoring tool. Due to the nature of subtle progressive cognitive decline in AD, individuals may conflate pathological cognitive decline with normative age-related changes and consequently delay seeking timely medical care [72] or wait until symptoms have worsened [73]. One overarching goal of this study was to develop an optimized version of the Color Shapes task that is effective for unsupervised, high-frequency cognitive testing. Our results showed evidence that mobilizing smartphone technology to administer ambulatory cognitive assessments in this manner is a viable approach for sampling a representative measure of cognitive status variability in daily contexts. A recent survey reported that 70% of Americans want to know if they were at risk for preclinical AD for seeking earlier treatment, however, only 60% of Americans were reluctant to discuss mild cognitive impairment symptoms and early signs with their health care providers [73]. Commonly cited reasons included concerns with receiving incorrect diagnosis, treatment, or waiting until symptoms did not resolve. Our results were promising as single probe trials may be less cognitively loading and may support long-term adherence. This study demonstrates an effective and feasible design to address related to effective screening and earlier intervention. Finally, although this study focused on deriving cognitive process parameters using the Color Shapes task, future research should explore how these parameters relate to performance on standard neuropsychological assessments. Mapping these associations would help contextualize DDM parameters and facilitate their application in cognitive and clinical settings.

Conclusions

Collecting data with the Color Shapes task in ambulatory settings and analyzing it using a drift diffusion modeling approach yielded digital markers of key cognitive processes. We found that the single-probe (test array with 1 colored shape) version of the Color Shapes task with 500 ms (short) study time, even presentation (50% to 50%) of *different* and *same* trials, and minimal response urgency exhibited expected performance and associations. These novel digital cognitive features could facilitate more sensitive and earlier screening of individuals for secondary AD and related dementia, enable earlier intervention, and support efforts to target modifiable risk factors [1,3].

Acknowledgments

ZO, MJS, JGH, and KDH were supported by the NIH (National Institutes of Health; R01 AG074208, R56 AG074208, and U2 CAG060408). JGH was supported by the NIH under award numbers R00AG056670 and R33AG078084. JV was supported by a National Science Foundation grant 2051186. KDH was supported by a T32 AG049676 training grant. This work used assessments developed under the National Institute on Aging grant U2 CAG060408.

Conflicts of Interest

KDH is a full-time employee of CogState Ltd, which develops cognitive assessments. This study did not use any CogState products. All other authors declare no conflicts of interest.

Multimedia Appendix 1
The Color Shapes task shapes and color palette.
[DOCX File , 61 KB - formative_v9i1e66300_app1.docx]

Multimedia Appendix 2
Drift diffusion model.
[DOCX File , 19 KB - formative_v9i1e66300_app2.docx]

Multimedia Appendix 3
Age distribution of participants (N=68).
[DOCX File , 52 KB - formative_v9i1e66300_app3.docx]

Multimedia Appendix 4
Summary of included and excluded trials.
[DOCX File , 15 KB - formative_v9i1e66300_app4.docx]

Multimedia Appendix 5
Open-ended responses on cognitive strategy use collected during the debrief survey.
[DOCX File , 27 KB - formative_v9i1e66300_app5.docx]

References

XSL•FO
RenderX

1. Sperling RA, Aisen PS, Beckett LA, Bennett DA, Craft S, Fagan AM, et al. Toward defining the preclinical stages of Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement 2011 May 22;7(3):280-292 [FREE Full text] [doi: 10.1016/j.jalz.2011.03.003] [Medline: 21514248]
2. Tsoi KK, Chan JY, Hirai HW, Wong SY, Kwok TC. Cognitive tests to detect dementia: a systematic review and meta-analysis. JAMA Intern Med 2015 Sep 01;175(9):1450-1458. [doi: 10.1001/jamainternmed.2015.2152] [Medline: 26052687]
3. Ratcliff R, McKoon G. Can neuropsychological testing be improved with model-based approaches? Trends Cogn Sci 2022 Nov;26(11):899-901 [FREE Full text] [doi: 10.1016/j.tics.2022.08.015] [Medline: 36153231]
4. Hakun JG, Elbich DB, Roque NA, Yabiku ST, Sliwinski M. Mobile Monitoring of Cognitive Change (M2C2): high-frequency assessments and protocol reporting guidelines. PsyArXiv Preprint posted online on August 23, 2024 [FREE Full text] [doi: 10.31234/osf.io/34ux5]
5. Sliwinski MJ, Mogle JA, Hyun J, Munoz E, Smyth JM, Lipton RB. Reliability and validity of ambulatory cognitive assessments. Assessment 2018 Jan 15;25(1):14-30 [FREE Full text] [doi: 10.1177/1073191116643164] [Medline: 27084835]
6. Parra MA, Abrahams S, Logie RH, Méndez LG, Lopera F, Della Sala S. Visual short-term memory binding deficits in familial Alzheimer's disease. Brain 2010 Sep;133(9):2702-2713. [doi: 10.1093/brain/awq148] [Medline: 20624814]
7. Parra MA, Calia C, Pattan V, Della Sala S. Memory markers in the continuum of the Alzheimer's clinical syndrome. Alzheimers Res Ther 2022 Sep 30;14(1):142 [FREE Full text] [doi: 10.1186/s13195-022-01082-9] [Medline: 36180965]
8. Nicosia J, Aschenbrenner AJ, Balota DA, Sliwinski MJ, Tahan M, Adams S, et al. Unsupervised high-frequency smartphone-based cognitive assessments are reliable, valid, and feasible in older adults at risk for Alzheimer's disease. J Int Neuropsychol Soc 2023 Jun;29(5):459-471 [FREE Full text] [doi: 10.1017/S135561772200042X] [Medline: 36062528]
9. Oravecz Z, Sliwinski M, Kim SH, Williams L, Katz MJ, Vandekerckhove J. Partially observable predictor models for identifying cognitive markers. Comput Brain Behav 2025 Mar 24;97(6):1038-1050. [doi: 10.1007/s42113-025-00238-8] [Medline: 39891430]
10. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol 2008;4:1-32. [doi: 10.1146/annurev.clinpsy.3.022806.091415] [Medline: 18509902]
11. Patrick K, Griswold WG, Raab F, Intille SS. Health and the mobile phone. Am J Prev Med 2008 Aug;35(2):177-181 [FREE Full text] [doi: 10.1016/j.amepre.2008.05.001] [Medline: 18550322]
12. Koo BM, Vizer LM. Mobile technology for cognitive assessment of older adults: a scoping review. Innov Aging 2019 Jan;3(1):igy038 [FREE Full text] [doi: 10.1093/geroni/igy038] [Medline: 30619948]
13. Milner T, Brown MR, Jones C, Leung AW, Brémault-Phillips S. Multidimensional digital biomarker phenotypes for mild cognitive impairment: considerations for early identification, diagnosis and monitoring. Front Digit Health 2024 Mar 6;6:1265846 [FREE Full text] [doi: 10.3389/fdgth.2024.1265846] [Medline: 38510280]
14. De Boeck P, Jeon M. An overview of models for response times and processes in cognitive tests. Front Psychol 2019 Feb 6;10:102 [FREE Full text] [doi: 10.3389/fpsyg.2019.00102] [Medline: 30787891]
15. Lee MD, Wagenmakers EJ. Bayesian Cognitive Modeling: A Practical Course. Cambridge, MA: Cambridge University Press; 2014.
16. Logie RH, Parra MA, Della Sala S. From cognitive science to dementia assessment. Policy Insights Behav Brain Sci 2015 Oct 01;2(1):81-91. [doi: 10.1177/2372732215601370]
17. Vandekerckhove J, Tuerlinckx F, Lee MD. Hierarchical diffusion models for two-choice response times. Psychol Methods 2011 Mar;16(1):44-62. [doi: 10.1037/a0021765] [Medline: 21299302]
18. Batchelder WH. Cognitive psychometrics: using multinomial processing tree models as measurement tools. In: Embretson SE, editor. Measuring Psychological Constructs: Advances in Model-Based Approaches. Washington, DC: American Psychological Association; 2010:71-93.
19. Wiecki TV, Poland J, Frank MJ. Model-based cognitive neuroscience approaches to computational psychiatry. Clin Psychol Sci 2015 Mar 10;3(3):378-399. [doi: 10.1177/2167702614565359]
20. Aschenbrenner AJ, Balota DA, Gordon BA, Ratcliff R, Morris JC. A diffusion model analysis of episodic recognition in preclinical individuals with a family history for Alzheimer's disease: The adult children study. Neuropsychology 2016 Feb;30(2):225-238 [FREE Full text] [doi: 10.1037/neu0000222] [Medline: 26192539]
21. Ging-Jehli NR, Kuhn M, Blank JM, Chanthrakumar P, Steinberger DC, Yu Z, et al. Cognitive signatures of depressive and anhedonic symptoms and affective states using computational modeling and neurocognitive testing. Biol Psychiatry Cogn Neurosci Neuroimaging 2024 Jul;9(7):726-736. [doi: 10.1016/j.bpsc.2024.02.005] [Medline: 38401881]
22. Souza AS, Frischkorn GT. A diffusion model analysis of age and individual differences in the retro-cue benefit. Sci Rep 2023 Oct 13;13(1):17356 [FREE Full text] [doi: 10.1038/s41598-023-44080-z] [Medline: 37833420]
23. Oravecz Z, Harrington KD, Hakun JG, Katz MJ, Wang C, Zhaoyang R, et al. Accounting for retest effects in cognitive testing with the Bayesian double exponential model via intensive measurement burst designs. Front Aging Neurosci 2022 Sep 26;14:897343 [FREE Full text] [doi: 10.3389/fnagi.2022.897343] [Medline: 36225891]
24. Thapar A, Ratcliff R, McKoon G. A diffusion model analysis of the effects of aging on letter discrimination. Psychol Aging 2003 Sep;18(3):415-429 [FREE Full text] [doi: 10.1037/0882-7974.18.3.415] [Medline: 14518805]

XSL•FO
RenderX

25. Parra MA, Saarimäki H, Bastin ME, Londoño AC, Pettit L, Lopera F, et al. Memory binding and white matter integrity in familial Alzheimer's disease. Brain 2015 May 11;138(Pt 5):1355-1369 [FREE Full text] [doi: 10.1093/brain/awv048] [Medline: 25762465]
26. Didic M, Barbeau EJ, Felician O, Tramoni E, Guedj E, Poncet M, et al. Which memory system is impaired first in Alzheimer's disease? J Alzheimers Dis 2011;27(1):11-22 [FREE Full text] [doi: 10.3233/JAD-2011-110557] [Medline: 21799246]
27. Parra MA, Gazes Y, Habeck C, Stern Y. Exploring the association between amyloid-β and memory markers for Alzheimer's disease in cognitively unimpaired older adults. J Prev Alzheimers Dis 2024 Mar;11(2):339-347 [FREE Full text] [doi: 10.14283/jpad.2024.11] [Medline: 38374740]
28. Butler J, Watermeyer TJ, Matterson E, Harper EG, Parra-Rodriguez M. The development and validation of a digital biomarker for remote assessment of Alzheimer's diseases risk. Digit Health 2024 Jan 23;10:20552076241228416 [FREE Full text] [doi: 10.1177/20552076241228416] [Medline: 38269369]
29. Butler J, Owo AS, Watermeyer TJ, Danso SO, Parra‐Rodriguez MA. Developing an AI algorithm to detect predictors of poor performance in a self-administered, web-based digital biomarker for Alzheimer's disease: proof of concept. Alzheimers Dement 2025 Jan 09;20(S2):1. [doi: 10.1002/alz.092887]
30. Cerino ES, Katz MJ, Wang C, Qin J, Gao Q, Hyun J, et al. Variability in cognitive performance on mobile devices is sensitive to mild cognitive impairment: results from the Einstein aging study. Front Digit Health 2021 Dec 3;3:758031 [FREE Full text] [doi: 10.3389/fdgth.2021.758031] [Medline: 34927132]
31. Thompson LI, Strenger JR, Emrani S, De Vito A, Harrington K, Roque N, et al. 91 remote smartphone-based assessment predicts standard neuropsychological test performance and cerebral amyloid status in cognitively normal older adults. J Int Neuropsychol Soc 2023 Dec 21;29(s1):493-494. [doi: 10.1017/S1355617723006392]
32. Ratcliff R. A theory of memory retrieval. Psychol Rev 1978;85(2):59-108. [doi: 10.1037//0033-295x.85.2.59]
33. Stafford T, Pirrone A, Croucher M, Krystalli A. Quantifying the benefits of using decision models with response time and accuracy data. Behav Res Methods 2020 Oct 30;52(5):2142-2155 [FREE Full text] [doi: 10.3758/s13428-020-01372-w] [Medline: 32232739]
34. Ratcliff R, McKoon G. The diffusion decision model: theory and data for two-choice decision tasks. Neural Comput 2008 Apr;20(4):873-922 [FREE Full text] [doi: 10.1162/neco.2008.12-06-420] [Medline: 18085991]
35. Ratcliff R, Scharre DW, McKoon G. Discriminating memory disordered patients from controls using diffusion model parameters from recognition memory. J Exp Psychol Gen 2022 Jun;151(6):1377-1393 [FREE Full text] [doi: 10.1037/xge0001133] [Medline: 34735185]
36. Make a positive impact by volunteering for research. ResearchMatch & Vanderbilt University Medical Center. URL: https://www.researchmatch.org/ [accessed 2025-05-31]
37. Harris PA, Scott KW, Lebo L, Hassan N, Lightner C, Pulley J. ResearchMatch: a national registry to recruit volunteers for clinical research. Acad Med 2012;87(1):66-73. [doi: 10.1097/acm.0b013e31823ab7d2]
38. Vandekerckhove J. Extensions and applications of the diffusion model for two-choice response times. Katholieke Universiteit Leuven. 2009. URL: http://www.cidlab.com/prints/vandekerckhove2009extensions.pdf [accessed 2025-05-29]
39. Kim S, Vandekerckhove J, Oravecz Z, Sliwinski M, Hakun JG, Li Y, et al. Digital biomarkers project. Open Science Framework. URL: https://osf.io/ty4uf/ [accessed 2025-08-14]
40. Nilsson H, Rieskamp J, Wagenmakers EJ. Hierarchical Bayesian parameter estimation for cumulative prospect theory. J Math Psychol 2011 Feb;55(1):84-93. [doi: 10.1016/j.jmp.2010.08.006]
41. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A. Bayesian Data Analysis. New York, NY: Chapman and Hall/CRC; 2013.
42. R: a language and environment for statistical computing. The R Core Team. 2025. URL: https://cran.r-project.org/doc/manuals/r-release/fullrefman.pdf [accessed 2025-05-29]
43. Plummer M. JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. DSC 2003 Working Papers. 2003. URL: https://www.r-project.org/conferences/DSC-2003/Drafts/Plummer.pdf [accessed 2025-05-29]
44. Plummer M, Stukalov A, Denwood M. Package 'rjags': Bayesian graphical models using MCMC. CRAN R. 2016. URL: https://cran.r-project.org/web/packages/rjags/rjags.pdf [accessed 2025-05-29]
45. Wabersich D, Vandekerckhove J. Extending JAGS: a tutorial on adding custom distributions to JAGS (with a diffusion model example). Behav Res Methods 2014 Mar 20;46(1):15-28 [FREE Full text] [doi: 10.3758/s13428-013-0369-3] [Medline: 23959766]
46. Brooks SP, Gelman A. General methods for monitoring convergence of iterative simulations. J Comput Graph Stat 1998 Dec;7(4):434-455. [doi: 10.1080/10618600.1998.10474787]
47. Chávez De la Peña AF, Vandekerckhove J. An EZ Bayesian hierarchical drift diffusion model for response time and accuracy. Psychon Bull Rev (Forthcoming) 2025 Jul 25. [doi: 10.3758/s13423-025-02729-y] [Medline: 40715684]
48. Ratcliff R, Smith PL, Brown SD, McKoon G. Diffusion decision model: current issues and history. Trends Cogn Sci 2016 Apr;20(4):260-281 [FREE Full text] [doi: 10.1016/j.tics.2016.01.007] [Medline: 26952739]
49. Voss A, Nagler M, Lerche V. Diffusion models in experimental psychology: a practical introduction. Exp Psychol 2013 Jan 01;60(6):385-402. [doi: 10.1027/1618-3169/a000218] [Medline: 23895923]

XSL•FO
RenderX

50. Ratcliff R, Thapar A, McKoon G. Individual differences, aging, and IQ in two-choice tasks. Cogn Psychol 2010 May;60(3):127-157 [FREE Full text] [doi: 10.1016/j.cogpsych.2009.09.001] [Medline: 19962693]
51. Schmiedek F, Oberauer K, Wilhelm O, Süss HM, Wittmann WW. Individual differences in components of reaction time distributions and their relations to working memory and intelligence. J Exp Psychol Gen 2007 Aug;136(3):414-429. [doi: 10.1037/0096-3445.136.3.414] [Medline: 17696691]
52. Ashby NJ, Fukuda K, Vogel EK. Estimating visual working memory capacity with whole and single probe test arrays. J Vis 2009 Aug 01;9(8):590. [doi: 10.1167/9.8.590]
53. Boehm U, Marsman M, Matzke D, Wagenmakers EJ. On the importance of avoiding shortcuts in applying cognitive models to hierarchical data. Behav Res Methods 2018 Aug 12;50(4):1614-1631 [FREE Full text] [doi: 10.3758/s13428-018-1054-3] [Medline: 29949071]
54. JASP (Version 0.17.3). JASP. URL: https://jasp-stats.org/ [accessed 2025-05-29]
55. Starns JJ, Ratcliff R. The effects of aging on the speed-accuracy compromise: boundary optimality in the diffusion model. Psychol Aging 2010 Jun;25(2):377-390 [FREE Full text] [doi: 10.1037/a0018022] [Medline: 20545422]
56. Gür E, Duyan YA, Türkakın E, Arkan S, Karson A, Balcı F. Aging impairs perceptual decision-making in mice: integrating computational and neurobiological approaches. Brain Struct Funct 2020 Jul 22;225(6):1889-1902. [doi: 10.1007/s00429-020-02101-x] [Medline: 32566973]
57. Rabbitt P. How old and young subjects monitor and control responses for accuracy and speed. British J of Psychology 2011 Apr 13;70(2):305-311. [doi: 10.1111/j.2044-8295.1979.tb01687.x]
58. Ratcliff R. A diffusion model account of response time and accuracy in a brightness discrimination task: fitting real data and failing to fit fake but plausible data. Psychon Bull Rev 2002 Jun;9(2):278-291. [doi: 10.3758/bf03196283] [Medline: 12120790]
59. Smith PL. Attention and luminance detection: effects of cues, masks, and pedestals. J Exp Psychol Hum Percept Perform 2000 Aug;26(4):1401-1420. [doi: 10.1037//0096-1523.26.4.1401] [Medline: 10946722]
60. Forstmann BU, Ratcliff R, Wagenmakers EJ. Sequential sampling models in cognitive neuroscience: advantages, applications, and extensions. Annu Rev Psychol 2016 Jan 04;67(1):641-666 [FREE Full text] [doi: 10.1146/annurev-psych-122414-033645] [Medline: 26393872]
61. Ratcliff R, Thapar A, McKoon G. The effects of aging on reaction time in a signal detection task. Psychol Aging 2001;16(2):323-341. [doi: 10.1037//0882-7974.16.2.323]
62. Theisen M, Lerche V, von Krause M, Voss A. Age differences in diffusion model parameters: a meta-analysis. Psychol Res 2021 Jul 13;85(5):2012-2021 [FREE Full text] [doi: 10.1007/s00426-020-01371-8] [Medline: 32535699]
63. Erhan C, Balcı F. Obsessive compulsive features predict cautious decision strategies. Q J Exp Psychol (Hove) 2017 Jan 01;70(1):179-190. [doi: 10.1080/17470218.2015.1130070] [Medline: 26652727]
64. Erhan C, Bulut G, Gökçe S, Ozbas D, Turkakin E, Dursun OB, et al. Disrupted latent decision processes in medication-free pediatric OCD patients. J Affect Disord 2017 Jan 01;207:32-37. [doi: 10.1016/j.jad.2016.09.011] [Medline: 27690351]
65. Gupta A, Bansal R, Alashwal H, Kacar AS, Balci F, Moustafa AA. Neural substrates of the drift-diffusion model in brain disorders. Front Comput Neurosci 2021 Jan 7;15:678232. [doi: 10.3389/fncom.2021.678232] [Medline: 35069160]
66. Shapiro Z, Huang-Pollock C. A diffusion-model analysis of timing deficits among children with ADHD. Neuropsychology 2019 Sep;33(6):883-892 [FREE Full text] [doi: 10.1037/neu0000562] [Medline: 31094550]
67. Vrahatis AG, Skolariki K, Krokidis MG, Lazaros K, Exarchos TP, Vlamos P. Revolutionizing the early detection of Alzheimer's disease through non-invasive biomarkers: the role of artificial intelligence and deep learning. Sensors 2023 Apr 22;23(9):4184. [doi: 10.3390/S23094184]
68. Vrahatis AG, Vlamos P, Avramouli A, Exarchos T, Gonidi M. Emerging machine learning techniques for modelling cellular complex systems in Alzheimer's disease. In: Proceedings of the 2020 International Conference on Computational Biology and Bioinformatics. 2021 Presented at: GeNeDis '20; September 21-24, 2020; Bali, Indonesia p. 199-208 URL: https://link.springer.com/chapter/10.1007/978-3-030-78775-2_24 [doi: 10.1007/978-3-030-78775-2_24]
69. Wong B. Points of view: color blindness. Nat Methods 2011 Jun 27;8(6):441. [doi: 10.1038/nmeth.1618] [Medline: 21774112]
70. Rajan KB, Weuve J, Barnes LL, McAninch EA, Wilson RS, Evans DA. Population estimate of people with clinical Alzheimer's disease and mild cognitive impairment in the United States (2020-2060). Alzheimers Dement 2021 Dec 27;17(12):1966-1975 [FREE Full text] [doi: 10.1002/alz.12362] [Medline: 34043283]
71. Price JL, Morris JC. Tangles and plaques in nondemented aging and "preclinical" Alzheimer's disease. Ann Neurol 1999 Mar;45(3):358-368. [doi: 10.1002/1531-8249(199903)45:3<358::aid-ana12>3.0.co;2-x] [Medline: 10072051]
72. Harada CN, Natelson Love MC, Triebel KL. Normal cognitive aging. Clin Geriatr Med 2013 Nov;29(4):737-752 [FREE Full text] [doi: 10.1016/j.cger.2013.07.002] [Medline: 24094294]
73. Alzheimer's disease facts and figures. World Health Organization. URL: https://www.alz [accessed 2023-07-20]

Abbreviations

AD: Alzheimer disease

XSL•FO
RenderX

DDM: drift diffusion model
RT: response time

Edited by J Sarvestan; submitted 09.09.24; peer-reviewed by E Fox, F Balci; comments to author 14.02.25; revised version received 11.06.25; accepted 14.07.25; published 01.10.25.

Please cite as:
Kim SH, Hakun JG, Li Y, Harrington KD, Elbich DB, Sliwinski MJ, Vandekerckhove J, Oravecz Z
Optimizing the Color Shapes Task for Ambulatory Assessment and Drift Diffusion Modeling: A Factorial Experiment
JMIR Form Res 2025;9:e66300
URL: https://formative.jmir.org/2025/1/e66300
doi:10.2196/66300
PMID:

©Sharon Haeun Kim, Jonathan G Hakun, Yanling Li, Karra D Harrington, Daniel B Elbich, Martin J Sliwinski, Joachim Vandekerckhove, Zita Oravecz. Originally published in JMIR Formative Research (https://formative.jmir.org), 01.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Identification of Syndrome Types in Patients With Pancreatic Cancer From Free Text in Electronic Medical Records: Model Development and Validation

He Ba[1,2*], PhD; Haojie Du[3,4*], MEng; Chienshan Cheng[2,5], MD; Yuan Zhang[2,5], MD; Linjie Ruan[2,5], MD; Zhen Chen[2,5], Prof Dr Med

[1]Qingdao Institute, Department of Integrative Oncology, Fudan University Shanghai Cancer Center, Qingdao, China

[2]Department of Oncology, Shanghai Medical College, Fudan University, Shanghai, China

[3]Key Laboratory of Surface & Interface Science of Polymer Materials of Zhejiang Province, School of Chemistry and Chemical Engineering, Zhejiang Sci-Tech University, Hangzhou, China

[4]Shengzhou Innovation Research Institute of Zhejiang Sci-Tech University, Shengzhou, China

[5]Department of Integrative Oncology, Fudan University Shanghai Cancer Center, Shanghai, China

[*]these authors contributed equally

Corresponding Author:
Zhen Chen, Prof Dr Med
Department of Integrative Oncology
Fudan University Shanghai Cancer Center
No. 270 Dongan Rd., Xuhui district
Shanghai
China
Phone: 86 13311965818
Email: zchenzl@fudan.edu.cn

Abstract

Background: Syndrome differentiation is crucial in traditional Chinese medicine (TCM) diagnosis and treatment, but it heavily relies on expert experience, limiting systematic standardization.

Objective: This study developed and validated a BERT (bidirectional encoder representations from transformers)–based model, the traditional Chinese medicine pancreatic cancer syndrome differentiation bidirectional encoder representations from transformers (TCMPCSD-BERT), using in-house pancreatic cancer medical records, to digitalize expert knowledge and support standardized syndrome differentiation in TCM.

Methods: A retrospective dataset of pancreatic cancer cases (2011-2024) from Fudan University Shanghai Cancer Center was annotated into 4 TCM syndrome types by 2 experts (Cohen κ=0.913). The proposed TCMPCSD-BERT model was compared with conventional models (long short-term memory and text convolutional neural network) embedded in TCM diagnostic tools and with large language models (LLMs; ChatGPT-4o, Kimi, Ernie Bot 4.0 Turbo, and Zhipu Qingyan) under a prompt engineering framework. Performance evaluation on in-house data was supplemented with attention visualizations and integrated gradients analyses for interpretability. The McNemar test assessed classification accuracy differences, while bootstrap 95% CIs quantified statistical uncertainty and stability. The Welch *t* test (2-tailed) was used to evaluate mean differences between TCMPCSD-BERT and the comparator models.

Results: Among 6830 records, case counts were damp-heat syndrome (n=1694), spleen-deficiency syndrome (n=1185), damp-heat with spleen-deficiency syndrome (n=1178), and others (n=2773). On the test set, McNemar test showed significantly higher accuracy for TCMPCSD-BERT than the 3 baseline models and generally better performance than LLMs. In all comparisons, TCMPCSD-BERT achieved higher mean macroprecision, macrorecall, macro–F_1-score, and accuracy, with nonoverlapping 95% bootstrap CIs and significant Welch *t* test results (P<.01). The model achieved a macroprecision of 0.935 (95% CI 0.918-0.951), macrorecall of 0.921 (95% CI 0.900-0.942), macro–F_1-score of 0.927 (95% CI 0.908-0.945), and accuracy of 0.919 (95% CI 0.899-0.939). Attention visualizations suggested the model could capture less common TCM term associations, while integrated gradients highlighted high-attribution diagnostic features (eg, “gray-white stool” 0.933 in damp-heat syndrome; “indigestion” 1.204 in spleen-deficiency syndrome). Misclassification analyses indicated challenges in handling overlapping or atypical symptom

XSL•FO
RenderX

presentations. Compared with LLMs, web-based platforms, and diagnostic instruments, TCMPCSD-BERT appeared to provide relatively higher accuracy, interpretability, and efficiency in processing long unstructured texts for syndrome differentiation.

Conclusions: The TCMPCSD-BERT model shows potential for automated syndrome differentiation from unstructured clinical texts and broader application in TCM. Based on this study, it appears to improve operability over 4-diagnostic instruments and web-based platforms, and offers greater stability and accuracy than LLMs in specific tasks. However, these findings should be interpreted cautiously, given the subjectivity of syndrome definitions, data imbalance, and reliance on preprocessed, expert-annotated data. Further studies involving larger and more diverse populations are needed to validate its generalizability and support its broader application in real-world settings.

(JMIR Form Res 2025;9:e70602) doi:10.2196/70602

KEYWORDS

pancreatic cancer; traditional Chinese medicine; syndrome differentiation; BERT; natural language processing; bidirectional encoder representations from transformers

Introduction

Background

In recent years, traditional Chinese medicine (TCM) has been explored as a potential adjunctive approach in the integrative management of pancreatic cancer, particularly within clinical settings in China and other East Asian countries [1,2]. Some small-scale clinical studies have suggested that TCM interventions may help relieve symptoms and improve quality of life for patients with pancreatic cancer under specific conditions [3-5]. A core principle of TCM is “treatment based on syndrome differentiation,” which emphasizes tailoring therapies to the patient’s individualized symptom profile. In clinical practice, this process involves evaluating a patient’s tongue appearance, pulse characteristics, mental state, and other observable signs to classify them into specific syndrome types, forming the basis for treatment selection [6]. However, the process of syndrome differentiation in TCM is still largely grounded in traditional Chinese philosophical theories, which are abstract and holistic in nature. It remains heavily dependent on the clinician’s subjective interpretation, thereby introducing considerable variability into diagnostic outcomes. While TCM continues to play a role in culturally embedded health practices, reducing the subjectivity inherent in syndrome differentiation and improving its standardization and reproducibility remain key challenges in the ongoing development of TCM diagnostic methodology.

In recent years, the application of deep learning technologies within the field of TCM has grown significantly, opening up new possibilities for the digital transformation of TCM diagnostics and treatment. In the field of natural language processing (NLP), the BERT (bidirectional encoder representations from transformers) model has emerged as a highly effective tool for semantic understanding due to its bidirectional encoding and self-attention mechanisms [7,8]. As a third-generation NLP model, BERT marks a considerable advancement over earlier approaches. Traditional first-generation models, such as multilayer perceptrons [9], support vector machines [10], and XGBoost (extreme gradient boosting) [11], perform well with structured data, particularly when dealing with explicit textbook features. However, they struggle to handle the contextual semantics present in unstructured clinical records. Second-generation models, such as long short-term memory (LSTM) [12] and text convolutional neural networks (text-CNN) [13], introduced sequence-processing capabilities. While LSTM preserves long-term dependencies through memory cells, and text-CNN focuses on local features for tasks such as phrase extraction, both models fall short in providing a global understanding of semantics. In contrast, BERT’s self-attention mechanism constructs semantic associations between words, capturing long-range dependencies and achieving a comprehensive understanding of context, critical for processing unstructured TCM case texts. This allows BERT to significantly enhance the accuracy and precision of differentiation tasks, treating words as though “distance does not matter.”

In recent years, several general-purpose large language models (LLMs), such as ChatGPT [14], Ernie Bot [15], and Kim [16], as well as TCM-specific LLMs, including Zhipu Qingyan [17], have been developed and widely adopted. These models, leveraging powerful computational capabilities and extensive knowledge bases, have demonstrated significant potential in advancing the digital transformation of clinical diagnostics and treatment. However, challenges such as the “hallucination problem” [18] and “long-text forgetting” [19] remain common issues faced by large models across various application domains, significantly affecting their accuracy and stability in syndrome differentiation tasks. Compared to modern medicine, the availability of public clinical case corpora in the field of TCM is relatively scarce, and most TCM case records have not been systematically annotated. For LLMs, training requires a large amount of high-quality annotated data to enhance their performance in specific domains, but the limited availability of TCM-specific data severely restricts the application and accuracy of models in TCM-related tasks. Additionally, the complexity of TCM terminology and its unique linguistic style pose additional challenges for the debugging and application of LLMs. Many TCM terms are rarely encountered in modern medical literature, and some terms appear in a semiclassical Chinese format, making it difficult for models to accurately understand and process these terms. As a result, LLMs often struggle to correctly recognize and comprehend TCM terms within clinical cases, leading to potential inaccuracies in syndrome differentiation results, which in turn affects the reliability of clinical decision-making.

XSL•FO
RenderX

Furthermore, despite the release of several syndrome differentiation platforms [20] and 4-diagnostic instruments [21-23] in the field of TCM, these web-based platforms and offline hardware still rely primarily on structured input due to limitations in computational power and memory. This structured input method often requires manual intervention and preprocessing, which results in limited operability and flexibility. In TCM clinical practice, different diseases possess unique diagnostic characteristics, meaning that the same symptoms may have significantly different implications for syndrome differentiation depending on the disease. For example, when inputting syndrome differentiation elements for pancreatic cancer "damp-heat syndrome (DH)," the platform might mistakenly identify it as a different disease, such as a hepatitis-related syndrome, failing to accurately assess the pancreatic cancer syndrome content, leading to a lack of diagnostic specificity in the output. In contrast, the BERT-based end-to-end approach holds the promise of achieving more accurate, efficient, and specific performance for the task of syndrome differentiation in pancreatic cancer.

Objective

This study aims to explore the feasibility of a BERT-based model, traditional Chinese medicine pancreatic cancer syndrome differentiation bidirectional encoder representations from transformers (TCMPCSD-BERT), trained on 6830 clinical records, in supporting syndrome classification in pancreatic cancer. The model was developed to explore whether deep learning techniques may help mitigate subjectivity and enhance the operability of syndrome differentiation from unstructured TCM narratives. Its performance was evaluated against conventional neural network models commonly embedded in existing TCM diagnostic tools [21-23], as well as several representative LLMs under a prompt-based framework.

Methods

TCMPCSD-BERT Model Training Pipeline

The training pipeline of the TCMPCSD-BERT model is illustrated in Figure 1, which outlines the key procedures involved in the model's development. Each module shown in the figure corresponds to a specific methodological component, which is described in further detail in the subsequent subsections of the Methods section. The process began with text cleaning, during which clinical narratives were manually reviewed to remove historical descriptions and resolved symptoms considered irrelevant to the current syndrome classification task. As indicated by the red strikethroughs in Figure 1, only current symptoms and select historical details with potential diagnostic relevance were retained to facilitate the model's learning of syndrome-related patterns (see "Data Source and Patient Population" and Multimedia Appendix 1 for details). Subsequently, relevant diagnostic features were extracted from the cleaned unstructured texts and mapped to standardized syndrome elements based on established TCM diagnostic guidelines. These steps are illustrated in Figure 1 through shaded text segments and color-coded arrows (see "Data Source and Patient Population" and "Multimedia Appendix 1" for details). Based on these extracted and mapped features, all records were labeled according to corresponding syndrome categories. The annotated records were then transformed into 1-hot encoded vectors to meet the model's input requirements. During the training phase, the model was fine-tuned using a fully connected output layer followed by a binary cross-entropy loss function, which optimized the model parameters for a multilabel classification task. As shown in Figure 1, the loss function computed discrepancies between predicted and actual labels and updated the model weights via backpropagation (highlighted in orange), gradually improving classification performance over successive iterations (see "TCMPCSD-BERT Model Architecture and Training Strategy" and Multimedia Appendix 2 for details). Following training, the optimized model was preliminarily applied to infer syndrome types from previously unseen TCM clinical texts. When presented with previously unseen, unlabeled clinical narratives, the model inferred syndrome labels along with associated confidence scores, using the optimized parameters fixed during training (highlighted in red in Figure 1).

As all clinical narratives used in training were originally written in Chinese, the examples in Figure 1 have been translated into English for illustrative clarity. A detailed Chinese-English mapping of figure components is available in Table S1 (Multimedia Appendix 3).

XSL•FO
RenderX

Figure 1. Training pipeline of TCMPCSD-BERT for syndrome differentiation in pancreatic cancer. TCMPCSD-BERT: traditional Chinese medicine pancreatic cancer syndrome differentiation bidirectional encoder representations from transformers.

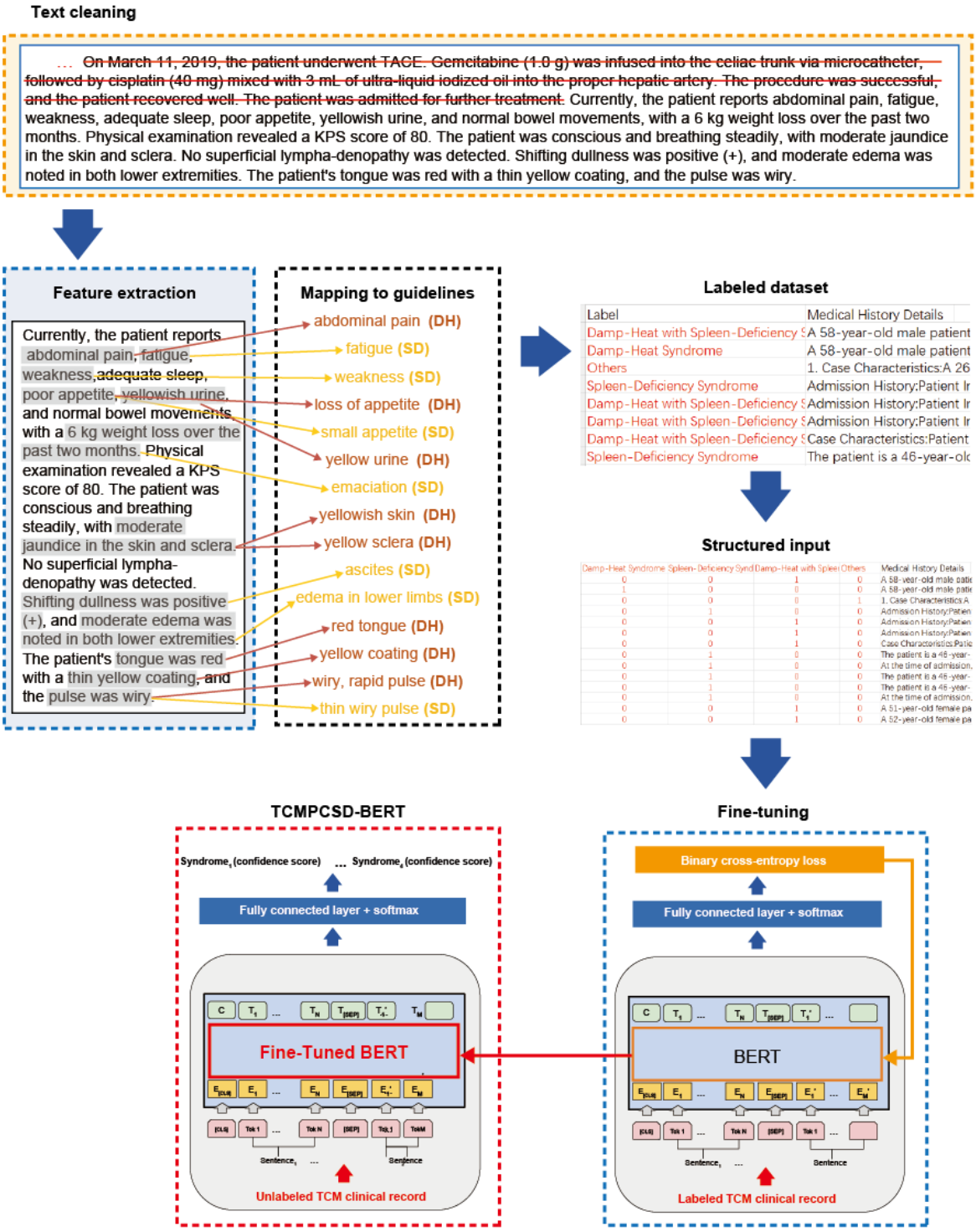

Data Source and Patient Population

This study retrospectively collected clinical records of patients with pancreatic cancer hospitalized between January 2011 and September 2024 from the electronic health record system of the Department of Integrative Oncology at Fudan University Shanghai Cancer Center. To ensure patient privacy, all records were rigorously deidentified and anonymized, removing personally identifiable information such as names, ID numbers, contact details, and medical record numbers. The data were used exclusively for research purposes, with no commercial use or public release.

This study included only patients with pathologically confirmed pancreatic cancer. Individuals diagnosed with pancreatic neuroendocrine tumors or nonpancreatic primary malignancies were excluded. In constructing the TCM syndrome classification model, we selected only those syndrome types that were represented by at least 500 cases (for example, DH was included only if it had 500 or more records). Cases that did not clearly exhibit a recognizable syndrome pattern according to established

XSL•FO
RenderX

diagnostic criteria were grouped under the category "others." Likewise, syndrome types with insufficient sample sizes—such as blood-stasis syndrome, which was represented by fewer than 500 records—were also collectively categorized as "others." This inclusion threshold was primarily based on the potential modeling limitations that deep learning architectures may encounter under conditions of class imbalance.

Previous studies [24-26] have suggested that in BERT-based text classification tasks, the semantic features of a minority class may not be adequately learned when the number of training examples for that class is substantially lower than that of the others. Specifically, because the BERT model is typically trained by minimizing an overall loss function such as cross-entropy, classes with fewer samples contribute relatively little to the gradient updates. This may lead the model to focus disproportionately on majority classes during training, while underrepresenting or even neglecting minority classes during inference, thereby resulting in notably reduced recall for those categories. Moreover, underrepresented classes often lack sufficient contextual diversity to support robust representation learning, which may compromise the model's ability to capture distinctive linguistic patterns associated with these categories and, in turn, affect its overall classification stability and generalizability. In light of these considerations, and to ensure sufficient training signal for learning reliable representations, this study prioritized the inclusion of relatively well-represented syndrome types as classification targets.

This study followed the syndrome differentiation criteria for pancreatic cancer outlined in "Integrated Chinese and Western Medicine in Oncology" [27] and "Principles of Syndrome Differentiation and Treatment in TCM" [28]. Two TCM physicians, each with over 10 years of clinical experience, independently reviewed each case and performed feature annotation. To evaluate the consistency between the 2 experts' annotations, this study, following relevant research [29], used Cohen κ coefficient as a quantitative measure. In instances where discrepancies or ambiguities arose between the 2 experts' annotations, a third TCM physician—also with more than 10 years of clinical experience—was invited to participate in the review. The 3 physicians jointly re-examined the case and engaged in thorough discussion until a consensus was reached, which was then adopted as the final annotation for that case. Detailed information on feature engineering and data preprocessing can be found in Multimedia Appendix 1.

TCMPCSD-BERT Model Architecture and Training Strategy

The TCMPCSD-BERT model is based on the BERT architecture and effectively captures semantic relationships and contextual information of vocabulary through input embeddings, positional encoding, and self-attention mechanisms. In the masked language model task, the model randomly masks certain tokens, forcing it to predict the missing words using contextual information, thereby enhancing its understanding of the text's context. Through iterative optimization of parameters, the model progressively improves its accuracy in identifying TCM syndrome features. Ultimately, by incorporating fine-tuning and a multiclass classification output layer, TCMPCSD-BERT achieves the automatic identification and classification of syndrome features in the clinical texts of patients with pancreatic cancer. Detailed information on the TCMPCSD-BERT model architecture and training strategy can be found in Multimedia Appendix 2.

Experiment Setup

All algorithms in this study were implemented using Python (version 3.12; Python Software Foundation). The experimental hardware configuration included an NVIDIA RTX 4090 GPU (25.2 GB VRAM), a 16-core AMD EPYC 9354 CPU, and 60 GB of memory.

Based on the proportion of case records for each syndrome type, the dataset was stratified and randomly split into a training and validation set (6147 entries) and a test set (683 entries) at a ratio of 9:1, ensuring that the distribution of classes in the training, validation, and test sets was consistent with the original dataset. Given that BERT model training requires a substantial amount of data, the 6147 text entries constitute a relatively limited dataset. To improve the generalization capability and stability of the model, 10-fold cross-validation was used. Specifically, the dataset was randomly divided into 10 subsets, with 9 subsets used for training and 1 subset for validation in each iteration. This process was repeated 10 times, allowing each subset to serve as the validation dataset once. The final model performance was calculated as the average of all validation results.

The training procedure followed the BERT multiclass classification fine-tuning strategy proposed by Sun et al [30] in 2019, incorporating their methods for data preprocessing and hyperparameter adjustment. The Adam optimizer was selected for optimization, as it dynamically adjusts the learning rate based on gradient variations, facilitating faster convergence and improved model performance. To prevent overfitting, we implemented dropout regularization, which enhances model generalizability by randomly deactivating a subset of neurons during training.

For hyperparameter tuning, we used grid search to systematically adjust parameters such as learning rate, batch size, and number of training epochs. Each combination was evaluated on the validation dataset, with the combination yielding the highest accuracy and F_1-score selected to optimize the model's performance in the syndrome differentiation task.

To assess the model's performance in syndrome differentiation for this multiclass classification task, the metrics of macroprecision, macrorecall, macro–F_1-score, and accuracy were used, as recommended in previous studies [31].

Model Baseline

In line with previous studies [21-23], this research uses LSTM and text-CNN as baseline models. To ensure consistency in experimental design, these baseline models were trained using the same 10-fold cross-validation approach as TCMPCSD-BERT, with identical training and validation datasets. Additionally, a BERT model without fine-tuning on clinical data was included as a baseline. We assessed the performance of TCMPCSD-BERT, the non–fine-tuned BERT,

XSL•FO
RenderX

LSTM, and text-CNN models on the same test dataset, comparing macroprecision, macrorecall, macro–F_1-score, and accuracy metrics to provide a comprehensive evaluation of each model's effectiveness in the TCM syndrome differentiation task. The detailed definitions and calculation procedures for these 4 evaluation metrics are provided in Multimedia Appendix 4.

To further examine whether the observed performance differences between models were statistically significant, and in line with prior studies using similar evaluation approaches [32], we applied McNemar test for pairwise comparison of classification results. Given the multiclass nature of the task, a "one-vs-rest" strategy was used to convert each syndrome type into a binary classification problem, allowing for class-wise analysis. For each syndrome label, a 2×2 contingency table was constructed to compare TCMPCSD-BERT with each baseline model. When the number of discordant predictions (b+c) was ≥25, the chi-square approximation without continuity correction was used [33]; otherwise, the exact binomial test was applied for small sample sizes (b+c<25) [34]. Further statistical details are provided in Multimedia Appendix 4.

To more comprehensively assess the stability and statistical uncertainty of each model's classification performance, we adopted a stratified bootstrap resampling strategy based on prior studies [35,36]. Specifically, for each model, we conducted 5000 iterations of resampling with replacement from the test set predictions. For each bootstrap sample, we computed 4 evaluation metrics: macroprecision, macrorecall, macro–F_1-score, and accuracy. Based on the resulting 5000 estimates per metric, we used the percentile bootstrap method to construct 95% CIs (bootstrap 95% CIs) and calculated the corresponding bootstrap means, thereby quantifying the variability and uncertainty in model performance. Further statistical details are provided in Multimedia Appendix 4.

Furthermore, given that all models were evaluated on the same test set and their prediction results were mutually independent, we conducted independent samples *t* tests to examine whether the observed differences between TCMPCSD-BERT and other models were statistically significant across the evaluation metrics [37]. When the bootstrap estimates approximately satisfied the assumptions of normality and homogeneity of variance, the standard independent samples *t* test was applied. In cases where unequal variances were observed, Welch *t* test was used instead. For metrics whose distributions substantially deviated from normality and could not be normalized through transformation, the nonparametric Mann-Whitney *U* test was used as an alternative. Further statistical details are provided in Multimedia Appendix 4.

Visualization of the Self-Attention Mechanism in TCMPCSD-BERT for Syndrome Differentiation

In order to gain a deeper insight into how the TCMPCSD-BERT model captures word associations in the task of syndrome differentiation, this study adopts a visualization approach based on the self-attention mechanism, as proposed in previous research [38]. The model's performance was examined by processing clinical texts associated with 3 distinct syndrome types: DH, "spleen-deficiency syndrome," and "damp-heat with spleen-deficiency syndrome (DH-SD)." The visualization specifically focuses on the attention distribution within the model's 0th layer, across all 12 layers, to provide a detailed understanding of how the model attends to different vocabulary elements in these syndrome differentiation tasks.

Explanation of Syndrome Differentiation Decisions

To further quantify the model's reliance on specific symptoms and signs in the syndrome differentiation task, we applied the integrated gradients algorithm [39] to assess each word's contribution to syndrome prediction in the clinical record text. This method generates an attribution score for each word in the case, reflecting its importance in predicting the syndrome type. Words that are more critical to the syndrome differentiation will receive higher attribution scores, while symptoms with lower or negligible impact on the prediction will receive lower attribution scores. A more detailed explanation, setup, and interpretation of the integrated gradients algorithm can be found in Multimedia Appendix 5.

LLMs Prompt Engineering

In terms of prompt design, this study drew inspiration from the MedPrompt strategy proposed by the Microsoft team in 2023 [40] and recent similar studies [41,42], integrating the characteristics of TCM syndrome differentiation tasks to develop an initial prompt scheme. The key components include "role-play prompting," "in-context learning," "chain-of-thought," and "choice shuffling ensemble." To address issues related to context forgetting and model hallucination when handling long text tasks, this study reconstructs the dialogue interface with every 15 questions, reloads the task context and rules, and repeats the processes of in-context learning, chain-of-thought, and choice shuffling ensemble until all cases have completed syndrome differentiation [43].

To evaluate the performance of LLMs in the syndrome differentiation task, this study adopted the same stratified test set as used for TCMPCSD-BERT and other specialized models, ensuring that the class distribution in the test set was consistent with the original dataset. The test set split followed the procedure described in the "Experiment Setup" section of the Methods. Detailed settings can be found in Multimedia Appendices 6 and 7.

To assess whether the performance differences between TCMPCSD-BERT and the LLMs were statistically significant, we adopted the same evaluation framework as described for the baseline model comparisons. McNemar test was applied using a "one-vs-rest" strategy to conduct pairwise comparisons for each syndrome category. Depending on the number of discordant cases, either the chi-square approximation or the exact binomial test was used, as appropriate.

Additionally, stratified bootstrap resampling (5000 iterations) was used to quantify the uncertainty of model performance. For each resampled dataset, macroprecision, macrorecall, macro–F_1-score, and accuracy were computed. The percentile method was used to construct bootstrap 95% CIs, and bootstrap means were calculated to summarize the metric distributions.

XSL•FO
RenderX

Further statistical details are provided in Multimedia Appendix 4.

Subsequently, independent samples *t* tests were performed on the 5000 bootstrap estimates to examine whether the observed differences between TCMPCSD-BERT and the LLMs were statistically significant. Welch *t* test or the Mann-Whitney *U* test was used when assumptions of normality or homogeneity of variance were not met. Further statistical details are provided in Multimedia Appendix 4.

Comparison of Unstructured Long-Text Processing and Syndrome Differentiation Specificity for Pancreatic Cancer Between TCMPCSD-BERT and Web-Based Platforms

We selected a recently published online syndrome differentiation platform [20] and input structured syndrome differentiation elements from clinical cases corresponding to DH, "spleen-deficiency syndrome," and DH-SD into the platform to observe whether it could accurately differentiate pancreatic cancer. Subsequently, we input the unstructured text of each clinical case into the platform to evaluate its performance in syndrome differentiation. This process allowed us to assess the specificity and accuracy of syndrome differentiation for pancreatic cancer by the web-based platform in comparison to TCMPCSD-BERT.

Ethical Considerations

This study was approved by the Ethics Committee of the Department of Oncology, Fudan University (approval number 2012228-17 and 2509-Exp282). All procedures were conducted in accordance with the ethical standards of the institutional research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. Given the retrospective design of this study and the complete anonymization of the data used, the requirement for informed consent was waived according to institutional policy. Furthermore, as no direct patient participation or interventions were involved, the requirement for participant compensation was also waived.

Results

Dataset and Model Training Details

This study used a total of 6830 clinical records from 1980 patients with pancreatic cancer for model training and evaluation. The dataset included 1694 records labeled as DH, 1185 records as spleen-deficiency syndrome, and 1178 records as DH-SD. An additional 2773 records were categorized as "others," which encompassed cases without vivid syndrome differentiation and syndromes with fewer than 500 records, thus not fulfilling the inclusion criteria. The mean age of the patients was 61 (SD 8.4) years in the DH group, 62 (SD 7.7) years in the spleen-deficiency syndrome group, 62 (SD 8.6) years in the DH-SD group, and 61 (SD 8.5) years in the others group. Among all patients, 39.4% (2688/6830) were female, 80.9% (5526/6830) had stage III–IV disease, and 46.7% (3189/6830) received chemotherapy. Detailed baseline characteristics are presented in Table 1. Based on the interrater agreement analysis using Cohen κ coefficient, the value for the syndrome differentiation task in this study was 0.913, indicating a high level of agreement between the 2 experts. The detailed contingency table for the Cohen κ analysis of the experts' annotations is provided in Table S4 of Multimedia Appendix 8.

Throughout the model training process, extensive hyperparameter tuning was performed for each model, with performance evaluated using the validation dataset. This process allowed for the identification of optimal hyperparameter configurations for each model, achieving the highest accuracy and F_1-score for the syndrome differentiation task. The optimal training parameters for each model are detailed in Table S2 of Multimedia Appendix 8.

XSL•FO
RenderX

Table 1. Descriptive statistics of the cancer cohort.

Descriptive statistics	Damp-heat syndrome (1694/6830, 24.8%)	Spleen-deficiency syndrome (1185/6830, 17.3%)	Damp-heat with spleen-deficiency syndrome (1178/6830, 17.2%)	Others (2773/6830, 40.6%)
Demographics				
Sex (female), n/N (%)	678/1694 (40)	484/1185 (40.8)	447/1178 (37.9)	1079/2773 (38.9)
Age (years), mean (SD)	61 (8.4)	62 (7.7)	62 (8.6)	61 (8.5)
Cancer stage, n/N (%)				
Stage I	110/1694 (6.5)	55/1185 (5)	87/1178 (7)	263/2773 (9.5)
Stage II	200/1694 (11.8)	122/1185 (10.3)	95/1178 (8)	372/2773 (13.4)
Stage III	776/1694 (45.8)	560/1185 (47.3)	531/1178 (45)	918/2773 (33.1)
Stage IV	608/1694 (35.9)	448/1185 (37.8)	465/1178 (39.5)	1220/2773 (44)
Chemotherapy, n/N (%)				
Yes	791/1694 (46.7)	511/1185 (43.1)	487/1178 (41.3)	1400/2773 (50.5)
No	903/1694 (53.3)	674/1185 (56.9)	691/1178 (58.7)	1373/2773 (49.5)
Diabetes, n/N (%)				
Yes	251/1694 (14.8)	145/1185 (12.2)	129/1178 (11)	385/2773 (13.9)
No	1443/1694 (85.2)	1040/1185 (87.8)	1049/1178 (89)	2388/2773 (86.1)
Hypertension, n/N (%)				
Yes	249/1694 (14.7)	182/1185 (15.4)	270/1178 (22.9)	613/2773 (22.1)
No	1445/1694 (85.3)	1003/1185 (84.6)	908/1178 (77.1)	2160/2773 (77.9)
Smoking status, n/N (%)				
Never	302/1694 (17.8)	198/1185 (16.7)	226/1178 (19.2)	348/2773 (12.6)
Former	1279/1694 (75.5)	916/1185 (77.3)	843/1178 (71.6)	2246/2773 (81)
Current	113/1694 (6.7)	71/1185 (6)	109/1178 (9.2)	179/2773 (6.4)
Alcohol consumption, n/N (%)				
Never	120/1694 (7.1)	137/1185 (11.6)	144/1178 (12.2)	187/2773 (6.7)
Occasionally	115/1694 (6.8)	56/1185 (5)	80/1178 (7)	68/2773 (3)
Regularly	1459/1694 (86.1)	992/1185 (83.7)	954/1178 (81)	2518/2773 (90.8)

Performance Comparison of TCMPCSD-BERT and Baseline Models

As illustrated in Figure 2, TCMPCSD-BERT demonstrated superior syndrome differentiation performance in both training and validation phases. With increasing epochs, the accuracy of the training and validation datasets improved steadily, with validation accuracy closely approaching training accuracy. This pattern indicates the model's strong generalization capacity and the absence of overfitting (Figure 2A). Simultaneously, the training and validation losses decreased in parallel, reflecting good convergence and the model's effectiveness in capturing critical features for the TCM syndrome differentiation task (Figure 2B). The F_1-score, recall, and precision metrics on the validation dataset progressively improved throughout training, ultimately reaching approximately 0.97, underscoring the model's high accuracy and reliability in syndrome differentiation (Figure 2C).

XSL•FO
RenderX

Figure 2. Training and validation performance curves of TCMPCSD-BERT, LSTM, and text-CNN models. (A) Training and validation accuracy curves for TCMPCSD-BERT. (B) Training and validation loss curves for TCMPCSD-BERT. (C) Validation F1-score, recall, and precision curves for TCMPCSD-BERT. (D) Training and validation accuracy curves for LSTM. (E) Training and validation loss curves for LSTM. (F) Validation F1-score, recall, and precision curves for LSTM. (G) Training and validation accuracy curves for text-CNN. (H) Training and validation loss curves for text-CNN. (I) Validation F1-score, recall, and precision curves for text-CNN. LSTM: long short-term memory; TCMPCSD-BERT: traditional Chinese medicine pancreatic cancer syndrome differentiation bidirectional encoder representations from transformers; text-CNN: text convolutional neural network.

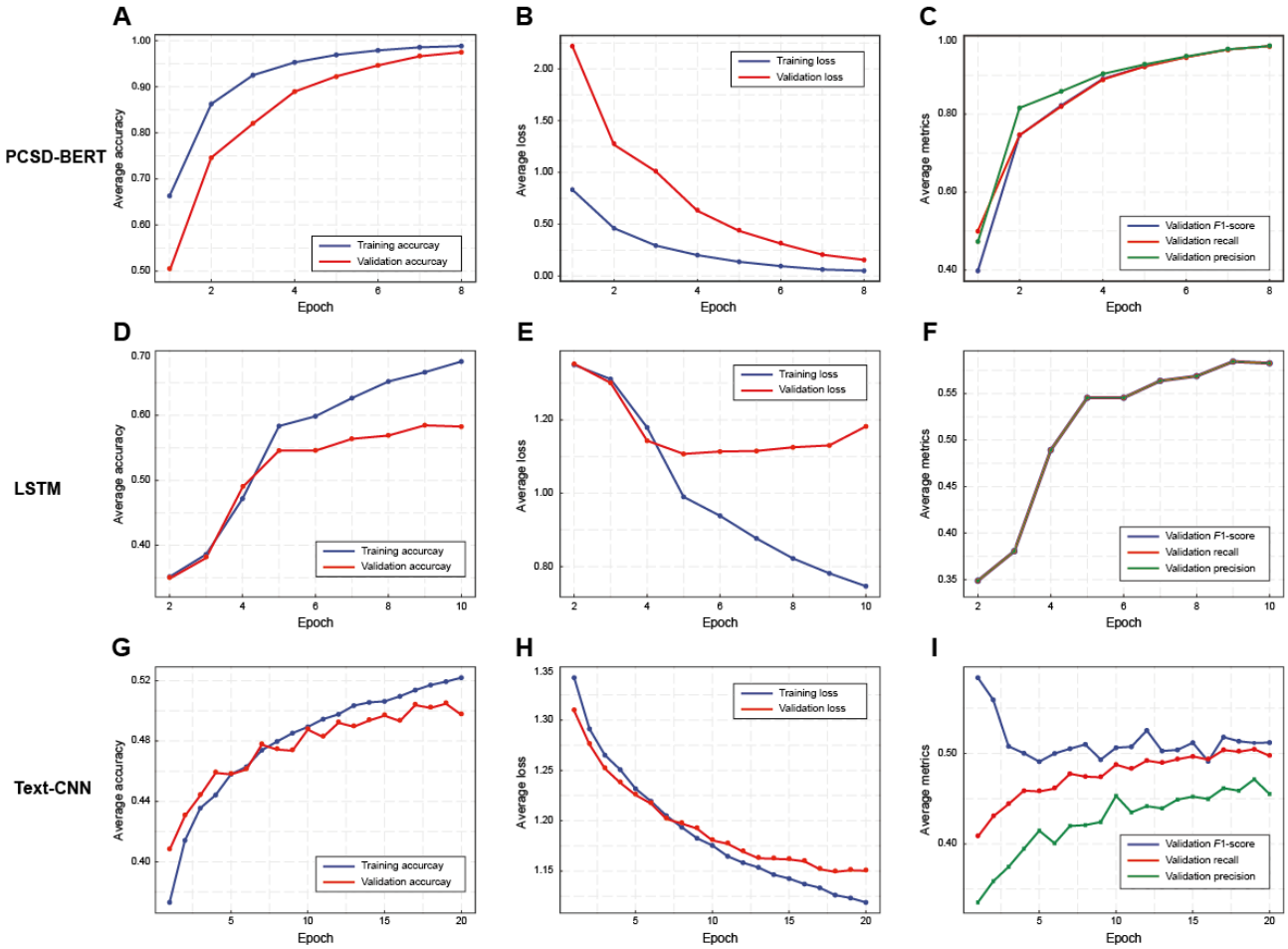

In contrast, the LSTM model displayed clear signs of overfitting. As the number of epochs increased, training accuracy approached 0.7, while validation accuracy rose slowly and consistently remained lower than training accuracy, indicating limited generalization on the validation dataset (Figure 2D). Training loss continued to decrease, whereas validation loss plateaued or slightly increased, further confirming overfitting (Figure 2E). The F_1-score, recall, and precision of the LSTM model stabilized at around 0.55 after an initial increase, revealing limited classification capability (Figure 2F).

For the text-CNN model, training accuracy increased with additional epochs but plateaued around 0.52 after approximately 20 epochs. Validation accuracy exhibited limited improvement and fluctuated considerably, peaking near 0.48 but consistently remaining below training accuracy, suggesting a tendency toward overfitting (Figure 2G). Although training loss declined significantly, validation loss leveled off after an initial drop, remaining slightly higher than training loss and indicating insufficient generalization on the validation dataset (Figure 2H). The F_1-score, recall, and precision on the validation dataset were both low and unstable, highlighting text-CNN's limited classification capability in syndrome differentiation. The substantial fluctuations in these metrics suggest inconsistent performance across different validation folds, hindering text-CNN's ability to establish stable differentiation rules (Figure 2I).

As shown in Table 2, McNemar test suggests that TCMPCSD-BERT demonstrated significantly better classification performance than the 3 baseline models in distinguishing all 4 TCM syndrome types. As shown in Table 3, when compared with the traditional baseline models, TCMPCSD-BERT achieved higher mean values across all 4 evaluation metrics. The 95% bootstrap CIs for TCMPCSD-BERT did not overlap with those of the comparator models. The results from Welch t test indicated that these differences were statistically significant in all comparisons (P<.01).

XSL•FO
RenderX

Table 2. McNemar test comparing TCMPCSD-BERT[a] and baseline models across syndrome categories.

Variable	TCMPCSD-BERT[a] correct only/total	Baseline models correct only/total	Chi-square (*df*)	*P* value
TCMPCSD-BERT versus BERT[b] without fine-tuning				
Damp-heat syndrome	67/169	9/169	42.75 (1)	<.001[c]
Spleen-deficiency syndrome	49/119	8/119	28.07 (1)	<.001[c]
Damp-heat with spleen-deficiency syndrome	48/118	7/118	29.09 (1)	<.001[c]
Others	107/277	10/277	78.77 (1)	<.001[c]
TCMPCSD-BERT versus LSTM[d]				
Damp-heat syndrome	76/169	4/169	63.01 (1)	<.001[c]
Spleen-deficiency syndrome	56/119	6/119	38.73 (1)	<.001[c]
Damp-heat with spleen-deficiency syndrome	58/118	8/118	36.38 (1)	<.001[c]
Others	126/277	7/277	104.69 (1)	<.001[c]
TCMPCSD-BERT versus text-CNN[e]				
Damp-heat syndrome	80/169	4/169	66.96 (1)	<.001[c]
Spleen-deficiency syndrome	57/119	4/119	44.33 (1)	<.001[c]
Damp-heat with spleen-deficiency syndrome	56/118	3/118	45.83 (1)	<.001[c]
Others	136/277	11/277	104.60 (1)	<.001[c]

[a]TCMPCSD-BERT: traditional Chinese medicine pancreatic cancer syndrome differentiation bidirectional encoder representations from transformers.
[b]BERT: bidirectional encoder representations from transformers.
[c]The *P* values were calculated using the McNemar test without continuity correction.
[d]LSTM: long short-term memory.
[e]Text-CNN: text convolutional neural network.

XSL•FO
RenderX

Table 3. Performance metrics with bootstrap 95% CIs and statistical comparisons between TCMPCSD-BERT[a] and baseline models.

Models and metrics	Mean (SD)	95% CIs[b]	*P* value[c]
TCMPCSD-BERT			
Macroprecision	0.935 (0.008)	0.918-0.951	—[d]
Macrorecall	0.921 (0.011)	0.900-0.942	—
Macro–F_1-score	0.927 (0.009)	0.908-0.945	—
Accuracy	0.919 (0.010)	0.899-0.939	—
BERT[e] without fine-tuning			
Macroprecision	0.531 (0.019)	0.495-0.567	<.01[f]
Macrorecall	0.540 (0.021)	0.500-0.580	<.01[f]
Macro–F_1-score	0.527 (0.020)	0.489-0.565	<.01[f]
Accuracy	0.548 (0.019)	0.510-0.586	<.01[f]
LSTM[g]			
Macroprecision	0.481 (0.019)	0.443-0.519	<.01[f]
Macrorecall	0.492 (0.020)	0.451-0.531	<.01[f]
Macro–F_1-score	0.484 (0.020)	0.446-0.522	<.01[f]
Accuracy	0.491 (0.019)	0.454-0.529	<.01[f]
Text-CNN[h]			
Macroprecision	0.483 (0.020)	0.444-0.522	<.01[f]
Macrorecall	0.483 (0.020)	0.443-0.523	<.01[f]
Macro–F_1-score	0.477 (0.019)	0.437-0.515	<.01[f]
Accuracy	0.475 (0.019)	0.436-0.511	<.01[f]

[a]TCMPCSD-BERT: traditional Chinese medicine pancreatic cancer syndrome differentiation bidirectional encoder representations from transformers.
[b]Bootstrap 95% CIs.
[c]Compared with traditional Chinese medicine pancreatic cancer syndrome differentiation bidirectional encoder representations from transformers model.
[d]Not applicable.
[e]BERT: bidirectional encoder representations from transformers.
[f]*P* values are obtained from the Welch *t* test.
[g]LSTM: long short-term memory.
[h]Text-CNN: text convolutional neural network.

Validation of TCMPCSD-BERT Model's Syndrome Differentiation Performance in Practical Applications

Upon completion of the training and performance assessment of the TCMPCSD-BERT model, its practical effectiveness in syndrome differentiation was further examined using real-world clinical cases. Cases were selected based on clinical expert diagnoses, representing 4 categories: DH, "spleen-deficiency syndrome," DH-SD, and cases without clear syndrome differentiation (classified as "others"). These cases were input into the TCMPCSD-BERT model to evaluate its classification accuracy and interpretability. Figure 3 illustrates the practical implementation of TCMPCSD-BERT within the Jupyter Notebook environment. The "sample text" section below the code block displays the input clinical case text, and the translated clinical case records on the right provide an English version of the input clinical case text on the left for reference. The "probabilities for each label" section shows the confidence scores generated by TCMPCSD-BERT for each syndrome label. Diagnostic features within the clinical text were manually annotated under expert guidance, with primary symptoms highlighted in dark orange and secondary symptoms highlighted in light orange to enhance interpretability. For each input case, the model generated confidence scores for all syndrome labels ("probabilities for each label"), with the highest confidence score indicating the final predicted classification. To provide interpretability, diagnostic features within the clinical text were annotated under expert guidance, with primary symptoms highlighted in dark orange and secondary symptoms in light orange.

XSL•FO
RenderX

Figure 3. Syndrome differentiation of clinical texts by TCMPCSD-BERT. (A) A case diagnosed as “damp-heat syndrome” by clinical experts. (B) A case diagnosed as “spleen-deficiency syndrome.” (C) A case diagnosed as “damp-heat with spleen-deficiency syndrome.” (D) A case without clear syndrome differentiation features, classified as “others.” TCMPCSD-BERT: traditional Chinese medicine pancreatic cancer syndrome differentiation bidirectional encoder representations from transformers.

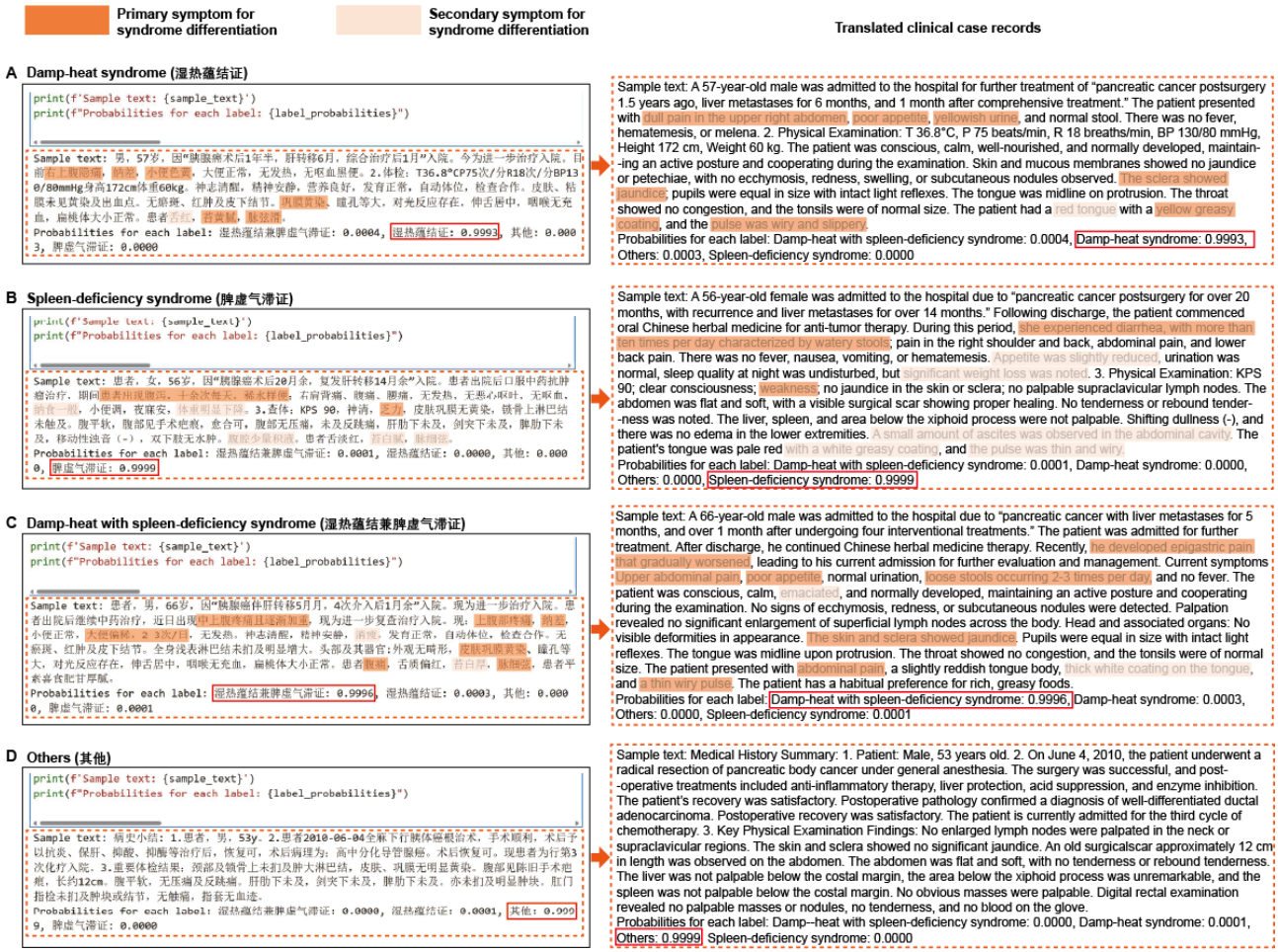

For instance, in a case diagnosed as DH, the clinical text included features such as “右上腹隐痛” (abdominal pain), “纳差” (loss of appetite), and “舌红” (red tongue), corresponding to the primary and secondary symptoms of DH. The model predicted this case as DH with a confidence score of 0.9993, consistent with the clinical expert’s diagnosis (Figure 3A). In a case diagnosed as “spleen-deficiency syndrome,” features such as “腹泻” (loose stools), “纳食一般” (small appetite), and “乏力” (fatigue) were identified. The model classified this case as “spleen-deficiency syndrome” with a confidence score of 0.9999 (Figure 3B). For a case exhibiting characteristics of both DH and “spleen-deficiency syndrome,” diagnostic features such as “腹痛” (abdominal pain) and “纳差” (loss of appetite or small appetite) were highlighted. The model predicted this case as DH-SD with a confidence score of 0.9996 (Figure 3C). Lastly, for a case without significant diagnostic features, the model correctly excluded other syndrome labels and classified it as “others” with a confidence score of 0.9999 (Figure 3D).

Table S6 in Multimedia Appendix 9 provides a detailed summary of the diagnostic features extracted from each clinical case and their correspondence with the TCM syndrome differentiation guidelines, along with their classification as primary or secondary symptoms. The table includes examples for DH and “spleen-deficiency syndrome,” listing the diagnostic features in the clinical text, their corresponding entries in the guidelines, and their classification within the differentiation framework. In all evaluated cases, the predictions generated by TCMPCSD-BERT were consistent with expert diagnoses, validating its effectiveness in practical TCM syndrome differentiation tasks.

Visualization of Self-Attention Mechanism in TCMPCSD-BERT

As illustrated in Figure 4, the TCMPCSD-BERT model captures associations between vocabulary elements in clinical case texts for different syndrome types in syndrome differentiation tasks. The lines between words represent semantic associations, with the thickness of the lines reflecting the strength of these associations. Thicker lines indicate higher attention weights assigned by the model, highlighting the interdependence and significance of specific terms in syndrome differentiation. The translated clinical case records on the left provide an English version of the input clinical case text on the right for reference. Diagnostic features within the clinical text were manually annotated under expert guidance, with primary symptoms highlighted in dark orange and secondary symptoms highlighted in light orange to enhance interpretability. The [SEP] token shown in the figure is a special marker in the BERT model, used to distinguish sentences or paragraphs during text

XSL•FO
RenderX

processing. This token ensures that the model can correctly identify textual boundaries and accurately interpret contextual relationships.

Figure 4. Visualization of attention distributions in TCMPCSD-BERT across syndrome texts. This figure illustrates the attention distributions in layer 0 of the TCMPCSD-BERT model for three syndrome types: (A) damp-heat syndrome, (B) spleen-deficiency syndrome, and (C) damp-heat with spleen-deficiency syndrome. CLS: classification token; SEP: separator token; TCMPCSD-BERT: traditional Chinese medicine pancreatic cancer syndrome differentiation bidirectional encoder representations from transformers.

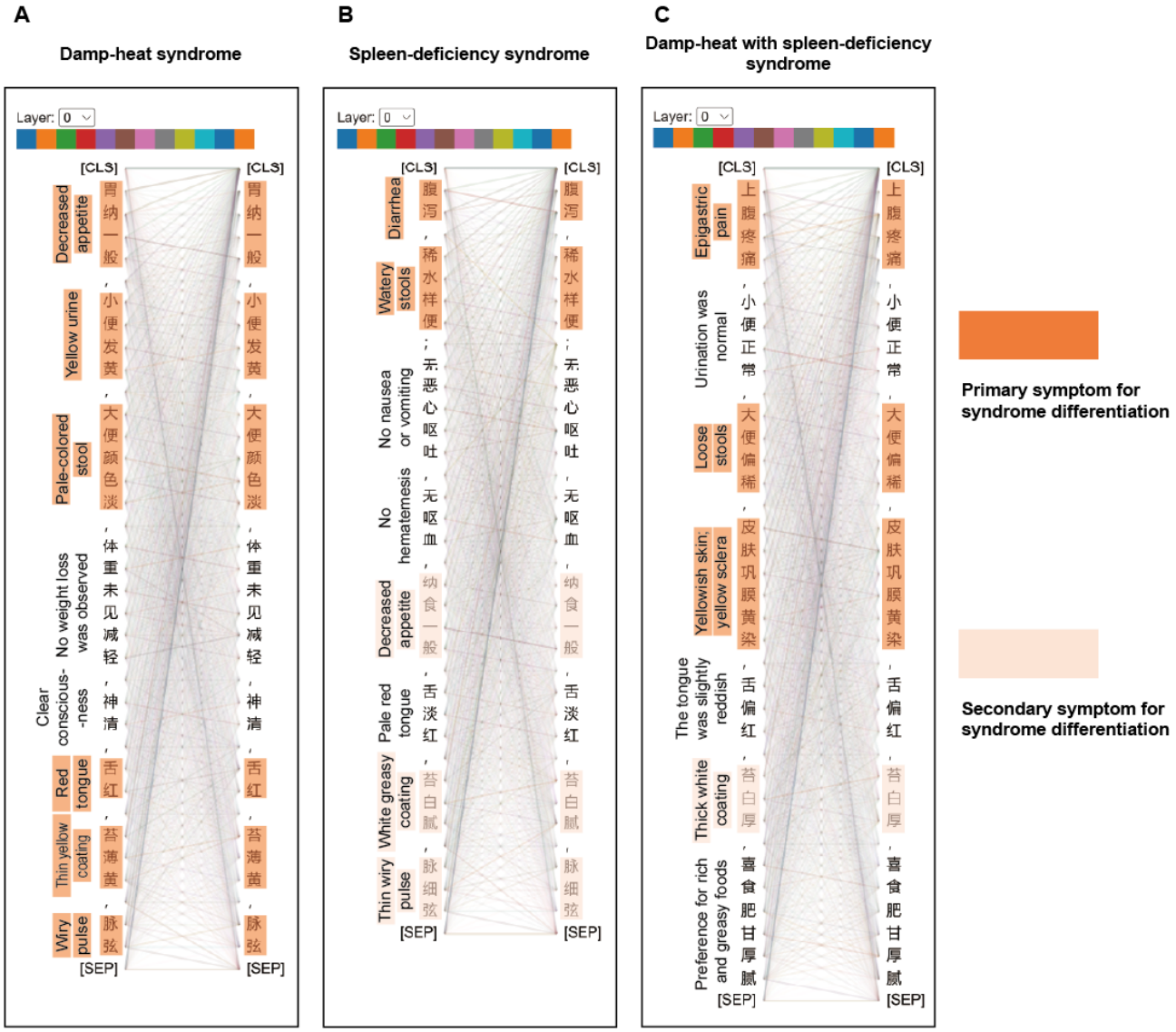

As shown in Figure 4A, in the text for DH, the model strengthens associations among terms such as “胃纳一般” (loss of appetite), “小便发黄” (yellow urine), “大便颜色淡” (gray-white stool), “舌红” (red tongue), “苔薄黄” (yellow coating), and “脉弦” (wiry and rapid pulse). In Figure 4B, for “spleen-deficiency syndrome,” the model highlights associations among terms such as “腹泻” (loose stools), “纳食一般” (small appetite), “舌淡红” (pale and swollen tongue), “苔白腻” (white slippery coating), and “脉细弦” (thin wiry pulse). In Figure 4C, for DH-SD, the model reinforces associations between terms such as “上腹疼痛” (abdominal pain), “大便偏稀” (loose stools), “皮肤巩膜黄染” (yellowish skin and yellow sclera), and “苔白厚” (white thick coating). This attention distribution demonstrates the model’s ability to effectively identify structural relationships between TCM-specific terms, thereby constructing a precise framework for syndrome differentiation.

Beyond associations between specific terms, we also observe that all words within the same passage exhibit a degree of interrelation. This suggests that the model considers not only the relationships among key terms but also integrates contextual links across all words, contributing to a comprehensive understanding of the overall context.

Attribution Analysis of Word-Syndrome Associations in TCMPCSD-BERT

To evaluate the model’s ability to identify the associations between words and syndrome labels in syndrome differentiation tasks, integrated gradients were applied to clinical texts representing 3 typical syndromes: DH, “spleen-deficiency syndrome,” and DH-SD. This analysis aimed to reveal how the model assesses the importance of key diagnostic terms when interpreting clinical texts. As shown in Figure 5, the first row (translation) provides the English translation of the second row

XSL•FO
RenderX

(token), which contains the original clinical case text in Chinese. The color of each token in the second row indicates its importance in real-world TCM syndrome differentiation: dark orange denotes primary diagnostic features, while light orange represents secondary features. The third row (attribution score) shows the attribution scores assigned to each token, with the values color-coded by magnitude; higher scores are represented in darker orange, indicating greater contributions to syndrome differentiation. By comparing token colors with their corresponding attribution scores, the alignment between the model's judgments and clinical standards can be intuitively analyzed.

Figure 5. Gradient attribution analysis of TCMPCSD-BERT on syndrome texts. (A) A case diagnosed as "damp-heat syndrome" by clinical experts. (B) A case diagnosed as "spleen-deficiency syndrome." (C) A case diagnosed as "damp-heat with spleen-deficiency syndrome." (D) Content from a case diagnosed as "damp-heat syndrome" that does not include any diagnostic features related to syndrome differentiation. TCMPCSD-BERT: traditional Chinese medicine pancreatic cancer syndrome differentiation bidirectional encoder representations from transformers.

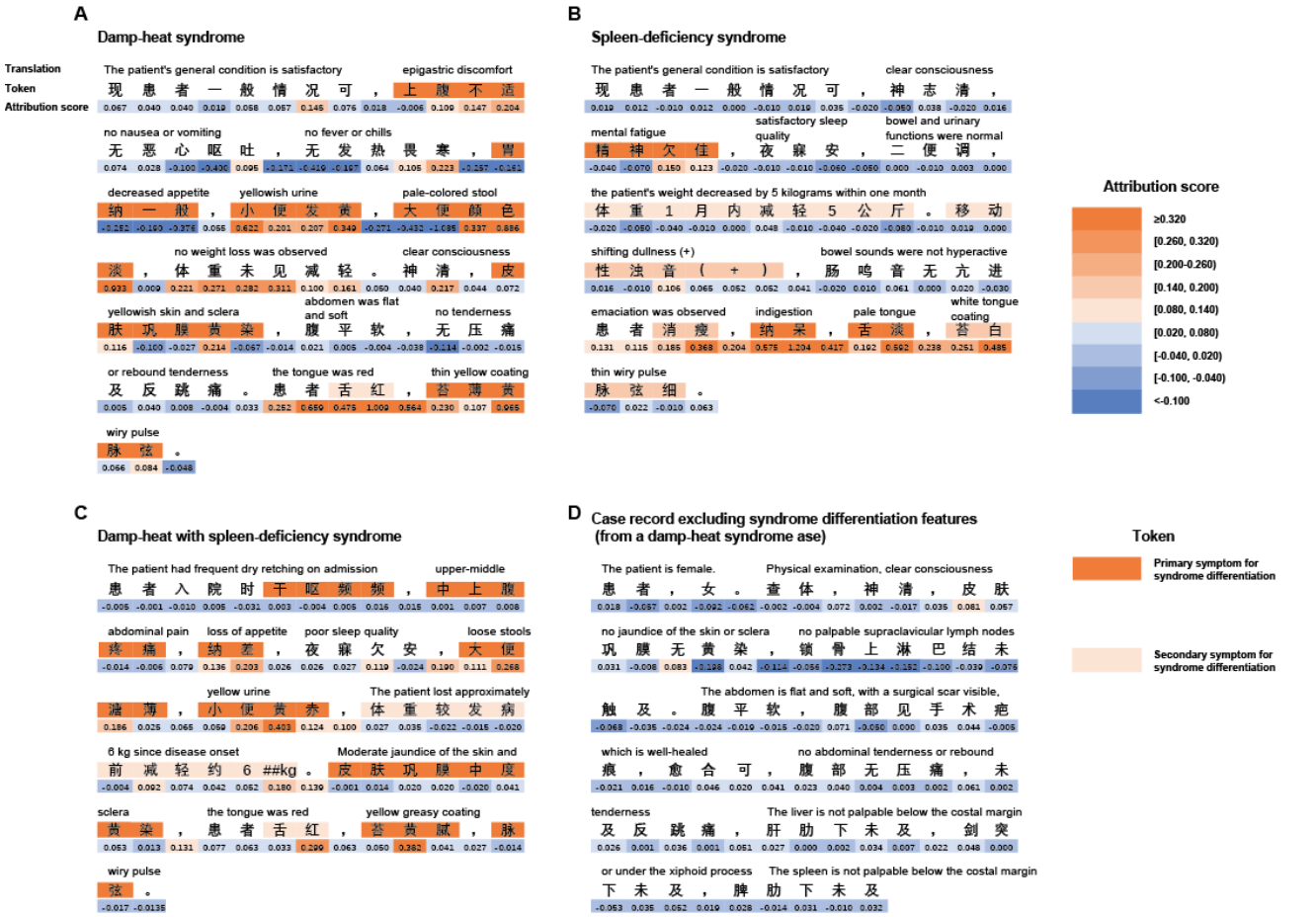

The TCMPCSD-BERT model assigns higher attribution scores to certain key diagnostic features. For instance, in DH, the highest attribution score for "大便颜色淡" (gray-white stool) was 0.933 (Figure 5A); in "spleen-deficiency syndrome," the highest score for "纳呆" (indigestion) reached 1.204 (Figure 5B); and in DH-SD, the highest score for "小便黄赤" (yellow urine) was 0.403 (Figure 5C). Additionally, Figure 5D shows the attribution scores for content in the DH case that does not include diagnostic elements related to syndrome differentiation, where these terms generally exhibit lower attribution scores.

Table S7 in Multimedia Appendix 10 provides a detailed summary of the key diagnostic features extracted from each clinical text, their correspondence with the TCM syndrome differentiation guidelines, and their classification within the syndrome differentiation framework, along with attribution scores. The table highlights the highest attribution scores assigned to different diagnostic features and compares them with the average attribution scores of nondiagnostic contents. It is observed that the attribution scores for most diagnostic features are higher than those of nondiagnostic content. This distinction further validates the model's ability to differentiate critical diagnostic features from irrelevant content. However, the attribution scores for some features deviate from clinical standards, reflecting the model's limitations in assigning appropriate importance weights to certain features.

Case Studies of Specific Error Patterns in TCMPCSD-BERT

The TCMPCSD-BERT model exhibited misclassification in certain cases from the test set. To gain a deeper understanding of the model's error patterns, we selected misclassified cases from 3 typical syndromes and presented the attribution analysis results for these cases in Figure S1 (Multimedia Appendix 11).

Figure S1A in Multimedia Appendix 11 presents the attribution analysis result where the TCMPCSD-BERT model misclassified a DH case as "spleen-deficiency syndrome." It can be observed that the model assigned lower attribution scores to key diagnostic features of DH, such as "上腹疼痛" (abdominal

XSL•FO
RenderX

pain), “低热” (persistent low-grade fever), “脉弦数” (wiry and rapid pulse); while assigning higher scores to the diagnostic feature of “近2个月体重减轻2.5kg左右” (emaciation) which is associated with “spleen-deficiency syndrome.” Despite the case exhibiting multiple secondary symptoms of “spleen-deficiency syndrome” (3 secondary symptoms), the absence of a primary symptom prevented it from meeting the diagnostic criteria for “spleen-deficiency syndrome.”

Figure S1B in Multimedia Appendix 11 displays the attribution analysis results where the TCMPCSD-BERT model misclassified a “spleen-deficiency syndrome” case as DH-SD. The model assigned lower attribution scores to key diagnostic features of “spleen-deficiency syndrome,” such as “移动性浊音（+）” (ascites), “双下肢无凹陷性水肿” (edema in lower limbs), and “苔薄腻” (thin and greasy coating); while assigning higher scores to features associated with DH, such as “大便黏腻不爽、味重” (foul-smelling loose stools), and “脉弦数” (wiry and rapid pulse). Despite the case presenting 1 primary symptom and 1 secondary symptom related to DH, it failed to meet the diagnostic criteria for DH.

Figure S1C in Multimedia Appendix 11 shows the attribution analysis results where the TCMPCSD-BERT model misclassified a DH-SD case as “spleen-deficiency syndrome.” It can be observed that the model assigned lower attribution scores to the diagnostic features of “低热” (persistent low-grade fever) and “感上腹疼痛” (abdominal pain), which are key features of DH.

Table S8 in Multimedia Appendix 11 provides a detailed summary of the key diagnostic features extracted from each misclassified case’s clinical text, along with their correspondence to the TCM syndrome differentiation guidelines and their classification within the syndrome differentiation framework, as well as the attribution scores. Furthermore, a comparison is made between the attribution scores of these diagnostic features and the average attribution scores of nondiagnostic content, and the table indicates whether the model correctly assigned higher attribution scores to the diagnostic features.

Performance Comparison Between TCMPCSD-BERT and LLMs Guided by Prompt Engineering Framework

This study further evaluated the efficacy of LLMs in differentiating syndromes in pancreatic cancer cases based on TCM principles, using a prompt engineering framework. Detailed descriptions of the prompt engineering frameworks and corresponding test cases for English- and Chinese-based LLMs are provided in the second and third sections of Multimedia Appendix 6, respectively.

As shown in Table 4, McNemar test indicates that TCMPCSD-BERT generally exhibited statistically superior classification performance compared to the evaluated LLMs across most TCM syndrome types. Although some comparisons did not reach conventional significance thresholds (eg, P=.07), the overall results suggest that TCMPCSD-BERT may offer more consistent differentiation performance in this task. As shown in Table 5, when compared with the LLMs, TCMPCSD-BERT achieved higher mean values across all 4 evaluation metrics, with no overlap between its bootstrap 95% CIs and those of the comparator models. Results from Welch t test indicated that these differences were statistically significant in all comparisons (P<.01).

XSL•FO
RenderX

Table 4. McNemar test comparing TCMPCSD-BERT[a] and LLMs[b] across syndrome categories.

Variable	TCMPCSD-BERT correct only/total	LLMs correct only/total	Chi-square (*df*)	*P* values
TCMPCSD-BERT versus ChatGPT-4				
Damp-heat syndrome	38/169	11/169	13.80 (1)	<.001[c]
Spleen-deficiency syndrome	27/119	8/119	10.31 (1)	<.001[c]
Damp-heat with spleen-deficiency syndrome	24/118	5/118	12.45 (1)	<.001[c]
Others	66/277	20/277	24.60 (1)	<.001[c]
TCMPCSD-BERT versus ChatGPT-4o				
Damp-heat syndrome	27/169	14/169	4.12 (1)	<.05[c]
Spleen-deficiency syndrome	17/119	8/119	3.24 (1)	.07[c]
Damp-heat with spleen-deficiency syndrome	18/118	9/118	3.00 (1)	.08[c]
Others	40/277	17/277	9.28 (1)	.002[c]
TCMPCSD-BERT versus Ernie Bot 4.0 Turbo				
Damp-heat syndrome	35/169	9/169	15.36 (1)	<.001[c]
Spleen-deficiency syndrome	25/119	7/119	10.13 (1)	<.001[c]
Damp-heat with spleen-deficiency syndrome	25/118	7/118	10.13 (1)	<.001[c]
Others	58/277	14/277	26.89 (1)	<.001[c]
TCMPCSD-BERT versus Kimi				
Damp-heat syndrome	41/169	11/169	17.31 (1)	<.001[c]
Spleen-deficiency syndrome	27/119	4/119	13.36 (1)	<.001[c]
Damp-heat with spleen-deficiency syndrome	27/118	6/118	13.36 (1)	<.001[c]
Others	66/277	15/277	32.11 (1)	<.001[c]
TCMPCSD-BERT versus Zhipu Qingyan				
Damp-heat syndrome	26/169	12/169	5.16 (1)	<.05[c]
Spleen-deficiency syndrome	19/119	9/119	3.57 (1)	.06[c]
Damp-heat with spleen-deficiency syndrome	16/118	6/118	—[d]	.05[e]
Others	45/277	20/277	8.86 (1)	.002[c]

[a]TCMPCSD-BERT: traditional Chinese medicine pancreatic cancer syndrome differentiation bidirectional encoder representations from transformers.
[b]LLM: large language model.
[c]The *P* values were calculated using the McNemar test without continuity correction.
[d]Not applicable.
[e]The *P* values were calculated using the exact McNemar test.

XSL•FO
RenderX

Table 5. Performance metrics with bootstrap 95% CIs and statistical comparisons between TCMPCSD-BERT[a] and baseline models.

Models and metrics	Mean (SD)	95% CIs[b]	*P* value[c]
TCMPCSD-BERT			
Macroprecision	0.935 (0.008)	0.918-0.951	—[d]
Macrorecall	0.921 (0.011)	0.900-0.942	—
Macro–F_1-score	0.927 (0.009)	0.908-0.945	—
Accuracy	0.919 (0.010)	0.899-0.939	—
ChatGPT-4			
Macroprecision	0.740 (0.017)	0.706-0.774	<.001[e]
Macrorecall	0.752 (0.017)	0.717-0.787	<.001[e]
Macro–F_1-score	0.744 (0.017)	0.710-0.778	<.001[e]
Accuracy	0.748 (0.017)	0.716-0.780	<.001[e]
ChatGPT-4o			
Macroprecision	0.841 (0.015)	0.811-0.869	<.001[e]
Macrorecall	0.846 (0.015)	0.817-0.875	<.001[e]
Macro–F_1-score	0.843 (0.014)	0.813-0.871	<.001[e]
Accuracy	0.845 (0.014)	0.817-0.873	<.001[e]
Ernie Bot 4.0 Turbo			
Macroprecision	0.734 (0.018)	0.699-0.769	<.001[e]
Macrorecall	0.745 (0.018)	0.710-0.780	<.001[e]
Macro–F_1-score	0.738 (0.018)	0.704-0.773	<.001[e]
Accuracy	0.749 (0.017)	0.716-0.782	<.001[e]
Kimi			
Macroprecision	0.755 (0.017)	0.723-0.788	<.001[e]
Macrorecall	0.764 (0.017)	0.732-0.797	<.001[e]
Macro–F_1-score	0.758 (0.016)	0.725-0.790	<.001[e]
Accuracy	0.761 (0.016)	0.729-0.794	<.001[e]
Zhipu Qingyan			
Macroprecision	0.838 (0.015)	0.808-0.865	<.001[e]
Macrorecall	0.846 (0.014)	0.816-0.873	<.001[e]
Macro–F_1-score	0.841 (0.014)	0.812-0.868	<.001[e]
Accuracy	0.846 (0.014)	0.818-0.873	<.001[e]

[a]TCMPCSD-BERT: traditional Chinese medicine pancreatic cancer syndrome differentiation bidirectional encoder representations from transformers.
[b]Bootstrap 95% CIs.
[c]Compared with traditional Chinese medicine pancreatic cancer syndrome differentiation bidirectional encoder representations from transformers model.
[d]Not applicable.
[e]*P* values are obtained from the Welch *t* test.

Comparison of Unstructured Long-Text Processing and Syndrome Differentiation Specificity for Pancreatic Cancer Between TCMPCSD-BERT and Web-Based Platforms

This study presents a preliminary comparison between TCMPCSD-BERT and online syndrome differentiation

XSL•FO
RenderX

platforms in terms of unstructured long-text processing performance and the specificity of pancreatic cancer syndrome differentiation. As shown in Figure S2A (Multimedia Appendix 12), when we input a medical case diagnosed by a clinical expert as DH into the TCMPCSD-BERT model, it accurately predicted the case as DH with a confidence of 0.9994, closely matching the expert's diagnosis. In contrast, as shown in Figure S2B (Multimedia Appendix 12), when entering the syndrome differentiation elements for pancreatic cancer with damp-heat accumulation, such as "腹部胀痛" (corresponding to the syndrome differentiation element in the guidelines: abdominal pain), "小便色黄" (yellow urine), "皮肤黄染" (yellowish skin), "舌红" (red tongue), "脉滑数" (rapid pulse) from the case in Figure S2A in Multimedia Appendix 12, the platform's output is "邪热蕴蒸证" (evil-heat obstruction syndrome), rather than "湿热蕴结证" (DH). This result highlights that while the platform can perform basic syndrome differentiation, it fails to adequately recognize the specific pathological features of pancreatic cancer, resulting in a lack of specificity in its output. Furthermore, as shown in Figure S2C (Multimedia Appendix 12), when the original case's long text was input, although the platform was able to process the text and provide a syndrome differentiation result, the output was "出血证" (bleeding syndrome), and key syndrome differentiation elements such as "腹部胀痛" (corresponding to the syndrome differentiation element in the guidelines: abdominal pain), "小便色黄" (yellow urine), "皮肤黄染" (yellowish skin), "舌红" (red tongue), "脉滑数" (rapid pulse) were not included in the "symptoms" (syndrome differentiation basis) sections. This indicates that the platform has limitations in processing unstructured long-text data and fails to effectively extract and recognize important syndrome differentiation features.

Additionally, in Figures S3 and S4 (Multimedia Appendix 12), we present comparison examples of TCMPCSD-BERT and the web-based platform in processing unstructured long texts and syndrome differentiation specificity for cases of spleen-deficiency syndrome and DH-SD. TCMPCSD-BERT also demonstrated stronger long-text processing performance and more accurate syndrome differentiation specificity in these pancreatic cancer syndrome tasks. Finally, in Figure S5 (Multimedia Appendix 12), we present the consultation interface of a certain 4-diagnostic instrument. Although the 4-diagnostic instrument supports basic syndrome differentiation operations, some devices still require users to manually select syndrome differentiation elements. This operational approach cannot directly process unstructured long-text clinical cases, resulting in relatively lower data processing efficiency and a more cumbersome operation process.

Discussion

Principal Findings

This study trained the TCMPCSD-BERT model using 6830 TCM medical records from pancreatic cancer cases. Compared to online syndrome differentiation platforms and the 4-diagnostic instrument, TCMPCSD-BERT demonstrates superior performance and efficiency in automatically identifying syndrome-related features from unstructured medical records, particularly in processing long-text, unstructured data. In comparison with LLMs guided by the prompt engineering framework, TCMPCSD-BERT shows significant advantages in terms of stability, accuracy, and operability. Furthermore, TCMPCSD-BERT is more capable of precisely identifying clinical features relevant to the syndrome differentiation of pancreatic cancer, showcasing improved specificity and customization, thus further confirming its potential for application in automated TCM diagnosis.

Cultural and Conceptual Relevance of TCM Syndrome Patterns in Pancreatic Cancer

In TCM theory, DH and spleen-deficiency syndrome are considered 2 of the more commonly observed foundational patterns in patients with pancreatic cancer. Although these concepts have not been systematically validated within the framework of modern biomedical science, they have been applied in TCM clinical practice for many years and may offer an alternative perspective for understanding the general health status of patients with this disease.

Within the TCM theoretical system, the term "spleen" does not correspond precisely to the anatomical spleen as defined in modern medicine. Instead, it refers to a set of physiological functions related to the transformation of food, the regulation of fluids, and the generation of qi and blood [44]. It is believed that spleen function plays a vital role in maintaining internal homeostasis. Spleen-deficiency syndrome typically refers to a weakened state of the spleen and stomach functions, potentially resulting in impaired transformation and distribution of nutrients and subsequent systemic imbalances. Clinically, it is often associated with symptoms such as poor appetite, digestive difficulties, fatigue, and weakness [45-48]. Among patients with pancreatic cancer, particularly in the middle to advanced stages of disease progression, similar manifestations are frequently observed, including progressive weight loss, malnutrition, and reduced immune function [45]. While the underlying mechanisms remain to be fully elucidated, some clinical observations suggest that TCM interventions aimed at strengthening spleen function—such as treatments based on the principle of invigorating the spleen and replenishing qi—may lead to partial improvements in patient-reported symptoms and general condition, potentially enhancing tolerance to antitumor therapies [49]. These findings are still preliminary and require further high-quality studies for confirmation; however, they tentatively indicate that spleen-deficiency syndrome may represent a common clinical phenotype in patients with pancreatic cancer. Under the guidance of TCM's holistic and pattern-differentiation principles, regulation of spleen function is viewed as a potentially supportive therapeutic approach that may contribute to improving patients' functional status and mitigating treatment-related burdens.

DH, however, is often considered to be closely associated with dietary and lifestyle factors. TCM holds that excessive consumption of greasy, rich foods, alcohol, or unhygienic dietary practices may impair spleen and stomach function, leading to internal accumulation of dampness and stagnation of qi, which over time may transform into heat, giving rise to damp-heat pathology. For instance, alcohol is believed to be both "hot in

XSL•FO
RenderX

nature and damp in substance," and its overconsumption may exacerbate damp-heat accumulation. Similarly, rich foods are thought to contribute to the generation of dampness and heat [44]. In clinical practice, symptoms such as loss of appetite, abdominal distension, nausea, and jaundice—commonly observed in pancreatic cancer—are often interpreted within the TCM framework as indicative of DH [50]. Although systematic mechanistic studies on the relationship between DH and pancreatic cancer are still lacking, some clinical observations have indicated that patients with DH in advanced, unresectable pancreatic cancer may experience partial symptom relief—such as improvements in appetite, fatigue, and abdominal distension—when treated with a combination of chemotherapy and herbal formulas aimed at clearing heat and resolving dampness [51]. As such, this pattern is also regarded as relatively common among patients with this malignancy. It is worth noting that although some exploratory studies have begun to investigate the possible molecular, cellular, or microbiome-related underpinnings of these syndromes in the context of pancreatic cancer, the overall research in this area remains in an early stage, and the underlying biological mechanisms have yet to be clearly defined [47,48,52-55]. Against this backdrop, leveraging deep learning and other artificial intelligence techniques to conduct more objective and standardized phenotypic assessments of TCM syndromes in clinical populations may help advance syndrome research toward greater precision and empirical rigor. Such efforts may also facilitate the identification of potential biological correlates underlying TCM diagnostic categories, offering new avenues for integrative research.

Feature Labeling in Expert Knowledge Learning for the TCMPCSD-BERT Model and Analysis of Study Results

In this study, TCMPCSD-BERT is designed to learn expert knowledge by identifying the key elements of syndrome differentiation in the training data. The model's learning focus is influenced by several factors, including "knowledge framework construction," "syndrome feature extraction," "syndrome differentiation element weighting and symptom severity assessment," and "syndrome differentiation thresholds setting." A more detailed discussion on how these factors impact the focus of the TCMPCSD-BERT model's learning can be found in Multimedia Appendix 13 (feature labeling in expert knowledge learning for the TCMPCSD-BERT model).

In comparative analyses with the 3 traditional baseline models, McNemar test indicated that TCMPCSD-BERT achieved statistically significantly higher classification accuracy across all 4 TCM syndrome types. Consistently, the mean values for macroprecision, macrorecall, macro–F_1-score, and accuracy were greater for TCMPCSD-BERT in all comparisons, with its 95% bootstrap CIs showing no overlap with those of the comparator models. Welch *t* test further indicated that these differences were statistically significant in every comparison ($P<.01$). All pairwise comparisons exhibited evidence of heteroscedasticity, reflected in unequal variances between the distributions of performance metrics. In such contexts, Welch *t* test is generally considered appropriate, as it is designed to address heteroscedasticity by avoiding the pooling of variances and by adjusting the degrees of freedom, thereby helping to maintain control over type I error rates. On this basis, the statistical inferences from Welch *t* test remain interpretable even in the presence of unequal variances. In the present analysis, "variance" refers to the dispersion of metric distributions across the 5000 bootstrap samples for each model. Unequal variances may arise when models differ in the stability of their predictive performance—such as when 1 model shows minimal variation across resamples while another exhibits greater fluctuations. Such differences may be influenced by the randomness inherent in bootstrap resampling and by the class distribution of the test set, particularly under class imbalance. The presence of unequal variances does not, in itself, imply that the comparative results are unreliable. Rather, the persistence of statistically significant differences under the more conservative conditions of Welch *t* test may indicate that the observed differences are not solely attributable to variability in performance stability across models [56].

In the visualization of the self-attention mechanism of TCMPCSD-BERT, we observed that the model effectively identifies the relational structure between words within terms. For example, in Figure 4A, the terms "苔-薄-黄" (thin yellow coating) and "脉-弦" (wiry pulse), in Figure 4B, "纳-食" (appetite), "苔-白-腻" (white greasy coating), and "脉-细-弦" (thin wiry pulse), and in Figure 4C, "苔-白-厚" (thick white coating), represent combinations of words that are not typical in everyday language. However, TCMPCSD-BERT successfully "notices" the relationships between these words, as indicated by the thicker lines in the visualization, demonstrating the model's ability to construct a vocabulary framework related to syndrome differentiation in pancreatic cancer and effectively identify the structural relationships of these specialized terms.

We used the integrated gradients algorithm to interpret the decision-making process of the TCMPCSD-BERT model. This algorithm assigns attribution scores to each word in the clinical text, with higher scores indicating greater importance of the term in syndrome differentiation. This helps us understand how the model processes and interprets key information in clinical texts. In Figures 5A-5C, we can see that the model assigns higher attribution scores to the diagnostic features corresponding to each syndrome type, emphasizing the critical role these symptoms play in the differentiation process. In Figure 5D, when we select a clinically diagnosed DH case and input the portion of the case that does not include syndrome differentiation features, the model assigns lower attribution scores to these irrelevant contents. This contrasts sharply with the higher scores assigned to the relevant diagnostic symptoms. These results demonstrate that the TCMPCSD-BERT model effectively identifies key diagnostic features across different syndrome types and assigns an appropriate weight to them. At the same time, the model effectively distinguishes irrelevant content from syndrome differentiation, minimizing its attention to such content.

Although the TCMPCSD-BERT model performed well on the test set, some misclassifications still occurred. To further understand the model's error patterns, we selected misclassified cases from 3 typical syndrome types and conducted attribution

XSL•FO
RenderX

analysis to explore the model's shortcomings when handling complex symptoms. In 1 case where DH was misclassified as "spleen-deficiency syndrome" (Figure S1 in Multimedia Appendix 11), the attribution analysis showed that the model assigned lower attribution scores to several key symptoms of DH while assigning higher scores to symptoms associated with spleen-deficiency syndrome. Despite the case exhibiting multiple secondary symptoms related to spleen-deficiency syndrome, the absence of primary symptoms led to an incorrect diagnosis. In Figure S1B in Multimedia Appendix 11, the model misclassified a "spleen-deficiency syndrome" case as DH-SD. The model assigned lower attribution scores to several key symptoms of spleen-deficiency syndrome and higher scores to features associated with DH. Although the case displayed 1 primary symptom and 1 secondary symptom related to DH, it did not meet the diagnostic criteria for DH. In Figure S1C in Multimedia Appendix 11, the model misclassified a DH-SD case as "spleen-deficiency syndrome." The attribution analysis showed that the model assigned lower attribution scores to key symptoms of DH (eg, "low-grade fever" and "abdominal pain"), which could be the cause of the misclassification. Furthermore, the model also assigned higher attribution scores to some clinically irrelevant features, such as "大便秘结" (constipation, 0.451) in Figure S1A in Multimedia Appendix 11, "小便调" (normal urination, 1.399) in Figure S1B in Multimedia Appendix 11, and "二便正常" (normal bowel and urination, 0.162). These misclassifications reflect the model's insensitivity to less common symptoms (eg, low-grade fever) and its potential overreliance on secondary symptoms or confusion between subtle differences in syndrome features. While the model can capture some associations between syndrome features, it may fail to accurately distinguish the weighting of primary and secondary symptoms when faced with ambiguous boundaries between syndrome types, complex multisyndrome symptoms, or atypical symptom presentations, leading to misdiagnosis. Potential causes of misclassification may stem from the wide and sometimes vague definitions of syndromes and symptoms in TCM, which can blur the boundaries between different syndrome types. When the boundaries between syndromes are ambiguous, the model might not be able to differentiate between primary and secondary symptoms effectively. Additionally, the model's reliance on secondary symptoms and its insensitivity to less common symptoms can contribute to misdiagnosis, highlighting the need for more precise and objective symptom definitions in future model improvements.

We compared the performance of the TCMPCSD-BERT model with that of English, Chinese, and TCM-specialized LLMs under the prompt engineering framework for syndrome differentiation tasks. McNemar test results showed that TCMPCSD-BERT generally demonstrated statistically superior classification accuracy across most TCM syndrome differentiation tasks. In particular, when compared with general-purpose LLMs (such as ChatGPT-4, Ernie Bot 4.0 Turbo, and Kimi), TCMPCSD-BERT exhibited significant advantages across all 4 syndrome categories ($P<.01$). However, when compared with more powerful or domain-specialized models (eg, ChatGPT-4o and the TCM-oriented model Zhipu Qingyan), the statistical advantage of TCMPCSD-BERT was not consistent across all tasks. For example, in the classification of "spleen-deficiency syndrome" and DH-SD, several comparisons yielded P values (eg, $P=.072$ and $P=.083$ for ChatGPT-4o; $P=.059$ and $P=.053$ for Zhipu Qingyan) that did not meet the conventional threshold for statistical significance. On the 1 hand, the intrinsic complexity of the 2 syndromes—characterized by overlapping or ambiguous symptom presentations—may increase the difficulty of accurate differentiation, thus posing challenges for all models. However, it is worth noting that spleen-deficiency syndrome (n=1185, 17.3%) and DH-SD (n=1178, 17.2%) represent the 2 syndrome types with the smallest number of cases in the test set. Compared to DH (n=1694) and other syndromes (n=2773), this imbalance in sample distribution may not only affect the sensitivity of statistical analyses but could also limit the TCMPCSD-BERT model's ability to effectively learn syndrome-specific features during training. These factors may have contributed to the model's relatively lower classification performance in these categories. Based on these observations, future studies should consider expanding the sample size for these less-represented syndrome types to enhance the model's robustness and discriminatory capacity in imbalanced classification tasks.

In most cases, these LLMs were able to process the input clinical case content with reasonable accuracy. However, it is important to note that LLMs are sometimes affected by issues such as "large model hallucination" [18], "long-text forgetting" [19], and "instability," which can result in incorrect outputs. In contrast, the TCMPCSD-BERT model demonstrated superior stability in the specific task of syndrome differentiation for pancreatic cancer. The model maintained high diagnostic accuracy when handling longer texts or complex symptoms and effectively mitigated the impact of the "forgetting" issue, ensuring reliable and consistent diagnoses. Furthermore, due to the inherent "long-text forgetting" issue in LLMs, it is often necessary to restart a new conversation and re-enter the prompt framework after 10 to 20 queries. This process makes the use of LLMs cumbersome in practical applications, limiting their ability to provide continuous and stable service. As a result, the TCMPCSD-BERT model exhibited greater stability, accuracy, and operational efficiency in the specific task of pancreatic cancer syndrome differentiation compared to LLMs guided by the prompt engineering framework.

In this study, we conducted a preliminary comparison between the TCMPCSD-BERT model, online syndrome differentiation platforms, and certain 4-diagnostic instruments in terms of unstructured long-text processing performance and specificity in pancreatic cancer syndrome differentiation. In TCM clinical practice, different diseases have their own diagnostic characteristics, meaning that the same symptoms may have significantly different implications for syndrome differentiation depending on the disease. For instance, when inputting the syndrome differentiation elements for pancreatic cancer "damp-heat accumulation syndrome," the platform may mistakenly identify it as a different disease, such as a hepatitis-related syndrome, and fail to assess the pancreatic cancer syndrome content accurately. Moreover, web-based platforms have not yet been able to effectively process long, unstructured clinical case texts. While they can perform basic syndrome differentiation tasks, they have significant limitations

XSL•FO
RenderX

in identifying the specific pathological features of pancreatic cancer, leading to a lack of diagnostic specificity in their outputs. These platforms and 4-diagnostic instruments often require users to manually select syndrome differentiation elements, making them cumbersome when handling unstructured long texts and resulting in lower data processing efficiency. The root cause of this issue may lie in the limitations of computational power and memory in web-based platforms and 4-diagnostic instruments, which typically use lightweight models that are not well-suited for handling unstructured long-text data [57,58]. In contrast, the TCMPCSD-BERT model performs more accurately and efficiently in this task. Furthermore, by building on the TCMPCSD-BERT model and incorporating more powerful models, such as vision transformer, we can create a more efficient and precise syndrome quantification framework. This scalability and potential remain unmatched by current web-based platforms and hardware-based syndrome differentiation tools.

Limitations and Future Directions

Compared to previous NLP studies on TCM syndrome differentiation [14-17], this study demonstrates several key advantages. First, it trained the TCMPCSD-BERT model on a larger dataset comprising 6830 clinical cases, which enhances its generalization ability in real-world applications and allows it to effectively handle the uncertainty and complexity of actual clinical scenarios. Additionally, this study conducted a more comprehensive evaluation and analysis of the TCMPCSD-BERT model, including self-attention visualization and integrated gradient analysis, revealing its alignment with clinical standards. Furthermore, through the analysis of misclassified cases, this study explored the model's limitations in handling complex symptoms. Moreover, this research also performed an initial evaluation of several LLMs with English and Chinese backgrounds, as well as TCM-specific LLMs, assessing their performance in TCM syndrome differentiation tasks for pancreatic cancer under the guidance of the prompt engineering framework, and compared them with TCMPCSD-BERT. The results indicated that, especially in research contexts and specific problem evaluations, customized models currently retain irreplaceable advantages in terms of precision and stability. Additionally, several studies support this view, demonstrating that for specific tasks, small- to medium-scale customized models can achieve and, in some cases, surpass the performance of large-scale models with relatively fewer data requirements [31,59-61].

However, this study does have several limitations. Foremost, it is important to acknowledge that the definitions of "syndrome types" and "syndrome elements" in this study remain relatively subjective and broad. While existing literature has demonstrated a close association between DH and inflammatory microenvironments, as well as a strong link between spleen-deficiency syndrome and energy metabolism, syndrome differentiation continues to predominantly rely on the observation of clinical symptoms and signs. Currently, effective clinical diagnostic indicators to guide syndrome differentiation are still in development. In light of these limitations, there are several potential solutions and improvements for future research. First, it is important to address the broad nature and potential overlap of many syndrome elements in the current differentiation process. We believe that future studies should incorporate the insights of more authoritative experts across various fields to refine and clarify the definitions of the symptoms and signs associated with each syndrome type. Furthermore, incorporating a graded approach, such as differentiating symptoms into mild, moderate, and severe categories, or developing a concise scoring system, would improve the precision of syndrome differentiation. This stratification can help mitigate the broadness of the syndrome elements and reduce overlap, ultimately leading to more accurate and distinct syndrome definitions. Second, this study used independent evaluations by 2 experts and quantified the consistency of their annotations using Cohen κ coefficient, which yielded a value of 0.913, indicating a high level of agreement between the 2 experts in syndrome differentiation. However, only a single round of annotation was conducted. Future research could benefit from implementing multiple rounds of evaluation, which would provide a more robust and reliable means of mitigating subjective biases. By incorporating iterative assessments, we may achieve even more precise evaluations of syndrome characteristics, further enhancing the reliability of the differentiation process. The syndrome differentiation model developed in this study, based on the BERT architecture, aims to assist in automating the identification of syndrome-related information embedded within patient case data. By integrating the BERT architecture with image processing models, a more objective and standardized approach to quantifying syndrome information can be achieved. This method will not only significantly enhance the accuracy of syndrome differentiation but also provide an important tool for subsequent research, assisting in the identification of clinical diagnostic biomarkers for pancreatic cancer and promoting the objectification and precision of TCM syndrome differentiation.

Furthermore, the BERT model requires a substantial amount of data for training. The 6830 clinical cases used in this study are relatively limited for a 4-category classification task, and as the number of classification labels increases, the data requirements grow significantly. As shown in the "Dataset and Model Training Details" section and Table 1, the number of cases corresponding to the 4 syndrome labels is not balanced, and there are noticeable differences in baseline characteristics—such as gender distribution, tumor staging, and treatment plans—across these syndrome categories. Additionally, data collection should ensure the adequate representation and comprehensiveness of symptoms to reflect expert diagnostic thought processes and avoid biases in the model's learning and judgment. The current dataset only includes a subset of common clinical symptoms and signs, leaving many syndrome-related symptoms insufficiently learned. Additionally, to ensure adequate representation of each label and avoid potential modeling bias due to severe class imbalance, this study included only syndrome types with at least 500 corresponding cases as classification targets. Cases with unclear syndrome manifestations or those belonging to underrepresented categories were collectively assigned to the "others" group. While this strategy was intended to improve the model's training stability and reduce the risk of overfitting to poorly represented classes, it may inevitably constrain the model's ability to learn and

XSL•FO
RenderX

recognize rarer but clinically relevant syndrome patterns. This limitation could affect the model's adaptability and generalizability in more diverse or nuanced clinical contexts. Furthermore, the 4 syndrome categories adopted in this study exhibit certain limitations in diagnostic breadth and are insufficient to encompass the complex and diverse syndrome patterns and reasoning processes involved in TCM clinical practice for pancreatic cancer. To address these limitations, future work will aim to incorporate a larger volume of high-quality clinical records and explore transfer learning strategies to extend the diagnostic breadth of the TCMPCSD-BERT model and enhance its potential applicability in real-world clinical scenarios.

Moreover, another important limitation of this study is that all training, validation, and testing data were obtained from a single institution—Fudan University Shanghai Cancer Center. Although we used internal 10-fold cross-validation and withheld a portion of the data as an independent test set, this internal evaluation may not fully capture the model's generalizability to external clinical settings. Given that syndrome-related language expressions, documentation habits, and diagnostic styles can vary across institutions and regions, center-specific data patterns may lead to potential overfitting or bias, thus limiting the external applicability of the TCMPCSD-BERT model. Unfortunately, due to institutional restrictions and data privacy regulations, we were unable to obtain external datasets from other hospitals for cross-institutional validation in this study. Acknowledging this limitation, we plan to expand our collaboration network and actively seek access to deidentified data from additional centers in future work. Such efforts will facilitate a more comprehensive evaluation of model robustness across diverse clinical environments and further enhance the reliability and applicability of the proposed framework.

Conclusions

The TCMPCSD-BERT model demonstrates potential for identifying syndrome-related elements from unstructured clinical case texts and supporting automated, scalable syndrome differentiation in pancreatic cancer. Based on the findings of this study, it appears to offer improved operability over 4-diagnostic instruments and web-based platforms, as well as greater stability and task-specific accuracy than LLMs. Nonetheless, these results should be interpreted with caution due to the subjectivity of syndrome definitions, data imbalance, and reliance on preprocessed and expert-annotated inputs. Future research involving larger and more representative populations is needed to further validate the model and explore its potential for broader application in real-world clinical settings.

Acknowledgments

This research was funded by the National Natural Foundation of China (81930115 and 82174155). The funders had no role in the design of this study; in the collection, analyses, or interpretation of data; in the writing of this paper; or in the decision to publish the results.

Data Availability

The source code underlying this paper is available on GitHub [62]. The datasets generated and analyzed during this study are not publicly available due to patient privacy considerations. However, deidentified data may be made available from the corresponding author upon reasonable request and with approval from the institutional ethics committee.

Authors' Contributions

HB contributed to topic design, paper drafting, data collection, and data annotation. HD was responsible for data analysis, model training, and the design of the analytical framework. CC assisted with this paper's drafting and data collection. YZ and LR annotated the data. ZC supervised the topic selection, provided guidance throughout this study, and facilitated access to clinical data. All authors critically reviewed this paper and approved the final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Workflow of feature extraction, expert annotation, and data preprocessing.
[DOCX File , 40 KB - formative_v9i1e70602_app1.docx]

Multimedia Appendix 2
TCMPCSD-BERT model architecture and training strategy. TCMPCSD-BERT: traditional Chinese medicine pancreatic cancer syndrome differentiation bidirectional encoder representations from transformers.
[DOCX File , 20 KB - formative_v9i1e70602_app2.docx]

Multimedia Appendix 3
Chinese-English comparison for Figure 1.

XSL•FO
RenderX

[DOCX File , 27 KB - formative_v9i1e70602_app3.docx]

Multimedia Appendix 4
Evaluation metrics, McNemar test, and stratified bootstrap statistical analysis for multi-class syndrome differentiation.
[DOCX File , 35 KB - formative_v9i1e70602_app4.docx]

Multimedia Appendix 5
Explanation of syndrome differentiation decisions.
[DOCX File , 24 KB - formative_v9i1e70602_app5.docx]

Multimedia Appendix 6
LLMs prompt engineering. LLM: large language model.
[DOCX File , 31 KB - formative_v9i1e70602_app6.docx]

Multimedia Appendix 7
The Prompt Framework for LLMs in English and Chinese. LLM: large language model.
[DOCX File , 53 KB - formative_v9i1e70602_app7.docx]

Multimedia Appendix 8
Interrater agreement and optimal model parameters.
[DOCX File , 18 KB - formative_v9i1e70602_app8.docx]

Multimedia Appendix 9
Mapping clinical features to TCM guideline terms for syndrome differentiation. TCM: traditional Chinese medicine.
[DOCX File , 19 KB - formative_v9i1e70602_app9.docx]

Multimedia Appendix 10
Attribution analysis of diagnostic features in TCM syndromes. TCM: traditional Chinese medicine.
[DOCX File , 20 KB - formative_v9i1e70602_app10.docx]

Multimedia Appendix 11
Attribution analysis of misclassified cases.
[DOCX File , 560 KB - formative_v9i1e70602_app11.docx]

Multimedia Appendix 12
Comparison of TCMPCSD-BERT, web-based platforms, and diagnostic instruments in pancreatic cancer syndrome differentiation. TCMPCSD-BERT: traditional Chinese medicine pancreatic cancer syndrome differentiation bidirectional encoder representations from transformers.
[DOCX File , 2166 KB - formative_v9i1e70602_app12.docx]

Multimedia Appendix 13
Feature labeling in expert knowledge learning for the TCMPCSD-BERT model. TCMPCSD-BERT: traditional Chinese medicine pancreatic cancer syndrome differentiation bidirectional encoder representations from transformers.
[DOCX File , 23 KB - formative_v9i1e70602_app13.docx]

References

1. Xiang Y, Guo Z, Zhu P, Chen J, Huang Y. Traditional Chinese medicine as a cancer treatment: modern perspectives of ancient but advanced science. Cancer Med 2019;8(5):1958-1975 [FREE Full text] [doi: 10.1002/cam4.2108] [Medline: 30945475]
2. Ng JY, Bhatt HA, Raja M. Complementary and alternative medicine mention and recommendations in pancreatic cancer clinical practice guidelines: a systematic review and quality assessment. Integr Med Res 2023;12(1):100921 [FREE Full text] [doi: 10.1016/j.imr.2023.100921] [Medline: 36684828]
3. Wu J, Tang G, Cheng C, Yeerken R, Chan Y, Fu Z, et al. Traditional Chinese medicine for the treatment of cancers of hepatobiliary system: from clinical evidence to drug discovery. Mol Cancer 2024;23(1):218 [FREE Full text] [doi: 10.1186/s12943-024-02136-2] [Medline: 39354529]

XSL•FO
RenderX

4. Zhang Y, Xu H, Li Y, Sun Y, Peng X. Advances in the treatment of pancreatic cancer with traditional Chinese medicine. Front Pharmacol 2023;14:1089245 [FREE Full text] [doi: 10.3389/fphar.2023.1089245] [Medline: 37608897]
5. Wang J, Wang Q, Zhang P, Zhang R, He J. Efficacy and safety of traditional Chinese medicine for the treatment of pancreatic cancer: an overview of systematic reviews and meta-analyses. Front Pharmacol 2022;13:896017 [FREE Full text] [doi: 10.3389/fphar.2022.896017] [Medline: 36120323]
6. Jiang M, Lu C, Zhang C, Yang J, Tan Y, Lu A, et al. Syndrome differentiation in modern research of traditional Chinese medicine. J Ethnopharmacol 2012;140(3):634-642 [FREE Full text] [doi: 10.1016/j.jep.2012.01.033] [Medline: 22322251]
7. Vaswani A, Shazeer N, Parmar N. Attention is all you need. : ACM; 2017 Presented at: Proceedings of the 31st International Conference on Neural Information Processing Systems; December 4-9, 2017; Long Beach, CA p. 6000-6010.
8. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. : Association for Computational Linguistics; 2019 Presented at: Proceedings of NAACL-HLT; June 2-7, 2019; Minneapolis, MN p. 4171-4186.
9. Bengio Y, Ducharme R, Vincent P, Janvin C. A neural probabilistic language model. J Mach Learn Res 2003;3:1137-1155. [doi: 10.1007/10985687_6]
10. Gatto J, Seegmiller P, Johnston G, Preum SM. Identifying the perceived severity of patient-generated telemedical queries regarding COVID: developing and evaluating a transfer learning-based solution. JMIR Med Inform 2022;10(9):e37770 [FREE Full text] [doi: 10.2196/37770] [Medline: 35981230]
11. Zhang R, Liu Y, Zhang Z, Luo R, Lv B. Interpretable machine learning model for predicting postpartum depression: retrospective study. JMIR Med Inform 2025;13:e58649 [FREE Full text] [doi: 10.2196/58649] [Medline: 39864955]
12. Baker W, Colditz JB, Dobbs PD, Mai H, Visweswaran S, Zhan J, et al. Classification of Twitter vaping discourse using BERTweet: comparative deep learning study. JMIR Med Inform 2022;10(7):e33678 [FREE Full text] [doi: 10.2196/33678] [Medline: 35862172]
13. Kim Y. Convolutional neural networks for sentence classification. 2014 Presented at: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); August 22, 2025; Doha, Qatar p. 1746-1751. [doi: 10.3115/v1/d14-1181]
14. Achiam O, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, et al. GPT-4 technical report. arXiv Preprint posted online on March 4, 2024. [doi: 10.48550/arXiv.2303.08774]
15. Zhang H, Yin W, Fang Y. ERNIE-ViLG: unified generative pre-training for bidirectional vision-language generation. arXiv Preprint posted online on December 31, 2021. [doi: 10.48550/arXiv.2112.15283]
16. Qin R, Li Z, He W. Mooncake: a KVCache-centric disaggregated architecture for LLM serving. arXiv Preprint posted online on July 9, 2024. [doi: 10.48550/arXiv.2407.00079]
17. ChatGLM3: Open Bilingual Chat LLMs. GitHub. 2024. URL: https://github.com/THUDM/ChatGLM3 [accessed 2025-08-28]
18. Kim J, Cho ES, Paik JY. Feature engineering using file layout for malware detection. arXiv Preprint posted on April 5, 2023. [doi: 10.48550/arXiv.2304.02260]
19. Dai Z, Yang Z, Yang Y, Carbonell J, Le Q, Salakhutdinov R. Transformer-XL: attentive language models beyond a fixed-length context. 2019 Presented at: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; August 22, 2025; Florence, Italy p. 2978-2988. [doi: 10.18653/v1/p19-1285]
20. Huang L, Wang Q, Duan Q, Shi W, Li D, Chen W, et al. TCMSSD: a comprehensive database focused on syndrome standardization. Phytomedicine 2024;128:155486 [FREE Full text] [doi: 10.1016/j.phymed.2024.155486] [Medline: 38471316]
21. Tian D, Chen W, Xu D, Xu L, Xu G, Guo Y, et al. A review of traditional Chinese medicine diagnosis using machine learning: inspection, auscultation-olfaction, inquiry, and palpation. Comput Biol Med 2024;170:108074. [doi: 10.1016/j.compbiomed.2024.108074] [Medline: 38330826]
22. Pan D, Guo Y, Fan Y, Wan H. Development and application of traditional Chinese medicine using AI machine learning and deep learning strategies. Am J Chin Med 2024;52(3):605-623. [doi: 10.1142/S0192415X24500265] [Medline: 38715181]
23. Hu C, Zhang S, Gu T, Yan Z, Jiang J. Multi-task joint learning model for Chinese word segmentation and syndrome differentiation in traditional Chinese medicine. Int J Environ Res Public Health 2022;19(9):5601 [FREE Full text] [doi: 10.3390/ijerph19095601] [Medline: 35564995]
24. Haibo He, Garcia E. Learning from imbalanced data. IEEE Trans Knowl Data Eng 2009;21(9):1263-1284. [doi: 10.1109/TKDE.2008.239]
25. Krawczyk B. Learning from imbalanced data: open challenges and future directions. Prog Artif Intell 2016;5(4):221-232. [doi: 10.1007/s13748-016-0094-0]
26. Buda M, Maki A, Mazurowski MA. A systematic study of the class imbalance problem in convolutional neural networks. Neural Netw 2018;106:249-259. [doi: 10.1016/j.neunet.2018.07.011] [Medline: 30092410]
27. Hou W. Integrated Chinese and Western Medicine in Oncology. Beijing: People's Medical Publishing House; 2022.
28. Fang Z. Principles of Syndrome Differentiation and Treatment in TCM. Shanghai: Shanghai University of Traditional Chinese Medicine Press; 2008.

XSL•FO
RenderX

29. Borelli JL, Wang Y, Li FH, Russo LN, Tironi M, Yamashita K, et al. Detection of depressive symptoms in college students using multimodal passive sensing data and light gradient boosting machine: longitudinal pilot study. JMIR Form Res 2025;9:e67964 [FREE Full text] [doi: 10.2196/67964] [Medline: 40460426]
30. Sun C, Qiu X, Xu Y, Huang X. How to fine-tune BERT for text classification? In: Chinese Computational Linguistics. Cham: Springer International Publishing; 2019:194-206.
31. Li M, Wang L, Wu Q, Zhu J, Zhang M. Diagnosis knowledge constrained network based on first-order logic for syndrome differentiation. Artif Intell Med 2024;147:102739. [doi: 10.1016/j.artmed.2023.102739] [Medline: 38044249]
32. Campos S, Nuñez D, Pérez JC, Robinson J. Characterization of psychopathology in latin American adolescents using a web-based screening tool: cross-sectional Study. JMIR Form Res 2024;8:e57038 [FREE Full text] [doi: 10.2196/57038] [Medline: 39116425]
33. Hazra A, Gogtay N. Biostatistics series module 4: comparing groups - categorical variables. Indian J Dermatol 2016;61(4):385-392 [FREE Full text] [doi: 10.4103/0019-5154.185700] [Medline: 27512183]
34. Chaudhari AS, Grissom MJ, Fang Z, Sveinsson B, Lee JH, Gold GE, et al. Diagnostic accuracy of quantitative multicontrast 5-minute knee MRI using prospective artificial intelligence image quality enhancement. AJR Am J Roentgenol 2021;216(6):1614-1625 [FREE Full text] [doi: 10.2214/AJR.20.24172] [Medline: 32755384]
35. Jung H, Lee H, Nam EW. Mediating effect of social capital on the association between digital literacy and life satisfaction among older adults in South Korea: cross-sectional study. JMIR Form Res 2025;9:e68163 [FREE Full text] [doi: 10.2196/68163] [Medline: 39916363]
36. Sinci KA, Koska IO, Cetinoglu YK, Erdogan N, Koc AM, Eliyatkin NO, et al. Deep learning-based classification of parotid gland tumors: integrating dynamic contrast-enhanced MRI for enhanced diagnostic accuracy. BMC Med Imaging 2025;25(1):264 [FREE Full text] [doi: 10.1186/s12880-025-01814-x] [Medline: 40615947]
37. Chu C, Lu C, Yeung WK. Systematic evaluation of atrous spatial pyramid pooling in U-net for pore segmentation in plasma electrolytic oxidation coatings. Langmuir 2025;41(25):16368-16377. [doi: 10.1021/acs.langmuir.5c01673] [Medline: 40523154]
38. Yao L, Jin Z, Mao C, Zhang Y, Luo Y. Traditional Chinese medicine clinical records classification with BERT and domain specific corpora. J Am Med Inform Assoc 2019;26(12):1632-1636 [FREE Full text] [doi: 10.1093/jamia/ocz164] [Medline: 31550356]
39. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. In: ICML'17: Proc 34th Int Conf Mach Learn. 2017 Presented at: International Conference on Machine Learning 2017; August 6-11, 2017; Sydney, Australia p. 3319-3328. [doi: 10.48550/arXiv.1703.01365]
40. Nori H, Lee Y, Zhang S. Can generalist foundation models outcompete special-purpose tuning? Case study in medicine. arXiv Preprint posted online on November 28, 2023. [doi: 10.48550/arXiv.2311.16452]
41. Wang X, Wei J, Schuurmans D, Le Q, Chi EH, Zhou D. Self-consistency improves chain of thought reasoning in language models. arXiv Preprint posted online on March 7, 2023. [doi: 10.48550/arXiv.2203.11171]
42. Zhang J, Sun K, Jagadeesh A, Falakaflaki P, Kayayan E, Tao G, et al. The potential and pitfalls of using a large language model such as ChatGPT, GPT-4, or LLaMA as a clinical assistant. J Am Med Inform Assoc 2024;31(9):1884-1891. [doi: 10.1093/jamia/ocae184] [Medline: 39018498]
43. Liu NF, Lin K, Hewitt J. Lost in the middle: how language models use long contexts. Trans Assoc Comput Linguist 2023;12:157-173. [doi: 10.1162/tacl_a_00638]
44. Maciocia G. The Foundations of Chinese Medicine: A Comprehensive Text. London, England: Churchill Livingstone; 2015.
45. Chen Z, Wang P. Clinical distribution and molecular basis of traditional Chinese medicine ZHENG in Cancer. Evid Based Complement Alternat Med 2012;2012:783923 [FREE Full text] [doi: 10.1155/2012/783923] [Medline: 22829858]
46. Gao S, Chen L, Wang P, Liu L, Chen Z. MicroRNA expression in salivary supernatant of patients with pancreatic cancer and its relationship with ZHENG. Biomed Res Int 2014;2014:756347 [FREE Full text] [doi: 10.1155/2014/756347] [Medline: 25126577]
47. Dai H, Wang P, Feng L, Liu L, Meng Z, Zhu X, et al. The molecular mechanisms of traditional Chinese medicine ZHENG syndromes on pancreatic tumor growth. Integr Cancer Ther 2010;9(3):291-297 [FREE Full text] [doi: 10.1177/1534735410373922] [Medline: 20702498]
48. Chen Z, Chen L, Wang P, Dai H, Gao S, Wang K. Tumor microenvironment varies under different TCM ZHENG models and correlates with treatment response to herbal medicine. Evid Based Complement Alternat Med 2012;2012:635702 [FREE Full text] [doi: 10.1155/2012/635702] [Medline: 22690248]
49. Cao N, Zhao A, Zhao G, Wang X, Han B, Lin R, et al. Survival analysis of 272 patients with pancreatic cancer undergoing combined treatment. Integr Cancer Ther 2015;14(2):133-139 [FREE Full text] [doi: 10.1177/1534735414564185] [Medline: 25567328]
50. Wang J, Wu C, Liu B, Zhang R. Analysis on the difference of median survival time of advanced pancreatic cancer patients with different traditional Chinese medicine syndromes and intervention times of Chinese herbal medicine： a retrospective study. J Tradit Chin Med 2024;65:2332-2337. [doi: 10.13288/j.11-2166/r.2024.22.009]

XSL•FO
RenderX

51. Yang Z, Li T, Bao X, Fan Y, Qiu D. Evaluation of efficacy of Qingre Huashi decoction on advanced pancreatic cancer patients with damp heat accumulation. Clin J Tradit Chin Med 2022;34(03):517-522. [doi: 10.16448/j.cjtcm.2022.0324]
52. Ji Q, Luo Y, Wang W, Liu X, Li Q, Su S. Research advances in traditional Chinese medicine syndromes in cancer patients. J Integr Med 2016;14(1):12-21. [doi: 10.1016/S2095-4964(16)60237-6] [Medline: 26778224]
53. Yin J, Shi W, Zhu X, Chen Z, Liu L. Qingyihuaji formula inhibits progress of liver metastases from advanced pancreatic cancer xenograft by targeting to decrease expression of Cyr61 and VEGF. Integr Cancer Ther 2012;11(1):37-47 [FREE Full text] [doi: 10.1177/1534735411400315] [Medline: 21382954]
54. Jiao J, Cheng C, Xu P, Yang P, Ruan L, Chen Z. A mouse model of damp-heat syndrome in traditional Chinese medicine and its impact on pancreatic tumor growth. Front Oncol 2022;12:947238 [FREE Full text] [doi: 10.3389/fonc.2022.947238] [Medline: 35957897]
55. Tong Y, Han F, Liu M, Xu T, Zhang A, Qin J, et al. Characteristics of gut microbiome in the murine model of pancreatic cancer with damp-heat syndrome. Biomedicines 2024;12(10):2360 [FREE Full text] [doi: 10.3390/biomedicines12102360] [Medline: 39457673]
56. Ruxton GD. The unequal variance t-test is an underused alternative to Student's t-test and the Mann-Whitney U test. Behavioral Ecology 2006;17(4):688-690. [doi: 10.1093/beheco/ark016]
57. Han X, Cai Z, Zhang Y, Fan C. Hermes: memory-efficient pipeline inference for large models on edge devices. 2024 Presented at: 2024 IEEE 42nd International Conference on Computer Design (ICCD); November 18-20, 2024; Milan, Italy. [doi: 10.1109/ICCD63220.2024.00076]
58. Li C, Lammie C, Gallo M, Rajendran B. Efficient deployment of transformer models in analog in-memory computing hardware. arXiv:2411.17367 2024. [doi: 10.48550/arXiv.2411.17367]
59. Pang H, Wei S, Zhao Y, He L, Wang J, Liu B, et al. Effective attention-based network for syndrome differentiation of AIDS. BMC Med Inform Decis Mak 2020;20(1):264 [FREE Full text] [doi: 10.1186/s12911-020-01249-0] [Medline: 33059709]
60. Liu Z, He H, Yan S, Wang Y, Yang T, Li G. End-to-end models to imitate traditional Chinese medicine syndrome differentiation in lung cancer diagnosis: model development and validation. JMIR Med Inform 2020;8(6):e17821 [FREE Full text] [doi: 10.2196/17821] [Medline: 32543445]
61. Chen Z, Zhang D, Liu C, Wang H, Jin X, Yang F, et al. Traditional Chinese medicine diagnostic prediction model for holistic syndrome differentiation based on deep learning. Integr Med Res 2024;13(1):101019 [FREE Full text] [doi: 10.1016/j.imr.2023.101019] [Medline: 38298865]
62. PCSD_Bert. GitHub. 2024. URL: https://github.com/wangzaimieyoutou/PCSD_Bert [accessed 2025-08-29]

Abbreviations

BERT: bidirectional encoder representations from transformers
DH: damp-heat syndrome
DH-SD: damp-heat with spleen-deficiency syndrome
LLM: large language model
LSTM: long short-term memory
NLP: natural language processing
TCM: traditional Chinese medicine
TCMPCSD-BERT: traditional Chinese medicine pancreatic cancer syndrome differentiation bidirectional encoder representations from transformers
Text-CNN: text convolutional neural network
XGBoost: extreme gradient boosting

Edited by J Sarvestan; submitted 26.12.24; peer-reviewed by S Matsuda, LR Guo; comments to author 23.01.25; revised version received 13.04.25; accepted 12.08.25; published 03.10.25.

Please cite as:
Ba H, Du H, Cheng C, Zhang Y, Ruan L, Chen Z
Identification of Syndrome Types in Patients With Pancreatic Cancer From Free Text in Electronic Medical Records: Model Development and Validation
JMIR Form Res 2025;9:e70602
URL: https://formative.jmir.org/2025/1/e70602
doi:10.2196/70602
PMID:

XSL•FO
RenderX

©He Ba, Haojie Du, Chienshan Cheng, Yuan Zhang, Linjie Ruan, Zhen Chen. Originally published in JMIR Formative Research (https://formative.jmir.org), 03.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Exploring the Use of Digital Technology to Support Health Behavior Change in Young People Under the Care of Complications of Excess Weight (CCEW) Clinics: Qualitative Patient-Centered Design Study

Brioney Gee[1], PhD; Bonnie Teague[1,2], PhD; Matt Farrar[3], BA; Victoria Farrar[3], BA; Dorothy Szinay[4], PhD; Li F Chan[5,6], PhD; Ken K Ong[7], PhD; Ben Jackson[8]; Sydney Wylie[9]; Felix Naughton[4], PhD; Jon Wilson[1,2], MPhil; Emma Alice Webb[2,10], BSc, MBBS, PhD

[1]Norfolk and Suffolk NHS Foundation Trust Research, Norfolk and Suffolk NHS Foundation Trust, Norwich, United Kingdom

[2]Norwich Medical School, University of East Anglia, Norwich, United Kingdom

[3]LovedBy Group, Cheshire, Loved By Group, Chester, United Kingdom

[4]School of Health Sciences, University of East Anglia, University of East Anglia, Norwich, United Kingdom

[5]Centre for Endocrinology, William Harvey Research Institute, Barts and the London School of Medicine,, Queen Mary University, London, United Kingdom

[6]Department of Paediatric Endocrinology, The Children's Hospital at the Royal London Hospital, Royal London Hospital, London, United Kingdom

[7]Medical Research Council Epidemiology Unit & Department of Paediatrics, Institute of Metabolic Science, University of Cambridge, Institute of Metabolic Science, University of Cambridge, Cambridge, United Kingdom

[8]Health Innovation East, Health Innovation, Cambridge, United Kingdom

[9]Public Contributor, public contributer, Norwich, United Kingdom

[10]Jenny Lind Children's Hospital, Norfolk and Norwich University Hospital NHS Foundation Trust, Norwich, United Kingdom

Corresponding Author:
Emma Alice Webb, BSc, MBBS, PhD
Norwich Medical School
University of East Anglia
Rosalind Franklin Rd
Norwich, NR4 7UQ
United Kingdom
Phone: 44 1603286286
Email: emma.webb@nnuh.nhs.uk

Abstract

Background: Specialist multidisciplinary clinics have been established to provide care for the burgeoning number of young people presenting with comorbidities related to severe obesity in childhood. Digital technology, an integral component of most young people's lives, may enable clinics to offer accessible, ongoing support between appointments to the patients, thereby increasing the likelihood of successful health behavior change. However, while short-term engagement with technology-based behavior change interventions is good, engagement tends to decrease over time, limiting their overall impact. Little is known about the views of young people living with obesity on the role of digital technology as an adjunct to current traditional care pathways.

Objective: This study aims to explore the views of adolescent patients and their families on whether digital technology should be used by obesity services to support health behavior change.

Methods: Participants included patients aged between 10 and 16 years from an obesity clinic, along with their adult family members. Four focus groups and co-design workshops, facilitated by a cross-disciplinary team of clinicians, academics, and technology innovators, explored young people's health priorities, identified the barriers to and facilitators of health behavior change, and co-designed ways in which technology could be used to support them in overcoming these barriers to achieving their health goals. Data were analyzed using inductive content analysis, with findings integrated with key co-design workshop outputs.

XSL•FO
RenderX

Results: In total, 37 individuals participated, including 19 (51%) adolescents (n=11, 58% female) and 18 (49%) family members. The young participants, on average, were aged 13.4 (SD 1.68; range 10-16) years; the mean BMI was 36.6 (SD 0.3; range 34-47) kg/m^2. The mean socioeconomic decile was 4.3 (SD 2; range 1-8). Participants did not mention weight as an important aspect of their health. Instead, mental health, sleep, and peer support were identified as the domains where patients felt they would most benefit from additional support. Addressing these aspects of health was viewed as foundational to all other aspects of health, with poor mental health, sleep, and social support reducing young people's ability to engage in the process of health behavior change. Participants reported that technology could help provide this support as an adjunct to in-person support. Participants expressed a preference for technologies able to individually tailor content to the young person's needs, including relatable peer-produced content. The need for support for both the young people and their family members was highlighted, along with the need to integrate in-person strategies to maintain engagement with any technological offering.

Conclusions: There is clear potential for digital technology to support the holistic health priorities of young people receiving specialist care for the comorbidities of excess weight. This study's findings will serve as a foundation for developing innovative approaches to the use of technology to support this high-need population.

(JMIR Form Res 2025;9:e64947) doi:10.2196/64947

KEYWORDS

digital technology; health behavior change; obesity; young people; adolescents; participatory research

Introduction

Background

The growing prevalence of childhood obesity is among the most important threats to public health globally [1]. In addition to being at a higher risk of significant childhood comorbidities [2], children living with overweight and obesity are more likely to experience obesity in adulthood [3]. Adults living with obesity are at markedly increased risk of a wide range of noncommunicable diseases, including cardiovascular diseases, diabetes, musculoskeletal disorders, and certain types of cancer [4].

In England, 22.7% of children aged 10 to 11 years were living with obesity in 2022 and 2023, with 5.7% classed as severely obese [5]. This equates to more than 2.5 million children eligible for specialist treatment according to the National Institute for Health and Care Excellence guidance [6]. Obesity currently costs the UK National Health Service (NHS) around £6 (US $8.1) billion annually, and, without a significant step-change in health intervention effectiveness, these costs are estimated to rise to over £9.7 (US $13.1) billion per year by 2050 [7].

Furthermore, childhood obesity is a significant factor in the maintenance of health inequalities [8]. Children growing up in the least economically advantaged neighborhoods are more than twice as likely to experience obesity and 4 times as likely to experience severe obesity as their peers growing up in the most advantaged neighborhoods [5]. In addition, obesity is more prevalent in children from Black and Asian (excluding Chinese) ethnic backgrounds than in children from White ethnic backgrounds [5].

The NHS Long Term Plan [9] is committed to improving the care available to children and young people experiencing health complications related to excess weight. As part of this commitment, 21 new Complications of Excess Weight (CEW) clinics located across England were commissioned in 2021 [10]. CEW clinics provide specialist biopsychosocial care, delivered by a multidisciplinary team comprising consultants, nurses, and allied health professionals (eg, dietitian, physiotherapist, psychologist, family therapist, and social worker). Given the limited evidence base for what constitutes optimal care for young people experiencing severe obesity, a key objective of the pilot clinics was to contribute to generating evidence on the most promising interventions and service models for this cohort.

During their treatment, young people under the care of a CEW clinic are encouraged to make multiple behavioral changes to improve their health, such as increasing activity, changing dietary patterns, and taking regular medication. However, adolescents face multiple barriers to successful and sustained behavioral change [11]. While CEW clinicians aim to support young people and their families to overcome these barriers, this support is only provided during scheduled clinic appointments.

Digital technology is now an integral component of nearly all young people's lives [12]. Consequently, the increased use of digital technology may enable CEW clinics to offer accessible, ongoing support between appointments to their patients to increase the likelihood of successful health behavior change. Systematic reviews of technology-based interventions for overweight and obesity in adolescents have found some evidence of short-term improvement in dietary behaviors and physical activity [13] and significant reductions in BMI [14-16].

However, while short-term engagement with technology-based behavior change interventions is often good, this engagement tends to decrease over time [17-19], likely limiting their impact [20]. Furthermore, despite the acknowledged importance of involving intended end users in the development of digital health technologies [21], most currently available digital weight-management interventions were not developed with the involvement of patients or health care professionals [22].

Objectives

This study aimed to explore the views of young patients and their family members on the potential for digital technology to enhance the support offered by CEW clinics to adolescents living with severe obesity. This study aimed to address the following research questions: (1) What are the most important unaddressed health priorities for adolescent CEW clinic patients

XSL•FO
RenderX

from the perspective of patients and their family members? (2) What do adolescent CEW clinic patients and their family members perceive to be the key barriers to and facilitators of health behavior change, and to what extent do they believe that digital technology could assist in overcoming barriers? (3) What preferences do adolescent CEW clinic patients and their family members have for the design and delivery of digital technology to support health behavior change?

Methods

Design

This study used a participatory design involving a series of focus groups and co-design workshops (Figure 1) with adolescent CEW clinic patients and their adult family members. Each focus group and workshop was conducted twice on the same day to facilitate a larger number of families to participate while maintaining small group sizes (maximum 5 young people and 5 parents or guardians) to encourage participants to feel comfortable sharing their views openly. On the basis of feedback from potential participants and our young advisor, adolescents attended sessions together with their family members to help increase their confidence to participate. However, due to the potential for the presence of adult family members to inhibit the views young people were willing to share (and vice versa), participants were invited to break into separate groups during parts of the later sessions.

Figure 1. Study design schematic for a user-centered design study of digital technology to support young people being treated for obesity. (A) Focus groups and workshops shown in blue were conducted with the first participant group; (B) those shown in green were conducted with the second participant group.

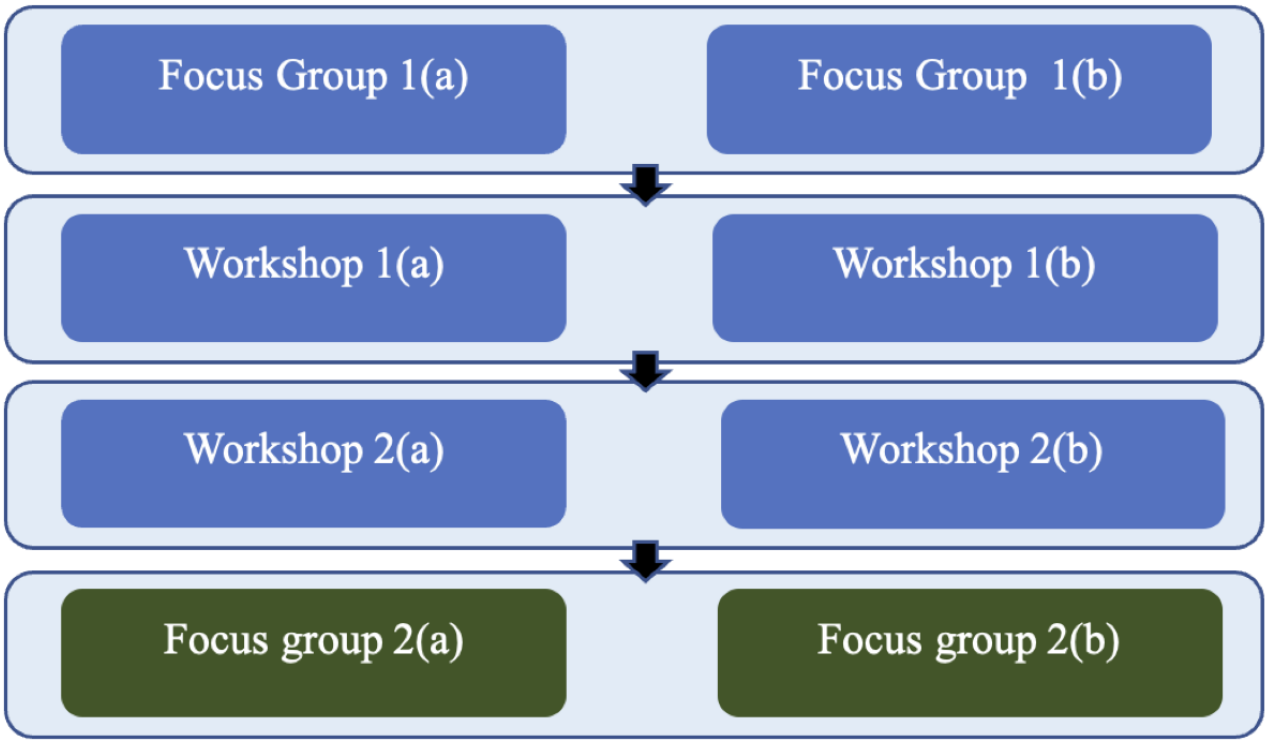

Focus group 1 and workshops 1 and 2 were conducted as an iterative process, with analysis of session 1 being used to inform discussion in the subsequent workshops. The same group of families participated in the first 3 sessions (focus group 1 and workshops 1 and 2). The final focus groups (focus group 2) reviewed the conclusions of focus group 1 and workshops 1 and 2. For the final focus groups, a different group of young people was recruited.

All focus groups and workshops were carried out face-to-face at nonclinical venues to maximize engagement. Each session lasted approximately 90 minutes. The study was carried out over a 12-month period from January to December 2023.

Ethical Considerations

This study was approved by the NHS Health Research Authority following a favorable ethical opinion from the Wales Research Ethics Committee 4 (22/WA/0340). Informed consent (or informed parental consent and the young person's assent in the case of those aged <16 y) was obtained and documented before participation in the study. To maintain confidentiality, each participant was assigned a numerical code. Each family was offered £50 (US $67.50) in high street shopping vouchers per focus group or workshop they attended and reimbursement of any travel expenses in recognition of the time and effort involved in their participation.

Setting and Participants

The Healthy Futures clinic at Norfolk and Norwich University Hospital (NNUH) was established in 2022 as 1 of the 21 pilot CEW clinics commissioned by NHS England. The clinic provides specialist medical, dietetic, psychological, and family support to approximately 100 new patients each year. Children and young people eligible for the service are aged between 2 and 17 years, have a BMI 3.5 SDs above the mean, and have a

XSL•FO
RenderX

suspected or diagnosed physical (eg, hypertension, joint or mobility problems, abnormal glucose metabolism, fatty liver disease, and sleep apnea) or psychological comorbidity.

Young people eligible to participate were patients of the NNUH CEW clinic aged between 10 and 16 years (N=67), who were willing and able to provide informed consent or assent to participate (and for those aged <16 y, whose parent or carer was willing and able to provide informed parental consent for their participation). Participating family members were consenting adult caregivers (aged >18 y) of eligible young participants, largely parents or carers, and also included a small number of other relatives (ie, grandparents and adult siblings).

All eligible families under the care of the NNUH CEW clinic at the time of recruitment were invited to participate via letter, followed up by conversations in face-to-face clinic appointments held in the pediatric department at NNUH. Participants were informed that the research was being conducted by a team of clinical and academic researchers in partnership with a technology company with the aim of improving the support offered by the Healthy Futures Clinic and other similar services.

As the study used a qualitative methodology, a formal sample size calculation was not appropriate. Instead, the sample size was selected based on empirical guidance regarding the number of focus groups and participants needed for a high likelihood of reaching theoretical saturation [23] while also taking into account pragmatic considerations such as the time frame of the study and resources available.

Focus Groups

Focus groups were facilitated by authors BG (female) and BT (female), clinical researchers (both holding PhDs) experienced in facilitating focus groups regarding sensitive topics with young people and their families. In addition, one or more members of the CEW clinical team were present at each focus group and available to support participants. Focus groups were semistructured with content guided by approved topic guides, supported by visual aids presented via a slide deck. Each participant was provided with a notepad and encouraged to write or draw their ideas if they felt more comfortable communicating their views in this way than contributing verbally.

Topic guides were informed by the capability, opportunity, motivation, and behavior (COM-B) model of behavior change [24], which identifies 3 factors—capability, opportunity, and motivation—as necessary for successful behavior change. Topics covered in focus group 1 included (1) participants' understanding of what "living a healthier life" meant, (2) perceived barriers to and facilitators of making changes toward a healthier life, and (3) what additional support participants believed young people accessing a CEW clinic would benefit from to help them make changes toward a healthier life (including the potential role of technology in providing this support).

Topics covered in focus group 2 included (1) the extent to which participants agreed with the health domains prioritized in previous focus groups and workshops as the most important unaddressed needs of CEW clinic patients, (2) participants' views of existing digital health technologies designed to support the health domains prioritized in previous focus groups and workshops, and (3) participants' views of existing health technology platforms through which young people accessing a CEW clinic could be provided with access to additional digital support. The technology platforms presented were chosen based on the priorities young people had highlighted in the preceding focus groups and workshops.

All focus groups were audio recorded and transcribed by members of the research team. All written or drawn contributions were collected and transcribed or photographed.

Co-Design Workshops

Creative co-design workshops were facilitated by authors MF (male) and VF (female), experienced technology design professionals at LovedBy with previous experience of working in collaboration with adolescents to co-design digital health behavior change interventions. In addition, as with the focus groups, one or more members of the CEW clinical team were present at each workshop to support participants as required.

Workshop content was designed to encompass the first 2 stages of the Design Council's Double Diamond design process [25]: the "discover" phase, involving gaining a detailed understanding of the design problem to be addressed through close engagement with the experiences and views of those impacted, and the "define" phase, involving using the insights gathered during the discovery phase to clearly define the design challenge and develop hypothesized solutions to be tested and refined in future phases. Activities involved emotional journey mapping of the young person's day, paper-based cocreation exercises, initial prototyping of ideas generated, and gathering feedback on potential design solutions.

Due to the creative and nonlinear nature of the workshops, these were not audio recorded. Instead, the research team made observational field notes to capture key discussions and collected or photographed all written and visual outputs (ie, notes, drawings, and diagrams) produced by participants during the sessions.

Data Analysis

As co-design involves collaborative real-time identification of challenges and cocreation of solutions between workshop participants and researchers [26], analysis in co-design is not a discrete process occurring after data collection is complete. Instead, initial phases of analysis take place concurrently with data generation through a collaborative process of participants and researchers working together to agree on the focus and outputs of the co-design process [27]. For instance, in workshop 1, the emotional journey mapping exercise was used to generate potential health domains to focus on, which were then discussed and prioritized by participants. In workshop 2, prototype design solutions based on ideas codeveloped in workshop 1 were collaboratively reviewed and refined to generate new insights into participant preferences for design and delivery of digital health technologies.

Analysis of the focus groups followed more conventional qualitative data analysis procedures. This involved inductive content analysis [28,29] of focus group transcripts (by BT and

XSL•FO
RenderX

BG), involving data familiarization, iterative coding, and creating and refining content categories and subcategories.

Following the completion of the co-design process and inductive content analysis, the workshop outputs and focus group findings were reviewed and synthesized to generate overall answers to each of the study's research questions. This process included mapping findings to the COM-B model and the theoretical domains framework of behavior change techniques [30] to support identification of potential intervention components.

Patient and Public Involvement

A young person with personal experience of being supported by a CEW clinic acted as a young advisor and was an integral member of the study team. SW was involved in developing the recruitment strategy, advised on topic guides and workshop schedules, attended focus groups and workshops, and contributed to the interpretation and dissemination of findings.

Results

Overview

In total, 19 young people and 18 family members (n=37) took part in the study; 25 (68%) in the first set of focus groups and workshops 13 (35%) in the final focus groups (one young person was part of both cohorts). Across both cohorts, on average, the young participants were aged 13.4 (SD 1.68; range 10-16) years; mean BMI was 36.6 (SD 0.3; range 34-47) kg/m^2. Of the 19 young people, 11 (58%) were female and 8 (42%) were male, and all were White except for 1 (5%) participant who identified as Asian. The demographic characteristics of participating young people and the wider patient cohort are presented in Table 1. On the basis of participants' home postcodes, the mean socioeconomic decile was 4.3 (SD 2; range 1-8); 7 participants and their parents or guardians came from the lowest 3 socioeconomic deciles. No demographic data were collected for family members.

Table 1. Demographic characteristics of young people who participated in the user-centered design study and the wider cohort of patients under the care of Norfolk and Norwich University Hospital, Complications of Excess Weight clinic.

Characteristic	Study participants (n=19)	Wider patient cohort (n=48)
Age (y), mean (SD; range)	13.4 (1.68; 10-16)	13.6 (1.6; 10.2-15.8)
BMI (kg/m^2), mean (SD; range)	36.6 (0.3; 34-47)	40 (0.3; 32.5-50)
Female sex, n (%)	11 (58)	22 (46)

Research Question 1: What Are the Most Important Unaddressed Health Priorities for Adolescent CEW Clinic Patients and Their Family Members?

When asked to define what living a "healthy life" meant to them, participants' responses fell into 3 categories: achieving good mental health, implementing healthier behavior patterns, and having positive health outcomes. The category of achieving good mental health encompassed feeling happy and "good in yourself" as well as feeling confident and satisfied with their appearance. Health behaviors viewed as important to living a healthier life included eating a healthy diet, incorporating exercise or other enjoyable activities into their daily lives, and regularly getting enough sleep. Participants mentioned wanting to achieve health outcomes such as having more energy and living longer. Participants did not mention losing weight or maintaining a healthy weight as an important aspect of health, except as it related to their mental health, ability to make health behavior changes, or improve their health outcomes.

Three domains emerged as the most important health priorities for both young people and their family members (areas for which adequate support is currently lacking): mental health, sleep, and peer support.

Mental health was viewed as the most important unaddressed health priority by nearly all young people and their family members. The mental health support currently available was often viewed as inadequate due to difficulties accessing specialist mental health services and long waitlists for treatment:

> *Persistent [mental health] problems I like wrote about and described to like doctors for years, nothing happens because the waitlists are that long.... I think it's the thing I'd like to see the most done about, because there seems to be the worst area in my experience.* [Young person]

> *She's had two referrals to CAMHS [child and adolescent mental health services] and they've both been rejected both times. And they're when we've been right at the worst moments as well. And that's where you need the support.* [Family member]

Participants spoke about schools as often being the only source of mental health support available to them. The quality of the mental health support provided by schools was described by participants as highly variable. Some families were highly satisfied with the support they had received from their child's school. However, others were less than satisfied with the support available; several participants reported that the responsibility for pastoral support in their school fell to staff with little training in mental health who were therefore ill-equipped to provide the specialist support they or their child needed.

Mental health was viewed as foundational to all other aspects of health, with poor mental health reducing the young people's ability to engage in the process of health behavior change:

> *Well for me mental your mental health is key, isn't it? You're happy. Everything else falls into place. Then it's come to light that if your if your mental health isn't great, and then everything else drops off, you know.* [Family member]

XSL•FO
RenderX

In addition, sleep was viewed as an important area of unmet need by some young people. This emerged particularly strongly in the emotional journey mapping activity (Multimedia Appendix 1) during which young people and their family members highlighted the impact of poor-quality sleep on the young person's daytime mood, energy, and functioning. The young person's mood and ability to function (for instance, to attend school regularly and on time), in turn, impacted the well-being of the wider family. None of the participants reported having previously received professional support focused specifically on sleep, but some had tried using self-help resources with varied success:

> *She found it so difficult to sleep. We've tried all sorts, at the minute she's listening to Whale noises and stuff like that just to try and block everything else what's going on in her head to go to sleep. But even then, sometimes that's not working and it's difficult for her. And then when it comes to school, it's like "I can't go to school. I don't wanna go to school."* [Family member]

The final area of unmet need highlighted by participants was peer support. Peer support was seen as an important factor in helping young people successfully change their behaviors by nearly all adult family members; good peer support was also seen as a protective factor for mental health. Young people were more divided on the importance of this area depending on the quality of their existing friendships. Some young people were satisfied with the support provided by their existing friendships and therefore did not view peer support as something from which they would benefit:

> *I've always had like really helping like relationships with my friends and my family and that...if I didn't have friends, I don't know how will it feel.* [Young person]

However, others described having few friendships and finding it hard to relate to peers:

> *I've had worse than everyone who I've been in my school with, so there's no point in asking how they feel.* [Young person]

Experiences of bullying and neurodiversity were viewed as key barriers to young people forming supportive peer networks.

The opportunity to meet people facing similar challenges was viewed as being potentially beneficial both for the young people themselves and their family members:

> *At school I think young people are so judged by what they look like. And I think that [opportunities for peer support] would kind of make them not feel as bad. Yeah, if you know what I mean. Because they've got the support, and they'll realise that they're not the only one who struggles.* [Family member]

Adult participants reported that they valued the opportunity to do this during the focus groups and workshops and would appreciate further opportunities to connect with parents and carers in similar situations to themselves.

Research Question 2: What Do Adolescent CEW Clinic Patients and Their Family Members Perceive to Be the Key Barriers to and Facilitators of Health Behavior Change, and to What Extent Do They Believe That Digital Technology Could Assist in Overcoming Barriers?

Poor mental health was cited as a significant barrier to health behavior change. Anxiety related to social interactions and engagement in group health activities, sometimes related to previous experience of bullying or low self-esteem, was viewed as particularly problematic. Poor sleep was also noted as impacting motivation to engage in health behavior change. Disruption to routines, for instance, due to school holidays, and seasonal pressures, such as Christmas, Easter, and summer breaks, were viewed as making it more difficult to maintain healthy behaviors. In addition, financial considerations were important barriers to health behavior change for many families. The high cost of healthy food and activities and limited access to affordable transport, particularly for those living in rural areas, were the main financial barriers identified by the participants:

> *I was trying to find, you know, the money to go and do lots of [active] things, do those things. Time is not an issue...I can give her the time that she needs. Yeah, it's whether I've got the funds to do it.* [Family member]

Social support was viewed as the most important facilitator of health behavior change. The support and encouragement of family were most often seen as imperative, with changes undertaken as a family considered easier to maintain than those made by the young person in isolation:

> *They [family] can encourage you to go out and do like exercise and things like that, and encourage you to eat healthier options, but also they kind of do it as well with you.* [Young person]

Support from peers and school staff was also an important facilitator of making and maintaining changes for some participants.

Digital technology was viewed both as a barrier to and a potential facilitator of health behavior change. The opportunity costs imposed by intensive use of digital technology were seen as a barrier to health behavior change, with time spent engaging with screen-based activities (eg, playing video games, using social media, and watching online content) detracting from the time available for physical activity and encouraging snacking behavior:

> *It makes one want to eat as well, you're sat there doing nothing, well looking at screen, you want, yeah, they, they do because...it's like "oh, I'm sitting here, I can just snack on something." Whereas if they're out and about, they're not thinking about it.* [Family member]

The content young people were exposed to online was also sometimes viewed as creating barriers to health behavior change. For instance, seeing images of "ultrafit" celebrities and

XSL•FO
RenderX

influencers on social media was viewed as creating unrealistic expectations, impacting motivation to change behaviors.

Many young people reported having used health apps, for instance, step trackers or food diaries, and found these helpful, at least to some extent. However, adult family members spoke about the costs associated with some health technologies, creating barriers to sustained use, for instance, health apps that charged a monthly subscription after an initial free trial period:

> *Even working families struggling and even having to buy an app, that's maybe only £4.99 but that's £4.99 that's coming out of the food bill.* [Family member]

Some aspects of social media were also viewed as facilitators of health behavior change both by young people and their family members. Helpful online content that the families spoke about included inspiring stories of health behavior change, ideas for healthy recipes, and suggestions for enjoyable movement, particularly activities and challenges that could be completed together as a family.

Research Question 3: What Preferences Do Adolescent CEW Clinic Patients and Their Family Members Have for the Design and Delivery of Digital Technology to Support Health Behavior Change?

Extensive use of digital technology was reported to be a universal aspect of the lives of all the young participants. As such, support delivered or facilitated via technology was viewed as a potentially useful adjunct to face-to-face support. However, participants were also clear that it should not be a replacement for the support currently available or indeed seen as a substitute for increasing the availability of face-to-face support in areas of unmet need.

Key perceived benefits of using digital technology to provide support included more immediate access than face-to-face support and increased privacy and autonomy for young people in relation to the support they receive:

> *They can dip in and out of that they can actually access for themselves, yeah, when they need it, when they want it and nobody else knows what they're doing.* [Family member]

> *I know that [being able to access support via technology] would help me and I know it would definitely help a lot of my friends that can't speak out, and their friends and their like parents won't give them that support.* [Young person]

Participants viewed digital technology as having the potential to help address all 3 of the unmet health priorities identified: mental health, sleep, and peer support. The privacy and autonomy offered by digital interventions were viewed as particularly valuable in areas that might be more sensitive, such as mental health. Technology was also seen as a practical way of bringing people with similar experiences together to provide peer support and build a sense of community. In addition to facilitating direct content sharing with other families, participants expressed that they would value having access to digital content featuring others who have been through similar experiences. Peer-led content was seen as more relatable than other digital content, lessening feelings of isolation and boosting motivation.

Participants expressed largely positive views of the examples they were shown of existing digital health technologies designed to support mental health (Lumi Nova) [31], sleep (digital cognitive behavioral therapy [Sleepio] [32] for insomnia), and peer support (Kooth) [33]. Suggested adaptations to the example technologies were largely related to the needs or preferences of the specific young person (eg, to make the technology more appropriate for their age range) rather than CEW clinic users more widely.

Most participants were not previously aware of the example technologies discussed, and parents and carers reported that they often had difficulty finding out about the resources available to them. Participants felt that there is a need for a platform bringing together available resources in an easily accessible format. They expressed a preference for a platform that is appealing and appropriate for the targeted age range and able to tailor content to the specific health needs of individual young people, including relatable peer-produced content.

Participants expressed the need to maintain the young person's motivation to engage with a digital platform. They suggested that this could be achieved through integration with online applications the young person already uses frequently or through support from family members and clinicians. Family members felt that they would also benefit from access to digital resources and communities that could help them support the young person.

During data analysis, concepts were mapped under the components of the COM-B model and the constructs of the theoretical domains framework. Behavior change techniques were identified to support future work in the development of intervention components. The concepts mapped under these models and the behavior change techniques identified are presented in Multimedia Appendix 2.

SW is a coauthor of this paper, and her reflections on being involved in the study are as follows:

> *I was interested in getting involved in the project as I need to make some changes to my own health habits. As a young person I think it's a very important project as access to good and accurate advice about health and lifestyle can be hard and daunting. There is often conflicting advice online and things aren't always presented in a way that makes them accessible. It's also not always that young people don't know the information, but that they don't know how to act on it. Particularly if there are circumstances outside their control such as living or financial situation. I know people say it doesn't cost money to be healthy but it can definitely help make it easier.*
>
> *Also, as someone who enjoys programming, the technical side of the project interested me to see how an app could be used and how interactive it could be. During the focus groups I found hearing several young people's perspectives interesting. We had different views sometimes on who would be best to provide information and it was good to discuss our*

XSL•FO
RenderX

ideas and work together to think about what would have the broadest appeal to as many young people as possible.

I think it's important to understand that young people have access to lots of information and that this project needed to focus on what will really draw young people to make changes. For me personally, I would find it most engaging to hear from professionals in short videos/blogs about different topics. I also especially thought the close work with the company, LovedBy, was helpful as they listened to our thoughts and then helped us shape them to what was possible technically.

The challenges I saw with the project were drawing enough information from participants in a short time scale and it would be really good to have follow up sessions even if online as I think people often think of things after the event that would be useful. Success to me on this project would be a resource for young people to find good and accurate information about getting and maintaining a healthy lifestyle. Real advice but with thought about how lifestyles are different for young people depending on their circumstances. Advice on how to make the healthier habits both realistic but also interesting to achieve would be good. Also, a way to set goals and measure achievements within an app/site would encourage people with forming better habits.

Discussion

Principal Findings and Comparison With Prior Work

Digital technology offers the potential to provide young people with severe obesity with access to additional support between CEW clinic appointments to increase the likelihood of successful health behavior change. This study aimed to explore the views of young CEW clinic patients and their family members on the use of digital technology to support them in achieving their health priorities.

Young people and their family members were supportive of the idea of increased use of digital technology by CEW clinics to facilitate access to additional support to address currently unmet needs. However, participants expressed the view that this support should be an adjunct to, not a replacement for, face-to-face support provided by clinicians. This view aligns with the evidence base, as there is currently little evidence for the effectiveness of digital interventions for pediatric obesity as stand-alone interventions outside of a comprehensive package of support [15,34].

When participants talked about health, they did not focus on behavior change that directly affected their weight but on changes that would help them feel better and would indirectly enable them to engage with behavior change. We believe this interesting finding was made possible by the nondirective, participant-led approach adopted in this study. This insight should shape the content of any digital intervention for adolescents under the care of a CEW clinic, ensuring adequate focus on concerns beyond weight to maximize young people's engagement.

Participants identified 3 domains in which they would particularly benefit from increased support: mental health, sleep, and peer support. Improved mental health, sleep, and increased social support (both from peers and family members) were seen as facilitators of wider health behavior change as well as overcoming financial barriers and balancing the benefits of technology against the negative impacts of excessive screen time. Participants identified several ways in which the use of digital technology could be a barrier to health behavior change, including exposure to harmful online content and excessive screen time promoting sedentary behavior and snacking. Care must be taken to minimize these potential negative effects when developing digital technology for this patient group. Many of the barriers to and facilitators of health behavior change identified by participants in this study overlap with those identified in a Canadian study of adolescents living with severe obesity receiving multidisciplinary clinical care [35]. This study highlighted 3 factors that impact behavior change across lifestyle domains: perceived controllability, the impact of mental health, and social relationships and interactions. In line with this study, the Canadian team identified both mental health and sleep as important priority areas for young people living with obesity and concluded that "lifestyle-based interventions for behavior change should evolve to emphasize outcomes beyond weight status to include mental health as a primary intervention focus and outcome."

All 3 domains prioritized by the participants in our study as unmet needs were areas where technology has previously been found to positively impact clinical outcomes. Kooth provides a web-based service that gives children and young people access to an online community of peers and experienced counselors [33]. Accessing Kooth is associated with reductions in psychological distress, suicidal ideation, loneliness, and reported self-harm. Lumi Nova is a mobile app that has been shown to reduce anxiety through the delivery of exposure-based cognitive behavioral therapy strategies via immersive gaming technology [31]. Sleepio is a cognitive behavioral therapy program recommended by the National Institute for Health and Care Excellence for improving sleep and can be accessed through a website or an app [32].

These existing digital interventions were viewed as potentially helpful, but current awareness and the use of digital resources were low. Participants expressed enthusiasm for the creation of a platform bringing together digital resources tailored to the needs of individual young people and their family members, including content created by peers. The use of digital technology to connect young patients and their family members with others with similar experiences was also supported.

In line with previous research [11,16,36], the findings of this study make the role of family support in the success or failure of adolescents' health behavior change efforts clear. As such, any digital technology developed in this area must take into account the needs and preferences of not only the young patient but also their wider family support network. For instance, a digital intervention could include a forum for parents and carers

XSL•FO
RenderX

to share ideas and seek support from both professionals and peers and include lifestyle interventions aimed at the family unit. Young people expressed enthusiasm for activities and challenges aimed at the whole family, suggesting this approach is likely to be more acceptable than an intervention targeting the young patient alone.

Strengths and Limitations

To the best of our knowledge, this is the first study to seek the views of adolescents under the care of a specialist pediatric obesity clinic and their adult family members regarding the use of digital technology to support health behavior change. We were able to engage with a diverse range of family characteristics, accessing the clinic through a creative participatory design facilitated by a cross-disciplinary team of clinicians, academics, innovators, and a young service user. The benefits of involving the intended users of a health technology early in the design process are well established [37].

The limitation of this study is that it was conducted in a single site with patients from only 1 CEW clinic and with a self-selecting sample, possibly limiting the transferability of the findings. Furthermore, the geographical area in which this clinic is located (Norfolk, United Kingdom) has a population with limited ethnic diversity, which was reflected in the demographics of the sample recruited. However, aligned with the population profile of Norfolk, participating families were from diverse socioeconomic and urban-rural backgrounds. Given the higher prevalence of obesity in young people from non-White ethnic backgrounds [5], it will be important that future research explores the relevance of this study's findings to families from a diverse range of cultural backgrounds. Multisite studies will likely be needed to adequately address this limitation in future research.

In addition, young people and their adult family members participated in the focus groups and workshops alongside one another. While this was intended to increase the confidence of young people to participate, and we believe it was successful in this regard, it may have inhibited some participants from both groups from expressing their views as openly as they would have if their family members had not been present. However, young people and family members participating together was also a benefit as it allowed both to respond to each other's views and also allowed for family dynamics likely to be important to the use of digital interventions to be observed.

Conclusions

This study highlights the potential for digital technology to be used to provide young people experiencing health complications related to obesity with additional support to make changes to improve their health. The findings suggest that such technology-based interventions should focus on outcomes beyond weight, provide individually tailored content, and support the needs of the wider family and not just the young patient alone to maximize their impact. Technological support should be offered as an adjunct to multidisciplinary, clinic-based interventions and may require ongoing clinician support to ensure engagement is maintained over time for potential benefits to be realized.

Acknowledgments

The study was funded by the University of East Anglia Health and Social Care Partners and supported by Health Innovation East (formerly Eastern Academic Health Sciences Network). The views expressed in this paper are those of the authors and not necessarily of the University of East Anglia Health and Social Care Partners or Health Innovation East.

The authors would like to thank the young people and family members who spent their time sharing their experiences and ideas with us as part of this study. The authors are also grateful to Jasmine Nabarro (Norfolk and Norwich University Hospital National Health Service Foundation Trust) and Dr Sara Karrar (Norfolk County Council) for their support for the study. Generative artificial intelligence was not used in any part of the manuscript writing.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

FN is an unpaid member of the scientific committee for the Smoke Free app. All other authors declare no conflicts of interest.

Multimedia Appendix 1
Digital representation of the output of an emotional journey mapping activity carried out in the first co-design workshop conducted as part of a user-centered design study of digital technology to support young people being treated for obesity.
[PNG File , 160 KB - formative_v9i1e64947_app1.png]

Multimedia Appendix 2
Recommendations for future development of digital technology to support young people being treated for obesity based on the application of the capability, opportunity, motivation, and behavior model and the theoretical domains framework.
[DOCX File , 18 KB - formative_v9i1e64947_app2.docx]

XSL•FO
RenderX

References

1. Di Cesare M, Sorić M, Bovet P, Miranda J, Bhutta Z, Stevens G, et al. The epidemiological burden of obesity in childhood: a worldwide epidemic requiring urgent action. BMC Med 2019 Nov 25;17(1):212 [FREE Full text] [doi: 10.1186/s12916-019-1449-8] [Medline: 31760948]
2. Davies DS. Time to solve childhood obesity: an independent report by the chief medical officer. e Department of Health and Social Care. 2018. URL: https://assets.publishing.service.gov.uk/media/5d9ddd6be5274a596f829bbd/cmo-special-report-childhood-obesity-october-2019.pdf [accessed 2025-05-29]
3. Simmonds M, Llewellyn A, Owen CG, Woolacott N. Predicting adult obesity from childhood obesity: a systematic review and meta-analysis. Obes Rev 2016 Feb 23;17(2):95-107 [FREE Full text] [doi: 10.1111/obr.12334] [Medline: 26696565]
4. Obesity and overweight. World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight [accessed 2025-05-29]
5. National child measurement programme. Office for National Statistics. URL: https://digital.nhs.uk/data-and-information/publications/statistical/national-child-measurement-programme/2022-23-school-year [accessed 2025-05-29]
6. Obesity: clinical assessment and management: quality standard: reference number: QS127. National Institute for Health and Care Excellence. URL: https://www.nice.org.uk/guidance/qs127 [accessed 2025-05-29]
7. Holmes J. Tackling obesity: the role of the NHS in a whole-system approach. The King's Fund. URL: https://www.kingsfund.org.uk/insight-and-analysis/reports/tackling-obesity-nhs [accessed 2025-05-29]
8. Bhutta ZA, Norris SA, Roberts M, Singhal A. The global challenge of childhood obesity and its consequences: what can be done? Lancet Glob Health 2023 Aug;11(8):e1172-e1173. [doi: 10.1016/S2214-109x(23)00284-X]
9. NHS long term plan. Department of Health and Social Care. URL: https://tinyurl.com/4e9ktk83 [accessed 2025-05-29]
10. Complications from Excess Weight (CEW) clinics for children. National Health Service. URL: https://www.england.nhs.uk/get-involved/cyp/specialist-clinics-for-children-and-young-people-living-with-obesity/ [accessed 2025-05-29]
11. Kebbe M, Perez A, Buchholz A, McHugh TF, Scott SS, Richard C, et al. Barriers and enablers for adopting lifestyle behavior changes in adolescents with obesity: a multi-centre, qualitative study. PLoS One 2018 Dec 18;13(12):e0209219 [FREE Full text] [doi: 10.1371/journal.pone.0209219] [Medline: 30562377]
12. Internet users, UK: 2020. Office for National Statistics. URL: https://www.ons.gov.uk/businessindustryandtrade/itandinternetindustry/bulletins/internetusers/2020 [accessed 2024-01-04]
13. Rose T, Barker M, Maria Jacob C, Morrison L, Lawrence W, Strömmer S, et al. A systematic review of digital interventions for improving the diet and physical activity behaviors of adolescents. J Adolesc Health 2017 Dec;61(6):669-677 [FREE Full text] [doi: 10.1016/j.jadohealth.2017.05.024] [Medline: 28822682]
14. Alotaibi M, Alnajjar F, Cappuccio M, Khalid S, Alhmiedat T, Mubin O. Efficacy of emerging technologies to manage childhood obesity. Diabetes Metab Syndr Obes 2022 Apr;Volume 15:1227-1244. [doi: 10.2147/dmso.s357176]
15. Kouvari M, Karipidou M, Tsiampalis T, Mamalaki E, Poulimeneas D, Bathrellou E, et al. Digital health interventions for weight management in children and adolescents: systematic review and meta-analysis. J Med Internet Res 2022 Feb 14;24(2):e30675 [FREE Full text] [doi: 10.2196/30675] [Medline: 35156934]
16. Azevedo LB, Stephenson J, Ells L, Adu-Ntiamoah S, DeSmet A, Giles EL, et al. The effectiveness of e-health interventions for the treatment of overweight or obesity in children and adolescents: a systematic review and meta-analysis. Obes Rev 2022 Feb 07;23(2):e13373. [doi: 10.1111/obr.13373] [Medline: 34747118]
17. Hawkes RE, Miles LM, Ainsworth B, Ross J, Meacock R, French DP. Engagement with a nationally-implemented digital behaviour change intervention: Usage patterns over the 9-month duration of the National Health Service Digital Diabetes Prevention Programme. Internet Interv 2023 Sep;33:100647 [FREE Full text] [doi: 10.1016/j.invent.2023.100647] [Medline: 37502122]
18. Lipschitz JM, Pike CK, Hogan TP, Murphy SA, Burdick KE. The engagement problem: a review of engagement with digital mental health interventions and recommendations for a path forward. Curr Treat Options Psychiatry 2023 Sep;10(3):119-135 [FREE Full text] [doi: 10.1007/s40501-023-00297-3] [Medline: 38390026]
19. Yardley L, Spring BJ, Riper H, Morrison LG, Crane DH, Curtis K, et al. Understanding and promoting effective engagement with digital behavior change interventions. Am J Prev Med 2016 Nov;51(5):833-842. [doi: 10.1016/j.amepre.2016.06.015] [Medline: 27745683]
20. Tu AW, Watts AW, Chanoine JP, Panagiotopoulos C, Geller J, Brant R, et al. Does parental and adolescent participation in an e-health lifestyle modification intervention improves weight outcomes? BMC Public Health 2017 Apr 24;17(1):352 [FREE Full text] [doi: 10.1186/s12889-017-4220-0] [Medline: 28438202]
21. Partridge SR, Redfern J. Strategies to engage adolescents in digital health interventions for obesity prevention and management. Healthcare (Basel) 2018 Jun 21;6(3):70 [FREE Full text] [doi: 10.3390/healthcare6030070] [Medline: 29933550]
22. Rivera J, McPherson A, Hamilton J, Birken C, Coons M, Iyer S, et al. Mobile apps for weight management: a scoping review. JMIR Mhealth Uhealth 2016 Jul 26;4(3):e87 [FREE Full text] [doi: 10.2196/mhealth.5115] [Medline: 27460502]
23. Guest G, Namey E, McKenna K. How many focus groups are enough? Building an evidence base for nonprobability sample sizes. Field Methods 2016 Jul 24;29(1):3-22. [doi: 10.1177/1525822X16639015]

XSL•FO
RenderX

24. Michie S, van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. Implement Sci 2011 Apr 23;6(1):42 [FREE Full text] [doi: 10.1186/1748-5908-6-42] [Medline: 21513547]
25. The double diamond. Design Council. URL: https://www.designcouncil.org.uk/our-resources/the-double-diamond/ [accessed 2025-05-29]
26. Knowles SE, Ercia A, Caskey F, Rees M, Farrington K, Van der Veer SN. Participatory co-design and normalisation process theory with staff and patients to implement digital ways of working into routine care: the example of electronic patient-reported outcomes in UK renal services. BMC Health Serv Res 2021 Jul 18;21(1):706 [FREE Full text] [doi: 10.1186/s12913-021-06702-y] [Medline: 34273978]
27. Langley J, Wolstenholme D, Cooke J. 'Collective making' as knowledge mobilisation: the contribution of participatory design in the co-creation of knowledge in healthcare. BMC Health Serv Res 2018 Jul 25;18(1):585 [FREE Full text] [doi: 10.1186/s12913-018-3397-y] [Medline: 30045726]
28. Elo S, Kyngäs H. The qualitative content analysis process. J Adv Nurs 2008 Apr 18;62(1):107-115. [doi: 10.1111/j.1365-2648.2007.04569.x] [Medline: 18352969]
29. Vears DF, Gillam L, Vears D. Inductive content analysis: a guide for beginning qualitative researchers. Focus Health Prof Educ 2022 Mar 31;23(1):111-127. [doi: 10.11157/fohpe.v23i1.544]
30. Cane J, O'Connor D, Michie S. Validation of the theoretical domains framework for use in behaviour change and implementation research. Implement Sci 2012 Apr 24;7:37 [FREE Full text] [doi: 10.1186/1748-5908-7-37] [Medline: 22530986]
31. Lockwood J, Williams L, Martin JL, Rathee M, Hill C. Effectiveness, user engagement and experience, and safety of a mobile app (Lumi Nova) delivering exposure-based cognitive behavioral therapy strategies to manage anxiety in children via immersive gaming technology: preliminary evaluation study. JMIR Ment Health 2022 Jan 24;9(1):e29008 [FREE Full text] [doi: 10.2196/29008] [Medline: 35072644]
32. Sleepio to treat insomnia and insomnia symptoms: medical technologies guidance: reference number: MTG70. National Institute for Health and Care Excellence. 2022. URL: http://www.nice.org.uk/guideance/mtg70 [accessed 2025-05-29]
33. Coote L, Kelly L, Graham C, Curtis-Gretton L, Green M, Salhi L, et al. An early economic evaluation of Kooth, a web-based mental health platform for children and young people with emerging mental health needs. Internet Interv 2024 Jun;36:100748 [FREE Full text] [doi: 10.1016/j.invent.2024.100748] [Medline: 38803649]
34. Wickham CA, Carbone ET. Who's calling for weight loss? A systematic review of mobile phone weight loss programs for adolescents. Nutr Rev 2015 Jun;73(6):386-398. [doi: 10.1093/nutrit/nuu018] [Medline: 26011913]
35. Rivera J, McPherson AC, Hamilton J, Birken C, Coons M, Peters M, et al. User-centered design of a mobile app for weight and health management in adolescents with complex health needs: qualitative study. JMIR Form Res 2018 Apr 04;2(1):e7 [FREE Full text] [doi: 10.2196/formative.8248] [Medline: 30684409]
36. Niemeier BS, Hektner JM, Enger KB. Parent participation in weight-related health interventions for children and adolescents: a systematic review and meta-analysis. Prev Med 2012 Jul;55(1):3-13. [doi: 10.1016/j.ypmed.2012.04.021] [Medline: 22575353]
37. Blower S, Swallow V, Maturana C, Stones S, Phillips R, Dimitri P, et al. Children and young people's concerns and needs relating to their use of health technology to self-manage long-term conditions: a scoping review. Arch Dis Child 2020 Nov 22;105(11):1093-1104 [FREE Full text] [doi: 10.1136/archdischild-2020-319103] [Medline: 32444448]

Abbreviations

CEW: Complications of Excess Weight
COM-B: capability, opportunity, motivation, and behavior
NHS: National Health Service
NNUH: Norfolk and Norwich University Hospital

Edited by A Mavragani; submitted 01.08.24; peer-reviewed by T Baranowski, H Bhurawala; comments to author 29.10.24; revised version received 26.11.24; accepted 12.01.25; published 15.10.25.

Please cite as:
Gee B, Teague B, Farrar M, Farrar V, Szinay D, Chan LF, Ong KK, Jackson B, Wylie S, Naughton F, Wilson J, Webb EA
Exploring the Use of Digital Technology to Support Health Behavior Change in Young People Under the Care of Complications of Excess Weight (CCEW) Clinics: Qualitative Patient-Centered Design Study
JMIR Form Res 2025;9:e64947
URL: https://formative.jmir.org/2025/1/e64947
doi:10.2196/64947
PMID:

XSL•FO
RenderX

©Brioney Gee, Bonnie Teague, Matt Farrar, Victoria Farrar, Dorothy Szinay, Li F Chan, Ken K Ong, Ben Jackson, Sydney Wylie, Felix Naughton, Jon Wilson, Emma Alice Webb. Originally published in JMIR Formative Research (https://formative.jmir.org), 15.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Toward Comprehensive Assessment of Beliefs and Attitudes Related to Physical Activity in Young Adults: Pilot Study

Theodorus D B Noordover[1,2], MSc; Aave Hannus[1,2], PhD; Kenn Konstabel[1,2], PhD

[1]Department of Psychology, University of Tartu, Tartu, Estonia

[2]National Institute for Health Development, Tallinn, Estonia

Corresponding Author:
Kenn Konstabel, PhD
National Institute for Health Development
Paldiski mnt 80
Tallinn, 10617
Estonia
Phone: 372 56676445
Email: kenn.konstabel@tai.ee

Abstract

Background: Studies show that despite the positive effects of physical activity (PA), most university students are not active enough. For interventions, it is necessary to understand the determinants of PA behavior. Theory of planned behavior (TPB) is one of the most widely used frameworks to describe the psychological determinants of health behavior. Research has shown that in addition to the determinants included in TPB (attitudes, subjective norms, perceived control, and intention), fear of negative outcomes (eg, discomfort or pain) is a major barrier to increasing one's PA. It has been shown that accepting the possibility of unpleasant outcomes may help in reaching one's PA goals.

Objective: The purpose of this study was to create a questionnaire of PA determinants based on TPB and complemented by the topic of acceptance of unpleasant outcomes. The questionnaire is thus meant for evaluating the effectiveness of health psychological PA interventions in university students and young adults.

Methods: This study was carried out using qualitative and quantitative methods and consisted of three phases: (1) an elicitation study for item generation, (2) pretesting for clarity and understanding, and (3) item selection using conceptual and psychometric criteria (based on a pretest with N=447) to maximize domain coverage and avoid redundancy.

Results: A questionnaire covering the core topics of TPB plus acceptance of negative outcomes was constructed, using a combination of qualitative and quantitative methods. The final shortened questionnaire consists of the following question blocks: positive and negative behavioral beliefs, acceptance of negative outcomes, subjective norms (injunctive and descriptive norms, and motivation to comply), and control beliefs. In terms of predicting self-reported PA, the shortened questionnaire was equal to the unabridged version. Notably, negative behavioral beliefs and acceptance of negative outcomes had opposite-signed correlations with self-reported PA (–0.22 and 0.32, $P<.001$). Despite the aim of avoiding redundancy, several item bundles (eg, positive and negative behavioral beliefs, and acceptance of negative outcomes) were highly homogenous in the final version, and are thus usable as psychometric scales.

Conclusions: This questionnaire can assess a range of PA determinants and has good psychometric properties. The questionnaire can be used to assess the beliefs and attitudes (behavioral beliefs, perceived norms, control beliefs, and acceptance of negative consequences) related to PA in young adults when planning interventions, as well as evaluating the effects of health psychological interventions aiming to increase PA.

(JMIR Form Res 2025;9:e69094) doi:10.2196/69094

KEYWORDS

theory of planned behavior; physical activity; university students; questionnaire; acceptance and commitment therapy

XSL•FO
RenderX

Introduction

Background

Physical inactivity is associated with several health risks, which increased physical activity (PA) can reduce [1]. In students, PA has been shown to relate to better overall (mental) health, happiness, and quality of life [2-4] as well as better academic performance [5].

Despite these benefits, many people, including university students, are not physically active enough. Depending on the criteria, the percentage of students not meeting the World Health Organization PA recommendations varies from 22% to 80% [3,5-7], and they spend an average of 7 to 10 hours per day being sedentary [8].

Increasing students' PA can thus be beneficial for physical and mental health, as well as academic performance. For planning interventions, one needs systematic information about the determinants of PA, including attitudes, beliefs, perceived barriers, etc. A review of PA barriers in undergraduate students [9] suggested external barriers (eg, lack of time and facilities) are more prevalent than internal ones. The generality of this conclusion, however, is difficult to test: studies have used different methods (the method of qualitative studies is often not described; in quantitative studies, the choice of barriers to select from differs across studies); several barriers are not easily classifiable as internal versus external (eg, stress or work-related fatigue); the geographical variability in barriers is not systematically considered. A more recent review cited lack of time, motivation, and of accessible places as the most important barriers for high school and university students [10].

Notably, male students generally show higher levels of PA compared to female students [6,11], indicating the need for more tailored interventions.

Attitudes and Beliefs as Determinants of PA

The theory of planned behavior (TPB) is a widely used framework for understanding health behavior and developing interventions [12,13]. The central component in the TPB is the intention to carry out a certain behavior. Intentions, in turn, are influenced by attitudes, subjective norms, and perceived behavioral control [14].

Attitudes refer to "the degree to which a person has a favourable or unfavourable evaluation or appraisal of the behaviour in question" [15]; subjective norms refer to "perceived social pressure to perform or not perform the behaviour" [15]. Perceived behavior control refers to "the perceived ease or difficulty of performing the behaviour" [15]. The more positive these 3 factors are, the stronger the intention to perform the specific behavior [15].

In TPB, attitudes are determined by behavioral beliefs and outcome evaluations [16]. Behavioral beliefs are related to the belief that carrying out the behavior in question is associated with certain results or consequences, while outcome evaluations refer to the valuation of these consequences [16]. For example, one may believe that engaging in PA leads to feeling good (behavioral belief), and this feeling is evaluated as positive (outcome evaluation).

Subjective norms are decomposable into normative beliefs (whether people in general, or some specific referents, approve or disapprove of a behavior) and motivation to comply. However, the actual behaviors of the referents (descriptive norm) may also be important. As a hypothetical example, an individual's parents may approve their being physically active (normative beliefs), engage in PA themselves (descriptive norms), and the individual is somewhat motivated to align with their parents' expectations (motivation to comply) [16].

Perceived control, also called personal agency, is subdivided into control beliefs and perceived power. Control beliefs relate to the likelihood of factors helping or hindering a behavior, while perceived power refers to how much these factors impact the ease or difficulty of performing the behavior [16]. As a hypothetical example, one may believe that one's equipment for skiing is poor (strength of control belief) and that without good equipment, skiing is extremely hard (control belief power). Ajzen [17] argues that the concept of self-efficacy is fundamentally similar to that of control beliefs: self-efficacy and perceived controllability can be conceptualized as partially overlapping but distinct aspects of perceived behavioral control. However, self-efficacy can be thought of as a direct way to ask about one's ability to perform a behavior, whereas the TPB framework offers an indirect measure (strength of control belief multiplied by the power of control belief). The indirect measure cannot be implemented for all beliefs: for example, one can easily ask about the perceived power of changeable factors (eg, how much does bad weather make it more difficult to be active), but the perceived strength of such factors (eg, how bad is the weather) can only be asked about a specific moment or situation.

While TPB has been criticized on different grounds [18], our main goal here is to avoid leaving out any concept that has important consequences on PA and is feasible to be asked in a self-report questionnaire.

In addition to the determinants included in TPB, fear of negative outcomes is found to be an important barrier to increasing PA in several studies. For example, students in Colombia [19] but not so much in the United States [20] see the threat of injuries as an important barrier to PA. Fatigue brought by exercise, as well as similar items, has been found to be an important barrier in several studies [21-23]. However, acceptance and commitment therapy (ACT) has shown promise as a framework for PA intervention [24]. ACT promotes psychological flexibility by encouraging present-moment awareness and acceptance of internal experiences and acting in line with personal values [25].

In ACT, acceptance means embracing one's experiences without trying to change them [25]. That is, one may engage in a behavior even if one believes it is likely to lead to negative outcomes, provided that one is willing to accept these outcomes. In TPB, conversely, a likely negative outcome should directly lead to a decrease in the intention to perform the behavior.

XSL•FO
RenderX

Current TPB Questionnaires: What Is Missing

Excluding the questionnaires targeting specific health conditions (eg, [26]), there are not many TPB-based PA questionnaires available; two could be applicable in a general population of university students: (1) a questionnaire based on extended TPB by Cheng et al [27] comprising subscales on attitudes (eight items), subjective norms (three items), perceived behavioral control (five items), and intention (three items), and (2) a TPB-based questionnaire by González et al [28], comprising seven items on behavioral beliefs, three on subjective norms, eight on intention or self-efficacy, and two on perceived control. Both questionnaires, however, share a few shortcomings. Many items are paraphrases of the same question: for example, one of the injunctive norm questions states that "the majority of people important to me want me to exercise […]"; in a second item, "want me" is replaced by "think I should," and in a third item by "expect me to" [28]. Having 3 almost identical questions about a topic has an illusory benefit of elevated Cronbach α, but otherwise offers no more information than a single question. Moreover, the construct of attitudes in both questionnaires has a very narrow meaning, essentially, "good" as opposed to "bad," without differentiating between behavioral outcomes and their evaluations [29]. The norm questions in both questionnaires are overly general, referring to "most people I care about." This is a good solution in a short questionnaire, but it does not enable one to differentiate between sources of social influence. Finally, behavioral control in both questionnaires includes only perceived self-efficacy, whereas in TPB, it "[…] is assumed to reflect past experience as well as anticipated impediments and obstacles factors they [i.e., the respondents] believe could make it easier or more difficult for them to perform the behaviour" [15].

In this paper, we take a bottom-up approach to construct a questionnaire of psychological determinants of PA aimed at university students. Besides being a research instrument, a detailed questionnaire of PA determinants can help develop and evaluate interventions, as well as counsel individuals about health behavior. Our primary aim is content coverage; additionally, we need to consider time limitations: respondents tend to be less attentive and cooperative when asked a large number of similar questions [30].

This Study

The present study aimed to create a TPB-based questionnaire for university students, in accordance with good practices of creating such questionnaires [29].

Among the available TPB-based questionnaires, none offers a comprehensive coverage of all TPB constructs. In addition to the TPB constructs, we decided to assess 1 additional facet: acceptance of negative consequences. This is potentially important for complex behaviors that may bring about both positive and negative consequences. In TPB, the behavioral attitude is determined by behavioral beliefs and outcome evaluations; if a consequence is believed to follow a behavior with high certainty (behavioral belief) and this consequence is evaluated negatively (outcome evaluation), it should reduce the motivation to engage in that behavior. If the occurrence of a consequence (such as fatigue) cannot be avoided, the only way to protect one's motivation would thus be to change the evaluation of that consequence. An alternative—accepting a consequence (eg, fatigue) without changing the evaluation—is not explicitly described in TPB. A person who accepts fatigue as a consequence of PA would not avoid PA because of the behavioral belief, but would still avoid other behaviors that bring about similar amounts of fatigue (eg, working too hard or going to sleep too late). One should emphasize that this addition is in principle compatible with the general framework of TPB: one might describe acceptance of negative consequences, for instance, in terms of decision weights to be applied to different perceived consequences—negative consequences that one is willing to accept, matter less to the decision to engage in the behavior.

This study consisted of 3 phases. We started with a qualitative elicitation study to collect ideas and potential items. Thematic analysis [31] of the responses was carried out jointly by the 3 authors of this paper, the results of which were condensed into a draft questionnaire. This was followed by pretesting for clarity and readability. The respondents also provided free-text comments, which were considered in modifying the questionnaire. The modified questionnaire was used in quantitative pretesting. In the following analysis, the aim was to shorten each block of the questionnaire by keeping the most informative items. For this purpose, psychometric analyses, as well as conceptual analysis, were used as tools.

In sum, this study aimed to construct a questionnaire of beliefs and attitudes regarding PA among university students, using thematic analysis of free-response descriptions from the target group as a starting point. The questionnaire will be structured around TPB constructs, complemented by the acceptance of negative outcomes. The questionnaire can be used to assess the PA motivation of young adults when planning interventions (for example, to find out the most salient perceived obstacles to PA), as well as to evaluate the effects of health psychological interventions aiming to increase PA.

Methods

Study Design

This study lasted from January 2021 to February 2022. The stages of questionnaire construction in this study are depicted in Figure 1. The data were collected using Google Docs (phase 1) and LimeSurvey (phases 2 and 3), and analyzed using SPSS (IBM Corp; version 28.0.1.1), and R (R Foundation; version 4.3.3) [32].

XSL•FO
RenderX

Figure 1. Stages of questionnaire construction in this study.

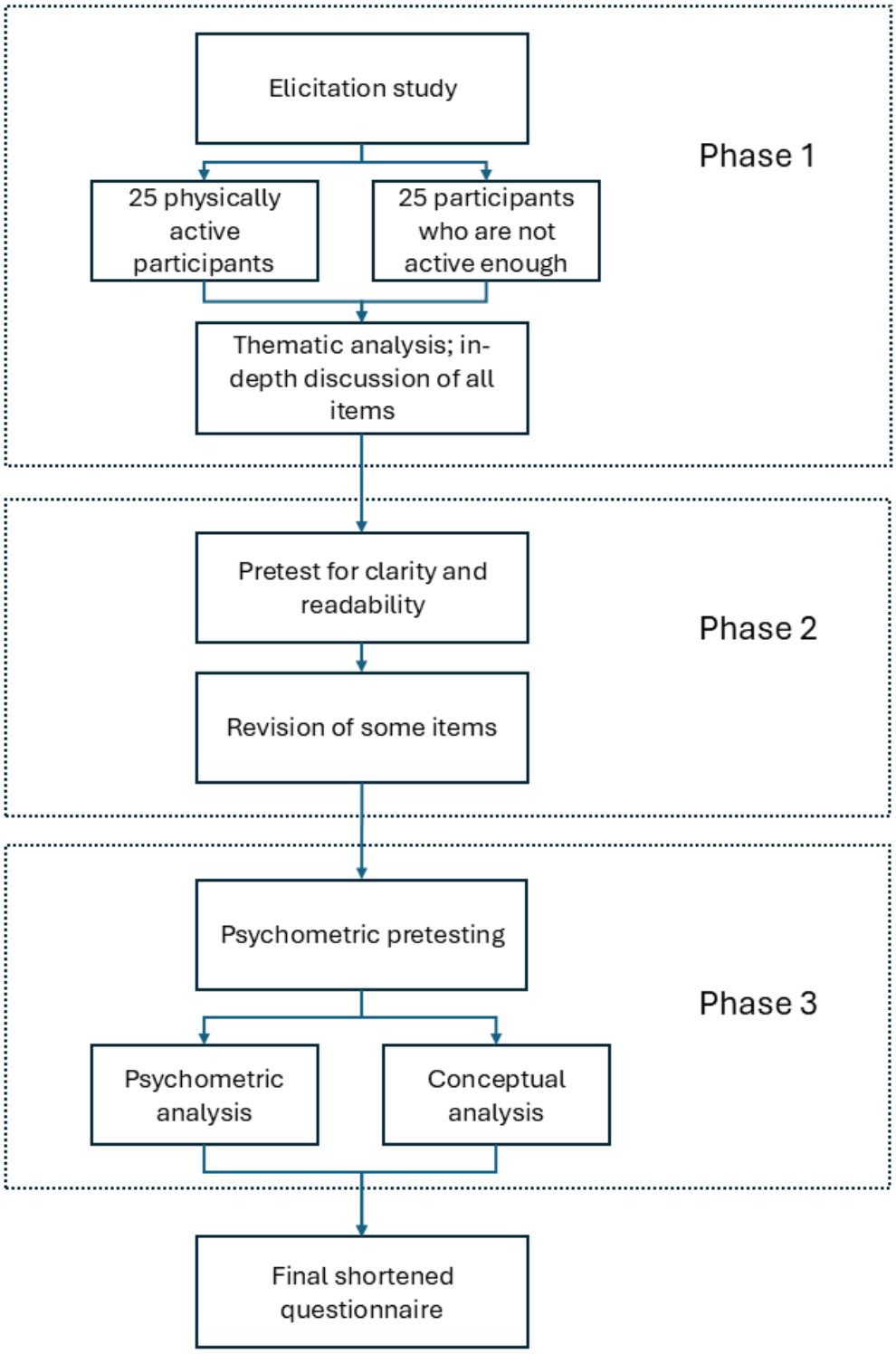

Ethical Considerations

This study was conducted in accordance with the Declaration of Helsinki. As per applicable regulations, ethical approval was not required: according to its statute, the Research Ethics Committee of the University of Tartu assesses research projects aiming to collect personally identifiable data or having potential health impact. The committee does not issue opinions on whether or not a study requires ethical approval; in cases of doubt, the researchers can consult their faculty's research integrity counselor. The present authors' assessment that ethical approval was not required was confirmed post hoc by the research integrity counselor of the Faculty of Social Sciences.

Participants remained anonymous, and no sensitive or potentially identifying information was collected. Invitations, sent via public university mailing lists, briefly explained this study's purpose, estimated completion time, and ensured anonymity and

XSL•FO
RenderX

research-only use of data. Participants could contact the first author with questions. No compensation was provided.

Phase 1: Elicitation Study

Overview

With the elicitation study, we aimed to generate potential items for the questionnaire, using a web-based survey with free-text answers. The answers were thematically analyzed to create a draft questionnaire covering the most important determinants of PA mentioned by the target group. The elicitation survey, as well as analytic steps in thematic analysis and item generation, were based on guidelines by Francis et al [29].

Data Collection

Francis et al [29] recommended a sample size of 25 for the elicitation study. We set the goal of collecting the same number of responses (N=25) from respondents who self-identify as physically active, and from respondents who report not being physically active enough. Data were collected in 2021 by sending an invitation and a link to an anonymous web-based survey to student mailing lists in a public university in Estonia, as well as posting it on social media. Initially (January 15-26), only 25 answers were received, some of which were very brief. Therefore, a second round was carried out (February 19-26), spreading the invitation somewhat more widely, resulting in a higher-than-expected number of answers (n=260). Of the total of 285 respondents, 99 were male, 183 female, 2 nonbinary, and 1 unspecified. The first 25 complete responses from each group (self-identified as active and inactive) were selected for thematic analysis, excluding respondents who provided mostly very short textual responses: for example, “well” in response to the question about one’s feelings after PA, or “none” for the question about the disadvantages of PA. Of the 50 respondents selected for analysis, 35 (70%) were female and 15 were male, with the age ranging from 19 to 41 (mean 24.2, SD 5.0) years.

Questionnaire

The questionnaire (Multimedia Appendix 1), based on Francis et al [29], included the following sections: (1) background: gender, age, university, study area, work, and PA habits; (2) perceived benefits of PA for people in general, importance of these benefits to the respondent (rating scale from 1 to 7); (3) disadvantages of being physically active for people in general. importance thereof for the respondent (rating scale from 1 to 7); (4) benefits of being physically active (for self-admitted active respondents) or hypothetical benefits of becoming more active (for self-admitted low active respondents); (5) feelings during and after PA, feelings about the idea of being physically active, what one enjoys or hates about being physically active; (6) people who think the respondent should be more or less physically active; (7) factors that make it easier or more difficult for the respondent to be physically active; (8) other comments on PA; and (9) whether the respondent wants and thinks they need to be more physically active.

Thematic Analysis

Thematic analysis [31] was conducted to identify recurrent themes in the answers to each question. After omitting empty responses, the text pool contained 572 separate answers (4642 words), some of which were split into 2 or 3 based on content (eg, when a respondent reported being “tired” and “happy” after PA). As a first step, the first author (TDBN) assigned a code to each answer and grouped synonymous or near-synonymous answers. All coding was reviewed by the second (AH) and third (KK) authors; discrepancies were discussed and resolved by consensus. As a second step, items were written for each section of the questionnaire, corresponding to the constructs of TPB. This was done in group sessions with the participation of all authors; every answer was considered as a potential item. During this process, some items were paraphrased to improve clarity; ambiguous or idiosyncratic answers were omitted. In some cases, item categories were combined into a single item: for example, “(feeling) energized” was not included as a separate item because “more energy” was already included; “benefit to health” was not included because there was already “better health.” All such cases were documented with the reasons for excluding a response from the item pool.

We rephrased some responses to ensure broad applicability. For instance, “less sleep disorders” became “better sleep”—so that respondents with normal sleep would still be able to relate to the item. Similarly, “better mood when you have achieved something” was shortened to “better mood” to cover instances where PA itself was not a major achievement. Responses about PA “taking time” were merged with the category of “less time for other activities” (all activities consume time; the issue is how that time impacts other pursuits). Examples of responses that were omitted from the questionnaire: (1) an evolutionary explanation that all organisms tend to conserve energy and thus avoid effort without a clear need, and (2) an explanation of how PA contributes to the oxygenation of the brain. These were judged to be too specific for most respondents.

Phase 2: Pretesting for Clarity and Understanding

The second phase involved pilot-testing the 112-item questionnaire (Multimedia Appendix 2) for clarity and understandability, and collecting improvement recommendations. After each of the 9 question blocks, the respondents were prompted: “We would like to receive your feedback on the questions (numbers of relevant questions) that you answered. Were the questions clear? Do you have any recommendations to improve the questions?” In total, there were 49 completed responses, which were all analyzed. The sample consisted of Estonian university students, 85.7% female, with the age range of 19 to 53 (mean 30.5, SD 10.3) years.

Phase 3: Psychometric Pretesting and Shortening the Questionnaire

Overview

The third phase had two goals: (1) reducing items while retaining the maximum amount of information. As the intention was to measure all TPB concepts, item reduction could be used only within a block of questions, not across all blocks, and (2) to show the psychometric properties of the shortened scales. In some cases, this could not be done directly, as some of the items were combined or generalized in the final scale. For example, the “mother” and “father” items in the perceived norms section were combined into a single item, “parents.” In such cases, we

XSL•FO
RenderX

used the upward-rounded average of the 2 items. Psychometric methods were used in combination with rational analysis of item content (eg, semantic and logical relationships between items). Additionally, somewhat different criteria had to be used in different blocks.

It should also be noted that the TPB is not a psychometric theory: it makes no predictions about the correlations between items within a given domain (eg, behavioral beliefs or injunctive norms), or between the domains. This means that even though one might be tempted to use an overall confirmatory factor analysis to test the psychometric properties of the questionnaire, such an analysis would be meaningless due to the absence of relevant theoretical predictions. Latent variable models are commonly seen as an essential in questionnaire construction, but their use is too often unjustified and solves fewer problems than expected [33,34]. In TPB, the constructs are not treated as latent variables: the theory does not state that, for instance, the different behavioral beliefs reflect a common cause and should be correlated. For this reason, psychometric analyses (which in most cases imply a form of reflective latent variable model) serve a supporting role in this paper. For example, good internal consistency and clean factor structure are not our primary aims; instead, we aim at maximizing the informativeness and content coverage.

Sample and Procedure

Self-report data with the refined questionnaire (Multimedia Appendix 3) were collected from November 2021 to February 2022. Participants were recruited through the several mailing lists of 6 Estonian universities. Of the 797 responses, only complete ones (where the respondent had reached the last page) from people currently studying in a university were used; 349 incomplete responses, as well as one respondent who reported as not to be currently studying, were discarded. The final sample included 447 responses (83% female, 15.4% male, and 1.6% other or unreported gender). The age ranged between 18 and 61 (mean 25.0, SD 7.7) years.

Strategy of Item Reduction

In most questionnaire blocks, the following analytic strategy was used for item reduction (an exception is described below):

1. Parallel analysis was used to test dimensionality, with the goal of making sure that each nonrandom dimension of variability in the item set was represented in the final, shortened questionnaire.
2. To interpret the meaning of these dimensions, exploratory factor analysis (EFA) was used in combination with rational analysis of item content.
3. Internal consistency was computed for each block of items. As the previous steps aim at content breadth rather than homogeneity, a high internal consistency is not necessarily expected, and is not a reflection of the quality of the scale. It may happen that there are several distinct and weakly correlated classes within a domain, each of which contributes to the prediction of the outcome. However, it is also possible that the item set is relatively homogenous, which would justify, for example, combining the items into a summary score. This situation is neither hypothesized in TPB nor necessarily desirable, but it is useful to know and can inform future research. We therefore report internal consistency statistics for the final version of each block of items.

This general scheme had to be adapted or modified in a few cases.

1. In the case of behavioral beliefs, an aim was to cover both positive and negative perceived consequences. For negative consequences, conceptual considerations were given more weight than factor analysis, because these items concerned mostly specific physiological conditions (eg, sweating, fatigue, and injuries). Conversely, positive consequences were mostly subjective, with several items being similar in content.
2. Rational analysis occasionally suggested more than one item to represent a factor. For instance, “being satisfied with myself” was included even though there was already an item (“feeling better”) representing the first factor of behavioral beliefs because being satisfied with oneself refers to cognitive evaluation, which is semantically different from just feeling better (affective reaction).
3. For logically related item blocks, the same items were selected for all blocks, considering the empirical results within each block. First, the section of negative perceived consequences had the same items as the section of acceptance of (negative) consequences. Second, the three norm-related item blocks all contained the same items.
4. In phase 2, several respondents perceived the two control-related sections (control beliefs and power of control) as very similar; this was confirmed by high correlations between corresponding items. It was thus decided to conduct a joint factor analysis of these items and to use the reduced set of items only once in the final questionnaire.

Instead of factor analysis, the individual correlations between selected pairs of items were considered in the norm-related blocks. Where possible, we combined the empirically related items into one. This had to be done considering the conceptual relations between categories: for example, merging the items “mother” and “father” into a more general item “parents” makes sense, whereas merging “mother” and “sister” into “female family members” would be less intuitive and “children” have no common denominator with “colleagues,” even if empirically related.

Strategy for Showing the Psychometric Properties of the Shortened Scales

We used principal axes factoring, which, differently from ML estimation, does not require multivariate normality [35]. The results of Bartlett test, the Kaiser-Meier-Olkin criterion, as well as skewness and kurtosis of variables were examined before analysis. For domains that can arguably be treated as unidimensional, we examined the data with item response theory (IRT; graded response model [36]) to compare the informativeness of items at different levels of the latent variable. Instead of separate threshold (“difficulty”) parameters, we use the IRF-based location index [37] implemented in R’s *mirt* package [38]. Roughly, this index refers to the value of the

XSL•FO
RenderX

latent variable where the expected response of an item is at the scale midpoint. Instead of IRT discrimination parameters, we present factor loadings, which convey essentially the same information but have the advantage of being directly comparable to the EFA factor loadings.

Results

Phase 1: Elicitation Study

Overview

First, we analyzed each question separately. For instance, the question "What do you think are the benefits of being physically active for people in general?" yielded 117 distinct answers (many respondents provided several answers), which were grouped into 22 categories (each with at least 2 mentions); 21 answers could not be coded. For example, the category "better overall health" included answers such as "stronger health," "improves overall health," and "beneficial for health"; the category "better mood" included answers such as "good mood," "better mood," and "improved mood."

Behavioral Beliefs and Acceptance of Negative Consequences

The answers referring to beliefs about the consequences of PA were combined into 34 items. Several perceived consequences of PA (such as better mood and fatigue) appeared across multiple questions. Of these 34 items, 26 were positive consequences of PA, and 8 were negative. Initially, we planned to follow the guidelines by Francis et al [29], asking 2 questions for each item: behavioral beliefs (likelihood of PA leading to this outcome) and outcome evaluations (desirability of the outcome). However, we could not effectively incorporate the desirability ratings for 2 reasons. First, all items were either clearly positive or clearly negative, making half of the response scale seem unnatural to many respondents. Think, for instance, of the degree of "undesirability" of good health or "desirability" of injuries. Francis et al [29] noticed the same issue, suggesting omitting "downright silly" questions. Second, the perceived outcomes of PA seem mostly to be matters of degree rather than all-or-none questions. Pain is undesirable, but mild pain is mildly undesirable, and extreme pain is extremely undesirable. Similarly, a good mood is more desirable if it lasts longer. Modifying the items so that the desirability question would make sense would make the section excessively long and complex, requiring additional pretesting. This might involve adding modifiers (eg, "5 extra years of life" or "serious injuries") or introducing comparisons (eg, desirability of good health compared to good salary). We thus decided to omit the outcome evaluations. As discussed in the Introduction, however, we introduced a section on accepting negative outcomes (8 items).

Subjective Norms

For the question about individuals who think the respondent should be more physically active, we identified 32 distinct answers, categorized into 8 groups (mother, father, sisters, friends, colleagues, partner, parents, and family), excluding some unique answers (myself, nobody, and everyone). For a related question about people who think the respondent should not be more physically active, we found 18 distinct answers, categorized into 6 groups (nobody [the most frequent answer]; mother, father, friend, partner, and parents). Ultimately, 8 items were included in the preliminary questionnaire (mother, father, sisters, brothers, friends, colleagues, fellow students, and relatives [excluding parents and siblings]). The same items were used across all 3 norm-related question blocks.

Control Beliefs

We received 88 distinct responses for factors making PA easier and 85 for factors making it harder, categorized respectively into 19 and 23 groups, and combined into 24 items. Initially, the items included two instructions to assess control beliefs: (1) strength of belief (eg, "to what extent does the time you spend being physically active, depend on weather"), and (2) power of control belief (eg, "how difficult does bad weather make you to be physically active?").

We also added two general PA-related self-efficacy and controllability items: (1) "The amount of time that I spend on being physically active is completely up to me," and (2) "I am confident that I can be more physically active in future."

Phase 2: Pretesting for Clarity and Understanding

We reviewed every specific comment to determine if changes were needed in items or instructions; some changes were made in most sections. For instance, we clarified that questions on perceived consequences referred to the respondent personally, not people in general, following feedback. When questions were unclear, we rephrased them to improve readability. In several cases, no changes were made. For example, some of the behavioral belief items were perceived to be redundant by 1 respondent; no items were omitted, though, because that was the task for the third phase.

Most of the changes were related to readability of the items, potential ambiguity, or compatibility among items, or between instructions and items. For example, the item "relatives (except for parents and siblings)" was changed to "relatives (except for parents, siblings and children)" because the intention was to ask about the "other" relatives, not mentioned in previous items (parents, siblings, and children).

Phase 3: Item Selection and Psychometric Pretesting

Overview

Detailed analyses and the final questionnaire are included in online supplements (Multimedia Appendices 4 and 5). Here, we report only the main considerations for each question block.

Behavioral Beliefs and Acceptance

We ran parallel analysis on 35 behavioral belief items, suggesting 6 factors that we retained in factor analysis (principal axes method, Varimax rotation; details in Multimedia Appendix 5, Section S2).

We aimed to select no more than half of the items, using the following criteria: (1) each factor should be represented by at least one high-loading item; (2) factors with heterogenous content should be represented by more than one item; and (3) preference is to be given to shorter, more general, less ambiguous and universally applicable items.

XSL•FO
RenderX

Based on these criteria, 6 positive and 8 negative behavioral consequences were selected. For example, better mood, feeling better, and feeling good all had high loadings on the first factor, and are all conceptually similar, so we retained only the highest-loading item. Conversely, "self-satisfaction" loaded on both first (n=0.6) and the second (n=0.4) factors and is conceptually distinct from the 3 items mentioned above (referring to the respondent as the object of the feeling), so we kept it. A detailed description of the considerations in item selection is given in Multimedia Appendix 5, Section S2.

An IRT analysis revealed that while the positive behavioral consequences can be treated as a unidimensional latent variable, the selected 6 indicators all had a relatively low value of the location parameter. Therefore, to use the positive beliefs items as a scale, it will be useful to add a few items with a higher location parameter (eg, "better focus"); Table 1 includes 3 such items in addition to the initial 6 selected for content validity.

Table 1. Negative behavioral beliefs and acceptance of negative consequences: factor loadings, IRT[a] parameters, and correlations with PA[b].

	Negative behavioral beliefs			Acceptance of the negative consequences of PA		
	λ[c]	*gl* [d]	*r* [e]	λ	*gl*	*r*
Feeling pain	0.57	1.35	–0.17[f]	0.48	0.96	0.06
Muscle pain	0.63	–0.51	–0.14[f]	0.65	–1.29	0.25[f]
Injury or injuries	0.39	1.77	0.10[g]	0.41	0.90	0.18[f]
Feeling tired	0.68	0.14	–0.21[f]	0.74	–0.65	0.25[f]
Feeling exhausted	0.84	0.45	–0.24[f]	0.64	0.07	0.15[f]
Sweating	0.37	–3.15	0.09	0.68	–1.68	0.20[f]
Feeling uncomfortable	0.74	0.54	–0.33[f]	0.68	–0.34	0.12[g]
Less time for other activities	0.48	–0.01	–0.14[f]	0.53	–0.29	0.43[f]

[a]IRT: item response theory.

[b]PA: physical activity.

[c]λ: item response theory loading.

[d]*gl*: generalized location parameter.

[e]*r*: correlations with self-reported physical activity.

[f]*P*<.01.

[g]*P*<.05.

Negative consequences showed mostly positive and significant correlations, averaging 0.30 (SD 0.13), and ranging from 0.08 (sweating vs less time for other activities) to 0.59 (feeling exhausted vs feeling uncomfortable). Positive consequences were all positively correlated with an average of 0.49 (SD 0.13) and ranged from 0.04 (stable weight vs weight gain) to 0.84 (feeling better vs peace of mind). Notably, items referring to subjective experiences were highly intercorrelated: for example, the correlation between "more energy" and "better mood" (arguably distinct concepts) was 0.69. Cross-correlations between negative and positive consequences were mostly negative, averaging –0.12 (SD 0.12) and ranging from –0.39 (more energy vs feeling pain) to 0.24 (muscle pain vs weight gain due to an increase in muscle mass). See Multimedia Appendix 6 for a full table of correlations.

Finally, acceptance of negative consequences items were all positively and significantly correlated, averaging 0.31 (SD 0.10) and ranging from 0.12 (sweating vs feeling pain) to 0.49 (sweating vs feeling uncomfortable and feeling tired vs feeling exhausted).

For positive behavioral beliefs, the first principal axes factor explained 51.9% of variance, to which the next unrotated factor added 4.1%. The shortened scale closely approximates the full version, with a strong correlation of 0.94 (*P*<.001). Table 2 presents the IRT factor loadings and location parameters.

XSL•FO
RenderX

Table 2. Positive behavioral beliefs: IRT[a] factor loading and generalized location parameter.

	λ[b]	*gl* [c]
Feeling better	0.87	–1.49
Better physical health	0.87	–1.77
Better long-term health[d]	0.82	–0.91
Stable weight	0.49	–1.43
Staying in shape	0.83	–1.39
Self-satisfaction	0.80	–1.50
More energy	0.76	–0.85
Better focus	0.78	–0.77
Greater work capacity	0.84	–0.73

[a]IRT: item response theory.

[b]λ: item response theory factor loading.

[c]*gl*: generalized location parameter.

[d]This item was a generalization of 2 items: preventing health problems and a longer lifespan. The upward-rounded average of these items was used in the analysis.

We omitted "weight gain" from negative behavioral beliefs and acceptance of negative consequences due to its ambiguity, low correlation with other negative items, and limited relevance for typical PA intervention targets. The remaining 8 items had an average correlation of 0.3, with a Cronbach α of 0.78 and McDonald ω of 0.83. The first factor explained 32%, to which the next unrotated factor added 6%.

The average interitem correlation of acceptance items was 0.31, with a Cronbach α of 0.79 and McDonald ω of 0.85. The first factor explained 32.1% of the variance, to which the second factor added 13%.

Correlations between negative behavioral beliefs and acceptance of the corresponding outcome were generally low, ranging from –0.15 to 0.26. Sum scores of negative behavioral beliefs and acceptance had opposite correlations with self-reported PA: –0.22 and 0.33 (both *P*s<.001). Table 1 presents the IRT parameters of negative behavioral beliefs and acceptance items and their correlations with self-reported PA.

Subjective Norms

Correlations between items of injunctive norms and motivation to comply were high, positive, and statistically significant (average *r*=0.62, SD 0.11 and *r*=0.57, SD 0.11, respectively). Correlations between conceptually related item pairs (such as mother vs father, brothers vs sisters, and colleagues vs other students) were somewhat higher than average. However, descriptive norms had generally lower positive correlations (average *r*=0.24, SD 0.15), and not all were statistically significant (Multimedia Appendix 5, Section S3).

Average correlations calculated across 3 matrices are shown in Table 3.

XSL•FO
RenderX

Table 3. Average correlations between subjective norms (mean of 3 correlation matrices: injunctive norms, descriptive norms, and motivation to comply).[a]

	1. My mother	2. My father	3. My brothers	4. My sisters	5. My friends	6. My colleagues	7. Other (university) students	8. My partner	9. My children
1. My mother									
r	1	0.59^{IDM}	0.49^{IDM}	0.54^{IDM}	0.37^{IM}	0.33^{IdM}	0.30^{IM}	0.41^{IdM}	0.36^{IM}
P value	—[b]	—	—	—	—	—	—	—	—
2. My father									
r	0.59^{IDM}	1	0.55^{IDM}	0.46^{IdM}	0.37^{IM}	0.36^{IdM}	0.35^{IdM}	0.39^{IM}	0.34^{IM}
P value	—	—	—	—	—	—	—	—	—
3. My brothers									
r	0.49^{IDM}	0.55^{IDM}	1	0.67^{IDM}	0.48^{IDM}	0.45^{IM}	0.41^{IM}	0.42^{IdM}	0.55^{IDM}
P value	—	—	—	—	—	—	—	—	—
4. My sisters									
r	0.54^{IDM}	0.46^{IdM}	0.67^{IDM}	1	0.49^{IDM}	0.47^{IdM}	0.41^{IdM}	0.50^{IDM}	0.65^{IDM}
P value	—	—	—	—	—	—	—	—	—
5. My friends									
r	0.37^{IM}	0.37^{IM}	0.48^{IDM}	0.49^{IDM}	1	0.60^{IDM}	0.55^{IDM}	0.54^{IDM}	0.64^{IDM}
P value	—	—	—	—	—	—	—	—	—
6. My colleagues									
r	0.33^{IdM}	0.36^{IdM}	0.45^{IM}	0.47^{IdM}	0.60^{IDM}	1	0.66^{IDM}	0.34^{IM}	0.64^{IDM}
P value	—	—	—	—	—	—	—	—	—
7. Other (university) students									
r	0.30^{IM}	0.35^{IdM}	0.41^{IM}	0.41^{IdM}	0.55^{IDM}	0.66^{IDM}	1	0.35^{IM}	0.46^{IdM}
P value	—	—	—	—	—	—	—	—	—
8. My partner									
r	0.41^{IdM}	0.39^{IM}	0.42^{IdM}	0.50^{IDM}	0.54^{IDM}	0.34^{IM}	0.35^{IM}	1	0.55^{IM}
P value	—	—	—	—	—	—	—	—	—
9. My children									
r	0.36^{IM}	0.34^{IM}	0.55^{IDM}	0.65^{IDM}	0.64^{IDM}	0.64^{IDM}	0.46^{IdM}	0.55^{IM}	1
P value	—	—	—	—	—	—	—	—	—
10. Relatives (other than parents, siblings, and children)									
r	0.50^{IDM}	0.48^{IDM}	0.53^{IDM}	0.49^{IDM}	0.50^{IDM}	0.57^{IDM}	0.55^{IDM}	0.38^{IdM}	0.53^{IDM}
P value	—	—	—	—	—	—	—	—	—

[a]Subscript letters denote significance levels (*P* values) for each correlation matrix. I: injunctive norms; D/d: descriptive norms; M: motivation to comply. Lowercase letters: *P*<.05; capital letters: *P*<.01. For example, the superscript IDM means that the correlation was significant at 0.01 level in all 3 matrices; the superscript IM means that the correlation was significant at 0.01 level in injunctive norms and motivation to comply, but not in descriptive norms; the superscript IdM means that the correlation was significant at 0.01 level for injunctive norms and motivation to comply, but at 0.05 level for descriptive norms.

[b]Not available.

For the final questionnaire, we combined items based on their correlations, as well as common sense. For example, the items “my mother” and “my father” had a moderate correlation (n=0.59), making it reasonable to combine them as “parents.” However, despite a moderate correlation between “my friends” and “my children” (n=0.64), combining them would not be appropriate, as they do not form a natural category. We omitted the item “other relatives” due to its heterogeneity and varying

XSL•FO
RenderX

personal meanings, as well as its moderate-to-high correlations with other categories. In sum, the number of items in all 3 sections was reduced from 10 to 6: parents, siblings, friends, colleagues/costudents, partner, and children.

For injunctive norms, the first factor explained 64.4% of the variance, with the second factor adding 7.1%. The average correlation among the 8 items was 0.61, with Cronbach α of 0.91 and McDonald ω of 0.96. For descriptive norms, the first factor explains 32% of the variance, with the second factor adding 11.8%. The average correlation among the items was 0.29, with a Cronbach α of 0.67 and McDonald ω of 0.84. The Kaiser-Meier-Olkin value was 0.56, indicating that the data may not be suitable for factor analysis. For motivation to comply, the first factor explained 58.4% of the variance, whereas the second one added 30.1%. The average correlation among the items was 0.67, with a Cronbach α of 0.89 and a McDonald ω of 0.96. Table 4 presents the IRT factor loadings and location parameters for these constructs.

Table 4. Perceived norms and motivation to comply: factor loadings and IRT[a] parameters.

	Injunctive norms		Descriptive norms		Motivation to comply	
	λ[b]	*gl* [c]	λ	*gl*	λ	*gl*
Parents	0.74	0.51	0.27	–0.16	0.86	–0.10
Siblings	0.86	0.68	0.44	–0.67	0.88	0.15
Friends	0.91	0.68	0.67	–0.50	0.83	–0.17
Colleagues and fellow students	0.94	0.89	0.55	–0.32	0.72	0.85
Partner	0.80	0.30	0.42	–0.60	0.83	–0.98
Children	0.86	0.56	0.83	–0.95	0.82	–0.64

[a]IRT: item response theory.

[b]λ: item response theory loading.

[c]*gl*: generalized location parameter.

Control Beliefs and Self-Efficacy

The 2 self-efficacy and controllability items were moderately correlated (r=0.38, P<.01); both were included in the final questionnaire.

Parallel analysis of 48 control belief and perceived power items suggested 12 factors. Based on this, a factor analysis was carried out using 12 factors (principal axes with Varimax rotation; details Multimedia Appendix 5, Section S4). Most factors were clearly identifiable and interpretable; it was therefore decided to keep at least 1 item for each factor. A few items were rewritten in a more general way: for example, instead of the "gym," the item in the final version refers to the "place to be physically active" (subsuming, eg, gyms, parks, and tennis courts). We sometimes retained more than 1 item per factor: for example, the items "weather" and "the place where I live" loaded on the same factor but refer to clearly distinct determinants (one is changeable and not under the person's control, whereas the second is rather stable, and depends on the respondent's own choices). Two items referring to safety ("living in an unsafe environment" and "whether the environment is safe") were at best weakly related to any of the factors; however, it is known that safety concerns can be an important obstacle to PA [39]. It was therefore decided to keep a paraphrase of the initial items ("the safety of the place where I live") in the final questionnaire. The paraphrase was intended to make the item less ambiguous, referring to the safety of one's neighborhood, and making other interpretations of the original items (eg, air pollution or geopolitical concerns) less likely.

In the shortened version, parallel analysis suggested 4 factors. Unlike in other domains, the factors beyond the first one were of substantial size (the first 4 Promax-rotated EFA factors explained, respectively, 14%, 7%, 11%, and 8% of the variance). It is therefore not meaningful to treat these responses as indicators of a single latent variable, even though the correlations were mostly positive (average 0.23, ranging from –0.16 to 0.82). The factors had 1 to 6 indicators (loadings above 0.4). Table 5 presents the factor loadings from a 4D graded response model analysis.

XSL•FO
RenderX

Table 5. Control beliefs: IRT[a] loadings and factor correlations (Promax rotation)[b].

Item	F1[c]	F2	F3	F4
The weather	0.40	0.18	–0.28	0.16
How I am feeling	–0.01	0.68	0.07	0.15
My health	–0.14	0.74	0.07	–0.10
My planning skills	–0.03	0.46	0.13	–0.04
How fast I see the results	0.63	0.02	–0.17	0.04
Whether I have time or not	–0.04	0.07	0.08	0.97
Whether there is a gym nearby	0.10	0.04	0.90	0.07
Whether the gym is open or not	0.06	0.04	0.98	0.01
Whether I have a car or not [...]	0.62	–0.24	0.19	–0.02
Having the necessary equipment	0.62	0.00	0.19	–0.05
My financial situation	0.45	0.10	0.26	–0.06
Where I live	0.66	–0.03	–0.02	–0.01
Whether the environment is safe	0.29	0.38	–0.22	–0.07
Whether I have someone to do it with or not [...]	0.41	0.05	0.03	–0.03
Factor correlations				
F1				
r	1.00	0.53	0.34	0.30
P value	__[d]	—	—	—
F2				
r	0.53	1.00	–0.09	0.43
P value	—	—	—	—
F3				
r	0.34	–0.09	1.00	–0.06
P value	—	—	—	—
F4				
r	0.30	0.43	–0.06	1.00
P value	—	—	—	—

[a]IRT: item response theory.
[b]Some items were shortened for presentation.
[c]F: factor.
[d]Not available.

Behavioral Intentions

Two items referring to behavioral intentions correlated r=0.50 (P<.01) and were both included in the final questionnaire.

Predicting Self-Reported PA

Table 6 shows the correlations between the item bundles and self-reported PA.

XSL•FO
RenderX

Table 6. Correlations of item bundles with self-reported physical activity[a].

	Full		Shortened	
	r	*P* value	*r*	*P* value
Positive beliefs	0.29	<.001	0.26	<.001
Positive beliefs (2)	N/A[b]	N/A	0.29	<.001
Negative beliefs	–0.18	<.001	–0.22	<.001
Acceptance	0.32	<.001	0.32	<.001
Injunctive norms	–0.40	<.001	–0.39	<.001
Descriptive norms	0.11	.02	0.15	.002
Motivation to comply	0	.98	0	.94
Control beliefs (4 EFA[c] factors and multivariate prediction)	0.48	.004	0.48	<.001

[a]For positive behavioral beliefs, 2 shortened versions were considered: 6 items selected to optimize content validity (first line) and 9 items, including the aforementioned 6, and adding 3 with the highest location parameter from IRT. For perceived norms, person means across all completed items were used (eg, if the respondent does not have children, then the mean of all other items was used). For control beliefs, 4 factor scores were used to predict the criterion, and the *r* refers thus to the multivariate model (square root of R^2).

[b]N/A: not applicable.

[c]EFA: exploratory factor analysis.

Discussion

Principal Findings

This study aimed to create a questionnaire based on the TPB for university students and evaluate its psychometric properties. Additionally, the questionnaire is meant for evaluating the effectiveness of health psychological PA interventions in university students and young adults.

This study consisted of three different phases: (1) an elicitation study for item generation; (2) pretesting for clarity and readability; and (3) item reduction and evaluation of psychometric properties. This resulted in a questionnaire consisting of the following sections: behavioral beliefs regarding PA (14 items), acceptance of negative consequences of PA (6 items), injunctive norms, descriptive norms, and motivation to comply (6 items each), self-efficacy and subjective controllability (2 items), control beliefs (14 items), and behavioral intention (2 items).

Our focus was on covering as many of the different aspects of each construct as possible. This led us, for example, to reduce the number of positive behavioral beliefs more than the number of negative behavioral beliefs. Of the 13 items loading over 0.5 on the first rotated factor, most were related to subjective experiences, and although many of them (eg, better mood and reduced stress) are distinct concepts, the correlations between any pairs of items were considerably higher than the average correlation between the negative perceived consequences.

Our shortened questionnaire is thus usable in situations where the goal is to get a comprehensive overview of the perceived determinants of PA behavior. If the goal is to assess some of the determinants (eg, perceived positive consequences of PA) in more detail or with a greater level of accuracy, one might select more items from these domains.

Before proceeding to a more general discussion, we briefly summarize the most important considerations regarding the 3 phases of this study.

Phase 1: Elicitation Study

We used thematic analysis to identify recurrent themes mentioned in free-text responses. In thematic analysis, it is important that the authors acknowledge the role of their own theoretical position [31]; we used TPB as a starting point, but used common sense and consistency where appropriate. For example, “brothers” were added to injunctive norms despite not being mentioned in the elicitation study; the same items were used for injunctive and descriptive norms, as well as negative perceived consequences and acceptance. In making judgments about which responses to combine, or whether or not to include an item in the questionnaire, every such item was discussed between the 3 authors until consensus was reached. For example, the answer “counterbalance to the sedentary lifestyle” was deemed ambiguous because (1) it was interpreted as a reason for engaging in PA rather than a benefit of it, and (2) the specific positive effect of this counterbalancing was unclear (eg, whether the respondent meant a positive health effect, or reduction in the negative effect of prolonged sitting, or a feeling of self-satisfaction derived from doing something good for one’s health, etc). Finally, some subjectivity was inevitable in the judgment of thematic saturation. We considered that the analysis of 50 responses yielded more distinct items in each category than could be reasonably included in a final questionnaire (eg, 34 behavioral belief items, 20 of which were based on more than one free-text answer, and the remaining 14 were in most cases closely related variants or nuances of the more frequently mentioned consequences of PA). However, the frequent responses (mentioned at least 3 times) were almost all included in the draft questionnaire, whereas infrequent ones (mentioned only once) were more often judged to be too unusual or ambiguous (with a few exceptions: eg, pain and avoiding weight

XSL•FO
RenderX

gain as consequences of PA, and sisters and colleagues as sources of social influence were mentioned only once or twice in the elicitation study). It was thus decided not to apply thematic analysis to the rest of the responses, as it was unlikely to result in new important themes that could realistically be included in the questionnaire.

Phase 2: Pretesting for Clarity and Readability

This phase aimed to check the clarity and readability of the whole questionnaire, including the instructions, and relationships between item blocks. Participants' comments often highlighted issues with interitem relationships (eg, redundancy within item blocks, or comparing 1 item block with another) or instructions (eg, linguistic compatibility between an item and the instructions). The resulting changes were focused on clarifying the content of the items and instructions, thus reducing the error variance due to different understandings of the items.

Phase 3: Shortening the Questionnaire While Preserving Content Validity

In this phase, we used psychometric methods to reduce the number of items in the questionnaire, prioritizing content validity over other psychometric considerations. While none of the psychometric properties were an aim in themselves, it should be noted that the resulting scales do have a moderate to high degree of internal consistency, which may justify the use of sum-scores when this is needed (except for control beliefs).

We presented evidence regarding construct validity and reliability of the shortened scales. Item bundles of positive and negative behavioral beliefs, acceptance of unfavorable outcomes, injunctive and descriptive norms, as well as motivation to comply, can be meaningfully used as sum scores, even if not always strictly unidimensional. Notably, these shortened item bundles match or even exceed the unshortened versions in predicting self-reported PA.

Comparison With Previous Studies

For behavioral beliefs, the results in our elicitation study largely mirrored earlier research [13], showing commonly cited advantages such as improvements in physical and psychological health, weight control, increased energy, stress relief, and relaxation. There were, however, a few differences. First, the theme of "improved daily functioning" emerged in more than 2/3 studies reviewed by Downs and Hausenblas [13] but was represented by a single item ("greater working capacity") in our study. Similarly, improvements in social life (eg, making friends, meeting new people, and socializing) were often mentioned in other studies, but were not found in our study. This is likely a characteristic of the student sample: the participants have generally no major issues with daily functioning, and live in a varied social environment with plenty of opportunities to make friends and meet new people, making these less salient as advantages of PA. If, however, one wants to use the questionnaire in a more general sample, one should consider either a new elicitation study, at least for behavioral beliefs. Several perceived consequences of PA mentioned in the review by Downs and Hausenblas [13] were highly correlated in our empirical pretest (phase 3): for example, improved mood, feeling good, having more energy, and decreased stress, had all high loadings on the same factor, and were represented by a single item in our shortened questionnaire (feeling better). Some researchers might be interested in assessing more specific beliefs about subjective consequences (related to mood, focus, energy, etc) of PA and might thus need to include more items from factor 1 of behavioral beliefs (Multimedia Appendix 5 and Table 4).

The most frequent disadvantages of PA in earlier research [13] were health issues, mostly related to pain and injury. This and other frequent categories (fatigue and lack of time) were also represented in our study. The notion of PA being dangerous (eg, fear of another heart attack) did not emerge in our sample, probably because of the low prevalence of any major health issues. Additionally, the everyday inconveniences (eg, interference with family routine or changes in meal times) did not emerge in our study as a separate theme but would likely be more prevalent in other age groups.

For normative beliefs, most elicitation studies reviewed by Downs and Hausenblas [13] reported family members and friends, as well as health care professionals. The latter category (and several others that emerged in some elicitation studies: boss, teacher, coach, other exercisers, church members, media, and other people with cancer) did not emerge in our study, and, again, might be important for other target groups.

Downs and Hausenblas [13] identified health issues, inconvenience, lack of motivation or energy, lack of time, and lack of social support as common categories of control beliefs. Feelings of motivation and energy were loaded, in our phase 3, on the same factor as mood, and were thus not included in the shortened questionnaire.

Limitations and Directions for Future Research

The strengths of our approach include a reasonably large sample, as well as having a specific target group, allowing us to compile an extensive list of PA determinants. The list is, however, not necessarily comprehensive for other target groups. Due to the age, lifestyle, and typically good health condition of students, some themes did not emerge in the elicitation study; for example, medical professionals were not mentioned as relevant for injunctive or descriptive norms. As another example, items related to daily functioning, and potential dangers related to vigorous PA (eg, experiencing a heart attack) were not mentioned in the elicitation phase. Finally, inconveniences related to everyday routine and family routine were not mentioned and thus not included in the questionnaire. These themes would be likely to be elicited in different (eg, older) target groups.

Some researchers may be interested in making more fine-grained distinctions among the positive consequences of PA [40]. In these cases, we would recommend complementing the shortened questionnaire with extra items based on either our initial questionnaire or the extant literature.

In retrospect, our initial instructions for the sections on control beliefs and strength of control were too similar. Control beliefs refer to a factor's influence on PA (eg, weather's impact on exercise), while strength of control refers to the difficulty posed by an unfavorable factor (eg, bad weather making PA difficult).

XSL•FO
RenderX

In theory, these are different questions: for example, one may perceive the difficulty of exercising in bad weather but still do it, resulting in weather having little influence on PA. Questions about control beliefs are typically framed as a likelihood of a favorable or unfavorable factor. This would have been cumbersome in our case: for instance, without going into specifics, it is not possible to say how likely "bad weather" is (there are various degrees and kinds of "bad weather"). After the quantitative pretesting, it became evident that asking these different questions would not make sense, but also that it is relatively easy to ask about the perceived power of the factors influencing one's PA, but more difficult to ask about the perceived likelihood of these factors. This is an area that needs further investigation, as there is substantial evidence about the role of perceived behavioral control in, for example, moderating the intention-behavior relationship [41], but in most studies, the relative contribution of control belief strength and power of control factors is not investigated (a product of these 2 terms is used; O'Sullivan et al [42] have described the methodological problems that this practice creates).

Future studies could explore alternative framings of the behavioral control questions: for example, self-efficacy of overcoming a specific barrier (such as exercising in bad weather), self-efficacy of creating a favorable situation, temptation to use a given obstacle as an excuse (eg, not exercising and referring to bad weather), and perhaps also self-efficacy of overcoming this temptation. It is an empirical question whether these alternative framings improve the prediction of intention or behavior, and whether the respondents would perceive these as distinct questions.

Another potentially important avenue of research is expanding the concept of intention in TPB. For example, in self-determination theory [43], there is a crucial distinction between intrinsic and extrinsic motivation, as well as a taxonomy of psychological needs that may underlie different intentions.

Even though several potentially important themes are missing, our shortened questionnaire is considerably broader in content coverage than most of the readily available alternatives. This, on the one hand, gives the researcher more confidence that important determinants are included; however, long questionnaires may be a source of frustration and boredom to participants [30] and may not be necessary for each purpose. Therefore, an even shorter inventory might be useful—but we would suggest constructing it empirically, looking for the most important predictors of behavior, or the strongest moderators of the intention-behavior relationship. An even shorter questionnaire is, thus, beyond the scope of this study.

In this study, we aimed at creating a comprehensive inventory, focusing thus on content validity [44], or, more specifically, on readability, relevance, and content coverage of the items. Among the other aspects of validity to be investigated in future studies, we would prefer to highlight 2. First, if a measure is to be useful in the context of intervention studies, it needs to be sensitive to change in attitudes. One way to investigate this would be a combination of qualitative and quantitative methods to determine whether the questionnaire captures the change in attitudes that people perceive as important. Second, predictive and construct validity need to be investigated: for example, the degree to which the measure predicts behavior or behavioral intention.

The present questionnaire was developed in Estonian; the English translation is provided only for reference and needs additional pretesting before use. Additionally, the cultural variability of PA-related attitudes and beliefs was not addressed in the present research but needs attention in future studies. Besides culture, other potential sources of influence include age, gender, health and weight status, and socioeconomic background. This study was conducted with university students; even though education in public universities in Estonia is mostly free, socioeconomic disparities exist [45], making it likely that persons with the lowest socioeconomic status were excluded from the sample. An adaptation of the present questionnaire to another cultural, demographic, or socioeconomic context should ideally include a new elicitation study, possibly using the present questionnaire as a starting point, focussing on new topics that are not yet included.

Finally, some of the TPB constructs were not assessed in this study: evaluation of negative outcomes was replaced by acceptability, and evaluation of positive outcomes was omitted. We made these decisions because the outcomes in question (eg, good mood or pain) come in degrees (a very good mood is more desirable than just a little bit good mood, etc) and a general desirability or undesirability question does not make sense. Asking about the desirability of specific degrees of such outcomes would make the questionnaire longer; moreover, the degrees of subjective states are, by definition, also subjective. We could, however, ask about the acceptability of negative outcomes, which is related but not identical to the concept of outcome evaluations, and also related to the TPB construct of controllability: if a given outcome is acceptable, it has therefore less control over the respondent's behavior.

This decision makes it impossible to compute the combined TPB scores for attitudes. We should, however, note that the use of product scores (eg, likelihood of an outcome multiplied by its desirability) is questionable on methodological grounds, creating a so-called "expectancy-value muddle" [42] and blurring the contribution of the individual components: the multiplication makes good sense conceptually but may fail for measurement reasons. First, it is a good practice to include interaction (product) terms in a regression analysis only after including the corresponding 2-order terms [46]. Second, subjective probabilities are essentially different from actual probabilities [47], and the question of whether they can be meaningfully multiplied needs additional research.

Conclusions

In this paper, we used a bottom-up approach to construct a questionnaire of determinants of PA in university students. To develop the initial pool of items retrieved from the elicitation study into a questionnaire, we used a combined approach based on free-response comments from the target group, factor analysis, and expert discussions considering match with theoretical constructs, as well as logical and common-sense considerations, with the principal aim of reducing redundancy while retaining informativeness. The shortened questionnaire

XSL•FO
RenderX

includes a reasonably comprehensive set of determinants as described by TPB (with a few justified omissions and additions) for this target group. For other target groups, additional factors might need to be considered (eg, medical experts as a source of injunctive norms or heart attack as a possible adverse result of PA). Further research is needed to investigate the predictive and construct validity of the questionnaire, as well as its sensitivity to change in beliefs and attitudes. Finally, some types of studies may need a shorter questionnaire, or more nuanced measurement of some specific determinants (eg, more differentiation in mood states believed to result from PA).

Acknowledgments

When writing this paper, generative artificial intelligence (Gemini) was consulted on English usage and for suggestions on improving the conciseness of expression. This study was supported by the Estonian Research Council (PRG1656 to KK). The funder had no involvement in this study's design, data collection, analysis, interpretation, or the writing of this paper.

Data Availability

Anonymous data from the third phase are available online [48]. The data from the first and second phases are available from the authors upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Phase 1 (elicitiation study) questionnaire.
[DOCX File , 17 KB - formative_v9i1e69094_app1.docx]

Multimedia Appendix 2
Initial version of the questionnaire.
[DOCX File , 31 KB - formative_v9i1e69094_app2.docx]

Multimedia Appendix 3
Questionnaire modified according to the results of phase 2 (pretesting for clarity and understandability).
[DOCX File , 27 KB - formative_v9i1e69094_app3.docx]

Multimedia Appendix 4
Final shortened questionnaire.
[DOCX File , 29 KB - formative_v9i1e69094_app4.docx]

Multimedia Appendix 5
Descriptive statistics and factor analysis results.
[DOCX File , 98 KB - formative_v9i1e69094_app5.docx]

Multimedia Appendix 6
Correlations among positive and negative perceived consequences of physical activity.
[XLSX File (Microsoft Excel File), 35 KB - formative_v9i1e69094_app6.xlsx]

Multimedia Appendix 7
Shortened questionnaire: assumption checks and psychometric analyses.
[DOCX File , 169 KB - formative_v9i1e69094_app7.docx]

References

1. Global Status Report on Physical Activity 2022. World Health Organization. 2022. URL: https://www.who.int/publications/i/item/9789240059153 [accessed 2025-09-17]
2. Cihan BB, Bozdağ B, Var L. Examination of physical activity and life quality levels of university students in terms of related factors. JEL 2018;8(1):120. [doi: 10.5539/jel.v8n1p120]

XSL•FO
RenderX

3. Murphy M, Carlin A, Woods C, Nevill A, MacDonncha C, Ferguson K, et al. Active students are healthier and happier than their inactive peers: the results of a large representative cross-sectional study of university students in Ireland. J Phys Act Health 2018;15(10):737-746 [FREE Full text] [doi: 10.1123/jpah.2017-0432] [Medline: 30119614]
4. Snedden TR, Scerpella J, Kliethermes SA, Norman RS, Blyholder L, Sanfilippo J, et al. Sport and physical activity level impacts health-related quality of life among collegiate students. Am J Health Promot 2019;33(5):675-682 [FREE Full text] [doi: 10.1177/0890117118817715] [Medline: 30586999]
5. Lipošek S, Planinšec J, Leskošek B, Pajtler A. Physical activity of university students and its relation to physical fitness and academic success. Ann Kin 2018;9(2):89-104. [doi: 10.35469/ak.2018.171]
6. Edelmann D, Pfirrmann D, Heller S, Dietz P, Reichel JL, Werner AM, et al. Physical activity and sedentary behavior in university students-the role of gender, age, field of study, targeted degree, and study semester. Front Public Health 2022;10:821703 [FREE Full text] [doi: 10.3389/fpubh.2022.821703] [Medline: 35784227]
7. Bednarek J, Pomykała S, Bigosińska M, Szyguła Z. Physical activity of polish and Turkish university students as assessed by IPAQ. Cent Eur J Sport Sci Med 2016;16:13-22. [doi: 10.18276/cej.2016.4-02]
8. Castro O, Bennie J, Vergeer I, Bosselut G, Biddle SJH. How sedentary are university students? A systematic review and meta-analysis. Prev Sci 2020;21(3):332-343. [doi: 10.1007/s11121-020-01093-8] [Medline: 31975312]
9. Mazuera C, Arteaga AP, Manrique A. Perceived barriers to physical activity practice in undergraduate students. Int J Educ Inf Technol 2018;12:163-170.
10. Silva RMF, Mendonça CR, Azevedo VD, Raoof Memon A, Noll PRES, Noll M. Barriers to high school and university students' physical activity: a systematic review. PLOS One 2022;17(4):e0265913 [FREE Full text] [doi: 10.1371/journal.pone.0265913] [Medline: 35377905]
11. Espada M, Romero-Parra N, Bores-García D, Delfa-De La Morena JM. Gender differences in university students' levels of physical activity and motivations to engage in physical activity. Educ Sci 2023;13(4):340. [doi: 10.3390/educsci13040340]
12. Shafieinia M, Hidarnia A, Kazemnejad A, Rajabi R. Effects of a theory based intervention on physical activity among female employees: a quasi-experimental study. Asian J Sports Med 2016;7(2):e31534 [FREE Full text] [doi: 10.5812/asjsm.31534] [Medline: 27625759]
13. Downs DS, Hausenblas HA. Elicitation studies and the theory of planned behavior: a systematic review of exercise beliefs. Psychol Sport Exercise 2005;6(1):1-31. [doi: 10.1016/j.psychsport.2003.08.001]
14. Ajzen I. The theory of planned behavior: frequently asked questions. Human Behav and Emerg Tech 2020;2(4):314-324. [doi: 10.1002/hbe2.195]
15. Ajzen I. The theory of planned behavior. Organ Behav Hum Decis Process 1991;50(2):179-211. [doi: 10.1016/0749-5978(91)90020-T]
16. Montano D, Kasprzyk D. Theory of reasoned action, theory of planned behavior, and the integrated behavioral model. In: Health Behavior: Theory, Research, and Practice. San Francisco, CA: Jossey-Bass; 2015:95-124.
17. Ajzen I. Perceived behavioral control, self-efficacy, locus of control, and the theory of planned behavior. J Appl Soc Pyschol 2002;32(4):665-683. [doi: 10.1111/j.1559-1816.2002.tb00236.x]
18. Sniehotta FF, Presseau J, Araújo-Soares V. Time to retire the theory of planned behaviour. Health Psychol Rev 2014;8(1):1-7. [doi: 10.1080/17437199.2013.869710] [Medline: 25053004]
19. Ramírez-Vélez R, Tordecilla-Sanders A, Laverde D, Hernández-Novoa JG, Ríos M, Rubio F, et al. The prevalence of barriers for Colombian college students engaging in physical activity. Nutr Hosp 2014;31(2):858-865 [FREE Full text] [doi: 10.3305/nh.2015.31.2.7737] [Medline: 25617574]
20. Kulavic K, Hultquist CN, McLester JR. A comparison of motivational factors and barriers to physical activity among traditional versus nontraditional college students. J Am Coll Health 2013;61(2):60-66. [doi: 10.1080/07448481.2012.753890] [Medline: 23409855]
21. Chung-Yan Chan J. Psychological determinants of exercise behavior of nursing students. Contemp Nurse 2014;49(1):60-67. [doi: 10.1080/10376178.2014.11081954]
22. Blake H, Stanulewicz N, Mcgill F. Predictors of physical activity and barriers to exercise in nursing and medical students. J Adv Nurs 2017;73(4):917-929. [doi: 10.1111/jan.13181] [Medline: 27731886]
23. Grubbs L, Carter J. The relationship of perceived benefits and barriers to reported exercise behaviors in college undergraduates. Fam Community Health 2002;25(2):76-84. [doi: 10.1097/00003727-200207000-00009] [Medline: 12010117]
24. Pears S, Sutton S. Effectiveness of acceptance and commitment therapy (ACT) interventions for promoting physical activity: a systematic review and meta-analysis. Health Psychol Rev 2021;15(1):159-184. [doi: 10.1080/17437199.2020.1727759] [Medline: 32036768]
25. Hayes SC, Luoma JB, Bond FW, Masuda A, Lillis J. Acceptance and commitment therapy: model, processes and outcomes. Behav Res Ther 2006;44(1):1-25. [doi: 10.1016/j.brat.2005.06.006] [Medline: 16300724]
26. Mulero-Portela AL, Santaella CLC, Gómez CC. Development of a questionnaire to assess the determinants of exercise among breast cancer survivors in Puerto Rico. Physiother Theory Pract 2013;29(5):379-392 [FREE Full text] [doi: 10.3109/09593985.2012.746763] [Medline: 23244037]
27. Cheng OY, Yam CLY, Cheung NS, Lee PLP, Ngai MC, Lin C. Extended theory of planned behavior on eating and physical activity. Am J Health Behav 2019;43(3):569-581. [doi: 10.5993/AJHB.43.3.11] [Medline: 31046887]

XSL•FO
RenderX

28. González ST, López MCN, Marcos YQ, Rodríguez-Marín J. Development and validation of the theory of planned behavior questionnaire in physical activity. Span J Psychol 2012;15(2):801-816. [doi: 10.5209/rev_sjop.2012.v15.n2.38892] [Medline: 22774454]
29. Francis J, Eccles MP, Johnston M, Walker AE, Grimshaw JM, Foy R, et al. Constructing questionnaires based on the theory of planned behaviour: a manual for health services researchers. City St George's University of London. Newcastle, Australia: University of Newcastle; 2004. URL: https://openaccess.city.ac.uk/id/eprint/1735/ [accessed 2025-09-17]
30. Konstabel K, Lönnqvist J, Walkowitz G, Konstabel K, Verkasalo M. The 'Short Five' (S5): measuring personality traits using comprehensive single items. Eur J Pers 2012;26(1):13-29. [doi: 10.1002/per.813]
31. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
32. R: a language and environment for statistical computing. R Foundation for Statistical Computing. 2024. URL: https://www.R-project.org/ [accessed 2025-09-17]
33. Cooper C. What is wrong with individual differences research? Pers Individ Differ 2024;221:112550. [doi: 10.1016/j.paid.2024.112550]
34. Revelle W. The seductive beauty of latent variable models: or why I don't believe in the Easter Bunny. Pers Individ Differ 2024;221:112552. [doi: 10.1016/j.paid.2024.112552]
35. Fabrigar LR, Wegener DT. Exploratory Factor Analysis. Oxford, United Kingdom: Oxford University Press; 2012.
36. Samejima F. Estimation of latent ability using a response pattern of graded scores. Psychometrika 1969;34(S1):1-97. [doi: 10.1007/bf03372160]
37. Ali US, Chang H, Anderson CJ. Location indices for ordinal polytomous items based on item response theory. ETS Res Rep Ser 2015;2015(2):1-13. [doi: 10.1002/ets2.12065]
38. Chalmers RP. mirt: a multidimensional item response theory package for the R environment. J Stat Soft 2012;48(6):1-29. [doi: 10.18637/jss.v048.i06]
39. Rachele JN, Ghani F, Loh VH, Brown WJ, Turrell G. Associations between physical activity and the neighbourhood social environment: baseline results from the HABITAT multilevel study. Prev Med 2016;93:219-225. [doi: 10.1016/j.ypmed.2016.06.034] [Medline: 27370165]
40. Thayer RE, Newman JR, McClain TM. Self-regulation of mood: strategies for changing a bad mood, raising energy, and reducing tension. J Pers Soc Psychol 1994;67(5):910-925. [doi: 10.1037//0022-3514.67.5.910] [Medline: 7983582]
41. Hagger MS, Cheung MW, Ajzen I, Hamilton K. Perceived behavioral control moderating effects in the theory of planned behavior: a meta-analysis. Health Psychol 2022;41(2):155-167. [doi: 10.1037/hea0001153] [Medline: 35143225]
42. O' Sullivan B, McGee H, Keegan O. Comparing solutions to the expectancy-value muddle in the theory of planned behaviour. Br J Health Psychol 2008;13(Pt 4):789-802. [doi: 10.1348/135910708X278306] [Medline: 18215337]
43. Ryan RM, Deci EL. Intrinsic and extrinsic motivation from a self-determination theory perspective: definitions, theory, practices, and future directions. Contemp Educ Psychol 2020;61:101860. [doi: 10.1016/j.cedpsych.2020.101860]
44. Haynes SN, Richard DCS, Kubany ES. Content validity in psychological assessment: a functional approach to concepts and methods. Psychol Assess 1995;7(3):238-247. [doi: 10.1037/1040-3590.7.3.238]
45. Põder K, Lauri T. The paradox of state-funded higher education: does the winner still take it all? Educ Sci 2021;11(12):812. [doi: 10.3390/educsci11120812]
46. Aiken L, West S, Reno R. Multiple Regression: Testing and Interpreting Interactions (Reprinted). Thousand Oak, California: SAGE; 2010.
47. Kahneman D, Tversky A. Subjective probability: a judgment of representativeness. Cognit Psychol 1972;3(3):430-454. [doi: 10.1016/0010-0285(72)90016-3]
48. OSF Home. URL: https://osf.io/hqgxn/ [accessed 2025-09-17]

Abbreviations

ACT: acceptance and commitment therapy
EFA: exploratory factor analysis
IRT: item response theory
PA: physical activity
TPB: theory of planned behavior

XSL•FO
RenderX

Edited by A Mavragani; submitted 21.11.24; peer-reviewed by H Tilga, T Obuobisa-Darko; comments to author 10.01.25; revised version received 17.04.25; accepted 13.08.25; published 16.10.25.

Please cite as:
Noordover TDB, Hannus A, Konstabel K
Toward Comprehensive Assessment of Beliefs and Attitudes Related to Physical Activity in Young Adults: Pilot Study
JMIR Form Res 2025;9:e69094
URL: https://formative.jmir.org/2025/1/e69094
doi:10.2196/69094
PMID:

©Theodorus D B Noordover, Aave Hannus, Kenn Konstabel. Originally published in JMIR Formative Research (https://formative.jmir.org), 16.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

An mHealth App and System Architecture for Respiratory Disease Management: Design Principles, Tool Development, and Pilot Usability Study

Andrew Chao[1,2], MSc; Lisa Martignetti[1], MSc; René Groh[3], PhD; Andreas M Kist[3], PhD; Nicole YK Li-Jessen[1,2,4,5,6], PhD

[1]School of Communication Sciences and Disorders, Faculty of Medicine and Health Sciences, McGill University, Montreal, QC, Canada
[2]Department of Biomedical Engineering, McGill University, Montreal, QC, Canada
[3]Department of Artificial Intelligence in Biomedical Engineering, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
[4]Department of Otolaryngology—Head and Neck Surgery, McGill University, Montreal, QC, Canada
[5]Translational Research in Respiratory Diseases Program, Research Institute of McGill University Health Center, Montreal, QC, Canada
[6]The Centre for Research on Brain, Language and Music, McGill University, Montreal, QC, Canada

Corresponding Author:
Nicole YK Li-Jessen, PhD
School of Communication Sciences and Disorders
Faculty of Medicine and Health Sciences
McGill University
2001 McGill College, 8th Floor
Montreal, QC, H3A1G1
Canada
Phone: 1 5143985933
Email: nicole.li@mcgill.ca

Abstract

Background: Mobile health (mHealth) apps are software interfaces that enable users to access and manage wearable technology through smartphones and tablet devices for health improvement purposes. However, many respiratory disease mHealth apps lack transparent development documentation, compromising user confidence in their quality, functionality, and usability.

Objective: This study aimed to develop and evaluate AIrway, a companion mHealth app designed to interface with an in-house wearable device for monitoring airway symptoms following established mHealth development and reporting standards.

Methods: The development cycle of AIrway comprised 2 study phases. In phase 1, AIrway, a native Android app, was developed following academic and industrial standards (Android material design and Morville's design principles) and privacy regulations (Personal Information Protection and Electronic Documents Act). Core functionalities included location-based environmental monitoring, a clinical diary with action plans, Bluetooth connectivity, and real-time data storage. In phase 2, the usability of AIrway was evaluated by software app developers using standardized assessment tools, namely, the User Version of the Mobile Application Rating Scale survey and the IQVIA questionnaire.

Results: AIrway successfully fulfilled 7 of 8 development criteria on usability, privacy, security, appropriateness, transparency, safety, and technical support, with only the technology aspects requiring refinement. Accessibility assessments confirmed that AIrway's content and interface were comprehensible to the general population (grade 9-10 reading level). Technical testing demonstrated reliable Bluetooth data transmission for up to 10 minutes without interruption. User evaluation scores for the User Version of the Mobile Application Rating Scale (3.6/5.0) and IQVIA (8/11) were comparable to those of similar mHealth apps on the market.

Conclusions: By adhering to established mHealth app design principles, AIrway achieved the necessary accessibility standards and wireless communication capabilities for wearable device integration. Future development will focus on expanding cross-platform compatibility and conducting usability evaluation with intended patient populations to validate its clinical effectiveness and support ongoing improvements.

(JMIR Form Res 2025;9:e73584) doi:10.2196/73584

XSL•FO
RenderX

KEYWORDS

asthma; chronic obstructive pulmonary disease; COPD; development guidelines and standards; mobile health apps; usability assessment; wearable devices

Introduction

Background

Asthma and chronic obstructive pulmonary disease (COPD) each affect approximately 330 million people worldwide, representing major public health challenges with substantial health and economic burdens [1,2]. Self-management interventions are key to mitigate the risk of symptom exacerbations in patients with asthma and COPD from escalating into life-threatening emergencies. To better support self-management, wearable technology can enable remote monitoring and early exacerbation detection, allowing patients to act before symptoms worsen [3-5].

Effective respiratory monitoring requires specialized sensing technologies that address unique challenges. While surface body sensors accurately capture vital signs including heart rate, skin temperature, and breathing patterns, continuous acoustic monitoring raises privacy concerns due to incidental speech recording [6-8]. Neck surface accelerometers (NSAs) provide an alternative approach, capturing purely mechano-acoustic signals without recording speech, making them suitable for long-term chronic disease monitoring [9-14]. Recent studies demonstrate that NSA-based systems can detect voice physiology, swallowing events, and COPD severity indicators with 93% to 96% accuracy using machine learning approaches [15-17].

The success of wearable systems depends on reliable mobile interfaces. In ubiquitous health solutions, raw sensor data (eg, from NSA) are stored onboard before transmission via Bluetooth Low Energy (BLE) to mobile devices for local processing or cloud-based analysis. Results are delivered through mobile health (mHealth) apps that often include interactive features such as self-assessment questionnaires and chatbots to promote healthier behaviors [18-21]. However, despite the increasing use of Android platforms for mHealth development, many studies lack comprehensive documentation of development workflows, design decisions, and validation procedures [22]. This transparency gap hinders clinical adoption, as mHealth apps serve as the primary interface between patients and wearable devices.

Related Work and Research Gaps

Clinical research shows that wearable tools can increase medication adherence by 50% and improve disease-related knowledge by 52% [23]. However, Cochrane systematic reviews report mixed benefits for mHealth interventions targeting patients with asthma and COPD specifically.

Belisario et al [24] evaluated smartphone and tablet apps versus paper-based asthma action plans in 2 randomized controlled trials (n=408 participants) over 6 months. While emergency visits decreased significantly, symptom control showed no improvement on the patient-reported Asthma Control Questionnaire (ACQ). Likewise, mixed results are also noted in COPD research. Janjua et al [25] reviewed 14 randomized controlled trials (n=1518 participants) comparing mobile-based self-management to usual care. Evidence to date suggested that digital technology might provide short-term but not long-term improvements in patients' quality of life. Nevertheless, these 2 Cochrane reports noted key methodological limitations that undermined the evidence quality in asthma and COPD literature, such as heavy reliance on self-reported outcomes, unregistered protocols, and underpowered statistics from small sample sizes [24,25].

Despite that many mHealth solutions are available, most of them remain inadequate to fully support self-management interventions in asthma and COPD [26]. General-purpose wearables such as the Apple Watch [27] and Fitbit [28] provide basic physiological metrics but lack the disease-specific algorithms and personalized interventions essential for respiratory disease management. Specialized devices such as Afflo [29] and AcuPebble [30] capture respiratory signals but do not integrate data into self-management plans or provide timely exacerbation alerts. Stand-alone apps like Breathe [31], Sonde Health [32], and CoughPro [33] rely on smartphone microphones and questionnaires for symptom detection. These methods capture only intermittent snapshots of respiratory status, rather than providing continuous monitoring capabilities.

More importantly, most mHealth solutions did not disclose how well their apps adhere to regulatory frameworks. Multiple standards govern mHealth apps, including the European Commission's Green Paper on Mobile Health [34] and the US Food and Drug Administration's Mobile Medical Applications Guidance [35], which establish privacy, encryption, and data-transfer requirements. Beyond regulatory compliance, developers shall follow established design principles found in industrial guidelines such as Apple's Human Interface Guidelines [36] and Android's Material Design Guidelines [37]. Besides, apps handling protected health information must comply with the Health Insurance Portability and Accountability Act (HIPAA) or equivalent regional privacy regulations [38]. To address these multiple requirements, Llorens-Vernet and Miró [39] developed a comprehensive framework that consolidates these diverse requirements into 8 unified development criteria: usability, privacy, security, appropriateness and suitability, transparency and content, safety, technical support and updates, and technology.

The aforesaid research gaps, namely, suboptimal clinical evidence, technical constraints, and inadequate development documentation, underscored the need for more rigorously developed mHealth solutions in respiratory disease monitoring. Our team previously developed privacy-preserving NSA-enabled wearable sensors for airway symptom detection [12,13,25,40]. However, these sensors require a companion mobile app to enable data visualization, cloud storage, user interaction, and so on. This study aimed to address these technological gaps by

XSL•FO
RenderX

developing and evaluating a mobile app using transparent, framework-guided development processes.

Research Objectives

This study presented the development and evaluation of AIrway, an mHealth app for asthma and COPD management. Research objectives were (1) to develop the app that would collect patient-reported outcomes, environmental data, and peripheral accelerometer sensor data; (2) to evaluate the functionality of the app including wireless connectivity, cloud storage, and literacy analysis; and (3) to perform a pilot usability study with technical raters, namely, app developers.

Methods

Study Design

This study consisted of 2 phases, namely, frontend and backend development (phase 1) and a pilot usability test (phase 2). The AIrway development followed rigorous academic and industry best practices to ensure compliance with the mHealth development standard. Comprehensive documentation of app requirements, design, implementation and validation, and benchmark were reported against the framework of Llorens-Vernet and Miró [39].

Phase 1a: Frontend Development

App Interface Implementation and Workflow

The AIrway user interface (UI) was set up using XML (version 1.0) [41] in Android Studio (Chipmunk 2021.2.1 version) [42], a Google native Android development integrated development platform. The integrated development platform allowed developers to build Android apps on Windows, Mac, and Linux operating systems. The app supported respiratory disease management with 2 types of interfaces. Task-based interfaces allowed users to create a personalized account and log in, manage location permissions, input medication details, log symptoms via a clinical diary, and set diary reminders. Meanwhile, visualization interfaces displayed local weather, action plans, and health data summaries and provided account and app privacy information (Figure 1).

Figure 1. The design flowchart of the AIrway app. The login and registration interface allowed users to input their login information and create an account. The location permission interface allowed users to turn on or off the device's location for parsing local weather information. The account interface allowed users to complete the profile setup. The app also had interfaces today, report, profile, and help for data visualization and information delivery. ACQ: Asthma Control Questionnaire; BCSS: Breathlessness, Cough and Sputum Scale; CAT: COPD Assessment Test; COPD: chronic obstructive pulmonary disease; FAQ: frequently asked question; NRS: Numeric Rating Scale.

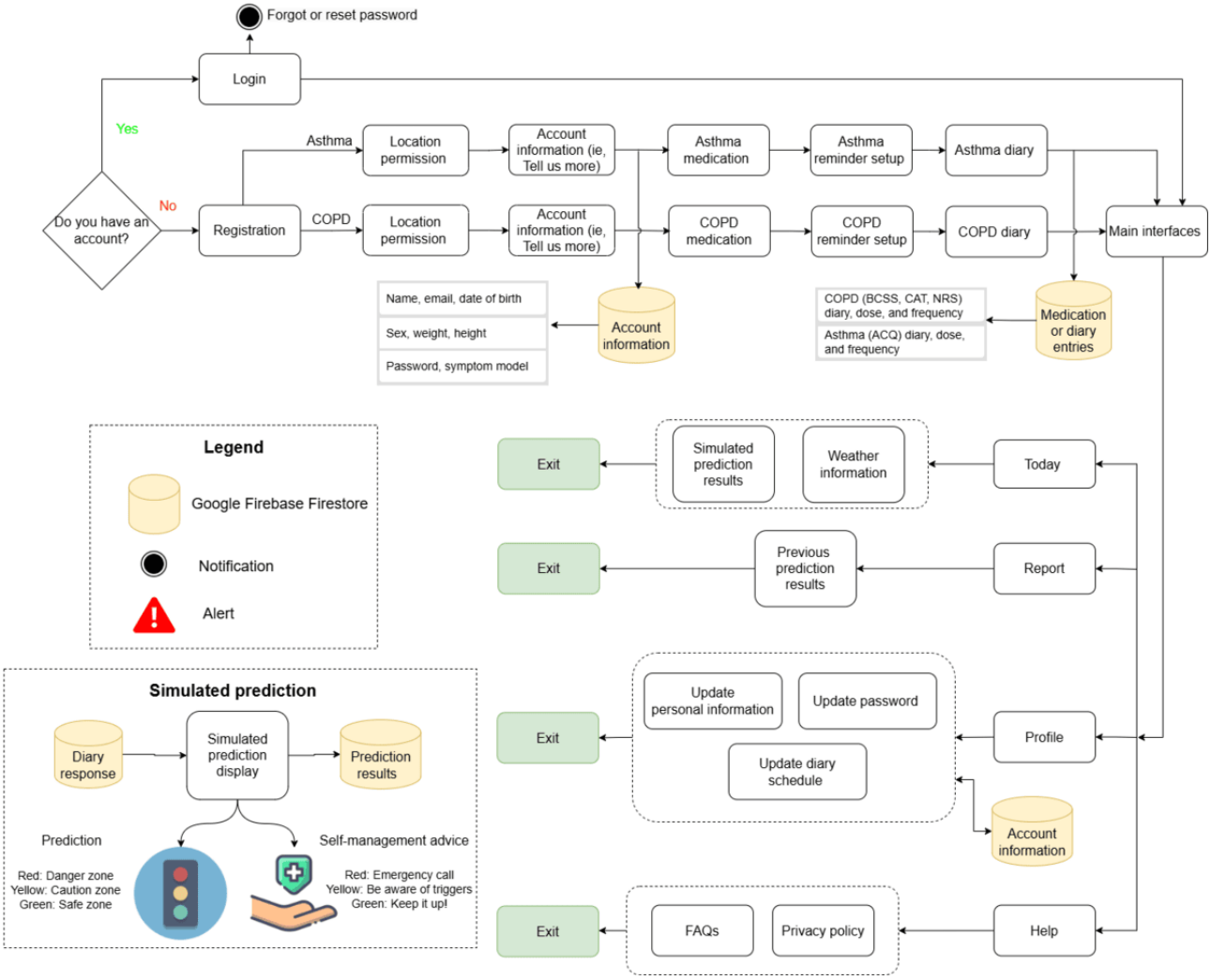

XSL•FO
RenderX

Design Principles

The design principles of Morville [43] were followed to guide the user experience (UX) and UI design of AIrway (Table 1). In brief, Morville's principles include 7 aspects of UX for mobile apps and websites: usefulness, value, findability, credibility, accessibility, desirability, and usability. These principles have been widely adopted in mHealth apps such as an emotional and physical health support app for older people [44] and a posttraumatic stress monitoring app [45].

Table 1. Design principles of the AIrway appa.

Design principle	Question (Q) and answer
Usefulness	Q1: Does the app have practical value for the target users? • The app allowed the target users (ie, patients with asthma or COPDb) to create a personalized account, monitor their conditions, and predict airway conditions through a clinical diary. It also provided reminders and alerts for diary completion. The app stored the data in a cloud service.
Value	Q2: Does the app advance the mission of the organization behind it? • The AIrway app was developed by the Voice and Upper Airway Research Lab at McGill University, Canada, in collaboration with researchers at the University of Erlangen-Nürnberg, Germany. We are improving personalized medicine in voice and upper airway dysfunctions with advanced technology. For this project, we hope to support asthma and COPD management and provide recommended actions to the user.
Findability	Q3: Can users locate what they are looking for? • The app was developed using large fonts, buttons, and a navigation menu to assist the user to switch between interfaces. The app also incorporated some frequently asked questions to support the user.
Credibility	Q4: Is the app trustworthy? • The app components were developed based on other mHealthc apps, such as Fitbit [28], Afflo [29], and Breathe [31]. The clinical diary and prediction results were developed based on clinical guidelines and literature (such as GINAd and GOLDe). Data are safely stored in Google Cloud Services, which are eligible for HIPAAf compliance.
Accessibility	Q5: Are there barriers that may prevent the target users from using the app? • The app was developed based on the WCAGg [46] and readability metrics. For instance, our app used user-friendly colors for the app background and easily understandable written materials for the app content.
Desirability	Q6: Do the target users want to use the app? What are the responses? • A usability study was conducted with app developers. The study protocol was outlined, and the responses were summarized in the Results section.

[a]A question and answer format was presented to summarize the adaption of Morville's design principles to the app's user experience and user interface elements in this study [43].

[b]COPD: chronic obstructive pulmonary disease.

[c]mHealth: mobile health.

[d]GINA: Global Initiative for Asthma.

[e]GOLD: Global Initiative for Chronic Obstructive Lung Disease.

[f]HIPAA: Health Insurance Portability and Accountability Act.

[g]WCAG: Web Content Accessibility Guidelines.

For AIrway, the end users are individuals with asthma or COPD, who are often older people. To apply Morville's principles for the app, the UX and UI were built with simplified UIs, large texts, easy-to-use navigation buttons, and motivational functions, such as reminders to keep users engaged and encouraged for continued use [47].

Optimal Element Layout, Text Font, and Color Contrast for App Interfaces

The Android operating system was chosen for deploying AIrway across different devices, such as LG, Google Pixel, and Samsung. As these devices can vary significantly in their screen size, the app content display can be altered, and information may become misaligned.

To address this development issue, the app's UI was built using the constraint layout [48] following the Android Material Design Guidelines [37]. Constraint layout allows UI elements (eg, buttons and texts) to be positioned relative to each other. For instance, the elements can always be specified relative to the top, center, or bottom of the screen using constraint layout in the XML (version 1.0) file [48]. That way, the app's UI components were flexible enough to be adjusted and fit different screen sizes. Besides, Android Guidelines [37] recommend a minimum text size of 20 scalable pixels for titles, 16 scalable pixels for subtitles, and 14 scalable pixels for buttons and body to improve readability. These recommendations were adhered to in the development of the app.

Color contrast is critical for user accessibility and usability. According to Web Content Accessibility Guidelines [46], texts

XSL•FO
RenderX

and images should have a minimum contrast ratio of 7:1 (ie, AAA compliance). The app's UI components, such as backgrounds, buttons, drop-down menus, and input fields, were adjusted using an open-source color contrast checker [49] to meet or exceed this requirement. All interface elements achieved contrast ratios between 12:1 and 21:1, ensuring high visibility for users with low vision or reduced contrast sensitivity.

User Login and Registration

Upon using the app, users first interacted with the login interface. They could either log in with their credentials or create an account on the register interface by entering their name, email, and password and selecting a condition to monitor. The email input field included validation checks for proper formatting, and passwords were masked for security. Based on their selection, users were directed to either asthma or COPD-specific interfaces. If users forgot their password, they could request a reset link via a valid email and then log in with the new password (Multimedia Appendix 1).

Location Permission

After registration, the app requested access to the device's location to retrieve local weather data. This location permission was declared in the AndroidManifest.xml file, with ACCESS_COARSE_LOCATION enabling access to the approximate location without storing it [50]. The app requested the permission at runtime, offering 3 options: "While using the app," "Only this time," or "Deny" (Multimedia Appendix 1).

User Account Information

Users provided personal information including their biological sex, date of birth, weight, and height. Biological sex was a drop-down menu with 3 options: "Male," "Female," and "Prefer not to say." A DatePicker [51] component allowed users to easily scroll through dates for their date of birth. Input fields for weight and height accepted only numbers, with selectable units: "kg" or "lbs" for weight and "m" or "ft" for height (Multimedia Appendix 1).

Medication Profile

The medication profile design was based on the guidelines from the Canadian Lung Association [52], the Ontario Lung Association [53], and the Breathe app [31]. Depending on the condition, users filled out either the asthma or COPD profile including the medication name, amount, and frequency. To handle over 10 options without cluttering the UI, a drop-down menu was implemented. This design allowed users to easily select a suitable option with a single click, after which the menu collapsed to display only the selected option, providing a clean UI to users (Multimedia Appendix 1).

Clinical Diary

Effective self-management and improved health care outcomes for asthma and COPD rely on clinically validated diaries that are user-friendly, relevant to the patient's symptoms and medications, and capable of identifying the predictors of exacerbations [54,55]. The app prompted users to fill out the clinical diary (asthma or COPD) daily either after the initial registration or at the scheduled time.

For asthma, it used the ACQ [56], a clinically validated tool recommended by the Global Initiative for Asthma (GINA) [57], comprising 5 items for assessing asthma symptoms and 1 item for rescue inhaler bronchodilator use. For COPD, it incorporated 3 questionnaires: the COPD Assessment Test [58], the Breathlessness, Cough and Sputum Scale [59], and a Numeric Rating Scale [60] for a total of 12 items.

The diary interface presented 1 question per page with "Back" and "Next" buttons along with a progress bar. Responses used RadioButton or Likert-scale styles for single-choice responses, in which users could change or deselect answers. If users tried to move on without answering, a pop-up message "Please select one choice" would appear (Figure 2). A TimePicker [61] component allowed users to schedule a daily reminder. At the scheduled time, a notification would appear with the message "Please complete the diary now!" offering 2 buttons: "Complete" (to open the diary) and "Dismiss" (Multimedia Appendix 1).

XSL•FO
RenderX

Figure 2. Clinical diary interfaces of AIrway. (A) An Asthma Control Questionnaire question is presented in the asthma diary interface. (B) The Numerical Rating Scale question is presented in the chronic obstructive pulmonary disease diary interface. The diary interface presented 1 question per page with "Back" and "Next" buttons along with a progress bar.

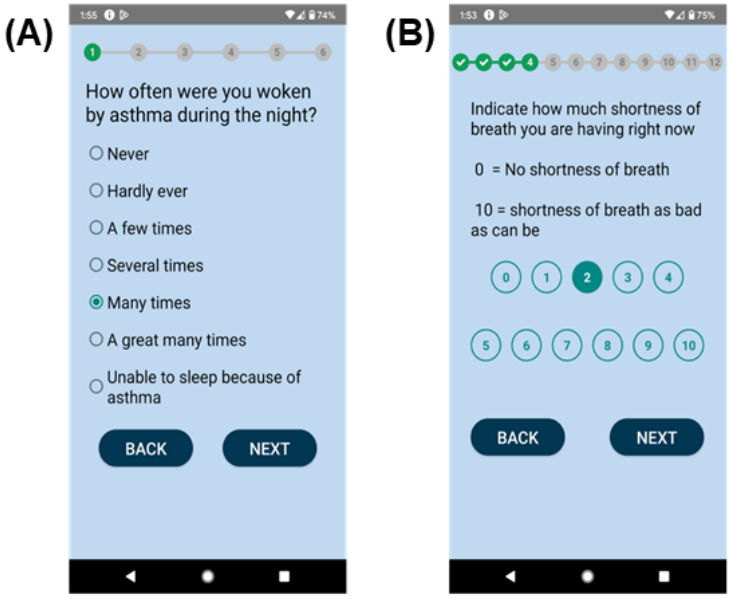

Today Interface: Local Weather Information and Control Zone Results

The today interface featured both local weather information and the asthma or COPD control zone of the day (Figure 3A). Since asthma or COPD symptoms can be impacted by air pollution, temperature, and humidity [62], the interface incorporated key air quality indicators outlined by the World Health Organization guidelines, including particulate matter (PM), ozone (O_3), sulfur dioxide (SO_2), and nitrogen dioxide (NO_2) [63]. In Canada, Environment Canada's Air Quality Health Index (AQHI) is used to convey the health impact of air quality. This AQHI information was retrieved and displayed on the today interface [64].

XSL•FO
RenderX

Figure 3. AIrway visualization interfaces. (A) Today, (B) report, (C) profile, and (D) help interfaces were designed for information delivery.

Report Interface: Clinical Symptom Prediction

The report interface (Figure 3B) featured a pie chart representing the percentages of each symptom. These numbers, for instance, could represent symptom data sourced from an external wearable device or body sensor via wireless BLE communication technology.

Profile Interface: Account Information Management

The profile interface (Figure 3C) allowed users to view the information they entered during the previous profile setup by clicking the corresponding texts or arrow icons. In addition, users could manage their accounts and update the diary schedule as well as log out of their accounts by clicking the "Logout" button.

XSL•FO
RenderX

Help Interface: Support, Resources, and Data Privacy

The help interface (Figure 3D) provided users with detailed information on the app's privacy policy and frequently asked questions (FAQs). The privacy policy, developed in line with the Personal Information Protection and Electronic Documents Act [65] and Quebec Privacy Law [66], explained the type of data collected as well as how it would be stored and used (Multimedia Appendix 2). The FAQs provided additional explanations on the development of asthma and COPD zone calculations and technical support (eg, reset password procedure). It also included the research team's contact email for additional inquiries.

In addition, the Google Firebase Firestore database [67] used for app data storage is HIPAA [68] compliant. Google will strictly adhere to the US national standards to protect health information and data through the Business Associate Agreement. The privacy policy also included the Business Associate Agreement information.

Literacy Analysis for the App's Written Materials

The app's written materials, including the privacy policy, FAQs, the asthma action plan, and the COPD action plan, were evaluated using an open-source readability calculator, Online-Utility [69]. The target reading level was set to grade 9, matching most English-educated adults in the United States [70]. The Flesch Reading Ease Score, a standard metric for research, educational, and digital content [71], was applied (70-80=grade 7, 60-70=grades 8-9, and 50-60=grades 10-12) [72]. All text was iteratively revised until it consistently met the ninth-grade threshold without losing essential detail. Definitions for each readability metric appear in Multimedia Appendix 3.

Mobile Accessibility Analysis for App Interfaces

Based on the World Wide Web Consortium guideline [73,74], the content of websites and apps needs to be accessible to people with disabilities and the general population. The Accessibility Scanner [75], the official Android accessibility testing tool available on the Google Play Store, was used to verify the accessibility of the app. The scanner took a snapshot of the target interface and evaluated its components, such as content labels, target size, clickable items, and text and image contrast. The scanner then provided content-based descriptions and recommendations for improvement. For example, one suggestion was to adjust the item's height from 32 to 48 density-independent pixels or larger.

Phase 1b: Backend Development

The backend development was programmed using Java (version 11.0.12) [76] in Android Studio (Chipmunk 2021.2.1 version) [42] to implement (1) local weather data retrieval, (2) asthma and COPD control zone calculations, (3) cloud database storage and retrieval, and (4) BLE data transmission.

Local Weather Data Retrieval

The app retrieved the current local weather information (temperature and humidity) and used the current $PM_{2.5}$, O_3, and NO_2 to compute the AQHI. The information was displayed on the today interface using the OpenWeatherMap application programming interface (One Call Version 3.0), an online service that provides the latest forecast data and air pollution information in over 200,000 cities for web and mobile app developers [77].

Control Zone Calculations and Generation of Action Plans

For asthma, the inputs for control zones (red, yellow, and green) included (1) user responses from the ACQ [56], (2) air quality data from the AQHI [64], as well as (3) the temperature and humidity levels of user's current city location. The ACQ provides a mean score determining asthma severity, while the AQHI indicates health risk levels, with high readings exacerbating symptoms [56,78]. Extreme temperature and humidity can also trigger asthma symptoms [79]. Additionally, the app categorized asthma severity into "Uncontrolled," "Partly controlled," and "Well controlled" based on GINA guidelines, providing corresponding self-management action items from the Centers for Disease Control and Prevention asthma action plan [80] (Multimedia Appendix 4).

For COPD, the inputs for control zones (red, yellow, and green) included (1) user responses from the 3 validated questionnaires: the Breathlessness, Cough and Sputum Scale [60], the COPD Assessment Test [58], and the Numeric Rating Scale [81]; (2) air quality data from the AQHI; as well as (3) the temperature and humidity levels of user's current city location. These questionnaires assess the impact of COPD symptoms, which can be worsened by extreme temperature and humidity. Self-management action items for each control zone were provided based on the Canadian Thoracic Society COPD action plan [82] (Multimedia Appendix 4).

Cloud Database

mHealth apps often integrate cloud databases for data storage, processing, and sharing to ensure ease of access and scalability. In the AIrway app, Google Firebase Firestore [67] was used to store large volumes of real-time user data, such as account information and clinical diary responses. Firestore was considered an ideal cloud database for the app for 3 reasons.

Firestore was chosen for 3 main reasons. First, it provides real-time updates and integrates directly into common mobile development frameworks. Researchers have also used Firestore in an eHealth management system for cardiovascular diseases [83] and a humidity and temperature monitoring system [84]. Second, its Not Only Structured Query Language design requires no fixed schema: collections and documents can grow independently and hold diverse data types, from text and numbers to arrays and objects [85-87]. Third, unlike traditional Structured Query Language databases that need hardware upgrades and incur downtime, Firestore scales horizontally on demand, adding capacity without service interruption [88].

Cloud Database Access Feature

To access Firestore, users must authenticate via the Firebase Authentication Service using credentials such as email and password, phone number, and third-party providers (eg, Google and Facebook) [89]. In AIrway, the register interface collected email and password inputs. Once the user input fields were

XSL•FO
RenderX

filled, the createUserWithEmailAndPassword method [89] was used to create an account. Upon successful registration, the message "User successfully created!" was displayed, while a failure showed "This account already exists. Please register another email!" preventing users from moving to the next interface. The registered email and password then served as the user's sign-in key for logging in and out of Firestore.

Cloud Data Storage and Data Retrieval

Storing app data in Firestore involved 3 steps. First, read or write access was enabled by setting the rules to "True" for authenticated users. Second, in Android Studio, a storage path was defined with a collection named "users" and a document named after the user ID generated during authentication. User data were written within this collection, with each user's specific data stored in a separate document (Figure 4). Third, new data objects were added using the Map <String, Object> structure and push methods. For example, an asthma diary question like "How bad were your asthma symptoms when you woke up in the morning?" was stored with the selected RadioButton response under the header "Asthma Diary Response."

Figure 4. Firestore cloud architecture and data flow of the AIrway app. The storage structure comprised a "users" collection containing individual documents identified by authentication UIDs. UID: user ID.

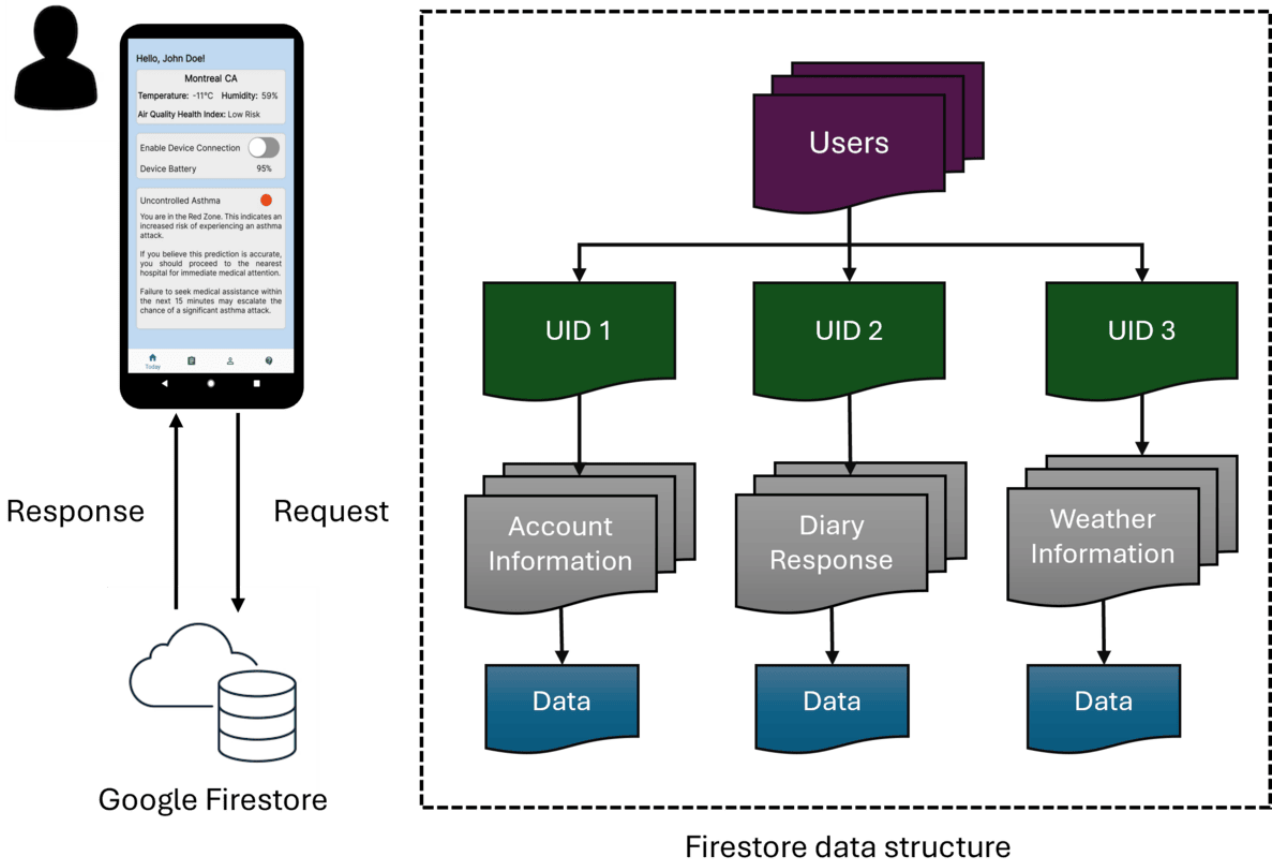

The app also retrieved user data, such as the account details in the profile interface, to enhance data visualization and increase user retention. This feature allowed users to quickly access personalized data and review previous entries before making updates. For example, when updating a password, the app reminded users of their current password to prevent them from setting the same one again. To achieve this dynamic data retrieval feature, the app first checked if the user document existed. If it did, the getString method fetched the corresponding data in the app. The setHint method then displayed the stored data.

BLE Communication Features and Testing

BLE data transmission enables the app to receive data from an external wearable device, that is, NSA in this study with minimal power use [90,91]. The BLE protocol stack [92] has 2 sections. The controller is at the lower level of the stack that handles radio transmission and manages the data connection between 2 devices, while the physical layer is responsible for analog communication operations [93]. The host operates at a higher level of the stack that defines the data transmission between a device (peripheral) and a mobile app (client) using the Attribute Protocol and the Generic Attribute Profile. The Attribute Protocol allows for data request between the client and the peripheral, whereas the Generic Attribute Profile organizes data into services and characteristics [91,94,95].

In the AIrway app, the Android BLE application programming interface [96] handled scanning, connecting, and reading functions. Users clicked a button to scan for nearby devices, and the app listed available devices, showing each device's name, Media Access Control address (ie, a hardware identifier that uniquely identifies each device), and its Received Signal Strength Indicator measured in decibel-milliwatts (dBm). After

XSL•FO
RenderX

selecting a target device, the app established a connection using the connectGatt method [97], read the characteristics (ie, numerical data) using the readCharacteristic method [98], and displayed the characteristic values in real time. To optimize storage, only the first bit of each 16-bit characteristic representing the actual value was stored in Firestore.

For testing, a prototype printed circuit board based on the nRF52840 development board [99] was used to simulate 2 services, namely, symptom event classification (sending random integers every half second) and battery level (0 to 100). This setup verified reliable BLE data transmission to the app.

Phase 2: Functionality Usability Tests With Technical Raters

A pilot usability study with technical raters (app developers) was performed to evaluate the app's pilot functionality.

Implementation of App Usability Survey

The User Version of the Mobile Application Rating Scale (uMARS) [100] and IQVIA [101] were used. The uMARS questionnaire, widely used for evaluating mHealth apps [100], is structured into 5 sections: engagement, functionality, aesthetics, information quality, and subjective quality. In this study, the original uMARS questionnaire was modified for clarity and to ensure its relevance to the specific functions of the app with an open-text prompt for feedback (Multimedia Appendix 5). In addition, the IQVIA questionnaire was used to assess the presence of specific app functions, with a particular focus on self-management features such as alerts, guidance, and communication [101] (Multimedia Appendix 6).

The usability survey was deployed using McGill's IT-managed LimeSurvey platform (version 3; LimeSurvey GmbH) and contained 51 items in total. The survey included prescreening (5 items), consent (1 item), and 4 separate sections. Sections A and B measured technical raters' demographics (6 items) and mobile app development experience (5 items). Section C included the uMARS questionnaire (23 items), while Section D included the IQVIA questionnaire (11 items).

Study Recruitment and Procedure

A convenient sample of app developers was recruited via an electronic advertisement that was disseminated through Canadian universities' computer science student mailing lists and social media. Inclusion criteria were (1) being aged 18 years and older; (2) at least 1 year of mobile app development experience or, for computer science PhD students and postdoctoral fellows, completion of a mobile app development course or related thesis research; (3) Android smartphone user with internet access; and (4) use of English as the primary daily language. Exclusion criteria included severe cognitive and psychiatric conditions that may prevent study completion. Technical raters were compensated with a US $35 Amazon gift card.

During the 1.5-hour Zoom (Zoom Video Communications) session, a study investigator (AC) explained the workflow. The raters received a fictitious Gmail address to access an instruction manual. After completing prescreening, consent, demographics, and mobile app development sections in LimeSurvey, they received an email from Firebase App Distribution [102] to download the app. Once the tasks were completed, they filled out the uMARS and IQVIA questionnaires in LimeSurvey.

Ethical Considerations

The human study was approved by the institutional review board of the Faculty of Medicine and Health Sciences at McGill University (protocol A12-E39-22B). Informed consent was obtained from all participants. Participant data were deidentified and anonymized for privacy protection. Each received a US $35 Amazon gift card as compensation.

Results

Verification of mHealth Apps' Development Standards

AIrway met 7 of the 8 best practices for mHealth app development (Table 2). Usability compliance was confirmed through a technical-expert evaluation and strict adherence to Android design guidelines. As the app targeted English-speaking adults with respiratory conditions, multilingual support was not included.

XSL•FO
RenderX

Table 2. Comparative results of AIrway app against established mobile health development standards [39].

	AIrway
Usability criterion	
The app has been tested by potential users before being made available to the public.	Yes (usability study)
It has instructions or some kind of assistance for use.	Yes (Help page)
It is easy to use (ie, navigation is intuitive).	Yes (usability study)
It follows the recommendations, patterns, and directives in the official manuals of the different operating systems (Android, iOS, or others).	Yes (Android development guidelines)
The interface design follows the same pattern. That is, all graphic elements (typographies, icons, and buttons) have a consistent appearance. The function of each element (navigation menu, lists, and photo gallery) is clearly identified.	Yes
The functionality is adapted to the purpose of the app.	Yes
The information of the app must be able to be accessed in the shortest possible time. All users must be able to access all resources regardless of their capabilities.	Yes
The app can be consulted in more than 1 language. All languages adapt appropriately to the content interface.	Not applicable
Privacy criterion	
The app gives information about the terms and conditions of purchases in the app and the personal data recorded.	Yes
It gives information about the kind of user data to be collected and the reason (the app must only ask for user data that are essential for the app to operate). It gives information about access policies and data treatment and ensures the right of access to recorded information. It describes the maintenance policy and the data erasure procedure. It gives information about possible commercial agreements with third parties.	Yes (consent form or privacy policy)
It guarantees the privacy of the information recorded. It requires users to give their express consent. It warns of the risks of using the app.	Yes (privacy policy)
It tells users when it accesses other resources of the device, such as their accounts or their social network profile.	Not applicable
It takes measures to protect minors in accordance with the current legislation.	Not applicable
Confidential user data are protected and anonymized, and there is a privacy mechanism so that users can control their data.	Yes (anonymized ID or fictitious names)
Security criterion	
The app has encryption mechanisms for storing, collecting, and exchanging information. It has password management mechanisms.	Yes (password protection in the cloud)
The cloud services used have the relevant security measures. It states the terms and conditions of cloud services.	Yes (Firestore cloud terms or conditions)
The authorization and authentication mechanisms protect the users' credentials and give access to their data. It limits access to data that are only necessary for the user.	Yes (Firestore account credentials or login)
It detects and identifies cybersecurity vulnerabilities, possible threats, and the risk of being exploited. It applies the appropriate security measures to cybersecurity vulnerabilities in the face of possible threats.	Not applicable
Appropriateness and suitability criteria	
The end users for whom the app is designed are explicitly indicated or actually intuitable (the name identifies the app) to the audience to whom it is set out.	Yes
The benefits and advantages of using the app are explained.	Yes
The app has been validated or created by experts (eg, a group of specialized professionals, a health organization, or a scientific society).	Partially yes (usability study with developers)
Transparency and content criteria	
The app identifies the authors of the content and their professional qualifications.	Yes (Help page or FAQ[a])
It gives transparent information about the owners' identity and location.	Yes (Help page or FAQ)
It gives information about its sources of funding, promotion, and sponsorship, and possible conflicts of interest. Any third parties or organizations that have contributed to the app development are clearly identified.	Yes (Help page or FAQ)
It uses scientific evidence to guarantee the quality of the content. It is based on ethical principles and values.	Yes
The sources of the information are indicated. Concise information is given about the procedure used to select the content.	Yes

XSL•FO
RenderX

	AIrway
Safety criterion	
The possible risks to users are identified. Users are warned that the app does not intend to replace the services provided by a professional.	Yes (action plan)
Potential risks for users caused by bad use or possible adverse effects are explained.	Yes
Technical support and updates criteria	
It gives a warning if updates modify or affect how the app functions. It gives a warning if updates can influence insensitive data.	Yes (Firebase App Distribution can send updates)
Frequent security updates are guaranteed. Every time an update of a third-party component is published, the change is inspected, and the risk evaluated.	Yes (Firebase App Distribution can send updates)
The frequency with which the content of the app is revised or updated.	Yes (Firebase App Distribution can send updates)
Users have support mechanisms (email, phone, and contact form) for solving doubts, problems, or issues related to the health content and technical support.	Yes
Technology criterion	
It works correctly. It does not fail during use (eg, blocks). Functions are correctly retrieved after context changes (eg, switch to another app and return), external interruptions (eg, incoming calls or messages), and switching off the terminal.	Yes
It does not waste resources excessively: battery, central processing unit, memory, data, or network.	Partially yes (battery use with BLE[b], memory use)
It can work in flight mode and deal with network delays and any loss of connection.	No (need internet to log in and connect to the cloud)
It supports multiple versions of data structures.	Yes (graph, text, and button)
It supports multiple formats (eg, to support different operating systems).	No (Android only)

[a]FAQ: frequently asked question.

[b]BLE: Bluetooth Low Energy.

Privacy is protected by a comprehensive policy that covers data collection, terms of use, and secure cloud storage. Security relies on password-protected Google Firestore access, although broader cyberthreat defenses remain to be implemented. The app's appropriateness for English-speaking adults with respiratory conditions was demonstrated via targeted performance testing.

Transparency requirements were satisfied through dedicated Help and FAQ sections providing author credentials, mission statements, funding information, and scientific evidence supporting the asthma and COPD control zone calculations. Safety protocols included explicit disclaimers regarding clinical diagnosis limitations and emergency response guidance for users experiencing severe symptoms.

Technical support is provided via monitored email channels and app updates distributed through Firebase App Distribution. The only area needing enhancement is technology—specifically, adding cross-platform compatibility and offline functionality, since AIrway is currently Android-only and requires a constant internet connection.

Verification of Literacy Analysis

Based on the readability calculator [69], the Flesch Reading Ease score [103] for the privacy policy, FAQs, the asthma action plan, and the COPD action plan were 58.63, 56.69, 57.72, and 59.91, respectively. With an average score near 60 (1.4), these materials are accessible to users with a formal education level of grades 9-10. The Gunning Fog Index [104] indicated an 8th- to 10th-grade reading level, while the Coleman Liau Index [105] suggested suitability for a 7th- to 10th-grade audience. The Flesch-Kincaid Grade Level [106] indicated a 6th- to 8th-grade range, and the Automated Readability Index [72] estimated a 5th- to 7th-grade level. The Simple Measure of Gobbledygook Index measured the texts based on syllables [107] and concluded that the materials were suitable for a 9th- to 10th-grade audience (Table 3).

XSL•FO
RenderX

Table 3. Results of literacy analysis on AIrway app materials based on 5 standard readability indices [108].

Readability metrics	Privacy policy	FAQ[a]	Asthma action plan	COPD[b] action plan	Interpretation of the metric
Gunning Fog Index	8.97	9.38	8.44	8.49	8th to 10th grade
Coleman Liau Index	9.18	9.39	7.47	8.35	7th to 10th grade
Flesch-Kincaid Grade Level	7.57	8.03	7.17	6.80	6th to 8th grade
Automated Readability Index	5.94	6.33	4.86	5.38	5th to 7th grade
SMOG[c] Index	9.32	9.67	9.06	9.71	9th to 10th grade

[a]FAQ: frequently asked question.

[b]COPD: chronic obstructive pulmonary disease.

[c]SMOG: Simple Measure of Gobbledygook.

Accessibility Results

The accessibility analysis identified 3 primary areas for further enhancing the app's UIs (Multimedia Appendix 7). First, it recommended modifying the width layout of test view components from a fixed value to an adjustable parameter, such as wrap_content. This modification enabled dynamic text expansion, which could prevent component overlap when screen sizes change. Second, it suggested increasing the text size of input items from the current range of 24-44 density-independent pixels to a minimum of 48 density-independent pixels for better readability. Third, it recommended enhancing the contrast colors of the "Today," "Report," "Profile," and "Help" icons to achieve a minimum ratio of 3.00:1 as well as the contrast of unselected time-scrolling texts within the diary reminder interface to at least 4.50:1 with respect to the background.

BLE Experiment Results

The BLE experiment evaluated data transmission between AIrway (client) and the nRF52840 development board (peripheral; Tables 4 and 5). First, the app's scan feature accurately detected the board by its device name, "NSA BLE," along with the unique manufacturing MAC address (D5:D2:5D:62:AF:48). Signal strength, measured in decibel-milliwatts on a logarithmic scale (with values closer to 0 dBm indicating stronger signals [109]), showed that the board was advertising the highest signal value of –19 dBm among the neighboring BLE devices, which confirmed its closest proximity to the app.

Table 4. Results of Bluetooth Low Energy scan range.

Scan range (dBm)	Corresponding MAC address
–94	1E:73:99:6D:30:E7
–19	D5:D2:5D:62:AF:4B
–100	57:97:8B:22:DC:5F
–86	5B:7B:F2:E7:2C:3C
–81	6F:AE:90:EE:00:95

Table 5. Results of Bluetooth Low Energy data transmission.

Data transmission timestamp	Read characteristics
May 18, 16:33:28	0x0D (equivalent to integer 13)
May 18, 16:33:29	0x0D
May 18, 16:33:30	0x0C (equivalent to integer 12)
May 18, 16:33:31	0x0C
May 18, 16:33:32	0x0B (equivalent to integer 11)

Once connected, the data read characteristic feature was successfully displayed and verified visually. As described in the experiment setup, simulated battery level data were transmitted using random integers and set to decrement by 1 every 2 seconds. The app accurately retrieved the real-time data timestamp along with the corresponding hexadecimal data value (eg, 0x0D representing integer 13, followed by 0x0C representing integer 12 after 2 seconds). This result confirmed the successful integration of the scan, connect, and read characteristic features.

Furthermore, physical distance, connection time, and obstacles were found to impact BLE performance (Table 6). The best performance for stable and continuous data streaming occurred when the devices were adjacent without obstacles. However, after connections exceeded 10 minutes, some packet loss was observed. Additionally, occasional disconnections, particularly

XSL•FO
RenderX

when they were farther apart, were likely due to interference from nearby BLE devices. In these cases, the app was able to manually reconnect to the board.

Table 6. Results of Bluetooth Low Energy performance test in terms of physical distance, connection time, run time, and physical obstacles.

Task	Physical distance	Scan signal strength (dBm)	Run time (hours:minutes:seconds)	Obstacle
Put the board next to the phone for 5 minutes	0	–26	14:37:35 to 14:42:37	None
Trial 1: put the board next to the phone for 10 minutes	0	–26	16:23:50 to 16:29:19 (self-disconnected)	None
Trial 2: put the board next to the phone for 10 minutes	0	–26	15:40:18 to 15:51:10	None
Put the board next to the phone for 20 minutes	0	–26	14:01:37 to 14:20:07	None
Put the board next to the phone for 30 minutes	0	–26	15:10:17 to 15:42:09	None
Put the board next to the phone for 40 minutes	0	–26	13:13:23 to 13:53:11	None
Put the board and the phone on 2 different laboratory tables for 5 minutes	~2 m apart	–70	14:50:24 to 14:52:49 (self-disconnected)	None
Put the board on the laboratory table and the phone outside the laboratory for 5 minutes	~4 m apart	–91	13:47:24 to 13:48:13 (self-disconnected)	Door
Put the board and the phone next to a microwave (off) for 5 minutes	~0.6 m apart	–58	14:49:44 to 14:55:01	Microwave
Put the board and the phone next to a microwave (on) for 1 minute	~0.6 m apart	–58	15:00:14 to 15:01:31	Microwave

App Usability Evaluation

A total of 5 app developers participated in the evaluation of AIrway usability. Most technical raters were female, identified as visible minorities, and resided in Canada. They held at least a bachelor degree, worked part-time, and had an average age of 29.0 (SD 5.61) years. The majority were computer science PhD students or postdoctoral fellows with 1-2 years of Android mobile app development experience, and each had developed at least 1 app before (Table 7).

XSL•FO
RenderX

Table 7. Demographics and mobile app development experience of technical raters.

Variable and category	Values
Sex identity, n (%)	
Male	2 (40)
Female	3 (60)
Prefer not to answer	0 (0)
Visible minority status, n (%)	
Yes	4 (80)
No	0 (0)
Prefer not to answer	1 (20)
Reside location, n (%)	
United States	0 (0)
Canada	5 (100)
Education level, n (%)	
High school diploma	0 (0)
Apprenticeship or trades certificate or diploma	0 (0)
College or CEGEP[a] degree or diploma, or university degree lower than bachelor	1 (20)
Bachelor degree	20 (20)
Graduate degree (master or doctorate)	3 (60)
Job status, n (%)	
Full-time (>30 hours per week) employed	3 (60)
Part-time (<30 hours per week) employed	2 (40)
Self-employed	0 (0)
Unemployed	0 (0)
Age (years)	
Mean (SD)	29.0 (5.61)
Primary role, n (%)	
Designing products (eg, UI[b] designer and interaction designer)	0 (0)
Developing software (eg, programmer, developer, and software engineer)	1 (20)
Testing software (eg, tester and quality analyst)	0 (0)
Managing software development (eg, project manager and IT manager)	1 (20)
Computer science PhD student or postdoctoral fellow	2 (40)
Other: software and hardware support analyst	1 (20)
Experience in mobile development (years), n (%)	
1-2	5 (100)
2-3	0 (0)
3-4	0 (0)
More than 5	0 (0)
Mobile app development platforms (multiple selected), n (%)	
iOS	2 (40)
Android	5 (100)
Windows	2 (40)
BlackBerry	0 (0)

XSL•FO
RenderX

Variable and category	Values
Apps developed, n (%)	
1	4 (80)
2-5	1 (20)
More than 5	0 (0)
Size of the mobile app development team in the organization, n (%)	
1-9 employees	3 (60)
10-99 employees	1 (20)
100-999 employees	0 (0)
1000-9999 employees	0 (0)
10,000+ employees	0 (0)
Not applicable	1 (20)

[a]CEGEP: Collège d'Enseignement Général et Professionnel.
[b]UI: user interface.

uMARS Results

The overall mean uMARS score for AIrway was 3.6 of 5.0 (SD 0.2), indicating above-average quality (>3.0; Figure 5A). The information quality section received the highest mean score of 4.4 of 5.0 (SD 0.2), which implied that the app's quality and visual information were highly appealing to users. Similarly, the functionality section also received the highest mean score of 4.2 of 5.0 (SD 0.4), which indicated excellent performance, ease of use, navigation, and gestural design for the app. The aesthetics section received a mean score of 3.5 of 5.0 (SD 0.6), while the engagement section received a mean score of 3.4 of 5.0 (SD 0.4). Both the aesthetics and engagement sections indicated appropriate layout elements and customization aspects. Among all the sections, the subjective quality section received a score of 2.6 of 5.0 (SD 0.3), which indicated a moderate willingness to use the app in the next 12 months and pay for its use.

XSL•FO
RenderX

Figure 5. Results of AIrway app usability and functionality. (A) Technical raters' means and SDs of uMARS and (B) technical raters' IQVIA functionality scores (n=5).

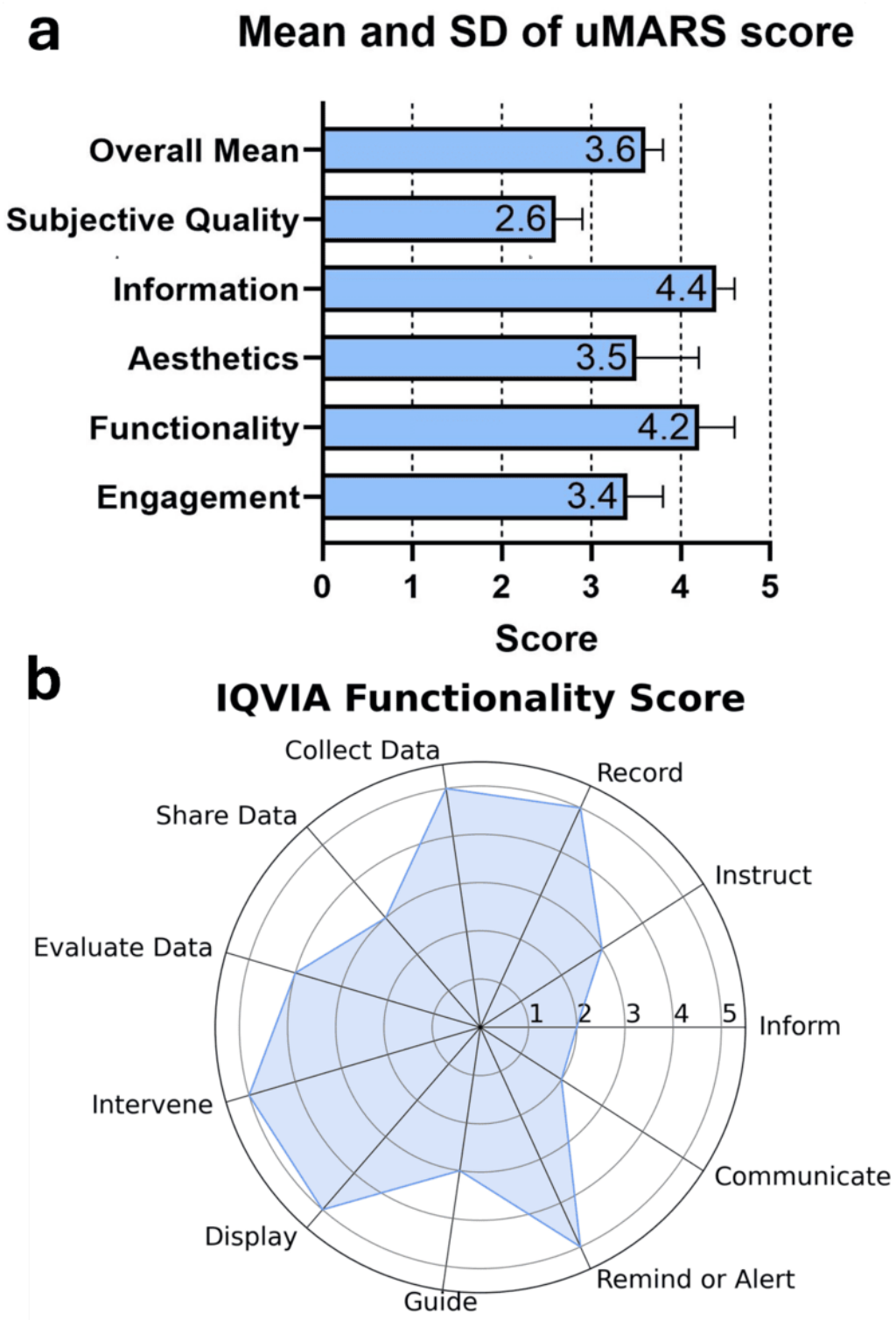

IQVIA Results

The IQVIA functionality score ranged from 6 to 11, with a median of 8 (IQR 7-10; Figure 5B). All technical raters confirmed that the app included recording, data collection, intervention, display, and alert functions (5/5, 100%). Most agreed (4/5, 80%) that the user data could be evaluated by others, and over half (3/5, 60%) expressed that the app supported instructing, guiding, and sharing data functions. However, only 1 (20%) rater observed a communication feature that allowed users to interact with others via social networks and provided information in a variety of formats (eg, text, photo, and video).

Open-Ended Feedback From Technical Raters

Feedback was collected through 5 open-ended questions regarding the app's engagement, functionality, aesthetics,

XSL•FO
RenderX

information quality, and potential improvements. Overall, the technical raters expressed positive impressions of the app and a clear understanding of its purpose. The results are summarized in Multimedia Appendix 8. A key suggestion was to add an option to show and hide passwords in the "Enter a password" and "Retype Your Password" fields on the register interface, which would help users verify that their passwords match.

Discussion

Overview

In recent years, mHealth apps targeting asthma monitoring have doubled from 2015 to 2019 [110], and COPD apps increased 6-fold between 2017 and 2023 [111,112]. The SARS-CoV-2 pandemic has notably increased clinicians' and patients' familiarity and positive attitudes toward self-monitoring and home-monitoring apps [113]. COVID-19 pandemic restrictions, such as social distancing mandates, shortages of protective equipment, and concerns about disease exposure, had increased the use and adoption of telehealth and remote monitoring technology from 11% to 46% among patients and health care providers [114]. In a study of 162 patients with acute COVID-19 infections, almost 79% of them were reported to engage in the digital platform of remote patient monitoring [114]. Further, to respond to this pandemic, the US Centers for Medicare and Medicaid Services approved over 80 new telehealth services and began reimbursement for telehealth visits since 2020 [114,115].

Principal Findings and Comparison to Prior Work

In this study, AIrway was designed to support NSA wearable devices for active asthma and COPD self-management purposes. AIrway did not provide diagnostic recommendations or treatment prescriptions to users and was not intended for medical use. Target user characteristics and needs were incorporated into the AIrway app UI design. For instance, the content of AIrway including the privacy policy, FAQ, and action plan was written at a 9th- to 10th-grade reading level that would enhance the comprehensibility and accessibility to general and geriatric users. Meanwhile, about 6.4% of patients with COPD or asthma using bronchodilator medications (eg, β_2 -adrenergic agonist bronchodilators) experience hand tremors, which can impair their ability to navigate touchscreen interfaces effectively [116]. The AIrway app has implemented the recommended text sizes (20 scalable pixels for titles, 16 scalable pixels for subtitles, and 14 scalable pixels for buttons and body) and constraint layout to enhance accessibility and adaptability across various screen sizes, following Android Material Design Guidelines. However, these measures were limited, as the app did not include a dynamic font sizing feature to adapt to individual user needs.

BLE tests with the nRF52840 development board demonstrated reliable real-time data exchange and accurate decoding of hexadecimal values. These results support future integration with NSA wearables using nRF5 series microcontrollers. Meanwhile, performance tests revealed that distance, connection time, and obstacles affect BLE stability. Future work will prioritize optimizing critical parameters like the maximum transmission unit and transmission frequency while implementing auto-reconnection functionality to minimize packet loss.

In the usability evaluation, AIrway scored 3.6 of 5.0 on the uMARS and a median IQVIA functionality score of 8 of 11 (IQR 7-10). These results were compared favorably with commercial mHealth apps for respiratory, cardiac, and sleep disorders, which typically report scores ranging from 3.0 to 4.2 on the Mobile App Rating Scale and 6 to 10 on IQVIA [101,116-118]. However, the lower uMARS subjective quality score (2.6/5.0) indicated areas for enhancement. Technical raters valued AIrway's utility for condition management but showed limited interest in long-term use or payment. Future work will prioritize clinical validation studies co-designed with people living with asthma or COPD and their clinicians, using appropriately powered trials to assess safety and demonstrate meaningful clinical benefits before broader deployment.

Limitations and Future Directions

The usability study was also limited by a small sample size (n=5) compared to the 15-40 participants typically included in similar research [119-121]. This sampling limitation was attributed to recruitment challenges through university mailing lists and paid research groups on Facebook, which did not effectively reach the intended audience. In addition, the usability evaluation relied on subjective uMARS and IQVIA ratings, which are prone to perception variability and bias. A more comprehensive assessment could incorporate objective use metrics such as system interaction patterns (eg, login frequency and session duration), error occurrence rates, feature use statistics, and configuration preferences to provide a more accurate view of UXs and app performance.

At present, AIrway is Android-specific; thus, future development is needed for cross-platform frameworks like Flutter to improve accessibility and equity by extending availability to a broader user base. Our further accessibility review of the UIs recommended adjusting layout widths to accommodate dynamic text and increasing text sizes to a minimum of 48 density-independent pixels. Future studies should also include patients with asthma and COPD as well as their attending clinicians and caretakers to collect direct feedback from end users, ensuring that AIrway addresses the specific needs and preferences of its target populations.

Conclusions

This AIrway app followed mHealth best practices to ensure accessibility and seamless wireless integration with wearables. Future work will include cross-platform support and assess usability with patients to validate its clinical effectiveness.

Acknowledgments

This study was supported by the Fonds de Recherche du Québec—Santé (#322737), National Institute on Deafness and Other Communication Disorders (R01 DC021461), the Centre for Research on Brain, Language Incubator Award, and Canada Research

XSL•FO
RenderX

Chair research stipend (NYKL-J). This work was in part supported by the Bavarian State Ministry of Science and the Arts (AMK). The presented content is solely the responsibility of the authors and does not necessarily represent the official views of the above funding agencies. The authors would also like to thank Meghana Munipalle for her assistance in figure generation. The authors did not use generative artificial intelligence at any stage of the research process.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: AC (lead), NYKL-J (lead), RG, AMK
Formal analysis: AC
Fund acquisition: NYKL-J, AMK (equal)
Investigation: AC (lead), NYKL-J
Methodology: AC (lead), NYKL-J
Project administration: AC (lead), LM, NYKL-J
Resources: NYKL-J
Software: AC
Supervision: NYKL-J
Visualization: AC (lead), NYKL-J
Writing—original draft: AC (lead), NYKL-J
Writing—review and editing: AC (lead), NYKL-J, LM, RG, AMK

Conflicts of Interest

None declared.

Multimedia Appendix 1
App interfaces.
[DOCX File , 913 KB - formative_v9i1e73584_app1.docx]

Multimedia Appendix 2
Privacy policy for the app.
[DOCX File , 26 KB - formative_v9i1e73584_app2.docx]

Multimedia Appendix 3
Readability metrics and definition.
[DOCX File , 25 KB - formative_v9i1e73584_app3.docx]

Multimedia Appendix 4
Summary of zone calculation and action plan in AIrway.
[DOCX File , 26 KB - formative_v9i1e73584_app4.docx]

Multimedia Appendix 5
User Version of the Mobile Application Rating Scale questionnaire.
[DOCX File , 28 KB - formative_v9i1e73584_app5.docx]

Multimedia Appendix 6
IQVIA questionnaire.
[DOCX File , 24 KB - formative_v9i1e73584_app6.docx]

Multimedia Appendix 7
Accessibility analysis on AIrway app interfaces.
[DOCX File , 25 KB - formative_v9i1e73584_app7.docx]

Multimedia Appendix 8
Open-ended question responses of technical raters.

XSL•FO
RenderX

[DOCX File , 27 KB - formative_v9i1e73584_app8.docx]

References

1. Papi A, Brightling C, Pedersen SE, Reddel HK. Asthma. Lancet 2018;391(10122):783-800. [doi: 10.1016/S0140-6736(17)33311-1] [Medline: 29273246]
2. Ye C, Yuan L, Wu K, Shen B, Zhu C. Association between systemic immune-inflammation index and chronic obstructive pulmonary disease: a population-based study. BMC Pulm Med 2023;23(1):295 [FREE Full text] [doi: 10.1186/s12890-023-02583-5] [Medline: 37563621]
3. Wu C, Li G, Huang C, Cheng Y, Chen C, Chien J, et al. Acute exacerbation of a chronic obstructive pulmonary disease prediction system using wearable device data, machine learning, and deep learning: development and cohort study. JMIR Mhealth Uhealth 2021;9(5):e22591 [FREE Full text] [doi: 10.2196/22591] [Medline: 33955840]
4. Jafleh E, Alnaqbi F, Almaeeni H, Faqeeh S, Alzaabi MA, Al Zaman K. The role of wearable devices in chronic disease monitoring and patient care: a comprehensive review. Cureus 2024;16(9):e68921. [doi: 10.7759/cureus.68921] [Medline: 39381470]
5. Sanchez-Morillo D, Fernandez-Granero MA, Leon-Jimenez A. Use of predictive algorithms in-home monitoring of chronic obstructive pulmonary disease and asthma: a systematic review. Chron Respir Dis 2016;13(3):264-283. [doi: 10.1177/1479972316642365] [Medline: 27097638]
6. Kadambi P, Mohanty A, Ren H. Towards a wearable cough detector based on neural networks. 2018 Presented at: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); April 15-20, 2018; Calgary, AB, Canada p. 2161-2165. [doi: 10.1109/icassp.2018.8461394]
7. Wu Y, Li F, Xie Y, Wang Y, Yang Z. SymListener: detecting respiratory symptoms via acoustic sensing in driving environments. ACM Trans Sen Netw 2023;19(1):1-21. [doi: 10.1145/3517014]
8. Sun X, Lu Z, Hu W, Cao G. SymDetector: detecting sound-related respiratory symptoms using smartphones. 2015 Presented at: UbiComp '15: The 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing; September 7-11, 2015; Osaka, Japan p. 97-108. [doi: 10.1145/2750858.2805826]
9. Lei Z, Martignetti L, Ridgway C, Peacock S, Sakata JT, Li-Jessen NYK. Wearable neck surface accelerometers for occupational vocal health monitoring: instrument and analysis validation study. JMIR Form Res 2022;6(8):e39789 [FREE Full text] [doi: 10.2196/39789] [Medline: 35930317]
10. Zañartu M, Ho JC, Kraman SS, Pasterkamp H, Huber JE, Wodicka GR. Air-borne and tissue-borne sensitivities of bioacoustic sensors used on the skin surface. IEEE Trans Biomed Eng 2009;56(2):443-451. [doi: 10.1109/TBME.2008.2008165] [Medline: 19272887]
11. Mehta DD, Zañartu M, Feng SW, Cheyne HA, Hillman RE. Mobile voice health monitoring using a wearable accelerometer sensor and a smartphone platform. IEEE Trans Biomed Eng 2012;59(11):3090-3096 [FREE Full text] [doi: 10.1109/TBME.2012.2207896] [Medline: 22875236]
12. Groh R, Lei Z, Martignetti L, Li-Jessen NYK, Kist AM. Efficient and explainable deep neural networks for airway symptom detection in support of wearable health technology. Adv Intell Syst 2022;4(7):2100284. [doi: 10.1002/aisy.202100284]
13. Lei Z, Kennedy E, Fasanella L, Li-Jessen NY, Mongeau L. Discrimination between modal, breathy and pressed voice for single vowels using neck-surface vibration signals. Appl Sci (Basel) 2019;9(7):1505 [FREE Full text] [doi: 10.3390/app9071505] [Medline: 32133204]
14. Landry V, Matschek J, Pang R, Munipalle M, Tan K, Boruff J, et al. Audio-based digital biomarkers in diagnosing and managing respiratory diseases: a systematic review and bibliometric analysis. Eur Respir Rev 2025;34(176):240246 [FREE Full text] [doi: 10.1183/16000617.0246-2024] [Medline: 40368428]
15. Song Y, Yun I, Giovanoli S, Easthope CA, Chung Y. Multimodal deep ensemble classification system with wearable vibration sensor for detecting throat-related events. NPJ Digit Med 2025;8(1):14 [FREE Full text] [doi: 10.1038/s41746-024-01417-w] [Medline: 39775108]
16. Kang YJ, Arafa HM, Yoo J, Kantarcigil C, Kim J, Jeong H, et al. Soft skin-interfaced mechano-acoustic sensors for real-time monitoring and patient feedback on respiratory and swallowing biomechanics. NPJ Digit Med 2022;5(1):147 [FREE Full text] [doi: 10.1038/s41746-022-00691-w] [Medline: 36123384]
17. Lee HK, Park SU, Kong S, Ryu H, Kim HB, Lee SH, et al. Real-time deep learning-assisted mechano-acoustic system for respiratory diagnosis and multifunctional classification. npj Flex Electron 2024;8(1):69. [doi: 10.1038/s41528-024-00355-7]
18. Vijayan V, Connolly JP, Condell J, McKelvey N, Gardiner P. Review of wearable devices and data collection considerations for connected health. Sensors (Basel) 2021;21(16):5589. [doi: 10.3390/s21165589] [Medline: 34451032]
19. Ahmed A, Ali N, Aziz S, Abd-alrazaq AA, Hassan A, Khalifa M, et al. A review of mobile chatbot apps for anxiety and depression and their self-care features. Comput Methods Programs Biomed Update 2021;1:100012. [doi: 10.1016/j.cmpbup.2021.100012]
20. Liang Z, Tatha O, Andersen L. Developing mHealth app for tracking academic stress and physiological reactions to stress. 2020 Presented at: IEEE 2nd Global Conference on Life Sciences and Technologies (LifeTech); March 10-12, 2020; Kyoto, Japan p. 147-150. [doi: 10.1109/lifetech48969.2020.1570618580]

XSL•FO
RenderX

21. Beierle F, Allgaier J, Stupp C, Keil T, Schlee W, Schobel J, et al. Self-assessment of having COVID-19 with the corona check mHealth app. IEEE J Biomed Health Inform 2023;27(6):2794-2805. [doi: 10.1109/JBHI.2023.3264999] [Medline: 37023154]
22. Giebel GD, Abels C, Plescher F, Speckemeier C, Schrader NF, Börchers K, et al. Problems and barriers related to the use of mHealth apps from the perspective of patients: focus group and interview study. J Med Internet Res 2024;26:e49982 [FREE Full text] [doi: 10.2196/49982] [Medline: 38652508]
23. Rowland SP, Fitzgerald JE, Holme T, Powell J, McGregor A. What is the clinical value of mHealth for patients? NPJ Digit Med 2020;3:4 [FREE Full text] [doi: 10.1038/s41746-019-0206-x] [Medline: 31970289]
24. Belisario JSM, Huckvale K, Greenfield G, Car J, Gunn L. Smartphone and tablet self management apps for asthma. Cochrane Database Syst Rev 2013;2013(11):CD010013. [doi: 10.1002/14651858.CD010013.pub2] [Medline: 24282112]
25. Janjua S, Banchoff E, Threapleton C, Prigmore S, Fletcher J, Disler R. Digital interventions for the management of chronic obstructive pulmonary disease. Cochrane Database Syst Rev 2021;4(4):CD013246 [FREE Full text] [doi: 10.1002/14651858.CD013246.pub2] [Medline: 33871065]
26. Quach S, Michaelchuk W, Benoit A, Oliveira A, Packham TL, Goldstein R, et al. Mobile heath applications for self-management in chronic lung disease: a systematic review. Netw Model Anal Health Inform Bioinform 2023;12(1):25 [FREE Full text] [doi: 10.1007/s13721-023-00419-0] [Medline: 37305790]
27. Apple Watch. Apple. URL: https://www.apple.com/ca/watch/ [accessed 2023-05-10]
28. Fitbit. 2022. URL: https://www.fitbit.com/global/en-ca/home [accessed 2022-03-02]
29. Bernbaum A. Afflo. 2022. URL: https://www.jamesdysonaward.org/en-US/2019/project/afflo [accessed 2022-03-01]
30. Devani N, Pramono RXA, Imtiaz SA, Bowyer S, Rodriguez-Villegas E, Mandal S. Accuracy and usability of AcuPebble SA100 for automated diagnosis of obstructive sleep apnoea in the home environment setting: an evaluation study. BMJ Open 2021;11(12):e046803 [FREE Full text] [doi: 10.1136/bmjopen-2020-046803] [Medline: 34933855]
31. Morita PP, Yeung MS, Ferrone M, Taite AK, Madeley C, Stevens Lavigne A, et al. A patient-centered mobile health system that supports asthma self-management (breathe): design, development, and utilization. JMIR Mhealth Uhealth 2019;7(1):e10956 [FREE Full text] [doi: 10.2196/10956] [Medline: 30688654]
32. Unleash the power of voice as a vital sign. Sonde. 2023. URL: https://www.sondehealth.com/ [accessed 2023-04-15]
33. Harness the potential of cough as a vital sign. Hyfe. 2023. URL: https://www.hyfe.ai/ [accessed 2023-04-06]
34. Green paper on mobile health ("mHealth"). European Commission. 2024. URL: https://digital-strategy.ec.europa.eu/en/library/green-paper-mobile-health-mhealth [accessed 2024-06-30]
35. Policy for device software functions and mobile medical applications. US Food & Drug Administration. 2024. URL: https://tinyurl.com/bdhm2j2x [accessed 2024-06-30]
36. Human interface guidelines. Apple. 2023. URL: https://developer.apple.com/design/human-interface-guidelines [accessed 2023-06-08]
37. Typography. Android. 2023. URL: https://m1.material.io/style/typography.html [accessed 2023-06-08]
38. Health app use scenarios and HIPAA. HHS. URL: https://www.hhs.gov/sites/default/files/ocr-health-app-developer-scenarios-2-2016.pdf [accessed 2024-08-05]
39. Llorens-Vernet P, Miró J. Standards for mobile health-related apps: systematic review and development of a guide. JMIR Mhealth Uhealth 2020;8(3):e13057 [FREE Full text] [doi: 10.2196/13057] [Medline: 32130169]
40. Lei Z, Fasanella L, Martignetti L, Li-Jessen NY, Mongeau L. Investigation of vocal fatigue using a dose-based vocal loading task. Appl Sci (Basel) 2020;10(3):1192 [FREE Full text] [doi: 10.3390/app10031192] [Medline: 32133203]
41. Create XML layouts for android. Android. 2023. URL: https://developer.android.com/codelabs/basic-android-kotlin-training-xml-layouts#0 [accessed 2023-03-20]
42. Android Studio Chipmunk | 2021.2.1. Android. 2023. URL: https://developer.android.com/studio/releases/past-releases/as-chipmunk-release-notes [accessed 2023-03-03]
43. Morville P. User experience honeycomb. UXPin. 2022. URL: https://www.uxpin.com/studio/blog/ux-honeycomb-definition-and-use/ [accessed 2022-03-08]
44. Göransson C, Eriksson I, Ziegert K, Wengström Y, Langius-Eklöf A, Brovall M, et al. Testing an app for reporting health concerns—experiences from older people and home care nurses. Int J Older People Nurs 2018;13(2):e12181. [doi: 10.1111/opn.12181] [Medline: 29210218]
45. Creighton JR. Yogic Breathing for Post-Traumatic Stress Disorder: Designing an Application to Supplement Learning and Overcome a Stress State [Dissertation]. The University of Arizona. 2014. URL: https://repository.arizona.edu/bitstream/handle/10150/344450/azu_etd_13601_sip1_m.pdf?sequence=2 [accessed 2025-10-17]
46. Caldwell B, Cooper M, Reid LG. Web Content Accessibility Guidelines (WCAG) 2.0. WWW Consortium (W3C). 2008. URL: https://www.w3.org/TR/WCAG20/ [accessed 2025-10-17]
47. Vaghefi I, Tulu B. The continued use of mobile health apps: insights from a longitudinal study. JMIR Mhealth Uhealth 2019;7(8):e12983 [FREE Full text] [doi: 10.2196/12983] [Medline: 31469081]
48. Build a responsive UI with ConstraintLayout. Android. URL: https://developer.android.com/develop/ui/views/layout/constraint-layout [accessed 2022-03-03]
49. Color contrast checker. Coolors. 2022. URL: https://coolors.co/contrast-checker/112a46-acc8e5 [accessed 2022-03-05]

XSL•FO
RenderX

50. Request location permissions. Android. 2023. URL: https://developer.android.com/training/location/permissions#approximate-request [accessed 2023-03-05]
51. DatePicker. Android. 2023. URL: https://developer.android.com/reference/android/widget/DatePicker [accessed 2023-04-02]
52. Long R. The Canadian Lung Association/Canadian Thoracic Society and tuberculosis prevention and control. Can Respir J 2007;14(7):427-431 [FREE Full text] [doi: 10.1155/2007/612974] [Medline: 17948145]
53. Respiratory medications. The Lung Association Ontario. 2023. URL: https://taddlecreekfht.ca/wp-content/uploads/2014/10/Respiratory-Medications.pdf [accessed 2023-05-10]
54. van Kruijssen V, van Staa A, Dwarswaard J, In 't Veen JC, Mennema B, Adams SA. Use of online self-management diaries in asthma and COPD: a qualitative study of subjects' and professionals' perceptions and behaviors. Respir Care 2015;60(8):1146-1156. [doi: 10.4187/respcare.03795] [Medline: 25922542]
55. O'Connell S, McCarthy VJC, Queally M, Savage E. The preferences of people with asthma or chronic obstructive pulmonary disease for self-management support: a qualitative descriptive study. J Clin Nurs 2021;30(19-20):2832-2841. [doi: 10.1111/jocn.15790] [Medline: 33982317]
56. Juniper E, O'Byrne PM, Guyatt G, Ferrie P, King D. Development and validation of a questionnaire to measure asthma control. Eur Respir J 1999;14(4):902-907 [FREE Full text] [doi: 10.1034/j.1399-3003.1999.14d29.x] [Medline: 10573240]
57. Reddel HK, Bateman ED, Becker A, Boulet L, Cruz AA, Drazen JM, et al. A summary of the new GINA strategy: a roadmap to asthma control. Eur Respir J 2015;46(3):622-639 [FREE Full text] [doi: 10.1183/13993003.00853-2015] [Medline: 26206872]
58. Jones PW, Harding G, Berry P, Wiklund I, Chen W, Kline Leidy N. Development and first validation of the COPD Assessment Test. Eur Respir J 2009;34(3):648-654. [doi: 10.1183/09031936.00102509] [Medline: 19720809]
59. Gómez FP, Rodriguez-Roisin R. Global Initiative for Chronic Obstructive Lung Disease (GOLD) guidelines for chronic obstructive pulmonary disease. Curr Opin Pulm Med 2002;8(2):81-86. [doi: 10.1097/00063198-200203000-00001] [Medline: 11845001]
60. DeVries R, Kriebel D, Sama S. Validation of the breathlessness, cough and sputum scale to predict COPD exacerbation. NPJ Prim Care Respir Med 2016;26:16083 [FREE Full text] [doi: 10.1038/npjpcrm.2016.83] [Medline: 27906157]
61. TimePicker. Android. 2023. URL: https://developer.android.com/reference/android/widget/TimePicker [accessed 2023-05-20]
62. See K, Phua J, Lim T. Trigger factors in asthma and chronic obstructive pulmonary disease: a single-centre cross-sectional survey. Singapore Med J 2016;57(10):561-565 [FREE Full text] [doi: 10.11622/smedj.2015178] [Medline: 26768322]
63. What are the WHO Air quality guidelines? World Health Organization. URL: https://www.who.int/news-room/feature-stories/detail/what-are-the-who-air-quality-guidelines [accessed 2023-03-05]
64. Air Quality Health Index. Government of Canada. 2024. URL: https://www.canada.ca/en/environment-climate-change/services/air-quality-health-index/about.html [accessed 2024-05-10]
65. Personal Information Protection and Electronic Documents Act (S.C. 2000, c. 5). Government of Canada. 2023. URL: https://laws-lois.justice.gc.ca/eng/acts/p-8.6/ [accessed 2023-05-12]
66. Act respecting the protection of personal information in the private sector. Québec. 2023. URL: https://www.legisquebec.gouv.qc.ca/en/document/cs/p-39.1 [accessed 2023-05-13]
67. Firebase Firestore. Google. 2023. URL: https://firebase.google.com/docs/storage [accessed 2023-02-02]
68. HIPAA compliance on Google Cloud platform. Google. 2023. URL: https://cloud.google.com/security/compliance/hipaa [accessed 2023-02-04]
69. Adamovic M. Readability Calculator. Online-Utility.org. 2023. URL: https://www.online-utility.org/english/readability_test_and_improve.jsp [accessed 2023-03-10]
70. DuBay WH. Smart Language: Readers, Readability, and the Grading of Text. Germany: ERIC; 2007.
71. Readability formulas. Readable. 2023. URL: https://readable.com/readability/flesch-reading-ease-flesch-kincaid-grade-level/ [accessed 2023-03-14]
72. The Automated Readability Index (ARI). Readability. 2023. URL: https://readabilityformulas.com/automated-readability-index.php [accessed 2023-05-01]
73. Mobile accessibility at W3C. W3C. 2022. URL: https://www.w3.org/WAI/standards-guidelines/mobile/ [accessed 2022-06-28]
74. Di Gregorio M, Di Nucci D, Palomba F, Vitiello G. The making of accessible Android applications: an empirical study on the state of the practice. Empir Softw Eng 2022;27(6):145 [FREE Full text] [doi: 10.1007/s10664-022-10182-x] [Medline: 35949521]
75. Accessibility Scanner. Google. 2024. URL: https://play.google.com/store/apps/details?id=com.google.android.apps.accessibility.auditor&hl=en_CA&gl=US [accessed 2024-05-15]
76. Android Studio release notes. Android. 2023. URL: https://developer.android.com/studio/releases [accessed 2023-03-03]
77. Current weather and forecasts collection. OpenWeather. 2023. URL: https://openweathermap.org/price#current [accessed 2023-05-14]
78. Living with asthma and Air Quality Health Index. Asthma Canada. 2024. URL: https://tinyurl.com/749whjc6 [accessed 2024-05-11]

XSL•FO
RenderX

79. Han A, Deng S, Yu J, Zhang Y, Jalaludin B, Huang C. Asthma triggered by extreme temperatures: from epidemiological evidence to biological plausibility. Environ Res 2023;216(Pt 2):114489 [FREE Full text] [doi: 10.1016/j.envres.2022.114489] [Medline: 36208788]
80. Asthma Action Plan. Centers for Disease Control and Prevention. 2023. URL: https://www.cdc.gov/asthma/action-plan/documents/asthma-action-plan-508.pdf [accessed 2023-05-15]
81. Gift AG, Narsavage G. Validity of the numeric rating scale as a measure of dyspnea. Am J Crit Care 1998;7(3):200-204. [doi: 10.4037/ajcc1998.7.3.200]
82. My COPD Action Plan. Canadian Thoracic Society. 2023. URL: https://cts-sct.ca/wp-content/uploads/2019/03/5491_THOR_COPDActionPlanUpdate_2019_Editable_Eng_v2.pdf [accessed 2023-06-11]
83. Varshney H, Allahloh A, Sarfraz M. IoT based eHealth management system using Arduino and Google Cloud Firestore. 2019 Presented at: International Conference on Electrical, Electronics and Computer Engineering (UPCON); November 8-10, 2019; Aligarh, India p. 1-6. [doi: 10.1109/upcon47278.2019.8980238]
84. Fornari G, Minato R, Pilotto G, Palazzi C. Applying frugal innovation to humidity and temperature monitoring. 2020 Presented at: MobiCom '20: The 26th Annual International Conference on Mobile Computing and Networking; September 21, 2020; London, United Kingdom p. 12-17. [doi: 10.1145/3410670.3410860]
85. SQL vs. NoSQL databases: what's the difference? IBM. 2023. URL: https://www.ibm.com/cloud/blog/sql-vs-nosql [accessed 2023-06-01]
86. Ganesh Chandra D. BASE analysis of NoSQL database. Future Gener Comput Syst 2015;52:13-21. [doi: 10.1016/j.future.2015.05.003]
87. Sukmana Y, Rosmansyah Y. The use of Cloud Firestore for handling real-time data updates: an empirical study of gamified online quiz. 2021 Presented at: 2nd International Conference on Electronics, Communications and Information Technology (CECIT); December 27-29, 2021; Sanya, China p. 1239-1244. [doi: 10.1109/cecit53797.2021.00220]
88. Slingerland C. Horizontal vs. vertical scaling: which should you choose? Cloudzero. 2024. URL: https://tinyurl.com/3txkea5j [accessed 2024-07-31]
89. Firebase Authentication. Firebase. 2023. URL: https://firebase.google.com/docs/auth [accessed 2023-05-16]
90. Bluetooth technology overview. Bluetooth. 2023. URL: https://www.bluetooth.com/learn-about-bluetooth/tech-overview/ [accessed 2023-05-24]
91. Tosi J, Taffoni F, Santacatterina M, Sannino R, Formica D. Performance evaluation of bluetooth low energy: a systematic review. Sensors (Basel) 2017;17(12):2898 [FREE Full text] [doi: 10.3390/s17122898] [Medline: 29236085]
92. CC2540 and CC2541 Bluetooth Low Energy software developers guide. Texas Instruments. 2009. URL: https://www.ti.com/lit/ug/swru271i/swru271i.pdf [accessed 2025-09-18]
93. BLE protocol stack—controller. Pc Os. 2022. URL: https://pcng.medium.com/ble-protocol-stack-controller-2d2d5371deec [accessed 2022-12-18]
94. The Bluetooth Low Energy protocol stack: understanding the layers. NovelBits. 2022. URL: https://novelbits.io/bluetooth-low-energy-protocol-stack-layers/ [accessed 2022-12-14]
95. Chapter 4. GATT (services and characteristics). Oreilly. 2022. URL: https://www.oreilly.com/library/view/getting-started-with/9781491900550/ch04.html [accessed 2022-12-09]
96. Bluetooth Low Energy. Android Developers. 2023. URL: https://developer.android.com/guide/topics/connectivity/bluetooth/ble-overview [accessed 2023-06-02]
97. BluetoothDevice. Android Developers. 2023. URL: https://developer.android.com/reference/android/bluetooth/BluetoothDevice [accessed 2023-06-02]
98. BluetoothGatt. Android Developers. 2023. URL: https://developer.android.com/reference/android/bluetooth/BluetoothGatt [accessed 2023-06-02]
99. nRF52840 DK. Nordic Semiconductor. 2023. URL: https://www.nordicsemi.com/Products/Development-hardware/nrf52840-dk [accessed 2023-06-05]
100. Stoyanov SR, Hides L, Kavanagh DJ, Wilson H. Development and validation of the user version of the Mobile Application Rating Scale (uMARS). JMIR Mhealth Uhealth 2016;4(2):e72 [FREE Full text] [doi: 10.2196/mhealth.5849] [Medline: 27287964]
101. Choi YK, Demiris G, Lin S, Iribarren SJ, Landis CA, Thompson HJ, et al. Smartphone applications to support sleep self-management: review and evaluation. J Clin Sleep Med 2018;14(10):1783-1790 [FREE Full text] [doi: 10.5664/jcsm.7396] [Medline: 30353814]
102. Firebase App Distribution. Firebase. 2023. URL: https://firebase.google.com/docs/app-distribution [accessed 2023-06-18]
103. Flesch reading ease. Flesch-Kincaid Calculator. 2023. URL: https://charactercalculator.com/flesch-reading-ease/ [accessed 2023-06-10]
104. What is a Gunning Fog Index readability score? Readable. 2023. URL: https://readable.com/readability/gunning-fog-index/ [accessed 2023-03-14]
105. The Coleman-Liau Readability Index. Readable. 2023. URL: https://readable.com/readability/coleman-liau-readability-index/ [accessed 2023-03-14]

XSL•FO
RenderX

106. Flesch-Kincaid grade level. Readable. 2023. URL: https://readable.com/readability/flesch-reading-ease-flesch-kincaid-grade-level/ [accessed 2023-03-14]
107. The SMOG Readability Formula. Readability. 2023. URL: https://readabilityformulas.com/smog-readability-formula.php [accessed 2023-05-01]
108. Readability Formulas. Readability. 2023. URL: https://readabilityformulas.com/aboutus.php [accessed 2023-05-01]
109. Understanding RSSI. metageek. URL: https://www.metageek.com/training/resources/understanding-rssi/ [accessed 2025-09-18]
110. Himes BE, Leszinsky L, Walsh R, Hepner H, Wu AC. Mobile health and inhaler-based monitoring devices for asthma management. J Allergy Clin Immunol Pract 2019;7(8):2535-2543 [FREE Full text] [doi: 10.1016/j.jaip.2019.08.034] [Medline: 31706485]
111. Quach S, Benoit A, Oliveira A, Packham TL, Goldstein R, Brooks D. Features and characteristics of publicly available mHealth apps for self-management in chronic obstructive pulmonary disease. Digit Health 2023;9:20552076231167007 [FREE Full text] [doi: 10.1177/20552076231167007] [Medline: 37065541]
112. Sobnath DD, Philip N, Kayyali R, Nabhani-Gebara S, Pierscionek B, Vaes AW, et al. Features of a mobile support app for patients with chronic obstructive pulmonary disease: literature review and current applications. JMIR Mhealth Uhealth 2017;5(2):e17 [FREE Full text] [doi: 10.2196/mhealth.4951] [Medline: 28219878]
113. Alsahli S, Hor SY, Lam M. Factors influencing the acceptance and adoption of mobile health apps by physicians during the COVID-19 pandemic: systematic review. JMIR mHealth and uHealth 2023;e50419:11. [doi: 10.2196/50419] [Medline: 37938873]
114. Hegde S, Eid NS. Telehealth and remote patient monitoring after the COVID pandemic. Pediatr Allergy Immunol Pulmonol 2021;34(4):130-131 [FREE Full text] [doi: 10.1089/ped.2021.0162] [Medline: 34724847]
115. Mobbs RJ, Ho D, Choy WJ, Betteridge C, Lin H. COVID-19 is shifting the adoption of wearable monitoring and telemedicine (WearTel) in the delivery of healthcare: opinion piece. Ann Transl Med 2020;8(20):1285 [FREE Full text] [doi: 10.21037/atm-20-3678] [Medline: 33209865]
116. Sunjaya AP, Sengupta A, Martin A, Di Tanna GL, Jenkins C. Efficacy of self-management mobile applications for patients with breathlessness: systematic review and quality assessment of publicly available applications. Respir Med 2022;201:106947 [FREE Full text] [doi: 10.1016/j.rmed.2022.106947] [Medline: 36037561]
117. Schmeelk S, Davis A, Li Q, Shippey C, Utah M, Myers A, et al. Monitoring symptoms of COVID-19: review of mobile apps. JMIR Mhealth Uhealth 2022;10(6):e36065 [FREE Full text] [doi: 10.2196/36065] [Medline: 35609313]
118. Diaz-Skeete YM, McQuaid D, Akinosun AS, Ekerete I, Carragher N, Carragher L. Analysis of apps with a medication list functionality for older adults with heart failure using the mobile app rating scale and the IMS Institute for Healthcare Informatics functionality score: evaluation study. JMIR Mhealth Uhealth 2021;9(11):e30674 [FREE Full text] [doi: 10.2196/30674] [Medline: 34726613]
119. Balebako R, Marsh A, Lin J, Hong J, Cranor L. The privacy and security behaviors of smartphone app developers. 2014 Presented at: Workshop on Usable Security; August 10-14, 2014; Seattle, WA. [doi: 10.14722/usec.2014.23006]
120. Fuller-Tyszkiewicz M, Richardson B, Klein B, Skouteris H, Christensen H, Austin D, et al. A mobile app-based intervention for depression: end-user and expert usability testing study. JMIR Ment Health 2018;5(3):e54 [FREE Full text] [doi: 10.2196/mental.9445] [Medline: 30139722]
121. Mohamad Marzuki MF, Yaacob NA, Bin Yaacob NM, Abu Hassan MR, Ahmad SB. Usable mobile app for community education on colorectal cancer: development process and usability study. JMIR Hum Factors 2019;6(2):e12103 [FREE Full text] [doi: 10.2196/12103] [Medline: 30990454]

Abbreviations

ACQ: Asthma Control Questionnaire
AQHI: Air Quality Health Index
BLE: Bluetooth Low Energy
COPD: chronic obstructive pulmonary disease
FAQ: frequently asked question
GINA: Global Initiative for Asthma
HIPAA: Health Insurance Portability and Accountability Act
mHealth: mobile health
NSA: neck surface accelerometer
PM: particulate matter
UI: user interface
uMARS: User Version of the Mobile Application Rating Scale
UX: user experience

XSL•FO
RenderX

Edited by A Mavragani; submitted 07.03.25; peer-reviewed by C Baxter; comments to author 01.05.25; revised version received 13.07.25; accepted 21.07.25; published 29.10.25.

Please cite as:
Chao A, Martignetti L, Groh R, Kist AM, Li-Jessen NYK
An mHealth App and System Architecture for Respiratory Disease Management: Design Principles, Tool Development, and Pilot Usability Study
JMIR Form Res 2025;9:e73584
URL: https://formative.jmir.org/2025/1/e73584
doi:10.2196/73584
PMID:

©Andrew Chao, Lisa Martignetti, René Groh, Andreas M Kist, Nicole YK Li-Jessen. Originally published in JMIR Formative Research (https://formative.jmir.org), 29.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Detecting Laterality Errors in Combined Radiographic Studies by Enhancing the Traditional Approach With GPT-4o: Algorithm Development and Multisite Internal Validation

Kung-Hsun Weng[1], MSc, MD; Yi-Chen Chou[1], MD; Yu-Ting Kuo[1,2,3], MD, PhD; Tsyh-Jyi Hsieh[1], MD; Chung-Feng Liu[4], PhD

[1]Department of Medical Imaging, Chi Mei Medical Center, Tainan, Taiwan
[2]Institute of Precision Medicine, College of Medicine, National Sun Yat-sen University, Kaohsiung, Taiwan
[3]Department of Radiology, School and College of Medicine, Kaohsiung Medical University, Kaohsiung, Taiwan
[4]Department of Medical Research, Chi Mei Medical Center, Tainan, Taiwan

Corresponding Author:
Chung-Feng Liu, PhD
Department of Medical Research
Chi Mei Medical Center
No 901, Zhonghua Road, Yongkang District
Tainan, 710
Taiwan
Phone: 886 62812811 ext 52590
Email: chungfengliu@gmail.com

Abstract

Background: Laterality errors in radiology reports can endanger patient safety. Effective methods for screening for laterality errors in combined radiographic reports, which combine multiple studies into one, remain unexplored.

Objective: First, we define and analyze the unstudied combined radiographic report format and its challenges. Second, we introduce a clinically deployable ensemble method (rule-based+GPT-4o), evaluated on large-scale, real-world, imbalanced data. Third, we demonstrate significant performance gaps between real-world imbalanced and synthetic balanced datasets, highlighting limitations of the benchmarking methodology commonly used in current studies.

Methods: This retrospective study analyzed deidentified English radiology reports containing laterality terms in order. We split the data into TrainVal (combined training and validation dataset), Test-1 (both real-world, imbalanced), and Test-2 (synthetic, balanced). Test-1 comes from a distinct branch. Experiment 1 compared the baseline, workaround, and GPT-4o-augmented rule-based methods. Experiment 2 compared the rule-based method with the highest recall to fine-tuned RoBERTa, ClinicalBERT, and GPT-4o models.

Results: As of July 2024, our dataset included 10,000 real-world and 889 synthetic radiology reports. The laterality error rate in real-world reports was 1.20% (120/10,000), significantly higher in combined (103/7000, 1.47%) than in noncombined reports (17/3000, 0.57%; difference=0.90%; z=3.81; P<.001). In experiment 1, recall differed significantly among the 3 versions of rule-based methods (Q=6.0; P=.0498, Friedman test). The rule-based+GPT-4o method had the highest recall (average rank=1), significantly better than the baseline (average rank=3; P=.04, Nemenyi test). Most (5/6) of the false positives introduced by the GPT-4o information extraction were due to parser limitations hidden by error cancellation. In experiment 2, on Test-1, rule-based+GPT-4o (precision=0.696; recall=0.889; F_1-score=0.780) outperformed GPT-4o (precision=0.219; recall=0.889; F_1-score=0.352), ClinicalBERT (precision=0.047; recall=0.667; F_1-score=0.088), and RoBERTa (F_1-score=0.000). On Test-2, rule-based+GPT-4o (precision=0.996; recall=0.925; F_1-score=0.959) and GPT-4o (precision=0.979; recall=0.953; F_1-score=0.966) outperformed ClinicalBERT (precision=0.984; recall=0.749; F_1-score=0.851) and RoBERTa (F_1-score=0.013). Both ClinicalBERT and GPT-4o exhibited notable declines in precision on TrainVal and Test-1 compared to Test-2. Both Test-1 data membership (GPT-4o: odds ratio [OR] 239.89, 95% CI 111.05-518.01; P<.001; ClinicalBERT: OR 1924.07, 95% CI 687.46-5383.99; P<.001) and order count per study (GPT-4o: OR 1.79, 95% CI 1.38-2.31; P<.001; ClinicalBERT: OR 2.50, 95% CI 1.64-3.80; P<.001) independently predicted false positive errors in multivariate logistic regression. In subgroup analysis, all models showed reduced precision and F_1 in combined-study subgroups.

XSL•FO
RenderX

Conclusions: The combined radiographic report format poses distinct challenges for both radiology report quality assurance and natural language processing. The combined rule-based and GPT-4o method effectively screens for laterality errors in imbalanced real-world reports. A significant performance gap exists between balanced synthetic datasets and imbalanced real-world data. Future studies should also include real-world imbalanced data.

(JMIR Form Res 2025;9:e76384) doi:10.2196/76384

KEYWORDS

natural language processing; large language model; radiographic reports; laterality error; radiology report; quality assurance; rule-based methods; artificial intelligence; deep learning; electronic health records

Introduction

Background

Radiographic reports play a crucial role in modern medical diagnostics. Although imaging technologies like computed tomography (CT) and magnetic resonance imaging (MRI) have advanced significantly, radiographic studies still account for the most significant proportion of examinations performed in the radiology department, with chest x-ray remaining the most common radiology study worldwide [1]. Meanwhile, errors in radiographic reports are also an important source of diagnostic errors in radiology [2], highlighting the importance of reporting quality.

Errors in radiology reports include more than missed lesions. Previous studies have systematically classified potential errors in radiology reports. For example, Kim and Mansfield [2] identified 12 primary error types, while Pinto et al [3] grouped them into 4 distinct categories. A common type of error is when radiologists identify lesions but make mistakes in interpreting them.

Laterality error is a type of interpretation error that can occur in radiology reports. This error occurs when radiologists confuse the sides of a lesion, such as mistaking the left side for the right side or vice versa. Such errors can be misleading and lead to wrong-side invasive procedures or even wrong-side surgeries [4-10]. Laterality errors have the potential to cause patient disability or even death, such as removing a patient's only functioning kidney [11], performing a wrong-side amputation [12], or removal of the healthy, cancer-free testicle [13]. In surgical practice, wrong-side surgery is considered a "never event" [14]. For the health care system, such errors are often classified as malpractice, leading to lawsuits, compensation, and diminished trust between patients and providers [15].

Detecting laterality errors in radiographic reports can be complicated by the adaptation of combined reporting. A combined report integrates the results of multiple radiology examinations into a single assessment. For example, in breast cancer screening, combining mammography and ultrasound offers a complementary approach to cancer detection, and combining all findings into 1 report helps reduce confusion [16]. Medical institutions can also integrate radiographic studies according to their policies and requirements. For example, a trauma study could include simultaneous knee, leg, femur, and pelvis radiographs, all interpreted by 1 radiologist in 1 report, providing a comprehensive view of the patient's injuries.

Although the combined results can be more concise for human readers to understand and communicate, they pose challenges for natural language processing (NLP), including the detection of laterality errors. Images of different anatomical parts, views, and even sides may mix in a single radiology report. Such complexity can render naive algorithms ineffective, such as checking that only "left" appears in the report or vice versa. The problem is further exacerbated by the free-typing nature of radiology reports. Radiologists have the flexibility to write combined reports without using a paragraph heading, standardized keywords, or mandatory paragraph segmentation. Currently, no literature addresses laterality errors in combined radiographic reports or their detection using NLP.

Review of Literature

Current research on laterality errors in radiology reports is sparse. Existing research has used various algorithms to identify laterality error in radiology reports, including keyword matching or regular expression [17-19] and deep learning approaches based on ClinicalBERT [20], GPT-4 [21], and multimodal large language model (LLM) [22].

Previous studies have relied on keyword- or regular expression–based screening of text, often paired with checks against structured metadata. Sangwaiya et al [4] reported addendum-corrected laterality error rate and error rate in a separate 1-month cohort identified by simple "left or right" keyword filtering, followed by manual review. Minn et al [19] developed a keyword- and regular expression–based system that compares report text against Health Level 7 metadata to flag gender and laterality mismatches for immediate correction. Sheehan et al [18] developed an automated Structured Query Language e-trigger applied to electronic health record–derived imaging-order data, which matched structured laterality modifiers against the free-text clinical history to detect potential wrong-side orders. Landau et al [23] developed a syntax-highlighting approach that allows radiologists to review laterality-indicating terms before finalizing the report. Collectively, these approaches use relatively simple rule-based matching in structured data and free text.

Recent studies have applied deep learning and LLMs to detect errors in reports, including laterality errors. Min et al [20] used a fine-tuned ClinicalBERT [24] binary classifier to detect inconsistencies, including lesion locations, between findings and conclusions in chest x-ray reports, evaluating largely on synthetically generated errors plus a smaller, less balanced, manually edited set. Gertz et al [21] introduced 5 common error types, including side confusion, into radiographic and

XSL•FO
RenderX

cross-sectional reports and found that GPT-4 underperformed the best-performing radiologist overall but was comparable to other radiologists. Kathait et al [22] used a multimodal GPT-4 family model to validate CT, MRI, and ultrasound reports flagged by commercial NLP tools for potential laterality errors, serving as secondary validation rather than a primary detector. These studies show promise of deep learning and LLMs in analyzing complex radiology reports, such as CT and MRI, but often rely on synthetic or balanced datasets, with limited evaluation as primary screening tools in real-world, imbalanced settings.

To the best of our knowledge, existing studies are limited to single-modality radiology reports and do not specifically address scenarios involving combined reports [17-23], in which different anatomical parts and laterality can mix together.

Aim of Our Study

Our study aims to advance the detection of laterality errors in radiology reports in 3 ways. First, we define and characterize the combined radiographic report, a significant yet unstudied format in our region. We showed that combined radiographic reports are associated with a significantly higher laterality error rate and lower detection performance of state-of-the-art NLP models compared to standard noncombined reports. Second, we introduce a clinically deployable ensemble approach integrating rule-based methods with GPT-4o, rigorously evaluated on 10,000 authentic, highly imbalanced reports. Third, we quantify and contrast the performance gap between synthetic and real-world datasets and elucidate the limitations of the evaluation methodologies commonly used in prior studies.

Methods

Study Design

We conducted a retrospective, multisite, internal validation study using deidentified electronic health records, including radiographic reports and metadata, from 3 hospital branches between 2012 and 2024.

Participant Flow

Eligible studies included English radiographic reports with orders specifying laterality, and Chinese labor health examinations were excluded. The data were split into 3 datasets without overlapping: TrainVal (combined training and validation dataset), Test-1 (both real-world, imbalanced), and Test-2 (balanced, synthetic). Test-1 came from a distinct branch.

Intervention

All evaluations followed a fixed batch evaluation pipeline: inputs were the full radiology report and the corresponding order, processing used either rule-based methods (baseline, manual workarounds, or GPT-4o–assisted information extraction) or pure deep-learning comparators (fine-tuned RoBERTa, ClinicalBERT, or GPT-4o few-shot), and the output was a report-level binary prediction (positive if the report contained at least 1 laterality error).

Data Collection and Analysis

The primary outcomes were report-level precision, recall, and F_1-score, both overall and by combined versus noncombined subgroups. Secondary analyses included qualitative error reviews and a multivariable logistic regression to identify predictors of false positives for GPT-4o and ClinicalBERT on the Test-1 and Test-2 datasets.

Feasibility and Acceptability Assessment

We assessed feasibility by performing end-to-end execution of our methods using both synthetic, balanced data and cross-site, real-world, imbalanced data. As a method-oriented formative study, we did not conduct formal acceptability assessments.

Ethical Considerations

The institutional review board at Chi Mei Hospital approved this study (11305-008). This study involves a retrospective analysis of deidentified electronic medical records (no direct identifiers were retained), eliminating the need to obtain informed consent from participants. No participant compensation was provided.

Data Collection

We obtained the data in this study from the 3 branches of the Chi Mei hospital: the Yongkang branch (medical center, 1296 beds), the Liuying branch (regional teaching hospital, 895 beds), and the Jiali branch (regional teaching hospital, 341 beds). The 3 branches employ 25, 11, and 3 full-time radiologists, respectively. In 2024, the approximate number of radiographic studies for the 3 hospital branches was approximately 250,000, 160,000, and 70,000, respectively.

Figure 1 depicts the flowchart of the study. The inclusion criteria for this study include radiological examinations performed in the 3 branches of our institution from 2012 to 2024. We included radiographic (x-ray) reports that are written in English. In addition, the relevant medical orders must indicate the laterality of the imaging site using the keywords “Lt,” “Rt,” “left,” or “right.” Only cases that meet all these conditions were included. Conversely, the exclusion criteria are Chinese labor health examination reports; all cases meeting this criterion were excluded from the study.

Prior to initiating our experimental procedures, all data were subjected to a deidentification process. This involved the removal of sensitive information that could potentially identify patients, such as medical record numbers and accession numbers.

Our research includes radiographic studies with single or multiple medical orders per study. For each radiographic study that involved multiple medical orders, the radiologists identified whether the report described the findings for each order separately or combined findings related to multiple medical orders into a single paragraph.

XSL•FO
RenderX

Figure 1. Research flow. TrainVal: combined training and validation dataset.

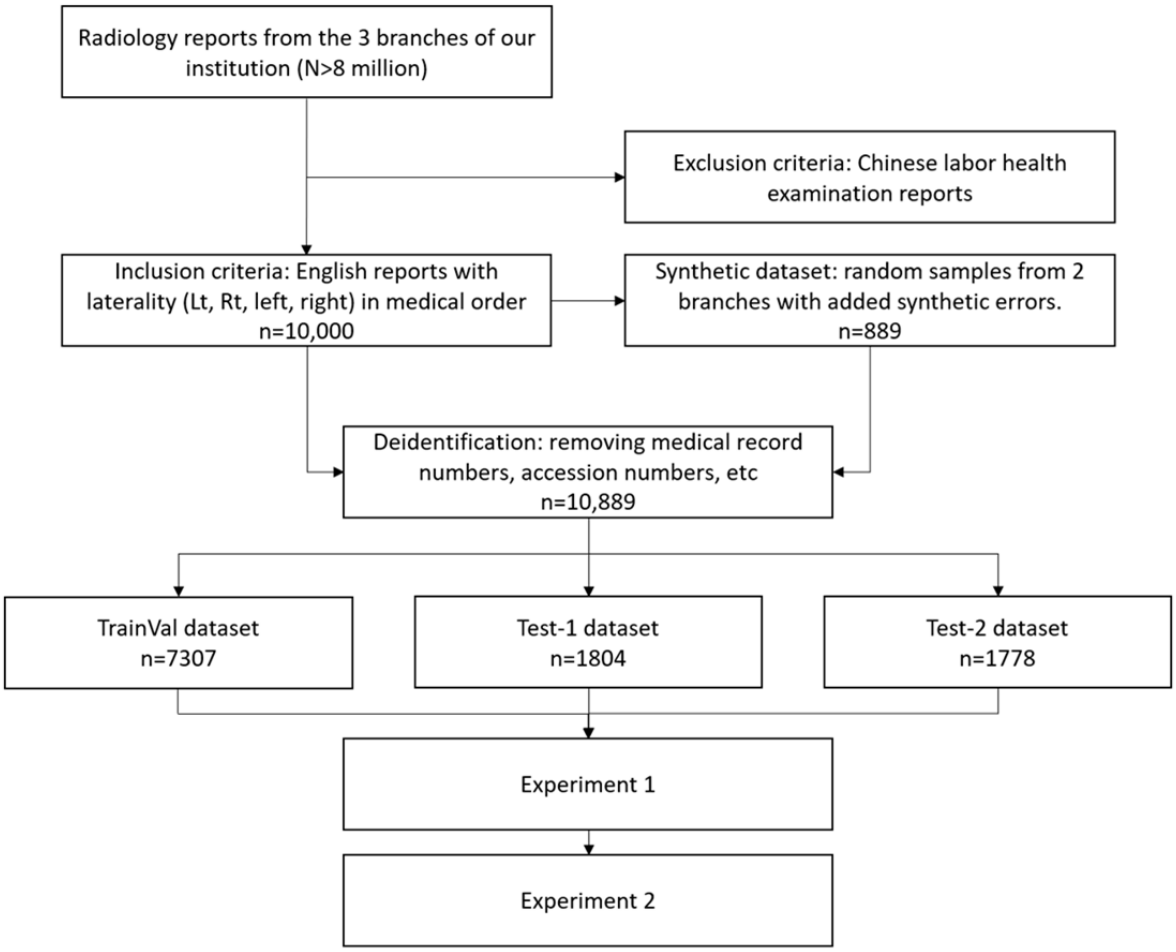

Dataset Preparation

Overview

We categorized the data into 3 datasets: TrainVal, the first test dataset (Test-1), and the second test dataset (Test-2). We used 2 test datasets to investigate the impact of different factors on the model's performance. There is no overlap between the data in the 3 datasets.

TrainVal Dataset

The TrainVal datasets include data from the Yongkang and Jiali branches of the Chimei Hospital, but not from the Liuying branch.

Test-1 Dataset

The Test-1 dataset contains data solely from the Liuying branch to evaluate the model's generalization capability across different hospitals and variations in physician reporting styles.

Test-2 Dataset

The Test-2 dataset was created with additional simulated data to improve class balancing and explore the effect of class balancing on model performance. We selected a subset of data free of laterality errors from the Yongkang and Jiali branches. KHW manually introduced errors by randomly reversing one occurrence of "Lt," "left," "Rt," and "right" in the report text. These manually patched data were combined with the unmodified versions to form the Test-2 dataset. The panel of experts discussed the size of the Test-2 dataset.

Data Annotation

A board-certified radiologist annotated all reports in this study. We did not perform double annotation or adjudication due to our sample size (10,000 real-world studies) and resource constraints. To assess interrater agreement, a second radiologist independently annotated subsamples, which we selected using label-based stratified random sampling with supersampling of the minority class, without access to the original labels or class distributions.

We focused on laterality errors that were clearly suspicious from reviewing physicians' orders and radiology reports in accordance with our institution's x-ray imaging protocols, such as reporting a right tibia fracture on a left knee x-ray. We classified ambiguous cases as having no laterality error because confirming them would require an image review that is beyond the scope of this study. Multimedia Appendix 1 provides examples of clearly suspicious errors.

Our research investigated discrepancies between the laterality of the lesion detailed in the reports and the laterality specified in the physicians' orders, as well as discrepancies of laterality within the reports themselves. We do not identify errors other than laterality errors.

XSL•FO
RenderX

Our study examines the combined radiographic reports. Each report in our study could contain multiple medical orders. Some medical orders may contain multiple anatomical parts. For example, "L-S Spine anteroposterior+Lat" contains both the lumbar and sacral spines.

The format of radiographic reports is entirely free typing in our hospital. There is no requirement to use headings to indicate which medical order or anatomical part corresponds to the findings in each paragraph.

Table 1 presents a simulated example illustrating the detection of laterality errors in combined reports in this study. This example presumes the physician's orders encompassed a chest x-ray in posteroanterior view and a left shoulder in anteroposterior view. The right clavicle fracture at line 2 is reliable and does not represent a laterality error, as a chest x-ray inherently captures both clavicles. Consequently, it is justifiable for the radiologist to document a fracture of the right clavicle. The right clavicle fracture at line 4 is incorrect, as it is in the paragraph referring to the left shoulder x-ray, which cannot depict a right clavicle fracture. The heading "Rt shoulder anteroposterior view" at line 5 represents a laterality error, as the original order indicated a left shoulder anteroposterior view, and the validity of line 6 cannot be determined.

Table 1. Simulated example of a combined radiographic report[a].

Line number	X-ray report
1	Chest posteroanterior view and left shoulder anteroposterior view shows:
2	Right clavicle fracture.
3	Left shoulder anteroposterior view shows:
4	Right clavicle fracture.
5	Right shoulder anteroposterior view shows:
6	Right clavicle fracture.[b]
7	Right rib fracture.

[a]Assuming the order for this single study includes a chest x-ray posteroanterior view and a left shoulder anteroposterior view. Right clavicle at line 4 is incompatible with the left shoulder at paragraph heading at line 3, and right shoulder at line 5 is incompatible with the original order of left shoulder and are both invalid. Line 7 does not use a paragraph heading, but the right rib fracture can be explained by the chest x-ray and therefore is valid.

[b]New paragraph begins below (end of right shoulder paragraph).

Model Development

Rule-Based Dictionary Look-Up Method

We implemented the baseline parser using an approach similar to a finite state machine (see Multimedia Appendix 2 for details on the parser implementation and an example). The parser scans the orders and reports term by term, identifying terms that indicate laterality, anatomy, or findings. It then checks for laterality based on the scanned terms and the expert rules.

We encoded our expert rules as key-value pairs in a dictionary format. The dictionary key represents an anatomical part, while the value is a list of possible body parts or findings associated with that part. For example, the key "knee" includes the values "femur" and "tibia," while the key "chest" includes "clavicle" and "ribs." Values in different keys may overlap; for example, "Fibula" may appear in values of keys "knee" and "leg."

Radiologist KHW manually curated and refined the expert rules based on TrainVal results (see Multimedia Appendix 2 for examples of expert rule modification). For each false positive, the radiologist identified and adjusted problematic terms in the expert rules, such as removing the term or assigning it to the appropriate anatomical part. For each false negative, the radiologist reviewed the report text to identify missed errors and updated the rules with relevant terms.

To reduce the expert workload, automated methods were used to generate initial drafts of expert rules, which the experts then manually refined and validated. We segmented reports into paragraphs using keyword-matched subheadings (eg, lines ending with a colon) and inferred the relevant anatomical part for each section from subheading keywords (eg, "chest" and "knee"). After lemmatizing and removing stop words with Natural Language Toolkit, we identified potential laterality-dependent terms by extracting words from 2 words before and 3 words after laterality indicators (eg, "Lt," "Rt," and "bilateral"). These terms, paired with their respective anatomical parts, formed key-value pairs for the draft expert rules.

Besides the baseline version, we implemented 2 additional approaches to improve the rule-based method: first, using GPT-4o information extraction for report preprocessing, and second, having experts create workarounds in the parser based on individual physician writing styles.

In the first approach, GPT-4o was used to extract anatomical locations, such as the names of bones, joints, and organs, along with the laterality from radiology reports, standardize terminology, and remove extraneous, potentially confounding details. We also instructed the GPT-4o to express all results in singular nouns whenever possible.

We hypothesize that using GPT for anatomical information extraction can standardize expressions across various styles of report texts, thereby enhancing the robustness of expert rules. For instance, terms such as "Rt genu valgum," "Rt genu varum," and even misspelled versions like "genuvarum, Rt." and "genuvalgum, Rt." can be accurately interpreted by GPT as pertaining to the right knee and captured by the same expert

XSL•FO
RenderX

rule of "knee." After extracting anatomical information from the reports using GPT-4o, we used this extracted information as parser input when appropriate.

In the second approach, KHW analyzed the TrainVal report dataset to identify common nonstandard expressions, specifically inverted writing styles such as "genuvarum, Rt," and create specific workarounds in the parser. These workarounds may not apply to other hospitals or radiologists.

Fine-Tuned RoBERTa and ClinicalBERT With Noisy, Automatically Generated Labels

In this experiment, we approached the task as a sentence contradiction task. We used radiologists' reports and combined them with medical orders, with the view information removed, leaving only the anatomical part, to form sentence pairs. We mapped the laterality errors as contradictions in the sentences. If there is no laterality error, it is considered a noncontradiction and vice versa. In addition to the gold labels annotated by radiologists, we included automatically generated samples with laterality errors in the model training to ensure proper class balancing.

For each report in the training data, we patched 1 of every 2 words indicating left and right, such as "Lt" and "Rt," alternatively, changing from left to right and vice versa. This modification can create inconsistencies between the paragraph text and the heading and between the paragraph heading and the medical orders. The training program labeled these automatically generated samples as having laterality errors and included them in the training data.

We used the RoBERTa large model fine-tuned on the Multi-Genre Natural Language Inference corpus, which focuses on the sentence contradiction task and evaluates a model's ability to determine whether a premise sentence entails, contradicts, or is neutral concerning a hypothesis sentence [25]. The hyperparameters used in this study are as follows: the learning rate is 2×10^{-5}, the batch size for training per device is 16, and the batch size for evaluation per device is also 16. The Adam optimizer was used with the weight decay parameter set to 0.01. We capped training at 60 epochs and used early stopping: training stopped if the validation metric failed to improve for 3 consecutive epochs. We used 20% of the TrainVal for validation.

For ClinicalBERT, the hyperparameters are as follows: the learning rate is 5×10^{-5}, and the batch size for training and evaluation per device is 32. All other hyperparameters remain consistent with those used in the RoBERTa experiments.

GPT-4o With Few-Shot Prompting

In our experiments with LLMs, we used GPT-4o and approached the task as a binary text classification problem with few-shot and chain-of-thought prompting [26] and in-context learning. The task of the LLM was to categorize each report as either containing or not containing laterality errors.

We provided examples for assessing laterality errors, including abstract conceptual and 4 concrete examples [27] using simulated reports and physician orders. We provided the radiologist's hand-crafted step-by-step reasoning alongside responses for each example. We required the model to use the chain-of-thought technique and output its step-by-step reasoning process, reflecting the examples provided by the radiologist in the prompt, regardless of the final prediction of the model.

Common Model Implementation Details

The models were optimized solely based on the TrainVal, without using the Test-1 or Test-2 datasets, including the creation and optimization of algorithms and dictionaries for expert rule-based methods, the training data for the RoBERTa and ClinicalBERT models, and the radiology reports referenced for GPT prompting.

To improve the GPT's performance and reproducibility of the results, we used the following methodology in all cases: first, we instructed the GPT to role-play as a radiologist [28]. Next, we specified the abbreviations and conventions commonly used in our institution. Finally, we set the temperature to 0 during the application programming interface call and used a fixed seed of 5566.

The view information was removed, leaving only the anatomical part whenever the models needed to refer to medical order information in the experiments, either through chain-of-thought reasoning guided by our prompts in the GPT-4o with few-shot prompting experiment or via dictionary 1-way mapping in other experiments. For example, different views of a chest x-ray, such as posteroanterior, lordotic, oblique, and decubitus, were all simplified to "chest."

This design greatly simplified the creation of expert rules for the expert rule-based methods. Rules only need to be created for each anatomical part without considering different combinations of views. This approach forced the deep learning methods to focus on the anatomical part rather than the view information.

Evaluation

Overview

We frame laterality error detection as a binary text classification task. For each x-ray study, the model ingests the full radiology report together with all associated medical orders and predicts whether at least 1 laterality error is present; a single study may be linked to one or more orders. The rule-based method additionally requires expert rules. The model outputs whether the report contains any laterality error.

We conducted 2 experiments.

First Experiment

In the first experiment, we developed a rule-based dictionary look-up method and compared it with 2 optimized versions using hand-crafted workarounds and GPT-4o information extraction for text preprocessing.

Second Experiment

In the second experiment, we compared the method with the best recall from the first experiment with pure deep learning approaches, including fine-tuned RoBERTa and ClinicalBERT with noisy labels and GPT-4o with few-shot prompting.

XSL•FO
RenderX

We conducted subgroup analyses to compare model performance in noncombined and combined studies. Test-1 and Test-2 data were merged to avoid complete class absence in specific subgroups while maintaining the imbalanced distribution characteristic of real-world data. We reported primary metrics separately for each test set and used the combined data only for subgroup analyses.

We gathered all cases, in which ClinicalBERT made highly confident predictions (softmax probability≥0.9) that were incorrect. Using the transformer-interpret package, the 5 most attributed words per case were identified. A board-certified radiologist then reviewed these words for relevance to laterality errors.

We additionally used multivariate logistic regression models to investigate the factors underlying the performance differences in false positive prediction between ClinicalBERT and GPT-4o across the Test-1 and Test-2 datasets after adjusting for potential confounders.

Statistical Analysis

We used Python (version 3.12.3; Python Software Foundation) with SciPy (version 1.12.0) and scikit-posthocs (version 0.11.3) for statistical analysis. The significance level (α) was set at .05, and all analyses were 2-tailed unless otherwise specified.

Demographics

We compared error rates and anatomical part distributions among TrainVal, Test-1, and Test-2 using chi-square tests. All observations were independent. For error rates, all expected cell counts exceeded 5. For analysis of anatomical parts, we merged temporomandibular joint and neck categories, resulting in all expected counts exceeding 1, with approximately 2.47% (2/81) of cells below 5. We used Bonferroni correction for pairwise comparisons (α=.0167) and reported Cramér *V* as the effect size.

A 2-proportion *z* test was used to compare laterality error rates between combined and noncombined reports. We verified that the expected number of events and nonevents per group exceeded 10 (based on the pooled proportion). Errors in reports were independent.

First Experiment

In the first experiment, we used the Friedman test to compare the recall ranks of the baseline and 2 optimized rule-based dictionary look-up methods. For each of the 3 datasets, the recall for each method was computed, after which the methods were ranked within each dataset (rank 1=highest recall).

After finding significant differences, we used the Nemenyi post-hoc test to identify which versions differed. The nonparametric, repeated measures rank data and paired observations of the 3 methods across the 3 datasets satisfy the assumptions of both tests. Scikit-posthocs provides direct access to pairwise *P* values, eliminating the need to calculate critical differences for the Nemenyi test manually.

Second Experiment

In the second experiment, multivariate logistic regression models were additionally used to investigate the factors underlying the false positive prediction of ClinicalBERT and GPT-4o across the Test-1 and Test-2 datasets. We constructed separate logistic regression models for the evaluation of ClinicalBERT and GPT-4o, with each observation corresponding to a positive prediction made by the model (ClinicalBERT or GPT-4o) on a sample from either dataset (Test-1 or Test-2). All observations were independent, with no duplicates within each analysis.

The 2 primary independent variables were dataset membership (Test-1 or Test-2) and order count per examination. The models included anatomical location (ankle, femur, finger, foot, forearm, hip, humerus, leg, scapula, shoulder, wrist, and others) as covariates. Anatomical part variables are not mutually exclusive; each anatomical part covariate is included in the model as a separate binary variable indicating presence or absence in the observation. The outcome variable indicates whether a prediction was a false positive. All independent variables are binary, except for the order count per examination, which is a continuous variable. We performed the sensitivity analysis using the *E* value [29] to evaluate the robustness of our findings to potential unmeasured confounding, with the calculation based on the observed odds ratio (OR) point estimate for rare outcome (<15%) or the square root of OR otherwise. Due to the imbalance in the outcome variable, we applied frequency weights to prevent the dominant class from skewing the results.

The following diagnostic checks were performed to ensure that our model met the assumptions of logistic regression. The outcome variable was defined as a dichotomous variable. We excluded binary independent variables with fewer than 10 observations in the least frequent category. We checked for linearity of the continuous covariate (order count per examination) using the Box-Tidwell test. We retained outliers identified using Cook distance (4/N) and standardized residuals ($|z|>3$) because they reflected valid data. Boolean OR was used to combine specific dichotomous covariates for anatomical locations into an "others" category, reducing multicollinearity, decreasing the variance inflation factor, and improving model stability. As our focus is on the association between the primary independent variables and the outcome rather than the specific impact of each anatomical location, the combination is justified.

Results

Demographics

Table 2 shows the demographic information of the datasets. As of July 2024, our study includes 10,889 studies comprising 10,000 real-world radiographic studies from 3 hospital branches corresponding to 20,635 orders, 22,523 distinct anatomical parts, and 889 synthetic studies containing manually generated errors. The anatomical parts were derived by removing the view information and laterality from the original medical orders. For example, "Rt knee anteroposterior+lateral+Merchant view, L-spine anteroposterior+lateral view" was simplified to "Knee, L-spine."

XSL•FO
RenderX

Table 2. Demographics.

	TrainVal[a]	Test-1	Test-2	Test-2 (before augmentation)	*P* value[b]
Studies, n	7307	1804	1778	889	<.001
Without laterality error	7205	1786	889	889	
With laterality error	102	18	889	0	
Attending physicians, n	27	12	27	27	__[c]
Orders, n	14,399	4503	3466	1733	—
Orders per study, mean (SD)	1.97 (0.97)	2.50 (0.86)	1.95 (1.02)	1.95 (1.02)	<.001
Anatomical parts involved, n (%)[d]					<.001
Abdomen	174 (1.11)	56 (1.13)	2 (0.05)	1 (0.05)	
Ankle	1023 (6.54)	298 (6)	274 (7.16)	137 (7.16)	
Cervical spine	241 (1.54)	103 (2.07)	64 (1.67)	32 (1.67)	
Calcaneus	62 (0.40)	16 (0.32)	18 (0.47)	9 (0.47)	
Chest	2228 (14.24)	762 (15.35)	344 (8.99)	172 (8.99)	
Clavicle	357 (2.28)	111 (2.24)	104 (2.72)	52 (2.72)	
Coccyx	44 (0.28)	7 (0.14)	6 (0.16)	3 (0.16)	
Elbow	601 (3.84)	221 (4.45)	158 (4.13)	79 (4.13)	
Femur	528 (3.38)	188 (3.79)	128 (3.34)	64 (3.34)	
Finger	108 (0.69)	8 (0.16)	32 (0.84)	16 (0.84)	
Foot	1029 (6.58)	293 (5.90)	246 (6.43)	123 (6.43)	
Forearm	226 (1.44)	57 (1.15)	86 (2.25)	43 (2.25)	
Hand	893 (5.71)	283 (5.70)	260 (6.79)	130 (6.79)	
Hip	368 (2.35)	188 (3.79)	106 (2.77)	53 (2.77)	
Humerus	234 (1.50)	74 (1.49)	56 (1.46)	28 (1.46)	
Knee	2035 (13.01)	577 (11.62)	506 (13.22)	253 (13.22)	
Lumbar spine	611 (3.91)	194 (3.91)	184 (4.81)	92 (4.81)	
Leg	579 (3.70)	156 (3.14)	158 (4.13)	79 (4.13)	
Lower limb	37 (0.24)	1 (0.02)	4 (0.10)	2 (0.10)	
Neck	6 (0.04)	7 (0.14)	4 (0.10)	2 (0.10)	
Pelvis	999 (6.39)	427 (8.6)	248 (6.48)	124 (6.48)	
Sacrum	542 (3.46)	149 (3)	162 (4.23)	81 (4.23)	
Scapula	126 (0.81)	18 (0.36)	30 (0.78)	15 (0.78)	
Shoulder	1144 (7.31)	375 (7.55)	304 (7.94)	152 (7.94)	
Skull	141 (0.90)	44 (0.89)	24 (0.63)	12 (0.63)	
Thoracic spine	125 (0.80)	55 (1.11)	30 (0.78)	15 (0.78)	
Temporomandibular joint	3 (0.02)	1 (0.02)	0 (0)	0 (0)	
Wrist	1180 (7.54)	296 (5.96)	290 (7.58)	145 (7.58)	

[a]TrainVal: combined training and validation dataset.

[b]Comparison for TrainVal, Test-1, and postaugmentation Test-2 datasets.

[c]Not available.

[d]Anatomical parts are obtained by removing view information from order (eg, chest x-ray posteroanterior becomes chest); each combined report may correspond to multiple anatomic parts.

Of the 10,000 real-world, nonsynthetic studies, 3000 contain only 1 medical order (noncombined studies). Among 7000 studies containing multiple orders (combined studies), 4325 reported each order's findings separately, and 2675 merged

XSL•FO
RenderX

findings in reporting, with the results corresponding to multiple medical orders combined into a single paragraph.

The error rates of the 3 datasets differed significantly (χ^2_2=4193.9; *P*<.001; Cramér *V*=0.62). Bonferroni-adjusted post hoc tests (α=0.0167) revealed no significant difference between TrainVal and Test-1 (χ^2_1=1.5; *P*=.23). All comparisons with artificially balanced Test-2 remained highly significant (*P*<.001), reflecting its intentional difference in class ratio.

The anatomical site distributions across the 3 datasets (TrainVal, Test-1, and Test-2) were statistically significantly different (χ^2_{52}=322.7; *P*<.001). However, the effect size was small (Cramér *V*=0.081), reflecting only minimal practical differences in distributions.

Distribution of Laterality Errors

The real-world data collected contains 120 cases of laterality errors, resulting in an error rate of 1.20% (120/10,000). Of these cases, 97.50% (117/120) contain 1 laterality error per report, and 2.50% (3/120) contain 2 errors per report.

The laterality error rate was 1.47% (103/7000) for combined reports compared to 0.57% (17/3000) for noncombined reports. The difference in proportions was 0.90% (95% CI 0.52%-1.29%), which was statistically significant (*z*=3.81; *P*<.001, 2-proportion *z* test).

Experiment 1

Primary Performance Metrics

Table 3 shows the impact of different optimization strategies on the rule-based dictionary look-up method. The rank of recall of 3 versions of the rule-based method shows significant differences (*Q*=6.0; *P*=.0498; *df*=2, Friedman test). The version enhanced with GPT information extraction (average rank=1) achieved the highest recall across all datasets, significantly outperforming the baseline (average rank=3; *P*=.04, Nemenyi test). There was no significant difference in recall between the GPT and workaround versions (average rank=2; *P*=.43, Nemenyi test).

Table 3. Impact of optimization strategies on the performance of rule-based methods.

	Precision	Recall	F_1-score
TrainVal[a] dataset			
Baseline	0.885	0.676	0.767
+Workaround[b]	*0.912*[c]	0.912	*0.912*
+GPT-4o[d]	0.874	*0.951*	0.911
Testing (Test-1 dataset)			
Baseline	0.833	0.556	0.667
+Workaround	*0.875*	0.778	*0.824*
+GPT-4o	0.696	*0.889*	0.780
Testing (Test-2 dataset)			
Baseline	0.992	0.882	0.934
+Workaround	0.993	0.897	0.942
+GPT-4o	*0.996*	*0.925*	*0.959*

[a]TrainVal: combined training and validation dataset.

[b]Workarounds: implement additional workarounds for individual radiologists' grammatically incorrect free-text reports.

[c]The top result for each metric is reported in italics format.

[d]GPT-4o: use GPT-4o to extract information before applying the rule-based method.

Subjective Error Analysis of Combined Rule-Based Method With GPT-4o

The rule-based method combined with GPT information extraction yielded 9 incorrect predictions (7 false positives and 2 false negatives) on Test-1. Contralateral report comparison contributed to 55.55% (5/9) of errors, misconfigured expert rules accounted for 22.22% (2/9), and other unaddressed scenarios in algorithm design—such as dictating the same side twice in a bilateral report—made up 22.22% (2/9) of the errors. None of the errors resulted from the omission of view information or terms misattributed to unrelated orders.

GPT information extraction reduced 2 false negatives but added 6 false positives in the Test-1 dataset compared to the workaround version. One of the reduced false negatives resulted from GPT's terminology standardization, where plural terms were converted to singular forms so that expert rules could correctly capture them. In total, 5 of the 6 additional false positives were due to limitations in the parser design rather than errors in the GPT information extraction. In these cases, radiologists simultaneously mentioned both the left and right sides for comparison, which was intentional and not a laterality error. In the experiment of a rule-based parser with a workaround, the expert rules coincidentally lacked plural versions of keywords, leading to error cancellation that masked

XSL•FO
RenderX

these hidden errors. The parser could trigger these false positives only after the GPT information extraction standardized the terms.

Experiment 2

Primary Performance Metrics

Table 4 compares the rule-based method enhanced with GPT information extraction and those based solely on deep learning. The rule-based method enhanced with GPT information extraction consistently performed well on both imbalanced, real-world TrainVal and Test-1 datasets and the balanced, synthetic Test-2 dataset. In the balanced, synthetic Test-2 dataset, both rule-based methods enhanced with GPT information extraction and GPT-4o performed well.

ClinicalBERT, trained on data from the artificial error generator, achieved a recall of 0.667 on the Test-1 dataset, detecting only two-thirds of laterality errors. RoBERTa achieved a recall of 0.000, suggesting that it did not effectively learn from the error patterns.

ClinicalBERT trained with noisy labels and GPT-4o using few-shot prompting showed substantial performance degradation when applied to the 2 real-world, imbalanced datasets, especially in terms of precision.

Table 4. Performance of different methods with subgroup analysis.

	Precision	Recall	F_1-score
TrainVal[a] dataset			
Rule-based+GPT-4o[b]	*0.874*[c]	*0.951*	*0.911*
GPT-4o[d]	0.423	0.892	0.574
RoBERTa[e]	0.500	0.010	0.019
ClinicalBERT[f]	0.142	0.843	0.244
Testing (Test-1 dataset)			
Rule-based+GPT-4o	*0.696*	*0.889*	*0.780*
GPT-4o	0.219	*0.889*	0.352
RoBERTa	0.000	0.000	0.000
ClinicalBERT	0.047	0.667	0.088
Testing (Test-2 dataset)			
Rule-based+GPT-4o	0.996	0.925	0.959
GPT-4o	0.979	*0.953*	*0.966*
RoBERTa	*1.000*	0.007	0.013
ClinicalBERT	0.984	0.749	0.851
Test-1+Test-2 (noncombined subgroup)			
Rule-based+GPT-4o	*1.000*	0.929	0.963
GPT-4o	0.997	*0.943*	*0.969*
RoBERTa	*1.000*	0.017	0.034
ClinicalBERT	0.966	0.798	0.874
Test-1+Test-2 (combined subgroup)			
Rule-based+GPT-4o	*0.981*	0.921	*0.950*
GPT-4o	0.878	*0.957*	0.916
RoBERTa	0.000	0.000	0.000
ClinicalBERT	0.619	0.715	0.664

[a]TrainVal: combined training and validation dataset.

[b]Rule-based+GPT-4o: combined rule-based dictionary look-up and GPT-4o extraction.

[c]The top result for each metric is reported in italics format.

[d]GPT-4o: GPT-4o with few-shot prompting.

[e]RoBERTa: fine-tuned RoBERTa with noisy, automatically generated labels.

[f]ClinicalBERT: fine-tuned ClinicalBERT with noisy, automatically generated labels.

XSL•FO
RenderX

Subgroup Analysis

Subgroup analysis showed that all models had lower precision and F_1-scores in the combined-study subgroups than in the noncombined-study subgroups.

Multivariable Logistic Regression for Predictors of False Positives in ClinicalBERT and GPT-4o

Table 5 summarizes the multivariable logistic regression results. We found that order count per examination or anatomical part differences cannot explain away the effect of data membership (in Test-1) on increased false positive predictions for GPT-4o and ClinicalBERT.

Table 5. Multivariate logistic regression analysis examining the associations between false positive prediction errors of the GPT-4o and ClinicalBERT models on the Test-1 and Test-2 datasets and the characteristics of combined reports.

Variable	GPT-4o		ClinicalBERT	
	OR[a] (95% CI)	*P* value	OR (95% CI)	*P* value
In Test-1	239.89 (111.05-518.01)	<.001	1924.07 (687.46-5382.99)	<.001
Order count	1.79 (1.38-2.31)	<.001	2.50 (1.64-3.80)	<.001
Anatomical parts				
Ankle	1.89 (0.97-3.67)	.06	0.90 (0.24-3.35)	.88
Femur	3.31 (1.69-6.49)	<.001	0.53 (0.09-3.05)	.48
Finger	8.72 (2.54-29.90)	<.001	0.95 (0.01-71.59)	.98
Foot	1.25 (0.63-2.48)	.52	0.43 (0.11-1.75)	.24
Forearm	0.27 (0.06-1.26)	.10	1.21 (0.24-6.25)	.82
Hip	3.21 (1.52-6.75)	<.001	0.17 (0.02-1.32)	.09
Humerus	0.19 (0.03-1.03)	.05	0.90 (0.15-5.53)	.91
Leg	2.01 (1.04-3.88)	.04	0.43 (0.10-1.86)	.26
Scapula	2.77 (0.74-10.46)	.13	0.51 (0.00-86.14)	.80
Shoulder	0.90 (0.45-1.82)	.77	0.23 (0.06-0.89)	.03
Wrist	1.24 (0.59-2.61)	.57	0.88 (0.27-2.92)	.84
Others	0.86 (0.42-1.77)	.68	2.84 (0.66-12.24)	.16

[a]OR: odds ratio.

After adjusting for anatomical parts, both Test-1 membership and order count remained significant independent predictors of false positive errors for GPT-4o (Test-1: OR 239.89, 95% CI 111.05-518.01; *P*<.001; order count: OR 1.79, 95% CI 1.38-2.31; *P*<.001) and ClinicalBERT (Test-1: OR 1924.07, 95% CI 687.46-5382.99; *P*<.001; order count: OR 2.50, 95% CI 1.64-3.80; *P*<.001). Associations for anatomical parts were variable and inconsistent across models. We included these variables primarily for covariate adjustment rather than causal inference.

E value sensitivity analysis indicated substantial robustness to unmeasured confounding (GPT-4o: in Test-1: 479.29, order count: 2.97; and ClinicalBERT: in Test-1: 87.2, order count: 2.53).

Subjective Error Analysis of ClinicalBERT on the Test-1 Dataset

Figure 2 shows a representative example of word attribution analysis. Of the 44 high-confidence (softmax probability≥0.9) incorrect predictions (38 false positives and 6 false negatives), we identified the 5 most attributed words per case, totaling 220 words. Only 46.82% (103/220) of these words were relevant to laterality error assessment. Therefore, even in high-confidence cases, ClinicalBERT often relies on features unrelated to the target task.

XSL•FO
RenderX

Figure 2. A high-confidence false-positive prediction from ClinicalBERT (softmax probability≥0.9). Corresponding orders were anteroposterior and lateral views of the bilateral ankles, wrists, and cervical spine. Word attribution analysis shows that ClinicalBERT overlooked important keywords, such as "Lt" and assigned attention to irrelevant words, such as "otherwise" and "obvious.".

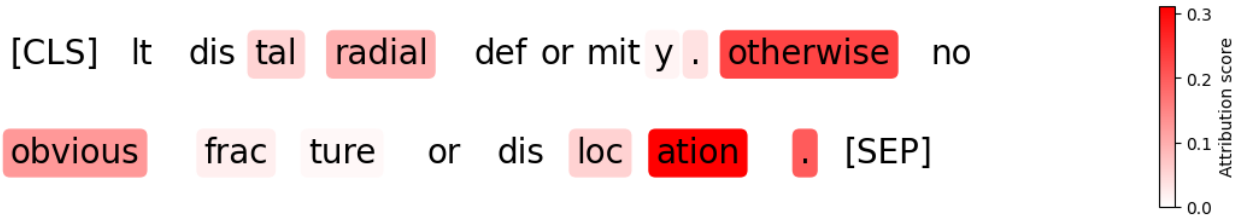

Subjective Error Analysis of GPT-4o on the Test-1 Dataset

In our analysis of GPT-4o's wrong predictions on the Test-1 dataset, 47.46% (28/59) resulted from misclassifying valid reporting styles, such as splitting or merging descriptions of both sides or legitimate omission, as laterality errors. Additionally, 18.64% (11/59) of incorrect predictions involved misclassifying nonlaterality errors (eg, partial order omissions) as laterality errors, 15.25% (9/59) reflected inconsistent or self-contradictory reasoning (eg, stating that no laterality error was present while still flagging the case as positive), and 11.86% (7/59) reflected a mismatch between the anatomical part of the findings and the clinical orders. Report quality issues accounted for 3.39% (2/59), while discrepant but legitimate error classifications and false negatives each comprised 1.69% (1/59). Multimedia Appendix 3 presents an example for each type of error, accompanied by GPT-4o's reasoning and our explanation.

Interrater Agreement in Data Annotation

We selected 500 samples for independent annotation, including 100 of 120 samples with laterality errors identified by the main annotator and 400 of 9880 samples without laterality errors. Independent annotation by a second annotator yielded a κ value of 0.998, with disagreement in one case, which the main annotator had labeled as positive.

Discussion

Principal Findings

Distinct Challenges Presented by Combined Reports

Combined reports pose distinct challenges for radiology report quality assurance. Our results show a significantly higher laterality error rate in combined reports compared to noncombined reports.

In addition to increased reporting errors, combined reports also present new challenges for NLP. In experiment 2, models that perform well on noncombined reports, such as GPT-4o and ClinicalBERT, exhibit substantial drops in performance on the combined reports subgroup, particularly in precision (Table 4). Furthermore, multivariable logistic regression (Table 5) identifies the order count per study as an independent predictor of false positive errors for both models, even after adjusting for covariates. Our findings show that although methods for noncombined reports remain applicable, their effectiveness declines in the context of combined reports.

Performance Analysis of Pure Deep Learning–Based Methods

Our results point to class imbalance as a potentially important factor in reducing the precision of pure deep learning methods. Both GPT-4o and ClinicalBERT demonstrated much lower precision on the imbalanced Test-1 dataset compared to the balanced Test-2 dataset, while recall remained similar—a pattern consistent with the behavior seen in imbalanced datasets dominated by negative samples (samples without laterality errors). Furthermore, in our multivariate logistic regression analysis (Table 5), the order count per study and anatomical distribution did not fully account for the increased false positives. Sensitivity analysis indicated substantial robustness of our results to unmeasured confounding. These findings support class imbalance as the primary explanation for the observed reduction in precision.

ClinicalBERT trained with artificial error generator data achieved high precision (>0.9) on the balanced, synthetic Test-2 dataset while attaining moderate recall on both the Test-1 (0.667) and Test-2 (0.749) datasets. These results suggest that the synthetic errors from our error generator used for training partially overlap with real-world error patterns, enabling the model to identify a subset of clinically relevant cases. However, these artificial errors fail to capture the full diversity and complexity of real-world laterality errors. Our word attribution analysis (Figure 2) further revealed that only 46.82% (103/220) of the top 5 most attributed words were relevant to laterality error detection, indicating that ClinicalBERT may rely on spurious correlations or shortcut features unrelated to the target task.

These results underscore both the value and the limitations of synthetic data augmentation. While our error generator can supplement model training with clinically plausible erroneous patterns, it cannot fully substitute for diverse, real-world cases. Expanding and diversifying the training data, especially by incorporating realistic clinical errors or improving the diversity of synthetically generated samples, will be essential for achieving robust and generalizable model performance.

Additionally, the consistently superior performance of ClinicalBERT over RoBERTa in terms of recall across all datasets demonstrates that domain-specific pretrained models better capture relevant patterns in our screening task.

XSL•FO
RenderX

Role and Limitations of the Synthetic Test-2 Dataset

We use the synthetic, class-balanced Test-2 as a controlled diagnostic dataset and rely on the real-world, imbalanced, cross-site Test-1 for deployment-facing conclusions. By design, Test-2 was created manually by flipping exactly 1 laterality term per report, matching the per-report error count in 97.50% (117/120) of real-world positive cases. Such random flips can produce contradictions across orders, subheadings, and paragraph text. We did not attempt to align the distribution of error contexts with heterogeneous real-world reporting conditions (eg, missing subheadings, merged descriptions in combined studies, legitimate contralateral comparisons, and difficult-to-align anatomies), nor did we model reports with multiple laterality errors or nonrandom-flip error patterns (eg, copy-paste with an unedited side), including patterns we may not have identified. Thus, while Test-2 captures single-term lexical substitutions, it does not fully capture the more complex structural contexts and error etiologies observed in practice, which can alter task difficulty and interfere with model performance. Consistent with these differences, all models showed lower recall on Test-1 than on Test-2 in experiment 2 (Table 4), suggesting that Test-1 may contain more challenging cases that are underrepresented in Test-2.

Therefore, our results suggest that while synthetic data grounded in real-world observations can yield optimistic estimates of model performance on common, easily observed laterality error manifestations, it cannot entirely replace real-world evaluation, given its constrained scope and the distributional differences from real reports. Future work could broaden the scope of synthetic patterns and expand real-world, nonsynthetic cases to more accurately assess model performance in clinical settings.

Proposed Solution for Laterality Error Screening in Combined Radiographic Reports

We found that the combination of GPT-4o with rule-based methods shows strong performance in real-world screening. The integration of GPT-4o significantly improves the recall of the rule-based parser (Table 3), while maintaining acceptable sensitivity and overall performance. In Test-1, the GPT-optimized version had 6 more false positives compared to the workaround version, mostly (5/6) due to parser limitations triggered by GPT. Thus, GPT extraction did not significantly affect precision, and our hybrid approach improves case detection with minimal loss of specificity in real-world screening.

Our method can be embedded into the radiology reporting workflow as a presignature, soft-block quality assurance step within the radiology information system. Figure 3 illustrates a conceptual message box after integration. When a radiologist attempts to finalize a report, the client can call a backend web service with the order information and the full draft report. Our methods allow line-level alignment among the original sentences, the GPT-extracted entities, and the parser-mapped anatomical parts and laterality. The user interface can present returned results with inline highlighting and 1-click options to revise or proceed with finalization. All alerts, GPT and parser outputs, user actions, and the versions of the ruleset, prompt, and model can be stored to support reproducibility, root-cause analysis of false alerts, and retrospective review.

Figure 3. A conceptual message box for detecting laterality errors using our method for the example in Table 1. Line 4 is incompatible with the paragraph heading at line 3, which is left shoulder only. Our method permits a human-readable error message with line-level alignment among the problematic sentences, the GPT-extracted entities (not shown), and the parser-mapped anatomical parts with laterality.

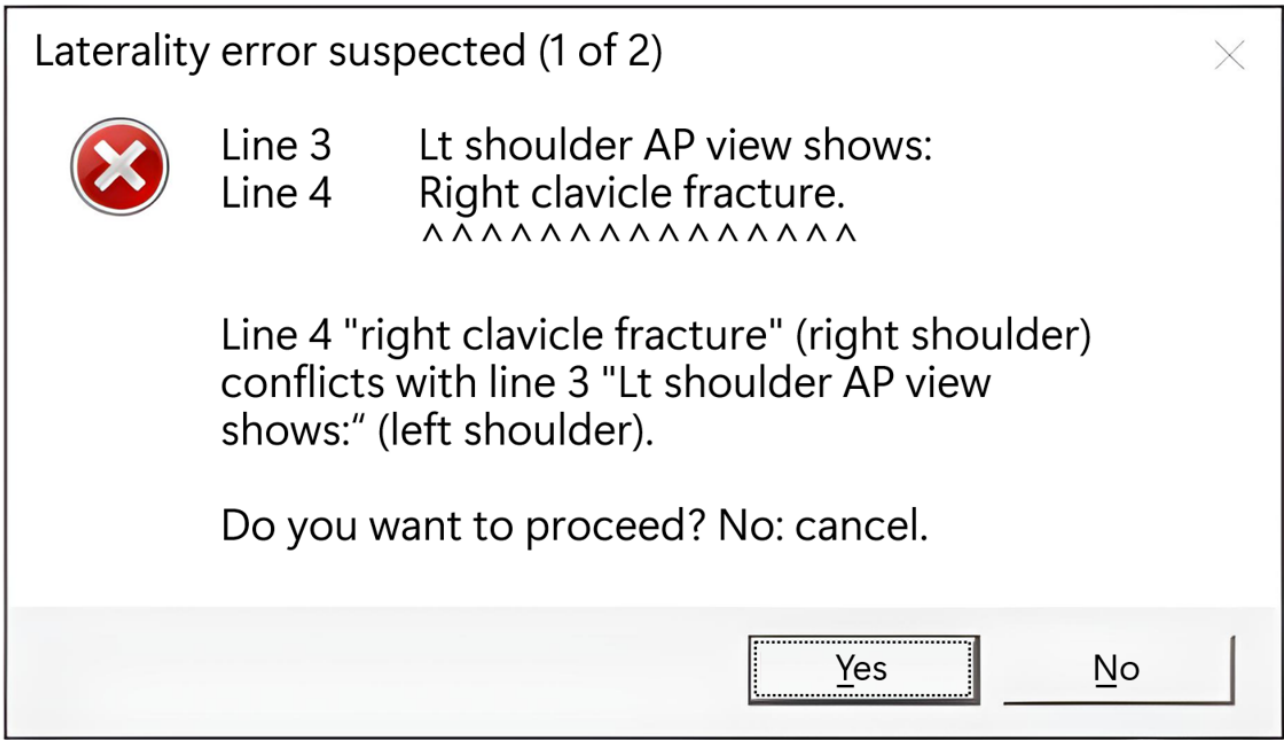

XSL•FO
RenderX

In daily use, this review-before-release check aims to reduce preventable harm by flagging likely laterality errors prior to sign-off, without adding undue burden to the workflow, given the performance of our method on real-world, imbalanced data. For safety improvement beyond prospective use, batch analysis of historical reports can detect previously missed laterality errors to enable timely correction and remediation. For sites that cannot deploy LLMs due to cost or privacy constraints, a rule-based parser with hand-crafted workarounds is a practical fallback; however, its effectiveness relies on continuous implementation modifications (not only rule updates) to keep pace with shifting local reporting styles over time, imposing an ongoing maintenance burden. By contrast, our combined rule-based parser and GPT-4o approach standardizes heterogeneous language upstream automatically, leading to improved recall and reduced manual maintenance efforts simultaneously, resulting in a more sustainable and scalable clinical NLP pipeline.

Comparison to Prior Work

Table 6 provides a comparative summary of related studies that have used NLP methods to detect laterality errors.

Table 6. Comparison of studies using natural language processing (NLP) methods to detect laterality errors.

	NLP algorithm	Laterality comparison	Values, n	Synthetic data	Partial label validation
Lee et al [17]	Keyword matching	Examination name, report	300	Yes	No
Lee et al [17]	Keyword matching	Examination name, report	29,257	No	No
Sheehan et al [18]	Keyword matching	Image side, clinical history or order	76,468	No	Yes
Minn et al [19]	Keyword matching regular expression	HL7[a] metadata, impression of report	391,657	No	Yes
Min et al [20]	Fine-tuned ClinicalBERT	Finding, impression of report	187,673	Yes	No
Kathait et al [22]	Multimodal GPT-4	Finding, impression of report, image	898	No	No
Gertz et al [21]	GPT-4	Finding and impression of report	200	Yes	No
Our study	Ruled-based method with or without GPT, RoBERTa, GPT-4o	Order, full report	10,000[b]	No	No
Our study	Ruled-based method with or without GPT, RoBERTa, GPT-4o	Order, full report	886[b]	Yes	No

[a]HL7: Health Level 7.

[b]Combined report included.

Our study is the first to define and systematically evaluate the combined report as a distinct category of complex radiology reports. By evaluating both existing and newly proposed NLP methods for these reports, we identified and addressed unique challenges that have not been previously reported. Furthermore, we demonstrated a significant performance gap between commonly used synthetic, balanced datasets and large-scale, real-world imbalanced data, underscoring the importance of real-world evaluation for the clinically robust deployment.

Although previous studies by Gertz et al [21] and Kathait et al [22] have also focused on complex reports, such as CT and MRI, they did not address the challenges posed by the combined report format or real-world evaluation using highly imbalanced data. In addition, Kathait et al [22] evaluated the performance of the multimodal GPT-4 in a second look at data flagged as laterality errors by commercial NLP tools, in contrast to our study's focus on first-line screening.

Sheehan et al [18] and Minn et al [19] used a more significant number of real-world samples, but did not fully evaluate all included data for laterality errors. They only assessed whether the data identified as positive by the NLP algorithm were true or false positives. Similar to other research [18-22], the study by Lee et al [17] does not address the complexities that arise from combined reports. Their study, published in 2015, also lacks an evaluation of modern approaches based on deep learning and LLMs.

Future Directions

Algorithmic Improvement for Rule-Based Method

Our results show that expert rule-based approaches are not without limitations, as experts may not consider all possible exceptions when designing algorithms and formulating expert rules. For example, we did not consider the scenario in the Test-1 dataset where radiologists intentionally referenced the opposite side for comparison, resulting in false positive predictions.

Despite the effectiveness of our method on real-world data, these edge cases reveal opportunities for further improvement in clinical deployment. Addressing potential mismatches between the development and test data and improving data diversity can help experts identify issues early in the development process. Algorithmic refinements such as context-aware postprocessing may also further enhance the robustness of our approach. Our findings underscore the importance of ongoing optimization and iterative expert review to ensure real-world applicability and support the continued adaptation to the evolving clinical environment.

Automatic extraction of expert rules is also a viable direction for future research. The significant manual effort to analyze

XSL•FO
RenderX

samples and formulate rules limits the application of rule-based methods. If automated methods could support the extraction of expert rules, rule-based approaches would become significantly more feasible.

Finally, our method still assumes specific reporting styles (eg, using "shows:" as a subheading); thus, external adaptation requires prior analysis of local reporting conventions. Future work could focus on enhancing the parser's flexibility to accommodate heterogeneous reporting formats better and improve generalizability across institutions.

Future LLM Applications

Currently, LLMs have shown varying degrees of success in many radiology-related applications. The systematic review by Keshavarz et al [30] categorized ChatGPT applications in radiology into 5 broad clinical domains. Notably, ChatGPT excels at simplifying complex radiology reports into more straightforward, more structured formats. However, LLMs also have well-documented limitations and potential pitfalls, such as difficulties with the complex, multistep reasoning tasks that are often required in advanced applications [31].

Our study encountered notable difficulties in GPT-4o reasoning. In experiment 2, most (55/59, 93.22%) GPT-4o's errors on the Test-1 dataset were reasoning-related, including misclassification of valid reporting styles or nonlaterality errors as laterality errors, inconsistent or self-contradictory reasoning, and mismatched anatomical parts. These results highlight limitations in GPT-4o's reasoning capabilities in our task. To address these difficulties, we divided the task into anatomical parts extraction using LLMs and matching with traditional algorithms, achieving significantly better performance than using LLMs alone. Based on previous literature [30] and our experimental results, we infer that our hybrid approach is more effective because it leverages GPT's strengths in transforming and summarizing reports into clearer formats while mitigating its weaknesses in the complex multistep reasoning required for our task.

This strategy may provide insights for applications beyond the detection of laterality errors. Rather than relying on LLMs for all tasks end-to-end, future NLP applications may achieve better performance by using LLMs for language understanding and assigning high-precision, context-dependent, algorithmic subtasks, such as logical decision-making for rare error screening, to more accurate methods. Such modular approaches use the complementary strengths of each method and may enable broader applications in NLP.

Future Studies Using Deep Learning for Laterality Error Detection

Our study shows that high performance on class-balanced, synthetic datasets does not guarantee similar outcomes in real-world, imbalanced scenarios where clinical deployment ultimately occurs. Deep learning methods such as GPT-4o and ClinicalBERT performed well on our balanced Test-2 dataset, a finding similar to the results of Min et al [20] and Kathait et al [22]. However, their performance severely degraded in the real-world imbalanced Test-1 dataset. The findings of Min et al [20] also showed a downward trend in the area under the precision-recall curve, precision, and recall of the model's performance in imbalanced datasets compared to more balanced datasets. We also found that real-world data may contain more challenging, underrepresented cases, whereas controlled synthetic data may capture only more obvious error patterns. These findings raise concerns about relying solely on balanced or synthetic datasets for benchmarking and about the real-world performance of pure deep learning methods.

Therefore, we recommend that future studies additionally include imbalanced real-world datasets without simulated samples into their evaluation frameworks, moving beyond synthetic, balanced samples. Our results demonstrate that real-world validation, which accurately reflects the clinical environment, is essential for developing robust and reliable clinical NLP applications suitable for deployment in actual workflows.

Limitations of Our Study

Our study has several limitations. First, our method for detecting laterality errors references the laterality of medical orders without the original images. Thus, our method cannot identify some laterality errors, such as mislabeling a left rib fracture as a right rib fracture on a chest x-ray. Our metadata-based method also cannot entirely eliminate false positives, as we cannot fully exclude the possibility that some studies may include minimal contralateral-side images. However, both laterality error detection based on metadata and images of studies are important and represent different but complementary approaches.

Second, several factors may affect the comparability of our study with other studies. For example, our research compares physician orders and free-text reports rather than findings and impressions. In addition, our study does not include more complex CT, MRI, or ultrasound reports. The varying complexity of free-text reports could influence the performance of the methods. However, our research is clinically relevant because radiographic reports are among the most common studies in practice, and discrepancies between the laterality of free text and physician orders are a major source of laterality errors. Our institution does not have combined CT or MRI studies available for research.

Third, the quantity and proportion of data with laterality errors are limited in real-world datasets, meaning that even minor differences in model predictions can significantly impact performance metrics. Nevertheless, our team has collected sufficient samples to the best of our ability, and our study included class-balanced samples for performance evaluation, similar to previous research.

In addition, we have not addressed all the known issues of the expert rule-based models. Our initial attempts to add rules for edge cases increased system complexity and reduced reliability due to unintended rule interactions. Introducing test-specific fixes could also affect unbiased evaluation of the model's performance on new data. Despite such limitations, the current results show that the rule-based model applies to this problem.

Moreover, given the large sample size (10,000 real-world studies), resource limitations, and the pilot or feasibility nature of our study, only 1 board-certified radiologist performed all

XSL•FO
RenderX

primary annotations. Single-annotator designs risk subtle systematic bias such as same-directional handling of borderline or ambiguous cases, intrarater drift, and fatigue-related undercalling and incorporation bias, as the same radiologist defined the guidelines and expert rules. Our study mitigated incorporation bias by grounding a high-level guideline—including an a priori definition of ambiguity—in imaging standards rather than providing detailed instructions specific to anatomy-view combinations and finalizing it before evaluation. We developed expert rules only on TrainVal and froze them before evaluating model performances on Test-1 and Test-2. For label reliability, an independent radiologist—uninvolved in guideline or rule design—double-annotated a stratified subsample enriched for the minority class based on their own knowledge of imaging standards, blinded to primary labels, model outputs, and class prevalence. Agreement was near-perfect (Cohen κ=0.998). Nonetheless, with only 2 raters, the possibility of residual common-direction bias cannot be completely excluded, especially for ambiguous cases.

Furthermore, our rule-based method does not use view information. While the rule-based approach combined with GPT-4o exhibited no errors attributable to missing view information in Test-1, we acknowledge that, due to the complexity of x-ray imaging, view information may be essential for laterality error detection in some instances.

Finally, our study is an internal validation study, and we selected Test-1 from a different institution. We cannot exclude the possibility that unquantified data drift or systematic institutional differences not included in our analysis (eg, reporting styles or clinical practices) may also exert their effects through data membership, together with the effect of class imbalance. Our sensitivity analysis cannot exclude these factors. Our approach does not substitute for stratified splitting at the partitioning stage to control data drift and ensure unbiased model development and evaluation. Further studies with larger and more diverse external datasets address these issues. Nonetheless, cross-site validation with different branches directly demonstrates model performance in a new clinical setting, which is crucial for real-world applications.

Conclusions

Combined radiographic reports pose distinct challenges for both clinical radiology report quality assurance and NLP. The ensemble method that integrates GPT-4o information extraction with expert rule-based methods detects laterality errors effectively in combined radiographic reports on real-world, highly imbalanced data. Carefully and manually optimized rule-based methods without GPT-based information extraction are also practical alternatives.

Our study found a significant performance gap between the results using well-balanced augmented data and real-world highly imbalanced data. We suggest additional inclusion of imbalanced real-world datasets without simulated samples in future studies.

Acknowledgments

This work was supported by the Chi Mei Medical Center (grant CMFHR113105). During the preparation of this work, the authors used DeepL Write, Grammarly, and ChatGPT in order to polish the English text and translate Chinese into academic English. After using these services, the authors reviewed and edited the content as needed and take full responsibility for the content of the published paper.

Data Availability

The software code, expert rules for the rule-based method, fine-tuned ClinicalBERT and RoBERTa models, and GPT prompts in this study are available in the author KHW's repository [27,32]. The deidentified patient data are not publicly available due to institutional policies but are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: YTK, KHW
Data curation: KHW
Formal analysis: KHW, CFL
Funding acquisition: YTK
Investigation: KHW
Methodology: KHW, CFL
Project administration: CFL
Resources: KHW
Software: KHW
Supervision: CFL, YTK
Writing—original draft: KHW
Writing—review and editing: CFL, TJH, YCC, YTK
Validation: CFL, TJH, YCC

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Examples of clearly suspicious laterality errors in x-ray of the knee.
[DOCX File , 44 KB - formative_v9i1e76384_app1.docx]

Multimedia Appendix 2
The parser implementation details of rule-based dictionary look-up method.
[DOCX File , 47 KB - formative_v9i1e76384_app2.docx]

Multimedia Appendix 3
One representative example of each error type produced by GPT-4o on the Test-1 dataset, together with GPT-4o's reasoning and our analysis.
[DOCX File , 49 KB - formative_v9i1e76384_app3.docx]

References

1. Bennani S, Regnard N, Ventre J, Lassalle L, Nguyen T, Ducarouge A, et al. Using AI to improve radiologist performance in detection of abnormalities on chest radiographs. Radiology 2023;309(3):e230860. [doi: 10.1148/radiol.230860] [Medline: 38085079]
2. Kim YW, Mansfield LT. Fool me twice: delayed diagnoses in radiology with emphasis on perpetuated errors. AJR Am J Roentgenol 2014;202(3):465-470. [doi: 10.2214/AJR.13.11493] [Medline: 24555582]
3. Pinto A, Brunese L, Pinto F, Reali R, Daniele S, Romano L. The concept of error and malpractice in radiology. Semin Ultrasound CT MRI 2012;33(4):275-279. [doi: 10.1053/j.sult.2012.01.009] [Medline: 22824117]
4. Sangwaiya MJ, Saini S, Blake MA, Dreyer KJ, Kalra MK. Errare humanum est: frequency of laterality errors in radiology reports. Am J Roentgenol 2009;192(5):W239-W244. [doi: 10.2214/AJR.08.1778] [Medline: 19380530]
5. Kwaan MR, Studdert DM, Zinner MJ, Gawande AA. Incidence, patterns, and prevention of wrong-site surgery. Arch Surg 2006;141(4):353-7; discussion 357. [doi: 10.1001/archsurg.141.4.353] [Medline: 16618892]
6. Seiden SC, Barach P. Wrong-side/wrong-site, wrong-procedure, and wrong-patient adverse events: are they preventable? Arch Surg 2006;141(9):931-939. [doi: 10.1001/archsurg.141.9.931] [Medline: 16983037]
7. Jhawar BS, Mitsis D, Duggal N. Wrong-sided and wrong-level neurosurgery: a national survey. J Neurosurg Spine 2007;7(5):467-472. [doi: 10.3171/SPI-07/11/467] [Medline: 17977186]
8. Christensen JD, Hutchins GC, McDonald CJ. Computer automated detection of head orientation for prevention of wrong-side treatment errors. AMIA Annu Symp Proc 2006;2006:136-140 [FREE Full text] [Medline: 17238318]
9. Clarke JR, Johnston J, Blanco M, Martindell DP. Wrong-site surgery: can we prevent it? Adv Surg 2008;42:13-31. [doi: 10.1016/j.yasu.2008.03.004] [Medline: 18953807]
10. Elghrably I, Fraser SG. An observational study of laterality errors in a sample of clinical records. Eye (Lond) 2008;22(3):340-343. [doi: 10.1038/sj.eye.6702590] [Medline: 16980925]
11. Dyer O. Doctors suspended for removing wrong kidney. BMJ 2004;328(7434):246 [FREE Full text] [doi: 10.1136/bmj.328.7434.246-a] [Medline: 14751886]
12. Asunción Márquez J, López Gutiérrez A, Pérez Martínez V, Poggio Cano D, Combalia A. Surgery on the wrong side: implication for the patient and the professional. Experience in a major ambulatory surgery of the foot unit. Rev Esp Cir Ortop Traumatol 2012;56(2):104-114. [doi: 10.1016/j.recot.2011.08.005] [Medline: 23594751]
13. Tuncer FB, Frautschi RS, Lundy SD, Haber G, Gastman B. First report of testicular replantation following wrong-site surgery. Urology 2022;164:267-272. [doi: 10.1016/j.urology.2021.11.044] [Medline: 35114199]
14. Austin JM, Pronovost PJ. "Never events" and the quest to reduce preventable harm. Jt Comm J Qual Patient Saf 2015;41(6):279-288. [doi: 10.1016/s1553-7250(15)41038-4] [Medline: 25990894]
15. Younis A. Enhancing clarity and patient safety: evaluating the inclusion of laterality in orthopaedic outpatient correspondence in Galway, Ireland. Cureus 2024;16(5):e60782 [FREE Full text] [doi: 10.7759/cureus.60782] [Medline: 38903340]
16. Eghtedari M, Chong A, Rakow-Penner R, Ojeda-Fournier H. Current status and future of BI-RADS in multimodality imaging, from the AJR special series on radiology reporting and data systems. AJR Am J Roentgenol 2021;216(4):860-873. [doi: 10.2214/AJR.20.24894] [Medline: 33295802]
17. Lee YH, Yang J, Suh J. Detection and correction of laterality errors in radiology reports. J Digit Imaging 2015;28(4):412-416 [FREE Full text] [doi: 10.1007/s10278-015-9772-x] [Medline: 25712502]
18. Sheehan SE, Safdar N, Singh H, Sittig DF, Bruno MA, Keller K, et al. Detection and remediation of misidentification errors in radiology examination ordering. Appl Clin Inform 2020;11(1):79-87 [FREE Full text] [doi: 10.1055/s-0039-3402730] [Medline: 31995835]

XSL•FO
RenderX

19. Minn MJ, Zandieh AR, Filice RW. Improving radiology report quality by rapidly notifying radiologist of report errors. J Digit Imaging 2015;28(4):492-498 [FREE Full text] [doi: 10.1007/s10278-015-9781-9] [Medline: 25694167]
20. Min D, Kim K, Lee JH, Kim Y, Park C. RRED : A radiology report error detector based on deep learning framework. 2022 Presented at: Proceedings of the 4th Clinical Natural Language Processing Workshop; September 30, 2025; Seattle, WA, United States p. 41-52. [doi: 10.18653/v1/2022.clinicalnlp-1.5]
21. Gertz RJ, Dratsch T, Bunck AC, Lennartz S, Iuga A, Hellmich MG, et al. Potential of GPT-4 for detecting errors in radiology reports: implications for reporting accuracy. Radiology 2024;311(1):e232714. [doi: 10.1148/radiol.232714] [Medline: 38625012]
22. Kathait AS, Garza-Frias E, Sikka T, Schultz TJ, Bizzo B, Kalra MK, et al. Assessing laterality errors in radiology: comparing generative artificial intelligence and natural language processing. J Am Coll Radiol 2024;21(10):1575-1582. [doi: 10.1016/j.jacr.2024.06.014] [Medline: 38960083]
23. Landau E, Hirschorn D, Koutras I, Malek A, Demissie S. Preventing errors in laterality. J Digit Imaging 2015;28(2):240-246 [FREE Full text] [doi: 10.1007/s10278-014-9738-4] [Medline: 25273506]
24. Huang K, Altosaar J, Ranganath R. ClinicalBERT: modeling clinical notes and predicting hospital readmission. ArXiv Preprint posted online on November 29, 2020 [FREE Full text] [doi: 10.48550/arXiv.1904.05342]
25. Williams A, Nangia N, Bowman S. A broad-coverage challenge corpus for sentence understanding through inference. 2018 Jun Presented at: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers); September 30, 2025; New Orleans, LA, United States p. 1112-1122. [doi: 10.18653/v1/n18-1101]
26. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F. Chain-of-thought prompting elicits reasoning in large language models. 2022 Presented at: NIPS'22: 36th International Conference on Neural Information Processing Systems; November 28-December 9, 2022; New Orleans, LA, United States URL: https://dl.acm.org/doi/10.5555/3600270.3602070 [doi: 10.5555/3600270.3602070]
27. Weng KH. d8888/GPT-study-2024-public_data. GitHub. 2025. URL: https://github.com/d8888/GPT-study-2024-public_data [accessed 2025-10-04]
28. Kong A, Zhao S, Chen H, Li Q, Qin Y, Sun R. Better zero-shot reasoning with role-play prompting. 2024 Presented at: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers); September 30, 2025; Mexico City, Mexico p. 4099-4113. [doi: 10.18653/v1/2024.naacl-long.228]
29. VanderWeele TJ, Ding P. Sensitivity analysis in observational research: introducing the E-value. Ann Intern Med 2017;167(4):268-274. [doi: 10.7326/M16-2607] [Medline: 28693043]
30. Keshavarz P, Bagherieh S, Nabipoorashrafi SA, Chalian H, Rahsepar AA, Kim GHJ, et al. ChatGPT in radiology: a systematic review of performance, pitfalls, and future perspectives. Diagn Interv Imaging 2024;105(7-8):251-265 [FREE Full text] [doi: 10.1016/j.diii.2024.04.003] [Medline: 38679540]
31. Bhayana R. Chatbots and large language models in radiology: a practical primer for clinical and research applications. Radiology 2024;310(1):e232756. [doi: 10.1148/radiol.232756] [Medline: 38226883]
32. Weng KH. d8888 (Kung Hsun Weng). Hugging Face. 2025. URL: https://huggingface.co/d8888/ [accessed 2025-10-04]

Abbreviations

CT: computed tomography
LLM: large language model
MRI: magnetic resonance imaging
NLP: natural language processing
OR: odds ratio
TrainVal: combined training and validation dataset

Edited by J Sarvestan; submitted 22.04.25; peer-reviewed by S Sivarajkumar, Z Yang; comments to author 20.05.25; revised version received 06.09.25; accepted 02.10.25; published 29.10.25.

Please cite as:
Weng KH, Chou YC, Kuo YT, Hsieh TJ, Liu CF
Detecting Laterality Errors in Combined Radiographic Studies by Enhancing the Traditional Approach With GPT-4o: Algorithm Development and Multisite Internal Validation
JMIR Form Res 2025;9:e76384
URL: https://formative.jmir.org/2025/1/e76384
doi:10.2196/76384
PMID:

XSL•FO
RenderX

©Kung-Hsun Weng, Yi-Chen Chou, Yu-Ting Kuo, Tsyh-Jyi Hsieh, Chung-Feng Liu. Originally published in JMIR Formative Research (https://formative.jmir.org), 29.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Impact of Detailed Versus Generic Instructions on Fine-Tuned Language Models for Patient Discharge Instructions Generation: Comparative Statistical Analysis

Muneerah Alqahtani[1,2], MS; Abdullah Al-Barakati[1], PhD; Fahd Alotaibi[1], PhD; Mohammed Al Shibli[3], MD; Saad Almousa[4], MD

[1]Information System Department, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia
[2]Information Systems Department, King Khalid University, Abha, Saudi Arabia
[3]Primary Health Care Center, Aseer Health Cluster, Abha, Saudi Arabia
[4]College of Medicine, King Khalid University, Abha, Saudi Arabia

Corresponding Author:
Muneerah Alqahtani, MS
Information System Department
Faculty of Computing and Information Technology
King Abdulaziz University
PO Box 80200
University Road, Al-Sulaymaniyah District
Jeddah, 21589
Saudi Arabia
Phone: 966 557574844
Email: malqahtani0680@stu.kau.edu.sa

Abstract

Background: Discharge instructions are essential for patients after hospital care but are time-consuming to write. With the rise of large language models (LLMs), there is a strong potential to automate this process. This study explores the use of open-source LLMs for generating discharge instructions.

Objective: We investigated whether a Mistral model can reliably generate patient-oriented discharge instructions. Two distinct instruction-tuning paradigms were compared, each using a different mechanism for embedding guidance during fine-tuning.

Methods: In our experiment, we applied Mistral-NeMo-Instruct, an LLM, in combination with 2 distinct instruction strategies for fine-tuning. The first were detailed instructions tailored to the task of discharge instruction generation. The second was a basic instruction with minimal guidance and no task-specific detail. The independent variable in this study is the instruction strategy (detailed vs generic), while the dependent variables are the evaluation scores of the generated discharge instructions. The generated discharge instructions were evaluated against 3621 ground-truth references. We used Bilingual Evaluation Understudy (BLEU-1) to BLEU-4, Recall-Oriented Understudy for Gisting Evaluation (ROUGE-1, ROUGE-2, and Recall-Oriented Understudy for Gisting Evaluation—Longest Common Subsequence), SentenceTransformer similarity, and Bidirectional Encoder Representations From Transformers Score as evaluation metrics to assess the quality of the generated outputs in comparison to the corresponding ground-truth instructions for the same discharge summaries.

Results: The detailed instruction model demonstrated superior performance across all automated evaluation metrics compared with the generic instruction model. Bidirectional Encoder Representations From Transformers Score increased from 78.92% to 87.05%, while structural alignment measured by Recall-Oriented Understudy for Gisting Evaluation—Longest Common Subsequence improved from 8.59% to 26.52%. N-gram precision (BLEU-4) increased from 0.81% to 21.24%, and the Metric for Evaluation of Translation With Explicit Ordering scores rose from 15.33% to 18.47%. Additional metrics showed consistent gains: ROUGE-1 improved from 16.59% to 42.72%, and ROUGE-2 increased from 1.97% to 45.84%. All improvements were statistically significant ($P<.001$), indicating that detailed, task-specific instruction design substantially enhances model performance.

Conclusions: The use of detailed, task-specific instruction strategies significantly enhances the effectiveness of open-source LLMs in generating discharge instructions. These findings indicate that carefully designed instructions during fine-tuning substantially improve model performance.

XSL•FO
RenderX

(JMIR Form Res 2025;9:e80917) doi:10.2196/80917

KEYWORDS

discharge instructions; large language models; clinical text generation; instruction tuning; medical natural language processing; open-source large language models; patient safety; prompt engineering

Introduction

Background

The quality of writing discharge instructions represents a challenge in health care hospitals. Discharge instructions serve as essential guides for posthospitalization care, but their crafting needs significant physician time within already demanding clinical workflows. Studies show that spending too much time on documentation is directly linked to physician burnout and poor work-life balance. Some providers spend over 6 hours each week on after-hours documentation [1,2]. Physicians showed low satisfaction with their electronic health records (EHRs) and computerized order systems (Computerized Physician Order Entry) because of the large amount of time spent on clerical work to write patient records. This growing burden highlights the urgent need for supportive tools to automate the process [3].

Current Solutions and Their Limitations

The use of predefined templates within EHRs is a widespread strategy for producing discharge instructions. However, these existing solutions have their own limitations. While they are intended to improve efficiency and ensure key information is included, they often fall short in practice. Evidence shows that even when disease-specific templates are available, they are frequently underused and fail to enhance the readability of patient instructions [4]. As a result, each discharge requires manual customization, where doctors frequently modify or rewrite discharge instructions to reflect individualized care plans. They add personalized medications, tailor activity restrictions, and simplify medical terminology for patient understanding [5,6]. This time-consuming process leads to inconsistent quality between providers. Under discharge time pressure, important information may be missing or written in ways patients cannot understand [6].

Language Models: A Promising Solution

According to a recent scoping review by Meng et al [7], large language models (LLMs) have been actively explored in several areas of medicine, including clinical documentation support. The review, which analyzed over 550 studies, found that LLMs have the potential to reduce clinician workload and streamline routine writing tasks. By leveraging their ability to understand and generate medically relevant content, LLMs present a promising solution to address the time burden associated with clinical documentation [7].

Optimizing LLMs Through Instruction Design for Discharge Instructions

While LLMs show a great promising solution to generate discharge instructions, their effectiveness depends largely on how they are guided, especially through prompt design. Two primary optimization approaches have emerged: (1) prompt engineering at inference time, where users modify input prompts without changing the model [8]; and (2) instruction tuning during training, where models are fine-tuned with task-specific instructions [9].

While prompt engineering has shown significant effects on LLM performance in general domains [8,10]. Instruction tuning offers deeper adaptation by incorporating domain knowledge directly into model parameters. In medical contexts, studies have demonstrated that carefully designed prompts can improve diagnostic reasoning and clinical documentation quality [8,11].

Aim of This Study

While LLMs are gaining popularity in health care, it remains unclear how much their performance depends on the way they are guided, especially in the context of generating discharge instructions. To investigate this, we conducted a comparative study evaluating 2 instruction strategies. We used 2 instruction strategies for generating discharge instructions. The tailored method used detailed instructions specifying medical roles (eg, doctors or physicians) and clear guidelines tailored to hospital-specific needs. The generic method used short, open-ended instructions. Both models were trained on the same set of real discharge summaries from the Medical Information Mart for Intensive Care IV (MIMIC-IV) database [12] and evaluated using standard metrics, including Bilingual Evaluation Understudy (BLEU) [13], Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [14], Metric for Evaluation of Translation With Explicit Ordering (METEOR) [15], and Bidirectional Encoder Representations From Transformers Score (BERTScore) [16], to assess performance.

The primary objective was to assess whether more structured instructions lead to more accurate, relevant, and patient-friendly discharge instructions. This study aimed to determine if structured instructions generate accurate, relevant, and patient-friendly discharge instructions. More broadly, the study explores providing practical insights for implementing LLMs in hospital settings.

The integration of LLMs into hospitals' workflows is expected to offer a solution to reduce the documentation burden. LLMs could help physicians focus more on patient care.

Methods

Ethical Considerations

This study used the MIMIC-IV database [12], a publicly available dataset containing deidentified health information from critical care patients. The database was approved by the institutional review board of the Massachusetts Institute of Technology (IRB #2001P001699) and was granted a waiver of informed consent in accordance with the Health Insurance Portability and Accountability Act (HIPAA) Safe Harbor

XSL•FO
RenderX

provisions. All authors completed the Collaborative Institutional Training Initiative "Data or Specimens Only Research" training and signed the PhysioNet Credentialed Health Data Use Agreement. The lead author (MA, credential approved on July 8, 2024) maintained an active credential for database access. To protect participant privacy, all data were used strictly in compliance with the data use agreement. No attempts were made to reidentify individuals, and all analyses were performed on secure, access-controlled systems. Because the dataset is fully deidentified and publicly available, no participant contact or compensation was involved.

Study Design

This was a comparative study based on an experiment of testing 2 instruction strategies for generating discharge instructions: generic instructions (21 words) versus tailored, detailed instructions (587 words). We tested these strategies by fine-tuning the same base model (Mistral-NeMo-Instruct) twice. A schematic of the workflow is shown in Figure 1.

Figure 1. Study workflow. (A) Data preparation (35,851 pairs from MIMIC-IV). (B) Distinct fine-tuning on Mistral-NeMo-Instruct with generic versus tailored instructions. (C) Evaluation with BLEU, ROUGE, METEOR, BERTScore, and SentenceTransformer similarity. BERTScore: Bidirectional Encoder Representations From Transformers Score; BLEU: Bilingual Evaluation Understudy; LLM: large language model; METEOR: Metric for Evaluation of Translation With Explicit Ordering; MIMIC-IV: Medical Information Mart for Intensive Care IV; ROUGE: Recall-Oriented Understudy for Gisting Evaluation.

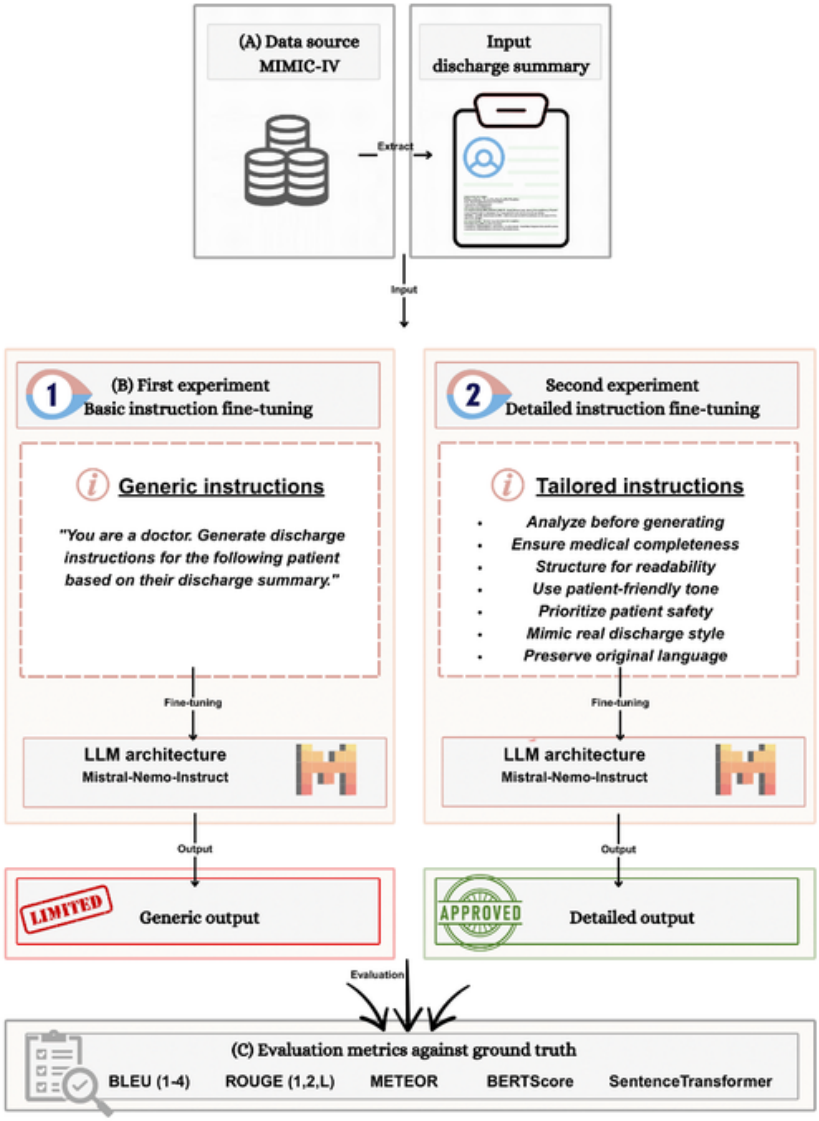

The experiment had 2 sides, one with generic instructions and the other with tailored instructions. During inference, each model used its corresponding instruction type when generating discharge instructions. For example, the model trained with generic instructions received the same 21-word generic prompt at inference time, while the model trained with tailored instructions received the same 587-word detailed prompt. This approach allowed us to evaluate the full impact of instruction design on model performance throughout both training and inference phases.

Our experiment followed 3 steps. First, we prepared the data by extracting 35,851 discharge summary-instruction pairs from MIMIC-IV. Second, we fine-tuned 2 models using the Mistral-NeMo-Instruct architecture. One was fine-tuned with generic instructions, and one with tailored instructions. Third, we evaluated the generated discharge instructions using multiple

XSL•FO
RenderX

metrics (BLEU, ROUGE, METEOR, BERTScore, and SentenceTransformer similarity). This design lets us measure how much tailored instructions improved discharge instruction generation compared to generic ones.

Data

Data Source and Description

This study used the MIMIC-IV database (version 2.2, released July 23, 2024), a publicly available collection of 364,627 deidentified health records from patients admitted to intensive care units or emergency departments between 2008 and 2022. Managed by Massachusetts Institute of Technology researchers, MIMIC-IV was a widely respected resource in medical research due to its continuous maintenance and evidence-based clinical documentation.

We built our data from the discharge.csv file, using notes labeled as "Discharge summary." These summaries included comprehensive patient details such as symptoms, patient history, medications before and after admission, laboratory results, and most importantly, the discharge instruction section. In our experiment, each summary served as input for our model, while the discharge instruction sections extracted from these summaries were used as expected output for model training. We used discharge instruction sections as the ground truth in the evaluation phase.

We chose MIMIC-IV for its reliability because patients' records were real and written by health care professionals. Additionally, the dataset contained diverse, real-world patient cases. It was a widely respected resource in medical research due to its continuous maintenance and evidence-based clinical documentation [12].

Preprocessing and Formatting

Our preprocessing pipeline extracted full discharge summary text from the report column of the MIMIC-IV database [12]. We constructed input or output pairs by separating the whole discharge summary documentation (input) and the "Discharge Instructions" section (output) using regular expressions Regex [17].

We applied some filtering steps. We excluded records of patients with discharge conditions labeled as "Expired," "Deceased," or "Gone Away," retaining only those of living patients, who were eligible for discharge instructions. We also excluded records of patients who were transferred to other medical facilities such as another hospital, because their discharge instructions would be issued by the facility that received them. Additionally, we excluded summaries that exceeded 20,000 characters to reduce the noise from outliers. We dropped all records with instructions that had fewer than 5 words because they may not have carried meaningful information for the patient.

Given the large size of the original dataset (364,627 records), we performed proportional sampling to create a smaller but representative dataset. We ensured that the dataset still reflected the original distribution of medical services. Analysis of the service distribution revealed that the department of medicine dominated with 195,533 cases (59.1% of total admissions). Our proportional sampling reduced the dataset from 364,627 records to 35,851 records while preserving the relative distribution across medical services. Therefore, the final dataset contained 35,851 records, as illustrated in Figure 2, which provides a high-level overview of the preprocessing pipeline.

Figure 2. High-level overview of the preprocessing pipeline. MIMIC-IV: Medical Information Mart for Intensive Care IV.

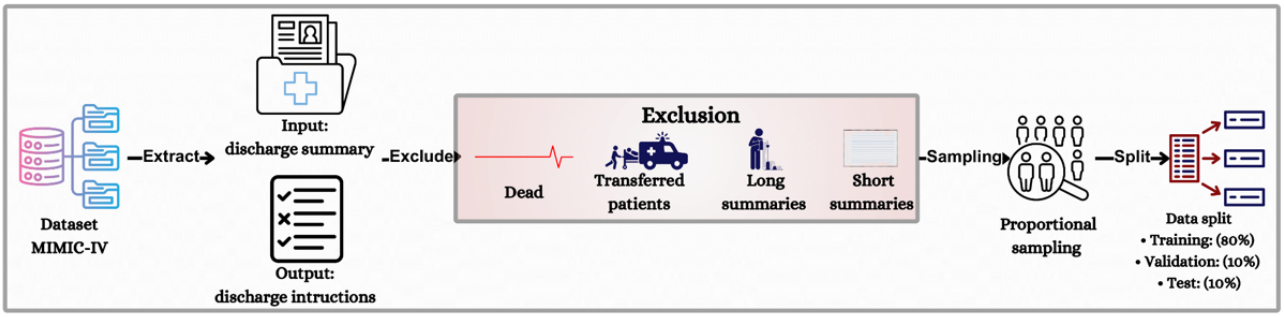

For instruction fine-tuning, we used predefined templates based on the Mistral-NeMo-Instruct format. The template comprised 4 core components: role (asking the model to act as a health care provider), instructions (guidelines for structuring output), input (the discharge summary), and output (the discharge instructions). To assess how the complexity of instructions affected the performance of the model when asked to generate discharge instructions, we developed 2 template variants: a simple, generic version and a detailed, task-specific version. The detailed template incorporated advanced prompting techniques, including acting as a physician, analyzing discharge summaries to identify medical details. The detailed template also stressed using simple, easy-to-understand instruction that is friendly for patients.

Final Dataset

Our final dataset comprised 35,851 high-quality discharge summary-instruction pairs. The dataset was split into training (n=28,673, 80%), validation (n=3557, 9.9%), and test sets (n=3621, 10.1%). Table 1 summarizes the number and percentage of samples in each split.

XSL•FO
RenderX

Table 1. Dataset splits for model development and evaluation.

Dataset	Samples, n (%)
Training	28,673 (80)
Validation	3557 (9.9)
Test	3621 (10.1)
Total	35,851 (100)

Statistical analysis of token counts revealed important characteristics of our data. Discharge summaries ranged from 276 to 3066 tokens with a mean of 1437 (SD 550; median 1414, IQR 1012-1876) tokens, while discharge instructions were considerably shorter, ranging from 68 to 420 tokens, with a mean of 176 (SD 72; median 157, IQR 118-214) tokens. The 90th percentile was 2193 tokens for discharge summaries and 291 tokens for discharge instructions, indicating that most documents fit comfortably within our model's context window, as detailed in Table 2.

Table 2. Token distribution statistics for discharge summaries and instructions.

Statistic	Discharge summaries	Discharge instructions
Mean (SD)	1437 (550)	176 (72)
Range	276-3066	68-420
Median (IQR)	1414 (1023-1812)	157 (118-223)
90th percentile	2193	291

Instructions Design

Overview

The core innovation of this study lay in comparing 2 distinct prompting strategies for instruction fine-tuning: a generic minimal instruction versus a tailored comprehensive instruction. Both instructions were used during model training and inference. This section details the design rationale and specific content of each instruction (Figure 3).

Figure 3. Generic versus tailored instruction design used during training and inference. The generic prompt (21 words) provides minimal task guidance. The tailored prompt (587 words) encodes structure, tone, safety checks, and completeness criteria.

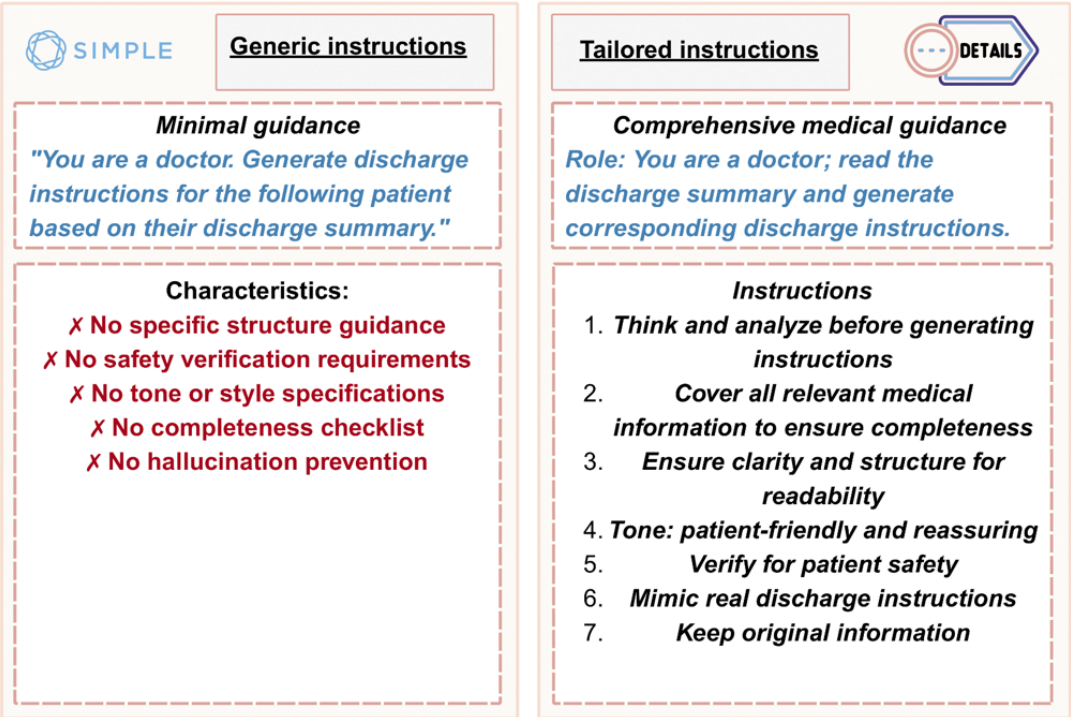

Generic Instructions

The generic instructions were designed to represent a minimal instruction approach. It provided only an essential task definition without domain-specific guidance. For example, "You are a doctor. Generate discharge instructions for the following patient based on their discharge summary."

These 21-word instructions were deliberately designed to test whether LLMs could generate adequate discharge instructions

XSL•FO
RenderX

with minimal guidance. It provided only 3 key elements: role identification ("You are a doctor"), task specification ("Generate discharge instructions"), and structural markers for input and output. We did not add additional constraints or formatting requirements.

Tailored Instructions

The tailored instructions were designed to incorporate essential domain-specific guidance based on analysis of the final dataset. We recognized that discharge instructions lacked universal standardization across health care institutions. Thus, these 587-word instructions were designed to capture common features and maintain diverse documentation styles.

The instruction text was derived from MIMIC-IV [12] discharge instructions using manual observation rather than rule-based extraction. A qualitative review was conducted to identify recurring elements and their typical order. The instructions were drafted from these observations. Then, we ran pilot trials on 100 summaries, monitored metric behavior, and made minor edits to the instructions. The final instruction text was fixed before fine-tuning and final evaluation.

Our analysis of discharge instructions revealed regular patterns that helped formulate the tailored instructions. These instructions often started with a greeting, such as "Dear [Patient]," followed by the reasons behind admission, including presenting symptoms, a summary of treatments provided, and a plan for ongoing care. Additionally, physicians consistently used clear, simple, and patient-friendly language to ensure readability.

We confirmed that most instructions included essential components such as (1) reason for admission, whether diagnosis or presenting symptoms; (2) key treatments and procedures performed during hospitalization; (3) medication changes with clear dosing instructions; (4) follow-up appointments and monitoring requirements; and (5) any warning signs that needed immediate medical attention.

Both instructions were integrated into the Mistral-NeMo-Instruct instruction template format during training, which included four core components: (1) the role, (2) instructions, (3) input, and (4) output. Each discharge summary was paired with its corresponding instructions and target output, allowing the model to learn distinct generation patterns. The full 21-word generic prompt and the complete 587-word tailored prompt are provided in Multimedia Appendix 1.

Model

Model Selection Criteria

We scanned existing models in order to select an efficient base model for discharge instruction generation. We considered the following factors: (1) demonstrated performance on medical and instructional tasks, (2) computational efficiency enabling training on available hardware, (3) support for extended context windows to accommodate lengthy discharge summaries, and (4) open open-source model. Based on these criteria, Mistral-NeMo-Instruct was selected as our foundation model.

Mistral NeMo Instruct

Mistral NeMo 12B, developed by Mistral AI in collaboration with NVIDIA and released in March 2024, represented a significant advancement in efficient LLM design. This 12-billion parameter model introduced enhanced reasoning abilities while maintaining computational efficiency, which made it accessible to researchers. A key feature was its expanded context window supporting 128,000 tokens, allowing it to process even the longest discharge summaries in our dataset without truncation concerns. An important practical consideration was that the entire 12-billion parameter model could run on a single high-end consumer graphics processing unit (GPU).

Model Configuration

We used Unsloth training speedup, parameter-efficient fine-tuning for low-rank adaptation implementation and bitsandbytes for 4-bit quantization. This stack enabled efficient training on a single GPU while maintaining model performance.

Training

Hardware and Environment

Model training and evaluation were conducted on Vast.ai using an NVIDIA H200 GPU with 140 GB of video random access memory. The machine environment included 192 central processing unit cores, 258 GB RAM, and CUDA (version 12.6). Training was implemented using the Unsloth framework built on top of Transformers, TRL, and Safetensors libraries. All experiments were executed in a Linux environment with Python (version 3.10; Python Software Foundation).

Training Configuration

We fine-tuned the base model unsloth/Mistral-Nemo-Instruct-2407-bnb-4bit using low-rank adaptation with the configurations presented in Table 3.

XSL•FO
RenderX

Table 3. Training configuration.

Setting	Values
Training samples	28,673
Validation samples	3557
Epochs	7 (with different convergence rates for basic and detailed models)
Batch size	1
Learning rate	4×10^{-5}
Maximum sequence length	15,000 tokens
Context window optimization	128,000
Instructions fine-tuning format	Mistral-NeMo-Instruct style (21-word generic or 587-word detailed instruction)
Few-shot examples	We used only 2 examples during inference time

Evaluation

Overview

To evaluate models, we used multiple automated metrics to assess the quality of generated discharge instructions. The evaluation was applied to 3621 records, which represented 10% of the dataset. All metrics compared model-generated instructions against the ground truth, physician-written reference instructions.

Evaluation Metrics

We selected evaluation metrics that captured different aspects of discharge instruction quality. Surface-level similarity was assessed using BLEU [10] scores (BLEU-1 through BLEU-4), measuring n-gram precision at increasing granularity levels, and ROUGE scores [11] (ROUGE-1, ROUGE-2, and Recall-Oriented Understudy for Gisting Evaluation—Longest Common Subsequence [ROUGE-L]) evaluated both unigram or bigram overlap and longest common subsequences. Additionally, METEOR scores [12] provided a synonym-aware evaluation particularly valuable for medical terminology variations where “hypertension” and “high blood pressure” conveyed identical meaning.

Semantic similarity was measured using the official BERTScore [13] implementation (bert-score library), which computed F_1-scores based on contextualized embeddings from roberta large model, providing a more nuanced assessment than simple word overlap, and SentenceTransformer [18] similarity using all-MiniLM-L6-v2 to generate dense vector representations and calculate cosine similarity.

Cross-Model Validations

The external validation was conducted with 200 cases to verify that the benefit of detailed task instructions is not specific to one architecture. We used 4 additional models from different LLM families. We compared basic versus detailed prompting (without fine-tuning) for GPT-4 [19], GPT-3.5-turbo [20], Mistral-7B-Instruct [21], Qwen2.5-3B [22], and Llama-3.2-3B [23]. Preprocessing, prompts, and scoring followed the same protocol as the main study.

Statistical Analysis

The Wilcoxon signed rank [24] statistical test was used because our evaluation involved paired, nonparametric text data and did not assume a normal distribution. This test allowed us to effectively compare discharge instructions generated by 2 different models on the same sample of 3621 discharge summaries. Additionally, we calculated the rank-biserial correlation, which is necessary for computing the effect size. This metric helps quantify the practical significance of the observed differences. All analyses were performed using Python’s SciPy library, evaluating performance metrics (BLEU [10], ROUGE [11], METEOR [12], and BERTScore [13,15]) across the full evaluation dataset.

Qualitative Analysis

A targeted, case-based qualitative review was conducted. We first identified candidate instances using quantitative signals (eg, low ROUGE-L). From these samples, cases that reflected safety-relevant errors were purposively sampled. The illustrative cases were reported to contextualize model behavior. We did not estimate prevalence or conduct formal frequency estimates because of the difficulty of reliably labeling possible generation errors. Although various failure modes occurred in the discharge instructions generated by both models, our analysis was focused on 2 particularly concerning error types (omission and hallucination) because of their potential impact on patient safety and continuity of care. We defined omission as the absence of clinically salient information in the generated discharge instructions that was present in the reference text and was necessary for safe and effective patient care. We also defined hallucination as the introduction of clinical information in the generated discharge instructions that was not present in the reference text and was not entailed by it, especially when such additions could affect patient care. We presented discharge-instruction cases, where the basic and detailed models produced outputs demonstrating these errors.

Human Evaluation

Two consultant physicians evaluated 40 paired discharge instructions, with each assessing 20 cases within their specialty expertise. A plastic surgeon with microsurgery subspecialty evaluated surgery and neurology cases, while a family medicine

XSL•FO
RenderX

specialist evaluated medicine and orthopedics cases. Cases were drawn from 4 services (medicine, neurology, surgery, and orthopedics), 10 samples from each. Safety was rated in 3 categories: safe, safe but needs minor changes, and not safe at all. Ratings required explicit red-flag instructions. Outputs with clearly unsafe statements (eg, dose doubling, mixing alcohol with opioids or sedatives, advice to ignore worsening symptoms, or unrestricted immediate activity) were classified as not safe at all. For each case, physicians received the deidentified discharge summary and both model outputs and recorded their ratings. Analyses used descriptive counts and proportions only. Raters were blinded to model identity; both followed a predefined rubric (Multimedia Appendix 2). We did not compute interrater agreement due to split-by-specialty assignment.

Results

Model Performance

We evaluated the performance of 2 instruction-tuned models: the generic and tailored models. Table 4 presents the comprehensive evaluation results comparing the generic and tailored models across all metrics on the test set of 3621 discharge instructions.

Table 4. Comparison of automatic evaluation metrics for discharge instruction generation using basic versus detailed fine-tuning strategies[a].

Metric	Generic instructions+fine-tuning	Detailed instructions+fine-tuning[b]
METEOR[c]	15.33	18.47
BERTScore[d]	78.92	87.05
ROUGE-1[e]	16.59	42.72
ROUGE-2	1.97	45.84
ROUGE-L[f]	8.59	26.52
BLEU-1[g]	13.63	32.86
BLEU-2	4.46	33.35
BLEU-3	1.76	25.13
BLEU-4	0.81	21.24
SentenceTransformer similarity	11.91	74.90

[a]The detailed+fine-tuning model consistently outperforms the generic+fine-tuning model across all metrics, with notable improvement.

[b]Statistically significant improvement (P<.001) across all evaluation.

[c]METEOR: Metric for Evaluation of Translation With Explicit Ordering.

[d]BERTScore: Bidirectional Encoder Representations From Transformers Score.

[e]ROUGE: Recall-Oriented Understudy for Gisting Evaluation.

[f]ROUGE-L: Recall-Oriented Understudy for Gisting Evaluation—Longest Common Subsequence.

[g]BLEU: Bilingual Evaluation Understudy.

Performance by Metric Category

We analyzed performance across 3 metric categories: lexical overlap, semantic similarity, and embedding-based similarity.

Lexical Overlap (BLEU and ROUGE)

Surface-level metrics showed dramatic differences between the 2 models. The generic model achieved very low scores on higher-order n-gram metrics, with BLEU-4 at only 0.81% compared to 21.24% for the tailored model. However, for 4-word phrases (BLEU-4), the generic model dropped to only 0.81%, while the tailored model maintained 21.24%. This shows that the generic model struggles to generate correct long phrases (BLEU-4), while the tailored model performs much better.

Semantic Similarity (METEOR)

The detailed model on METEOR, which accounts for stemmed variants, achieved 18.47% versus 15.33% for the generic model. It indicates that detailed instructions enhance linguistic matches.

Embedding-Based Similarity (BERTScore and SentenceTransformer)

Our analysis showed clear differences between the 2 instruction strategies. The generic model achieved 78.92%, while the detailed model reached 87.05%. The high score of 87.05% achieved by the detailed model suggests that its outputs are clinically reliable and closely align with how physicians typically document discharge instructions. This finding is strongly confirmed by SentenceTransformer results, where the detailed model achieved 74.90% compared to only 11.91% for the generic model. Together, these metrics indicate that detailed instructions produce clinically appropriate discharge instructions.

Cross-model validations revealed that phrase-level (BLEU-4) and bigram overlap (ROUGE-2) improved for all models. However, ROUGE-L gains were modest but consistent. BERTScore changes were very small. Detailed instructions improved phrase-level accuracy and structural alignment consistently, while semantic similarity stayed stable (Table 5).

XSL•FO
RenderX

Table 5. Cross-model validation (n=200 per model) comparing basic and detailed prompting using GPT-4, GPT-3.5-turbo, Mistral-7B, Qwen2.5-3B, and Llama-3.2-3B.

Model and prompt	BLEU-4[a]	ROUGE-L[b]	BERTScore[c]
GPT-4 [19]			
Basic	0.018	0.131	0.809
Detailed	0.045	0.169	0.821
GPT-3.5-turbo [20]			
Basic	0.023	0.139	0.818
Detailed	0.037	0.157	0.819
Mistral-7B [21]			
Basic	0.024	0.131	0.811
Detailed	0.035	0.146	0.811
Qwen2.5-3B [22]			
Basic	0.014	0.114	0.806
Detailed	0.017	0.119	0.807
Llama-3.2-3B [23]			
Basic	0.019	0.128	0.802
Detailed	0.020	0.128	0.803

[a]BLEU: Bilingual Evaluation Understudy.

[b]ROUGE-L: Recall-Oriented Understudy for Gisting Evaluation—Longest Common Subsequence.

[c]BERTScore: Bidirectional Encoder Representations From Transformers Score.

Subgroup Performance by Service Category

We performed subgroup analysis across 6 clinical service categories representing 3621 discharge instructions: general medicine, general surgery, specialized surgery, neurosciences, obstetrics and gynecology, and psychiatry or emergency. Table 6 presents the performance improvements for each service category.

All service categories showed BERTScore gains with the tailored model, particularly surgical departments, where general surgery rose by +0.140 and specialized surgery by +0.152 (Table 6). Three categories showed that BLEU-1 declines, which we interpret as the model demonstrating true language understanding through paraphrasing rather than memorization. Without exception, all subgroups demonstrated enhanced semantic accuracy. These findings confirm that the detailed instruction approach generalizes well. It is not specialty-specific but rather improves documentation across the clinical spectrum.

Table 6. Performance improvements by clinical service category.

Service category	BLEU-1[a] (tailored-generic)	BLEU-4 (tailored-generic)	ROUGE-L[b] (tailored-generic)	BERTScore[c] (tailored-generic)
General medicine	–0.021	+0.003	+0.014	+0.075
General surgery	+0.024	+0.003	+0.086	+0.140
Specialized surgery	+0.015	+0.002	+0.072	+0.152
Neurosciences	–0.003	+0.008	+0.052	+0.111
OB/GYN[d]	+0.018	+0.004	+0.073	+0.116
Psychiatry or emergency	–0.027	+0.002	–0.002	+0.051

[a]BLEU: Bilingual Evaluation Understudy.

[b]ROUGE-L: Recall-Oriented Understudy for Gisting Evaluation—Longest Common Subsequence.

[c]BERTScore: Bidirectional Encoder Representations From Transformers Score.

[d]OB/GYN: obstetrics and gynecology.

XSL•FO
RenderX

Statistical Interpretation

Overview

The Wilcoxon signed rank test [24] was used to compare the performance of the generic and tailored models across key evaluation metrics. Results showed statistically significant improvements in the tailored model for all tested metrics ($P<.001$), with consistently large effect sizes as measured by the rank-biserial correlation. Table 7 displays the detailed results for all metrics.

Table 7. Statistical comparison of generic and tailored model performance on discharge instruction generation (n=3621).

Metric	*P* value	RBC[a] (effect size)
BLEU-1[b]	<.001	0.976
BLEU-2	<.001	0.999
BLEU-3	<.001	0.999
BLEU-4	<.001	0.999
ROUGE-1-F[c]	<.001	0.995
ROUGE-2-F	<.001	0.999
ROUGE-L-F	<.001	0.999
METEOR[d]	<.001	0.998
BERTScore[e]-F_1	<.001	0.881
BERTScore-Official	<.001	0.999

[a]RBC: rank-biserial correlation.

[b]BLEU: Bilingual Evaluation Understudy.

[c]ROUGE: Recall-Oriented Understudy for Gisting Evaluation.

[d]METEOR: Metric for Evaluation of Translation With Explicit Ordering.

[e]BERTScore: Bidirectional Encoder Representations From Transformers Score.

Results Interpretation

The interpretation of results is as follows:

- BLEU-4 scores, which measure 4-word phrase accuracy: The tailored model achieved a median score of 10.85% (IQR 0.07-0.27) compared to 0.39% (IQR 0.00-0.01) for the generic model ($P<.001$; $r=0.9997$), indicating a substantial improvement in generating accurate longer medical phrases.
- ROUGE-L scores: The tailored model had a median score of 23.73% (IQR 0.19-0.49) versus 8.69% (IQR 0.06-0.11) for the generic model (>$P<.001$; $r=0.9990$), reflecting better structural alignment with reference instructions.
- BERTScore scores: The tailored model reached a median similarity of 92.17% 0.86 (IQR 0.85-0.89) compared to 84.10% (IQR 76.0-81.0) for the generic model (>$P<.001$; $r=0.8807$), confirming improved semantic faithfulness.

The large effect sizes indicate that the observed improvements are not only statistically significant but also practically meaningful in real-world clinical contexts.

Distribution Analysis of Model Performance

We examined score distributions across all 3621 test samples. Table 8 reports percentile values (0%, 25%, 50%, 75%, 100%) for the 4 metrics: BLEU-1 (word overlap), BLEU-4 (phrase overlap), ROUGE-L-F (structural alignment), and BERTScore-Official (semantic similarity). Figure 4 visualizes the full percentile curves across metrics performance.

XSL•FO
RenderX

Table 8. Percentile distribution of BLEU-1[a], BLEU-4, ROUGE-L[b], and BERTScore[c] metrics for generic and tailored models (n=3621).

Metric and model	Range	Median (IQR)
BLEU-1		
Generic	0.0-0.61	0.13 (0.07-0.19)
Tailored	0.06-0.99	0.31 (0.24-0.48)
BLEU-4		
Generic	0.0-0.37	0.0 (0.0-0.01)
Tailored	0.0-0.99	0.11 (0.07-0.27)
ROUGE-L		
Generic	0.0-0.63	0.09 (0.06-0.11)
Tailored	0.06-1.0	0.24 (0.19-0.49)
BERTScore		
Generic	0.69-0.88	0.81 (0.76-0.81)
Tailored	0.78-0.98	0.86 (0.85-0.89)

[a]BLEU: Bilingual Evaluation Understudy.

[b]ROUGE-L: Recall-Oriented Understudy for Gisting Evaluation—Longest Common Subsequence.

[c]BERTScore: Bidirectional Encoder Representations From Transformers Score.

Figure 4. Percentile distributions of test-set scores (n=3621): (A) BERTScore, (B) ROUGE-L, (C) BLEU-4, and (D) BLEU-1. The detailed model (solid line) consistently exceeds the generic model (dashed line) across percentiles. BERTScore: Bidirectional Encoder Representations From Transformers Score; BLEU: Bilingual Evaluation Understudy; ROUGE-L: Recall-Oriented Understudy for Gisting Evaluation—Longest Common Subsequence.

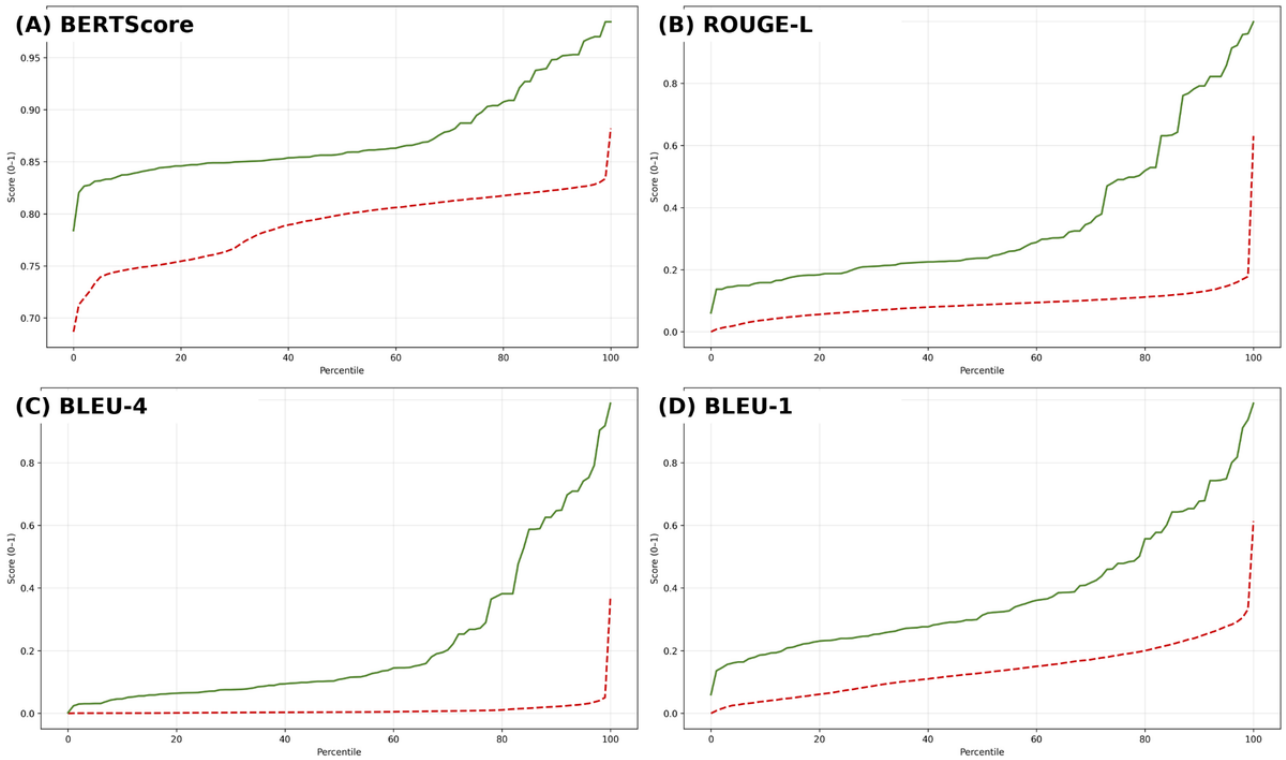

Clinical Relevance

From a medical standpoint, the improvements observed with the tailored model are substantial:

- BLEU-4 captures the model's ability to reproduce clinically relevant multiword expressions. The higher scores suggest that the tailored model better replicates the language typically used by health care providers in discharge documentation.
- ROUGE-L, which reflects the overall structural similarity to the reference text, indicates that the tailored model produces more coherent and well-organized discharge instructions, crucial for patient comprehension and safety.
- BERTScore, a semantic similarity measure, shows that the tailored model generates outputs that are more meaningfully

XSL•FO
RenderX

aligned with physician-written references, reducing the risk of ambiguity or misinformation.

These results suggest that fine-tuning with detailed, medically grounded instructions tailored to the hospital setting produces discharge instructions that are more accurate and clinically appropriate. This supports improved communication between care teams and patients. For a comparative example of model versus physician-written discharge instructions, see Multimedia Appendix 3. Illustrative best, worst, and typical cases, aligned with physician references and metric outputs, are also included.

Qualitative Error Analysis

We reviewed representative cases, in which the basic and detailed models demonstrated omission and hallucination errors. To illustrate the safety implications of these errors, we first present a medication-omission case (Textboxes 1-3).

Textbox 1. Case 1: comparison of basic and detailed model outputs showing improved capture of medication adjustments.

In the reference written by the physician, multiple medication changes were documented, including warfarin, Lovenox, metoprolol, atenolol, amlodipine, and hydrochlorothiazide. The following excerpt shows these critical modifications: "We made the following changes to your medications: • lowered the dose of the blood pressure medication Amlodipine • stop taking Hydrochlorothiazide (HCTZ) • stop taking Atenolol 100 mg daily (we started Metoprolol instead) • take Warfarin (the blood thinner) as directed • use Lovenox injections until Warfarin is at a therapeutic level ..."

The basic model's output contained no medication instructions, omitting all medication-related information entirely. Although the detailed model removed the explicit information about discontinuing atenolol, it captured the essential substitution with metoprolol and the critical anticoagulation therapy: "You were started on a new medication called metoprolol… continue your blood thinners called lovenox and warfarin." We then examined a hallucination case, where the output included clinical features absent from the reference.

Textbox 2. Case 2: example of physician-documented vertigo presentation demonstrating that the detailed model maintained diagnostic accuracy while the generic model introduced hallucinated symptoms.

In the ground truth, the physician documented a straightforward presentation of vertigo and nausea episodes occurring at the primary care office, with normal examination findings and negative imaging for stroke. The reference explicitly stated: "You were admitted with a few episodes of vertigo and nausea and vomiting while at your PCPs office ... MRI of your brain which did not show any evidence of acute stroke ... This symptom likely represents a peripheral vertigo."

The generic model fabricated entirely different symptoms. Instead of the vertigo case, it produced: "Chief Complaint: worsening aphasia and hallucination ... left-handed man with non-ischemic cardiomyopathy ..." This output contained multiple false features not found in the original reference, such as aphasia, hallucinations, and a complex cardiac problem. These features represent hallucinations that could misdirect care.

In contrast, the detailed model captured the actual clinical presentation: "You were admitted to the hospital for dizziness ... evaluated for a stroke but we did not find any evidence of this on your brain imaging. We also started you on a medication called valium which helped with your symptoms. You should only take this medication as needed for your dizziness." While the detailed model simplified the output, it remained aligned with the reference-documented vertigo presentation; however, it hallucinated a new medication (Valium) not present in the reference.

Textbox 3. Case 3: comparison showing the generic model's incorrect dietary recommendations versus the detailed model's clinically consistent output.

This case demonstrates diet guidance errors. In the ground truth, the physician documented a patient with Crohn disease recovering from small bowel obstruction. The key instructions included monitoring bowel function, and the reference noted: "You have tolerated a regular diet, are passing gas and your pain is controlled with pain medications by mouth."

The generic model generates a special diet: "gradually increase your intake of clear liquids and light foods ... Avoid foods that are high in fat, fiber, or are spicy." This contradicted the reference, which stated the patient was already "tolerated a regular diet."

The detailed model generated the clinical condition correctly and warning signs but simply did not address diet at all, which aligns with the reference document that also provided no specific dietary guidance beyond, noting the patient had tolerated a regular diet.

Beyond content errors, we observed that the models' writing styles often differed significantly from how physicians write discharge instructions. In our dataset, physician-authored instructions consistently used plain text without special formatting or Markdown symbols; accordingly, the generated text should adhere to institutional formatting conventions.

However, the basic model consistently ignored this convention. For example, it used "###" Markdown headers in 100% of outputs, elements that never appeared in the references. The detailed model performed notably better at matching clinical writing patterns. It generated text similar to physician notes and completely avoided the use of "###" Markdown headers (0%).

Clinical Safety Evaluation by Experts

Overall, the tailored model markedly improved safety compared with the generic model. For the generic model, ratings were safe in 3 of 40 (7.5%) cases, minor safety issues in 3 of 40 (7.5%) cases, and not safe in 34 of 40 (85%) cases. For the tailored model, ratings were safe in 16 of 40 (40%) cases, minor safety issues in 22 of 40 (55%) cases, and not safe in 2 of 40 (5%) cases.

Per-service analysis revealed consistent improvements across all departments (Table 9). Most notably, surgery services showed complete elimination of unsafe instructions with the tailored model compared to the generic model.

XSL•FO
RenderX

Table 9. Human evaluation of discharge instruction safety by clinical service and model type.

Service and model	Safe, n (%)	Safe but needs minor changes, n (%)	Not safe at all, n (%)
Orthopedics			
Generic	0 (0)	0 (0)	10 (100)
Tailored	3 (7.5)	7 (70)	0 (0)
Neurology			
Generic	3 (7.5)	0 (0)	7 (70)
Tailored	5 (50)	4 (40)	1 (10)
Surgery			
Generic	0 (0)	0 (0)	10 (100)
Tailored	4 (40)	6 (60)	0 (0)
Medicine			
Generic	0 (0)	3 (7.5)	7 (70)
Tailored	4 (40)	5 (50)	1 (10)
Totals (n=40)			
Generic	3 (7.5)	3 (7.5)	34 (85)
Tailored	16 (40)	22 (55)	2 (5)

Across the 4 clinical services, the tailored model produced discharge instructions ready for clinical use in 16 of 40 (40%) cases, with 22 of 40 (55%) cases requiring only minor edits. Despite this improvement, the 5% (2/40) unsafe rate confirms that physician review remains a critical safety step.

Discussion

Principal Findings

Despite using the same underlying base model (Mistral-NeMo-Instruct), the model fine-tuned with detailed, medically tailored instructions substantially outperformed the model fine-tuned with basic, generic instructions across all evaluation metrics. This highlights the critical importance of specific, clinically grounded instruction tuning that is aligned with the practices of the target hospital.

The improvements were observed across all the BLEU scores, where the tailored instruction model consistently outperformed the generic model. The tailored model generates words and phrases that closely match those found in human-written reference texts. It indicated its ability to generate context-appropriate discharge instructions.

The higher ROUGE-L score means that the ordering of words in the generated text closely matches the reference. This helps ensure that the discharge instructions are organized and easier to follow. The final result of ROUGE-L suggests that the tailored model produces discharge instructions with better logical flow, improved readability, and a style that closely follows how doctors write instructions for patients.

BERTScore reflects whether the generated text preserves the meaning of the reference text. For example, it recognizes that “take medicine twice daily” and “take medication 2 times per day” convey the same instruction. The high score (87%) shows that the tailored model captures medical meaning correctly. This semantic accuracy is crucial for patient safety. When the model generates instructions about drug interactions or warning signs, it must convey the correct medical message. Our results show that the tailored model does this reliably. In practical terms, physicians can trust that the generated instructions maintain clinical accuracy while reviewing and finalizing them.

The human evaluation findings provide important validation of our automated metrics while highlighting important safety considerations for clinical deployment. The physician safety assessment revealed that only 40% of tailored model outputs were immediately safe for patient use, with an additional 55% requiring minor modifications. Most concerning is that the 5% rated by physicians as unsafe raises important patient-safety concerns, underscoring the need for clinician oversight and continuous monitoring prior to clinical use.

Although both models used the same training data, the tailored model performed much better. The only difference was the design of the instructions. This shows how we instruct the model is important. The detailed instructions gave the model clear guidance about discharge instructions. It specified what information to include. The generic instructions missed this guidance.

This finding has practical value for health care systems. Each hospital can craft detailed prompts (instructions) that match its specific documentation standards and use these instructions for fine-tuning. The same base model adapts to each setting simply by changing the instructions and fine-tuning the model with the desired detailed instructions. This highlights an important shift in strategy: instead of collecting more training data or using larger models, health care systems can improve artificial intelligence performance through better instruction design. This approach is faster and less expensive than other methods.

XSL•FO
RenderX

These encouraging results prompted us to investigate service-specific performance. We examined whether the benefits of detailed instructions extended across 6 clinical service categories. Improvements were observed in all groups, but the size of the gains varied. Surgical specialties achieved BERTScore gains roughly twice that of general medicine. These gains make clinical sense because surgical discharge instructions mostly follow consistent procedural descriptions and template-like postoperative protocols, contrasting sharply with internal medicine, which needs to address multiple conditions, medication interactions, and individualized care plans.

Additionally, a revealing pattern appeared in general medicine, neurosciences, and psychiatry or emergency, where lower BLEU-1 scores coincided with higher BERTScore performance. This apparent contradiction actually represents a strength: the model learned to express medical concepts accurately without rigid adherence to source text phrasing.

All departments converged to the same performance range. Such consistent results justify implementing this across the entire institution.

Comparison With Earlier Works

Early approaches to discharge summary generation relied heavily on templates. While this method ensured consistency, it often lacked personalization and failed to capture patient-specific needs. The introduction of neural language models marked a significant shift in solving the problem of personalized patient instructions. An early study on natural language generation for EHRs established that encoder-decoder models could produce realistic emergency chief complaints text. These are short texts that describe why a patient came to the emergency room. This study demonstrated that neural networks could produce realistic text. However, this paper worked with short chief complaints, not full discharge instructions. In addition, the author used a very simple instruction with just 21 words: “You are a doctor. Generate discharge instructions for the following patient based on their discharge summary” [25].

Recent research has focused on improving generated discharge summaries using medical language models. The MEDISCHARGE study [26] fine-tuned Meditron-7B, a medical language model, to write discharge instructions. They made several technical improvements: extending the model’s reading capacity from 2000 to 6000 words and adding methods to extract information from different clinical sections. Their results on MIMIC-IV showed BLEU-4 of 0.103, ROUGE-L of 0.284, and BERTScore of 41.7%. However, their approach needs complex data processing and powerful computers, making it difficult for many hospitals to use [26].

Our work takes a different path by focusing on instruction quality rather than architectural complexity. We show that writing better instructions for the model can achieve improved results. Our instruction fine-tuning achieved BERTScore of 87.05% versus MEDISCHARGE’s 41.7% [26], showing the effectiveness of detailed instructions. We achieved this by giving the model clear instructions, which included specific medical guidelines, defined roles, and example texts. These results show that detailed instruction engineering alone can yield substantial improvements, offering a practical alternative for clinical text generation without requiring system-level modifications.

Another advanced approach is Re3Writer, developed by Liu et al [27]. This approach works in 3 steps. First, it finds similar patient cases. Second, it uses a medical knowledge graph database to understand medical relationships. Third, it combines all this information to write patient instructions. It achieved a BLEU-4 of 30.5 on the transformer model as a backbone. Despite these achievements, their approach requires maintaining a database of historical cases and constructing medical knowledge graphs, which added considerable complexity to the deployment pipeline [27]. A summary comparison of performance across earlier works and our approach is presented in Table 10.

XSL•FO
RenderX

Table 10. Performance comparison of discharge instruction generation methods.

Study or approach	Model or method	Dataset	BLEU-4[a]	ROUGE-L[b]	BERTScore[c] (%)	Key Limitations
Lee (2018) [25]	Encoder-decoder (early NLG[d] for EHR[e])	Chief complaints	__[f]	—	—	Worked only with short texts
MEDISCHARGE (Wu et al, 2024) [26]	Fine-tune a medical model (Meditron-7B) with dynamic section-based extraction	MIMIC-IV[g]	0.103	0.284	41.7	Complex preprocessing
Re3Writer (Liu et al, 2024) [27]	Retrieval+medical knowledge graph transformer backbone	MIMIC-III[h]	30.5	42.2	—	Requires maintaining knowledge graphs; deployment complexity
Our work (2025)—generic instructions+fine-tuning	Fine-tuned Mistral-NeMo fine-tuned with simple instructions	MIMIC-IV	0.81	8.59	78.92	Generated text lacked structure
Our work (2025)—detailed instructions+fine-tuning	Fine-tuned Mistral-NeMo fine-tuned with detailed instructions	MIMIC-IV	21.24	26.52	87.05	Generated text required clinician oversight

[a]BLEU: Bilingual Evaluation Understudy.

[b]ROUGE-L: Recall-Oriented Understudy for Gisting Evaluation—Longest Common Subsequence.

[c]BERTScore: Bidirectional Encoder Representations From Transformers Score.

[d]NLG: natural language generation.

[e]EHR: electronic health record.

[f]Not available.

[g]MIMIC-IV: Medical Information Mart for Intensive Care IV.

[h]MIMIC-III: Medical Information Mart for Intensive Care III.

Feasibility of Open-Source Models for Discharge Instruction Generation

The results achieved by our tailored model are particularly promising, demonstrating that high-quality discharge instruction generation is possible using an open-source, fine-tuned language model. Despite not relying on proprietary large-scale systems like GPT-4 [19], our tailored model produced outputs with significantly higher lexical, structural, and semantic accuracy across all metrics. Remarkably, our model achieved a BERTScore of 87.2%, which gives a strong indication of better alignment with doctor-written instructions. These findings underscore the potential of fine-tuned, open-source models to deliver clinically useful results. This makes the approach not only cost-effective and accessible but also compliant with health care–specific data standards. It offered a path for hospitals aiming to integrate generative AI into documentation workflows without compromising control, privacy, or quality.

Cross-Model Validation: Small-Model Benefits in Resource-Limited Settings

To explore whether smaller language models could benefit from our instruction design, we conducted a preliminary evaluation across multiple models using 200 discharge instructions. The preliminary results suggest promising possibilities for resource-constrained settings. In 200 cases, GPT-4 [19] showed the largest improvements (150% for BLEU-4), revealing that it has advanced instruction-following capabilities. Mistral-7B [21] reached ROUGE-L 0.146, close to GPT-3.5-turbo [20] at 0.157, while using much less compute. Qwen2.5-3B [22] improved BLEU-4 by 20.4% and ROUGE-2 by 20.8%. These results suggest the possibility that careful instruction design helps small models produce more accurate medical phrases and term pairs. Although these initial findings on 200 samples cannot confirm broad generalizability, they indicate that our approach of detailed instructions with compact models may have a promising direction to be deployed in resource-limited settings.

These findings are consistent with the broader understanding that larger instruction-tuned models tend to capture complex instructions more effectively than their smaller counterparts, while smaller models can still benefit significantly from optimized instruction formatting [9].

Workflow Integration and Time Efficiency

During our inference, the tailored model generated discharge instructions for a single patient case within 30-40 seconds. This represents a promising direction for assisting physicians with documentation tasks. In contrast, current manual discharge processes typically require 15-30 minutes of physician time per patient [6].

These inference times (30-40 seconds per case) are dependent on several technical factors, including network latency and the computational hardware used in our experimental setup. Importantly, our model examined cases one at a time in a controlled experimental environment. However, the total time savings in practice will depend on multiple factors beyond generation speed, including the hospital's available hardware resources, the time required for physicians to review and modify the drafted generated text, the frequency of cases requiring substantial versus minor edits during revision, and system

XSL•FO
RenderX

performance under real-world conditions with multiple simultaneous users.

This efficiency gain may ease documentation burdens, providing particular value in high-volume discharge settings. However, it is important that implementation efforts emphasize scalable system design and include comprehensive evaluations to assess real-world time savings and physician satisfaction.

The Critical Role of Clinician Oversight

Our evaluation revealed a 5% unsafe rate, demonstrating that oversight frameworks are not optional but essential in clinical deployment. This result illustrates that even with advanced LLMs, certain cases, particularly those involving complex patient needs or specific clinical reasoning, require close expert review. These medical cases provide a clear reminder that human judgment cannot be replaced and that LLM should be positioned as a supportive tool.

This finding affirms the value of human-LLM collaboration in drafting patient discharge instructions. The LLM accelerates the drafting process, while physicians refine the outputs by applying their clinical expertise. Such collaboration mitigates risk and transforms the LLM into an opportunity to save physicians' time. In addition, routine physician reviews create a feedback cycle that fosters both the model's strengths and areas for improvement. These reviews can enhance future model refinements and provide a guide for the model's optimization.

The persistent problem of LLM hallucinations further reinforces the necessity of human oversight. Without physicians' review, there is a chance to generate discharge instructions that misrepresent patient care plans, potentially compromising safety. At the current time, success depends on developing interaction models in which LLMs contribute efficiency, while physicians ensure safety and accuracy in the care plan.

Limitations

Our study was conducted under significant resource constraints, which limited the scope of experimentation. We were only able to fine-tune and evaluate a single open-source language model due to limited computing time and hardware access. As a result, the effects of instruction variation were not tested across different model sizes or architectures.

The inference time measurements were obtained on limited hardware resources during the experiment, where the model processed cases sequentially, one at a time. We did not conduct stress testing to evaluate performance under multiple concurrent cases, as would occur in a busy hospital with several physicians generating discharge instructions simultaneously. Furthermore, inference times in our setup depend on network and hardware conditions and may not generalize to clinical environments.

Our human evaluation was limited to 40 discharge summaries across 4 specialties. It was not large enough to capture the full variability of clinical practice. While this sample provided valuable potential insights, future studies should expand human evaluation with a larger sample size and more diverse specialties and calculate the formal interrater agreement metrics to better establish the reliability of clinical safety assessments.

Our evaluation is physician-centered, focusing on the goal of helping clinicians reduce their documentation burden. Future studies should therefore integrate patient-centered assessments. Such evaluations can capture aspects beyond clinical safety, including readability, freedom from criticism, avoidance of bias, encouragement to follow instructions, a welcoming tone that motivates patients to seek medical advice when needed, and clarity in presenting guidance on medications, activities, warning signs, and when to seek further help.

Conclusions

This study demonstrated that fine-tuning open-source LLMs with detailed, tailored instructions significantly enhanced the quality of generated discharge instructions compared to generic prompts. The statistical analysis across multiple evaluation metrics revealed substantial and practically meaningful improvements. The tailored, detailed model generated outputs closely aligned with real instructions written by doctors. These findings suggested the importance of customizable instructions in leveraging open-source language models in hospitals. The tailored instruction tuning method offered a promising, cost-effective, efficient, and scalable strategy to automate discharge instruction generation, thereby potentially reducing clinician workload and improving patient safety and comprehension in hospital care settings.

Acknowledgments

The authors thank the Massachusetts Institute of Technology Laboratory for Computational Physiology for providing access to the MIMIC-IV (version 2.2) database through PhysioNet [12] and many clinicians whose deidentified documentation made this research possible. No generative artificial intelligence tools were used in the writing, analysis, or generation of this manuscript. All content was authored by the research team.

Data Availability

The data used in this study were derived from the publicly available Medical Information Mart for Intensive Care IV database, which is hosted on PhysioNet [12]. Access to Medical Information Mart for Intensive Care IV requires completion of the Collaborative Institutional Training Initiative "Data or Specimens Only Research" course and acceptance of the PhysioNet Credentialed Health Data Use Agreement. Due to data use agreement restrictions and patient privacy considerations, the raw data cannot be shared directly. However, the datasets generated and analyzed during this study are reproducible using the described

XSL•FO
RenderX

cohort extraction criteria. The code for model fine-tuning and evaluation, along with implementation scripts for both the generic and detailed instruction-tuning approaches, is openly available on Zenodo [28].

Conflicts of Interest

None declared.

Multimedia Appendix 1
Complete text of the detailed instruction prompt (587 words) used for model fine-tuning.
[DOCX File , 24 KB - formative_v9i1e80917_app1.docx]

Multimedia Appendix 2
Discharge instruction safety rating rubric.
[DOCX File , 28 KB - formative_v9i1e80917_app2.docx]

Multimedia Appendix 3
Example discharge instructions comparing physician-written reference text with generic and tailored model outputs across best, median, and worst-case scenarios.
[DOCX File , 51 KB - formative_v9i1e80917_app3.docx]

References

1. Robertson SL, Robinson MD, Reid A. Electronic health record effects on work-life balance and burnout within the I3 Population Collaborative. J Grad Med Educ 2017 Aug;9(4):479-484 [FREE Full text] [doi: 10.4300/JGME-D-16-00123.1] [Medline: 28824762]
2. Overhage JM, McCallie D. Physician time spent using the electronic health record during outpatient encounters: a descriptive study. Ann Intern Med 2020 Feb 04;172(3):169-174. [doi: 10.7326/M18-3684] [Medline: 31931523]
3. Shanafelt TD, Dyrbye LN, Sinsky C, Hasan O, Satele D, Sloan J, et al. Relationship between clerical burden and characteristics of the electronic environment with physician burnout and professional satisfaction. Mayo Clin Proc 2016 Jul;91(7):836-848. [doi: 10.1016/j.mayocp.2016.05.007] [Medline: 27313121]
4. Mueller SK, Giannelli K, Boxer R, Schnipper JL. Readability of patient discharge instructions with and without the use of electronically available disease-specific templates. J Am Med Inform Assoc 2015 Jul;22(4):857-863. [doi: 10.1093/jamia/ocv005] [Medline: 25882030]
5. Kripalani S, LeFevre F, Phillips CO, Williams MV, Basaviah P, Baker DW. Deficits in communication and information transfer between hospital-based and primary care physicians: implications for patient safety and continuity of care. JAMA 2007 Feb 28;297(8):831-841. [doi: 10.1001/jama.297.8.831] [Medline: 17327525]
6. Cain CH, Neuwirth E, Bellows J, Zuber C, Green J. Patient experiences of transitioning from hospital to home: an ethnographic quality improvement project. J Hosp Med 2012;7(5):382-387. [doi: 10.1002/jhm.1918] [Medline: 22378714]
7. Meng X, Yan X, Zhang K, Liu D, Cui X, Yang Y, et al. The application of large language models in medicine: a scoping review. iScience 2024;27(5):109713 [FREE Full text] [doi: 10.1016/j.isci.2024.109713] [Medline: 38746668]
8. Reynolds L, McDonell K. Prompt programming for large language models: beyond the few-shot paradigm. 2021 Presented at: CHI '21: CHI Conference on Human Factors in Computing Systems; May 8-13, 2021; Yokohama, Japan p. 1-7. [doi: 10.1145/3411763.3451760]
9. Chung H, Hou L, Longpre S, Zoph B, Tai Y, Fedus W, et al. Scaling instruction-finetuned language models. J Mach Learn Res 2024;25:1-53 [FREE Full text]
10. Sanh V, Webson A, Raffel C, Bach S, Sutawika L, Alyafeai Z, et al. Multitask prompted training enables zero-shot generalization. ArXiv Preprint posted online on March 17, 2022 [FREE Full text] [doi: 10.48550/arXiv.2110.08207]
11. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature 2023;620(7972):172-180 [FREE Full text] [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
12. Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data 2023;10(1):1 [FREE Full text] [doi: 10.1038/s41597-022-01899-x] [Medline: 36596836]
13. Papineni K, Roukos S, Ward T, Zhu W. BLEU: a method for automatic evaluation of machine translation. 2002 Presented at: ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics; July 7-12, 2002; Philadelphia, PA, United States p. 311-318. [doi: 10.3115/1073083.1073135]
14. Lin CY. ROUGE: a package for automatic evaluation of summaries. 2004 Presented at: Text Summarization Branches Out; Septembert 30, 2025; Barcelona, Spain p. 74-81 URL: https://aclanthology.org/W04-1013/
15. Banerjee S, Lavie A. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. 2005 Presented at: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization; September 30, 2025; Ann Arbor, MI, United States p. 65-72 URL: https://aclanthology.org/W05-0909/

XSL•FO
RenderX

16. Zhang T, Kishore V, Wu F, Weinberger K, Artzi Y. BERTScorevaluating text generation with BERT. 2020 Presented at: International Conference on Learning Representations (ICLR); April 24-28, 2020; Singapore URL: https://arxiv.org/pdf/1904.09675
17. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform 2008:128-144. [Medline: 18660887]
18. Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using siamese BERT-networks. 2019 Presented at: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing; November 3-7, 2019; Hong Kong, China p. 3982-3992 URL: https://aclanthology.org/D19-1410/ [doi: 10.18653/v1/d19-1410]
19. OpenAI. GPT-4 technical report. ArXiv Preprint posted online on March 4, 2024 [FREE Full text] [doi: 10.48550/arXiv.2303.08774]
20. OpenAI. GPT-3.5-turbo model. OpenAI API Documentation. 2023. URL: https://platform.openai.com/docs/models [accessed 2025-10-03]
21. Mistral AI. Mistral-7B-Instruct-v0.2. Hugging Face. 2024. URL: https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2 [accessed 2025-10-03]
22. Qwen Team. Qwen2.5-3B-Instruct. Hugging Face. 2025. URL: https://huggingface.co/Qwen/Qwen2.5-3B-Instruct [accessed 2025-10-03]
23. meta-llama/Llama-3.2-3B-Instruct. Hugging Face. 2024. URL: https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct [accessed 2025-10-03]
24. Wilcoxon F. Individual comparisons by ranking methods. Biometrics Bull 1945;1(6):80-83. [doi: 10.2307/3001968]
25. Lee SH. Natural language generation for electronic health records. NPJ Digit Med 2018;1:63 [FREE Full text] [doi: 10.1038/s41746-018-0070-0] [Medline: 30687797]
26. Wu H, Boulenger P, Faure A, Céspedes B, Boukil F, Morel N, et al. EPFL-MAKE at “Discharge Me!”: an LLM system for automatically generating discharge summaries of clinical electronic health record. 2024 Presented at: Proceedings of the 23rd Workshop on Biomedical Natural Language Processing; September 30, 2025; Bangkok, Thailand p. 696-711. [doi: 10.18653/v1/2024.bionlp-1.61]
27. Liu F, Yang B, You C, Wu X, Ge S, Liu Z, et al. Retrieval-augmented and knowledge-grounded language models for faithful clinical medicine. ArXiv Preprint posted online on July 21, 2024 [FREE Full text]
28. Alqahtani M, Alotaibi FS, Al-Barakati A. Impact of detailed versus generic instructions on fine-tuned language models for patient discharge instructions generation: comparative statistical analysis. Zenodo. 2025. URL: https://doi.org/10.5281/zenodo.16035170 [accessed 2025-10-20]

Abbreviations

BLEU: Bilingual Evaluation Understudy
BERTScore: Bidirectional Encoder Representations From Transformers Score
EHR: electronic health record
GPU: graphics processing unit
HIPAA: Health Insurance Portability and Accountability Act
LLM: large language model
METEOR: Metric for Evaluation of Translation With Explicit Ordering
MIMIC-IV: Medical Information Mart for Intensive Care IV
ROUGE: Recall-Oriented Understudy for Gisting Evaluation
ROUGE-L: Recall-Oriented Understudy for Gisting Evaluation—Longest Common Subsequence

Edited by J Sarvestan; submitted 18.07.25; peer-reviewed by R Cheng, H Xu; comments to author 15.08.25; revised version received 02.09.25; accepted 30.09.25; published 30.10.25.

Please cite as:
Alqahtani M, Al-Barakati A, Alotaibi F, Al Shibli M, Almousa S
Impact of Detailed Versus Generic Instructions on Fine-Tuned Language Models for Patient Discharge Instructions Generation: Comparative Statistical Analysis
JMIR Form Res 2025;9:e80917
URL: https://formative.jmir.org/2025/1/e80917
doi:10.2196/80917
PMID:41032874

XSL•FO
RenderX

©Muneerah Alqahtani, Abdullah Al-Barakati, Fahd Alotaibi, Mohammed Al Shibli, Saad Almousa. Originally published in JMIR Formative Research (https://formative.jmir.org), 30.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Expectant Mothers' Reflections on Childhood and Parenting: Cross-Sectional Study

Bente Prytz Mjølstad[1], MD, PhD; Gunnhild Åberge Vie[1], MD, PhD; Linn Okkenhaug Getz[1], MD, PhD; Mari Hegnes Hansen[1], MD; Hanna Sandbakken Mørkved[1], MD; Bjarne Austad[1], MD, PhD

General Practice Research Unit, Department of Public Health and Nursing, Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology, Trondheim, null, Norway

Corresponding Author:
Bente Prytz Mjølstad, MD, PhD
General Practice Research Unit, Department of Public Health and Nursing
Faculty of Medicine and Health Sciences
Norwegian University of Science and Technology
Postboks 8905
Trondheim, null, N-7491
Norway
Phone: 47 91614182
Email: bente.mjolstad@ntnu.no

Abstract

Background: The significance of parenting practices for children's development and health has gained increased attention, aligning with life course perspectives on health. Adverse childhood experiences are widespread and linked to impaired parenting in adulthood. Understanding how expectant parents reflect on their childhoods, and how such reflections can be fostered as part of antenatal care, is essential for supporting healthier caregiving and helping to break cycles of intergenerational adversity.

Objective: This study aimed to explore pregnant women's perspectives on parenting and sources of support in light of their childhood experiences, including responses to a key question about what they want to pass on to their child.

Methods: A cross-sectional online survey was conducted in 2022 among pregnant women recruited through social media platforms, including Facebook groups and Instagram accounts. The Pearson chi-square test was used to examine associations between the assessment of childhood and 4 parenting-related variables. Free-text responses were analyzed for content, categorized, and quantified.

Results: Among 1402 pregnant women across Norway, 10% (139/1393) reported difficult childhoods. Among them, 28.1% (39/139) felt that their upbringing would significantly influence their parenting, compared to 15.9% (171/1074) of those with good childhoods ($P<.001$). Most participants expressed a desire to pass on positive values. For some, the key question prompted disclosure of difficult childhoods and a strong wish to break with their past. Approximately 10% (139/1389) of the participants reported they lacked someone to turn to for parenting advice. This was more common among those with difficult childhoods (37/139, 26.6%) than those with average (28/176, 15.9%) or good childhoods (74/1067, 6.9%; $P<.001$).

Conclusions: Expectant women's upbringing shapes their parenting expectations and support needs. Many, especially those with difficult pasts, expect these experiences to affect parenting. These findings highlight the role of intergenerational transmission and the value of inviting such reflection as part of antenatal care to identify support needs.

(JMIR Form Res 2025;9:e80060) doi:10.2196/80060

KEYWORDS

pregnancy; adverse childhood experiences; parenting; primary health care; general practice; person-centered care; key questions

Introduction

The growing recognition of the many factors that influence health across the life course [1] has led to increased emphasis on parenting as a key factor in children's development and well-being. Reflecting this trend, expectant parents are increasingly seeking antenatal education that offers early, realistic insights into parenting, along with access to support and guidance [2]. To foster healthy and supportive family environments, the Norwegian government's parental support

XSL•FO
RenderX

strategy places particular focus on meeting the needs of vulnerable families [3].

Early, unplanned pregnancies, often predisposed by adversity in childhood and adolescence [4], can increase the risk of both socioeconomic disadvantage and medical complications related to pregnancy, birth, and child health [5]. Moreover, parental adverse childhood experiences (ACEs) [6] have been associated with elevated parenting stress, with attachment style identified as a potential mediator [7-10]. Conversely, positive childhood experiences, such as feeling supported and perceiving the family as a safe haven, have cumulative benefits and can buffer the negative effects of adversities [11].

ACEs are prevalent, even in high-income countries. According to Norwegian studies, 4% to 7% of adults report having a difficult childhood, and approximately 20% lacked access to trusted adults during their upbringing [12-14]. ACEs are consistently associated with long-term challenges in health, education, and social functioning [6,15]. In Norway, the cumulative economic burden due to ACEs is estimated to account for approximately 2.7% of the gross domestic product [16], underscoring the critical importance of supportive parenting as a focus for public health promotion.

Reducing the long-term and intergenerational impact of ACEs is essential to prevent the past from becoming the prologue to the unborn child's future [17]. Traditional programs aimed at preventing child maltreatment have shown limited effects [18], and more effective approaches are warranted. A recent meta-analysis showing associations between expectant women's reflections about their unborn child and later parent-infant interaction quality indicates the relevance of fostering reflection on future parenting strategies during pregnancy [19].

In clinical settings, discussions about parenting and childhood adversity are often emotionally sensitive and require a thoughtful approach. Many first-time parents feel unprepared, frequently expressing concerns about relationships, parenting, and finances [20,21]. While health care providers are expected to discuss parenthood during antenatal care, how they do so is typically left to their discretion [22]. Experts recommend screening for maternal ACEs during pregnancy. Even if only a few cases may be identified, the potential benefits are substantial [23]. Standardized tools such as the ACE questionnaire can be used [5]. Alternatively, or in addition, a more open, nonjudgmental verbal invitation to reflect on one's upbringing may foster supportive parenting and help build trust, particularly for those in need of additional support. This study marks an initial step in exploring such an approach within the Norwegian context.

In general practice, key questions are a well-established method for initiating dialogues on complex topics in a noninvasive and reflective manner [24,25]. Rooted in patient-centered communication [24-26], such questions are designed to explore complex and sensitive issues with empathy and openness. Key questions are typically single, open-ended, inviting questions that offer structured support for nuanced conversations in general practice. Grounded in person-centered medicine, humanistic psychology, and the biopsychosocial model, these questions reflect a narrative orientation by inviting patients to share their stories, broaden perspectives, deepen reflection, and inspire meaningful change [24,25,27]. Despite the growing awareness of ACEs and the importance of early intervention, structured, person-centered tools remain limited in routine health care. In emotionally significant contexts such as the transition to parenthood, a thoughtfully framed single item may offer a simple yet meaningful entry point for dialogue and support, consistent with narrative and person-centered approaches in health care [27].

Inspired by the use of key questions, we developed and introduced a single open-ended question for expectant parents: "What do you want to pass on from your childhood?" While a single question may not capture the full nuances of multi-item instruments, it can prompt personal reflection and emotionally relevant responses. Previous studies have explored associations between ACEs and parenting, but few have examined these links during pregnancy in a general population sample using both structured and open-ended questions.

This study aimed to examine pregnant women's perspectives on parenting and sources of support in light of their childhood experiences. As part of this, we explored how they respond to a reflective key question about what they wanted to pass on to their child. To guide the analysis, we asked the following question: "How are self-reported childhood experiences associated with parenting expectations and perceived support–and what aspects of their upbringing do participants want to pass on to their child?"

Methods

Study Setting

In Norway, antenatal care and the public child health program are universal, free of charge, and primarily delivered through primary care services. Under the Regular General Practice Scheme, each citizen is assigned to a regular general practitioner (GP), and most GPs provide antenatal care in collaboration with municipal midwives. The Norwegian Public Child Health Program (for children aged 0-5 years) offers regular checkups, vaccinations, and parental support, mainly delivered by public health nurses in cooperation with GPs. Participation in the program is mandatory. Additionally, informal postnatal groups, organized by public nurses, bring together parents of similarly aged infants to share experiences and offer peer support. These groups often continue independently, fostering community and mutual learning over time. Some municipalities also offer birth and parenting preparation courses, though availability varies and is not universal.

Study Design, Participants, and Data Collection

Between January and March 2022, pregnant women in Norway who were aged 18 years and above were invited to participate in an open, anonymous, cross-sectional online survey. The questionnaire was distributed using Nettskjema, a data collection platform developed by the University of Oslo, which is user friendly and compatible with a wide range of digital devices. The study was promoted through an introductory video and various social media posts. Recruitment was conducted entirely through social media, leveraging 6 Facebook (Meta Platform) due groups and 5 (Meta Platform) accounts. These platforms

XSL•FO
RenderX

had a combined follower base of approximately 109,000, although an overlap between followers was likely. For context, 51,480 women gave birth in Norway in 2022. Following the survey launch on a new platform, all responses were received within 4 days. Data collection for each group was closed after 2 to 4 weeks, as no additional responses were submitted. A total of 1402 women completed the survey, with 62% (n=868) recruited via Facebook and 38% (n=534) via Instagram. The study adhered to the Checklist for Reporting Results of Internet E-Surveys (CHERRIES) [28], ensuring transparency and methodological rigor in online recruitment and data collection. The checklist is attached in the Multimedia Appendix 1.

The Questionnaire

The questionnaire included up to 57 items, depending on the respondent's answers, and covered five main areas: (1) demographic information; (2) the pregnancy; (3) antenatal care; (4) sources of advice and information; and (5) expectations for the postnatal period and the parental role including reflections on the respondent's childhood. Of the total items, 12 were open-ended questions with free-text responses. This paper reports findings related to demographics, pregnancy and antenatal care, sources of advice and support, and parental role. The questionnaire included both multiple-answer options and open-ended questions. The other findings are published separately [29,30]. An English translation of the complete questionnaire is attached in the Multimedia Appendix 2.

As no existing validated questionnaire fully aligned with the broader aims of the study, a new instrument was developed. It incorporated selected validated items from previous surveys, including the Cambridge Worry Scale, and questions on demographics, self-rated health, and childhood experiences from the large population-based Norwegian Trøndelag Health Study (HUNT) [29-31]. The single-item question assessing childhood experiences was originally developed for the HUNT study and has been validated, demonstrating good construct validity in capturing the impact of ACEs on adult quality of life [31]. Additionally, inspired by the concept of key questions, we developed a neutral, open-ended single-item question, "What do you want to pass on from your childhood?" This was done to encourage reflection on their childhood experiences and intergenerational transfer. Questions regarding parenting perspectives and supportive networks were developed specifically for this survey. The survey was pilot-tested among 6 pregnant women and did not result in any changes to the variables used in this analysis. Participants provided informed consent by submitting the completed questionnaire, which included an introductory information letter detailing the study's purpose, estimated completion time, and the voluntary nature of participation. The survey included a back button for answer revision.

Data Analysis and Study Variables

Data were analyzed using SPSS Statistics (IBM Corp). Descriptive statistics were calculated for the total sample and relevant subgroups. For analysis, we merged the assessment of a very good or good childhood into good and a difficult or very difficult childhood into difficult. Similarly, we merged very great or great extent into great extent and small or not at all into small extent for the inquiry about their thoughts and feelings regarding becoming a parent and how well they felt prepared through antenatal care. The Pearson chi-square test was used to examine associations between the assessment of childhood and the 4 questions concerning influence on raising children, preparation for parenthood, views on becoming a parent, and having someone to seek advice from. Most respondents answered all questions; data on childhood experiences were missing for 9 of 1402 (0.6%) participants. We performed complete case analyses and excluded those with missing responses from each analysis. We analyzed all free-text responses for content, categorized them, and counted the number of responses in each category. The open-ended key question was designed to encourage reflection rather than to identify ACEs directly. Responses were categorized thematically, focusing on values and experiences participants wished to pass on. While coding, we did not systematically assess the presence of ACEs; some responses were interpreted as reflecting adverse experiences, such as neglect, abuse, or instability, based on explicit language used by participants. We additionally used Pearson chi-square tests to assess whether responding to the open-ended key question was associated with demographic characteristics or assessment of own childhood.

Ethical Considerations

A remit assessment was sent to the Central Norway Regional Committee for Medical and Health Research Ethics, which determined that a formal application was not necessary (reference number 267956). The Norwegian Center for Research Data approved the study (reference number 134424), confirming that no direct or indirect personal identifiers were processed, including no use of cookies or collection of IP addresses. All methods were carried out following relevant national and institutional guidelines and regulations. Informed written consent was obtained from all participants.

Participation was voluntary, and participants could withdraw at any time without providing a reason. No compensation was offered for participation. Data were stored securely on encrypted servers with access restricted to the research team.

Results

Overview

A total of 1402 pregnant women, geographically distributed across all regions of Norway, participated in the survey; the demographic characteristics are presented in Table 1.

XSL•FO
RenderX

Table 1. Demographics and characteristics of the study population (N=1393-1402).

Variables	Participants, n (%)
Age (y)	
<25	119 (8.5)
25-37	1226 (87.4)
>37	57 (4.1)
Parity	
Primipara	722 (51.5)
Multipara	680 (48.5)
Gestational age[a]	
First trimester	99 (7.1)
Second trimester	776 (55.4)
Third trimester	525 (37.5)
Living with a spouse or partner	
Yes	1369 (97.6)
No	33 (2.4)
Education	
Lower secondary school	28 (2)
Upper secondary school	315 (22.5)
College or university (<4 y)	495 (35.3)
College or university (≥4 y)	564 (40.2)
Residence[b]	
Northern Norway	129 (9.2)
Central Norway	230 (16.4)
Western Norway	323 (23.1)
Eastern Norway	631 (45.1)
Southern Norway	87 (6.2)
Size of municipality[b]	
Rural area (<5000)	270 (19.3)
Town (5000-20,000)	331 (23.6)
City (>20,000)	800 (57.1)
Birthplace	
Norway	1336 (95.3)
Outside Norway	66 (4.7)
Cultural background[c]	
Norwegian	1332 (95.2)
Not Norwegian	15 (1.1)
Multicultural	52 (3.7)
Childhood[d]	
Very good	579 (41.3)
Good	497 (35.4)
Average	178 (12.7)
Difficult	102 (7.3)

XSL•FO
RenderX

Variables	Participants, n (%)
Very difficult	37 (2.6)

[a]Missing values: n=2.
[b]Missing value: n=1.
[c]Missing values: n=3.
[d]Missing values: n=9.

Perceptions of Parenthood in Light of Childhood Experiences

The majority (1076/1393, 77.2%) reported their childhood as very good or good, while 10% (139/1393) described it as difficult or very difficult. Figure 1 shows to what extent these women reported that their childhood would influence their parenting, whether antenatal visits prepared them for parenthood, and their views on becoming a parent. Of those with a difficult childhood, 28.1% (39/139) said their upbringing would influence their parenting to a great extent. In comparison, 6.7% (12/178) of the participants with average childhoods and 15.9% (171/1074) of the participants with good childhoods said their upbringing would influence their parenting to a great extent (χ^2_4=69.0, P<.001). Most participants (921/1337, 68.9%) claimed that antenatal checkups had not prepared them for parenthood, with minor differences based on childhood evaluation (χ^2_4=5.2, P=.27). First-time mothers reported this more often than those who had given birth before (506/696, 72.7% vs 415/641, 64.7%, P=.007, not shown in Figure 1). Half (740/1401, 52.8%) of the participants were looking forward to becoming parents, while nearly half (635/1401, 45.3%) had mixed feelings, and a small minority group (24/1401, 1.7%) felt apprehensive. More women with a good childhood (607/1075, 56.5%) looked forward to parenthood compared to those with average (74/178, 41.6%) or difficult (57/138, 41.3%) childhoods (χ^2_4=23.4, P<.001). There were no differences based on whether they were first-time mothers or not (P=.43).

Figure 1. To what extent pregnant women reported that their childhood would influence how they would raise their children, whether the antenatal visits had prepared them for parenthood, and how they viewed becoming a parent, according to how they viewed their own childhood.

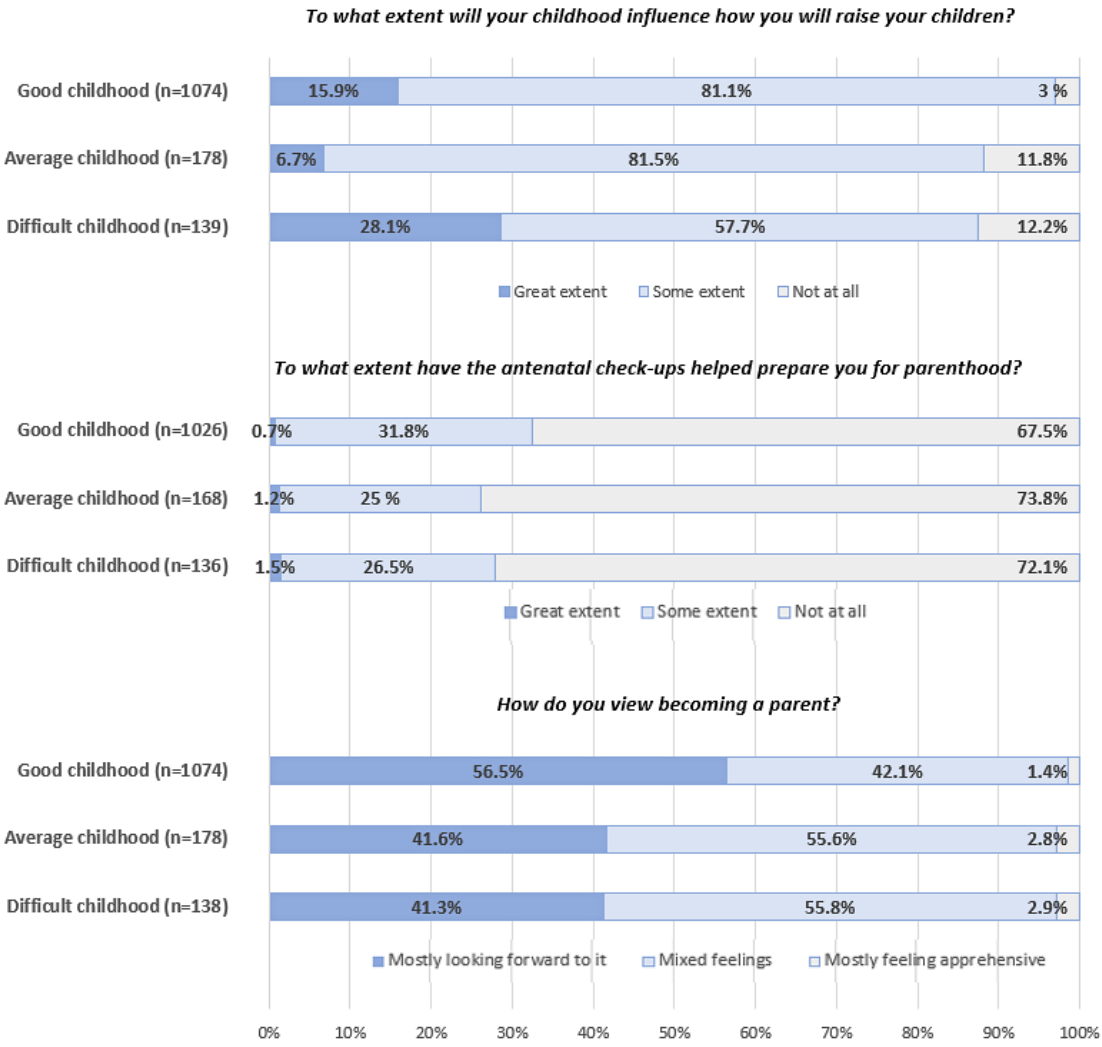

XSL•FO
RenderX

Supportive Networks

In total, 10% (139/1389) of respondents reported not having anyone to seek advice from about the parental role, while 90% (1250/1389) had at least 1 person. The proportion who had someone to consult was larger among those with a good childhood (993/1067, 93.1%) compared to those with an average childhood (148/176, 84.1%) and those with a difficult childhood (102/139, 73.4%; χ^2_2=60.3, P<.001).

Among the 826 (N=1402, 58.9%) women who answered an open-ended question on sources for advice, most mentioned family members and friends, with multiple sources reported. Women with a good childhood more frequently cited family members (574/648, 88.6%) compared to those with a difficult childhood (39/78, 50%). Conversely, women with a difficult childhood more often mentioned family-in-law (23/78, 30%) or health personnel (22/78, 28%) compared to those with a good childhood (95/648, 14.7% and 134/648, 20.7%) when asked whom they would seek advice from regarding parenting.

"What Do You Want to Pass on From Your Childhood?"

Approximately half (669/1402, 47.7%) of the participants responded to the open-ended key question, "What do you want to pass on from your childhood?" Most described positive experiences, such as safety (n=205, 30.6%), love (n=195, 29.1%), care (n=108, 16.1%), openness (n=60, 9%), and being "a present parent" (n=50, 7.5%; ie, being emotionally and physically available, as opposed to being absent).

However, among a smaller group of those who responded (40/669, 6%), the question led to the opposite response; they wrote about what they did not want to pass on, citing negative experiences such as abuse, neglect, and instability. Some participants strongly distanced themselves from their upbringing. A few respondents answered the question with a single word, "nothing," while others provided more detailed reflections, stating that their childhood would serve as a template for what they did not want for their child (Textbox 1).

Textbox 1. Selected responses to the open-ended question, "What do you want to pass on from your childhood?"

- "I've had an upbringing marked by neglect, so I know very well how I DON'T want to be a parent, and I've formed a clear picture of how I want to do the opposite!" [Participant 381]
- "I know what I shouldn't carry forward from a turbulent upbringing. What I do know is that it should be open to everything and filled with love and care." [Participant 374]
- "My upbringing serves as a template for what I don't want for my child." [Participant 396]
- "I haven't grown up in safety, so I need to provide my children with that. Stable relationships; I experienced how important it is for development, so I'm taking it with me. Be strong and steady in interactions with others is important for me, even though I wasn't met that way as a child." [Participant 183]
- "Minimal exposure to neglect and a violent father. Good and close care from grandparents, presence, and availability." [Participant 275]
- "The opposite. My child should avoid moving around, having multiple stepfathers, and attending different schools. It should have security and stability with me and the child's father." [Participant 220]
- "My child should never experience having their parent involuntarily committed to psychiatric care or something similar. It should never almost lose a parent due to attempted suicide." [Participant 236]
- "What was good, and, above all, do the opposite of what my parents failed at during my upbringing." [Participant 229]

Suggestions for Parenting Support in Antenatal Care

When participants were asked an open-ended question about how the health care system could best support parenting as part of antenatal care, responses varied widely. Approximately one-fourth of the total sample (334/1402, 23.8%) responded. Among them, several participants (110/334, 32.9%) requested more information—either in general or about specific topics (ie, breastfeeding, mental health, and partner relationships)—or access to parenting courses. Others (32/334, 9.6%) emphasized the need for increased dialogue or more personalized follow-up. Greater involvement of the partner was also explicitly requested by 19 (5.7%) participants, and an increase in the number of or longer antenatal checkups was wanted by 18 (5.4%) participants.

Discussion

Principal Findings

In this large, digitally recruited sample of pregnant women in Norway, approximately 1 in 10 participants reported having had a difficult or very difficult childhood. Among these, 28.1% (39/139) indicated that their upbringing would strongly influence their approach to parenting, compared to 15.9% (171/1074) of those with a good childhood. Approximately 10% (139/1389) of all respondents reported having no one to turn to for parenting advice, a situation more common among those with difficult childhoods.

Women with positive childhood experiences frequently cited family members as sources of support. Those with difficult childhoods more often mentioned in-laws and health professionals, although some also named specific relatives. Most participants responded to the open-ended key question, "What do you want to pass on from your childhood?" by expressing a desire to transmit positive values and experiences. However, for a smaller group of participants (6%), the question acted as an evocative prompt for disclosing difficult or traumatic aspects of their upbringing, often accompanied by a strong desire to break intergenerational patterns.

XSL•FO
RenderX

Participants also offered suggestions for improving support in antenatal care, including greater involvement of partners, enhanced postnatal follow-up, and more direct conversations about parenting. These findings suggest that the open-ended key question holds promise as a simple yet meaningful tool for initiating reflective and supportive dialogue in routine antenatal care.

Findings in Relation to Comparable Studies

One out of 10 participating pregnant women reported their childhood to have been difficult or very difficult. This is comparable to another Norwegian study by Haugland et al [13], where 10% of the women assessed their childhood as difficult or very difficult in response to the same question. The proportion of adults reporting the same in the HUNT study has been somewhat lower (4.1%-4.4%) but relatively stable over the last 2 decades [12,14]. Internationally, studies involving adults from various countries found that approximately 1 out of 4 had experienced at least 1 ACE [32,33].

The observed association between reporting a difficult childhood and indicating that their upbringing might significantly influence how they would raise their child aligns well with studies suggesting that the foundation for the parental role is established in one's upbringing. Furthermore, parents' ACEs can impact their parenting style in various ways; unresolved emotional issues can affect their ability to respond to their children's needs and create difficulties in forming secure attachment with their children [8,34].

One out of 10 respondents stated that they had no one to seek advice from regarding their parenting role, while 90% (1250/1389) had at least 1 person. This proportion was significantly lower among pregnant women who had a difficult childhood, and when asked whom they would seek advice from regarding parenting, they mentioned others outside their immediate family, such as in-laws and health care professionals, more frequently. This can be seen in light of previously presented figures from the HUNT study, where 21% reported growing up without a reliable person to seek support from [14]. These women might need more professional advice and guidance from health care personnel. Importantly, nearly 70% (921/1337) of respondents reported that the antenatal checkups had not helped prepare them for the parental role, regardless of their childhood background. This points to a broader potential for improvement in how health care services support expectant parents. Furthermore, only 1 in 4 respondents with a difficult childhood mentioned health care professionals as a source of importance, suggesting that vulnerable groups may not fully recognize or access the help available to them.

Knowledge about the childhood experiences of expectant parents can provide valuable insights for health care providers, allowing them to tailor follow-up and advice about the parental role during antenatal care, as part of preventive health care in public child health centers. Identifying resources within the pregnant women's networks is crucial, and health care personnel can also take initiatives to strengthen support systems such as postnatal groups.

Participants in our study also shared suggestions for greater involvement of partners. This aligns with a study by Solberg and Glavin [35], which demonstrated that many Norwegian fathers feel sidelined in antenatal care and the public child health care program and express a desire for more direct engagement and recognition of their caregiving role. Despite policy efforts to promote equality in parenting and family-centered services, structural and cultural barriers continue to limit fathers' participation [3,35].

Our study demonstrated that a digitally delivered, open-ended, and neutrally framed question encouraged expectant parents to reflect on their upbringing and, in some cases, disclose ACEs, a topic often considered sensitive. This approach appeared especially evocative for women with difficult childhoods, introducing a rare and safe space to express concerns and articulate a desire to break negative intergenerational patterns. The phrase thereby appears promising as a key question in the sense introduced earlier, that is, a meaningful entry point for person-centered dialogues [24,26]. With further development and validation, it could become part of a communicative approach aimed at identifying support needs and fostering reflective conversations in antenatal care.

Methodological Strengths and Limitations

This study has several strengths aligned with the CHERRIES [28], including a large sample size and the use of an anonymous and accessible online format that likely encouraged honest responses, even to sensitive questions. The questionnaire combined previously validated items with newly developed questions, which were pilot-tested for clarity and relevance.

The sample was broadly comparable to the general population of pregnant women in Norway, although there was a slight overrepresentation of first-time mothers, participants from central Norway, and those with higher education (Multimedia Appendix 3). Only 4.7% (66/1402) of participants were born outside Norway, limiting the generalizability of the findings. Missing data were minimal and considered negligible for the interpretation of results.

Participants were recruited through digital platforms and social media, which provided efficient access to the target population. However, this approach may have introduced self-selection bias, a recognized limitation in web-based surveys [28]. Due to the nature of social media recruitment, a traditional response rate could not be calculated.

Despite these limitations, the large and diverse sample enhances the reliability of findings. The broad scope of the questionnaire reduces the likelihood that participation was driven solely by interest in parenting or childhood experiences.

Considering both strengths and limitations, we regard the inclusion of a new single-item open question in the digital survey as a methodologically promising approach. Although only half (699/1402, 47.7%) of the patients responded, and response rates were higher among women with greater awareness of their upbringing and those with higher educational attainment, those who engaged provided meaningful reflection and elicited distilled insights into complex personal experiences. These findings suggest its applicability as a key question in

XSL•FO
RenderX

clinical settings, such as antenatal care and child welfare services. Overall, the study provides valuable insights into how digital tools can support the development of antenatal care and highlights the potential of social media for recruiting and engaging expectant parents in research on this matter.

Conclusions

This study underscores how pregnant women's childhood experiences shape their expectations and needs regarding parenting. Many, particularly those with difficult pasts, anticipated that these experiences would influence their parenting approach. While some reported a lack of adequate support, many also expressed a strong desire to pass on positive values, reflecting both vulnerability and resilience. On the basis of a digital survey, our findings suggest that reflective key questions may be a promising tool for exploring intergenerational influences among parents. Future research should focus on adapting and testing such questions in clinical practice to support early identification and guide tailored interventions.

Acknowledgments

The authors thank clinical psychologist and PhD candidate Ingvill Engvik Øvsthus for valuable advice and assistance with recruiting participants via the various social media platforms. BPM and BA are practicing general practitioners, specialists in general practice, and associate professors in the same field. GAV is a physician and associate professor of health services research. LOG is a physician, professor of behavioral science in medicine, and head of the General Practice Research Unit at Norwegian University of Science and Technology. We used the artificial intelligence tool Microsoft Copilot for language editing during manuscript preparation. The tool was not used for content generation, data analysis, or interpretation.

The research received no external funding.

Data Availability

The datasets used and analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

BPM, BA, LOG, MHH, and HSM collaboratively developed the questionnaire. MHH and HSM conducted the study as part of their Doctor of Medicine student thesis, under the supervision of BPM, BA, and LOG. BPM drafted the initial manuscript and conducted the qualitative analysis together with GAV. BA and GAV carried out statistical analysis. BPM, BA, GAV, and LG contributed to writing and critical revision of the manuscript. MHH and HSM contributed to this study as medical students at Norwegian University of Science and Technology, collaborating until their graduation in June 2023. All authors provided feedback and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
CHERRIES checklist.
[DOCX File , 20 KB - formative_v9i1e80060_app1.docx]

Multimedia Appendix 2
English questionnaire.
[DOCX File , 27 KB - formative_v9i1e80060_app2.docx]

Multimedia Appendix 3
Birth demographics for Norway in 2021.
[DOCX File , 16 KB - formative_v9i1e80060_app3.docx]

References

1. Halfon N, Larson K, Lu M, Tullis E, Russ S. Lifecourse health development: past, present and future. Matern Child Health J 2014 Mar;18(2):344-365 [FREE Full text] [doi: 10.1007/s10995-013-1346-2] [Medline: 23975451]
2. Entsieh AA, Hallström IK. First-time parents' prenatal needs for early parenthood preparation-A systematic review and meta-synthesis of qualitative literature. Midwifery 2016 Aug;39:1-11 [FREE Full text] [doi: 10.1016/j.midw.2016.04.006] [Medline: 27321714]
3. Helleland LH, Wara TM, Høie B, Sanner JT, Michaelsen Å. Trygge foreldre – trygge barn: regjeringens strategi for foreldrestøtte (2018–2021). Safe parents—safe children; government strategy for parental support (2018–2021). Regjeringens

XSL•FO
RenderX

Strategi for Foreldrestøtte. 2018. URL: https://www.regjeringen.no/contentassets/23fff4c08d0440ca84543ff74786c6a0/bld_foreldrestotte_strategi_21juni18.pdf [accessed 2025-10-30]
4. Drevin J, Hallqvist J, Sonnander K, Rosenblad A, Pingel R, Bjelland EK. Childhood abuse and unplanned pregnancies: a cross-sectional study of women in the Norwegian Mother and Child Cohort Study. BJOG 2020 Mar;127(4):438-446. [doi: 10.1111/1471-0528.16037] [Medline: 31802594]
5. Olsen JM. Integrative review of pregnancy health risks and outcomes associated with adverse childhood experiences. J Obstet Gynecol Neonatal Nurs 2018 Nov;47(6):783-794. [doi: 10.1016/j.jogn.2018.09.005] [Medline: 30308147]
6. Felitti VJ, Anda RF, Nordenberg D, Williamson DF, Spitz AM, Edwards V, et al. Relationship of childhood abuse and household dysfunction to many of the leading causes of death in adults. The Adverse Childhood Experiences (ACE) Study. Am J Prev Med 1998 May;14(4):245-258. [doi: 10.1016/s0749-3797(98)00017-8] [Medline: 9635069]
7. Iyengar U, Kim S, Martinez S, Fonagy P, Strathearn L. Unresolved trauma in mothers: intergenerational effects and the role of reorganization. Front Psychol 2014 Sep 01;5:966 [FREE Full text] [doi: 10.3389/fpsyg.2014.00966] [Medline: 25225490]
8. Lange BC, Callinan LS, Smith MV. Adverse childhood experiences and their relation to parenting stress and parenting practices. Community Ment Health J 2019 May;55(4):651-662 [FREE Full text] [doi: 10.1007/s10597-018-0331-z] [Medline: 30194589]
9. Steele H, Bate J, Steele M, Dube SR, Danskin K, Knafo H, et al. Adverse childhood experiences, poverty, and parenting stress. Canadian Journal of Behavioural Science / Revue canadienne des sciences du comportement 2016 Jan;48(1):32-38. [doi: 10.1037/cbs0000034]
10. Moe V, von Soest T, Fredriksen E, Olafsen KS, Smith L. The multiple determinants of maternal parenting stress 12 months after birth: the contribution of antenatal attachment style, adverse childhood experiences, and infant temperament. Front Psychol 2018 Oct 23;9:1987 [FREE Full text] [doi: 10.3389/fpsyg.2018.01987] [Medline: 30405481]
11. Bethell C, Jones J, Gombojav N, Linkenbach J, Sege R. Positive childhood experiences and adult mental and relational health in a statewide sample: associations across adverse childhood experiences levels. JAMA Pediatr 2019 Nov 01;173(11):e193007 [FREE Full text] [doi: 10.1001/jamapediatrics.2019.3007] [Medline: 31498386]
12. Tomasdottir MO, Sigurdsson JA, Petursson H, Kirkengen AL, Krokstad S, McEwen B, et al. Self reported childhood difficulties, adult multimorbidity and allostatic load. A cross-sectional analysis of the Norwegian HUNT Study. PLoS One 2015 Jun 18;10(6):e0130591 [FREE Full text] [doi: 10.1371/journal.pone.0130591] [Medline: 26086816]
13. Haugland SH, Dovran A, Albaek AU, Sivertsen B. Adverse childhood experiences among 28,047 Norwegian adults from a general population. Front Public Health 2021 Jul 26;9:711344 [FREE Full text] [doi: 10.3389/fpubh.2021.711344] [Medline: 34381754]
14. HUNT databank. URL: https://hunt-db.medisin.ntnu.no/hunt-db/variablelist [accessed 2025-10-30]
15. Kreshpaj B, Elsenburg LK, Andersen SH, De Vries TR, Thielen K, Rod NH. Association between childhood adversity and use of the health, social, and justice systems in Denmark (DANLIFE): a nationwide cohort study. Lancet Public Health 2025 Jan;10(1):e29-e35 [FREE Full text] [doi: 10.1016/S2468-2667(24)00242-1] [Medline: 39681123]
16. Hughes K, Ford K, Bellis MA, Glendinning F, Harrison E, Passmore J. Health and financial costs of adverse childhood experiences in 28 European countries: a systematic review and meta-analysis. Lancet Public Health 2021 Nov;6(11):e848-e857 [FREE Full text] [doi: 10.1016/S2468-2667(21)00232-2] [Medline: 34756168]
17. SmithBattle L. The past is prologue? The long arc of childhood trauma in a multigenerational study of teen mothering. Soc Sci Med 2018 Nov;216:1-9. [doi: 10.1016/j.socscimed.2018.09.013] [Medline: 30245301]
18. Euser S, Alink LR, Stoltenborgh M, Bakermans-Kranenburg MJ, van IJzendoorn MH. A gloomy picture: a meta-analysis of randomized controlled trials reveals disappointing effectiveness of programs aiming at preventing child maltreatment. BMC Public Health 2015 Oct 18;15(1):1068 [FREE Full text] [doi: 10.1186/s12889-015-2387-9] [Medline: 26476980]
19. Foley S, Hughes C. Great expectations? Do mothers' and fathers' prenatal thoughts and feelings about the infant predict parent-infant interaction quality? A meta-analytic review. Developmental Review 2018 Jun;48:40-54. [doi: 10.1016/j.dr.2018.03.007]
20. Deave T, Johnson D, Ingram J. Transition to parenthood: the needs of parents in pregnancy and early parenthood. BMC Pregnancy Childbirth 2008 Jul 29;8:30 [FREE Full text] [doi: 10.1186/1471-2393-8-30] [Medline: 18664251]
21. Wilska A, Rantanen A, Botha E, Joronen K. Parenting fears and concerns during pregnancy: a qualitative survey. Nurs Rep 2021 Nov 07;11(4):891-900 [FREE Full text] [doi: 10.3390/nursrep11040082] [Medline: 34968276]
22. Nasjonal faglig retningslinje: svangerskapsomsorgen. National clinical practice guideline. The Norwegian Directorate of Health. 2019. URL: https://www.helsedirektoratet.no/retningslinjer/svangerskapsomsorgen [accessed 2025-10-30]
23. Watson C, Kathryn RK, Nancy G, Kelly YW. Promising practices for implementing adverse childhood experiences and resilience screening in obstetric care. J Womens Health (Larchmt) 2022 Oct;31(10):1377-1379 [FREE Full text] [doi: 10.1089/jwh.2022.0260] [Medline: 36108330]
24. Hunskår S, editor. Allmennmedisin.4.utg. Oslo, Norway: Gyldendal Akademisk; 2023.
25. Malterud K. Key questions--a strategy for modifying clinical communication. Transforming tacit skills into a clinical method. Scand J Prim Health Care 1994 Jun;12(2):121-127 [FREE Full text] [doi: 10.3109/02813439409003686] [Medline: 7973190]

XSL•FO
RenderX

26. Mjølstad BP. Knowing patients as persons. A theory-driven, qualitative study of the relevance of person-related knowledge in primary health care. Norwegian University of Science and Technology. 2015. URL: https://tinyurl.com/mmkyk6yc [accessed 2025-10-30]
27. Launer J. Narrative-Based Practice in Health and Social Care: Conversations Inviting Change. London, UK: Routledge; Feb 06, 2018.
28. Eysenbach G. Improving the quality of Web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34 [FREE Full text] [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
29. Hansen MH, Mørkved HS, Austad B, Vie G, Getz LO, Mjølstad BP. Worries and information seeking during pregnancy: a cross-sectional study among 1402 expectant Norwegian women active on social media platforms. Scand J Prim Health Care 2025 Jun;43(2):488-499 [FREE Full text] [doi: 10.1080/02813432.2025.2461036] [Medline: 39953944]
30. Austad B, Vie G, Hansen MH, Mørkved HS, Getz LO, Mjølstad BP. Association of self-rated health in pregnancy with maternal childhood experiences, socioeconomic status, parity, and choice of antenatal care providers: cross-sectional study. JMIR Form Res 2025 Jun 03;9:e68811 [FREE Full text] [doi: 10.2196/68811] [Medline: 40460308]
31. Vederhus JK, Timko C, Haugland SH. Adverse childhood experiences and impact on quality of life in adulthood: development and validation of a short difficult childhood questionnaire in a large population-based health survey. Qual Life Res 2021 Jun;30(6):1769-1778 [FREE Full text] [doi: 10.1007/s11136-021-02761-0] [Medline: 33534031]
32. Madigan S, Deneault AA, Racine N, Park J, Thiemann R, Zhu J, et al. Adverse childhood experiences: a meta-analysis of prevalence and moderators among half a million adults in 206 studies. World Psychiatry 2023 Oct;22(3):463-471 [FREE Full text] [doi: 10.1002/wps.21122] [Medline: 37713544]
33. Bellis MA, Hughes K, Ford K, Ramos Rodriguez G, Sethi D, Passmore J. Life course health consequences and associated annual costs of adverse childhood experiences across Europe and North America: a systematic review and meta-analysis. Lancet Public Health 2019 Oct;4(10):e517-e528 [FREE Full text] [doi: 10.1016/S2468-2667(19)30145-8] [Medline: 31492648]
34. Wattanatchariya K, Narkpongphun A, Kawilapat S. The relationship between parental adverse childhood experiences and parenting behaviors. Acta Psychol (Amst) 2024 Mar;243:104166 [FREE Full text] [doi: 10.1016/j.actpsy.2024.104166] [Medline: 38295656]
35. Solberg B, Glavin K. Fedre ønsker en mer aktiv rolle i svangerskapsomsorgen og på helsestasjonen. Fathers want a more active role in antenatal care and at child health centres. Sykepleien Forskning 2018;13(72006):e72006 [FREE Full text] [doi: 10.4220/sykepleienf.2018.72006]

Abbreviations

ACE: adverse childhood experience
CHERRIES: Checklist for Reporting Results of Internet E-Surveys
GP: general practitioner
HUNT: Trøndelag Health Study

Edited by A Mavragani; submitted 03.07.25; peer-reviewed by T Harpel; comments to author 19.09.25; revised version received 25.09.25; accepted 20.10.25; published 07.11.25.

Please cite as:
Mjølstad BP, Vie GÅ, Getz LO, Hansen MH, Mørkved HS, Austad B
Expectant Mothers' Reflections on Childhood and Parenting: Cross-Sectional Study
JMIR Form Res 2025;9:e80060
URL: https://formative.jmir.org/2025/1/e80060
doi:10.2196/80060
PMID:

©Bente Prytz Mjølstad, Gunnhild Åberge Vie, Linn Okkenhaug Getz, Mari Hegnes Hansen, Hanna Sandbakken Mørkved, Bjarne Austad. Originally published in JMIR Formative Research (https://formative.jmir.org), 07.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Co-Designing a Digital Solution for Decreasing Loneliness and Social Isolation Among Older People in Sweden: Explorative Study

Caroline Eklund[1], PhD; Viktoria Zander[2], PhD; Annelie K Gusdal[2], PhD; Charlotta Åkerlind[1], PhD; Petra von Heideken Wågert[1], PhD

[1]School of Health, Care and Social Welfare, Mälardalen University, Västerås, Sweden

[2]School of Health, Care and Social Welfare, Mälardalen University, Eskilstuna, Sweden

Corresponding Author:
Caroline Eklund, PhD
School of Health, Care and Social Welfare
Mälardalen University
Box 883
Västerås 72123
Sweden
Phone: 46 (0)736484426
Email: caroline.eklund@mdu.se

Abstract

Background: Older people are particularly vulnerable to loneliness and social isolation due to common age-related changes. The ability to maintain social relationships is considered important for health and well-being and is an essential aspect of healthy aging. The use of information and communication technology has been shown to promote social connectedness and social support among older people; however, many existing solutions require already established contacts and are not developed based on expressed needs among older people experiencing loneliness or social isolation.

Objective: The overarching aim of this study was to develop a prototype of a health and welfare technology intervention for decreasing loneliness and social isolation among older people.

Methods: This study describes an iterative participatory design process conducted in 3 phases to create a prototype of a digital intervention for decreasing loneliness and social isolation in older people through participatory design procedures with several key stakeholders. Phase I explored social service staff's perceptions of how social isolation and experiences of loneliness among older people can be reduced by health and welfare technology. Data were collected by conducting 2 workshops (WSs) with social service staff. Phase II explored needs related to loneliness and social isolation perceived by older people and how these needs may be targeted by using health and welfare technology. Data were collected through 2 semistructured interviews and 2 WSs. In phase III, 3 co-design WSs brought together older people with experience of loneliness or social isolation, social service staff, and IT experts to collaboratively generate and refine design ideas.

Results: The main result of the study was the development of a paper prototype of the Fik@ room, which is a digital solution supporting social interaction among older people experiencing loneliness or social isolation. Key needs identified include the desire for low-threshold opportunities to connect, flexible modes of communication, and a safe and welcoming environment. These needs directly informed the final design. The main feature of the prototype is to facilitate social interaction in a digital break room with coffee tables where users can interact via video, audio, or chat. Other features are a bulletin board for announcements about interactions inside or outside of the Fik@ room, as well as a profile page.

Conclusions: The iterative co-design process ensured that the solution was grounded in user needs, which may have contributed to the usability and acceptability of the Fik@ room and potentially effective in reducing loneliness and social isolation among older people. Future research should focus on evaluating effects in a real-world setting.

(JMIR Form Res 2025;9:e78213) doi:10.2196/78213

KEYWORDS

action research; digitalization; eHealth; healthy aging; internet intervention; participatory research design

XSL•FO
RenderX

Introduction

The rapidly growing aging population poses a global demographic challenge [1,2]. This challenge is strongly tied to the health and well-being of the older population, and these qualities are influenced by multiple factors, including psychosocial factors such as social isolation and loneliness [3-5]. Loneliness and social isolation are multidimensional concepts that are sometimes used interchangeably, yet they have distinct meanings. Although there is no single universally agreed-upon definition of loneliness, one of the most widely cited definitions is provided by Perlman and Peplau [6], who conceptualize loneliness as "the unpleasant experience that occurs when a person's network of social relations is significantly deficient in quantitatively or qualitatively". This definition highlights that loneliness is inherently subjective and negative, which clearly differentiates it from objective social isolation, which may also be voluntary. While loneliness relates to a subjective feeling of lost friendship and support as a result of few or insufficient social relationships and interactions [7], social isolation is usually characterized as a lack of social networking and social support [8]. Although social isolation may be voluntary, loneliness is always involuntary by definition. Regardless of which term is used, loneliness in the older population elicits concern, and loneliness has been described as being more prominent among people aged 75-85 years than among people aged 57-65 years [9]. The estimated prevalence of loneliness in older adults (aged 60 years and older) has varied from 7% [10] to as much as 49% in the oldest adults (aged 80 years and older) [11].

The ability to maintain social relationships is considered important for health and well-being and is an essential part of healthy aging [1,2]. Few or insufficient social relationships are threats to the mental and physical health of the older population and are associated with lower levels of self-rated physical health [9], as well as with a range of health outcomes, such as malnutrition, hypertension, disrupted sleep, frailty, and increased mortality [12-16]. The biological mechanisms that have been proposed to mediate the link between loneliness and poor health conditions include genetic, neural, and hormonal mechanisms related to various physiological functions, including dysregulated hypothalamic pituitary adrenal axis function and immune function [13,17]. Moreover, experienced loneliness has been shown to predict nursing home admissions [18] and is related to poor emotional and cognitive health, ranging from psychoses, personality disorders, depression, and suicide to impaired cognitive performance and the risk of Alzheimer disease [19].

Older people are particularly vulnerable to loneliness and social isolation due to common age-related changes and losses, such as poor health status, poor functional status, poor vision, and loss of hearing [20]. Moreover, older people use the internet to a lesser degree than younger people do, as they tend to prefer face-to-face and telephone contact [21]. However, this is only one of several factors influencing lower levels of social technology use. Barriers also include low perceived self-efficacy, fear of technology, lack of social capital, and physical limitations, which may hinder older peoples' motivation and ability to engage with social technology [22]. This indicates that those individuals not able to be active outside of their homes have risks of feeling excluded in two manners, including both by physical contact and by digital exclusion from a digitally dominated society [23]. The use of information and communication technology (ICT) has been shown to promote social connectedness and social support among older people through the development and maintenance of social networks. ICT use has been shown to decrease feelings of loneliness [24] and social isolation and foster well-being, life satisfaction, and self-worth, as well as enhance social support, social contacts, and social networks [3,25], with a particular favoring of videoconferencing [26]. However, it is important that the development of ICT considers the specific needs of older people [22,25,27]. Models such as the Unified Theory of Acceptance and Use of Technology [28] and the Senior Technology Acceptance Model [29] highlight that perceived usefulness and ease of use are important for adoption of new technologies among older people, but that social influence, prior experience, trust, and facilitating conditions (eg, self-efficacy for use, attitudes, age-specific considerations like physical and cognitive abilities, support, and accessibility) also play crucial roles in determining whether older people will engage with new technology. These models are particularly relevant when designing interventions aimed at reducing loneliness and social isolation, as successful adoption and implementation rely not only on addressing functional needs but also on aligning with users' expectations, capabilities, and social contexts.

One possible method of achieving success when developing interventions for a specific group is to involve potential end users during the development process [30]. Participatory approaches, such as co-design by involvement of end users and stakeholders collaborating in the solution development process, are common in health care and have been applied in the development of a number of ICT tools to support self-care for people with chronic conditions [31,32]. Co-design in research has been described as "doing research 'with' or 'by' the public rather than 'to', 'about', or 'for' the public" [33] and has been argued to have the potential to improve the quality, relevance, and impact of health research [30,33]. Furthermore, there are also moralistic arguments for participatory research and co-design, as the public has the right to be involved in any publicly funded research that may impact them [33]. Although co-design principles have been used less with older people, there is currently a growing interest in involving older people in the development of solutions to maintain well-being and independence, as well as to promote healthy aging [32]. ICT can support safe and active aging; however, its acceptance and use can be suboptimal without consideration of the end users' needs and preferences [34]. Research suggests that ICT solutions that have been developed in collaboration with end users can have a positive impact on a number of health outcomes, such as the ability to cope with a disease, disease control, reductions in falls, improvements in the care experience, better access to health care, improved patient satisfaction, and reductions in costs [32]. Moreover, the solutions have been perceived to be more applicable and acceptable to end users [30], which may facilitate implementation.

XSL•FO
RenderX

A previous study identified loneliness and social isolation as being important contributing factors to decreased health and well-being among older persons; furthermore, health and welfare technology developed for older people was identified as having the potential to reduce loneliness [35]. Health and welfare technology comprises the umbrella term ICT and is defined as "a technology-based intervention that aims at maintaining or promoting health, well-being, quality of life, and increasing efficiency in the service delivery system of welfare, social and health care services, while improving working conditions of the staff" [36]. The development of a health and welfare technology solution for decreasing loneliness and social isolation via co-design may increase the acceptability of this solution to potential end users. Moreover, by involving the staff who meet these people in the design, the solution may be easier to implement in the organization. Experts within the field, such as researchers and system developers, may also contribute to broadening the perspective to suggest solutions and to ensure an evidence base.

The overarching aim of this study was to develop a prototype of a health and welfare technology intervention for decreasing loneliness and social isolation among older people. The specific aims of the different phases of the study were (1) to explore social service staffs' perceptions of how experiences of loneliness and social isolation among older people can be reduced by health and welfare technology; (2) to explore needs related to loneliness and social isolation perceived by older people and how these needs may be targeted by using health and welfare technology; and (3) to, in co-design with older people, social service staff, system developers and researchers, develop a prototype of a health and welfare technology intervention for social interaction to reduce experienced loneliness and social isolation among older people.

Methods

Overview

This study had an explorative design using a participatory research design based on a model for experience-based co-design by Robert [37]. According to the 3 specific aims, the study was conducted in 3 phases with separate data collection methods and analyses of the data and results. We used the COREQ (Consolidated Criteria for Reporting Qualitative Research) item checklist for reporting the study (Multimedia Appendix 1 [38]).

Robert [37] described the process of experience-based co-design in six stages: (1) setting up the design, (2) engaging staff and gathering experiences, (3) engaging patients/staff and collecting experiences, (4) gathering patients and staff to share experiences and begin the co-design process, (5) providing detailed co-design activities, and (6) gathering together again for celebration, review and renewal. Stage 1 included establishing governance and project management arrangements and was conducted prior to data collection by a core team consisting of researchers, social service staff, a system developer, and the intended workshop (WS) leader. Stage 6 was not included in this study.

The data were mainly collected through WSs with key stakeholders, including social service staff, older persons, system developers, and researchers. Interviews in phase II were conducted by the second author, the WSs in phase I were conducted by authors 3 and 4, and the WSs in phases II and III were led by an experienced WS leader from an IT company. All researchers/authors have previous experience in qualitative methodology and conducting interviews. Participants received no compensation for participating in any of the phases.

In phase I, the WSs were conducted at the researchers' university and lasted approximately 3 hours each. In phase II, the WSs were held in a conference room at the municipality's Family Caregiver Centre, which is separate from the university, with a duration of approximately 2 hours per WS. The individual interviews in phase II were conducted at the researchers' university and lasted for approximately one hour each. The phase III WSs were again conducted at the researchers' university, each lasting about 2 hours.

The process was iterative, and the results were synthesized between WSs and phases. Figure 1 shows an overview of the WSs included in each phase. Methods for each phase are presented below.

XSL•FO
RenderX

Figure 1. The 3 phases for data collection, the number of participants in each workshop, and how the phases are related to Robert's process of experience-based co-design.

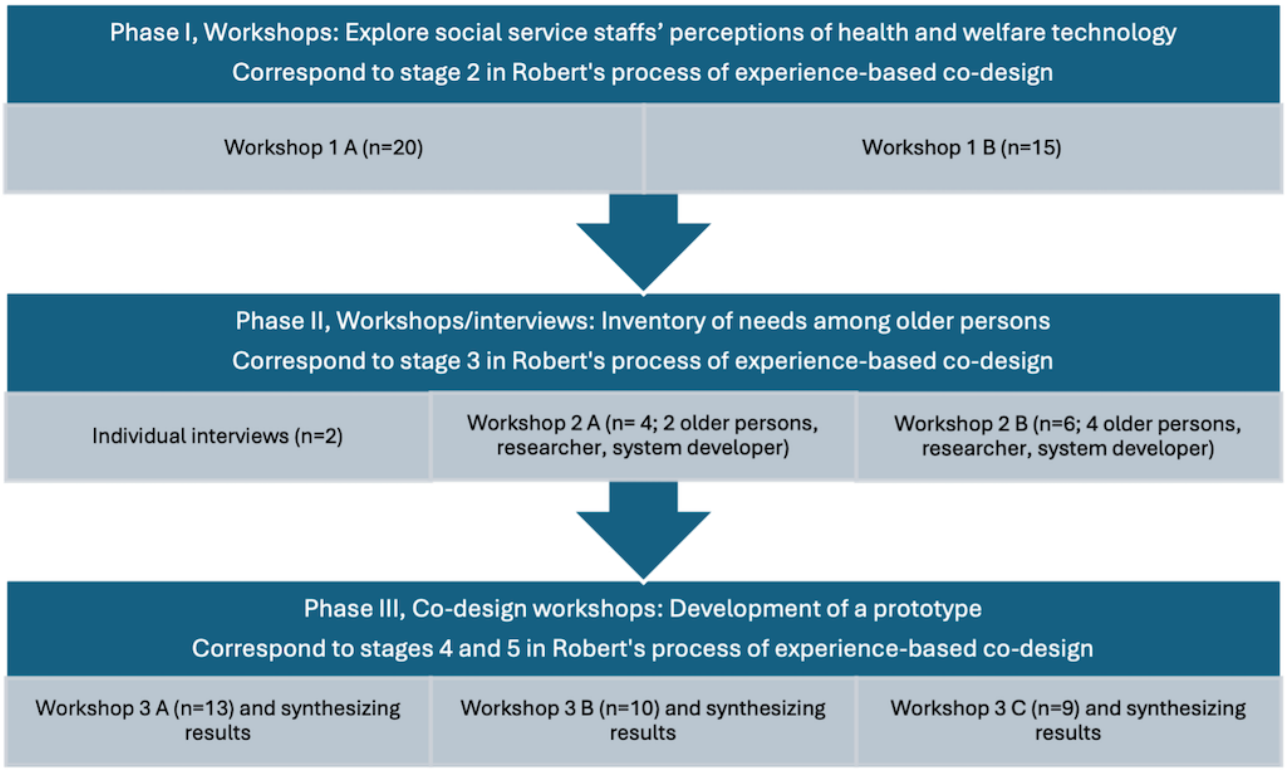

Phase I

The aim of phase I was to explore social service staff's perceptions of how experiences of loneliness and social isolation among older people can be reduced by health and welfare technology. An invitation to participate in 2 WSs was sent to social service organizations in 2 municipalities in the Mideast of Sweden. The invitation was directed to staff in the home care service, care managers, and care unit managers to discuss current research and brainstorm what health and welfare technology is suitable and feasible for use in the municipal care setting. In the invitation, it was expressed that it was desirable that the participants be present on both occasions. Interested people contacted the researchers. Twelve people participated in both WSs 1 A and 1 B. Figure 1 provides the total number of participants. Table 1 shows the distribution of participants according to municipality and profession. The 2 WSs were conducted during the autumn of 2018 by authors 3 and 4.

In WS 1 A, a presentation introduced the participants to the body of research within digital interventions for decreasing loneliness and social isolation among older people through health and welfare technology. The researchers who were leading the WS presented an overview of the published research in the field, including interventions organized into 4 categories: videoconference/telephone, education (eg, in the use of computers and social platforms, eg, Facebook), robots for social interaction, and TV games. Afterwards, the participants were divided into 4 groups that discussed how these digital interventions could be implemented in their organizations and the potential content and use of digital interventions for decreasing experienced loneliness and social isolation among older people. The groups were created to include persons of different sexes, professions, and years of employment. Notes were taken regarding key messages presented at the end of the WS by one of the researchers.

WS 1 B started with a summary of WS 1 A, in which the participants concluded that the intervention of interest was a web-based platform for socializing, with the opportunity for physical meetings, after which the WS continued with discussions in 4 groups. The discussion departed from 4 questions: "What should a web-based platform for decreasing experienced loneliness contain?" "How should the intervention be implemented?" "Which layout and name could be appropriate for the web-based platform?" and "How could the target population be reached?" Additionally, there was time for general discussions on the topic.

The WSs were not audio recorded because some participants opposed recording, as they felt it would inhibit their participation. Data were collated through field notes. The collected data were analyzed via categorization of the notes that were taken during the WSs.

XSL•FO
RenderX

Table 1. Characteristics of participants in the workshops and individual interviews conducted in phases I-III of the study. Participants were social service staff working with older people and contributed with their perceptions of how health and welfare technology could help reduce loneliness and social isolation among older people, older people (68-91 years of age with experience of loneliness or social isolation), system developer, researchers, and an IT pedagogue.

Profession	Municipality X/Y, n/n	Participants, n
Phase I: Workshop 1 A (n=20, female/male: 16/4)		
Care manager/care unit manager	4/6	10
Occupational therapist	0/1	1
Home care staff	3/5	8
Technical coach	1/0	1
Phase I: Workshop 1 B (n=15, female/male: 11/4)		
Care unit manager	2/4	6
Home care staff	3/3	6
Technical coach	1/0	1
Strategist	0/2	2
Phase II: Individual interviews (female/male: 1/1)		
Older people	N/A[a]	2
Phase II: Workshop 2 A (female/male: 1/1)		
Older people	N/A	2
Phase II: Workshop 2 B (female/male: 1/1)		
Older people	N/A	4
Phase III: Workshop 3 A (n=13, female/male: 12/1)		
Older people	N/A	4
Social service staff	N/A	4
IT pedagogue	N/A	1
Researchers	N/A	2
Research administrator	N/A	1
System developer	N/A	1
Phase III: Workshop 3 B (n=10, female/male: 10/0)		
Older people	N/A	2
Social service staff	N/A	3
IT pedagogue	N/A	1
Researchers	N/A	2
Research administrator	N/A	1
System developer	N/A	1
Phase III: Workshop 3 C (n=9, female/male: 9/0)		
Older people	N/A	2
Social service staff	N/A	3
Researchers	N/A	2
Research administrator	N/A	1
System developer	N/A	1

[a]Not applicable.

XSL•FO
RenderX

Phase II

The purpose of phase II was to explore needs related to loneliness and social isolation perceived by older people and how these needs may be targeted by using health and welfare technology. The data were collected during the spring of 2019.

For this phase, 6 older people with experiences of loneliness or social isolation were recruited (Table 1). One of the older persons was recruited through another ongoing project where she had expressed loneliness. Three persons were recruited after the researchers had participated and presented the project in a seminar held by a nongovernmental organization (NGO) for persons in retirement, and 2 persons were recruited through contact persons in the municipalities that were included in phase I. They varied in their experiences and proficiency regarding the use of the internet and technology.

Two of the participants were interviewed individually (individual interviews 1 and 2) by the second author and thereafter together in a WS (WS 2 A), along with one researcher and one system developer. Individual interviews were performed since recruitment took longer than expected and only 2 older persons had agreed to participate at the time data collection commenced. The other 4 participants participated in a WS (WS 2 B) together with the system developer and one researcher. Additionally, a research administrator was present at both WSs. Both WS 2 A and WS 2 B were led by the same experienced WS leader who was not part of the research team. The interview and WS questions can be found in Multimedia Appendix 2. During the WSs, the WS leader created a mind map on a whiteboard based on the ongoing discussions in the WSs on a whiteboard for all the participants to see and reflect upon. This was a visualization that was visible to all participants for reflection and served as field notes. The interviews and WSs were audio-recorded. The recorded interviews, as well as the discussions in WS 2 A and B, were transcribed verbatim. Afterwards, the first author read the texts several times. The texts were analyzed by thematic analysis [39]. Meaning units were marked, and codes were noted in the margin of the text. All codes that were relevant to the aim of this phase were combined, and themes were formed. The results of the thematic analysis were compared with the mind maps to identify themes, functions, or other important aspects that were not captured in the mind maps or if the mind maps contained information not found in the thematic analysis, in order to ensure consistency between the participants' expressed needs in the WSs and in the deeper patterns that emerged from the analysis. The mind maps were thus used as a complementary tool to cross-check and validate the important aspects from the discussion and were reflected in the thematic structure.

Phase III

In the third and last phase, the purpose was to, in co-design with older people, social service staff, system developers, and researchers, develop a prototype of a health and welfare technology intervention for social interaction to reduce experienced loneliness and social isolation among older people. The older people who were included in phase II were invited to participate in the WSs. Invitations were sent to the 2 municipalities that were involved in phases I and II. One person working as an IT pedagogue for older people who were working in a third municipality contacted the research team, expressing interest in the project, and was also invited to participate. The same experienced WS leader and system developer as in phase II participated in the WSs. The data were collected during the autumn of 2019.

A core team of researchers, along with the WS leader, system developer, and social service administrative staff, conducted pre-WSs before the WSs in phase III, wherein the findings from prior phases or WSs were discussed, and suggestions for a prototype and development of the prototype were made. The first pre-WS discussions took departure from the findings in phases I and II. After WS 3 A, the core team continued the development of the prototype based on the discussions and field notes from the previous WSs, and the improved prototype represented a point of departure in WS 3 B. This was further developed after WS 3 B, and this prototype was used as a point of departure in WS 3 C, wherein the discussions were more focused on details and functions. Figure 2 provides an overview of phase III WS.

Table 1 provides an insight for participants in WS 3 A, WS 3 B, and WS 3 C. In WS 3 B, all of the participants had participated in WS 3 A except for one of the older persons. In WS 3 C, all of the participants also participated in WS 3 B.

The data were collected by audio recordings, to which the researchers and system developers could listen again if something was not clear; for example, if a feature was discussed during the WSs that needed clarification. The audio recordings were transcribed verbatim and used for validating the final version of the prototype. Additionally, field notes were taken during both the WSs and the pre-WS by one of the attending researchers.

XSL•FO
RenderX

Figure 2. Overview of phase III of the study: 3 participatory co-design workshops conducted in Västerås, Sweden. Participants included older persons with experience of loneliness or social isolation, social service staff, and IT experts. The workshops focused on collaboratively generating, discussing, and refining ideas for the digital solution Fik@ room. The figure illustrates the stepwise procedure, including ideation, prototype development, and feedback iterations (adapted from Hultman et al [40], which is published under Creative Commons Attribution 4.0 International License [41]).

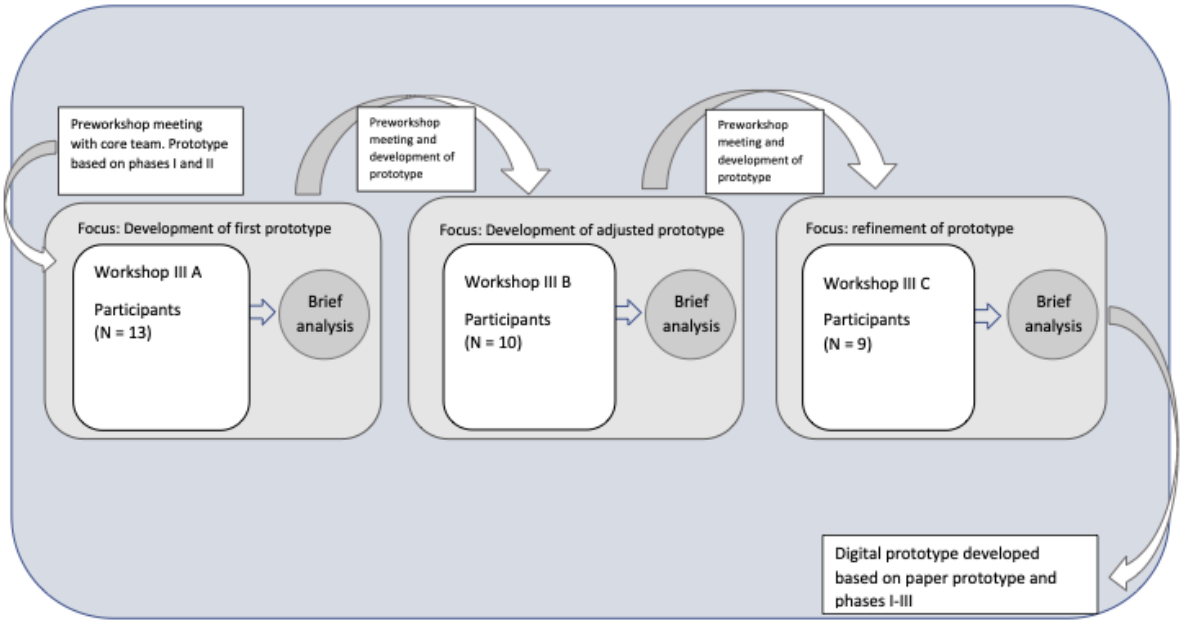

Ethical Considerations

This study followed the research ethics guidelines of the Declaration of Helsinki [42]. An application for ethical review was submitted to the Swedish Ethical Review Authority (Dnr 2019-01870). However, as no sensitive personal data were to be collected, the study design did not fall under the scope of the Act on Ethical Review of Research Involving Humans (SFS 2003:460) [42]. The authority, therefore, granted an exemption from full ethical review and provided an advisory opinion stating that no ethical concerns were identified regarding the study. All participants were provided with information about the study and verbally consented to participate before the data were collected. Older participants were verbally informed by municipal social service staff or the research members at the NGO meetings, as well as in writing through an information letter, and then again verbally through a follow-up telephone call, where informed consent was given. The data were stored to maintain the confidentiality of the participants according to the General Data Protection Regulation. No financial or other compensation was provided to participants for their participation in the study.

Results

The results are presented for each phase, including a short conclusion for each phase.

Phase I

Overview

The results for phase I are organized by first presenting the results from WS 1 A, including a conclusion, followed by the results of the discussions from WS 1 B, which are summarized for each question discussed. The results in phase I are summarized in a conclusion for both WSs combined. Categories identified in WS 1 A and topics discussed in WS 1 B are summarized in Textbox 1. The notes from the presentations from the 4 groups in WS 1 A were categorized into 3 categories.

Textbox 1. Overview of results from phase I: perceptions of social service staff on how loneliness and social isolation among older people can be reduced by health and welfare technology. Data were collected during 2 workshops (WSs).

Categories identified in WS 1 A

- Identification of intended users
- A digital meeting venue
- Technical advancements

Topics discussed in WS 1 B

- What should a digital meeting space for decreasing experienced loneliness contain?
- How could the digital meeting space be implemented?
- Layout aspects and name of the digital meeting space
- How the population could be reached

XSL•FO
RenderX

Identification of Intended Users

Participants expressed a wish for a solution that reached persons with home care services and persons in residential care facilities who experienced loneliness. The intervention was discussed as something complementing (but not replacing) human interaction in real life (IRL). It was also discussed as an intervention for those individuals who do not have the ability or possibility to leave their own homes and meet other people due to various conditions, such as immobility or anxiety. This group included people without home care services, but it was also mentioned that these people may be difficult to reach.

A Digital Meeting Venue

The intervention was discussed as a digital meeting venue that included activities such as bingo, physical exercise, or games. However, a place for digital meetings where information about activities for the intended group could be found, including a digital calendar in which older people can document upcoming events and which events others will attend, would also be beneficial. The digital meeting space may also include social interaction at any point in time (eg, the possibility of eating together via the web). This type of meeting venue was described as more important than care robots. Nevertheless, the digital meeting venue must be simple to use. Video was mentioned as being a medium for meeting health care personnel and relatives, and could be used for attending larger social events, such as weddings.

A central theme in the discussions was that the digital meeting venue would facilitate new acquaintances with whom the older people could also meet IRL. Dating and building new networks were mentioned. The digital meeting venue could connect people with similar interests and enable them to locate old friends or to discuss existential questions and talk about their lives. This scenario could fulfill the longing for belonging, fellowship, and coherence that was described as something that many older people have.

Technical Advancements

It was suggested that a GPS function could be useful for people with dementia, as it could help them leave the house and find their way home. Additionally, alarm systems for falling were discussed as encouraging older people to go outside, which could thereby decrease loneliness and social isolation. Virtual reality (VR) was also discussed as being a factor that could lead to older people having smelling, visual, and hearing experiences that they would otherwise not be able to have. VR could be used for web-based meetings and to visit places. However, VR was also discussed as being very expensive and may lead to economic exclusion. The use of TV screens for planning and controlling daily schedules and as a tool involving speech synthesis to support people with impaired visual and hearing functions. The needs were discussed as differing between men and women.

Conclusions From WS 1 A

All groups discussed some type of digital meeting venue for older people, with the possibility of meeting physically after an initial digital contact, as well as a meeting venue for older people to tell stories and be able to find both new and old friends. This meeting venue could be accessed from home and complements human interaction IRL. A digital meeting venue was the point of departure for WS 1 B.

The Content of a Digital Meeting Venue for Decreasing Experienced Loneliness and Social Isolation

The digital meeting venue could be built with rooms for groups to meet in for activities such as bingo, quizzes, online games, physical exercises, or church masses (both live and recorded). It may also include lectures, opportunities for text-based chats, matchmaking, reading, collaborative cooking, movies, study circles, craft groups or book circles or singing together. It was also suggested that there should be separate rooms available for men. Fixed time for digital meetings and weekly programs was suggested, as was the possibility of seeing who was available online.

It was also discussed to include professional support to help make plans for the future, and a suggestion involved including links for future meeting times to family counseling, legal counseling, and housing agency support. The meeting venue was suggested to include local, national, and international news relevant to the users. In digital meeting venues, people who experience social isolation due to physical restraints may have the opportunity to meet others and participate.

Strategies to Implement a Digital Meeting Venue

Digital literacy was expressed as being crucial among both staff and users for successful implementation. It was further discussed as to whether it should be an individual responsibility to have the necessary equipment to participate in the digital meeting venue or if it should be provided to the intended users. Ownership of the digital meeting venue may be for the municipality or NGO. Links to technical support should be available at all times, where NGOs may be a resource.

The digital meeting venue must be a safe place. Security may be obtained by using a log-in procedure. Young people may experience that using social media affects their mental health, and this scenario should also be considered among older people when implementing digital meeting venues.

Layout Aspects and Name of the Digital Meeting Venue

The staff expressed that the digital meeting venue should have a pleasing and appealing design to arouse curiosity. When considering the intended target population, it should be easy to use and have contrasting colors, few, and large buttons. Speech synthesis was preferred. The digital meeting venue should be able to be used on both tablets and computers, with information available in several languages. It was also discussed that VR glasses and gloves would be interesting options to include. A personal page should also be available.

Several names were discussed to reflect the imagined benefit of the meeting venue: “Atle,” as the ice-breaker boat; “Simul,” as Latin for social interaction; and “the golden middle age,” “it said click,” “mature dating,” “senior spot/app/channel,” as the digital meeting space. The name of this program should not be associated with loneliness and aging. It was also suggested that the intended target group should choose the name of the web-based platform.

XSL•FO
RenderX

Strategies to Reach Intended Users

Several strategies to distribute information were discussed, such as flyers to distribute in primary care (as part of the customary information to all people older than 75 years), advertisements in local newspapers, on NGOs and municipality web pages, on Swedish health care direct (1177), and at gyms and churches where the target group can be contacted.

Conclusion Phase I

The conclusion from WS 1 A and WS 1 B was that the focus for further research should be on a digital meeting venue where older people can meet online for various activities that also include information directed to older people's specific life situations. The program should also be easy and safe to use. The intended users would include older people experiencing loneliness or social isolation, which could be identified by staff within several occupations and fields. The meeting venue should facilitate the discovery of both new and old friends, and the name may indicate this effect (but should be chosen by the target group).

Phase II

Overview

In phase II, the thematic analysis resulted in 6 themes reflecting the expressed needs related to loneliness and social isolation, as well as how these needs may be targeted by using health and welfare technology (Textbox 2). In Figures S1 and S2 in Multimedia Appendix 3, translated mind maps from WS 2 A and B are provided. Quotations from participants are included to illustrate and support the findings.

Textbox 2. Overview of the themes identified in phase II: needs and preferences related to loneliness and social isolation perceived by older people. Data were collected through 2 semistructured interviews and 2 workshops.

Themes

- A gateway for finding new acquaintances
- Provide opportunity for both light-hearted and deeper conversations
- Being part of a community
- Enabling new ways for older people with barriers to interaction
- Available, safe, and varied forms of social interaction at any time
- An easy-to-use and appealing web-based platform

A Gateway for Finding New Acquaintances

A central theme in the participants' descriptions of loneliness was the challenge of finding new acquaintances. This was seen as crucial for reducing feelings of loneliness but increasingly difficult with age. In earlier life stages, especially before retirement, new social connections often emerged naturally, for example, through work. After retirement, such contexts disappeared. Additionally, aging was associated with losing friends and spouses, which led to a shrinking social network and reduced opportunities to meet new people.

> *But then… then, what I'm experiencing now is that several friends… you lose them, they die, they die… and it feels a bit… uh, sad, and it has actually hit me very hard because they are close friends. So in that sense, loneliness strikes. But… but at the same time, I also see opportunities because I'm still curious about life and want to be part of it.* [participant WS 2 B]

Participants expressed that finding new acquaintances required initiative and courage, especially IRL. Activities such as walking with others or meeting people during daily walks were mentioned as potential methods for making contact. However, it was often difficult to find such groups or to approach strangers. In this context, web-based interaction emerged as both a barrier and an opportunity. Digital tools were often used to maintain existing relationships, but seldom for meeting new people. Still, participants saw potential in digital platforms for enabling initial connections that could lead to in-person meetings later.

> *It's like this - most people, if we put it that way, might start with something digital. But in the long run, it's kind of like a dating site - you start there, but eventually, you still want to meet in real life.* [participant individual interview 2]

Participants emphasized that initiating new conversations via the web should feel natural, but few platforms were known to support this. Existing digital communication was typically limited to known contacts, and reaching out to someone new, especially without a clear reason, felt awkward or even inappropriate.

> *If you have the contact, you can maintain it. But getting a new contact, I think that's difficult.* [participant WS 2 A]

The internet was seen as a tool for finding others with similar, or even rare, interests, but this was often difficult both digitally and IRL. The lack of accessible, interest-based meeting spaces contributed to a sense of isolation for those whose hobbies or passions were not widely shared. There was also a clear longing among some participants for deeper friendships and meaningful new connections, even with people who had different backgrounds or interests, or from the opposite sex. This was seen as both desirable and particularly challenging.

> *There has to be something that makes people meet, you know. You need a reason to start talking, maybe*

XSL•FO
RenderX

> *several times. The first time, nothing really happens. It's hard to get to that point.* [participant WS 2 B]

To facilitate new connections, the participants suggested the creation of a digital "gateway"—a platform where people experiencing loneliness could meet regularly via the web. This would allow for repeated interactions and the gradual building of trust, which could eventually lead to offline friendships.

> *So, some kind of catalyst is needed to get these kind of discussions started.* [participant WS 2 B]

Provide Opportunity for Both Light-Hearted and Deeper Conversations

The participants expressed a need for a digital platform that provided opportunities for different types of encounters, such as a meeting place for light-hearted conversations and a place to have fun together to find joy in both fellowship and conversations, but also as a place for deeper conversations about various subjects, such as existential matters. Deeper conversations were expressed as being the one topic that provides meaning to social encounters. The ability to express significant feelings to others when alone was considered very important, not only to be listened to, but also to listen to others.

> *[...] I didn't say a word, and she said, 'You were very nice to talk to,' and I hadn't said anything". "You were listening, and that's what's needed.* [2 participants in WS 2 B]

However, it was also expressed as being important that a digital platform should also provide the opportunity not only to talk about difficult matters but also to find people to laugh with and joke about things together. Some of the participants stated that it took some time to express their true selves and how they could contribute to a conversation, and sometimes it could be easier to write in a chat.

> *That there are those, I'm the kind of person, who wants a slightly deeper conversation. But then there are others who don't want that, and both need to exist, I think.* [participant WS 2 A]

The digital platform should also provide an opportunity to talk about special interests, such as specific subjects that one has deep knowledge about and may have worked with earlier in life.

It was also expressed that they desired to help others in similar situations and wanted to share these experiences with others. It was considered desirable to be both able to give advice and provide support for others, as well as to find support and guidance for oneself. Being important to others was expressed as valuable and could make one feel significant to others.

> *Yes, I mean, [...] I'm used to being a widow and knows how… one is treated and how… how friends react when you are a widow. Then it can be good to be able to guide someone else who has just become a widow, that's how I see it.* [participant WS 2 B]

The digital platform was discussed as being something similar to the breakrooms from former workplaces or, as one participant referred to, a digital smoking room that was described as the place where "everything happened," with a special fellowship for those who attended these locations.

> *Yeah, but that sense of community, the new community, it was very much built in the smoking room, as I said, because that was how it was back then.* [participant WS 2 A]

The participants missed the natural context at work for meetings such as those that occurred in the breakroom. In retirement, the participants described experiencing involuntary exclusion. In the breakroom at work, one could simply enter, sit down at any table, and discuss whatever topic they wished to discuss, sometimes light-hearted, sometimes deeper conversations, or sometimes about specific interests with like-minded individuals.

When in retirement, it was described as more difficult to meet new people for conversations, as one may have previously done in the breakroom. However, no individual enters an empty breakroom; there must be people in the breakroom to talk to. In a web-based setting, it was expressed that it is important that there is a maximum number of the people allowed in this setting, as too many people in a digital format would make it difficult to communicate without disrupting each other. It was also suggested that someone should be responsible for the conversations that occurred. In the breakrooms at work, there were discussions about important things, and agreements on a specific topic were not necessary, which resulted in stimulating conversations. To gather more people in one location could stimulate the initiation of those kinds of conversations.

It was also expressed that those participating in web-based meetings should share similar experiences of loneliness, and that no professionals should be involved.

> *When it comes to loneliness, I think it's better to be in groups where people themselves have experience of what is being discussed, so no professionals.* [participant WS 2 B]

Being Part of a Community

Participants described how retirement often leads to a loss of the natural social interactions that come with working life. Without the daily structure of colleagues, meetings, and casual conversations, many older persons experienced a shift from being active participants in a community to becoming observers from the sidelines. This change was perceived as contributing to feelings of loneliness and exclusion. One participant reflected poignantly on this transition:

> *Well, when you get as old as I am now, you feel there's a kind of boundary, and I notice it with other friends who are the same age. They think seventy is some kind of threshold. It's a bit of both, you're on the side now because you don't work anymore. You visit your children and grandchildren, and then there are other young families nearby, but you are on the side. And that's a bit sad.* [participant WS 2 B]

This experience of being "on the side" highlights the social rupture that retirement can bring, underscoring the need for new ways to maintain a sense of community and meaningful involvement later in life.

XSL•FO
RenderX

The internet was described as a resource for finding information on where to seek help when feeling lonely, as well as for sharing information and links with others. Simply seeing other people, whether online or offline, could help a person experiencing loneliness feel a sense of belonging.

> *It's just like that, in some way, that there are people around you.* [participant individual interview 1]

This scenario could occur during a walk or while being online. For example, it may be appealing to attend a concert alone, but it does not automatically lead to feelings of belonging. Loneliness was described as being alone with no opportunities to share with or talk to others. The participants expressed that different people may have different perceptions of what makes one feel lonely; however, Swedish culture was described as being a factor that contributes to older people feeling lonely, and men were described as being particularly vulnerable.

> *Well, our immigrants, they socialize much, much more within the family and among relatives. So, they are not alone in the same way we are.* [participant WS 2 B]

Neighbors and community members who look out for one another were highlighted as important for people experiencing loneliness. Having no one to talk to for more than a day was described as threatening to one's well-being. However, the encounters needed to counteract loneliness did not have to be significant; even brief conversations, for example, with home care staff, could help alleviate the feeling.

> *But I notice that when I'm alone - well, one day is fine, but after two days or so, I go downhill. So... meeting someone, and it doesn't have to be anything special really, just talking a little - not just about the weather either, but something a little deeper - then you feel better afterwards. At least I do. So, I'm probably more social than I thought. Yes. And when you were working, it was so natural, meeting people, and for me who had students and so on, it could be nice to come home and have it be quiet.* [participant WS 2 A]

The participants also spoke about the importance of connecting with like-minded individuals; people who share similar levels of knowledge and interests.

> *Among like-minded people, they should have roughly the same knowledge. My colleagues had a certain way of understanding things, and that's the kind of discussion you move within.* [participant WS 2 B]

Finally, participants noted the challenges of sharing personal interests that others may not understand.

> *It's hard for me to get people in general to understand my interest. What it is about. Many just say, 'That's not something I'm into.'* [participant Individual interview 2]

Finding like-minded individuals was thus important for feeling part of a community.

Enabling New Ways for Older People With Barriers to Interact

Meeting digitally was described as an important opportunity for older individuals facing barriers that make it difficult to leave their homes. These barriers included a lack of initiative, physical limitations, medical conditions, or being naturally reserved or shy.

> *I can't attend courses because there's air conditioning, and I can't go to the movies, I can't take the bus, train, or fly... So, I become very lonely.* [participant WS 2 B]

Web-based activities, such as playing games or cards, were described as tools for reducing loneliness. These games could be both synchronous (played together in real-time) or asynchronous (played over a longer period), offering flexible ways to stay socially engaged. Although playing alone could provide a temporary distraction from feelings of loneliness, participants emphasized that digital interaction with others was more valuable for their sense of connection.

> *But for example, if you say that many people who are lonely also have difficulty moving and end up sitting alone at home. Then, instead of meeting in person, you could have a conversation in a group like that.* [participant individual interview 1]

Available, Safe, and Varied Forms of Social Interaction at Any Time

Participants described the emotional highs and lows associated with loneliness. When the telephone rang, it could initially spark excitement, a hope that something meaningful was about to happen. However, this feeling often turned to disappointment, for example, when the caller turned out to be a salesman. Such experiences could reinforce feelings of loneliness and lead to further withdrawal from social interactions. One participant shared an experience of reaching out to a friend during a difficult time, only to be rejected. This highlighted a broader need expressed by the participants: to have access to someone to talk to when feelings of loneliness become overwhelming and urgent.

> *Yesterday, I needed someone to talk to, so I wrote to a friend. 'I'm watching [a TV-show], can I get back to you tomorrow morning?' That was the response. By then, it was too late.* [participant individual interview 2]

At the same time, concerns were raised about using the internet for social connections. Participants discussed the risks of being tricked or swindled via web and emphasized the importance of safe, trustworthy digital platforms.

> *I'm bad with Facebook and bank passwords because my account was hacked during my first visit. I became very, very scared. Neither were the police interested, and I didn't know what to do. And I got really scared.* [participant WS 2 B]

A key aspect of feeling secure was being able to see who one is communicating with, such as through video calls. Visual contact not only fostered trust but also made the interaction feel more genuine. However, participants also acknowledged

XSL•FO
RenderX

that speaking could sometimes feel too demanding, and that writing, via e-mail or chat, could be a gentler way to reach out when feeling vulnerable. That said, while writing could offer a temporary outlet, participants noted that chatting alone did not significantly reduce feelings of loneliness, and direct interaction remained crucial for emotional connection.

An Easy-to-Use and Appealing Web-Based Platform

All participants had experience in using digital tools; however, some of them preferred to use the telephone, which was a method that they were more accustomed to. They often used the telephone to decrease feelings of loneliness. Facebook was also used, mostly for reading what others had published. Some of the participants had computers that they did not use and perceived themselves to lack knowledge about how to use them. Others used computers every day. Curiosity and willingness to try new things were described as being important for learning about digital tools. The fact that older people often lack the knowledge to use digital tools, as well as the need for support to use digital tools, was discussed as being important when designing digital tools for older people. The tool itself should be simple to use and easy to obtain, and the layout and included images must fit the intended population. Images were preferred over long texts.

> *Well, as I said, I think it should be easily accessible and clearly laid out.* [participant individual interview 2]

The design should also accommodate the specific needs of older people with disabilities.

> *I have some difficulty writing, I'm shaky, I have an illness, so I find it hard to press the... buttons, quite simply. Does that fit in here somehow?* [participant WS 2 B]

Conclusion Phase II

The analysis of the recorded material, together with the mind maps, formed the basis for the development of the first prototype of a web-based platform with functions that captured the expressed needs of the participants in phase II.

The web-based platform should be designed to enable people to find new acquaintances and act as a gateway to reach people with the same (but also different) interests. The web-based platform should enable both light-hearted and deeper conversations, with the ability to help each other in difficult situations and share experiences. However, the participants did not want any “professionals” to participate on the web-based platform. The web-based platform was described as desired to be similar to the breakroom at work in one of the WSs. One advantage of a web-based platform is that it enables older people, who sometimes have barriers to leaving their apartments to participate, to interact with others, as well as the fact that interactions are available at any time. However, having scheduled time for meetings was suggested to be optimal. It was important that the web-based platform was safe to use and provided different ways to communicate, but also be easy to use and appealing to the intended users. Options of continuing the relationships outside of the web-based platform and to “make appointments” were discussed in both WSs. The mind maps, together with the identified needs from the thematic analysis, were the basis for the first prototype presented in WS 3 A in phase III.

The name of the web-based platform was also determined. We did not decide on “the smoking room” option that was suggested; instead, “the Fik@ room” was chosen as the name for the web-based platform, as “fika” is a Swedish word for having a break (preferably together with others), while having a cup of coffee or tea.

Phase III

Overview

In the first pre-WS meeting, before the first WS in phase III, a digital prototype (prototype 1) was developed based on the analysis of the transcribed interviews and WS recordings from phases I and II, as well as the mind maps from phase II (Figure 3). Figuratively, prototype 1 was built as a house with different rooms where different activities occurred, and was the point of departure for WS 3 A. All of the different boxes in Figure 3 were operationalized as a web page/screen. One example is provided in Figure 4, which shows the living room that was the starting page from which the different functions or rooms could be reached. This user exemplified in Figure 4 had chosen to have “What do you want to do today” and the “Lecture room” as shortcuts from the Living room.

XSL•FO
RenderX

Figure 3. Prototype 1, translated from Swedish, of the web-based platform presented and discussed in workshop 3 A (phase III). The workshop was conducted with older persons with experience of loneliness or social isolation, social service staff, and IT experts. The figure illustrates the first version of an overview of the Fik@ room prototype, including the digital break. Feedback from the participants informed the subsequent development of the prototype.

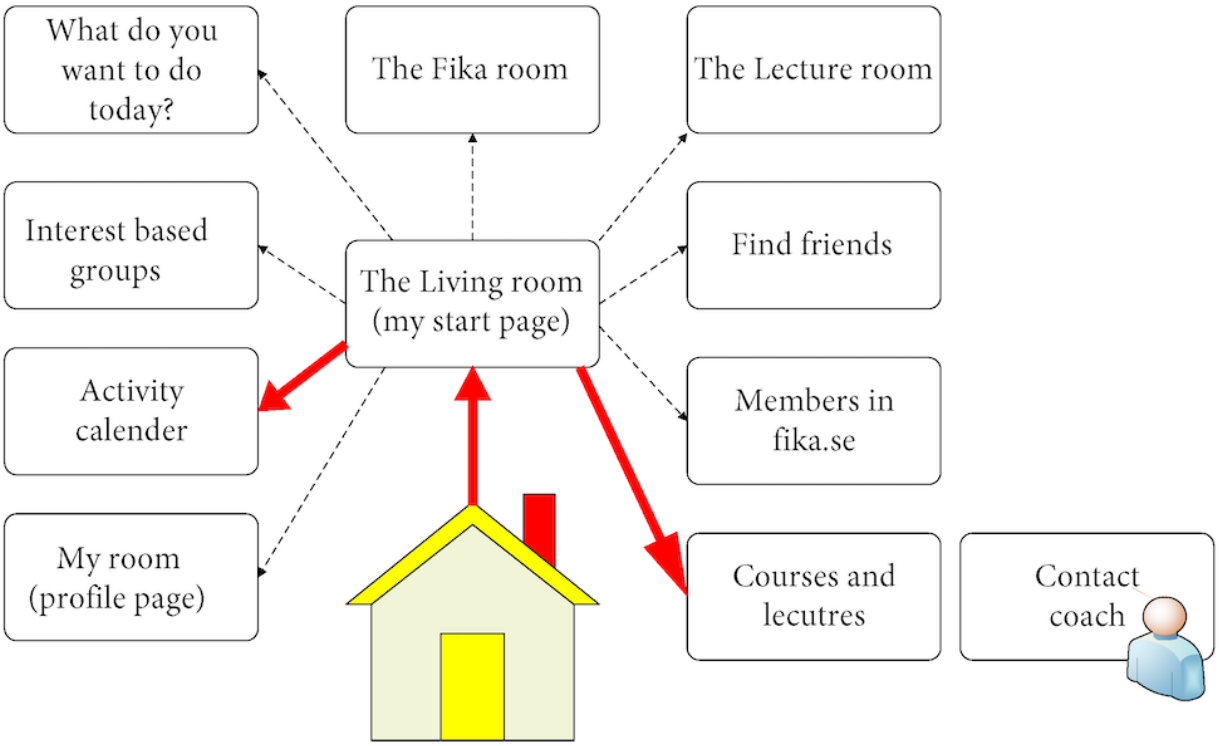

Figure 4. Example screen from prototype 1 of the Fik@ room, showing ”Your Living Room” (Ditt vardagsrum). This screen was presented and discussed during workshop 3 A (phase III) with older persons, social service staff, and IT experts. ”The Living Room” provided links to “What do you want to do today?” (Vad vill du göra idag?), ”the Lecture room” (Lektionssalen) and a link to ”External transportation service” (Färdtjänst). Feedback from participants on the usability and relevance of this design informed later iterations of the prototype design.

XSL•FO
RenderX

Results of WS 3 A and the Core Team Meeting in the Pre-WS Meetings

The topics brought to discussion by the WS 3 A participants can be found in Textbox 3.

Textbox 3. Topics brought to discussion by the workshop 3 A participants.

- Need for a digital coach as part of the web platform.
- There is a need for general support for using web-based platforms.
- Provision of information on what activities are occurring in the community for the target population.
- Distribution of software and hardware for older persons.
- A web-based platform should be easy to use, and all functions and rooms in prototype 1 could be difficult to navigate. The web-based platform should be stable and not change over time in terms of functions.
- Web-based platforms could be connected to other important functions and organizations of society, such as transportation services.
- Some type of calendar to find activities for older persons.
- Security of and access to the web-based platform, and a wish that all persons should be in attendance on equal terms.
- How can we connect with and find new acquaintances, such as based on geographical factors or by interests?
- Fixed appointments in time for joining the digital platform.

In the pre-WS meeting before WS 3 B, the core team members had an impression that prototype 1 may have been too extensive for the participants to grasp, and that the discussion mainly focused on support for using the web-based platform. It was subsequently decided to scale down the functions of the web-based platform, but to retain functions that could still meet the expressed needs. Therefore, Prototype 2 included the following functions: a start page with coffee tables as a central function that supported different preferences for communication (chat, audio, or video conversations); a profile page; a bulletin board; and a support and information function, as the point of departure for WS 3 B. The topics brought to discussion by the WS 3 B participants can be found in Textbox 4.

Textbox 4. Topics brought to discussion by the workshop 3 B participants.

- The support function for technical issues or if someone misbehaves. Ethical principles, such as a code of conduct, could be stated. Instructions could be included in video format. One page for support and instructions was suggested.
- The profile page could have optional choices of photo and text for the description of interests. It should be possible to see who else is online.
- Fixed appointments in time for “fika” could increase the possibility of meeting if a few individuals are online/use the digital platform. However, the web platform should be open 24 hours a day.
- The overview of the coffee tables should include who has joined the tables and what topics are discussed at the tables. It should also be simple to join and leave tables if the topic is not interesting. The individual who starts the table can pick a topic and form for communication (chat, audio, or video conversations).
- The bulletin board should be a place for advertising activities such as meetings on the web-based platform (such as specific topics), in real life, or general notes with tips. Topics need to be relevant, and no old posts should be kept. Contact information for the individual who posted and a link to that person’s profile page should also be provided.
- The suggested activity calendar could include the possibility of inviting others to events, both online on the web-based platform and in real life, as well as seeing what other events will be attended by others. Herein, it could be a dedicated page for local activities from the municipality and nongovernmental organizations to be advertised.
- The design must be simple, and the web-based platform must be easy to use and navigate, as well as not be intimidating. There should not be extensive functions and information.
- Security issues, such as the application of the General Data Protection Regulation and access to the web-based platform, were discussed.
- There was a request for private spaces for deep conversations. These places/spaces could be advertised on the bulletin board. These “rooms” could be used for structured meetings and to follow up on topics such as existential questions with “homework.”
- The possibility of sending messages (both private and more official) was requested.

In the pre-WS before WS 3 C, minor adjustments were made to Prototype 2 with a focus on the functions discussed in WS 3 B. The discussions in WS 3 B were perceived as being more focused on the functions; moreover, due to the fact that it was discussed how the web-based platform should be easy to use and navigate, it was decided to have the coffee tables (the fika tables) as a central function and start page. Additionally, due to the fact that the activity calendar may be difficult to use and expensive to develop, it was not included in Prototype 3. Prototype 3 included profile pages, a bulletin board, a support function, an information page, and a page with links to important societal functions. During WS 3 C, a discussion of whether some functions were lost and whether some functions would be tested, such as how to write messages on the bulletin board

XSL•FO
RenderX

and how to conduct a video conversation, should be performed. The topics brought to discussion by the WS 3 C participants can be found in Textbox 5.

Textbox 5. Topics brought to discussion by the workshop 3 C participants.

- An overview of online users. Additionally, the possibility of having a "friend-list" and seeing who is online.
- Profile picture should be used to identify who sits at the tables.
- A reasonable number of participants should be 4 individuals at one table.
- The design must support the use of aids for those with loss of function, such as hearing or vision impairments.
- If many people are online, how can we group them; specifically, should there be one fika room for each municipality? Will this be a national app or a regional app?
- Fixed times for fika sessions each day could help people to log in at the same time. Many people must be logged in for the web platform to work.
- One table is ready to use when logging in to make it easy to start.
- A table overview was suggested, in order for individuals to see who is at the other tables.
- Empty tables should be removed.
- The participants were positive about the video meeting test.
- A suggestion to be able to share screens in meetings was expressed.
- Rooms/tables for private conversations were suggested. Additionally, it should be possible to send private messages.
- The need for technical support.
- The possibility of inviting other people to use the platform.
- The advertisements on the bulletin board must be up to date. There should also be a possibility to answer advertisements and see who has answered them. Advertisements should be sorted by topics or how to meet, digital or in real life.
- Pages with links that are important for an older person experiencing loneliness.

Conclusion Phase III

The 3 phases resulted in the development of a paper prototype of a web-based platform that was named the Fik@ room, aiming at reducing experienced loneliness and social isolation. The main function of the Fik@ room involves coffee tables where people can interact similarly to how breakrooms functioned at work before retirement. Participants had discussed in phase II examples of separate groups, such as for men or widows. This was interpreted as an indication of the importance of enabling users to choose with whom they interact. In the prototype, digital coffee tables can therefore be arranged for specific groups (eg, "men only" or other interest-based discussions) if users wish, and such tables can also be announced on the bulletin board or labeled with the theme of the discussion.

One important finding was that the web-based platform must be easy to use to not exclude older people with lower digital literacy. Therefore, a prioritization among the functions from Prototype 1 was developed for Prototype 2. The needs expressed in phase II were still perceived as being met (to a large extent) in the final prototype. Except for the coffee tables, the main functions in Prototype 2 were the bulletin board, the profile page, a support and information page, and a page for relevant links for the target group.

The need to meet new people, as well as to identify old acquaintances, has both lighthearted and deeper conversations, and find people for activities IRL, in addition to the other needs that were expressed, could be fulfilled by the Fik@ room.

In phase I, one finding was that the web-based platform could include different activities and professionals. In phase II, activities were also mentioned, both of which can occur on the platform and IRL. The more prominent need to keep the platform simple, easy, and intuitive to use was nevertheless prioritized. Herein, the bulletin board may play an important role, and instructions will be developed regarding how the bulletin board may be used for arranging meetings both inside and outside of the platform. The log-in procedures, as well as who has access to the Fik@ room, will enable a secure and safe platform.

In Multimedia Appendix 4, a URL to one of the developed instruction videos regarding how to navigate in the Fik@ room is included. This video was also included in the final design of the application and used during the feasibility and pilot studies. It shows the final visual interface of the Fik@ room [43].

Discussion

Principal Results

This study describes the iterative design process that resulted in the application of the Fik@ room for older persons experiencing loneliness or social isolation through participatory design procedures with different key stakeholders. The main result of the study is the production of the paper prototype version of the Fik@ room, supporting social interaction among people experiencing loneliness or social isolation. The name of the application (the Fik@ room) is a direct translation of the Swedish suggestion "Fikarummet," which was proposed by participants in the WSs in 2019. The term "fika" is the Swedish

XSL•FO
RenderX

word for having coffee/tea, taking a break, and being socially interactive, which represents a common phenomenon in Swedish workplaces. The letter "a" was replaced by @ as a symbol for the digital form of this fika room. The participants in the WSs drew a parallel to the coffee/break room at their former working place, where they would sit down at a table and were socially interactive by having conversations with their colleagues. They missed this fika break from their working life and wanted the web-based platform to fill this gap. The results from the different phases contributed to fulfilling this need in the Fik@ room. The developed prototype was designed to primarily target social rather than emotional loneliness [44] as the latter is challenging to address through digital solutions only. The platform is designed to offer opportunities for initial contact and social interaction, allowing users to meet new people and, if they wish, continue interacting with the same individuals over time. As expressed by the participants, such a platform may serve as a starting point for developing relationships that could later transition to face-to-face meetings or other communication channels. The platform should therefore be regarded as a complement rather than a substitute for in-person social interaction.

Comparison With Prior Work

General successful interventions for decreasing loneliness and social isolation among older people include one-to-one or group interventions, as well as community activities, that are performed as both face-to-face and technical interventions [27]. A recent umbrella review showed that ICT interventions can be effective at increasing social support, increasing contact with others, and increasing self-confidence [26]. Although many of the successful interventions for social connectedness that are presented in the reviews are based on communication with family and friends, the effectiveness of developing new friendships needs to be focused on in future studies, such as this study. The use of the Fik@ room does not require an established network of contacts and can also be used at any time, as long as others are logged into the app. Of course, this implies that there must be a critical minimum number of users of the Fik@ room, which should be considered when implementing the application of the Fik@ room. Otherwise, an empty Fik@ room may contribute to feelings of loneliness and social isolation.

The Fik@ room was evaluated in a feasibility study [45] and a pilot study [46], and the implementation of the Fik@ room was evaluated from an organizational perspective [47]. The web-based platform has been shown to provide older persons with new contacts and reduce experienced loneliness, thus serving as a complement to other relationships and activities [45,46]. It has been shown to be feasible to develop new relationships and thereby be suitable for those who lack family and friends. Moreover, over time, relationships tended to undergo a transformation into deepened conversations among friends [45]. However, it may not be suitable for all individuals. Not all individuals appreciate meeting and getting to know new people, especially not in a web-based setting. A prerequisite for using the web-based platform involved motivation and a perceived need to meet new acquaintances via the web [46]. In earlier studies, it has been shown that web-based social activities have the potential to improve older people's social contacts and increase participation in activities, which can reduce their experiences of loneliness [48].

The web-based platform was designed to be easy to use, which has been confirmed [46] and may have been achieved through the involvement of key stakeholders [30]. However, some of the older persons considered digital technology devices to be too complicated to use due to joint pain in their hands or because touchscreens do not respond to older people's dry fingertips. Hence, the use of a web-based platform on a personal computer instead of a tablet should be optional. The development of new technology could preferably include a dictation function as an option for writing, and this was planned to be used in the chat function in the Fik@ room. Moreover, as we experienced, a lack of digital confidence and skills may be a barrier [47]. Providing continuous support with login advice and handling the different features of the web-based platform could overcome these barriers.

Regardless of how good the product is, it may be difficult to reach older people in need of social interaction. According to the abovementioned studies, care professionals and social services personnel may be the appropriate people for recruiting and introducing the intended target group [45,46], thereby increasing the chances of reaching potential users.

The Fik@ room was developed in co-design, with the notion that by involving older persons and the staff who meet these persons every day in the design, the web-based platform would be tailored to the end users' needs, thereby being easier to implement. Experts within the field, such as researchers and system developers, were also involved in contributing to broadening the perspective to find new solutions. The implementation of interventions is complex and is, therefore, important to consider during the development process. In addition to recruitment and being able to reach the intended target group, personnel must be able to introduce the web-based platform [47]. This indicates that they will need knowledge about the web-based platform and be able to show it to the intended user. If not perceived as being a useful tool by personnel, they will be reluctant to introduce it to older people. To feel comfortable introducing digital products, management will have to increase time and costs to improve digital skills among personnel, thus allowing for time to use and become comfortable with the product. Moreover, there is a need for IT support. It has been shown that implementation involves an entire organization (to varying degrees) and that all levels need to be involved to create the necessary conditions for successful implementation. Future work will also need to address implementation strategies for reaching a critical minimum number of users to ensure meaningful engagement. This may include partnerships with NGOs, community organizations, and municipal services to facilitate user onboarding. In the feasibility and pilot study, a suggested time for "fika" was communicated to the users in order to ensure a critical minimum of users [45,46].

Strengths and Limitations

The strengths of this study included the involvement of several different key stakeholders in the development of the Fik@ room. To better understand users' needs, creating an intervention that

XSL•FO
RenderX

is engaging and easy to use is suggested to have the involvement of key stakeholders in development, and should be a standard procedure [30]. The valuable input from different sources may have strongly contributed to the successful implementation of the Fik@ room presented in other studies [46], and the issues during implementation were not associated with the use of the Fik@ room but with other aspects [47]. The co-design included key stakeholders and experts from several fields and in different phases of development, thus contributing to a broad base of knowledge in the development process.

A potential limitation of our approach is that the process started with WSs with social service staff, which may have implicitly framed the project and the input from the older persons in phase II. Even though phases I and II were conducted and presented in chronological order, they should be regarded as parallel co-design events rather than sequential phases building on one another, to put the needs of older people first and in line with Robert's co-design model [37]. To minimize this influence, the interview guide used in phase II was designed to explore older people's experience and preferences broadly, rather than being based on the discussion in phase I. In addition, different members of the research team collected data in these 2 phases, which further minimized the risk of bias. Nevertheless, the focus on a digital solution was predefined in this study, which may have hindered the participant in all phases from expressing needs for a nondigital solution. It is also possible that the older persons who consented to participate were those who were already positively inclined toward digital solutions.

At some points, we obtained conflicting suggestions in the phases and from different participants. Recommendations suggest that co-design studies be conducted in small, frequent consultations with stakeholders to understand their preferences [49]; by the time of the last WSs, it was perceived that all of the participants agreed on the final prototype. One further limitation of the study was the small sample size and representation within the different groups of key stakeholders. There was only one man who participated (he represented the intended target group), and there were also small variations regarding cultural background among the participants. The small sample size and the homogeneity of the sample may have contributed to the final design not being sufficiently designed for other groups that were not included in this study. Recruitment for this study was challenging, and participation of older persons was limited to 1 man and 4 women, all ethnically Swedish and born in Sweden. This homogeneity restricts the transferability of our findings, particularly regarding potential cultural differences in experiences of loneliness and social isolation. Future studies and iterations of the Fik@ room should strive to include underrepresented groups, such as more men and individuals from immigrant communities, to better capture diverse perspectives and address potential cultural barriers. Regarding sample size, there is no consensus regarding the group size on co-design development studies [30], and the studies evaluating the Fik@ room [45,46] did not demonstrate that it was not gender or culturally inappropriate. Another limitation was that the participants in phase I opposed being recorded, which may have influenced the depth of the data analysis in phase I.

Conclusions

The study process led to the development of the prototype of the Fik@ room, which is a digital venue for older people experiencing loneliness or social isolation to meet and interact with others with the same experiences. The features of the Fik@ room are based on the needs expressed by social service staff with professional experience working with this group and older people with first-hand experience of loneliness or social isolation. The central aspects that were included in the development and the final digital prototype were an easy-to-use application (meaning a program including a few functions and intuitive use). The layout and colors should focus on being appealing to an older population. The main features include a digital fika room (break room) with tables where the user of the Fik@ room can "sit down" for a chat, an advertisement board for announcing meetings in or outside the Fik@ room, and a profile page for each user. Only persons experiencing loneliness and social isolation should have access to the Fik@ room. The development of the Fik@ room led to an application that was perceived as being easy to use and that provided the opportunity for new contacts, thus implying that the development was successful when implemented with support for users.

Acknowledgments

The authors would like to acknowledge the members of the research group PrILiv at Mälardalen University, Sweden, for thoughtful comments on the manuscript, the municipalities and staff that contributed to the development of the WSs, and the older persons with experiences of loneliness or social isolation for generously sharing their needs and wishes. The authors also want to thank Lena Bernegård and Lotten Blomberg for their valuable work with leading WSs and developing the first prototypes of the Fik@ room. The study was funded by the Mälardalen Competence Centre for Health and Welfare (part of the Social Contract), Mälardalen University, and the IT company Evry. The funder was not involved in the study design, data collection, analysis, or the writing of the manuscript. Generative artificial intelligence (AI) tools (Microsoft CoPilot) were used to assist with the translation of citations and to refine the language of smaller portions of the text that were added after professional language editing. All AI-assisted outputs were carefully reviewed by the authors to ensure accuracy and to preserve the original meaning of the citations. No new scientific content was generated de novo by AI.

Data Availability

The datasets generated and analyzed during this study are not publicly available because data sharing was not included in the agreement with participants. Datasets are available from the corresponding author upon reasonable request.

XSL•FO
RenderX

Authors' Contributions

All authors were employed by Mälardalen University. CE contributed to the conception and design of the work for phases 2 and 3, analyzed the data from phases 1 to 3, collected data for phases 2 and 3, and served as the main author of the manuscript. VZ contributed to the conception and design of the work for phases 1 to 3, collected data for phases 2 and 3, analyzed data across all 3 phases, and drafted the manuscript. AKG was involved in the conception and design of the work for phases 1 to 3, collected and analyzed data for phase 1, and critically revised the manuscript with substantial intellectual input. CÅ contributed to the conception and design of the work for phase 1, collected and analyzed data for that phase, and critically revised the manuscript, adding intellectual content. PHW, as the project leader, contributed to the conception and design of the work for phases 1 to 3, analyzed data across all 3 phases, and drafted the manuscript. All of the authors have agreed to be accountable for all aspects of the work.

Conflicts of Interest

CE, VZ, and PHW have equity in the company Avatara AB, which now owns the Fik@ room and is a company that may benefit from this study's findings. The company was founded after the data collection of this study and aims to make the Fik@ room available for its target audience.

Multimedia Appendix 1
COREQ Checklist.
[PDF File (Adobe PDF File), 71 KB - formative_v9i1e78213_app1.pdf]

Multimedia Appendix 2
Interview and workshop gudie phase II.
[PDF File (Adobe PDF File), 45 KB - formative_v9i1e78213_app2.pdf]

Multimedia Appendix 3
Mind maps developed during WS in phase II (Figure S1 and Figure S2).
[PDF File (Adobe PDF File), 1961 KB - formative_v9i1e78213_app3.pdf]

Multimedia Appendix 4
URL to one of the developed instruction videos regarding how to navigate in the Fik@ room. This video was also included in the final design of the application and used during the feasibility and pilot studies and shows the final visual interface of the Fik@ room.
[PDF File (Adobe PDF File), 38 KB - formative_v9i1e78213_app4.pdf]

References

1. Global strategy and action plan on ageing and health. World Health Organization (WHO). Geneva: World Health Organization; 2017. URL: https://www.who.int/publications/i/item/9789241513500 [accessed 2025-10-21]
2. World report on ageing and health. World Health Organization (WHO). Geneva: World Health Organization; 2015. URL: https://apps.who.int/iris/handle/10665/186463 [accessed 2025-10-21]
3. Francis J, Rikard RV, Cotten SR, Kadylak T. Does ICT use matter? how information and communication technology use affects perceived mattering among a predominantly female sample of older adults residing in retirement communities. Inf Commun Soc 2017;22(9):1281-1294. [doi: 10.1080/1369118x.2017.1417459]
4. Holwerda TJ, Beekman ATF, Deeg DJH, Stek ML, van Tilburg TG, Visser PJ, et al. Increased risk of mortality associated with social isolation in older men: only when feeling lonely? results from the amsterdam study of the elderly (AMSTEL). Psychol Med 2012;42(4):843-853. [doi: 10.1017/S0033291711001772] [Medline: 21896239]
5. Khosravi P, Rezvani A, Wiewiora A. The impact of technology on older adults' social isolation. Comput Hum Behav 2016;63:594-603. [doi: 10.1016/j.chb.2016.05.092]
6. Perlman D, Peplau LA. Loneliness. In: Friedman HS, editor. Encyclopedia of Mental Health. San Diego, CA: Academic Press; 1998:571-581.
7. Nicholson NR. Social isolation in older adults: an evolutionary concept analysis. J Adv Nurs 2009;65(6):1342-1352. [doi: 10.1111/j.1365-2648.2008.04959.x] [Medline: 19291185]
8. Dickens AP, Richards SH, Greaves CJ, Campbell JL. Interventions targeting social isolation in older people: a systematic review. BMC Public Health 2011;11:647 [FREE Full text] [doi: 10.1186/1471-2458-11-647] [Medline: 21843337]
9. Cornwell EY, Waite LJ. Social disconnectedness, perceived isolation, and health among older adults. J Health Soc Behav 2009;50(1):31-48 [FREE Full text] [doi: 10.1177/002214650905000103] [Medline: 19413133]
10. Victor C, Scambler S, Bond J, Bowling A. Being alone in later life: loneliness, social isolation and living alone. Rev Clin Gerontol 2001;10(4):407-417. [doi: 10.1017/s0959259800104101]

XSL•FO
RenderX

11. Holmén K, Ericsson K, Winblad B. Loneliness and living conditions of the oldest old. Scand J Soc Med 1994;22(1):15-19. [doi: 10.1177/140349489402200103] [Medline: 8029661]
12. Aanes MM, Hetland J, Pallesen S, Mittelmark MB. Does loneliness mediate the stress-sleep quality relation? the hordaland health study. Int Psychogeriatr 2011;23(6):994-1002 [FREE Full text] [doi: 10.1017/S1041610211000111] [Medline: 21338549]
13. Cacioppo JT, Hawkley LC, Crawford LE, Ernst JM, Burleson MH, Kowalewski RB, et al. Loneliness and health: potential mechanisms. Psychosom Med 2002;64(3):407-417. [doi: 10.1097/00006842-200205000-00005] [Medline: 12021415]
14. Momtaz YA, Hamid TA, Yusoff S, Ibrahim R, Chai ST, Yahaya N, et al. Loneliness as a risk factor for hypertension in later life. J Aging Health 2012;24(4):696-710. [doi: 10.1177/0898264311431305] [Medline: 22422758]
15. Ramic E, Pranjic N, Batic-Mujanovic O, Karic E, Alibasic E, Alic A. The effect of loneliness on malnutrition in elderly population. Med Arh 2011;65(2):92-95. [Medline: 21585182]
16. Tabue Teguo M, Simo-Tabue N, Stoykova R, Meillon C, Cogne M, Amiéva H, et al. Feelings of loneliness and living alone as predictors of mortality in the elderly: the PAQUID study. Psychosom Med 2016;78(8):904-909. [doi: 10.1097/psy.0000000000000386]
17. Hawkley LC, Cacioppo JT. Loneliness matters: a theoretical and empirical review of consequences and mechanisms. Ann Behav Med 2010;40(2):218-227 [FREE Full text] [doi: 10.1007/s12160-010-9210-8] [Medline: 20652462]
18. Cacioppo S, Capitanio JP, Cacioppo JT. Toward a neurology of loneliness. Psychol Bull 2014;140(6):1464-1504 [FREE Full text] [doi: 10.1037/a0037618] [Medline: 25222636]
19. Alpert PT. Self-perception of social isolation and loneliness in older adults. Home Health Care Manag Pract 2017;29(4):249-252. [doi: 10.1177/1084822317728265]
20. Savikko N, Routasalo P, Tilvis RS, Strandberg TE, Pitkälä KH. Predictors and subjective causes of loneliness in an aged population. Arch Gerontol Geriatr 2005;41(3):223-233. [doi: 10.1016/j.archger.2005.03.002] [Medline: 15908025]
21. Yuan S, Hussain SA, Hales KD, Cotten SR. What do they like? communication preferences and patterns of older adults in the United States: the role of technology. Educ Gerontol 2015;42(3):163-174. [doi: 10.1080/03601277.2015.1083392]
22. Wilson G, Gates JR, Vijaykumar S, Morgan DJ. Understanding older adults' use of social technology and the factors influencing use. Ageing Soc 2021;43(1):222-245 [FREE Full text] [doi: 10.1017/s0144686x21000490]
23. Seifert A, Cotten SR, Xie B. A double burden of exclusion? digital and social exclusion of older adults in times of COVID-19. J Gerontol B Psychol Sci Soc Sci 2021;76(3):e99-e103 [FREE Full text] [doi: 10.1093/geronb/gbaa098] [Medline: 32672332]
24. Cotten SR, Anderson WA, McCullough BM. Impact of internet use on loneliness and contact with others among older adults: cross-sectional analysis. J Med Internet Res 2013;15(2):e39 [FREE Full text] [doi: 10.2196/jmir.2306] [Medline: 23448864]
25. Chen YR, Schulz PJ. The effect of information communication technology interventions on reducing social isolation in the elderly: a systematic review. J Med Internet Res 2016;18(1):e18 [FREE Full text] [doi: 10.2196/jmir.4596] [Medline: 26822073]
26. Balki E, Hayes N, Holland C. Effectiveness of technology interventions in addressing social isolation, connectedness, and loneliness in older adults: systematic umbrella review. JMIR Aging 2022;5(4):e40125 [FREE Full text] [doi: 10.2196/40125] [Medline: 36279155]
27. Stojanovic J, Collamati A, Mariusz D, Onder G, La Milia DI, Ricciardi W, et al. Decreasing loneliness and social isolation among the older people: systematic search and narrative review. Epidemiol Biostat Public Health 2022;14(2). [doi: 10.2427/12408]
28. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view. MIS Q 2003;27(3):425. [doi: 10.2307/30036540]
29. Chen K, Chan AHS. A review of technology acceptance by older adults. Gerontechnology 2011;10(1):1-12. [doi: 10.4017/gt.2011.10.01.006.00]
30. Slattery P, Saeri AK, Bragge P. Research co-design in health: a rapid overview of reviews. Health Res Policy Syst 2020;18(1):17 [FREE Full text] [doi: 10.1186/s12961-020-0528-9] [Medline: 32046728]
31. Revenäs Å, Hvitfeldt Forsberg H, Granström E, Wannheden C. Co-designing an eHealth service for the co-care of parkinson disease: explorative study of values and challenges. JMIR Res Protoc 2018;7(10):e11278 [FREE Full text] [doi: 10.2196/11278] [Medline: 30377143]
32. Sumner J, Chong LS, Bundele A, Wei Lim Y. Co-designing technology for aging in place: a systematic review. Gerontologist 2021;61(7):e395-e409 [FREE Full text] [doi: 10.1093/geront/gnaa064] [Medline: 32506136]
33. Boote J, Baird W, Sutton A. Public involvement in the systematic review process in health and social care: a narrative review of case examples. Health Policy 2011;102(2-3):105-116. [doi: 10.1016/j.healthpol.2011.05.002] [Medline: 21641075]
34. Cook EJ, Randhawa G, Sharp C, Ali N, Guppy A, Barton G, et al. Exploring the factors that influence the decision to adopt and engage with an integrated assistive telehealth and telecare service in Cambridgeshire, UK: a nested qualitative study of patient 'users' and 'non-users'. BMC Health Serv Res 2016;16:137 [FREE Full text] [doi: 10.1186/s12913-016-1379-5] [Medline: 27095102]
35. Gusdal AK, Johansson-Pajala RM, Zander V, von Heideken Wågert P. Prerequisites for a healthy and independent life among older people: a Delphi study. Ageing Soc 2020;41(9):2171-2187. [doi: 10.1017/s0144686x20000306]

XSL•FO
RenderX

36. Andersson SW, Richardson MX, Cozza M, Lindén M, Redekop K. Addressing evidence in health and welfare technology interventions from different perspectives. Health Policy Technol 2021;10(2):100519. [doi: 10.1016/j.hlpt.2021.100519]
37. Robert G. Participatory action research: using experience-based co-design to improve the quality of healthcare services. In: Ziebland S, Coulter A, Calabrese J, Locock L, editors. Understanding and Using health Experiences: Improving Patient Care. United Kingdom: Oxford University Press; 2013.
38. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
39. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2008;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
40. Hultman L, Eklund C, von Heideken Wågert P, Söderlund A, Lindén M, Elfström ML. Development of an eHealth intervention including self-management for reducing sedentary time in the transition to retirement: participatory design study. JMIR Form Res 2025;9:e63567 [FREE Full text] [doi: 10.2196/63567] [Medline: 39832361]
41. Attribution 4.0 International (CC BY 4.0). Creative Commons. URL: https://creativecommons.org/licenses/by/4.0/ [accessed 2025-11-07]]
42. World Medical Association. World medical association declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA 2013;310(20):2191-2194. [doi: 10.1001/jama.2013.281053] [Medline: 24141714]
43. Instruction video: how to navigate in the Fik@ room. Mälardalen University. URL: https://play.mdu.se/media/t/0_qtvfmgib [accessed 2025-10-07]
44. Yanguas J, Pinazo-Henandis S, Tarazona-Santabalbina FJ. The complexity of loneliness. Acta Biomed 2018;89(2):302-314 [FREE Full text] [doi: 10.23750/abm.v89i2.7404] [Medline: 29957768]
45. Johansson-Pajala RM, Gusdal A, Eklund C, Florin U, von Heideken Wågert P. A codesigned web platform for reducing social isolation and loneliness in older people: a feasibility study. Inform Health Soc Care 2023;48(2):109-124 [FREE Full text] [doi: 10.1080/17538157.2022.2070068] [Medline: 35579407]
46. Gusdal AK, Florin U, Johansson-Pajala RM, Eklund C, Fritz J, von Heideken Wågert P. Older adults' use of a research-based web platform for social interaction. Healthcare 2023;11(3):408 [FREE Full text] [doi: 10.3390/healthcare11030408] [Medline: 36766983]
47. Fritz J, von Heideken Wågert P, Gusdal AK, Johansson-Pajala R, Eklund C. Determinants of implementing an information and communication technology tool for social interaction among older people: qualitative content analysis of social services personnel perspectives. JMIR Aging 2024;7:e43999 [FREE Full text] [doi: 10.2196/43999] [Medline: 38407955]
48. Larsson E, Padyab M, Larsson-Lund M, Nilsson I. Effects of a social internet-based intervention programme for older adults: an explorative randomised crossover study. Br J Occup Ther 2016;79(10):629-636. [doi: 10.1177/0308022616641701]
49. Thabrew H, Fleming T, Hetrick S, Merry S. Co-design of eHealth interventions with children and young people. Front Psychiatry 2018;9:481 [FREE Full text] [doi: 10.3389/fpsyt.2018.00481] [Medline: 30405450]

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Research
IRL: in real life
ICT: information and communication technology
NGO: nongovernmental organization
VR: virtual reality
WS: workshop

Edited by A Stone, A Mavragani; submitted 02.06.25; peer-reviewed by J Janssen, X Zheng; comments to author 24.07.25; revised version received 29.09.25; accepted 13.10.25; published 21.11.25.

Please cite as:
Eklund C, Zander V, Gusdal AK, Åkerlind C, von Heideken Wågert P
Co-Designing a Digital Solution for Decreasing Loneliness and Social Isolation Among Older People in Sweden: Explorative Study
JMIR Form Res 2025;9:e78213
URL: https://formative.jmir.org/2025/1/e78213
doi:10.2196/78213
PMID:

©Caroline Eklund, Viktoria Zander, Annelie K Gusdal, Charlotta Åkerlind, Petra von Heideken Wågert. Originally published in JMIR Formative Research (https://formative.jmir.org), 21.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use,

XSL•FO
RenderX

distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Comparing Patient Simulation With a Humanoid Robot or a Human Actor in Terms of Training Success and Acceptance: Pilot Questionnaire Study

Patricia Schwarz[1], MEng; Sandra Hellmers[1], Dr ing; Sebastian Spanknebel[2], Dr phil; Diana Immel[2], MSc; Rene Hurlemann[2], Dr Dr med; Andreas Hein[1], Dr ing

[1]Assistance Systems and Medical Device Technology, Department of Health Services Research, Carl von Ossietzky Universität Oldenburg, Oldenburg, Germany

[2]School of Medicine and Health Sciences, Department of Psychiatry and Psychotherapy, Carl von Ossietzky Universität Oldenburg, Oldenburg, Germany

Corresponding Author:
Patricia Schwarz, MEng
Assistance Systems and Medical Device Technology
Department of Health Services Research
Carl von Ossietzky Universität Oldenburg
Ammerländer Heerstr. 114-118
Oldenburg, 26129
Germany
Phone: 49 04419722133
Email: patricia.schwarz@uol.de

Abstract

Background: The addition of simulated patients to medical and nursing training makes it possible to create a link between theory and practice. This makes what has been learned more realistic and allows the complexity and multilayered nature of many illnesses to be reflected in a real-life setting. However, the selection, training, and supervision of actors as simulated patients is time consuming and expensive. In this study, we investigated how differently students and nurses perceive 2 different methods of patient simulation.

Objective: The aim of this pilot study was to investigate whether patient behavior simulated by a humanoid robot is comparable to patient simulation by actors in videos in terms of training success and user acceptance. Participants were asked to recognize the symptoms presented by the humanoid robot and make a diagnosis. For comparison purposes, we asked a second group of participants to make a diagnosis based on a video featuring a human patient actor.

Methods: We asked the participants (medical students and nursing staff; N=21) to conduct a psychopathological assessment. Group 1 (n=11) used the humanoid robot as a patient simulator, and group 2 (n=10) watched the identical symptoms in a video with a human actor as patient.

Results: The participants had a mean age of 28.7 (SD 3.5) years. The students were in their sixth semester and had, on average, 7.6 (SD 3.3) years of professional experience in the medical field. The correct diagnosis was made 90% (9/10) of the time based on the video with the human patient actor and 91% (10/11) of the time based on the robot. One participant in each group made the wrong diagnosis, constituting a total error rate of 10% (2/21). In general, participants with the humanoid robot as patient simulator felt more confident that their diagnosis was correct compared to those with the human actor as patient (humanoid robot: 9/11, 82% were neutral to very confident and 2/11, 18% were uncertain to very uncertain; human actor: 9/10, 90% were neutral to very confident and 1/10, 10% were uncertain or unsure).

Conclusions: The simulations of the human actor in the video were judged to be more realistic overall than those of the humanoid robot as patient simulator. However, the differences between the simulation methods in relation to the result (diagnosis) were very small. The results of our pilot study show a good performance of the robot in the simulation of selected psychiatric patient cases. We conclude that a humanoid robot could be a useful addition to patient simulators in medical education and discuss future directions.

(JMIR Form Res 2025;9:e70363) doi:10.2196/70363

XSL•FO
RenderX

KEYWORDS

humanoid robot; simulated patient; robot patient; medical education; psychiatry; schizophrenia; artificial intelligence; AI; psychiatric nursing

Introduction

Background

Learning and practicing the collection of psychopathological findings is an essential aspect of teaching as part of medical studies. This practice summarizes the results of a psychiatric examination and forms the basis for diagnostic decisions and therapeutic measures [1]. Assessing findings is an important part of physicians' everyday work as they spend up to 80% of their time communicating with patients [2]. In the field of psychiatry in general, having various training opportunities is important so that students can understand complex behaviors and use medical conversation techniques. The ability to determine a diagnosis can be learned in practical training units. In addition to theoretical lessons, human patient actors are used to simulate realistic situations [3]. The role of the patient is played by human actors who have been prepared for the role in advance by the lecturer and a corresponding script. Symptoms can be objectively visible or subjectively experienced [4]. Recognition of symptoms and correct classification requires knowledge of psychopathological terminology and training in perception [5]. Some signs are almost pathognomonic for a diagnosis, but most signs can also occur with various clinical symptoms and sometimes in healthy people [6].

Recognizing and identifying symptoms is crucial not only for medical students but also for nurses. Accurate symptom assessment forms the basis for effective interdisciplinary collaboration. Nurses are often the first to notice changes in a patient's health status and, thus, provide crucial information for medical decision-making. In multidisciplinary teams, such as in psychiatry or intensive care, nurses contribute to the early detection of complications and the adjustment of treatment strategies through their continuous observation of patients. Trained diagnostic perception among nursing staff not only improves the quality of care but also strengthens communication between professional groups and promotes joint, patient-centered action.

Previous Training Methods and Their Limitations

Human actors have traditionally simulated patient behavior by following scripts, but their performance can still vary due to environmental factors, time, conditions, or participant interactions [7]. The representation of difficult and complex patient behavior by human actors is often stressful and difficult to standardize, leading to improvised and, therefore, varying interactions between actor and student. In addition, the preparation of roles and their variations is very time-consuming [8]. These seminars also require a great organizational and financial effort, which allows students only very limited practice opportunities [2]. Role-playing with fellow students is an alternative to acting patients in videos but has drawbacks such as familiarity, lack of feedback training, and limited empathy. In particular, complex psychiatric behaviors are difficult to simulate realistically using case studies or videos [9].

Potential of Humanoid Robots as Simulated Patients

So-called patient simulators [10] have revolutionized medical training, assessment, and research for several reasons. They can be used to recreate high-risk situations in controlled and safe environments. Patient simulators are consistent across users because their performance is not affected by individual trainee differences or other environmental variables that influence human behavior. This is especially true given recent advances in robotics [11]. New-generation humanoid robots attract interest due to their humanlike appearance and social-interactive capabilities, enabling physical interaction as inanimate patient simulators [12,13]. The abilities and strengths of modern humanoid robots are particularly well suited to the representation of facial expressions and gestures. Therefore, when designing the training scenario, we focused on the imitation of psychiatric disorders by the robot. In a prestudy, we investigated the technical requirements and acceptance of the students [14].

In the current simulated patient program at the University of Oldenburg, the clinical picture of schizophrenia is also practiced with acting patients. Existing records of acting patients enabled the robot to reconstruct patient behavior. In collaboration with colleagues from the Department of Psychiatry, the clinical picture of schizophrenia was selected as a clinical case study for the robot simulated patient in our training scenario. In 2021, approximately 0.3% of the population worldwide had schizophrenia; this value has been stable in recent decades [15]. Commonly affected symptom domains include thought processes, emotions, body perception, and interpersonal contact, often leading to difficulties distinguishing reality from subjective experience [16,17]. Schizophrenia can manifest in a wide variety of symptoms. We will use some example symptoms to examine the feasibility of patient simulation by humanoid robots.

Aim of the Study

On the basis of our previous studies [18,19] and the limitations of current humanoid robots for the patient simulation program, we developed our own approach.

The aim of this pilot study was to investigate whether patient behavior simulated by a humanoid robot is comparable to patient simulation by actors in videos in terms of training success and user acceptance. Two groups of participants with a medical background were asked to conduct a psychopathological assessment on the humanoid robot (group 1-R) and the human actor in the video (group 2-V) and make a diagnosis based on this. The participants were instructed to recognize the symptoms presented by the humanoid robot or the human actor and make a diagnosis.

The training concept is generally based on the taxonomy by Bloom [20] and promotes a cognitive learning process. On this basis, a study design was developed that measured the diagnostic performance and subjective perception of participants when using a humanoid robot compared to a video with an actor playing a patient. To evaluate training outcomes, this pilot study

XSL•FO
RenderX

examined participants' acceptance of the robot as a learning tool and their perception of the robot's simulated behavior realism. The following research questions were answered:

1. Symptom identification: are the participants in both groups able to identify the symptoms presented by the given simulated patient?
2. Correct and incorrect diagnoses and influencing factors: what is the percentage of correct and incorrect diagnoses in the 2 groups?
3. Reliability of diagnosis: do the participants consider their own diagnoses to be reliable (self-assessment), and were there differences in relation to each group?
4. Realism in patient presentation: is it possible to simulate and present the patient behavior shown by the humanoid robot as realistically as possible?

The first research question investigates whether the participants in both groups can recognize the symptoms presented by the simulated patient, whether there is a difference in the ability to make a correct diagnosis between groups 1-R and 2-V, and how realistic the presented symptoms were from the participants' perspective. The second research question determines the proportion of correct and incorrect diagnoses in both groups, as well as criteria for determining user acceptance of the robot-assisted procedure. The comparison between robot and actor patients enabled an objective assessment of standardization, reproducibility, and acceptance. The third research question examines whether participants consider their own diagnoses to be reliable and whether this self-assessment differs between the 2 groups. The fourth research question observes whether a humanoid robot can realistically represent patient behavior. Technical limitations such as limited facial expressions and mouth movements were taken into account, and the transferability of complex behaviors, such as those associated with schizophrenia, to the robot was evaluated.

Methods

Overview

The Study Design section describes the study procedure. As part of the pilot study, the robot presented a use case through facial expressions, gestures, and speech. The Case Design section describes the use case that we evaluated. For this purpose, the most important requirements for the humanoid robot as a simulated patient were analyzed (for a full analysis of requirements, refer to Multimedia Appendix 1).

Study Design

One objective was to investigate whether participants from the medical field are able to identify psychopathological symptoms in a simulated patient. For this purpose, both a humanoid robot as a simulated patient and videos of an actor patient were used, and 2 groups of participants were formed: group 1-R conducted the assessment using the humanoid robot patient, and group 2-V did so using the human patient actor in the video. The participants were assigned to 1 of the 2 groups through block randomization (Figure 1).

Figure 1. Flowchart of the study procedure and individual steps. R: robot; V: video.

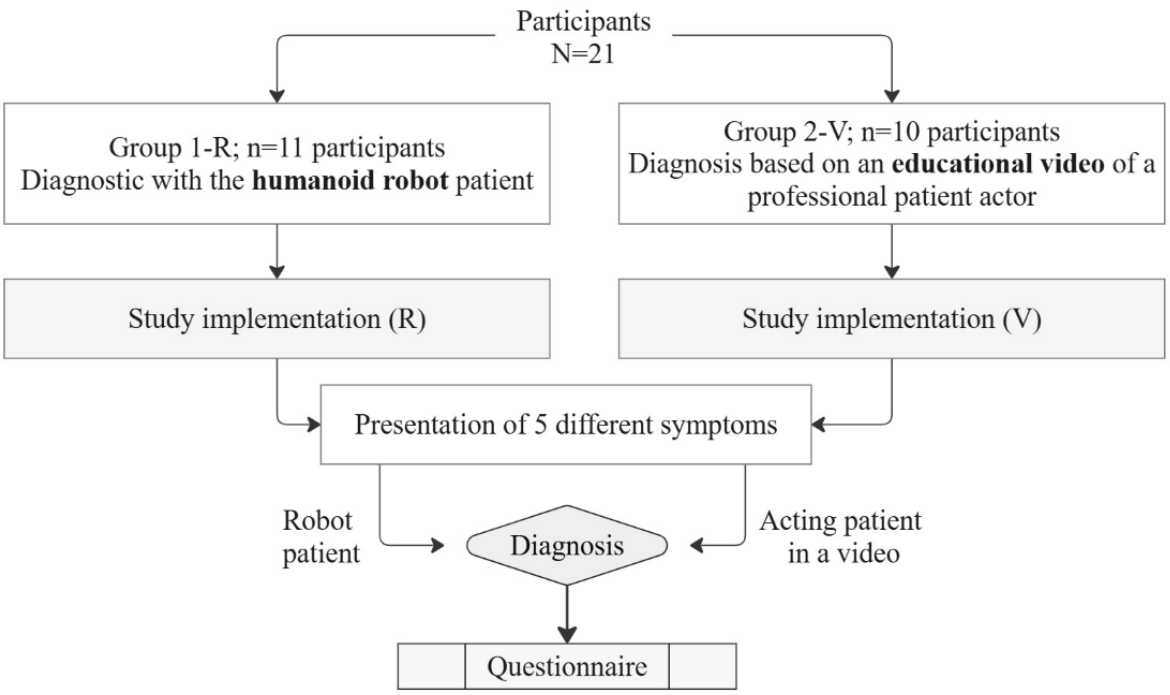

At the beginning of the pilot study, participant inclusion and exclusion criteria were applied. Students or professionals in medicine, nursing, or psychology aged ≥18 years were eligible to participate; individuals who did not understand the study content or were unable to perform the exercises due to severe visual impairments were excluded. Participants were then informed about the aim and procedure of the study and gave their consent.

XSL•FO
RenderX

The procedure was explained to the participants, and the aim of the study was presented. Following the introduction, the participants were asked to complete a questionnaire (Multimedia Appendix 2). This involved a standardized assessment and documentation of psychopathological findings (for the psychopathological categories and associated symptoms covered in this study, see Multimedia Appendix 3). Depending on group allocation, participants observed selected symptom presentations in the robot or the human actor. They conducted a psychopathological assessment after up to 5 symptoms, which were identical in content and sequence for both simulations, each lasting 0.3 to 2 minutes. Participants could diagnose early if they were confident and take notes during the presentations. Following the simulations, participants recorded their findings and completed a questionnaire. Figure 1 illustrates the study process and steps.

Case Design

Overview

The overview graphic in Figure 2 shows the procedure and the individual components required to conduct a psychopathological assessment using a humanoid robot patient. The three phases were (1) theoretical design of the patient case, (2) technical implementation, and (3) study implementation.

Figure 2. Overview of the components required to conduct a psychopathological assessment using a humanoid robot patient (RP). GUI: graphical user interface.

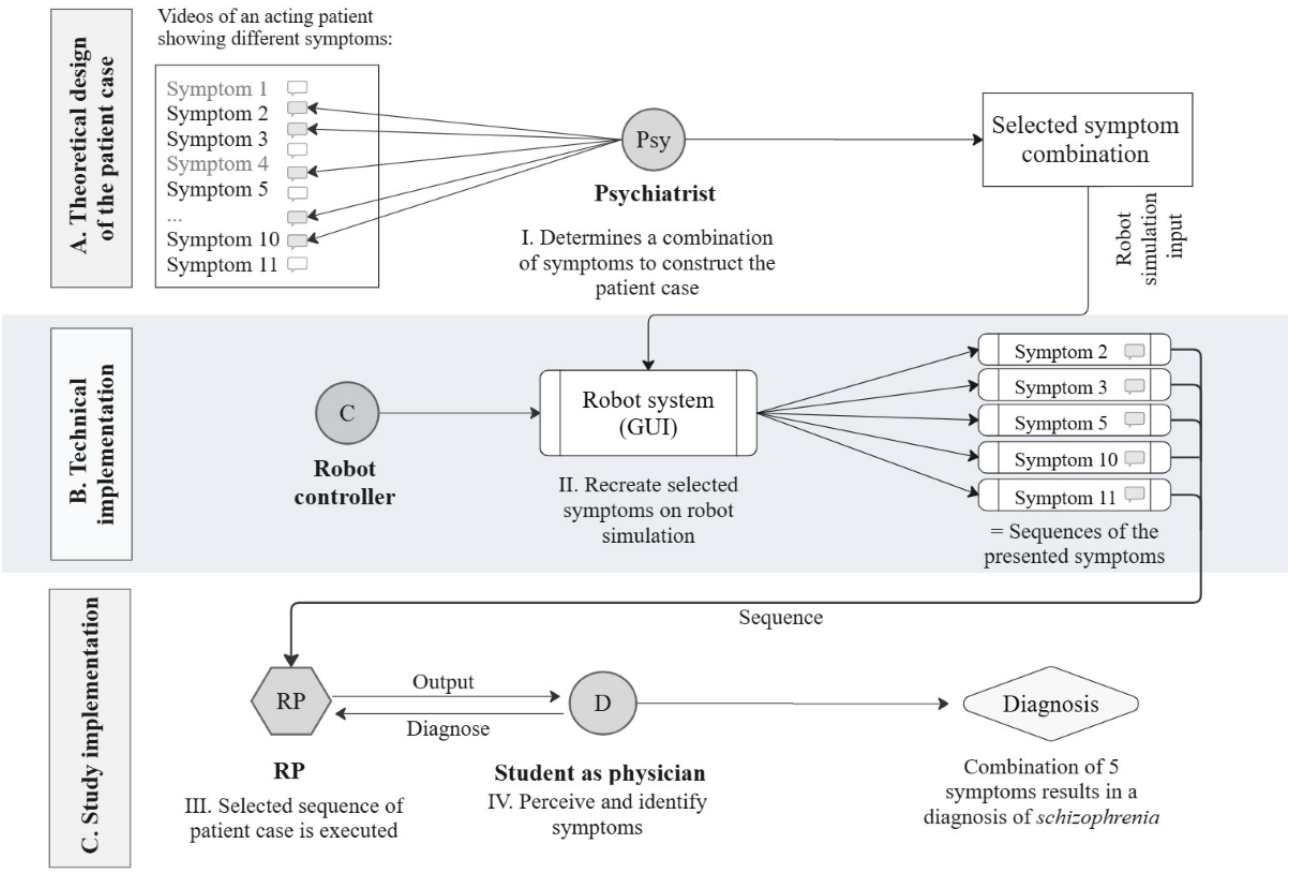

Videos of actor patients presenting various psychopathological symptoms served as a reproducible and precise basis for the robot simulation. This approach enabled direct comparison between the robot and video simulations by participants. In total, 12 videos depicting different symptom combinations, creating distinct clinical profiles, were used to develop the humanoid robot patient. This provided a very detailed basis for transferring the behavior shown in the video to the robot step by step. The individual symptoms were selected by an expert (psychiatrist) and represented the diagnosis of schizophrenia.

All data were transferred to a simulation via a graphical user interface (Figure 2). This was done by reproducing the actors' mood, thoughts, and feelings (as an audio track). In combination with matching facial expressions and gestures (as movement sequences), this resulted in a sequence of symptom representations. The outcome was 5 different sequences, each mapping the content of the videos (the actors' presentation of symptoms) onto the robot. Altogether, these elements produced the robotic presentation of the patient case (Figure 2).

In the next step, the specific audio and movements for each symptom were transferred to the robot management system. Movements, speech, and facial expressions from the video (section I in Figure 2) were transferred the robot's graphical user interface (section II in Figure 2) and then played back on the humanoid robot (section III in Figure 2). Thus, the videos

XSL•FO
RenderX

served as *input* for the programmed sequence, and the simulation result was the *output* of the robot.

The individual symptoms were stored in the robot management system and, thus, could be presented in various combinations by the humanoid robot. The robot exhibited the programmed movements and spoke to the participant acting as the patient. The participant (in the role of the physician) observed the robot patient, listened to the symptom descriptions, and then attempted to make the appropriate diagnosis based on the presented symptoms. For each symptom of schizophrenia—flight of ideas, withdrawal of thoughts, delusions, delusional mood, and thought insertion—there was a short video showing a typical scene with a human actor portraying the patient.

Materials

The humanoid robot (Ameca; Engineered Arts) that we used in our study was specially designed as a platform for the development of future robot technologies and, therefore, is very well suited as a humanoid robot platform for human-robot interaction [20]. Facial expressions, movements, and the style of speech of patients can be programmed in the robot and, thus, imitated. Ameca can show different emotions and gestures, such as anger, joy, surprise, or sadness (Figure 3).

Figure 3. Ameca showing different emotions: anger, joy, surprise, and sadness.

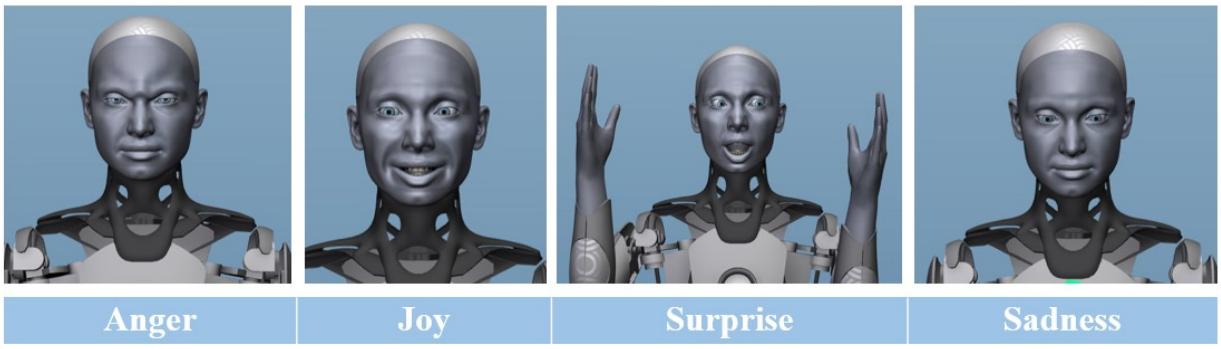

Participants

The pilot study included both medical and nursing students to capture diverse perspectives on patient simulation effectiveness. These groups differ in education, clinical experience, and professional roles, which may affect their symptom perception and diagnostic approaches. Including both cohorts enhanced the generalizability of the results across health care professions and allowed for analysis of how educational background influences symptom identification. Evaluating simulation performance across these groups helped identify specific strengths and limitations, guiding targeted improvements in patient simulation.

Flyers and the student portal were used to recruit students of medicine, psychology, or nursing. General information about the study and contact details of the study director were published on the student portal. Interested individuals had the opportunity to voluntarily register as study participants by contacting the study director via email. Participation was open to students aged >18 years who met the inclusion criteria.

A total of 21 participants were recruited. All participants gave consent for their data to be used in the study. The sessions took place in a university laboratory equipped with necessary assessment materials, including a table, chairs, humanoid robot, screen, psychopathology note sheets, and questionnaires. Participants were randomly assigned to 1 of 2 groups.

Questionnaire

The self-developed questionnaire contained 5 questions (for the full questionnaire, refer to Multimedia Appendix 2 in accordance with research question 1, question 2 of the questionnaire asked whether all symptoms could be identified and how many symptoms the participants viewed before making their diagnosis (question 3). To address research question 2, the diagnosis made was recorded (question 5). Subsequently, the diagnosis made was checked by experts and categorized as correct or incorrect. The questionnaire also recorded the age of the participants, their occupational group (medical student or nurse), and their professional experience in the medical field. In this way, the effect of these data on the correct or incorrect diagnosis was investigated. Question 4 measured how certain the participants were that the diagnosis they had given was correct. This answered research question 3. To answer research question 4, the results of questions 1 and 2 from the questionnaire, as well as the expert review, were used.

Ethical Considerations

The Commission for Research Impact Assessment and Ethics at Carl von Ossietzky University of Oldenburg reviewed and approved this study (2024-016). The participants were fully informed about the study objectives and methodology. General participant information about the study was provided to the participants well in advance of the study start date. After the participants had been informed, their consent was obtained. The consent form contained all necessary information and was attached to the ethics application. The participants were informed about the procedures and objective of the study via a participant information sheet (sent via email). They had at least 1 day to decide to participate. Participants did not experience any consequences if they withdrew. There was no financial compensation for participating in the study. The privacy and confidentiality of the participants' data generated or analyzed during the study will be protected.

Data Collection and Analyses

The data were collected and processed anonymously using a code. The code list will be destroyed once data analysis has

XSL•FO
RenderX

been completed. The data collected as part of this study will be used exclusively for research purposes.

In this study, descriptive data were collected using a questionnaire to evaluate the training method. The questionnaire contained both closed-ended questions (multiple-choice or dichotomous answer formats) and open-ended questions that gave participants the opportunity to formulate their answers freely. The aim of this survey method was to gain an overview of the existing attitudes, experiences, and perceptions of the respondents. The data were evaluated using the statistical software JASP (The JASP Team).

Results

Characteristics of the Participants

In total, 21 participants (n=4, 19% male and n=17, 81% female) took part in this study. All of them had a medical background (n=10, 48% medical students and n=11, 52% nurses; 15 (71%) participants stated that they had medical experience in practice, and 6 (29%) participants stated that they had no experience in practice. Table 1 shows the professional background, age, sex, work experience, and given diagnoses of the participants.

Table 1. Characteristics of the participants (N=21).

Characteristic	Values
Age (y), mean (SD)	28.7 (3.5)
Sex, n (%)	
Male	4 (19)
Female	17 (81)
Profession, n (%)	
Medical student	10 (48)
Nurse	11 (52)
Professional experience (y), mean (SD)	7.6 (3.3)
Professional medical experience in practice, n (%)[a]	
Yes	15 (71)
No	6 (29)
Diagnosis rating, n (%)	
Correct	19 (90)
Incorrect	2 (10)
Given diagnosis, n (%)	
Schizophrenia or paranoid schizophrenia	17 (81)
Psychosis	2 (10)
Anxiety disorder	1 (5)
Delusions	1 (5)

[a]Does the participant have a professional background in the medical/nursing field?

In agreement with experts in the field of psychiatry, the diagnoses of schizophrenia, paranoid schizophrenia, and psychosis were determined to be correct. The diagnoses of anxiety disorder and delusion were rated as incorrect.

Survey Results

Table 2 summarizes the responses to the questionnaire. Most participants rated the human actor patient video as more realistic, with 80% (8/10) selecting “very realistic” and 20% (2/10) selecting “realistic.” In contrast, ratings for the robot patient were more varied: 27% (3/11) rated it as “very realistic,” 45% (5/11) rated it as “realistic,” and 27% (3/11) rated it as “neutral.”

XSL•FO
RenderX

Table 2. Survey results by group.

	Humanoid robot simulation (group 1; n=11)	Video simulation (group 2; n=10)
Realism of patient simulation[a]		
1=very realistic, n (%)	3 (27)	8 (80)
2=realistic, n (%)	5 (45)	2 (20)
3=neutral, n (%)	3 (27)	0 (0)
4=unrealistic, n (%)	0 (0)	0 (0)
5=very unrealistic, n (%)	0 (0)	0 (0)
Score, mean (SD)	2 (0.74)	1.2 (0.4)
All symptoms identified[b]**, n (%)**		
Yes	10 (91)	9 (90)
No	1 (9)	1 (10)
Number of symptoms presented until diagnosis was made[c]		
2, n (%)	2 (18)	2 (20)
3, n (%)	5 (45)	1 (10)
5, n (%)	4 (36)	7 (70)
Mean (SD)	3.5 (1.16)	4.2 (1.3)
Self-confidence in diagnosis[d]		
1=very sure, n (%)	8 (73)	4 (40)
2=sure, n (%)	1 (9)	5 (50)
3=neutral, n (%)	0 (0)	1 (10)
4=unsure, n (%)	1 (9)	1 (10)
5=very unsure, n (%)	1 (9)	0 (0)
Score, mean (SD)	1.73 (1.35)	1.9 (0.94)
Correct diagnosis, n (%)	10 (91)	9 (90)
Incorrect diagnosis, n (%)	1 (9)	1 (10)

[a]"How realistically were the symptoms presented?"

[b]"Were you able to identify all the symptoms presented?"

[c]"How many symptoms were presented to you by the simulated patient before you made a diagnosis?"

[d]"How sure are you that your diagnosis is correct?"

During the patient simulation by the robot, 18% (2/11) of the participants made the diagnosis after only 2 presented symptoms, and 45% (5/11) of the participants made the diagnosis after 3 presented symptoms. The other participants had all 5 symptoms presented to them before they made a diagnosis (Table 2). In the patient simulation by the human actor in the video, 20% (2/10) of the participants made a diagnosis after 2 symptoms were presented, and 10% (1/10) of the participants made the diagnosis after 3 symptoms. The remaining 70% (7/10) of the participants were shown all 5 symptoms before they made a diagnosis. In total, 90% (9/10) of the participants were able to identify all symptoms shown in the video. A total of 91% (10/11) of the participants were able to identify all symptoms shown by the robot.

The participants who diagnosed the actor patient reported that they were neutral (1/10, 10%), sure (4/10, 40%), or very sure (4/10, 40%) of the correct diagnosis. The participants who diagnosed the robot patient said that they were very sure (8/11, 73%), sure (1/11, 9%), unsure (1/11, 9%), or very unsure (1/11, 9%). Finally, we also evaluated the diagnoses made by the participants and checked whether these were correct. Overall, 90% (19/21) of the participants made the correct diagnosis, and 10% (2/21) made an incorrect diagnosis.

Comparison Between the Established Procedure and the New Procedure

Another aspect that was examined was whether the new procedure with the humanoid robot patient was worse or better than using the video with the actor patient (standard procedure). For this purpose, a limit must be defined in advance. In our case, we set a maximum deviation of 20% from the standard procedure. We calculated the deviation based on the average values of the questionnaire responses (Figure 4; average values: Multimedia Appendix 3).

XSL•FO
RenderX

Figure 4. Feasibility and usability study: comparison between the results of the established procedure and the new procedure (A) and CIs (B).

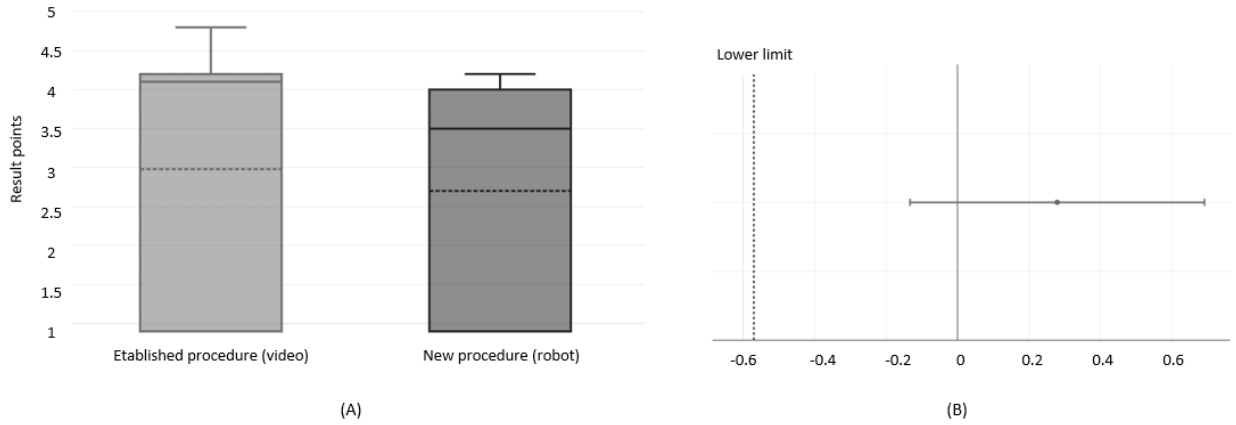

Discussion

Principal Findings

The results show that, although participants found the video simulation more realistic, diagnostic accuracy and self-confidence were comparably high in both groups. Thus, the humanoid robot was able to successfully represent patient behavior in a way that enabled reliable symptom assessment and diagnosis, making it appear to be a suitable learning tool for patient simulation.

Interpretations

The responses from participants in group 1 (robot patient) were more variable. Human actors can portray patients more realistically because they can authentically use complex emotions and subtle nonverbal signals, such as facial expressions, gestures, and body language. These subtle forms of expression are difficult for robots to implement due to technical limitations such as movement speed and flexibility. Humans also adapt their portrayal to reactions and context depending on the situation, which is currently only possible to a limited extent in robots.

The results indicate that only 1 participant per group made an incorrect diagnosis, showing minimal difference between groups. Although human actors' realistic symptom presentation, nonverbal cues, interactive flexibility, and emotional credibility may aid diagnosis accuracy, overall results for the robot and video simulations were comparable. This suggests that the robot's patient portrayal was not significantly inferior to the human actor videos. Using a 20% deviation criterion, the robot simulation was deemed noninferior to the standard procedure, demonstrating equivalence in realism, symptom recognition, and diagnostic accuracy. Thus, in our pilot study, the training method using humanoid robots did not have any significant disadvantages in terms of training success and acceptance compared to video and can offer advantages in certain areas of application.

The results also indicate no direct correlation between older age or greater professional experience and correct diagnosis. Accurate diagnoses were evenly distributed between medical students and nurses, suggesting that professional background is not a decisive factor. Overall, the participants felt less confident about assessing the human actor in the video. This uncertainty could be attributed to various factors. Possible reasons may include a lack of emotional intelligence and empathy in the video. However, the difference between the 2 groups in terms of self-perception was very small. The so-called uncanny valley effect describes the phenomenon in which humanlike robots that look and behave almost but not quite human evoke a feeling of discomfort or aversion in people. A low to medium human likeness is associated with increased positive responses, whereas a shift to negative responses is expected for highly anthropomorphic robots [21]. This feeling can make people feel more uncertain about how to assess the robot.

To address the research questions, participants evaluated the realism of symptom presentation and their ability to identify symptoms. Humanoid robots such as Ameca offer realistic movements and facial expressions through advanced sensors and actuators but remain limited to preprogrammed responses and lack spontaneity. While human actors present symptoms with greater emotional nuance, the robot showed comparable suitability due to its humanlike appearance and complex behavior. However, fully replicating human adaptability and expression remains technically challenging. In the future, large language models could generate patient-specific responses and replace video-based templates.

Comparisons to Existing Literature

The results of our study show that the use of humanoid robots in training to simulate medical scenarios offers learners a more practical and low-risk training opportunity and is met with a high level of acceptance.

Comparable studies on the use of humanoid robots in medical scenarios or in training show similarly positive results in terms of learning success and acceptance among learners [22-24]. In a study by [23], the acceptance, effectiveness, and user-friendliness of an android robot patient was compared with that of a human actor patient for diagnostic teaching, training, and assessment. For this purpose, the android robot patient was

XSL•FO
RenderX

implemented with features for simulating delirium. The results showed that an android robot patient can be used to provide safe training in assessing delirium diagnosis to participants with no previous knowledge. The android robot patient was not difficult for participants to use and led to greater confidence than using a human actor. Takanobu et al [24] developed a humanoid robot for training in dental therapy that can influence treatment through altered facial expressions or active movements of the neck and hands. This enabled students to undergo dental training that was closer to actual practice and avoid risks. Hashimoto et al [25] developed a humanoid robot named Saya. Saya's speech and behavior can be remotely controlled by an operator. A field trial was conducted at an elementary school, where Saya led a science lesson on the principle of levers and interacted with the students.

Limitations

The sample size of 21 participants was small, which limits the generalizability of the results. Furthermore, this was a pilot study with a specific study design that only allowed for a comparison between a specific humanoid robot and a video simulation. Other factors influencing the participants, such as previous technical experience, individual approaches to robots, or differences in clinical experience, were not systematically measured. The robot's limited expressiveness due to the limitation of its facial muscles (number of motors in the face) may also have influenced the perception of the patient simulation. The assessment of acceptance and realism was based on participants' individual subjective opinions, which are sensitive to bias.

Implications

The results of this study show that humanoid robots achieved comparable acceptance and diagnostic accuracy to those of videos featuring actors playing patients. Humanoid robots offer a platform for individualized, artificial intelligence–supported learning processes and the simulation of rare or complex clinical pictures. At the same time, the results highlight the need for further technical development in terms of movement flexibility, facial expressions, and interaction capabilities. Future studies should include larger samples and different clinical pictures to comprehensively test effectiveness and acceptance.

Conclusions

The results of our small pilot study show that the robot was not rated significantly worse by the participants than the video with the actor patient for training in the conduct of psychopathological assessments. In addition to the further technical development of the robot, future development should also take into account corresponding didactic theories and the design of the scenarios in which they are used. Furthermore, the training scenario can be transferred to various areas of application and institutions using the taxonomy by Bloom [20].

The humanoid robot patient embodies didactic concepts, and working with it could increase the motivation and commitment of learners and, at the same time, contribute to the development of problem-solving and teamwork skills. Compared to instructional videos, the robot offers the advantage that learning occurs constructively through working on real problems and knowledge is produced through interaction between people and with the environment.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization and data curation: PS and AH. Writing—original draft: PS. Software: PS. Methodology: PS. Formal analysis: PS, SH, and AH. Writing—review and editing: PS, DI, and SH. Supervision: AH, RH, and SS. All authors approved the final manuscript as submitted and agreed to be accountable for all aspects of the work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Analysis of requirements.
[PDF File (Adobe PDF File), 94 KB - formative_v9i1e70363_app1.pdf]

Multimedia Appendix 2
Full questionnaire.
[PDF File (Adobe PDF File), 97 KB - formative_v9i1e70363_app2.pdf]

Multimedia Appendix 3
Psychopathological categories and associated symptoms assessed in the study and calculation for feasibility and usability study.
[PDF File (Adobe PDF File), 141 KB - formative_v9i1e70363_app3.pdf]

References

XSL•FO
RenderX

1. Fähndrich E, Stieglitz RD. Guidelines for the Assessment of Psychopathological Findings: Semi-Structured Interview Using the AMDP System. Göttingen, Germany: Hogrefe Verlag; 2022.
2. Bachmann C, Roschlaub S, Scherer M. Practicing communication with simulation patients. Ärztekammer Hamburg. URL: https://www.aerztekammer-hamburg.org/files/aerztekammer_hamburg/wissenswertes/kommunikation/HAEB_052016_Kommunikation_mit_Simulationspatienten_ueben.pdf [accessed 2025-05-29]
3. St. Pierre M, Breuer G. Simulation in der Medizin: Grundlegende Konzepte - Klinische Anwendung. Berlin, Germany: Springer; 2013.
4. Linden M. Psychopathologie, deskription und diagnostik psychischer erkrankungen. In: Gastpar MT, Kasper S, Linden M, editors. Psychiatrie und Psychotherapie. Vienna, Austria: Springer; 2003:1-17.
5. Stieglitz RD, Haug A, Kis B, Kleinschmidt S, Thiel A. Praxisbuch AMDP: Psychopathologische Befunderhebung – Grundlagen und Anwendungsbeispiele. Göttingen, Germany: Hogrefe Verlag; 2022.
6. Psychopathological findings. Amboss. URL: https://next.amboss.com/de/article/pM0Lpg#Z924bfdcf251b945a5f6f9c3dd3b8d087 [accessed 2025-05-29]
7. Miller JR, Lerman DC, Fritz JN. An experimental analysis of negative reinforcement contingencies for adult-delivered reprimands. J Appl Behav Anal 2010;43(4):769-773 [FREE Full text] [doi: 10.1901/jaba.2010.43-769] [Medline: 21541165]
8. Schnabel K. Flesh and blood simulation: acting patients. In: St.Pierre M, Breuer G, editors. Simulation in Medicine: Basic Concepts. Berlin, Germany: Springer; 2013:115-119.
9. Falk M. Schauspielpatienten der JLU üben mit Medizinstudenten Arzt-Patienten-Gespräche. Giessener Anzeiger. 2021. URL: https://www.giessener-anzeiger.de/stadt-giessen/schauspielpatienten-der-jlu-ueben-mit-medizinstudenten-arzt-patienten-gespraeche-91007335.html [accessed 2025-05-29]
10. Falk M. Schauspielpatienten der JLU üben mit Medizinstudenten Arzt-Patienten-Gespräche. Gießener Anzeiger. 2021 Sep 27. URL: https://www.giessener-anzeiger.de/stadt-giessen/schauspielpatienten-der-jlu-ueben-mit-medizinstudenten-arzt-patienten-gespraeche-91007335.html [accessed 2025-05-29]
11. Denson JS, Abrahamson S. A computer-controlled patient simulator. JAMA 1969 Apr 21;208(3):504-508 [FREE Full text] [doi: 10.1001/jama.1969.03160030078009] [Medline: 5818529]
12. Kazemi E, Stedman-Falls LM. Can humanoid robots serve as patient simulators in behavior analytic research and practice? Behav Anal Res Pract 2016 Aug;16(3):135-146 [FREE Full text] [doi: 10.1037/bar0000046]
13. Kanda T, Ishiguro H, Imai M, Ono T. Development and evaluation of interactive humanoid robots. Proc IEEE 2004 Nov 16;92(11):1839-1850. [doi: 10.1109/JPROC.2004.835359]
14. Kaneko K, Harada F, Kanehiro G, Miyamori K, Akachi K. Humanoid robot HRP-3. In: Proceedings of the 2008 International Conference on Intelligent Robots and Systems. 2008 Presented at: IROS '08; September 22-26, 2008; Nice, France p. 2471-2478 URL: https://ieeexplore.ieee.org/document/4650604 [doi: 10.1109/IROS.2008.4650604]
15. Radtke R. Global proportion of the population suffering from schizophrenia from 1990 to 2021. Statista. 2021. URL: https://de.statista.com/statistik/daten/studie/1078922/umfrage/anteil-der-weltbevoelkerung-mit-schizophrenie/ [accessed 2025-05-29]
16. William Jr F, Miller DR, Wiener M. What Is Schizophrenia?. Berlin, Germany: Springer; 2012.
17. Karow A, Rietschel L, Bock T, Lambert M. Information on psychosis. psychenet. URL: https://www.psychenet.de/de/psychische-gesundheit/informationen/psychosen.htmlPD [accessed 2025-05-29]
18. Schwarz P, Hellmers S, Spanknebel S, Hurlemann R, Hein A. Humanoid patient robot for diagnostic training in medical and psychiatric education. Front Robot AI 2024 Oct 09;11:1424845 [FREE Full text] [doi: 10.3389/frobt.2024.1424845] [Medline: 39445149]
19. Schwarz P, Hein A. Conception of a humanoid-robot-patient in education to train and practice. In: Proceedings of the 2nd German Education Conference. 2023 Presented at: GECon '23; August 2-4, 2023; Berlin, Germany p. 1-5 URL: https://ieeexplore.ieee.org/document/10295118 [doi: 10.1109/gecon58119.2023.10295118]
20. Ulbrich H. Prozessmerkmale guten unterrichts mit flipped classroom: didaktische ansprüche an homeschooling. R&E source 2021 Oct 15;16:1-10 [FREE Full text] [doi: 10.53349/resource.2021.i16.a1003]
21. Ameca - the future face of robotics. Engineered Arts Limited. URL: https://engineeredarts.com/robot/ameca/ [accessed 2025-05-29]
22. Mara M, Appel M, Gnambs T. Human-like robots and the uncanny valley: a meta-analysis of user responses based on the godspeed scales. Z Psychol 2022 Jan;230(1):33-46 [FREE Full text] [doi: 10.1027/2151-2604/a000486]
23. Röhl JH, Klausen AD, Feldmann N, Diekmann R, Hellmers S, Günther U, et al. Android robot-patient for teaching and training of delirium assessment instruments: a pilot study. In: Proceedings of the 2023 IEEE International Conference on Advanced Robotics and Its Social Impacts. 2023 Presented at: ARSO '23; June 5-7, 2023; Berlin, Germany p. 78-83 URL: https://ieeexplore.ieee.org/document/10187448 [doi: 10.1109/arso56563.2023.10187448]
24. Takanobu H, Okino A, Takanishi A, Madokoro M, Miyazaki Y, Maki K. Dental patient robot. In: Proceedings of the 2006 International Conference on Intelligent Robots and Systems. 2006 Presented at: RSJ '06; October 9-15, 2006; Beijing, China p. 1273-1278 URL: https://ieeexplore.ieee.org/document/4058544 [doi: 10.1109/iros.2006.281888]
25. Hashimoto T, Kobayashi H, Kato N. Educational system with the android robot SAYA and field trial. In: Proceedings of the 2011 IEEE International Conference on Fuzzy Systems. 2011 Presented at: FUZZ-IEEE '11; June 27-30, 2011; Taipei, Taiwan p. 766-771 URL: https://ieeexplore.ieee.org/document/6007430 [doi: 10.1109/fuzzy.2011.6007430]

XSL•FO
RenderX

Edited by A Stone; submitted 20.12.24; peer-reviewed by A Hudon, A Asghar; comments to author 07.07.25; revised version received 01.09.25; accepted 14.10.25; published 05.12.25.

Please cite as:
Schwarz P, Hellmers S, Spanknebel S, Immel D, Hurlemann R, Hein A
Comparing Patient Simulation With a Humanoid Robot or a Human Actor in Terms of Training Success and Acceptance: Pilot Questionnaire Study
JMIR Form Res 2025;9:e70363
URL: https://formative.jmir.org/2025/1/e70363
doi:10.2196/70363
PMID:

©Patricia Schwarz, Sandra Hellmers, Sebastian Spanknebel, Diana Immel, Rene Hurlemann, Andreas Hein. Originally published in JMIR Formative Research (https://formative.jmir.org), 05.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Patterns of Prior Induced Abortions and the Likelihood of Subsequent Natural Pregnancy Loss: Exploratory Application of Pregnancy Outcome Sequencing

Francois K Domagni[1*], PhD; Daniel J Tsang[1*], MS; James Studnicki[2,3*], ScD, MPH, MBA; John W Fisher[3*], PhD, JD, MS, MA; David C Reardon[3,4], PhD; Christopher Craver[3], MA

[1]California State University, Northridge, Northridge, CA, United States
[2]University of North Carolina at Charlotte, Charlotte, NC, United States
[3]Charlotte Lozier Institute, Arlington, VA, United States
[4]Elliot Institute, Gulf Breeze, FL, United States
[*]these authors contributed equally

Corresponding Author:
James Studnicki, ScD, MPH, MBA
Charlotte Lozier Institute
2776 S. Arlington Mill Dr.
#803
Arlington, VA, 22206
United States
Phone: 1 2022238073
Email: jstudnicki@lozierinstitute.org

Abstract

Background: Research concerning the long-term health consequences of induced abortion is constrained by both the limitations in the availability of data necessary to construct complete reproductive histories, as well as the limitations in the analytical methods necessary to interpret them.

Objective: This study aims to determine the association of induced abortion and the likelihood of a subsequent natural loss by applying the pregnancy outcome sequence (POS), a research construct that defines the number and order of all pregnancy outcomes (births, induced abortions, natural losses) in each woman's reproductive history.

Methods: Using the Medicaid Analytic eXtract files from the Centers for Medicare and Medicaid Services Chronic Conditions Warehouse, we identified a study population of 508 unique POSs, representing 5455 women, each of whom had 1 to 16 pregnancy outcomes, for a total of 14,198 pregnancies. We applied an exploratory iterative sequential analytical approach, which included aggregate POS correlation analysis, logistic multiple regression, and simultaneous CIs (Agresti and Tukey-Kramer methods). We also established counting methods to populate the data tables for each analytical phase.

Results: Overall, we found evidence to conclude that both prior abortions and natural losses are significantly associated with the risk of subsequent natural losses. For abortion, there is evidence of a dose-response relationship from 0 to 3 abortions and the likelihood of subsequent natural loss. For natural loss, the risk of a subsequent natural loss is significant after 2 natural losses and between the first and second. There is no association of prior births, or any combination of birth/abortion or birth/natural loss, with the risk of subsequent natural loss. Limitations imposed by the skewed distribution of the number of total pregnancies per reproductive history, resulting in small cell sizes, and exclusion of important covariates, restrain assurance in the results.

Conclusions: The POS demonstrates that the order and combinations of pregnancy outcomes may result in varying conclusions that were previously undetectable. Therefore, further methodological development is indicated.

(JMIR Form Res 2025;9:e78489) doi:10.2196/78489

KEYWORDS

reproductive history; pregnancy outcomes; abortion; natural loss; miscarriage; birth

XSL•FO
RenderX

Introduction

The Knowledge Gap: Toward a Comprehensive Interpretation of Reproductive History

A pregnancy can end in one of 3 ways: an induced abortion, a live birth, or a natural loss. The terms spontaneous abortion, miscarriage, and early natural pregnancy loss are often used interchangeably. In studies of various domains of reproduction, the pregnancy outcomes are considered as both dependent and independent variables. In the former case, the pregnancy outcome is seen as the result of a factor or group of factors that influence the likelihood that the outcome will occur; for example, chromosomal abnormalities and miscarriage, and poverty and abortion [1,2]. In the latter case, the pregnancy outcome is perceived as a precipitating or causative factor for some other outcome; for example, abortion and subsequent repeat abortions, and miscarriage and anxiety and depression [3,4]. Most of these studies are limited by the data available to a single pregnancy outcome as the focus of the analysis, such as the woman's first pregnancy or another identified "index" outcome. However, for most women, a single pregnancy outcome fails to capture the full extent of variation in their reproductive years.

The number, type, timing, and order of all pregnancy outcomes comprise an important component of a woman's reproductive history. This is illustrated by conflicting research on the association between induced abortion and subsequent pregnancy loss. The science related to the various influences and associations of induced abortion to other outcomes is a more contemporary component of the concept of a woman's "reproductive history," a term used by researchers for generations and introduced as a PubMed search term in 1997 [5]. Reproductive history has been defined as "An important aggregate factor in epidemiological studies of women's health. The concept usually includes the number and timing of pregnancies and their outcomes, the incidence of breastfeeding, and may include age of menarche and menopause, regularity of menstruation, fertility, gynecological or obstetric problems, or contraceptive usage" [6]. Research on the association between abortion and subsequent natural loss has produced conflicting results [7-14]. A major methodological limitation confronting analyses of pregnancy outcomes is the reliance on patient recall rather than the objective documentation of one or more abortions. Since randomized controlled trials are not appropriate for studying the risks of abortion, investigators usually use observational study designs (eg, cross-sectional studies or cohorts), and use some form of survey presenting the possibility of both recall and selection bias [15]. Women with abortions tend to underreport having had an abortion when interviewed or responding to a questionnaire [16]. An important finding concluded that women actually identify less than half (47%) of the abortions they experienced [17]. The recall bias may also be selective; that is, it has been demonstrated that healthy participants with an exposure (abortion) are less likely to report their abortion if they have not experienced long-term negative health effects [18]. Actual selection bias occurs when some factor or variable is associated with the outcome of interest. Investigators now routinely include levels of postpregnancy outcomes in the period prior to the pregnancy as a means to control selection bias [19]. Of course, recall bias and other problems associated with survey data are best addressed by using data extracted from a comprehensive registry or medical record summary, which, if complete, provides an accurate record of all pregnancy outcomes within a reproductive history.

We consider the pregnancy outcome sequence (POS) as an analytical construct. A research construct represents an attempt to represent some abstract phenomenon that is not directly measurable. For example, "intelligence" (which is not directly measurable itself) is inferred from other variables that are measurable and observable, such as scores on IQ tests, formal educational attainment, and problem-solving skills. These observable variables are selected to validate the underlying concept represented by the construct; that is, they operationalize it. The POS construct must be operationalized to capture or validate the concept it represents and may have many measurable variables; that is, the total number of outcomes in the POS; the first pregnancy outcome; the time interval between the first and last pregnancy outcome; the percentage of all outcomes represented by a particular outcome; and so on. The POS construct may then enable an explanation or understanding of complex relationships that are hidden in the patterns among the variables that compose the construct.

However, the ordering of pregnancy outcomes is exponentially complex; that is, the number of unique combinations possible increases more and more with each additional pregnancy. A woman with 4 pregnancies during her reproductive years will have 81 different possible orderings of these pregnancy outcomes, 6561 if she has 8 pregnancies, and more than 43 million if she is pregnant 16 times. Methods which attempt to describe the characteristics of this complexity and also interpret its influence on or association with other factors or events are largely absent from the research domain. The availability of and access to data sources necessary to inform these analyses are similarly limited and restricted. This exploratory research attempts an explanation of the likelihood of natural loss given the number, type, and patterns of the order of all pregnancy outcomes.

Benefits of Sequential Analysis

Longitudinal data is typically composed of sets of categorical sequences. Sequential data is arranged in sequences where the order matters. In a period of hospitalization, for example, the medical record is a repository of tests, procedures, and other treatment events, of which patterns emerge with analysis. In the typical descriptive ordering analysis of the association of a particular pregnancy outcome (abortion) with the likelihood of another (miscarriage), the outcome that was closest in time to the subsequent outcome is used for classifying the type of outcome and determining the interpregnancy interval [20]. In sequential analysis, incremental findings may allow a clearer understanding of complex relationships, which may also enable investigators to be more flexible during the research project. In this analysis, for example, a consistent lack of association between birth and subsequent natural loss enabled investigators to pursue a set of strategies to increase sample size. In sum, determining within-sequence patterns for every pair of

XSL•FO
RenderX

pregnancy outcomes in every reproductive history may reveal associations among outcomes that were undetectable without this level of outcome granularity and order.

In this exploratory analysis, we introduce the POS in order to meet 3 analytical objectives. First, we will comprehensively describe the POSs that are extant in a defined population. To our best knowledge, such a description has never been previously accomplished. Second, we will provide a preliminary assessment of the applicability and utility of the POS as a research construct within this research domain. What are the challenges and apparent inadequacies in this application? What are the next steps and possible future applications? Third, we will determine whether this initial analysis supports, refutes, or is inconclusive regarding the association of patterns of abortion and the likelihood of subsequent natural pregnancy loss. Given the exploratory and developmental aspects of this project, no finding would be considered confirmatory or conclusive.

Methods

Data Source and Study Population Coding

This study used data drawn from the Medicaid Analytic eXtract files licensed through the Centers for Medicare and Medicaid Services Chronic Condition Data Warehouse. The total dataset comprises Medicaid-eligible women 13 years of age and older, with at least one pregnancy, from 17 states where state funds provided coverage for induced abortions ineligible for coverage by federal Medicaid. Medicaid eligibility is determined by financial status relative to a designated poverty level, and, therefore, the study results may not be generalizable to more affluent women. In this analysis, the study population was limited to 7 states that continually submitted their full Medicaid claims data for all years of the study period (1999-2015): Connecticut, New Jersey, New Mexico, New York, Oregon, Vermont, and West Virginia. Additionally, to be included in the study population, each patient had to maintain Medicaid eligibility for at least one month in each of the 17 years (1999-2015). This criterion was implemented to prevent pregnancies from being missed due to nonreporting by the states or interruptions in patient eligibility. Study participants were drawn from all women who were or turned 16 years old in 1999. In a similarly determined population of 15-year-old women in 1999, less than 0.47% had a pregnancy, providing high confidence that this study reflects the outcomes of women's first and subsequent pregnancies. These inclusion criteria also provide a control on the range of women's age at the pregnancy outcome since women are all less than 33 years of age during the observation period, largely avoiding the increasing risk of natural loss in later years [21]. We identified all unique pregnancy outcomes using *ICD-9* (*International Classification of Diseases, Ninth Revision*) and *ICD-10* (*International Statistical Classification of Diseases, Tenth Revision*) codes. Pregnancy outcomes were confirmed using Current Procedural Terminology, Fourth Edition (CPT4) and Healthcare Common Procedure Coding System codes.

The following codes were used to identify pregnancy outcomes in 3 categories: live birth (*ICD-9* V27.0, V27.2, V27.5 and *ICD-10* Z370, Z372, and V375); induced abortion (*ICD-9* 635.xx; *ICD-10* O04; CPT4: 59840, 59841, 59850, 59851, 59852, 59855, 59856, and 59857; and Healthcare Common Procedure Coding System: S0199, S2260, S2265, S2266, S2267, X7724, X7726, S0190, and S0191); and natural pregnancy loss, a category that includes miscarriage, stillbirth, ectopic pregnancy, and other pregnancies not ending in induced abortion or live birth (*ICD-9* V27.1, V27.4, V27.7, 630, 631, 633, and 634; ICD-10 O00, O01, O02, and O03). Multiple codes occurring within 30 days of an induced abortion or 180 days of a live birth were collapsed into a single pregnancy outcome based on the first date associated with that group of claims. Twins and higher-order pregnancies that resulted in a combination of live birth and pregnancy loss were excluded.

In the typical multivariate model application, the addition of important variables is encouraged, consistent with model parsimony; that is, simpler models are preferred over more complex ones, provided the models fit the data equally as well. This approach implies that key variables such as comorbid conditions, age at pregnancy outcome, race, type of abortion (medication vs surgical), type of natural loss (miscarriage, stillbirth, and ectopic and molar pregnancy), type of delivery (vaginal vs c-section) and others could influence the likelihood of the subsequent (ie, next pregnancy) outcome, and should therefore be included in the analysis. This research is atypical for multiple reasons. The distribution of the number of pregnancy outcomes within a population of reproductive histories is highly skewed, with nearly three-quarters of women experiencing 3 or fewer pregnancies. At the same time, the number of unique POSs increases exponentially for each additional pregnancy. From a research design perspective, this means that sample sizes become smaller and smaller as the number of possible unique sequences increases. This enables more possible combinations, which are rarely populated or are absent completely from the study population. Adding variables to the model in these circumstances aggravates the problems of statistically underpowered cell sizes and the instability of estimates. Therefore, we concluded that this first exploratory application of sequential analysis to pregnancy outcomes should be parsimonious so as to enable an iterative application of multiple analytical methods that reflect interim results [22].

Ethical Considerations

This study has been exempted from institutional review board review by Sterling Institutional Review Board (IRB ID 7269).

POS as a Research Construct

The POS is a numeric construct representing the occurrence of each of these possible outcomes in the order in which they occur, coding an abortion as a 1, a live birth as a 2, and a natural loss as a 3. For instance, if a woman has had a total of 7 pregnancies with an outcome sequence of birth-abortion-natural loss-birth-abortion-abortion-natural loss appearing in that order, her history would be coded as a POS 2-1-3-2-1-1-3. All women in the data set who experienced this same series of pregnancy outcomes in the same order were assigned this POS. The number of women with the same POS was recorded, and their age at each outcome event was averaged and included in the POS table.

XSL•FO
RenderX

Counting Methods

POS Description

Coding the Medicaid data using the POS structure resulted in a dataset of 5455 women and 14,198 pregnancies grouped in 508 distinct POS groups. The distribution of the number of pregnancies is highly skewed to the right, however. Table S1 in Multimedia Appendix 1 and Figure 1 show the very rapid reduction in Patients/POS with increasing numbers of pregnancies. Note that a cell size minimum of 30 women is no longer met beyond 3 pregnancies and accounts for less than 51% of all pregnancies in the data set. To account for at least 80% of all pregnancies requires including all women who have at least 6 pregnancies, with an average POS cell size of only 1.75 women. Note also that all 14,198 pregnancies populate only 508 POS groups out of a possible 64.6 million. This skewed concentration effectively precludes analyzing at the POS level the effect of sequences on pregnancy outcomes. Instead, we use an exploratory iterative sequential analytical approach to extract information from the disaggregated sequence segments through various ordering counting methods, each method addressing a different question.

Figure 1. Data set POS distribution. POS: pregnancy outcome sequence.

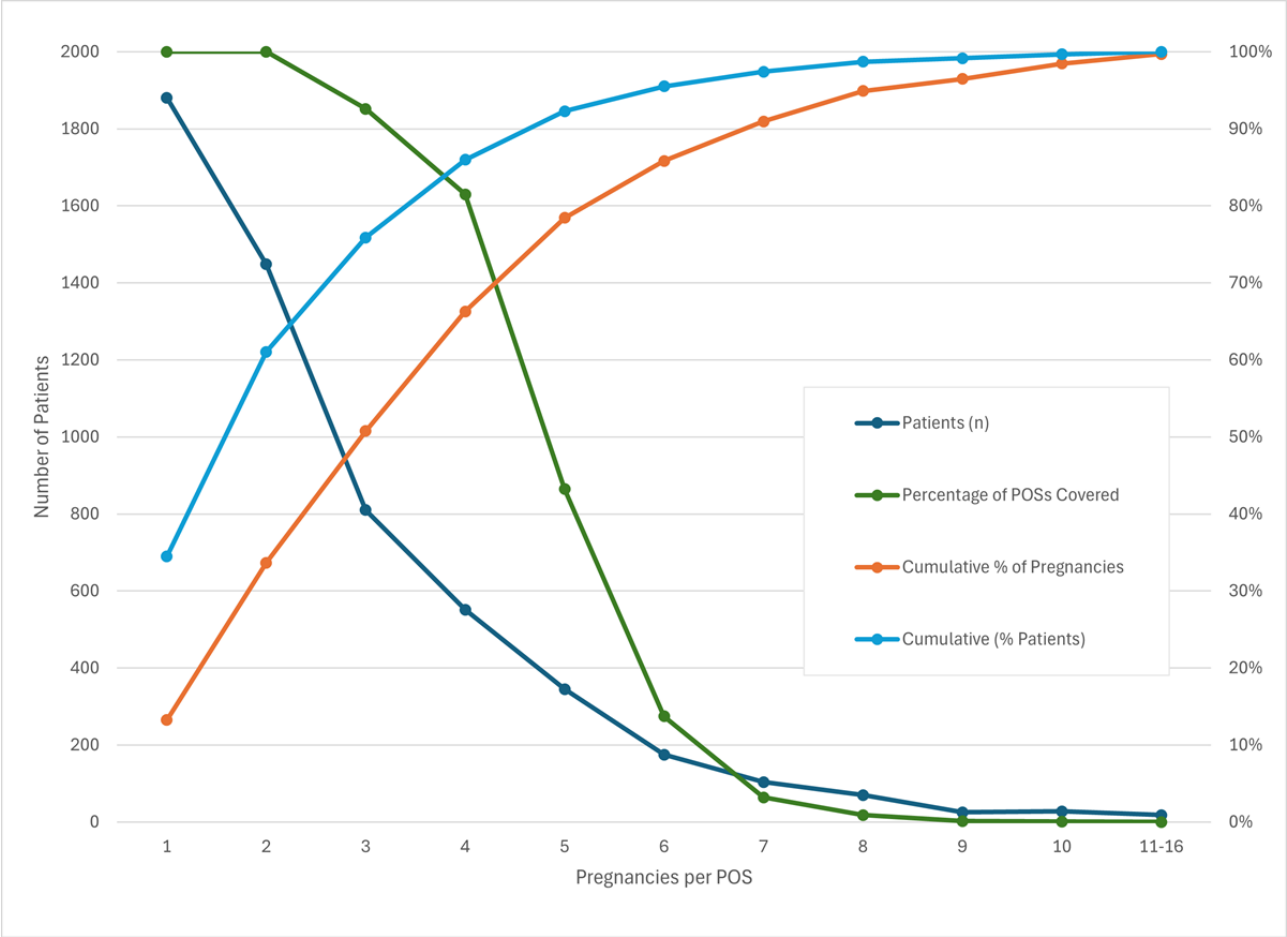

Counting Method 1: Data Aggregation for the Correlation Matrix

The length of the POS is the number of pregnancies that a woman experiences in her lifetime. For example, suppose a woman has a total of 4 pregnancies: birth-abortion-natural loss-birth. This determines a POS of 2-1-3-2.

For each POS sequence, we created a row vector of length 4 to store the frequencies of each specific outcome and the total number of outcomes without regarding the order in which each outcome occurred. In this example, the frequency of abortion is 1, birth is 2, natural loss is 1, and the number of pregnancies is 4. Thus, the 4D row vector underlying the sequence is (1, 2, 1, 4). We compute the row vector for each sequence in the dataset and use it to compute the correlation matrix. In this application, the correlation coefficient measures the strength and direction of the association of each pair of pregnancy variables represented by the 4D row vectors. It is important to note that the unit of analysis is the count of the 4 outcomes within a given POS, disregarding their order. Therefore, the correlation coefficients are a measure of the pairwise associations between the different pregnancy variables of the POS. The coefficients are not influenced by the order, so that, for example, it does not matter whether any abortion comes before or after any natural loss. Thus, the correlation coefficients do not address the influence of the ordering of various total outcomes. Note that the statistical software R was used to obtain the correlation matrix between the 4 columns of the data set.

XSL•FO
RenderX

Counting Method 2: Data Aggregation for Maximum Likelihood Estimation of Specific Sequences of Pregnancy Outcomes

To produce a table of natural losses and births given the patterns of prior abortions for all of the POSs, we started with the sequences where the first pregnancy outcome is not an abortion. For each such sequence, we counted the number of births and natural losses that occur in the sequence before the first abortion occurs. If the sequence did not contain any abortion at all, we counted the number of births and natural losses in the entire sequence. We recorded the number of births and natural losses separately for each sequence. After we coded all the sequences, we aggregated the records of births and natural losses. The aggregate results of births and natural losses represent the frequency and proportions of births and natural losses given no (0) prior abortions. As an example, we compared 2 separate sequences: abortion, birth, birth, natural loss (1-2-2-3), and birth, abortion, abortion, natural loss (2-1-1-3). Thus, in the count of births and natural losses given 0 prior abortions, we skipped sequence 1-2-2-3, but we counted sequence 2-1-1-3 with a tally of one birth and no natural losses. To count the aggregate number of births and natural losses given only one prior abortion, we used the same method as that used to count births and natural losses given no (0) prior abortion. The only difference is that we started counting the births and natural losses right after the first abortion occurred. We counted until right before the second abortion occurred, if any. For instance, for sequence 1-2-2-3, the tally of births and natural losses after one prior abortion is 2 births and 1 natural loss. For sequence 2-1-1-3, the tally of births and natural losses after one (only) prior abortion is no (0) births and no (0) natural losses. For any specific (and only) prior abortion, the same method was used to count the number of natural losses and births. These counts resulted in tables that give the proportions of natural losses and births for circumstances in which there were no abortions in the POS, following any specific abortion within a POS, or in determining dose-response relationships with given numbers of prior abortions. Here, note that the estimation of the proportions obtained via this counting method is equivalent to the maximum likelihood estimation (MLE) of the proportions assuming that the POS sequences follow a multinomial distribution [23].

Counting Method 3: Data Aggregation for a Logistic Regression

We used logistic regression, where the dependent variable (response) is the proportion of natural losses, and the independent variables (doses) are the number of prior abortions, births, and natural losses, respectively, allowing for the detection of any dose-response relationships. A comprehensive logistic regression will include all possible effects of prior abortions, births, and natural losses, including the possible interactions between these outcomes. To populate the table for this, we counted the number of immediate births and natural losses under all the possible combinations of prior abortions, prior births, and prior natural losses. For instance, the tally of the sequence 3-2-1-1-3-1-3 is displayed in Table S2 in Multimedia Appendix 1, which should be understood as follows. The first pregnancy in the sequence 3-2-1-1-3-1-3 is 3, which is a natural loss. Prior to that natural loss, the woman had no pregnancies. Therefore, the combination (Prior Abortions, Prior Births, Prior Natural Losses) is (0, 0, 0). Consequently, the first row of the table is 0-0-0-0-1. Likewise, the second pregnancy of the sequence 3-2-1-1-3-1-3 is 2, which is a birth. Before that birth, the woman had only one pregnancy, which was a natural loss. Thus, the combination (Prior Abortions, Prior Births, Prior Natural Losses) is (0, 0, 1). Consequently, the second row of Table S2 in Multimedia Appendix 1 is 0-0-1-1-0. We skipped the third and fourth pregnancies of the sequence because these pregnancies ended in abortion. The fifth pregnancy was a 3, a natural loss. Before this natural loss, the woman had 2 abortions, one birth, and one natural loss. The combination of (Prior Abortions, Prior Births, Prior Natural Losses) is (2, 1, 1). Thus, the third row of the table is 2-1-1-0-1. We skip the sixth pregnancy since it is an abortion. The seventh pregnancy is a natural loss. Before that natural loss, the combination (Prior Abortions, Prior Births, Prior Natural Losses) is (3, 1, 2). Therefore, the fourth row of the table is 3-1-2-0-1. We used this counting method on all the sequences in the dataset and obtained an aggregation of all the results. Table S3 in Multimedia Appendix 1 is a sample of the aggregation. Note that the goal in this section is to understand the relationship between a woman's history of pregnancy outcomes and her subsequent risk of natural loss. Thus, this counting method is a natural way of associating a woman's pregnancy outcome with her prior history of pregnancy.

Counting Method 4: Data Aggregation for Simultaneous CIs

Next, we analyzed the effects of abortion, birth, and natural loss on the risk of subsequent natural loss separately, using simultaneous CIs (SCI). To do so, we got all of the row vectors from Table S3 in Multimedia Appendix 1 for which exactly 1 entry of (Prior Abortions, Prior Births, Prior Natural Losses) is different from 0. In other words, these are the rows with only prior abortions, only prior births, or only prior natural losses. We also included the row vector where all entries of (Prior Abortions, Prior Births, Prior Natural Losses) were 0 as a control. For instance, the row vector (3, 0, 0, 42, 13, 55, 0.2364) would be included because the number of prior abortions is 3, and the number of prior births and prior natural losses are 0 for each. The meaning of this row vector is that there have been 55 women whose first 3 pregnancies resulted in abortion. However, on their fourth pregnancy, out of the 55 women, 42 had a birth, and the remaining 13 had a natural loss. This corresponds to a 23.64% rate of natural loss given exactly 3 prior abortions.

All such vectors with more than 20 patients are displayed in Table S4 in Multimedia Appendix 1. This is equivalent to the MLE of the proportion of natural loss given a certain history of pregnancy.

We built SCIs for the differences in the rate of natural loss and odds ratios (ORs) underlying each level of prior abortions to determine any statistically significant difference between the different rates. Using the same methodology, we also compared the rate of natural loss and ORs underlying each level of prior births or natural losses.

XSL•FO
RenderX

We applied the Agresti et al [24] SCIs method to obtain SCIs for the differences in the proportions and the ORs between different dose levels.

Counting Method 5: Increasing Statistical Power of SCIs

As will be shown in the logistic regression results, the effect of prior births on subsequent natural loss is not significant. Furthermore, the SCI displayed in section 3.4 shows there is no statistically significant difference in the rate of natural loss or ORs given different levels of prior births. Note that the sample sizes for different levels of prior births used to arrive at this conclusion are reasonably large (4113, 1370, 560, 230, and 103 for 0, 1, 2, 3, and 4 prior births, respectively). Thus, we are highly confident that birth does not have a significant effect on subsequent natural loss. Therefore, we can include sequences with births that occur before the effect level we are analyzing. For instance, if we are looking at the effects of the second abortion, the sequences 1-2-1-2, 2-1-1-3, and 1-1-2 will all be included in the analysis. Therefore, we counted the outcome of the pregnancy right after the second abortion, whether it was a birth or a natural loss. For these 3 sequences given above, the tally of births after 2 abortions is 2, and the tally of natural losses is 1. The same methodology was also used when analyzing the effect of natural loss. We did so without confounding any effect, and by doing so, we increased the statistical power of the SCI and OR in detecting any statistically significant difference in the effects of prior abortions/natural losses on the rate of subsequent natural losses. The results of this counting method are displayed in Tables S5 and S6 in Multimedia Appendix 1. This counting method is equivalent to the MLE of the proportion of natural loss given a certain history of pregnancy.

Results

Correlation Matrix for the Dataset

The correlation matrix (Figure 2) shows a positive correlation between the number of pregnancies and abortions, births, and natural losses. Specifically, the correlation between the number of pregnancies and abortion is 0.73, which is the strongest among all other pairwise correlations. The correlation coefficient between pregnancy and birth is 0.53, a moderate association compared with that of pregnancy and abortion. Between pregnancy and natural loss, the correlation coefficient is 0.36, positive but relatively weak. The correlation coefficient between abortion and birth is –0.11. The correlation coefficients birth-natural loss and abortion-natural loss are respectively 0.018 and 0.061. This suggests weak associations between birth and natural loss and between abortion and natural loss at the aggregate level. Note that the results of the correlation matrix indicate that increasing pregnancies are strongly correlated to abortion, but this method is insufficiently sensitive to explain the association of prior pregnancy outcomes on the likelihood of subsequent natural loss. Therefore, we applied methods to reveal the separate marginal effect of prior pregnancies on the risk of a natural loss.

Figure 2. Correlation matrix.

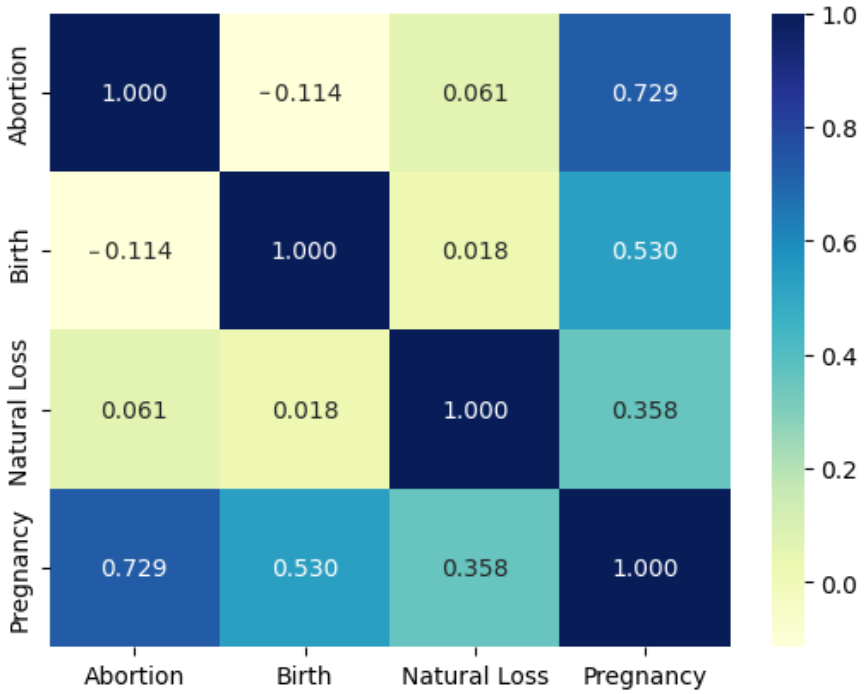

MLE of Specific Sequences of Pregnancy Outcomes

We determined the likelihood of the outcome of the next pregnancy by identifying the pattern of pregnancy outcomes prior to it. In the simplest example, the 2030 women who had only one pregnancy outcome in their reproductive years, a birth, showed the following distribution of proportions for their next outcome: birth 55.7% (1131/2030); abortion 32.5% (660/2030); natural loss 11.8% (239/2030). For POSs that had 2 or more outcomes, the order of prior outcomes was important in determining the proportions of the third outcome. For the 359 women who had a birth first, then an abortion (2-1), the estimated proportions of the third outcome are as follows: birth

XSL•FO
RenderX

30.6% (110/359); abortion 58.2% (209/359); natural loss 11.1% (40/359). For the 289 women who had the same 2 outcomes but in the opposite order (1-2), the estimated proportions of the third outcome are: birth 45.0% (130/289); abortion 40.0% (115/289); natural loss 15.2% (44/289). Even at this 2-outcome level of analysis, the order of events was a significant determinant in the likelihood of a subsequent event.

Our research objective for this analytical phase was to determine whether there was a statistically significant difference between the proportion of natural losses for women with no prior abortions and women with at least one prior abortion. Similarly, we sought to determine the aggregate proportion of natural losses for women with 1, 2, 3, or 4 prior abortions to detect any existing dose-response association. We performed a parallel analysis for the association between prior birth and natural losses, including a no-prior-abortion requirement to isolate any effects of birth from any effects of abortion. We performed a similar analysis for the association between a prior natural loss and subsequent natural losses. Finally, we investigated the effect that the interaction of prior births, prior abortions, and prior natural losses may have on the rate of natural losses.

Logistic Regression

Based on the aggregation result in Table S3 in Multimedia Appendix 1 we build a logistic regression model given by the following equation:

where the predictors X_{1i}, X_{2i}, and X_{3i} are, respectively, the count of prior abortions, prior births, and prior miscarriages. The response, p_i, is the proportion of natural loss for the pregnancy following the sequences containing exactly X_{1i} prior abortions, X_{2i} prior births, and X_{3i} prior natural losses. The observations X_{1i}, X_{2i}, X_{3i}, and p_i are displayed in the ith row of the aggregation result table. We only displayed a sample of the aggregation results in Table S3 in Multimedia Appendix 1 because the full table is huge. $X_{1i}X_{2i}$, $X_{1i}X_{3i}$, and $X_{2i}X_{3i}$ are the interactions between the predictors. Furthermore, is the log odds for the proportion of subsequent natural loss p_i. Note that the goal of this model is to assess the effects and interactions of prior abortions, prior births, and prior natural losses on the risk of subsequent natural loss simultaneously. Table S7 in Multimedia Appendix 1 is the correlation matrix of the predictors. The correlation between prior abortions and prior births is 0.07, the correlation between prior abortions and prior natural losses is 0.09, and the correlation between prior births and prior natural losses is 0.15. All the correlations between the predictors are positive but weak. Thus, we detect no multicollinearity problem and retain all the predictors in our regression analysis. Table 1 displays the logistic regression results. The most significant effects are the intercept and the prior abortions, with P values of <.001. In addition, the P value for the effect of prior natural losses is .007 which is also very small compared with the threshold of .05. All the remaining effects and interactions (Prior births and the interactions of abortion and birth, abortion and natural loss, and birth and natural loss) have P values greater than .05 (P=.51, P=.60, P=.85, P=.13, respectively). We concluded that the number of prior abortions and prior natural losses has a statistically significant effect on the risk of subsequent natural loss. All the remaining effects and interactions (prior births and all the interactions) are not significant. Alternatively, we ran a reduced logistic regression model with the 2 significant effects (prior abortions and prior natural losses) while omitting all the remaining effects and interactions that are not significant. The reduced model is given by

$$\text{logit}(p_i) = \beta_0 + \beta_1 X_{1i} + \beta_3 X_{3i} + \varepsilon_i. \quad (2)$$

Table 1. Results of logistic regression for equation 1.

	Estimate	Exp(Estimate)	SE	z value	*P* value
Intercept	–1.7694	0.1704	0.0385	–45.92	<.001
Prior abortions	0.2085	1.2319	0.0353	5.903	<.001
Prior births	0.0213	1.0216	0.0323	0.66	.51
Prior natural loss	0.2290	1.2573	0.0862	2.656	.008
Prior abortions × prior births	0.0123	1.0124	0.0234	0.526	.60
Prior abortions × prior natural losses	–0.0086	0.9914	0.0452	–0.19	.85
Prior births × prior natural losses	–0.0845	0.9190	0.0557	–1.518	.13

We ran the reduced logistic regression model, and the results are displayed in Table 2. The intercept, the effects of both prior abortions and prior natural losses, are all statistically significant. Next, we determined whether the explanatory power of the reduced model (equation 2) was significantly different from that of the full model (equation 1). We performed an F test and the results are displayed in Table S8 in Multimedia Appendix 1.

XSL•FO
RenderX

Table 2. Results of logistic regression for equation 2.

	Estimate	SE	z value	*P* value
Intercept (β_0)	–1.75555	0.03216	–54.586	<.001
Prior abortions (β_1)	0.21849	0.02521	8.667	<.001
Prior natural losses (β_2)	0.14112	0.06091	2.317	.02

The *P* value of the *F* test is .62. This is a large *P* value compared with the threshold of .05, therefore, we fail to reject the null hypothesis. Thus, the explanatory power of the full model displayed in equation 1 is quasi-equal to the explanatory power of the reduced model shown in equation 2. Note that the parsimony of the reduced model, combined with its great explanatory power, makes it a model of choice to understand the association between a woman's history of pregnancy and her subsequent risk of natural loss. Comparing the 2 models, the intercepts of the full and reduced models are very close to each other (–1.769 and –1.7555, respectively). The coefficients of prior abortions are slightly different (0.2085 and 0.2184). The coefficient of prior natural losses looks different (0.229 vs 0.14112). Using the reduced model, the results indicate that the odds of natural loss are $e^{-1.7555}$, which equals 0.17. This implies that the risk of natural loss for a woman who has no prior pregnancies is 14.5%. The coefficient underlying prior abortion is 0.2184. This implies that a unit increase in abortion changes the odds of natural loss by $e^{0.2184}$, which equals 1.24. This corresponds to a 24% increase in the odds of natural loss for every additional unit of abortion, provided everything else is constant. Likewise, a unit increase in natural loss changes the odds of subsequent natural loss by $e^{0.1411}$, which equals 1.15. This corresponds to a 15% increase in the odds of natural loss for every additional unit of natural loss, provided everything else is constant.

Next, we tested whether the coefficient of prior abortions, β_1, in the reduced model is greater than the coefficient of prior natural losses, β_2. Therefore, let $\gamma=\beta_1-\beta_0$, we solve for β_1 and substitute into the logistic regression, resulting in the following equation.

$$\text{logit}(p_i) = \beta_0 + \gamma X_{1i} + \beta_2(X_{1i} + X_{2i}) + \varepsilon_i \quad (3)$$

The results displayed in Table 3 indicate that the coefficient γ is not statistically significant. Therefore, we do not have significant evidence that the coefficient of prior abortions, β_1, is significantly greater than the coefficient of natural losses, β_2.

Table 3. Results of logistic regression for equation 3.

Estimate	Estimate	SE	z value	*P* value
Intercept (β_0)	–1.75555	0.03216	–54.586	<.001
Difference (γ)	0.07737	0.06812	1.136	.26
Prior abortions + prior natural losses (β_2)	0.14112	0.06091	2.317	.02

Analysis of the SCIs

In this section, we take a different approach to analyzing the data by performing a comparative analysis of the risk of natural loss based on various categories of pregnancies. We carried out the analysis using SCIs for the difference in proportions and the ORs between the categories using 2 methods, the Agresti method (as previously described) and the Tukey-Kramer method, which tests the significance between the means for all pairs of outcomes [24,25]. Before discussing the results, we will briefly explain both methods.

Let [illegible] be respectively the MLEs estimators of the proportions $p_1,\ldots,p_k$ of k independent binomial populations. Define $p_{ij}=p_i - p_j$ and [illegible] as the difference in proportion and the OR between population i and population j, respectively, for $i\neq j$. Clearly, the MLEs of p_{ij} and or_{ij} are respectively [illegible] and [illegible]. Here, [illegible], where x_i is the number of successes in a sample of n_i observations from the population i. The 100(1 − α)% SCI for the pairwise differences p_{ij} and the ORs or_{ij} using Agresti are given by [illegible] and [illegible] respectively, where [illegible] and [illegible]

Using the Tukey-Kramer method, the 100(1−α)% SCI for the pairwise difference in proportion for [illegible], where $Q_k(\alpha)$ is the upper α point of the range of k independent and identically distributed standard normal variates. v_i is the sample variance of $\hat{p}_i$ given by [illegible].

Next, we assessed the effect of a woman's history of pregnancy on her subsequent risk of natural loss. We considered the data analysis from 3 perspectives. The first perspective consisted of carrying out an analysis of the effect of the number of prior abortions on the risk of natural loss. Similar to the first perspective, in the second, we analyzed the effect of the number of prior births on the risk of natural loss. In the third, we analyze the effect of the number of prior natural losses on the risk of another natural loss. In Table S9 in Multimedia Appendix 1, the aggregate numbers of births and natural losses are provided along with their corresponding proportions based on the number of prior abortions. The sample sizes (the total count of births and natural losses) underlying 0, 1, 2, 3, and 4 prior abortions are 4113, 678, 116, 55, and 33, respectively. These sample sizes are reasonably large as compared with 12, 9, 1, and 2, which are the respective sample sizes underlying 5, 6, 7, and 8 prior

XSL•FO
RenderX

abortions. To avoid small sample bias, we omitted the effects of 5 or more prior abortions from our analysis. From 0 to 2 prior abortions, the proportion of natural losses tends to increase slightly (equivalently, the proportion of births tends to decrease slightly) from 14.51% (597/4113) to 15.52% (18/116). From 2 to 3 prior abortions, the proportion of natural losses increases much more, from 15.52% (18/116) to 23.64% (13/55). It then decreases from 23.64% (13/55) to 12.12% (4/33) from 3 to 4 prior abortions (Table S9 in Multimedia Appendix 1). To confirm the trend observed, we constructed 95% SCIs for the pairwise difference in proportions and the ORs using the Agresti method. For the sake of comparison, we also constructed 95% SCIs for the pairwise difference in proportions using the Tukey-Kramer method. Note that both the Agresti method and the Tukey-Kramer method for SCIs are based on the independence and normality assumptions of the sample proportions. Thus, the construction of the SCIs included those cases with 0, 1, 2, 3, and 4 prior abortions. This is because, for those cases, the sample sizes are reasonably large and we can assume that their corresponding sample proportions $p_1,\ldots, p_4$ distributions are roughly normal by the Central Limit Theorem [26]. The 5, 6, and 7 prior abortion cases were omitted from the SCIs' analysis due to small sample sizes.

The results of the SCIs are displayed in Tables 4 and 5. A comparison of the SCIs for the difference in proportions using the Agresti method versus the Tukey-Kramer method shows that neither method is universally superior to the other method. As shown in Tables 4 and 5, in some cases, the SCIs for the difference in proportions using the Agresti method are shorter than those of the Tukey-Kramer method and vice versa. However, all SCIs for the difference in proportions using both methods contain 0. Furthermore, all the SCIs for the ORs displayed in Table 4 contain 1. This indicates that we have no statistically significant evidence that the true proportions of natural loss p_0, p_1, p_2, p_3, and p_4 differ from one another. We may interpret these results in 2 ways. The first implication could be that the number of prior abortions does not actually affect the subsequent risk of natural loss. The second implication could be that since some of the sample sizes are not that large and both tests are conservative, they could not detect any difference in the proportions of natural losses p_0, p_1, p_2, p_3, and p_4. To overcome this problem, we increased the sample sizes and thereby increased the power of both tests to detect any difference in the proportions p_0, p_1, p_2, p_3, and p_4, as seen in the next section.

Table 4. Simultaneous CIs (SCIs) for the difference in proportions and odds ratio (OR; Agresti method [24]) for the pregnancy after the ith abortion, no prior births or natural losses.

Rows	Difference in proportion	95% SCI	OR	95% SCI
(0,1)	0.005	–0.035 to 0.045	1.042923	0.762 to 1.427
(0,2)	0.010	–0.081 to 0.101	1.081735	0.536 to 2.181
(0,3)	0.091	–0.04 to 0.222	1.822924	0.779 to 4.264
(0,4)	–0.024	–0.192 to 0.144	0.8123375	0.191 to 3.46
(1,2)	0.005	–0.093 to 0.103	1.037215	0.49 to 2.194
(1,3)	0.086	–0.053 to 0.225	1.747899	0.718 to 4.255
(1,4)	–0.029	–0.202 to 0.144	0.7789047	0.179 to 3.398
(2,3)	0.081	–0.091 to 0.253	1.685185	0.567 to 5.005
(2,4)	–0.034	–0.225 to 0.157	0.7509579	0.151 to 3.723
(3,4)	–0.115	–0.352 to 0.122	0.4456233	0.084 to 2.37

Table 5. Simultaneous CIs (SCIs) for the difference in proportions and odds ratio (Tukey-Kramer method [25]) for the pregnancy after the ith abortion, no prior births or natural losses.

Rows	Difference in proportion	95% SCI
(0,1)	0.005	–0.035 to 0.046
(0,2)	0.010	–0.083 to 0.103
(0,3)	0.091	–0.066 to 0.248
(0,4)	–0.024	–0.18 to 0.132
(1,2)	0.005	–0.094 to 0.104
(1,3)	0.086	–0.075 to 0.247
(1,4)	–0.029	–0.189 to 0.13
(2,3)	0.081	–0.1 to 0.262
(2,4)	–0.034	–0.214 to 0.146
(3,4)	–0.115	–0.335 to 0.105

XSL•FO
RenderX

Similar to the abortion/natural loss analysis, we aggregated the count of the immediate births and natural losses based on the number of prior births, and with no prior abortion or natural loss. Here, the goal was to assess the effect of prior births on the risk of subsequent natural losses. The aggregate results displayed in Table S10 in Multimedia Appendix 1 show that the sample sizes for the total observations are respectively 4113, 1370, 560, 230, and 103 for 0, 1, 2, 3, and 4 prior births. These sample sizes are reasonably large compared with 28 and 5, which are the sample sizes for the 5 and 6 prior births. We removed the 5 and 6 prior abortion cases from our analysis. The proportions of natural losses are respectively 14.51% (597/4113), 17.45% (239/1370), 13.04% (73/560), 11.30% (26/230), and 15.53% (16/103) under 0, 1, 2, 3, or 4 prior births (Table S10 in Multimedia Appendix 1). To gain more insight from the results, we used the Agresti method and the Tukey-Kramer method to construct the 95% SCI for the difference in proportions. Note that the Agresti method also provides 95% SCI for the ORs for natural losses. As we see from Tables 6 and 7, all of the SCI for the differences in the proportion contain 0. In addition, all the SCIs for the ORs contain 1. Thus, there is no statistically significant difference between the proportions of subsequent natural losses under 0, 1, 2, 3, or 4 prior births. Here, the sample sizes used to construct the SCI are all large; therefore, we may conclude that there is no association between the number of prior births of a woman and her subsequent risk of natural loss.

Table 6. Simultaneous CIs (SCIs) for the difference in proportions and odds ratio (OR; Agresti method [24]) for the pregnancy after the ith birth, no prior abortions or natural losses.

Rows	Difference in proportion	95% SCI	OR	95% SCI
(0,1)	0.029	–0.001 to 0.060	1.245	0.993 to 1.559
(0,2)	–0.015	–0.058 to 0.028	0.883	0.616 to 1.265
(0,3)	–0.032	–0.097 to 0.033	0.751	0.422 to 1.336
(0,4)	0.010	–0.086 to 0.106	1.083	0.515 to 2.276
(1,2)	–0.044	–0.094 to 0.006	0.709	0.480 to 1.047
(1,3)	–0.061	–0.134 to 0.011	0.603	0.332 to 1.094
(1,4)	–0.019	–0.125 to 0.086	0.870	0.408 to 1.856
(2,3)	–0.017	–0.088 to 0.053	0.850	0.440 to 1.643
(2,4)	0.025	–0.075 to 0.125	1.227	0.547 to 2.752
(3,4)	0.042	–0.065 to 0.150	1.443	0.572 to 3.639

Table 7. Simultaneous CIs (SCIs) for the difference in proportions and odds ratio (Tukey-Kramer method [25]) for the pregnancy after the ith birth, no prior abortions or natural losses.

Rows	Difference in proportion	95% SCI
(0,1)	0.029	–0.002 to 0.061
(0,2)	–0.015	–0.056 to 0.027
(0,3)	–0.032	–0.091 to 0.027
(0,4)	0.01	–0.088 to 0.109
(1,2)	–0.044	–0.092 to 0.004
(1,3)	–0.061	–0.125 to 0.002
(1,4)	–0.019	–0.12 to 0.082
(2,3)	–0.017	–0.086 to 0.052
(2,4)	0.025	–0.08 to 0.13
(3,4)	0.042	–0.07 to 0.155

Similar to the first and second perspectives, here the aggregation of the count of births and natural losses is done with respect to the number of prior natural losses. Table S11 in Multimedia Appendix 1 shows that for 0 and 1 prior natural losses, the sample sizes for the total count of births and natural losses are, respectively, 4113 and 377. These sample sizes are relatively large as compared with 24 and 4, which are the sample sizes underlying 2 and 3 prior natural losses, respectively. We omitted the case with 3 prior natural losses from our analysis due to its small sample size. From 0 prior natural losses to one prior natural loss, the proportion of subsequent natural losses goes from 14.51% (597/4113) to 9.02% (34/377). From 1 to 2 prior abortions, that proportion goes from 9.02% (34/377) to 25% (6/24) (Table S11 in Multimedia Appendix 1). The constructions of the SCI using both the Agresti and the Tukey-Kramer methods show that the SCI for the difference in proportion of

XSL•FO
RenderX

subsequent natural loss under 0 prior natural losses and one prior natural loss is significant. The corresponding OR is significant as well. That is, we see a significant drop in the risk of natural loss when a woman goes from 0 prior natural loss to one prior natural loss (Tables 8 and 9). However, the test could not detect any difference in the proportion of subsequent natural loss for the remaining pairs (0,2) and (1,2). This could be due to the fact that the sample size under 2 prior natural losses, which is only 24, is relatively small.

Table 8. Simultaneous CIs (SCIs) for the difference in proportions and odds ratio (OR; Agresti method [24]) for the pregnancy after the ith natural loss, no prior abortions or births.

Rows	Difference in proportion	95% SCI	OR	95% SCI
(0,1)	–0.055	–0.103 to –0.007	0.584	0.364 to 0.937
(0,2)	0.105	–0.081 to 0.29	1.963	0.604 to 6.376
(1,2)	0.160	–0.002 to 0.322	3.363	0.954 to 11.85

Table 9. Simultaneous CIs (SCIs) for the difference in proportions and odds ratio (Tukey-Kramer method [25]) for the pregnancy after the ith natural loss, no prior abortions or births.

Rows	Difference in proportion	95% SCI
(0,1)	–0.055	–0.092 to –0.018
(0,2)	0.105	–0.103 to 0.312
(1,2)	0.160	–0.05 to 0.37

The conclusion from this subsection can be summarized as follows. We found no difference in the risk of subsequent natural loss under 0, 1, 2, 3, and 4 prior births. The sample sizes used to arrive at that conclusion are large for each case (Table S10 in Multimedia Appendix 1). Therefore, we are confident to assume that there is no association between a woman's prior history of birth and her subsequent risk of natural loss. The analysis and the statistical test performed for the risk of subsequent natural loss under prior abortions could not detect any difference in the risk of subsequent natural loss based on prior abortion history. Likewise, the statistical test could not detect any difference in the risk of subsequent natural loss for the pairs (0,2) and (1,2). In the next subsection, we performed the same analysis for prior abortions and prior natural losses. However, this time, when counting the number of prior abortions of a sequence, we allowed the prior abortions to be interspersed with some births but not with natural loss. For instance, for the following 3 sequences 1-1-2, 2-1-2-2-1-3, and 2-3-1-1-3, when we tallied the births and natural losses after 2 prior abortions, we did not take into account sequence 2-3-1-1-3 because there is a natural loss in the sequence before the 2 abortions occurred. Note that for sequences 1-1-2 and 2-1-2-2-1-3, the tally of the immediate birth and natural loss after 2 prior abortions is respectively 1 and 1. Recall, we did not find any significant effect of prior births on the risk of natural loss. Therefore, by including sequences that have a prior birth before the target number of prior abortions occurred, we are not confounding the effects of prior abortions and prior births since the effect of prior births is negligible. By using this counting method, we increased the sample sizes of the observations in our analysis and thereby increased the power of our test in detecting any statistically significant evidence about the risk of natural loss based on the number of prior abortions or the number of prior natural losses.

Analysis of the Effect of Prior Abortions and Prior Natural Losses Under Increased Sample Sizes

By using the counting technique just described, the sample sizes of the count of immediate births and natural losses underlying 0, 1, 2, 3, and 4 prior abortions go from 4113, 678, 116, 55, and 33 to 4113, 886, 202, 109, and 53, respectively (Table S5 in Multimedia Appendix 1). Here again, we constructed 95% SCI under the increased sample sizes for the pairwise difference in proportions and the ORs for 0, 1, 2, 3, and 4 prior abortions using the Agresti method. For comparison, we also constructed 95% SCIs under the increased sample sizes for the pairwise difference in proportions using the Tukey-Kramer method. As we see from Tables 10 and 11, the Agresti method detected 4 significant effects, and the Tukey-Kramer method detected 2 significant effects. According to the Agresti method, there is a statistically significant increase in the risk of natural loss when a woman goes from 0 prior abortions to 1, 2, or 3 prior abortions. There is also a significant increase in the risk of natural loss when a woman goes from one prior abortion to 3 prior abortions. All other pairwise differences are not significant using the Agresti method. Using the Tukey-Kramer method, we found that the risk of natural loss increases significantly for women who go from 0 prior abortions to 2 or 3 prior abortions. All remaining pairwise differences are not significant. With the same counting technique, the sample sizes of the total immediate births and natural losses go from their original values of 4113, 377, and 24 to 4113, 543, and 54 under 0, 1, and 2 prior natural losses, respectively (Table S6 in Multimedia Appendix 1). To confirm these results, the Agresti and Tukey-Kramer 95% SCIs are displayed in Tables 12 and 13. The Agresti method did not detect any difference in the proportion of natural losses under 0 prior abortions and the proportion of natural losses under one prior natural loss. The Tukey-Kramer method, on the other hand, shows that the proportion of natural losses under one prior natural loss is significantly lower than that under 0 prior natural losses, suggesting that women tend to have more natural losses

XSL•FO
RenderX

in their first pregnancies. Both methods show that the risk of natural loss is significantly higher for women with 2 prior natural losses as compared with women with 0 prior natural losses. Likewise, the risk of natural loss is significantly higher for women with 2 prior natural losses as compared with women with one prior natural loss. Recall that with the original sample sizes, we could not detect any difference between the proportion of natural losses between 0 and 2 prior natural losses, and also between 1 and 2 prior natural losses. Here, by increasing the sample sizes, we gain more power to detect these differences.

Table 10. Simultaneous CIs (SCIs) for the difference in proportions and odds ratio (OR; Agresti method [24]) for the pregnancy the ith abortion, no prior natural losses.

Rows	Difference in proportion	95% SCI	OR	95% SCI
(0,1)	0.038	0.001 to 0.074	1.318	1.014 to 1.712
(0,2)	0.092	0.022 to 0.163	1.836	1.166 to 2.89
(0,3)	0.158	0.063 to 0.252	2.557	1.468 to 4.455
(0,4)	0.100	–0.033 to 0.233	1.914	0.816 to 4.49
(1,2)	0.055	–0.029 to 0.139	1.393	0.848 to 2.288
(1,3)	0.120	0.01 to 0.23	1.941	1.076 to 3.501
(1,4)	0.062	–0.088 to 0.213	1.452	0.605 to 3.487
(2,3)	0.065	–0.077 to 0.207	1.393	0.694 to 2.797
(2,4)	0.008	–0.172 to 0.187	1.043	0.403 to 2.699
(3,4)	–0.057	–0.263 to 0.149	0.748	0.274 to 2.041

Table 11. Simultaneous CIs (SCIs) for the difference in proportions and odds ratio (Tukey-Kramer method [25]) for the pregnancy after the ith abortion, no prior natural losses.

Rows	Difference in proportion	95% SCI
(0,1)	0.038	–0.001 to 0.076
(0,2)	0.092	0.009 to 0.176
(0,3)	0.158	0.037 to 0.279
(0,4)	0.1	–0.062 to 0.262
(1,2)	0.055	–0.034 to 0.144
(1,3)	0.12	–0.005 to 0.245
(1,4)	0.062	–0.103 to 0.227
(2,3)	0.065	–0.08 to 0.21
(2,4)	0.008	–0.173 to 0.188
(3,4)	–0.057	–0.258 to 0.144

Table 12. Simultaneous CIs (SCIs) for the difference in proportions and odds ratio (OR; Agresti method [24]) for the pregnancy after the ith natural loss, no prior abortions.

Rows	Difference in proportion	95% SCI	OR	95% SCI
(0,1)	–0.038	–0.079 to 0.002	0.704	0.485 to 1.022
(0,2)	0.207	0.082 to 0.332	3.197	1.597 to 6.403
(1,2)	0.245	0.122 to 0.368	4.539	2.098 to 9.821

Table 13. Simultaneous CIs (SCIs) for the difference in proportions and odds ratio (Tukey-Kramer method [25]) for the pregnancy after the ith natural loss, no prior abortions.

Rows	Difference in proportion	95% SCI
(0,1)	–0.038	–0.072 to –0.005
(0,2)	0.207	0.054 to 0.36
(1,2)	0.245	0.09 to 0.4

XSL•FO
RenderX

Discussion

Value of Sequential Analysis

This exploratory application of sequential analyses to the formation of a reproductive history has demonstrated that the variation in the order of sequential outcomes may reveal associations not otherwise apparent using other methods. It has also revealed that sequence ordering analysis results in extraordinary complexity at scale and that the pilot application is unwieldy. Our preliminary results are internally consistent. However, limitations related to inadequate sample sizes in comparisons beyond 4 total pregnancy outcomes, exclusion of explanatory variables of likely importance, and the use of a data source selected because it enabled us to construct reproductive histories based upon certifiable, claims-based pregnancy outcomes—but without the benefit of representativeness, constrain our ability to consider the findings as confirmatory or decisive.

Limits of Correlation Analysis

While the correlation analysis indicated that the number of total lifetime pregnancies is strongly associated with the likelihood of abortion, the correlation coefficient does not consider the ordering of outcomes within the POS nor the marginal effects of prior pregnancies on the risk of a natural loss. Therefore, it is a limited metric for determining the association of abortion with subsequent natural loss.

Logistic Regression Findings

The logistic regression model indicated an association between the number of prior abortions or prior natural losses and an increased likelihood of subsequent natural loss. The remaining outcome (prior birth) and all interaction terms are not significant. The model indicates a 24% increase in the odds of natural loss for every additional abortion and a 15% increase for every additional natural loss, although the coefficient differences are not significant.

SCI Analysis

An iterative analysis of the SCI for differences in the proportions and ORs using 2 separate approaches (Agresti and Tukey-Kramer methods) and increased sample size indicated that there is no association between the number of prior births and subsequent risk of natural loss; that there is evidence of a dose-response association between prior abortion and subsequent natural loss when a woman goes from 0 to 1, 2, or 3 prior abortions; and that the risk of natural loss is significantly higher for women with 2 prior natural losses and also between 1 and 2 prior natural losses. Overall, these findings suggest that a history of pregnancy loss via either induced abortion or natural loss may be associated with a significantly greater likelihood of experiencing a subsequent natural loss. The effect size on the variation is larger for prior induced abortion than for prior natural loss. For all analyses in this study, no significant association of prior births to the likelihood of natural loss is evident.

While logistic regression results suggest a dose-response association between prior abortions and subsequent natural pregnancy loss, the SCI analyses often failed to detect statistically significant differences. This inconsistency weakens the strength of the conclusions and indicates that the observed associations should be interpreted with caution.

Study Limitations

There are serious limitations to this study. The results may not be generalizable to other populations since Medicaid eligibility is determined by financial status, which may confer other systematic biases, such as for comorbid conditions and age. Moreover, the data were restricted to Medicaid-eligible women in 7 states. Additionally, the requirement for continuous Medicaid eligibility excluded women with interrupted coverage and could have introduced selection bias if the excluded women exhibited systematically different outcome sequences than those who were continuously enrolled. Important study variables such as maternal age, comorbid conditions, and between-outcome time intervals were excluded from the parsimonious model and could risk confounding and misinterpretation of results. For example, the time intervals between repeated pregnancy outcomes are likely to be associated with the risk of any outcome in some nonlinear pattern. Complications related to consecutive births, for example, are more likely to occur if the interval between births is less than 18 months or more than 5 years. Each of the 3 basic pregnancy outcomes has a level of descriptive granularity that could result in significant variation: birth (vaginal or c-section); abortion (medication or surgical); natural loss (miscarriage, stillbirth, ectopic, and molar pregnancy). This analysis did not specify the type of induced abortion; that is, medical versus surgical, nor maternal age [20,21]. It did not distinguish between types of natural pregnancy losses, that is, miscarriage, stillbirth, ectopic pregnancies, nor neonatal losses. Without granularity, these results may mask potentially meaningful differences. To illustrate, research from a Chinese population of 2953 pregnant women concluded that women whose pregnancy was terminated by vacuum aspiration experienced an increased risk of first-trimester miscarriage in a subsequent pregnancy [9]. Lack of granularity regarding the specific type of abortion in our POS analysis limits the ability to claim consistency with these results. Similarly, recent research has concluded that in women with a history of 3 or more miscarriages, the subsequent risk may increase with the presence and number of previous caesarean births [27]. Again, lack of granularity regarding the type of birth in the POS analysis limits comparability of the findings.

The use of administrative data for claims payment is not ideal for research purposes and is subject to coding errors, the exclusion of codes considered nonessential for billing purposes, and inconsistent coding practices [28,29].

The application of the POS construct revealed other issues which could influence the analysis and the validity of these results. The dataset is highly skewed toward women with 3 or fewer pregnancies. For women with 4 or more pregnancies, the very small sample sizes produce unstable estimates and limit the robustness of the sequential outcome analysis. POSs with 3 or fewer pregnancies account for 75.9% (4139/5455) of all women and 50.8% (7207/14,198) of all pregnancies. Nearly all of the possible POSs for 3 or fewer pregnancies are represented in our

XSL•FO
RenderX

sample (37 out of 39 possible). However, at 5 total pregnancies, our sample contains only 43.2% (105/243) of possible POSs, and at 8 pregnancies we capture less than 1% (60/6561) of possible POSs. As a result, for this population, we have confidence in our analysis up to 4 total pregnancies but not beyond. Cohorts of women with 4 or more pregnancies could enable valid assessment of the effects of pregnancy outcome ordering in reproductive histories with higher numbers of pregnancies. Similarly, other dimensions of the POS, aside from counts by type, are likely to affect outcomes. For example, the time intervals between outcomes and the mother's age at each outcome are obvious candidates for inclusion.

Future Applications of the POS Methodology

Finally, the POS may have significant utility in determining the likelihood of a range of adverse outcomes as defined by disease (eg, heart disease), health care use (eg, emergency department visits or hospital admissions), behavioral events (eg, suicidal ideation), or prescription drug use history. The seminal analysis of the POS presented here reveals the unique analytical potential of the construct but also its considerable conceptual and analytical complexity. Future development of the POS construct will address issues of scalability beyond 4 pregnancies and pursue a practical, efficient, and versatile methodology.

The POS also provides a serious cautionary note regarding research focused on a particular pregnancy outcome of interest (eg, first pregnancy and "index pregnancy") or some count or average of all outcomes (eg, "multiple abortions"). This exploratory research suggests that the specific order of outcomes within the POS may shape the observed likelihood of subsequent outcomes. To realize the considerable research potential of the POS, investigators must have access to data which is extracted from a surveillance system in which all pregnancy outcomes are comprehensively reported. The collection of population-based health registries in Finland is a good example and widely used in medical and health services research [30]. In the United States, by contrast, valid and certified pregnancy outcome histories require the aggregation of claims payment data with all their attendant inadequacies. Nonetheless, this exploratory application of the POS demonstrates its potential capability to reveal and interpret the influence of heretofore hidden patterns of reproductive histories on various outcomes of interest. Despite this potential, substantial methodological and interpretive shortcomings identified in this pilot application significantly affect the reliability, generalizability and validity of the findings.

Acknowledgments

This work was funded by the Charlotte Lozier Institute. The sponsor had no role in the review or approval of the manuscript for publication. The authors thank Tessa Cox for her assistance with the investigation; reviewing and editing; and project administration.

Data Availability

The dataset analyzed during this study can be obtained through a data use agreement from the Centers for Medicare and Medicaid Services (CMS).

Authors' Contributions

JS conceptualized the study, acquired funding, and provided supervision. The methodology was developed by JS, JWF, FKD, and DT. FKD, DT, and JWF conducted the formal analysis, and all authors contributed to the investigation. Software was managed by CC and JWF. FKD, DT, JWF, and CC provided validation. The original draft was written by FKD, DT, and JS, with review and editing by the remaining authors. JS provided project administration.

Conflicts of Interest

In accordance with JMIR Formative Research's definition of "non-financial competing interests," which includes relationship with research institutions, JS, JWF, DCR, and CC declare affiliation with the Charlotte Lozier Institute, which educates policymakers, the media, and the public on the value of life from fertilization to natural death; DCR declares affiliation with the Elliot Institute, which conducts original research on the impact of abortion; and FKD and DJT declare personal fees from the Charlotte Lozier Institute.

Multimedia Appendix 1
Supplementary materials.
[DOCX File , 33 KB - formative_v9i1e78489_app1.docx]

References

1. Hardy PJ, Hardy K. Chromosomal instability in first trimester miscarriage: a common cause of pregnancy loss? Transl Pediatr 2018;7(3):211-218 [FREE Full text] [doi: 10.21037/tp.2018.03.02] [Medline: 30159247]
2. Jones RK, Jerman J. Population group abortion rates and lifetime incidence of abortion: United States, 2008-2014. Am J Public Health 2022;112(9):1284-1296. [doi: 10.2105/AJPH.2017.304042r] [Medline: 35969818]

XSL•FO
RenderX

3. Kc S, Gissler M, Virtanen SM, Klemetti R. Risks of adverse perinatal outcomes after repeat terminations of pregnancy by their methods: a nationwide register-based cohort study in Finland 1996-2013. Paediatr Perinat Epidemiol 2017;31(6):485-492. [doi: 10.1111/ppe.12389] [Medline: 28815662]
4. Nynas J, Narang P, Kolikonda MK, Lippmann S. Depression and anxiety following early pregnancy loss: recommendations for primary care providers. Prim Care Companion CNS Disord 2015;17(1) [FREE Full text] [doi: 10.4088/PCC.14r01721] [Medline: 26137360]
5. Beall CM, Leslie PW. Collecting women's reproductive histories. Am J Hum Biol 2014;26(5):577-589 [FREE Full text] [doi: 10.1002/ajhb.22543] [Medline: 24665016]
6. Reproductive history. National Library of Medicine. 2025. URL: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/reproductive-history [accessed 2025-05-29]
7. Kline J, Stein Z, Susser M, Warburton D. Induced abortion and spontaneous abortion: no connection? Am J Epidemiol 1978;107(4):290-298. [doi: 10.1093/oxfordjournals.aje.a112544] [Medline: 645698]
8. Chung CS, Smith RG, Steinhoff PG, Mi MP. Induced abortion and spontaneous fetal loss in subsequent pregnancies. Am J Public Health 1982;72(6):548-554. [doi: 10.2105/ajph.72.6.548] [Medline: 7072872]
9. Sun Y, Che Y, Gao E, Olsen J, Zhou W. Induced abortion and risk of subsequent miscarriage. Int J Epidemiol 2003;32(3):449-454. [doi: 10.1093/ije/dyg093] [Medline: 12777435]
10. Catalano R, Bruckner TA, Karasek D, Adler NE, Mortensen LH. Shared risk aversion in spontaneous and induced abortion. Hum Reprod 2016;31(5):1113-1119. [doi: 10.1093/humrep/dew031] [Medline: 26965433]
11. Xu S, Zhao W, Zhang Y, Qiang C, Zhang C. The effect of previous induced abortion history on the assisted reproduction outcomes. Arch Gynecol Obstet 2024;309(2):469-474. [doi: 10.1007/s00404-023-06928-7] [Medline: 36708427]
12. Thorp JM, Hartmann KE, Shadigian E. Long-term physical and psychological health consequences of induced abortion: review of the evidence. Obstet Gynecol Surv 2003;58(1):67-79. [doi: 10.1097/00006254-200301000-00023] [Medline: 12544786]
13. Lowit A, Bhattacharya S, Bhattacharya S. Obstetric performance following an induced abortion. Best Pract Res Clin Obstet Gynaecol 2010;24(5):667-682. [doi: 10.1016/j.bpobgyn.2010.02.015] [Medline: 20362515]
14. Zou L, Hong M, Dai Z, Zhu J, Peng Q, Wang W. The association between previous induced abortion and in vitro fertilization outcomes: a retrospective cohort study in Hefei, China. Eur J Obstet Gynecol Reprod Biol 2021;262:124-128. [doi: 10.1016/j.ejogrb.2021.05.020] [Medline: 34015639]
15. National Academies of Sciences, Engineering, and Medicine, Health and Medicine Division, Board on Health Care Services, Board on Population Health and Public Health Practice. The Safety and Quality of Abortion Care in the United States. Washington (DC): National Academies Press; 2018.
16. Lindberg LD, Maddow-Zimet I, Mueller J, VandeVusse A. Randomized experimental testing of new survey approaches to improve abortion reporting in the United States. Perspect Sex Reprod Health 2022;54(4):142-155 [FREE Full text] [doi: 10.1363/psrh.12217] [Medline: 36511507]
17. Jones RK, Kost K. Underreporting of induced and spontaneous abortion in the United States: an analysis of the 2002 National Survey of Family Growth. Stud Fam Plann 2007;38(3):187-197. [doi: 10.1111/j.1728-4465.2007.00130.x] [Medline: 17933292]
18. Anderson B, Katus K, Puur A, Silver B. The validity of survey responses on abortion: evidence from Estonia. Demography 1994;31(1). [doi: 10.2307/2061911]
19. Studnicki J, Longbons T, Fisher J, Reardon DC, Skop I, Cirucci CA, et al. A cohort study of mental health services utilization following a first pregnancy abortion or birth. Int J Womens Health 2023;15:955-963 [FREE Full text] [doi: 10.2147/IJWH.S410798] [Medline: 37342485]
20. Virk J, Zhang J, Olsen J. Medical abortion and the risk of subsequent adverse pregnancy outcomes. N Engl J Med 2007;357(7):648-653. [doi: 10.1056/NEJMoa070445] [Medline: 17699814]
21. Magnus MC, Wilcox AJ, Morken N, Weinberg CR, Håberg SE. Role of maternal age and pregnancy history in risk of miscarriage: prospective register based study. BMJ 2019;364:l869 [FREE Full text] [doi: 10.1136/bmj.l869] [Medline: 30894356]
22. Lakens D. Performing high‐powered studies efficiently with sequential analyses. Euro J Social Psych 2014;44(7):701-710. [doi: 10.1002/ejsp.2023]
23. Tsang D. An exploratory application of pregnancy outcome sequencing: patterns of prior pregnancy outcomes and the likelihood of subsequent natural pregnancy loss. In: Master's Thesis. Northridge: California State University Northridge; 2025.
24. Agresti A, Bini M, Bertaccini B, Ryu E. Simultaneous confidence intervals for comparing binomial parameters. Biometrics 2008;64(4):1270-1275. [doi: 10.1111/j.1541-0420.2008.00990.x] [Medline: 18266891]
25. Piegorsch WW. Multiple comparisons for analyzing dichotomous response. Biometrics 1991;47(1):45. [doi: 10.2307/2532494]
26. Kwak SG, Kim JH. Central limit theorem: the cornerstone of modern statistics. Korean J Anesthesiol 2017;70(2):144-156 [FREE Full text] [doi: 10.4097/kjae.2017.70.2.144] [Medline: 28367284]

XSL•FO
RenderX

27. Chung Y, Easter C, Coomarasamy A, Devall A. P-546 History of caesarean births and risk of miscarriage - secondary analysis of three randomised controlled trials. Human Reproduction 2023;38((Supplement_1)). [doi: 10.1093/humrep/dead093.886]
28. Romano PS. Using administrative data to identify associations between implanted medical devices and chronic diseases. Ann Epidemiol 2000;10(4):197-199. [doi: 10.1016/s1047-2797(00)00041-7] [Medline: 10854953]
29. Virnig B, Parsons H. Strengths and limitations of CMS administrative data in research. ResDAC. 2021. URL: https://resdac.org/articles/strengths-and-limitations-cms-administrative-data-research [accessed 2025-05-29]
30. Helle N, Niinimäki M, Linnakaari R, But A, Gissler M, Heikinheimo O, et al. National register data are of value in studies on miscarriage-validation of the healthcare register data in Finland. Acta Obstet Gynecol Scand 2022;101(11):1245-1252 [FREE Full text] [doi: 10.1111/aogs.14445] [Medline: 36056916]

Abbreviations

CPT4: Current Procedural Terminology, Fourth Edition
ICD-10: International Statistical Classification of Diseases, Tenth Revision
ICD-9: International Classification of Diseases, Ninth Revision
MLE: maximum likelihood estimation
OR: odds ratio
POS: pregnancy outcome sequence
SCI: simultaneous CI

Edited by J Sarvestan; submitted 04.06.25; peer-reviewed by A Di Nallo; comments to author 10.08.25; revised version received 28.08.25; accepted 26.09.25; published 09.12.25.

Please cite as:
Domagni FK, Tsang DJ, Studnicki J, Fisher JW, Reardon DC, Craver C
Patterns of Prior Induced Abortions and the Likelihood of Subsequent Natural Pregnancy Loss: Exploratory Application of Pregnancy Outcome Sequencing
JMIR Form Res 2025;9:e78489
URL: https://formative.jmir.org/2025/1/e78489
doi:10.2196/78489
PMID:

©Francois K Domagni, Daniel J Tsang, James Studnicki, John W Fisher, David C Reardon, Christopher Craver. Originally published in JMIR Formative Research (https://formative.jmir.org), 09.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Large Language Model Evaluation in Traditional Chinese Medicine for Stroke: Quantitative Benchmarking Study

Hulin Long[1], BSc; Yang Deng[2], MSc; Yaoguang Guo[1], MSc; Zifan Shen[3], BSc; Yuzhu Zhang[4], BSc; Ji Bao[2], PhD, MD; Yang He[1], MSc

[1]Hospital of Chengdu University of Traditional Chinese Medicine, Chengdu, Sichuan Province, China
[2]Institute of Clinical Pathology, West China Hospital of Sichuan University, Chengdu, China
[3]Physical Education Institute, Southwest Medical University, Luzhou, China
[4]Chengdu University of Traditional Chinese Medicine, Chengdu, China

Corresponding Author:
Yang He, MSc
Hospital of Chengdu University of Traditional Chinese Medicine
No. 39 Shi-er-qiao Road
Chengdu, Sichuan Province, 610072
China
Phone: 86 13648085638
Email: yangyanghe123@163.com

Abstract

Background: The application of large language models (LLMs) in medicine is rapidly advancing. However, evaluating LLM capabilities in specialized domains such as traditional Chinese medicine (TCM), which possesses a unique theoretical system and cognitive framework, remains a sizable challenge.

Objective: This study aimed to provide an empirical evaluation of different LLM types in the specialized domain of TCM stroke.

Methods: The Traditional Chinese Medicine-Stroke Evaluation Dataset (TCM-SED), a 203-question benchmark, was systematically constructed. The dataset includes 3 paradigms (short-answer questions, multiple-choice questions, and essay questions) and covers multiple knowledge dimensions, including diagnosis, pattern differentiation and treatment, herbal formulas, acupuncture, interpretation of classical texts, and patient communication. Gold standard answers were established through a multiexpert cross-validation and consensus process. The TCM-SED was subsequently used to comprehensively test 2 representative LLM models: GPT-4o (a leading international general-purpose model) and DeepSeek-R1 (a large model primarily trained on Chinese corpora).

Results: The test results revealed a differentiation in model capabilities across cognitive levels. In objective sections emphasizing precise knowledge recall, DeepSeek-R1 comprehensively outperformed GPT-4o, achieving an accuracy lead of more than 17% in the multiple-choice section (96/137, 70.1% vs 72/137, 52.6%, respectively). Conversely, in the essay section, which tested knowledge integration and complex reasoning, GPT-4o's performance notably surpassed that of DeepSeek-R1. For instance, in the *interpretation of classical texts* category, GPT-4o achieved a scoring rate of 90.5% (181/200), far exceeding DeepSeek-R1 (147/200, 73.5%).

Conclusions: This empirical study demonstrates that Chinese-centric models have a substantial advantage in static knowledge tasks within the TCM domain, whereas leading general-purpose models exhibit stronger dynamic reasoning and content generation capabilities. The TCM-SED, developed as the benchmark for this study, serves as an effective quantitative tool for evaluating and selecting appropriate LLMs for TCM scenarios. It also offers a valuable data foundation and a new research direction for future model optimization and alignment.

(JMIR Form Res 2025;9:e81545) doi:10.2196/81545

KEYWORDS

large language models; traditional Chinese medicine; stroke; evaluation dataset; model evaluation; artificial intelligence

XSL•FO
RenderX

Introduction

In recent years, artificial intelligence technologies, particularly large language models (LLMs), have demonstrated immense potential in the health care sector [1]. LLMs have achieved substantial progress in applications, ranging from aiding clinical decision-making to optimizing patient communication [1]. The outstanding performance of general-purpose LLMs, such as ChatGPT (OpenAI), in natural language processing tasks suggests broad application prospects in complex medical scenarios [2]. However, traditional Chinese medicine (TCM) presents unique challenges for the application of LLMs [3,4]. TCM is a complex knowledge system with a unique theoretical framework, an emphasis on a holistic view (harmony between man and nature and pattern differentiation and treatment), and a reliance on highly individualized diagnosis and treatment plans [3,4].

To address these challenges, domain-specific TCM LLMs (eg, Huatuo and ShenNong) have been developed. Studies indicate that the performance of these specialized models on the Chinese National Medical Licensing Examination for TCM Practitioners is significantly superior to that of general-purpose models, highlighting the necessity of domain specialization [3]. This need for specialization, however, extends beyond technical data adaptation; at its core lie the profound epistemological differences between Chinese and Western medicine. The theoretical cornerstones of TCM, such as Yin-Yang and the Five Elements; Qi, Blood, and Body Fluids; and Zang-fu organs and meridians, constitute a unique cognitive framework with no direct correspondence to the ontology of modern medicine [4].

As an LLM's understanding is fundamentally derived from statistical learning of text patterns, an effective evaluation dataset cannot be limited to recalling factual knowledge. Such a dataset must also be able to deeply assess a model's ability to perform clinical reasoning within the specific framework of TCM. Therefore, the construction of this evaluation dataset is, in itself, a crucial step in translating the abstract TCM knowledge system into a modern, operational format. This process is vital for the digital representation, inheritance, and innovation of TCM knowledge.

General medical LLM evaluation datasets, such as those based on the United States Medical Licensing Examination [5], are insufficient for assessing a model's depth and accuracy in understanding TCM-specific stroke concepts. Therefore, a specialized evaluation dataset for TCM stroke is of paramount importance. Such a dataset serves as both a necessary means to validate the accuracy and safety of models and a fundamental guarantee for guiding model iteration and building user trust.

Existing benchmark test sets often cover only a small fraction of a model's potential output, thus limiting their ability to ensure the safety of health care LLMs [6]. It is crucial to evaluate LLMs in specific application scenarios; for TCM stroke, this requires testing them in specialized TCM contexts. The performance disparity between TCM LLMs and general LLMs on professional examinations [3] indicates that *domain adaptation* in highly specialized fields such as TCM is not a simple task. Merely fine-tuning with domain-specific texts may be insufficient [3], potentially necessitating architectural adjustments or innovative training methods to better capture the subtle, context-dependent knowledge of TCM. The abstract nature of TCM theory and its individualized features [4] may require LLMs to develop a deeper, more contextualized understanding than standard fine-tuning can provide. Consequently, a robust evaluation dataset can not only assess existing models but also drive research into more effective methods for imbuing LLMs with specialized, culturally rich medical knowledge.

The primary objective of this study is to conduct a rigorous empirical evaluation of leading LLMs within the specialized domain of TCM stroke. To enable this evaluation, we first undertook the systematic development of a high-quality, specialized benchmark, the Traditional Chinese Medicine-Stroke Evaluation Dataset (TCM-SED).

This paper describes the development methodology for this new dataset—covering data sourcing, question construction, and multiexpert validation (Figure 1)—and presents the results of our empirical study, which used the TCM-SED to compare the capabilities of GPT-4o (OpenAI) and DeepSeek-R1 (High-Flyer).

This research provides a direct, quantitative comparison of current LLM performance in TCM stroke. Furthermore, the resulting TCM-SED benchmark (and its described methodology) serves as a validated tool and foundation for future model development and evaluation in this specialized field.

XSL•FO
RenderX

Figure 1. Flowchart for the construction of the Traditional Chinese Medicine-Stroke Evaluation Dataset. TCM: traditional Chinese medicine; TCM-SED: Traditional Chinese Medicine-Stroke Evaluation Dataset.

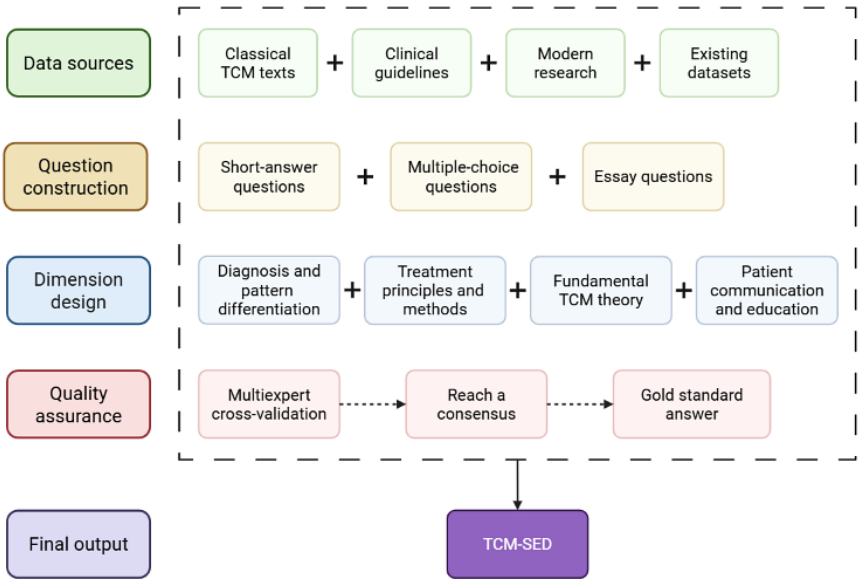

Methods

Collection and Selection of Data Sources

The construction of a high-quality evaluation dataset required a multisource data system that was authoritative, classic, and cutting edge. The data sources for this study were derived from 2 levels, as described in the following paragraphs.

First, the Guidelines for the Diagnosis and Treatment of Cerebral Hemorrhage with Integrated Chinese and Western Medicine published by the Chinese Medical Association [7] were used as the core knowledge framework. These guidelines provided standardized diagnostic criteria, pattern differentiation, and treatment plans, serving as the cornerstone for ensuring the dataset's content accuracy. Concurrently, to trace theoretical origins and ensure cultural depth, classical TCM texts (eg, Huangdi Neijing [Yellow Emperor's Inner Canon] [8] and Jingui Yaolue [Synopsis of the Golden Chamber] [9]) as well as case records from renowned physicians were systematically reviewed. Pathological discussions, formula analyses, and treatment plans related to true stroke and stroke-like syndromes were extracted to ensure the dataset was deeply rooted in the theoretical foundations of TCM.

Second, modern scientific research findings and existing data resources were actively integrated. Modern research papers on the treatment of stroke with TCM were incorporated, with special attention paid to cutting-edge studies that correlate TCM patterns with biomedical indicators or imaging features (eg, computed tomography angiography and computed tomography perfusion) [10], thereby enhancing the scientific and contemporary nature of the dataset. Furthermore, existing public datasets were prudently reviewed. General medical question-answering datasets (eg, WebMedQA [11] and Huatuo-26M [12]) were examined to draw inspiration from their question structures. Specialized TCM datasets (eg, ShenNong-TCM-Dataset [13]) and question banks from the Chinese National Medical Licensing Examination for TCM Practitioners were analyzed to acquire professional question styles and content. TCM clinical case databases, if accessible and structurally processed, would also represent an invaluable supplementary data source [14].

To address the balance and sourcing of the 203 questions, our knowledge dimensions were systematically mapped to the primary data sources described in the previous two paragraphs. This process ensured that questions testing foundational theory were drawn from classical texts, while questions assessing clinical application were derived from modern guidelines and real-world cases. The primary data sources used for each knowledge dimension are summarized in Table 1.

Table 1. Primary data sources for the Traditional Chinese Medicine-Stroke Evaluation Dataset knowledge dimensions.

Knowledge dimension	Classical TCM[a] books	Clinical guidelines	Existing question banks	Real clinical cases
Diagnosis and pattern differentiation		✓	✓	
Treatment principles and methods		✓	✓	
Fundamental TCM theory	✓			
Patient communication and education			✓	✓
Interpretation of classical texts	✓			
Clinical case analysis				✓

[a]TCM: traditional Chinese medicine.

XSL•FO
RenderX

Evaluation Dimensions and Question Construction Paradigms

To comprehensively evaluate the capabilities of LLMs in the field of TCM stroke, 3 core question paradigms were designed: short-answer questions, multiple-choice questions, and essay questions.

Short-answer questions were primarily used to assess the model's ability to master and recall core factual knowledge. These questions required the model to provide precise and concise answers (eg, "What are the main constituent herbs of Buyang Huanwu Tang?"). This paradigm served as a fundamental test of the model's knowledge accuracy and has been applied in several TCM datasets, such as ShenNong [13].

Multiple-choice questions, including single-choice and multiple-choice formats, focused on evaluating the model's ability to discriminate between similar or easily confused concepts. By designing distractors based on common misconceptions in TCM, the model's precise grasp of knowledge points, such as diagnostic criteria and formula compatibility, could be effectively tested. This paradigm was widely adopted in mainstream evaluations such as MedQA [15] and ShenNong-TCM-EB [13] due to its objectivity and ease of automated scoring. As an example of this design principle, a question assessing the identification of TCM patterns in acute ischemic stroke was constructed to include a specific distractor based on a common misconception. While the correct options described valid acute patterns (eg, wind, phlegm, or blood stasis), the distractor was formulated as "Liver Qi stagnation is the most common pattern in the acute phase of stroke." This distractor was specifically designed because liver Qi stagnation, while relevant to stroke, is typically associated with poststroke depression or chronic management, not the primary acute phase pattern. This design principle was applied throughout the dataset to test the model's ability to accurately differentiate nuanced, context-dependent TCM concepts beyond simple recall.

Essay questions were designed to assess higher-order skills, such as knowledge integration, clinical reasoning, and text generation. These questions were typically based on a clinical case or a specific pathogenesis and required the model to conduct an in-depth analysis (eg, "Please discuss the pathogenesis, primary symptoms, and treatment principles and formulas for the pattern of liver-yang transforming into wind."). By analyzing the long-form text answers generated by the model, its logical coherence, the rigor of its pattern-differentiation thinking, and its ability to apply knowledge could be examined, which were crucial for evaluating its potential as an assistant in real clinical scenarios [1,4].

In summary, by combining these 3 complementary question paradigms, a multilevel evaluation system was constructed to conduct a comprehensive and in-depth examination of LLMs across multiple dimensions, from knowledge recall and discrimination to integrated application.

Design and Coverage of Evaluation Dimensions

To ensure the comprehensiveness of the evaluation, the question design of this dataset covered multiple key dimensions of knowledge in the field of TCM stroke. The core dimension was pattern differentiation and treatment (Bian-Zheng-Lun-Zhi), which required the model to perform pattern identification and differential diagnosis based on clinical manifestations, recommend formulas and make modifications, and select appropriate acupuncture points and explain their combination principles. On a theoretical level, the evaluation covered the understanding of pathological factors (eg, phlegm, blood stasis, wind, fire, and deficiency) and the ability to interpret relevant discussions in classical TCM texts, such as the Huangdi Neijing [8]. At the application and extension level, the dimensions were extended to rehabilitation, nursing care, and patient education. Among these, the evaluation of patient education and communication skills was considered particularly important, as the ability of an LLM to explain abstract TCM concepts in simple and accurate language was a key aspect of its value in real clinical settings [1].

Expert Validation and Ground Truth Construction

The scientific validity and reliability of this evaluation dataset relied on the authoritativeness and uniqueness of its ground truth answers. To this end, a rigorous mechanism for expert validation and consensus building was established. The core of this process involved a review team composed of senior experts with extensive clinical experience in TCM stroke. To minimize individual bias and ensure the uniqueness of the final answers, the multireviewer cross-validation model advocated by the Real-World Evaluation of Large Language Models in Healthcare framework [5] was adopted.

This process was particularly critical for establishing the unique *gold standard answer* for all objective (short-answer and multiple-choice) questions. During the review, any initial disagreements on these answers were resolved through group discussion to reach a final consensus or were adopted directly from authoritative guidelines [1,7].

For the essay questions, which required subjective judgment, a separate process of independent expert scoring was used, as detailed in the Testing and Scoring Criteria section. To ensure consistency, detailed annotation guidelines and scoring rubrics were formulated in advance for the experts. Through this series of quality control measures, we ensured that the answer to every question in the dataset possessed the highest level of accuracy and authority.

Testing and Scoring Criteria

The testing and scoring in this study adhered to the principles and standards outlined in the following paragraphs.

First, regarding the testing language, the entire evaluation dataset was presented in Chinese. Using Chinese was considered essential to ensure the highest degree of accuracy and professionalism, particularly for content involving pattern differentiation, formula rationale, and interpretation of classical texts, thereby assessing the model's depth of understanding of core TCM concepts more effectively.

Second, regarding the test participants, this dataset was used to test 2 representative LLMs: GPT-4o and DeepSeek-R1. GPT-4o was chosen because, as one of the most advanced general-purpose large models internationally, it represents the

XSL•FO
RenderX

current technological frontier. DeepSeek-R1 was chosen because, as a large model primarily trained on Chinese corpora, it has a unique advantage in handling tasks related to Chinese language and culture. Both models were the current mainstream versions of their respective series, and their test results hold broad reference value. All tests for both models were conducted within a controlled time frame (from June 16, 2025, to July 10, 2025) to ensure comparability of model versions.

Finally, regarding the scoring criteria, differentiated scoring methods were established for different question types to ensure objectivity and reliability:

- For short-answer and multiple-choice questions, a binary scoring method was used. A model's answer was scored 1 point if it perfectly matched the unique gold standard answer, and 0 points otherwise.
- For essay questions, a rigorous dual-expert independent scoring method was used. Several key scoring points were predefined in the standard answer. Two senior TCM rehabilitation experts were then invited to independently assign a dynamic score (0-10 points) based on the model's coverage and accuracy, without consulting each other.
- To quantitatively assess the consistency of this independent scoring, 2 measures of interrater reliability were calculated across all 40 responses. First, as a formal statistic accounting for chance agreement, Cohen weighted κ (with linear weights) was calculated due to the ordinal nature of the 0 to 10 scale. The analysis yielded κ value of 0.78, which indicates substantial agreement [16]. Second, as a descriptive measure of agreement, it was found that the exact agreement (ie, identical scores) was 38% (15/40), and the within 1-point agreement (ie, scores differed by 0 or 1 point) was 88% (35/40). Both metrics confirmed a robust and high degree of consistency in the independent scoring process.
- The final score for each category reported in the Essay Question Results section (with a maximum of 200 points) was calculated by summing the scores from both experts across all 10 questions in that category (10 questions×2 experts×10 points/question). This method allowed a comprehensive evaluation of the models' integrated abilities by incorporating the independent judgments of both experts.

Ethical Considerations

Institutional Review Board Approval and Exemption

This study was reviewed by the Scientific Research Department of the Hospital of Chengdu University of Traditional Chinese Medicine, which confirmed that the research complies with all medical ethics standards. The study was a secondary analysis and did not involve direct interaction with human participants or access to private patient-identifiable information. Thus, formal institutional review board approval was not required.

Informed Consent: Patients

The requirement for informed consent from patients was waived because the "real clinical cases" used in the clinical case analysis section were compiled from established medical textbooks, published case reports, and fully anonymized public datasets. The study involved no direct interaction with human participants.

Informed Consent and Compensation: Experts

The senior TCM experts who participated in the dataset validation and scoring process were provided with detailed information regarding the study's objectives and provided their informed consent to participate. Their participation was voluntary and uncompensated.

Privacy and Confidentiality

All data used in this study were fully deidentified and anonymized to ensure strict protection of privacy and confidentiality.

Results

Composition of the Evaluation Dataset

Through the systematic construction and validation process described in the Methods section, a specialized evaluation dataset for the field of TCM stroke was finalized and named the TCM-SED.

The TCM-SED comprised a total of 203 questions, structured into 3 core paradigms. The specific composition and purpose of each paradigm were as follows:

- Short-answer questions (46/203, 22.6%)—this section was designed to evaluate the accurate recall of core factual knowledge, covering diagnosis and pattern differentiation, treatment principles and methods, fundamental TCM theory, and patient communication.
- Multiple-choice questions (137/203, 67.4%)—this section, comprising both single- and multiple-choice formats, was focused on testing the model's ability to discriminate between similar or easily confused concepts across the same key knowledge areas.
- Essay questions (20/203, 9.8%)—this section was used to assess higher-order analytic and reasoning skills. It was divided into 2 categories: interpretation of classical texts, which required models to analyze discussions from classical TCM literature, and clinical case analysis, which required models to conduct a full diagnostic and treatment process for a given scenario (including TCM diagnosis, treatment principle, prescription, and formula rationale).

This multilevel and multidimensional structure provided a solid data foundation for the subsequent comprehensive evaluation of LLM capabilities.

Short-Answer Question Results

Overall, in the 46-question short-answer section, DeepSeek-R1 outperformed GPT-4o. DeepSeek-R1 achieved a total accuracy of 83% (38/46), while GPT-4o's total accuracy was 76% (35/46). The performance of the 2 models varied across dimensions, as detailed in Figure 2A.

XSL•FO
RenderX

Figure 2. Performance comparison of GPT-4o and DeepSeek-R1 on the Traditional Chinese Medicine-Stroke Evaluation Dataset: (A) accuracy comparison on short-answer questions across different dimensions, (B) accuracy comparison on multiple-choice questions across different dimensions, and (C) grouped score comparison on essay questions, showing the separate total scores (out of 100) from each of the 2 independent expert raters to visualize interrater consistency. TCM: traditional Chinese medicine.

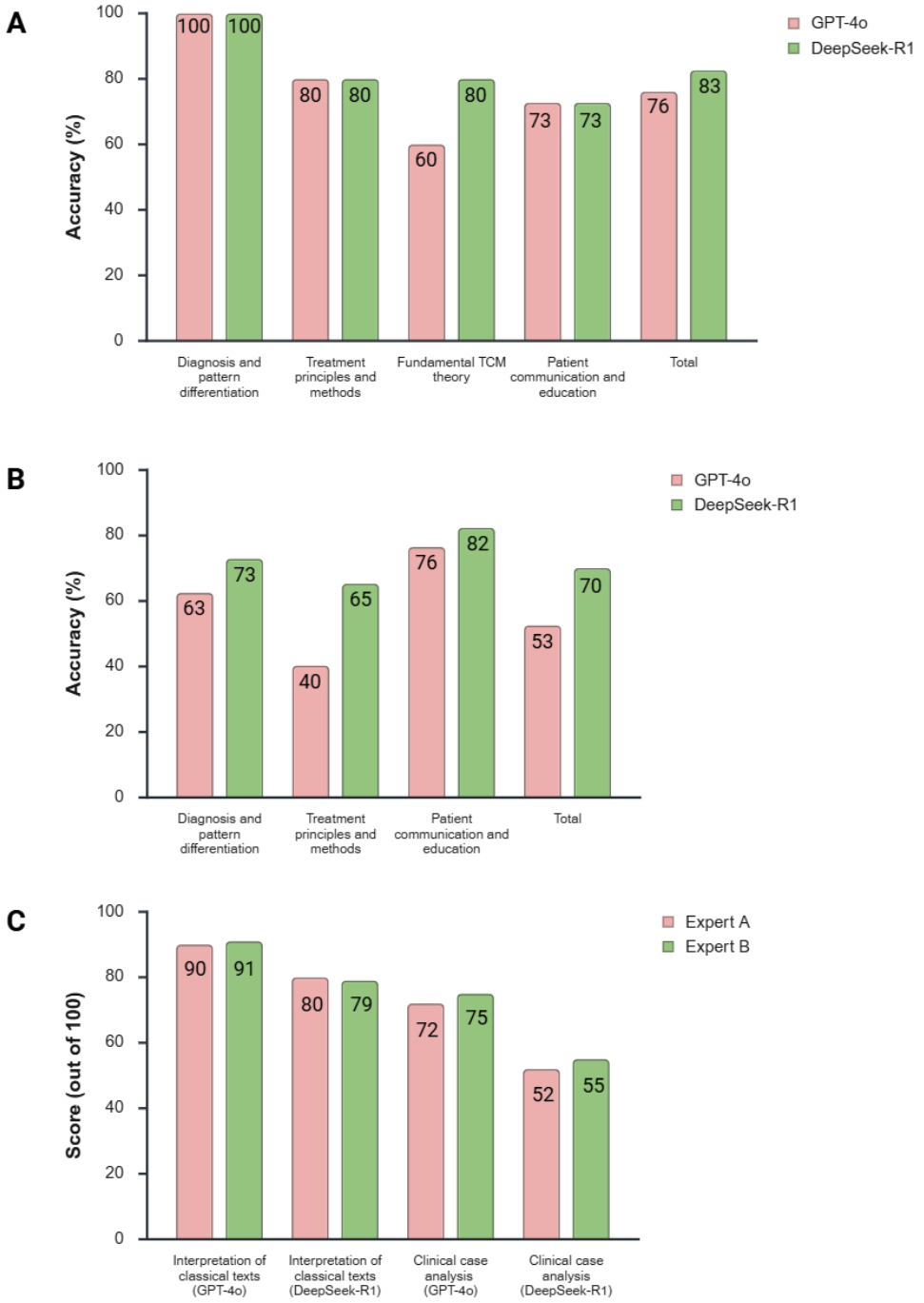

The test results indicated that in the diagnosis and pattern differentiation dimension, both models performed perfectly, achieving 100% (10/10) accuracy. In the treatment principles and methods and patient communication and education dimensions, the models' performances were also identical, with accuracies of 80% (8/10) and 73% (8/11), respectively.

Notably, although the total scores in these latter categories were the same, the specific questions the models answered incorrectly were not identical, indicating distinct knowledge gaps. For instance, DeepSeek-R1 erred on a question regarding acupuncture point selection for poststroke speech impediment, which GPT-4o answered correctly. Conversely, GPT-4o erred on a question regarding the treatment method for a "phlegm-heat binding in the fu-organs" pattern, which DeepSeek-R1 answered correctly.

The most notable performance difference was observed in the *Fundamental TCM Theory* dimension. In this area, DeepSeek-R1's accuracy was 80% (12/15), substantially higher than GPT's 60% (9/15). This result suggested that while both models possessed similar capabilities in handling direct clinical application questions, the DeepSeek-R1 model demonstrated a stronger grasp of deeper, fundamental TCM theoretical knowledge.

XSL•FO
RenderX

Multiple-Choice Question Results

In the 137-question multiple-choice section, the DeepSeek-R1 model's performance was comprehensively superior to that of the GPT-4o model. DeepSeek-R1 achieved an overall accuracy of 70.1% (96/137), whereas GPT-4o's overall accuracy was only 52.6% (72/137). The specific performance of the 2 models across different knowledge dimensions are detailed in Figure 2B.

The results showed that in the multiple-choice section, the performance gap between the 2 models widened further. DeepSeek-R1's accuracy was approximately 17.5% higher than GPT-4o's, indicating its clear advantage in handling TCM knowledge questions that required precise discrimination among distractors. The results from the multiple-choice questions further highlighted the advantage of domain-specific models (or those trained on a specific linguistic and cultural corpus) when dealing with highly specialized and nuanced knowledge.

Among the categories, the performance gap was most pronounced in the *Treatment Principles and Methods* dimension. DeepSeek-R1's accuracy (47/72, 65%) was 25% higher than GPT-4o's (29/72, 40%). This suggested that for clinical knowledge involving specific treatment plans and formula modifications that required detailed analysis, the DeepSeek-R1 model, trained on a core of Chinese corpora, possessed a deeper and more accurate understanding.

The performance of both models in the *patient communication and education* dimension was the highest among their respective categories, reaching 76% (13/17) for GPT-4o and 82% (14/17) for DeepSeek. This reflected that general-purpose large models possessed a relatively good foundational ability in language organization and layperson explanation.

Essay Question Results

The essay question section, consisting of 20 questions (10 for *interpretation of classical texts* and 10 for *clinical case analysis*), was designed to assess higher-order skills. As described in the Testing and Scoring Criteria section, 2 senior TCM experts independently scored each question (from 0 to 10 points). The total score for each category (maximum 200 points) represented the sum of the independent scores from both experts across all 10 questions. This scoring method, which demonstrated *substantial* interrater reliability (κ=0.78), was used to generate the final results.

The test results showed a trend completely opposite to that of the short-answer and multiple-choice questions, as GPT-4o's performance was comprehensively superior to that of DeepSeek-R1's performance. Detailed visual comparison of the scores from each independent expert is shown in Figure 2C.

In the assessment of comprehensive exposition skills, GPT-4o performed markedly better than DeepSeek-R1. In the *interpretation of classical texts* category, GPT-4o achieved a scoring rate of 90.5% (181/200), far exceeding DeepSeek-R1's scoring rate of 73.5% (147/200). In the more complex *clinical case analysis*, GPT-4o's scoring rate (159/200, 79.5%) was also substantially ahead of DeepSeek-R1's scoring rate (107/200, 53.5%).

This result highlighted the differences in the core capabilities of the 2 models. Although DeepSeek-R1 had an advantage in the precise recall of factual knowledge, especially in multiple-choice questions, GPT-4o demonstrated stronger abilities in long-text generation, logical organization, knowledge integration, and complex reasoning. Essay questions required not only accurate knowledge points but also the ability of the model to effectively organize these points into a coherent and rigorous analysis. As a leading international general-purpose large model, GPT-4o's powerful underlying generation and reasoning capabilities enabled it to perform better on such tasks.

Discussion

This study presented an empirical evaluation of 2 representative LLMs, GPT-4o and DeepSeek-R1, within the specialized domain of TCM stroke. The evaluation, which was conducted using the newly developed TCM-SED benchmark, revealed notable differences in their performance capabilities and limitations within this professional domain.

Core Findings and Interpretation

The most central finding of this evaluation was that the DeepSeek-R1 model, primarily trained on Chinese corpora, demonstrated a comprehensively and substantially superior performance compared to the internationally leading general-purpose large model, GPT-4o. This advantage was particularly prominent in the multiple-choice section, which required precise discrimination, with a total accuracy gap of more than 17%.

This difference was attributed primarily to the underlying disparities in their training data. TCM is a knowledge system deeply intertwined with Chinese language, culture, and philosophical thought. The training of the DeepSeek-R1 model on massive Chinese corpora provided it with a more *native* understanding of TCM terminology, context, and unique logical paradigms (eg, pattern-differentiation thinking). In contrast, the general-purpose model GPT-4o appeared to rely more on a *translational* understanding when handling these issues. This *translational* approach, while capable of managing direct, factual short-answer questions, proved insufficient when required to distinguish subtleties in multiple-choice questions, especially those with carefully designed distractors based on TCM theory.

This analogy was used to describe the models' differing proficiency in statistical pattern matching, rather than implying genuine cognitive comprehension. This distinction is critical because core TCM concepts, such as "Yin-Yang and Five Elements" and "Qi," lack direct, one-to-one equivalents in English. A model relying on a *translational* framework might fail to grasp the nuanced, context-dependent application of these foundational concepts. In contrast, a *native* model, trained extensively on the source Chinese corpus, could more accurately replicate the correct contextual patterns and theoretical linkages, even without genuine cognitive comprehension.

The performance differences across evaluation dimensions further corroborated this judgment. In the short-answer questions on *diagnosis and pattern differentiation*, both models performed perfectly, indicating that both possessed strong memory and

XSL•FO
RenderX

recall abilities for clear, structured knowledge from guidelines. However, when it came to the deeper *Fundamental TCM Theory* and the more nuanced *Treatment Principles and Methods*, especially in the multiple-choice questions, the superiority of DeepSeek-R1 became evident. This suggested a capability gap between *knowing what* (ie, the diagnosis) and *knowing why* (ie, understanding the underlying theory and principles), a gap identified as crucial for evaluating TCM LLMs.

However, the results of the essay questions showed a completely opposite trend, revealing a deep-seated differentiation in the capabilities of the 2 models. In both *interpretation of classical texts* and *clinical case analysis* tasks, GPT-4o's performed substantially better than DeepSeek-R1. This finding was interpreted to highlight the core capability differences between the 2 models. Short-answer and multiple-choice questions focused more on the precise recall and discrimination of *static knowledge*, where DeepSeek-R1 excelled due to its *native* advantage in Chinese corpora. In contrast, essay questions required a strong *dynamic capability* to logically organize and integrate disparate knowledge points and to generate well-structured, coherent long-form text. As a leading general-purpose large model, GPT-4o's powerful underlying reasoning and text generation capabilities enabled it to perform better on such complex, open-ended analytical tasks. This finding provided important insights into the applicable scenarios for different models: one model was more akin to a knowledgeable *database*, while the other was more like an adept *reasoning engine*.

Importance and Application Value

The importance of this study is 2-fold.

First, for model development, the results provided clear evidence regarding the development of large models in highly specialized and culturally rich fields such TCM. This evidence suggests that merely fine-tuning a general-purpose model may be insufficient and that the foundational pretraining data are of critical importance. Future research and development should focus more on constructing high-quality, Chinese pretraining corpora that contain deep domain knowledge.

Second, for model evaluation, the TCM-SED, which was constructed for this study, proved to be both effective and necessary. It successfully identified subtle differences in the professional capabilities of different models, providing a quantifiable tool for assessing the safety and reliability of TCM LLMs. This dataset can serve as a benchmark for future iterations of related models, helping developers identify their models' shortcomings.

Limitations and Future Outlook

This study also had certain limitations. First, while the dataset's size (N=203) was sufficient to reveal trends, it was relatively limited. This scale was determined as a pragmatic balance between achieving comprehensive coverage of all core knowledge dimensions (as shown in Table 1) and the substantial time and resource investment required for the rigorous multiexpert validation process. A larger-scale dataset would provide more statistically significant results. Second, this study focused solely on the disease of stroke, and its conclusions require further validation to determine their generalizability to other areas of TCM (eg, internal medicine, gynecology, or pediatrics). Finally, regarding the scoring of essay questions, while our dual-expert method demonstrated *substantial* reliability (κ=0.78), a degree of subjectivity in human annotation cannot be entirely eliminated. This limitation directly informs our proposed future work on exploring methods that combine LLM-assisted evaluation with expert assessment to further improve scoring consistency.

Looking ahead, several directions are proposed:

- Dataset expansion—the question bank can be further expanded in both depth and breadth, incorporating more diseases and introducing more diverse question types, such as multimodal diagnosis.
- Optimization of scoring mechanisms—methods that combine LLM-assisted evaluation with expert assessment should be explored to improve the efficiency and consistency of scoring open-ended questions.
- Dynamic and adversarial evaluation—more challenging dynamic evaluation benchmarks could be developed, such as simulating continuous, multiturn physician-patient dialogues or introducing adversarial questions designed to attack a model's knowledge gaps, to more comprehensively assess model robustness and safety.
- Empowering model optimization and alignment—the TCM-SED is not only a measuring stick but can also serve as a high-quality training set to provide feedback for model improvement. Its expert-validated question-answer pairs can be directly used for supervised fine-tuning to correct model errors on specific knowledge points. In particular, the expert scoring results from the essay questions provide valuable data for building reward models and conducting reinforcement learning from human feedback, which can help achieve a deeper alignment between the model's output and the clinical thinking patterns and value judgments of TCM experts, thereby genuinely enhancing their professional capabilities.

Conclusions

This empirical study rigorously evaluated leading LLMs (GPT-4o and DeepSeek-R1) in the specialized domain of TCM stroke, using the newly developed and expert-validated TCM-SED benchmark.

The findings revealed a clear differentiation in capabilities: the Chinese-centric model (DeepSeek-R1) demonstrated superior performance in *static knowledge* tasks requiring precise recall, whereas the general-purpose model (GPT-4o) excelled in *dynamic reasoning* and complex generation tasks. This highlights the critical importance of culturally and linguistically specific pretraining data for specialized medical domains. The TCM-SED, validated through this process, is now established as an effective tool for future benchmarking, selection, and alignment of LLMs in TCM.

XSL•FO
RenderX

Acknowledgments

The authors attest that generative artificial intelligence (AI) tools (eg, ChatGPT) were not used in the generation or writing of the original manuscript. Generative AI was used only to assist with language editing (such as checking spelling and grammatical errors) and polishing during the presubmission revision and the peer-review revision process. All AI-assisted content was reviewed, edited, and validated by the authors for accuracy and appropriateness.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this paper.

Data Availability

The datasets generated or analyzed during this study are available in the GitHub repository [17].

Conflicts of Interest

None declared.

References

1. Bedi S, Liu Y, Orr-Ewing L, Dash D, Koyejo S, Callahan A, et al. Testing and evaluation of health care applications of large language models: a systematic review. JAMA 2025 Jan 28;333(4):319-328. [doi: 10.1001/jama.2024.21700] [Medline: 39405325]
2. Omiye JA, Gui H, Rezaei SJ, Zou J, Daneshjou R. Large language models in medicine: the potentials and pitfalls. Ann Intern Med 2024 Feb;177(2):210-220. [doi: 10.7326/m23-2772] [Medline: 38285984]
3. Ren Y, Luo X, Wang Y, Li H, Zhang H, Li Z, ADVANCED Working Group. Large language models in traditional Chinese medicine: a scoping review. J Evid Based Med 2025 Mar 09;18(1):e12658. [doi: 10.1111/jebm.12658] [Medline: 39651543]
4. Huang T, Lu L, Chen J, Liu L, He J, Jhao Y, et al. TCM-3CEval: a triaxial benchmark for assessing responses from large language models in traditional Chinese medicine. ArXiv Preprint posted online March 10, 2025 [FREE Full text] [doi: 10.48550/arXiv.2503.07041]
5. Earle WB, O'Loughlin L, Smith CC. Revisiting the utility of U.S. medical licensing examination step 3. Ann Intern Med 2023 Jul;176(7):989-990. [doi: 10.7326/M23-0695] [Medline: 37335993]
6. Wu C, Qiu P, Liu J, Gu H, Li N, Zhang Y, et al. Towards evaluating and building versatile large language models for medicine. NPJ Digit Med 2025 Jan 27;8(1):58 [FREE Full text] [doi: 10.1038/s41746-024-01390-4] [Medline: 39865143]
7. Guideline for the diagnosis and treatment of intracerebral hemorrhage with integrated traditional Chinese and Western medicine. Chinese Medical Association. URL: https://tinyurl.com/ykkkdsw3 [accessed 2023-05-21]
8. The yellow emperor's inner canon. Guoxue Meng. URL: https://www.guoxuemeng.com/guoxue/huangdineijing/ [accessed 2025-05-29]
9. Gushiwen classical texts. Guwen Island. URL: https://www.gushiwen.cn/guwen/book.aspx?id=193 [accessed 2025-05-29]
10. Wu Y, Lu T, Wang X, Xu X, Sun B. The correlation between imaging manifestations and traditional Chinese medicine syndrome diagnosis in patients with cerebral infarction: a systematic review of the literature. Adv Clin Med 2025;15(04):1895-1903 [FREE Full text] [doi: 10.12677/acm.2025.1541135]
11. Awesome-medical-dataset/resources/webMedQA. GitHub. URL: https://github.com/openmedlab/Awesome-Medical-Dataset/blob/main/resources/webMedQA.md [accessed 2019-04-09]
12. Li J, Wang X, Wu X, Zhang Z, Xu X, Fu J, et al. Huatuo-26M, a large-scale Chinese medical QA dataset. arXiv Preprint posted online May 2, 2023 [FREE Full text] [doi: 10.18653/v1/2025.findings-naacl.211]
13. Awesome medical dataset. GitHub. URL: https://github.com/openmedlab/Awesome-Medical-Dataset [accessed 2025-05-29]
14. Duan Y, Zhou Q, Li Y, Qin C, Wang Z, Kan H, et al. Research on a traditional Chinese medicine case-based question-answering system integrating large language models and knowledge graphs. Front Med (Lausanne) 2024 Jan 7;11:1512329 [FREE Full text] [doi: 10.3389/fmed.2024.1512329] [Medline: 39839612]
15. MedQA benchmark. Vals AI. URL: https://www.vals.ai/benchmarks/medqa [accessed 2025-02-25]
16. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159-174. [Medline: 843571]
17. LongHL0729/TCM-SED. GitHub. URL: https://github.com/LongHL0729/TCM-SED [accessed 2025-12-03]

Abbreviations

LLM: large language model
TCM: traditional Chinese medicine
TCM-SED: Traditional Chinese Medicine-Stroke Evaluation Dataset

XSL•FO
RenderX

Edited by A Schwartz, M Balcarras; submitted 30.07.25; peer-reviewed by F Lin, NR Mannuru; comments to author 20.10.25; accepted 21.11.25; published 11.12.25.

Please cite as:
Long H, Deng Y, Guo Y, Shen Z, Zhang Y, Bao J, He Y
Large Language Model Evaluation in Traditional Chinese Medicine for Stroke: Quantitative Benchmarking Study
JMIR Form Res 2025;9:e81545
URL: https://formative.jmir.org/2025/1/e81545
doi:10.2196/81545
PMID:

©Hulin Long, Yang Deng, Yaoguang Guo, Zifan Shen, Yuzhu Zhang, Ji Bao, Yang He. Originally published in JMIR Formative Research (https://formative.jmir.org), 11.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Evaluating the Accuracy of Medical Information Generated by ChatGPT and Gemini and Its Alignment With International Clinical Guidelines From the Surviving Sepsis Campaign: Comparative Study

Dina Kutbi[1,2], PharmD, MSc, MBA, PhD; Ehab Abou-Bakr[3,4], MSc, PhD; Hassan Mousa Haidar[2], MSc, PhD

[1]Department of Pharmacy, King Fahad Armed Forces Hospital, Jeddah, Saudi Arabia

[2]Department of Business, Jeddah International College, Jeddah, Saudi Arabia

[3]Department of Computer Science and Information Technology, Jeddah International College, Jeddah, Saudi Arabia

[4]Department of Computer Engineering, The Higher Institute of Engineering and Technology, Elbehera, Egypt

Corresponding Author:
Dina Kutbi, PharmD, MSc, MBA, PhD
Department of Pharmacy
King Fahad Armed Forces Hospital
Al Kurnaysh Br Rd
Jeddah, 23311
Saudi Arabia
Phone: 966 12 232 8888
Email: dina.t.kutbi@gmail.com

Abstract

Background: Assessment of medical information provided by artificial intelligence (AI) chatbots like ChatGPT and Google's Gemini and comparison with international guidelines is a burgeoning area of research. These AI models are increasingly being considered for their potential to support clinical decision-making and patient education. However, their accuracy and reliability in delivering medical information that aligns with established guidelines remain under scrutiny.

Objective: This study aims to assess the accuracy of medical information generated by ChatGPT and Gemini and its alignment with international guidelines for sepsis management.

Methods: ChatGPT and Gemini were asked 18 questions about the Surviving Sepsis Campaign guidelines, and the responses were evaluated by 7 independent intensive care physicians. The responses generated were scored as follows: 3=correct, complete, and accurate; 2=correct but incomplete or inaccurate; and 1=incorrect. This scoring system was chosen to provide a clear and straightforward assessment of the accuracy and completeness of the responses. The Fleiss κ test was used to assess the agreement between evaluators, and the Mann-Whitney *U* test was used to test for the significance of differences between the correct responses generated by ChatGPT and Gemini.

Results: ChatGPT provided 5 (28%) perfect responses, 12 (67%) nearly perfect responses, and 1 (5%) low-quality response, with substantial agreement among the evaluators (Fleiss κ=0.656). Gemini, on the other hand, provided 3 (17%) perfect responses, 14 (78%) nearly perfect responses, and 1 (5%) low-quality response, with moderate agreement among the evaluators (Fleiss κ=0.582). The Mann-Whitney *U* test revealed no statistically significant difference between the two platforms (P=.48).

Conclusions: ChatGPT and Gemini both demonstrated potential for generating medical information. Despite their current limitations, both showed promise as complementary tools in patient education and clinical decision-making. The medical information generated by ChatGPT and Gemini still needs ongoing evaluation regarding its accuracy and alignment with international guidelines in different medical domains, particularly in the sepsis field.

(JMIR Form Res 2025;9:e84251) doi:10.2196/84251

KEYWORDS

sepsis; chatbots; artificial intelligence; AI; ChatGPT; medical information; Gemini; large language model

XSL•FO
RenderX

Introduction

In the medical field, precise diagnoses are essential to managing patients' situations. When making clinical choices, clinicians often depend on their experience and a variety of case scenarios [1]. Sepsis and septic shock are two severe pathological conditions that are defined by the body's reaction to an infection; they may result in organ failure and they have a high death rate. To increase the odds of survival, early diagnosis and prompt treatment are essential. Due to its vague symptoms and the wide range of patient reactions to infections, sepsis is difficult to detect [2]. Improved diagnostic assistance is thus required as the complexity of patients increases, especially those who require referrals to specialized departments like general internal medicine [1].

The assessment of artificial intelligence (AI)–driven models for providing medical information on sepsis and septic shock, as well as comparison of this information with international guidelines, is a burgeoning area of research. AI technologies have shown promise in enhancing diagnostic precision and therapeutic efficacy in sepsis management, yet their alignment with established guidelines like those of the Surviving Sepsis Campaign remains a critical area of evaluation [3]. AI models have been integrated into sepsis management for early detection, subtyping analysis, and prognosis assessment. These systems rely on high-granularity data from intensive care unit (ICU) settings to improve recognition and intervention capabilities, potentially leading to more precise treatment strategies [3]. The performance of these AI models can vary significantly depending on the complexity of the questions and the medical domain being addressed. For instance, ChatGPT's accuracy is lower in responding to more complex medical science questions compared to simpler fracture prevention queries [4].

A previous study comparing ChatGPT (GPT-3.5) and Gemini found that the platforms had similar accuracy rates (71% and 70.5%, respectively) when answering microbiology questions [5]. However, their performance varied across different sections, with Gemini excelling in general microbiology and immunology, while ChatGPT performed better in applied microbiology [5]. In the context of local anesthesia for eye surgery, ChatGPT demonstrated slightly higher accuracy than Gemini (the scores were 4.71 and 4.61). However, traditional patient information leaflets outperformed both AI models in terms of accuracy and completeness [6]. For questions about immune-related adverse events, ChatGPT scored higher than Gemini in accuracy (3.87 vs 3.5) and completeness (3.83 vs 3.46), indicating generally reliable performance in this specific medical domain [7].

Even though several generative AI models have recently been built, ChatGPT remains one of the most popular [8]. Gemini is a user-friendly and efficient AI tool that has transformed ways to access and engage with various types of information by providing advanced, accurate, and relevant responses [9]. While AI platforms like ChatGPT and Gemini show promise in providing medical information, they are not without limitations. Their accuracy and reliability can vary, and they often lack the completeness and adherence to guidelines that traditional sources provide. For example, both ChatGPT and Gemini have been noted to produce factual inaccuracies, including fabricated citations and summaries, which raises concerns about their reliability as standalone sources of clinical information [10,11]. This study, therefore, aims to assess the accuracy of medical information generated by ChatGPT and Gemini and determine how they align with international guidelines for sepsis management.

Methods

Study Design

In this cross-sectional study, the 2021 Surviving Sepsis Guidelines [12] were used as the gold standard reference. The guidelines were rephrased as 18 questions (Multimedia Appendix 1), varying from simple, direct questions to questions about more complex clinical scenarios. The questions encompassed screening, diagnosis, management, and follow-up planning in order to cover all aspects of the guidelines.

Then, ChatGPT (GPT-4o) and Gemini 1.5 were prompted to answer the questions (Multimedia Appendices 2 and 3), and the responses generated by each of these large language models (LLMs) were evaluated for accuracy and alignment with the 2021 Surviving Sepsis Guidelines.

Seven independent intensive care physicians with 15 to 20 years of critical care experience were recruited for the evaluation. The recruited physicians blindly and independently scored the answers generated by ChatGPT and Gemini on a scale of 3 to 1 (3=correct, complete, and accurate; 2=correct but incomplete or inaccurate; 1=incorrect). For blinding, ChatGPT was assessed as "AI-Platform-1" while Gemini was assessed as "AI-Platform-2," without naming the models.

Data Analysis

The data in each group were analyzed with the Fleiss κ test to measure interrater agreement within the group. All responses were assessed independently by 7 physicians using a standardized 3-point scoring rubric (1 to 3). This rubric evaluated (1) scientific accuracy, (2) completeness of the response, and (3) correctness of the clinical reasoning. Responses scoring 3 points were categorized as perfect answers, those scoring 2 points as nearly perfect answers, and those scoring 1 point as low-quality answers. The percentages of perfect answers, nearly perfect answers, and low-quality answers were calculated for each platform. The Mann-Whitney *U* test was then used to assess any significant differences between the two platforms. Analysis used SPSS (version 18.0; IBM Inc) and differences were considered significant at a *P* value of less than .05.

Ethical Considerations

Participating physicians gave their verbal approval and informed consent to participate. However, ethical approval does not apply to this study based on the laws and regulations of the Saudi National Committee of Bioethics guidelines (version 3; 2022).

XSL•FO
RenderX

Results

The scores assigned by the 7 physicians to ChatGPT's and Gemini's responses to the 18 sepsis and septic shock questions are presented in Table 1 and Table 2, respectively. ChatGPT provided 5 (28%) perfect responses to questions about sepsis management, which were fully aligned with guidelines and recommendations. These included questions about source control, nutrition and glycemic control, antibiotic coverage for patients at high risk of multidrug resistance, antimicrobial stewardship in patients with an unconfirmed infectious diagnosis, and antibiotic treatment in patients with community-acquired pneumonia (CAP) regardless of procalcitonin (PCT) level. ChatGPT provided nearly perfect responses to 12 (67%) of the 18 questions; these responses had issues that included missing the time frame for ICU admission, the antibiotics administration time (within the golden hour), or the dosage for corticosteroids; recommending an inappropriate route of administration for neuromuscular blocking agents; not specifying Ringer lactate as the preferred resuscitation fluid of choice; and not mentioning ventilation measures or information about other screening and assessment tools. However, ChatGPT responded to a question about venous thromboembolism (VTE) prophylaxis with a low-quality answer that contradicted the guideline and included incorrect information about the renal adjustment dose of enoxaparin.

Among its 18 responses, Gemini only provided 3 (17%) that were perfect and fully aligned with guideline recommendations about source control and antimicrobial stewardship in patients with an unconfirmed infectious diagnosis and provided comprehensive information about quick sequential organ failure assessment and other assessment tools. Among the 18 responses, Gemini provided 14 nearly perfect answers (78%; Table 3); these answers were missing information on dosage, admission, drug timing, ventilator measures, and nutrition timing approaches. However, in response to a question on antibiotic coverage in patients with a high risk of multidrug resistance, while Gemini's answer included information that aligned with the guideline recommendations, it also suggested irrelevant and incorrect alternative regimen examples, and it also provided 1 overall misleading answer about antibiotic treatment for CAP according to PCT level (Table 2).

The Fleiss κ values for both AI models highlighted the differences in interrater agreement levels and showed that ChatGPT had a score of 0.656, indicating a substantial agreement between the evaluators. However, Gemini had a Fleiss κ score of 0.582, indicating moderate agreement between the evaluators. The Mann-Whitney *U* test returned a *U* statistic of 184 and a *P* value of .48, indicating that there was no statistically significant difference between the average scores of the two AI models.

Table 1. Scores for ChatGPT.

Questions	Evaluator							Average score	Comments
	1	2	3	4	5	6	7		
Q1	3	2	3	3	2	3	3	2.7	Missed information—no time frame (within 6 hours).
Q2	2	2	2	2	2	2	2	2	Inaccurate answer according to guideline recommendations.
Q3	2	2	2	2	2	2	2	2	Missed information—no time frame (within 1 hour).
Q4	3	3	3	3	3	3	3	3	None.
Q5	2	2	2	2	2	2	2	2	Inaccurate answer according to guideline recommendations.
Q6	2	2	2	2	2	2	2	2	Inaccurate answer as the doses were missed.
Q7	3	2	3	3	2	3	3	2.7	Inaccurate answer according to guideline recommendations, but the correct guideline was highlighted in the answer.
Q8	3	3	3	3	2	3	3	2.8	Inaccurate answer according to guideline recommendations.
Q9	3	3	3	3	3	3	3	3	None.
Q10	2	2	3	3	3	2	3	2.6	Inaccurate answer according to guideline recommendations.
Q11	3	2	3	3	2	2	3	2.6	Missed information—posthospital rehabilitation program.
Q12	3	3	3	3	3	3	3	3	None.
Q13	3	2	3	3	2	3	3	2.7	Missed information—some measures and treatment duration (no plateau measures or prone time mentioned).
Q14	2	2	2	2	2	2	2	2	Inaccurate answer according to guideline recommendations.
Q15	3	3	3	3	3	3	3	3	None.
Q16	3	3	3	3	3	3	3	3	None.
Q17	2	2	2	2	2	2	2	2	Inaccurate answer according to guideline recommendations.
Q18	2	2	2	2	2	2	1	1.8	Inaccurate answer according to guideline recommendations. Also, the dose mentioned was not accurate.

XSL•FO
RenderX

Table 2. Scores for Gemini.

Question	Evaluator							Average score	Comments
	1	2	3	4	5	6	7		
Q1	3	2	3	3	2	3	3	2.7	Missed information—no time frame (admission within 6 hours).
Q2	2	2	2	2	2	2	2	2	Inaccurate answer according to guideline recommendations.
Q3	2	2	2	2	2	2	2	2	Missed information—no time frame (within 1 hour).
Q4	3	3	3	3	3	3	3	3	None.
Q5	3	2	3	3	2	2	3	2.6	Missed information—no doses mentioned.
Q6	2	2	2	2	2	2	2	2	Missed information—no doses or time frame mentioned.
Q7	3	2	3	3	2	2	2	2.4	Inaccurate answer according to guideline recommendations.
Q8	2	2	2	2	2	2	2	2	Inaccurate answer according to guideline recommendations.
Q9	3	2	3	3	3	3	3	2.8	Inaccurate answer according to guideline recommendations.
Q10	3	2	3	3	3	2	3	2.7	Inaccurate answer according to guideline recommendations.
Q11	3	2	3	3	2	3	3	2.7	Missed information—post-hospital rehabilitation program.
Q12	3	2	3	3	3	3	1	2.6	Answer is fully aligned with guideline recommendation but contains one incorrect choice of alternative drug regimen mentioned in the explanation (piperacillin/tazobactam is not used for organisms resistant to meropenem, while ceftazidime/avibactam is used).
Q13	2	2	3	3	3	3	3	2.7	Missed information—duration of prone position was not mentioned. Also, the preference of high PEEP[a] over low PEEP.
Q14	3	3	3	3	3	3	3	3	None.
Q15	3	3	3	3	3	3	3	3	None.
Q16	1	1	2	1	1	1	1	1.1	Overall answer was not aligned with guideline recommendation (PCT[b] is not the goal measure for deescalation of antibiotics and CAP[c] treatment should be started regardless of PCT level).
Q17	2	2	2	2	2	2	2	2	Inaccurate answer according to guideline recommendations.
Q18	2	2	2	2	2	2	2	2	Inaccurate answer according to guideline recommendations and missed information (no doses mentioned).

[a]PEEP: positive end-expiratory pressure.

[b]PCT: procalcitonin.

[c]CAP: community-acquired pneumonia.

Table 3. Percentages of perfect, nearly perfect, and low-quality answers to questions (n=18).

Scores	ChatGPT, n (%)	Gemini, n (%)
3 (perfect)	5 (28)	3 (17)
2 to <3 (nearly perfect)	12 (67)	14 (78)
1 to <2 (low quality)	1 (5)	1 (5)

Discussion

This study evaluated the responses generated by ChatGPT and Gemini to 18 questions on septic shock and sepsis and compared the responses with guideline recommendations. Although ChatGPT responded perfectly to 5 questions and Gemini to 3 questions, some responses were inadequate and most were inaccurate. They could not provide accurate information on the timeframe for ICU admission or the administration of antibiotics, and most responses were missing information on the dosage of certain vasoactive drugs, ventilation measures, neuromuscular blocking agents, and the best route of administration according to the guidelines, suggesting that their medical knowledge is limited. These results corroborate the observations of Cascella et al [13], who reported that although these AI technologies may facilitate the acceleration of therapeutic interventions and patient education, there is a chance that they could provide incorrect suggestions.

AI has seen huge advancements in recent years, leading to impressive progress in a variety of forms, including AI chatbots. Because AI chatbots such as ChatGPT and Gemini can respond to almost any inquiry with thorough and human-like replies, they have drawn a large user base in a variety of industries. According to reports, ChatGPT is used for many medical

XSL•FO
RenderX

purposes, including medical teaching, writing, and recording. It has recently been announced that ChatGPT can pass the United States Medical Licensing Examination, which is considered the gold standard, indicating that it may have important medical uses [14].

However, the growing volume and accessibility of medical data pose significant difficulties for physicians. A growing number of physicians and patients are using chatbots to help make medical information easier to understand and more accessible. It is thus critical to assess the accuracy and dependability of chatbot responses, particularly as patients are increasingly depending on these responses to guide their medical decisions [15]. Serious pathological conditions such as sepsis and septic shock are defined by the body's reaction to an infection, which may result in organ failure and a high death rate. To increase the odds of survival, early diagnosis and prompt treatment are essential. Due to its vague symptoms and the wide range of patient reactions to infections, sepsis is difficult to diagnose [2]. According to several previous studies, initial iterations of chatbots provided information that was easy to understand and somewhat accurate. Still, they also gave inadequate, erroneous, or outdated responses [16]. For example, ChatGPT was shown to be generally accurate in answering broad inquiries regarding osteoporosis; however, the replies to queries based on the National Osteoporosis Criteria Group criteria were only 61.3% correct [17].

There are subtle differences between ChatGPT and Gemini when it comes to generating clinical information. Due to its deep learning capabilities, Gemini may perform very well in specialized fields like ophthalmology [18]. However, for wider applications, ChatGPT's adaptability and accessibility make it very useful for general patient education in a variety of medical fields. Its capacity to produce understandable and customized content is especially appreciated in patient education, providing a more dynamic and captivating approach than conventional techniques and some AI alternatives [19]. This observation aligns with the results of this study, which showed that ChatGPT outperformed Gemini in its recommendations for the choice of antibiotics for an ICU patient diagnosed with septic shock suspected of being due to gram-negative infection and for an ICU patient with possible CAP. In contrast, Gemini provided a complete response and alternate assessment tools that differed from the quick sequential organ failure assessment score for diagnosing sepsis.

Furthermore, when asked for recommendations on VTE prophylaxis in critically ill patients, apart from the fact that both AI chatbots failed to recommend proper prophylaxis according to the guidelines, which discourage the use of mechanical VTE plus pharmacological prophylaxis in favor of pharmacological prophylaxis alone, ChatGPT added incorrect information on the renal adjustment dose of enoxaparin (30 mg subcutaneously twice a day instead of 30 mg once daily). These differences between the two AI chatbots might be due to differences in their training databases, design patterns, application algorithms, and updates [20].

The statistical analysis using Fleiss κ indicated substantial agreement for ChatGPT and moderate agreement for Gemini in the quality of their responses, reflecting their relative consistency in adhering to clinical guidelines. Additionally, the Mann-Whitney *U* test supported the absence of a statistically significant difference in their performance. These findings align with previous research that directly compared ChatGPT and Gemini in delivering effective medical information, demonstrating no significant statistical difference between the two platforms [21]. This underlines the comparable potential of both general-purpose and specialized AI systems in supporting clinical decision-making, provided their outputs are critically evaluated and supplemented with human oversight.

Although this study provides a robust set of evidence on the accuracy of ChatGPT and Gemini in generating medical information, its limitations need to be considered. It primarily examined the accuracy and alignment of generalist and commonly used LLM-generated medical information with international clinical guidelines, and it did not include medically fine-tuned and domain-specific LLMs, such as Med-PaLM and BioGPT, which might perform differently on domain-specific tasks.

In conclusion, despite current limitations, AI chatbots like ChatGPT and Gemini show promise as complementary tools in patient education and clinical decision-making, potentially enhancing health care delivery by providing quick access to medical information [22]. The medical information generated by ChatGPT and Gemini still needs continuous evaluation for accuracy, reliability, and alignment with international guidelines in different medical domains. Thus, their use for medical information and treatment should only occur after consultation with a human health care professional. Finally, there is still a need for improvement and further research through training of the AI chatbots using evidence-based sepsis management datasets, including clinical guidelines (such as the Surviving Sepsis Campaign) and real-world case studies.

Acknowledgments

The authors would like to thank the following physicians of the Intensive Care Unit, King Fahad Armed Forces Hospital, Jeddah, Saudi Arabia, for their kind evaluations of the answers generated by ChatGPT and Gemini. Dr. Omnia Hagag, Dr. Mohamed Zohir, Dr. Mohamed Elsakran, Dr. Mohamad Hany, Dr. Mohamed Aid, Dr. Tarek Mokhtar, and Dr. Abdulrahman Nasser. The authors acknowledge the use of ChatGPT for language editing and clarity. Open access funding was provided by Jeddah International College, Saudi Arabia.

XSL•FO
RenderX

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Authors' Contributions

DK conceptualized the overall study. DK, EA-B, and HMH organized and processed the data. DK, EA-B, and HMH conducted the data analysis. HMH secured funding. DK, EA-B, and HMH designed the research methods. DK managed the overall project and coordination. EA-B and HMH validated the results. DK draft the original manuscript. DK, EA-B, and HMH provided critical review and revised the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
2021 Surviving Sepsis Guidelines, rephrased as 18 questions.
[PDF File (Adobe PDF File), 226 KB - formative_v9i1e84251_app1.pdf]

Multimedia Appendix 2
Responses to questions by ChatGPT.
[PDF File (Adobe PDF File), 301 KB - formative_v9i1e84251_app2.pdf]

Multimedia Appendix 3
Responses to questions by Gemini.
[PDF File (Adobe PDF File), 339 KB - formative_v9i1e84251_app3.pdf]

References

1. Hirosawa T, Kawamura R, Harada Y, Mizuta K, Tokumasu K, Kaji Y, et al. ChatGPT-generated differential diagnosis lists for complex case-derived clinical vignettes: diagnostic accuracy evaluation. JMIR Med Inform 2023 Oct 09;11:e48808 [FREE Full text] [doi: 10.2196/48808] [Medline: 37812468]
2. Gorecki G, Tomescu D, Ple L, Panaitescu A, Dragosloveanu, Scheau C, et al. Implications of using artificial intelligence in the diagnosis of sepsis/sepsis shock. Germs 2024 Mar;14(1):77-84. [doi: 10.18683/germs.2024.1419] [Medline: 39169980]
3. Yang J, Hao S, Huang J, Chen T, Liu R, Zhang P, et al. The application of artificial intelligence in the management of sepsis. Med Rev (2021) 2023 Oct;3(5):369-380 [FREE Full text] [doi: 10.1515/mr-2023-0039] [Medline: 38283255]
4. Meng J, Zhang Z, Tang H, Xiao Y, Liu P, Gao S, et al. Evaluation of ChatGPT in providing appropriate fracture prevention recommendations and medical science question responses: a quantitative research. Medicine (Baltimore) 2024 Mar 15;103(11):e37458 [FREE Full text] [doi: 10.1097/MD.0000000000037458] [Medline: 38489735]
5. Ranjan J, Ahmad A, Subudhi M, Kumar A. Assessment of artificial intelligence platforms with regard to medical microbiology knowledge: an analysis of ChatGPT and Gemini. Cureus 2024 May;16(5):e60675 [FREE Full text] [doi: 10.7759/cureus.60675] [Medline: 38770053]
6. Gondode P, Duggal S, Garg N, Lohakare P, Jakhar J, Bharti S, et al. Comparative analysis of accuracy, readability, sentiment, and actionability: artificial intelligence chatbots (ChatGPT and Google Gemini) versus traditional patient information leaflets for local anesthesia in eye surgery. Br Ir Orthopt J 2024;20(1):183-192 [FREE Full text] [doi: 10.22599/bioj.377] [Medline: 39183761]
7. Burnette H, Pabani A, von Itzstein MS, Switzer B, Fan R, Ye F, et al. Use of artificial intelligence chatbots in clinical management of immune-related adverse events. J Immunother Cancer 2024 May 30;12(5):e008599 [FREE Full text] [doi: 10.1136/jitc-2023-008599] [Medline: 38816231]
8. Ray PP. ChatGPT: a comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet Things Cyber-Phys Syst 2023;3:121-154. [doi: 10.1016/j.iotcps.2023.04.003]
9. Imran M, Almusharraf N. Google Gemini as a next generation AI educational tool: a review of emerging educational technology. Smart Learn Environ 2024 May 23;11(1):22. [doi: 10.1186/s40561-024-00310-z]
10. Caspi R, Karp PD. An evaluation of ChatGPT and Bard (Gemini) in the context of biological knowledge retrieval. Access Microbiol 2024;6(6):0793. [doi: 10.1099/acmi.0.000790.v3] [Medline: 39045247]
11. Labrague LJ. Utilizing artificial intelligence-based tools for addressing clinical queries: ChatGPT versus Google Gemini. J Nurs Educ 2024 Aug;63(8):556-559. [doi: 10.3928/01484834-20240426-01] [Medline: 39120501]
12. Evans L, Rhodes A, Alhazzani W, Antonelli M, Coopersmith CM, French C, et al. Surviving Sepsis Campaign: International Guidelines for Management of Sepsis and Septic Shock 2021. Crit Care Med 2021 Nov 01;49(11):e1063-e1143. [doi: 10.1097/CCM.0000000000005337] [Medline: 34605781]

XSL•FO
RenderX

13. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. J Med Syst 2023 Mar 04;47(1):33 [FREE Full text] [doi: 10.1007/s10916-023-01925-4] [Medline: 36869927]
14. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
15. Pan A, Musheyev D, Bockelman D, Loeb S, Kabarriti AE. Assessment of artificial intelligence chatbot responses to top searched queries about cancer. JAMA Oncol 2023 Oct 01;9(10):1437-1440. [doi: 10.1001/jamaoncol.2023.2947] [Medline: 37615960]
16. Chen S, Kann BH, Foote MB, Aerts HJWL, Savova GK, Mak RH, et al. Use of artificial intelligence chatbots for cancer treatment information. JAMA Oncol 2023 Oct 01;9(10):1459-1462 [FREE Full text] [doi: 10.1001/jamaoncol.2023.2954] [Medline: 37615976]
17. Cinar C. Analyzing the performance of ChatGPT about osteoporosis. Cureus 2023 Sep;15(9):e45890 [FREE Full text] [doi: 10.7759/cureus.45890] [Medline: 37885522]
18. Masalkhi M, Ong J, Waisberg E, Lee AG. Google DeepMind's gemini AI versus ChatGPT: a comparative analysis in ophthalmology. Eye (Lond) 2024 Jun;38(8):1412-1417 [FREE Full text] [doi: 10.1038/s41433-024-02958-w] [Medline: 38355668]
19. Sallam M, Salim N, Barakat M, Al-Tammemi A. ChatGPT applications in medical, dental, pharmacy, and public health education: a descriptive study highlighting the advantages and limitations. Narra J 2023 Apr;3(1):e103 [FREE Full text] [doi: 10.52225/narra.v3i1.103] [Medline: 38450035]
20. Tong L, Zhang C, Liu R, Yang J, Sun Z. Comparative performance analysis of large language models: ChatGPT-3.5, ChatGPT-4 and Google Gemini in glucocorticoid-induced osteoporosis. J Orthop Surg Res 2024 Sep 18;19(1):574 [FREE Full text] [doi: 10.1186/s13018-024-04996-2] [Medline: 39289734]
21. Dursun D, Bilici Geçer R. Can artificial intelligence models serve as patient information consultants in orthodontics? BMC Med Inform Decis Mak 2024 Jul 29;24(1):211 [FREE Full text] [doi: 10.1186/s12911-024-02619-8] [Medline: 39075513]
22. Souza LLD, Fonseca FP, Martins MD, de Almeida OP, Pontes HAR, Coracin FL, et al. ChatGPT and medicine: a potential threat to science or a step towards the future? J Med Artif Intell 2023 Oct;6:19-19. [doi: 10.21037/jmai-23-70]

Abbreviations

AI: artificial intelligence
CAP: community-acquired pneumonia
LLM: large language model
PCT: procalcitonin
VTE: venous thromboembolism

Edited by A Stone; submitted 16.09.25; peer-reviewed by H Zhang; comments to author 03.11.25; accepted 18.11.25; published 17.12.25.

Please cite as:
Kutbi D, Abou-Bakr E, Haidar HM
Evaluating the Accuracy of Medical Information Generated by ChatGPT and Gemini and Its Alignment With International Clinical Guidelines From the Surviving Sepsis Campaign: Comparative Study
JMIR Form Res 2025;9:e84251
URL: https://formative.jmir.org/2025/1/e84251
doi:10.2196/84251
PMID:

©Dina Kutbi, Ehab Abou-Bakr, Hassan Mousa Haidar. Originally published in JMIR Formative Research (https://formative.jmir.org), 17.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Feasibility and Usability of an mHealth App (mLab+) to Guide Users Through HIV and Syphilis Self-Testing: Pilot Randomized Controlled Trial

Maeve Brin[1], BA; Thomas F Scherr[2], PhD; Janejira Chaiyasit[3], DNP; Jianfang Liu[1], PhD; Maura Abbott[3], PhD; Robert Garofalo[4,5], MD; Lisa M Kuhns[4,5], PhD; Tess Sky[3]; Ian Esliker[1], BS; Rebecca Schnall[1,6], PhD

[1]School of Nursing, Columbia University, 560 W 168 Street, New York, NY, United States
[2]Vanderbilt University, Nashville, United States
[3]MGH Institute of Health Professions School of Nursing, Washington DC, United States
[4]Division of Adolescent & Young Adult Medicine, Ann & Robert H. Lurie Children's Hospital of Chicago, Chicago, IL, United States
[5]Department of Pediatrics, Feinberg School of Medicine, Northwestern University, Chicago, IL, United States
[6]Mailman School of Public Health, Columbia University, New York, NY, United States

Corresponding Author:
Rebecca Schnall, PhD
School of Nursing, Columbia University, 560 W 168 Street, New York, NY, United States

Abstract

Background: HIV self-testing is an important strategy in the US Ending the HIV Epidemic initiative. To facilitate uptake of self-testing, we developed the mLab app, which complements existing self-test options to support the potential for higher uptake of the HIV self-test. Syphilis, a sexually transmitted infection with currently rising prevalence and overlap in risk profiles with HIV, could similarly benefit from the advantages of companion diagnostic mobile apps such as mLab. Due to the success of the mLab app in promoting HIV self-testing during a randomized controlled trial and the scientific evidence of need for at-home syphilis testing, our team developed the mLab+ app, which supports both HIV and syphilis testing through an image processing algorithm that incorporates a duplex HIV and syphilis point-of-care test.

Objective: We conducted a pilot study to assess the feasibility and usability of the mLab+ app for HIV and syphilis testing.

Methods: We recruited participants who were assigned male sex at birth and reported sex with another man. Participants came to the Nurse Practitioner Group clinic for baseline and follow-up visits. Participants rated the usability of the app using the Health Information Technology Usability Evaluation Scale and the Post-Study System Usability Questionnaire at their 3-month follow-up visit. The primary outcome was the number of participants who were able to self-administer the DPP HIV-Syphilis test with the assistance of the mLab+ app. Feasibility was measured through recruitment pace, retention over 3 months, app usability, and paradata.

Results: Of the 20 participants, 19 identified as male and 1 identified as nonbinary. Most participants (n=16) were able to complete the DPP HIV-Syphilis test with facilitation support from the mLab+ app. The average duration of an app session, from after authentication until log-out or abandonment, was 30 minutes and 33 seconds (SD 21 minutes and 40 seconds). Apart from the 27% (13/48) of sessions that were 5 minutes or less, the distribution of session durations was approximately normal. Users spent the longest time viewing testing screens (ie, timer screens, initial testing screen, test guided walkthroughs, test results, and picture and result upload). The overall mean scores on the Post-Study System Usability Questionnaire (2.65, SD 1.06) and Health Information Technology Usability Evaluation Scale (3.62, SD 1.07) indicated medium to high usability. The retention rate for the 3-month trial was 80% (16/20).

Conclusions: The findings support the use of the mLab+ app as a tool for assisting consumers in self-testing for HIV and syphilis. The limitations of the study design warrant further examination outside of clinic settings to better understand the utility of these tools for improving consumer health outcomes.

Trial Registration: ClinicalTrials.gov NCT06059443; https://clinicaltrials.gov/study/NCT06059443

(JMIR Form Res 2025;9:e72955) doi:10.2196/72955

KEYWORDS

HIV; syphilis; self-test; mobile health; mHealth; sexually transmitted infections; men who have sex with men; MSM

XSL•FO
RenderX

Introduction

HIV self-testing is recommended for diverse populations worldwide [1] and is considered an important strategy in the US Ending the HIV Epidemic initiative [2,3]. HIV self-tests can facilitate access to antiretroviral therapy, HIV preexposure prophylaxis (PrEP), and other prevention services [4,5]. Aligned with this evidence, and to facilitate uptake of HIV self-testing, our study team developed the mLab app through user-centered design including a youth advisory board and usability testing with informatics experts and end users [6]. The mLab app provides HIV prevention information, notification reminders for testing, step-by-step instructions for using OraQuick HIV tests, and an image upload function so individuals can send an image of their OraQuick HIV test for automated analysis. Individuals report their visual result of the OraQuick test on the app; the app then interprets the results of the OraQuick tests independent of the visually reported result. If an individual tests positive for HIV on the OraQuick HIV test, either via self-reported visual result or through mLab's image processing functionality, the study team is notified and coordinates follow-up confirmatory testing with the participant.

The mLab app has advantages over stand-alone self-test options and supports the potential for higher uptake of the HIV self-test [7]. It addresses many of the current barriers to self-testing kits to overcome ambiguous test interpretations and provide immediate result reporting. Further, the mLab user interface promotes a holistic testing experience as it provides step-by-step error checking with clear picture instructions for completing the self-test. Although rapid tests such as the OraQuick seem simple to interpret, weak positive bands and weak control lines sometimes make it difficult for users to accurately interpret the test results, which we found in our own research on the OraQuick home test [8]. Another study found a correct interpretation rate of 70% for weak positives and 59% for invalid tests [9]. Beyond support for HIV testing, the mLab app provides information for those who test positive for HIV and educates users on the importance of follow-up testing and prevention services for those who test negative to promote linkage to care.

Implementation of HIV self-testing also provides an opportunity for self-testing of another sexually transmitted infection (STI) that has seen recent increases in prevalence in the United States: syphilis. From 2018 to 2022, there was an 80% increase in reported syphilis cases in the United States [10]. The rate of syphilis acquisition has increased steadily since 2000, especially among men, and men who have sex with men (MSM) constitute a disproportionate number of cases (36% of all cases in 2021 were among MSM) [11,12]. Moreover, the rate of syphilis among MSM is 100 times higher than that in men who have sex with women and even higher among young MSM [13,14]. In 2023, the highest rates of syphilis acquisition in male individuals were among men between the ages of 20 and 39 years (with rates per 100,000 ranging between 44.5% and 62% within each age subdivision) [15]. Additionally, disproportionate rates of HIV and STIs among transgender women and nonbinary people suggest the need for greater access to self-testing among these populations as well [16-20].

Importantly, syphilis makes it easier to both acquire and transmit HIV, and approximately half of MSM who have syphilis are coinfected with HIV [14,21]. The biological nature of syphilis, specifically syphilitic ulcer proliferation, facilitates transmission of HIV [22]. In addition, syphilis infections among people living with HIV are associated with higher HIV viral load and lower CD4 cell counts and, therefore, worsen the severity of HIV-related symptoms [23].

Prior studies have shown the need for at-home testing options for MSM and transgender women, not only for HIV but for STIs as well due to testing frequency recommendations and concerns about privacy and confidentiality [20,24,25]. Considering the higher level of coinfection with syphilis among people living with HIV, there is a need for increased concurrent testing for HIV and syphilis. One study showed that uptake of concurrent HIV and syphilis testing in a health care setting was highest among people between the ages of 25 and 34 years, suggesting potential acceptability of concurrent testing among this age group [26]. Moreover, the study acknowledged the social barriers to concurrent HIV and syphilis testing acceptance within the health care setting and called for the development of interventions that support decision-making between emergency department staff and high-risk patients [26].

Due to the success of the mLab app in a randomized controlled trial (RCT) and the scientific evidence of need for at-home syphilis point-of-care (POC) testing among young MSM and young transgender women, our team developed the mLab+ app (Figure 1), which supports HIV and syphilis testing at the POC. It facilitates the use of a duplex HIV and syphilis POC test and has an image processing algorithm that automatically and objectively interprets the diagnostic test. We conducted a rigorous pilot study to assess the feasibility and usability of the mLab+ app. The primary objective of this study was to evaluate the use of the mLab+ app in a clinical setting with 20 participants assigned male sex at birth who reported sex with a man and identified as male, transgender, or nonbinary. The aim of this study was to demonstrate the feasibility and usability of the mLab+ app to support users through HIV and syphilis self-testing.

XSL•FO
RenderX

Figure 1. Screenshots from the mLab+ app: (A) authentication, (B) home or landing page, (C) step-by-step guided walkthrough of taking the test, (D) in-app timer with mindfulness exercises, (E) automatic timer-based navigation and upload of a photo of the test, (F) automated image processing review of photograph quality, (G) upload of user-reported visual result, and (H) test result history. This figure does not include information on any participants enrolled in the study.

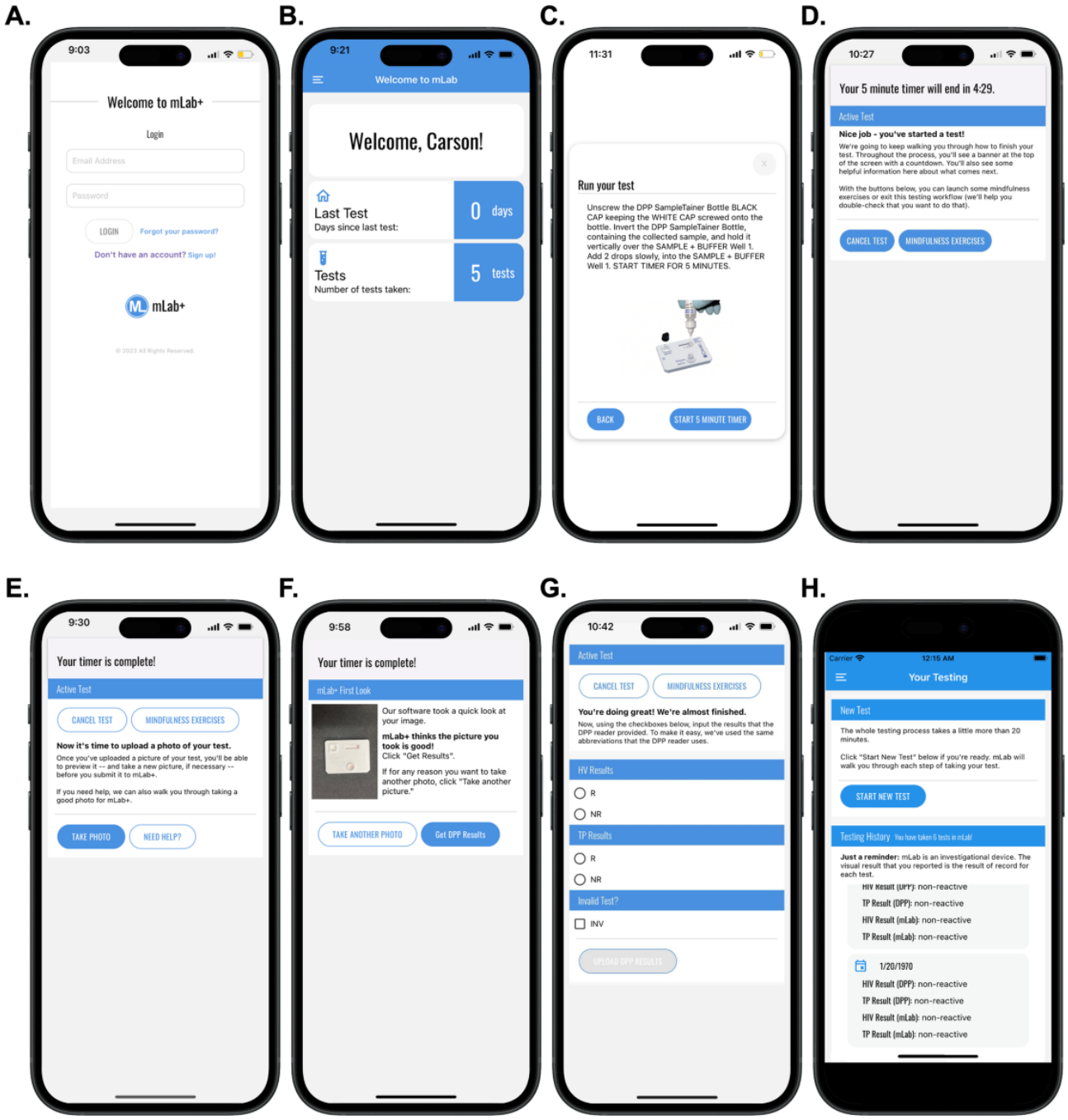

Methods

Study Design

We conducted a 3-month pilot RCT with 40 participants from October 2023 to June 2024. Participants in both arms received STI counseling and condoms (baseline) and completed a survey (baseline and 3-month follow-up). Participants who were randomized to the intervention arm (mLab+ app) were given app access and took an HIV and syphilis test while using the app at both their baseline and 3-month follow-up appointments. This paper only reports the findings of the intervention arm.

Recruitment Procedures

We used a multimodal recruitment strategy. We recruited through sexual networking apps (eg, Grindr and Scruff) and through community partners. Online advertisements included images of diverse MSM and were targeted to MSM through dating apps. A public link to a screener form was embedded within the advertisement image so that participants could fill out a brief survey to determine eligibility prior to being contacted by the research team. Regarding recruitment through community partners, research staff visited several clinics, organizations, gay bars, and shops throughout New York City to post physical flyers. Physical flyers included the study team's

XSL•FO
RenderX

phone number and a QR code that directed interested participants to the public screener link. The eligibility criteria can be found in Textbox 1. Eligible participants came to the Columbia University Irving Medical Center Nurse Practitioner Group (NPG) clinic for their study visit. Because all study procedures took place in person, participants were all from the New York City metropolitan area. They had to be able to travel to the study office for visits to be enrolled in the study.

Textbox 1. Eligibility criteria for participation in the mLab+ app study.

Inclusion criteria

- Between 18 and 39 years of age
- Assigned male sex at birth
- Ability to understand and read English
- Self-identifying as any race or ethnicity
- Substantial risk of acquiring HIV per Centers for Disease Control and Prevention guidance (eg, sexual partner with HIV, recent bacterial sexually transmitted disease, high number of sexual partners, history of inconsistent or no condom use, or sex work)
- Report of having sex with a man or men
- Smartphone ownership
- Self-report of being HIV negative or having an unknown status
- Self-report of having tested negative for syphilis or having an unknown status
- Not having been tested for HIV or syphilis in the previous 3 months
- Understanding the limitations of the duplex lateral flow test and the mLab+ app (eg, a confirmatory test is needed, and the self-test must be performed in the presence of a qualified clinician)

Exclusion criteria

- Persons with a known diagnosis of HIV or syphilis
- Persons for whom the investigators determine that participation may be detrimental to the participant or to the study (eg, those with severe cognitive deficit)
- Persons diagnosed with systemic lupus erythematosus as their medical condition could affect the results of the DPP HIV-Syphilis test
- Persons unable or unwilling to provide consent for study participation

Ethical Considerations

All study activities were approved by the Columbia University Institutional Review Board under approval AAAU3380 prior to the recruitment or enrollment of study participants. This study was registered on ClinicalTrials.gov (NCT06059443). Study participants signed electronic informed consent forms through REDCap (Research Electronic Data Capture; Vanderbilt University) that stated the procedures for both the intervention and control arms of the study. The consent form outlined analysis procedures following the study's completion, including the individuals and agencies who had permission to look through research records, the deidentification of all pieces of information that may identify them, and the use of their deidentified data for future research purposes. The consent form also included a statement of confidentiality. This detailed authorization for Columbia University Irving Medical Center, the New York-Presbyterian Hospital, and the principal investigator and her staff to use and disclose protected health information (PHI) in connection with the research study; an explanation of what PHI consists of; and for what purposes the PHI may be used or disclosed. Furthermore, it stated that we obtained a certificate of confidentiality from the National Institutes of Health. Participants were compensated with US $40 for their baseline appointment and US $60 for their 3-month follow-up appointment.

Study Procedures

Baseline Visit

At the baseline visit, participants received HIV and STI risk reduction counseling, a box of condoms, a PrEP assessment, and referral information for clinics that provide PrEP. Participants then completed a survey via REDCap, a web-based survey software, that included questions on demographics; sexual risk behaviors, including number of sexual partners (ie, number of people they engaged in anal or oral sex with) and condomless anal intercourse within the previous 3 months; HIV and STI testing history; opinions regarding HIV and STI testing; postexposure prophylaxis and PrEP use and adherence; drug and alcohol use [27]; and HIV risk index [28]. Study staff were available to answer any questions while the participants were taking the survey.

Following completion of the survey, participants were instructed on how to log into and use the mLab+ app (Figure 1) by study staff. Next, participants verbally completed a DPP HIV-Syphilis testing assessment to ensure that they were aware of the functions and limitations of the test.

Participants completed the DPP HIV-Syphilis test guided by the mLab+ app with step-by-step images and instructions on how to complete the test. The DPP HIV-Syphilis assay, assigned a moderate level of complexity, is not yet approved by the Food

XSL•FO
RenderX

and Drug Administration (FDA) for self-administration. As a result, participants completed their self-tests at the NPG clinic under the supervision of a trained clinician who was qualified to use the device according to the approved instructions for use. If the participant was unable to complete the test on their own, the clinician helped administer and interpret the test.

Testing using a DPP HIV-Syphilis kit required that the participant obtain a small drop of blood using a sterile lancet and sample loop and insert it into the SampleTainer bottle that was prefilled with a buffer solution. Participants then applied 2 drops of the sample and buffer from the SampleTainer bottle to the test. After a 5-minute wait, 4 drops of the DPP running buffer solution were added to the test by the participant. After a 10-minute waiting period, the test was ready for analysis first by the mLab+ app.

Participants took a picture of their diagnostic test on the app, which was then processed through automated image analysis software. Next, participants attached the included DPP Micro Reader, a small optical device provided with the testing kits, to the cassette to interpret the test results as described in the manufacturer's instructions. Once the DPP Micro Reader finished running the test, participants reported the DPP Micro Reader results on the app and then were directed to a result summary screen on their mobile devices. The test took approximately 20 minutes to complete. Once complete, the result summary simultaneously displayed the results of the HIV and syphilis test.

After finishing the test, a study team member discussed the test results with the participants. Test results were securely transmitted and stored in a secure database. Participants with a positive test result were referred to care and offered confirmatory testing at the clinic.

3-Month Follow-Up Visit

At the 3-month follow-up visit, participants returned to the NPG clinic for a survey and testing using the DPP HIV-Syphilis assay guided by the mLab+ app. The 3-month follow-up survey included the same measures as the baseline survey with the addition of usability assessments. The Health Information Technology Usability Evaluation Scale (Health-ITUES) and Post-Study System Usability Questionnaire (PSSUQ) were used to evaluate the mLab+ app in its ability to facilitate access to health information and guide users through testing. The Health-ITUES is a 20-item customizable instrument with 4 subscales: quality of work life (items 1‑3), perceived usefulness (items 4-12), perceived ease of use (items 13-17), and user control (items 18‑20), all scored from 1 ("strongly disagree") to 5 ("strongly agree") [29]. The PSSUQ is a 16-item instrument divided into 3 subscales: system usefulness (items 1-6), information quality (items 7-12), and interface quality (items 13-15), all scored from 1 ("strongly agree") to 7 ("strongly disagree") [30]. Both scales assess user satisfaction and perceptions of the system's usability and have strong evidence of reliability and content and construct validity [29,30]. Participants self-administered the DPP HIV-Syphilis test at the clinic in the presence of a clinician following the same procedures as for baseline testing.

Study Outcomes and Measures

The primary outcome was the number of participants who were able to self-administer the DPP HIV-Syphilis test using the mLab+ app. A successful self-test was defined as having completed the DPP HIV-Syphilis test by themselves and having accurately identified and interpreted their HIV and syphilis statuses as determined by the supervising clinician. If participants were unable to self-administer the test, were apprehensive, or incorrectly used the DPP HIV-Syphilis test as per the instructions for use, the clinician intervened and helped the participants with using the test. If the clinician, rather than the participant, administered the DPP HIV-Syphilis test, the test was not recorded as a successful self-test. A research staff member from Columbia University was always present with the study participants. The primary outcome was recorded by staff running the visit, who indicated through REDCap whether the participants successfully completed the testing procedures on the app without the help of a clinician.

In addition, we measured the feasibility of this tool through several approaches: recruitment pace, retention over 3 months, usability of the app, and paradata. Paradata allowed us to quantify participants' use of the app and are considered "free" in that they are collected in the background of app use and do not require any additional effort from the user [31]. The following paradata were collected: (1) unique participant code, (2) page accessed, (3) action performed on the page, and (4) time stamp. From these data, derivative metrics were calculated for each user and in aggregate, including duration on each page, page progression through the app, time from log-in to result, and total time from log-in to log-out.

Results

Sample Characteristics

Of a total of 707 people who filled out an eligibility screener, 122 (17.3%) were screened as eligible, and 40 (5.7%) were enrolled and randomized to the control (n=20, 50%) or intervention (n=20, 50%) arms (Figure 2).

In total, 95% (19/20) of the participants who used the mLab+ app identified as male, and 5% (1/20) identified as nonbinary. Age ranged from 21 to 35 years, with an average of 28 (SD 3.94) years. Participants reported their race as Asian (5/20, 25%), Black (6/20, 30%), White (8/20, 40%), and other (1/20, 5%) and their ethnicity as not Hispanic or Latino (14/20, 70%), Dominican (2/20, 10%), Mexican (1/20, 5%), Puerto Rican (2/20, 10%), and another Hispanic identity (1/20, 5%).

XSL•FO
RenderX

Figure 2. CONSORT (Consolidated Standards of Reporting Trials) diagram. This figure includes the number of total participants assessed for eligibility, randomized, and enrolled. NYC: New York City.

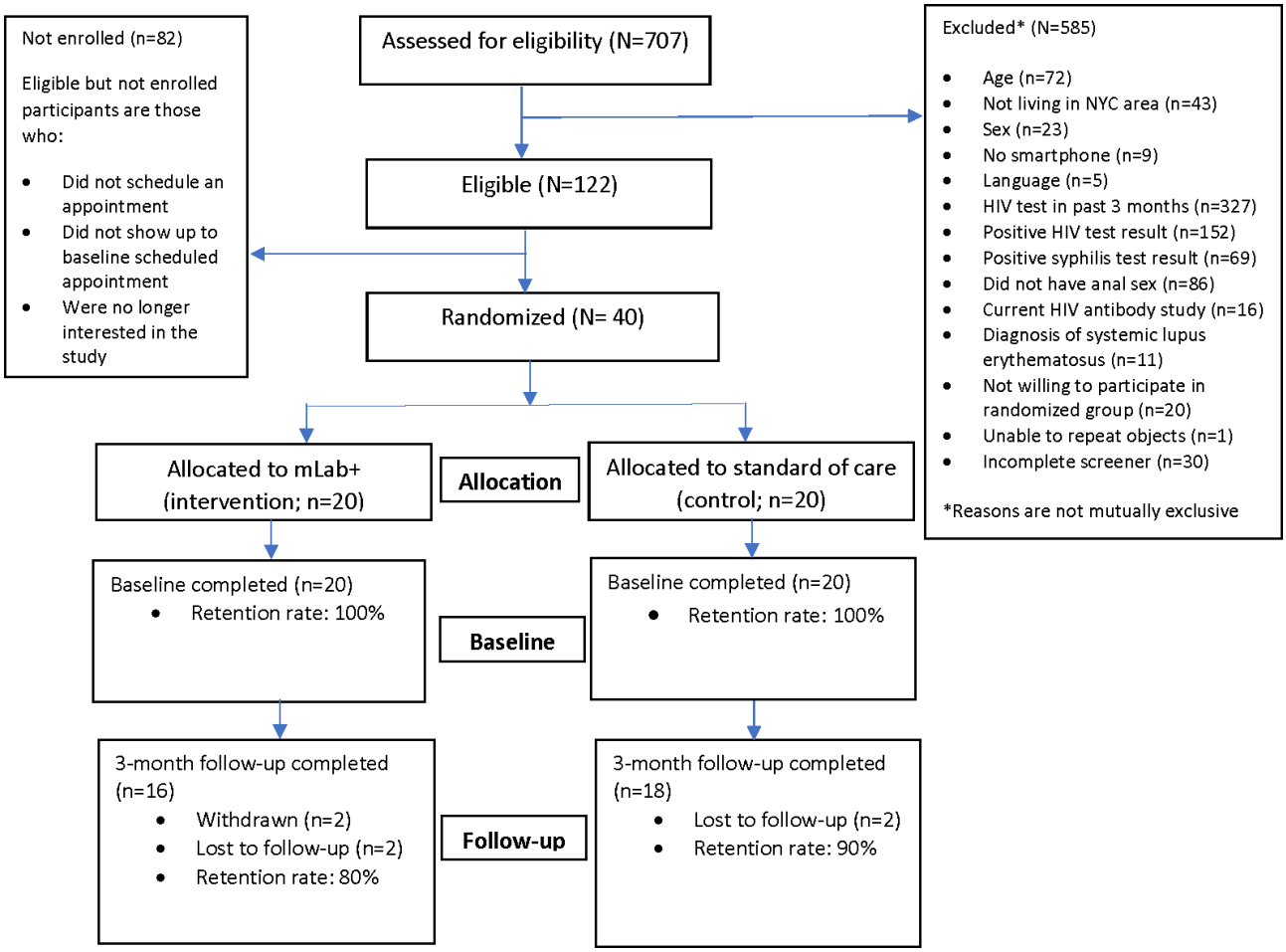

Primary Outcome

Most participants (16/20, 80%) were able to complete the test on their own at both time points. Table 1 shows the number of participants who completed testing independently (ie, "yes") and with the help of a clinician (ie, "no"). Table 2 shows the test results for both HIV and syphilis at each time point.

Table . Participants who conducted the HIV and syphilis test on their own at baseline and the 3-month follow-up while using the mLab+ app at the clinic.

	Participants, n/N (%)
Baseline	
Yes	16/20 (80)
No	4/20 (20)
3-month follow-up[a]	
Yes	12/15 (80)
No	3/15 (20)

[a]One of the participants who was retained at the 3-month follow-up visit did not complete testing at this time point.

XSL•FO
RenderX

Table . Reactive and nonreactive HIV and syphilis test results among young men who have sex with men on the mLab+ app study at both study time points (baseline and 3-month follow-up).

	Participants, n/N (%)
Baseline	
Nonreactive for HIV and nonreactive for syphilis treponemal test	15/20 (75)
Nonreactive for HIV and reactive for syphilis treponemal test	3/20 (15)
Reactive for HIV and nonreactive for syphilis treponemal test	1/20 (5)
Reactive for HIV and reactive for syphilis treponemal test	1/20 (5)
3-month follow-up[a]	
Nonreactive for HIV and nonreactive for syphilis treponemal test	12/15 (80)
Nonreactive for HIV and reactive for syphilis treponemal test	3/15 (20)

[a]One participant did not complete testing at the follow-up visit.

Linkage to Care

All participants who tested reactive for HIV or syphilis at baseline (5/20, 25%) or follow-up (3/20, 15%) were referred to care. Only 40% (2/5) of those participants reported linkage to care after their baseline appointment. The others were withdrawn due to a previous HIV diagnosis (1/5, 20%), were scheduled for testing and did not show up (1/5, 20%), or did not respond to follow-up calls about linkage to care (1/5, 20%).

Paradata

Paradata were collected during baseline and follow-up visits. The average duration of a session, from after authentication until log-out or abandonment, was 30 minutes and 33 seconds (SD 21 minutes and 40 seconds; median 36 minutes and 2 seconds; IQR 0 minutes 42 seconds-44 minutes and 15 seconds). Apart from the 27% (13/48) of sessions that were 5 minutes or less, the distribution of session durations was approximately normal (Figure 3). The distribution of duration viewing each screen (Figure 4) shows that users spent the longest time viewing testing screens (ie, timer screens, the initial testing screen, the test guided walkthroughs, test results, and picture and result upload). Auxiliary screens (ie, privacy policy and help) were typically only briefly visited. Notably, the 2 timer screens (“Timer: running buffer” and “Timer: sample”) both had very narrow distributions, indicating that time spent viewing the timer screens was brief, and the automated timers functioned as expected, moving participants along the testing workflow at the completion of each timer.

XSL•FO
RenderX

Figure 3. Distribution of session duration across all users enrolled in the mLab+ app study between their first visit and the 3-month follow-up visit.

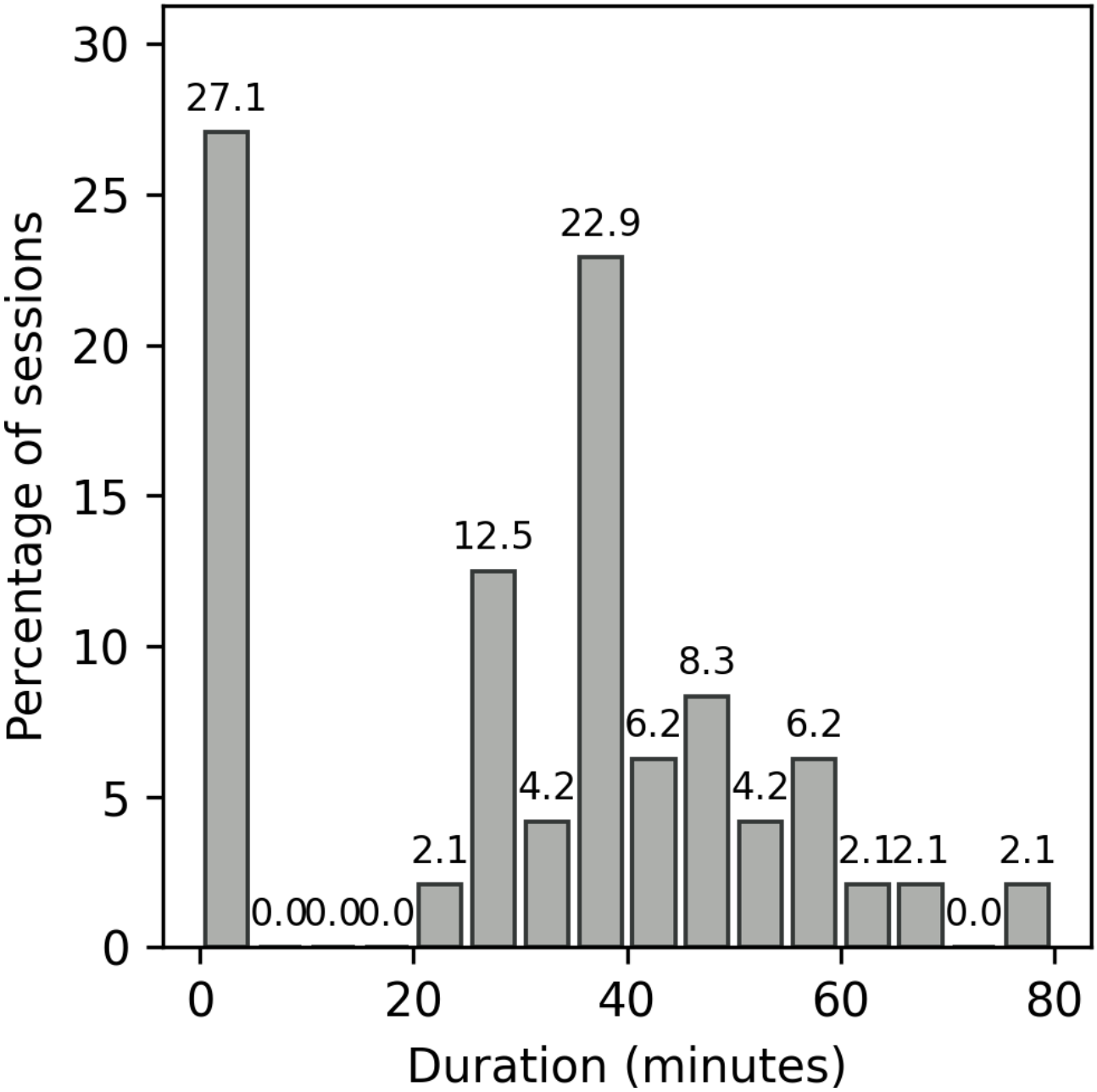

XSL•FO
RenderX

Figure 4. Distribution of time spent on each screen on the mLab+ app across all users during the 3-month study period (outliers greater than 1.5 times the IQR were removed for visualization purposes, but statistical calculations were not removed).

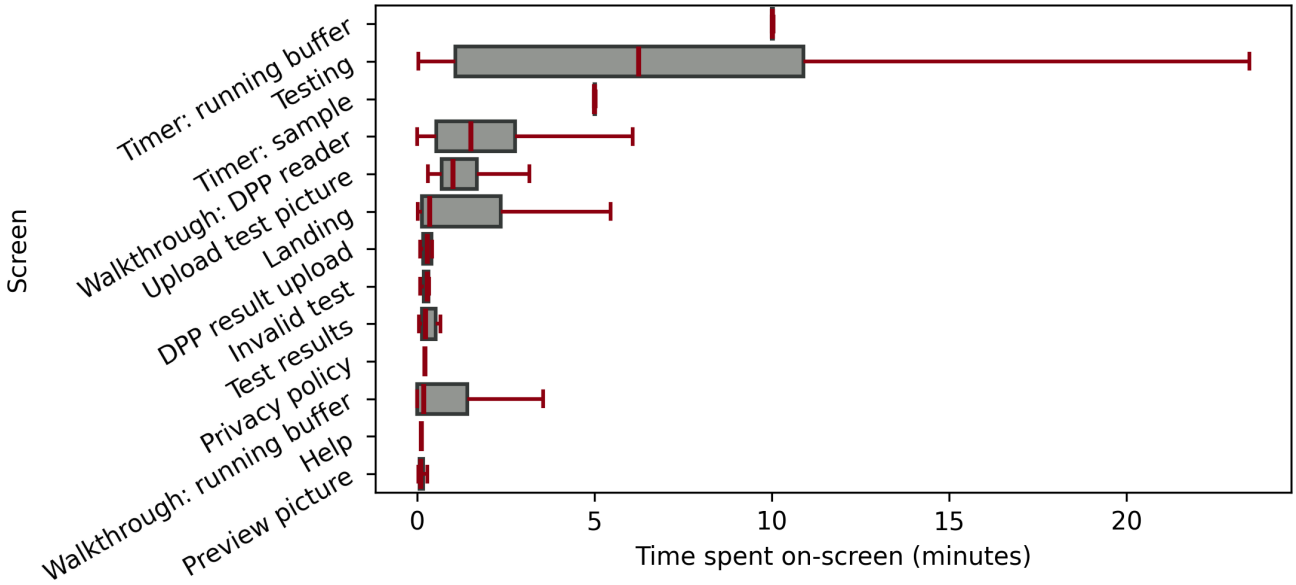

As test taking was the primary objective of app use, it is informative to look at paradata related to the testing pathway. The workflow conversion funnel (Table 3) shows the percentage of time during which a screen was visited in an individual session. These data indicate that 77% (37/48) of sessions resulted in a user visiting the screen where a test could be initiated ("Testing") and 71% (34/48) of sessions started a test. There was no drop-off in conversion until the "Preview picture" screen, which 67% (32/48) of the sessions reached. This conversion remained constant using the DPP Micro Reader and only decreased to 63% (30/48) of sessions that ended with a test result.

Of the sessions that resulted in a completed test, the distribution of the time it took from log-in to obtaining the test results (Figure 5) was approximately normal (mean 39 minutes and 6 seconds, SD 10 minutes and 1 second; median 38 minutes and 14 seconds).

Table . Conversion rate for each mLab+ app screen, that is, the percentage of time each screen was visited in an individual session across all users enrolled in the mLab+ app study at baseline and the 3-month follow-up.

Step	Screen	Screen sessions (n=48)	Conversion rate (%)
1	Landing	47	97.92
2	Testing	37	77.08
3	Timer: sample	34	70.83
4	Walkthrough: running buffer	34	70.83
5	Timer: running buffer	34	70.83
6	Upload test picture	34	70.83
7	Preview picture	32	66.67
8	Walkthrough: DPP Micro Reader	32	66.67
9	DPP result upload	32	66.67
10	Test results	30	62.50

XSL•FO
RenderX

Figure 5. The distribution of time to reach the HIV and syphilis result screen from the first screen across all mLab+ app participant sessions that resulted in a completed HIV and syphilis test at both the baseline and 3-month follow-up visits.

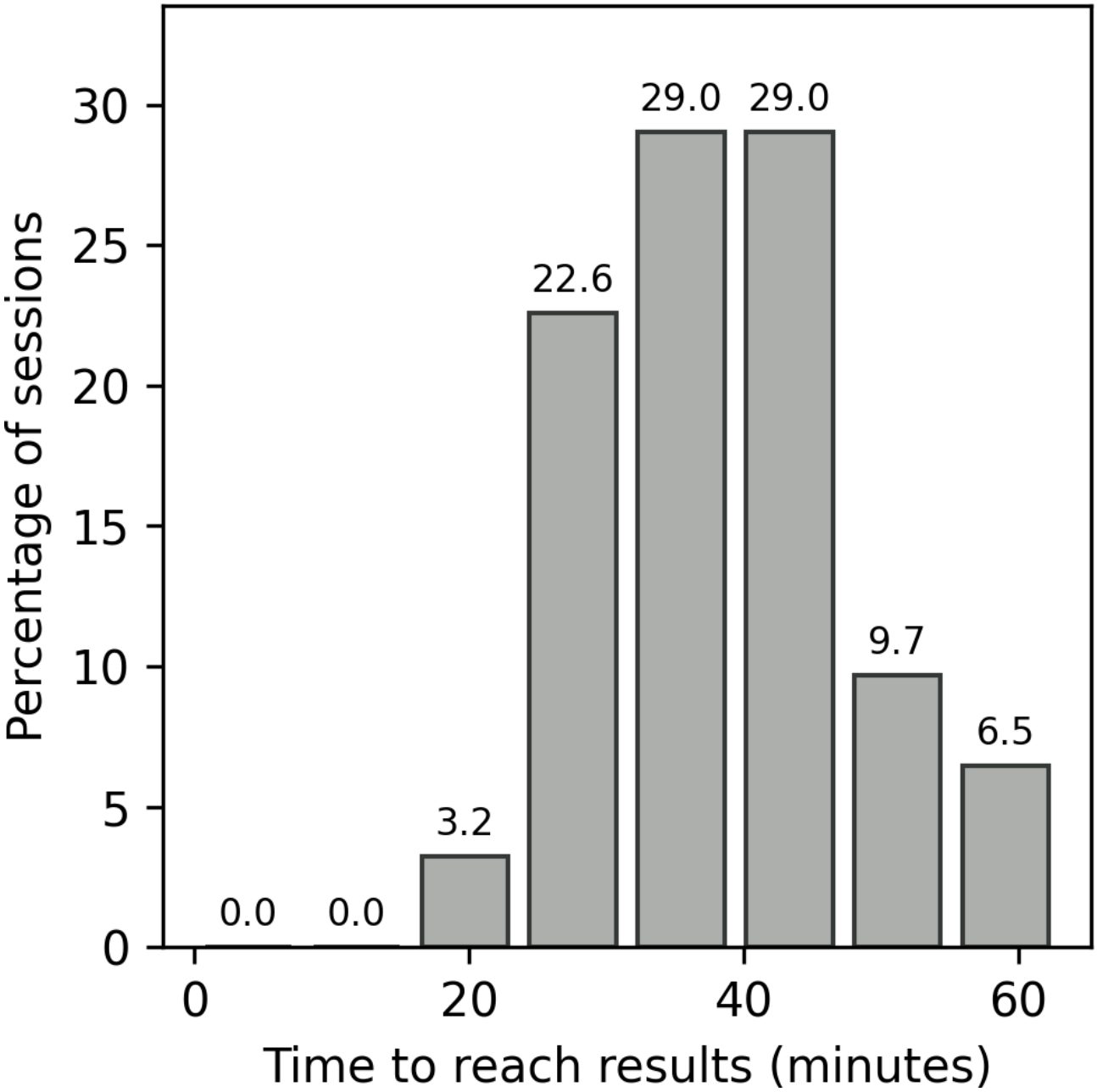

Usability

The mLab+ app was rated by participants in the intervention arm using the Health-ITUES [29] and PSSUQ [32] (Table 4). The overall mean score on the Health-ITUES was 3.62 (SD 1.07; range 1-5), with higher scores indicating greater perceived usefulness. The overall mean score on the PSSUQ was 2.65 (SD 1.06; range 1-4.19), with higher scores indicating lower perceived usefulness. The scores on both measures indicate medium to high usability.

XSL•FO
RenderX

Table . Health Information Technology Usability Evaluation Scale (Health-ITUES) and Post-Study System Usability Questionnaire (PSSUQ) usability scores as rated by participants who used the mLab+ app.

Instrument	Score, mean (SD)
PSSUQ (strongly agree=1 to strongly disagree=7)	
Overall	2.65 (1.06)
System usefulness	2.51 (1.15)
Information quality	2.77 (1.03)
Interface quality	2.67 (1.24)
Health-ITUES (Strongly agree=5 to strongly disagree=1)	
Overall	3.62 (1.07)
Quality of work life	3.52 (1.08)
Perceived usefulness	3.64 (1.23)
Perceived ease of use	3.76 (1.05)
User control	3.46 (1.07)

Recruitment

During the recruitment and enrollment period, 707 people were screened for eligibility, with 82.7% (585/707) being screened as ineligible and 17.3% (122/707) being screened as eligible. Having had an HIV test within the previous 3 months (n=327, 46.8%) and an HIV diagnosis (n=152, 21.5%) were the most frequently cited reasons for ineligibility. Exclusion criteria were not mutually exclusive. In addition, 7% (6/82) of the people who were screened as eligible did not show up to their study visit, 9% (7/82) of the participants showed up to their appointment but were found to be ineligible and not enrolled, and others (69/82, 84%) never scheduled a baseline appointment. In total, 50% (20/40) of the enrolled participants were randomized to the control arm and, thus, did not use the mLab+ app. During the study's 6-month recruitment and enrollment period, an average of 5 participants (SD 4.57) were enrolled per month.

Retention

Of the 20 participants who enrolled in the study and used the mLab+ app, 16 (80%) completed their 3-month follow-up study assessment. One participant had a known HIV diagnosis before enrolling in the study and was administratively withdrawn. Another participant voluntarily withdrew from the study. The remaining 2/4 (50%) were lost to follow-up.

Discussion

Principal Findings

The findings of this study focused on the feasibility and usability of the mLab+ app. Most, but not all, participants (16/20, 80%) were able to use the test on their own without the assistance of the clinician. This points to some of the challenges for consumers using these tests at home without the supervision of a clinician. Notably, resistance to finger stick blood collection is often a barrier to self-testing [33]. Nonetheless, past work has similarly demonstrated difficulties among people using at-home testing, such as the OraQuick HIV test, which requires fewer steps than the DPP HIV-Syphilis test and has already been FDA approved for at-home use [9]. Future studies should collect data on participant self-testing history, including experience with and acceptance of at-home tests, which may highlight additional barriers associated with successful completion of a self-test through the mLab+ app.

The overall usability of the app was medium to high, with lower scores on the usefulness subscale. Given that the mLab+ app was originally designed for in-home use, these scores reflect the use of this app in a clinical setting, which was not a part of its intended design. Furthermore, the usefulness of the app after the study visit was limited given that the app was designed to support using the HIV and syphilis test at home and participants did not have access to the test in their homes.

The recruitment and retention rate were acceptable within the context of this small sample size and would warrant further consideration in a fully powered trial. Retention would likely be improved in a future trial if research visits took place remotely [7,34], as is evidenced by the original mLab study, which largely took place over Zoom and resulted in a retention rate of 89.5% at the 6-month follow-up and 84.6% at the 12-month follow-up; this trial was much larger (N=525) and longer (12-month study period) than the mLab+ app study, demonstrating the potential for improved retention in a remote setting [7,34]. Given the FDA restrictions at the time of this study's start, we were not able to conduct the mLab+ app study remotely. Furthermore, to improve linkage to care among HIV-positive participants, future studies should consider integrating nearby resources, including addresses and phone numbers, into the app, as well as next steps to facilitate making a physician's appointment, as demonstrated in the SMARTtest mobile health study for HIV and syphilis testing [35]. This feature would be helpful for participants who refused on-site testing the day they received their reactive result.

Finally, the paradata highlight key aspects of user behavior and the effectiveness of the testing workflow. Users predominantly spent time on testing-related screens, such as timers, guided walkthroughs, and result screens, reflecting the primary purpose of the app. Auxiliary screens such as the privacy policy and

XSL•FO
RenderX

help were visited briefly, which is expected given the scope of the study; these screens, particularly the help screen, would be expected to have more use in a larger trial or in an environment where study team members were not immediately available. The consistency in time spent on both timer screens suggests that the automated workflows function as intended, guiding users smoothly through the testing process by closely following the manufacturer's guidelines. Importantly, the analysis of the workflow conversion funnel indicates minimal drop-off once a test was initiated. This, particularly when combined with the usability results, suggests that the app effectively supported users in completing the testing process. Furthermore, the time to complete the tests aligns well with the expected duration, although slight deviations may occur due to session resets when users resume testing after logging out. Overall, the findings underscore the usability and robustness of the app in facilitating the testing workflow and its suitability for use in a larger study.

Limitations

This study was designed as a pilot RCT, but due to the limitations of conducting the study in a clinical setting, we are only reporting on the intervention arm, who had access to the mLab+ app and the combination test for HIV and syphilis. Although participants were randomized into intervention versus control arms, we did not include control arm data in this paper. As originally designed, study participants would use the mLab+ app and the combination test for HIV and syphilis for 3 months in a pilot RCT. The original conceptualization of the study design was to assess testing uptake and referral to services in the participants who were given access to the mLab+ app and HIV and syphilis test as compared to a control arm, which would only receive information on HIV and syphilis testing services. The FDA denied the request for study participants to use the combination test for HIV and syphilis in their homes, so the study design was modified, and participants were only able to use the test in a clinical setting under the supervision of a nurse practitioner. Future studies should consider the feasibility and acceptability of the mLab+ app outside of the clinical setting.

Conclusions

Findings from this study support the use of the mLab+ app as a tool for assisting consumers in self-testing for HIV and syphilis. This is especially important in the context of the growing HIV and syphilis epidemic among MSM in the United States. Nonetheless, the limitations of the study design warrant further examination outside of the clinic setting to better understand the utility of these tools for improving consumer health outcomes.

Funding

This research was supported by the National Institutes of Health (NIH) under awards R01MH118151, UG3AI169658, UH3AI169658, and K24NR018621. TFS would like to acknowledge support from the NIH-funded Tennessee Center for AIDS Research (P30AI110527). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Data Availability

The datasets generated or analyzed during this study are not publicly available but are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Checklist 1
CONSORT-EHEALTH checklist (V 1.6.1).
[PDF File, 1171 KB - formative_v9i1e72955_app1.pdf]

References

1. Guidelines on HIV self-testing and partner notification: supplement to consolidated guidelines on HIV testing services. Pan American Health Organization. 2016. URL: https://www.paho.org/en/documents/guidelines-hiv-self-testing-and-partner-notification-supplement-consolidated-guidelines [accessed 2025-11-10]
2. Delaney KP, DiNenno EA. HIV testing strategies for health departments to end the epidemic in the U.S. Am J Prev Med 2021 Nov;61(5 Suppl 1):S6-S15. [doi: 10.1016/j.amepre.2021.06.002] [Medline: 34686292]
3. Fauci AS, Redfield RR, Sigounas G, Weahkee MD, Giroir BP. Ending the HIV epidemic: a plan for the United States. JAMA 2019 Mar 5;321(9):844-845. [doi: 10.1001/jama.2019.1343] [Medline: 30730529]
4. Hecht J, Sanchez T, Sullivan PS, DiNenno EA, Cramer N, Delaney KP. Increasing access to HIV testing through direct-to-consumer HIV self-test distribution - United States, March 31, 2020-March 30, 2021. MMWR Morb Mortal Wkly Rep 2021 Sep 24;70(38):1322-1325. [doi: 10.15585/mmwr.mm7038a2] [Medline: 34555001]
5. MacGowan RJ, Chavez PR, Borkowf CB, et al. Effect of internet-distributed HIV self-tests on HIV diagnosis and behavioral outcomes in men who have sex with men: a randomized clinical trial. JAMA Intern Med 2020 Jan 1;180(1):117-125. [doi: 10.1001/jamainternmed.2019.5222] [Medline: 31738378]

XSL•FO
RenderX

6. Sanabria G, Scherr T, Garofalo R, et al. Usability evaluation of the mLab app for improving home HIV testing behaviors in youth at risk of HIV infection. AIDS Educ Prev 2021 Aug;33(4):312-324. [doi: 10.1521/aeap.2021.33.4.312] [Medline: 34370566]
7. Schnall R, Scherr TF, Kuhns LM, et al. Efficacy of the mLab app: a randomized clinical trial for increasing HIV testing uptake using mobile technology. J Am Med Inform Assoc 2025 Feb 1;32(2):275-284. [doi: 10.1093/jamia/ocae261] [Medline: 39560363]
8. Schnall R, John RM, Carballo-Dieguez A. Do high-risk young adults use the HIV self-test appropriately? Observations from a think-aloud study. AIDS Behav 2016 Apr;20(4):939-948. [doi: 10.1007/s10461-015-1240-6] [Medline: 26518679]
9. Galli RA, Maraj D, McBain K, et al. A prospective multi-site study to evaluate the performance and usability of an oral fluid-based HIV self-test in Canada. BMC Public Health 2025 Jan 11;25(1):125. [doi: 10.1186/s12889-024-21228-8] [Medline: 39794710]
10. U.S. syphilis cases reach highest numbers since the 1950s. Centers for Disease Control and Prevention. 2024. URL: https://www.cdc.gov/nchhstp/newsroom/releases/2024/STI-Surveillance-Report-2022.html [accessed 2025-11-10]
11. Sexually transmitted diseases surveillance. Centers for Disease Control and Prevention. 2017. URL: https://stacks.cdc.gov/view/cdc/59237 [accessed 2025-11-10]
12. Syphilis & MSM (men who have sex with men). Centers for Disease Control and Prevention. 2023. URL: https://www.cdc.gov/std/treatment-guidelines/msm.htm [accessed 2025-11-10]
13. Pathela P, Braunstein SL, Schillinger JA, Shepard C, Sweeney M, Blank S. Men who have sex with men have a 140-fold higher risk for newly diagnosed HIV and syphilis compared with heterosexual men in New York City. J Acquir Immune Defic Syndr 2011 Dec 1;58(4):408-416. [doi: 10.1097/QAI.0b013e318230e1ca] [Medline: 21857351]
14. Sexually transmitted disease surveillance 2014. Centers for Disease Control and Prevention. 2015. URL: https://stacks.cdc.gov/view/cdc/36835 [accessed 2025-11-10]
15. Sexually transmitted infections surveillance 2023. Centers for Disease Control and Prevention. 2023. URL: https://www.cdc.gov/sti-statistics/media/pdfs/2025/09/2023_STI_Surveillance_Report_FINAL_508.pdf [accessed 2025-11-10]
16. Rael CT, Kutner BA, Lopez-Rios J, Lentz C, Dolezal C, Balán IC. Experiences of transgender women who used a dual HIV/syphilis rapid self-test to screen themselves and potential sexual partners (the SMARTtest study). AIDS Behav 2022 Apr;26(4):1229-1237. [doi: 10.1007/s10461-021-03478-4] [Medline: 34559351]
17. Shover CL, De Vost MA, Beymer MR, Gorbach PM, Flynn RP, Bolan RK. Using sexual orientation and gender identity to monitor disparities in HIV, sexually transmitted infections, and viral hepatitis. Am J Public Health 2018 Nov;108(S4):S277-S283. [doi: 10.2105/AJPH.2018.304751] [Medline: 30383431]
18. Becasen JS, Denard CL, Mullins MM, Higa DH, Sipe TA. Estimating the prevalence of HIV and sexual behaviors among the US transgender population: a systematic review and meta-analysis, 2006-2017. Am J Public Health 2019 Jan;109(1):e1-e8. [doi: 10.2105/AJPH.2018.304727] [Medline: 30496000]
19. Vyavaharkar M, Moneyham L, Corwin S, Saunders R, Annang L, Tavakoli A. Relationships between stigma, social support, and depression in HIV-infected African American women living in the rural southeastern United States. J Assoc Nurses AIDS Care 2010;21(2):144-152. [doi: 10.1016/j.jana.2009.07.008] [Medline: 19879778]
20. Zalazar V, Frola CE, Gun A, et al. Acceptability of dual HIV/syphilis rapid test in community- and home-based testing strategy among transgender women in Buenos Aires, Argentina. Int J STD AIDS 2021 May;32(6):501-509. [doi: 10.1177/0956462420979852] [Medline: 33533303]
21. Syphilis and people with HIV. HIV.gov. 2024. URL: https://www.hiv.gov/hiv-basics/staying-in-hiv-care/other-related-health-issues/syphilis [accessed 2025-11-10]
22. Guidelines for the prevention and treatment of opportunistic infections in adults and adolescents with HIV. HIV.gov. 2023. URL: https://clinicalinfo.hiv.gov/en/guidelines/hiv-clinical-guidelines-adult-and-adolescent-opportunistic-infections/syphilis [accessed 2025-11-10]
23. Buchacz K, Patel P, Taylor M, et al. Syphilis increases HIV viral load and decreases CD4 cell counts in HIV-infected patients with new syphilis infections. AIDS 2004 Oct 21;18(15):2075-2079. [doi: 10.1097/00002030-200410210-00012] [Medline: 15577629]
24. Kularadhan V, Gan J, Chow EPF, Fairley CK, Ong JJ. HIV and STI testing preferences for men who have sex with men in high-income countries: a scoping review. Int J Environ Res Public Health 2022 Mar 4;19(5):3002. [doi: 10.3390/ijerph19053002] [Medline: 35270694]
25. Screening recommendations and considerations referenced in treatment guidelines and original sources. Centers for Disease Control and Prevention. 2024. URL: https://www.cdc.gov/std/treatment-guidelines/screening-recommendations.htm [accessed 2025-11-10]
26. Klein PW, Martin IBK, Quinlivan EB, Gay CL, Leone PA. Missed opportunities for concurrent HIV-STD testing in an academic emergency department. Public Health Rep 2014;129 Suppl 1(Suppl 1):12-20. [doi: 10.1177/00333549141291S103] [Medline: 24385644]
27. Mimiaga MJ, Reisner SL, Grasso C, et al. Substance use among HIV-infected patients engaged in primary care in the United States: findings from the Centers for AIDS Research Network of Integrated Clinical Systems cohort. Am J Public Health 2013 Aug;103(8):1457-1467. [doi: 10.2105/AJPH.2012.301162] [Medline: 23763417]

XSL•FO
RenderX

28. Liu AY, Vittinghoff E, von Felten P, et al. Randomized controlled trial of a mobile health intervention to promote retention and adherence to preexposure prophylaxis among young people at risk for human immunodeficiency virus: the EPIC study. Clin Infect Dis 2019 May 30;68(12):2010-2017. [doi: 10.1093/cid/ciy810] [Medline: 30239620]
29. Schnall R, Cho H, Liu J. Health information technology usability evaluation scale (Health-ITUES) for usability assessment of mobile health technology: validation study. JMIR Mhealth Uhealth 2018 Jan 5;6(1):e4. [doi: 10.2196/mhealth.8851] [Medline: 29305343]
30. Lewis JR. IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use. Int J Hum Comput Interact 1995 Jan;7(1):57-78. [doi: 10.1080/10447319509526110]
31. Kreuter F. Improving Surveys with Paradata: Analytic Uses of Process Information: John Wiley & Sons; 2013, Vol. 581. [doi: 10.1002/9781118596869]
32. Lewis J. Psychometric evaluation of the PSSUQ using data from five years of usability studies. Int J of Human-Comp Interaction 2002 Sep 1;14(3):463-488. [doi: 10.1207/S15327590IJHC143&4_11]
33. Wang C, Cheng W, Li C, et al. Syphilis self-testing: a nationwide pragmatic study among men who have sex with men in China. Clin Infect Dis 2020 May 6;70(10):2178-2186. [doi: 10.1093/cid/ciz603] [Medline: 31260513]
34. Schnall R, Kuhns LM, Pearson C, et al. Efficacy of MyPEEPS mobile, an HIV prevention intervention using mobile technology, on reducing sexual risk among same-sex attracted adolescent males: a randomized clinical trial. JAMA Netw Open 2022 Sep 1;5(9):e2231853. [doi: 10.1001/jamanetworkopen.2022.31853] [Medline: 36129712]
35. Balán IC, Lopez-Rios J, Nayak S, et al. SMARTtest: a smartphone app to facilitate HIV and syphilis self- and partner-testing, interpretation of results, and linkage to care. AIDS Behav 2020 May;24(5):1560-1573. [doi: 10.1007/s10461-019-02718-y] [Medline: 31679075]

Abbreviations

CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
FDA: Food and Drug Administration
Health-ITUES: Health Information Technology Usability Evaluation Scale
MSM: men who have sex with men
NPG: Nurse Practitioner Group
PHI: protected health information
POC: point of care
PrEP: preexposure prophylaxis
PSSUQ: Post-Study System Usability Questionnaire
RCT: randomized controlled trial
REDCap: Research Electronic Data Capture
STI: sexually transmitted infection

Edited by A Mavragani; submitted 22.02.25; peer-reviewed by J Mitchell; revised version received 11.07.25; accepted 14.07.25; published 30.12.25.

Please cite as:
Brin M, Scherr TF, Chaiyasit J, Liu J, Abbott M, Garofalo R, Kuhns LM, Sky T, Esliker I, Schnall R
Feasibility and Usability of an mHealth App (mLab+) to Guide Users Through HIV and Syphilis Self-Testing: Pilot Randomized Controlled Trial
JMIR Form Res 2025;9:e72955
URL: https://formative.jmir.org/2025/1/e72955
doi:10.2196/72955

© Maeve Brin, Thomas F Scherr, Janejira Chaiyasit, Jianfang Liu, Maura Abbott, Robert Garofalo, Lisa M Kuhns, Tess Sky, Ian Esliker, Rebecca Schnall. Originally published in JMIR Formative Research (https://formative.jmir.org), 30.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

A Novel Primary Care Planning Informatics Tool Informed by Data-Driven Multimorbidity Grouping: User-Centered Design and Feasibility Testing

Linnaea Schuttner[1,2], MD, MS; Karin Daniels[3], PhD; Srinidhi Venkatesh[4], MD; Franya Hutchins[3,5], PhD; Karley Atchison[3], MA; Rebecca Piegari[6], MS; Timothy Bober[4], MD, MS; Xinhua Zhao[3], PhD; Ann-Marie Rosland[3,4], MD, MS

[1]Health Services Research & Development, VA Puget Sound Health Care System, 1660 S Columbian Way, Seattle, WA, United States

[2]School of Medicine, University of Washington, Seattle, WA, United States

[3]VA Center for Health Equity Research and Promotion (CHERP), VA Pittsburgh Healthcare System, Pittsburgh, PA, United States

[4]Department of Medicine and Caring for Complex Chronic Conditions Research Center, School of Medicine, University of Pittsburgh, Pittsburgh, PA, United States

[5]Department of General Internal Medicine, Pearlman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States

[6]Analytics and Performance Integration, Veterans Health Administration, Washington, DC, United States

Corresponding Author:
Linnaea Schuttner, MD, MS
Health Services Research & Development, VA Puget Sound Health Care System, 1660 S Columbian Way, Seattle, WA, United States

Abstract

Background: Patients with multimorbidity have complex health care needs and are at high risk for adverse health outcomes. Primary care teams need tools to effectively and proactively plan care for these patients. We developed VET-PATHS (Veteran Panel Management Tool for High-Risk Subgroups), a novel care planning informatics tool for complex primary care patients. VET-PATHS groups patients by chronic condition via latent class analysis of electronic health record data, then jump-starts care planning by suggesting “care steps” based on data-driven high-priority care for the group.

Objective: The study aimed to iteratively adapt VET-PATHS with user input, then test feasibility and acceptability by frontline primary care teams for empaneled patients at high risk.

Methods: Three rounds of user-centered design sessions with 17 primary care providers and registered nurses at 5 sites from 2019 to 2021 were conducted to obtain feedback on the VET-PATHS layout, content, and user interface. Feedback was summarized into 4 user experience domains (useful, desirable, credible, and usable), leading to progressively updated prototypes. After the national tool release, we conducted a pilot intervention study in 2023‐2024 with 6 primary care teams at 4 sites using VET-PATHS during asynchronous regular meetings. Tool use and resulting care plans were assessed by templated observation during meetings, postpilot chart review, and administrative data. Individual qualitative interviews were analyzed by rapid template analysis for feasibility, acceptability, and utility.

Results: User-centered feedback led to updated tool content, context (eg, use in proactive panel management), targeted users (eg, focusing on primary care providers), and display layout. Pilot teams used VET-PATHS over 4 to 8 weekly meetings (mean length 24, range 16‐49 min), actively reviewing 80% (280/351) of empaneled patients at high risk visible in the tool. Tool use prompted 127 new actions for 91 unique patients (33% of patients reviewed) and documentation of >1 new care plan for 19% of patients reviewed. Common actions included requests to return to the clinic (n=34, 27%), referrals (n=25, 20%), or vaccinations (n=24, 19%). Of the 127 actions planned, 53 (42%) were received by patients. Difference-in-difference trends for acute hospitalizations declined post pilot, while outpatient utilization was stable or increased for pilot team patients compared to all patients at high risk at pilot sites (per-patient counts: acute hospitalizations −0.15; primary care visits 0.00; mental health visits 0.99). Four generalist teams (n=11 interviews) described higher acceptability. Two “focused” teams with more homogenous panels, for example, substance use disorder (n=3 interviews), found care steps less useful. Teams described how VET-PATHS improved efficiency of care planning through automated patient grouping and identification of care gaps and increased multidisciplinary involvement.

Conclusions: User-centered improvements to VET-PATHS were designed to help clinicians process and use complex information about patient multimorbidity to efficiently create new care plans. Subsequently, VET-PATHS was acceptable and feasible to frontline primary care teams, particularly with more general patient panels, and led to concrete changes to clinical care delivery.

XSL•FO
RenderX

(JMIR Form Res 2025;9:e75081) doi:10.2196/75081

KEYWORDS

complex care; panel management; population health; primary care; multimorbidity; care planning

Introduction

The number of patients with multimorbidity (ie, multiple chronic health conditions) and complex health needs is increasing as chronic diseases accumulate at younger ages [1,2]. In the Veterans Health Administration (VHA) and other health care systems, the top 10% of patients who are most complex account for 80% of health care utilization [3]. Patients with multimorbidity are also at high risk for worse health outcomes, including functional decline [4,5] and death [6]. These patients require greater time and effort from health care staff in primary care visits [4,7,8], potentially adding to clinician burnout [9,10]. Delivering care that effectively improves health outcomes for these patients is an ongoing but important challenge requiring innovative solutions.

Many health care systems, including VHA, have risk-prediction scores that identify patients at high risk for hospitalization or other adverse events; however, clinicians are left wondering what to do once a patient is identified as high risk [11]. “One-size-fits-all” interventions for high-risk patients have not been successful [12,13], in part due to heterogeneity in concurrent multimorbidity. In VHA, primary care teams have access to tools that highlight which of their patients are at high risk but have described feeling overwhelmed by the volume of patient health information to review and the lack of guidance on what care needs are most important for high-risk patients [14-16]. Other tools are focused on specific health needs or diseases, such as preventive care, to identify care gaps [17]. However, to our knowledge, no tools within or outside VHA currently exist that identify and prioritize care needs across health conditions for high-risk patients with complex multimorbidity.

One promising approach is to develop individualized care plans for patients with multimorbidity. Care planning involves a comprehensive review of patient health and care history and the development of a multidisciplinary action plan [18]. Care planning has shown promising results in specialized settings (eg, geriatrics) and populations (eg, diabetes) [19-23] and is widely recommended for high-risk patients and those with complex multimorbidity [13,24,25]. However, individualized care planning has not been successfully implemented in primary care due to the amount of time, effort, and resources required for this patient population, as well as varied care needs [12,23,26].

Population segmentation for primary care patients with multimorbidity could help bridge the gap between time-consuming individualized care planning and ineffective one-size-fits-all interventions. Population segmentation is a popular analytic technique used by health care systems to identify groups of patients more similar to each other within the wider patient population [27-30]. Data-driven segmentation applied to patients with complex multimorbidity could identify groups of patients for tailored, group-specific interventions. To date, population segmentation has not been bridged beyond observational uses to identify patients for health care interventions within the care delivery of high-risk patients.

In this paper, we describe the user-centered design and pragmatic pilot testing of VET-PATHS (Veteran Panel Management Tool for High-Risk Subgroups), a novel clinical informatics tool to facilitate care planning for high-risk primary care patients with multimorbidity. VET-PATHS uses population segmentation to sort patients with high-risk scores for hospitalization into groups of patients with similar patterns of multimorbidity and identify priority care gaps based on group-specific risks for hospitalization and a patient’s prior receipt of care. In this study, frontline primary care teams gave input into the VET-PATHS design, then used VET-PATHS to generate care plans for their empaneled patients who were at high risk, tailored to multimorbidity groupings and individualized patient needs. Our study objective was to describe the user-centered design process and results of a feasibility and acceptability pilot of VET-PATHS for team-based management of patients at high risk with multimorbidity within the VHA.

Methods

Primary Care Setting

In the VHA, primary care is delivered to over 6 million veterans in over 1100 facilities through multidisciplinary care teams (*Patient Aligned Care Teams* [PACTs]) [31]. Fully staffed PACTs include a primary care clinician (PCP), registered nurse (RN) care manager, clinical associate (CA; eg, licensed vocational nurse), and medical support administrator (MSA) at a ratio of 1 full-time equivalent PCP to 3 non-PCP “core” PACT members. Patients are empaneled to a specific PACT at a ratio of 1200 patients per 1.0 full-time equivalent PCP. “Extended” PACT multidisciplinary providers are shared across core PACTs and include embedded clinical pharmacists, social workers, mental health providers, and nutritionists. PACTs are encouraged to use “panel management” tools during asynchronous (nonpatient time) team meetings to identify and meet health needs for subpopulations of patients on their panel, such as identifying and contacting all empaneled patients currently due for cancer screening [17].

VET-PATHS Patient Population

VET-PATHS displays a team’s empaneled primary care patients with a high VHA Care Assessment Need (CAN) score. The CAN score is a validated risk prediction score for adverse outcomes used nationally in VHA [32]. Patients are included in VET-PATHS if they have a CAN score ≥90th percentile at any time within the past 1 year, corresponding to an approximate 25% risk of hospitalization; patients with newly high CAN scores are added quarterly. Once included in VET-PATHS, patients remain visible within the tool for 12 months after their most recent qualifying CAN score.

XSL•FO
RenderX

Tool Description

VET-PATHS is available on a VHA intranet website alongside other panel management tools and population health registries. The tool draws from databases of electronic health records (EHRs) to display patients actively empaneled with the team and the patients' up-to-date medical and health care information. Patients are first displayed in a panel view, sortable by multimorbidity group label (described later) and key demographic and clinical information. A second individual patient-level view provides suggested care plan elements ("care steps," described below) tailored to the patient's group, alongside additional relevant patient health and utilization information (see Multimedia Appendix 1 for tool screenshots).

1. Multimorbidity groups: Patient assignment to multimorbidity groups is derived from previously described and validated population segmentation [33-36]. Briefly, these models identify latent multimorbidity groups among high-risk Veterans using 31 chronic conditions [36] identified via health care encounters within the prior 24 months, based on *International Classification of Diseases, 10th Edition* codes from the US Department of Health and Human Services' multiple chronic conditions framework [37] and expanded for Veteran care (eg, inclusion of posttraumatic stress disorder; Multimedia Appendix 2). At the time of the pilot intervention reported here, there were 6 multimorbidity groups used in the tool, named for the most common diagnoses among patients assigned to that group: cancer, cardiometabolic, liver, mental health, substance use, and low comorbidity. Veterans are uniquely assigned to a group if the models show a probability of matching ≥70% to the group; patients are not required to have all of the group conditions to be assigned and often have differing but overlapping sets of conditions. Around 6% to 8% of patients at high risk do not meet the threshold for assignment to any group and are labeled "unassigned."
2. Care steps: VET-PATHS also provides suggested care plan components for the primary care team to consider for each patient. These are called "care steps" in the tool and are tailored to each multimorbidity group's unique needs. Within each group, the individual patient-level view displays each care step as (1) applicable to the patient but not yet received (*due*); (2) applicable and already received, with date of receipt (*done*); or (3) not applicable to the individual patient (*N/A*). Care step status is automatically drawn from administrative databases linked to the patient's EHR for care already received (including encounters, diagnoses, and laboratory results) but is not able to capture data from "free-text" EHR notes. Care step status based on receipt of health care is updated daily. Care steps are presented if relevant to the patient's multimorbidity group; an individual patient may not have all steps "due" or "applicable" and may have other individual care needs not part of the group's care steps and thus not displayed by the tool.
3. Steps are presented within the tool as optional suggestions. The need for clinician review and agreement was emphasized in tool training, graphics, and documentation. During the pilot, 4 of 6 multimorbidity groups had care steps programmed and available for preliminary testing and were the focus of the pilot intervention (cardiometabolic, liver, mental health, and substance use groups).

Care steps target three main types of care plan components for each group. First, some care steps target common gaps in quality metrics for a multimorbidity group. For example, a care step may suggest hemoglobin A_{1C} testing in patients with diabetes, if due, based on prominent group-level gaps in this metric for mental health group patients. Other care steps highlight common gaps in health care services within the group in response to adverse medical events [34], for example, screening for fall risk factors in groups with comparatively high rates of medically documented falls or nephrology referral in groups with a frequent diagnosis of advanced kidney disease but underuse of this guideline-recommended service. Third, care steps address unique risk factors for hospitalization identified for each group based on their EHR data [34], for example, addressing alcohol use above recommended limits for groups with comparatively higher hospitalization risk in patients with alcohol use disorder.

We initially developed and iteratively refined 3 to 5 care steps for each group within VET-PATHS. Initial candidate care steps were chosen based on gaps in care identified from observed and predicted differences in clinical conditions, utilization patterns, and health outcomes specific to each group [34,38]. Specific care steps were aligned with relevant VHA/Department of Defense Clinical Practice Guidelines when applicable. Care step candidates then went through three rounds of clinical expert review: two multispecialty clinical panels conducted with 39 unique experts representing VHA program offices (eg, Primary Care and Population Health), clinical specialties (eg, cardiology, substance use disorder, geriatrics), and special populations (homeless care, women's health); then input from frontline clinicians during preliminary testing. Care steps were designed to not overlap with existing population health management, registry tools, or clinical "reminders" in the EHRs; although developed for VET-PATHS prior to release as a clinical reminder, one care step (hepatitis B vaccination) for liver and substance use group patients was available to some sites as a VHA national clinical reminder released in October 2018 (for at-risk adults, universal screening was recommended in 2023 [39]).

Iterative User-Centered Design and Prototype Development

From 2019 to 2021, we conducted three rounds of user-centered design to develop the VET-PATHS content and user interface [40]. Target primary care clinical users were recruited via staff meetings and purposeful outreach to clinicians already serving as champions in related projects. Eligible users were either PCPs or RN care managers within an active primary care role at a VHA site; RNs and PCPs were recruited separately, not as team dyads. For each round, we conducted 20 - to 45-minute virtual or in-person individual semistructured interviews. Prototypes were updated based on interviewee input between each round. Round one (2019) used paper and static screen mock-ups, displaying fictional patient data to interview participants. Participants were asked about the interpretation of displayed clinical information, design features, and perceived usefulness.

XSL•FO
RenderX

Round two (2019) was completed using electronic visuals (screenshots, pictures) of an updated prototype completed with simulated patient information. Participants were asked in this phase about clinical interpretation, design features, and potential workflows for tool use. Round three (2021) used an interactive web-based version of the tool. This version displayed the clinician's actual patient panel and patients' real-time health information. In this round, participants were asked to "walk through" how they would navigate, interpret, and use different parts of the tool and describe how the content related to the care or understanding of their patients.

Interviews and walk-throughs were audio-recorded and conducted by a PhD-level staff member, with a research coordinator taking detailed notes verified via recording review for accuracy and completeness. Feedback was summarized into deductive categories guided by 4 user experience domains (useful, desirable, credible, and usable) [41], as well as validation of the tool content for real-world patient care (round 3 only). Updates based on the summary reports were made to the tool between each design round, and user interview guides were refined in response to results and Veteran stakeholder board (2019) feedback. Following these three rounds of user-centered improvements, the tool was released for nationwide clinical use in 2021.

Intervention Pilot Study

In 2023 - 2024, we conducted a pilot intervention study with frontline PACTs to examine the feasibility, acceptability, and perceived utility of using the tool to develop care plans for empaneled patients with complex multimorbidity. Outreach to potentially interested teams was based on recommendations by site leadership and peer referral. After a clinician "champion" from a team agreed to participate, the team was trained on tool use. Study facilitators then scheduled 30-minute weekly or biweekly meetings, during which the team met to review and discuss their patients displayed in VET-PATHS by multimorbidity group, for a maximum of 8 weeks. Champions could include any core or extended PACT members in their meetings at their discretion. Teams were provided with optional adaptable documentation templates to track care planned for a patient in the main EHR. Study facilitators supported team meetings, including answering technological questions if asked (eg, how to navigate screens within the tool), but were instructed not to provide unsolicited prompts on tool use, advise teams on clinical interpretation or application of information, or otherwise direct how teams used the tool for patient care. Two study facilitators attended each meeting for data adjudication and fidelity review.

Data on tool use and resulting clinical activity were collected from multiple sources. First, study facilitators directly observed team discussions and actions intended during their meetings, recording observations into a structured data collection template. Second, manual chart reviews were completed by separate, nonfacilitator study staff approximately 60 days after the team's final meeting. For both facilitator observation and chart review, actions considered related to VET-PATHS meetings, which included (1) tool-suggested care steps discussed and related intended actions (eg, based on VET-PATHS care steps, team decided a patient should receive a vaccine; PCP stated that a patient should get updated laboratory tests); (2) new individualized care plans documented in the EHR after the team meeting tracking intended actions, using a recommended "Care Plan" note template or note addendum; (3) intended actions carried out or ordered (eg, phone call by nurse to discuss vaccination documented; laboratory order placed in the EHR by the PCP); and (4) actions received (eg, patient actually received the vaccine; patient completed laboratory test). Following the chart review for data collection, two other study members (MD and PhD level) completed data abstraction for clinical activity independently, using prespecified case definitions. Agreement between abstractors was 92%, with remaining differences resolved by discussion. Third, at the conclusion of the 8-week pilot period, team members were sent an electronic confidential System Usability Scale (SUS) [42], adapted to the VET-PATHS tool and its functions. SUS questions are scored on a 1 to 5 Likert scale, with a maximum possible total of 100 points (higher is more usable); a score of 68.0 (industry average) or higher is considered "acceptable" usability [42,43]. Fourth, qualitative baseline interviews with the PACT champion and postintervention individual interviews with each participating team member were conducted by trained study staff. Postintervention interviews were conducted by staff uninvolved in the pilot or facilitation. Interviews were audio-recorded and transcribed using Microsoft Teams; then, transcripts were verified by another study team member. Interview topics included contextual aspects (eg, typical patient care and panel management workflows, usual team workload division, staffing shortages) that informed prepilot team training and meeting logistics and other factors that influenced team acceptability [44] and usability of the tool (eg, competing clinic initiatives) in both pre- and postpilot interviews. Qualitative data were analyzed by team and by clinical role. Two study qualitative analysts not involved in the user-design testing or pilot intervention completed rapid template analysis of transcripts [45], with result interpretation supported by discussion and consensus among the larger study team. Finally, administrative data outcomes for utilization were collected as exploratory findings. This pilot study was not designed with the power to detect statistically significant differences in outcomes; therefore, we limit exploratory outcome analyses to descriptive statistics only [46,47]. Patient baseline outcomes for 3 months prior to the index date of team enrollment were compared within 3 to 6 months post final team meeting, using an intention-to-treat difference-in-difference approach. Comparator trends were for similar time periods among patients with high CAN scores within each pilot site but not enrolled on a pilot team.

Ethical Considerations

This work was carried out as a quality improvement evaluation and designated nonresearch under the terms of a signed attestation from the VHA Office of Primary Care, which exempts this work from institutional review board review, in accordance with the national VHA Office of Research and Development's nonresearch policy under the US Department of Veterans Affairs. This VHA documentation and review process ensures that work not carried out under a human

XSL•FO
RenderX

subject's protocol is part of institutionally sanctioned quality improvement activities. Primary participant data (eg, qualitative interviews) specific to project activities are collected with individual verbal informed consent and handled in accordance with ethical and privacy guidelines consistent with research-designated data. Data from participants were deidentified for analysis and anonymized for reporting. No additional compensation was provided to individual participants outside of their usual salary; however, reimbursement of time for team participation in pilot testing was covered by the project to offset patient care time needed for quality improvement activities.

Results

User-Centered Design (Prototype Development) Phase

A total of 17 clinical users (Tables 1 and 2), including PCPs (n=5) and RN or nursing supervisors (n=12), completed interviews at 5 VHA primary care sites across the United States (Northeast: n=8; Midwest: n=3; South: n=4; and West: n=2). Round one of user interviews (n=6 RN, paper prototype) led to refining the tool and workflows to target PCPs for care recommendations within care steps (usefulness). Users felt the tool was desirable for some contexts, including perceptions that using the tool would be helpful for proactive panel management, improving understanding of new patients to a provider, and identifying patients in need of unscheduled proactive outreach to prevent missed care and health decline, and avoid reactive and acute care. Tool credibility was perceived as boosted by specific factors, including integration into a currently used portal and use of established diagnosis and risk-prediction scores. Round two (n=2 PCPs, n=6 RNs, electronic prototype) revealed user perception of the tool as useful in specific contexts: panel management for a group of patients, providing feedback to a patient on why care steps might be recommended, or in preparation for a patient visit. Tool content was seen as desirable if information did not overlap with existing tools, the tool was integrated with the EHR, and the care steps were nonredundant, which current "clinical reminders" staff were already asked to complete during encounters. Credibility was felt to be enhanced by regular and visible updates. In round three of interviews (n=3 PCPs, web-based interactive prototype), users were interviewed as they "walked-through" the tool. They perceived usefulness as related to the type of information displayed and integration with the EHR. The care steps were perceived as desirable to cover gaps in care not previously considered for the patients. Credibility was felt to be related to the transparency of the data origins underlying the care steps. All three rounds led to updates made to the tool content, format, graphics, and text (see Multimedia Appendix 3 for overview of changes, such as views and content updates, terminology revisions, and formatting changes).

Table . Participant characteristics in VET-PATHS[a] development.

Round and role	Type of prototype test	Site/region
User-centered prototype development, n=17		
Round 1 (6 nurses; RN[b])	Paper prototype	Urban VAMC[c] (Northeast)
Round 2 (2 PCPs[d])	Paper prototype	Urban VAMC (Northeast)
Round 2 (6 nurses; RN)	Screenshot visuals with simulated patient data	2 urban VAMCs (Midwest, South)
Round 3 (3 PCPs)	Web-based interactive prototype with actual patient data	2 urban VAMCs (South, West)

[a]VET-PATHS: Veteran Panel Management Tool for High-Risk Subgroups.

[b]RN: registered nurse.

[c]VAMC: Veterans Health Administration medical center–affiliated clinic.

[d]PCP: primary care clinician.

XSL•FO
RenderX

Table . Participant characteristics in VET-PATHS[a] pilot testing.

Teams[b]	Patient panel size	Site/region
Feasibility pilot intervention (n=23; 6 teams)		
General, median (IQR)	342 (192-573)	
A, n	476	Urban VAMC[c] (Northeast)
B, n	862	Rural CBOC[d] (Northeast)
C, n	143	Urban VAMC (Western)
D, n	208	Rural CBOC (Western)
Focused, median (IQR)	108 (105-110)	
E, n	112	Urban VAMC (Northeast)
F, n	103	Urban VAMC (Northeast)

[a]VET-PATHS: Veteran Panel Management Tool for High-Risk Subgroups.

[b]General primary care teams care for heterogeneous patient panels. Focused teams provide medical or topic-specific care to more homogenous patients (eg, patients living with substance use disorders).

[c]VAMC: Veterans Health Administration medical center–affiliated clinic.

[d]CBOC: community-based outpatient clinic.

Feasibility Pilot Intervention

Six teams (23 clinical staff) participated in the pilot testing of VET-PATHS within real-world clinical settings at 4 VHA sites (2 community clinics and 2 VHA medical centers) across two US regions (Tables 1 and 2). Four teams represented typical PACTs in the VHA, providing general primary care to larger panels (median 342, IQR 192-573 total patients per team, 15% of panel appearing as high-risk in VET-PATHS). Two teams provided primary care focused on patient populations with specific medical conditions and had smaller patient panels (median 108, IQR 105-110 total patients, 56% of panel as high-risk within VET-PATHS). Teams conducted between 4 and 8 meetings each across an 8-week period. The mean length of meetings was 24 (range 16‑49) minutes. A range of multidisciplinary roles attended meetings, with PCP Champions attending all meetings. General teams had attendance at most meetings from “core” team members (RN, CA, and MSA), while focused teams more often had other extended PACT members attend (Table 3).

XSL•FO
RenderX

Table . VET-PATHS[a] feasibility results among 6 pilot teams (4 general and 2 focused).

	General					Focused		
	Team A	Team B	Team C	Team D	Mean (SD)	Team E	Team F	Mean (SD)
Site	Urban VAMC[b]	Rural CBOC[c]	Urban VAMC	Rural CBOC	__[d]	Urban VAMC	Urban VAMC	—
Total team panel size, n	476	862	150	411	475 (294)	112	103	108 (6)
HR[e] patients in tool, n	88	91	24	27	58 (32)	60[f]	61	61 (0.5)
HR patients reviewed in meetings, n/N (%)	77/88 (88)	87/91 (96)	23/24 (96)	27/27 (100)	94.7 (4.5)	39/60 (65)	27/61 (44)	54.6 (10.4)
Team meetings, n	8	6	6	5	6.3 (1.1)	6	4	5.0 (1.0)
Time reviewing patients per meeting (min)	21	49	21	17	27.0 (12.8)	16	19	17.6 (1.4)
Meeting attendance by role (%)	100% PCP[g,h] 13% RN[i], 88% MSA[j,h]	100% PCP[h], 100% RN[h], 50% CA[k], 100% MSA[h]	100% PCP[h], 83% RN, 83% CA[h], 67% MSA[h]	100% PCP[h], 100% RN[f], 100% CA, 80% SW[l,h]	74% core, 20% not core	100% PCP[h], 60% Psych[m], 40% Pharm[n], 40% SW	100% PCP[h], 100% Pharm, 100% RN, 100% CA[h]	50% core, 40% not core
Care steps "due" and discussed, n	130	165	36	41	93 (55.9)	43	36	39.5 (3.5)
Total new actions intended, n/N (%)	40/130 (31)	47/165 (28)	21/36 (58)	8/41 (10)	29.5 (16.0)	11/43 (26)	0/36 (0)	—
Of above, additional actions (not care steps), n	12	14	6	1	8.3 (5.6)	3	0	—
Care plans documented, n/N (%)	17/77 (22)	22/87 (25)	11/23 (39)	4/27 (15)	52/214 (24)	2/39 (5)	0/27 (0)	2/66 (3)

[a]VET-PATHS: Veteran Panel Management Tool for High-Risk Subgroups.

[b]VAMC: Veterans Health Administration medical center–affiliated clinic.

[c]CBOC: community-based outpatient clinic.

[d]Not applicable.

[e]HR: high-risk; predicted ≥90th percentile of risk for hospitalization within 1 year.

[f]Number shown in tool exceeds high-risk patients at baseline due to update in patients at the time of first pilot team meeting, after baseline demographic data were pulled.

[g]PCP: primary care clinician.

[h]Provided posttest qualitative interview.

[i]RN: registered nurse.

[j]MSA: medical support administrator (administrative clerk).

[k]CA: clinical associate (eg, licensed practical nurse).

[l]SW: social worker.

[m]Psych: mental health provider (eg, team clinical psychologist or psychiatrist).

[n]Pharm: pharmacist.

Teams actively reviewed 80% (280/351) of the patients visible within the VET-PATHS, with more patients reviewed by general teams (214/230, 93%) than focused teams (66/121, 55%; Table 3). A total of 1343 care steps were possible for the 351 patients within 4 multimorbidity groups, and 596 (44%) care steps were "due" care suggestions for patients. For the 280 patients

XSL•FO
RenderX

reviewed, 451 care steps were "due" and discussed by the team. Teams verbally discussed and intended 127 actions and documented these actions in 76 new care plans (plans were counted once only, per patient) in the EHR among patients reviewed following a team meeting. Of actions, most (91/127, 72%) were based on care steps recommended within the tool; the remainder (36/127, 28%) were additional actions that the teams added after reviewing the patient's information in the tool, for example, a team decision to schedule a return visit for patients not seen recently in primary care. Of 127 actions planned, 69 (54%) actions led to an order placed or the team documenting an action to provide the care to the patient, and 53 (42%) actions were documented as either received (41/53) or declined (12/53) by the patient (Figure 1). Of 127 actions planned, these included return to clinic orders (34/127, 27%), referrals other than to mental health (25/127, 20%), vaccinations (24/127, 19%), other actions (21/127, 16%), referrals to mental health (16/127, 13%), and laboratory orders (7/127, 5%).

Of 280 patients discussed, actions were intended for 91 (32%) unique patients, ordered on 56 (20%), and received by 45 (16%). Of specific care steps within the tool, intended uptake (team discussion with planned action on suggestions from VET-PATHS care steps) varied by multimorbidity group and care step. Intended uptake for each group's care steps ranged from a mean proportion of 11% (12/109 actions intended "due" and discussed by team, Liver Group) to 37% (30/82 actions intended "due" and discussed by team, Substance Use Group), with greater uptake of items with less effort to deliver, such as vaccination or laboratory orders, and less uptake of items requiring more complex decision-making or discussion with patients, such as referrals to palliative care or specialists. When teams opted to take no action on care steps "due" for patients discussed (n=363 care steps), the most common reasons stated in the team discussion were PCPs' judgment that a care step was not a priority for the patient (92/363, 25%) or the patient already was scheduled for care that had not yet occurred (42/363, 12%; Multimedia Appendix 4).

Tool usability surveys were completed by 15 of 19 (79%) participating users. The mean SUS score was 69.9 (SD 7.8). Usability was scored similarly between general and focused teams (general: mean 69.3, SD 7.9; focused: mean 71.3, SD 8.5).

XSL•FO
RenderX

Figure 1. Care steps and actions received by patients during the pilot test. *Twelve of 53 (23%) did not show or declined. FOC: focused team; GEN: general team; MH: mental health.

XSL•FO
RenderX

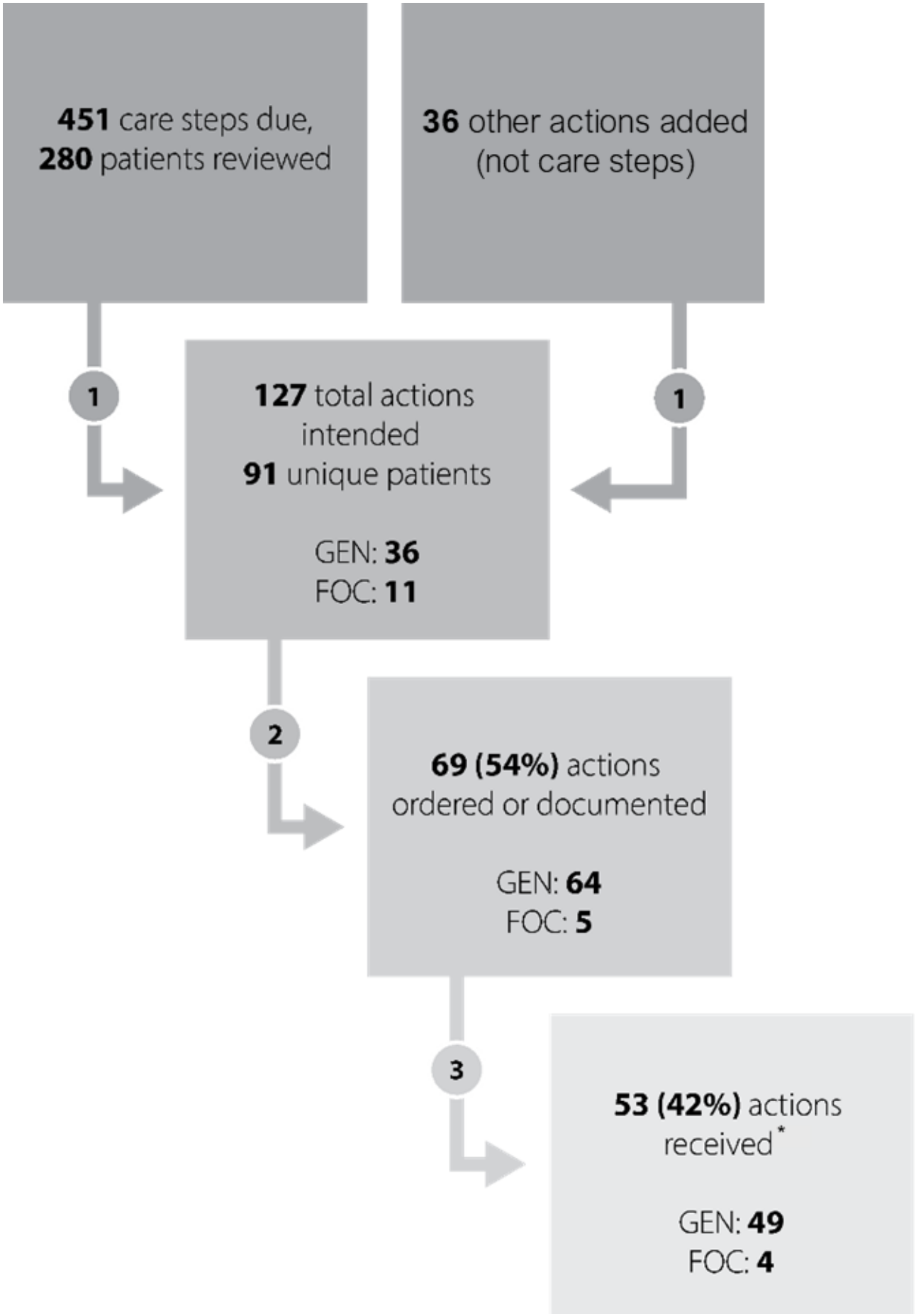

XSL•FO
RenderX

Utilization was examined among patients at high risk on pilot teams compared to nonpilot patients at high risk at the same clinic during similar time periods. Trends in hospitalization, 30-day readmissions, and urgent care use were lower in pilot patients 3 to 6 months post intervention, with increases in primary care (telephone), mental health, and specialty outpatient use (Table 4).

Table . Difference-in-difference (DID) utilization outcomes, high-risk usual care versus VET-PATHS[a] pilot patients.

	Usual care (n=12,230)[b]		VET-PATHS (n=341)[c]		DID: VET-PATHS vs usual care
	Pre (0 - 3 m)	Post (3 - 6 m)	Pre (0 - 3 m)	Post (3 - 6 m)	
Acute hospitalizations	1.4 (0.8)	1.4 (0.9)	1.8 (1.2)	1.6 (1.0)	−0.15
Readmissions (30-d)	1.5 (1.0)	1.5 (1.2)	1.7 (0.9)	1.6 (0.8)	−0.08
ED[d] visits	1.6 (3.2)	1.5 (3.4)	2.0 (3.61)	2.1 (3.5)	0.18
Urgent care visits	0.04 (0.3)	0.05 (0.4)	0.04 (0.2)	0.03 (0.4)	−0.02
Primary care, in-person visits	1.5 (1.8)	1.4 (1.6)	1.4 (1.8)	1.3 (2.1)	0.00
Primary care, phone visits	1.2 (1.8)	1.1 (1.8)	0.9 (1.6)	1.4 (1.7)	0.56
Mental health, all clinic visits	8.4 (14.3)	7.3 (13.2)	7.8 (11.4)	7.8 (15.6)	0.99
Specialty care, all clinic visits	4.9 (4.6)	4.3 (4.4)	5.5 (5.9)	5.5 (5.3)	0.57

[a]VET-PATHS: Veteran Panel Management Tool for High-Risk Subgroups.

[b]Mean age 70.1 (SD 12.9) y; 69% (8239/12,002) White; 8% (935/12,230) female.

[c]Mean age 68.3 (SD 13.7) y; 67% (225/336) White; 12% (41/341) female.

[d]ED: emergency department.

Experiences with the tool and use in team meetings were described by 14 team members within posttest interviews (Table 4). All teams contributed at least one postpilot interview, and all team roles were represented (interview participants are noted in Table 3). Three main themes were identified. First, participants described efficiency related to the tool's grouping of patients:

> *This was new for me and my team to have a way to characterize groups further and have these buckets to think about. […] It helped me get to know some of my patients* [PCP, general team]

Second, teams described the tool, making them aware of care "gaps" that had been previously hard to identify:

> *it focused our efforts on certain aspects of their care that […] maybe we have neglected* [PCP, general team]

Third, participants described VET-PATHS as enabling the whole team's involvement in patient care planning:

> *it definitely helped us to work together as a team to sort through some of our higher risk panel […] to reevaluate some of our Veterans a little bit more thoroughly* [RN, general team]

When examined by teams and roles, general teams overall described high acceptability of the tool content and mixed acceptance of care steps, depending on the type of care step and group. They desired more automatic integration with the primary EHR, for example, one-click placement of orders recommended by care steps. Focused teams reported lower usability of the tool for their more specific patient needs and often felt they had already addressed or did not find the care steps suggested appropriate for their specific patient panels. Acceptability of the tool and specific care steps also varied across roles and teams (Tables 5 and 6). Team members newer to panel management, less familiar with their patients, or serving more heterogeneous patients perceived more utility than those with more established panel management processes or well-known or more homogeneous patient panels. Among roles, PCPs and RNs reported overall high acceptability of the tool, with positive perceptions of the patient grouping functions and particular support for advancing the multidisciplinary team member contributions to care planning. MSAs and CAs described the acceptability of the tool as they perceived it facilitating their more support-based role to the PCP.

XSL•FO
RenderX

Table . Postpilot qualitative interview results by role.

Role	Themes	Example quotes
PCP[a] (n=6)	Helpful to have groups to characterize patients; positive about team-based discussion. Some mixed acceptability of care steps	*Maybe I'm thinking their problems, their core issues lie here. But according to VET-PATHS profiling, they're more of "this flavor," right? And that was helpful to crosscheck […] my characterization of certain patients and their risk.* [PCP GEN[b]]
RN[c] (n=2)	Helpful for organizational process (sorting, tracking) and exposing tasks not previously considered for patients. Both described processes improving team collaboration	*Having the high-risk patients there and what to consider approaching, it's almost like a checklist, like "did we do this, did we do that?" It's nice to have.* [RN FOC[d]]
CA[e] (n=2)	Acceptable if used in conjunction with PCP, to reduce PCP burden	*As the [CA], it falls under my category…to get the provider all the information that she needs at a glance to make her job easier.* [CA FOC]
MSA[f] (n=3)	Acceptable if in conjunction with PCP. Tool increased understanding of patient clinical needs, helped clerks prioritize their tasks according to clinical urgency, which they often did not have insight into	*It's a very helpful… because for us, we schedule the patients and we were getting informed and that the patients' needs are our priorities.* [MSA GEN]
Other (n=1)	Improved ability of team to collaboratively discuss patients	*Was helpful to go over the high-risk patients as a team… it's good to have a list just so nobody falls through the cracks.* [SW[g] FOC]

[a]PCP: primary care clinicians (MD, nurse practitioner).

[b]GEN: general team.

[c]RN: registered nurse.

[d]FOC: focused team.

[e]CA: clinical associate (eg, licensed vocational nurse, care technician).

[f]MSA: medical support administrator (administrative clerk).

[g]SW: social worker.

XSL•FO
RenderX

Table . Postpilot qualitative interview results by team.

Team	Contextual factors	Acceptability/utility	Quotes
Team A (n=2)	• General team, understaffed. Non-PCP[a] staff turnover during pilot • No prior panel management process	Positive in helping identify most at-risk patients and highlighting need; some concerns about documenting care step assignment/tracking.	• *I think it was a nice tool to pull out who are my high-risk patients. I think it helped me identify who had not been seen in some time [.]. [One care step] for a lot of my patients did not seem appropriate [.] but put it in the forefront.* [PCP GEN[b]]
Team B (n=3)	• General team, fully staffed • No prior panel management process	Positive about tool content (especially subgrouping), overall felt context and navigation were acceptable; positive reflection on increase in interdisciplinary discussion.	• *It was helpful in bringing the team together and focusing on Veterans who are medically and psychosocially complicated.* [PCP GEN] • *It broke out very straightforward on the things that we need to concentrate on.* [MSA[c] GEN]
Team C (n=3)	• General team, fully staffed. Recently inherited ~50% of high-risk patients • Significant panel management experience	Positive about tool, felt it was useful for both subgroups and care steps. Some challenges with administrative implementation of care steps (eg, reaching patients post-meeting).	• *Having sort of clear buckets and care steps that are evidence based or evidence informed, I think was really innovative and that's [what] I'm very excited about.* [PCP GEN]
Team D (n=3)	• General team, understaffed. Significant prior experience in panel management	Positive about tool, especially subgrouping and increasing interdisciplinary discussion. Felt care steps less useful, primarily due to feeling these were not applicable or already done for patients.	• *It kind of validated that we were actually on the right focus. Not necessarily that it made us. Like we were already doing it. I felt like it matched what we were doing with the for the patients.* [RN[d] FOC[e]]
Team E (n=1)	• Focused team for specific care population, understaffed • Doing regular previsit huddles, but not full team meetings	Some concerns about applicability of care steps to more specific population, many of whom already received suggested care.	• *It's not that helpful to see a list of your sickest Veterans and then have them just check off all of the boxes.* [PCP FOC]
Team F (n=2)	• Already doing multidisc. team meetings, but not with panel tools • Focused team for specific care population, understaffed	Some concerns about utility of care steps for more specific population, many of whom already received suggested care.	• *The care steps that [were] listed on there, those are things that we'd already done, and we needed to go beyond those initial care steps to what's next.* [CA[f] FOC]

[a]PCP: primary care clinicians.

[b]GEN: general team.

[c]MSA: medical support administrator (administrative clerk).

[d]RN: registered nurse.

[e]FOC: focused team.

[f]CA: clinical associate (eg, licensed vocational nurse, care technician).

Discussion

Principal Findings

VET-PATHS is an innovative informatics tool designed to improve health outcomes for patients with complex multimorbidity, using a team- and data-driven patient grouping approach. In this formative pilot, we found that among participating VHA teams, the unique features of the tool, including the application of population segmentation models to a clinical team's patient panel, automatic identification of data-driven "care step" suggestions, and workflows that enabled multidisciplinary input into patient care, were perceived as useful and desirable during both the user design and preliminary testing. In the preliminary testing with frontline teams, general

XSL•FO
RenderX

primary care teams with variation in staffing, site context, and prior panel management experience reported high VET-PATHS acceptability and utility during facilitated team meetings, and were able to review 44% to 96% of their empaneled patients in the tool over 4 to 8 brief team meetings, planning a change in care for 1 in every 3 patients reviewed, such as new vaccines, laboratory tests, or evidence-based consults. In contrast, focused "specialized" primary care teams had lower acceptability and feasibility. While individual care steps directly prompted by the tool had variable utility, participating teams enacted new care changes not directly from the tool and perceived value in the multidisciplinary discussion and review process, supporting that both the tool use context and content were important. We observed favorable trends in most utilization outcomes for pilot team patients, though this feasibility study was not designed nor powered to evaluate for statistically significant changes in these outcomes. This pilot covers the user-centered design and preliminary testing of a VHA-specific tool developed and evaluated via a learning health system approach (applying health system data patterns to inform current patient care) [48].

Based on the observations and feedback from this pilot, we plan to focus future testing of the tool on VHA general primary care teams with heterogeneous Veteran patient panels, which care for almost 90% of patients at high risk in the VHA [49]. We are currently revising the tool with the clinical group, care step, and functionality updates, including contemporary data-driven multimorbidity groups, expanded and alternative care steps based on perceived relevance and clinical practice guidelines, increased comprehensive care summarized data based on team feedback, and greater explanation of data sources and methodology within the tool. Given variable uptake of some care steps, prompts for shared decision-making and patient interface with care recommendations may increase the usefulness of care steps that require more complex discussion with patients. Furthermore, we note that most of the patients included in the tool are being seen by PCPs regularly. Most "due" care steps represent interventions indicated for a long time, with possible prior recommendation by providers, but not otherwise received. Therefore, even modest uptake of care suggested by the tool among these patients, if translated using a population health perspective, could have significance. For example, 5 of 27 (18.5%) patients "due" for hepatitis B vaccination in the substance use group received it, following tool use. Among the 819,000 Veterans enrolled in the VHA with a prior substance use disorder diagnosis still susceptible to hepatitis B infection [50], similar proportions would lead to over 151,000 additional Veterans immunized.

Comparison to Prior Work

Patients with multimorbidity are increasingly prevalent and costly for health systems to care for [1,2,4]. In the face of primary care staffing shortages [51], there is a growing mismatch between available clinical resources and the effort needed to provide personalized and effective care to patients with multimorbidity [7,8]. Tools to facilitate care planning could improve the efficiency and effectiveness of care for this population, but barriers to use exist. A 2018 survey of 1804 VHA PACT clinicians found that informatics tool use for complex patients was impeded by feelings of being "overwhelmed" by the amount and variety of information displayed and lack of guidance from tools on what patient needs to prioritize [14]. In VHA, as in other health systems, registries of patients predicted to have high risk for adverse events are provided to teams but do not help clinicians identify actions to take to lower risk [15]. Two main takeaways from this work are relevant to other health systems. First, this pilot presents a novel bridging of validated population-level prediction and population segmentation analytics [15,36] with individualized clinical care recommendations in a team-based, primary care intervention [15,27]. Other health systems can use similar methods to segment their patient populations and incorporate the results into their panel management tools or apply similar workflows in their primary care settings. Second, our development and evaluation process is a useful example of tool co-design with frontline health care staff within a learning health system. Few detailed examples spanning early development to real-world testing for digital innovations in primary care have been published [52].

Strengths and Limitations

There are strengths and limitations to this study. User-centered design and acceptability interviews were conducted with interested, participating frontline clinical staff in multiple team roles in VHA medical centers, following the principles of co-design with target users. This process can be replicated outside the VHA, but the specific feedback given by VHA users may not generalize to non-VHA clinical settings. Teams within VHA, however, participated from a variety of clinic settings, staffing levels, and prior panel management experience. Pilot teams were supported by trained facilitators, which may not be pragmatic to offer throughout health systems. However, brief training and "champion user" facilitation are commonly used approaches by health systems helping health care teams adopt new technology tools [53]. Additionally, as previously described, facilitators had key roles in training at the onset, followed by smaller technical assistance roles for subsequent meetings that could be filled by lower-resource methods (eg, on-demand assistance services or more in-depth reference materials). VET-PATHS is focused on Veterans in primary care settings at high risk of hospitalization, so results may not fully generalize to patient populations in other health care settings. While we used several methods to check the accuracy and validity of facilitator observations, we acknowledge that this data could be susceptible to confirmation bias, although we attempted to mitigate this via structured collection templates and dually staffed observations. Given these limitations, the next crucial step is a full, randomized trial of the effectiveness of using VET-PATHS in broader, real-world primary care settings with evaluation of patient and clinician outcomes. Testing tools like VET-PATHS in real-world settings will also build evidence for how panel management tools for patients with multimorbidity can be best integrated into typical primary care team workflows. Finally, we do not have follow-up data on the subsequent patient experience or shared decision-making that will be integral to future capture of the full impact of personalized care planning for primary care patients at high risk.

XSL•FO
RenderX

Conclusions and Future Directions

VET-PATHS, developed through user-centered design and testing, was feasible and acceptable for participating primary care teams in the VHA to use to develop new care plans for their patients with complex multimorbidity and led to several documented clinical impacts, including the delivery of new evidence-based care to patients. VET-PATHS demonstrates a tool and workflow linking population-segmentation analytics and clinical care, a novel preliminary study of an approach for teams managing complex patients at high risk. In our learning health system, multimorbidity group and care step information within VET-PATHS is being iteratively updated, in response to changing EHR data patterns and clinician feedback. Future directions include a planned full effectiveness trial of VET-PATHS with primary care teams and ongoing study of the impact of iterative updates to VET-PATHS.

Acknowledgments

We thank the Veterans Health Administration (VHA) clinicians who participated in these studies and the VHA Clinical Assessment Reporting and Tracking Program office for invaluable programming and support, as well as Mariah Theis (VA Puget Sound, Primary Care Analytics Team) and Nicole Beyer, Terrance Litam, Janke Mains-Mason, and Zachary Michaels (VA Pittsburgh) for their assistance with this study.

Funding

This work was supported by the VHA Primary Care Analytics Team, funded by the VHA Office of Primary Care and the VHA Spark-Seed-Spread Innovation Investment Program. LS is supported by a Veterans Affairs Health Services Research Career Development Award (K2HX003462), with prior support from grant K12HS026369 from the Agency for Healthcare Research and Quality. The views expressed are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs or the US Government

Data Availability

Data for this study were collected as part of a national evaluation effort through the VHA Office of Primary Care. As such, no datasets will be available for dissemination or sharing outside the institution. Analytic plans and detailed methods, outside those shared through scientific publications and conference proceedings, can be made available upon reasonable request.

Authors' Contributions

Conceptualization: AMR
Data curation: FH, XZ, KD, LS, RP, SV
Investigation: LS, KD, KA, AMR, SV
Analysis: LS, KD, TB, XZ, AMR
Methodology: RP, FH, KD, AMR
Funding acquisition: KD, AMR
Supervision: AMR, LS, KD
Administration: KA, KD
Writing - original draft: LS
Writing - review & editing: KD, SV, FH, KA, RP, TB, XZ, AMR

Conflicts of Interest

None declared.

Multimedia Appendix 1
Tool snapshots.
[DOCX File, 1560 KB - formative_v9i1e75081_app1.docx]

Multimedia Appendix 2
Chronic conditions used to define multimorbidity groups.
[DOCX File, 44 KB - formative_v9i1e75081_app2.docx]

Multimedia Appendix 3
User-centered design phase results.
[DOCX File, 45 KB - formative_v9i1e75081_app3.docx]

Multimedia Appendix 4

XSL•FO
RenderX

Per–care step uptake and activity by multimorbidity group (4 of 6 groups with active care steps at the time of the pilot). [DOCX File, 28 KB - formative_v9i1e75081_app4.docx]

References

1. Cassell A, Edwards D, Harshfield A, et al. The epidemiology of multimorbidity in primary care: a retrospective cohort study. Br J Gen Pract 2018 Apr;68(669):e245-e251. [doi: 10.3399/bjgp18X695465] [Medline: 29530918]
2. Fortin M, Stewart M, Poitras ME, Almirall J, Maddocks H. A systematic review of prevalence studies on multimorbidity: toward a more uniform methodology. Ann Fam Med 2012;10(2):142-151. [doi: 10.1370/afm.1337] [Medline: 22412006]
3. Zulman DM, Pal Chee C, Wagner TH, et al. Multimorbidity and healthcare utilisation among high-cost patients in the US Veterans Affairs Health Care System. BMJ Open 2015 Apr 16;5(4):e007771. [doi: 10.1136/bmjopen-2015-007771] [Medline: 25882486]
4. Yoon J, Zulman D, Scott JY, Maciejewski ML. Costs associated with multimorbidity among VA patients. Med Care 2014 Mar;52 Suppl 3(Suppl 3):S31-S36. [doi: 10.1097/MLR.0000000000000061] [Medline: 24561756]
5. Marengoni A, Angleman S, Melis R, et al. Aging with multimorbidity: a systematic review of the literature. Ageing Res Rev 2011 Sep;10(4):430-439. [doi: 10.1016/j.arr.2011.03.003] [Medline: 21402176]
6. Lee TA, Shields AE, Vogeli C, et al. Mortality rate in veterans with multiple chronic conditions. J Gen Intern Med 2007 Dec;22 Suppl 3(Suppl 3):403-407. [doi: 10.1007/s11606-007-0277-2] [Medline: 18026809]
7. Hwang AS, Atlas SJ, Hong J, et al. Defining team effort involved in patient care from the primary care physician's perspective. J Gen Intern Med 2017 Mar;32(3):269-276. [doi: 10.1007/s11606-016-3897-6] [Medline: 27770385]
8. Roskos SE, Fitzpatrick L, Arnetz B, Arnetz J, Shrotriya S, Hengstebeck E. Complex patients' effect on family physicians: high cognitive load and negative emotional impact. Fam Pract 2021 Jul 28;38(4):454-459. [doi: 10.1093/fampra/cmaa137] [Medline: 33367637]
9. Okunogbe A, Meredith LS, Chang ET, Simon A, Stockdale SE, Rubenstein LV. Care coordination and provider stress in primary care management of high-risk patients. J Gen Intern Med 2018 Jan;33(1):65-71. [doi: 10.1007/s11606-017-4186-8] [Medline: 28971306]
10. Whitebird RR, Solberg LI, JaKa MM, et al. Patient experiences and perceptions of care coordination in primary care. J Nurs Care Qual 2024;39(3):239-245. [doi: 10.1097/NCQ.0000000000000759] [Medline: 38198659]
11. Stockdale SE, Rose DE, McClean M, et al. Factors associated with patient-centered medical home teams' use of resources for identifying and approaches for managing patients with complex needs. J Ambul Care Manage 2022;45(3):171-181. [doi: 10.1097/JAC.0000000000000418] [Medline: 35612388]
12. Niles J, Litton T, Mechanic R. An initial assessment of initiatives to improve care for high-need, high-cost individuals in accountable care organizations. Health Affairs Blog. 2019. URL: https://www.healthaffairs.org/do/10.1377/hblog20190411.143015/full/ [accessed 2205-11-26]
13. Effective care for high-need patients: opportunities for improving outcomes, value, and health. National Academy of Medicine. 2017. URL: https://nam.edu/wp-content/uploads/2017/06/Effective-Care-for-High-Need-Patients.pdf [accessed 2025-11-14]
14. McGowan M, Rose D, Paez M, Stewart G, Stockdale S. Frontline perspectives on adoption and non-adoption of care management tools for high-risk patients in primary care. Healthcare (Basel) 2023 Dec;11(4):100719. [doi: 10.1016/j.hjdsi.2023.100719]
15. Nelson KM, Chang ET, Zulman DM, Rubenstein LV, Kirkland FD, Fihn SD. Using predictive analytics to guide patient care and research in a national health system. J Gen Intern Med 2019 Aug;34(8):1379-1380. [doi: 10.1007/s11606-019-04961-4] [Medline: 31011959]
16. Schuttner L, Hockett Sherlock S, Simons C, et al. Factors affecting primary care physician decision-making for patients with complex multimorbidity: a qualitative interview study. BMC Prim Care 2022 Feb 5;23(1):25. [doi: 10.1186/s12875-022-01633-x] [Medline: 35123398]
17. Wheat CL, Gunnink EJ, Rojas J, et al. Changes in primary care quality associated with implementation of the Veterans Health Administration preventive health inventory. JAMA Netw Open 2023 Apr 3;6(4):e238525. [doi: 10.1001/jamanetworkopen.2023.8525] [Medline: 37067799]
18. Baker A, Cronin K, Conway P, DeSalvo K, Rajkumar R, Press M. Making the comprehensive shared care plan a reality. NEJM Catal 2016 May 18. [doi: 10.1056/CAT.16.0838]
19. Coulter A, Entwistle VA, Eccles A, Ryan S, Shepperd S, Perera R. Personalised care planning for adults with chronic or long-term health conditions. Cochrane Database Syst Rev 2015 Mar 3;2015(3):CD010523. [doi: 10.1002/14651858.CD010523.pub2] [Medline: 25733495]
20. Counsell SR, Callahan CM, Clark DO, et al. Geriatric care management for low-income seniors: a randomized controlled trial. JAMA 2007 Dec 12;298(22):2623-2633. [doi: 10.1001/jama.298.22.2623] [Medline: 18073358]
21. Ohar JA, Loh CH, Lenoir KM, Wells BJ, Peters SP. A comprehensive care plan that reduces readmissions after acute exacerbations of COPD. Respir Med 2018 Aug;141:20-25. [doi: 10.1016/j.rmed.2018.06.014] [Medline: 30053968]

XSL•FO
RenderX

22. Aiken LS, Butner J, Lockhart CA, Volk-Craft BE, Hamilton G, Williams FG. Outcome evaluation of a randomized trial of the PhoenixCare intervention: program of case management and coordinated care for the seriously chronically ill. J Palliat Med 2006 Feb;9(1):111-126. [doi: 10.1089/jpm.2006.9.111] [Medline: 16430351]
23. Reeves D, Hann M, Rick J, et al. Care plans and care planning in the management of long-term conditions in the UK: a controlled prospective cohort study. Br J Gen Pract 2014 Sep;64(626):e568-e575. [doi: 10.3399/bjgp14X681385] [Medline: 25179071]
24. Chronic care management toolkit. Centers for Medicare and Medicaid Services. 2024. URL: https://www.cms.gov/files/document/chronic-care-management-toolkit.pdf [accessed 2025-11-21]
25. Sharing the care plan: brief. Agency for Healthcare Research and Quality. 2023. URL: https://www.ahrq.gov/teamstepps-program/curriculum/team/tools/briefs.html [accessed 2024-03-13]
26. Agarwal SD, Basu S, Landon BE. The underuse of Medicare's prevention and coordination codes in primary care: a cross-sectional and modeling study. Ann Intern Med 2022 Aug;175(8):1100-1108. [doi: 10.7326/M21-4770] [Medline: 35759760]
27. Arnold J, Thorpe J, Mains-Mason J, Rosland AM. Empiric segmentation of high-risk patients: a structured literature review. Am J Manag Care 2022 Feb 1;28(2):e69-e77. [doi: 10.37765/ajmc.2022.88752] [Medline: 35139299]
28. Yan S, Kwan YH, Tan CS, Thumboo J, Low LL. A systematic review of the clinical application of data-driven population segmentation analysis. BMC Med Res Methodol 2018 Nov 3;18(1):121. [doi: 10.1186/s12874-018-0584-9] [Medline: 30390641]
29. Chong JL, Lim KK, Matchar DB. Population segmentation based on healthcare needs: a systematic review. Syst Rev 2019 Aug 13;8(1):202. [doi: 10.1186/s13643-019-1105-6] [Medline: 31409423]
30. Vuik SI, Mayer EK, Darzi A. Patient segmentation analysis offers significant benefits for integrated care and support. Health Aff (Millwood) 2016 May 1;35(5):769-775. [doi: 10.1377/hlthaff.2015.1311] [Medline: 27140981]
31. Rosland AM, Nelson K, Sun H, et al. The patient-centered medical home in the Veterans Health Administration. Am J Manag Care 2013 Jul 1;19(7):e263-e272. [Medline: 23919446]
32. Wang L, Porter B, Maynard C, et al. Predicting risk of hospitalization or death among patients receiving primary care in the Veterans Health Administration. Med Care 2013 Apr;51(4):368-373. [doi: 10.1097/MLR.0b013e31827da95a] [Medline: 23269113]
33. Hutchins F, Thorpe J, Zhao X, Zhang H, Rosland AM. Two-year change in latent classes of comorbidity among high-risk Veterans in primary care: a brief report. BMC Health Serv Res 2022 Nov 12;22(1):1341. [doi: 10.1186/s12913-022-08757-x] [Medline: 36371216]
34. Hutchins F, Thorpe J, Maciejewski ML, et al. Clinical outcome and utilization profiles among latent groups of high-risk patients: moving from segmentation towards intervention. J Gen Intern Med 2022 Aug;37(10):2429-2437. [doi: 10.1007/s11606-021-07166-w] [Medline: 34731436]
35. Hutchins F, Rosland AM, Zhao X, Zhang H, Thorpe JM. The impact of dual VA-Medicare use on a data-driven clinical management tool for older Veterans with multimorbidity. J Am Geriatr Soc 2024 Jan;72(1):69-79. [doi: 10.1111/jgs.18608] [Medline: 37775961]
36. Prenovost KM, Fihn SD, Maciejewski ML, Nelson K, Vijan S, Rosland AM. Using item response theory with health system data to identify latent groups of patients with multiple health conditions. PLOS ONE 2018;13(11):e0206915. [doi: 10.1371/journal.pone.0206915] [Medline: 30475823]
37. Parekh AK, Goodman RA, Gordon C, Koh HK, HHS Interagency Workgroup on Multiple Chronic Conditions. Managing multiple chronic conditions: a strategic framework for improving health outcomes and quality of life. Public Health Rep 2011;126(4):460-471. [doi: 10.1177/003335491112600403] [Medline: 21800741]
38. Hutchins F, Schuttner L, Zhang H, et al. Using health system data to prioritize primary care strategies to lower risk of hospitalization for high risk/complex patients. Presented at: 2024 Annual Research Meeting; Jun 29 to Jul 2, 2024 URL: https://vmx.m-anage.com/us/2024arm/en-US/presentation/652477 [accessed 2025-11-21]
39. Conners EE, Panagiotakopoulos L, Hofmeister MG, et al. Screening and testing for hepatitis B virus infection: CDC recommendations—United States, 2023. MMWR Recomm Rep 2023 Mar 10;72(1):1-25. [doi: 10.15585/mmwr.rr7201a1] [Medline: 36893044]
40. Abstracts from the 2021 Annual Meeting of the Society of General Internal Medicine. J Gen Intern Med 2021 Jul;36(S1):1-469. [doi: 10.1007/s11606-021-06830-5]
41. Morville P. User experience design. Semantic Studio. 2004. URL: https://semanticstudios.com/user_experience_design/ [accessed 2020-10-31]
42. Brooke J. System usability scale (SUS): a quick-and-dirty method of system evaluation user information. In: Jordan PM, Thomas B, McClelland IL, Weerdmeester B, editors. Usability Evaluation in Industry, 1st edition: CRC Press; 1996:252. [doi: 10.1201/9781498710411]
43. Sauro J. Measuring usability with the System Usability Scale (SUS). MeasuringU. 2011. URL: https://measuringu.com/sus/ [accessed 2024-08-09]

XSL•FO
RenderX

44. Sekhon M, Cartwright M, Francis JJ. Acceptability of healthcare interventions: an overview of reviews and development of a theoretical framework. BMC Health Serv Res 2017 Jan 26;17(1):88. [doi: 10.1186/s12913-017-2031-8] [Medline: 28126032]
45. Hamilton AB, Finley EP. Qualitative methods in implementation research: an introduction. Psychiatry Res 2019 Oct;280:112516. [doi: 10.1016/j.psychres.2019.112516]
46. Leon AC, Davis LL, Kraemer HC. The role and interpretation of pilot studies in clinical research. J Psychiatr Res 2011 May;45(5):626-629. [doi: 10.1016/j.jpsychires.2010.10.008] [Medline: 21035130]
47. Lancaster GA, Dodd S, Williamson PR. Design and analysis of pilot studies: recommendations for good practice. J Eval Clin Pract 2004 May;10(2):307-312. [doi: 10.1111/j..2002.384.doc.x] [Medline: 15189396]
48. Greene SM, Reid RJ, Larson EB. Implementing the learning health system: from concept to action. Ann Intern Med 2012 Aug 7;157(3):207-210. [doi: 10.7326/0003-4819-157-3-201208070-00012] [Medline: 22868839]
49. Chang ET, Zulman DM, Nelson KM, et al. Use of general primary care, specialized primary care, and other veterans affairs services among high-risk veterans. JAMA Netw Open 2020 Jun 1;3(6):e208120. [doi: 10.1001/jamanetworkopen.2020.8120] [Medline: 32597993]
50. Hoggatt KJ, Harris AHS, Washington DL, Williams EC. Prevalence of substance use and substance-related disorders among US Veterans Health Administration patients. Drug Alcohol Depend 2021 Aug 1;225(108791):108791. [doi: 10.1016/j.drugalcdep.2021.108791] [Medline: 34098384]
51. Zhang D, Son H, Shen Y, et al. Assessment of changes in rural and urban primary care workforce in the United States from 2009 to 2017. JAMA Netw Open 2020 Oct 1;3(10):e2022914. [doi: 10.1001/jamanetworkopen.2020.22914] [Medline: 33112401]
52. Kilfoy A, Hsu TCC, Stockton-Powdrell C, Whelan P, Chu CH, Jibb L. An umbrella review on how digital health intervention co-design is conducted and described. NPJ Digit Med 2024 Dec 23;7(1):374. [doi: 10.1038/s41746-024-01385-1] [Medline: 39715947]
53. Santos WJ, Graham ID, Lalonde M, Demery Varin M, Squires JE. The effectiveness of champions in implementing innovations in health care: a systematic review. Implement Sci Commun 2022 Jul 22;3(1):80. [doi: 10.1186/s43058-022-00315-0] [Medline: 35869516]

Abbreviations

CA: clinical associate
CAN: Care Assessment Need
CBOC: community-based outpatient clinic
EHR: electronic health record
MSA: medical support administrator
PACT: Patient Aligned Care Teams
PCP: primary care clinician
RN: registered nurse
SUS: System Usability Scale
VAMC: VHA medical center–affiliated clinic
VET-PATHS: Veteran Panel Management Tool for High-Risk Subgroups
VHA: Veterans Health Administration

Edited by J Sarvestan; submitted 27.03.25; peer-reviewed by M Kanter; accepted 24.10.25; published 04.12.25.

Please cite as:
Schuttner L, Daniels K, Venkatesh S, Hutchins F, Atchison K, Piegari R, Bober T, Zhao X, Rosland AM
A Novel Primary Care Planning Informatics Tool Informed by Data-Driven Multimorbidity Grouping: User-Centered Design and Feasibility Testing
JMIR Form Res 2025;9:e75081
URL: https://formative.jmir.org/2025/1/e75081
doi:10.2196/75081

© Linnaea Schuttner, Karin Daniels, Srinidhi Venkatesh, Franya Hutchins, Karley Atchison, Rebecca Piegari, Timothy Bober, Xinhua Zhao, Ann-Marie Rosland. Originally published in JMIR Formative Research (https://formative.jmir.org), 4.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium,

XSL•FO
RenderX

provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Mobile-Based Ecological Momentary Intervention for Improving Physical Activity in Adults Without Regular Physical Activity: Pilot Randomized Controlled Trial

Takeyuki Oba[1*], PhD; Chihiro Moriishi[1*], PhD; Keisuke Takano[1], PhD; Kentaro Katahira[1], PhD; Kenta Kimura[1,2], PhD

[1]Human Informatics and Interaction Research Institute, National Institute of Advanced Industrial Science and Technology, 1-1-1 Higashi, Tsukuba, Ibaraki, Japan

[2]Faculty of Letters, Hosei University, Chiyoda-ku, Japan

[*]these authors contributed equally

Corresponding Author:
Keisuke Takano, PhD
Human Informatics and Interaction Research Institute, National Institute of Advanced Industrial Science and Technology, 1-1-1 Higashi, Tsukuba, Ibaraki, Japan

Abstract

Background: The ecological momentary intervention (EMI) is one of the most promising digital—primarily mobile—interventions to enhance physical activity (PA) and other health behaviors. It is a combination of ecological momentary assessment (EMA), in which participants are prompted to indicate their momentary states and ongoing behaviors in daily life, and the just-in-time delivery of interventions tailored to the EMA responses. The EMI has typically been implemented in message-based interventions (eg, activity recommendations tailored to users' physical locations), but its efficacy and feasibility have not been sufficiently established because of the variability in design and implementation.

Objective: This pilot, two-arm, parallel-group randomized controlled trial aimed to be an exemplar of EMI for improving PA and establishing efficacy and feasibility among adults without a habit of PA.

Methods: A total of 40 participants (23 women; mean age 45.40, SD 10.50 years) were recruited from among community dwellers in northeast Japan and randomly allocated to the EMI or control group. Each participant wore an activity tracker to monitor their daily step count and heart rate (HR) for 4 weeks (fully automated). Simultaneously, they responded to EMA questions about the current weather, location, and social context 3 times during the daytime and an additional evening question about motivations for PA each day. Only the EMI group received messages tailored to their responses to EMA, recommending more active alternative behaviors suited to EMA-reported contexts.

Results: Participants wore a Fitbit device for 90.3% (21.66/24 hours per day) of the study period (mean 90.3, SD 10.0), and no dropouts were observed. The EMI group showed no significant improvement in the self-reported amount of PA (P=.44), step count (P=.24), or motivation for PA (from P=.11 to P=.91) compared with the control group. However, the EMI group showed a significantly larger increase in the minutes of 40% HR reserve, a measure of moderate or high intensity of PA (mean 16.03, 95% CI 3.76-28.29; Cohen d=0.20 - 0.41; P=.02 for the follow-up weeks). The intervention was rated as marginally useful and satisfactory, and approximately half of the participants expressed a willingness to continue the intervention. The timing of the EMA prompting was considered inappropriate.

Conclusions: These findings suggest that the EMI with messages tailored to EMA-reported contexts was not effective in increasing the amount or motivation for PA but may increase the intensity as assessed by the HR. The intervention aimed to help individuals implement small but slightly more active behaviors in their daily routine, which may not accompany prominent body movements but may be reflected in the increased HR. Marginal feasibility indicates that the intervention has sufficient scope for improvement, particularly in terms of prompt timing.

Trial Registration: OSF Registries osf.io/4zcvx; https://osf.io/4zcvx/overview

(JMIR Form Res 2025;9:e79360) doi:10.2196/79360

KEYWORDS

ecological momentary intervention; ecological momentary assessment; physical activity; motivation; heart rate reserve

XSL•FO
RenderX

Introduction

Background

The rapid growth of mobile technologies has expanded the accessibility and applicability of health behavior change interventions that are delivered remotely, automatically, and ubiquitously to users in need. The ecological momentary intervention (EMI) is a mobile intervention that offers real-time support tailored to the dynamically changing states of users in daily life [1,2]. It often incorporates ecological momentary assessment (EMA) to determine the support offered at each moment, for which users make real-time or near–real-time self-reports via a mobile device about their current behavior, context, emotions, beliefs, and experiences [3]. An EMI system then selects the seemingly best intervention according to the EMA responses; for example, a system generates a message recommending outdoor activities to enhance physical activity (PA) when the weather is good and a user reports feeling energetic. A systematic review [1] found that coping strategy suggestions, including behavioral activation (eg, doing a short productive chore, such as loading the laundry or cleaning a room) [4] and restructuring negative thoughts [5], are the most typical interventions offered by EMIs, followed by motivational (eg, positive reinforcement and supportive messages) and information (eg, customized graphs) feedback [1]. For instance, an EMI trial delivered congratulatory and supportive messages when participants reported, in response to an EMA prompt, that they had successfully enacted a healthy behavior (eg, avoided sugar-sweetened beverages) [6]. Another trial [7] delivered 3 text messages each day on how to replace sitting time with a more active behavior (eg, standing up 5 times) with the sending schedules as well as the message contents and tones being adjusted to individual participants' preferences. Participants provided real-time feedback about how they responded to each message, which was used to update the tailoring rule to generate messages. To date, various forms of implementations, designs, and system architectures have been explored and tested across trials, targeting different health issues, including dietary behavior, PA, and substance use [8-11]. Studies have provided preliminary support for EMI's feasibility and efficacy; however, EMI remains an emerging area of research that calls for more empirical evidence [1].

The EMI has gained particular attention from researchers and practitioners working on PA. It is widely recognized that PA reduces the risk of serious health conditions [12], but a lot of people still fail to initiate and maintain PA. Providing health information may not suffice, and theories have articulated key aspects to trigger health behavior change. The capability, opportunity, motivation, and behavior (COM-B) model is one of the most pronounced frameworks, highlighting the following three determinants of PA: (1) physical and psychological capability (eg, having sufficient physical stamina to sustain PA and understanding why PA helps with health management), (2) opportunity (eg, having safe environments for PA and economic stability and receiving support from family and friends), and (3) motivation (eg, considering the long-term health benefits of PA and reducing emotional barriers like fear, stress, or fatigue) [13]. A health behavior change is triggered when these 3 are successfully implemented, and EMI is an ideal design to achieve it. This includes sending a prompt when a moment of opportunity arises, and prompting can be repeated as much as needed with almost no cost for human experts. Additionally, the adaptive nature of EMI concerns the appropriateness of the prompt, which is tailored to the moment or situation that the users are in (eg, not recommending outdoor activities during bad weather). Such tailoring would improve the acceptability of prompts and efficiently motivate individuals to engage in healthy behavior [14].

Regardless of its theoretical advantages, mixed evidence has been found for the efficacy of EMI for PA [15]. The EMI trials, typically designed to remind participants of their current (inactive) states and alternative active behaviors, have established positive effects on proximate outcomes, such as awareness [16], motivation, capability [17], and reduced computer use associated with sedentary behavior [18]; however, these effects did not transfer to PA; participants became more motivated, capable, and aware of opportunities, but their PA levels remained unchanged even after completing the intervention [19].

An important question arises here: How can EMI have an effect on a distal physical outcome? We kept two key aspects in mind when conceptualizing this study. First, regarding the intervention design, we followed Van Dantzig et al [20], who highlighted three principles for a successful EMI: (1) the messages should be scheduled to shoot actionable or reflective moments of the participants; (2) the message content and recommendations should be relevant and feasible given the current context and state of the user; and (3) the message format and delivery channels should fit the real-time, anytime-anywhere nature of the intervention (eg, concise and digestive texts and easy instructions resulting in an immediate reaction). These aspects were implemented in the EMI. Messages were sent 3 times a day (morning, afternoon, and evening) to increase the likelihood of accessing a moment of opportunity. Each message included activity recommendations tailored to the participants' weather and location, which have been used as tailoring variables in former just-in-time adaptive intervention (JITAI) trials [21]. In addition, we incorporated the participants' social contexts into the EMA to tailor the messages, encouraging participants to work together if someone else was present at the moment. Studies have shown that social components (eg, collaborating with or competing against peers) significantly contribute to behavior change interventions [22] but have rarely been implemented in EMI. We expected that the combination of these three tailoring variables (ie, location, weather, and social context) would allow for capturing individual participants' contexts in higher resolution than former trials, which is, thereby, an opportunity to tailor each recommendation message to fit in individual contexts more closely. Also, the social addition would help participants become aware of the opportunities for using social resources to initiate and maintain their engagement with PA [21].

Second, we focused on the fact that the effect of EMI may not always be reflected in changes in the amount of PA or step count. The amount (often assessed by an accelerometer) is not the only determinant of PA intensity. For example, a study

XSL•FO
RenderX

suggested that the percentage heart rate reserve (HRR), which is closely related to oxygen consumption and PA intensity [23,24], is significantly higher during occupational than leisure time PA. Furthermore, the percentage HRR shows a pronounced difference between sitting and standing, but no prominent difference between standing and moving or walking [25]. Existing EMIs typically suggest that participants implement small changes in daily behaviors requiring less effort (eg, walking faster and standing when a long sedentary time is detected) [15,26]. In other words, EMIs are not necessarily designed to increase the amount of PA or steps; however, such small behavioral changes may induce cardiovascular responses, interpreted as increased PA intensity. Therefore, we included heart rate (HR) as an exploratory outcome to test the efficacy of the EMI.

Objectives

This study was a pilot, 2-arm, parallel-group, randomized controlled trial testing the efficacy and feasibility of message-based EMI for improving PA among adults without a PA habit. Our EMI was characterized by messages and activity recommendations tailored to participants' responses to momentary contextual questions embedded in the EMA. In response to each EMA prompt, the participants reported the weather, location, and social interaction with others at that moment, and this information was used to immediately generate a message and recommendation delivered on a smartphone. We hypothesized that participants in the EMI group would show larger increases in the amount and intensity of PA as well as motivations for PA (as assessed by self-report questionnaires and wearable activity trackers) than those in the control group, who received real-time feedback on step count and HR but no tailored messages. In the exploratory analyses to clarify the appropriateness of the EMI recommendations, we tested (1) how likely participants were to accept the activity recommendations, (2) what activities were most likely to be accepted, and (3) whether the frequency of acceptance was associated with an increase in PA and motivation.

Methods

Participants

The participants were recruited via online advertisements and emails distributed among the inhabitants of Tsukuba City and its environs (a north-middle region in Japan) from November 2024 to February 2025. Emails were sent to those who had been registered in a database of potential participants. The advertisement informed participants that the study would take place at the National Institute of Advanced Industrial Science and Technology and assess psychological and behavioral dynamics in daily life. A total of 232 individuals were assessed for eligibility, of which 40 (mean age 45.40, SD 10.50 years) were enrolled. Recruitment occurred as part of an overarching project that encompassed smaller trials to test the feasibility and efficacy of different mHealth interventions [27]. For this trial, recruitment was combined with that for another trial [28], which required participation with a family or friend as a pair. Participants who were not eligible for that trial or could not find a time that worked for each of the pairs were invited to this trial. The inclusion criteria for this study were the same as those of the other trials in the project (see the paper by Takano et al [27] for the complete list): aged 18 - 64 years, expecting no major life changes right before or during the study, not pregnant or lactating, in good health and receiving no medication (eg, tranquilizers or antihypertensive drugs), and not exercising regularly. We assumed that participants had good internet literacy, and necessary support was provided at the briefing session by a researcher. All participants were paid for their participation.

The sample size was determined pragmatically by considering the resources and capacity of the recruiting institute. With no suitable previous information on the expected size of the intervention effect when the study was conceptualized, we decided not to perform a priori power calculations. However, meta-analytic studies have estimated the effect size of text message interventions for PA as Cohen d=0.30 - 0.40 (with or without message tailoring) [29]. A pilot trial reported an effect size of Cohen d=0.90, favoring personalized messages tailored to users' PAs tracked by an app [26,30]. With the current sample size (n=40), an effect of Cohen d>0.91 can be detected reliably (with α=.05 and power=0.80). Although this size may be somewhat overly optimistic, rules of thumb suggest sampling n=30 - 40 as the minimum for a pilot trial [31,32]; and if the main trial is considered together, the sample size for a pilot trial could be even smaller [33].

Randomization and Blinding

Eligible participants were randomly allocated to either the intervention or control group in a 1:1 ratio. Block randomization was used (permutations within a batch of 4 participants) to balance the sample size between the 2 groups. A random allocation sequence was generated by a project member using R before recruitment. Experimenters (who conducted the recruitment and assessments) were not involved in this process. Single blinding was used; all participants received the same instructions (participants allocated to the control condition were also told that they might receive messages after responding to a prompt) and did not know what condition they were assigned to. Experimenters handed lab smartphones (programmed with the prompt schedules and messages to deliver) to participants, so that they knew the condition allocations. Yet, messages were automatically delivered, and researchers were not involved in the actual intervention.

Interventions

Each participant received a smartphone and a Fitbit device at the beginning of the study and was asked to wear the Fitbit device on their nondominant wrist for 24 hours. The Fitbit device was linked to the phone via the Fitbit app, via which participants could monitor their step count and other physical data (eg, HR) at any time. Each evening, the participants were asked to self-report their daily step count and motivation for PA on an EMA app (m-Path; KU-Leuven [34]). The self-reported step count was meant to be a manipulation of self-monitoring and feedback (behavior change techniques taxonomy v1 [35]), controlling for the variability in attention (or preventing participants from paying no attention) to the PA and activity tracker.

XSL•FO
RenderX

Over the 4-week study duration, each participant was prompted by the EMA app to indicate (1) the current location (at home vs work or school vs other), (2) weather (sunny or cloudy vs rainy or snowy), and (3) social context (alone or with a family member, friend, colleague, or other). The EMA prompts were sent 3 times a day (at 8 AM, 12 PM, and 5 PM), and the participants could respond until the next prompt arrived. Only the EMI group received tailored messages upon completing each EMA questionnaire. This EMI began in the second week of the study period (the first week served as the baseline, with no EMI).

Messages were generated by combining sentences and phrases tagged with particular responses to the EMA questions. Specifically, when a participant indicated that the weather was nice, the EMI system used the stem sentence (Why not go out for some light exercise that…), followed by:

- … you can do when the weather is good? (at home)
- … you can do at work or school when the weather is good? (at the workplace or school)
- … can be done anywhere when the weather is good? (at other places)

Similarly, when a participant indicated that the weather was not good, the stem (Why not incorporate some light exercise that…) was combined with

- … you can do at home? (at home)
- … you can do at work or school? (at the workplace or school)
- … can be done anywhere? (other places)

Each message was followed by a mention of the social context as follows:

- Exercising with your family, friends, and colleagues is also effective in forming an exercise habit (being with a family, friend, or colleague).
- The following activities are as strenuous as or more strenuous than walking (being with other people or being alone).

For example, when a participant indicated that the weather was nice and they were alone at home, a message was generated as follows: “Why not go out for some light exercise that you can do when the weather is good? The following activities are strenuous or more strenuous than walking: (1) weeding the garden, cleaning around the entrance, (2) walking to the nearest park, and (3) hanging the futon (bed mattress) out in the sun, changing the bedsheet or linen.

The participants were presented with a set of activity recommendations that were deemed suitable for the indicated context (eg, outdoor activities recommended for good weather response; see Table 1), and the app asked them to indicate any activity that they would like to try then. If the participants were too busy to exercise at that moment, they could choose an option of “not now, maybe later” or “try another activity” for the sake of users’ autonomy [18]. The recommended activities were selected from an existing activity list [36] with known metabolic equivalents (METs), some of which were slightly modified to fit the EMI context (eg, “Walking for pleasure,” rephrased as “Walking around office or school, or shopping at a store that is a bit far”). We selected activities that (1) were not too hard, ranging from 3.0 to 7.0 hourly metabolic equivalents (METs-h/w), (2) could be performed either at home or at the office or school, and (3) were not too time-consuming and could be completed in 30 minutes or less. Also, all these activities could be performed alone or together with someone else; therefore, we did not select the type of activity depending on the social context specified in EMA (simply recommended co-working on the activity that fits the location and weather whenever a family or friend was present).

XSL•FO
RenderX

Table . Recommended activities for different locations and weather conditions.

Location	Weather	Activity[a]
Home	Sunny or cloudy	• Weeding the garden, cleaning around the entrance (3.5 METs[b]). • Walking to the nearest park (3.5 METs). • Hanging the futon (bed mattress) out in the sun, changing the bedsheet or linen (3.3 METs).
Home	Rainy or snowy	• Vacuuming (3.3 METs). • Scrubbing the bathroom and bathtub (3.5 METs). • General kitchen activity (eg, cooking, washing dishes, and cleaning up; 3.3 METs).
Work or school	Sunny or cloudy	• Walking around the office or school, shopping at a store that is a bit far (3.5 METs). • Workout in the nearest park or playground, with or without equipment (eg, park bench workout; 3.8 METs). • Jogging (7 METs).
Work or school	Rainy or snowy	• Walking fast in the office or school hallways (3.5 - 4.3 METs). • Going up and down the stairs (4 METs). • Organizing or cleaning your desk area in the office (4.8 METs).
Other	__[c]	• Using the stairs (instead of an elevator or escalator; 4 METs). • Taking a short detour to the destination (3.5 METs). • Walking faster than usual (3.5 - 4.3 METs).

[a]The activity recommendations were not tailored to different social contexts, as each listed activity could be performed alone or with someone else.

[b]MET: metabolic equivalent.

[c]For the other locations, weather conditions were not used to tailor the activity recommendations. Instead, the recommendations were meant to be tips to increase PA when traveling from one place to another (eg, on the way home).

Measures

We conducted 2 types of measurements, namely pre- and postintervention assessments, and the EMA. The primary outcome was self-reported PA before and after the intervention. Secondary outcomes included Fitbit measures (step count and HR) and motivations for PA. Self-reported PA was treated as the primary outcome because (1) the International Physical Activity Questionnaire—Short Form (IPAQ-SF) was a widely used measure that secures good comparability with other trials in the research field; and (2) we expected that the EMI would enhance motivation for PA, which would be more likely to be transferred to subjective than objective PA. Self-reported PA is not always consistent with the objective (eg, accelerometer-based) PA [37], and is known to be vulnerable to social desirability and other reporting biases [38]. Therefore, we included the Fitbit measures as the secondary outcomes to mitigate these limitations or even to cover the aspects that the subjective measure could not detect (eg, small changes that cannot be perceived).

International Physical Activity Questionnaire—Short Form

Subjective PA levels were assessed as the primary outcome using IPAQ-SF [39,40]. This self-report scale has three activity categories: (1) walking, (2) moderate-intensity activity, and (3) vigorous-intensity activity (the inactivity category or sedentary time was not used in this study). The participants were asked to report the number of days and hours spent on each category in a typical week. The reported weekly hours of activities (days×hours) were aggregated across the 3 categories to represent total PA in the form of hourly metabolic equivalents (METs-h/w).

Self-Determined Motivation Scale for Exercise

Motivation for PA was assessed as the secondary outcome pre- and postintervention using the revised version of the Self-Determined Motivation Scale for Exercise (SMSE) [41]. This scale has 22 items representing 6 types of motivations and regulatory styles for PA and exercises with the self-determination theory as the theoretical basis [42]. Some items were as follows: for intrinsic motivation, “exercising itself is fun;” for external regulation, “I exercise because other people

XSL•FO
RenderX

will be pleased with me;" for introjected regulation, "I feel guilty if I do not exercise;" for identified regulation, "I think it is a good way to improve myself;" for integrated regulation, "it is essential to my identity and sense of self;" and for amotivation, "I do not know why I exercise." Each item was rated on a 5-point scale (1="Not at all true" and 5="Very true"), and the sum of the scores was calculated for each motivation type. The Cronbach alphas for the current data were >0.64.

Stage-of-Change Questionnaire

To assess the readiness for PA at baseline, we used the Stage-of-Change Questionnaire for PA [43,44]. This questionnaire asked participants to indicate the most applicable statement among the following 5 statements: "I currently do not exercise and do not intend to start exercising in the future" (precontemplation); "I currently do not exercise but I am thinking about starting to exercise in the next six months" (contemplation); "I currently exercise some but not regularly" (preparation); "I currently exercise regularly but have only begun doing so within the last six months" (action); and "I currently exercise regularly and have done so for longer than six months" (maintenance).

Fitbit Measures

Daily step count and HR were assessed using a Fitbit device (Sense or Sense 2). Valid wear time was estimated by counting the frequency of valid HR values recorded (in minutes) over the 4-week study period. The HR was also analyzed as an outcome measure to evaluate the efficacy of this EMI, but this was not included in the preregistration protocol. We decided to analyze this as a secondary outcome because our EMI-recommended activities were physically more intensive than a sedentary daily routine but did not necessarily increase the step count (eg, kitchen activity). In processing the HR data, we first determined the target HR ranges with a threshold of 40% HRR or higher for each participant and then indexed the intensity of PA with the minutes of HR exceeding the threshold each day. The guidelines of the American College of Sports Medicine [45] classify 40% - 59% HRR as moderate intensity (METs 3 - 5.9), which corresponds to the target PA intensity this EMI aimed to encourage. The threshold of 40% HRR was calculated using the following equation: (resting HR + [maximum HR–resting HR] × 0.4). The maximum HR was estimated using the Tanaka equation (208–participants' age in years ×0.7) [46], in which the resting HR was the average defined by the Fitbit over the study period.

Modified Situational Motivational Scale

The participants were prompted to respond to an EMA signal each evening (at 7 PM) in addition to the 3 EMA signals in the morning and afternoon. The evening EMA included items on current motivations for PA, each representing identified regulation (ie, "I'm feeling that exercising is important for me"), intrinsic motivation (ie, "I'm feeling that exercising is fun"), and external motivation (ie, "I'm feeling that I have to exercise"). These items were adapted from the Situational Motivational Scale [47] for physical exercise [48], with the wording modified for the EMA [27]. Each item was rated on a 7-point scale (1="Not at all" and 7="Very much").

Feasibility

At the postintervention assessment, participants in the EMI group were asked to rate the intervention as follows: the recommended activities and exercises were (1) satisfactory, (2) useful, (3) helpful, (4) suited to me, (5) conducted at an appropriate time, (6) practical, and (7) "I would like to continue using this intervention system" [49]. Each item was rated on a 7-point scale (1="Not at all" and 7="Very much").

Procedure

Figure 1 shows the study flowchart. Each eligible participant was invited to the laboratory, where they received the relevant study information and provided written informed consent. At this time, they completed the preintervention assessment, including the IPAQ-SF and SMSE-2. The day after this briefing session, the participants started the baseline week (week 1), during which they wore a Fitbit and responded to 3 daytime and 1 evening EMA prompts. The EMI started in week 2 and lasted until week 4; only the intervention group received a tailored message upon completing the EMA. The control group continued the EMA for this 3-week intervention period, but did not receive any messages from the app. The EMA prompts and EMI messages were sent automatically. Technical support was offered by a researcher whenever appropriate (eg, for app login and online questionnaires). After week 4, the participants completed the postintervention assessment for the outcomes (IPAQ-SF and SMSE-2) and feasibility (only the intervention group). The postintervention assessment took place 1.35 days (range 0 - 9) after the completion of the intervention.

XSL•FO
RenderX

Figure 1. Study flowchart. EMI: ecological momentary intervention; EMA: ecological momentary assessment; HRR: heart rate reserve; IPAQ: International Physical Activity Questionnaire; SMSE: Self-Determined Motivation Scale for Exercise.

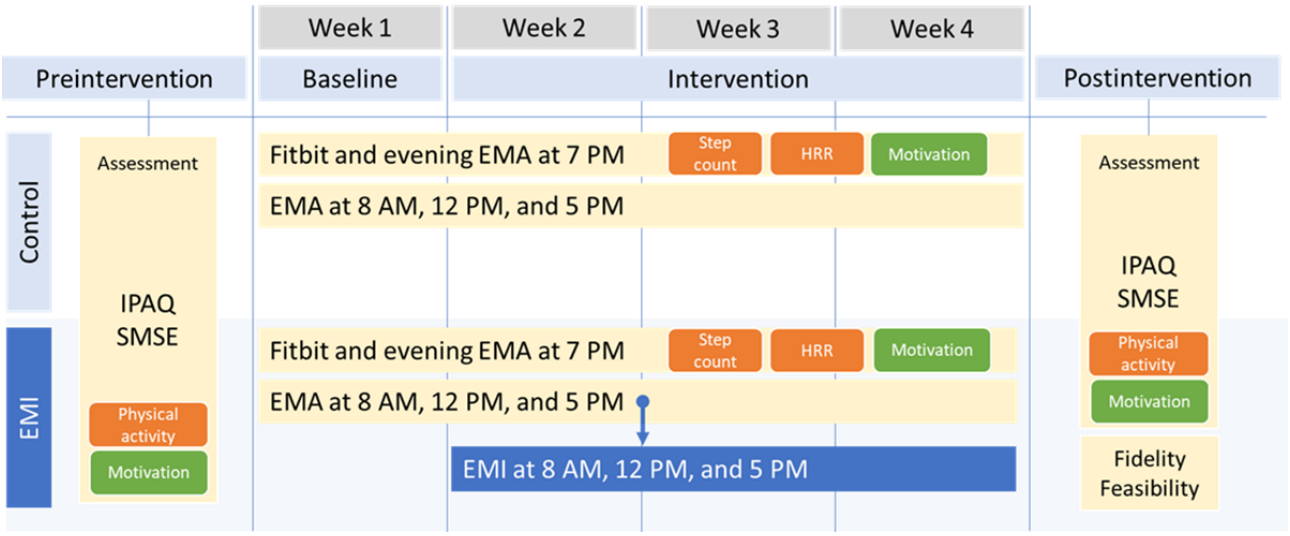

Ethical Considerations

The protocol of the overarching study that included this study was approved by the Ethics Committee of the National Institute of Advanced Industrial Science and Technology (ID 2022 - 1240). The protocol used in this study is available in the Open Science Framework [50]. We adhered to the CONSORT-EHEALTH (Consolidated Standards for Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) guidelines (see Checklist 1) [51] when reporting the study findings.

The participants received information about the procedure and aims of the study as well as their rights before completing the preintervention assessment. They were also explicitly informed that they had the right to withdraw from the study without any consequences. Subsequently, they provided written informed consent. Pseudo-anonymization was used to ensure data privacy and confidentiality. Participants' personal identities were masked using unique participant IDs for the data collected through smartphones, activity trackers, and online questionnaire forms (for pre- and postintervention assessments). A reference list linking the participant IDs and personal information (eg, names and email) was created digitally. This list was only accessible by the first author and was destroyed at the end of data collection. Each participant was compensated at the end of the study (JPY 15,000 - 20,000; US $100 - 140).

Statistical Analysis

The primary outcome (IPAQ total score) was analyzed using a multilevel model with the group (control vs EMI) as the between-person factor and time as the within-person factor (pre- vs postintervention). A random intercept and slope were assumed to account for individual differences in baseline levels and changes in PA. The same multilevel model was applied to the secondary outcomes assessed pre- and postintervention. Another set of multilevel models was estimated for the day-level outcomes assessed over the 3-week intervention period. Each model included groups as the between-person predictor, time (days since the beginning of the intervention) as the within-person predictor, and their interaction. In each model, the mean over the baseline week was controlled for, and a random slope was assumed to explain individual differences in the changes in outcomes. Restricted maximum likelihood estimation was used. We calculated Cohen *d* for group differences in (1) the postintervention scores and (2) means during the intervention period per week (ie, weeks 2 - 4). All analyses were conducted using R (version 4.2.2; R Foundation for Statistical Computing), with the relevant packages of *lme4*, *lmerTest*, *emmeans*, and *fitbitr* [52-54].

Analyses were performed on an intention-to-treat basis; all randomized participants were included in each test. As no participants dropped out of the study, we had complete data for the variables assessed pre- and postintervention. For the EMA variables, we explored the variables that informed missingness before running the multilevel modeling analyses. Specifically, we calculated the compliance rates of the EMA responses for each participant, which were then tested for correlations with participant characteristics at baseline. As we found no significant correlations (see the compliance section), we did not use a covariate in each multilevel model.

Results

Recruitment, Demographics, and Baseline Characteristics

Figure 2 illustrates the recruitment process. The data collection was conducted between November 2024 and March 2025. All randomized participants (n=40) completed the postintervention assessment and were included in the analyses. No dropouts were observed, and no adverse events were reported (14 participants reported a light skin rash likely due to wearing a Fitbit). The baseline characteristics of the participants are summarized in Table 2. Most participants were middle-aged, had received higher education (graduated from a 2-year college, university, or higher), and had a job. Their BMI was in the normal range (18.5 - 25.0 kg/m^2), but they reported little to no habit of PA at baseline (ie, most were at a precontemplation or contemplation stage as per the Stage-of-Change Questionnaire). No significant differences were detected between the EMI and

XSL•FO
RenderX

control groups in baseline characteristics, except for household income (higher in the EMI group), which was controlled for in subsequent analyses.

Figure 2. CONSORT (Consolidated Standards of Reporting Trials) flow diagram. EMI: ecological momentary intervention.

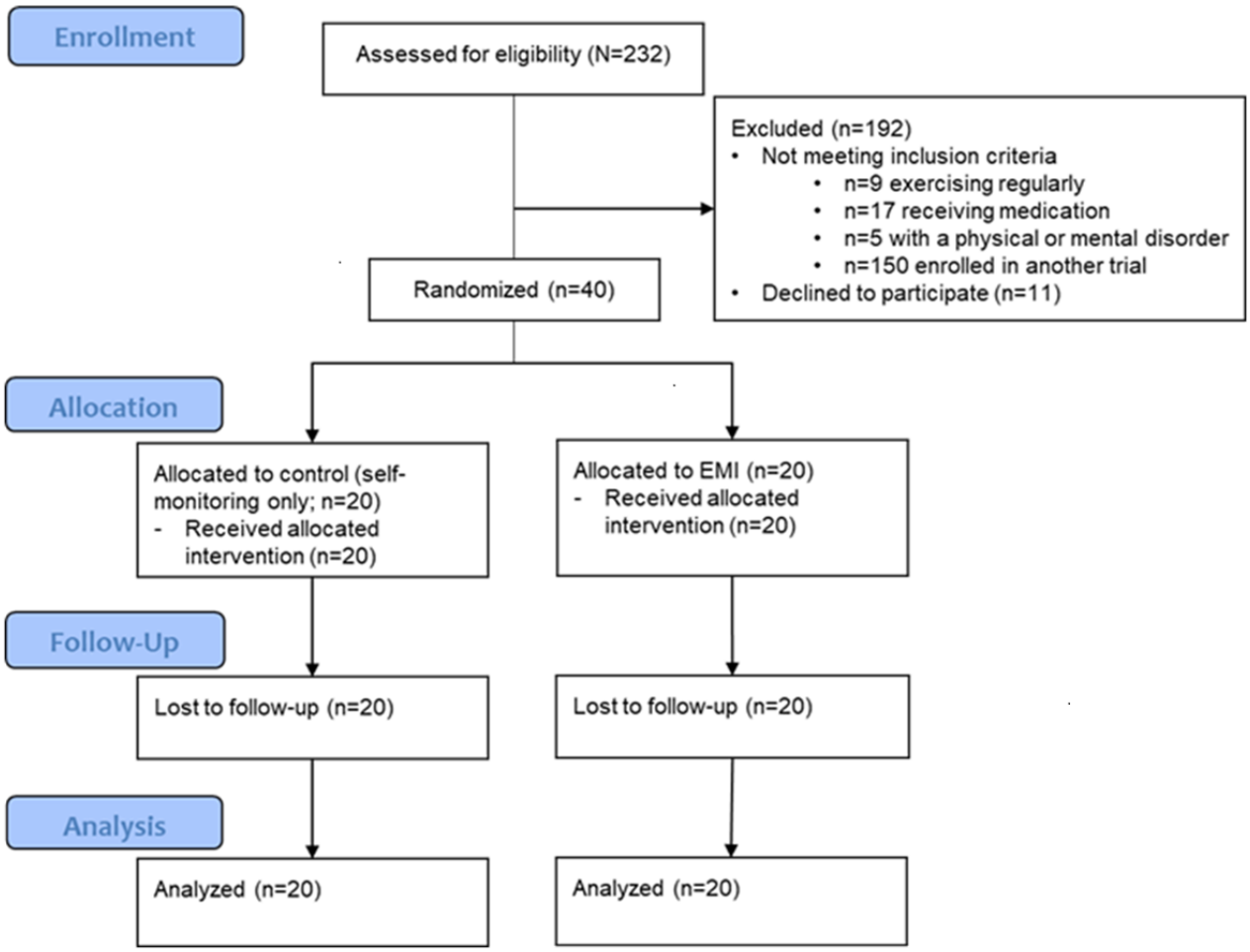

XSL•FO
RenderX

Table . Demographics and descriptive statistics at baseline.

Variables	Control (n=20)	EMI[a] (n=20)	*t* test (*df*)	Chi-square (*df*)	*P* value
Age, mean (SD)	45.2 (8.08)	45.6 (12.69)	0.12 (38)	—[b]	.91
Women, n (%)	13 (65)	10 (50)	—	0.41 (1)	.52
BMI, mean (SD)	24.5 (5.06)	22.26 (3.69)	1.60 (38)	—	.12
Married, n (%)	13 (65)	12 (60)	—	0 (1)	≥.99
One or more children, n (%)	14 (70)	11 (55)	—	0.43 (1)	.51
Education level, n (%)			—	3.59 (2)	.17
Middle or high school	5 (25)	5 (25)			
College	12 (60)	7 (35)			
University and above	3 (15)	8 (40)			
Job, n (%)	19 (95)	17 (85)	—	0.28 (1)	.60
Household income, n (%)			—	12.75 (4)	.01
JPY <3 million (US $2000)	5 (25)	3 (15)			
JPY 3 - 5 million (US $2000 - 3333)	2 (10)	7 (35)			
JPY 5 - 7 million (US $3333 - 4667)	10 (50)	2 (10)			
JPY 7 - 10 million (US $4667 - 6667)	3 (15%)	4 (20)			
JPY 10 million and above (US $6667 and above)	0 (0)	4 (20)			
PA[c] (METs-h/w)[d], mean (SD)	21.63 (42.56)	9.45 (10.56)	1.24 (21.33)	—	.23
Stage of change, n (%)			—	4.40 (4)	.36
Precontemplation	6 (30)	5 (25)			
Contemplation	10 (50)	7 (35)			
Preparation	2 (10)	7 (35)			
Action	1 (5)	0 (0)			
Maintenance	1 (5)	1 (5)			
SMSE-2[e], mean (SD)					
Intrinsic motivation (Cronbach α=0.75)	11.50 (3.56)	12.30 (3.34)	0.73 (38)	—	.47
Integrated regulation (Cronbach α=0.85)	9.70 (3.20)	10.50 (3.82)	0.72 (38)	—	.48
Identified regulation (Cronbach α=0.88)	14.25 (3.40)	14.20 (4.15)	0.04 (38)	—	.97
Introjected regulation (Cronbach α=0.86)	10.80 (4.18)	11.45 (3.97)	0.51 (38)	—	.62
External regulation (Cronbach α=0.64)	7.35 (2.54)	6.50 (2.44)	1.08 (38)	—	.29
Amotivation (Cronbach α=0.83)	6.90 (2.47)	6.40 (2.66)	0.62 (38)	—	.54

XSL•FO
RenderX

Variables	Control (n=20)	EMI[a] (n=20)	*t* test (*df*)	Chi-square (*df*)	*P* value
Baseline phase (week 1)					
Steps, mean (SD)	7639.84 (3767.33)	7071.18 (2843.90)	0.54 (38)	—	.59
Motivation, mean (SD)					
Intrinsic	3.79 (2.00)	4.35 (1.91)	0.91 (38)	—	.37
Identified	4.57 (1.96)	5.04 (1.82)	0.79 (38)	—	.44
External	4.43 (1.84)	5.24 (1.87)	1.37(38)	—	.18
HRR[f] (min/d), mean (SD)	33.86 (29.52)	35.21 (23.13)	0.16 (38)	—	.87

[a]EMI: ecological momentary intervention.

[b]Not available.

[c]PA: physical activity.

[d]METs-h/w: hourly metabolic equivalents per week

[e]SMSE-2: Self-Determined Motivation Scale for Exercise.

[f]HRR: heart rate reserve.

Compliance

The mean compliance rate of the EMA was 79.5% (SD 23.5%; 66.78 valid responses out of 3 prompts ×28 days), and that of the evening prompts was 87.2% (SD 18.81%). We detected valid HR signals with the Fitbit for 90.3% of the study duration (SD 10), indicating that the participants wore a Fitbit conscientiously during study participation. We found no statistically significant correlation between the compliance rates and baseline characteristics (see Table S1 in Multimedia Appendix 1).

Efficacy of the Ecological Momentary Intervention

We performed multilevel modeling analyses with time and group to predict each outcome assessed at the pre- and postintervention time points. The estimated fixed effects can be found in Table S2 in Multimedia Appendix 1. The marginal mean changes and effect sizes (ie, standardized group differences in the change) are presented in Table 3. Although the EMI group showed slightly larger increases in PA than the control group (group difference=7.01), the statistical tests detected no significant group-time interaction effect on the primary (PA assessed by the IPAQ-SF; *P*=.44, Cohen *d*=0.21, 95% CI −0.35 to 0.78) or secondary (motivations assessed by the SMSE; from *P*=.11 to *P*=.91) outcomes.

Table . Marginal mean changes in physical activity and motivations between the control and ecological momentary intervention groups. Changes in pre- and postintervention assessments were calculated.

Outcome	Control (n=20), Estimate (95% CI)	EMI[a] (n=20), Estimate (95% CI)	Group difference, Cohen *d* (95% CI)
PA[b] (METs-h/w)[c]	4.66 (−12.55 to 21.87)	11.67 (−5.54 to 28.88)	0.21 (−0.35 to 0.08)
Motivation			
Intrinsic	0.15 (−1.17 to 1.47)	−1.00 (−2.32 to 0.32)	−0.28 (−0.63 to 0.01)
Integrated	−0.30 (−2.00 to 1.40)	−0.40 (−2.10 to 1.30)	−0.03 (−0.49 to 0.04)
Identified	0.10 (−1.48 to 1.68)	−0.45 (−2.03 to 1.13)	−0.12 (−0.49 to 0.03)
Introjected	−0.20 (−2.17 to 1.77)	−0.75 (−2.72 to 1.22)	−0.12 (−0.59 to 0.04)
External	−0.20 (−1.71 to 1.31)	−0.40 (−1.91 to 1.11)	−0.07 (−0.61 to 0.05)
Amotivation	0.55 (−1.07 to 2.17)	0.10 (−1.52 to 1.72)	−0.15 (−0.71 to 0.041)

[a]EMI: ecological momentary intervention.

[b]PA: physical activity.

[c]METs-h/w: hourly metabolic equivalents per week

Another set of multilevel modeling analyses was performed for the EMA outcomes, namely daily step count, motivation, and minutes of 40% HRR. In each model, the baseline (ie, week 1) levels were controlled, and the main and interaction effects of days and groups were tested. We found no significant main or interaction effects on any outcome (see Table S3 in Multimedia Appendix 1), except for a significant main effect of group on HRR (fixed effect=16.03; SE 6.44; *P*=.02; 95% CI 3.76 to 28.29; Cohen *d*=0.35, 95% CI 0.06 to 0.64). This effect implies that, compared with the control group, the EMI group showed larger increases in minutes of 40% HRR for the intervention (relative to the baseline) weeks (see Table 4). This group difference was

XSL•FO
RenderX

prominent in the first 2 weeks (Cohen *d*=0.41, 95% CI 0.11 to 0.72; *P*=.01 and Cohen *d*=0.30, 95% CI 0.01 to 0.60; *P*=.04) but diminished in the last week of the intervention (Cohen *d*=0.20, 95% CI −0.17 to 0.56; *P*=.28).

We expected that tailored recommendations would enhance participants' engagement with and motivation for PA. Therefore, we investigated how frequently each participant accepted the EMI-recommended activities and how acceptance frequency was associated with levels and motivations for PA. The EMI group reported that they would engage with (ie, accept) a recommended activity in response to 17.3 prompts (SD 17.02; 27.5% out of 3 prompts×21 days) over the intervention phase; they indicated "not now, maybe later" 28.25 times (SD 18.25; 44.8% out of 3 prompts×21 days) and "try another activity" 6.40 times (SD 7.29; 10.2% out of 3 prompts×21 days). The most frequently accepted activities were (1) weeding the garden and cleaning around the entrance (73/346, 21.1%), (2) using the stairs (61/346, 17.63%), and (3) hanging the futon (bed mattress) out in the sun and changing the bedsheet or linen (46/346, 13.3%; see Table S4 in Multimedia Appendix 1). The acceptance frequency showed a positive correlation with increases in the EMA-assessed identified regulation (*r*=.45, *P*=.047). Similarly, acceptance was positively correlated with increases in intrinsic motivation and integrated regulation, but this was not statistically significant (*r*=.41 and *r*=.36; *P*=.08 and *P*=.09, respectively; see Table S5 in Multimedia Appendix 1). No significant correlations were found with the changes in PA amount and intensity.

Table . Marginal means of the ecological momentary assessment outcomes for the control and intervention groups at each week of the intervention phase.

Outcome	Control (n=20), Estimate (95% CI)	EMI[a] (n=20), Estimate (95% CI)	Group difference, Cohen *d* (95% CI)
Steps/day			
Week 2	7327.68 (6576.67 to 8078.70)	7960.28 (7209.23 to 8711.34)	0.17 (−0.12 to 0.46)
Week 3	7114.60 (6344.60 to 7884.61)	7666.41 (6897.38 to 8435.44)	0.15 (−0.15 to 0.45)
Week 4	6901.53 (5730.75 to 8072.30)	7372.54 (6203.75 to 8541.33)	0.13 (−0.32 to 0.58)
Motivation			
Intrinsic			
Week 2	4.28 (4.03 to 4.53)	4.14 (3.89 to 4.39)	−0.16 (−0.60 to 0.27)
Week 3	4.21 (3.93 to 4.49)	4.11 (3.83 to 4.39)	−0.11 (−0.59 to 0.37)
Week 4	4.14 (3.79 to 4.49)	4.08 (3.74 to 4.43)	−0.06 (−0.65 to 0.53)
Identified			
Week 2	4.75 (4.49 to 5.02)	4.96 (4.70 to 5.22)	0.24 (−0.20 to 0.68)
Week 3	4.74 (4.44 to 5.03)	4.97 (4.67 to 5.27)	0.27 (−0.22 to 0.77)
Week 4	4.72 (4.36 to 5.08)	4.98 (4.63 to 5.33)	0.30 (−0.29 to 0.89)
External			
Week 2	4.97 (4.73 to 5.21)	4.92 (4.68 to 5.16)	−0.05 (−0.45 to 0.35)
Week 3	4.98 (4.73 to 5.23)	4.88 (4.63 to 5.13)	−0.12 (−0.54 to 0.29)
Week 4	5.00 (4.67 to 5.33)	4.84 (4.51 to 5.16)	−0.19 (−0.73 to 0.35)
HRR[b]			
Week 2	22.95 (13.16 to 32.73)	41.70 (31.96 to 51.45)	0.41 (0.11 to 0.72)
Week 3	23.78 (14.34 to 33.23)	37.68 (28.38 to 46.98)	0.30 (0.01 to 0.60)
Week 4	24.62 (12.73 to 36.51)	33.66 (22.07 to 45.25)	0.20 (−0.17 to 0.56)

[a]EMI: ecological momentary intervention.

[b]HRR: heart rate reserve; minutes of 40% HRR per day.

Feasibility

For feasibility, each item was rated on a 7-point scale, and we interpreted scores of 4 or higher as an indication that participants agreed on that point. Approximately half of the participants in the EMI group reported that the intervention was satisfactory (10/20, 50%), useful (9/20, 45%), helpful (9/20, 45%), suitable (8/20, 40%), and practical (10/20, 50%; see Table 5). Only a quarter (5/20, 25%) rated the prompt timing as appropriate, implying that typical EMA prompting (here, 3 times a day with a fixed schedule) may not be suitable for the user experience. Despite the timing issue, more than half (11/20, 55%) of the participants were willing to continue the intervention.

XSL•FO
RenderX

Table . Feasibility rated by participants in the ecological momentary intervention group (N=20).

Item[a]	Values, mean (SD)	Endorsement, n (%)
Satisfactory	3.10 (1.21)	10 (50)
Useful	3.30 (1.08)	9 (45)
Helpful	3.10 (1.21)	9 (45)
Suitable	3.15 (1.04)	8 (40)
Timing	2.85 (1.04)	5 (25)
Practical	3.30 (1.03)	10 (50)
Willingness to continue	3.35 (1.46)	11 (55)

[a]Each item was rated on a 7-point scale (1="Not at all" and 7="Very much"). Scores of 4 or higher were regarded as endorsements for each item.

Discussion

Principal Findings

This pilot randomized controlled trial investigated the efficacy and feasibility of the EMI in improving PA, with messages and recommendations tailored to participants' EMA responses. Overall, EMI had null effects on self-reported PA, motivation for PA, or step count as assessed by an activity tracker. However, the minutes of 40% HRR (moderate- or high-intensity PA) showed a larger increase in the EMI group than in the control group. Regarding feasibility, approximately half of the participants found the intervention useful and satisfactory and expressed a willingness to continue. However, the prompt timing was perceived as inappropriate, leaving room for improvement in the design and implementation.

Efficacy

Contrary to our hypothesis, no significant improvement was observed in the self-reported PA or step count. These null findings are in line with the results of previous EMI trials [16-18], which failed to establish a transfer effect on PA. This result suggests that fully automated, low-intensity EMI may not be strong enough to induce a significant change in PA. Our EMI was delivered for 3 weeks, which was comparable to the average duration of EMIs in the literature; a systematic review documented that the duration of EMI trials ranged from 2 to 15 weeks, with an average of 4 weeks [1]. However, a JITAI study [55] showed that participants became bored even with contextually tailored messages (eg, "Is there anything better than weekend afternoons? You could enjoy the fresh air by taking a walk around your neighborhood!") so that the effect of messages declined as the intervention progressed. That means EMI (or messaging in general) might have low intensity, needing some add-ons that maintain and facilitate active user engagement, like team competition or other forms of gamification [22]. Some more reflections in light of the principles proposed by Van Dantzig et al [20] can help in the future development of an EMI. First, we sent out 3 EMA prompts across the day, which was also comparable to the average sending frequency reported in the literature (average of 4 to 5 times [1]). However, this prompting might be too sparse to spot a moment of opportunity for individuals to engage in a behavior. Similar to other EMI trials (eg, using accelerometers to monitor and detect sedentary behavior [19,56]), integrating accelerometers and other wearable sensors into the EMI decision system to determine what messages to send and when might have improved acceptability and efficacy. Another possible technical improvement would be to link an EMI app to an external database through an application performance interface to fetch relevant momentary information (eg, weather and location) when needed, without relying on a user's self-report [21,26]. Second, the recommended activities were carefully selected when designing the intervention, and we were convinced that each activity was actionable in each target context. However, we did not tailor the recommendations to individual preferences; the same decision rules and activity lists were used across participants. To improve acceptability, a user-defined activity list that aligns with user preferences [7] or frequent behaviors [26] may be more appropriate.

Notably, our EMI failed to establish an effect on motivation for PA, unlike earlier EMI trials [6,7,16,17]. We assumed that short, concise text messages would be sufficient to motivate participants as long as they were context-aware [16]. However, our findings may suggest that the prompting messages should contain explicit motivational and congratulatory words (delivering congratulatory and supportive messages when users engage in healthy behavior [6] and delivering personalized feedback and encouragement based on individuals' predetermined goals [7]) to exert a positive psychological influence on message recipients. The causal relationship remains unclear (ie, whether EMI-induced motivation leads to an improvement in PA [16,17]), but motivation plays an important role in aspects other than efficacy, such as sustained user engagement and retention [57]. Developers of EMI should consider embedding a motivational source in their intervention if it is not personalized or tailored. In addition to motivation, the COM-B model highlights the roles of capability and opportunity as active components of an effective intervention for behavior change [13]. Our EMI was designed to help find an opportunity to perform PA in daily life, and the recommended activities did not require high levels of capability. However, half of the participants rated prompt timing as inappropriate, implying that the opportunity component did not work as intended. Inappropriate prompt timing may affect motivation to maintain the engagement—that means, through the lens of the COM-B model, the motivation and opportunity components would be the key challenges for the EMI.

XSL•FO
RenderX

We found that the EMI group showed a significantly larger increase in minutes of 40% HRR than the control group. Although the HRR analysis was exploratory (not preregistered), this significant increase may imply that the EMI group successfully integrated the recommended activities into their daily routine, which were small but slightly physically intensive behavioral changes. It should be noted that the EMI had no significant effect on step count or motivation outcomes, meaning that HRR changed independently of the amount of PA and motivation for that. A meta-analysis showed that autonomous motivation is not associated with the amount of PA (and likely the intensity as well) when assessed as an outcome in intervention studies [58]. In the context of physical education, HR-based intensity was found to have even a negative association with autonomous motivation [59]. These findings may suggest that individuals do not have to be intrinsically motivated to perform PA in an intervention. Our EMI recommended low-intensity PA that could be part of house choring (eg, cleaning, kitchen activity, and weeding the garden), which were necessary for living but did not have to be liked. Also, these activities did not require massive body movement, directly increasing step count. Although this narrative includes some speculation, it is possible that engaging in the recommended activities led to changes in HRR, while these changes were not reflected in motivation or step count.

Feasibility

Compliance was excellent, and no dropouts were observed. However, the EMI was rated low for satisfaction and other feasibility items. Around half of the EMI participants found that the intervention was satisfactory, useful, helpful, suitable, and practical and showed a willingness to continue the intervention. For one-third of the EMA prompts, the participants indicated that they would engage in a recommended activity; for the other prompts, they typically indicated “not now, maybe later.” Indeed, prompt timing was perceived as suboptimal (no more than 25% of the participants rated the timing as appropriate). A pilot EMI trial [7] reported that participants found most (>80%) EMI messages enjoyable and motivating. Other trials have also documented that EMI messages and reminders are liked and accepted by participants [15,16]; however, several studies have noted that just-in-time messages might be perceived as too early or too late in a day [60] and that the activity recommendations might not always reflect participants’ preferences [26]. As we found a preliminary association between the frequency of accepting activity recommendations and the motivation outcome of integrated regulation (r=0.45), it is crucial to improve the acceptability of each message and recommendation. In this sense, using a more positive and encouraging tone in each message and tailoring recommendations per users’ preferences might be helpful. In addition, a machine learning approach may be able to resolve the optimization issues for improving acceptability. A recent study [61] trained a machine learning model on EMA data to predict momentary engagements in PA using situational and contextual information. Such a model’s prediction would allow for tailoring the content and timing of intervention prompts (eg, delivering motivational messages with specific suggestions when the highest message acceptability is predicted).

Limitations

These results should be interpreted in light of the several limitations of this study. First, the sample size might not have been sufficiently large to detect several important effects, particularly for the correlation analyses within the EMI group. Second, the generalizability of the findings might have been limited by the homogeneity of the sample. We did not find any effects unique to the Japanese population, but some aspects of the EMI, such as the selection of recommended activities, might require appropriate adaptations for users with a specific cultural background and lifestyle. Third, the participants were paid volunteers, which might explain their good compliance and zero dropout rate. In a naturalistic setting, low retention rates are one of the most important challenges for mobile devices, especially for commercial health care apps [62,63]. A motivational source for incentivizing users is crucial when introducing the EMI to the market [64]. Fourth, we relied on self-reported EMA to assess the 3 tailoring variables. The self-reporting could be replaced with passive sensing (eg, GPS and audio recording), which would allow for just-in-time adaptive triggers without any cost for users to respond to EMA signals. For this study, we could not, however, clear data protection issues concerning passive sensing, especially for handling personal information (eg, GPS signals and voices). We see technical and ethical challenges in implementing a ubiquitous sensing system, but this will also resolve most of the timing issues in EMI. We used fixed-time prompting to improve the response rate of participants, but there would be no need to fix or define any prompt timing if real-time sensing becomes available.

Conclusion

Regardless of these limitations, we believe that this pilot trial serves as an EMI exemplar. It highlights several important challenges in designing a successful EMI program, such as prompt timing and other technical or technological mechanisms to effectively detect a moment of opportunity, the message tone to be used (positive and motivational), and activity recommendations tailored to momentary contextual information and individual preferences. As noted above, machine learning may resolve part of the optimization issues for the timing and content of messaging. Another important direction is the use of passive sensing, which would require less or even no self-reporting to detect the moment of opportunity for intervention. Self-reported EMA may feel, even if carefully personalized, somewhat distracting and invasive due to the high frequency of prompting. These 2 approaches could be used in combination and have a large potential to improve the acceptability of intervention messages. This study also had an implication for the trial design, particularly for outcome selection. The EMI often targets daily routine behaviors (eg, sedentary time) and aims to replace them with slightly more active alternatives (eg, taking a break or standing). Such small behavioral changes may not be reflected in accelerometer-based outcomes but could be captured by HR. Although HRR has not been widely used in mobile assessments, future EMI trials may benefit from it, increasing the chances of identifying an undetected effect with step count and other PA amount measures.

XSL•FO
RenderX

Acknowledgments

We thank Daichi Watabayashi for his support in data collection.

Funding

This study was supported by JSPS KAKENHI (23H01051).

Data Availability

The data sets are available from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary tables for each analysis.
[DOCX File, 52 KB - formative_v9i1e79360_app1.docx]

Checklist 1
CONSORT checklist.
[PDF File, 1375 KB - formative_v9i1e79360_app2.pdf]

References

1. Dao KP, De Cocker K, Tong HL, Kocaballi AB, Chow C, Laranjo L. Smartphone-delivered ecological momentary interventions based on ecological momentary assessments to promote health behaviors: systematic review and adapted checklist for reporting ecological momentary assessment and intervention studies. JMIR Mhealth Uhealth 2021 Nov 19;9(11):e22890. [doi: 10.2196/22890] [Medline: 34806995]
2. Heron KE, Smyth JM. Ecological momentary interventions: Incorporating mobile technology into psychosocial and health behaviour treatments. British J Health Psychol 2010 Feb;15(1):1-39. [doi: 10.1348/135910709X466063]
3. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol 2008;4:1-32. [doi: 10.1146/annurev.clinpsy.3.022806.091415] [Medline: 18509902]
4. Bakker D, Kazantzis N, Rickwood D, Rickard N. Development and pilot evaluation of smartphone-delivered cognitive behavior therapy strategies for mood- and anxiety-related problems: MoodMission. Cogn Behav Pract 2018 Nov;25(4):496-514. [doi: 10.1016/j.cbpra.2018.07.002]
5. Goldstein SP, Thomas JG, Foster GD, et al. Refining an algorithm-powered just-in-time adaptive weight control intervention:a randomized controlled trial evaluating model performance and behavioral outcomes. Health Informatics J 2020 Dec;26(4):2315-2331. [doi: 10.1177/1460458220902330]
6. Mundi MS, Lorentz PA, Grothe K, Kellogg TA, Collazo-Clavell ML. Feasibility of smartphone-based education modules and ecological momentary assessment/intervention in pre-bariatric surgery patients. Obes Surg 2015 Oct;25(10):1875-1881. [doi: 10.1007/s11695-015-1617-7] [Medline: 25702141]
7. Bronas UG, Marquez DX, Fritschi C, et al. Ecological momentary intervention to replace sedentary time with physical activity to improve executive function in midlife and older latino adults: pilot randomized controlled trial. J Med Internet Res 2024 Sep 5;26:e55079. [doi: 10.2196/55079] [Medline: 39235836]
8. Rodgers A, Corbett T, Bramley D, et al. Do u smoke after txt? Results of a randomised trial of smoking cessation using mobile phone text messaging. Tob Control 2005 Aug;14(4):255-261. [doi: 10.1136/tc.2005.011577] [Medline: 16046689]
9. Schwerdtfeger AR, Schmitz C, Warken M. Using text messages to bridge the intention-behavior gap? A pilot study on the use of text message reminders to increase objectively assessed physical activity in daily life. Front Psychol 2012;3(270):270. [doi: 10.3389/fpsyg.2012.00270] [Medline: 22876237]
10. Weitzel JA, Bernhardt JM, Usdan S, Mays D, Glanz K. Using wireless handheld computers and tailored text messaging to reduce negative consequences of drinking alcohol. J Stud Alcohol Drugs 2007 Jul;68(4):534-537. [doi: 10.15288/jsad.2007.68.534] [Medline: 17568957]
11. Carreiro S, Newcomb M, Leach R, Ostrowski S, Boudreaux ED, Amante D. Current reporting of usability and impact of mHealth interventions for substance use disorder: A systematic review. Drug Alcohol Depend 2020 Oct 1;215(108201):108201. [doi: 10.1016/j.drugalcdep.2020.108201] [Medline: 32777691]
12. Anderson E, Durstine JL. Physical activity, exercise, and chronic diseases: a brief review. Sports Med Health Sci 2019 Dec;1(1):3-10. [doi: 10.1016/j.smhs.2019.08.006] [Medline: 35782456]

XSL•FO
RenderX

13. Michie S, van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. Implement Sci 2011 Apr 23;6(1):42. [doi: 10.1186/1748-5908-6-42] [Medline: 21513547]
14. Carey RL, Le H, Coffman DL, et al. mHealth-based just-in-time adaptive intervention to improve the physical activity levels of individuals with spinal cord injury: protocol for a randomized controlled trial. JMIR Res Protoc 2024 Jun 28;13(1):e57699. [doi: 10.2196/57699] [Medline: 38941145]
15. Hardeman W, Houghton J, Lane K, Jones A, Naughton F. A systematic review of just-in-time adaptive interventions (JITAIs) to promote physical activity. Int J Behav Nutr Phys Act 2019 Apr 3;16(1):31. [doi: 10.1186/s12966-019-0792-7] [Medline: 30943983]
16. Ding X, Xu J, Wang H, Chen G, Thind H, Zhang Y. WalkMore: promoting walking with just-in-time context-aware prompts. Presented at: 2016 IEEE Wireless Health (WH); Oct 25-27, 2016. [doi: 10.1109/WH.2016.7764558]
17. Sporrel K, Wang S, Ettema DDF, et al. Just-in-time prompts for running, walking, and performing strength exercises in the built environment: 4-week randomized feasibility study. JMIR Form Res 2022 Aug 1;6(8):e35268. [doi: 10.2196/35268] [Medline: 35916693]
18. van Dantzig S, Geleijnse G, van Halteren AT. Toward a persuasive mobile application to reduce sedentary behavior. Pers Ubiquit Comput 2013 Aug;17(6):1237-1246. [doi: 10.1007/s00779-012-0588-0]
19. Bond DS, Thomas JG, Raynor HA, et al. B-MOBILE--a smartphone-based intervention to reduce sedentary time in overweight/obese individuals: a within-subjects experimental trial. PLOS One 2014;9(6):e100821. [doi: 10.1371/journal.pone.0100821] [Medline: 24964010]
20. van Dantzig S, Bulut M, Krans M, van der Lans A, de Ruyter B. Enhancing physical activity through context-aware coaching. 2018 May 21 Presented at: PervasiveHealth '18; May 21-24, 2018. [doi: 10.1145/3240925.3240928]
21. Ismail T, Al Thani D. Design and evaluation of a just-in-time adaptive intervention (JITAI) to reduce sedentary behavior at work: experimental study. JMIR Form Res 2022 Jan 26;6(1):e34309. [doi: 10.2196/34309] [Medline: 35080498]
22. Patel MS, Small DS, Harrison JD, et al. Effectiveness of behaviorally designed gamification interventions with social incentives for increasing physical activity among overweight and obese adults across the United States: the STEP UP randomized clinical trial. JAMA Intern Med 2019 Dec 1;179(12):1624. [doi: 10.1001/jamainternmed.2019.3505]
23. Karvonen J, Vuorimaa T. Heart rate and exercise intensity during sports activities. Practical application. Sports Med 1988 May;5(5):303-311. [doi: 10.2165/00007256-198805050-00002] [Medline: 3387734]
24. Wu HC, Wang MJJ. Relationship between maximum acceptable work time and physical workload. Ergonomics 2002 Mar;45(4):280-289. [doi: 10.1080/00140130210123499]
25. Coenen P, Korshøj M, Hallman DM, et al. Differences in heart rate reserve of similar physical activities during work and in leisure time – A study among Danish blue-collar workers. Physiol Behav 2018 Mar;186:45-51. [doi: 10.1016/j.physbeh.2018.01.011]
26. Rabbi M, Pfammatter A, Zhang M, Spring B, Choudhury T. Automated personalized feedback for physical activity and dietary behavior change with mobile phones: a randomized controlled trial on adults. JMIR Mhealth Uhealth 2015 May 14;3(2):e42. [doi: 10.2196/mhealth.4160] [Medline: 25977197]
27. Takano K, Oba T, Katahira K, Kimura K. Deconstructing Fitbit to specify the effective features in promoting physical activity among inactive adults: pilot randomized controlled trial. JMIR Mhealth Uhealth 2024 Jul 12;12:e51216. [doi: 10.2196/51216] [Medline: 38996332]
28. Takano K, Oba T, Kimura K, Katahira K, Moriishi C, MacLellan A. Collaborative and competitive features in mobile health intervention for improving physical activity. OSF. 2024 Jul 19. URL: https://osf.io/2abg9/overview?view_only=15cbd6283d174ccb9f0c569b7fed494e [accessed 2025-10-22]
29. Head KJ, Noar SM, Iannarino NT, Grant Harrington N. Efficacy of text messaging-based interventions for health promotion: a meta-analysis. Soc Sci Med 2013 Nov;97(41–48):41-48. [doi: 10.1016/j.socscimed.2013.08.003] [Medline: 24161087]
30. Smith DM, Duque L, Huffman JC, Healy BC, Celano CM. Text message interventions for physical activity: a systematic review and meta-analysis. Am J Prev Med 2020 Jan;58(1):142-151. [doi: 10.1016/j.amepre.2019.08.014] [Medline: 31759805]
31. Julious SA. Sample size of 12 per group rule of thumb for a pilot study. Pharm Stat 2005 Oct;4(4):287-291. [doi: 10.1002/pst.185]
32. Kieser M, Wassmer G. On the use of the upper confidence limit for the variance from a pilot sample for sample size determination. Biometrical J 1996 Jan;38(8):941-949. [doi: 10.1002/bimj.4710380806]
33. Whitehead AL, Julious SA, Cooper CL, Campbell MJ. Estimating the sample size for a pilot randomised trial to minimise the overall trial sample size for the external pilot and main trial for a continuous outcome variable. Stat Methods Med Res 2016 Jun;25(3):1057-1073. [doi: 10.1177/0962280215588241] [Medline: 26092476]
34. Mestdagh M, Verdonck S, Piot M, et al. m-Path: an easy-to-use and highly tailorable platform for ecological momentary assessment and intervention in behavioral research and clinical practice. Front Digit Health 2023;5:1182175. [doi: 10.3389/fdgth.2023.1182175] [Medline: 37920867]
35. Michie S, Richardson M, Johnston M, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013 Aug;46(1):81-95. [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]

XSL•FO
RenderX

36. Ainsworth BE, Haskell WL, Herrmann SD, et al. 2011 Compendium of Physical Activities: a second update of codes and MET values. Med Sci Sports Exerc 2011 Aug;43(8):1575-1581. [doi: 10.1249/MSS.0b013e31821ece12] [Medline: 21681120]
37. Lee PH, Macfarlane DJ, Lam TH, Stewart SM. Validity of the International Physical Activity Questionnaire Short Form (IPAQ-SF): a systematic review. Int J Behav Nutr Phys Act 2011 Oct 21;8(1):115. [doi: 10.1186/1479-5868-8-115] [Medline: 22018588]
38. Adams SA, Matthews CE, Ebbeling CB, et al. The effect of social desirability and social approval on self-reports of physical activity. Am J Epidemiol 2005 Feb 15;161(4):389-398. [doi: 10.1093/aje/kwi054] [Medline: 15692083]
39. Craig CL, Marshall AL, Sjöström M, et al. International physical activity questionnaire: 12-country reliability and validity. Med Sci Sports Exerc 2003 Aug;35(8):1381-1395. [doi: 10.1249/01.MSS.0000078924.61453.FB] [Medline: 12900694]
40. Murase N, Katsumura T, Ueda C, Inoue S, Shimomitsu T. Validity and reliability of Japanese version of International Physical Activity Questionnaire. J Health Welf Stat 2002(11):1-9 [FREE Full text]
41. Matsumoto H, Taniguchi A, Nishida J. A revised self-determined motivation scale for exercise with integrated regulation inclusion. Journal of Health Psychology Research 2021;34(1):13-22. [doi: 10.11560/jhpr.200608129]
42. Ryan RM, Deci EL. Intrinsic and extrinsic motivations: classic definitions and new directions. Contemp Educ Psychol 2000 Jan;25(1):54-67. [doi: 10.1006/ceps.1999.1020] [Medline: 10620381]
43. Marcus BH, Simkin LR. The stages of exercise behavior. J Sports Med Phys Fitness 1993 Mar;33(1):83-88. [Medline: 8350613]
44. Reed GR, Velicer WF, Prochaska JO, Rossi JS, Marcus BH. What makes a good staging algorithm: examples from regular exercise. Am J Health Promot 1997;12(1):57-66. [doi: 10.4278/0890-1171-12.1.57] [Medline: 10170436]
45. Liguori G, Yuri F, Fountaine C. ACSM's Guidelines for Exercise Testing and Prescription, 11th edition: Wolters Kluwer; 2022.
46. Tanaka H, Monahan KD, Seals DR. Age-predicted maximal heart rate revisited. J Am Coll Cardiol 2001 Jan;37(1):153-156. [doi: 10.1016/s0735-1097(00)01054-8] [Medline: 11153730]
47. Lonsdale C, Sabiston CM, Taylor IM, Ntoumanis N. Measuring student motivation for physical education: examining the psychometric properties of the Perceived Locus of Causality Questionnaire and the Situational Motivation Scale. Psychol Sport Exerc 2011 Jun;12(3):284-292. [doi: 10.1016/j.psychsport.2010.11.003]
48. Rodrigues F, Cid L, Faustino T, Monteiro D. The situational motivation scale in the exercise context: construct validity, factor structure, and correlational analysis. Curr Psychol 2023 Feb;42(6):4811-4820. [doi: 10.1007/s12144-021-01824-2] [Medline: 36684465]
49. Weiner BJ, Lewis CC, Stanick C, et al. Psychometric assessment of three newly developed implementation outcome measures. Implementation Sci 2017 Dec;12(1). [doi: 10.1186/s13012-017-0635-3]
50. Takano K, Oba T, Katahira K, Kimura K, Moriishi C. Ecological momentary intervention for improving physical activity: a pilot randomized controlled trial to tailor recommendation messages for inactive adults. OSF. 2024 Nov. URL: https://osf.io/nzkyf/overview?view_only=e23804b941dd4843a3ba0902e51fec20 [accessed 2025-05-19]
51. Eysenbach G, Group CE. CONSORT-EHEALTH: improving and standardizing evaluation reports of Web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126. [doi: 10.2196/jmir.1923] [Medline: 22209829]
52. Kuznetsova A, Brockhoff PB, Christensen RHB. lmerTest package: tests in linear mixed effects models. J Stat Soft 2017;82(13):1-26. [doi: 10.18637/jss.v082.i13]
53. Lenth RV, Banfai B, Bolker B, Buerkner P, Giné-Vázquez I, Herve M, et al. Emmeans: estimated marginal means, aka least-squares means. 2025. URL: https://cran.r-project.org/web/packages/emmeans/index.html [accessed 2025-04-28]
54. Kaye KM. Fitbitr: interface with the "fitbit" API. 2023. URL: https://cran.r-project.org/web/packages/fitbitr/index.html [accessed 2025-04-28]
55. Klasnja P, Smith S, Seewald NJ, et al. Efficacy of contextually tailored suggestions for physical activity: a micro-randomized optimization trial of HeartSteps. Ann Behav Med 2019 May 3;53(6):573-582. [doi: 10.1093/abm/kay067] [Medline: 30192907]
56. Pellegrini CA, Hoffman SA, Daly ER, Murillo M, Iakovlev G, Spring B. Acceptability of smartphone technology to interrupt sedentary time in adults with diabetes. Transl Behav Med 2015 Sep;5(3):307-314. [doi: 10.1007/s13142-015-0314-3] [Medline: 26327936]
57. Vandelanotte C, Trost S, Hodgetts D, et al. Increasing physical activity using an just-in-time adaptive digital assistant supported by machine learning: a novel approach for hyper-personalised mHealth interventions. J Biomed Inform 2023 Aug;144:104435. [doi: 10.1016/j.jbi.2023.104435] [Medline: 37394024]
58. Knittle K, Nurmi J, Crutzen R, Hankonen N, Beattie M, Dombrowski SU. How can interventions increase motivation for physical activity? A systematic review and meta-analysis. Health Psychol Rev 2018 Sep;12(3):211-230. [doi: 10.1080/17437199.2018.1435299] [Medline: 29385950]
59. Jaakkola T, Liukkonen J, Laakso T, Ommundsen Y. The relationship between situational and contextual self-determined motivation and physical activity intensity as measured by heart rates during ninth grade students' physical education classes. Eur Phy Educ Rev 2008 Feb;14(1):13-31. [doi: 10.1177/1356336X07085707]

XSL•FO
RenderX

60. Lin Y, Jessurun J, de Vries B, Timmermans H. Motivate: towards context-aware recommendation mobile system for healthy living. 2012 Presented at: 5th International ICST Conference on Pervasive Computing Technologies for Healthcare; May 23-26, 2011. [doi: 10.4108/icst.pervasivehealth.2011.246030]
61. Kumar D, Haag D, Blechert J, Niebauer J, Smeddinck JD. Feature selection for physical activity prediction using ecological momentary assessments to personalize intervention timing: longitudinal observational study. JMIR Mhealth Uhealth 2025 Jan 24;13(1):e57255. [doi: 10.2196/57255] [Medline: 39865572]
62. Amagai S, Pila S, Kaat AJ, Nowinski CJ, Gershon RC. Challenges in participant engagement and retention using mobile health apps: literature review. J Med Internet Res 2022 Apr 26;24(4):e35120. [doi: 10.2196/35120] [Medline: 35471414]
63. Di Sebastiano KM, Lau EY, Yun L, Faulkner G. An evaluation of a commercialized mHealth intervention to promote physical activity in the workplace. Front Public Health 2022;10(740350):740350. [doi: 10.3389/fpubh.2022.740350] [Medline: 35372253]
64. Mitchell M, Lau E, White L, Faulkner G. Commercial app use linked with sustained physical activity in two Canadian provinces: a 12-month quasi-experimental study. Int J Behav Nutr Phys Act 2020 Feb 25;17(1):24. [doi: 10.1186/s12966-020-00926-7] [Medline: 32098625]

Abbreviations

COM-B: capability, opportunity, motivation, and behavior
CONSORT-EHEALTH: Consolidated Standards for Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
EMA: ecological momentary assessment
EMI: ecological momentary intervention
HR: heart rate
HRR: heart rate reserve
JITAI: just-in-time adaptive intervention
MET: metabolic equivalent
PA: physical activity
SMSE: Self-Determined Motivation Scale for Exercise

Edited by A Mavragani; submitted 22.06.25; peer-reviewed by H Tilga; revised version received 22.10.25; accepted 23.10.25; published 04.12.25.

Please cite as:
Oba T, Moriishi C, Takano K, Katahira K, Kimura K
Mobile-Based Ecological Momentary Intervention for Improving Physical Activity in Adults Without Regular Physical Activity: Pilot Randomized Controlled Trial
JMIR Form Res 2025;9:e79360
URL: https://formative.jmir.org/2025/1/e79360
doi:10.2196/79360

©Takeyuki Oba, Chihiro Moriishi, Keisuke Takano, Kentaro Katahira, Kenta Kimura. Originally published in JMIR Formative Research (https://formative.jmir.org), 4.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Feasibility of the Social Media–Based Prevention Program "Leduin" for German Adolescents on Instagram: Mixed Methods Pilot Study

Elizabeth Zimmermann[1,2], MSc; Samuel Tomczyk[1,2], MSc, Dr phil

[1]Department Health and Prevention, Institute of Psychology, University of Greifswald, Robert-Blum-Str. 13, Greifswald, Germany

[2]German Center for Child and Adolescent Health (DZKJ), Partner Site Greifswald/Rostock, Greifswald, Germany

Corresponding Author:
Elizabeth Zimmermann, MSc
Department Health and Prevention, Institute of Psychology, University of Greifswald, Robert-Blum-Str. 13, Greifswald, Germany

Abstract

Background: Digital platforms, particularly social media, including Instagram, present unique opportunities for health promotion among adolescents due to their widespread use with interactive features supporting high user engagement. However, the feasibility of effectively utilizing platforms like Instagram for health interventions requires careful consideration of adolescent engagement patterns.

Objective: This pilot study evaluated the leduin program—designed to foster essential life skills and functional social media use among adolescents—while also exploring the broader feasibility of using Instagram to deliver complex social and psychological interventions in this population.

Methods: The study adapted the feasibility framework by Bowen et al and used a mixed methods approach. Quantitatively, Instagram interaction metrics of 99 participants (women: 62/99, 63%; men: 37/99, 37%; aged 14 - 18 years; mean age 15.2, SD 0.74 years) were analyzed descriptively (means, medians, SDs) and inferentially (Welch ANOVA, Kruskal-Wallis, Pearson and Spearman correlations, linear and segmented regression analyses) using RStudio. Metrics included story views, retention rates, feature engagement (eg, polls, question stickers, quizzes), and drop-off rates. Recruitment efforts were also analyzed descriptively. Qualitatively, 13 postprogram semistructured interviews were conducted with 11 women (11/13, 65%) and 6 men (6/13, 35%; mean age 15.29, SD 0.99 years). Participants were sampled to reflect varying engagement levels (6 high, 5 medium, 6 low). The mean interview duration was 25 minutes 11 seconds (SD 6 minutes 34 seconds). Content analysis, with high intercoder reliability (κ=0.90), comprehensively explored participants' experiences and the program's impact.

Results: Quantitative results indicated that the recruitment process was challenging, with 101 schools and 10 youth centers contacted, resulting in a participation rate of 12.8% (99/775 students). On Instagram, story views ranged from 34 to 81 per post, with an average daily retention rate of 87.7% (SD 7.8%). By week 4, 76% of the total drop in views had occurred (mean views declined from 66.1 to 53.4); by week 6, 97.3% of the drop had been reached (declined from 66.1 to 49.9 views), indicating sustained viewer interest over the 14-week program. Features requiring minimal user effort, including polls (mean 56.8% - 54.4%), quizzes (mean 56.6%), and sliders (mean 51.2%), showed significantly higher interaction rates than more demanding features such as challenges (mean 21.7%) and question stickers (mean 20.6%; P<.001). Qualitative findings revealed that adolescents valued the program, its design and methods for its relevance to their daily lives, and its support in developing essential life skills. Suggestions for improvements were made.

Conclusions: The study underlines the potential of various Instagram features and content posting schedules for health interventions to meet adolescent preferences and interests. Challenges with reaching the target group effectively emphasize the need for targeted recruitment strategies and optimizing initial content to boost engagement, underscoring the critical implications for prevention research and policy in leveraging digital platforms to enhance adolescent health.

(JMIR Form Res 2025;9:e78774) doi:10.2196/78774

KEYWORDS

social media; adolescents; feasibility; intervention; life skills; mixed methods; user engagement

XSL•FO
RenderX

Introduction

Opportunities, Risks, and Skills in Adolescent Social Media Use

Digitalization has profoundly transformed modern culture, by blending physical and virtual worlds. Consequently, how individuals communicate and engage with information has been reshaped. A central element of this shift is the rise of social media, which now connects more than 5 billion users worldwide [1]. Adolescents, defined by the World Health Organization (WHO) as individuals aged 10 years to 19 years [2], are among the most active social media users [3]. Accordingly, they are particularly affected by the dual nature it presents. On one hand, platforms like Instagram, YouTube, and TikTok enable the creation and sharing of content on an unprecedented scale, facilitating social interaction, identity exploration, and self-expression [4-7]. In doing so, social media can support key developmental tasks for adolescents such as building autonomy, self-efficacy, and social capital while also providing representation and empowerment for marginalized groups [6,8-10]. On the other hand, social media entails risks including misinformation, privacy risks, and the need for critical consumption skills [11,12]. Due to their still-developing cognitive, emotional, and digital competencies, adolescents are highly vulnerable to risks such as exposure to harmful content, cyberbullying, and platform mechanisms that may support addiction [13-16]. Accordingly, problematic social media use—characterized by preoccupation and compulsion—has been linked to mental health issues, including depression, anxiety, and stress [17-19]. This highlights the need for the development of skills to functionally use social media [15,20].

The Role of Social Media in Promoting Adolescent Health

Although social media poses benefits and risks, it also presents unique opportunities for health promotion and prevention: Its ability to reach adolescents within their daily routines—among teens aged 13 years to 17 years, YouTube is used by 90%, TikTok is used by 63%, and Instagram is used by 61% [3]—makes it a low-threshold, cost-effective, accessible, and engaging medium for interventions [21-25]. For example, a recent meta-analysis found that social media recruitment nearly doubled cost-efficiency over traditional methods and outperformed other online approaches by 66% [23].

Under these favorable conditions, social media interventions have effectively promoted health behaviors among adolescents: Reported successes include improving physical activity and diet, increasing safe sexual practices, raising mental health awareness, and promoting skin health with a small but significant overall effect size of a Hedges *g* of 0.24 [26-28]. These findings indicate the potential of social media as a tool for health promotion, though effectiveness varies by health topic and study.

Interactivity and User Engagement

Interactive elements play a crucial role in social media interactions: Research underscores that interactivity boosts engagement, symptomatic improvements, and learning outcomes [29] by promoting a feeling of connectedness with the program [30]. Interactive features such as social forums, peer interaction, and gamification elements including challenges and rewards significantly enhance user involvement, reduce dropout rates, and encourage behavior change [31-37]. Additionally, effective program design that utilizes multimedia, regular interaction by hosts, and relevant links supports sustained user engagement [37,38].

Instagram, known for its rich interactive features, is increasingly being researched as an effective platform for health promotion [39-42]. It offers significantly higher interaction rates than platforms like Facebook or Twitter, with interaction rates 30 to 200 times higher in Instagram-based health campaigns [43]. The possibilities of the Instagram story continue to evolve, incorporating advanced features such as polls, quizzes, shopping links, countdowns, and swipe-up links. These stories can be saved beyond the typical 24-hour period using the “Highlights” feature, enhancing their utility for prolonged engagement [44]. Recent research underscores the importance of using relatable visuals, authentic language, humor, and culturally resonant messaging to align with the audience’s social and cultural identities for effective health promotion on Instagram [45-47]. Further, the necessity of addressing barriers such as digital literacy and privacy concerns and incorporating interactive elements to boost the accessibility and effectiveness of these interventions have been highlighted [48].

Despite these promising developments, however, evidence remains limited on how to use Instagram to deliver complex social and psychological interventions, since most existing interventions focus on single behaviors or brief interventions. To address this gap, we initiated a broader research project to explore how Instagram can be used as a delivery platform for structured and complex, interactive interventions (such as the leduin program) [49]. This paper focused on one component of that project: the feasibility evaluation of the leduin program. It built on the feasibility framework by Bowen et al [50], which has been widely applied across diverse health and digital intervention research fields [51-57].

Feasibility studies aim to assess whether the development and evaluation of an intervention can be realistically undertaken and whether it should proceed to further testing. They provide insights into implementation strategies and inform the design of larger trials [50,58,59]. Specifically, feasibility studies help determine the conditions under which an intervention might work prior to testing its actual effectiveness [60].

Social Media Feasibility Studies

Existing research provides mixed but instructive evidence on the feasibility of delivering interventions via Instagram. Some studies highlight the platform’s potential to reach adolescents effectively: For instance, Instagram ads recruited a diverse adolescent sample with over 96% retention at the 16-week follow-up [61], and brief video-based interventions targeting depression and anxiety symptoms were successfully delivered to youth [62]. Additionally, social media metrics have proven to be a practical tool for measuring engagement [46].

XSL•FO
RenderX

However, other findings underscore substantial limitations. In a human-centered outreach campaign, fewer than 25% of adolescents followed the Instagram account after 2 months, and only 7% of followers matched the intended target group [47]. A 12-week exercise intervention for young women showed high satisfaction but only 56% retention and low active engagement—just 30% consistently liked posts, and few commented or tagged others [63]. Similarly, a 13-week physical activity program reported good reception but poor survey completion rates, dropping from 33.3% at 4 weeks to 22.2% after 3 months [64].

The Leduin Program

Recognizing both the opportunities and risks of social media, we aimed to harness its potential while addressing its challenges. Accordingly, we developed the *leduin* program (short for "Lebenskompetent durch Instagram," meaning "Life skills through Instagram") for adolescents delivered via Instagram. It teaches life skills for use in both analog and digital settings (see Zimmermann and Tomczyk [65] for examples). It draws on the Ottawa Charter for Health Promotion [66] and the definition of life skills as outlined in the WHO's life skills framework [67], including decision-making and problem-solving, creative and critical thinking, communication and interpersonal skills, self-awareness and empathy, and coping with emotions and stress. In sum, these are fundamental skills to navigate the challenges of everyday life [15,16].

The development of the leduin program preceded this feasibility study. It followed the intervention mapping framework [68] and drew on established psychological theories as well as co-creation to guide content and structure (see Multimedia Appendix 1). Adolescent perspectives on program design and Instagram features guided the co-creation process [69]. The intervention lasts 14 weeks and includes weekly modules addressing individual, social, and health-related skills as shown in Figure 1. Each module introduces a real-world challenge and links it to digital life skills. It is delivered through Instagram stories and feed posts aligning with WHO life skills education guidelines [70]. The program is conveyed via a private Instagram account. Only adolescents who actively decided to participate in the program are accepted as followers, creating a safe space and relevant social comparison among peers. Features such as polls, sliders, question stickers, and quizzes enable active participation. Weekly self-care tasks, journaling activities (via an optional analog workbook), and interactive challenges with raffles are included. These also promote sustained engagement and spaced learning [71].

Figure 1. Module structure of the 14-week life skills program leduin.

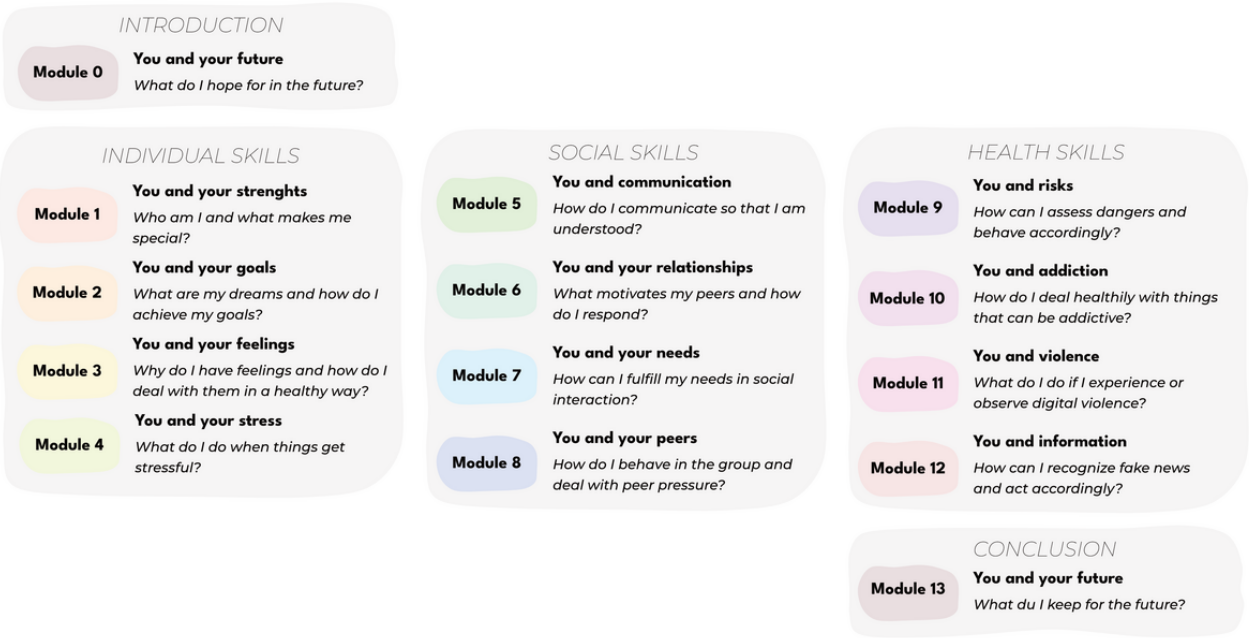

In sum, to the best of our knowledge, the leduin program is the first prevention initiative to be co-created and delivered entirely on Instagram, designed to strengthen both life skills and social media skills in adolescents. It provides an exclusive environment for this age group and leverages Instagram's interactive features to facilitate anonymous peer exchange and deliver multimedia content. The modular structure offers daily inputs, reflection tasks, and challenges, with learning seamlessly integrated into daily life and habits, thereby supporting behavior change.

Objectives

The aim of this study was to evaluate the feasibility of delivering a complex, theory-driven life skills intervention for adolescents via Instagram using an iterative process with active involvement of the target group [72]. As part of a broader research initiative exploring the use of social media for digital prevention, this pilot study focused on the leduin program—a 14-week intervention designed to promote digital life skills in adolescents through interactive, platform-native content.

XSL•FO
RenderX

Methods

Study Design and Reporting Standards

This paper was written in accordance with APA Style Mixed Methods Article Reporting Standards (MMARS) including the Journal Article Reporting Standards for Qualitative Research (JARS-Qual) and Quantitative Research (JARS-Quan) [73,74]. We further used the Consolidated Criteria for Reporting Qualitative Research (COREQ) [75].

Research Design and Theoretical Framework

This study used a pragmatic mixed methods approach, integrating both quantitative and qualitative data to provide a comprehensive analysis of the program [76]. Quantitative data were derived from Instagram usage metrics, directly reflecting participant behavior on the platform throughout the program. Qualitative data were gathered through interviews with program participants after program end. These interviews were analyzed using qualitative content analysis to deeply explore adolescents' perceptions of the program [77]. The qualitative content analysis was guided by the feasibility framework by Bowen et al [50], which comprises 8 key domains for evaluating whether and how an intervention can work: acceptability, demand, implementation, practicality, adaptation, integration, limited efficacy testing, and expansion potential.

The unique nature of a social media–facilitated intervention required adaptations to this framework to reflect the specific conditions of digital delivery. Unlike traditional interventions, Instagram-based programs are centrally administered and do not rely on repeated implementation by local practitioners. Although traditional interventions often depend on face-to-face facilitation and organizational infrastructure, social media interventions are disseminated digitally, engage users asynchronously, and can also be evaluated through platform-based engagement metrics. In the adapted version of the framework by Bowen et al [50] used in this study, the core meaning of each domain was retained but reinterpreted to align with the Instagram delivery format (see Table 1). Please see Multimedia Appendix 2 for more details.

Table . Overview of the feasibility framework by Bowen et al [50], including the domains and their adaptation for the Instagram-facilitated life skills program leduin.

Domain	What the domain evaluates	How it can be assessed
Acceptability	How well the target audience receives the intervention methods and design	Participant feedback on the program usability; engagement metrics of various features
Demand	Evidence of interest and participation by the target population of the intervention content	Participant feedback on the content and its relevance to their lives; engagement metrics on usage over time and the demand of different topics
Implementation	Whether the digital program delivery met adolescents' expectations and when engagement declined during the intervention	Participant feedback on expectations; analysis of drop-offs across the intervention period
Practicality	Feasibility of integrating the program into adolescents' daily routines	Participant feedback on time investment and use; quantitative weekday/view analysis
Adaptation	Flexibility of the intervention to meet target group needs or platform requirements	Considered conceptually; program intentionally fixed
Integration	Fit with existing social and institutional structures	Evaluation of the recruitment process to motivate program participation
Limited efficacy testing	Preliminary evidence that the intervention produces desired outcomes	Participants' subjective learning experiences and perceived changes in behavior
Expansion	Scalability of the intervention to other groups or contexts	Considered conceptually; scalability inferred from platform reach and modular program structure; calculation of staff hours and cost/participant

Ethical Considerations

The study was approved by the Ethics Committee at University Medicine Greifswald (BB 060/22). Approved activities included implementing the 14-week prevention program leduin on Instagram and conducting subsequent interviews with the participants postprogram.

All participants and their guardians were fully informed about the study's objectives, methodologies, and data protection measures. To guarantee voluntary participation, informed consent was obtained within the school setting, with consent forms collected by the teachers accompanying the program. These forms were completed and submitted before the beginning of the leduin program and subsequent interviews.

Data collection and storage were handled in strict compliance with German data protection laws, ensuring the anonymity of all data to protect participant privacy. The publication omits all personal identifiers, maintaining participant confidentiality throughout the study.

There was no financial compensation nor other incentives offered to participants for their involvement in the study. Recruitment aimed to foster genuine interest and engagement in the program. Participation in certain more demanding

XSL•FO
RenderX

challenges was encouraged by the opportunity to win prizes in a raffle.

No photographs of participants were captured. Accordingly, images are absent from both the manuscript and any supplementary materials. Consequently, no separate consent for images was necessary.

Research Flow and Team

The leduin program began in October 2022 and ran for 14 weeks, with a 2-week break over the Christmas holidays, concluding in February 2023. Usage data were collected throughout, and interviews were conducted from February 2023 to April 2023 by 4 female researchers with psychology backgrounds, all affiliated with the University of Greifswald. The team included 2 undergraduate students and 2 students with bachelor's degrees, part of a research group studying social media's impact on health care.

These researchers were closely involved in the planning and implementation of the leduin program, which provided valuable contextual knowledge for the evaluation. At the same time, the team was aware of their dual role as program implementers and evaluators and the potential influence this could have on data collection and interpretation. To mitigate bias and minimize power dynamics, especially between adult researchers and adolescent participants, the main account holder—the visible figure during the program—did not conduct any interviews. Furthermore, the research team engaged in regular debriefings and reflexive discussions throughout the study to reflect on their positionality, discuss ethical considerations, train communication styles with adolescents, and critically examine how their roles and assumptions might shape interactions with participants and the analysis of data.

Participants and Recruitment

Our study used a purposive sampling method, targeting adolescents in 9th and 10th grades aged 14 years to 18 years across schools in North Rhine-Westphalia and youth centers throughout Germany. This demographic was chosen because it represents a critical phase of development: Social norms strongly shape behavior and healthy habits formed during this time often persist into adulthood. Moreover, these grades are especially important as they prepare students for the transition to secondary education or graduation, highlighting their significance for timely intervention efforts.

To facilitate recruitment, the secretariats of several schools and youth centers were contacted via phone and email to engage teachers and social workers interested in implementing our program. Although the program was delivered primarily through Instagram, teachers and social workers played a key organizational role in recruiting participants. Additionally, they received weekly newsletters on program content and participant engagement, along with optional teaching materials for each module block.

Information sessions were conducted with interested classes and youth groups to thoroughly explain the program's objectives, methods, and data collection procedures. Adolescents willing to participate had to return a jointly signed consent form that included their own informed assent and the written consent of their legal guardians, submitted via their educational or social work facilitators to the leduin team. Participation was restricted to those who were at least 14 years old and proficient in German to ensure comprehension of the program content.

In sum, 99 adolescents took part in the program: 62 (63%) identified as women, and 37 (37%) identified as men. The mean age was 15.2 (SD 0.74) years, ranging from 14 years to 18 years. Adolescents came from highly diverse transnational family contexts. Of the 99 participants, 60 (61%) had a migration background: 50 (50/99, 51%) had both parents born outside of Germany, and 10 (10/99, 10%) had 1 parent born abroad. The remaining 39 participants (39/99, 39%) had no migration background; both parents were born in Germany. Among those with a migration background, the most frequently reported countries of parental origin were Kosovo (mentioned 16 times), Ukraine (11 times), Morocco (11 times), Serbia (10 times), and Russia (5 times). Other countries mentioned more than once included Kazakhstan, Pakistan, Armenia, Iran, and the Netherlands. In total, parents were reported to have been born in over 25 different countries, covering regions across Eastern and Southeastern Europe, Central and South Asia, North Africa, and Western Europe. Of the 99 participating adolescents, 10 (10%) were recruited via a youth center and 89 (90%) via 6 different schools. These included 4 secondary schools (63/99, 64%) and 2 grammar schools (26/99, 26%).

Data Collection

Quantitative Data Collection

Quantitative data were collected on both the recruitment process and program engagement. During recruitment, the research team documented process indicators, including the number of schools and youth centers contacted, participation in information events, and the number of students who voluntarily enrolled in the program. Throughout the 14-week intervention, 2 trained researchers manually recorded Instagram usage data in structured Excel (Microsoft Corp) sheets. This approach was necessary because the intervention was delivered via a private Instagram account, which limited access to Instagram's native analytics features. We deliberately prioritized this privacy setting due to ethical considerations: The program was designed to be accessible only to adolescents who had actively consented to participation, along with consent from their caregivers. The private account ensured that only enrolled participants could view and interact with the content. This setup also guaranteed that peer exchange remained confined to the target group, with no possibility for teachers, parents, or external observers to access posts or responses. This privacy was essential to fostering a sense of trust, safety, and openness, especially since many of the program topics touched on emotionally sensitive or personally relevant issues. To address the limitations introduced by manual tracking, 2 trained researchers jointly recorded all engagement data using a standardized protocol to ensure accuracy and consistency. The resulting data informed the evaluation of 4 feasibility domains: acceptability, demand, implementation, and practicality (see Table 2 for an overview).

XSL•FO
RenderX

Table . Instagram usage data reflecting the interactions of participants with the program content and relevant feasibility framework domains.

Data	What the data say	Additional information	Relevant to evaluate the following domain(s)
Follower amount	Number of users who followed the Instagram account on a given day	Participants could join or unfollow the private account throughout the intervention due to organizational flexibility.	Integration
Module	Weekly module number (0 - 13)	Each week addressed a different life skills topic.	Practicality
Day of the week	Day on which content was published	Content was distributed daily, from Monday through Sunday.	Practicality
Story views	Number of views each story slide received	Stories consisted of multiple slides, each with unique content.	All
Quizzes	Number of quiz responses and number of correct answers	Multiple-choice quizzes were related to each module's content.	Acceptability
Polls	Number of votes submitted in polls	Polls allowed anonymous voting on module-relevant questions.	Acceptability
Poll by team	Number of responses to polls posted by the leduin team	Polls were created and shared by the program team.	Acceptability
Poll by users	Number of responses to polls based on peer answers	Participants' anonymous answers were shared, and peers could vote on shared experiences.	Acceptability
Poll for challenge	Number of responses to polls asking for preferred challenge rewards	Polls asked participants to choose among possible raffle rewards.	Acceptability
Question stickers	Number of short answers submitted via question stickers	These were used to collect open-text responses to prompts or questions.	Acceptability
Challenge	Number of challenge submissions received	Participants edited screenshots of story slides and submitted them via direct message.	Acceptability
Slider	Number of times the slider feature was used	Sliders were used for self-assessment of skills or mood.	Acceptability

Qualitative Data Collection

Qualitative data were collected through semistructured interviews conducted after the conclusion of the leduin program. The primary objective was to gather adolescents' perspectives to inform the feasibility evaluation and guide further program refinement. Interviews were conducted individually or in a group setting and followed ethical procedures. These included verbal reminders of participants' rights to voluntary participation, withdrawal at any time without consequences, and assured confidentiality.

To capture a range of experiences, participants were purposively sampled based on their level of engagement with the program's interactive challenges: high (more than 5 participations), medium (1 to 5 participations), and low or no participation. Only adolescents with prior consent for follow-up were contacted via direct Instagram messages. In total, 13 individual interviews were conducted: 8 orally via video conferencing (audio recorded and transcribed) and 4 in written form via Instagram direct messages. One group interview with 5 adolescents (3 women, 2 men) was held at a youth center, organized by a social worker. Despite efforts to balance gender representation, the final sample consisted of 11 women and 6 men (see Table 3 for an overview). Participants' ages ranged from 14 years to 18 years (mean 15.29, SD 0.99 years). Engagement levels varied: 6 participants showed high engagement, 5 had medium engagement, and 6 had low engagement. The interviews lasted on average 25 minutes and 11 seconds (SD 6 minutes and 34 seconds; see Table 3 for an overview).

XSL•FO
RenderX

Table . Participants of semistructured interviews with 17 adolescents (11 women, 6 men) aged 14 years through 18 years with different engagement levels during the leduin program.

Interview	Type	Duration (minutes)	Gender	Age (years)	Level of engagement
01	Oral	20:29	Woman	14	High
02	Oral	24:38	Man	15	High
03	Oral	21:13	Woman	15	High
04	Oral	34:51	Woman	15	Medium
05	Oral	21:35	Man	16	Medium
06	Oral	14:46	Woman	14	High
07	Oral	25:20	Woman	16	High
08	Oral	31:43	Man	15	Medium
09	Written	N/A[a]	Man	15	Medium
10	Written	N/A	Woman	15	Low
11	Written	N/A	Woman	15	High
12	Written	N/A	Woman	16	Low
13	Group video conference	32:12	3 women, 2 men	Women: 16, 16, 18; men: 14, 15	Women: low, medium, low; men: low, low

[a]N/A: not applicable.

Interviews were guided by a semistructured protocol designed according to the recommendations for qualitative research with adolescents, allowing for both flexibility and structured inquiry [78,79]. This approach allowed spontaneous questions and topic shifts, fostering authentic dialogue and aiding the evaluation of complex subjects [80]. Interviewers could adapt to participants' responses, ask clarifying questions, and pursue emerging themes, which supported the collection of authentic and context-rich data.

The interview guidelines focused on 5 of the 8 feasibility domains by Bowen et al [50] relevant to adolescent perspectives: acceptance, demand, implementation, practicality, and limited efficacy testing. Questions were designed to align with the theoretical framework while remaining accessible to adolescent participants. An overview of interview topics and their corresponding feasibility domains can be found in Multimedia Appendix 3, and the corresponding interview guide can be found in Multimedia Appendix 4.

Individual interviews were conducted online in private one-on-one settings to protect confidentiality. Adolescents joined the interviews from home. In the session held at a youth center, a social worker was also present to provide supervision, adhering to the protocols for conducting research with minors.

To enhance authenticity and reduce social desirability bias, all interviewers received intensive training based on the methodological approach to interviewing in social work by Widulle [81] and Reinders' [78] recommendations for conducting qualitative research with adolescents [82]. Training emphasized relationship-sensitive interviewing, active listening, conversational warm-up techniques, mirroring responses, and empathetic communication to build trust and promote open dialogue.

Audio recordings were transcribed according to the guidelines by Dresing and Pehl [83]. Due to logistical constraints, transcripts were not returned to participants for verification. All materials used or generated in the qualitative phase, including anonymized transcripts and the coding framework, are available in the Open Science Framework (OSF) online repository [84].

Data Analysis

To evaluate the feasibility of the leduin program, we applied a mixed methods design guided by the adapted framework by Bowen et al [50]. Quantitative analyses of Instagram usage data addressed 6 of the 8 framework domains, while qualitative data from semistructured interviews provided in-depth insights into 5 domains. Together, these complementary approaches enabled a comprehensive assessment of the program's delivery, engagement, relevance, and perceived impact.

Quantitative Data Analysis

Quantitative analyses were conducted in RStudio (Posit Software, PBC) using R (Version 2024.12.0+467). Statistical analyses were supported by the following R packages: dplyr for data wrangling, ggplot2 for data visualization, car for Levene test of homogeneity of variance, rstatix for Welch ANOVA and Games-Howell post hoc tests, and lmtest for linear model diagnostics (eg, Breusch-Pagan test). These packages were used to perform descriptive statistics, regression models, nonparametric tests, and visualization of engagement metrics. The full analysis script is available via the OSF [84]. We analyzed the feasibility domains of acceptability, demand, implementation, practicality, integration, and expansion. Descriptive statistics and inferential methods were applied as appropriate. All significance tests were 2-tailed with an alpha level of $P<.05$ unless stated otherwise. Detailed analytic procedures for each domain are provided in Multimedia Appendix 5.

XSL•FO
RenderX

Qualitative Data Analysis

Qualitative data underwent content analysis based on the methods outlined by Mayring and Fenzl [77], utilizing MAXQDA (VERBI software). This content analysis focused on the dimensions of acceptability, demand, implementation, practicality, and limited efficacy testing. A deductive-inductive coding approach was used: A preliminary category system based on the framework by Bowen et al [50] guided the initial coding, and new categories were added inductively after reviewing 25% of the data. This approach ensured both theoretical grounding and responsiveness to participant insights.

The unit of analysis was the individual interview. Adolescent statements were analyzed at a moderate level of abstraction to maintain depth and thematic clarity. The selection criteria focused on statements from young people regarding their evaluation of the program and its feasibility.

The coding team consisted of the first author and 2 trained research assistants. Intercoder reliability reached κ=0.90 after iterative refinement, indicating a high level of agreement among coders [85,86]. Disagreements were resolved through discussion until consensus.

Reflexivity was emphasized throughout the coding process. Coders were trained to identify their own assumptions and ensure neutrality, particularly when interpreting subjective feedback [82]. The final category system included domains, subcategories, definitions, and anchor examples. It was continually adjusted to best represent the data (see Multimedia Appendix 6).

Finally, we mapped summarized results onto the original feasibility domains to confirm that our analysis addressed the study's core research questions. Figure 2 summarizes the content analysis process.

XSL•FO
RenderX

Figure 2. Qualitative content analysis process for adolescent preferences on the feasibility of the Instagram-facilitated life skills program leduin.

XSL•FO
RenderX

Results

Overview

The following results draw on both quantitative and qualitative data to assess the feasibility of the leduin program. Quantitative findings are based on Instagram usage metrics and reflect behavioral engagement patterns across the entire sample (n=99). In contrast, qualitative findings stem from in-depth interviews with a purposively selected subsample (n=13), capturing participants' subjective perceptions. Interview participants were selected based on their level of challenge participation—categorized as high (>5 participations), medium (1 to 5 participations), and low or none—to reflect varying degrees of engagement.

Within the qualitative results, high and medium challenge engagement was often associated with more differentiated feedback, greater motivation, and stronger signs of impact. However, consistently low or no engagement tended to co-occur with limited feedback and interest, reduced program consumption, and an overall weaker connection to the program. These patterns help contextualize the qualitative findings and offer a nuanced understanding of how adolescents experienced the program.

Acceptability

Retention Rates

Retention rates were calculated daily as the proportion of accounts that viewed the entire story, determined using the viewing of the final story frame. The rates ranged from 64.5% to 100% across the 14-week program, with views of each story ranging from a minimum of 34 to a maximum of 81. The average daily retention rate was 87.7% (SD 7.8%), indicating consistently high viewer engagement. Visual inspection suggested a slight dip in retention during the early weeks, followed by a stabilization in later weeks. Weekly boxplots confirmed stable medians with minor fluctuations and no substantial drop-offs, supporting the interpretation of sustained engagement throughout the intervention (see Figure 3).

Figure 3. Retention rate distribution by week for the 14-week life skills program leduin.

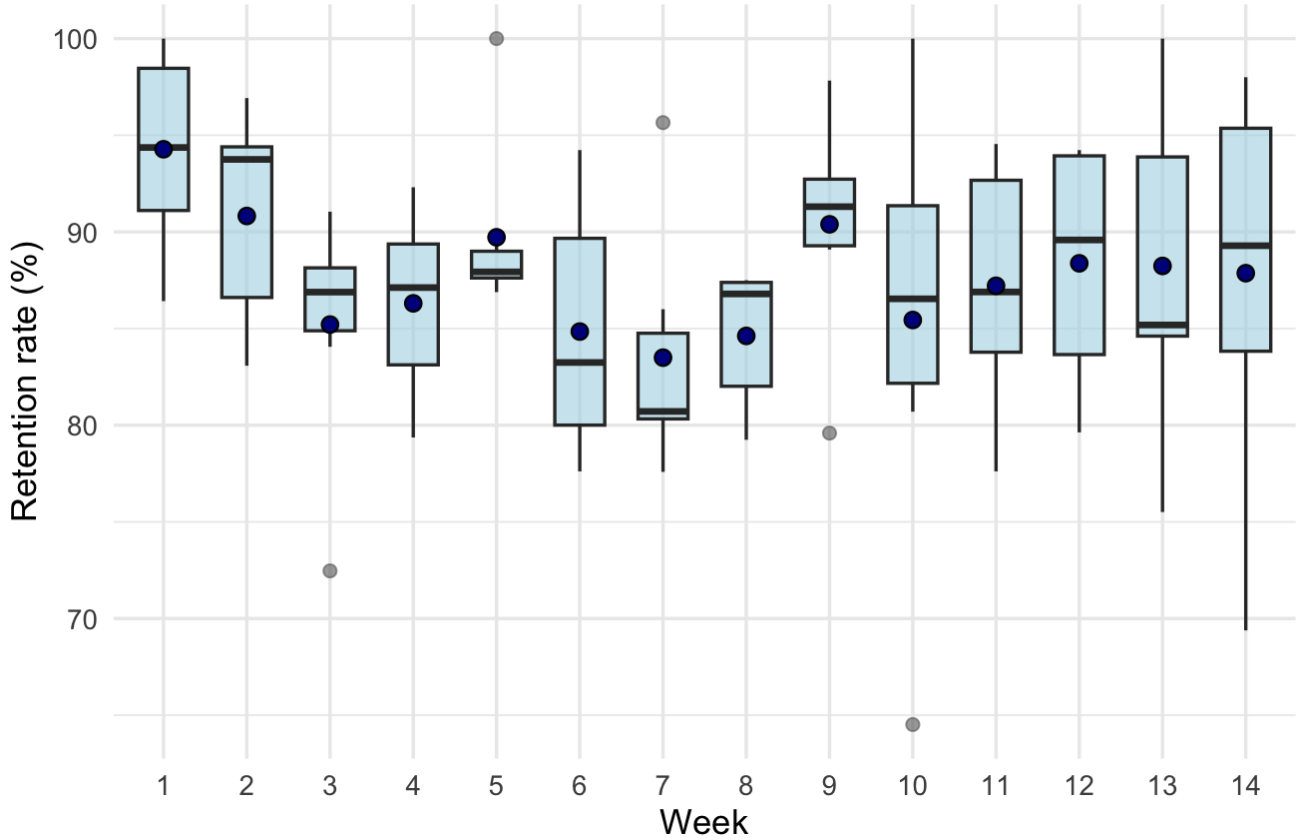

Interaction Rates

Interaction rates were determined as the proportion of accounts viewing the feature who actually interacted with it, and mean interaction rates were determined across all instances of each feature over the 14 weeks. Poll by team (mean 56.8%), quiz (mean 56.6%), poll by users (mean 54.4%), and slider (mean 51.2%) formed the group with the highest interaction rates, with no significant differences among them (poll by team vs quiz: P=.99; poll by team vs poll by users: P=.98; poll by team vs slider: P=.99; quiz vs poll by users: P=.97; quiz vs slider: P=.98; poll by users vs slider: P=.98). Poll for challenge (mean 44.2%) was significantly lower than these 4 (P<.001) but higher than challenge (mean 21.7%) and question sticker (mean 20.6%), which did not differ significantly from each other (challenge vs question sticker: P≥.99). Question sticker Sunday (mean 1.5%) had the lowest interaction rate, significantly lower than all other features (P<.001). These results demonstrate high content engagement through retention and substantial variability in interaction, depending on feature type, as visualized in Figure 4.

XSL•FO
RenderX

Figure 4. Interaction rates by story feature.

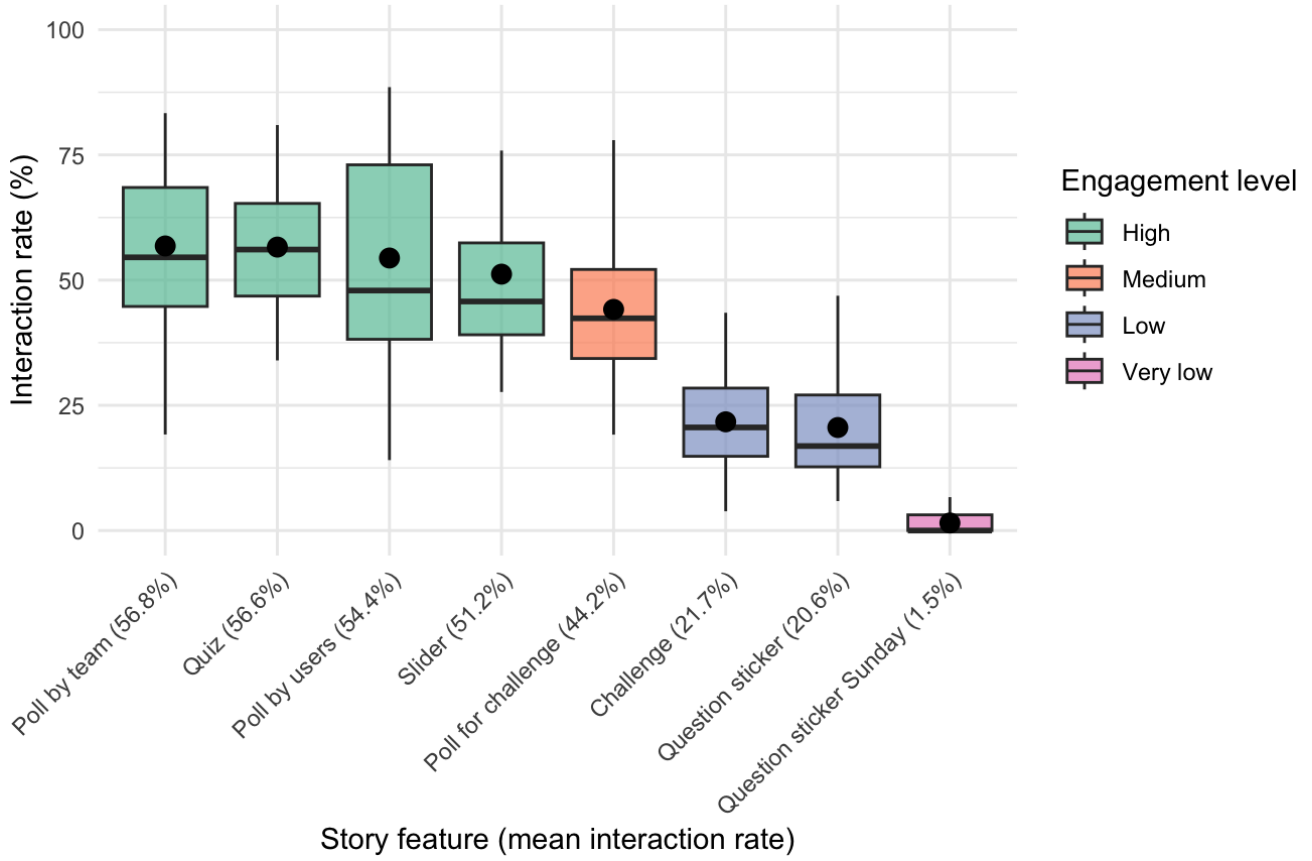

Qualitative Acceptability Findings

General Evaluation of the Program

Participants perceived the program as highly valuable and relevant to their everyday challenges, including planning for the future, managing relationships, and regulating emotions. The promotion of life skills was seen as particularly important, with several participants noting a lack of such content in traditional education. The program was experienced as a safe space fostering self-reflection and personal growth. One participant summarized the following:

> *We live in times when young people struggle to find their way...I think this program is a great starting point....We don't learn this in school....Maybe it will evolve or serve as a foundation for helping young people develop essential life skills....People my age, around 16, can find in this program a valuable retreat to grow and cope with their struggles.* [interview 3, section 103]

Many participants stated that they would recommend the program to their peers.

Evaluation of the Account Holder or Team

Participants described the program team as competent, trustworthy, and supportive. They reported feeling taken seriously, describing the team as inclusive and treating participants with equality and respect. At the same time, some adolescents expressed a desire for a more personalized and less research-influenced approach.

Language

Overall, adolescents found the program's language appropriate and easy to understand. As one participant noted: "The language was well-suited – not overly sophisticated, but exactly at our level" (interview 06, sections 11 - 12). This view reflected the majority opinion. However, 2 contrasting views emerged: One found the tone too formal, while another found it too simplistic and suggested more technical terms for better engagement.

Other Design Aspects

Participants praised the program's well-structured and engaging design, which was free of unnecessary or dull elements. The playful and colorful layout—with vibrant graphics, illustrations, images, and practical tips—was seen as visually appealing and helpful for maintaining attention. One participant stated:

> [The program was creatively designed, visually appealing, and colorful, which made reading enjoyable and helped maintain focus.] [interview 04, section 35]

Interactivity and Anonymous Exchange

Interactive Instagram features such as question stickers, polls, and quizzes were highly appreciated, especially for enabling anonymous participation. In this context, anonymity not only meant expressing one's own thoughts without revealing identity but also engaging in anonymous social comparison—seeing aggregated group responses without knowing who gave which answer. This allowed participants to compare their experiences with others, recognize commonalities, and feel less alone in their struggles. Many noted that anonymity enabled engagement,

XSL•FO
RenderX

as discussing sensitive topics with strangers felt uncomfortable—an openness that they felt was impossible in a school context. One participant noted the following:

> *Being anonymous is great because it allows you to be more open and honest. You can complete the tasks based on what you truly think, without worrying that someone will recognize you. I also liked that questions were answered and that we received a message after each challenge. It was nice to see that people appreciated our participation.* [interview 01, sections 28-40]

Interaction with the program team was also valued. Participants appreciated receiving responses to their questions and personalized messages following challenge completion. Although the direct communication was initially surprising to some, it became a welcomed and appreciated feature.

Suggestions for improvement included expanding anonymous interaction options, such as incorporating anonymous video calls, and offering opportunities for one-on-one conversations with the team to discuss personal matters more deeply.

Instagram Features in Detail

Adolescents interacted with the program's Instagram features in diverse ways, ranging from occasional to frequent use. The variety of available formats catered to different preferences and engagement styles. Low-threshold features like polls and quizzes were particularly well received for their ease of use, while more involved formats such as challenges or open-ended questions appealed to those seeking deeper engagement. Many participants valued the overall variety rather than preferring a single feature.

Despite this positive reception, several usability barriers emerged across features. Adolescents reported that cognitively demanding or time-consuming formats could lead to disengagement. Suggestions included simplifying language, improving visual clarity, and limiting content length.

Reels and Stories

The program's reels and stories were regularly consumed, reflecting Instagram's integration into adolescents' daily routines. Reels were generally perceived as informative and concise; however, some lacked engaging visuals or felt cognitively demanding. Participants suggested making them shorter and more visually dynamic. Similarly, long or content-heavy stories were sometimes skipped, which disrupted comprehension. Some participants preferred static posts for their clearer structure and easier readability. Reducing the number of story slides per topic was a common suggestion to enhance accessibility.

Question Stickers, Quizzes, and Polls

The interactive features in the Instagram story were evaluated positively, although engagement levels varied. Polls stood out for their simplicity and one-click ease. Their immediacy and low demand made them a consistent tool for engagement. Quizzes were seen as fun and educational, with instant feedback enhancing motivation. However, mismatched answer options occasionally caused confusion. Question stickers encouraged self-expression and peer engagement but required significant effort. Although some used them actively, others ignored them due to the mental load of formulating answers.

Challenges

By those who participated in them, challenges were enthusiastically received. Many enjoyed the creative tasks like fill-in-the-template exercises, and small prizes boosted motivation. Still, others found multistep tasks too effortful. One participant noted: "It was exhausting to screenshot the stories and transfer them to my iPad because I couldn't edit them on my phone" (interview 06, section 18). Others missed challenge posts or felt less motivated after winning once. Confusing instructions, especially when introductory content was missed, also posed barriers.

Demand

Story views showed a significant negative correlation with time (r=–0.70, P<.001), indicating a steady decline in engagement, with a maximum of 81 views and a minimum of 34 views. A linear regression analysis confirmed this trend ($ß$=–1.24, SE=0.13, P<.001, R^2=0.49). However, residual diagnostics indicated nonlinearity, prompting a segmented regression analysis. The segmented model revealed a sharp decline during the first 4 weeks ($ß$=–4.20, SE=0.74, P<.001), followed by a nonsignificant slope thereafter ($ß$=–0.29, SE=0.18, P=.10). Figure 5 visualizes this segmented pattern clearly, showing the steep drop in story views during the early weeks and the relative stabilization of engagement thereafter. This model explained 69% of the variance and met the assumption of homoscedasticity (Breusch-Pagan test, P=.33). These results indicate that engagement declined primarily in the early phase and then stabilized, suggesting that demand was shaped more by timing than by specific content.

XSL•FO
RenderX

Figure 5. Comparison of story view decline between early weeks (1-4) and later weeks (5-14).

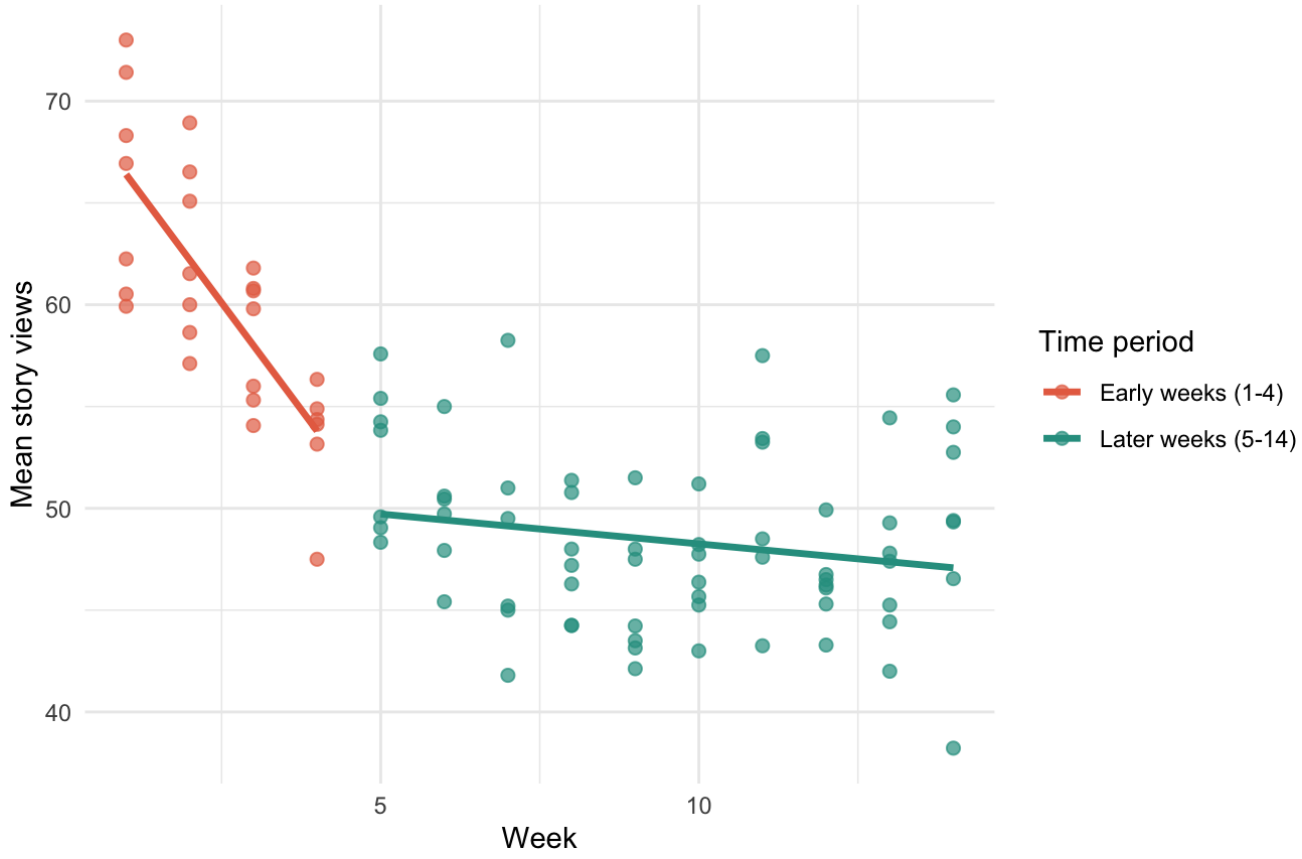

Qualitative Demand Findings

Personal Relevance

Many adolescents reported a strong personal connection to the program content. Topics such as hate speech, social media use, cyberbullying, and substance use were perceived as especially relevant. Social media was frequently cited as an area where participants struggled and appreciated guidance. One participant reflected: "I found the content about social media the most helpful because so many young people, including myself, struggle with it. It was really valuable to receive education on this topic" (interview 08, section 2). Others noted how the content helped them better understand the blurry line between jokes and harmful behavior online.

Stress management was another core concern, with 10th grade described as a particularly stressful time. Adolescents also valued guidance on goal setting and future planning, often expressing uncertainty about their goals. Furthermore, interpersonal relationships, emotional regulation, and communication were considered highly relevant. As one participant put it: "The program focused a lot on friendships, social interactions, handling emotions, and conflict resolution, which I found really, really interesting" (interview 04, section 3). Self-care, self-reflection, and identity development were also frequently mentioned as particularly impactful. One participant shared: "The self-care challenge...really opened my eyes" (interview 05, section 4).

However, not all topics resonated equally. Some adolescents found content such as source evaluation or addiction awareness less relevant, particularly if these were already covered in other settings. Although some found all topics manageable, others described certain subjects such as hate speech or family conflicts as emotionally challenging or overwhelming.

Repetition

Opinions on the program's variety were mixed among participants. Several participants appreciated the rotation of weekly topics and the structured, balanced presentation, describing the content as well-paced and thoughtfully sequenced. One participant commented: "I had the feeling that all the topics were somehow treated equally well and that it was well timed and built on each other" (interview 06, section 23).

Others felt that the program included too much repetition. The recurring Monday self-care challenge and continuous emphasis on behavioral change were seen by some as redundant or even demotivating. As one participant stated: "At some point, it gets boring to keep seeing or hearing the same thing — that you're doing something wrong or that you need to improve" (interview 07, section 23).

Implementation

Consistent with the demand analysis, story views declined sharply in the early weeks and stabilized thereafter. By week 4, 76% of the total drop in views had occurred, decreasing from a mean view of 66.1 in week 1 to 53.4 in week 4. By week 6, this figure rose to 97.3% (mean views 49.9). Although minor fluctuations were observed in later weeks, weekly averages remained stable, indicating that most disengagement happened early in the intervention.

XSL•FO
RenderX

Qualitative Implementation Findings

Overall, adolescents reported that the program met their expectations. Most felt well-informed about its structure and content in advance, which helped avoid surprises during participation. They had anticipated a focus on social interactions, personal challenges, and mental health support, along with learning opportunities—areas they felt were well addressed. Some participants who were initially skeptical noted that the program exceeded their expectations.

Practicality

Content was successfully published according to the predefined plan. Posts were scheduled around 2 PM to align with school dismissal times, ensuring availability throughout the afternoon and evening.

Mean views were distributed equally across weekdays (Kruskal-Wallis χ^2_6=0.92, *P*=.99). To examine whether content structure influenced retention, we explored associations between the number of story slides and viewer retention, as well as differences across weekdays. Story slides varied significantly during the program, ranging from 3 to 25 slides (mean 9.5, SD 4.9 slides). The Spearman rank correlation between the number of story slides and retention rate, however, was not statistically significant (ρ=0.11, *P*=.28), indicating no monotonic relationship between content length and viewer retention. Similarly, the correlation between weekday (Monday to Sunday) and retention rate was nonsignificant (ρ=–0.07, *P*=.50). These findings suggest that neither the amount of content per day nor the weekday of publication significantly influenced retention.

Qualitative Practicality Findings

Total Duration and Daily Time Expenditure

Adolescents had mixed views on the program's overall duration. Many found the 3-month period manageable and felt time passed quickly due to regular engagement and other commitments. One participant stated: "I found it very entertaining, it was very short. Yes, it was over quicker than expected" (interview 06, section 35). Others, however, felt it was too long, which led to decreased motivation and drop-off.

The daily time commitment, however, was generally seen as appropriate. Many noted that tasks could be completed in a few minutes and fit well into their Instagram routines. One participant explained: "The time required was fine since the program only took a few minutes per day" (interview 05, section 14). Still, some adolescents found it harder to keep up during busy periods due to academic demands.

General Frequency of Use

Engagement levels varied. Some participants followed the program closely, interacting several times per week or almost daily. One participant stated: "I saw most of it – I didn't really miss anything" (interview 01, section 4). Others had moderate engagement, while a few interacted with only a small portion of the content and completed few or no tasks.

Pace of Topic Delivery

Feedback on pacing was also mixed. Some adolescents appreciated the structure and felt the time for each topic was sufficient. One participant noted: "I think the timing of the topics was appropriate. They weren't covered too quickly, and there was enough time between them" (interview 01, section 16). Others found certain topics too long or overwhelming when presented with too much information at once.

Difficulty

Most participants found the content accessible and well-designed. Challenges were seen as doable, and videos were considered clear and easy to follow. However, tasks requiring deep reflection or decision-making in ambiguous situations were perceived as more difficult. One participant explained: "Too difficult? Well, there were challenges where I had to think more, but it was still appropriate, yes. So it was still doable." (interview 02, section 42).

Integration

To recruit adolescents for the leduin program, we contacted schools in North Rhine-Westphalia and youth centers across Germany. In total, we contacted 101 schools: 52 grammar schools, 35 secondary schools, and 14 free schools. We contacted 10 youth centers, and we held 31 information events in 19 schools and 1 youth center (reflecting 20, or 18.2%, of the initially contacted 111 institutions) with approximately 775 students, with an average of 25 students per class.

Of the 775 students reached through the information events, 99 decided to participate in the study, 10 (10%) coming from 1 youth center and 89 (90%) from 6 different schools (4 secondary schools and 2 grammar schools), resulting in an overall participation rate of 12.8% (99/775). At Instagram, a maximum of 98 followers was noted, showing that nearly all participants who agreed to participate in the study and the program also followed the account. These results indicate that initially reaching the target group was hard despite high interest from teachers and social workers.

Limited Efficacy Testing

Subjective Learning Success

Many adolescents reported meaningful personal growth from participating in the program. They described increased self-confidence, enhanced stress management, and a more constructive approach to everyday challenges. Participants noted greater self-awareness, authenticity, and attentiveness to their own needs, often applying strategies such as writing to-do lists to stay organized. The program also encouraged reflection, normalized mistakes, and supported the development of clearer future goals and a more optimistic outlook.

Social benefits were equally prominent. Adolescents described improved communication and emotional expression, greater empathy, and stronger conflict resolution skills. Many felt more comfortable forming new friendships and more capable in social situations. The program also influenced how participants engaged with social media. Increased awareness of cyberbullying, hate speech, social comparison, and digital well-being led to behavioral change, including unfollowing negative accounts, limiting screen time, reporting harmful content, and focusing on inspirational profiles. One participant summed up their experience:

XSL•FO
RenderX

> *I'm now more mindful of my Instagram use. I set time limits — sometimes 10 minutes, sometimes just 5. I only follow people who inspire and uplift me, which has helped my mental health. In my daily life, I apply new techniques. When I feel stressed, I create a to-do list and work through it. I've also started thinking about my future and setting goals. I try to stay positive, even in conflicts. I see mistakes as learning opportunities and try to grow from them. My communication has improved, both in real life and on Instagram. I can express my feelings better, which has strengthened my relationships. What I've taken away from the leduin program is that making mistakes is okay and that you should always keep your goals in mind.* [interview 07, sections 40-55]

Limitations and Lack of Perceived Impact

Although most participants reported positive outcomes, some felt the program had limited relevance or impact. Some attributed this to prior knowledge, noting they learned little that was new. One person reported no significant change in their behavior or social interactions, stating the following:

> *I didn't notice any changes in my behavior on Instagram that could be directly linked to the program. Maybe it's because I didn't follow along consistently. Unfortunately, I also didn't find any real-life applications for what was taught — neither in how I treat myself nor in how I interact with others. There's nothing I personally took away from the leduin program or benefited from.* [interview 03, sections 66-77]

Adaptation

Adaptation considers the extent to which an intervention can be modified to suit different contexts or audiences. In this study, the domain was not empirically assessed but considered conceptionally, as the program was intentionally designed as a fixed 14-week intervention. This structure was grounded in established theory and prior evidence from behavioral change research and life skills education and developed using a co-design approach to ensure alignment with adolescents' needs and preferences. Although the program was not designed for real-time modification, it demonstrated responsiveness through interactive features and anonymous exchange, including reposting of participant input. These elements created space for individualized engagement, even though the content itself remained unchanged. Accordingly, the program's limited flexible adaptation options were a conscious decision to maintain its theoretical integrity and consistency.

In addition to conceptual scalability, we estimated the staff resources required for recruitment and delivery. The calculation included initial and follow-up contact with 111 institutions (15 minutes each; ≈27.75 hours); 31 online information events including preparation (1 hour each; ≈31 hours), coordination, and consent form processing for 99 participants (10 minutes each; ≈16.5 hours); daily content posting and moderation during the 14-week program (1 hour/day; 98 hours); and weekly newsletters to institutions (15 minutes/wk; ≈3.5 hours). In total, these activities accounted for ≈155 hours, corresponding to an average of about 94 minutes (1 hour, 34 minutes) of staff time per participant, excluding evaluation activities.

Discussion

Study Overview

In this pilot study, we used a mixed methods approach to evaluate the feasibility of leduin, a 14-week Instagram-based prevention program designed to promote life skills and functional social media use among adolescents. A total of 99 participants aged 14 years to 18 years were recruited through schools and youth centers in Germany. The program delivered short, interactive daily content via a private Instagram account. Feasibility was assessed through an adapted version of the framework by Bowen et al [50], combining quantitative platform usage metrics with qualitative in-depth postprogram interviews to capture adolescents' experiences and perspectives.

Integration of Feasibility Domains

Taken together, the leduin program demonstrated high feasibility, with its theory-based and co-designed format aligning closely with adolescents' preferences and daily habits. Findings indicate high acceptability, supported by a mean retention rate of 87.7% across the 14-week program, surpassing industry benchmarks of approximately 73% retention by the 10th frame, which corresponded to the average story length in this study [87]. Retention serves as a strong indicator of acceptance, especially given that Instagram stories function independently, do not compete with other content, and, once started, remain unaffected by algorithmic influence. These rates remained stable regardless of the number of story slides (ranging from 3 to 25) or the weekday of posting and did not follow the expected linear decline. In some cases, view counts even increased toward later story frames, suggesting that adolescents engaged selectively with content rather than dropping off systematically. This highlights the importance of interpreting retention within context rather than presuming a continuous decrease.

High engagement was further reflected in interaction rates, which revealed a clear hierarchy based on the effort required. Low-effort features showed the highest interaction: team polls (56.8%), quizzes (56.6%), user-generated polls (54.4%), and sliders (51.2%), with adolescents favoring brief, relatable formats that enabled peer comparison. The lack of significant difference between team- and user-generated polls suggests that topic relevance outweighed the importance of source. Polls linked to challenge prizes (44.2%) showed lower engagement, likely due to reduced emotional or social resonance. Interaction dropped further for more cognitively and technically demanding features such as challenges (21.7%) and question stickers (20.6%). The open-ended Sunday questions recorded the lowest interaction rate (1.5%), likely due to limited perceived relevance.

Qualitative data further supported the high acceptability of the program's language, design, and structure. Participants described the content as visually appealing, accessible, and thoughtfully presented. Interactive elements, particularly those allowing anonymous engagement, were valued for fostering openness and inclusion. These findings align with prior research

XSL•FO
RenderX

highlighting the effectiveness of relatable visuals, authentic language, culturally resonant messaging, and interactive features in social media-based health interventions [45-48].

Moreover, adolescents considered the daily time requirement manageable, the pacing appropriate, and most tasks feasible, even though perceived difficulty varied individually. This indicates that the program integrated well into daily routines. Its flexible and low-effort design aligns with research emphasizing the unique advantages of social media as a platform for youth health promotion such as cost-effectiveness, accessibility, and seamless integration into everyday behavior [3,21-25]. In sum, the program's methods and structure enabled interactivity and individualized engagement without requiring real-time content adjustments, thereby maintaining theoretical integrity and consistency.

In doing so, the program addressed key developmental tasks and interests. After an initial drop in story views during the first 4 weeks, interest stabilized without topic-specific spikes, indicating sustained and evenly distributed demand across all content. This pattern likely reflects a common phenomenon in digital health interventions: Adolescents may begin with a willingness to participate but disengage when content does not feel immediately relevant [88,89]. Content interaction is further influenced by Instagram's algorithm: However, as academic evaluations of Instagram's algorithmic principles are scarce, such insights rely largely on industry analyses [90]. These indicate that content visibility is strongly influenced by the platform's algorithm, which prioritizes posts and stories from accounts with which users interact most frequently. Consequently, low interaction early in the program may have triggered reduced algorithmic ranking, meaning fewer followers were shown subsequent posts in their feeds or story lists. This is because Instagram's algorithm weighs both an individual's past engagement with the account and the overall interaction rates across the audience, so lower aggregate engagement can diminish visibility even for followers who have not actively disengaged. This creates a feedback loop in which reduced exposure further limits collective and individual opportunities for re-engagement, potentially accelerating dropout. Although this dynamic challenges scalability, the stabilization of views in later weeks suggests that participants who found the content meaningful continued engaging despite algorithmic filtering. That supports the feasibility of sustained delivery for an interested core audience.

Qualitative data corroborated this interpretation, as most adolescents described the topics as relevant to their daily lives. However, preferences for specific content areas and views on content repetition varied individually. Functional social media use, cyberbullying, and digital violence were frequently cited as particularly meaningful. These concerns align with existing evidence on adolescents' vulnerability to problematic social media use, including exposure to harmful content, cyberbullying, and addictive platform mechanisms [13-16]. These risks correlate with mental health issues such as depression, anxiety, and stress [17-19], reinforcing the need for interventions that promote emotional resilience and digital literacy [20]. Accordingly, many participants reported increased awareness of their media use and a growing desire to adopt more intentional digital habits.

In line with the WHO life skills framework [67] and the Ottawa Charter for Health Promotion [66], participants also identified interpersonal relationships, emotional regulation, and school-related stress as central concerns. This alignment indicates the program's developmental relevance. Delivered in 10th grade—a transitional phase marked by academic pressure, peer dynamics, and important life decisions—the program was well timed, consistent with findings showing that life skills interventions are particularly effective during this stage [70,91].

Despite high engagement and perceived relevance during the program, recruiting adolescents initially proved challenging. Outreach through schools and youth centers resulted in a participation rate of only 12.8% among the 775 students directly approached through information events, with the research context likely serving as an additional barrier. Unlike traditional school-based programs that reach adolescents within the classroom [92], leduin relied on voluntary participation during adolescents' free time. Although this approach limited the sample size, it generated a focused but self-selected group. Nearly all who enrolled followed the Instagram account, indicating alignment between expressed interest and actual behavior. Qualitative data confirmed this: Most participants stated that the program met or exceeded their expectations, which likely contributed to continued engagement. However, the recruitment challenge remains. Digital outreach, although an intuitive alternative, does not ensure better access. For example, even a well-targeted Instagram campaign failed to reach their intended audience, with less than 7% of followers ultimately belonging to the target group [47]. Engaging adolescents in voluntary prevention programs thus remains a key concern for future efforts. At the same time, despite the challenges in recruitment, the resources invested per participant—about 94 minutes for recruitment, program implementation, and organizational support—can be considered economical given that each adolescent received a 14-week intervention with daily input and first indications of positive impact.

Overall, the findings suggest that a structured, Instagram-based life skills program offers a feasible and promising approach to adolescent health promotion. Its modular structure, developmental fit, and flexible delivery model provide a strong foundation for expansion to other target groups or more specific health-related topics.

Strengths and Limitations

This study offers several notable strengths that enhance its methodological integrity and relevance for advancing research on digital prevention. First, the mixed methods design, based on the feasibility framework by Bowen et al [50], enabled a comprehensive evaluation by combining behavioral Instagram engagement data with adolescents' subjective perceptions of their motivations, the program's perceived impact, and usability.

Although usage data reflect direct engagement with the content and program methods, the semistructured interviews allowed for an in-depth, participant-driven exploration of program

XSL•FO
RenderX

perceptions, barriers, and outcomes [80]. The deductive-inductive coding strategy ensured that analysis was grounded in theory while remaining open to emergent insights. High intercoder reliability (κ=0.90) confirms the consistency of the qualitative analysis [86], and the reflexive approach adopted by the research team helped to minimize bias during interpretation [82].

However, the study also presents limitations that restrict the generalizability and interpretability of the findings. Most notably, the sample was limited to adolescents in Germany, largely drawn from schools and youth centers in specific regions, which limits external validity. This is because health perceptions and social media use are influenced by cultural and regional differences regarding, for example, usage patterns, privacy concerns, or self-disclosure habits [93-96]. Voluntary participation may have led to self-selection bias, overrepresenting adolescents with higher motivation or interest in psychological topics [97]. Although valuable for understanding feasibility, this limits insights into broader population-level uptake and effectiveness.

To address this, it is useful to consider cross-cultural perspectives on digital literacy and social media use. A recent overview synthesizes global evidence, showing that, in many majority-world countries, social media plays an even greater role in young people's daily lives than the internet in general [98]. At the same time, access is less universal than in Western countries, digital skills are often lower, and social media is more heavily relied upon for information seeking. The overview also highlights gender inequalities, fewer opportunities for parental or school-based digital literacy support, and less regulation for youth. Taken together, these conditions suggest that using social media to impart skills that protect and empower young people could be even more critical in these contexts than in settings like Germany.

Further, the integration of the intervention into school settings also proved difficult, with a low participation rate of 12.8% despite strong institutional support. This challenge in reaching adolescents outside of class time highlights a key barrier to scalability and suggests that more targeted recruitment strategies are needed in future studies as well as further insights into what factors support program participation.

Additionally, social desirability bias may have influenced interview responses, despite the team's efforts to create a safe, trusting environment. The dual role of the researchers as program implementers and interviewers could have subtly shaped how adolescents expressed their views [99], even though the main account holder did not conduct interviews to reduce this risk. Furthermore, member checking was not performed, meaning that participants did not have the opportunity to review or confirm the researchers' interpretations of their interviews, potentially affecting the accuracy of qualitative findings [100].

Finally, although the study demonstrates promising results in terms of perceived impact and engagement, no long-term follow-up was conducted, and effectiveness was only evaluated through subjective self-report. This limits conclusions about sustained behavioral change or broader health outcomes. Although we deliberately chose to use a private Instagram account to ensure confidentiality, trust, and privacy, incorporating direct Instagram analytics in future studies would be valuable. Such data, alongside smartphone-based passive sensing combined with self-report, could enrich feasibility evaluations by providing objective verification of engagement patterns, contextualizing self-reported experiences, and quantifying relationships between online activity and well-being. Although these approaches come with ethical and practical challenges, particularly regarding privacy and compliance, they hold considerable potential for capturing adolescents' behaviors more comprehensively and for linking subjective perceptions with observable usage data [101,102].

Implications for Future Research, Policy, and Practice

Despite its limitations, this study provides robust preliminary evidence on the feasibility of using Instagram to promote life skills among adolescents and offers important methodological guidance for digital prevention research. Future research should focus on strategies to improve initial recruitment and reduce early dropout, both of which pose key challenges in voluntary, free-time programs. To address self-selection bias and enhance generalizability, future studies should include more diverse and geographically broader samples. Beyond subjective reports, controlled studies are needed to assess the program's efficacy and long-term impact. Research should also explore algorithm-informed content strategies to optimize visibility and engagement, as well as the expansion potential of Instagram-based interventions for other age groups and health topics.

In practice, this study provides actionable insights for designing engaging social media interventions. Usage metrics revealed clear preferences: Low-effort features (eg, polls, sliders) encourage routine interaction, while more cognitively demanding features (eg, challenges, open questions) are best used strategically when deeper reflection is desired. This underscores the importance of feature sensitivity and intentional design. The sustained engagement observed supports the feasibility of delivering even complex interventions via Instagram when content is relevant and well-integrated into adolescents' daily routines. Co-design was essential for aligning with adolescents' preferences, reinforcing the importance of participatory development. Notably, the high demand for emotionally and socially relevant topics confirms the need for prevention programs that directly address the realities of adolescents' lives.

Policy should acknowledge social media platforms as viable channels for health promotion. This includes supporting the development and evaluation of evidence-based, culturally sensitive digital interventions and providing funding frameworks that support innovative formats to go beyond traditional school-based models. National prevention strategies should explicitly include digital interventions as low-threshold, scalable complements to in-person services particularly for reaching youth where they already are: on social media.

Conclusions

This pilot study demonstrates that delivering a structured, theory-based life skills intervention via Instagram is both

XSL•FO
RenderX

feasible and promising. High engagement rates, positive user feedback, and meaningful self-reported outcomes suggest that social media can serve as a powerful platform for adolescent health promotion, especially when interventions are co-designed, culturally aligned, and intentionally integrated into daily routines. Although initial recruitment and early disengagement remain key challenges, the overall findings underscore Instagram's potential for complex, sustained interventions. Building on this foundation, future research and policy should continue exploring how to effectively harness digital environments to support adolescent development and mental well-being.

Acknowledgments

We thank the Heinrich Böll Foundation for supporting our research. We further thank Anna-Lea Voigt, Leandra Heußmann, Elisabeth Seidel, Simon Barth, Maren Güßmann, Emine Celebi, Sina Jünemann, Lea Weiß, and Hannah Venz for supporting the implementation of the study, including recruitment, information events, program execution on Instagram, and conducting and evaluating the interviews. R Wizard (Open AI) was used to improve the R codes for analyzing the data, and DALL-E was used to create the table of contents image.

Data Availability

All data used in this study can be found in the open repository OSF [84].

Conflicts of Interest

None declared.

Multimedia Appendix 1
Description of the development of the leduin program.
[DOCX File, 39 KB - formative_v9i1e78774_app1.docx]

Multimedia Appendix 2
Adaptation of the feasibility framework by Bowen et al [50] for this study.
[DOCX File, 15 KB - formative_v9i1e78774_app2.docx]

Multimedia Appendix 3
Overview of interview topics and their corresponding feasibility domains.
[DOCX File, 18 KB - formative_v9i1e78774_app3.docx]

Multimedia Appendix 4
Interview guide for semistructured interviews with participants of the leduin program.
[PDF File, 276 KB - formative_v9i1e78774_app4.pdf]

Multimedia Appendix 5
Detailed analytic procedures for each domain.
[DOCX File, 16 KB - formative_v9i1e78774_app5.docx]

Multimedia Appendix 6
Category system for qualitative content analysis of semistructured interviews with adolescents.
[XLSX File, 12 KB - formative_v9i1e78774_app6.xlsx]

References

1. Number of internet and social media users worldwide as of april 2024. Statista. 2024. URL: https://www.statista.com/statistics/617136/digital-population-worldwide/ [accessed 2024-05-23]
2. The second decade: improving adolescent health and development. World Health Organization. 2021. URL: https://iris.who.int/server/api/core/bitstreams/4f346197-f9aa-435d-91bf-570f145d8f19/content [accessed 2025-11-09]
3. Faverio M, Sidoti O. Teens, social media and technology 2024. Pew Research Center. 2024. URL: https://www.pewresearch.org/internet/2024/12/12/teens-social-media-and-technology-2024/ [accessed 2025-04-30]
4. Korda H, Itani Z. Harnessing social media for health promotion and behavior change. Health Promot Pract 2013 Jan;14(1):15-23. [doi: 10.1177/1524839911405850] [Medline: 21558472]

XSL•FO
RenderX

5. Nesi J, Choukas-Bradley S, Prinstein MJ. Transformation of adolescent peer relations in the social media context: part 1-a theoretical framework and application to dyadic peer relationships. Clin Child Fam Psychol Rev 2018 Sep;21(3):267-294. [doi: 10.1007/s10567-018-0261-x] [Medline: 29627907]
6. Moorhead SA, Hazlett DE, Harrison L, Carroll JK, Irwin A, Hoving C. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. J Med Internet Res 2013 Apr 23;15(4):e85. [doi: 10.2196/jmir.1933] [Medline: 23615206]
7. Subrahmanyam K, Smahel D. Digital Youth: The Role of Media in Development: Springer; 2011. [doi: 10.1007/978-1-4419-6278-2]
8. Wallace B, Varcoe C, Holmes C, et al. Towards health equity for people experiencing chronic pain and social marginalization. Int J Equity Health 2021 Feb 2;20(1):53. [doi: 10.1186/s12939-021-01394-6] [Medline: 33531018]
9. Robards F, Kang M, Usherwood T, Sanci L. How marginalized young people access, engage with, and navigate health-care systems in the digital age: systematic review. J Adolesc Health 2018 Apr;62(4):365-381. [doi: 10.1016/j.jadohealth.2017.10.018] [Medline: 29429819]
10. Gudka M, Gardiner KLK, Lomas T. Towards a framework for flourishing through social media: a systematic review of 118 research studies. J Posit Psychol 2023 Jan 2;18(1):86-105. [doi: 10.1080/17439760.2021.1991447]
11. Zuboff S. The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power: PublicAffairs; 2019. URL: https://www.publicaffairsbooks.com/titles/shoshana-zuboff/the-age-of-surveillance-capitalism/9781610395694/ [accessed 2025-11-18]
12. Allcott H, Braghieri L, Eichmeyer S, Gentzkow M. The welfare effects of social media. American Economic Review 2020 Mar 1;110(3):629-676. [doi: 10.1257/aer.20190658]
13. Crone EA, Konijn EA. Media use and brain development during adolescence. Nat Commun 2018 Feb 21;9(1):588. [doi: 10.1038/s41467-018-03126-x] [Medline: 29467362]
14. Livingstone S, Helsper E. Balancing opportunities and risks in teenagers' use of the internet: the role of online skills and internet self-efficacy. New Media & Society 2010 Mar;12(2):309-329. [doi: 10.1177/1461444809342697]
15. Festl R. Social media literacy & adolescent social online behavior in Germany. J Child Media 2021 Apr 3;15(2):249-271. [doi: 10.1080/17482798.2020.1770110]
16. iTM. Digital skills gap index 2021. Initiative D21. URL: https://initiatived21.de/publikationen/digital-skills-gap-2021 [accessed 2025-11-10]
17. Huang C. A meta-analysis of the problematic social media use and mental health. Int J Soc Psychiatry 2022 Feb;68(1):12-33. [doi: 10.1177/0020764020978434] [Medline: 33295241]
18. Shannon H, Bush K, Villeneuve PJ, Hellemans KGC, Guimond S. Problematic social media use in adolescents and young adults: systematic review and meta-analysis. JMIR Ment Health 2022 Apr 14;9(4):e33450. [doi: 10.2196/33450] [Medline: 35436240]
19. Vannucci A, Simpson EG, Gagnon S, Ohannessian CM. Social media use and risky behaviors in adolescents: a meta-analysis. J Adolesc 2020 Feb;79:258-274. [doi: 10.1016/j.adolescence.2020.01.014] [Medline: 32018149]
20. Polanco-Levicán K, Salvo-Garrido S. Understanding social media literacy: a systematic review of the concept and its competences. Int J Environ Res Public Health 2022 Jul 20;19(14):8807. [doi: 10.3390/ijerph19148807] [Medline: 35886657]
21. Welch V, Petkovic J, Pardo Pardo J, Rader T, Tugwell P. Interactive social media interventions to promote health equity: an overview of reviews. Health Promot Chronic Dis Prev Can 2016 Apr;36(4):63-75. [doi: 10.24095/hpcdp.36.4.01] [Medline: 27077792]
22. Yardley L, Choudhury T, Patrick K, Michie S. Current issues and future directions for research into digital behavior change interventions. Am J Prev Med 2016 Nov;51(5):814-815. [doi: 10.1016/j.amepre.2016.07.019] [Medline: 27745680]
23. Tsaltskan V, Sanchez Baez R, Firestein GS. Cost-effectiveness of social media advertising as a recruitment tool: a systematic review and meta-analysis. J Clin Transl Sci 2023;7(1):e180. [doi: 10.1017/cts.2023.596] [Medline: 37745929]
24. Saleem M, Kühne L, De Santis KK, Christianson L, Brand T, Busse H. Understanding engagement strategies in digital interventions for mental health promotion: scoping review. JMIR Ment Health 2021 Dec 20;8(12):e30000. [doi: 10.2196/30000] [Medline: 34931995]
25. Borghouts J, Eikey E, Mark G, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. J Med Internet Res 2021 Mar 24;23(3):e24387. [doi: 10.2196/24387] [Medline: 33759801]
26. Guse K, Levine D, Martins S, et al. Interventions using new digital media to improve adolescent sexual health: a systematic review. J Adolesc Health 2012 Dec;51(6):535-543. [doi: 10.1016/j.jadohealth.2012.03.014] [Medline: 23174462]
27. Laranjo L, Arguel A, Neves AL, et al. The influence of social networking sites on health behavior change: a systematic review and meta-analysis. J Am Med Inform Assoc 2015 Jan;22(1):243-256. [doi: 10.1136/amiajnl-2014-002841] [Medline: 25005606]
28. Brzozowska JM, Gotlib J. Social media potential and impact on changing behaviors and actions in skin health promotion: systematic review. J Med Internet Res 2025 Jan 6;27:e54241. [doi: 10.2196/54241] [Medline: 39761100]
29. Tibaijuka M. Influence of interactive learning tools on student engagement in online courses in Tanzania. Am J Online Distance Learn 2024;6(2):47-57. [doi: 10.47672/ajodl.2322]

XSL•FO
RenderX

30. Bitrián P, Buil I, Catalán S. Enhancing user engagement: the role of gamification in mobile apps. J Bus Res 2021 Aug;132:170-185. [doi: 10.1016/j.jbusres.2021.04.028]
31. Morris RR, Schueller SM, Picard RW. Efficacy of a web-based, crowdsourced peer-to-peer cognitive reappraisal platform for depression: randomized controlled trial. J Med Internet Res 2015 Mar 30;17(3):e72. [doi: 10.2196/jmir.4167] [Medline: 25835472]
32. Carter L, Rogith D, Franklin A, Myneni S. NewCope: a theory-linked mobile application for stress education and management. Stud Health Technol Inform 2019 Aug 21;264:1150-1154. [doi: 10.3233/SHTI190406] [Medline: 31438105]
33. Fleming TM, Bavin L, Stasiak K, et al. Serious games and gamification for mental health: current status and promising directions. Front Psychiatry 2017 Jan 10;7:215. [doi: 10.3389/fpsyt.2016.00215] [Medline: 28119636]
34. Bamidis PD, Gabarron E, Hors-Fraile S, Konstantinidis E, Konstantinidis S, Rivera O. Gamification and behavioral change: techniques for health social media. In: Syed-abdul S, Gabarron E, Lau AYS, editors. Participatory Health Through Social Media: Elsevier Inc; 2016:112-135. [doi: 10.1016/B978-0-12-809269-9.00007-4]
35. Khadjesari Z, Murray E, Kalaitzaki E, et al. Impact and costs of incentives to reduce attrition in online trials: two randomized controlled trials. J Med Internet Res 2011 Mar 2;13(1):e26. [doi: 10.2196/jmir.1523] [Medline: 21371988]
36. Bakker D, Kazantzis N, Rickwood D, Rickard N. Development and pilot evaluation of smartphone-delivered cognitive behavior therapy strategies for mood- and anxiety-related problems: MoodMission. Cogn Behav Pract 2018 Nov;25(4):496-514. [doi: 10.1016/j.cbpra.2018.07.002]
37. Linardon J, Messer M, Shatte A, et al. Does the method of content delivery matter? Randomized controlled comparison of an internet-based intervention for eating disorder symptoms with and without interactive functionality. Behav Ther 2022 May;53(3):508-520. [doi: 10.1016/j.beth.2021.12.001] [Medline: 35473653]
38. Veale HJ, Sacks-Davis R, Weaver ER, Pedrana AE, Stoové MA, Hellard ME. The use of social networking platforms for sexual health promotion: identifying key strategies for successful user engagement. BMC Public Health 2015 Feb 6;15(1):85. [doi: 10.1186/s12889-015-1396-z] [Medline: 25884461]
39. Chan V, Allman-Farinelli M. Young Australian adults prefer video posts for eissemination of nutritional information over the social media platform Instagram: a pilot cross-sectional survey. Nutrients 2022 Oct 19;14(20):4382. [doi: 10.3390/nu14204382] [Medline: 36297066]
40. Caliandro A, Graham J. Studying Instagram beyond selfies. Social Media + Society 2020 Apr;6(2). [doi: 10.1177/2056305120924779]
41. Koinig I. Picturing mental health on Instagram: insights from a quantitative study using different content formats. Int J Environ Res Public Health 2022 Jan 30;19(3):1608. [doi: 10.3390/ijerph19031608] [Medline: 35162631]
42. Goodyear VA, Wood G, Skinner B, Thompson JL. The effect of social media interventions on physical activity and dietary behaviours in young people and adults: a systematic review. Int J Behav Nutr Phys Act 2021 Jun 5;18(1):72. [doi: 10.1186/s12966-021-01138-3] [Medline: 34090469]
43. Edney S, Bogomolova S, Ryan J, Olds T, Sanders I, Maher C. Creating engaging health promotion campaigns on social media: observations and lessons from Fitbit and Garmin. J Med Internet Res 2018 Dec 10;20(12):e10911. [doi: 10.2196/10911] [Medline: 30530449]
44. Stories. Instagram Help Center. 2024. URL: https://help.instagram.com/1660923094227526 [accessed 2024-10-22]
45. Johnson O, Delk DW. #Wematter: creating culturally tailored health promotion content for Black and Latina college women on Instagram. Digit Health 2024;10:20552076241241919. [doi: 10.1177/20552076241241919] [Medline: 38596403]
46. O'Kane N, McKinley MC, Gough A, Hunter RF. Investigating the feasibility and acceptability of using Instagram to engage post-graduate students in a mass communication social media-based health intervention, #WeeStepsToHealth. Pilot Feasibility Stud 2022 Dec 12;8(1):254. [doi: 10.1186/s40814-022-01207-9] [Medline: 36510310]
47. Thomas VL, Chavez M, Browne EN, Minnis AM. Instagram as a tool for study engagement and community building among adolescents: a social media pilot study. Digit Health 2020;6:2055207620904548. [doi: 10.1177/2055207620904548] [Medline: 32215216]
48. Malloy JA, Kemper JA, Partridge SR, Roy R. Empowering young women: a qualitative co-design study of a social media health promotion programme. Nutrients 2024 Mar 9;16(6):780. [doi: 10.3390/nu16060780] [Medline: 38542691]
49. Zimmermann E, Tomczyk S. Fostering digital life skills through social media with adolescents in 6 German states: protocol for an accessibility study according to the RE-AIM framework. JMIR Res Protoc 2024 Apr 17;13(1):e51085. [doi: 10.2196/51085] [Medline: 38631035]
50. Bowen DJ, Kreuter M, Spring B, et al. How we design feasibility studies. Am J Prev Med 2009 May;36(5):452-457. [doi: 10.1016/j.amepre.2009.02.002] [Medline: 19362699]
51. Schaap FD, Fokkens AS, Dijkstra GJ, Reijneveld SA, Finnema EJ. Dementia care mapping to support staff in the care of people with intellectual disability and dementia: a feasibility study. J Appl Res Intellect Disabil 2018 Nov;31(6):1071-1082. [doi: 10.1111/jar.12464] [Medline: 29691956]
52. Bourboulis S, Kemp E, Beatty L, the Finding My Way-Advanced Authorship Group. A qualitative exploration of factors that influence engagement with a digital mental health intervention for women with metastatic breast cancer: Finding My Way-Advanced. Support Care Cancer 2025 Mar 31;33(4):333. [doi: 10.1007/s00520-025-09379-9] [Medline: 40164760]

XSL•FO
RenderX

53. Laird B, Puzia M, Larkey L, Ehlers D, Huberty J. A mobile app for stress management in middle-aged men and women (Calm): feasibility randomized controlled trial. JMIR Form Res 2022 May 24;6(5):e30294. [doi: 10.2196/30294] [Medline: 34989677]
54. Massaro AF, Lecuona O, García-Rubio C, Castro-Paredes A. Bringing mindfulness-based relapse prevention for substance use disorders into Individual therapy with Spanish population: a feasibility and effectiveness study. Mindfulness (N Y) 2022 Mar;13(3):766-785. [doi: 10.1007/s12671-022-01835-5]
55. van der Smissen D, Overbeek A, van Dulmen S, et al. The feasibility and effectiveness of web-based advance care planning programs: scoping review. J Med Internet Res 2020 Mar 17;22(3):e15578. [doi: 10.2196/15578] [Medline: 32181750]
56. Al Rawwad T, Tata V, Wanat MA, Campbell D, Thornton D. Evaluating the feasibility of implementing a prescription drug misuse prevention intervention in the community: a mixed methods study. BMC Public Health 2023 Apr 21;23(1):728. [doi: 10.1186/s12889-023-15608-9] [Medline: 37085838]
57. Parker L, Maman S, Pettifor A, et al. Feasibility analysis of an evidence-based positive prevention intervention for youth living with HIV/AIDS in Kinshasa, Democratic Republic of the Congo. AIDS Educ Prev 2013 Apr;25(2):135-150. [doi: 10.1521/aeap.2013.25.2.135] [Medline: 23514081]
58. O'Cathain A, Hoddinott P, Lewin S, et al. Maximising the impact of qualitative research in feasibility studies for randomised controlled trials: guidance for researchers. Pilot Feasibility Stud 2015;1(1):32. [doi: 10.1186/s40814-015-0026-y] [Medline: 27965810]
59. Tickle-Degnen L. Nuts and bolts of conducting feasibility studies. Am J Occup Ther 2013;67(2):171-176. [doi: 10.5014/ajot.2013.006270] [Medline: 23433271]
60. Orsmond GI, Cohn ES. The distinctive features of a feasibility study: objectives and guiding questions. OTJR (Thorofare N J) 2015 Jul;35(3):169-177. [doi: 10.1177/1539449215578649] [Medline: 26594739]
61. Kutok ER, Doria N, Dunsiger S, et al. Feasibility and cost of using Instagram to recruit adolescents to a remote intervention. J Adolesc Health 2021 Nov;69(5):838-846. [doi: 10.1016/j.jadohealth.2021.04.021] [Medline: 34059428]
62. Amsalem D, Martin A, Dixon LB, Haim-Nachum S. Using Instagram to promote youth mental health: feasibility and acceptability of a brief social contact–based video intervention to reduce depression stigma. JAACAP Open 2025 Feb. [doi: 10.1016/j.jaacop.2024.12.004]
63. Curtis RG, Ryan JC, Edney SM, Maher CA. Can Instagram be used to deliver an evidence-based exercise program for young women? A process evaluation. BMC Public Health 2020 Oct 6;20(1):1506. [doi: 10.1186/s12889-020-09563-y] [Medline: 33023559]
64. Lewis ZH, Danayan S. The protocol and feasibility results of a preliminary Instagram-based physical activity promotion study. Technologies (Basel) 2021;9(4):70. [doi: 10.3390/technologies9040070]
65. Zimmermann E, Tomczyk S. Using social media to promote life skills among adolescents: a debate on opportunities, challenges, and implications for health and education. J Prev (2022) 2025 Apr;46(2):201-211. [doi: 10.1007/s10935-025-00826-1] [Medline: 39873826]
66. Ottawa Charter for Health Promotion. World Health Organization (WHO). 1986. URL: https://www.who.int/publications/i/item/ottawa-charter-for-health-promotion [accessed 2025-03-13]
67. Partners in life skills education - conclusions from a United Nations inter-agency meeting. Department of Mental Health World Health Organization Geneva. 1999. URL: https://pdf4pro.com/view/partners-in-life-skills-education-who-world-287ba2.html [accessed 2025-11-18]
68. Bartholomew Eldridge LK, Parcel GS, Kok G, Gottlieb N. Planning health promotion programs. In: Barholomew LK, editor. An Intervention Mapping Approach: Wiley; 2006.
69. Zimmermann E, Tomczyk S. The ways of using social media for health promotion among adolescents: qualitative interview and focus group study. J Med Internet Res 2025 Jun 9;27:e71510. [doi: 10.2196/71510] [Medline: 40489771]
70. Life skills education for children and adolescents in schools. World Health Organization. 1994. URL: https://apps.who.int/iris/handle/10665/63552 [accessed 2025-11-10]
71. Smolen P, Zhang Y, Byrne JH. The right time to learn: mechanisms and optimization of spaced learning. Nat Rev Neurosci 2016 Feb;17(2):77-88. [doi: 10.1038/nrn.2015.18] [Medline: 26806627]
72. O"Cathain A, Croot L, Duncan E, et al. Guidance on how to develop complex interventions to improve health and healthcare. BMJ Open 2019 Aug;9(8):e029954. [doi: 10.1136/bmjopen-2019-029954]
73. Appelbaum M, Cooper H, Kline RB, Mayo-Wilson E, Nezu AM, Rao SM. Journal article reporting standards for quantitative research in psychology: the APA Publications and Communications Board task force report. Am Psychol 2018 Jan;73(1):3-25. [doi: 10.1037/amp0000191] [Medline: 29345484]
74. Levitt HM, Bamberg M, Creswell JW, Frost DM, Josselson R, Suárez-Orozco C. Journal article reporting standards for qualitative primary, qualitative meta-analytic, and mixed methods research in psychology: the APA Publications and Communications Board task force report. Am Psychol 2018 Jan;73(1):26-46. [doi: 10.1037/amp0000151] [Medline: 29345485]
75. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]

XSL•FO
RenderX

76. Morgan DL. Pragmatism as a paradigm for mixed methods research. In: Integrating Qualitative and Quantitative Methods: A Pragmatic Approach: SAGE Publications, Inc; 2014:25-44 URL: https://methods.sagepub.com/book/integrating-qualitative-and-quantitative-methods-a-pragmatic-approach [accessed 2025-11-18] [doi: 10.4135/9781544304533.n2]
77. Mayring P, Fenzl T. In: Baur N, Blasius J, editors. Handbuch Methoden Der Empirischen Sozialforschung: Springer Fachmedien; 2019:633-648 URL: http://link.springer.com/10.1007/978-3-658-21308-4 [accessed 2025-11-18] [doi: 10.1007/978-3-658-21308-4]
78. Reinders H. Qualitative Interviews Mit Jugendlichen Führen: Ein Leitfaden: Oldenbourg Verlag München; 2012. [doi: 10.1524/9783486717600]
79. Adeoye‐Olatunde OA, Olenik NL. Research and scholarly methods: semi‐structured interviews. J Am Coll Clin Pharm 2021 Oct;4(10):1358-1367. [doi: 10.1002/jac5.1441]
80. Gill P, Stewart K, Treasure E, Chadwick B. Methods of data collection in qualitative research: interviews and focus groups. Br Dent J 2008 Mar 22;204(6):291-295. [doi: 10.1038/bdj.2008.192] [Medline: 18356873]
81. Widulle W. Gesprächsführung in Der Sozialen Arbeit: VS Verlag für Sozialwissenschaften; 2012. [doi: 10.1007/978-3-531-19118-8]
82. Finlay L. “Outing” the researcher: the provenance, process, and practice of reflexivity. Qual Health Res 2002 Apr;12(4):531-545. [doi: 10.1177/104973202129120052] [Medline: 11939252]
83. Dresing T, Pehl T. Transkriptionen qualitativer Daten. In: Mey G, Mruck K, editors. Handbuch Qualitative Forschung in Der Psychologie: Springer Fachmedien; 2017. URL: http://link.springer.com/10.1007/978-3-658-18387-5 [doi: 10.1007/978-3-658-18387-5_56-1]
84. Zimmermann E, Tomczyk S. Pilotstudy leduin material. Open Science Framework (OSF). 2025. URL: https://osf.io/umtga/overview [accessed 2025-11-09]
85. Lombard M, Snyder-Duch J, Bracken CC. Content analysis in mass communication: assessment and reporting of intercoder reliability. Human Comm Res 2002 Oct;28(4):587-604. [doi: 10.1111/j.1468-2958.2002.tb00826.x]
86. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159-174. [doi: 10.2307/2529310] [Medline: 843571]
87. 2024 Instagram stories benchmark report. RivalIQ. 2024 Jan 9. URL: https://www.rivaliq.com/blog/instagram-stories-benchmark-report/ [accessed 2025-05-20]
88. Hernández-Rodríguez JC, García-Muñoz C, Ortiz-Álvarez J, Saigí-Rubió F, Conejo-Mir J, Pereyra-Rodriguez JJ. Dropout rate in digital health interventions for the prevention of skin cancer: systematic review, meta-analysis, and metaregression. J Med Internet Res 2022 Dec 9;24(12):e42397. [doi: 10.2196/42397] [Medline: 36485027]
89. Torous J, Lipschitz J, Ng M, Firth J. Dropout rates in clinical trials of smartphone apps for depressive symptoms: a systematic review and meta-analysis. J Affect Disord 2020 Feb 15;263:413-419. [doi: 10.1016/j.jad.2019.11.167] [Medline: 31969272]
90. Hostrander R. The surprising truth about how Instagram story views are sorted. Clixie AI. 2025. URL: https://www.clixie.ai/blog/the-surprising-truth-about-how-instagram-story-views-are-sorted [accessed 2025-04-10]
91. Pedditzi ML, Nonnis M, Fadda R. Self-efficacy in life skills and satisfaction among adolescents in school transitions. J Public Health Res 2023 Oct;12(4):22799036231211420. [doi: 10.1177/22799036231211420] [Medline: 38020217]
92. Hale DR, Fitzgerald-Yau N, Viner RM. A systematic review of effective interventions for reducing multiple health risk behaviors in adolescence. Am J Public Health 2014 May;104(5):e19-e41. [doi: 10.2105/AJPH.2014.301874] [Medline: 24625172]
93. Polit DF, Beck CT. Generalization in quantitative and qualitative research: myths and strategies. Int J Nurs Stud 2010 Nov;47(11):1451-1458. [doi: 10.1016/j.ijnurstu.2010.06.004] [Medline: 20598692]
94. Trepte S, Masur PK. Cultural differences in social media use, privacy, and self-disclosure. Deutsche Nationalbibliothek. 2016. URL: http://opus.uni-hohenheim.de/volltexte/2016/1218/ [accessed 2025-11-10]
95. AlAfnan MA. Cultural and behavioral insights into European social media users: platform preferences and personality types. Stud Media Commun 2025;13(1):17. [doi: 10.11114/smc.v13i1.7306]
96. Aubel J, Chibanda D. The neglect of culture in global health research and practice. BMJ Glob Health 2022 Sep;7(9):e009914. [doi: 10.1136/bmjgh-2022-009914]
97. Hill LG, Rosenman R, Tennekoon V, Mandal B. Selection effects and prevention program outcomes. Prev Sci 2013 Dec;14(6):557-569. [doi: 10.1007/s11121-012-0342-x] [Medline: 23417667]
98. Magis-Weinberg L, Dodel M, Pangrazio L, Pathak-Shelat M, Tao S, Uzuegbunam CE. Global perspectives on youth and the digital environment: learnings from majority world countries. In: Christakis DA, Hale L, editors. Handbook of Children and Screens: Digital Media, Development, and Well-Being from Birth Through Adolescence: Springer, Cham; 2024:507-514. [doi: 10.1007/978-3-031-69362-5_69]
99. Bispo Júnior JP. Social desirability bias in qualitative health research. Rev Saude Publica 2022;56:101. [doi: 10.11606/s1518-8787.2022056004164] [Medline: 36515303]
100. Birt L, Scott S, Cavers D, Campbell C, Walter F. Member checking: a tool to enhance trustworthiness or merely a nod to validation? Qual Health Res 2016 Nov;26(13):1802-1811. [doi: 10.1177/1049732316654870] [Medline: 27340178]

XSL•FO
RenderX

101. Slade C, Sun Y, Chao WC, Chen CC, Benzo RM, Washington P. Current challenges and opportunities in active and passive data collection for mobile health sensing: a scoping review. JAMIA Open 2025 Aug;8(4):ooaf025. [doi: 10.1093/jamiaopen/ooaf025] [Medline: 40688708]
102. Keusch F, Conrad FG. Using smartphones to capture and combine self-reports and passively measured behavior in social research. J Surv Stat Methodol 2021 Sep 27;10(4):863-885. [doi: 10.1093/jssam/smab035]

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Research
JARS-Qual: Journal Article Reporting Standards for Qualitative Research
JARS-Quan: Journal Article Reporting Standards for Quantitative Research
MMARS: Mixed Methods Article Reporting Standards
OSF: Open Science Framework
WHO: World Health Organization

Edited by D Gan; submitted 19.06.25; peer-reviewed by C Okerulu, R Ndukwe, X Zheng; revised version received 26.08.25; accepted 10.10.25; published 27.11.25.

Please cite as:
Zimmermann E, Tomczyk S
Feasibility of the Social Media–Based Prevention Program "Leduin" for German Adolescents on Instagram: Mixed Methods Pilot Study
JMIR Form Res 2025;9:e78774
URL: https://formative.jmir.org/2025/1/e78774
doi:10.2196/78774

© Elizabeth Zimmermann, Samuel Tomczyk. Originally published in JMIR Formative Research (https://formative.jmir.org), 27.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Feasibility and Acceptability of a mHealth Patient Navigation Intervention to Increase Pre-Exposure Prophylaxis Uptake in Racially and Ethnically Diverse Sexual and Gender Minority Youth in Los Angeles (PrEPresent): Pilot Randomized Controlled Trial

Sam Calvetti[1], BA; Bryan Lei[2], MS; Jacob B Stocks[3], MSc; Matthew T Rosso[3,4], MPH; Manuel Puentes[1], BA; Ramon Durazo-Arvizu[2], PhD; Lindsay Slay[1], MSW; Michele D Kipke[1,5], PhD; Lisa B Hightow-Weidman[3], MD, MPH

[1]Pediatric Research Administration, Children's Hospital Los Angeles, 4650 Sunset Boulevard, Los Angeles, CA, United States

[2]Biostatistics and Data Management Core, The Saban Research Institute, Children's Hospital Los Angeles, Los Angeles, CA, United States

[3]Institute on Digital Health and Innovation, College of Nursing, Florida State University, Tallahassee, FL, United States

[4]Duke Clinical Research Institute, Duke University School of Medicine, Durham, NC, United States

[5]Department of Population and Public Health Sciences, Keck School of Medicine, University of Southern California, Los Angeles, CA, United States

Corresponding Author:
Sam Calvetti, BA
Pediatric Research Administration, Children's Hospital Los Angeles, 4650 Sunset Boulevard, Los Angeles, CA, United States

Abstract

Background: Pre-exposure prophylaxis (PrEP) is a powerful tool to prevent the transmission of HIV. Interventions promoting PrEP must focus on populations most impacted by systemic barriers to uptake. Historically, young sexual minority men and transgender women have the highest demonstrated rates of new HIV diagnoses, but prevalence within other gender minority populations is now being studied. Few interventions have focused on addressing PrEP uptake with sexual and gender minority (SGM) youth, particularly through mobile health (mHealth) technologies. Built on the successful foundation of the HealthMpowerment Platform, PrEPresent aimed to engage SGM youth across diverse gender, racial, and ethnic identities in the Greater Los Angeles area.

Objective: The aim of this study was to evaluate the feasibility and acceptability of a digital peer patient-navigation PrEP uptake app.

Methods: PrEPresent incorporated patient activation theory into an mHealth intervention. The study took place over a 6-month period with visits at baseline, 3 months, and 6 months. The intervention period lasted from baseline to 3 months. Control participants received an information-only app. Intervention participants received an enhanced app and access to an interventionist, the PrEPresentative. Intervention participants could meet with the PrEPresentative four times over the 3-month period via phone, Health Insurance Portability and Accountability Act–compliant videoconferencing, or an in-app text messaging. PrEP uptake was measured through survey responses and the UrSure rapid urine test of tenofovir.

Results: PrEPresent comprised of 147 sexual and gender-diverse participants—75 participants were randomized into the control arm and 72 into the intervention arm. A total of 48% (71/147) were Latinx and 18% (27/147) were Black or African American. Most (98/147, 67%) were transgender or gender diverse, and the remaining (49/147, 33%) were cisgender men. PrEP was initiated by 25% (14/56) of intervention participants and 19% (11/58) of control participants. In total, 50% (36/72) of intervention participants completed two or more sessions with the interventionist. Intervention participants had an average of 15.93 (SD 15.85) logins compared to 6.31 (SD 9.27) logins for control participants. Average use of the mHealth platform was 9.51 (SD 11.47) minutes for intervention participants and 3.03 (SD 5.70) minutes for control participants.

Conclusions: PrEPresent met primary outcome measures of feasibility and acceptability. Despite this, PrEP uptake was low, and use of the platform was low compared to other HealthMpowerment projects. While mHealth offers promising HIV prevention outcomes, fostering active app engagement is crucial in promoting behavior change. Mixed success in initiating PrEP uptake across mHealth interventions involving SGM youth warrants further inquiry into how these platforms can address prevention barriers with this population. Interventions targeting uptake and adherence will need to adapt as the landscape of PrEP delivery evolves with the adoption of on-demand and long-acting injectable modalities.

Trial Registration: ClinicalTrials.gov NCT05281393; https://clinicaltrials.gov/study/NCT05281393

International Registered Report Identifier (IRRID): RR2-10.2196/50866

XSL•FO
RenderX

(JMIR Form Res 2025;9:e69255) doi:10.2196/69255

KEYWORDS

PrEP; mobile health; LGBTQ; adolescent; transgender; HIV; pre-exposure prophylaxis

Introduction

Background

Despite a decrease in new HIV acquisitions within cisgender men (those who identify their gender with their sex assigned at birth) in the LGBTQ+ community, 67% of new HIV acquisitions are still attributed to sexual interactions within this group. Young sexual minority men, those between the ages of 13 and 34, make up just over half (56%) of new diagnoses [1]. Historical erasure of expansive gender identities has contributed to a lack of understanding of HIV likelihood and prevalence within transgender and nonbinary people (those who identify their gender differently from their sex assigned at birth) [2]. The most well-documented incidences are in transgender women, who have a higher rate of HIV acquisition than any other population subgroup. One count within major US cities in a 2019 - 2020 Center for Disease Control Special Report cited a 42% positivity rate within this population, with 18.9% of the sample living with HIV aged 18 - 29 years [3]. HIV prevention strategies must address the needs of populations like young sexual minority men and young transgender women while working to prioritize the inclusion of all sexual and gender minority (SGM) youth.

While similar HIV prevention issues have been noted for both young sexual minority men and gender minority youth populations, each group has unique needs to overcome barriers to care [4,5]. The intersection of marginalized identities can differently impact their ability to engage in the HIV care continuum [6]. Individual level issues (internalized stigmas, isolation, mental health, or substance use), social level issues (stigmas perpetuated by social relations through friends, family, or relationships), and institutional level issues (transphobia, homophobia, and racism) create health disparities among SGM people that impact their ability to access and continually engage in HIV prevention and treatment [7,8]. More research is needed to understand the individualized needs of SGM youth. Understanding overlapping needs could tailor engagement strategies to best engage people with SGM identities.

Pre-Exposure Prophylaxis Awareness, Utilization, and Uptake-Advancements in Pre-Exposure Prophylaxis

An important factor to end the HIV epidemic is increasing the uptake of pre-exposure prophylaxis (PrEP), a medication that can prevent the transmission of HIV [9]. In the United States, PrEP awareness among SGM youth continues to be high, but uptake is low [10,11]. More research is needed to understand how these factors impact SGM populations. Low uptake in this group has been attributed to different factors, such as a lack of trust in the medical system by transgender people interested in PrEP [12] and social stigma around PrEP disclosure due to negative connotations of PrEP users among Black SMM communities [13]. In a 2021 review of interventions targeting PrEP uptake, adherence, or persistence in the United States, 20 published interventions were identified, but only two listed uptake as an intended impact [14]. Additional studies are necessary to understand how interventions can increase this health behavior, particularly in SGM youth.

Not all forms or methods of PrEP are approved to be used by all people. The current approved forms of PrEP for people assigned female at birth who have receptive vaginal sex are daily oral Emtricitabine/Tenofovir Disoproxil Fumarate and Cabotegravir injectable [15]. Injectable PrEP was approved to be used in December 2021 [16]. Emtricitabine/Tenofovir Alafenamide (TAF), an alternative daily medication approved for PrEP in October 2019, is only for use in people assigned male at birth [17,18]. Although not US Food and Drug Administration–approved, guidance exists for on-demand PrEP (also called 2-1-1), but its efficacy has only been well-studied in gay and bisexual cisgender men [19]. Patient navigation, empowerment, and education around uptake are essential within SGM youth because of the availability and accessibility of different types of PrEP for different types of people.

Mobile Health Interventions

Mobile health (mHealth) interventions can play a key role in mitigating factors that contribute to HIV acquisition among SGM youth in the United States. These interventions have been shown to be acceptable in many contexts: in a national sample of PrEP adherence with a majority population of young sexual minority men of color [20], by increasing self-HIV testing within a cohort of 202 transgender youth recruited digitally [21], and in access to HIV/sexually transmitted infection (STI) at home testing kits, condoms, lube, and prevention information among transmasculine people living in Atlanta, Georgia, and Washington, DC [22]. mHealth has impacted HIV-related stigma in both people living with HIV and those not living with HIV [23]. There is a lack of information about how mHealth solutions can affect the uptake of PrEP, particularly in populations including both young sexual minority men and transgender and nonbinary people. PrEPresent, a mHealth app built on the successful foundation of the HealthMpowerment (HMP) Platform [24], prioritized PrEP uptake as a health behavior change and aimed to engage with SGM youth across diverse gender and racial and ethnic identities in the Greater Los Angeles area.

Project Aims

A pilot randomized control trial (RCT; ClinicalTrials.gov NCT05281393) was conducted to evaluate a tailored mHealth intervention, PrEPresent, which sought to increase PrEP uptake among SGM youth at a greater likelihood of HIV acquisition currently outside of the care continuum. PrEPresent connected users with a mobile app and a trained intervention peer support navigator, the PrEPresentative, and provided educational and motivational support to SGM youth in their decision to initiate PrEP. Information about the session content and other components of the intervention is outlined in the research

XSL•FO
RenderX

protocol [25]. PrEPresent incorporated patient activation theory, which focuses on one's ability to manage their own wellness and health care [26]. The aim of this project was to evaluate the feasibility and acceptability of the app and intervention with a racially and ethnically diverse sample of 150 SGM youth (aged 16‑26) in Los Angeles.

Methods

Study Design and Population

Formative research was first conducted to inform app features, content, and resources (Phase 1A), including 13 qualitative interviews with key informants providing PrEP care in the Los Angeles area. These interviews were transcribed and reviewed by a coding team of two infectious disease physicians. Overarching themes from the interviews were: location of services or descriptions of clinics, description of PrEP-specific provider training, basic provider duties around PrEP uptake, adherence, and persistence, source of clinic referrals, client motivators and facilitators, client barriers, clinical motivators, PrEP care specific to transgender populations, conversations about PrEP or sexual health, PrEP prescribing practices, client and provider stigma, COVID-19 impacts, and app suggestions or resources. These interview findings informed the formation of working groups and contributed to the addition of Los Angeles–specific resources and studies to the app (Phase 1B). The working groups included four weekly Zoom (Zoom Video Communications, Inc) sessions with 20 SGM youth with a goal of understanding PrEP awareness, acceptance, readiness, and hesitation. Each group consisted of 10 SGM youths who met study criteria; the working group sessions were facilitated by a member of our research team. The four sessions covered: health and sexuality, PrEP, mHealth apps, and a guided feedback session for the proposed mHealth for PrEP app. Working group findings were summarized into an information packet. This was shared with app developers to inform app tailoring to youth preferences, including the desire for a rewards system, varied avatar representations, and the general structure of the app's design. Youth Advisory Board (YAB) members also reviewed these findings to inform SGM youth–specific study and resource creation. The overall takeaway from the youth working groups was PrEP information sources, barriers and activators to accessing health care and PrEP, experiences in STI or HIV testing, prescription or adherence concerns, and familiarity with PrEP navigators. Phase 2 involved preliminary usability testing of the PrEPresent platform with 7 SGM youths, and Phase 3 involved 1:1 randomization of racially and ethnically diverse SGM youth into a pilot RCT.

In Phase 3, participants completed study visits at baseline, 3 months, and 6 months. The intervention was delivered between the baseline and 3-month visits. Most study activities took place digitally, but participants could elect to complete procedures in person. Participants completed computer-assisted self-interviewing via a web-based survey at baseline. At the 3- and 6-month time points, all participants completed computer-assisted self-interviewing surveys and a biomarker test of PrEP adherence, the UrSure point-of-care rapid urine test of tenofovir. UrSure is a validated urine assay with high sensitivity and specificity for tenofovir use within the previous 48 hours [27]. UrSure test kits were either mailed to participants or completed on-site at the research location. If all study components were completed (three surveys and two UrSure biomarker tests), participants received a bonus incentive.

A total of 147 participants were recruited into the study between August 2022 and July 2023. Participants were recruited using a variety of strategies, including recruitment from an existing cohort of young sexual minority men and transgender and nonbinary youth of color in Los Angeles [28], distribution of fliers at community-based organizations and tabling at community events, and through web-based advertising via social media and dating apps (eg, Instagram, Facebook, and Jack'd). Eligibility criteria for participants were as follows: (1) aged 16‑26 years; (2) cisgender men, transgender or gender nonconforming people, or identified differently from the gender picked for them at birth; (3) gay, bisexual, or some other same-sex identity, or reported having had sex with anyone with a penis during the previous 12 months; (4) White or Caucasian, Black or African American, Hispanic or Latino or Latinx, or multiracial with one of these identities; (5) living in the Los Angeles metro area; (6) had daily access to an iOS or Android smartphone or tablet with internet access; (7) reported having insertive or receptive sex in the previous 6 months or reported a positive STI result in the previous 6 months; (8) not currently on PrEP and had no plan to start or restart PrEP in the following 7 days; and (9) not currently enrolled in another HIV prevention study. Participants who participated in Phases 1 and 2 of the research were not eligible to participate in Phase 3. Individuals were not eligible to participate if they were living with HIV or were non-English speaking. Contact information for potential participants, such as name, phone number, email address, and date of birth, was stored in a recruitment database. This information was used to ensure participants could only enroll once in the project.

Randomization Mechanism

A Microsoft Excel–based simple random assignment approach was used for the PrEPresent Study. The RANDBETWEEN function with a range of 0 to 1 was used to generate a list of 150 continuous random numbers. Randomly generated numbers with a value greater than or equal to 0.5 were denoted as intervention allocations, while numbers with a value less than 0.5 were denoted as control allocations. This resulted in a 1:1 list of 75 assignment allocations per study arm. No stratification or blocking was used to generate the random assignments. The study data manager generated the assignment list independently of other study team members. All study personnel beyond the data manager were blinded to the order of assignments prior to individual participant randomization. Upon completion of consent and the baseline survey, study research assistants used the participant management system to prospectively assign participants to intervention or control, at which point they became aware of their assignment. From there, participant arm assignment was unblinded.

XSL•FO
RenderX

Study Procedures

Overview

Following verification of eligibility, participants completed in-person or virtual electronic consent processes via videoconferencing, then performed a baseline survey. After survey completion, participants were randomized into the intervention (PrEPresent app) or control (information-only app) arm and were prompted to download the study app. Baseline visits concluded with a tour of the app and its features. Participants were encouraged to use the app daily, and SMS text message push reminders were sent to those who had not logged into the app for 7, 14, 30, or 60 days.

Intervention Arm

Intervention participants were encouraged to schedule their first session with the PrEPresentative, the interventionist peer support navigator. Participants received SMS text message reminders for their confirmed appointments 24 hours prior and 1 hour prior. Participants were encouraged to meet with the interventionist monthly, with weekly check-ins occurring via text or phone call between sessions. The PrEPresentative was trained on an intervention curriculum. The curriculum included guided session topics, a reference library, motivational messaging examples, common solutions to known PrEP barriers, and additional information about PrEP. Sessions were conducted via an in-app chat feature, phone call, or Health Insurance Portability and Accountability Act–compliant video call. Each intervention session had to be completed before moving on to the next. Following intervention sessions, the PrEPresentative completed a case report form to track intervention statistics. Session content and referrals were recorded, and multiple session topics could be chosen. Session topics were broken into three overarching categories: individual, sociocultural, and structural. After each session, participants were sent a brief satisfaction survey. Intervention participants had full access to the PrEPresent app. App features included a health tracker for habit monitoring, an anonymous user forum, activities and goal-setting functions, avatar customization, achievement badges, a care locator, an ask the expert section, and an interventionist appointment scheduler.

Control Arm

Control participants were provided with an information-only version of the app. This app contained a media library with resources regarding PrEP and did not allow access to other features. Control participants could not interact with the PrEPresentative.

Outcomes

During this initial work to evaluate the PrEPresent intervention, the primary outcomes were acceptability and feasibility. Outcomes were prespecified in the protocol and trial registration as being assessed at the 3-month follow-up. This time point was selected based on the anticipated window of initial behavior change following exposure to the digital intervention. We defined acceptability as participants' use of the app at higher rates than the control arm. Feasibility, in turn, was defined as at least 50% of intervention participants completing two peer navigation sessions with the PrEPresentative. A brief satisfaction assessment was administered at the end of each navigation session to identify challenges experienced by the participant, the most and least helpful aspects of the session, challenges using the system, satisfaction with information provided, and any additional needs the navigator did not address. The intervention sessions were defined as feasible if an overall postsession score of 4 or higher (out of 7) was reported on these surveys. Metrics on feedback about components of the app, such as trustworthiness of information, ease of use, and future app use, were gathered to understand app acceptance. The secondary outcome measure was defined as higher levels of PrEP uptake within intervention participants compared to control participants. PrEP uptake was defined as an adherent UrSure test result or self-report PrEP use on survey measures.

Statistical Analysis

All statistical analysis was performed using R (version 4.2.3, R Core Team). Descriptive statistics were calculated and stratified by treatment arm. Continuous variables were summarized using medians and IQR. Categorical variables were reported using frequencies and percentages. The Fisher exact test and the Wilcoxon rank sum test were used to compare variables of interest between arms. Cronbach α was calculated to assess the internal consistency of nonvalidated instruments. The total scores of these instruments were calculated by summing the individual question scores. Due to the descriptive nature of our pilot trial, statistical analysis did not include power calculations or tests of statistical significance. Instead, 95% CIs are provided when applicable.

Ethical Considerations

All phases of PrEPresent were approved by the Children's Hospital Los Angeles Institutional Review Board (CHLA-20‐00596) under expedited review. The study involved three phases of research, as mentioned above. For Phase 1A, a waiver of signed informed consent was approved. Participants received an information sheet, and verbal consent was obtained. For Phases 1B, 2, and 3, participants signed informed consent and Health Insurance Portability and Accountability Act authorization documents. A waiver of permission for parental consent was obtained for minor participants due to the sensitive nature of the study eligibility criteria, in line with California-based protections for minors accessing sexual health services. To maintain patient privacy and confidentiality, only minimal identifying information necessary to maintain contact with participants during study procedures was collected and stored in a password-protected database. All study data were deidentified and monitored under a pre-established data safety and monitoring plan. Participants were prompted to use an anonymous username to enroll in the app and were discouraged from posting identifying information. Participants were compensated via Cash App, PayPal, Venmo, or Visa e-gift cards. In Phase 1A, PrEP care providers and PrEP navigators participating in qualitative interviews were compensated US $35 for their participation in a 60-minute interview. In Phase 1B, SGM youth participants received US $50 per working group session for 4 sessions, with a chance to receive an additional US $50 bonus for up to a total of US $250. In Phase 2, SGM youth participants pilot-testing the app received US $40 for the

XSL•FO
RenderX

pretest survey measure and focus group attendance and US $40 for a posttest survey measure and one-on-one interview participation for a total of US $80. In Phase 3, pilot RCT participants received US $50 for completing each survey and US $20 for biomarker test completion. If participants successfully completed all study activities, they received an additional US $50 bonus. Intervention arm participants earned US $10 for completing each postsession survey. Participants in the intervention arm could earn up to US $280, and participants in the control arm could earn up to US $240.

Results

Demographics and Comparisons

Of those determined eligible via digital screener, 147 participants were enrolled in the study, and 274 participants were excluded. Most (n=232) were excluded due to not confirming eligibility and scheduling an initial consent visit. A total of 42 participants confirmed their eligibility but did not attend a consent visit due to other reasons. Of the 147 participants enrolled in the study, 72 participants were randomly assigned to the intervention arm and 75 to the control arm (Figure 1). The mean age for the sample was 22.57 (SD 2.48) years. Gender was measured using a two-part question assessing sex assigned at birth and gender identity, with gender being measured with a “select all that apply” approach that also allowed for a write-in. If participants selected more than one gender, they were prompted to select which gender they most identified with. These responses were collapsed, as presented in Table 1. Gender varied greatly among participants, with 33% (47/147) of the participants identifying as cisgender men, followed by nonbinary individuals (42/147, 29%), and trans men (19/147, 13%). Gay (41/147, 28%) and queer (39/147, 27%) were the most common sexual orientations, making up more than 50% (80/147) of the participants. Almost 62% (91/147) of participants reported a sexual attraction to more than one gender. Hispanic or Latinx was the most represented group (71/147, 48%), predominantly in the control arm (42/75, 56%) compared to the intervention arm (29/72, 40%). Black or African American-identified participants made up 21% (15/72) and 16% (12/75) of the intervention and control arms, respectively. A detailed breakdown of the demographic characteristics is summarized in Table 1.

XSL•FO
RenderX

Figure 1. PrEPresent CONSORT flow diagram. CONSORT: Consolidated Standards of Reporting Trials.

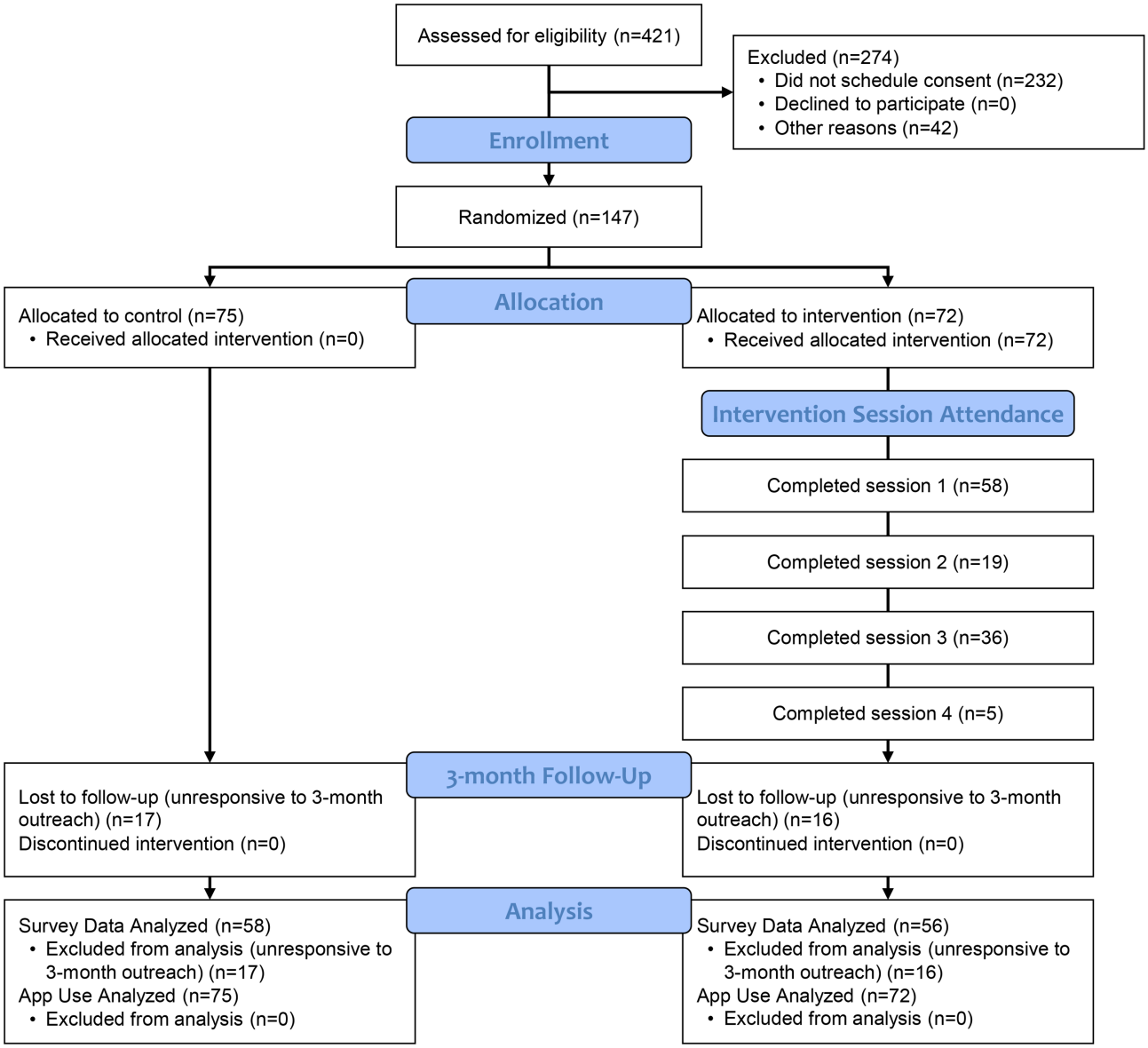

XSL•FO
RenderX

Table . Baseline demographic characteristics of sexual and gender minority youth in the PrEPresent study[a,b].

Characteristic	Total (N=147)	Control (n=75)	Intervention (n=72)
Age (in years), mean (SD)	22.57 (2.48)	22.55 (2.52)	22.60 (2.45)
Sex at birth[a], n (%)			
Male	78 (53)	40 (53)	38 (53)
Female	66 (45)	32 (43)	34 (47)
Intersex	1 (1)	1 (1)	0 (0)
Gender, n (%)			
Cisgender man	49 (33)	26 (35)	23 (32)
Gender nonconforming	6 (4)	0 (0)	6 (8)
Genderfluid	8 (5)	4 (5)	4 (6)
Genderqueer	6 (4)	5 (7)	1 (1)
Nonbinary	42 (29)	20 (27)	22 (31)
Trans woman	10 (7)	5 (7)	5 (7)
Trans man	19 (13)	10 (13)	9 (13)
Two-spirit	1 (1)	1 (1)	0 (0)
Other	6 (4)	4 (5)	2 (3)
Sexuality, n (%)[b]			
Bisexual	28 (19)	10 (13)	18 (25)
Demisexual	1 (1)	0 (0)	1 (1)
Gay	41 (28)	21 (28)	20 (28)
Heterosexual (straight)	1 (1)	0 (0)	1 (1)
Homosexual	3 (2)	2 (3)	1 (1)
Lesbian	3 (2)	3 (4)	0 (0)
Men who have sex with other men	2 (1)	1 (1)	1 (1)
Pansexual	24 (16)	13 (17)	11 (15)
Queer	39 (27)	22 (29)	17 (24)
Same gender-loving	1 (1)	1 (1)	0 (0)
Other	3 (2)	2 (3)	1 (1)
Race or ethnicity			
Black or African American	27 (18)	12 (16)	15 (21)
Hispanic or Latinx	71 (48)	42 (56)	29 (40)
White	49 (33)	21 (28)	28 (39)

[a]2 “Decline to answer” in control.

[b]1 “Don’t know” in intervention.

Patient Activation

The Patient Activation Measure was used, with a Cronbach α of 0.93, to measure participants’ activation at each survey timepoint [29]. Both study arms scored within the high-level activation category (category 4, defined as 67.1 - 100 points) at baseline. Patient activation scores remained stable across both arms over the 3 months (Table 2).

XSL•FO
RenderX

Table . PrEP[a] eligibility, readiness, knowledge, and uptake among participants from baseline to 3 months.

Variables	Baseline control (n=75)	Baseline intervention (n=72)	3-month control (n=58)	3-month intervention (n=56)
PrEP readiness ruler				
On a scale from 1 to 10, how important is starting PrEP to you right now?, n (%)				
1 - 5	51 (68)	50 (69)	__[b]	—
6 - 10	24 (32)	22 (31)	—	—
On a scale from 1 to 10, how motivated do you feel to start PrEP right now?, n (%)				
1 - 5	48 (64)	50 (69)	—	—
6 - 10	27 (36)	22 (31)	—	—
PrEP Knowledge				
Before today, what types of PrEP had you heard about? (check all that apply), n (%)				
Oral PrEP	65 (87)	62 (86)	—	—
On-demand PrEP (2-1-1)	24 (32)	28 (39)	—	—
Injectable	21 (28)	15 (21)	—	—
Other	1 (1)	0 (0)	—	—
I have never heard of PrEP	7 (9)	7 (10)	—	—
Where did you hear about or get information on how to use PrEP in the last 6 months?[c], n (%)				
Pharmacist	3 (4)	2 (3)	—	—
Doctor	24 (35)	24 (37)	—	—
Friends	26 (38)	38 (59)	—	—
HIV counselor	12 (18)	20 (31)	—	—
Social media	26 (38)	20 (31)	—	—
Other	15 (22)	11 (17)	—	—
Patient Activation Score				
Mean (SD)	78.02 (16.61)	77.74 (20.55)	77.06 (18.79)	80.13 (16.54)
PrEP uptake, n (%)				
Prebaseline PrEP uptake (lifetime)	14 (19)	16 (22)	—	—
PrEP uptake (3 mo)	—	—	11 (19)	14 (25)
PrEP type (select all that apply)[d]				
Daily oral pill	14 (100)	15 (88)	7 (64)	14 (100)
Injectable PrEP	0 (0)	1 (6)	1 (9)	0 (0)
On-demand PrEP (2-1-1)	0 (0)	1 (6)	0 (0)	0 (0)
Undetermined[e]	—	—	3 (27)	—
PrEP UrSure Adherence Test results[f]				
Adherent	—	—	5 (14)	4 (12)
Nonadherent	—	—	28 (80)	30 (88)
Intermediate	—	—	2 (6)	0 (0)
UrSure adherence among PrEP users[g]				
Adherent	—	—	5 (56)	4 (40)
Nonadherent	—	—	4 (44)	6 (60)

[a]PrEP: pre-exposure prophylaxis.
[b]Data not collected.

XSL•FO
RenderX

[c]n=68 for control, n=65 for intervention due to skip logic.

[d]n=14 for baseline control; n=17 for baseline intervention; n=11 for 3-month control; and n=14 3-month intervention.

[e]Undetermined participants tested adherent on UrSure Biomarker but did not report PrEP use at 3 months

[f]n=35 for 3-month control; and n=34 for 3-month intervention.

[g]n=9 for 3-month control; and n=10 for 3-month intervention.

PrEP Readiness, Knowledge, Uptake, and Adherence

Only 14 participants reported not having heard of PrEP before participating in this study (7 in the intervention and 7 in the control arm). Although most participants had heard about oral PrEP (62/72, 86% and 65/75, 87% in the intervention and control arms, respectively), there was less overall knowledge of on-demand or injectable PrEP indications. Both arms had similar rates of lifetime prior PrEP use, with 22% (16/72) of intervention and 19% (14/75) of control arm participants reporting prior use. Participants were most likely to have heard about PrEP from their friends (64/147, 44%), a doctor (48/147, 33%), social media (46/147, 31%), or an HIV counselor (32/147, 22%). In an assessment of PrEP readiness, almost 70% (101/147) of both study arms reported the importance of starting PrEP between 1 and 5 on a 10-point scale. A total of 69% (50/72) of intervention participants and 68% (51/75) of control participants rated their motivation to start PrEP as between 1 and 5 on a 10-point scale. PrEP uptake was initiated by 14 (25%) intervention participants at 3 months and 11 (19%) participants in the control arm. Nearly all participants reported using daily oral PrEP. However, one control participant reported using injectable PrEP. Information on PrEP type is missing for three control participants, as they tested adherent on UrSure but did not report PrEP use in their 3-month assessment.

PrEP biomarker testing had a low rate of return across both arms. At the 3-month assessment, only 34 of the 56 (61%) intervention participants and 35 of the 58 (60%) control participants had completed their testing. Reported adherence levels according to tests were low across the sample, with 88% (30/34) of intervention participants and 80% (28/35) of control participants demonstrating nonadherence based on test results at 3 months. Of the 14 intervention participants who reported PrEP uptake, 10 completed biomarker testing and 4 (40%) were found to be adherent, compared to 11 control participants, of whom 9 completed biomarker testing and 5 (56%) were found to be adherent.

Intervention

Session Content and Information

Intervention acceptability and feasibility are summarized in Table 3, including session delivery method, topic coverage, participant referrals, session satisfaction, and app satisfaction. In total, 58 (81%) participants attended the first session, 36 (50%) participants attended 2 sessions, 19 (26%) attended three sessions, and 5 (7%) attended four sessions. Phone call and in-app message-based intervention sessions were similarly used (50/118, 42% and 49/118, 42%, respectively), but most sessions took place outside of the app (video or phone call; 69/118, 59%). Motivation was discussed in 88% (104/118) of sessions, with PrEP knowledge as the second most discussed topic. HIV knowledge was discussed in 20% (23/118) of sessions, with sexual health being discussed in 35% (41/118). Referrals were provided to participants in 34% (40/118) of sessions. The most common referrals were to PrEP clinics, with the second to PrEP educational materials. Nine referrals to PrEP navigators were made.

XSL•FO
RenderX

Table . Session attendance, delivery, acceptability, and feasibility of the PrEPresent intervention (n=72).

Variables	Total sessions (n=118)	Session 1 (n=58)	Session 2 (n=36)	Session 3 (n=19)	Session 4 (n=5)
Attendance, n (%)	118 (100)	58 (81)	36 (50)	19 (26)	5 (7)
Session delivery, n (%)					
Video call	19 (16)	10 (17)	4 (11)	3 (16)	2 (40)
In-app messaging	49 (42)	20 (35)	17 (47)	9 (47)	3 (60)
Phone call	50 (42)	28 (48)	15 (42)	7 (37)	0 (0)
Session topics[a], n (%)					
Individual					
Motivation	104 (88)	53 (91)	32 (89)	15 (79)	4 (80)
Adherence	12 (10)	2 (3)	5 (14)	3 (16)	2 (40)
Sexual health	41 (35)	28 (48)	7 (19)	5 (26)	1 (20)
Self-efficacy	24 (20)	13 (22)	6 (17)	4 (21)	1 (20)
Patient activation	49 (42)	26 (45)	13 (36)	9 (47)	1 (20)
PrEP[b] Knowledge	66 (56)	56 (97)	8 (22)	1 (5)	1 (20)
HIV Knowledge	23 (20)	17 (29)	4 (11)	1 (5)	1 (20)
Other	55 (47)	26 (45)	16 (44)	10 (53)	3 (60)
Sociocultural					
Relationships	15 (13)	10 (17)	4 (11)	1 (5)	0 (0)
Other	4 (3)	2 (3)	1 (3)	1 (5)	0 (0)
Structural					
Stigma	2 (2)	2 (3)	0 (0)	0 (0)	0 (0)
Insurance	19 (16)	14 (24)	4 (11)	1 (5)	0 (0)
Work or employment	3 (3)	0 (0)	1 (3)	2 (11)	0 (0)
Other	5 (4)	2 (3)	1 (3)	2 (11)	0 (0)
Referrals[a,c]	40 (34)	24 (41)	13 (36)	3 (16)	0 (0)
PrEP navigator	9 (8)	6 (10)	3 (8)	0 (0)	0 (0)
PrEP clinic	21 (18)	15 (26)	8 (22)	1 (5)	0 (0)
PrEP educational materials	13 (11)	8 (14)	4 (11)	1 (5)	0 (0)
Sexual health resource	1 (1)	1 (2)	0 (0)	0 (0)	0 (0)
STI[d] or HIV testing	4 (3)	2 (3)	1 (3)	1 (5)	0 (0)
Community clinic	1 (1)	0 (0)	1 (3)	0 (0)	0 (0)
Postintervention session satisfaction					
Mean (SD)	6.23 (1.09)	6.23 (1.01)	6.25 (1.15)	6.15 (1.20)	6.35 (1.04)
PrEPresent Care Navigator Satisfaction (n=55), n (%)					
Very satisfied	31 (56)	—[e]	—	—	—
Mostly satisfied	16 (29)	—	—	—	—
Indifferent or mildly dissatisfied	5 (9)	—	—	—	—
I did not talk to the care navigator	3 (6)	—	—	—	—

[a]Substance use, transportation, food insecurity, and housing were either not discussed or discussed across only one session.

[b]PrEP: pre-exposure prophylaxis.

XSL•FO
RenderX

[c]No referrals were made to the following: PrEP Income Support, PEP, Gender Affirming Resources, Transportation Services, Housing Services, Substance Use Services, Domestic Violence Support, Mental Health Resources, Legal Aid, COVID Services, Social/Community Support, Food Pantry, Insurance Navigation, or Immigration Services.

[d]STI: sexually transmitted infection.

[e]Data not collected.

Intervention Satisfaction

The postsession satisfaction mean across sessions was 6.23 (SD 1.09) on a 7-point scale with a Cronbach α of 0.85. Satisfaction for the interventionist was high, with 56% (31/55) of participants indicating they were very satisfied and 29% (16/55) indicating they were mostly satisfied. While a majority remained engaged and satisfied, about 9% (5/55) of participants were indifferent or mildly dissatisfied.

Platform Use and Satisfaction

Overall, intervention participants had higher mean total logins and mean total minutes in the PrEPresent app than control participants. The mean total logins on the platform was 6.31 (SD 9.27) in the control arm and 15.93 (SD 15.85) in the intervention arm (95% CI 5.01-13.58; Table 4). A similar trend was observed in the mean total minutes spent on the platform, with control participants spending 3.03 (SD 5.70) minutes in the app on average compared to 9.51 (SD 11.47) minutes in the app in the intervention arm (95% CI 3.51-9.51). Platform satisfaction was also rated more highly among intervention participants, with 41% (23/56) stating they were very satisfied compared to 26% (15/58) of control participants. More than two-thirds (34/56) of intervention participants stated PrEPresent helped them deal with the challenges they have faced getting on PrEP compared to 43% (25/58) of control participants (Table 4). Satisfaction around the app platform was universally high across both RCT arms, though overall use of the platform was low. Around 70% (81/114) of participants agreed or strongly agreed that the platform was easy to use. At least 70% (82/114) of participants in both arms stated PrEPresent helped them make healthier choices at a 4 or above on a 0 - 7-point sliding scale. There were high ratings on levels of trust for the information presented in the app, with no intervention participants disagreeing or strongly disagreeing that they trust the information in PrEPresent. Most participants would use the app in the future.

XSL•FO
RenderX

Table . The 3-month mHealth[a] platform engagement and usability measurements.

Variables	Total control (n=75)	Total intervention (n=72)	3-month control (n=58)	3-month intervention (n=56)
Platform use				
Total logins				
Mean (SD)	6.31 (9.27)	15.93 (15.85)	—[b]	—
Median (IQR)	2.00 (1.00-7.50)	11.00 (4.00-22.25)	—	—
Total minutes				
Mean (SD)	3.03 (5.70)	9.51 (11.47)	—	—
Median (IQR)	0.86 (0.04-2.73)	5.30 (2.01-13.06)	—	—
Overall, how satisfied are you with PrEPresent?, n (%)				
Very satisfied	—	—	15 (26)	23 (41)
Somewhat satisfied	—	—	30 (52)	22 (39)
Somewhat unsatisfied	—	—	12 (21)	10 (18)
Not satisfied	—	—	1 (2)	1 (2)
Has PrEPresent helped you deal with the challenges you've faced getting on PrEP[c]?, n (%)				
Decline to answer	—	—	20 (35)	13 (23)
No, it really didn't help	—	—	13 (22)	9 (16)
Yes, it helped	—	—	17 (29)	28 (50)
Yes, it helped a great deal	—	—	8 (14)	6 (11)
PrEPresent was easy to use, n (%)				
Strongly disagree	—	—	0 (0)	2 (4)
Disagree	—	—	4 (7)	5 (9)
Undecided	—	—	13 (22)	9 (16)
Agree	—	—	24 (41)	23 (41)
Strongly agree	—	—	17 (29)	17 (30)
PrEPresent was hard to use, n (%)				
Strongly disagree	—	—	15 (26)	12 (21)
Disagree	—	—	36 (62)	35 (63)
Agree	—	—	5 (9)	7 (13)
Strongly agree	—	—	2 (3)	2 (4)
I trust the information in PrEPresent				
Strongly disagree	—	—	1 (2)	0 (0)
Disagree	—	—	1 (2)	0 (0)
Agree	—	—	26 (45)	26 (46)
Strongly agree	—	—	30 (52)	30 (54)
I would use PrEPresent in the future if it were available, n (%)				
Strongly disagree	—	—	3 (5)	6 (11)
Disagree	—	—	14 (24)	7 (13)
Agree	—	—	34 (59)	32 (57)
Strongly agree	—	—	7 (12)	11 (20)
PrEPresent helps me make healthier choices in my life[d], n (%)				
0 - 3	—	—	15 (26)	17 (30)
4 - 7	—	—	43 (75)	39 (70)

XSL•FO
RenderX

[a]mHealth: mobile health.
[b]Data not available.
[c]PrEP: pre-exposure prophylaxis.
[d]0-7-point sliding scale.

Discussion

Principal Findings

Overall, within a sample of racially and ethnically diverse SGM youth, PrEPresent met predetermined objectives for feasibility and acceptability in the pilot RCT portion of the research. The project did not see high rates of PrEP uptake when compared to other PrEP uptake interventions [29-34] or when compared against documented PrEP use within similar cohorts outside of intervention settings [11,35]. Additionally, use of the mHealth platform was overall low across both study arms compared to other iterations of the HMP platform [24,36-38]. The outcomes of this project are valuable in understanding how future projects should engage SGM youth in HIV prevention efforts, particularly using health technology.

PrEP Uptake and Adherence

Although PrEP uptake within the intervention arm was 25% (14/56) and 19% (11/58) in the control arm, these are lower uptake percentages compared to other interventions assessing PrEP initiation [30-32]. A few studies have noted intervention-level PrEP uptake levels similar to PrEPresent but with a greater difference between control and intervention arms [29,34]. The outcomes from PrEPresent add to previous literature demonstrating the mixed success that PrEP uptake and adherence projects have had in activating SGM youth in the PrEP care cascade. In the TechStep project, conducted across 5 major US cities, there was no difference between intervention arms (text plus, text only, web app plus, and web app only) and control arms in PrEP uptake within 254 transgender and gender-expansive youth [39]. In a mobile messaging mHealth app study for 1226 SMM, high-risk HIV-negative groups had a higher odds of HIV testing and PrEP use, but these behavioral changes were not noted for the participants with low-risk HIV-negative groups [33]. In the theory-driven MyChoices app for cisgender young sexual minority men, HIV testing was higher within the intervention arm, but no difference in PrEP uptake was noted [40]. High acceptability across these projects indicates that while SGM youth find mHealth interventions usable, more work needs to be done to understand how to translate app engagement to behavior change. Future mHealth projects targeting PrEP uptake powered for efficacy are required to understand how mHealth can help SGM youth overcome barriers they face to starting PrEP.

Although adherence was not an outcome for this project, we used the UrSure test as an indicator of PrEP uptake in conjunction with self-reported survey data. Biological markers of PrEP use continue to be important in conjunction with self-reported PrEP use, as evidenced in both clinical and research settings [41,42]. In a study assessing the feasibility of urine testing for PrEP adherence in a sample of young people, it was found to be theoretically more acceptable than phlebotomy draw, finger prick, or hair test [43]. Despite this, completion of the at-home UrSure urine test for PrEP adherence in this study was low across both arms. Of those who completed the 3-month surveys, only around 60% (69/114) also completed the UrSure testing, accounting for 47% (69/147) of the total initial sample from baseline. The study team noted several challenges in biological result collection. Due to privacy concerns and short follow-up windows for specimen delivery, study team members required participant verification of address before sending test kits. Future projects may consider automating check-in processes for verifying participant addresses before their testing window opens or obtaining consent to send materials without verification. Many participants received test kits but did not complete them. Increasing the specimen completion incentive instead of prioritizing a bonus incentive may have increased test kit return.

It is possible that, due to low PrEP uptake in our cohort, participants were not as interested in understanding their PrEP adherence levels. In a 2023 qualitative feedback study on self-collected STI testing from a similar participant base, low levels of return were noted for those who had less concerns about a positive HIV and STI diagnosis [44]. Additional language about the importance of test kit completion to the study aims, regardless of PrEP uptake, may have helped increase biomarker return. Postintervention qualitative feedback with PrEPresent participants around PrEP adherence testing may provide insights into motivators and barriers to test completion. Biological markers of adherence delivered key insights into PrEP uptake across the cohort, and future studies may consider pilot testing adherence testing protocols with members of YABs to understand acceptability and to troubleshoot sample collection.

Rates of adherence to the UrSure tests were also low among our participants. The UrSure point-of-care test measures adherence in the last 48 hours. While positive adherence results on these biological markers can deliver important PrEP uptake information, it is difficult to interpret levels of protection against HIV based on these results. It has been established that 4 (or more) doses of oral PrEP per week over time create adequate protection in rectal tissues [45], whereas 6 - 7 doses of oral PrEP per week are needed to create adequate protection in vaginal tissues [46]. Participants could have protective levels of PrEP in their system but test nonadherent on the UrSure point of care test, depending on when pills were taken. As noted in a review of PrEP adherence testing methods, understanding sexual activity type and sex assigned at birth is important in contextualizing urine adherence results, particularly those that measure dosage in the last 48 hours [47]. Additionally, though rates of injectable PrEP were low in our cohort, injectable Cabotegravir is not detected by the UrSure test. Future research and clinical settings using adherence measures must note that evolving PrEP landscapes and the use of on-demand PrEP may impact results regarding adherence across different forms of testing.

XSL•FO
RenderX

Intervention Feasibility and Acceptability

A total of 50% (36/72)of participants attended a second intervention session, meeting the metric for feasibility. Completion of the fourth closing session was low, at only 5 (7%) participants. Four intervention sessions, spaced every 4 weeks, were offered to participants within the 3-month active intervention period. Scheduling this number of sessions may not have seemed practical for our participant population. Referrals dropped in the third and fourth sessions, possibly indicating that a reduced number of sessions may still meet participant needs [32,34,48]. An initial session delivering PrEP information to SGM youth may be supportive in their decision to pursue a PrEP prescription, as motivation and PrEP knowledge were the most discussed intervention session topics during the first session. Postsession satisfaction across sessions remained high and scored well above our target average of 4 for feasibility. Overall satisfaction with the PrEPresentative was also high, with the majority of participants rating their satisfaction as "very satisfied." These results demonstrate that in a population of SGM youth, a peer-support navigation intervention for PrEP uptake is acceptable and feasible.

Delivery of the intervention sessions varied across participants and sessions. YAB feedback dictated offering a variety of session modalities because of the overall increase in video meetings throughout the COVID-19 pandemic. Future interventions should use community feedback to understand desires for intervention delivery. One study citing clinical strategies to adjust to pandemic-constrained PrEP access noted that newly offered telehealth access for services remained in use even after COVID-19 cases had stabilized [49]. An overview of PrEP access via telehealth published in 2022 found both publicly funded clinical settings and private companies using telehealth for PrEP delivery and highlighted the use of a California state-funded tele-PrEP delivery program [50]. Given that the study was deployed in the post-COVID-19 pandemic period in a state offering digital PrEP access, the novelty of an app-delivered peer support navigator for PrEP uptake may have been less appealing during this time.

One reason the acceptability and feasibility of the intervention may have been high despite low platform use is that PrEPresent prioritized community-based collaboration throughout all phases of the project. App feedback was overwhelmingly positive, with most participants trusting the information in the app and agreeing that it helped them make better choices. Youth working groups and YABs comprised of racially and ethnically diverse SGM youth engaged with the project to understand the similarities and differences between young sexual minority men and transgender and nonbinary populations and how a mHealth intervention could best serve them both. Members of the research team, from research assistants to the interventionist, represent different components of the community represented within this sample. Uplifting community voices should continue to be prioritized in prevention strategies to meet community members' needs.

Platform Use and Satisfaction

Use of the PrEPresent intervention was overall low compared to other iterations of mHealth projects built on the HMP Platform [23,24,36]. Primary login to the app was verified for all participants. Intervention participants did have similar login rates to the app as other mHealth platforms [24,38], which may be due to using scheduling features to meet with the interventionist. Most intervention participants chose to engage with the PrEPresentative outside of the app (through phone call or video call), but this alone does not fully explain the low time spent within the app. The project team worked to adapt the platform to drive digital interaction, but certain features shown to increase engagement in other mHealth interventions were not integrated. Features such as use-based points or digital currency [36], app-based health monitoring and tailored messaging [22,51], serious gaming or enhanced gamification [52,53], peer mentor engagement [54], and access to HIV prevention materials such as home tests or condoms and lube [23,51] have been demonstrated to engage similar participant populations in mHealth studies. Reviews of published mHealth findings have found links between higher app use and behavior change [55]. mHealth interventions hoping to combine app-based features with external features such as two-way SMS text messaging or peer health navigators may consider additional in-app incentives to encourage user engagement and should prioritize the collection of app paradata to explore participant use.

Limitations

Several limitations of this study should be noted. Randomization for the pilot RCT was conducted on a 1:1 basis with no stratification, and as a result, some imbalances may have occurred. While this is an important consideration, it is not unique to this study and reflects a common limitation in randomized trials with modest sample sizes. Additionally, PrEPresent was not powered to look at comparisons between subgroups.

This study population was drawn from the Los Angeles metropolitan area, and it reflects the demographic profile in Los Angeles. While we did recruit a racially diverse sample, we cannot compare outcomes within different racial or ethnic groups or generalize results to these groups. Because recruitment for the project took place in a large-scale US city, results may not be generalizable to other geographic areas where factors impacting PrEP uptake may differ, such as access to insurance coverage or clinical spaces specializing in PrEP services. In addition, a subset of participants was recruited from a previous observational cohort study focused on HIV prevention with SGM youth of color. While participants had not received any digital health intervention or app-based support related to PrEP during their involvement, our sample size constrained our ability to conduct subgroup analysis to understand potential impacts on intervention engagement. Future studies involving existing cohorts may include prior participation as a stratification variable or covariate where appropriate.

Additionally, while we intended to include the validated System Usability Scale (SUS) in our survey instrument, we discovered during data analysis that supplementary usability questions were mistakenly identified as the SUS in the survey documentation. As a result, we were unable to calculate SUS scores using the validated measure. Given this error, we used an alternative

XSL•FO
RenderX

breakdown of usability constructs as presented in Table 4, which aligns with the domains outlined in our grant proposal and captures key facets of user experience (eg, engagement, ease of use, and satisfaction). Although not based on the SUS, this adapted usability assessment still provides valuable insights into intervention performance and supports the interpretability of our findings. Future mHealth work with this population should further psychometric evaluation of existing validated measures while also prioritizing the development of tools that reflect the lived experiences of diverse young populations at risk for HIV.

Finally, attrition by 6 months was higher than anticipated, leading to reduced power and potential bias in outcome estimates measured at that time point. As such, we felt it most appropriate to focus on the more complete and rigorously collected 3-month data for this initial report, as including 6-month outcomes would not have strengthened the interpretation of the primary study aims. We are currently exploring the 6-month data in secondary analyses and plan to report those findings—particularly related to longer-term patterns of engagement and behavioral trajectories—in a subsequent paper.

Future Directions

As demonstrated in the outcomes of this project, one limitation of many mHealth interventions is engagement in the app or web-based tool [55]. In a similar mHealth study, while community events, group sessions, access to services, and use of their mHealth app were all used, app use averaged about 3.35 minutes per user, and PrEP outcomes were not significant [56]. Understanding the best features to drive long-term use and interest in app platforms and to track and analyze the use of those features will move app-based PrEP interventions toward higher rates of use, which may lead to better prevention outcomes. Engagement linked to features available through mHealth platforms will also need to consider scalability (the ability to offer the app and features to larger amounts of people) in terms of translational efforts in being integrated into clinical care [57,58]. Future studies will need to prioritize the components of their interventions that work best within their participant population, such as interventionist visits in this study, to understand how to guide engaging mHealth projects into translatable tools for HIV prevention.

Throughout the course of this project's development, changes in the types of PrEP available from foundational research to pilot RCT execution occurred. Future mHealth apps supporting PrEP uptake and adherence will need to be developed within the evolving context of PrEP modalities used among SGM communities [59,60]. Across the United States, injectable PrEP use alone increased from 1.1% to 2.5% of all PrEP prescriptions from 2022 to 2023 [61]. Recently, a longer-acting injectable PrEP option delivered in-clinic twice-yearly, Lenacapavir, was approved for all adolescents and adults [62]. Injectable PrEP may have a lower barrier to uptake and adherence, particularly for those accessing the medication through Medicare coverage, as a new National Coverage Determination expanded coverage to both oral and injectable PrEP medications with no cost-sharing [63]. With the introduction of injectable PrEP and the popularity of on-demand PrEP, daily PrEP adherence may not be as relevant of a metric for HIV prevention moving forward. Instead, initiation, persistence, and clinical appointment attendance may need to be addressed through innovative app features.

Conclusions

Results from the PrEPresent study indicate that SGM youth at a greater likelihood of HIV acquisition are willing to engage with a peer-support navigator through a mHealth platform to understand more about PrEP and PrEP uptake. Access to this type of mHealth app in combination with peer navigation may increase PrEP uptake in individuals outside the PrEP care continuum. Among those in the intervention arm, the app and PrEPresentative were satisfactory, although overall engagement was low. Additional mHealth studies around prevention should continue to include and elevate the voices of their participant populations through community-engagement processes to ensure they meet the needs of SGM people.

Acknowledgments

Generative artificial intelligence was not used in any portion of the writing of this manuscript. The authors would like to acknowledge the contributions of the staff members who contributed to collection, management, analysis, and review of this data: Alicia Bolton, Michael Smith, Lucas Lara Rojas, Ali Johnson, Gerald Branden Mabe, and Laura Rachel. The authors would also like to acknowledge the insightful and practical commentary of the members of the Community Advisory Board—Carlos D. Mena and Donta Morrison: AIDS Project Los Angeles; Dr. Harold Glenn San Agustin: JWCH Wesley Health Centers; Harold C. Sarmiento: Kaiser Los Angeles Medical Center; Lorenzo Banda: Watts Healthcare Foundation; Maria Roman: Trans Latina Coalition; Percival Pandy and Steven Campa: The LGBT Center South; Miguel Martinez: Division of Adolescent and Young Adult Medicine, Children's Hospital Los Angeles; and Otis Harris: The LGBT Center Long Beach. Finally, the authors would like to acknowledge the meaningful feedback of the PrEPresent Youth Community Advisory Board on their app and resource guide feedback. The work presented in this study was supported by the National Institute of Drug Abuse of the National Institutes of Health (award R34 DA050531). The funder had no involvement in the study design, data collection, analysis, interpretation, or writing of the manuscript. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This work was supported by the Biostatistics and Data Management Core at The Saban Research Institute, Children's Hospital Los Angeles, and by grants UL1TR001855 and UL1TR000130 from the National Center for Advancing Translational Science of the US National Institutes of Health.

XSL•FO
RenderX

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

Checklist 1
CONSORT e-Health Checklist (V1.6.1)
[PDF File, 552 KB - formative_v9i1e69255_app1.pdf]

References

1. Diagnoses, deaths, and prevalence of HIV in the united states and 6 territories and freely associated states, 2022. Centers for Disease Control and Prevention. 2022. URL: http://www.cdc.gov/hiv-data/nhss/hiv-diagnoses-deaths-prevalence.html [accessed 2024-09-09]
2. Stutterheim SE, van Dijk M, Wang H, Jonas KJ. The worldwide burden of HIV in transgender individuals: an updated systematic review and meta-analysis. PLoS ONE 2021;16(12):e0260063. [doi: 10.1371/journal.pone.0260063] [Medline: 34851961]
3. HIV infection, risk, prevention, and testing behaviors among transgender women—national HIV behavioral surveillance, 7 u.s. Centers for Disease Control and Prevention. 2019. URL: https://web.archive.org/web/20221001232700/https://www.cdc.gov/hiv/pdf/library/reports/surveillance/cdc-hiv-surveillance-special-report-number-27.pdf [accessed 2024-11-26]
4. Eisenberg ME, Lawrence SE, Gower AL, et al. Are HIV prevention services reaching all LGBTQ+ youth? An intersectional analysis in a national sample. AIDS Behav 2024 Apr;28(4):1435-1446. [doi: 10.1007/s10461-023-04230-w] [Medline: 38085427]
5. Arrington-Sanders R, Hailey-Fair K, Wirtz AL, et al. Role of structural marginalization, HIV stigma, and mistrust on HIV prevention and treatment among young Black Latinx men who have sex with men and transgender women: perspectives from youth service providers. AIDS Patient Care STDS 2020 Jan;34(1):7-15. [doi: 10.1089/apc.2019.0165] [Medline: 31944853]
6. Kerr J, Lelutiu-Weinberger C, Nelson LE, et al. Addressing Intersectional stigma in programs focused on ending the HIV epidemic. Am J Public Health 2022 Jun;112(S4):S362-S366. [doi: 10.2105/AJPH.2021.306657] [Medline: 35763743]
7. Magno L, Silva LD, Veras MA, Pereira-Santos M, Dourado I. Stigma and discrimination related to gender identity and vulnerability to HIV/AIDS among transgender women: a systematic review. Cad Saude Publica 2019;35(4):e00112718. [doi: 10.1590/0102-311X00112718] [Medline: 30994744]
8. Storholm ED, Siconolfi DE, Campbell CK, et al. Structural inequities, syndemics, and resilience: the critical role of social support in overcoming barriers and empowering engagement in HIV care for young Black sexual-minority men in the US South. J Racial Ethn Health Disparities 2025 Feb;12(1):250-261. [doi: 10.1007/s40615-023-01869-y] [Medline: 38095825]
9. Fauci AS, Redfield RR, Sigounas G, Weahkee MD, Giroir BP. Ending the HIV epidemic. JAMA 2019 Mar 5;321(9):844. [doi: 10.1001/jama.2019.1343] [Medline: 30730529]
10. Rodriguez A, Horvath KJ, Dowshen N, et al. Awareness and utilization of pre-exposure prophylaxis and HIV prevention services among transgender and non-binary adolescent and young adults. Front Reprod Health 2023;5:1150370. [doi: 10.3389/frph.2023.1150370] [Medline: 38318604]
11. Warus J, Hidalgo MA, Wong CF, Kipke M, Olson-Kennedy J. 165. Relationships between self-perceived risk of HIV, behavioral risk of HIV, and self-reported pre-exposure prophylaxis (PREP) utilization among young men who have sex with men of color: preliminary findings from a prospective cohort study. Journal of Adolescent Health 2024 Mar;74(3):S88. [doi: 10.1016/j.jadohealth.2023.11.364]
12. Dang M, Scheim AI, Teti M, et al. Barriers and facilitators to HIV pre-exposure prophylaxis uptake, adherence, and persistence among transgender populations in the United States: A systematic review. AIDS Patient Care STDS 2022 Jun;36(6):236-248. [doi: 10.1089/apc.2021.0236] [Medline: 35687813]
13. Brooks RA, Nieto O, Landrian A, Fehrenbacher A, Cabral A. Experiences of pre-exposure prophylaxis (PrEP)-related stigma among Black MSM PrEP users in Los Angeles. J Urban Health 2020 Oct;97(5):679-691. [doi: 10.1007/s11524-019-00371-3] [Medline: 31214977]
14. Garrison LE, Haberer JE. Pre-exposure prophylaxis uptake, adherence, and persistence: a narrative review of interventions in the U.S. Am J Prev Med 2021 Nov;61(5 Suppl 1):S73-S86. [doi: 10.1016/j.amepre.2021.04.036] [Medline: 34686294]
15. Pre-exposure prophylaxis (PrEP) for women: what you need to know. Centers for Disease Control and Prevention. 2022. URL: https://web.archive.org/web/20221130091612/https://www.cdc.gov/stophivtogether/library/topics/prevention/brochures/cdc-lsht-prevention-brochure-prep-is-for-women-patient.pdf [accessed 2024-11-26]
16. FDA approves first injectable treatment for HIV pre-exposure prevention. FDA. URL: https://www.fda.gov/news-events/press-announcements/fda-approves-first-injectable-treatment-hiv-pre-exposure-prevention [accessed 2024-11-26]

XSL•FO
RenderX

17. FDA approves second drug to prevent HIV infection as part of ongoing efforts to end the HIV epidemic. US Food and Drug Administration. 2019. URL: https://web.archive.org/web/20221205071543/https://www.fda.gov/news-events/press-announcements/fda-approves-second-drug-prevent-hiv-infection-part-ongoing-efforts-end-hiv-epidemic [accessed 2024-11-26]
18. Mayer KH, Molina JM, Thompson MA, et al. Emtricitabine and tenofovir alafenamide vs emtricitabine and tenofovir disoproxil fumarate for HIV pre-exposure prophylaxis (DISCOVER): primary results from a randomised, double-blind, multicentre, active-controlled, phase 3, non-inferiority trial. Lancet 2020 Jul 25;396(10246):239-254. [doi: 10.1016/S0140-6736(20)31065-5] [Medline: 32711800]
19. Molina JM, Charreau I, Spire B, et al. Efficacy, safety, and effect on sexual behaviour of on-demand pre-exposure prophylaxis for HIV in men who have sex with men: an observational cohort study. Lancet HIV 2017 Sep;4(9):e402-e410. [doi: 10.1016/S2352-3018(17)30089-9] [Medline: 28747274]
20. Horvath KJ, Helm JL, Black A, et al. A pilot randomized controlled trial of an mHealth intervention to improve PREP adherence among young sexual minority men. AIDS Behav 2024 Aug;28(8):2804-2820. [doi: 10.1007/s10461-024-04374-3] [Medline: 38816592]
21. Stephenson R, Todd K, Kahle E, et al. Project moxie: results of a feasibility study of a telehealth intervention to increase HIV testing among binary and nonbinary transgender youth. AIDS Behav 2020 May;24(5):1517-1530. [doi: 10.1007/s10461-019-02741-z] [Medline: 31760536]
22. Jones J, Butler G, Woody M, et al. Preferences for and experiences of an HIV-prevention mobile app designed for transmasculine people: pilot feasibility trial and qualitative investigation. JMIR Form Res 2023 Sep 21;7:e51055. [doi: 10.2196/51055] [Medline: 37733430]
23. Mulawa MI, Rosengren AL, Amico KR, Hightow-Weidman LB, Muessig KE. mHealth to reduce HIV-related stigma among youth in the United States: a scoping review. Mhealth 2021;7:35. [doi: 10.21037/mhealth-20-68] [Medline: 33898604]
24. Hightow-Weidman LB, LeGrand S, Muessig KE, et al. A randomized trial of an online risk reduction intervention for young Black MSM. AIDS Behav 2019 May 15;23(5):1166-1177. [doi: 10.1007/s10461-018-2289-9] [Medline: 29995205]
25. Stocks JB, Calvetti S, Rosso MT, et al. Evaluating the feasibility and acceptability of a digital pre-exposure prophylaxis navigation and activation intervention for racially and ethnically diverse sexual and gender minority youth (PrEPresent): protocol for a pilot randomized controlled trial. JMIR Res Protoc 2023 Sep 29;12:e50866. [doi: 10.2196/50866] [Medline: 37773616]
26. Hibbard JH, Stockard J, Mahoney ER, Tusler M. Development of the patient activation measure (PAM): conceptualizing and measuring activation in patients and consumers. Health Serv Res 2004 Aug;39(4 Pt 1):1005-1026. [doi: 10.1111/j.1475-6773.2004.00269.x] [Medline: 15230939]
27. Daughtridge G, Hebel S, Fischer M, Hashim J, Kahn-Woods E, Kardashian K. Development and validation of a point-of-care, urine assay to measure adherence to prep and ART. Presented at: International AIDS Society Conference on HIV Science 2019; Jul 21-24, 2019; Mexico CIty, Mexico URL: https://www.ias2019.org/Portals/5/IAS2019_Abstracts.pdf [accessed 2025-10-20] [doi: 10.1002/jia2.25327]
28. Azucar D, Rusow JA, Slay L, et al. Combining the HYM (healthy young men's) cohort study and the TRUTH (A Trans Youth of Color Study): protocol for an expanded mixed methods study renewal. JMIR Res Protoc 2022 Nov 3;11(11):e39232. [doi: 10.2196/39232] [Medline: 36326811]
29. Desrosiers A, Levy M, Dright A, et al. A randomized controlled pilot study of a culturally-tailored counseling intervention to increase uptake of HIV pre-exposure prophylaxis among young Black men who have sex with men in Washington, DC. AIDS Behav 2019 Jan;23(1):105-115. [doi: 10.1007/s10461-018-2264-5] [Medline: 30171452]
30. Doblecki-Lewis S, Butts S, Botero V, Klose K, Cardenas G, Feaster D. A randomized study of passive versus active PrEP patient navigation for a heterogeneous population at risk for HIV in South Florida. J Int Assoc Provid AIDS Care 2019;18:2325958219848848. [doi: 10.1177/2325958219848848] [Medline: 31131679]
31. Rolle CP, Rosenberg ES, Siegler AJ, et al. Challenges in translating PREP interest into uptake in an observational study of young Black MSM. J Acquir Immune Defic Syndr 2017 Nov 1;76(3):250-258. [doi: 10.1097/QAI.0000000000001497] [Medline: 28708811]
32. Chan PA, Nunn A, van den Berg JJ, et al. A randomized trial of a brief behavioral intervention for PREP uptake among men who have sex with men at increased risk for HIV infection. J Acquir Immune Defic Syndr 2021 Jul 1;87(3):937-943. [doi: 10.1097/QAI.0000000000002671] [Medline: 33734099]
33. Sullivan PS, Stephenson R, Hirshfield S, et al. Behavioral efficacy of a sexual health mobile app for men who have sex with men: randomized controlled trial of mobile messaging for men. J Med Internet Res 2022 Feb 2;24(2):e34574. [doi: 10.2196/34574] [Medline: 35025755]
34. Teixeira da Silva D, Bouris A, Ramachandran A, et al. Embedding a linkage to preexposure prophylaxis care intervention in social network strategy and partner notification services: results from a pilot randomized controlled trial. J Acquir Immune Defic Syndr 2021 Feb 1;86(2):191-199. [doi: 10.1097/QAI.0000000000002548] [Medline: 33109935]
35. Grov C, Patel VV, D'Angelo AB, Zohra F, Appenroth MN. Close, but not quite: one in twenty trans and gender diverse individuals assigned female at birth are being prescribed the wrong kind of PREP. J Acquir Immune Defic Syndr 2021 Oct 1;88(2):e10-e11. [doi: 10.1097/QAI.0000000000002756] [Medline: 34506363]

XSL•FO
RenderX

36. Baltierra NB, Muessig KE, Pike EC, LeGrand S, Bull SS, Hightow-Weidman LB. More than just tracking time: complex measures of user engagement with an internet-based health promotion intervention. J Biomed Inform 2016 Feb;59:299-307. [doi: 10.1016/j.jbi.2015.12.015] [Medline: 26732997]
37. Mulawa MI, Mtukushe B, Knippler ET, et al. Supporting adolescents with HIV in South Africa through an adherence-supporting app: mixed methods beta-testing study. JMIR Form Res 2023 Jun 1;7:e47575. [doi: 10.2196/47575] [Medline: 37261883]
38. Choi SK, Golinkoff J, Michna M, Connochie D, Bauermeister J. Correlates of engagement within an online HIV prevention intervention for single young men who have sex with men: randomized controlled trial. JMIR Public Health Surveill 2022 Jun 27;8(6):e33867. [doi: 10.2196/33867] [Medline: 35759333]
39. Reback CJ, Cain D, Rusow JA, et al. Technology-based interventions, with a stepped care approach, for reducing sexual risk behaviors and increasing PrEP initiation among transgender and gender expansive youth and young adults. AIDS Behav 2024 Dec;28(12):3956-3969. [doi: 10.1007/s10461-024-04513-w] [Medline: 39304589]
40. Biello KB, Daddario SR, Hill-Rorie J, et al. Uptake and acceptability of MyChoices: results of a pilot RCT of a mobile app designed to increase HIV testing and PrEP uptake among young American MSM. AIDS Behav 2022 Dec;26(12):3981-3990. [doi: 10.1007/s10461-022-03724-3] [Medline: 35829971]
41. Daughtridge G, Kahn-Woods E, Enghuus C, et al. The time is now for disruptive innovation in pre-exposure prophylaxis adherence monitoring. HIV/AIDS Res Treat Open J 2020;7(1):5-9. [doi: 10.17140/HARTOJ-7-133]
42. Hebel S, Kahn-Woods E, Malone-Thomas S, et al. Brief report: discrepancies between self-reported adherence and a biomarker of adherence in real-world settings. J Acquir Immune Defic Syndr 2020 Dec 1;85(4):454-457. [doi: 10.1097/QAI.0000000000002486] [Medline: 33136744]
43. Hunt T, Lalley-Chareczko L, Daughtridge G, Swyryn M, Koenig H. Challenges to PrEP use and perceptions of urine tenofovir adherence monitoring reported by individuals on PrEP. AIDS Care 2019 Oct;31(10):1203-1206. [doi: 10.1080/09540121.2019.1587369] [Medline: 30821473]
44. Sallabank G, Stephenson R, Gandhi M, Merrill L, Sharma A. Lessons learned from the implementation of a pilot study on self-collected specimen return by sexual minority men (Project Caboodle!): qualitative exploration. JMIR Form Res 2023 Apr 6;7:e43539. [doi: 10.2196/43539] [Medline: 37023442]
45. Grant RM, Anderson PL, McMahan V, et al. Uptake of pre-exposure prophylaxis, sexual practices, and HIV incidence in men and transgender women who have sex with men: a cohort study. Lancet Infect Dis 2014 Sep;14(9):820-829. [doi: 10.1016/S1473-3099(14)70847-3] [Medline: 25065857]
46. Cottrell ML, Yang KH, Prince HMA, et al. A translational pharmacology approach to predicting outcomes of preexposure prophylaxis against HIV in men and women using tenofovir disoproxil fumarate with or without emtricitabine. J Infect Dis 2016 Jul 1;214(1):55-64. [doi: 10.1093/infdis/jiw077] [Medline: 26917574]
47. Hannaford A, Arens Y, Koenig H. Real-time monitoring and point-of-care testing: a review of the current landscape of PrEP adherence monitoring. Patient Prefer Adherence 2021;15:259-269. [doi: 10.2147/PPA.S248696] [Medline: 33574659]
48. Raifman J, Nunn A, Oldenburg CE, et al. An evaluation of a clinical pre-exposure prophylaxis education intervention among men who have sex with men. Health Serv Res 2018 Aug;53(4):2249-2267. [doi: 10.1111/1475-6773.12746] [Medline: 28744983]
49. Rogers BG, Tao J, Maynard M, et al. Characterizing the impact of COVID-19 on pre-exposure prophylaxis (PREP) care. AIDS Behav 2021 Nov;25(11):3754-3757. [doi: 10.1007/s10461-021-03337-2] [Medline: 34114166]
50. Dawson L, Frederiksen B, Gomez I. PREP access in the United States: the role of telehealth. KFF. 2022. URL: https://www.kff.org/hivaids/issue-brief/prep-access-in-the-united-states-the-role-of-telehealth [accessed 2024-10-29]
51. Biello KB, Hill-Rorie J, Valente PK, et al. Development and evaluation of a mobile app designed to increase HIV testing and pre-exposure prophylaxis use among young men who have sex with men in the United States: open pilot trial. J Med Internet Res 2021 Mar 24;23(3):e25107. [doi: 10.2196/25107] [Medline: 33759792]
52. Whiteley L, Craker L, Haubrick KK, et al. The impact of a mobile gaming intervention to increase adherence to pre-exposure prophylaxis. AIDS Behav 2021 Jun;25(6):1884-1889. [doi: 10.1007/s10461-020-03118-3] [Medline: 33483897]
53. Jitmun W, Palee P, Choosri N, Surapunt T. The success of serious games and gamified systems in HIV prevention and care: scoping review. JMIR Serious Games 2023 Sep 5;11:e39915. [doi: 10.2196/39915] [Medline: 37669098]
54. Mulawa MI, Hoare J, Knippler ET, et al. MASI, a smartphone app to improve treatment adherence among South African adolescents and young adults with HIV: protocol for a pilot randomized controlled trial. JMIR Res Protoc 2023 Sep 19;12:e47137. [doi: 10.2196/47137] [Medline: 37725409]
55. Choi SK, Muessig KE, Hightow-Weidman LB, Bauermeister JA. Paradata: Measuring engagement in digital HIV interventions for sexual and gender minorities. Curr HIV/AIDS Rep 2023 Dec;20(6):487-501. [doi: 10.1007/s11904-023-00679-5] [Medline: 37930613]
56. Arnold EA, Saberi P, Wong JO, et al. We are family: a feasibility and acceptability study of an HIV prevention intervention with the house ball and gay family communities. J Acquir Immune Defic Syndr 2021 Dec 15;88(S1):S6-S11. [doi: 10.1097/QAI.0000000000002805] [Medline: 34757987]
57. Paul ME, Castillo M, Emmanuel P, et al. Scale up mHealth HIV interventions: site and public health perspectives and lessons learned from P3. Mhealth 2021. [doi: 10.17615/a3rb-mj69]

XSL•FO
RenderX

58. Sullivan PS, Hightow-Weidman L. Mobile apps for HIV prevention: how do they contribute to our epidemic response for adolescents and young adults? Mhealth 2021;7:36. [doi: 10.21037/mhealth-20-71] [Medline: 33898605]
59. Horvath KJ, Ma J, Storholm ED, Black A, Klaphake J, Baker JV. The use of daily and on-demand oral pre-exposure prophylaxis dosing strategies among young adult gay, bisexual and other men who have sex with men enrolled in an mHealth adherence intervention. AIDS Behav 2023 Nov;27(11):3632-3644. [doi: 10.1007/s10461-023-04077-1] [Medline: 37195471]
60. Celum C, Grinsztejn B, Ngure K. Preparing for long-acting PrEP delivery: building on lessons from oral PrEP. J Int AIDS Soc 2023 Jul;26 Suppl 2(Suppl 2):e26103. [doi: 10.1002/jia2.26103] [Medline: 37439077]
61. Mann LM, Zhu W, Huang YLA, Kourtis AP, Fanfair RN, Hoover KW. Trends in oral and injectable HIV preexposure prophylaxis prescriptions in the US, 2013-2023. JAMA 2024 Nov 12;332(18):1580-1583. [doi: 10.1001/jama.2024.21493] [Medline: 39401236]
62. Injectable lenacapavir for PREP. PREPWatch. 2025. URL: https://www.prepwatch.org/products/lenacapavir-for-prep [accessed 2025-07-18]
63. Preexposure prophylaxis (PREP) using antiretroviral therapy to prevent human immunodeficiency virus (HIV) infection. CMS.gov. 2024. URL: https://www.cms.gov/medicare-coverage-database/view/ncacal-decision-memo.aspx?proposed=N&ncaid=310 [accessed 2024-11-27]

Abbreviations:

HMP: HealthMpowerment
mHealth: mobile health
PrEP: pre-exposure prophylaxis
RCT: randomized control trial
SGM: sexual and gender minority
STI: sexually transmitted infection
SUS: System Usability Scale
YAB: Youth Advisory Board

Edited by A Mavragani; submitted 06.12.24; peer-reviewed by G Rebchook, S Maccarthy; revised version received 29.08.25; accepted 30.08.25; published 31.10.25.

Please cite as:
Calvetti S, Lei B, Stocks JB, Rosso MT, Puentes M, Durazo-Arvizu R, Slay L, Kipke MD, Hightow-Weidman LB
Feasibility and Acceptability of a mHealth Patient Navigation Intervention to Increase Pre-Exposure Prophylaxis Uptake in Racially and Ethnically Diverse Sexual and Gender Minority Youth in Los Angeles (PrEPresent): Pilot Randomized Controlled Trial
JMIR Form Res 2025;9:e69255
URL: https://formative.jmir.org/2025/1/e69255
doi:10.2196/69255

© Sam Calvetti, Bryan Lei, Jacob B Stocks, Matthew T Rosso, Manuel Puentes, Ramon Durazo-Arvizu, Lindsay Slay, Michele D Kipke, Lisa B Hightow-Weidman. Originally published in JMIR Formative Research (https://formative.jmir.org), 31.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Sleep and Activity Patterns as Transdiagnostic Behavioral Biomarkers in Psychiatry: Longitudinal Observational Study From the DeeP-DD Study

Dylan Hamitouche[1,2]; Tihare Zamorano[1,3], BSc; Youcef Barkat[1,4], MSc; Deven Parekh[1], MSc; Lena Palaniyappan[1,5,6,7], MD, PhD; Sara Jalali[1], MSc; David Benrimoh[1,5], MSc, MD

[1]Douglas Mental Health University Institute, 6875 LaSalle boulevard, Verdun, QC, Canada
[2]Department of Medicine, Faculty of Medicine and Health Sciences, McGill University, Montreal, QC, Canada
[3]Integrated Program in Neuroscience (IPN), Faculty of Medicine and Health Sciences, McGill University, Montreal, QC, Canada
[4]Department of Biochemistry, Faculty of Medicine, Université de Montréal, Montreal, QC, Canada
[5]Department of Psychiatry, McGill University, Montreal, QC, Canada
[6]Robarts Research Institute, Western University, London, ON, Canada
[7]Department of Psychology, McGill University, Montreal, QC, Canada

Corresponding Author:
Dylan Hamitouche
Douglas Mental Health University Institute, 6875 LaSalle boulevard, Verdun, QC, Canada

Abstract

Background: Despite widespread use of symptom rating scales in psychiatry, these tools are limited by reliance on self-report, infrequent administration, and lack of predictive power. This constrains clinicians' ability to monitor illness trajectories or anticipate adverse outcomes like relapse. Actigraphy, a passive wearable-based method for measuring sleep and physical activity, offers objective, high-resolution behavioral data that may better reflect symptom fluctuations. Prior research has shown associations between actigraphy features and mood or psychosis symptoms, but most studies have focused on narrow diagnostic groups or fixed time windows, limiting clinical translation.

Objective: This study aims to examine whether actigraphy-derived sleep and activity features correlate with psychiatric symptom severity in a transdiagnostic psychiatric sample, and to identify which features are most clinically relevant across multiple temporal resolutions.

Methods: We present a feasibility case series study analyzing preliminary data from 8 outpatients (ages 18 - 52 years) enrolled in the Deep Phenotyping and Digitalization at Douglas (DeeP-DD) study, a prospective transdiagnostic study of digital phenotyping. Participants wore wrist-based actigraphy devices (GENEActiv) for up to 5 months. Symptom severity was measured using a variety of self- and clinician-rated scales. We performed intraindividual Spearman correlations and interindividual repeated measures correlations across daily, weekly, monthly, and full-duration averages.

Results: Intraindividual analyses revealed that later rise times were significantly associated with higher weekly 9-item Patient Health Questionnaire (PHQ-9) scores in participant 7 (ρ=0.74, P<.001) and participant 4 (ρ=0.78, P=.02), as well as higher weekly 7-item General Anxiety Disorder (GAD-7) scores in participant 7 (ρ=0.59, P=.03). While similar trends were observed at daily and monthly timescales, the weekly resolution yielded the most robust significance. Interindividual analyses showed that weeks with later average rise time correlated with higher PHQ-9 (r=0.48, P<.001) and GAD-7 scores (r=0.38, P=.03), with the PHQ-9 association remaining significant after Bonferroni correction (Bonferroni-corrected P=.02). Increased light physical activity was linked to lower PHQ-9 scores weekly (r=−0.44, P=.001) and monthly (r=−0.53, P=.01). Over the whole duration of the study, increased levels of sedentary activity were associated with lower GAD-7 scores (ρ=0.74; P<.001).

Conclusions: Our findings highlight actigraphy-derived sleep and activity features, particularly rise time and physical activity, as promising transdiagnostic markers of psychiatric symptom burden. Their consistent associations across temporal scales and diagnostic groups underscore their potential utility for scalable, real-world clinical monitoring. Future work should validate these findings in larger cohorts and explore advanced analytical methods to capture circadian rhythmicity and symptom dynamics more precisely.

(JMIR Form Res 2025;9:e81107) doi:10.2196/81107

KEYWORDS

actigraphy; digital biomarkers; digital phenotyping; physical activity; sleep

XSL•FO
RenderX

Introduction

Despite decades of research into the neurobiology of various mental illnesses, there remains a lack of clinically reliable tools for identifying, monitoring, and staging most psychiatric conditions [1-4]. In major depression, best practices involve standardized rating scales as part of measurement-based care, which can help patients achieve remission more quickly [5-8]. These scales are clinically useful for tracking symptom severity and guiding treatment decisions. However, there are no widely accepted protocols for monitoring patients post-remission, and in severe mental illnesses such as schizophrenia, patient self-report is less reliable, while rating scales add to clinicians' workload [9-12]. Moreover, rating scales may fail to capture what matters most to patients [13]—for instance, the commonly used 9-item Patient Health Questionnaire (PHQ-9) [14,15] does not distinguish between hypersomnia and insomnia. While rating scales are helpful for assessing illness status or guiding treatment adjustments, they are seldom used to predict events such as rehospitalization. This may be due in part to their limited predictive utility, as well as their infrequent and inconsistent use in clinical settings. Therefore, this gap in our ability to anticipate and prevent adverse outcomes highlights the need for more reliable, actionable predictors in clinical care.

In response, and in parallel with the rise of smartphones and wearables, interest has grown in "digital biomarkers" [16,17]—sensor-derived data collected passively by patient devices. Among those, digital monitoring and predictive biomarkers often use both smartphones and wearables, capturing data such as app usage, communication, location, light, movement, sleep, activity, and heart rate [18,19]. These tools have been used to infer mood states or behavior by combining multiple data streams—for example, call logs and GPS for sociability and mobility [19]. Integrating passive (sensor) and active (questionnaire) digital tools with clinical, historical, and biological data is one approach to "deep phenotyping"—a detailed characterization of patients. Among these tools, actigraphy, a wearable-based method for measuring rest-activity cycles, has emerged as particularly relevant. It provides objective, continuous data on sleep and activity, which are critical in conditions like bipolar disorder [20,21] and are often poorly captured by self-report [22].

Actigraphy has shown clinical relevance as a monitoring and predictive tool by detecting changes linked to symptom trajectories across psychiatric disorders. In schizophrenia, sleep disturbances captured by actigraphy may worsen symptoms via mood and attention-related pathways [23-25]. Studies show daily fluctuations in sleep and activity track symptoms, with physical activity linked to improved same-day mood but sometimes worse next-day symptoms [26]. In mood disorders, machine learning models using actigraphy can predict next-day depressive or manic episodes [27]. Depressed patients show reduced daytime activity, with age-specific sleep difficulties suggesting circadian disruption as a core mechanism [28-30]. Finally, in bipolar disorder, sleep and activity disturbances persist even during euthymia and are tied to episode onset, severity, and long-term outcomes [31-33].

Despite their promise, existing studies of actigraphy-derived sleep and activity features have often examined narrow diagnostic groups or short, fixed observation windows, limiting generalizability [34,35]. This limitation is significant given the multitude of available actigraphy measures, making it essential to identify which features demonstrate reliability and clinical relevance across diagnostic categories to optimize data collection and enhance clinical interpretation. Furthermore, since actigraphy data can be examined over multiple temporal scales, from days to months, understanding which measures maintain their utility across these timescales is critical. For instance, measures that are only informative over extended periods restrict their applicability for early or real-time clinical inference, as a substantial accumulation of data would be necessary before meaningful conclusions can be drawn. Addressing these limitations will improve the clinical utility and efficiency of actigraphy-based monitoring in psychiatric research and clinical practice.

This gap in the literature motivates the exploration of actigraphy-derived sleep and physical activity features as candidate monitoring and predictive biomarkers that transcend diagnostic categories and temporal scales. As part of the Deep Phenotyping and Digitalization at Douglas (DeeP-DD) project, a prospective transdiagnostic study aiming to identify clinically useful, scalable, and interpretable digital markers of mental illness, we present a case series study aiming to demonstrate the feasibility of actigraphy-based monitoring across diagnostic categories and temporal scales. We report data completeness rates, highlight the relevance of actigraphy for tracking subjective mood within and between individuals, and explore early signals of clinical relevance across multiple timescales in a real-world, diagnostically heterogeneous sample. Identifying robust, objective markers of illness could support clinical psychiatry by reducing reliance on disorder-specific tools and providing utility even when diagnostic clarity is limited, such as in youth mental health populations.

Methods

Overview

DeeP-DD is an ongoing feasibility study at the Douglas Mental Health University Institute in Montreal, Canada. Its objective is to test the feasibility and acceptability of multimodal digital phenotyping in a realistic transdiagnostic clinical population, including individuals with early psychosis, schizophrenia or schizoaffective disorder, bipolar disorder, anxiety, unipolar depression, and other conditions such as personality or substance use disorders. The study evaluates a wide range of digital tools, including actigraphy, smartphone sensors, ecological momentary assessments, and self-report questionnaires, and aims to develop interpretable "return of results" clinical reports to support personalized care.

DeeP-DD is designed as a feasibility and demonstration study rather than to test specific quantitative hypotheses. The planned target sample size is approximately 50 participants per diagnostic group (psychosis, bipolar disorder, depression, and personality or addiction disorder), for a total of about 200 participants. This sample size was chosen to enable assessment

XSL•FO
RenderX

of transdiagnostic feasibility and data completeness across multiple modalities. As data collection progresses, analyses across all DeeP-DD components will inform the required sample size for a future interventional trial evaluating the impact of a combined digital phenotyping intervention and personalized feedback report. In this study, we focus on preliminary actigraphy and questionnaire data from the initial pilot phase, as these modalities were available for most participants.

Ethical Considerations

The study was approved by the research ethics board of the West-Central Montreal health authority (Project No. 2023 - 816) and conducted in line with the Declaration of Helsinki and the Tri-Council Policy Statement. Participants met with researchers who explained the study before providing written informed consent. Informed consent was obtained from all participants, and those who agreed to provide actigraphy data were provided with a GENEActiv wristband (Activinsights, Cambridge, UK) and given instructions on its use. Participant compensation was explicitly tied to engagement with study procedures to encourage adherence: participants received CAD $5 (US $3.55) per week for consistently wearing the actigraphy watch and CAD $7 (US $4.96) per week for completing the questionnaires. This approach was designed to motivate sustained participation. While this was a small compensation per week, it is still compensation, which would not be available in clinical settings; while this is fairly standard in research designs, the reader should keep this in mind when interpreting our results, especially with respect to data contribution rates. Participants provided written informed consent prior to enrollment. All data were deidentified and stored on secure servers, accessible only to authorized study personnel. No clinical care decisions were influenced by participation.

Participant Recruitment

Participants are outpatients recruited from the Clinical High-Risk for Psychosis Clinic, First-Episode Psychosis Clinic, and the Bipolar Disorder Clinic; 1 patient with a primary personality disorder diagnosis was recruited from the general Neuropsychiatry Clinic. The study was advertised by clinicians directly to their patients during routine clinical visits and through posters displayed in the clinics. Participants are referred to the study by their clinicians or self-referred to the study after seeing study advertisements at their clinic.

Data Collection

Patient- and clinician-rated questionnaire data were collected using the REDCap platform [36]. For every participant, a medical chart review was performed at the Douglas Mental Institute assessing overall symptomatology, medication use, and other clinically relevant information. Actigraphy data were extracted from the GENEActiv wristbands using GENEActiv PC Software (version 3.3; ActivInsights) [37], a validated approach to monitor sleep in adults [31-33]. The REDCap questionnaires included in the analyses were the PHQ-9 [14], which assesses depressive symptoms; 7-item General Anxiety Disorder 7 (GAD-7) [38], which evaluates anxiety severity; Clinical Global Impression - Severity (CGI-S) [39], which measures overall illness severity; Scale for the Assessment of Positive Symptoms (SAPS) [40], which examines psychotic symptoms; and Scale for the Assessment of Negative Symptoms (SANS) [41], which assesses deficits in normal emotional and behavioral functioning. The SAPS, SANS, and CGI-S were clinician-rated, while the PHQ-9 and GAD-7 were self-reported. The PHQ-9 and GAD-7 were administered weekly, the SAPS and SANS monthly, and the CGI-S at each clinical visit. Actigraphy data were inaccessible to participants and clinicians during the monitoring month but were made available to both after each monthly monitoring period concluded. As this was a real-world clinical feasibility study, both the timing of clinician-administered questionnaires and return of results depended on clinical follow-up schedules and as such were subject to variation in timing.

Actigraphy Data Processing

Sleep and physical activity features extraction were processed from the raw movement data using GENEActiv default R markdown analysis tools [42]. Participants wore the actigraphy device for 1 month at a time (the length of a single charge), after which it was replaced with a newly charged device. The sleep features extracted and included in the analyses were as follows: total sleep time, sleep efficiency (ie, time spent asleep divided by time spent in bed), number of active periods per night, median length of those active periods, sleep onset time, rise time, day-to-day sleep onset time variability, and day-to-day rise time variability. The GENEActiv PC Software provided the daily step count and activity modes classified as “sedentary,” “light physical activity,” “moderate physical activity,” and “vigorous physical activity” [37]. We excluded vigorous activity from our analyses, as participants spent very little time in this mode (often less than 5 min per day), and these brief episodes were likely due to artifacts or misclassification rather than true vigorous exertion. We also included the daily duration of actigraphy nonwear in our analyses, as these periods may reflect participant behavior or symptom fluctuations and could provide meaningful insights into health status [43].

No filtering was performed on these features to maintain a straightforward preprocessing pipeline and preserve all available data, reflecting the exploratory nature of our analyses and the intent to apply these data in routine clinical practice using existing analysis tools. Due to the limited sample size, missing data were not imputed to avoid introducing bias. Actigraphy features and questionnaire scores were averaged by time unit (daily, weekly, monthly, and overall duration of study) to enable consistent comparisons between behavioral data and psychiatric outcomes across multiple time scales. Days with no data output despite the participant wearing the watch were excluded, as these were considered to result from sensor malfunction. To simplify preprocessing, enhance translatability, and given high participant compliance, we considered a day valid if any actigraphy data were recorded, and a week or month valid if it contained at least 1 such day. The results from sensitivity analyses using stricter thresholds (≥3 or ≥4 valid days for a valid week, and ≥10 or ≥15 valid days for a valid month) are available in Multimedia Appendix 1.

XSL•FO
RenderX

Statistical Analyses

We conducted analyses at the intraparticipant level to capture how actigraphy features relate to psychiatric symptoms within each individual over time. For each participant, we computed Spearman rank correlations between actigraphy features and questionnaire scores at daily, weekly, and monthly resolutions, but only if at least 5 data points were available to ensure sufficient data for a stable and interpretable correlation estimate. This nonparametric method was chosen due to its robustness to nonlinear associations and ordinal or skewed data distributions. Correlations were computed independently for each participant and time scale. Bonferroni correction was applied across all tests to control for multiple comparisons.

Then, we performed analyses at the interparticipant level to identify consistent associations across the sample, accounting for repeated measures within individuals [44]. To do so, we used repeated measures correlation (Pingouin package, Python) [45] to assess associations between daily, weekly, and monthly actigraphy features and questionnaire scores across individuals, while accounting for nonindependence of within-subject data. Additionally, we computed Spearman correlations across participants using data aggregated over the full study period to capture stable between-subject associations and overall group-level trends not dependent on repeated measurements. To ensure reliable estimation, only correlations with at least 10 data points for daily analyses and 4 for weekly and full-study summaries were included, as interindividual analyses provided more data than intraindividual ones. These thresholds were chosen to balance data availability and reliability.

Results

Description

Data were collected from 8 participants, yielding a combined total of 33 months of actigraphy recordings. Participants were aged 18-30 years, included both males and females, and represented 3 primary psychiatric diagnoses with various comorbidities (Table S1 in Multimedia Appendix 1). Most participants met criteria for a first episode of psychosis, except for participant 5, who had a primary personality disorder, and participant 14, who had a primary bipolar disorder. Primary diagnoses included schizophrenia spectrum disorders (eg, schizophrenia, schizoaffective disorder), bipolar disorder, and personality disorder. Comorbid conditions commonly included depression, anxiety, attention-deficit or hyperactivity disorder, posttraumatic stress disorder, and personality disorders. Participants differed in clinical severity and care utilization: participants 4 and 8 had histories of hospitalization and multiple emergency room visits, while others (eg, participants 1, 5, and 13) had no acute care history. Medication use ranged from none (patient 1) to complex regimens involving antipsychotics, mood stabilizers, antidepressants, and stimulants (eg, participants 5 and 8) (Table S1 in Multimedia Appendix 1). Participants 1 and 6 did not consent to complete self-reported questionnaires.

The mean number of days of actigraphy data collected per participant was 110 (SD 43), ranging from 58 to 156 days. A total of 57 days were missing due to sensor malfunction, distributed across 3 participants (5, 7, and 14). Any additional missing days were due to the device being rotated among participants. Compliance with wearing the GeneActiv sensor was very high (97.27% of days showed no nonwear time; mean 1.58 min, median 0 min, SD 22.4 min). Across all recorded days of actigraphy (n=843), 820 showed no nonwear time, 23 had some nonwear time, and only 5 exceeded 1 hour of nonwear (maximum 9 h). On average, a valid week included 6.16 (median 7.00, SD 1.65, range 1 - 7) days of data, while a valid month included 19.65 (median 22.00, SD 9.50, range 3 - 31) days. Data missingness is reported in Figure 1 and Table 1.

XSL•FO
RenderX

Figure 1. Visualization of actigraphy compliance and questionnaire completion across participants. Each green box represents a day with completed data. In the actigraphy panel, a color gradient indicates compliance (with green representing perfect wear and red indicating high nonwear of the GeneActiv device), while gray areas indicate days without data, which may result from sensor malfunction, participants returning the device, or the device being unavailable due to maintenance or the device being given to another participant due to a limited number of devices. Day 1 marks the start of wristband use for each participant, and the final day corresponds to the last day of actigraphy recording across all participants. Participant numbers on the y-axis are not sequential, as they represent individual patient IDs. CGI-S: Clinical Global Impression–Severity; GAD-7: 7-item General Anxiety Disorder; PHQ-9: 9-item Patient Health Questionnaire; SANS: Scale for the Assessment of Negative Symptoms; SAPS: Scale for the Assessment of Positive Symptoms.

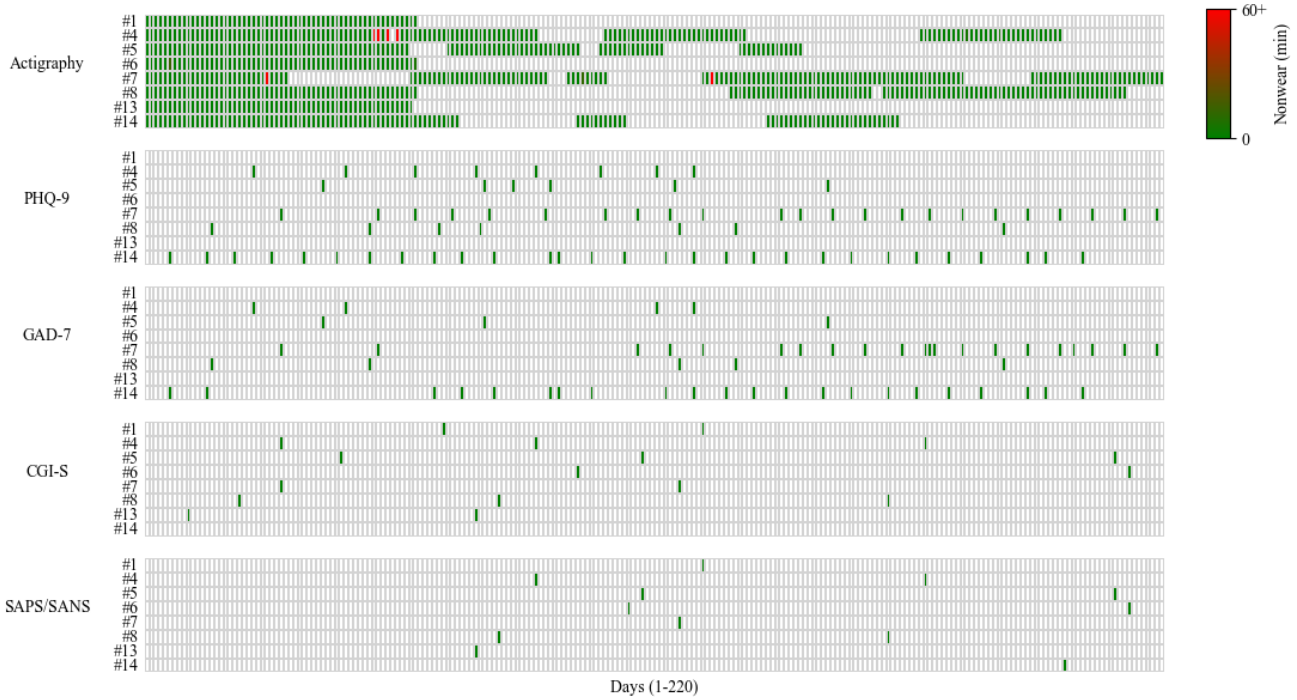

Table . Summary of actigraphy adherence and questionnaire completion.

	Values, mean (SD)	Min	Max
Days of actigraphy	110.1 (40.4)	58	156
Missing days of actigraphy	7.1 (11.5)	0	35
Completed actigraphy (%)	95.4 (7.5)	78	100
PHQ-9[a] completed	9.8 (11.0)	0	29
GAD-7[b] completed	7.5 (9.5)	0	25
CGI-S[c] completed	2.8 (1.2)	0	4
SAPS[d] or SANS[e] completed	2.1 (0.8)	1	3

[a]PHQ-9: 9-item Patient Health Questionnaire.

[b]GAD-7: 7-item General Anxiety Disorder.

[c]CGI-S: Clinical Global Impression–Severity.

[d]SAPS: Scale for the Assessment of Positive Symptoms.

[e]SANS: Scale for the Assessment of Negative Symptoms.

Intraparticipant Patterns

We analyzed same-day associations between questionnaire scores and actigraphy features within participants using Spearman correlations. In participant 7, higher GAD-7 scores were associated with a later sleep onset the night before (ρ=0.54, P=.04) and later same-day rise time (ρ=0.53, P=.04). However, these daily associations did not remain significant after Bonferroni correction. Figure 2 summarizes all significant associations established in this study.

XSL•FO
RenderX

Figure 2. Heatmap of significant associations between actigraphy features and clinical questionnaires across multiple time scales at both intra- and interparticipant levels. "+" and "–" indicate the direction of the correlation. Cells with bold borders represent features with significant associations at multiple time scales (or across different participants for intraindividual analyses). Each cell displays only the most significant time scale. CGI-S: Clinical Global Impression–Severity; GAD-7: 7-item General Anxiety Disorder; PHQ-9: 9-item Patient Health Questionnaire; SANS: Scale for the Assessment of Negative Symptoms; SAPS: Scale for the Assessment of Positive Symptoms.

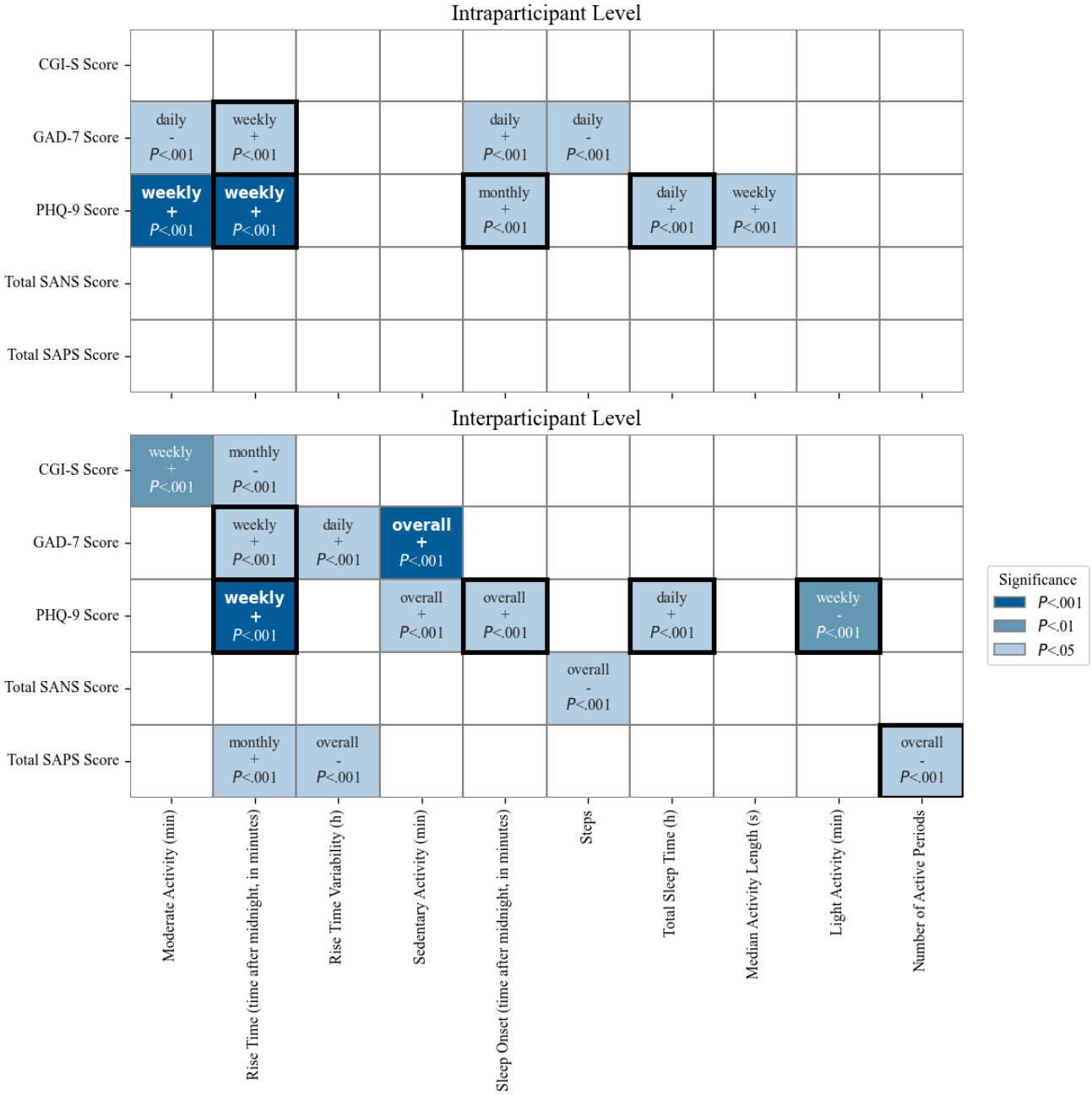

On a weekly time scale, we found that for participant 4 (patient with a schizoaffective disorder and a personality disorder), a later rise time was associated with higher PHQ-9 scores, although the significance did not survive Bonferroni correction (ρ=0.8; uncorrected *P*=.02 Bonferroni-corrected *P*=.3). For participant 7 (patient with bipolar disorder, personality disorder, and a first-episode psychosis), a later rise time was associated with both higher PHQ-9 (ρ=0.74; *P*<.001) and GAD-7 scores (ρ=0.59; *P*=.03), although only the association with PHQ-9 stayed significant after correction (*P*=.007). We found that for participant 5 (patient who has a primary personality disorder), a longer time of moderate physical activity was associated with higher PHQ-9 scores (ρ=1.0; *P*<.001); see discussion below. This association correlates clinically, as it was noted in the participant's clinical chart that they tended to go on long walks in the evening when feeling depressed. The higher level of significance for this participant is explained by the perfect correlation and limited number of overlapping data points for this participant (5 wk), which can inflate correlation coefficients.

Participant 7 showed that later average monthly rise time was associated with higher monthly average GAD-7 score (ρ=0.81; *P*=.05), while later average monthly sleep onset was associated with higher monthly average PHQ-9 scores (ρ=0.77; *P*=.04). None of these correlations remained significant after Bonferroni correction.

In summary, several associations emerged between mood scores and sleep timing or physical activity. After correction, only

XSL•FO
RenderX

weekly associations involving rise time and PHQ-9, and moderate activity and PHQ-9 (which was most likely idiosyncratic to participant 5), remained significant—highlighting rise time and activity patterns as potential behavioral markers of mood (Figure 3).

Figure 3. Significant intraindividual associations between actigraphy features and questionnaire scores across time scales. Asterisk (*) indicates fewer than 10 data points; Spearman correlation may be less reliable [46]. PHQ-9: 9-item Patient Health Questionnaire; GAD-7: 7-item General Anxiety Disorder.

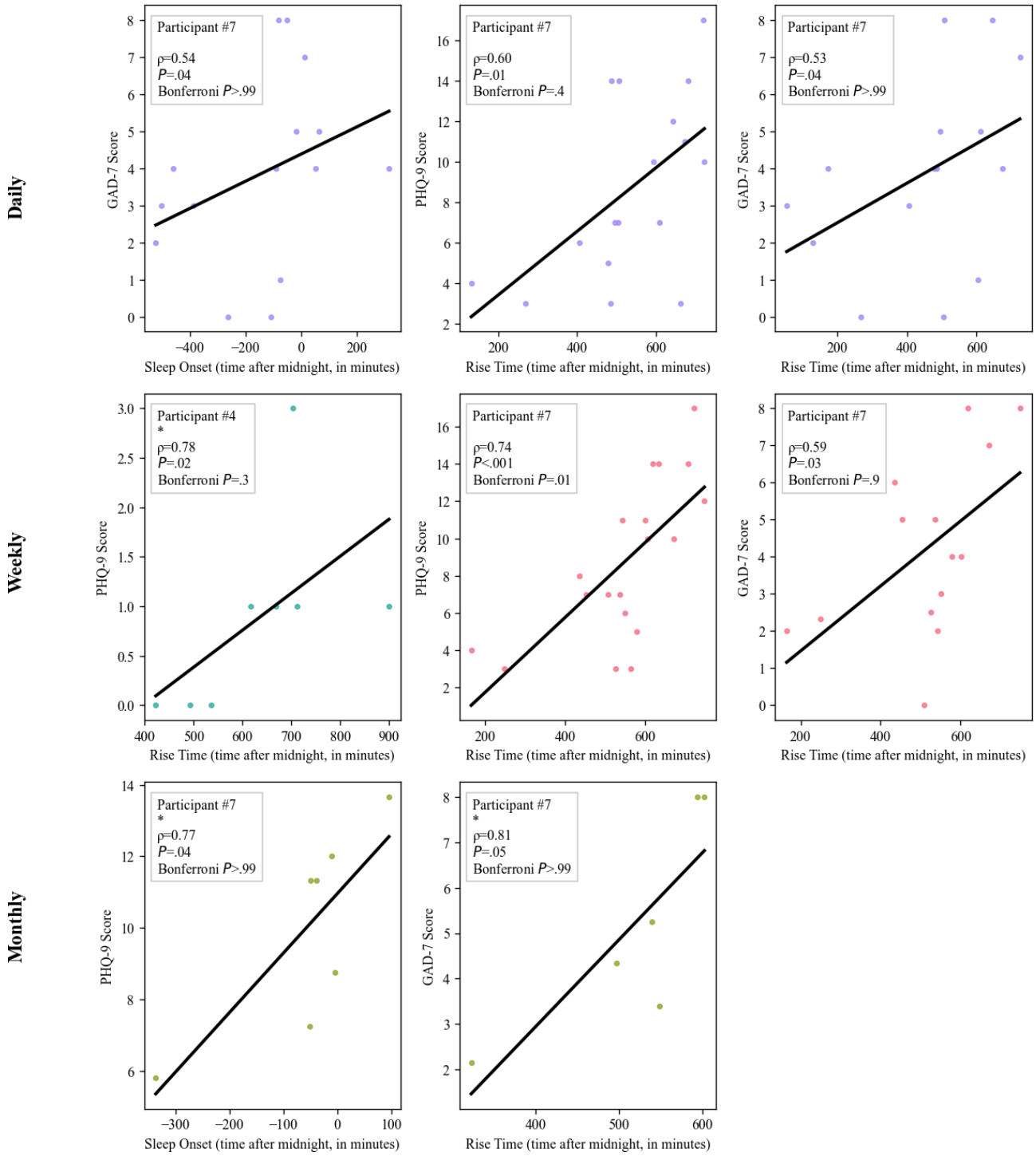

Interparticipant Trends

Same-day associations at the interindividual level showed longer total sleep duration the night before (r=0.30; P=.03). Increased rise-time variability, defined as the absolute difference in rise time between consecutive days, was correlated with higher GAD-7 scores (r=0.39, P=.02). However, none of the same-day associations survived correction for multiple comparisons.

Interestingly, the intraparticipant associations between rise time, sleep onset time, and questionnaire scores remained consistent in our interparticipant repeated measures correlation analyses conducted at a weekly time scale, such that weeks with a later average rise time were associated with a higher GAD-7 (r=0.38;

XSL•FO
RenderX

P=.03) and PHQ-9 score (*r*=0.49; *P*<.001). Furthermore, weeks of increased time spent doing light physical activity were associated with lower PHQ-9 (*r*=−0.44; *P*=.001). Only the association between PHQ-9 scores and rise time remained significant after Bonferroni correction (*P*=.02).

On a monthly time scale, increased light physical activity was associated with lower PHQ-9 scores (*r*=−0.53; *P*=.01). Later average monthly rise time was associated with both higher average CGI-S (*r*=−0.68; *P*=.05) and lower total SAPS (*r*=0.96; *P*=.04). The association with higher CGI-S appears contradictory to the rest of the study's findings linking later rise time with better outcomes, but this is likely driven by participant clustering—specifically, participant 4, who had consistently later rise times but low CGI-S scores. However, none of the monthly associations remained significant after Bonferroni correction.

When comparing participants' average questionnaire scores and actigraphy features collected over the study period (ranging from 1 to several months, depending on each participant's enrollment date), Spearman correlation analyses showed that participants with a later average sleep onset had a higher PHQ-9 scores average (ρ=0.90; *P*=.04). Participants with more average daily time spent in sedentary activity had higher PHQ-9 score (ρ=0.90; *P*=.04) and GAD-7 score averages (ρ=1.00; *P*<.001); however, after Bonferroni correction, only the GAD-7 association remained significant (*P*<.001), though this unusually strong correlation should be interpreted with caution given the small sample size (Figure 4).

XSL•FO
RenderX

Figure 4. Significant interindividual associations between actigraphy features and questionnaire scores across time scales. Repeated measures correlations were conducted at the daily, weekly, and monthly time scales; Spearman correlations were used for overall averages across the full study duration. CGI-S: Clinical Global Impression–Severity; GAD-7: 7-item General Anxiety Disorder; PHQ-9: 9-item Patient Health Questionnaire; SANS: Scale for the Assessment of Negative Symptoms; SAPS: Scale for the Assessment of Positive Symptoms.

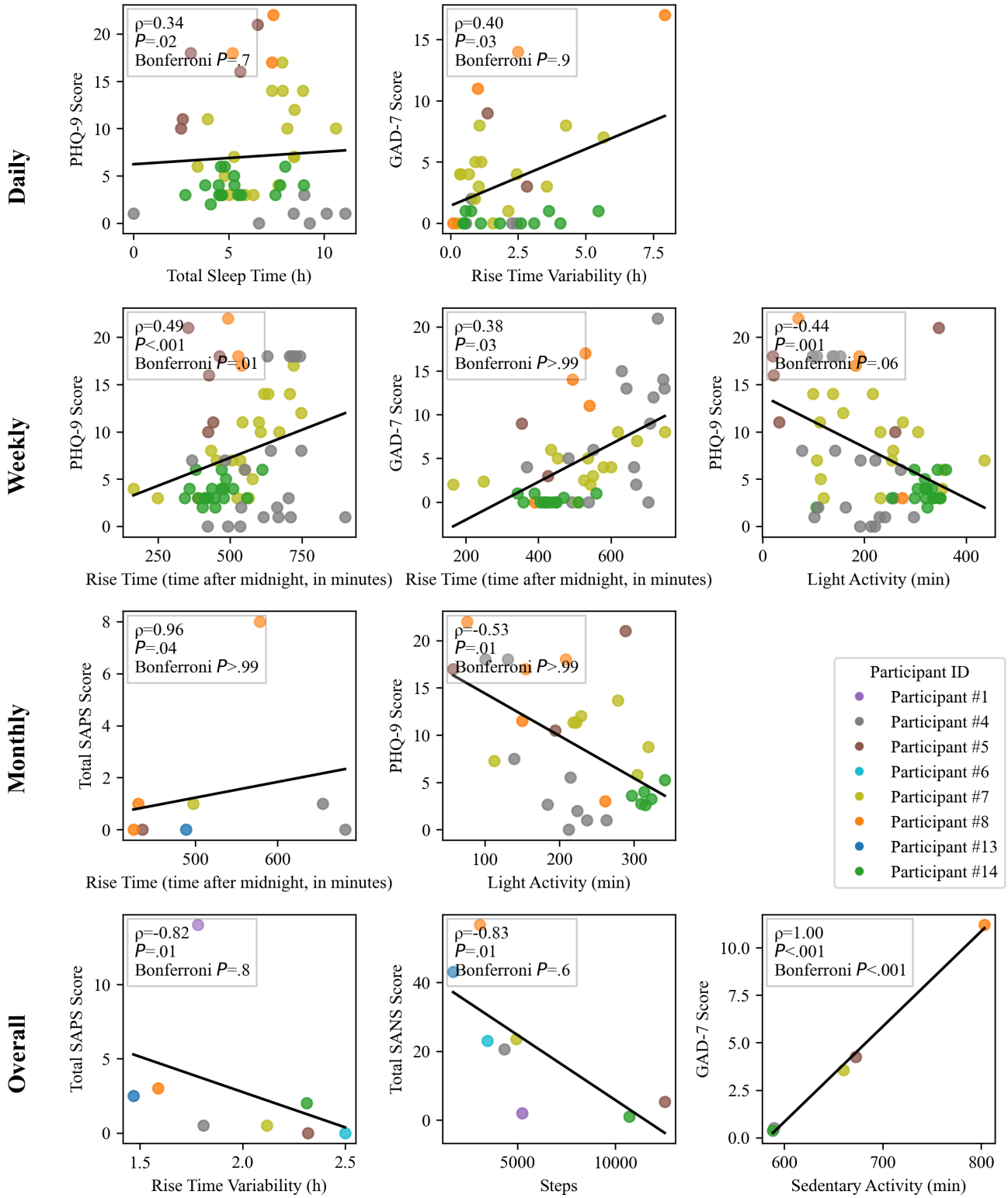

Sensitivity analyses using stricter validity thresholds (≥3 or ≥4 valid days per week, and ≥10 or ≥15 valid days per month) were conducted to assess the robustness of our findings and are available in Multimedia Appendix 1. These analyses showed similar patterns of associations, though with reduced statistical significance, as expected with decreased data availability and statistical power.

In summary, findings from interparticipant analyses aligned with intraparticipant results, showing consistent trends where symptom improvement was linked to earlier rise times, longer sleep duration, and increased time spent doing light physical activity, while symptom worsening was associated with increased sedentary behavior and delayed sleep onset. These

XSL•FO
RenderX

associations were stronger at the weekly level, with some remaining significant after correction for multiple comparisons.

Discussion

This case series, focused on the feasibility of the use of actigraphy in a realistic clinical setting, explored the relationship between sleep and activity patterns and psychiatric symptom severity across multiple diagnostic categories and time scales. Despite our small and heterogeneous sample, several meaningful intra- and interindividual associations emerged. Notably, sleep timing (particularly rise time) showed consistent associations with mood scores at the individual level, but also at the group level and at different time scales. While these results align with prior evidence linking delayed sleep-wake cycles to mood disorders and psychosis, our study is novel in showing these associations within a realistic transdiagnostic population and highlighting their persistence across different temporal scales [47,48]. This holds clinical significance, as it underscores the potential of sleep timing as a modifiable biomarker and intervention target for improving mood symptoms across a range of psychiatric disorders.

At the individual level, higher depressive and anxious symptoms were consistently associated with delayed circadian rhythm, observed at both daily and weekly time scales. For participant 5, who we know clinically to become more agitated when depressed, the presence of worse depressive symptoms during weeks of increased moderate physical activity likely reflects this pattern. This highlights the importance of accounting for individual clinical context when interpreting behavioral data. Interindividual analyses echoed these findings, showing that weeks with more severe depressive and anxious symptoms were also characterized by later rise times. The persistence of these associations after correction in this small, heterogeneous sample strengthens the case for delayed sleep-wake phase (ie, a regular sleep schedule that is considerably later than the conventional or desired time) as a transdiagnostic marker of psychiatric symptom burden. This characteristic has been previously linked with worse depressive symptoms in young adults [49] and worse outcomes in patients at clinical high risk of psychosis, as well as in those with early psychosis or schizophrenia [48,50], but we provide new evidence that these significant changes may occur across temporal scales and diagnostic groups [51,52]. We also found that during weeks and months of increased physical activity, participants reported less severe depressive and anxiety symptoms, consistent with findings from previous studies [53].

Same-day analyses, although not surviving correction for multiple comparisons, showed that shorter total sleep duration the night before, earlier rise time, and a rise time that has less variability from the previous day are each associated with lower same-day symptom severity. The unexpected association between total shorter sleep duration the preceding night and lower same-day symptom severity should be interpreted with caution, as the relationship was weak and largely driven by a small number of participants (notably participants 5 and 7). The sleep timing findings further support prior evidence that circadian phase is a key predictor of same-day mood episodes [26]. These trends point toward possible short-term responsiveness of mood and anxiety symptoms to daily behavioral patterns.

Additionally, when comparing participants' data across the entire study period (ie, averaged across all months), those with later average sleep onset and more sedentary behavior tended to report worse symptoms, while those with higher average daily step counts reported fewer negative psychotic symptoms. Due to very high compliance with wearing the GeneActiv sensor (97.27% of recorded days of actigraphy showed no nonwear time), we did not find any significant associations between non-compliance and questionnaire scores.

Despite the promising findings described above, several limitations must be acknowledged. The small sample, as well as missing data due to sensor malfunctions and variable adherence to actigraphy and questionnaire protocols, may have limited statistical power, introduced bias, and reduced sensitivity to detect certain effects. Furthermore, it is important to note that inferring temporal causality from these observations remains challenging; incorporating ecological momentary assessments data could provide a more precise temporal resolution, thereby improving the ability to disentangle the directionality of the relationship between sleep timing and symptom trajectory [54]. In addition, the use of GENEActiv devices requires physical docking to download data, meaning that data are not available in real time. As such, real-time monitoring of adverse events or symptom burden (eg, hospitalizations) was not possible in this study. Alternative wearable devices with real-time data transmission capabilities may be better suited for future studies aiming to evaluate actigraphy as a biomarker for timely clinical intervention. However, the actigraphy data were able to be processed and included in return-of-results reports to clinicians and patients on a monthly basis, which was preliminarily found to be clinically useful; the process of developing these reports will be discussed elsewhere. Nonetheless, the consistency of our results across time scales and methods highlights sleep timing and physical activity as a promising behavioral biomarker in real-world psychiatric populations.

The preliminary findings from the DeeP-DD study reveal complex associations between sleep features and clinical symptoms across various diagnostic groups, time scales, and data collection methods within a realistic clinical population. Earlier sleep and wake times, along with higher physical activity levels, were consistently associated with better clinical outcomes, including lower anxiety and depressive symptoms, both within and between individuals and across multiple time scales. The weekly time scale was particularly interesting, as multiple associations involving both sleep timing and physical activity were significant across participants. Although some associations did not remain significant after correction for multiple comparisons, the consistent patterns observed at both the inter- and intraindividual levels align with prior literature and offer novel insights into which actigraphy metrics, for different durations, may serve as digital biomarkers of monitoring and prediction or psychiatric symptom trajectories in a transdiagnostic population. However, given the small sample size, these conclusions should be interpreted with caution. The next steps in our project involve further exploring these relationships using larger datasets to assess their

XSL•FO
RenderX

consistency and investigate the longitudinal dynamics between sleep and symptom changes. Additionally, we aim to incorporate more sophisticated approaches such as Fourier transformations at an intraindividual level to capture rhythmicity and periodic patterns in sleep and activity data, which may reveal subtle disruptions in circadian cycles linked to symptom fluctuations [55]. These methods could help identify individual-specific signatures, ultimately informing personalized interventions.

Acknowledgments

The authors acknowledge financial support from the Mach-Gaensslen Foundation of Canada, the Dr Clarke K McLeod Memorial Scholarship, the Saputo Foundation, the Emerging Challenges Modelling Project (an initiative led by the Centre de Recherches Mathématiques in partnership with GERAD and UNIQUE, and funded by the Quebec Research Fund), the Strategia Program at the Centre de Recherches Mathématiques, the McGill Computational and Data Systems Initiative, the Douglas Research Center, the Fonds de recherche du Québec – Nature et technologies (STRATÉGIA grant), and the Fonds de Recherche du Québec - Santé (FRQS; Junior 1 Grant).

DB is supported by a NARSAD Young Investigator Award. LP's research is supported by Monique H. Bourgeois Chair in Developmental Disorders, the Graham Boeckh Foundation, the Mirella and Lino Saputo Foundation, and a Wellcome Trust Discretionary Grant (226168/Z/22/Z to Dr Iris Sommer and LP). He receives a salary award from the Fonds de recherche du Québec-Santé (FRQS: 366934).

Funding

This research protocol was funded by the Douglas Research Center, FRQS Junior 1 Grant, McGill Computational and Data Systems Initiative, Strategia FRQ Program at the Centre de Recherches Mathématiques, Healthy Brains, Healthy Lives grant from McGill, Saputo Foundation, and Brain and Behavior Research Foundation Young Investigator Grant.

Disclaimer

We occasionally used the generative artificial intelligence tool ChatGPT (OpenAI) to assist with grammar and phrasing during manuscript preparation.

Data Availability

Data available on reasonable request from the authors and subject to review board approval.

Authors' Contributions

Conceptualization: DB (lead), DH (equal)
Data curation: DH (lead), DP (equal), YB (equal), TZ (equal)
Formal analysis: DB (supporting), DH (lead), YB (supporting)
Funding acquisition: DB (lead), LP (equal)
Methodology: DB (lead), DH (supporting), SJ (supporting)
Supervision: DB
Validation: YB
Visualization: DH
Writing – original draft: DB (equal), DH (lead)
Writing – review & editing: DB (equal), DH (lead), SJ (supporting), DP (supporting), LP (supporting), TZ (supporting), YB (supporting)

Conflicts of Interest

DB is a founder and shareholder of Aifred Health, a digital mental health company. Aifred Health was not involved in this research. LP reports personal fees for serving as chief editor from the *Canadian Medical Association Journal*, speaker honorarium from Janssen Canada and Otsuka Canada, SPMM Course Limited, UK; book royalties from Oxford University Press; and investigator-initiated educational grants from Otsuka Canada outside the submitted work in the last 5 years.

Multimedia Appendix 1
Clinical and sociodemographic characteristics, and figures showing significant associations between actigraphy features and clinical questionnaire scores across multiple time scales, at both the intra- and interindividual level.
[PDF File, 1295 KB - formative_v9i1e81107_app1.pdf]

References

XSL•FO
RenderX

1. Fava GA, Kellner R. Staging: a neglected dimension in psychiatric classification. Acta Psychiatr Scand 1993 Apr;87(4):225-230. [doi: 10.1111/j.1600-0447.1993.tb03362.x] [Medline: 8488741]
2. Muneer A. Staging models in bipolar disorder: a systematic review of the literature. Clin Psychopharmacol Neurosci 2016 May 31;14(2):117-130. [doi: 10.9758/cpn.2016.14.2.117] [Medline: 27121423]
3. Compton MT, Bernardini F, Attademo L, et al. Risk prediction models in psychiatry: toward a new frontier for the prevention of mental illnesses. Psychiatrist.com. URL: https://www.psychiatrist.com/jcp/risk-prediction-models-in-psychiatry/ [accessed 2025-05-11]
4. Horowitz MA, Macaulay A, Taylor D. Limitations in research on maintenance treatment for individuals with schizophrenia. JAMA Psychiatry 2022 Jan 1;79(1):83-85. [doi: 10.1001/jamapsychiatry.2021.3400] [Medline: 34817553]
5. Coley RY, Boggs JM, Beck A, Hartzler AL, Simon GE. Defining success in measurement-based care for depression: a comparison of common metrics. Psychiatr Serv 2020 Apr 1;71(4):312-318. [doi: 10.1176/appi.ps.201900295]
6. Trivedi MH, Rush AJ, Wisniewski SR, et al. Evaluation of outcomes with citalopram for depression using measurement-based care in STAR*D: implications for clinical practice. Am J Psychiatry 2006 Jan;163(1):28-40. [doi: 10.1176/appi.ajp.163.1.28] [Medline: 16390886]
7. Guo T, Xiang YT, Xiao L, et al. Measurement-based care versus standard care for major depression: a randomized controlled trial with blind raters. Am J Psychiatry 2015 Oct;172(10):1004-1013. [doi: 10.1176/appi.ajp.2015.14050652] [Medline: 26315978]
8. Benrimoh D, Armstrong C, Mehltretter J, et al. Development and validation of a deep-learning model for differential treatment benefit prediction for adults with major depressive disorder deployed in the artificial intelligence in depression medication enhancement (AIDME) study. arXiv. Preprint posted online on Jun 7, 2024. [doi: 10.48550/arXiv.2406.04993]
9. Elliott CS, Fiszdon JM. Comparison of self-report and performance-based measures of everyday functioning in individuals with schizophrenia: implications for measure selection. Cogn Neuropsychiatry 2014;19(6):485-494. [doi: 10.1080/13546805.2014.922062] [Medline: 24901357]
10. Mote J, Fulford D. Ecological momentary assessment of everyday social experiences of people with schizophrenia: a systematic review. Schizophr Res 2020 Feb;216:56-68. [doi: 10.1016/j.schres.2019.10.021]
11. Abel DB, Minor KS. Social functioning in schizophrenia: comparing laboratory-based assessment with real-world measures. J Psychiatr Res 2021 Jun;138:500-506. [doi: 10.1016/j.jpsychires.2021.04.039] [Medline: 33971484]
12. Xiao L, Qi H, Zheng W, et al. The effectiveness of enhanced evidence-based care for depressive disorders: a meta-analysis of randomized controlled trials. Transl Psychiatry 2021 Oct 16;11(1):531. [doi: 10.1038/s41398-021-01638-7] [Medline: 34657142]
13. Zimmerman M, Martinez J, Attiullah N, Friedman M, Toba C, Boerescu DA. Symptom differences between depressed outpatients who are in remission according to the Hamilton Depression Rating Scale who do and do not consider themselves to be in remission. J Affect Disord 2012 Dec 15;142(1-3):77-81. [doi: 10.1016/j.jad.2012.03.044] [Medline: 22980402]
14. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613. [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
15. Levis B, Benedetti A, Thombs BD, DEPRESSion Screening Data (DEPRESSD) Collaboration. Accuracy of Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: individual participant data meta-analysis. BMJ 2019 Apr 9;365:l1476. [doi: 10.1136/bmj.l1476] [Medline: 30967483]
16. Barron DS. Decision models and technology in psychiatry. Biol Psychiatry 2021 Aug 15;90(4):208-211. [doi: 10.1016/j.biopsych.2021.06.012] [Medline: 34325803]
17. Benoit J, Onyeaka H, Keshavan M, Torous J. Systematic review of digital phenotyping and machine learning in psychosis spectrum illnesses. Harv Rev Psychiatry 2020;28(5):296-304. [doi: 10.1097/HRP.0000000000000268] [Medline: 32796192]
18. Spinazze P, Rykov Y, Bottle A, Car J. Digital phenotyping for assessment and prediction of mental health outcomes: a scoping review protocol. BMJ Open 2019 Dec;9(12):e032255. [doi: 10.1136/bmjopen-2019-032255]
19. Mendes JPM, Moura IR, Van de Ven P, et al. Sensing apps and public data sets for digital phenotyping of mental health: systematic review. J Med Internet Res 2022 Feb 17;24(2):e28735. [doi: 10.2196/28735] [Medline: 35175202]
20. Dunster GP, Swendsen J, Merikangas KR. Real-time mobile monitoring of bipolar disorder: a review of evidence and future directions. Neuropsychopharmacology 2021 Jan;46(1):197-208. [doi: 10.1038/s41386-020-00830-5] [Medline: 32919408]
21. Plante DT, Winkelman JW. Sleep disturbance in bipolar disorder: therapeutic implications. Am J Psychiatry 2008 Jul;165(7):830-843. [doi: 10.1176/appi.ajp.2008.08010077] [Medline: 18483132]
22. Girschik J, Fritschi L, Heyworth J, Waters F. Validation of self-reported sleep against actigraphy. J Epidemiol 2012;22(5):462-468. [doi: 10.2188/jea.je20120012] [Medline: 22850546]
23. Ferrarelli F. Sleep disturbance in schizophrenia spectrum disorders: more than just a symptom? Int Clin Psychopharmacol 2023 May 1;38(3):187-188. [doi: 10.1097/YIC.0000000000000467] [Medline: 36866853]
24. Nordholm D, Jensen MA, Glenthøj LB, et al. Sleep disturbances and the association with attenuated psychotic symptoms in individuals at ultra high-risk of psychosis. J Psychiatr Res 2023 Feb;158:143-149. [doi: 10.1016/j.jpsychires.2022.12.041]
25. Reich N, Delavari F, Schneider M, Thillainathan N, Eliez S, Sandini C. Multivariate patterns of disrupted sleep longitudinally predict affective vulnerability to psychosis in 22q11.2 Deletion Syndrome. Psychiatry Res 2023 Jul;325:115230. [doi: 10.1016/j.psychres.2023.115230] [Medline: 37201254]

XSL•FO
RenderX

26. Pieters LE, Deenik J, Hoogendoorn AW, van Someren EJW, van Harten PN. Sleep and physical activity patterns in relation to daily-life symptoms in psychosis: an actigraphy and experience sampling study. Psychiatry Res 2025 Feb;344:116320. [doi: 10.1016/j.psychres.2024.116320] [Medline: 39673966]
27. Lim D, Jeong J, Song YM, et al. Accurately predicting mood episodes in mood disorder patients using wearable sleep and circadian rhythm features. NPJ Digit Med 2024 Nov 18;7(1):1-13. [doi: 10.1038/s41746-024-01333-z]
28. Burton C, McKinstry B, Szentagotai Tătar A, Serrano-Blanco A, Pagliari C, Wolters M. Activity monitoring in patients with depression: a systematic review. J Affect Disord 2013 Feb 15;145(1):21-28. [doi: 10.1016/j.jad.2012.07.001] [Medline: 22868056]
29. Abad VC, Guilleminault C. Sleep and psychiatry. Dialogues Clin Neurosci 2005 Dec 31;7(4):291-303. [doi: 10.31887/DCNS.2005.7.4/vabad]
30. de Leeuw M, Verhoeve SI, van der Wee NJA, van Hemert AM, Vreugdenhil E, Coomans CP. The role of the circadian system in the etiology of depression. Neurosci Biobehav Rev 2023 Oct;153:105383. [doi: 10.1016/j.neubiorev.2023.105383]
31. Jonasdottir SS, Minor K, Lehmann S. Gender differences in nighttime sleep patterns and variability across the adult lifespan: a global-scale wearables study. Sleep 2021 Feb 12;44(2):zsaa169. [doi: 10.1093/sleep/zsaa169] [Medline: 32886772]
32. van Hees VT, Sabia S, Jones SE, et al. Estimating sleep parameters using an accelerometer without sleep diary. Sci Rep 2018 Aug 28;8(1):12975. [doi: 10.1038/s41598-018-31266-z]
33. te Lindert BHW, Van Someren EJW. Sleep estimates using microelectromechanical systems (MEMS). Sleep 2013 May 1;36(5):781-789. [doi: 10.5665/sleep.2648]
34. Krane-Gartiser K, Asheim A, Fasmer OB, Morken G, Vaaler AE, Scott J. Actigraphy as an objective intra-individual marker of activity patterns in acute-phase bipolar disorder: a case series. Int J Bipolar Disord 2018 Mar 7;6(1):8. [doi: 10.1186/s40345-017-0115-3] [Medline: 29511876]
35. Tazawa Y, Wada M, Mitsukura Y, et al. Actigraphy for evaluation of mood disorders: a systematic review and meta-analysis. J Affect Disord 2019 Jun 15;253:257-269. [doi: 10.1016/j.jad.2019.04.087] [Medline: 31060012]
36. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381. [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
37. Dodge L. GENEActiv support 11. Activinsights. URL: https://activinsights.com/resources/geneactiv-support-1-1/ [accessed 2025-06-22]
38. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder. Arch Intern Med 2006 May 22;166(10):1092. [doi: 10.1001/archinte.166.10.1092]
39. Busner J, Targum SD. The clinical global impressions scale: applying a research tool in clinical practice. Psychiatry (Edgmont) 2007 Jul;4(7). [Medline: 20526405]
40. Andreasen NC. Scale for the assessment of positive symptoms. APA PsychTests. 2018. URL: https://psycnet.apa.org/doiLanding?doi=10.1037%2Ft48377-000 [accessed 2025-11-12]
41. Andreasen NC. Scale for the assessment of negative symptoms. APA PsychTests. 2014. URL: https://psycnet.apa.org/doiLanding?doi=10.1037%2Ft12696-000 [accessed 2025-11-12]
42. Selbie HR. Markdown training. Activinsights. URL: https://activinsights.com/research-study-services/r-markdown-training/ [accessed 2025-06-22]
43. Di J, Demanuele C, Kettermann A, et al. Considerations to address missing data when deriving clinical trial endpoints from digital health technologies. Contemp Clin Trials 2022 Feb;113:106661. [doi: 10.1016/j.cct.2021.106661] [Medline: 34954098]
44. Bakdash JZ, Marusich LR. Repeated measures correlation. Front Psychol 2017;8:456. [doi: 10.3389/fpsyg.2017.00456] [Medline: 28439244]
45. Vallat R. Pingouin: statistics in Python. JOSS 2018 Nov 19;3(31):1026. [doi: 10.21105/joss.01026]
46. Spearman's rank correlation coefficient Rs and probability (p) value calculator. GeographyFieldwork.com. URL: https://geographyfieldwork.com/SpearmansRankCalculator.html [accessed 2025-07-06]
47. Dollish HK, Tsyglakova M, McClung CA. Circadian rhythms and mood disorders: time to see the light. Neuron 2024 Jan 3;112(1):25-40. [doi: 10.1016/j.neuron.2023.09.023] [Medline: 37858331]
48. Meyer N, Lok R, Schmidt C, et al. The sleep–circadian interface: a window into mental disorders. Proc Natl Acad Sci USA 2024 Feb 27;121(9):e2214756121. [doi: 10.1073/pnas.2214756121]
49. Crouse JJ, Carpenter JS, Song YJC, et al. Circadian rhythm sleep-wake disturbances and depression in young people: implications for prevention and early intervention. Lancet Psychiatry 2021 Sep;8(9):813-823. [doi: 10.1016/S2215-0366(21)00034-1] [Medline: 34419186]
50. Zanini M, Castro J, Coelho FM, et al. Do sleep abnormalities and misaligned sleep/circadian rhythm patterns represent early clinical characteristics for developing psychosis in high risk populations? Neurosci Biobehav Rev 2013 Dec;37(10 Pt 2):2631-2637. [doi: 10.1016/j.neubiorev.2013.08.012] [Medline: 24096189]
51. Lunsford-Avery JR, Gonçalves BDS, Brietzke E, et al. Adolescents at clinical-high risk for psychosis: circadian rhythm disturbances predict worsened prognosis at 1-year follow-up. Schizophr Res 2017 Nov;189:37-42. [doi: 10.1016/j.schres.2017.01.051] [Medline: 28169087]

XSL•FO
RenderX

52. Wong TR, Hickie IB, Carpenter JS, et al. Dynamic modelling of chronotype and hypo/manic and depressive symptoms in young people with emerging mental disorders. Chronobiol Int 2023 Jun 3;40(6):699-709. [doi: 10.1080/07420528.2023.2203241] [Medline: 37132360]
53. Peluso MAM, Guerra de Andrade LHS. Physical activity and mental health: the association between exercise and mood. Clinics (Sao Paulo) 2005 Feb;60(1):61-70. [doi: 10.1590/s1807-59322005000100012] [Medline: 15838583]
54. De la Barrera U, Arrigoni F, Monserrat C, Montoya-Castilla I, Gil-Gómez JA. Using ecological momentary assessment and machine learning techniques to predict depressive symptoms in emerging adults. Psychiatry Res 2024 Feb;332:115710. [doi: 10.1016/j.psychres.2023.115710] [Medline: 38194800]
55. Li X, Kane M, Zhang Y, et al. Circadian rhythm analysis using wearable device data: novel penalized machine learning approach. J Med Internet Res 2021 Oct 14;23(10):e18403. [doi: 10.2196/18403] [Medline: 34647895]

Abbreviations

CGI-S: Clinical Global Impression–Severity
DeeP-DD: Deep Phenotyping and Digitalization at Douglas
GAD-7: 7-item General Anxiety Disorder
PHQ-9: 9-item Patient Health Questionnaire
SANS: Scale for the Assessment of Negative Symptoms
SAPS: Scale for the Assessment of Positive Symptoms

Edited by A Stone, A Mavragani; submitted 28.07.25; peer-reviewed by S Avey; revised version received 22.10.25; accepted 23.10.25; published 14.11.25.

Please cite as:
Hamitouche D, Zamorano T, Barkat Y, Parekh D, Palaniyappan L, Jalali S, Benrimoh D
Sleep and Activity Patterns as Transdiagnostic Behavioral Biomarkers in Psychiatry: Longitudinal Observational Study From the DeeP-DD Study
JMIR Form Res 2025;9:e81107
URL: https://formative.jmir.org/2025/1/e81107
doi:10.2196/81107

© Dylan Hamitouche, Tihare Zamorano, Youcef Barkat, Deven Parekh, Lena Palaniyappan, Sara Jalali, David Benrimoh. Originally published in JMIR Formative Research (https://formative.jmir.org), 14.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Formative Evaluation of an HIV Prevention App tailored for Latino Men Who Have Sex With Men: Acceptability and Usability Study

Valeria D Cantos[1], MD; Humberto Posada-Orozco[1], MPH; Isabella Batina[1], MPH, MD; Natalie Sanchez[2], BA; Eric Rangel[3]; Patrick S Sullivan[4], DVM, PhD; Andres Camacho-Gonzalez[5], MSc, MD; Aaron J Siegler[6], MHS, PhD

[1]Division of Infectious Diseases, Department of Medicine, Emory University School of Medicine, 68 Armstrong St. SE. Suite 210, Atlanta, GA, United States

[2]Duke University School of Medicine, Durham, NC, United States

[3]Latino Linq, Atlanta, GA, United States

[4]Department of Epidemiology, Rollins School of Public Health, Emory University, Atlanta, GA, United States

[5]Division of Pediatric Infectious Diseases, Department of Pediatrics, Emory University School of Medicine, Atlanta, GA, United States

[6]Department of Behavioral, Social, and Health Education Sciences, Rollins School of Public Health, Emory University, Atlanta, GA, United States

Corresponding Author:
Valeria D Cantos, MD
Division of Infectious Diseases, Department of Medicine, Emory University School of Medicine, 68 Armstrong St. SE. Suite 210, Atlanta, GA, United States

Abstract

Background: HIV incidence is increasing among Latino gay, bisexual, and other men who have sex with men (MSM) in the Atlanta metropolitan area. Mobile phone apps represent an innovative tool to promote pre-exposure prophylaxis (PrEP) use, HIV testing, and condom use.

Objective: This study aimed to assess the acceptability and usability of Saludfindr, an Android-based HIV prevention app tailored to the needs of Latino MSM in the Atlanta area.

Methods: We recruited adult Latino MSM to interact with the app for 4 months. Saludfindr included initial and periodic health assessments; provision of suggestions regarding PrEP, HIV testing, and condom use; in-app product ordering; customized motivational messages; a customized sexual health clinic list; and a “Contact Us” button. To assess acceptability, we measured use of each app feature, PrEP and HIV testing uptake, and participant ratings of the app’s usefulness. We assessed usability using the System Usability Scale.

Results: We enrolled 31 participants; the median age was 27 (IQR 24.5-32) years, 97% (30/31) were cisgender men, 81% (25/31) identified as MSM, and 61% (19/31) used the app in Spanish. All participants completed the initial health screening, with 84% (26/31) and 77% (24/31) completing the 2- and 4-month health screenings, respectively. Of all participants, 52% (16/31) and 23% (7/31) ordered condoms and home HIV tests through the app at least once, respectively. During the study period, 71% (22/31) of the participants got tested for HIV, of whom 68% (15/22) accessed it through clinic-based HIV testing. Of the participants not on PrEP at baseline, 41% (7/17) initiated PrEP during the study, and all of them did so at one of the clinics listed on the app. Saludfindr reached a System Usability Scale score of 74.5/100 (excellent).

Conclusions: Saludfindr was highly acceptable and usable among Latino MSM participants in the Atlanta area. In-app assistance to access PrEP and clinic-based HIV testing referrals was well received. Further efforts are needed to increase users’ self-efficacy with home HIV self-testing.

(JMIR Form Res 2025;9:e74208) doi:10.2196/74208

KEYWORDS

mobile health; mHealth; pre-exposure prophylaxis; PrEP; acceptability; usability; pilot; Latino; HIV

Introduction

Background

Between 2018 and 2022, HIV incidence in the United States decreased 12% due at least in part to an overall increase in pre-exposure prophylaxis (PrEP) use [1]. These advancements toward ending the HIV epidemic have not reached all racial and ethnic groups equally. Nationally, HIV incidence among Latino gay, bisexual, and other men who have sex with men (MSM) remained unchanged between 2018 and 2022 while decreasing 16% and 20% among Black and White MSM, respectively [2]. In 2022, Latino MSM ranked first in number of new HIV

XSL•FO
RenderX

diagnoses among MSM, accounting for 36% of all new diagnoses among MSM [2].

The state of Georgia has the third highest rate of HIV diagnosis among Latina, Latino, or Latinx individuals in the United States, with 80% of cases occurring in the Atlanta metropolitan area [3]. Between 2010 and 2019, HIV diagnosis rates among Latino men in Georgia, 80% of them MSM, increased by 13% while decreasing among Black and White men (−12.3% and −21%, respectively) [4].

Despite comparable PrEP awareness among White, Black, and Latino MSM [2], significant inequities in PrEP uptake are observed [5]. The PrEP-to-need ratio (PnR), a measurement assessing PrEP use compared to HIV diagnoses in the same period for the same population or region, is widely used as an estimation of PrEP equity, with a lower PnR representing a higher unmet need [6]. In 2021, the PnR for Latina, Latino, or Latinx individuals was over 4 times lower compared to the PnR for White individuals (6.0 vs 25.9) [5].

Latino MSM face significant barriers to PrEP uptake in metropolitan Atlanta [7-9]. Some of these barriers, such as low perceived PrEP eligibility [7,10-12], intersectional stigma [7,13-15], and PrEP misconceptions [7,16], are also common among other racial and ethnic groups in the United States. However, many PrEP uptake barriers among Latino MSM are related to its limited accessibility [7-9]. These factors include scarce availability of reliable PrEP information in Spanish, limited self-efficacy on how to access PrEP care, perceived or real PrEP costs, perceived ineligibility for free or low-cost PrEP care due to having an undocumented migratory status, competing job obligations, and a limited pool of available clinics that are equipped to provide language- and culturally concordant PrEP care to Latino MSM [7-9].

Mobile health (mHealth) technology represents an innovative set of tools to increase PrEP uptake in different populations [17-22]. HealthMindr [19] is an evidence-based mobile app that has been added to the Centers for Disease Control and Prevention's compendium of evidence-based interventions [23]. It features in-app health screenings that direct users to appropriate services: HIV testing and PrEP service locator; mail delivery of condoms, lubricant, and home HIV and sexually transmitted infection tests; and customizable HIV testing and PrEP care maintenance plans [19]. To date, no comprehensive HIV prevention app has been customized to address the PrEP accessibility barriers faced by Latino MSM, representing an important research gap.

Objectives

The overall objective of this study was to develop and pilot-test Saludfindr, a comprehensive HIV prevention app for Latino MSM, with the ultimate goal of increasing PrEP uptake in this group. Using HealthMindr as a template app and following a modified version of the assessment, decision, adaptation, production, topical experts–integration, training, and testing framework [24], we first conducted formative work with Latino MSM to guide the adaptation process to produce Saludfindr [7]. In this paper, we present the results of the 4-month pilot study of Saludfindr, which assessed the app's acceptability and usability.

Methods

Saludfindr Development

Following the assessment, decision, adaptation, production, topical experts–integration, training, and testing framework [24], we first decided to use HealthMindr as a template platform to develop Saludfindr as it focused on MSM and the outcomes of interest included PrEP initiation. We then conducted a formative assessment to guide the adaptation process. We conducted 13 in-depth interviews with Latino MSM, during which we identified the main barriers to and facilitators of PrEP uptake in this population, gauged opinions on existing HealthMindr app features, and proposed de novo Saludfindr features [7]. In addition to general PrEP uptake barriers such as low perceived PrEP eligibility, intersectional stigma, cost concerns, and PrEP misconceptions, we identified barriers that predominantly impact Latino MSM's access to PrEP. These include scarcity of local PrEP clinics that can provide culturally and language-concordant care, limited availability of information about PrEP access in Spanish, distrust of peers as a reliable source of PrEP-related information, perceived ineligibility for low-cost PrEP services stemming from having an undocumented immigration status, anti-immigrant policies and rhetoric, and competing work obligations limiting attendance to PrEP clinic visits. As PrEP uptake facilitators, we identified that health care providers represented a trusted source of PrEP information, especially if they were familiar with prescribing PrEP and were bilingual Latino individuals themselves [7].

On the basis of these data, we decided to keep the following HealthMindr features: periodic health assessments to determine PrEP eligibility, in-app home HIV test and condom ordering to facilitate access, and frequently asked questions to increase information. We added 3 de novo Saludfindr features: customized motivational messages to incentivize users to adopt HIV prevention strategies, a customized sexual health clinic list to facilitate access, and a "Contact Us" button to provide additional information that users may need. Of note, all clinics included in the list were known to offer free or low-cost services to Latino MSM regardless of immigration status and had Spanish-language service availability. The process of determining which clinics to include is described elsewhere [7].

All app content was developed in English and Spanish. Members of our partner community-based organizations reviewed the content in both languages and provided feedback to optimize readability and cultural concordance. VDC, IB, HO, and NS developed all app content. The National Cancer Institute–designated Stephenson Cancer Center mHealth Shared Resource at the University of Oklahoma Health Sciences Center produced the app through the Insight mHealth platform.

App Features

Figure 1 shows a full list of app features and descriptions.

XSL•FO
RenderX

Figure 1. Saludfindr HIV prevention app features for Latino men who have sex with men. FAQ: frequently asked question; PReP: pre-exposure prophylaxis.

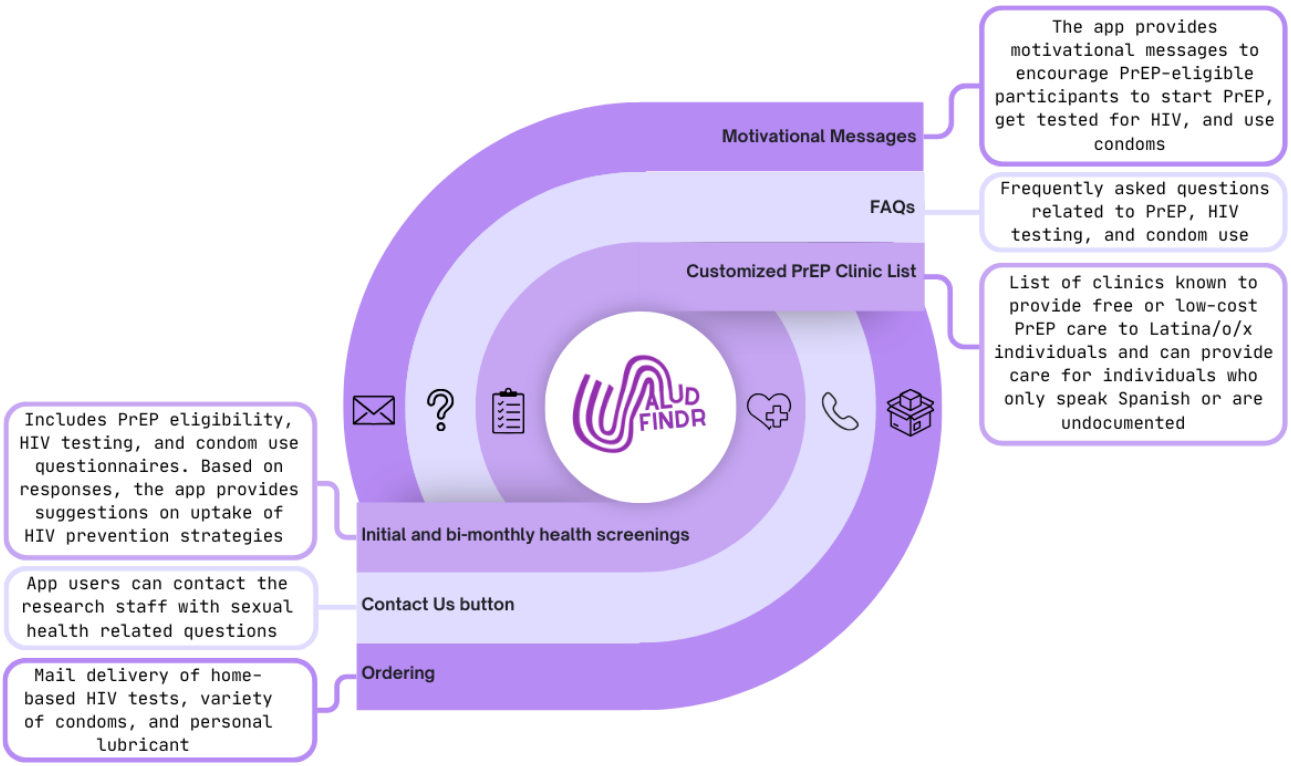

Pilot Study Overview

The purpose of this 4-month pilot study was to evaluate the acceptability and usability of Saludfindr among Latino MSM living in metropolitan Atlanta. Enrolled participants downloaded Saludfindr onto their Android cell phone or study-provided Android cell phone. Participants were asked the language they preferred to use the app in (English or Spanish) and interacted with it during the study period. They first completed a baseline sociodemographic survey and an initial health screening that included PrEP eligibility, HIV testing, and condom use patterns. On the basis of each individual's responses to the initial health assessment, the app provided suggestions on PrEP use and HIV testing frequency. It also sent periodic motivational messages to encourage PrEP uptake, HIV testing, and condom use. Participants then completed 2- and 4-month health screenings that assessed needs for interim PrEP initiation and HIV testing. The 2- and 4-month health screenings also assessed the use of app features to help participants access PrEP or HIV testing services, as well as the use of the home HIV test and condom ordering feature. This was followed by customized app-based suggestions and messages. At the end of the 4-month period, participants completed a survey to evaluate the app's acceptability and usability in assisting Latino MSM in adopting HIV prevention strategies such as PrEP uptake, HIV testing, and condom use. Throughout the study period, we gathered order tracking and app-based use data.

Study Population, Recruitment, and Enrollment

We recruited participants from October 2022 to April 2023 using in-person and web-based recruitment strategies. All recruitment materials were available in English and Spanish. We partnered with Latino LinQ, a local community-based organization that supports, advocates for, and provides direct services to the Latina, Latino, and Latinx lesbian, gay, bisexual, transgender, and queer communities living in metropolitan Atlanta. We visited bars and clubs popular among Latino MSM with the Latino LinQ staff, who had established HIV testing stations at these venues. At these locations, trained research staff provided information about the study to attendants and distributed physical study flyers. These included a QR code that directed interested individuals to a HIPAA (Health Insurance Portability and Accountability Act)–compliant document with further information about the study and a brief eligibility survey. For web-based recruitment, Latino LinQ posted institutional review board (IRB)–approved electronic flyers in their social media accounts with a link that directed individuals to the online materials as described previously. We also conducted a social media recruitment campaign using Meta Ads Manager. Advertisements were displayed on Facebook and Instagram for 4 months. Targeted parameters included individuals aged 18 years and older residing in the Atlanta metropolitan area. Advertisements consisted of static images with a call to action directing interested individuals to the aforementioned materials.

To be eligible for this study, individuals needed to meet the following inclusion criteria (all per self-report): (1) assigned male at birth, (2) age of 18 years or older, (3) identifying as Latino, (4) having had sex with cisgender men, (5) negative or unknown HIV status, and (6) residence in the Atlanta metropolitan area.

Trained research staff (HO and IB) reviewed all electronic eligibility forms from all sources (in-person events, Latino LinQ social media, and our Meta recruitment campaign). They contacted eligible individuals, verified their identity and contact information, provided in-depth information about the study,

XSL•FO
RenderX

and went over a verbal informed consent process with them over the phone. Consented participants received an email or SMS text message with instructions and a unique registration code to download the app either on their own Android cell phone or on a study-provided Android phone. Once they downloaded the app, participants completed the brief sociodemographic survey and initial health screening, which completed their enrollment into the study.

Ethical Considerations

This study was approved by the Emory University IRB (MOD002-IRB00097743). We obtained verbal informed consent over the phone for this study. The Emory University IRB waived the requirement to obtain a written documentation of consent as this research was considered to present no more than minimal risk of harm to participants and did not involve procedures for which written consent was required. Study participants were assigned a participant identification number, and all data was deidentified prior to analysis. Participants were compensated with US $50 for the completion of the baseline sociodemographic survey and initial health screening, US $50 for the completion of the 2-month health screening, and US $50 for the completion of the 4-month health screening and app evaluation survey.

Measures and Outcomes

In the baseline sociodemographic survey, we assessed the following: age in years, gender identity, sexual orientation, sexual partners' gender identities, region or country of birth, immigration status, and employment status. During the initial health screening, we measured participants' recent HIV testing history and their PrEP eligibility, which was defined as meeting at least one of the following criteria in the 6 months before enrollment: (1) engagement in condomless anal sex; (2) having had a bacterial sexually transmitted infection; (3) use of drugs such as poppers, cocaine, methamphetamines, or heroin; and (4) having had a partner living with HIV or with an unknown HIV status. During the 2- and 4-month health screenings, participants were asked whether they had started taking PrEP. Those who had started PrEP were asked whether they used the in-app customized clinic list to access PrEP care, whereas participants who had not started PrEP were asked about their reasons for not doing so, and their PrEP eligibility was reassessed. Similarly, participants were asked about interim HIV testing, test results, use of the in-app clinic list to access HIV testing services, and use of the at-home HIV test ordering app feature. Finally, we asked participants whether they ordered condoms through the app.

To evaluate the app, we assessed Saludfindr's acceptability and usability. For acceptability, we measured the use of each app feature (from health screenings, app-based metadata, and order tracking data) and surveyed participants about the app's acceptability in assisting Latino MSM in using HIV prevention strategies (PrEP uptake, HIV testing, and condom use). We also measured PrEP and HIV testing uptake via participant self-report through the health screenings. Usability was measured using the System Usability Scale (SUS), a validated scale that measures user ratings of product usability through 10 statements [25]. Its final score ranges from 0 to 100, with higher scores indicating higher usability. Finally, we measured uptake of HIV prevention strategies during the study period, including PrEP initiation and HIV testing.

As the main outcomes of this pilot study were acceptability and usability rather than efficacy in increasing PrEP uptake, HIV testing, or condom use, the sample size did not need to provide statistical power for detecting the app's effects on these outcomes. A sample size of 25 to 35 participants offers enough precision to judge acceptability and usability and gather information for planning a future clinical trial [26].

A priori, we defined the following thresholds to be met by Saludfindr: (1) a 50% or higher proportion completing the 2- and 4-month health screenings; (2) a 50% or higher proportion using the at-home HIV test and condom ordering feature; (3) a 50% or higher proportion using the in-app customized clinic list to access PrEP or HIV testing services among participants who were not on PrEP at baseline or who were due for HIV testing, respectively (defined as those who had never been tested for HIV or who had been tested over 6 months before study enrollment); and (4) SUS mean score of 60/100 (good) or higher. These cutoff points informed a set of go or no-go criteria based on which Saludfindr would (1) be ready for future research during a clinical trial (all thresholds met; go), (2) need further modifications (some thresholds met; expert consultation needed), or (3) not merit further research (no go).

Data Analysis

All analyses were conducted using the R software (R Foundation for Statistical Computing). Descriptive statistics summarized baseline characteristics, Saludfindr use, HIV prevention outcomes, perceived usefulness, and system usability. Continuous variables such as age and scores were summarized using medians and IQRs or means and SDs, as appropriate. Categorical variables were presented as frequencies and percentages. Data were analyzed from baseline sociodemographic data, in-app health assessments, the final survey, and app tracking data.

Results

Baseline Characteristics

We enrolled 31 Latino individuals assigned male at birth for the 4-month pilot study. The median age was 27 (IQR 24.5-32) years, 97% (30/31) were cisgender men, 81% (25/31) identified as MSM, and 19% (6/31) identified as bisexual. Approximately two-thirds (20/31, 65%) were born outside of the United States, 35% (11/31) either had an undocumented immigration status or were under humanitarian parole or asylum, and 61% (19/31) used the app in Spanish. Of all participants, 84% (26/31) had gotten tested for HIV in the previous 6 months, and 77% (24/31) were eligible for PrEP (Table 1).

XSL•FO
RenderX

Table . Baseline sociodemographic characteristics of the Saludfindr app pilot study participants (N=31).

Characteristic	Values
Age (y), median (IQR)	27 (24.5-32)
Current gender identity, n (%)	
Cisgender man	30 (97)
Sexual orientation, n (%)	
MSM[a]	25 (81)
Bisexual	6 (19)
HIV testing status, n (%)	
Within the previous 6 months	26 (84)
Over 6 months before or never tested	5 (16)
PrEP[b] eligibility criteria met (in the previous 6 months)[c], n (%)	
Condomless anal sex	23 (74)
Bacterial STI[d]	6 (19)
Drug use	12 (39)
Partner living with HIV or with an unknown HIV status	9 (29)
PrEP status, n (%)	
Eligible for PrEP	24 (77)
Currently on PrEP	14 (45)
Region or country of birth, n (%)	
United States	11 (35)
Mexico	8 (26)
Central America	3 (10)
South America	9 (29)
Immigration status, n (%)	
US citizen or permanent resident	17 (55)
Undocumented or on humanitarian parole or asylum	11 (35)
On visa or DACA[e]	3 (10)
Used the app in Spanish, n (%)	19 (61)

[a]MSM: men who have sex with men.

[b]PrEP: pre-exposure prophylaxis.

[c]PrEP eligibility was defined as meeting at least one of the following criteria in the previous 6 months: condomless anal sex, bacterial sexually transmitted infection, drug use, or having a partner living with HIV or with an unknown HIV status (participants who were found to be PrEP noneligible did not respond to the partner question and the drug use question, so we were unable to make a PrEP eligibility determination).

[d]STI: sexually transmitted infection.

[e]DACA: Deferred Action for Childhood Arrivals.

Acceptability of Saludfindr Features

All participants completed the initial health screening, and 84% (26/31) and 77% (24/31) completed the 2- and 4-month health screenings, respectively (Table 2). Of all participants, 52% (16/31) ordered condoms through the app at least once, and 62% (10/16) of this group placed repeat orders during the pilot study. While 23% (7/31) of the participants ordered home HIV tests through the app at least once, 29% (2/7) of them placed repeat orders. Finally, 16% (5/31) of participants used the "Contact Us" button to reach out to research staff with sexual health–related questions (Table 2).

XSL•FO
RenderX

Table . Acceptability of the Saludfindr app features.

Feature	Participants, n/N (%)
Health screenings[a]	
Initial health screening	31/31 (100)
Month 2 health screening	26/31 (84)
Final health screening	24/31 (77)
Condom orders[b]	
At least once	16/31 (52)
Repeat orders	10/16 (62)
Home HIV test orders[b]	
At least once	7/31 (23)
Repeat orders	2/7 (29)
"Contact Us" button use[a]	5/31 (16)

[a]Data collected from app use metadata.

[b]Data collected from product shipping tracking.

PrEP and HIV Testing Uptake

During the 4-month study period, 71% (22/31) of the participants got tested for HIV, 32% (7/22) of them using the in-app home HIV test ordering function and 68% (15/22) through clinic-based HIV testing. Of the 16% (5/31) of the participants who had never gotten tested for HIV or who had been tested over 6 months before study enrollment, 60% (3/5) used the in-app customized clinic list to find an HIV testing site. The remaining 40% (2/5) of the participants were lost to follow-up, so we could not document whether they got tested for HIV. At baseline, 55% (17/31) of the participants were not on PrEP. Of these, 41% (7/17) initiated PrEP during the study, and all of them did so at one of the clinics listed on the app (Table 3).

Table . Pre-exposure prophylaxis (PrEP) and HIV testing uptake among Saludfindr study participants.

Outcome	Participants, n/N (%)
HIV testing	22/31 (71)
Using in-app home HIV test ordering feature	7/22 (32)
Clinic-based testing	15/22 (68)
At clinic from in-app list	10/15 (67)
Positive HIV test	0 (0)
HIV testing among those who had never been tested or had been tested >6 months before at baseline	3/5 (60)
At one of the clinics listed on the app	3/3 (100)
PrEP initiation among those off PrEP at enrollment	7/17 (41)
At one of the clinics listed on the app	7/7 (100)

When we asked participants about Saludfindr's usefulness in assisting Latino MSM with the uptake of HIV prevention strategies, the proportion of people who rated it as "useful" or "very useful" was 92% (22/24) for PrEP initiation, 96% (23/24) for HIV testing, and 79% (19/24) for condom use (Table 4).

Table . Perceived usefulness of the Saludfindr app for uptake of HIV prevention strategies among Latino men who have sex with men (N=24).

HIV prevention strategy	Not useful at all, n (%)	Not useful, n (%)	Neutral, n (%)	Useful, n (%)	Very useful, n (%)
PrEP[a] uptake	1 (4)	0 (0)	1 (4)	4 (17)	18 (75)
HIV testing	0 (0)	1 (4)	0 (0)	5 (21)	18 (75)
Condom use	0 (0)	2 (8)	3 (12)	2 (8)	17 (71)

[a]PrEP: pre-exposure prophylaxis.

XSL•FO
RenderX

Saludfindr Usability

Overall, Saludfindr reached an SUS score of 74.5 (excellent; Table 5). The SUS statements that obtained the highest scores included "The app was easy to use" (4.21/5), "most people would learn to use this app very quickly" (4.25/5), and "I felt very confident using the app" (4.21/5). The 2 statements with the lowest scores were "The app was unnecessarily complex" (3.12/5) and "there was too much inconsistency in this app" (3.25/5).

Table . System Usability Scale scores for the Saludfindr app during the pilot study (total calculated score: 74.5/100).

Statement	Mean (SD)[a]	Absolute[b]
I would like to use this app frequently	3.79 (1.28)	3.79 (1.28)
The app was unnecessarily complex[c]	2.88 (1.48)	3.12 (1.48)
The app was easy to use	4.21 (1.02)	4.21 (1.02)
I would need support from a technical person to be able to use this app[c]	2.46 (1.38)	3.54 (1.38)
Various functions in the app were well integrated	3.88 (1.19)	3.88 (1.19)
There was too much inconsistency in this app[c]	2.75 (1.39)	3.25 (1.39)
Most people would learn to use this app very quickly	4.25 (1.03)	4.25 (1.03)
The app was very cumbersome to use[c]	2.33 (1.43)	3.67 (1.43)
I felt very confident using the app	4.21 (1.06)	4.21 (1.06)
I had to learn many things before I could get going with this app[c]	2.67 (1.34)	3.33 (1.34)

[a]Scoring based on a scale from 1="totally disagree" to 5="totally agree."

[b]Adjusts the scores of negative statements so that larger numbers are associated with positive statements.

[c]Negative statements.

Discussion

Principal Findings

In this pilot study assessing the acceptability and usability of an HIV prevention app tailored for Latino MSM living in metropolitan Atlanta, most enrolled participants were young, non–US born, primarily Spanish speakers, and eligible for PrEP. The intervention, Saludfindr, met or surpassed prespecified targets for acceptability and usability, including the percentage of health screenings completed, in-app condom ordering, use of in-app customized clinic lists to access PrEP services among participants off PrEP at baseline, and HIV testing among participants due for it. It also surpassed the predetermined SUS score. Saludfindr did not meet the target for use of in-app home HIV test ordering. This indicates that further modifications to this feature are needed to address potential user-related uptake barriers. Taken together, these findings indicate that a clinical trial to determine the app's efficacy to promote PrEP uptake, HIV testing, and condom use is merited.

Compared with the HealthMindr pilot study, Saludfindr had a higher percentage of PrEP uptake among PrEP-naive users (7/17, 41% vs 9% during the former study [19]). PrEP use at baseline was also higher among Saludfindr participants than among HealthMindr participants (14/31, 45% vs 12%, respectively). These differences could be explained at least in part by the 6-year gap between the HealthMindr and Saludfindr pilot studies. Between 2017, when the HealthMindr pilot study was conducted, and 2023, PrEP awareness and initiation and willingness to use it among MSM increased overall across races and ethnicities in the United States [16,27,28]. In addition, all Saludfindr participants who started PrEP during the study did so at one of the clinics we listed on the app. This new feature to highlight clinics friendly to Latino MSM was not available in the original HealthMindr study, suggesting the importance of this newly added feature. Further facilitating PrEP access for Latino MSM through the app may be a more focused area of intervention to increase PrEP uptake in this group. For example, a future study could combine the customized clinic list feature with other evidence-based support such as PrEP navigation assistance [29,30].

In terms of HIV testing, the proportion of Saludfindr participants who ordered a home HIV test through the app was lower than that of HealthMindr participants and lower than those of other interventions for home HIV testing among Latino men, where uptake of home HIV testing ranged between 30% and 70% [31-34]. This finding may be due in part to the high uptake of PrEP, which entails regular HIV testing and, as such, additional home testing would not be needed. Among Saludfindr participants who did use HIV testing, most HIV tests were conducted at a clinic, often one of the clinics listed on the app. Our research staff reported that, during conversations about the app and its functionalities, some participants voiced hesitancy in ordering home HIV tests out of concerns about making mistakes while getting tested, which could lead to inaccurate test results. They were also concerned about interpreting test results correctly and what to do if they had a positive result. These results and insights suggest that, for future Saludfindr

XSL•FO
RenderX

versions, additional in-app content may need to be added to increase users' self-efficacy in using home HIV tests.

Saludfindr reached comparable SUS scores to those of HealthMindr (74.5 vs 73, respectively) and DOT Diary (SUS score of 76.4), a mobile app focused on PrEP retention in care and adherence and adapted for Latino MSM and transgender women's needs [35]. To optimize usability, future Saludfindr versions will need to include automatic data upload from the app to the cloud. The study version required users to tap the "Sync" button every time they used the app, which may explain why some participants found the app "unnecessarily complex."

Limitations

Our study had several limitations. First, Saludfindr's software was only compatible with Android phones. Even though we provided study cell phones to participants who did not have an Android phone, it represented a barrier to how seamlessly we could enroll participants into the study. Second, similar to all prevention studies, there was a risk of selection bias because individuals more familiar with prevention modalities may have been more motivated to participate in the study. Third, we only enrolled participants living in the Atlanta area as our formative work to develop the app and its features was based in this area. As such, our study results do not necessarily generalize to Latino MSM living in other areas of the country, including those living in rural areas or in states with enhanced support services for this population. Fourth, our sample size was relatively small compared to other mobile app–based studies. That said, we were able to enroll a population with multiple barriers to accessing PrEP and historically identified as "hard to reach." As such, their specific feedback related to an HIV prevention app is extremely valuable to optimize Saludfindr and guide other investigators working on digital HIV interventions for this group.

Conclusions

Despite these limitations, Saludfindr represents the first bilingual mobile app aiming to increase PrEP uptake, HIV testing, and condom use that has been tailored to the needs of primarily Spanish-speaking Latino MSM in the United States. There has been some movement to develop adherence support apps for Latino MSM, such as DOT Diary by Liu et al [35], a mobile app to support PrEP retention in care and adherence by combining electronic pill taking technology and a sex diary to provide PrEP protection-level feedback to users. Liu et al [35] adapted and piloted DOT Diary to meet the needs of Spanish-speaking MSM and transgender women in San Francisco and Miami, attaining high SUS scores, and it is currently being assessed for effectiveness in the setting of a clinical trial.

Overall, Saludfindr represents an innovative, culturally responsive tool that, once optimized, could contribute to efforts to curb the HIV epidemic among Latino MSM in the United States. Saludfindr, if proven effective in increasing PrEP uptake, HIV testing, and condom use among Latino MSM in a future clinical trial, could be deployed by health departments in other southern US cities and rural areas to assist in narrowing the existing racial and ethnic gap in PrEP uptake. We believe that this would be feasible as the app was well accepted by future users and it is easy to install and use. As with any tool, its effectiveness in decreasing HIV incidence in this population will likely be enhanced through combination with other strategies, mHealth-based or of other nature, as barriers to accessibility to HIV prevention services among this group are complex and multifactorial.

Acknowledgments

The authors thank the commitment and collaboration of the Latino LinQ team, without whom this work would not have been possible.

This work was supported by a CFAR Adelante Program grant from the National Institutes of Health (NIH) Center for AIDS Research at Emory University (P30 AI050409). Research reported in this publication was also supported by the National Cancer Institute Cancer Center Support Grant P30CA225520 awarded to the University of Oklahoma Stephenson Cancer Center. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Data Availability

The datasets generated or analyzed during this study are not publicly available as they contain potentially sensitive information and the study population is vulnerable; however, they are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: VDC (lead), AJS (supporting)
Data curation: HP-O (lead), IB (supporting), NS (supporting)
Formal analysis: VDC (lead), IB (supporting)
Funding acquisition: VDC
Methodology: VDC (lead), AJS (supporting)
Project administration: HP-O (lead), IB (supporting), VDC (supporting)
Investigation: VDC (lead), HP-O (supporting), IB (supporting), NS (supporting)
Writing - original draft: VDC

XSL•FO
RenderX

Writing - review & editing: VDC (lead), HP-O (supporting), IB (supporting), ER (supporting), PSS (supporting), AC-G (supporting), AJS (supporting)

Conflicts of Interest

VDC receives research funds paid to Emory University by Gilead Sciences.
AC-G receives research funds paid to Emory University by Gilead Sciences, Merck, ViiV Healthcare, and GSK.
PSS receives research funds paid to Emory University by Gilead Sciences, Merck, and ViiV Healthcare and consulting fees from Gilead Sciences and Merck.
AJS receives research funds paid to Emory University by Merck and ViiV Healthcare.

References

1. HIV surveillance supplemental report: estimated HIV incidence and prevalence in the United States, 2018–2022. Centers for Disease Control and Prevention. URL: https://stacks.cdc.gov/view/cdc/156513 [accessed 2025-10-02]
2. Fast facts: HIV and gay and bisexual men. Centers for Disease Control and Prevention Fast Facts. URL: https://www.cdc.gov/hiv/data-research/facts-stats/gay-bisexual-men.html [accessed 2025-10-02]
3. HIV/AIDS among Hispanic/Latino populations in Georgia: epidemiologic profile summary, 2014-2020. Georgia Department of Public Health. 2020. URL: https://dph.georgia.gov/document/document/hivepifact-sheethivaids-among-hispaniclatino-populations-georgia-2014-2020pdf/download [accessed 2025-10-7]
4. HIV among Hispanic/Latino populations in Georgia: an epidemiological profile, 2014-2021. Georgia Department of Public Health. 2023. URL: https://dph.georgia.gov/document/document/hivepifactsheethivaidsamonghispaniclatinopopulationsga2014-2021pdf/download [accessed 2025-10-7]
5. Sullivan PS, DuBose SN, Castel AD, et al. Equity of PrEP uptake by race, ethnicity, sex and region in the United States in the first decade of PrEP: a population-based analysis. Lancet Reg Health Am 2024 May;33(12):100738. [doi: 10.1016/j.lana.2024.100738] [Medline: 38659491]
6. Siegler AJ, Mouhanna F, Giler RM, et al. The prevalence of pre-exposure prophylaxis use and the pre-exposure prophylaxis-to-need ratio in the fourth quarter of 2017, United States. Ann Epidemiol 2018 Dec;28(12):841-849. [doi: 10.1016/j.annepidem.2018.06.005] [Medline: 29983236]
7. Cantos VD, Hagen K, Duarte AP, et al. Development of a mobile app to increase the uptake of HIV pre-exposure prophylaxis among Latino sexual minority men: qualitative needs assessment. JMIR Form Res 2023 Feb 8;7(6):e43844. [doi: 10.2196/43844] [Medline: 36625855]
8. Hassan R, Saldana CS, Garlow EW, et al. Barriers and facilitators to HIV service access among Hispanic/Latino gay, bisexual, and other men who have sex with men in Metropolitan Atlanta-a qualitative analysis. J Urban Health 2023 Dec;100(6):1193-1201. [doi: 10.1007/s11524-023-00809-9] [Medline: 38012505]
9. Saldana C, Philpott DC, Mauck DE, et al. Public health response to clusters of rapid HIV transmission among Hispanic or Latino gay, bisexual, and other men who have sex with men - Metropolitan Atlanta, Georgia, 2021-2022. MMWR Morb Mortal Wkly Rep 2023 Mar 10;72(10):261-264. [doi: 10.15585/mmwr.mm7210a3] [Medline: 36893048]
10. Blumenthal J, Jain S, Mulvihill E, et al. Perceived versus calculated HIV risk: implications for pre-exposure prophylaxis uptake in a randomized trial of men who have sex with men. J Acquir Immune Defic Syndr 2019 Feb 1;80(2):e23-e29. [doi: 10.1097/QAI.0000000000001888] [Medline: 30422909]
11. Haukoos JS, White DA, Rowan SE, et al. HIV risk and pre-exposure prophylaxis eligibility among emergency department patients. AIDS Patient Care STDS 2021 Jun;35(6):211-219. [doi: 10.1089/apc.2021.0012] [Medline: 34097464]
12. Lockard A, Rosenberg ES, Sullivan PS, et al. Contrasting self-perceived need and guideline-based indication for HIV pre-exposure prophylaxis among young, black men who have sex with men offered pre-exposure prophylaxis in Atlanta, Georgia. AIDS Patient Care STDS 2019 Mar;33(3):112-119. [doi: 10.1089/apc.2018.0135] [Medline: 30844305]
13. Rosengren AL, Lelutiu-Weinberger C, Woodhouse EW, Sandanapitchai P, Hightow-Weidman LB. A scoping review of HIV pre-exposure prophylaxis stigma and implications for stigma-reduction interventions for men and transwomen who have sex with men. AIDS Behav 2021 Jul;25(7):2054-2070. [doi: 10.1007/s10461-020-03135-2] [Medline: 33389319]
14. Brooks RA, Landrian A, Nieto O, Fehrenbacher A. Experiences of anticipated and enacted pre-exposure prophylaxis (PrEP) stigma among Latino MSM in Los Angeles. AIDS Behav 2019 Jul;23(7):1964-1973. [doi: 10.1007/s10461-019-02397-9] [Medline: 30649635]
15. Brooks RA, Nieto O, Landrian A, Fehrenbacher A, Cabral A. Experiences of pre-exposure prophylaxis (PrEP)-related stigma among Black MSM PrEP users in Los Angeles. J Urban Health 2020 Oct;97(5):679-691. [doi: 10.1007/s11524-019-00371-3] [Medline: 31214977]
16. Jones JT, Smith DK, Thorne SL, Wiener J, Michaels S, Gasparac J. Community members' pre-exposure prophylaxis awareness, attitudes, and trusted sources for PrEP information and provision, context matters survey, 2015-2016. AIDS Educ Prev 2020 Apr;32(2):102-1S6. [doi: 10.1521/aeap.2020.32.2.102] [Medline: 32539481]
17. Li C, Xiong Y, Maman S, et al. An instant messaging mobile phone application for promoting HIV pre-exposure prophylaxis uptake among Chinese gay, bisexual and other men who have sex with men: a mixed methods feasibility and piloting

XSL•FO
RenderX

randomized controlled trial study. PLoS ONE 2023;18(11):e0285036. [doi: 10.1371/journal.pone.0285036] [Medline: 37956177]
18. Jones J, Butler G, Woody M, et al. Preferences for and experiences of an HIV-prevention mobile app designed for transmasculine people: pilot feasibility trial and qualitative investigation. JMIR Form Res 2023 Sep 21;7(3):e51055. [doi: 10.2196/51055] [Medline: 37733430]
19. Sullivan PS, Driggers R, Stekler JD, et al. Usability and acceptability of a mobile comprehensive HIV prevention app for men who have sex with men: a pilot study. JMIR Mhealth Uhealth 2017 Mar 9;5(3):e26. [doi: 10.2196/mhealth.7199] [Medline: 28279949]
20. Shrestha R, Altice FL, Khati A, et al. Clinic-integrated smartphone app (JomPrEP) to improve uptake of HIV testing and pre-exposure prophylaxis among men who have sex with men in Malaysia: mixed methods evaluation of usability and acceptability. JMIR Mhealth Uhealth 2023 Feb 16;11(12):e44468. [doi: 10.2196/44468] [Medline: 36795465]
21. Biello KB, Daddario SR, Hill-Rorie J, et al. Uptake and acceptability of MyChoices: results of a pilot RCT of a mobile app designed to increase HIV testing and PrEP uptake among young American MSM. AIDS Behav 2022 Dec;26(12):3981-3990. [doi: 10.1007/s10461-022-03724-3] [Medline: 35829971]
22. Biello KB, Hill-Rorie J, Valente PK, et al. Development and evaluation of a mobile app designed to increase HIV testing and pre-exposure prophylaxis use among young men who have sex with men in the United States: open pilot trial. J Med Internet Res 2021 Mar 24;23(3):e25107. [doi: 10.2196/25107] [Medline: 33759792]
23. Collins CB Jr, Wilson KM. CDC's dissemination of evidence-based behavioral HIV prevention interventions. Transl Behav Med 2011 Jun;1(2):203-204. [doi: 10.1007/s13142-011-0048-9] [Medline: 24073042]
24. Wingood GM, DiClemente RJ. The ADAPT-ITT model: a novel method of adapting evidence-based HIV interventions. J Acquir Immune Defic Syndr 2008 Mar 1;47 Suppl 1(6):S40-S46. [doi: 10.1097/QAI.0b013e3181605df1] [Medline: 18301133]
25. Bangor A, Kortum PT, Miller JT. An empirical evaluation of the system usability scale. Int J Hum Comput Interact 2008 Jul 29;24(6):574-594. [doi: 10.1080/10447310802205776]
26. Billingham SAM, Whitehead AL, Julious SA. An audit of sample sizes for pilot and feasibility trials being undertaken in the United Kingdom registered in the United Kingdom Clinical Research Network database. BMC Med Res Methodol 2013 Aug 20;13(1):104. [doi: 10.1186/1471-2288-13-104] [Medline: 23961782]
27. King H, Thornton N, Evans KN, et al. Factors associated with the awareness of and willingness to Use HIV pre-exposure prophylaxis among gay, bisexual, and other men who have sex with men, Baltimore, MD, 2017-2019. J Racial Ethn Health Disparities 2025 Apr;12(2):1202-1215. [doi: 10.1007/s40615-024-01954-w] [Medline: 38436888]
28. Sullivan PS, Sanchez TH, Zlotorzynska M, et al. National trends in HIV pre-exposure prophylaxis awareness, willingness and use among United States men who have sex with men recruited online, 2013 through 2017. J Int AIDS Soc 2020 Mar;23(3):e25461. [doi: 10.1002/jia2.25461] [Medline: 32153119]
29. Matos LA, Janek SE, Holt L, Ledbetter L, Gonzalez-Guarda RM. Barriers and facilitators along the PrEP continuum of care among Latinx sexual minoritized men and transgender women: a systematic review. AIDS Behav 2024 Nov;28(11):3666-3709. [doi: 10.1007/s10461-024-04434-8] [Medline: 39083153]
30. Pathela P, Jamison K, Blank S, Daskalakis D, Hedberg T, Borges C. The HIV pre-exposure prophylaxis (PrEP) cascade at NYC sexual health clinics: navigation is the key to uptake. J Acquir Immune Defic Syndr 2020 Apr 1;83(4):357-364. [doi: 10.1097/QAI.0000000000002274] [Medline: 31904700]
31. Rosenthal EM, Tesoriero JM, Flavin SM, Yuan Y, Sullivan TT, Johnson MC. Geospatial prioritization to reach Hispanic or Latino and other priority populations through HIV home testing services. J Public Health Manag Pract 2023;29(3):336-344. [doi: 10.1097/PHH.0000000000001700] [Medline: 36693389]
32. Edelstein ZR, Wahnich A, Purpura LJ, et al. Five waves of an online HIV self-test giveaway in New York City, 2015 to 2018. Sex Transm Dis 2020 May;47(5S Suppl 1):S41-S47. [doi: 10.1097/OLQ.0000000000001144] [Medline: 32149955]
33. Merchant RC, Clark MA, Liu T, et al. Comparison of home-based oral fluid rapid HIV self-testing versus mail-in blood sample collection or medical/community HIV testing by young adult Black, Hispanic, and White MSM: results from a randomized trial. AIDS Behav 2018 Jan;22(1):337-346. [doi: 10.1007/s10461-017-1802-x] [Medline: 28540562]
34. Chiu CJ, Young SD. Correlates of requesting home HIV self-testing kits on online social networks among African-American and Latino men who have sex with men. AIDS Care 2016;28(3):289-293. [doi: 10.1080/09540121.2015.1090533] [Medline: 26444956]
35. Liu AY, Alleyne CD, Doblecki-Lewis S, et al. Adapting mHealth interventions (PrEPmate and DOT Diary) to support PrEP retention in care and adherence among English and Spanish-speaking men who have sex with men and transgender women in the United States: formative work and pilot randomized trial. JMIR Form Res 2024 Mar 27;8:e54073. [doi: 10.2196/54073] [Medline: 38536232]

Abbreviations

HIPAA: Health Insurance Portability and Accountability Act
IRB: institutional review board

XSL•FO
RenderX

mHealth: mobile health
MSM: men who have sex with men
PnR: pre-exposure prophylaxis–to-need ratio
PrEP: pre-exposure prophylaxis
SUS: System Usability Scale

Edited by A Mavragani; submitted 22.04.25; peer-reviewed by O Igunma; revised version received 11.09.25; accepted 12.09.25; published 28.10.25.

Please cite as:
Cantos VD, Posada-Orozco H, Batina I, Sanchez N, Rangel E, Sullivan PS, Camacho-Gonzalez A, Siegler AJ
Formative Evaluation of an HIV Prevention App tailored for Latino Men Who Have Sex With Men: Acceptability and Usability Study
JMIR Form Res 2025;9:e74208
URL: https://formative.jmir.org/2025/1/e74208
doi:10.2196/74208

© Valeria D Cantos, Humberto Posada-Orozco, Isabella Batina, Natalie Sanchez, Eric Rangel, Patrick S Sullivan, Andres Camacho-Gonzalez, Aaron J Siegler. Originally published in JMIR Formative Research (https://formative.jmir.org), 28.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Behavioral Change Intervention to Promote a Healthier Postpartum Lifestyle: Mixed Methods Pilot Study

Pernille Kjærgaard Christiansen[1,2,3,4*], PhD; Trine Kjær[4*], PhD; Christina Anne Vinter[5,6,7*], PhD; Mette Juel Rothmann[3,5,6*], PhD; Mette Maria Skjøth[3,6,8*], PhD; Eva Draborg[4*], PhD

[1]Applied Business Research, UCL University College, Seebladsgade 1, Odense, Denmark
[2]OPEN, Clinical Institute, University of Southern Denmark, Odense, Denmark
[3]Center for Innovative Medical Technology (CIMT), Odense, Denmark
[4]Department of Public Health, Danish Centre for Health Economics, University of Southern Denmark, Odense, Denmark
[5]Steno Diabetes Center, Odense, Denmark
[6]Department of Clinical Research, University of Southern Denmark, Odense, Denmark
[7]Department of Gynecology and Obstetrics, Odense University Hospital, Odense, Denmark
[8]Department of Dermatology and Allergy, Odense University Hospital, Odense, Denmark
[*]all authors contributed equally

Corresponding Author:
Pernille Kjærgaard Christiansen, PhD
Applied Business Research, UCL University College, Seebladsgade 1, Odense, Denmark

Abstract

Background: Research has shown that many mothers lack tools needed to motivate and support themselves in a healthy lifestyle after giving birth. A mobile health app (mHealth) has the potential to become a tool to accommodate this need. Accordingly, Healthy Together—a module in a general mHealth app, My Hospital including Podcasts, weight tracking, and exercise videos—was developed.

Objective: The aim was to assess mothers' use of the behavioral intervention, Healthy Together, which aims to support new mothers in a healthier lifestyle. Further, it evaluates mothers' experiences and attitudes when using the intervention.

Methods: A mixed method pilot study was conducted, and 34 women were included. From 3 weeks to 6 months postpartum, the women were granted access to Healthy Together. App activity was registered during the intervention period. All of the women received a questionnaire at the end of the intervention period; of these, 28 responded. In addition, 18 women participated in an online, semi-structured interview.

Results: On average, each invited participant accessed the module 37 times. Push notifications and podcasts were used by 65% (n=34), and thus the content used the most. One-third found push notifications motivating. Half used exercise videos, while slightly fewer utilized weight tracking. A total of 70% (n=28) of those who answered the questionnaire had used Healthy Together. About half of the users reported that the intervention had a positive influence on their health status, and 70% (n=20) of the users stated they would recommend Healthy Together to others. The mean age was 29.8 years for users and 32 years for non-users. Pre-pregnancy body mass index averaged 25 and 24.8, respectively, increasing to 25.6 and 25.7 at 6 months postpartum. The body mass index difference was 0.6 for users and 0.9 for non-users, corresponding to total increases of 2.4 and 3.6, respectively. By gradually introducing the mothers to new content, the mothers could more easily digest the information. Podcasts were, in general, the preferred information channel. Weight tracking reminders motivated some, while they had the opposite effect on others. The users seemed to be those who were more physically active and had a healthier diet prior to their pregnancy, compared to non-users. About one-third of the users experienced technical problems.

Conclusions: This study demonstrates that Healthy Together potentially is a feasible tool to assist women in improving their postpartum lifestyle. Further research with a larger sample is needed.

(JMIR Form Res 2025;9:e69391) doi:10.2196/69391

KEYWORDS

diet; exercise; health technology; motherhood; mixed methods; postpartum health

XSL•FO
RenderX

Introduction

Pregnancy is a strong contributor to women gaining weight [1]. According to a study by Rogozińska et al [2], more than one-third of women exceed the Institute of Medicine's (IOM) recommendations regarding weight gain during pregnancy. These women have an increased risk of not returning to their pre-pregnancy weights [3,4]. Following Villamor and Cnattingus [5], postpartum weight retention is associated with maternal obesity and adverse birth outcomes in subsequent pregnancies. As shown by Robinson [6] and Muttarek [7], being overweight has become so common that some overweight or obese people do not recognize themselves as being overweight, and health-related risks as the result of having a high BMI are often underestimated. Studies of the postpartum period show that women generally feel unprepared for the physical changes that come with pregnancy and birth and lack support of how to cope with these changes [8-11]. Although the educational preparations that women receive mainly focus on breastfeeding and infant care, there is not much focus on lifestyle and the postpartum woman's physical and emotional functioning [10,12]. Lack of knowledge about the physical changes that birth causes can make mothers afraid of exercising after birth, causing them to be inactive [12]. In addition, focus on the baby can cause mothers to give their own lifestyles less priority [12]. As pointed out by Walker et al [13], and our previous studies [12,14,15], there seems to be an unmet need for lifestyle support among women in the postpartum period. Mothers need better knowledge about the importance of a healthy postpartum lifestyle, as well as guidance on how to maintain a healthy lifestyle.

There is evidence that health care professionals (HCPs)—in particular, hospital doctors—have a high rate of being trusted [16]. Therefore, there is the potential to make information and guidance on postpartum lifestyles available by having relevant HCPs deliver the information. Most mothers of childbearing age use information technology daily [17], giving it the potential as a tool to deliver important information and guidance. Mobile health (mHealth) apps have increasingly been used to promote health, and the number of mHealth apps is expected to increase even further in the years to come [18,19]. However, as Eysenbach [20] has pointed out, it is important that the app is reliable and established for its purpose, as technical issues can have an impact on whether a person will use an app.

The hospitals in the Region of Southern Denmark use the mHealth app called My Hospital [21]. Each department can have one or more modules in the app. The app is free of charge and follows European GDPR standards [22]. To support women in maintaining a healthy postpartum lifestyle, Healthy Together was developed as a module within My Hospital [23] and pilot tested at the Gynecological Department at Odense University Hospital.

The aim of this paper was to examine postpartum women's experiences with, and attitudes toward, the use of Healthy Together, along with their frequency of use of the module, during the initial intervention period. In this paper, we examine whether Healthy Together enabled change in lifestyle behaviors in terms of diet and exercise and further identify possible barriers and facilitators to the module's use, including preferences for individual components: push notifications, exercise videos, weight tracking, and podcasts. A mixed methods approach was used to gather more comprehensive data.

Methods

Setting and Participants

The pilot study on implementation of Healthy Together was carried out in the Region of Southern Denmark between June and December 2020 at the Department of Gynecology and Obstetrics at Odense University Hospital. It included women who had just given birth and were over 18 years old, Danish-speaking, and agreed to participate. Exclusion criteria were mothers from vulnerable families who are cared for by a family center, which provides support for families that face challenges related to abuse, distress, and other social problems. The study did not consider any medical or surgical interventions, such as cesarean or medication. Breastfeeding was also not considered.

Healthy Together Module

Pregnant women in the Region of Southern Denmark are encouraged to use the My Hospital app through which they have access to all their appointments with the hospital and midwives, as well as leaflets [24]. Through the app, they can also communicate with a midwife. A new module, Healthy Together, was developed and added to the existing content [23]. The module's aim is to support new mothers to a healthy lifestyle that focuses on diet and exercise as delivered through podcasts, weekly push notifications, exercise videos, and self-monitored weight tracking. A healthy lifestyle is conceptualized as the mothers' self-assessment of the beneficial influence of the intervention on their individual health.

Following Kushniruk and Nøhr [25], the module was co-created with HCPs, mothers (intended users), and IT experts to define an intervention that met the women's actual needs [24] and was usable. Mothers who participated in developing the intervention expressed a wish for podcasts as the preferred information channel. They already listened to podcasts and found it convenient that they could listen while keeping their hands and eyes free [24]. Their view is in accordance with the increasing use of podcasts in Denmark, and in particular, among young people, with 60% of 20 - to 39-year-olds listening to podcasts in 2020 [26]. The mothers wished for exercise videos as they allowed them to exercise from home when the time would best suit them [24]. In a study by Laframboise and colleagues, virtual exercise videos have been used, with promising results, to reduce diastasis recti abdominis in postpartum women [27]. Previous studies have found push notifications to be effective for adhering to behavioral change interventions [28,29]. Push notifications were therefore proposed to remind the mothers about the presence of Healthy Together, while allowing them to gradually be introduced to the content in the module [24]. Finally, self-monitoring of weight gain/loss was added, as the mothers wanted a way to follow their own weight development. Studies show that regularly self-monitoring can improve weight loss outcomes [30,31]. Research also shows that new technologies, such as smartphone apps, may support adherence to

XSL•FO
RenderX

self-monitoring [32]. Healthy Together sent the women weekly push notifications containing advice related to a healthy postpartum lifestyle and information about where in the module they could find further information on the topic. The module contained 6 podcasts and 12 exercise videos. The exercise videos acknowledge that women are advised to delay physical activity for 4-6 weeks postpartum, with an extended recommendation of 6-8 weeks for those who have undergone a cesarean section. The exercises provided were precisely the same as those utilized in in-person postpartum exercise classes.

From 8 weeks postpartum, the women also had the opportunity to track their own weight by self-monitoring it on a regular basis and inputting the information under My data in the module. Every third week, they received a reminder about weight tracking in the module. The women could download the app on their own smartphones. They could also access the content in the app on a tablet or a computer.

Data

Overview

Data were gathered in a cross-sectional study. A mixed method containing information from data on actual use of the app, a questionnaire, and individual semi-structured interviews was used for the analysis following Creswell and Clark [33]. Hence, a mixed method concurrent design was applied. Data from the semi-structured interviews were triangulated with some of the questions from the questionnaire. From June 5 to July 10, 2020, a medical doctor and two nurses at the maternity wards at Odense University Hospital consecutively invited 40 new mothers to receive further information about the pilot study prior to their hospital discharge after the birth. Due to COVID-19, the hospital only had resources to recruit 40 new mothers. All women gave their consent to be contacted and to receive more information. Between 2 and 3 weeks postpartum, the first author contacted the women by phone or email. They were informed about the study and were asked to participate.

Extraction of Data From Healthy Together

From the app, we gathered summary data on the number of times each content page in the module was visited over a period of 6 months. Data were retrieved once all the women finished the intervention period. The statistics do not contain individual-level information but provide information on the total number of times that all participants visited during the intervention period.

Questionnaire

The questionnaire consisted of 33 questions about the women's experiences using Healthy Together and its components, their general use of technology, sociodemographics (income, education, employment, marital status, and number of children in the household), subjective well-being and health behaviors (weight, weight gained during pregnancy, diet, exercise, and general mode). The questions about well-being are from the WHO-5 well-being index, using 6-point Likert scales [34]. The respondents were divided into users and non-users and compared. Respondents' BMI and deviations from recommended weight gain according to the IOM were calculated based on self-reported information about height and weight. A translated version of the full questionnaire can be found in Multimedia Appendix 1. Data were gathered and stored in My Hospital, from which they were retrieved.

The questionnaire was pretested on 6 postpartum women, resulting in minor modifications. At 6 months postpartum, each woman received the questionnaire. Respondents participated in a lottery for a 500 Danish Krones (about US $75) voucher to shop for baby equipment. Reminders were sent after a week, and again after 2 weeks to those who had not previously replied.

Semi-Structured Interviews

A semi-structured interview guide with open-ended questions was developed to gather information about the women's experiences using Healthy Together [35]. The questions included whether they used the module, what their experiences were from using it, and if it had an impact on their lifestyles. The women were also asked if they had any suggestions for improvements and if they would recommend the app to others. The interview guide was pretested on a postpartum woman and is presented in Multimedia Appendix 2.

Due to restrictions related to the COVID-19 pandemic, the interviews were conducted by telephone or Zoom. All participants were invited to participate in an interview. When invited by phone, those who said they did not use or only scarcely used Healthy Together were asked if there were any specific reasons for this. Interview statements used in this publication were translated from Danish to English. The first author, who was a female PhD student, conducted and transcribed all interviews. The participants were informed that she was a PhD student at the University of Southern Denmark, who took part in the study.

Methodological Orientation

The study consisted of iterations, where each iteration gave new insight into the next iteration, in which the hermeneutic cycle was applied. The questionnaire was used to gather information on the women's attitudes toward the intervention and their experiences of using it. For this evaluation, a positivistic approach was applied [36].

Analysis

Data from the app were available on the aggregate level and therefore could not be merged with the questionnaire data. Descriptive results from the questionnaire were reported by user status. Unfortunately, due to the limited sample size of the questionnaire data, the statistical testing of differences was not applicable and was therefore not performed.

Field notes were made, and the interviews were recorded on a dictaphone and transcribed in NVivo12 (version 12.6.1; Lumivero) and analyzed by the first author [37]. The interviews were about 30-45 minutes each. The interview data were analyzed using systematic text condensation (STC) [38,39]. An STC was made in 4 steps, where the first step was obtaining a total impression by reading the transcripts several times. In the second step, meaning units were identified. In the third step, overall themes were defined. Finally, in the fourth step, the content from each meaning unit was abstracted and synthesized.

XSL•FO
RenderX

Ethical Considerations

The study complied with the Declaration of Helsinki and Danish Code of Conduct for Research Integrity [40]. The Danish Data Protection Agency (case no. 17/43487) approved the study, and it was submitted to the Scientific Ethics Committee for assessment. It was determined that the study did not require their approval, according to Danish legislation because no biological material or treatment was involved [41]. All women were given written and verbal information about the project. They were given an informed consent in which they were informed that their participation was voluntary, their data and statements would be anonymized, and they could withdraw their participation at any time. Participants did not receive any compensation.

Results

Patient Characteristics

In total, 85% (n=34) of the 40 mothers invited to the study accepted to participate. The first author granted them access to the module. All of the women had access to the module between 3 weeks postpartum and 6 months postpartum. One participant dropped out after a few weeks, due to technical problems with the app. A total of 28 women replied to the questionnaire, and 18 participated in an online, semi-structured interview after the intervention period (some did both). Interviews were conducted between December 2020 and January 2021. A flowchart of the study is shown in Figure 1.

Figure 1. Flowchart of recruitment of postpartum women from the Region of Southern Denmark, who participated in the study Healthy Together from June to December 2020.

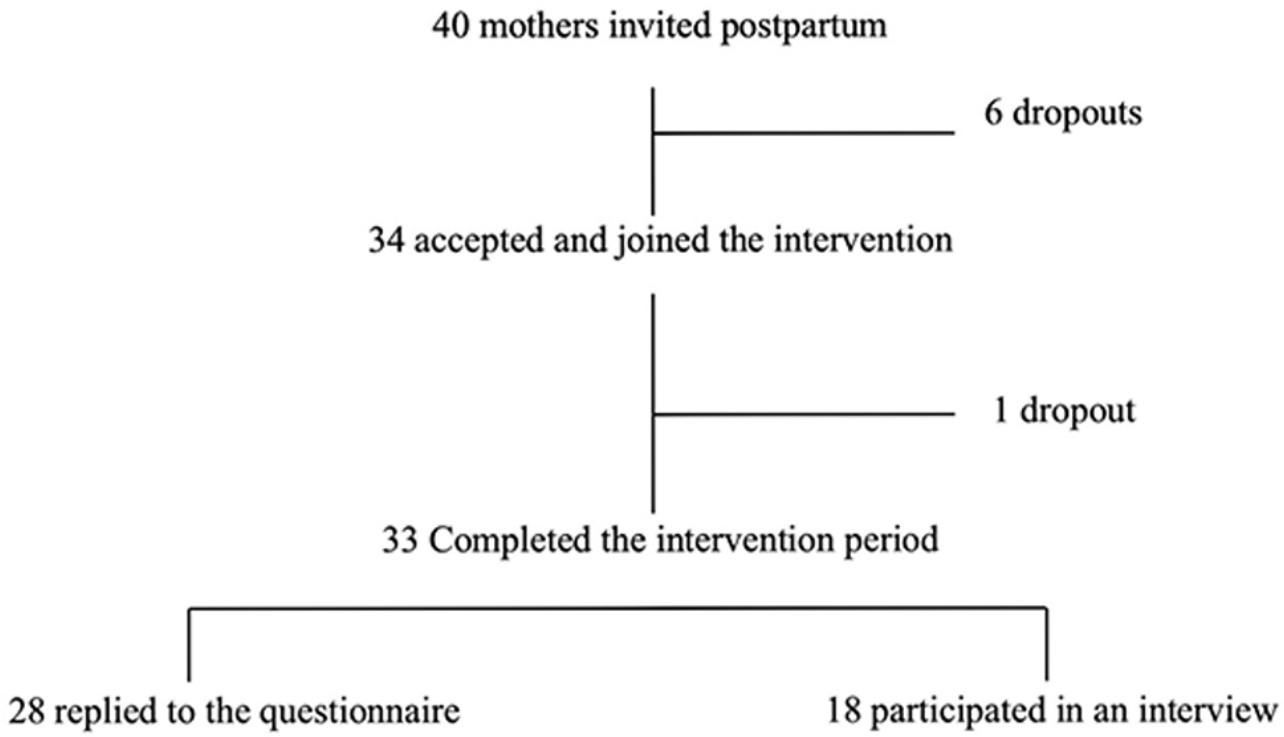

Data on Use of Healthy Together

The complete use of the module during the intervention period is displayed in Table 1. There was a variation in the use of the different elements in the module, with push notifications being used the most and contact information the least. On average, each of the invited participants visited the module 37 times. As we are not able to contingent on user status, this number is likely to cover a large variation in use (from zero to high use).

XSL•FO
RenderX

Table . Use of the module Healthy Together by postpartum women in the Region of Southern Denmark, who participated in the study (users and non-users), retrieved from the app (June-December 2020; n=34).

Menus	Number of visits	Average visits per participant (n=34)	Average visits per week per participant (n=34)
Push notifications	586	17	0.7
Exercise videos	167	5	0.2
Podcasts	164	5	0.2
My data (weight tracking)	151	4.5	0.2
About	112	3.3	0.1
Contact information	65	2	0
Total	1264	37	__[a]

[a]not applicable.

Questionnaire Findings

A total of 28 of the 33 participants (85%) responded to the questionnaire. The characteristics of the respondents at 6 months postpartum are shown in Table 2, grouped by user app status.

Overall, 70% (n=28) of the respondents stated they used Healthy Together. The respondents were between 23 and 40 years old at 6 months postpartum. The results seem to suggest some differences across user status (although not tested and therefore cannot be verified). The mean age of users was 29.8 years, whereas it was 32 years for non-users. Mean prepregnancy BMI for users and non-users at 6 months postpartum was 25 and 24.8, respectively. None of the respondents were underweight. During their pregnancies, half of the users exceeded the IOM's weight gain recommendations, compared to 25% (n=8) of those who did not use the app. At 6 months postpartum, the mean BMI was increased in both groups to 25.6 for users and 25.7 for non-users. The mean difference of BMI among users and non-users was 0.6 and 0.9, respectively, corresponding to an increase of BMI of 2.4 for users and 3.6 for non-users.

Prior to pregnancy, the users stated on average a healthier lifestyle, as measured by physical activity and diet compared to the non-users. At 6 months postpartum, physical activity and healthy diet had decreased in both groups compared to prepregnancy values. A total of 90% (n=20) of the users and all of the non-users stated they wanted to be more physically active, and 75% (n=28) of the users in both groups stated they would like to have a healthier diet. Among the users, 75% (n=20) felt they had been cheerful and in good spirits over the past 4 weeks compared to 25% (n=8) of the non-users, and a larger share of the users than non-users stated that they had control of their own health (90%, n=20 vs 62.5%, n=8). The median household income for users was between DKK 500,000 (US $81,587.5) and DKK 599,999 (US $97,905) and between DKK 400,000 (US $65,270) and DKK 499,999 (US $81,587) for non-users. The median category of education was medium cycle higher education for both users and non-users. All respondents were in a relationship.

XSL•FO
RenderX

Table . Characteristics of a sample of postpartum women, chosen consecutively in the Region of Southern Denmark between June and December 2020[a].

Questions	Users (n=20), n (%)	Non-users (n=8), n (%)	All (n=28), n (%)
Age (y)			
20 - 29	10 (50)	3 (37.5)	13 (46.4)
30 - 39	10 (50)	4 (50)	14 (50)
40+	0 (0)	1 (12.5)	1 (3.6)
Mean age (y)	29.8	32	30.4
Occupation prior to the maternity leave			
Under education/employed	17 (85)	6 (75)	23 (82.1)
Unemployed/other	3 (15)	2 (25)	5 (17.6)
Education level			
Primary/upper secondary education	2 (10)	1 (12.5)	3 (10.7)
Short cycle higher education	4 (20)	1 (12.5)	5 (17.6)
Medium/long cycle higher education	13 (65)	6 (75)	19 (67.9)
Other	1 (5)	0 (0)	1 (3.6)
Yearly household income (DKK)			
Under 300,000 (US $48,952)	4 (20)	1 (12.5)	5 (17.6)
300,000 - 499,000 (US $48,952-81,587)	4 (20)	4 (50)	8 (28.6)
500,000 (US $81,587.5) or more	12 (60)	3 (37.5)	15 (53.6)
Number of children in the household			
1 (vs more than 1)	13 (65)	4 (50)	17 (60.7)
Civil status			
Married or in a relationship	20 (100)	8 (100)	28 (100)
Hours spent on screen (phone) per day			
0 - 2	11 (55)	3 (37.5)	14 (50)
>2	9 (45)	5 (62.5)	14 (50)
Days of minimum 30-minute physical activity			
0 - 2	15 (75)	8 (100)	23 (82.1)
>2	5 (25)	0 (0)	5 (17.6)
Physical activity prior to pregnancy			
0 - 2	10 (50)	8 (100)	18 (64.3)
>2 days	10 (50)	0 (0)	10 (35.7)
Would you like to be more physically active?			
Yes	18 (90)	8 (100)	26 (92.9)
Diet prior to pregnancy			
Very healthy/healthy	13 (65)	3 (37.5)	16 (57.1)
Somewhat healthy/very unhealthy/unhealthy	7 (35)	5 (62.5)	12 (42.9)
Diet at present			

XSL•FO
RenderX

Questions	Users (n=20), n (%)	Non-users (n=8), n (%)	All (n=28), n (%)
Very healthy/healthy	9 (45)	1 (12.5)	10 (35.7)
Somewhat healthy/very unhealthy/unhealthy	11 (55)	7 (87.5)	18 (64.3)
Would you like to eat healthier?			
Yes	15 (75)	6 (75)	21 (75)
Over the past 4 weeks			
I have felt cheerful and in good spirits			
All the time/more than half of the time/some of the time	15 (75)	2 (25)	17 (60.7)
Less than half of the time/at no time	5 (25)	6 (75)	11 (39.3)
Over the past 4 weeks			
I have felt tired			
All the time/more than half of the time/some of the time	8 (40)	4 (50)	12 (42.9)
Less than half of the time/at no time	12 (60)	4 (50)	16 (57.1)
I have control of my health			
Strongly agree/agree	18 (90)	5 (62.5)	23 (82.1)
Neither, disagree, strongly disagree	2 (10)	3 (37.5)	5 (17.6)
What influences my health			
is what I do myself			
Strongly agree/agree	16 (80)	7 (87.5)	23 (82.1)
Neither, disagree, strongly disagree	4 (20)	1 (12.5)	5 (17.6)
BMI prior to pregnancy			
>25 (vs 25 and below)	6 (30)	3 (37.5)	9 (32.1)
Mean score	25	24.8	
Calculated BMI 6 months postpartum			
>25 (vs 25 and below)	6 (30)	4 (50)	10 (35.7)
Mean score	25.6	25.7	
IOM[b] weight gain recommendations (based on BMI)			
Above	10 (50)	2 (25)	12 (42.9)
Normal	5 (25)	4 (50)	9 (32.1)
Below	5 (25)	2 (25)	7 (25)

[a]Data collected by a questionnaire, n=28.

[b]IOM: Institute of Medicine.

The self-reported use of and satisfaction with Healthy Together is displayed in Table 3.

Weekly push messages and podcasts were rated as the most often used content. They were used by 65% (n=20) of the users. About one-third of those who used the module stated that the push notifications motivated them. Half of the respondents stated that they used the exercise videos, while slightly less used the weight tracking. More than half of those who used the module stated it worked well with podcasts and exercise videos. In total, 40% (n=20) of the women tracked their own weight development in the module. Only 15% (n=20) stated they regularly used Healthy Together. The majority stated that Healthy Together was easy to use, and all users found the

XSL•FO
RenderX

information in Healthy Together easy to understand. However, more than one-third of the participants answered that they experienced technical issues with the module. When asked if they had any comments, the interface was mentioned as an issue as it was too complicated to log on to the messages, and it seemed stiff and boring. In addition, another platform was recommended by several participants due to technical issues. Only 25% (n=20) of the participants used other apps related to postpartum health. More than half (55%, n=20) stated that Healthy Together had a positive effect on their health status, and 70% (n=20) stated they would recommend it to others.

XSL•FO
RenderX

Table . Use of the app Healthy Together by a sample of postpartum women chosen consecutively in the Region of Southern Denmark between June and December 2020.[a]

Question and response	N (%)
Which elements have you used in the module?	
Podcasts	13 (65)
Push notifications	13 (65)
Exercise videos	9 (45)
Weight tracking	8 (40)
I have used the module regularly	
Strongly agree	0 (0)
Agree	3 (15)
Neither agree nor disagree	8 (40)
Disagree	5 (25)
Strongly disagree	4 (20)
Don’t know	0 (0)
The module has been easy to use	
Strongly agree	6 (30)
Agree	8 (40)
Neither agree nor disagree	4 (20)
Disagree	1 (5)
Strongly disagree	1 (5)
Don’t know	0 (0)
The information in the module has been easy to understand	
Strongly agree	15 (75)
Agree	5 (25)
Neither agree nor disagree	0 (0)
Disagree	0 (0)
Strongly disagree	0 (0)
Don’t know	0 (0)
The module has had a positive influence on my health	
Strongly agree	1 (5)
Agree	10 (50)
Neither agree nor disagree	8 (40)
Disagree	0 (0)
Strongly disagree	0 (0)
Don’t know	1 (5)
It has worked well with podcasts	
Strongly agree	5 (25)
Agree	7 (35)
Neither agree nor disagree	5 (25)
Disagree	1 (5)
Strongly disagree	0 (0)
Don’t know	2 (10)
It has worked well with exercise videos	

XSL•FO
RenderX

Question and response	N (%)
Strongly agree	5 (25)
Agree	6 (30)
Neither agree nor disagree	2 (10)
Disagree	2 (10)
Strongly disagree	1 (5)
Don't know	4 (20)
The push notifications have motivated me	
Strongly agree	2 (10)
Agree	5 (25)
Neither agree nor disagree	8 (40)
Disagree	1 (5)
Strongly disagree	1 (5)
Don't know	3 (15)
I would recommend the module to others	
Strongly agree	4 (20)
Agree	10 (50)
Neither agree nor disagree	2 (10)
Disagree	1 (5)
Strongly disagree	0 (0)
Don't know	3 (15)
Have you experienced technical issues?	
Yes	7 (35)
No	12 (60)
Don't know	1 (5)
Do you use other apps related to your health?	
Yes	5 (25)
No	15 (75)
Don't know	0 (0)

[a]Non-users have been excluded. Data collected by a questionnaire, n=20.

Semi-Structured Interviews

Overview

In total, 18 semi-structured interviews were conducted. When invited for an interview, mothers who did not use or only scarcely used the module, according to their own perception, explained that they lacked time and energy to focus on it. A total of six overall key themes were identified from the STC: (1) information flow, (2) podcasts as a preferred information channel, (3) information from a trustworthy source, (4) technical barriers, (5) user interface, and (6) weight tracking reminders.

Key Theme 1: Information Flow

The women generally felt that the push notification reminded them about the module's presence and contents and were happy to receive reminders about the information in the module, as it gave them time to obtain and comprehend the information. "*The notifications reminded me about the topics in the module (Healthy Together). It made me more aware of my diet and also that I had to remember to exercise the pelvic floor"* (Mother, participant). When asked about the number of push notifications, most preferred to receive more messages during the first weeks, specifically those concerning the intake of water and healthy food. About two-thirds of the women would like to be introduced to the module at the end of their pregnancy, rather than after the baby was born, as they felt they still had time and energy to focus on new information.

Key Theme 2: Podcasts—A Preferred Information Channel

The women were happy to receive information through podcasts, as it was more personal than a leaflet, making them a more preferred information channel. As one mother said, "*A leaflet is way more factual…like a list of things I should keep in mind or do….I do not read them…A podcast on the other hand is*

XSL•FO
RenderX

more empathetic…It is like they are saying 'I understand how you feel'…and they are nice to listen to…and it is like I am the fly on the wall…" (Mother, participant). Some women mentioned that podcasts were convenient, as they could listen while having their hands and eyes free. However, not all the mothers used the podcasts, though they still preferred to receive information in this manner. The women who listened to the podcasts would like more of them; for instance, with information for those who had had a cesarean or a complicated delivery, as their bodies were more vulnerable. In addition, some wished for podcasts on mental well-being and solutions such as meditation.

Key Theme 3: Information From a Trustworthy Source

The women who used the exercise videos or listened to the podcasts said they gained new information they felt they could rely on regarding postpartum health and lifestyle, as the information was provided by the hospital, and thus from a trustworthy source. One mother said, "*It was as if someone held my hand"* (Mother, participant). The women liked the combination of the chosen information channels (push notifications, videos, and podcasts) and found them appropriate for delivering the provided information.

Key Theme 4: Technical Barriers

Several mothers expressed having experienced technical issues. In most cases, they had been logged off the app and did not receive the weekly push notifications. Consequently, they were not reminded about the module and its content. This would happen if they did not update the app or their phone. One participant said, "*I did not receive the messages…I forgot about the app…and then it became an app among the others…"* (Mother, participant). Hence, the technical problems had a negative impact on the women's use of Healthy Together.

Key Theme 5: User Interface

The women were generally positive about the content, though the user interface was an issue. They did not find the application user-friendly and found it a bit boring. For instance, when receiving push notifications, they would be informed that they could log in to see the message, rather than receiving the message directly on the screen. The main recommendation was using another app for Healthy Together. Finally, some mothers preferred a message that popped up with their own name in it. To further motivate usage, a participant said, "*It would make me feel that they are talking to me, rather than to a broad audience in a standard message"* (Mother, participant). Finally, adding an image of each exercise, next to the video, was recommended so the mothers could do the exercises at their own pace once they had already seen the videos. A minor theme was some participants suggested customizing the app so that the mothers could choose the content they wished to use.

Key Theme 6: Weight Tracking Reminders

There was great variation among those who liked the weight tracking reminders and those who did not. For some of the women, being asked about their weight elicited negative feelings, while others found it motivating. This may be related to the individual woman's progress related to her weight loss. One mother stated, "*It was like being slapped in the face…I had not lost weight, and I knew that…and then it made me feel less motivated to use the app…"* (Mother, participant).

Triangulation

Overall, there is a concurrence in the data from the semi-structured interviews, the questionnaire, and the data retrieved from the app, My Hospital. The women who participated in an interview had all used the module, whereas 70% (n=20) of those who answered the questionnaire (Table 3). When asked if they would recommend Healthy Together to others, 70% of the users in the questionnaire said yes, which is in line with the results from the semi-structured interviews.

In the interviews, when asked about their experience of using Healthy Together, the mothers were in general happy with the content provided. They highlighted technical problems as an issue in both the questionnaire and the interviews. When asked about suggestions for improvements, it was recommended in both the questionnaire and the interviews to use an independent app for Healthy Together, as the user interface was boring and some experienced technical issues.

In the questionnaire, podcasts and push notifications were rated the most used content, followed by exercise videos and weight tracking reminders, which is in line with the results from the interviews and the data retrieved from the app. According to Table 1, which presents the data retrieved from the app, push notifications were used the most, while podcasts and exercise videos were on average used an equal number of times. However, mothers typically engaged with podcasts only once, whereas they tended to revisit exercise videos multiple times.

In the interviews, mothers who already participated in private postpartum exercise classes explained that they, in general, did not use the exercise videos. The weight tracking was used by less than half of the users. Those who did lose weight were happy to track their weight, while those who did not felt demotivated.

A bit more than half of the users in both the interviews and questionnaire stated that the module had a positive impact on their individual health.

Discussion

Principal Findings and Comparison With Previous Works

This study demonstrated that the module Healthy Together in My Hospital has the potential to enable change in lifestyle behaviors in terms of diet and exercise, as it provides continuous reminders with trustworthy guidance and motivation to live healthier. The results did, however, also show that technical issues and the interface in My Hospital were considered barriers for use. The findings further suggest that the specific intervention might not target those in most need for information and guidance.

When looking at physical activity and diet, the intervention has been used more by those who were already physically active, had a healthy diet prior to pregnancy, and were in higher self-reported self-control. As both diet and physical activity

XSL•FO
RenderX

deteriorated in both groups, the intervention could benefit all users. The results from the questionnaire indicate that Healthy Together cannot be a stand-alone solution when aiming to assist all new mothers, as it was not used by all. A multicomponent intervention may thus be a way forward to reach a broader group of women postpartum [42]. Consequently, Healthy Together could serve as a supplement to the current services provided to the group of new "unproblematic" mothers who are given standard care. Whether the module should rely more on consultations with HCPs should also be considered. Finally, and according to our previous study, partner support has an impact on mothers' lifestyles. Thus, how to involve them should be considered further in future studies [24,43,44].

The mean BMI for both users and non-users was close to overweight both before and after the intervention period. Less than half of the participants used the weight tracking in the module, which also included weight tracking reminders. The reminders were intended to work as a "nudge" motivating to behavioral change [45]. Unfortunately, the inclusion of weight tracking reminders appeared to be two-sided, as it motivated those who applied the weight tracking component and did lose weight but had the opposite effect on those who put on weight. For the latter group, it also had a negative impact on their overall use of Healthy Together. Some of the women proposed to handle this problem by making the intervention more customized, thereby letting the women freely choose if they would like to receive weight tracking reminders or not and helping avoid pushing users who are not prepared. Our finding is in keeping with Eysenbach [20], who addresses the issue of push messages that are perceived as being too pushy and the fact that they can have a negative impact on the use of an app.

Push notifications referring to the same content can also decrease motivation to use an app, and thus have an impact on adherence or cause some users to uninstall the app [46]. Some said that a way to accommodate this could be by adding more content (podcasts and videos), for instance on mental well-being, while referring less to already presented content. Some mothers recommended making the module more personal by addressing the users by their name in the push notifications. This is relevant to bring forward as there is evidence that messages that include the receiver's name compared to standard messages can improve uptake and adherence [47]. In addition, the user interface should be improved by using fewer steps to access the push notifications.

Technical problems had a major impact on the mothers' use of the module. The results underline the importance of testing on a small scale, when using push notifications, as the sender does not receive information on whether a notification has been received or not. This is especially important when communication goes one way, there is no feedback on the reception and interpretation of a message [48]. Moreover, functionality has been found to affect app usage and to have a direct impact on dropouts if people feel annoyed or lack competencies to use the provided tool [20]. My Hospital may be a mature technology for other modules, though for Healthy Together, which uses push notifications, it has proven to be immature, which should be dealt with prior to further testing (pursuant to EU guidelines [49]).

Data from the summary statistics on average visits to the module showed that all content has been used, but to varying degrees and likely not by all participants. By gradually providing the mothers with reminders and information through push notifications, they expressed there was time to obtain and comprehend the information. In addition, the mothers' general wish was to be introduced to the module prior to giving birth, as then they still had time to take in new information. In line with Hoq, too much information handed out at the same time and lack of time to understand the information can cause information overload [50]. The use of podcasts was also expressed as being a convenient information channel, as the mothers could absorb the information while having their eyes and hands free. The health care system should reconsider the way it communicates and delivers information, for instance, by using podcasts [16]. In line with the summary statistics from the app, the mothers rated podcasts and push notifications as the content they had used the most. The study raises a broader question about how to communicate and deliver information so that it reaches those who need it the most.

Strengths and Weaknesses

The reliability of the results from the questionnaire may be limited due to the small sample size, and generalizability may be limited due to dropouts [51]. For the same reason, statistical hypothesis testing was not conducted. In contrast, the qualitative data material is comprehensive, which is a strength. After 18 interviews, an impression of information satiety had been reached [52].

To increase the response rate, all of the women who answered the questionnaire participated in the lottery for a 500 DKK (US $81.6) voucher for baby equipment. Using a lottery can motivate people with low motivation to participate [53], which is a strength. The women who participated in the study had to accept their participation. Therefore, they do not represent all who gave birth, but only those who willingly participated. However, it cannot be concluded whether this has caused a bias compared to all women who gave birth, and in which direction. The mothers were all in a relationship, almost all were employed or students, and were well educated.

The study is likely to suffer from selection bias. Women who were not motivated may have declined the offer to participate. In addition, there may be a selection bias with the overrepresentation of non-users among those who did not respond to the questionnaire, as the questionnaire was sent out via the module and not directly to their addresses. Unfortunately, we do not have information on the non-users. Finally, the questionnaire was self-reported, resulting in a possible bias related to under-reporting of weight and hours spent on the screen (phone), while over-reporting lifestyle (diet and exercise) and the use of Healthy Together.

The combined use of a questionnaire, interviews, and summary statistics from the app in a triangulation is a strength, as it provides more comprehensive information with different perspectives. In total, 18 respondents were interviewed, giving a better understanding of the experiences and attitudes. The respondents in the questionnaire and interviews were not completely identical, as those who accepted an invitation for

XSL•FO
RenderX

an interview had mostly used Healthy Together, while the respondents in the questionnaire contained a higher share of non-users.

Credibility was ensured by using a semi-structured interview guide with follow-up questions [54]. A semi-structured interview guide was also used to aim to gather similar data across all interviews. The study was conducted in a local setting. However, literature suggests that postpartum women in general lack similar support [8-10]. Thus, transferability to another setting [54] with similar users seems possible. Confirmability was aimed for by using notes and transcripts from all the interviews. In addition, the results were discussed with the co-authors. Regarding dependability, all interviews were conducted online due to COVID-19 restrictions. Authenticity was aimed for by selecting the appropriate people (postpartum women) and presenting their views and deeper understanding of their attitudes from using Healthy Together. The interviews were conducted via phone or Zoom, which may have caused data loss, as behavior and body language could not be observed. There is, however, as shown by Novick, only limited evidence of data loss when conducting phone interviews [55].

Healthy Together was tested at the beginning of the COVID-19 lockdown. Consequently, exercise groups for new mothers were, in most cases, canceled or broadcast online. Thus, the women had even less physical contact with HCPs. This may have had an impact on their use of Healthy Together, but whether this has increased or decreased use and satisfaction is uncertain. Medical or surgical interventions were not considered in the inclusion of the women, and the impact of this is unknown.

Healthy Together was developed in a tax-financed health care system. The module can potentially be adapted and implemented in other health care systems that focus on promoting a healthy lifestyle. While there are some costs related to the development in another setting, the marginal cost of offering the module is close to zero. Therefore, it has potential to be used on a large scale.

Conclusion

The study demonstrates that the app module Healthy Together is a potentially feasible tool to assist women in improving their lifestyle postpartum. The information flow with weekly reminders on podcasts, videos, and weight tracking gave the mothers time to obtain and comprehend the information and has most likely increased the overall use of the module. Reactions to the reminders of weight tracking were heterogeneous and, in some cases, negative, suggesting that a customized solution should be considered to avoid demotivation.

Though the intervention did not reach all, the intervention is still relevant for those who use it. The module could benefit from a more personalized and customized approach where the users can choose the content that interests them. Finally, the women should be introduced to the module at the end of their pregnancy, as they still have time and energy to obtain new information.

The study demonstrates that it is important to use a platform that is ready for the purpose. In addition, improving the user interface should be considered. Further research with a larger sample is needed in order to measure the effects related to the women's lifestyles and health statuses.

Acknowledgments

We would like to thank all the women who took part in the study and shared their experiences using the app, Healthy Together. We would also like to thank the nurses at the maternity ward at Odense University Hospital for assisting with recruitment.

Data Availability

All data are presented in the manuscript.

Authors' Contributions

Conducting interviews: PKC
Development of the interview guide: MJR (equal), PKC (equal)
Development of questionnaires: CAV (equal), ED (equal), MJR (equal), MMS (equal), PKC (equal), TK (equal)
Formal analysis: CAV (supporting), ED (supporting), MJR (supporting), MMS (supporting), PKC (lead), TK (supporting)
Recruitment of women: CAV
Supervision: CAV (equal), ED (equal), MJR (equal), MMS (equal), TK (equal)
Writing the original draft: CAV (supporting), ED (supporting), MJR (supporting), MMS (supporting), PKC (lead), TK (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Questionnaire.
[DOCX File, 17 KB - formative_v9i1e69391_app1.docx]

Multimedia Appendix 2

XSL•FO
RenderX

Semi-structured interview guide.
[DOCX File, 15 KB - formative_v9i1e69391_app2.docx]

References

1. Umberson D, Liu H, Mirowsky J, Reczek C. Parenthood and trajectories of change in body weight over the life course. Soc Sci Med 2011 Nov;73(9):1323-1331. [doi: 10.1016/j.socscimed.2011.08.014] [Medline: 21925781]
2. Rogozińska E, Zamora J, Marlin N, et al. Gestational weight gain outside the Institute of Medicine recommendations and adverse pregnancy outcomes: analysis using individual participant data from randomised trials. BMC Pregnancy Childbirth 2019 Dec;19(1):322. [doi: 10.1186/s12884-019-2472-7]
3. Bogaerts AFL, Devlieger R, Nuyts E, Witters I, Gyselaers W, Van den Bergh BRH. Effects of lifestyle intervention in obese pregnant women on gestational weight gain and mental health: a randomized controlled trial. Int J Obes (Lond) 2013 Jun;37(6):814-821. [doi: 10.1038/ijo.2012.162] [Medline: 23032404]
4. Villamor E, Tedroff K, Peterson M, et al. Association between maternal body mass index in early pregnancy and incidence of cerebral palsy. JAMA 2017 Mar 7;317(9):925-936. [doi: 10.1001/jama.2017.0945] [Medline: 28267854]
5. Villamor E, Cnattingius S. Interpregnancy weight change and risk of adverse pregnancy outcomes: a population-based study. Lancet 2006 Sep 30;368(9542):1164-1170. [doi: 10.1016/S0140-6736(06)69473-7] [Medline: 17011943]
6. Robinson E. Overweight but unseen: a review of the underestimation of weight status and a visual normalization theory. Obes Rev 2017 Oct;18(10):1200-1209. [doi: 10.1111/obr.12570] [Medline: 28730613]
7. Muttarak R. Normalization of plus size and the danger of unseen overweight and obesity in England. Obesity (Silver Spring) 2018 Jul;26(7):1125-1129. [doi: 10.1002/oby.22204] [Medline: 29932517]
8. Sacks A, Birndorf C. What No One Tells You: A Guide to Your Emotions From Pregnancy to Motherhood: Simon & Schuster Paperbacks; 2019. URL: https://www.simonandschuster.ca/books/What-No-One-Tells-You/Alexandra-Sacks/9781501112560 [accessed 2025-10-07]
9. Tully KP, Stuebe AM, Verbiest SB. The fourth trimester: a critical transition period with unmet maternal health needs. Am J Obstet Gynecol 2017 Jul;217(1):37-41. [doi: 10.1016/j.ajog.2017.03.032] [Medline: 28390671]
10. Martin A, Horowitz C, Balbierz A, Howell EA. Views of women and clinicians on postpartum preparation and recovery. Matern Child Health J 2014 Apr;18(3):707-713. [doi: 10.1007/s10995-013-1297-7] [Medline: 23775250]
11. Howell EA. Lack of patient preparation for the postpartum period and patients' satisfaction with their obstetric clinicians. Obstet Gynecol 2010 Feb;115(2 Pt 1):284-289. [doi: 10.1097/AOG.0b013e3181c8b39b] [Medline: 20093900]
12. Christiansen PK, Skjøth MM, Lorenzen LE, et al. Barriers to a healthy postpartum lifestyle and the possibilities of an information technology-based intervention: a qualitative study. Midwifery 2021 Jul;98:102994. [doi: 10.1016/j.midw.2021.102994] [Medline: 33845388]
13. Walker LO, Murphey CL, Nichols F. The broken thread of health promotion and disease prevention for women during the postpartum period. J Perinat Educ 2015;24(2):81-92. [doi: 10.1891/1058-1243.24.2.81] [Medline: 26957891]
14. Christiansen PK, Skjøth MM, Rothmann MJ, Vinter CA, Lamont RF, Draborg E. Lifestyle interventions to maternal weight loss after birth: a systematic review. Syst Rev 2019 Dec 16;8(1):327. [doi: 10.1186/s13643-019-1186-2] [Medline: 31842988]
15. Christiansen PK, Rothmann MJ, Skjøth MM, et al. The development of an IT-based intervention to support a healthy postpartum lifestyle through behavioral change. Health Informatics J 2023;29(1):14604582231153523. [doi: 10.1177/14604582231153523] [Medline: 36745514]
16. Gopal DP. Viewpoint: why I prescribe podcasts to my patients. Br J Gen Pract 2019 Dec;69(689):621. [doi: 10.3399/bjgp19X707009] [Medline: 31780490]
17. Statistics Denmark. IT-anvendelse I Befolkningen 2021. URL: https://www.dst.dk/Site/Dst/Udgivelser/GetPubFile.aspx?id=39431&sid=itbef2021 [accessed 2022-05-10]
18. Global market insight. mHealth Market - By Platform (Mobile Apps, Wearable Devices), By Application (Communication and Training, Education and Awareness, Diagnostics and Treatment, Remote Data Collection, Remote Monitoring), By End-use & Forecast, 2023-2032. URL: https://www.gminsights.com/industry-analysis/mhealth-market [accessed 2022-05-20]
19. Raft CF, Høffner M, Laursen PB, Hougaard IB. Når mere ikke er bedre – 2021. URL: https://vaelgklogt.dk/sites/danskepatienter.dk/files/media/document/VaelgKlogt_rapport_final_0.pd f [accessed 2022-05-20]
20. Eysenbach G. The law of attrition. J Med Internet Res 2005 Mar 31;7(1):e11. [doi: 10.2196/jmir.7.1.e11] [Medline: 15829473]
21. CIMT. Overview of patient use of the app My Hospital. URL: https://cimt.dk/en/projects/archive-old-projects/overview-of-patient-use-of-the-app-my-hospital [accessed 2022-03-10]
22. Rothmann MJ, Mouritsen JD, Ladefoged NS, et al. The use of telehealth for psychological counselling of vulnerable adult patients with rheumatic diseases or diabetes: explorative study inspired by participatory design. JMIR Hum Factors 2022 Mar 21;9(1):e30829. [doi: 10.2196/30829] [Medline: 35311690]
23. OUH. App: Mit sygehus. URL: https://ouh.dk/forskning-og-innovation/innovation-og-sundhedsteknologi/innovation-en-del-af-hverdagen/appen-mit-sygehus [accessed 2022-03-10]
24. OUH. Appen Mit Sygehus. URL: https://ouh.dk/til-patienter-og-parorende/odense/afdelinger/d-gynaekologisk-obstetrisk-afdeling/graviditet-fodsel-og-barsel/app-mit-sygehus [accessed 2022-03-25]

XSL•FO
RenderX

25. Kushniruk A, Nøhr C. Participatory design, user involvement and health IT evaluation. Stud Health Technol Inform 2016;222:139-151. [doi: 10.3233/978-1-61499-635-4-139] [Medline: 27198099]
26. Statistics denmark. Weekly podcast listening in Denmark in 2020, by age. 2020. URL: https://www.statista.com/statistics/1011819/weekly-podcast-listening-in-denmark-by-age/ [accessed 2022-02-10]
27. Laframboise FC, Schlaff RA, Baruth M. Postpartum exercise intervention targeting diastasis recti abdominis. Int J Exerc Sci 2021;14(3):400-409. [doi: 10.70252/GARZ3559] [Medline: 34055160]
28. Fjeldsoe BS, Marshall AL, Miller YD. Behavior change interventions delivered by mobile telephone short-message service. Am J Prev Med 2009 Feb;36(2):165-173. [doi: 10.1016/j.amepre.2008.09.040] [Medline: 19135907]
29. Free C, Phillips G, Galli L, et al. The effectiveness of mobile-health technology-based health behaviour change or disease management interventions for health care consumers: a systematic review. PLoS Med 2013;10(1):e1001362. [doi: 10.1371/journal.pmed.1001362] [Medline: 23349621]
30. Butryn ML, Phelan S, Hill JO, Wing RR. Consistent self-monitoring of weight: a key component of successful weight loss maintenance. Obesity (Silver Spring) 2007 Dec;15(12):3091-3096. [doi: 10.1038/oby.2007.368] [Medline: 18198319]
31. Brockmann AN, Eastman A, Ross KM. Frequency and consistency of self-weighing to promote weight-loss maintenance. Obesity (Silver Spring) 2020 Jul;28(7):1215-1218. [doi: 10.1002/oby.22828] [Medline: 32437055]
32. Carpenter CA, Ugwoaba UA, Cardel MI, Ross KM. Using self-monitoring technology for nutritional counseling and weight management. Digit Health 2022;8:20552076221102774. [doi: 10.1177/20552076221102774] [Medline: 35663238]
33. Creswell JW, Plano Clark VL. Designing and Conducting Mixed Methods Research, 3rd edition: Sage Publications; 2011.
34. Mental Health Services. The WHO-5 website. 2020. URL: https://www.psykiatri-regionh.dk/who-5/who-5-questionnaires/Pages/default.aspx [accessed 2021-01-25]
35. Holloway I, Kathleen G. Qualitative Research in Nursing and Healthcare, 4th edition: Wiley Blackwell; 2017. URL: https://www.wiley.com/en-ca/Qualitative+Research+in+Nursing+and+Healthcare%2C+4th+Edition-p-9781118874486 [accessed 2025-10-07]
36. Videnskabsteori, 3rd edition: Lindhardt og Ringhof; 2011.
37. Alfasoft. NVIVO. URL: https://tinyurl.com/2ka4eyvc [accessed 2022-05-12]
38. Malterud K. Systematic text condensation: a strategy for qualitative analysis. Scand J Public Health 2012 Dec;40(8):795-805. [doi: 10.1177/1403494812465030] [Medline: 23221918]
39. Malterud K. Kvalitative Forskningsmetoder for Medisin Og Helsefag, 4th edition: Universitetsforlaget. URL: https://www.universitetsforlaget.no/kvalitative-forskningsmetoder-for-medisin-og-helsefag-4 [accessed 2025-10-07]
40. World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA 2013 Nov 27;310(20):2191-2194. [doi: 10.1001/jama.2013.281053] [Medline: 24141714]
41. Danish Code of Conduct for Research Integrity. 2014. URL: https://ufm.dk/en/publications/2014/files-2014-1/the-danish-code-of-conduct-for-research-integrity.pdf [accessed 2022-12-12]
42. Schoeppe S, Alley S, Van Lippevelde W, et al. Efficacy of interventions that use apps to improve diet, physical activity and sedentary behaviour: a systematic review. Int J Behav Nutr Phys Act 2016 Dec 7;13(1):127. [doi: 10.1186/s12966-016-0454-y] [Medline: 27927218]
43. Maindal HT, Timm A, Dahl-Petersen IK, et al. Systematically developing a family-based health promotion intervention for women with prior gestational diabetes based on evidence, theory and co-production: the face-it study. BMC Public Health 2021 Sep 3;21(1):1616. [doi: 10.1186/s12889-021-11655-2] [Medline: 34479526]
44. Nielsen KK, Dahl-Petersen IK, Jensen DM, et al. Protocol for a randomised controlled trial of a co-produced, complex, health promotion intervention for women with prior gestational diabetes and their families: the face-it study. Trials 2020 Feb 7;21(1):146. [doi: 10.1186/s13063-020-4062-4] [Medline: 32033613]
45. Thaler RH, Sunstein CR. Nudge—Improving Decisions About Health, Wealth and Happiness: Penguin Books Ltd; 2008. URL: https://www.researchgate.net/publication/235413094_NUDGE_Improving_Decisions_About_Health_Wealth_and_Happiness [accessed 2025-10-07]
46. Pham XL, Nguyen TH, Hwang WY, Chen GD. Effects of push notifications on learner engagement in a mobile learning app. Presented at: 2016 IEEE 16th International Conference on Advanced Learning Technologies (ICALT); Jul 25-28, 2016; Austin, TX, USA p. 90-94. [doi: 10.1109/ICALT.2016.50]
47. Nyatsanza F, McSorley J, Murphy S, Brook G. “It’s all in the message”: the utility of personalised short message service (SMS) texts to remind patients at higher risk of STIs and HIV to reattend for testing-a repeat before and after study. Sex Transm Infect 2016 Aug;92(5):393-395. [doi: 10.1136/sextrans-2015-052216] [Medline: 26670912]
48. Silvius G. Strategic Integration of Social Media into Project Management Practice Global Advances in IT Personnel and Project Management: IGI Global; 2016. [doi: 10.4018/978-1-4666-9867-3]
49. European Commission, Directorate-General for Research and Innovation, Strazza C, Olivieri N, et al. Technology readiness level: guidance principles for renewable energy technologies: final report. : Publications Office; 2017 URL: https://data.europa.eu/doi/10.2777/577767 [accessed 2022-05-16]
50. Hoq KMG. Information overload: causes, consequences and remedies—a study. Philos Prog 2016 Feb;55(1-2):49-68. [doi: 10.3329/pp.v55i1-2.26390]

XSL•FO
RenderX

51. Carmines EG. Reliability and Validity Assessment—Quantitative Applications in Social Sciences: SAGE University Paper; 1979. [doi: 10.4135/9781412985642]
52. Kvale S, Brinkmann S. Interview: Interview—Introduktion Til et Håndværk: Hans Reitzels Forlag; 2009. URL: https://vbn.aau.dk/en/publications/interview-introduktion-til-et-h%C3%A5ndv%C3%A6rk [accessed 2025-10-07]
53. Göritz AS, Luthe SC. Effects of lotteries on response behavior in online panels. Field Methods 2013 Aug;25(3):219-237. [doi: 10.1177/1525822X12472876]
54. Lincoln YS, Guba EG. Naturalistic Inquiry: Sage Publications; 1985. URL: https://www.researchgate.net/profile/Jan-Armstrong/publication/256294652_Naturalistic_Inquiry/links/02e7e5222605c57278000000/Naturalistic-Inquiry.pdf [accessed 2025-10-07]
55. Novick G. Is there a bias against telephone interviews in qualitative research? Res Nurs Health 2008 Aug;31(4):391-398. [doi: 10.1002/nur.20259] [Medline: 18203128]

Abbreviations

BMI: body mass index
HCP: health care professional
IOM: Institute of Medicine
STC: systematic text condensation

Edited by A Mavragani; submitted 13.12.24; peer-reviewed by KC Stryczek; revised version received 31.05.25; accepted 03.07.25; published 22.10.25.

Please cite as:
Christiansen PK, Kjær T, Vinter CA, Rothmann MJ, Skjøth MM, Draborg E
Behavioral Change Intervention to Promote a Healthier Postpartum Lifestyle: Mixed Methods Pilot Study
JMIR Form Res 2025;9:e69391
URL: https://formative.jmir.org/2025/1/e69391
doi:10.2196/69391

© Pernille Kjærgaard Christiansen, Trine Kjær, Christina Anne Vinter, Mette Juel Rothmann, Mette Maria Skjøth, Eva Draborg. Originally published in JMIR Formative Research (https://formative.jmir.org), 22.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Thematic Analysis of Smoking Cessation and Future Cessation Interventions for Cancer Survivors: Convenience Sampling Study

Kinsey Pebley[1,2], PhD, MPH; Tianna Williams[1], MPH; Kathryn Moody[1], BS; Alana M Rojewski[1,2], PhD

[1]Department of Public Health Sciences, Medical University of South Carolina, Charleston, SC, United States

[2]Hollings Cancer Center, Medical University of South Carolina, 86 Jonathan Lucas St, Charleston, SC, United States

Corresponding Author:
Kinsey Pebley, PhD, MPH
Department of Public Health Sciences, Medical University of South Carolina, Charleston, SC, United States

Abstract

Background: Smoking after a cancer diagnosis has significant health consequences, and there are substantial benefits if cancer survivors cease smoking. However, there are few smoking cessation interventions for cancer survivors that have been effective, and they are highly resource intensive. Thus, novel low resource cessation interventions are needed. One such intervention modality that has not been tested among cancer survivors is using pre-recorded videos to deliver information about skills and resources to achieve smoking cessation.

Objective: This study aimed to assess barriers to smoking cessation, interest in a video-based smoking cessation intervention, and video content preferences (eg, topic, video length, age and gender of video presenter, presenter type [peer, medical professional]) among individuals with cancer who smoke cigarettes.

Methods: Participants were recruited from a cancer center in the southeast United States that houses an opt-out tobacco treatment program in which tobacco treatment specialists proactively call all patients who have a current smoking status in their medical record. Patients were informed about the study, and their contact information was sent to the study team. Verbal consent was obtained from individuals who were cancer survivors, who were currently smoking, and who agreed to participate; semi-structured qualitative interviews were conducted (n=10). Participants were asked questions about smoking history, previous quit attempts, barriers to quitting smoking, previous experience with and openness to watching videos about quitting smoking, and participant preferences related to video content and presentation. Interviews were transcribed and coded by two reviewers, and a codebook was developed. A thematic analysis was then conducted.

Results: Results indicated that all participants had previously tried to quit smoking, but other substance use (5/10; 50%), negative affect (eg, depression [4/10; 40%], anxiety [1/10; 10%], stress [2/10; 20%], and loneliness [1/10; 10%]), the social environment (eg, being around others who smoke [2/10; 20%] and risky social situations [2/10; 20%]), and habits surrounding smoking were significant barriers for cessation. All participants were open to watching smoking cessation videos and expressed a desire to see a peer with lived experience in the videos. Most (8/10; 80%) participants did not have preferences related to the age or gender of video presenters. Many participants had difficulty articulating content preferences for the videos.

Conclusions: Cancer survivors who smoke may not be aware of their knowledge gaps related to smoking cessation, and videos may be an acceptable way to provide evidence-based information to fill knowledge gaps.

(JMIR Form Res 2025;9:e76792) doi:10.2196/76792

KEYWORDS

smoking; smoking cessation; cancer; video; video intervention

Introduction

Smoking remains the leading cause of preventable death in the United States, and smoking cessation is critical to improving health [1]. The health consequences of smoking are exacerbated after a cancer diagnosis, as cancer survivors (defined as any living person who has received a cancer diagnosis) who smoke are at increased risk of morbidity and mortality, while those who quit have improved treatment outcomes and survival [2-8]. However, an estimated 11.4% of survivors continue to smoke after diagnosis despite the health risks and impact to their cancer treatment outcomes [9]. Thus, there is significant room for improvement related to smoking cessation among cancer survivors.

Numerous smoking cessation interventions for cancer patients have been developed and tested, yet the majority have not resulted in increased smoking cessation rates compared to control interventions. One review found that 36 smoking cessation interventions for cancer survivors were tested by year 2020, and only 4 of these interventions resulted in statistically

XSL•FO
RenderX

significant abstinence rates compared to control conditions [10]. Intervention strategies included methods such as contingency management and intensive counseling [10], meaning these interventions were also resource-intensive and required significant monetary and staff resources to deliver. Such interventions may not be feasible for under-resourced clinics, individuals living in rural areas without access to dedicated tobacco treatment services, or individuals who are interested in quitting on their own. Thus, there is a need for novel, effective, and less resource-intensive smoking cessation interventions for cancer survivors.

One intervention modality that has not been explored in the context of smoking cessation among cancer survivors, to our knowledge, is the delivery of smoking cessation content via video. Online informational videos are easily accessible via websites such as YouTube and thousands of online videos on these websites address smoking cessation. However, many of these videos do not depict evidence-based smoking cessation strategies [11]. Video-based interventions hold promise for a variety of other health considerations (eg, diabetes, chronic obstructive pulmonary disease, HIV, etc) by improving patient knowledge, facilitating health behavior changes, and increasing self-efficacy. However, many video intervention studies have been found to be at moderate to high risk of bias [12], and video interventions in the context of smoking cessation have not been as widely tested. One study conducted in the Netherlands found that a video intervention produced significantly higher long-term abstinence rates than a text-based intervention or generic advice [13]. Another study among individuals experiencing mental illness also found that a smoking cessation video intervention was associated with an increase in patient knowledge about quitting, and that this intervention was acceptable to patients [14]. Thus, video interventions for smoking cessation are promising, but there is a need for more work in this area across different populations.

The current study used semi-structured interviews with cancer survivors who smoke cigarettes to assess previous smoking cessation experiences, barriers to smoking cessation, interest in a novel video-based smoking cessation intervention, and preferences for video content and video presenter characteristics. The study was exploratory in nature and aimed to learn more about how videos should look (eg, who is presenting the information) and what survivors would like to learn more about. This qualitative study was planned to act as a guide to begin the creation of a video-based smoking cessation intervention tailored to the unique needs of cancer survivors.

Methods

Participants and Recruitment

Participants were recruited from the National Cancer Institute-designated Hollings Cancer Center at the Medical University of South Carolina, which is an academic medical center located in the southern United States. The cancer center houses an opt-out tobacco treatment program where all individuals receiving care at the cancer center who are designated as smoking in their medical chart receive a phone call to offer tobacco treatment services by a tobacco treatment specialist (TTS). The TTSs were asked to inform all of their patients that there was a study that they may be eligible for, reflecting a convenience sampling approach. The TTSs provided the patient's contact information to the study team if they were interested in receiving more information. Inclusion criteria for participation included being 18 years or older, diagnosed with cancer (self-reported), and currently smoking cigarettes (self-reported). Exclusion criteria included not being able to speak English and no access to a telephone.

Procedures

Results are reported following the Standards for Reporting Qualitative Research (SRQR) guidelines [15]. Semi-structured qualitative interviews were conducted by the primary investigator (PI, author KP) via telephone with participants using an interview guide created to lead the conversation related to patient experiences, barriers, and preferences (see Multimedia Appendix 1). Semi-structured interview methods were selected to allow for flexibility in the discussion, while also allowing for consistent gathering of data related to points of interest (eg, patient experiences and preferences). Sample questions from the interview guide include "What information would you like to learn more about if you were to watch videos about quitting smoking?" and "Do you think being able to watch videos on your personal device or computer about how to quit smoking could be helpful?" All data were collected via participant interview.

Interviews were conducted until thematic saturation was achieved (ie, no new ideas were discussed, and ideas participants brought up were repetitive), and the average duration was 20 minutes and 30 seconds.

Participants were asked questions about smoking history, previous quit attempts, barriers to quitting smoking, previous experience with and openness to watching videos about quitting smoking, and participant preferences related to video content and presentation. This included asking preferences about the perceived characteristics of people they would like to see present the content (eg, doctor versus peer, man or woman, older or younger person, etc). Interviews were recorded. They were then transcribed by two individuals (a trained medical student and a trained Master of Public Health [MPH] student[TW]) and coded by two coders (the PI [KP] and MPH student [TW]) based on an iteratively developed codebook. Specifically, recurring themes and ideas were identified by the PI and trained MPH student and listed as codes. These codes were refined and new codes were added as the PI and MPH student continued to review interview transcriptions until new themes and ideas were no longer identified. Coders had 86.5% agreement for the original codings, and a trained third coder (research assistant [KM]) was brought in to resolve discrepancies by coding the portion of the interview in which there was disagreement. An inductive thematic analysis was then conducted by the PI. Software was not used to facilitate the analysis.

Ethical Considerations

Study procedures were approved by the Medical University of South Carolina Institutional Review Board (Pro00131304). Patients were called by the PI and given information about the

XSL•FO
RenderX

study and provided verbal consent to participate prior to the interview. Given the low risk of the study, a waiver of written consent was granted by the Medical University of South Carolina Institutional Review Board. Participants received a US $30 Amazon gift card via email for their time. Data were deidentified, as identifying information was redacted in transcriptions.

Results

Participant Characteristics and Smoking History

The study team received information for 17 patients who were then contacted. One patient was ineligible (not actually diagnosed with cancer) and 6 could not be reached. Ten participants were interviewed, and 50% (5/10) reported White racial identity, 30% (3/10) reported Black racial identity, and 10% (1/10) reported biracial identity; 1 participant did not report their race. Sixty percent were women and 40% were men. The mean (SD) age was 55.9 (10.6) years (range 35 - 72 years). Demographics of the current sample were similar to those of individuals who participate in the cancer center's Tobacco Treatment Program (40% Black, 60% women, mean age of 58.8 y) [16]. Cancer diagnoses varied considerably and included breast, cervical, lung, bone, liver, prostate, thyroid, and brain cancers, with half of the participants reporting multiple cancers. Most (7/10) did not know their disease stage. All participants started smoking before age 18 years. The modal number of cigarettes smoked was 20 cigarettes per day (range: 2 - 30).

Methods for Quitting

All participants had made attempts to quit smoking previously. Five out of 10 participants had previously used nicotine patches, 2/10 had used nicotine gum, 2/10 had tried varenicline, and 2/10 tried to quit with no aid (ie, cold turkey). Only 1 participant had tried counseling before, and 1 had called a Quit Line.

Barriers to Quitting

Substance Use

Half (5/10) of the participants endorsed other substance use as a barrier to quitting. Four participants reported that alcohol was a barrier, and 1 participant reported that the previous use of another undisclosed drug had previously been a barrier.

> *There are two things that trigger it: the coffee and the drinking.* [Participant 10]
>
> *The first time I tried to quit, I had quit for three weeks and then we went out drinking for my 21st birthday, and I started back.* [Participant 8]

Negative Affect

Another commonly reported barrier to quitting included negative affect, such as depression (4/10), anxiety (1/10), stress (2/10), or loneliness (1/10).

> *It's just that today was my last day of work. I had a hard day and this [is] just depression not being able to go back to work [due to the cancer diagnosis].* [Participant 4]
>
> *…In between me and the treatments and being extra depressed and lonely and all of this other stuff going on with me, it's been kinda hard to wanna focus on something different.* [Participant 3]

Social Environment

Four of the 10 participants also reported that being around others who smoked (2/4) or being in risky social situations (2/4) was a barrier to quitting and prompted relapses in the past.

> [I was living with my brother and he was a smoker. Participant 5]
>
> [And what happened when you went back to smoking? Is there something in particular that made you go back? Interviewer]
>
> [Being around people that smoked. Participant 4]

Habits

Participants also reported different specific habits or daily routines related to smoking that acted as a barrier to quitting, such as coffee consumption.

> [Make the coffee, let the dog out, light a cigarette. The hardest part is like, um, old behavior and old, uh, routines. Uh, like after I ate or getting up with a coffee in the morning. Um, when I drank alcohol, all that. It pushed me more to want a cigarette. Participant 1]

Past Video Experience and Openness to Watching Videos

Participants were asked if they had ever previously watched videos about quitting smoking, and if they would be open to watching videos in the future. Only 3/10 participants had seen videos about quitting smoking previously. Two participants reported that these videos were commercials; one of these participants stated she had seen commercials for nicotine replacement therapy such as nicotine gum, while the other did not identify the types of commercials they had seen. The third participant who reported seeing videos about quitting smoking stated these videos provided information about how to quit, such as strategies to distract. All participants said they were open to watching videos.

Video Presentation and Content

Most participants also reported not having a preference related to gender (8/10; 2 did not respond to related to gender preferences) or age (8/10; 2 did not respond related to age) of the video presenter. Two participants reported that they would not want to see a younger adult delivering information in the videos. Some participants endorsed several different people they would be open to see presenting information (eg, peer with lived experience or doctor), but the majority endorsed wanting to see a peer with lived experience (8/10).

When asked what information they would like to receive in the videos, many participants had a difficult time generating ideas and there was a wide variety of responses. However, several participants suggested behavioral strategies to avoid smoking, including finding alternative activities (3/10), getting rid of cigarettes (2/10), and replacing the habit of smoking with another activity (2/10).

XSL•FO
RenderX

> [Techniques to replace a habit I guess. Replace it with good habits, ya know? Like taking a walk… taking a walk on the beach, ya know? Participant 1]

Other participants suggested discussing how to be ready "mentally" (2/10) and address emotions that may pose a challenge as they quit smoking (3/10).

> [I think first you gotta prepare your mind mentally to wanna do it. It's like with any addiction, you have to be willing to know that it's bad for you and that it'll end up killing you, and that you have to really be willing to let it go and turn it over. The mood swings and stuff, ya know? Because you do get moody when ya quit. I used to get really angry at everybody. Participant 1]

> [… When I don't have a cigarette I start getting antsy… and then I start that urge to like just, ya know. What's the word… I get aggro! Put it that way. Participant 2]

Of note, only 1 participant mentioned wanting to know more about medications and stated that they wanted to know about the difference between nicotine replacement therapy and cigarettes. Additionally, only 2/10 individuals stated that they would like to hear more about the impact of smoking on their lungs. No other participants endorsed wanting to hear about health impacts, including health benefits of quitting. One participant reported that she did not want to hear about health detriments at all.

> [I mean I'm already eat up with cancer so I don't need to know anything about that, ya know? Participant 4]

Discussion

Principal Findings and Comparison With Previous Works

In semi-structured interviews with cancer survivors who smoked cigarettes, participants indicated that they had previously attempted to quit smoking but had difficulty with making their abstinence last long-term. Participants reported numerous barriers, including routines or habits, other substance use, negative affect, and exposure to smoking in their social environment. This is consistent with previous literature indicating that alcohol use [17], depression [18-20], anxiety [21], and having more people who smoke in their social network [22] reduce the likelihood of someone successfully quitting smoking in non-cancer samples. Some findings are also consistent with previous research conducted among people with cancer. Specifically, individuals diagnosed with cancer with more depressive symptoms are more likely to continue smoking after diagnosis [23,24]. Additionally, among individuals with head and neck cancer, individuals who are exposed to smoking in their social environment (ie, peer or spousal smoking, others smoking at home) are less likely to achieve cessation [25]. Thus, addressing these barriers in smoking cessation interventions tailored for cancer survivors may be important.

When asked about preferences for future video-based interventions, participants largely endorsed not having preferences related to the gender presentation or age of the individual in the video. Participants may not be aware of a preference regarding age or gender of the presenter or do not wish to express that preference due to a social desirability bias. However, most participants did want to see someone with lived experience giving information. A recent systematic review and meta-analysis indicated that interventions that included peer support, particularly those led by individuals who had successfully quit smoking, seem to be helpful in improving smoking abstinence [26]. Additionally, videos created and aired by the Centers for Disease Control as part of the "Tips from Former Smokers" campaign that featured individuals who were affected by smoking-related diseases were effective in increasing awareness of health risks attributable to smoking [27]. It may be that having someone relatable in smoking cessation videos is important to make the information feel salient.

When asked about specific content they would like to see or topics they would like to learn more about, they had a difficult time generating ideas. Ideas that were presented varied and were largely related to behavioral strategies they could achieve on their own rather than information about medication or external resources. Survivors may not know where the gaps in their knowledge lie or may not be able to articulate the knowledge they do have regarding successful smoking cessation strategies that may be helpful for others. There may also be incorrect beliefs about smoking cessation strategies such as medication use that prevented participants from expressing interest in learning more, particularly as it relates to the safety or efficacy of medications like nicotine replacement therapy [28-30]. Previous studies have found that these concerns about nicotine replacement therapy act as a barrier to use [31,32]. While participants in the current study did not voice active opposition to learning more about medication options, misperceptions may have decreased their interest and providing accurate information about medications may increase interest in using them [33].

Overall, participants reported wanting to hear about actionable steps they could take to meet their goal of quitting smoking rather than hearing about the health implications. Many videos publicly available online do discuss the health impact or call for quitting smoking, but do not explain how to actually achieve this [11]. Thus, individuals may be aware that smoking is detrimental for health and can cause cancer, but given that they already have cancer, they may not see the relevance of hearing more about health implications of smoking. Indeed, cancer fatalism, or the belief that death is inevitable when someone has cancer, has been shown to be higher among individuals who smoke compared to those who do not [34], and higher cancer fatalism is associated with an increased likelihood of engaging in risky health behaviors such as smoking [35]. Higher cancer fatalism may also be associated with not seeking cancer-related health information [36]. Thus, it may be that participants do not wish to hear more about health implications of smoking due to fatalistic beliefs. It may be important to present possible health benefits to be gained with smoking cessation, even after a cancer diagnosis, to increase knowledge about the health benefits that come with smoking cessation.

XSL•FO
RenderX

Limitations

While the current study has strengths such as expanding current knowledge about the unique needs and preferences of cancer for a novel tailored smoking cessation intervention, the current study has limitations worth noting. First, while the sample was diverse in terms of racial and gender identity, all participants were recruited from a cancer center at an academic medical center. Individuals receiving care from community clinics or health care providers in rural areas may have different preferences about content for videos for smoking cessation or the feasibility of watching these videos. It is also a small sample size, and may not be generalizable to all individuals with cancer who smoke. Additionally, participants were asked to recall quit attempts and barriers to cessation, which relied on patient memory. Participants may not have been able to accurately recall some of the resources they used to assist in smoking cessation attempts, or reasons for relapse. We also did not gather data related to education, income, or insurance status, which are important considerations in the context of smoking cessation. Lastly, the PI of the study conducted interviews and also served as one of the coders in the thematic analysis, which may introduce bias. However, all interviews were coded by two trained individuals to reduce the risk of bias.

Conclusions

Cancer survivors who smoke are open to novel intervention options such as videos that relay evidence-based information and would like to see peers with lived experience represented in these videos. While participants were able to clearly articulate their own experiences with and barriers to cessation, they had difficulty generating ideas about topics they would like to hear more about in future video interventions. Results from the current study provide considerations for researchers in the future who may be exploring the use of videos for smoking cessation in this vulnerable population. Information presented here, in combination with previously studied interventions and cessation guidelines, may be helpful when planning a video-based intervention. Videos presenting evidence-based information that guides actionable steps may be helpful to inform cancer survivors about quitting smoking while also being low touch and not resource-intensive. A clear next step is to use the current study results to create videos and test their acceptability, feasibility, and efficacy in changing smoking behaviors among cancer survivors, and explore optimal pathways for intervention delivery within current clinical or survivorship pathways.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Semi-structured interview guide used to facilitate telephone interviews.
[DOCX File, 19 KB - formative_v9i1e76792_app1.docx]

References

1. US Department of Health and Human Services. The health consequences of smoking—50 years of progress: a report of the surgeon general. : Centers for Disease Control and Prevention. National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health, Atlanta, GA; 2014 URL: https://www.hhs.gov/sites/default/files/consequences-smoking-exec-summary.pdf [accessed 2025-10-16]
2. Kawahara M, Ushijima S, Kamimori T, et al. Second primary tumours in more than 2-year disease-free survivors of small-cell lung cancer in Japan: the role of smoking cessation. Br J Cancer 1998 Aug;78(3):409-412. [doi: 10.1038/bjc.1998.507] [Medline: 9703291]
3. Demark-Wahnefried W, Aziz NM, Rowland JH, Pinto BM. Riding the crest of the teachable moment: promoting long-term health after the diagnosis of cancer. J Clin Oncol 2005 Aug 20;23(24):5814-5830. [doi: 10.1200/JCO.2005.01.230] [Medline: 16043830]
4. Tseng TS, Lin HY, Moody-Thomas S, Martin M, Chen T. Who tended to continue smoking after cancer diagnosis: the national health and nutrition examination survey 1999-2008. BMC Public Health 2012 Sep 14;12(1):1-9. [doi: 10.1186/1471-2458-12-784] [Medline: 22974404]
5. Gritz ER, Fingeret MC, Vidrine DJ, Lazev AB, Mehta NV, Reece GP. Successes and failures of the teachable moment: smoking cessation in cancer patients. Cancer 2006 Jan 1;106(1):17-27. [doi: 10.1002/cncr.21598] [Medline: 16311986]
6. Li WHC, Chan SSC, Lam TH. Helping cancer patients to quit smoking by understanding their risk perception, behavior, and attitudes related to smoking. Psychooncology 2014 Aug;23(8):870-877. [doi: 10.1002/pon.3486] [Medline: 24493624]
7. Krutz M, Acharya P, Elliott J, Zhao D, Mhawej R, Queimado L. Tobacco cessation following laryngeal cancer diagnosis predicts response to treatment and laryngectomy-free survival. Otolaryngol Head Neck Surg 2024 Feb;170(2):438-446. [doi: 10.1002/ohn.567] [Medline: 37890055]
8. Services, smoking cessation: a report of the surgeon general. : U.S. Department of Health and Human Services; 2020 URL: https://www.hhs.gov/sites/default/files/2020-cessation-sgr-full-report.pdf [accessed 2025-10-16]
9. National cancer institute. Cancer Survivors and Smoking. 2024 Mar. URL: https://progressreport.cancer.gov/after/smoking [accessed 2024-04-04]

XSL•FO
RenderX

10. Yingst JM, Carrillo M, Chan KH, et al. Effectiveness of smoking cessation interventions among persons with cancer: A systematic review. Psychooncology 2023 Aug;32(8):1147-1162. [doi: 10.1002/pon.6171] [Medline: 37226331]
11. Backinger CL, Pilsner AM, Augustson EM, Frydl A, Phillips T, Rowden J. YouTube as a source of quitting smoking information. Tob Control 2011 Mar;20(2):119-122. [doi: 10.1136/tc.2009.035550] [Medline: 20966132]
12. Deshpande N, Wu M, Kelly C, et al. Video-based educational interventions for patients with chronic illnesses: systematic review. J Med Internet Res 2023 Jul 19;25:e41092. [doi: 10.2196/41092] [Medline: 37467015]
13. Stanczyk NE, de Vries H, Candel M, Muris JWM, Bolman CAW. Effectiveness of video- versus text-based computer-tailored smoking cessation interventions among smokers after one year. Prev Med 2016 Jan;82:42-50. [doi: 10.1016/j.ypmed.2015.11.002] [Medline: 26577867]
14. Sharma-Kumar R, Puljević C, Morphett K, Meurk C, Gartner C. The acceptability and effectiveness of videos promoting smoking cessation among australians experiencing mental illness. Health Educ Behav 2022 Jun;49(3):506-515. [doi: 10.1177/10901981211034738] [Medline: 34496656]
15. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014 Sep;89(9):1245-1251. [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]
16. Cummings KM, Toll BA, Talbot V, et al. Implementation, enrollment, and engagement in an opt-out telehealth pharmacist-assisted tobacco treatment program for patients seen in oncology outpatient clinics. Cancer 2024 Jul 15;130(14):2482-2492. [doi: 10.1002/cncr.35291] [Medline: 38546445]
17. van Amsterdam J, van den Brink W. The effect of alcohol use on smoking cessation: a systematic review. Alcohol 2023 Jun;109:13-22. [doi: 10.1016/j.alcohol.2022.12.003] [Medline: 36690220]
18. Heckman BW, Carpenter MJ, Correa JB, et al. Effects of experimental negative affect manipulations on ad libitum smoking: a meta-analysis. Addiction 2015 May;110(5):751-760. [doi: 10.1111/add.12866] [Medline: 25641624]
19. Lawrence D, Mitrou F, Zubrick SR. Smoking and mental illness: results from population surveys in Australia and the United States. BMC Public Health 2009 Aug 7;9:1-14. [doi: 10.1186/1471-2458-9-285] [Medline: 19664203]
20. Cook BL, Wayne GF, Kafali EN, Liu Z, Shu C, Flores M. Trends in smoking among adults with mental illness and association between mental health treatment and smoking cessation. JAMA 2014 Jan 8;311(2):172-182. [doi: 10.1001/jama.2013.284985] [Medline: 24399556]
21. Becoña E, Vázquez FL, del Carmen Mı´guez M. Smoking cessation and anxiety in a clinical sample. Pers Individ Dif 2002 Feb;32(3):489-494. [doi: 10.1016/S0191-8869(01)00050-2]
22. Blok DJ, de Vlas SJ, van Empelen P, van Lenthe FJ. The role of smoking in social networks on smoking cessation and relapse among adults: A longitudinal study. Prev Med 2017 Jun;99:105-110. [doi: 10.1016/j.ypmed.2017.02.012] [Medline: 28216381]
23. Berg CJ, Thomas AN, Mertens AC, et al. Correlates of continued smoking versus cessation among survivors of smoking-related cancers. Psychooncology 2013 Apr;22(4):799-806. [doi: 10.1002/pon.3077] [Medline: 22488864]
24. Park ER, Japuntich SJ, Rigotti NA, et al. A snapshot of smokers after lung and colorectal cancer diagnosis. Cancer 2012 Jun 15;118(12):3153-3164. [doi: 10.1002/cncr.26545] [Medline: 22271645]
25. Kashigar A, Habbous S, Eng L, et al. Social environment, secondary smoking exposure, and smoking cessation among head and neck cancer patients. Cancer 2013 Aug 1;119(15):2701-2709. [doi: 10.1002/cncr.28088] [Medline: 23765604]
26. Yuan P, Westmaas JL, Thrul J, Toussaert S, Hilton JF, White JS. Effectiveness of peer-support interventions for smoking cessation: a systematic review and meta-analysis. Nicotine Tob Res 2023 Aug 19;25(9):1515-1524. [doi: 10.1093/ntr/ntad059] [Medline: 37042206]
27. Huang LL, Thrasher JF, Abad EN, et al. The U.S. national tips from former smokers antismoking campaign: promoting awareness of smoking-related risks, cessation resources, and cessation behaviors. Health Educ Behav 2015 Aug;42(4):480-486. [doi: 10.1177/1090198114564503] [Medline: 25588934]
28. Smith SY, Curbow B, Stillman FA. Harm perception of nicotine products in college freshmen. Nicotine & Tobacco Res 2007 Sep;9(9):977-982. [doi: 10.1080/14622200701540796]
29. Wikmans T, Ramström L. Harm perception among Swedish daily smokers regarding nicotine, NRT-products and Swedish Snus. Tob Induc Dis 2010 Aug 13;8(1):9. [doi: 10.1186/1617-9625-8-9] [Medline: 20707898]
30. Bansal M, Cummings KM, Hyland A, Giovino G. Stop-smoking medications: who uses them, who misuses them, and who is misinformed about them? Nicotine & Tobacco Res 2004 Dec;6(6):303-310. [doi: 10.1080/14622200412331320707]
31. Balmford J, Borland R, Hammond D, Cummings KM. Adherence to and reasons for premature discontinuation from stop-smoking medications: data from the ITC four-country survey. Nicotine & Tobacco Research 2011 Feb 1;13(2):94-102. [doi: 10.1093/ntr/ntq215]
32. Silla K, Beard E, Shahab L. Nicotine replacement therapy use among smokers and ex-smokers: associated attitudes and beliefs: a qualitative study. BMC Public Health 2014 Dec 22;14(1):1311. [doi: 10.1186/1471-2458-14-1311] [Medline: 25535404]
33. Ferguson SG, Gitchell JG, Shiffman S, Sembower MA, Rohay JM, Allen J. Providing accurate safety information may increase a smoker's willingness to use nicotine replacement therapy as part of a quit attempt. Addict Behav 2011 Jul;36(7):713-716. [doi: 10.1016/j.addbeh.2011.02.002] [Medline: 21371825]

XSL•FO
RenderX

34. Hall MB, Vos P. Comparison of cancer fatalism among rural smokers and nonsmokers. J Community Health 2019 Apr;44(2):215-221. [doi: 10.1007/s10900-018-0576-z] [Medline: 30267351]
35. Zaidi M, Sarkar S, Arakelyan S, Poghosyan H. Relationship between fatalistic cancer beliefs and risky health behaviors. West J Nurs Res 2024 Oct;46(10):757-765. [doi: 10.1177/01939459241273388] [Medline: 39161288]
36. Kobayashi LC, Smith SG. Cancer fatalism, literacy, and cancer information seeking in the American public. Health Educ Behav 2016 Aug;43(4):461-470. [doi: 10.1177/1090198115604616] [Medline: 26377524]

Abbreviations

MPH: Master of Public Health
PI: private investigator
TTS: tobacco treatment specialist

Edited by J Sarvestan; submitted 30.04.25; peer-reviewed by K Wiseman, N Machado; revised version received 03.10.25; accepted 05.10.25; published 23.10.25.

Please cite as:
Pebley K, Williams T, Moody K, Rojewski AM
Thematic Analysis of Smoking Cessation and Future Cessation Interventions for Cancer Survivors: Convenience Sampling Study
JMIR Form Res 2025;9:e76792
URL: https://formative.jmir.org/2025/1/e76792
doi:10.2196/76792

© Kinsey Pebley, Tianna Williams, Kathryn Moody, Alana M Rojewski. Originally published in JMIR Formative Research (https://formative.jmir.org), 23.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Eye and Systemic Disease Management Changes After Teleophthalmology Screening in Primary Care: Retrospective Cross-Sectional Pilot Study of 200 Consecutive Patients

Alexander Pinhas[1,2,3], MD; Boris Pinhas[4], BA; Egor Dmitruk[2], BA; Stella Pinhas[4], MD

[1]Department of Ophthalmology, Icahn School of Medicine at Mount Sinai, 1 Gustave L. Levy Place, New York, NY, United States

[2]myeye.healthcare LLC, Forest Hills, NY, United States

[3]Retina & Ophthalmology Services of New York PLLC, Elmhurst, NY, United States

[4]Internal Medicine & Geriatrics PLLC, Elmhurst, NY, United States

Corresponding Author:
Alexander Pinhas, MD
Department of Ophthalmology, Icahn School of Medicine at Mount Sinai, 1 Gustave L. Levy Place, New York, NY, United States

Abstract

Background: Undiagnosed ocular diseases and ocular complications from systemic diseases are common in primary care populations, and many can be detected through retinal imaging before symptoms develop. Asynchronous store-and-forward teleophthalmology offers a scalable way to integrate eye screening into primary care, yet its broader impact beyond diabetes and diabetic retinopathy detection remains underexplored.

Objective: This study evaluated the outcomes of asynchronous store-and-forward teleophthalmology screening in a primary care clinic, including detection and triage of ocular conditions and subsequent changes in eye and systemic management.

Methods: This was a retrospective cross-sectional analysis of the first 200 patients screened in a single primary care clinic in Elmhurst, New York, between January and May 2025. Each patient underwent nonmydriatic external and posterior eye imaging, which was reviewed by a remote reading eye clinician. Reports included eye findings, triage decisions (routine monitoring vs in-person referral), and management recommendations. Subsequent changes in care were extracted from primary care and in-person specialist consult notes.

Results: Of 200 patients (mean age 62.1, SD 19.0, range 11‑100 years), 71.5% (143/200, 95% CI 64.9-77.3) had positive eye findings, and 40% (80/200, 95% CI 33.5‑46.9) were referred for in-person eye examinations. Only 8.8% (7/80, 95% CI 4.3-17.0) of referrals were for diabetic retinopathy; most were for glaucoma suspects, age-related macular degeneration, cataracts, and other eye diseases. Image quality was high, with 98.2% (390/397, 95% CI 96.4-99.1) of fundus images being at least partially adequate. Of the 32 patients with documented in-person eye follow-up, 87.5% (28/32) of evaluations confirmed the screening findings. Eye management changes were initiated in 11 patients, whereas systemic management changes occurred in 70 patients, including new prescriptions for Age-Related Eye Disease Study 2 supplements, antihypertensives, diabetes medications, and lipid-lowering agents.

Conclusions: Asynchronous teleophthalmology screening in a primary care setting effectively identified both ocular diseases and ocular complications from systemic diseases, leading to meaningful changes in eye and systemic management. The low rate of diabetic retinopathy among referrals highlights the broader diagnostic value of retinal imaging beyond diabetes management. This care model offers a scalable, high-yield strategy for proactive disease detection and interdisciplinary intervention at the primary care level.

(JMIR Form Res 2025;9:e81918) doi:10.2196/81918

KEYWORDS

teleophthalmology; asynchronous telemedicine; primary health care; diabetic retinopathy screening; age-related macular degeneration; interdisciplinary care

XSL•FO
RenderX

Introduction

Background

The US population is steadily growing and aging, with projections rising from 333 million in 2020 to 355 million by 2030. Adults aged 65 and older are expected to comprise 20% of the population by then, up from 17% in 2020 [1]. As a result, the prevalence of chronic medical conditions, including hypertension, diabetes mellitus, dyslipidemia, and heart disease, is projected to reach 171 million [2], while chronic eye diseases such as cataracts, diabetic retinopathy, glaucoma, and age-related macular degeneration are expected to affect 58 million individuals by 2030 [3]. In response to these trends, the CDC has called for a shift from reactive care to proactive, preventive healthcare [4]. This underscores the urgent need for scalable, cost-effective strategies for early detection, triage, and intervention in both systemic and ocular disease.

As a myriad of systemic diseases, including common cardiovascular and neurodegenerative disorders, manifest in the eyes, ocular biomarkers have long been proposed as noninvasive tools for detecting both ocular and systemic conditions [5]. Landmark studies such as the Diabetes Control and Complications Trial, the UK Prospective Diabetes Study, and the UK Prospective Diabetes Study-Hypertension in Diabetes Study demonstrated this potential, using serial color fundus photography to show that intensive glycemic and blood pressure control reduces microvascular complications in diabetes [6-8]. In 2001, EyePACS launched one of the earliest asynchronous store-and-forward teleophthalmology programs, enabling non–eye care settings to capture and upload fundus images for remote interpretation of diabetic retinopathy [9]. To support this model, the Centers for Medicare & Medicaid Services (CMS) and the American Medical Association introduced Current Procedural Terminology (CPT) codes 92227 and 92228 in 2011 for remote diabetic retinopathy screening in primary care. The National Committee for Quality Assurance Healthcare Effectiveness Data and Information Set also recognized nonmydriatic fundus photography that covers the macula and optic nerve as an acceptable form of eye examination in diabetics. However, early adoption was limited, likely because of insufficient imaging infrastructure and the lack of widespread, integrated telehealth ecosystems.

The ideas of remote eye screening and oculomics (the study of ocular biomarkers for systemic disease) have recently embarked on a renaissance because of advances in smaller, automated, and more affordable eye imaging technologies, such as color fundus cameras, alongside developments in data science, cloud computing, and artificial intelligence (AI) [10]. Between 2018 and 2024, 3 autonomous AI algorithms (ie, IDx-DR, EyeArt, and AEYE-DS) capable of diabetic retinopathy screening via color fundus photography without clinician input received Food and Drug Administration (FDA) approval [11-16]. In response, CMS approved CPT code 92229 in 2021, covering point-of-care AI-based retinal imaging. Despite this, real-world adoption has been limited. By 2023, the use of CPT 92229 remained low (59 per 100,000), significantly trailing the traditional remote interpretation codes 92227 and 92228 (156 per 100,000), which have nearly tripled in usage since 2021. Still, even these remain far below conventional in-clinic fundus photography (CPT 92250), which had a 6 times higher usage rate of 897 per 100,000 in 2023 [17].

Adoption of autonomous AI for diabetic retinopathy screening faces several barriers. Economically, the appeal is limited because only approximately 15% to 20% of the primary care population has diabetes [18,19], making it difficult for non–eye clinics to justify investing in imaging equipment solely for this purpose [20]. The expansion of indications for AI beyond diabetic retinopathy screening is pushing forward, with glaucoma, age-related macular degeneration, retinopathy of prematurity, and ocular oncology identified as priority areas [21]. Recent studies also highlight the growing potential of retinal imaging combined with AI to detect and risk stratify a range of systemic chronic conditions through the eye, including hypertension [22], diabetes mellitus [23], metabolic syndrome [24], atherosclerotic cardiovascular disease [25,26], chronic kidney disease [27], and neurodegenerative disorders such as Alzheimer dementia [28-30]. However, each new indication requires separate FDA approval tied to specific devices, involving time- and cost-intensive clinical trials. In addition, fully autonomous algorithms often yield higher false-positive rates when no human oversight is involved, potentially increasing unnecessary referrals and placing further strain on already limited eye care resources [31,32].

Between 2021 and 2022, CMS expanded the range of International Classification of Diseases, 10th Revision, codes eligible for CPT codes 92227 and 92228, reflecting growing evidence that tele-ophthalmology can detect a variety of retinal diseases beyond diabetic retinopathy. The COVID-19 pandemic further accelerated telehealth adoption, prompting rapid regulatory shifts, expanded reimbursement, and widespread integration of remote care technologies [33]. Other CPT codes such as 99451, which reimburse for written virtual interprofessional consultations, were allowed to be used alongside CPT code 92228 to support human-led virtual eye care coordination [34]. These policy shifts likely contributed to the observed 3-fold increase in the usage of CPT codes 92227 and 92228 since 2021.

The asynchronous store-and-forward model, where images are captured in non–eye clinics, uploaded to the cloud, and interpreted by remote eye clinicians, has proven both clinically effective and economically viable [34,35]. Unlike autonomous AI, this model allows broader screening across ocular pathologies and may yield higher specificity, reducing unnecessary referrals. Emerging research suggests that co-piloting, or combining assistive AI with human oversight, offers the most balanced performance in sensitivity, specificity, and cost-effectiveness [32]. As the field evolves, the optimal future may lie in hybrid models that integrate both approaches. For now, asynchronous human-led screening provides a practical and scalable solution.

Objectives

This study aimed to evaluate the outcomes of a pilot asynchronous teleophthalmology program in a primary care setting, assessing the prevalence of ocular and systemic findings

XSL•FO
RenderX

beyond diabetic retinopathy. In addition, it examined how screening results influenced subsequent eye and systemic disease management.

Methods

Overview

This was a retrospective, cross-sectional, descriptive pilot study reviewing the first 200 consecutive patients enrolled in a teleophthalmology program at a single primary care clinic (Internal Medicine & Geriatrics PLLC, Elmhurst, NY) between January and May 2025 (Figure 1). Reporting follows Strengthening the Reporting of Observational Studies in Epidemiology, and a completed checklist with page references is provided (Checklist 1). Six weeks were allowed to pass after the 200th virtual eye consult report was received, before performing the chart review. Data sources included virtual eye consult reports by a single remote reading eye clinician, as well as follow-up documentation in non–eye clinician progress notes and in-person eye consult records. No additional inclusion or exclusion criteria were applied.

XSL•FO
RenderX

Figure 1. Asynchronous store-and-forward tele-ophthalmology workflow diagram.

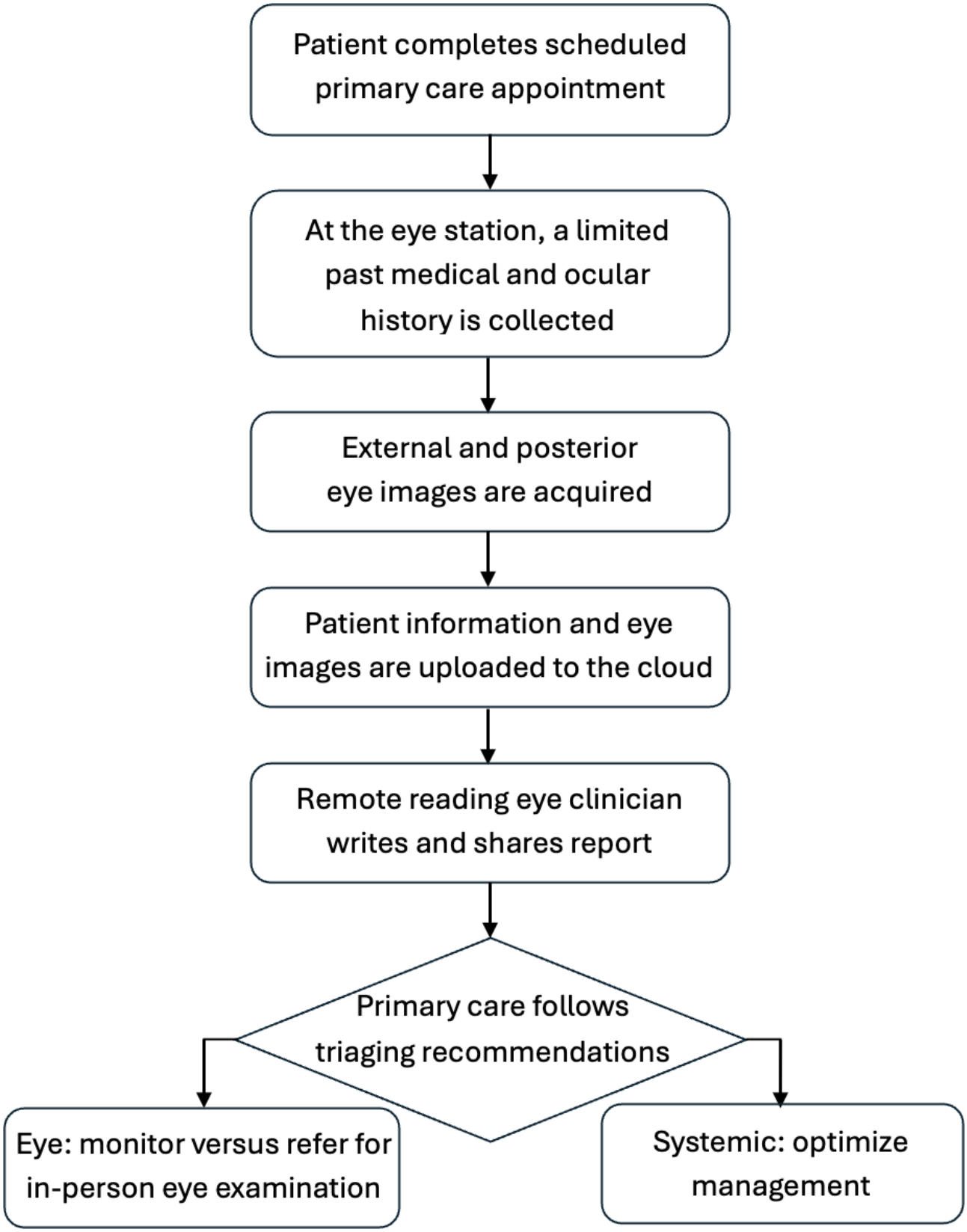

Ethical Considerations

This study was reviewed and granted exempt status by the Castle Institutional Review Board (IRB), as only deidentified patient data were analyzed. Castle IRB is an independent IRB, which has full accreditation with the Association for the Accreditation of Human Research Protection Programs and is registered with both the US Office for Human Research Protections and the FDA as IRB Organization number: IORG0010151 and IRB Registration number: IRB00012054. As deidentified patient data were used in this retrospective review, informed consent was not required, and no compensation was provided to patients.

XSL•FO
RenderX

The study adhered to the tenets of the Declaration of Helsinki and all applicable Health Insurance Portability and Accountability Act regulations. It was ensured that no identification of individual patients in any images of the manuscript is possible.

Variables

Descriptive variables in this study included age (years, continuous), sex (female, male), race or ethnicity (White, Asian, Hispanic, Black), past medical history (diabetes mellitus, hypertension, dyslipidemia, obesity), self-reported past ocular history (cataract, diabetic retinopathy, glaucoma, age-related macular degeneration), and active visual symptoms at screening (decreased vision, floaters, or flashes).

Prespecified key outcomes with 95% Wilson CIs included image quality adequacy, positive eye findings, referral for in-person eye examination, distribution of referable eye diagnoses, and systemic management changes initiated. Other descriptive outcomes included in-person eye referral completion, screening concordance with in-person eye consultations, in-person non–eye referral completion, eye management changes initiated, and patient compliance with new plans. Number needed to screen (NNS=1/proportion) for select outcomes were calculated, including referral for in-person eye examination and systemic management changes initiated. We did not perform adjusted or interaction analyses because this pilot was descriptive and not designed to test associations.

Sample Size Rationale

The sample size was predefined to ensure acceptable descriptive precision for key proportions (eg, positive screens and referral rate) with an acceptable margin of error between 5% and 10% at 95% CI. Using $n=[z^2p(1-p)]/E^2$, where n is the sample size, z is 1.96 at 95% CI, and the conservative anticipated proportion p is 0.50, our target margin of error (E) between 5% and 10% implied a necessary sample size between 97 (for E=10%) and 385 (for E=5%). The sample size of 200 was chosen as a pragmatic midpoint.

Virtual Eye Consult Reports

All virtual eye consults in this pilot were interpreted by 1 remote eye clinician, a board-certified ophthalmologist fellowship trained in medical and surgical diseases of the vitreous and retina. The reader had 4 years of clinical experience, with routine use of color fundus photography in practice. The reader applied prespecified diagnostic (International Clinical Diabetic Retinopathy [ICDR] for diabetic retinopathy, Wong–Mitchell for hypertensive retinopathy, Age-Related Eye Disease Study [AREDS] for age-related macular degeneration, and standard optic nerve features for glaucoma suspects) and referral criteria as described in the Referable Eye Disease Definitions section. The reader had access to basic demographics and past medical and ocular history embedded in the consult request. The reader was unaware of later in-person eye examination findings.

The virtual eye consult reports included demographic information (age, sex, race/ethnicity), past medical history (diabetes mellitus, hypertension, dyslipidemia, obesity), self-reported past ocular history (cataracts, diabetic retinopathy, glaucoma, age-related macular degeneration), and active visual symptoms at screening (decreased vision, floaters, or flashes). Each report featured a photo analysis section assessing image quality and documenting image findings based on 1 nonmydriatic 45° center-nasal single-field color fundus photograph (centered between the fovea and the disc center) and 1 external eye photograph per eye (DRSplus; iCare USA Inc, Raleigh, North Carolina). The reports concluded with an assessment and plan, including a problem list, management recommendations for the primary care clinician, and guidance on whether in-person eye or systemic referrals were warranted.

Image Quality Review

The virtual eye consult reports included an assessment of image quality, classified as adequate, partially adequate, or inadequate. Grading was subjective and followed criteria adapted from prior studies [36]. Quality consideration considered four main domains: sharpness, illumination, field of view, and artifacts. Sharpness was assessed based on image focus. For external images, this included iris patterns and eyelash clarity, whereas for posterior images, it included foveal detail, tertiary retinal vessel, and cup and disc margin clarity. Illumination was deemed suboptimal if dark or overexposed areas interfered with visualization of standard anatomical features. Field of view was considered adequate when both the foveal center and the entire optic nerve head were within the field of imaging. Artifacts included dust spots, sensor defects, or obstruction by lids and lashes that impaired image interpretation.

Referable Eye Disease Definitions

Overview

The remote reading eye clinician used standardized criteria to determine whether a patient should be referred for an in-person eye examination. Definitions for each condition were as follows.

Cataracts

External eye color photographs were used to screen for cataracts, although evaluation was limited by nonmydriatic imaging. As a result, only more advanced cataracts, typically dense, brunescent, or white, were reliably detected and considered referable.

Diabetic Retinopathy

Referable diabetic retinopathy was defined as more-than-mild nonproliferative diabetic retinopathy, which is defined by the ICDR scale as the presence of signs beyond microaneurysms [37]. This includes dot-blot hemorrhages, exudates, venous beading, intraretinal microvascular abnormalities, neovascularization, preretinal hemorrhages, and macular edema. Macular edema was defined as exudates or apparent thickening within one-disc diameter of the fovea.

Hypertensive Retinopathy

Referable hypertensive retinopathy was defined as more-than-mild hypertensive retinopathy, which is defined by the Wong and Mitchell scale as the presence of signs beyond arteriolar narrowing, copper wiring, and arteriovenous nicking [38]. This includes dot-blot hemorrhages, flame-shaped

XSL•FO
RenderX

hemorrhages, microaneurysms, cotton-wool spots, hard exudates, and optic disc swelling.

Glaucoma Suspect

As glaucoma grading requires imaging and testing beyond fundus photography, including intraocular pressure measurement, visual fields, and optical coherence tomography retinal nerve fiber measurements, screening with fundus photography was only able to detect glaucoma suspects (unless the patient self-reported a history of glaucoma).

Glaucoma suspect, or referable possible glaucoma, was defined as a large cup-to-disc ratio (a ratio of vertical cup to disc diameter of 0.6 or greater), asymmetry in cup-to-disc ratio between eyes (a difference >0.2), neuroretinal rim thinning or notching, optic disc pallor, and optic disc hemorrhages (eg, Drance hemorrhages) [39].

Age-Related Macular Degeneration

Referable age-related macular degeneration was defined using the AREDS grading system as the presence of more-than-mild age-related macular degeneration, including at least 20 medium-sized drusen (63‑124 microns), at least one large drusen (≥125 microns), pigmentary changes associated with large drusen, central or noncentral geographic atrophy, and neovascular membranes [40].

Other Conditions

For other pathologies, clinical judgment guided referral decisions. Patients were also referred if they reported concerning active visual symptoms, such as decreased vision, floaters, or flashes.

Referrals Based on Image Quality

If image quality was deemed partially adequate or inadequate, and referable disease could not be confidently ruled out, patients were referred for in-person eye evaluation.

Primary Care Clinician Progress Notes and In-Person Consult Notes

Progress notes from the primary care clinician and notes from in-person eye and noneye consultations were reviewed to determine whether an eye referral was made and completed, whether a noneye referral was made and completed, whether the initial remote report findings were confirmed by the referring clinician or consulting providers, whether changes were made to eye management, whether changes were made to systemic management, and whether the patient was compliant with recommended management changes.

Bias and Data Quality

We took several steps to limit common biases in this retrospective cross-sectional pilot. To minimize selection bias, we used consecutive sampling of patients with no post hoc exclusions. We also used a fixed abstraction window, beginning chart review 6 weeks after imaging the 200th patient. To minimize classification bias, standardized protocols were used to acquire images, grade image quality, and grade ocular findings and referability. When disease could not be ruled out because of poor image quality, patients were referred, favoring sensitivity over missed disease. To minimize observer bias and incorporation bias, the remote reader rendered reports before any in-person eye referral outcomes existed.

To ensure data abstraction quality, we used a structured abstraction template with explicit variable definitions and maintained a data provenance log. Summary counts and CIs were independently checked by a second author. Effort was taken not to count missing data as an outcome verification. Concordance with in-person assessments was calculated only where consultant notes were available. Referral completion and compliance were counted as completed, not completed, or unknown rather than treating missing data as negative. As this pilot was descriptive with small event counts and incomplete follow-up for some referrals, we did not perform adjusted or interaction analyses, reducing the risk of overfitting or spurious inference.

Results

Demographics and Systemic Disease Prevalence

The study included 200 patients, with a mean age of 62.1 years (SD 19.0; range: 11‑100 years). Of these, 122 (61%) were female, and 78 (39%) were male. Self-identified race or ethnicity was as follows: 140 (70%) White, 31 (15.5%) Asian, 26 (13%) Hispanic, and 3 (1.5%) Black. Of the 200 patients, 52 (26%) had type 2 diabetes mellitus, 97 (48.5%) had hypertension, 146 (73%) had dyslipidemia, and 35 (17.5%) were classified as obese (Table 1). Self-reported past ocular histories included 18 (9%) patients reporting a history of cataract, 38 (19%) patients reporting having had cataract surgery (this group was mutually exclusive from the history of cataract group), 3 (1.5%) patients reporting a history of diabetic retinopathy, 11 (5.5%) patients reporting a history of glaucoma, 4 (2%) patients reporting a history of age-related macular degeneration, and 13 (6.5%) patients reporting having some active visual symptoms.

XSL•FO
RenderX

Table . Demographics and systemic disease prevalence.

Variable	Results (N=200 patients)
Age (years)	
Mean (SD)	62.1 (19.0)
Range	11 - 100
Sex, n (%)	
Female	122 (61.0)
Male	78 (39.0)
Race/ethnicity, n (%)	
White	140 (70.0)
Asian	31 (15.5)
Hispanic	26 (13.0)
Black	3 (1.5)
Past medical history, n (%)	
Diabetes mellitus	52 (26.0)
Hypertension	97 (48.5)
Dyslipidemia	146 (73.0)
Obesity	35 (17.5)

Imageability Metrics

Strict imageability was defined as the proportion of images subjectively graded as adequate out of the total number of images assessed. Lenient imageability included both adequate and partially adequate images. For external eye images, 400 images were expected (2 per eye×200 patients), but 3 were missing due to monocular status. Of the 397 images obtained, 396 were adequate, and 1 was partially adequate, resulting in strict imageability of 99.7% (95% CI 98.6-100) and lenient imageability of 100% (95% CI 99.0-100). For posterior eye images, 400 were expected, with the same 3 missing due to monocular status. Of the 397 images obtained, 342 were adequate, 48 were partially adequate, and 7 were inadequate. This resulted in a strict imageability of 86.1% (95% CI 82.4-89.2) and a lenient imageability of 98.2% (95% CI 96.4-99.1; Table 2; Figure 2).

Table . Eye screening and triaging key outcomes.

Outcome	Images (N1=397) and patients (N2=200), n (%; 95% CI).	NNS[a] (95% CI)
Posterior image lenient adequacy	390 (98.2; 96.4-99.1)	-[b]
Positive eye findings	143 (71.5; 64.9-77.3)	-
Referred for in-person eye examination	80 (40.0; 33.5-46.9)	2.50 (2.13-2.99)
Positive eye findings not referred	63 (31.5; 25.5-38.2)	-
Systemic management changes	70 (35.0; 28.7-41.8)	2.86 (2.39-3.48)

[a] NNS: number needed to screen.

[b] -: not applicable.

XSL•FO
RenderX

Figure 2. Image quality assessment examples of posterior eye images. The first column showcases adequate images, the second column showcases partially adequate images, and the third column showcases inadequate images. Panel A shows the average image, without any defects. Panel B shows an illumination defect with an inferior shadow that is not too dark, still allowing discernment of blood vessels in the area. Panel C shows a sharpness defect with mild blurriness that still allows discernment of anatomic features. Panel D shows a sharpness defect with moderate blurriness that makes it more difficult to discern anatomic features. Panel E shows an illumination defect with an inferior shadow that is darker, making discernment of blood vessels in the area difficult. Panel F shows a field of view defect, where the fovea is visible but the optic nerve head is not visible. Panel G shows an illumination defect with an inferior shadow that covers visibility of the fovea and optic nerve head. Panel H shows an inferior illumination defect and a superior sharpness defect. Panel I shows an artifact defect allowing only a small portion of the posterior segment to be visualized. The bottom of each image is labeled with age and sex of the patient, and eye laterality (M=male; F=female; OD=right eye; OS=left eye).

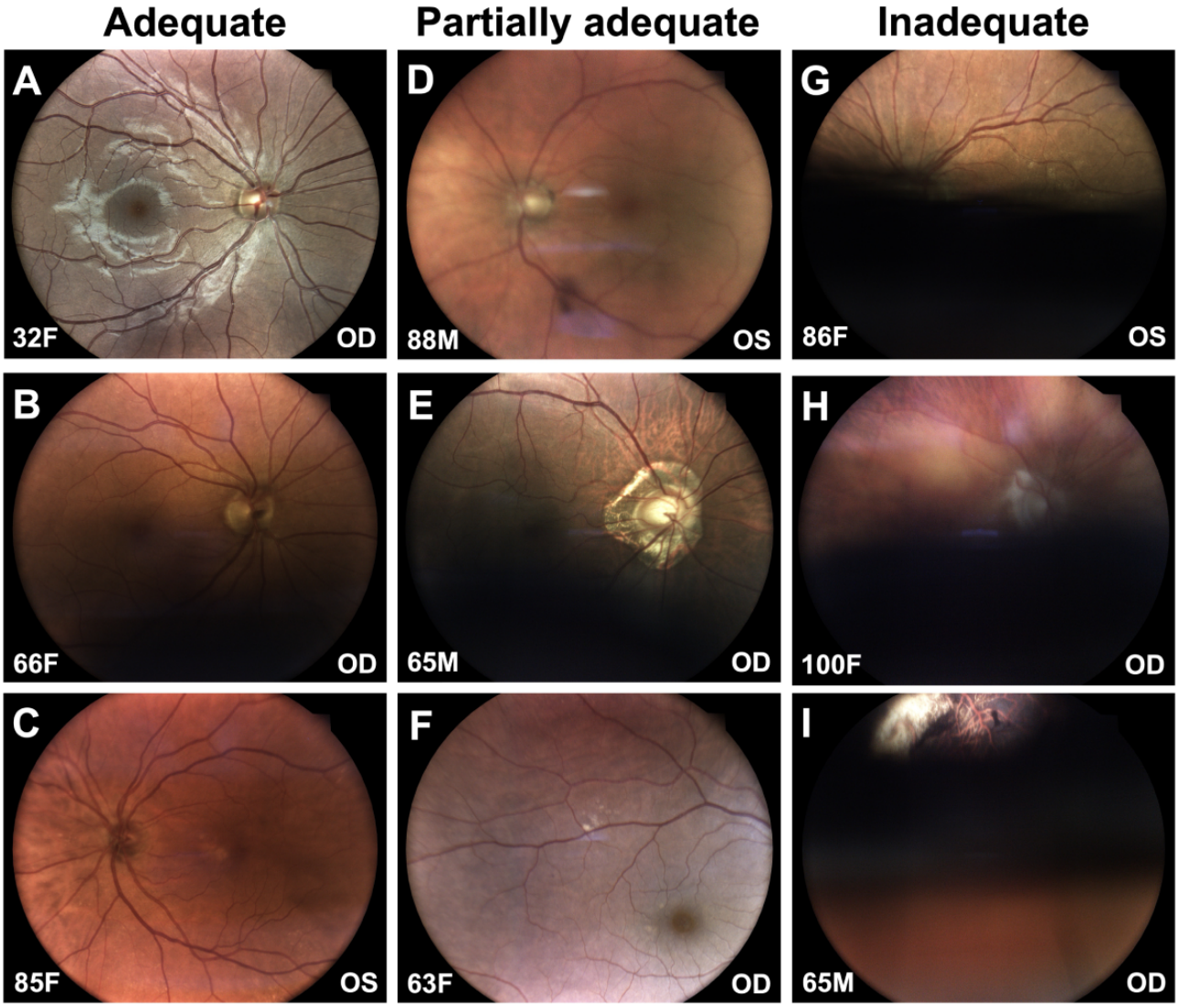

Eye Screening and Triaging Outcomes

The following percentages and 95% CIs are presented relative to all 200 patients screened. Of the 200 patients screened, 57 (28.5%; 95% CI 22.7-35.1) had negative eye screens, whereas 143 (71.5%; 95% CI 64.9-77.3) had positive findings (Table 2). Among those with positive findings, 76 (38%; 95% CI 31.6-44.9) had 1 finding, 48 (24.0%; 95% CI 18.6-30.4) had 2 findings, and 19 (9.5%; 95% CI 6.2%-14.4%) had 3 findings. A total of 80 (40%) patients (95% CI 33.5 - 46.9) had referable disease and were referred for in-person eye examinations, corresponding to screening 2.5 patients per referral (NNS 2.50; 95% CI 2.13-2.99; Table 2). The remaining 63 (31.5%) patients (95% CI 25.5-38.2) with positive findings that did not meet the threshold for referral were managed by the primary care clinician and scheduled for annual remote monitoring.

There were 62 cases (31%; 95% CI 25.0-37.7) of hypertensive retinopathy (1 referable, 0.5%; 95% CI 0.1-2.8), 17 cases (8.5%; 95% CI 5.4-13.2) of diabetic retinopathy (7 referable, 3.5%; 95% CI 1.7-7.0), 47 cases (23.5%; 95% CI 18.2-29.8) of age-related macular degeneration (18 referable, 9.0%; 95% CI 5.8-13.8), 27 cases (13.5%; 95% CI 9.4-18.9) of glaucoma suspect (26 referable, 13.0%; 95% CI 9.0-18.4), 10 cases of referable cataract (5.0%; 95% CI 2.7-9.0), and 66 cases (33.0%; 95% CI 26.9-39.8) of other findings (33 referable, 16.5%; 95% CI 12.0-22.3; Figures 3 and 4). Of note, these categories were not mutually exclusive, so percentages sum to more than 100%. Other findings included dry eye, pterygium, xanthelasma, age-related ptosis, ectropion, iris nevus, retinal vascular occlusion, long-term hydroxychloroquine use, choroidal nevus, myopic degeneration, epiretinal membrane, vascular changes in nondiabetics/nonhypertensives, retinal aneurysm, blurring

XSL•FO
RenderX

of disc margins, nonglaucomatous optic neuropathy, complaints of subjective visual disturbances (including floaters, flashes, and decreased vision), and low vision or monocular status. Remote eye clinician findings did not align with 13 self-reported cataracts, likely due to the difficulty assessing for cataract in the nonmydriatic setting.

Notably, only 7 (8.8%) of the 80 referrals (95% CI 4.3-17.0) were for diabetic retinopathy. Comparing remote reading eye clinician findings to self-reported past ocular histories, patients were unaware of their eye conditions in 193 (84.3%) of the 229 unique findings (95% CI 79.0-88.4), highlighting the value of proactive screening.

Figure 3. Distribution of eye findings and referability.

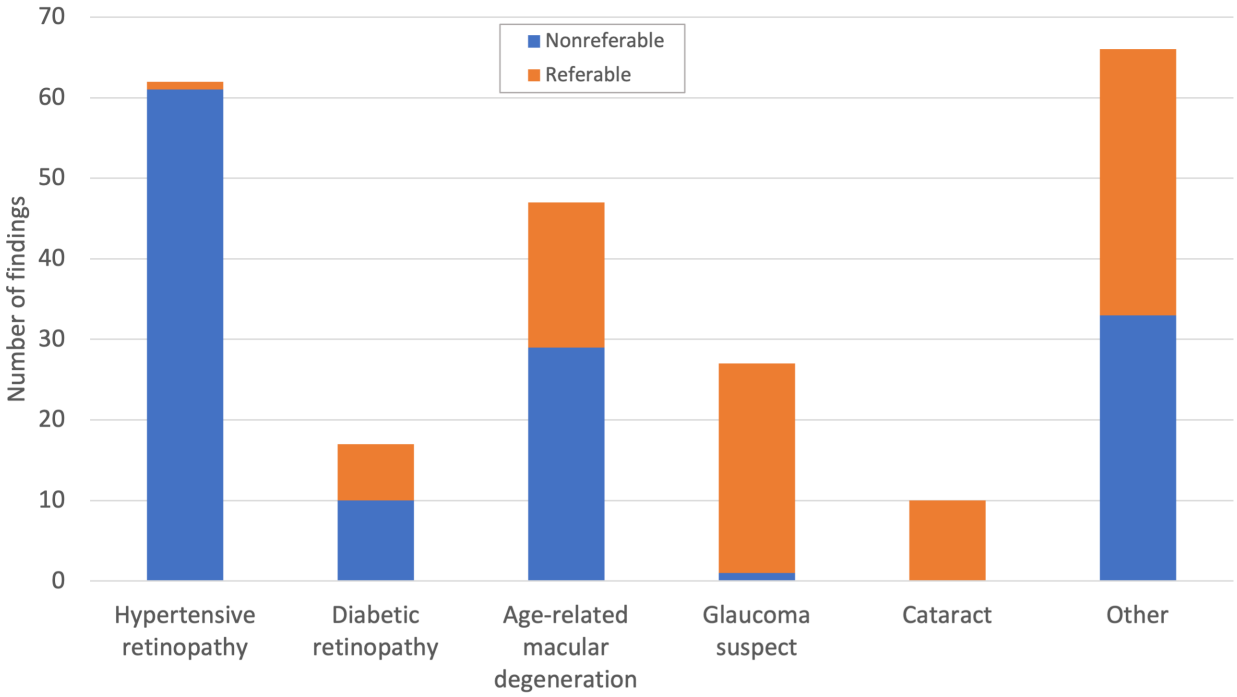

XSL•FO
RenderX

Figure 4. Referable eye disease examples. (A) Moderate hypertensive retinopathy: A 90-year-old female with hypertension and dyslipidemia was found to have diffuse arteriolar attenuation with arteriovenous nicking (arrowheads), increased venous tortuosity, and multiple dot-and-blot intraretinal hemorrhages (arrows). (B) Moderate nonproliferative diabetic retinopathy: A 76-year-old female with diabetes mellitus, hypertension, dyslipidemia, and obesity was found to have scattered dot-and-blot intraretinal hemorrhages (arrows). (C) Intermediate age-related macular degeneration: A 75-year-old female with hypertension and dyslipidemia was found to have numerous medium-sized (arrow) and large-sized (arrowhead) drusen. (D) Glaucoma suspect: A 76-year-old male was found to have asymmetric cupping of his optic nerve head, right greater than left (arrows demarcate the thinned optic nerve rim). He was subsequently confirmed to have glaucoma at his in-person eye visit. (E) Cataract: An 89-year-old male was found to have brunescence of his lens that was visible even without pupillary dilation (arrow). Cataract was subsequently confirmed at his in-person eye visit, and cataract surgery is being planned. (F) A 21-year-old otherwise healthy male was found to have dot-and-blot intraretinal hemorrhages (arrows) and perivenular sheathing (arrowheads). He was subsequently diagnosed with idiopathic retinal vasculitis at his in-person eye visit and was started on systemic immune modulators. The bottom of each image is labeled with age and sex of the patient and eye laterality (M=male; F=female; OD=right eye; OS=left eye).

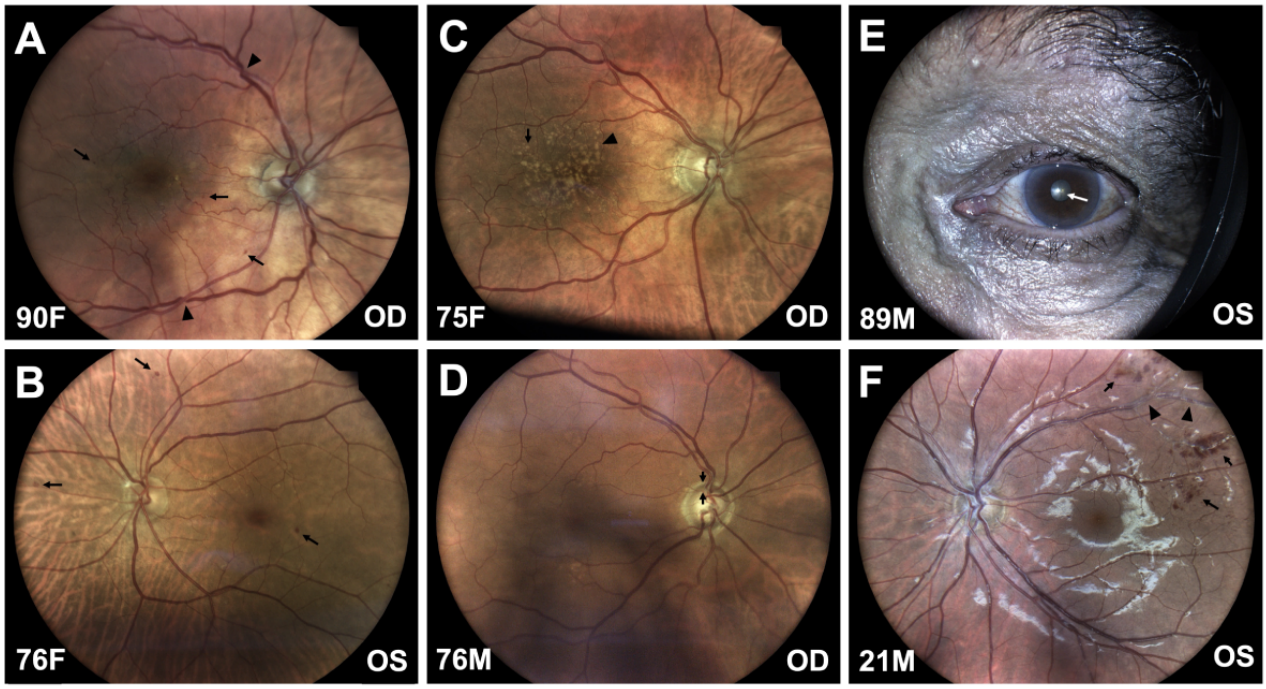

Referrals and Changes in Management

In-Person Eye Consults Triggered From Eye Screening Findings

The remote reading eye clinician recommended in-person eye examinations for 80 of the first 200 patients based on findings of referable eye disease. The primary care clinician adhered to all recommendations, initiating referrals for each of the 80 patients. Referral communication varied and included telephone calls, face-to-face discussions, and written prescriptions. Referrals were either general or directed to specific eye care providers. Nineteen (23.8%) referrals were nonspecific, advising patients to see an eye clinician (optometry or ophthalmology). Sixty-one (76.3%) referrals were specialty specific, with 15 (18.8%) to comprehensive ophthalmology, 42 (52.5%) to retina specialists, 12 (15%) to glaucoma specialists, 1 (1.3%) to uveitis, 1 (1.3%) to oculoplastics, and 1 (1.3%) to low vision optometry. Of the 80 patients, 71 (88.8%) patients were referred to 1 eye care specialty, 7 (8.8%) patients were referred to 2 eye specialties, and 2 (2.5%) patients were referred to 3 eye specialties.

Of the 80 patients who were recommended for in-person eye evaluations, the records indicated that 35 patients (43.8%) reported completing at least one in-person eye visit, 12 patients (15%) reported not completing any in-person eye visit, and 44 patients (55%) had no documentation of compliance (there was no in-person eye consult note or a patient report in the primary care clinic chart). Among the 35 patients who reported completing a visit, 32 in-person eye consult reports were available for review. Of these 32 consult reports, 28 reports (87.5%) confirmed agreement with the findings of the virtual eye consult, and 4 reports (12%) documented disagreement. In these 32 reviewed consults, 11 patients (34.4%) had documented new eye management changes, which included 4 cases of closer observation or additional workup, 3 recommendations for topical eye drops, 2 instructions to perform Amsler grid testing at home, 1 planned glaucoma laser procedure, and 1 planned cataract surgery. Regarding compliance with these new management plans, 8 patients (72.7%) were documented as compliant, 1 patient (9.1%) was documented as noncompliant, and 2 patients (18.2%) had no compliance documentation. Given the small denominator, these figures are descriptive, and CIs are not presented.

In-Person Non–Eye Consults Triggered From Eye Screening Findings

On the basis of virtual eye consult report findings and recommendations, the primary care clinician initiated 7 in-person non–eye specialist consults. Of these, 6 consults were completed, and 1 consult had no documentation of completion (ie, no consult note or compliance record in the primary care clinic chart). In-person non–eye consults included 3 endocrinology consults for improved glycemic control, 2

XSL•FO
RenderX

cardiology consults (1 for a patient with suspected central retinal vein occlusion to improve blood pressure control and assess for carotid artery stenosis and 1 for a patient with retinal aneurysm and concurrent left arm pain), 2 rheumatology consults (1 for the central retinal vein occlusion patient to evaluate for a hypercoagulable state and 1 for a patient with idiopathic retinal vasculitis), and 1 pediatric and 1 genetic counseling consult for an 11-year-old patient with retinal tortuosity to evaluate for connective tissue disease. The central retinal vein occlusion patient had no significant carotid artery stenosis and a negative hypercoagulable workup and was started on aspirin. The patient with the retinal aneurysm and arm pain was found to have a bridging left anterior descending coronary artery and was started on metoprolol. The pediatric patient was found not to have a connective tissue disorder.

Systemic Management Changes Triggered by Eye Imaging Findings

On the basis of the findings from eye imaging, systemic management changes were made for 70 patients (35%; 95% CI 28.7 - 41.8), corresponding to screening ~3 patients per systemic management change (NNS 2.86; 95% CI 2.39-3.48; Table 2). These changes included 34 initiations of oral AREDS2 supplements (17%); 17 hypertension optimizations (8.5%), involving adjustments or initiations of medications, such as minoxidil, losartan, enalapril, carvedilol, metoprolol, hydrochlorothiazide, and spironolactone; 11 diabetes mellitus optimizations (5.5%), involving semaglutide, tirzepatide, metformin, insulin, and empagliflozin; 7 dyslipidemia optimizations (3.5%) primarily involving statins; 8 weight optimizations (4%) involving semaglutide and tirzepatide; 1 initiation on aspirin; 1 dietary intervention; and 1 initiation of mycophenolate mofetil. Some patients received more than 1 systemic intervention. In terms of compliance, 41 (58.6%) of these 70 patients were documented as compliant, 12 patients (17.1%) were documented as noncompliant, and 17 patients (24.3%) had no available compliance documentation.

Discussion

Principal Findings and Relevance

This retrospective cross-sectional analysis of the first 200 patients screened via asynchronous teleophthalmology at a primary care clinic provides early real-world insights into the feasibility, diagnostic yield, and triage outcomes of integrating remote eye screening into non–eye clinical settings. The findings support the study's central objective to evaluate whether asynchronous store-and-forward telehealth technology can identify a meaningful prevalence of both ocular disease and systemic manifestations through eye imaging in a primary care setting and appropriately triage patients for either continued monitoring or in-person evaluation. The findings also support the idea that teleophthalmology screening results can guide changes in eye and systemic disease management.

In our study, 40% of the total cohort was referred for in-person eye examinations based on referable pathology, indicating that every 2.5 patients screened prompted 1 in-person eye referral. This referral rate is consistent with prior teleophthalmology studies, which have reported referral rates ranging from approximately 20% to 48%, depending on the population screened and the range of ocular pathologies assessed [41]. Notably, only 8.8% of our referrals were for diabetic retinopathy, reinforcing the central hypothesis that the ocular disease burden in primary care populations extends far beyond diabetes-related pathology. A mobile teleophthalmology screening program in New York City reported screening 957 individuals in a low socioeconomic minority community, finding that only 29 (3%) individuals had diabetic retinopathy, whereas 305 (32%) had glaucoma, 124 (13%) had cataract, 9 (1%) had macular degeneration, and 97 (10%) had other eye diseases [42]. The higher cumulative prevalence of findings related to eye diseases other than diabetic retinopathy, including hypertensive retinopathy, age-related macular degeneration, glaucoma suspicion, and cataract, underscores the broader potential of eye screening programs as a general tool for chronic disease surveillance, not solely limited to diabetes care.

In our study, patients were unaware of their eye conditions 84.3% of the time. Prior literature aligns with this finding, showing substantial patient unawareness of having eye diseases, with diabetic retinopathy at a ~70% unawareness rate [43], age-related macular degeneration at a >80% unawareness rate [44], and open-angle glaucoma at a ~50% unawareness rate [45]. These statistics underscore the value of proactive, point-of-care eye screening and patient education in primary care settings.

The strong imageability metrics, 98.2% lenient and 86.1% strict adequacy for posterior images, were consistent with those reported in prior literature [46] and confirm that high-quality, nonmydriatic imaging is achievable in routine primary care workflows using minimal training and without the need for pupil dilation. Follow-through from primary care clinicians on virtual consult recommendations was high, with all patients flagged for referral receiving the appropriate communication. Of those who kept their in-person eye appointments and had documentation available, a strong concordance rate was observed, indicating an appropriate clinical accuracy of this asynchronous model.

The relatively high proportion of patients undergoing systemic management changes based on ocular findings in our study underscores the close linkage between eye and systemic disease and that a significant portion of interventions for common eye diseases, such as diabetic retinopathy, hypertensive retinopathy, and age-related macular degeneration, is systemic rather than purely ocular. While AI-based cardiovascular risk calculators that incorporate retinal imaging are not yet FDA-approved in the United States, primary care clinicians can still apply clinical gestalt to interpret retinal findings when managing "gray zone" patients. This includes optimizing blood pressure, lipid levels, and glycemic control based on suggestive ocular markers. In the absence of validated AI tools, this human-guided approach remains a practical and immediately implementable strategy for enhancing chronic disease management.

Nearly half of the systemic management changes involved initiating oral AREDS2 supplements, which contain vitamin C (500 mg), vitamin E (400 IU), zinc (80 mg with 2 mg copper),

XSL•FO
RenderX

lutein (10 mg), and zeaxanthin (2 mg). These supplements have been shown to reduce the risk of progression to advanced age-related macular degeneration by about 78 cases per 1000 people with intermediate (referable) age-related macular degeneration and 4 cases per 1000 with early (nonreferable) age-related macular degeneration [47-49]. Additional studies have reported improvements in visual acuity and contrast sensitivity in patients with early age–related macular degeneration taking AREDS2 [50].

With an estimated 30 million Americans affected by age-related macular degeneration, 25.6 million with early or intermediate forms, and 5.4 million with advanced disease [51], there is a critical need for early intervention. Furthermore, there is mounting evidence that a major pathophysiological component of age-related macular degeneration is vascular disease, and the presence of advanced age-related macular degeneration may serve as a biomarker for systemic vascular health [52]. Thus, primary care clinicians, when informed of their patients' age-related macular degeneration status, have the potential to play a valuable role in early management, including through AREDS2 supplementation, dietary and lifestyle modifications, and systemic vascular health interventions. It should be considered that AREDS2 supplements are over the counter and are typically not covered by insurance.

Limitations

Despite these encouraging findings, the study had several limitations. First, the retrospective nature and reliance on electronic chart review may have introduced information bias due to incomplete documentation, particularly regarding compliance with referrals or recommended management changes. A considerable proportion of referred patients had no documented follow-up status, which limited our ability to fully assess downstream clinical impact. A chart review was conducted only 6 weeks after the 200th patient was screened, which may not have allowed sufficient time for some of the later patients to complete in-person eye consultations and for those outcomes to be documented in the medical record. In addition, the absence of a control group limited causal inference, as systemic management changes may have been implemented without eye imaging information.

Furthermore, the study's setting, a single clinic in Elmhurst, New York, may have introduced selection bias, given the specific demographic and socioeconomic context of the patient population. Most patients self-identified as White, which may not reflect broader urban or national diversity. Also, the geriatrics setting is enriched for older adults and higher chronic disease prevalence, limiting the generalizability of our findings to other primary care settings. For example, our clinic had a higher diabetes prevalence (26%) compared to reported averages in the primary care setting (15% - 20%).

While referability criteria were based on established grading systems, the interpretation of images and decision for referral was left to the discretion of a single reading eye clinician. A single-reader design may have introduced reader-level bias from the subjectivity of image grading and disease classification and may limit reproducibility. Furthermore, the magnitude of potential false positives or negatives, especially in cases graded as partially adequate, remains unknown without a systematic rereview of all cases by a second grader or adjudication panel. The impact of these biases was likely moderate; however, it should be acknowledged that both over- and under-referral could have occurred, affecting the downstream burden on specialty care or the risk of missed pathology. Other key limitations of this study involved limitations in digitizing the eye examination, including lack of dilation, lack of slit lamp visualization of anterior eye structures, and lack of peripheral retinal visualization outside of the single 45° field, all of which could have contributed to missed referable and vision-threatening pathology.

Interpretation of these results must be cautious, particularly in the context of emerging AI-based screening alternatives. While FDA-cleared autonomous AI tools for diabetic retinopathy offer scalability, their narrow indication, lower specificity, and limited adoption to date contrast with the broader diagnostic scope and nuanced triage possible via asynchronous human grading. Our findings support the view that human-led teleophthalmology remains a valuable, perhaps superior, solution in its current form, especially for complex, multipathology detection in high-risk populations. However, these results also align with growing evidence, suggesting a hybrid "clinician+AI" model may ultimately yield the best balance of efficiency, sensitivity, and specificity [32,53,54].

Conclusions

This study provides early evidence that asynchronous teleophthalmology screening integrated into primary care settings can successfully identify a wide range of ocular and systemic diseases, with high patient imageability, diagnostic yield, and clinician and patient follow-through. The generalizability of this study is limited by the single-center design, yet the model itself is inherently scalable. The success of the imaging device used (a nonmydriatic fundus camera) and the asynchronous written communication between the remote reading eye clinician and the primary care clinician suggest that this approach could be feasibly replicated in other primary care settings, particularly those serving aging or chronically ill populations. While the model requires further validation through prospective, multicenter studies and longer-term outcomes tracking [55], its potential to expand access, reduce missed diagnoses, and enhance care coordination across disciplines is compelling. As health care continues to evolve toward proactive, interdisciplinary, and digitally-enabled care models, asynchronous eye screening at the frontlines of medicine deserves further investment and evaluation.

Acknowledgments

The authors received no specific funding for this work.

XSL•FO
RenderX

Data Availability

The datasets generated or analyzed during this study are available in the corresponding author's Google Drive [56].

Authors' Contributions

AP contributed to conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, software, supervision, validation, visualization, and writing, reviewing, and editing the article. BP contributed to data curation, investigation, methodology, project administration, resources, and software. ED contributed to conceptualization, data curation, resources, software, and visualization. SP contributed to conceptualization, formal analysis, investigation, methodology, resources, supervision, validation, visualization, and writing, reviewing, and editing the article.

Conflicts of Interest

AP is the owner of Retina & Ophthalmology Services of New York PLLC and Myeye.healthcare LLC and a part-time clinical employee of Mount Sinai Health System. SP is the owner of Internal Medicine & Geriatrics PLLC. BP is the co-owner of Proceed Medical Billing LLC. ED is an independant contractor of Myeye.healthcare LLC.

Checklist 1
STROBE checklist.
[PDF File, 104 KB - formative_v9i1e81918_app1.pdf]

References

1. U.S. Census Bureau. Demographic turning points for the United States: population projections for 2020 to 2060. : U.S. Department of Commerce; 2020 P25-1144 URL: https://www.census.gov/content/dam/Census/library/publications/2020/demo/p25-1144.pdf [accessed 2025-05-08]
2. Health for life: focus on wellness. American Hospital Association. 2007. URL: https://www.aha.org/system/files/content/00-10/071204_H4L_FocusonWellness.pdf [accessed 2025-05-08]
3. National Eye Institute. Eye disease statistics fact sheet. version 10. : National Institutes of Health; 2014 URL: https://www.nei.nih.gov/sites/default/files/2019-04/NEI_Eye_Disease_Statistics_Factsheet_2014_V10.pdf [accessed 2025-05-08]
4. Levine S, Malone E, Lekiachvili A, Briss PA. Health care industry insights: why the use of preventive services is still low. Prev Chronic Dis 2019 Mar 14;16:E30. [doi: 10.5888/pcd16.180625] [Medline: 30873937]
5. Gowers WR. A Manual and Atlas of Medical Ophthalmoscopy: J. & A. Churchill; 1879.
6. Diabetes Control and Complications Trial Research Group, Nathan DM, Genuth S, et al. The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. N Engl J Med 1993 Sep 30;329(14):977-986. [doi: 10.1056/NEJM199309303291401] [Medline: 8366922]
7. Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33). UK Prospective Diabetes Study (UKPDS) Group. Lancet 1998 Sep 12;352(9131):837-853. [doi: 10.1016/S0140-6736(98)07019-6] [Medline: 9742976]
8. Hypertension in Diabetes Study (HDS): I. Prevalence of hypertension in newly presenting type 2 diabetic patients and the association with risk factors for cardiovascular and diabetic complications. J Hypertens 1993 Mar;11(3):309-317. [doi: 10.1097/00004872-199303000-00012] [Medline: 8387089]
9. Cuadros J, Bresnick G. EyePACS: an adaptable telemedicine system for diabetic retinopathy screening. J Diabetes Sci Technol 2009 May 1;3(3):509-516. [doi: 10.1177/193229680900300315] [Medline: 20144289]
10. Wagner SK, Fu DJ, Faes L, et al. Insights into systemic disease through retinal imaging-based oculomics. Transl Vis Sci Technol 2020 Feb 12;9(2):6. [doi: 10.1167/tvst.9.2.6] [Medline: 32704412]
11. Abràmoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. NPJ Digit Med 2018;1(39):39. [doi: 10.1038/s41746-018-0040-6] [Medline: 31304320]
12. Savoy M. IDx-DR for diabetic retinopathy screening. Am Fam Physician 2020 Mar 1;101(5):307-308. [Medline: 32109029]
13. Bhaskaranand M, Ramachandra C, Bhat S, et al. The value of automated diabetic retinopathy screening with the EyeArt system: a study of more than 100,000 consecutive encounters from people with diabetes. Diabetes Technol Ther 2019 Nov;21(11):635-643. [doi: 10.1089/dia.2019.0164] [Medline: 31335200]
14. Ipp E, Liljenquist D, Bode B, et al. Pivotal evaluation of an artificial intelligence system for autonomous detection of referrable and vision-threatening diabetic retinopathy. JAMA Netw Open 2021 Nov 1;4(11):e2134254. [doi: 10.1001/jamanetworkopen.2021.34254] [Medline: 34779843]
15. Lik Au SC. Artificial intelligence systems for diabetic retinopathy screening: appraisal on the 3rd US FDA Approved Algorithms- AEYE-DS. JOAR 2023;4(2):1-3. [doi: 10.46889/JOAR.2023.4202]

XSL•FO
RenderX

16. Rajalakshmi R, PramodKumar TA, Naziyagulnaaz AS, et al. Leveraging artificial intelligence for diabetic retinopathy screening and management: history and current advances. Semin Ophthalmol 2025 Nov;40(8):719-726. [doi: 10.1080/08820538.2024.2432902] [Medline: 39580713]
17. Shah SA, Sokol JT, Wai KM, et al. Use of artificial intelligence-based detection of diabetic retinopathy in the US. JAMA Ophthalmol 2024 Dec 1;142(12):1171-1173. [doi: 10.1001/jamaophthalmol.2024.4493] [Medline: 39480408]
18. Nichols GA, McBurnie M, Paul L, et al. The high prevalence of diabetes in a large cohort of patients drawn from safety net clinics. Prev Chronic Dis 2016 Jun 16;13:E78. [doi: 10.5888/pcd13.160056] [Medline: 27309415]
19. O'Brien MJ, Bailey SC, Gregory DL, et al. Screening for prediabetes and diabetes in a national network of federally qualified health centers: an observational study. J Gen Intern Med 2023 Dec;38(16):3541-3548. [doi: 10.1007/s11606-023-08402-1] [Medline: 37731136]
20. Yogesan K, Wilcock A, Mehrotra A. Retinal screening of patients with diabetes in primary care clinics: why has uptake of this promising idea been so low? In: Yogesan K, Cuadros J, Goldschmidt L, editors. Digital Eye Care and Teleophthalmology: Springer International Publishing; 2020:33-42.
21. Blumenkranz MS, Tarver ME, Myung D, Eydelman MB, Collaborative Community on Ophthalmic Imaging Executive Committee. The collaborative community on ophthalmic imaging: accelerating global innovation and clinical utility. Ophthalmology 2022 Feb;129(2):e9-e13. [doi: 10.1016/j.ophtha.2021.10.001] [Medline: 34774340]
22. Tan W, Yao X, Le TT, Tan B, Schmetterer L, Chua J. The new era of retinal imaging in hypertensive patients. Asia Pac J Ophthalmol (Phila) 2022;11(2):149-159. [doi: 10.1097/APO.0000000000000509] [Medline: 35533334]
23. Dumitrescu AG, Istrate SL, Iancu RC, Guta OM, Ciuluvica R, Voinea L. Retinal changes in diabetic patients without diabetic retinopathy. Rom J Ophthalmol 2017;61(4):249-255. [Medline: 29516043]
24. Zhu Z, Liu D, Chen R, et al. The association of retinal age gap with metabolic syndrome and inflammation. J Diabetes 2023 Mar;15(3):237-245. [doi: 10.1111/1753-0407.13364] [Medline: 36919192]
25. Yi JK, Rim TH, Park S, et al. Cardiovascular disease risk assessment using a deep-learning-based retinal biomarker: a comparison with existing risk scores. Eur Heart J Digit Health 2023 May;4(3):236-244. [doi: 10.1093/ehjdh/ztad023] [Medline: 37265875]
26. Vaghefi E, Squirrell D, Yang S, et al. Development and validation of a deep-learning model to predict 10-year atherosclerotic cardiovascular disease risk from retinal images using the UK Biobank and EyePACS 10K datasets. Cardiovasc Digit Health J 2024 Apr;5(2):59-69. [doi: 10.1016/j.cvdhj.2023.12.004] [Medline: 38765618]
27. Joo YS, Rim TH, Koh HB, et al. Non-invasive chronic kidney disease risk stratification tool derived from retina-based deep learning and clinical factors. NPJ Digit Med 2023 Jun 17;6(1):114. [doi: 10.1038/s41746-023-00860-5] [Medline: 37330576]
28. Ashayeri H, Jafarizadeh A, Yousefi M, Farhadi F, Javadzadeh A. Retinal imaging and Alzheimer's disease: a future powered by Artificial Intelligence. Graefes Arch Clin Exp Ophthalmol 2024 Aug;262(8):2389-2401. [doi: 10.1007/s00417-024-06394-0] [Medline: 38358524]
29. Wu JH, Liu TYA. Application of deep learning to retinal-image-based oculomics for evaluation of systemic health: a review. J Clin Med 2022 Dec 24;12(1):152. [doi: 10.3390/jcm12010152] [Medline: 36614953]
30. Poplin R, Varadarajan AV, Blumer K, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nat Biomed Eng 2018 Mar;2(3):158-164. [doi: 10.1038/s41551-018-0195-0] [Medline: 31015713]
31. Cuadros J. The real-world impact of artificial intelligence on diabetic retinopathy screening in primary care. J Diabetes Sci Technol 2021 May;15(3):664-665. [doi: 10.1177/1932296820914287] [Medline: 32329352]
32. Xie Y, Nguyen QD, Hamzah H, et al. Artificial intelligence for teleophthalmology-based diabetic retinopathy screening in a national programme: an economic analysis modelling study. Lancet Digit Health 2020 May;2(5):e240-e249. [doi: 10.1016/S2589-7500(20)30060-1] [Medline: 33328056]
33. Olayiwola JN, Magaña C, Harmon A, et al. Telehealth as a bright spot of the COVID-19 pandemic: recommendations from the virtual frontlines ("Frontweb"). JMIR Public Health Surveill 2020 Jun 25;6(2):e19045. [doi: 10.2196/19045] [Medline: 32479413]
34. Palmer CF. Billing, coding, and reimbursement in eye telehealth programs. In: Maa A, editor. Ocular Telehealth: A Practical Guide: Elsevier Health Sciences; 2022:175-184. [doi: 10.1016/B978-0-323-83204-5.00014-7]
35. Walsh L, Hong SC, Chalakkal RJ, Ogbuehi KC. A systematic review of current teleophthalmology services in New Zealand compared to the four comparable countries of the United Kingdom, Australia, United States of America (USA) and Canada. Clin Ophthalmol 2021;15:4015-4027. [doi: 10.2147/OPTH.S294428] [Medline: 34675470]
36. Saha SK, Fernando B, Cuadros J, Xiao D, Kanagasingam Y. Automated quality assessment of colour fundus images for diabetic retinopathy screening in telemedicine. J Digit Imaging 2018 Dec;31(6):869-878. [doi: 10.1007/s10278-018-0084-9] [Medline: 29704086]
37. Wu L, Fernandez-Loaiza P, Sauma J, Hernandez-Bogantes E, Masis M. Classification of diabetic retinopathy and diabetic macular edema. World J Diabetes 2013 Dec 15;4(6):290-294. [doi: 10.4239/wjd.v4.i6.290] [Medline: 24379919]
38. Wong TY, Mitchell P. Hypertensive retinopathy. N Engl J Med 2004 Nov 25;351(22):2310-2317. [doi: 10.1056/NEJMra032865] [Medline: 15564546]

XSL•FO
RenderX

39. Ahmad SS. Glaucoma suspects: a practical approach. Taiwan J Ophthalmol 2018;8(2):74-81. [doi: 10.4103/tjo.tjo_106_17] [Medline: 30038885]
40. Age-Related Eye Disease Study Research Group. The Age-Related Eye Disease Study system for classifying age-related macular degeneration from stereoscopic color fundus photographs: the Age-Related Eye Disease Study Report Number 6. Am J Ophthalmol 2001 Nov;132(5):668-681. [doi: 10.1016/s0002-9394(01)01218-1] [Medline: 11704028]
41. Dolar-Szczasny J, Barańska A, Rejdak R. Evaluating the efficacy of teleophthalmology in delivering ophthalmic care to underserved populations: a literature review. J Clin Med 2023 Apr 27;12(9):3161. [doi: 10.3390/jcm12093161] [Medline: 37176602]
42. Al-Aswad LA, Elgin CY, Patel V, et al. Real-time mobile teleophthalmology for the detection of eye disease in minorities and low socioeconomics at-risk populations. Asia Pac J Ophthalmol (Phila) 2021;10(5):461-472. [doi: 10.1097/APO.0000000000000416] [Medline: 34582428]
43. Land MR, Patel PA, Bui T, et al. Examining the role of telemedicine in diabetic retinopathy. J Clin Med 2023 May 18;12(10):3537. [doi: 10.3390/jcm12103537] [Medline: 37240642]
44. Cheng BT, Kim AB, Lyon AT, Mirza RG. Disease awareness among patients with age-related macular degeneration: patterns and predictors. J Vitreoretin Dis 2024;8(1):58-66. [doi: 10.1177/24741264231210078] [Medline: 38223766]
45. Friedman DS, Wolfs RCW, O'Colmain BJ, et al. Prevalence of open-angle glaucoma among adults in the United States. Arch Ophthalmol 2004 Apr;122(4):532-538. [doi: 10.1001/archopht.122.4.532] [Medline: 15078671]
46. Padua W, Butola S, Nithyanandam S, et al. Performance of a new tabletop non-mydriatic fundus camera for single-field diabetic retinopathy detection: a pilot study. Cureus 2025 Jul;17(7):e88198. [doi: 10.7759/cureus.88198] [Medline: 40827150]
47. Chew EY, Clemons TE, Agrón E, et al. Long-term effects of vitamins C and E, β-carotene, and zinc on age-related macular degeneration: AREDS report no. 35. Ophthalmology 2013 Aug;120(8):1604-1611. [doi: 10.1016/j.ophtha.2013.01.021] [Medline: 23582353]
48. Age-Related Eye Disease Study 2 Research Group. Lutein + zeaxanthin and omega-3 fatty acids for age-related macular degeneration: the Age-Related Eye Disease Study 2 (AREDS2) randomized clinical trial. JAMA 2013 May 15;309(19):2005-2015. [doi: 10.1001/jama.2013.4997] [Medline: 23644932]
49. Evans JR, Lawrenson JG. Antioxidant vitamin and mineral supplements for slowing the progression of age-related macular degeneration. Cochrane Database Syst Rev 2023 Sep 13;9(9):CD000254. [doi: 10.1002/14651858.CD000254.pub5] [Medline: 37702300]
50. Beatty S, Chakravarthy U, Nolan JM, et al. Secondary outcomes in a clinical trial of carotenoids with coantioxidants versus placebo in early age-related macular degeneration. Ophthalmology 2013 Mar;120(3):600-606. [doi: 10.1016/j.ophtha.2012.08.040] [Medline: 23218821]
51. García-Layana A, Cabrera-López F, García-Arumí J, Arias-Barquet L, Ruiz-Moreno JM. Early and intermediate age-related macular degeneration: update and clinical review. Clin Interv Aging 2017;12:1579-1587. [doi: 10.2147/CIA.S142685] [Medline: 29042759]
52. Mauschitz MM, Finger RP. Age-related macular degeneration and cardiovascular diseases: revisiting the common soil theory. Asia Pac J Ophthalmol (Phila) 2022;11(2):94-99. [doi: 10.1097/APO.0000000000000496] [Medline: 35213420]
53. Dow ER, Khan NC, Chen KM, et al. AI-human hybrid workflow enhances teleophthalmology for the detection of diabetic retinopathy. Ophthalmol Sci 2023 Dec;3(4):100330. [doi: 10.1016/j.xops.2023.100330] [Medline: 37449051]
54. Shin JY, Son J, Kong ST, et al. Clinical utility of deep learning assistance for detecting various abnormal findings in color retinal fundus images: a reader study. Transl Vis Sci Technol 2024 Oct 1;13(10):34. [doi: 10.1167/tvst.13.10.34] [Medline: 39441571]
55. Horton MB, Brady CJ, Cavallerano J, et al. Practice Guidelines for Ocular Telehealth-Diabetic Retinopathy, Third Edition. Telemed J E Health 2020 Apr;26(4):495-543. [doi: 10.1089/tmj.2020.0006] [Medline: 32209018]
56. Pinhas et al teleophthalmology pilot datasheet. Google Drive. URL: https://docs.google.com/spreadsheets/d/1xHIRep9Lge0cbThy07JdRWIKxrYjHX3r9ww2smNsHdY/edit?gid=0#gid=0 [accessed 2025-10-16]

Abbreviations:

AI: artificial intelligence
AREDS: Age-Related Eye Disease Study
CMS: Centers for Medicare & Medicaid Services
CPT: Current Procedural Terminology
FDA: Food and Drug Administration
ICDR: International Clinical Diabetic Retinopathy
IRB: institutional review board

XSL•FO
RenderX

Edited by A Stone, A Mavragani; submitted 05.08.25; peer-reviewed by SCL Au; revised version received 30.09.25; accepted 30.09.25; published 24.10.25.

Please cite as:
Pinhas A, Pinhas B, Dmitruk E, Pinhas S
Eye and Systemic Disease Management Changes After Teleophthalmology Screening in Primary Care: Retrospective Cross-Sectional Pilot Study of 200 Consecutive Patients
JMIR Form Res 2025;9:e81918
URL: https://formative.jmir.org/2025/1/e81918
doi:10.2196/81918

© Alexander Pinhas, Boris Pinhas, Egor Dmitruk, Stella Pinhas. Originally published in JMIR Formative Research (https://formative.jmir.org), 24.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

A Robot-Delivered Training Program to Improve Children's Mental Health and Resilience in Dutch Primary Schools: Pilot Intervention Study

Anne Zijp[1], MSc; Jiska J Aardoom[1], PhD; Olivier A Blanson Henkemans[2], PhD; Sylvia van der Pal[2], PhD; Eline Vlasblom[2], MSc; Anke Versluis[1], PhD

[1]Department of Public Health and Primary Care, Leiden University Medical Center, Albinusdreef 2, Leiden, The Netherlands
[2]Netherlands Organisation for Applied Scientific Research (TNO), Leiden, The Netherlands

Corresponding Author:
Anne Zijp, MSc
Department of Public Health and Primary Care, Leiden University Medical Center, Albinusdreef 2, Leiden, The Netherlands

Abstract

Background: Mental health problems often start at an early age and can persist into adulthood, leading to physical and mental health problems such as substance abuse, sleep problems, depressive disorders, and suicidal tendencies. Therefore, it is important to invest in the mental health of young people through, for example, initiatives focused on mental health promotion and prevention. The ePartners robot buddy offers children training modules focused on enhancing resilience and mental health, specifically targeting self-image and social skills and/or addressing unhelpful feelings and thinking patterns in children's daily life situations.

Objective: The study primarily aims to assess the feasibility, acceptability, and usability of the intervention according to the children and their teachers and secondarily aims to evaluate its potential effects on the mental well-being (general mental well-being, quality of life [QoL], and self-efficacy) of children.

Methods: A single-arm, 6-week, pre-post pilot intervention study involving children and their teachers was conducted in 3 primary schools in the Netherlands. Outcomes were assessed using questionnaires. Primary outcomes were assessed postintervention and included feasibility and acceptability for teachers and acceptability and usability for children. Secondary outcomes included self-reported general mental well-being and self-efficacy and teacher-reported general mental well-being and were assessed at baseline and postintervention.

Results: Data showed that the intervention was generally perceived as moderately feasible according to the Feasibility of Intervention Measure (mean 17.3, SD 2.6; on a scale from 4 to 20) and showed relatively high acceptability (mean 16.9, SD 3.8; on a scale from 4 to 20) according to teachers (n=7). Additional feasibility questions showed that teachers found it generally feasible to guide children who had few questions about using the robot. Feasibility was moderate due to limited time for integration, many content-related questions from children, and the substantial learning needed to select themes. Children (n=73) reported high perceived usability of the intervention (mean 15.2, SD 2.4; on a scale from 4 to 20). The perceived acceptability of the intervention by children was also relatively high, with a mean of 12.0 (SD 2.3) on a scale from 3 to 15. Teacher-reported QoL of children improved significantly from baseline (mean 36.0, SD 4.6) to postintervention (mean 37.2, SD 3.8; t_{64}=2.77; P=.01); however, the children's self-reported QoL did not significantly change over time. No significant changes in general mental well-being and self-efficacy scores were found.

Conclusions: This study provides valuable insights into the feasibility, acceptability, and usability of a robot-delivered mental health–promoting intervention within a primary school setting. Further research is needed to fully understand its potential benefits and address existing limitations associated with the implementation of such interventions in the school setting.

(JMIR Form Res 2025;9:e66797) doi:10.2196/66797

KEYWORDS

well-being; resilience; mental health; primary school children; robot-delivered intervention; robot; training program; child; mental; primary school; adulthood; depressive disorder; promotion; prevention; suicidal ideation; feasibility; acceptability; usability; pilot intervention; single-arm; suicide

XSL•FO
RenderX

Introduction

Mental disorders represent a substantial global burden and rank among the top 10 leading causes of worldwide disease burden [1]. Concerning adolescents, nearly one-third (30.5%) of Dutch adolescents aged 12 to 16 years struggle with mental issues, with rates being higher for girls (39.8%) than for boys (21.5%) [2]. These mental health issues are already present in primary school (26.1%) and include hyperactivity/attention problems, emotional problems, problems with peers, and behavioral problems. A recent large-scale meta-analysis showed that almost two-thirds of mental disorders begin before the age of 25 years, and one-third start before the age of 14 years [3]. Experiencing mental health issues in youth often extends into adulthood, leading to physical and mental problems such as substance use, sleep issues, depressive disorders, and suicidal tendencies [4-6]. Therefore, it is important to invest in the mental health of young people through, for example, initiatives focused on mental health promotion and prevention.

Fostering resilience, defined as the ability to cope with stressful situations and successfully adapt to setbacks [7], can be considered a key factor in mental health promotion. Resilience encompasses the ability to bounce back from adversity, adapt positively to challenging circumstances, and maintain well-being in the face of stress or trauma. Resilience is a positive capacity that can be learned and, as such, can improve over time [8], making it a suitable target for mental health promotion initiatives. Indeed, resilience has been found to influence individuals' health, well-being, and quality of life (QoL) [9-11]. Moreover, resilience-focused interventions have been found to improve mental health outcomes, such as depressive symptoms and general psychological distress, in children and adolescents [12]. The school environment provides a convenient setting for promoting mental health and fostering resilience in children. Previous studies on school-based mental health promotion and prevention programs have shown positive effects on children's mental health. More specifically, a systematic review showed positive effects of resilience-focused mental health promotion programs in primary schools with regard to children's ability to cope with daily stressors [13]. The reviewed programs were designed to enhance social and emotional skills with a strong emphasis on enhancing coping skills and strategies, such as mindfulness-based stress reduction techniques. These programs also aimed to improve social relationships, with the ultimate goal of improving mental well-being and safeguarding against future risks.

Social robots that can communicate and interact with people offer novel opportunities for mental health promotion. Social robots are increasingly being used in clinical and health settings to provide information, monitor health outcomes, and deliver psychosocial interventions [14-18]. Dawe et al [18] conducted a scoping review regarding social robots to help children in health care contexts. Fifty studies were identified, most of which were small feasibility studies. In these studies, the social robots were primarily used to provide companionship, support, and well-being through distraction and emotional support. They also served educational purposes (eg, providing information, demonstrating exercises, and delivering feedback) and provided assistance in stimulating social interactions and play among children. The results showed that social robots are generally readily and well accepted among children and that their use in health care is promising.

Building on research in this area, this pilot study examined a social robot–delivered mental health–promoting intervention in primary school. The intervention offers children training modules focused on enhancing resilience and mental health, specifically targeting self-image and social skills and/or addressing unhelpful feelings and thinking patterns in their daily life situations. The main aim of the pilot study was to investigate the feasibility, acceptability, and usability of the intervention according to the children and their teachers. A secondary aim was to evaluate its potential effects on the mental well-being of children.

Methods

Design and Population

A single-arm, 6-week, pre-post pilot intervention study was conducted. Data were collected from June to October 2023. The study population consisted of children and their teachers. Participants were recruited at 3 primary schools in the Netherlands: a special education school, a school in a low socioeconomic status neighborhood, and a school in a higher socioeconomic status neighborhood. In the Dutch primary school system, 1 teacher typically provides instruction to the same group of children throughout the entire week. Children remain in the same classroom, and the teacher covers most subjects. Grades 5, 6, and 7 generally include children aged 8 - 10, 9 - 11, and 10 - 12 years, respectively. Children were eligible when they (1) attended grades 5, 6, or 7; (2) were able to speak, read, and understand the Dutch language; and (3) when their parents or guardians mastered the Dutch language to understand the information letter and informed consent form. Teachers were eligible to participate when they were teaching the participating children.

Ethical Considerations

The Medical Ethical Committee of the Leiden University Medical Center in the Netherlands approved the protocol of the study (decision reference number WSC-2023 - 05). According to the Medical Ethics Committee of Leiden University Medical Center, the study did not fall within the scope of the Dutch Medical Research Involving Human Subjects Act (23 - 3026). All procedures were conducted in accordance with institutional guidelines and the Declaration of Helsinki. Informed consent was obtained from all parents of participating children before data collection. Participation was voluntary, and participants were free to withdraw at any time without consequence. Data were anonymized before analysis. No personal identifiers were retained, and access to the data was restricted to the research team members to ensure confidentiality. Participants did not receive individual compensation. As part of the study activities, the participating classes received an educational game as an incentive. No images or supplementary materials included in this paper allow for the identification of individual participants.

XSL•FO
RenderX

Intervention

Each child individually followed a psychological training developed by Therapieland, a Dutch organization that offers evidence-based digital solutions for mental health care, focusing on one or two of the following themes: (1) self-image, (2) social skills, and (3) thinking, feeling, and doing. A short description of each theme can be found in Textbox 1, and a more elaborate description in Multimedia Appendix 1. The training was offered via a robot. The robot delivered the training by speaking text and asking questions to help children reflect on the module's content, make connections to their own experiences, and create assignments for the upcoming week, as well as by showing educational videos via a tablet connected to the robot. An overview of the setup can be found in Figure 1. The child responded verbally to the robot or answered the robot's question via the tablet. The robot provided reinforcing feedback on the process (eg, "Thank you for your answer").

Textbox 1. Three intervention themes.

- Self-image: designed for children to develop a positive self-image. Recommended for children who are experiencing a negative self-image, anxiety, gloom, and/or stress.
- Social skills: designed for children to learn social skills techniques. Recommended for children who are experiencing insecurity and/or have limited social skills.
- Thinking, feeling, doing: designed for children to develop coping strategies for various situations. Recommended for children who are experiencing gloom, anxiety, negative self-image, and/or problems in emotion regulation.

Figure 1. Intervention setup of the ePartners robot buddy with the tablet.

XSL•FO
RenderX

The theme(s) the child worked on was based on the needs of the child as determined by either the parent(s), teacher, and/or researchers (see *Procedure* section). Within each theme, the teacher, child, and/or researchers decided which specific sessions were done (see *Procedure* section). An overview of the themes and an example of underlying sessions can be found in Figure 2. The child completed 1 session with the robot per week, for a total of 6 weeks. At the end of each session, the robot gave the child an exercise to complete at home and in the classroom (eg, compliment others this week and write down how it went). In the following session, the robot inquired whether the child practiced with the exercise and how they felt about it. The robot did not provide content-related feedback on the exercise but delivered process-related feedback, such as "yes it was a difficult exercise" or "let's practice some more."

Figure 2. Overview of the themes of the psychological intervention ePartner robot buddy and examples of sessions within the themes.

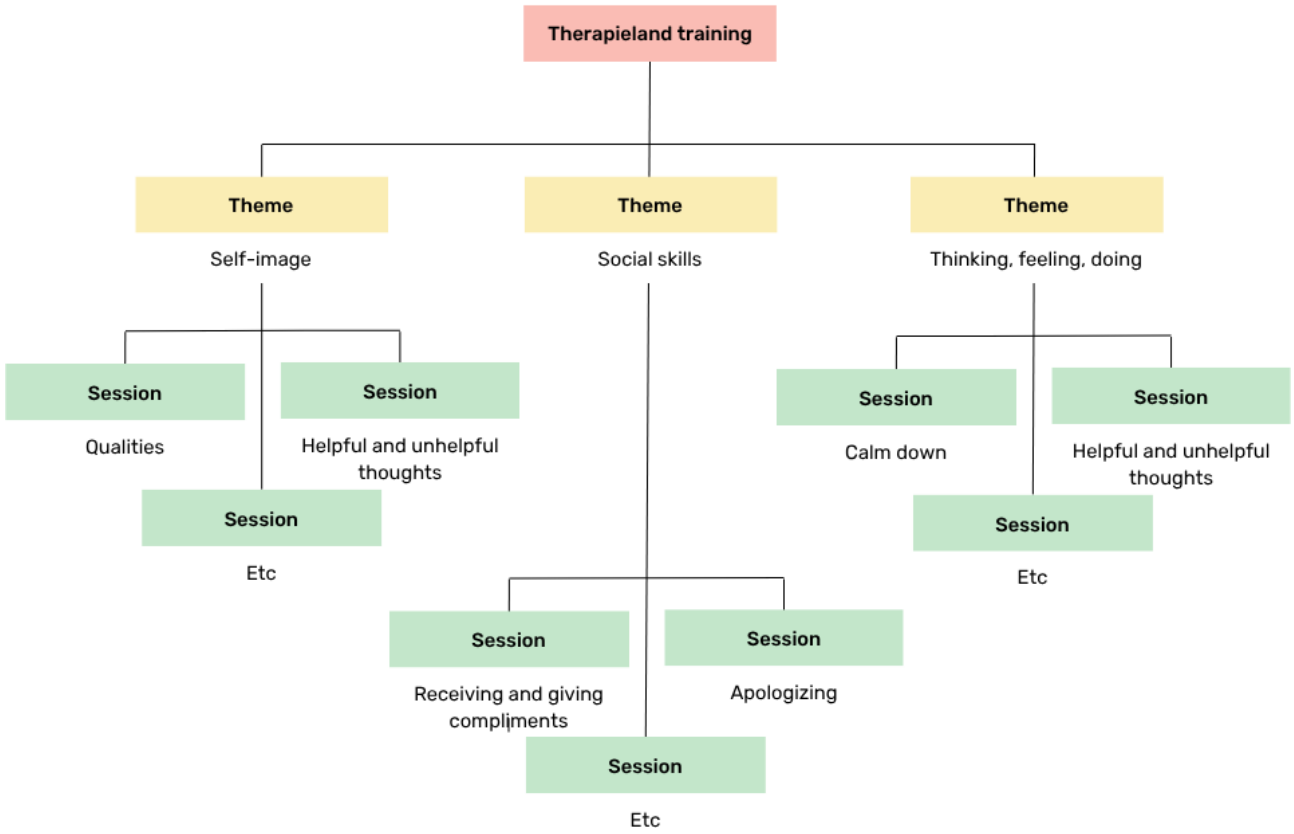

Procedure

Teachers of grades 5, 6, and 7 of the 3 participating schools received a digital information letter; researchers visited the school for a presentation of the robot and the research and invited teachers to participate. Interested teachers were asked to sign the informed consent form. Teachers were provided with an explanation of the intervention and procedure. Children were recruited through the participating teacher; they received an informative flyer. In addition, parents of the children received an official invitation through a school newsletter, a school app for parents, or on paper. This invitation included a QR code for the information letter, screening, and digital informed consent form. In case of questions, the parents were instructed to contact the teacher or the researchers.

After a child was included, the teacher determined which theme(s) was most suitable for the child based on an existing checklist that is periodically completed by the teacher as part of a student tracking system. This system, already implemented in the schools, is used for tracking student progress, results, and behaviors. The checklist provides scores on the child's self-confidence, self-image, and social interactions, and these scores were used as a guideline for the selection of themes. The teacher then provided the selected suggestions for the theme(s) to the parent(s), and parents could respond in a face-to-face meeting (which would then be scheduled) or via email. If parents did not respond, the teacher decided on the theme(s). In some cases, teachers did not have time to select the theme(s), and the researcher chose the themes instead. By using this approach, the intervention could be tailored to the individual needs of each child. Next, the teacher completed the baseline KIDSCREEN-10 for the child (see *Measures* section).

At the first session, the child was asked by the researcher to complete the baseline questionnaire on general mental well-being and self-efficacy (see *Measures* section). Next, the researcher introduced the robot, and the robot also introduced itself using an automated introduction text. The researcher explained that each theme comprised different sessions, and the teacher and child together decided which sessions the child would complete in that week. There were a couple of exceptions where researchers chose the sessions because the teacher did not have sufficient time to do so.

The child received 1 session per week, each lasting up to 20 minutes, over a period of 6 weeks during school hours. Each session took place in a dedicated room where a robot and a

XSL•FO
RenderX

tablet were prearranged, along with any required materials such as forms. To preserve the autonomous character of the intervention, the researcher remained present but out of sight. Upon arrival, the child was welcomed by the researcher and invited to sit in front of the robot. Once seated, the researcher stepped aside, and the robot independently initiated the session. During the session, the robot guided the child through various components, including educational animations or videos, reflective questions, and weekly assignments. The researcher intervened only in case of technical difficulties or if the child encountered confusion regarding the next steps. At the end of each session, an evaluation form was completed by the child to assess their experience. Sessions lasted approximately 15 minutes, consisting of a brief welcome and optional recap (3 - 4 min), followed by an educational component (4 - 5 min), a reflection and assignment phase (3 - 4 min), and a short evaluation (1 min). Most children were assigned to the theme Self-image (n=32), followed by Feeling, thinking, doing (n=20), and Social skills (n=17), and some were assigned to sessions from both Feeling, thinking, doing and Self-image (n=4). At the end of the 6 weeks, the teacher completed the postintervention questionnaire on general mental well-being for the child, and the child completed the postintervention questionnaires assessing acceptability, usability, general mental well-being, and self-efficacy. After all the children in the school had completed the intervention, the corresponding teachers of that school completed the postintervention questionnaires assessing feasibility, acceptability, and additional measures. All participants completed pre- and postmeasurements.

Measures

Feasibility

The intervention's feasibility for teachers was assessed by the validated 4-item Feasibility of Intervention Measure [19]. Additionally, teachers were asked how feasible it was to determine the themes for the child, whether they had to learn a lot to determine the themes, and whether they had enough time to integrate the intervention in the school setting. All items were answered on a 5-point rating scale ranging from 1 ("completely disagree") to 5 ("completely agree"). The feasibility of recruitment was assessed by the response rate, that is, the proportion of invited children and parents who provided informed consent.

Acceptability

For teachers, the intervention's acceptability was assessed using the 4-item Acceptability of Intervention Measure [19]. Two additional questions were added to determine whether they (1) were satisfied with using the intervention in the school setting and (2) would use the intervention again. All items were answered on a 5-point Likert scale ranging from 1 ("completely disagree") to 5 ("completely agree").

In children, the intervention's acceptability was assessed using 3 questions. It assessed whether they were satisfied with the robot, how helpful the robot was, and their willingness to use the robot again. The items were answered on a 5-point Smileyometer scale. The Smileyometer scale was presented to the children in a horizontal row with supporting words under the smiley faces [20]. Children were asked to pick 1 smiley face.

Usability

The intervention's usability was assessed in children using 4 items. The items assessed whether the intervention was easy to use and could be carried out independently, as well as the difficulty level of the questions of the robot and the children's ability to understand what the robot was saying. The items were answered on a 5-point Smileyometer scale.

General Mental Well-Being and Self-Efficacy as Reported by Children

Children's self-reported general mental well-being and self-efficacy of children were assessed using 2 items from the validated Child Outcome Rating Scale [21]. Specifically, children were asked to answer the questions "How am I doing?" and "How am I doing at school?" The items were answered on a 10-point smiley face Likert scale.

Next, the Dutch version of KIDSCREEN-10 for children was used to assess children's self-reported QoL [22]. The 10 questions ask about the child's experiences over the past week, with an example item being "Have you felt fit and well?" One additional question aimed to assess the child's general health (ie, "In general, how would you say your health is?"). All questions were answered on a 5-point Likert scale. Higher scores indicate a better QoL.

Additionally, the 8-item Self-Efficacy Questionnaire for Children was administered [23]. An example item is "How well can you become friends with other children?" The items were answered on a 5-point smiley face Likert scale. Higher scores indicate a better self-efficacy.

Children's General Mental Well-Being as Reported by Teachers

To complement the children's self-reported QoL, proxy reports from the teachers were used [24]. Specifically, children's mental health was assessed with the Dutch KIDSCREEN-10 questionnaire completed by their teacher. The teacher questionnaire included the same questions as the child version, but from the teacher's perspective (eg, "Has the child felt fit and well?").

Additional Measures

Sociodemographic characteristics, including age and gender, were assessed in children at baseline and postintervention in teachers. General experiences of teachers with the intervention in the school setting were assessed by 3 open-ended questions administered at postintervention (ie, How did you experience the training sessions with the robot in the school setting? How do you think the children experienced the training with the robot? Would you like to say anything else about the training with the robot in the school setting?). These questions aimed to get a better understanding of their experiences.

In addition, before conducting each session, children were asked to respond to 5 questions regarding their experiences with the latest session: what they thought of last week's session (pleasant/neutral/unpleasant), how they felt last week's session

XSL•FO
RenderX

went (good/neutral/bad), what they thought of the level of difficulty (too difficult/somewhat difficult/just right/somewhat easy/too easy), whether they did try practicing with it at home (yes/no), and how important they found it to learn the task with the robot from last week (yes/no).

Analyses

Descriptive analyses (eg, mean [SD], N, and percentages) were used to summarize the results regarding sociodemographic characteristics, feasibility, acceptability, usability, mental well-being (both children's self-report and teacher-reported), and self-efficacy. To identify potential change over time, the total scores on the Child Outcome Rating Scale, KIDSCREEN-10, and 8-item Self-Efficacy Questionnaire for Children were statistically compared between pre- and postintervention using paired samples (2-tailed) *t* tests. Data were only analyzed in case a child completed 5 or more sessions. The analyses were performed using SPSS (version 25; IBM Corp) [25]. The free-text answers to the open-ended questions were summarized.

Results

Primary Research Question: Feasibility, Acceptability, and Usability of the Robot

Feasibility and Acceptability According to Teachers

Seven teachers were included and completed the postintervention questionnaires. The mean age was 41.3 years (range 33 - 63 years); of 7 teachers, 3 were male and 4 were female. Table 1 shows the results regarding feasibility and acceptability. The intervention's feasibility was scored high. The additional feasibility questions showed that teachers found it generally feasible to guide the children and set themes for the children. Furthermore, children did not ask many questions about how to use the robot. Feasibility scores were moderate in terms of having enough time to integrate the intervention into the school setting. According to the teachers, the feasibility of the intervention was considered to be limited by children asking many questions about the content of the sessions and due to the considerable amount of learning required to be able to determine the themes for the child.

The general acceptability of the intervention was reasonably high. Furthermore, teachers generally reported being satisfied with the use of the intervention in the school setting and that they would like to use the intervention more often.

Three of the 7 teachers from 2 schools completed the open-ended questions, of whom 2 from the same school experienced the intervention positively. They indicated that it was a well-organized intervention and that children enjoyed working with the robot. One teacher was less positive, highlighting the amount of administration and the fact that the children did not give positive feedback on the intervention. When asked if they had anything else to say about the intervention in the school setting, one mentioned that the possibilities of the intervention were not clear for one teacher, whereas another teacher from another school felt that it could be managed well within the school setting. One teacher suggested areas for improvement:

> *I think a robot should communicate with every pupil, asking a student a predetermined number of questions to prompt an internal dialogue in student. This will increase the "coaching" effect of the robot.*

Table . Perceived feasibility and acceptability of the intervention ePartners robot buddy according to teachers.

	Mean (SD)
FIM[a] score (range 4 - 20)	17.3 (2.6)
Additional feasibility questions (score; range 1 - 5)	
Guiding the child/children was feasible	4.1 (0.4)
It is feasible to set themes for the children	3.9 (0.7)
It was feasible to define the themes for the children together with the parents	3.6 (1.5)
I feel I have enough time to integrate the intervention into the school setting	3.3 (1.4)
To determine the themes, I had to learn a lot[b]	2.8 (1.6)
The child/children asked many questions about the content of the sessions[b]	2.5 (2.0)
The child/children asked many questions about using the robot[b]	1.5 (0.8)
AIM[c] score (range 4 - 20)	16.9 (3.8)
Additional acceptability questions (score; range 1 - 5)	
I am satisfied with the use of the intervention in the school setting	4.1 (0.7)
I would like to use the intervention more in the school setting	4.0 (0.8)

[a]FIM: Feasibility of Intervention Measure.

[b]A higher score means lower feasibility.

[c]AIM: Acceptability of Intervention Measure.

XSL•FO
RenderX

Acceptability and Usability According to Pupils

A total of 73 children were included and completed the pre- and postintervention questionnaires. The mean age was 9.9 years (SD 1.1 years, range 8 - 12), and the ratio of boys versus girls was nearly equal (41/73, 55.4% boys). Children reported high perceived usability of the intervention (mean 15.2, SD 2.4). The perceived acceptability of the intervention was also scored relatively high with a mean of 12.0 (SD 2.3).

The additional questions administered before each session indicate that most children found the sessions pleasant, reported that the session went good, and perceived the difficulty level as just right. On most occasions, a small majority of children indicated that they had not practiced at home with the robot. However, the children expressed the importance of engaging in the session with the robot (Table 2).

Table . Children's experiences with the ePartners robot buddy sessions.

Question and answer options	n (%)
What did you think of last week's session?	
Pleasant	163 (75.8)
Neutral	26 (12.1)
Unpleasant	8 (3.7)
Missing	18 (8.4)
How did last week's session go?	
Good	161 (74.9)
Neutral	31 (14.4)
Bad	5 (2.3)
Missing	18 (8.4)
Did you find the session difficult/easy?	
Too difficult	0 (0)
Somewhat difficult	30 (14.0)
Just right	103 (47.9)
Somewhat easy	39 (18.1)
Too easy	25 (11.6)
Missing	18 (8.4)
Did you try practicing with it at home?	
Yes	83 (38.6)
No	112 (52.1)
Missing	20 (9.3)
Did you find it important to learn the session from last week with the robot?	
Yes	185 (86.0)
No	12 (5.6)
Missing	18 (8.4)

Secondary Research Question: Preliminary Effects of the Robot

The results regarding mental well-being can be found in Table 3. The teacher-reported QoL of the children was significantly improved at postintervention as compared to baseline with a small to medium effect size. However, the children's self-reported QoL did not significantly change over time and had a small effect size. In addition, no significant changes in general mental well-being and self-efficacy scores of children were found.

XSL•FO
RenderX

Table . Data of secondary outcome measures at baseline and postintervention, along with the results of paired *t* tests of the preliminary effects of the ePartners robot buddy.

	Baseline, mean (SD)	Postintervention, mean (SD)	*t* test		Cohen *d*
			t (*df*)	*P* value	
General mental well-being					
How are you doing?	8.8 (1.6)	8.8 (1.5)	0 (64)	>.99	0
How are you doing at school?	8.3 (1.7)	8.6 (1.7)	1.10 (64)	.28	0.14
Quality of life					
Teacher-reported KIDSCREEN-10	36.0 (4.6)	37.2 (3.8)	2.77 (64)	.01	0.34
Children's self-reported KIDSCREEN-10	37.8 (3.5)	38.5 (3.6)	1.56 (61)	.10	0.21
Self-efficacy					
Self-reported SEQ-C[a]	28.9 (4.8)	29.8 (5.6)	1.62 (57)	.11	0.21

[a]SEQ-C: 8-item Self-Efficacy Questionnaire for Children.

Discussion

Principal Findings

The pilot study primarily investigated the feasibility, acceptability, and usability of a robot-delivered mental health–promoting intervention within a primary school setting. The results showed that the intervention was generally perceived as moderately feasible and acceptable according to both teachers and children, although several challenges were identified with the implementation and integration of the intervention into the school setting. More specifically, teachers were not always able to find the time to select the most suitable themes for the children or to consult with the child and parents about which sessions within these themes he/she would be best suited to attend. Furthermore, addressing children's questions regarding the content of the intervention sessions, as well as the amount of learning required to determine suitable intervention themes for the children, required a substantial time investment by the teachers.

The identified challenges with implementation and integration underscore the importance of considering practical barriers when introducing new programs into educational environments, such as primary schools. Both on a national and international scale, there is a shortage of teachers, leading to increased workload and corresponding limited resources [26,27]. Teachers' limited time availability for tasks related to adequate implementation of mental health–promoting interventions highlights the need for streamlined and efficient intervention protocols that minimize the burden on educators. Possibly, the social domain could provide support through training and coaching or coordinating resources, thereby supporting educational institutions and their staff. Another way to address such implementation challenges within school contexts is to actively engage teachers and school personnel in the development and implementation process of such interventions to ensure that the intervention aligns with the school's curriculum and resources. The significant time investment required from teachers to manage children's inquiries about the robot and the intervention content furthermore highlights the importance of carefully designing such interventions in co-creation with the target population, so that such programs meet their needs and abilities effectively. Another possibility is to enable the robot to select and conduct tasks with the child without the assistance of a human facilitator. Addressing these challenges is essential to ensure the successful integration and sustainability of mental health interventions within school settings, ultimately maximizing their potential benefits for children's well-being.

Regarding the preliminary effects of the intervention on children's mental well-being, the findings suggest that while the intervention may have a positive impact on teacher-perceived aspects of children's well-being, its effects on self-reported measures and broader mental health outcomes may be limited or require further investigation. The latter has been highlighted as a future research topic as well in recent reviews on robot-delivered interventions in the field of mental health [14,15,18]. Furthermore, it should be noted that there was no information available about which specific sessions the children attended within the intervention themes. It is possible that different sessions could have different impacts; this aspect warrants consideration for future research. In addition to studying the effects of mental health–promoting interventions delivered by robots, a more thorough investigation of specific needs of children and teachers is needed, for example, using a qualitative study design exploring end-user needs in more detail.

Strengths and Limitations

A strength of this study was its multischool design. However, the small sample size and potential selection bias due to children being selected for the intervention by their teachers limit the generalizability of the findings. In addition, the fact that 2 of

XSL•FO
RenderX

the 3 schools started the intervention after the summer holidays may have caused a bias when teachers completed questionnaires about the child, as their knowledge about the children's well-being may have been limited at that time and improved over the duration of the pilot.

Conclusions

This study provides valuable insights into the feasibility, acceptability, and usability of a robot-delivered mental health–promoting intervention within a primary school setting. However, further research is needed to fully understand its potential benefits and address existing limitations in terms of the implementation of such interventions in the primary school setting. In addition, the burden on teachers is a crucial factor in the feasibility of the intervention within the primary school setting. Therefore, the robot must be intelligent and personalized, enabling students to use it independently.

Acknowledgments

The study was funded by Top Consortium for Knowledge and Innovation - Life Sciences & Health (HealthHealth~Holland) as part of the EUREKA eHealth Call for R&D Project Proposals Switzerland – Netherlands. We would like to thank the participating schools for supporting and participating in the current pilot study. We would furthermore like to thank Olmo van der Mast and Maxime Postema for their work during the pilot tests at the schools in the Netherlands.

Data Availability

The datasets generated and/or analyzed during the current study are not publicly available due to ethical and privacy considerations. Anonymized data may be made available from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Description of the intervention themes.
[DOCX File, 14 KB - formative_v9i1e66797_app1.docx]

References

1. GBD 2019 Mental Disorders Collaborators. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet Psychiatry 2022 Feb;9(2):137-150. [doi: 10.1016/S2215-0366(21)00395-3]
2. Boer M, Dorsselaer SV, Looze MD, Roos SD, Brons H, Eijnden RVD, et al. HBSC 2021 Gezondheid en welzijn van jongeren in Nederland. : Universiteit Utrecht; 2022 URL: https://www.trimbos.nl/wp-content/uploads/2022/09/AF2022-HBSC-2021-Gezondheid-en-welzijn-van-jongeren-in-Nederland.pdf [accessed 2025-08-05]
3. Solmi M, Radua J, Olivola M, et al. Age at onset of mental disorders worldwide: large-scale meta-analysis of 192 epidemiological studies. Mol Psychiatry 2022 Jan;27(1):281-295. [doi: 10.1038/s41380-021-01161-7] [Medline: 34079068]
4. Fichter MM, Kohlboeck G, Quadflieg N, Wyschkon A, Esser G. From childhood to adult age: 18-year longitudinal results and prediction of the course of mental disorders in the community. Soc Psychiatry Psychiatr Epidemiol 2009 Sep;44(9):792-803. [doi: 10.1007/s00127-009-0501-y] [Medline: 19212695]
5. Otto C, Reiss F, Voss C, et al. Mental health and well-being from childhood to adulthood: design, methods and results of the 11-year follow-up of the BELLA study. Eur Child Adolesc Psychiatry 2021 Oct;30(10):1559-1577. [doi: 10.1007/s00787-020-01630-4]
6. McGee R, Williams S, Poulton R, Moffitt T. A longitudinal study of cannabis use and mental health from adolescence to early adulthood. Addiction 2000 Apr;95(4):491-503. [doi: 10.1046/j.1360-0443.2000.9544912.x] [Medline: 10829326]
7. Barnett K, Mercer SW, Norbury M, Watt G, Wyke S, Guthrie B. Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. The Lancet 2012 Jul;380(9836):37-43. [doi: 10.1016/S0140-6736(12)60240-2]
8. Magnano P, Craparo G, Paolillo A. Resilience and emotional intelligence: which role in achievement motivation. Int J Psychol Res (Medellin) 2016;9(1):9-20. [doi: 10.21500/20112084.2096]
9. Moisoglou I, Katsiroumpa A, Kolisiati A, et al. Resilience and social support improve mental health and quality of life in patients with post-COVID-19 syndrome. Eur J Investig Health Psychol Educ 2024;14(1):230-242. [doi: 10.3390/ejihpe14010015] [Medline: 38248135]
10. Keener TA, Hall K, Wang K, Hulsey T, Piamjariyakul U. Relationship of quality of life, resilience, and associated factors among nursing faculty during COVID-19. Nurse Educ 2021;46(1):17-22. [doi: 10.1097/NNE.0000000000000926] [Medline: 32941307]

XSL•FO
RenderX

11. Cal SF, de Sá LR, Glustak ME, Santiago MB. Resilience in chronic diseases: a systematic review. Cogent Psychol 2015 Dec 31;2(1):1024928. [doi: 10.1080/23311908.2015.1024928]
12. Dray J, Bowman J, Campbell E, et al. Systematic review of universal resilience-focused interventions targeting child and adolescent mental health in the school setting. J Am Acad Child Adolesc Psychiatry 2017 Oct;56(10):813-824. [doi: 10.1016/j.jaac.2017.07.780] [Medline: 28942803]
13. Fenwick-Smith A, Dahlberg EE, Thompson SC. Systematic review of resilience-enhancing, universal, primary school-based mental health promotion programs. BMC Psychol 2018 Jul;6(1):30. [doi: 10.1186/s40359-018-0242-3] [Medline: 29976252]
14. Robinson NL, Cottier TV, Kavanagh DJ. Psychosocial health interventions by social robots: systematic review of randomized controlled trials. J Med Internet Res 2019 May;21(5):e13203. [doi: 10.2196/13203] [Medline: 31094357]
15. Scoglio AAJ, Reilly ED, Gorman JA, Drebing CE. Use of social robots in mental health and well-being research: systematic review. J Med Internet Res 2019 Jul;21(7):e13322. [doi: 10.2196/13322]
16. Kabacińska K, Prescott TJ, Robillard JM. Socially assistive robots as mental health interventions for children: a scoping review. Int J of Soc Robotics 2021 Aug;13(5):919-935. [doi: 10.1007/s12369-020-00679-0]
17. Guemghar I, Pires de Oliveira Padilha P, Abdel-Baki A, Jutras-Aswad D, Paquette J, Pomey MP. Social robot interventions in mental health care and their outcomes, barriers, and facilitators: scoping review. JMIR Ment Health 2022 Apr 19;9(4):e36094. [doi: 10.2196/36094] [Medline: 35438639]
18. Dawe J, Sutherland C, Barco A, Broadbent E. Can social robots help children in healthcare contexts? A scoping review. BMJ Paediatr Open 2019 Jan;3(1):e000371. [doi: 10.1136/bmjpo-2018-000371] [Medline: 30815587]
19. Weiner BJ, Lewis CC, Stanick C, et al. Psychometric assessment of three newly developed implementation outcome measures. Implement Sci 2017 Aug;12(1):108. [doi: 10.1186/s13012-017-0635-3] [Medline: 28851459]
20. Borgers N. Response quality in survey research with children and adolescents: the effect of labeled response options and vague quantifiers. Int J Public Opin Res 2003 Mar;15(1):83-94. [doi: 10.1093/ijpor/15.1.83]
21. Casey P, Patalay P, Deighton J, Miller SD, Wolpert M. The Child Outcome Rating Scale: validating a four-item measure of psychosocial functioning in community and clinic samples of children aged 10-15. Eur Child Adolesc Psychiatry 2020 Aug;29(8):1089-1102. [doi: 10.1007/s00787-019-01423-4] [Medline: 31659441]
22. Ravens-Sieberer U, Gosch A, Rajmil L, et al. The KIDSCREEN-52 quality of life measure for children and adolescents: psychometric results from a cross-cultural survey in 13 European countries. Value Health 2008;11(4):645-658. [doi: 10.1111/j.1524-4733.2007.00291.x] [Medline: 18179669]
23. Muris P. A brief questionnaire for measuring self-efficacy in youths. J Psychopathol Behav Assess 2001 Sep;23(3):145-149. [doi: 10.1023/A:1010961119608]
24. White-Koning M, Grandjean H, Colver A, Arnaud C. Parent and professional reports of the quality of life of children with cerebral palsy and associated intellectual impairment. Dev Med Child Neurol 2008 Aug;50(8):618-624. [doi: 10.1111/j.1469-8749.2008.03026.x] [Medline: 18754901]
25. IBM SPSS statistics for windows, version 25.0. IBM Corp. 2017. URL: https://www.ibm.com/products/spss-statistics [accessed 2025-08-05]
26. Transforming education from within: current trends in the status and development of teachers. World Teachers' Day 2022. United Nations Educational, Scientific and Cultural Organization. 2022. URL: https://unesdoc.unesco.org/ark:/48223/pf0000383002 [accessed 2025-08-05]
27. Adriaens H, Elshout M, Elshout S. Personeelstekort primair onderwijs. : Centerdata Research Institute; 2022 URL: https://www.aanpaklerarentekort.nl/documenten/2022/12/8/personeelstekorten-primair-onderwijs-peildatum-1-oktober-2022 [accessed 2025-08-05]

Abbreviations

QoL: quality of life

Edited by A Mavragani; submitted 23.09.24; peer-reviewed by MJ Sarmiento-Suarez; revised version received 27.05.25; accepted 28.05.25; published 29.08.25.

Please cite as:
Zijp A, Aardoom JJ, Blanson Henkemans OA, van der Pal S, Vlasblom E, Versluis A
A Robot-Delivered Training Program to Improve Children's Mental Health and Resilience in Dutch Primary Schools: Pilot Intervention Study
JMIR Form Res 2025;9:e66797
URL: https://formative.jmir.org/2025/1/e66797
doi:10.2196/66797

XSL•FO
RenderX

© Anne Zijp, Jiska J Aardoom, Olivier A Blanson Henkemans, Sylvia van der Pal, Eline Vlasblom, Anke Versluis. Originally published in JMIR Formative Research (https://formative.jmir.org), 29.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Feasibility, Acceptability, and Preliminary Outcomes of a Mobile Adaptation of a Relational Savoring Intervention to Prevent Loneliness in College Students: Mixed Methods Pilot Study

Brenda Nguyen[1], BA; Jocelyn Lai[1], PhD; Hana Qureshi[1], BA; Christopher Marcotullio[1], BA; Sina Labbaf[2], PhD; Yuning Wang[3], PhD; Salar Jafarlou[2], PhD; Nikil Dutt[2], PhD; Amir M Rahmani[2,4], PhD; Jessica L Borelli[1], PhD

[1]Department of Psychological Science, School of Social Ecology, University of California, Irvine, 4201 Social and Behavioral Sciences Gateway, Irvine, CA, United States

[2]Department of Computer Science, Donald Bren School of Information and Computer Sciences, University of California, Irvine, California, United States

[3]Department of Computing, University of Turku, Turku, Finland

[4]Sue and Bill Gross School of Nursing, University of California, Irvine, California, United States

Corresponding Author:
Jessica L Borelli, PhD
Department of Psychological Science, School of Social Ecology, University of California, Irvine, 4201 Social and Behavioral Sciences Gateway, Irvine, CA, United States

Abstract

Background: Rates of loneliness have risen sharply since the onset of the COVID-19 pandemic, largely due to disruptions in social relationships and daily routines, with college students experiencing some of the greatest increases. While prevention programs targeting loneliness have been developed, their success has been limited. One promising approach may lie in enhancing the quality of existing relationships rather than simply increasing social interactions during periods of acute loneliness. Relational savoring, an intervention rooted in attachment theory and positive psychology, aims to deepen feelings of connection by encouraging individuals to reflect on positive interpersonal experiences.

Objective: This study aimed to evaluate the feasibility, acceptability, and preliminary outcomes of a mobile health adaptation of relational savoring, termed mSavorUs (developed by Amir Rahmani), designed to prevent and reduce loneliness among college students.

Methods: A randomized controlled pilot study was conducted with a diverse sample of 29 college students (43.3% Latinx, 40% Asian American, and 16.7% White). The intervention leveraged a smart ring, smartwatch, and smartphone app to enable just-in-time delivery of relational savoring prompts, alongside continuous monitoring of loneliness-related indicators (eg, physiological activity, sleep, and behavior). Aim 1 involved a thematic analysis of participant feedback regarding the utility, benefits, and challenges of both mSavorUs and the monitoring tools. Aim 2 examined the intervention's effects on loneliness and perceived connectedness.

Results: For aim 1, qualitative findings suggested that participants found the content of mSavorUs (developed by Amir Rahmani) rewarding and helpful; however, the timing of the intervention was often experienced as disruptive. For aim 2, quantitative analyses revealed no significant reductions in loneliness or increases in connectedness, indicating the need for adjustments to the intervention delivery method.

Conclusions: Although participants found the intervention content valuable, the just-in-time delivery format may have limited its effectiveness. Future iterations should consider alternative timing or delivery strategies to maximize program benefits.

(JMIR Form Res 2025;9:e70528) doi:10.2196/70528

KEYWORDS

relational savoring; loneliness; connectedness; mobile; college students

Introduction

Background

As of 2019, approximately 60% of American adults met the clinical criteria for loneliness [1], marking it as a significant public health concern [2,3]. Loneliness is a subjective experience that may stem from a lack of social contact [4] or from the perception that one's social relationships are insufficient in quality or quantity [5]. It is associated with a range of negative physical and mental health outcomes, including increased risk of mortality [6,7]. Although research on loneliness has

XSL•FO
RenderX

traditionally focused on older adults, recent global data indicate that young adults report higher levels of loneliness than older individuals [8]. Young adulthood is a developmental period marked by both opportunity and vulnerability. Challenges related to identity development, forming new relationships, living independently, managing financial responsibilities, and navigating academic pressures may heighten feelings of loneliness during this stage of life [9]. At the same time, this period presents a valuable window for establishing healthy coping and relational habits. Accordingly, this study aimed to evaluate the effectiveness of a novel, in-the-moment intervention designed to reduce loneliness and enhance connectedness among young adults.

Existing Loneliness Prevention and Intervention Programs

Interventions aimed at preventing or reducing loneliness have shown some effectiveness among youth and emerging adults [10]. Many existing programs focus on promoting successful social interactions and fostering new social connections [11]. However, despite their intuitive appeal, a meta-analysis found that the overall effect sizes of these interventions on reducing loneliness are small [12]. These findings suggest that simply increasing the quantity of social interactions or perceived social support may not sufficiently address the subjective experience of loneliness. Instead, individuals struggling with loneliness may benefit more from interventions that help them derive deeper meaning from existing relationships, particularly when delivered at moments of heightened need.

Loneliness can persist even in the presence of frequent social interaction if negative social cognitions diminish the perceived quality of relationships. If individuals do not feel safe, secure, or supported by others, they may continue to feel lonely despite being socially engaged. Therefore, interventions that enhance perceptions of relational quality, by fostering feelings of connection, security, and emotional closeness, may represent a more effective approach. However, limited research has explored the efficacy of such strategies for alleviating loneliness, underscoring the need for further investigation in this area.

Relational Savoring

Grounded in attachment theory, relational savoring is a brief, guided intervention designed to help individuals reflect on positive memories with relational and attachment-relevant content [13,14]. This technique encourages participants to re-experience moments in which they either provided or received safety, security, and support, thereby amplifying the positive emotions and cognitions associated with those experiences [14]. These reflections may center on significant milestones (eg, saying goodbye to parents before leaving for college) or everyday interactions that might otherwise go unnoticed (eg, assisting an older person with loading groceries). The relational savoring protocol follows 5 standardized steps, including sensory focus, emotional focus, meaning-making, future orientation, and mind-wandering, to facilitate deep and intentional engagement with the memory.

Relational savoring has been evaluated in multiple randomized controlled trials across diverse populations, including parents of young children [15], individuals in long-distance relationships [16], adolescents in residential treatment [17], and older adults [18]. The intervention has been delivered both via a website [15,16] and in person [17-19], demonstrating benefits, such as increased positive emotion, enhanced feelings of closeness, greater relationship satisfaction, and reduced physiological reactivity [15-18]. Importantly, relational savoring appears effective even when the other person in the target relationship is not physically present [20]. By guiding individuals to focus on meaningful moments of connection, relational savoring promotes a deeper sense of closeness and increases help-seeking behavior [16]. In doing so, it may help individuals appreciate the strengths of their existing relationships, thereby offering a promising strategy for preventing and alleviating loneliness.

Increasing Accessibility Through Digital Just-in-Time Interventions

Intervention accessibility is important to young adults; with the use of digital mental health tools on the rise over the past few years [21], several technology-based interventions addressing loneliness among young adults have also been explored by adapting existing interventions into online platforms. These digital interventions have been shown to provide similar benefits to the original interventions, particularly increasing access to low-income individuals and young adults who may prefer digital tools [21,22]. Just-in-time (JIT) interventions that leverage digital tools are designed to carefully time the delivery of interventions to moments when they may be relevant to increase the efficacy of the intervention [23]. Such approaches aim to provide appropriate real-time support in moments of need by tracking changes in individuals' internal state based on various cues, such as physiological changes [23]. Studies have shown that JIT interventions can have greater efficacy than non-JIT interventions [24,25].

Currently, the use of JIT interventions to alleviate feelings of loneliness and increase feelings of connectedness has not been tested. With 38% of US teenagers and young adults reporting that they already use mobile apps for well-being [26], there is reason to believe that the development of effective mobile-based interventions for loneliness may be feasible, familiar, and cost-effective.

Overview of Study Aims

Given the growing risks associated with heightened feelings of loneliness and the limited effectiveness of existing interventions, a more accessible and targeted approach focusing on individuals' subjective perceptions of their relationships may offer a promising path forward. To this end, we developed mSavorUs, a smartphone app designed to deliver relational savoring in a mobile format aimed at preventing and reducing loneliness among emerging adults. Building on prior work developing personalized, predictive models of mental health and well-being [27-30], we integrated a ubiquitous monitoring system that combined ecological momentary assessments (EMAs), passive mobile sensing, and wearable physiological sensors to capture real-time indicators of health and emotional state. The design of this system was predicated on well-established links between sleep disturbances [31], physical activity [32], and physiology [33,34] with loneliness. This system was paired with a proactive

XSL•FO
RenderX

mobile intervention that delivered relational savoring exercises JIT—specifically when the monitoring system detected potential episodes of loneliness. We pursued two primary aims:

1. Aim 1: Qualitatively assess the feasibility and acceptability of mSavorUs by evaluating participants' perceptions of the utility, benefits, and challenges of each feature of the intervention, including the monitoring systems.
2. Aim 2: Quantitatively evaluate the preliminary impact of mSavorUs on loneliness and social connectedness. We hypothesized that participants randomized to the mSavorUs intervention group would report greater reductions in loneliness and greater increases in feelings of social connectedness compared to those in the monitoring-only control group.

Methods

Participants

Undergraduate students were recruited from a large West Coast university in the United States. Recruitment was accomplished through the distribution of flyers and posting announcements about the study on class web pages for large undergraduate courses. Participants were eligible if they were between the ages of 18 and 22 years and were fluent in English. Due to the system requirements of the app designed for the study, participants were also required to use an Android smartphone with an operating system of 6.0 or higher. To create a more homogenous sample reflecting emerging adulthood, participants could not be parents, married, or returning to school after a 3-year or longer absence. Individuals who had participated in a similar previous study were also ineligible. We recruited a pilot sample size. A total of 37 students were enrolled in the study. Eight withdrew participation over the course of the study (n=29; mean 19.93, SD 1.22 years; 13 male, 16 female). Of the remaining participants, 43.3% (n=13) of participants identified as Latinx, 43.3% (n=12) as Asian American, and 17.2% (n=5) of participants identified as White.

Procedure

Participants were enrolled in the study between late January and early February of 2021. Study procedures lasted 22 weeks. First, participants completed an in-person baseline assessment, reporting on baseline levels of loneliness and social connection. They were provided with wearable smart devices (ie, Oura Ring and Samsung Watch) and installed study-relevant phone apps (ie, *m*SavorUs, AWARE [developed by AR], and wearable device apps). Table S1 in Multimedia Appendix 1 provides a brief overview of the apps and devices. Each participant was provided with a study ID that had a preassigned group through a randomization generator. Participants were assigned their IDs based on their enrollment order and were unaware of what group they were placed into.

After enrollment, participants completed a 6-week-long monitoring period, wearing smart devices and completing 5 interval-based EMAs daily. The EMAs consisted of brief questions that assessed emotions, feelings of loneliness, and social connectedness. Participants were asked between 20 and 40 questions in each EMA assessment (eg, morning surveys asked about sleep quality), but this study focuses on just 2 questions asked in each EMA assessment.

The algorithm was developed using objective data collected from smartphones, smartwatches, and smart rings (for algorithm details, see [35]). During the monitoring period, these data were analyzed for associations with self-reported loneliness, as measured via EMAs. Objective features included various sleep metrics, heart rate variability, phone usage patterns, physical location, and types of phone notifications. A binary classification label (lonely vs not lonely) was generated based on median EMA-reported loneliness scores. Personalized random forest models were trained for each participant, using the first 50% of their individual data, while incorporating data from other participants to train the model and the remaining 50% for testing. This hybrid approach captured both individual and group-level patterns, yielding an average model accuracy of 82% (see [35] for further details). During the intervention phase, the algorithm was used to trigger mSavorUs prompts when participants' objective data indicated potential loneliness risk. Prompts could be delivered between 9 AM and 9 PM, regardless of the participant's current context (eg, at home, in class, shopping, or resting).

Participants then entered the next phase of the study, which lasted 4 weeks. During this time, the monitoring group (n=14) continued with the same procedures while the relational savoring group (n=15) additionally completed the JIT intervention. Specifically, the mSavorUs group was prompted through the app to complete the relational savoring intervention during periods of suspected loneliness elevations, determined by sensing-based indicators that had previously been associated with their own above-median level of loneliness during the monitoring phase (ie, personalized algorithm). The timing of the JIT intervention was guided by continuous monitoring of each participant's psychological and physical state using a personalized mental health navigator framework, described above [29,35]. When objective indicators or EMA loneliness reports exceeded a participant-specific threshold usually above the median within a defined time window, the system delivered a JIT prompt to complete the mSavorUs intervention. If the participant did not respond to the JIT prompt by completing the intervention, they would receive a second and a third prompt within that hour-long interval.

To reduce participant burden, if a participant's sensing data indicated the potential for elevated loneliness multiple times within an hour-long interval but they had already completed the intervention, they would not be prompted to complete it again. At the end of the 4-week intervention period, both groups completed a midpoint assessment, like that of the baseline assessment. All participants then completed 12 weeks of continued monitoring. During this phase, no intervention was provided, and the number of daily EMAs was reduced to 3 in order to reduce participation burnout. An exit assessment was completed at the end of the study during an in-person session where participants returned their devices and uninstalled all study-related phone apps. The mSavorUs group completed an interview about their experiences with the intervention. The study timeline is depicted in Figure 1.

XSL•FO
RenderX

Figure 1. Study Timeline. EMA: ecological momentary assessment; JIT: Just-in-Time.

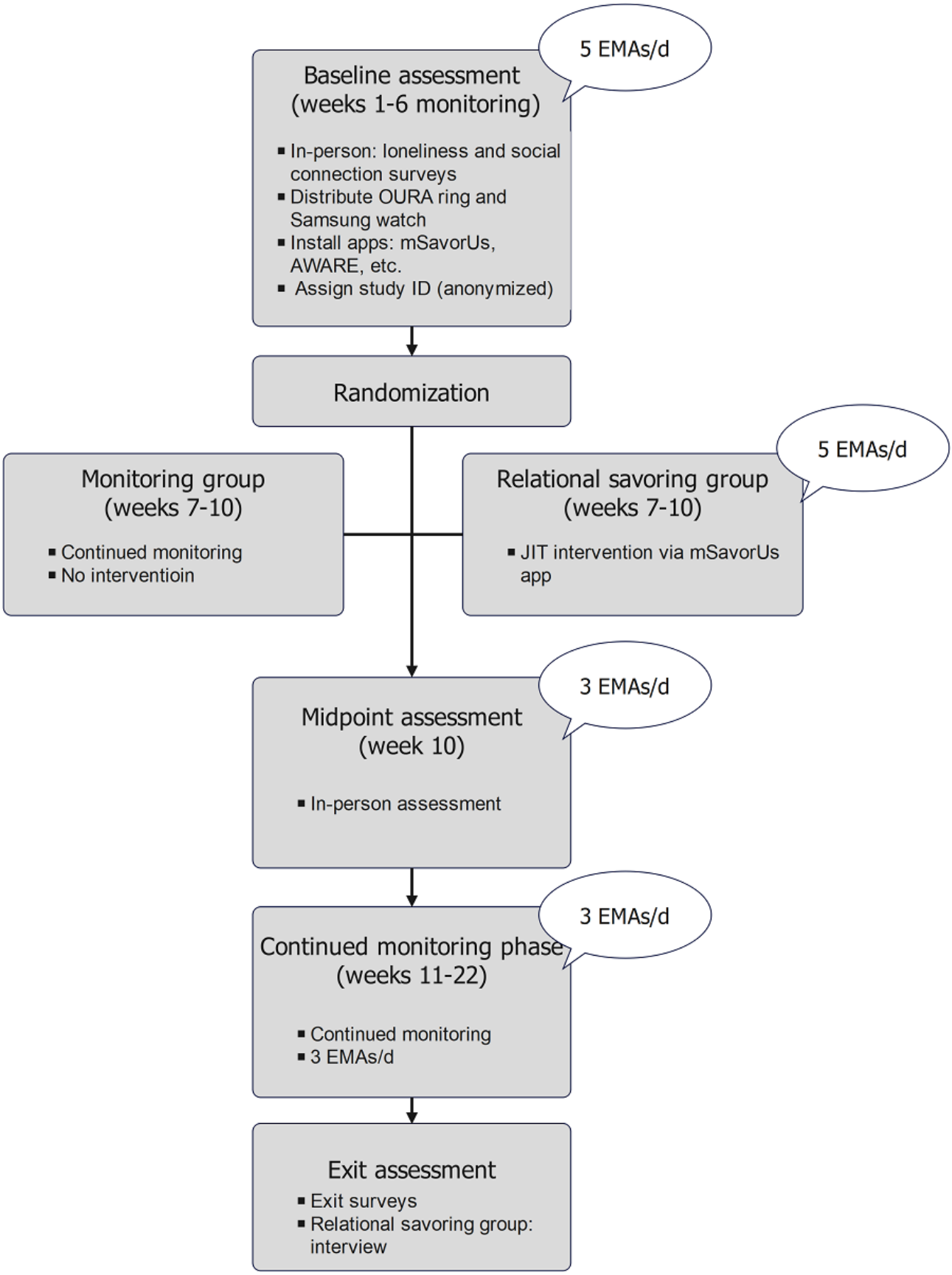

mSavorUs Intervention

Once randomized to the mSavorUs condition, participants met for an individual session with a researcher in which they were guided through an in-person administration of relational savoring, following the protocol established in prior research [13]. Following this in-person session, researchers demonstrated the use of the mSavorUs app. Researchers explained how the

XSL•FO
RenderX

app notifications would work and how participants could respond to the notifications by completing the intervention.

Each time they completed the intervention on the mSavorUs app, pursuant to the protocol of previous administrations of relational savoring [13], participants first completed a brief mindfulness breathing exercise. Then they provided written responses to a series of 5 prompts designed to build appreciation for moments of relational connection. The 5 prompts corresponded to the 5 reflection steps used in prior trials of relational savoring [13]. In this study, participants were asked to think of different memories to reflect on each time they used the mSavorUs app. Participants also had the option to upload pictures related to the memories they shared and had access to them at all times while using the app.

Measures

Feasibility and Acceptability of Relational Savoring Intervention.

At the exit assessment (after week 22 of study participation), the mSavorUs group completed brief, semistructured individual interviews in which they answered open-ended questions about their experience. Interviews were audiorecorded, transcribed, and had notes taken during the process. Interviews consisted of 22 questions probing feasibility (eg, What barriers did you encounter while engaging in this activity?) and acceptability (eg, What did you like about this activity? What did you dislike?). Trained researchers conducted the interviews, which focused on questions regarding the utility, benefits, and challenges of each feature of the intervention along with their experience with all the features of the monitoring systems (for complete list of questions, see Table S2 in Multimedia Appendix 1).

Baseline, Midpoint, Exit Assessment of Loneliness and Social Connection

University of California Los Angeles Loneliness Scale

The University of California Los Angeles (UCLA) Loneliness Scale is a widely used 3-item measure of perceived loneliness (“How often do you feel you are left out?” [36]). Items were rated on a 3-point Likert scale (1=hardlyever, 2=some of the time, 3=often) and summed, where higher scores reflect greater loneliness (Cronbach α at baseline=0.69, at midpoint=0.58, at endpoint=0.68).

The Social Connectedness Scale

This 8-item scale assesses the degree to which individuals feel connected to others in their social environment “I feel disconnected from the world around me” [37]. Items are rated using a 5-point Likert scale (1=strongly agree, 6=strongly disagree, reverse-scored and summed, where greater scores indicate higher levels of perceived social connectedness ($\alpha_{1baseline}$=0.95, $\alpha_{2midpoint}$=0.97, and $\alpha_{3endpoint}$=0.95).

Multimodal Assessments of Health, Feelings, and Behavior

Physiology, Sleep, and Behavioral Patterns

Participants were fitted with the Oura Ring and Samsung Gear Sport smartwatch and downloaded the corresponding mobile apps in an effort to capture an accurate depiction of their daily physical habits, sleep, and health. The Oura ring was used in this study to assess sleep quality by measuring sleep duration, average heart rate during sleep, and heart rate variability during sleep [38]. The Samsung Gear Sport smartwatch was used in this study to measure daily heart rate and daily heart rate variability using sensors and a pedometer system [39].

Mobile Activity

The AWARE app [40,41] passively collected biometric data and logged daily routines on participants’ smartphones. The data collected includes social relationship data (eg, amount of time spent with other people), daily rhythms (eg, routines such as going out vs staying in one’s home), and phone interactions (eg, amount of time spent texting, calling, and app browsing). The AWARE framework has been studied and validated [42]. The connection between the AWARE app and research servers was encrypted to ensure the privacy of participants’ data.

Momentary Loneliness and Social Connectedness

Throughout the EMA portion of the study, participants completed assessments of loneliness and perceived social connectedness daily on the mSavorUs app. Loneliness was assessed with a single item, “How lonely do you feel right now?” which was rated on a sliding scale (1-100). Social connectedness was assessed with a single item, “How connected do you feel to others right now?” rated on the same scale.

Data Analytic Plan

Aim One: Feasibility and Accessibility

To address “aim one” and evaluate the feasibility of the JIT mSavorUs intervention, we completed thematic analysis (TA; [43]) applied to the qualitative data. The team of coders (n=6), which comprised postbaccalaureate research staff and undergraduate research assistants, reviewed themes generated from the transcribed interviews. The team followed these steps: (1) identifying unique responses, (2) agreeing upon common themes and subthemes, (3) assigning subthemes to each response, (4) calculating agreement among raters, and (5) reviewing responses.

Aim Two: Preliminary Quantitative Outcomes

To evaluate changes in feelings of loneliness and connectedness, we first examined whether self-reported loneliness and connectedness varied across the 3 assessment points (ie, initial monitoring phase [baseline], intervention phase [midpoint], and continued monitoring phase [endpoint]). A repeated measures ANOVA was used to compare changes in scores across assessment points and groups (see Section A in Multimedia Appendix 1 provides information regarding procedures for missing data).

XSL•FO
RenderX

We also examined whether there were changes in momentary loneliness and connectedness across the study period. Multilevel models were conducted in RStudio (Posit, PBC; [44,45]). Our model consisted of 3 levels, with multiple assessments (Level-1) nested within day (Level-2) within persons (Level-3). Group assignment (control group vs mSavorUs group), study phase (initial monitoring, intervention, and continued monitoring), day, and assessment number were included as predictors. An interaction term was included for the group and study phase to test changes over the study phase by group. Day was the total number of days assessed, which enabled us to consider trends over time (154 d in total). The EMA number referred to the assessment for that day, which enabled us to consider trends over the day (5 assessments, later changed to 3, per day). Both survey day and EMA were included as a fixed effect (covariate) to account for change over the assessment as well as how much people's loneliness and social connectedness may vary within a day.

Ethical Considerations

The study protocol was approved by the institutional review board at the University of California, Irvine (IRB number 2019‑5153). Participants provided informed consent to participate. All participants were assigned a participant identifier, and all data were stored with that identifier. Identifying data were collected from participants for the purpose of contacting them to schedule visits. Identifiable data were stored on a secure network to which only the research team had access. Participants were compensated according to the following schedule (US $30 for the baseline assessment US $60 for the 4-week observation period, US $120 for the 4-week intervention period, US $30 for the postintervention assessment, US $80 for weeks 1‑4 of postintervention monitoring, US $100 for weeks 5‑8 of postintervention monitoring, US $120 for weeks 9‑12 of postintervention monitoring, and US $30 for the final follow-up assessment). In total, participants could earn up to US $645 for completing all study activities. Participants could also earn up to US $75 in bonuses (US $15 every 4 weeks) for maintaining at least 90% adherence (eg, device use and survey completion) during each of the 5 monitored periods. Compensation was distributed via email.

Results

Technological Challenges and Participant Adherence

Throughout the study, several technological challenges impacted data collection and participant engagement. These included temporary device disconnections, battery depletion, and conflicts with mobile app permissions, which occasionally disrupted passive data streams. The most significant technical issue occurred following the initial 4-week monitoring period when a technical error delayed the intervention rollout, necessitating a 2-week extension of the monitoring phase for all participants.

In addition, some participants encountered difficulties submitting their EMAs and intervention responses via the mSavorUs app. These challenges were likely due to factors such as unavailable Wi-Fi during submission, limited data storage capacity in the cloud database, or technical issues associated with the mSavorUs app and related apps (eg, AWARE and the Oura Mobile app). For instance, on at least 2 occasions, data from AWARE were not being received, requiring the research team to maintain close communication with participants to troubleshoot the problem. In one case, a participant withdrew from the study due to the burden of managing technical difficulties. Other participants experienced issues submitting surveys or intervention responses through mSavorUs, prompting the computer science team to investigate and resolve the issues through direct interaction with affected individuals.

Additional technical challenges included difficulties syncing the Oura ring and situations where participants who obtained new phones during the study needed assistance reconnecting their devices to the apps and integrating data sources. Generally, an increase in technological issues appeared to correspond with greater participant fatigue and burden.

To address these challenges, we implemented a proactive data monitoring strategy via a centralized dashboard. This approach allowed for real-time tracking of data integrity and facilitated prompt issue resolution. A dedicated team of psychology research assistants and computer science technicians collaborated closely, addressing most technical issues as they arose and maintaining participant engagement throughout the study.

Adherence Rates

The total number of possible observations per individual was 607, comprising 210 prompts during the monitoring period, 140 during the intervention period, and 257 during the continued monitoring period. The overall adherence rate for EMA assessments across the study was 69% (mean 417.96, SD 137.63, range 87‑579 completed out of 602 possible observations). During the initial monitoring phase, adherence was 79% (mean 166.04, SD 35.00, range 60‑202 completed out of 210 possible observations). During the intervention phase, adherence was 75% (mean 105.57, SD 35.97, range 14‑139 completed out of 140 possible observations). During the continued monitoring phase, adherence declined to 57% (mean 146.36, SD 76.09, range 7‑245 completed out of 257 possible observations), indicating some decrease over the extended 12-week period.

Intervention Completion

Among intervention group participants, adherence to completing the intervention was moderate. Once a JIT prompt was triggered, participants who did not complete the mSavorUs intervention received the prompt up to 3 additional times to maximize completion. During the 4-week intervention period, 1205 notifications were sent, with 219 responses submitted, yielding an 18.17% response rate. The response rate across the 3 JIT prompts was relatively consistent: first prompt (16.02%), second prompt (21.25%), and third prompt (19.59%).

Aim One: Feasibility and Accessibility

Overall, 2 of 11 participants opted out of being recorded for the postintervention interviews. As a result, 5 research members completed a review of all 385 video transcriptions and notes and compared them. A total of 10 responses only had transcripts with no notes. They confirmed the transcripts and notes matched in content and confirmed the themes and percentage agreements

XSL•FO
RenderX

were not affected by either version. As a result, notes were used for the TA procedure while transcripts were mainly used for quoting purposes.

The team identified a total of 223 unique responses and agreed upon common overarching themes and subthemes. Once the themes were identified, all 223 responses were assigned a subtheme. A single response could be assigned to more than one subtheme. The percentage of agreement was computed as the number of matching subthemes present for a single response divided by the number of raters (eg, one response was assigned the subtheme "helpful" by 4/6 raters, yielding a percentage agreement of 67%). In addition, responses with a percent agreement of 50% were discussed to resolve the tie, and ratings were then recalculated. Only data that achieved 80% agreement and higher were included.

Of the 223 responses, 167 (74.89%) unique responses reached a minimum 80% interrater agreement while 56 (25.11%) responses did not and thus were excluded. After factoring in that some responses were coded with multiple themes, a total of 269 responses were included. The TA procedure revealed 3 core themes: benefits, barriers, and improvements. Some of the questions covered during the interviews contained responses that did not fall under any core theme, so they are presented separately (see Section B in Multimedia Appendix 1). A complete list of quotes for each theme and subtheme and their frequencies can be found in Tables S3 and S4 in Multimedia Appendix 1.

Benefits

Sample quotes illustrating the core theme of Benefits can be found in Table 1. The subthemes that underlie the overarching theme of benefits included "helpful," "positive feelings," "calming, relaxing, meditative," "focused on the moment," and "reflect and appreciate good memories." "Helpful" was a general term participants used to describe a variety of experiences with the mSavorUs intervention (ie, it helped participants feel more positive when they were experiencing stress). Interesting, the majority of the comments that were coded as helpful referred to the introductory session led by researchers where they introduced mSavorUs. "Positive feelings" reflected times when participants felt happier, more joyful, or generally more positive after completing the mSavorUs intervention. "Calming, relaxing, meditative" referred to the state of being elicited by the mSavorUs app—it induced a sense of peacefulness in the participants. Participants' responses indicated that they enjoyed the mSavorUs intervention because it prompted them to take a moment to step away from daily life and reflect on positive memories, reflected in the code, "focused on the moment." Finally, "reflect and appreciate good memories" described how participants were able to reflect on positive memories in a more detailed fashion that made them appreciate and enjoy the memory more than before.

XSL•FO
RenderX

Table . Benefits core theme and subthemes.

Subthemes	Quotes	Frequency (%)
Helpful	*I forgot who I had done it with before, but he helped me understand how I should identify more happy moments and make sure there's no negative and only has to do with positive and so when I think about those moments even if there's a little bit of negative I make sure to only focus on the positive.*	13.75
Positive feelings	*It's like having a tough day, you know. It helps to like kinda force yourself to think more positively, because I know at least in these past 2 weeks it's been a little bit hectic for me so it's kinda nice to take a moment to think more positively so I think it was kinda helpful to just [improve] the overall mood.*	3.35
Calming, relaxing, meditative	*What I liked about it was, I liked the mediation aspects. I mean I realised that like, through doing it that I don't ever really sit down with my thoughts, so it was nice to do that.*	1.86
Focused on the moment	*I liked how it focused on focusing on different senses, and like what emotions you're feeling. I feel like I don't usually focus on that stuff in everyday life.*	0.74
Reflect and appreciate good memories	*Oh yeah I think it was because you know I have all these memories in my head, you know. And I can think about them, whenever I want, but I think putting it like, actually writing it down and writing down every single like part of it, and it felt like I was explained the memory to someone else that it was just a lot more fun and it was much more easier to actually like make the memory tangible, I guess, I don't know if that's the right word, but it is a lot more, felt more real than just like thinking about [it] in my head like actually getting to like write it down and explain it to someone else made it more interesting.*	3.35

Barriers

Quotes illustrating the core theme of "Barriers" can be found in Table 2. A variety of subthemes relating to barriers included "high frequency," "tech issues," "lack of memories," "length of activity," "time-window," and "busy." "High frequency" expressed the participants' concern with the number of times they were prompted to complete the mSavorUs intervention. "Tech issues" entailed any problems relating to the mSavorUs app (eg, activity submission problems or the app inaccurately stating an activity is available vs unavailable). "Lack of memories" reflected participants having difficulties generating new positive memories to reflect on. "Length of activity" described the activity as taking too much of the participants' time to complete. Similarly, "time-window" was the concern that participants were unable to complete the activity within the given 1-hour window starting from the first prompt notification they received to complete it. Finally, "busy" represented the participants' inability to complete the activity due to their daily schedules.

XSL•FO
RenderX

Table . Barriers core theme and subthemes.

Subthemes	Quotes	Frequency (%)
High frequency	*…when I was getting five or six a day… I was getting a little bit irritated because I like couldn't think of new memories… But once they reduced it to three like I was feeling better about it.*	5.95
Tech issues	*I tried to do them every single time but there always seemed to be problems, with, like, I would go through and write a super detailed response to each, and then at the end it wouldn't submit, so like I would go through again, write shorter answers, submit again, and a lot of the time it wouldn't even send through, and then the ones that did were often like super short responses.*	4.83
Lack of memories	*One thing that I did find was that it was hard for me to keep coming up with positive memories, I don't know, it was hard for me to brainstorm them after a while. Or they became really vague, like if they happened a long time ago, because I don't remember specific details. So I think because it was so frequent it was hard to come up with different memories.*	9.29
Length of activity	*Probably just how long it took yeah because you need to be pretty detailed.*	1.49
Time window	*As I mentioned before, um, giving only an hour to complete it. And so I'm getting that notification while I was still asleep, and not even done in the morning check-in um. It made it so I wasn't able to complete the activity every time I got the notification.*	6.32
Busy	*Sometimes I dislike the timing. I understand that it was supposed to be random but sometimes it came at inconvenient times like I might be texting my friends to go do something and I see it pop up but like I am in the middle of doing something at the same time like a class so it's a little bit inconvenient at times within that like hour.*	14.13

Improvements

Quotes illustrating the core theme of "Improvements" can be found in Table 3. The subthemes relating to improvement were related to the barriers listed above, such as "less tech issues," "lower frequency," "wider time window," "decrease activity length," and "change activity questions." "Less tech issues" described the participants' desire to encounter fewer problems when engaging with the mSavorUs app, such as being able to submit responses without any issues. "Lower frequency" represented the desire to reduce the number of prompt notifications participants received within a single day. "Wider time-window" entailed the participants wanting to have more time to complete the activity than the allotted 1-hour window. On a similar note, "decreased activity length" expressed the desire for the mSavorUs intervention to be shorter and to take up less time to complete (eg, reducing the number of questions asked and requiring fewer details for each response). "Change activity questions" involved participants wanting to modify some of the questions in the mSavorUs intervention to have more inclusive answer choices or for the activity to feel less redundant upon each completion.

XSL•FO
RenderX

Table . Improvements core theme and subthemes.

Subthemes	Quotes	Frequency (%)
Less tech issues	*I don't know, I did like the formatting of it, it felt very easy, so that's good. If it could be, you know, the technical issues worked out.*	4.46
Lower frequency	*I think it was a good feature, I think maybe just like decreasing the frequency.*	8.92
Wider time-window	*Maybe giving a broader um amount of time to complete the activity, because I don't know I couldn't do it during my class or stuff like that, because it takes like a moment of like just pure reflection. So giving up like a period to do it within longer than an hour I think, would be helpful.*	7.81
Decrease activity length	*I think shorter activities might have been nice.*	6.32
Change activity questions	*Maybe something like changing out the problem solving. I think probably the reason it felt like a chore was it was the exact same thing over and over and over. Just different name for my memory so maybe like changing out the questions, a little bit like think of a memory of a time you specifically felt this way or think of a time you specifically did this thing or you know, a time you felt excited versus spontaneous so you know different different questions, I think, would have made it more intriguing.*	7.43

Aim Two: Preliminary Quantitative Outcomes

Table S5 in the Multimedia Appendix 1 includes the means across groups and study phases. For the 2×3 repeated measures ANOVAs, we did not find group differences in reported loneliness over the 3 assessment periods ($F_{2,54}$=0.50; P=.61; η^2=0.02), nor were there any differences in reported loneliness across the assessment period overall ($F_{2,54}$=0.17; P=.84; η^2=0.006). We also did not find group differences in social connectedness ($F_{2,54}$=0.72; P=.49; η^2=0.03), nor were there any differences in social connectedness across the period overall ($F_{2,54}$=0.73; P=.49; η^2=0.03).

Changes in Momentary Loneliness Over the Assessment Period

Loneliness did not change over time or over the course of the entire assessment period (β=−0.02, 95% CI −0.04 to −0.00; P=.06), but loneliness did change over the assessments per day (β=1.66, 95% CI 1.47-1.85; P<.001). Participants reported greater feelings of loneliness by the evening EMA or in EMA surveys sent later in the day.

Groups did not differ in momentary loneliness (β=−1.17, 95% CI −9.72 to 7.38; P=.79). We did, however, find significant group-by-study phase interaction in momentary loneliness between the initial monitoring period and intervention period (β=2.00, 95% CI 0.17-3.83; P=.03), and the intervention period and continued monitoring period (β=−2.40, 95% CI −4.21 to −0.59; P=.009). In probing the simple slopes, for the control group, there were decreases in loneliness during the intervention period compared to the initial monitoring period (β=−3.66; P<.001). Furthermore, there was an actual increase in loneliness for the control group during the continued monitoring period compared to the intervention period (β=1.90; P=.02). The control group did not differ in their loneliness at the initial monitoring phase and the continued monitoring phase (β=−1.76; P=.09).

For the mSavorUs group, they also reported decreases in loneliness during the intervention period compared to the initial monitoring period (β=−1.66; P=.03) as well as during the continued monitoring phase compared to the initial monitoring period (β=−2.16; P=.04). There was no difference in loneliness in the intervention period and continued monitoring period for the Relational savoring group (β=−0.50; P=.55). In other words, the mSavorUs group reported an overall decrease in loneliness from the start of the study (see Tables S6-S9 in Multimedia Appendix 1 for full results). Mean levels of loneliness (Figure 2) for each group are shown across the 3 study phases. Shaded error bands reflect the 95% CIs based on group-wise SEs.

XSL•FO
RenderX

Figure 2. Loneliness across the study period. RS: Relational Savoring.

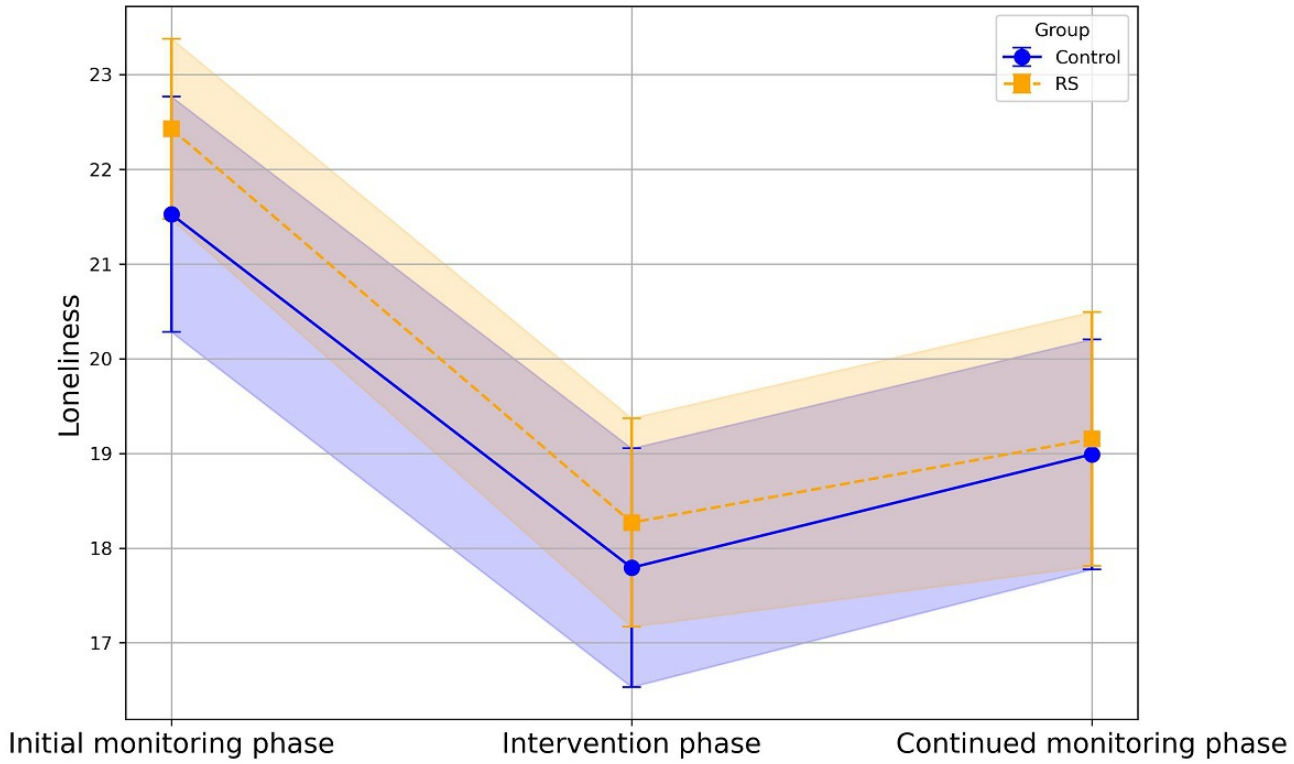

Changes in Momentary Connectedness Over the Assessment Period

Connectedness did change over time (*unstandardized* β=0.05, 95% CI 0.02-0.07; *P*=.001) but did not change over the assessments per day (*unstandardized* β=−0.04, 95% CI −0.27 to 0.19; *P*=.74). In other words, participants had a small increase in feelings of connectedness over the study duration, but they did not report differences in feelings of connectedness over the course of the day.

Groups did differ in momentary connectedness (*unstandardized* β=−42.38, 95% CI −53.47 to −31.29; *P*<.001). When testing for interactions with the study phase, we found significant group differences in change across all study phases (the initial monitoring period and intervention period [*unstandardized* β=−4.82, 95% CI −7.17,−2.46; *P* <.001] intervention period and continued monitoring period [*unstandardized* β=−3.87, 95% CI −6.20 to −1.53; *P* <.001] initial monitoring period and continued monitoring period [*unstandardized* β=−8.68, 95% CI −10.77 to −6.60; *P* <.001]). In probing the simple slopes, the control group had increased connectedness during the intervention period compared to the initial monitoring period (*unstandardized* β=4.78; *P*<.001) yet decreased connectedness during the continued monitoring period compared to the intervention period (*unstandardized* β=−3.27; *P*=.003). The control group did not differ in their connectedness at the initial monitoring phase and the continued monitoring phase (*unstandardized* β=1.51; *P*=.27). In summary, the control group reported an increase in connectedness during the intervention period but a decrease in connectedness during the continued monitoring phase.

There was no difference in connectedness comparing the intervention period from the initial monitoring period in the mSavorUs group (*unstandardized* β=−0.04; *P*=.97). Furthermore, contrary to expectations, the mSavorUs group reported a decrease in connectedness during the continued monitoring period compared to the initial monitoring period (*unstandardized* β=−7.18; *P*<.001) as well as a decrease in connectedness during the continued monitoring period compared to the intervention period (*unstandardized* β=−7.14; *P*<.001). Overall, there was a decline in connectedness (see Tables S6-S9 in Multimedia Appendix 1 for full results). Mean levels of connectedness (Figure 3) for each group are shown across the 3 study phases. Shaded error bands reflect the 95% CIs based on group-wise SEs.

XSL•FO
RenderX

Figure 3. Connectedness across the study period. RS: Relational Savoring.

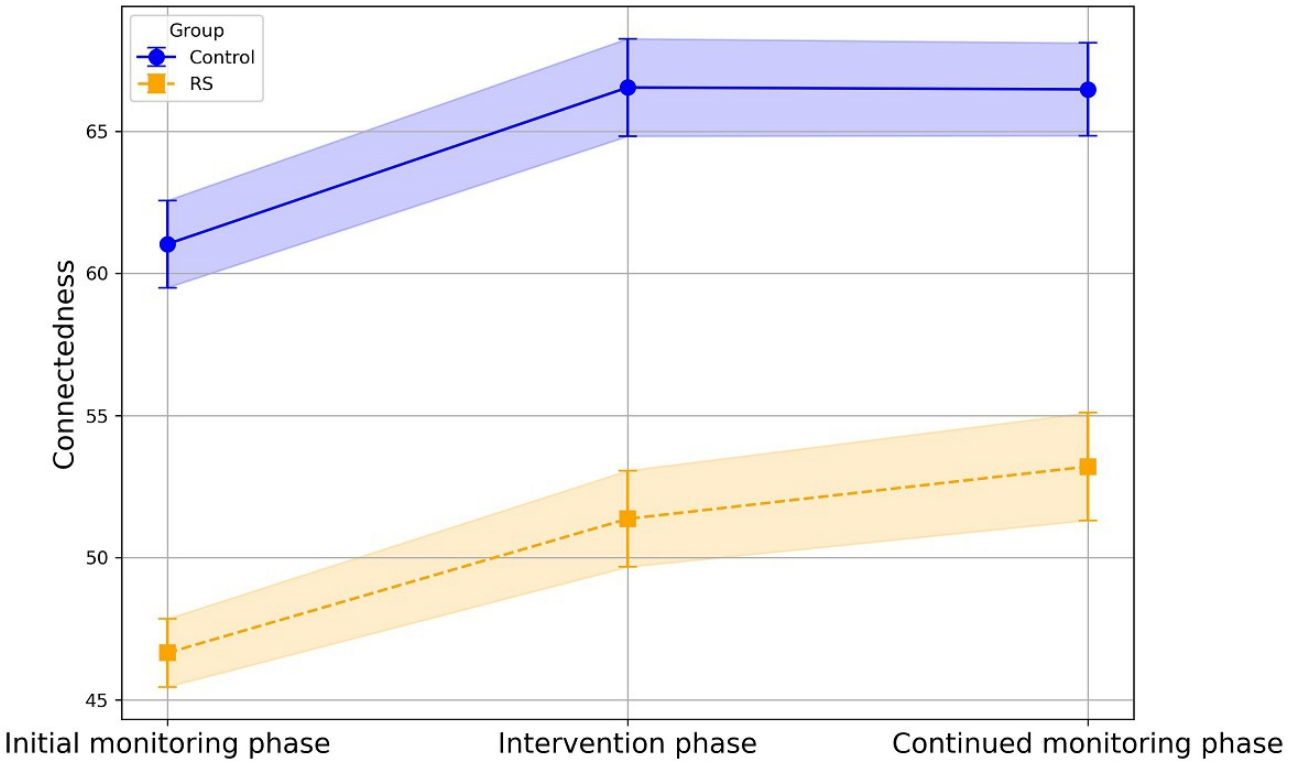

Discussion

Principal Findings

Loneliness poses significant health risks [6,7], but identifying effective interventions to reduce loneliness has remained challenging [12]. To address this issue, we evaluated the feasibility and efficacy of mSavorUs, a just-in-time tool designed to prevent loneliness and enhance social connectedness. The app leverages relational savoring—an attachment-based intervention aimed at fostering feelings of connectedness within existing relationships [14]. By integrating passive sensing tools, mSavorUs objectively identifies moments of loneliness and delivers the intervention at the most relevant times [35]. Our mixed methods study revealed an interesting pattern of results that are useful in informing future adaptations of mSavorUs. The qualitative findings revealed that while participants enjoyed the substance of the intervention, they found the JIT delivery and technological challenges detracted from the experience. The quantitative findings did not reveal significant impacts of mSavorUs on loneliness or connectedness. In the next 2 sections, we discuss the findings and their implications in greater depth, as well as the limitations of the study.

Participant Feedback and Challenges

Participants generally viewed the mSavorUs intervention positively. They reported feeling more positive and relaxed during the activity, as it encouraged focusing on the present rather than past or future stressors. However, certain aspects of the JIT format hindered effectiveness. Participants did not always complete the intervention when prompted, often due to busy schedules (eg, academic, work, and extracurricular commitments) or the short time window for response. The long duration of the intervention and the frequency of prompts also posed challenges, particularly during periods of high activity.

Previous studies have similarly noted that a high number of prompts can reduce engagement over time [46]. This finding suggests that while JIT interventions are theoretically promising, they may be less practical when integrated into busy daily lives. Allowing participants to choose when they receive prompts, while accounting for data-driven indicators of loneliness, could increase adherence. For instance, some participants might prefer completing the intervention at bedtime, while others might choose lunchtime. Enhancing user agency in scheduling could improve engagement.

Quantitative Outcomes: Loneliness and Connectedness

Despite the positive reception, we did not observe significant reductions in loneliness throughout the study. A closer examination revealed that loneliness decreased between the initial monitoring and intervention phases, but unexpectedly, the control group showed a more pronounced reduction compared to the intervention group. Notably, the control group's loneliness increased during the continued monitoring phase, while the mSavorUs group exhibited a steady decline, suggesting potential benefits from ongoing engagement with the intervention.

Regarding connectedness, significant changes emerged over time. The control group's connectedness increased between the monitoring and intervention phases but decreased afterward, while the relational savoring group showed a gradual decline across all phases. External factors, such as academic calendar changes, may have influenced these trends, as shifts in routine could affect social dynamics.

XSL•FO
RenderX

Although JIT interventions aim to address individual variability (as seen with health behaviors [47]), reducing loneliness may require broader social initiatives. A meta-analysis reported a linear increase in loneliness among emerging adults over the past 4 decades [48], suggesting that individual interventions, while helpful, may be insufficient on their own.

Limitations and Future Directions

This study focused on college students, an important population to target given their heightened risk for loneliness [8] and mental health challenges, such as depression [49], particularly during the COVID-19 pandemic. College students' familiarity with mobile devices made them a suitable group to pilot a mobile-based intervention [21,22]. However, the reliance on a college sample limits generalizability to other groups, especially those less comfortable with technology. Future research should explore the intervention among young adults not in college, as well as middle-aged and older adults who may have lower digital proficiency. In addition, the requirement for an Android device (OS 6.0 or higher) excluded users of other operating systems, such as iPhones and Google phones. Expanding compatibility would improve accessibility and inclusiveness.

Technical challenges, including connectivity issues and app incompatibility, likely contributed to participant fatigue, reducing engagement. Improved integration of app functions, flexible data connection options (Wi-Fi or mobile data), and seamless navigation are priorities for future iterations. Ensuring that the app properly opens and closes and permits consistent connectivity would prevent confusion and increase usability. Moreover, analyzing participants' preferred engagement times could help optimize prompt delivery, enhancing adherence by aligning intervention prompts with moments of likely availability.

Future studies should also use more reliable metrics to assess loneliness, such as detailed EMA data, given that small sample sizes may compromise the reliability of brief scales like the UCLA Loneliness Scale. Exploring additional analytic techniques could help capture subtle changes in loneliness.

Conclusion

The mSavorUs intervention shows promise as a mobile tool for reducing loneliness, with potential for sustained benefits even after users discontinue active engagement. Modifying the app's features, increasing device compatibility, and allowing for more user control over prompt timing could enhance feasibility and user satisfaction. Efforts to modify the intervention may wish to proceed incrementally and with plenty of opportunity for participants to provide feedback. Microtrials could be a useful method for evaluating the short-term impact of mSavorUs app use, while focus groups may provide meaningful participant feedback to inform future improvements. With these improvements, mobile JIT interventions may become a practical and effective method to alleviate loneliness among emerging adults.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

BN assisted with project administration, conceptualization. and writing-original draft. JL assisted with project administration, conceptualization, writing-original draft and writing-review & editing. HQ assisted with project administration and writing-review & editing. CM assisted with project administration and writing-review & editing. SL assisted with project administration and methodology. YW conducted formal analysis and writing – original draft. SJ assisted with project administration and methodology. ND worked on conceptualization and methodology. AR contributed to conceptualization, funding acquisition and methodology. JLB contributed to funding acquisition, conceptualization, methodology, resources, writing-original draft and writing-review & editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supporting information regarding the methodology of the study.
[DOCX File, 36 KB - formative_v9i1e70528_app1.docx]

References

1. Cigna I. Loneliness and the Workplace 2020 US Report. : Cigna; 2020 URL: https://www.cigna.com/static/www-cigna-com/docs/about-us/newsroom/studies-and-reports/combatting-loneliness/cigna-2020-loneliness-report.pdf [accessed 2024-12-31]
2. Health risks of social isolation and loneliness. Centers for Disease Control and Prevention. 2023. URL: https://www.cdc.gov/social-connectedness/risk-factors/index.html [accessed 2024-12-31]
3. Holt-Lunstad J, Smith TB, Baker M, Harris T, Stephenson D. Loneliness and social isolation as risk factors for mortality: a meta-analytic review. Perspect Psychol Sci 2015 Mar;10(2):227-237. [doi: 10.1177/1745691614568352] [Medline: 25910392]

XSL•FO
RenderX

4. Schrempft S, Jackowska M, Hamer M, Steptoe A. Associations between social isolation, loneliness, and objective physical activity in older men and women. BMC Public Health 2019 Jan 16;19(1):74. [doi: 10.1186/s12889-019-6424-y] [Medline: 30651092]
5. Heinrich LM, Gullone E. The clinical significance of loneliness: a literature review. Clin Psychol Rev 2006 Oct;26(6):695-718. [doi: 10.1016/j.cpr.2006.04.002] [Medline: 16952717]
6. Leigh-Hunt N, Bagguley D, Bash K, et al. An overview of systematic reviews on the public health consequences of social isolation and loneliness. Public Health (Fairfax) 2017 Nov;152:157-171. [doi: 10.1016/j.puhe.2017.07.035] [Medline: 28915435]
7. Holt-Lunstad J, Smith TB, Layton JB. Social relationships and mortality risk: a meta-analytic review. PLoS Med 2010 Jul 27;7(7):e1000316. [doi: 10.1371/journal.pmed.1000316] [Medline: 20668659]
8. Barreto M, Victor C, Hammond C, Eccles A, Richins MT, Qualter P. Loneliness around the world: age, gender, and cultural differences in loneliness. Pers Individ Dif 2021 Feb 1;169:110066. [doi: 10.1016/j.paid.2020.110066] [Medline: 33536694]
9. Goosby BJ, Bellatorre A, Walsemann KM, Cheadle JE. Adolescent loneliness and health in early adulthood. Sociol Inq 2013 Nov 1;83(4):505-536. [doi: 10.1111/soin.12018] [Medline: 24187387]
10. Eccles AM, Qualter P. Review: Alleviating loneliness in young people - a meta-analysis of interventions. Child Adolesc Ment Health 2021 Feb;26(1):17-33. [doi: 10.1111/camh.12389] [Medline: 32406165]
11. Cacioppo S, Grippo AJ, London S, Goossens L, Cacioppo JT. Loneliness: clinical import and interventions. Perspect Psychol Sci 2015 Mar;10(2):238-249. [doi: 10.1177/1745691615570616] [Medline: 25866548]
12. Masi CM, Chen HY, Hawkley LC, Cacioppo JT. A meta-analysis of interventions to reduce loneliness. Pers Soc Psychol Rev 2011 Aug;15(3):219-266. [doi: 10.1177/1088868310377394] [Medline: 20716644]
13. Borelli JL. Relational Savoring: Using Guided Reflection to Strengthen Relationships and Improve Outcomes in Therapy: American Psychological Association; 2024. [doi: 10.1037/0000372-000]
14. Borelli JL, Smiley PA, Kerr ML, et al. Relational savoring: an attachment-based approach to promoting interpersonal flourishing. Psychotherapy (Chic) 2020 Sep;57(3):340-351. [doi: 10.1037/pst0000284] [Medline: 31999188]
15. Burkhart ML, Borelli JL, Rasmussen HF, Sbarra DA. Cherish the good times: relational savoring in parents of infants and toddlers. Pers Relatsh 2015 Dec;22(4):692-711 [FREE Full text] [doi: 10.1111/pere.12104]
16. Borelli JL, Rasmussen HF, Burkhart ML, Sbarra DA. Relational savoring in long-distance romantic relationships. J Soc Pers Relat 2015 Dec;32(8):1083-1108. [doi: 10.1177/0265407514558960]
17. Wang BA, Bouche V, Hong K, Eriksen DE, Rice R, Borelli JL. Investigating the efficacy of relational savoring among male adolescents in residential treatment. Resid Treat Child Youth 2021 Jul 3;38(3):307-323. [doi: 10.1080/0886571X.2019.1707146]
18. Borelli JL, Bond DK, Fox S, Horn-Mallers M. Relational savoring reduces physiological reactivity and enhances psychological agency in older adults. J Appl Gerontol 2020 Mar;39(3):332-342. [doi: 10.1177/0733464819866972] [Medline: 31402745]
19. Borelli JL, Kerr ML, Smiley PA, Rasmussen HF, Hecht HK, Campos B. Relational savoring intervention: positive impacts for mothers and evidence of cultural compatibility for Latinas. Emotion 2023 Mar;23(2):303-320. [doi: 10.1037/emo0001102] [Medline: 35549365]
20. Straszewski T, Siegel JT. From writing tasks to a public service announcement: experimentally assessing savoring as a means of increasing help-seeking for depression. Soc Sci Med 2021 Oct;287:114362. [doi: 10.1016/j.socscimed.2021.114362] [Medline: 34530219]
21. Lattie EG, Stiles-Shields C, Graham AK. An overview of and recommendations for more accessible digital mental health services. Nat Rev Psychol 2022 Feb;1(2):87-100. [doi: 10.1038/s44159-021-00003-1] [Medline: 38515434]
22. Marshall JM, Dunstan DA, Bartik W. Effectiveness of using mental health mobile apps as digital antidepressants for reducing anxiety and depression: protocol for a multiple baseline across - individuals design. JMIR Res Protoc 2020 Jul 5;9(7):e17159. [doi: 10.2196/17159] [Medline: 32623368]
23. Nahum-Shani I, Smith SN, Spring BJ, et al. Just-in-Time Adaptive Interventions (JITAIs) in Mobile Health: key components and design principles for ongoing health behavior support. Ann Behav Med 2018 May 18;52(6):446-462. [doi: 10.1007/s12160-016-9830-8] [Medline: 27663578]
24. Smyth J, Heron KE. Is providing mobile interventions “just-in-time” helpful? an experimental proof of concept study of just-in-time intervention for stress management. Presented at: 2016 IEEE Wireless Health (WH); Oct 25-27, 2016; Bethesda, MD, USA. [doi: 10.1109/WH.2016.7764561]
25. Wang L, Miller LC. Just-in-the-Moment Adaptive Interventions (JITAI): a meta-analytical review. Health Commun 2020 Nov;35(12):1531-1544. [doi: 10.1080/10410236.2019.1652388] [Medline: 31488002]
26. Rideout V, Fox S, Well Being Trust. Digital health practices, social media use, and mental well-being among teens and young adults in the US. 2018 URL: https://plu.mx/plum/a/?repo_url=https://digitalcommons.providence.org/publications/1093&theme=plum-bigben-theme [accessed 2024-12-31]
27. Jafarlou S, Lai J, Azimi I, et al. Objective prediction of next-day’s affect using multimodal physiological and behavioral data: algorithm development and validation study. JMIR Form Res 2023 Mar 15;7(1):e39425. [doi: 10.2196/39425] [Medline: 36920456]

XSL•FO
RenderX

28. Lai J, Rahmani A, Yunusova A, et al. Using multimodal assessments to capture personalized contexts of college student well-being in 2020: case study. JMIR Form Res 2021 May 11;5(5):e26186. [doi: 10.2196/26186] [Medline: 33882022]
29. Rahmani AM, Lai J, Jafarlou S, et al. Personal mental health navigator: harnessing the power of data, personal models, and health cybernetics to promote psychological well-being. Front Digit Health 2022;4:933587. [doi: 10.3389/fdgth.2022.933587] [Medline: 36213523]
30. Yunusova A, Lai J, Rivera AP, et al. Assessing the mental health of emerging adults through a mental health app: protocol for a prospective pilot study. JMIR Res Protoc 2021 Mar 2;10(3):e25775. [doi: 10.2196/25775] [Medline: 33513124]
31. Hom MA, Chu C, Rogers ML, Joiner TE. A meta-analysis of the relationship between sleep problems and loneliness. Clin Psychol Sci 2020 Sep;8(5):799-824. [doi: 10.1177/2167702620922969]
32. Pels F, Kleinert J. Loneliness and physical activity: a systematic review. Int Rev Sport Exerc Psychol 2016 Jan;9(1):231-260. [doi: 10.1080/1750984X.2016.1177849]
33. Vasan S, Lim MH, Eikelis N, Lambert E. Investigating the relationship between early cardiovascular disease markers and loneliness in young adults. Sci Rep 2024 Jun 20;14(1):14221. [doi: 10.1038/s41598-024-65039-8] [Medline: 38902298]
34. Xia N, Li H. Loneliness, social isolation, and cardiovascular health. Antioxid Redox Signal 2018 Mar 20;28(9):837-851. [doi: 10.1089/ars.2017.7312] [Medline: 28903579]
35. Jafarlou S, Azimi I, Lai J, et al. Objective monitoring of loneliness levels using smart devices: a multi-device approach for mental health applications. PLoS ONE 2024;19(6):e0298949. [doi: 10.1371/journal.pone.0298949] [Medline: 38900745]
36. Hughes ME, Waite LJ, Hawkley LC, Cacioppo JT. A short scale for measuring loneliness in large surveys: results from two population-based studies. Res Aging 2004;26(6):655-672. [doi: 10.1177/0164027504268574] [Medline: 18504506]
37. Lee RM, Robbins SB. Measuring belongingness: the social connectedness and the social assurance scales. J Couns Psychol 1995;42(2):232-241. [doi: 10.1037//0022-0167.42.2.232]
38. Oura health. Oura ring. 2023. URL: https://ouraring.com/ [accessed 2024-12-31]
39. Samsung. Galaxy Watch4 Classic. 2023. URL: https://www.samsung.com/us/watches/galaxy-watch4-classic/ [accessed 2024-12-31]
40. Ferreira D, Kostakos V, Dey AK. AWARE: Mobile Context Instrumentation Framework. Front ICT 2015;2:6. [doi: 10.3389/fict.2015.00006]
41. Nishiyama Y, Ferreira D, Eigen Y, et al. IOS crowd–sensing won't hurt a bit!: AWARE framework and sustainable study guideline for IOS platform. 2020 Presented at: Distributed, Ambient and Pervasive Interactions: 8th International Conference, DAPI 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Proceedings; Jul 17-19, 2020; Copenhagen, Denmark p. 223-243. [doi: 10.1007/978-3-030-50344-4_17]
42. Nishiyama Y, Ferreira D, Sasaki W, et al. Using ios for inconspicuous data collection: a real-world assessment. 2020 Presented at: Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp'20); Sep 12-17, 2020 p. 261-266. [doi: 10.1145/3410530.3414369]
43. Braun V, Clarke V. Thematic analysis. In: Cooper H, Camic PM, Long DL, Panter AT, Rindskopf D, Sher KJ, editors. APA Handbook of Research Methods in Psychology, Vol 2: Research Designs: Quantitative, Qualitative, Neuropsychological, and Biological: American Psychological Association; 2012, Vol. 2:57-71.
44. Kleiman E. Understanding and analyzing multilevel data from real-time monitoring studies: an easily-accessible tutorial using r. PsyArXiv. Preprint posted online on Oct 10, 2017. [doi: 10.31234/osf.io/xf2pw]
45. R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing. 2023. URL: https://www.R-project.org/ [accessed 2024-12-31]
46. Morrison LG, Hargood C, Pejovic V, et al. The effect of timing and frequency of push notifications on usage of a smartphone-based stress management intervention: an exploratory trial. PLoS ONE 2017;12(1):e0169162. [doi: 10.1371/journal.pone.0169162] [Medline: 28046034]
47. Schembre SM, Liao Y, Robertson MC, et al. Just-in-time feedback in diet and physical activity interventions: systematic review and practical design framework. J Med Internet Res 2018 Mar 22;20(3):e106. [doi: 10.2196/jmir.8701] [Medline: 29567638]
48. Buecker S, Mund M, Chwastek S, Sostmann M, Luhmann M. Is loneliness in emerging adults increasing over time? A preregistered cross-temporal meta-analysis and systematic review. Psychol Bull 2021 Aug;147(8):787-805. [doi: 10.1037/bul0000332] [Medline: 34898234]
49. Wang C, Wen W, Zhang H, et al. Anxiety, depression, and stress prevalence among college students during the COVID-19 pandemic: a systematic review and meta-analysis. J Am Coll Health 2023 Oct;71(7):2123-2130. [doi: 10.1080/07448481.2021.1960849] [Medline: 34469261]

Abbreviations

EMA: ecological momentary assessment
JIT: just-in-time
TA: thematic analysis
UCLA: University of California Los Angeles

XSL•FO
RenderX

Edited by A Mavragani; submitted 23.12.24; peer-reviewed by MS Herbert, M Adnan; revised version received 27.05.25; accepted 28.05.25; published 08.09.25.

Please cite as:
Nguyen B, Lai J, Qureshi H, Marcotullio C, Labbaf S, Wang Y, Jafarlou S, Dutt N, Rahmani AM, Borelli JL
Feasibility, Acceptability, and Preliminary Outcomes of a Mobile Adaptation of a Relational Savoring Intervention to Prevent Loneliness in College Students: Mixed Methods Pilot Study
JMIR Form Res 2025;9:e70528
URL: https://formative.jmir.org/2025/1/e70528
doi:10.2196/70528

© Brenda Nguyen, Jocelyn Lai, Hana Qureshi, Christopher Marcotullio, Sina Labbaf, Yuning Wang, Salar Jafarlou, Nikil Dutt, Amir M Rahmani, Jessica L Borelli. Originally published in JMIR Formative Research (https://formative.jmir.org), 8.9.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Previsit Preparation for Shared Decision-Making in Lung Cancer Screening in Primary Care Using a Paper Decision Aid and an Automated Text Messaging Program: Quasi-Experimental Pilot Study

Mayuko Ito Fukunaga[1], MSc, MD; Renda Soylemez Wiener[2,3,4], MPH, MD; Shaun Toomey[1], MD; Joann Wagner[1], MSW; Qiming Shi[5], MS; Kavitha Balakrishnan[6], BS; Alexandra Nguyen[7,8], MD; Dan Nguyen[8], BS; M Diane McKee[9], MD; Alexander A Bankier[10], MD, PhD; Rajani S Sadasivam[6], PhD; Sybil L Crawford[11], PhD; Paul KJ Han[12,13,14], MPH, MD; Thomas K Houston[15], MPH, MD; Kathleen M Mazor[1], EdD

[1]Department of Medicine, University of Massachusetts Chan Medical School, Worcester, MA, United States
[10]Department of Radiology, University of Massachusetts Chan Medical School, Worcester, MA, United States
[11]Tan Chingfen Graduate School of Nursing, University of Massachusetts Chan Medical School, Worcester, MA, United States
[12]Center for Outcomes Research and Evaluation, Maine Medical Center, Portland, ME, United States
[13]Tufts University School of Medicine, Boston, MA, United States
[14]Division of Cancer Control and Population Sciences, National Cancer Institute, Bethesda, MD, United States
[15]Department of Internal Medicine, Wake Forest University School of Medicine, Winston-Salem, NC, United States
[2]Center for Health Optimization and Implementation Research, VA Boston Healthcare System, Boston, MA, United States
[3]National Center for Lung Cancer Screening, Veterans Health Administration, Washington, DC, United States
[4]Pulmonary Center, Boston University Chobanian & Avedisian School of Medicine, Boston, MA, United States
[5]Center for Clinical and Translational Science, University of Massachusetts Chan Medical School, Worcester, MA, United States
[6]Department of Population and Quantitative Health Sciences, University of Massachusetts Chan Medical School, Worcester, MA, United States
[7]Department of Obstetrics and Gynecology, University of Arizona, Tucson, AZ, United States
[8]University of Massachusetts Chan Medical School, Worcester, MA, United States
[9]Department of Family Medicine and Community Health, University of Massachusetts Chan Medical School, Worcester, MA, United States

Corresponding Author:
Mayuko Ito Fukunaga, MSc, MD
Department of Medicine, University of Massachusetts Chan Medical School, Worcester, MA, United States

Abstract

Background: Patient-provider discussions and shared decision-making (SDM) are essential for tailoring lung cancer screening (LCS) decisions to individual patients. However, the implementation of SDM in primary care settings is challenging. Innovative approaches are needed to reach and prepare patients eligible for LCS for SDM in primary care settings and increase LCS uptake.

Objective: We piloted previsit preparation comparing 2 strategies: a paper decision aid (DA; DA group) and an enhanced comparator strategy consisting of the paper DA plus an automated text message program (DA+TM group) designed to promote patient-provider LCS discussions. We explored feasibility and gathered preliminary data on its potential effects on LCS discussions, decision-making, and LCS uptake in primary care settings.

Methods: In a sequential quasi-experimental pilot study, we recruited patients who were eligible for LCS in a single academic health care system. Prior to an upcoming visit, participants in both groups received a paper-based DA by mail. In the DA+TM group, participants also received a series of automated text messages to help them prepare for their LCS discussions. We monitored participant recruitment and retention, as well as patient engagement in DA and text messages. In exploratory analyses, we assessed patient-provider discussion of LCS, SDM, patient knowledge, decision conflict at baseline and in follow-up telephone surveys, and LCS completion measured by electronic health records.

Results: We enrolled and included 48 participants (DA group=19 and DA+TM group=29) in the final analysis. Participants were predominantly White, with a median age of 61.0 (IQR, 57.0 - 65.0), and 58% (28/48) of them were female. Engagement was high in both groups. LCS knowledge significantly improved in the DA+TM group (4.5 baseline vs 6.0 follow-up; P=.003), while there was no change in the DA group (5.0 baseline vs 5.0 follow-up, P=.23). Median LCS knowledge change from baseline to follow-up was 0.5 (IQR –1.0 to 2.5) in the DA group and 1.5 (IQR 0 - 3.0) in the DA+TM group (P=.24). Decision conflict in both groups significantly decreased (DA group: 37.5 baseline vs 0 follow-up, P<.001; DA+TM group: 50.0 baseline vs 20.0

XSL•FO
RenderX

follow-up, *P*=.003). The median SDM process score (a measure of SDM) was 3.0 (IQR 1.5-4.0) in the DA group and 2.0 (IQR 1.0-3.0) in the DA+TM group (*P*=.11). The LCS completion rates were 5% (1/19) in the DA group and 31% (9/29) in the DA+TM group at 3 months (*P*=.07), and 26% (5/19) in the DA group and 34% (10/29) in the DA+TM group at 6 months (*P*=.75).

Conclusions: Previsit preparation was feasible in primary care settings. An enhanced, text message–based strategy has the potential to reach and engage broader LCS-eligible populations and prepare patients for LCS discussions with their primary care providers, which may ultimately improve informed decision-making and LCS uptake.

(JMIR Form Res 2025;9:e69044) doi:10.2196/69044

KEYWORDS

text messaging; decision-making, shared; lung cancer; cancer screening; telemedicine; health

Introduction

Shared decision-making (SDM) is an essential component of lung cancer screening (LCS) with annual low-dose computed tomography (LDCT). Many medical professional societies, such as the US Preventive Services Task Force, recommend and the Centers for Medicare and Medicaid Services (CMS) has mandated SDM between patients and providers using a decision aid (DA) before enrolling in LCS. SDM is a process in which the patient and health care provider collaboratively review the best available evidence, consider the patient's values and preferences, and then identify the most suitable option for the patient [1,2]. While LCS reduces lung cancer mortality, it also carries potential harms (eg, false positives, downstream invasive procedures and complications, radiation exposure, and anxiety). Furthermore, the balance of benefits and harms varies depending on an individual's risks for lung cancer and comorbidities; thus, SDM is essential.

However, high-quality SDM for LCS rarely happens during clinical visits [3-8]. LCS is still relatively new compared to other cancer screenings. Many patients are not aware of LCS, including the screening process and its potential harms. Providers often discuss the potential benefits but not the potential harms of LCS, and few providers use DAs [3-8]. Due to multiple competing priorities, providers and patients often lack sufficient time to review LCS in depth and fully engage in SDM during primary care visits. [3,7]. Despite good intentions, the requirement for SDM for LCS is considered to be a barrier to LCS. Thus, an intervention to support and streamline SDM discussions is urgently needed.

One strategy to overcome some of the posed challenges is previsit preparation to introduce patients to key LCS facts and allow time for patients to prepare for SDM. By preparing patients, this strategy may also reduce the provider's work to go over the information during the visit. Various DAs for LCS were developed, which improved patient knowledge and decreased decisional conflict in study settings [9,10]. Text messaging has been shown to increase patient knowledge and uptake of screenings for breast, cervical, and colon cancer and holds significant potential for effectively reaching and engaging the broader populations eligible for LCS. [11-13]

Thus, we developed two previsit preparation strategies of different intensities: (1) a mailed paper-based DA and (2) a mailed paper-based DA paired with a previsit automated text messaging program to support SDM for LCS. We then conducted a pilot study to test the feasibility of the 2 strategies and explore their potential effects on LCS discussions, decision-making, and LCS uptake at 3 and 6 months.

Methods

Study Design

We conducted a sequential quasi-experimental pilot study of previsit preparation using a paper-based DA (DA group) or the combination of paper-based DA and an automated text messaging program (DA+TM group). This study was conducted in accordance with CONSORT (Consolidated Standards of Reporting Trials) guidelines (Checklist 1).

Study Setting and Participants

This study was conducted at 6 primary care clinics within a large academic health care system in the Northeastern United States from November 2020 to December 2022, during the COVID-19 pandemic. This health care system has an alert in the EPIC (Epic Systems Corporation) electronic health record (EHR) health maintenance section to inform providers of their patient's potential eligibility for LCS based on age and smoking history. Primary care providers (PCPs) are encouraged to review the health maintenance section during visits.

Participants in this study were individuals eligible for LCS who never had an LCS before and who had an upcoming primary care visit. The inclusion criteria were (1) age 55-77 years, (2) cigarette smoking within the past 15 years with a minimum of 30 pack-years (ie, the 2015 CMS eligibility criteria for LCS) [14,15], (3) English-speaking, (4) owning a cell phone with text messaging capability, and (5) having an upcoming scheduled primary care visit. The exclusion criteria were (1) PCP's exclusion of the patient based on their medical conditions (eg, under hospice care and severe mental illness), (2) active cancer, (3) history of lung cancer, (4) previous LCS, and (5) chest CT within the past 12 months.

Recruitment Procedures

We identified patients potentially eligible for LCS who had an upcoming primary care appointment based on their age, histories of tobacco use, and previous chest imaging studies via EHR data. We recruited patients who had a primary care visit within 1 - 4 weeks for the DA group and within 2 - 5 weeks for the DA+TM group to ensure that the participants had enough time to receive the interventions prior to their visits and minimize the difference in the duration between the last day of the interventions (the day when the participants received DA in the

XSL•FO
RenderX

DA group and the day when the participants received the last text message in the DA+TM group) and their scheduled visit days between the 2 groups.

We mailed invitation letters to all potentially eligible patients. One week after the mailing, a research staff member made a telephone call to those who did not opt out and confirmed study eligibility.

Interventions

We delivered the paper-based DA developed by the Agency for Healthcare Research and Quality, which was publicly available during the study period. We developed an LCS text message program based on existing DAs in collaboration with patients eligible for LCS [16]. The text message program consisted of 14 messages providing key information about LCS (eg, eligibility, benefits, and harms) and a link to the DA for more detailed information. This program emphasized the importance of understanding both the benefits and harms of LCS, SDM, and smoking cessation and encouraged patients to discuss LCS with their provider at their next visit (Table 1). We delivered 1 text message per day for 14 days, always in the same order.

Table . Example of lung cancer screening text messages and learning objectives.

Learning objective of each text message	Text message
Introduction	Hi [FIRST NAME], this is the Lung Cancer Screening Message Center at xxxx. Your appointment with your primary care doctor is coming up soon. Based on your smoking history, we would like to send you text messages about lung cancer screening for the next 14 days. We would like you to learn about lung cancer screening before your next visit.
Pros and cons of lung cancer screening	Lung cancer screening has both pros and cons. It is important to learn about the pros and cons before you decide whether to have lung cancer screening.
What are cons of lung cancer screening?	What are the cons of lung cancer screening? One con is that you could get a false alarm. Lung cancer screening may find spots in the lungs that are NOT due to cancer. If this happens, you could need more testing to find out if the spots are due to cancer or NOT due to cancer.
Your value and preference	Should you be screened? Talk with your doctor to decide whether lung cancer screening is right for you. It will depend on your overall health and what is most important to you. Refer to the pamphlet that we sent to you if you would like to learn more. [a link to a decision aid].
Smoking cessation	Lung cancer screening does not replace the need for quitting smoking. As you know, quitting smoking could reduce your chance of getting lung cancer.
Closing message	We hope that you feel ready to talk about lung cancer screening with your doctor at your next visit. Think about some questions that you would like to ask. You can go to the website: [a link to a decision aid] for more information. We hope you found these messages helpful!

Study Protocol

After obtaining informed consent, the study staff administered a baseline survey over the phone. Then, the study staff mailed the paper-based DA to all participants and started a text messaging intervention for the participants in the DA+TM group. The study staff confirmed the completion of the PCP visit in the EHR and aimed to conduct a follow-up telephone survey within 2 business days of the index primary care visit. The EHR was reviewed 3 and 6 months after the index visit for the completion of LDCT for LCS and regular chest computed tomography (CT).

We chose a sequential quasi-experimental design so that we could incrementally learn the process of the primary care flow, monitor the timing of recruitment, interventions, and PCP visits, and revise our study protocol if necessary. We originally planned to enroll the first 30 participants in the DA group and the next 30 participants in the DA+TM group. Due to the COVID-19 pandemic, recruitment was slower than expected. Thus, after we enrolled the first 16 participants in the DA group, we switched our enrollment efforts from the DA group to the DA+TM group. After enrolling 30 participants in the DA+TM group, we switched back to the DA group and enrolled 3 additional participants. After enrolling a total of 49 participants, we ended recruitment since we ran out of the recruitment pool from the 6 primary care clinics.

Measures

Recruitment and Retention

We monitored the rates of recruitment and retention from study enrollment to the completion of follow-up surveys. We also monitored whether participants kept the original PCP appointments or rescheduled their appointments using the EHR.

Participant Characteristics

During recruitment calls, we collected data on participants' ages and detailed smoking history to confirm study eligibility. During the baseline survey call, we collected data on the remaining demographics, family history of lung cancer, previous technology use, and perceived lung cancer risk [17]. We measured reading health literacy with the Single-Item Literacy Screener and listening health literacy with the item developed

XSL•FO
RenderX

by the Health Literacy Cancer Prevention research team (Table S1 in Multimedia Appendix 1) [18,19].

Patient Engagement in DA and Text Messages

To evaluate intervention feasibility, we measured patient engagement with the mailed DA and text messages by self-report. We asked how much of these materials participants read and whether they clicked the link to a DA in the text messages in a follow-up survey (Table S1 in Multimedia Appendix 1).

LCS Discussion and SDM

Patient self-reported LCS discussion and SDM were measured at follow-up after the PCP visit. We asked participants whether they talked about LCS with their providers. If they had an LCS discussion, we asked participants who initiated the LCS discussion, whether their providers made LCS recommendations, and what recommendations they made (Table S1 in Multimedia Appendix 1).

The degree of SDM was measured with the shared decision-making process survey that consists of 4 items [20]. Participants were asked how much they discussed with their providers about the reason to have or not to have LCS (recorded as A lot or Some=1, A little or Not at all=0), whether their provider explained the options (to have or not to have LCS or to take more time to decide), and whether their providers asked the participant's desire to proceed with LCS (recorded as Yes=1, or No=0). The total SDM process score ranges from 0 to 4, and a higher score indicates more SDM.

LCS Knowledge

Participants' LCS knowledge, such as benefits and harms of LCS, was measured at baseline and follow-up using 9 questions adapted from a previous study (Table S2 in Multimedia Appendix 1) [10]. We created the LCS knowledge score, with the number of correct responses ranging from 0 to 9.

Decisional Conflict

Decisional conflict was measured with the 10-item Decisional Conflict Scale (DCS) in baseline and follow-up surveys [21,22]. The DCS score ranges from 0 (no conflict) to 100 (high conflict), and a score lower than 25 is associated with implementing a decision.

Stage of Decision-Making

The stage of decision-making was assessed with the "Stage of Decision Making" tool at baseline and in follow-up surveys [23]. The stage of decision-making is the individual's readiness to engage in decision-making, progress in making a choice, and willingness to consider or reconsider options. We measured the stage of decision-making because the individual may respond differently to decision support or need different decision support depending on their stage of decision-making.

LCS Uptake

LDCT order, appointment, and completion data were collected from the EHR at 3 and 6 months after the PCP visit. We also measured regular chest CT completion at 6 months after the PCP visit.

Statistical Analysis

Since this is a pragmatic trial in a real-world primary care setting, we included all participants regardless of whether they received text messages, rescheduled their appointments, or had an LCS discussion with their providers during their visits, and conducted an intention-to-treat analysis. Descriptive statistics were computed for participants' demographics, responses to surveys, and LDCT outcomes. Continuous variables were summarized as median and IQR. Categorical variables were summarized as frequencies and proportions. The proportions of smoking status, health literacy, and the baseline stage of decision-making were compared by Fisher exact test. We first compared the LCS knowledge and DCS scores between baseline and follow-up surveys in each group using the Wilcoxon rank-sum test. Then, we compared the changes in LCS knowledge and DCS scores from baseline to follow-up surveys and SDM process scores between the groups by the Wilcoxon rank-sum test, and LCS discussion and LCS uptake between the arms by Fisher exact test. In addition, we conducted sensitivity analyses of SDM, LCS knowledge, and DCS scores among participants who reported LCS discussions with their providers. All statistical analyses were conducted using Stata 18 (StataCorp LLC).

Ethical Considerations

This study was approved by the University of Massachusetts Chan Medical School institutional review board. Consent forms were mailed to all interested potential participants. A waiver of documentation of consent was obtained from the Institutional Review Board; all participants provided verbal informed consent before participation. Participants were offered a US $25 gift card for completion of each study survey (baseline and follow-up). Study data were handled in pseudonymized form using a coded key, accessible only to the research team. All personal data were stored securely on encrypted University of Massachusetts Chan Medical School servers. No identifiable images or personal information are included in the paper or any supplemental materials.

Results

Recruitment and Retention

The recruitment pool of 324 patients yielded 49 participants (19 in the DA group and 30 in the DA+TM group) with a recruitment rate of 15.4% (Figure 1). One participant in the DA+TM group could not receive text messages due to discontinuation of their text messaging plan. All participants had PCP visits. Notably, 4 (21%) participants in the DA group and 8 (27%) participants in the DA+TM group rescheduled their appointments. A total of 2 (11%) participants in the DA group had their appointment more than 4 weeks after the baseline survey. In the DA+TM group, 3 (10%) participants had their appointment within 2 weeks of the baseline survey, and 4 (13%) participants had their appointment more than 5 weeks after the baseline survey. Overall, 16/19 (84%) participants in the DA group and 28/30 (93%) participants in the DA+TM group completed the follow-up survey. Due to appointment changes and participants' availability, 2/16 participants in the DA group and 4/28 participants in the DA+TM group completed a

XSL•FO
RenderX

follow-up survey more than 7 days after the PCP visit. For LCS uptake, that is, LDCT order and completion, we excluded 1 participant in the DA + TM group who passed away within 3 months after the PCP visit and included 19 participants in the DA group and 29 participants in the DA + TM group.

XSL•FO
RenderX

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) flowchart. DA: decision aid; EHR: electronic health record; PCP: primary care provider; TM: text message.

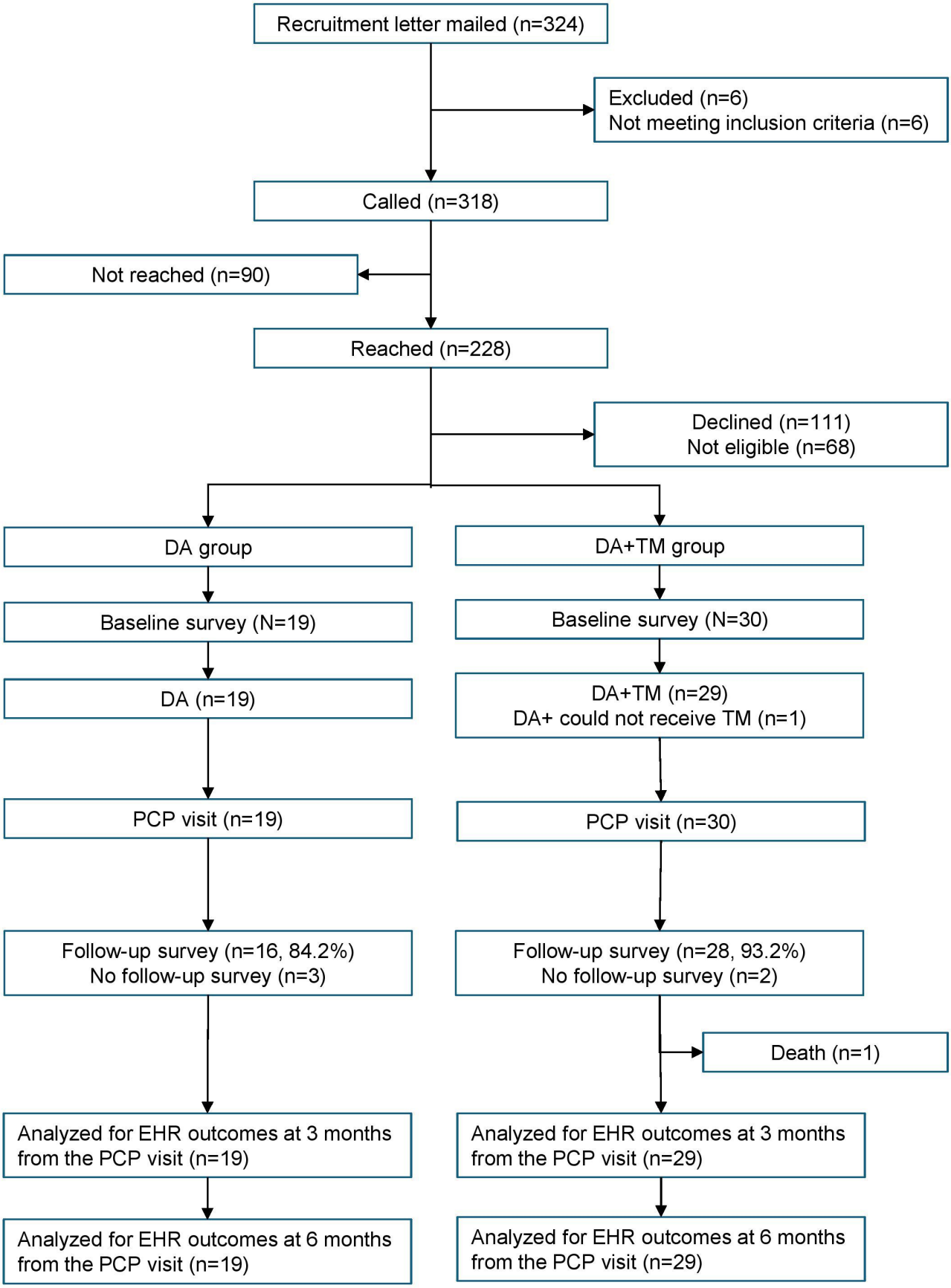

XSL•FO
RenderX

Study Population

Table 2 describes participant characteristics. Participants were predominantly White, with 58% (28/48) female and a median age of 61 (IQR 57 - 65) years. In terms of smoking status, 47% (9/19) of the participants in the DA group and 69% (20/29) of the participants in the DA+TM group reported currently smoking (*P*=.23). Regarding health literacy, 16% (3/19) of the participants in the DA group and 31% (9/29) of the participants in the DA+TM group reported limited reading health literacy (*P*=.32), and 53% (10/19) of the participants in the DA group and 76% (22/29) of the participants in the DA+TM group reported limited listening literacy (*P*=.12).

XSL•FO
RenderX

Table . Participant characteristics.

Characteristics	DA[a] group (n=19)	DA+TM[b] group (n=29)
Age in years, median (IQR)	60 (57-64)	62 (56-68)
Sex, n (%)		
Female	13 (68)	15 (52)
Male	6 (32)	14 (48)
Race, n (%)		
White	18 (95)	27 (93)
Black	0 (0)	0 (0)
Native American or Alaska Native	0 (0)	1 (3)
Other	1 (5)	1 (3)
Ethnicity, n (%)		
Hispanic or Latino	1 (5)	2 (7)
Annual income (US $), n (%)		
<30,000	8 (42)	12 (41)
30,000-59,999	6 (32)	9 (31)
60,000-99,999	3 (16)	2 (7)
>100,000	0 (0)	0 (0)
Prefer not to answer	2 (11)	6 (21)
Education level, n (%)		
High school graduate or less	6 (32)	13 (45)
Some college	8 (42)	8 (28)
Associate degree	3 (16)	1 (3)
Bachelor's degree	2 (11)	6 (21)
Advanced college degree	0 (0)	1 (3)
Limited reading health literacy, n (%)	3 (16)	9 (31)
Limited listening health literacy, n (%)	10 (53)	22 (76)
Technology use, n (%)		
Smartphone ownership	18 (95)	23 (79)
Regular email usage	14 (74)	19 (66)
Text message usage	18 (95)	27 (93)
Previous patient portal usage	14 (74)	21 (72)
Smoking status, n (%)		
Current	9 (47)	20 (69)
Smoking duration by pack-year, median (IQR)	44 (33-52.5)	52 (33.5-52)
General health, n (%)		
Excellent	3 (16)	1 (3)
Very good	1 (5)	4 (14)
Good	7 (37)	9 (31)
Fair	6 (32)	12 (41)
Poor	2 (11)	3 (10)
Mental health, n (%)		
Excellent	3 (16)	5 (17)
Very good	4 (21)	5 (17)

XSL•FO
RenderX

Characteristics	DA[a] group (n=19)	DA+TM[b] group (n=29)
Good	8 (42)	9 (31)
Fair	3 (16)	6 (21)
Poor	1 (5)	4 (14)
Family history of lung cancer, n (%)	10 (53)	13 (45)
Perceived lung cancer risk[c], n (%)		
Very low	3 (16)	5 (18)
Somewhat low	5 (26)	2 (7)
Moderate	5 (26)	10 (36)
Somewhat high	4 (21)	9 (32)
Very high	2 (11)	2 (7)

[a]DA: decision aid.
[b]TM: text message.
[c]n=1 missing data in the DA+TM group.

DA and Text Message Engagement

In the DA group, 63% (10/16) of the participants reported reading all or most of the DA while in the DA+TM group, 57% (16/28) of the participants reported reading all or most of the DA, and 21% (6/28) of the participants reported reading none of it (Table 3). In the DA+TM group, 75% (21/28) of the participants reported reading all or most of the text messages, and only 39% (11/28) of the participants reported using the link to access the DA. Among the 6 participants who did not read a mailed DA at all in the DA+TM group, none accessed DA from the links provided in the text messages. However, 5 participants read all or most of the text messages, while 1 participant read only a few of the text messages. In summary, 63% (10/16) of the participants in the DA group reported reading all or most of the mailed DA and 82% (23/28) of the participants in the DA+TX group reported reading all or most of the mailed DA or the text messages (*P*=.17).

Table . Patient self-report engagement in the intervention.

DA[a] and TM[b] engagement	DA group (n=16)	DA+TM group (n=28)
Decision-aid reading, n (%)		
All or most	10 (63)	16 (57)
Some or a little	6 (38)	4 (14)
None	0 (0)	6 (21)
Did not receive, do not remember	0 (0)	2 (7)
Text message reading, n (%)		
All or most	—[c]	21 (75)
Some or a few	—	6 (21)
None	—	0 (0)
Could not receive text messages	—	1 (4)

[a]DA: decision aid.
[b]TM: text message.
[c]Not applicable.

LCS Discussion and SDM

Most participants (94%, 15/16 in the DA group and 82%, 23/28 in the DA+TM group) reported having an LCS conversation with their PCPs during the visit (*P*=.39). Furthermore, 56% (9/16) of the participants in the DA group and 57% (16/28) of the participants in the DA+TM group stated they initiated the LCS conversation (*P*=1.00). In terms of the provider's LCS recommendation, 63% (10/16) of the participants in the DA group and 61% (17/28) of the participants in the DA+TM group reported receiving recommendations from their providers (*P*=1.00). With the exception of 1 participant in the DA+TM group who was advised not to undergo LCS, 26/27 participants in the 2 groups were recommended for LCS.

The median of the SDM process score measuring degree of SDM (with 0 being low and 4 being high) was 3.0 (IQR 1.5 - 4.0) in the DA group (n=16) and 2.0 (IQR 1.0 - 3.0) in the DA+TM group (n=26 due to 2 missing data; *P*=.11).

XSL•FO
RenderX

LCS Knowledge

The median patient LCS knowledge score in the DA group was 5 of 9 (IQR 4 - 6) at baseline and 5 of 9 (IQR, 4.5 - 7) at follow-up, without significant change (*P*=.23). On the other hand, the median patient LCS knowledge in the DA+TM group significantly improved from 4.5 (IQR 3 - 6) at baseline to 6 (IQR 5 - 6) at follow-up (*P*=.003; Figure 2). However, there was no statistically significant difference in the change of LCS knowledge score from baseline to follow-up between the 2 groups (DA: 0.5, IQR –1 to 2.5) and DA+TM: 1.5, IQR 0 - 3; *P*=.24). Details about performance by group on specific LCS knowledge questions are provided in Table S2 in Multimedia Appendix 1.

Figure 2. Lung cancer screening knowledge at the baseline and follow-up surveys. Boxes indicate 25th to 75th percentile. Dots indicate an outlier. DA: decision aid; LCS: lung cancer screening; TM: text messages. Lines indicate a median.

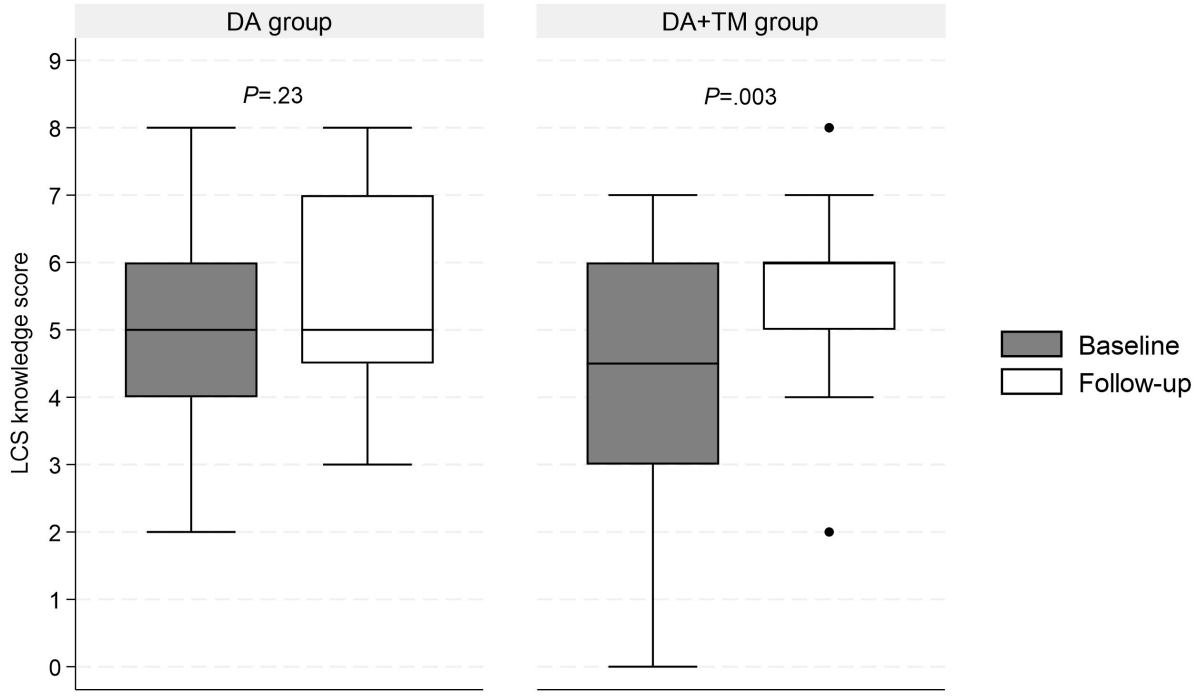

Decisional Conflict

The total median DCS decreased significantly from 40 (IQR 20.0 - 70.0) at baseline to 0 (IQR 0 - 15.0) at follow-up in the DA group (n=15 with 1 missing data; *P*<.001) and from 50.0 (IQR 32.5 - 65.0) to 20 (IQR 0 - 42.5) in the DA+TM group (*P*<.001; Figure 3). The median change of the total DCS was 25.0 (IQR 15.0 - 50.0) in the DA group and 20.0 (IQR 10.0 - 47.5) in the DA+TM group (*P*=.50). A DCS of less than 25 means the participants were ready to make a decision. At baseline, 31.3% (5/16) of the participants in the DA group and 14.3% (4/18) of the participants in the DA+TM group had a total DCS less than 25 (*P*=.25). At follow-up, 80% (12/15, n=1 missing data) of the participants in the DA group and 50% (14/28) of the participants in the DA+TM group had the total DCS less than 25 (*P*=.10). Figures S1 and Table S3 in Multimedia Appendix 1 describe the Informed, Value, Support, and Uncertainty subscales.

XSL•FO
RenderX

Figure 3. Decisional conflict scale at the baseline and follow-up surveys. Lines indicate a median. Boxes indicate 25th to 75th percentile. Dots indicate an outlier. DA: decision aid, DCS: decisional conflict scale; TM: text messages.

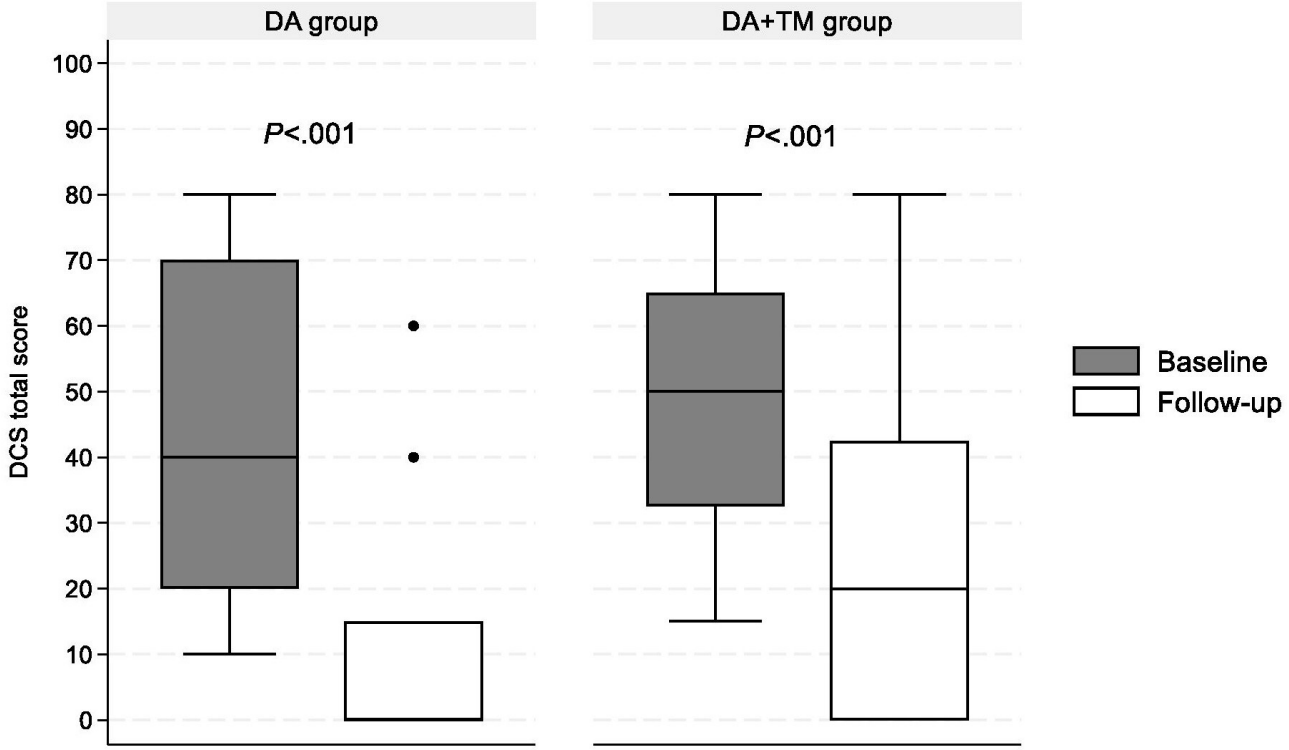

Stage of Decision-Making

At baseline, the participants were at diverse stages of decision-making about LCS. Among the participants in the DA group, 19% (3/16) of the participants had not begun to think about the choices, 50% (8/16) of the participants were considering the options, and 25% (4/16) of the participants had already made a decision and were unlikely to change their minds (Figure 4). Among the participants in the DA+TM group, 43% (12/28) of the participants had not begun to think about the choices, 25% (7/28) of the participants were considering the options, and 11% (3/28) of the participants had already made a decision and were unlikely to change their minds. There was no statistically significant difference in the baseline stages of decision-making between the 2 groups (P=.07). The stage of decision-making shifted toward decision-making at follow-up in both groups, and 79% (11/14, 2 missing data) of the participants in the DM group and 64% (18/28) of the participants in the DA+TM group stated that they had already made a decision and were unlikely to change their mind.

XSL•FO
RenderX

Figure 4. Stage of decision making at the baseline and follow-up surveys. DA: decision aid, TM: text messages.

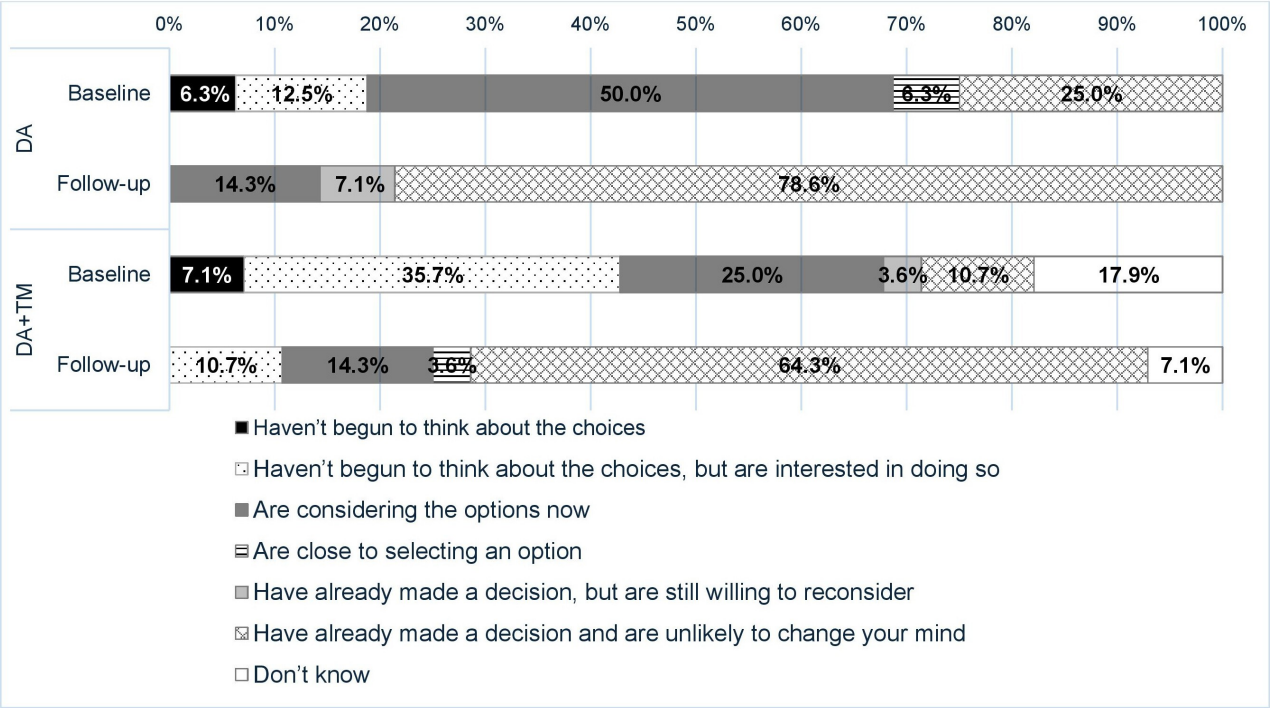

LCS Uptake

Table 4 describes LDCT ordering, scheduling, and completion rates at 3 and 6 months after the index PCP visit. Within 3 months after the index PCP visit, 26% (5/19) of the participants in the DA group and 48% (14/29) of the participants in the DA+TM group had LDCT orders (*P*=.15), and 5% (1/19) of the participants in the DA group and 31% (9/29) of the participants in the DA+TM group completed LDCT for LCS (*P*=.07). Within 6 months after the index PCP visit, 26% (5/19) of the participants in the DA group and 34% (10/29) of the participants in the DA+TM group completed LDCT for LCS (*P*=.75). Also, 2 participants in each group had chest CT for other indications during the study period.

Table . Lung cancer screening uptake.

LDCT[a] for LCS[b]	DA[c] group (n=19)	DA+TM[d] group (n=29)	*P* value
3 months after PCP[e] visit, n (%)			
Order	5 (26)	14 (48)	.15
Appointment	3 (16)	12 (41)	.11
Completion	1 (5)	9 (31)	.07
6 months after PCP visit, n (%)			
Order	7 (37)	14 (48)	.56
Appointment	7 (37)	13 (45)	.77
Completion	5 (26)	10 (34)	.75
Chest CT[f], n (%)			
6 months from PCP visit	2 (11)	2 (7)	1.00

[a]LDCT: low-dose computed tomography.
[b]LCS: lung cancer screening.
[c]DA: decision aid.
[d]TM: text message.
[e]PCP: primary care provider.
[f]CT: computed tomography.

XSL•FO
RenderX

Sensitivity Analysis of the Participants Who Reported LCS Discussion

Tables S4-S6 and Figures S2 and S3 in Multimedia Appendix 1 describe the results of the sensitivity analysis of SDM, LCS knowledge, decision conflict, and LCS uptake among participants who reported LCS discussion during their PCP visit. Our findings of SDM, LCS knowledge, decisional conflict, and LCS uptake were similar between our sensitivity and primary analyses among all participants.

Discussion

Principal Findings

We successfully conducted a sequential quasi-experimental pilot study to evaluate the feasibility and to explore potential outcomes of the previsit preparation strategies for SDM in LCS in primary care settings. Our study protocol appeared to be feasible, given that our recruitment rate was similar to that of other studies evaluating the effectiveness of a DA for LCS in primary care settings [24,25] and the retention rate was high. We found that previsit preparation using a mailed DA with or without automated text messaging was feasible. Most participants in both groups reported discussing LCS with their provider. Patient decision conflict was significantly reduced in both groups from baseline to follow-up. We observed significant improvement in LCS knowledge among the participants in the DA+TM group, but not among those in the DA group. These findings provide promising preliminary signals on the potential of the enhanced previsit preparation using text messaging to broadly reach patients eligible for LCS and to improve informed decision-making and uptake of LCS. However, the degree of SDM remained limited in both groups.

LCS discussion rates in our study exceeded 80% in both groups, surpassing the 60% rates reported in a previous study about SDM in LCS [25]. Furthermore, approximately 60% of the participants in our study reported initiating the LCS conversation. Although the exact reason for the high LCS discussion rates is unclear, receiving the information in advance rather than immediately before their PCP visit might have provided more time for patients to think about and prepare for LCS discussions with their providers.

In our pilot study, the 6 participants who did not read DAs reported reading text messages. This finding supports the high acceptability of text messages observed during the development of the LCS text messages and aligns with prior studies showing similar acceptability for text messages to support SDM for colon cancer screening among individuals diverse in income, race, and health literacy [26]. Individuals with low income and limited health literacy often need assistance to review DAs [10]. Thus, text messages have the potential to improve the reach and engagement of broader LCS-eligible populations in SDM.

In the DA+TM group, patient LCS knowledge significantly improved at follow-up, despite 31% reporting limited health literacy. Presenting new information in small doses in a sequential manner may have decreased cognitive load for learners, especially for those with limited health literacy [27,28]. However, the participants in both groups still had a deficit of detailed knowledge on the potential harms of LCS and the limited degree of SDM, which may reflect more on providers' practice than patients' factors. Improving patient understanding and SDM may require a provider-level intervention in addition to this patient-facing previsit intervention.

Comparison With Prior Work

Despite the CMS mandate, how best to reach and support patients eligible for LCS in SDM in primary care settings is unknown. Various DAs for LCS have been shown to improve patient knowledge and decrease decisional conflict in study settings [9]. Web-based DAs increased the uptake of LCS in primary care settings but did not improve the quality of SDM in previous randomized controlled trials (RCTs) [24,25]. Still, adoption of DAs in real-world clinical settings remains challenging [29].

Our pilot study differed from previous studies in 2 main ways [24,25]. First, we used an automated text messaging system as a channel to reach and deliver information to patients. Second, we delivered a DA and text messaging program to the participants outside of clinical and study visits, whereas previous RCTs delivered DAs during the study visit right before the primary care visit [24,25]. This approach is potentially more feasible in real-world clinical settings with limited resources.

Limitations

This study has some limitations. First, our participants were predominantly White and from a single health care system, limiting generalizability. However, our participants were diverse in socioeconomic status. Second, the sample size of this pilot study was small and there was an imbalance between the 2 groups. Third, there may have been differences in participants' characteristics and stages of decision-making between the 2 groups that could have affected the outcomes. In addition, the temporal effect of the COVID-19 pandemic and unmeasured providers' characteristics and SDM practices may have affected the outcomes due to the sequential design of the study. Thus, comparability and statistical power for between-group analyses of participants' characteristics and outcomes are limited. A RCT with a larger sample size powered to detect clinically meaningful differences between groups is needed for further evaluation. Still, our pilot study demonstrated interesting preliminary findings to explore how to help patients prepare for LCS discussions prior to their primary care visits.

Conclusions

An enhanced previsit preparation strategy using an automated text messaging program is feasible in primary care settings and potentially effective in reaching and engaging the broader LCS-eligible population in SDM for LCS. Our pilot study suggests that a previsit automated text messaging program combined with a mailed DA may improve basic LCS knowledge without exacerbating disparities related to the digital divide or health literacy level. Previsit preparation enhanced by a text messaging program may facilitate SDM discussions about LCS and ultimately increase LCS uptake.

XSL•FO
RenderX

Acknowledgments

This study was supported by the National Heart, Lung, and Blood Institute (grant 1K12HL138049-01) and the National Center for Advancing Translational Science (grant UL1TR001453). The authors thank Dr Kristina Crother for sharing the knowledge questions. Grammarly was used to correct grammatical errors. MIF's effort is supported by the National Heart, Lung, and Blood Institute (grant 1K12HL138049-01) and the National Cancer Institute of the National Institutes of Health (award K08CA283304). RSW is supported in part by resources from the VA Boston Healthcare System.

The views expressed in this article are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs or the US government. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Data Availability

The datasets for this manuscript are available from the corresponding author upon reasonable request and institutional review board approval.

Conflicts of Interest

AAB is a consultant to Daiichi Pharmaceuticals, AstraZeneca, and AbbVie.

Multimedia Appendix 1
Additional materials.
[DOCX File, 106 KB - formative_v9i1e69044_app1.docx]

Checklist 1
CONSORT checklist.
[PDF File, 223 KB - formative_v9i1e69044_app2.pdf]

References

1. Légaré F, Witteman HO. Shared decision making: examining key elements and barriers to adoption into routine clinical practice. Health Aff (Millwood) 2013 Feb;32(2):276-284. [doi: 10.1377/hlthaff.2012.1078] [Medline: 23381520]
2. Elwyn G, Laitner S, Coulter A, Walker E, Watson P, Thomson R. Implementing shared decision making in the NHS. BMJ 2010 Oct 14;341:c5146. [doi: 10.1136/bmj.c5146] [Medline: 20947577]
3. Wiener RS, Koppelman E, Bolton R, et al. Patient and clinician perspectives on shared decision-making in early adopting lung cancer screening programs: a qualitative study. J Gen Intern Med 2018 Jul;33(7):1035-1042. [doi: 10.1007/s11606-018-4350-9] [Medline: 29468601]
4. Brenner AT, Malo TL, Margolis M, et al. Evaluating shared decision making for lung cancer screening. JAMA Intern Med 2018 Oct 1;178(10):1311-1316. [doi: 10.1001/jamainternmed.2018.3054] [Medline: 30105393]
5. Nishi SPE, Lowenstein LM, Mendoza TR, et al. Shared decision-making for lung cancer screening: how well are we "sharing"? Chest 2021 Jul;160(1):330-340. [doi: 10.1016/j.chest.2021.01.041] [Medline: 33556362]
6. Golden SE, Ono SS, Thakurta SG, et al. "I'm Putting My Trust in Their Hands": a qualitative study of patients' views on clinician initial communication about lung cancer screening. Chest 2020 Sep;158(3):1260-1267. [doi: 10.1016/j.chest.2020.02.072] [Medline: 32278782]
7. Martinez MC, Stults CD, Li J. Provider and patient perspectives to improve lung cancer screening with low-dose computed tomography 5 years after Medicare coverage: a qualitative study. BMC Prim Care 2022 Dec 20;23(1):332. [doi: 10.1186/s12875-022-01925-2] [Medline: 36539693]
8. Hill PA. Current state of shared decision-making for CT lung cancer screening and improvement strategies. J Patient Exp 2020 Feb;7(1):49-52. [doi: 10.1177/2374373518817340] [Medline: 32128371]
9. Fukunaga MI, Halligan K, Kodela J, et al. Tools to promote shared decision-making in lung cancer screening using low-dose CT scanning: a systematic review. Chest 2020 Dec;158(6):2646-2657. [doi: 10.1016/j.chest.2020.05.610] [Medline: 32629037]
10. Crothers K, Kross EK, Reisch LM, et al. Patients' attitudes regarding lung cancer screening and decision aids: a survey and focus group study. Ann Am Thorac Soc 2016 Nov;13(11):1992-2001. [doi: 10.1513/AnnalsATS.201604-289OC] [Medline: 27652509]
11. Uy C, Lopez J, Trinh-Shevrin C, Kwon SC, Sherman SE, Liang PS. Text messaging interventions on cancer screening rates: a systematic review. J Med Internet Res 2017 Aug 24;19(8):e296. [doi: 10.2196/jmir.7893] [Medline: 28838885]

XSL•FO
RenderX

12. Ruco A, Dossa F, Tinmouth J, et al. Social media and mHealth technology for cancer screening: systematic review and meta-analysis. J Med Internet Res 2021 Jul 30;23(7):e26759. [doi: 10.2196/26759] [Medline: 34328423]
13. Muller CJ, Robinson RF, Smith JJ, et al. Text message reminders increased colorectal cancer screening in a randomized trial with Alaska Native and American Indian people. Cancer 2017 Apr 15;123(8):1382-1389. [doi: 10.1002/cncr.30499] [Medline: 28001304]
14. Screening for lung cancer with low dose computed tomography (LDCT). CMS.gov. URL: https://www.cms.gov/medicare-coverage-database/details/nca-decision-memo.aspx?NCAId=274 [accessed 2023-12-20]
15. Lung cancer screening with low dose computed tomography (LDCT). CMS.gov. 2022. URL: https://www.cms.gov/medicare-coverage-database/view/ncd.aspx?NCDId=364 [accessed 2023-12-20]
16. Ito Fukunaga M. Empowering patients for shared decision making in lung cancer screening via text messages. : University of Massachusetts Chan Medical School; 2020 URL: https://hdl.handle.net/20.500.14038/31341
17. Health information national trends survey 2003 (HINTS 2003). National Cancer Institute. 2003. URL: https://hints.cancer.gov/docs/Instruments/HINTS_2003_Instrument-English.pdf [accessed 2025-07-29]
18. Morris NS, MacLean CD, Chew LD, Littenberg B. The Single Item Literacy Screener: evaluation of a brief instrument to identify limited reading ability. BMC Fam Pract 2006 Mar 24;7:21. [doi: 10.1186/1471-2296-7-21] [Medline: 16563164]
19. Mazor KM, Roblin DW, Williams AE, et al. Health literacy and cancer prevention: two new instruments to assess comprehension. Patient Educ Couns 2012 Jul;88(1):54-60. [doi: 10.1016/j.pec.2011.12.009] [Medline: 22244323]
20. Valentine KD, Vo H, Fowler FJ, Brodney S, Barry MJ, Sepucha KR. Development and evaluation of the shared decision making process scale: a short patient‐reported measure. Med Decis Making 2021 Feb;41(2):108-119. [doi: 10.1177/0272989X20977878] [Medline: 33319648]
21. O'Connor AM. Validation of a decisional conflict scale. Med Decis Making 1995;15(1):25-30. [doi: 10.1177/0272989X9501500105] [Medline: 7898294]
22. O'Connor AM. User manual – decisional conflict scale (10-item question format). Ottawa Hospital Research Institute. 1993. URL: https://decisionaid.ohri.ca/docs/develop/User_Manuals/UM_decisional_conflict.pdf [accessed 2024-11-20]
23. O'Connor AM. User mannual - stage of decsion making. Ottawa Hospital Research Institute. 2000. URL: https://decisionaid.ohri.ca/docs/develop/User_Manuals/UM_Stage_Decision_Making.pdf [accessed 2024-11-20]
24. Schapira MM, Hubbard RA, Whittle J, et al. Lung cancer screening decision aid designed for a primary care setting: a randomized clinical trial. JAMA Netw Open 2023 Aug 1;6(8):e2330452. [doi: 10.1001/jamanetworkopen.2023.30452] [Medline: 37647070]
25. Walsh JME, Karliner L, Smith A, et al. LungCARE: encouraging shared decision-making in lung cancer screening. J Gen Intern Med 2023 Nov;38(14):3115-3122. [doi: 10.1007/s11606-023-08189-1] [Medline: 37653203]
26. Hwang S, Lazard AJ, Reffner Collins MK, et al. Exploring the acceptability of text messages to inform and support shared decision-making for colorectal cancer screening: online panel survey. JMIR Cancer 2023 May 5;9:e40917. [doi: 10.2196/40917] [Medline: 37145859]
27. Clark RC, Nguyen F, Sweller J. Efficiency in Learning: Evidence-Based Guidelines to Manage Cognitive Load: John Wiley & Sons; 2011.
28. Baker DW, DeWalt DA, Schillinger D, et al. "Teach to goal": theory and design principles of an intervention to improve heart failure self-management skills of patients with low health literacy. J Health Commun 2011;16 Suppl 3(Suppl 3):73-88. [doi: 10.1080/10810730.2011.604379] [Medline: 21951244]
29. Kukhareva PV, Li H, Caverly TJ. Implementation of lung cancer screening in primary care and pulmonary clinics: pragmatic clinical trial of electronic health records. Chest 2023 Nov;164(5):1325-1338. [doi: 10.1016/j.chest.2023.04.040] [Medline: 37142092]

Abbreviations

CMS: Centers for Medicare & Medicaid Services
CONSORT: Consolidated Standards of Reporting Trials
CT: computed tomography
DA: decision aid
DCS: Decisional Conflict Scale
EHR: electronic health record
LCS: lung cancer screening
LDCT: low‐dose computed tomography
PCP: primary care provider
RCT: randomized controlled trial
SDM: shared decision-making

XSL•FO
RenderX

Edited by A Mavragani; submitted 14.05.25; peer-reviewed by F Yeh, M Gasmi; revised version received 03.07.25; accepted 05.07.25; published 17.09.25.

Please cite as:
Ito Fukunaga M, Wiener RS, Toomey S, Wagner J, Shi Q, Balakrishnan K, Nguyen A, Nguyen D, McKee MD, Bankier AA, Sadasivam RS, Crawford SL, Han PKJ, Houston TK, Mazor KM
Previsit Preparation for Shared Decision-Making in Lung Cancer Screening in Primary Care Using a Paper Decision Aid and an Automated Text Messaging Program: Quasi-Experimental Pilot Study
JMIR Form Res 2025;9:e69044
URL: https://formative.jmir.org/2025/1/e69044
doi:10.2196/69044

© Mayuko Ito Fukunaga, Renda Soylemez Wiener, Shaun Toomey, Joann Wagner, Qiming Shi, Kavitha Balakrishnan, Alexandra Nguyen, Dan Nguyen, M Diane McKee, Alexander A Bankier, Rajani S Sadasivam, Sybil L Crawford, Paul KJ Han, Thomas K Houston, Kathleen M Mazor. Originally published in JMIR Formative Research (https://formative.jmir.org), 17.9.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Implementation, Challenges, and Outlook of an Intergenerational, Layperson-led, Health Coaching Program (HealthStart): A Pilot Case Study

Xiaoting Huang[1,2*], MBBS, MMed; Ka Shing Yow[1,3*], MBBS, MMed; Audrey Shu Ting Kwan[1,4,5], BSc, MPH; Jin Ye Yeo[1], BSc; Haikel A Lim[1,6,7], MD, MSc; Jie Xin Lim[1], PhD; Meng Han Lim[1,8], BSocSci; Lynn Pei Zhen Teo[1,4], BSc; Nerice Heng Wen Ngiam[1,9]; Si Qi Lim[1,5], M.Ger; Kharuna Jaichandra[4], MA; Kai Wen Aaron Tang[3], MBBS; Angeline Jie-Yin Tey[1,10], MBBS, MMed; Lian Leng Low[4,11*], MBBS, MMed; Kennedy Yao Yi Ng[1,4,7,12,13*], MBBS, MMed

[1]TriGen, Singapore, Singapore
[10]Department of Respiratory and Critical Care Medicine, Tan Tock Seng Hospital, Singapore, Singapore
[11]Research and Translational Innovation, SingHealth Community Hospitals, Singapore, Singapore
[12]Division of Medical Oncology, National Cancer Centre Singapore, Singapore, Singapore
[13]Department of Global Health and Population, Harvard T.H. Chan School of Public Health, Boston, MA, United States
[2]Department of Geriatric Medicine, Singapore General Hospital, Singapore, Singapore
[3]Department of Internal Medicine, National University Health System, Singapore, Singapore
[4]Division of Population Health and Integrated Care, Singapore General Hospital, Outram Road, Singapore, Singapore
[5]Department of Physiotherapy, Singapore General Hospital, Singapore, Singapore
[6]Department of Psychiatry, National Healthcare Group, Singapore, Singapore
[7]Duke-NUS Medical School, Singapore, Singapore
[8]Primary and Community Care Division, Ministry of Health, Singapore, Singapore
[9]Department of Internal Medicine, Singapore General Hospital, Singapore, Singapore
[*]these authors contributed equally

Corresponding Author:
Kennedy Yao Yi Ng, MBBS, MMed
TriGen, Singapore, Singapore

Abstract

Background: As rapidly aging populations become a worldwide phenomenon, early detection and prompt management of chronic disease become essential to support healthy aging. Community-based health screenings, a key component of this strategy, often struggle with poor follow-up rates, limiting their long-term impact. Given the untapped potential of youth volunteers and the urgent need for a scalable approach to improve continuity of care post health screenings, we developed HealthStart: a structured, theory-based program that empowers these older adults to take greater ownership of their health and their chronic conditions with the support of youth community health volunteers (CHVs).

Objective: This study aimed to describe the development, implementation, and early outcomes of HealthStart—an intergenerational, layperson-led health coaching program—and summarize operational lessons to guide similar models in Asian communities.

Methods: HealthStart adopted an intergenerational service-learning approach modeled on a self-determination theory–based layperson-led health coaching framework. Each HealthStart team consisted of 1 health care volunteer (HCV) and 4 youth CHVs. All volunteers underwent blended training and were assessed for layperson-led health coaching readiness. Between September 2022 and June 2023 in Singapore, youth CHVs empowered adult participants aged 40 years and older after their health screening to (1) learn about their chronic diseases, (2) learn at least one digital health app, (3) enroll with a primary care provider, and (4) set a lifestyle goal (based on the Specific, Measurable, Achievable, Realistic/Relevant, and Time-bound [SMART] framework for goal setting) and achieve it. We used an implementation-focused case study design using descriptive statistics and volunteer-participant feedback to evaluate feasibility and outcomes.

Results: Of 236 eligible individuals, 192 enrolled. Participants had a mean age of 67 (SD 9.6) years; 52.1% (n=100) of participants were female, with a majority of Chinese ethnicity, having completed primary or secondary school education, residing in self-owned flats, and living in 3-room public housing. Follow-up rate with primary care increased from 42.7% (82/148) preprogram to 84.5%

XSL•FO
RenderX

(125/148) postprogram (χ^2_1=43; P<.001). In total, 58 HCVs were recruited, comprising 26 nurses and 6 doctors, with the remainder as allied health professionals. A total of 33 were trained and deployed. The mean age of HCVs was 37 years old, and 24 (72.7%) were female. Furthermore, 149 youth CHVs were recruited, 138 trained, and 102 deployed. The mean age of the youth CHVs who were deployed was 24 years, and 75 (73.5%) were female. Reflections included the importance of volunteer competency and selection criteria, tiering of participant intervention, tapping on community assets, adoption of a social prescription framework, importance of alignment with population health policies, and cultivating intergenerational relationships.

Conclusions: HealthStart demonstrates the feasibility and acceptability of a structured, intergenerational, layperson-led health coaching model embedded in primary care. We identify key lessons learned in the conceptualization and implementation of the program that may inform the design of similar volunteer-enabled initiatives for harnessing laypersons, an often-underused asset, to promote health in the community.

(JMIR Form Res 2025;9:e76592) doi:10.2196/76592

KEYWORDS

healthy aging; health coaching; aging in place; intergenerational; chronic disease management; youth community health volunteers

Introduction

Population aging is accelerating worldwide: by 2030, the number of people above the age of 60 will reach an estimated 1.4 billion [1]. As most years lived with chronic disease accrue after mid-life, timely management of hypertension, diabetes mellitus, and dyslipidemia is pivotal to healthy aging [2]. Effective follow-up after screening enables medication titration, lifestyle counseling, and complication surveillance—cornerstones of chronic-disease control [3].

Singapore, a fast-aging, multi-ethnic, cosmopolitan city-state in Southeast Asia, is no exception. Within the next 5 years, 1 in 4 residents is projected to be 65 years and older, with a corresponding rise in multimorbidity [4]. Community health screenings conducted to identify chronic diseases early report high dropout rates and a lack of follow-up after diagnosis: 1 in 4 people who were screened had not returned for a doctor's follow-up a year later for their newly diagnosed chronic conditions [5,6]. In response to these challenges, the Singapore government introduced the Healthier SG (HSG) initiative in 2024, which focuses on 5 key features: mobilizing a network of family doctors for preventive care, developing health plans, activating community partners, launching a national enrollment program, and using IT enablers [7]. While welcomed, HSG's success remains contingent on sustained patient activation, digital literacy, and care navigation gaps that disproportionately affect older adults.

Youth community health volunteers (youth CHVs), as digital natives, are well-placed to support older adults' use of digital platforms, which play an increasingly important role in health promotion and access to health care [8-10]. In addition, intergenerational service-learning projects in Singapore using youth CHVs such as TriGenerational HomeCare (TriGen) demonstrated reductions in hospital use among older adults [11]. A recent meta-analysis demonstrates that peer coaching by laypersons improved glycemic control and patient activation in chronic disease management [12]. Furthermore, training youth CHVs facilitates the intergenerational transmission of health promotion knowledge and encourages health behaviors during a formative life stage, benefiting both volunteers and their communities [13].

In response to poor postscreening follow-up and the benefits of using lay CHVs, we developed the HealthStart, an intergenerational, youth-led, layperson health-coaching program that provides longitudinal follow-up support to participants after screening. HealthStart mobilizes youth CHVs to help adults aged 40 years and older in achieving the following goals: (1) increase follow-up with a primary care physician, (2) increase ownership of one's health, (3) increase knowledge of their chronic conditions, (4) increase digital health literacy, and (5) set specific health goals related to their condition and follow through on them. By fostering health literacy, digital engagement, and continuity of care, HealthStart aims to enhance chronic disease follow-up while building intergenerational relationships and bolstering community capacity.

HealthStart mobilizes community assets, including grassroots organizations and trained youth CHV, as part of a scalable, layperson-led model to reinforce continuity of care. In this way, HealthStart functions as a community-based program that extends Healthier SG's population health objectives into the neighborhoods, particularly among hard-to-reach older adults identified through screening programs.

In this paper, we present the HealthStart program's conceptual foundations, describe its implementation, and reflect on key operational and training lessons. We also share the experiences of youth CHVs and health care volunteers (HCVs) involved, with the aim of informing similar interventions targeting community-dwelling older adults.

Methods

Overview

This study used an implementation-focused case study design to examine the feasibility, operational processes, and outcomes of the HealthStart program over 2 cycles between September 2022 and June 2023. Quantitative and qualitative data were collected to describe participant outcomes and volunteer deployment. The study followed the StaRI (Standards for Reporting Implementation Studies) checklist and the TIDieR (Template for Intervention Description and Replication) extension for complex interventions [14,15].

XSL•FO
RenderX

HealthStart is an initiative by TriGen Ltd [16], a nonprofit organization in Singapore, in collaboration with the Singapore General Hospital Division of Population Health and Integrated Care. The Singapore General Hospital Division of Population Health and Integrated Care is responsible for the population health needs of the approximately 300,000 residents living in the Southeast region [17] and conducts health screening in the community as part of the government's "Screen for Life" initiative [18]. Individuals aged 40 years and older are invited to attend health screenings conducted in community spaces such as community centers or active aging centers. After undergoing the health screenings, those with ≥1 newly detected or suboptimally controlled chronic condition were invited to participate in the HealthStart program.

Conceptual Framework

HealthStart is a layperson-led health coaching program grounded in self-determination theory (SDT), incorporating the SMART (Specific, Measurable, Achievable, Realistic/Relevant, and Time-bound) framework for goal setting, and delivered through an intergenerational service-learning model. SDT postulates that there are 3 innate psychological needs—competence, autonomy, and relatedness—which, when satisfied, yield enhanced self-motivation and, when thwarted, lead to diminished motivation and well-being [19]. Goal setting via the SMART framework has been theorized to facilitate behavior change by assisting individuals in defining clear and actionable goals [20]. HealthStart aims to increase the engagement and follow-up rates of participants attending a chronic disease community health screening by developing, training, and empowering CHVs to serve as health coaches. Over a 3-month period, CHVs support participants in their journey to improve their health. The program is designed to promote participants' autonomy by allowing them to identify goals that are important to them and develop personalized action plans. It enhances competency through SMART goal setting and motivational interviewing, which is taught to all CHVs [21]. Finally, the program's longitudinal and relational nature emphasizes building connections and fulfilling participants' need for relatedness.

Volunteer training in HealthStart is anchored upon an intergenerational and service-learning approach. Intergenerational theory focuses equal attention on the potential development of all members of each intergenerational dyad to learn as a function of social interaction [22]. Contact theory, a widely cited theory guiding intergenerational program development and study, describes 4 main tenets that increase the likelihood that people will experience positive contact: support of governing bodies, equal group status, intergroup cooperation, and common goals. Service learning encompasses the teaching approach that connects theory and practice by giving learners the opportunity both to participate in a service that meets community needs and to reflect on the experience to gain a deeper understanding and an enhanced sense of civic engagement [23]. Youth CHVs have the opportunity to acquire health care and caregiving skills that could improve their own health and promote transfer of this health-promoting knowledge to their own communities. In the design of the HealthStart program, we built upon the aforementioned frameworks and lessons learned from a previous intergenerational program that was demonstrated to increase health literacy and reduce ageist attitudes [11]. HealthStart activates 2 additional contact theory principles. Equal status is fostered by training all volunteers together and holding them jointly accountable for participant outcomes, while cooperation toward shared goals is achieved when youth CHVs and participants co-create individualized action plans. Throughout the 3-month journey, HCVs serve as clinical mentors, providing guidance while enabling youth CHVs to lead participant engagement.

SDT served as the overarching framework, providing the central mechanism for fostering self-directed, value-driven motivation. Within this framework, contact theory was used to promote equal-status interactions, service-learning pedagogy to encourage civic engagement, and the transtheoretical model to guide stage-matched health coaching. These components were deliberately integrated into a unified theory of change, with each reinforcing the others, to inform both volunteer training and participant-facing activities and ensure a synergistic approach to sustaining behavior change (see Figure 1).

XSL•FO
RenderX

Figure 1. Theory of change as applied to the HealthStart program, with the 3 components—self-determination theory, SMART framework, and intergenerational and service learning—demarcated by different colored arrows across the 4 stages. HCV: health care volunteer; CHV: community health volunteer; SMART: Specific, Measurable, Achievable, Realistic/Relevant, and Time-bound.

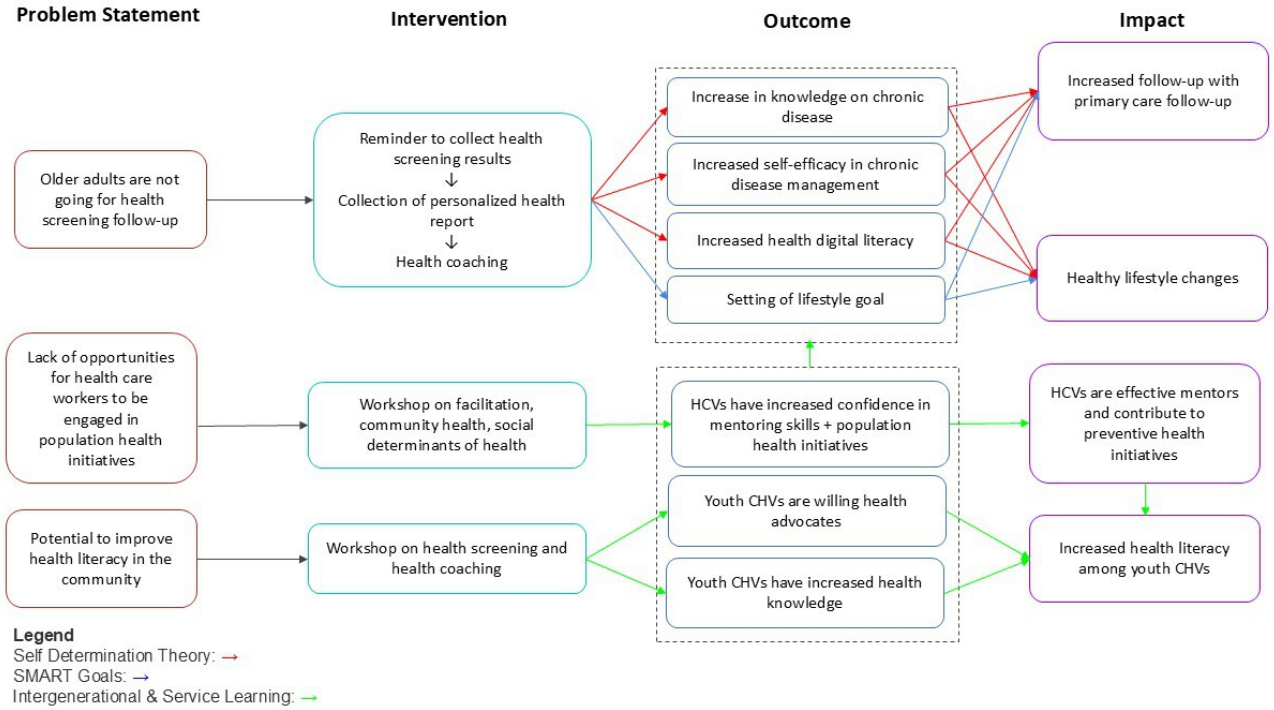

Program Description

Each HealthStart group consists of 1 HCV and 4 youth CHVs who commit to a period of 3 months. Each youth CHV is required to attend the HealthStart training day and is assigned 2 participants subsequently. They are expected to follow up with participants through either physical or internet-based visits (see Figure 2). With each visit, youth CHVs are required to monitor participants' progress toward their health goals. HCVs oversee the teams, provide avenues for content clarification, and professional support as necessary. This multimodal, layperson-led health coaching aims to facilitate participants' need for competence (via greater knowledge about their condition and health-efficacy and increased eHealth knowledge), relatedness (via continual interaction with volunteers and health care professionals), and autonomy (via SMART goal setting).

Participants' baseline health literacy and digital skills were assessed during the initial layperson-led health coaching session using the HealthStart First Visit Form (Multimedia Appendix 1). This structured form captured participants' understanding of their chronic conditions, experience with digital devices, familiarity with health apps (eg, Healthy365 by the Health Promotion Board), and self-reported confidence in managing their health. The form also captured social history and lifestyle behaviors to guide personalized goal setting. These baseline assessments enabled youth CHVs to tailor their coaching approaches according to each participant's readiness, capacity, and specific areas of support required.

Youth CHVs support participants to fulfill the following objectives through the HealthStart program:

1. Learn about their chronic diseases, such as diabetes mellitus, hypertension, or hyperlipidemia, with the aid of the Health Promotion Board (HPB) booklets.
2. Learn at least one digital health app such as HealthHub or Healthy 365:
   - HealthHub is a digital platform that interfaces with medical health record systems across various public health care institutions and national source systems, such as the National Electronic Health Record and National Immunization Registry in Singapore [24].
   - Healthy 365 is a mobile app by the Health Promotion Board (HPB) Singapore that aims to encourage users to adopt a healthier lifestyle. Through the use of gamification and rewards, users are encouraged to sign up for in-app challenges and health programs to earn Healthpoints [25].
3. Enroll with a primary care provider (PCP).
4. Set a lifestyle goal and achieve it.

XSL•FO
RenderX

Figure 2. Structure of the HealthStart program for the youth community health volunteers and residents across 3 months with recommended visit types and fixed team follow-ups for team bonding.

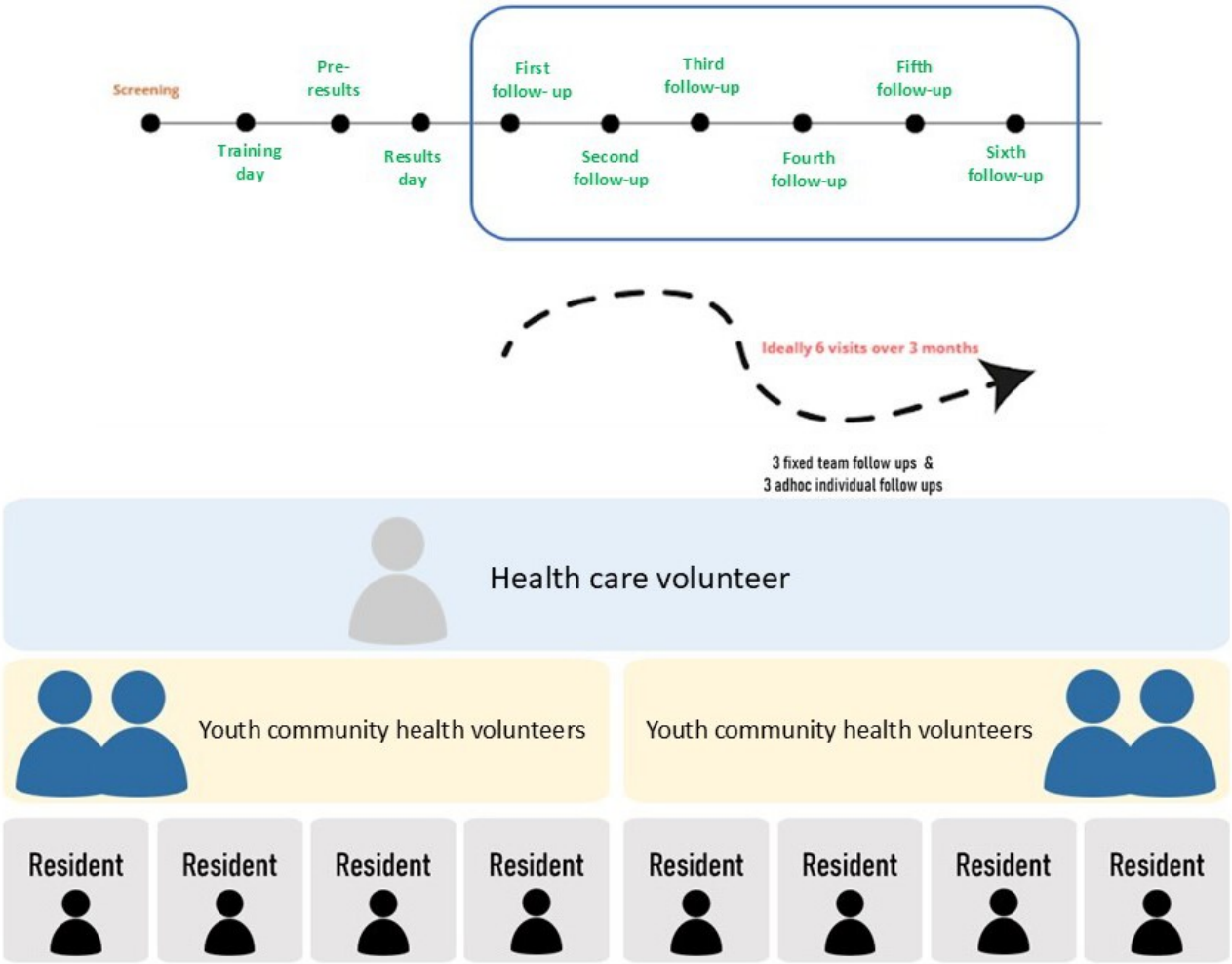

The participants first meet their youth CHV at a community event where health screening reports are distributed and explained by a community nurse. The youth CHVs then conduct their first session of layperson-led health coaching with the participants and build rapport via questions that cover sociodemographic factors and lifestyle aspects using tools such as a Time Wheel and a Healthy Plate (see Multimedia Appendix 1) to guide SMART goal setting.

Youth CHVs follow up with their participants longitudinally over a period of 3 months through in-person or internet-based visits (ie, video calls) every fortnight (for a minimum of six visits). They use online forms (form.sg) to track the participant's progress through HealthStart. By fostering meaningful relationships and using motivational interviewing techniques, youth CHVs guide participants in aligning their core values and purpose with their health aspirations, facilitating deeper self-integration.

Youth CHVs also provide digital literacy training using components from a previously described program [8-10]. After teaching the basics of using a smartphone or alternative digital devices, HCVs will guide their participants to download and use health apps to achieve Objective 2 of HealthStart as described above. This is further described in Multimedia Appendix 1 (see volunteer protocol). To close the digital divide, specific guidance is provided to youth CHVs to help participants without digital devices (see Annex S7 in Multimedia Appendix 1). Youth CHVs are provided a step-by-step guide on what to do during each visit (see Table S1 in Multimedia Appendix 2).

Participants who complete the above goals may be discharged from the program before the 3-month mark. When this occurs, youth CHVs are offered the opportunity to take on additional participants. In addition, those with greater experience are encouraged to support more than 2 participants if they are willing. A total of 10% - 20% of youth CHVs return as recurrent volunteers, gradually building a pool of experienced CHVs that expands the program's scale and capacity. HCVs and youth CHVs convene monthly to reflect on their intergenerational and service-learning experience. During these meetings, youth CHVs report on their participants' progress, while the HCVs provide guidance and advice.

Description of Volunteer Recruitment and Training

Volunteer Recruitment

Youths [26], aged 15 - 35 years old, with no formal health care training or clinical experience, were recruited through volunteer portals, school outreach, and youth networks. While formal health care knowledge was not required, candidates were expected to demonstrate strong interpersonal communication skills, a willingness to learn, and a commitment to engage over

XSL•FO
RenderX

a 3-month period. Selection involved completion of a pretraining knowledge quiz and participation in a full-day in-person training session. Only candidates who passed both formative assessments (content comprehension and role-play evaluations) were deployed. Those who did not meet the criteria were either redirected to other volunteering roles or offered additional training. HCVs, comprising doctors, nurses, and allied health professionals, were recruited via publicity emails and newsletters sent out to relevant professional networks.

Volunteer Training

HealthStart adopted a blended learning approach [27], combining self-paced online modules, a pretraining quiz, and an 8-hour in-person workshop (see Table S2 in Multimedia Appendix 2). Before training, volunteers received educational materials from the HPB on hypertension, hyperlipidemia, and diabetes mellitus, along with a protocol guidebook (see Multimedia Appendix 1). The workshop covered the Singapore health care system, chronic disease management, SMART goal setting, motivational interviewing, and the role of social determinants of health (see Multimedia Appendix 1). From the outset, volunteers were assigned to intergenerational teams to foster collaboration.

The training concluded with four standardized role-play scenarios (see Multimedia Appendix 1) simulating common participant engagement challenges. Using a standardized checklist (see Multimedia Appendix 1), HCVs assessed youth CHVs on communication skills, clinical understanding, and application of motivational interviewing techniques. Only those meeting the competency thresholds were deployed; others were offered remediation or reassigned to alternative roles. This multitiered training and evaluation process ensured consistent quality and preparedness across the volunteer cohort.

Data Collection

Quantitative data were collected between September 2022 and June 2023. For older adult participants, data were obtained at program entry and at the 3-month endpoint using a secure FormSG digital survey administered by trained volunteers. Variables included sociodemographic, primary care follow-up status, and open-ended program feedback.

Quantitative and qualitative feedback from volunteers (CHVs and HCVs) was collected at three milestones: (1) completion of training, (2) the youth CHV's first participant meeting, and (3) program conclusion. After-action reviews involving the organizing team and community partners were conducted at each milestone, with minutes taken and subsequently analyzed alongside volunteer and participant feedback to identify key themes.

Data Analysis

Quantitative analyses were conducted using SPSS Statistics (version 27; IBM Corp). The primary outcome was the change in primary care follow-up status before and after participation in HealthStart. Descriptive statistics were reported with SD for calculated means and with percentages for absolute numbers. The McNemar test was used to assess paired categorical data. Missing data was handled via listwise deletion; no imputation methods were applied.

Fidelity was monitored at various levels: (1) volunteer training completion and pass rates, (2) logging of all volunteer-participant encounters on the electronic survey platform (capturing mode, duration, and coaching elements delivered), (3) cumulative visit counts per participant, and (4) participant attrition at various stages of the program. Program fidelity and outcomes will be reported in detail in a separate mixed-methods outcomes paper.

Ethical Considerations

Ethics approval was obtained from the SingHealth Centralized Institutional Review Board (CIRB references 2022/2700 and 2023/2480). All participants provided informed consent and were given a participant information sheet. All data were anonymized and deidentified in the study. No compensation was provided to the participants.

Results

Program Outreach and Impact

HealthStart was implemented alongside 2 community health screening events in 2022 and 2023. Of 236 eligible individuals, 192 enrolled (79%). The mean age of the participants was 66.9 (SD 9.6) years; 47.9% (92/192) were male and 52.1% (100/192) female. Most participants were of Chinese ethnicity, had completed primary or secondary school education, resided in self-owned flats, and in 3-room public housing (see Table 1).

XSL•FO
RenderX

Table . Participant characteristics.

Characteristics	HealthStart participants (N=192)
Age (years), mean (SD)	66.9 (9.6)
Sex, n (%)	
Male	92 (47.9)
Female	100 (52.1)
Race, n (%)	
Chinese	170 (88.5)
Indian	12 (6.3)
Malay	9 (5.0)
Others	1 (0.2)
Marital status, n (%)	
Never married	31 (16.2)
Married	132 (69.1)
Divorced	7 (4.0)
Widowed	20 (10.5)
Others	1 (0.2)
Status of residential, n (%)	
Own	165 (85.9)
Lodge	9 (4.7)
Rent	18 (9.4)
Type of housing[a], n (%)	
1-room housing	15 (7.9)
2-room public housing	11 (5.8)
3-room public housing	78 (41.1)
4-room public housing	39 (20.5)
5-room public housing	43 (22.6)
Private housing	4 (2.1)

[a]A total of 70% to 80% percent of Singaporeans live in public housing provided by the Singapore Government Housing Development Board (HDB) [28]. The eligibility of these public housing is means-tested, and the combined monthly income of the household must be below approximately US $8500 - $10,500. In addition, most of the 1-room public housing units are rental units and are only eligible for individuals with monthly incomes of below approximately US $1120. As such, housing type is often used as a surrogate for socioeconomic status in Singapore, with larger and private units representing higher socio-economic status [29].

Follow-up rate with primary care increased from 42.7% (82/148) preprogram to 84.5% (125/148) postprogram (χ^2_1=43; P<.001). In models adjusting for demographics and baseline questionnaire scores, effects were greater for older adults (odds ratio [OR] 1.8, 95% CI 1.2 - 2.7) and minority ethnic groups (OR 1.6, 95% CI 1.1 - 2.3). Postprogram feedback was submitted by 109 participants, including open-ended responses.

Across two cycles, 207 volunteers were recruited (58 HCVs and 149 youth CHVs). Following training, 48 HCVs and 138 youth CHVs were assessed for deployment readiness. Ultimately, 33 HCVs and 102 youth CHVs were deployed. Nondeployment (15 HCVs and 36 youth CHVs) reflected nonattendance, role-expectation mismatch, or suboptimal formative assessment performance (see Multimedia Appendix 1). Figure 3 summarizes the number of volunteers at each stage of the program (recruitment, training, and deployment). The mean age of the HCVs was 37 years, 24 (72.7%) were female. A total of 26 were nurses, 6 were doctors, and the rest were allied health professionals. The mean age of the CHVs who were deployed was 24 years, and 75 (73.5%) were female. A total of 58 (56.9%) CHVs were students, and the majority were postsecondary students. In addition, 31 volunteers provided postprogram feedback.

XSL•FO
RenderX

Figure 3. Volunteer deployment flow chart showing total recruitment of volunteers, number of volunteers excluded after training as well as final number of volunteers deployed.

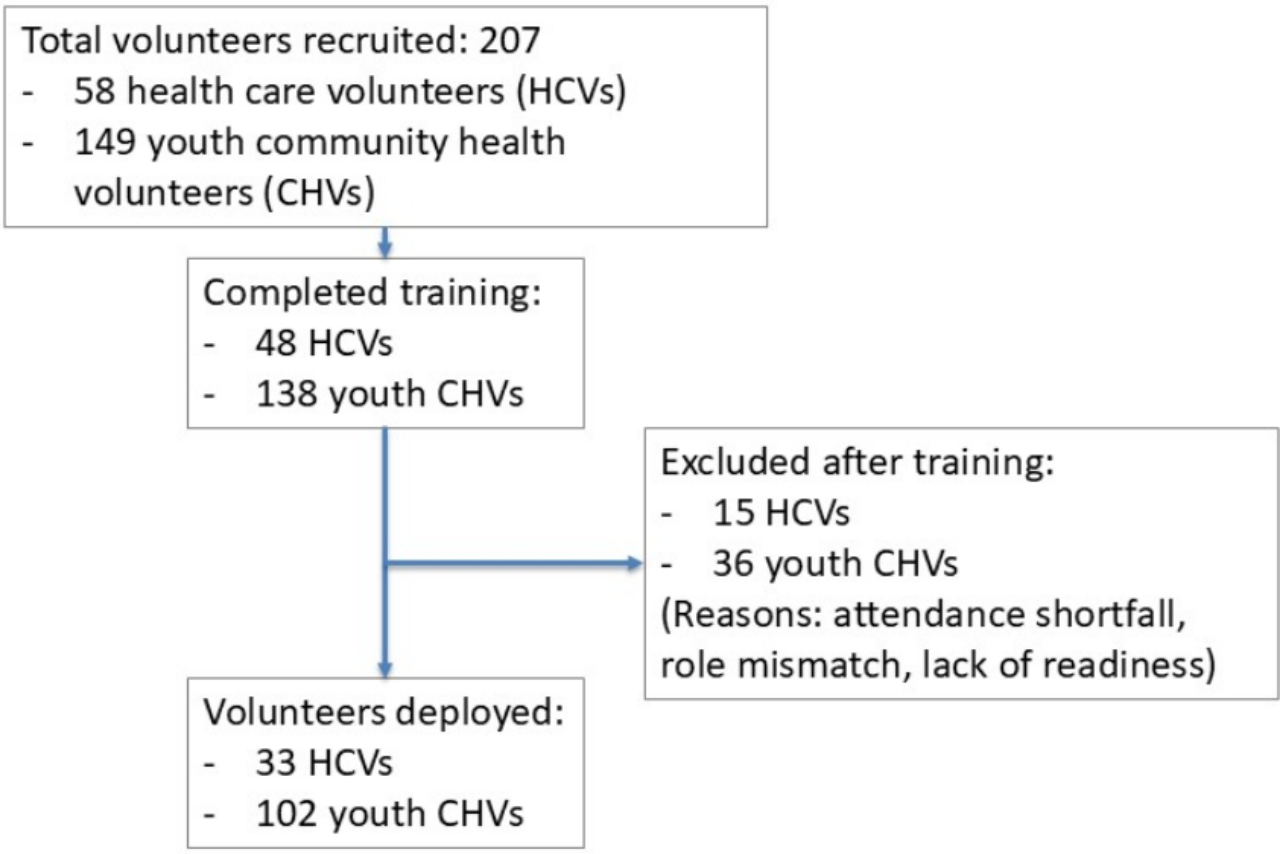

Program Level Reflections

Volunteer Competency and Selection Criteria

A multistep, competency-based selection process ensured standardized and quality layperson-led coaching. Attendance at all training components was mandatory; volunteers were required to pass a pretraining quiz and achieve competency on an HCV-rated checklist across 4 role-play scenarios. Despite completing training, 26% (36/138) of youth CHVs were not suitable for deployment in the actual program.

Tiering of Participant Intervention

The program involved a wide spectrum of participants, some of whom were complex with significant psychosocial issues, whereas some were already taking initiative and self-managing their conditions, thus requiring limited input from the program. To mitigate this, our program provided multiple levels of support for volunteers, including the tiering of participants. This was in keeping with what we know in the transtheoretical model, which posits that health behavior change involves progress through 6 stages of change: precontemplation, contemplation, preparation, action, maintenance, and termination [30]. Through the initial contact on the result collection day, the participants were redistributed to ensure that each group had participants of varied complexities. Each group was led by an HCV who would guide youth CHVs and address any complex questions the participants may have. Finally, an organizing committee member supervised every group of volunteers. These check-ins were crucial in helping youth CHVs refine their approach to distinct groups of participants. Furthermore, to maximize the usability of the program, a tiered model based on the participant's stage of change may be used to tailor each volunteer's intervention and approach.

The transtheoretical model of behavioral change has been used in this setting to gauge the stages of change and readiness for behavioral health change. In this process, participants pass through 5 stages: precontemplation, contemplation, preparation, action, and maintenance. Based on the participant's stage of change and factors controlling the transfer between stages, youth CHVs and HCVs can better apply motivational interviewing and health coaching principles to facilitate the transition to the next stage of change. Service learning pedagogy provides the educational scaffold that enhances youth CHV's self-efficacy and role identity. During program delivery, youth CHVs use autonomy-supportive behaviors; these behaviors, rooted in SDT, are hypothesized to enhance client autonomous motivation. SMART goal setting translates this motivation into concrete, achievable behavioral targets, thereby reinforcing competence. High-quality intergenerational interaction, as specified by contact theory, fosters relatedness and mitigates ageist stereotypes, which the transtheoretical model positions as catalysts for progression through stages of change. The combined effect of these mechanisms is expected to produce stage-matched action plans, sustained self-monitoring, and ultimately improved linkage to primary care and biometric outcomes.

XSL•FO
RenderX

Community Level Reflections

Tapping on Community Assets

There are many preexisting health initiatives in the community, and an initiative such as HealthStart needs to be aligned with community health needs and existing programs to ensure sustainability and constructive collaboration. Before the launch of the program, local grassroots organizations (GROs) were engaged to obtain an understanding of community health needs and existing initiatives.

HealthStart aligned itself with the GROs' goals of improving community health care and obtained buy-in from these stakeholders. As a result of this early alignment, the HealthStart team was able to secure practical support that directly facilitated implementation. This included logistical support such as access to venue spaces (eg, community clubs and activity centers) for resident meet-ups during result collection. Manpower support was also enhanced through on-site assistance from GRO-linked volunteers. In addition, co-branded communication materials helped to improve publicity efforts and build credibility with residents, contributing to stronger engagement during dissemination activities. Despite strong community interest, several challenges emerged in operationalizing partnerships. Space availability and volunteer scheduling also posed logistical barriers. In early cycles, inconsistent communication channels between organizing committees and community venues led to conflicting instructions. To mitigate these issues, a single-point liaison system and standardized briefing packages were introduced in later cycles, improving coordination and reducing friction.

Adoption of a Social Prescription Framework

To effectively improve the health outcomes of participants, the social determinants of the health of each participant need to be addressed in a deliberate manner. Social prescription aims to provide a framework to improve well-being by linking individuals to community assets [31]. During the initial cycles of HealthStart, while volunteers organically attempted to link participants to community exercise and wellness programs, there was no standardized referral pathway or tool to support this effort. This represented a key implementation gap. To address this, future iterations will incorporate a structured social prescribing referral directory and a guided conversation template. This tool will be co-developed with local partners and integrated into the volunteer protocol. Baseline assessments in the First Visit Form will be updated to capture social isolation risk and community engagement levels, allowing for more intentional linkage to social care assets during the program journey. We plan to introduce an applied version of this in HealthStart for youth CHVs in hopes that it can better guide them to address the holistic needs of the participants.

Societal Level Reflections

Importance of Alignment With Population Health Policies

The success of this layperson-led health coaching program is dependent on the broader environment in which the program operates. HealthStart was launched right after HSG, allowing the program to leverage publicity, incentives for primary care enrollment, and subsidized preventive services. Furthermore, the program's goals aligned with those of HSG, which leverages promoting health instead of just health care, as well as focusing on intergenerational bonding between participants and youth CHVs.

Cultivating Intergenerational Relationships

Singapore places great emphasis on the importance of intergenerational ties and intergenerational support [32,33]. Participants and volunteers reported meaningful intergenerational exchanges through HealthStart; the health screening participants shared life experiences and practical challenges with healthy living with volunteers, and the volunteers contributed their energy and fresh perspectives. Intergenerational contact is important in aging populations, as it allows society to address health-related issues through the resources provided by the younger population and intergenerational transfer of health literacy [31].

Known challenges in intergenerational programs (intergenerational conflicts, role conflict and ambiguity, inactive engagement, and difficulty in establishing meaningful connections between generations) were observed [34]. These were mitigated through the availability of HCV mentorship (eg, communication of technical terms and facilitated difficult conversations), language-matched volunteer-participants matching, proactive leadership by the organizing committee.

Discussion

Principal Findings

HealthStart demonstrates the feasibility of an intergenerational, layperson-led health coaching model embedded in primary care. Engaging youth CHVs with no previous health care experience (an underused resource in the community) was associated with substantial gains in primary care follow-up. Integration of SDT, SMART goals, and service-learning concepts provided a coherent framework for participant engagement and volunteer development. Participation of youth CHVs has the additional benefit of improving their adoption of healthy behaviors and diffusing health knowledge within their own communities, amplifying the impact of the program.

Comparison With Previous Work

There has been growing interest in the potential of laypersons as health coaches [35]. Peer-led programs such as the Whole Health Action Management (WHAM) model have increased patient activation among adults with behavioral health conditions, outperforming a clinician-delivered self-management program [36]. Among college students, an 8-week peer-coaching intervention raised physical activity, positive affect, and overall well-being [37]. Smartphone-enabled coaching by nonclinicians has also lowered HbA_{1c} in type 2 diabetes [38].

HealthStart extends these peer-coaching models by adding an intergenerational component, which fosters social cohesion, empowers and expands the scope of volunteers, and introduces mentorship and guidance of these laypersons by the health care professionals. By layering structured goal setting, social

XSL•FO
RenderX

prescribing, and digital health coaching onto this intergenerational platform, HealthStart addresses gaps—particularly digital literacy deficits—seldom tackled in earlier lay-coach designs.

Study protocols are being developed evaluating the comparative effectiveness of health coaching by laypersons versus health care professionals; HealthStart's model may inform such evaluations [39,40].

Limitations

The study has several limitations. First, the study was conducted within a specific geographical region in Singapore, which may limit its generalizability to other cultural or health care settings. Differences in population demographics and community engagement practices could influence the feasibility and effectiveness of similar programs elsewhere.

Second, the recruitment and retention of youth CHVs posed a challenge. Despite rigorous selection and training processes, a proportion of youth CHVs were found to be unsuitable for deployment, requiring additional training or reassignment to other roles. As a relatively new field, layperson-led health coaching programs are challenged by the lack of a structured framework, reimbursement, and community coalition [41].

Third, the study relied on self-reported outcomes from participants and volunteers, which may introduce response biases, particularly social desirability bias. Although objective measures such as enrollment with a primary care provider were tracked, triangulation with objective metrics was not feasible in this pilot due to limitations in data sharing agreements and system interoperability.

Finally, the absence of a control group limits causal inference. Improvement in primary care follow-up seen may reflect the effect of concurrent initiatives such as HSG.

Future Directions

Building on initial learnings, the next phase of HealthStart will focus on improving scalability, integration, and evaluation rigor. First, integration with national platforms under HSG will be deepened through formalized referral pathways and alignment of volunteer coaching with individualized health plans. Greater data connectivity will be pursued to enable linkage with electronic medical records and backend analytics (eg, HealthHub and Healthy365) for more robust outcome tracking.

Second, a structured social prescribing framework will be introduced. This includes enhanced baseline assessments to identify social needs (eg, isolation and financial insecurity), and a digital tool to guide volunteers in referring participants to community resources.

Third, comparative research will examine the effectiveness of youth-led versus professional-led coaching models, as well as the cultural adaptability of HealthStart across diverse settings. Digital adaptations for multilingual, underserved populations will also be explored to extend program reach and promote health equity.

Finally, to sustain scale-up, efforts will focus on reducing coordination burden via digital tools and supporting volunteer retention through micro-incentives (eg, digital badges and learning credits), structured recognition, and clearer role progression. Cost-effectiveness analyses and workforce planning will be essential to assess the long-term viability of integrating layperson-led coaching into primary care. These have to be tailored to the resource demands and contextual constraints of the specific community.

Conclusion

HealthStart demonstrates the feasibility and promise of a structured, intergenerational, layperson-led health coaching model integrated within a primary health system. Designed with a foundation in behavior change theory, aligned with community needs, and supported by digital health tools, the program addresses key barriers to chronic disease follow-up. We identified key lessons learned in the conceptualization and implementation of the program. These insights may support efforts to replicate this model of care across diverse health systems.

Acknowledgments

This research is supported by the Singapore Ministry of Health's National Medical Research Council under the Fellowship Program by SingHealth Regional Health System, the Population-Based, Unified, Learning System for Enhanced and Sustainable (PULSES) Health Centre Grant (NMRC/CG/C027/2017_SHS), the Healthy, Empowered and Active Living (HEAL) fund, and the Infocomm Media Development Authority Digital for Life Fund.

No generative artificial intelligence tools (eg, ChatGPT or similar models) were used in the drafting, editing, or writing of this manuscript. All content was produced and reviewed by the authors without automated generation assistance.

Authors' Contributions

Conceptualization: XH, KSY, KYYN
Methodology: XH, KSY, KYYN
Investigation: KSY, JXL, JYY, HAL, LPZT, SQL, KJ, KWAT, AJYT
Data curation: KSY, JXL, JYY, HAL, NHWN, LPZT
Formal analysis: KSY, ASTK, JXL
Writing – original draft: XH, KSY, KYYN, ASTK
Writing – review & editing: XH, KSY, KYYN, ASTK, MHL

XSL•FO
RenderX

Supervision: LLL, KYYN
Project administration: LPZT

Conflicts of Interest

KYYN is supported by the Singapore Ministry of Health's National Medical Research Council under the Fellowship Program by SingHealth Regional Health System, the Population-Based, Unified, Learning System for Enhanced and Sustainable (PULSES) Health Center Grant (NMRC/CG/C027/2017_SHS), and the SingHealth Academic Medicine HEART grant (AM/HRT025/2023). The other authors declare no other conflicts of interest.

Multimedia Appendix 1
Detailed HealthStart program outline and protocol.
[DOCX File, 1917 KB - formative_v9i1e76592_app1.docx]

Multimedia Appendix 2
Suggested guide for HealthStart visits and content of volunteer training.
[DOCX File, 23 KB - formative_v9i1e76592_app2.docx]

Checklist 1
Standards for Reporting Implementation Studies (StaRI) checklist.
[DOCX File, 81 KB - formative_v9i1e76592_app3.docx]

Checklist 2
The Template for Intervention Description and Replication (TIDieR) checklist.
[PDF File, 467 KB - formative_v9i1e76592_app4.pdf]

References

1. United Nations decade of healthy ageing (2021–2030). World Health Organization. 2020. URL: https://www.who.int/initiatives/decade-of-healthy-ageing [accessed 2025-08-31]
2. Griffin SJ. Lost to follow-up: the problem of defaulters from diabetes clinics. Diabet Med 1998;15 Suppl 3:S14-S24. [doi: 10.1002/(sici)1096-9136(1998110)15:3+3.3.co;2-9]
3. Haggstrom DA, Taplin SH, Monahan P, Clauser S. Chronic care model implementation for cancer screening and follow-up in community health centers. J Health Care Poor Underserved 2012 Aug;23(3 Suppl):49-66. [doi: 10.1353/hpu.2012.0131] [Medline: 22864487]
4. National Population Health Survey 2020: Ministry of Health Singapore, Health Promotion Board; 2021.
5. Many not following up after health screening. Singapore General Hospital. 2017. URL: https://www.sgh.com.sg/news/others/many-not-following-up-after-health-screening [accessed 2024-06-26]
6. Hernandez DC, Johnston CA. Individual and environmental barriers to successful aging. Am J Lifestyle Med 2017 Jan;11(1):21-23. [doi: 10.1177/1559827616672617]
7. What is healthier SG? Healthier SG. URL: https://www.healthiersg.gov.sg/about/what-is-healthier-sg/ [accessed 2024-06-26]
8. Ngiam NHW, Yee WQ, Teo N, et al. Building digital literacy in older adults of low socioeconomic status in Singapore (project wire up): nonrandomized controlled trial. J Med Internet Res 2022 Dec 2;24(12):e40341. [doi: 10.2196/40341] [Medline: 36459398]
9. Lu SY, Yoon S, Yee WQ, Heng Wen Ngiam N, Ng KYY, Low LL. Experiences of a community-based digital intervention among older people living in a low-income neighborhood: qualitative study. JMIR Aging 2024 Apr 25;7:e52292. [doi: 10.2196/52292] [Medline: 38662423]
10. Lim HA, Lee JSW, Lim MH, et al. Bridging connectivity issues in digital access and literacy: reflections on empowering vulnerable older adults in Singapore. JMIR Aging 2022 May 3;5(2):e34764. [doi: 10.2196/34764] [Medline: 35503520]
11. Ng KYY, Leung GYC, Tey AJY, et al. Bridging the intergenerational gap: the outcomes of a student-initiated, longitudinal, inter-professional, inter-generational home visit program. BMC Med Educ 2020 Dec;20(1):148. [doi: 10.1186/s12909-020-02064-x]
12. Verma I, Gopaldasani V, Jain V, et al. The impact of peer coach-led type 2 diabetes mellitus interventions on glycaemic control and self-management outcomes: A systematic review and meta-analysis. Prim Care Diabetes 2022 Dec;16(6):719-735. [doi: 10.1016/j.pcd.2022.10.007]
13. Ballard PJ, Ozer EJ. In: Rogers J, Terriquez V, Russell ST, editors. Contemporary Youth Activism: Advancing Social Justice in the United States: University of California Press; 2020:223-244.
14. Hoffmann TC, Glasziou PP, Boutron I, et al. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. BMJ 2014 Mar 7;348(mar07 3):g1687. [doi: 10.1136/bmj.g1687] [Medline: 24609605]

XSL•FO
RenderX

15. Pinnock H, Barwick M, Carpenter CR, et al. Standards for Reporting Implementation Studies (StaRI) Statement. BMJ 2017 Mar 6;356:i6795. [doi: 10.1136/bmj.i6795] [Medline: 28264797]
16. HealthStart. TriGen. URL: https://www.trigen.sg/healthstart [accessed 2024-06-26]
17. About PHICO. Singapore General Hospital. URL: https://www.sgh.com.sg/PHICO/Pages/About-PHICO.aspx [accessed 2024-06-26]
18. Screen for life. HealthHub. URL: https://www.healthhub.sg/programs/screen_for_life#home [accessed 2024-06-26]
19. Ryan RM, Deci EL. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. Am Psychol 2000 Jan;55(1):68-78. [doi: 10.1037//0003-066x.55.1.68] [Medline: 11392867]
20. Bailey RR. Goal setting and action planning for health behavior change. Am J Lifestyle Med 2019;13(6):615-618. [doi: 10.1177/1559827617729634] [Medline: 31662729]
21. Rollnick S, Miller WR, Butler CC. Motivational Interviewing in Health Care: Helping Patients Change Behavior: New York: The Guilford Press; 2008.
22. Vanderven K. Intergenerational theory: the missing element in today's intergenerational programs. Child Youth Serv 1999;20:33-47. [doi: 10.1300/J024v20n01_04]
23. Resch K, Schrittesser I. Using the service-learning approach to bridge the gap between theory and practice in teacher education. Int J Incl Educ 2023 Aug 24;27(10):1118-1132. [doi: 10.1080/13603116.2021.1882053]
24. About us. HealthHub. URL: https://www.healthhub.sg/about-us [accessed 2025-02-28]
25. The Healthy365 app. HealthHub. URL: https://www.healthhub.sg/programmes/healthyliving [accessed 2025-02-28]
26. Youth. Ministry of Culture, Community & Youth. 2025. URL: https://www.mccy.gov.sg/sectors/youth
27. Hoic-Bozic N, Mornar V, Boticki I. A blended learning approach to course design and implementation. IEEE Trans Educ 2009;52(1):19-30. [doi: 10.1109/TE.2007.914945]
28. Households. Department of Statistics Singapore. URL: https://www.singstat.gov.sg/find-data/search-by-theme/households/households/latest-data [accessed 2025-06-06]
29. Wong FY, Wong RX, Zhou S, et al. Effects of housing value and medical subsidy on treatment and outcomes of breast cancer patients in Singapore: A retrospective cohort study. Lancet Reg Health West Pac 2021 Jan;6:100065. [doi: 10.1016/j.lanwpc.2020.100065] [Medline: 34327401]
30. Prochaska JO, Velicer WF. The transtheoretical model of health behavior change. Am J Health Promot 1997;12(1):38-48. [doi: 10.4278/0890-1171-12.1.38] [Medline: 10170434]
31. World Health Organization Regional Office for the Western Pacific. A Toolkit on How to Implement Social Prescribing: Manila: WHO Regional Office for the Western Pacific; 2022. URL: https://www.who.int/publications/i/item/9789290619765 [accessed 2025-08-31]
32. Intergenerational and extended family ties increasingly important as population ages. Ministry of Social and Family Development. URL: https://www.msf.gov.sg/media-room/article/Intergenerational-and-extended-family-ties-increasingly-important-as-population-ages [accessed 2024-06-26]
33. Promoting intergenerational harmony. Ministry of Health. URL: https://www.moh.gov.sg/ifeelyoungsg/how-can-we-build-stronger-ties/promoting-inter-generational-harmony [accessed 2024-06-26]
34. Combatting ageism. World Health Organization. URL: https://www.who.int/teams/social-determinants-of-health/demographic-change-and-healthy-ageing/combatting-ageism [accessed 2024-06-26]
35. Tseng JT, Kuo T, Huang CK, et al. Examining the impact of the design-thinking intergenerational service-learning model on older adults' self-care behaviors and well-being. J Appl Gerontol 2025 May;44(5):679-687. [doi: 10.1177/07334648241286327] [Medline: 39297722]
36. Cook JA, Jonikas JA, Burke-Miller JK, et al. Whole health action management: a randomized controlled trial of a peer-led health promotion intervention. PS (Wash DC) 2020 Oct 1;71(10):1039-1046. [doi: 10.1176/appi.ps.202000012]
37. Yan Z, Peacock J, Cohen JFW, et al. An 8-week peer health coaching intervention among college students: a pilot randomized study. Nutrients 2023 Mar 4;15(5):1284. [doi: 10.3390/nu15051284] [Medline: 36904282]
38. Wayne N, Ritvo P. Smartphone-enabled health coach intervention for people with diabetes from a modest socioeconomic strata community: single-arm longitudinal feasibility study. J Med Internet Res 2014 Jun 6;16(6):e149. [doi: 10.2196/jmir.3180] [Medline: 24907918]
39. Chung EKH, Wong ELY, Leung HHW, Yeung DY, Yeoh EK, Chen FY. Lay health coaching intervention for older adults with chronic diseases: study protocol for a pragmatic randomised controlled trial. Trials 2024 Dec 18;25(1):817. [doi: 10.1186/s13063-024-08649-x] [Medline: 39696468]
40. Federman AD, Barry M, Moas E, et al. Protocol for a feasibility randomized trial of self-management support for people with chronic obstructive pulmonary disease using lay health coaches. Contemp Clin Trials 2021 Nov;110:106570. [doi: 10.1016/j.cct.2021.106570] [Medline: 34560265]
41. Patil SJ. Task sharing chronic disease self-management training with lay health coaches to reduce health disparities. Ann Fam Med 2020 Jan;18(1):2-3. [doi: 10.1370/afm.2513] [Medline: 31937525]

XSL•FO
RenderX

Abbreviations

CHV: community health volunteer
GRO: grassroots organization
HCV: health care volunteer
HPB: Health Promotion Board
HSG: Healthier SG
SMART: Specific, Measurable, Achievable, Realistic/Relevant, and Time-bound

Edited by J Sarvestan; submitted 27.04.25; peer-reviewed by A Ramadas, P Pichayapinyo; revised version received 11.08.25; accepted 13.08.25; published 22.09.25.

Please cite as:
Huang X, Yow KS, Kwan AST, Yeo JY, Lim HA, Lim JX, Lim MH, Teo LPZ, Ngiam NHW, Lim SQ, Jaichandra K, Tang KWA, Tey AJY, Low LL, Ng KYY
Implementation, Challenges, and Outlook of an Intergenerational, Layperson-led, Health Coaching Program (HealthStart): A Pilot Case Study
JMIR Form Res 2025;9:e76592
URL: https://formative.jmir.org/2025/1/e76592
doi:10.2196/76592

© Xiaoting Huang, Ka Shing Yow, Audrey Shu Ting Kwan, Jin Ye Yeo, Haikel A Lim, Jie Xin Lim, Meng Han Lim, Lynn Pei Zhen Teo, Nerice Heng Wen Ngiam, Si Qi Lim, Kharuna Jaichandra, Kai Wen Aaron Tang, Angeline Jie-Yin Tey, Lian Leng Low, Kennedy Yao Yi Ng. Originally published in JMIR Formative Research (https://formative.jmir.org), 22.9.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Development of a Patient-Centered Digital Health Care Technology for Young Adults in Opioid Use Disorder Treatment: Qualitative Study

Karen Alexander, PhD; Madison Scialanca, MPH

Friends Research Institute, 1040 Park Avenue, Suite 103, Philadelphia, PA, United States

Corresponding Author:
Karen Alexander, PhD
Friends Research Institute, 1040 Park Avenue, Suite 103, Philadelphia, PA, United States

Abstract

Background: Young adults, defined as individuals between the ages of 18 and 29 years, drop out of opioid use disorder (OUD) treatment more often than older adults. Premature treatment drop-out substantially increases fatal overdose risk. Self-monitoring through text messaging has been researched extensively among people with OUD to identify drop-out risk factors. Self-monitoring could potentially improve methadone treatment engagement among young adults, who are a population that is both hard to reach and more likely to use technology compared to older adults. Self-monitoring can increase risk factor awareness and help patients and counselors develop targeted coping strategies and treatment plans. However, embedding a discussion of risk factor information into existing counseling sessions has been limited and may offer a promising opportunity to improve engagement among young adults.

Objective: This pilot proof-of-concept study examined the implementation of self-monitoring intervention, AWARE (Awareness and Response to the Environment), designed to bring attention to treatment drop-out risk factors among young adults and create discussion about risk factors with their existing treatment counselor.

Methods: In this formative research, a convenience sample (N=8) of young adults (n=3, 38%) in methadone treatment, their counselors (n=3, 38%), and clinic leadership (n=2, 25%) were recruited from an opioid treatment program after referral from treatment staff. Participants were interviewed to obtain feedback as AWARE was developed. In semistructured interviews, perspectives regarding barriers to treatment for young adults and AWARE utility were obtained. Concurrently, 3 dyads of young adults (n=3, 38%) and counselors (n=3, 38%) piloted the intervention daily for 4 weeks.

Results: The 3 consented young adults with OUD participants (n=2, 67% female; n=2, 67% Latino/a) were sent daily surveys for 28 days (53% overall completion rate). Young adults and counselors found AWARE relevant to their treatment experience and acceptable to complete over 4 weeks. The most reported daily stressors included concerns about the health and well-being of a family member, challenges with staying organized, and feeling overwhelmed by responsibilities without adequate support. In qualitative interviews, counselors and clinic leadership reported that AWARE presented a relevant, new way to engage young adults daily, in addition to weekly counseling sessions. Young adults felt that prompts sent by AWARE offered a type of social support they lacked, like “someone checking in on them.”

Conclusions: Overall, young adult and counselor participants were able to engage in AWARE in a busy clinic environment, and participants and clinic leadership found it valuable. By addressing common stressors and providing a sense of social connection, AWARE may help fill a gap in support between counseling sessions. However, the study was limited by the small number of young adults engaging in methadone treatment. Further research is needed to refine the measures and methods of AWARE and evaluate its effectiveness.

(JMIR Form Res 2025;9:e67401) doi:10.2196/67401

KEYWORDS

ecological momentary assessment; treatment barriers; opioid use disorder; opioid treatment programs; qualitative study; patient-centered; digital health care; digital care; digital technology; young adult; teenager; juvenile; proof-of-concept; self monitor; EMA; overdose; formative research; counselor; qualitative interview

XSL•FO
RenderX

Introduction

Background

Young adults, aged 18 - 29 years, have the highest opioid use rate per capita [1]. Despite the efficacy of medications (eg, methadone and buprenorphine) to prevent morbidity and mortality related to opioid use, young adults with opioid use disorder (OUD) are less likely to initiate treatment [2,3] and more likely to drop out of treatment [4] compared to older age groups. A personalized behavioral treatment responsive to the individual needs of young adults may increase retention by addressing coping skills associated with young adults with OUD stressors and thoughts of leaving treatment [5-8].

Addressing Risk Factors Through Ecological Momentary Assessment

The scientific understanding of risk assessment in a real-world context has been advanced by ecological momentary assessment (EMA) research methods [9,10]. EMA is a data collection technique requiring frequent self-monitoring of symptoms, thoughts, and social interactions, enabling the identification of unfolding patterns of mental states and behavior [11]. Self-monitoring through text messaging (an EMA method) has been used extensively among people with OUD to identify treatment drop-out risk factors through pattern recognition [12-17].

People with OUD have differing retention trajectories based on levels of stress [18,19]. The variability of stress, not the averages at points in time, best predicts craving and future opioid use [20]. According to self-management and cognitive relapse prevention theories, greater attention to and awareness of mental states may lead to improvements [21,22]. Drop-out risk factors like stress may normally be unconscious to a person, but self-monitoring can bring them to the forefront and, thus, increase the opportunity for the patient and counselor to develop more specific plans, coping strategies, and more control [23]. It is likely that automatic responses to stress, when brought to cognitive awareness, become easier to address and extinguish [24].

Previous research has incorporated EMA data within OUD treatment to engage participants through text messages to remind participants of appointments [25], mobile health (mHealth) applications to deliver therapy [26-28], and web-based software to access and deliver cognitive behavioral therapy (CBT) [29]. However, few digital interventions are described in the literature that incorporate EMA data into counseling sessions within an existing opioid treatment program (OTP). Integrating real-time patient-generated data at the point of care has the potential for a significant impact by informing risk assessment and augmenting existing treatment.

Study Purpose

The present study developed and pilot-tested an intervention tailored for young adults with OUD called Awareness and Response to the Environment (AWARE). AWARE used Research Electronic Data Capture (REDCap; Vanderbilt University) software to send EMA surveys on a study-provided phone, collecting patient-generated data on antecedents of treatment drop-out that could be integrated into weekly counseling sessions. The purpose of this paper is to (1) describe the use of AWARE by young adults with OUD and treatment staff, (2) detail the perceived barriers and facilitators of treatment engagement, and (3) describe how AWARE may address identified barriers. This pilot, proof-of-concept project will lay the foundation for a future clinical trial of AWARE.

Theoretical Framework

The conceptual framework for AWARE draws from self-management and cognitive-behavioral relapse prevention theories, which are applied to the retention stage of the OUD Cascade of Care [30]. To target improvements in the retention stage, self-efficacy beliefs (“I’m going to the clinic today”) and adherence (“I received medication today”) to medication are crucial [31,32]. Interventions underpinned by self-management theory in HIV and substance use research have shown success in increasing medication receipt and adherence [33,34]. Self-management theory emphasizes that improvements in self-efficacy beliefs can result in better decision-making and intention to adhere to healthy behaviors [33-35]. Cognitive-behavioral relapse prevention theory demonstrates that relapse (defined as a setback in progressing toward recovery) is not a binary event but, rather, a dynamic, fluctuating process (see Figure 1) [7,26]. The rationale for integrating these 2 theories stems from the utility of self-management–based interventions in chronic disease and the chronic, relapsing nature of addiction [34,36]. Improved self-management skills for someone with OUD may prevent future relapses but also could improve daily cravings, negative mood, and stress, which may reflect a greater benefit to an individual than abstinence from drugs alone [37]. These personal benefits may also positively reinforce treatment adherence, thereby making the work of recovery more appealing to participants [38].

XSL•FO
RenderX

Figure 1. Integration of AWARE components with theoretical frameworks. AWARE: Awareness and Response to the Environment; EMA: ecological momentary assessment.

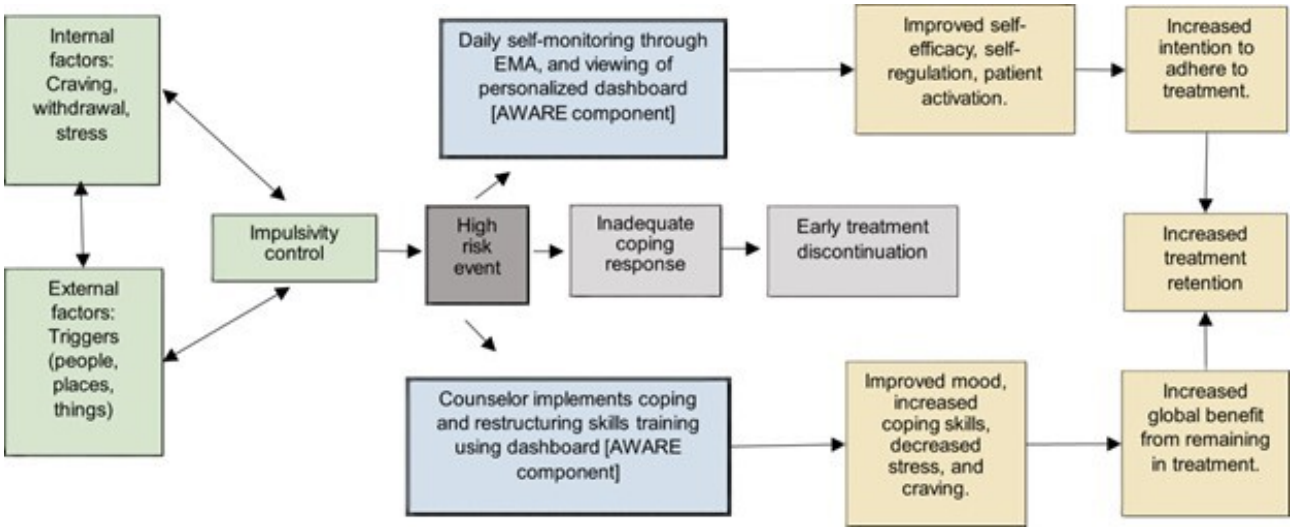

Methods

Setting and Design

Young adults with OUD and treatment staff recruitment took place at an urban, Northeastern United States OTP that offers methadone or buprenorphine and individual and group counseling informed by CBT, using the National Institute on Drug Abuse (NIDA) CBT manual [39]. Each incoming patient receives an individual counselor with a caseload of 25 - 30 patients. The OTP treated 426 patients in 2023 for OUD with either methadone (93%) or buprenorphine (7%). Approximately 10% of the patient population were young adults in 2023. The program does not currently use patient-centered digital health care technology. Medications administered and dispensed, as well as doses and drug screening results, are recorded in an electronic health record (EHR). Urine samples are obtained randomly at least twice per month, and patients are not discharged for drug use alone. Rather, continued opioid or other drug use may impact clinical decisions regarding take-home medication, dosage, and counseling approaches. Reasons for discharge are noted in the EHR. During our pilot study, we capitalized on already existing resources, including existing counselors, to adapt AWARE into the opioid treatment center structure. This study used a qualitative approach, which is appropriate for formative work. This article adheres to the COREQ (Consolidated Criteria for Reporting Qualitative Research) regarding key data collection and analysis aspects.

AWARE Intervention

AWARE was developed as a self-monitoring, mobile phone–based intervention that couples EMA surveys with counselor feedback during weekly CBT sessions. The goal of AWARE is to augment standard care, which includes medication, through identifying and responding to high-risk events or situations known to predict drop-out. The EMA structure, content, and timing in this project are informed by longitudinal studies that sought to predict early treatment dropout using EMA [18,40-46]. For this pilot study, the Daily Hassles Survey [18], a list of external stressors, was used to identify stressors particular to young adults with OUD. Weekly data summaries were provided to the participant and their counselor via email (see Table 1).

Table . Awareness and Response to the Environment intervention integration of EMA data into counseling sessions for young adults with opioid use disorder.

Participation and roles of participant and counselor	Activities	Theoretical construct
Participant (daily) self-monitors via EMA, views dashboard, real-world practice of skills	Identifies and examines thoughts, symptoms, triggers of drop-out; practices coping skills; practices problem solving	Self-efficacy, self-regulation, patient activation; coping, relapse or drop-out prevention; self-efficacy, relapse prevention;
Study counselor (weekly) views dashboard, analyzes patterns	Identifies and examines thoughts, symptoms, triggers of drop-out; identifies contextual issues affecting intention to stay in treatment; teaches and leads rehearsal of coping skills (emotional and trigger coping); engages in problem solving	Self-efficacy, self-regulation; self-efficacy, self-regulation; coping, relapse or drop-out prevention; coping, relapse, or drop-out prevention

Eligibility Criteria

The study recruited young adults with OUD who were (1) aged 18 - 29 years, (2) receiving counseling during OUD treatment, and (3) English speaking. We excluded young adults with OUD if they exhibited (1) active suicidal ideation and/or (2) active psychosis. Inclusion criteria for treatment provider/staff interviews included (1) counselors with a client enrolled in AWARE or (2) clinic leadership.

XSL•FO
RenderX

Recruitment and Screening Procedures

The research assistant approached eligible young adults with OUD participants during dosing, asking them if they would like to participate in a study regarding a mobile phone–based intervention. The research assistant then obtained written informed consent and conducted the baseline interview. Participants were then oriented to the EMA survey procedures. Treatment center staff were recruited at a kick-off breakfast event at the treatment center. Staff provided verbal consent to the interviews before their interview began. Of the 14 young adult clients enrolled in the opioid treatment program, 8 were contacted and the research team never received a response; 3 were discharged from the treatment center before contact could be made, and 3 were successfully enrolled.

Measures

All study participants completed a brief demographic questionnaire at baseline. In addition, young adults with OUD participants completed the daily stressors questionnaire. Qualitative interviews were performed using a semistructured guide before and after AWARE implementation. All participants were asked key questions regarding perceived barriers to treatment for young adults with OUD and the utility of AWARE in improving treatment engagement. Suggestions for improvement, as well as content or delivery methods, were obtained. Participants received a US$50 gift card upon completion of the qualitative interview. Young adults with OUD participants were paid US$50 in cash at the beginning and end of the AWARE intervention, and they also received a study phone if desired.

Data Analysis

Interview data were analyzed using thematic analysis methods: generating initial codes, searching, reviewing, defining, and naming subthemes, and identifying basic and global themes. The research team consisted of a master's level Research Assistant and a PhD-prepared nurse scientist, each with extensive research and clinical experience working with young adults with OUD. Transcripts were reviewed independently by a master's prepared research assistant and a PhD-prepared investigator using rigorous and accelerated qualitative data reduction (RADaR) and content analysis [47]. A series of spreadsheets was created to produce short, concise data tables. From the discussion of these tables, a consensus was reached on the content and relevance of themes identified. The identified daily stressors were tabulated using frequency of identification and ranked from most identified to least identified.

Ethical Considerations

The study was approved by both WCG Institutional Review Board and the City of Philadelphia Department of Public Health Institutional Review Board (IRB# 2024-07). We obtained written, informed consent from all patient participants and verbal consent from the counselor and clinic leadership to be interviewed. All study data were deidentified before analysis. Participants were assigned unique study IDs, and no identifying information (eg, names, addresses, and phone numbers) was stored with survey or interview responses. Data collected through electronic platforms (eg, REDCap) were stored on secure, password-protected cloud-based storage systems compliant with institutional and Health Insurance Portability and Accountability Act (HIPAA) standards. Access to identifiable data was restricted to key research personnel and monitored through audit logs. All qualitative interview recordings were transcribed with identifying details removed and securely deleted after transcription. Findings are reported in aggregate form to ensure individual confidentiality. Patient and staff participants were paid US$50 for completing baseline and follow-up interviews at the beginning and end of the 4-week intervention period.

Results

Overview

The 3 consented young adult participants with OUD (n=2, 67% female; n=2, 67% Latino/a) were sent daily surveys for 28 days (53% overall completion rate), and only 2 young adult participants with OUD finished the study. All 3 counselors received a weekly data summary. Young adults with OUD and counselor ratings of the usability of the platform were favorable; all agreed with statements that the daily surveys were clinically useful, easy to navigate, and relevant to the treatment experience. The young adults with OUD and treatment center staff interviews centered on the previously identified themes of treatment engagement stressors and barriers and the feasibility and usability of AWARE to promote treatment engagement.

Participant Characteristics

Three young adult participants with OUD consented to participate in the study. The first participant was a Latina woman, aged 27 years, living in her own home with her partner and child. The second participant was a Latino man living in a family member's house. The third participant was a Black woman experiencing ongoing homelessness; however, she was lost to follow-up after responding to one week of surveys. All participants received a study phone to complete the AWARE surveys.

Five treatment center staff were recruited. All but one of the counselors had less than 6 years of experience as a drug and alcohol counselor. One clinic administrator had over 20 years of experience working in therapy within opioid treatment programs. Two treatment center staff identified as Black, one identified as Asian, and 2 identified as White. Treatment center staff interviewed primarily identified as women, except for one man.

Treatment Engagement Stressors and Barriers for Young Adults

Over the 28-day period, young adults with OUD most frequently reported stressors related to family-related obligations (10 responses), having to do things without help (9 responses), and being organized (9 responses). Other common stressors included the health of a family member, physical care, and medical abilities (each 8 responses). In contrast, children, partners, and boredom ("having enough to do") were reported least often (0–3 responses). Overall, stressors spanned family, financial, health, and daily-life domains, with practical and relational demands emerging as the most prominent. In qualitative interviews, 2

XSL•FO
RenderX

young adults with OUD participants identified "not having much to do" and not having any social connections outside of the clinic as a frequent stressor. The most frequently reported daily stressors were identified as "the health and well-being of a family member," "being organized," and "having too many things to do without help."

One participant described the care she provides daily for her child and her partner, who works night shift and is also a patient at the methadone clinic.

> *I have to be up and ready by 7:30 AM. [My partner is] outside and then I know we're coming to the clinic and he's tired, so he has to go home and sleep till he goes to work. So it's like I know I can't do nothing Monday through Friday but cater to him and the baby.* [Participant 1]

Each young adult with OUD reported in their interviews that they had difficulty in the past with clinic regulations and the clinic environment. This was not the first treatment episode for any young adults with OUD participants. The social environment around the clinic was not seen as positive. Most participants reported going to and from the clinic and not spending time with peers.

> *Well, it's like with this time, this isn't my first rodeo, so I know what to do this time. So I know this might not be the best answer, but this works for me. I isolate and it's not good. I isolate because I don't live in a nice neighborhood. There's drugs everywhere, but I don't like, it's definitely my mindset. I just go straight home and I try to keep a one track mind, but I mean, that wasn't good for me to do it back then.* [Participant 3]

Young adults with OUD participants also reported that they did not have much to do and making money at a job posed a problem, as they feared they might use the money for drugs.

> *Because you got to understand the last 10 years of my life getting high was my hobby, my interest. So when I just stop, you do feel lost, then that's when depression and being miserable, that all kicks in. You just don't know what to do. But I just got to find new shit to do, to keep me occupied, and I don't even know what to do right now. You know what I mean? It's weird. I feel like I'm learning myself. I feel like how I felt when I was in high school. I am learning, trying to find myself, [trying to find things] that interest me instead of getting high.* [Participant 3]

Treatment center staff had a similar perspective to young adults with OUD regarding stressors and barriers to treatment engagement. All 5 treatment staff participants mentioned housing as a distinct barrier to stay in treatment. However, treatment center staff also highlighted issues related to childhood trauma, motivation, and family disconnection as barriers to maintaining recovery.

> *A lot of [clients] are not close with their families anymore, so I mean they don't have anything to look forward to. Holidays, everything is just a regular day to them, so because they don't have the motivation or the push, it makes it very difficult to keep them engaged on sobriety. They don't see a reason to be sober.* [Staff 2]

Difficulties staying away from drug use were linked by several treatment center staff to the environment in which participants lived and socialized. According to treatment center staff, there was little motivation to stay away from drugs and staying in treatment was lacking for many participants, due to a lack of positive relationships. Treatment staff also mentioned that the relationships that they do have at the treatment center are not always positive and that many times staying away from people is a strategy young adults employ in order to stay engaged in treatment.

AWARE Feasibility, Acceptability, and Usability

Young adults and treatment center staff found AWARE to be relevant, acceptable, and feasible to complete over 4 weeks, although completion rates varied. Of the 3 young adults with OUD participants who began the AWARE intervention, only 2 completed over 90% of the 28 surveys sent. The third participant completed less than half of the surveys sent. Participants reported that it took 5 - 10 minutes to complete their daily survey and that AWARE was "easy to use" and "didn't take much time."

> *I think it pretty much everything made sense the way [AWARE] was set up because there wasn't too little questions or too much. It was kind of perfect to just reflect or more or less.* [Participant 1]

Young adults with OUD participants felt that AWARE was offering a type of social support, like "someone checking in on them." One participant mentioned they were able to keep in contact with their parole officer as a result of having a study phone.

Clinic leadership and counselors reported that they felt AWARE presented a relevant, new way to engage young adults on a daily basis, in addition to weekly meetings. They also mentioned that AWARE allowed for anonymous reflection on the part of their clients. Counselors also appreciated that AWARE enabled weekly discussion of stressors based on patient-generated, real-time data.

> *[You] sent me a summary of my client's information, and I thought that was really helpful. I can actually bring this up in session and see what exactly going on. Because usually in session, it's kind of hard to pull that out of him or it's the generic, "how are you doing," "how's your week been," and "it's fine, fine."* [Staff 4]

Treatment center staff also identified AWARE as an intervention that provides check-ins with their clients, addressing the scarcity of relationships in their clients' lives. AWARE extended the reach of a positive relationship at the treatment center through automated messaging and required response.

> *And I think it's great because it's daily, because in a way it gives them something to feel a part of. Like somebody cares, there's a regular check-in. A lot of them don't have anything, don't have anybody asking*

XSL•FO
RenderX

them anything. Don't have any conversations, nobody checking in. [Staff 2]

Discussion

Principal Findings

This is the first study to develop a patient-centered digital health care technology specifically for young adults in OUD treatment to promote retention through client-counselor engagement with EMA data. This pilot study sought to explore the use of AWARE among young adults with OUD and counselors to address the barriers and facilitators to treatment engagement. We aimed to identify barriers to treatment engagement and assess the potential of AWARE to address these barriers through real-time, technology-based support. Although preliminary, our findings provide valuable insights that lay the groundwork for a future clinical trial designed to rigorously evaluate the effectiveness of AWARE in improving treatment engagement and outcomes for this hard-to-reach population.

Initial findings from this pilot project indicate that delivery of daily surveys to young adult participants and the delivery of weekly summaries to counselors regarding stressors is not only possible but relevant to all involved. Previous research has found that people with OUD want a more personalized approach to treatment [48]. Methadone treatment delivery is largely unchanged since initial federal regulations and policies were established in 1972 [49]. New delivery methods are needed to promote treatment retention among young adults who demonstrate the highest opioid use rate per capita and the highest treatment drop-out rates [50-52]. To engage in treatment, young adults and others need treatment to be personalized and relevant to their circumstances (grounded in real-world experiences) [48]. Previous mHealth research for people with OUD has not specifically targeted young adults who are hard to reach and underrepresented in treatment [53-55]. Future research should focus on engaging out-of-treatment populations of young adults with new delivery innovations.

Young adult participants with OUD identified several stressors that may interfere with treatment engagement, including housing, transportation, drug use, and family obligations. Improved self-management in response to stress, craving, and negative mood is particularly crucial for young adults who are in a stage of exploration and identity formation and who often experience high levels of impulsivity in response to external and internal triggers [56,57]. Two young adults with OUD participants identified "not having much to do" and not having any social connections outside of the clinic as frequent stressors. Among individuals with OUD, loneliness has been recognized as a factor that drives substance use and increases cravings, and it is linked to significant triggers that can lead to relapse [58]. AWARE was able to reach out to participants on a daily basis and make them feel supported and less alone. It is possible that more contacts with positive relationships, through an intervention like AWARE, could not only increase treatment engagement but also begin to address social isolation.

Interventions that integrate a response to patient-generated data within existing substance use disorder treatment are more effective than self-monitoring alone at improving treatment retention [59]. However, patient-generated data has not been integrated within a direct point-of-care system workflow to inform treatment response to address drop-out risk factors at an opioid treatment center. AWARE presents an opportunity to integrate real-time patient-generated data at the point of care, which has the potential for significant impact by informing risk assessment and increasing opportunities for patient engagement and shared decision-making regarding treatment options. Such an approach promises to increase early treatment retention and thereby improve the quality of care and reduce the risk of overdose death.

Limitations

All participants received the intervention. Future studies will be trials of AWARE and will include control and intervention groups to test the effect of AWARE. The small sample size of this specific study tested the recruitment strategy and highlighted the small number of young adults in treatment for OUD at a large, urban opioid treatment center at one point in time. As a result, findings are likely influenced by selection bias, particularly toward young adults who were more motivated to engage in research. The results may not reflect the broader population of young adults in OUD treatment. In addition, due to the limited sample size, thematic saturation was not achieved in qualitative interviews; findings should be interpreted as preliminary and exploratory, intended to inform future refinement and testing of the intervention.

Conclusion

Overall, young adults with OUD and treatment staff participants were able to engage in AWARE in a busy OTP environment, and participants found it valuable. By addressing common stressors and providing a sense of social connection, AWARE may help fill a gap in support between counseling sessions for a hard-to-reach and underrepresented population in methadone treatment. Further research is needed to refine the measures and methods of AWARE and evaluate its effectiveness.

Acknowledgments

This work was supported by the Friends Research Institute Pilot Award. KA was supported by the National Institute of Drug Abuse (R01DA057608-01).

Informed written consent was obtained from all young adult participants in the study. Informed verbal consent was obtained from all treatment providers. The authors obtained permission to publish from the City of Philadelphia.

XSL•FO
RenderX

Data Availability

The deidentified data sets generated and analyzed during this study are not publicly available due to privacy and confidentiality considerations but are available from the corresponding author on reasonable request.

Authors' Contributions

KA conceived and designed the study, developed the methodology, and conducted the formal analysis. MS carried out the investigation, including material preparation and data collection, and contributed to data curation. KA drafted the initial version of the manuscript, and all authors contributed to reviewing and editing subsequent drafts. KA provided overall supervision of the project. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Singh GK, Kim IE, Girmay M, et al. Opioid epidemic in the United States: empirical trends, and a literature review of social determinants and epidemiological, pain management, and treatment patterns. Int J MCH AIDS 2019;8(2):89-100. [doi: 10.21106/ijma.284] [Medline: 31723479]
2. Choi S, Yerneni R, Healy S, Goyal M, Neighbors CJ. Predictors of medication utilization for opioid use disorder among medicaid-insured HIV patients in New York. Am J Addict 2020 Mar;29(2):151-154. [doi: 10.1111/ajad.12998] [Medline: 31951083]
3. Hadland SE, Wharam JF, Schuster MA, Zhang F, Samet JH, Larochelle MR. Trends in receipt of buprenorphine and naltrexone for opioid use disorder among adolescents and young adults, 2001-2014. JAMA Pediatr 2017 Aug 1;171(8):747-755. [doi: 10.1001/jamapediatrics.2017.0745] [Medline: 28628701]
4. Alexander K, Reed MK, Sterling RC. The interaction of race and age in methadone treatment retention outcomes: a single-center analysis. J Subst Use Addict Treat 2023 May;148:209020. [doi: 10.1016/j.josat.2023.209020] [Medline: 36933661]
5. Dugosh K, Abraham A, Seymour B, McLoyd K, Chalk M, Festinger D. A systematic review on the use of psychosocial interventions in conjunction with medications for the treatment of opioid addiction. J Addict Med 2016;10(2):93-103. [doi: 10.1097/ADM.0000000000000193] [Medline: 26808307]
6. Krawczyk N, Williams AR, Saloner B, Cerdá M. Who stays in medication treatment for opioid use disorder? A national study of outpatient specialty treatment settings. J Subst Abuse Treat 2021 Jul;126:108329. [doi: 10.1016/j.jsat.2021.108329] [Medline: 34116820]
7. Lin LA, Lofwall MR, Walsh SL, Knudsen HK. Perceived need and availability of psychosocial interventions across buprenorphine prescriber specialties. Addict Behav 2019 Jun;93:72-77. [doi: 10.1016/j.addbeh.2019.01.023] [Medline: 30690416]
8. Martínez-Vispo C, Martínez Ú, López-Durán A, Fernández Del Río E, Becoña E. Effects of behavioural activation on substance use and depression: a systematic review. Subst Abuse Treat Prev Policy 2018 Sep 29;13(1):36. [doi: 10.1186/s13011-018-0173-2] [Medline: 30268136]
9. Burgess-Hull A, Epstein DH. Ambulatory assessment methods to examine momentary state-based predictors of opioid use behaviors. Curr Addict Rep 2021;8(1):122-135. [doi: 10.1007/s40429-020-00351-7] [Medline: 33425652]
10. McDevitt-Murphy ME, Luciano MT, Zakarian RJ. Use of ecological momentary assessment and intervention in treatment with adults. Focus (Am Psychiatr Publ) 2018;16(4):370-375. [doi: 10.1176/appi.focus.20180017] [Medline: 31191181]
11. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol 2008;4:1-32. [doi: 10.1146/annurev.clinpsy.3.022806.091415] [Medline: 18509902]
12. Alexander K, Sanjuan P, Terplan M. The use of ecological momentary assessment methods with people receiving medication for opioid use disorder: a systematic review. Curr Addict Rep 2023;10(3):366-377. [doi: 10.1007/s40429-023-00492-5]
13. Miranda R, Treloar Padovano H, Gray JC, Wemm SE, Blanchard A. Real-time assessment of alcohol craving and naltrexone treatment responsiveness in a randomized clinical trial. Addict Behav 2018 Aug;83:72-78. [doi: 10.1016/j.addbeh.2018.01.009] [Medline: 29395188]
14. Burgess-Hull AJ, Smith KE, Schriefer D, Panlilio LV, Epstein DH, Preston KL. Longitudinal patterns of momentary stress during outpatient opioid agonist treatment: a growth-mixture-model approach to classifying patients. Drug Alcohol Depend 2021 Sep 1;226:108884. [doi: 10.1016/j.drugalcdep.2021.108884] [Medline: 34229153]
15. Chandra S, Scharf D, Shiffman S. Within-day temporal patterns of smoking, withdrawal symptoms, and craving. Drug Alcohol Depend 2011 Sep 1;117(2-3):118-125. [doi: 10.1016/j.drugalcdep.2010.12.027] [Medline: 21324611]
16. Kennedy AP, Epstein DH, Phillips KA, Preston KL. Sex differences in cocaine/heroin users: drug-use triggers and craving in daily life. Drug Alcohol Depend 2013 Sep 1;132(1-2):29-37. [doi: 10.1016/j.drugalcdep.2012.12.025] [Medline: 23357742]

XSL•FO
RenderX

17. Panlilio LV, Stull SW, Bertz JW, et al. Beyond abstinence and relapse II: momentary relationships between stress, craving, and lapse within clusters of patients with similar patterns of drug use. Psychopharmacology (Berl) 2021 Jun;238(6):1513-1529. [doi: 10.1007/s00213-021-05782-2] [Medline: 33558983]
18. Preston KL, Schroeder JR, Kowalczyk WJ, et al. End-of-day reports of daily hassles and stress in men and women with opioid-use disorder: relationship to momentary reports of opioid and cocaine use and stress. Drug Alcohol Depend 2018 Dec 1;193:21-28. [doi: 10.1016/j.drugalcdep.2018.08.023] [Medline: 30336389]
19. Moran LM, Kowalczyk WJ, Phillips KA, et al. Sex differences in daily life stress and craving in opioid-dependent patients. Am J Drug Alcohol Abuse 2018;44(5):512-523. [doi: 10.1080/00952990.2018.1454934] [Medline: 29641291]
20. Panlilio LV, Stull SW, Kowalczyk WJ, et al. Stress, craving and mood as predictors of early dropout from opioid agonist therapy. Drug Alcohol Depend 2019 Sep 1;202:200-208. [doi: 10.1016/j.drugalcdep.2019.05.026] [Medline: 31357121]
21. Hendershot CS, Witkiewitz K, George WH, Marlatt GA. Relapse prevention for addictive behaviors. Subst Abuse Treat Prev Policy 2011 Jul 19;6:17. [doi: 10.1186/1747-597X-6-17] [Medline: 21771314]
22. Larimer ME, Palmer RS, Marlatt GA. Relapse prevention. An overview of Marlatt's cognitive-behavioral model. Alcohol Res Health 1999;23(2):151-160. [Medline: 10890810]
23. Tiffany ST. A cognitive model of drug urges and drug-use behavior: role of automatic and nonautomatic processes. Psychol Rev 1990 Apr;97(2):147-168. [doi: 10.1037/0033-295x.97.2.147] [Medline: 2186423]
24. Karoly P. Mechanisms of self-regulation: a systems view. Annu Rev Psychol 1993 Jan;44(1):23-52. [doi: 10.1146/annurev.ps.44.020193.000323]
25. Tofighi B, Badiei B, Badolato R, et al. Integrating text messaging in a low threshold telebuprenorphine program for New York City residents with opioid use disorder during COVID-19: a pilot randomized controlled trial. J Addict Med 2023;17(5):e281-e286. [doi: 10.1097/ADM.0000000000001161] [Medline: 37788603]
26. Monico LB, Eastlick M, Michero D, Pielsticker P, Glasner S. Feasibility and acceptability of a novel digital therapeutic combining behavioral and pharmacological treatment for opioid use disorder. Digit Health 2024;10:20552076241258400. [doi: 10.1177/20552076241258400] [Medline: 38812851]
27. Marsch LA, Guarino H, Acosta M, et al. Web-based behavioral treatment for substance use disorders as a partial replacement of standard methadone maintenance treatment. J Subst Abuse Treat 2014 Jan;46(1):43-51. [doi: 10.1016/j.jsat.2013.08.012] [Medline: 24060350]
28. Ashrafioun L, Allan NP, Stecker TA. A randomized clinical trial to assess feasibility, acceptability and preliminary effects of telehealth-delivered cognitive-behavioral therapy for perceived social isolation among individuals with opioid use disorders. Drug Alcohol Depend 2024 May 1;258:111268. [doi: 10.1016/j.drugalcdep.2024.111268] [Medline: 38520791]
29. Carroll KM, Kiluk BD, Nich C, et al. Computer-assisted delivery of cognitive-behavioral therapy: efficacy and durability of CBT4CBT among cocaine-dependent individuals maintained on methadone. Am J Psychiatry 2014 Apr;171(4):436-444. [doi: 10.1176/appi.ajp.2013.13070987] [Medline: 24577287]
30. Williams AR, Nunes EV, Bisaga A, et al. Developing an opioid use disorder treatment cascade: a review of quality measures. J Subst Abuse Treat 2018 Aug;91:57-68. [doi: 10.1016/j.jsat.2018.06.001] [Medline: 29910015]
31. Sheeran P, Maki A, Montanaro E, et al. The impact of changing attitudes, norms, and self-efficacy on health-related intentions and behavior: a meta-analysis. Health Psychol 2016 Nov;35(11):1178-1188. [doi: 10.1037/hea0000387] [Medline: 27280365]
32. Barta WD, Kurth ME, Stein MD, Tennen H, Kiene SM. Craving and self-efficacy in the first five weeks of methadone maintenance therapy: a daily process study. J Stud Alcohol Drugs 2009 Sep;70(5):735-740. [doi: 10.15288/jsad.2009.70.735] [Medline: 19737498]
33. Gass JC, Funderburk JS, Shepardson R, Kosiba JD, Rodriguez L, Maisto SA. The use and impact of self-monitoring on substance use outcomes: a descriptive systematic review. Subst Abus 2021;42(4):512-526. [doi: 10.1080/08897077.2021.1874595] [Medline: 33617740]
34. Swendeman D, Ramanathan N, Baetscher L, et al. Smartphone self-monitoring to support self-management among people living with HIV: perceived benefits and theory of change from a mixed-methods randomized pilot study. J Acquir Immune Defic Syndr 2015 May 1;69 Suppl 1(1):S80-S91. [doi: 10.1097/QAI.0000000000000570] [Medline: 25867783]
35. Svendsen TS, Bjornestad J, Slyngstad TE, et al. "Becoming myself": how participants in a longitudinal substance use disorder recovery study experienced receiving continuous feedback on their results. Subst Abuse Treat Prev Policy 2020 Jan 23;15(1):8. [doi: 10.1186/s13011-020-0254-x] [Medline: 31973763]
36. Dunbar-Jacob J, Mortimer-Stephens MK. Treatment adherence in chronic disease. J Clin Epidemiol 2001 Dec;54 Suppl 1:S57-S60. [doi: 10.1016/s0895-4356(01)00457-7] [Medline: 11750211]
37. Carver CS, Scheier MF, Weintraub JK. Assessing coping strategies: a theoretically based approach. J Pers Soc Psychol 1989;56(2):267-283. [doi: 10.1037/0022-3514.56.2.267] [Medline: 2926629]
38. McKay JR. Making the hard work of recovery more attractive for those with substance use disorders. Addiction 2017 May;112(5):751-757. [doi: 10.1111/add.13502] [Medline: 27535787]
39. Carroll KM. Therapy Manuals for Drug Addiction, Manual 1: A Cognitive-Behavioral Approach: Treating Cocaine Addiction: National Institute on Drug Abuse; 1998:55-65.

XSL•FO
RenderX

40. Preston KL, Kowalczyk WJ, Phillips KA, et al. Before and after: craving, mood, and background stress in the hours surrounding drug use and stressful events in patients with opioid-use disorder. Psychopharmacology (Berl) 2018 Sep;235(9):2713-2723. [doi: 10.1007/s00213-018-4966-9] [Medline: 29980821]
41. Preston KL, Kowalczyk WJ, Phillips KA, et al. Context and craving during stressful events in the daily lives of drug-dependent patients. Psychopharmacology (Berl) 2017 Sep;234(17):2631-2642. [doi: 10.1007/s00213-017-4663-0] [Medline: 28593441]
42. Preston KL, Epstein DH. Stress in the daily lives of cocaine and heroin users: relationship to mood, craving, relapse triggers, and cocaine use. Psychopharmacology (Berl) 2011 Nov;218(1):29-37. [doi: 10.1007/s00213-011-2183-x] [Medline: 21336579]
43. Kowalczyk WJ, Moran LM, Bertz JW, et al. Using ecological momentary assessment to examine the relationship between craving and affect with opioid use in a clinical trial of clonidine as an adjunct medication to buprenorphine treatment. Am J Drug Alcohol Abuse 2018;44(5):502-511. [doi: 10.1080/00952990.2018.1454933] [Medline: 29634425]
44. Moran LM, Phillips KA, Kowalczyk WJ, et al. Aripiprazole for cocaine abstinence: a randomized-controlled trial with ecological momentary assessment. Behav Pharmacol 2017 Feb;28(1):63-73. [doi: 10.1097/FBP.0000000000000268] [Medline: 27755017]
45. Preston KL, Vahabzadeh M, Schmittner J, Lin JL, Gorelick DA, Epstein DH. Cocaine craving and use during daily life. Psychopharmacology (Berl) 2009 Dec;207(2):291-301. [doi: 10.1007/s00213-009-1655-8] [Medline: 19777216]
46. Preston KL, Jobes ML, Phillips KA, Epstein DH. Real-time assessment of alcohol drinking and drug use in opioid-dependent polydrug users. Behav Pharmacol 2016 Oct;27(7):579-584. [doi: 10.1097/FBP.0000000000000250] [Medline: 27579810]
47. Watkins DC. Rapid and Rigorous Qualitative Data Analysis: The “RADaR” technique for applied research. Int J Qual Methods 2017;16(1):1609406917712131. [doi: 10.1177/1609406917712131]
48. Kourounis G, Richards BDW, Kyprianou E, Symeonidou E, Malliori MM, Samartzis L. Opioid substitution therapy: lowering the treatment thresholds. Drug Alcohol Depend 2016 Apr 1;161:1-8. [doi: 10.1016/j.drugalcdep.2015.12.021] [Medline: 26832931]
49. Schwartz RP, Kelly SM, Mitchell SG, et al. Patient-centered methadone treatment: a randomized clinical trial. Addiction 2017 Mar;112(3):454-464. [doi: 10.1111/add.13622] [Medline: 27661788]
50. Bagley SM, Chavez L, Braciszewski JM, et al. Receipt of medications for opioid use disorder among youth engaged in primary care: data from 6 health systems. Addict Sci Clin Pract 2021 Jul 7;16(1):46. [doi: 10.1186/s13722-021-00249-3] [Medline: 34233750]
51. Chang DC, Klimas J, Wood E, Fairbairn N. Medication-assisted treatment for youth with opioid use disorder: current dilemmas and remaining questions. Am J Drug Alcohol Abuse 2018;44(2):143-146. [doi: 10.1080/00952990.2017.1399403] [Medline: 29190156]
52. Fishman M, Wenzel K, Scodes J, et al. Young adults have worse outcomes than older adults: secondary analysis of a medication trial for opioid use disorder. J Adolesc Health 2020 Dec;67(6):778-785. [doi: 10.1016/j.jadohealth.2020.07.038] [Medline: 32873500]
53. Gustafson DH, Landucci G, Vjorn OJ, et al. Effects of bundling medication for opioid use disorder with an mHealth intervention targeting addiction: a randomized clinical trial. Am J Psychiatry 2024 Feb 1;181(2):115-124. [doi: 10.1176/appi.ajp.20230055] [Medline: 37789744]
54. King VL, Siegel G, Priesmeyer HR, Siegel LH, Potter JS. Development and evaluation of a digital app for patient self-management of opioid use disorder: usability, acceptability, and utility study. JMIR Form Res 2024 Apr 1;8:e48068. [doi: 10.2196/48068] [Medline: 38557501]
55. Nuamah J, Mehta R, Sasangohar F. Technologies for opioid use disorder management: mobile app search and scoping review. JMIR Mhealth Uhealth 2020 Jun 5;8(6):e15752. [doi: 10.2196/15752] [Medline: 32501273]
56. Schuman-Olivier Z, Weiss RD, Hoeppner BB, Borodovsky J, Albanese MJ. Emerging adult age status predicts poor buprenorphine treatment retention. J Subst Abuse Treat 2014 Sep;47(3):202-212. [doi: 10.1016/j.jsat.2014.04.006] [Medline: 24953168]
57. Reynolds BW, Basso MR, Miller AK, Whiteside DM, Combs D. Executive function, impulsivity, and risky behaviors in young adults. Neuropsychology 2019 Feb;33(2):212-221. [doi: 10.1037/neu0000510] [Medline: 30589284]
58. Herczyk JM, Zullig KJ, Davis SM, et al. Association of loneliness and mindfulness in substance use treatment retention. Int J Environ Res Public Health 2023 Aug 13;20(16):6571. [doi: 10.3390/ijerph20166571] [Medline: 37623157]
59. Reback CJ, Rünger D, Fletcher JB, Swendeman D. Ecological momentary assessments for self-monitoring and counseling to optimize methamphetamine treatment and sexual risk reduction outcomes among gay and bisexual men. J Subst Abuse Treat 2018 Sep;92:17-26. [doi: 10.1016/j.jsat.2018.06.005] [Medline: 30032940]

Abbreviations

AWARE: Awareness and Response to the Environment
CBT: cognitive behavioral therapy
COREQ: Consolidated Criteria for Reporting Qualitative Research
EHR: electronic health record

XSL•FO
RenderX

EMA: ecological momentary assessment
HIPAA: Health Insurance Portability and Accountability Act
mHealth: mobile health
NIDA: National Institute on Drug Abuse
OTP: opioid treatment program
OUD: opioid use disorder
RADaR: rigorous and accelerated data reduction
REDCap: Research Electronic Data Capture

Edited by A Mavragani; submitted 10.10.24; peer-reviewed by K Lunze, K Dainty, K Yoshiuchi; revised version received 04.09.25; accepted 05.09.25; published 23.09.25.

Please cite as:
Alexander K, Scialanca M
The Development of a Patient-Centered Digital Health Care Technology for Young Adults in Opioid Use Disorder Treatment: Qualitative Study
JMIR Form Res 2025;9:e67401
URL: https://formative.jmir.org/2025/1/e67401
doi:10.2196/67401

© Karen Alexander, Madison Scialanca. Originally published in JMIR Formative Research (https://formative.jmir.org), 23.9.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Feasibility of a Randomized Controlled Trial of Large AI-Based Linguistic Models for Clinical Reasoning Training of Physical Therapy Students: Pilot Randomized Parallel-Group Study

Raúl Ferrer-Peña[1,2]; Silvia Di-Bonaventura[1,3]; Alberto Pérez-González[1,2]; Alfredo Lerín-Calvo[1,2,4]

[1]Grupo de Investigación Clínico-Docente sobre Ciencias de la Rehabilitación (INDOCLIN), Centro Superior de Estudios Universitarios La Salle, Universidad Autónoma de Madrid, La Salle 10, Aravaca, Madrid, Spain

[2]Departamento de Fisioterapia, Centro Superior de Estudios Universitarios La Salle, Aravaca, Madrid, Spain

[3]Department of Physical Therapy, Occupational Therapy, Rehabilitation and Physical Medicine, Rey Juan Carlos University, Alcorcón, Spain

[4]Grupo de Investigación de Neurociencias Aplicadas a la Neurorehabilitación (GINARE), Neuron, Madrid, Spain

Corresponding Author:
Alfredo Lerín-Calvo
Grupo de Investigación Clínico-Docente sobre Ciencias de la Rehabilitación (INDOCLIN), Centro Superior de Estudios Universitarios La Salle, Universidad Autónoma de Madrid, La Salle 10, Aravaca, Madrid, Spain

Abstract

Background: Clinical reasoning is a critical skill for physical therapists, involving the collection and interpretation of patient information to form accurate diagnoses. Traditional training often lacks the diversity of clinical cases necessary for students to develop these skills comprehensively. Large language models (LLMs) like GPT-4 have the potential to simulate a wide range of clinical scenarios, offering a novel approach to enhance clinical reasoning in physical therapy education.

Objective: The aim of the study is to explore the main barriers and facilitators that could be encountered in conducting a randomized clinical trial to study the effectiveness of the implementation of LLM models as tools to work on the clinical reasoning of physical therapy students.

Methods: This pilot randomized parallel-group study involved 46 third-year physical therapy students at La Salle Centre for Higher University Studies. Participants were randomly assigned to either the experimental group, which received LLM training, or the control group, which followed the usual curriculum. The intervention lasted for 4 weeks, during which the experimental group used LLM to solve weekly clinical cases. Digital competencies, satisfaction, and costs were evaluated to explore the feasibility of this intervention.

Results: The recruitment and participation rates were high, but active engagement with the LLM was low, with only 5.75% (5/23) of the experimental group actively using the model. No significant difference in overall satisfaction was found between the groups, and the cost analysis reflected an initial cost of US $1738 for completing the study.

Conclusions: While LLMs have the potential to enhance specific digital competencies in physical therapy students, their practical integration into the curriculum faces challenges. Future studies should focus on improving student engagement with LLMs and extending the training period to determine the feasibility of integrating this tool into physical therapy education and maximize benefits.

Trial Registration: ClinicalTrials.gov NCT 06809634; https://tinyurl.com/48nf3zks

(JMIR Form Res 2025;9:e66126) doi:10.2196/66126

KEYWORDS

clinical reasoning; physical therapy education; artificial intelligence; pilot study; feasibility; physical therapy; randomized controlled trial; artificial intelligence-based; linguistic model; physical therapy student; critical skill; large language models; LLM; barriers; facilitator; effectiveness; implementation; physiotherapy

Introduction

Clinical reasoning, a vital skill in physical therapist training, is essential to ensure safe and efficacious practice [1]. This multifaceted process relies on the collection and interpretation of patient information to formulate an accurate diagnostic hypothesis [2]. A crucial part of this process is instant patient feedback, which can help refine hypotheses and tailor interventions according to the patient's individual needs [3].

However, a significant limitation in the training of physical therapy students is the lack of access to a sufficient variety of clinical cases during their internships. Students often encounter

XSL•FO
RenderX

a limited range of clinical conditions during their training, which can limit their ability to develop clinical reasoning skills in a wide range of settings [4]. This lack of diversity in clinical experience can be especially problematic given that physical therapists often encounter a wide range of conditions in their subsequent practice [5]. In Spain, the physical therapy degree follows a competency-based model in which students are expected to progressively develop clinical reasoning skills throughout their education. However, these skills are not always formally assessed beyond traditional written exams, which may not adequately reflect real-world decision-making.

To address these challenges, different pedagogical approaches have been implemented in physical therapy education to enhance clinical reasoning skills. Some of the most widely used methodologies include problem-based learning, case-based discussions, and simulated patient encounters [6]. While these strategies have shown positive results in improving students' decision-making skills, their implementation often requires significant resources, such as trained facilitators or access to standardized patients, making them challenging to scale [7]. This highlights the need for alternative approaches that can provide physical therapy students with diverse clinical scenarios to enhance their clinical reasoning in a more accessible and structured manner.

One such alternative is the use of artificial intelligence (AI)–based large language models (LLMs), such as GPT-4, which have emerged as promising tools for improving this critical skill in physical therapy students [8].

LLMs can simulate realistic and online clinical cases, allowing students to engage in patient scenarios without the logistical constraints of traditional training methods. Previous studies have shown that LLMs can improve diagnostic accuracy, clinical decision-making, and knowledge retention, particularly in medical and nursing students [9,10].

Despite this potential, the use of LLMs in physical therapy education remains underexplored, and it is unclear whether they can provide similar benefits for clinical reasoning in this field. Moreover, to our knowledge, no studies have directly replicated an existing protocol for integrating LLMs into physical therapy education, making it difficult to assess their impact systematically.

In physical therapy education, the assessment of clinical reasoning is conducted through a combination of formative and summative evaluations. Formative evaluation assesses students' skills and attitudes throughout their coursework, while summative evaluation consists of final written and oral-practical tests to measure the acquisition of clinical competencies [11].

Although these methods are effective in measuring theoretical knowledge, they may not fully capture the complexity of real-world clinical reasoning. LLMs could help bridge this gap by providing adaptive, interactive case simulations that allow students to refine their decision-making in a controlled, risk-free environment.

While artificial intelligence has already demonstrated benefits in health care applications such as pathology, where it has improved productivity, diagnostic accuracy, cost-efficiency, and staff satisfaction [12], its application in physical therapy remains largely theoretical. More research is needed to determine whether these models can effectively enhance physical therapy students' clinical reasoning skills in a structured educational setting.

Therefore, the aim of the study is to explore the main barriers and facilitators that could be encountered in conducting a randomized clinical trial to study the effectiveness of the implementation of LLM models as tools to work on the clinical reasoning of physical therapy students.

Methods

Ethical Considerations

This study was approved by the local ethics committee of the La Salle University Center for Advanced Studies, Madrid, Spain (CSEULS-PI-002/2025). This trial was registered in clinicaltrials.org (NCT06809634). This study followed the CONSORT guideline for randomized controlled trials (RCTs) [13]. All information and consent forms were provided in an easy-read format, and written informed consent was obtained from all participants. Participants were assured that their personal data would be kept confidential and only used for the purposes of this study. All data were anonymized and stored in accordance with privacy regulations. Participants were informed of their right to withdraw from the study at any time without any consequences. No compensation was provided for participation, but participants were informed of any potential benefits associated with the study outcomes.

Design

A pilot randomized parallel-group study was conducted in La Salle Centre for Higher University Studies (LCHUS). In total, 46 students from different third-year physical therapy classes participated in the study. These 2 classes were randomly assigned to the experimental or control group using a simple randomization after signing the informed consent by an external researcher who did not participate in the study. Participants in the experimental group received the LLM training in addition to attending course lessons, while the participants in the control group received only the course lessons. The study was carried out between September 2023 and January 2024.

In the context in which this study was carried out (Spain), physical therapists have the capacity to carry out evaluations, formulate functional diagnoses and establish treatment plans without the need for a medical referral in most cases, except in those pathologies that require previous medical intervention, as established in Order CIN 2135/2008 [14]. In this study, students were trained to apply this professional framework in a simulated environment, where they had to [1] formulate diagnostic hypotheses based on the anamnesis and clinical examination of the online patient [2], make a physiotherapeutic diagnosis in line with the WHO's (World Health Organization) International Classification of Functioning, Disability and Health (ICF) and [3] design a treatment plan according to the patient's functional assessment of the patient. This approach sought to replicate the physical therapist's clinical reasoning process in a controlled context, favoring the development of critical skills

XSL•FO
RenderX

in clinical decision-making. The trial follows the CONSORT (Consolidated Standards of Reporting Trials) guidelines (the CONSORT checklist is provided in Checklist 1).

Development of the Theoretical Framework of Learning

The physical therapist's program at the LCHUS (Autonomous University of Madrid) follows a clinical reasoning approach based on hypothetical-deductive theory and decision-making in real and simulated clinical environments [2,3]. This model integrates 2 main approaches: 1, the hypothetical-deductive model, in which students generate diagnostic hypotheses based on the medical history and clinical examination, contrasting them with new tests to adjust their reasoning [1]; 2, narrative reasoning, which emphasizes the patient's history and their biopsychosocial context for individual-centered decision-making [15]; and 3, reflective reasoning and metacognition, which encourages self-assessment of clinical thinking through tutorials and structured feedback [10]. The acquisition and consolidation of clinical reasoning in the program is evaluated through a system of formative and summative assessment. Formative assessment is based on observation and feedback during face-to-face and simulated practical, written reflection in learning portfolios and classroom discussion of clinical cases. On the other hand, summative assessment includes structured practical exams, where students must justify their diagnostic and therapeutic reasoning in standardized clinical scenarios, as well as written evaluations of clinical cases and self-assessments using tools such as the Lasater Clinical Judgment Rubric [10].

Sample Size

This was a feasibility study, so a formal sample size calculation was not required. In total, 46 participants were recruited for the study and allocated in equal numbers in the control and experimental groups.

Blinding

Students were not blinded to the group allocation because they knew if they had to work with or without using the LLM. Therefore, the person applying the treatment (in this case the lessons) was aware of the group to which each of the participants had been assigned. Assessors were blinded to the assignment of the participants.

Interventions

After being randomized, for those students belonging to the experimental group (LLM Group), a personal LLM ChatGPT account in version 3.5 was generated for them for a period of 1 month. Using this account, the participants solved a total of 4 clinical cases for 4 weeks, one per week, in which the LLM will serve as a virtual patient, answering the questions that the student asked and based on a physical therapy diagnosis, participants proposed a treatment for the virtual patient. It is important to note that students were able to perform the same clinical case as many times as they wanted, which would be supervised for the researchers after ending the study. No formal training was provided to students on how to use the LLM before the intervention. This decision was intentional, as the aim was to observe how students engaged with the tool naturally, identifying potential barriers and facilitators without prior guidance influencing their interaction.

The control group, on the other hand, carried out the usual practices prescribed by the LCHUS, in which, during the same 4 weeks, they participated in the physical therapy sessions carried out by their practice tutor, intervening with him in the clinical reasoning of the patients treated. In turn, they carried out the same cases as in the experimental group, in written format, which they had to resolve in the same way with the practice tutor.

At the end of the 4 weeks, the students in the experimental and control groups were evaluated with a final clinical case study in written format.

For the participants in the experimental group to execute the clinical cases, they had to follow a procedure and write predefined prompts in the LLM, which were provided to them via a PDF file. The procedure for the execution of the clinical cases and the prompts entered at the beginning of the cases are shown in Figure 1. Students who did not complete the required four weeks of LLM clinical scenario cases were not included in the final analysis, as their incomplete participation could bias the results. However, their academic performance in the course was assessed through the standard evaluation methods established by the institution, ensuring that all students met the competency requirements of the physical therapy curriculum.

The clinical cases presented in the LLM consisted of a person with low back pain with facet characteristics, a patient with knee pain with characteristics related to chondromalacia patella, a patient with rheumatoid arthritis, and a person presenting to the clinic for postsurgical rehabilitation of a hip prosthesis. An example of the prompts used in the cases and the expected interactions can be found in Multimedia Appendix 1.

XSL•FO
RenderX

Figure 1. Predefined prompts for clinical cases. Figure developed by the authors based on previous literature on artificial intelligence in physical therapy and medical education [8,16]. IAS: Institutional Academic System.

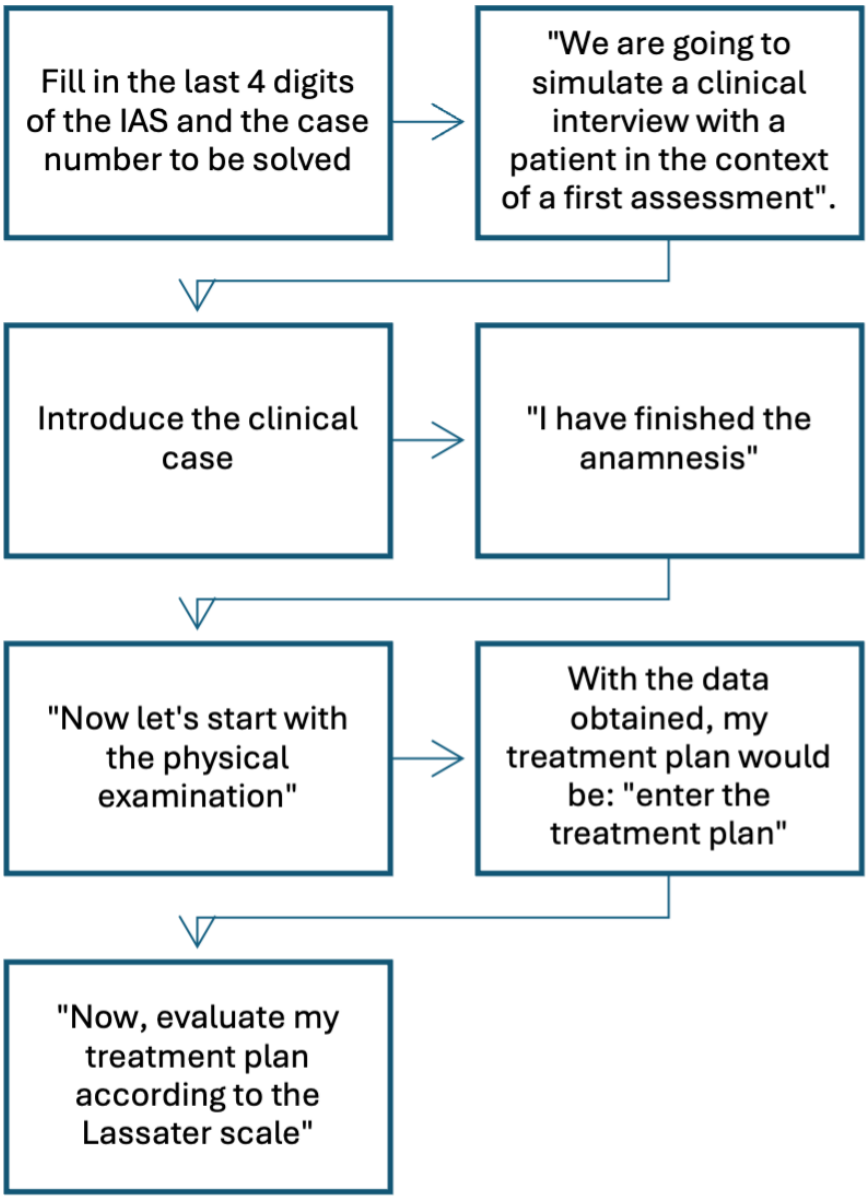

Outcomes

Recruitment and Participation

Study recruitment and retention data include the number of students who expressed interest to take part in the study, participants' enrollment, and participants completing treatment.

Cost Analysis

The cost analysis of the intervention was calculated based on the cost of the ChatGPT access plan from OpenAI, which provides ChatGPT version 4.0 rather than the free ChatGPT 3.5 account. To accurately capture the effective cost, we also considered the teaching period during which students had access to the program, in alignment with the course calendar.

However, in order to reflect the practical realities when evaluating the feasibility of integrating LLM into teaching practice, it is necessary to consider other associated costs, both direct and indirect. Among the most significant is the need to train instructors who will incorporate this model as part of their teaching methodology. This entails expenses related to instructional hours (eg, introduction to LLM, design and engineering of prompts, and methods for evaluating the information generated), as well as preparation time, rehearsal of the proposed routine or procedure, and the acquisition, comparison, or assessment of the results obtained. Since traditional assessment systems will be modified, it is also important to review interactions with the AI, which may increase workload hours unless automated systems are established to record the data generated.

Likewise, the costs of integrating these models into the academic setting should be assessed, including potential changes to academic guidelines, the provision of necessary resources for students, the required technical support, and relevant training to ensure effective use of the technology. However, the costs associated with integrating these models into teaching practice could reduce future expenses after the initial investment, since

XSL•FO
RenderX

in the short term—and once the model is integrated—direct or indirect benefits are expected to be obtained in the long run.

Suitability of Outcome Measures

The outcome measurements tool used in the feasibility study was an ad hoc questionnaire for digital competencies [17]. A visual analog scale was used to evaluate the satisfaction of participants with the educative intervention.

The choice to include a digital competencies questionnaire was closely tied to our main objective: evaluating how effectively students could adopt and benefit from an LLM-based intervention. Because the experimental group interacted with ChatGPT, it was crucial to determine whether participants possessed (or could develop) the necessary digital skills to navigate and critically assess AI-generated content.

The questionnaire developed by Montero-Delgado et al (2020) [17] assesses the 26 digital competencies for health care professionals, divided into six domains. Other similar ad hoc questionnaires have shown that it is a valid and reliable tool. For example, in the study carried out in patients with type 1 diabetes by Barajas-Galindo et al [18], the questionnaire was shown to be valid, with a Cronbach α coefficient of .77. By measuring these competencies, we aimed to identify whether insufficient digital literacy might hinder the effective use of AI-based tools and to observe any potential improvements in key digital skills that could support clinical reasoning in physical therapy. The Spanish version of this questionnaire is included in Multimedia Appendix 2.

The visual analog satisfaction scale consisted of a 100 mm line on which the participant had to indicate the final satisfaction with the educational program in a range between very dissatisfied (0 mm) and very satisfied (100 mm).

Statistical Analysis

Once the data collection period was over, data analysis was carried out using SPSS statistical software version 29.0 (SPSS Inc). A descriptive analysis of the demographic characteristics and the main and secondary variables in the sample was performed, presenting continuous variables as mean SD, while categorical variables are presented as number (n) and percentage (relative frequency, %). Descriptive statistics on recruitment and retention rates, baseline participants' characteristics, and acceptability of the treatment were reported, and differences between groups were examined with *t* and chi-square tests.

Results

Study Participants

After randomization, 23 participants were assigned to the LLM group, and 23 participants were included in the control group. Out of 23 participants, there were 17 (73.9%) females in the LLM group and 9 (39.1%) females in the control group. Baseline sociodemographic data for both groups are reported in Table 1. Statistically significant differences were found between groups for gender, nationality, and language variables.

XSL•FO
RenderX

Table . Baseline characteristics of all randomized participants.

Category		Control	Intervention	*P* value	Rosenthal R
Age, mean (SD)		20.48 (1.78)	21.29 (1.3)	.08[a]	0.257
Nationality, n (%)				.005[b]	0.419
	Spanish	11 (47.8)	20 (87.0)		
	French	12 (52.2)	3 (13.0)		
Language, n (%)				.005[b]	0.419
	Spanish	11 (47.8)	20 (87.0)		
	French	12 (52.2)	3 (13.0)		
Gender, n (%)				.02[b]	0.353
	Man	6 (26.1)	14 (60.9)		
	Woman	17 (73.9)	9 (39.1)		
Educational level, n (%)				—[c]	—
	Bachelor	23 (100)	23 (100)		
Ethnic group, n (%)				.54[d]	0.153
	White	11 (47.8)	11 (47.8)		
	Latin	12 (52.2)	10 (43.5)		
	Asian	0 (0)	1 (4.3)		
	Black	0 (0)	1 (4.3)		
Civil status, n (%)				—	—
	Unmarried	23 (100)	23 (100)		
Employment, n (%)				1.000[d]	0.149
	Unemployed	23 (100)	22 (95.7)		
	Employed	0 (0)	1 (4.3)		

[a]Student *t* test.
[b]Chi-square test.
[c]Not applicable.
[d]Fisher exact test.

Recruitment and Participation

All individuals agreed to participate in the study. All participants completed the initial digital competence questionnaire and the visual analog satisfaction scale at the end of the study. However, of the 23 students who were assigned to the ChatGPT group, only 4 (5.75%) of the students actively participated and interacted with the model for the resolution of the clinical cases. These 4 students completed the 4-week intervention and the final clinical case, but no additional students engaged with the LLM (Figure 2).

XSL•FO
RenderX

Figure 2. CONSORT (Consolidated Standards of Reporting Trials) flow diagram of the study.

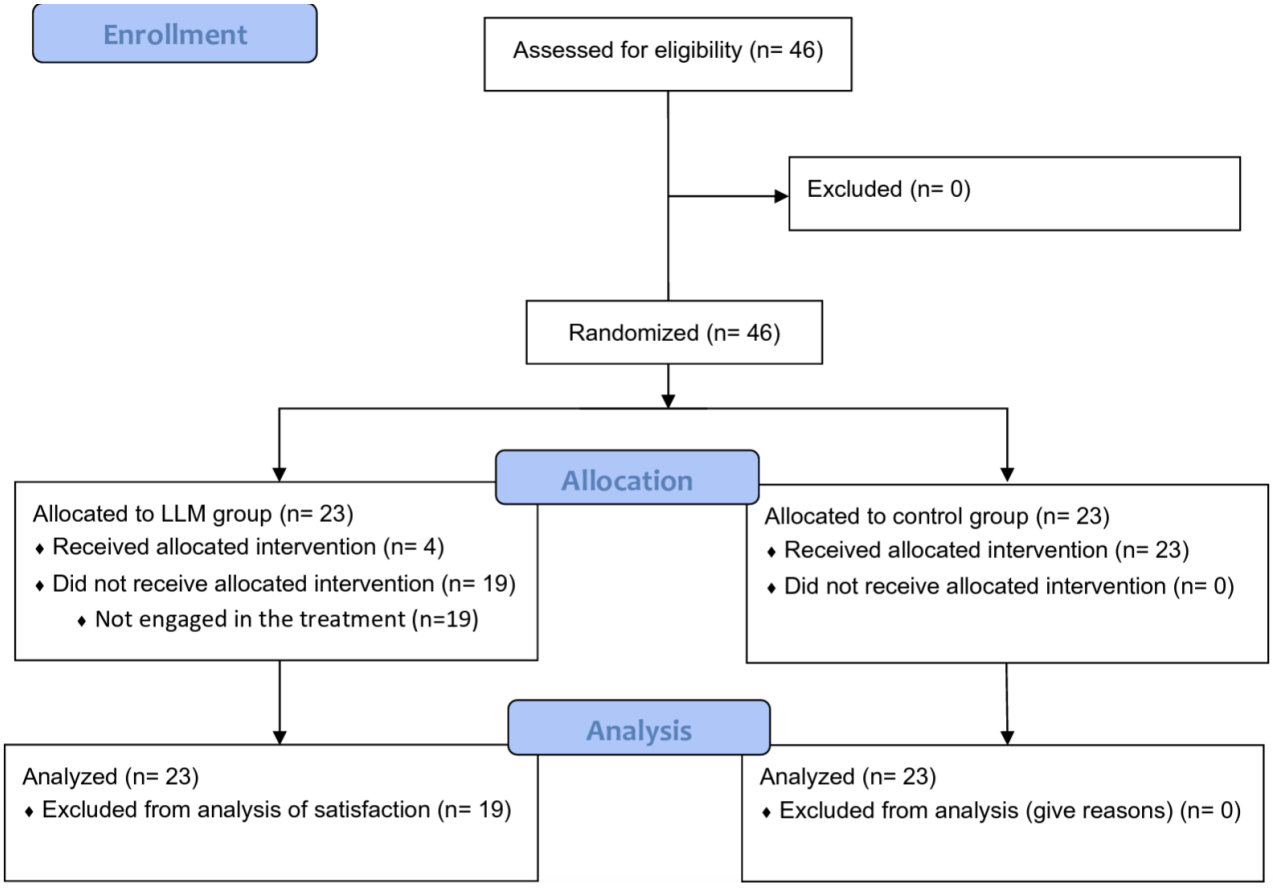

Cost Analysis

The cost analysis of the intervention was calculated based on direct and indirect expenses. Direct costs included the OpenAI ChatGPT Team Plan (version 4.0) at US $26 per user/month. For 23 students over 4 weeks, this totaled US $530. Indirect costs accounted for (1) clinical case preparation: faculty spent 2 hours per 4 cases total designing LLM-compatible scenarios, equivalent to 8 hours at US $30.22/hour (average academic hourly rate), totaling US $60; (2) teacher training: a 4-hour workshop to familiarize educators with LLM tools, costing US $138; and (3) assessment time: the experimental group required additional time to evaluate LLM interaction logs (1 h/student), adding 23 hours (US $794). Total estimated cost: US $1738 (direct+indirect).

Secondary Variables: Suitability of Outcome Measures

In terms of overall satisfaction, no significant differences were observed between the groups (P=.53), with a median of 77.00 (IQR 61.00 - 84.00) for the control group and 77.00 (IQR 66.5 - 85.5) for the intervention group. For this outcome, only four students in the experimental group who completed the trial were included for the analysis.

However, with respect to digital competencies, the item “Data analysis and management” showed a statistically significant difference between the groups, with a median of 6.00 (IQR 5.00 - 6.75) in the control group and 7.00 (IQR 6.33 - 7.30) in the intervention group (P=.042), with the intervention group standing out with higher skills in this area. Table 2 shows the overall data of these variables.

XSL•FO
RenderX

Table . Characteristics of outcome measures.

Variable	Control, median (IQR)	Intervention, median (IQR)	*P* value	Rosenthal R
Satisfaction	77 (61.00-84.00)	77.0 (66.5-85.5)	.71	0.056
Digital health literacy	7.30 (6.50-7.80)	7.30 (6.80-8.00)	.27	0.164
Effective management of scientific and health information	6.67 (5.67-7.00)	7.00 (6.33-7.67)	.15	0.211
Health communication 2.0	6.80 (5.80-7.40)	7.40 (6.60-7.80)	.27	0.162
Creation of scientific-health digital content	6.20 (5.20-7.00)	7.20 (6.40-7.60)	.05	0.286
Collective networking with health teams	6.50 (5.30-7.00)	7.00 (6.33-8.00)	.18	0.198
Data analysis and management	6.00 (5.00-6.75)	7.00 (6.33-7.30)	.04	0.299

Discussion

Principal Findings

The main objective of this study was to assess the feasibility of implementing ChatGPT as an LLM for improving clinical reasoning in physical therapy students. This is the first study that tries to implement the use of LLM for university students, and its findings offer preliminary insights into both the potential benefits and challenges associated with this innovative educational approach. It will serve as a foundation for designing a larger RCT that will assess the effectiveness of LLMs in enhancing clinical reasoning skills in physical therapy students.

The results obtained reflect some barriers for the implementation of this intervention in university centers, which must be addressed in future RCTs. While other studies have shown the effectiveness of ChatGPT for other health careers like ophthalmology [19], our study demonstrated a low active participation rate within the experimental group, with only 5.75% of students actively engaged with the model to solve clinical cases. The frequency of use of this tool for educational purposes in other studies with similar samples has been shown to be higher among students in other health care fields [20]. This suggests that while there is interest, the practical integration of LLMs into the learning routine of physical therapy students may face challenges, such as their comfort with new technologies, their perceived usefulness of the tool, or possible technical issues. One of the barriers identified by other studies is the absence of human interaction that can negatively affect students who prefer a personal connection with their educator, as shown in the study conducted by D'Mello et al [21].

Another issue for implementing LLM in physical therapy students could be the level of digital competencies shown by the participants, as, despite demonstrating a good level in the different items, in none of them was the average score excellent, and probably it may not be at a sufficient level to benefit from the use of this technology in a comfortable way for the learner. In fact, one of the main barriers to the implementation of this technology seems to be this lack of knowledge [16] in the field.

Specifically, the item "data analysis and management" was the one reported with the lowest score by the participants. It can therefore be identified that physical therapy students self-assess themselves as having a poor ability to use health databases and to handle analysis of large amounts of data [17]. While the rest of the health professions are making progress in this recent field of research, it is probably desirable that physical therapy universities change their teaching models so that their students are adapted to these new technologies once they finish their degree, which seems to be an issue in other health care careers [22].

The level of satisfaction with the study was high in both groups, with no significant differences between them. This fact provides a strong background to carry out the full study, since our hypothesis is that the overall satisfaction of students using LLM to improve their clinical reasoning will be higher than those using traditional methods, as we have already observed in other studies in medical students [23,24].

This statement is supported by the theoretical basis that, despite the low level of regular interaction within the LLM group (5/23, 5.75%), there were no differences between the groups. Thus, it could be inferred that low participation does not equate to low satisfaction, given that in both cases it was very similar. Therefore, we can consider two possible scenarios.

On the one hand, it might be that the students who did use the LLM found it so useful that it compensated for the dissatisfaction of those who did not use it. Alternatively, although few used it, those who did had no negative expectations, and the rest remained neutral.

Hence, the premise of the hypothesis arises from the notion of a latent or unexpressed acceptability, wherein—despite the low participation—there was no active rejection of the tool. This suggests that there is no psychological resistance to the use of LLMs as a learning method. There is a clear theoretical potential for students to adopt these tools once the previously mentioned barriers (lack of training, difficulty in understanding and managing interaction-derived data, insufficient usage time, etc) are overcome.

In consequence, we cannot currently assert superiority, nor can we dismiss its feasibility, given that the comparable level of satisfaction indicates that, under optimal conditions, LLMs

XSL•FO
RenderX

could be accepted and integrated as a supplement to traditional methods.

Therefore, the intervention protocol appears to be valid and usable; however, it is important to better train the students in ChatGPT in order to obtain greater adherence in the intervention group. With the aim of improving the methodology, the entire study should be conducted for 8 weeks instead of the 4 weeks initially proposed. In this way, the first 4 weeks of the study would be dedicated to increasing the students' knowledge and skills in the use of ChatGPT, followed by the 4 weeks of intervention. The cost of the study would therefore double from the US $530 initially proposed to a total of €920 (US $1074.24) to carry out the study, increasing the total number of weeks of the study to 8.

One of the main points highlighted in this pilot feasibility study is the low adherence of students to the LLM intervention, which suggests that pedagogical factors can significantly influence the acceptance and use of these tools in physical therapy education. The integration of technologies in the teaching of clinical reasoning requires not only access to the tool, but also adequate pedagogical preparation that fosters active learning and confidence in the use of AI [16]. Previous studies have shown that self-directed, simulation-based learning improves the acquisition of health science skills, provided that students understand their applicability and feel comfortable with the methodology [25]. In this sense, the incorporation of an analysis of the learning styles of the participants before the intervention would allow the introduction of the LLM to be personalized, increasing the adherence and effectiveness of the program. Tools such as Kolb's Learning Style Inventory [26] or the VARK Learning Modality Preference Scale [27] have been successfully used to adapt pedagogical strategies in similar educational settings. As we have previously mentioned, for the future randomized clinical trial, we propose extending the intervention to eight weeks, incorporating an initial period of specific training on the use of LLMs and their application in clinical reasoning, which could improve participation and the impact of the tool in the training of physical therapy students.

Limitations

This study has several limitations. One of the main limitations of this study was the lack of a structured evaluation of the barriers perceived by the students who participated in the experimental group. While the initial objective was to evaluate the feasibility of the intervention, future research should include a formal evaluation of the obstacles encountered by students to improve adherence to the tool. Through informal conversations with participants, factors such as lack of familiarity with AI, preference for human interaction, perception of additional workload, and technological difficulties were identified. In a future RCT, questionnaires and semistructured interviews will be included to identify and mitigate these factors, ensuring that the intervention can be optimally implemented.

Also, the small sample size and short duration limit the generalizability of the findings. The low engagement rate in the experimental group also poses a significant challenge to interpreting the results. In addition, students were not blinded to group allocation, which could introduce bias. Despite these limitations, the study provides valuable preliminary data on the feasibility and potential impact of LLMs in physical therapy education.

Conclusions

This pilot study investigated the feasibility of using LLMs like GPT-4 to enhance clinical reasoning skills in physical therapy students. While recruitment and participation were high, actual engagement with the LLM was low, so strategies for increasing the engagement should be implemented to carry on the complete study.

The cost of providing ChatGPT Premium for four weeks was US $530, raising questions about cost-effectiveness given the low engagement. However, future studies should consider a training period and indirect costs, increasing the total amount to US $1738.

Acknowledgments

We would like to thank all students that participated in this study.

This research was supported by La Salle Centre for Higher University Studies (LCHUS), which covered the costs associated with the publication of this article.

Data Availability

The datasets generated and used in this study are publicly available in Zenodo [28].

Conflicts of Interest

None declared.

Multimedia Appendix 1
Clinical cases example.
[DOCX File, 255 KB - formative_v9i1e66126_app1.docx]

XSL•FO
RenderX

Multimedia Appendix 2
Digital competencies questionnaire in English.
[DOCX File, 20 KB - formative_v9i1e66126_app2.docx]

Checklist 1
CONSORT (Consolidated Standards of Reporting Trials) checklist.
[PDF File, 128 KB - formative_v9i1e66126_app3.pdf]

References

1. Edwards I, Jones M, Carr J, Braunack-Mayer A, Jensen GM. Clinical reasoning strategies in physical therapy. Phys Ther 2004 Apr;84(4):312-330. [Medline: 15049726]
2. Jones MA, Rivett DA. Clinical Reasoning in Musculoskeletal Practice: Elsevier; 2019. URL: https://linkinghub.elsevier.com/retrieve/pii/C20130142675 [accessed 2025-07-10]
3. Higgs J, Jensen GM, Loftus S, Christensen N, editors. Clinical Reasoning in the Health Professions, 4th edition: Elsevier; 2019:511.
4. Longworth MK. An exploration of the perceived factors that affect the learning and transfer of skills taught to student midwives. Midwifery 2013 Aug;29(8):831-837. [doi: 10.1016/j.midw.2012.07.013] [Medline: 23079869]
5. Ståhl M, Kautiainen H, El-Metwally A, et al. Non-specific neck pain in schoolchildren: prognosis and risk factors for occurrence and persistence. A 4-year follow-up study. Pain 2008 Jul 15;137(2):316-322. [doi: 10.1016/j.pain.2007.09.012] [Medline: 17964722]
6. Huhn K, Gilliland SJ, Black LL, Wainwright SF, Christensen N. Clinical Reasoning in Physical Therapy: A Concept Analysis. Phys Ther 2019 Apr 1;99(4):440-456. [doi: 10.1093/ptj/pzy148] [Medline: 30496522]
7. Willis BW, Campbell AS, Sayers SP, Gibson K. Integrated clinical experience with concurrent problem-based learning is associated with increased clinical reasoning of physical therapy students in the United States. J Educ Eval Health Prof 2018;15:30. [doi: 10.3352/jeehp.2018.15.30] [Medline: 30586954]
8. Chaudhari A, Walke R. Role of artificial intelligence and Machine learning in musculoskeletal physiotherapy. J Pharm Negat Results 2022:2868-2870. [doi: 10.47750/pnr.2022.13.S06.369]
9. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. : OpenAI; 2019.
10. Lasater K. Clinical judgment development: using simulation to create an assessment rubric. J Nurs Educ 2007 Nov;46(11):496-503. [doi: 10.3928/01484834-20071101-04] [Medline: 18019107]
11. Olmos Migueláñez S. Evaluación formativa y sumativa de estudiantes universitarios: aplicación de las tecnologías a la evaluación educativa [article in spanish]. University of Salamanca. 2008. URL: http://hdl.handle.net/10366/18453 [accessed 2025-07-10]
12. Hollon TC, Pandian B, Adapa AR, et al. Near real-time intraoperative brain tumor diagnosis using stimulated Raman histology and deep neural networks. Nat Med 2020 Jan;26(1):52-58. [doi: 10.1038/s41591-019-0715-9] [Medline: 31907460]
13. Cobos-Carbó A, Augustovski F. Declaración CONSORT 2010: actualización de la lista de comprobación para informar ensayos clínicos aleatorizados de grupos paralelos. Medicina Clínica 2011 Jul;137(5):213-215. [doi: 10.1016/j.medcli.2010.09.034]
14. Orden CIN 2135/2008, por la que se establece el perfil de competencias del título de fisioterapeuta [article in spanish]. Consejo de Ministerios del Gobierno de España. 2008. URL: https://www.boe.es/eli/es/o/2008/07/03/cin2135 [accessed 2025-07-10]
15. Mattingly C. What is clinical reasoning? Am J Occup Ther 1991 Nov;45(11):979-986. [doi: 10.5014/ajot.45.11.979] [Medline: 1793121]
16. Boscardin CK, Gin B, Golde PB, Hauer KE. ChatGPT and Generative Artificial Intelligence for Medical Education: Potential Impact and Opportunity. Acad Med 2024 Jan 1;99(1):22-27. [doi: 10.1097/ACM.0000000000005439] [Medline: 37651677]
17. Montero Delgado JA, Merino Alonso FJ, Monte Boquet E, Ávila de Tomás JF, Cepeda Díez JM. Competencias digitales clave de los profesionales sanitarios. Educación Médica 2020 Sep;21(5):338-344. [doi: 10.1016/j.edumed.2019.02.010]
18. Barajas Galindo DE, Martínez Pillado M, Ballesteros Pomar MD, et al. Validación de un cuestionario para analizar la capacidad digital de pacientes con diabetes mellitus tipo 1. J Healthc Qual Res 2022 Nov;37(6):374-381. [doi: 10.1016/j.jhqr.2022.04.001]
19. Milad D, Antaki F, Milad J, et al. Assessing the medical reasoning skills of GPT-4 in complex ophthalmology cases. Br J Ophthalmol 2024 Sep 20;108(10):1398-1405. [doi: 10.1136/bjo-2023-325053] [Medline: 38365427]
20. Biri SK, Kumar S, Panigrahi M, Mondal S, Behera JK, Mondal H. Assessing the utilization of large language models in medical education: Insights from undergraduate medical students. Cureus 2023;15(10) [FREE Full text]
21. D'Mello S, Lehman B, Pekrun R, Graesser A. Confusion can be beneficial for learning. Learn Instr 2014 Feb;29:153-170. [doi: 10.1016/j.learninstruc.2012.05.003]

XSL•FO
RenderX

22. Khamisy-Farah R, Gilbey P, Furstenau LB, et al. Big Data for Biomedical Education with a Focus on the COVID-19 Era: An Integrative Review of the Literature. Int J Environ Res Public Health 2021 Aug 26;18(17):8989. [doi: 10.3390/ijerph18178989] [Medline: 34501581]
23. Hu JM, Liu FC, Chu CM, Chang YT. Health Care Trainees' and Professionals' Perceptions of ChatGPT in Improving Medical Knowledge Training: Rapid Survey Study. J Med Internet Res 2023 Oct 18;25:e49385. [doi: 10.2196/49385] [Medline: 37851495]
24. Park J. Medical students' patterns of using ChatGPT as a feedback tool and perceptions of ChatGPT in a Leadership and Communication course in Korea: a cross-sectional study. J Educ Eval Health Prof 2023;20:29. [doi: 10.3352/jeehp.2023.20.29] [Medline: 38096895]
25. Cook DA, Hatala R, Brydges R, et al. Technology-enhanced simulation for health professions education: a systematic review and meta-analysis. JAMA 2011 Sep 7;306(9):978-988 [FREE Full text] [doi: 10.1001/jama.2011.1234] [Medline: 21900138]
26. Kolb DA. Experimental Learning: Experience as the Source of Learning and Development: Prentice-Hall; 1984:256.
27. Fleming ND, Mills C. Not Another Inventory, Rather a Catalyst for Reflection. To Improve the Academy 1992 Jun;11(1):137-155. [doi: 10.1002/j.2334-4822.1992.tb00213.x]
28. Feasibility of a Randomized Controlled Trial of Large Artificial Intelligence-Based Linguistic Models for Clinical Reasoning Training of Physical Therapy Students. Zenodo. URL: https://zenodo.org/records/12600488 [accessed 2025-07-10]

Abbreviations

AI: artificial intelligence
CONSORT: Consolidated Standards of Reporting Trials
ICF: International Classification of Functioning, Disability and Health
LCHUS: La Salle Centre for Higher University Studies
LLM: large language model
RCT: randomized controlled trial
WHO: World Health Organization

Edited by A Mavragani; submitted 04.09.24; peer-reviewed by F Babatunde, I Salinas, J López-Marcos; revised version received 08.04.25; accepted 10.04.25; published 23.07.25.

Please cite as:
Ferrer-Peña R, Di-Bonaventura S, Pérez-González A, Lerín-Calvo A
Feasibility of a Randomized Controlled Trial of Large AI-Based Linguistic Models for Clinical Reasoning Training of Physical Therapy Students: Pilot Randomized Parallel-Group Study
JMIR Form Res 2025;9:e66126
URL: https://formative.jmir.org/2025/1/e66126
doi:10.2196/66126

© Raúl Ferrer Peña, Silvia Di Bonaventura, Alberto Pérez González, Alfredo Lerín Calvo. Originally published in JMIR Formative Research (https://formative.jmir.org), 23.7.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Just-in-Time Delivery of Cognitive Behavioral Therapy–Based Exercises: Single-Case Experimental Design With Random Multiple Baselines

Takeyuki Oba[1], PhD; Keisuke Takano[1], PhD; Daichi Sugawara[2], PhD; Kenta Kimura[1], PhD

[1]Human Informatics and Interaction Research Institute, National Institute of Advanced Industrial Science and Technology, 1-1-1 Higashi, Tsukuba, Japan

[2]Institute of Human Sciences, University of Tsukuba, Tsukuba, Japan

Corresponding Author:
Keisuke Takano, PhD
Human Informatics and Interaction Research Institute, National Institute of Advanced Industrial Science and Technology, 1-1-1 Higashi, Tsukuba, Japan

Abstract

Background: Just-in-time adaptive interventions (JITAIs) are a promising approach in mental health care given the potential scalability (ie, interventions are offered automatically and remotely) and preciseness (ie, the right interventions are offered at the right moments). Typically, a smartphone app is programmed to assess users' psychological states in daily life; when a particular state is detected, the app prompts users to engage in specific behaviors. Conceptually, JITAIs hold significant potential for precision health, although there is currently limited evidence in the literature.

Objective: We implemented this scheme as a smartphone intervention for daily stress management, based on cognitive behavioral therapy (CBT), and evaluated its feasibility and efficacy using a single-case experimental design.

Methods: A total of 8 Japanese adults (community sample: 4 women; mean 37.6, SD 13.1 y) were recruited. An AB phase design with multiple random baselines was used, where "A" represents the baseline phase and "B" represents the intervention phase. Throughout the study period (28 d), participants were prompted to indicate their momentary levels of stress (range 0 - 100) using a smartphone thrice a day. The baseline phase duration was randomly varied among participants, lasting between 7 and 14 days. The remaining period was used as the intervention phase (14 - 21 d), where 6 CBT-based exercises (ie, breath control, mindfulness, relaxation, self-talk, cognitive defusion, and cognitive restructuring) were offered depending on the reported levels of stress.

Results: Approximately 70% (6/8) of the participants perceived the intervention to be useful and helpful. A randomization test detected a statistically significant decrease in reported stress levels after the intervention began (P=.005), though this effect was less pronounced when analyzed individually for each participant. Multilevel model analysis detected a significant acute reduction in the momentary level of stress right after completing a CBT-based exercise (pre-exercise: mean 47.98, SD 21.65; post exercise: mean 42.13, SD 19.88; P=.03; Cohen dz=0.58). Also, a significant reduction in depressive rumination was observed in the postintervention assessment (preintervention: mean 13.00, SD 3.21; post intervention: mean 9.25, SD 2.60; P=.01, Cohen dz=1.17).

Conclusions: The intervention was feasible and effective in reducing subjective stress (and rumination) in the study sample. The small sample size and the nonclinical nature of the sample may limit the generalizability and implications of the study findings for clinical practice. More evidence should be collected to draw solid conclusions for technical and technological as well as clinical aspects of mobile interventions. Accumulating exemplars with different implementations will clarify how a JITAI can be designed and developed on a mobile platform and how the program can be delivered in the prevention and treatment of mental ill health.

Trial Registration: OSF Registries kgf8m; https://osf.io/kgf8m

(JMIR Form Res 2025;9:e69556) doi:10.2196/69556

KEYWORDS

ecological momentary assessment; just-in-time adaptive intervention; stress; rumination; cognitive behavior therapy; mindfulness

XSL•FO
RenderX

Introduction

Background

Digital technologies have become familiar not only in daily life but also in clinical practice for mental health. The electronic and mobile implementation of psychological interventions allows for the remote and ubiquitous delivery of evidence-based care, which is of particular importance to people who have limited access to clinics due to financial or physical constraints [1]. The COVID-19 pandemic has served as a catalyst [2], accelerating the development of digital health care services. These services can be delivered in various formats and platforms, such as computer- or internet-based interventions, via chat systems or smartphone apps, with or without human therapeutic guidance. Stakeholders appear to see mobile implementations as one of the most promising directions, given the number of apps available in the market and the huge amount of investment in recent years [3].

An increasing number of efficacy and effectiveness trials have been published in recent decades, and the results have been synthesized from different aspects. For example, a meta-analysis of digital interventions in general showed a moderate effect on improving depression [4]. Another meta-analysis, with a particular focus on app-based mobile health interventions [5], found small effects on anxiety and depression. A meta-analysis of app-based interventions for moderate to severe depression reported a larger, moderate-sized effect [6]. Overall, the evidence has been mixed, which may be attributed to heterogeneity in the intervention designs and how each intervention (or technique) is delivered.

The rapid development of mobile technologies has expanded the degrees of freedom in implementing care and interventions offered digitally and automatically. Ecological momentary interventions (EMIs) are a general scheme for delivering psychological interventions in daily life, typically for which an app is programmed to obtain information from users and prompt them to engage in specific behaviors at opportunistic times [7-9]. Many EMIs are designed to be just-in-time adaptive interventions (JITAIs), which deliver interventions remotely that are personally tailored and implemented on the basis of an individual's context (eg, with the GPS in a mobile device), psychophysiology, or responses to questions. As such, a JITAI provides the right treatment (ie, "adaptive") at the right moment (ie, "just-in-time") [10]. For the management of mental health, a JITAI typically delivers cognitive behavioral and other specific messages (eg, for behavior activation and cognitive reappraisal) tailored to an individual's present context and state [11]. The just-in-time feature is of particular importance for patients to become aware of the context in which a maladaptive psychological process (eg, depressive rumination) is triggered and the consequences that maintain it. Appropriate, timely guidance would help them perform functional analyses of their maladaptive behavior and more accurately implement alternative coping [12]. The concept and theoretical basis are not at all novel and have been implemented in existing, in-person psychological treatments. However, the most critical limitation of the traditional approaches is that interventions are offered in brief and infrequent windows of time (eg, in a session at a clinic), which assumes patients' memory and effort to apply the learned techniques in daily life. Thereby, the JITAI scheme is expected to overcome the therapy-real world gap while delivering relevant support at the time, place, and frequency needed to intervene in symptoms as they arise [13].

Decision rules, which define what interventions to deliver and when [10], constitute the core of EMI. A variety of interventions are offered in EMIs, such as cognitive behavioral therapy (CBT), acceptance and commitment therapy, mindfulness, behavioral activation, and relaxation (for a review, see [14]). Furthermore, different trials have used different logic to define delivery interventions. For example, a JITAI trial assessed participants' experiences of stressful events and related ruminative thinking several times a day, which were then used to determine which messages and CBT-based support are needed to be delivered [15]. Another JITAI system referred to contextual information (eg, the number of emails and meetings of the day) as well as passive sensing of facial expressions as tailoring variables to determine stress levels [16].

Objectives

However, EMI and JITAI are still in their infancy, and to the best of our knowledge, there are no established designs (particularly decision rules) that promise the success of an intervention. In other words, an evidence gap exists in the way to determine when and what interventions should be delivered. Interventions were offered, for example, at fixed or semirandom moments [17,18], when stressful events were reported [15], or when environmental and contextual information indicated a high workload of users [16]. Also, the degree of optimization varies across trials—the same interventions (or messages) can be delivered across different occasions, or different interventions can be scheduled according to the values of tailoring variables. Given that no standard or consensus has been established so far, it is pivotal to accumulate exemplars with different implementations and test various decision rules, as there are a number of parameters to optimize when designing a JITAI. The key parameters include the following questions:

1. What interventions or techniques should the system offer in its pool?
2. What psychological and behavioral states should be monitored, and how?
3. What conditions should be implemented to trigger or inhibit an intervention?

This study was designed as a pilot example of JITAI, implementing decision rules to offer different CBT techniques for stress management, depending on the momentary states of app users.

Specifically, we aimed to evaluate the feasibility and efficacy of an app-based intervention with just-in-time prompts using CBT-based exercises for stress reduction. The key idea was to provide an opportunity to engage in CBT exercises (including relaxation, body scans, self-compassion, distancing, and cognitive restructuring) when needed, that is, when elevated stress levels are detected. These CBT techniques (each had an evidence base with or without mobile implementations [13,19])

XSL•FO
RenderX

were offered in the form of stand-alone, self-guided exercises, which could be completed in a few minutes without any support from therapists or other health care professionals. This quick-to-go format is crucial for JITAI, similar to homework in traditional CBT [20], as participants are prompted to engage with it in daily life. Furthermore, our pilot app was designed to offer different exercises according to the detected levels of stress. This tailoring was important because stress has intraindividual, temporal dynamics, and appropriate coping strategies may differ across the levels and types of stress [21,22]. It is not always possible for individuals to make a rational decision on how to cope with their stress. A survey on general practitioners showed that even health care professionals use ineffective coping strategies (eg, avoidance strategies and self-blame) under highly stressful situations [23], which highlights the need for tailored, context-specific guidance.

To this end, we conducted this pilot, proof-of-concept trial to test the feasibility (eg, adherence, acceptability, and usability) of the JITAI approach as well as its efficacy in reducing stress over a 2- to 3-week intervention. A single-case experimental (AB phase) design was used with random multiple baselines (baseline for 7 - 14 d vs intervention for 14 - 21 d). For the feasibility, we were particularly interested in how participants would feel about the timing and frequency of the prompts, which may affect the perception of usefulness and willingness to continue the app use. One of the known challenges for mental health apps is the low retention rate [24]—this issue could be resolved by the regular prompting of JITAIs (eg, to maintain users' interests), but too much prompting or inappropriate delivery timing may have adverse effects leading to dropout. For efficacy, we hypothesized that participants would show significant decreases in subjective stress over time.

Methods

Participants

Eight adults (mean age 37.6, SD 13.1 y; 4 women) were recruited from inhabitants of Ibaraki Prefecture and its environs (middle-north regions in Japan) early in 2024. The study was advertised online (eg, via portal sites and emails), and those interested were invited to an online eligibility assessment. Participants were assessed for their current levels of subjective stress by responding to an item from the Perceived Stress Scale (PSS; "In the last week, how often have you felt nervous and stressed?") [25] by indicating the most applicable of the 5 response options: 0=never, 1=almost never, 2=sometimes, 3=fairly often, and 4=very often. Those who scored 3 or 4 on the PSS, indicating they felt nervous and stressed fairly often or very often, were eligible to participate. Other inclusion criteria were as follows: (1) aged 20 - 64 years, (2) expecting no major life changes right before or during the study participation, (3) not pregnant or lactating, (4) being in good health with no history of dysautonomia, depression, mania, bipolar disorder, or facial paralysis, (5) no medication with tranquilizers or antihypertensive agents, (6) not under hormone replacement therapy, (7) no oral contraceptives, and (8) no habit of physical activity and exercise (this criterion was used for an overarching study). Recruitment continued until the planned sample size was achieved. The sample size was determined pragmatically (considering the resources of the recruiting institution) due to the pilot nature of this study. However, a simulation study [26] examining the power of randomization tests in multiple baseline designs suggested that a sample size (N) of 6 - 10 participants would be sufficient to detect an effect of Cohen d=0.60 - 1.0 with a sufficient power (0.80).

Study Design

An "AB" design with random baselines was used (Figure 1). The entire study period was 28 days (4 wks) for each participant. The baseline phase duration was randomly varied among participants, lasting between 7 and 14 days. The remaining period was used as the intervention phase (14 - 21 d). Participants were not informed when the intervention would be started (single blinding). Throughout the 28-day study period, participants received 3 signals each day at fixed times (11 AM, 4 PM, and 9 PM) on their smartphones with an EMA app, m-Path [27], being installed. These momentary cues prompted participants to indicate their current levels of subjective stress using a 0 - 100 slider (0=not at all, 100=very much). The exact prompt states the following:

> *We would like to ask you about your current stress level. Please rate your current stress using the slider below.*

During the intervention phase, participants received in-app guidance for CBT exercises (see the next section "CBT Exercises") depending on the reported momentary levels of subjective stress. For example, if participants scored 80 or above on the stress slider, they received a recommendation for muscle relaxation, guided by a tutorial video (3 - 9 min long each). Participants were allowed to engage in an exercise that was not recommended at this moment (eg, performing breath control instead of or in addition to muscle relaxation) or skip the opportunity if they were too busy. In addition, if participants indicated low levels of stress (ie, <40 on the slider), they received no specific recommendation for exercise.

At the end of the exercises, participants were asked to specify the exercises in which they were actually engaged (this prompt appeared even when no recommendation was presented). Specifically, they indicated any applicability from the list of exercises (ie, breath control, mindfulness, progressive muscle relaxation, compassionate self-talk, cognitive defusion, and cognitive restructuring) with "other and skipped (too busy)" response options. Subsequently, participants rated their current stress levels again using a 0 - 100 slider. Postexercise stress levels were subjected to randomization tests as the primary outcome.

XSL•FO
RenderX

Figure 1. Schematic flow diagram of the study. An AB design was used (28 d in total for each participant) with the baseline length randomized across participants (7 - 14 d). CBT: cognitive behavioral therapy; GAD-7: Generalized Anxiety Disorder-7 Scale; PHQ-9: Patient Health Questionnaire-9; PSS: Perceived Stress Scale.

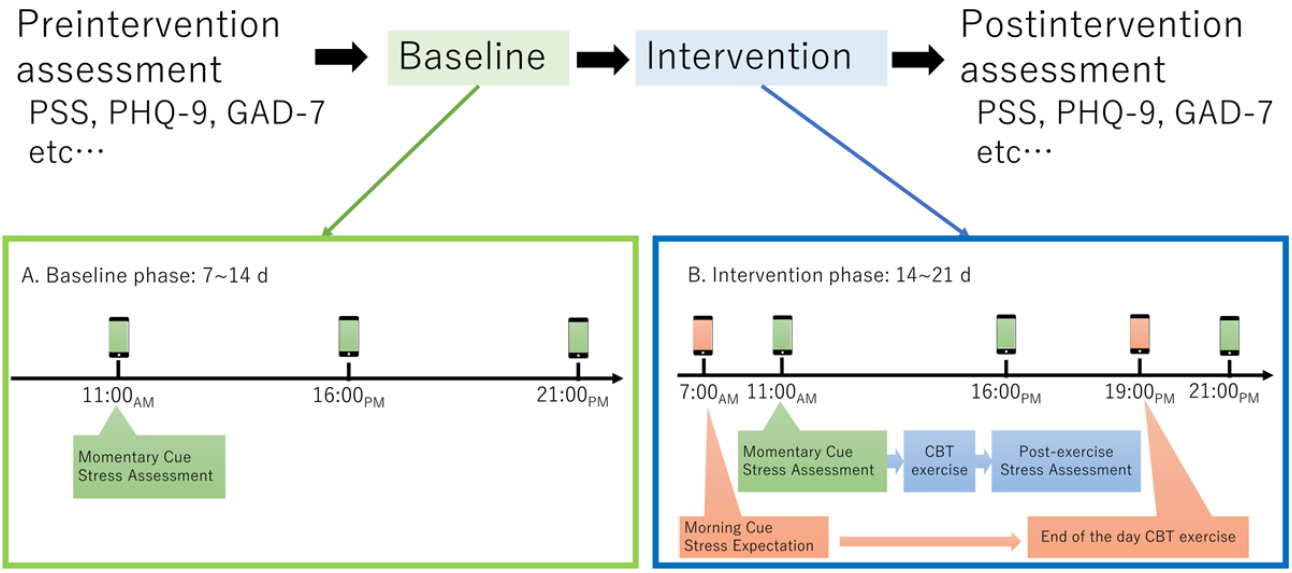

In addition to the momentary cues, participants were prompted each morning to estimate how busy they would be today; depending on the estimated levels, they received another in-app guidance for CBT exercise at 7 PM, on the day. The morning cue stated:

> *How busy do you expect to be today? Use the slider below to indicate how busy you expect to be today.*

We asked how busy they would be instead of how stressful the day would be because we were not sure how accurately participants could estimate the stress level of the day; instead, estimating the extent of busyness was deemed more reliable and easier for participants, as they could refer to, for example, their schedules.

CBT Exercises

Different CBT exercises were prompted depending on the reported levels of stress in response to the momentary and morning cues. Table 1 summarizes the individual exercises and decision rules. Among the techniques used in CBT and related programs [28,29], we implemented the following 6 exercises into the current program: breath control [30], mindfulness [31], progressive muscle relaxation [32], compassionate self-talk [33], leaves on a stream exercises [34], and cognitive restructuring [29]. These techniques are suitable for mobile implementation as stand-alone exercises, each instructed via videos lasting between 3.5 and 9.0 minutes.

XSL•FO
RenderX

Table . Decision rules defining exercises prompted for different levels of stress.

Cue, stress, or busyness level threshold	Exercise	Description, example instruction
Momentary cues[a]		
Stress level		
0 - 39	No exercise prompted	No exercise prompted
40 - 59	Breath control	Take deep breaths while counting the numbers in your mind
60 - 79	Mindfulness (focusing on breath)	Pay attention to your breathing and savor the sensation
80 - 100	Progressive muscle relaxation	Reduce stress and anxiety by progressively tensing and then relaxing different muscle groups in the body
Morning cues[b]		
Busyness level		
0 - 39	No exercise prompted	No exercise prompted
40 - 59	Compassionate self-talk	Repeat in your mind a phrase that is warm and compassionate toward self
60 - 79	Leaves on a stream (cognitive defusion)	Imagine leaves flowing down a river and observe the thoughts and feelings that arise
80 - 100	Cognitive restructuring	Recall negative events, write down thoughts and feelings, and assess whether your evaluations are accurate

[a] Participants responded to momentary cues thrice a day, and exercise was offered right after each momentary cue.

[b] Participants responded to a morning cue each day (7 AM) and indicated how busy they would be today; exercise was offered at 7 PM each day.

The first 3 exercises (breath control, mindfulness, and muscle relaxation) were prompted by momentary cues. These quick-to-go exercises (less cognitive, typically with a particular focus on the body and meditation) were selected because momentary cues were scheduled thrice a day and each exercise should not be too demanding. Another focus was on the levels of stress, and we programmed breath control for low levels of stress, mindfulness for moderate levels, and muscle relaxation for high levels. We expected that the participants (a community sample) would experience low levels of stress most frequently, for which a hands-free, light-touch exercise was deemed appropriate. Therefore, the breath-control tutorial video was the shortest (3.8 min) of the three (9.0 and 5.5 min for mindfulness and progressive muscle relaxation, respectively).

The other three exercises (compassionate self-talk, leaves on a stream exercises, and cognitive restructuring) were programmed for morning cues and were to be completed in the evening (at 7 PM or later). These imaginative or cognitive exercises were performed once a day (which is less frequent than those for the momentary cues) and thus were meant to be more cognitively effortful and typically required a longer time to complete. Studies suggest that the successful use of cognitive coping, particularly reappraisal, may produce significant cognitive costs, and people choose reappraisal less frequently than other strategies (eg, suppression and distraction) to regulate their emotions [35,36]. Therefore, we programmed cognitive exercises to be offered in the evening (deemed to be a quieter moment with fewer distractors than midday), not right after experiencing stress (ie, momentary cues). Compassionate self-talk was prompted when participants indicated that they would be somewhat busy today in response to a morning cue. Leaves on a stream, exercises, and cognitive restructuring were recommended when participants expected that the day would be tougher. Also, cognitive restructuring requires individuals to recall a specific negative event, and thus, this exercise was regarded as being suited to the busiest day when they could easily spot a stressor to target. These decision rules were determined somewhat arbitrarily, but compassionate self-talk was assumed to be less effortful and a better fit on the day when light stress was foreseen compared to the other two.

Measures

We administered two types of assessments: momentary and pre- and postintervention assessments. Our primary outcome was the momentary postexercise levels of subjective stress reported following the momentary cues, which were sent three times daily over the 28-day study period. The pre- and postintervention assessments included questionnaires on subjective stress (over the past month), psychopathological symptoms and traits, and well-being. The postintervention assessment also included feasibility measures to clarify how acceptable and appropriate the intervention was for each participant. Note that we did not examine the internal consistency of each measure, given the small sample size. However, all the measures used here have been shown to have good psychometric properties.

XSL•FO
RenderX

The PSS

Subjective stress levels were assessed using the PSS, which consists of 10 items [25]. The scale asks how respondents felt and thought during the past month (eg, "how often have you been upset because of something that happened unexpectedly?"). Each item is rated on a 5-point scale (0=never and 4=very often), and the total score is calculated. A validation study documented a mean score of 20.3 (SD 5.4) in a Japanese sample [37], in which a PSS score > mean 31, SD 2) can be interpreted as a high level of subjective stress.

Brief Patient Health Questionnaire

Depressive symptoms were assessed using the brief Patient Health Questionnaire-9 (PHQ-9) depression module [38]. This scale consists of 9 items representing depressive symptoms rated using a 4-point scale (0=not at all and 3=nearly every day). The sum score was calculated and interpreted with cutoff scores of 5, 10, 15, and 20, indicating mild, moderate, moderately severe, and severe depressive symptoms, respectively.

Seven-Item Anxiety Scale

Anxiety and symptoms of generalized anxiety disorder were assessed using the Generalized Anxiety Disorder-7 Scale (GAD-7) [39]. Each item represents typical generalized anxiety disorder symptoms (eg, nervousness and excessive worry), rated using a 4-point scale (0=not at all and 3=nearly every day). The sum score was calculated and interpreted with cutoff scores of 5, 10, and 15, indicating mild, moderate, and severe levels of anxiety, respectively.

Response Styles Questionnaire

Ruminative response styles to depressed mood were assessed using the Response Style Questionnaire (RSQ) [40,41]. The brooding subscale was the target of our interest, which is known to be predictive of depression and anxiety (eg, [42]). This subscale has 5 items (eg, Think “Why can’t I handle things better?”), and each is rated using a 4-point scale (1=almost never and 4=almost always). The sum score was calculated.

Hedonic and Eudaimonic Motives for Activities Scale

Hedonic and eudaimonic conceptions of well-being were assessed using the Hedonic and Eudaimonic Motives for Activities (HEMA) scale [43]. The HEMA scale has 2 subscales: hedonic and eudaimonic. The former assesses pleasure and the absence of pain and is subdivided into 2 factors: seeking relaxation (2 items) and seeking pleasure (3 items). The eudaimonic subscale consists of 4 items, reflecting, for example, authenticity and excellence. Each item is rated using a 7-point scale (1=not at all and 7=very much), and the sum scores are calculated for the eudaimonic, seeking-relaxation, and seeking-pleasure subscales.

Feasibility

Immediately after the intervention phase, participants rated the feasibility of the intervention online. Specifically, they rated the following aspects using a 7-point scale (1=not at all and 7=very much): the intervention was (a) satisfactory, (b) useful, (c) helpful, (d) suited to me, (e) appropriate timing, (f) practical, and (g) I would like to continue using this intervention system [44].

Procedure

Eligible participants were individually invited to the laboratory for a briefing session. They received study instructions and mobile devices (a smartphone and activity tracker) for assessment and intervention. After providing written informed consent, participants completed questionnaires on stress and symptomatology (eg, the PSS and PHQ-9). The baseline phase began on the day after the briefing. The randomization sequence for the baseline length was generated prior to recruitment and masked to the participants. Upon completion of the intervention phase, participants completed the postintervention questionnaires (online) and returned their laboratory devices in exchange for compensation for study participation.

Statistical Analyses

Momentary levels of stress, reported 3 times daily (for the intervention phase, those reported immediately after the exercises), were aggregated within each day to reduce the proportion of missing values when passing the data to the randomization tests. The aggregated scores were then subjected to randomization tests to establish the effect of the intervention, which was operationalized as the mean difference in the stress score between the baseline and intervention phases. A randomization test permutes the condition assignments of each assessment occasion for a person, which mimics the random assignment used to test a null hypothesis about the treatment effect in a randomized experiment [45]. For example, for a participant who received an assignment sequence of AABBBB for T1-T6 assessment occasions (ie, the first 2 occasions were the baseline and the following 4 were in the intervention phase), the observed test statistic is given by the mean of T1-T2 minus that of T3-T6. Then, a randomization test calculates the test statistics for each possible permuted sequence (eg, ABBBBB, AAABB, and AAAAB) to generate a randomization distribution, on which the observed test statistic (here: T1-T2 minus T4-T6) is located to test the null hypothesis, for example, no treatment effect exists. The P value is given as the proportion of the observed value equal to or smaller than the values from the permutated sequences. We set the α level to .05 and generated random permuted sequences 5000 times. Randomization tests were conducted on each participant and group of participants [46]. The R (R Core Team) package Single-Case Randomization Tests (SCRT) [45] was used to conduct the randomization tests.

As an exploratory analysis (not preregistered), a multilevel model was estimated to test whether the momentary level of stress was acutely decreased right after each CBT exercise. The model specified stress as the outcome and time (pre- vs postexercise) as the predictor with the random intercept to allow the intercept to vary across participants. Similarly, we tested whether the stress level during the baseline phase would differ from the pre-exercise level during the intervention phase. Multilevel modeling was conducted using “lmer4” [47] and “lmerTest” packages [48].

We did not specify in the preregistered protocol how missing values would be handled, as we did not have a clear prospect

XSL•FO
RenderX

for how good the compliance would be among the studied population with this specific study design. After analyzing the data, we found that several participants had missing responses for more than half of the momentary cues (see the section "Dropout and Dose"). Therefore, we concluded that imputations would not be helpful at the within-person, momentary level of the data. Also, the randomization tests were meant for the efficacy but not for the effectiveness of our pilot intervention, and therefore, a per-protocol analysis was regarded to be more appropriate than an intention-to-treat approach. To this end, the randomization tests were performed on the day-level data, with participants' momentary responses being aggregated for each day. The exploratory, multilevel modeling analysis was conducted with the maximum likelihood estimation while leaving out a participant who showed extremely low compliance (eg, dropout and dose). Although this estimation approach is robust under the assumption of missing at random, we would not argue that our results were unbiased as we did not have sufficient contextual information to gauge the randomness or underlying mechanisms of missingness; like other EMIs, it was highly likely that participants responded to momentary (and morning) cues only when they were available for that.

Ethical Considerations

The study protocol was approved by the Ethics Committee of the National Institute of Advanced Industrial Science and Technology (ID 2022‐1240). The registered protocol is available from Open Science Framework [49], and we adhered to the SCRIBE (Single-Case Reporting Guideline In Behavioural Interventions) checklist [50] (see Checklist 1). Each participant received detailed study information about the procedure, aim, and their rights in the briefing session. They were also informed that they had the option to withdraw from study participation without any consequences at any point in the study. Afterward, they provided written informed consent prior to completing the baseline questionnaires. The privacy and confidentiality of participants were ensured by pseudoanonymizing all collected data. Each participant was recognized by a unique participant ID in the collected data, and personal information (eg, participants' names and emails) was saved in a separate file. A reference list was made digitally to link the ID and personal information, which was accessible only by the first author and was deleted at the end of the data collection. Participants received JPY 15,000‐20,000 (US $100‐$140) as compensation at the end of the study.

Results

Demographics and Baseline Characteristics

Eight participants who met the inclusion criteria were enrolled, five of whom were married and had a child or children. Four participants were currently unemployed. All participants had received higher education, namely, graduated from a 2-year college, university, or above. Regarding baseline levels of stress and psychopathology, all participants had experienced high levels of stress, four had mild levels of depressive and anxiety symptoms, and one had moderate levels of symptoms (Table 2).

Table . Preintervention characteristics of each participant.

ID	Baseline length (day)	Age (years)	Gender[a]	PSS[b]	PHQ-9[c]	GAD-7[d]	RSQ[e]	HEMA scale[f] Eud[g]	HEMA scale Ple[h]	HEMA scale Relx[i]
1	7	52	Men	50	7	9	12	21	18	18
2	8	43	Men	42	4	1	17	22	18	16
3	9	22	Men	54	3	7	14	28	20	18
4	10	23	Men	46	9	8	17	19	16	22
5	11	23	Women	58	7	4	13	19	18	21
6	12	52	Women	31	4	2	7	21	18	24
7	13	39	Women	53	13	12	12	22	18	26
8	14	47	Women	37	4	6	12	17	17	13
Mean (SD)	—[j]	37.6 (13.1)	—	46.4 (9.2)	6.4 (3.4)	6.1 (3.7)	13.0 (3.2)	21.1 (3.3)	17.9 (1.1)	19.8 (4.3)

[a]Gender: self-identified.

[b]PSS: Perceived Stress Scale.

[c]PHQ-9: Patient Health Questionnaire-9.

[d]GAD-7: Generalized Anxiety Disorder-7 Scale.

[e]RSQ: Response styles questionnaire, brooding scale.

[f]HEMA: Hedonic and Eudaimonic Motives for Activities.

[g]Eud: eudaimonia.

[h]Ple: pleasure.

[i]Relx: relaxation.

[j]Not applicable.

XSL•FO
RenderX

Dropout and Dose

All participants completed the pre- and postintervention assessments (no dropouts were observed). No adverse events were reported. One participant (ID 4) showed low compliance (responded to 9/30, 30% momentary cues during the baseline phase and 2/72, 3% momentary cues during the intervention phase). While there is no established standard for the minimum number of observations needed for a randomization test as far as we know, 3 - 5 observations per phase may be a practical minimum [51]. The test on ID4, which had only 2 observations in the intervention phase, was thus deemed unreliable, and their data were excluded from statistical analyses. The remaining 7 participants responded to 75.6% (range 44.4% - 90.5%) of the momentary cues in the baseline phase and 59.7% (range 28.1% - 79.4%) in the intervention phase. They showed lower compliance with morning cues: 44.3% (range 0% - 92.9%). The frequency of recommendations and number of engagements for each type of exercise are shown in Table S1 in Multimedia Appendix 1. Breath control and mindfulness were recommended and performed most frequently. In other words, participants hardly experienced the highest (80 or above) levels of stress during the intervention period. Overall, 4 participants performed breath control more frequently than recommended. Similarly, 3 participants practiced muscle relaxation spontaneously even when not explicitly recommended.

Feasibility

Each aspect of feasibility was rated using a 7-point scale (1=not at all to 7=very much), and we counted the number of participants who indicated 4 or higher as an indication of endorsement (Table S2 in Multimedia Appendix 1). Most participants perceived the current intervention as satisfactory (63%), useful (75%), helpful (75%), and practical (63%). In addition, most participants reported that they would like to continue the intervention even after the study (75%). However, 7 out of 8 (88%) participants disagreed that the momentary and morning cues were delivered at appropriate times, suggesting that timing remains a challenge.

Intervention Effect

Figure 2 illustrates how the stress levels (for the intervention phase, stress reported immediately after each exercise) changed over time. Participants did not show a clear, acute change in the reported stress levels; however, 3 participants (ID 1, 2, and 7) appeared to have experienced reductions, whereas 2 participants (ID 5 and 6) showed slight increases during the intervention phase.

XSL•FO
RenderX

Figure 2. Changes in stress level as a function of time.

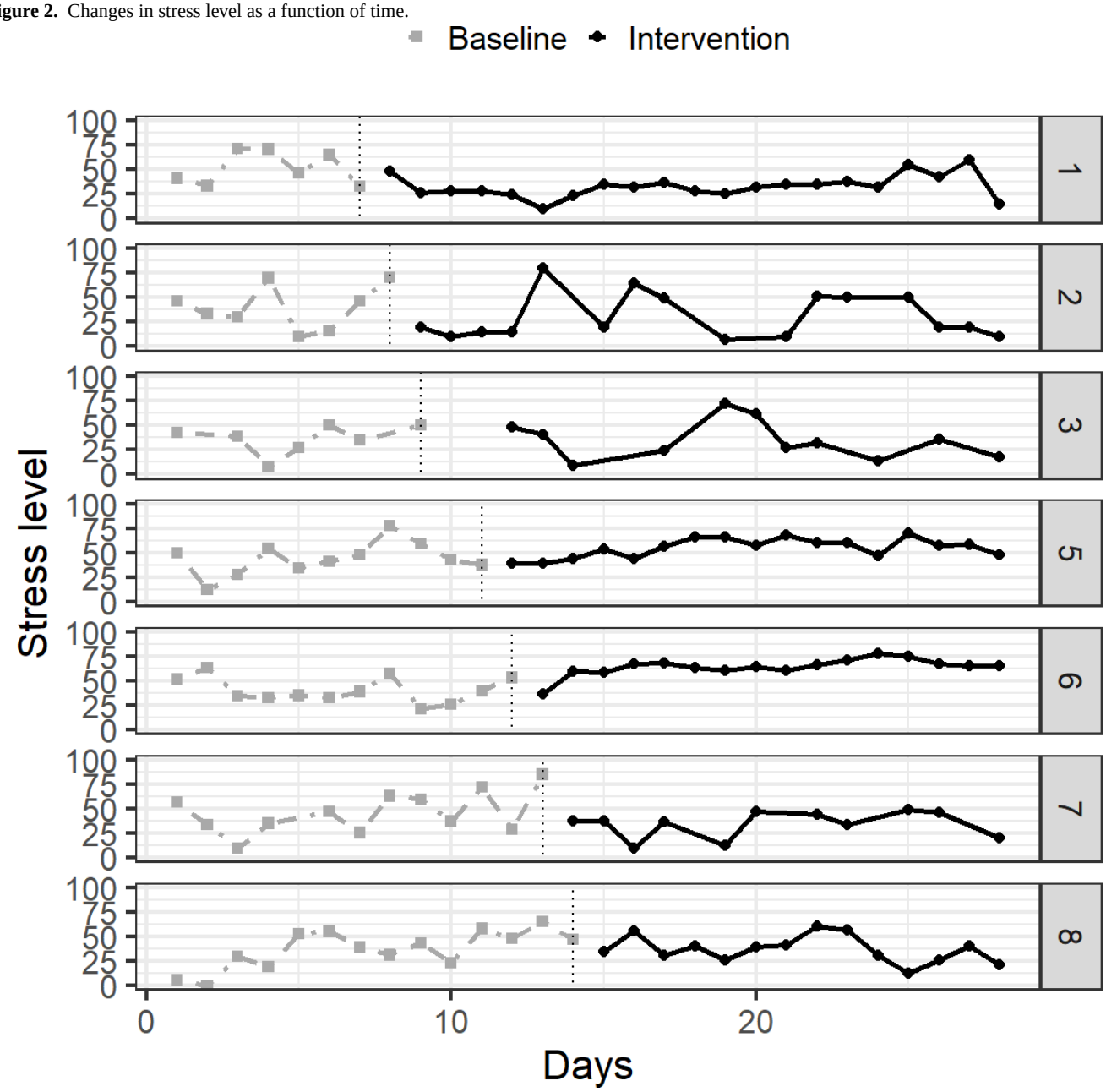

The randomization test across all participants (Table 3) showed a significant decrease in subjective stress from the baseline to the intervention phase (P=.005), rejecting the null hypothesis of no change. Note that P values were calculated for the test statistics, M (intervention)–M (baseline), with the null hypothesis, M (intervention)=M (baseline); and that ID4 was not analyzed because of low compliance. However, randomization tests for individual participants did not show significant reductions (all P>.05), consistent with the visual inspection of Figure 2, where no clear changes or inflection points were observed. In Figure 2, each panel indicates the mean stress levels per day for each participant (ID 1 - 8, except for ID 4, which was excluded due to low compliance).

XSL•FO
RenderX

Table . Means (SDs) of stress levels reported by each participant and the results of randomization tests.

ID	Baseline phase, mean (SD)	Intervention phase, mean (SD)	*P* value
1	51.8 (17.2)	33.0 (11.8)	.25
2	40.3 (22.5)	30.8 (23.0)	.13
3	35.9 (14.8)	34.8 (19.8)	.63
4	—[a]	—	—
5	44.5 (17.1)	55.4 (9.90)	.52
6	40.6 (13.1)	64.2 (9.06)	.63
7	46.1 (21.7)	34.0 (13.8)	.13
8	37.0 (19.8)	36.8 (13.9)	.12
All participants	42.3 (5.6)	41.3 (13.0)	.01

[a]Not applicable.

As an exploratory analysis, we tested whether the stress level was acutely decreased from the pre- to postexercise moments. A multilevel model analysis detected a significant decrease, fixed effect=−5.77; $t_{4.57}$=3.038; *P*=.03; Cohen *dz*=0.582 (pre-exercise: mean 47.98, SD 21.65; post exercise: mean 42.13, SD 19.88). Another multilevel model revealed no significant change from the baseline phase to the pre-exercise moments during the intervention phase (fixed effect=1.62; $t_{6.06}$=0.315; *P*=.76; Cohen *dz*=0.013; baseline: mean 44.34, SD 20.46).

We calculated the means and SDs for each measure administered at the pre- and postintervention assessments (Table S3 in Multimedia Appendix 1). Neither well-being nor psychopathology symptoms showed notable changes (Cohen *dz*<0.35; *P*>.36), except for rumination, which showed a statistically significant reduction over time (Cohen *dz*=1.17; *P*=.01). Although this result should be interpreted carefully, given the small sample size, the current intervention may help control depressive cognitive styles.

Discussion

Feasibility and Adherence

This pilot trial aimed to establish the feasibility and efficacy of the JITAI in reducing subjective stress among 8 Japanese adults experiencing high levels of subjective stress at enrollment. Participants perceived the intervention as satisfactory and helpful; overall, it is appropriate to conclude that the intervention was acceptable. Breath control and mindfulness were the most frequently recommended exercises by the system (because participants typically indicated that they were experiencing low-to-moderate levels of stress in response to momentary cues) and thus were the most frequently performed by participants. Several participants spontaneously performed breath control and relaxation, even when these exercises were not explicitly recommended. These findings are in line with the feasibility and acceptability reported in the literature for mobile (particularly mindfulness) interventions. A JITAI study for smoking cessation [52] delivered brief mindfulness strategies for stress management via smartphone, and participants reported that those strategies were helpful and they were likely to continue to use the strategies. Another trial that offered brief body scans via smartphone documented excellent adherence and continued interest of participants in mindfulness practice [53]—more than half of the participants spontaneously maintained mindfulness practice even after the active intervention phase. Good retention and engagement have also been reported in a JITAI intervention targeting repetitive negative thinking [13].

However, the feasibility evaluation revealed a practical issue: participants found that the cues were not sent at the appropriate time. There is an ongoing debate on how to optimize and personalize the intervention timing for an electronic and mobile health care system. In general, it is thought that delivering interventions at the right moments (eg, considering participants' schedules, circadian rhythms, and lifestyles) would increase the likelihood of accepting and adopting the intervention's recommendations, thereby increasing the adherence rate [54,55]. In this study, one participant failed to respond to most (around 90%) of the momentary cues and therefore missed almost all the opportunities to perform the CBT-based exercises. The remaining 7 participants showed good compliance in responding to momentary (sent at 11 AM, 4 PM, and 9 PM each day) and morning cues (at 7 AM), even though they perceived that these were not the best moments to do so. Participants may have been intrinsically motivated to engage in CBT-based exercises whenever the cues were sent; however, it should be noted that all participants were paid volunteers.

Efficacy

The overall effect of the intervention on momentary subjective stress levels merged across all participants was statistically significant. This result implies that the intervention is effective in reducing subjective stress. However, the randomization tests on individual participants failed to detect significant effects, which may suggest that the size of the effect was too small to be detected with the current design, the statistical power was not sufficient with the number of observations per participant, or both. This study used across-participants random multiple baselines, which requires at least two participants to randomize by nature. As such, the test on multiple participants fits the study design better than the individual-participant tests, which do not consider the manipulation (ie, randomization) in the baseline length. Importantly, it is known that both the larger number of

XSL•FO
RenderX

participants and the wider range of possible start moments of intervention (=possible lengths of baseline) lead to a higher power [26]. Another possibility is that the effect of the exercises was short-lived. Indeed, we found that there was a statistically significant decrease in stress right after completing a CBT exercise, but this decrease did not appear to stay for the entire intervention period.

It is also noteworthy that 2 participants (ID 5 and 6) experienced increases (but did not reach statistical significance) in the intervention phase. We do not have data that readily explain these unexpected increases; both participants received a fair dose, which was comparable to that of other participants (ID 5: breath control 17 times, ID 6: mindfulness, and cognitive defusion 11 times). It is too early to conclude that these exercises had adverse effects on the participants, but the data suggest that they were not responsive to these CBT techniques. In general, single-case experimental design, combined with repeated measurements, helps better understand how an outcome (here: stress) changes within an individual over an intervention course, compared with a typical experimental (eg, parallel group) design that averages participants' behaviors as a group [56]. Such contextualization may be the first step to explore how the outcome change is manifested in daily life and to explore who is responsive to what interventions. Establishing the optimal prescription of CBT techniques was beyond our scope, but integrating the P4 health spectrum into a JITAI system is crucial to address the variability in how individuals respond to psychological interventions [57].

Another interesting finding was that participants experienced significant reductions in depressive rumination. This effect was assessed as the difference between the two (pre- vs postintervention) time points, and thus, should be interpreted carefully, given the low statistical power. That said, it is fair to say that the current intervention might include a factor that helps reduce cognitive vulnerability (ie, rumination), such as mindfulness. Mindfulness interventions, even in a brief format or embedded in an EMI, are known to be effective not only in improving symptoms but also in reducing depressive rumination and repetitive negative thinking [53,58]. Given that rumination is understood as a mental habit sustained by learned context-response associations (eg, rumination triggered when being alone), JITAI may be a good platform to identify cuing contexts in daily life and to replace an unhelpful ruminative response to the cues with a more helpful response [59]. The just-in-time nature would well suit this scheme (ie, counterconditioning to establish a new context-response association), which allows patients to perform repeated practice at using an alternative incompatible coping strategy immediately after a triggering contextual cue is detected. However, our pilot system was not designed to collect contextual and environmental information; also, we cannot conclude which factors (mindfulness, cognitive restructuring, relaxation, etc) were the active ingredients to facilitate the learning of new compelling behaviors. Future research should explore the contextual cues to target and the (combinations of) techniques to offer for further optimization (eg, using a fractional factorial design [60]).

Limitations

This study had some limitations. First, the nonclinical nature of the sample may have limited the clinical implications of the study findings. The mobile implementation of CBT techniques, especially with a just-in-time approach, can guide daily stress management for nonclinical or at-risk individuals. However, it remains necessary to examine whether this approach works in clinical settings as a stand-alone self-help intervention or as an add-on to therapist-guided treatment. Moreover, since our sample consisted only of Japanese adults, this may limit the generalizability and replicability of our findings across different populations. Future research should target a wider range of users, such as children, adolescents, older adults, and those with different cultural backgrounds. This may involve system-wise modifications (eg, digital literacy and technology acceptance [61]) and content-wise adaptations of the implemented CBT techniques (eg, used language, acronyms, and vignettes; see for a meta-analysis of cultural adaptations [62]). Second, although the single-case design is a powerful tool for investigating the intervention effect in a small sample, it is appropriate to proceed to a larger-scale trial with appropriate controls and long-term follow-ups to establish the efficacy and effectiveness of the intervention. Third, a sampling bias may have affected the results. Overall, we found good compliance, and it is possible that the participants enjoyed the intervention and were intrinsically motivated to participate. However, all participants were paid volunteers and adherence (or user engagement) may not be readily generalized to real-life settings.

Conclusions

The current JITAI implementation of CBT techniques was generally accepted by the participants and was effective in reducing subjective stress and depressive rumination. An important direction for future research is to develop a reliable system to automatically detect psychological distress and stress (eg, [63]) that can be embedded in the decision rule (ie, if the sensor detects elevated stress, it triggers a particular intervention). Technological sophistication will improve the usability, adherence, and effectiveness of the intervention.

Acknowledgments

We thank RTC for their support in the data collection. We did not use any generative AI at any stage of the study, including preparation of materials and manuscript writing.

This study was supported by the funding program of the University of Tsukuba: Project for Industry-University Collaboration Strengthening in Tsukuba.

XSL•FO
RenderX

Data Availability

The data is available from the repository of Open Science Framework [45].

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional tables.
[DOCX File, 40 KB - formative_v9i1e69556_app1.docx]

Checklist 1
Completed SCRIBE (Single-Case Reporting Guideline In Behavioural Interventions) checklist.
[DOCX File, 21 KB - formative_v9i1e69556_app2.docx]

References

1. Andrade LH, Alonso J, Mneimneh Z, et al. Barriers to mental health treatment: results from the WHO World Mental Health surveys. Psychol Med 2014 Apr;44(6):1303-1317. [doi: 10.1017/S0033291713001943] [Medline: 23931656]
2. Torous J, Jän Myrick K, Rauseo-Ricupero N, et al. Digital mental health and COVID-19: using technology today to accelerate the curve on access and quality tomorrow. JMIR Ment Health 2020 Mar 26;7(3):e18848. [doi: 10.2196/18848] [Medline: 32213476]
3. Torous J, Wisniewski H, Bird B, et al. Creating a digital health smartphone app and digital phenotyping platform for mental health and diverse healthcare needs: an interdisciplinary and collaborative approach. J technol behav sci 2019 Jun;4(2):73-85. [doi: 10.1007/s41347-019-00095-w]
4. Moshe I, Terhorst Y, Philippi P, et al. Digital interventions for the treatment of depression: a meta-analytic review. Psychol Bull 2021 Aug;147(8):749-786. [doi: 10.1037/bul0000334] [Medline: 34898233]
5. Lu SC, Xu M, Wang M, et al. Effectiveness and minimum effective dose of app-based mobile health interventions for anxiety and depression symptom reduction: systematic review and meta-analysis. JMIR Ment Health 2022 Sep 7;9(9):e39454. [doi: 10.2196/39454] [Medline: 36069841]
6. Bae H, Shin H, Ji HG, et al. App-based interventions for moderate to severe depression: a systematic review and meta-analysis. JAMA Netw Open 2023 Nov 1;6(11):e2344120. [doi: 10.1001/jamanetworkopen.2023.44120] [Medline: 37983028]
7. Heron KE, Smyth JM. Ecological momentary interventions: incorporating mobile technology into psychosocial and health behaviour treatments. Br J Health Psychol 2010 Feb;15(Pt 1):1-39. [doi: 10.1348/135910709X466063] [Medline: 19646331]
8. McDevitt-Murphy ME, Luciano MT, Zakarian RJ. Use of ecological momentary assessment and intervention in treatment with adults. Focus (Am Psychiatr Publ) 2018;16(4):370-375. [doi: 10.1176/appi.focus.20180017] [Medline: 31191181]
9. Schueller SM, Aguilera A, Mohr DC. Ecological momentary interventions for depression and anxiety. Depress Anxiety 2017 Jun;34(6):540-545. [doi: 10.1002/da.22649] [Medline: 28494123]
10. Nahum-Shani I, Smith SN, Spring BJ, et al. Just-in-Time Adaptive Interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. Ann Behav Med 2018 May 18;52(6):446-462. [doi: 10.1007/s12160-016-9830-8] [Medline: 27663578]
11. Zainal NH, Liu X, Leong U, et al. Bridging innovation and equity: advancing public health through just-in-time adaptive interventions. Annu Rev Public Health 2025 Apr;46(1):43-68. [doi: 10.1146/annurev-publhealth-071723-103909] [Medline: 39656954]
12. Kanter JW, Manos RC, Bowe WM, et al. What is behavioral activation? A review of the empirical literature. Clin Psychol Rev 2010 Aug;30(6):608-620. [doi: 10.1016/j.cpr.2010.04.001] [Medline: 20677369]
13. Bell I, Arnold C, Gilbertson T, et al. A personalized, transdiagnostic smartphone intervention (Mello) targeting repetitive negative thinking in young people with depression and anxiety: pilot randomized controlled trial. J Med Internet Res 2023 Dec 13;25:e47860. [doi: 10.2196/47860] [Medline: 38090786]
14. Versluis A, Verkuil B, Spinhoven P, et al. Changing mental health and positive psychological well-being using ecological momentary interventions: a systematic review and meta-analysis. J Med Internet Res 2016 Jun 27;18(6):e152. [doi: 10.2196/jmir.5642] [Medline: 27349305]
15. Wang L, Miller L. Assessment and disruption of ruminative episodes to enhance mobile cognitive behavioral therapy just-in-time adaptive interventions in clinical depression: pilot randomized controlled trial. JMIR Form Res 2023 Jan 5;7:e37270. [doi: 10.2196/37270] [Medline: 36602841]
16. Suh J, Howe E, Lewis R, et al. Toward tailoring just-in-time adaptive intervention systems for workplace stress reduction: exploratory analysis of intervention implementation. JMIR Ment Health 2024 Sep 12;11:e48974. [doi: 10.2196/48974] [Medline: 39264703]

XSL•FO
RenderX

17. Versluis A, Verkuil B, Spinhoven P, et al. Effectiveness of a smartphone-based worry-reduction training for stress reduction: a randomized-controlled trial. Psychol Health 2018 Sep 2;33(9):1079-1099. [doi: 10.1080/08870446.2018.1456660]
18. Beute F, de Kort YAW. Stopping the train of thought: a pilot study using an ecological momentary intervention with twice-daily exposure to natural versus urban scenes to lower stress and rumination. Appl Psychol Health Well Being 2018 Jul;10(2):236-253. [doi: 10.1111/aphw.12128] [Medline: 29722135]
19. Martinengo L, Stona AC, Griva K, et al. Self-guided cognitive behavioral therapy apps for depression: systematic assessment of features, functionality, and congruence with evidence. J Med Internet Res 2021 Jul 30;23(7):e27619. [doi: 10.2196/27619] [Medline: 34328431]
20. Kazantzis N, Miller AR. A comprehensive model of homework in cognitive behavior therapy. Cogn Ther Res 2022 Feb;46(1):247-257. [doi: 10.1007/s10608-021-10247-z]
21. Achnak S, Vantilborgh T. Do individuals combine different coping strategies to manage their stress in the aftermath of psychological contract breach over time? A longitudinal study. J Vocat Behav 2021 Dec;131:103651. [doi: 10.1016/j.jvb.2021.103651]
22. Haines SJ, Gleeson J, Kuppens P, et al. The wisdom to know the difference: strategy-situation fit in emotion regulation in daily life is associated with well-being. Psychol Sci 2016;27(12):1651-1659. [doi: 10.1177/0956797616669086]
23. Aragonès E, Fernández-San-Martín M, Rodríguez-Barragán M, et al. Gender differences in GPs' strategies for coping with the stress of the COVID-19 pandemic in Catalonia: a cross-sectional study. Eur J Gen Pract 2023 Dec;29(2):2155135. [doi: 10.1080/13814788.2022.2155135] [Medline: 36533663]
24. Baumel A, Muench F, Edan S, et al. Objective user engagement with mental health apps: systematic search and panel-based usage analysis. J Med Internet Res 2019 Sep 25;21(9):e14567. [doi: 10.2196/14567] [Medline: 31573916]
25. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. J Health Soc Behav 1983 Dec;24(4):385-396. [doi: 10.2307/2136404] [Medline: 6668417]
26. Bouwmeester S, Jongerling J. Power of a randomization test in a single case multiple baseline AB design. In: Gasparini M, editor. PLOS ONE 2020;15(2):e0228355. [doi: 10.1371/journal.pone.0228355] [Medline: 32027683]
27. Mestdagh M, Verdonck S, Piot M, et al. m-Path: an easy-to-use and highly tailorable platform for ecological momentary assessment and intervention in behavioral research and clinical practice. Front Digit Health 2023;5:1182175. [doi: 10.3389/fdgth.2023.1182175] [Medline: 37920867]
28. Beck JS. Cognitive Behavior Therapy: Basics and Beyond, 3rd edition: Guilford Press; 2021.
29. Leahy RL. Cognitive Therapy Techniques: A Practitioner's Guide, 2nd edition: Guilford Press; 2018.
30. Chien HC, Chung YC, Yeh ML, et al. Breathing exercise combined with cognitive behavioural intervention improves sleep quality and heart rate variability in major depression. J Clin Nurs 2015 Nov;24(21-22):3206-3214. [doi: 10.1111/jocn.12972] [Medline: 26404039]
31. Williams M, Teasdale J, Segal Z, et al. The Mindful Way through Depression: Freeing Yourself from Chronic Unhappiness: Guilford Press; 2007.
32. Johnson HM. You must relax. Psychol Bull 1936;33(7):563-567. [doi: 10.1037/h0052011]
33. Neff K, Germer C. The Mindful Self-Compassion Workbook: A Proven Way to Accept Yourself, Find Inner Strength, and Thrive: Guilford Press; 2018.
34. Hayes SC, Strosahl KD, Wilson KG. Acceptance and Commitment Therapy: The Process and Practice of Mindful Change, 2nd edition: Guilford Press; 2012.
35. Troy AS, Shallcross AJ, Brunner A, et al. Cognitive reappraisal and acceptance: effects on emotion, physiology, and perceived cognitive costs. Emotion 2018 Feb;18(1):58-74. [doi: 10.1037/emo0000371] [Medline: 29154585]
36. Brans K, Koval P, Verduyn P, et al. The regulation of negative and positive affect in daily life. Emotion 2013 Oct;13(5):926-939. [doi: 10.1037/a0032400] [Medline: 23731436]
37. Sumi K. Reliability and validity of the Japanese version of the Perceived Stress Scale. The Japanese Journal of Health Psychology 2006;19(2):44-53. [doi: 10.11560/jahp.19.2_44]
38. Kroenke K, Spitzer RL, Williams JBW. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613. [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
39. Spitzer RL, Kroenke K, Williams JBW, et al. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
40. Nolen-Hoeksema S, Morrow J. A prospective study of depression and posttraumatic stress symptoms after a natural disaster: the 1989 Loma Prieta Earthquake. J Pers Soc Psychol 1991 Jul;61(1):115-121. [doi: 10.1037//0022-3514.61.1.115] [Medline: 1890582]
41. Treynor W, Gonzalez R, Nolen-Hoeksema S. Rumination reconsidered: a psychometric analysis. Cognit Ther Res 2003;27(3):247-259. [doi: 10.1023/A:1023910315561]
42. Raes F, Hermans D. On the mediating role of subtypes of rumination in the relationship between childhood emotional abuse and depressed mood: brooding versus reflection. Depress Anxiety 2008;25(12):1067-1070. [doi: 10.1002/da.20447] [Medline: 18839403]

XSL•FO
RenderX

43. Huta V, Waterman AS. Eudaimonia and its distinction from hedonia: developing a classification and terminology for understanding conceptual and operational definitions. J Happiness Stud 2014 Dec;15(6):1425-1456. [doi: 10.1007/s10902-013-9485-0]
44. Weiner BJ, Lewis CC, Stanick C, et al. Psychometric assessment of three newly developed implementation outcome measures. Implement Sci 2017 Aug 29;12(1):108. [doi: 10.1186/s13012-017-0635-3] [Medline: 28851459]
45. Bulté I, Onghena P. Randomization tests for multiple-baseline designs: an extension of the SCRT-R package. Behav Res Methods 2009 May;41(2):477-485. [doi: 10.3758/BRM.41.2.477] [Medline: 19363188]
46. Onghena P, Edgington ES. Customization of pain treatments: single-case design and analysis. Clin J Pain 2005;21(1):56-68. [doi: 10.1097/00002508-200501000-00007] [Medline: 15599132]
47. Bates D, Maechler M, Bolker B, et al. Fitting linear mixed-effects models using lme4. J Stat Softw 2015;67:1-48. [doi: 10.18637/jss.v067.i01]
48. Kuznetsova A, Brockhoff PB, Christensen RHB. lmerTest Package: Tests in Linear Mixed Effects Models. J Stat Soft 2017;82(13):1-26. [doi: 10.18637/jss.v082.i13]
49. Takano K, Sugawara D, Oba T, et al. Just-in-time delivery of CBT-based exercises: a single case experimental design with random baselines (Protocol). URL: https://osf.io/3b5tp/?view_only=822e6357a8f442409285921e4b15e060 [accessed 2023-12-23]
50. Tate RL, Perdices M, Rosenkoetter U, et al. The Single-Case Reporting Guideline In BEhavioural Interventions (SCRIBE) 2016 Statement. Phys Ther 2016 Jul;96(7):e1-e10. [doi: 10.2522/ptj.2016.96.7.e1] [Medline: 27371692]
51. Michiels B, Onghena P. Randomized single-case AB phase designs: prospects and pitfalls. Behav Res Methods 2019 Dec;51(6):2454-2476. [doi: 10.3758/s13428-018-1084-x] [Medline: 30022457]
52. Yang MJ, Sutton SK, Hernandez LM, et al. A Just-In-Time Adaptive intervention (JITAI) for smoking cessation:feasibility and acceptability findings. Addict Behav 2023 Jan;136:107467. [doi: 10.1016/j.addbeh.2022.107467] [Medline: 36037610]
53. Sommerhoff A, Ehring T, Takano K. Effects of induced mindfulness at night on repetitive negative thinking: ecological momentary assessment study (preprint). JMIR Ment Health. Preprint posted online on Nov 21, 2022. [doi: 10.2196/preprints.44365]
54. Chaudhari S, Ghanvatkar S, Kankanhalli A. Personalization of intervention timing for physical activity: scoping review. JMIR Mhealth Uhealth 2022 Feb 28;10(2):e31327. [doi: 10.2196/31327] [Medline: 35225811]
55. Solbrig L, Jones R, Kavanagh D, et al. People trying to lose weight dislike calorie counting apps and want motivational support to help them achieve their goals. Internet Interv 2017 Mar;7:23-31. [doi: 10.1016/j.invent.2016.12.003] [Medline: 28286739]
56. Bulkeley K, Bundy A, Roberts J, et al. ASD intervention research in real world contexts: refining single case designs. Res Autism Spectr Disord 2013 Oct;7(10):1257-1264. [doi: 10.1016/j.rasd.2013.07.014]
57. Sagner M, McNeil A, Puska P, et al. The P4 health spectrum – A predictive, preventive, personalized and participatory continuum for promoting healthspan. Prog Prev Med 2017;2(1):e0002. [doi: 10.1097/pp9.0000000000000002]
58. Howarth A, Smith JG, Perkins-Porras L, et al. Effects of brief mindfulness-based interventions on health-related outcomes: a systematic review. Mindfulness (N Y) 2019 Oct;10(10):1957-1968. [doi: 10.1007/s12671-019-01163-1]
59. Watkins ER, Nolen-Hoeksema S. A habit-goal framework of depressive rumination. J Abnorm Psychol 2014 Feb;123(1):24-34. [doi: 10.1037/a0035540] [Medline: 24661156]
60. Watkins ER, Newbold A. Factorial designs help to understand how psychological therapy works. Front Psychiatry 2020;11:429. [doi: 10.3389/fpsyt.2020.00429] [Medline: 32477195]
61. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989 Sep;13(3):319. [doi: 10.2307/249008]
62. Hall GCN, Ibaraki AY, Huang ER, et al. A meta-analysis of cultural adaptations of psychological interventions. Behav Ther 2016 Nov;47(6):993-1014. [doi: 10.1016/j.beth.2016.09.005] [Medline: 27993346]
63. Burns MN, Begale M, Duffecy J, et al. Harnessing context sensing to develop a mobile intervention for depression. J Med Internet Res 2011 Aug 12;13(3):e55. [doi: 10.2196/jmir.1838] [Medline: 21840837]

Abbreviations

CBT: cognitive behavioral therapy
EMI: Ecological momentary intervention
GAD-7: Generalized Anxiety Disorder-7 Scale
HEMA: Hedonic and Eudaimonic Motives for Activities
JITAI: Just-in-time adaptive intervention
PHQ-9: Patient Health Questionnaire-9
PSS: Perceived Stress Scale
RSQ: Response Style Questionnaire
SCRIBE: Single-Case Reporting Guideline In Behavioural Interventions
SCRT: Single-Case Randomization Tests

XSL•FO
RenderX

Edited by A Mavragani; submitted 02.12.24; peer-reviewed by NH Zainal; revised version received 09.04.25; accepted 11.04.25; published 24.07.25.

Please cite as:
Oba T, Takano K, Sugawara D, Kimura K
Just-in-Time Delivery of Cognitive Behavioral Therapy–Based Exercises: Single-Case Experimental Design With Random Multiple Baselines
JMIR Form Res 2025;9:e69556
URL: https://formative.jmir.org/2025/1/e69556
doi:10.2196/69556

© Takeyuki Oba, Keisuke Takano, Daichi Sugawara, Kenta Kimura. Originally published in JMIR Formative Research (https://formative.jmir.org), 24.7.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Environmental Impact of Online Versus in-Person Critical Care Education Through the Carbon Footprint Analysis of the CERTAIN Program: Cross-Sectional Study

Baiyong Wang[1,2], MD; Claudia Castillo Zambrano[2], MPH, MD; Nasrin Nikravangolsefid[2], MD; Ricardo Machado Carvalhais[3], MD; Alexander Niven[2], MD; Ognjen Gajic[2], MD; Yue Dong[4], MD

[1]Department of Critical Care Medicine, the Affiliated Hospital of Hangzhou Normal University, Hangzhou, China
[2]Division of Pulmonary and Critical Care Medicine, Department of Medicine, Mayo Clinic, Rochester, MN, United States
[3]Department of Internal Medicine, Montefiore New Rochelle Hospital, New Rochelle, NY, United States
[4]Department of Anesthesiology and Perioperative Medicine, Mayo Clinic, 1216 2nd St SW, Rochester, MN, United States

Corresponding Author:
Yue Dong, MD
Department of Anesthesiology and Perioperative Medicine, Mayo Clinic, 1216 2nd St SW, Rochester, MN, United States

Abstract

Background: Climate change is a pressing public health issue, with the US health care sector contributing about 479 million tons of carbon dioxide (CO_2) annually. Online continuing medical education offers an alternative solution to increase global education delivery while reducing CO_2 emissions associated with traditional teaching methods.

Objective: This study aimed to evaluate the carbon dioxide equivalent (CO_2e) emissions associated with different delivery methods of the CERTAIN (Checklist for Early Recognition and Treatment of Acute Illness and Injury) global critical care education program. Specifically, we aimed to compare the climate impact of local in-person courses in Rochester, MN, international in-person courses, and online courses to determine the potential environmental benefits of transitioning to digital education platforms.

Methods: A cross-sectional analysis of CO_2e emissions linked to the CERTAIN program was conducted from 2016 to 2022. We compared the climate impact of 3 different course offerings: local in-person at Rochester, MN, international in-person courses, and online courses. The international conferences were conducted in the host country with faculty traveling there to provide the educational content. CO_2e emissions were calculated using the "My Climate Flight Calculator" and "Environmental Protection Agency Emission Factors" formulas for travel, conference venues, and online course–related emissions. Learner satisfaction was assessed via validated 5-point Likert surveys.

Results: Local courses had the highest emissions: 52.7 tons/course (2.5 tons/participant), 96% from air travel (*50.6tons*, $P<.001$), versus other formats. International courses showed 20.2 tons/course (0.4 tons/participant), of which 93%(18.8 tons) were travel-related. Online courses reduced emissions by 96% per capita (0.1 tons/participant, $P<.001$) and 89% per course (5.6 tons, $P<.001$) versus local format. Overall course ratings were either excellent (live 50%, n=136) vs online 44%, n=11) or very good (live 30.9%, n=84 vs online 53%, n=12) for both live and online courses.

Conclusions: The transition to online delivery of our CERTAIN global education program has led to a substantial reduction in CO_2 emissions, mainly by eliminating travel, with similar levels of learner satisfaction. These findings support a strategic shift toward digital medical education platforms to promote environmental responsibility and broaden global educational access.

(JMIR Form Res 2025;9:e63524) doi:10.2196/63524

KEYWORDS

carbon footprint; CO2e emissions; continuing medical education; climate change; public health; medical education; environment; global health; digital education; eLearning; carbon dioxide equivalent emissions

Introduction

Climate change has been recognized as the leading threat to public health around the globe [1,2]. It is well established that rising levels of greenhouse gas (GHG) emissions cause and exacerbate a wide range of health problems due to air pollution, severe weather, wildfires, extreme temperatures, changes in vector ecology, and disturbances in the food supply, among others [3]. Health care systems are large contributors to global emissions, and intensive care units (ICUs) are a complex and resource-intensive component of these systems [4]. The US

XSL•FO
RenderX

health care industry emits an estimated 479 million tons of carbon dioxide equivalent emissions (CO_2e) each year; nearly 8% of the country's total [5]. These emissions stem directly from health care facility operations, and indirectly from purchased sources of energy, heating, and cooling, and health care services and goods supply chain [3]. If the global health care sector were a country, it would be the world's fifth-largest emitter of GHGs [6]. Agency for Healthcare Research and Quality identifies the climate crisis as the 21st century's top public health challenge, urging the health care sector to urgently reduce its carbon footprint by the administration's goals to halve emissions in 8 years and reach net zero by 2050, highlighting the necessity for medical professionals to quickly address their industry's environmental impact [7,8].

Higher education institutions are leading in sustainability by reporting GHG emissions through the American College and University Presidents Climate Commitment, which requires them to account for their emissions and create a climate action plan with reduction goals [9]. Online medical education reduces travel and energy usage at physical venues, which benefits the environment [10]. The expansion and adoption of online learning during the COVID-19 pandemic offers a more sustainable alternative to traditional, classroom-based education [11-13].

CERTAIN (Checklist for Early Recognition and Treatment of Acute Illness and Injury) is a global critical care training program focused on systematic evaluation and continuous management of critically ill patients [14-16]. Its pedagogical structure is tailored for global dissemination, including online simulation–based activities and weekly sessions, facilitated by a diverse panel of local and international experts [17]. We aimed to compare and quantify the carbon emissions associated with our local and international in-person courses, which took place prior to the COVID-19 pandemic, with our online offering that we developed and implemented over the past 4 years.

Methods

Recruitment

The CERTAIN program started in 2013 as an online global quality improvement project with demonstrated improved care process and patient outcomes from 34 ICUs from 15 countries [14]. The CERTAIN education programs (2 d live workshop) were conducted in Rochester, MN from 2016 to 2018. From 2019, this program was adapted to an international in-person offering, which was offered 6 times in 5 different countries (India, Croatia, Slovenia, China, and Vietnam). In 2020, due to the COVID-19 pandemic, the program was converted to an exclusively online offering, including both asynchronous learning using a flipped classroom model and synchronous virtual sessions that included critical care topic updates, case-based discussions, simulation-based and quality improvement training [18]. We analyzed the CO_2e associated with these 3-course delivery formats using the data of faculty and learner engagement between 2016 and 2022 recorded in our database, including (1) 24 faculty and 57 learners in Rochester, MN, (2) 26 faculty and 290 learners in our international, in-person courses, and (3) 12 faculty and 282 learners in our online offerings (Table 1).

XSL•FO
RenderX

Table . Comparative characteristics of 3 instructional modalities in the Checklist for Early Recognition and Treatment of Acute Illness and Injury critical care training program (2016 - 2022).

Course type, location, and year[a]	Faculty members (n)	Learners (n)	Course duration (h)
Rochester course (in-person)[b]			
2016/Rochester	6	11	8
2017/Rochester	6	11	8
2018/Rochester	6	19	8
2019/Rochester	6	16	8
Subtotal	24	57	32
International course (in-person)[c]			
2019/New Delhi, India	3	40	8
2019/Zagreb, Croatia	3	28	8
2019/Ljubljana, Slovenia	5	61	8
2019/Guangzhou, China	5	93	8
2019/Hanoi, Vietnam	5	41	8
2019/Beijing, China	5	27	8
Subtotal	26	290	48
Online[d]			
2020/China	3	32	40
2021/Bosnia	3	107	40
2021/Montenegro	3	74	40
2022/Vietnam	3	69	40
Subtotal	12	282	160

[a]Study design: cross-sectional analysis; population: 629 learners/62 faculty members; locations: Rochester (US), India, Croatia, China, Slovenia, Vietnam, Bosnia, and Montenegro; timeframe: 2016 - 2022

[b]Rochester course: in-person training conducted annually in Rochester.

[c]International course: in-person training conducted in various international locations.

[d]Online: courses delivered virtually in the specified years and regions.

We performed a comprehensive analysis of the CERTAIN Course's carbon footprint, including CO2e emission factors for travel, hotel, conference facility venues, local transportation, and educational technology (personal computers, network data transfer, servers, and organizer meetings) (Figure 1).

XSL•FO
RenderX

Figure 1. Carbon footprint calculation framework for the CERTAIN program using my climate calculator and EPA standards. carbon dioxide (CO_2)

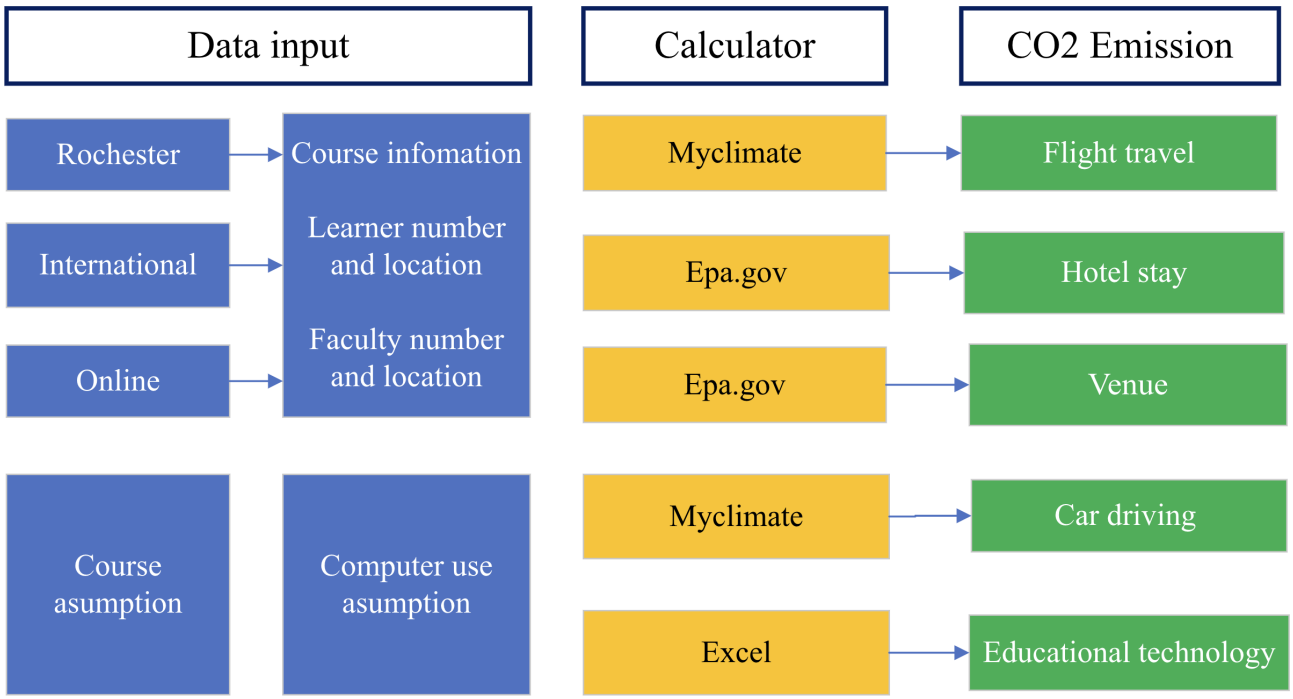

The assumption was used for the following categories:

1. Flight travel: We used the My Climate flight calculator [19], a publicly accessible internet-based tool, to determine direct and indirect CO_2e emissions per airplane passenger for each specified flight distance. This calculation used the average fuel burn rates of common aircraft. We assumed that all passengers flew economy class and traveled from their respective capital cities to Rochester and back.
2. Hotel stay: To calculate emissions produced by hotel stays. We used energy consumption data from the Environmental Protection Agency's (EPA) resources, “Indirect Emissions from Events and Conferences” [20]. The accommodation provided fell under the “Midscale with food and beverage” category, and the event was hosted in the Midwest region. We assumed registrants to stay in the “Midscale” hotel hosting the conference for 2 nights.
3. Venue: To calculate venue-related emissions, we used emission factors sourced from Environmental Protection Agency (EPA) datasets, “Indirect Emissions from Events and Conferences” [19]. We assumed using 1500 square feet of meeting space at the venue.
4. Car driving: we used “My Climate” to calculate driving by car emissions [21]. Our method was grounded in defined assumptions: The car with a petrol fuel type, consuming 8 liters per 100 kilometers, emits 0.033 metric tons (33 kilograms) of CO_2. Each attendee was assumed to make a round trip from their residence to the event location, a distance estimated at 5 kilometers separated from the hotel and the venue, with faculty members incurring an equivalent travel distance. Additionally, the transit from the airport to the hotel, a 20-kilometer journey, was anticipated to be undertaken twice by each participant.
5. Educational technology: We used methodologies adapted from Faber’s research [22,23] to calculate learners’ computer CO_2e emissions from various components, including computer use, network data transfer, and electricity consumption associated with both personal devices and servers. We calculated the carbon footprint of the computers used by the participants and the organizers of the online conference.

In the Rochester courses and international courses, we assumed about one-fourth of learners used computers. We assumed that everyone used the same model of computer for 8 hours on the conference day and 1 hour for each course meeting. The faculty also attended an organizer meeting for 8 hours.

We calculated the emissions for the online course assuming each learner used the same model of computer. The course duration was 40 hours, and the organizer meeting was 1 hour. We also assumed that 2 faculty members attended the organizer meeting and used the same computer model as the learners.

Data Collection

This study used a secure online database for data collection, gathering detailed information on faculty and learners of the CERTAIN course, including course duration and learner origins. We used Microsoft Excel to calculate flight emissions, hotel stay emissions, venue emissions, car driving emissions, and educational technology (computer use, network data transfer, and electricity consumption associated with both personal devices and servers).

Statistical Analysis

Continuous variables are presented as mean (SD). Differences in average carbon dioxide equivalent (CO_2 e) emissions (per course and per participant) across the 3 delivery formats—local in-person (Rochester, MN; n=4 courses, 81 participants), international in-person (n=6 courses, 316 participants), and online (n=4 courses, 294 participants)—were analyzed using one-way ANOVA in IBM SPSS Statistics (version 28.0; IBM Corp). Normality of data distribution was confirmed via the

XSL•FO
RenderX

Shapiro-Wilk test (P>.05 for all groups), and homogeneity of variances was verified using Levene test (P>.10). Post-hoc pairwise comparisons were conducted with Tukey Honest Significant Difference test to adjust for multiple comparisons (α=.05).

Ethical Considerations

The Mayo Clinic Institutional Review Board (IRB) acknowledges that based on the responses submitted for this new activity through the Mayo Clinic IRBe Human Subjects Research Wizard tool, and in accordance with the Code of Federal Regulations, 45 CFR 46.102, the above noted activity does not require IRB review.

Results

CERTAIN Program Course Modalities

Course Implementation and Participation

The CERTAIN program was delivered through three distinct modalities between 2016 and 2022, reaching a total of 629 learners across 98 faculty members.

Rochester Course (In-person)

The Rochester in-person course was conducted annually from 2016 to 2019. A total of 24 faculty members delivered the program to 57 learners over four years, with each course lasting 8 hours. Annual enrollment ranged from 11 to 19 participants, with the highest enrollment recorded in 2018 (n=19).

International Course (In-person)

Six international in-person courses were implemented across five countries in Asia and Europe. These courses engaged 26 faculty members and trained 290 learners, representing the largest participant cohort among the three modalities. Each course maintained the standard 8-hour duration. The highest enrollment was observed in Guangzhou, China (n=93), followed by Ljubljana, Slovenia (n=61). Course sizes varied considerably, ranging from 27 participants (Beijing, China) to 93 participants (Guangzhou, China).

Online Course

Four online courses were delivered between 2020 and 2022, primarily targeting participants in Asia and the Balkans. A total of 12 faculty members facilitated these courses for 282 learners. Unlike the in-person modalities, online courses were extended to 40 hours duration. The largest online cohort was recorded in Bosnia (2021) with 107 participants, while the smallest was in China (2020) with 32 participants (Table 1).

Calculated CO_2e emissions for each of the 3-course offerings are summarized in Table 2 with detailed calculations in Multimedia Appendix 1. The local Rochester, MN course, involving a consistent faculty number and varying international learner enrollments, was responsible for a total of 210.8 tons of CO_2e emissions. The results showed a year-on-year increase in emissions, from 22.4 tons CO_2e in 2016 to a peak of 78.8 tons CO_2e in 2018, with only a slight reduction to 65.7 tons CO_2e in 2019. This difference was due to the variation in learners' number. Course participation resulted in an average emission of 2.5 tons CO_2e per participant and 52.7 tons CO_2e per course (Table 2).

XSL•FO
RenderX

Table . Breakdown of carbon dioxide equivalent emissions by source across instructional modalities unit (metric tons).

Course type, location, and year[ab]	Carbon dioxide equivalent emissions by category (kg)[c]						Total/average emission per course (tons)	Average CO_2 emission per person (tons)
	Flight travel	Hotel stays	Venue	Driving by car	Educational technology[d]	Total		
Rochester course (in-person)								
2016	21,500	398.2	64	178.2	229.3	22,369.7	22.4	1.3
2017	43,000	398.2	64	244.2	229.3	43,935.7	43.9	2.6
2018	77,400	687.8	64	402.6	243.9	78,798.3	78.8	3.2
2019	64,500	579.2	64	343.2	236.6	65,723	65.7	3
Total	206,400	2063.4	256	1168.2	939	210,826.6	210.8/52.7	2.5
International course (in-person)								
New Delhi	15,600	108.6	64	323.4	258.5	16,354.5	16.4	0.4
Zagreb	9000	108.6	64	244.2	236.6	9653.4	9.7	0.3
Ljubljana	14,500	181	64	501.6	309.7	15,556.3	15.6	0.2
Guangzhou	27,000	181	64	712.8	368.2	28,326	28.3	0.3
Ha Noi	28,500	181	64	369.6	273.1	29,387.7	29.4	0.6
Beijing	21,500	90.5	32	171.6	251.2	22,045.3	22	0.7
Total	116,100	850.7	352	2323.2	1697.3	121,323.2	121.4/20.2	0.4
Online								
2020, China	0	0	0	0	4170	4170	4.2	0.1
2021, Bosnia	0	0	0	0	6911.6	6911.6	6.9	0.1
2021, Montenegro	0	0	0	0	5705.3	5705.3	5.7	0.1
2022, Vietnam	0	0	0	0	5522.5	5522.5	5.5	0.1
Total	0	0	0	0	22,309.4	22,309.4	22.3/5.6	0.1

[a]Study design: cross-course comparison.

[b]Geographic scope: multinational; timeframe: 2016 - 2022.

[c]Emissions: local courses (52.7 metric tons/course, 2.5 metric tons/per person), international courses (20.2 metric tons/course ,0.4 metric tons/per person), and online courses (5.6 metric tons/course, 0.1 metric tons/per person).

[d]Educational technology emissions include learner's computers, network data transfer, server emissions, and faculty meetings.

The collective carbon emissions from the CERTAIN international courses in 2019 amounted to 121.3 metric ton of CO_2e. Air travel was again identified as the primary source of emissions, accounting for over 93% in each case. Course participation resulted in an average emission of 0.4 tons CO2e per participant and 20.2 tons CO_2e per course (Table 2).

CO_2e emissions of online CERTAIN courses totaled 22.3 tons over the 3 years. Despite the scale of these courses—spanning multiple countries and involving a significant number of learners and faculty—the average CO_2e emission per participant was only 0.1 tons CO_2e, and per course was 5.6 tons CO_2e emissions (Table 2).

Post-hoc analysis using Tukey HSD test indicated that the CO_2 emissions of the Rochester course were significantly higher than those of the international and online courses (P<.001). Additionally, the CO_2 emissions of the international course were significantly higher than those of the online course (P<.001) (Table 3).

XSL•FO
RenderX

Table . Statistical comparison of carbon dioxide equivalent emissions across instructional modalities (ANOVA).

Variables[a, b]	In-person course ,Rochester	International course	Online course	*F* test (*df*)[c]	*P* value[c]	Post-hoc[d]
Average carbon dioxide equivalent emissions per course (tons);[e] n, mean (SD)	4, 52.71 (25.5)	6, 18.55 (8.39)	4, 5.58 (1.2)	24.67 (2,11)	<.001	R[f]>I[g]> O[h]
Average carbon dioxide equivalent emissions per person[i] (tons); n, mean (SD)	81, 1.30 (0.25)	316, 0.40 (0.15)	294, 0.10 (0.02)	18.45 (2,688)	<.001	R>I > O

[a]Study design: variance analysis.

[b]Locations: United States, India, and 3 other countries; timeframe: 2016-2022.

[c]One-way ANOVA results are reported with *F* test and *P* value.

[d]Post-hoc (Tukey honest significant difference): symbols indicate pairwise comparisons (“>” denotes significant differences). Post-hoc tests confirmed Rochester >International >Online.

[e]Average per course (kg): Significant differences were observed between course types ($F_{2,11}$=24.67, *P*<.001).

[f]R: Rochester.

[g]I: international.

[h]O: online.

[i]Average CO_2 per person (tons): Significant differences were also found ($F_{2,688}$=18.45, *P*<.001), with the same hierarchical pattern (Rochester>International>Online).

A comparison of the average emissions per course for each of the 3 CERTAIN program formats is summarized in Table 4. Rochester leads in CO_2e emissions, both in total and per capita, significantly surpassing International and Online, which show much lower emissions in comparison. Notably, online shows the lowest emissions in comparison.

Table . Post-hoc analysis of carbon dioxide equivalent emission differences across instructional modalities using Tukey Honest Significant Difference test.

Comparison[a, b] variables	Emissions, mean difference (tons)	*P* value[c]
Average emissions per course		
Rochester versus international	3.42	<.001
Rochester versus online	4.71	<.001
International versus online	1.30	<.001
Average CO_2 per person		
Rochester versus international	0.90	.002
Rochester versus online	1.20	<.001
International versus online	0.30	.005

[a]Study design: multiple comparisons.

[b]Geographic scope: multinational; timeframe: 2016 - 2022.

[c]Significant differences: local versus international (*P*.001), local versus online (*P*<.001), international versus online (*P*=.005).

Discussion

Principal Findings

Online education has been adopted by the medical field and has been growing after the COVID-19 pandemic [13]. Health care professionals are also aware of the impact of the health care industry on the environment [24]. Our analysis underscores the considerable environmental benefits of transitioning academic meetings and conferences to online platforms. In comparing the carbon footprints of CERTAIN courses delivered in Rochester, internationally, and online, we observed online delivery resulted in 98% and 61% lower GHG impact compared to the in-person Rochester and international courses, respectively, due to reduced travel requirements. Our findings are consistent with a growing body of evidence suggesting that increasing the use of online platforms for academic meetings and conferences offers an effective strategy to lessen the environmental impact of these activities [25]. Roy et al [26] demonstrated that distance learning in higher education consumes 87% less energy and produces 85% lower CO_2e emissions than traditional full-time campus-based courses.

XSL•FO
RenderX

Similarly, part-time campus courses are also more efficient, reducing energy and CO_2e emissions by 65% and 61%, respectively. The primary contributing factors to these reductions are decreased student travel, reduced energy consumption for student housing, and more efficient campus site use [26].

Transportation already accounts for up to 26% of global CO_2e emissions and is one of the few industrial sectors where emissions continue to increase [27]. Traveling by airplane has been considered the main factor contributing to the high carbon footprint of medical congresses around the globe [28]. Climate experts warn that the Paris Agreement's target of limiting global warming to 1.5°C by 2050 requires drastic cuts in greenhouse gas emissions, with target individual emissions of 2.5, 1.4, and 0.7 tons of CO_2e per year by 2030, 2040, and 2050, respectively [29]. The average American emits 14.9 tons in 2022, and the global average is 4.6 tons of CO_2e [2,30]. For example, a single conference trip can account for as much as 7% of an average individual's total annual CO2e emissions [31]. Multiple authors have demonstrated a dramatic improvement in CO2e leveraging virtual continuing medical education delivery and eliminating these costly travel requirements [13,23,32-35]. Adopting online and hybrid educational strategies aligns the medical field with sustainability goals, setting a precedent for environmentally responsible practices [36].

Online platforms also ensure that high-quality and equitable learning opportunities are accessible to a global audience, regardless of geographical location or economic status [37]. This accessibility improves the delivery of high-quality evidence-based care to patients, as health care professionals who engage in simulation courses would be able to apply their learned skills directly to their clinical practice [38-40]. Additionally, it fosters international collaboration and broadens global clinical and research opportunities by sponsoring exchanges of research scholars and intensivists among institutions [37,38,41].

However, the available platforms of online classes do not provide a classroom-like feel, thereby creating hindrances in students' learning process and diminishing the quality of education [42]. While many have raised valid concerns about the quality of online compared to in-person educational experiences, especially on factors like poor Internet access, student engagement, interaction, and access to resources [43], there is also a lot of promise in the initial findings. Research by Almendingen et al and Clark et al [44,45] found that students adapted well to online learning during the COVID-19 pandemic, showing improved academic performance. Specifically, Clark et al [45] reported that students who accessed higher-quality, externally recorded lessons significantly outperformed peers who used their own school's recordings, showing the benefits of quality educational resources. Gonzalez et al [46] found that a large cohort of students from the Universidad Autónoma de Madrid developed more consistent learning habits using an online learning model during the COVID-19 pandemic, which positively influenced academic performance [46]. This finding is reinforced by observations in India, where online tests and open-book exams have also been correlated with better student scores [46]. In our experience, transitioning to a digital platform for medical education posed challenges but achieved similar rates of learner satisfaction [17].

Online learning also offers a solution by eliminating the need to travel, thus preventing the risk of acute sleep deprivation and its effects on performance and memory processing [47,48]. This enables participants to engage fully and efficiently in their educational activities with greater flexibility by joining live or watching video recordings at the time of their convenience.

While the benefits of international online learning are evident, its limitations, including technological and language barriers, must also be acknowledged [40]. However, in-person, hands-on experience is still necessary in many aspects of medical education, especially in procedural training [49]. Online education not only aligns with worldwide efforts to mitigate climate change and global warming but also offers a pathway to enhance inclusive, flexible, and accessible education. This suggests a hybrid model may be a promising future direction, integrating the strengths of both online and in-person learning platforms [50] to create an optimal balance, particularly for practical skill acquisition. Additionally, it suggests the need for further research into the environmental impacts of incorporating advanced technologies like virtual reality into medical training. Finally, our findings underscore the need for broader awareness and analyses using these methods on the long-term effects of online medical education on global sustainability and, to better understand its overall impact [51].

Study Limitations

There are some limitations of our study. Emission calculators require a variety of different assumptions, which can result in variability in the resulting estimates of carbon emissions and their environmental impact. Our approach might not fully capture actual flight emissions due to missing variables like flight class or route changes. Car travel calculations, based on straight-line distances and gasoline-powered vehicles, may not reflect actual fuel consumption. We also were not able to include emissions from nonparticipant companion travel and ancillary activities outside of conference participation, which resulted in an underestimation of the total emissions impact. Additionally, we found it was impossible to account for all online activity–related emissions and could not capture such common sources as search engine queries, monitor usage, desk lamp usage, and website visits. The study also assumed uniform computer usage among participants, oversimplifying diverse technology use and energy consumption. In general, however, these unaccounted emissions were considered minor relative to the overall footprint of online learning. Our study is a retrospective analysis of the previous course. There is a great need to develop an online calculator like in the airline industry as a planning tool for medical education activities [52].

Conclusions

Online medical education offers significant environmental benefits and increased global availability over traditional in-person methods. With high course ratings and comparable learner satisfaction, our findings join a growing body of literature suggesting a strategic shift toward digital or hybrid

XSL•FO
RenderX

learning platforms. This transition is seen as key to enhancing sustainability, accessibility, and global reach in medical education, aligning with global environmental goals and evolving educational needs.

Acknowledgments

The authors thank Tekin Aysun for her contributions to Table 4. During the preparation of this manuscript, the authors used ChatGPT (OpenAI) for limited language polishing and grammatical refinement of certain sections. The artificial intelligence (AI) was exclusively used to enhance readability and ensure linguistic accuracy, while all scientific content, data interpretation, and conclusions remain entirely the authors' original work. The use of AI was carefully supervised, with all outputs verified for factual accuracy and compliance with academic standards. No AI tool was used for content generation, data analysis, or literature synthesis.

Authors' Contributions

BYW and YD participated in study design, data collection and analysis, drafting of the initial manuscript, and manuscript revision. CCZ, NN, and RMC were responsible for critical revision of the paper. OG and AN contributed to study design and manuscript review and provided key feedback and edits. All authors reviewed and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Characteristics of three instructional modalities in the Checklist for Early Recognition and Treatment of Acute Illness and Injury critical care training program
[XLSX File, 31 KB - formative_v9i1e63524_app1.xlsx]

References

1. Nordberg LB, Pohl H, Haavardsholm EA, Lillegraven S, Bretthauer M. Carbon-footprint analyses in RCTs - toward sustainable clinical practice. N Engl J Med 2024 Jun 27;390(24):2234-2236. [doi: 10.1056/NEJMp2402018] [Medline: 38749030]
2. Climate change. World Health Organization. 2023. URL: https://www.who.int/news-room/fact-sheets/detail/climate-change-and-health [accessed 2025-06-22]
3. Dzau VJ, Levine R, Barrett G, Witty A. Decarbonizing the U.S. health sector - a call to action. N Engl J Med 2021 Dec 2;385(23):2117-2119. [doi: 10.1056/NEJMp2115675] [Medline: 34644470]
4. Masud FN, Sasangohar F, Ratnani I, et al. Past, present, and future of sustainable intensive care: narrative review and a large hospital system experience. Crit Care 2024 May 9;28(1):154. [doi: 10.1186/s13054-024-04937-9] [Medline: 38725060]
5. Richie C. Environmental sustainability and the carbon emissions of pharmaceuticals. J Med Ethics 2022 May;48(5):334-337. [doi: 10.1136/medethics-2020-106842] [Medline: 33853877]
6. Wang H, Horton R. Tackling climate change: the greatest opportunity for global health. Lancet 2015 Nov 7;386(10006):1798-1799. [doi: 10.1016/S0140-6736(15)60931-X] [Medline: 26111438]
7. Sampath B. Reducing healthcare carbon emissions: a primer on measures and actions for healthcare organizations to mitigate climate change. Institute for Healthcare Improvement. 2022. URL: https://doi.org/10.13074/jent.2025.03.2511154 [accessed 2025-06-22]
8. WHO guidance for climate-resilient and environmentally sustainable health care facilities. World Health Organization. 2020. URL: https://www.who.int/publications/i/item/9789240012226 [accessed 2025-06-22]
9. Klein-Banai C, Theis TL. Quantitative analysis of factors affecting greenhouse gas emissions at institutions of higher education. J Clean Prod 2013 Jun;48:29-38. [doi: 10.1016/j.jclepro.2011.06.004]
10. Sharma D, Rizzo J, Nong Y, et al. Virtual learning decreases the carbon footprint of medical education. Dermatol Ther (Heidelb) 2024 Apr;14(4):853-859. [doi: 10.1007/s13555-024-01120-4] [Medline: 38532067]
11. Tsagkaris C, Hoian AV, Ahmad S, et al. Using telemedicine for a lower carbon footprint in healthcare: a twofold tale of healing. The Journal of Climate Change and Health 2021 Mar;1:100006. [doi: 10.1016/j.joclim.2021.100006]
12. Lameirão ADS, Meiriño MJ, Fortes MZ. Correlation between electric energy use and CO 2 emissions in a university campus in Brazil. International Journal of Emerging Electric Power Systems 2022 Jun 16;23(3):435-450. [doi: 10.1515/ijeeps-2021-0186]
13. Gandhi AP, Padhi BK, Goel K, Singh A, Kansal OP, Al-Ahdal T. Impact of virtual continued medical education on carbon footprint and awareness of digital sobriety: a retrospective cross-sectional study among public health professionals in India. Front Public Health 2023;11:1118685. [doi: 10.3389/fpubh.2023.1118685] [Medline: 36992884]

XSL•FO
RenderX

14. Vukoja M, Dong Y, Adhikari NKJ, et al. Checklist for early recognition and treatment of acute illness and injury: an exploratory multicenter international quality-improvement study in the ICUs with variable resources. Crit Care Med 2021 Jun 1;49(6):e598-e612. [doi: 10.1097/CCM.0000000000004937] [Medline: 33729718]
15. Vukoja M, Kashyap R, Gavrilovic S, Dong Y, Kilickaya O, Gajic O. Checklist for early recognition and treatment of acute illness: International collaboration to improve critical care practice. World J Crit Care Med 2015 Feb 4;4(1):55-61. [doi: 10.5492/wjccm.v4.i1.55] [Medline: 25685723]
16. Rovati L, Zec S, Dziuba D, et al. Implementation of a multimodal knowledge-exchange platform to provide trauma critical care education during the ongoing conflict in Ukraine. JAMA Netw Open 2023 Feb 1;6(2):e230050. [doi: 10.1001/jamanetworkopen.2023.0050] [Medline: 36763355]
17. Sun Y, Kashyap R, Heise K, et al. Digital delivery of a global critical care education program. ATS Sch 2023 Jun;4(2):198-206. [doi: 10.34197/ats-scholar.2022-0086IN] [Medline: 37538073]
18. Flurin L, Tekin A, Bogojevic M, et al. International virtual simulation education in critical care during COVID-19 pandemic: preliminary description of the virtual checklist for early recognition and treatment of acute illness and injury program. Simul Healthc 2022 Jun 1;17(3):205-207. [doi: 10.1097/SIH.0000000000000656] [Medline: 35439791]
19. CO2 flight calculator. myclimate. URL: https://co2.myclimate.org/en/flight_calculators/new [accessed 2025-06-22]
20. Greenhouse gas inventory guidance indirect emissions from events and conferences. EPA Center for Corporate Climate Leadership. 2018. URL: https://www.epa.gov/sites/default/files/2018-12/documents/indirectemissions_draft2_12212018_b_508pass_3.pdf [accessed 2025-06-22]
21. CO2 emissions calculator for your car. myclimate. 2024. URL: https://co2.myclimate.org/en/car_calculators/new [accessed 2025-06-22]
22. Faber G. A framework to estimate emissions from virtual conferences. International Journal of Environmental Studies 2021 Jul 4;78(4):608-623. [doi: 10.1080/00207233.2020.1864190]
23. Parker EB, Bluman A, Pruneski J, et al. American Orthopaedic Foot and Ankle Society Annual Meeting all-in-person attendance results in immense carbon expenditure. Clin Orthop Relat Res 2023 Dec 1;481(12):2469-2480. [doi: 10.1097/CORR.0000000000002764] [Medline: 37493467]
24. Mominkhan D, Alamri FA, Balharith M, et al. The current state and potential evolution of carbon emissions in the healthcare sector: a narrative review article. Front Sustain Energy Policy 2023;2. [doi: 10.3389/fsuep.2023.1230253]
25. Bousema T, Selvaraj P, Djimde AA, et al. Reducing the carbon footprint of academic conferences: the example of the American Society of Tropical Medicine and Hygiene. Am J Trop Med Hyg 2020 Nov;103(5):1758-1761. [doi: 10.4269/ajtmh.20-1013] [Medline: 33069267]
26. Roy R, Potter S, Yarrow K. Designing low carbon higher education systems: environmental impacts of campus and distance learning systems. International Journal of Sustainability in Higher Education 2008;9:116-130. [doi: 10.1108/14676370810856279]
27. Macintosh A, Wallace L. International aviation emissions to 2025: can emissions be stabilised without restricting demand? Energy Policy 2009 Jan;37(1):264-273. [doi: 10.1016/j.enpol.2008.08.029] [Medline: 32287868]
28. Storz MA. Medical conferences and climate change mitigation: challenges, opportunities, and omissions. J Occup Environ Med 2019 Oct;61(10):e434-e437. [doi: 10.1097/JOM.0000000000001681] [Medline: 31348418]
29. Ryu K. 1.5-degree lifestyles: targets and options for reducing lifestyle carbon footprints. In: Institute for Global Environmental Strategies (IGES) 2019. [doi: 10.57405/iges-6719]
30. global Carbon Atlas: A platform to explore and visualize the most up-to-date data on carbon fluxes resulting from human activities and natural processes. global Carbon Atlas. 2023. URL: https://globalcarbonatlas.org [accessed 2025-06-22]
31. Spinellis DD, Louridas P. The carbon footprint of conference papers. PLoS ONE 2013;8(6):e66508. [doi: 10.1371/journal.pone.0066508] [Medline: 23840496]
32. Jäckle S. The carbon footprint of travelling to international academic conferences and options to minimise it. In: Academic Flying and the Means of Communication: Springer Nature; 2022:19-52. [doi: 10.1007/978-981-16-4911-0_2]
33. Kitamura Y, Karkour S, Ichisugi Y, Itsubo N. Carbon footprint evaluation of the business event sector in Japan. Sustainability 2020;12(12):5001. [doi: 10.3390/su12125001]
34. Neugebauer S, Bolz M, Mankaa RN, Traverso M. How sustainable are sustainability conferences? – Comprehensive life cycle assessment of an international conference series in Europe. J Clean Prod 2020 Jan;242:118516. [doi: 10.1016/j.jclepro.2019.118516]
35. Pickard Strange M, Booth A, Akiki M, Wieringa S, Shaw SE. The role of virtual consulting in developing environmentally sustainable health care: systematic literature review. J Med Internet Res 2023 May 3;25:e44823. [doi: 10.2196/44823] [Medline: 37133914]
36. Daniel M, Gordon M, Patricio M, et al. An update on developments in medical education in response to the COVID-19 pandemic: A BEME scoping review: BEME Guide No. 64. Med Teach 2021 Mar;43(3):253-271. [doi: 10.1080/0142159X.2020.1864310] [Medline: 33496628]
37. Fowler RA, Adhikari NKJ, Bhagwanjee S. Clinical review: critical care in the global context--disparities in burden of illness, access, and economics. Crit Care 2008;12(5):225. [doi: 10.1186/cc6984] [Medline: 19014409]

XSL•FO
RenderX

38. Posever N, Sehdev M, Sylla M, Mashar R, Mashar M, Abioye A. Addressing equity in global medical education during the COVID-19 pandemic: the Global Medical Education Collaborative. Acad Med 2021 Nov 1;96(11):1574-1579. [doi: 10.1097/ACM.0000000000004230] [Medline: 34261867]
39. Goldberg LR, Crocombe LA. Advances in medical education and practice: role of massive open online courses. Adv Med Educ Pract 2017;8(603):603-609. [doi: 10.2147/AMEP.S115321] [Medline: 28860891]
40. Cheng C, Papadakos J, Umakanthan B, et al. On the advantages and disadvantages of virtual continuing medical education: a scoping review. Can Med Educ J 2023 Jun;14(3):41-74. [doi: 10.36834/cmej.75681] [Medline: 37465741]
41. Bollinger RC, McKenzie-White J, Gupta A. Building a global health education network for clinical care and research. The benefits and challenges of distance learning tools. Lessons learned from the Hopkins Center for Clinical Global Health Education. Infect Dis Clin North Am 2011 Jun;25(2):385-398. [doi: 10.1016/j.idc.2011.02.006] [Medline: 21628053]
42. Jindal M, Bajal E, Singh P, Diwakar M, Arya C, Sharma K. Online education in covid-19: limitations and improvements. Presented at: 2021 IEEE 8th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON); Nov 11-13, 2021; Dehradun, India p. 1-5. [doi: 10.1109/UPCON52273.2021.9667605]
43. Jiang B, Li X, Liu S, Hao C, Zhang G, Lin Q. Experience of online learning from COVID-19: preparing for the future of digital transformation in education. Int J Environ Res Public Health 2022 Dec 14;19(24):16787. [doi: 10.3390/ijerph192416787] [Medline: 36554668]
44. Almendingen K, Morseth MS, Gjølstad E, Brevik A, Tørris C. Student's experiences with online teaching following COVID-19 lockdown: a mixed methods explorative study. PLoS ONE 2021;16(8):e0250378. [doi: 10.1371/journal.pone.0250378] [Medline: 34464386]
45. Clark AE, Nong H, Zhu H, Zhu R. Compensating for academic loss: online learning and student performance during the COVID-19 pandemic. China Econ Rev 2021 Aug;68:101629. [doi: 10.1016/j.chieco.2021.101629] [Medline: 36570716]
46. Gonzalez T, de la Rubia MA, Hincz KP, et al. Influence of COVID-19 confinement on students' performance in higher education. PLoS ONE 2020;15(10):e0239490. [doi: 10.1371/journal.pone.0239490] [Medline: 33035228]
47. Yoo SS, Hu PT, Gujar N, Jolesz FA, Walker MP. A deficit in the ability to form new human memories without sleep. Nat Neurosci 2007 Mar;10(3):385-392. [doi: 10.1038/nn1851] [Medline: 17293859]
48. Weingarten JA, Collop NA. Air travel: effects of sleep deprivation and jet lag. Chest 2013 Oct;144(4):1394-1401. [doi: 10.1378/chest.12-2963] [Medline: 24081353]
49. AlQhtani A, AlSwedan N, Almulhim A, et al. Online versus classroom teaching for medical students during COVID-19: measuring effectiveness and satisfaction. BMC Med Educ 2021 Aug 28;21(1):452. [doi: 10.1186/s12909-021-02888-1] [Medline: 34454493]
50. Mustafa A, Psarikidou K, Pranjol MZI. Beyond the COVID-19 pandemic: can online teaching reduce the carbon footprint of the internationalisation of UK higher education? IME 2022;1(2):85-96. [doi: 10.3390/ime1020011]
51. The climate action & innovation matrix: a guidebook to level up your climate strategy. Climate Collaborative. 2023. URL: https://assets.nationbuilder.com/climatecollaborative/pages/2645/attachments/original/1712071072/Climate-Action-Matrix-FINALv1.pdf?1712071072 [accessed 2025-06-22]
52. Emissions calculator. Travel Impact Model. 2024. URL: https://travelimpactmodel.org/lookup/flight [accessed 2025-06-22]

Abbreviations

CERTAIN: Checklist for Early Recognition and Treatment of Acute Illness and Injury
CO_2e: carbon dioxide equivalent
GHG: greenhouse gas
ICU: intensive care unit

Edited by A Mavragani; submitted 01.07.24; peer-reviewed by E Bluman, Y Yang; revised version received 12.05.25; accepted 12.05.25; published 31.07.25.

Please cite as:
Wang B, Zambrano CC, Nikravangolsefid N, Carvalhais RM, Niven A, Gajic O, Dong Y
Environmental Impact of Online Versus in-Person Critical Care Education Through the Carbon Footprint Analysis of the CERTAIN Program: Cross-Sectional Study
JMIR Form Res 2025;9:e63524
URL: https://formative.jmir.org/2025/1/e63524
doi:10.2196/63524

© Baiyong Wang, Claudia Castillo Zambrano, Nasrin Nikravangolsefid, Ricardo Carvalhais, Alexander Niven, Ognjen Gajic, Yue Dong. Originally published in JMIR Formative Research (https://formative.jmir.org), 31.7.2025. This is an open-access

XSL•FO
RenderX

article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Feasibility of a Digital Coaching Program for Improving Mental Well-Being and Emotional Intelligence: Pragmatic Retrospective Cohort Study

Nicholas Chalmer Peiper[1], PhD; Adam Pettitt[2], PhD; Bela Shah[3], JD; Oliver Attwood[3], MS; Ella Sivonen[3], LLM; Jeffrey Pfeffer[4], PhD

[1]Department of Epidemiology and Population Health, School of Public Health and Information Sciences, University of Louisville, 485 East Gray Street, Louisville, United States

[2]Center for Digital Mental Health, University of Oregon, Eugene, United States

[3]Meru Health, San Mateo, CA, United States

[4]Graduate School of Business, Stanford University, Stanford, United States

Corresponding Author:
Nicholas Chalmer Peiper, PhD
Department of Epidemiology and Population Health, School of Public Health and Information Sciences, University of Louisville, 485 East Gray Street, Louisville, United States

Abstract

Background: Within the past decade, digital coaching programs (DCPs) have emerged as an evidence-based modality to improve mental well-being and emotional intelligence (EI), although there is limited evidence in real-world contexts.

Objective: This pragmatic retrospective cohort study aims to determine the preliminary effectiveness of a DCP in improving mental well-being and EI within a real-world context. We hypothesized that there would be a significant increase in mental well-being and EI.

Methods: This study included 588 people who voluntarily enrolled in an 8-week, blended care DCP offered through their employers from October 2021 to August 2024. The DCP included routine check-ins and consultations with certified coaches. Participants completed the World Health Organization-Five Well-Being Index (WHO-5) at baseline and then weekly until the end of the program, as well as the Brief Emotional Intelligence Scale-10 (BEIS-10) at baseline and the end of the program. Multivariable linear mixed models examined changes in WHO-5 (biweekly) and BEIS-10 (pre-post) scores, adjusting for age, gender, program engagement, and program completion. Multivariable logistic regression models evaluated correlates of clinically meaningful improvements on the WHO-5 (ie, at least a 10-point improvement). We calculated a reliable change index (RCI) for the BEIS-10 and the proportion of participants meeting the RCI criterion from baseline to end of treatment.

Results: In multivariate linear mixed models adjusting for demographics and program characteristics, we observed a significant increase in WHO-5 scores (baseline $\bar{x}$=45.6; week 8 $\bar{x}$=66.3; Cohen's *d*=1.98; *P*<.001). Over half of the sample (55.4%) experienced a clinically meaningful improvement on the WHO-5. Multivariable logistic regression found that higher engagement was associated with an increased odds of a clinically meaningful improvement on the WHO-5 (adjusted odds ratio [aOR] 1.002, 95% CI 1.001 - 1.003), while program noncompletion (aOR 0.27, 95% CI 0.15 - 0.50) and higher baseline well-being (aOR 0.91, 95% CI 0.89 - 0.92) were associated with reduced odds. BEIS-10 scores also significantly increased from baseline to the end of the program after adjusting for relevant correlates (baseline $\bar{x}$=37.6; week 8 $\bar{x}$=41.2; Cohen's *d*=1.32; *P*<.001). The estimated RCI on the BEIS-10 was approximately 5, with 19.7% experiencing a meaningful improvement.

Conclusions: These results demonstrate that DCPs can be a viable option for individuals looking to improve their mental well-being. Additional efforts should focus on establishing reliable change metrics for EI measures. Studies using hybrid effectiveness-implementation trial designs are now needed to further evaluate the real-world effectiveness of this program.

(JMIR Form Res 2025;9:e71828) doi:10.2196/71828

KEYWORDS

mental health services; implementation science; epidemiology; digital health; quality of life; emotional intelligence

XSL•FO
RenderX

Introduction

The United States is currently experiencing an unprecedented mental health crisis that was exacerbated by the acute phase of the COVID-19 pandemic [1,2]. US federal surveillance systems estimate that over 40% of adults are experiencing clinically significant levels of anxiety and depression symptoms in the past week [2]. The pervasiveness of mental health problems at the population level contributes to significant functional impairments, reduced workplace productivity, and disability [3-5], even among those experiencing serious psychological distress (SPD) who do not necessarily meet a clinical threshold for a formal diagnosis [6]. Moreover, research demonstrates that anxiety, depression, and SPD are highly comorbid with unhealthy behaviors and a variety of medical conditions [7,8]. As such, comprehensive prevention and treatment efforts are urgently needed to address population-level mental health problems.

Unfortunately, the US mental health care system has been experiencing significant structural barriers for over a decade [9]. The combination of stigma associated with mental health problems, inadequate reimbursement by health insurers, and a shortage of mental health providers has resulted in mental health being inadequately addressed [9,10]. It is estimated that only two-thirds of adults with depression receive any care each year, with only a small fraction of those receiving care getting high-quality, evidence-based treatment [11,12]. Even before the onset of the COVID-19 pandemic, long wait times for mental health care were common. This has been exacerbated by widespread shortages in the mental and behavioral health workforce [13]. Given the magnification of these structural barriers due to the pandemic, there remains an urgent need for new mental health care models.

Digital mental health interventions (DMHI) have recently emerged as a promising model of mental health care. DMHIs have been shown to address a number of structural barriers to care, including access, cost, and stigma [14,15]. Recent meta-analyses have found that DMHIs incorporating evidence-based modalities like cognitive behavioral therapy (CBT), mindfulness, and behavioral activation are as effective as traditional CBT in reducing depression and anxiety symptoms [16]. Because some DMHIs focus more on the treatment of clinically significant mental health symptoms rather than promoting mental well-being, they may not be the best fit for people experiencing SPD and other subclinical problems [17,18].

Recently, digital coaching programs (DCP) have emerged as viable interventions grounded in behavioral science principles to help people with subclinical mental health problems improve their mental well-being, reduce stress, and prevent the onset of mental disorders [19]. DCPs function to help people improve their well-being by building skills that increase emotional intelligence (EI), resilience, and self-actualization. Yet, studies evaluating DCPs and other coaching programs frequently lack an operational definition of what techniques were included in the interventions [20,21]. Moreover, information about the effectiveness of DCPs in real-world settings remains limited [22]. Given these gaps in knowledge and the importance of early intervention to prevent the onset of mental health problems, there is an urgent need for studies that investigate how DCPs may improve mental health and well-being.

This study evaluated the preliminary effectiveness of an 8-week DCP implemented in a real-world setting to improve mental well-being and increase EI. The primary hypothesis was that participation in the DCP would be associated with increased mental well-being and EI after adjusting for demographic and program-related factors. Exploratory analyses calculated reliable change scores for the EI measure and examined factors associated with meaningful improvements in EI.

Methods

Study Sample and Procedures

Program participants were enrolled in a DCP through a provider of employer-sponsored mental health care (Meru Health) between October 1, 2021, and August 19, 2024. Participants were offered the DCP through their employer. During this period, there were 45 employers offering the DCP to approximately 100,000 people. Interested participants could schedule an initial screening assessment through a digital platform. As part of this process, potential participants completed the Patient Health Questionnaire-2 and Generalized Anxiety Disorder-2 to report on depression and anxiety symptoms, respectively. To be eligible, all participants were required to have a mobile phone, be at least 18 years old, and have scores of less than 3 on the Patient Health Questionnaire-2 and Generalized Anxiety Disorder-2 [23,24]. Those reporting current or recent suicidality, substance use disorders, eating disorders, or psychotic disorders were ineligible for the program. After the screening assessment, eligible participants participated in an intake video call with a certified coach to develop a personalized care plan. Those not meeting program eligibility requirements were provided information about the 12-week treatment program that is designed for clinically significant levels of depression and anxiety symptoms.

Ethical Considerations

This study was deemed exempt by the Pearl Institutional Review Board (21-MERU-114) for secondary analyses of previously collected and deidentified data from a program evaluation. As part of the DCP, participants consented to have their anonymized data used for evaluation and quality improvement purposes. We conducted a retrospective analysis with program data after the DCP took place. All data and protected health information were stored in a cloud storage environment compliant with the Health Insurance Portability and Accountability Act. All data were encrypted in transit, end-to-end, and at rest. Deidentified data from this study are not available in a public archive or for distribution to ensure privacy of the program participants.

Program Description

The Meru Health DCP is a blended care intervention designed to improve well-being among people experiencing subclinical problems through a mobile phone app over an 8-week period, including face-to-face sessions, skills-building content, regular chat engagement, and digital workshops. The intervention is grounded in Daniel Goleman's EI Framework, which posits

XSL•FO
RenderX

that mental well-being and functioning may be increased by building competencies in 4 core domains of emotional intelligence: self-awareness, self-regulation, social awareness and empathy, and relationship management skills [25,26]. Each week focuses on building competencies in the 4 core EI domains, with the first 4 weeks focusing on individual-level skills and the latter 4 weeks focusing on relational skills (Multimedia Appendix 1). Participants receive a total of 32 learning videos, 40 CBT exercises, and 24 meditations spread evenly across the 8-week program. In rare situations, a participant will be referred to the therapist-supported program if they determine with their coach that therapy will be more appropriate.

The certified coaches provided personalized, daily chat support to encourage and motivate participants. Coaches are certified (Emotional Intelligence Coaching Certification, International Coaching Federation, National Board Certified Health & Wellness Coach, and Certified International Health Coach) and have training in EI, CBT, motivational interviewing, and mindfulness-based stress reduction. Participants were able to communicate with coaches via a chat support function and regular face-to-face sessions. After their intake session, on average, participants spent about 45 minutes per week on the app and engaging with their coach on various activities. Coaches used a secure, web-based dashboard developed by Meru Health to communicate with participants and monitor participant engagement metrics.

Measures

The primary outcome was mental well-being, which was measured with the World Health Organization-5 Well-Being Index (WHO-5). The WHO-5 is a psychometrically validated screening instrument that assesses mental well-being in the past 2 weeks [27]. There are a total of 5 items that are scored on a 0 - 5 Likert scale, with raw scores ranging from 0 to 25. A composite score ranging from 0 to 100 is calculated by multiplying the raw score by 4. Higher scores indicate better mental well-being. A score of 50 or less indicates poor mental well-being and suggests further clinical investigation. Participants completed the WHO-5 on a biweekly basis during the DCP. The secondary outcome was EI and measured with the Brief Emotional Intelligence Scale-10 (BEIS-10), which has established validity and reliability [28]. There are a total of 10 items scored on a 1 - 5 Likert scale, with scores ranging from 10 to 50. The BEIS-10 captures dimensions of self-appraisal of emotions, appraisal of others' emotions, self-regulation of emotions, regulation of others' emotions, and use of emotions. Higher scores indicate better EI. Participants completed the BEIS-10 at baseline and week 8.

Program-related measures included engagement and completion. Engagement was defined as the total number of minutes participants used the app during the program. Completion was defined as the proportion of participants who actively engaged in the program for at least 4 weeks [29]. Demographics included age and gender identity (male and female). The version of the DCP (versions 1 and 2) was also included as a covariate to adjust for any potential effects of intervention modifications.

Statistical Analyses

We plotted the average level of engagement during the intervention and fit a mixed-effects model using the *lme4* package to examine whether there was a significant change during the intervention [30]. A mixed-effects model was also used to explore significant changes in BEIS-10 scores pre- and post-intervention. Both models used age, gender, engagement, program completion, and program version as covariates. Cohen's *d* was used to assess effect sizes in pre- and post-program WHO-5 and BEIS-10 scores. Effect sizes of 0.2, 0.5, and 0.8 were considered small, medium, and large effect sizes, respectively [31].

To determine clinically significant improvements in mental well-being and emotional intelligence [32], we calculated the proportion of participants experiencing at least a 10-point improvement on the WHO-5 from baseline to the end of the intervention [33]. Because there is no documented threshold for clinically significant improvement on the BEIS-10, we calculated a reliable change index (RCI) for the BEIS-10 and the proportion of participants meeting the RCI criterion from baseline to end of treatment [34]. We also fit a multivariable logistic regression model to determine the association between clinically significant improvement in mental well-being with program-related factors and demographics while controlling for baseline well-being, EI, and the program version.

All statistical analyses were conducted with RStudio (version 1.3.959). Statistical significance was defined as a 2-tailed P value≤ .05. Our analytical approach used intent-to-treat analyses whereby all participants who enrolled in the program were included regardless of engagement or completion [35]. Full information maximum likelihood was used to account for missing data within the mixed-effects models, and multiple imputation with chained equations was used to account for missingness in the modeling procedures for clinically significant improvements [36,37].

Results

Overall, there were 588 people who enrolled in the Meru Health DCP between January 1, 2022, and August 19, 2024. As illustrated in Table 1, the majority of participants were female (77%), and the average age was approximately 42 (SD 11) years. The average WHO-5 score at baseline was 43 (SD 18), and the average BEIS-10 score was 37.5 (SD 4.1). A large majority of the sample (77%) participated in the DCP for at least 4 out of 8 weeks, and the total number of minutes active during the DCP was 367 (SD 343). The estimated marginal mean (EMM) number of minutes participants were active in the DCP app was 49.15 (95% CI 44.1 - 54.2) during the first week, then decreased in a linear fashion to 7.67 (95% CI 2.61 - 12.7) minutes at week 8 (Figure 1). No significant differences between program versions were found in the demographic characteristics and engagement.

XSL•FO
RenderX

Table . Characteristics of participants in an 8-week digital coaching program (N=588).

Characteristic	Overall	V1[a] program (n=96)	V2 program (n=492)	*P* value[b]
Completion, n/N (%)				.049
Completer	450/588 (77)	66/96 (69)	384/492 (78)	
Dropout	138/588 (23)	30/96 (31)	108/492 (22)	
Age (years), mean (SD)	42 (11)	42 (11)	42 (11)	.70
Gender, n/N (%)				.50
Female	451/588 (77)	76/96 (79)	375/492 (76)	
Male	137/588 (23)	20/96 (21)	117/492 (24)	
Total minutes active, mean (SD)	367 (343)	212 (178)	398 (359)	<.001
Baseline WHO-5[c], mean (SD)	43 (18)	49 (16)	42 (18)	<.001
Baseline BEIS-10[d], mean (SD)	37.5 (4.1)	38.1 (4.1)	37.4 (4.1)	.11

[a]V1 and V2 refer to the 2 versions of the digital coaching program.

[b]Pearson chi-square test and Wilcoxon rank sum test.

[c]WHO-5: World Health Organization-5 Well-Being Index.

[d]BEIS-10: Brief Emotional Intelligence Scale-10.

Figure 1. Patterns of engagement during an 8-week digital coaching program (N=588).

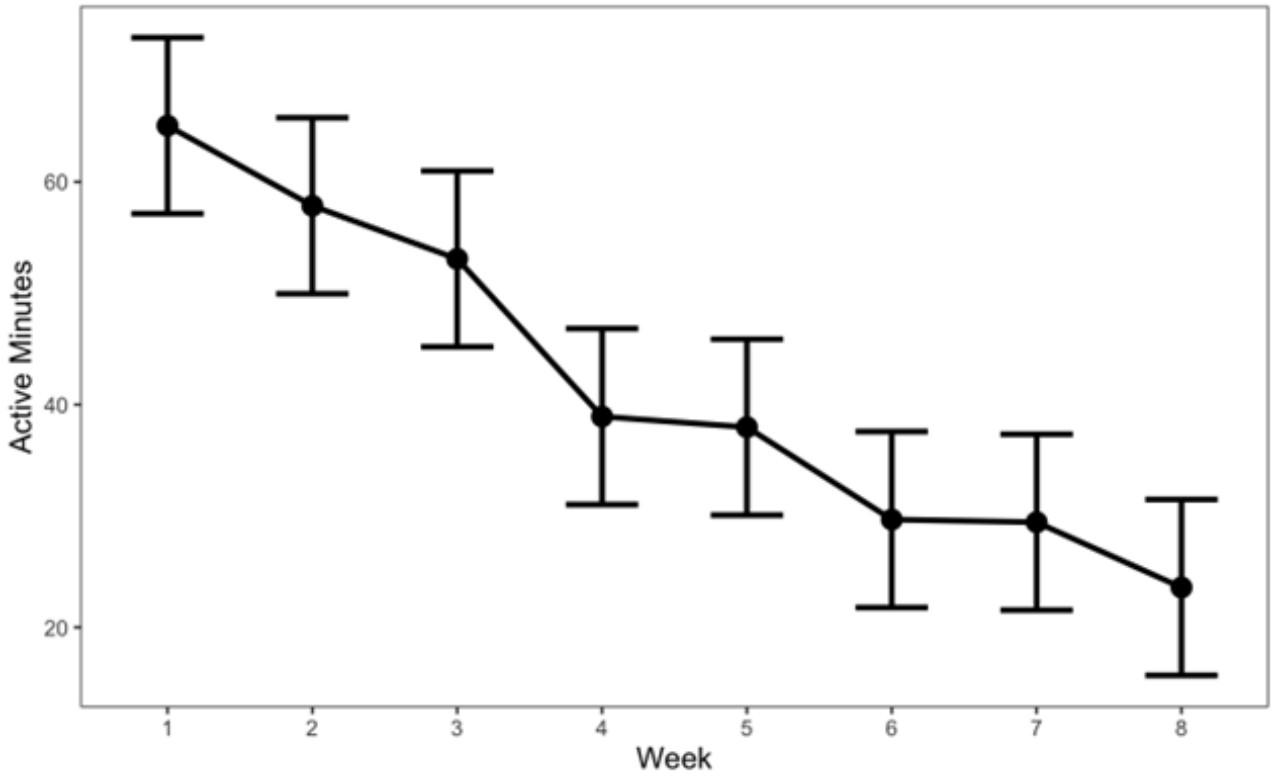

Mental Well-Being Outcomes

We conducted mixed-effects modeling to determine WHO-5 outcomes across the coaching program. Figure 2 and Table 2 depict the results of the linear mixed-effects model to determine changes in WHO-5 scores during the DCP. At baseline, the EMM for the WHO-5 was 45.6 (95% CI 43.2 - 48.0), which met the established clinical threshold of 50 or less for poor mental well-being requiring clinical attention. At the end of the program, a significant increase in the WHO-5 score was observed (EMM 66.3, 95% CI 63.6 - 69.0). This represented a large improvement (Cohen's *d*=1.98, 95% CI 1.81 - 2.15) that brought participants' average WHO-5 scores outside of the clinical threshold of 50 or less. There were 55.4% who experienced a clinically meaningful improvement on the WHO-5

XSL•FO
RenderX

(ie, at least a 10-point improvement). Figure 3 shows the results from the multivariable logistic regression model examining correlates of clinically meaningful improvements in mental well-being. Higher engagement was associated with an increased odds of a clinically meaningful improvement (adjusted odds ratio [aOR] 1.002, 95% CI 1.001 - 1.003). That is, the odds of a clinically meaningful improvement increased by 0.2% for every unit increase in engagement. Program noncompletion (aOR 0.27, 95% CI 0.15 - 0.50) and higher baseline well-being (aOR 0.91, 95% CI 0.89 - 0.92) were associated with reduced odds of a clinically meaningful improvement.

Figure 2. Change in WHO-5 scores during an 8-week digital coaching program (N=588). Error bands indicate 95% CIs. WHO-5 scores range from 0 to 100, with higher scores indicating better mental well-being. Scores of 50 or less indicate poor mental health requiring clinical attention. WHO-5: World Health Organization-Five Well-Being Index.

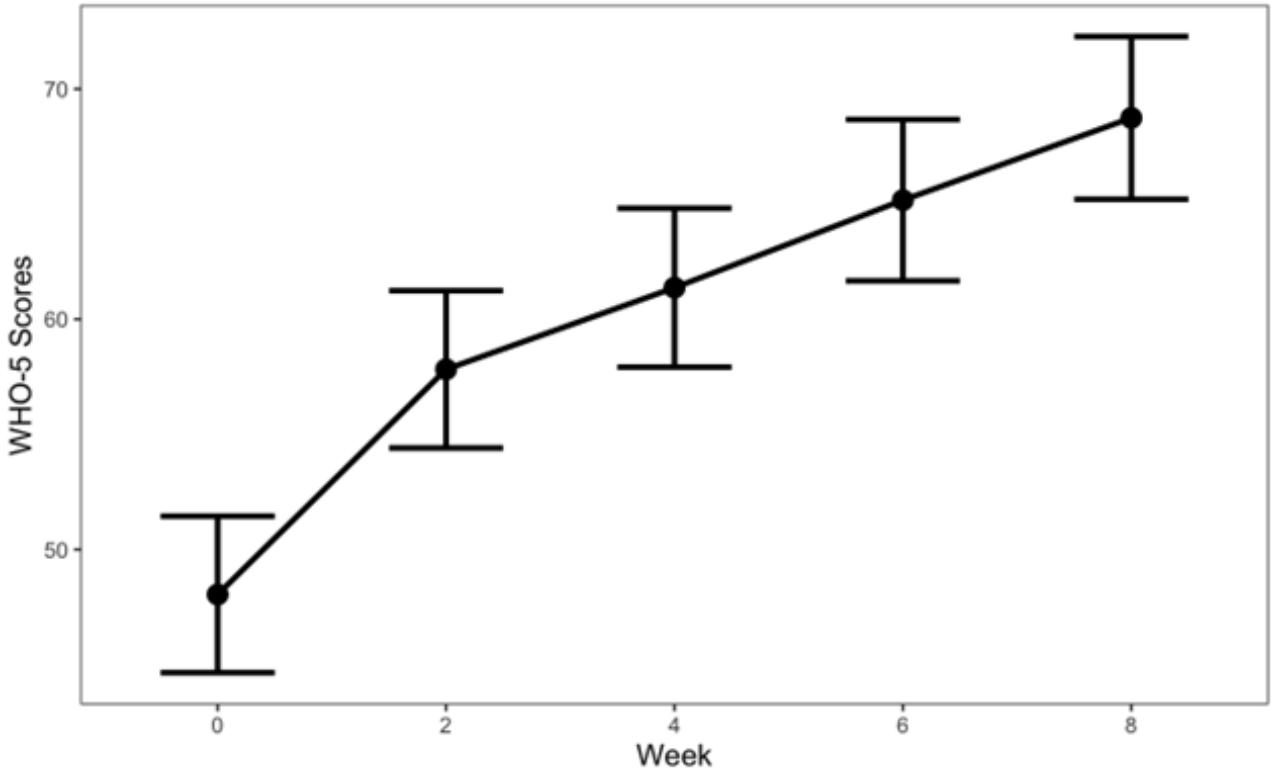

Table . Multivariable mixed-effects model results for WHO-5[a] scores across time among participants in an 8-week digital coaching program (N=588). Reference groups are included parenthetically. For intervention weeks, the baseline assessment served as the reference group.

Predictors	WHO-5[a]				
	Estimates	Standardized beta	95% CI	Standardized 95% CI	*P* value
Intercept	47.35	−0.45	37.94 - 56.77	−0.66 to−0.24	<.001
Week 2	7.61	0.46	5.00 - 10.22	0.30 - 0.62	<.001
Week 4	9.03	0.55	5.98 - 12.08	0.36 - 0.73	<.001
Week 6	15.45	0.93	12.10 - 18.80	0.73 - 1.14	<.001
Week 8	17.37	1.05	13.80 - 20.94	0.84 - 1.27	<.001
Age	0.02	0.01	−0.20 to 0.24	−0.13 to 0.16	.87
Male gender (female)	−0.63	−0.04	−5.94 to 4.67	−0.36 to 0.28	.82
Total minutes active	0.00	0.06	−0.01 to 0.01	−0.09 to 0.22	.43
Version 2 (version 1)	0.07	0.00	−4.93 - 5.07	−0.30 - 0.31	.98

[a]WHO-5: World Health Organization-5 Well-Being Index.

XSL•FO
RenderX

Figure 3. Multivariate associations with clinically significant improvement in mental well-being (N=588) Program noncompletion was defined as engaging in the program for less than 4 weeks. Reference groups are included parenthetically. Total active minutes, age, baseline WHO-5, and baseline BEIS-10 do not show visible CIs due to their extremely narrow ranges relative to other predictors. * indicates continuous variables. WHO-5: World Health Organization-Five Well-Being Index; BEIS-10: Brief Emotional Intelligence Scale-10.

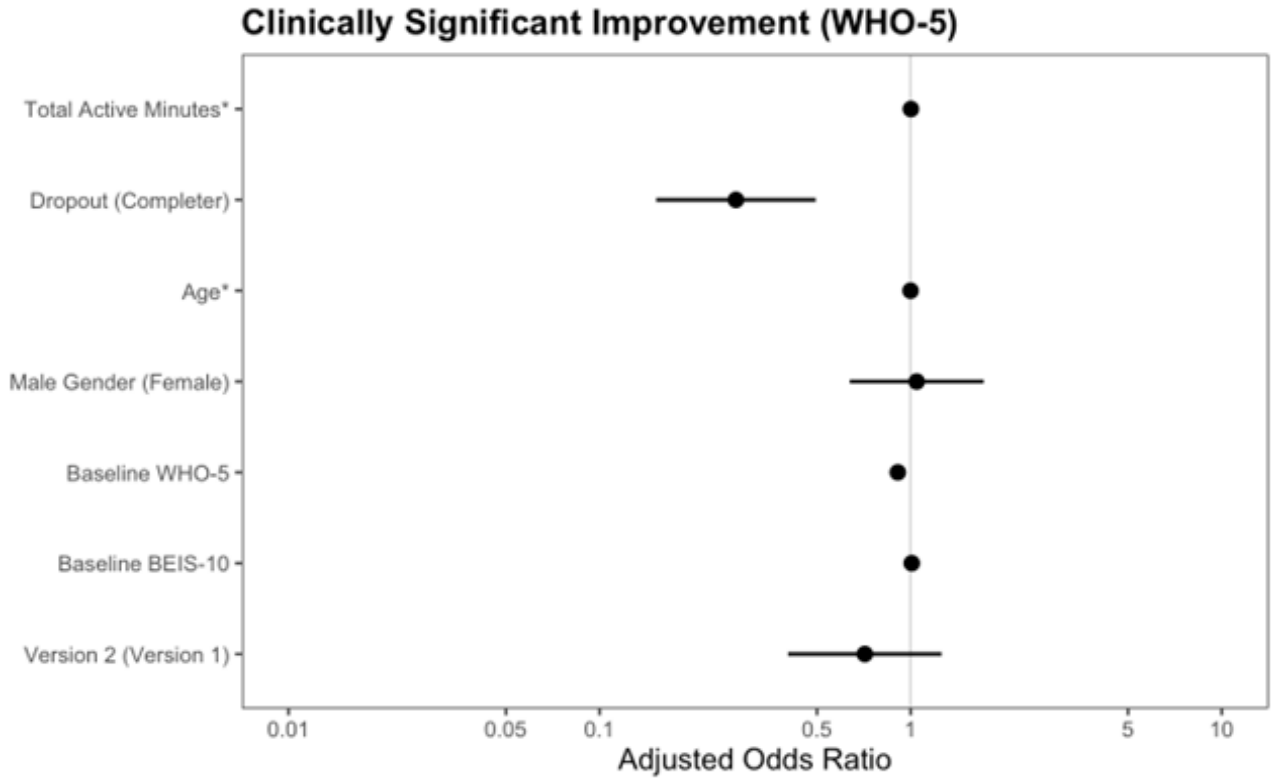

Emotional Intelligence Outcomes

We also fit a separate mixed effects model with BEIS-10 scores as the outcome to determine how these scores changed from baseline to the end of treatment. Table 3 and Multimedia Appendix 2 depict the results of the model and demonstrate that from baseline to end of treatment, there was a significant increase in BEIS-10 scores. The EMM for baseline was 37.6 (95% CI 37.0 - 38.1), while at the end of treatment, the EMM was 41.2 (95% CI 40.5 - 41.9). This represented a large effect size (Cohen's *d*=1.32, 95% CI 1.14 - 1.50). Finally, we found an RCI of approximately 5 for the BEIS-10. Based on these criteria, 19.7% of individuals experienced a meaningful improvement in EI during the DCP.

Table . Multivariable mixed-effects model results for BEIS-10[a] scores across time among participants in an 8-week digital coaching program (N=588). Reference groups are included parenthetically.

Predictors	BEIS-10[a]				
	Estimates	Standardized beta	95% CI	Standardized 95% CI	*P* value
Intercept	38.92	−0.15	36.55 - 41.39	−0.36 to 0.06	<.001
Week 8 (baseline)	3.80	0.94	2.81 - 4.80	0.69 - to 1.18	<.001
Age	−0.02	−0.04	−0.07 to 0.04	−0.19 to 0.11	.59
Male gender (female)	−0.85	−0.21	−2.26 to 0.56	−0.56 to 0.14	.24
Total minutes active	0.00	0.02	−0.00 to 0.00	−0.15 to 0.19	.79
Version 2 (version 1)	−0.38	−0.09	−1.71 to 0.94	−0.42 to 0.23	.59

[a]BEIS-10: Brief Emotional Intelligence Scale-10.

Discussion

This study investigated the preliminary effectiveness of an 8-week DCP among 588 people with subclinical mental health problems. Consistent with our primary hypothesis, we observed that participation in the DCP was associated with large increases in mental well-being as measured by the WHO-5 (Cohen's *d*=1.98). Similarly, we found that participants who did not complete the program (less than 4 weeks) had 73% lower odds of experiencing a clinically meaningful improvement in mental well-being (ie, at least a 10-point improvement on the WHO-5) compared to program completers. Exploratory analyses also revealed large increases in EI during the DCP (Cohen's *d*=1.32) and that nearly a fifth of participants (19.7%) experienced a

XSL•FO
RenderX

meaningful improvement based upon an estimated RCI of 5 on the BEIS-10 [34].

Overall, this study demonstrates that participation in a blended care DCP was associated with improvements in mental well-being among a sample of 588 adults. The results are consistent with other research studies and support the further investigation of DCPs as viable interventions to prevent the onset of common mental health problems like depression and anxiety [38]. For example, a study evaluating a blended care cognitive behavioral program delivered via video observed large improvements in depression (Cohen's d=1.08) and anxiety (Cohen's d=1.33) [39], which strongly aligns with the improvements in mental well-being (Cohen's d=1.98) and EI (Cohen's d=1.32) found in this study. Similarly, a randomized controlled trial of 146 primary care patients found that those randomized to a self-guided DCP were more likely to experience clinically significant improvements in depression and anxiety compared to a waitlist control [40]. In this study, there were large, statistically significant increases in mental well-being during the program, although the inverse association between baseline WHO-5 scores and clinically significant improvement suggests that the program had the greatest effect on participants who started with the lowest mental well-being. Thus, our results provide additional evidence that DCPs targeting people with subclinical distress may help reduce the onset of mental disorders that cause significant disability and role impairments [4,5].

Another notable finding was the relatively higher levels of completion (77%) compared to other studies. In particular, a dose-response relationship between engagement and clinically meaningful improvements in mental well-being was observed. It is possible that the continuous involvement of trained coaches over the course of the program led to higher engagement and more positive outcomes compared to self-guided or text-based DCPs [41-43]. Direct communications and involvement of trained coaches may have facilitated a remote, continuous care approach that establishes trust with participants at program intake and fosters a therapeutic alliance over the duration of the program [44,45]. This approach may have influenced the higher rate of completion observed in this study compared to other DCPs [39,46], although additional studies are needed to better evaluate how interactions with coaches (eg, frequency of coach interactions, average amount of coach time per participant) may impact adherence and program outcomes [45].

Several limitations are acknowledged. The lack of an experimental design precludes inferences about the DCP's effectiveness in improving mental well-being and EI compared to other forms of care. Without a control group, our results are generalizable to other observational studies that cannot fully rule out natural remission of the documented changes in well-being and EI. Because participants voluntarily agreed to participate in an employer-sponsored program, we cannot rule out motivation to improve mental health and EI as an alternative explanation for the results. While self-selection may have impacted our results, the dose-response effect we observed in a real-world context is consistent with meaningful improvements in mental well-being and EI found in experimental and observational studies [18,20,39,46]. Similarly, the modest sample size represents another limitation to the study's external validity, although our multivariate analyses were robust and highly comparable to other studies. In addition, this DCP only captured age and gender to reduce participant burden during intake, limiting inferences about race, ethnicity, and other key demographic categories. Finally, our approach to addressing missing data raises the potential for methodological artifacts. These methods assume that data for participants with missing responses on the WHO-5 and BEIS-10 may be predicted with data from those who provided responses. Because those with poorer outcomes are less likely to respond, outcomes imputed based upon the observed characteristics of participants may include inaccurate assumptions. Nonetheless, we used an intent-to-treat framework and rigorous missing data approaches rather than performing listwise deletion and conducting completer-only analyses that can lead to selection biases and inflated effect sizes [35-37,47].

Our findings indicated that participation in an 8-week DCP was associated with large improvements in mental well-being and EI. Follow-up studies with larger samples are now needed to replicate these findings and investigate the effectiveness of the DCP in preventing common mental health disorders. Hybrid implementation-effectiveness trials are also needed to generate stronger evidence of the real-world effectiveness of DCPs [48,49]. In addition, mixed methods approaches may help elucidate mechanisms related to acceptability, satisfaction, and participant-coach interactions [50]. In sum, this study provides preliminary evidence that DCPs may be viable, evidence-based interventions for improving mental well-being at the population level.

Data Availability

The datasets generated or analyzed during this study are not available in a public archive or through request to ensure privacy of the program participants.

Conflicts of Interest

AP, BS, OA, and ES report employment at Meru Health during the conduct of this study. NCP reports scientific consultancy fees from Meru Health and research support from the CDC Foundation, National Institutes of Health, Abbott Laboratories, and Gilead Sciences. JP has no financial disclosures or conflicts of interest. The content is solely the responsibility of the authors and does not necessarily represent the official views of Meru Health. All decisions about the research were made by the authors and were unrestricted.

XSL•FO
RenderX

Multimedia Appendix 1
Weekly Themes and Competencies of an 8-Week Digital Coaching Program.
[DOCX File, 14 KB - formative_v9i1e71828_app1.docx]

Multimedia Appendix 2
Change in BEIS-10 Scores During an 8-Week Digital Coaching Program (N=588).
[PNG File, 5 KB - formative_v9i1e71828_app2.png]

References

1. Kola L, Kumar M, Kohrt BA, Fatodu T, Olayemi BA, Adefolarin AO. Strengthening public mental health during and after the acute phase of the COVID-19 pandemic. Lancet 2022 May 14;399(10338):1851-1852. [doi: 10.1016/S0140-6736(22)00523-2] [Medline: 35339230]
2. Vahratian A, Blumberg SJ, Terlizzi EP, Schiller JS. Symptoms of anxiety or depressive disorder and use of mental health care among adults during the COVID-19 pandemic—United States, August 2020–February 2021. MMWR Morb Mortal Wkly Rep 2020;70(13):490-494. [doi: 10.15585/mmwr.mm7013e2]
3. Beck A, Crain AL, Solberg LI, et al. Severity of depression and magnitude of productivity loss. Ann Fam Med 2011;9(4):305-311. [doi: 10.1370/afm.1260] [Medline: 21747101]
4. Hasin DS, Sarvet AL, Meyers JL, et al. Epidemiology of adult DSM-5 major depressive disorder and Its specifiers in the United States. JAMA Psychiatry 2018 Apr 1;75(4):336-346. [doi: 10.1001/jamapsychiatry.2017.4602] [Medline: 29450462]
5. Hendriks SM, Spijker J, Licht CMM, et al. Long-term work disability and absenteeism in anxiety and depressive disorders. J Affect Disord 2015 Jun 1;178:121-130. [doi: 10.1016/j.jad.2015.03.004] [Medline: 25805404]
6. Olfson M, Wang S, Wall M, Marcus SC, Blanco C. Trends in serious psychological distress and outpatient mental health care of US Adults. JAMA Psychiatry 2019 Feb 1;76(2):152-161. [doi: 10.1001/jamapsychiatry.2018.3550] [Medline: 30484838]
7. Bierman AS, Tinetti ME. Precision medicine to precision care: managing multimorbidity. Lancet 2016 Dec 3;388(10061):2721-2723. [doi: 10.1016/S0140-6736(16)32232-2] [Medline: 27924764]
8. Nelson BW, Peiper NC, Aschbacher K, Forman-Hoffman VL. Evidence-based therapist-supported digital mental health intervention for patients experiencing medical multimorbidity: a retrospective cohort intent-to-treat study. Psychosom Med 2024;86(6):547-554. [doi: 10.1097/PSY.0000000000001319] [Medline: 38718176]
9. Alegría M, Alvarez K, Cheng M, Falgas-Bague I. Recent advances on social determinants of mental health: looking fast forward. Am J Psychiatry 2023 Jul 1;180(7):473-482. [doi: 10.1176/appi.ajp.20230371] [Medline: 37392038]
10. Peterson E, Busch S. Achieving mental health and substance use disorder treatment parity: a quarter century of policy making and research. Annu Rev Public Health 2018 Apr 1;39:421-435. [doi: 10.1146/annurev-publhealth-040617-013603] [Medline: 29328871]
11. Nagata JM, Ganson KT, Bonin SL, et al. Prevalence and sociodemographic correlates of unmet need for mental health counseling among adults during the COVID-19 pandemic. Psychiatr Serv 2022 Feb 1;73(2):206-209. [doi: 10.1176/appi.ps.202100111] [Medline: 34189929]
12. Walker ER, Cummings JR, Hockenberry JM, Druss BG. Insurance status, use of mental health services, and unmet need for mental health care in the United States. Psychiatr Serv 2015 Jun;66(6):578-584. [doi: 10.1176/appi.ps.201400248] [Medline: 25726980]
13. Erikson C, Schenk E, Westergaard S, Salsberg ES. New behavioral health workforce database paints a stark picture. Health Affairs Forefront 2022. [doi: 10.1377/forefront.20220829.640971]
14. Connolly SL, Kuhn E, Possemato K, Torous J. Digital clinics and mobile technology implementation for mental health care. Curr Psychiatry Rep 2021 May 7;23(7):38. [doi: 10.1007/s11920-021-01254-8] [Medline: 33961135]
15. Lipschitz JM, Van Boxtel R, Torous J, et al. Digital mental health interventions for depression: scoping review of user engagement. J Med Internet Res 2022 Oct 14;24(10):e39204. [doi: 10.2196/39204] [Medline: 36240001]
16. Goldberg SB, Lam SU, Simonsson O, Torous J, Sun S. Mobile phone-based interventions for mental health: a systematic meta-review of 14 meta-analyses of randomized controlled trials. PLOS Digit Health 2022;1(1):e0000002. [doi: 10.1371/journal.pdig.0000002] [Medline: 35224559]
17. Karyotaki E, Efthimiou O, Miguel C, et al. Internet-based cognitive behavioral therapy for depression: a systematic review and individual patient data network meta-analysis. JAMA Psychiatry 2021 Apr 1;78(4):361-371. [doi: 10.1001/jamapsychiatry.2020.4364] [Medline: 33471111]
18. Reins JA, Buntrock C, Zimmermann J, et al. Efficacy and moderators of internet-based interventions in adults with subthreshold depression: an individual participant data meta-analysis of randomized controlled trials. Psychother Psychosom 2021;90(2):94-106. [doi: 10.1159/000507819] [Medline: 32544912]
19. Hrynyschyn R, Hafner J, Schönfeld S, Dockweiler C. Digital health interventions in prevention, relapse, and therapy of mild and moderate depression: scoping review. JMIR Ment Health 2021 Apr 16;8(4):e26268. [doi: 10.2196/26268] [Medline: 33861201]

XSL•FO
RenderX

20. Ammentorp J, Uhrenfeldt L, Angel F, Ehrensvärd M, Carlsen EB, Kofoed PE. Can life coaching improve health outcomes?—a systematic review of intervention studies. BMC Health Serv Res 2013 Oct 22;13(1):428. [doi: 10.1186/1472-6963-13-428] [Medline: 24148189]
21. Newnham-Kanas C, Gorczynski P, Morrow D, Irwin J. Annotated bibliography of life coaching and health research. Int J Evid Based Coach Mentor 2009;7(1):39-103 [FREE Full text]
22. Theeboom T, Beersma B, van Vianen AEM. Does coaching work? A meta-analysis on the effects of coaching on individual level outcomes in an organizational context. J Posit Psychol 2014 Jan 2;9(1):1-18. [doi: 10.1080/17439760.2013.837499]
23. Löwe B, Kroenke K, Gräfe K. Detecting and monitoring depression with a two-item questionnaire (PHQ-2). J Psychosom Res 2005 Feb;58(2):163-171. [doi: 10.1016/j.jpsychores.2004.09.006] [Medline: 15820844]
24. Plummer F, Manea L, Trepel D, McMillan D. Screening for anxiety disorders with the GAD-7 and GAD-2: a systematic review and diagnostic metaanalysis. Gen Hosp Psychiatry 2016;39:24-31. [doi: 10.1016/j.genhosppsych.2015.11.005] [Medline: 26719105]
25. Boyatzis RE, Goleman D, Dhar U, Osiri JK. Thrive and survive: assessing personal sustainability. Consulting Psychology Journal: Practice and Research 2021;73(1):27-50. [doi: 10.1037/cpb0000193]
26. Goleman D. What makes a leader? Harv Bus Rev 1998;76(6):93-102 [FREE Full text] [Medline: 10187249]
27. Topp CW, Østergaard SD, Søndergaard S, Bech P. The WHO-5 Well-Being Index: a systematic review of the literature. Psychother Psychosom 2015;84(3):167-176. [doi: 10.1159/000376585] [Medline: 25831962]
28. Davies KA, Lane AM, Devonport TJ, Scott JA. Validity and reliability of a Brief Emotional Intelligence Scale (BEIS-10). J Individ Differ 2010 Jan;31(4):198-208. [doi: 10.1027/1614-0001/a000028]
29. Erekson DM, Lambert MJ, Eggett DL. The relationship between session frequency and psychotherapy outcome in a naturalistic setting. J Consult Clin Psychol 2015 Dec;83(6):1097-1107. [doi: 10.1037/a0039774] [Medline: 26436645]
30. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. J Stat Softw 2015;67(1):1-48. [doi: 10.18637/jss.v067.i01]
31. Fritz CO, Morris PE, Richler JJ. Effect size estimates: current use, calculations, and interpretation. J Exp Psychol Gen 2012 Feb;141(1):2-18. [doi: 10.1037/a0024338] [Medline: 21823805]
32. Kounali D, Button KS, Lewis G, et al. How much change is enough? Evidence from a longitudinal study on depression in UK primary care. Psychol Med 2022 Jul;52(10):1875-1882. [doi: 10.1017/S0033291720003700] [Medline: 33138872]
33. Bech P, Lindberg L, Moeller SB. The reliable change index (RCI) of the WHO-5 in primary prevention of mental disorders. A measurement-based pilot study in positive psychiatry. Nord J Psychiatry 2018 Aug;72(6):404-408. [doi: 10.1080/08039488.2018.1463393] [Medline: 29656671]
34. Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. J Consult Clin Psychol 1991;59(1):12-19. [doi: 10.1037/0022-006X.59.1.12]
35. Ten Have TR, Normand SLT, Marcus SM, Brown CH, Lavori P, Duan N. Intent-to-treat vs. non-intent-to-treat analyses under treatment non-adherence in mental health randomized trials. Psychiatr Ann 2008 Dec;38(12):772-783. [doi: 10.3928/00485713-20081201-10] [Medline: 20717484]
36. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? Int J Methods Psychiatr Res 2011 Mar;20(1):40-49. [doi: 10.1002/mpr.329] [Medline: 21499542]
37. Matta TH, Flournoy JC, Byrne ML. Making an unknown unknown a known unknown: missing data in longitudinal neuroimaging studies. Dev Cogn Neurosci 2018 Oct;33:83-98. [doi: 10.1016/j.dcn.2017.10.001] [Medline: 29129673]
38. Grant F, Guille C, Sen S. Well-being and the risk of depression under stress. PLoS One 2013;8(7):e67395. [doi: 10.1371/journal.pone.0067395] [Medline: 23840872]
39. Lungu A, Jun JJ, Azarmanesh O, Leykin Y, Chen CEJ. Blended care-cognitive behavioral therapy for depression and anxiety in real-world settings: pragmatic retrospective study. J Med Internet Res 2020 Jul 6;22(7):e18723. [doi: 10.2196/18723] [Medline: 32628120]
40. Graham AK, Greene CJ, Kwasny MJ, et al. Coached mobile app platform for the treatment of depression and anxiety among primary care patients: a randomized clinical trial. JAMA Psychiatry 2020 Sep 1;77(9):906-914. [doi: 10.1001/jamapsychiatry.2020.1011] [Medline: 32432695]
41. Fleming T, Bavin L, Lucassen M, Stasiak K, Hopkins S, Merry S. Beyond the trial: systematic review of real-world uptake and engagement with digital self-help interventions for depression, low mood, or anxiety. J Med Internet Res 2018 Jun 6;20(6):e199. [doi: 10.2196/jmir.9275] [Medline: 29875089]
42. Karyotaki E, Kemmeren L, Riper H, et al. Is self-guided internet-based cognitive behavioural therapy (iCBT) harmful? An individual participant data meta-analysis. Psychol Med 2018 Nov;48(15):2456-2466. [doi: 10.1017/S0033291718000648] [Medline: 29540243]
43. Boucher EM, Raiker JS. Engagement and retention in digital mental health interventions: a narrative review. BMC Digit Health 2024;2(1):1-21. [doi: 10.1186/s44247-024-00105-9]
44. D'Alfonso S, Lederman R, Bucci S, Berry K. The digital therapeutic alliance and human-computer interaction. JMIR Ment Health 2020 Dec 29;7(12):e21895. [doi: 10.2196/21895] [Medline: 33372897]
45. Schiepek G, Aichhorn W, Gruber M, Strunk G, Bachler E, Aas B. Real-time monitoring of psychotherapeutic processes: concept and compliance. Front Psychol 2016;7:604. [doi: 10.3389/fpsyg.2016.00604] [Medline: 27199837]

XSL•FO
RenderX

46. Lungu A, Boone MS, Chen SY, Chen CE, Walser RD. Effectiveness of a cognitive behavioral coaching program delivered via video in real world settings. Telemed J E Health 2021 Jan;27(1):47-54. [doi: 10.1089/tmj.2019.0313] [Medline: 32311301]
47. Salim A, Mackinnon A, Christensen H, Griffiths K. Comparison of data analysis strategies for intent-to-treat analysis in pre-test-post-test designs with substantial dropout rates. Psychiatry Res 2008 Sep 30;160(3):335-345. [doi: 10.1016/j.psychres.2007.08.005] [Medline: 18718673]
48. Kessler RC, Luedtke A. Pragmatic precision psychiatry—a new direction for optimizing treatment selection. JAMA Psychiatry 2021 Dec 1;78(12):1384-1390. [doi: 10.1001/jamapsychiatry.2021.2500] [Medline: 34550327]
49. Curran GM, Bauer M, Mittman B, Pyne JM, Stetler C. Effectiveness-implementation hybrid designs: combining elements of clinical effectiveness and implementation research to enhance public health impact. Med Care 2012 Mar;50(3):217-226. [doi: 10.1097/MLR.0b013e3182408812] [Medline: 22310560]
50. Gould CE, Carlson C, Ma F, Forman-Hoffman V, Ranta K, Kuhn E. Effects of mobile app–based intervention for depression in middle-aged and older adults: mixed methods feasibility study. JMIR Form Res 2021 Jun 29;5(6):e25808. [doi: 10.2196/25808] [Medline: 34185000]

Abbreviations

aOR: adjusted odds ratio
BEIS-10: Brief Emotional Intelligence Scale-10
CBT: cognitive behavioral therapy
DCP: digital coaching program
DMHI: digital mental health intervention
EI: emotional intelligence
EMM: estimated marginal mean
RCI: reliable change index
SPD: serious psychological distress
WHO-5: World Health Organization-Five Well-Being Index

Edited by A Mavragani; submitted 27.01.25; peer-reviewed by P Wu, T Tillmann; revised version received 09.06.25; accepted 10.06.25; published 07.08.25.

Please cite as:
Peiper NC, Pettitt A, Shah B, Attwood O, Sivonen E, Pfeffer J
Feasibility of a Digital Coaching Program for Improving Mental Well-Being and Emotional Intelligence: Pragmatic Retrospective Cohort Study
JMIR Form Res 2025;9:e71828
URL: https://formative.jmir.org/2025/1/e71828
doi:10.2196/71828

© Nicholas Chalmer Peiper, Adam Pettit, Bela Shah, Oliver Atwood, Ella Sivonen, Jeffrey Pfeffer. Originally published in JMIR Formative Research (https://formative.jmir.org), 7.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Examining Nonverbal Communication in Dyadic Interactions With Virtual Humans Using an Integrated Coding System: Mixed Methods Analysis

Analay Perez, PhD; Rae Sakakibara, BS; Srikar Baireddy, BS; Michael D Fetters[†], MD, MPH, MA; Timothy C Guetterman, PhD

Department of Family Medicine, University of Michigan, 1018 Fuller St, Ann Arbor, MI, United States

[†]deceased

Corresponding Author:
Analay Perez, PhD
Department of Family Medicine, University of Michigan, 1018 Fuller St, Ann Arbor, MI, United States

Abstract

Background: The patient-physician dyad involves both verbal and nonverbal communication. Traditional methods use quantitative or qualitative coding when analyzing dyadic data of nonverbal communication. Quantitative coding methods can capture the frequency of nonverbal communication, while qualitative coding methods can provide descriptive information on the context and nuance of the nonverbal communication expressed. Yet, limited research has examined the integration of quantitative and qualitative coding methods of nonverbal communication between a patient-physician dyad through video recordings.

Objective: The objective of this formative study was to demonstrate how nonverbal communication data can be analyzed using a mixed methods analysis approach and propose an integrated coding system using a subset of the original dataset.

Methods: A secondary analysis was conducted from the intervention study with a sample of 32 pairs of randomly selected video recordings based on first and second interactions after receiving feedback from a virtual human. A 2-minute segment was used to code nonverbal communication, and a codebook was developed, informed by the literature and inductive qualitative approaches. For the mixed methods analysis, we purposefully selected 2 participants from the sample of 32 who demonstrated high frequency in quantitative and qualitative coding of nonverbal behaviors. We developed a joint display to visually represent the integration of quantitative and qualitative coding methods and developed person-level meta-inferences.

Results: This formative study demonstrates an approach to nonverbal communication analysis that mixes qualitative and quantitative methods. The mixed methods results indicated the frequency of participants' (n=32) nonverbal behaviors increased after repeated interactions, including eyebrow raise, nodding, and smiling, in addition to the increased average duration of nonverbal behaviors across interactions. Illustrated through an in-depth example of integrated mixed methods coding of 2 participants from the sample, the integration of quantitative and qualitative data provided insights into nonverbal communication. Quantitatively, we captured the frequency of nonverbal behaviors while qualitatively expanding on the context for nonverbal behaviors and generating person-level meta-inferences. The joint display informed our integrated coding system for mixed methods analysis of nonverbal communication.

Conclusions: The resultant integrated coding system may be helpful to researchers engaging in nonverbal communication data of dyads by providing a step-by-step method using a mixed methods analysis approach. This approach can help us to advance methods for analyzing nonverbal communication to enhance the patient-physician dyad and education on nonverbal communication. We encourage applying the integrated coding system across several subdisciplines in health sciences research to identify how it can be further expanded and refined.

International Registered Report Identifier (IRRID): RR2-10.2196/46601

(JMIR Form Res 2025;9:e59328) doi:10.2196/59328

KEYWORDS

nonverbal communication; nonverbal behaviors; mixed methods analysis; joint display; patient-physician dyad; patient-doctor; patient-physician; dyad; dyadic; interaction; simulation; virtual human; quantitative; qualitative; secondary analysis; non-verbal

XSL•FO
RenderX

Introduction

Background

Verbal and nonverbal communication are important characteristics of the patient-physician dyad. Verbal communication can include but is not limited to language, dialects, and the tone of voice, whereas nonverbal communication includes eye contact, gaze gesture, facial expressions, movement, stance, body position, and spatial distancing [1,2]. These behaviors are only a subset of verbal and nonverbal behaviors included in communication. Researchers have also identified more specific nonverbal behaviors including emotions, embodiment, and interpersonal perception [3]. In clinical practice, it is vital for physicians to effectively communicate with patients using verbal and nonverbal behaviors to improve health outcomes and the overall patient-physician dyadic encounter [4]. Physicians who can appropriately interpret patients' nonverbal cues lead to increased patient satisfaction, and patients are more likely to return for follow-up visits than physicians who are less likely to read patients' nonverbal cues correctly [5]. Appropriate use of nonverbal cues between a patient and a physician can contribute to a positive visit, increase patient self-disclosure, understanding, and recall of information discussed during the visit, and influence a patient's adherence to a physician's recommendations [6]. Not only is verbal and nonverbal communication critical to the patient-physician dyad, but nonverbal communication is pivotal for accreditation of residency programs and physician certifications as these assess physicians' competency in interpersonal skills, which involve verbal and nonverbal communication [7]. For example, the Accreditation Council for Graduate Medical Education includes a core competency of interpersonal and communication skills comprising both verbal and nonverbal communication. Although a wealth of literature exists on the benefits of enhancing verbal and nonverbal communication in health sciences research and the types of gestures commonly associated with each type, additional methodological guidance is needed when coding, analyzing, and integrating verbal and nonverbal using a mixed methods approach. Hillen et al [8] conducted a systematic review synthesizing quantitative and qualitative coding methods for assessing the interaction between health care professionals and patients and/or caregivers. However, none of the articles included in their review used mixed methods research, that is, the intentional integration of quantitative and qualitative research to better understand the research phenomenon of health care professionals and patient and/or caregiver interactions. Therefore, research is needed to understand the potential of using mixed methods analysis to holistically explore verbal and nonverbal communication.

Traditional Methods for Analyzing Nonverbal Communication

Traditionally, both quantitative and qualitative methods are used to analyze nonverbal communication. Coding is a process of identifying segments of data, such as a brief video snippet, and applying a code as a label for what is occurring. From a quantitative perspective, the most common method to code and analyze facial expressions involves manual coding of specific facial expressions and movements using frequencies [9]. The Facial Affect Coding System (FACS) [10,11] has been widely used by several researchers and has demonstrated strong psychometric properties [12]. The FACS was created to identify facial expressions, including details on their intensity and timing. Trained coders are directed to code all possible facial expressions into action units, which are based on underlying muscle movements mapped to specific and observable facial expressions. However, 2 limitations of the FACS are that it does not account for body language and only focuses on one person rather than the interactions.

Another common approach to analyzing nonverbal behaviors is using quantitative content analysis. This process includes developing an established codebook based on manifest behavior codes and recurring codes to guide coders' decisions [13]. Quantitative content analysis uses a similar approach as the FACS; however, it is less structured and based on an established codebook developed by the research team. Both approaches use frequencies to determine the total number of specified behaviors expressed by each participant. In either case, researchers can conduct statistical analyses if a large enough sample size exists. Though quantitative methods for coding nonverbal communication can help quantify the frequency of specific behaviors, they fail to provide context for how or why individuals may be expressing those behaviors.

Qualitatively, researchers can expand on the quantitative codes by exploring additional qualitative descriptors of each nonverbal communication. Qualitative content analysis can be implemented by developing codes and themes using an inductive process. In other words, researchers do not begin with a pre-established codebook but let the data drive what codes they develop and identify. Quantitative content analysis can help uncover what common nonverbal behaviors are, particularly when examining the patient-physician dyad, while qualitative content analysis can help reveal why by providing additional verbal and nonverbal details to complement the quantitative coding [13].

A more specific qualitative coding method is live coding [14]. Live coding is a method that allows researchers to code both verbal and nonverbal behaviors while listening to and watching video recordings [14]. The Rotor Interaction Analysis System (RIAS) uses a variation of live coding by accounting for some nonverbal behaviors qualitatively, though it primarily uses a quantitative coding system [15]. Given the advent of digital data such as video, audio, social media, music, photographs, and films in qualitative research, and the advancements of computerized assisted qualitative data analysis software, researchers are using digital software tools to facilitate this process and simultaneously code both video and audio data. As noted by Parameswaran and colleagues [14], this allows researchers to "gather contextual information from the video clips to better situate the themes when coding."

Nonetheless, qualitatively coding nonverbal communication fails to provide the magnitude of a particular nonverbal behavior and may be less precise than quantitative coding. Therefore, merging quantitative and qualitative coding may be an ideal solution to further deepen understanding by providing

XSL•FO
RenderX

quantifiable and contextual descriptions of nonverbal communication between dyads. Yet, limited research has explored the potential benefits of integrating quantitative and qualitative coding methods using a mixed methods analysis approach to examine nonverbal communication.

Mixed Methods Analysis of Nonverbal Communication

Mixed methods research is a methodology that intentionally integrates quantitative and qualitative research methods to obtain a holistic understanding of the research questions and objectives [16]. When integrating quantitative and qualitative coding at the methods level, a single video dataset can be analyzed using an inherently mixed methods analysis approach. Onwuegbuzie and Abrams [1] proposed several reasons for integrating quantitative and qualitative coding. One reason is to reduce the dimensionality of qualitative nonverbal communication data to subsequently further explore using quantitative analysis, such as exploratory factor analysis. Another approach for integrating quantitative and qualitative coding can be to compare quantitative and qualitative nonverbal data through a joint display to generate new conclusions called meta-inferences. Alternatively, using the same dataset, another approach is to transform quantitative nonverbal data to analyze qualitatively by exploring salient themes or link the qualitative nonverbal data with quantitized nonverbal or verbal communication data to examine its correlation. Quantitizing data refers to transforming qualitative coded data into a quantitative representation, while qualitizing data refers to transforming numerical results into qualitative codes and themes [17,18]. This process can be viewed as a mixed methods coding technique [19].

Onwuegbuzie and Abrams [1] have conducted recent research on analyzing nonverbal communication using mixed methods analysis and provided useful examples from empirical research to guide researchers. To demonstrate the applications of mixed methods research for nonverbal communication, researchers [1] present several examples of studies, most involving qualitative interviews as a method for coding nonverbal communication, as well as in-person observation coding, and using a combination of digital technologies including video recordings and photographs/field notes, with examples primarily derived from the field of education. Given the role and importance of nonverbal communication in health care and health professions education, it is imperative to explore how researchers can use mixed methods research to integrate quantitative and qualitative coding of nonverbal communication at the methods level. Thus, our aim was to develop and pilot a mixed methods approach to analyze nonverbal communication in health care.

Research Aims

The overarching purpose of this formative study was to develop and pilot an integrated coding system for mixed methods analysis of nonverbal communication and illustrate how researchers can integrate quantitative and qualitative coding methods of nonverbal communication. Particularly, this study used data from an existing video interaction dataset of simulated health care encounters between a physician and a virtual human to explore nonverbal behaviors. For this formative study, we developed an integrated coding system using a subset of data from the original sample. The aims were to: (1) understand how nonverbal behaviors could be improved through learner interactions with a virtual human across repeated scenarios and (2) use a novel mixed methods approach to analyze quantitative and qualitative coded data of nonverbal communication to provide greater context for each nonverbal behavior exhibited by participants.

Methods

Participants

This analysis relied on secondary data collected as part of a larger trial [20]. Participants in this study consisted of second-year medical students recruited from 3 institutions in the United States.

Data Sources

We conducted a secondary analysis from the intervention arm of a randomized controlled trial (n=210 medical students and 840 video recordings of interactions) of a communication simulation using virtual humans. Our sample for developing and piloting a novel mixed methods analysis of nonverbal communication consisted of n=32 pairs of randomly selected video recordings of first and second interactions after receiving feedback from the virtual human program, representing approximately 15% of the entire dataset.

The duration of interactions was typically between a range of 4 and 8 minutes; however, we used the “thin slice” method for coding, which involved coding short excerpts of the video recordings. Evidence shows that short observations of 1 to 3 minutes can represent a participant’s nonverbal behavior and predict clinical and social-psychological outcomes [21]. We chose to code 2-minute segments to account for the pauses in conversation while learners read prompts on the screen. We began coding 2 minutes into the videos because of the likelihood of capturing the interaction at a point where the learner is confronted by virtual humans in emotionally charged states after some initial exchanges. The virtual human simulation, MPathic-VR (Medical Cyberworlds), provides automated feedback after the first interaction, and the learner is given another opportunity to practice in their second interaction with the same virtual human(s). Additional information on the simulation and larger study can be found in a protocol paper [22].

Virtual Human Scenarios and Procedures

Participants in the intervention arm interacted with the MPathic-VR and participated in two related scenarios. One scenario focused on intercultural communication between a virtual human patient, the virtual human patient’s mother, and the learner. The second scenario focused on interprofessional communication with a nurse virtual human and the learner [20]. Participants in the intervention arm received personalized feedback on their performance (both verbal and nonverbal) after each scenario and a video recording of their conversations with the virtual human.

XSL•FO
RenderX

Quantitative Analysis

Using the established codebook, we first identified which codes would be treated quantitatively using frequencies. After the 2 analysts coded the nonverbal quantitative behaviors by counting the total number of times each behavior was expressed in the 2-minute video segment for the first and second interactions with a virtual human. We summed the number of expressions for each nonverbal behavior across all videos to produce a total number for interactions one and two. We also summed the number of nonverbal behaviors expressed by the participant across the learner and virtual human talk time. Multimedia Appendix 1 provides the codebook describing each code, its definition, and whether the code was considered a quantitative or qualitative code.

Qualitative Analysis

We began our analysis by reviewing the literature on measures and coding schemes for analyzing nonverbal behaviors. Specifically, we included in the codebook behaviors emphasized by the MPathic-VR system (eg, smiling, nodding, and eyebrow raises), in addition to nonverbal behaviors identified as expressions that may indicate compassion as described by the FACS [23]. These included eye gaze, head orientation, forward lean, oblique eyebrows, furrowed eyebrows, lower eyelid raise, slight mouth press, and lip corner puller. Coding of these behaviors was guided by the FACS [10,11]. We used the FACS in this investigation because the action units have a specific anatomical basis, which allowed for reliable application of nonverbal behavior codes across coders.

Qualitative codes were identified using open coding of the data to generate additional codes. The 2 coders (TCG and RS) also created new codes (not part of FACS) inductively through qualitative open coding. Both were experienced and trained in thematic and content analysis, one at the PhD level and the other, a research assistant with over three years of experience with qualitative data. These qualitative codes (eg, intonation and nodding) were necessary because of their prominence in the data and because the FACS does not capture such behavior. We imported video recordings into MAXQDA 2022 qualitative software for coding [24]. The two individuals conducting the initial open coding reviewed three video segments second-by-second to compare and discuss codes through consensus and reach a final codebook. Each code was defined for consistency, and we ensured all analysts were familiar with the codebook. All authors participated in the coding and review of videos and were blinded as to whether the video was the first or second interaction. We applied the qualitative codes to the relevant segment of video, which is a time duration with a start and end point precise to a second using the MAXQDA software. Any discrepancies in coding were resolved through discussions after each video was coded.

Ethical Considerations

The original study was approved by the University of Michigan's Institutional Review Board (HUM00067336) as an exempt study under the educational research category. Written informed consent was waived for participants. The secondary analysis reported in this paper was also approved by the University of Michigan's Institutional Review Board (HUM00134766). Participants did not receive compensation. It was not possible to completely anonymize the video data, but we removed all participant identifiers and assigned a random study identifier. After coding, the videos were unlinked to develop a deidentified dataset.

Results

Nonverbal Behaviors Demonstrated by Learners

To provide a more thorough examination of nonverbal behaviors, we first investigated whether behaviors emphasized by the MPathic-VR system and other nonverbal behaviors that may convey empathy increased after repeat scenarios by examining the frequency and duration of nonverbal behaviors across videos from 32 pairs. As demonstrated in Table 1, the frequency of nonverbal behaviors emphasized by MPathic-VR increased in the second interactions compared to the first, including eyebrow raise (T_1=111 instances, T_2=128 instances; here T_1=Time 1 or first interaction and T_2=Time 2 or second interaction), nodding (T_1=112 instances and T_2=211 instances), and smiling (T_1=33 instances and T_2=38 instances). The frequency of all other nonverbal behaviors also increased in the second interaction compared to the first, except for lower eyelid raises and furrowed eyebrows, which remained consistent across scenarios (73 instances and 53 instances, respectively). Table 2 shows that the average duration of nonverbal behaviors regarding orientation increased in the second interaction compared to the first, including the average duration of eye gaze toward the virtual human (T_1=61.3 s and T_2=76.3 s), forward lean (T_1=37.5 s and T_2=38.9 s), and head tilting (T_1=34.2 s and T_2=48.9 s).

XSL•FO
RenderX

Table . Frequency of nonverbal behaviors in the first and second interactions.

Nonverbal behaviors	Code frequency (instances across n=32 videos), n	
	First interaction	Second interaction
Behaviors emphasized by MPathic-VR		
Eyebrow raise	111	128
Nodding	112	211
Smiling	33	38
Other facial expressions		
Furrowed eyebrows	73	73
Lip corner puller	90	120
Lower eyelid raise	53	53
Oblique eyebrows	45	62
Slight lip press	90	115

Table . Duration of nonverbal behaviors in the first and second interactions.

Orientation	Average code duration per video (seconds)	
	First interaction	Second interaction
Eye gaze toward virtual human	61.3	76.3
Forward lean	37.5	38.9
Head tilting	34.2	48.9

We also examined whether the frequency across all nonverbal behaviors increased after repeated scenarios while either talking or listening to the virtual human (see Figures 1 and 2). The frequency of nonverbal behaviors across learner and virtual human talk time increased from the first interactions to the second interactions, except for lower eyelid raise and smiling. The frequency of lower eyelid raise decreased between the first and second interaction for learner talk time (T_1=17 and T_2=16 instances) and smiling (T_1=3 and T_2=2 instances). The frequency of smiling was the only nonverbal behavior that did not increase after repeated scenarios during the virtual human talk time (T_1=19 instances and T_2=19). Figure 1 shows the nonverbal behaviors of the first interactions, and Figure 2 displays these behaviors across the second interactions. These figures are extracted from the MAXQDA software (VERBI Software) demonstrating a Code Relations Browser with the respective frequencies of each nonverbal behavior between the learner and virtual human talk time.

XSL•FO
RenderX

Figure 1. First interaction with virtual human: Frequency of simultaneous occurrences of nonverbal behavior codes when the learner or virtual human was talking during video scenario. VH: virtual human.

Code System	Learner talk time	VH talk time
Orientation		
Forward lean	49	47
Eye gaze towards VH	187	237
Head and body orientation to VH		
Head tilt left	43	33
Head tilt right	36	24
Facial expression		
Lip corner puller		44
Eyebrow raise	71	39
Furrowed eyebrows	49	25
Oblique eyebrows	33	14
Lower eyelid raised	17	13
Slight lip press	1	37
Smile	3	19
Nodding	61	34

Figure 2. Second interaction with virtual human: Frequency of simultaneous occurrences of nonverbal behavior codes when the learner or virtual human was talking during video scenario. VH: virtual human.

Code System	Learner talk time	VH talk time
Orientation		
Forward lean	67	62
Eye gaze towards VH	218	267
Head and body orientation to VH		
Head tilt left	61	59
Head tilt right	53	39
Facial expression		
Lip corner puller	2	82
Eyebrow raise	100	53
Furrowed eyebrows	56	39
Oblique eyebrows	42	27
Lower eyelid raised	16	30
Slight lip press	2	71
Smile	2	19
Nodding	73	118

Mixed Methods Analysis of Nonverbal Communication

To illustrate the application of mixed methods analysis of nonverbal communication, 2 participants were purposefully selected from the sample based on the highest frequency of quantitative and qualitative coded segments of nonverbal behaviors. To analyze the quantitative and qualitative codes and integrate the data, we found concordance between quantitative and qualitative nonverbal communication codes across segments that were close in time. Table 3 displays the joint display of integrated findings across quantitative and qualitative coded nonverbal communication data. A joint display is a visual display used to facilitate and represent the integration of the quantitative and qualitative phases in mixed methods research [25]. The joint display presented is categorized by the time segment and its respective quantitative codes, the time segment and its respective qualitative codes, and the person-level meta-inferences. Meta-inferences, or integrated conclusions from the quantitative and qualitative strands, were developed by merging quantitative and qualitative coding inferences to fully understand participants' nonverbal communication and further augment the verbal data.

XSL•FO
RenderX

We generated meta-inferences at the person-level by aggregating quantitative and qualitative findings across all time points for each person to generate a rich description of each participant's overall nonverbal communication (see Table 3 for person-level meta-inferences). For example, participant 047 for 2:06-2:30 engaged in forward lean while at the same time nodding their head upwards and downwards. When integrating these results within the context of the video scenario, we can conclude the participant was engaged in the conversation and displayed nonverbal behaviors consistent with the verbal data, affirming that they knew what was best for the patient. Collectively, these were integrated to generate a person-level meta-inference representing the nonverbal communication displayed by each participant for the entire time segment.

Table 4 displays the frequencies for each quantitative nonverbal behavior code across each participant. Compared to other quantitative nonverbal codes, participant ID 047 engaged in forward lean and learner talk time at the highest frequency (four and three instances, respectively). Participant ID 057 displayed eye gaze toward the virtual human at the highest frequency (7 instances), followed by learner talk time and virtual human talk time (3 instances, respectively). These results help to further contextualize the qualitative findings from the joint display by providing quantifiable data across nonverbal behaviors expressed by the participants.

XSL•FO
RenderX

Table . Joint display of integrated nonverbal coding analysis and person-level meta-inferences.

Time	Quantitative codes	Time	Qualitative codes	Person-level meta-inference
ID 047				
2:06-2:30	Forward lean	2:09-2:12	Nodding head up and down	The participant was engaged in the conversation and displayed nonverbal communication in response to the virtual human regarding steps they took to alleviate tension between the patient and their mother. Participant's nonverbal cues demonstrated that they cared about the outcome of the conversation and explained the actions they took to alleviate tension between the patient and their mother. The participant's gaze was focused on the virtual human, and they adequately responded all questions posed by the virtual human. The participant displayed nonverbal behaviors such as shaking their head while agreeing with the virtual human's response when learning about confidential patient information.
2:06-2:35	Forward lean and learner talk time	2:30-2:32	Shakes head side-to-side	
2:34-2:58	Forward lean and learner talk time	2:54-2:58	Shakes head	
2:55-3:48	Forward lean	3:31-3:32	Nodding head	
2:57-3:00	Eye gaze toward virtual human			
3:47-4:15	Learner talk time	3:47-3:48	Inconsistent behavior	
ID 057				

XSL•FO
RenderX

Time	Quantitative codes	Time	Qualitative codes	Person-level meta-inference
2:03-2:06	Learner talk time and eye gaze toward virtual human	2:05-2:07	Nodding head up and down	
2:24-2:43	Virtual human talk time and eye gaze toward virtual human	2:43-2:44	Nodding head forward and backward	
2:45-2:55	Learner talk time and eye gaze toward virtual human	2:46-2:47	Nodding head side-to-side	
2:52- 3:22	Virtual human talk time and eye gaze toward virtual human	3:01-3:23	Nodding heading, nodding head front to back, frowns, and presses lips to the corner of mouth	
3:24-3:46	Virtual human talk time and eye gaze toward virtual human	3:24- 3:29	Nodding head forward and backward	
3:47-3:51	Eye gaze toward virtual human and learner talk time	3:45-3:48	Nodding head forward and backward and inconsistent behavior	
3:57-3:59	Eye gaze toward virtual human	3:57-3:58	Nodding head forward and backward	

XSL•FO
RenderX

Time	Quantitative codes	Time	Qualitative codes	Person-level meta-inference
				The participant focused on the virtual human's response while nodding their head up and down to affirm patient's autonomy. During this process, the participant nodded their head forward and backward, closed their eyes, and their lips read "okay" as if to imply that there could be an issue with the information that was disclosed about the patient. The participant nodded their head back and forth while quietly under breath stating, "I see," based on confidential patient information disclosed by the virtual human. Then, they nodded their head and affirmed they did not understand the situation, implying the participant was admitting to making a mistake. The participant was focused on the response from the virtual human and nodded their head in a manner that demonstrated they understood the situation and closed their eyes as if to sympathize with the patient. As the virtual human continued to share information, the participant nodded their head back and forth in agreement. The participant frowned and pressed their lips to the corner of their mouth to show signs of sympathy and disbelief. The participant nodded their head forward and backward to show that they understood what was happening based on the information shared with them about the patient by the virtual human. The participant's nonverbal behaviors aligned with their agreement to the virtual human's comment on the importance of working as a team in health care with the pa-

XSL•FO
RenderX

Time	Quantitative codes	Time	Qualitative codes	Person-level meta-inference
				tient and doctors.

Table . Frequencies of quantitative nonverbal codes expressed by participants.

Quantitative code	Frequency, n	
	Participant ID 047	Participant ID 057
Forward lean	4	0
Learner talk time	3	3
Eye gaze toward virtual human	1	7
Virtual human talk time	0	3

Discussion

Principal Findings

The findings from this study demonstrate that the frequency of nonverbal behaviors increased after repeated scenarios between the first and second interactions. Specifically, all nonverbal behaviors emphasized by the MPathic-VR, including eyebrow raise, nodding, and smiling, increased from the first to second interactions. The average duration (in seconds) of nonverbal behaviors, including eye gaze toward virtual human, forward lean, and head tilting, also increased from the first to second interactions. Overall, these results provide some evidence of the validity of our coding approach, given the behaviors expected to increase did increase based on the nonverbal coding.

These results are particularly promising regarding patient-physician interactions. Physicians who display empathic nonverbal behaviors increase patients' perceptions of physician empathy, warmth, and competence [26]. Furthermore, nonverbal behaviors such as patient-centered gaze and body orientation positively affect a patient's perceived level of physician empathy [27]. Therefore, these findings hold clinical implications for training effective nonverbal communication skills among medical students and professionals in clinical practice. Furthermore, these results are supported in the literature demonstrating that students who interacted with MPathic-VR and received feedback from a simulated scenario improved their communication scores, particularly nonverbal communication [20]. The findings from this study used the same virtual human, the MPathic-VR, and reported the benefits of using virtual humans as it allowed for a standardized method for comparison across participants [20]. Further research is needed to ensure the findings from a virtual human are replicated in real-world settings.

Through this formative study, we highlight the role of using video recordings and mixed methods analysis to obtain a more comprehensive understanding of how learners improve their nonverbal communication through virtual human video interactions. Using video rather than photos has been shown to be a more ecologically robust method for analyzing nonverbal communication [26]. From a mixed methods analysis perspective, this study sheds light on a novel approach by coding quantitative and qualitative nonverbal data to generate a more holistic understanding with descriptions and context for each behavior. By analyzing and integrating quantitative and qualitative coding of nonverbal behaviors, we generated meta-inferences that illuminated more nuance of each nonverbal behavior displayed by 2 participants from the sample. As such, one of the contributions of this study was the development of a step-by-step approach elucidating an integrated coding system for mixed methods analysis of nonverbal communication and strategies to guide researchers in analyzing nonverbal communication data. Below we provide a step-by-step approach for an integrated coding system:

First, determine the duration of the coded segment: The first step is to identify the duration of the video that will be coded. As noted, we used the thin slice method, but if there is a rationale for a longer duration, it is important to justify the length of time that will be coded. It is also critical to identify whether coding will occur toward the beginning, middle, or end of the video, explicitly outlining the start and end times for each clip.

Second, developing a codebook of nonverbal behaviors: Afterwards, it is important to identify the quantitative and qualitative nonverbal codes that will be incorporated into the codebook. These codes may be informed by literature, theory, or the study. Quantitative codes involve using frequencies or counts to document nonverbal behaviors such as eye gaze, head tilt, and smile. Qualitative codes can provide further details of each behavior. For example, if a qualitative code is based on positive or negative intonation, researchers may expand on this code by describing whether participants appeared uninterested during the interaction, the context of the discussion, and the point at which the participant starts to show signs of disinterest. When developing a codebook, it is encouraged to collaborate with all members of the research team who will be engaging in the coding and data analysis phase to ensure consensus on the codebook, determine distinctions between quantitative and qualitative nonverbal codes, and ensure each code is operationalized to increase reliability among coders.

Third, the coding process: Determine how many individuals from the research team will be involved in the coding process, how many video clips each individual on the research team will code, the number of times individuals will meet to discuss codes, and strategies to ensure the validity and reliability of the data

XSL•FO
RenderX

throughout the coding process. The team should also decide whether software such as MAXQDA, NVivo (QSR International), or any other qualitative software programs will be used to code video data or whether other alternatives, such as Microsoft Excel, will be used to organize coded information.

Fourth, integrating quantitative and qualitative coded data of nonverbal behaviors: Once all videos have been coded quantitatively and qualitatively, researchers will be tasked with determining the most appropriate ways to integrate the data. Given that nonverbal communication data can be analyzed quantitatively and qualitatively, we encourage researchers to use mixed methods analysis approaches [1]. When using mixed methods analysis to analyze nonverbal data, consider the purpose of conducting a mixed methods analysis. These can include but are not limited to reducing data, using a joint display, data transformation, correlating data, merging verbal with nonverbal data to generate new codes or variables, comparing data, confirming data, or using findings based on the qualitative coded data of nonverbal behavior to inform the quantitative analysis [1]. Determining the reason for conducting a mixed methods analysis of nonverbal data will depend upon the research questions and the study's objectives. Thus, researchers are encouraged to revisit their research questions and aims frequently to determine the most adequate approach for the mixed methods analysis of nonverbal data.

Fifth, presenting a visual display and developing meta-inferences: To visually depict the integration of quantitative and qualitative coding of nonverbal data and findings, a joint display can be used to represent this information and aid in generating meta-inferences. A joint display can be used to integrate qualitative and quantitative data, methods, or results using a visual representation [25]. To identify meta-inferences, a researcher can match verbal and nonverbal domains to get a better understanding of the phenomenon. For example, if integrating the quantitative and qualitative data using matching, a researcher may code the type of verbal interactions and the context of verbal responses with the corresponding nonverbal behaviors to identify the meta-inferences. When developing meta-inferences, the context of the interaction will be particularly helpful. Analysts familiar with the coding scheme should work together to develop these meta-inferences. As a method of member check, researchers may share the developed meta-inferences with participants to ensure the validity of the findings. This will allow for participants to provide feedback, revisions, and additions to the proposed meta-inferences. We encourage researchers to use creativity to explore the best strategies to visually represent the data and findings, reinforcing thoughtful consideration of the generated meta-inferences.

Limitations

This study has several limitations. First, the nonverbal analysis is based on a small, formative probabilistic sample of only 32 individuals. Results are not likely generalizable, as the goal was to develop methods of nonverbal analysis. A second limitation of this study is that data in the joint display are only displayed for 2 participants. However, we purposefully selected these participants to demonstrate how data can be integrated and analyzed using mixed methods research. Future research can incorporate more participant examples in a joint display to further highlight nuances of integration and analysis of quantitative and qualitative coding of nonverbal behaviors. Furthermore, we present and code several nonverbal behaviors, but we also recognize that this is not an exhaustive list of all nonverbal behaviors present during a patient-physician interaction. In addition, although we provided a codebook and defined each nonverbal behavior, there may be issues of coder drift. However, we aimed to capture relevant nonverbal behaviors in healthcare that appear related to empathic communication and explained these behaviors in the codebook to ensure behaviors were coded consistently. Yet, linking specific behaviors to emotions or underlying intent is tenuous and beyond the scope of this formative study. We encourage researchers to expand on these nonverbal codes by examining the literature in their respective fields to determine additional codes that can be incorporated.

Conclusions

This study aimed to develop and pilot a mixed methods coding system to analyze verbal and nonverbal behaviors using video data to obtain a more comprehensive understanding of nonverbal behaviors. Although nonverbal communication may contradict verbal communication, being cognizant of nonverbal behaviors can help physicians better express their nonverbal communication responses [4]. As a result, enhancing nonverbal communication skills in health settings can help to build trust, foster healing, and improve health outcomes [4]. Furthermore, we demonstrated how using a mixed methods analysis approach to code nonverbal communication can be particularly beneficial when analyzing video data of patient-physician communication to expand on both verbal and nonverbal behaviors exhibited by each individual. Using mixed methods analysis, we developed a novel approach to analyzing nonverbal communication that builds on existing literature by providing a formative demonstration and proposing an integrated coding system. We aim for this integrated coding system to shed light for researchers analyzing nonverbal communication data across various subdisciplines within health sciences research. We encourage researchers to investigate creative paths for analyzing nonverbal communication data using video recordings through a mixed methods analysis approach that advances discussions on enhancing the patient-physician dyad across multiple clinical settings.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to the protection of participants, given that the data are in video format.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Codebook excerpt derived from Facial Affect Coding System (FACS) and inductive qualitative coding.
[DOCX File, 18 KB - formative_v9i1e59328_app1.docx]

References

1. Onwuegbuzie AJ, Abrams SS. Nonverbal communication analysis as mixed analysis. In: The Routledge Reviewer's Guide to Mixed Methods Analysis: Routledge; 2021. [doi: 10.4324/9780203729434]
2. Purnell L. Cross cultural communication: verbal and non-verbal communication, interpretation and translation. In: Marty DM, Pacquiao D, Purnell L, editors. Global Applications of Culturally Competent Health Care: Guidelines for Practice: Springer International Publishing; 2018:132-142. [doi: 10.1007/978-3-319-69332-3_14]
3. Hall JA, Knapp ML, editors. Nonverbal Communication: De Gruyter, Inc; 2013. [doi: 10.1515/9783110238150]
4. Dalonges DA, Fried JL. Creating immediacy using verbal and nonverbal methods. J Dent Hyg 2016 Aug;90(4):221-225. [Medline: 27551142]
5. Mast MS. On the importance of nonverbal communication in the physician-patient interaction. Patient Educ Couns 2007 Aug;67(3):315-318. [doi: 10.1016/j.pec.2007.03.005] [Medline: 17478072]
6. Robinson JD. Nonverbal communication and physician-patient interaction: review and new directions. In: The Sage Handbook of Nonverbal Communication: Sage Publications, Inc; 2006:437-459. [doi: 10.4135/9781412976152.n23]
7. Duffy FD, Gordon GH, Whelan G, et al. Assessing competence in communication and interpersonal skills: the Kalamazoo II report. Acad Med 2004 Jun;79(6):495-507. [doi: 10.1097/00001888-200406000-00002] [Medline: 15165967]
8. Hillen MA, Mitchell KAR, Schouten BC, Cyrus JW, Brown RF, Pieterse AH. Systematic review of observational coding systems to assess patient-healthcare professional interactions. Patient Educ Couns 2025 Jun;135:108718. [doi: 10.1016/j.pec.2025.108718] [Medline: 40037145]
9. Harrigan JA. Methodology: coding and studying nonverbal behavior. In: Nonverbal Communication Handbooks of Communication Science: De Gruyter Mouton; 2013:35-68. [doi: 10.1515/9783110238150.35]
10. Ekman P, Friesen WV. Facial Action Coding System: Consulting Psychologists Press; 1978. [doi: 10.1037/t27734-000]
11. Ekman P, Rosenberg EL, editors. What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS): Oxford University Press; 1997. [doi: 10.1093/oso/9780195104462.001.0001]
12. Cohn JF, Ambadar Z, Ekman P. Observer-based measurement of facial expresion with the facial action coding system. In: Handbook of Emotion Elicitation and Assessment: Oxford University Press; 2007:203-221. [doi: 10.1093/oso/9780195169157.003.0014]
13. Pennington DR. Coding of non-text data. In: The SAGE Handbook of Social Media Research Methods 2017:232-251. [doi: 10.4135/9781473983847.n15]
14. Parameswaran UD, Ozawa-Kirk JL, Latendresse G. To live (code) or to not: a new method for coding in qualitative research. Qual Soc Work 2020 Jul;19(4):630-644. [doi: 10.1177/1473325019840394]
15. Roter D, Larson S. The Roter Interaction Analysis System (RIAS): utility and flexibility for analysis of medical interactions. Patient Educ Couns 2002 Apr;46(4):243-251. [doi: 10.1016/s0738-3991(02)00012-5] [Medline: 11932123]
16. Creswell JW, Plano Clark VL. Designing and Conducting Mixed Methods Research, 3rd edition: Sage; 2018.
17. Sandelowski M, Voils CI, Knafl G. On quantitizing. J Mix Methods Res 2009 Jul 1;3(3):208-222. [doi: 10.1177/1558689809334210] [Medline: 19865603]
18. Onwuegbuzie AJ, Leech NL. On qualitizing. Int J Multiple Res Approaches 2019 Aug 31;11(2):98-131. [doi: 10.29034/ijmra.v11n2editorial2]
19. Saldaña J. The Coding Manual for Qualitative Researchers, 4th edition: SAGE; 2021.
20. Kron FW, Fetters MD, Scerbo MW, et al. Using a computer simulation for teaching communication skills: a blinded multisite mixed methods randomized controlled trial. Patient Educ Couns 2017 Apr;100(4):748-759. [doi: 10.1016/j.pec.2016.10.024] [Medline: 27939846]
21. Ambady N, Rosenthal R. Thin slices of expressive behavior as predictors of interpersonal consequences: a meta-analysis. Psychol Bull 1992;111(2):256-274. [doi: 10.1037/0033-2909.111.2.256]
22. Perez A, Fetters MD, Creswell JW, et al. Enhancing nonverbal communication through virtual human technology: protocol for a mixed methods study. JMIR Res Protoc 2023 Jun 6;12(1):e46601. [doi: 10.2196/46601] [Medline: 37279041]
23. Goetz JL, Keltner D, Simon-Thomas E. Compassion: an evolutionary analysis and empirical review. Psychol Bull 2010 May;136(3):351-374. [doi: 10.1037/a0018807] [Medline: 20438142]
24. VERBI software. MAXQDA 2022. 2021. URL: https://www.maxqda.com/ [accessed 2025-06-27]
25. Guetterman TC, Fàbregues S, Sakakibara R. Visuals in joint displays to represent integration in mixed methods research: a methodological review. Methods Psychol 2021 Dec;5:100080. [doi: 10.1016/j.metip.2021.100080]

XSL•FO
RenderX

26. Kraft-Todd GT, Reinero DA, Kelley JM, Heberlein AS, Baer L, Riess H. Empathic nonverbal behavior increases ratings of both warmth and competence in a medical context. PLoS ONE 2017;12(5):e0177758. [doi: 10.1371/journal.pone.0177758] [Medline: 28505180]
27. Brugel S, Postma-Nilsenová M, Tates K. The link between perception of clinical empathy and nonverbal behavior: the effect of a doctor's gaze and body orientation. Patient Educ Couns 2015 Oct;98(10):1260-1265. [doi: 10.1016/j.pec.2015.08.007] [Medline: 26320820]

Abbreviations

FACS: Facial Affect Coding System
RIAS: Rotor Interaction Analysis System

Edited by A Mavragani; submitted 10.04.24; peer-reviewed by AC Dantas, D Hartigan; revised version received 16.05.25; accepted 21.05.25; published 07.08.25.

Please cite as:
Perez A, Sakakibara R, Baireddy S, Fetters MD, Guetterman TC
Examining Nonverbal Communication in Dyadic Interactions With Virtual Humans Using an Integrated Coding System: Mixed Methods Analysis
JMIR Form Res 2025;9:e59328
URL: https://formative.jmir.org/2025/1/e59328
doi:10.2196/59328

© Analay Perez, Rae Sakakibara, Srikar Baireddy, Michael D Fetters, Timothy C Guetterman. Originally published in JMIR Formative Research (https://formative.jmir.org), 7.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Evaluating a Customized Version of ChatGPT for Systematic Review Data Extraction in Health Research: Development and Usability Study

Jayden Sercombe*, BPsych; Zachary Bryant*, BCom, MPH; Jack Wilson, PhD

The Matilda Centre for Research in Mental Health and Substance Use, University of Sydney, Jane Foss Russell Building (G02), Level 6, Sydney, Australia

*these authors contributed equally

Corresponding Author:
Zachary Bryant, BCom, MPH
The Matilda Centre for Research in Mental Health and Substance Use, University of Sydney, Jane Foss Russell Building (G02), Level 6, Sydney, Australia

Abstract

Background: Systematic reviews are essential for synthesizing research in health sciences; however, they are resource-intensive and prone to human error. The data extraction phase, in which key details of studies are identified and recorded in a systematic manner, may benefit from the application of automation processes. Recent advancements in artificial intelligence, specifically in large language models (LLMs) such as ChatGPT, may streamline this process.

Objective: This study aimed to develop and evaluate a custom Generative Pre-Training Transformer (GPT), named *Systematic Review Extractor Pro*, for automating the data extraction phase of systematic reviews in health research.

Methods: OpenAI's GPT Builder was used to create a GPT tailored to extract information from academic manuscripts. The Role, Instruction, Steps, End goal, and Narrowing (RISEN) framework was used to inform prompt engineering for the GPT. A sample of 20 studies from two distinct systematic reviews was used to evaluate the GPT's performance in extraction. Agreement rates between the GPT outputs and human reviewers were calculated for each study subsection.

Results: The mean time for human data extraction was 36 minutes per study, compared to 26.6 seconds for GPT generation, followed by 13 minutes of human review. The GPT demonstrated high overall agreement rates with human reviewers, achieving 91.45% for review 1 and 89.31% for review 2. It was particularly accurate in extracting study characteristics (review 1: 95.25%; review 2: 90.83%) and participant characteristics (review 1: 95.03%; review 2: 90.00%), with lower performance observed in more complex areas such as methodological characteristics (87.07%) and statistical results (77.50%). The GPT correctly extracted data in 14 instances (3.25% in review 1) and four instances (1.16% in review 2) when the human reviewer was incorrect.

Conclusions: The custom GPT significantly reduced extraction time and shows evidence that it can extract data with high accuracy, particularly for participant and study characteristics. This tool may offer a viable option for researchers seeking to reduce resource demands during the extraction phase, although more research is needed to evaluate test-retest reliability, performance across broader review types, and accuracy in extracting statistical data. The tool developed in the current study has been made open access.

(JMIR Form Res 2025;9:e68666) doi:10.2196/68666

KEYWORDS

artificial intelligence; systematic reviews; data extraction; LLM; ChatGPT; AI; large language models

Introduction

The application of artificial intelligence (AI) in health research has the potential to innovate and optimize various research tasks [1]. Systematic reviews are now increasingly conducted as a gold-standard method for evaluating and synthesizing the evidence base [2]. However, this research methodology often takes up significant time, and therefore can be costly, and is subject to human error [3-5]. As large language models (LLMs) and their capabilities rapidly improve, they are now being used to assist with systematic reviews, streamlining a significant portion of health research methodology.

There has been a dramatic rise in the production of systematic reviews, as demonstrated by a 20-fold increase over the past 20-years, equivalent to 80 publications per day [6,7]. Although they have become a valuable resource for informing policy and clinical practice [8], systematic reviews require significant labor and financial costs. A recent analysis concluded that the mean estimated time to complete and publish a systematic review was 67.3 weeks, with an average of five authors per published text

XSL•FO
RenderX

[3]. In addition to software and dissemination fees, organizations can expect to pay around USD $140,000 per review [4]. There is further concern for the potential of human error during the screening and extractions stages [5].

In response to this significant burden, there have been strong calls for the use of AI in systematic reviews, making the process more efficient [9]. The oldest applications have focused on the automation of title and abstract screening. Machine learning classifiers, such as support vector machines or complement naive Bayes have been trained to replicate human inclusion/exclusion decisions based on label training data [10,11]. Active learning methods, such as those implemented in softwares such as Covidence [12] or ASReview [13], prioritize the most relevant citations for initial screening. However, these traditional methods still require human involvement to make final decisions on study inclusion, and they do not automate other stages of the review (eg, extraction). Rule-based and natural language processing (NLP)–driven named entity recognition approaches have been used for structured data extraction from full-text articles [14]. Recent developments in deep learning, particularly transformer-based models such as BERT and SciBERT, have improved the performance of such tasks by capturing more nuanced linguistic features [15-17]. The recent advancement in NLP have further contributed toward the development of much-needed extraction tools [18-20]. While these pilot programs demonstrated strong performance, they were largely designed for the extraction of clinical trial data, limiting their application across study designs. Recent focus has shifted toward the popular platform, OpenAI's ChatGPT [21]. This accessible LLM software is proficient at understanding and processing human language with high speed and accuracy. As the tool can effectively interpret information from complex texts, it may have strong potential for application in the extraction of systematic reviews. In addition, OpenAI recently released the functionality to build customizable versions of ChatGPT. Users can create a Generative Pre-training Transformer (GPT) and tailor it for certain tasks by providing it instructions and assumed knowledge [22]. These applications can be saved and accessed by other users. With the recent GPT functionality, health researchers can now create and share tools specifically designed to carry out systematic review data extraction.

Despite recent advances in AI technology, only three studies to date have reported using LLMs for data extraction. One study briefly reported on the feasibility of ChatGPT for data extraction [23], while others evaluated the performance of ChatGPT-4 [24] and ChatGPT-3.5 [25] ; these studies found moderate performance for extracting complex information but high accuracy for simpler extraction fields [24,25]. While these studies are the first to evaluate the performance of ChatGPT as an extraction tool, further efforts are needed to improve the availability of these programs.

We present a custom GPT program tailored for systematic review extraction, made available as open access. In the current study, we aim to provide a pilot evaluation of the extraction tool, using data from two systematic reviews. Performance of the custom GPT will be compared to that of a human extractor, examining rates of agreement and time taken to extract data.

Methods

GPT Building

OpenAI's GPT (version 3.5) Builder allows users to create custom versions of ChatGPT that perform specific tasks by combining instructions, knowledge, and capabilities. The GPT builder interface was used to develop a specialized GPT to extract information from academic manuscripts to assist with the extraction phase of two systematic reviews: a methodological review and a systematic review of interventions. The developed GPT is named *Systematic Review Extractor Pro (*Figure 1*)*.

XSL•FO
RenderX

Figure 1. The OpenAI custom GPT Builder interface used to develop and test automated data extraction (1–Instructions, 2–Knowledge, and 3–Capabilities).

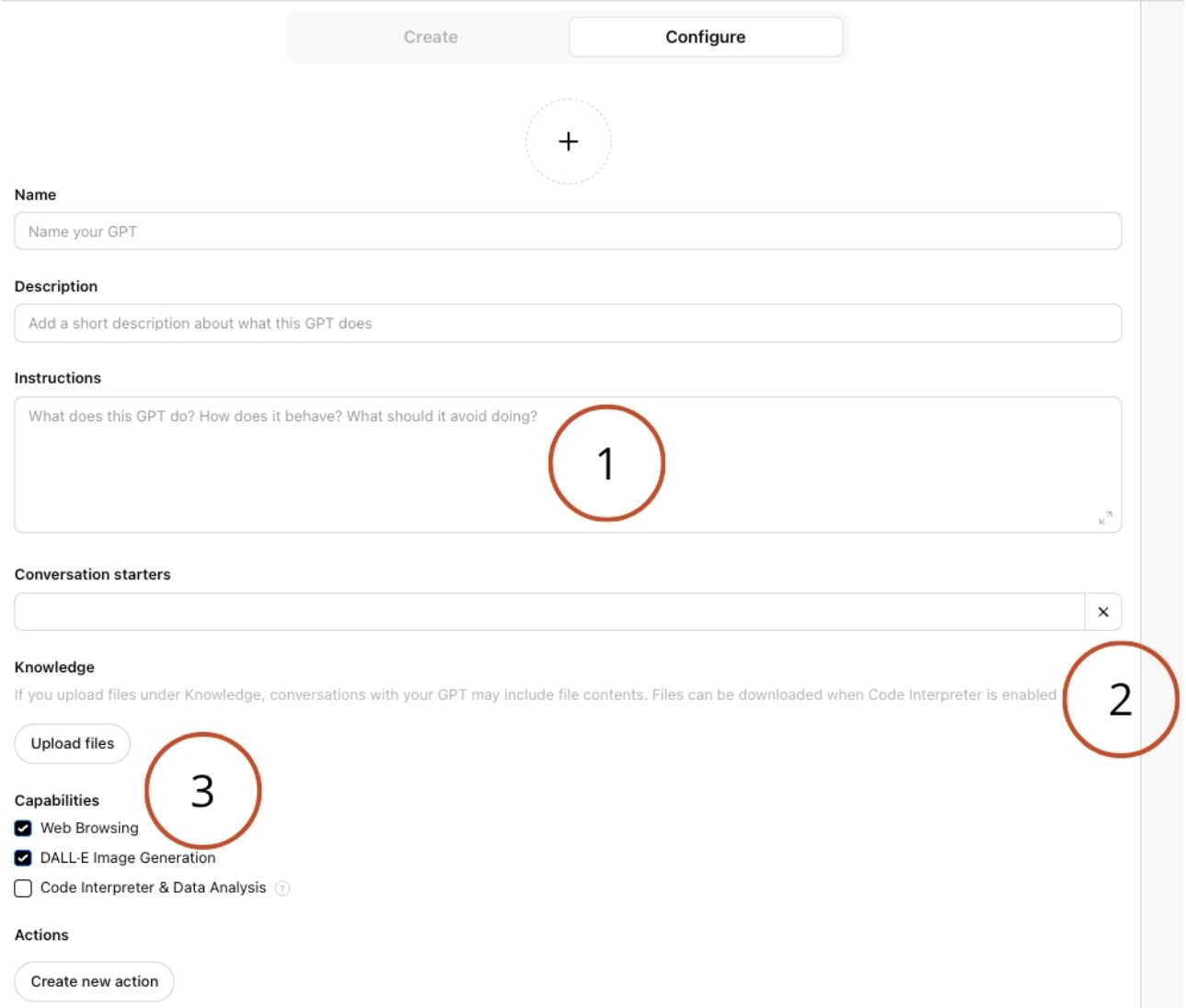

Prompt Engineering

Detailed prompt engineering guidelines were followed under the Role, Instruction, Steps, End goal, and Narrowing (RISEN) framework [22] to manually configure the GPT's base context (purpose) and stepped actions to improve its performance in completing the required tasks. Each term refers to specific prompt components that improve the output quality. *Role* provides an identity to the GPT dictating the manner in which it acts, that is,"As an expert in the conduct of systematic reviews you are to...." *Instructions* inform the GPT what task it is to perform, while *Steps* provide a hierarchy of instructions that the GPT is to follow to perform the desired task. *End goal* informs the GPT on the format of desired output and content. Finally, *Narrowing* provides constraints to the GPT including key requirements for performing the given task (eg, setting a word limit on the GPT's response output).

Iterative Development of the Tool

The RISEN prompt used to instruct the GPT was iteratively fine-tuned to improve its performance by working through five iterations of the detailed instructions. Differences between iterations were generally focused on providing greater specificity to the prompt and detailed steps to carry out the extraction. Any vagueness in the *Steps* would lead to diverse response ranges for extraction variables, and therefore required the greatest number of changes and troubleshooting. Any changes made to the prompt were systematically tested on dummy health research manuscripts to optimize performance. Due to higher inaccuracy in early stages of extracting statistical results in the template, OpenAI's user guidelines were followed, to iteratively improve the tool [22]. These guidelines suggest 'splitting' or 'chunking' complex tasks into smaller ones.

Reviews for Extraction

A random sample of 10 studies from each of the two systematic reviews (total of 20 studies) were selected to evaluate the performance of the GPT. Review 1 is a systematic review of ecological momentary assessment (EMA) research methodologies for measuring substance use and associated behaviors. Full details are provided in the review protocol which was preregistered on the International Prospective Register of Systematic Reviews (PROSPERO: CRD42023400418). Of specific importance to this review was the extraction of methodological characteristics relating to ecological momentary

XSL•FO
RenderX

assessment techniques. Review 2 is a systematic review of the effectiveness of wellbeing interventions for helping professionals. For comprehensive details, refer to the review protocol preregistered on PROSPERO (CRD42023422224). As a review assessing intervention effectiveness, this study included the extraction of statistical results.

Agreement

Total proportion of agreement was calculated by summing the number of instances where the GPT and human reviewer agreed on extracted data, divided by the total number of applicable fields. Agreement was computed for each study according to subsection, which included study characteristics, participant characteristics, methodological characteristics (only relevant for review 1), and statistical results (only relevant for review 2). See Multimedia Appendices 1 and 2 for the full list of variables in each subsection. A brief qualitative synthesis of the nature of errors was also conducted.

Timing

The time taken to extract data for each study was recorded for both the GPT and human reviewer. To enhance practicality, we also recorded the time taken for a human to review the GPT output.

Ethical Considerations

This study is a methodological paper which conducts secondary analysis on deidentified and aggregated data and as such, is exempt from requiring ethical approval as per the Australian National Statement on Ethical Conduct in Human Research 2023 [26].

Results

Time Taken

The mean extraction time for the human reviewers was 36 minutes. For GPT, the mean time for extraction was 26.6 seconds, and an additional 13 minutes for human review.

Agreement

For the completed extraction templates for the GPT and human reviewers, see Multimedia Appendix 1 (review 1) and Multimedia Appendix 2 (review 2). As seen in Table 1, overall agreement between the GPT and human reviewer across the 10 studies in review 1 was 91.45%; agreement was the highest for study (95.25%) and participant characteristics (95.03%) and the lowest for methodological characteristics (87.07%). For the 10 studies in review 2, overall agreement was 89.31%, similar to review 1. The highest agreement was observed in study (90.83%) and participant (90.00%) characteristics. Lower agreement was recorded in extraction fields relating to statistical results (77.50%).

Table . Agreement between human reviewers and GPT for data extraction by manuscript subsections.

Manuscript subsection	Review 1 (agreement %)	Review 2 (agreement %)
Study characteristics	95.25	90.83
Participant characteristics	95.03	90.00
Methodological characteristics	87.07	__[a]
Results	__[a]	77.50
Overall agreement	91.45	89.31

[a]Data not applicable for review.

Qualitative Summary of Errors

Across both Review 1 and 2, the most common error made by the GPT was reporting an extraction field as “not specified,” despite the relevant information being present in the study manuscript (see Multimedia Appendices 1 and 2). These types of errors accounted for a high proportion of the inaccuracies, particularly where only p-values were sometimes reported rather than the required measures of effect.

While there were instances of errors resembling hallucination (eg, describing a randomized controlled trial as a pre-post study), these were few in number. When the GPT erroneously inserted information, it typically mistakenly drew data from other parts of the paper, rather than fabricating the information entirely, For example, several errors occurred in sample size extraction, such as reporting the baseline sample as the follow-up sample.

Instances Where the GPT Was More Accurate That the Human Reviewer

In review 1, there were 14 instances where the GPT correctly extracted fields, and the human reviewer was incorrect (equating to 3.25% of the overall number of fields). In Review 2, the GPT was correct in four fields where the human reviewer was incorrect with a rate of 1.16%.

Discussion

Principal Findings

This study evaluated the performance of a custom-configured LLM, specifically a GPT, in automating the data extraction phase of two systematic reviews. This application of AI for extraction provides initial insights into addressing challenges in systematic review methodology, such as time consumption, labor intensity, and potential for human error.

XSL•FO
RenderX

Even when accounting for human review, the custom GPT took less than 14 minutes to complete extraction of each study, 22 minutes faster than human extraction. The 26.6 seconds it took for extraction alone was comparable to a study by Gue et al [25], where ChatGPT-3.5 took approximately 17 seconds per study compared to 77 minutes by a reviewer. As such, the findings of prior research alongside the current study emphasize the processing speed of AI and its potential to significantly reduce the time demands and resultant costs associated with systematic reviews. This efficiency could facilitate more timely summaries of evidence-based practices and support researchers in managing large volumes of literature.

The high overall agreement rates observed in this study between the GPT and human extraction for review 1 (91.45% overall) and 2 (89.31%), suggest that LLMs may have potential to support automation of the extraction of data from academic manuscripts. Consistent across review 1 and 2, the GPT was particularly effective at extracting study (95.25% and 90.83%) and participant characteristics (95.05% and 90%, respectively). This is likely due to the homogeneity in participant and study characteristics that are reported in the extracted manuscripts, reflecting reporting standards upheld by academic journals.

However, agreement rates were lower when extracting methodological characteristics (77.88%) and results (77.50%) sections. Early iterations were improved by 'splitting' the task by extracting results separately to the rest of the extraction template. Results are challenging for GPT to extract, as studies used varied methods of analysis and some statistical results were only reported in tables. As LLMs favor extracting data from text information, additional prompting was required to ensure that tables were thoroughly and consistently incorporated into the information extraction sequence.

In this study, a small percentage of instances where the GPT outperformed the human extractor were observed. This demonstrates the potential for AI to identify patterns and insights that may be overlooked in human extraction. Human extractors can be prone to fatigue which may have led to these small inconsistencies. The preliminary findings of this study suggest that LLMs may have utility in identifying errors made by human extractors.

Overall, in accordance with previous research, LLMs excel in quickly extracting simple data (eg, participant demographics [25]. The purpose-built GPT evaluated in the current study was less accurate while extracting more complex fields compared to simple fields (agreement ranged from 77.50% to 95.25%) but appears to be relatively accurate in comparison to extraction using the generic ChatGPT model where agreement rates were as low as 41.2% [25]. A study of 87 human reviewers found error rates of 28.3% to 31.2%, corresponding to agreement rates of 68.8% and 71.7% with a gold-standard reviewer [27]. These rates remained relatively consistent across reviewer experience level and, similar to the findings of the current study, were more accurate for the extraction of participants and study characteristics compared to statistical results. Overall, the GPT evaluated in the current study demonstrated comparable error rates to those reported by human reviewers.

In its current format, researchers may consider the use of this tool as a second extractor, when they lack funding, resourcing, or time for a team member to perform the task (Textbox 1). The overall agreement rate of the custom GPT was approximately 90%, and the GPT had utility in identifying some human errors. Caution should be exercised while using the tool to extract information in areas where it showed lower accuracy, in particular results and methodological characteristics.

Textbox 1. A guide to using the Systematic Review Extractor Pro for data extraction.

How to use the tool

1. Log into a ChatGPT account and access the 'Systematic Review Extractor Pro' GPT [28].
2. Copy and paste your extraction template variable headings into the GPT message bar (any text format).
3. Upload a PDF file of the study from which you want to extract data.
4. Copy the output to an Excel file. If needed, use the 'Convert Text to Table' functionality, separating by '|'.

Optimizing use of this tool

- Be explicit with the outcomes in your extraction template. For example, "Mental health outcomes (specify scales and subscales used in bullet point format)" or "modality (in-person; digital; telephone)." For an example, see the authors templates in Multimedia Appendices 1 and 2.
- With more complex aspects of the template such as statistical results, you may have better accuracy by 'chunking' the task. This involves splitting complex tasks into simpler ones. To do this, provide the tool with the results section of the extraction template separate to the rest of the template.

Citing this tool and other considerations

- To reference the use of this tool in publications, please cite this paper.
- Consider the copyright limitations of articles when uploading them to the tool.

Strengths and Limitations

Building on previous work exploring the capacity of ChatGPT 3, 3.5 and 4 to perform data extraction [23-25], this study tested the performance of a custom-built GPT across two distinct systematic reviews. To the best of our knowledge, this study highlights the first empirical investigation of a custom OpenAI GPT to perform data extraction in a systematic review.

XSL•FO
RenderX

Our study has several limitations. When determining agreement between extractors, there is always a degree of subjectivity between coding any qualitative responses, which can either overestimate or underestimate agreement rates. The current study observed agreement rates across 20 studies. Testing a larger sample size would improve confidence in the tool. In terms of limitations of the tool itself, the tool's construction on the GPT Builder may pose replicability challenges of the output due to the constantly evolving nature of the LLM. The benefit of this aspect is that the tool will automatically update as ChatGPT releases improved versions of their model [29]. However, future research should evaluate the test-retest validity of this custom GPT.

Conclusions

This study serves as promising early evidence for the application of a custom-built GPT in conducting systematic review data extraction. The findings of this study are consistent with the growing sentiment suggesting AI can enhance the efficiency of systematic reviews and reduce cost and time while maintaining accuracy [4]. By encouraging open discourse regarding the role of AI in research, researchers can contribute to the development of robust, transparent, and reproducible scientific practice. This is a rapidly advancing field of technology and OpenAI frequently releases new versions of ChatGPT that show improved performance and efficiency [29]. It is likely that with further advancement of AI, LLMs may be relied on as a sole reviewer for data extraction.

Acknowledgments

JS and ZB are supported by Postgraduate Scholarship through the University of Sydney. JW is supported by a NHMRC fellowship funding. While this manuscript describes the use of generative AI as a tool for systematic review data extraction, the authors attest that no generative AI was used in the writing of this manuscript.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Authors' Contributions

Conceptualization: JB, JS
Formal analysis: JB, JS
Investigation: JS
Writing–original draft: JB, JS, JW
Writing–review & editing: JS, JW

Conflicts of Interest

None declared.

Multimedia Appendix 1
Systematic review 1 - agreement file.
[XLSX File, 27 KB - formative_v9i1e68666_app1.xlsx]

Multimedia Appendix 2
Systematic review 2 - agreement file.
[XLSX File, 45 KB - formative_v9i1e68666_app2.xlsx]

References

1. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med 2022 Jan;28(1):31-38. [doi: 10.1038/s41591-021-01614-0] [Medline: 35058619]
2. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ. In: Welch VA, editor. Cochrane Handbook for Systematic Reviews of Interventions Version 65: Cochrane; 2024. URL: https://www.training.cochrane.org/handbook [accessed 2024-09-15]
3. Borah R, Brown AW, Capers PL, Kaiser KA. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. BMJ Open 2017 Feb 27;7(2):e012545. [doi: 10.1136/bmjopen-2016-012545] [Medline: 28242767]
4. Michelson M, Reuter K. The significant cost of systematic reviews and meta-analyses: a call for greater involvement of machine learning to assess the promise of clinical trials. Contemp Clin Trials Commun 2019 Dec;16:100443. [doi: 10.1016/j.conctc.2019.100443] [Medline: 31497675]

XSL•FO
RenderX

5. Mathes T, Klaßen P, Pieper D. Frequency of data extraction errors and methods to increase data extraction quality: a methodological review. BMC Med Res Methodol 2017 Nov 28;17(1):152. [doi: 10.1186/s12874-017-0431-4] [Medline: 29179685]
6. Hoffmann F, Allers K, Rombey T, et al. Nearly 80 systematic reviews were published each day: observational study on trends in epidemiology and reporting over the years 2000-2019. J Clin Epidemiol 2021 Oct;138:1-11. [doi: 10.1016/j.jclinepi.2021.05.022] [Medline: 34091022]
7. Johnson BT, Hennessy EA. Systematic reviews and meta-analyses in the health sciences: best practice methods for research syntheses. Soc Sci Med 2019 Jul;233:237-251. [doi: 10.1016/j.socscimed.2019.05.035]
8. Chalmers I, Fox DM. Increasing the incidence and influence of systematic reviews on health policy and practice. Am J Public Health 2016 Jan;106(1):11-13. [doi: 10.2105/AJPH.2015.302915] [Medline: 26562111]
9. van Dijk SHB, Brusse-Keizer MGJ, Bucsán CC, van der Palen J, Doggen CJM, Lenferink A. Artificial intelligence in systematic reviews: promising when appropriately used. BMJ Open 2023 Jul 7;13(7):e072254. [doi: 10.1136/bmjopen-2023-072254] [Medline: 37419641]
10. O'Mara-Eves A, Thomas J, McNaught J, Miwa M, Ananiadou S. Using text mining for study identification in systematic reviews: a systematic review of current approaches. Syst Rev 2015 Jan 14;4(1):5. [doi: 10.1186/2046-4053-4-5] [Medline: 25588314]
11. Wallace BC, Small K, Brodley CE, Trikalinos TA. Active learning for biomedical citation screening. Presented at: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining; Jul 25, 2010; Washington DC USA p. 173-182. [doi: 10.1145/1835804.1835829]
12. Covidence systematic review software. 2025. URL: https://www.covidence.org/ [accessed 2024-09-09]
13. van de Schoot R, de Bruin J, Schram R, et al. ASReview: Open source software for efficient and transparent active learning for systematic reviews. Int J Digit Curation 2021;16(1):123-133. [doi: 10.2218/ijdc.v16i1.726]
14. Jonnalagadda SR, Goyal P, Huffman MD. Automating data extraction in systematic reviews: a systematic review. Syst Rev 2015 Dec;4(1):78. [doi: 10.1186/s13643-015-0066-7]
15. Beltagy I, Lo K, Cohan A. SciBERT: a pretrained language model for scientific text. Presented at: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); Nov 7-10, 2019; Hong Kong, China p. 3615-3620. [doi: 10.18653/v1/D19-1371]
16. Khalil H, Ameen D, Zarnegar A. Tools to support the automation of systematic reviews: a scoping review. J Clin Epidemiol 2022 Apr;144:22-42. [doi: 10.1016/j.jclinepi.2021.12.005] [Medline: 34896236]
17. van de Schoot R, de Bruin J, Schram R, et al. An open source machine learning framework for efficient and transparent systematic reviews. Nat Mach Intell 2021 Feb;3(2):125-133. [doi: 10.1038/s42256-020-00287-7]
18. Golinelli D, Nuzzolese AG, Sanmarchi F, et al. Semi-Automatic systematic literature reviews and information extraction of COVID-19 scientific evidence: description and preliminary results of the COKE project. Information 2022;13(3):117. [doi: 10.3390/info13030117]
19. Panayi A, Ward K, Benhadji-Schaff A, Ibanez-Lopez AS, Xia A, Barzilay R. Evaluation of a prototype machine learning tool to semi-automate data extraction for systematic literature reviews. Syst Rev 2023 Oct 6;12(1):187. [doi: 10.1186/s13643-023-02351-w] [Medline: 37803451]
20. Zhang T, Yu Y, Mei J, Tang Z, Zhang X, Li S. Unlocking the power of deep PICO extraction: step-wise medical NER identification. arXiv. 2020. URL: http://arxiv.org/abs/2005.06601 [accessed 2024-09-10]
21. ChatGPT. Open AI. 2023. URL: https://openai.com/chatgpt/overview/ [accessed 2025-08-07]
22. Prompt engineering. Open AI. 2024. URL: https://platform.openai.com/docs/guides/prompt-engineering [accessed 2024-07-03]
23. Mahuli SA, Rai A, Mahuli AV, Kumar A. Application ChatGPT in conducting systematic reviews and meta-analyses. Br Dent J 2023 Jul;235(2):90-92. [doi: 10.1038/s41415-023-6132-y] [Medline: 37500847]
24. Khraisha Q, Put S, Kappenberg J, Warraitch A, Hadfield K. Can large language models replace humans in systematic reviews? Evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. Res Synth Methods 2024 Jul;15(4):616-626. [doi: 10.1002/jrsm.1715] [Medline: 38484744]
25. Gue CCY, Rahim NDA, Rojas-Carabali W, et al. Evaluating the OpenAI's GPT-3.5 Turbo's performance in extracting information from scientific articles on diabetic retinopathy. Syst Rev 2024 May 16;13(1):135. [doi: 10.1186/s13643-024-02523-2] [Medline: 38755704]
26. Council and Universities Australia. National Statement on Ethical Conduct in Human Research: National Health and Medical Research Council; 2023. URL: https://www.nhmrc.gov.au/about-us/publications/national-statement-ethical-conduct-human-research-2023 [accessed 2024-09-10]
27. Horton J, Vandermeer B, Hartling L, Tjosvold L, Klassen TP, Buscemi N. Systematic review data extraction: cross-sectional study showed that experience did not increase accuracy. J Clin Epidemiol 2010 Mar;63(3):289-298. [doi: 10.1016/j.jclinepi.2009.04.007] [Medline: 19683413]
28. Systematic review extractor pro. ChatGPT. URL: https://chatgpt.com/g/g-5IRhGHEcc-systematic-review-extractor-pro [accessed 2025-08-07]

XSL•FO
RenderX

29. OpenAI, Achiam J, Adler S, Agarwal S, et al. GPT-4 technical report. arXiv. 2024 Mar 4. URL: http://arxiv.org/abs/2303.08774 [accessed 2024-09-10]

Abbreviations

AI: artificial intelligence
GPT: Generative Pre-Training Transformer
LLM: large language model
RISEN: Role, Instruction, Steps, End goal, and Narrowing

Edited by A Mavragani; submitted 11.11.24; peer-reviewed by AK Dubey, J Zheng; revised version received 01.07.25; accepted 03.07.25; published 11.08.25.

Please cite as:
Sercombe J, Bryant Z, Wilson J
Evaluating a Customized Version of ChatGPT for Systematic Review Data Extraction in Health Research: Development and Usability Study
JMIR Form Res 2025;9:e68666
URL: https://formative.jmir.org/2025/1/e68666
doi:10.2196/68666

© Jayden Sercombe, Zachary Bryant, Jack Wilson. Originally published in JMIR Formative Research (https://formative.jmir.org), 11.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Acceptability and Feasibility of a Prototype Regional Disaster Teleconsultation System for COVID-19 Pandemic Response: Pilot Field Test

Stephanie Ludy[1,2], MD; Mari-Lynn Drainoni[3,4,5], PhD; Lauren Schmidt[6], MPH; Mark Litvak[7], MBA; Julianne Dugas[8], MPH; Eric Goralnick[1,9], MD, MS; Paul D Biddinger[2,9], MD; Tehnaz Parakh Boyle[7,9], MD, PhD

[1]Department of Emergency Medicine, Harvard Medical School, Brigham and Women's Hospital, Boston, MA, United States

[2]Department of Emergency Medicine, Harvard Medical School, Massachusetts General Hospital, Boston, MA, United States

[3]Department of Medicine, Section of Infectious Diseases, Boston University Chobanian and Avedisian School of Medicine, Boston Medical Center, Boston, MA, United States

[4]Department of Health Law, Policy and Management, Boston University School of Public Health, Boston, MA, United States

[5]Evans Center for Implementation and Improvement Sciences, Boston University Chobanian and Avedesian School of Medicine, 801 Albany Street, Boston, MA, United States

[6]Department of Pediatrics, Boston Medical Center, Boston, MA, United States

[7]Department of Pediatrics, Boston University Chobanian and Avedisian School of Medicine, Boston Medical Center, 801 Albany Street, Rm 4027, Boston, MA, United States

[8]Department of Emergency Medicine, Boston Medical Center, Boston, MA, United States

[9]Region 1 Disaster Health Response System, Boston, MA, United States

Corresponding Author:
Tehnaz Parakh Boyle, MD, PhD
Department of Pediatrics, Boston University Chobanian and Avedisian School of Medicine, Boston Medical Center, 801 Albany Street, Rm 4027, Boston, MA, United States

Abstract

Background: Disaster telehealth can be used to provide rapid access to remote specialty expertise and virtual surge capacity for overwhelmed local clinicians. The Regional Disaster Health Response System (RDHRS) is developing a disaster teleconsultation system for cross-jurisdictional care in the United States. In 2020, the Region 1 RDHRS provided Massachusetts hospitals access to disaster teleconsultation services with out-of-state critical care experts during the first wave of the COVID-19 pandemic response.

Objective: We aimed to field-test (1) the acceptability and feasibility of using a prototype, web-based disaster teleconsultation platform with minimal-to-no user training and (2) the feasibility of deploying a national volunteer expert pool to access out-of-state expertise.

Methods: This was a prospective, mixed methods, observational study. We recruited field clinicians from Massachusetts hospitals and out-of-state critical-care physicians as experts for a 2-week pilot (June 2020). Experts were trained to use a prototype platform, while field clinicians received a just-in-time tool. Field clinicians requested teleconsultations for hospitalized patients with COVID-19 (clinical call) or simulated patients (test call). We collected demographics, call performance data, and Telehealth Usability Questionnaire (TUQ) ratings to measure acceptability (primary outcome; total usability score ≥6 of 7) and feasibility (secondary outcome; interface, interaction quality, and reliability items), and interviewed participants. We report descriptive statistics and key themes using the Technology Acceptance Model framework.

Results: Ten experts from 6 states and 17 field clinicians from 4 hospitals participated. All experts and 10 field clinicians completed postpilot questionnaires (74% response overall). Of these, 20% had previously used telemedicine in a disaster. In total, 50 test calls and no clinical calls were logged. Most (70%) made ≥1 call; 22% (95% CI 10% - 34%) connected successfully. The median time to connect was 1.6 (IQR 3.2) minutes. Among field clinician respondents, 50% used smartphone devices, 40% hospital desktop computers, and 10% laptop computers to access RDHRS teleconsultation services. Calls failed due to platform routing errors (49%), hospital computers without cameras or microphones (10%), firewalls (8%), and expert notification failures (5%). The mean total usability score was 5.6 (SD 1.3). TUQ item scores were highest in usefulness (mean 6.0, SD 1.1) and ease-of-use (mean 6.0, SD 1.4), and lowest in reliability (mean 2.4, SD 1.4). Participants were comfortable using the platform. Those with difficulty identified discomfort with technology as the cause. All experts were willing to participate in a national expert registry and obtain emergency licensure, and most (80%) were willing to serve on a volunteer, unpaid basis.

XSL•FO
RenderX

Conclusions: Clinicians found the prototype platform acceptable, but the workflow requires revision to reduce call failure and improve feasibility and reliability for future use with minimal-to-no training. Using familiar clinical workflows for emergency consultation and mobile devices with camera and microphone capabilities could improve call performance and reliability.

(JMIR Form Res 2025;9:e73078) doi:10.2196/73078

KEYWORDS

disaster medicine; teleconsultation; regional medical programs; COVID-19; tele-critical care

Introduction

In 2018, the Administration for Strategic Preparedness and Response launched the Regional Disaster Health Response System (RDHRS) program as part of the national strategic response to disasters in the United States [1]. The goal of an RDHRS is to establish multi-state partnerships and enhance local health care assets to support care delivery when catastrophic events overwhelm local and state capacity and capability. Telehealth is an important RDHRS tool to expand access to specialty care and increase surge capacity. Disaster telehealth services during hurricanes, wildfires, and pandemics have helped maintain access to health care when traditional in-person services are disrupted [2-5]. Delivering telehealth services across state lines and jurisdictions has been limited by poor interoperability of telehealth platforms, variable infrastructure, and regulatory barriers related to licensure, credentialing, and reimbursement [6].

The Region 1 RDHRS, which serves the 6 New England states, is developing a disaster teleconsultation system to provide rapid, temporary, cross-jurisdictional access to clinical specialists to support disaster medical response. We previously demonstrated proof-of-concept for a model teleconsultation system operating within a regional medical emergency operations center (MEOC) in simulated disasters [7]. In this model system, the regional MEOC coordinates access to an expert clinician pool to provide temporary disaster teleconsultation services to field clinicians in affected hospitals. While this model had high clinician acceptance, it required a simpler, more intuitive communication interface to support efficient response in real-world settings. The initial prototype model used a telehealth platform in clinical use at one academic medical center to connect to hospitals outside its health system. It required a trained telehealth coordinator to triage patients by telephone, connect bedside clinicians with remote expert teleconsultants, and guide just-in-time installation and use of software. However, dependence on external guidance for technology use and manual triage hampered workflow efficiency and system throughput.

In February 2020, as the first COVID-19 cases skyrocketed, 3 factors motivated the Region 1 RDHRS team to develop and field-test another prototype platform for real-world use. First, large surges in critical patients produced severe equipment and workforce shortages, overwhelming regional critical care capacity [8]. Second, COVID-19 led to the widespread adoption of video technology that made patients, providers, and health care organizations more receptive to using telehealth tools [9]. Third, the Coronavirus Aid, Relief, and Economic Security Act and state emergency laws and waivers lifted legal, regulatory, and reimbursement barriers to the implementation of regional telehealth systems [10,11].

In early March 2020, we collaborated with industry partners to configure a new disaster telehealth platform for pilot field implementation. Our primary objective was to test the feasibility and acceptability of providing Massachusetts hospitals temporary access to disaster teleconsultation services with out-of-state critical care experts during the first wave of the COVID-19 pandemic response. Our secondary objectives were to test the feasibility of deploying a national volunteer expert pool and identify best practices and policy recommendations to support implementation in future disasters.

Methods

Study Design

This was a prospective observational study. We used mixed methods to capture clinician ratings and feedback after a 2-week pilot field implementation of the prototype RDHRS teleconsultation system. We included a test call strategy to evaluate call performance in conditions approximating real-world use if clinical use lagged. We followed STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines for reporting observational studies.

Ethical Considerations

The Boston Medical Center/Boston University Medical Campus institutional review board approved this study and granted an exemption under Category 2 of the Human Research Protection Program policies and procedures. Although we used an informed consent process, the study qualified for a *Waiver of Documentation of Informed Consent* due to the minimal risk of harm to participants and the nature of the study procedures, which typically do not require written consent outside of a research context. Participant data was deidentified and assigned a unique study identifier, with the master code stored separately and accessible only to the study team. Participation was voluntary and unremunerated.

Setting

The study was conducted in Massachusetts during the first wave of COVID-19 response in 2020. We gave Massachusetts hospitals temporary access to disaster teleconsultations using a national volunteer pool of critical care experts. Based on participating hospital needs, disaster teleconsultation services were available on demand for 12 hours/day (5 PM-5 AM) from June 1 to June 15, 2020.

XSL•FO
RenderX

Eligibility and Recruitment

We recruited subject matter experts (SMEs) from the Society for Critical Care Medicine as RDHRS teleconsultants via email invitation to a list of 100 member volunteers interested in testing new technologies for remote care delivery provided by the Society. Eligible SMEs had a full medical license in good standing in any US state, territory, or district. Participation was part-time.

All Massachusetts hospitals were invited to participate as field test sites by email invitation sent by the Massachusetts Hospital Association and local health care coalitions on behalf of the RDHRS. Hospitals interested in accessing RDHRS teleconsultation services were introduced to the goals of the pilot in follow-up virtual meetings with the RDHRS team. There was no fee for service use. SME and hospital participation was voluntary.

Prototype RDHRS Teleconsultation Platform and Workflow

After testing off-the-shelf video-conferencing platforms, we selected a Health Insurance Portability and Accountability Act (HIPAA)–compliant telehealth platform (Bluestream Health, Inc.) for disaster-relevant capabilities. This commercially available, web-based application had low bandwidth requirements, did not require software or hardware installation, and was device-agnostic and interoperable to support a distributed care delivery model. The platform was configured for relevant workflows (virtual interpreter and tele-emergency services) with remote high-throughput triage capability, virtual team functionality, and rapid scalability, and it was FirstNet-certified for priority network access. Key security features included end-to-end encryption for all data transmissions, secure communication channels to protect patient information during teleconsultations, robust authentication protocols to verify user identities, and restricted access to safeguard sensitive health data.

Pilot Field Implementation Process

SME Onboarding

SMEs were onboarded as RDHRS teleconsultants as follows: (1) each out-of-state volunteer obtained an emergency temporary medical license to comply with Massachusetts law, (2) physicians unable to use their existing malpractice insurance signed a volunteer physician practice agreement with an RDHRS partner institution (Boston Medical Center) to receive secondary liability coverage during pilot field implementation, and (3) all SMEs tested device and internet connections and were trained to technical competency in platform use. SMEs also received a just-in-time user manual with graphic instructions on how to receive and answer video consultations for reference during and after training.

Hospital Onboarding

Each participating hospital received an implementation toolkit describing (1) RDHRS program goals, (2) existing emergency legislation and regulatory waivers, (3) SME licensure and credentials, (4) minimum technology requirements, and (5) teleconsultation service parameters. Hospitals used emergency credentialing procedures written in their bylaws. Hospital technology teams were given information to bypass firewall protections by whitelisting.

User Registration

To maintain cybersecurity standards, each SME and hospital user was registered by the technology vendor prior to service activation. This included configuring the platform to route consult requests from registered hospitals through the MEOC to the SME.

Field clinicians were given password-protected access and electronic copies of the just-in-time user manual but otherwise received no training or exposure to the platform. Field clinicians were instructed to request teleconsultation for hospitalized adult patients with COVID-19 (clinical call) or simulated patients (test call) on demand.

Assessment Tools

Quantitative Tools

After the pilot period, all participants completed data forms to report demographics and rate the platform on the Telehealth Usability Questionnaire (TUQ). The TUQ has 21 items validated to assess telehealth platform usefulness, ease of use, interface quality, interaction quality, reliability, and satisfaction on 7-point Likert scales [12]. A total usability score (TUS) can be calculated as the mean user rating for all 21 items (maximum score=7).

Field clinicians also received a brief questionnaire to query user comfort using the RDHRS prototype system without prior training or exposure, device types used to request teleconsultations, success completing a video call on the first attempt, reasons for call failure, success completing a video call during any attempt, comfort making future teleconsultation requests, user preferences for device type to receive video consults during disasters, willingness to use personal devices, and preferences for notification of teleconsultant availability.

Qualitative Tools

To supplement survey assessments and contextualize our findings, we also conducted individual interviews with those who completed surveys. We designed a semistructured interview guide to explore participants' overall perception and comfort level with the platform, device preferences, challenges with use, and recommendations for future implementation.

Outcomes

The primary outcome, acceptability, was measured as TUS≥6 of 7. The secondary outcome, feasibility, was measured via select TUQ items in interaction quality and reliability domains.

Call performance measures were recorded and downloaded from the platform, including the frequency of calls attempted, calls connected, video calls completed successfully, reasons for call failure, and wait times. A connection was successful when a field clinician completed a synchronous video call with the SME. Reasons for call failure included user cancellation (call canceled manually by the requester), orphaned call (call canceled by the server when the SME could not be contacted or the server lost contact with the requester before the SME answered), call

XSL•FO
RenderX

routing error (error in call routing from the hospital to the MEOC), expert notification error, device problem, and connection problem.

Data Collection

Questionnaires were administered electronically immediately after the pilot period ended and stored using the REDCap (Research Electronic Data Capture; Vanderbilt University) platform hosted by Boston University [13]. All participants were given 2 weeks with 3 email reminders to complete questionnaires.

After survey completion, brief interviews (30 min) with willing participants were audio-recorded using Zoom Pro by 2 study investigators. Audio recordings were transcribed verbatim by a third-party transcription service.

Sample Size

As a pilot field test, this study was not powered to detect statistically significant outcomes. We determined that a minimum of 30 experts would be required to ensure 24/7 on-demand teleconsultation. Our staffing model assumed that each physician would volunteer for three 4-hour shifts during the 2-week pilot period. Based on preliminary data, we needed approximately 20 participants to estimate acceptability with adequate precision. Assuming each encounter involved 1 field clinician and 1 expert, we anticipated a minimum of 10 teleconsultation encounters during the pilot period.

Analysis

Quantitative Analysis

We used SAS (version 9.4; SAS Institute) software for descriptive statistics and NVivo (version 12; Lumivero) software for qualitative coding and analysis. We report means with SD, medians with IQR, and proportions with 95% CI as appropriate.

Qualitative Analysis

To code transcripts, we revised an inductive codebook developed previously [7]. Two investigators (SL and LS) independently reviewed and coded 2 transcripts using the prior codebook and met to establish consensus and revise codes for the final codebook. The 2 investigators then independently coded each transcript using the final codebook, with frequent meetings to reconcile discrepancies by consensus. After coding was complete, themes were created and mapped to the Technology Acceptance Model framework to identify factors influencing acceptance of disaster teleconsultation services by SMEs and field clinicians [14]. In this theoretical framework, how providers perceive the *usefulness* (use enhances or diminishes care delivery) and *ease-of-use* (use is easy or difficult) of a technology solution affects their *attitude* (positive or negative emotion) and *behavioral intention* (willingness) to adopt and use it in the future.

Results

User Statistics

Of 100 Society of Critical Care Medicine members who volunteered to test new technologies, 66 SMEs were eligible for recruitment based on board certification. Among these, 27/66 (41%) responded to an invitation to volunteer as an RDHRS teleconsultant. Ultimately, 10/66 (15%) eligible SMEs from 6 US states completed the onboarding process and participated in the study.

In total, 4 community hospitals and 3 tertiary care centers in Massachusetts expressed interest in participating in the pilot. Of these, 2 community hospitals and 2 tertiary care centers completed their administrative review and onboarding in time to participate in the pilot. A total of 17 board-certified internal and emergency physicians from these 4 hospitals participated as field clinicians.

Of these, 10 field clinicians and 10 SMEs completed post-pilot questionnaires (75% response rate). Table 1 summarizes participant demographics.

XSL•FO
RenderX

Table . Demographic data for participating field clinicians (n=10) and expert teleconsultants (n=10), including self-reported sex, age, specialty, years of clinical experience, primary practice setting, and prior utilization of telemedicine either clinically or in a prior disaster setting.

Demographics	Values
Sex, n (%)	
Male	11 (55)
Female	9 (45)
Age in years, median (IQR)	44 (14)
Medical specialty, n (%)	
Anesthesiology critical care	4 (20)
Emergency medicine	7 (35)
Internal medicine/hospitalist	1 (5)
Internal medicine/critical care medicine	3 (15)
Pulmonary/critical care medicine	4 (20)
Surgical critical care	2 (10)
Other	0 (0)
Years of clinical experience, n (%)	
<5 years	0 (0)
5 - 9 years	6 (30)
10 - 19 years	7 (35)
≥20 years	7 (35)
Primary practice setting, n (%)	
Academic tertiary care hospital	9 (45)
Community hospital	10 (50)
Other	1 (5)
Best estimate of how many times you have previously used telemedicine to care for or consult on a patient in any setting, n (%)	
Never	8 (40)
<5 times	2 (10)
6 - 10 times	1 (5)
>10 times	9 (45)
Best estimate of how many times you have previously used telemedicine to care for or consult on a patient in a disaster setting, n (%)	
Never	16 (80)
<5 times	0 (0)
6 - 10 times	2 (10)
>10 times	2 (10)

Evaluation Outcomes

In total, 50 test calls and no clinical calls were logged. While only 3 of the 10 field clinician respondents (30%) reported success on their first attempt, most (70%) successfully completed at least 1 video call during the pilot period. Among the 50 test calls, only 22% (95% CI 10% - 34%) connected successfully (Table 2). The median time to connect to an expert teleconsultant was 1.6 (IQR 3.2) minutes.

XSL•FO
RenderX

Table . Call frequency and performance across all field sites (Hospital 1, Hospital 2, Hospital 3, and Hospital 4) for all attempted calls (n=50) and the subset of calls routed through the virtual Region 1 RDHRS[a] MEOC[b] (n=23).

	Total calls attempted	Calls routed through virtual MEOC
Call frequency by site, n (%)		
Hospital 1	28 (56)	5 (22)
Hospital 2	12 (24)	8 (35)
Hospital 3	5 (10)	5 (22)
Hospital 4	5 (10)	5 (22)
Patient registered, n (%)	50 (100)	23 (100)
Consult request initiated, n (%)	50 (100)	23 (100)
Call routed to RDHRS, n (%)	23 (46)	23 (100)
Expert notified, n (%)	23 (46)	21 (91)
Expert responded, n (%)	21 (42)	19 (83)
Connection established, n (%)	15 (30)	15 (65)
Video call successful, n (%)	11 (22)	11 (48)
Video problem during call, n (%)	5 (10)	5 (22)
Audio problem during call, n (%)	7 (14)	7 (26)

[a]RDHRS: Regional Disaster Health Response System.

[b]MEOC: medical emergency operations center.

Table 3 details the reasons for call failure (n=39) recorded by the Region 1 RDHRS platform. Half (49%) of the failed calls were due to a call routing error in Hospital 1, and 14 (36%) failed for >1 reason. Among 23 test calls routed correctly to the MEOC, 48% (95% CI 27% - 69%) were successful video call connections (Table 2).

Table . Reasons for call failure (N=39) recorded by the platform. Counts exceed 100% as calls can fail for multiple reasons.

Reasons for call failure	Reasons, n (%)
User canceled call	12 (31)
Orphaned call	8 (21)
Call routing error	19 (49)
Expert notification error	2 (5)
No connection logged	4 (10)
Device problem	4 (10)
Connection problem	3 (8)

Among field clinician respondents (n=10), 50% used smartphone devices, 40% used hospital desktop computers, and 10% used a laptop computer to access RDHRS teleconsultation services. All hospital desktop computer users failed to connect their first call. Lack of camera or microphone capability (30%) and lack of video (40%) or audio (20%) during the call were common reasons for call failure reported by field clinicians. Fewer reported difficulties logging in to the platform (20%), adding a patient (10%), or initiating a video call (20%).

The mean TUS was 5.6 (SD 1.3; Table 4). TUQ item scores were highest in usefulness (mean 6.0; SD 1.1) and ease of use (mean 6.0, SD 1.4) and lowest in reliability (mean 2.4, SD 1.4).

XSL•FO
RenderX

Table . Telehealth Usability Questionnaire (TUQ) item mean scores with SD and median scores with IQR. Participants (N=20) rated agreement with each item on a 7-point Likert scale (1=disagree, 7=agree). The total usability score (TUS) is the average for all items (maximum score=7).

Domain and item	Participants, n	Mean (SD)	Median	Quartile	
				Lower	Upper
Usefulness					
Telemedicine improves my ability to provide or receive remote medical consultation during a disaster.	20	6.0 (1.1)	6.0	5.0	7.0
Telemedicine saves me time when communicating with another provider.	16	5.0 (1.8)	5.5	4.0	6.5
Telemedicine provides for my needs as a health care provider.	16	5.3 (1.4)	5.5	4.0	6.5
Ease of use and learnability					
It is simple to use this system.	18	5.5 (1.5)	6.0	5.0	7.0
It is easy to learn to use the system.	18	6.0 (1.4)	6.5	6.0	7.0
I believe I could become productive quickly using this telemedicine system.	18	5.8 (1.3)	6.0	5.0	7.0
Interface quality					
The way I interact with this system is agreeable.	16	5.6 (1.3)	6.0	5.0	6.5
I like using this telemedicine system.	17	5.3 (1.4)	5.0	5.0	6.0
The telemedicine system is simple and easy to understand.	19	5.7 (1.6)	6.0	4.0	7.0
This telemedicine system is able to do everything I would want it to be able to do.	15	4.4 (2.1)	5.0	2.0	6.0
Interaction quality					
I can easily talk to the clinician using this telemedicine system.	15	5.4 (1.6)	6.0	5.0	7.0
I can hear the clinician clearly using this telemedicine system.	15	5.5 (1.6)	6.0	5.0	7.0
Using this telemedicine system, I am able to express myself effectively.	16	5.8 (1.6)	6.0	5.5	7.0
Using this telemedicine system, I can see the clinician as well as if we met in person.	15	5.3 (1.6)	6.0	4.0	6.0
Reliability					

XSL•FO
RenderX

Domain and item	Participants, n	Mean (SD)	Median	Quartile	
				Lower	Upper
I think the consultations I give or receive over this telemedicine system are the same as hospital-based consultations.	17	4.2 (1.8)	4.0	3.0	6.0
Whenever I make a mistake using this system, I can recover easily and quickly.	11	4.2 (2.1)	4.0	2.0	6.0
This telemedicine system gives error messages that clearly told me how to fix problems.	19	2.4 (1.4)	2.0	1.0	3.0
Satisfaction and future use					
I feel comfortable communicating with the clinician using this telemedicine system.	15	5.6 (1.6)	6.0	5.0	7.0
This telemedicine system is an acceptable way to give or receive online consultation.	16	5.9 (1.5)	6.0	6.0	7.0
I would use this disaster telemedicine service again.	16	6.0 (1.5)	6.0	6.0	7.0
Overall, I am satisfied with this telemedicine system.	16	5.4 (1.8)	6.0	4.5	7.0
Total usability score	20	5.6 (1.3)	5.9	4.9	6.5

All participants felt comfortable using the RDHRS platform to access disaster teleconsultation services. Most (80%) field clinicians preferred to use personal or institutional smartphones or tablet devices. All SMEs were willing to participate in a national expert registry, including on short notice. All were willing to obtain out-of-state emergency licensure, and 8 in 10 were willing to serve on a volunteer, unpaid basis. Almost all (90%) SMEs were willing to provide teleconsultation services from a personal device.

In total, 6 field clinicians and 6 SMEs completed interviews. Table 5 summarizes key themes with representative quotes. Most physicians perceived the platform as simple and easy to learn, and those who had difficulty learning the system identified discomfort with technology as the primary cause. When video was limited, providers could still communicate via an audio call or real-time chat function. Most field clinicians agreed that simplifying the 2-step process of connecting to SMEs would improve system workflow. Almost all SMEs desired more platform training to support real-time technical assistance for field clinicians during the call.

XSL•FO
RenderX

Table . Key themes and representative quotes from physicians regarding perceived usefulness and ease-of-use, attitudes, behavioral intention to use, and future recommendations for using the prototype model system for disaster teleconsultation services.

Theme	Representative quotes from physicians
Usefulness	
Access to teleconsultants can enhance bedside physician capabilities	• "I think it's especially useful when there are novel issues for which providers are not familiar with." • "I think having that ability to have this type of a resource would be invaluable in that situation." • "I would clearly just use it for um, for uh, for a disease or a skill that I don't have." • "Sort of the value proposition here is completely access to a bunch of specialty services that you don't have at a small hospital or wherever that I will be in the middle of nowhere. Um, just that safety net feels great."
I thef workflow is not efficien,t it may be underused	• "…the login for sure. I mean, especially in a disaster setting, I think people really have a very little patience to try to figure out how to get into the system in a short amount of time" • "The greater, the disaster situation, the sort of greater time crunch and acuity and volume, you have, you may feel less inclined to, um, to stop it, utilize this"
Internal Context	
Individual provider comfort level with telemedicine may limit usability	• "For me everything technical is a challenge." • "Not everyone is as familiar with smartphones or applications or even computers" • "I think it's gonna also come down to like my, my degree of comfort or familiarity with like what it is that I'm treating in the first place." • "I'll say that a lot of our older attendings who probably will be using this to when a disaster, right, um, are less comfortable with it."
Access to proper devices will be essential	• "I had no audio capabilities on my computer at the hospital and my phone wasn't working." • "I use my desktop, which does not have a microphone or video capabilities. I was able to log in, um, which I could from my phone for some reason."
Site-specific limitations mus bet addressed in advance	• "I think that I had roadblocks within the hospital for various reasons." • "…it's a running joke here about our WiFi" • "There were a large number of your calls that were going to, that were routing to the wrong call center, essentially." • "I imagine it was the firewall that they gave me trouble with, um, the audio and video on the first call when I was at MGH. The one that I did that was successful was at home."
Ease of use	
System was simpl;e howeve,r logging in to the system proved difficult in practice	• "It was pretty self-explanatory to me once I got into the system." • "…there weren't, you know, there weren't too many options or buttons, which is a good thing" • "You know, from the initial training, it seems like it should have gone very easily and it didn't seem to be quite as intuitive" • "The login password would be the most difficult part to manage." • "The technical parts of trying to log in and having separate passwords and having a different user login, anything to simplify that would be amazing." • "Whatever that extra button is for the request to make it, to alert me. Uh, I found that challenging"

XSL•FO
RenderX

Theme	Representative quotes from physicians
The interface quality was strong but needs improvements	• "The quality of the video and the audio is really good even on my phone." • "Overall the interface is pretty user friendly with the very few buttons and the buttons on the side were pretty self-explanatory" • "…trying to have a conversation when every three or four seconds the screens are freezing, you're just going to say, 'Uh, I don't have time for this." • "I had challenges with the video function again, not knowing um, exactly what, uh, how to activate it from the other end."
Workflow improvements will assist just-in-time training for future disasters	• "I did have to read through the emails and the sheets very carefully to make sure I was doing everything right. Um, so that was a little bit of a barrier." • "…would take some practice ahead of time because again, I mean, for me to just jump in and start doing screen-sharing, I'm not sure how easy that would be." • "…because it's all web based, it was seamless and, you know, very easy from that regard." • "If I didn't have the PDF, I'm not sure. I haven't tried it without following the instructions you sent"
Attitudes and behavioral intention to use	
Institutional support and liability protections could incentivize use	• "…as long as the institutional support and, and liability reasons are there, you know, supported, then I think it'd be really easy to use"
Access and awareness of system availability could influence use	• "There's nothing that would make me not want to use the system, other than the inability to access it or forgetting that I had it."
Recommendations for future use	
Workflow improvements to improve simplicity	• "Something as simple as the password reset" • "There's that one extra step to make it an active consult." • "…whatever that extra button is for the request to make it, to alert me. Uh, I found that challenging" • "There's the completed consult and then request a consult and then like active consult. Um, it was sitting in sort of that holding middle." • "…so the less steps the better, or at least the more guided steps, the better as well, too."
Expert teleconsultant's role needs clarification	• "Not having a clear line of communication in terms of what's being expected of the consultation." • "So potentially just more expertise on the side of, of us. Um, as the, as the supposed experts on implementing the system."
Institutional support and "super users" at each site could be invaluable	• "The challenge in getting something like this up in, you know, a couple of days without super users at the various sites." • "I think overall institutional support, societal support, medical, legal, liability…" • "Eventually the legal or public with legal ramifications in public view." • "…probably honestly a champion in every institution who can bring this about and to let the not only their staff know, but make sure it's a discussed in the faculty meetings"

Overall, the biggest barrier to successful connections was issues related to the lack of audiovisual capabilities on hospital computers, access to reliable and consistent mobile or wireless internet connections in clinical workspaces, and circumventing institutional firewall restrictions. Despite technical limitations, most expressed positive attitudes and future intent to use the platform. Recommendations for future use included clarifying the role and expectations of teleconsultants, streamlining the workflow for field clinicians to minimize risk of call failures, and improving the SME's knowledge of the system to facilitate real-time communication with field clinicians.

Discussion

Principal Results

This was a field test of the acceptability and feasibility of implementing a prototype RDHRS disaster teleconsultation system to deliver tele-critical care services to Massachusetts

XSL•FO
RenderX

hospitals during the first wave of the COVID-19 pandemic. Clinicians found the prototype platform acceptable, but the workflow requires revision to reduce call failure and improve feasibility and reliability for future use with minimal-to-no training. We expanded access across hospital systems in real-world conditions and deployed SMEs from a national volunteer pool using emergency waivers and state emergency licensure processes, although this delayed implementation and impacted clinical use.

We were able to improve our initial prototype model in important ways [7]. We avoided software or hardware needs by deploying a web-based application. Using typical emergency department workflows and automating triage and expert notification processes that were previously done manually improved call center efficiency by reducing call wait times. However, these metrics are based on small samples under low volume conditions and did not involve patient care. Future evaluations should include broader system performance indicators, such as throughput, consult duration and quality, and user experience, ideally measured in real-world conditions.

Our workflow analysis identified potential sources of call failure at platform, user, and hospital levels that require technology and system solutions to improve call center performance and reliability. Initially, overall call connection success through the MEOC was low because the technology vendor accidentally introduced a call routing error while registering one hospital. Call failures due to expert notification errors, use of hospital computers lacking cameras and microphones, and institutional firewalls were less common. Though many participants could complete a call, the major sources of call failure—orphaned and canceled calls, and call routing errors—occurred early during steps where the consult was requested and routed to the MEOC. In the initial configuration (Figure 1A), the field clinician initiated the video call connection, which required them to remain at the location of the device used to initiate the call until the SME answered. Thus, call routing errors and delayed expert response produced orphaned and canceled calls. Field clinicians also reported confusion with the 2-step process, often mistaking patient registration with consult request submission.

We have since streamlined the workflow to eliminate these errors (Figure 1B). Field clinicians (or support staff) now submit consult requests in a single step and receive confirmation. The consult request is automatically routed to the MEOC, which alerts the SME to a pending consult. The SME then initiates the video call connection by sending a text link and email invitation to the field clinician's chosen device. The video call connection is established when the field clinician accepts either invitation. This frees the field clinician to remain active in care and may reduce orphaned or canceled calls. These changes improved call connection success in a subsequent large-scale functional exercise conducted by the RDHRS, but these data are preliminary and not yet externally validated. Future third-party assessment is needed to validate system performance.

Each hospital was onboarded individually to institutional cybersecurity standards in this pilot field test. We found this process may hinder scalability and was vulnerable to error. Registering hospitals in advance could mitigate this but requires central coordination and administrative support. Another option is to use a central web portal with an event-specific code that can be distributed via existing emergency notification systems for just-in-time access. This feature could enable an RDHRS to control system activation and deactivation centrally, make services accessible immediately but restricted to affected facilities, and scale system use to demand.

XSL•FO
RenderX

Figure 1. Summary of the prototype workflow. We used tele-emergency workflows to register and room patients. This involved a 2-step process where the field clinician first added one or more patients and then requested consultation. Triage and expert notification processes were automated within a virtual medical emergency operations center where consults were prioritized by illness severity and timing. The prototype R1/RDHRS (Regional Disaster Health Response System) teleconsultation system: (A) initial workflow and (B) revised workflow.

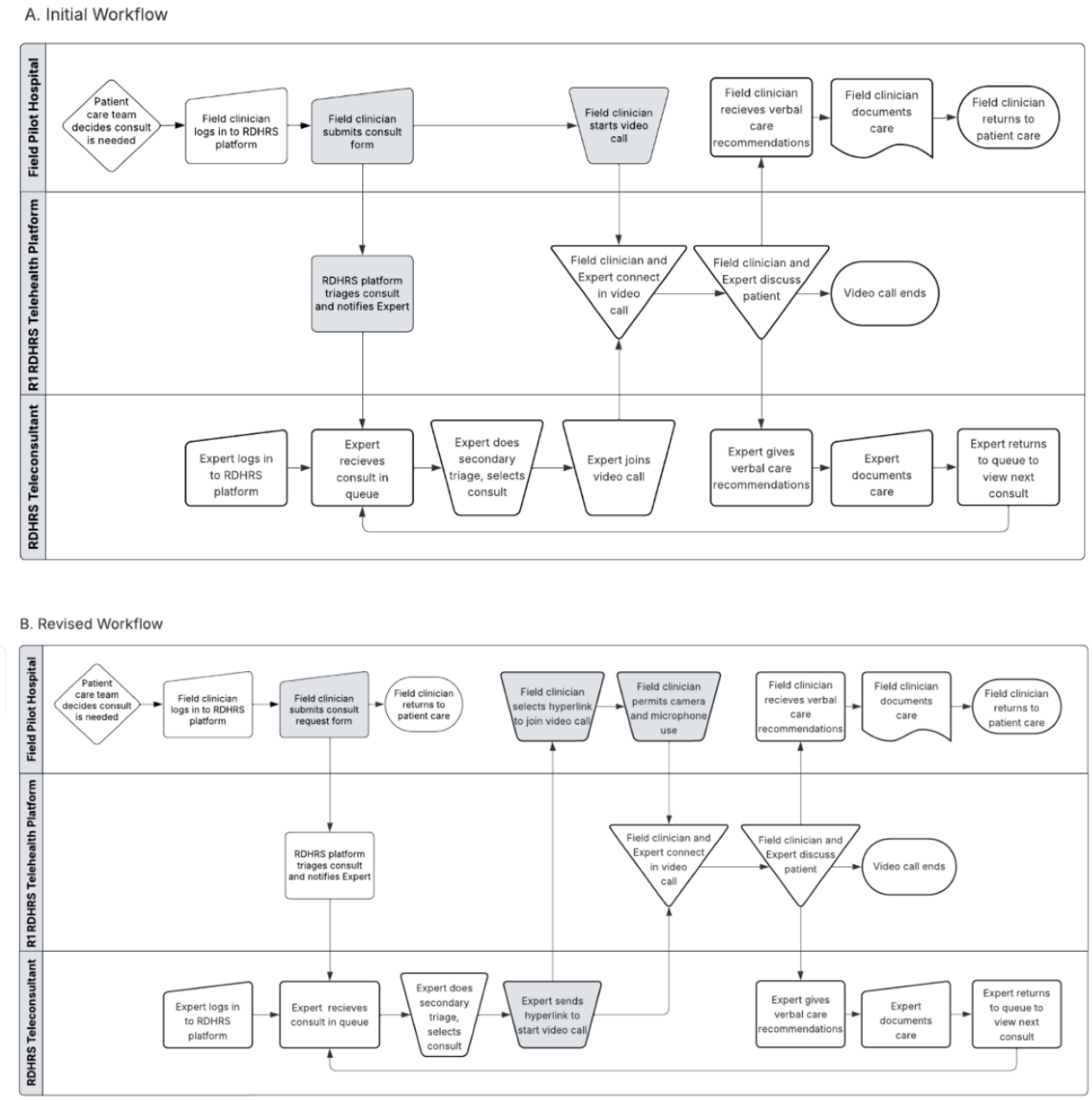

Comparison With Prior Work

Our findings support prior studies highlighting the importance of access to suitable devices [15,16]. While the platform was device-agnostic, hospital computers without cameras or microphones accounted for many failed calls. Every participant who used a desktop computer failed their first call, and device problems constituted a third of call failures. Using designated institutional mobile devices, tablets, or personal devices could improve call reliability. A 2021 systematic review identified lack of workflow integration, outdated technology infrastructure, and user acceptance among the most commonly cited barriers for telehealth services implemented globally during COVID [15]. Postpandemic institutional technology upgrades may benefit the future implementation of disaster teleconsultation systems [17]. We recommend hospital emergency operation plans include devices with audio-visual capabilities and telecommunication network redundancy to ensure seamless access to disaster teleconsultation services.

Our experience also highlights how technology, workforce, and operational system readiness are essential to harness digital health tools for disaster response. We accelerated prototype system development and tested implementation in March 2020 as the need to expand critical care access emerged. Working with an industry partner, we configured a prototype platform for clinical use within 2 weeks. Despite federal and state

XSL•FO
RenderX

emergency laws and waivers [10,11], it took an additional 6 weeks to navigate out-of-state physician licensure, hospital credentialing, and liability coverage for teleconsultants. By the time pilot implementation began in June 2020, COVID-19 cases had declined in Massachusetts, so only simulated calls were performed.

This study mirrors prior research demonstrating how variation in licensing and credentialing procedures can delay service provision, despite emergency declarations [6,18,19]. The United States lacks a national medical license. Instead, state medical boards grant physician licenses to practice medicine in that state. Hospitals and healthcare organizations use credentialing procedures to verify qualifications before permitting physicians to provide care within their institutions. The Interstate Medical Licensure Compact is 1 legal pathway to expedite licensure [20]. However, some states are not members, and full licensure is still required in each state, so the financial and administrative burden for physician volunteers would be prohibitive. Model legislation, like the Emergency Volunteer Health Practitioners Act [21], is an alternate pathway to allow license reciprocity, credentialing waivers, and liability protections during declared emergencies, but only 20 states have enacted such laws. Without adoption of national solutions, RDHRS and similar response entities will need to rely on state emergency waivers and limit scope of practice to advisory only when providing access to disaster teleconsultation services across jurisdictions.

We and others have argued that disaster telehealth uptake requires proactive rather than reactive system development [18,22]. Disasters requiring specialty care are often no-notice events, so disaster teleconsultation systems need simple, reliable, field-tested platforms and preverified, rostered clinicians to support virtual deployment within hours rather than days to weeks. A range of specialists may need to be onboarded quickly because service needs are dictated by the medical consequences of the disaster. Therefore, these systems need to be flexible to support various use cases and easy to use regardless of discipline, training, or exposure. Emulating key features of the workflow and infrastructure of established emergency call centers and providing a clear scope of practice for disaster teleconsultants could encourage adoption.

Participants referenced US Poison Centers [23] as an effective model for delivering on-demand emergency consultations. Field clinicians felt comfortable consulting Poison Center services to access medical experts by telephone. However, once the expert could view the patient via video, providers were unclear whether the interactions constituted the broader practice of medicine (telemedicine) versus offering recommendations (teleconsultation) at a distance.

While teleconsultation is a type of telemedicine encounter, there are important distinctions in provider responsibilities, licensure, liability, and patient privacy [24]. Providers must adhere to standards of in-person practice, but responsibility and liability depend on whether the teleconsultant is directly involved in care or offering advice. Many states provide a “consultation exception” to licensure, but the scope and details of these exceptions vary by state [25]. Clinicians need to verify their malpractice coverage extends to teleconsultation and includes relevant jurisdictions. They should also include the limitations of remote assessment in any medical documentation.

While HIPAA provisions were temporarily relaxed during the COVID-19 public health emergency, using off-the-shelf platforms without encryption introduces the risk of privacy breaches. Disaster telehealth platforms should use robust technical and administrative safeguards aligned with best practices in health information security, including end-to-end data encryption, access control mechanisms, audit trails, and data anonymization techniques [26,27]. Strict adherence to HIPAA and privacy laws and careful attention to cyber security risks are essential to maintain patient and provider trust and legal compliance, as consultations may occur in non-secure environments [28].

Finally, clinicians in our study highlighted how administrative and legislative buy-in were necessary for real-world implementation, a finding supported by pediatric disaster groups attempting statewide implementation of specialty telehealth services [29]. Establishing trust and acceptance of disaster teleconsultation systems by the medical community and the public will require socialization and integration of the RDHRS program within the national strategic health response.

Limitations

This study has limitations. First, the small, regionally restricted sample may limit generalizability. Still, we captured important perspectives from community and tertiary care providers in real-world conditions to guide future system refinements. Second, lack of opportunity to provide clinical care limited any exploration of how disaster teleconsultation services could impact patient outcomes, so more research is needed. Third, we focused on pilot implementation in Massachusetts, as this was the initial location of clinical need, so our analysis of barriers and facilitators may not apply to states or regions where laws or regulations differ. For example, states that allow license reciprocity or participate in regional compacts could streamline access to disaster teleconsultation services. This could benefit health professional shortage areas and rural communities where disasters are likely to exacerbate underlying disparities in healthcare access. To support future pilots in different state contexts, RDHRS and other regional disaster programs should establish processes to authorize disaster telehealth service delivery in coordination with local and state regulatory bodies and relevant government agencies. Last, hospital and clinician willingness to use regional disaster teleconsultation services for pandemics may differ from no-notice events where care delivery is time-sensitive and specialist supply is even more limited (eg, burn surgeons).

Conclusions

Implementing RDHRS disaster teleconsultation services to connect field clinicians to specialists across state lines and hospital systems was feasible in real-world conditions. We improved system efficiency using a simple, device-agnostic, web-based platform to create a virtual MEOC that automatically routes, triages, and notifies experts of requested consultations. While user attitudes were generally positive, further technology and system refinements are needed to align call performance,

XSL•FO
RenderX

reliability, and usability to desired standards. Although emergency waivers enabled the RDHRS to mobilize and deploy a virtual volunteer workforce for short-term disaster response, liability concerns and variable hospital credentialing practices delayed implementation. We encourage government and emergency management agencies to advance policies that expand liability protections for volunteer teleconsultants and incentivize hospitals to use standard procedures to support timely, effective regional disaster telehealth solutions.

Acknowledgments

We are grateful for the contributions of Rachel Bedenbaugh from the Regional Disaster Health Response System (RDHRS), Gerard Dubois from the Northeast Telehealth Resource Center, and the Society for Critical Care Medicine. This research was funded by the Administration for Strategic Preparedness and Response (ASPR; grant #HITEP180042-01-00). TPB is funded by the National Institutes of Health (NIH; grant 1K23HL145126-01A1). The contents herein are solely the authors' responsibility and do not necessarily represent the official views of ASPR or NIH. The authors attest that there was no use of generative artificial intelligence technology in the generation of text, figures, or other informational content of this manuscript.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

References

1. Administration for Strategic Preparedness and Response. Regional disaster health response system: building regional solutions. Secondary Regional Disaster Health Response System: Building Regional Solutions. 2021. URL: https://www.phe.gov/Preparedness/planning/RDHRS/Pages/default.aspx [accessed 2025-08-14]
2. Der-Martirosian C, Griffin AR, Chu K, Dobalian A. Telehealth at the US Department of Veterans Affairs after Hurricane Sandy. J Telemed Telecare 2019 Jun;25(5):310-317. [doi: 10.1177/1357633X17751005] [Medline: 29384428]
3. Grover JM, Smith B, Williams JG, Patel MD, Cabanas JG, Brice JH. Novel use of telemedicine by hurricane evacuation shelters. Prehosp Emerg Care 2020;24(6):804-812. [doi: 10.1080/10903127.2020.1723756] [Medline: 32011202]
4. Koonin LM, Hoots B, Tsang CA, et al. Trends in the use of telehealth during the emergence of the COVID-19 Pandemic - United States, January-March 2020. MMWR Morb Mortal Wkly Rep 2020 Oct 30;69(43):1595-1599. [doi: 10.15585/mmwr.mm6943a3] [Medline: 33119561]
5. Friedman RSC, Carpenter DM, Shaver JM, McDermott SC, Voelkel J. Telemedicine familiarity and post-disaster utilization of emergency and hospital services for ambulatory care sensitive conditions. Am J Prev Med 2022 Jul;63(1):e1-e9. [doi: 10.1016/j.amepre.2021.12.025] [Medline: 35300889]
6. Curfman A, Hackell JM, Herendeen NE, et al. Telehealth: opportunities to improve access, quality, and cost in pediatric care. Pediatrics 2022 Mar 1;149(3):e2021056035. [doi: 10.1542/peds.2021-056035] [Medline: 35224638]
7. Boyle TP, Ludy S, Meguerdichian D, et al. Feasibility and acceptability of a model disaster teleconsultation system for regional disaster health response. Telemed J E Health 2023 Apr;29(4):625-632. [doi: 10.1089/tmj.2022.0103] [Medline: 36036805]
8. Kadri SS, Sun J, Lawandi A, et al. Association between caseload surge and COVID-19 survival in 558 U.S. hospitals, March to August 2020. Ann Intern Med 2021 Sep;174(9):1240-1251. [doi: 10.7326/M21-1213]
9. Hong YR, Lawrence J, Williams D Jr, Mainous III A. Population-level interest and telehealth capacity of US hospitals in response to COVID-19: cross-sectional analysis of Google Search and National Hospital Survey Data. JMIR Public Health Surveill 2020 Apr 7;6(2):e18961. [doi: 10.2196/18961] [Medline: 32250963]
10. Coronavirus aid, relief, and economic security (CARES) act: s 3548 - 116th congress. : GovTrack.us; 2020.
11. Keesara S, Jonas A, Schulman K. Covid-19 and health care's digital revolution. N Engl J Med 2020 Jun 4;382(23):e82. [doi: 10.1056/NEJMp2005835] [Medline: 32240581]
12. Parmanto B, Lewis AN, Graham KM, Bertolet MH. Development of the Telehealth Usability Questionnaire (TUQ). Int J Telerehabil 2016;8(1):3-10. [doi: 10.5195/ijt.2016.6196] [Medline: 27563386]
13. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381. [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
14. Holden RJ, Karsh BT. The technology acceptance model: its past and its future in health care. J Biomed Inform 2010 Feb;43(1):159-172. [doi: 10.1016/j.jbi.2009.07.002] [Medline: 19615467]

XSL•FO
RenderX

15. Khoshrounejad F, Hamednia M, Mehrjerd A, et al. Telehealth-based services during the COVID-19 pandemic: a systematic review of features and challenges. Front Public Health 2021;9:711762. [doi: 10.3389/fpubh.2021.711762] [Medline: 34350154]
16. Phuong J, Ordóñez P, Cao J, et al. Telehealth and digital health innovations: a mixed landscape of access. PLOS Digit Health 2023 Dec;2(12):e0000401. [doi: 10.1371/journal.pdig.0000401] [Medline: 38100519]
17. Bains J, Greenwald PW, Mulcare MR, et al. Utilizing telemedicine in a novel approach to COVID-19 management and patient experience in the emergency department. Telemed J E Health 2021 Mar;27(3):254-260. [doi: 10.1089/tmj.2020.0162] [Medline: 32821027]
18. Boyle TP, Mehra A, Smith NJ, et al. The first statewide implementation of a regional disaster teleconsultation system to expand critical care surge capacity: a case study in Vermont. Telemed J E Health 2024 May;30(5):1495-1498. [doi: 10.1089/tmj.2023.0339] [Medline: 38039352]
19. Pamplin JC, Gray B, Quinn MT, et al. National emergency tele-critical care in a pandemic: barriers and solutions. Crit Care Explor 2024 May;6(5):e1091. [doi: 10.1097/CCE.0000000000001091] [Medline: 38746740]
20. The IMLC Commission. Interstate medical licensure compact. Secondary Interstate Medical Licensure Compact. URL: https://www.imlcc.org [accessed 2025-08-14]
21. Emergency volunteer health practitioners act. : National Conference of Commissioners on Uniform State Laws; 2007.
22. Smith AC, Thomas E, Snoswell CL, et al. Telehealth for global emergencies: Implications for coronavirus disease 2019 (COVID-19). J Telemed Telecare 2020 Jun;26(5):309-313. [doi: 10.1177/1357633X20916567] [Medline: 32196391]
23. Poison centers secondary poison centers. Health resources and services administration. URL: https://poisonhelp.hrsa.gov/poison-centers [accessed 2025-08-14]
24. Nittari G, Khuman R, Baldoni S, et al. Telemedicine practice: review of the current ethical and legal challenges. Telemed J E Health 2020 Dec;26(12):1427-1437. [doi: 10.1089/tmj.2019.0158] [Medline: 32049608]
25. Schwamm LH, Chumbler N, Brown E, et al. Recommendations for the implementation of telehealth in cardiovascular and stroke care: a policy statement from the American Heart Association. Circulation 2017 Feb 14;135(7):e24-e44. [doi: 10.1161/CIR.0000000000000475] [Medline: 27998940]
26. Tukur M, Saad G, AlShagathrh FM, Househ M, Agus M. Telehealth interventions during COVID-19 pandemic: a scoping review of applications, challenges, privacy and security issues. BMJ Health Care Inform 2023 Aug;30(1):e100676. [doi: 10.1136/bmjhci-2022-100676] [Medline: 37541739]
27. He Y, Aliyu A, Evans M, Luo C. Health care cybersecurity challenges and solutions under the climate of COVID-19: scoping review. J Med Internet Res 2021 Apr 20;23(4):e21747. [doi: 10.2196/21747] [Medline: 33764885]
28. Lieneck C, McLauchlan M, Phillips S. Healthcare cybersecurity ethical concerns during the COVID-19 global pandemic: a rapid review. Healthcare (Basel) 2023 Nov 18;11(22):2983. [doi: 10.3390/healthcare11222983] [Medline: 37998475]
29. Marcus BS, Newton C, Rafton SA, et al. Regional pediatric disaster network guides statewide telehealth initiation during COVID-19 pandemic. Am J Disaster Med 2022;17(2):163-169. [doi: 10.5055/ajdm.2022.0430] [Medline: 36494886]

Abbreviations

HIPAA: Health Insurance Portability and Accountability Act
MEOC: medical emergency operations center
RDHRS: Regional Disaster Health Response System
REDCap: Research Electronic Data Capture
SME: subject matter expert
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
TUQ: Telehealth Usability Questionnaire
TUS: total usability score

Edited by A Mavragani; submitted 24.02.25; peer-reviewed by J Marcin, N Hu; revised version received 09.06.25; accepted 11.06.25; published 19.08.25.

Please cite as:
Ludy S, Drainoni ML, Schmidt L, Litvak M, Dugas J, Goralnick E, Biddinger PD, Boyle TP
Acceptability and Feasibility of a Prototype Regional Disaster Teleconsultation System for COVID-19 Pandemic Response: Pilot Field Test
JMIR Form Res 2025;9:e73078
URL: https://formative.jmir.org/2025/1/e73078
doi:10.2196/73078

XSL•FO
RenderX

© Stephanie Ludy, Mari-Lynn Drainoni, Lauren Schmidt, Mark Litvak, Julianne Dugas, Eric Goralnick, Paul Biddinger, Tehnaz Parakh Boyle. Originally published in JMIR Formative Research (https://formative.jmir.org), 19.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Piloting the Extension for Community Healthcare Outcomes (ECHO) Pediatric Oncology Telehealth Education Program in Western Kenya: Implementation Study

Tyler Severance[1,2*], MD, MEd; Gilbert Olbara[3*], MD; Festus Njuguna[4], MD, PhD; Martha Kipng'etich[5], MSc; Sandra Lang'at[5], MPH; Maureen Kugo[5]; Jesse Lemmen[6,7], MD; Marjorie Treff[8], EdD; Patrick Loehrer[9], MD; Terry Vik[2], MD

[1]Department of Pediatrics, University of Missouri, 400 N Keene St, Suit 111, Columbia, MO, United States
[2]Department of Pediatrics, Indiana University, Indianapolis, IN, United States
[3]Moi Teaching and Referral Hospital, Eldoret, Kenya
[4]Moi University, Eldoret, Kenya
[5]Academic Model Providing Access to Healthcare, Eldoret, Kenya
[6]Department of Pediatric Oncology, Emma Children's Hospital, Amsterdam UMC, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
[7]Princess Máxima Center for Pediatric Oncology, Utrecht, The Netherlands
[8]School of Education, Indiana University, Bloomington, IN, United States
[9]School of Medicine, Indiana University, Indianapolis, IN, United States
[*]these authors contributed equally

Corresponding Author:
Tyler Severance, MD, MEd
Department of Pediatrics, University of Missouri, 400 N Keene St, Suit 111, Columbia, MO, United States

Abstract

Background: Childhood cancer has an annual incidence of 150 - 160 cases per million children worldwide but remains vastly underdiagnosed in low- to middle-income countries such as those in Sub-Saharan Africa. The Moi Teaching and Referral Hospital (MTRH) serves a population of 25 million people, including 10 million children. The average number of pediatric cancer diagnoses was 216 cases annually in 2017 - 2019, which was well below the anticipated 1500 cases based on epidemiology data. The remaining 75% - 80% of pediatric cancer cases remain undiagnosed, and these patients are not likely to survive. Prior outreach and needs assessments demonstrated a lack of medical knowledge related to pediatric cancer as a primary barrier to improved referrals, diagnoses, and ultimately, cure.

Objective: This study aimed to address disparities in medical knowledge contributing to low diagnostic rates of cancer in children. We implemented Project ECHO (Extension for Community Healthcare Outcomes)—a validated virtual guided practice and telementoring model—to connect multidisciplinary specialists at MTRH with staff in medically underserved communities in western Kenya for training, technical assistance, and mentorship.

Methods: Sessions were freely available on Zoom twice monthly and featured an expert-led didactic topic followed by a learner-led, case-based discussion. The discussion used dialogue education to promote learning and engagement among participants, with mentorship from the expert team. Information on ECHO participation was tracked, and electronic surveys were sent to the participants at the end of the pilot year. The ECHO program was run in parallel with the pediatric oncology cancer registry to monitor trends in diagnostic rates within the referral region.

Results: The ECHO program launched successfully in January 2020 with a curriculum focused on pediatric oncology for health care providers. A total of 22 sessions were conducted, with an average of 23 learners per session. A total of 148 participants attended at least one session, with the majority (n=80, 54.1%) attending multiple sessions. The year-end analysis in January 2021 demonstrated that 286 new pediatric patients were diagnosed with cancer at MTRH, representing a 33% increase over the 3-year average.

Conclusions: The Project ECHO platform created a dynamic virtual platform to continue to engage stakeholders across western Kenya. The implementation of this telehealth education platform in Kenya represents an effective model for increasing the recognition and earlier referral of childhood cancer in low- to middle-income countries.

(JMIR Form Res 2025;9:e59776) doi:10.2196/59776

XSL•FO
RenderX

KEYWORDS

pediatric cancer; global health; education; Extension for Community Healthcare Outcomes; project ECHO; pediatric oncology extension; pediatric; oncology; health care; outcome; telehealth; children; cancer; diagnosis; epidemiology; telementoring; technical assistance; virtual platform; effectiveness; equitable access; early diagnosis; collaboration; training; medical education; remote education; online education; low-middle income country; LMIC; community health care outcomes

Introduction

Pediatric Oncology in Kenya

Over the past decade, efforts have been made to identify health care disparities affecting children in Sub-Saharan Africa [1]. It is recognized that pediatric cancer rates are far below what would be anticipated based on treatment records at large referral centers [2]. The western regions of Kenya are home to approximately 24 million citizens, which corresponds to 10 million children under the age of 15 [3]. Based on epidemiological data from similar patient populations in the United States, approximately 1000 - 1200 patients were diagnosed with pediatric cancer each year [4]. As the only national multidisciplinary referral center in this region, Moi Teaching and Referral Hospital (MTRH) is uniquely positioned to provide multidisciplinary care for these patients, including chemotherapy, oncologic surgery, and radiotherapy. However, in the past 3 years, an average of only 216 pediatric patients were diagnosed with cancer annually [5]. This suggests that the remaining 75% - 80% of patients remain undiagnosed and likely do not survive (Figure 1). One example is pediatric leukemia, the most common type of pediatric cancer in this region. Although this malignancy should represent 300 - 480 cases (ie, 30% - 40% of all pediatric cancer diagnosis), only 40 cases per year have been seen over the last several years [5].

Figure 1. Number of cases of pediatric cancer diagnosed at Moi Teaching and Referral Hospital as a percentage of the expected total number of cases based on cancer epidemiology for the referral region. The pediatric oncology program within Academic Model Providing Access To Healthcare began in 2010, which is when diagnostic data was first available.

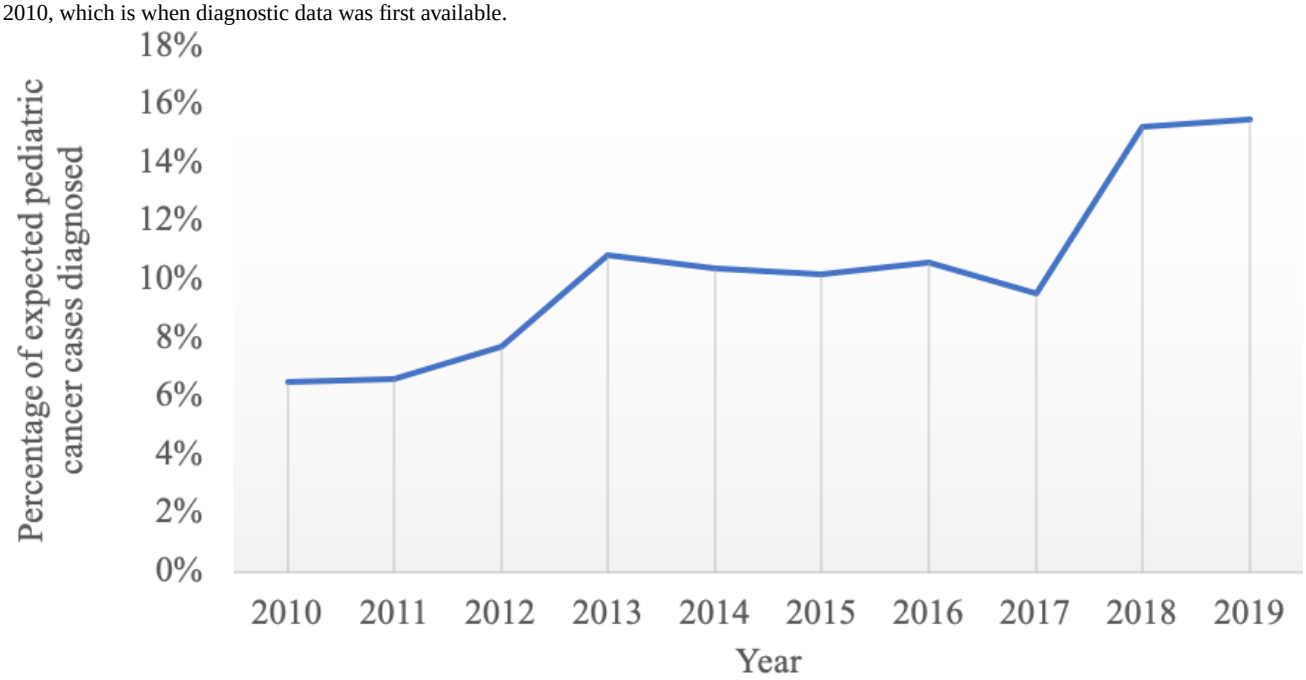

Kenyan Health care and Academic Model Providing Access To Health care (AMPATH)

To better address the disparities impacting western Kenya, an international partnership has grown to educate and empower Kenyan clinicians. This partnership began more than 30 years ago, initially included Indiana University in collaboration with Moi University College of Health Sciences and MTRH, with an emphasis on treating HIV [6,7]. Since then, AMPATH (Academic Model Providing Access To Health care) has expanded to include multiple North American institutions and a broader focus of acute and chronic diseases, including both pediatric and adult oncology [7]. Pediatric oncology training was further bolstered through partnerships with Princess Máxima Center in the Netherlands and Dutch pediatric oncology colleagues. The educational initiatives initially focused on educating clinicians at MTRH, but have since broadened to include specialty training, fellowships, regional outreach programs, and disease-specific conferences for providers in medically isolated communities [5]. The pediatric oncology program within AMPATH includes a comprehensive fellowship program, more than a decade of outreach training, and international collaboration in the United States and the Netherlands [5,8]. These efforts to improve childhood cancer care were critical in highlighting the significant gap between the actual and anticipated diagnostic rates of pediatric cancer based on population and epidemiologic information available in the referral region. Further, engaging with health care providers (HCPs) and medical leaders at the local county level offered insights into diagnostic challenges such as lack of awareness of signs and symptoms of cancer, and knowing when to refer patients for further evaluation [5].

XSL•FO
RenderX

Project ECHO (Extension for Community Healthcare Outcomes)

Project ECHO (Extension for Community Healthcare Outcomes) was developed in 2003 by Dr. S. Arora, a hepatologist at the University of New Mexico in Albuquerque [9]. He observed that patients with hepatitis C were arriving too late for treatment due to long wait times, especially in the rural, underserved, and socially disadvantaged population in New Mexico. Additionally, the isolated and overworked primary care practitioners required a new health care delivery model to remedy this problem. In response, Arora et al [9] developed a virtual hepatitis C outreach clinic using videoconferencing and didactic lectures by specialists during weekly interprofessional rounds. In this setting, remote practitioners could present their hepatitis C patients, receive guidance from the specialist team at the University of New Mexico, and subsequently treat patients themselves. Their outcomes showed an identical level of care in the rural and university clinics and led to the development of Project ECHO [9].

The ECHO programs use a telementoring approach for the education of clinicians in a hub-and-spoke model. Providers at spoke sites present cases, but no patients are ever on the network. TeleECHO clinics typically convene twice monthly for 60‑75 minutes, during which case presentations, demonstrations, and didactics are provided and recorded for both synchronous and asynchronous observation. All knowledge is shared in a “learning loop” at no cost to participants [10,11]. Continuing medical education and nursing credits are provided, and education certificates are given based on clinical specialty. The goal of each ECHO varies based on the needs of the spoke sites and the disease type. For cancer related care in a high-income country, similar cancer ECHO programs have focused on prevention, screening, and management of survivorship care [12]. However, little data exists to support the use of the ECHO program in Sub-Saharan Africa for cancer care. This paper outlines the pilot implementation of the Project ECHO model to enhance pediatric cancer care in western Kenya.

Methods

Framework of Implementation

The first step in building the pediatric oncology ECHO program was to recruit an expert hub team [13]. This team incorporated personnel from both Kenya and the United States, including American pediatric oncologists and pediatric hematology-oncology fellows from Kenya. Additional Kenyan personnel included pediatric oncology nurse navigators, as well as a dedicated coordinator and facilitator for the program. Key ECHO hub team personnel, including both Kenyan and North American program leaders, participated in virtual or in-person immersion training, which is recommended for all new programs [11,13].

The implementation of the ECHO program was consistent with the traditional model [9,14]. Each session comprised a 20‑to 25-minute didactic lecture following a curriculum based on the previously expressed needs of rural providers. A case-based discussion typically followed the traditional didactic. This required advanced preparation, as spokes were asked to submit a case for discussion at least 24 hours prior to the ECHO session. The information was summarized into a single page document consisting of pertinent information related to the patient’s presentation. All personalized or identifying information was intentionally removed from all presentation materials. During the session, the spoke presented the case and received follow-up clarifying questions from both the spokes and then the hub team. Subsequently, the spokes and hub participants then offered suggestions on how best to address the proposed clinical questions. Occasionally, when no cases were submitted, the hub team identified actively admitted patients from MTRH inpatient wards for discussion.

Recruitment and Participation

Prior to launch, spoke sites were recruited based on previously established connections in the region. Outreach attempts were made to county referral hospitals and faith-based hospitals in western Kenya. Any HCP who had recently participated in the annual AMPATH Pediatric Oncology Conference, had participated in localized training efforts, or had recently graduated from medical training programs at MTRH was invited to join the ECHO sessions. Primary recruitment efforts mainly consisted of email and phone calls. Participants were also invited to join a mobile phone messaging service (ie, WhatsApp) to allow for continued engagement and recruitment. These relationships established a core of 10‑12 clinics to participate the initial pediatric oncology ECHO sessions.

Attendance at each ECHO session was captured automatically by the parallel iECHO software. This software, which is available at no cost to all ECHO programs, tracked participation at each session and other variables, including recurrent participation and the total number of unique attendees. This information was readily available both during the pilot year and at the conclusion of the curriculum.

Evaluation

At the end of the year, a survey was sent to all participants recorded in the iECHO database. The survey used a modified Likert scale to assess multiple variables related to the program including enjoyment, comfort presenting a case, continued participation, ability to diagnose pediatric cancer, recognition of signs and symptoms of cancer, and knowledge growth. Initial survey requests were sent via electronic communication with additional reminders sent via email, mobile app, and phone calls. The results of this survey were accumulated in the Research Electronic Data Capture platform.

In the decade prior to ECHO launch, a formalized process was used to track diagnosis and referral patterns of pediatric patients with cancer. This process was continued throughout the duration of the ECHO curriculum. At the conclusion of the pilot year, the total number of new diagnoses was obtained.

Ethical Considerations

For patients requiring diagnosis and treatment, informed consent was obtained for each case, and institutional review board approval was granted for the entire duration of the project including the multiyear cancer registry (Moi Teaching and

XSL•FO
RenderX

Referral Hospital/Moi University College of Health Sciences Institutional Research and Ethics Committee approval number FAN: 0004025). Written consent from parents or guardians and assent from patients was obtained at the time of initial diagnosis in person by a member of the study team. An interpreter was used when the native language of the patient or guardian was different from that of the consenting team member. All patient information for the registry was deidentified at enrollment. Participants and their families were not financially compensated for enrollment.

For the Project ECHO telehealth sessions, this was considered an educational program, and institutional review board approval was not required for this program, as all surveys were optional and patient information was deidentified. Participants were not paid for their time or survey completion, although continuing education credits were offered to active participants in the ECHO program throughout the year.

Results

Framework of Implementation

Following the training of hub team personnel, the ECHO program launched successfully in January 2020 with a curriculum focused on pediatric oncology for general health care workers. The program was conducted continuously during the pilot year with omitted sessions only during national holidays. Feedback sessions were held at 8-week intervals throughout the year to continue to adapt to learner needs.

A curriculum was designed based on interpretations of the prior needs assessment [13]. This included an emphasis on the typical presentations of common pediatric malignancies in the region (Table 1). Of note, general review sessions were incorporated into the didactic schedule, which consisted of high-yield information from preceding sessions followed by an opportunity for spokes to provide feedback.

Table . Final didactic curriculum for the pilot year of the ECHO[a] program. The didactic schedule was designed by the hub team, with two sessions left open for topics chosen by the participants.

Session	Didactic title
1	Anatomy of a successful ECHO: yearlong goals for this program
2	Pediatric leukemia: basics of disease
3	Pediatric leukemia: history, exam, and lab findings consistent with leukemia (Part 1)
4	Pediatric leukemia: recent updates in epidemiology and outcomes
5	Pediatric lymphoma: basics of disease
6	Pediatric leukemia: next steps in evaluating (bone marrow testing, flow cytometry, tissue biopsy)
7	Transfusions in children: anemia or something else
8	Session open: didactic topic at request of spoke sites[b]
9	Pediatric lymphoma: differential diagnosis and next steps in evaluating
10	Solid tumors: bone tumors–basics and next steps
11	Pediatric leukemia: physical exam findings consistent with leukemia
12	Oncology emergencies at diagnosis
13	Brain tumors in pediatrics
14	Pediatric leukemia: lab findings consistent with leukemia
15	Abdominal masses: differential diagnosis and next steps
16	Session open: didactic topic at request of spoke sites[c]
17	Pediatric leukemia: history findings consistent with leukemia (Part 2)
18	Pediatric leukemia: standard treatment, timing, recent outcomes
19	Kidney tumors in pediatrics: basics and next steps
20	Difficulties enrolling with national health insurance fund
21	Special patient populations: adolescents and young adults
22	Retinoblastoma: what to know
23	Pediatric leukemia: maintenance therapy and survivorship basics

[a]ECHO: Extension for Community Healthcare Outcomes.

[b]Topic selected: splenomegaly–physiology and evaluation.

[c]Topic selected: Wilms tumor–diagnosis and management.

XSL•FO
RenderX

The curriculum implemented throughout the pilot year closely mirrored the proposed outline at the beginning of the year. Participants ultimately requested that lectures on the physiology and evaluation of splenomegaly and on the diagnosis and management of Wilms tumor be added to the curriculum for sessions 8 and 16, respectively. These topics were common themes relevant in preceding case discussions.

Recruitment and Participation

A total of 22 sessions were conducted during the pilot year, with an average of 23 learners per session (Figure 2). This program consisted primarily of staff at hospitals within the referral region. There were 148 total participants who attended at least one session. Among the participants, 80 (54.1%) learners returned and attended more than one session and 38 (25.7%) attended at least four sessions. In addition, during the year, a total of 31 (20.9%) patient cases were presented for discussion, with an average of 1.4 cases per session.

Figure 2. The total participants at each ECHO session during the pilot year are shown. Participation was captured electronically for all individuals logging into the Zoom platform at any time during each of the 22 sessions.

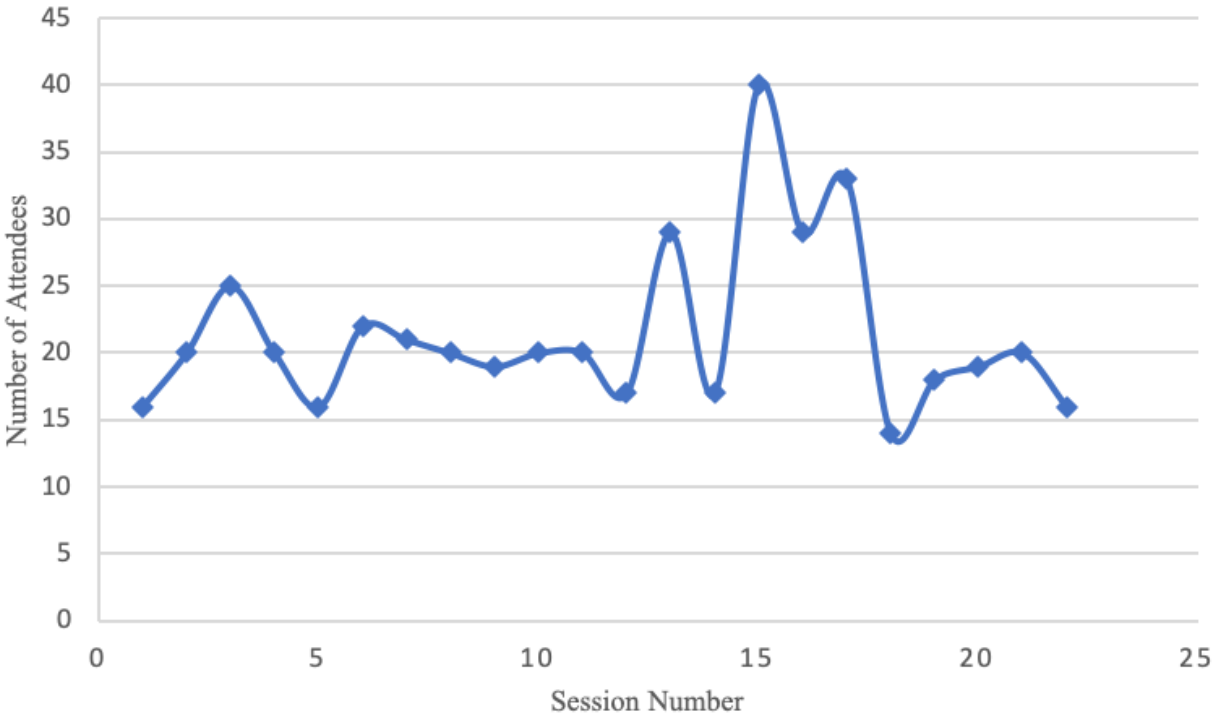

Evaluation

At the conclusion of the pilot year, all survey data was combined and analyzed. A total of 23 (15.5%) participants completed the electronic survey among the 151 total participants. Overall, the majority of responders reflected a favorable impression of the ECHO program (Figure 3).

XSL•FO
RenderX

Figure 3. Participant survey results from the pediatric oncology ECHO program following the initial pilot year are shown. Surveys were administered electronically to all participants in the program after the final session and answers were anonymously recorded.

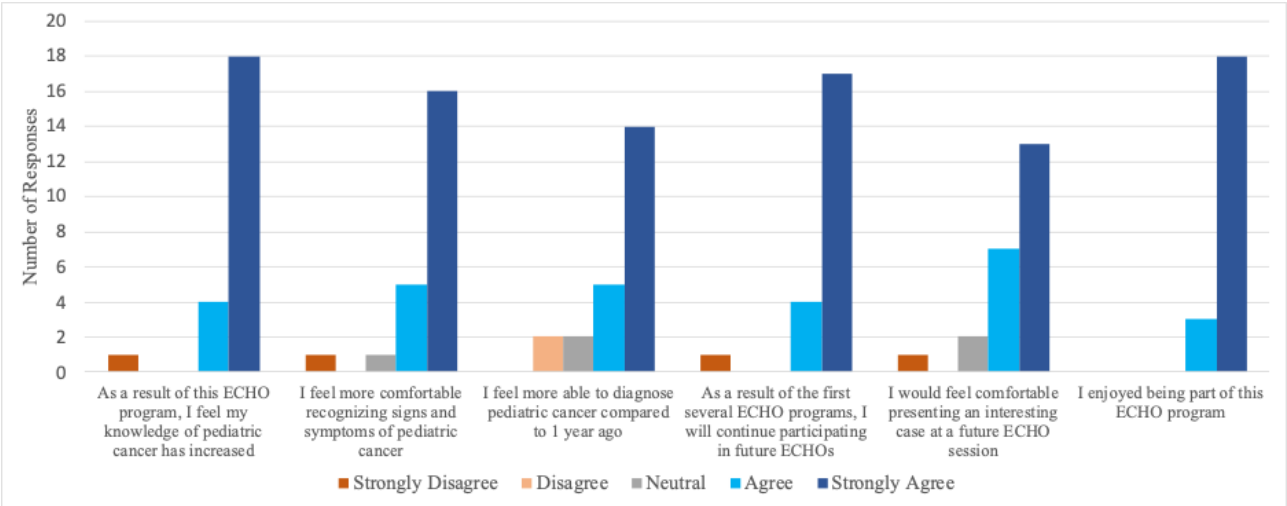

Finally, the pediatric cancer registry continued to track diagnostic rates during the course of the ECHO pilot. In the calendar year of the pediatric cancer ECHO, a total of 286 new pediatric patients were diagnosed with cancer at MTRH, which represents a 33% increase over the preceding 3-year average.

Discussion

Principal Findings

The Project ECHO platform created a dynamic virtual platform to continue engaging stakeholders across western Kenya. The primary goal was to deliver effective education regarding the recognition, diagnosis, and timely referral for pediatric patients with cancer presenting to local health care settings. The combination of strong attendance and favorable participant survey results suggests that the program was effective, particularly within the context of rising pediatric cancer diagnostic rates.

The ECHO program was initially started in New Mexico as a mechanism to improve access to quality care at the population level in high-income countries. While it achieved great success in this capacity, the use of Project ECHO as a mechanism to address disparities in care in LMICs remains a novel prospect. In this setting, the ECHO program was successfully implemented and maintained throughout the pilot year.

The AMPATH collaboration with MTRH, Moi University, and Indiana University, coupled with the expertise offered by the ECHO Institute and the recommended immersion training, was sufficient to launch the program. The prior decade of community outreach and academic conferences provided a strong foundation for the recruitment of spokes from medically isolated communities. In particular, county hospital providers constituted the bulk of participants during the initial pilot year. By using participant-led, case-based discussion alongside a standardized curriculum tailored to the needs of learners, the Project ECHO model led to increased engagement and participation.

As a voluntary educational program, AMPATH Project ECHO placed a large emphasis on measuring recurrent participation as a marker of perceived value from spoke participants. This became particularly valuable in the second half of the year, as attendance continued to rise (Figure 2). This increase also coincided with the start of the COVID-19 pandemic and its eventual spread to rural Kenya. A by-product of the pandemic was a general shift toward virtual platforms in medical communities [15]. With a general increase in comfort with Zoom and telehealth, additional increase in participation was noted. Further, the virtual nature of Project ECHO allowed the program to continue without interruption.

Although the program was available to learners without charge, there were still costs associated with running the platform during the pilot year. The experts on the Hub Team volunteered to participate, and the coordinator of the program received partial funding support through grant sponsorship. Long-term, additional sources of funding will be required to continue running the program effectively or broaden its impact. While the participants could join sessions free of charge, they were not compensated for their time. Many ECHO programs will award continuing education credits or the international equivalent for participation. However, due to funding limitations and logistical challenges, these were not approved for in time to be incorporated during the pilot period.

Comparison With Prior Work

AMPATH is among the several organizations using the ECHO model to address disparities in cancer care over the past decade [16]. For example, starting in 2018, the ECHO Institute partnered with health care leaders from various African nations to support implementation of the National Cancer Control Plan [17,18]. These efforts included using the ECHO model to support LMICs in navigating the rising cancer incidence. Although there was Kenyan representation at this monthly ECHO, it focused on health care leadership within the Ministry of Health rather than on clinicians at referring hospitals [17].

Other ECHO programs localized within Sub-Saharan Africa, supported by international collaboration have also shown recent success. One such ECHO program focuses on providing education to nursing team members in Cameroon, as they look to improve care for patients with gynecologic cancers and suggests successful implementation based on participation and

XSL•FO
RenderX

survey results [19]. Another program focused on the impact of lung cancer and improving recognition of early signs and symptoms by educating doctors and nurses in rural regions of the Northern Cape Province. This group also reported success using the Project ECHO model for training and improving awareness [20].

Expanding beyond acute cancer diagnosis and treatment, additional Project ECHO efforts have focused on supportive care such as the Palliative Care ECHO. A team from the MD Anderson Cancer Center collaborated with colleagues in 5 countries in Sub-Saharan Africa to deliver telehealth education via the ECHO model, with favorable feedback from participants [21]. Another ECHO led by Columbia University and partners throughout Africa focused on improving laboratory practice, collaboration, and resource sharing and has experienced strong participation and growth since its launch in 2018 [22]. Both programs demonstrated successful implementation based on participation and survey results [21,22].

Limitations

The survey results offer additional insight into the participant experience, although they must be interpreted in the context of the limited response rate. Those participants that completed the survey reflected a general positive experience, as indicated by a majority of respondents selecting “agree” or “strongly agree” for each of the six prompts. This suggests that the knowledge, comfort, and confidence of recognizing and diagnosing pediatric cancer improved during the pilot Project. It is recognized that those with a positive experience may be more likely to complete the voluntary survey associated with the program. However, with voluntary education, enjoyment of the program remains critical to sustained success. The emphasis on creating a pleasant and collegial environment is supported by favorable responses to the final question.

In the preceding years, typical educational interventions focused on annual 1‑ to 2- day conferences targeted at medically underserved communities. Although pre- and post-testing suggest there is potential benefit to be gained from instruction, this knowledge can wane over time–particularly without continued mentorship and active application [23]. The ECHO model was ideal for addressing this gap in instruction as the regularly occurring sessions combined with dialogue-based collaboration and telementorship further augment the growth of local HCPs. However, the generalized registry data reflect trends across the referral region. It is possible that the cumulative effect of other outreach or engagement programs may impact the interpretation of increase in diagnostic rates.

Finally, it is recognized that true assessment of the impact of educational programs such as this one and the impact on public health outcomes is difficult, particularly in a single year of implementation. Future evaluations will aim to correlate involvement and participation in the Project ECHO platform with regional oncology data. For example, trends in diagnosis at the county level, early mortality in cancer care, and even survivorship care could be impacted by sustained education and engagement with local health care teams. Future investigations would also be better positioned to provide comparison or controls by including counties that do not actively participate in the Project ECHO program.

Conclusions

Given the generalized success of early implementation of the ECHO model, there is significant scope to expand its potential benefits to impact medically isolated communities in Kenya as well as across Sub-Saharan Africa. Further, current evaluations of ECHO programs are largely limited to participation and survey results which are common in educational program assessments. The unique position of MTRH as the primary pediatric oncology treatment center in the region allows for rapid assessment of impact based on direct availability of diagnostic data and allows additional insight into the impact on public health outcomes. During the pilot year, analyses showed an increase in new oncology diagnoses at MTRH and support continued implementation with additional evaluations and analysis in the coming years. Finally, a favorable impact on the pediatric oncology care also suggests the potential to expand ECHO implementation beyond childhood cancer to address other impactful diseases at the regional level.

Acknowledgments

This paper was made possible, in part, with support from the Indiana Clinical and Translational Sciences Institute funded, in part by Grant Number UL1TR002529 from the National Institutes of Health, National Center for Advancing Translational Sciences, Clinical and Translational Sciences Award. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Funding support was made available through our partnership with Takeda Pharmaceutical Company

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

References

XSL•FO
RenderX

1. Geel JA, Challinor J, Ranasinghe N, et al. Pediatric cancer care in Africa: SIOP Global Mapping Program report on economic and population indicators. Pediatr Blood Cancer 2021 Nov;68(11):e29345. [doi: 10.1002/pbc.29345] [Medline: 34519435]
2. Bhakta N, Force LM, Allemani C, et al. Childhood cancer burden: a review of global estimates. Lancet Oncol 2019 Jan;20(1):e42-e53. [doi: 10.1016/S1470-2045(18)30761-7] [Medline: 30614477]
3. Kenya population and housing census, volume I: population by county and sub-county. Kenya National Bureau of Statistics. 2019. URL: https://housingfinanceafrica.org/app/uploads/VOLUME-I-KPHC-2019.pdf [accessed 2025-05-14]
4. Howlader N, Noone AM, Miller D, et al. SEER cancer statistics review (CSR) 1975-2018. : National Cancer Institute; 2021 URL: https://seer.cancer.gov/archive/csr/1975_2018/index.html? [accessed 2025-05-14]
5. Severance TS, Njuguna F, Olbara G, et al. An evaluation of the disparities affecting the underdiagnosis of pediatric cancer in Western Kenya. Pediatr Blood Cancer 2022 Oct;69(10):e29768. [doi: 10.1002/pbc.29768] [Medline: 35593641]
6. Inui TS, Nyandiko WM, Kimaiyo SN, et al. AMPATH: living proof that no one has to die from HIV. J Gen Intern Med 2007 Dec;22(12):1745-1750. [doi: 10.1007/s11606-007-0437-4] [Medline: 17972138]
7. Quigley F. Walking Together, Walking Far: How a US and African Medical School Partnership Is Winning the Fight against HIV/AIDS2009: Indiana University Press; 2009.
8. Severance T, Olbara GK, Njuguna F, Treff M, Loehrer PJ, Vik TA. Novel approach to improve the diagnosis of pediatric cancer in Kenya via telehealth education. JCO 2021 May 20;39(15_suppl):11000-11000. [doi: 10.1200/JCO.2021.39.15_suppl.11000]
9. Arora S, Kalishman S, Dion D, et al. Partnering urban academic medical centers and rural primary care clinicians to provide complex chronic disease care. Health Aff (Millwood) 2011 Jun;30(6):1176-1184. [doi: 10.1377/hlthaff.2011.0278] [Medline: 21596757]
10. Zhou C, Crawford A, Serhal E, Kurdyak P, Sockalingam S. The Impact of Project ECHO on participant and patient outcomes: a systematic review. Acad Med 2016 Oct;91(10):1439-1461. [doi: 10.1097/ACM.0000000000001328] [Medline: 27489018]
11. Project ECHO: a revolution in medical education and care delivery: UNM School of Medicine; 2021. URL: https://echo.unm.edu [accessed 2023-07-25]
12. Severance TS, Milgrom Z, Carson A, et al. Cancer prevention, screening, and survivorship ECHO: a pilot experience with an educational telehealth program. Cancer Med 2022 Jan;11(1):238-244. [doi: 10.1002/cam4.4421] [Medline: 34816614]
13. Kruk ME, Lewis TP, Arsenault C, et al. Improving health and social systems for all children in LMICs: structural innovations to deliver high-quality services. Lancet 2022 May 7;399(10337):1830-1844. [doi: 10.1016/S0140-6736(21)02532-0] [Medline: 35489361]
14. Arora S, Thornton K, Murata G, et al. Outcomes of treatment for hepatitis C virus infection by primary care providers. N Engl J Med 2011 Jun 9;364(23):2199-2207. [doi: 10.1056/NEJMoa1009370] [Medline: 21631316]
15. Hasson SP, Waissengrin B, Shachar E, et al. Rapid implementation of telemedicine during the COVID-19 pandemic: perspectives and preferences of patients with cancer. Oncologist 2021 Apr;26(4):e679-e685. [doi: 10.1002/onco.13676] [Medline: 33453121]
16. Eldridge L, Cira MK, Duncan K, Pearlman P, Gopal S. The 8th Symposium on Global Cancer Research: recognizing creativity and collaboration to support global cancer research and control. JCO Glob Oncol 2020 Jul;6:1-3. [doi: 10.1200/GO.20.00199] [Medline: 32716655]
17. Nakaganda A, Cira MK, Abdella K, Uwinkindi F, Mugo-Sitati C, Duncan K. Expanding best practices for implementing evidence-based cancer control strategies in Africa: The 2019-2020 Africa Cancer Research and Control ECHO Program. J Cancer Policy 2021 Jun;28:100286. [doi: 10.1016/j.jcpo.2021.100286] [Medline: 35559915]
18. Duncan KCM, Ng'ang'a A. Use of telementoring to advance cancer control: The 2018 Africa Cancer Research and Control ECHO Programme: Cancer Control; 2018. URL: https://www.cancercontrol.info/wp-content/uploads/2019/07/Kalina-Duncan.pdf [accessed 2023-07-25]
19. Fokom Domgue J, Baker E, Manjuh F, et al. Connecting frontline providers in Africa with distant experts to improve patients' outcomes through Project ECHO: a successful experience in Cameroon. Int J Gynecol Cancer 2019 Nov;29(9):1446-1447. [doi: 10.1136/ijgc-2019-000405] [Medline: 31118215]
20. Osei-Fofie D. Effective collaborative project to reduce disparities and inequities in cancer care and for cancer control in low-resource setting countries: South African experience. JCO Global Oncology 2020 Jul;6(Supplement_1):65-65. [doi: 10.1200/GO.20.62000]
21. Yennurajalingam S, Amos CE Jr, Weru J, et al. Extension for community healthcare outcomes-palliative care in Africa program: improving access to quality palliative care. J Glob Oncol 2019 Jul;5:1-8. [doi: 10.1200/JGO.19.00128] [Medline: 31335237]
22. Lehmer J, Struminger B. Virtual communities of practice strengthen public health: using technology for mentoring and sharing of knowledge and best practices across africa and beyond. Lab culture: advancing the laboratory profession and networks in Africa. 2019. URL: https://aslm.org/wp-content/uploads/2019/11/ASLM_Lab_Culture_Issue_22.pdf?x26552 [accessed 2023-07-25]
23. Young JQ, Van Merrienboer J, Durning S, Ten Cate O. Cognitive load theory: implications for medical education: AMEE Guide No. 86. Med Teach 2014 May;36(5):371-384. [doi: 10.3109/0142159X.2014.889290] [Medline: 24593808]

XSL•FO
RenderX

Abbreviations

AMPATH: Academic Model Providing Access to Healthcare
ECHO: Extension for Community Healthcare Outcomes
HCP: health care provider
LMIC: low- to middle-income country
MTRH: Moi Teaching and Referral Hospital

Edited by A Mavragani; submitted 24.04.24; peer-reviewed by J(Thomas, R Kohli, S Rayala; revised version received 04.03.25; accepted 05.03.25; published 03.06.25.

Please cite as:
Severance T, Olbara G, Njuguna F, Kipng'etich M, Lang'at S, Kugo M, Lemmen J, Treff M, Loehrer P, Vik T
Piloting the Extension for Community Healthcare Outcomes (ECHO) Pediatric Oncology Telehealth Education Program in Western Kenya: Implementation Study
JMIR Form Res 2025;9:e59776
URL: https://formative.jmir.org/2025/1/e59776
doi:10.2196/59776

© Tyler Severance, Gilbert Olbara, Festus Njuguna, Martha Kipng'etich, Sandra Lang'at, Maureen Kugo, Jesse Lemmen, Marjorie Treff, Patrick Loehrer, Terry Vik. Originally published in JMIR Formative Research (https://formative.jmir.org), 3.6.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Community Health Worker Diabetes Prevention Awareness Training in an Immersive Virtual World Environment: Mixed Methods Pilot Study

Laurie Ruggiero[1], PhD; Lauretta Quinn[2], RN, PhD; Amparo Castillo[3], MD, PhD; Colleen Monahan[4], DC, MPH; Leticia Boughton Price[5], MSW; Wandy Hernandez[5], BA

[1]Health Behavior and Nutrition Sciences, College Of Health Sciences, University of Delaware, Rm 315, 100 Discovery Blvd, Newark, DE, United States

[2]Biobehavioral Nursing Science, College of Nursing, University of Illinois at Chicago, Chicago, IL, United States

[3]Community Health Sciences, School of Public Health, University of Illinois at Chicago, Chicago, IL, United States

[4]School of Public Health, University of Illinois at Chicago, Chicago, IL, United States

[5]Illinois Community Health Workers Association, Chicago, IL, United States

Corresponding Author:
Laurie Ruggiero, PhD
Health Behavior and Nutrition Sciences, College Of Health Sciences, University of Delaware, Rm 315, 100 Discovery Blvd, Newark, DE, United States

Abstract

Background: The burden of diabetes and obesity are greater for some racial-ethnic minority groups in the United States, including non-Hispanic blacks, underscoring the importance of raising community awareness of diabetes prevention. Community health workers (CHWs) play a critical role in extending our reach into communities to raise awareness of diabetes prevention. Systematic training and support are central to their work. Remote approaches have been helpful in delivering training to overcome common participation barriers. One remote approach, immersive 3D virtual worlds (VW) offer a unique approach to providing remote training incorporating engaging interactive contextual learning opportunities.

Objective: This study aimed to implement and evaluate an internet-based 3D VW model to remotely deliver an adapted CHW training program on diabetes prevention awareness for racial-ethnic minority communities.

Methods: A sequential mixed methods design, including a pre-post pilot and explanatory phase, examined the feasibility, acceptability, and impact of the VW training. Female CHWs who self-identified as African American or Black or African Ancestry, between 21 - 65 years of age, fluent in English, and with risk factors for diabetes were recruited. CHW input was gathered to adapt a Centers for Disease Control and Prevention's CHW diabetes prevention awareness training and the VW environment for this study. The final adapted training was standardized for delivery over 10 weeks. Quantitative and qualitative data were collected to examine acceptability, feasibility, and impact of the training model. Primary quantitative pre-post outcomes included training content knowledge and confidence; and secondary behavioral outcomes included motivation for lifestyle change and eating habits. Focus group feedback was collected on acceptability and feasibility during the explanatory phase. Quantitative descriptive and qualitative thematic analysis approaches were used to examine the acceptability, feasibility, and impact of the VW training model.

Results: A total of 26 CHWs initiated the study and 22 completed the postassessment. The majority of participants reported that their expectations were met across all sessions and content topics. Participants generally reported satisfaction with the information provided (20/22, 91% rated very good-excellent) and high levels of interactivity in the training (17/22, 77% rated very good-excellent). Results of the posttraining acceptability and feasibility quantitative survey and qualitative feedback were generally positive. Mean pre-post values improved across all quantitative outcomes for the VW training group (eg, 92% [11/12] improved in knowledge; 62% [8/13] - 77% [10/13] improved across eating habits measures). Explanatory focus group findings were generally positive, highlighting satisfaction with the overall training, its interactivity, and content. The main constructive feedback was related to providing more training and support in using the avatar.

Conclusions: Findings on the acceptability, feasibility, and preliminary impact of the VW training model are promising and support continued use, development, and research on this approach.

Trial Registration: ClinicalTrials.gov NCT04161846; https://clinicaltrials.gov/study/NCT04161846

(JMIR Form Res 2025;9:e64051) doi:10.2196/64051

XSL•FO
RenderX

KEYWORDS

diabetes prevention; virtual world; community health workers; African American; training; community health worker; pilot study; obesity; community awareness; remote training; mixed-methods; acceptability; feasibility

Introduction

Obesity and Diabetes Risk

The burden of diabetes and obesity are greater for racial-ethnic minority groups in the United States, including non-Hispanic Blacks [1,2]. Estimates (2019‑2021) of the prevalence of diagnosed diabetes in the United States indicate that 12.5% of Black non-Hispanic adults (>18 years) have diabetes compared with 8.5% of white non-Hispanic adults [3]. Obesity is a major risk factor for the development of type 2 diabetes mellitus (T2DM). Estimates from the National Health and Nutrition Examination Survey (2017‑2020) indicated that Black non-Hispanic adults (>20 years) had the highest prevalence rates of obesity (49.9%) compared with non-Hispanic Whites (41.4%), non-Hispanic Asian (16.1%), and Hispanic individuals (45.5%) [4]. The greater prevalence of obesity and diabetes underscores the importance of raising awareness of diabetes and its risk in this racial-ethnic group and connecting individuals with available evidence-based programs, such as the National Diabetes Prevention Program [5].

Community Health Workers Extend Our Reach in Underserved Communities

The Community health worker (CHW) Section of the American Public Health Association describes CHWs in the following way: “frontline public health worker who is a trusted member of and/or has an unusually close understanding of the community served. This trusting relationship enables the worker to serve as a liaison or link or intermediary between health or social services and the community to facilitate access to services and improve the quality and cultural competence of service delivery. A CHW also builds individual and community capacity by increasing health knowledge and self-sufficiency through a range of activities such as outreach, community education, informal counseling, social support and advocacy” [6]. CHWs generally share characteristics of the community they serve, such as language, race-ethnicity, and lived experience. Evidence has been accumulating on the role and impact of CHWs involvement in both supporting people with diabetes [7-10] and promoting diabetes prevention [11-13]. Equipping and supporting CHWs in raising community awareness about diabetes risk factors and prevention in higher risk and underserved communities can support the public health effort to improve health equity.

Training and Empowering CHWs to Support Diabetes Prevention Efforts

Systematic training and support are central to the work of CHWs [14]. Studies have described different training approaches, and some have examined the link between training and outcomes [15]. Research comparing training approaches is limited and a recent systematic review underscored the lack of research specifically examining CHW training approaches and their impact with African American communities [16].

Remote approaches have been helpful in delivering training that maximizes the reach and may help overcome common barriers to participation (eg, transportation and childcare) and novel unanticipated participation barriers, such as experienced during the COVID-19 pandemic. The value of remote approaches to working and providing training was clearly recognized during the COVID-19 pandemic [17]. The use of video conferencing has become a regular tool to reach CHWs and help remove participation barriers to training. Another remote approach, immersive 3D virtual worlds (ie, using avatars), offers a unique opportunity to provide remote training, while also offering additional engaging interactive contextual training opportunities.

Leveraging the Virtual World to Provide Remote Training

Internet-based immersive 3D virtual worlds (eg, Second Life) offer the opportunity to reach populations with standardized and tailored training that can overcome common barriers to participation and allow for engaging, interactive, and contextual or experiential educational opportunities. Individuals can synchronously interact with each other (using voice or text chat) and the learning environment through avatars. A virtual world (VW) model has many potential advantages in offering trainings for CHWs: (1) scalable to disseminate to diverse populations remotely; (2) supports internal validity (eg, standardized delivery); (3) removes common barriers to participation (eg, transportation time or expense, child care); (4) can be tailored (eg, virtual environment, activities, content) for diverse groups; (5) provides engaging interactive contextual or experiential learning opportunities to build knowledge, skill, confidence, and empathy; and (6) allows for social interaction and peer-to-peer learning; and participation does not require special equipment (other than a computer). Research has demonstrated the use of Second Life to deliver health education in general [18] and a few published studies have shown promise of its use to help people make healthy lifestyle changes for weight loss or to provide diabetes self-care education [19-23]. To our knowledge, no published research has examined the use of this immersive 3D VW delivery model for implementing a CHW training, in general, or specifically to raise awareness about diabetes prevention to help extend efforts in underserved at-risk racial-ethnic minority communities.

Objective and Aims

Our long-term objective is to identify effective strategies to promote healthy lifestyle change to reduce obesity and diabetes risk in underserved populations and communities. The overall aims of this study were to adapt, implement, and evaluate a VW model to remotely deliver an adapted best-practice CHW training (based on Center for Disease Control and Prevention’s “Road to Health Toolkit”) [Road to Health; RTH] to support CHWs efforts to raise awareness about diabetes prevention in racial-ethnic minority communities.

XSL•FO
RenderX

This study was conducted before and during the COVID-19 pandemic (details described later). The original prepandemic aims were to conduct a randomized implementation pilot study to examine change in primary outcomes (0-12 wk) after the in-person (RTH) or VW delivery (RTH-VW) of the training program; and examine feasibility and acceptability of the VW delivery of the RTH training. The original hypotheses were: the 2 training approaches would result in similar preliminary outcomes (ie, equivalency); and the VW delivery would be feasible and acceptable to CHWs. We were unable to complete the randomized controlled trial as planned because of COVID-19 pandemic restrictions, therefore, all subsequently enrolled participants were assigned to the RTH-VW. This paper describes all available study participants (RTH and RTH-VW) and focuses on the RTH-VW group.

Methods

Design

A randomized 2-group (RTH and RTH-VW) repeated measures design (baseline and posttraining) was initially implemented (prepandemic). COVID-19 pandemic-related study delays (ie, 9 mo) and challenges (eg, delayed groups; inability to hold in person meetings) led to the modification of the experimental design to a quasi-experimental single-group pre-post design (RTH-VW only).

A sequential mixed methods research design [24] incorporating both quantitative and qualitative approaches was used. This paper describes a pre-post pilot phase examining acceptability, feasibility, and the preliminary impact of the training, followed by an explanatory phase clarifying and expanding upon the findings of the pilot. The primary quantitative variables included primary training-related outcomes (ie, knowledge and confidence). An additional objective of the study was to include CHWs who experience risk factors similar to the population they serve and to offer the opportunity to personally implement the lifestyle changes learned to help facilitate experiential learning and empathy for the challenges that may be experienced by others making similar lifestyle changes. Therefore, we also examined secondary behavior change outcomes.

Training Description

Adapted CHW Diabetes Prevention Training

We chose the Center for Disease Control and Prevention's (CDC) Division of Diabetes Translation's (DDT) RTH Toolkit [25] as the CHW training program. The 2008 version was used since it was the only version available at the start of the study (updated version now available). It was chosen because it was developed with CHW collaboration and focus group testing; incorporates storytelling, literacy tailoring and cultural tailoring for African American communities; and includes a comprehensive set of resources to support standard implementation. The overall RTH training goals are to raise awareness of diabetes, its risk factors, and prevention activities (making healthy food choices and physical activity) to reduce weight in those overweight or obese to help lower risk of T2DM. The RTH Toolkit content and resources adapted for this study served as the core CHW training.

The 3 main diabetes prevention messages of the RTH toolkit were used and can be seen in Figure 1 (based on 2008 version). Resources to support CHW participation and facilitate diabetes prevention community awareness efforts were included: (1) original resources from the RTH Toolkit [26] (eg, training and resource guides, posters, eating and activity tracking tools, and flip chart) and (2) supplemental content-related resources (eg, American Diabetes Association and CDC prediabetes risk test [27], US Department of Agriculture MyPlate resources [28], and US Food and Drug Administration nutrition facts label information [29].

Based on an intensive community-engaged development phase incorporating CHW input over several months to gather feedback on iterative versions of adapted training and VW environment, the final training was adapted and standardized for delivery over 10 weeks with an orientation or overview session, 8 content sessions delivered biweekly (2 topics or session) and final summary session. The 10-week duration of the training was designed to allow CHWs the opportunity to try the strategies learned in their own lives, if desired, to gain personal experience in lifestyle change.

XSL•FO
RenderX

Figure 1. Diabetes prevention awareness primary content areas and sample topics from the "Road to Health Toolkit" adapted for the CHW training. Note: This study used the 2008 CDC "Road to Health Toolkit" and graphics are from that version; an updated version is available. T2DM: type 2 diabetes mellitus.

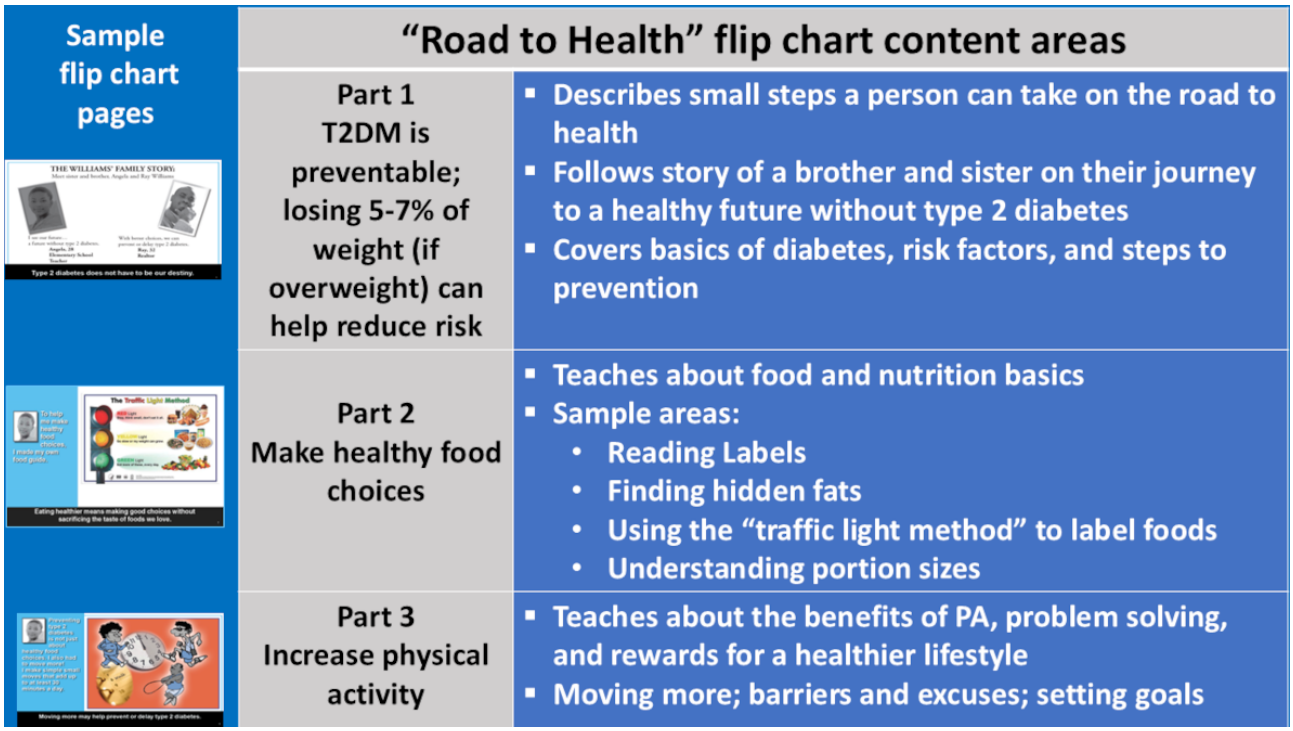

VW Description

The VW environment used was Second Life, a computer-based simulated 3D environment, intended for users to inhabit and interact via an avatar (VW representation of the user). Users are identified by their assigned avatar name and can communicate with other users (through avatars) in real time using voice and text chat tools. Real life anonymity is possible unless the individual chooses to share their real-world identity. Based on CHW input in the development phase of this project, we adapted the VW environment to support delivery of the RTH training. The Second Life VW created for the study island was private. Only study participants and research staff (as avatars) had access and study avatars could not travel to other spaces in Second Life.

The final adapted environment included synchronous educational sessions in the VW classroom or in other VW locations (eg, fast food restaurant), immersive experiential activities (eg, comparing food labels in grocery store), access to educational resources, educational messages posted throughout the virtual environment (eg, MyPlate messages in grocery store), peer interactions, and knowledge self-checks opportunities. Example scenes from the VW environment are shown in Figure 2.

XSL•FO
RenderX

Figure 2. Examples of the virtual world environment and activities. This study used the 2008 CDC "Road to Health Toolkit" and graphics are from that version; an updated version is available. RTH: Road to Health Toolkit; VW: virtual worlds.

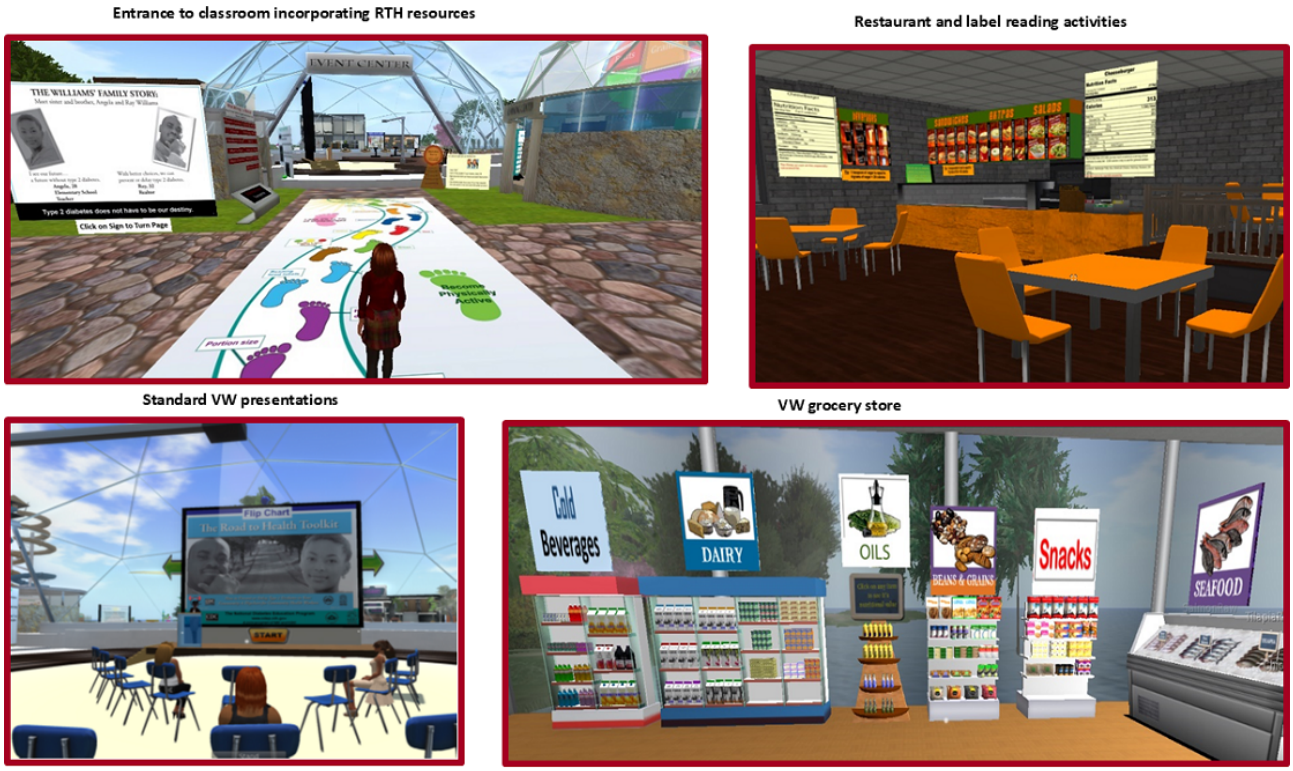

Study Population and Recruitment.

The eligibility criteria included: female, self-identify as African American or Black or African Ancestry; adult (21 - 65 y old); self-report as overweight or obese based on BMI; currently inactive, defined as not participating in regular moderate physical activity (150 min/wk), self-identify as a CHW, fluent in English, and regular access to a computer and the internet. The CHWs were purposefully recruited to share characteristics (eg, diabetes risk factors) to facilitate an understanding of the lived experience of the community members they are working to reach with messages about diabetes prevention awareness. This approach offered an opportunity for participants to implement the knowledge and lessons learned in their own lives to make lifestyle changes and gain understanding of and empathy for the process and challenges of making healthy lifestyle changes. Exclusion criteria included: (1) household member already participating in study, (2) plans to travel or move from the city during the study period, and (3) self-reported inability to participate in regular moderate physical activity.

CHW participants were recruited in collaboration with the Illinois Community Health Worker Association (ILCHWA). Project recruitment information was shared with potential participants through researcher-led informational sessions and distribution of study flyers at CHW professional meetings in the Chicago-area, and using various communication strategies (eg, email and social networking approaches) implemented by the ILCHWA to share the study flyer. Similar recruitment approaches also occurred through research team connections and communications with various CHW community and academic groups, health systems employing CHWs, and CHW training programs in the Chicago area.

Ethical Considerations

Interested individuals were instructed to call a study phone number to get more information about the study and to determine their eligibility for study participation. University of Illinois at Chicago (UIC) research team members conducted phone eligibility screening. Eligible and interested individuals attended an in-person meeting at UIC (prepandemic) or via teleconferencing (during pandemic) to complete the study orientation, written paper consent (eg, completed digitally using REDCap [30,31] during pandemic). Study staff fully informed individuals of the study (eg, procedures, incentives, and assessments), addressed questions, and obtained consent. Participants received US $25 for completing the first in-person assessment and US $100 for completing the postintervention in-person assessment. An additional US $25 was provided for focus group (described below) participation. To maximize retention at assessments, announcements were made during program sessions and follow-up prompts were conducted. Efforts were made to minimize barriers to participation (eg, parking reimbursement) for in-person meetings. This study involved human participants; the study protocol was reviewed and approved by the University of Delaware (#1106606 - 20) and University of Illinois (IRB 2018 - 0986) institutional review boards.

Implementation

Following the baseline assessment and (pre-COVID-19 pandemic), CHWs were randomized to either the in-person or VW CHW training program. After assessment and randomization, participants took part in the CHW training either virtually (RTH-VW) or in person (RTH). For both the RTH and RTH-VW, the training sessions were delivered in small cohorts (ie, up to 11) and led by the same facilitator (ie, Registered

XSL•FO
RenderX

Dietitian) who had extensive experience in delivering group health education programs with diverse populations, both in person and in the 3D VW environment. Before COVID-19 restrictions, the facilitator and CHWs participated as avatars in the VW for the RTH-VW groups and the RTH (in person) groups met in a university building.

The training was designed to be delivered in a standard way across the 2 delivery formats (in-person and VW), including consistent content delivery, supplemental activities, follow-up home activities (eg, review favorite fast-food menu to find healthier choices) and opportunity for peer interaction. For the RTH-VW implementation, participants were provided with an initial technology-focused training session (eg, downloading application, logging in, and use of avatars) and offered a variety of options for follow-up technical support, (eg, videoconferencing, meetings in virtual environment, and telephone support), where needed. Consistent with the RTH delivery, core content sessions in the RTH-VW were generally held in a VW classroom-like setting using the same slide presentations as the RTH to maintain consistency in delivery. Some RTH-VW sessions also included facilitated discussions in immersive locations in the VW to maximize experiential learning (eg, discussing making healthier choices in the fast-food restaurant). The RTH-VW also included opportunities for contextual activities (eg, reading nutrition facts labels in a grocery store), on-demand activities (eg, knowledge check kiosks; activity comparing nutrition facts labels for calories) and on-demand access to educational resources and VW physical activity options (eg, riding a bike, refer to Figure 2).

COVID-19-Related Modifications

As noted earlier, the COVID-19 pandemic and restrictions resulted in substantial (eg, 9 mo) delay in the study, thereby greatly impacting the implementation (eg, delayed groups, inability to hold in-person sessions, and multiple protocol and IRB revisions). To continue the study during the pandemic and maintain the primary goal of examining the VW implementation, the design was modified to assign participants only to the RTH-VW group and all study activities occurred remotely (ie, videoconference consent process, orientation, postprogram meeting, and training in use of VW application, VW training, and online assessments).

Assessment Measures and Methods

Background and Sociodemographic Information

A baseline survey collected basic background, health-related, and sociodemographic information.

Feasibility and Acceptability

Quantitative and qualitative approaches were used to gather participant feedback regarding acceptability and feasibility, especially related to the VW approach. Brief quantitative session feedback surveys tailored for each session's topics were collected following each session to examine content satisfaction. Participants were asked to rate how well their expectations were met for the session topics with the following response choices: all were met, some were met, very few or none were met. A posttraining survey included quantitative and qualitative questions designed to examine acceptability and feasibility of the training implementation. Example multiple choice questions included: Overall, how would you rate your satisfaction with the information presented during the program? and "Overall, how would you rate the level of interactivity of the program?" with response choices of "poor," "fair," "good," "very good," and "excellent." Example open-ended qualitative questions included: "What did you think about the program?"; "What was the most helpful part of the program?"; and "What part of the program needs to be improved?."

Quantitative Outcomes

The primary quantitative pre-post training constructs included training content knowledge and confidence in content and delivery skills. In addition, a secondary exploratory quantitative assessment of behavioral outcomes was conducted. Self-report assessments were conducted electronically throughout the study using Qualtrics [32]software. The pre-post assessments (0, 10 - 12 wk) were conducted in-person using a mobile device at a university location before the COVID-19 pandemic and were collected remotely with a survey link during the pandemic. Some measures were omitted due to differences in the measurement approaches (eg, lab-measured BMI vs self-report) and restrictions on daily activities (eg, physical activity measure) due to the evolving COVID-19 pandemic guidance (eg, stay at home, gyms closed, and social distancing).

Content Knowledge

Content knowledge was assessed with 22 adapted true or false and multiple-choice knowledge questions taken from the RTH "Toolkit Training Guide" (2008 version) and specific content of the training. Example questions were: "A light bulb size portion equals three servings of vegetables" (true or false) and "In the DPP study, how many minutes of physical activity (for example, brisk walking) did people have to do to prevent or delay type 2 diabetes?" At least 30 minutes every day; 30 minutes 5 days/week; 60 minutes every day; 60 minutes 5 days/week; none of the above (multiple choice). A total correct score was used for analyses (range 0 - 22).

Confidence in RTH Content and Delivery Skills

This score was computed as an index based on 5 items that asked about the confidence in the training content and implementation of community initiatives and response choices ranged from "not at all confident" to "extremely confident" (range=5 - 25). Examples of questions include: "How confident are you that you could present a community awareness event on diabetes prevention?" and "How confident are you in your ability to make content engaging and interactive when delivering a community-based presentation?."

Eating Habits

CHW eating habits were assessed using the Visually-Enhanced Food Behavior Checklist [33] and Rate Your Plate (RYP)-Heart questionnaire [34]. The Visually-Enhanced Food Behavior Checklist aims to "enhance usefulness of food behavior surveys for low-literate populations using visual information processing theories." It is intended to be used with adults and measures medium-term changes in dietary habits. For the purposes of this study, two variables were examined: (1) daily intake (cups) of

XSL•FO
RenderX

fruits and vegetables (ie, composite score of daily fruits and daily vegetable intake in cups; range=none to 6 cups or more) and (2) self-rated eating habits (range=0‑10; poor=0 to excellent=10). The RYP questionnaire includes 24 items (range=24‑72) designed to assess qualitative nutrition information related to the typical food choices, specifically examines intake of 16 food categories, and includes items on food preparation, eating out, and serving sizes (lower score represents healthier eating habits).

Stage of Change and Self-Efficacy

Stage of change (SOC) items assessed CHW behavioral motivation (ie, precontemplation, contemplation, and preparation) or achievement of eating habit and physical activity behaviors (ie, action or maintenance stages) across multiple areas, including reducing high fat dairy and protein foods; avoiding sugary beverages; reading Nutrition Facts labels; increasing fruit or vegetables, and getting moderate physical activity. A multi-item index of SOC was used to examine behavioral motivation or change based on coding stage as “Precontemplation=1” to “Maintenance=5”. The SOC items were combined into an overall multi-item SOC index by summing the scores across all 6 SOC items (range=6 to 30).

A multi-item index of behavioral self-efficacy was used to examine confidence in making lifestyle changes (ie, same behavioral areas as SOC). Confidence items had a 5-point response scale, ranging from “not at all confident=1” to “extremely confident=5”. The confidence individual item scores were combined into an overall confidence index by summing the score across all confidence items (range=6 to 30).

Adherence and Fidelity

Researchers regularly observed program sessions or activities to monitor fidelity, observe attendance, and examine feasibility aspects of delivering the training. Make-up sessions were offered for each missed session.

Posttraining Explanatory Focus Groups

Semistructured focus groups were conducted following the training to gather information on acceptability (eg, likability, tailoring, and suggested refinements), feasibility (eg, facilitators or barriers to participation or VW use), and general feedback. Examples of core structured questions include: (1) What did you think about the VW application? (2) How would you describe this program or VW to a friend or a family member? (3) What was your favorite part and why? (4) What did you think of the different activities in the VW? (5)What did you think of presentations in the VW? And (6) what were your expectations and were they met?

Analyses

Quantitative and qualitative analysis approaches were used to examine the acceptability, feasibility, and preliminary impact of the VW training model.

Quantitative Analyses

The primary focus of the research was learning about the VW approach; however, there were COVID-19-related design or implementation modifications limiting randomization. The descriptive results (eg, mean, SD, sample size, and percentage) are presented for the RTH (in person, n=8) and the combined RTH-VW subsample (n=18). Due to the small sample size, only descriptive statistics are included. For the RTH-VW group, the percent improved and percent unchanged are also described.

Explanatory Qualitative Analysis

Posttraining focus group recordings were transcribed and entered into the Dedoose analysis [35] application for coding and analysis. Thematic analysis was used to analyze this data [36]. First, qualitative data was deductively coded across training and implementation-related questions by 2 team members using a priori general categories of positive, negative, and neutral feedback. The total frequency of each global category was calculated. In addition, subthemes were then inductively identified from the positive and negative categories across questions. The coding was reviewed and discussed by 2 team members and inconsistencies were reviewed and resolved.

Results

Sample Characteristics

A total of 74 individuals were screened, 25 were ineligible, and 23 eligible individuals did not enroll or were unable to participate. A total of 26 CHWs initiated the study (randomized prepandemic: 8 in-person, 7 VW; during pandemic: 11 VW) and 22 completed the postassessment. The baseline characteristics (N=26 total sample; n=18 combined VW sample) include: all female and African American or Black or African Ancestry; average age of 47 years; education: 3.8% (1/26) high school, 19.2% (5/26) some college, 11.5% (3/26) associate’s degree; 65.4% (17/26) bachelors or advanced degree; and 84.6% (22/26) employed. In addition, 76.9% (20/26) reported daily computer use; 84.6% (22/26) daily internet use; 76.9% (20/26) were very comfortable using a computer; and 80.8% (21/26) were very comfortable using the internet.

Acceptability and Feasibility Findings

Individual Training Session Postsurveys (All Groups)

The findings from the individual postsession surveys combined across all training sessions indicated that greater than 95% of each group (RTH and VW-RTH), on average, reported that their expectations were met across all sessions and content topics. Findings regarding satisfaction with the information provided and level of interactivity of the training sessions across groups indicated that CHWs generally reported satisfication with the information provided (ie, poor or fair=0% [0/22]; good=9.1% [2/22]; very good=40.9% [9/22]; excellent=50% [11/22]) and reported a high level of interactivity of the training (ie, poor=0% [0/22]; fair=4.6% [1/22]; good=18.2% [4/22]; very good=36.4% [8/22]; excellent=40.9% [9/22]). These findings support overall satisfaction with the training experience.

Posttraining Survey Questions

Results of the quantitative posttraining acceptability and feasibility survey questions found very positive responses (eg, all respondents rated the general quality of the training as very good-excellent and nearly all (90-100%) liked the number of sessions, length of the sessions, and timing of sessions). All

XSL•FO
RenderX

participants agreed or strongly agreed that they would recommend the training to others; and that the training in general and specifically, the healthy eating and physical activity information, was helpful in learning about diabetes prevention to raise awareness in their communities. Participant satisfaction ratings of the information presented during the training were consistent with the other findings (ie, all rated satisfaction as good to excellent).

Exploration of 3 qualitative explanatory survey questions provided additional insights into the acceptability and feasibility of the training. When asked "what did you think about the program?," the majority of comments were positive with the main subtheme indicating that participants liked the training (eg, "good," "educational," and "informative"). Examples of specific feedback on the VW approach included "learning in the VW was enjoyable as you had the opportunity to bond and share in a unique way" and "the online session information presented was good." Constructive feedback was also offered with the main subtheme focused on improving supplemental resources (eg, bigger font handouts) and one comment suggesting providing more opportunities for avatars to work together in VW activities. When asked about the "most helpful" part of the training, the main subtheme focused on the information provided, especially healthy eating related topics (eg, reading nutrition facts labels) and the 3 core RTH diabetes prevention messages, while other common responses focused on engaging with peers, and positive characteristics of the facilitator. For the question about what "needs to be improved", the main themes included session scheduling (eg, length and timing) and challenges with using technology (eg, navigating avatar). Regarding barriers experienced in attending the training, the main themes were scheduling challenges or technology challenges.

Quantitative Pre-Post Training Outcomes

RTH training-related and exploratory behavioral outcomes are presented in Table 1. Descriptive results are presented for the randomized RTH group and combined RTH-VW group. As can be seen in Table 1, mean pre-post mean values improved for all variables. In addition, for the RTH-VW group, the majority (11/12, 91.7%) improved in knowledge (0/12, 0% unchanged) and half reported increased confidence (2/12, 16.7% unchanged). The majority also experienced improvements across all behavioral outcomes examined.

XSL•FO
RenderX

Table . Descriptive statistics for pre-post CHW behavioral and training-related outcomes for the in-person training group and the combined VW training groups. Eating habits and fruit and vegetable intake were measured with the Visually-Enhanced Food Behavior Checklist; higher scores represent better outcomes for all outcomes except for the RYP Total; total n for RTH = 8 and total n for RTH-VW =18.

Behavioral and training-related outcomes	Baseline mean (SD, n)	Posttraining mean (SD, n)	% RTH-VW improved pre-post	% RTH-VW unchanged pre-post
RTH knowledge (range 0-22)				
RTH[a]	16.13 (2.42, 8)	17.43 (2.23, 7)	—[c]	—
RTH-VW[b]	15.72 (2.76, 18)	18.58 (1.24, 12)	92%	0%
RTH content and delivery confidence (range 5-25)				
RTH	17.38 (3.81, 8)	20.29 (2.50, 7)	—	—
RTH-VW	16.17 (3.85, 18)	18.00 (2.66, 12)	50%	17%
Stage of change index (range 6-30)				
RTH	22.63 (3.81, 8)	24.43 (4.80, 7)	—	—
RTH-VW	20.00 (5.22, 18)	22.75 (3.41, 12)	75%	8%
Self-efficacy index (range 6-30)				
RTH	22.88 (4.19, 8)	26.43 (3.69, 7)	—	—
RTH-VW	20.11 (4.16, 18)	22.42 (3.34, 12)	67%	8%
Eating habits rating (range 0-10)				
RTH	5.63 (2.33, 8)	6.29 (1.70, 7)	—	—
RTH-VW	5.28 (1.65, 18)	6.00 (1.68, 13)	62%	23%
Fruit and vegetable intake (range 0 to 6 cups)				
RTH	1.94 (1.76, 8)	3.00 (1.73, 7)	—	—
RTH-VW	1.06 (0.80, 18)	2.00 (0.91, 13)	69%	31%
Rate your plate total (range 24–72)				
RTH	41.25 (11.02, 8)	40.43 (12.01,7)	—	—
RTH-VW	47.22 (8.75, 18)	40.00 (7.68, 13)	77%	8%

[a]RTH: Road to Health.

[b]RTH-VW: Road to Health- virtual world.

[c]not applicable.

Posttraining (Explanatory) Qualitative Findings

The posttraining focus groups included 10 participants from the RTH-VW group. Of the overall positive and negative feedback, 79% (71/90) of the comments were coded positive (19/90, 21% negative). The main positive subthemes focused on the overall experience, interactivity, and content learned, along with the facilitator characteristics. Of the constructive feedback, the main theme was related to navigating the avatar. Table 2 shows the subthemes and sample quotes for each theme.

XSL•FO
RenderX

Table . Posttraining focus group subthemes and example quotes for each theme. Note: The comments in brackets were added by authors for clarification.

Subthemes	Example quotes
Positive feedback subthemes	
Overall experience	• "It was innovative because I know we can't meet in person, so having class this way I can have my toddlers here playing; I know I can watch them at the same time I am doing something here…I can do two things at once, I don't have to leave the house". • "Everything. You can go in any time of the night". [Favorite part] • "I actually loved the avatar".
Interactivity	• "Coming together and learning from each other". • "In the evening times, we would just meet up…and go through the things, discuss, and basically well just have fun." • "I like in the classroom when we talked about the meal we had, and how she engaged with everyone."
Content learned	• "I really liked it [fast food restaurant] because, you know, it helped you create this awareness about choosing your options…like to take an opportunity to choose a healthier meal among fast food choices". • "I thought the slide presentations [in virtual world] were pretty good, we are able to follow along with the facilitator". • "Label reading". [Favorite part]
Constructive feedback subtheme	
Avatar navigation	• "In the beginning I believe that it was kind of challenging, I was excited about the virtual world, but navigating around, I think that was my biggest challenge." • "I enjoyed using the application in terms of, you know, moving around and everything…it takes a little getting used to but that was why there was an orientation process at the beginning to help you know and teach you how to move". • "Basically, the biggest problem was learning how to navigate."

When asked "What were your expectations and were they met?," all responses were positive. Two example quotes include: "I've learned a lot" and "It did in every aspect…I want to help myself because of this program…Now I'm ready to do more for myself." Examples of quotes about how they would describe the program or VW to a friend or a family member include:

> *we are represented there…it's like a village, everything you need is right there…you can go to the grocery store, you can go out in the morning to do some exercise, you can go to the fast-food restaurant to get your food*
>
> *it is like a classroom…we are not physically present…we all come together in the VW to learn from each other and learn from the doctor [facilitator].*

Attendance and Fidelity

Observations at sessions confirmed fidelity of content delivery and clarified aspects of feasibility. Nearly all participants of both delivery approaches either attended all sessions or completed make-up sessions.

Discussion

Principal Findings

The overall aims of this study were to adapt, implement, and evaluate a VW model to remotely deliver an adapted best-practice CHW training to support CHW's efforts to raise awareness about diabetes prevention in racial-ethnic minority communities. Integrated quantitative and qualitative findings on the acceptability, feasibility, and preliminary impact of the RTH-VW training model implemented with CHWs shows initial promise of this immersive remote approach.

The findings regarding the acceptability of the training overall were encouraging. The respondents from both delivery approaches (pre-COVID-19 randomized groups) generally provided positive feedback regarding their training experience. For example, strong positive feedback was offered about the general quality, specific content, and interactivity of the training. The postsession feedback for individual sessions for both delivery approaches indicated that nearly all respondents felt their expectations were met across all training topics. VW group participants enjoyed the interactive immersive (eg, label reading in grocery store) aspects of this approach underscoring the unique opportunities available to support learning. Finally, all respondents of the posttraining quantitative survey agreed that they would recommend the training to others. The feedback on acceptability from various sources supports the value of the training overall and both implementation approaches.

While the overall delivery of both training approaches was feasible from a research and implementation perspective (excluding pandemic-related barriers), there remains room for improvement to support participants in being able to fully participate. For example, both groups (prepandemic) continued

XSL•FO
RenderX

to experience challenges with attending scheduled sessions related to work and family schedules, possibly contributing to the inconsistency found in the feedback across assessment approaches regarding the length and timing of sessions. This feedback suggests that scheduling is an important consideration in planning any CHW training programs. Suggestions for strategies to remove potential scheduling barriers include allowing each group to identify a mutually convenient time for group members; offering multiple training session options on different days and times, including evenings and weekends; and, as done in this study, offering make-up sessions to support full participation.

Although remote VW delivery may provide unique remote experiential learning opportunities (eg, reading labels in grocery store) and overcome some barriers (eg, transportation, travel time, and parking), constructive feedback was provided on other challenges and areas to further improve feasibility of this virtual approach to maximize interactivity. The explanatory focus group provided a more nuanced understanding of this feedback. For example, some participants experienced challenges with navigating their avatars initially with improvement over time, while a few had continued challenges. Future work should offer more preliminary training in the use of the VW application, as well as continued ongoing support, to enhance feasibility. One useful approach in this study was to meet people, as avatars, in the VW shortly before sessions to review features needed for VW participation (eg, navigating avatar and communicating with other avatars). This required only a few minutes of time and was valuable in helping participants fully engage in the training.

While the results of training-specific outcomes (ie, knowledge, confidence in delivering content and activities) were encouraging, confidence in applying the knowledge could be improved. Future work related to this CHW training and other skill-based trainings might benefit from offering CHWs greater opportunity within the training program to work on tailoring community awareness approaches and activities for their communities, practicing them to enhance confidence, and gathering feedback from other CHWs and facilitators to refine their approaches before community implementation. This may be a useful strategy to support CHWs in applying new content and skills across various CHW training programs.

An exploratory and unique aspect of this study was to include CHWs with similar diabetes risk factors to the community members they serve. The training format (eg, duration, and spacing) was designed to allow participants the opportunity to implement the knowledge and skills learned in the training in their own lives to gain a better understanding of the process and complexity of making healthy lifestyle changes. The exploration of behavior change outcomes showed positive changes in a variety of areas (eg, readiness for change, confidence, fruit, and vegetable intake), thereby underscoring the potential positive impact of the training on the participant's own lifestyle behaviors. This approach may support CHWs in enhancing empathy and identifying possible behavior change tips for supporting other individuals working on making similar healthy lifestyle changes to reduce the risk of diabetes.

Strengths and Limitations

A strength of the study was the ability to pivot during the pandemic to continue to deliver the VW approach when in-person restrictions were implemented. This demonstrated that the VW approach was well-suited to implementation during this challenging unanticipated situation and underscored its promise in terms of feasibility and potential to overcome participation barriers. As previously noted, the main limitation in the study was the inability to fully implement the planned randomized design to allow statistical comparison of the 2 approaches (ie, in-person and VW). Another study limitation was the small sample size. The inability to examine all planned outcomes due pandemic-related modifications and influences was also a limitation. Therefore, the results should be interpreted with these caveats in mind.

Conclusions

Findings on the acceptability, feasibility, and preliminary impact of the RTH-VW training are promising and support continued use of this approach. As noted in the introduction and generally supported in this study, a VW model has many potential advantages in offering trainings for CHWs, such as the potential to be tailored and standardized to disseminate to diverse groups remotely to minimize some common barriers to participation; provides opportunity to deliver engaging interactive contextual or experiential learning activities to help facilitate knowledge, skill, confidence, and empathy; and allows for social interaction and peer-to-peer learning.

Future research will continue to examine this VW training approach incorporating the revised version of the CDC RTH Toolkit resources, additional immersive on-demand activities, constructive feedback from participants, and lessons learned during the implementation experience (eg, greater initial training in using the VW application).

Acknowledgments

The research team would like to acknowledge and thank the Chicago CHW community, especially the ILCHWA, for their support throughout this project; our great undergraduate and graduate research assistants from the University of Delaware and the UIC who helped implement the study; our phenomenal facilitator, Natoya Baker, MS, RD, LDN, CDCES; and our wonderful participants for their valuable contributions to the project. We would also like to thank the original developers of the "Road to Health Toolkit". Research reported in this publication was supported by the National Institute on Minority Health and Health Disparities of the National Institutes of Health under Award Number R21MD011413. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or Centers for Disease Control and Prevention.

XSL•FO
RenderX

Data Availability

The data sets generated during and analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

All authors were involved in conceptualizing the study. LR, LQ, and AC designed the study and led the training delivery, data collection, analysis, and interpretation. AC, WH, and LBP facilitated the recruitment efforts. CM supported the virtual world implementation. LR wrote the initial draft of the manuscript with consultation and input from all authors. All authors provided critical feedback on the manuscript and approved the final version.

Conflicts of Interest

None declared.

References

1. Thornton PL, Kumanyika SK, Gregg EW, et al. New research directions on disparities in obesity and type 2 diabetes. Ann N Y Acad Sci 2020 Feb;1461(1):5-24. [doi: 10.1111/nyas.14270] [Medline: 31793006]
2. Haw JS, Shah M, Turbow S, Egeolu M, Umpierrez G. Diabetes complications in racial and ethnic minority populations in the USA. Curr Diab Rep 2021 Jan 9;21(1):2. [doi: 10.1007/s11892-020-01369-x] [Medline: 33420878]
3. Centers for Disease Control and Prevention. National Diabetes Statistics Report website. URL: https://www.cdc.gov/diabetes/php/data-research/index.html [accessed 2024-12-19]
4. Stierman B, Afful J, Carroll MD, National Center for Health Statistics (U.S.), et al. National Health and Nutrition Examination Survey 2017–March 2020 Prepandemic Data Files Development of Files and Prevalence Estimates for Selected Health Outcomes. National Health Statistics Reports, NHSR 2021 Jun 14 [FREE Full text]
5. Gruss SM, Nhim K, Gregg E, Bell M, Luman E, Albright A. Public health approaches to type 2 diabetes prevention: the US national diabetes prevention program and beyond. Curr Diab Rep 2019 Aug 5;19(9):78. [doi: 10.1007/s11892-019-1200-z] [Medline: 31385061]
6. American Public Health Association CHW Section. Community Health Workers. URL: https://www.apha.org/apha-communities/member-sections/community-health-workers/ [accessed 2024-12-19]
7. Egbujie BA, Delobelle PA, Levitt N, Puoane T, Sanders D, van Wyk B. Role of community health workers in type 2 diabetes mellitus self-management: A scoping review. PLoS One 2018;13(6):e0198424. [doi: 10.1371/journal.pone.0198424] [Medline: 29856846]
8. Guide to community preventive services. community health workers help patients manage diabetes. URL: https://www.thecommunityguide.org/findings/diabetes-management-interventions-engaging-community-health-workers.html [accessed 2024-12-19]
9. Palmas W, March D, Darakjy S, et al. Community health worker interventions to improve glycemic control in people with diabetes: a systematic review and meta-analysis. J Gen Intern Med 2015 Jul;30(7):1004-1012. [doi: 10.1007/s11606-015-3247-0] [Medline: 25735938]
10. Trump LJ, Mendenhall TJ. Community health workers in diabetes care: a systematic review of randomized controlled trials. Families, Systems, & Health 2017;35(3):320-340. [doi: 10.1037/fsh0000283]
11. Hill J, Peer N, Oldenburg B, Kengne AP. Roles, responsibilities and characteristics of lay community health workers involved in diabetes prevention programmes: A systematic review. PLoS One 2017;12(12):e0189069. [doi: 10.1371/journal.pone.0189069] [Medline: 29216263]
12. Ruggiero L, Castillo A, Quinn L, Hochwert M. Translation of the diabetes prevention program's lifestyle intervention: role of community health workers. Curr Diab Rep 2012 Apr;12(2):127-137. [doi: 10.1007/s11892-012-0254-y] [Medline: 22350806]
13. Ruggiero L, Oros S, Choi YK. Community-based translation of the diabetes prevention program's lifestyle intervention in an underserved Latino population. Diabetes Educ 2011;37(4):564-572. [doi: 10.1177/0145721711411107] [Medline: 21690435]
14. CDC National Center for Chronic Disease Prevention and Health Promotion. Addressing chronic disease through community health workers: A policy and systems-level approach. 2015 URL: https://www.cdc.gov/dhdsp/docs/chw_brief.pdf [accessed 2024-12-19]
15. Seneviratne S, Desloge A, Haregu T, et al. Characteristics and outcomes of community health worker training to improve the prevention and control of cardiometabolic diseases in low and middle-income countries: a systematic review. Inquiry 2022;59:469580221112834. [doi: 10.1177/00469580221112834] [Medline: 35916447]
16. Adams LB, Richmond J, Watson SN, et al. Community health worker training curricula and intervention outcomes in African American and Latinx communities: a systematic review. Health Educ Behav 2021 Aug;48(4):516-531. [doi: 10.1177/1090198120959326] [Medline: 34254544]

XSL•FO
RenderX

17. Salve S, Raven J, Das P, et al. Community health workers and Covid-19: Cross-country evidence on their roles, experiences, challenges and adaptive strategies. PLOS Glob Public Health 2023;3(1):e0001447. [doi: 10.1371/journal.pgph.0001447] [Medline: 36962877]
18. Beard L, Wilson K, Morra D, Keelan J. A survey of health-related activities on second life. J Med Internet Res 2009 May 22;11(2):e17. [doi: 10.2196/jmir.1192] [Medline: 19632971]
19. Nosek MA, Robinson-Whelen S, Ledoux TA, et al. A pilot test of the GoWoman weight management intervention for women with mobility impairments in the online virtual world of Second Life®. Disabil Rehabil 2019 Nov;41(22):2718-2729. [doi: 10.1080/09638288.2018.1473511] [Medline: 29889580]
20. Johnson C, Feinglos M, Pereira K, et al. Feasibility and preliminary effects of a virtual environment for adults with type 2 diabetes: pilot study. JMIR Res Protoc 2014 Apr 8;3(2):e23. [doi: 10.2196/resprot.3045] [Medline: 24713420]
21. Johnston JD, Massey AP, Devaneaux CA. Innovation in weight loss programs: a 3-dimensional virtual-world approach. J Med Internet Res 2012 Sep 20;14(5):e120. [doi: 10.2196/jmir.2254] [Medline: 22995535]
22. Ruggiero L, Moadsiri A, Quinn LT, et al. Diabetes island: preliminary impact of a virtual world self-care educational intervention for african americans with type 2 diabetes. JMIR Serious Games 2014;2(2):e10. [doi: 10.2196/games.3260] [Medline: 25584346]
23. Sullivan DK, Goetz JR, Gibson CA, et al. Improving weight maintenance using virtual reality (second life). J Nutr Educ Behav 2013 May;45(3):264-268. [doi: 10.1016/j.jneb.2012.10.007]
24. Creswell JW, Plano Clark VL. Designing and Conducting Mixed Methods Research, 3rd edition: SAGE Publications, Inc; 2018.
25. Centers for Disease Control and Prevention. Road to health user's guide. : U.S. Department of Health and Human Services, Centers for Disease Control and Prevention; 2022 URL: https://www.cdc.gov/diabetes/professional-info/pdfs/toolkits/User-Guide-508.pdf [accessed 2024-12-19]
26. Centers for Disease Control and Prevention. Road to Health Toolkit (2008 version used for study). URL: https://www.cdc.gov/diabetes/php/toolkits/road-to-health.html [accessed 2024-12-19]
27. ADA/CDC Prediabetes Risk Test. URL: https://www.cdc.gov/diabetes/prevention/pdf/Prediabetes-Risk-Test-Final.pdf [accessed 2025-06-04]
28. U.S. Department of Agriculture. MyPlate. URL: https://www.myplate.gov/ [accessed 2024-12-19]
29. U.S. Food and Drug Administration. How to understand and use the nutrition facts label. URL: https://www.fda.gov/food/new-nutrition-facts-label/how-understand-and-use-nutrition-facts-label [accessed 2024-12-19]
30. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381. [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
31. Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: Building an international community of software platform partners. J Biomed Inform 2019 Jul;95:103208. [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]
32. Qualtrics.: Provo, Utah; 2024. URL: https://www.qualtrics.com [accessed 2025-06-04]
33. Townsend MS, Sylva K, Martin A, Metz D, Wooten-Swanson P. Improving readability of an evaluation tool for low-income clients using visual information processing theories. J Nutr Educ Behav 2008;40(3):181-186. [doi: 10.1016/j.jneb.2007.06.011] [Medline: 18457787]
34. Gans KM, Sundaram SG, McPhillips JB, Hixson ML, Linnan LA, Carleton RA. Rate your plate: An eating pattern assessment and educational tool used at cholesterol screening and education programs. J Nutr Educ 1993 Jan;25(1):29-36. [doi: 10.1016/S0022-3182(12)80186-5]
35. Dedoose version 9.2.006, cloud application for managing, analysing, and presenting qualitative and mixed method research data. 2024. URL: www.dedoose.com [accessed 2025-06-04]
36. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005 Nov;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]

Abbreviations

3D: three-dimensional
CDC: Centers for Disease Control and Prevention
CHW: community health worker
DDT: Division of Diabetes Translation
ILCHWA: Illinois Community Health Worker Association
RTH: Road to Health
RTH-VW: Road to Health Toolkit- virtual world
RYP: Rate Your Plate
SOC: stage of change
T2DM: type 2 Diabetes Mellitus
UIC: University of Illinois at Chicago

XSL•FO
RenderX

VW: Virtual World

Edited by A Mavragani; submitted 08.07.24; peer-reviewed by K Baucom; revised version received 06.03.25; accepted 07.03.25; published 10.06.25.

Please cite as:
Ruggiero L, Quinn L, Castillo A, Monahan C, Boughton Price L, Hernandez W
Community Health Worker Diabetes Prevention Awareness Training in an Immersive Virtual World Environment: Mixed Methods Pilot Study
JMIR Form Res 2025;9:e64051
URL: https://formative.jmir.org/2025/1/e64051
doi:10.2196/64051

© Laurie Ruggiero, Lauretta Quinn, Amparo Castillo, Colleen Monahan, Leticia Boughton Price, Wandy Hernandez. Originally published in JMIR Formative Research (https://formative.jmir.org), 10.6.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Smartphone Ecological Momentary Assessment and Wearable Activity Tracking in Pediatric Depression: Cohort Study

Jimena Unzueta Saavedra[1], BS; Emma A Deaso[1], BS; Margot Austin[1], BS; Laura Cadavid[1], BS; Rachel Kraff[1], BS; Emma E M Knowles[1,2], PhD

[1]Boston Children's Hospital, 300 Longwood Avenue, Boston, MA, United States

[2]Harvard Medical School, 25 Shattuck Street, Boston, MA, United States

Corresponding Author:
Emma E M Knowles, PhD
Boston Children's Hospital, 300 Longwood Avenue, Boston, MA, United States

Abstract

Background: Adolescent depression is a significant public health concern. The presentation of depressive symptoms varies widely among individuals, fluctuating in intensity over time. Ecological momentary assessment (EMA) offers a unique advantage by enhancing ecological validity and reducing recall bias, allowing for a more accurate and nuanced understanding of major depressive disorder (MDD) symptoms. This methodology provides valuable insights into the fluctuating nature of depression, which could inform more personalized and timely interventions.

Objective: This study aims to (1) evaluate the feasibility of collecting smartphone-based EMA data alongside activity and sleep tracking in adolescents with depression; (2) investigate the severity and variability of mood symptoms reported over time; and (3) explore the relationship between mood, activity, and sleep.

Methods: Thirty-six participants (23 with MDD, 13 unaffected controls; 75% [n=27] female, mean age 19.50 y) completed twice-daily EMA check-ins over 2 weeks, complemented by continuous activity and sleep monitoring using FitBit Charge 3 devices. The study examined feasibility, usability of the EMA app, symptom severity and variability, and relationships between mood, activity, and sleep. We applied linear mixed-effects regression to the data to examine relationships between variables.

Results: Participants completed a total of 923 unique check-ins (mean check-ins per participant=25.60). Overall compliance rates were high (91.57%), indicating the approach is highly feasible. MDD participants demonstrated greater symptom severity and variability over time compared with controls (β=34.48, *P*<.001). Individuals with MDD exhibited greater diurnal variation (β=−2.54, *P*<.001) with worse mood in the morning and worse mood than anxiety scores over time (β=−6.93, *P*<.001). Life stress was a significant predictor of more severe EMA scores (β=24.50, *P*<.001). MDD cases exhibited more inconsistent sleep patterns (β=32.14, *P*<.001), shorter total sleep times (β=−94.38, *P*<.001), and a higher frequency of naps (β=14.05, *P*<.001). MDD cases took fewer steps per day (mean 5828.64, SD 6188.85) than controls (mean 7088.47, SD 5378.18) over the course of the study, but this difference was not significant (*P*=.33), and activity levels were not significantly predictive of EMA score (*P*=.75).

Conclusions: This study demonstrates the feasibility of integrating smartphone-based EMA with wearable activity tracking in adolescents with depression. High compliance rates support the practicality of this approach, while EMA data provide valuable insights into the dynamic nature of depressive symptoms, particularly in relation to sleep and life stress. Future studies should validate these findings in larger, more diverse samples. Clinically, EMA and wearable tracking may enhance routine assessments and inform personalized interventions by capturing symptom variability and external influences in real time.

(JMIR Form Res 2025;9:e66187) doi:10.2196/66187

KEYWORDS

depression; ecological momentary assessment; activity; sleep; mood; adolescent

Introduction

Adolescent depression is a significant mental health crisis; 14.7% of the adolescent population reports at least one major depressive episode with severe impairment [1]. This trend predates the COVID-19 pandemic [2] and has continued apace [3]. The impact of adolescent depression is severe; a depressive episode leads to immediate debilitating effects plus long-term consequences [4], for example, impaired academic performance [5] and challenges in forming interpersonal relationships [6]. Depression that begins in adolescence often follows a recurrent pattern [7], persists into adulthood, and is associated with higher levels of anxiety, substance abuse, and impaired functioning in later life [8]. Importantly, depression does not manifest uniformly across individuals; major depressive disorder (MDD) is known for its heterogeneous symptomatology [9]. The course

XSL•FO
RenderX

and severity of depression can vary widely among individuals [10]. Moreover, depressive symptoms can fluctuate within the same individual over time [11,12].

Despite the fluctuating and dynamic nature of MDD, typical clinical assessment methods are cross-sectional. Most standard symptom assessments often require patients to recall how they have felt for the previous 1 - 2 weeks [13]. Recall of mood states in the previous weeks may be inaccurate depending on cognitive style and illness severity [14-18]. Ecological momentary assessment (EMA) bridges this gap. EMA refers to the repeated and brief assessments of the same person over multiple days. EMA enables a granular view of depressive symptoms in real-world settings, overcoming the temporal and contextual limitations of standard assessments [19]. In particular, EMA provides increased ecological validity by decreasing recall bias [20]. It also enables the measurement and modeling of within-person variability since changes can be tracked over time. Thus, this tool facilitates the assessment of symptom fluctuations, offering insights into the dynamic nature of the disorder. In addition, repeated assessment of one or more variables (eg, mood and sleep) allows for the examination of covariation between variables over time [21].

Thanks to growth in technology, EMA can be conveniently completed using the ubiquitous smartphone and can be integrated with wearables, such as activity trackers (eg, FitBit) to monitor sleep and physical activity [19]. Despite the many advantages of EMA and its potential clinical utility [22-24], literature on the topic is limited by sample type, and also by EMA method and schedule intensity [25]. The present manuscript adds to the extant literature in a number of ways. First, most EMA mood studies have been conducted in adults [25,26], while the present study focuses on adolescents with MDD. Second, many studies that focused on adolescents used phone-based EMA [27-35] where interviewers called participants on a schedule and recorded responses, or paper-based EMA [36,37] where participants fill in paper diaries on a schedule, leaving the utility of smartphone-based EMA, which we used in the present study, still mostly unknown [25,38]. To our knowledge, only one study has measured mood over time in depressed patients using smartphone, app-based EMA, which comprised a single mood question administered over the course of 6 days [39]. Third, there is a notable gap in terms of moderate duration studies with high-frequency check-ins, the present study had 2 weeks of twice-daily check-ins. Previous studies have either been much shorter (eg, 4 days) [29] or longer studies but with less frequent check-ins (eg, 8 weeks of check-ins with 4-day intervals) [30,33]. Table 1 shows details of cited studies. More data are needed to demonstrate that momentary assessment of depressive symptoms in adolescents using smartphone-based apps is possible. Additionally, the collected data behave as expected, that is, depressed subjects show greater severity and variability of symptoms than healthy controls, and related measures covary with those symptoms.

The present study had 3 primary goals, all of which contribute to assessing the feasibility of smartphone-based EMA in adolescents with depression. First, we aimed to evaluate whether participants could be retained in the study and engage consistently with smartphone-based EMA and wearable data collection. Second, we assessed whether mood reports captured expected symptom variation, specifically whether adolescents with depression demonstrated greater severity and variability of mood and anxiety symptoms compared with healthy controls, supporting the validity and potential clinical utility of this method. Third, we examined whether participants found the app-based EMA approach acceptable by evaluating self-reported satisfaction and usability of the EMA methods. By addressing these key aspects of feasibility, this study provides critical insights into the viability of smartphone-based EMA for real-time symptom monitoring and personalized intervention development in adolescent depression. As this is a pilot study with a small sample size, the findings should be viewed as preliminary. Future studies with larger, more diverse samples will be necessary to confirm and expand upon the relationships observed here, including potential subgroup-specific patterns within the heterogeneous construct of MDD.

XSL•FO
RenderX

Table . Characteristics of existing studies of ecological momentary assessment (EMA) of mood and associated variables in adolescent samples with and without major depressive disorder (MDD).

Study	Sample	Age range (years)	EMA focus	EMA duration	Check-in frequency	EMA method	Actigraphy
Axelson et al [27]	• 5 controls • 11 MDD • 5 BP	10 - 17	Location, social context, mood, media use, future plans, significant events	5 four-day blocks (Fri-Mon)	12 calls per block	Answer-only cellular phone	No
Silk et al [32]	• 16 controls • 19 MDD	8 - 17	Positive and negative affect, and companions	5 four-day blocks (Fri-Mon)	12 calls per block	Answer-only cellular phone	No
Whalen et al [35]	• 23 controls • 30 MDD	7 - 17	Positive and negative affect, caffeine consumption	5 four-day blocks (Fri-Mon)	12 calls per block	Answer-only cellular phone	No
Mor et al [37]	• 278 healthy individuals	__[a]	Mood, stress, self-focus, social activity	3 days	6 per day	Paper diary	No
Forbes et al [29]	• 28 controls • 15 MDD	8 - 17	Positive and negative affect	4 days	1 per day	Cell phone	No
Silk et al [33]	• 32 controls • 47 MDD	7 - 17	Behavior, emotion, social context	5 four-day blocks (Fri-Mon)	12 calls per block	Answer-only cellular phone	No
Cousins et al [28]	• 23 controls • 23 anxiety • 42 MDD	8 - 16	Positive and negative affect	Two 4 day blocks	12 calls per block	Answer-only cellular phone	Yes
Primack et al [31]	• 60 controls • 46 MDD	7 - 17	Media exposures (eg, internet)	5 four-day blocks (Fri-Mon)	12 calls per block	Answer-only cellular phone	No
Forbes et al [30]	• 31 MDD and anxiety • 23 anxiety • 12 MDD	8 - 16	Positive and negative affect, and companions	4 days	12 calls total	Answer-only cellular phone	No
Waller et al [34]	• 31 controls • 29 MDD	11 - 17	Behavior, emotion, social context	3 five-day blocks (Thurs-Mon)	42 calls total	Answer-only cellular phone	No
Frost et al [36]	• 353 healthy individuals	11 - 18	Feelings and activities	7 days	8 per day	Paper diary	No
Bickham et al [40]	• 125 healthy individuals	12 - 15	Media use	Two weeks	48 assessments	Handheld computer	No
Cushing et al [41]	• 20 healthy individuals	13 - 18	Affect and energy	20 days	4 assessments per day	Cell-phone app on study device	No
Minich and Moreno [39]	• 253 healthy individuals	12 - 17	Mood	6 days	30 assessments total	Text message	No

[a]Not available.

Methods

Procedures

Figure 1 depicts the study procedures. Study procedures were identical across participants regardless of case status. Participants completed the protocol across a mean time span of 7 weeks.

XSL•FO
RenderX

Figure 1. Study protocol. After being consented, the presence (or absence) of major depressive diagnoses (and other Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition [DSM-V] diagnoses) was confirmed using a structured clinical interview. Symptom scales for depression, general psychopathology, anxiety, suicidality and self-injury, anhedonia, and trauma were administered. A smartphone app that enabled ecological momentary assessment (EMA) to take place (LifeData) was set up on the participant's phone. Each participant was given a FitBit, and the device was linked to the participant's smartphone (for data syncing). Participants completed 2 weeks of EMA and actigraphy concurrently. At the end of the 2 weeks, participants returned the FitBit device, and they completed an EMA usability assessment. BPRS: Brief Psychiatric Rating Scale; CTQ: Childhood Trauma Questionnaires; SHAPS: Snaith-Hamilton Pleasure Scale; SITBI: Self-Injurious Thoughts and Behaviors Interview.

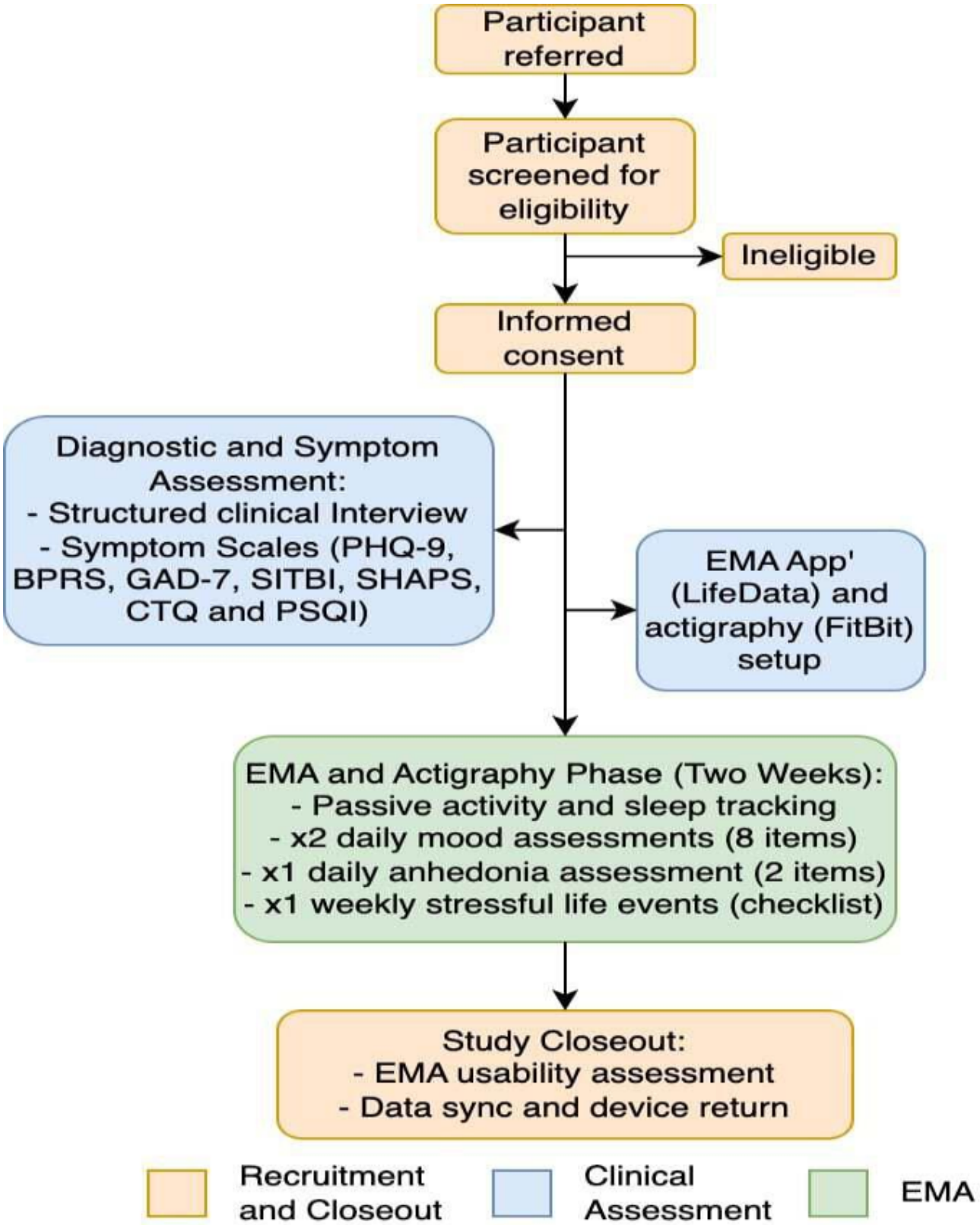

XSL•FO
RenderX

Participant Recruitment

The sample comprised 36 individuals (75% female [n=27], mean age 19.50 y, SD 3.92 y, range 14 - 27 y) recruited from the Boston, MA, area (Table 2). The majority of the sample identified as White (67% [n=29]), with 2 individuals identifying as Hispanic or Latino, 4 individuals identifying as Black or African American, and 1 individual identifying as Asian. Twenty-three individuals had a MDD and were recruited from outpatient psychiatry services at Boston Children's Hospital. Providers requested consent from the participant and their family for the research team to contact them about the study, and research staff followed up to explain the study and, if appropriate, arrange for a consent visit to take place. The control group (N=13) was recruited from the community via flyers and advertisements placed in Boston. Exclusion criteria for individuals with MDD included a severe neurodevelopmental disorder or other impairment that impacts the participant's ability to provide the required information for the study (eg, symptom reports), a substance or medication-induced affective disorder, an affective disorder secondary to a medical condition, and a current Axis I psychotic or bipolar disorder. Exclusion criteria for controls included a severe neurodevelopmental disorder or a current or past psychiatric diagnosis as defined by DSM-5 (Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition) criteria.

XSL•FO
RenderX

Table . Baseline characteristics of the depressed and control groups.

	Depressed	Controls
Age		
Mean (SD)	17.91 (2.98)	22.31 (3.90)
Range	14 - 26	14 - 27
Sex		
Female, n (%)	15 (65)	11 (92)
Parental education, n (%)		
Graduate professional training (graduate degree)	12 (53)	—[a]
Partial college or specialized training	5 (20)	—
Standard college or university graduation	6 (27)	—
Parental occupation, n (%)		
Higher executives, proprietors of large businesses, major professionals	6 (27)	—
Administrators, lesser professionals, proprietors of medium-sized businesses	8 (36)	—
Small business owners, farm owners, managers, minor professionals	6 (27)	—
Technicians, semi-professionals, small business owners (US $50,000-US $70,000)	2 (10)	—
Clinical assessments		
PHQ-9[b], mean (SD)	9.05 (5.81)	1.69 (1.32)
GAD-7[c], mean (SD)	6.65 (6.34)	2.31 (2.72)
BPRS[d], mean (SD)	32.60 (5.97)	26.20 (2.45)
PSQI[e], mean (SD)	8.67 (3.31)	5.27 (3.29)
Medications		
Benzodiazepine	1	0
Betablocker	1	0
Birth Control	2	7
IUD[f]	0	2
Cognitive Enhancer	3	0
Anticonvulsant	2	0
Antidepressant (SSRI)	12	1
Antidepressant (NDRI)	0	1
Antidepressant (Tricyclic)	1	0
Atypical Antipsychotic	3	0
Stimulant	3	1
Depression diagnoses		
Major depressive disorder, lifetime	2	0
Major depressive disorder, recurrent	11	0
Major depressive disorder, single episode, remission	2	0
Other specified depressive disorder (insufficient symptoms), current	1	0
Past major depressive disorder	2	0

XSL•FO
RenderX

	Depressed	Controls
Persistent depressive disorder	5	0
Comorbidities		
Eating disorders		
Anorexia nervosa, in remission	1	0
Bulimia nervosa, in remission	1	0
Anxiety disorders		
GAD	2	0
OCD[g]	1	0
PTSD[h]	1	0
Social anxiety disorder	4	1
Neurodevelopmental disorders		
ADHD	6	0

[a]Not applicable.
[b]PHQ: patient health questionnaire.
[c]GAD: generalized anxiety disorder.
[d]BPRS: Brief Psychiatric Rating Scale.
[e]PSQI: Pittsburgh Sleep Quality Index.
[f]IUD: intrauterine device.
[g]OCD: obsessive-compulsive disorder.
[h]PTSD: posttraumatic stress disorder.

Ethical Considerations

This study was reviewed and approved by the Boston Children's Hospital Institutional Review Board (BCH IRB-P00031981). The study involved the collection and analysis of data from human subjects, and all procedures were conducted in accordance with the IRB-approved protocol. All participants provided informed consent/assent using forms approved by the institutional review board at Boston Children's Hospital. This ensured that participants were fully informed about the study's purpose, procedures, potential risks, and benefits before agreeing to participate. To maintain participant confidentiality, all collected data were stored in secure databases and deidentified prior to analysis. All data included in the present manuscript or any supplementary materials in Multimedia Appendix 1 are completely anonymous. There was no cost or fee associated with study participation. Participants were compensated for their time and effort in completing the study procedures. They were eligible to receive up to US $144, with payments provided via ClinCard, a secure electronic payment system. Payment breakdown was as follows: US $20 for Structured Clinical Interview for DSM-5; US $20 for questionnaire completion; US $14 (US $1 per day) for FitBit wearing; US $14 for FitBit return; US $14 for usability questionnaire; and US $56 (US $2 per check-in for EMA). Participants were paid for all aspects of the study that they completed, including individual EMA check-ins. No identifiable images of individual participants are included in this manuscript or supplementary materials in Multimedia Appendix 1.

Diagnostic and Symptom Assessment

The presence or absence of major depression diagnoses (and comorbid disorders) was confirmed using the Structured Clinical Interview for DSM-5 [42] (with additional modules from the Kiddie Schedule for Affective Disorders and Schizophrenia interview [43] if<18 years of age) in order to confirm past or current MDD (and other DSM diagnoses). Depressive symptoms were assessed at study intake using the patient health questionnaire (PHQ-9) [44]. Participants were also administered the Brief Psychiatric Rating Scale [45], generalized anxiety disorder (GAD-7) [46], and the Pittsburgh Sleep Quality Index [47]. Parents of probands were administered the Hollingshead Index of Socioeconomic Status [48], and 25 parents agreed to complete the scale.

Ecological Momentary Assessment (EMA)

EMA data were recorded using LifeData [49]. LifeData is a smartphone-based EMA program that runs on both Android and iPhone smartphones. Check-ins are delivered via the LifeData mobile app RealLifeExp. EMA lasted 2 weeks, during this time participants received a push notification twice per day (6 AM and 4 PM), reminding them to complete a check-in. Notifications were sent 4 times (at 1 h intervals) per check-in, allowing each participant a 4-hour window to complete the check-in (6-10 AM and 4-10 PM).

For each check-in, participants were asked the same 8 questions. Participants rated the items using a slider, which translated to a score (0 - 100): "I am sad."; "I feel bothered by every little thing."; "I have no interest in things I would usually enjoy (eg, food, TV, games, spending time with friends/family)."; "I do not have enough energy to get going."; "I have no appetite or

XSL•FO
RenderX

feel much hungrier than usual."; "I feel physically tense and/or jittery."; "I am nervous, anxious or on edge."; and "I can't stop worrying." Items were designed to be reflective of established measures of depression (PHQ-9) [44] and anxiety (GAD-7) [46], with the adaptation that they could be administered multiple times per day (Table S1 in Multimedia Appendix 1 for an explanation of each item). For a subset of statistical models (see below), the first 5 items were classed as measuring mood and the rest as measuring anxiety. Once per week, during an afternoon check-in, participants completed a checklist of stressful life events. The question reads "Did any of the following happen to you since the last time you completed one of these surveys? Check all that apply". The participant selects as many of the following options as appropriate: "I argued with a friend or family member."; "I was not allowed to do something I wanted to do."; "I got a bad grade in school."; "My parents have been arguing a lot."; "Somebody in my family got a serious illness."; "Someone in my family was arrested."; "Somebody teased or threatened me."; "I teased somebody else."; "I did something that made me feel embarrassed."; "Someone commented negatively on the way I look."; "I was excluded from a group event."; and "I got disciplined or suspended from school."

Physical Activity (Steps) and Sleep Features

Participants were issued a FitBit Charge 3 and were instructed to wear it for the duration of the EMA portion of the study. Twenty-two individuals (11 controls (mean age 22.90, SD 3.27 y, 10 females) and 11 MDD cases (mean age 18.20, SD 2.86 y, 7 females) wore the FitBit with sufficient regularity for data to be analyzed. Data were collected on daily activity (number of steps) and sleep.

Fitbit data were obtained via the Fitbit API, which provides preprocessed JSON files reflecting proprietary algorithms for step counting and sleep [50]. For sleep, each participant had one JSON file that spanned the duration of the study, which included a sleep log with timestamps for sleep onset and offset and a breakdown of sleep stages (eg, light, rapid eye movement [REM], deep). For steps, each JSON file contained minute-level step count data linked to a timestamp. Data in the JSON files were converted to csv using a combination of standard and pandas [51] libraries in PyCharm [52] (Python version 3.9; Python Software Foundation) and rearranged for analysis using tidyr [53] and dplyr [54] in R [55]. Following others' work [56] in the field, we extracted numerous sleep features from the data collected using the Fitbit device. These are related to sleep architecture, quality, and stability (Table S2 in Multimedia Appendix 1 contains details or individual metrics). A nap was defined as time spent in bed lasting<180 minutes, FitBit does not calculate deep, light, and REM sleep during these shorter sleeps. Therefore, naps were included in analyses focused on total sleep time (TST) and time in bed (TIB) but were excluded from analyses focused on sleep stage. Sleep quality metrics were calculated per all logged sleeps per individual, and sleep stability metrics were derived across all logged sleeps. Steps that were recorded during a documented period of sleep were recoded to 0 as they were likely due to movement in bed rather than steps.

Usability and Tolerability Questionnaire

At the end of participation, all participants completed a short questionnaire that evaluated their experience using the EMA app (eg, if the app was easy to use; see in Multimedia Appendix 1 for a list of questions). All but one of the statements were rated from 1 to 7, with 1 being strongly disagree, 4 neutral, and 7 strongly agree. Responses were recorded as disagree (1-3) and agree (5-7). Analysis included frequencies and proportions of responses per item. One item, relating to where participants used the app, included a free text response (unique responses are included in the Multimedia Appendix 1).

Statistical Analyses

All analyses were conducted in R (version 4.2.1; R Foundation for Statistical Computing) [55].

How Does Depressive Symptomatology Vary Over Time?

In line with others' EMA work [57], we opted to examine variability in depressive symptoms both visually and statistically. Time series data were examined using standardized (group-mean-centered) and raw data. Group-mean-centered plots depict the relative variability over time across the sample. Before plotting data were group-mean centered (by subtracting each participant's mean from each individual score), which results in standardization of the figure such that units are in units of deviation from each individual's average (set to 0). Per subject, raw score time-series plots depict the fine-grained differences between each individual's data. Data were plotted using ggplot2 [58].

We used several statistics to quantify variability, including the root mean square of successive differences (RMSSD) [59], coefficient of variation (CoV) [60], and intraclass correlations (ICC). The RMSSD reflects variability over time. The CoV reflects the dispersion of the data and is useful for comparing variability when means differ (eg, in cases vs controls). The ICC reflects the proportion of variance attributable to between-person variability, thus, 1-ICC indicates the proportion of variability attributable to within-person variability. Variability statistics were calculated using a combination of base R and the psych [61] package.

Predictors of Depression Severity and Variability Over Time

The EMA data were nonindependent, for example, we might expect data collected within the same person to be more strongly related than those collected from 2 different people. This structure is sometimes referred to as multilevel; observations are nested within item type (mood or anxiety), which are nested within the time of day (AM or PM) and day, which are nested within the participant. Therefore, when evaluating the impact of various predictors on EMA scores; we applied linear mixed-effects regression to the data using the package lme4 [62]. Effect sizes for fixed effects ($\eta p2$) were calculated using the effectsize package [63], and were interpreted as small effect: $\eta p2 \approx 0.01$, medium effect: $\eta p2 \approx 0.06$, large effect: $\eta p2 \approx 0.14$. Similarly, observations for sleep architecture and quality were nested within a day (where some individuals logged sleep with

XSL•FO
RenderX

onset times within the same day), within the sleep stage, and within the participant. Multilevel modeling is robust to missing data, and therefore we did not impute the small amount of missing data that was present in the EMA, of 1008 expected check-ins, 923 were observed (overall compliance=92%). Prior to testing the effect of various predictors on depression, we tested an intercept-only, or unconditioned model to evaluate the need for a random effect. Significant random effects (determined using the lmerTest [64] package) were retained in subsequent models. Given the different distributions of age and sex in case and control groups, these variables were included as covariates in all models.

First, we asked whether individuals with a MDD diagnosis demonstrated greater symptom severity (indexed using EMA score) and variability than controls. We also evaluated the impact of sex and age on severity, as well as fixed effects of Time of Day and Item Type. Second, we examined whether anxiety and mood covaried over time and whether anxiety or mood at the preceding check-in significantly predicted mood at the following one. Third, we asked whether life stress was associated with increased symptom severity. Fourth, we examined whether activity levels (number of steps) were associated with MDD diagnosis or symptom severity. Fifth, we assessed whether number of minutes asleep, sleep onset, or sleep-waking was associated with MDD diagnosis or symptom severity. Sixth, we examined the results of the usability questionnaire.

Results

Overview

The total unique check-ins completed by participants=923 (mean check-ins per participant 25.60, SD 2.78). Assessments were completed across a total of 486 unique days (mean days per participant 13.50, SD 1.25), and yielded 1.90 mean check-ins per participant per day. For the 2 required check-ins per day compliance rate of 91.57%, cases demonstrated slightly lower compliance (89.94%) than controls. Across individual EMA items, this yielded 7336 data points for regression modeling.

How Does Depressive Symptomatology Vary Over Time?

Table 3 shows the descriptive and variability statistics for the sum score of all EMA items as well as for the individual items. In terms of reported symptom severity, MDD cases scored higher than controls on all items. Examination of variability statistics (Table 3) indicates that there was considerable variability in both the sum score and the individual items. The RMSSD and CoV suggest greater variability in MDD cases than in controls. Examination of the ICC across items suggests that ≈50% of the variance in item-level responses was due to within-person variance in MDD cases.

XSL•FO
RenderX

Table . Descriptive and variability statistics for individual ecological momentary assessment items in depressed and control groups.

Item	Group	Total responses	Non-zero responses, n (%)	Mean (SD)	Range	Skewness	RMSSD[a]	CoV[b]	ICC[c]
Anxious	Ctrl	340	44.41	9.77 (8.93)	0 - 90	1.57	11.93	0.91	0.46
Anxious	MDD	572	76.05	36.1 (19.91)	0 - 100	0.4	24.59	0.55	0.55
Appetite	Ctrl	342	17.54	4.23 (7.41)	0 - 80	3	9.72	1.75	0.17
Appetite	MDD	575	76	41.65 (20.38)	0 - 100	0.03	25.89	0.49	0.58
Bothered	Ctrl	346	30.35	4.94 (5.89)	0 - 80	2.27	7.53	1.19	0.35
Bothered	MDD	576	68.58	35.97 (20.07)	0 - 100	0.54	24.55	0.56	0.56
Energy	Ctrl	342	39.77	8.43 (9.91)	0 - 80	1.72	14.03	1.18	0.39
Energy	MDD	575	80.52	46.04 (22.7)	0 - 100	0.29	30.75	0.49	0.5
Interest	Ctrl	344	13.95	2.17 (3.11)	0 - 60	2.84	4.37	1.43	0.45
Interest	MDD	575	71.13	32.14 (18.38)	0 - 100	0.6	23.88	0.57	0.5
Sad	Ctrl	347	27.38	4.38 (5.52)	0 - 50	2.72	7.26	1.26	0.31
Sad	MDD	576	65.80	28.71 (17.01)	0 - 100	1.15	22.41	0.59	0.59
Tense	Ctrl	341	34.90	6.98 (7.82)	0 - 80	1.87	10.13	1.12	0.36
Tense	MDD	574	65.85	30.73 (20.22)	0 - 100	0.72	24.03	0.66	0.54
Worry	Ctrl	340	22.35	5.5 (5.73)	0 - 100	2.48	6.43	1.04	0.5
Worry	MDD	571	69	31.27 (18.05)	0 - 100	0.77	21.3	0.58	0.57
Sum score	Ctrl	340	65.29	46.58 (34.95)	0 - 400	1.64	44.39	0.75	0.51
Sum score	MDD	571	93.35	282.52 (99.06)	0 - 800	0.38	115.58	0.35	0.7

[a]RMSSD: root mean square of successive differences.

[b]CoV: coefficient of variation.

[c]ICC: intraclass correlation.

Figure 2 shows the group-mean centered time series plot of the sum score. At the group level, scores fluctuated in a saw-tooth pattern, with repeated rises and falls but no clear linear trend of improvement or worsening over time. This pattern reflects short-term variability in EMA data and highlights overall fluctuations in symptom severity across the sample. Item-level data followed a similar pattern (Figure S1 in Multimedia Appendix 1); in MDD cases, variability in symptom reporting did not appear to vary as a function of item. Controls appear to report a greater degree of variability to the irritability and anhedonic items (Bothered and Interest) than they did on mood or anxiety symptoms.

XSL•FO
RenderX

Figure 2. Time-series plot of the group-mean centered ecological momentary assessment (EMA) data (sum score across all 8 items). Colored lines represent individuals with a major depressive disorder (MDD) diagnosis, gray lines represent controls.

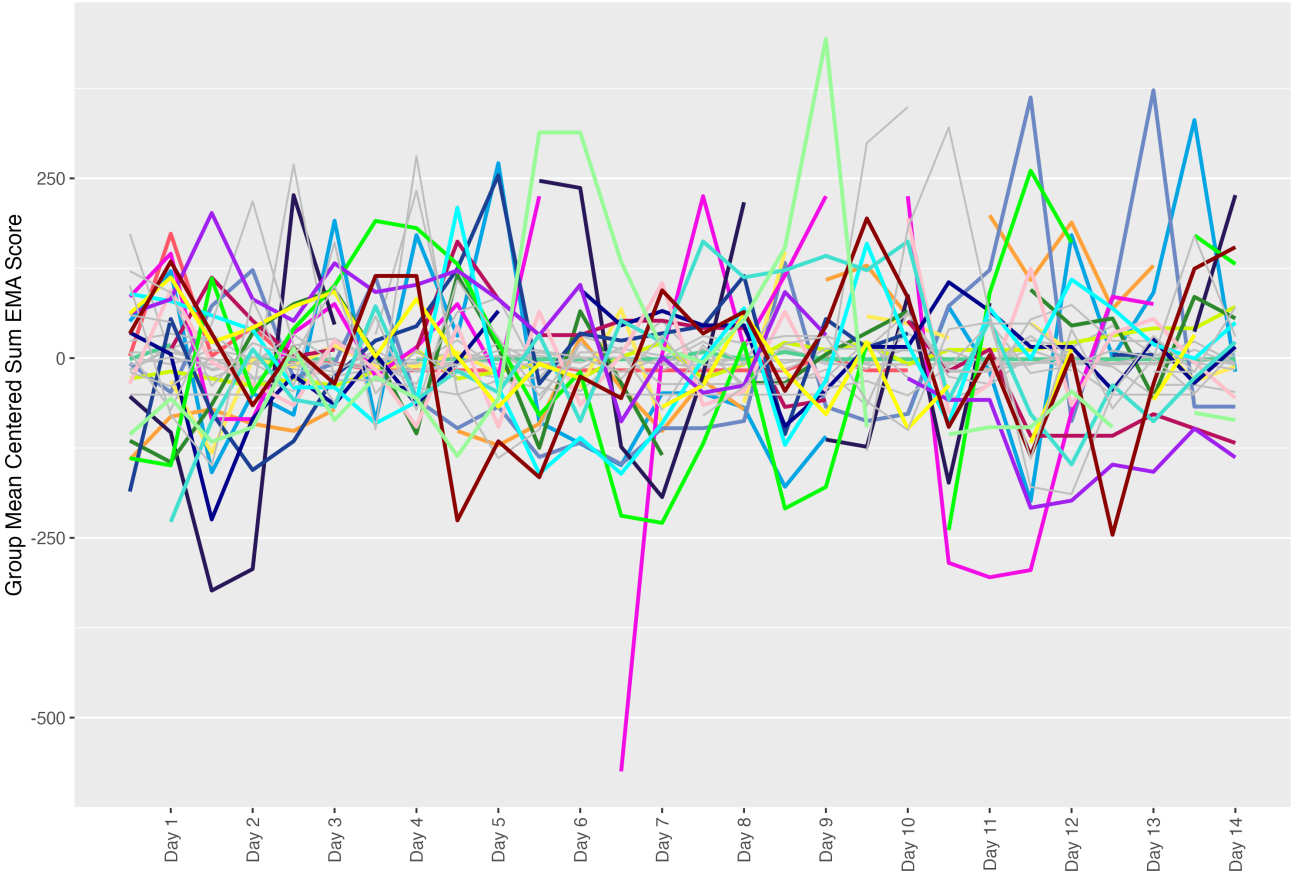

At the individual level, we observe heterogeneity (Figure 3) where some individuals with a MDD diagnosis demonstrate worsening throughout the course of the study (eg, 30,001, 30,015, 30,028, 30,030, 30,044, and 30,048), and others appear to improve (eg, 30,035, 30,036, and 30,047). Interestingly, not all controls demonstrated floor effects. While controls reported more zero responses than cases (Table 3) some endorsed minimal symptoms (eg, 30,011, 30,033, 30,038, 30,046, and 30,057) throughout the course of the study; admittedly these symptoms would likely amount to sub-threshold symptoms for a MDD diagnosis but this suggests that healthy controls can demonstrate variation in depressive symptomatology. Three MDD cases (30,003, 30,012, and 30,019) demonstrated minimal symptoms; diagnoses for these individuals were in line with low symptoms (Other Specified Depressive Disorder [Insufficient Symptoms], Recurrent MDD, and Past MDD Single Episode, respectively).

XSL•FO
RenderX

Figure 3. Time series plots of the raw ecological momentary assessment (EMA) data (sum score across all 8 items). The black line represents the reported score, dashed lines are controls, and solid lines are those with an major depressive disorder (MDD) diagnosis. Mean score (per individual) is represented by the red horizontal line, and blue dashed lines represent SD of 1.

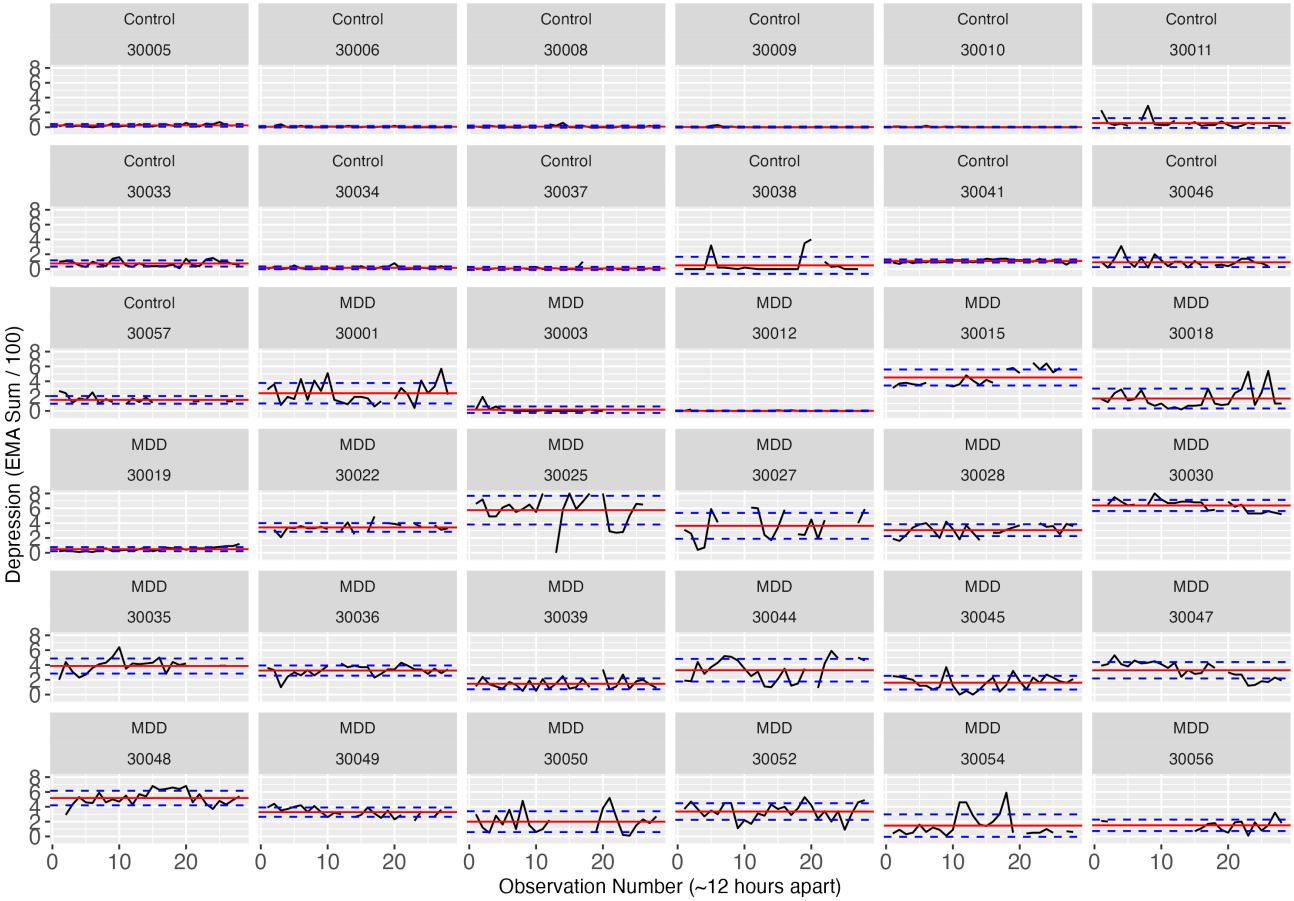

Predictors of Depression Severity and Variability Over Time

The unconditioned model indicated significant between-subject nested effects of Participant, Day, and Item Type but not Time of Day (Table S3 in Multimedia Appendix 1). All significant random effects were retained in subsequent modeling.

Effect of MDD on EMA Score Severity Over Time

We examined the impact of MDD diagnosis on summed EMA score using a linear mixed model (Rmarginal2=0.26, Rconditional2=0.65) with fixed effects of MDD (plus age, sex, item type, and time of day). The model indicated significant main effects of MDD (ηp2=0.41, β=34.48, 95% CI 20.22 to 48.74, *P*<.001), sex (ηp2=0.16, β=16.81, 95% CI 3.55 to 30.07, *P*=.01), and a small but significant effects of item type (ηp2=0.01, β=−1.69, 95% CI –2.27 to –0.55, *P*=.01) and time of day (ηp2=0.002, β=−1.41, 95% CI –2.27 to –0.55, *P*<.001). The effect of age (*P*=.85) was not significant.

We fit an additional linear mixed model (Rmarginal2=0.27, Rconditional2=0.65) with MDD*time of day and MDD*item type interactions to establish whether there was greater diurnal variation in mood in cases than controls and whether there was a greater effect of item type in cases than in controls. There was a significant MDD*time of day interaction (ηp2=0.001, β=−2.54, 95% CI –4.32 to –0.76, *P*<.001; Figure S2 in Multimedia Appendix 1) where cases exhibited diurnal mood variation (higher scores in the morning) but controls did not, ηp2 indicates that the effect was minimal. There was also a significant MDD*item type interaction (ηp2=0.001, β=−6.93, 95% CI –9.64 to –4.21, *P*<.001; Figure S3 in Multimedia Appendix 1) where cases demonstrated worse mood than anxiety scores and controls demonstrated the opposite effect; however, ηp2 indicates that the effect was minimal.

PHQ-9 score at baseline (ηp2=0.58, β=2.74, 95% 1.92 to 3.57, *P*<.001) significantly predicted EMA scores over the course of the study, suggesting that the items we selected for EMA have good convergent validity with an established measure of depression.

Cross-Sectional Covariation of Mood and Anxiety Over Time

We examined cross-sectional relationships between mood and anxiety over time using a linear mixed model (Rmarginal2=0.33, Rconditional2=0.87) of summed EMA mood items with fixed effects of summed EMA anxiety items and MDD (plus age and sex). This model indicated significant covariation between mood and anxiety over time (Figure S4 in Multimedia Appendix 1 shows raw data per participant), main effects of anxiety (ηp2=0.25, β=0.64, 95% CI 0.56 to 0.71, *P*<.001) and MDD (ηp2=0.45, β=132.14, 95% CI 81.24 to 183.05, *P*<.001) were significant, but neither age (*P*=.22) nor sex were (*P*=.39).

XSL•FO
RenderX

Impact of Preceding Anxiety Score on Mood

We examined whether prior anxiety ratings predicted later mood scores using a linear mixed model with the same predictors as above, plus anxiety score from the preceding check-in (where *t*=time, T-1) as a fixed effect (Rmarginal2=0.63, Rconditional2=0.89). The preceding anxiety score was a nominally significant fixed effect, but the size of the effect was small (ηp2=0.01, β=0.11, 95% CI 0.04 to 0.19, *P*<.001).

Impact of Preceding Mood on Subsequent Mood Score

We examined whether prior mood score predicted later mood score using a linear mixed model (Rmarginal2=0.46, Rconditional2=0.88) with T-1, T-2, and T-3 mood scores as a fixed effect in addition to MDD (plus age and sex). The immediately preceding mood score (T-1; ηp2=0.02, β=0.11, 95% CI 0.04 to 0.18, *P*<.001) was a significant predictor of subsequent mood score, as was mood score that was separated by one check-in (T-2; ηp2=0.03, β=0.16, 95% CI 0.09 to 0.23, *P*<.001) and 2 (T-3; ηp2=0.02, β=0.11, 95% CI 0.04 to 0.18, *P*<.001) check-ins. MDD (ηp2=0.53) remained a significant predictor of mood score, age (*P*=.66), and sex (*P*=.55) were not significant predictors in this model.

Effect of Life Stress on EMA Score Severity

Both MDD cases and controls reported life stress (Table S4 in Multimedia Appendix 1). Most stressful life events in control are related to arguments with friends or family and embarrassment. These stressors were also common in MDD cases, but they also endorsed stress related to instances of not being allowed to do what they wanted, being teased (and teasing others), or having negative comments made about their appearance. MDD cases reported 91 stressful life events while controls reported 35. Because controls were older than MDD cases and because the type of events included in the checklist were focused on younger individuals, differences in the frequency of life stress are likely not meaningful. A linear model comparing the number of stressful life events between groups while covarying for age and sex indicated a nonsignificant effect of group (*P*=.43). Notwithstanding this confound, looking at the entire sample EMA scores were higher in individuals that reported life stress, this was true in both controls and MDD cases (Figure S5 in Multimedia Appendix 1). We examined cross-sectional and lagged relationships between self-reported stress and summed EMA scores across each week (stress was reported once per week) using linear mixed models, including MDD, age, and sex as predictors. Concurrent self-reported stress, occurring in the same week, significantly predicted EMA scores (ηp2=0.16, β=24.50, 95% CI 6.07 to 42.93, *P*<.001), whereas lagged stress (T-1; ηp2=0.07, β=6.15, 95% CI –1.52 to 13.82, *P*=.28) did not, suggesting an immediate rather than cumulative impact of stress on mood and anxiety. MDD (ηp2=0.16, *P*<.001) and sex (ηp2=0.14, *P*<.001) remained significant predictors, but age (*P*=.48) was not a significant predictor.

Relationship Between MDD and Mood and Activity Levels

Number of steps taken per day (and across the course of the study) ranged widely between participants (Figure S6 in Multimedia Appendix 1). MDD cases took fewer steps per day (mean 5828.64, SD 6188.85) than controls (mean 7088.47, SD 5378.18) over the course of the study. However, MDD was not a significant predictor of steps taken (ηp2=0.18, *P*=.33). We also examined cross-sectional relationships between EMA summed score and steps taken each day over time using a linear mixed model, steps were not a significant predictor of EMA score (ηp2=−0.01, *P*=.75). Thus, activity, indexed using number of steps taken per day, and mood were not significantly related in the present study.

Relationship Between MDD and Mood and Sleep

In both cases and controls, sleep onset was most commonly between 8 PM and 4 AM (Figure S7 in Multimedia Appendix 1), and sleep offset was most commonly between 5 AM and 11 AM (Figure S8 in Multimedia Appendix 1), though there was more spread in sleep offset hours in cases. Descriptive statistics of sleep metrics are shown in Table 4.

XSL•FO
RenderX

Table . Descriptive statistics of sleep metrics in depressed and control groups.

	Control	MDD[a]
Nights, n	160	232
Individuals, n	11	11
Total sleep time (TST), mean (SD)	388.62 (108.56)	277.19 (163.31)
Time in bed (TIB), mean (SD)	439.84 (124.20)	315.11 (186.48)
Time awake in bed (TAB), mean (SD)	51.12 (20.95)	37.41 (27.57)
Deep sleep		
Mean (SD)	79.24 (23.69)	66.47 (27.90)
Percentage	17	15
Light sleep		
Mean (SD)	264.05 (55.20)	250.49 (82.77)
Percentage	57	58
REM[b] sleep		
Mean (SD)	94.80 (34.44)	83.43 (39.12)
Percentage	20	19
Efficiency (TST/TIB)	0.89	0.89
5 min wakes	390	370
Hypersomnia	2	5
Number of naps	13	88
Coefficient of variation (CoV) of TST	27.93	58.92
Mode onset hour	23:00	23:00
SD onset time (minutes from 12 PM)	123.54	202.99
Mode offset hour	7:00	06:00
SD offset time (minutes from 12 PM)	113.48	166.14

[a]MDD: major depressive disorder

[b]REM: rapid eye movement.

In terms of sleep staging, an inspection of Table 4 shows that the amount of deep, light, and REM sleep between cases and controls did not differ, nor did sleep efficiency (how much TIB was spent asleep). However, there were differences in (1) TST, (2) number of naps, and (3) variability in sleep length. We tested differences between cases and controls using linear mixed models, and age and sex were included as covariates in all models. Cases spent significantly less time asleep (TST; $\eta p2=0.20$, $\beta=-94.38$, 95% CI –178.62 to –10.14, $P<.001$) and TIB ($\eta p2=0.18$, $\beta=-103.39$, 95% CI –199.85 to –6.94, $P<.001$). Cases took significantly more naps ($\beta=14.05$, 95% CI 2.32 to 85.02, $P<.001$). Cases had significantly more variability in their TST (the length of each sleep was more inconsistent than that taken by controls) as indexed by the CoV of TST ($\eta p2=0.53$, $\beta=32.14$, 95% CI 16.04 to 48.23, $P<.001$). We examined the relationship between TST and EMA data using a linear mixed model. The effect of TST on EMA score the following day was not significant ($P=.69$).

Did Participants Find Using the App to Report on Their Mood Acceptable?

Figure 4 shows the results of the usability questionnaire. Inspection of Figure 4 shows that the majority agreed that the EMA app was easy to use (96%) and that they were comfortable using the app in social settings (77%); participants reported using the app in a wide variety of places (Table S5 in Multimedia Appendix 1 contains a complete list of unique responses). A total of 100% of individuals found the app an acceptable way to report on their mood, 89% of them said they felt the app improved their access to research participation, and 63% of individuals agreed that they would like to use an app like this to communicate with their health care provider.

XSL•FO
RenderX

Figure 4. Usability data regarding the ecological momentary assessment (EMA) app. The majority of participants agreed that the app was ”easy” to use, that they liked the “interface” of the app, that they felt comfortable using the app in “social” settings, that they were “satisfied” with the app, and that the app provided them with an “acceptable” way to report on their mood. A small minority of participants disagreed that the app increased their “access” to research participation (N=1) and that they would like to use a similar app to communicate with their “healthcare” provider (N=4), but the majority were in agreement. MDD: major depressive disorder.

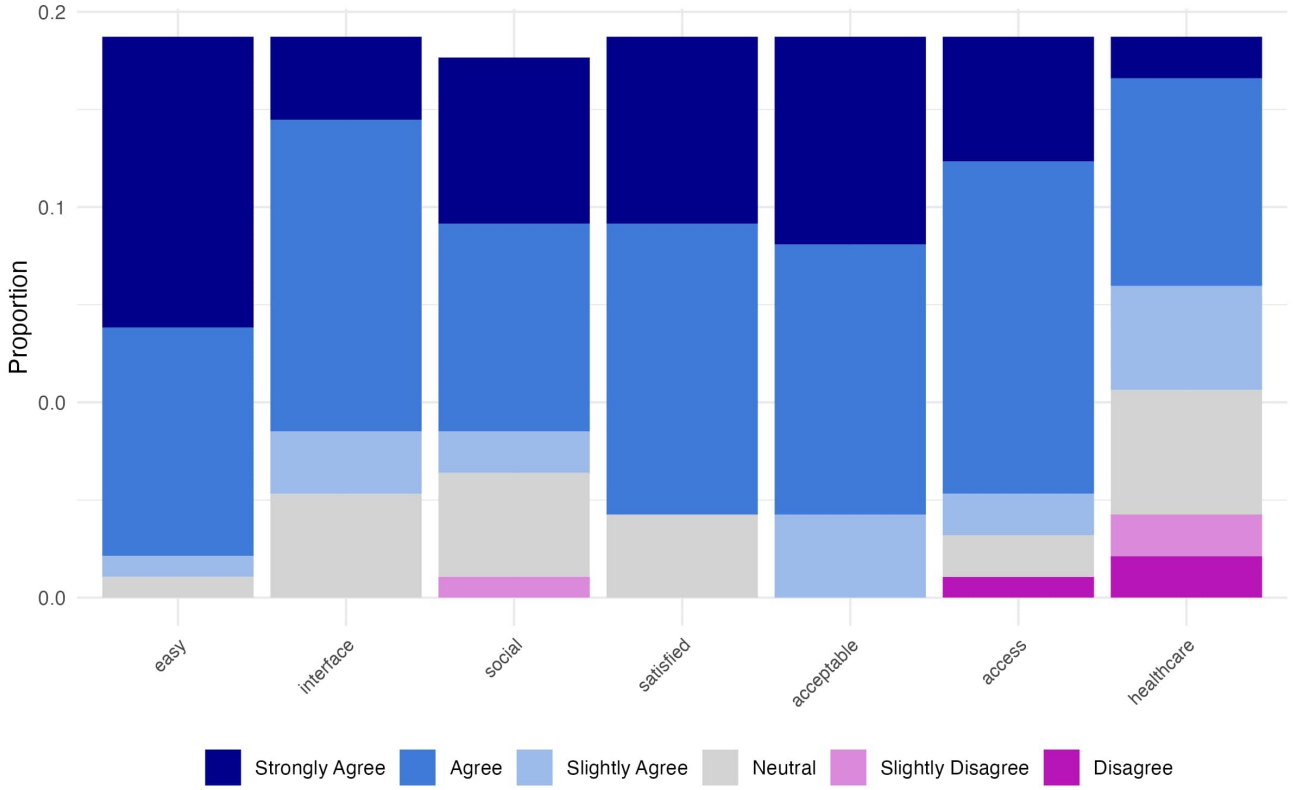

Discussion

The present study comprehensively assessed the feasibility of smartphone-based EMA in adolescents with depression by (1) evaluating participant retention and engagement, (2) examining whether mood reports captured expected symptom variation, and (3) determining the acceptability of the app-based approach. Together, these findings demonstrate that smartphone-based EMA is a feasible, viable, and effective method for real-time symptom monitoring in adolescents with depression.

Engagement with EMA was high, with exceptional compliance rates across participants. This suggests that adolescents were able and willing to complete frequent mood assessments over an extended period, reinforcing the feasibility of implementing this method in both research and clinical settings. The usability data indicate that participants found the app easy and satisfying to use, that they used the app wherever they were, including in social settings, and that most participants would be amenable to using such an app in a clinical setting. These responses indicate that adolescents find smartphone-based apps for mood reporting convenient and further suggest that approaches using them, whether that be in the context of research or in the clinic, are likely to be well tolerated.

Beyond high compliance and self-reported acceptability, feasibility was further supported by the clinical distinctions captured in the collected data. Adolescents with depression reported significantly more severe and variable symptoms than healthy controls, reinforcing EMA’s ability to differentiate between clinical and nonclinical populations. This supports the discriminant validity of our approach, further reinforcing its utility and feasibility as a real-time assessment tool. The EMA scores were significantly predicted by baseline PHQ-9 scores, supporting its criterion validity against an established depression measure. While this alignment was expected given that EMA items were adapted from the PHQ-9 and GAD-7, it is notable that these adaptations retained the ability of those items to reflect depression severity while also being suitable for repeated, momentary assessment. Furthermore, individual-level trajectories revealed that while some participants experienced worsening symptoms, others improved. These findings underscore EMA’s ability to capture dynamic mood changes [9,22-24,65] that would be difficult to assess with retrospective measures [14-19]. Notably, some healthy controls also reported transient depressive symptoms, underscoring the importance of including control groups in EMA studies and highlighting the potential for identifying individuals with emerging depressive episodes. Interestingly, increased life stress was associated with greater mood and anxiety symptom severity across both groups. The effect of stress on EMA scores was observed concurrently, but not lagged, indicating an immediate rather than cumulative effect of stress on mood and anxiety symptoms.

Activity data revealed that MDD cases spent less time asleep and in bed but took more naps than controls, suggesting that their sleep was more fragmented and variable than in controls. These findings are consistent with previous research showing an association between sleep disturbances and depression and anxiety [66,67], particularly short sleep duration and symptoms

XSL•FO
RenderX

of depressed mood [68] and increased napping in adults [69,70] and adolescents [71] with depression. Although we did not find a significant relationship between daily mood variations and the previous night's sleep or daily physical activity levels, our study supports the notion that sleep disturbances are a critical component of adolescent depression [72].

The study should also be viewed in the context of some important limitations. First, our sample size limits the generalizability of our results. The sample size is small, and so future replication of our main findings is key. Second, the racial homogeneity and relatively high Socioeconomic Status of our sample, with 67% of participants identifying as White, may limit the generalizability of our findings to more diverse populations. Third, there is an age disparity between the studied groups, with the controls being significantly older on average than the MDD cases. This could have introduced an age-related bias, particularly for age-specific assessments like the life stress questionnaire. Fourth, while fitbit is a cost-effective way to collect sleep and activity information, it may overestimate sleep time (particularly sleep staging estimates [73,74] and activity levels [75] compared with gold standard methods (eg, polysomnography for sleep). We did not find differences between cases and controls in terms of number of steps taken or in terms of sleep staging; these might be true findings, or the lack of difference might be clouded by the measurement error associated with Fitbit devices. Future research should aim to include a larger and more age-matched sample using alternate devices to mitigate these limitations.

When considering the clinical applications of EMA, it is important to acknowledge that this study was conducted as a paid research project. As a result, the high compliance rates observed may have been influenced by financial compensation, which does not necessarily translate to similar real-world clinical settings. Additionally, the implementation of EMA in clinical practice presents several challenges that must be addressed. These include logistical barriers, such as ensuring access to the necessary technology, as well as concerns about privacy and data confidentiality, particularly when working with vulnerable populations like children.

The findings of this study have strong implications for both research and, potentially, clinical practice. From a research perspective, this study is one of the first to explore the integration of smartphone-based EMA and wearable activity tracking in adolescents with depression. The findings highlight the potential of these technologies to capture real-time data and provide valuable insights into the dynamic nature of depressive symptoms. Future studies should explore the use of EMA and activity-tracking methods in a larger and more diverse population to validate our findings and ultimately expand this novel method to clinical practices. In terms of clinical practice, the demonstrated feasibility of EMA and activity tracking in adolescents suggests that these methods could be integrated into routine clinical assessments to provide a more comprehensive and dynamic picture of symptomatology. Moreover, the development of personalized intervention strategies that take into account individual patterns of symptom fluctuations could lead to more effective management of adolescent depression [57]. A more comprehensive understanding of the external factors influencing symptom variability would provide insight into this relationship, particularly by looking at external factors such as environmental stressors and social interactions.

Acknowledgments

This work was supported by a Tommy Fuss Independent Investigator Award (grant Number: 92188)

Data Availability

The data supporting the findings of this study are available upon request. Interested researchers may contact the corresponding author at emma.knowles@childrens.harvard.edu for access.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional methodological details and supplementary materials related to the main analyses.
[DOCX File, 2171 KB - formative_v9i1e66187_app1.docx]

References

1. Major depression. National Institute of Mental Health (NIMH). URL: https://www.nimh.nih.gov/health/statistics/major-depression [accessed 2023-06-14]
2. Twenge JM, Cooper AB, Joiner TE, Duffy ME, Binau SG. Age, period, and cohort trends in mood disorder indicators and suicide-related outcomes in a nationally representative dataset, 2005-2017. J Abnorm Psychol 2019 Apr;128(3):185-199. [doi: 10.1037/abn0000410] [Medline: 30869927]
3. Data and statistics on children's mental health. CDC: Centers for Disease Control and Prevention. URL: https://www.cdc.gov/children-mental-health/data-research/index.html [accessed 2024-07-17]

XSL•FO
RenderX

4. Clayborne ZM, Varin M, Colman I. Systematic review and meta-analysis: adolescent depression and long-term psychosocial outcomes. J Am Acad Child Adolesc Psychiatry 2019 Jan;58(1):72-79. [doi: 10.1016/j.jaac.2018.07.896] [Medline: 30577941]
5. Fletcher JM. Adolescent depression and educational attainment: results using sibling fixed effects. Health Econ 2010 Jul;19(7):855-871. [doi: 10.1002/hec.1526] [Medline: 19582699]
6. Rao U, Hammen C, Daley SE. Continuity of depression during the transition to adulthood: a 5-year longitudinal study of young women. J Am Acad Child Adolesc Psychiatry 1999 Jul;38(7):908-915. [doi: 10.1097/00004583-199907000-00022] [Medline: 10405510]
7. Wilson S, Dumornay NM. Rising rates of adolescent depression in the United States: challenges and opportunities in the 2020s. J Adolesc Health 2022 Mar;70(3):354-355. [doi: 10.1016/j.jadohealth.2021.12.003] [Medline: 35183317]
8. Copeland WE, Alaie I, Jonsson U, Shanahan L. Associations of childhood and adolescent depression with adult psychiatric and functional outcomes. J Am Acad Child Adolesc Psychiatry 2021 May;60(5):604-611. [doi: 10.1016/j.jaac.2020.07.895] [Medline: 32758528]
9. Fried EI. Moving forward: how depression heterogeneity hinders progress in treatment and research. Expert Rev Neurother 2017 May;17(5):423-425. [doi: 10.1080/14737175.2017.1307737] [Medline: 28293960]
10. Schoevers RA, van Borkulo CD, Lamers F, et al. Affect fluctuations examined with ecological momentary assessment in patients with current or remitted depression and anxiety disorders. Psychol Med 2021 Aug;51(11):1906-1915. [doi: 10.1017/S0033291720000689] [Medline: 32234092]
11. Clasen PC, Fisher AJ, Beevers CG. Mood-reactive self-esteem and depression vulnerability: person-specific symptom dynamics via smart phone assessment. PLoS ONE 2015;10(7):e0129774. [doi: 10.1371/journal.pone.0129774] [Medline: 26131724]
12. Hosenfeld B, Bos EH, Wardenaar KJ, et al. Major depressive disorder as a nonlinear dynamic system: bimodality in the frequency distribution of depressive symptoms over time. BMC Psychiatry 2015 Sep 18;15:222. [doi: 10.1186/s12888-015-0596-5] [Medline: 26385384]
13. Ben-Zeev D, Young MA. Accuracy of hospitalized depressed patients' and healthy controls' retrospective symptom reports: an experience sampling study. J Nerv Ment Dis 2010 Apr;198(4):280-285. [doi: 10.1097/NMD.0b013e3181d6141f] [Medline: 20386257]
14. Ben-Zeev D, Young MA, Madsen JW. Retrospective recall of affect in clinically depressed individuals and controls. Cogn Emot 2009 Aug;23(5):1021-1040. [doi: 10.1080/02699930802607937]
15. Ruhe HG, Mocking RJT, Figueroa CA, et al. Emotional biases and recurrence in major depressive disorder. results of 2.5 years follow-up of drug-free cohort vulnerable for recurrence. Front Psychiatry 2019;10:145. [doi: 10.3389/fpsyt.2019.00145] [Medline: 30984039]
16. Bianchi R, Laurent E, Schonfeld IS, Bietti LM, Mayor E. Memory bias toward emotional information in burnout and depression. J Health Psychol 2020 Sep;25(10-11):1567-1575. [doi: 10.1177/1359105318765621] [Medline: 29600730]
17. Marsh L, Edginton T, Conway MA, Loveday C. Positivity bias in past and future episodic thinking: relationship with anxiety, depression, and retrieval-induced forgetting. Q J Exp Psychol (Hove) 2019 Mar;72(3):508-522. [doi: 10.1177/1747021818758620] [Medline: 29364056]
18. Urban EJ, Charles ST, Levine LJ, Almeida DM. Depression history and memory bias for specific daily emotions. PLoS ONE 2018;13(9):e0203574. [doi: 10.1371/journal.pone.0203574] [Medline: 30192853]
19. Russell MA, Gajos JM. Annual research review: ecological momentary assessment studies in child psychology and psychiatry. J Child Psychol Psychiatry 2020 Mar;61(3):376-394. [doi: 10.1111/jcpp.13204] [Medline: 31997358]
20. Armey MF, Schatten HT, Haradhvala N, Miller IW. Ecological momentary assessment (EMA) of depression-related phenomena. Curr Opin Psychol 2015 Aug;4:21-25. [doi: 10.1016/j.copsyc.2015.01.002]
21. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol 2008;4(1):1-32. [doi: 10.1146/annurev.clinpsy.3.022806.091415] [Medline: 18509902]
22. Price M, Yuen EK, Goetter EM, et al. mHealth: a mechanism to deliver more accessible, more effective mental health care. Clin Psychol Psychother 2014;21(5):427-436. [doi: 10.1002/cpp.1855] [Medline: 23918764]
23. Beames JR, Kikas K, Werner-Seidler A. Prevention and early intervention of depression in young people: an integrated narrative review of affective awareness and ecological momentary assessment. BMC Psychol 2021 Aug 16;9(1):113. [doi: 10.1186/s40359-021-00614-6] [Medline: 34392830]
24. Schueller SM, Aguilera A, Mohr DC. Ecological momentary interventions for depression and anxiety. Depress Anxiety 2017 Jun;34(6):540-545. [doi: 10.1002/da.22649] [Medline: 28494123]
25. Baltasar-Tello I, Miguélez-Fernández C, Peñuelas-Calvo I, Carballo JJ. Ecological momentary assessment and mood disorders in children and adolescents: a systematic review. Curr Psychiatry Rep 2018 Aug 1;20(8):66. [doi: 10.1007/s11920-018-0913-z] [Medline: 30069650]
26. Yim SJ, Lui LMW, Lee Y, et al. The utility of smartphone-based, ecological momentary assessment for depressive symptoms. J Affect Disord 2020 Sep 1;274:602-609. [doi: 10.1016/j.jad.2020.05.116] [Medline: 32663993]

XSL•FO
RenderX

27. Axelson DA, Bertocci MA, Lewin DS, et al. Measuring mood and complex behavior in natural environments: use of ecological momentary assessment in pediatric affective disorders. J Child Adolesc Psychopharmacol 2003;13(3):253-266. [doi: 10.1089/104454603322572589] [Medline: 14642013]
28. Cousins JC, Whalen DJ, Dahl RE, et al. The bidirectional association between daytime affect and nighttime sleep in youth with anxiety and depression. J Pediatr Psychol 2011 Oct;36(9):969-979. [doi: 10.1093/jpepsy/jsr036] [Medline: 21795377]
29. Forbes EE, Hariri AR, Martin SL, et al. Altered striatal activation predicting real-world positive affect in adolescent major depressive disorder. AJP 2009 Jan;166(1):64-73. [doi: 10.1176/appi.ajp.2008.07081336]
30. Forbes EE, Stepp SD, Dahl RE, et al. Real-world affect and social context as predictors of treatment response in child and adolescent depression and anxiety: an ecological momentary assessment study. J Child Adolesc Psychopharmacol 2012 Feb;22(1):37-47. [doi: 10.1089/cap.2011.0085] [Medline: 22339611]
31. Primack BA, Silk JS, DeLozier CR, et al. Using ecological momentary assessment to determine media use by individuals with and without major depressive disorder. Arch Pediatr Adolesc Med 2011 Apr;165(4):360-365. [doi: 10.1001/archpediatrics.2011.27] [Medline: 21464384]
32. Silk JS, Dahl RE, Ryan ND, et al. Pupillary reactivity to emotional information in child and adolescent depression: links to clinical and ecological measures. AJP 2007 Dec;164(12):1873-1880. [doi: 10.1176/appi.ajp.2007.06111816]
33. Silk JS, Forbes EE, Whalen DJ, et al. Daily emotional dynamics in depressed youth: a cell phone ecological momentary assessment study. J Exp Child Psychol 2011 Oct;110(2):241-257. [doi: 10.1016/j.jecp.2010.10.007] [Medline: 21112595]
34. Waller JM, Silk JS, Stone LB, Dahl RE. Co-rumination and co-problem solving in the daily lives of adolescents with major depressive disorder. J Am Acad Child Adolesc Psychiatry 2014 Aug;53(8):869-878. [doi: 10.1016/j.jaac.2014.05.004] [Medline: 25062594]
35. Whalen DJ, Silk JS, Semel M, et al. Caffeine consumption, sleep, and affect in the natural environments of depressed youth and healthy controls. J Pediatr Psychol 2008 May;33(4):358-367. [doi: 10.1093/jpepsy/jsm086] [Medline: 17947257]
36. Frost A, Hoyt LT, Chung AL, Adam EK. Daily life with depressive symptoms: gender differences in adolescents' everyday emotional experiences. J Adolesc 2015 Aug;43:132-141. [doi: 10.1016/j.adolescence.2015.06.001] [Medline: 26107805]
37. Mor N, Doane LD, Adam EK, et al. Within-person variations in self-focused attention and negative affect in depression and anxiety: a diary study. Cogn Emot 2010 Jan;24(1):48-62. [doi: 10.1080/02699930802499715]
38. Sequeira L, Battaglia M, Perrotta S, Merikangas K, Strauss J. Digital phenotyping with mobile and wearable devices: advanced symptom measurement in child and adolescent depression. J Am Acad Child Adolesc Psychiatry 2019 Sep;58(9):841-845. [doi: 10.1016/j.jaac.2019.04.011] [Medline: 31445619]
39. Minich M, Moreno M. Real-world adolescent smartphone use is associated with improvements in mood: an ecological momentary assessment study. PLoS ONE 2024;19(5):e0298422. [doi: 10.1371/journal.pone.0298422] [Medline: 38809836]
40. Bickham DS, Hswen Y, Rich M. Media use and depression: exposure, household rules, and symptoms among young adolescents in the USA. Int J Public Health 2015 Feb;60(2):147-155. [doi: 10.1007/s00038-014-0647-6] [Medline: 25586816]
41. Cushing CC, Marker AM, Bejarano CM, Crick CJ, Huffhines LP. Latent variable mixture modeling of ecological momentary assessment data: implications for screening and adolescent mood profiles. J Child Fam Stud 2017 Jun;26(6):1565-1572. [doi: 10.1007/s10826-017-0689-5]
42. First MB. Structured clinical interview for the DSM (SCID). In: Cautin RL, Lilienfeld SO, editors. The Encyclopedia of Clinical Psychology: Wiley-Blackwell; 2015:1-6. [doi: 10.1002/9781118625392]
43. Geller B, Zimerman B, Williams M, et al. Reliability of the Washington University in St. Louis kiddie schedule for affective disorders and schizophrenia (WASH-U-KSADS) mania and rapid cycling sections. J Am Acad Child Adolesc Psychiatry 2001 Apr;40(4):450-455. [doi: 10.1097/00004583-200104000-00014] [Medline: 11314571]
44. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613. [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
45. Overall JE, Gorham DR. The brief psychiatric rating scale. Psychol Rep 1962 Jun;10(3):799-812. [doi: 10.2466/pr0.1962.10.3.799]
46. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
47. Buysse DJ, Reynolds CF III, Monk TH, Berman SR, Kupfer DJ. The Pittsburgh sleep quality index: a new instrument for psychiatric practice and research. Psychiatry Res 1989 May;28(2):193-213. [doi: 10.1016/0165-1781(89)90047-4] [Medline: 2748771]
48. Hollingshead A. Four factor index of social status. : Yale University; 1975.
49. LifeData: experience sampling app for research. LifeData Corp. URL: https://www.lifedatacorp.com [accessed 2024-07-17]
50. Fitbit web API data dictionary. fitbit. 2023. URL: https://enterprise.fitbit.com/wp-content/uploads/Fitbit-Web-API-Data-Dictionary-Downloadable-Version-2023.pdf [accessed 2025-02-04]
51. Pandas. The pandas development team. : Zenodo; 2020.
52. PyCharm. JetBrains. 2024. URL: https://www.jetbrains.com/pycharm/download/?section=windows [accessed 2025-04-30]
53. Wickham H, Vaughan D, Girlich M. Tidy messy data. : tidyr; 2023.
54. Wickham H, François R, Henry L, Vaughan D. A grammar of data manipulation. dplyr. 2023. URL: https://dplyr.tidyverse.org/ [accessed 2025-04-30]

XSL•FO
RenderX

55. R Core Team. R: A language and environment for statistical computing. : R Foundation for Statistical Computing; 2017.
56. Zhang Y, Folarin AA, Sun S, et al. Relationship between major depression symptom severity and sleep collected using a wristband wearable device: multicenter longitudinal observational study. JMIR Mhealth Uhealth 2021 Apr 12;9(4):e24604. [doi: 10.2196/24604] [Medline: 33843591]
57. Kleiman EM, Turner BJ, Fedor S, Beale EE, Huffman JC, Nock MK. Examination of real-time fluctuations in suicidal ideation and its risk factors: results from two ecological momentary assessment studies. J Abnorm Psychol 2017 Aug;126(6):726-738. [doi: 10.1037/abn0000273] [Medline: 28481571]
58. Wickham H. Ggplot2: Elegant Graphics for Data Analysis: Springer-Verlag New York; 2016. [doi: 10.1007/978-3-319-24277-4_9]
59. von Neumann J, Kent RH, Bellinson HR, Hart BI. The mean square successive difference. Ann Math Statist 1941 Jun;12(2):153-162. [doi: 10.1214/aoms/1177731746]
60. Brown CE. Coefficient of variation. In: Applied Multivariate Statistics in Geohydrology and Related Sciences: Springer; 1998:155-157. [doi: 10.1007/978-3-642-80328-4]
61. Revelle W. psych: procedures for psychological, psychometric, and personality research. R package version 1.0–95. Psychology (Irvine) 2013.
62. Bates D, Maechler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. J Stat Softw 2015;67:1-48. [doi: 10.18637/jss.v067.i01]
63. Ben-Shachar M, Lüdecke D, Makowski D. Effectsize: estimation of effect size indices and standardized parameters. JOSS 2020;5(56):2815. [doi: 10.21105/joss.02815]
64. Kuznetsova A, Brockhoff PB, Christensen RHB. lmerTest package: tests in linear mixed effects models. J Stat Soft 2017;82(13). [doi: 10.18637/jss.v082.i13]
65. Fried EI, Nesse RM. Depression is not a consistent syndrome: an investigation of unique symptom patterns in the STAR*D study. J Affect Disord 2015 Feb 1;172:96-102. [doi: 10.1016/j.jad.2014.10.010] [Medline: 25451401]
66. Alvaro PK, Roberts RM, Harris JK. A systematic review assessing bidirectionality between sleep disturbances, anxiety, and depression. Sleep 2013 Jul 1;36(7):1059-1068. [doi: 10.5665/sleep.2810] [Medline: 23814343]
67. Gradisar M, Kahn M, Micic G, et al. Sleep's role in the development and resolution of adolescent depression. Nat Rev Psychol 2022;1(9):512-523. [doi: 10.1038/s44159-022-00074-8] [Medline: 35754789]
68. Roberts RE, Duong HT. The prospective association between sleep deprivation and depression among adolescents. Sleep 2014 Feb 1;37(2):239-244. [doi: 10.5665/sleep.3388] [Medline: 24497652]
69. Li L, Zhang Q, Zhu L, et al. Daytime naps and depression risk: a meta-analysis of observational studies. Front Psychol 2022;13:1051128. [doi: 10.3389/fpsyg.2022.1051128]
70. Foley DJ, Vitiello MV, Bliwise DL, Ancoli-Israel S, Monjan AA, Walsh JK. Frequent napping is associated with excessive daytime sleepiness, depression, pain, and nocturia in older adults: findings from the national sleep foundation "2003 Sleep in America" poll. Am J Geriatr Psychiatry 2007 Apr;15(4):344-350. [doi: 10.1097/01.JGP.0000249385.50101.67] [Medline: 17384317]
71. Murray CB, Murphy LK, Palermo TM, Clarke GM. Pain and sleep-wake disturbances in adolescents with depressive disorders. J Clin Child Adolesc Psychol 2012;41(4):482-490. [doi: 10.1080/15374416.2012.658613] [Medline: 22420746]
72. Daly M. Prevalence of depression among adolescents in the U.S. from 2009 to 2019: analysis of trends by sex, race/ethnicity, and income. J Adolesc Health 2022 Mar;70(3):496-499. [doi: 10.1016/j.jadohealth.2021.08.026] [Medline: 34663534]
73. Haghayegh S, Khoshnevis S, Smolensky MH, Diller KR, Castriotta RJ. Accuracy of wristband Fitbit models in assessing sleep: systematic review and meta-analysis. J Med Internet Res 2019 Nov 28;21(11):e16273. [doi: 10.2196/16273] [Medline: 31778122]
74. Renerts K, Doheny EP, Baumann C, et al. F63 validation of fitbit charge 4 for sleep monitoring in participants with huntington's disease. Presented at: EHDN 2022 Plenary Meeting; Sep 2022; Bologna, Italy. [doi: 10.1136/jnnp-2022-ehdn.154]
75. Feehan LM, Geldman J, Sayre EC, et al. Accuracy of Fitbit devices: systematic review and narrative syntheses of quantitative data. JMIR Mhealth Uhealth 2018 Aug 9;6(8):e10527. [doi: 10.2196/10527] [Medline: 30093371]

Abbreviations

CoV: coefficient of variation
DSM-5: Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition
EMA: ecological momentary assessment
GAD: generalized anxiety disorder
ICC: intraclass correlation
MDD: major depressive disorder
PHQ: patient health questionnaire
REM: rapid eye movement
RMSSD: root mean square of successive differences

XSL•FO
RenderX

TIB: time in bed
TST: total sleep time

Edited by A Mavragani; submitted 09.09.24; peer-reviewed by D Benrimoh; revised version received 10.02.25; accepted 11.02.25; published 11.06.25.

Please cite as:
Unzueta Saavedra J, Deaso EA, Austin M, Cadavid L, Kraff R, Knowles EEM
Smartphone Ecological Momentary Assessment and Wearable Activity Tracking in Pediatric Depression: Cohort Study
JMIR Form Res 2025;9:e66187
URL: https://formative.jmir.org/2025/1/e66187
doi:10.2196/66187

©Jimena Unzueta Saavedra, Emma A Deaso, Margot Austin, Laura Cadavid, Rachel Kraff, Emma E M Knowles. Originally published in JMIR Formative Research (https://formative.jmir.org), 11.6.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Impacts of the Mindfulness Meditation Mobile App Calm on Undergraduate Students' Sleep and Emotional State: Pilot Randomized Controlled Trial

Tovan Lew[1], BSc; Natnaiel M Dubale[1], BSc; Erik Doose[1], M BEMA; Alex Adenuga[2], MSc; Holly E Bates[1], PhD; Sarah L West[1,3], PhD

[1]Department of Biology, Trent University, 1600 W Bank Drive, Peterborough, ON, Canada
[2]School of Graduate Studies, Trent University, Peterborough, ON, Canada
[3]Department of Kinesiology, Trent University, Peterborough, ON, Canada

Corresponding Author:
Sarah L West, PhD
Department of Biology, Trent University, 1600 W Bank Drive, Peterborough, ON, Canada

Abstract

Background: Undergraduate students frequently experience negative emotional states and sleep quality, which is believed to have worsened following the COVID-19 pandemic.

Objective: This study piloted the use of a popular mobile mindfulness app (Calm) as a potential intervention to improve state depression, anxiety, stress, and sleep quality in undergraduate students attending a Canadian university, following the COVID-19 pandemic.

Methods: Undergraduate students were randomly assigned to a control or treatment group and completed a series of 3 questionnaires to evaluate baseline state emotional health (Depression Anxiety Stress Scale 42-Item Version [DASS-42], Perceived Stress Scale 10-Item Version [PSS-10], and Pittsburgh Sleep Quality Index). Treatment group participants were instructed to engage with the fully-automated Calm app's sleep section for 30 days: 20 minutes daily, 5 days a week, along with an additional 30 minutes of interaction with other app sections each week, resulting in a goal of 130 minutes per week. The control participants were instructed to continue with everyday life and refrain from the use of mindfulness-based apps for 30 days. Following the 30-day treatment period, all participants repeated the 3 questionnaires. The impact of the treatment on all outcomes was examined using linear mixed model analyses. Independent samples *t* tests were used to determine if psychosocial health or sleep scores differed between baseline and follow-up and if differences in such scores were present between the groups.

Results: A total of 80 students met the inclusion criteria and were randomly assigned to the control (n=40) or treatment (n=40) group. One control participant was lost to follow-up and 3 treatment participants discontinued engaging with the Calm app. Both control (n=39) and treatment (n=37) groups began with similar demographic, emotional state, and sleep characteristics. Treatment participants engaged with the Calm app's sleep section for an average of 234 minutes per week; however, 54% (20/37) met the minimum prescribed interaction time across all 4 weeks. Following the 30-day treatment period, compared to the control group, the treatment group's state anxiety (mean 14, SD 7.4 vs mean 12, SD 7.8; P=.002), state stress (DASS-42: mean 20, SD 8.8 vs mean 15, SD 8.5; P<.001; PSS-10: mean 22, SD 5.9 vs mean 19, SD 5.9; P=.02), and sleep quality (mean 7.7, SD 2.7 vs mean 6.4, SD 3.5; P<.001) improved. Posttreatment, state stress and perceived stress severity was lower in the treatment versus control group (DASS-42: P=.02; PSS-10: P=.03, respectively).

Conclusions: These pilot findings indicate that a mindfulness app may be an effective tool for reducing state anxiety and stress, as well as enhancing sleep quality among undergraduate university students. A larger, randomized controlled trial should confirm these findings.

(JMIR Form Res 2025;9:e66131) doi:10.2196/66131

KEYWORDS

mindfulness; mobile apps; students; mHealth; sleep; emotional state; pilot study; depression; anxiety; stress; university; mobile phone; mobile health

XSL•FO
RenderX

Introduction

Poor emotional health is a considerable problem among university students, with current evidence suggesting depression, anxiety, and stress are frequently reported among this population [1-4]. For instance, in a World Health Organization survey of students from 19 universities across 8 countries, 35% of participants met the diagnostic criteria for either a mood, substance-use, or anxiety disorder [1]. Recent systematic reviews suggest the COVID-19 pandemic led to a considerable worsening of psychosocial health among postsecondary students [5,6]. Furthermore, a growing consensus in the literature suggests that, unlike in the general population, postsecondary student mental health has not returned to prepandemic levels [7-15].

Poor sleep quality is another common observance among postsecondary students [16], with conditions such as delayed sleep phase syndrome being 50% more prevalent among postsecondary students compared to the general population [17]. A distinctive combination of social, work, and academic pressures renders this population particularly susceptible to poor sleep quality [18,19]. Sleep disturbances can have enduring effects on mental and physical well-being, having been associated with conditions including cardiometabolic diseases, cancer, and chronic respiratory diseases [20]. Sleep issues have also been linked to decreased academic productivity [21], performance [21], and increased dropout rates in postsecondary students [22].

While bidirectional relationships are believed to exist between each of depression, anxiety, and stress [23-26], a recent meta-analysis of randomized controlled trials does suggest poor sleep plays a causal role in the development of reduced psychosocial health [27]. Accordingly, poor sleep quality may create a cyclic effect wherein worse sleep quality negatively impacts psychosocial health, thereby further complicating the maintenance of healthy sleep habits [23,26,28,29]. As increasing levels of depression, anxiety, and stress have been linked to lower levels of help-seeking intention [30], additional targeted resources aimed at improving state emotional health and sleep quality are needed among postsecondary students.

Mindfulness interventions aim to extend upon the levels of attention and awareness in normal functioning by fostering sustained, nonjudgmental present-moment awareness [31]. These practices, which commonly include focused attention meditation, body scanning, and mindful breathing, aim to mitigate cognitive distractions and promote a deliberate focus on internal and external experiences [31]. Recently, mindfulness meditation has advanced to include mobile applications offering users the convenience of guided practice at their preferred time and location. These resources offer a potential alternative to traditional mental health supports and may address the most common barriers (such as long wait times and confidentiality concerns) to accessing mental health care as reported by a recent random sample of approximately 2000 Canadian students [15]. Current meta-analyses suggest mindfulness-based online or app-based tools are effective in significantly improving both mental health and sleep quality in postsecondary students [32-35]. While reviews generally agree that online and app-based mindfulness tools can significantly reduce anxiety and stress in students, the evidence regarding their impact on depression is mixed [32-34,36]. Moreover, few studies to date have explored the impact of an app-based mindfulness intervention on the critical intersection of emotional well-being and sleep quality in postsecondary students.

Therefore, the following pilot study aimed to explore the impact of a 1-month stand-alone treatment using the mindfulness-based app Calm on state emotional health and sleep quality in a sample of undergraduate students attending a Canadian university post the COVID-19 pandemic. We hypothesized that using the Calm app would improve undergraduate students' sleep quality and symptoms of state stress, anxiety, and depression relative to a control group. As this was a pilot investigation, we also hypothesized that using the Calm app for the duration of this study would be feasible, as determined by meeting the minimum suggested minutes of app interaction per day.

Methods

Study Design and Participants

A pilot nonblinded randomized controlled study was conducted among undergraduate students attending a university in Ontario, Canada. Inclusion criteria were as follows: (1) be 17 years of age or older, (2) be enrolled in full-time undergraduate studies, (3) possess a smartphone or tablet capable of effectively running the Calm app, (4) not be regular users or subscribers of the Calm app, and (5) provide informed consent. Internet literacy was an implicit eligibility requirement. Past or current diagnoses of clinical depression or anxiety were not considered an exclusion criteria for this study. Participant recruitment occurred during 2 periods: between January 31 and February 13, 2023 (winter term), and September 21 and October 13, 2023 (fall term). Participants were recruited on a first-come, first-served basis using an approved script distributed via email to student mailing lists across various academic departments at the university. Word of mouth among students also contributed to participant recruitment.

The Calm app is a smartphone app available from developer Calm.com on iOS (Apple Inc) and Android (Open Handset Alliance [led by Google]) operating systems, that includes a variety of mindfulness features such as guided meditation, sleep stories, breathing exercises, and mindfulness-based stress reduction techniques [37]. The app's sleep section houses several resources designed to improve sleep quality, such as guided relaxation exercises, sleep stories, and calming soundscapes [37]. The Calm app was selected due to its comprehensive approach and popularity. Study participants were randomized to either the treatment or control group by study personnel via the random allocation functions in Microsoft Excel (version 16.83, Microsoft Corporation). An identical in-person questionnaire was administered to participants by study personnel at 2 time points: before (baseline) and after (follow-up) the 30-day Calm app treatment period.

Participants randomized to the treatment group received 1-month of access to the Calm app, accessible via a unique URL. The

XSL•FO
RenderX

30-day length of intervention was selected for the current pilot study as this was thought to be a realistic, shorter timeframe to allow for an initial assessment of the potential benefits of using the Calm app. The treatment group was instructed to use the app's sleep section for 30 days: 20 minutes daily, 5 days a week, along with an additional 30 minutes of interaction with other app sections each week, resulting in a total of 130 minutes of app use per week. This amount of time was selected for the intervention as it was thought to be similar to one 30-minute meditation class per day. We advised use for 5 days per week, instead of 7, to provide students with greater flexibility and enhance the probability of sustained engagement. Compliance was tracked using a logbook; participants filled out the total daily time spent using the Calm app during the 30-day treatment period. The control group was instructed to continue with their usual daily routines, refraining from the use of any mindfulness apps such as Calm. After the 30-day study period, and following the completion of the follow-up questionnaires, the control group received free access to the Calm app for 1 month. Figure 1 describes this study timeline.

Figure 1. Study design line schematic including recruitment, baseline, intervention, and follow-up.

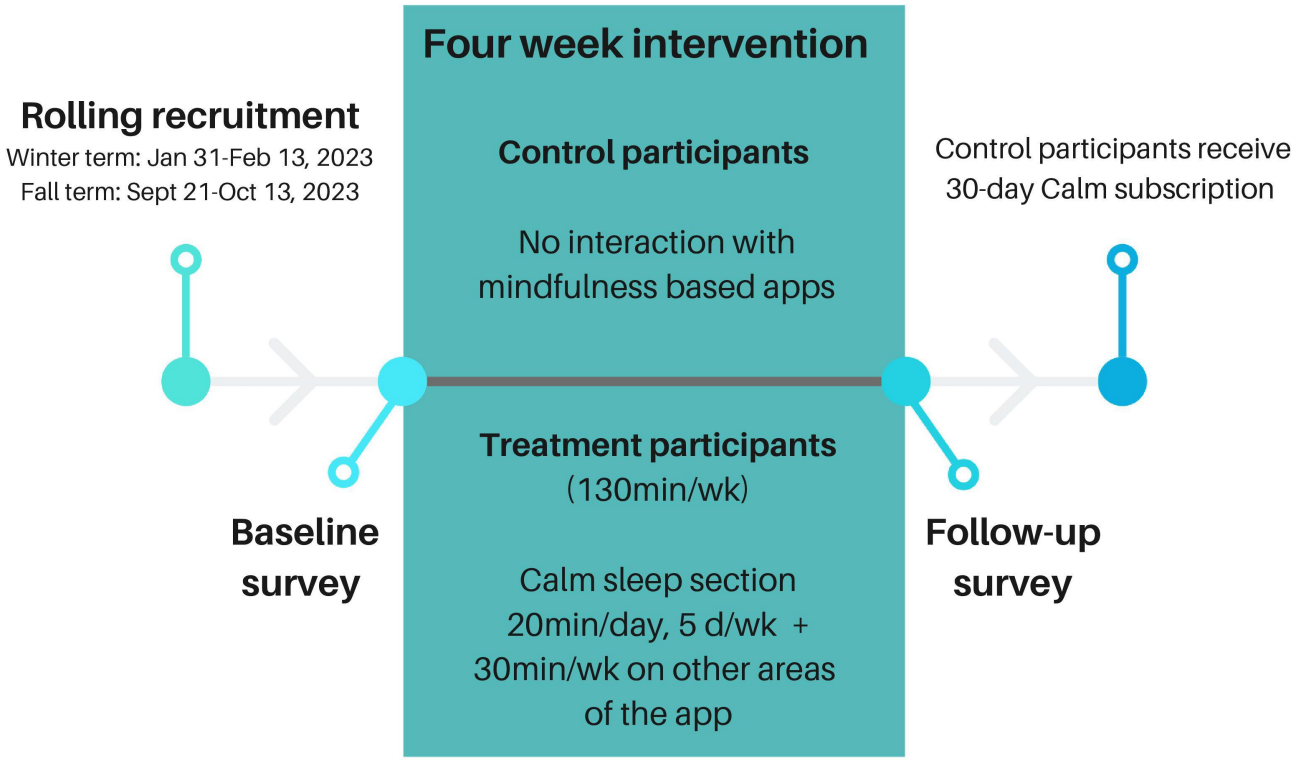

Ethical Considerations

This study was reviewed and received approval from Trent University's Research Ethics Board (28256). Given that this study was a pilot investigation, the protocol was not registered within a clinical trials database. Following ethics approval, interested participants were provided with information about this study via a consent form and were provided with the opportunity to review the form and ask questions about this study to a member of the research team. Written informed consent was obtained from all participants. Each participant was assigned a unique study ID number to ensure privacy and confidentiality. Access to 1 month of the Calm app served as the incentive for study participation, provided to the treatment group during their study interval and the control group for 1 month following their 30-day study interval.

Questionnaire and Measures

The questionnaire consisted of 4 components including a demographic questionnaire which collected age, gender, year of study, and program of study. Components 2 to 5 consisted of validated self-assessed psychological tools: the DASS-42 (Depression Anxiety Stress Scale 42-Item Version) [38], the Pittsburgh Sleep Quality Index (PSQI) [39], and the Perceived Stress Scale 10-Item Version (PSS-10) [40]. The DASS-42 is a validated tool consisting of 3 sections, each containing 14 subgrouped questions assessing state depression, anxiety, and stress [41-43] and has been used to assess emotional states among postsecondary students previously [44-46]. The DASS-42 produces 3 scores ranging from 0 to 42, each corresponding to either state depression, anxiety, or stress [47]. Such scores can be categorized as normal to extremely severe according to the following scale for anxiety, stress, and depression, respectively: normal (0 - 7, 0 - 14, and 0 - 9), mild (8-9, 15-18, and 10-13), moderate (10-14, 19-25, and 14-20), severe (15-19, 26-33, and 21-27), extremely severe (≥20, ≥34, and ≥28) [47].

The PSQI is a 10-item self-report that evaluates sleep over the past month on 7 domains, producing a global sleep quality score. Higher scores correspond to poorer sleep quality [39]. Such domains include sleep quality, latency, duration, efficiency, disturbance, medication use, and daytime dysfunction [39]. The PSQI has been validated through previous studies [48,49] and has been used to examine sleep in undergraduate students [50-52]. The PSS-10 assesses self-perceived stress over the past month using 10 questions on a 5-point Likert scale [40]. Stress

XSL•FO
RenderX

levels can then be categorized as follows: scores from 0 to 13 indicate low stress, scores from 14 to 26 indicate moderate stress, and scores from 27 to 40 indicate high stress [53]. The PSS-10 has been validated among university students [54], and a systematic review of studies assessing the PSS-10's validity supports its use [55].

Statistical Analyses

Demographic data was analyzed using descriptive statistics. Two-sample *t* test with equal variances was used to compare mean age between treatment and control groups, and Pearson chi-squared tests were used to determine if significant differences in categorical demographic variables existed between these groups. Independent *t* tests were used to determine if baseline and posttreatment scores differed significantly for each of the 5 measures (DASS-depression, DASS-anxiety, DASS-stress, PSS-stress, and PSQI-sleep) between treatment and control groups. Linear mixed models were computed for each of this study's outcomes, examining the interaction between time point (baseline or follow-up) and study group (treatment or control). Effect sizes (Cohen *d*) were calculated for linear mixed model outcomes, with thresholds defined as large ($d \geq 0.80$), medium ($0.50 \leq d < 0.80$), and small ($0.20 \leq d < 0.50$). Statistical significance was set at $P < .05$. The term of data collection (winter 2023 or fall 2023) was also included in the models to examine if the term of study had an impact on the treatment and study outcomes. Statistical analyses were conducted in R (version 4.3.2, R Foundation for Statistical Computing), and the following packages were used for analyses: lme4, lmerTest, and effsize. Figures were generated using Microsoft Excel.

Results

Participant Exclusion and Demographic Characteristics

A total of 80 participants, equally distributed between fall and winter recruitment terms, were randomly assigned to either the treatment (n=40) or control (n=40) group. One participant during the fall recruitment term and 2 participants during the winter term withdrew from this study, leading to their exclusion from the treatment group. Additionally, one participant was excluded from the control group due to loss at follow-up during the fall data collection period. Therefore, data from 76 undergraduate students (38 from each collection period; 39 in the control, and 37 in the treatment) were included in the analysis (Figure 2).

Figure 2. CONSORT flow diagram of participants' eligibility, enrollment, group allocation, treatment, and data analyzed in this pilot study. CONSORT: Consolidated Standards of Reporting Trials.

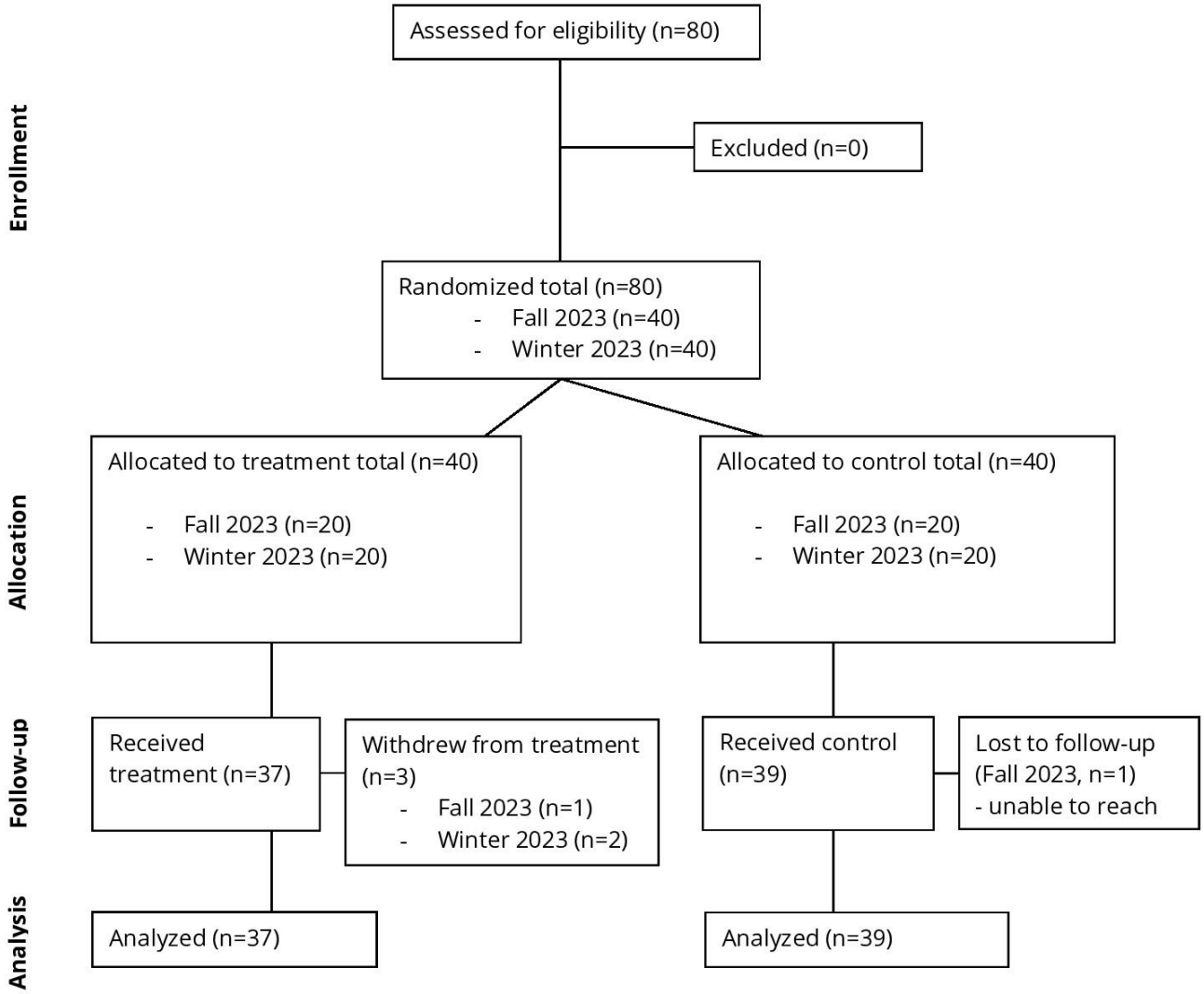

XSL•FO
RenderX

Table 1 presents demographic data by study group. The mean age of participants did not differ significantly (*P*=.27) between the 2 groups (mean 20.6, SD 3.1 and mean 21.8, SD 5.8 years in the control and treatment groups, respectively). Most participants identified as female, including 82% (32/76) and 81% (30/76) of participants in the control and treatment groups, respectively (*P*=.80). Students' year of study ranged from year 1 to year 5+. There was also an equal distribution of students by year of study (*P*=.59) and program of study (*P*=.21) between the control and treatment groups.

Table . Demographic characteristics of participants, comparing the control and treatment groups (N=76). All programs are bachelor's degrees.

Characteristic	Control group (n=39)	Treatment group (n=37)	*P* value
Age (years), mean (SD)	20.64 (3.09)	21.81 (5.76)	.27
Gender, n (%)			.80
Female	32 (82)	30 (81)	
Male	5 (13)	6 (16)	
Gender nonconforming	2 (5)	1 (3)	
Year of study, n (%)			.59
First year	11 (28)	10 (27)	
Second year	8 (21)	5 (14)	
Third year	11 (28)	12 (32)	
Fourth year	9 (23)	8 (22)	
Fifth year or higher	0 (0)	2 (5)	
Program of study, n (%)			.21
BA[a]	2 (5)	4 (11)	
BAS[b]	2 (5)	1 (3)	
BESS[c]	1 (3)	0 (0)	
BSc[d]	18 (46)	24 (65)	
BScFS[e]	10 (26)	3 (8)	
BScKin[f]	0 (0)	2 (5)	
BScN[g]	4 (10)	2 (5)	
BSW[h]	2 (5)	1 (3)	

[a]BA: Bachelor of Arts.
[b]BAS: Bachelor of Arts and Science.
[c]BESS: Bachelor of Environment Science or Studies.
[d]BSc: Bachelor of Science.
[e]BScFS: Bachelor of Science in Forensic Science.
[f]BScKin: Bachelor of Science in Kinesiology.
[g]BScN: Bachelor of Science in Nursing.
[h]BSW: Bachelor of Social Work.

State Depression, Anxiety, and Stress

Baseline

Table 2 presents pretreatment outcome data. The mean state depression scores pretreatment were similar in the control and treatment groups (*P*=.58) and represented moderate severity state depression at baseline in both groups. The mean state anxiety scores at baseline were also similar between the control and treatment groups (*P*=.11), however, scores represented moderate-severity state anxiety in the control group and severe state anxiety in the treatment group. Baseline mean DASS-42 (*P*=.85) and PSS-10 (*P*=.10) scores were comparable between groups, both indicating moderate stress severity. Overall, 26% - 44% (10/39 to 17/39) of those in the control group experienced severe to extremely severe levels of state depression, anxiety, and stress. Comparatively, 24% - 54% (9/37 to 20/37) of participants in the treatment group reported similar levels.

XSL•FO
RenderX

Table . Mean baseline state emotional health scores, comparing the control and treatment groups (N=76).

Measure	Control group (n=39)		Treatment group (n=37)		*P* value
	Mean (SD)	Score category	Mean (SD)	Score category	
DASS[a]-depression	15.41 (11.05)	Moderate	14.08 (9.91)	Moderate	.58
DASS-anxiety	14 (7.02)	Moderate	16.81 (8.27)	Severe	.11
DASS-stress	21.38 (8.4)	Moderate	21.73 (7.09)	Moderate	.85
PSS[b] -stress	22.15 (6.88)	Moderate	22.16 (6.39)	Moderate	.10

[a]DASS: Depression Anxiety Stress Scale 42-Item Version.

[b]PSS: Perceived Stress Scale 10-Item Version.

Pre- Versus Posttreatment

Table 3 summarizes the interaction of time point and study group for all study outcomes. State depression showed no significant interaction (*P*=.37; Figure 3A), and Cohen *d* effect size of the treatment group's changes in state depression was small (*d*=0.4, 95% CI 0.16 to 0.73). The interaction of time point and study group for state anxiety demonstrated a significant interaction (*P*=.002; Figure 3B), with a medium effect size within the treatment group (*d*=0.6, 95% CI 0.26-0.95).

The interaction of time point and study group for stress also showed a significant interaction for both the DASS-42 stress score (*P*<.001; Figure 3C) and PSS-10 stress score (*P*=.02; Figure 3D). The treatment group's change in stress had a large effect size for DASS-42 stress score (*d*=0.81, 95% CI 0.53-1.10) and a medium effect size for PSS-10 stress score (*d*=0.57, 95% CI 0.24-0.89).

There was no impact on outcome measures between the 2023 winter term dataset and the 2023 fall term dataset (*P*>.16; Table 4).

Table . Linear mixed model of the interaction of time point × study group for state emotional health outcomes.

Variables	Precontrol (n=39), mean (SD)	Postcontrol (n=39), mean (SD)	Control effect size (95% CI)	Pretreatment (n=37), mean (SD)	Posttreatment (n=37), mean (SD)	Treatment effect size (95% CI)	*t* value (*df*)	*P* value
DASS[a]-depression	15.4 (11.05)	12.9 (8.37)	0.25 (-0.03 to 0.53)	14.1 (9.91)	9.81 (9.08)	0.45 (0.16 to 0.73)	–0.9 (74)	.37
DASS-anxiety	14 (7.02)	14.1 (7.44)	–0.004 (–0.23 to 0.22)	16.8 (8.28)	11.95 (7.82)	0.6 (0.26 to 0.95)	–3.24 (74)	.002
DASS-stress	21.4 (8.4)	20.1 (8.84)	0.1 (–0.05 to 0.34)	21.7 (7.09)	15.2 (8.54)	0.81 (0.53 to 1.1)	–4.05 (74)	<.001
PSS[b]-stress	22.12 (6.88)	21.74 (5.92)	0.06 (–0.19 to 0.32)	22.2 (6.4)	18.7 (5.91)	0.57 (0.24 to 0.89)	–2.47 (74)	.02

[a]DASS: Depression Anxiety Stress Scale 42-Item Version.

[b]PSS: Perceived Stress Scale 10-Item Version.

XSL•FO
RenderX

Figure 3. Depression, anxiety, and stress scores among the control (n=39) and treatment (n=37) groups pre- and posttreatment with the Calm app. Peaks of the bar columns represent the mean score, and error bars represent the SD. Statistical analysis by linear mixed model. (**A**) Mean DASS-42 state depression. (**B**) Mean DASS-42 state anxiety. (**C**) Mean DASS-42 state stress. (**D**) Mean PSS-10 state stress. Mean state anxiety and state stress measured by the DASS-42 and PSS-10 decreased significantly in the treatment group following 30 days of treatment. There were no changes in any of the state emotional health outcomes in the control group between baseline and 30 days later. DASS-42: Depression Anxiety Stress Scale-42; PSS-10: Perceived Stress Scale-10. *P<.05, **P<.01, ***P<.001.

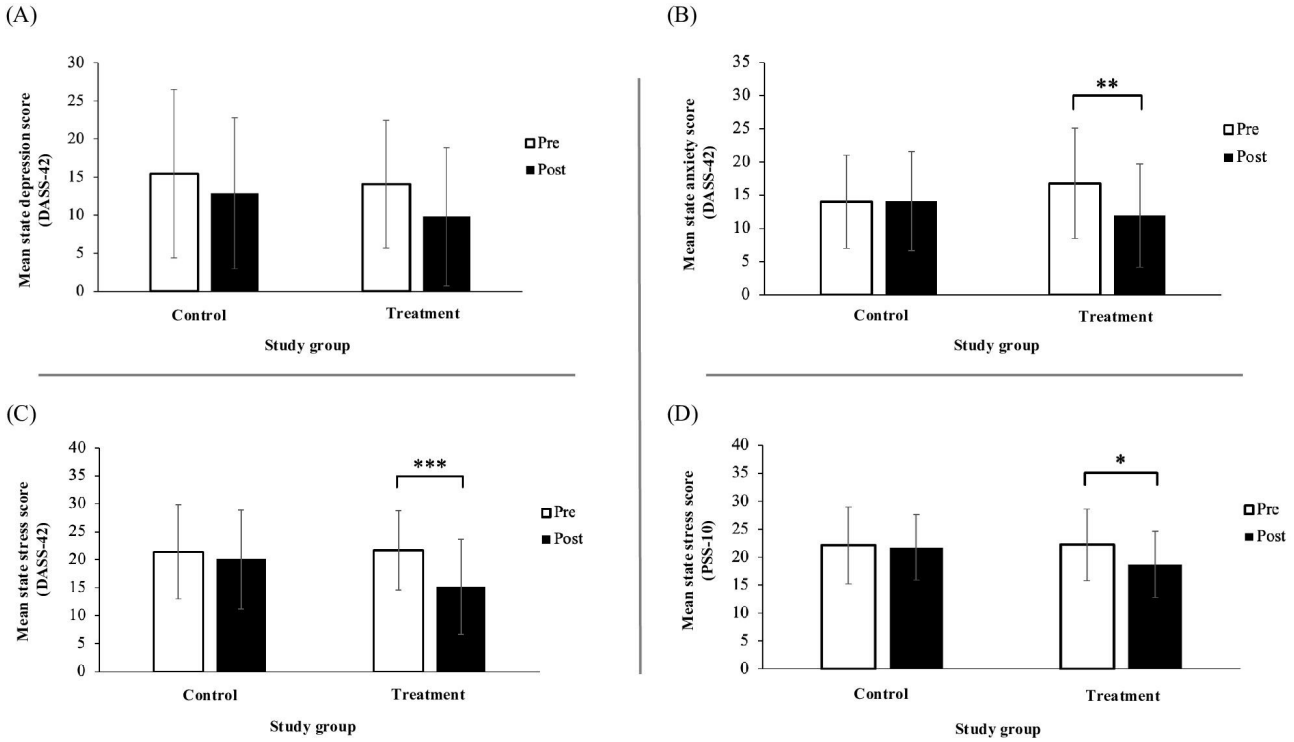

Table . Linear mixed model of interaction of fall and winter term data collection × all outcome measures.

Outcome measure	*t* value (*df*)	*P* value
DASS[a]-depression	-0.24 (72)	.81
DASS-anxiety	-1.43 (72)	.16
DASS-stress	0.86 (72)	.39
PSQI[b]	-0.23 (72)	.82
PSS[c]-stress	-0.33 (72)	.75

[a]DASS: Depression Anxiety Stress Scale 42-Item Version.

[b]PSQI: Pittsburgh Sleep Quality Index.

[c]PSS: Perceived Stress Scale 10-Item Version.

Posttreatment Follow-Up

Mean follow-up state emotional health scores are summarized in Table 5. The mean state depression follow-up scores were similar (P=.13) for the control and treatment groups, both of which were in the mild category. The mean follow-up state anxiety scores were similar between groups and corresponded to a moderate level of state anxiety in both groups. The mean DASS-42 stress scores at follow-up were mean 20.13 (SD 8.84) and mean 15.22 (SD 8.54) in the control and treatment groups, respectively, corresponding to a moderate severity in the control group and a mild severity in the treatment group. The mean PSS-10 stress scores at follow-up were mean 21.74 (SD 5.92) and mean 18.68 (SD 5.91) in the control and treatment groups, respectively, falling in the moderate stress category [53]. Both the DASS-42 and PSS-10 stress scores were higher in the control versus treatment group at study end (P=.02 and P=.03, respectively).

XSL•FO
RenderX

Table . Mean scores of state emotional health in control compared to treatment groups at posttreatment or 30 day follow-up.

Variables	Control group (n=39)		Treatment group (n=37)		*P* value
	Mean (SD)	Score category	Mean (SD)	Score category	
DASS[a]-depression	12.87 (8.37)	Mild	9.81 (9.08)	Mild	.13
DASS-anxiety	14.03 (7.44)	Moderate	11.95 (7.84)	Moderate	.24
DASS-stress	20.13 (8.84)	Moderate	15.22 (8.54)	Mild	.02
PSS[b]-stress	21.74 (5.92)	Moderate	18.68 (5.91)	Moderate	.03

[a]DASS: Depression Anxiety Stress Scale 42-Item Version.

[b]PSS: Perceived Stress Scale 10-Item Version.

Sleep Quality

Baseline

Baseline sleep quality scores were not significantly different (*P*=.14) between the control (mean 8.18, SD 3.15) and treatment (mean 9.46, SD 4.26) groups. This represents a categorization of poor sleep quality in both groups as indicated by a score greater than 5 [39]. Most participants in both groups displayed poor sleep quality, with 82% (32/39) in the control group and 84% (31/37) in the treatment group falling into this category. Over three-quarters of the treatment (28/37, 76%) and control (31/39, 79%) groups exhibited insomnia according to previously established cutoffs.

Pre- Versus Posttreatment

The interaction of time point and study group for sleep quality showed a significant interaction (*P*<.001; Figure 4), with a medium effect size within the treatment group (*d*=0.78, 95% CI 0.45 to 1.10).

Figure 4. Mean PSQI scores among the control (n=39) and treatment (n=37) groups pre- or posttreatment with the Calm app. Peaks of the bar columns represent the mean score, and error bars represent the SD. Statistical analysis by linear mixed mode. Mean PSQI score decreased (improved) significantly in the treatment group following 30 days of treatment with the Calm app. There were no changes in PSQI score in the control group between baseline and 30 days later. PSQI: Pittsburgh Sleep Quality Index. ***P<.001.

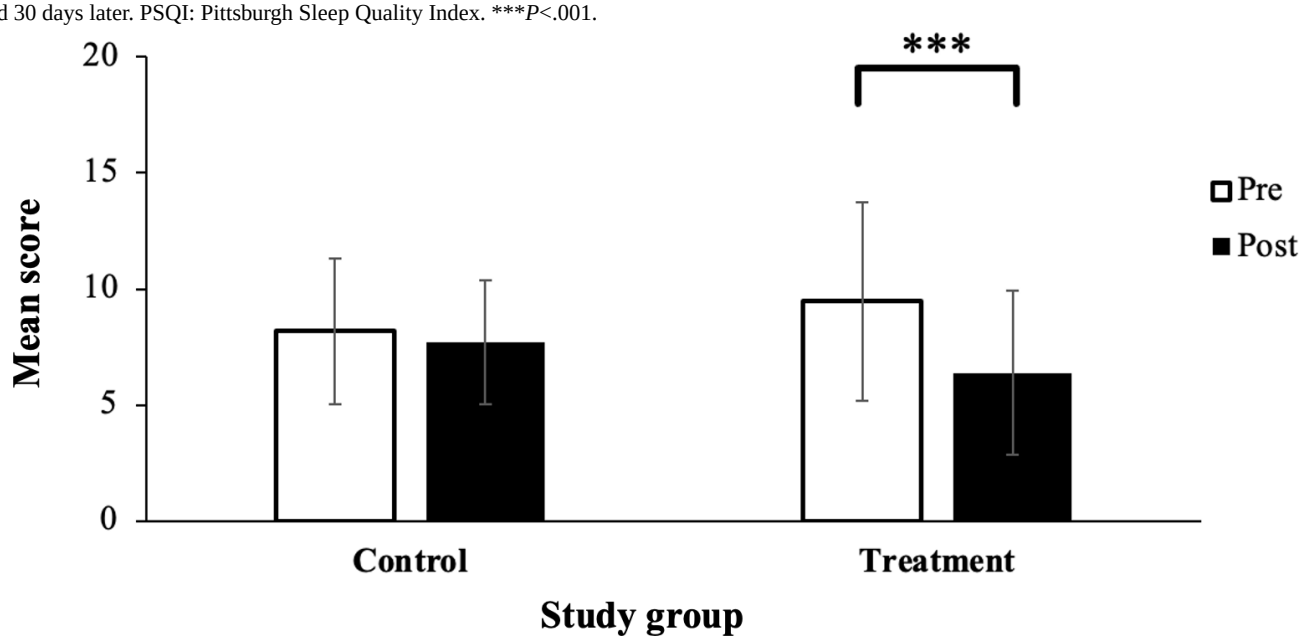

Posttreatment Follow-Up

The mean sleep quality scores at follow-up were mean 7.69 (SD 2.66) and mean 6.38 (SD 3.53) in the control and treatment groups, respectively. This represents a categorization of poor sleep quality in both groups [39]. An independent sample *t* test found no difference between the control and treatment sleep quality scores at follow-up (*P*=.07).

Adherence

Over half (20/37, 54.1%) of treatment group participants completed the minimum requested interaction time with Calm (130 min/wk) across all 4 weeks. Differences in treatment group adherence (how many weeks the minimum interaction time was met) between participants recruited during the fall ($\bar{x}$=3.3, SD=1.2) and winter ($\bar{x}$=2.6, SD=1.7) terms were not significant (*P*=.14), with a small effect size (*d*=0.48). The total average time that treatment group participants interacted with the Calm app's sleep section was 663.6 minutes in the fall group and 1208.7 minutes in the winter group (*P*=.18), representing a small effect size (*d*=0.4). Importantly, the average minimum number of minutes fall participants spent using the sleep section (144.1 min/wk) was similar to the winter participants (106.4 min/wk; *P*=.1), with a medium effect size (*d*=0.55). The most used sections of the app outside of sleep were music (10/33, 30%), daily calm (5/33, 15%), and soundscapes (5/33, 15%).

XSL•FO
RenderX

Discussion

Principal Findings

This pilot study found that following a 30-day treatment period with the Calm app, undergraduate students assigned to the treatment group displayed significant reductions in state anxiety and stress severity as well as improvements in sleep quality relative to the control group. While both groups had similar outcome measures at baseline, state stress levels were significantly lower in the treatment group compared to the control group at follow-up. This represents a decrease in state stress levels that are similar to the lower range found in worldwide representative samples pre–COVID-19 pandemic [56-62]. Moreover, the treatment group experienced a 16% reduction in the prevalence of poor sleep quality. Such improvements in state anxiety, stress, and sleep quality, despite only 54% adherence to the full-prescribed use of the Calm app, are particularly encouraging. These findings suggest that even partial engagement with mindfulness interventions can lead to meaningful benefits. They also indicate that the app's effects may be more substantial or easier to attain than expected, potentially enhancing its feasibility and appeal among university students.

State Emotional Health in Undergraduate Students

Baseline findings from our sample of undergraduate students indicate that more than one-third of participants experienced severely heightened levels of state depression, anxiety, or stress. This is higher than the worldwide representative scores of state depression, anxiety, and stress among university students documented in the literature. For instance, studies conducted in countries including the United States, Saudi Arabia, Ethiopia, Spain, Turkey, Pakistan, and Malaysia have reported severe or extremely severe prevalence rates of depression in university student samples ranging from 9.7% to 27%, state anxiety from 15% to 41.4%, and stress from 5% to 34% [56-62]. These studies, however, were conducted before the COVID-19 pandemic. Since then, several reports have indicated a decline in psychosocial health among undergraduate students, attributed to factors such as social isolation [63], decreased physical activity [64], and uncertainty about the future [65] as a result of the COVID-19 pandemic. Indeed, in a study of students at the same postsecondary institution, our research group found that over half of the students surveyed displayed moderate-to-extremely severe state anxiety, stress, or depression, and that state anxiety and stress worsened in students assessed in 2021 compared to 2020 [7].

Digital mindfulness-based interventions have been shown to improve a variety of psychosocial health indicators in student populations, including depression, stress, anxiety, and general mental well-being [66-70]. Like the findings of this study, in a systematic review and meta-analysis, Chen et al [33] reported that mobile-device-based mindfulness interventions appeared to significantly improve stress and anxiety but not depression among postsecondary students. In contrast, Gong et al [34], Alrashdi et al [36], and Zuo et al [35] found that online mindfulness interventions improved anxiety, stress, as well as depression among university students. Thus, the impact of such mindfulness interventions on depressive symptoms in this population remains unclear. Nevertheless, given that online and app-based meditation tools have shown capacity to improve anxiety and stress severity [32-34,36], the improvements in state anxiety and stress that we observed following Calm app use may be related to the collective content and structure of mindfulness-based apps rather than being attributable specifically to the Calm app itself. This is promising as there may be the potential for developing independent and accessible mobile resources that consider the unique needs of specific postsecondary populations, which can be integrated into the current framework of mental health support available to university students. Nevertheless, further studies are needed to reinforce relationships demonstrated in the extant literature.

Sleep Quality in Undergraduate Students

At baseline, over 80% of the control and treatment group participants exhibited poor sleep quality (indicated by a PSQI global score >5) [39], with a mean PSQI global score of 8.2 and 9.5 in the control and treatment groups, respectively. Over three-quarters of all participants exhibited insomnia (indicated by a PSQI global score ≥6) [50]. A systematic review of 23 studies conducted before the COVID-19 pandemic revealed that undergraduate students typically exhibited poor sleep, with a mean PSQI global score of 6.3 among the studies [71]. This suggests participants in our sample displayed worse baseline sleep quality than what has been reported in the prepandemic literature. Namely, participants displayed insufficient sleep duration (<7 hours of sleep per night), prolonged sleep onset, increased sleep disturbances (such as restlessness), heightened reliance on sleep medication, and increased sleep-related daytime dysfunction. The presence of elevated state depression, anxiety, and stress at baseline in our population may aid in accounting for the poor sleep quality, because evidence suggests that sleep and mental health influence each other bidirectionally [23-26,71]. Experiencing depression, anxiety, and stress during the day may lead to delayed sleep onset and sleep disruption [71]. Poor sleep may then impair one's ability to cope with daytime stressors, and increase maladaptive emotional regulation, further exacerbating the experience of negative emotional states and poor sleep and psychosocial health over time [26,28,29,71]. Ultimately, such relationships highlight the value of interventions that simultaneously target psychosocial health and sleep quality among university students.

Previous research on the impact of mobile mindfulness-based interventions on sleep has yielded comparable results to this study. A recent meta-analysis suggests online mindfulness-based interventions can significantly improve sleep quality among postsecondary students [35]. In 1 study involving adults with sleep disturbances, the Calm app was found to significantly reduce fatigue, daytime sleepiness, and presleep arousal compared to controls [72]. Similarly, physician assistant students using the Happier app showed marked improvements in sleep impairment relative to the control group [73], and a study with university students using the Calm app revealed a negative correlation between overall mindfulness and sleep disturbance [74].

XSL•FO
RenderX

Mobile mindfulness interventions may have a multifaceted impact on student well-being, positively influencing both emotional state health and sleep quality. This finding is particularly notable given the complex and interrelated challenges faced by university students, who frequently experience heightened levels of stress, anxiety, and disrupted sleep. A dual impact is especially valuable in an academic context, where students often encounter difficulties in accessing or committing to traditional mental health resources. Therefore, the integration of mobile mindfulness interventions could offer a promising avenue for enhancing overall student well-being and academic performance, thus contributing to a more holistic approach to mental health support within higher education settings.

Limitations and Future Directions

Our findings should be considered in the context of certain limitations. As our sample consisted of students exclusively from a single university campus in Canada, caution should be applied when generalizing our findings to the broader undergraduate population. Participants were not evenly distributed across demographic categories; there was unequal representation among genders, years of study, and programs of study. The timing of data collection in the fall and winter may have also introduced confounding variables, notably examination periods and seasonal affective disorder (SAD). These potential confounders were not controlled, as the periods in which students are at risk of being impacted by such factors comprise most of the fall and winter terms (September to April of each year). Thus, this was impractical to control in a pilot study. Nevertheless, as these factors have been found to impact undergraduate psychosocial health [75-78], future studies should aim to determine the influence of examination-related stress and SAD on the efficacy of mobile psychosocial health interventions among students. Given the pilot nature of this study, recruitment occurred only during the fall and winter terms (and not the summer term, which occurs from May to August of each year) to maximize the number of potential recruits, as the percentage of full-time students who are taking on-campus classes during the summer term is lower. This study may have also lacked sufficient power due to a small sample size, which may have limited the ability to detect smaller, potentially meaningful differences in app usage patterns, such as the increased time spent in specific sections of the app by the winter group. This could suggest that the positive return on invested in-app time within certain app sections diminishes beyond a minimum threshold. Future studies should aim to recruit larger cohorts during the fall, winter, and summer sessions who are demographically representative of the broader undergraduate population, to aid in better understanding the influence of SAD and other nuances of such mindfulness interventions.

Treatment group adherence must also be considered. Only 54% (20/37) of treatment group participants indicated that they completed the minimum prescribed weekly interaction time with the Calm app. Although this self-reported moderate level of adherence was evidently adequate in impacting psychosocial health and sleep quality, it may have negatively impacted the strength of associations between the treatment and outcome measures. This level of engagement is consistent with previous findings; a systematic review of web- and app-based mindfulness interventions across the general population, including university student cohorts, reported a mean adherence rate of 56% (SD 15%), based on the authors' established definitions of compliance [79]. Strategies to improve adherence, such as text reminders, have been positively received among university students [80], with program reminders being associated with increased adherence [79]. Without a detailed analysis of the control group's measured engagement with alternative coping strategies (nonadherence) or placebo treatment, it remains unclear whether these improvements can be attributed solely to the current treatment. As well, we did not consider whether participants had previous experience with mindfulness app usage; this may have impacted participants' expectations, and future studies should consider accounting for this. If feasible, obtaining access to screen time spent on the app itself, while ensuring confidentiality, would offer a more direct quantification of adherence. Future studies may benefit from implementing engagement strategies such as integrated daily reminders and an active control group.

Given the use of self-reported questionnaires, it must also be noted that self-report biases, including social desirability bias and recall bias, may have impacted our results [81]. This is particularly relevant for the PSQI and PSS-10, which require participants to retrospectively report their sleep habits and feelings of stress over the preceding month [82]. Additionally, variability in bedtimes and wake-up times complicates the accurate reporting of typical sleep patterns and durations [82,83]. The use of a social desirability scale can improve the validity of self-reports in future investigations [81]. Finally, as this is a pilot study with a relatively small sample size, this research is limited by the potential for bias due to incomplete data, as an intent-to-treat analysis was not performed. The small sample size itself also reduces the generalizability of the findings. These results should be validated through larger-scale preregistered clinical trials to confirm their robustness.

Conclusions

The use of the Calm app over a 4-week period resulted in a significant reduction in state anxiety and stress, while also improving the sleep quality in full-time undergraduate students attending a Canadian university post-COVID-19 pandemic. A moderate level of treatment adherence (54%) suggests that even partial engagement with a mindfulness-based mobile app may lead to improvements in state emotional health and sleep quality. Future studies should explore the effects of mindfulness-based apps across larger samples while implementing interventions during summer terms. Direct quantification methods of treatment adherence may also improve our understanding of these relationships. This study highlights the potential value of mindfulness-based apps as a resource to support both the emotional state and sleep quality of undergraduate students.

XSL•FO
RenderX

Acknowledgments

The authors would like to thank all participants for their time and engagement in this study.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to participant confidentiality, but are available from the corresponding author upon reasonable request.

Authors' Contributions

SLW was the principal investigator for this study and led research planning, design, and implementation. TL and ED were responsible for data collection. AA conducted linear mixed methods analysis, and TL conducted additional analyses. TL was responsible for paper drafting. SLW, HEB, and NMD were responsible for supervision, content expertise, and paper editing.

Conflicts of Interest

The authors have no conflicts of interest to declare. The Calm app was not involved in this study in any capacity. No specific funding was obtained for this study, and funds to purchase Calm app access for participants were provided by SLW.

Checklist 1
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File, 1202 KB - formative_v9i1e66131_app1.pdf]

References

1. Auerbach RP, Alonso J, Axinn WG, et al. Mental disorders among college students in the World Health Organization World Mental Health Surveys. Psychol Med 2016 Oct;46(14):2955-2970. [doi: 10.1017/S0033291716001665] [Medline: 27484622]
2. Blanco C, Okuda M, Wright C, et al. Mental health of college students and their non-college-attending peers: results from the National Epidemiologic Study on Alcohol and Related Conditions. Arch Gen Psychiatry 2008 Dec;65(12):1429-1437. [doi: 10.1001/archpsyc.65.12.1429] [Medline: 19047530]
3. Ibrahim AK, Kelly SJ, Adams CE, Glazebrook C. A systematic review of studies of depression prevalence in university students. J Psychiatr Res 2013 Mar;47(3):391-400. [doi: 10.1016/j.jpsychires.2012.11.015] [Medline: 23260171]
4. Stallman HM. Psychological distress in university students: a comparison with general population data. Aust Psychol 2010 Dec 1;45(4):249-257. [doi: 10.1080/00050067.2010.482109]
5. Hamza CA, Ewing L, Heath NL, Goldstein AL. When social isolation is nothing new: a longitudinal study on psychological distress during COVID-19 among university students with and without preexisting mental health concerns. Can Psychol / Psychol Can 2021;62(1):20-30. [doi: 10.1037/cap0000255]
6. Buizza C, Bazzoli L, Ghilardi A. Changes in college students mental health and lifestyle during the COVID-19 pandemic: a systematic review of longitudinal studies. Adolesc Res Rev 2022;7(4):537-550. [doi: 10.1007/s40894-022-00192-7] [Medline: 35966832]
7. Dubale NM, McGinnis E, McEwen B, Bates HE, West SL. State depression, anxiety, stress, and sleep among students attending a Canadian university during the COVID-19 pandemic. Can J Behav Sci / Rev Can Sci Comportement 2024 Jul 25. [doi: 10.1037/cbs0000427]
8. Fuse-Nagase Y. Mental health status among university students during COVID-19 pandemic in Japan: a longitudinal study with one year interval. Asian J Psychiatr 2022 Feb;68:102973. [doi: 10.1016/j.ajp.2021.102973] [Medline: 34953220]
9. Höhne E, Fekete SP, Schild J, Zander L. Perceived stress, individual psychological resources, and social resources among computer science students during the COVID-19 pandemic. Front Educ 2022 Jun 9;7. [doi: 10.3389/feduc.2022.840216]
10. Linden B, Stuart H, Ecclestone A. Trends in post-secondary student stress: a pan-canadian study. Can J Psychiatry 2023 Jul;68(7):521-530. [doi: 10.1177/07067437221111365] [Medline: 35791667]
11. Lorenzo A, Kthupi A, Liu W, et al. The mental health impact of the COVID-19 pandemic on post-secondary students: a longitudinal study. Psychiatry Res 2023 Sep;327:115401. [doi: 10.1016/j.psychres.2023.115401] [Medline: 37567112]
12. McLeish AC, Walker KL, Hart JL. Changes in internalizing symptoms and anxiety sensitivity among college students during the COVID-19 pandemic. J Psychopathol Behav Assess 2022;44(4):1021-1028. [doi: 10.1007/s10862-022-09990-8] [Medline: 35915667]
13. Nomura K, Yamazaki T, Maeda E, et al. Longitudinal survey of depressive symptoms among university students during the COVID-19 pandemic in Japan. Front Psychol 2022;13:863300. [doi: 10.3389/fpsyg.2022.863300] [Medline: 36092090]
14. Weber M, Schulze L, Bolzenkötter T, Niemeyer H, Renneberg B. Mental health and loneliness in university students during the COVID-19 pandemic in Germany: a longitudinal study. Front Psychiatry 2022;13:848645. [doi: 10.3389/fpsyt.2022.848645] [Medline: 35492687]
15. The new abnormal: student mental health two years into COVID-19. : Canadian Alliance of Student Associations; 2022 Sep.

XSL•FO
RenderX

16. Jiang XL, Zheng XY, Yang J, et al. A systematic review of studies on the prevalence of insomnia in university students. Public Health (Fairfax) 2015 Dec;129(12):1579-1584. [doi: 10.1016/j.puhe.2015.07.030] [Medline: 26298588]
17. Brown FC, Soper B. Prevalence of delayed sleep phase syndrome in university students. Coll Stud J 2001;35(3):472 [FREE Full text]
18. Gaultney JF. The prevalence of sleep disorders in college students: impact on academic performance. J Am Coll Health 2010;59(2):91-97. [doi: 10.1080/07448481.2010.483708] [Medline: 20864434]
19. Owens J, Adolescent Sleep Working Group, Committee on Adolescence. Insufficient sleep in adolescents and young adults: an update on causes and consequences. Pediatrics 2014 Sep;134(3):e921-e932. [doi: 10.1542/peds.2014-1696] [Medline: 25157012]
20. Sleep: a neglected public health issue. Lancet Diabetes Endocrinol 2024 Jun;12(6):365. [doi: 10.1016/S2213-8587(24)00132-3]
21. Hysenbegasi A, Hass SL, Rowland CR. The impact of depression on the academic productivity of university students. J Ment Health Policy Econ 2005 Sep;8(3):145-151 [FREE Full text] [Medline: 16278502]
22. Megivern D, Pellerito S, Mowbray C. Barriers to higher education for individuals with psychiatric disabilities. Psychiatr Rehabil J 2003;26(3):217-231. [doi: 10.2975/26.2003.217.231] [Medline: 12653444]
23. Alvaro PK, Roberts RM, Harris JK. A systematic review assessing bidirectionality between sleep disturbances, anxiety, and depression. Sleep 2013 Jul 1;36(7):1059-1068. [doi: 10.5665/sleep.2810] [Medline: 23814343]
24. Dinis J, Bragança M. Quality of sleep and depression in college students: a systematic review. Sleep Sci 2018;11(4):290-301. [doi: 10.5935/1984-0063.20180045] [Medline: 30746048]
25. Yap Y, Slavish DC, Taylor DJ, Bei B, Wiley JF. Bi-directional relations between stress and self-reported and actigraphy-assessed sleep: a daily intensive longitudinal study. Sleep 2020 Mar 12;43(3):zsz250. [doi: 10.1093/sleep/zsz250] [Medline: 31608395]
26. Niu X, Snyder HR. The role of maladaptive emotion regulation in the bidirectional relation between sleep and depression in college students during the COVID-19 pandemic. Anxiety Stress Coping 2023 Jan;36(1):83-96. [doi: 10.1080/10615806.2022.2073441] [Medline: 35536737]
27. Scott AJ, Webb TL, Martyn-St James M, Rowse G, Weich S. Improving sleep quality leads to better mental health: a meta-analysis of randomised controlled trials. Sleep Med Rev 2021 Dec;60:101556. [doi: 10.1016/j.smrv.2021.101556] [Medline: 34607184]
28. O'Leary K, Bylsma LM, Rottenberg J. Why might poor sleep quality lead to depression? A role for emotion regulation. Cogn Emot 2017 Dec;31(8):1698-1706. [doi: 10.1080/02699931.2016.1247035] [Medline: 27807996]
29. Zhang J, Lau EYY, Hsiao JW. Using emotion regulation strategies after sleep deprivation: ERP and behavioral findings. Cogn Affect Behav Neurosci 2019 Apr;19(2):283-295. [doi: 10.3758/s13415-018-00667-y] [Medline: 30460483]
30. Zochil ML, Thorsteinsson EB. Exploring poor sleep, mental health, and help‐seeking intention in university students. Aust J Psychol 2018 Mar 1;70(1):41-47. [doi: 10.1111/ajpy.12160]
31. Cullen M. Mindfulness-based interventions: an emerging phenomenon. Mindfulness (N Y) 2011 Sep;2(3):186-193. [doi: 10.1007/s12671-011-0058-1]
32. Reangsing C, Abdullahi SG, Schneider JK. Effects of online mindfulness-based interventions on depressive symptoms in college and university students: a systematic review and meta-analysis. J Integr Complement Med 2023 May;29(5):292-302. [doi: 10.1089/jicm.2022.0606] [Medline: 36576393]
33. Chen B, Yang T, Xiao L, Xu C, Zhu C. Effects of mobile mindfulness meditation on the mental health of university students: systematic review and meta-analysis. J Med Internet Res 2023 Jan 3;25:e39128. [doi: 10.2196/39128] [Medline: 36596239]
34. Gong XG, Wang LP, Rong G, Zhang DN, Zhang AY, Liu C. Effects of online mindfulness-based interventions on the mental health of university students: a systematic review and meta-analysis. Front Psychol 2023;14:1073647. [doi: 10.3389/fpsyg.2023.1073647] [Medline: 36844353]
35. Zuo X, Tang Y, Chen Y, Zhou Z. The efficacy of mindfulness-based interventions on mental health among university students: a systematic review and meta-analysis. Front Public Health 2023;11:1259250. [doi: 10.3389/fpubh.2023.1259250] [Medline: 38146476]
36. Alrashdi DH, Chen KK, Meyer C, Gould RL. A systematic review and meta-analysis of online mindfulness-based interventions for university students: an examination of psychological distress and well-being, and attrition rates. J Technol Behav Sci 2023;9(2):211-223. [doi: 10.1007/s41347-023-00321-6]
37. Calm app home page. Calm. 2024. URL: https://www.calm.com/app [accessed 2025-05-10]
38. Lovibond PF, Lovibond SH. The structure of negative emotional states: comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories. Behav Res Ther 1995 Mar;33(3):335-343. [doi: 10.1016/0005-7967(94)00075-u] [Medline: 7726811]
39. Buysse DJ, Reynolds CF, Monk TH, Berman SR, Kupfer DJ. The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research. Psychiatry Res 1989 May;28(2):193-213. [doi: 10.1016/0165-1781(89)90047-4] [Medline: 2748771]
40. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. J Health Soc Behav 1983 Dec;24(4):385-396. [doi: 10.2307/2136404] [Medline: 6668417]

XSL•FO
RenderX

41. Antony MM, Bieling PJ, Cox BJ, Enns MW, Swinson RP. Psychometric properties of the 42-item and 21-item versions of the Depression Anxiety Stress Scales in clinical groups and a community sample. Psychol Assess 1998;10(2):176-181. [doi: 10.1037//1040-3590.10.2.176]
42. Crawford JR, Henry JD. The Depression Anxiety Stress Scales (DASS): normative data and latent structure in a large non-clinical sample. Br J Clin Psychol 2003 Jun;42(Pt 2):111-131. [doi: 10.1348/014466503321903544] [Medline: 12828802]
43. Depression Anxiety Stress Scales (DASS). The School of Psychology - UNSW. 2023. URL: https://www2.psy.unsw.edu.au/groups/dass/ [accessed 2025-05-10]
44. Osman A, Wong JL, Bagge CL, Freedenthal S, Gutierrez PM, Lozano G. The Depression Anxiety Stress Scales-21 (DASS-21): further examination of dimensions, scale reliability, and correlates. J Clin Psychol 2012 Dec;68(12):1322-1338. [doi: 10.1002/jclp.21908] [Medline: 22930477]
45. Mohammadzadeh J, Mami S, Omidi K. Mean scores of depression, anxiety and stress in Iranian university students based on DASS-21: a systematic review and meta-analysis. Int J Epidemiol Res 2019;6(1):42-48. [doi: 10.15171/ijer.2019.08]
46. Syed A, Ali SS, Khan M. Frequency of depression, anxiety and stress among the undergraduate physiotherapy students. Pak J Med Sci 2018;34(2):468-471. [doi: 10.12669/pjms.342.12298] [Medline: 29805428]
47. Lovibond SH, Lovibond PF. Manual for the Depression Anxiety Stress Scales, 2nd edition: Psychology Foundation; 1995.
48. Manzar MD, BaHammam AS, Hameed UA, et al. Dimensionality of the Pittsburgh Sleep Quality Index: a systematic review. Health Qual Life Outcomes 2018 May 9;16(1):89. [doi: 10.1186/s12955-018-0915-x] [Medline: 29743066]
49. Mollayeva T, Thurairajah P, Burton K, Mollayeva S, Shapiro CM, Colantonio A. The Pittsburgh sleep quality index as a screening tool for sleep dysfunction in clinical and non-clinical samples: a systematic review and meta-analysis. Sleep Med Rev 2016 Feb;25:52-73. [doi: 10.1016/j.smrv.2015.01.009] [Medline: 26163057]
50. Dietch JR, Taylor DJ, Sethi K, Kelly K, Bramoweth AD, Roane BM. Psychometric evaluation of the PSQI in U.S. college students. J Clin Sleep Med 2016 Aug 15;12(8):1121-1129. [doi: 10.5664/jcsm.6050] [Medline: 27166299]
51. Liu D, Kahathuduwa C, Vazsonyi AT. The Pittsburgh Sleep Quality Index (PSQI): psychometric and clinical risk score applications among college students. Psychol Assess 2021 Sep;33(9):816-826. [doi: 10.1037/pas0001027] [Medline: 34014747]
52. Deng J, Zhou F, Hou W, et al. The prevalence of depressive symptoms, anxiety symptoms and sleep disturbance in higher education students during the COVID-19 pandemic: a systematic review and meta-analysis. Psychiatry Res 2021 Jul;301:113863. [doi: 10.1016/j.psychres.2021.113863] [Medline: 33984824]
53. Perceived Stress Scale. : NH Department of Administrative Services; 2020 URL: https://www.das.nh.gov/wellness/docs/percieved%20stress%20scale.pdf [accessed 2025-05-10]
54. Anwer S, Manzar MD, Alghadir AH, Salahuddin M, Abdul Hameed U. Psychometric analysis of the perceived stress scale among healthy university students. Neuropsychiatr Dis Treat 2020;16:2389-2396. [doi: 10.2147/NDT.S268582] [Medline: 33116538]
55. Lee EH. Review of the psychometric evidence of the perceived stress scale. Asian Nurs Res (Korean Soc Nurs Sci) 2012 Dec;6(4):121-127. [doi: 10.1016/j.anr.2012.08.004] [Medline: 25031113]
56. Lemma S, Gelaye B, Berhane Y, Worku A, Williams MA. Sleep quality and its psychological correlates among university students in Ethiopia: a cross-sectional study. BMC Psychiatry 2012 Dec 28;12:237. [doi: 10.1186/1471-244X-12-237] [Medline: 23270533]
57. Bahhawi TA, Albasheer OB, Makeen AM, et al. Depression, anxiety, and stress and their association with khat use: a cross-sectional study among Jazan University students, Saudi Arabia. Neuropsychiatr Dis Treat 2018;14:2755-2761. [doi: 10.2147/NDT.S182744] [Medline: 30425493]
58. Shamsuddin K, Fadzil F, Ismail WSW, et al. Correlates of depression, anxiety and stress among Malaysian university students. Asian J Psychiatr 2013 Aug;6(4):318-323. [doi: 10.1016/j.ajp.2013.01.014] [Medline: 23810140]
59. Asif S, Mudassar A, Shahzad TZ, Raouf M, Pervaiz T. Frequency of depression, anxiety and stress among university students. Pak J Med Sci 2020;36(5):971-976. [doi: 10.12669/pjms.36.5.1873] [Medline: 32704273]
60. Bayram N, Bilgel N. The prevalence and socio-demographic correlations of depression, anxiety and stress among a group of university students. Soc Psychiatry Psychiatr Epidemiol 2008 Aug;43(8):667-672. [doi: 10.1007/s00127-008-0345-x] [Medline: 18398558]
61. Ramón-Arbués E, Gea-Caballero V, Granada-López JM, Juárez-Vela R, Pellicer-García B, Antón-Solanas I. The prevalence of depression, anxiety and stress and their associated factors in college students. Int J Environ Res Public Health 2020 Sep 24;17(19):7001. [doi: 10.3390/ijerph17197001] [Medline: 32987932]
62. Beiter R, Nash R, McCrady M, et al. The prevalence and correlates of depression, anxiety, and stress in a sample of college students. J Affect Disord 2015 Mar 1;173:90-96. [doi: 10.1016/j.jad.2014.10.054] [Medline: 25462401]
63. Rasmussen B, Hutchinson A, Lowe G, et al. The impact of covid-19 on psychosocial well-being and learning for australian nursing and midwifery undergraduate students: a cross-sectional survey. Nurse Educ Pract 2022 Jan;58:103275. [doi: 10.1016/j.nepr.2021.103275] [Medline: 34922092]
64. Wilson OWA, Holland KE, Elliott LD, Duffey M, Bopp M. The impact of the COVID-19 pandemic on US college students' physical activity and mental health. J Phys Act Health 2021 Mar 1;18(3):272-278. [doi: 10.1123/jpah.2020-0325] [Medline: 33601332]

XSL•FO
RenderX

65. Oral M, Karakurt N. The impact of psychological hardiness on intolerance of uncertainty in university students during the COVID-19 pandemic. J Community Psychol 2022 Sep;50(8):3574-3589. [doi: 10.1002/jcop.22856] [Medline: 35415919]
66. Lee RA, Jung ME. Evaluation of an mHealth app (DeStressify) on university students' mental health: pilot trial. JMIR Ment Health 2018 Jan 23;5(1):e2. [doi: 10.2196/mental.8324] [Medline: 29362209]
67. Davies EB, Morriss R, Glazebrook C. Computer-delivered and web-based interventions to improve depression, anxiety, and psychological well-being of university students: a systematic review and meta-analysis. J Med Internet Res 2014 May 16;16(5):e130. [doi: 10.2196/jmir.3142] [Medline: 24836465]
68. Malinauskas R, Malinauskiene V. Meta-analysis of psychological interventions for reducing stress, anxiety, and depression among university students during the COVID-19 pandemic. Int J Environ Res Public Health 2022 Jul 27;19(15):9199. [doi: 10.3390/ijerph19159199] [Medline: 35954553]
69. El Morr C, Ritvo P, Ahmad F, Moineddin R, MVC Team. Effectiveness of an 8-week web-based mindfulness virtual community intervention for university students on symptoms of stress, anxiety, and depression: randomized controlled trial. JMIR Ment Health 2020 Jul 17;7(7):e18595. [doi: 10.2196/18595] [Medline: 32554380]
70. Huberty J, Green J, Glissmann C, Larkey L, Puzia M, Lee C. Efficacy of the mindfulness meditation mobile app "Calm" to reduce stress among college students: randomized controlled trial. JMIR mHealth uHealth 2019 Jun 25;7(6):e14273. [doi: 10.2196/14273] [Medline: 31237569]
71. Gardani M, Bradford DRR, Russell K, et al. A systematic review and meta-analysis of poor sleep, insomnia symptoms and stress in undergraduate students. Sleep Med Rev 2022 Feb;61:101565. [doi: 10.1016/j.smrv.2021.101565] [Medline: 34922108]
72. Huberty JL, Green J, Puzia ME, et al. Testing a mindfulness meditation mobile app for the treatment of sleep-related symptoms in adults with sleep disturbance: a randomized controlled trial. PLoS ONE 2021;16(1):e0244717. [doi: 10.1371/journal.pone.0244717] [Medline: 33411779]
73. Smith JL, Allen JW, Haack CI, et al. Impact of app-delivered mindfulness meditation on functional connectivity, mental health, and sleep disturbances among physician assistant students: randomized, wait-list controlled pilot study. JMIR Form Res 2021 Oct 19;5(10):e24208. [doi: 10.2196/24208] [Medline: 34665153]
74. Glissmann C. Calm college: testing a brief mobile app meditation intervention among stressed college students. : Arizona State University; 2018. [PMC: 29995981]
75. Tonetti L, Barbato G, Fabbri M, Adan A, Natale V. Mood seasonality: a cross-sectional study of subjects aged between 10 and 25 years. J Affect Disord 2007 Jan;97(1-3):155-160. [doi: 10.1016/j.jad.2006.06.008] [Medline: 16860875]
76. Melrose S. Seasonal affective disorder: an overview of assessment and treatment approaches. Depress Res Treat 2015;2015:178564. [doi: 10.1155/2015/178564] [Medline: 26688752]
77. Fritz J, Stochl J, Kievit RA, van Harmelen AL, Wilkinson PO. Tracking stress, mental health, and resilience factors in medical students before, during, and after a stress-inducing exam period: protocol and proof-of-principle analyses for the RESIST cohort study. JMIR Form Res 2021 Jun 8;5(6):e20128. [doi: 10.2196/20128] [Medline: 34100761]
78. Turner L, Galante J, Vainre M, Stochl J, Dufour G, Jones PB. Immune dysregulation among students exposed to exam stress and its mitigation by mindfulness training: findings from an exploratory randomised trial. Sci Rep 2020 Apr 2;10(1):5812. [doi: 10.1038/s41598-020-62274-7] [Medline: 32242145]
79. Winter N, Russell L, Ugalde A, White V, Livingston P. Engagement strategies to improve adherence and retention in web-based mindfulness programs: systematic review. J Med Internet Res 2022 Jan 12;24(1):e30026. [doi: 10.2196/30026] [Medline: 35019851]
80. Moffitt-Carney KM, Duncan AB. Evaluation of a mindfulness-based mobile application with college students: a pilot study. J Am Coll Health 2021;69(2):208-214. [doi: 10.1080/07448481.2019.1661420] [Medline: 31513476]
81. van de Mortel TF. Faking it: social desirability response bias in self-report research. Aust J Adv Nurs 2008;25(4):40-48. [doi: 10.37464/2008.254.1817]
82. Kreutz C, Müller J, Schmidt ME, Steindorf K. Comparison of subjectively and objectively assessed sleep problems in breast cancer patients starting neoadjuvant chemotherapy. Support Care Cancer 2021 Feb;29(2):1015-1023. [doi: 10.1007/s00520-020-05580-0] [Medline: 32556623]
83. Lovibond SH, Lovibond PF. Depression Anxiety Stress Scales (DASS--21, DASS--42): APA PsycTests; 1995. [doi: 10.1037/t01004-000]

Abbreviations

DASS-42: Depression Anxiety Stress Scale 42-Item Version
PSQI: Pittsburgh Sleep Quality Index
PSS-10: Perceived Stress Scales 10-Item Version
SAD: seasonal affective disorder

XSL•FO
RenderX

Edited by A Mavragani; submitted 04.09.24; peer-reviewed by T Neher, V Sicard; revised version received 08.04.25; accepted 10.04.25; published 11.06.25.

Please cite as:
Lew T, Dubale NM, Doose E, Adenuga A, Bates HE, West SL
Impacts of the Mindfulness Meditation Mobile App Calm on Undergraduate Students' Sleep and Emotional State: Pilot Randomized Controlled Trial
JMIR Form Res 2025;9:e66131
URL: https://formative.jmir.org/2025/1/e66131
doi:10.2196/66131

© Tovan Lew, Natnaiel M Dubale, Erik Doose, Alex Adenuga, Holly E Bates, Sarah L West. Originally published in JMIR Formative Research (https://formative.jmir.org), 11.6.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Using Ecological Momentary Assessment to Assess Family Functioning in Spanish-Speaking Parent and Adolescent Dyads: Daily Questionnaire Study

Alejandra Fernandez[1], PhD; Savannah Bernal[2], BS; Lana Kim[3], BS; Subodh Potla[4], BS

[1]Department of Social and Behavioral Sciences, Peter O'Donnell Jr. School of Public Health, The University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd, Dallas, United States

[2]Department of Health Promotion and Behavior Sciences, UTHealth Houston School of Public Health, Dallas, TX, United States

[3]Department of Integrative Biology, Rice University, Houston, TX, United States

[4]Deparment of Epidemiology, UTHealth Houston School of Public Health, Dallas, TX, United States

Corresponding Author:
Alejandra Fernandez, PhD
Department of Social and Behavioral Sciences, Peter O'Donnell Jr. School of Public Health, The University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd, Dallas, United States

Abstract

Background: Family functioning is associated with several adolescent health outcomes, and many family-based interventions (FBIs) exist to improve family functioning. However, most FBIs assess family functioning retrospectively at baseline and post intervention, thereby overlooking the daily fluctuations in family functioning throughout the intervention. Ecological momentary assessment (EMA) is a method involving a high frequency of assessments and has been underused to assess family functioning across parent and adolescent dyads. Further, limited research exists on the use of EMA in bilingual populations.

Objective: The purpose of this study was to assess an EMA protocol's feasibility and acceptability and to analyze within-person and between-person variance in family functioning reports in a sample of primarily Spanish-speaking parent and adolescent dyads.

Methods: Participants completed a baseline assessment (including demographics and family functioning assessment), a 7-day protocol with a once-daily family assessment questionnaire using an EMA app, and an acceptability questionnaire at the conclusion of the study.

Results: We recruited 7 mothers (mean age 37.29, SD 3.82 years) and 8 adolescents (n=7, 88% females; mean age 11.86, SD 1.07 years) who identified themselves as Hispanic/Latinx. The participants showed overall satisfaction with the EMA protocol. The daily assessments were completed relatively quickly (mean 3 minutes and 16 seconds, SD 11 minutes and 5 seconds) after the prompt notification was received, and the response rate across the daily assessments was 90% (87/97). The reported family functioning was relatively high across both adolescents (mean 4.57) and parents (mean 4.59). The variance across adolescents (SD 0.459) was larger than that within their individual reports of family functioning (SD 0.122). Alternatively, the variance across parents was smaller (SD 0.132) than that reported among parents' individual reports of family functioning (SD 0.286). Our findings highlight the heterogeneity between adolescent and parent responses. Finally, the visual inspection of data underscored the individualized patterns and reported differences in the family functioning reports across parents and adolescents.

Conclusion: Our findings emphasize the value of EMA in studying family (eg, adolescent-caregiver) behaviors. EMA's ability to capture immediate experiences presents a nuanced picture of daily interactions and offers suggestions for practice when using the EMA methodology in populations such as the one included in this study (ie, primarily Spanish-speaking parent-adolescent dyads).

(JMIR Form Res 2025;9:e60073) doi:10.2196/60073

KEYWORDS

adolescence; ecological momentary assessment; family functioning; family function; feasibility; acceptability; adolescent; family-based; family-based interventions; EMA; community-based; smartphone

XSL•FO
RenderX

Introduction

Background

Family-based interventions (FBIs) are effective tools for targeting several adolescent behavioral outcomes [1]. The mechanism underlying the effectiveness of FBIs is the improvement in family functioning behaviors such as parent-adolescent communication, family cohesion, and parental monitoring [2]. For example, families exhibiting strong family functioning behaviors tend to have a decreased likelihood of adolescents participating in harmful behaviors [3]. Oftentimes, FBIs target family functioning behaviors by engaging parents or primary caregivers in individual parent sessions, where parents or primary caregivers can learn healthy parenting skills, and in dyadic parent-child sessions, where they can practice these skills [4]. The effects of the intervention on family functioning are then assessed at intermittent times (eg, baseline, that is, before intervention implementation; after the intervention; 3 months post intervention; 6 months post intervention). Few FBIs measure and evaluate family functioning behaviors in a daily (momentary) manner [5]. Rather, most FBIs utilize retrospective measurements, which limit the ability to capture data that indicate when targeted messaging tailored to participants may be needed through FBIs [5].

Ecological momentary assessment (EMA) is a collection of methods involving a high frequency of assessments (eg, receiving one or more assessments each day) that asks participants to report on individual, interpersonal, or contextual factors in real time, thereby minimizing recall bias and maximizing ecological validity [6,7]. Additionally, EMA encompasses a variety of approaches, including daily diaries [7]. EMA has been implemented with adolescents as young as 9 years by using smartphone apps and SMS text messaging [8-10]. The advantages of EMA over retrospective assessments include a higher level of temporal detail that captures changes in real time, ecological validity, and reliability [7]. Additionally, constructs assessed through EMA are often dynamic and can change quickly, and it is unfeasible to use face-to-face or in-person approaches to identify optimal opportunities to intervene [11], for example, based on changing family dynamics.

Of the existing research using EMA in the context of family functioning, most are focused on studying behavior fluctuations and understanding dynamic behavioral variations [12]. Experts in the field argue that very little is known regarding daily parent-adolescent interactions [13]. Although limited, current evidence suggests that EMA is feasible to use with early adolescents [14]; however, relatively few studies include both parent and adolescent reports when applying an EMA approach [15], with even fewer assessing family functioning behaviors [16,17]. Thus, more EMA-focused research is needed to shed light on the variability in family functioning across parents and adolescents. Capturing the dynamic changes in family functioning via EMAs could lead to adaptive intervention messaging based on daily reported family functioning, which in turn influences parent-adolescent family functioning and ultimately, adolescent behavioral outcomes.

An additional limitation of existing EMA research is the limited evidence available on EMA protocols being implemented and tested for feasibility and acceptability in Spanish-speaking (monolingual and bilingual) populations (ie, English and Spanish) [18]. One previous study using EMA methodology assessed suicidal thoughts among Spanish-speaking Hispanic/Latino adults and indicated high adherence (74.5% of EMAs completed) in their sample [19]; however, that study provided participants with smartphones to complete EMAs, possibly influencing their high adherence. Further, other studies with Spanish-speaking participants have shown relatively lower adherence (40%) [20]. Existing EMA research has made recommendations related to the implementation of EMA protocols with special populations [14], including adaptations of assessment protocols. The EMA platform technology may also need to adapt to multiple languages and EMA protocols, including protocol instructions and assessments, to have better acceptability if available in the participants' preferred language. Therefore, more studies are needed that include Spanish-speaking parents and adolescents and that describe barriers and facilitators to protocol adherence.

Objectives

By utilizing an approach that investigates family functioning behaviors using the EMA protocol and framework, we intend to evaluate models to develop stronger tools for FBIs. This study has 3 aims: (1) to examine the feasibility and acceptability of the developed EMA protocol, (2) to examine the within- and between-variance in daily reports of family functioning, and (3) to visually inspect changes across daily family functioning for each participant.

Methods

Study Participants

Adolescent inclusion criteria were (1) Hispanic/Latino/Latina/Latinx descent, (2) access to a smartphone, (3) having either English or Spanish proficiency, and (4) living most of the time with the participating parent who provided consent. Parents provided informed consent, and adolescents provided assent. All recruited participants remained in the study throughout its entirety.

Participant Recruitment

This study uses convenience sampling via dissemination of recruitment study materials to community-based organizations that primarily serve minority populations in the Dallas/Fort Worth area. Study participants were not recruited via a clinical setting, and recruitment was primarily community-based. Interested participants completed a recruitment survey and provided their contact information. Study personnel reached out to interested individuals to set up a virtual meeting to determine study eligibility, consent, and assent, and to describe study procedures (eg, use of the EMA smartphone app, compensation structure). The eligibility criteria for the study were (1) adolescent aged between 10 and 17 years, (2) reported Hispanic/Latinx descent, (3) resided in Dallas county, (4) had a primary caregiver who was also willing to participate, (5) both adolescent and primary caregiver had access to their own

XSL•FO
RenderX

smartphones, and (6) read and understood either English or Spanish. Participants were enrolled in the study on a rolling basis.

Procedures

During the initial virtual meeting, both parent and adolescent participants downloaded the LifeData smartphone app (an existing app platform that was used to deliver the daily EMAs, available via Samsung or Apple app store for free) to their personal smartphone and viewed a training video on how to complete daily EMAs. The LifeData app delivered daily assessments via in-time notifications. The delivery of daily assessments began the day after the initial virtual meeting. Thereafter, once per day, for a week (7 days, that is, 5 weekdays and 2 weekend days), participants were asked to complete a daily assessment of family function behaviors related to that day's (ie, the day the assessment was delivered) parent-adolescent behaviors. Only one wave of data was collected. The daily assessment was relatively brief (ie, 11 questions) to reduce participant burden, and all items were available in either English or Spanish based on the participant's preference. Both parent and adolescent participants were advised to complete the daily assessment independently. We planned to deliver one prompt per day. If no response was received after the initial prompt notification, a second automated notification was sent 15 minutes after the first to elicit a response, with 2 additional notifications in 15-minute increments sent if no response was received. If no response was received after the notifications, the assessment for that day was considered incomplete. Participant data were reviewed by study personnel daily, and if a participant appeared to have incomplete data, study personnel followed up with the participant to determine any barrier to completion. After the 7-day EMAs, participants were sent a post-EMA, consisting of open-ended questions on their experience completing the daily assessments. Participants were compensated using online gift cards. For the initial virtual visit, parents and adolescents (ie, participants) received a US $15 gift card each. Further, participants received US $3 for each daily assessment completed. Finally, participants received US $15 for completing the post-EMA. Study data were collected and managed using the Research Electronic Data Capture (REDCap) tools hosted at UTHealth Houston [21,22].

Measures

Demographics

Parents reported the sex of the child, caregiver's relationship to the child, age of the child, ethnicity, Hispanic origin, race, nationality, years lived in the United States, caregiver's age, marital status, education, working status, and household composition.

Acceptability of the EMA Protocol

Both parents and adolescents reported on the acceptability of the EMA protocol by responding to several questions (both multiple-response option and open-ended questions) related to the ease of use of the smartphone app, technical difficulties experienced, understandability of questions, and so forth.

Feasibility of the EMA Protocol

The feasibility of the EMA protocol was assessed by examining the average amount of time taken to complete the daily assessments, the number of daily assessments completed on average across participants, and the number of reminders delivered.

Daily Family Functioning

For the daily assessments, items were adapted from existing family functioning measures (most were adapted from measures used in a large national cohort study [23]) to be brief and represent various dimensions of family functioning, including communication, warmth, and monitoring. The original items and adapted items are shown in Multimedia Appendix 1. Participants were to respond based on the respective day the daily assessment was delivered. Items were worded differently toward parent and adolescent participants. For example, parents received items such as "I knew where my child was today," while adolescents received items such as "My parent(s) knew where I was today." There was a total of 11 items asked during the daily assessments; however, for this paper, only 9 items with the same Likert scale response option were used to create an aggregate score. Response options for the 9 included items ranged from 1=strongly disagree to 5=strongly agree. Items were averaged, with higher scores indicating better perceived family functioning (Cronbach α=0.824).

Analytical Plan

To address aim 1, we used descriptive statistics (eg, mean, standard deviation) to assess the feasibility and acceptability data. To address aim 2, we used the *xtsum* function in R Studio (Posit Software, PBC) to examine the overall, between-, and within-variance for parent and adolescent reports of daily family functioning. Additionally, we calculated the coefficient of variation (CV) for both adolescent and parent daily family functioning to enable comparison to previous reports. Finally, to address aim 3, we examined individual and difference score plots of parent and adolescent daily family functioning by using the *ggplot* function in R Studio.

Ethical Considerations

All study procedures were approved by the UTHealth Houston Committee for the Protection of Human Subjects (HSC-SPH-23-0129). Parents were asked to consent to study procedures, and adolescents were asked to assent. Consent procedures included a virtual visit with both parents and adolescents, where a link to the consent and assent forms was shared. Parents and adolescents were then asked to read the consent or assent form (in their preferred language, either English or Spanish) and to ask questions, if any, which the study personnel could answer. All study data were deidentified prior to analyses.

Results

The participants in this study were 8 adolescents (7/8, 88% females; mean age 11.86, SD 1.07 years) and 7 mothers (mean age 37.29, SD 3.82 years). Additional participant information is found in Table 1.

XSL•FO
RenderX

Table . Demographics of the parents (n=7) and adolescents (n=8) in this study.

	Value
Sex of child, n (%)	
Male	1 (12)
Female	7 (88)
Age of child (years), mean (SD)	11.86 (1.07)
Parent's relationship to child, n (%)	
Mother	7 (100)
People in the household, mean (SD)	4.71 (2.21)
Parent's age (years), mean (SD)	37.29 (3.82)
Parent Hispanic status, n (%)	7 (100)
Parent Hispanic origin, n (%)	
Mexican	5 (71)
Other Latin American or Hispanic	2 (29)
Race, n (%)	
White	3 (43)
Other race	1 (14)
Refuse to answer	3 (43)
Parent born in the United States, n (%)	
No	7 (100)
Yes	0 (0)
Parent years in the United States, mean (SD)	11.0 (6.63)
Marital status, n (%)	
Married	6 (86)
Living with a partner	1 (14)
Parent education, n (%)	
Less than high school diploma	2 (29)
High school diploma or equivalent	2 (29)
Some college or some type of degree	2 (29)
Don't know	1 (14)

Acceptability

Most participants rated the overall experience with the EMA protocol as either positive (4/15, 27%) or very positive (10/15, 67%). Further, all participants rated the use of the LifeData app as either easy or very easy. Most participants felt positive (5/15, 33%) or very positive (9/15, 60%) about the questions asked as part of the daily assessments and the understandability of the questions. When asked about the burden of completing the daily assessments, most rated the burden as somewhat high or high (8/15, 53%). Some participants reported experiencing problems when receiving notifications (3/15, 20%). Finally, most participants reported that they felt that they received the right number of notifications (12/15, 80%) and at the right time (13/15, 87%). When asked to provide any additional comments in the form of open-ended questions, some participants stated that the questions became repetitive across the 7-day assessment period, the daily assessments reminded parents to communicate with their children, and the overall experience of the daily assessments was positive.

Feasibility

On average, participants spent 3 minutes and 16 seconds (SD 11 minutes and 5 seconds) completing the daily assessment once they started the survey. The day-level response rate across the 97 daily assessment notifications sent (including all the participants) was 90% (n=87). Finally, an average of 1.89 (SD 1.79) reminders were sent to remind participants to complete the daily assessment. A large portion of the participants completed the daily assessment without needing a reminder notification sent (40/97, 41%), with adolescents more likely to complete the daily assessment without a reminder (29/53, 55% of adolescents vs 11/44, 25% of parents). When asked to provide additional comments in the form of open-ended questions regarding barriers to completing the daily assessments, some participants indicated that they experienced problems with their

XSL•FO
RenderX

internet connection, not receiving daily assessment notifications, having to download the smartphone app several times, and having to check the smartphone app to see if there were any incomplete daily assessments due to the lack of notifications received. Most barriers to completion were related to technical difficulties with the smartphone app.

Daily Family Functioning Variance

The reported standard deviations in Table 2 indicate that the variations in daily family functioning across adolescents (SD 0.459) are similar to the mean and standard deviations (mean 4.568, SD 0.441) observed across all person-days that had a survey response. However, when examining within-adolescent differences, adolescents were relatively more consistent in their reports of daily family functioning (SD 0.122) than when compared to the reports of the entire adolescent sample. Alternatively, across or between parents, the reported standard deviation indicated that daily family functioning was relatively more consistent (SD 0.132) compared to daily family functioning within individual parent reports of family functioning (SD 0.286), which was similar to the overall mean and standard deviation across all parents (mean 4.587, SD 0.311). The results for CV indicate a relatively low coefficient (ie, <0.10), indicating low variability and more consistent reports of family functioning for adolescents (CV=0.097) and parents (CV=0.068), with parent reports being more closely clustered around the mean when compared to adolescents. Fluctuations in the daily reports of family functioning across parents and adolescents are shown in Figure 1. Additionally, Figure 2 plots the difference scores between adolescent and parent dyads on their daily reported family functioning. Negative scores indicate that adolescents reported lower levels of family functioning compared to their parents.

Table . Overall, within-, and between-variance and coefficient of variation for parent and adolescent daily family functioning.

Adolescent or parent, variance	Value	Coefficient of variation
Adolescent daily family functioning		0.097
Overall, mean (SD)	4.568 (0.411)	
Between, SD	0.459	
Within, SD	0.122	
Parent daily family functioning		0.068
Overall, mean (SD)	4.587 (0.311)	
Between, SD	0.132	
Within, SD	0.286	

XSL•FO
RenderX

Figure 1. Variance in the daily family functioning across parents (n=7) and adolescents (n=8). Missing data points are daily assessments without response. Each graph represents a participant ID. Adol: adolescent.

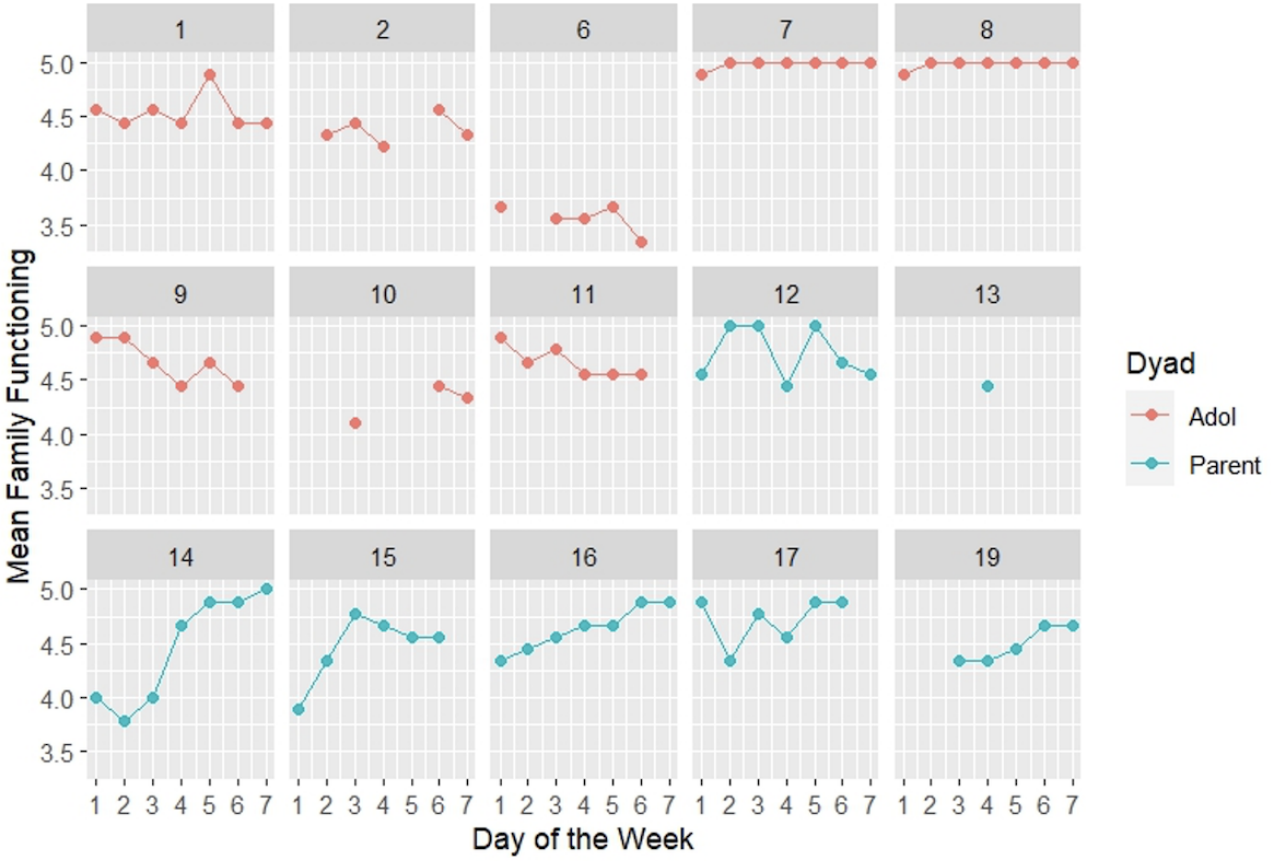

XSL•FO
RenderX

Figure 2. Dyadic difference scores (adolescent-parent) across 7 days. Missing points indicate that the adolescent or the parent or both had missing data for the day. Negative values indicate adolescent-reported lower levels of family functioning compared to their parents.

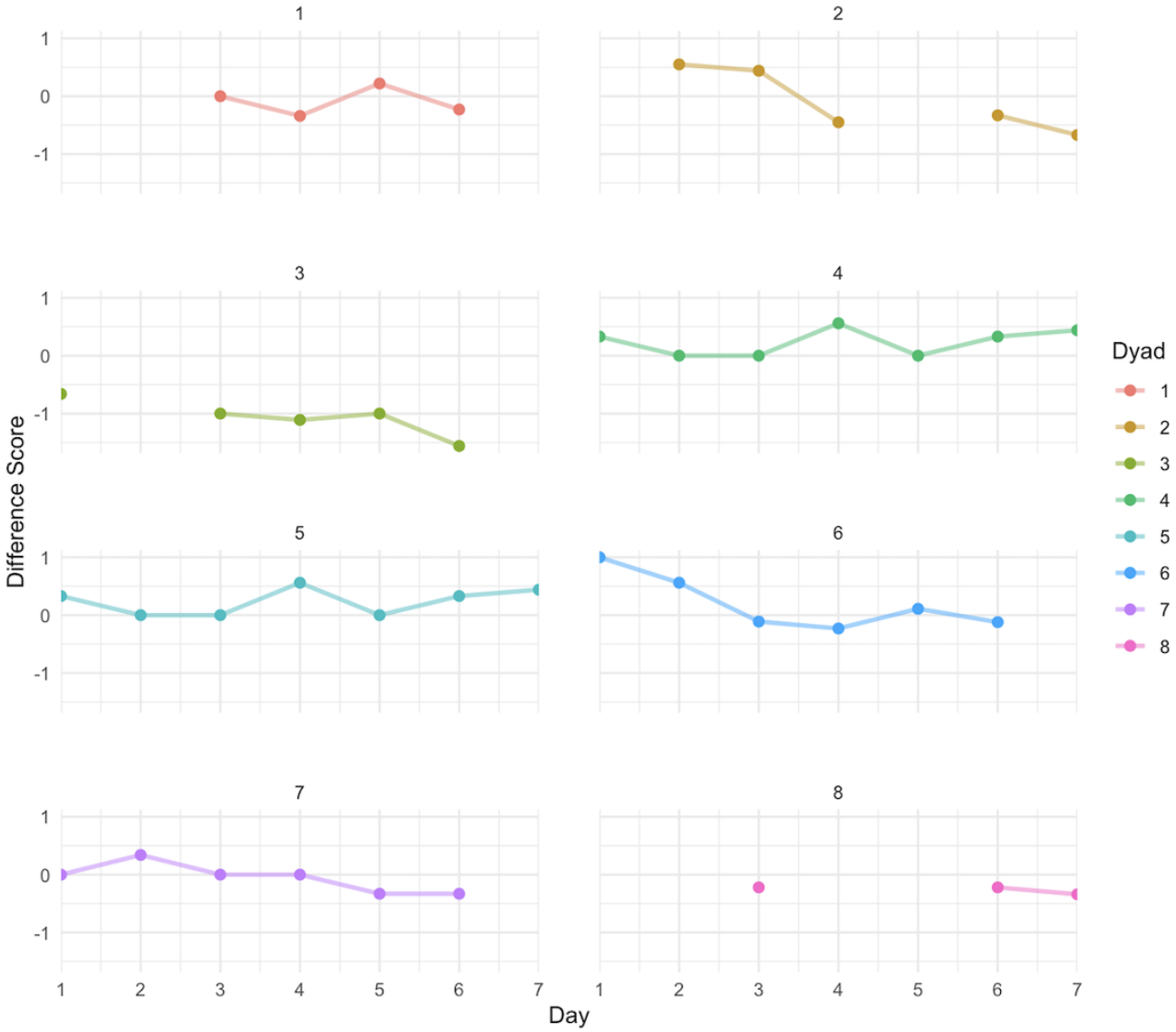

Discussion

Principal Findings

Our study is one of the few studies that has implemented an EMA protocol in a bilingual (ie, primarily Spanish speaking) population and has examined family functioning across parent and adolescent dyads. The purpose of our study was to assess the feasibility and acceptability of the EMA protocol in parent and adolescent dyads, examine the within- and between-variance in daily reports of family functioning, and assess changes across daily family functioning for each participant. Our findings show preliminary acceptability and feasibility among adolescents and parents in the implementation of the EMA protocol to assess family functioning; however, technical barriers related to the smartphone app hindered data collection. Additionally, findings indicated minimal variance in the reports of family functioning within participant reports.

Most participants showed a high level of acceptability toward the use of the LifeData app to administer the EMA family functioning protocol. However, it is important to note that the protocol was delivered once a day at the same time each day for a week (ie, 7 PM CST). This may have increased the burden among the participants, particularly if the set time was inconvenient for the participants. Previous EMA research with parents and adolescents implemented random notification triggers throughout a particular day [16], which may have lessened the monotony of the questionnaires and increased the chances of response across multiple time points. While considering the reported burden, overall, participants reported acceptability of the EMA protocol. Additionally, qualitatively and in response to open-ended questions, parents reported that the EMA questions served as a reminder to communicate with their children. Although previous research suggests that repeated EMA does not significantly alter behaviors [24], it may serve as a subtle prompt for parents to engage with their children [25], even if this effect is statistically minor or short-lived.

In terms of feasibility, the time to completion of the EMA family functioning protocol across participants was less than 5 minutes, and compliance was approximately 90% (87/97). These findings are consistent with those of existing EMA studies, including those implemented among youth [14,24]. A critique of EMA

XSL•FO
RenderX

is the feasibility of implementing it among special populations (eg, early adolescents), with recommendations including providing proper training on EMA procedures, considering participant capabilities, adapting survey language, and scheduling accommodations, among other recommendations [14]. We applied these recommendations in our study, which may have played a role in the high compliance, while also taking into consideration the bilingual nature of our study sample. Our entire parent sample preferred assessments in Spanish, which we prepared by translating and backtranslating our assessments and providing our training materials, including our one-on-one virtual training session and training video in Spanish. Future studies implementing EMA protocols among diverse populations should make necessary adaptations to assessments as well.

In an assessment of between-participant and within-participant variance in reported family functioning, our findings suggest that the variability in reported daily family functioning varied across parents and adolescents. Although adolescents were uniform in their individual (within) reports of family functioning, parents had relatively more variation. It is possible that due to the nature of the early adolescent age (mean 11.9 years), adolescents may report consistent family functioning, as conflict may not manifest between parents and adolescents until later in adolescence when adolescents begin to explore their autonomy [26,27]. Additionally, it is important to consider the EMA protocol length (ie, 7 days). Perhaps, if the assessment length had been extended to a longer period (eg, 30 days), there may have been more variability with a probability of external factors occurring, such as academic circumstances or both immediate and extended family conflict. In the future, we intend to assess parents and adolescents by using a measurement burst design such that an EMA protocol is implemented once a month for 7 days across multiple months to allow for a more in-depth snapshot of parent and adolescent family functioning experiences.

Limitations

Although this study provides formative evidence for the continued testing and development of an EMA protocol assessing daily family functioning, there are some limitations to consider. First, the sample size for the analysis was small (n=15; 7 mothers and 8 adolescents). The small sample size limited the variability across and between participants that may have appeared with a larger sample size. Second, most participants were recruited via a local community organization offering parenting classes, which may have influenced how the daily reported family functioning data were skewed toward more favorable reports of family functioning (skewness: adolescent=−1.16; parent=−0.77), and reports may be higher than those reported from the general public. The parents and adolescents recruited may already present better family functioning because of their potential attendance in community-based class involvement. Finally, the EMA daily assessment items were delivered at the same time (7 PM) every day on the 7 study days, which may impact recall bias. Although this was a potential limitation, our aims were primarily to assess the acceptability of the protocol (eg, English language) and the feasibility (eg, use of own smartphone for LifeData app download and use). This was our reasoning for choosing a time that would appear to be convenient for both adolescents (ie, at home from school) and parents (ie, at home from work).

Implications

The adoption of EMA in this study highlights its practicality for real-time observation of family functioning. The variability in daily familial interactions, although relatively small, may point toward the utility of EMA in personalizing FBIs, offering a pathway for tailored strategies that align with the fluctuating dynamics of daily life. We plan to use the lessons learned in this study to fine-tune our final EMA protocol and implement within culturally responsive interventions for diverse adolescent populations and their families. Finally, these findings advocate for further research to explore longitudinal applications of EMA, potentially enhancing adolescent health outcomes through informed, dynamic support systems.

Acknowledgments

We would like to acknowledge the adolescents and parents who participated in this study. AF is supported by funding from the National Heart, Lung, and Blood Institute (grant 1K01HL166439-01A1). We did not use any generative artificial intelligence for any portion of this manuscript.

Data Availability

The datasets generated or analyzed and the code generated during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: AF
Data curation: AF
Formal analysis: AF (lead), SB (supporting), LK (supporting)
Funding acquisition: AF
Investigation: AF (lead), SB (supporting)
Methodology: AF (lead), SB (supporting)

XSL•FO
RenderX

Project administration: AF
Resources: AF
Software: AF
Supervision: AF
Validation: AF
Visualization: AF (lead), LK (supporting), SP (supporting)
Writing – original draft: AF (lead), SB (supporting), LK (supporting)
Writing – review & editing: AF (lead), SB (supporting), LK (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Original and adapted measures used in the ecological momentary assessment of daily family functioning.
[DOCX File, 16 KB - formative_v9i1e60073_app1.docx]

References

1. Hale DR, Fitzgerald-Yau N, Viner RM. A systematic review of effective interventions for reducing multiple health risk behaviors in adolescence. Am J Public Health 2014 May;104(5):e19-e41. [doi: 10.2105/AJPH.2014.301874] [Medline: 24625172]
2. Hsieh HF, Mistry R, Kleinsasser MJ, et al. Family functioning within the context of families with adolescent children in urban India. Fam Process 2023 Mar;62(1):287-301. [doi: 10.1111/famp.12784] [Medline: 35638112]
3. Cordova D, Heinze J, Mistry R, et al. Family functioning and parent support trajectories and substance use and misuse among minority urban adolescents: a latent class growth analysis. Subst Use Misuse 2014 Dec;49(14):1908-1919. [doi: 10.3109/10826084.2014.935792] [Medline: 25033377]
4. Sturge-Apple ML, Davies PT, Cummings EM. Typologies of family functioning and children's adjustment during the early school years. Child Dev 2010;81(4):1320-1335. [doi: 10.1111/j.1467-8624.2010.01471.x] [Medline: 20636698]
5. Prado G, Fernandez A, St George SM, et al. Results of a family-based intervention promoting healthy weight strategies in overweight Hispanic adolescents and parents: an RCT. Am J Prev Med 2020 Nov;59(5):658-668. [doi: 10.1016/j.amepre.2020.06.010] [Medline: 33011010]
6. Smyth JM, Heron KE. Ecological momentary assessment (EMA) in family research. In: Emerging Methods in Family Research: Springer International Publishing; 2015:145-161. [doi: 10.1007/978-3-319-01562-0_9]
7. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol 2008;4:1-32. [doi: 10.1146/annurev.clinpsy.3.022806.091415] [Medline: 18509902]
8. Dunton GF, Intille SS, Wolch J, Pentz MA. Children's perceptions of physical activity environments captured through ecological momentary assessment: a validation study. Prev Med 2012 Aug;55(2):119-121. [doi: 10.1016/j.ypmed.2012.05.015] [Medline: 22659225]
9. Dunton GF, Liao Y, Intille S, Wolch J, Pentz MA. Physical and social contextual influences on children's leisure-time physical activity: an ecological momentary assessment study. J Phys Act Health 2011 Jan;8 Suppl 1:S103-S108. [doi: 10.1123/jpah.8.s1.s103] [Medline: 21350250]
10. Liao Y, Skelton K, Dunton G, Bruening M. A systematic review of methods and procedures used in ecological momentary assessments of diet and physical activity research in youth: an adapted STROBE checklist for reporting EMA studies (CREMAS). J Med Internet Res 2016 Jun 21;18(6):e151. [doi: 10.2196/jmir.4954] [Medline: 27328833]
11. Nahum-Shani I, Smith SN, Spring BJ, et al. Just-in-Time Adaptive Interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. Ann Behav Med 2018 May 18;52(6):446-462. [doi: 10.1007/s12160-016-9830-8] [Medline: 27663578]
12. Trofholz A, Tate A, Janowiec M, et al. Ecological momentary assessment of weight-related behaviors in the home environment of children from low-income and racially and ethnically diverse households: development and usability study. JMIR Res Protoc 2021 Dec 1;10(12):e30525. [doi: 10.2196/30525] [Medline: 34855612]
13. Keijsers L, Boele S, Bülow A. Measuring parent-adolescent interactions in natural habitats. The potential, status, and challenges of ecological momentary assessment. Curr Opin Psychol 2022 Apr;44:264-269. [doi: 10.1016/j.copsyc.2021.10.002] [Medline: 34788708]
14. Heron KE, Everhart RS, McHale SM, Smyth JM. Using mobile-technology-based ecological momentary assessment (EMA) methods with youth: a systematic review and recommendations. J Pediatr Psychol 2017 Nov 1;42(10):1087-1107. [doi: 10.1093/jpepsy/jsx078] [Medline: 28475765]
15. Thunnissen MR, Aan Het Rot M, van den Hoofdakker BJ, Nauta MH. Youth psychopathology in daily life: systematically reviewed characteristics and potentials of ecological momentary assessment applications. Child Psychiatry Hum Dev 2022 Dec;53(6):1129-1147. [doi: 10.1007/s10578-021-01177-8] [Medline: 34075490]

XSL•FO
RenderX

16. Janssen LHC, Verkuil B, van Houtum L, Wever MCM, Elzinga BM. Perceptions of parenting in daily life: adolescent-parent differences and associations with adolescent affect. J Youth Adolesc 2021 Dec;50(12):2427-2443. [doi: 10.1007/s10964-021-01489-x] [Medline: 34482492]
17. Janssen LHC, Verkuil B, van Houtum L, Wever MCM, Wentholt WGM, Elzinga BM. A closer look into the affect dynamics of adolescents with depression and the interactions with their parents: an ecological momentary assessment study. Eur Child Adolesc Psychiatry 2024 Dec;33(12):4259-4272. [doi: 10.1007/s00787-024-02447-1] [Medline: 38762682]
18. Connelly K, Stein KF, Chaudry B, Trabold N. Development of an ecological momentary assessment mobile app for a low-literacy, Mexican American population to collect disordered eating behaviors. JMIR Public Health Surveill 2016 Jul 14;2(2):e31. [doi: 10.2196/publichealth.5511] [Medline: 27418020]
19. Silva C, Cero I, Ricci N, Pérez A, Conwell Y, Van Orden K. The feasibility and acceptability of using smartphones to assess suicide risk among Spanish-speaking adult outpatients. Suicide Life Threat Behav 2022 Oct;52(5):918-931. [doi: 10.1111/sltb.12889] [Medline: 35674249]
20. Pratap A, Anguera JA, Renn BN, et al. The feasibility of using smartphones to assess and remediate depression in Hispanic/Latino individuals nationally. Presented at: 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Symposium on Wearable Computers; Sep 11-15, 2017; Maui, Hawaii. [doi: 10.1145/3123024.3127877]
21. Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: building an international community of software platform partners. J Biomed Inform 2019 Jul;95:103208. [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]
22. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)-a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381. [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
23. Zucker RA, Gonzalez R, Feldstein Ewing SW, et al. Assessment of culture and environment in the Adolescent Brain and Cognitive Development Study: rationale, description of measures, and early data. Dev Cogn Neurosci 2018 Aug;32:107-120. [doi: 10.1016/j.dcn.2018.03.004] [Medline: 29627333]
24. Law MK, Furr RM, Arnold EM, Mneimne M, Jaquett C, Fleeson W. Does assessing suicidality frequently and repeatedly cause harm? A randomized control study. Psychol Assess 2015 Dec;27(4):1171-1181. [doi: 10.1037/pas0000118] [Medline: 25894705]
25. Bamberger KT. The application of intensive longitudinal methods to investigate change: stimulating the field of applied family research. Clin Child Fam Psychol Rev 2016 Mar;19(1):21-38. [doi: 10.1007/s10567-015-0194-6] [Medline: 26541560]
26. Montemayor R. Parents and adolescents in conflict: all families some of the time and some families most of the time. J Early Adolesc 1983 Mar;3(1-2):83-103. [doi: 10.1177/027243168331007]
27. Smetana JG. Adolescents' and parents' reasoning about actual family conflict. Child Dev 1989 Oct;60(5):1052-1067. [doi: 10.1111/j.1467-8624.1989.tb03536.x] [Medline: 2805883]

Abbreviations

CV: coefficient of variation
EMA: ecological momentary assessment
FBI: family-based intervention
REDCap: research electronic data capture

Edited by A Mavragani; submitted 30.04.24; peer-reviewed by G Ash, S Boele; revised version received 03.04.25; accepted 30.04.25; published 11.06.25.

Please cite as:
Fernandez A, Bernal S, Kim L, Potla S
Using Ecological Momentary Assessment to Assess Family Functioning in Spanish-Speaking Parent and Adolescent Dyads: Daily Questionnaire Study
JMIR Form Res 2025;9:e60073
URL: https://formative.jmir.org/2025/1/e60073
doi:10.2196/60073

© Alejandra Fernandez, Savannah Bernal, Lana Kim, Subodh Potla. Originally published in JMIR Formative Research (https://formative.jmir.org), 11.6.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Reactivity to Smoking Cues in a Social Context: Virtual Reality Experiment

Katharina Eidenmueller[1,2], Dr sc hum; Sabine Hoffmann[1,2], MA; Kornelius Kammler-Sücker[1,3], Dr sc hum; Leonard Wenger[1,2], MSc; Massimiliano Mazza[2], Dott; Christiane Mühle[4], PD Dr rer nat; Manuel Stenger[1,2], Dr med; Gerrit Meixner[5], Prof Dr; Falk Kiefer[1,2], Prof Dr med; Bernd Lenz[1,2], Prof Dr med

[1]Partner Site Mannheim-Heidelberg-Ulm, Deutsches Zentrum für Psychische Gesundheit (German Center for Mental Health), Mannheim, Germany
[2]Department of Addictive Behavior and Addiction Medicine, Central Institute of Mental Health, J5, Mannheim, Germany
[3]Center for Innovative Psychiatric and Psychotherapeutic Research (CIPP), Virtual Reality Lab, Central Institute of Mental Health, Mannheim, Germany
[4]Department of Psychiatry and Psychotherapy, Friedrich-Alexander University Erlangen-Nürnberg (FAU), Erlangen, Germany
[5]UniTyLab, Faculty of Informatics, Heilbronn University, Heilbronn, Germany

Corresponding Author:
Katharina Eidenmueller, Dr sc hum
Partner Site Mannheim-Heidelberg-Ulm, Deutsches Zentrum für Psychische Gesundheit (German Center for Mental Health), Mannheim, Germany

Abstract

Background: Social contextual factors influence the onset and maintenance of substance abuse. Virtual reality (VR) provides a standardized method to present social stimuli and is increasingly used in addiction research.

Objective: This study examines the influence of a smoking versus a nonsmoking agent in VR on craving in nicotine-dependent male participants. Our primary hypothesis was that the interaction with a smoking agent is associated with increased craving compared to a nonsmoking agent. We expected higher craving in the presence of an agent regardless of the agent's smoking status.

Methods: Using a head-mounted display (Oculus Rift), 50 nicotine-dependent smokers were exposed to four VR conditions on a virtual marketplace: first without an agent, second and third with an agent who either smoked or did not smoke in randomized order, and fourth without an agent as a follow-up condition. Before the follow-up condition, participants smoked a cigarette. Craving was assessed with the Questionnaire of Smoking Urges and a visual analog scale within VR and after each session. We also examined anxiety and agitation (visual analog scale), immersion and presence with the igroup Presence Questionnaire, and salivary cortisol levels.

Results: Results showed no significant difference in the participants' craving, anxiety, or agitation between the smoking and nonsmoking agent conditions. However, craving, anxiety, and agitation increased from the marketplace without an interacting agent to the conditions with an interacting agent, and decreased after smoking a cigarette. Immersion was low in all conditions and decreased over time. Salivary cortisol levels were highest at baseline and decreased over the course of the experiment.

Conclusions: These findings suggest that the presence of an agent (as a contextual factor) may override the specific influence of proximal stimuli (burning cigarette). The low immersion highlights the challenges in developing effective VR environments for cue exposure.

Trial Registration: Deutsches Register Klinischer Studien DRKS00025746; https://drks.de/search/de/trial/DRKS00025746

(JMIR Form Res 2025;9:e71285) doi:10.2196/71285

KEYWORDS

virtual reality; smoking; cue reactivity; craving; social context

Introduction

Despite its decreasing prevalence, cigarette smoking remains a global health challenge. According to the latest report of the Global Burden of Disease study from 2019, a total of 1.24 billion people were current smokers, and 1.69 million deaths occurred as a consequence of tobacco use globally [1]. One of the diagnostic criteria for tobacco use disorder is craving [2], which describes the strong urge or desire to smoke. Craving is a well-established driver of smoking behavior [3,4] and relapse after smoking cessation [5].

A meta-analysis on cue-reactivity in smokers shows that the presentation of smoking-related cues can induce craving for cigarettes and that self-reported craving is a valuable index for

XSL•FO
RenderX

cue-reactivity, as effect sizes were robust across different moderators. In contrast, the effects of physiological measures of cue reactivity remained small or nonsignificant [6]. Social context was found to modulate cue-induced craving [7,8]. Dimoff and Sayette [9] make a plea for including social contextual factors in laboratory experiments on smoking in order to gain a deeper understanding of the mechanisms underlying smoking behavior and craving. The authors point out that only a few studies examine the influence of social context on smoking behavior, but those that do suggest an influence of the presence of others on smoking behavior, the effects of smoking, and self-regulatory and perceptive processes related to smoking. Adolescents with peers who are smokers are more likely to smoke or to start smoking [10]. Field studies using ecological momentary assessment found that smokers are particularly likely to smoke when socializing [11] or in the presence of other people who are smoking [12,13]. The presence of people who are not smoking, on the other hand, appears to suppress smoking behavior [14]. Social context was found to enhance cue-induced craving compared to cues presented in a neutral context in one study [7], whereas it seemed to reduce cue-induced craving for cigarettes in a study using virtual reality (VR) [8]. Which specific processes are involved in the modulation of smoking behavior and cue-induced craving in a social context has yet to be determined.

Stress hormones like cortisol take effect in the context of social stress, and increased cortisol levels in response to social stressors have been shown to be associated with increased craving in smokers [15,16] and increased reactivity to smoking cues [17]. This could indicate cortisol levels as a potential mediator between social context and cue-induced craving.

VR is a promising tool for studying social context, as it facilitates the standardized and controlled presentation of complex social situations in experimental designs without sacrificing ecological validity [18,19]. Moreover, VR simulations have been successfully used to induce cigarette craving [20] and the accompanying physiological reactions [21] in addicted smokers. Smoking cues in VR appear to induce stronger cravings than the presentation of smoking-related images [22]. Different VR environments can induce cigarette craving (bar, restaurant, street, and home scenarios) [23]. Virtual party scenarios that include social interactions with virtual agents who smoke and offer cigarettes induce higher levels of craving than a virtual paraphernalia room with smoking cues [20]. Furthermore, VR-based cue exposure therapy has been successfully implemented for smoking cessation [24], but there are also null findings of VR cue exposure on self-reported craving [25]. The subjective level of presence in the VR is positively correlated with craving for cigarettes [26]. As VR has been shown to be an appropriate tool for studying social context, as well as for the induction of craving in smokers, the use of VR appears to be a suitable approach for examining the influence of social contextual factors on craving for cigarettes.

In this study, our primary aim was to examine the influence of the interaction with a smoking versus a nonsmoking virtual agent on craving and stress hormone levels in smokers with tobacco use disorder. Study participants were immersed in a VR setting in which they interacted with a virtual agent who was, depending on the condition, either smoking a cigarette or not. Craving, anxiety, and agitation were then assessed within the VR, and saliva samples were collected after each trial. We hypothesized that the presence of a smoking agent would be associated with an increase in craving compared to the nonsmoking agent. We also expected a difference in craving levels in the presence of an agent compared to the absence of an agent, regardless of smoking status, and a decrease in craving after smoking a cigarette. Moreover, we explored the effects on anxiety and agitation and the association between salivary stress hormone levels and cue-elicited craving in VR.

Methods

Participants

The study sample consisted of 50 male smokers aged between 18 and 64 (mean 30.38, SD 10.97) years. Sample characteristics are depicted in Table 1. Only smokers who met the Diagnostic and Statistical Manual of Mental Disorders, 5th edition criteria for tobacco use disorder were included. Because the menstrual cycle has been associated with substance use behaviors [27] and controlling for the phases would have required a large number of participants, no female participants were recruited for this pilot study. Further exclusion criteria were psychiatric comorbidities, having had more than one psychiatric or psychotherapeutic treatment session within 10 years prior to recruitment, current intake of psychoactive medication, severe somatic disorders or neurological disorders, epilepsy, and illegal drug use. Participants who had ametropia of more than three diopters and could not wear contact lenses were also excluded to ensure a clear vision of the VR scenario. Participants were questioned about these criteria in a telephone screening before being included in the study.

XSL•FO
RenderX

Table . Demographic and clinical characteristics of study participants (N=50).

	12-hour abstinence (n=20)	30-minute abstinence (n=30)	*t* test (df)	*P* value
Age (years), mean (SD)	28.05 (10.39)	31.93 (11.24)	–1.23 (48)	.22
Severity of nicotine dependence (according to the FTND[a] sum score)[b], %				.18
Low	10	3.3	N/A[c]	N/A
Moderate	45	33.3	N/A	N/A
High	30	40	N/A	N/A
Very high	15	23.3	N/A	N/A
FTND sum score, mean (SD)	5.0 (2.18)	5.97 (1.65)	–1.79 (48)	.08
Age at first cigarette (years), mean (SD)	15.10 (3.64)	14.93 (2.49)	1.93 (48)	.85
Cigarettes per day, mean (SD)	12.20 (7.44)	12.87 (5.69)	–0.36 (48)	.72
Years of smoking, mean (SD)	12.05 (11.70)	14.73 (9.94)	–0.87 (48)	.39
PSS[d] score, mean (SD)	15.10 (5.48)	13.74 (6.01)	0.80 (48)	.43

[a]Fagerström Test for Nicotine Dependence.

[b]For nicotine dependence: U=41.38; Z=–1.33.

[c]N/A: not applicable.

[d]Perceived Stress Scale.

Of the 50 participants, the first 20 were asked to remain abstinent from smoking for 12 hours before the study appointment. Adherence to this was checked with a carbon monoxide breath test. The remaining 30 participants were instructed to abstain from smoking for 30 minutes prior to the study appointment. The two groups did not differ significantly in age, number of cigarettes per day, age of first cigarette, smoking years, and severity of nicotine dependence (Table 1).

Ethical Considerations

Study participation was voluntary, and all participants provided written informed consent prior to participation. The study was carried out in accordance with the Declaration of Helsinki and was approved by the Ethics Committee II of the Medical Faculty Mannheim, Heidelberg University, Germany (#2021-504). Participants received a compensation of 50€ (US$ 56.23) for study participation. The data was anonymized.

Measures

The severity of nicotine dependence was assessed with the Fagerström Test for Nicotine Dependence [28]. The Perceived Stress Scale [29] was used to assess participants' self-reported levels of stress. Cigarette craving was assessed with the Questionnaire of Smoking Urges (QSU). The QSU has two subscales, which represent the expected positive consequences of smoking (factor 1) and relief from unpleasant states (factor 2) [30]. The 7-point visual analog scales (VAS) were used to assess craving, anxiety, and agitation during the VR trials. To measure the level of immersion in the virtual environment, the igroup Presence Questionnaire (iPQ) was used, assessing participants' sense of realism, spatial presence, and involvement in the VR in distinct scales [31]. As VR simulations have the potential to induce motion sickness in participants, the Simulator Sickness Questionnaire [32] was used to be able to control symptoms of motion sickness. Saliva samples were analyzed for cortisol levels by ELISA (RE52611, IBL International GmbH, Hamburg, Germany).

Virtual Environment and Apparatus

The virtual environment was presented on a head-mounted display (Oculus Rift, Reality Labs/Meta Platforms, Menlo Park, California), together with the accompanying Touch Controllers for user interaction. The experiment was programmed using the Python-based software Vizard 6 (WorldViz). The 3D environment was a customized own design, based on templates provided with Vizard and other models purchased from various sources and manually adapted with 3d Studio (Autodesk). The virtual companion character was created with Fuse CC beta (Adobe Inc) and rigged and animated with Mixamo (Adobe Inc), with its own adaptations to the animation to include smoking movements (created with MotionBuilder, Autodesk). We used voice recordings of a volunteer for the character's dialogue parts when chatting with the participant.

Within the virtual environment, participants found themselves seated under a sunshade in an outdoor café on a plaza of roughly Mediterranean style. In the distance, two virtual characters sat on a bench, chatting with each other, and a fountain on the square would provide a constantly burbling background sound. Besides these aspects, the further environment was free of distractions and designed to provide a comfortable atmosphere. Placed on a table directly in front of the participants, they found a virtual computer screen, which displayed all the following instructions and visual reaction tasks. When in either one of the agent conditions, the virtual character faced the participants while leisurely sitting on a neighboring table to their right

XSL•FO
RenderX

(Figure 1). The character then involved the participants in a short dialogue, which included the character commenting on how stressed out the participant appeared (both agent conditions), and offered a cigarette two times (smoking agent condition). During the dialogue, participants were asked via screen instructions to respond, and then to indicate that they had finished with a button press on the handheld controller. After the dialogue, the virtual character remained seated, either comfortably idling (nonsmoking condition) or taking a drag from her smoking cigarette from time to time (smoking condition). The smoking animation was repeated with nonregular (randomized), yet high frequency, with the character dragging or puffing out for roughly 47% of the time. The smoking animation included an appreciative inhalation sound and a prolonged puffing out of virtual smoke with accompanying sound effects. In the smoking agent condition, this continued throughout all tasks.

Figure 1. Screenshot of the market square in the smoking agent condition. On the computer screen within the VR, the craving visual analog scale is depicted ("How strong is your craving for a cigarette?" in German language). VR: virtual reality.

Procedure

Prior to the assessment, a telephone screening was conducted to ensure participants met the diagnostic criteria for tobacco use disorder and to check for exclusion criteria. At the study appointment, a carbon monoxide breath test was conducted on arrival for the participants who were asked to remain abstinent from smoking for 12 hours prior to the assessment. At the beginning of the appointment, all participants were asked to fill in an initial set of questionnaires on a laptop (sociodemographic information, Fagerström Test for Nicotine Dependence, and Perceived Stress Scale). Subsequently, four VR trials were conducted. For the baseline assessment, participants were presented with the virtual marketplace with no agent present. The second and third were the trials that included the interaction with the virtual agent, who was either smoking and offering cigarettes, or not smoking and asking participants about daily stressors. The order in which these conditions were presented was randomized. In the fourth trial, participants were once again presented with the empty marketplace with no agent present as a follow-up measurement. In each trial, after interacting with the agent or after getting some time to familiarize themselves with the marketplace (for trials with no agent present), the VAS were conducted on the computer screen within the VR. Each trial took about 20 minutes. After each VR trial, participants filled in a set of questionnaires on a laptop (QSU, iPQ, and Simulator Sickness Questionnaire) and provided saliva samples, which took around 10 minutes each time. Before the follow-up trial, participants smoked a cigarette during a 10-minute cigarette break. Figure 2 shows a flowchart of the experimental procedure.

XSL•FO
RenderX

Figure 2. Flowchart of the experimental procedure. iPQ: igroup Presence Questionnaire; QSU: Questionnaire of Smoking Urges; VAS: visual analog scale.

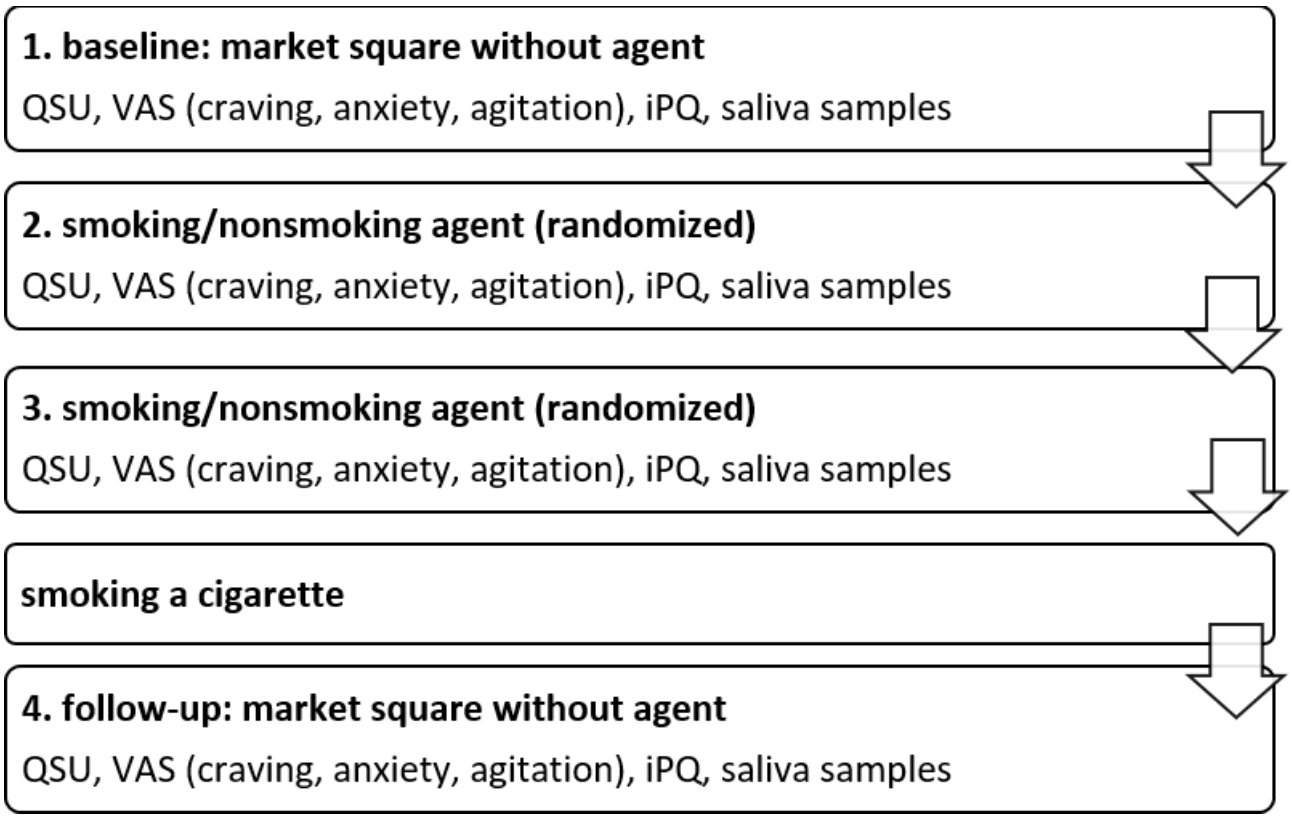

Statistical Analysis

In order to compare the smoking with the nonsmoking agent conditions, linear mixed models were fitted for both QSU scales, VAS (craving, anxiety, and agitation), iPQ total score, and cortisol as dependent variables with age, duration of prior abstinence, sequence (smoking or nonsmoking agent first), time point (trials 2 or 3), and condition (smoking vs nonsmoking agent) as predictors. Cortisol levels were logarithmized prior to the analysis. To analyze changes over the course of the experiment, multilevel models were fitted with both the QSU scale, VAS (craving, anxiety, and agitation), cortisol, and the iPQ total score as the dependent variables and age, duration of prior abstinence, sequence (smoking or nonsmoking agent first), and condition (baseline (T0), smoking|nonsmoking (T1|T2), follow-up (T3)) as predictors. We estimated multilevel (mixed) models with random intercepts to account for the clustered structure of the data points within participants (level 1: VR condition; level 2: participant). We report *F* tests for main effects in the generalized linear mixed models, followed by 2-tailed *t* tests for the parameter estimates β, indicating differences from the reference category in categorical variables; *t* values are shown for predictors with three or more categories. Statistical analyses were conducted using SPSS Statistics version 29; IBM Corp) for Windows. Results were regarded as significant when the 2-sided *P* value was below .05.

Results

Differential Effects Between Smoking and Nonsmoking Agents

There was no significant difference between the effects of the smoking agent compared to the effects of the nonsmoking agent on explicit craving in both QSU scales and the VAS (Table 2). For the QSU factor 2 subscale and the VAS, the craving was higher in the group with longer abstinence, and for the craving VAS, the time point (trials 2 or 3) was a significant predictor, with higher craving for the third trial. The effects shown also remain when the duration of abstinence is taken into account. The duration of abstinence increases craving but is not a moderator for the effects of the VR in our models (Table S1 in Multimedia Appendix 1).

XSL•FO
RenderX

Table . Results of a multilevel model with craving as the dependent variable and age, prior abstinence, sequence, time point (trials 2 or 3), and condition as predictors.

	β (95% CI)	*F* test (*df*)	*P* value
QSU[a] factor 1 (cases n=50; observations n=93)			
Age	.196 (–0.138 to 0.529)	0.514 (1, 45)	.48
Prior abstinence	__[b]	5.548 (1, 45)	.24
12 hours	8.654 (1.259 to 16.049)	—	—
30 minutes	[reference]		
Sequence	—	0.045 (1, 46)	.83
Smoking agent first	–.759 (–7.975 to 6.456)	—	—
Nonsmoking agent first	[reference]		
Time point	—	0.883 (1, 42)	.35
Second trial	–1.039 (–3.273 to 1.194)	—	—
Third trial	[reference]		
Condition	—	0.131 (1, 42)	.72
Smoking Agent	–.401 (–2.635 to 1.833)	—	—
Nonsmoking Agent	[reference]		
QSU factor 2 (cases n=50; observations n=99)			
Age	.155 (–0.235 to 0.545)	0.641 (1, 46)	.43
Prior abstinence	—	5.792 (1, 46)	.02[c]
12 hours	10.304 (1.685 to 18.922)	—	—
30 minutes	[reference]		
Sequence	—	0.338 (1, 46)	.56
Smoking agent first	2.431 (–5.993 to 10.856)	—	—
Nonsmoking Agent First	[reference]		
Time point	—	0.970(1, 47)	.33
Second trial	–.861 (–2.620 to 0.898)	—	—
Third trial	[reference]		
Condition	—	1.422 (1, 47)	.24
Smoking agent	–1.043 (–2.802 to 0.716)	—	—
Nonsmoking agent	[reference]		
Craving VAS[d] (cases n=50; observations n=100)			
Age	–.008 (–0.044 to 0.028)	0.190 (1, 46)	.66
Prior abstinence	—	6.699 (1, 46)	.01[c]
12 hours	1.017 (0.226 to 1.809)	—	—
30 minutes	[reference]		
Sequence	—	0.308 (1, 46)	.58
Smoking Agent First	–.213 (–0.987 to 0.560)	—	—
Nonsmoking agent first	[reference]		
Time point	—	4.384 (1, 48)	.04[c]
Second trial	–.451 (–0.885 to 0.018)	—	—
Third trial	[reference]		

XSL•FO
RenderX

	β (95% CI)	*F* test (*df*)	*P* value
Condition	—	0.191 (1, 48)	.66
Smoking agent	.094 [–0.339 to 0.528)	—	—
Nonsmoking agent	[reference]		

[a]QSU: Questionnaire of Smoking Urges.

[b]Not applicable.

[c]$P<.05$.

[d]VAS: visual analog scale.

There were also no significant differential effects of the smoking versus the nonsmoking agent on the anxiety or agitation VAS, the IPQ score (presence and immersion), or the saliva cortisol concentrations (Table 3). There was a significant increase in cortisol concentrations from the second to the third trial.

XSL•FO
RenderX

Table . Results of a multilevel model with anxiety, agitation, presence, and cortisol as the dependent variable and age, prior abstinence, sequence, time point (trials 2 or 3), and condition as predictors[a].

	β (95% CI)	*F* test (df)	*P* value
Anxiety VAS[b]			
Age	.002 (–0.027 to 0.031)	0.013 (1, 46)	.91
Prior abstinence	__[c]	1.454 (1, 46)	.23
12 hours	–.385 (–1.027 to 0.258)	—	—
30 minutes	[reference]		
Sequence	—	0.694 (1, 46)	.41
Smoking agent first	–.260 (–0.888 to 0.368)	—	—
Nonsmoking agent first	[reference]		
Time point	—	0.176 (1, 46)	.68
Second trial	.060 (–0.228 to 0.348)	—	—
Third trial	[reference]		
Condition	—	1.362 (1, 46)	.25
Smoking agent	–.167 (–0.455 to 0.121)	—	—
Nonsmoking agent	[reference]		
Agitation VAS			
Age	–.30 (−0.067 to 0.006)	2.746 (1, 46)	.10
Prior abstinence	—	0.588 (1, 46)	.45
12 hours	–.308 (−0.501 to 1.117)	—	—
30 minutes	[reference]		
Sequence	—	0.411 (1, 46)	.53
Smoking agent first	–.252 (–1.04 to 0.539)	—	—
Nonsmoking agent first	[reference]		
Time point	—	0.043 (1, 46)	.84
Second trial	–.044 (–0.381 to 0.468)	—	—
Third trial	[reference]		
Condition	—	2.993 (1, 48)	.09
Smoking agent	–.365 (–0.790 to 0.059)	—	—
Nonsmoking agent	[reference]		
iPQ[d] score			
Age	–.012 (–0.128 to 0.103)	0.047 (1, 46)	.83
Prior abstinence	—	0.001 (1, 46)	.97
12 hours	–.048 (–2.600 to 2.505)	—	—
30 minutes	[reference]		
Sequence	—	4.6517 (1, 46)	.04[e]
Smoking agent first	2.634 (0.139 to 5.129)	—	—
Nonsmoking agent first	[reference]		
Time point	—	0.756 (1, 48)	.39
Second trial	–.261 (–0.342 to 0.863)	—	—
Third trial	[reference]		
Condition	—	0.023 (1, 48)	.88

XSL•FO
RenderX

	β (95% CI)	*F* test (df)	*P* value
Smoking agent	–.045 (–0.557 to 0.648)	—	—
Nonsmoking agent	[reference]		
Cortisol concentrations			
Age	–.010 (–0.009 to 0.029)	1.133 (1, 46)	.29
Prior abstinence	—	0.042 (1, 46)	.84
12 hours	–.042 (–0.374 to 0.459)	—	—
30 minutes	[reference]		
Sequence	—	0.022 (1, 46)	.88
Smoking agent first	–.030 (–0.437 to 0.377)	—	—
Nonsmoking agent first	[reference]		
Time point	—	18.463 (1, 48)	<.001[e]
Second trial	–.269 (0.143 to 0.396)	—	—
Third trial	[reference]		
Condition	—	0.307 (1, 48)	.58
Smoking agent	–.035 (–0.161 to 0.091)	—	—
Nonsmoking agent	[reference]		

[a]Cases n=50; Observations n=100.
[b]VAS: visual analog scale.
[c]Not applicable.
[d]iPQ: igroup Presence Questionnaire.
[e]$P<.05$

Longitudinal Effect From Baseline to Marketplace With Interacting Agent to Follow-Up

From baseline to the conditions with an agent present, there was an increase in craving in both QSU scales. For the craving VAS, the increase compared to the baseline was not significant. After smoking a cigarette, there was a reduction in craving in QSU and VAS. Anxiety and agitation increased from baseline to the agent conditions and decreased after smoking a cigarette. The iPQ score decreased from the baseline and agent conditions to the follow-up, and the cortisol concentrations reduced from baseline to the agent and follow-up conditions. Figure 3 depicts changes in the craving measures, anxiety, agitation, immersion, and salivary cortisol over the different conditions. Results tables are provided in Tables S2-S8 in Multimedia Appendix 1.

XSL•FO
RenderX

Figure 3. Bar graphs of estimated marginal means (+/– standard errors of the means) for QSU scales, VAS (craving, anxiety, agitation), iPQ sum scores, and salivary cortisol (logarithmized values) over the course of the experiment. The timeline is structured as baseline, agents (smoking|nonsmoking merged), and follow-up. Asterisks indicate significant differences between time points: * for *P*<.05, ** for *P*<.01, and *** for *P*<.001.

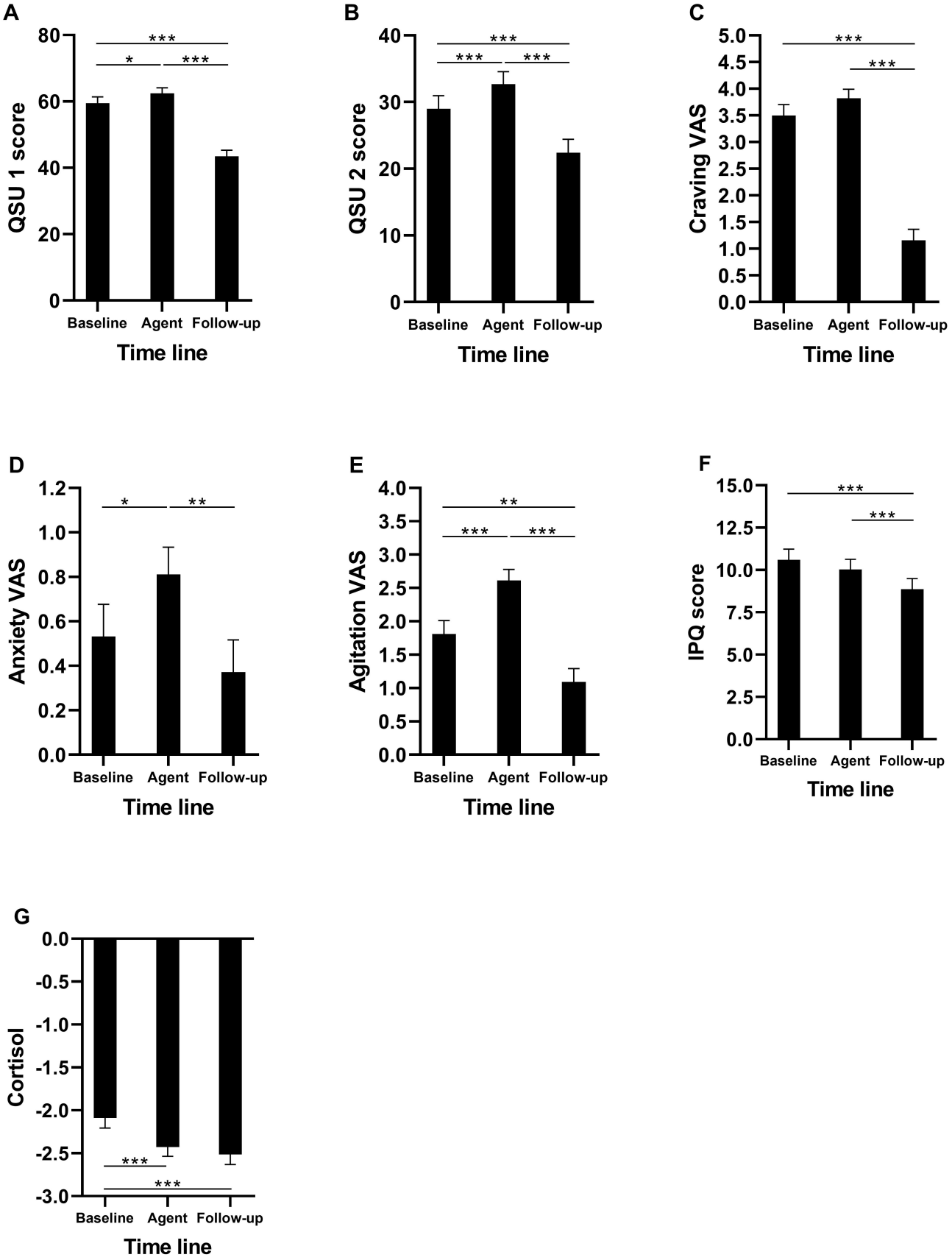

Participants who were abstinent for 12 hours reported overall higher cravings in VAS and QSU compared to participants who had been abstinent for 30 minutes. For salivary cortisol levels, immersion, anxiety, and agitation, the duration of prior abstinence was no significant predictor.

XSL•FO
RenderX

According to the qualitative grading system for iPQ scores by Melo and Gonçalves [33], the mean immersion scores were low for all three subscales. Over the course of the experiment, there was a decrease in immersion, with higher iPQ sum scores at baseline and lower scores at follow-up.

Salivary cortisol concentrations were highest at the beginning of the experiment and decreased from baseline to the agent and follow-up conditions. Follow-up cortisol levels did not differ significantly from the agent condition (Table S8 in Multimedia Appendix 1).

Discussion

Principal Findings

This study provides new insight into the interaction of contextual factors with proximal stimuli in relation to cigarette craving in a VR environment. Self-reported craving did not significantly differ in the presence of a smoking compared to a nonsmoking virtual agent. However, in the longitudinal design, the presence of an agent per se was associated with higher craving in the QSU compared to baseline, unrelated to the agent's smoking status. As a positive control, smoking a cigarette was followed by a decrease in craving, which speaks for the validity of this study setup and highlights the applicability of the VAS within VR as a valid measure for explicit craving. The craving results were paralleled by the anxiety and agitation VAS, which were higher in the agent conditions compared to baseline and decreased after smoking a cigarette.

Our findings suggest that social context can induce craving independently of the presence of smoking-related cues. The presence of and interaction with a virtual agent was associated with more craving compared to baseline, but the agent smoking, who offered a cigarette, did not lead to a further increase in self-reported craving. Similarly, a VR study with patients with alcohol dependence [34] found that social pressure applied by virtual agents (offering drinks and trying to persuade participants to drink) without environmental alcohol cues induced a significant increase in craving. Social pressure with the addition of alcohol cues was not associated with more craving than social pressure on its own. While Lee et al [34] included agents pressuring participants to drink alcohol, the interaction in the nonsmoking agent condition did not reference the topic of smoking but rather referred to the participants as appearing quite stressed to the virtual agent. A possible explanation for this study's results could be stimulus generalization [35]. When addictive substances like cigarettes and alcohol are often consumed in the presence of others, a social context that would normally be associated with substance use can elicit craving without the presence of specific substance-related cues. Studies with pictorial stimuli show that the combination of different drug-related stimuli does not necessarily have additive effects on cue-induced responses [36], which might explain why we did not find higher craving in the smoking compared to the nonsmoking agent condition.

This study's findings differ from those of Winkler et al [8], who found that cue-induced craving in smokers was attenuated in the social context in a VR experiment. The authors provided the explanation that the presence of an agent who neither smokes nor applies social pressure to smoke might reduce the perceived availability of smoking. In contrast to our experiment, Winkler et al [8] did not include interactions with the virtual agents but had agents face the participants at a table with a neutral facial expression. It can be argued that the availability of smoking might be perceived as higher when the agent shows interest and interacts in a friendly way compared to noninteractive neutrality. Furthermore, the outdoor setting of our virtual environment might make smoking in the presence of another person, unrelated to their smoking status, appear less rude compared to the indoor setting of Winkler et al [8], which would also increase the perceived availability of smoking, and therefore, not lead to an attenuation of craving as argued by the authors. However, the smoking agent condition, which arguably had higher availability of smoking as the agent repeatedly offered a cigarette, did not lead to increased craving, which speaks against the perceived availability hypothesis. Generally, the discrepancy between our findings and those of Winkler et al [8] suggests that social context affects craving differently depending on the presence or absence of social interactions. Assessing how different aspects of social context might influence craving and modulate the response to smoking cues could be an interesting topic for further research.

We found an increase in craving from baseline to the agent conditions in the QSU, but not in the VAS. This is somewhat surprising, as the literature review shows high correlations between the QSU and craving VAS [37]. One possible explanation is that the VAS was administered within the VR, whereas the QSU was filled in immediately after each VR trial on a laptop, resulting in differences in the environment, but also the time participants had been exposed to the VR condition at the time of the two craving assessments. In addition, it can be argued that compared to a validated psychometric questionnaire like the QSU, the VAS assesses craving more broadly and may be less sensitive to changes in specific aspects of experienced craving.

Cortisol concentrations did not differ significantly between the smoking and nonsmoking agent conditions and decreased over the course of the experiment. Smoking a cigarette did not appear to have a significant effect on salivary cortisol concentration at follow-up. It is possible that participants experienced more stress at the beginning of the experiment due to the unfamiliar environment and experimental tasks, and decreased as they familiarized themselves with the VR environment over time. Considering the latency of poststress changes in salivary cortisol concentration [38], it is possible that potential cortisol peaks after cue exposure could not have been detected in our experiment due to the interval between cue exposure and saliva sampling being too short.

Limitations

A limiting factor of this study is the weak immersion in the VR environment. The depth of immersion, as assessed with the iPQ, was classified as low and decreased over the course of the experiment. A study testing different VR environments for the induction of craving for cigarettes found a positive correlation between the subjective level of presence in the VR and craving

XSL•FO
RenderX

[26]. In patients with alcohol use disorder, the perceived realism of alcoholic beverages in a VR environment is a predictor for craving [39]. It is possible that craving did not increase in the presence of the smoking agent because the smoking cues were not perceived as realistic enough. However, the same study found no correlation between the perceived realism of the overall VR environment and induced craving [39]. As participants only rated the overall immersion, but not the realism of specific aspects (eg, smoking cues) of the environment, it is unclear if the craving results were likely to be influenced by low immersion. Moreover, several studies indicate that while immersion is a relevant factor, its importance appears not to be central to study outcomes [40-42]. For this reason, we do not expect the low immersion to severely limit the validity of our results, albeit it certainly is a limiting factor.

Another limitation of this study is the lack of measures to evoke embodiment. Sense of embodiment in VR describes the sense of ownership, self-location, and agency over the virtual body [43] and is being discussed as one of the key factors that make VR applications effective, as it is suggested to mirror the brain mechanism of embodied simulations [44]. It has been shown to impact responses to virtual stimuli [45], and therefore, might influence cue-induced craving in VR. However, a study on VR gambling did not find an association between the level of embodiment and craving [46]. As ours was a small-budget pilot study, we were not able to implement measures to evoke embodiment, like programming a virtual avatar for participants or the inclusion of visuotactile stimulation. We also did not assess participants' sense of embodiment. The role of embodiment in cue-induced craving in VR is a relevant topic for future research.

Conclusions

In conclusion, this study shows that inducing cigarette craving in VR may not be as easy as some studies suggest. We were able to elicit craving with our VR environment, however, the presence of a smoking agent did not lead to the increase in craving we had expected when compared to a nonsmoking agent. This highlights the importance of manipulation checks in VR settings targeting craving and cue reactivity. According to our findings, social context in a virtual environment does have an effect on cigarette craving. However, which specific contextual factors are relevant or necessary for the induction and modulation of craving in smokers is not fully clear and requires further research. As VR allows the standardized presentation and isolated manipulation of social contextual stimuli, VR experiments present an exciting approach to address this question.

Acknowledgments

We would like to thank PD Dr Tagrid Leménager for providing the voice recordings. We would also like to thank Berkan Akin for his help in data collection. Furthermore, we would like to thank all participants for taking part in our study. The study was presented at the German Addiction Conference (Deutscher Suchtkongress) 2024.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary materials.
[DOCX File, 29 KB - formative_v9i1e71285_app1.docx]

References

1. Reitsma MB, Kendrick PJ, Ababneh E. Spatial, temporal, and demographic patterns in prevalence of smoking tobacco use and attributable disease burden in 204 countries and territories, 1990–2019: a systematic analysis from the Global Burden of Disease Study 2019. Lancet 2021 Jun 19;397(10292):2337-2360. [doi: 10.1016/S0140-6736(21)01169-7] [Medline: 34051883]
2. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders 2013:947. [doi: 10.1176/appi.books.9780890425596]
3. Conklin CA, Vella EJ, Joyce CJ, Salkeld RP, Perkins KA, Parzynski CS. Examining the relationship between cue-induced craving and actual smoking. Exp Clin Psychopharmacol 2015 Apr;23(2):90-96. [doi: 10.1037/a0038826]
4. Gass JC, Motschman CA, Tiffany ST. The relationship between craving and tobacco use behavior in laboratory studies: a meta-analysis. Psychol Addict Behav 2014 Dec;28(4):1162-1176. [doi: 10.1037/a0036879] [Medline: 25134054]
5. Vafaie N, Kober H. Association of drug cues and craving with drug use and relapse. JAMA Psychiatry 2022 Jul 1;79(7):641. [doi: 10.1001/jamapsychiatry.2022.1240] [Medline: 35648415]
6. Betts JM, Dowd AN, Forney M, Hetelekides E, Tiffany ST. A meta-analysis of cue reactivity in tobacco cigarette smokers. Nicotine Tob Res 2021 Jan 22;23(2):249-258. [doi: 10.1093/ntr/ntaa147] [Medline: 32772094]

XSL•FO
RenderX

7. Vollstädt-Klein S, Nees F, Wieland A, et al. Contexts enhance ratings of craving and psychophysiological responses of cue-reactivity in tobacco use disorder. medRxiv. Preprint posted online on Jul 13, 2022. [doi: 10.1101/2022.07.12.22277347]
8. Winkler MH, Li Y, Pauli P, Mühlberger A. Modulation of smoking cue reactivity by social context—implications for exposure therapy in virtual reality. Front Virtual Real 2023;4. [doi: 10.3389/frvir.2023.926679]
9. Dimoff JD, Sayette MA. The case for investigating social context in laboratory studies of smoking. Addiction 2017 Mar;112(3):388-395. [doi: 10.1111/add.13503] [Medline: 27503776]
10. Simons-Morton BG, Farhat T. Recent findings on peer group influences on adolescent smoking. J Prim Prev 2010 Aug;31(4):191-208. [doi: 10.1007/s10935-010-0220-x] [Medline: 20614184]
11. Hatsukami DK, Morgan SF, Pickens RW, Champagne SE. Situational factors in cigarette smoking. Addict Behav 1990;15(1):1-12. [doi: 10.1016/0306-4603(90)90002-f] [Medline: 2316408]
12. Shiffman S, Gwaltney CJ, Balabanis MH, et al. Immediate antecedents of cigarette smoking: an analysis from ecological momentary assessment. J Abnorm Psychol 2002 Nov;111(4):531-545. [doi: 10.1037//0021-843x.111.4.531] [Medline: 12428767]
13. Shiffman S, Paty JA, Gnys M, Kassel JA, Hickcox M. First lapses to smoking: within-subjects analysis of real-time reports. J Consult Clin Psychol 1996 Apr;64(2):366-379. [doi: 10.1037//0022-006x.64.2.366] [Medline: 8871421]
14. Shiffman S, Rathbun SL. Point process analyses of variations in smoking rate by setting, mood, gender, and dependence. Psychol Addict Behav 2011 Sep;25(3):501-510. [doi: 10.1037/a0022178] [Medline: 21480683]
15. Childs E, de Wit H. Effects of acute psychosocial stress on cigarette craving and smoking. Nicotine Tob Res 2010 Apr 1;12(4):449-453. [doi: 10.1093/ntr/ntp214]
16. Buchmann AF, Laucht M, Schmid B, Wiedemann K, Mann K, Zimmermann US. Cigarette craving increases after a psychosocial stress test and is related to cortisol stress response but not to dependence scores in daily smokers. J Psychopharmacol 2010 Feb;24(2):247-255. [doi: 10.1177/0269881108095716] [Medline: 18957475]
17. Wanger TJ, de Moura FB, Ashare R, et al. Brain and cortisol responses to smoking cues are linked in tobacco-smoking individuals. Addict Biol 2023 Dec;28(12):e13338. [doi: 10.1111/adb.13338] [Medline: 38017638]
18. Pan X, Hamilton AFDC. Why and how to use virtual reality to study human social interaction: the challenges of exploring a new research landscape. Br J Psychol 2018 Aug;109(3):395-417. [doi: 10.1111/bjop.12290] [Medline: 29504117]
19. Yaremych HE, Persky S. Tracing physical behavior in virtual reality: a narrative review of applications to social psychology. J Exp Soc Psychol 2019 Nov;85:32831397. [doi: 10.1016/j.jesp.2019.103845] [Medline: 32831397]
20. Thompson-Lake DGY, Cooper KN, Mahoney JJ III, et al. Withdrawal symptoms and nicotine dependence severity predict virtual reality craving in cigarette-deprived smokers. Nicotine Tob Res 2015 Jul;17(7):796-802. [doi: 10.1093/ntr/ntu245]
21. Choi JS, Park S, Lee JY, et al. The effect of repeated virtual nicotine cue exposure therapy on the psychophysiological responses: a preliminary study. Psychiatry Investig 2011 Jun;8(2):155-160. [doi: 10.4306/pi.2011.8.2.155] [Medline: 21852993]
22. Lee JH, Ku J, Kim K, et al. Experimental application of virtual reality for nicotine craving through cue exposure. CyberPsychol Behav 2003 Jun;6(3):275-280. [doi: 10.1089/109493103322011560]
23. Mazza M, Kammler-Sücker K, Leménager T, Kiefer F, Lenz B. Virtual reality: a powerful technology to provide novel insight into treatment mechanisms of addiction. Transl Psychiatry 2021 Dec 6;11(1):617. [doi: 10.1038/s41398-021-01739-3] [Medline: 34873146]
24. Park CB, Choi JS, Park SM, et al. Comparison of the effectiveness of virtual cue exposure therapy and cognitive behavioral therapy for nicotine dependence. Cyberpsychol Behav Soc Netw 2014 Apr;17(4):262-267. [doi: 10.1089/cyber.2013.0253] [Medline: 24555521]
25. Moon J, Lee JH. Cue exposure treatment in a virtual environment to reduce nicotine craving: a functional MRI study. Cyberpsychol Behav 2009 Feb;12(1):43-45. [doi: 10.1089/cpb.2008.0032] [Medline: 19025463]
26. Ferrer-García M, García-Rodríguez O, Gutiérrez-Maldonado J, Pericot-Valverde I, Secades-Villa R. Efficacy of virtual reality in triggering the craving to smoke: its relation to level of presence and nicotine dependence. Stud Health Technol Inform 2010;154:123-127. [Medline: 20543283]
27. Hoffmann S, Gerhardt S, Mühle C, et al. Associations of menstrual cycle and progesterone-to-estradiol ratio with alcohol consumption in alcohol use disorder: a sex-separated multicenter longitudinal study. Am J Psychiatry 2024 May 1;181(5):445-456. [doi: 10.1176/appi.ajp.20230027] [Medline: 38196336]
28. Heatherton TF, Kozlowski LT, Frecker RC, Fagerström KO. The Fagerström Test for Nicotine Dependence: a revision of the Fagerström Tolerance Questionnaire. Br J Addict 1991 Sep;86(9):1119-1127. [doi: 10.1111/j.1360-0443.1991.tb01879.x] [Medline: 1932883]
29. Klein EM, Brähler E, Dreier M, et al. The German version of the Perceived Stress Scale—psychometric characteristics in a representative German community sample. BMC Psychiatry 2016 May 23;16:159. [doi: 10.1186/s12888-016-0875-9] [Medline: 27216151]
30. Tiffany ST, Drobes DJ. The development and initial validation of a questionnaire on smoking urges. Br J Addict 1991 Nov;86(11):1467-1476. [doi: 10.1111/j.1360-0443.1991.tb01732.x] [Medline: 1777741]
31. Schubert TW. [Experience of presence in virtual exercises: a scale for measuring spatial presence, involvement, and reality judgment]. Z Medienpsychol 2003. [doi: 10.1026/1617-6383.15.2.69]

XSL•FO
RenderX

32. Kennedy RS, Lane NE, Berbaum KS, Lilienthal MG. Simulator Sickness Questionnaire: an enhanced method for quantifying simulator sickness. Int J Aviat Psychol 1993 Jul;3(3):203-220. [doi: 10.1207/s15327108ijap0303_3]
33. Melo M, Gonçalves G, Vasconcelos-Raposo J, Bessa M. How much presence is enough? qualitative scales for interpreting the Igroup Presence Questionnaire score. IEEE Access 2023;11:24675-24685. [doi: 10.1109/ACCESS.2023.3254892]
34. Lee JS, Namkoong K, Ku J, et al. Social pressure-induced craving in patients with alcohol dependence: application of virtual reality to coping skill training. Psychiatry Investig 2008 Dec;5(4):239-243. [doi: 10.4306/pi.2008.5.4.239] [Medline: 20046344]
35. Andreatta M, Pauli P. Generalization of appetitive conditioned responses. Psychophysiology 2019 Sep;56(9):e13397. [doi: 10.1111/psyp.13397] [Medline: 31152454]
36. Mucha RF, Pauli P, Weber M, Winkler M. Smoking stimuli from the terminal phase of cigarette consumption may not be cues for smoking in healthy smokers. Psychopharmacology (Berl) 2008 Nov;201(1):81-95. [doi: 10.1007/s00213-008-1249-x] [Medline: 18704373]
37. Acquadro C, Desvignes-Gleizes C, Mainy N, Hankins M, Weitkunat R, Chrea C. Measurement properties of the translations of instruments evaluating the subjective effects of tobacco- and nicotine-containing products: a systematic review of the literature. F1000Res 2019;8:2056. [doi: 10.12688/f1000research.20595.1]
38. Lopez-Duran NL, Mayer SE, Abelson JL. Modeling neuroendocrine stress reactivity in salivary cortisol: adjusting for peak latency variability. Stress 2014 Jul;17(4):285-295. [doi: 10.3109/10253890.2014.915517] [Medline: 24754834]
39. Hernández-Serrano O, Ghiţă A, Fernández-Ruiz J, et al. Determinants of cue-elicited alcohol craving and perceived realism in virtual reality environments among patients with alcohol use disorder. J Clin Med 2021 May 21;10(11):2241. [doi: 10.3390/jcm10112241] [Medline: 34064120]
40. Kovacs M, Campos F, Vugts V. What we don't know the effect of realism in virtual reality on experience and behaviour augmented reality and virtual reality: the power of AR and VR for business. In: Augmented Reality and Virtual Reality: The Power of AR and VR for Business: Springer, Cham; 2019:45-57. [doi: 10.1007/978-3-030-06246-0_4]
41. Bowman DA, McMahan RP. Virtual reality: how much immersion is enough? Computer (Long Beach Calif) 2007;40(7):36-43. [doi: 10.1109/MC.2007.257]
42. Rose T, Nam CS, Chen KB. Immersion of virtual reality for rehabilitation—review. Appl Ergon 2018 May;69:153-161. [doi: 10.1016/j.apergo.2018.01.009] [Medline: 29477323]
43. Kilteni K, Groten R, Slater M. The sense of embodiment in virtual reality. Presence: Teleoperators Virtual Environ 2012 Nov;21(4):373-387. [doi: 10.1162/PRES_a_00124]
44. Riva G, Wiederhold BK, Mantovani F. Neuroscience of virtual reality: from virtual exposure to embodied medicine. Cyberpsychol Behav Soc Netw 2019 Jan;22(1):82-96. [doi: 10.1089/cyber.2017.29099.gri] [Medline: 30183347]
45. Gall D, Roth D, Stauffert JP, Zarges J, Latoschik ME. Embodiment in virtual reality intensifies emotional responses to virtual stimuli. Front Psychol 2021;12(674179):674179. [doi: 10.3389/fpsyg.2021.674179] [Medline: 34552525]
46. Oberdörfer S, Schraudt D, Latoschik ME. Embodied gambling—investigating the influence of level of embodiment, avatar appearance, and virtual environment design on an online VR slot machine. Front Virtual Real 2022;3. [doi: 10.3389/frvir.2022.828553]

Abbreviations

iPQ: iGroup Presence Questionnaire
QSU: Questionnaire of Smoking Urges
VAS: visual analog scale
VR: virtual reality

Edited by A Mavragani; submitted 17.01.25; peer-reviewed by A Beck, A Capobianco; revised version received 07.03.25; accepted 11.03.25; published 26.05.25.

Please cite as:
Eidenmueller K, Hoffmann S, Kammler-Sücker K, Wenger L, Mazza M, Mühle C, Stenger M, Meixner G, Kiefer F, Lenz B
Reactivity to Smoking Cues in a Social Context: Virtual Reality Experiment
JMIR Form Res 2025;9:e71285
URL: https://formative.jmir.org/2025/1/e71285
doi:10.2196/71285

© Katharina Eidenmueller, Sabine Hoffmann, Kornelius Kammler-Sücker, Leonard Wenger, Massimiliano Mazza, Christiane Mühle, Manuel Stenger, Gerrit Meixner, Falk Kiefer, Bernd Lenz. Originally published in JMIR Formative Research (https://formative.jmir.org), 26.5.2025. This is an open-access article distributed under the terms of the Creative Commons

XSL•FO
RenderX

Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Feasibility and Cultural Adaptation of a Community-Engaged Physical Activity Intervention for Hispanic Older Adults: Pilot Study

Zvinka Z Zlatar[1], PhD; Mikael Anne Greenwood-Hickman[2], MPH; Lazaro N Martinez Lujan[2], MD; Julie Cooper[2], MPA; Stefani Florez-Acevedo[2,3], MPH; David X Marquez[4], PhD; Rosa Gutierrez Aceves[1], BA; Andrea Paula Vargas[1]; Dori E Rosenberg[2], PhD

[1]Department of Psychiatry, University of California, San Diego, 9500 Gilman Drive MC 0811, La Jolla, CA, United States

[2]Kaiser Permanente Washington Health Research Institute, Seattle, WA, United States

[3]Department of Health Systems and Population Health, School of Public Health, University of Washington, Seattle, WA, United States

[4]Department of Kinesiology and Nutrition, University of Illinois Chicago, Chicago, IL, United States

Corresponding Author:
Zvinka Z Zlatar, PhD
Department of Psychiatry, University of California, San Diego, 9500 Gilman Drive MC 0811, La Jolla, CA, United States

Abstract

Background: The adult Hispanic population of the United States faces an increased risk of dementia compared to non-Hispanic White adults. Physical activity (PA) can help reduce dementia risk, but culturally adapted interventions for Hispanic populations are lacking. Culturally adapted interventions are needed to increase uptake and adherence to brain health promotion strategies in the Hispanic community.

Objective: The De Pie y a Movernos intervention aims to reduce barriers to participation in clinical research by culturally adapting a remotely based PA intervention for middle-aged and older Hispanic adults and establish its feasibility and acceptability through a pre-post pilot study. Findings from the cultural adaptation process will inform a stage II PA randomized controlled trial.

Methods: The adaptation process followed Barrera and Castro's 2006 cultural adaptation framework and included a literature review, translation of intervention materials, review by a Hispanic-comprised community advisory board and bilingual staff, and a pre-post pilot study (N=10) with subsequent focus groups to refine the intervention. The pilot intervention included the use of Fitbit activity trackers and 2 individualized goal-setting calls with a health coach over a period of 3 weeks. Feasibility and acceptability were assessed using both quantitative methods and qualitative focus groups. Primary quantitative outcomes included enrollment, recruitment, and completion rates, as well as acceptability (predetermined satisfaction survey scores ≥3). Focus groups were thematically coded to identify themes for participants' opinions about several aspects of the intervention and explore key barriers and facilitators to PA engagement to improve the planned stage II trial.

Results: Ten Hispanic adults (age: mean 62.7, SD 5.3 years; education: mean 11.8, SD 3.8 years; n=9, 90% female; n=9, 90% Spanish-speaking) participated in the pre-post pilot, with a 100% completion rate, 50% enrollment rate, and a recruitment rate of 5 participants per month. Acceptability was high (mean score 4.6, SD 0.3; range 1 - 5). Qualitative analyses indicated that participants had high satisfaction with the intervention. They expressed a preference for adding group-based activities and increased interaction with study staff. Key barriers to PA included lack of awareness about the benefits of PA, low self-efficacy, time constraints, health conditions, and weather, while facilitators included awareness of PA's cognitive benefits, social support from family or friends, accountability, enjoyable activities, self-efficacy, and Fitbit use. Insights from participants and community advisory board recommendations led to modifications for the larger trial, such as incorporating group-based elements for those who want them and adding an option for teleconference coaching calls.

Conclusions: The cultural adaptation process was essential in refining the intervention to align with the preferences of older Hispanic adults, which resulted in a feasible and acceptable intervention. Findings will inform a planned stage II randomized controlled trial aimed at promoting PA and reducing dementia risk in older Hispanic adults.

(JMIR Form Res 2025;9:e65489) doi:10.2196/65489

KEYWORDS

lifestyle; dementia risk reduction; brain health; sedentary; exercise

XSL•FO
RenderX

Introduction

Older adults from the Hispanic or Latino community (henceforth referred to as Hispanic) face numerous health disparities [1-4], including higher rates of Alzheimer disease and related dementias (ADRD) [5-7] and cardiometabolic conditions like diabetes, hypertension, and obesity, which are more prevalent in Hispanic populations compared to their non-Hispanic White counterparts [8-13]. Notably, cardiometabolic conditions increase the risk of developing ADRD, which is a growing public health concern [5,14-18]. Projections indicate that by the year 2060, there will be 3.2 million Hispanic older adults living with ADRD [7], highlighting the urgency for targeted and culturally adapted risk reduction strategies to promote cognition and brain health for this population.

The 2024 Lancet Commission on Dementia Prevention highlighted that addressing 14 modifiable risk factors, including physical inactivity, could account for up to 45% of dementia cases worldwide [19]. Regular physical activity (PA) is associated with improved cognitive function [20,21], reduced neurodegeneration [22], higher cerebral blood flow [23], and a lower risk of ADRD [24,25]. However, older Hispanic adults engage in PA at significantly lower rates than their White counterparts [26] in part due to institutional (less information or education about PA, financial constraints, and lower access to health care), societal (caregiving responsibilities, lack of childcare, and poor spousal support), and environmental barriers (weather, less access to green spaces, facilities, or safe places to exercise) [27,28]. These disparities highlight the need to develop novel and culturally adapted PA interventions that address these barriers and support engagement in PA among older Hispanic adults to improve brain health. Using community-engaged approaches such as cultural adaptation, involving a community advisory board (CAB), and including bicultural/bilingual staff can increase adherence and satisfaction by making PA and other lifestyle behavior interventions more culturally relevant, acceptable, and accessible to the target communities [29-34].

To address this gap, we developed the culturally adapted De Pie y a Movernos intervention, a fully remote and individualized PA intervention designed to gradually increase PA to reduce ADRD risk among middle-aged and older Hispanic adults. This manuscript describes the cultural adaptation process informed by community engagement strategies [35] and presents feasibility and acceptability outcomes from a brief pre-post pilot study and follow-up participant focus groups. We also summarize qualitative findings about participants' opinions of the study as well as perceived barriers and facilitators. Findings from this pilot trial will inform the development of a larger stage II randomized controlled trial (RCT) to determine the efficacy of the De Pie y a Movernos intervention to increase objectively measured PA while investigating changes in mechanisms of adherence and cognitive function.

Methods

Study Description

The goal of the De Pie y a Movernos study was to culturally adapt a PA intervention for the older and middle-aged Hispanic community based on participant and community feedback gathered during a short pre-post pilot study, with the ultimate goal to later test the adapted intervention in a larger, fully remote, stage II RCT. The pre-post pilot intervention lasted 3 weeks and was comprised of 2 phone calls with the study health coach. The Fitbit Inspire 2, a commercially available activity tracker with heart rate monitoring, was used as an intervention tool that helped the health coach guide individualized PA goals. See Figure 1 for a timeline of the study and its components.

XSL•FO
RenderX

Figure 1. Depiction of the De Pie y a Movernos pilot study timeline and main procedures. The study lasted 6 weeks in total when taking into consideration the baseline and post assessments. The De Pie y a Movernos intervention itself lasted 3 weeks and included a tech phone call to set up the activity tracker at week 1 (Fitbit Inspire 2), followed by 2 phone calls from a health coach at weeks 2 and 3 of the intervention (N=10). Focus groups were conducted after the postintervention assessments (n=9). Nine out of the 10 participants who completed the intervention participated in the focus groups. AD=Alzheimer disease; CAB: community advisory board; PA: physical activity.

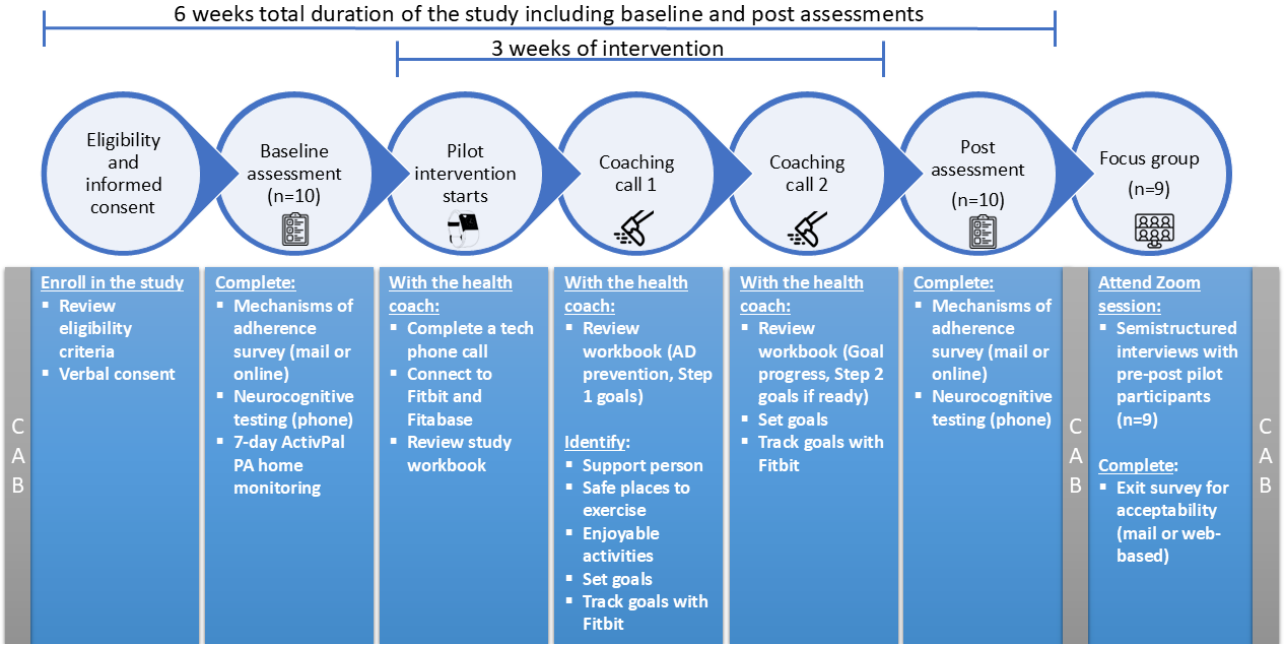

Cultural Adaptation Process

Overview

We followed the Barrera and Castro four-phased cultural adaptation approach [36,37]: (1) information gathering, (2) preliminary adaptation design, (3) preliminary adaptation tests, and (4) adaptation refinement. We approached each step in a community-engaged way [35] and considered several cultural dimensions underscored by Bernal et al [38] (eg, language, study team members from similar cultural backgrounds, incorporating cultural values and activities).

Information Gathering

We reviewed the literature for existing PA interventions developed for older Hispanic adults and convened a CAB comprised of members of the target population. The CAB was carefully crafted to represent middle-aged and older Hispanic adults with individuals from different sociodemographic backgrounds. The CAB was comprised of 8 members from different countries, all bicultural and some bilingual (with several monolingual Spanish speakers). One of the members is a PhD-level scientist with years of experience developing PA interventions for older Hispanic adults, 4 members are community health workers (promotores de salud) working with Hispanic older adults and Hispanic organizations, 1 member is involved in public health and has extensive experience working with the Hispanic community, another member has expertise in linguistics, and 1 member is a community-dwelling retired bicultural adult. Furthermore, the De Pie y a Movernos study team, which provided further feedback about translations, materials, and cultural relevance, was comprised of bilingual and bicultural faculty (PhD-level) and staff (graduate students, research assistants, and a community health worker) from Chile, Mexico, and Colombia.

Preliminary Adaptation Design

Based on findings and input from the information-gathering phase, we customized recruitment strategies and materials for the Hispanic community. The CAB met with the study team prior to conducting the pre-post pilot intervention, and all members of the CAB were offered a Fitbit Inspire 2 so they may provide feedback regarding the use and acceptability of the Fitbit as an intervention tool. The CAB helped the study team develop the study name and logo, review recruitment flyers/brochures, devise a suitable recruitment strategy, and review other participant-facing materials to ensure they were relevant to the Hispanic population.

Preliminary Adaptation Tests

We conducted a 3-week pre-post pilot study of the De Pie y a Movernos intervention designed and preliminarily adapted following the steps described above. Details of the pilot study design are outlined in the following section, “Intervention Pilot Study.”

Adaptation Refinement

Following the pre-post pilot study, we assessed feasibility and acceptability using a mixed methods approach through follow-up surveys, analysis of pilot intervention study data, and 2 focus groups with all pilot participants focused on gathering feedback about the intervention. We additionally conducted a debrief meeting with the CAB where these findings were presented, and additional adaptation feedback was shared with the study team. Details on feasibility and acceptability assessment, including focus group and qualitative analysis methods, can be found in the “Feasibility and Acceptability Measures” section.

XSL•FO
RenderX

Intervention Pilot Study

Overview

All participants were recruited between March 21, 2023, and May 26, 2023, from the San Diego County community and surrounding regions in Southern California following recommendations from the CAB. The members of the CAB provided ideas of places that could be targeted and input regarding the study pamphlet and flyers. Community outreach was conducted via our trained promotora de investigación (community research worker), who has established relationships with several community partners, agencies, and other community health worker networks that are knowledgeable of the Hispanic community of San Diego County. Our community research worker attended events such as health fairs and education sessions at community centers which are organized by several community organizations. During these community-facing events, the study community research worker recruited participants directly from the community.

Inclusion and Exclusion Criteria

Individuals ages 55 - 89 years who self-identified as Hispanic and who were able to have a phone conversation, able to walk 1 block unassisted, had no planned surgeries or travel that would interfere with participation for their 6 weeks in the study, were able to speak and read Spanish or English, were willing to wear a Fitbit and activPAL accelerometer, and had a smartphone were eligible to participate. Individuals who self-reported more than 90 minutes per week of PA, less than 6 hours per day sitting, employment in a physically active or standing job, a fall that resulted in hospitalization in the past year, having been told by a doctor they should not engage in PA, a new cancer diagnosis in the past year, or a myocardial infarction or arrhythmia in the past year were excluded. Additionally, individuals were ineligible if they scored ≤4 on the 6-item screener [39] administered by study staff at the initial screening call. Those who self-reported having a dementia diagnosis were also excluded.

Intervention Design

The pilot intervention was based on social cognitive theory and habit formation theory [40,41] and targeted the promotion of self-efficacy (gradually building behavior from sitting less and moving more to staying fit), habit strength (creating strong connections between existing daily habits and new PA behaviors), social support (finding a "buddy" to help achieve PA goals and increase accountability), and enjoyment of PA (finding individualized activities that are enjoyable for each participant to sustain adherence). To support these mechanisms, this pilot used a modified "staircase approach" [42] as seen in Figure 2. Step 1 involved focusing on sitting less and moving more throughout the day. Once participants felt ready, they could move to step 2, which encouraged structured moderate-intensity PA. Additionally, participants received a Fitbit Inspire 2 to help them track their progress and to inform goal setting. All participant Fitbits were connected to the Fitabase platform, a research-focused data management platform for Fitbit and other wearable devices [43], which allowed the health coach to access participant data to inform coaching and goal setting. The abbreviated intervention was administered over 3 weeks and consisted of a tech check-in call, where the coach provided technical assistance with the Fitbit device and Fitabase connection, and 2 coaching sessions focused on reviewing study workbook content and setting progressive step 1 and, if the participant was ready, step 2 goals.

Figure 2. Conceptual model to help participants gradually move toward moderate intensity activity by sitting less and moving more throughout the day. This conceptual model guided the intervention coaching calls, which were individualized for each participant.

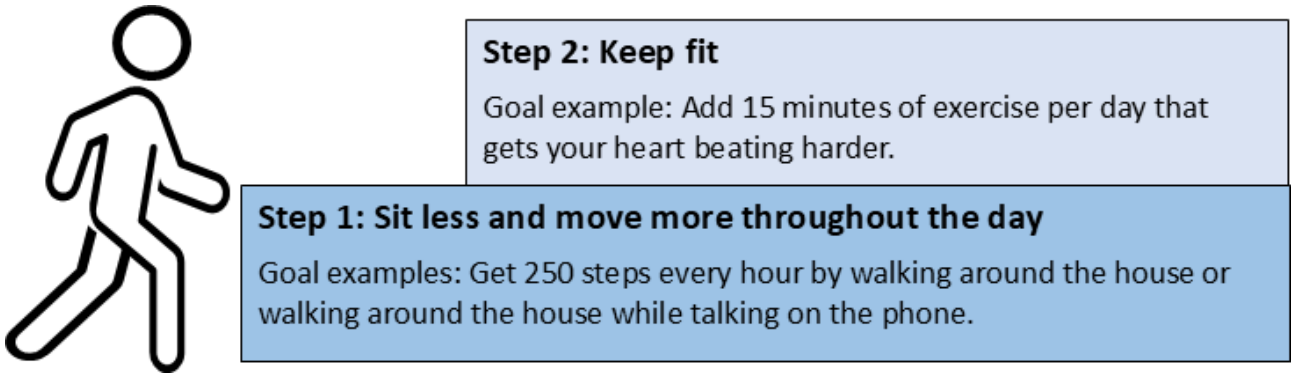

Health Screening

Based on guidance from the project's safety officer, who is a practicing physician, participants were screened for safety to engage in PA using the Exercise and Screening for You screening tool [44]. Only participants who answered "no" to all items were allowed to set goals for step 2 (keeping fit; see Figure 2). Participants who answered "yes" to any of the questions were encouraged to set step 1 goals only.

Ethical Considerations

This study was reviewed and approved by the Kaiser Permanente - Interregional Institutional Review Board (IRB) as the IRB of record. The IRB at the University of California, San Diego, relied on the IRB of record (UCSD IRB: #806367; KPiIRB #1987547-1). Because this study was carried out via telephone, the IRB approved a waiver of signed informed consent. All participants received a study information sheet via mail or email which was discussed with study staff via telephone. After a full explanation of all study procedures and allowing plenty of time for participants to ask questions about the study, participants

XSL•FO
RenderX

provided verbal consent to formally enroll in the study. All data were kept confidential and used only for study purposes. All participants were assigned a study ID for the data being collected. All identifiable data, such as contact information, were safely stored using Research Electronic Data Capture (REDCap) managed by Kaiser Permanente Washington Health Research Institute. REDCap is a password-protected and Health Insurance Portability and Accountability Act–compliant platform heavily used in research. As compensation for their participation in the pre-post intervention pilot study, participants received US $40 in cash. Additionally, they received US $50 for completing the follow-up focus group and kept the Fitbit Inspire 2 band (approximately US $100 in value).

Feasibility and Acceptability Measures

We used a mixed methods approach to assess feasibility and acceptability and to inform needed refinements of the culturally adapted intervention program. Below we detail both the qualitative and quantitative methods we used in this study.

Qualitative Measures and Analysis

Qualitative feedback on program acceptability and adaptations needed were collected via focus group with participants who completed the pre-post pilot intervention. All pilot intervention participants were invited to participate in a focus group discussion as part of their overall study participation. After completing the intervention, participants confirmed their interest in attending the focus group session during their final measurement phone call. Based on extensive input from all members of the study team, we developed a semistructured focus group discussion guide, a written document that included key questions and possible prompts. To inform ongoing cultural adaptation and refinement, our goal in this discussion was to elicit feedback about participant experiences with the De Pie y a Movernos pilot intervention program and how it resonated with their lives. Discussion topics included (1) aspects of the intervention or study measurement activities participants most enjoyed, (2) parts of the intervention or measurements participants disliked or would change, (3) reactions to completing study activities remotely by phone, (4) interest in and comfort with using video platforms for study activities, (5) barriers to participating in the intervention or study measurement activities, (6) types of physical activity participants most enjoy and resonate with, (7) reactions to the stair step structure (Figure 2) of the intervention, and (8) participants motivators to participate in the study. The final discussion guide is available in Multimedia Appendix 1.

Two study team members (ZZZ and SF-A) cofacilitated two 2-hour digital focus group sessions using the Zoom platform. ZZZ was a multiple principal investigator on the study and a bilingual Hispanic woman from Chile, trained in neuropsychology (PhD). SF-A was the primary measurement staff member and bilingual Hispanic woman from Colombia, trained in psychology and health services research. Based on participant preference, both focus groups were conducted in Spanish, though a separate English session was offered. Two additional study team members observed each focus group discussion, and team members kept informal notes during the focus group discussion to inform team debrief discussions.

Thematic analysis followed the recursive process outlined by Braun and Clarke [45,46], which consists of six key steps: (1) familiarization with the data, (2) code generation, (3) constructing themes from related codes, (4) reviewing and refining themes, (5) describing each theme, and (6) writing an analytic narrative. Two team members, bilingual in Spanish and English, reviewed translated transcripts along with audio recordings to ensure retained meaning. Each member of a 3-person coding team (the Hispanic woman from Colombia described earlier [SF-A], a Hispanic man from Mexico trained in medicine [LML], and a non-Hispanic woman from the United States [MAG-H] trained in anthropology and public health) carefully reviewed translated transcripts and independently coded each transcript (step 1). Using a primarily deductive approach to thematic analysis [45,47], we developed an initial code list informed by focus group questions. Codes were iteratively refined during coding and reconciliation by the coding team (final codebook available in Multimedia Appendix 2). After 1 round of independent coding, the coders met to reconcile discrepancies, refine code definitions, and assign final codes (step 2). Data from each code were reviewed, and themes were collaboratively constructed and refined by the coding team (steps 3 and 4). Final themes were each summarized along with supporting quotations in a summary table, and a summary of findings was presented to the larger study team and CAB (steps 5 and 6). Due to the small volume of data, qualitative coding was managed using the comment and search functions in Microsoft Word. We followed the Consolidated Criteria for Reporting Qualitative Research (COREQ) guidelines for reporting our qualitative methods and findings [48].

Quantitative Measures

Overview

We developed a study satisfaction survey in-house to obtain quantitative data related to participant's opinions of the pre-post pilot intervention. The satisfaction survey was collected at the end of the study and before the focus groups. All questions were scored on a Likert-type scale from 1=strongly disagree to 5=strongly agree, and covered satisfaction with study logistics (phone contacts and alternative contact modes), study measurement devices (Fitbit Inspire 2 and activPAL), measurement procedures, clarity of study materials, and the intervention's relevance to participants' lives. Example questions included (1) "Overall, I liked the study and found it helpful," (2) "The Fitbit device was helpful and easy to use," and (3) "I liked that I didn't have to travel anywhere for study activities and could participate by phone and mail." The complete satisfaction survey is available in Multimedia Appendix 3. For each item, an average score ≥3 (neutral, agree, or strongly agree) was the threshold for acceptability.

Feasibility

The recruitment rate was defined as the total number of participants recruited divided by the total number of months the trial was actively recruiting. The enrollment rate was defined as the number of enrolled participants divided by the number of prescreened participants x 100. Completion rate was defined as the number of participants who completed the intervention divided by the number of participants who enrolled in the

XSL•FO
RenderX

intervention x 100. Furthermore, to assess the feasibility of collecting key measures of interest in our target population, we collected a full measurement battery (objective neurocognitive testing, accelerometer assessment, and mechanisms of adherence) from all pilot participants at the baseline session and after the intervention was completed. Successful completion of these measures was considered an indicator of feasibility. We present descriptive statistics to demonstrate the successful collection and completion of these assessments. We hypothesized that any change in these measures would trend toward increased PA, promotion of mechanisms of adherence, and maintenance of cognition. However, we did not expect significant intervention effects on these outcomes of interest in the abbreviated 3-week pre-post pilot intervention period.

PA Assessment

We collected activPAL accelerometer data at baseline only to test the feasibility of deploying research-grade accelerometers in the community. The activPAL is a very small and lightweight inclinometer [49], which is waterproofed in a plastic casing and adhered to participants' thighs using medical-grade adhesive. To facilitate the appropriate application and wearing of the activPAL, study staff checked in with participants and provided guidance during the tech call described previously. activPAL is valid for measuring steps and sitting time [50,51] and has been used in studies with Hispanic populations [52]. Participants wore the device for 7 days and kept a sleep log. Data were processed in PALanalysis using activPAL's VANE algorithm to estimate daily steps, standing time (minutes/day), stepping time (minutes/day) and sit-to-stand transitions averaged over the number of wear days.

Additionally, data from each participant's Fitbit Inspire 2 were extracted using the "Daily Activities" file from the Fitabase platform. Average daily steps, lightly active minutes, fairly or very active minutes, and sedentary minutes were calculated across all days in the first week after their completed tech call (baseline) and during week 6 after the coaching call (post intervention). Because we had access only to daily summary-level Fitbit data, it was not possible to identify and remove individual periods of device nonwear. To approximate this using the available summary data, the accumulation of fewer than 500 recorded steps in a calendar day was considered an indication that the device was not worn for a significant portion of the day. In these instances, the summary data from that day were considered invalid and removed from the analytic sample. Notably, the Fitbit was provided to participants after their baseline activPAL wear. Therefore, due to differences in when participants received and were instructed to wear each device, the summarized baseline data for the Fitbit did not align with the week of baseline activPAL wear, and the 2 summaries should not be directly compared.

Mechanisms of Adherence

Hypothesized mechanisms of adherence were measured at baseline and postintervention via REDCap electronic survey unless participants requested the materials be mailed to them. All mechanisms of adherence were assessed with scales that have validated and comparable Spanish and English versions [53-56] and were assessed as point-in-time measures that reflect the individual's state at the time of assessment. All assessments were conducted in the participant's language of preference. The Self-Efficacy for PA questionnaire is a 5-item scale (1=not confident and 5=extremely confident) that assesses confidence for engaging in PA in different contexts [54]. The Self-Report Habit Index is a 12-item index that assesses habit-building strength on a scale from 1 to 5 (1=never and 5=always) [55]. The Social Support for Exercise Scale has 3 subscales that assess support from family (10 items) and friends (10 items), and family rewards or punishments (3 items) around engaging in PA with responses ranging from 1=none to 5=very often [53]. The Physical Activity Enjoyment Scale consists of 16 items (1=totally disagree and 5=totally agree) that measure the enjoyment of PA [56].

Objective Neurocognition

Objective neurocognition was measured via telephone with the following subtests from the National Alzheimer's Coordinating Center Uniform Data Set version 3 neuropsychological assessment [57]: Craft story 21 Immediate and Delayed Recall, Number Span, Category Fluency, and Phonemic Verbal Fluency. The Uniform Data Set was translated into Spanish with cultural relevance in mind and linguistic equivalence to the English versions [58]. We also collected the Brief-Spanish English Verbal Learning Test as a detailed measure of verbal episodic memory available in English and Spanish [59].

Results

Participant Characteristics

A total of 20 participants expressed interest in the study and scheduled a screening call. Four individuals declined participation prior to screening, 5 were determined ineligible for screening, and 1 completed screening but withdrew prior to baseline assessment. Participants were between the ages of 55 and 70 years, 90% (n=9) were women, 90% preferred Spanish (n=9), and they had a mean education of 11.8 (SD 3.8) years (Table 1).

XSL•FO
RenderX

Table . Pre-post pilot participant characteristics (N=10).

Characteristics	Values
Age (years)	
Mean (SD)	62.7 (5.3)
Range	55 - 70
Sex (female), n (%)	9 (90)
Education (years)[a]	
Mean (SD)	11.8 (3.8)
Range	6 - 17
Race[a], n (%)	
White or Caucasian	5 (50)
Other	1 (10)
More than 1 race	3 (30)
Employment status[a], n (%)	
Retired	4 (40)
Employed part-time	1 (10)
Self-employed part-time	1 (10)
Other	3 (30)
Marital status, n (%)	
Married	6 (60)
Divorced	1 (10)
Separated	2 (20)
Widowed	1 (10)
Ethnic identity (multiple categories permitted), n (%)	
Hispanic or Latina	10 (100)
Mexican or Mexican American	9 (90)
No origin shared	1 (10)
Diabetes, n (%)	4 (40)
Hypertension, n (%)	4 (40)
High cholesterol[a], n (%)	5 (50)
Fall in the last 12 months, n (%)	1 (10)
Smoking, n (%)	
Ever smoked	4 (40)
Current Smoker	0 (0)

[a]Some values are missing from the participant characteristics table due to participant nonresponse: education (n=2); race (n=1); employment status (n=1), and high cholesterol (n=1). All health indicators are based on participant self-report during the baseline screening phone call.

Qualitative Measures Findings

Overview

A total of 9 out of 10 participants who completed the De Pie y a Movernos pilot study participated in a focus group session in July 2023 (session 1: n=2; session 2: n=7). Only 1 participant did not attend due to scheduling conflicts. Below, we summarize the key participant feedback and themes that arose in the group discussion. A summary table of key themes and feedback, along with supporting quotations, can be found in Multimedia Appendix 4.

PA Devices

Participants found that using the Fitbit Inspire 2 was informative, motivating, easy, and enjoyable, with some challenges associated with setting it up, skin irritation, and connection difficulties. Having available tech support phone calls was identified as helpful. The participants identified that having instructional videos for the Fitbit would be helpful. Regarding

XSL•FO
RenderX

the activPAL, many participants thought that it was easy to wear, whereas others identified that the device did not work as intended (did not record their data) and that it irritated the skin. There was confusion about whether the activPAL was continuously recording and sending data to the study team.

Study Materials

Study materials were found to be relevant and helpful. Participants thought the materials made program goals clear and easy to connect with. However, participants suggested that our materials could be more organized with a clear order for the participants to follow seamlessly. They also described that it would be helpful to present the same information in multiple learning formats to accommodate varying learning styles (eg, narrative and pictures or videos). Participants liked getting to know the study team through the printed biographies and photographs provided to them (which were mailed or emailed to them before their baseline session). Participants wanted more information about cognitive health broadly and its relationship with PA.

Study Activities

Participants expressed that the health coaching calls were helpful in setting goals, generating ideas, breaking down content into achievable units, and allowing them to ask questions and receive support. They also identified that the coaching calls were helpful in providing accountability and follow-up on their goals. Participants were satisfied with the amount and content of communication with the study team. They expressed that the communications with the study team made it clear that they cared about the Hispanic culture and provided support, and some participants appreciated that the program was carried out via telephone as a low-tech option that is accessible to them. Some participants expressed the desire for videoconferencing options for the coaching calls and other aspects of the study. Furthermore, there was a strong desire for more in-person opportunities to connect with the community and meet other study participants.

Program Structure and Purpose

The goal of the program to develop interventions to reduce the risk of ADRD in Hispanics resonated with participants. This made them feel appreciated as members of the Hispanic community. Moreover, they found that this focus on ADRD risk reduction in Hispanics was important to the Hispanic community. They appreciated that the program was available in Spanish and that the team members were bilingual or bicultural and well acquainted with the Hispanic culture.

Barriers and Facilitators to PA

While focus group discussion focused primarily on feedback on the study materials, tools and procedures, barriers and facilitators to PA engagement surfaced incidentally during discussion. These are summarized along with supporting quotations in Multimedia Appendix 4. However, because we did not design our sampling framework or discussion guide to explore these topics, it is probable that additional barriers and facilitators exist for this community that were not discussed. Overall, reported barriers to PA, in general, included a lack of awareness about its health benefits, low self-efficacy, health and weather issues, lack of enjoyment, insufficient social support, ingrained sedentary habits, finding the Fitbit annoying or troublesome, and limited time availability. Facilitators of PA engagement in general included increased awareness of its cognitive and physical health benefits, a sense of accountability through study monitoring, improved self-efficacy in one's ability to be active, and the need for self-care. Enjoyment of activities, social support from friends or family, habit formation, positive experiences with Fitbit, and perceived improvements in physical and mental health were also discussed as facilitators of PA engagement.

Quantitative Measures Findings

Acceptability

The satisfaction survey was completed by 7 participants. Figure 3 shows the participant endorsement of each survey item (all items included in the satisfaction survey are listed on the y-axis of Figure 3). The threshold of acceptability was met for all items (scores≥3) and the average score was 4.6 (SD 0.3). The score for the survey item "There were too many phone calls" (Multimedia Appendix 3) was reversed and renamed since it was the only item for which a higher score would indicate less acceptability. In Figure 3, the reversed and renamed item is listed as: "The number of phone calls was acceptable."

XSL•FO
RenderX

Figure 3. Participant satisfaction survey. Responses to the satisfaction survey are based on feedback from 7 participants. There was missingness from participant nonresponse on the following items: activPAL satisfaction (n=1) and materials were meaningful (n=1). Error bars denote the SD. The vertical red line indicates the threshold score (mean=3) for an item to be considered acceptable to participants.

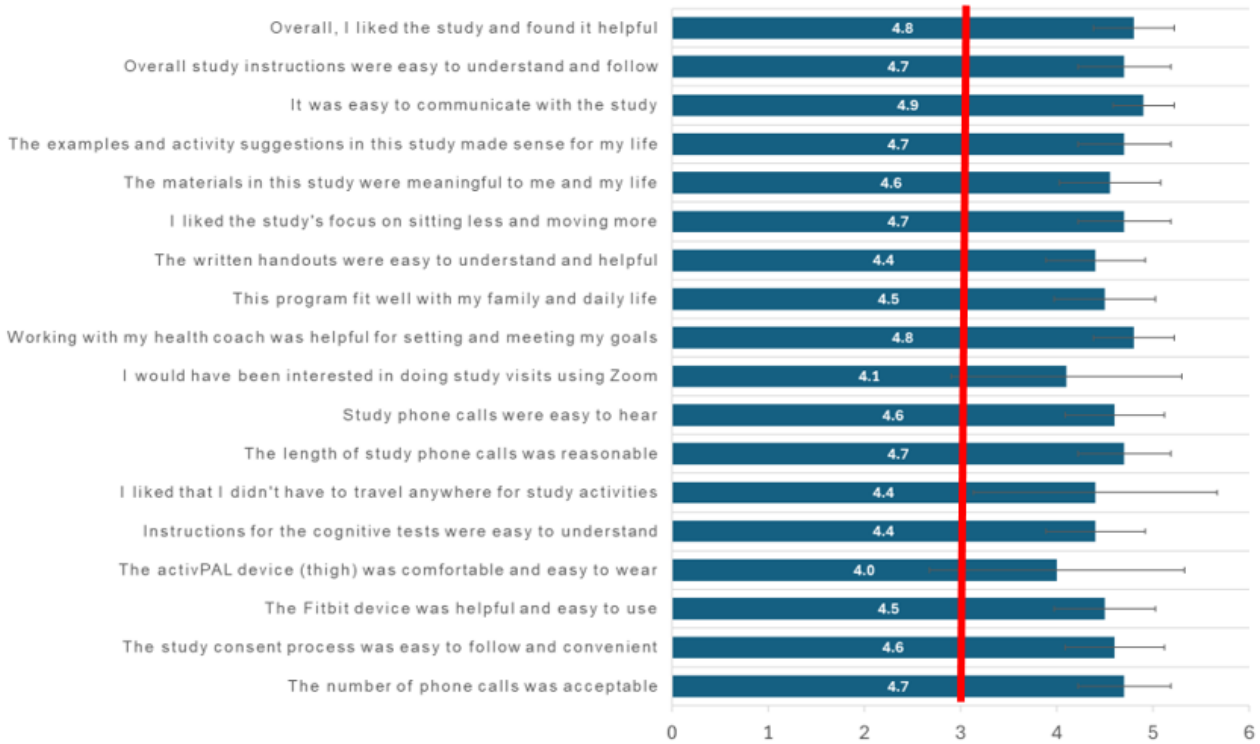

Feasibility

A total of 10 out of 20 prescreened Hispanic adults were eligible and consented to participate, yielding an enrollment rate of 50%. Recruitment took place between March 21, 2023, and May 26, 2023, yielding a recruitment rate of 5 participants per month. All participants who enrolled in the intervention completed the 3-week intervention, leading to a completion rate of 100%. All participants who completed a baseline assessment also completed the follow-up assessment (some missing items as per Table 2). We successfully obtained questionnaire data via REDCap survey or via mail from all 10 participants (9 in Spanish and 1 in English). We also successfully carried out telephone coaching calls without any complaints of hearing difficulties. Seven out of 10 participants were able to successfully complete the baseline activPAL measurement and return valid data. Missing activPAL data were due to device failure (n=1), mail loss (n=1), and participant nonwear (n=1). All participants were able to successfully install the Fitbit app on their smartphones and link their accounts to the Fitabase platform for the health coach to access the Fitbit data. Some participants needed substantial technology support from study staff and technology-savvy family members to set up their accounts, but they were all able to successfully complete this step. Table 2 shows a summary of pre-post means and SDs from the measures we collected at each assessment session. Although we did not expect any changes in our measures of interest given the short pilot intervention of 3 weeks, mean scores trended toward improvement for most neurocognitive tests, enjoyment of PA, social support for exercise, habit strength, and self-efficacy scores increased from pre- to postintervention (not tested for statistical significance given the focus on feasibility metrics and not effectiveness).

XSL•FO
RenderX

Table . Descriptive statistics for all the measures of interest (N=10). All measures were collected at baseline and postintervention (after 3 weeks), except for the activPAL which was only collected at baseline. These measures were not collected to determine efficacy of the intervention but rather to establish the feasibility of completing these assessments for a planned stage II randomized controlled trial.[a]

	Baseline, mean (SD)	Postintervention, mean (SD)
Fitbit metrics		
Steps/day	7872 (3723)	6932 (2531)
Lightly active minutes/day	267 (125)	245 (90)
Fairly or very active minutes/day	275 (128)	250 (90)
Sedentary minutes/day	747 (340)	804 (237)
activPAL measures[b]		
Steps/day	5184 (1379)	—[c]
Standing time (min/day)	319.5 (155.5)	—
Stepping time (min/day)	71.1 (20.4)	—
Sit-to-stand transitions	46 (8)	—
Mechanisms of adherence		
Physical activity enjoyment[d]	60.0 (15.2)	69.7 (9.7)
Social support for exercise[e]	18.5 (7.7)	26.8 (9.4)
Habit strength[f]	2.3 (0.5)	2.9 (0.3)
Self-efficacy for physical activity[g]	8.4 (2.5)	12.2 (3.8)
Objective neurocognition (National Alzheimer's Coordinating Center Uniform Data Set version 3)		
Craft Story 21 Recall Immediate – Verbatim[h]	15.7 (6.4)	20.9 (8.8)
Craft Story 21 Recall Immediate – Paraphrase[i]	6.3 (5.3)	5.4 (2.4)
Craft Story 21 Recall Delay – Verbatim[h]	13.8 (6.4)	18.8 (9.6)
Craft Story 21 Recall Delay – Paraphrase[i]	7.1 (5.1)	6.2 (3.0)
Verbal Fluency Total F and L words (P and M in Spanish)[j]	20.1 (4.8)	23.1 (8.4)
Animals[j]	19.0 (4.8)	19.3 (4.7)
Vegetables[j]	12.8 (4.2)	14.8 (3.5)
Number span test forward total correct sequences[k]	5.5 (2.0)	6.1 (1.9)
Number span test backward total correct sequences[k]	5.2 (1.5)	6.3 (1.3)
B-SEVLT trial 1[l]	7.7 (2.9)	9.1 (1.9)
B-SEVLT trial 2[l]	10.0 (2.4)	11.7 (1.9)
B-SEVLT trial 3[l]	11.7 (1.7)	13.0 (1.4)
B-SEVLT recall[l]	10.0 (2.6)	12.2 (2.6)

[a]Data missing due to participant nonresponse for the following items: baseline—physical activity enjoyment (n=2); postintervention—physical activity enjoyment (n=1), social support for exercise (n=2), self-efficacy for physical activity (n=1), and physical activity habit strength (n=1).

[b]Based on 7 participants who completed an activPAL wear with at least 6 days of valid data. Collected only at baseline.

[c]Not applicable.

[d]Range 16 - 80, higher scores indicate more enjoyment of physical activity.

[e]Range 10 - 50, higher scores indicate more social support for exercise from friends and family.

XSL•FO
RenderX

[f]Range 1 - 5, higher scores indicate stronger habit strength for physical activity.

[g]Range 5 - 25, higher scores indicate higher self-efficacy for physical activity.

[h]Range 0 - 44, higher scores indicate more story information was recalled verbatim.

[i]Range 0 - 25, higher scores indicate more story information was recalled while paraphrasing.

[j]Score represents the number of correct items provided within 1 minute for each letter and category. Higher numbers indicate greater verbal fluency.

[k]Range 0 - 14, higher scores indicate better performance

[l]B-SEVLT: Brief Spanish-English Verbal Learning Test; range 0 - 15 for each trial.

Discussion

We successfully developed and culturally adapted the De Pie y a Movernos intervention, based on established theoretical frameworks and using community-engaged strategies to determine feasibility and acceptability outcomes from a brief pre-post pilot intervention with 10 Hispanic older adults. Both qualitative and quantitative data indicated that the De Pie y a Movernos intervention was feasible and acceptable to participants. Regarding feasibility, there was a completion rate of 100%, an enrollment rate of 50%, and a recruitment rate of 5 participants per month. Acceptability criteria were well exceeded (mean 4.6, SD 0.3), with all participants rating the intervention as satisfactory (defined as scores≥3 on the satisfaction survey). Qualitative analyses indicated that participants overall had high satisfaction with the intervention. They expressed a preference for adding group-based activities and increased interaction with study staff. Some of the key barriers to performing PA generally included time constraints, health conditions, and weather, while facilitators included awareness of PA's cognitive benefits, social support from family or friends, accountability, enjoyable activities, self-efficacy, and Fitbit use.

Many elements of the study were well-liked by participants and will be kept in the next iteration, while some adaptations will be made to the next iteration based on feedback from the CAB and participants. These include new training videos (in English and Spanish) about how to set up technologies used in the study like the Fitbit, which will be shared via text message or email (depending on preference). Since older Hispanic adults may face challenges in adopting some technologies due to limited access, digital literacy, and language barriers [60,61], we will also encourage participants to seek help from a loved one or family member (partner, children, friends, and caregivers) to help them set up their Fitbit prior to reaching out to the study team for further assistance. Having a loved one support the technological aspects of the study, while the team provides layers of support, from written instructions to instructional videos in the participant's language of preference, should help reduce the technological barriers related to using the Fitbit.

Participants also reported wanting more contact with the other study participants as a group. Group sessions were not feasible because this study was designed to be fully remote to reduce transportation barriers (ie, lack of access to reliable transportation or long distances to research facilities) that are commonly faced by the Hispanic community [62,63], which limit their ability to participate in clinical research [64], exacerbating health disparities [65,66]. The remote format rather than a group setting is also more conducive to accommodating the individualized approach to health coaching that has been successful in our previous research [67] and aids in building self-efficacy through gradually sitting less and building toward becoming more physically active. However, to facilitate more group connection in the future iteration of the program, we will provide participants with personalized resources where they can exercise for free, or at low cost with groups of people in their community or neighborhoods. These resources will be shared with participants who value group activity and who need a sense of community during their health coaching calls. Participants also expressed a desire to have some study contacts occur via teleconference platform, which will be accommodated in the larger trial by offering coaching calls via telephone or teleconference. Moreover, the study will create a newsletter that will be sent to all participants regularly highlighting study progress, status, team updates, and finally, the results of the study.

Following Consolidated Standards of Reporting Trials (CONSORT) 2010 recommendations for pilot and feasibility studies [68], we present pre-post data on the proposed outcomes (objectively measured PA via activPAL and Fitbit), potential mechanisms of adherence, and other measures of interest to assess the feasibility of completion (for descriptive purposes). While PA pre- to posttest did not differ based on Fitbit, there was no true baseline period for the Fitbit as people started to use it immediately after receiving it. The activPAL was worn as an indicator of baseline PA and was worn prior to receipt of the Fitbit. We saw improvement in psychosocial mechanisms over the course of the study as well as some neurocognitive outcomes. However, given the small sample size, short duration, and lack of a comparison group, we cannot interpret these findings. As mentioned earlier, these measures were collected to assess the feasibility of completion, not for inferential purposes. The larger stage II trial will investigate the efficacy of the De Pie y a Movernos intervention on increasing PA, mechanisms of adherence, and potential effects on cognitive function.

This study is not without limitations. First, the pre-post pilot intervention was comprised of a small sample size of 10 participants and may not have been fully representative of the target community. Second, the duration of the intervention was short (3 weeks), which does not allow us to test the efficacy of the pilot intervention. That said, the goal of the pilot intervention was not to test efficacy, but rather to aid in the cultural adaptation process and determine the interventions' feasibility and acceptability to inform a planned stage II trial. Third, the Fitbit and activPal measurement periods were nonoverlapping, thus not allowing for a true comparison between these 2 PA measurement modalities. There are several strengths of this study that should be highlighted. There are few existing studies on culturally adapting PA interventions for brain health in older

XSL•FO
RenderX

Hispanics [69]. We involved a diverse CAB comprised of adults from the target population with several occupational and educational backgrounds, a fully bilingual and bicultural staff, and the development of the intervention based on established behavior change theory [40,41], which has been supported by previous studies conducted by the research team [67,70].

In the development and cultural adaptation of the De Pie y a Movernos intervention, we followed a community-engaged approach which led to a feasible and acceptable intervention with a 100% completion rate in our first iteration. These findings are consistent with previous studies showing good adherence and enjoyment of PA interventions that are culturally adapted and learn from participant feedback in the Hispanic community [33,69,71]. The CAB input was crucial to the development of a recruitment plan and recruitment materials, further highlighting the importance of community engagement in the design of research studies [32]. The combination of participants and CAB's input informed the development of the culturally adapted De Pie y a Movernos intervention to be further tested in a large stage II RCT. The ultimate goal of the larger trial is to help older Hispanic adults increase their PA engagement to preserve their brain health and reduce dementia risk.

In conclusion, the De Pie y a Movernos intervention was deemed feasible and acceptable, meeting the necessary milestones to justify efficacy testing in a larger stage II PA trial. We highlight the importance of including the target community in intervention development from study conception. Using a community-engaged approach and obtaining participant feedback from a small pilot study allowed us to develop a feasible and acceptable intervention that was culturally adapted to address the specific concerns of the Hispanic community. In the next iteration of the study, we will implement the changes highlighted during the focus groups and test the De Pie y a Movernos intervention for efficacy to increase PA while continuing to obtain participant and CAB feedback for further refinement.

Acknowledgments

The authors would like to acknowledge Rosa Gutierrez-Aceves for her recruitment efforts, all the pilot participants for their time and valuable feedback, and the members of the community advisory board (David Márquez, PhD, Cecilia Guzmán, María Martinez, Daisy Parra-Padilla, MPH, José Becerra, Tina Castillo, Deysi Merino González, and Silvia Martinez), all of whom were instrumental in the cultural adaptation process. Research reported in this publication was supported by the National Institute on Aging of the National Institutes of Health under award R61AG077969. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Data Availability

The data generated from the pre-post pilot study are available from the corresponding author upon request.

Authors' Contributions

ZZZ wrote the manuscript and facilitated focus groups. MAG-H provided data for Table 1 and qualitative data expertise. RGA contributed to participant recruitment and screening procedures. JC contributed to project management and overseeing study activities. DXM provided scientific consultation and was a member of the community advisory board. SF-A conducted the focus groups and helped with coding. LML was the health coach and contributed to qualitative coding. APV was part of the evaluation team. ZZZ and DER conceptualized the study aims, obtained funding, and supervised all aspects of the study. All authors reviewed drafts, assisted in writing, and provided review of the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Final focus group discussion questions.
[DOCX File, 164 KB - formative_v9i1e65489_app1.docx]

Multimedia Appendix 2
Final qualitative codebook.
[DOCX File, 15 KB - formative_v9i1e65489_app2.docx]

Multimedia Appendix 3
Study satisfaction survey.
[PDF File, 107 KB - formative_v9i1e65489_app3.pdf]

Multimedia Appendix 4

XSL•FO
RenderX

Summary of key themes and pilot program feedback from participant focus groups.
[PDF File, 295 KB - formative_v9i1e65489_app4.pdf]

References

1. Vega WA, Rodriguez MA, Gruskin E. Health disparities in the Latino population. Epidemiol Rev 2009;31:99-112. [doi: 10.1093/epirev/mxp008] [Medline: 19713270]
2. Vega IE, Cabrera LY, Wygant CM, Velez-Ortiz D, Counts SE. Alzheimer's disease in the Latino community: intersection of genetics and social determinants of health. J Alzheimers Dis 2017;58(4):979-992. [doi: 10.3233/JAD-161261]
3. McGrath RP, Snih SA, Markides KS, et al. The burden of health conditions across race and ethnicity for aging Americans: disability-adjusted life years. Medicine (Baltimore) 2019 Nov;98(46):e17964. [doi: 10.1097/MD.0000000000017964] [Medline: 31725658]
4. Stickel A, McKinnon A, Ruiz J, Grilli MD, Ryan L. The impact of cardiovascular risk factors on cognition in Hispanics and non-Hispanic whites. Learn Mem 2019 Jul;26(7):235-244. [doi: 10.1101/lm.048470.118] [Medline: 31209118]
5. Quiroz YT, Solis M, Aranda MP, et al. Addressing the disparities in dementia risk, early detection and care in Latino populations: highlights from the second Latinos & Alzheimer's Symposium. Alzheimers Dement 2022 Sep;18(9):1677-1686. [doi: 10.1002/alz.12589] [Medline: 35199931]
6. Matthews KA, Xu W, Gaglioti AH, et al. Racial and ethnic estimates of Alzheimer's disease and related dementias in the United States (2015-2060) in adults aged ≥65 years. Alzheimers Dement 2019 Jan;15(1):17-24. [doi: 10.1016/j.jalz.2018.06.3063] [Medline: 30243772]
7. U.S. burden of Alzheimer's disease, related dementias to double by 2060. Centers for Disease Control and Prevention. URL: https://archive.cdc.gov/#/details?url=https://www.cdc.gov/media/releases/2018/p0920-alzheimers-burden-double-2060.html [accessed 2025-03-10]
8. Aguayo-Mazzucato C, Diaque P, Hernandez S, Rosas S, Kostic A, Caballero AE. Understanding the growing epidemic of type 2 diabetes in the Hispanic population living in the United States. Diabetes Metab Res Rev 2019 Feb;35(2):e3097. [doi: 10.1002/dmrr.3097] [Medline: 30445663]
9. Statistics about diabetes. American Diabetes Association. 2023. URL: https://diabetes.org/about-diabetes/statistics/about-diabetes [accessed 2024-04-15]
10. National diabetes statistics report 2020: estimates of diabetes and its burden in the united states. : Centers for Disease Control and Prevention; 2020 URL: https://www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetes-statistics-report.pdf [accessed 2025-05-20]
11. Gomez S, Blumer V, Rodriguez F. Unique cardiovascular disease risk factors in Hispanic individuals. Curr Cardiovasc Risk Rep 2022;16(7):53-61. [doi: 10.1007/s12170-022-00692-0] [Medline: 35669678]
12. Parker ED, Lin J, Mahoney T, et al. Economic costs of diabetes in the U.S. in 2022. Diabetes Care 2024 Jan 1;47(1):26-43. [doi: 10.2337/dci23-0085] [Medline: 37909353]
13. Powell-Wiley TM, Baumer Y, Baah FO, et al. Social determinants of cardiovascular disease. Circ Res 2022 Mar 4;130(5):782-799. [doi: 10.1161/CIRCRESAHA.121.319811] [Medline: 35239404]
14. Daviglus ML, Talavera GA, Avilés-Santa ML, et al. Prevalence of major cardiovascular risk factors and cardiovascular diseases among Hispanic/Latino individuals of diverse backgrounds in the United States. JAMA 2012 Nov 7;308(17):1775-1784. [doi: 10.1001/jama.2012.14517] [Medline: 23117778]
15. Daviglus ML, Pirzada A, Durazo‐Arvizu R, et al. Prevalence of low cardiovascular risk profile among diverse Hispanic/Latino adults in the United States by age, sex, and level of acculturation: the Hispanic Community Health Study/Study of Latinos. JAHA 2016 Aug 8;5(8). [doi: 10.1161/JAHA.116.003929]
16. Lamar M, Durazo-Arvizu RA, Sachdeva S, et al. Cardiovascular disease risk factor burden and cognition: implications of ethnic diversity within the Hispanic Community Health Study/Study of Latinos. PLoS One 2019;14(4):e0215378. [doi: 10.1371/journal.pone.0215378] [Medline: 31009492]
17. Márquez F, Tarraf W, Stickel AM, et al. Hypertension, cognitive decline, and mild cognitive impairment among diverse Hispanics/Latinos: Study of Latinos-Investigation of Neurocognitive Aging Results (SOL-INCA). JAD 2024;97(3):1449-1461. [doi: 10.3233/JAD-230424]
18. Stickel AM, Tarraf W, Gonzalez KA, et al. Central obesity, cardiometabolic risk, and cognitive change in the study of Latinos: investigation of neurocognitive aging. J Alzheimers Dis 2021;82(3):1203-1218. [doi: 10.3233/JAD-210314] [Medline: 34151803]
19. Livingston G, Huntley J, Liu KY, et al. Dementia prevention, intervention, and care: 2024 report of the Lancet Standing Commission. The Lancet 2024 Aug;404(10452):572-628. [doi: 10.1016/S0140-6736(24)01296-0]
20. Carson Smith J, Callow DD, Pena GS, et al. Exercise and protection from age-related cognitive decline. Curr Top Behav Neurosci 2024;67(263-280):263-280. [doi: 10.1007/7854_2024_501] [Medline: 39080244]
21. Augusto-Oliveira M, Arrifano GP, Leal-Nazaré CG, et al. Exercise reshapes the brain: molecular, cellular, and structural changes associated with cognitive improvements. Mol Neurobiol 2023 Dec;60(12):6950-6974. [doi: 10.1007/s12035-023-03492-8] [Medline: 37518829]

XSL•FO
RenderX

22. Boa Sorte Silva NC, Barha CK, Erickson KI, Kramer AF, Liu-Ambrose T. Physical exercise, cognition, and brain health in aging. Trends Neurosci 2024 Jun;47(6):402-417. [doi: 10.1016/j.tins.2024.04.004] [Medline: 38811309]
23. Zlatar ZZ, Hays CC, Mestre Z, et al. Dose-dependent association of accelerometer-measured physical activity and sedentary time with brain perfusion in aging. Exp Gerontol 2019 Oct 1;125(110679):110679. [doi: 10.1016/j.exger.2019.110679] [Medline: 31382010]
24. Nguyen S, LaCroix AZ, Hayden KM, et al. Accelerometer-measured physical activity and sitting with incident mild cognitive impairment or probable dementia among older women. Alzheimers Dement 2023 Jul;19(7):3041-3054. [doi: 10.1002/alz.12908] [Medline: 36695426]
25. Iso-Markku P, Kujala UM, Knittle K, Polet J, Vuoksimaa E, Waller K. Physical activity as a protective factor for dementia and Alzheimer's disease: systematic review, meta-analysis and quality assessment of cohort and case-control studies. Br J Sports Med 2022 Jun;56(12):701-709. [doi: 10.1136/bjsports-2021-104981] [Medline: 35301183]
26. Adult physical inactivity. Centers for Disease Control and Prevention. 2023 Nov 20. URL: https://www.cdc.gov/physicalactivity/data/inactivity-prevalence-maps/index.html [accessed 2025-05-20]
27. Bantham A, Taverno Ross SE, Sebastião E, Hall G. Overcoming barriers to physical activity in underserved populations. Prog Cardiovasc Dis 2021;64:64-71. [doi: 10.1016/j.pcad.2020.11.002] [Medline: 33159937]
28. Ramirez A. Research: latino communities lack accessible green space. Salud America. 2019 May 10. URL: https://salud-america.org/research-latino-communities-lack-accessible-green-space/ [accessed 2025-05-20]
29. Aisenberg E, Dwight-Johnson M, O'Brien M, Ludman EJ, Golinelli D. Building a community-academic partnership: implementing a community-based trial of telephone cognitive behavioral therapy for rural Latinos. Depress Res Treat 2012;2012:257858. [doi: 10.1155/2012/257858] [Medline: 23050133]
30. Cyril S, Smith BJ, Possamai-Inesedy A, Renzaho AMN. Exploring the role of community engagement in improving the health of disadvantaged populations: a systematic review. Glob Health Action 2015;8(1):29842. [doi: 10.3402/gha.v8.29842] [Medline: 26689460]
31. Maafs-Rodríguez A, Folta SC. Effectiveness of theory-based physical activity and nutrition interventions in aging Latino adults: a scoping review. Nutrients 2023 Jun 18;15(12):2792. [doi: 10.3390/nu15122792] [Medline: 37375696]
32. Persad-Clem R, Ventura LM, Lyons T, et al. Community engagement in behavioral medicine: a scoping review. Int J Behav Med 2024 Dec;31(6):1018-1034. [doi: 10.1007/s12529-023-10242-6] [Medline: 38057655]
33. Balbim GM, Aguiñaga S, Ajilore OA, et al. The effects of the BAILAMOSTM dance program on physical activity levels and cognition of older Latino adults: a pilot study. J Aging Health 2022 Jan;34(1):25-40. [doi: 10.1177/08982643211020996] [Medline: 34027686]
34. Nierkens V, Hartman MA, Nicolaou M, et al. Effectiveness of cultural adaptations of interventions aimed at smoking cessation, diet, and/or physical activity in ethnic minorities: a systematic review. PLoS One 2013;8(10):e73373. [doi: 10.1371/journal.pone.0073373] [Medline: 24116000]
35. Balls-Berry JE, Acosta-Pérez E. The use of community engaged research principles to improve health: community academic partnerships for research. P R Health Sci J 2017 Jun;36(2):84-85. [Medline: 28622404]
36. Barrera M, Castro FG. A heuristic framework for the cultural adaptation of interventions. Clin Psychol Sci Pract 2006;13(4):311-316. [doi: 10.1111/j.1468-2850.2006.00043.x]
37. Castro FG, Barrera M, Holleran Steiker LK. Issues and challenges in the design of culturally adapted evidence-based interventions. Annu Rev Clin Psychol 2010;6(213-239):213-239. [doi: 10.1146/annurev-clinpsy-033109-132032] [Medline: 20192800]
38. Bernal G, Bonilla J, Bellido C. Ecological validity and cultural sensitivity for outcome research: issues for the cultural adaptation and development of psychosocial treatments with Hispanics. J Abnorm Child Psychol 1995 Feb;23(1):67-82. [doi: 10.1007/BF01447045] [Medline: 7759675]
39. Callahan CM, Unverzagt FW, Hui SL, Perkins AJ, Hendrie HC. Six-item screener to identify cognitive impairment among potential subjects for clinical research. Med Care 2002 Sep;40(9):771-781. [doi: 10.1097/00005650-200209000-00007] [Medline: 12218768]
40. Bandura A. A social cognitive perspective on positive psychology. Int J Soc Psychol 2011 Jan;26(1):7-20. [doi: 10.1174/021347411794078444]
41. Gardner B, Lally P, Wardle J. Making health habitual: the psychology of "habit-formation" and general practice. Br J Gen Pract 2012 Dec;62(605):664-666. [doi: 10.3399/bjgp12X659466] [Medline: 23211256]
42. Dogra S, Copeland JL, Altenburg TM, Heyland DK, Owen N, Dunstan DW. Start with reducing sedentary behavior: a stepwise approach to physical activity counseling in clinical practice. Patient Educ Couns 2022 Jun;105(6):1353-1361. [doi: 10.1016/j.pec.2021.09.019] [Medline: 34556383]
43. Fitabase FAQ. Fitabase. URL: https://www.fitabase.com/how-it-works/faq/ [accessed 2024-04-15]
44. Chodzko-Zajko WJ, Resnick B, Ory MG. Beyond screening: tailoring physical activity options with the EASY tool. Transl Behav Med 2012 Jun;2(2):244-248. [doi: 10.1007/s13142-012-0134-7] [Medline: 24073117]
45. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]

XSL•FO
RenderX

46. Clarke V, Braun V. Teaching thematic analysis: overcoming challenges and developing strategies for effective learning. Psychologist 2013;26(2) [FREE Full text]
47. Boyatzis RE. Transforming Qualitative Information: Thematic Analysis and Code Development: SAGE; 1998.
48. Tong A, Sainsbury P, Craig J. Consolidated Criteria for Reporting Qualitative Research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
49. PAL family. activPAL. URL: https://www.palt.com/pals/ [accessed 2024-04-15]
50. Kozey-Keadle S, Libertine A, Lyden K, Staudenmayer J, Freedson PS. Validation of wearable monitors for assessing sedentary behavior. Med Sci Sports Exerc 2011 Aug;43(8):1561-1567. [doi: 10.1249/MSS.0b013e31820ce174] [Medline: 21233777]
51. Lyden K, Keadle SK, Staudenmayer J, Freedson PS. The activPAL™ accurately classifies activity intensity categories in healthy adults. Med Sci Sports Exerc 2017 May;49(5):1022-1028. [doi: 10.1249/MSS.0000000000001177] [Medline: 28410327]
52. Takemoto M, Schechtman M, Villa N, et al. Arriba por la Vida Estudio (AVE): study protocol for a standing intervention targeting postmenopausal Latinas. Contemp Clin Trials 2019 Apr;79:66-72. [doi: 10.1016/j.cct.2019.02.004] [Medline: 30771560]
53. Walker TJ, Heredia NI, Reininger BM. Examining the validity, reliability, and measurement invariance of the Social Support for Exercise Scale among Spanish- and English-language Hispanics. Hisp J Behav Sci 2019 Aug;41(3):427-443. [doi: 10.1177/0739986319854144] [Medline: 32536744]
54. Mendoza-Vasconez AS, Marquez B, Benitez TJ, Marcus BH. Psychometrics of the self-efficacy for physical activity scale among a Latina women sample. BMC Public Health 2018 Sep 5;18(1):1097. [doi: 10.1186/s12889-018-5998-0] [Medline: 30185171]
55. Gutiérrez-Sánchez A, Pino-Juste M. Psychometric validation of the Spanish version of the scale properties Self-Report Habit Index (SRHI) measuring physical exercise habits. Rev Esp Salud Publica 2011 Aug;85(4):363-371. [doi: 10.1590/S1135-57272011000400005] [Medline: 22392467]
56. Moreno JA, González-Cutre D, Martínez C, Alonso N, López M. Propiedades psicométricas de la Physical Activity Enjoyment Scale (PACES) en el contexto español [Article in Spanish]. Estudios de Psicología 2008 Jan;29(2):173-180. [doi: 10.1174/021093908784485093]
57. Weintraub S, Besser L, Dodge HH, et al. Version 3 of the Alzheimer Disease Centers' neuropsychological test battery in the Uniform Data Set (UDS). Alzheimer Dis Assoc Disord 2018;32(1):10-17. [doi: 10.1097/WAD.0000000000000223] [Medline: 29240561]
58. Acevedo A, Krueger KR, Navarro E, et al. The Spanish translation and adaptation of the Uniform Data Set of the National Institute on Aging Alzheimer's Disease Centers. Alzheimer Dis Assoc Disord 2009;23(2):102-109. [doi: 10.1097/WAD.0b013e318193e376] [Medline: 19474568]
59. Breton J, Stickel AM, Tarraf W, et al. Normative data for the Brief Spanish-English Verbal Learning Test for representative and diverse Hispanics/Latinos: results from the Hispanic Community Health Study/Study of Latinos (HCHS/SOL). Alzheimers Dement (Amst) 2021;13(1):e12260. [doi: 10.1002/dad2.12260] [Medline: 34934802]
60. Dolan HR, Alvarez AA, Freylersythe SJ, et al. Barriers and facilitators for adopting a healthy lifestyle among Latina cancer survivors: a qualitative descriptive study. Support Care Cancer 2022 Mar;30(3):2649-2659. [doi: 10.1007/s00520-021-06596-w] [Medline: 34820710]
61. Philpot LM LM, Ramar P, Roellinger DL, McIntee MA, Ebbert JO. Digital health literacy and use of patient portals among Spanish-preferred patients in the United States: a cross-sectional assessment. Front Public Health ;12 [FREE Full text] [doi: 10.3389/fpubh.2024.1455395]
62. Arevalo M, Heredia NI, Krasny S, et al. Mexican-American perspectives on participation in clinical trials: a qualitative study. Contemp Clin Trials Commun 2016 Dec 15;4(52-57):52-57. [doi: 10.1016/j.conctc.2016.06.009] [Medline: 27570845]
63. Hildebrand JA, Billimek J, Olshansky EF, Sorkin DH, Lee JA, Evangelista LS. Facilitators and barriers to research participation: perspectives of Latinos with type 2 diabetes. Eur J Cardiovasc Nurs 2018 Dec;17(8):737-741. [doi: 10.1177/1474515118780895]
64. Aranda MP, Marquez DX, Gallagher-Thompson D, et al. A call to address structural barriers to Hispanic/Latino representation in clinical trials on Alzheimer's disease and related dementias: a micro-meso-macro perspective. Alzheimers Dement (N Y) 2023;9(2):e12389. [doi: 10.1002/trc2.12389] [Medline: 37287471]
65. Bodicoat DH, Routen AC, Willis A, et al. Promoting inclusion in clinical trials-a rapid review of the literature and recommendations for action. Trials 2021 Dec 4;22(1):880. [doi: 10.1186/s13063-021-05849-7] [Medline: 34863265]
66. Marquez DX, Perez A, Johnson JK, et al. Increasing engagement of Hispanics/Latinos in clinical trials on Alzheimer's disease and related dementias. Alzheimers Dement (N Y) 2022;8(1):e12331. [doi: 10.1002/trc2.12331] [Medline: 35910673]
67. Yaffe K, Vittinghoff E, Dublin S, et al. Effect of personalized risk-reduction strategies on cognition and dementia risk profile among older adults: the SMARRT randomized clinical trial. JAMA Intern Med 2024 Jan 1;184(1):54-62. [doi: 10.1001/jamainternmed.2023.6279] [Medline: 38010725]

XSL•FO
RenderX

68. Eldridge SM, Chan CL, Campbell MJ, et al. CONSORT 2010 statement: extension to randomised pilot and feasibility trials. Pilot Feasibility Stud 2016;2(1):64. [doi: 10.1186/s40814-016-0105-8] [Medline: 27965879]
69. Marquez DX, Bustamante EE, Aguiñaga S, Hernandez R. BAILAMOS: development, pilot testing, and future directions of a Latin dance program for older Latinos. Health Educ Behav 2015 Oct;42(5):604-610. [doi: 10.1177/1090198114543006] [Medline: 25108538]
70. Hays Weeks CC, Moore AA, Allison M, et al. The independent walking for brain health intervention for older adults: protocol for a pilot randomized controlled trial. JMIR Res Protoc 2023;12(1):e42980. [doi: 10.2196/42980]
71. Aguiñaga S, Marques IG, Kitsiou S, et al. BAILAMOS with mHealth technology! Improving physical activity and well-being in middle-aged and older Latinxs: a pre-post feasibility study. Health Educ Behav 2021 Oct;48(5):575-583. [doi: 10.1177/10901981211027517] [Medline: 34521228]

Abbreviations

ADRD: Alzheimer disease and related dementias
CAB: community advisory board
CONSORT: Consolidated Standards of Reporting Trials
COREQ: Consolidated Criteria for Reporting Qualitative Research
IRB: institutional review board
PA: physical activity
RCT: randomized controlled trial
REDCap: Research Electronic Data Capture

Edited by A Mavragani; submitted 20.08.24; peer-reviewed by M Lemonde, S Aguiñaga; revised version received 28.03.25; accepted 02.04.25; published 27.05.25.

Please cite as:
Zlatar ZZ, Greenwood-Hickman MA, Lujan LNM, Cooper J, Florez-Acevedo S, Marquez DX, Aceves RG, Vargas AP, Rosenberg DE Feasibility and Cultural Adaptation of a Community-Engaged Physical Activity Intervention for Hispanic Older Adults: Pilot Study JMIR Form Res 2025;9:e65489
URL: https://formative.jmir.org/2025/1/e65489
doi:10.2196/65489

© Zvinka Z Zlatar, Mikael Anne Greenwood-Hickman, Lazaro N Martinez Lujan, Julie Cooper, Stefani Florez-Acevedo, David X Marquez, Rosa Gutierrez Aceves, Andrea Paula Vargas, Dori E Rosenberg. Originally published in JMIR Formative Research (https://formative.jmir.org), 27.5.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Web-Based Stress Management for Working Adults With Attention-Deficit/Hyperactivity Disorder (ADHD): Single-Arm, Open Pilot Trial

Martin Oscarsson[1,2], MSc; Sandra Hammarbäck[1], MSc; Karolina Blom Wiberg[1], MSc; Alexander Rozental[3,4], PhD; Ylva Ginsberg[4], MD, PhD; Per Carlbring[1,5], PhD; Gerhard Andersson[4,6,7], PhD; Fredrik Jönsson[1], PhD

[1]Department of Psychology, Stockholm University, Albanovägen 12, Stockholm, Sweden
[2]Department of Psychiatry and Psychotherapy, University Hospital Bonn, Bonn, Germany
[3]Department of Health, Education and Technology, Luleå University of Technology, Luleå, Sweden
[4]Centre for Psychiatry Research, Department of Clinical Neuroscience, Karolinska Institutet, & Stockholm Health Care Services, Stockholm, Sweden
[5]School of Psychology, Korea University, Seoul, Republic of Korea
[6]Department of Behavioural Sciences and Learning, Linköping University, Linköping, Sweden
[7]Department of Biomedical and Clinical Sciences, Linköping University, Linköping, Sweden

Corresponding Author:
Martin Oscarsson, MSc
Department of Psychology, Stockholm University, Albanovägen 12, Stockholm, Sweden

Abstract

Background: National and international guidelines advocate for a multimodal approach to treating adult attention-deficit/hyperactivity disorder (ADHD), combining pharmacotherapy with psychological interventions. While recent reviews support cognitive behavioral therapy (CBT) as a viable treatment for ADHD in adults, evidence remains limited. Another challenge is the availability of psychological interventions, with stimulants remaining the primary treatment choice for adults with ADHD. One promising approach to increasing access to psychological interventions is the dissemination of internet-delivered CBT.

Objective: This study evaluated the feasibility, acceptability, and effects of a guided web-based stress management program specifically designed for working adults with ADHD. The intervention aimed to enhance quality of life by addressing stress, exhaustion, anxiety, and depression, commonly experienced by this population.

Methods: Thirty-six participants took part in a single-arm open trial, with assessments before, during, and after the intervention. The intervention consisted of 12 modules based on CBT principles, focusing on executive functioning, stress management, and emotion regulation, with clinician support on demand. Primary and secondary outcomes included quality of life (Adult ADHD Quality of Life Scale [AAQoL]), perceived stress (Perceived Stress Scale [PSS-10]), exhaustion (Karolinska Exhaustion Disorder Scale [KEDS]), anxiety (Generalized Anxiety Disorder 7-item Scale [GAD-7]), depression (Patient Health Questionnaire [PHQ-9]), and ADHD symptoms (the World Health Organization Adult ADHD Self-Report Scale [ASRS]).

Results: Results indicated a statistically and clinically significant improvement in quality of life (Cohen d=0.84), and a reduction in ADHD symptoms (d=0.98), as well as statistically significant reductions in perceived stress (d=0.83), exhaustion (d=1.12), anxiety (d=1.70), and depression (d=1.25). Improvements were sustained at a 12-week follow-up. A clinically significant improvement in quality of life was observed in 36% (13/36) of participants. Participants reported high satisfaction with the program and the guidance. Adherence was high, with an overall assessment response rate of 84%, a mean of 78% of modules opened, and no explicit dropouts. Twelve of the 36 participants reported negative effects. Qualitative content analysis of participants' written feedback revealed positive experiences and suggestions for improvement.

Conclusions: This study suggests promise for web-delivered interventions tailored to the needs of adults with ADHD, pending further research and development in controlled studies.

Trial Registration: OSF Registries osf.io/u2cdj; https://osf.io/u2cdj

(JMIR Form Res 2025;9:e66388) doi:10.2196/66388

KEYWORDS

attention deficit disorder with hyperactivity; cognitive behavioral therapy; quality of life; pilot projects; internet-based intervention

XSL•FO
RenderX

Introduction

Attention-deficit/hyperactivity disorder (ADHD) is a common neurodevelopmental disorder, affecting between 3% and 7% of adults worldwide [1]. Core symptoms are inattention, hyperactivity, and impulsivity. However, theoretical models propose a general deficiency in executive functions as a basis for the symptoms [2]. While the clinical presentation is well-defined in children, symptoms may have a more subtle and heterogeneous expression in adults. Despite growing evidence on the assessment and treatment of adult ADHD, many patients are underdiagnosed and undertreated, due to lack of recognition, misunderstandings, and poor access to specialized care [3,4]. Adults with ADHD have a general increased risk of experiencing mental health problems. A majority of adults with ADHD meet the criteria for at least 1 comorbid psychiatric disorder, most commonly mood, anxiety, or substance use disorders [5-7].

Adult ADHD has also been linked to stress and fatigue, and adults with ADHD generally score higher than those with no ADHD on measures of these constructs [8,9]. ADHD is also highly likely to be an underlying factor in cases of burnout [10]. In addition, adults with ADHD are more likely to experience stressful life events, such as divorce, loss of a job, and financial loss [11]. The inattentive nature of ADHD may lead to excessive proofreading, procrastination, and perfectionism, while impulsivity and lack of reflection may result in the acceptance of more tasks and responsibilities. This, combined with hyperactivity and difficulties relaxing, may lead to stress and exhaustion [12].

The impact of ADHD extends beyond individual well-being. Compared with their intellectual potential, many adults with ADHD underperform academically and professionally [13]. In most modern work environments, symptoms of inattention, hyperactivity, and impulsivity pose significant challenges [14]. In addition to impaired work performance and lower perceived effectiveness, adults with ADHD report significantly greater problems with attendance, teamwork, and social interaction at the workplace compared to controls [15]. The accumulation of performance impairment and interpersonal challenges, along with symptoms of psychiatric comorbidities, may significantly diminish the quality of life of adults with ADHD [16,17]. Additional negative outcomes include productivity losses and costs associated with sickness absence [18]. Several studies, such as the one by de Graaf et al [19], have indicated that adults with ADHD have more sickness absence days than controls, with some reporting rates twice [20] or several times as high [21].

National and international guidelines, such as those by Kooij et al [3] and the National Institute for Health and Care Excellence [22], advise a multimodal approach for the treatment of adult ADHD, combing pharmacotherapy with psychological interventions. Psychological interventions for adult ADHD include psychoeducation and cognitive behavioral therapy (CBT). While recent reviews, such as those by Fullen et al [23] and Nimmo-Smith et al [24], lend support to CBT as a viable treatment of ADHD in adults, evidence is limited due to a lack of available research and significant methodological issues with the included studies. Another limitation of CBT and other psychological interventions for adult ADHD is that of availability, with stimulants, in practice, remaining the treatment of choice for adults with ADHD [3].

While pharmacological treatment may be effective in reducing core symptoms of ADHD, functional impairment often remains. When discussing CBT for adults with ADHD and potential mechanisms of change, Ramsay [25] considered Barkley's theory of executive functioning and self-regulation [26,27]. Executive functions, such as behavioral inhibition, verbal and nonverbal working memory, and emotional self-regulation, are necessary for a person when organizing and executing goal-directed actions. Impaired executive functioning in adult ADHD interferes not only with the actual management of everyday tasks and activities, but also with the ability to modify dysfunctional behaviors associated with the disorder. In short, most adults with ADHD know what needs to change in terms of their behavior but fall short in implementing and maintaining changes. Consequently, Ramsay [25] characterized ADHD as an implementation disorder.

One way to increase access to psychological interventions is the dissemination of internet-delivered cognitive behavioral therapy (iCBT) [28]. iCBT with minimal clinician support has been shown to yield similar effects to face-to-face therapy for several psychiatric and somatic disorders, although repeated direct comparisons remain limited to a few specific conditions [29]. In a recent review and meta-analysis by Svärdman et al [30], iCBT was shown effective in reducing self-reported stress, as well as anxiety and depressive symptoms, among adults experiencing stress and stress-related disorders. iCBT has also been evaluated for adults with ADHD. In a recent randomized controlled trial (RCT) by Nasri et al [31], guided iCBT was found to yield significantly greater improvement on the main ADHD outcome measure than a treatment-as-usual control. In another recent RCT, Kenter et al [32] found that an unguided web intervention (MyADHD) based on CBT, dialectical behavior therapy and goal management training, was superior to a psychoeducation-only control in a sample of adults with ADHD. Earlier RCTs on adult ADHD include 1 by Pettersson et al [33], showing that unguided iCBT was superior to a waiting list control, and 1 by Moëll et al [34], showing that a CBT-inspired guided web-based course was superior to a waiting list control.

In a recent qualitative study on stress- and work-related psychological problems among adults with ADHD [35], participants reported few and unsatisfactory experiences of health care support beyond pharmacological treatment of the ADHD symptoms. Several participants specifically mentioned a desire for interventions designed for patients with a relatively high level of functioning, with an emphasis on executive functions. While psychoeducational interventions for adult ADHD may include information and exercises related to workplace functioning, such as in the study by Hirvikoski et al [36], it is not a primary focus in most common ADHD interventions. In fact, a recent review by Lauder et al [37] failed to identify any ADHD interventions specifically designed with

XSL•FO
RenderX

the workplace context and workplace functioning as a central component.

The limited offering and inadequate provision of psychological interventions for working adults with ADHD are particularly conspicuous when considering the high prevalence of ADHD and the potential of resource-efficient iCBT. Inspired by existing psychological interventions for ADHD, stress, and common psychiatric comorbidities, and informed by interviews with working adults with ADHD, we developed a guided web-based stress management program specifically for the adult population with ADHD. The program was based on CBT principles, aiming to enhance quality of life by addressing stress, exhaustion, anxiety, and depression. In this pilot study, we evaluate the feasibility, acceptability, and effects of this novel intervention.

Methods

Design

This was a single-arm open trial with repeated measures before, during, and after the intervention. The primary aim was to evaluate the feasibility, acceptability, and effects of a guided web-based stress management program for working adults with ADHD.

Preregistration

The research question, desired sample size, included variables, and planned main analysis were preregistered on Open Science Framework [38]. Preregistration was completed during the recruitment phase before any participants had started the intervention.

Participants

Information about the study was shared on social media by a large Swedish ADHD patient interest group (Riksförbundet Attention), who posted about the study on their Facebook page, and by the research team, who posted in relevant Facebook groups and on LinkedIn. The study was conducted entirely on the web, without involvement of any physical or clinical sites. All recruitment, screening, informed consent, and data collection took place via the secure web intervention platform iTerapi [39]. Accompanying the recruitment information was a link to a site on the iTerapi platform, where those interested were invited to read more about the study. The site provided details including participation criteria and a button to create an account and submit a notice of interest in participation. Provision of informed consent was required to submit a notice of interest.

Upon the creation of an account, prospective participants were sent a link to validate their email address and to complete a web-based screening questionnaire. The screening questionnaire included questions regarding sociodemographics and ADHD diagnosis, as well as self-report measures of quality of life, perceived stress, symptoms of exhaustion, anxiety, depression, and symptoms of ADHD. The questionnaire also included questions regarding possible concurrent psychological or pharmacological treatment of ADHD or mental health problems.

To be eligible for the study, participants needed to be aged 20 - 64 years (Swedish definition of working age), fluent in Swedish, have a diagnosis of ADHD, and work half-time or more. They also needed to score >19 on the Perceived Stress Scale (PSS-10; 75th percentile [40]), or >18 on the Karolinska Exhaustion Disorder Scale (KEDS; at risk for exhaustion disorder [41]), or >4 on the Generalized Anxiety Disorder 7-item Scale (GAD-7; mild anxiety [42]), or >9 on the Patient Health Questionnaire (PHQ-9; moderate depression [43]). Participants were excluded in cases of severe depression (PHQ-9>19), suicidality (PHQ-9 item 9>1), concurrent psychological treatment, start of psychological treatment scheduled for the next 3 months, psychotropic medication dosage not stable for at least 1 month, or change in psychotropic medication scheduled for the next 3 months.

In total, 47 individuals created an account on the treatment platform. One did not complete the screening questionnaire, while 4 met the exclusion criteria (1 on sick leave, 1 with a recent change in psychotropic medication, 1 with a scheduled change in psychotropic medication, and 1 with a score above 19 on the PHQ-9). The remaining 42 were invited to participate in the study, and 36 accepted the invitation. Table 1 shows the characteristics of the final sample.

XSL•FO
RenderX

Table . Demographic and clinical characteristics of participants (n=36) in a single-arm pilot trial of a web-based stress-management intervention for working adults with attention-deficit/hyperactivity disorder in Sweden.

Characteristics	Participants (n=36)
Age (years), mean (SD)	44.4 (8.29)
Gender, n (%)	
Women	28 (78)
Men	7 (19)
Other or did not want to assign	1 (3)
Education level, n (%)	
Elementary school	0 (0)
High school	4 (11)
University ≤3 years	7 (19)
University >3 years	25 (69)
Employment rate (percentage of full-time), mean (SD)	92.9 (11.7)
Psychotropic medication, n (%)	
Yes	31 (86)
No	5 (14)

Measures

The main outcome measure was quality of life, measured with the Adult ADHD Quality of Life Scale (AAQoL [44]). Secondary outcome measures were perceived stress, measured with the PSS-10; symptoms of exhaustion, measured with the KEDS; anxiety, measured with the GAD-7; depression, measured with the PHQ-9; and symptoms of ADHD, measured with the World Health Organization Adult ADHD Self-Report Scale (ASRS) Symptoms Checklist [45]. The frequency and time points of measurement are shown in Table 2. Week 12 is postintervention and week 24 is follow-up. Subscale results are reported for the PSS-10 and the ASRS. The ASRS was modified to reflect the last 4 weeks, and formatted without the shaded boxes, as is common in ADHD intervention research. The Negative Effects Questionnaire (NEQ-20 [46]) was used to explore negative effects and adverse events.

Table . Overview of outcome measures and assessment time points during a 12-week guided web-based stress-management intervention and 12-week follow-up for working adults with attention-deficit/hyperactivity disorder in Sweden.

Measure	Baseline	Week 2	Week 4	Week 6	Week 8	Week 10	Week 12	Week 24
AAQoL[a]	√	N/A[b]	√	N/A	√	N/A	√	√
ASRS[c]	√	N/A	√	N/A	√	N/A	√	√
PSS-10[d]	√	N/A	√	N/A	√	N/A	√	√
KEDS[e]	√	√	√	√	√	√	√	√
PHQ-9[f]	√	√	√	√	√	√	√	√
GAD-7[g]	√	√	√	√	√	√	√	√
NEQ-20[h]	N/A	N/A	N/A	N/A	N/A	N/A	√	N/A

[a]AAQoL: Adult Attention-Deficit/Hyperactivity Disorder Quality of Life.

[b]N/A: not applicable.

[c]ASRS: Adult ADHD Self-Report Scale.

[d]PSS-10: Perceived Stress Scale.

[e]KEDS: Karolinska Exhaustion Disorder Scale.

[f]PHQ-9: Patient Health Questionnaire.

[g]GAD-7: Generalized Anxiety Disorder 7-item Scale.

[h]NEQ-20: Negative Effects Questionnaire.

XSL•FO
RenderX

At week 12, participants were also invited to answer open-ended questions about their experiences of the intervention and rate their treatment satisfaction from 1 to 4 on 7 items.

Intervention

The intervention, called "Working with ADHD," was developed to increase the quality of life for working adults with ADHD by addressing and managing stress, exhaustion, anxiety, and depression. The program is based on CBT principles, incorporating psychoeducation, basic behavior analysis, goal setting, and simple behavioral experiments. The development process began with a comprehensive review of existing psychological interventions targeting ADHD, stress, and common psychiatric comorbidities.

As part of a previous qualitative study [35], 20 working adults with ADHD were asked about their interest in and attitudes toward a hypothetical web-based psychological intervention developed specifically for working adults with ADHD. In addition to being generally positive toward the concept, the participants underscored 3 things that would increase the likelihood of its success: treatment content presented in short, accessible modules with clear rationale; the option to choose modules based on needs and urgency; and the inclusion of clinician support.

The intervention had 3 main areas of focus. The first was compensatory strategies for executive function deficiencies, with a foundation in Barkley's theory of ADHD and Ramsay's proposed mechanisms of change. This included information and exercises regarding organization, planning, and prioritization, inspired by the CBT manual for adult ADHD by Safren et al [47]. Also included were information and exercises from CBT for procrastination, inspired by interventions developed by Rozental et al [48].

The second area of focus was stress management and recovery. Here, the theoretical basis was the cognitive behavioral model of clinical burnout by Almén [49]. Information and exercises revolved around the balance between stress reactivity and stress recovery, with an emphasis on recovery-facilitating behaviors. This also included information and exercises regarding sleep, inspired by the ADHD-specific behavioral insomnia intervention by Jernelöv et al [50].

The third and final area of focus was emotion regulation. The role of emotional dysregulation in the clinical presentation of adult ADHD has been extensively discussed in the literature, illustrated, for example, by the study by Hirsch et al [51]. Emotional dysregulation, as a component of adult ADHD, has also been linked to occupational impairment [52]. It may take different forms, including a hypersensitivity to criticism and failure [35], as well as impatience and irritability [52]. In this intervention, information and exercises revolved around assertiveness and self-esteem [53], as well as perfectionism [54] and self-compassion [55].

The intervention content was delivered in 12 modules over 12 weeks on the iTerapi platform. The platform is responsive and usable on computers, tablets, and smartphones. First, participants were granted access to a new module every Monday for 4 weeks. Then, participants were granted access to 6 new modules, to work on for 6 weeks, according to their preference and perceived needs. Finally, participants were granted access to 2 new modules, weeks 11 and 12, respectively. This is detailed in Table 3.

XSL•FO
RenderX

Table . Overview of the 12-module web-based stress management intervention ("Working with ADHD") delivered over 12 weeks to working adults with attention-deficit/hyperactivity disorder in Sweden.

Module	Theme	Content
1	ADHD[a], stress, self-assessment, and goals	ADHD, stress, and mental health; assessment of current situation; and goal formulation
2	Communication and boundary setting	Performance-based self-esteem, assessment of and working with self-esteem, self-compassion, regulating strong emotions, and communicating boundaries
3	Sleep	ADHD and sleep, sleep assessment, sleep hygiene, and sleep restriction
4	Recovery	The recovery paradox, assessment, recovery every day, practicing self-awareness, and finding balance in everyday life
5	Planning and prioritization	Use of calendar and to-do list, and prioritizing tasks
6	Problem-solving	Stepwise problem-solving, breaking tasks into smaller parts, and attention span
7	Perfectionism	Perfectionism, stress, and ADHD; assessment of perfectionist behaviors; practicing flexible thinking; and seeing the positive
8	Procrastination	Procrastination, stress, and ADHD; assessment; motivation; minimum effort; and managing distractions
9	Disclosing the diagnosis	Job seeking with ADHD and disclosing the diagnosis at the workplace
10	Rights and responsibilities at the workplace	General rights and responsibilities of employees; and the right to accommodation
11	Repetition and reflection	Summary of previous modules, assessment of what has been helpful, and noting progress
12	Beyond the program	Setbacks and relapses; and identifying risk situations

[a]ADHD: attention-deficit/hyperactivity disorder.

All modules were structured into brief sections, each alternating between information, assessment, and exercises. Most modules included 1500 - 2000 words in total. Participants were instructed to complete all assessments and exercises in writing, using their preferred medium (eg, notes app or spiral notebook). At the conclusion of each module, participants were offered suggestions for implementing new strategies in their daily lives. In addition, select sections within each module included prompts for participants to reach out to the research group for support.

Support was available on demand through asynchronous messaging on the treatment platform. All support was provided by a licensed psychologist with 5 years of clinical experience working with adults with ADHD. The support primarily consisted of encouragement and affirmation [56] and did not introduce new treatment content. Each participant also received 2 scheduled, prewritten messages per week: one on Mondays containing information about new modules or offering general encouragement to engage with the treatment content and another later in the week encouraging participants to reach out with questions or request feedback on the exercises. Messages from participants were answered at least once every weekday, that is, within 36 hours.

All participants received email notifications for new messages. If a participant had not logged in for a week, they were first contacted via email the following Monday and then twice by telephone later in the week. Participants were contacted via email and telephone in a similar manner if they did not respond to assessments. Study staff were immediately notified if a participant at any point scored PHQ-9 >19 (severe depression cutoff) or PHQ-9 item 9 >1 (suicidality cutoff), in which case the participant was contacted via telephone.

Analysis

Linear mixed models were fitted to estimate the fixed effects of time and the random effects of participant intercepts and participant slopes, using all available data from screening to postintervention. The models used unstructured covariance patterns and restricted maximum likelihood estimation for missing data. Paired-samples *t* tests (2-tailed) were used to compare full-scale follow-up scores with screening scores using last observation carried forward (LOCF) imputation. Effect sizes Cohen *d* were calculated on estimated random means for screening to postintervention, and raw means for screening to follow-up, using the mean differences and the pooled SDs.

XSL•FO
RenderX

The Jacobson and Truax [57] method was used to analyze clinical significance, using cutoff A (ie, 2 SDs in the direction of a functioning distribution) and reliable change index (RCI) on screening and postintervention AAQoL and full-scale ASRS data with LOCF imputation. Deterioration was defined as exceeding the RCI in a negative direction, while unchanged corresponded to a change score not exceeding the RCI. Analysis of missing data was done using one-way ANOVA. Adherence, treatment satisfaction, negative effects, and adverse events were explored descriptively. All statistical analyses were performed using jamovi (version 2.4; The jamovi project) [58,59]. Mixed models were run using the GAMLj module [60]. Jacobson-Truax classifications and RCI were calculated using the JTRCI R package [61].

Inductive qualitative content analysis [62] was used to explore data from participants' responses to the open-ended questions postintervention. The analysis was conducted by SH and KBW, under the supervision of MO. SH and KBW were final-year psychologist students at the time of the analysis. They had completed basic training in CBT and previously assisted MO in developing the treatment content. The analysis followed the procedure outlined by Krippendorff [63]. First, data were sampled from the participants, using the open-ended questions. Data were unitized, using the original wording of the participants and discarding data not relevant to the research questions. Then, the units of meaning were coded inductively, using the analysts' own understanding of the context. Finally, the codes were clustered, maintaining internal homogeneity and external heterogeneity between the final categories.

Ethical Considerations

This study was approved by the Swedish ethical review authority (diary nos.: 2022-04370-01 and 2022-07214-02), and all procedures followed relevant national and institutional guidelines. Informed consent was obtained digitally from all participants via the secure iTerapi platform, prior to the submission of interest in participation. Participants were provided with written information about the study, data handling, and their rights, and were encouraged to contact the principal investigator (FJ) with any questions before consenting. Study data were handled in pseudonymized form using a coded key, accessible only to the research team. All personal data were stored securely on encrypted servers or encrypted researcher devices (eg, university computers), either within the treatment platform or Stockholm University's approved storage systems. No financial compensation was provided to participants. No identifiable images or personal information is shown in the manuscript or Multimedia Appendix 1.

Results

Main Analyses

A mixed model revealed a statistically significant effect of time from screening to postintervention on the main outcome measure AAQoL ($F_{1,29.4}$=28.2; P<.001), corresponding to a large effect (d=0.84). This indicates an improved average quality of life among the participants. Mixed models also revealed statistically significant effects of time from screening to postintervention on the secondary outcome measures, all indicating lower average symptom levels among participants. The results of the mixed models, with corresponding effect sizes Cohen d, are shown in Table 4. Estimated marginal means are shown in Figure 1 and Table S1 in Multimedia Appendix 1.

Table . Results from linear mixed models estimating changes in all outcome measures from screening to postintervention in a single-arm pilot trial of a guided web-based stress management intervention for working adults with attention-deficit/hyperactivity disorder in Sweden.

Measure	*F* test (*df*)	*P* value	Cohen *d*
AAQoL[a]	28.2 (1, 29.4)	<.001	0.84
ASRS[b]	35.6 (1, 29.8)	<.001	0.98
ASRS (Inattention)	30.9 (1, 30.8)	<.001	1.12
ASRS (H/I)	28.8 (1, 28.8)	<.001	0.72
PSS-10[c]	17.3 (1, 31.5)	<.001	0.83
PSS-10 (Negative)	28.6 (1, 31.0)	<.001	1.01
KEDS[d]	26.1 (1, 33.5)	<.001	1.12
PHQ-9[e]	22.4 (1, 33.0)	<.001	1.25
GAD-7[f]	53.1 (1, 28.7)	<.001	1.70

[a]AAQoL: Adult Attention-Deficit/Hyperactivity Disorder Quality of Life.

[b]ASRS: Adult Attention-Deficit/Hyperactivity Disorder Self-Report Scale.

[c]PSS-10: Perceived Stress Scale-10.

[d]KEDS: Karolinska Exhaustion Disorder Scale.

[e]PHQ-9: Patient Health Questionnaire-9.

[f]GAD-7: Generalized Anxiety Disorder 7-item Scale.

XSL•FO
RenderX

Figure 1. Estimated marginal means with 95% CIs for all outcome measures from screening to postintervention in a single-arm pilot trial of a guided web-based stress-management intervention for working adults with attention-deficit/hyperactivity disorder in Sweden. AAQoL: Adult Attention-Deficit/Hyperactivity Disorder Quality of Life; ASRS: Adult Attention-Deficit/Hyperactivity Disorder Self-Report Scale; GAD-7: Generalized Anxiety Disorder 7-item Scale; KEDS: Karolinska Exhaustion Disorder Scale; PHQ-9: Patient Health Questionnaire-9; PSS-10: Perceived Stress Scale-10.

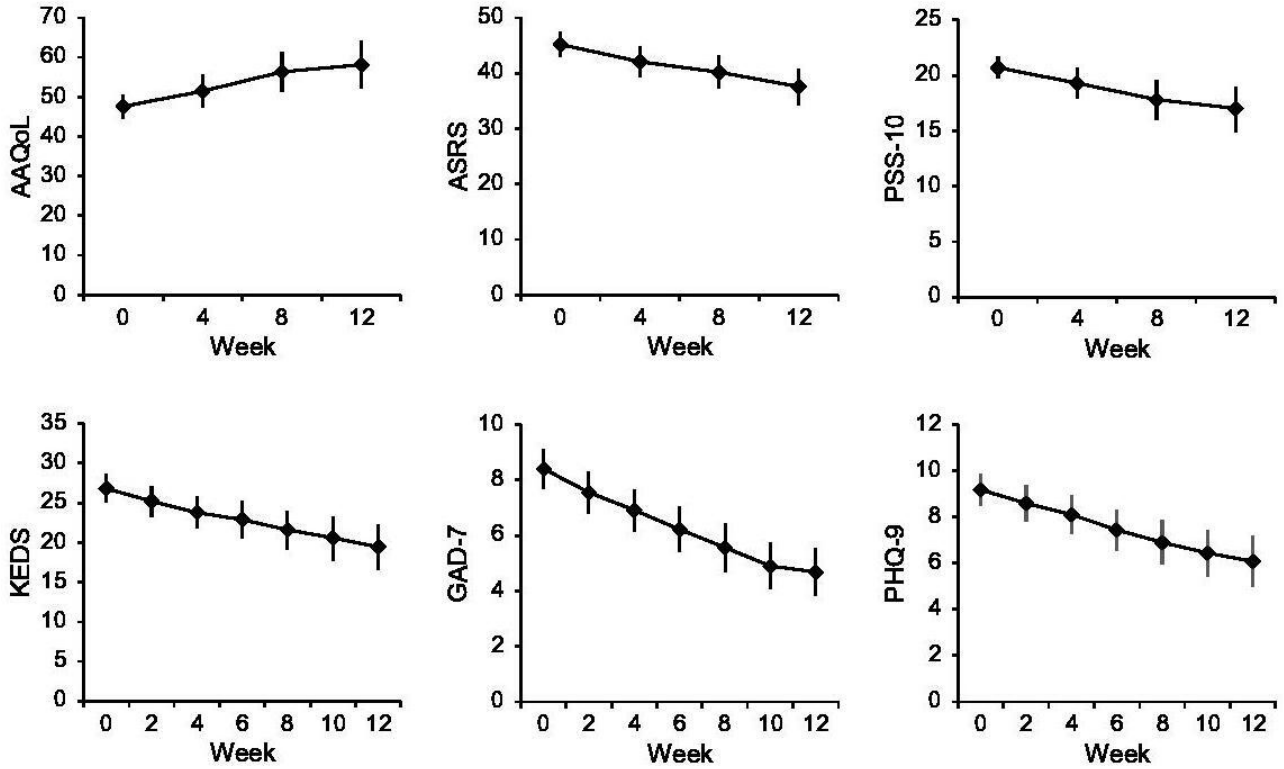

Paired-samples *t* tests indicate statistically significant differences between week 24 follow-up scores and screening scores on all outcome measures, corresponding to large effect sizes Cohen *d*. This indicates an improved average quality of life and lower average symptom levels among participants, 12 weeks postintervention. The results of the paired-samples *t* tests, with corresponding effect sizes Cohen *d*, are shown in Table 5. LOCF means are shown in Table S2 in Multimedia Appendix 1.

Table . Paired-samples *t* test results comparing outcome measures between screening and follow-up in a single-arm pilot trial of a guided web-based stress management intervention for working adults with attention-deficit/hyperactivity disorder in Sweden.

Measure	*t* test (*df*)	*P* value	Cohen *d*
AAQoL[a]	4.32 (35)	<.001	0.83
ASRS[b]	−5.36 (35)	<.001	0.82
PSS-10[c]	−5.89 (35)	<.001	1.12
KEDS[d]	−5.46 (35)	<.001	0.99
PHQ-9[e]	−5.36 (35)	<.001	1.31
GAD-7[f]	−5.32 (35)	<.001	1.13

[a]AAQoL: Adult Attention-Deficit/Hyperactivity Disorder Quality of Life.

[b]ASRS: Adult Attention-Deficit/Hyperactivity Disorder Self-Report Scale.

[c]PSS-10: Perceived Stress Scale-10.

[d]KEDS: Karolinska Exhaustion Disorder Scale.

[e]PHQ-9: Patient Health Questionnaire-9.

[f]GAD-7: Generalized Anxiety Disorder 7-item Scale.

Clinical Significance

On the primary outcome measure AAQoL, 6 of the 36 participants were recovered, 1 nonreliably recovered, 7 improved, 20 unchanged, and 2 deteriorated. On the ASRS, 3 participants were recovered, 1 nonreliably recovered, 10 improved, and 22 unchanged. Individual participant outcomes and Jacobson-Truax classifications are shown in Figures S1 and S2 in Multimedia Appendix 1.

XSL•FO
RenderX

Adherence

There were no explicit dropouts, that is, no participants communicated a premature discontinuation of treatment. However, adherence varied between participants in the case of login frequency, number of modules opened, and number of measurements completed. The mean number of logins was 17.7 (SD 9.28) and the total assessment response rate was 84%. The mean number of modules opened was 9.31 (SD 3.86). All participants but 1 opened the 2 introductory modules, and 29 of the 36 participants opened at least half of the modules. The mean number of messages sent by participants to study staff was 3.4 (SD 2.5).

Analysis of Missing Data

One-way ANOVA showed no statistically significant difference between screening scores of participants who did or did not respond to the measurement at week 12. This indicates that there were no apparent baseline differences between responders and nonresponders.

Negative Effects

Twenty-nine of the 36 participants completed the NEQ-20 at week 12, and 33% (12/36) of the participants reported at least 1 negative effect likely to have been caused by the intervention. The most common negative effects, caused by the intervention or not, were the feeling of more stress (n=10) and the feeling that the treatment did not produce any results (n=5).

Treatment Satisfaction

Twenty-nine of the 36 participants completed the treatment satisfaction questionnaire at week 12. The results are shown in Table 6.

Table . Self-rated treatment satisfaction at postintervention (week 12) among participants in a guided web-based stress management intervention for working adults with attention-deficit/hyperactivity disorder in Sweden.[a]

Items	Values, mean (SD)
1. How do you assess the quality of the treatment you have received?	2.90 (0.67)
2. To what extent have the modules been relevant to you and your areas of concern?	3.03 (0.63)
3. How satisfied are you with the opportunity to ask questions about the content of the modules?	3.48 (0.63)
4. How satisfied are you with the length of the treatment?	3.14 (0.74)
5. Has the treatment helped you to better cope with your problems?	3.00 (0.60)
6. Overall, how satisfied are you with the treatment you have received?	3.03 (0.63)
7. If a friend needed similar help, would you recommend our treatment to him or her?	3.17 (0.60)

[a]1=dissatisfied, 2=indifferent, 3=mostly satisfied, and 4=very satisfied.

Qualitative Results

Twenty-seven of the 36 participants answered the open-ended questions about their experiences of the intervention. Qualitative content analysis yielded 2 main categories and 7 subcategories that describe the participants' experiences. These are shown in Table 7.

XSL•FO
RenderX

Table . Main categories and subcategories from qualitative content analysis of participants' written feedback at postintervention in a pilot trial of a guided web-based stress management intervention for adults with attention-deficit/hyperactivity disorder in Sweden.

Categories and subcategories	Content
Helpful parts of the intervention	
Communication	Good, inviting, effective, appreciated reminders.
	"Very good. Generous and inviting, with personal feedback provided when necessary in the follow-up surveys" (Participant G).
Module content	Good, relevant, concise, easy to understand, some particularly appreciated content.
	"That I could see connections between performance-based self-esteem and stress. Reminded of the importance of recovery, which I apparently had completely skipped" (Participant B).
Treatment platform	Good, easy to navigate, modules (continuous publication, optional modules), login (secure).
	"Great setup regarding how the modules were released and the opportunity to choose what felt relevant" (Participant C).
Desired development of the program	
More structure	Checklists, guidance, templates, check-ins, clearer guidance, one thing at a time, control questions, too little structure toward the end, emphasize that tasks are not followed up, difficult without a deadline.
	"Having treatment over the internet was both good and bad for me. It would have been helpful if what I did was somehow checked off. One of my difficulties is getting things done without a deadline. At the same time, perhaps I wouldn't have had the opportunity to complete all tasks if there were deadlines" (Participant C).
More audiovisual	Layout, more user-friendly, visual materials, audio for tired eyes, being able to listen, images.
	"In a way, it's nice that the graphics are incredibly stripped down and boring, although maybe not so inspiring. Could certainly have benefited from a little more love" (Participant I).
Clinician support	Video calls, personal contact, short digital check-ins, clinician support for structure.
	"A brief conversation that provides drive to continue" (Participant J).
Technical functionality	Easier navigation, writing directly on the platform, filling in directly on the platform, digital workbook, more interactive.
	"More templates and mandatory check-ins. Checklists are very helpful" (Participant K).

Discussion

The purpose of this study was to evaluate the feasibility, acceptability, and effects of a web-based stress management program developed specifically for adults with ADHD. The results showed statistically significant improvements on all outcome measures, from preintervention screening to postintervention and follow-up assessments. The change in mean self-reported quality of life from screening to postintervention corresponds to a large effect, while changes in mean scores on secondary outcome measures correspond to medium-to-large effects. Participants expressed satisfaction with the program content, platform, and communication. More structure, support, audiovisual content, and technical improvements were requested.

The statistically significant increase in self-reported quality of life, corresponding to a large effect, indicates that the participants experienced increased quality of life over the course of the intervention. As previously noted, the cumulative impact of performance impairment, interpersonal challenges, and psychiatric comorbidities significantly diminishes the quality of life of adults with ADHD. The current intervention targets common psychiatric comorbidities among adults with ADHD, with a specific focus on workplace functioning that could affect work performance. Thus, the observed increase in quality of life among the participants, over the course of the intervention, may signify a reversal of this cumulative burden, as supported by concurrent improvements on measures of perceived stress and psychiatric symptoms, corresponding to medium-to-large effects. These findings imply that the intervention's content and delivery are well suited and beneficial for working adults with ADHD experiencing elevated levels of perceived stress, exhaustion, anxiety, and depression.

The observed pre-post effect sizes on both the AAQoL and the secondary outcome measures of psychiatric symptoms were notably larger than within-group effect sizes reported in previous studies. In the MyADHD studies by Nordby et al [64] and Kenter et al [32], improvements on the AAQoL corresponded

XSL•FO
RenderX

to medium effects (*d*=0.62 and *d*=0.60, respectively), compared with *d*=0.84 in this study. Here, it should be noted that the MyADHD intervention was unguided, without clinician support. This could indicate the potential added value of clinician involvement in web-delivered interventions, enhancing their efficacy beyond what is achievable in unguided formats [65]. On PSS measures, Nordby et al [64] reported *d*=0.35 (PSS-14), Kenter et al [32] reported *d*=0.20 (PSS-14), and Moëll et al [34] reported *d*=0.03 (PSS-10), while the means observed by Nasri et al [31] correspond to an effect size of *d*=0.37 (PSS-4), compared with *d*=0.83 (PSS-10) in this study. Similarly, on the GAD-7 and the PHQ-9, Nordby et al [64] reported *d*=0.04 and *d*=0.32, respectively, compared with *d*=1.70 and *d*=1.25 in this study.

Although ADHD symptoms were not a primary target of the current intervention, the ASRS means indicate a statistically significant improvement pre-post, corresponding to large effects on the total scale (*d*=0.98) and the inattention subscale (*d*=1.12) and a medium effect on the hyperactivity subscale (*d*=0.72). These effect sizes are more in-line with within-group effect sizes reported in previous studies. Nordby et al [64] and Kenter et al [32] reported total ASRS effect sizes of *d*=0.93 and *d*=0.80, respectively, while the means reported by Nasri et al [31] correspond to a smaller effect size of *d*=0.68. Moëll et al [34] reported similarly large effects on the inattention (*d*=1.18) and hyperactivity (*d*=0.55) subscales. Again, MyADHD was unguided, and both MyADHD and the web-based course by Moëll et al [34] focused primarily on symptoms of inattention. Considering that clinician support may have played a role in the superior outcomes observed in this study compared with unguided interventions, future research could directly compare the efficacy of interventions with and without support, or investigate the potential of generative artificial intelligence acting as a clinician. This could determine whether such technology could enhance intervention effectiveness through personalized support and real-time engagement, particularly for adults with ADHD [66].

Participants were invited to rate their satisfaction and provide feedback through open-ended questions regarding their experiences with the intervention. Overall, satisfaction scores were positive, particularly regarding the opportunity to inquire about treatment content. Responses to the open-ended questions highlighted participants' positive perceptions of the communication with the research group, describing it as good, inviting, and effective. Ramsay [25] underscores the therapist's role in CBT for adult ADHD, validating frustrations and aiding patients in reengaging with the change process. The incorporation of support on demand in the current intervention aimed to mitigate premature discontinuation of treatment due to frustration among the participants. Encouragingly, none of the participants explicitly withdrew from the study, although adherence varied. However, the mean number of modules opened (78%) exceeds the reported number of completed modules for the unguided intervention of Nordby et al (62%) [64] and Kenter et al (65%) [32], as well as for the guided interventions of Moëll et al (53%) [34], Pettersson et al (61%) [33], and Nasri et al (56%) [31]. The average number of messages sent by the participants to study staff was 3.4 messages. This is in-line [67] or higher [68] than in previous iCBT trials with support on demand. However, postintervention, several participants suggested improved and increased support in future development of the program.

When discussing the results of this study, it is essential to acknowledge the absence of any control group. This means the design did not control for the expectation of benefit, other common factors of digital psychological interventions, or the natural disease progression [69]. Also important to note are the characteristics of the sample, which primarily consisted of highly educated women, with a large majority on psychotropic medication. This may limit the generalizability of the findings to other adult populations with ADHD, particularly men and nonbinary individuals, those with lower levels of education, and those not on psychotropic medication. Predominately female samples are common in iCBT trials, as exemplified by the study by Lindner et al [70]. In Sweden, women also show a small majority among adults with ADHD [71]. However, future trials would benefit from exploring other recruitment strategies, including ones targeting specific populations, to achieve more representative samples. Future trials would also benefit from higher-resolution measures of pharmacological treatment among participants, for example, separating ADHD prescriptions from other psychotropic medication.

Considering the study's strengths, inclusion criteria were generous and did not exclude individuals with autism, learning disabilities, or psychiatric comorbidities other than severe depression or suicidality. While this makes for a sample more representative of the heterogeneous adult ADHD population, the demographic range is still limited. The outcome measures are well established and psychometrically evaluated. In light of recent criticism of the PSS [40], subscale outcomes were reported in the results. The potential issues with the KEDS, recently shown by Lindsäter et al [41], should be mentioned. However, the scale remains the benchmark for web intervention research in Sweden, and it is supplemented by the other measures included in the study. Concerning the measures, inclusion criteria required only an above-cutoff score on one of the secondary outcome measures. This means that not all participants had elevated levels of stress, exhaustion, anxiety, and depression at screening. This could introduce floor effects, as some participants lacked room for improvement on 1 or more measures. This is also why RCI was calculated and presented for only the AAQoL and the ASRS, relevant for all participants.

The intervention was based on previous interventions that have been shown efficacious for ADHD symptoms and common ADHD comorbidities. The intervention content was designed with ADHD theory and mechanisms of change in mind. The development of the intervention was also preceded by interviews with adults with ADHD, regarding workplace functioning and the intervention proposition. From the results of this study, it is clear that the intervention has the potential to be helpful for adults with ADHD who are experiencing stress, exhaustion, anxiety, and depression. Based on this, and the feedback provided by the participants, the intervention will be revised before being further evaluated in an RCT. If proven helpful and effective, the intervention could be developed further and possibly disseminated to a broader population through routine

XSL•FO
RenderX

health care or as a stand-alone product, such as an app, as exemplified in the studies by Selaskowski et al [72] and Seery et al [73].

Acknowledgments

We would like to thank Riksförbundet Attention for help in recruitment to the study. We used ChatGPT solely for language refinement and not for content creation. All intellectual contributions, including study design, data analysis, and interpretation of findings, are our own. We take full responsibility for the accuracy and originality of the manuscript.

Data Availability

Data from this study are not publicly available because they contain information that could compromise research participant privacy. However, data are archived for at least 10 years in compliance with Stockholm University regulations. Requests for any data should be addressed to the corresponding author.

Authors' Contributions

MO and AR initiated the idea for the study. Funding was secured by MO, AR, YG, PC, and FJ, with FJ as the main applicant and principal investigator for the project. The purpose, research aims, and design were further conceptualized by YG, PC, and FJ. GA provided resources and software for recruitment and data collection. Investigation and formal analysis were conducted by MO, KBW, and SH, with support from the remaining authors. The original draft was written by MO, KBW, and SH. Important revisions were suggested by AR, YG, PC, GA, and FJ. The final manuscript was approved by all authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Estimated marginal means, follow-up means, and Jacobson-Truax plots.
[PDF File, 165 KB - formative_v9i1e66388_app1.pdf]

References

1. Song P, Zha M, Yang Q, et al. The prevalence of adult attention-deficit hyperactivity disorder: a global systematic review and meta-analysis. J Glob Health 2021 Feb 11;11:04009. [doi: 10.7189/jogh.11.04009] [Medline: 33692893]
2. Willcutt EG. Theories of ADHD. In: Barkley RA, editor. Attention-Deficit Hyperactivity Disorder: A Handbook for Diagnosis and Treatment: Guilford Press; 2015:391-404.
3. Kooij JJS, Bijlenga D, Salerno L, et al. Updated European Consensus Statement on diagnosis and treatment of adult ADHD. Eur Psychiatry 2019 Feb;56:14-34. [doi: 10.1016/j.eurpsy.2018.11.001] [Medline: 30453134]
4. Kooij SJJ, Bejerot S, Blackwell A, et al. European consensus statement on diagnosis and treatment of adult ADHD: the European Network Adult ADHD. BMC Psychiatry 2010 Sep 3;10(1):67. [doi: 10.1186/1471-244X-10-67] [Medline: 20815868]
5. Kessler RC, Adler L, Barkley R, et al. The prevalence and correlates of adult ADHD in the United States: results from the National Comorbidity Survey Replication. Am J Psychiatry 2006 Apr;163(4):716-723. [doi: 10.1176/ajp.2006.163.4.716] [Medline: 16585449]
6. Polyzoi M, Ahnemark E, Medin E, Ginsberg Y. Estimated prevalence and incidence of diagnosed ADHD and health care utilization in adults in Sweden—a longitudinal population-based register study. Neuropsychiatr Dis Treat 2018;14:1149-1161. [doi: 10.2147/NDT.S155838] [Medline: 29765219]
7. Wilens TE, Biederman J, Faraone SV, Martelon M, Westerberg D, Spencer TJ. Presenting ADHD symptoms, subtypes, and comorbid disorders in clinically referred adults with ADHD. J Clin Psychiatry 2009 Nov;70(11):1557-1562. [doi: 10.4088/JCP.08m04785pur] [Medline: 20031097]
8. Combs MA, Canu WH, Broman-Fulks JJ, Rocheleau CA, Nieman DC. Perceived Stress and ADHD symptoms in adults. J Atten Disord 2015 May;19(5):425-434. [doi: 10.1177/1087054712459558] [Medline: 23034340]
9. Rogers DC, Dittner AJ, Rimes KA, Chalder T. Fatigue in an adult attention deficit hyperactivity disorder population: a trans-diagnostic approach. Br J Clin Psychol 2017 Mar;56(1):33-52. [doi: 10.1111/bjc.12119] [Medline: 27918087]
10. Brattberg G. PTSD and ADHD: underlying factors in many cases of burnout. Stress Health 2006 Dec;22(5):305-313. [doi: 10.1002/smi.1112]
11. Friedrichs B, Igl W, Larsson H, Larsson JO. Coexisting psychiatric problems and stressful life events in adults with symptoms of ADHD--a large Swedish population-based study of twins. J Atten Disord 2012 Jan;16(1):13-22. [doi: 10.1177/1087054710376909] [Medline: 20686099]

XSL•FO
RenderX

12. Andreassen CS, Griffiths MD, Sinha R, Hetland J, Pallesen S. The relationships between workaholism and symptoms of psychiatric disorders: a large-scale cross-sectional study. PLOS ONE 2016;11(5):e0152978. [doi: 10.1371/journal.pone.0152978] [Medline: 27192149]
13. Biederman J, Petty CR, Fried R, et al. Educational and occupational underattainment in adults with attention-deficit/hyperactivity disorder. J Clin Psychiatry 2008 Aug;69(8):1217-1222. [doi: 10.4088/jcp.v69n0803] [Medline: 18681752]
14. Adamou M, Arif M, Asherson P, et al. Occupational issues of adults with ADHD. BMC Psychiatry 2013 Feb 17;13(1):59. [doi: 10.1186/1471-244X-13-59] [Medline: 23414364]
15. Fuermaier ABM, Tucha L, Butzbach M, Weisbrod M, Aschenbrenner S, Tucha O. ADHD at the workplace: ADHD symptoms, diagnostic status, and work-related functioning. J Neural Transm 2021 Jul;128(7):1021-1031. [doi: 10.1007/s00702-021-02309-z]
16. Asherson P, Akehurst R, Kooij JJS, et al. Under diagnosis of adult ADHD: cultural influences and societal burden. J Atten Disord 2012 Jul;16(5 Suppl):20S-38S. [doi: 10.1177/1087054711435360] [Medline: 22377849]
17. Pawaskar M, Fridman M, Grebla R, Madhoo M. Comparison of quality of life, productivity, functioning and self-esteem in adults diagnosed with ADHD and with symptomatic ADHD. J Atten Disord 2020 Jan;24(1):136-144. [doi: 10.1177/1087054719841129] [Medline: 31043096]
18. Kleinman NL, Durkin M, Melkonian A, Markosyan K. Incremental employee health benefit costs, absence days, and turnover among employees with ADHD and among employees with children with ADHD. J Occup Environ Med 2009 Nov;51(11):1247-1255. [doi: 10.1097/JOM.0b013e3181bca68c] [Medline: 19858744]
19. de Graaf R, Kessler RC, Fayyad J, et al. The prevalence and effects of adult attention-deficit/hyperactivity disorder (ADHD) on the performance of workers: results from the WHO World Mental Health Survey Initiative. Occup Environ Med 2008 Dec;65(12):835-842. [doi: 10.1136/oem.2007.038448] [Medline: 18505771]
20. Kessler RC, Lane M, Stang PE, Van Brunt DL. The prevalence and workplace costs of adult attention deficit hyperactivity disorder in a large manufacturing firm. Psychol Med 2009 Jan;39(1):137-147. [doi: 10.1017/S0033291708003309] [Medline: 18423074]
21. Daley D, Jacobsen RH, Lange AM, Sørensen A, Walldorf J. The economic burden of adult attention deficit hyperactivity disorder: a sibling comparison cost analysis. Eur Psychiatry 2019 Sep;61:41-48. [doi: 10.1016/j.eurpsy.2019.06.011] [Medline: 31288209]
22. NICE. Attention deficit hyperactivity disorder: diagnosis and management. 2018. URL: https://www.nice.org.uk/guidance/cg72 [accessed 2025-04-21]
23. Fullen T, Jones SL, Emerson LM, Adamou M. Psychological treatments in adult ADHD: a systematic review. J Psychopathol Behav Assess 2020 Sep;42(3):500-518. [doi: 10.1007/s10862-020-09794-8]
24. Nimmo-Smith V, Merwood A, Hank D, et al. Non-pharmacological interventions for adult ADHD: a systematic review. Psychol Med 2020 Mar;50(4):529-541. [doi: 10.1017/S0033291720000069] [Medline: 32036811]
25. Ramsay JR. CBT for adult ADHD: adaptations and hypothesized mechanisms of change. J Cogn Psychother 2010 Feb;24(1):37-45. [doi: 10.1891/0889-8391.24.1.37]
26. Barkley RA. Behavioral inhibition, sustained attention, and executive functions: constructing a unifying theory of ADHD. Psychol Bull 1997 Jan;121(1):65-94. [doi: 10.1037/0033-2909.121.1.65] [Medline: 9000892]
27. Barkley RA. Etiologies of ADHD. In: Barkley R, editor. Attention-Deficit Hyperactivity Disorder: A Handbook for Diagnosis and Treatment: Guilford Press; 2015:356-390.
28. Andersson G, Carlbring P. Internet interventions in clinical psychology. In: Comprehensive Clinical Psychology: Elsevier; 2022:194-205. [doi: 10.1016/B978-0-12-818697-8.00211-9]
29. Hedman-Lagerlöf E, Carlbring P, Svärdman F, Riper H, Cuijpers P, Andersson G. Therapist-supported Internet-based cognitive behaviour therapy yields similar effects as face-to-face therapy for psychiatric and somatic disorders: an updated systematic review and meta-analysis. World Psychiatry 2023 Jun;22(2):305-314. [doi: 10.1002/wps.21088] [Medline: 37159350]
30. Svärdman F, Sjöwall D, Lindsäter E. Internet-delivered cognitive behavioral interventions to reduce elevated stress: a systematic review and meta-analysis. Internet Interv 2022 Sep;29:100553. [doi: 10.1016/j.invent.2022.100553] [Medline: 35781929]
31. Nasri B, Cassel M, Enhärje J, et al. Internet delivered cognitive behavioral therapy for adults with ADHD —a randomized controlled trial. Internet Interv 2023 Sep;33:100636. [doi: 10.1016/j.invent.2023.100636] [Medline: 37483263]
32. Kenter RMF, Gjestad R, Lundervold AJ, Nordgreen T. A self-guided internet-delivered intervention for adults with ADHD: Results from A randomized controlled trial. Internet Interv 2023 Apr;32:100614. [doi: 10.1016/j.invent.2023.100614] [Medline: 36969389]
33. Pettersson R, Söderström S, Edlund-Söderström K, Nilsson KW. Internet-based cognitive behavioral therapy for adults with ADHD in outpatient psychiatric care. J Atten Disord 2017 Apr;21(6):508-521. [doi: 10.1177/1087054714539998] [Medline: 24970720]

XSL•FO
RenderX

34. Moëll B, Kollberg L, Nasri B, Lindefors N, Kaldo V. Living SMART—a randomized controlled trial of a guided online course teaching adults with ADHD or sub-clinical ADHD to use smartphones to structure their everyday life. Internet Interv 2015 Mar;2(1):24-31. [doi: 10.1016/j.invent.2014.11.004]
35. Oscarsson M, Nelson M, Rozental A, Ginsberg Y, Carlbring P, Jönsson F. Stress and work-related mental illness among working adults with ADHD: a qualitative study. BMC Psychiatry 2022 Nov 30;22(1):751. [doi: 10.1186/s12888-022-04409-w] [Medline: 36451126]
36. Hirvikoski T, Lindström T, Carlsson J, Waaler E, Jokinen J, Bölte S. Psychoeducational groups for adults with ADHD and their significant others (PEGASUS): a pragmatic multicenter and randomized controlled trial. Eur Psychiatry 2017 Jul;44:141-152. [doi: 10.1016/j.eurpsy.2017.04.005] [Medline: 28641216]
37. Lauder K, McDowall A, Tenenbaum HR. A systematic review of interventions to support adults with ADHD at work—implications from the paucity of context-specific research for theory and practice. Front Psychol 2022;13:893469. [doi: 10.3389/fpsyg.2022.893469] [Medline: 36072032]
38. Oscarsson M. Internet-based stress management for working adults with ADHD: preregistration. Open Science Framework. 2023 Feb 2. URL: https://osf.io/u2cdj [accessed 2025-05-22]
39. Vlaescu G, Alasjö A, Miloff A, Carlbring P, Andersson G. Features and functionality of the Iterapi platform for internet-based psychological treatment. Internet Interv 2016 Nov;6:107-114. [doi: 10.1016/j.invent.2016.09.006]
40. Rozental A, Forsström D, Johansson M. A psychometric evaluation of the Swedish translation of the Perceived Stress Scale: a Rasch analysis. BMC Psychiatry 2023 Sep 22;23(1):690. [doi: 10.1186/s12888-023-05162-4] [Medline: 37736719]
41. Lindsäter E, van de Leur JC, Rück C, Hedman-Lagerlöf E, Bianchi R. Psychometric and structural properties of the Karolinska Exhaustion Disorder Scale: a 1,072-patient study. BMC Psychiatry 2023 Sep 2;23(1):642. [doi: 10.1186/s12888-023-05138-4] [Medline: 37660017]
42. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
43. Kroenke K, Spitzer RL, Williams JBW. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613. [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
44. Brod M, Johnston J, Able S, Swindle R. Validation of the Adult Attention-Deficit/Hyperactivity Disorder Quality-of-Life Scale (AAQoL): a disease-specific quality-of-life measure. Qual Life Res 2006 Feb;15(1):117-129. [doi: 10.1007/s11136-005-8325-z] [Medline: 16411036]
45. Kessler RC, Adler L, Ames M, et al. The World Health Organization Adult ADHD Self-Report Scale (ASRS): a short screening scale for use in the general population. Psychol Med 2005 Feb;35(2):245-256. [doi: 10.1017/s0033291704002892] [Medline: 15841682]
46. Rozental A, Kottorp A, Forsström D, et al. The Negative Effects Questionnaire: psychometric properties of an instrument for assessing negative effects in psychological treatments. Behav Cogn Psychother 2019 Sep;47(5):559-572. [doi: 10.1017/S1352465819000018] [Medline: 30871650]
47. Safren SA, Sprich S, Mimiaga MJ, et al. Cognitive behavioral therapy vs relaxation with educational support for medication-treated adults with ADHD and persistent symptoms: a randomized controlled trial. JAMA 2010 Aug 25;304(8):875-880. [doi: 10.1001/jama.2010.1192] [Medline: 20736471]
48. Rozental A, Forsell E, Svensson A, Andersson G, Carlbring P. Internet-based cognitive-behavior therapy for procrastination: a randomized controlled trial. J Consult Clin Psychol 2015 Aug;83(4):808-824. [doi: 10.1037/ccp0000023] [Medline: 25939016]
49. Almén N. A cognitive behavioral model proposing that clinical burnout may maintain itself. Int J Environ Res Public Health 2021 Mar 26;18(7):3446. [doi: 10.3390/ijerph18073446] [Medline: 33810358]
50. Jernelöv S, Larsson Y, Llenas M, Nasri B, Kaldo V. Effects and clinical feasibility of a behavioral treatment for sleep problems in adult attention deficit hyperactivity disorder (ADHD): a pragmatic within-group pilot evaluation. BMC Psychiatry 2019 Jul 24;19(1):226. [doi: 10.1186/s12888-019-2216-2] [Medline: 31340804]
51. Hirsch O, Chavanon M, Riechmann E, Christiansen H. Emotional dysregulation is a primary symptom in adult attention-deficit/hyperactivity disorder (ADHD). J Affect Disord 2018 May;232:41-47. [doi: 10.1016/j.jad.2018.02.007] [Medline: 29477097]
52. Barkley RA, Fischer M. The unique contribution of emotional impulsiveness to impairment in major life activities in hyperactive children as adults. J Am Acad Child Adolesc Psychiatry 2010 May;49(5):503-513. [doi: 10.1097/00004583-201005000-00011] [Medline: 20431470]
53. Hagberg T, Manhem P, Oscarsson M, Michel F, Andersson G, Carlbring P. Efficacy of transdiagnostic cognitive-behavioral therapy for assertiveness: a randomized controlled trial. Internet Interv 2023 Apr;32:100629. [doi: 10.1016/j.invent.2023.100629] [Medline: 37273933]
54. Rozental A, Shafran R, Wade T, et al. A randomized controlled trial of internet-based cognitive behavior therapy for perfectionism including an investigation of outcome predictors. Behav Res Ther 2017 Aug;95:79-86. [doi: 10.1016/j.brat.2017.05.015] [Medline: 28586713]
55. Diedrich A, Grant M, Hofmann SG, Hiller W, Berking M. Self-compassion as an emotion regulation strategy in major depressive disorder. Behav Res Ther 2014 Jul;58:43-51. [doi: 10.1016/j.brat.2014.05.006] [Medline: 24929927]

XSL•FO
RenderX

56. Berg I, Hovne V, Carlbring P, et al. “Good job!”: Therapists’ encouragement, affirmation, and personal address in internet-based cognitive behavior therapy for adolescents with depression. Internet Interv 2022 Dec;30:100592. [doi: 10.1016/j.invent.2022.100592] [Medline: 36439193]
57. Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. J Consult Clin Psychol 1991 Feb;59(1):12-19. [doi: 10.1037//0022-006x.59.1.12] [Medline: 2002127]
58. R core team. R: A language and environment for statistical computing. 2022. URL: https://www.R-project.org [accessed 2025-04-21]
59. Jamovi. The jamovi project. 2023. URL: https://jamovi.org [accessed 2025-04-21]
60. Gallucci M. GAMLj: general analyses for linear models. 2024. URL: https://gamlj.github.io [accessed 2025-04-21]
61. Kruijt AW. Obtain and plot Jacobson-Truax and reliable change indices. 2019. URL: https://rdrr.io/github/AWKruijt/JT-RCI/man/JTRCI.html [accessed 2025-05-23]
62. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005 Nov;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
63. Krippendorff K. Content Analysis: An Introduction to Its Methodology: SAGE Publications, Inc; 2019. [doi: 10.4135/9781071878781]
64. Nordby ES, Kenter RMF, Lundervold AJ, Nordgreen T. A self-guided Internet-delivered intervention for adults with ADHD: a feasibility study. Internet Interv 2021 Sep;25:100416. [doi: 10.1016/j.invent.2021.100416] [Medline: 34401375]
65. Baumeister H, Reichler L, Munzinger M, Lin J. The impact of guidance on Internet-based mental health interventions — A systematic review. Internet Interv 2014 Oct;1(4):205-215. [doi: 10.1016/j.invent.2014.08.003]
66. Carlbring P, Hadjistavropoulos H, Kleiboer A, Andersson G. A new era in Internet interventions: the advent of Chat-GPT and AI-assisted therapist guidance. Internet Interv 2023 Apr;32:100621. [doi: 10.1016/j.invent.2023.100621] [Medline: 37273936]
67. Rheker J, Andersson G, Weise C. The role of “on demand” therapist guidance vs. no support in the treatment of tinnitus via the internet: a randomized controlled trial. Internet Interv 2015 May;2(2):189-199. [doi: 10.1016/j.invent.2015.03.007]
68. Käll A, Olsson Lynch C, Sundling K, Furmark T, Carlbring P, Andersson G. Scheduled support versus support on demand in internet-delivered cognitive behavioral therapy for social anxiety disorder: randomized controlled trial. Clin Psychol Eur 2023 Sep;5(3):e11379. [doi: 10.32872/cpe.11379] [Medline: 38356895]
69. Lutz J, Offidani E, Taraboanta L, Lakhan SE, Campellone TR. Appropriate controls for digital therapeutic clinical trials: a narrative review of control conditions in clinical trials of digital therapeutics (DTx) deploying psychosocial, cognitive, or behavioral content. Front Digit Health 2022;4:823977. [doi: 10.3389/fdgth.2022.823977] [Medline: 36060538]
70. Lindner P, Nyström MBT, Hassmén P, Andersson G, Carlbring P. Who seeks ICBT for depression and how do they get there? Effects of recruitment source on patient demographics and clinical characteristics. Internet Interv 2015 May;2(2):221-225. [doi: 10.1016/j.invent.2015.04.002]
71. Diagnostik och läkemedelsbehandling vid ADHD [Diagnosis and pharmacological treatment of ADHD]. Socialstyrelsen. 2023. URL: https://www.socialstyrelsen.se/globalassets/sharepoint-dokument/artikelkatalog/ovrigt/2023-11-8862.pdf [accessed 2025-05-22]
72. Selaskowski B, Reiland M, Schulze M, et al. Chatbot-supported psychoeducation in adult attention-deficit hyperactivity disorder: randomised controlled trial. BJPsych Open 2023 Oct 13;9(6):e192. [doi: 10.1192/bjo.2023.573] [Medline: 37827996]
73. Seery C, Cochrane RH, Mulcahy M, Kilbride K, Wrigley M, Bramham J. “A one-stop shop”: real-world use and app-users’ experiences of a psychoeducational smartphone app for adults with ADHD. Internet Interv 2025 Mar;39:100807. [doi: 10.1016/j.invent.2025.100807] [Medline: 39996148]

Abbreviations

AAQoL: Adult Attention-Deficit/Hyperactivity Disorder Quality of Life Scale
ADHD: attention-deficit/hyperactivity disorder
ASRS: Adult Attention-Deficit/Hyperactivity Disorder Self-Report Scale
CBT: cognitive behavioral therapy
GAD-7: Generalized Anxiety Disorder 7-item Scale
iCBT: internet-delivered cognitive behavioral therapy
KEDS: Karolinska Exhaustion Disorder Scale
LOCF: last observation carried forward
NEQ: Negative Effects Questionnaire
PHQ-9: Patient Health Questionnaire
PSS: Perceived Stress Scale
RCI: reliable change index
RCT: randomized controlled trial

XSL•FO
RenderX

Edited by A Mavragani; submitted 11.09.24; peer-reviewed by L Sorensen, V Nimmo-Smith; revised version received 23.03.25; accepted 02.04.25; published 29.05.25.

Please cite as:
Oscarsson M, Hammarbäck S, Blom Wiberg K, Rozental A, Ginsberg Y, Carlbring P, Andersson G, Jönsson F
Web-Based Stress Management for Working Adults With Attention-Deficit/Hyperactivity Disorder (ADHD): Single-Arm, Open Pilot Trial
JMIR Form Res 2025;9:e66388
URL: https://formative.jmir.org/2025/1/e66388
doi:10.2196/66388

© Martin Oscarsson, Sandra Hammarbäck, Karolina Blom Wiberg, Alexander Rozental, Ylva Ginsberg, Per Carlbring, Gerhard Andersson, Fredrik Jönsson. Originally published in JMIR Formative Research (https://formative.jmir.org), 29.5.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Methods of Piloting an Abstraction Tool to Describe Family Engagement in the Hospital Setting: Retrospective Chart Review

Jennifer Morgan[1,2], PhD, RN-BC; Jennifer Cahill[3], PhD, RN; Christine Ritchie[4,5], MD, MSPH; Lingling Zhang[1], ScD; Priscilla Gazarian[6], PhD, CNS, RN

[1]Manning College of Nursing and Health Sciences, University of Massachusetts Boston, Boston, MA, United States
[2]Department of Nursing, Massachusetts General Hospital, 55 Fruit Street, Boston, MA, United States
[3]Yvonne L. Munn Center for Nursing Research, Massachusetts General Hospital, Boston, MA, United States
[4]Division of Palliative Care and Geriatric Medicine, Harvard Medical School, Boston, MA, United States
[5]Mongan Institute Center for Aging and Serious Illness, Massachusetts General Hospital, Boston, MA, United States
[6]Tan Chingfen Graduate School of Nursing, University of Massachusetts Chan Medical School, Worcester, MA, United States

Corresponding Author:
Jennifer Morgan, PhD, RN-BC
Manning College of Nursing and Health Sciences, University of Massachusetts Boston, Boston, MA, United States

Abstract

Background: Family engagement in hospitals is crucial for improving outcomes and ensuring holistic, patient-centered care. However, there is limited understanding of how providers document family engagement in electronic medical records (EMR) and how factors such as race and health disparities influence engagement practices. The absence of standardized EMR templates complicates tracking engagement and assessing its impact on patient outcomes. Retrospective chart review (RCR) is an effective method for investigating clinical practice and how family engagement is documented, using both structured and unstructured data from patient records. Despite its potential, gaps remain in the literature regarding distinctions between the prepilot and pilot phases in RCR studies.

Objective: This article describes the prepiloting and piloting stages in the development of an abstraction tool for an RCR study, highlighting how these phases refined the tool for extracting family engagement data from the EMRs.

Methods: A cohort of 2032 medical records was selected using the Research Patient Database Registry and EMRs. Initially, a draft tool was tested during the prepilot phase to assess its stability. To optimize diversity, the sample was then stratified by race. The modified tool was subsequently piloted on a subset of the sample.

Results: The prepilot phase tested the tool on 9 records. In the pilot phase, the tool was applied to 39 records, representing approximately 10% of the sample. After the prepiloting and piloting phases, 293 of the 405 patient records were deemed eligible for inclusion. More than three-quarters of patients had documentation of presence and communication; whereas, only about one-third had documentation of shared decision-making involving families.

Conclusions: The prepilot phase helped standardize the abstraction tool, align it with the EMRs, and address potential biases. The pilot phase provided insights into data availability and highlighted areas for refinement before finalizing the tool for the remaining records. Together, these phases ensured the tool's effectiveness for use in large-scale RCR studies.

(JMIR Form Res 2025;9:e66549) doi:10.2196/66549

KEYWORDS

retrospective chart review; medical record review; patient family centered care; care transitions; family-centered care; medical record; family engagement; patient care; abstraction tool; hospital setting; electronic medical record; EMR; data extraction; decision-making

Introduction

Family engagement refers to the active involvement of family members in patient care and decision-making processes. This includes their participation in specific aspects or domains of engagement, defined as presence, communication, identification and acknowledgment of family needs, shared decision-making, and contribution to care [1,2]. Family engagement in the care of seriously ill patients is more important than ever in the postpandemic era, especially given concerns about capacity and labor force shortages in acute care settings [3]. Family, as defined by the patient, may be a formal, legal, or informal relationship. Family engagement improves patient outcomes, enhances the patient experience, and ensures that care remains holistic and centered on the patient and their support network [4-7]. Without engagement, hospital stays may be prolonged

XSL•FO
RenderX

[8,9], and readmission rates and overall health care utilization tend to increase [10,11]. At the end of a patient's life, family engagement facilitates healthy bereavement for families and increases job satisfaction among providers [12].

The literature lacks clear insight into how family engagement occurs in hospital settings, which specific domains are experienced, the roles families play during inpatient care, and how these roles are documented in electronic medical records (EMR) [13-17]. Additionally, there is a gap in understanding of how race, ethnicity, and health disparities influence clinicians' initiation of family engagement [15-17]. The absence of a standardized EMR template further complicates the documentation of family engagement, limiting our understanding of how it affects patient outcomes in both acute care and the transition from acute care [17].

Retrospective chart review (RCR) is an effective methodology for investigating clinical practices, such as the current standard for documenting family engagement in medical records. As a foundational approach in clinical research, RCR involves extracting both structured data (eg, demographic data, age, or length of stay) and unstructured data (eg, nursing notes) from patient records [18-21]. This method provides insights into how clinicians initiate and sustain family engagement, as seen in unstructured data such as progress notes. As RCR research relies on both rigor and replication, careful preplanning, adherence to defined steps, and development of a well-structured abstraction tool is critical [19-23]. The abstraction tool standardizes data collection, reduces bias, and ensures consistency, making replication by other researchers feasible. However, there is a gap in the existing literature in making a distinction between the prepilot and pilot phases of the abstraction tool development. The abstraction tool in this study guided an RCR to assess current documentation standards of family engagement in hospitalized patients at a single academic medical center, while exploring potential variations in documentation based on race or ethnicity, social determinants of health, and patient characteristics, with the study published separately [17]. This paper describes the development of an abstraction tool through two distinct phases—prepiloting and piloting—and how each phase contributed to refining the tool for extracting data on family engagement from the EMRs.

Methods

Design, Setting, and Participants

The RCR extracted clinical and sociodemographic data from the EMRs of patients aged 18 years and older who were admitted to one of three neuroscience units at an academic medical center over six months in 2023. These units were selected as their health care teams had received training for serious illness conversations, which encompass communication, presence, needs assessment, and shared decision-making [24]. Exclusion criteria included hospital employees; patients admitted for fewer than three days (as unstructured notes often omit mention of family); those with a history of familial abuse or neglect (due to safety-related restrictions on family engagement); and patients who died during admission, as family engagement at the end of life, while vital for dignity, does not contribute to safe transitions between care settings [12].

Data Collection

Following institutional review board approval and a waiver of consent, a deidentified and encrypted list of eligible patients was extracted from the Research Patient Database Registry, a data warehouse containing EMR data from hospitals within the academic medical system. The database identified 2032 eligible patients. To optimize a diverse sample, the list was stratified by race into three groups [25]. The final convenience sample (N=405) included 135 Black, 135 White, and 135 Asian or Other patients [26]. Of these, 293 patients met inclusion criteria for data abstraction, exceeding the recommended sample size of 192, which calculated using G*Power (version 3.1.9.6; Heinrich Heine Universität Düsseldorf) with 80% power and a medium effect size to detect differences among the stratified racial groups [27].

A literature review of family engagement revealed significant gaps in existing research. It also helped identify and operationalize key variables of interest (ie,family engagement domains; Table 1). Understanding the structure of the medical record guided the identification of proxy measures for each domain. A draft abstraction tool was developed incorporating both structured and unstructured data using the Research Electronic Data Capture (REDCap, version 14.0.27), a secure web-based data management platform [28]. Nurse and physician notes from admission through two days post-admission, as well as two days prior to discharge through the day of discharge, were reviewed, as these are critical transition points for both patients and families [16,29]. To improve reliability and validity, observations and decisions were noted in an operating manual for both phases of the study [20]. Additionally, using intra-rater reliability, an accepted substitute for inter-rater reliability [22], the abstractor performed two passes through the records, followed by a third pass through any abstraction that had less than 95% agreement [17].

XSL•FO
RenderX

Table . Table identifying proxy measures of family engagement domains, where they can be found in the supporting literature. The table connects the proxy measures to specific family engagement domains.

Proxy measures	Family engagement domains with definition				
	Presence – Families physically bedside as attendants [1,30]	Family Needs – Family's distinct needs that must be met to form a partnership [1,30]	Communication – Bi-directional sharing of information between family and provider [1,30]	Shared-decision making – Decision-making related to choices about care [1,30]	Contribution to care – Tangible care (IADLs[a], ADLs[b]) and intangible care (emotional and moral support) [1,30]
Nursing note [31-33]	✓	✓	✓	✓	✓
Admission MD[c] note [32]	✓	✓	✓	✓	✓
SW[d] consult placed [32]		✓			
CM[e] note [34]	✓	✓	✓	✓	✓
EMR[f]-family education [32,35]	✓	✓	✓		✓
EMR-psychosocial family assessment [33,35]		✓	✓		✓
EMR-family coping [32,33]	✓	✓	✓		✓
EMR-serious illness conversation [33,35]	✓		✓	✓	
Interpreter services [32]		✓	✓	✓	✓
EMR-health care agent [30,32]		✓	✓	✓	
EMR-designated caregiver [30-32]	✓	✓	✓	✓	✓

[a]IADL: Instrumental activities of daily living.

[b]ADL: Activities of daily living.

[c]MD: Medical doctor.

[d]SW: Social work.

[e]CM: Case management.

[f]EMR: Electronic medical record.

Data Analysis

Data were analyzed using SAS software (version 9.4; SAS Institute). Structured and unstructured data were reviewed for the presence or absence of family engagement domains, rather than their magnitude. Consult notes (eg, physical therapy, occupational therapy, case management, social work, spiritual care, palliative care, hospice) were reviewed if available, particularly focusing on admission and discharge. Descriptive statistics were calculated for continuous variables and counts and percentages for categorical variables [17]. The χ^2 tests were used to assess differences between groups.

Ethical Considerations

This RCR required institutional review board approval as it involved the analysis of medical records without direct patient interaction and a waiver of consent was granted (Protocol # 2023P003145). The data was deidentified to ensure privacy. Additionally, a patient with an unusually long length of stay was excluded from the study to protect privacy. Confidentiality was maintained throughout the study by securing data behind institutional firewalls, and access to data was limited to study personnel.

Results

Figure 1 illustrates the flow of the RCR through the prepilot and pilot phase. These two phases clarified the types of data necessary to capture evidence of family engagement, identified where the data were located within the EMR, and showed how family engagement domains were documented. They also highlighted how to address inconsistent data, interpret missing documentation, and finalize the tool for use in the main study. Furthermore, the prepilot and pilot phases uncovered potential threats to the reliability of the tool and offered the opportunity to address them before the larger study.

XSL•FO
RenderX

Figure 1. Overview of the early stages of the RCR, illustrating the progression from literature review to pre-pilot and pilot phases, and the interaction with the abstraction tool and sample populations. RCR: retrospective chart review; RPDR: Research Patient Data Registry.

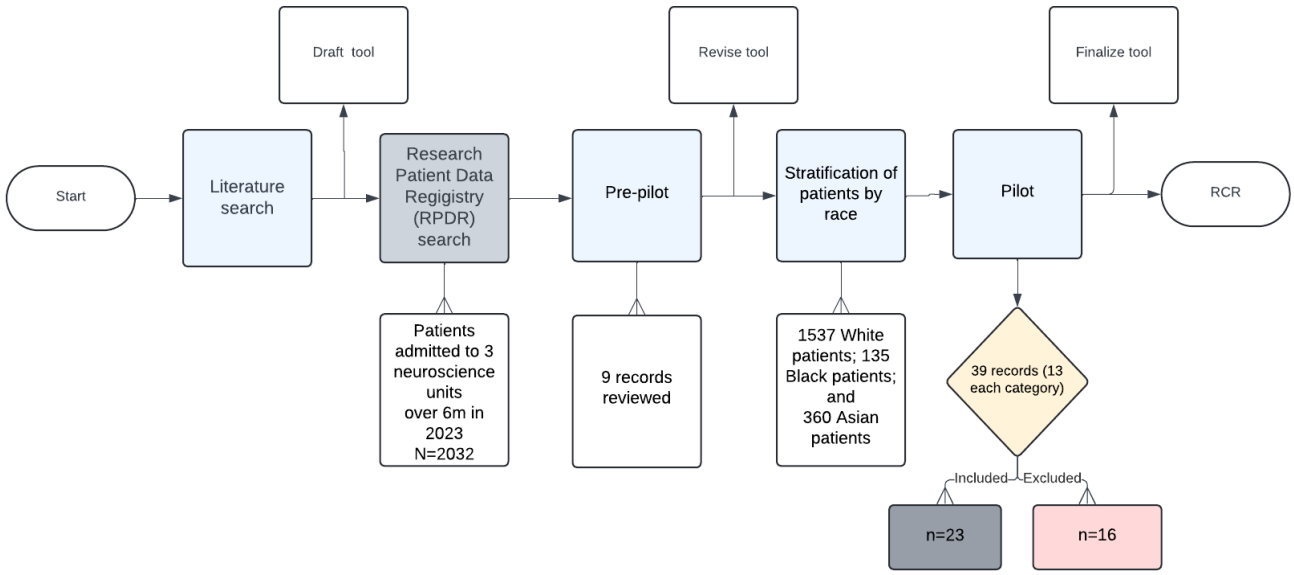

The RCR included structured and unstructured data from 293 patient records and identified which family engagement domains were captured in the EMR, who documented them, and where they appeared in the EMR [17]. The intra-rater reliability for abstraction ranged from 90% (n=1) to 100% (n=170) [17]. Abstraction revealed that the most frequently documented domains were presence (223/293) and communication (227/293), while the least documented domain was shared-decision making (87/293). Most of the family engagement documentation occurred in unstructured data (eg, nursing notes and physical or occupational therapists). There were no perceived differences in documentation in terms of race, social deprivation, or cognitive impairment [17]. However, there were differences in the domain documentation related to discharge placement and provider discipline [17]. These findings were directly related to the prepilot and pilot phase and the development of the abstraction tool.

Prepilot Phase

Standard guidelines suggest that the abstraction tool be tested on three to five records to align with the EMR workflow and to address potential reviewer biases and assumptions. During the prepilot, the draft abstraction tool was initially applied to three randomly selected records, where coding errors—such as issues with branching logic—were identified and corrected. Additionally, organizational and formatting changes were made to better align the tool with the EMR structure. The modified tool was then tested on four more records, which identified consult notes that did or did not contain family engagement documentation. Further modifications were made, and the tool remained stable when applied to two additional records. In total, 9 of the aggregate 2032 records were reviewed in the prepilot phase.

Pilot Phase

For the pilot phase, standard recommendations suggest testing the abstraction tool on 10% of the final sample [28,36,37]. The tool was applied to 39 patient records from the sample population: 13 Black, 13 White, and 13 Asian or Other patients. Six Black and 4 White patients were excluded for being admitted for fewer than 3 days and 6 patients from the Asian or Other group were excluded due to short stays, death during admission, or potential risk of identification due to length of stay. A total of 23 out of 39 patients (59%) were included for pilot phase analysis. There were no significant differences between included and excluded participants when stratified by race. A test of homogeneity indicated that the only distinguishing variable between the groups was the admitting service.

Discussion

Primary findings

Using the refined abstraction tool, the RCR identifies which domains of family engagement are documented in the EMR, who records the information, and where it is located [17]. Development of the tool occurs in two distinct but complementary phases—prepiloting and piloting—each playing a critical role in refining the data abstraction process and enhancing the tool's accuracy and usability.

Prepilot testing involved an initial evaluation to identify issues such as unclear instructions or data elements, with a small group providing feedback for adjustments [22]. Following this, pilot testing with a larger, more diverse sample assessed the tool's consistency, accuracy, and real-world feasibility, ensuring its readiness for full-scale data collection. These two distinct phases were essential for ensuring the abstraction tool's reliability and effectiveness in extracting relevant data.

Prepilot Phase

Improve Efficiency

The prepilot phase highlighted opportunities to streamline the abstraction process. Initially, the tool lacked an intake form to assess eligibility and additional items to guide the abstraction of proxy measures. The first revision integrated the intake form using branching logic to capture all reviewed records in one form while being able to distinguish patients who met the inclusion criteria from those who ultimately did not.

XSL•FO
RenderX

Additionally, the order of the items appearing within the tool was changed to more closely reflect the EMR workflow and maximize efficiency (eg, age was moved to the first item as it appears before other demographic items). Further modifications addressed gaps in instances where engagement domains were documented and by whom, including the omission of physical and occupational therapist notes, which consistently identified the patient's social support. A search term was identified to improve abstraction efficiency.

Identify Assumptions

The prepilot phase also challenged several assumptions in the draft tool. For example, it was initially assumed that the emergency room physician and nurse notes would be the primary sources of family engagement data. However, admitting service history and physical examination notes as well as the nurses' admission notes were found to provide more relevant documentation of family engagement.

Pilot Phase

Identify Potential Problems

The pilot phase raised concerns about excluding patients with stays shorter than 3 days. This exclusion criterion removed more patients than anticipated, particularly those admitted for 2 days. A posteriori, this criterion was re-evaluated. Allowing 2-day stays would have increased inclusion from 23 of 39 patients (59%) to 31 of 39 patients (79%). However, the rationale for excluding these patients—namely, the possibility that abbreviated admissions may not allow insufficient time to initiate family engagement in the hospital—remained compelling, and the exclusion criterion was retained. Additionally, "extended admission" was added as a new exclusion criterion due to confidentiality concerns.

Identify Missing Variables or Indicators

The pilot phase also identified missing variables that could inform data analysis and interpretation. For example, initially, the tool relied upon cognitive impairment being added to the patient's list of problems. As cognitive impairment is underdiagnosed and under-recognized in acute care settings, [38], broader criteria were adopted to identify patients who were experiencing cognitive impairment. The Confusion Assessment Method (CAM), a standardized tool for assessing delirium was added as a proxy measure of cognitive impairment, along with the use of direct observers and restraints. Being identified as cognitively impaired, "delirious" or Confusion Assessment Method-positive could influence family engagement, as these patients would likely need additional support from their families and be unable to initiate engagement themselves, and therefore need providers to proactively engage with families. The tool was revised to include whether a patient was delirious and whether sitters or restraints were used for safety. Additionally, documentation of family meetings for critical patients in one of the neuroscience units was included in the tool.

Identify Data That Need Clarification

Finally, the pilot study uncovered items that required clarification [20,28]. One purpose of the pilot was to ensure consistency of abstraction and coding [37]. In unstructured data such as nursing narrative notes, "supportive" was used to describe some family members but not others. Upon reflection and team consensus, terminologies such as "supportive family bedside" were considered documentation evidence for two domains, including presence and contributing to care. Similarly, some providers documented the existence of family support even when the family was not physically present. However, modification to the tool was made to acknowledge social support, allowing for the possibility that the support was not physically present at the patient's bedside.

Preplanning

Conducting prepilot and pilot phases allowed for preplanning. The two distinct phases helped refine the study timeline [39]. It was found that abstracting each record in the pilot phase took approximately 15 to 20 minutes. Knowing this enabled the team to estimate the data collection timeframe based on the known sample size and allocate resources accordingly.

Limitations and Strengths

Key strengths of RCR methodology include its flexibility, cost-effectiveness, ability to access unique information, reduction of recall bias, and potential for creating randomizable sample frames [19]. Beyond epidemiological and clinical research, RCR is useful for evaluating care patterns and informing training [19,20,28]. However, a notable limitation was that for an RCR, data are obtained twice from the patient, family, or clinical scenarios; first, the clinician gathers the data and second, they document the data. Finally, while every effort was made to optimize sample diversity, stratifying by race—a social construct—can introduce bias if patients did not self-identify their race. These characteristics should be kept in mind during the prepilot and pilot phases of the abstraction tool development.

Conclusion

The prepilot and pilot phases were essential for refining the abstraction tool used in this RCR. These phases allowed the tool to be optimized for efficiency and clarity, addressing potential problems, and improving data collection. The prepilot phase involved an initial evaluation to identify issues such as unclear instructions or missing data elements, with a small group providing feedback for adjustments. On the contrary, the pilot phase involved a more extensive and structured evaluation, with a larger sample size and more diverse set of users. The final abstraction tool integrated both structured and unstructured data, capturing family engagement across multiple disciplines, including nursing, case management, physicians, and ancillary consultants. The focus remained on critical transition points in care (ie, admission and discharge). The two distinct phases ensured that the tool was clear, efficient, and comprehensive, setting the stage for the larger study.

This study addresses a previously unexamined aspect of RCR. Describing the distinction between the prepilot and pilot phases increases transparency and increases the rigor and reproducibility of the methodology. The description of the iterative testing within each of the two phases, indicating that it was only adopted after careful systematic evaluation, also highlighted the learnings and rationale underlying specific

XSL•FO
RenderX

design decisions. This step-wise approach to RCR helps to build best practice standards and inform policy while documenting family engagement.

Acknowledgments

This research was supported by The Yvonne L. Munn Predoctoral Research Grant for PhD Nursing Students, Massachusetts General Hospital, Boston, MA and The Anne Kibrick Research Grant, Sigma Theta Tau International Honor Society of Nursing, Theta Alpha Chapter, University of Massachusetts, Boston, MA.

The content is solely the authors' responsibility and does not reflect the position or policy of the Massachusetts General Hospital or University of Massachusetts Boston.

Authors' Contributions

Conceptualization: JDM
Supervision: PG
Writing – original draft: JDM
Writing – review & editing: CR, JC, LJ, PG

Conflicts of Interest

None declared.

References

1. Goldfarb M. Family engagement in critical care cardiology: a guide for clinicians. Can J Cardiol 2023 Apr;39(4):562-565. [doi: 10.1016/j.cjca.2022.10.001] [Medline: 36216036]
2. Johnson B, Abraham M, Conway J, et al. Partnering with patients and families to design a patient- and family-centered health care system: Institute for Patient and Family-Centered Care; 2008. URL: https://betsylehmancenterma.gov/assets/uploads/PartneringwithPatientsandFamilies.pdf [accessed 2023-01-29]
3. Christantiello R. Amid staff shortages and capacity issues at hospitals, state paves way for quick discharges. The Boston Globe. 2024 Jan 17. URL: https://www.boston.com/news/health/2024/01/17/amid-staff-shortages-and-capacity-issues-at-hospitals-state-paves-way-for-quick-discharges [accessed 2024-05-14]
4. Engaging patients and families in their health care. Agency for Healthcare Research and Quality [AHRQ]. 2022. URL: https://www.ahrq.gov/patient-safety/patients-families/index.html [accessed 2023-03-15]
5. Person and family engagement. Center for Medicare and Medicaid Services. 2021 Dec 1. URL: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/QualityInitiativesGenInfo/Person-and-Family-Engagement [accessed 2023-03-15]
6. Leong MQ, Lim CW, Lai YF. Comparison of hospital-at-home models: a systematic review of reviews. BMJ Open 2021 Jan 29;11(1):e043285. [doi: 10.1136/bmjopen-2020-043285] [Medline: 33514582]
7. Rossinot H, Marquestaut O, de Stampa M. The experience of patients and family caregivers during hospital-at-home in France. BMC Health Serv Res 2019 Jul 9;19(1):470. [doi: 10.1186/s12913-019-4295-7] [Medline: 31288804]
8. Azad TD. Opinion & special articles: shared decision-making during the COVID-19 pandemic: three bullets in 3 hemispheres. Neurology (ECronicon) 2021 May 18;96(20):e2558-e2560. [doi: 10.1212/WNL.0000000000011811] [Medline: 33692167]
9. Segar NO. As hospitals restrict visitors, what constitutes a “good death”? Health Aff (Millwood) 2022 Jun;41(6):921-924. [doi: 10.1377/hlthaff.2021.01857] [Medline: 35666976]
10. Rodakowski J, Rocco PB, Ortiz M, et al. Caregiver integration during discharge planning for older adults to reduce resource use: a metaanalysis. J Am Geriatr Soc 2017 Aug;65(8):1748-1755. [doi: 10.1111/jgs.14873] [Medline: 28369687]
11. Herrin J, Harris KG, Kenward K, Hines S, Joshi MS, Frosch DL. Patient and family engagement: a survey of US hospital practices. BMJ Qual Saf 2016 Mar;25(3):182-189. [doi: 10.1136/bmjqs-2015-004006] [Medline: 26082560]
12. Morgan J, Gazarian P. A good death: a synthesis review of concept analyses studies. Collegian 2023 Apr;30(2):236-246. [doi: 10.1016/j.colegn.2022.08.006]
13. Morgan JD, Gazarian P, Hayman LL. An integrated review: connecting COVID-era hospital visiting policies to family engagement. Front Public Health 2023;11:1249013. [doi: 10.3389/fpubh.2023.1249013] [Medline: 37719727]
14. Burns KEA, Misak C, Herridge M, Meade MO, Oczkowski S, Patient and Family Partnership Committee of the Canadian Critical Care Trials Group. Patient and family engagement in the ICU. Untapped opportunities and underrecognized challenges. Am J Respir Crit Care Med 2018 Aug 1;198(3):310-319. [doi: 10.1164/rccm.201710-2032CI] [Medline: 29624408]
15. Cené CW, Johnson BH, Wells N, Baker B, Davis R, Turchi R. A narrative review of patient and family engagement. Med Care 2016;54(7):697-705. [doi: 10.1097/MLR.0000000000000548]

XSL•FO
RenderX

16. Gitlin LN, Wolff J. Family involvement in care transitions of older adults: What do we know and where do we go from here? Annu Rev Gerontol Geriatr 2011 Nov 1;31(1):31-64. [doi: 10.1891/0198-8794.31.31]
17. Morgan J, Cahill J, Ritchie C, Zhang L, Gazarian P. Family engagement on neuroscience units with post-COVID visiting policies: a retrospective chart review. Int J Nurs Knowl 2025 Mar 27. [doi: 10.1111/2047-3095.70009] [Medline: 40143628]
18. Teng A, Wilcox A. Simplified data science approach to extract social and behavioural determinants: a retrospective chart review. BMJ Open 2022 Jan 18;12(1):e048397. [doi: 10.1136/bmjopen-2020-048397] [Medline: 35042703]
19. Latifnejad Roudsari R. Methodological issues in conducting retrospective record reviews. JMRH 2020;8:2383-2384. [doi: 10.22038/jmrh.2020.16739]
20. Vassar M, Holzmann M. The retrospective chart review: important methodological considerations. J Educ Eval Health Prof 2013;10:12. [doi: 10.3352/jeehp.2013.10.12] [Medline: 24324853]
21. BUTLER JJ, QUINLAN JW. Internal audit in the department of medicine of a community hospital; two years' experience. J Am Med Assoc 1958 May 31;167(5):567-572. [doi: 10.1001/jama.1958.02990220037011] [Medline: 13538741]
22. Allison JJ, Wall TC, Spettell CM, et al. The art and science of chart review. Jt Comm J Qual Improv 2000 Mar;26(3):115-136. [doi: 10.1016/s1070-3241(00)26009-4] [Medline: 10709146]
23. Engel L, Henderson C, Fergenbaum J, Colantonio A. Medical record review conduction model for improving interrater reliability of abstracting medical-related information. Eval Health Prof 2009 Sep;32(3):281-298. [doi: 10.1177/0163278709338561] [Medline: 19679636]
24. Jain N, Bernacki RE. Goals of care conversations in serious illness: a practical guide. Med Clin North Am 2020 May;104(3):375-389. [doi: 10.1016/j.mcna.2019.12.001] [Medline: 32312404]
25. Acharya AS, Prakash A, Saxena P, Nigam A. Sampling: why and how of it? Indian J Med Spec 2013;4(2):330-333. [doi: 10.7713/ijms.2013.0032]
26. Furey E. Random number generator. Calculator Soup Online calculators. 2023 Sep. URL: https://www.calculatorsoup.com/calculators/statistics/random-number-generator.php [accessed 2024-02-26]
27. Erdfelder E, Faul F, Buchner A. GPOWER: A general power analysis program. Behavior Research Methods, Instruments, & Computers 1996 Mar;28(1):1-11. [doi: 10.3758/BF03203630]
28. Gearing RE, Mian IA, Barber J, Ickowicz A. A methodology for conducting retrospective chart review research in child and adolescent psychiatry. J Can Acad Child Adolesc Psychiatry 2006 Aug;15(3):126-134. [Medline: 18392182]
29. Hirschman KB, Hodgson NA. Evidence-based interventions for transitions in care for individuals living with dementia. Gerontologist 2018 Jan 18;58(suppl_1):S129-S140. [doi: 10.1093/geront/gnx152] [Medline: 29361067]
30. Olding M, McMillan SE, Reeves S, Schmitt MH, Puntillo K, Kitto S. Patient and family involvement in adult critical and intensive care settings: a scoping review. Health Expect 2016 Dec;19(6):1183-1202. [doi: 10.1111/hex.12402] [Medline: 27878937]
31. Baxter R, Fromme EK, Sandgren A. Patient identification for serious illness conversations: a scoping review. Int J Environ Res Public Health 2022 Mar 31;19(7):4162. [doi: 10.3390/ijerph19074162] [Medline: 35409844]
32. Davidson JE, Powers K, Hedayat KM, American College of Critical Care Medicine Task Force 2004-2005, Society of Critical Care Medicine, et al. Clinical practice guidelines for support of the family in the patient-centered intensive care unit: American College of Critical Care Medicine Task Force 2004-2005. Crit Care Med 2007 Feb;35(2):605-622. [doi: 10.1097/01.CCM.0000254067.14607.EB] [Medline: 17205007]
33. Frampton SB, Guastello S, Hoy L, et al. Harnessing evidence and experience to change culture: a guiding framework for patient and family engaged care. NAM Perspectives ;7(1). [doi: 10.31478/201701f]
34. Holland DE, Vanderboom CE, Dose AM, et al. Death and grieving for family caregivers of loved ones with life-limiting illnesses in the era of COVID-19: considerations for case managers. Prof Case Manag 2021;26(2):53-61. [doi: 10.1097/NCM.0000000000000485] [Medline: 33181608]
35. Dijkstra BM, Felten-Barentsz KM, van der Valk MJM, et al. Family participation in essential care activities: needs, perceptions, preferences, and capacities of intensive care unit patients, relatives, and healthcare providers-an integrative review. Aust Crit Care 2023 May;36(3):401-419. [doi: 10.1016/j.aucc.2022.02.003] [Medline: 35370060]
36. Banister G, Carroll DL, Dickins K, et al. Nurse-sensitive indicators during COVID-19. Int J Nurs Knowl 2022 Jul;33(3):234-244. [doi: 10.1111/2047-3095.12372] [Medline: 35582773]
37. Findley TW, Daum MC. Research in physical medicine and rehabilitation. III. The chart review or how to use clinical data for exploratory retrospective studies. Am J Phys Med Rehabil 1989 Jun;68(3):150-157. [Medline: 2730784]
38. Ritter SRF, Cardoso AF, Lins MMP, Zoccoli TLV, Freitas MPD, Camargos EF. Underdiagnosis of delirium in the elderly in acute care hospital settings: lessons not learned. Psychogeriatrics 2018 Jul;18(4):268-275. [doi: 10.1111/psyg.12324] [Medline: 30133935]
39. Siems A, Banks R, Holubkov R, et al. Structured chart review: assessment of a structured chart review methodology. Hosp Pediatr 2020 Jan;10(1):61-69. [doi: 10.1542/hpeds.2019-0225] [Medline: 31879317]

Abbreviations

EMR: electronic medical record

XSL•FO
RenderX

RCR: retrospective chart review

Edited by A Mavragani; submitted 16.09.24; peer-reviewed by M Cesare; revised version received 02.05.25; accepted 11.05.25; published 03.06.25.

Please cite as:
Morgan J, Cahill J, Ritchie C, Zhang L, Gazarian P
Methods of Piloting an Abstraction Tool to Describe Family Engagement in the Hospital Setting: Retrospective Chart Review
JMIR Form Res 2025;9:e66549
URL: https://formative.jmir.org/2025/1/e66549
doi:10.2196/66549

© Jennifer Morgan, Jennifer Cahill, Christine Ritchie, Lingling Zhang, Priscilla Gazarian. Originally published in JMIR Formative Research (https://formative.jmir.org), 3.6.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

A Portal-Based Intervention (PATTERN) Designed to Support Medication Use Among Older Adults: Feasibility and Acceptability Study

Allison Pack, MPH, PhD; Stacy C Bailey, MPH, PhD; Rachel O'Conor, MPH, PhD; Evelyn Velazquez, MPH; Guisselle Wismer, MPH; Fangyu Yeh, MPH; Laura M Curtis, MS; Kenya Alcantara, BA; Michael S Wolf, MA, MPH, PhD

Division of General Internal Medicine, Feinberg School of Medicine, Northwestern University, 750 North Lake Shore Drive, 10th Floor, Chicago, IL, United States

Corresponding Author:
Allison Pack, MPH, PhD
Division of General Internal Medicine, Feinberg School of Medicine, Northwestern University, 750 North Lake Shore Drive, 10th Floor, Chicago, IL, United States

Abstract

Background: Poor medication adherence among older adults with multiple chronic conditions and polypharmacy is a public health concern stemming from distinct challenges. Prior interventions have largely used a one-size-fits-all approach or resource-intensive approaches inappropriate for busy primary care clinics.

Objective: To address this, Phenotyping Adherence Through Technology-Enabled Reports and Navigation (PATTERN) was adapted from prior work. PATTERN is a portal-based intervention for monitoring self-reported medication adherence challenges among older adults in primary care. This study sought to implement and evaluate PATTERN's feasibility and acceptability.

Methods: We conducted a patient randomized study with a posttest design. Primary care physicians at the participating health center were informed of the study, and approval was obtained to contact their patients. Patient eligibility included being aged 60 years or older, having prescription medications for ≥8 chronic conditions, and an upcoming visit with a physician who had provided approval. Potentially eligible patients were identified using an electronic health record query, and a research coordinator phoned them to confirm eligibility, assess interest, obtain consent, and conduct enrollment. Randomization occurred following enrollment. Those randomized to PATTERN received a medication adherence assessment in their patient portal accounts several days ahead of their visit. The assessment identified whether a patient was experiencing a medication adherence challenge, and if so, the type (cognitive, psychological, medical, regimen-related, social, or economic). Identified challenges were sent to the patient's primary care physician. Assessment delivery several days ahead of a visit was thought to offer sufficient time for patients to complete it and clinicians to review any challenges. Approximately 2 weeks after visits, the coordinator recontacted participants to conduct posttest interviews. This ensured clinicians had sufficient time to respond to challenges during or after visits. Posttest interviews measured the self-reported use of the portal, demographic and health characteristics, and for those randomized to PATTERN, intervention satisfaction. Self-reported data were captured in REDCap and analyzed descriptively. Electronic health record data were also analyzed descriptively to objectively identify feasibility, that is, whether intervention arm participants completed the PATTERN assessment.

Results: We enrolled 64 participants (32 received usual care, and 32 received intervention). Most were female (66%, 42/64), not Hispanic or Latino (94%, 60/64), and identified as White (58%, 37/64). The average (SD) age was 75 (6.8) years. Most participants (80%) self-reported using the patient portal ≥12 times per year. However, electronic health record data revealed that less than half of all participants randomized to PATTERN (47%, 15/32) completed the medication adherence assessment. Of those who remembered completing it, 60% (3/5) were very satisfied with the experience and 20% (1/5) were a little satisfied.

Conclusions: PATTERN has the potential for use with older primary care patients experiencing multiple chronic conditions and polypharmacy. Yet, further adaptation is needed to ensure recipients access their patient portal accounts and complete assessments.

Trial Registration: ClinicalTrials.Gov NCT05766423; https://clinicaltrials.gov/study/NCT05766423

(JMIR Form Res 2025;9:e71676) doi:10.2196/71676

KEYWORDS

older adults; multiple chronic conditions; polypharmacy; primary care; medication adherence; patient portal

XSL•FO
RenderX

Introduction

Older adults with multiple chronic conditions and polypharmacy are at increased risk of suboptimal medication adherence [1,2], resulting in poor treatment outcomes, adverse drug events, and reduced quality of life [3]. In fact, poor medication adherence in this population is considered a substantial public health concern, impacting both morbidity and financial strain [4].

Barriers to medication adherence, or taking medication as prescribed, are well known in older adults with multiple chronic conditions and polypharmacy and comprise distinct challenges that can be mapped to the model of medication self-management; these include challenges classified as cognitive, psychological, medical, regimen-related, social, or economic [5]. Yet, the vast number of interventions designed to address medication adherence challenges do so using a "one-size-fits-all approach," as opposed to a more tailored approach, targeting the distinct challenges experienced by individual patients [6]. Furthermore, interventions that do tailor approaches are often resource-intensive, requiring clinical pharmacists, community health workers, or other individuals tasked with providing intensive one-on-one sessions [6-8]. These resource-intensive interventions and activities are not always available or appropriate for busy primary care settings. Yet, primary care remains the front line of health care for most older adults in the United States [9].

To address this, we created PATTERN: Phenotyping Adherence Through Technology-Enabled Reports and Navigation. Adapted from a previous study by our team [10,11], PATTERN uses the patient portal to identify and classify medication adherence challenges for review and resolution by a primary care physician or designated care team. It was designed to be low-cost, sustainable, and scalable using existing electronic health record (EHR) technology. We conducted a small feasibility study of this intervention to determine its feasibility and acceptability among older primary care patients with multiple chronic conditions and polypharmacy.

Methods

Study Design

We conducted a patient randomized feasibility study, using a posttest design to examine the effects of PATTERN. This paper largely presents data from the intervention arm pertaining to the feasibility and acceptability of PATTERN.

Eligibility Criteria, Recruitment, and Setting

We first sought approval, via email, from physicians affiliated with any primary care clinic (including geriatrics) within the participating academic health center in Chicago, IL, to contact their patients for recruitment into PATTERN. Patients were eligible to participate if they were aged 65 years or older, spoke English as their primary language, self-reported contending with at least two chronic health conditions, were prescribed eight or more medications they were primarily responsible for taking, and had a visit scheduled within the next 1 week with the physician who had provided contact approval. For intervention purposes, individuals also needed to have access to the internet, an active email address, and a patient portal account. Exclusion criteria included having participated in our prior, formative research or having any severe, uncorrectable visual, hearing or cognitive impairments that would have precluded study consent or participation.

Potentially eligible patient participants were identified using an EHR query, phoned by a trained research coordinator, informed about the study, and screened for eligibility. Those who met all eligibility criteria and expressed an interest in participating then provided electronic informed consent prior to enrollment. The research coordinator scheduled remote posttest interviews with enrolled participants; these occurred approximately 2 weeks after their scheduled clinic visit.

Randomization

Participants were randomized one-to-one to usual care or the PATTERN intervention using random block sizes of two and four, stratified by clinic type (geriatrics or other primary care clinic). The allocation table was created using SAS PROC PLAN and implemented using REDCap software [12]. Randomization took place immediately following enrollment (Multimedia Appendix 1). The purpose was to be able to compare differences between those randomized to PATTERN and those in usual care.

Intervention Activities

Participants who were randomized to the PATTERN intervention were sent an email 2 to 3 days ahead of their clinic visit. The email instructed them to log into their patient portal accounts to complete a brief medication adherence assessment prior to their upcoming primary care visit. Those who logged in were instructed to begin the two-tiered assessment adapted from a prior study [10]. The first three questions identified if the participant was experiencing any challenges taking their medications as prescribed. Specifically, these asked: (1) During the past month, did you decide to stop taking any of your medicines without telling your doctor? (2) During the past month, did you decide to take more or less of any of your medicines than you were supposed to without telling your doctor? (3) Are you having any problems with your medicines? Those who endorsed at least one adherence challenge were asked to respond to a subsequent set of 18 questions. These questions, informed by the model of medication self-management, were designed to "phenotype" or identify the distinct type of medication adherence challenge a participant might have been experiencing. Challenges could be classified as cognitive (eg, forgetfulness), psychological (eg, health literacy, depression, or motivation), medical (eg, acute illness), regimen-related (eg, side effects or complex dosing schedules), social (eg, transportation or support), and economic (eg, cost) [5].

Any identified and "phenotyped" challenges were designed to be automatically and securely forwarded, via the EHR, to the participant's primary care physician with whom they had an upcoming clinic visit. The intent was to ensure physicians received relevant and actionable data from patients regarding any medication challenges they might have been experiencing. This information could subsequently inform the scheduled visit

XSL•FO
RenderX

or referral to available support services such as social workers or clinical pharmacists, depending on clinic-specific resources.

Data Collection Activities

Data were collected between March and August 2024. Remote, structured study interviews were held with individual participants approximately 2 weeks after each participant's regularly scheduled clinic visit. The research coordinator phoned participants at previously agreed-upon times and administered a brief questionnaire. Data were captured using the REDCap software [12]. On average, it took participants 20 minutes to complete the questionnaire, and all participants were compensated US $50 for their time.

Measures

Participant Characteristics

All participants reported their age, sex, race, and ethnicity. Perceived health status was measured using the single-item patient-reported measurement information system (PROMIS) global health measure that asks participants to rate their health from poor to excellent using a 5-point Likert scale [13]. Depression was captured using the validated 6-item PROMIS depression measure, with higher scores representing greater symptom burden [13]. Patient activation was captured using the validated 10-item Consumer Health Activation Index (CHAI) [14]. Finally, the use of the patient portal was examined by asking participants how often they use their portal accounts and for what purposes. This self-reported information was supplemented by EHR data identifying the completion of any portal questionnaire in the past year.

Intervention Feasibility

We used data captured from self-reporting and from the EHR to examine intervention feasibility among participants randomized to PATTERN. Specifically, we were interested in identifying (1) the number of adherence assessments received (opened) among those sent; (2) the number of adherence assessments completed among those received; and (3) the number of participants who spoke with their clinical care teams about the adherence assessments in those who completed them.

Intervention Acceptability

Participants who completed the intervention were asked how satisfied they were with the PATTERN questionnaire. Response options ranged on a 4-point Likert scale from "not at all satisfied" to "very satisfied."

Analysis

Descriptive statistics were used to examine participant characteristics, overall and by study arm. All feasibility and acceptability data were also assessed using descriptive statistics in the intervention arm only. Analyses were conducted using SAS 9.4 software.

Ethical Considerations

All study procedures were approved by the Northwestern University's Institutional Review Board (STU00217555). All participants provided electronic informed consent prior to study participation and were compensated US $50 for their time. Data presented herein have all been deidentified.

Results

Participant Characteristics

Participant characteristics are presented in Table 1. In brief, a total of 64 patient participants were enrolled. The average (SD) age was 75 (6.8) years. Most indicated they were female (66%, 42/64), non-Hispanic (94%, 60/64), White (58%, 37/64), and in good or better health (64%, 41/64). A total of 16% (10/64) were classified as having mild or moderate depression, and 38% (24/64) as having low patient activation. A large majority of participants (80%, 51/54) reported they used the patient portal more than 12 times a year, and that when they access the portal, they view laboratory results (98%, 63/64), confirm appointments (89%, 57/64), and send messages to their care teams (91%, 58/64). Actual use, assessed via EHR data, revealed that most patients (72%, 46/64) had previously (within the past year) completed any type of assessment in the patient portal. No significant differences in participant characteristics were observed between study arms.

XSL•FO
RenderX

Table . Descriptive statistics of participant characteristics and technology use, overall and by study arm, captured via self-reporting and the electronic health records between March and August 2024.

Characteristics	Overall (n=64)	Usual care (n=32)	Intervention (n=32)
Gender, n (%)			
Male	22 (34.4)	12 (37.5)	10 (31.3)
Female	42 (65.6)	20 (62.5)	22 (68.8)
Age, years			
Mean (SD)	75.0 (6.8)	75.1 (7.5)	74.9 (6.1)
Median (range)	74.5 (66-93)	74 (66 - 93)	74.5 (66-90)
Ethnicity, n (%)			
Not Hispanic or Latino	60 (93.8)	29 (90.6)	31 (96.9)
Hispanic or Latino	1 (1.6)	1 (3.1)	N/A[a]
Refused	3 (4.7)	2 (6.3)	1 (3.1)
Race, n (%)			
White	37 (57.8)	19 (59.4)	18 (56.3)
Black or African-American	23 (35.9)	11 (34.4)	12 (37.5)
Other	3 (4.7)	2 (6.3)	1 (3.1)
Refused	1 (1.6)	N/A	1 (3.1)
Health status, n (%)			
Excellent	1 (1.6)	1 (3.1)	N/A
Very good	11 (17.2)	3 (9.4)	8 (25.0)
Good	29 (45.3)	17 (53.1)	12 (37.5)
Fair	20 (31.3)	10 (31.3)	10 (31.3)
Poor	3 (4.7)	1 (3.1)	2 (6.3)
PROMIS[b] depression, n (%)			
Within normal limits	54 (84.4)	29 (90.6)	25 (78.1)
Mild or moderate	10 (15.6)	3 (9.4)	7 (21.9)
CHAI,[c] n (%)			
Low	24 (38.1)	9 (28.1)	15 (48.4)
Moderate	27 (42.9)	16 (50.0)	11 (35.5)
High	12 (19.1)	7 (21.9)	5 (16.1)
Patient portal use			
Self-reported frequency, n (%)			
Never	2 (3.1)	N/A	2 (6.3)
1 to 6 times per year	1 (1.6)	1 (3.1)	N/A
7 to 12 times per year	10 (15.6)	4 (12.5)	6 (18.8)
More than 12 times per year	51 (79.7)	27 (84.4)	24 (75.0)
Self-reported activities, n (%)			
Requesting appointments	40 (62.5)	22 (68.8)	18 (56.3)
Viewing laboratory results	63 (98.4)	32 (100.0)	31 (96.9)
Refilling prescriptions	51 (79.7)	26 (81.3)	25 (78.1)
Sending a message to the care team	58 (90.6)	29 (90.6)	29 (90.6)

XSL•FO
RenderX

Characteristics	Overall (n=64)	Usual care (n=32)	Intervention (n=32)
Paying bills	32 (50.0)	13 (40.6)	19 (59.4)
Confirming appointments	57 (89.1)	29 (96.6)	28 (87.5)
Other	2 (3.1)	1 (3.1)	1 (3.1)
Actual use, n (%)			
Any questionnaire completion in the past year	46 (71.9)	25 (78.1)	21 (65.6)

[a]N/A: not applicable.

[b]PROMIS: patient-reported measurement information system.

[c]CHAI: Consumer Health Activation Index.

Intervention Feasibility and Acceptability

Our assessment of intervention feasibility revealed that 47% (15/32) of participants in the intervention arm completed the adherence assessments that were sent in their patient portal accounts. Of those, only 1 indicated they had any adherence challenges, according to EHR data, though this participant self-reported they did not remember having any challenges. Similarly, when asked by the research coordinator, 75% (24/32) of intervention arm participants self-reported they were unaware they had received an email or portal message asking them to complete the PATTERN medication adherence assessment, and therefore did not complete it. When probed for an explanation, most indicated they had accessed their email but did not notice the request (71%, 17/24). Of those who self-reported receiving the request (25%, 8/32), all but 1 found the adherence assessment to be somewhat or very easy to access, and most (63%, 5/8) reported they completed it. Of the 5 who said they completed the survey, 3 (60%) were very satisfied with the experience, 1 (20%) was a little satisfied, and 1 (20%) did not reply. Moreover, although none of these participants reported having any challenges with their medications, 2 revealed they had discussed the medication assessment and medication taking strategies with their clinical care teams and were satisfied with that conversation. Participants who accessed but did not complete the assessment (37%, 3/8), revealed it was not because they encountered challenges accessing or responding to it; rather, they indicated they did not want to complete it at that time.

Discussion

This small patient randomized feasibility study examined the feasibility and acceptability of the PATTERN intervention among older adults with multiple chronic conditions and polypharmacy from a large academic medical center in Chicago, IL. Results revealed that the previsit delivery of portal-based assessments has the potential to enhance patient-provider communication about medication adherence in primary care, as most of our sample self-reported being active users of the patient portal and had a record of a completed portal assessment in the EHR in the last year. This is an important finding as our studies have previously shown that while highly resource-intensive interventions can increase medication adherence among older adults, less-intensive options, such as portal-based assessments, are needed for busy primary care clinics [7]. This finding also has far-reaching implications, as health systems are increasingly relying on patient-reported outcomes collected via the patient portal [15-17].

However, less than half of all intervention arm participants completed the PATTERN adherence assessment ahead of their clinical care visit, as instructed. As such, the current approach likely requires additional nudges or supports to ensure all recipients are able to access and complete the assessments in a timely manner. This could include (1) information explaining how to know when assessments are available for completion in the patient portal; (2) more tailored information designed to entice recipients to open the portal message; and (3) detailed information on how to complete assessments on time. Prior portal-based studies conducted among older adults have also shown that this population could benefit from usability training [18,19]. Yet beyond a “how-to” training, information on why portal-based assessments are useful has also been identified as a need [20]. Future studies should explore the nudges, supports, or training older adults find acceptable, feasible, and appropriate. Furthermore, it may be that assessment results need to be sent to the patient ahead of their visit. These could be accompanied by a list of result-specific questions (perhaps even populated by artificial intelligence) that patients could ask during their visit [21]. Not only would this list serve as a reminder that they had completed the assessment, but as other studies using question prompt lists have shown, it can further care engagement [22,23].

The few participants in our study who reported completing the PATTERN assessment noted they were largely satisfied with it and the resulting conversations they had with their care teams. Had we conducted posttest interviews immediately following the clinical care visit, more participants would have likely been able to share their experiences and whether they found the intervention to be acceptable. Increasingly, studies are revealing that older adults typically have positive experiences with patient portals and technology-delivered assessments [24,25]

Results from this study need to be considered in light of the study limitations. First, this was a small feasibility study conducted within a single academic health system. It is possible that a larger, more diverse study would have resulted in different findings. However, the purpose of this study was to obtain preliminary data that could inform a larger, more diverse trial. There was also no age limit for participation in this study. It is

XSL•FO
RenderX

possible that PATTERN may be more feasible for the "younger" older adults, for example, those individuals who are between the ages of 65 and 75 years. Finally, we may need to account for "proxy users," other individuals appointed by the patient to access and respond to needs in their patient portal accounts. Although nearly all participants in our study had not officially authorized "proxy use" in their accounts, it is possible they relied on unofficial use. A recent national survey revealed that nearly half of all older adults between the ages of 50 and 80 years have allowed others to use their portal accounts [17]. This might occur if a patient's caregiver, for example, used the patient's login information to confirm appointments, update medication information, or use other portal-based features, including assessments [26]. It is possible that participants in the PATTERN intervention arm did not remember completing the medication adherence assessment because their proxy user had completed it for them.

In conclusion, PATTERN, a remote medication adherence monitoring intervention, shows promise for use with older primary care patients experiencing multiple chronic conditions and polypharmacy. However, further adaptation and training are needed to ensure participants are accessing their patient portal accounts in a timely manner ahead of scheduled clinic visits. Results from this small feasibility study may inform other research and clinical activities intending to capture self-reported data from older adults via the patient portal.

Acknowledgments

The work was funded in part by the National Institutes of Health's National Institute on Aging, Grant Number P30AG059988. REDCap software is supported by the National Institutes of Health's National Center for Advancing Translational Sciences, Grant Number UL1TR001422. The opinions expressed in this paper are those of the authors and do not necessarily represent that of the National Institutes of Health.

Data Availability

Deidentified data pertaining to this study may be obtained with reasonable request from the corresponding author.

Conflicts of Interest

APP reports grants from Merck, Pfizer, RRF Foundation for Aging, Lundbeck, Gilead, and Eli Lilly through her institution; and personal fees from Gilead and Remo Health. SCB reports grants from the NIH, Merck, Pfizer, RRF Foundation for Aging, Lundbeck, Gilead, and Eli Lilly through her institution; and personal fees from Gilead, Pfizer, University of Westminster, Lundbeck, and Luto outside the submitted work. RO reports grants from NIH through her institution. Dr. Wolf reports grants from the NIH (NIA, NIDDK, NINR, NHLBI, NINDS), Gordon and Betty Moore Foundation, RRF Foundation for Aging and Eli Lilly, Glaxo Smith Klein, and personal fees from Pfizer, Sanofi, Luto UK, University of Westminster, and Lundbeck. All other authors have no relationships to disclose.

Multimedia Appendix 1
CONSORT flow diagram.
[PDF File, 131 KB - formative_v9i1e71676_app1.pdf]

References

1. Brown MT, Bussell JK. Medication adherence: WHO cares? Mayo Clin Proc 2011 Apr;86(4):304-314. [doi: 10.4065/mcp.2010.0575] [Medline: 21389250]
2. Denneboom W, Dautzenberg MGH, Grol R, De Smet P. Analysis of polypharmacy in older patients in primary care using a multidisciplinary expert panel. Br J Gen Pract 2006 Jul;56(528):504-510. [Medline: 16834876]
3. Delara M, Murray L, Jafari B, et al. Prevalence and factors associated with polypharmacy: a systematic review and meta-analysis. BMC Geriatr 2022 Jul 19;22(1):601. [doi: 10.1186/s12877-022-03279-x] [Medline: 35854209]
4. Marcum ZA, Hanlon JT, Murray MD. Improving medication adherence and health outcomes in older adults: an evidence-based review of randomized controlled trials. Drugs Aging 2017 Mar;34(3):191-201. [doi: 10.1007/s40266-016-0433-7] [Medline: 28074410]
5. Bailey SC, Oramasionwu CU, Wolf MS. Rethinking adherence: a health literacy-informed model of medication self-management. J Health Commun 2013;18 Suppl 1(Suppl 1):20-30. [doi: 10.1080/10810730.2013.825672] [Medline: 24093342]
6. Xu HY, Yu YJ, Zhang QH, Hu HY, Li M. Tailored interventions to improve medication adherence for cardiovascular diseases. Front Pharmacol 2020;11:510339. [doi: 10.3389/fphar.2020.510339] [Medline: 33364935]
7. Federman AD, O'Conor R, Mindlis I, et al. Effect of a self-management support intervention on asthma outcomes in older adults: the SAMBA study randomized clinical trial. JAMA Intern Med 2019 Aug 1;179(8):1113-1121. [doi: 10.1001/jamainternmed.2019.1201] [Medline: 31180474]

XSL•FO
RenderX

8. O'Conor R, Arvanitis M, Wismer G, et al. Rationale and design of the regimen education and messaging in diabetes (REMinD) trial. Contemp Clin Trials 2019 Aug;83:46-52. [doi: 10.1016/j.cct.2019.06.016] [Medline: 31260791]
9. Realising the potential of primary health care. : OECD; 2020 May 30 URL: https://www.oecd.org/en/publications/realising-the-potential-of-primary-health-care_a92adee4-en.html [accessed 2025-04-17] [doi: 10.7748/phc.30.2.14.s14]
10. Yoon ES, Hur S, Curtis LM, et al. A multifaceted intervention to improve medication adherence in kidney transplant recipients: an exploratory analysis of the fidelity of the TAKE IT trial. JMIR Form Res 2022 May 5;6(5):e27277. [doi: 10.2196/27277] [Medline: 35511225]
11. Pack AP, Bailey SC, O'Conor R, et al. Phenotyping adherence through technology-enabled reports and navigation (the PATTERN Study): qualitative study for intervention adaptation using the exploration, preparation, implementation, and sustainment framework. JMIR Form Res 2024 Oct 17;8:e54916. [doi: 10.2196/54916] [Medline: 39418094]
12. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381. [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
13. Cella D, Riley W, Stone A, et al. The patient-reported outcomes measurement information system (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. J Clin Epidemiol 2010 Nov;63(11):1179-1194. [doi: 10.1016/j.jclinepi.2010.04.011] [Medline: 20685078]
14. Wolf MS, Smith SG, Pandit AU, et al. Development and validation of the consumer health activation index. Med Decis Making 2018 Apr;38(3):334-343. [doi: 10.1177/0272989X17753392] [Medline: 29436308]
15. Deshpande PR, Rajan S, Sudeepthi BL, Abdul Nazir CP. Patient-reported outcomes: a new era in clinical research. Perspect Clin Res 2011 Oct;2(4):137-144. [doi: 10.4103/2229-3485.86879] [Medline: 22145124]
16. Smith SM, Wallace E, O'Dowd T, Fortin M. Interventions for improving outcomes in patients with multimorbidity in primary care and community settings. Cochrane Database Syst Rev 2021 Jan 15;1(1):CD006560. [doi: 10.1002/14651858.CD006560.pub4] [Medline: 33448337]
17. Kirch M, Solway E, Roberts S, et al. Use and experiences with patient portals among older adults: University of Michigan national poll on healthy aging. : University of Michigan; 2023 URL: https://www.healthyagingpoll.org/reports-more/report/use-and-experiences-patient-portals-among-older-adults [accessed 2025-04-17]
18. Son H, Nahm ES, Zhu S, et al. Older adults' perception on and use of patient portals: a comparative analysis of two samples. Comput Inform Nurs 2022 Jan 1;40(1):61-68. [doi: 10.1097/CIN.0000000000000802] [Medline: 34347646]
19. Nahm ES, Son H, Yoon JM. Older adults' use of patient portals: Experiences, challenges, and suggestions shared through discussion board forums. Geriatr Nurs 2020;41(4):387-393. [doi: 10.1016/j.gerinurse.2019.12.001] [Medline: 31899003]
20. Price-Haywood EG, Harden-Barrios J, Ulep R, Luo Q. eHealth literacy: patient engagement in identifying strategies to encourage use of patient portals among older adults. Popul Health Manag 2017 Dec;20(6):486-494. [doi: 10.1089/pop.2016.0164] [Medline: 28384076]
21. Vordenberg SE, Nichols J, Marshall VD, Weir KR, Dorsch MP. Investigating older adults' perceptions of AI tools for medication decisions: vignette-based experimental survey. J Med Internet Res 2024 Dec 16;26:e60794. [doi: 10.2196/60794] [Medline: 39680885]
22. Brandes K, Linn AJ, Butow PN, van Weert JCM. The characteristics and effectiveness of question prompt list interventions in oncology: a systematic review of the literature. Psychooncology 2015 Mar;24(3):245-252. [doi: 10.1002/pon.3637] [Medline: 25082386]
23. Sansoni JE, Grootemaat P, Duncan C. Question prompt lists in health consultations: a review. Patient Educ Couns 2015 Jun 3:S0738-3991(15)00258-X. [doi: 10.1016/j.pec.2015.05.015] [Medline: 26104993]
24. Weir KR, Maitah Y, Vordenberg SE. Older adults' experiences with an online survey. JMIR Hum Factors 2025 Jan 8;12:e65684. [doi: 10.2196/65684] [Medline: 39773887]
25. Sakaguchi-Tang DK, Bosold AL, Choi YK, Turner AM. Patient portal use and experience among older adults: systematic review. JMIR Med Inform 2017 Oct 16;5(4):e38. [doi: 10.2196/medinform.8092] [Medline: 29038093]
26. Wolff JL, Dukhanin V, Burgdorf JG, DesRoches CM. Shared access to patient portals for older adults: implications for privacy and digital health equity. JMIR Aging 2022 May 4;5(2):e34628. [doi: 10.2196/34628] [Medline: 35507405]

ABBREVIATIONS

CHAI: Consumer Health Activation Index
EHR: electronic health record
PATTERN: Phenotyping Adherence Through Technology-Enabled Reports and Navigation
PROMIS: patient-reported measurement information system

XSL•FO
RenderX

Edited by A Mavragani; submitted 23.01.25; peer-reviewed by M Shade, S Leitch; revised version received 11.02.25; accepted 13.02.25; published 24.04.25.

Please cite as:
Pack A, Bailey SC, O'Conor R, Velazquez E, Wismer G, Yeh F, Curtis LM, Alcantara K, Wolf MS
A Portal-Based Intervention (PATTERN) Designed to Support Medication Use Among Older Adults: Feasibility and Acceptability Study
JMIR Form Res 2025;9:e71676
URL: https://formative.jmir.org/2025/1/e71676
doi:10.2196/71676

© Allison Pack, Stacy C Bailey, Rachel O'Conor, Evelyn Velazquez, Guisselle Wismer, Fangyu Yeh, Laura M Curtis, Kenya Alcantara, Michael S Wolf. Originally published in JMIR Formative Research (https://formative.jmir.org), 24.4.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Cardiac Self-Efficacy Improvement in a Digital Heart Health Program: Secondary Analysis From a Feasibility and Acceptability Pilot Study

Kimberly G Lockwood[1], PhD; Priya R Kulkarni[2], PhD; OraLee H Branch[1], PhD; Sarah A Graham[1], PhD

[1]Clinical Research Department, Lark Health, 809 Cuesta Dr. Suite B #1033, Mountain View, CA, United States

[2]Digital Health Technologies Department, Roche Information Solutions, Santa Clara, CA, United States

Corresponding Author:
Sarah A Graham, PhD
Clinical Research Department, Lark Health, 809 Cuesta Dr. Suite B #1033, Mountain View, CA, United States

Abstract

Background: Lifestyle modification programs play a critical role in preventing and managing cardiovascular disease (CVD). A key aim of many programs is improving patients' self-efficacy. In-person lifestyle modification programs can enhance self-efficacy in managing CVD risk, also known as cardiac self-efficacy (CSE). However, such programs are typically staffing and resource intensive. Digital lifestyle modification programs may offer a scalable and accessible way to improve CSE, but this has not been shown in prior research.

Objective: This study examined changes in CSE among individuals using a digital lifestyle modification program for cardiovascular health. Evaluation of improvement in CSE was a secondary goal of a feasibility and acceptability pilot study of a digital program for Heart Health.

Methods: Participants were individuals with elevated risk for CVD who enrolled in a 90-day pilot study that involved mobile app–based, artificial intelligence–powered health coaching and educational lessons focused on behaviors that promote cardiovascular health. Participants completed the 9-item CSE Scale at baseline and in month 2. Changes in confidence in participants' ability to manage their cardiovascular health were assessed.

Results: The sample included 273 (n=207, 61.2% female; mean age 59.3, SD 10.1 years) participants who submitted a complete CSE Scale at baseline and in month 2. The total CSE Scale score increased by 12.9% ($P<.001$) from baseline to month 2. Additionally, there were significant increases in mean score on each of the 9 individual CSE Scale items (all $P<.001$), with the largest increases in confidence "in knowing when to call or visit the doctor for your heart disease" (17% increase; $P<.001$), "in knowing how much physical activity is good for you" (16.3% increase; $P<.001$), and "that you can get regular aerobic exercise" (19% increase; $P<.001$).

Conclusions: The present analyses indicate that participants in a digital lifestyle modification program for cardiovascular health showed significant improvements in CSE within 2 months. This work adds to the growing literature examining ways to improve health-related self-efficacy and scalable access to programs for prevention and management of CVD.

(JMIR Form Res 2025;9:e60676) doi:10.2196/60676

KEYWORDS

digital health; self-efficacy; behavior change; prevention; cardiovascular health; digital heart health; lifestyle; heart health; health coaching; mobile phone

Introduction

Background

Cardiovascular disease (CVD) is the leading cause of death and the most common chronic disease in the United States, with 126.9 million adults (nearly 50% of the population) living with some form of CVD [1]. CVD is a key public health concern, as well as an economic burden, with costs related to CVD estimated to be US $378 billion, exceeding all other chronic health conditions [1,2]. Prevention and risk reduction for patients at risk for CVD is important to reduce mortality, disability, and economic burden [3]. According to the American Heart Association, engaging in a healthy lifestyle over the course of one's lifespan is the most important factor in CVD prevention [4]. As maintaining a healthy lifestyle is difficult for many adults [5], 1 method of CVD prevention and management is participation in behavioral interventions focused on healthy lifestyle modifications [4]. Position statements from the US Preventive Services Task Force state that lifestyle modification programs that emphasize a healthy diet and physical activity have a wide variety of cardiovascular health benefits and CVD

XSL•FO
RenderX

risk reduction among individuals with and without a diagnosis of CVD [6-8]. To enable individuals to make lasting behavior changes to improve their cardiovascular health, a key aim of many lifestyle modification programs is improving patients' self-efficacy to manage their health [7,9,10].

Importance of Self-Management in CVD

Self-efficacy, or confidence in one's ability to execute and control certain behaviors, is crucial for patient empowerment and self-management of cardiovascular health [11]. There are several evaluation tools for measuring self-efficacy in the context of disease self-management. For instance, general scales for managing chronic disease include the Chronic Disease Self-Efficacy Scale [12] and the Self-Efficacy for Managing Chronic Disease 6-Item Scale [13]. The Cardiac Self-Efficacy (CSE) Scale is a measure of self-efficacy specifically for self-management of CVD [14]. The CSE Scale is a self-report survey that evaluates individuals' beliefs in their ability to manage their cardiovascular health and the challenges that come with heart disease [14,15]. These self-efficacy evaluation tools provide insight into how confident an individual feels about carrying out behaviors that are beneficial to disease prevention and self-management.

Self-Efficacy in CVD Prevention and Management

Self-efficacy is a key target of lifestyle modification programs because it predicts a range of healthy behaviors and outcomes among individuals with CVD. For instance, higher self-efficacy is associated with greater engagement in self-care behaviors [16,17], such as higher levels of physical activity [18], healthier dietary habits [19], and better medication adherence [20]. Among cardiac patients, higher self-efficacy is also associated with better attendance in cardiac rehabilitation programs [21]. Meta-analytic results also indicate that stronger self-efficacy among those with CVD is linked with a range of metrics of health-related quality of life, such as physical functioning and mobility, emotional functioning, mental health, and social functioning [22].

A review of the literature indicated that in-person lifestyle modification programs for CVD have been successful in improving health-related self-efficacy [10]. One program involved 6 weeks of in-person, 2.5-hour educational sessions led by trained peer leaders who received 4 days of training; results of this study showed significant improvements in self-efficacy related to managing their health [23]. Another program that included 8 hour-long, individual sessions between each patient and a health professional found significant increases in self-efficacy related to self-care after 12 weeks [24].

There is also evidence that in-person lifestyle modification programs can improve self-efficacy specific to managing cardiovascular health, as measured by the CSE Scale. Recent studies show that intensive, in-person lifestyle modification programs for CVD led by nurses and health professionals can improve CSE scores over time [25,26], which have been shown to lead to downstream improvement of health behaviors and outcomes over time [14,27,28].

The evidence reviewed here shows that, although in-person programs have succeeded in improving self-efficacy, these intensive programs typically involve a high degree of human contact and staffing resources; this makes them difficult to implement or impossible to scale up to larger populations. Moreover, in-person programs can present access challenges to patients, such as having to travel long distances to attend [21]. As such, scalable and accessible lifestyle modification programs to improve CSE are needed.

Digital Health and CSE

Recent years have seen an increase in the availability of digital lifestyle modification programs for CVD prevention and management [29-31]. These digital health solutions have become increasingly available and can provide many benefits over in-person programs, such as increased accessibility, on-demand support, availability outside of typical working hours, lower cost, and scalability to larger populations [32]. These benefits have the potential to increase participation in lifestyle modifications for CVD [30]. Moreover, digital health programs for CVD prevention and management can significantly improve patient behaviors and show promise in delivering care that is accessible, cost-effective, and patient-focused [33].

Despite these benefits, many outcomes including CSE, remain unexplored in digital lifestyle modification programs and more evidence is required [31]. There is some evidence that digital interventions can increase general self-efficacy [34,35], suggesting that it may be possible for digital CVD programs to improve CSE. However, use of the CSE Scale in digital health studies is extremely limited. As such, the main objective of the present analyses was to test whether participation in an artificial intelligence (AI)–powered digital lifestyle modification program for heart health led to improved CSE.

Heart Health Program

The Lark Heart Health (Lark Technologies, Inc) program is a newly developed AI-powered lifestyle modification program that provides health behavior coaching to prevent and manage atherosclerotic CVD and coronary artery disease by targeting key CVD risk factors. We evaluated the acceptability and feasibility of this new program in a pilot study; a detailed study description and main outcome results for this pilot study are available elsewhere [36].

The Heart Health program targets individuals in primary prevention (ie, no history of CVD) and those in stable secondary prevention after a cardiac event. This program delivers fully remote, unlimited, real-time cardiovascular health coaching focused on digital nutrition therapy, medication adherence counseling, and personalized guidance on weight loss, physical activity, tobacco cessation, stress, and sleep. Heart Health is available to covered adult members of Lark's health care partners via a smartphone app. Full details on the content and design of the Heart Health program are provided in previous work [36]. Full detail on Lark's lifestyle change programs for diabetes prevention and hypertension care are provided elsewhere [37,38].

Study Objective

Beyond the primary acceptability and feasibility aims, a secondary goal of the Heart Health pilot study was to determine

XSL•FO
RenderX

whether CSE could improve early in this study. Specifically, we hypothesized that participants in the Heart Health pilot study who participated in the program for a minimum of 40 days would show improvement in CSE. This would indicate that a digital heart health program can help to improve CSE over a relatively short period, an important first step in developing scalable and accessible digital programs for improving self-efficacy among individuals at risk for CVD.

Methods

Study Design

This paper focuses on data from a real-world, single-arm, observational pilot study of a digital health app-based program called Lark Heart Health. The Heart Health program provides health behavior screeners and coaching to individuals at elevated risk for heart disease. The pilot study examined the feasibility of deploying screener surveys to participants of the program and the acceptability of coaching focused on building knowledge and improving self-management of atherosclerotic CVD risk. A detailed overview of study methods and primary outcomes [36] and predictors of interest in the program [39] have been reported in previously published works.

This study was 90 days in duration, during which participants received coaching on heart health and lifestyle modification. Participants could engage with the Heart Health program via (1) completing educational lessons, (2) engaging in coaching conversations with Lark's conversational AI-powered digital coach, (3) logging meals in the app, and (4) tracking their weight using a digital smart scale that automatically synced with the app.

Additionally, study participation involved completion of baseline screener surveys and assessment of CSE at study start and again in month 2 (after 40 days in the program). We measured CSE in month 2 for several reasons. First, self-efficacy has been shown to improve after relatively short digital health interventions for lifestyle change (eg, 6 week programs) [34], particularly compared to physical health outcomes that may take months or even years to improve. Second, our study involved completing several surveys and outcome measurements that were intentionally spread out over the course of the 90 days, rather than clustering all outcome measurements at the end of this study; this approach was aimed at reducing participant burden, a crucial factor for retention in digital health programs. Third, measuring CSE early in this study enabled us to maximize the sample size of individuals actively engaging with the program who completed the survey at both time points. Between study start date and the month 2 reassessment of CSE, the Heart Health program delivered 5 weekly educational lessons focused on risk factors for heart disease, understanding cholesterol, physical activity and how it impacts blood pressure, healthy eating, and how sodium impacts blood pressure.

Ethical Considerations

All participants provided informed consent to participate in this study through an electronic consent form. The pilot study received institutional review board approval (Advarra Pro00061694). Study personnel instituted appropriate safeguards to prevent any unauthorized use or disclosure of personal health information and implemented administrative, physical, and technical safeguards to protect the confidentiality, integrity, and availability of protected health information. Lark is compliant with HIPAA (Health Insurance Portability and Accountability Act) privacy and security rules and all applicable regulations. Lark is also SOC 2 (System and Organization Controls 2) and HITRUST (Health Information Trust Alliance) certified. As compensation for study participation, participants received a cellular smart scale and opportunities for nominal gift card incentives, as well as a Fitbit (Google LLC) upon study completion. None of these incentives were tied to completion of the CSE Scale.

Participants and Study Flow

The research team recruited participants through a health care partner and via online recruitment through 1nHealth using marketing emails or SMS messages and printed mailers (health care partner recruitment only). Exclusion criteria for the study included: BMI <25 and ≥50; serious uncontrolled health conditions that had been active in the last 6 months; current pregnancy or plans to become pregnant within the next 6 months; recent history of a medical professional advising against participation in a healthy lifestyle program; a medical reason preventing 10 consecutive minutes of moderate physical exercise; and not having a smartphone with internet connection. As the Heart Health program focused heavily on improving health behaviors (eg, diet and exercise), the study also excluded individuals who reported regularly engaging in strenuous physical activity and individuals who did not report unhealthy dietary behaviors.

Participants enrolled in the pilot study from March 31 to September 15, 2022, using the Lark app (version 5.2.6), downloading the Heart Health app and providing a first weight delineated study enrollment. To be included in the present analyses, participants had to be enrolled until the end of this study (ie, did not withdraw from this study), complete the first CSE Scale at baseline for descriptive analyses, and complete the second CSE Scale at month 2 for the CSE improvement analyses. Completion of the CSE Scale occurred outside of the Lark app and full completion was not required for study participation.

Data Collection and Measures

Participant Characteristics

We assessed basic demographic and health history characteristics at baseline using a modified version of the INTERHEART Modifiable Risk Score survey [40,41]. Items on this survey included age, sex, history of high blood pressure, tobacco use history, and typical physical activity in leisure time. Participants also provided their height and weight for calculation of BMI and the racial group they identified with.

CSE

Emails from the research team prompted participants to report on CSE shortly after study start (day 1 - 2 of this study) and again in month 2 (after day 40 of this study) using a modified version of the CSE Scale. The CSE Scale evaluates individuals'

XSL•FO
RenderX

beliefs in their ability to manage their cardiovascular health and the challenges that come with heart disease using 13 questions in 3 dimensions [14]. The Heart Health study used 2 dimensions of the CSE Scale that are most relevant to our study sample, “control illness” and “maintain functioning,” resulting in a 9-item survey. We did not include the “control symptoms” dimension, as the questions are not relevant to individuals without a CVD diagnosis (ie, most individuals in the present sample) [15]. Each item is scored on a 5-point Likert scale (0=not at all confident, 1=somewhat confident, 2=moderately confident, 3=very confident, and 4=completely confident) and the total score is calculated by summing scores on each item, for a maximum score of 36. As the program did not require participants to submit the CSE Scale and permitted them to skip items, calculation of the total sum score was not possible for all study participants.

Statistical Analysis

We conducted all statistical analyses with RStudio (version 2022.07.0; R Foundation). Continuous variables were normally distributed. The first set of analyses examined relationships between baseline participant characteristics and baseline CSE Scale score. Specifically, we tested the relationship of baseline CSE with demographic characteristics (age, sex, and race), health status and health history characteristics (high blood pressure, cardiac event, or CVD diagnosis—secondary prevention, and type II diabetes), and health behaviors (tobacco use and typical physical activity in leisure time). We used bivariate regressions for continuous variables (age and BMI), independent samples *t* tests for categorical variables with 2 levels (sex and health history variables), and 1-way ANOVAs for categorical variables with 3 levels (race and health behaviors). Baseline CSE analyses included participants who completed the baseline CSE Scale (n=341).

To address the primary study objective, the second set of analyses tested whether there was a statistically significant change in CSE from baseline to month 2, using paired *t* tests ($P<.05$). We first tested whether there was a significant change in the CSE sum score and then whether there was improvement in each of the 9 individual items on the CSE. These analyses included all participants who completed both the baseline CSE and month 2 CSE scales (n=273).

Results

Figure 1 shows the flow of participants through this study.

XSL•FO
RenderX

Figure 1. Flowchart of participants showing enrollment and inclusion in the analytic sample. CSE: cardiac self-efficacy.

XSL•FO
RenderX

Descriptive Statistics

We assessed descriptive statistics for baseline characteristics of participants who completed the baseline CSE Scale (n=341) and participants who completed both the baseline and month 2 CSE (n=273); both samples are shown in Table 1, as we conducted baseline descriptive analyses on the larger sample and improvement analyses on the smaller subsample (Table 1). As shown by the table, the 2 sets of participants were very similar with respect to key characteristics. Across both sets participants were, on average, aged 60 years (range=41 - 76 years) and classified as obese class I (BMI range=25 - 49). Just over 60% (n=207) of the participants were female, had a history of high blood pressure, and were never smokers. Approximately three-quarters of each sample reported typically being either mainly sedentary or engaging in low-effort mild exercise in their leisure time.

Table . Baseline characteristics of participants.

		Baseline CSE[a] complete (n=341)	Baseline and month 2 CSE complete (n=273)
Age (years), mean (SD)		60 (10.1)	59.3 (10.1)
Sex (female), n (%)		207 (60.7)	167 (61.2)
Race, n (%)			
	Black	26 (7.6)	22 (8.1)
	White	241 (70.7)	192 (70.3)
	Other	70 (20.5)	57 (20.9)
Recruitment source (health partner), n (%)		207 (60.7)	154 (56.4)
Baseline BMI, mean (SD)		32.8 (5.8)	32.9 (5.9)
High blood pressure history, n (%)		198 (58.1)	162 (59.3)
History of CVD[b] diagnosis or event, n (%)		41 (12.0)	34 (12.5)
Type II diabetes history, n (%)		52 (15.2)	37 (13.6)
Tobacco use status, n (%)			
	Never	203 (59.5)	171 (62.6)
	Former	110 (32.3)	82 (30.0)
	Current	27 (7.9)	19 (6.9)
Physical activity in leisure time, n (%)			
	Mainly sedentary	121 (35.5)	94 (34.4)
	Mild exercise, low effort	145 (42.5)	116 (42.5)
	Moderate exercise	72 (21.1)	61 (22.3)
Baseline CSE score, mean (SD)		25.1 (7.3)	25.1 (7.3)

[a]CSE: cardiac self-efficacy.

[b]CVD: cardiovascular disease.

Relationships Between Baseline CSE Scores and Participant Characteristics

Next, we examined relationships between baseline participant characteristics and baseline CSE Scale scores for all participants who completed the baseline CSE Scale. Of the demographic characteristics, age was significantly associated with CSE, such that CSE increased with age (n=341, β=.12, SE=0.04, P=.03). There was no statistically significant difference in baseline CSE by sex (t_{338}=0.19, P=.43) or racial group (F_{340}=2.72, P=.07).

Among the health history characteristics, participants with a history of type II diabetes tended to have higher baseline CSE (mean 27.48, SD 6.47, n=52) compared to participants without diabetes history (mean 24.71, SD 7.42, n=289; t_{77}=2.78, P=.003). There was no difference in baseline CSE between participants with a history of high blood pressure and those without (t_{322}=0.91, P=.36) or participants in primary prevention versus secondary prevention (t_{50}=0.57, P=.57). There was also a significant association between BMI and CSE, with lower BMI individuals reporting higher CSE (n=341, β=−.11, SE=0.07, P=.047).

For the health behavior variables, there was no difference in CSE based on tobacco use status (F_{339}=1.61, P=.20) or typical physical activity in leisure time (F_{337}=1.51, P=.22)

Change in CSE From Baseline to Month 2

In testing for changes in the CSE from baseline to month 2, we found a significant increase in the overall sum score on the CSE Scale from baseline to month 2 (Table 2). There were also significant increases in each individual item on the CSE (all P<.001). The individual CSE items with the greatest increases were “confidence in knowing when to call or visit the doctor,”

XSL•FO
RenderX

"confidence in knowing how much physical activity is good for you," and "confidence that you can get regular aerobic exercise."

Table . Changes in CSE[a] score from baseline to month 2 (n=273)[b]. The baseline CSE survey was surfaced to participants on day 2, and the month 2 CSE survey was deployed on day 40.

	Baseline, mean (SD)	Month 2, mean (SD)	Baseline to month 2, raw change	Baseline to month 2, change in %	*t* test (*df*)	*P* value
How confident are you that you know…						
When to call or visit your doctor about your heart disease?	2.7 (1.1)	3.1 (0.9)	+0.5	+17	7.9 (272)	<.001
How to make your doctor understand your concerns about your heart?	2.8 (1.1)	3.1 (0.9)	+0.3	+10.7	5.1 (272)	<.001
How to take your cardiac medications?	3.1 (1.1)	3.4 (0.9)	+0.4	+11.4	5.7 (272)	<.001
How much physical activity is good for you?	2.7 (1.1)	3.2 (0.8)	+0.5	+16.3	7.7 (272)	<.001
How confident are you that you can...						
Maintain your social activities?	2.8 (1)	3.2 (0.8)	+0.4	+13	6.4 (272)	<.001
Maintain your usual activities at home with your family?	3 (1)	3.3 (0.8)	+0.2	+7.7	4.2 (272)	<.001
Maintain your usual activities at work?	2.9 (1.1)	3.2 (0.9)	+0.3	+10.7	4.8 (272)	<.001
Maintain your sexual relationship with your spouse?	2.6 (1.3)	2.9 (1.1)	+0.3	+12.3	4.3 (272)	<.001
Get regular aerobic exercise?	2.5 (1.2)	2.9 (1.1)	+0.5	+19	6.4 (272)	<.001
Overall sum score on the CSE	25.07 (7.3)	28.3 (6.3)	+3.2	+12.9	9.2 (272)	<.001

[a]CSE: cardiac self-efficacy.

[b]n=273 participants submitted the cardiac self-efficacy with all items completed.

Discussion

Summary and Interpretation of Results

These analyses examined improvement in CSE as a secondary goal in a digital lifestyle modification program for heart health focused on AI-supported health coaching, educational lessons, behavioral tracking, and weight management. We found that the overall CSE Scale score improved significantly from baseline to month 2 of the program. Moreover, mean scores on each individual item of the CSE increased significantly over this short period, with the largest increases in patients' confidence in knowing when to contact their doctor about their heart health (+17%) and items related to confidence in knowing how much physical activity to do (+16.3%) and getting regular exercise (+19%). The results of these secondary analyses of the Heart Health program pilot study support the assertion that digital health programs can play an important role in increasing self-efficacy to better manage their health over a relatively short period.

This study is the first digital health study we are aware of to show improvements in CSE scores through participation in a digital lifestyle modification program for heart health. There is evidence that in-person lifestyle change programs for CVD can improve CSE [25,26], and literature showing that digital interventions can increase general self-efficacy [34,35]. These results are consistent with results from 2 prior studies with cardiac patients showing that in-person intervention programs

XSL•FO
RenderX

for CVD management can improve CSE scores; however, these in-person programs involved intensive, human-led interventions that required significant staffing and economic resources [25,26].

One recent study using a 60-day text message intervention implemented in patients recently discharged after acute coronary syndrome had no significant effect on the CSE [42]. The researchers implemented an automated 1-way text message campaign through which patients received messages about self-management, healthy living, and follow-up care. There are many key differences in the study design by Ross et al [42] compared to our study, but it is possible that we found improvements in CSE because the Heart Health program involves 2-way interaction with AI-supported coaching and active behavior tracking, rather than a 1-way messaging system.

It is also notable that 2 of the CSE items with the biggest improvement from baseline to month 2 were centered on confidence related to physical activity: "confidence in knowing how much physical activity is good for you" (+16%), and "confidence that you can get regular aerobic exercise" (+19%). Education on physical activity is a key component of the Heart Health program, with an educational lesson on Physical Activity for Low Blood Pressure delivered during the third weekly lesson in the program. The Heart Health digital coach also provides ongoing interactive coaching on physical activity, encouraging participants to log their physical activity in the app. As such, the present results suggest that the education and coaching in the first 2 months of the Heart Health program support improvements in confidence related to physical activity. Similarly, prior work has shown that digital health interventions focused on physical activity can improve exercise-related self-efficacy over 24 weeks [43]. Future research in this area could examine whether improvements in CSE Scale score during the first 2 months of the program predict longitudinal increases in physical activity.

Beyond the primary results of these analyses, we also found several relationships between baseline CSE score and participant characteristics. Specifically, we found that older age, lower BMI, and having a history of type II diabetes were all associated with having higher CSE scores at baseline. Importantly, a significant body of literature has shown that higher self-efficacy among older persons is associated with a range of positive health benefits, including increased self-care, better health behaviors, increased energy, and decreased pain and discomfort [44,45]. This is particularly relevant to our study, as participants in this sample tended to be older adults, with a mean age of 60 years and up to 76 years. Notably, the broader literature indicates that self-efficacy and closely related constructs, such as self-esteem, tend to increase with age due to increasing knowledge and maturity, greater control over life circumstances, higher likelihood of occupying positions of power and status, and improved coping resources and processes [46]. Indeed, a large meta-analysis of 164,868 participants showed that self-esteem tends to increase steadily with age, reaching peak levels between ages 60 - 70 before declining slightly after age 70 and steeply after age 90 [47].

Our finding that lower BMI was associated with high CSE scores aligns with prior work showing that cardiac patients with lower BMI also tended to have higher CSE [48] and with population-based research showing lower health-related self-efficacy among individuals with higher BMI [49]. This relationship may be due to individuals with higher BMI having a history of difficulty managing their weight and lower self-esteem [50], factors that are closely related to self-efficacy.

The link between the history of type II diabetes and baseline CSE is also consistent with prior literature showing that patients with a history of chronic disease diagnosis tend to have higher CSE [48,51]. This relationship is likely due to these individuals already having some experience with managing a health condition and potentially receiving prior education on managing cardiometabolic conditions. However, it is surprising that a history of high blood pressure and history of cardiac events or CVD diagnoses were not associated with differences in baseline CSE in this study's sample.

Limitations and Future Directions

There are several limitations to our results. First, data for these secondary analyses came from an observational pilot study focused on the acceptability and feasibility of a new digital heart health program. As such, these results should be considered preliminary, rather than conclusive. We were also unable to compare these results to a control group, and causal conclusions are limited. Additionally, participants could opt out of any survey, including the CSE survey, if they did not wish to complete it. As a result, we could only include individuals in these analyses if they submitted a complete CSE Scale. Completion of the CSE Scale also occurred outside of the Lark app to be used for pilot study purposes only. Thus, the sample may be biased toward participants who were more engaged in the program or more likely to check their emails. The CSE Scale is also limited as it includes questions about confidence in work settings and in marital life, which may not apply to all participants. Finally, this digital intervention was limited to those with a smartphone with internet access who had sufficient digital literacy to use the Lark app. As a result, this may exclude individuals of lower socioeconomic status. For instance, the 2024 Pew Research Center data showed that, although 90% of US adults have a smartphone, individuals in the lowest income bracket (US <$30,000/year) have lower rates of smartphone ownership (79%) compared to those in higher income brackets (US >$100,000/year; 98%) [52].

The pilot study results presented here provide a foundation for several lines of future work. For instance, future work including a control group can provide causal evidence for the impact of the Heart Health program on CSE. Additionally, a key future direction for this study will be assessing whether increased CSE predicts longitudinal clinical outcomes. Given that the duration of the pilot study was only a short period of 90 days, we did not expect to see a meaningful relationship between improved CSE Scale scores and clinical outcomes. This is particularly true given that we measured CSE improvements nearly halfway through the program. Future research could examine longitudinal changes in CVD risk factors, such as blood pressure or cholesterol levels. Based on the literature, we would expect that improvements in CSE would predict improved cardiovascular outcomes.

XSL•FO
RenderX

Conclusions

In summary, the present analyses indicate that participants in a digital lifestyle modification program for heart health showed significant improvements in CSE within 2 months. This work adds to the growing literature examining ways to improve health-related self-efficacy for the prevention and management of chronic disease and long-term health benefits.

Acknowledgments

The Heart Health pilot study was supported by funds from Roche Information Solutions and Lark Health. KGL, OHB, and SAG received salaries from Lark Health. PRK received a salary from Roche Information Solutions.

Data Availability

The dataset analyzed during this study is not publicly available as these analyses include sensitive, potentially identifiable information that was obtained through a collaboration between Lark Health and its health care partners, but data may be available from the corresponding author upon reasonable request.

Authors' Contributions

All authors contributed to the conception of the work and made critical revisions to the content of this paper. KGL contributed to data collection and management, led data analysis, and drafted and revised this paper. PRK and SAG provided critical reviews and interpretation of all statistical analyses. OHB was the principal investigator of the pilot study, supervised data collection and management, and supported analyses and interpretation of the data.

Conflicts of Interest

KGL, OHB, and SAG were employees of Lark Health at the time of manuscript development. PRK is an employee of Roche Information Solutions.

References

1. Tsao CW, Aday AW, Almarzooq ZI, et al. Heart disease and stroke statistics-2022 update: a report from the American Heart Association. Circulation 2022 Feb 22;145(8):e153-e639. [doi: 10.1161/CIR.0000000000001052] [Medline: 35078371]
2. McClellan M, Brown N, Califf RM, Warner JJ. Call to action: urgent challenges in cardiovascular disease: a presidential advisory from the American Heart Association. Circulation 2019 Feb 26;139(9):e44-e54. [doi: 10.1161/CIR.0000000000000652] [Medline: 30674212]
3. Rippe JM. Lifestyle strategies for risk factor reduction, prevention, and treatment of cardiovascular disease. Am J Lifestyle Med 2019;13(2):204-212. [doi: 10.1177/1559827618812395] [Medline: 30800027]
4. Arnett DK, Blumenthal RS, Albert MA, et al. 2019 ACC/AHA guideline on the primary prevention of cardiovascular disease: a report of the American College of Cardiology/American Heart Association task force on clinical practice guidelines. Circulation 2019 Sep 10;140(11):e596-e646. [doi: 10.1161/CIR.0000000000000678] [Medline: 30879355]
5. Kris-Etherton PM, Petersen KS, Velarde G, et al. Barriers, opportunities, and challenges in addressing disparities in diet-related cardiovascular disease in the United States. J Am Heart Assoc 2020 Apr 7;9(7):e014433. [doi: 10.1161/JAHA.119.014433] [Medline: 32200727]
6. US Preventive Services Task Force, Krist AH, Davidson KW, et al. Behavioral counseling interventions to promote a healthy diet and physical activity for cardiovascular disease prevention in adults with cardiovascular risk factors: US Preventive Services Task Force recommendation statement. JAMA 2020 Nov 24;324(20):2069-2075. [doi: 10.1001/jama.2020.21749] [Medline: 33231670]
7. Patnode CD, Redmond N, Iacocca MO, Henninger M. Behavioral counseling interventions to promote a healthy diet and physical activity for cardiovascular disease prevention in adults without known cardiovascular disease risk factors: updated evidence report and systematic review for the US Preventive Services Task Force. JAMA 2022 Jul 26;328(4):375-388. [doi: 10.1001/jama.2022.7408] [Medline: 35881116]
8. US Preventive Services Task Force, Mangione CM, Barry MJ, et al. Behavioral counseling interventions to promote a healthy diet and physical activity for cardiovascular disease prevention in adults without cardiovascular disease risk factors: US Preventive Services Task Force recommendation statement. JAMA 2022 Jul 26;328(4):367-374. [doi: 10.1001/jama.2022.10951] [Medline: 35881115]
9. Foroumandi E, Kheirouri S, Alizadeh M. The potency of education programs for management of blood pressure through increasing self-efficacy of hypertensive patients: a systematic review and meta-analysis. Patient Educ Couns 2020 Mar;103(3):451-461. [doi: 10.1016/j.pec.2019.09.018] [Medline: 31558325]
10. Katch H, Mead H. The role of self-efficacy in cardiovascular disease self-management: a review of effective programs. PI 2010;2010(2):33-44. [doi: 10.2147/PI.S12624]

XSL•FO
RenderX

11. Brady TJ, Murphy L, O'Colmain BJ, et al. A meta-analysis of health status, health behaviors, and health care utilization outcomes of the Chronic Disease Self-Management Program. Prev Chronic Dis 2013;10:120112. [doi: 10.5888/pcd10.120112] [Medline: 23327828]
12. Lorig K, Stewart A, Ritter P, González V, et al. Outcome Measures for Health Education and Other Health Care Interventions: Sage Publications, Inc; 1996.
13. Lorig KR, Sobel DS, Ritter PL, Laurent D, Hobbs M. Effect of a self-management program on patients with chronic disease. Eff Clin Pract 2001;4(6):256-262. [Medline: 11769298]
14. Sullivan MD, LaCroix AZ, Russo J, Katon WJ. Self-efficacy and self-reported functional status in coronary heart disease: a six-month prospective study. Psychosom Med 1998;60(4):473-478. [doi: 10.1097/00006842-199807000-00014] [Medline: 9710293]
15. Fors A, Ulin K, Cliffordson C, Ekman I, Brink E. The Cardiac Self-Efficacy Scale, a useful tool with potential to evaluate person-centred care. Eur J Cardiovasc Nurs 2015 Dec;14(6):536-543. [doi: 10.1177/1474515114548622] [Medline: 25149667]
16. Ea EE, Colbert A, Turk M, Dickson VV. Self-care among Filipinos in the United States who have hypertension. Appl Nurs Res 2018 Feb;39:71-76. [doi: 10.1016/j.apnr.2017.11.002] [Medline: 29422180]
17. Tan F, Oka P, Dambha-Miller H, Tan NC. The association between self-efficacy and self-care in essential hypertension: a systematic review. BMC Fam Pract 2021 Feb 22;22(1):44. [doi: 10.1186/s12875-021-01391-2] [Medline: 33618661]
18. Garland M, Wilbur J, Fogg L, Halloway S, Braun L, Miller A. Self-efficacy, outcome expectations, group social support, and adherence to physical activity in African American women. Nurs Res 2021;70(4):239-247. [doi: 10.1097/NNR.0000000000000516] [Medline: 33870956]
19. Kang A, Dulin A, Risica PM. Relationship between adherence to diet and physical activity guidelines and self-efficacy among black women with high blood pressure. J Health Psychol 2022 Mar;27(3):663-673. [doi: 10.1177/1359105320967105] [Medline: 33081506]
20. Warren-Findlow J, Seymour RB. Prevalence rates of hypertension self-care activities among African Americans. J Natl Med Assoc 2011 Jun;103(6):503-512. [doi: 10.1016/s0027-9684(15)30365-5] [Medline: 21830634]
21. Jackson L, Leclerc J, Erskine Y, Linden W. Getting the most out of cardiac rehabilitation: a review of referral and adherence predictors. Heart 2005 Jan;91(1):10-14. [doi: 10.1136/hrt.2004.045559] [Medline: 15604322]
22. Banik A, Schwarzer R, Knoll N, Czekierda K, Luszczynska A. Self-efficacy and quality of life among people with cardiovascular diseases: a meta-analysis. Rehabil Psychol 2018 May;63(2):295-312. [doi: 10.1037/rep0000199] [Medline: 29878834]
23. Lorig KR, Ritter PL, González VM. Hispanic chronic disease self-management: a randomized community-based outcome trial. Nurs Res 2003;52(6):361-369. [doi: 10.1097/00006199-200311000-00003] [Medline: 14639082]
24. DeWalt DA, Pignone M, Malone R, et al. Development and pilot testing of a disease management program for low literacy patients with heart failure. Patient Educ Couns 2004 Oct;55(1):78-86. [doi: 10.1016/j.pec.2003.06.002] [Medline: 15476993]
25. Bagheri H, Shakeri S, Nazari AM, et al. Effectiveness of nurse-led counselling and education on self-efficacy of patients with acute coronary syndrome: a randomized controlled trial. Nurs Open 2022 Jan;9(1):775-784. [doi: 10.1002/nop2.1129] [Medline: 34766453]
26. McKenzie KM, Park LK, Lenze EJ, et al. A prospective cohort study of the impact of outpatient intensive cardiac rehabilitation on depression and cardiac self-efficacy. Am Heart J Plus 2022 Jan;13:100100. [doi: 10.1016/j.ahjo.2022.100100] [Medline: 36407054]
27. Kang Y, Yang IS, Kim N. Correlates of health behaviors in patients with coronary artery disease. Asian Nurs Res (Korean Soc Nurs Sci) 2010 Mar;4(1):45-55. [doi: 10.1016/S1976-1317(10)60005-9] [Medline: 25030792]
28. Sarkar U, Ali S, Whooley MA. Self-efficacy as a marker of cardiac function and predictor of heart failure hospitalization and mortality in patients with stable coronary heart disease: findings from the Heart and Soul Study. Health Psychol 2009 Mar;28(2):166-173. [doi: 10.1037/a0013146] [Medline: 19290708]
29. Mahajan S, Lu Y, Spatz ES, Nasir K, Krumholz HM. Trends and predictors of use of digital health technology in the United States. Am J Med 2021 Jan;134(1):129-134. [doi: 10.1016/j.amjmed.2020.06.033] [Medline: 32717188]
30. Wongvibulsin S, Habeos EE, Huynh PP, et al. Digital health interventions for cardiac rehabilitation: systematic literature review. J Med Internet Res 2021 Feb 8;23(2):e18773. [doi: 10.2196/18773] [Medline: 33555259]
31. Zwack CC, Haghani M, Hollings M, et al. The evolution of digital health technologies in cardiovascular disease research. npj Digit Med 2023 Jan 3;6(1):1-11. [doi: 10.1038/s41746-022-00734-2]
32. Frederix I, Caiani EG, Dendale P, et al. ESC e-Cardiology Working Group position paper: overcoming challenges in digital health implementation in cardiovascular medicine. Eur J Prev Cardiol 2019 Jul;26(11):1166-1177. [doi: 10.1177/2047487319832394] [Medline: 30917695]
33. Gray R, Indraratna P, Lovell N, Ooi SY. Digital health technology in the prevention of heart failure and coronary artery disease. Cardiovasc Digit Health J 2022 Dec;3(6 Suppl):S9-S16. [doi: 10.1016/j.cvdhj.2022.09.002] [Medline: 36589760]
34. Devi R, Powell J, Singh S. A web-based program improves physical activity outcomes in a primary care angina population: randomized controlled trial. J Med Internet Res 2014 Sep 12;16(9):e186. [doi: 10.2196/jmir.3340] [Medline: 25217464]

XSL•FO
RenderX

35. Newby K, Teah G, Cooke R, et al. Do automated digital health behaviour change interventions have a positive effect on self-efficacy? A systematic review and meta-analysis. Health Psychol Rev 2021 Mar;15(1):140-158. [doi: 10.1080/17437199.2019.1705873] [Medline: 31847702]
36. Lockwood KG, Kulkarni PR, Paruthi J, et al. Evaluating a new digital app-based program for heart health: feasibility and acceptability pilot study. JMIR Form Res 2024 May 24;8:e50446. [doi: 10.2196/50446] [Medline: 38787598]
37. Branch OH, Rikhy M, Auster-Gussman LA, Lockwood KG, Graham SA. Relationships between blood pressure reduction, weight loss, and engagement in a digital app-based hypertension care program: observational study. JMIR Form Res 2022 Oct 27;6(10):e38215. [doi: 10.2196/38215] [Medline: 36301618]
38. Graham SA, Pitter V, Hori JH, Stein N, Branch OH. Weight loss in a digital app-based diabetes prevention program powered by artificial intelligence. Digit Health 2022;8:20552076221130619. [doi: 10.1177/20552076221130619] [Medline: 36238752]
39. Lockwood KG, Pitter V, Kulkarni PR, Graham SA, Auster-Gussman LA, Branch OH. Predictors of program interest in a digital health pilot study for heart health. PLOS Digit Health 2023 Jul 31;2(7):e0000303. [doi: 10.1371/journal.pdig.0000303]
40. McGorrian C, Yusuf S, Islam S, et al. Estimating modifiable coronary heart disease risk in multiple regions of the world: the INTERHEART Modifiable Risk Score. Eur Heart J 2011 Mar;32(5):581-589. [doi: 10.1093/eurheartj/ehq448] [Medline: 21177699]
41. Yusuf S, Rangarajan S, Teo K, et al. Cardiovascular risk and events in 17 low-, middle-, and high-income countries. N Engl J Med 2014 Aug 28;371(9):818-827. [doi: 10.1056/NEJMoa1311890] [Medline: 25162888]
42. Ross ES, Sakakibara BM, Mackay MH, et al. The use of SMS text messaging to improve the hospital-to-community transition in patients with acute coronary syndrome (Txt2Prevent): results from a pilot randomized controlled trial. JMIR mHealth uHealth 2021 May 14;9(5):e24530. [doi: 10.2196/24530] [Medline: 33988519]
43. Maddison R, Pfaeffli L, Whittaker R, et al. A mobile phone intervention increases physical activity in people with cardiovascular disease: results from the HEART randomized controlled trial. Eur J Prev Cardiolog 2015 Jun;22(6):701-709. [doi: 10.1177/2047487314535076]
44. Kostka T, Jachimowicz V. Relationship of quality of life to dispositional optimism, health locus of control and self-efficacy in older subjects living in different environments. Qual Life Res 2010 Apr;19(3):351-361. [doi: 10.1007/s11136-010-9601-0] [Medline: 20146007]
45. Anderson ES, Winett RA, Wojcik JR, Williams DM. Social cognitive mediators of change in a group randomized nutrition and physical activity intervention: social support, self-efficacy, outcome expectations and self-regulation in the guide-to-health trial. J Health Psychol 2010 Jan;15(1):21-32. [doi: 10.1177/1359105309342297] [Medline: 20064881]
46. Orth U, Trzesniewski KH, Robins RW. Self-esteem development from young adulthood to old age: a cohort-sequential longitudinal study. J Pers Soc Psychol 2010 Apr;98(4):645-658. [doi: 10.1037/a0018769] [Medline: 20307135]
47. Orth U, Erol RY, Luciano EC. Development of self-esteem from age 4 to 94 years: a meta-analysis of longitudinal studies. Psychol Bull 2018 Oct;144(10):1045-1080. [doi: 10.1037/bul0000161] [Medline: 30010349]
48. Kang Y, Yang IS. Cardiac self-efficacy and its predictors in patients with coronary artery diseases. J Clin Nurs 2013 Sep;22(17-18):2465-2473. [doi: 10.1111/jocn.12142] [Medline: 23441807]
49. Ovaskainen ML, Tapanainen H, Laatikainen T, Männistö S, Heinonen H, Vartiainen E. Perceived health-related self-efficacy associated with BMI in adults in a population-based survey. Scand J Public Health 2015 Mar;43(2):197-203. [doi: 10.1177/1403494814566263]
50. Kiviruusu O, Konttinen H, Huurre T, Aro H, Marttunen M, Haukkala A. Self-esteem and body mass index from adolescence to mid-adulthood. a 26-year follow-up. IntJ Behav Med 2016 Jun;23(3):355-363. [doi: 10.1007/s12529-015-9529-4]
51. Shrestha R, Rawal L, Bajracharya R, Ghimire A. Predictors of cardiac self-efficacy among patients diagnosed with coronary artery disease in tertiary hospitals in Nepal. J Public Health Res 2020 Oct 14;9(4):1787. [doi: 10.4081/jphr.2020.1787] [Medline: 33437751]
52. Gelles-Watnick R. Rates of smartphone ownership, broadband subscription vary across groups, including by household income and education. Pew Research Center. 2024. URL: https://www.pewresearch.org/internet/2024/01/31/americans-use-of-mobile-technology-and-home-broadband/ [accessed 2025-04-04]

Abbreviations

AI: artificial intelligence
CSE: cardiac self-efficacy
CVD: cardiovascular disease
HIPAA: Health Insurance Portability and Accountability Act
HITRUST: Health Information Trust Alliance
SOC 2: System and Organization Controls 2

XSL•FO
RenderX

Edited by A Mavragani; submitted 17.05.24; peer-reviewed by A Woodward, K Wac; revised version received 28.02.25; accepted 02.03.25; published 24.04.25.

Please cite as:
Lockwood KG, Kulkarni PR, Branch OH, Graham SA
Cardiac Self-Efficacy Improvement in a Digital Heart Health Program: Secondary Analysis From a Feasibility and Acceptability Pilot Study
JMIR Form Res 2025;9:e60676
URL: https://formative.jmir.org/2025/1/e60676
doi:10.2196/60676

© Kimberly G Lockwood, Priya R Kulkarni, OraLee H Branch, Sarah A Graham. Originally published in JMIR Formative Research (https://formative.jmir.org), 24.4.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Healing Through Ongoing Psychological Empowerment Telehealth Intervention With Two Spirit, Transgender, and Nonbinary Clients of Color in the United States: Open Clinical Trial Feasibility and Implementation Analysis

Stephanie Lynne Budge[1], PhD; Elliot Aaron Tebbe[2], PhD; Joonwoo Lee[2], MEd; Sergio Domínguez Jr[1], PhD; Em Matsuno[3], PhD; Louis Lindley[2], MA

[1]Department of Counseling Psychology, University of Wisconsin-Madison, 1000 Bascom Mall, 305 Education Building, Madison, WI, United States
[2]University of Wisconsin-Madison, Madison, WI, United States
[3]School of Counseling and Counseling Psychology, Arizona State University, Tempe, AZ, United States

Corresponding Author:
Stephanie Lynne Budge, PhD
Department of Counseling Psychology, University of Wisconsin-Madison, 1000 Bascom Mall, 305 Education Building, Madison, WI, United States

Abstract

Background: There is a notable lack of psychotherapeutic services tailored to the needs of Two Spirit, transgender, and nonbinary (2STNB) people of color; research indicates that 2STNB clients who are people of color report a lack of competence and cultural humility on the part of their therapists.

Objective: The purpose of this study was to report the feasibility and acceptability of the Healing Through Ongoing Psychological Empowerment (HOPE) teletherapy intervention using deductive content analysis.

Methods: We used an open clinical trial design (testing one intervention without a comparison group) to test the feasibility and acceptability of the HOPE intervention. At baseline, 51 clients were enrolled in the open clinical trial, with 49 2STNB clients who are people of color starting and completing the HOPE intervention. Clients were recruited primarily from social media and therapist waitlists. Clients completed up to 15 free face-to-face telehealth psychotherapy sessions that were provided by nine 2STNB therapists who are people of color. Feasibility and acceptability interviews were conducted prior to the intervention, immediately following the intervention, and at 6 months after completing the intervention.

Results: The HOPE intervention demonstrated high feasibility and acceptability, specifically regarding data collection, psychometric adequacy, interventionist recruitment or training or retention, delivery of the intervention, acceptability of the intervention to clients, and client engagement with the intervention.

Conclusions: These findings propose HOPE as a potentially feasible, culturally specific therapeutic approach for the 2STNB community who are people of color. Future randomized controlled trials comparing HOPE to existing evidence-based treatments are needed.

Trial Registration: ClinicalTrials.gov NCT05140174; https://clinicaltrials.gov/study/NCT05140174

(JMIR Form Res 2025;9:e64477) doi:10.2196/64477

KEYWORDS

Black, Indigenous, people of color; transgender; nonbinary; radical healing; internalized transnegativity; open clinical trial; psychotherapeutic; lack of competence; cultural; humility; psychotherapy; therapist; HOPE; Healing Through Ongoing Psychological Empowerment; gender identity; intervention; content analysis; treatment; medication; mental health

Introduction

Psychotherapy serves as an effective method for addressing mental health concerns for Two Spirit, transgender, and nonbinary (2STNB) clients who are people of color. However, 2STNB clients who are people of color experience numerous barriers to accessing competent and affirming psychotherapy services [1], with clients reporting various types of negative experiences with mental health providers, including gender- and race-based microaggressions [2], a general lack of competency [3,4], as well as gatekeeping dynamics when seeking access to medical gender-affirming care [5]. Participants in prior research have described how these negative experiences have undermined the therapeutic relationship, exacerbated

XSL•FO
RenderX

distress, and at times led to early termination of therapy [2,6,7]. While much of the existing research has documented these negative experiences in therapy, some studies have found that when barriers to accessing mental health care and oppression are addressed, and clients receive competent and affirming care, psychotherapy effectively reduces mental health symptoms and promotes healing [8,9]. In a recent psychotherapy randomized controlled trial with 19 transgender and nonbinary adults, researchers found that transgender and nonbinary adults readily participated in therapy when barriers to access were reduced (eg, financial barriers), and when therapists were trained to address the harmful impact of identity-based stigma and related stressors [8]. Although the study demonstrated that therapy improved clients' psychological well-being, and participants reported a strong working alliance with their therapist and overall satisfaction with the therapy, many also reported a desire to work with a transgender or nonbinary therapist. This request could not be accommodated in a trial that only included cisgender therapists.

In their review of race and gender identity dynamics between therapists and clients, Erby and White emphasized [6] the necessity of using an intersectional lens to address layers of systematic oppression within the therapeutic dynamic. Dominguez et al [10] highlighted the importance of using the radical healing framework in therapy to address the racial-gendered experiences of 2STNB clients who are people of color. The authors also stress the need to focus on unpacking the internalization of racial-gendered stigma. The psychological framework for radical healing (PRH) incorporates elements such as critical consciousness, cultural authenticity and self-knowledge, collectivism, radical hope, and cultivating strength and resistance for clients who are people of color [11,12]. The PRH specifically notes that the term "radical" derives from collective resistance that was deemed radical by those who were engaged in oppressive practices. Adames et al [11] defined radical as "a critical attitude or ideology that promotes the idea that complete change is necessary to reduce social problems"; therefore, radical healing in the psychotherapy context involves supporting clients in realizing their full potential in a context where they have experienced gendered and racialized oppression. While the PRH is robust in its conceptualization of healing for people of color, further clarity regarding gendered dimensions of healing could benefit clinicians. In contrast, Israel et al [13] developed an intervention to reduce internalized stigma for transgender and nonbinary people, but this approach lacked a racialized perspective on stigma. The literature clearly indicates a need for an approach that integrates intersectional understandings of both race and gender for 2STNB clients who are people of color.

In this study, we used community-based participatory research (CBPR) methods to design a psychotherapy study aimed to address some of the structural barriers that many 2STNB adults who are people of color face when seeking mental health care. Our CBPR process began at the study's conception and has continued through its completion and the dissemination of study findings, both in the scholarly literature and within the community. This approach is particularly important against the backdrop of historical and contemporary harms inflicted on 2STNB and people of color by both researchers and mental health professionals [14,15]. CBPR methods can effectively increase transparency of study goals and methods, improve study design to maximize participant experience, and build trust between researchers and community members [16]. Guided by our CBPR framework and processes, this psychotherapy trial implemented four main strategies for providing psychotherapy to 2STNB clients who are people of color: (1) we employed exclusively 2STNB therapists who are people of color to provide psychotherapy to 2STNB clients who are people of color; (2) all therapy sessions were provided at no cost to clients, reducing socioeconomic barriers; (3) by using telehealth, we reduced geographic barriers; and (4) all therapists received specialized training prior to beginning the study to increase their competence in effectively addressing the mental health needs of 2STNB clients who are people of color.

Our main research question for this study included: do results from the open clinical trial demonstrate the feasibility and acceptability of the Healing Through Ongoing Psychological Empowerment (HOPE) intervention? We used the framework of Teresi et al [17] to evaluate the feasibility and acceptability of our trial design and intervention implementation [17]. Specifically, we evaluated whether participants could adequately engage with the data collection protocols; their perceptions of the measures and psychometric adequacy of the measures; whether therapists could be recruited, trained, and retained; whether therapists provided the therapy as intended; and the overall acceptability of therapy to the clients. Our analytic approach for determining feasibility and acceptability is detailed below.

Methods

Participants

For this study, our target sample size for the open clinical trial design was 50 participants, each paired with one of the 10 therapists hired for the study (with therapists seeing 5 clients each). Following a screening call for study eligibility, 51 clients were enrolled, and 49 began and completed the psychotherapy intervention. The sample size was determined based on pilot clinical sample size recommendations [18]. All participants were located in the United States. The mean age of participants was 28.35 (SD 6.38; range 21 - 52) years. Participants identified their gender, race, and ethnicity using both forced-choice and open-response formats. The open-ended responses showed greater variability and multiple identities, which were used to contextualize participant quotes. Regarding responses to the forced-choice format, 43% (n=22) of the sample were Black or African American, 33% (n=17) were Asian or Asian American, 28% (n=14) were Latine or Latinx or Hispanic (non-White), 12% (n=6) were American Indian or Native American or Indigenous, and 2% (n=1) were Native Hawaiian or Pacific Islander (note: participants could note more than 1 identity; 14% (n=7) also listed a White identity alongside a people of color identity). Regarding forced-choice gender, most clients noted a nonbinary identity (n=39, 77%), while 13% (n=7) identified as transmen, and 10% (n=5) identified as transwomen..

XSL•FO
RenderX

Therapists

The training provided to therapists prior to participants enrolling in the study integrated approaches for addressing internalized transnegativity [13,19,20] into the PRH [11,12]. The design, development, and implementation of this training are described in greater detail elsewhere [21]. In total, 9 of the 10 intended outpatient community therapists were trained for this clinical trial, as 1 was unable to attend the required training due to health issues. Therapists' racial or ethnic identities were Asian or Asian American (n=2, 22%), Black or African American (n=3, 33%), Latine or Latinx or Hispanic (n=1, 11%), and multiracial (Indigenous or Latine, White or Latine, and Black or Latine; n=3, 33%). Therapists most often endorsed a nonbinary identity (n=6, 67%), and 3 (33%) identified as women. Therapists were licensed to provide therapy in all but 10 states throughout the United States. More detailed information on therapists' backgrounds and demographic information are presented in Multimedia Appendix 1.

Procedure

Procedures of this study were developed using a CBPR model. Specifically, a community advisory board (CAB) was established, consisting of 2STNB community members who are people of color and representatives from organizations serving 2STNB community members who are people of color. The CAB was integrally involved at each stage, providing critical feedback and guidance on various aspects of the study, including participant and therapist recruitment and retention, survey measures and interview protocols, and therapist training. This collaborative process was crucial for shaping the study design to maximize feasibility and acceptability for both 2STNB clients and therapists who are people of color. Furthermore, maintaining transparency throughout the process and ensuring open communication with both clients and therapists helped mitigate mistrust and uncertainty.

Participants were recruited within the United States and lived throughout 8 of 9 census regions (the only region not represented was East South Central [Mississippi, Alabama, Tennessee, and Kentucky]). Clients were recruited via study therapists' waitlists, social media, community organization listservs, and referrals from other community therapists in states where the study therapists were licensed. The inclusion criteria for clients were (1) being 18 years or older, (2) identifying as 2STNB or having sex assigned at birth different from their current gender, (3) fluency in English or Spanish, and (4) living in a US state where the therapists were licensed. Exclusion criteria included (1) the presence of psychiatric symptoms requiring inpatient treatment (eg, active psychosis) and (2) ongoing individual psychotherapy outside of the study. We kept the inclusion and exclusion criteria minimal to enhance the generalizability of our findings.

Once clients were screened and determined eligible for the study, they scheduled a 2-hour baseline session with study staff via a web-based platform. During this session, study procedures were discussed, informed consent was obtained, and a qualitative interview and baseline surveys were completed. Clients could choose to engage in data collection sessions either in Spanish or English. Additionally, if clients were matched with a bilingual therapist, they could choose to engage in psychotherapy in either Spanish or English. Throughout the study, participants completed qualitative interviews and surveys at 3 time points: prior to starting therapy, immediately following the 15-session limit, and 6 months after the trial ended. These time points were chosen to evaluate initial engagement, immediate outcomes, and longer-term effects of the intervention, providing comprehensive data on the feasibility and acceptability of the study design and therapeutic approach. As part of our CBPR process, we initially consulted with our CAB. The CAB recommended that we present all measures and qualitative questions to focus groups of 2STNB clients who are people of color to determine if items or questions are acceptable, appropriate, and relevant. We conducted 5 focus group meetings, during which we received important feedback on our qualitative interview questions. This feedback was incorporated into our interview protocol.

Training for Therapists

The HOPE intervention was developed to integrate PRH in psychotherapy for 2STNB clients who are people of color. Our intervention is grounded in the tenets of the PRH [11,12], and a web-based intervention aimed at reducing internalized transnegativity [13]. Initially, the training developed for this study was piloted with 100 therapists through a program evaluation study [21]. Based on feedback from this pilot, the training was adapted for the 9 therapists for this study. All therapists received 16 hours of training and participated in 12 monthly group supervisions throughout the trial. The 2 consultation group leaders were also 2STNB people of color.

Intervention

Detailed information about the intervention is included in the paper that demonstrates the pilot testing of the training [21]. The intervention incorporates the main tenants of the PRH [11,12] and also from a previously tested internalized transnegativity intervention [13]. The primary components of the PRH that were included in the intervention include critical consciousness, strength and resistance, emotional and social support, cultural authenticity and self-knowledge, and radical hope. The primary components of the internalized transnegativity intervention include opposing stereotypes, recognizing and rebuffing negative messages, strengthening the dismissal of negative messages, and improving identity affirmation. As this was not a manualized treatment, there is not an intervention manual.

Ethical Considerations

This study was approved by the Institutional Review Board (2021 - 1133) at the University of Wisconsin-Madison and was also registered at ClinicalTrials.gov (NCT05140174). Clients provided verbal consent and also electronically signed consent forms that were approved by the institutional review board. As part of our process of protecting the privacy of clients, all qualitative interviews were uploaded to a secure server and transcribed by a team member who deidentified all transcripts. All transcripts were stored on a secure drive and were coded using the secure drive. Clients were compensated US $100 at baseline and postintervention and US $125 at follow-up for their time spent filling out measures and engaging in the

XSL•FO
RenderX

interviews. Recruitment began in January 2022. Participants were offered 15 sessions of therapy free of charge, which could be scheduled flexibly (eg, weekly and biweekly) until December 31, 2022.

Instrument: Feasibility and Acceptability Questions

Clients were asked a series of questions at baseline, postintervention, and 6-month follow-up. These questions covered topics such as feasibility, acceptability, previous therapy experiences, therapy goals within the study, perceptions of the psychotherapy interventions for the study, and experiences of change from therapy. Each interview lasted approximately 60 minutes at each time point. For this paper, we will focus solely on the interview questions related to feasibility or acceptability. At baseline, participants were asked how they heard about the study, why they wanted to participate, any factors contributing to uncertainties about participation, and anticipated barriers to participation. At postintervention and follow-up, participants were asked about their overall experiences in the study, accessibility of telehealth, ease of scheduling sessions, completing measures, perceptions of study procedures, and their openness to participate in the study again if given the opportunity.

Data Analyses: Qualitative Data Analysis

All qualitative data for the feasibility and acceptability analyses were coded using deductive content analysis following transcription [22]. According to Elo and Kyngäs [22], there are several steps to qualitative data analysis, including the preparation phase, selecting the unit of analysis, and making sense to the data as a whole. In deductive content analysis, the coder develops a structured analysis matrix, codes the data according to categories, and then conducts hypothesis testing. Finally, the coder reports the analysis process and the results, which then leads to the interpretation of the analysis. For this study, the first author (SLB) developed a categorization matrix based on feasibility or acceptability definitions and coded the data accordingly, following the procedures outlined by Elo and Kyngäs [22].

Results

Measure Implementation

We conceptualized feasibility based on the guidelines of Teresi et al [17] for reporting pilot and feasibility studies. The first guideline includes assessing if participants could comply with data collection protocols. Clients who participated at each of the 3 main data collection points (baseline: n=51, postintervention: n=47, and 6-month follow-up: n=44) completed all measures and participated in qualitative interviews, demonstrating the feasibility of the main data collection time points (Table 1 and Figure 1). Additionally, we aimed to generate quantitative data at each of the psychotherapy sessions (ie, the Working Alliance Inventory [WAI-C] [23] and the Outcome Questionnaire-45 [24]). Therapists were instructed to remind participants to complete the presession measures before starting each session; however, not all clients filled out the presession surveys. At the postinterview, 10 of 47 (21%) participants indicated that their most frequent criticism of the study was that clients did not like filling out the surveys prior to the sessions, for example, when Alex (all names provided for quotes are pseudonyms; 29 years, Black, trans nonbinary client) was asked about any changes they would make to study procedures, they said: “Not taking the surveys before (laughs) each session um that was a little difficult to do beforehand especially because the timing that [therapist’s name] and I generally would just kinda roll out of bed and show up.” Frankie (26 years, Indigenous or Black, nonbinary trans client) said at the postinterview about the presession surveys: “I’m glad I don’t have to do that no more!” At 6-month follow-up, the criticism about the surveys was less of an issue, but 5 of 44 (11%) clients did indicate that they would have appreciated more reminders for surveys.

Table . Participant demographic variables prior to starting the Healing Through Ongoing Psychological Empowerment intervention[a].

	Values
Age (years), mean (SD)	28.35 (6.38)
Annual individual income (US $), mean (SD)	11,491 (18,384)
Race or ethnicity, n (%)	
American Indian or Native American or Indigenous	6 (12)
Asian or Asian American	17 (33)
Black or African American	22 (43)
Latine or Latinx or Hispanic	14 (28)
Native Hawaiian or Pacific Islander	1 (2)
Gender, n (%)	
Nonbinary	39 (77)
Transwoman	5 (10)
Transman	7 (13)

[a]The baseline sample included 51 clients. Participants could list more than 1 race or ethnicity; therefore, percentages add up to more than 100%. A total of 7 (14%) participants also noted a White identity along with a people of color identity.

XSL•FO
RenderX

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) diagram: enrollment, allocation, completed treatment, and follow-up assessments for the Healing Through Ongoing Psychological Empowerment intervention.

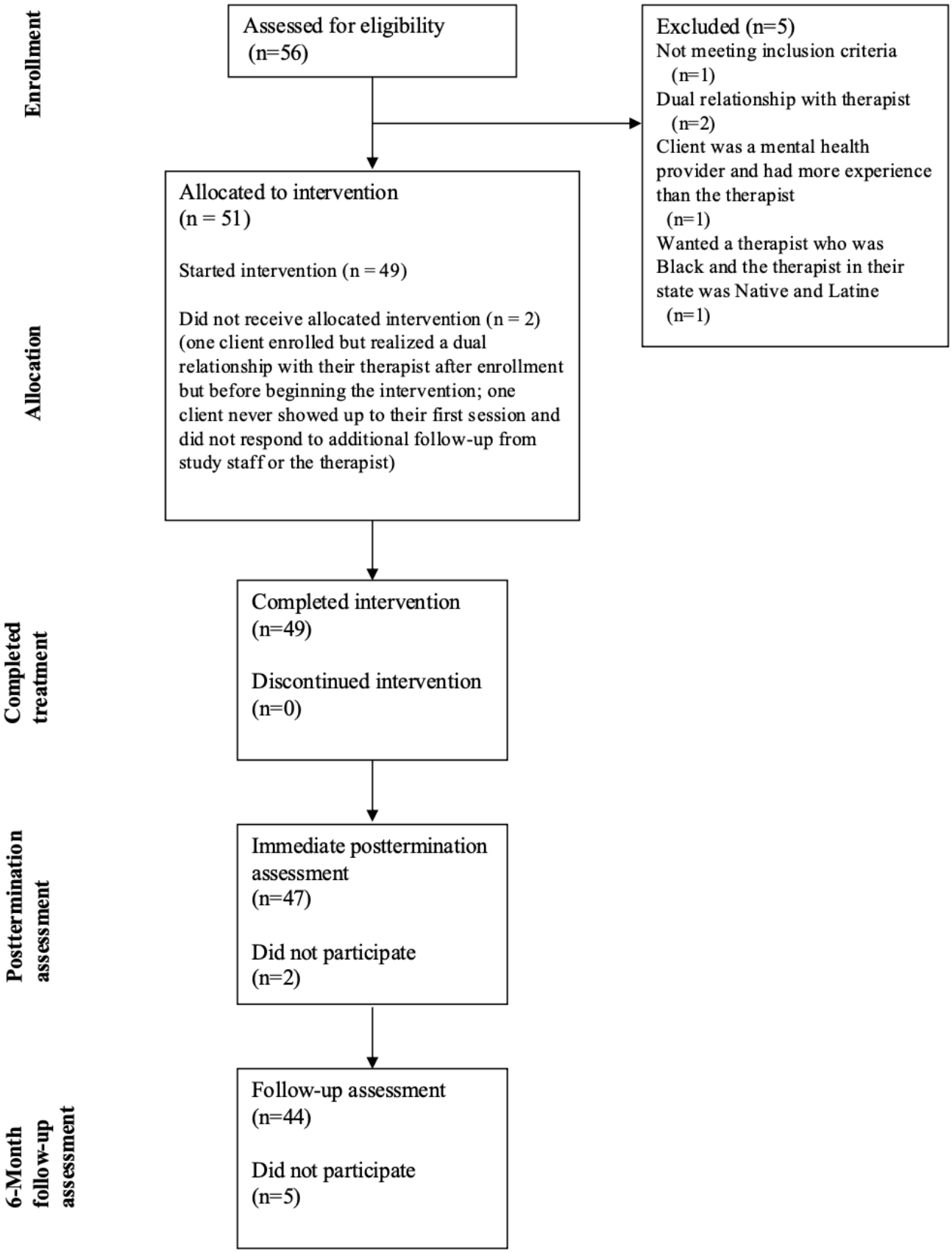

Regarding the conceptual and psychometric adequacy of the measures, we assessed whether the measures were acceptable, appropriate, and relevant to the clients in this study. Following completion of the study, 44 of 47 (94%) participants indicated that the measures were acceptable, appropriate, and relevant in the postinterview and 6-month follow-up interview, although some participants offered feedback common to survey research, such as wanting to answer “in-between” a numerical response,

XSL•FO
RenderX

or provided a more in-depth explanation for some survey items. At the postinterview, 3 of 47 (6%) clients mentioned that they did not like the questions from the WAI-C and found them challenging to fill out.

Intervention Implementation

For intervention implementation, we adhered to the framework established by Teresi et al [17] for assessing feasibility. Teresi et al [17] defined feasibility around 4 key aspects. The first aspect includes whether interventionists can be recruited, trained, and retained. We initially aimed to hire 10 community therapists. However, 1 therapist who was initially hired was unable to participate in the training and was therefore not included as a therapist in the study. The remaining 9 therapists were successfully trained and retained. However, during client recruitment, 1 of the 9 therapists was dealing with illness and grief and requested not to take on additional clients. In response, we increased the number of clients assigned to the other therapists in the study while ensuring equitable distribution and maintaining balance.

The second aspect of intervention implementation feasibility includes whether interventionists can deliver the intervention as intended. Therapists were asked about their ability to incorporate radical healing and internalized stigma interventions during the study. All therapists (100%) indicated that they used techniques from the training with each client in the study. When clients were asked about the incorporation of these techniques, 45 of 49 (92%) clients from the postintervention and follow-up indicated that their therapists used radical healing techniques in therapy. For internalized stigma techniques, 42 of 47 (90%) clients from the postintervention and follow-up indicated that their therapists incorporated these techniques into therapy (note: 1 client was not asked about this, resulting in missing data for both questions). Interestingly, 2 clients indicated at the 6-month follow-up that they were using tools from therapy to help reduce internalized stigma, despite saying their therapists did not use specific techniques to reduce internalized stigma at the postinterview.

The third aspect of feasibility for the intervention implementation includes whether the treatment conditions are acceptable to participants and interventionists. At postintervention, all but one participant indicated that they would participate in the study again and found the intervention useful; one participant—Rain (36 years, Black or Afro-American, gender-free, gender-empty, ungendered person)—who indicated they would not participate again stated:

> *I think I might be leaning towards a "maybe not"? Just because I'm thinking about the different kind of therapy that I'm thinking about moving into next, which would be like a more music therapy–based kind of thing. I have no idea where that would fit into how y'all are thinking about therapy but then also the options of people that you had as therapists were really limited for where I am in the world. So, I think that also would sway me towards a no.*

At the same time, this participant indicated that they experienced "so much positive change" because of the intervention and that "it just felt like a very simple, radical healing ... it just felt like this way to offer a simple invitation to living more fully." At the 6-month follow-up, 44 of 47 (94%) clients indicated they would participate again if they could, while 3 clients indicated that they were "50/50" or "not sure or probably" about participating again. Of these 3 participants, one shared that they felt as though they were in a better place following the completion of the study and would not want to take a spot from someone else who might benefit from the intervention. Despite improvement through the study, another client said that they would prefer to be able to choose who their therapist was rather than being assigned. Finally, the third client felt that they were "only able to scratch the surface" and would have wanted more sessions.

Finally, the last component of feasibility for intervention implementation focuses on adherence and engagement with the intervention. Adherence was not formally assessed in this study, as the therapists operated within a broader framework rather than following a strict protocol. This precluded the need for measuring strict adherence. Regarding engagement, clients were offered up to 15 sessions of psychotherapy but were informed that they did not need to use all of the sessions if they felt it was unnecessary. This approach aimed to assess the feasibility of a 15-session intervention. Of the 49 clients who began the intervention, no clients dropped out of the study, resulting in a 100% retention rate. Session attendance ranged from 7 to 15 sessions (mean 13.88, SD 2.14; median 15, IQR 14-15), with postassessments conducted after the client's final session. The vast majority of clients completed all 15 sessions (n=32, 65%). Only 5 clients completed 10 or fewer sessions, primarily due to issues with the clients' schedules. The client who completed the fewest sessions moved out of the country but was able to have a formal termination session with their therapist and complete postintervention and follow-up study sessions.

Discussion

Principal Findings

The primary aim of this study was to determine the feasibility and acceptability of the HOPE teletherapy intervention. All components assessed in this study, following the guidelines of Teresi et al [17] for pilot studies, demonstrated both feasibility and acceptability. For the intervention component of the study, we achieved a 100% retention rate; all 49 clients who started the intervention also completed it, though there were varying levels of session attendance. This retention rate is notably higher than the rates reported in meta-analyses of clinical trials, which typically range from 16% to 28% [24-29]. To date, there are no clinical trials that are directly comparable to this one for assessing retention rates. However, researchers indicate that attention to racial and ethnic identity components in interventions reduces attrition in clinical trials [30]. The only other published psychotherapy clinical trial that has focused on 2STNB clients also demonstrated high retention of clients (ie, 95.74%) [8]. We hypothesize that the high retention rate in this open trial was likely due to a combination of factors: our study's CBPR method and process, which strengthened the design for implementation with 2STNB populations who are people of

XSL•FO
RenderX

color; the provision of free, readily accessible outpatient therapy (ie, telehealth); and the flexible scheduling of the 15 sessions over an extended period.

In addition to the completion of the intervention, all but one client indicated at postintervention that they would participate in the study again. The large majority of clients were enthusiastic about the study, with several making unprompted remarks that it was the best research involvement they had experienced. This mirrors findings from a previous randomized clinical trial with 2STNB clients [8]. Besides the reasons listed earlier for high retention, clients appreciated the compensation structure for the 3 waves of data collection and the provision of free outpatient psychotherapy for the duration of the study. Additionally, clients felt that, in addition to benefiting from the intervention individually, they were contributing to scientific efforts to improve mental health interventions for future 2STNB clients who are people of color. This aligns with the critical consciousness component of the PRH, emphasizing the importance of contributing to bettering the experiences of community members [11,12].

Although the feasibility and acceptability data were strong in this study, the least feasible component was the session-by-session surveys (Outcome Questionnaire-45 and WAI-C) that clients were asked to fill out. Many clients indicated that they did not like taking the surveys because they felt the questions were rote or found it difficult to assign a number to their experiences. Studies show that clients in clinical trials rate their satisfaction highest when convenience is high [30,31], and it appears that completing surveys before each session was likely the least convenient aspect of the study design.

Limitations and Future Research

The present findings should be interpreted with the study's limitations in mind. A significant limitation of our pilot design is the absence of a control group, which limits our ability to conclusively determine whether the HOPE intervention is more effective than other treatments in improving mental health and reducing internalized stigma among 2STNB individuals who are people of color. Future research should involve a randomized controlled trial that compares the HOPE intervention with either treatment as usual or an additional intervention that has demonstrated efficacy. This future study will be crucial for disseminating evidence-based therapies tailored to the needs of historically marginalized 2STNB individuals who are people of color and can also determine the dose needed for optimal therapy to occur. In addition, this study's demographic composition, predominantly people located in the United States and nonbinary populations with a female sex at birth, indicates a need for future research to engage a more diverse array of participants. Specifically, the notable underrepresentation of transwomen of color highlights an important gap. Future studies should aim to better understand how to effectively reach and support the particularly underrepresented group within the broader community.

Conclusions

In this open clinical trial, we developed and implemented the HOPE intervention, the first psychotherapy intervention specifically developed for people of color who are 2STNB, based on PRH and an intervention aimed at reducing internalized transnegativity. We found broad enthusiasm, acceptability, and feasibility for the implementation of the HOPE intervention, specifically regarding data collection, psychometric adequacy, interventionist recruitment or training or retention, delivery of the intervention, acceptability of the intervention to clients, and client engagement with the intervention. Piloting the HOPE intervention is an exciting first step to determining effective interventions for 2STNB clients who are people of color and seeking psychotherapy that addresses stigma and oppression.

Acknowledgments

The authors would like to thank many people and organizations who have contributed to this project, including Marquel Norton, Yante Turner, Osha Towers, Kye Richards, Kelley Blair, Della Mosley, Hector Adames, Diverse & Resilient, the Wisconsin Office for Rural Health, and the Diversity Center for Oklahoma. Finally, the authors would especially like to thank the therapists and clients who participated in the Healing Through Ongoing Psychological Empowerment intervention.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to the sensitive nature of qualitative data, but quotes and excerpts are available from the corresponding author on reasonable request.

Authors' Contributions

SLB, EAT, SD, JL, and LL contributed to the design of the study. SLB conducted the primary analysis and wrote the Methods, Results, and Discussion sections. EAT wrote the literature review. JL and SD conducted interviews, and JL wrote the Limitations and Future Research section and the Abstract. EM trained the therapists in the study. LL, JL, and EM provided feedback and edits.

Conflicts of Interest

SLB has been hired by the American Civil Liberties Union, Lambda Legal, the National Center for Lesbian Rights, Gender Justice, and Wardenski P.C. to provide expert testimony in court cases.

XSL•FO
RenderX

Multimedia Appendix 1
Feasibility or acceptability interview protocol questions.
[DOCX File, 17 KB - formative_v9i1e64477_app1.docx]

References

1. James SE, Herman JL, Rankin S, Keisling M, Mottet L, Anafi M. The report of the 2015 U.S. transgender survey. : National Center for Transgender Equality; 2016 URL: https://transequality.org/sites/default/files/docs/usts/USTS-Full-Report-Dec17.pdf [accessed 2025-04-30]
2. Morris ER, Lindley L, Galupo MP. “Better issues to focus on”: transgender microaggressions as ethical violations in therapy. Couns Psychol 2020 Aug;48(6):883-915. [doi: 10.1177/0011000020924391]
3. Arora S, Gonzalez KA, Abreu RL, Gloster C. “Therapy can be restorative, but can also be really harmful”: therapy experiences of QTBIPOC clients. Psychotherapy (Chic) 2022 Dec;59(4):498-510. [doi: 10.1037/pst0000443] [Medline: 35617231]
4. Mizock L, Lundquist C. Missteps in psychotherapy with transgender clients: promoting gender sensitivity in counseling and psychological practice. Psychol Sex Orientat Gend Divers 2016;3(2):148-155. [doi: 10.1037/sgd0000177]
5. Wiegand A. Barred from transition: the gatekeeping of gender-affirming care during the gender clinic era. Intersect: Stanford J Sci Technol Soc 2021;15(1):1-11 [FREE Full text]
6. Erby AN, White ME. Broaching partially-shared identities: critically interrogating power and intragroup dynamics in counseling practice with trans people of color. Int J Transgend Health 2022;23(1-2):122-132. [doi: 10.1080/26895269.2020.1838389] [Medline: 35403107]
7. Freeman-Coppadge DJ, Langroudi KF. Beyond LGBTQ-affirmative therapy: fostering growth and healing through intersectionality. In: Nadal KL, Scharrón-del Río MR, editors. Queer Psychology: Intersectional Perspectives: Springer International Publishing; 2021:159-179. [doi: 10.1007/978-3-030-74146-4_9]
8. Budge SL, Sinnard MT, Hoyt WT. Longitudinal effects of psychotherapy with transgender and nonbinary clients: a randomized controlled pilot trial. Psychotherapy (Chic) 2021 Mar;58(1):1-11. [doi: 10.1037/pst0000310] [Medline: 32567869]
9. Wilcox MM. Oppression is not “culture”: the need to center systemic and structural determinants to address anti-Black racism and racial trauma in psychotherapy. Psychotherapy (Chic) 2023 Mar;60(1):76-85. [doi: 10.1037/pst0000446] [Medline: 35708949]
10. Dominguez J, Matsuno S, Adamez H. Using radical healing to address racial-gendered colonialism: a common factors approach to (imperfect) anti-colonial psychotherapy. Am Psychol (forthcoming). .
11. Adames HY, Chavez-Dueñas NY, Lewis JA, et al. Radical healing in psychotherapy: addressing the wounds of racism-related stress and trauma. Psychotherapy (Chic) 2023 Mar;60(1):39-50. [doi: 10.1037/pst0000435] [Medline: 35266783]
12. French BH, Lewis JA, Mosley DV, et al. Toward a psychological framework of radical healing in communities of color. Couns Psychol 2020 Jan;48(1):14-46. [doi: 10.1177/0011000019843506]
13. Israel T, Matsuno E, Choi AY, et al. Reducing internalized transnegativity: randomized controlled trial of an online intervention. Psychol Sex Orientat Gend Divers 2021;8(4):429-439. [doi: 10.1037/sgd0000447]
14. Buchanan NT, Wiklund LO. Intersectionality research in psychological science: resisting the tendency to disconnect, dilute, and depoliticize. Res Child Adolesc Psychopathol 2021 Jan;49(1):25-31. [doi: 10.1007/s10802-020-00748-y] [Medline: 33400076]
15. Veale JF, Deutsch MB, Devor AH, et al. Setting a research agenda in trans health: an expert assessment of priorities and issues by trans and nonbinary researchers. Int J Transgend Health 2022;23(4):392-408. [doi: 10.1080/26895269.2022.2044425] [Medline: 36324879]
16. Fleming PJ, Stone LC, Creary MS, et al. Antiracism and community-based participatory research: synergies, challenges, and opportunities. Am J Public Health 2023 Jan;113(1):70-78. [doi: 10.2105/AJPH.2022.307114] [Medline: 36516389]
17. Teresi JA, Yu X, Stewart AL, Hays RD. Guidelines for designing and evaluating feasibility pilot studies. Med Care 2022 Jan 1;60(1):95-103. [doi: 10.1097/MLR.0000000000001664] [Medline: 34812790]
18. Cocks K, Torgerson DJ. Sample size calculations for pilot randomized trials: a confidence interval approach. J Clin Epidemiol 2013 Feb;66(2):197-201. [doi: 10.1016/j.jclinepi.2012.09.002] [Medline: 23195919]
19. Matsuno E, Israel T. Psychological interventions promoting resilience among transgender individuals: transgender resilience intervention model (TRIM). Couns Psychol 2018 Jul;46(5):632-655. [doi: 10.1177/0011000018787261]
20. Matsuno E. Nonbinary-affirming psychological interventions. Cogn Behav Pract 2019 Nov;26(4):617-628. [doi: 10.1016/j.cbpra.2018.09.003]
21. Tebbe EA, Matsuno E, Lee J, Domínguez S, Budge SL. Implementation and evaluation of the gender resilience, resistance, empowerment, and affirmation training (GREAT) pilot program. Psychotherapy (Chic) 2025 Mar;62(1):32-43. [doi: 10.1037/pst0000552] [Medline: 40095968]
22. Elo S, Kyngäs H. The qualitative content analysis process. J Adv Nurs 2008 Apr;62(1):107-115. [doi: 10.1111/j.1365-2648.2007.04569.x] [Medline: 18352969]
23. Tracey TJ, Kokotovic AM. Factor structure of the Working Alliance Inventory. Psychol Assess J Consult Clin Psychol 1989;1(3):207-210. [doi: 10.1037/1040-3590.1.3.207]

XSL•FO
RenderX

24. Beckstead DJ, Hatch AL, Lambert MJ, Eggett DL, Goates MK, Vermeersch DA. Clinical significance of the Outcome Questionnaire (OQ-45.2). Behav Anal Today 2003;4(1):86-97. [doi: 10.1037/h0100015]
25. Cooper AA, Conklin LR. Dropout from individual psychotherapy for major depression: a meta-analysis of randomized clinical trials. Clin Psychol Rev 2015 Aug;40:57-65. [doi: 10.1016/j.cpr.2015.05.001] [Medline: 26067572]
26. Dixon LJ, Linardon J. A systematic review and meta-analysis of dropout rates from dialectical behaviour therapy in randomized controlled trials. Cogn Behav Ther 2020 May;49(3):181-196. [doi: 10.1080/16506073.2019.1620324] [Medline: 31204902]
27. Lewis C, Roberts NP, Gibson S, Bisson JI. Dropout from psychological therapies for post-traumatic stress disorder (PTSD) in adults: systematic review and meta-analysis. Eur J Psychotraumatol 2020;11(1):1709709. [doi: 10.1080/20008198.2019.1709709] [Medline: 32284816]
28. Iliakis EA, Ilagan GS, Choi-Kain LW. Dropout rates from psychotherapy trials for borderline personality disorder: a meta-analysis. Pers Disord Theory Res Treat 2021;12(3):193-206. [doi: 10.1037/per0000453]
29. Cardona ND, Ametaj AA, Cassiello-Robbins C, et al. Outcomes of people of color in an efficacy trial of cognitive-behavioral treatments for anxiety, depression, and related disorders: preliminary evidence. J Nerv Ment Dis 2023 Sep 1;211(9):711-720. [doi: 10.1097/NMD.0000000000001692] [Medline: 37432031]
30. Faulkner SD, Somers F, Boudes M, Nafria B, Robinson P. Using patient perspectives to inform better clinical trial design and conduct: current trends and future directions. Pharm Med 2023 Mar;37(2):129-138. [doi: 10.1007/s40290-022-00458-4]
31. Sine S, de Bruin A, Getz K. Patient engagement initiatives in clinical trials: recent trends and implications. Ther Innov Regul Sci 2021 Sep;55(5):1059-1065. [doi: 10.1007/s43441-021-00306-8] [Medline: 34097288]

Abbreviations

2STNB: Two Spirit, transgender, and nonbinary
CAB: community advisory board
CBPR: community-based participatory research
HOPE: Healing Through Ongoing Psychological Empowerment
PRH: psychological framework for radical healing
WAI-C: Working Alliance Inventory

Edited by A Mavragani; submitted 18.07.24; peer-reviewed by N Döring; revised version received 11.03.25; accepted 12.03.25; published 12.05.25.

Please cite as:
Budge SL, Tebbe EA, Lee J, Domínguez Jr S, Matsuno E, Lindley L
The Healing Through Ongoing Psychological Empowerment Telehealth Intervention With Two Spirit, Transgender, and Nonbinary Clients of Color in the United States: Open Clinical Trial Feasibility and Implementation Analysis
JMIR Form Res 2025;9:e64477
URL: https://formative.jmir.org/2025/1/e64477
doi:10.2196/64477

© Stephanie Lynne Budge, Elliot Aaron Tebbe, Joonwoo Lee, Sergio Dominguez Jr, Em Matsuno, Louis Lindley. Originally published in JMIR Formative Research (https://formative.jmir.org), 12.5.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Impact of a "Digital Health" Curriculum on Students' Perception About Competence and Relevance of Digital Health Topics for Future Professional Challenges: Prospective Pilot Study

Juliane Kröplin[1], MBA, MD, DMD; Leonie Maier[1]; Jan-Hendrik Lenz[1,2], MME, MD, DMD; Bernd Romeike[2], MME, MD

[1]Department of Oral and Maxillofacial Surgery, University Medical Centre Rostock, Schillingallee 35, Rostock, Germany

[2]Department of the Dean of Studies in Medical Didactics, University Medical Centre Rostock, Rostock, Germany

Corresponding Author:
Juliane Kröplin, MBA, MD, DMD
Department of Oral and Maxillofacial Surgery, University Medical Centre Rostock, Schillingallee 35, Rostock, Germany

Abstract

Background: The rapid integration of digital technologies in health care has emphasized the need to ensure that medical students are well-equipped with the knowledge and competencies related to digital health.

Objective: This study aimed to evaluate the impact of the "Digital Health" curriculum at our university on the perceptions of medical students regarding the relevance of digital health topics for their future professions and their self-assessed competence in these areas.

Methods: The "Digital Health" curriculum was introduced at a German university for 2 consecutive semesters. The perceived relevance of topics for their future careers and their subjective competence were evaluated before and after the curriculum using a Likert scale. Furthermore, the practical gamification-based teaching part of the robotics teaching unit was evaluated. In total, 6 months after completing the last semester, a follow-up analysis was performed with questions on the significance of the completed curriculum for current and future professional challenges regarding digital health and suggestions for improvement for innovative teaching. The study was meticulously planned and supported by an approved ethics vote of the local ethics committee to ensure that all ethical guidelines were adhered to (A 2022-0137).

Results: A total of 20 students participated, with 13 (65%) being women. In particular, data protection and information security were considered the most relevant topics both before and after the curriculum. Significant increases in perceived importance were observed for messenger apps (mean increase of 0.8 [SD 1.2]; $P<.01$). Regarding self-assessed competence, significant development was observed on almost all topics. The greatest development was observed in robotics (mean increase of 1.8 [SD 1.2]; $P<.001$), open educational resources (mean increase of 1.7 [SD 1.5]; $P<.001$), and simulation-training (mean increase of 1.6 [SD 1.3]; $P<.001$). The gamification-based, robot-related teaching was predominantly rated suitable and very enjoyable for the students.

Conclusions: The results highlight the potential to integrate more innovative teaching techniques, such as gamification, augmented reality, virtual reality, and simulation training, into a technologically advanced health care environment. Finally, the overarching importance of artificial intelligence and digital health applications signals the need to further integrate them, given their potential in remote and personalized medicine.

(JMIR Form Res 2025;9:e58940) doi:10.2196/58940

KEYWORDS

education; gamification; robotics; telemedicine; wearable electronic devices; artificial intelligence

Introduction

The term "digital health" closely links advanced medical technologies, disruptive innovations, and digital communication with the provision of optimal health care [1]. Providing medical students with a comprehensive understanding of the basic concepts and possibilities of eHealth is crucial to enabling future physicians to recognize and apply the benefits of these services in their daily practice. It is likely to increase their openness to digital health and reduce their reluctance to adopt digital health [2].

Medical students are well aware of the possible applications and effects of digitization, including artificial intelligence (AI) in medicine [3]. Despite their positive attitude and high expectations, they express concerns about patient acceptance and possible professional consequences [4]. Through an effective curriculum, students can gain learning benefits from this technology and understand both opportunities and

XSL•FO
RenderX

challenges [5]. Most medical students consider this training to be a valuable part of their education, as it helps them better understand and apply telemedicine approaches with greater confidence [6].

There is a general consensus in the current literature that future health care professionals must be prepared for the digital age [7]. Elective subjects appear to be suitable formats for prompt introduction, but long-term implementation in the compulsory curriculum should be the goal [8]. National initiatives could better meet the challenges of digitization in health care and offer solutions on how best to teach digital skills to students [9]. Innovative teaching content also requires innovative teaching methods. A promising approach is gamification, which refers to the integration of playful elements in a serious context. Educators in the health care sector are also increasingly relying on gamification to optimize students' learning success [10].

Current developments raise fundamental questions about the future of health care professions: (1) What digital skills will be needed in the future? (2) How will digitization affect the relationship between patients and health care professionals ? And (3) adapting education and training in health care professions will be crucial to meet the demands of an increasingly digitalized society [11].

The curriculum "Digital Health - Digitization and Digital Transformation of Medicine" was offered for the first time at a German university in the winter semester of 2022 or 2023. An analysis of the implementation process was published recently [12].

The curriculum is divided into four main topics: (1) digital transformation of medical education, (2) digital communication, (3) digital leadership, and Robotics or AI.

The study aims: (1) to analyze the impact of a transdisciplinary digital health curriculum on students' perceived relevance and competence regarding digital health topics and (2) to analyze the acceptance and suitability of a gamification-based teaching method integrated into the existing curriculum.

This was intended to evaluate when and how the opinions and subjective experiences of medical students can be optimally incorporated and integrated into the development of an effective curriculum. By integrating students into this process, the aim is to ensure that their individual perspectives are considered and a user-orientated curriculum is created.

Methods

Study Structure and Design

This prospective pilot study was designed as a pre-post analysis to assess the development in subjective competence and the subjectively perceived impact of our digital health curriculum on the perception about the relevance of digital health topics for future professional challenges. The curriculum content and implementation process have already been published [12].

Participation in both the curriculum and all evaluations was voluntary. The data were collected and analyzed anonymously.

Participants

The age and gender of the curriculum participants were analyzed. Students were also asked whether they were interested in doing a doctorate and what plans they had for a postdoc phase.

Relevance and Competence Acquisition Concerning Digital Health Topics: Pre-Post Comparison

The digital health curriculum included the following topics: big data, AI, digital health apps, messenger apps, wearables, telemedicine, data protection or information security, digital ethics, simulation training, augmented reality (AR), virtual reality (VR), economy, e-learning, robotics, digital leadership, mobile working, social media, and open educational resources (OER.

In the first step, students were asked to rate the relevance of individual course content for their future career on a Likert scale from 1 (very relevant) to 5 (irrelevant; Table 1). The subjective assessment of their own competence was also carried out on a Likert scale from 1 (very good) to 5 (not good at all) on the same topics (Table 2). The same survey was carried out again at the end of the course to achieve the aim of a pre-post comparison.

XSL•FO
RenderX

Table . Questionnaire of gamification-based robotic teaching.

Question	Answer options
RT[a]1: How did you complete the learning unit with the robot construction kit?	• Alone • In teams of two • With other teams of two of the digital health curriculum • With other people (parents, friends, etc)
RT2: How much time in hours did you spend with the robot construction kit?	• Building • Programming

RT3: What applications have you carried out?	• Building the robot
	  • Yes
	  • No
	• Programming movements
	  • Yes
	  • No
	• Programming of acoustic signals
	  • Yes
	  • No
	• Programming of visual signals
	  • Yes
	  • No

RT4: By using the robot construction kit, I programmed independently for the first time	• Yes • No
RT5: The course was a suitable way to reduce reservations about the topic of robotics in medicine	• Very true • Rather true • Partly/partly • Rather not applicable • Does not apply at all
RT6: I enjoyed working with the robot construction kit	• Very true • Rather true • Partly or partly • Rather not applicable • Does not apply at all

[a]RT: robotic teaching.

XSL•FO
RenderX

Table . Follow-up evaluation 6 months after the end of the course.

Question	Answer option
Follow-up question 1: Define the term "digital health."	__[a]
Follow-up question 2: How do you think your future work will be changed by developments in digital health?	—
Follow-up question 3: Do you see these changes in a positive or negative way?	• Positive • Negative • Neither (neutral)
Follow-up question 4: Give 2 examples each of positive and negative changes	• Positive: • Negative:
Follow-up question 5: How do you think teaching should be adapted?	—
Follow-up question 6: For example, have you found the elective's Instagram account useful and did you use it?	• Useful and used it • Useful, but didn't use it • Not useful, didn't use it • I did not know about the account
Follow-up question 7: Do you already use parts of Digital Health (both in general and for learning)? eg, wearables for interpreting your own medical data, AR[b] and VR[c] in the medical field, ChatGPT, LLM[d] (other generative AI[e])	• Yes (please describe): • No

[a]not applicable.

[b]AR: augmented reality.

[c]VR: virtual reality.

[d]LLM: large language model.

[e]AI: artificial intelligence.

Gamification in Robotic Teaching

In order to introduce the students to the subject of robotics playfully and reduce their fear of contact, they were provided with a robot construction kit for the entire duration of the course (Lego Mindstorms, Billund, and Denmark). Building instructions can be found online [13].

They were able to build a robot at home. During one course, the robots were demonstrated and reprogrammed live. Figure 1 shows 4 robots, that were built by the students during the semester. Students could program movements, and acoustic and visual signals with an application-based visual programming software, that can be downloaded for free (Lego Mindstorms app, Billund, and Denmark [14]). Due to the visual programming environment, users create programs by dragging and dropping adaptable command blocks into a sequence (Figure 2).

The gamification analysis took place at the beginning of the robotics teaching (RT) unit (course days 10 and 12). The Lego Mindstorms kit enabled students to built different types of robots. They had also the choice to build the robot alone or in teams. There were no specific instructions.

An intensive literature search found no validated, usable questionnaires, which is why our own questionnaires were developed as part of this pilot study.

The students' experiences were assessed using a 6-item questionnaire with open and closed questions (Robotics R-questions Table 1).

The application behavior was analyzed and the usefulness of the teaching method was assessed.

XSL•FO
RenderX

Figure 1. Different robots built with the Lego Mindstorms construction kit.

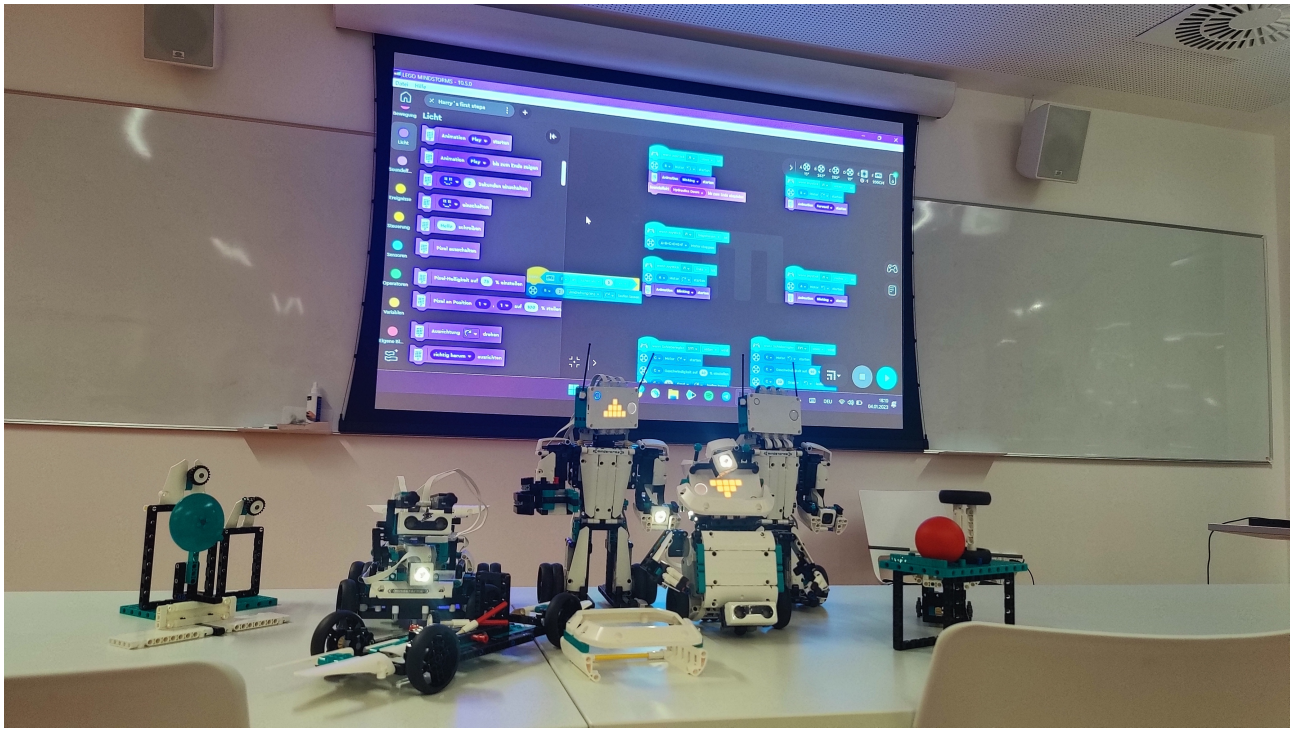

Figure 2. Screenshot of the Lego Mindstorms application-based visual programming software.

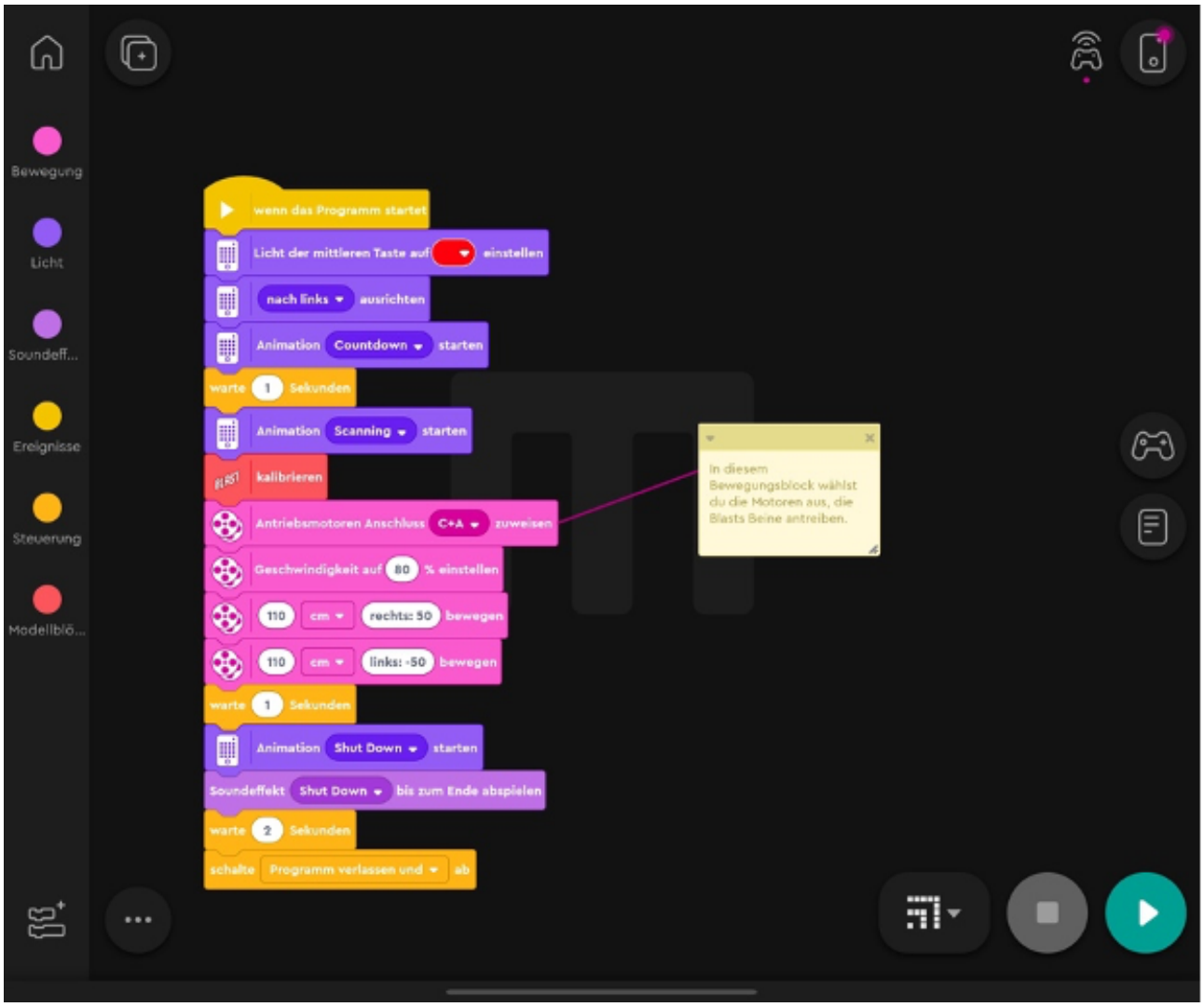

XSL•FO
RenderX

Follow-Up Evaluation

In total, 6 months (cohort 1) and 1 year (cohort 2) after completion of the curriculum, participants were asked to participate in a retrospective online survey using the online tool SurveyMonkey [15] (follow-up questions, Table 2).

This survey aimed to capture the sustainable subjective perception of various aspects related to the implementation of digital health in medical education and professional practice (Table 2). Participants were asked about their opinions on the future impact of digital health on their work and whether they felt positive or negative about these changes. Another focus was on the participants' opinions on possible adaptations of medical teaching to the developments in digital health. Furthermore, usage practices in the field of digital health were recorded both in general and in the context of learning to gain insight into the current integration of these technologies.

Statistics

Data were analyzed using IBM SPSS version 27. Quantitative data like pre-post-test results were analyzed with paired, 2-tailed Student's *t* test. A value of $P<.05$ was considered statistically significant. Open questions like the participants' career goals, definition of the term "digital health," their opinion about how future work would be changed by developments in digital health, or how teaching should be adapted were summarized with descriptive statistics.

Ethical Considerations

The study was meticulously planned and supported by an approved ethics vote of the Ethics Committee at the Medical Faculty of the University of Rostock to ensure that all ethical guidelines were adhered to (A 2022 - 0137). Participation in the evaluation was voluntary. Informed consent to participate was obtained from all of the participants.

Results

Participants

Within the first year, a total of n=20 students participated in the digital health curriculum. Out of 20 students, 13 (65%) participants were women and 7 (35%) men. During the pre-evaluation, the participants' academic career goals were evaluated. A total of 90% (18/20) of the participants aspire to a doctorate and 10% (2/20) of the participants also intend to obtain a PhD degree.

Relevance and Competence Concerning Digital Health Topics: Pre-Post Comparison

Table 3 shows the results of the pre-post comparison on the subjectively assessed relevance of the content of the digital health elective for the participants' later professional future. The topic of data protection and information security was rated the most important at the beginning of the curriculum (mean 1.2 [SD 0.5]). This assessment was also confirmed after the completion of the curriculum. A significant increase in importance was verified for the topic of messenger apps.

Table 4 shows the results of the pre-post comparison of the participants' subjective assessment of their own competence regarding the learning objectives of the curriculum. Except for the topics of messenger apps and social media, there was a significant subjective increase in knowledge in all topics. The greatest increase in competence was in the area of robotics; the least in the area of messenger apps. The topics in which no significant increase in knowledge could be generated were also those with the highest score in previous knowledge during the initial evaluation.

XSL•FO
RenderX

Table . Assessment of the relevance of curriculum content for the later professional future.

Relevance for the future	Pre, mean (SD)	Post, mean (SD)	Relevance increase, mean (SD)	*P* value
Messenger apps	2.9 (1.1)	2.1 (1.0)	0.8 (1.2)	.01
AR[a] and VR[b]	2.1 (1.0)	1.8 (1.2)	0.3 (1.3)	.33
Mobile working	2.4 (1.2)	2.1 (1.3)	0.3 (1.6)	.40
Wearables	2.1 (1.2)	2.4 (1.1)	0.3 (1.6)	.37
Digital health apps	1.8 (0.9)	1.6 (1.0)	0.2 (1.5)	.55
OER[c]	2.0 (1.0)	1.8 (1.0)	0.2 (1.5)	.57
Big data	2.1 (0.9)	2.0 (1.1)	0.1 (1.0)	.65
Economy	1.9 (1.0)	1.8 (1.0)	0.1 (1.9)	.58
AI[d]	1.5 (0.7)	1.5 (1.0)	0 (1.1)	.84
Social media	2.5 (1.0)	2.5 (1.4)	0 (1.3)	.87
Data protection or information security	1.2 (0.5)	1.3 (0.8)	−0.1 (1.0)	.65
Digital didactics	1.7 (1.0)	1.8 (1.2)	−0.1 (1.2)	.58
Digital ethics	1.4 (0.8)	1.5 (0.9)	−0.1 (0.9)	.61
Digital leadership	1.6 (0.7)	1.7 (1.0)	−0.1 (1.1)	.55
Simulation training	1.4 (0.5)	1.5 (1.0)	−0.1 (1.0)	.53
Robotics	1.5 (0.8)	1.8 (1.1)	−0.3 (1.0)	.13
Telemedicine	1.8 (1.1)	2.1 (1.4)	−0.3 (1.6)	.51

[a]AR: augmented reality.
[b]VR: virtual reality.
[c]OER: open educational resources.
[d]AI: artificial intelligence.

XSL•FO
RenderX

Table . Subjective competence development determined via a pre-post-comparison.

Competence	Precomparision, mean (SD)	Postcomparision, mean (SD)	Knowledge increase, mean (SD)	*P* value
Robotics	4.2 (1.1)	2.4 (1.0)	1.8 (1.2)	<.001
OER[a]	3.7 (1.3)	2.0 (0.8)	1.7 (1.5)	<.001
Simulation training	4.1 (1.1)	2.5 (1.2)	1.6 (1.3)	<.001
AI[b]	3.8 (1.1)	2.3 (0.9)	1.5 (1.0)	<.001
Digital leadership	4.2 (1.1)	2.7 (1.0)	1.5 (1.2)	<.001
AR[c] and VR[d]	3.6 (1.4)	2.2 (1.1)	1.4 (1.6)	<.001
Mobile working	3.9 (1.2)	2.5 (0.9)	1.4 (1.2)	<.001
Wearables	3.6 (1.2)	2.2 (1.2)	1.4 (1.9)	.004
Digital health apps	3.2 (1.4)	1.9 (0.9)	1.3 (1.4)	<.001
Digital didactics	3.0 (1.2)	1.8 (1.0)	1.2 (1.4)	.00
Digital ethics	3.8 (1.0)	2.7 (1.0)	1.1 (1.1)	<.001
Telemedicine	3.9 (1.0)	2.8 (1.2)	1.1 (1.3)	<.001
Data protection or information security	3.3 (1.3)	2.4 (1.1)	0.9 (1.3)	.01
Big data	4 (1.4)	3.2 (0.9)	0.8 (1.5)	.02
Economy	3.8 (1.2)	3.0 (1.2)	0.8 (1.5)	.02
Social media	2.1 (1.2)	1.9 (1.1)	0.2 (1.4)	.54
Messenger apps	1.8 (1.2)	1.6 (1.1)	0.2	.56

[a]OER: open educational resources.

[b]AI: artificial intelligence.

[c]AR: augmented reality.

[d]VR: virtual reality.

Gamification in Robotic Teaching

A total of 13 students participated in the robotics evaluation of the practical teaching unit. All students worked on the task in teams (RT1). The average time that participants invested in the task was 2 hours and 30 minutes (1.48 h building the robot; 0.42 h programming the robot; RT2). In Figure 1, 11 students had built their robot at the time of the survey; n=7 had already programmed the robot (RT3). For the vast majority of students (n=11), this was their first opportunity to generate code independently (RT4). Most of the participants (n=10) rated the application of the robotics teaching unit as suitable (RT5) and enjoyed using it (RT6).

Follow-Up Evaluation

In total, 13 out of 20 (65 %) students from both semesters of the curriculum participated in our follow-up evaluation. The first question of the participants, follow-up question 1, confirmed a basic understanding of the former participants' definition of digital health. The most frequently mentioned key points were digital transformation or digitization of medicine and the improvement of health care through digital technologies. Table 5 shows all responses to follow-up question 2 of how digital health will influence future work. The most common expectation is about new possibilities in diagnostics and therapy.

All survey participants are positive about the changes associated with digital health (follow-up question 3). The most frequently cited potential positive change was an increase in efficiency or increased flexibility and new possibilities in diagnostics and therapy (n=7 each; follow-up question 4). The most frequently cited potential negative change was an increased risk of misusing and stealing data (n=8) and the risk of digital addiction (n=5; follow-up question 4).

In follow-up question 5, former participants were asked to provide their assessment of the extent to which teaching should be adapted to the challenges of digital transformation. In particular an adaptation of the learning methodology (n=6); especially AI-supported, for more individualized or efficient or sustainable learning), familiarization with existing apps (n=5); especially digital documentation systems) and the general inclusion of digital health subjects (n=4) was indicated.

At the beginning of the curriculum, all participants were informed of the existence of a social media channel. As part of the follow-up evaluation, the students were able to evaluate both their usage behavior and the usefulness of this channel (follow-up question 6; Table 6).

The follow-up question (7 about the current use of digital tools in students' everyday medical lives was answered by all survey participants. When asked about specific applications, n=12

XSL•FO
RenderX

participants confirmed the use of ChatGPT [16] and the translation software DeepL [17]. Other participants indicated the use of wearables (n=5) and AR or VR for learning anatomy (n=1).

Table . Follow-up question 2: How do you think your future work will be changed by developments in digital health? Core messages; n=x.

Answer to follow-up question 2	Quantity (n)
New possibilities in diagnostics and therapy (eg, AI-assisted)	8
Easier data processing and documentation (eg, digital patient file)	4
Acceleration of processes or more efficient work	4
Home office or flexibility (eg, video consultation)	3
Generally positive enrichment (better patient contact)	2
Increased risk of data misuse and theft	1

Table . Follow-up question 6: Have you found the elective's Instagram account useful and did you use it, n=x?

Answer to follow-up question 6	Quantity (n)
Useful and used it	3
Useful but did not use it	8
Not useful, did not use it	1
I did not know about the account	1

Discussion

Principal Findings

This study provides an overview of the influence of our digital health curriculum on students' perception about their competence and the relevance of digital health topics for their future professional challenges. The results are based on 2 consecutive semesters in which the content of the elective subject "Digital Health" was taught at a German university. Although the topic of data protection was named the most relevant content, the students confirmed a significant increase in relevance for numerous topics during the curriculum, primarily with regard to the teaching unit messenger apps.

The analysis of the practical teaching unit (building and programming a LEGO robot) has shown that this approach is fun and was rated by the students as suitable for reducing fear of contact with the topic of robotics.

In a follow-up online analysis 6 months after completing the course, the students stated, among other things, that they would like to see a general adaptation of the teaching methods in the degree program, for example, to deepen individual, more efficient, and sustainable learning through AI-based apps.

Demographics

A particularity of our cohort appears to be a very high interest in science and in completing a doctoral thesis, while globally, medical students rather underscore the necessity and importance of research in medical education, with many students reporting positive attitudes and interest in research efforts [18].

Relevance and Competence Acquisition Concerning Digital Health Topics: Pre-Post Comparison and Follow-Up Evaluation

Data- and IT-Security

Data protection and IT security were the highest ranking in terms of relevance among our students. Students also reported a significant increase in competence after completing the curriculum.

As part of the follow-up analysis, the students rated data protection and information security as the most frequently cited potentially negative change. In particular, data misuse and data theft were identified as risk factors. This can be explained, among other things, by the presence of the subject and the actual threat scenarios [19]. Furthermore, the current literature shows the need to protect health care technology and patient data against unauthorized access [20]. In the context of relevance analyses on the topic of digital health, the topic of data protection generally has the highest priority [21]. The constant assessment of high relevance in the pre- and post-test survey indicates that we have adequately addressed the topic in the curriculum.

Evolving Telehealth Landscape

Although telemedicine has existed in various forms for decades, recent technological innovations, combined with changes in patient preferences and the global response to crises such as the COVID-19 pandemic, have greatly accelerated its adoption and importance [22]. Data from our curriculum suggest that, while students recognize the importance of many aspects of digital health, there seems to be an underestimate of the growing importance of telehealth. Telemedicine not only expands access to health care but also introduces efficiencies that traditional in-person visits may lack. Current and future health care professionals need a solid understanding of these tools, both

XSL•FO
RenderX

for their practical utility and for their wider impact on the health care system. The practical application of the interaction of telemedicine is an aspired goal for the further development of our curriculum. Simulated real-life experience would allow us to directly experience both the advantages and possible hurdles of communication.

Ethical Aspects of Digital Health

As our results show, the ethics subject area was rated as very important by the participants in the curriculum. Furthermore, the students confirmed a significant subjective increase in competence as a result of their participation. Ethical considerations of digital health are manifold. Issues such as consent in the age of digital monitoring, bias in AI algorithms, and ethical use of patient data for research are all pertinent discussions that must be ingrained in the medical curriculum. Furthermore, ethical perspectives shall not be limited to the technology itself. Instead, it is suggested to consider the broader sociotechnical system and also social justice for a comprehensive ethical analysis of digital health [23]. It can be assumed that the increasing use of AI in medical practice will also change the way future physicians understand their role. This also affects the development of as yet unforeseeable areas of responsibility, which require adaptation of medical training. Therefore, it is particularly important to prepare students for these changes in order to ensure high-quality health care [24].

Modern Pedagogies: AR VR, Simulation Training and OER

The digital health curriculum saw a significant increase in the perceived competence of students in areas such as simulation training. Simulation, particularly in medical education, offers a risk-free environment for students to hone their skills, make mistakes, and learn in a controlled setting . Furthermore, the rise of OER democratizes learning and offers students a plethora of resources that were not available to previous generations [25]. During the analysis period, we succeeded in significantly increasing the subjective competence gained on the topic of OER.

Taking into account the study results of Hassall et al [25], it can be assumed that students will also make use of OER resources in the future. In this study, a lack of awareness, motivation, and training regarding OERs were identified as the biggest obstacles to the use of OERs.

The students also confirmed a subjective increase in knowledge with regard to the topic of VR and AR. VR and AR technologies have particular potential with regard to modern pedagogies [26]. They address the need for sustainable medicine and enable a reproducible and objectifiable learning environment.

Taking the follow-up analysis into account, we feel that we will offer the compulsory elective subject in the future. However, the focus will be on further adaptation of teaching methods (digital transformation of medical education) as well as the increased teaching of existing apps for digital documentation systems.

Digital Health in Social Media

Social media has become an increasingly important source of information on medical content in recent years. This also applies to teaching [27]. This has prompted us to accompany our digital health curriculum with a social media channel. The results related to our digital health curriculum Instagram account are of particular interest to us. The high confirmation of the usefulness of such a digital communication tool is remarkable because the use of social media in medical education is an emerging field that needs further investigation and is still the subject of controversial debates [28,29]. The fact that only a minority used the channel can be explained by the fact that it was preferably used for general information on special offers on the compulsory elective subject. In the future, expanding the interaction with participants and the possibility of active participation to solve specific tasks would be a way to increase the frequency of use.

AI

AI plays an increasing role in medical education [30]. The great importance that students continue to attach to AI for their professional future in a pre-post comparison shows the special significance of this topic. This makes it even more significant that the students experienced a significant subjective increase in competence as a result of participating in the digital health curriculum. The follow-up analysis also provides an impressive insight into the current usage behavior of students with regard to AI. In particular, the use of ChatGPT and the translation software DeepL was mentioned. A development that is in agreement with the currently available literature and is of great interest for the further development of future teaching and research [31,32]. Furthermore, the follow-up analysis shows the students' desire for adapted teaching using AI-based support systems. These findings once again demonstrate the particular importance of analyzing the necessary framework conditions for constructive ethical use of AI.

Gamification in Robotic Teaching

It can be assumed that the topic of robotics was evaluated with the greatest subjective increase in competence due to the special focus on the implementation of a gamification-based teaching set.

Robot-assisted surgery is one of the most revolutionary developments in surgery. It can be assumed that demand will continue to increase in the future. Recently, the performance of robotic surgery has increased exponentially. This is primarily due to the benefits of robotics. Robotic surgery can help reduce hospitalization and recovery times. In addition, innovative robotic technology enables better visualization and fewer postoperative complications. Access to robotics for students is often limited by the fact that cost-intensive technology is not available at all locations [33]. Furthermore, robotics training is offered predominantly only to experienced surgeons and is mainly based primarily on workshops and fellowships [34]. Stockheim et al [34] have already addressed residents as a target group with their RoC curriculum, Robotic Curriculum for Young Surgeons. Introducing students at an early stage has the potential to reduce fears of contact even before they start their specialist

XSL•FO
RenderX

training and to get students excited about specialisms with robotic apps.

Considering the limited access for students, it was particularly important to develop solutions that enable low-threshold access to the topic of robotics. The success of this decision is reflected largely in the results of our evaluation.

Most of our participants considered the robot construction kit to be a suitable way of reducing the fear of contact with the subject of robotics. The majority of students also stated that working with the robot was fun. These results underline the findings of previously published literature on the subject of gamification [35].

In medical education, in particular, gamification has been shown to lead to a higher level of motivation, commitment, and challenge compared to traditional didactics [36]. This offers teachers alternative ways to optimize the teaching process. This way of imparting skills is widely received by learners and allows for an immersive experience that is perceived as effective, engaging, easy to understand, interesting, and instructive [37]. Research even suggests that gamification can improve the knowledge of health care professionals compared to traditional training [38].

Limitations

Limitations arise in particular from the single-center study design and the small number of participants at the time of analysis. However, the number of participants is quite normal for a pilot study and a compulsory elective, where students work in small groups in order to increase the quality of the lessons. Finally, it must be considered that the results are purely subjective assessments of the participants. A follow-up study will also ask about the acquisition of specialist knowledge.

Conclusions

The introduction of the “Digital Health: Digitization and Digital Transformation of Medicine” curriculum stands as proof of the changing face of medical education in the 21st century. Positive outcomes in terms of competence development and alignment with relevant digital health topics underscore its success. Although the curriculum effectively improved competencies on a range of topics, emphasis on data security and ethics is paramount given the vulnerabilities associated with digital health data.

The integration of gamification-based teaching methods can be seen as a useful addition to the teaching of basic skills and a low-threshold approach to complex digital topics.

As digital health continues to advance, the task is now to ensure that such curricula are not static but evolve in tandem with the technological and ethical landscape of modern healthcare.

Acknowledgments

We thank our students for participating in our study and the study dean's office of our medical school for technical support. We would also like to thank all lecturers who have supported the curriculum with their contributions.

Conflicts of Interest

None declared.

References

1. Meskó B, Drobni Z, Bényei É, Gergely B, Győrffy Z. Digital health is a cultural transformation of traditional healthcare. Mhealth 2017;3:38. [doi: 10.21037/mhealth.2017.08.07] [Medline: 29184890]
2. Baumgartner M, Sauer C, Blagec K, Dorffner G. Digital health understanding and preparedness of medical students: a cross-sectional study. Med Educ Online 2022 Dec;27(1):2114851. [doi: 10.1080/10872981.2022.2114851] [Medline: 36036219]
3. Pinto Dos Santos D, Giese D, Brodehl S, et al. Medical students’ attitude towards artificial intelligence: a multicentre survey. Eur Radiol 2019 Apr;29(4):1640-1646. [doi: 10.1007/s00330-018-5601-1] [Medline: 29980928]
4. Sultan I, Bardi MF, Baatta AM, Almaghrabi S, Mohammed RA. Medical students’ attitude towards robotic surgery: a cross-sectional survey. J Med Educ Curric Dev 2022;9:23821205211066483. [doi: 10.1177/23821205211066483] [Medline: 35036565]
5. Higgins RM, O’Sullivan P. The robotic surgery learning experience through the eyes of the medical student: what do they see? J Surg Educ 2020;77(3):549-556. [doi: 10.1016/j.jsurg.2019.12.011] [Medline: 31959582]
6. Waseh S, Dicker AP. Telemedicine training in undergraduate medical education: mixed-methods review. JMIR Med Educ 2019 Apr 8;5(1):e12515. [doi: 10.2196/12515] [Medline: 30958269]
7. Kuhn S, Huettl F, Deutsch K, Kirchgässner E, Huber T, Kneist W. Surgical education in the digital age - virtual reality, augmented reality and robotics in the medical school. Zentralbl Chir 2021 Feb;146(1):37-43. [doi: 10.1055/a-1265-7259] [Medline: 33588501]
8. Han ER, Yeo S, Kim MJ, Lee YH, Park KH, Roh H. Medical education trends for future physicians in the era of advanced technology and artificial intelligence: an integrative review. BMC Med Educ 2019 Dec 11;19(1):460. [doi: 10.1186/s12909-019-1891-5] [Medline: 31829208]

XSL•FO
RenderX

9. Khurana MP, Raaschou-Pedersen DE, Kurtzhals J, Bardram JE, Ostrowski SR, Bundgaard JS. Digital health competencies in medical school education: a scoping review and Delphi method study. BMC Med Educ 2022 Dec;22(1):129. [doi: 10.1186/s12909-022-03163-7]
10. van Gaalen AEJ, Brouwer J, Schönrock-Adema J, Bouwkamp-Timmer T, Jaarsma ADC, Georgiadis JR. Gamification of health professions education: a systematic review. Adv in Health Sci Educ 2021 May;26(2):683-711. [doi: 10.1007/s10459-020-10000-3]
11. Hege I, Tolks D, Kuhn S, Shiozawa T. Digital skills in healthcare. GMS J Med Educ 2020;37(6):Doc63. [doi: 10.3205/zma001356] [Medline: 33225055]
12. Kröplin J, Maier L, Lenz JH, Romeike B. Knowledge transfer and networking upon implementation of a transdisciplinary digital health curriculum in a unique digital health training culture: prospective analysis. JMIR Med Educ 2024 Apr 15;10:e51389. [doi: 10.2196/51389] [Medline: 38632710]
13. LEGO. URL: https://www.lego.com/de-de/service/buildinginstructions/51515?locale=de-de [accessed 2024-09-03]
14. LEGO. URL: https://www.lego.com/de-de/themes/mindstorms/app [accessed 2024-09-03]
15. SurveyMonkey. URL: https://de.surveymonkey.com [accessed 2024-09-03]
16. ChatGPT. URL: https://openai.com/chatgpt [accessed 2024-09-03]
17. DeepL. URL: https://www.deepl.com/de/translator [accessed 2024-09-03]
18. Stone C, Dogbey GY, Klenzak S, Van Fossen K, Tan B, Brannan GD. Contemporary global perspectives of medical students on research during undergraduate medical education: a systematic literature review. Med Educ Online 2018 Dec;23(1):1537430. [doi: 10.1080/10872981.2018.1537430] [Medline: 30372403]
19. von der Forst M, Popp E, Weigand MA, Neuhaus C. Special emergency situations and hazard control in German hospitals-a survey on the current state. Anaesthesiologie 2023 Nov;72(11):784-790. [doi: 10.1007/s00101-023-01349-2] [Medline: 37855945]
20. Kruse CS, Frederick B, Jacobson T, Monticone DK. Cybersecurity in healthcare: a systematic review of modern threats and trends. Technol Health Care 2017;25(1):1-10. [doi: 10.3233/THC-161263] [Medline: 27689562]
21. Kröplin J, Huber T, Geis C, Braun B, Fritz T. eSurgery—digital transformation in surgery, surgical education and training: survey analysis of the status quo in Germany. Eur Surg 2022 Oct;54(5):249-258. [doi: 10.1007/s10353-022-00747-x]
22. Omboni S, Padwal RS, Alessa T, et al. The worldwide impact of telemedicine during COVID-19: current evidence and recommendations for the future. Connect Health 2022 Jan 4;1:7-35. [doi: 10.20517/ch.2021.03] [Medline: 35233563]
23. Shaw JA, Donia J. The sociotechnical ethics of digital health: a critique and extension of approaches from bioethics. Front Digit Health 2021;3:725088. [doi: 10.3389/fdgth.2021.725088] [Medline: 34713196]
24. Masters K. Artificial intelligence in medical education. Med Teach 2019 Sep;41(9):976-980. [doi: 10.1080/0142159X.2019.1595557] [Medline: 31007106]
25. Hassall C, Lewis DI. Institutional and technological barriers to the use of open educational resources (OERs) in physiology and medical education. Adv Physiol Educ 2017 Mar 1;41(1):77-81. [doi: 10.1152/advan.00171.2016] [Medline: 28143826]
26. Kanwal L, Gulzar M, Idrees W, Ikram F, Sukhia RH, Fida M. The application of virtual reality and augmented reality in dentistry - a literature review. J Pak Med Assoc 2024 Apr;74(4 (Supple-4):S126-S131. [doi: 10.47391/JPMA.AKU-9S-19] [Medline: 38712420]
27. Hayman R, Smith EE. Social media in undergraduate teaching and learning: a scoping review protocol. PLoS ONE 2023;18(11):e0291306. [doi: 10.1371/journal.pone.0291306] [Medline: 38015905]
28. Guckian J, Utukuri M, Asif A, et al. Social media in undergraduate medical education: a systematic review. Med Educ 2021 Nov;55(11):1227-1241. [doi: 10.1111/medu.14567] [Medline: 33988867]
29. Cheston CC, Flickinger TE, Chisolm MS. Social media use in medical education: a systematic review. Acad Med 2013 Jun;88(6):893-901. [doi: 10.1097/ACM.0b013e31828ffc23] [Medline: 23619071]
30. Moritz S, Romeike B, Stosch C, Tolks D. Generative AI (gAI) in medical education: Chat-GPT and co. GMS J Med Educ 2023;40(4):Doc54. [doi: 10.3205/zma001636] [Medline: 37560050]
31. Sedaghat S. Early applications of ChatGPT in medical practice, education and research. Clin Med (Lond) 2023 May;23(3):278-279. [doi: 10.7861/clinmed.2023-0078] [Medline: 37085182]
32. Eysenbach G. The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers. JMIR Med Educ 2023 Mar 6;9:e46885. [doi: 10.2196/46885] [Medline: 36863937]
33. Naik R, Mandal I. Robotic simulation experience in undergraduate medical education: a perspective. J Robot Surg 2020 Oct;14(5):793-794. [doi: 10.1007/s11701-020-01059-6] [Medline: 32125601]
34. Stockheim J, Perrakis A, Sabel BA, Waschipky R, Croner RS. RoCS: robotic curriculum for young surgeons. J Robot Surg 2023 Apr;17(2):495-507. [doi: 10.1007/s11701-022-01444-3] [Medline: 35810233]
35. Deterding S, Sicart M, Nacke L, O'Hara K, Dixon D. Gamification. using game-design elements in non-gaming contexts. Presented at: CHI '11; May 7-12, 2011; Vancouver BC Canada. [doi: 10.1145/1979742.1979575]
36. Gue S, Ray J, Ganti L. Gamification of graduate medical education in an emergency medicine residency program. Int J Emerg Med 2022 Aug 30;15(1):41. [doi: 10.1186/s12245-022-00445-1] [Medline: 36042403]

XSL•FO
RenderX

37. Xu M, Luo Y, Zhang Y, Xia R, Qian H, Zou X. Game-based learning in medical education. Front Public Health 2023;11:1113682. [doi: 10.3389/fpubh.2023.1113682] [Medline: 36935696]
38. Gentry SV, Gauthier A, L'Estrade Ehrstrom B, et al. Serious Gaming and Gamification Education in Health Professions: Systematic Review. J Med Internet Res 2019 Mar 28;21(3):e12994. [doi: 10.2196/12994] [Medline: 30920375]

Abbreviations

AI: artificial intelligence
AR: augmented reality
OER: open educational resources
RT: robotic teaching
VR: virtual reality

Edited by A Mavragani; submitted 31.03.24; peer-reviewed by R Basu, R Rösch; revised version received 23.09.24; accepted 29.10.24; published 12.05.25.

Please cite as:
Kröplin J, Maier L, Lenz JH, Romeike B
Impact of a "Digital Health" Curriculum on Students' Perception About Competence and Relevance of Digital Health Topics for Future Professional Challenges: Prospective Pilot Study
JMIR Form Res 2025;9:e58940
URL: https://formative.jmir.org/2025/1/e58940
doi:10.2196/58940

© Juliane Kröplin, Leonie Maier, Jan-Hendrik Lenz, Bernd Romeike. Originally published in JMIR Formative Research (https://formative.jmir.org), 12.5.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Effects of Using a Digital Peer-Supported App on Glycated Hemoglobin Changes Among Patients With Type 2 Diabetes: Prospective Single-Arm Pilot Study

Shota Yoshihara[1,2], MSc, OTR; Kayoko Takahashi[1,3], ScD, OTR; Hiroyuki Kawaguchi[2]; Nozomi Harai[4], MD; Kyoichiro Tsuchiya[4], MD, PhD

[1]Department of Rehabilitation Sciences, Kitasato University Graduate School of Medical Sciences, 1-15-1 Kitasato, Minami-ku, Sagamihara, Kanagawa, Japan

[2]A10 Lab Inc, Tokyo, Japan

[3]Department of Occupational Therapy, School of Allied Health Science, Kitasato University, Kanagawa, Japan

[4]Department of Diabetes and Endocrinology, University of Yamanashi Hospital, Yamanashi, Japan

Corresponding Author:
Shota Yoshihara, MSc, OTR
Department of Rehabilitation Sciences, Kitasato University Graduate School of Medical Sciences, 1-15-1 Kitasato, Minami-ku, Sagamihara, Kanagawa, Japan

Abstract

Background: Controlling glycated hemoglobin (HbA_{1c}) levels can be challenging for patients with type 2 diabetes mellitus (T2DM). Peer support promotes HbA_{1c} control, and a digital peer-supported app designed for group interactions may enable patients with T2DM to encourage one another to achieve better HbA_{1c} outcomes. However, no studies have investigated the use of digital peer-supported apps to control HbA_{1c} levels in patients with T2DM.

Objective: This pilot study aimed to explore the effects of a digital peer-supported app on HbA_{1c} control in patients with T2DM.

Methods: This prospective single-arm pilot study enrolled patients with T2DM who owned smartphones and visited medical institutions in Japan. During the 3-month intervention, participants used a digital peer-supported app in addition to receiving standard care. This app allowed participants to share activity logs and concerns via a chat function to improve HbA_{1c} levels through mutual engagement and encouragement. The primary outcome was the change in HbA_{1c} levels, measured at health care facilities at baseline and after 3 months. The secondary outcomes were body weight and blood pressure, with the most recent data obtained from hospitals and clinics. Physical activity (≥1 hour/day) was assessed at the same time points using a self-reported questionnaire.

Results: The study included 21 participants with a median age of 56 (IQR 51 - 61) years, of which 13 (61.9%) were female. After using the digital peer-supported app for 3 months, the participants' HbA_{1c} levels significantly decreased from 7.1% (SD 0.6%) at baseline to 6.9% (SD 0.1%) (P=.04). Similarly, participants' body weight decreased from 70.7 (SD 12.7) kg to 69.9 (SD 12.4) kg (P =.004) through app use. Although blood pressure decreased slightly from 128.2 (SD 12.5) mm Hg to 126.0 (SD 12.9) mm Hg, this change was not statistically significant (P=.20). Additionally, the proportion of participants engaged in ≥1 hour of daily physical activity significantly increased from 23.5% (n=4) to 58.5% (n=10) (P=.03).

Conclusions: In addition to receiving standard clinical care, the use of a digital peer-supported app may significantly lower HbA_{1c} levels in patients with T2DM by promoting healthy behaviors.

(JMIR Form Res 2025;9:e72659) doi:10.2196/72659

KEYWORDS

diabetes type 2; T2DM; self-management; digital health; HbA1c; lifestyle intervention; digital intervention; type 2 diabetes mellitus; glycated hemoglobin

Introduction

Type 2 diabetes mellitus (T2DM) has become a global public health issue that significantly affects 90% of the 537 million patients with T2DM worldwide [1]. T2DM and its complications significantly impact patients and the society by increasing health care costs and reducing life expectancy and quality [2]. Glycated hemoglobin (HbA_{1c}) is the gold standard for measuring average blood glucose levels over 3 months and predicting the relative risk of diabetes complications [3,4]. A previous study revealed

XSL•FO
RenderX

that adopting healthy behaviors (ie, maintaining a healthy body weight, following a healthy diet, exercising, avoiding smoking, and avoiding drinking alcohol) is an effective strategy for controlling HbA_{1c} levels in patients with T2DM [5]. However, adopting and maintaining healthy behaviors independently can be challenging for patients with T2DM [6].

Peer support, defined as mutual assistance among individuals with shared experiences or challenges, may be an effective strategy for enhancing healthy behaviors in patients with T2DM [7]. A meta-analysis of 13 randomized controlled trials that focused on in-person peer support among patients with diabetes has indicated that patients receiving peer support experienced a significant reduction in HbA_{1c} levels of 0.57% (95% CI 0.36 - 0.78) on average, compared to those without peer support, across interventions durations ranging from 3 to 24 months [8]. However, traditional lifestyle counseling such as peer support in clinical settings, which relies heavily on face-to-face interactions and significant human and time resources, provides limited peer support for patients with T2DM that is often restricted to infrequent hospital visits [9].

With advancements in digital health technology, a digital peer-supported app may provide new opportunities for patients with T2DM to connect and support each other in virtual spaces, thus enabling those with T2DM to collaborate toward the goal of controlling HbA_{1c} levels. Using digital peer-supported apps, known as group-type apps, enable individuals to encourage each other to engage in healthy behaviors [10-12]. However, the effectiveness of such digital peer support interventions in T2DM management remains largely unclear.

This pilot study aimed to explore the effects of a digital, peer-supported app on HbA_{1c} control in patients with T2DM.

Methods

Study Design

This 3-month single-arm intervention study used a pre–post evaluation design. According to the transtheoretical model [13], a 3-month period is widely regarded in many studies as sufficient for individuals to progress from the “preparation stage” to the “action stage” and, subsequently, to the “maintenance stage.” In addition, the 3-month period is both physiologically justified and conceptually appropriate for evaluating behavioral and metabolic changes [3,4], as HbA_{1c} reflects the average blood glucose levels over the preceding 2 - 3 months, which corresponds to the approximately 120-day lifespan of red blood cells [14].

This study was conducted between December 2021 and June 2022 as an industry-government-academia collaboration among local governments, app-making companies, and universities. It was part of the “TRY! YAMANASHI! Pilot Experiment Support Project (FY2021, First Term)” in Yamanashi Prefecture. The program supports the social implementation of advanced technologies and services developed by startups and other innovators to address regional issues and create public value. Support also offered multifaceted assistance (eg, funding, expert consultation, partnerships with local resources, and public outreach) for projects that had advanced to the practical application stage through prior pilot testing.

Ethical Considerations

Informed consent was obtained from all participants before their enrollment. This study adhered to the ethical principles outlined in the Declaration of Helsinki and was approved by the independent Ethics Review Committee of Healthcare Systems Co., Ltd. (Approval No. 2130). Participants received a financial reward of 5000 JPY (US $34.70) in the form of a gift card when the end-of-study data were delivered (ie, HbA_{1c} levels at both time points and patient-reported outcomes at baseline and at 3 months). All collected data were anonymized prior to analysis to ensure participant confidentiality and privacy.

Participant Recruitment

Participants were recruited through convenience sampling and began using the digital peer-supported app upon recruitment. An explanation of the study and its objectives was provided by physicians when they visited the clinic and hospital. The referral process involved the following scenarios: (1) during a health consultation at a hospital; (2) as part of a diabetes program at a hospital; and (3) during a medical consultation at a clinic.

The inclusion criteria were (1) individuals who had been fully informed about the objectives and details of the study, possessed the capacity to consent, had understood the information thoroughly, and have freely agree to participate by providing written consent; (2) Japanese men and women aged 20-74 years at the time of consent; (3) patients with T2DM or prediabetes with an HbA_{1c} level of ≥6.0% and <8.5%; and (4) individuals who owned and could operate a smartphone (iOS or Android) capable of installing the app. The exclusion criteria were (1) diseases that could interfere with the study, as determined by the medical institution conducting the study, and (2) severe symptoms or disabilities due to lifestyle-related diseases, as determined by the medical institution.

Intervention Program

The experimental setup is shown in Figure 1. The participants began using the digital peer-supported app after being introduced to the app by physicians and downloading it. During the initial login, they received tutorials within the app. Participants were encouraged to join diabetes management teams within the digital peer-supported app and use this app for 3 months while continuing their treatment according to the standards of care, including any necessary adjustments to their diabetes pharmacotherapy. They were instructed to complete questionnaires at the beginning and end of the program to evaluate changes in their lifestyle habits and health behaviors. These questionnaires were designed to confirm participant sex, age, and health behaviors. HbA_{1c} levels, blood pressure, and body weight were measured at the beginning and end of the study, using the most recent data from hospitals and clinics.

XSL•FO
RenderX

Figure 1. Experimental flow of this study.

Digital Peer-Supported App

The study used "Minchalle," a commercially available digital peer-supported app [11]. This digital peer-supported app was developed by A10 Lab Inc. in June 2015 with an initial release in November 2015. It is available for both Android and iPhone devices. The digital peer-supported app created group chats of up to five participants that aimed at controlling their HbA_{1c} levels, and participants were anonymously assigned to each group (Figure 2). Once daily, the participants posted a set of daily activities to control their HbA_{1c} levels, along with photos and comments in a group chat box. The main functions of the digital peer-supported app used in this study were (1) posting photos, step counts, and comments about the day; (2) posting approvals from group members; (3) setting step-count goals on a group basis; and (4) providing feedback from assistant robots. Participants could post comments or photos more than once a day and enjoy interacting with other members; however, daily participation was not obligatory, and the participants were allowed to freely use the app. Examples of the user interfaces of digital peer-supported apps are provided in the Multimedia Appendix 1.

Figure 2. Examples of app screens. (A) Select a group. (B) Post a photo taken that day and comment on the day's events to the group. (C) The contents of the postings are displayed in the group. (D) Respond to posts by group members. (**E**) Check the records.

Measurements

We collected basic demographic information at baseline, including sex (male or female), age (in continuous years), alcohol consumption (categorized as nondrinker, former drinker, occasional drinker, or nearly daily drinker), and smoking status (categorized as nonsmoker, former smoker, or current smoker).

The primary outcome of this study was the difference in HbA_{1c} levels, which were evaluated using values measured at health care facilities immediately during the participants' visits at

XSL•FO
RenderX

baseline and at the 3-month follow up. All participants' baseline and 3-month follow-up values were measured at the same health care facilities using the same reagents. The secondary outcomes included differences in body weight and blood pressure, which were also obtained from the health care facilities. At the same time points, physical activity (≥1 hour/day) was evaluated using a self-reported questionnaire: "Do you engage in walking or equivalent physical activity for >1 hour per day in your daily life?" Responses were recorded as either "yes" or "no."

Statistical Analysis

We performed the Wilcoxon signed-rank test to assess the overall significance of differences in HbA_{1c} levels. Similarly, changes in secondary outcomes such as blood pressure and body weight were evaluated using the Wilcoxon signed-rank test to compare baseline and 3-month follow-up values. Additionally, McNemar test was used to analyze paired binary data, specifically to compare baseline and 3-month follow up values for physical activity, ≥1 hour/day (%).

All analyses were performed using the Stata software (version 18.0; StataCorp LLC). Statistical significance was set at $P<.05$ (2-tailed).

Results

The eligibility assessment and enrolment processes are shown in Figures 3 and 4, respectively. Based on HbA_{1c} levels, blood pressure, and weight (Figure 3), 22 participants were initially enrolled, and app usage was confirmed using the company database. One participant was excluded because of nonresponsiveness at follow-up, leaving 21 participants (95.5%) with complete HbA_{1c} levels, blood pressure, and weight data for analysis. Similarly, based on questionnaire data for physical activity ≥1 hour/day (%) (Figure 4), 22 participants were enrolled, and app usage was confirmed.

Figure 3. Flow diagram of study participant recruitment (glycated hemoglobin, blood pressure, and body weight).

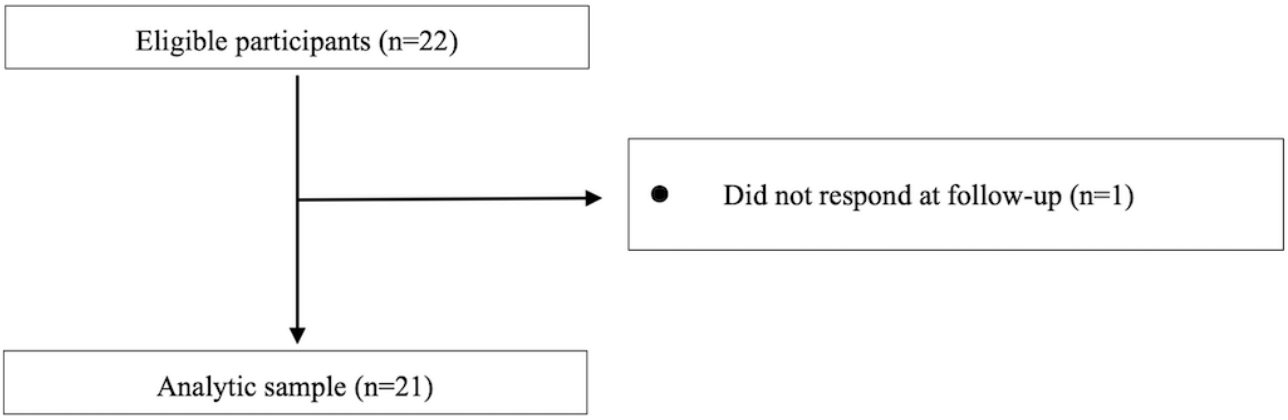

Figure 4. Flow diagram of study participant recruitment (physical activity ≥1 hour/day).

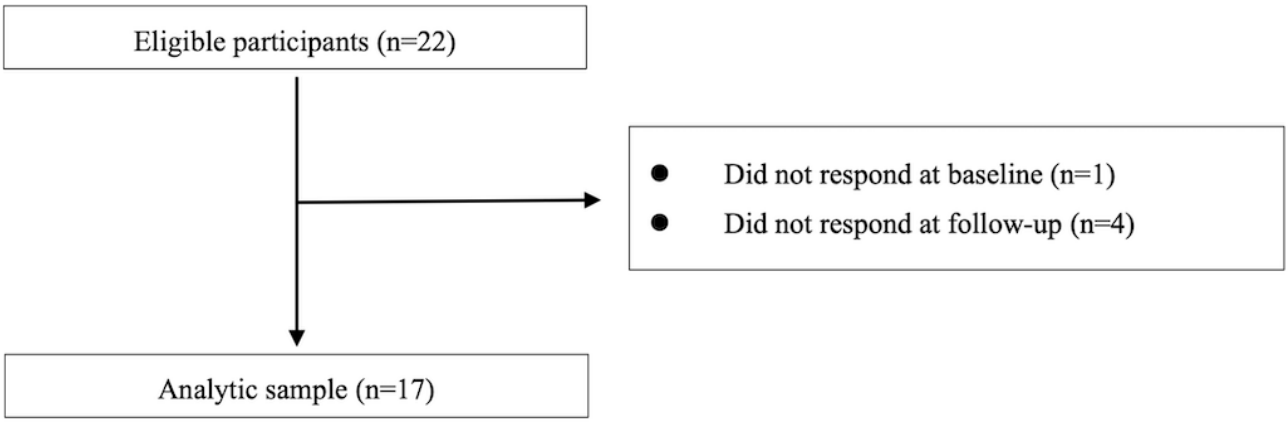

Table 1 shows the baseline characteristics of the 21 participants, including 3 (14.3%) individuals with prediabetes. The median participant age was 56 (IQR 51 - 61) years, and 61.9% (n=13) were females. Among the patients, 52.4% (n=11) were nonsmokers, 23.8% (n=5) were former smokers, and 23.8% (n=5) were current smokers. Regarding alcohol intake, 28.6% (n=6) of the study cohort were nondrinkers, 4.8% (n=1) was a former drinker, 61.9% (n=13) were occasional drinkers, and 4.8% (n=1) was a daily drinker. At baseline, 23.5% (n=4) of the participants reported ≥1 hour/day of physical activity. The median HbA_{1c}, blood pressure, and body weight were 7.1% (IQR 6.7% - 7.3%), 127 (IQR 121 - 135) mm Hg, and 70.4 (IQR 61.9 - 80.4) kg, respectively.

XSL•FO
RenderX

Table . Baseline participant characteristics.

Baseline characteristics	Participants[a] (N=21)
Sex, n (%)	
Male	8 (38.1)
Female	13 (61.9)
Age (year), median (IQR)	56 (51-61)
Smoking status, n (%)	
Nonsmoker	11 (52.4)
Former smoker	5 (23.8)
Current smoker	5 (23.8)
Alcohol intake, n (%)	
Nondrinker	6 (28.6)
Former drinker	1 (4.8)
Occasional drinker	13 (61.9)
Daily drinker	1 (4.8)
Physical activity ≥1 hour/day (n=17)[b], n (%)	4 (23.5)
HbA_{1c}[c] (%), median (IQR)	7.1 (6.7-7.3)
Blood pressure (mm Hg), median (IQR)	127 (121-135)
Body weight (kg), median (IQR)	70.4 (61.9-80.4)

[a]Data are shown as medians and quartiles for continuous variables and numbers (percentages) for categorical variables.

[b]Data were available only for 17 participants.

[c]HbA_{1c}: glycated hemoglobin.

Table 2 presents the results from the Wilcoxon signed-rank test, showing changes in HbA_{1c} levels from baseline to the 3-month follow-up. The mean baseline HbA_{1c} value for the intraindividual control group was 7.1% (SD 0.6%), which decreased by an average of 0.1% (SD 0.3%) to reach a 3-month follow-up value of 6.9% (SD 0.1%) (*P*=.04). As an additional metabolic parameter, patients reported their weight before and after 3 months. The groups' mean baseline blood pressure was recorded at 128.2 (SD 12.5) mm Hg and decreased on average by 2.1 (SD 9.4) mm Hg, resulting in a follow-up value of 126.0 (SD 12.9) mm Hg. However, this change in blood pressure was not statistically significant (*P*=.20). In addition, the mean baseline weight for the intraindividual control group was 70.7 (SD 12.7) kg, which decreased on average by 0.8 (SD 1.0) kg, resulting in a 3-month follow-up value of 69.9 (SD 12.4) kg (*P*.004). Finally, the proportion of participants engaging in ≥1 hour of daily physical activity significantly increased from 23.5% (n=4) to 58.5% (n=10) (*P*=.03).

Table . Changes in glycemic control, metabolic parameters, and physical activity.

Parameters	Baseline values (n=21)	Follow-up values (at 3 months)	Change	*P* value
HbA_{1c} (%), mean (SD)	7.1(0.6)	6.9 (0.1)	−0.1 (0.3)	.04[a]
Blood pressure (mm Hg), mean (SD)	128.2 (12.5)	126.0 (12.9)	−2.1 (9.4)	.20[a]
Body weight (kg), mean (SD)	70.7 (12.7)	69.9 (12.4)	−0.8 (1.0)	<.004[a]
Physical activity ≥1 hour/day (n=17), n (%)	4 (23.5)	10 (58.5)	__[b]	.03[c]

[a]Wilcoxon signed-rank test.

[b]Not applicable.

[c]McNemar test.

XSL•FO
RenderX

Discussion

Principal Findings

Our findings showed that a 3-month intervention using a digital peer-supported app led to a modest but statistically significant reduction in HbA_{1c} levels among patients with T2DM, indicating improved glycemic control. Additionally, self-reported frequency of physical activity improved significantly, highlighting the potential of this app to promote healthier behaviors. The findings of this study suggest that digital, peer-supported apps may serve as promising practical tools for the management of T2DM.

Moreover, the intervention using a digital peer-supported app significantly reduced HbA_{1c} levels among patients with T2DM, which is consistent with findings from previous studies that used nongroup-based, single-user apps for blood glucose control [15-17]. Peer support, grounded in the social cognitive theory, could effectively aid in blood glucose control when implemented through a digital peer-supported app [18].

Our results may be attributed to specific app design features, including reminder prompts and chat functions. One possible explanation is that communication via chats may also impact individual decision-making and behavior, shaping online community norms to maintain healthy behaviors [19]. These community norms may offer promising ways to improve adherence to diet and exercise recommendations by providing real-time reminders and emotional support from peers [20].

Previous studies on diabetes apps have highlighted the importance of chat-based communication in promoting healthy behavioral change, particularly by reinforcing positive actions such as initiating exercise, which can be more effective than restrictive measures [21]. Alternatively, real-time chats and reminders may help users remain frequently aware of their goals and consistently track their behaviors, which may facilitate effective acquisition, learning, and app of health information [15,22]. Although digital peer support platforms carry the risk of misinformation or inappropriate advice, no such incidents were reported during the study period. This may be attributed to the implementation of a safety-monitoring system, which included content oversight by the app provider and a dedicated research contact point. Future interventions should incorporate risk-management protocols to ensure safe and effective peer communication.

Our results showed a statistically significant change in HbA_{1c} levels, although the reduction did not reach a threshold considered meaningful in a clinical context [23]. One possible explanation for this finding is that the participants may have had low baseline HbA_{1c} levels. In a recent meta-analysis of online medicine, higher baseline HbA_{1c} levels (>7.5% or >8.0%), younger age (<55 years), and shorter duration of diabetes (<8.5 years and <7 years) were found to be associated with increased benefits [23]. Our study participants had lower baseline HbA_{1c} levels than those in other studies [15,16,24]. Future studies should aim to identify demographic or behavioral subgroups more likely to achieve HbA_{1c} reduction using digital peer-supported apps.

Strengths and Limitations

This study reported that a 3-month intervention using a digital peer-supported app led to a modest but statistically significant reduction in HbA_{1c} levels, which indicates improved glycemic control among patients with T2DM. Additionally, self-reported frequency of physical activity improved significantly, highlighting the potential of this app in promoting healthier behaviors. By minimizing interference from study personnel and addressing only technical aspects, the app was evaluated under real-life conditions to generate realistic and practical evidence. Therefore, digital peer-supported apps may offer cost-effective and scalable tools for managing diabetes and supporting lifestyle changes.

However, this study had several limitations that should be noted when interpreting our findings. First, as this was a single-arm interventional study with a small sample size, the analysis relied on classical statistical methods and a pre-post comparison, which makes it difficult to exclude the influence of confounding factors, such as baseline HbA_{1c}, age, or sex, through statistical adjustment. Although the reductions in HbA_{1c} and body weight were statistically significant, their relatively small magnitude along with the absence of detailed data on app usage, user engagement, and acceptability limit our ability to attribute these improvements solely to digital peer support interventions. These improvements may also have been mediated by increased adherence to treatment rather than by direct effects. To enhance the validity and generalizability of these findings, a multicenter randomized controlled trial with a larger and more diverse sample size and a parallel control group is currently underway (UMIN000056609). Second, this study used only HbA_{1c} level as the primary outcome measure. Although HbA_{1c} is a reliable indicator of long-term glycemic control and reflects average blood glucose levels over the past 2 - 3 months, it is subject to a time lag between behavioral changes and measurable outcomes. In contrast, continuous glucose monitoring devices can capture interstitial glucose levels in near real time, allowing immediate feedback and short-term trend analysis [25]. Future studies may benefit from combining long-term markers such as HbA_{1c} with continuous glucose monitoring to investigate the effects of interventions more comprehensively. Third, unmeasured factors such as medication adjustments and dietary habits during the study period may have influenced changes in HbA_{1c} levels. Although the duration since the diagnosis of diabetes may influence a participants' engagement with peer support and intervention outcomes, this information was not collected in the present study and should be considered in future studies. Fourth, this study did not collect information on personal or behavioral traits. As previous studies have shown that the stage of behavioral change may influence continued app usage [26], future studies should examine how individual characteristics affect long-term engagement in digital peer support. Finally, the sample size of 21 participants was relatively small and may have been insufficient to adequately detect significant changes in health outcomes. A formal sample size calculation was not conducted due to the lack of prior studies

XSL•FO
RenderX

on the effectiveness of digital peer support on health outcomes. Future studies should assess the impact of digital peer-supported apps on HbA_{1c} levels, enabling appropriate calculation of effect sizes.

Conclusions

The use of a digital peer-supported app intervention may significantly lower HbA_{1c} levels in patients with T2DM after 3 months. In T2DM populations capable of using the app, this intervention could promote healthier behaviors, in addition to regular clinical care, and may help reduce HbA_{1c} levels.

Acknowledgments

This study received financial support from A10 Lab Inc., the developer and operator of the habit-formation application, through self- and grant funding provided by the Yamanashi Prefecture. The authors thank the staff of the Yamanashi Prefecture Linear Future Creation Bureau and Linear Future Creation & Promotion Division (Hiroshi Saito, Hajime Hoya) for participant recruitment and project management. We also thank Makoto Tawada, Kunihiko Kubokawa, Makoto Tawada, Yoshifumi Kanemaru, Keisuke Kitamura, Sawako Koishi, and Minoru Nakazato for their invaluable assistance with data collection. Finally, we acknowledge Kanagawa Technology Innovations Inc. for overseeing and confirming the appropriateness of task execution and for conducting data analysis based on a commissioning contract. Portions of this manuscript were edited for grammar and clarity using ChatGPT [27] and Grammarly [28]. The final content was reviewed and approved by all authors, who take full responsibility for its accuracy and integrity. The authors would also like to thank Editage for editing the English language.

Authors' Contributions

Data curation: HK, NH, K Tsuchiya, SY
Formal analysis: Kanagawa Technology Innovations Inc.
Methodology: K Takahashi, HK, K Tsuchiya, SY
Writing – original draft – SY
Writing – review & editing – HK, K Takahashi, NH, K Tsuchiya, SY

Conflicts of Interest

HK was an employee of A10 Lab, Inc. at the time of this study. SY was an employee of A10 Lab Inc., when the manuscript was written but was not employed by the company when the research was conducted.

Multimedia Appendix 1
Examples of the user interfaces of digital peer-supported apps are provided.
[DOCX File, 70532 KB - formative_v9i1e72659_app1.docx]

References

1. Sun H, Saeedi P, Karuranga S, et al. IDF Diabetes Atlas: global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. Diabetes Res Clin Pract 2022 Jan;183:109119. [doi: 10.1016/j.diabres.2021.109119] [Medline: 34879977]
2. Association American Diabetes. Updates to the Standards of Medical Care in Diabetes-2018. Diabetes Care 2018 Sep;41(9):2045-2047. [doi: 10.2337/dc18-su09] [Medline: 30135199]
3. Diabetes Control and Complications Trial Research Group, Nathan DM, Genuth S, et al. The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. N Engl J Med 1993 Sep 30;329(14):977-986. [doi: 10.1056/NEJM199309303291401] [Medline: 8366922]
4. Stratton IM, Adler AI, Neil HA, et al. Association of glycaemia with macrovascular and microvascular complications of type 2 diabetes (UKPDS 35): prospective observational study. BMJ 2000 Aug 12;321(7258):405-412. [doi: 10.1136/bmj.321.7258.405] [Medline: 10938048]
5. Davies MJ, D'Alessio DA, Fradkin J, et al. Management of hyperglycemia in type 2 diabetes, 2018. a consensus report by the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD). Diabetes Care 2018 Dec;41(12):2669-2701. [doi: 10.2337/dci18-0033] [Medline: 30291106]
6. Ried-Larsen M, MacDonald CS, Johansen MY, et al. Why prescribe exercise as therapy in type 2 diabetes? We have a pill for that!. Diabetes Metab Res Rev 2018 Jul;34(5):e2999. [doi: 10.1002/dmrr.2999] [Medline: 29488311]
7. Riessman F. Restructuring help: a human services paradigm for the 1990s. American J Comm Psychol 1990 Apr;18(2):221-230. [doi: 10.1007/BF00931302]
8. Qi L, Liu Q, Qi X, Wu N, Tang W, Xiong H. Effectiveness of peer support for improving glycaemic control in patients with type 2 diabetes: a meta-analysis of randomized controlled trials. BMC Public Health 2015 May 6;15:471. [doi: 10.1186/s12889-015-1798-y] [Medline: 25943398]

XSL•FO
RenderX

9. Ricci-Cabello I, Ruiz-Pérez I, Rojas-García A, Pastor G, Rodríguez-Barranco M, Gonçalves DC. Characteristics and effectiveness of diabetes self-management educational programs targeted to racial/ethnic minority groups: a systematic review, meta-analysis and meta-regression. BMC Endocr Disord 2014 Jul 19;14:60. [doi: 10.1186/1472-6823-14-60] [Medline: 25037577]
10. Okamoto M, Saito Y, Nakamura S, et al. Smartphone-based digital peer support for a walking intervention among public officers in Kanagawa prefecture: single-arm pre- and postintervention evaluation. JMIR Form Res 2024 Sep 24;8:e53759. [doi: 10.2196/53759] [Medline: 39316793]
11. Tabira K, Oguma Y, Yoshihara S, et al. Digital peer-supported app intervention to promote physical activity among community-dwelling older adults: nonrandomized controlled trial. JMIR Aging 2024 May 30;7:e56184. [doi: 10.2196/56184] [Medline: 38814686]
12. Takebayashi M, Namba M, Koyama T, et al. Impact on step count by commitment-based health application. PLoS One 2024;19(8):e0305765. [doi: 10.1371/journal.pone.0305765] [Medline: 39133726]
13. Prochaska JO, Velicer WF. The transtheoretical model of health behavior change. Am J Health Promot 1997;12(1):38-48. [doi: 10.4278/0890-1171-12.1.38] [Medline: 10170434]
14. Sherwani SI, Khan HA, Ekhzaimy A, Masood A, Sakharkar MK. Significance of HbA1c test in diagnosis and prognosis of diabetic patients. Biomark Insights 2016;11:95-104. [doi: 10.4137/BMI.S38440] [Medline: 27398023]
15. Bretschneider MP, Klásek J, Karbanová M, Timpel P, Herrmann S, Schwarz PEH. Impact of a digital lifestyle intervention on diabetes self-management: a pilot study. Nutrients 2022 Apr 26;14(9):1810. [doi: 10.3390/nu14091810] [Medline: 35565782]
16. Bretschneider MP, Roth L, Schwarz PEH. Effectiveness of a digital health application for the treatment of diabetes type II-a pilot study. J Clin Med 2023 Sep 30;12(19):6317. [doi: 10.3390/jcm12196317] [Medline: 37834960]
17. Waki K, Fujita H, Uchimura Y, et al. DialBetics: a novel smartphone-based self-management support system for type 2 diabetes patients. J Diabetes Sci Technol 2014 Mar;8(2):209-215. [doi: 10.1177/1932296814526495] [Medline: 24876569]
18. Ginis KAM, Nigg CR, Smith AL. Peer-delivered physical activity interventions: an overlooked opportunity for physical activity promotion. Transl Behav Med 2013 Dec;3(4):434-443. [doi: 10.1007/s13142-013-0215-2] [Medline: 24294332]
19. Chancellor S, Hu A, De Choudhury M. Norms matter: contrasting social support around behavior change in online weight loss communities. Presented at: CHI '18: CHI Conference on Human Factors in Computing Systems; Apr 21-26, 2018; Montreal, QC, Canada. [doi: 10.1145/3173574.3174240]
20. Saffari M, Ghanizadeh G, Koenig HG. Health education via mobile text messaging for glycemic control in adults with type 2 diabetes: a systematic review and meta-analysis. Prim Care Diabetes 2014 Dec;8(4):275-285. [doi: 10.1016/j.pcd.2014.03.004] [Medline: 24793589]
21. Zhang Y, Li X, Luo S, et al. Use, Perspectives, and attitudes regarding diabetes management mobile apps among diabetes patients and diabetologists in China: national web-based survey. JMIR Mhealth Uhealth 2019 Feb 8;7(2):e12658. [doi: 10.2196/12658] [Medline: 30735147]
22. Lim SL, Tay MHJ, Ong KW, et al. Association between mobile health app engagement and weight loss and glycemic control in adults with type 2 diabetes and prediabetes (D'LITE Study): prospective cohort study. JMIR Diabetes 2022 Sep 30;7(3):e35039. [doi: 10.2196/35039] [Medline: 36178718]
23. Timpel P, Oswald S, Schwarz PEH, Harst L. Mapping the evidence on the effectiveness of telemedicine interventions in diabetes, dyslipidemia, and hypertension: an umbrella review of systematic reviews and meta-analyses. J Med Internet Res 2020 Mar 18;22(3):e16791. [doi: 10.2196/16791] [Medline: 32186516]
24. Pamungkas RA, Usman AM, Chamroonsawasdi K. A smartphone application of diabetes coaching intervention to prevent the onset of complications and to improve diabetes self-management: a randomized control trial. Diabetes Metab Syndr 2022 Jul;16(7):102537. [doi: 10.1016/j.dsx.2022.102537]
25. Battelino T, Alexander CM, Amiel SA, et al. Continuous glucose monitoring and metrics for clinical trials: an international consensus statement. Lancet Diabetes Endocrinol 2023 Jan;11(1):42-57. [doi: 10.1016/S2213-8587(22)00319-9] [Medline: 36493795]
26. Tong HL, Laranjo L. The use of social features in mobile health interventions to promote physical activity: a systematic review. NPJ Digit Med 2018;1(1):43. [doi: 10.1038/s41746-018-0051-3] [Medline: 31304323]
27. GPT-4. OpenAI. 2025. URL: https://openai.com/index/gpt-4-research/ [accessed 2025-02-24]
28. Grammarly. 2025. URL: https://app.grammarly.com/ [accessed 2025-02-24]

Abbreviations

HbA_{1c}: glycated hemoglobin
T2DM: type 2 diabetes mellitus

XSL•FO
RenderX

Edited by A Mavragani; submitted 14.02.25; peer-reviewed by C Baxter, L Roth; revised version received 11.04.25; accepted 23.04.25; published 20.05.25.

Please cite as:
Yoshihara S, Takahashi K, Kawaguchi H, Harai N, Tsuchiya K
Effects of Using a Digital Peer-Supported App on Glycated Hemoglobin Changes Among Patients With Type 2 Diabetes: Prospective Single-Arm Pilot Study
JMIR Form Res 2025;9:e72659
URL: https://formative.jmir.org/2025/1/e72659
doi:10.2196/72659

© Shota Yoshihara, Kayoko Takahashi, Hiroyuki Kawaguchi, Nozomi Harai, Kyoichiro Tsuchiya. Originally published in JMIR Formative Research (https://formative.jmir.org), 20.5.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Patient Acceptability and Technical Reliability of Wearable Devices Used for Monitoring People With Parkinson Disease: Survey Study

Tasmin Alanna Rookes[1], BSc, MSc; Amit Batla[2], MA, MBBS; Megan Armstrong[1], PhD; Gareth Ambler[3], BSc, PhD; Kate Walters[1], MBBS, MSc, PhD; Anette Schrag[2], PhD

[1]Research Department of Primary Care and Population Health, University College London, London, United Kingdom
[2]Department of Clinical and Movement Neurosciences, UCL Queen Square Institute of Neurology, University College London, London, United Kingdom
[3]Department of Statistical Science, University College London, London, United Kingdom

Corresponding Author:
Anette Schrag, PhD
Department of Clinical and Movement Neurosciences, UCL Queen Square Institute of Neurology, University College London, London, United Kingdom

Abstract

Background: Parkinson disease is a progressive neurodegenerative disorder with complex motor and nonmotor symptoms. To assess these, clinical assessments are completed, providing a snapshot of a person's experience. Monitoring Parkinson disease using wearable devices can provide continuous and objective data and capture information on movement patterns in daily life.

Objective: The aim of the study is to assess patient acceptability and technical reliability of 2 wearable devices used in clinical trials (ActivInsights and Axivity AX3).

Methods: Participants in a feasibility study testing a self-management toolkit (PD-Care) optionally wore a wearable device for 1 week, providing feedback through an open- and closed-question survey conducted over the telephone about the acceptability of wearing the device. The closed questions used a Likert scale from 1 to 5 (with 1=strongly agree and 5=strongly disagree) asking whether (1) the device was comfortable to wear, (2) the device was easy to put on, (3) the device was easy to wear, (4) the device was embarrassing to wear, and (5) if they were happy to wear the device for longer than 7 days. Differences in acceptability between devices were analyzed using Mann-Whitney *U* tests and Wilcoxon matched pairs signed rank tests. These were followed by open-ended questions asking (1) How did you find wearing the device? (2) How did you find putting the device on? (3) Did you take it off and why? (4) What was your overall impression? (5) Did you prefer the wrist- or trunk-worn device and why (Axivity AX3 only)?

Results: A total of 22 of 32 (69%) participants offered the device agreed to wear it. There were no significant differences in the demographic characteristics between those monitored and those who chose not to be. Acceptance with both devices was generally good. The ActivInsights device was more acceptable than the wrist- and trunk-worn Axivity AX3 devices, as more participants found it to be comfortable (n=15, 100% vs n=5, 71%; P=.02 and n=4, 57%; P=.004, respectively), easy to wear (n=15, 100% vs n=6, 86%; P=.048 and n=3, 43%; P=.004, respectively) and would wear for more than 7 days (n=13, 87% vs n=4, 57%; P=.02 and n=1, 14%; P<.001, respectively). The trunk-worn Axivity AX3 device had the lowest acceptance rates, but there were no statistical differences in acceptability between the wrist- and trunk-worn Axivity AX3 devices (all P>.05). There were issues with battery life and recording errors in 3 of 14 (21%) Axivity AX3 devices and upload failures in 3 of 15 (20%) ActivInsights devices.

Conclusions: Acceptability of wearables for monitoring Parkinson was satisfactory, especially when wrist-worn, although a few participants experienced difficulties in correct use, and there were some errors with the data upload.

Trial Registration: ISRCTN Registry ISRCTN92831552; https://www.isrctn.com/ISRCTN92831552

(JMIR Form Res 2025;9:e63704) doi:10.2196/63704

KEYWORDS

Parkinson disease; feasibility; remote monitoring; Parkinson; acceptability; reliability; wearable devices; wearable; self-management; quantitative assessments; quantitative; qualitative; monitoring; patient

Introduction

Parkinson disease (PD) is a neurodegenerative chronic disease, affecting around 145,000 people in the United Kingdom in 2018, equivalent to approximately 1 adult in every 350 [1]. The currently accepted monitoring of PD symptoms in clinical practice and trials is based on interviews and validated clinical scales and questionnaires [2]. However, the use of patient- and

XSL•FO
RenderX

clinician-rated scales is resource-intensive, and the results limited by subjectivity and inaccuracy and are either one-off assessments or dependent on the patient's recollection [3]. Continuous and accurate assessment of PD features, in both clinical practice and trials, would provide more relevant information and overcome the limitations of clinical scales and questionnaires through real-world data. With the increasing burden of care and limited medical resources, new assessment and monitoring approaches have the potential to improve clinical care and patient outcomes [4]. There are now a range of body-worn sensors using accelerometers or gyroscope devices (henceforth called wearable devices) that have been developed and validated to monitor parkinsonian clinical features, such as motor fluctuations, dyskinesia, tremor, bradykinesia, freezing of gait, or gait disturbances [2]. However, relatively little evidence exists on their use, acceptability, and feasibility in clinical practice and trials with patients with PD. In addition, the physical, technical, or practical limitations on prolonged use for real-time recording, uploading, and analysis of movement data are not clear [5].

The National Institute of Health and Care Excellence has conditionally recommended some wearable devices for remote monitoring of PD. However, the exact role, indication, and type of device is unclear, and evidence on the use of wearable devices is limited [6]. The cost, burden on patients and services, and effectiveness to improve clinical outcomes have not yet been assessed. National Institute of Health and Care Excellence has therefore recommended the collection of real-world evidence on devices that monitor people with PD [7] to enable the clinical and cost-effectiveness of the technologies to be fully assessed, ahead of further implementation into clinical practice. In practice, wearable devices are relatively rarely used due to several factors, such as unfamiliarity, unclear indications, cost, acceptability, and technical issues. Furthermore, although digital measures are considered to have high potential for future use to more accurately detect meaningful change in clinical trials, little data exist on their use and acceptability in clinical trials.

Data collection regarding utility and acceptability has largely been informal and focused on wearable devices in general [5]. A study in Finland reported that 88.9% (32/36) of patients with PD using a wearable device thought it was very easy or easy [8]. In the United States, the recent WATCH-PD study [9] explored the acceptability of wearables in people with PD and control participants over a 2-year window and found that participants with PD had generally positive views of the wearables, which were comparable to healthy controls. Conducting qualitative interviews with people with PD who had used multimodal sensors, including a wearable device, the wrist-worn wearable was the least acceptable, in comparison to cameras and passive ambient sensors [10]. Another small study explored perceptions of wearing devices in the home versus in public and found no differences in acceptability between these settings and acceptance of long-term symptom monitoring through wrist-worn devices [11]. However, these specific studies are likely to have recruited participants willing to wear a device, so there is a lack of real-world evidence from patients with PD wearing these devices in home and social settings. Eliciting opinions from patients on wearing such devices in real-life settings may provide important information to facilitate the implementation of wearables into clinical practice.

Here, we have used the opportunity to explore the views of participants taking part in a feasibility study from a large randomized controlled trial of a self-management tool for people with PD in the National Health Service (NHS) to obtain such information in a real-life setting. The objective of this study was to explore the acceptability and technical reliability of 2 wearable sensors that are commonly used by people with PD.

Methods

Overview

This study is reported in line with the CROSS checklist (Consensus-Based Checklist for Reporting of Survey Studies) [12] (Table S1 in Multimedia Appendix 1). We used a mixed methods survey design to obtain information from survey questions, both quantitative and semiqualitative, providing complementary information, within the context of a clinical feasibility trial in an NHS setting [13].

Study Design and Participant Recruitment

Participants in the PD-Care program feasibility study were offered to participate in this wearable substudy. The primary aim of the PD-Care program is to develop and trial a supported self-management tool for people with PD, the full methods of which have been described previously [13]. The study emphasizes inclusive participation with broad inclusion criteria, including no age limit, inclusion of participants with cognitive impairment as long as they have the capacity to consent, and in any stage of PD. Remote delivery via video or telephone calls increased accessibility to study participation. Following the development of the intervention, a planned feasibility study was conducted before the randomized controlled trial (ISRCTN92831552). The feasibility study included 35 participants with assessments at baseline and post-intervention at 3 months. Participants were recruited through general practice mailouts, secondary care neurology teams, and the neurology research register.

After the 3-month follow-up assessments, all participants were offered to participate in a substudy, wearing 1 of 2 devices for a one-off 7-day period. After wearing and returning the device, participants were asked to provide feedback on the acceptability of the device via a telephone survey. The researcher entered the participant's responses directly into the database while on the phone call. In addition, when downloading the data from the devices, the researcher recorded any errors that occurred. This substudy received input from the PD-Care patient and public engagement group, including where the device should be worn, how long it should be worn for, and which acceptability questions should be asked in the survey.

Devices

Two different devices, ActivInsights [14] and Axivity AX3 [15] were used. The choice of the 2 devices was guided by pragmatic reasons for the purpose of a larger, publicly funded trial in the NHS. Both devices have been used in previous clinical research in patients with PD [14,15]. Both devices were worn at the wrist.

XSL•FO
RenderX

In addition, as recommended by the patient and public advisory group, a sensor worn on the trunk was also tested to capture different clinical features. We therefore used 2 Axivity AX3 devices worn simultaneously, one on the wrist and the other on the trunk, whereas the ActivInsights device was worn on the wrist only.

The ActivInsights device provided output data on sleep, sedentary time, active time, and time spent exercising. The Axivity AX3 device output provided raw data on movement, recorded as a continuous wave accelerometer format, which was converted to a CSV file for analysis.

Data Collection

The wearable devices and instructions were sent to participants in the post, with a request to return in the post after 7 days using a stamped addressed envelope. Seven days was chosen due to the limit on the battery life of the Axivity AX3 device while still collecting enough data to have an average over several days. Participants were allocated to a device depending on what was available once they had completed study follow-up assessments. We had 8 wearable devices of each ActivInsights and Axivity AX3, but as 2 devices were needed for each participant wearing Axivity AX3, fewer participants used this device. The date to start and stop recording for participants was preprogrammed into the device before posting out, and participants were asked to start and stop wearing the device on a specific date and time. Both devices recorded continuously, and we asked participants to inform us of the 7-day time period during which they wore the device, depending on when it arrived. No feedback from the devices was provided to participants. Data upload was undertaken by the researchers upon the return of the devices.

Outcome Measures

Acceptability

The overall rate of participation in the substudy was recorded. After wearing and returning the device, participants were contacted via telephone to provide feedback about the acceptability of using their wearable device through a survey. We captured quantitative data to enable inferential statistical tests to be conducted to compare between devices as well as open-ended questions to ensure that the nuance of people's individual experiences was understood.

The open-ended questions were (1) How did you find wearing the device? (2) How did you find putting the device on? (3) Did you take it off and why? (4) What was your overall impression? (5) Did you prefer the wrist- or trunk-worn device (Axivity AX3 only) and why? The quantitative assessments included questions using a Likert scale from 1 to 5 (1=strongly agree and 5=strongly disagree) whether (1) the device was comfortable to wear, (2) the device was easy to put on, (3) the device was easy to wear, (4) the device was embarrassing to wear, and (5) if they were happy to wear the device for longer than 7 days.

The survey was designed by the research team to cover the core components of wearable acceptance while being short and quick to administer. The survey was not based on any specific framework or guideline. A pragmatic approach with short survey questions was adopted to minimize participant burden, to reduce the risk of attrition in the context of the overall feasibility study, which involved collecting a wide range of outcome measures.

Technical Reliability

We assessed whether issues of battery shortage, recording failure, or upload difficulties impacted on the availability of usable data.

Data Analysis

To ensure that the participants in the sample did not differ from those who chose to wear an activity monitor, the Fisher exact test and 2-tailed *t* test comparisons were conducted on demographic characteristics. To explore differences in acceptability between the different monitors, Mann-Whitney *U* tests and Wilcoxon matched pairs signed rank tests were conducted. All analyses were conducted in Stata (version 17; Stata Corp 2021). Open-ended question answers were written out, as participants provided their answers verbally, and responses were synthesized through content analysis to identify common concepts and views between participants.

Ethical Considerations

Participants were taking part in the feasibility study of the PD-Care trial (ISRCTN92831552) and provided optional consent to take part in this wearable substudy. This substudy received a favorable opinion by the London Queen Square Research Ethics Committee and Health Research Authority approval (18/LO/1470) on October 31, 2018, as part of the feasibility study application. All participant's data were pseudoanonymized through unique study ID numbers, and all study documentation, including wearable data and survey responses, were stored against this ID. Participants received a US $26 voucher for completing baseline assessments and US $13 voucher for completing 3-month follow-up assessments, as part of the feasibility study. No additional compensation was provided for taking part in this wearable substudy.

Results

Participants

Between October 2020 and February 2021, 35 participants with a clinical diagnosis of PD and living in a community setting were recruited to the PD-Care program from several secondary care sites in the United Kingdom. Of these, 1 moved permanently to a care home and was no longer eligible, and 2 were lost to follow-up in the overall study. The remaining 32 were offered participation in the device substudy. The average age of the 32 participants invited to wear a wearable device was 68.2 (SD 11.3) years, 11 (34%) were female, and 23 (72%) were White British. Of these, 22 (69%) agreed to wear a device for 7 days. Reasons for not participating were participants feeling burdened from the feasibility study and did not want to provide more data (n=7), participants had a deep brain stimulation device fitted and were uncertain about potential interactions with the wearable (n=2), and finding it embarrassing to wear (n=1) (Figure 1).

XSL•FO
RenderX

Figure 1. Flowchart of participants in the study. DBS: deep brain stimulation.

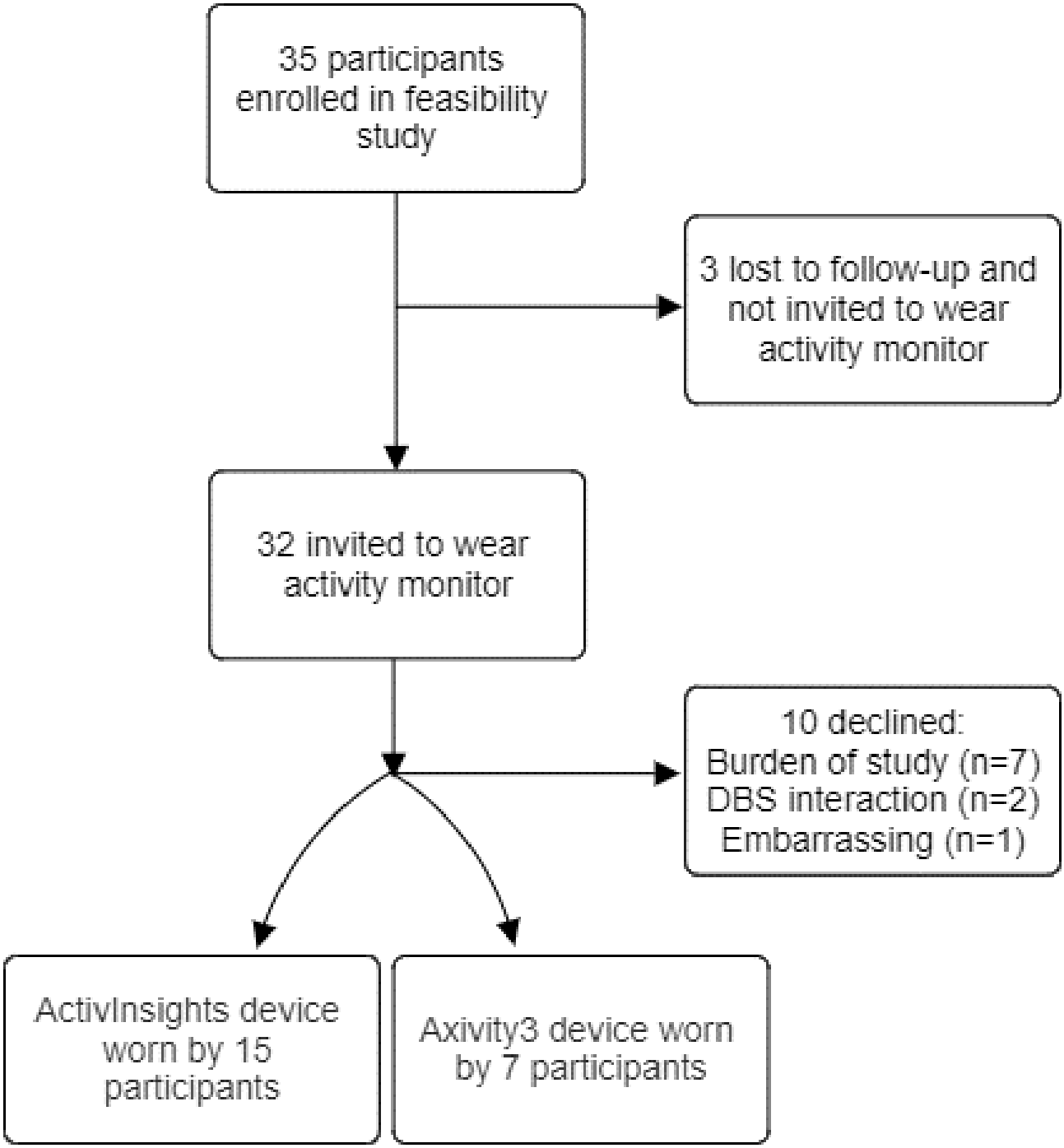

There were no statistically significant differences in age, sex, or disease duration or severity between participants and nonparticipants (Table S2 in Multimedia Appendix 1). In total, 15 of the 22 participants were allocated the ActivInsights device (wrist only) and 7 were asked to wear the Axivity AX3 device (wrist and trunk). There were no statistically significant differences in the characteristics of those who wore the Axivity AX3 and those who wore the ActivInsights devices (Table S3 in Multimedia Appendix 1).

Patient Acceptability

ActivInsights

Feedback to open-ended questions suggested that the device was easy to use, and participants quickly forgot they were wearing it (n=11). Some people commented that the strap was difficult to put on and needed help (n=5), and others suggested that it would have been nice if the device had a screen that gave them some feedback about their activity levels (n=3). Quantitative feedback suggested that overall the device was comfortable, easy to wear, and not embarrassing, and participants would be happy to wear it for longer than 7 days (Tables 1 and 2 and Multimedia Appendix 2).

XSL•FO
RenderX

Table . People with Parkinson disease who responded in a positive way to the survey questions about their experience of wearing different wearable devices in the home for 1 week.

Device	It was comfortable, n (%)	It was easy to put on, n (%)	It was easy to wear, n (%)	It was embarrassing to wear[a], n (%)	I would wear it for more than 7 days, n (%)
Axivity AX3 wrist (n=7)	5 (71)	5 (71)	6 (86)	7 (100)	4 (57)
Axivity AX3 trunk (n=7)	4 (57)	2 (29)	3 (43)	7 (100)	1 (14)
ActivInsights wrist (n=15)	15 (100)	10 (67)	15 (100)	15 (100)	13 (87)

[a]Percentage inversed.

Table . Mann-Whitney *U* tests comparing the acceptability responses between ActivInsights and both Axivity AX3 wrist and trunk and Wilcoxon matched pairs signed rank test comparing responses between Axivity AX3 wrist and trunk.

Device comparison	It was comfortable, *P* value	It was easy to put on, *P* value	It was easy to wear, *P* value	It was embarrassing to wear, *P* value	I would wear it for more than 7 days, *P* value
ActivInsights versus Axivity AX3 wrist	.02	.72	.048	N/A[a]	.02
ActivInsights versus Axivity AX3 trunk	.004	.38	.004	N/A	<.001
Axivity AX3 Wrist versus Axivity AX3 trunk	.50	.50	.25	N/A	.25

[a]N/A: not applicable.

Axivity AX3

All participants reported that they preferred the wrist-worn device to the trunk-worn device. Problems with the trunk-worn device centered around the tape not holding the device in place and having to be reapplied (n=7). This led to 3 of 7 participants not wearing the trunk-worn device but still provided feedback on both devices. The wrist-worn device was acceptable to all, with only a few comments about it being bulky and uncomfortable (n=3) and 1 participant wanted feedback on what the device was recording.

Quantitative feedback suggested that, on average, participants were neutral to the trunk-worn device being comfortable, easy to put on, and easy to wear but disagreed with being happy to wear it for longer than 7 days. For the wrist-worn wearable device, participants on average agreed that it was easy to put on and wear and were neutral about its comfort and whether they would wear it for more than 7 days (see Table 1 and Multimedia Appendix 2).

Patient Acceptability—Comparison Between Groups

Using Mann-Whitney *U* tests to compare the acceptability of the devices, overall, the ActivInsights device was more acceptable than both the Axivity AX3 wrist- and trunk-worn devices. The ActivInsights device was rated as more comfortable than the wrist-worn and trunk-worn Axivity AX3 devices (n=15, 100% vs n=5, 71%; *P*=.02 and n=4, 57%; *P*=.004, respectively), rated as easier to wear (n=15, 100% vs n=6, 86%; *P*=.048 and n=3, 43%; *P*=.004, respectively), and rated as more suitable to wear more than 7 days (n=13, 87% vs n=4, 57%; *P*=0.02 and n=1, 14%; *P*<.001, respectively). Using Wilcoxon matched pair signed rank tests, there were no significant differences between the Axivity AX3 trunk-worn and wrist-worn device, but the numbers were small. There were no significant differences between all 3 devices for the ease to put them on (Table 2) and for embarrassing to wear, and all participants responded positively, stating the devices were not embarrassing.

Technical Reliability

ActivInsights

There were no issues reported regarding battery shortage or recording failure. There were data upload errors with 3 (20%) of the devices, whereby the system failed when trying to upload the data, and when corrected, the data had been wiped from the device. Usable data were therefore available in 12 of 15 (80%) participants.

Axivity AX3

Of the 11 devices worn by 7 participants (7 wrist and 4 trunk), 6 (3 participants) were worn for shorter durations, resulting in missing data, for a median of 27 (IQR 8 - 48) hours, and this was due to hospitalization (n=1) and participant forgetting (n=2). In 3 other devices, the device ran out of battery during the data collection period. Despite the errors, some usable data was available for all participants with this device.

XSL•FO
RenderX

Discussion

Principal Findings

We assessed the acceptability and technical feasibility of 2 wearable devices used in trials with people with PD, which showed that wearable devices are acceptable to people with PD and feasible to administer and collect data with. Some data errors were found due to battery life, participant utility, and upload issues. Although several studies using wearable devices have reported validity for recording human movement in a free-living environment [16], including in people with PD [2,17,18], successful clinical adoption in monitoring movement in PD depends on acceptability and technical reliability [5,17]. To date, there is no fully validated system to monitor clinical features or activities in PD [5]. The bulk of data on technical reliability with the use of individual wearable devices comes from the studies led or supported by manufacturers of these devices [16]. This is the biggest study to date that reports the feedback of patients with PD from wearing 2 different monitoring devices using open and closed questions.

Patient Acceptability

Feedback from patients suggests that monitoring using both ActivInsights and Axivity AX3 is mostly acceptable, although there is a proportion of patients who do not wish to use a monitoring device due to burden and embarrassment. The burden was due to participants having to complete clinical assessments as part of the feasibility study. If the participants had had the option of clinical assessments or activity monitors, then uptake may have been higher. There were no demographic characteristics that differed between the participants who chose to be monitored and those who chose not to be monitored, including sex, age, and deprivation level. Largely, our findings are in line with those from a US population with PD, using a different device, suggesting ease of use and willingness to wear the device over a longer period when located on the wrist [9]. In addition, these findings reflect acceptability in other UK patient groups, such as wrist-worn alcohol sensors in those who are alcohol dependent, where people were positive about the devices, despite the perceptions of health care professionals that service users may have challenges using them [19].

We did not specifically examine motivational factors, but another study has reported on the motivations for and barriers against monitoring PD symptoms among patients with PD [20]. Key motivations include the wish to discuss findings with health care providers, obtain insight into the effect of medication and other treatments, and follow the progression of the disease. Key barriers were not wanting to focus too much on having PD, symptoms being relatively stable, and lacking an easy-to-use tool [17]. As all participants in our study were monitored as part of a research intervention, these motivating factors mentioned may have been less directly relevant. As no clinical feedback or reports were provided to participants, and the purpose was to supplement clinical assessments in a trial, the setting and motivation will differ from those who would consider monitoring for a clinical purpose. Nevertheless, 4 (18%) of the participants expressed the wish to know what was being monitored.

Social embarrassment and the feeling that wearing a device for monitoring signaled a person as “old” have been highlighted as major factors affecting the acceptability of body-worn sensors [11]. Similarly, the use of a wearable device was embarrassing in some of our study participants, resulting in them declining participation in this substudy. In another study using Axivity AX3 on each wrist, the participants wore the device for 4 hours in a research facility and then for 1 week at home. Overall, the study reported that monitoring with body-worn sensors was acceptable to patients with PD and that most participants were willing to wear the sensor both at home and in public [11]. Concerns regarding safety particularly with deep brain stimulation could also be an additional barrier that could be resolved with better communication and technical support.

In addition, overall, the wrist-worn devices were more acceptable than the trunk-worn devices, as many participants found the trunk-worn device to be uncomfortable and difficult to keep on. However, there may also be issues with wrist-worn devices. A previous study using Axivity AX3 on each wrist reported that, for prolonged wearing, participants were less likely to agree that the sensors were comfortable to wear [11]. Responses to open-ended questions revealed that the main source of sensor discomfort related to the strap [11]. Though the acceptance of the strap on the 7-day monitoring was not a problem in our cohort, 23% (n=5) needed help with the strap to put the device on. Ideally, if the wearable device is to be worn by a patient with PD, they should be able to do this independently despite poor dexterity.

Technical Reliability

The charging and preparation of the devices takes time and effort, with human errors leading to incomplete charging of the devices, and some participants forgetting to turn on the devices. The battery life on a single charge may not be long enough for long-term monitoring. There were also upload errors, with the upload failing in one-fifth of the participants using the ActivInsights device. Almost all wearable devices rely on a local computer interface, a stable internet connection for upload, and often an external storage system before being analyzed [2]. The upload to a cloud server relies on a good internet connection and can be a limiting factor in some settings [4]. Such technical problems not only add to costs and time pressure but can also lead to further anxiety for participants who might already have nonmotor symptoms of anxiety. More data are needed on failure rates with individual devices. On most occasions, device failure was identified after the recording had been completed, and the wearable devices had been sent back. Ideally, troubleshooting tools or a live screen display, as 4 of our participants suggested, could be more acceptable and help avoid recording failures. Furthermore, additional training of research staff and patients to ensure data upload could be helpful to overcome this issue.

Study Limitations and Future Research

The duration of monitoring used in this study was short, and we had a small sample size. Studies with larger sample sizes and longer monitoring periods are needed. As part of the PD-Care program of research, a large randomized controlled trial is now underway including a substudy using wearable devices in people with PD. These data will be used to further

XSL•FO
RenderX

validate the use of wearable devices as a clinical measure of PD symptoms as well as confirm the acceptability findings found in this feasibility study. Researchers should also explore differences in acceptability in people from underserved groups to ensure that the implementation of these devices does not widen health inequalities.

Conclusions

The study highlights some issues that can be addressed by device manufacturers such as difficulty with strapping single-handedly in someone with limited dexterity, help in uploading data, and clearer instructions. We found better acceptability of the ActivInsights device by our participant cohort, but there was a higher rate of usable recording when using Axivity AX3. The prolonged use of devices worn on the trunk was less acceptable. More real-world experience, including patient feedback and multidisciplinary coordinated effort involving all key stakeholders, is critical before the widespread adoption of wearable devices for monitoring PD in clinical trials.

Acknowledgments

The wider PD-Care team and collaborators involved in the development of the project. The authors thank the participants for their ongoing support of our research. The study received funding from National Institute for Health and Care Research program (grant RP-PG-1016-20001). Generative artificial intelligence was not used in any portion of the manuscript writing.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The CROSS (Consensus-Based Checklist for Reporting of Survey Studies) checklist and demographic characteristic comparisons.
[DOCX File, 27 KB - formative_v9i1e63704_app1.docx]

Multimedia Appendix 2
Graph showing responses to each question, for each device.
[DOCX File, 27 KB - formative_v9i1e63704_app2.docx]

References

1. The incidence and prevalence of Parkinson's in the UK: results from the Clinical Practice Research Datalink Summary report. Parkinson's UK. 2018. URL: https://www.parkinsons.org.uk/sites/default/files/2018-01/CS2960%20Incidence%20and%20prevalence%20report%20branding%20summary%20report.pdf [accessed 2025-03-10]
2. Goetz CG, Tilley BC, Shaftman SR, et al. Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results. Mov Disord 2008 Nov 15;23(15):2129-2170. [doi: 10.1002/mds.22340] [Medline: 19025984]
3. AlMahadin G, Lotfi A, Zysk E, Siena FL, Carthy MM, Breedon P. Parkinson's disease: current assessment methods and wearable devices for evaluation of movement disorder motor symptoms—a patient and healthcare professional perspective. BMC Neurol 2020 Nov 18;20(1):419. [doi: 10.1186/s12883-020-01996-7] [Medline: 33208135]
4. Moreau C, Rouaud T, Grabli D, et al. Overview on wearable sensors for the management of Parkinson's disease. NPJ Parkinsons Dis 2023 Nov 2;9(1):153. [doi: 10.1038/s41531-023-00585-y] [Medline: 37919332]
5. Del Din S, Godfrey A, Mazzà C, Lord S, Rochester L. Free-living monitoring of Parkinson's disease: lessons from the field. Mov Disord 2016 Sep;31(9):1293-1313. [doi: 10.1002/mds.26718] [Medline: 27452964]
6. Laar A, Silva de Lima AL, Maas BR, Bloem BR, de Vries NM. Successful implementation of technology in the management of Parkinson's disease: barriers and facilitators. Clin Park Relat Disord 2023;8:100188. [doi: 10.1016/j.prdoa.2023.100188] [Medline: 36864905]
7. Devices for remote monitoring of Parkinson's disease. NICE. 2023. URL: https://www.nice.org.uk/guidance/dg51 [accessed 2025-03-05]
8. Liikkanen S, Sinkkonen J, Suorsa J, et al. Feasibility and patient acceptability of a commercially available wearable and a smart phone application in identification of motor states in parkinson's disease. PLOS Digit Health 2023 Apr;2(4):e0000225. [doi: 10.1371/journal.pdig.0000225] [Medline: 37027348]
9. Kangarloo T, Latzman RD, Adams JL, et al. Acceptability of digital health technologies in early Parkinson's disease: lessons from WATCH-PD. Front Digit Health 2024 Aug 26;6:1435693. [doi: 10.3389/fdgth.2024.1435693] [Medline: 39253055]

XSL•FO
RenderX

10. Morgan C, Tonkin EL, Craddock I, Whone AL. Acceptability of an in-home multimodal sensor platform for Parkinson disease: nonrandomized qualitative study. JMIR Hum Factors 2022 Jul 7;9(3):e36370. [doi: 10.2196/36370] [Medline: 35797101]
11. Fisher JM, Hammerla NY, Rochester L, Andras P, Walker RW. Body-worn sensors in Parkinson's disease: evaluating their acceptability to patients. Telemed J E Health 2016 Jan;22(1):63-69. [doi: 10.1089/tmj.2015.0026] [Medline: 26186307]
12. Sharma A, Minh Duc NT, Luu Lam Thang T, et al. A Consensus-Based Checklist for Reporting of Survey Studies (CROSS). J Gen Intern Med 2021 Oct;36(10):3179-3187. [doi: 10.1007/s11606-021-06737-1] [Medline: 33886027]
13. Walters K, Armstrong M, Gardner B, et al. Clinical and cost-effectiveness of "Live Well with Parkinson's" self-management intervention versus treatment as usual for improving quality of life for people with Parkinson's: study protocol for a randomised controlled trial. Trials 2023 Dec 5;24(1):793. [doi: 10.1186/s13063-023-07700-7] [Medline: 38053209]
14. ActiveInsights band. ActiveInsights. URL: https://activinsights.com/technology/activinsights-band/ [accessed 2024-06-27]
15. AX3 band. Axivity. URL: https://axivity.com/product/ax3 [accessed 2024-06-27]
16. Narayanan A, Stewart T, Mackay L. A dual-accelerometer system for detecting human movement in a free-living environment. Med Sci Sports Exerc 2020 Jan;52(1):252-258. [doi: 10.1249/MSS.0000000000002107] [Medline: 31361712]
17. Maetzler W, Rochester L. Body-worn sensors—the brave new world of clinical measurement? Mov Disord 2015 Aug;30(9):1203-1205. [doi: 10.1002/mds.26317] [Medline: 26173577]
18. Morgan C, Rolinski M, McNaney R, et al. Systematic review looking at the use of technology to measure free-living symptom and activity outcomes in Parkinson's disease in the home or a home-like environment. J Parkinsons Dis 2020;10(2):429-454. [doi: 10.3233/JPD-191781] [Medline: 32250314]
19. Brobbin E, Parkin S, Deluca P, Drummond C. A qualitative exploration of the experiences of transdermal alcohol sensor devices amongst alcohol service practitioners (South London, UK). Addict Res Theory 2025;33(1):36-47. [doi: 10.1080/16066359.2024.2309869] [Medline: 39906322]
20. Evers LJW, Peeters JM, Bloem BR, Meinders MJ. Need for personalized monitoring of Parkinson's disease: the perspectives of patients and specialized healthcare providers. Front Neurol 2023;14:1150634. [doi: 10.3389/fneur.2023.1150634] [Medline: 37213910]

Abbreviations

CROSS: Consensus-Based Checklist for Reporting of Survey Studies
NHS: National Health Service
PD: Parkinson disease

Edited by A Mavragani; submitted 27.06.24; peer-reviewed by G Ymeri, J Ruokolainen, S Parkin; revised version received 31.01.25; accepted 03.02.25; published 25.03.25.

Please cite as:
Rookes TA, Batla A, Armstrong M, Ambler G, Walters K, Schrag A
Patient Acceptability and Technical Reliability of Wearable Devices Used for Monitoring People With Parkinson Disease: Survey Study
JMIR Form Res 2025;9:e63704
URL: https://formative.jmir.org/2025/1/e63704
doi:10.2196/63704

© Tasmin Alanna Rookes, Amit Batla, Megan Armstrong, Gareth Ambler, Kate Walters, Anette Schrag. Originally published in JMIR Formative Research (https://formative.jmir.org), 25.3.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Using Text Messaging Ecological Momentary Assessment to Record Changes in e-Cigarette and Combustible Cigarette Use: Pilot Randomized Clinical Trial

Tucker Morgan[1], BBA, MS; Michelle He[1], BS, MS; Andrew Nicholson[1], BS, MSPH; Omar El Shahawy[1,2], MD, PhD; Scott E Sherman[1,3], MPH, MD; Elizabeth R Stevens[1], MPH, PhD

[1]Department of Population Health, Grossman School of Medicine, New York University, 180 Madison Ave, New York, NY, United States

[2]School of Global Public Health, New York University, New York, NY, United States

[3]Department of Medicine, VA New York Harbor Health System, Veterans Health Administration, New York, NY, United States

Corresponding Author:
Elizabeth R Stevens, MPH, PhD
Department of Population Health, Grossman School of Medicine, New York University, 180 Madison Ave, New York, NY, United States

Abstract

Background: Ecological momentary assessment (EMA) provides insight into the effectiveness and feasibility of smoking-related interventions.

Objective: The objective of this paper was to assess adherence to an EMA protocol and compare EMA-derived responses with measures collected through multiple surveys.

Methods: A subanalysis was conducted using data from a 12-week, open-label, and 2-arm pilot randomized clinical trial among adult smokers with chronic obstructive pulmonary disease, coronary artery disease, peripheral vascular disease, or asthma in the last 12 months (n=109). Participants were randomized to either electronic cigarette (EC) or nicotine replacement therapy (NRT) treatment arms. We compared EMA data collected through automated SMS text message prompts sent to participants 4 times daily for 12 weeks, including cigarettes smoked per day (CPD), craving, and satisfaction, to survey data collected at 12 weeks. Convergent validity between survey- and EMA-reported measures was evaluated using Pearson correlation and paired *t* tests. CPD was modeled using negative binomial regression. Relative rates (RRs) of reaching at least 50%, 75%, and 100% CPD reduction between two arms were calculated using both EMA and survey data.

Results: The majority of participants were non-Hispanic White (63/109, 58%) and female (60/109, 55%), and had a median age of 60 (IQR 54 - 65) years. Among the 109 participants, 59.6% (n=65) were consistently adherent to the EMA protocol over the 12-week period. Median weekly EMA response rate remained high over the 12-week study period even though a modest decline was observed (week 1, 97.8% and week 12, 89.4%). The mean CPD declined significantly (week 1, mean 14.2, SD 9.9 and week 12, mean 4.6, SD 6.7; $P<.001$). EMA-derived and survey-based CPD measurements were positively correlated ($r=0.73$, 95% CI 0.6-0.82) as were measures of craving ($r=0.38$, 95% CI 0.17-0.56). No significant paired difference in CPD was observed between EMA measurements and surveys. A significant effect of time on CPD EMA data (incidence rate ratio [IRR] 1-week change 0.93; $P<.01$) and survey data was found (IRR 12-week change 0.36; $P<.01$). However, the treatment effect was not significant, which aligned with the RR results. An increase in the EC consumption was observed over time in the EC arm, with 12.1% (7/58) fully switched to EC (defined as CPD=0 and EC use>0) and 20.7% (12/58) mostly switched (defined as a reduction in CPD>75% and EC use>0) in week 12.

Conclusions: EMA is a suitable method to collect recall-based smoking-related data. Though results from mixed effect modeling and RR comparisons were similar using EMA or survey data, EMA provides unique advantages, namely greater granularity in the time and the capability to detect switching patterns in near real time. These findings provide the feasibility of using EMA in developing smoking cessation interventions in future tobacco harm reduction research.

Trial Registration: ClinicalTrials.gov NCT04465318; https://clinicaltrials.gov/study/NCT04465318

International Registered Report Identifier (IRRID): RR2-10.1186/s13722-021-00284-0

(JMIR Form Res 2025;9:e66709) doi:10.2196/66709

KEYWORDS

smoking; electronic cigarette; smoking cessation; ecological momentary assessment; EMA; adherence; real-time data; mobile phone; cigarette; smoker; e-cigarettes; pilot randomized clinical trial; clinical trial; pilot study; adult smoker; chronic condition

XSL•FO
RenderX

Introduction

In the United States, smoking remains the leading preventable cause of death, responsible for 480,000 deaths and US $170 billion in health care spending each year [1]. Despite considerable progress in the 20th century, in 2021, 30 million US adults were still cigarette users [2]. The benefits of smoking cessation are well-established [1]; however, there are high rates of relapse with existing smoking cessation therapies in some groups, like those with chronic obstructive pulmonary disease (COPD) [3]. Alternate smoking cessation tools, like electronic cigarettes (ECs), have been proposed as a more appealing harm reduction intervention for people who smoke [4-6].

In public health research, a primary interest is real-world behavior; however, this can be difficult to accurately assess in a laboratory or survey setting where measures, such as cigarettes smoked per day (CPD), typically rely on retrospective self-reports. Ecological momentary assessment (EMA), which is a data collection method that uses repeated sampling to collect data throughout a participant's everyday life, allows for the reporting of events in real-world settings and is much less time-consuming compared to traditional study visits [7]; EMA can limit issues associated with recall bias, particularly when participants are asked to recall the frequency of events over a long period [8]. Furthermore, EMA is often done remotely, thus serving as a useful tool for researchers considering remote assessments for their studies.

Interval-based EMA data collection methods have been used widely to study smoking habits and factors affecting use [9,10]. EMA data can provide useful insight into the changes in cigarette use and other factors, such as craving and satisfaction, over time and with greater granularity than study visits that are typically weeks or several months apart. Switching effects from combustible cigarettes (CCs) to ECs over an extended period, which has previously been observed through more traditional means of data collection [11], may be better understood using EMA data, particularly for vulnerable smokers with COPD and other chronic conditions.

This study assesses the comparability of smoking-related EMA measurements—CPD, craving, and satisfaction—with survey assessments completed as part of a 2-arm pilot randomized clinical trial (RCT). We also assessed the association between participant characteristics and adherence to the EMA protocol, visualized changes in EC use and CPD during program participation, and quantified the percent of individuals who switched completely to ECs over time. Furthermore, we compared the results of mixed effects negative binomial regression and relative rate comparisons of CPD reduction using EMA data versus using study visit data.

Methods

Study Setting and Participants

Participants were recruited via emails and phone calls between October 2020 and September 2022 from the electronic medical records of the New York University Langone Health (NYULH) system, a private hospital system serving New York, New Jersey, and Connecticut with approximately 6.8 million active patients. Patients were eligible for participation if they (1) had an ambulatory *International Statistical Classification of Diseases, Tenth Revision* (*ICD-10*) diagnosis code for COPD, cardiovascular disease, or asthma in the last 12 months; (2) for COPD patients only, a COPD Assessment Test (CAT) score ≥10; (3) were aged 21 to 75 years (the legal age for purchasing ECs in New York is 21); (4) were a current CC user (≥5 packs in a lifetime, smokes ≥4 days/week); (5) smoked at least 5 CCs per day on days they smoke CCs; (6) were motivated to quit smoking (at least a 5 on a 10-point Contemplation Ladder) [12]; and (7) possessed a phone with SMS text messaging capabilities. Potential participants were excluded if they (1), for COPD patients only, had a CAT score ≥30 (representing severe COPD) [13]; (2) reported NRT or EC use in the last 14 days; (3) had a medical condition (eg, unstable angina or heart disease) precluding use of nicotine patch or gum as determined by the study physician or by their treating physician; or (4) were pregnant or breastfeeding. Individuals who were not open to using NRT or ECs were also excluded from the study. Participants using other tobacco products were not excluded from this study.

Ethical Considerations

The NYU Langone Health Institutional Review Board (IRB) approved the study procedures (s20-00839) and written documentation of informed consent was received before starting data collection. The original informed consent allows for secondary analyses without additional consent. All participant data were deidentified prior to analysis. Participants did not receive any incentive for counseling calls but were incentivized to complete study visits and respond to the daily SMS text message smoking diary. Participants received US $10 for completion of a survey at baseline and at 6, 8, and 12 weeks, and US $20 for completion of a survey at 24 weeks. Participants could receive up to US $40 for completing a daily smoking diary via SMS text messages (US $20 if at least 60% of assessments are completed and US $40 if at least 80% are completed). In total, participants had the possibility of receiving a total of US $100 in incentives.

Study Design

This is a subanalysis of a 2-arm, open-label pilot RCT [14] designed to evaluate the harm-reducing effects of ECs in patients with COPD, coronary artery disease (CAD), peripheral vascular disease (PVD), or asthma. The methods of the parent study are detailed fully elsewhere [14]. Briefly, study participants were recruited from two large urban health centers' electronic health records (EHRs) and randomized in a 1:1 ratio to either standard smoking cessation care (nicotine replacement therapy [NRT] and counseling) or ECs and counseling [15]. Randomization was stratified by sex and baseline CPD (<20 vs ≥20). Participants were screened via phone or an internet-based screening tool before enrollment.

Participants completed study visits by telephone at baseline and at 1, 2, 3, and 6 months, with the intervention period lasting 3 months. At each visit, participants completed self-report survey assessments and were asked about their issued product supply and potential harm. Sociodemographic characteristics such as

XSL•FO
RenderX

age, race, ethnicity, place of birth, gender, marital status, education, work status, living situation, and homelessness were measured via self-report at baseline. In addition, condition-specific assessments (CAT [16], Clinical COPD Questionnaire [CCQ] [17,18], Seattle Angina Questionnaire-7 [SAQ-7] [19], Asthma Control Test [ACT] [20,21]) were also assessed as was motivation and confidence to quit smoking, tobacco use, and craving via the Minnesota Nicotine Withdrawal Scale (MNWS) [22].

Based on EHR diagnoses, participants were characterized into the following groups: COPD, asthma, and CAD or PVD. Participants did not receive any incentive for counseling calls, but were incentivized to complete study visits and respond to the daily SMS text message smoking diary.

EMA Protocol

In addition to study visit surveys, during the intervention period, participants were asked to record daily CC and EC use along with measures for CC craving and EC and CC satisfaction by SMS text messaging. Since the EMA protocol was embedded within the RCT, it allowed us to directly compare the message-derived outcomes with the survey assessments across intervention groups. The details of branching logic on the texting sequence participants would receive and the phrasing of each question are presented (Figure S1 in Multimedia Appendix 1). Certain responses triggered the follow-up questions about daily use, satisfaction, and craving. If the participant reported no CC or no EC use, only craving questions were triggered. Only participants in the EC arm received questions about EC use. Due to discordance between NRT use instructions (ie, 24 h patch) and the scheduled check-ins, the NRT arm was not asked about NRT use via SMS text message check-ins. This procedure employed time-based sampling with participants receiving SMS text message prompts four times per day, breaking the day into “Early, Lunch, Afternoon, Evening” periods such that approximately 4 hours had passed between check-ins.

Participants received brief instructions sent via an introductory text and response indicators in the body of SMS text message prompts but no formal training; of note, the EMA protocol was explained to participants during the baseline and consent process. Participants were given the opportunity to opt out of receiving SMS text messages before initiation of participation and during the study. Automated SMS text message prompts were sent to participants daily over the 12-week intervention.

EMA Data Processing and EMA-Based Measures

The SMS text message data were used to determine daily and weekly summaries for the following measures: response rate, number of CCs, EC uses (EC arm only), craving, satisfaction with the last CC, and satisfaction with the last EC (EC arm only).

Response Rate and CPD

We tabulated a participant’s weekly response rate as the number of responses from the participant divided by the total number of response-required texts (ie, those texts soliciting a response from the participant). To obtain an estimate of CPD—one that would be comparable to the traditional CPD measure recorded at study visits at baseline and week 12—we calculated the weekly average CPD by dividing the total number of reported cigarettes smoked in a week by 7 days. Participants with weekly response rates lower than 60% were removed from the calculation to obtain a more stable estimate of CPD and reduce downward bias. When calculating CPD for week 12, corresponding to the 3-month follow-up visit, the average CPD for the final week of EMA data collection was used rather than a total for the final day alone; we believe this average better approximates what participants would report in a visit setting compared with a 1-day total.

EC-Specific Measures

For the EC group, we used responses to the prompt, “Since your last report, how many separate times did you use an e-cig?” to calculate weekly averages for EC use. We defined mostly switched as >75% reduction in CPD from week 1 and use of ECs. We defined switching to ECs as those reporting 0 CPD smoked and EC use>0 in a given week.

Satisfaction and Craving

We calculated the average craving (range of 0 - 9 with higher values corresponding to higher craving) by averaging a person’s responses to the craving question (“On a scale of 0 - 9, how much do you want to smoke a cigarette right now?”) within a week. Similarly, we calculated the average satisfaction with the last cigarette and satisfaction with the last EC (range of 0 - 9 with higher values corresponding to greater satisfaction) by averaging a person’s responses to the satisfaction question (“On a scale of 0 - 9, how satisfying was your last cigarette/e-cig?”) within a week.

Statistical Analysis

To evaluate predictors of EMA adherence, participants were divided into two groups: (1) those with a weekly response rate >60% or more throughout the duration of the study period and (2) those with a weekly response rate <60% at any point in the study period. Descriptive statistics were calculated by group, including sociodemographic characteristics (age, gender, race and ethnicity, marital status, education, and work status) as well as condition-specific assessments, motivation, and confidence to quit smoking and tobacco use for each group. For variables not normally distributed, median and IQR were presented. Characteristics were evaluated using Wilcoxon rank sum test for continuous variables and Fisher exact test for categorical variables.

To assess the comparability between EMA measures of CPD and more traditional measurements gathered in a survey setting, we examined the Pearson correlation coefficient between CPD reported during the final week of EMA prompting (week 12) and CPD reported at the 3-month survey. We also examined the paired difference between EMA- and survey-reported CPD at 3 months using a paired *t* test. Pearson correlation coefficient was also applied to assess the convergent validity of EMA measurements for craving compared with survey-reported measures (first item of the Minnesota nicotine withdrawal scale: “Do you have a craving for cigarettes or desire to smoke?” with responses coded from 0 - 4, corresponding to “Not present” [0], “Slight” [1], “Mild” [2], “Moderate” [3], and “Severe” [4])

XSL•FO
RenderX

[23]. The EMA prompt “on a scale of 0 - 9, how much do you want to smoke a cigarette right now?” was used as the comparator for craving with higher numbers corresponding to greater craving.

To assess our primary outcome of change in CPD, we used a negative binomial mixed effects model, adjusting for the intervention arm, time (weeks from baseline), and a random intercept for each individual. We compared the results obtained from surveys by building a similar model, with the time variable expressed as time from baseline to week 12 visit. The negative binomial model has been shown to be more effective for count data than a Poisson model in the presence of overdispersion [24]. CPD reduction rate was evaluated by measuring the change in CPD from the baseline to week 12 from both EMA and survey-based data. To compare reduction rates by arm at week 12, we calculated the risk ratios of reduction rates of those reaching at least 50%, 75%, and 100% reduction as reported via EMA and survey. We used the same negative binomial mixed effects models to assess the satisfaction and craving scores.

For those in the EC arm, we tabulated the number and percentage of individuals each week that demonstrated evidence of switching (CPD=0 and EC use>0) as well as mostly switching (reduction in CPD >75% and EC use>0). We also summarized weekly EMA-based measures including response rates, the number and percentage of individuals with >60% response rate, average craving, average satisfaction with last CC, average satisfaction with last EC, average CPD, and average EC uses.

All analyses were conducted using Stata (version 18; Stata Corp) and R (version 2024.04.2+764; R Foundation for Statistical Computing), and all tests were 2-tailed, with a .05 significance level.

Results

Adherence to the EMA Protocol

In total, 109 (90%) individuals of the 121 randomized trial participants took part in the EMA protocol. Among the 109, 65 (59.6%) were considered adherent to the EMA protocol with a median weekly response rate that declined slowly from 97.8% in week 1% to 89.4% in week 12. The mean and median response rates over time for all participants are plotted in Figure S2 in Multimedia Appendix 1. Of the 12 individuals with no texting data, 5 withdrew from study activities or opted out, 3 never received the study welcome text, 3 never responded to the study welcome text, and 1 died during enrollment (Figure 1).

Table 1 displays participant characteristics by adherence to EMA protocol. The median age among adherent participants was 61 (IQR 54-65) years and 58 (IQR 52-65) tears among nonadherent individuals. The sample was mostly non-Hispanic white (63/108, 57.8%), and a plurality were married or dating (52/109, 47.7%), had some college or an associate’s degree (44/109, 40.4%), and were retired or on public assistance (47/109, 43.1%). Most participants had COPD (65/109, 59.6%), followed by asthma (36/109, 33%). The median motivation to quit smoking score was 10 (IQR 8 - 10), and the mean confidence to quit smoking was 6.5 (SD 2.2). At baseline, the mean number of CPD reported at the survey was 17.3 (SD 9.7), and the mean CPD from week 1 as assessed from EMA was 11.6 (SD 7.5). No significant differences were observed between the adherent and nonadherent groups.

XSL•FO
RenderX

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) flow diagram for trial recruitment and ecological momentary assessment (EMA) data analysis inclusion. EC: electronic cigarette; NRT: nicotine replacement therapy.

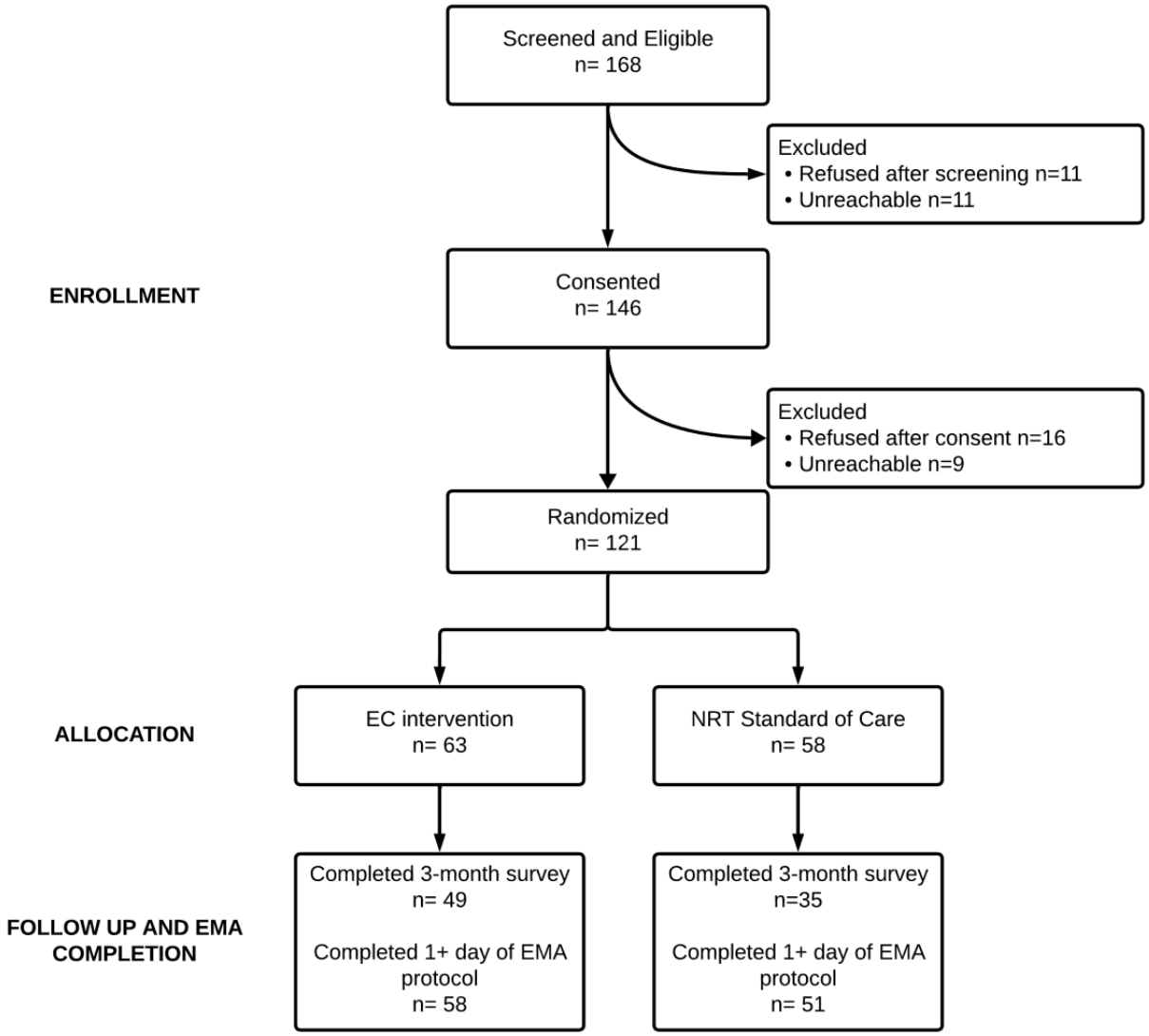

XSL•FO
RenderX

Table . Baseline characteristics of the patients with chronic obstructive pulmonary disease (COPD), coronary artery disease (CAD), peripheral vascular disease (PVD), or asthma in the 12-week, open-label, and 2-arm pilot randomized clinical trial. Distributions and statistical metrics for the comparison of adherent and nonadherent groups to the EMA protocol are presented (n=109). All values are presented as mean (SD) or n (%) or as otherwise stated. Characteristics were evaluated using Wilcoxon rank sum test for continuous variables and Fisher exact test for categorical variables. "Adherent" participants had a weekly response rate of 60% or more throughout the duration of the study period and "Non-adherent" participants had a weekly response rate less than 60% at any point in the study period.

Characteristic	Overall	Adherent	Nonadherent	*P* value
Participants, n (%)	109 (100)	65 (59.6)	44 (40.4)	__[a]
Study arm, n (%)				.82
e-Cigarette	58 (53.2)	34 (52.3)	24 (54.5)	
NRT[b]	51 (46.8)	31 (47.7)	20 (45.5)	
Sociodemographics				
Age (years), median (IQR)	60.0 (54.0 - 65.0)	61.0 (54.0 - 65.0)	58.0 (51.5 - 64.5)	.16
Sex, n (%)				.76
Male	49 (45)	30 (46.2)	19 (43.2)	
Female	60 (55)	35 (53.8)	25 (56.8)	
Race and ethnicity, n (%)				.78
Non-Hispanic White	63 (57.8)	40 (61.5)	23 (52.3)	
Non-Hispanic Black or African American	25 (22.9)	13 (20)	12 (27.3)	
Hispanic	16 (14.7)	9 (13.8)	7 (15.9)	
Non-Hispanic other[c]	5 (4.6)	3 (4.6)	2 (4.5)	
Marital status, n (%)				.67
Married or dating	52 (47.7)	28 (43.1)	24 (54.5)	
Separated, divorced, or widowed	25 (22.9)	16 (24.6)	9 (20.5)	
Single	32 (29.4)	21 (32.3)	11 (25.0)	
Educational history, n (%)				.56
Did not finish high school	11 (10.1)	6 (9.2)	5 (11.4)	
Completed high school or high school equivalency diploma (GED)	24 (22)	16 (24.6)	8 (18.2)	
Some college or associate's degree	44 (40.4)	28 (43.1)	16 (36.4)	
Completed college or graduate school	30 (27.5)	15 (23.1)	15 (34.1)	
Work status, n (%)				.45
Full-time or part-time work	42 (38.5)	21 (32.3)	21 (47.8)	
Public assistance or retirement saving	47 (43.1)	30 (46.2)	17 (38.6)	
Unemployed or other	20 (18.3)	14 (21.5)	6 (13.6)	
Screening diagnoses				
Asthma	36 (33.0)	18 (27.7)	18 (40.9)	.15
CAD or PVD	24 (22.0)	13 (20.0)	11 (25.0)	.54
COPD	65 (59.6)	42 (64.6)	23 (52.3)	.20
1 or more conditions	16 (14.7)	8 (12.3)	8 (18.2)	.40

XSL•FO
RenderX

Characteristic	Overall	Adherent	Nonadherent	*P* value
Condition-specific assessments [d], median (IQR)				
CAT[e]	20.5 (16.0 - 26.0)	19.5 (14.0 - 26.0)	22.0 (16.5 - 26.0)	.44
SAQ-7[f]	15.0 (13.0 - 19.0)	15.0 (13.0 - 19.0)	15.5 (13.5 - 19.5)	.65
ACT[g]	78.8 (69.7 - 87.9)	77.3 (69.7 - 89.4)	78.8 (69.7 - 84.8)	.79
Dyspnea and CCQ [h]				
Dyspnea (shortness of breath) Scale				.13
Dyspnea only with strenuous exercise	18 (16.5)	13 (20.0)	5 (11.4)	
1. Dyspnea when hurrying or walking up a slight hill	27 (24.8)	19 (29.2)	8 (18.2)	
2. Walks slower than people of the same age because of dyspnea or has to stop for breath when walking at own pace	24 (22.0)	10 (15.4)	14 (31.8)	
3. Stops for breath after walking 100 yards (91m) or after a few minutes	28 (25.7)	17 (26.2)	11 (25.0)	
4. Too dyspneic to leave house or breathless when dressing	10 (9.2)	6 (9.2)	4 (9.1)	
CCQ total score, median (IQR)	2.9 (2.1 - 3.9)	2.7 (2.1 - 3.8)	3.2 (2.2 - 4.0)	.51
CCQ symptom score, median (IQR)	3.5 (2.6 - 4.5)	3.2 (2.8 - 4.5)	3.8 (2.5 - 4.5)	.83
CCQ functional state score, median (IQR)	2.5 (1.5 - 3.5)	2.0 (1.5 - 3.2)	2.6 (1.8 - 3.6)	.27
CCQ mental state score, median (IQR)	2.5 (1.0 - 3.5)	2.0 (1.0 - 4.0)	2.5 (1.5 - 3.5)	.76
Motivation and confidence to quit smoking, median (IQR)				
Motivation to quit (1-10)	10.0 (8.0 - 10.0)	10.0 (8.0 - 10.0)	9.5 (7.0 - 10.0)	.72
Confidence to quit (1-10)	6.5 (2.2)	6.5 (2.2)	6.4 (2.2)	.71
Tobacco use and risk perception, mean (SD)				
Average number of CPD[i]	17.3 (9.7)	16.9 (9.4)	17.9 (10.2)	.60
CPD per SlickText EMA (week 1)	11.6 (7.5)	11.5 (7.1)	12.0 (8.3)	.76
Frequency of current cigarette use, n (%)				.09
Every day	105 (96.3)	61 (93.8)	44 (100)	

XSL•FO
RenderX

Characteristic	Overall	Adherent	Nonadherent	*P* value
Some days	4 (3.7)	4 (6.2)	0 (0)	

[a]Not applicable.

[b]NRT: nicotine replacement therapy.

[c]Non-Hispanic other include Arab or Middle Eastern (n=2), not reported (n=2), and Asian (n=1).

[d]Condition-specific assessments restricted to those with condition: COPD: Baseline (n=65; 23 nonadherent and 42 adherent), CAD or PVD: Baseline (n=24; 13 nonadherent and 11 adherent), asthma: Baseline (n=36; 18 nonadherent and 18 adherent).

[e]CAT: COPD Assessment Test.

[f]SAQ-7: Seattle Angina Questionnaire-7.

[g]ACT: Asthma Control Test.

[h]CCQ: Clinical COPD Questionnaire.

[i]CPD: cigarettes smoked per day.

Summary of EMA-Derived Measures

Consumption of CC and EC Over Time

Table 2 provides weekly summaries of response rates, CC and EC use, and satisfaction and craving. Using EMA, overall, the reduction of weekly average CPD was statistically significant, decreasing from 14.2 (SD 9.9) in week 1 to 4.6 (SD 6.7) in week 12 ($P<.001$). Among the EC group, the average ECs per day modestly increased from 4.3 (SD 5.5) in week 2 to 7.1 (SD 11.6) in week 10, followed by a minor decline to 5.8 (SD 9.6) in week 12.

XSL•FO
RenderX

Table . Weekly smoking-related metrics summarized from ecological momentary assessment (EMA) protocol over the 12-week study period. Satisfaction and craving values were based on a 0-9 point scale, ranging from 0 (not satisfied or no craving at all) to 9 (extremely satisfied or highest craving).

	Response rate			CC[a] and EC[b] use				Satisfaction and craving		
Week	Median response rate (IQR)	Mean response rate (SD)	Participants with >60% response rate, n (%)	Mean CPD[c] (SD)	Mean EC uses per day (SD)	Participants reporting mostly switching, n (%)[d,e]	Participants switching to ECs[d], n (%)	Mean Satisfaction with last CC (SD)	Mean EC satisfaction[d] (SD)	Mean craving (SD)
1	97.8 (94.7-100.0)	95.4 (8.7)	104 (95.4)	14.2 (9.9)	__[f]	—	—	6.2 (1.6)	—	4.7 (1.7)
2	98.0 (91.1-100.0)	94.9 (8.6)	102 (93.6)	10.9 (8.9)	4.3 (5.5)	3 (5.2)	1 (1.7)	5.7 (1.8)	4.9 (2.0)	4.6 (1.7)
3	96.9 (87.5-99.2)	94.1 (10.0)	100 (91.7)	8.8 (8.1)	6.4 (7.4)	5 (8.6)	1 (1.7)	5.7 (1.9)	4.9 (2.0)	4.4 (1.7)
4	96.2 (85.8-99.1)	93.4 (10.2)	96 (88.1)	7.6 (7.6)	6.3 (7.1)	8 (13.8)	4 (6.9)	5.6 (2.0)	4.9 (2.0)	4.5 (2.0)
5	96.3 (81.6-98.5)	92.5 (11.9)	93 (85.3)	7.0 (7.5)	7.0 (9.4)	11 (19.0)	6 (10.3)	5.6 (1.9)	5.0 (1.9)	4.5 (1.8)
6	94.6 (74.9-98.5)	91.8 (12.6)	87 (79.8)	6.3 (7.5)	6.7 (7.7)	9 (15.5)	3 (5.2)	5.5 (2.0)	5.1 (1.8)	4.4 (2.0)
7	93.2 (69.4-98.6)	91.7 (13.0)	83 (76.1)	5.9 (7.2)	7.3 (9.1)	7 (12.1)	5 (8.6)	5.5 (2.0)	5.2 (1.7)	4.3 (1.8)
8	91.6 (61.5-98.3)	91.1 (13.3)	80 (73.4)	5.0 (7.3)	7.0 (10.4)	8 (13.8)	6 (10.3)	5.4 (2.0)	5.2 (1.7)	4.2 (2.0)
9	89.0 (33.3-97.6)	90.2 (14.5)	74 (67.9)	5.7 (8.7)	6.1 (9.4)	12 (20.7)	5 (8.6)	5.4 (2.0)	5.2 (1.8)	4.1 (2.2)
10	88.9 (46.5-97.9)	89.8 (13.7)	76 (69.7)	5.2 (8.1)	7.1 (11.6)	12 (20.7)	7 (12.1)	5.4 (2.1)	5.3 (1.8)	4.3 (2.2)
11	87.7 (34.9-98.0)	89.9 (13.7)	74 (67.9)	5.1 (8.2)	6.2 (10.3)	10 (17.2)	6 (10.3)	5.4 (2.1)	5.4 (1.8)	4.0 (2.2)
12	89.4 (42.5-97.8)	89.1 (15.5)	70 (64.2)	4.6 (6.7)	5.8 (9.6)	12 (20.7)	7 (12.1)	5.3 (2.2)	5.4 (1.8)	4.1 (2.2)

[a]CC: combustible cigarette.

[b]EC: e-cigarette.

[c]CPD: cigarettes per day.

[d]Summary measures refer only to the EC group (n=58).

[e]Mostly switching is defined as >75% reduction in CPD from week 1 and use of e-cigarettes.

[f]Not applicable.

Craving and Satisfaction Over Time

We were also able to obtain weekly estimates for craving, satisfaction with the last CC, and satisfaction with the last EC (Table 2). Weekly average craving declined modestly from 4.7 (SD 1.7) in week 1 to 4.1 (SD 2.2) in week 12. Similarly, weekly average satisfaction with the last cigarette declined from 6.2 (SD 1.6) in week 1 to 5.3 (SD 2.2) in week 12. We report weekly satisfaction and craving by arm in Table 2. No treatment effect was found in the negative binomial mixed effects models for satisfaction and craving (Tables S1 and S2 in Multimedia Appendix 1). Among EC participants, average satisfaction with the last EC rose gradually from 4.9 (SD 2.0) in week 2 to 5.4 (SD 1.8) in week 12.

EMA Comparability With Survey Assessments

Reported CPD estimates from EMA correlated strongly with the survey setting at 3 months (Pearson r=0.73; 95% CI 0.60-0.82). Of note, there was one EC participant who reported a very high consumption of combustible cigarettes (>40 CPD) via EMA and at their 3-month visit; this individual's data was included in the current analysis as results did not differ meaningfully when excluding their data. There was no statistically significant paired difference between EMA-reported CPD and survey-reported CPD (paired difference=0.39, 95% CI −1.02 to 1.79). Craving measurements from EMA and surveys also showed a significant positive correlation (Pearson r=0.38, 95% CI 0.17-0.56). In Figure 2, plot A depicts a scatter plot visualizing the association between CPD use via survey and EMA at week 12. The estimated correlation (Pearson r) is 0.73 (95% CI 0.60-0.82). A fitted regression line and 95% CI

XSL•FO
RenderX

are included. Data points are stratified by treatment arm. Plot B depicts a histogram showing the distribution of paired CPD differences between EMA and survey at week 12 for both treatment arms. The estimated mean difference is 0.39 (95% CI −1.02 to 1.79). Plot C depicts a scatter plot showing the association between craving scores for CC via survey and EMA at week 12. The estimated correlation is 0.38 (95% CI 0.17-0.56). A fitted regression line and 95% CI are included. Data points are stratified by treatment arm.

Figure 2. Comparison of data from both ecological momentary assessment (EMA)- and survey-based method for cigarettes smoked per day (CPD) and craving scores for combustible cigarette (CC) by treatment arms. E-cig: e-cigarette; NRT: nicotine replacement therapy.

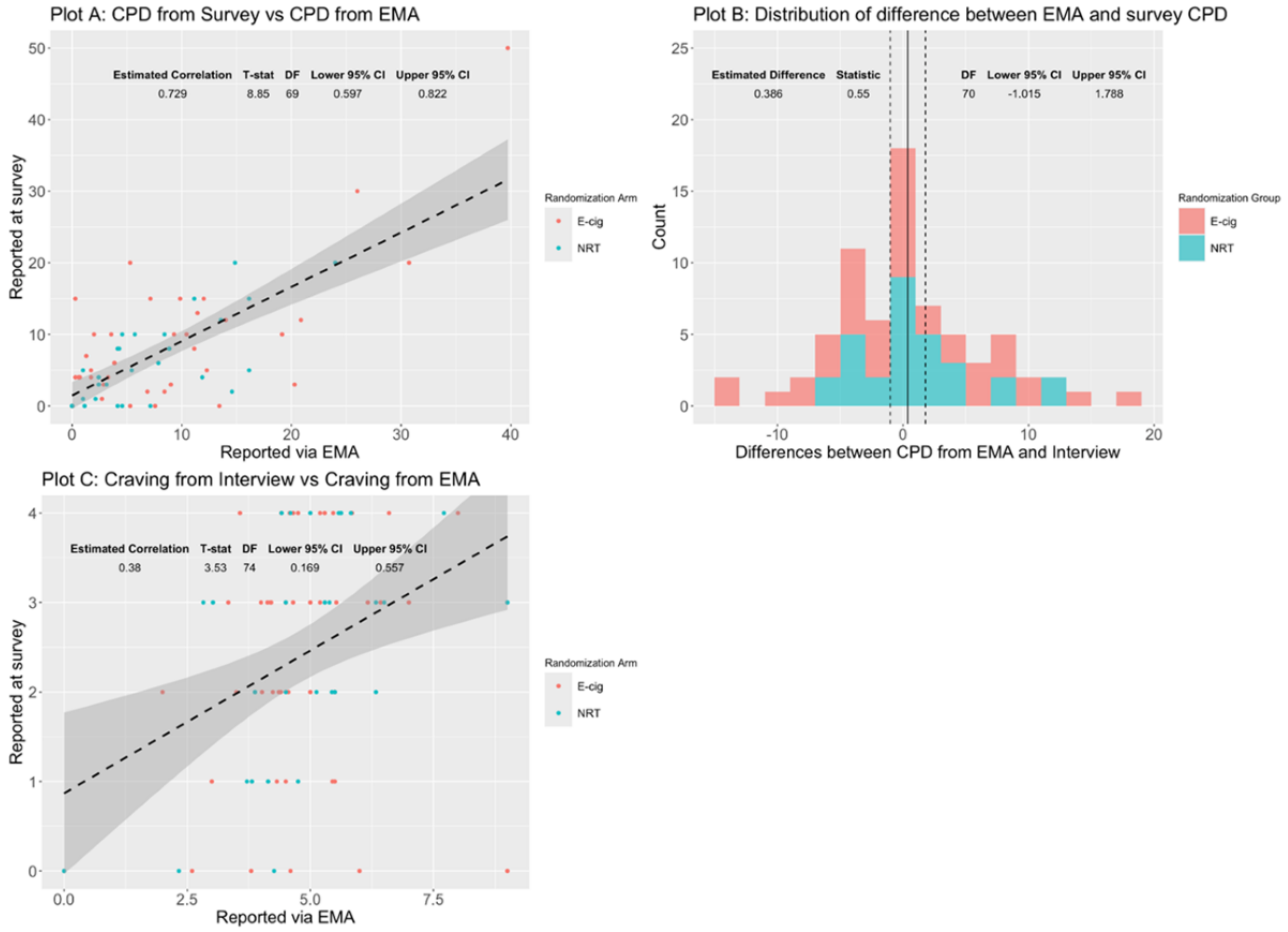

Reduction Rate Comparisons and Negative Binomial Regression Using EMA Versus Survey Data

Tables 3 and 4 show the comparisons between the reduction rate in CPD (50%, 75%, and 100% reduction) and negative binomial regression results using data reported through EMA and through surveys. Both treatment groups showed a significant decrease in the mean number of CPD at the end of the study period (overall; week 1: mean 14.2, SD 9.9; week 12: mean 4.6, SD 6.7) though there was no evidence that reduction in CPD was different between arms based on reduction rate. Of note, the point estimates for CPD reduction rate were all slightly higher than those computed from survey data.

The average CPD count had a mean of 8.2 (SD 8.03) and a variance of 64.5. With unequal mean and variance, our model must account for overdispersion, which the negative binomial model accomplishes. In both models, we found a significant effect of time on the average number of cigarettes per day. In the model using EMA data, we found a 7% decrease in CPD for each week (IRR for 1-week change 0.929, 95% CI 0.93-0.93), and in the model using survey data, the IRR for time across the 12-week intervention was 0.36 (95% CI 0.29-0.44). In both models, we found no statistically significant effect of randomization group on the rate of CPD; however, interestingly, the point estimates for the IRR for randomization group are in differing directions.

XSL•FO
RenderX

Table . Relative rates of reaching at least 50%, 75%, and 100% cigarettes smoked per day (CPD) reduction rates for time and treatment effects using ecological momentary assessment (EMA) and study surveys at week 12.

	EMA				Study survey			
	EC[a] (n=58), n (%)	NRT[b] (n=51), n (%)	RR[c] (95% CI)	*P* value	EC (n=58), n (%)	NRT (n=51), n (%)	RR (95% CI)	*P* value
CPD reduction rate								
50% or higher	34 (58.6)	24 (47.1)	1.25 (0.87 - 1.79)	.23	31 (53.4)	22 (43.1)	1.24 (0.83 - 1.84)	.28
75% or higher	26 (44.8)	15 (29.4)	1.52 (0.91 - 2.54)	.10	17 (29.3)	14 (27.5)	1.07 (0.59 - 1.94)	.83
100%	12 (20.7)	6 (11.8)	1.76 (0.71 - 4.35)	.21	10 (17.2)	8 (15.7)	1.10 (0.47 - 2.57)	.83

[a]EC: electronic cigarette.

[b]NRT: nicotine replacement therapy.

[c]RR: relative risk.

Table . Incidence rate ratios for time and treatment effects using ecological momentary assessment (EMA) and study surveys at week 12.

	EMA		Study survey	
	IRR[a] (95% CI)	*P* value	IRR (95% CI)	*P* value
Negative binomial regression				
Intercept	7.74 (5.34 - 11.23)	<.001	15.89 (12.92 - 19.54)	<.001
Time	0.929 (0.928 - 0.929)	<.001	0.36 (0.29 - 0.44)	<.001
Group				
NRT[b]	—[c]	—	—	—
EC[d]	0.87 (0.52 - 1.45)	.60	1.11 (0.85 - 1.46)	.43

[a]IRR: incidence rate ratio.

[b]NRT: nicotine replacement therapy.

[c]Not applicable.

[d]EC: electronic cigarette.

Cigarette and e-Cigarette Use Over Time

Plot A in Figure 3 shows the reported number of CCs smoked in each week by all participants. The trend lines were fitted using a generalized additive model and are included only for visualization. This visualization aligned with the hypothesis testing of negative binomial mixed effects modeling in that we see a general downward trend in the number of cigarettes smoked for participants in each group over time, with similar trends for each group. To visualize EC use more clearly, we presented EC use in days rather than weeks.

Among those in the EC arm, EC use increased gradually and stabilized after day 20, with daily CC consumption becoming slightly lower than daily EC use. The percentage of individuals who “switched” completely from CCs to ECs in each week ranged from 1.7% (1/58) in week 2 to 12.1% (7/58) in week 12. The percentage of mostly switching increased from 5.2% (3/58) in week 2 to 20.7% (12/58) in week 12.

XSL•FO
RenderX

Figure 3. Trends in electronic cigarette (EC) and combustible cigarette (CC) use over the 12-week study period. Plot A depicts the change in CC by treatment arms on a weekly basis over the study period. Data points represent the total number of CC per week. A fitted line for each trial arm is included, which was fit using a generalized additive model (GAM). Plot B depicts the comparison of change in total number of daily CC and EC uses among EC arm group participants over the 12-week study period. Data points represent the total number of CC and EC uses per day. Fitted lines for CC and EC were fitted using GAM method. CPD: cigarettes smoked per day; NRT: nicotine replacement therapy.

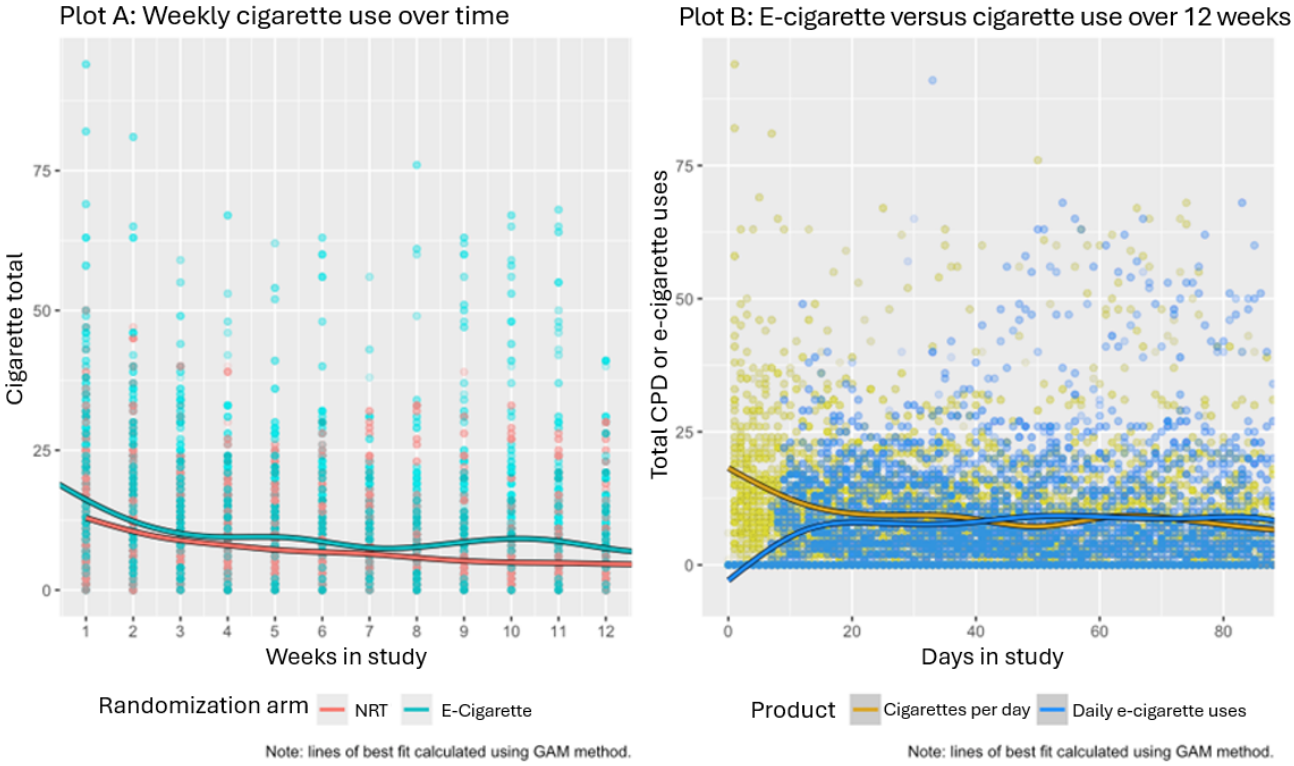

Discussion

Principal Findings

The results showed that participants in either study arm who reported greater CPD in the 3-month visit survey tended to report greater CPD in the final week of EMA prompting, and vice versa. Specifically, the EMA data indicated that EC use could lead to a reduction in CC use similar to that of NRT. These results suggest that ECs could be a potential alternative smoking cessation tool in the future. Similar to analyses using study-visit data, we found a statistically significant time effect on CPD reduction when analyzing the EMA data. Through the insights provided by EMA, patterns in dual use and switching could also be observed where participants in the EC arm appeared to consume fewer CPD over time and had a steady increase in EC use. Together, these results provide evidence that EMA is likely a suitable method to collect recall-based smoking data from participants with greater frequency and shorter recall periods compared with standard study visits.

In addition, the rigorous EMA protocol implemented in the study, with a high response rate over the 12-week intervention period, demonstrates the feasibility of assessing smoking behaviors in this manner. The relevance and validity of EMA data rely on the comparability with traditional methods of measurement (eg, surveys) and the overall adherence to the EMA protocol among participants. This study provides insight by directly comparing smoking-related measures collected through EMA and surveys. The nonsignificant paired difference found in this study contrasts with other similar studies that found large discrepancies between retrospective and EMA data [25,26]. These results indicate that an EMA protocol can achieve equal or more accurate data collection than traditional methods. Furthermore, levels of EMA adherence have been found to differ based on participant characteristics in some groups [27]. While previous studies have used EMA protocols to assess the prevalence of smoking CCs or ECs [9,26,28], as well as changing smoking patterns over time [29], this study is among the first to use a smoking EMA protocol in a population with chronic conditions, thus, demonstrating the feasibility of an SMS text messaging–based smoking EMA protocol in this population.

The visualization of CPD and EC use over time (see Figure 3) offer examples of the added information provided by EMA, giving researchers more precise data collection points. EMA can also help identify and quantify the switching (from CCs to ECs) patterns of participants closer to when switching actually occurs, which itself is an area of future research. For example, for this trial, using traditional surveys, we would only be able to assess switching at the time of surveys (at 3- and 6-month follow-ups); however, EMA allowed us to determine when and if participants did in fact switch from CC to EC within and immediately following the intervention period. If assessed in near real time, this information could even be leveraged to tailor counseling delivered to participants.

We note several limitations. In total, 109 (90%) of the 121 trial participants agreed and enrolled in SMS text messaging prompts; however, only 65 participants were adherent (EMA response rate greater than 60%) in the final week of prompting. All

XSL•FO
RenderX

responses were used in the negative binomial mixed effect model, whether they responded to all prompts in a day or only a subset. This type of missingness could bias the average number of cigarettes smoked downward, particularly as fewer participants responded consistently as the study period continued. While not statistically significant, exploratory analyses seemed to suggest that employed participants may be more likely to drop off in response rates compared with retired or unemployed participants; however, further work is required to better understand effects on response rate and retention. In general, there is a risk of an EMA protocol placing too much burden upon participants leading to fatigue or burnout [30]; in turn, participant fatigue could lead to higher drop-out rates and missingness, which could bias results. Similarly, there was variance in contact intensity between the two study arms, as the NRT participants were not asked about NRT or EC use limiting the total number of questions asked. While no statistically significant difference in EMA protocol adherence was measured, differential contact between the arms could potentially bias perceived burden or engagement with the EMA protocol. Despite the missingness and the potential for participant fatigue, this EMA protocol saw continued adherence and high response rates from a majority of enrolled participants while maintaining a fairly rigorous prompt frequency. Continued adherence indicates relatively low burden of text prompts (vs through other platforms; eg, phone app) and the familiarity most participants have with texting, which could be a more acceptable and feasible method compared with lengthier surveys. Future analyses of EMA data could seek to better understand additional smoking-related measures beyond craving and satisfaction. Due to limitations of measuring experiences using numerical scales [31,32], future EMA protocols may consider free-text responses and apply tools like natural language processing to EMA data to interpret and clean patient responses thereby transforming the data into meaningful information on human behaviors and sentiments. Better implementation of set responses can help ensure researchers are receiving the information desired in workable forms and reduce the data cleaning needed. Another future area of research is determining how well EMA data can detect the real-time switching from CCs to ECs.

Conclusion

SMS text messaging–based EMA represents a feasible and suitable method to collect recall-based smoking-related data. As compared with the survey data, the use of EMA provides unique advantages, namely greater granularity in the smoking timelines and a capability to detect switching patterns between CCs and ECs in near real time. EMA should be considered as a method for data collection when developing future tobacco harm reduction research.

Acknowledgments

Funding was received from internal funds provided to SES by New York University (NYU) Grossman School of Medicine. The study sponsor had no role in the study design; collection, analysis, and interpretation of data; writing the report; or the decision to submit the report for publication. The NYU Langone Health institutional review board approved the study (#s20-00839).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

All authors contributed to the study's conception and design. TM and MH performed data cleaning and analysis. TM contributed to material preparation and writing the first draft of the manuscript. All other co-authors reviewed the analysis and edited subsequent versions of the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary tables presenting (1) mean satisfaction with the last combustible cigarette (CC) and craving for CC over the 12-week study period stratified by treatment arms using ecological momentary assessment (EMA) data and (2) incidence rate ratios for craving and satisfaction over time and treatment effects using EMA data at week 12. Figures presenting (1) a flow diagram depicting the text messaging prompts and flow received by the patients from the electronic cigarette (EC) group (participants were asked whether they smoked a CC or used an EC since their last report and based on the responses they are directed to specific follow-up questions), and (2) trends of weekly means and medians of response rates (%) among all participants over the 12-week study period.
[DOCX File, 132 KB - formative_v9i1e66709_app1.docx]

Checklist 1
CONSORT-EHEALTH (version 1.6.1) checklist.
[PDF File, 1167 KB - formative_v9i1e66709_app2.pdf]

XSL•FO
RenderX

References

1. Smoking cessation: a report of the surgeon general. US Department of Health and Human Services. 2020. URL: https://www.hhs.gov/sites/default/files/2020-cessation-sgr-full-report.pdf [accessed 2025-03-11]
2. Cornelius ME, Loretan CG, Jamal A, et al. Tobacco product use among adults - United States, 2021. MMWR Morb Mortal Wkly Rep 2023 May 5;72(18):475-483. [doi: 10.15585/mmwr.mm7218a1] [Medline: 37141154]
3. Tashkin DP, Murray RP. Smoking cessation in chronic obstructive pulmonary disease. Respir Med 2009 Jul;103(7):963-974. [doi: 10.1016/j.rmed.2009.02.013]
4. Brown J, Beard E, Kotz D, Michie S, West R. Real-world effectiveness of e-cigarettes when used to aid smoking cessation: a cross-sectional population study. Addiction 2014 Sep;109(9):1531-1540. [doi: 10.1111/add.12623] [Medline: 24846453]
5. Cobb NK, Abrams DB. The FDA, e-cigarettes, and the demise of combusted tobacco. N Engl J Med 2014 Oct 16;371(16):1469-1471. [doi: 10.1056/NEJMp1408448] [Medline: 25317866]
6. Fiore MC, Schroeder SA, Baker TB. Smoke, the chief killer--strategies for targeting combustible tobacco use. N Engl J Med 2014 Jan 23;370(4):297-299. [doi: 10.1056/NEJMp1314942] [Medline: 24450888]
7. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol 2008;4:1-32. [doi: 10.1146/annurev.clinpsy.3.022806.091415] [Medline: 18509902]
8. Tversky A, Kahneman D. Availability: A heuristic for judging frequency and probability. Cogn Psychol 1973 Sep;5(2):207-232. [doi: 10.1016/0010-0285(73)90033-9]
9. Mead EL, Chen JC, Kirchner TR, Butler J 3rd, Feldman RH. An ecological momentary assessment of cigarette and cigar dual use among African American young adults. Nicotine Tob Res 2018 Aug 14;20(suppl_1):S12-S21. [doi: 10.1093/ntr/nty061] [Medline: 30125017]
10. Todd M. Daily processes in stress and smoking: effects of negative events, nicotine dependence, and gender. Psychol Addict Behav 2004 Mar;18(1):31-39. [doi: 10.1037/0893-164X.18.1.31] [Medline: 15008683]
11. Pratt SI, Ferron JC, Brunette MF, Santos M, Sargent J, Xie H. E-Cigarette provision to promote switching in cigarette smokers with serious mental illness-a randomized tTrial. Nicotine Tob Res 2022 Aug 6;24(9):1405-1412. [doi: 10.1093/ntr/ntac082] [Medline: 35363874]
12. Biener L, Abrams DB. The Contemplation Ladder: validation of a measure of readiness to consider smoking cessation. Health Psychol 1991;10(5):360-365. [doi: 10.1037//0278-6133.10.5.360] [Medline: 1935872]
13. Gupta N, Pinto LM, Morogan A, Bourbeau J. The COPD assessment test: a systematic review. Eur Respir J 2014 Oct;44(4):873-884. [doi: 10.1183/09031936.00025214]
14. Stevens ER, Lei L, Cleland CM, et al. Electronic cigarettes as a harm reduction strategy among patients with COPD: protocol for an open-label two arm randomized controlled pilot trial. Addict Sci Clin Pract 2022 Dec;17(1):2. [doi: 10.1186/s13722-021-00284-0]
15. Scott NW, McPherson GC, Ramsay CR, Campbell MK. The method of minimization for allocation to clinical trials. a review. Control Clin Trials 2002 Dec;23(6):662-674. [doi: 10.1016/s0197-2456(02)00242-8] [Medline: 12505244]
16. Gupta N, Pinto LM, Morogan A, Bourbeau J. The COPD assessment test: a systematic review. Eur Respir J 2014 Oct;44(4):873-884. [doi: 10.1183/09031936.00025214] [Medline: 24993906]
17. Ställberg B, Nokela M, Ehrs PO, Hjemdal P, Jonsson EW. Validation of the clinical COPD Questionnaire (CCQ) in primary care. Health Qual Life Outcomes 2009 Mar 25;7:1-9. [doi: 10.1186/1477-7525-7-26] [Medline: 19320988]
18. van der Molen T, Willemse BWM, Schokker S, ten Hacken NHT, Postma DS, Juniper EF. Development, validity and responsiveness of the Clinical COPD Questionnaire. Health Qual Life Outcomes 2003 Apr 28;1:1-10. [doi: 10.1186/1477-7525-1-13] [Medline: 12773199]
19. Chan PS, Jones PG, Arnold SA, Spertus JA. Development and validation of a short version of the Seattle angina questionnaire. Circ Cardiovasc Qual Outcomes 2014 Sep;7(5):640-647. [doi: 10.1161/CIRCOUTCOMES.114.000967] [Medline: 25185249]
20. Nathan RA, Sorkness CA, Kosinski M, et al. Development of the asthma control test: a survey for assessing asthma control. J Allergy Clin Immunol 2004 Jan;113(1):59-65. [doi: 10.1016/j.jaci.2003.09.008] [Medline: 14713908]
21. Soler X, Holbrook JT, Gerald LB, et al. Validity of the asthma control test questionnaire among smoking asthmatics. J Allergy Clin Immunol Pract 2018;6(1):151-158. [doi: 10.1016/j.jaip.2017.05.010] [Medline: 28669892]
22. Cappelleri JC, Bushmakin AG, Baker CL, Merikle E, Olufade AO, Gilbert DG. Revealing the multidimensional framework of the Minnesota nicotine withdrawal scale. Curr Med Res Opin 2005 May;21(5):749-760. [doi: 10.1185/030079905X43712] [Medline: 15969874]
23. Hughes JR, Hatsukami D. Signs and symptoms of tobacco withdrawal. Arch Gen Psychiatry 1986 Mar;43(3):289-294. [doi: 10.1001/archpsyc.1986.01800030107013] [Medline: 3954551]
24. Yirga AA, Melesse SF, Mwambi HG, Ayele DG. Negative binomial mixed models for analyzing longitudinal CD4 count data. Sci Rep 2020 Oct 7;10(1):16742. [doi: 10.1038/s41598-020-73883-7] [Medline: 33028929]
25. Poulton A, Pan J, Bruns LR Jr, Sinnott RO, Hester R. Assessment of alcohol intake: Retrospective measures versus a smartphone application. Addict Behav 2018 Aug;83:35-41. [doi: 10.1016/j.addbeh.2017.11.003] [Medline: 29128148]
26. Yang JJ, Ou TS, Lin HC, Kyung Nam J, Piper ME, Buu A. Retrospective and real-time measures of the quantity of e-cigarette use: an ecological momentary assessment study. Nicotine Tob Res 2023 Aug 23;25(10):1667-1675. [doi: 10.1093/ntr/ntad094] [Medline: 37327251]

XSL•FO
RenderX

27. Sokolovsky AW, Mermelstein RJ, Hedeker D. Factors predicting compliance to ecological momentary assessment among adolescent smokers. Nicotine Tob Res 2014 Mar;16(3):351-358. [doi: 10.1093/ntr/ntt154] [Medline: 24097816]
28. Ozga JE, Bays C, Haliwa I, et al. Measurement of cigarette smoking: comparisons of global self-report, returned cigarette filters, and ecological momentary assessment. Exp Clin Psychopharmacol 2022 Jun;30(3):365-370. [doi: 10.1037/pha0000449] [Medline: 33630647]
29. Piper ME, Baker TB, Benowitz NL, Jorenby DE. Changes in use patterns over 1 year among smokers and dual users of combustible and electronic cigarettes. Nicotine Tob Res 2020 Apr 21;22(5):672-680. [doi: 10.1093/ntr/ntz065] [Medline: 31058284]
30. van Genugten CR, Schuurmans J, Lamers F, et al. Experienced burden of and adherence to smartphone-based ecological momentary assessment in persons with affective disorders. J Clin Med 2020 Jan 23;9(2):322. [doi: 10.3390/jcm9020322] [Medline: 31979340]
31. Heishman SJ, Singleton EG, Pickworth WB. Reliability and validity of a Short Form of the Tobacco Craving Questionnaire. Nicotine Tob Res 2008 Apr;10(4):643-651. [doi: 10.1080/14622200801908174] [Medline: 18418787]
32. Nguyen C, Kortlever JTP, Gonzalez AI, Ring D, Brown LE, Somogyi JR. Attempts to limit censoring in measures of patient satisfaction. J Patient Exp 2020 Dec;7(6):1094-1100. [doi: 10.1177/2374373520930468] [Medline: 33457550]

Abbreviations

ACT: Asthma Control Test
CAD: coronary artery disease
CAT: COPD Assessment Test
CC: combustible cigarette
CCD: Clinical COPD Questionnaire
COPD: chronic obstructive pulmonary disease
CPD: cigarettes smoked per day
EC: electronic cigarette
EHR: electronic health record
EMA: ecological momentary assessment
IRR: incidence rate ratio
MNWS: Minnesota Nicotine Withdrawal Scale
NRT: nicotine replacement therapy
NYULH: New York University Langone Health
PVD: peripheral vascular disease
RCT: randomized clinical trial
RR: relative rate
SAQ-7: Seattle Angina Questionnaire-7

Edited by A Mavragani; submitted 24.09.24; peer-reviewed by S Cha; revised version received 11.02.25; accepted 12.02.25; published 21.03.25.

Please cite as:
Morgan T, He M, Nicholson A, El Shahawy O, Sherman SE, Stevens ER
Using Text Messaging Ecological Momentary Assessment to Record Changes in e-Cigarette and Combustible Cigarette Use: Pilot Randomized Clinical Trial
JMIR Form Res 2025;9:e66709
URL: https://formative.jmir.org/2025/1/e66709
doi:10.2196/66709

© Tucker Morgan, Michelle He, Andrew Nicholson, Omar El Shahawy, Scott E Sherman, Elizabeth R Stevens. Originally published in JMIR Formative Research (https://formative.jmir.org), 21.3.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Development and Implementation of Strong Foundations, a Digitally Delivered Fall Prevention Program: Usability and Feasibility Pilot Exercise Cohort Study

Ryan Moran[1,2], MPH, MD; David Wing[1], MS; Hope Davey[1], BS; Hava Barkai[1], MS; Jeanne Nichols[1], PhD

[1]Exercise and Physical Activity Resource Center (EPARC), University of California San Diego, 9500 Gilman Dr. La Jolla, San Diego, CA, United States

[2]Department of Medicine, University of California San Diego, San Diego, CA, United States

Corresponding Author:
Ryan Moran, MPH, MD
Exercise and Physical Activity Resource Center (EPARC), University of California San Diego, 9500 Gilman Dr. La Jolla, San Diego, CA, United States

Abstract

Background: Falls remain a major public health problem and a significant cause of preventable injury. Maintaining strength and balance by staying active can prevent falls in older adults, and public health advocates support referral to community exercise programs. Given the growth in use and acceptance of technological interfaces, there remains an interest in understanding the role of a synchronous exercise program designed to improve strength, postural alignment, and balance specifically designed to be delivered in a digital environment with respect to usability and feasibility.

Objective: This study aims to design and implement a synchronously delivered digital fall prevention program to adults aged 60 years and older, to understand the usability, feasibility, and attendance.

Methods: The "Strong Foundations" program, a 12-week, live, digitally delivered fall-prevention exercise program was informed from different existing in-person exercises and piloted to older adults who were considered a low fall risk by scores of 4 or less from the Centers for Disease Control and Prevention's (CDC's) Stopping Elderly Accidents and Deaths Initiative (STEADI) Staying Independent questionnaire. The System Usability Scale (SUS) measured usability and feasibility at the completion of this program, and digital measures of age-related function (timed up and go [TUG] and 30-second chair stand [30 CS]) were collected pre- and postintervention. Data were collected in 2021.

Results: A total of 39 older adults were recruited and 38 completed the 12-week program with an average age of 72 years. The average SUS was 80.6, with an 85% attendance rate and an 8.5 (out of 10) self-reported satisfaction score. Digitally collected TUG and 30 CS statistically improved pre- and postintervention by 9% and 24%, respectively; by week 12, 64% (23/36) of participants improved in the timed up and go and 91% (32/35) improved the chair stands.

Conclusion: There was excellent usability and acceptability for Strong Foundations, a novel fall-prevention program designed to be delivered digitally and promising improvement of objective measures of fall risk.

(JMIR Form Res 2025;9:e67406) doi:10.2196/67406

KEYWORDS

digital health; fall prevention; fall risk; older adults; geriatrics; system usability scale; Strong Foundations; feasibility; public health; user acceptance; exercise; usability; digital technology; mobile phone

Introduction

Fall-related injuries among older adults are a significant and growing public health concern. Over one-quarter of community-residing men and women 65 years of age and older fall annually [1,2]. Risk is compounded over time as a single fall predicts recurrent falls [3-7], and fall-related mortality has increased over 30% just between 2007 and 2016 [8]. Not only do falls represent a major cause of accidental death and mobility-related disability among older adults, but they also result in psychological distress, loss of independence, and a decrease in quality of life [9,10]. Despite the existence of various interventions and policies aimed at reducing fall risk in this population, the incidence of falls remains alarmingly high.

While fall risk is multifactorial, identification of risk factors, as well as referral to and participation in appropriate fall-risk reduction programs are established as an effective, evidence-based approach to reduce fall risk [2,11]. Specifically, targeted strength and balance exercises have consistently been shown to improve fall risk, and accordingly, the Centers for Disease Control and Prevention (CDC) has outlined an

XSL•FO
RenderX

evidence-based clinical approach to identify those at risk for falls to help assess known risk factors and refer for community-based fall-prevention programs [12,13]. The public health use of this approach is further supported by the United States Preventive Services Task Force (USPSTF) "Grade B" recommendation to refer older adults for fall prevention exercise programs [14,15]. However, despite these resources and recommendations, clinical identification and referrals have been insufficient [16]; only approximately one-third of older adults report being asked about fall-risk by their health care providers, and similarly, only around a third of those who fall report discussing their fall and fall circumstances with their health care provider [17,18]. At the population level, conservative estimates suggest far fewer than 50% older adults meet recommended physical activity levels [19]. These findings mirror the growing epidemiological evidence of increased rates of falls in older adults. While some existing programs that improve fall risk and activity for older adults, including EnhanceFitness PA and Silver Sneakers may be available to some older adults with certain insurance plans at no additional costs [20], others may face substantial fees; furthermore, many eligible seniors do not use these programs, citing several barriers. Such barriers include general dislike of gyms and a lack of awareness or instructor suitability and generalizability, problems accessing appropriate classes, concerns about bad weather and driving, and perceptions about being "old" in these environments [21]. These issues are even greater among older adults residing in areas and communities that have fewer facilities for guided physical activity, more limited transportation infrastructure and less exercise expertise [22]. Finally, there remain geographic and demographic disparities related to falls, including higher rates in rural areas [2], highlighting the interest and need for different approaches to address this public health and clinically relevant problem in older adults in novel ways that increase reach and are safe and effective.

Appropriate technologically driven approaches have promise to increase range and dissemination of fall prevention resources, including making them easier and safer to access [23]. Furthermore, digitally delivered programs that involve appropriately tailored exercises offer the opportunity to balance risks and benefits during times of social distancing, while improving dissemination and community uptake, and possibly improve fall risk [24]. While recent attention has highlighted concerns with broadband access in older adult populations, especially with attention to disparities and equity [25], digital interventions showcase the promise as those aged 65 years and older still represent the fasting growing demographic in technological uptake [26,27]. Thus, given the trend of increasingly ubiquitous access to smart-enabled technology in older adults, and gaining acceptance in this population [26], there is great potential to improve equitable access to high-quality exercise-based fall prevention resources by decreasing geographic boundaries—not only immediately, but increasingly in the future. Alternative approaches, including prerecorded, online, and asynchronous delivery, are an additional consideration but lack the ability for real-time guidance, which is possible with synchronous real-time delivery, and therefore offer the opportunity to provide guidance with risk mitigation and personalization when appropriate for participants. Therefore, approaches that leverage technology to improve access to fall-risk reduction exercise programs are imperative, and growing data suggest digitally formatted delivery may be feasible [28].

Zoom (Zoom Video Communications) is a user-friendly, web-based service that provides digital video-based connections between multiple users; Zoom-for-Telehealth is a HIPAA-compliant encrypted service used for clinically delivered care and permits conferencing between individuals. Zoom is readily available online, free to download, and supports a one-to-many viewing platform such that it can be set up so that only the host can see individual participants at a given time [29]. This application was selected based on the ubiquitous nature and availability to host professionally through our institution. The nature of this digital platform lends interest in understanding how fall-prevention curricula may be developed and delivered to older adults; therefore, we developed a novel digital platform, Strong Foundations, designed to leverage the advantages of digital technology with evidence-based fall-prevention strategies, along with real-time feedback to users based on appropriate or inappropriate practice, optimizing gains for a larger group and highlighting core exercises. This program includes features to promote adherence, engagement, and social connectivity enabled by real-time interaction, all important to promote efficacy and subsequent improvement in gains from any intervention [30]. Given the novelty of such a program, we aim to understand the feasibility of this 12-week program in terms of usability and acceptance using a validated scale and attendance rates and explore measures of digitally collected physiologic improvement in adults aged 60 years and older.

Methods

Program Description

Our program, Strong Foundations, is drawn from several rigorously tested fall and fracture prevention exercise programs [11,31-33] but is new and entirely designed to be delivered digitally. This design emphasized exercises that were able to be remotely instructed, witnessed, and performed in a limited physical space such that most users would be able to use readily available household spaces. The program was designed with physician input and by exercise physiologists and a Doctor of Physical Therapy, all with extensive training in both group and individualized exercise for older adult populations. Strong Foundations is a 12-week iterative curricular program with 3 core components: postural alignment and control, balance and mobility, and muscular strength and power. All the exercises offered over the course of the intervention are designed to be appropriate for the target population (ie, older adults) and are standardized so all participants receive the same basic instruction, but the level of difficulty is scaled to participant experience, capability, and musculoskeletal limitations. This individualization was enabled by real-time delivery, such that instructors are able to adapt, respond, and tailor specific aspects of any class to fit the needs of participants, including independently pulling our participants into a "Breakout Room" for direct one-on-one instruction before reassimilating into the larger group. The 12 weeks of instruction establish and build

XSL•FO
RenderX

such that throughout the program, posture and positions prioritized early in the program are used to build into more complicated maneuvers designed to improve the foundational techniques and activities that will allow participants to maximize learning opportunities and optimize gains related to balance, posture, and strength.

Real-Time Feedback

Differentiating this program from other streamed and real-time web-based interventions available, the novel feature of this program is the delivery of semi-individualized instruction in real time within a small group setting. This is accomplished with a specifically trained "support" instructor providing guidance and correction on optimal exercise form while the "lead" instructor demonstrates for the larger group. This semitailored approach allows the lead instructor to provide "best practice" instruction and provide a visual demonstration while assistant instructors assure that participants are not only gaining important skills related to fall-prevention technique, but also doing this safely and remotely, as numerous screens on a digital interface would otherwise compromise the integrity of a single instructor. All instructors are aware of the curriculum in its entirety and versed on adaptive approaches for a participant who may not be able to do specific components safely or without predisposing to injury. An example of this may include someone with a known right shoulder injury, whereby certain resistance band and posture actions need to be modified—thereby the lead instructor may continue providing instruction to the group at large, with uninterrupted group synchronous interaction, but the support instructor can provide individualized recommendations to a specific participant without disrupting the overall dynamic.

Program Development

The program was iteratively designed based on two, 4-week pilot classes with 9 - 12 participants each (total n=21), which refined and informed the curriculum to inform the design of a final total 12-week program, where foundational components are expanded in intensity and challenge, assuring continual growth and improvement of participants. Participant experience was surveyed, with respondents (n=14) noting that the interface was easy to follow, with sufficient audio and visual interface for safely following instruction. Furthermore, 100% (14/14) of participants agreed or strongly agreed that "instruction was clear and easy to follow" and "I felt safe doing the exercises." In addition, 86% (12/14) agreed or strongly agreed to "I could hear the instructors well" and "instruction was easy to see and follow." In total, 64% (9/14) reported awareness of instructors watching validated tools to establish usability, however, were not collected in this early phase of development.

Final program development was designed to incorporate approximately 1 hour of instruction for 12 weeks with key aspects to establish the weekly focus, and materials were generated for dissemination between classes to participants to encourage and underscore the importance of ongoing practice to support gains during this intervention (Table 1).

XSL•FO
RenderX

Table . Strong Foundations curriculum focus across 12 weeks.

Week	Foundational exercise 1	Foundational exercise 2	Foundational exercise 3	Foundational exercise 4	Focus of the week
1	Neutral spine and marionette pose	Pelvic tucks and tilts	Hip hikes	Hip hinge	Posture
2	Quick feet drills	Review hip hinge	"Traditional abs"	Self-test: static balance	Balance
3	Front and lateral arm raises (with band)	Hip hinge (review) with squat and chair sit, stand	Reverse lunge	Heel raises	Strength
4	Neutral spine and pelvic tuck and tilt—standing	Static balance	Reverse lunge with progression	Heel raises with tennis ball	Combine 3 pillars
5	Sit and stand transitions	Around the clock steps	Hip hinge with choice of arm extensions	Modified and regular jumping jacks	Floor to stand transition—floor exercises
6	Hip hinge (review) with squat and chair sit, stand	Lunge with chair assistance progression	Single leg heel raises	Hing hinge + picking up objects	Compound strength movements
7	Drinking bird	Squat to lateral leg raise	Single leg heel raises	Toe raise walk around chair	Increase balance challenge
8	Side squat	Wall sit with wood chop	High knee walk	Quick feet multidirectional	Odd impact and multidirectional movements
9	Weight transfer with split stance	Squat into a High knee and arm reach	High knee walk	Head rotations in a tandem stance	Combine balance with compound movements
10	Single leg stance with band pull downs and serial 7's	Lunge with chair support and head rotation	Standing superman with cognitive challenge	Goal post with arm slides	Introduce cognitive challenge
11	Reaching squats	Lateral jacks	Vertical push-ups	Knee drivers	Speed and power movements
12	Good morning	Bow and arrow	Hands up	Heel toe sitting with resistance band	Collaborative workout

Site, Eligibility, and Enrollment

Participants were recruited through established venues from the Exercise and Physical Activity Research Center (EPARC), an exercise science–based laboratory with experience in physiologic phenotyping based on the University of California, San Diego campus. Given that the intervention was developed and deployed during the COVID-19 pandemic, all recruitment, enrollment, and measurement was done remotely; no face-to-face interaction or measurement occurred throughout the pilot. Given the novelty of the delivery platform and development of this program, our primary aims centered on determining its' usability and feasibility in this context. Therefore, out of an abundance of caution and given the entirely remote recruitment, enrollment, delivery, and measurement through this program, this pilot included those with minimal-to-no-risk factors for falling based upon the CDC's Staying Independent brochure questionnaire [34]; for this pilot, a participant was considered no more than a moderate fall risk based on a score of ≤4 [16]. Textbox 1 shows the eligibility criteria. Informed consent was performed remotely through secure digital methods (Docusign) with a real-time Zoom-based meeting with staff to review protocols and consent documents and establish an appropriate room setup. This included ensuring that there was sufficient space to move while also being at a sufficient distance from the camera that their entire body was visible to the instructor. In addition, before remote measurement, participants were individually coached on how to measure an appropriate length course (ie, 3 meters) and have an appropriately sized chair (ie, approximately 17-inch chair height without arms). Enrollment occurred from August to September 2021, with 39 older adults recruited and cohorted based on predetermined times of the intervention such that participants could select the program time most convenient for their schedule.

XSL•FO
RenderX

Textbox 1. Eligibility criteria for Strong Foundations pilot.

Inclusion criteria:

- Age 60 years or older and ambulatory, including with the use of a cane or walker (1 participant enrolled age 56, which was an unintended protocol deviation).
- Completion of the Stopping Elderly Accidents and Deaths Initiative (STEADI) Stay Independent Brochure.
- Access to internet and computer and a Zoom-interface or broadband.

Exclusion criteria:

- Individuals who are wheelchair bound.
- Exclusively communicate in a language other than English.
- Score of 4 or more on the STEADI modified questionnaire.

Primary Feasibility Outcomes

Programmatic development and launch were aimed at determining feasibility and usability of a digital platform with an older adult population. Attendance was measured weekly as well as self-reported number of exercises performed outside of the class based on the previous week's instruction and usability determined at completion of the program based on the System Usability Scale [35] (SUS), a nonpriority validated questionnaire, which was designed to understand the ease of use of new systems or programs using a 5-element Likert scale.

In general, scores >70 on the SUS are considered to have appropriate acceptability of a program or platform. The SUS was selected as the primary outcome, given it is considered the most widely used scale in eHealth to determine usability [36]. Finally, a survey to understand subjective experience was developed and deployed at the end of the intervention.

Exploratory User Outcomes

To understand user experience and improvement, at baseline and at the completion of this curriculum, digitally collected objective measures of frailty and fall risk were collected to explore the opportunity to use the digital interface to collect these measures and understand changes from this intervention. All participants at baseline and after completing the intervention period (±1 week) were assessed remotely with the measures outlined and endorsed by national and international public health authorities, namely the timed up and go (TUG) [37-39] and 30-second chair rise [40]. Participants were then monitored and timed remotely while completing the assessments. TUG was completed with instruction to complete at both a normal speed (TUG normal) and "as fast as safely able to perform the test" (TUG-Fast). Given enrollment included those with few risk factors for falling, these physical measurements were selected because of their relative ease of administration and interpretability but remained exploratory from the lens of feasibility.

Sample Size Determination

Power calculation was not conducted given the aim to determine feasibility and usability for this program and the selected outcome, the SUS, has been shown to be valid in small samples [41]. Therefore, we used an assumed attendance rate of 70%, which has held consistent in national programs in both rural and urban areas for older adults in fall prevention programs [42] and similar fall prevention programs that have evaluated this time frame to be sufficient to improve measures, markers of frailty, and fitness in different health contexts between 20 and 60 individuals [43-45]. Given budgetary limitations, our plan to enroll 4 cohorts all with at least 8 - 10 older adults would be sufficient to determine both the feasibility and usability and explore physiologic changes.

Statistical Analysis

Primary outcomes to understand feasibility and usability were determined based on the validated SUS [35], which in general assumes an average score of 68 and higher indicates more usability. Attendance rates were determined by a participant joining the Zoom interface measured weekly for each individual, reported as averages (mean) and SD. Numerical averages (mean) of patient-reported questionnaire results (out of 10) were calculated. Associations with SUS scores were calculated by Pearson correlation coefficients. Secondary outcomes were measured using paired-sample 2-tailed *t* tests for differences (for the TUG and 30-second chair rise, collected digitally) with corresponding mean differences and 95% CIs. Statistical analysis was performed on SPSS (version 28; IBM Corp).

Ethical Considerations

This study was reviewed and approved by the University of California San Diego Institutional Review Board (#802148); informed consent was obtained by all participants. All study data were deidentified. Participants were not compensated to participate.

Results

Participants

Across 12 weeks, 38 of 39 (97.4%) recruited older adults completed the program. One participant left after 2 weeks of instruction because of a family emergency. Of the 38 who completed the program, the average age was 72 years (SD 5.5), and 37 were female (97.4%). Participants were spread across 4 cohorts, ranging in size from 9 - 10 individuals.

Feasibility and Usability

On average, participants attended 10 weekly sessions out of 12 (84%). The average number of home exercise sessions

XSL•FO
RenderX

performed weekly between classes was 2.3 (SD 0.7). The average SUS score was 80.6 (SD 15.4), with a range of 40 to 100. On a 10-point scale, the average self-reported overall experience of the intervention and exercise classes was 8.5 (SD 1.9). No participants experienced an injury or had any safety related concerns, and additional subjective feedback was excellent with most participants eager for ongoing involvement and instruction. SUS scores were statistically inversely associated with age (Pearson correlation r=−0.411; P=.10) and positively with reported satisfaction score (Pearson correlation r=0.654; P=.004) but not statistically associated with attendance or improvements in TUG or 30s chair rise scores. Table 2 shows the number of participants and their SUS score range.

Table . System Usability Scale (SUS) score range.

SUS Score	Participants (n=38), n (%)
<50	2 (5)
50 - 60	2 (5)
61 - 70	6 (16)
71 - 80	7 (18)
>80	21 (55)

Objective Measures

Pre- and postmeasures were collected for 36 of the participants for the TUG and 35 of the participants for the 30-second chair rise. Baseline TUG at the "normal" speed was 8.8 (SD 1.9) seconds, and at the fastest possible speed was 6.2 (SD 1.4) seconds. Baseline 30-second chair stands were 14.5 (SD 3.4) stands. Post intervention, TUG and TUG-Fast were 8.0 seconds (SD 1.3 seconds; difference of 0.8 [95% CI 0.24-1.21] seconds) and 5.7 seconds, respectively (SD 1.1 seconds; difference of 0.48 [95% CI 0.22-0.75] seconds) and the 30 second chair rise increased to 18 stands (SD 4.1 stands; difference of –4.8 [95% CI –2.6 to –6.9] stands) (Table 3). At 12 weeks, 64% of all participants improved in the TUG (23/36, for both measures) and 91% (32/35) improved the chair stands. Reasons reported for being unable to complete these were varied but included participants traveling internationally, unrelated knee pain, and a single participant dropping as noted above after 2 weeks.

Table . Digitally collected objective markers of fall-risk reduction.

Measure	Baseline, mean (SD)	Post intervention, mean (SD)	Difference (95% CI)	Change (%)
TUG[a] normal (seconds)	8.8 (1.9)	8.0 (1.3)	7.24 (0.24 - 1.21)	9
TUG-Fast (seconds)	6.2 (1.4)	5.7 (1.1)	0.48 (0.22 - 0.75)	8
30-second chair rise (stands)	14.5 (3.4)	18 (4.1)	−4.8 (−2.6 to 6.9)	24

[a]TUG: timed up and go.

Discussion

The development and rationale behind the need for a digital curriculum, Strong Foundations, is described here. In this 12-week pilot, we found a high level of usability and feasibility of this program based on the high attendance rates, SUS scores, and self-reported participant experience. In addition, we found impressive changes in digitally collected measures of functional improvement in most participants, suggesting that the program increased strength and mobility and may reduce the risk of falls. Clinically, the 8% - 20% improvement in measurement after 12 weeks is notable, especially considering recruitment generally was limited to those with limited risk factors for falling and have higher levels of function and therefore less scope for improvement or gain. While exploratory in nature, these gains are promising to show improvements across this intervention even in a relatively lower risk group of older adults.

This 12-week fall prevention program was well accepted and showed promise related to reducing fall risk in an older adult population. The usability and safety (as evidenced by no adverse events) of the program indicates that digitally delivered real-time exercise instruction has potential in this population to reduce barriers and increase availability of fall prevention programs. While those enrolled were relatively low risk for falls (as evidenced by the STEADI score used for eligibility), results should be contextualized that even those who were generally doing reasonably well have opportunity to improve. Indeed, in other populations, such as older adults with hip osteoarthritis, a change in TUG of around 1 second and a 30-second chair stand of around 1 second and 2 stands, respectively as meaningful [46], and between 0.9s and 6.0s depending upon the quality of life assessment tool in patients who were postoperative [47]. While our results of TUG change of 0.8 seconds (normal) is less than these noted changes, given the health status of our enrolled participants and limited 12-week intervention, the improvements are still notable. Furthermore, our improvement of 3.5 stands in 30 seconds, especially from a baseline of 14.5 stands—a notable good baseline—shows the resilience and opportunity for improvement in lower extremity strength regardless of starting health status.

However, this study is not without limitations. This project recruited heavily from an environment with a relatively high

XSL•FO
RenderX

socioeconomic status. As such, there may be a bias in enrolling individuals who already have digital tools like smartphones, tablets, and laptop computers and a high level of experience with digital technology, and therefore, a better understanding how to use these systems. As such, these participants may not be representative of the larger older adult population who may not have the same level of access to or understanding of such digital systems. We also did not ask specifically about their experience with previous technologies, though participants all did have sufficient equipment for enrollment. We also primarily enrolled women, which limits interpretability and generalizability of findings, especially given recent data suggesting higher fall-related impairment among older adult men [2]. Despite this, the use of a validated scale focused on usability commonly used across several domains related to eHealth and technology deployments [48], findings related to usability are supported; however SUS deployment in older adults has been challenged in some situations, namely those with cognitive impairment [49]. While we did not specifically screen for this in our enrollment, it is possible findings may be diluted. Furthermore, while the reasons remain unclear, the SUS average may be different and higher in digital physical activity interventions [48], making contextualizing results important; although, such data in older adult subpopulations remain limited. While we also did not only enroll participants who had experience with Zoom, it is possible that interested individuals may not have expressed willingness to enroll who were otherwise intimidated by such platforms, although we did not anecdotally experience this feedback. Subjective survey experience reports should be interpreted with this consideration [27]. Finally, it is not clear if digitally collected measures of physical performance, specifically the TUG and 30 second chair rise, are valid and reliable when compared with measures gathered in the clinic or laboratory. As such, the changes in physical performance demonstrated here, while promising, should be viewed as preliminary and not necessarily indicative of the same degree of change as would be expected were the data gathered in a more traditional context. However, given the rate of technology availability has continued to increase, attention is warranted to understand how these interfaces may be used for older adults [27].

In conclusion, this pilot showed that a digital interface is both usable and feasible but also suggested promise at improving functional measures associated with fall risk. Based on the increasing yearly number of falls and the relatively small number of participants who are engaging in fall-prevention programs, it seems clear that existing in-person exercise delivery modalities are insufficient. Strong Foundations offers a safe alternative to provide a real-time fall prevention program. Furthermore, the real-time feedback likely provides participants with added sense of purpose and value in their investment in improving their functional ability and therefore provide increased motivation for sustained behavior. Given the growing population and epidemiology of falls, having numerous options that enable engagement of older adults into high quality exercise programs designed to improve strength, postural alignment, and balance is increasingly important. Strong Foundations has promise to augment existing opportunities in this population and further leverages digital tools to increase the breadth of delivery potential. Ongoing attention and research is needed to both understand the scalability and sustainability of this program and to confirm that the observed changes in function and mobility through validated laboratory-based methods are replicated.

Acknowledgments

This pilot was funded by an internal grant obtained for pilots, partially supported by the National Institutes of Health (grant UL1TR001442). Generative artificial intelligence was not used in any aspect of this manuscript.

Data Availability

The datasets generated during or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

RM helped inform the design of the intervention, participated in study deployment, institutional review board (IRB) approval, manuscript generation, and data evaluation. DW helped develop the intervention; helped with IRB approval, manuscript review, data analysis; and performed the intervention and collected measures. JN helped develop the intervention, provided the intervention, informed and reviewed the manuscript, and provided guidance and support. HD helped inform the intervention, provided the intervention, and helped with data collection and cleaning. HB helped inform the intervention, provided the intervention, and helped collect measurements related to the intervention. The authors would like to acknowledge the Exercise and Physical Activity Resource Center (EPARC) group at large for contributing to this pilot, as well as the Altman Clinical and Translational Research Institute (ACTRI) Dissemination and Implementation Science Center (DISC) for providing funding to allow this pilot to occur.

Conflicts of Interest

None declared.

References

XSL•FO
RenderX

1. Burns E, Kakara R, Moreland B. A CDC Compendium of Effective Fall Interventions: What Works for Community-Dwelling Older Adults, 4th edition: Centers for Disease Control and Prevention (US); 2023. URL: https://stacks.cdc.gov/view/cdc/124200 [accessed 2025-02-17]
2. Kakara R, Bergen G, Burns E, Stevens M. Nonfatal and fatal falls among adults aged ≥65 Years - United States, 2020-2021. MMWR Morb Mortal Wkly Rep 2023 Sep 1;72(35):938-943. [doi: 10.15585/mmwr.mm7235a1] [Medline: 37651272]
3. Tinetti ME, Speechley M, Ginter SF. Risk factors for falls among elderly persons living in the community. N Engl J Med 1988 Dec 29;319(26):1701-1707. [doi: 10.1056/NEJM198812293192604] [Medline: 3205267]
4. Lundin-Olsson L, Nyberg L, Gustafson Y. Attention, frailty, and falls: the effect of a manual task on basic mobility. J Am Geriatr Soc 1998 Jun;46(6):758-761. [doi: 10.1111/j.1532-5415.1998.tb03813.x] [Medline: 9625194]
5. Baloh RW, Corona S, Jacobson KM, Enrietto JA, Bell T. A prospective study of posturography in normal older people. J Am Geriatr Soc 1998 Apr;46(4):438-443. [doi: 10.1111/j.1532-5415.1998.tb02463.x] [Medline: 9560065]
6. Espinosa de Los Monteros K, Gallo LC, Elder JP, Talavera GA. Individual and area-based indicators of acculturation and the metabolic syndrome among low-income Mexican American women living in a border region. Am J Public Health 2008 Nov;98(11):1979-1986. [doi: 10.2105/AJPH.2008.141903] [Medline: 18799765]
7. Gill TM. Assessment of function and disability in longitudinal studies. J Am Geriatr Soc 2010 Oct;58 Suppl 2(Suppl 2):S308-S312. [doi: 10.1111/j.1532-5415.2010.02914.x] [Medline: 21029059]
8. Burns E, Kakara R. Deaths from falls among persons aged ≥65 Years — United States, 2007–2016. MMWR Morb Mortal Wkly Rep 2018 May 11;67(18):509-514. [doi: 10.15585/mmwr.mm6718a1] [Medline: 29746456]
9. Vaishya R, Vaish A. Falls in older adults are serious. Indian J Orthop 2020 Feb;54(1):69-74. [doi: 10.1007/s43465-019-00037-x] [Medline: 32257019]
10. McAninch J, Greene C, Sorkin JD, Lavoie MC, Smith GS. Higher psychological distress is associated with unintentional injuries in US adults. Inj Prev 2014 Aug;20(4):258-265. [doi: 10.1136/injuryprev-2013-040958] [Medline: 24174466]
11. Judy A, Stevens P, Burns E. A CDC Compendium of Effective Fall Interventions: What Works for Community-Dwelling Older Adults, 3rd edition: Centers for Disease Control and Prevention (US); 2015. URL: https://stacks.cdc.gov/view/cdc/30689 [accessed 2025-02-27]
12. Stevens JA, Phelan EA. Development of STEADI: A fall prevention resource for health care providers. Health Promot Pract 2013 Sep;14(5):706-714. [doi: 10.1177/1524839912463576] [Medline: 23159993]
13. Sarmiento K, Lee R. STEADI: CDC's approach to make older adult fall prevention part of every primary care practice. J Safety Res 2017 Dec;63:105-109. [doi: 10.1016/j.jsr.2017.08.003] [Medline: 29203005]
14. Recommendation: falls prevention in community-dwelling older adults: interventions. United States Preventive Services Taskforce. URL: https://www.uspreventiveservicestaskforce.org/uspstf/recommendation/falls-prevention-community-dwelling-older-adults-interventions [accessed 2025-02-13]
15. US Preventive Services Task Force, Grossman DC, Curry SJ, et al. Interventions to prevent falls in community-dwelling older adults: US Preventive Services Task Force recommendation statement. JAMA 2018 Apr 24;319(16):1696-1704. [doi: 10.1001/jama.2018.3097] [Medline: 29710141]
16. Casey CM, Parker EM, Winkler G, Liu X, Lambert GH, Eckstrom E. Lessons learned from implementing CDC's STEADI falls prevention algorithm in primary care. Gerontologist 2017 Aug;57(4):787-796. [doi: 10.1093/geront/gnw074] [Medline: 27130270]
17. Wenger NS, Solomon DH, Roth CP, et al. The quality of medical care provided to vulnerable community-dwelling older patients. Ann Intern Med 2003 Nov 4;139(9):740-747. [doi: 10.7326/0003-4819-139-9-200311040-00008] [Medline: 14597458]
18. Stevens JA, Ballesteros MF, Mack KA, Rudd RA, DeCaro E, Adler G. Gender differences in seeking care for falls in the aged medicare population. Am J Prev Med 2012 Jul;43(1):59-62. [doi: 10.1016/j.amepre.2012.03.008] [Medline: 22704747]
19. Keadle SK, McKinnon R, Graubard BI, Troiano RP. Prevalence and trends in physical activity among older adults in the United States: A comparison across three national surveys. Prev Med 2016 Aug;89:37-43. [doi: 10.1016/j.ypmed.2016.05.009]
20. Belza B, Shumway-Cook A, Phelan EA, Williams B, Snyder SJ, LoGerfo JP. The effects of a community-based exercise program on function and health in older adults: the EnhanceFitness program. J Appl Gerontol 2006 Aug;25(4):291-306. [doi: 10.1177/0733464806290934]
21. Bethancourt HJ, Rosenberg DE, Beatty T, Arterburn DE. Barriers to and facilitators of physical activity program use among older adults. Clin Med Res 2014 Sep;12(1-2):10-20. [doi: 10.3121/cmr.2013.1171] [Medline: 24415748]
22. Pelletier CA, White N, Duchesne A, Sluggett L. Barriers to physical activity for adults in rural and urban Canada: A cross-sectional comparison. SSM Popul Health 2021 Dec;16:100964. [doi: 10.1016/j.ssmph.2021.100964] [Medline: 34841038]
23. Cosco TD, Fortuna K, Wister A, Riadi I, Wagner K, Sixsmith A. COVID-19, social isolation, and mental health among older adults: a digital catch-22. J Med Internet Res 2021 May 6;23(5):e21864. [doi: 10.2196/21864] [Medline: 33891557]
24. Raney L, Bergman D, Torous J, Hasselberg M. Digitally driven integrated primary care and behavioral health: how technology can expand access to effective treatment. Curr Psychiatry Rep 2017 Sep 30;19(11):86. [doi: 10.1007/s11920-017-0838-y] [Medline: 28965319]

XSL•FO
RenderX

25. Exposing the hidden connectivity crisis for older adults. : AARP; 2022 URL: https://oats.org/wp-content/uploads/2021/01/Aging-Connected-Exposing-the-Hidden-Connectivity-Crisis-for-Older-Adults.pdf [accessed 2025-02-17]
26. Mace RA, Mattos MK, Vranceanu AM. Older adults can use technology: why healthcare professionals must overcome ageism in digital health. Transl Behav Med 2022 Dec 30;12(12):1102-1105. [doi: 10.1093/tbm/ibac070] [Medline: 36073770]
27. Share of those 65 and older who are tech users has grown in the past. Pew Research Center. URL: https://www.pewresearch.org/short-reads/2022/01/13/share-of-those-65-and-older-who-are-tech-users-has-grown-in-the-past-decade/ [accessed 2025-02-13]
28. Jacobson CL, Foster LC, Arul H, Rees A, Stafford RS. A Digital Health Fall Prevention Program for Older Adults: Feasibility Study. JMIR Form Res 2021 Dec 23;5(12):e30558. [doi: 10.2196/30558] [Medline: 34837492]
29. S Ann Earon PD. Video in Healthcare Communications.: Zoom; 2022 URL: https://explore.zoom.us/docs/doc/Video_Communications_in_Healthcare.pdf
30. Orton E, Lafond N, Skelton DA, et al. Implementation fidelity of the Falls Management Exercise Programme: a mixed methods analysis using a conceptual framework for implementation fidelity. Public Health (Fairfax) 2021 Aug;197:11-18. [doi: 10.1016/j.puhe.2021.05.038] [Medline: 34271270]
31. Bansal S, Katzman WB, Giangregorio LM. Exercise for improving age-related hyperkyphotic posture: A systematic review. Arch Phys Med Rehabil 2014 Jan;95(1):129-140. [doi: 10.1016/j.apmr.2013.06.022] [Medline: 23850611]
32. Sherrington C, Michaleff ZA, Fairhall N, et al. Exercise to prevent falls in older adults: an updated systematic review and meta-analysis. Br J Sports Med 2017 Dec;51(24):1750-1758. [doi: 10.1136/bjsports-2016-096547] [Medline: 27707740]
33. Hopewell S, Adedire O, Copsey BJ, et al. Multifactorial and multiple component interventions for preventing falls in older people living in the community. Cochrane Database Syst Rev 2018 Jul 23;7(7):CD012221. [doi: 10.1002/14651858.CD012221.pub2] [Medline: 30035305]
34. Stay independent brochure. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/steadi/pdf/steadi-brochure-stayindependent-508.pdf [accessed 2025-02-13]
35. Brooke J. SUS—A quick and dirty usability scale. In: Jordan PW, Thomas B, Weerdmeester B, McClelland AL, editors. Usability Evaluation in Industry: Taylor and Francis; 1996.
36. Maramba I, Chatterjee A, Newman C. Methods of usability testing in the development of eHealth applications: A scoping review. Int J Med Inform 2019 Jun;126:95-104. [doi: 10.1016/j.ijmedinf.2019.03.018] [Medline: 31029270]
37. Assessment timed up and go (TUG). Centers for Disease Control and Prevention. URL: https://www.cdc.gov/steadi/media/pdfs/steadi-assessment-tug-508.pdf [accessed 2025-02-13]
38. Barry E, Galvin R, Keogh C, Horgan F, Fahey T. Is the Timed Up and Go test a useful predictor of risk of falls in community dwelling older adults: a systematic review and meta-analysis. BMC Geriatr 2014 Feb 1;14(1):14-14. [doi: 10.1186/1471-2318-14-14] [Medline: 24484314]
39. Montero-Odasso M, van der Velde N, Martin FC, et al. World guidelines for falls prevention and management for older adults: a global initiative. Age Ageing 2022 Sep 2;51(9):afac205. [doi: 10.1093/ageing/afac205] [Medline: 36178003]
40. 30-second chair stand. Centers for Disease Control and Prevention. 2017. URL: https://www.cdc.gov/steadi/media/pdfs/STEADI-Assessment-30Sec-508.pdf [accessed 2025-02-13]
41. Lewis JR. The System Usability Scale: past, present, and future. Int J Hum–Comp Interaction 2018 Jul 3;34(7):577-590. [doi: 10.1080/10447318.2018.1455307]
42. Smith ML, Towne SD, Herrera-Venson A, et al. Delivery of fall prevention interventions for at-risk older adults in rural areas: findings from a national dissemination. Int J Environ Res Public Health 2018 Dec 10;15(12):2798. [doi: 10.3390/ijerph15122798] [Medline: 30544658]
43. Marzuca-Nassr GN, Alegría-Molina A, SanMartín-Calísto Y, et al. Muscle mass and strength gains following resistance exercise training in older adults 65-75 years and older adults above 85 years. Int J Sport Nutr Exerc Metab 2024 Jan 1;34(1):11-19. [doi: 10.1123/ijsnem.2023-0087] [Medline: 37875254]
44. Lai X, Bo L, Zhu H, et al. Effects of lower limb resistance exercise on muscle strength, physical fitness, and metabolism in pre-frail elderly patients: a randomized controlled trial. BMC Geriatr 2021 Jul 30;21(1):447. [doi: 10.1186/s12877-021-02386-5] [Medline: 34330211]
45. Gusi N, Carmelo Adsuar J, Corzo H, Del Pozo-Cruz B, Olivares PR, Parraca JA. Balance training reduces fear of falling and improves dynamic balance and isometric strength in institutionalised older people: a randomised trial. J Physiother 2012;58(2):97-104. [doi: 10.1016/S1836-9553(12)70089-9] [Medline: 22613239]
46. Wright AA, Cook CE, Baxter GD, Dockerty JD, Abbott JH. A comparison of 3 methodological approaches to defining major clinically important improvement of 4 performance measures in patients with hip osteoarthritis. J Orthop Sports Phys Ther 2011 May;41(5):319-327. [doi: 10.2519/jospt.2011.3515] [Medline: 21335930]
47. Gautschi OP, Stienen MN, Corniola MV, et al. Assessment of the minimum clinically important difference in the timed up and go test after surgery for lumbar degenerative disc disease. Neurosurgery 2017 Mar 1;80(3):380-385. [doi: 10.1227/NEU.0000000000001320] [Medline: 27352275]
48. Hyzy M, Bond R, Mulvenna M, et al. System Usability Scale benchmarking for digital health apps: meta-analysis. JMIR Mhealth Uhealth 2022 Aug 18;10(8):e37290. [doi: 10.2196/37290] [Medline: 35980732]

XSL•FO
RenderX

49. Holden RJ. A Simplified System Usability Scale (SUS) for cognitively impaired and older adults. Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care 2020 Sep;9(1):180-182. [doi: 10.1177/2327857920091021]

Abbreviations

30 CS: 30-second chair stand
CDC: Centers for Disease Control and Prevention
EPARC: Exercise and Physical Activity Resource Center
STEADI: Stopping Elderly Accidents and Deaths Initiative
SUS: System Usability Scale
TUG: timed up and go
USPSTF: United States Preventive Services Task Force

Edited by A Mavragani; submitted 10.10.24; peer-reviewed by H Hawley-Hague, M Fraser; revised version received 28.01.25; accepted 29.01.25; published 28.02.25.

Please cite as:
Moran R, Wing D, Davey H, Barkai H, Nichols J
Development and Implementation of Strong Foundations, a Digitally Delivered Fall Prevention Program: Usability and Feasibility Pilot Exercise Cohort Study
JMIR Form Res 2025;9:e67406
URL: https://formative.jmir.org/2025/1/e67406
doi:10.2196/67406

© Ryan Moran, David Wing, Hope Davey, Hava Barkai, Jeanne Nichols. Originally published in JMIR Formative Research (https://formative.jmir.org), 28.2.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Evaluating the Knowledge Level, Practice, and Behavioral Change Potential of Care Managers in Pressure Injury Prevention Using a Mobile App Prototyping Model in the Home-Care Setting: Single-Arm, Pre-Post Pilot Study

Masushi Kohta[1*], PhD; Mayumi Takahashi[2*], CWOCN, RN; Hiroe Koyanagi[1*], CWOCN, RN, PhD; Junko Sugama[1*], RN, PhD

[1]Research Center for Implementation Nursing Science Initiative, Fujita Health University, Toyoake, Japan

[2]Department of Nursing, Akabane Central General Hospital, Kita-ku, Tokyo, Japan

[*]all authors contributed equally

Corresponding Author:
Masushi Kohta, PhD
Research Center for Implementation Nursing Science Initiative, Fujita Health University, Toyoake, Japan

Abstract

Background: The use of mobile apps to promote knowledge level, practice, and behavioral change potential has become increasingly common. However, studies on apps targeting social welfare employees working in the home-care setting to prevent pressure injury (PI) are lacking. The care manager (CM) plays a key role in connecting the demand and supply of home-care services. PI is more prevalent in the home-care setting, where resources are limited, than in acute settings.

Objective: The research hypothesis was that CMs who use a mobile app will have improved general knowledge and heightened practice for PI prevention, compared to that before using the app. This study aimed to assess the effectiveness of a PI prevention support mobile app prototyping model (Pips-Map) in improving the knowledge level, practice, and behavioral change potential of CMs in PI prevention in the home-care setting.

Methods: This was conducted between December 2021 and December 2023 as a single-arm, pre-post pilot study including 27 CMs who worked in a Japanese city. Pips-Map was used for 6 months in daily practice, and a self-administered test questionnaire was used to assess participants' knowledge and practice in PI prevention before or after using Pips-Map. At the end of the posttest, a validated App Behavior Change Scale was used to analyze behavioral change potential. This study followed the Consolidated Standards of Reporting Trials (CONSORT) extension to pilot and feasibility trials.

Results: In total, 19 participants were analyzed. Out of 55 points, the total mean knowledge score significantly increased from 30.9 (SD 5.9) in the pretest group to 36.1 (SD 5.9) in the posttest group (P=.0003). The number of participants with a total score of >70% (adequate knowledge level) increased from 2 (11%) to 7 (36.8%), but the difference was not statistically significant (P=.07). For the level of practice, out of 21 points, the total score increased from 15.2 (SD 3.1) in the pretest group to 16.2 (SD 3.0) in the posttest group, but no statistically significant differences were observed (P=.16). The behavior change scale revealed that participants positively evaluated the Pips-Map to provide information on PI prevention guidelines but had concerns regarding inadequate usability and financial incentives of Pips-Map.

Conclusions: The use of Pips-Map for 6 months in actual practice increased the knowledge level of Japanese CMs in PI prevention, but it did not change the level of practice. Considering the need for updating apps that aim to promote behavioral change, this study identified some limitations of Pips-Map. Thus, revisions must be made to adapt Pips-Map to home-based care needs.

Trial Registration: UMIN Clinical Trial Registry UMIN000048904; https://tinyurl.com/5bt6zv7t

(JMIR Form Res 2025;9:e57768) doi:10.2196/57768

KEYWORDS

behavioral change; home care; knowledge; mobile application; pressure injury; mHealth; mobile health; apps; practice; injury; prevention; prototype; effectiveness; care manager; Japan; Pips-Map; questionnaire; wound care; pilot study; women

XSL•FO
RenderX

Introduction

Backgrounds

Many developed countries are facing the challenge of a declining birthrate and aging population. Considering that Japan has the world's most advanced aging population, the Japanese government has shifted the focus of health care from hospital- to home- or community-based care provided in the home-care setting [1]. To provide comprehensive community-based care, the Japan-specific geriatric care system (Community-based Integrated Care System, a medical- and social-welfare networking model) works with the Japan-specific long-term care insurance system. The care manager (CM), who possesses specialized knowledge necessary to help individuals in need of assistance or care to achieve more independent lives, plays a key role in connecting the demand and supply of long-term care services [2]. A workflow of CMs in home-based care services was summarized in a recent study [3]. The CM is responsible for providing care management services, including checking the physical condition and living circumstances of individuals, establishing and adjusting care plans according to individual needs, and monitoring individual care. To obtain a CM license, an individual must have ≥5 years of geriatric care experience as a medical professional (physician or nurse) and social welfare professional (care worker or social worker) and should have passed the examination. Compared to CMs with medical backgrounds, the number of CMs with social welfare backgrounds is higher, because obtaining a CM license is a career advancement step. Compared with CMs who studied medicine, those from the social welfare background may not have sufficient knowledge or practice regarding the prevention and treatment of chronic diseases, such as pressure injury (PI) [2].

Although PIs are pervasive in all age groups, most of them occur in older individuals with immobility and chronic diseases, such as heart disease, respiratory illness, and dementia [4]. PI can result in poor quality of life for patients and socioeconomic burden [5,6]. PI is more prevalent in the home-care setting, where resources are limited, than in acute settings, with prevalence rates of 1.76% - 26% [7-9]. Although there is broad consensus that some PIs are unavoidable, most are considered preventable [10]. A paper-based risk assessment tool for preventing PIs that can be used by Japanese CMs has recently been developed and implemented [11-13]; however, its application to digital technology is still under consideration.

Prior Work

Digital technologies have become an emerging branch of medicine, with technology-based applications being used to improve disease knowledge and user activations being modified by relying on health-related information [14-16]. Mobile apps for PIs can be grouped into apps for PI treatment and apps for PI prevention. European wound management experts developed a clinical support mobile app based on a global wound care framework to implement care plans and promote wound healing [17]. Do Khac et al [18] evaluated a free-access mobile health app compared with standard technique to assess its reliability and validity of measurement of surface area of PIs in patients with spinal cord injury. In Japan, a PI treatment support mobile app was developed for sharing PI information with a multidisciplinary care team or for conducting remote consultation with wound care specialists [19,20]. A smartphone app focusing on PI prevention education was developed to evaluate the PI knowledge of informal carers and app acceptability and usability [21]. A guideline app was designed to present evidence-based guideline recommendations with easy access [22]. Amann et al [23] qualitatively investigated barriers and facilitators to the adoption of a self-management app for PI prevention in individuals with spinal cord injury. However, studies assessing a mobile app in improving knowledge and behavioral change within social welfare workers, such as CMs, who work at home-based care facilities, are lacking.

Aim of This Study

The research hypothesis is that CMs who experience a mobile app will improve general knowledge and heightened practice about PI prevention compared to before the experience. The primary objective was to investigate the effects of CMs on improving knowledge level when using a PI prevention support mobile app prototyping model (Pips-Map), which has been developed by the industry. In addition to improving the knowledge level, we surveyed whether the practice level of CMs for PI prevention was heightened when using Pips-Map as a secondary objective. We further investigated to determine whether the use of the app promoted behavior change in PI prevention among CMs and which specific features of the app positively or negatively influenced the behavior change.

Methods

Prototyping Model

Pips-Map was designed and developed by ALCARE Co. Ltd. (Tokyo, Japan) to support research and practice in wound care, home-based care, etc. According to the developer, the key desktop design concept of Pips-Map was to facilitate information sharing among home health care providers with different specialties and change their behavior for preventing PI development in patients who received assistance or nursing care at home. Pips-Map is unique because it facilitates information sharing among home health care providers, including CMs, home-care nurses, and primary physicians for each home-based patient. Displaying products formulary for preventing PI in Pips-Map helps in selecting products for each patient in a timely manner. Pips-Map has 5 functions: (a) photographing skin using a standard smartphone or tablet camera; (b) assessing PI risk according to the Japanese Ohura-Hotta scale; (c) entering data on prophylactic planning or comments based on PI risk assessment; (d) displaying products (eg, pressure redistribution mattress, skin care products, prophylactic dressings) formulary; and (e) data sharing between medical and social welfare professionals for each patient. Pips-Map includes a software system named Medical Care Station, which was developed by Embrace Co., Ltd. (Tokyo, Japan), to provide patient information. The system is a security-conscious, completely private social networking service that shares information in a timeline format. Patient data are stored on the system but not linked to the electronic medical records of each site. Selected

XSL•FO
RenderX

screenshots from Pips-Map are presented in Multimedia Appendix 1.

Study Design

This prospective, single-arm, interventional pilot study was conducted at a university from December 2021 to December 2023. There were no major changes in study methods after the trial commencement. This study followed the Consolidated Standards of Reporting Trials (CONSORT) extension to pilot and feasibility trials [24] from the Equator Network, which assess the feasibility of conducting a future definitive randomized controlled trial rather than the effectiveness or efficacy of an intervention. As an editorial, Lancaster et al have provided some general guidance on how to report non-randomized pilot and feasibility studies, including a pre-post test, using the CONSORT extension to pilot and feasibility trials [25]. The checklists for the reporting guideline in this study were provided in Multimedia Appendix 2.

Setting and Sample

The accessible population included individuals with a CM license and those who participated in the "Kita-ku Care Managers' Association" in Kita-ku, Tokyo (n=185). According to statistics from the Japanese Ministry of Health, Labor and Welfare in the year 2022, the number of CMs working full-time at home-care support facilities is 99,220. Of these, 8,943 live in Tokyo [26]. Kita-ku lies in the northeastern region of Tokyo and is one of the 23 districts. The Kita-ku Care Managers' Association works in Kita-ku, and the number of CMs accounts for approximately 2.1% (185/8943) of them in Tokyo. Japanese CMs could use an Android- and iOS-based mobile application. As power analysis could not be performed, this study was considered a pilot study. Pilot studies are needed to obtain preliminary data for calculating sample size for the primary outcome [27].

Study Procedure

Kita-ku Care Managers' Association holds regular meetings for information exchange and acquisition of knowledge about home-care management. In this meeting, we orally explained the details of this study after obtaining agreement for cooperation with this study from the member of the executive committee of the Kita-ku Care Managers' Association. A total of 27 individuals agreed to participate, and the pretest was completed. Pips-Map was downloaded and operated for 6 months in daily practice. The participants were instructed on how to use Pips-Map and record the details of PI prevention in Pips-Map through web-based guidance. Technical support on how to use Pips-Map was provided by the researcher through telephone and e-mail.

A pretest or posttest was performed to measure participants' knowledge level and practice level in PI prevention. The pretest was conducted on the same day when informed consent was obtained, and the posttest was conducted within a month after using Pips-Map. At the end of the posttest, an additional questionnaire survey was conducted to assess Pips-Map to promote behavioral change potential. All tests and questionnaires were made available via web-based Google Forms. The completion of pretest and posttest and additional questionnaire required approximately 20 and 10 minutes, respectively.

Measurement

General Characteristics of Participants

The language used for the questionnaire was Japanese. The questionnaire was divided into 4 parts. Part 1 contained general characteristics of participants, including age range, sex, years of experience as a CM, professional background, and education level.

Primary Outcome

In part 2, participant knowledge on PI prevention was evaluated. Although 2 validated knowledge tests for PI prevention have been reported [28,29], most respondents in these studies were medical professionals, such as nurses who worked at hospitals. The test questions were difficult to answer for social welfare professionals such as Japanese CMs. Therefore, the current self-administered test questionnaire was developed via collaboration between a nurse with certification in wound, ostomy, and continence working at the home-care division of a hospital and a senior assistant professor certified in wound, ostomy, and continence nursing. For questionnaire development, they referred to a previous report [30] and the Japanese PI guideline created by the Japanese Society of Pressure Ulcers [31]. This part consisted of 55 test questions reflecting 6 themes: (1) anatomy, (2) risk assessment, (3) pressure, (4) friction and shear, (5) moisture, and (6) nutritional support. Answering options and scoring rules varied depending on items. For items 1 - 4, 7, and 11 - 13, the participants were instructed to select "true," "false," or "I do not know." Each correct answer was assigned one point, and no points were assigned to wrong answers and "I do not know" answers. For items 5, 9, and 14, the participants selected "know" or "do not know." A score of 1 was assigned for "know" and 0 for "do not know." For items 6, 8, and 10, the multiple-choice questions had several options along with the "I do not know" option. A correct answer was assigned a score of 1, and an incorrect answer was assigned 0. The maximum score was 55 (Multimedia Appendix 3).

Secondary Outcomes

In part 3, the participants were asked about the practice of preventive strategies for PIs. This part consisted of 21 test questions reflecting 6 sections: (1) risk assessment, (2) pressure, (3) friction and shear, (4) moisture, (5) nutritional support, and (6) communication with other professionals. The test questionnaire in this section asked whether the participants implemented those actions or not. The answer choice for each question was "practice" or "do not practice." One or zero point was given when the participant answered "practice" or "do not practice," respectively. The maximum practice score was 21 in this section (Multimedia Appendix 4).

In part 4, behavioral change potential was measured using a validated App Behavior Change Scale (ABACUS) [32] with a minor modification. For the modification, the item "does the app provide a material or social reward or incentive?" in the original paper was divided into the following 2 items: "does the app provide a financial incentive?" and "does the app provide

XSL•FO
RenderX

a material or social reward?" The scale comprised 22 questions regarding a determinant of behavioral change potential of mobile apps, each on a 5-point Likert scale (1="strongly disagree" to 5="strong agree") with total score 22 - 110. A higher score indicated higher levels of promoting behavioral change by app usage. Previous literature showed that potential behavioral change of users of apps was evaluated by ABACUS in a wide range of research fields, including depression, diabetes, osteoarthritis, and physiotherapy care [33,34].

Data Analysis

For general characteristics, all data were categorical variables and expressed as frequencies and percentages. For level of knowledge, as a primary outcome, a dependent *t*-test (for continuous variables), the Wilcoxon signed-rank test (for ordinal variables), and McNemar's test (for categorical variables) were conducted to compare differences between each item and total score between the pretest and posttest groups. Ordinal variables with ≥5 categories were treated as continuous variables. A study on the learning outcome criteria for PI prevention considered an average total score of >70% satisfactory [35]. As a secondary outcome, similar analysis was conducted for the level of practice. For behavioral change potential, mean and SD were reported in each item and total score of ABACUS. All statistical analyses were conducted using the Statistical Package for the Social Sciences version 28.0 (IBM Corporation, Armonk, NY, USA). *P* values of <.05 were considered to indicate statistical significance.

Ethical Considerations

This study was approved by the university's ethics review board (approval number HM21-335). Electronic informed consent was obtained from all participants before participation. All experiments in this research were performed according to the principles of the Declaration of Helsinki. Participant data were kept confidential. This study has been registered in the Clinical Trials Registry (UMIN Clinical Trial Registry number: UMIN000048904) and is accessible online. Each participant who completed the questionnaire was given a gift certificate (5000 JPN yen) as a financial reward.

Results

General Characteristics of Participants

In total, 19 participants were analyzed (Figure 1). The recruitment of participants for this study trial started on September 11, 2022, and the follow-up was completed on August 31, 2023. Most participants were in their 40 and 50 seconds, and 12 out of 19 participants (63%) were female. All of the participants had a background in social welfare; that is, participants who studied medicine or nursing were not included in this study (Table 1).

XSL•FO
RenderX

Figure 1. Flowchart of participant progression in pretest and posttest study. A total of 27 care managers who participated in the "Kita-ku Care Managers' Association" in a district in Tokyo agreed to participate. As an intervention, a pressure injury prevention support mobile app prototyping model was operated for 6 months in daily practice. A pretest or posttest was performed to measure participants' knowledge level and practice level in pressure injury prevention. After the posttest was completed, an additional questionnaire survey was conducted to assess the mobile app to promote behavioral change potential. In total, 19 participants were included in data analysis. This study was conducted from December 2021 to December 2023.

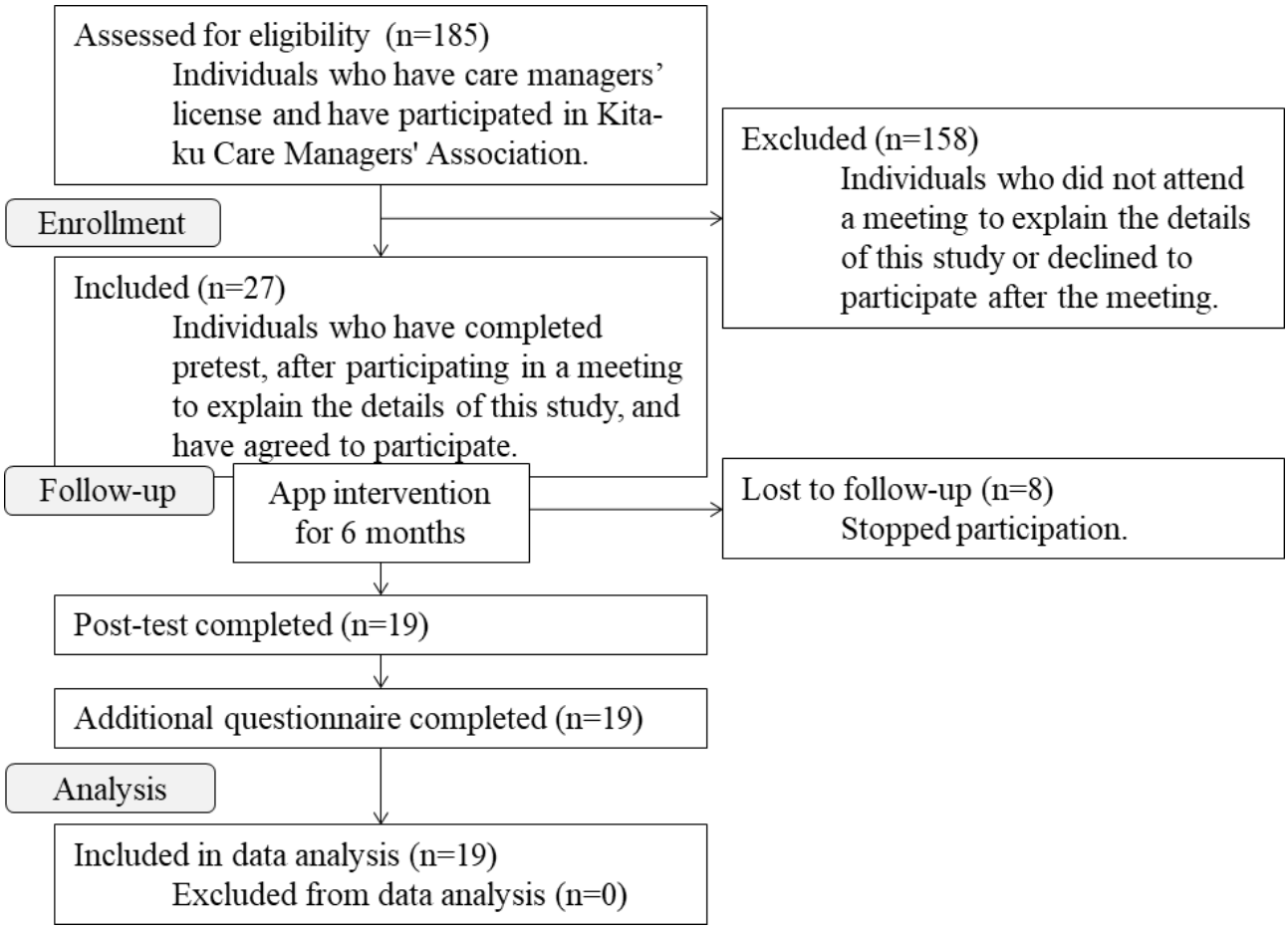

XSL•FO
RenderX

Table . General characteristics of participants. Under the Japan-specific long-term care insurance system, to obtain a care manager license, an individual must have ≥5 years of geriatric care experience as a medical professional (physician or nurse) and social welfare professional (care worker or social worker). Compared with care managers who studied medicine, those from the social welfare background may not have sufficient knowledge or practice regarding the prevention of pressure injury. All of the care managers who participated in this study had a professional background in social welfare, not medicine. As the primary outcome included knowledge levels of pressure injury prevention, the educational level for the study participants was an important characteristic to be confirmed.

Characteristics	Overall (N=19)
Age, n (%)	
30s	3 (16)
40s	7 (37)
50s	6 (32)
60s	3 (16)
Sex, n (%)	
Male	7 (37)
Female	12 (63)
Years of experience as a care manager, n (%)	
<5	3 (16)
≥5 and <10	4 (21)
≥10	12 (63)
Professional background, n (%)	
Social welfare	19 (100)
Medical	0 (0)
Education level, n (%)	
Junior high school	1 (6)
High school	8 (44)
Vocational school	5 (28)
Junior college	1 (6)
College	3 (17)

Primary Outcome Evaluation

As shown in Table 2, the total score of knowledge significantly increased in the posttest group compared with that in the pretest group (P=.0003). The number of participants with a total score >70% increased from 2 (11%) in the pretest group to 7 (37%) in the posttest group, but the differences were not statistically significant (P=.07). In each test questionnaire, “anatomy,” “risk assessment,” and “pressure” increased significantly, but the remaining items were similar between the 2 groups.

XSL•FO
RenderX

Table . Comparison of knowledge level between the pretest and posttest groups. for the questionnaire development to evaluate the participants' knowledge level in pressure injury prevention, the authors referred to a previous report [28] and the Japanese pressure injury guideline [29]. It consisted of 55 test questions reflecting 6 themes: (1) anatomy, (2) risk assessment, (3) pressure, (4) friction and shear, (5) moisture, and (6) nutritional support. The maximum score was 55. A pretest or posttest was performed to measure participants' knowledge level in pressure injury prevention. The pretest was conducted on the same day when informed consent was obtained, and the posttest was conducted within a month after intervention. The total score of knowledge significantly increased in the posttest group compared with that in the pretest group. Statistical analysis was conducted via those listed in footnotes b, d, and e.

Questionnaire	Pretest	Posttest	*P*-value
Anatomy, mean (SD[a])	16.2 (2.6)	19.2 (3.6)	.001[b]
Risk assessment, mean (SD)	0.3 (0.5)	1.2 (1.4)	.009[b]
Pressure, mean (SD)	6.3 (2.0)	7.4 (1.5)	.013[b]
Friction and shear, mean (SD)	4.7 (1.6)	4.8 (1.3)	.76[b]
Moisture, median (IQR[c])	2 (2–3)	3 (2–3)	.25[d]
Nutritional support, mean (SD)	1.2 (0.9)	1.1 (1.3)	.56[b]
Total score, mean (SD)	30.9 (5.9)	36.1 (5.9)	.0003[b]
Total score>70%, n (%)	2 (11)	7 (37)	.070[e]

[a]SD, standard deviation.

[b]dependent *t*-tests for continuous variables.

[c]IQR, interquartile range.

[d]Wilcoxon signed-rank test for ordinal variables, and.

[e]McNemar's test for categorical variables.

Secondary Outcome Evaluation

As shown in Table 3, the total score of practice increased from 15.2 (SD 3.1) in the pretest group to 16.2 (SD 3.0) in the posttest group, but no statistically significant differences were observed (*P*=.16). The score of "risk assessment" increased significantly (*P*=.030). The average total score for behavioral change potential using ABACUS was 71.04 (SD 11.55) (Multimedia Appendix 5). The highest score was achieved for the question "was the app created with expertise and does the app provide information consistent with national guidelines?" Lower scores (mean value of <3.0) were obtained for the questions "does the app provide financial incentives for the users and their organization?," "does the app have the ability to export data from app?," and "does the app allow or encourage practice or rehearsal, in addition to daily activities?"

XSL•FO
RenderX

Table . Comparison of level of practice between the pretest and posttest groups. The participants were asked about the practice of preventive strategies for pressure injuries. It consisted of 21 test questions reflecting 6 sections: (1) risk assessment, (2) pressure, (3) friction and shear, (4) moisture, (5) nutritional support, and (6) communication with other professionals. The maximum practice score was 21. The pretest was conducted on the same day when informed consent was obtained, and the posttest was conducted within a month after the intervention. No statistically significant differences were shown in the total score of practice between the posttest and posttest groups. Statistical analysis was conducted via a, c, and e.

Questionnaire	Pretest	Posttest	*P* value
Risk assessment, n (%)	4 (21)	6 (32)	.48[a]
Pressure, mean (SD[b])	6.8 (2.0)	7.9 (1.7)	.030[c]
Friction and shear, median (IQR[d])	2 (1–2)	2 (1–2)	.74[e]
Moisture, median (IQR)	2 (2–2)	2 (1–2)	.60[e]
Nutritional support, mean (SD)	2.7 (1.0)	2.6 (1.0)	.84[c]
Communication with other professionals, median (IQR)	2 (2–3)	2 (2–2)	.60[e]
Total score, mean (SD)	15.2 (3.1)	16.2 (3.0)	.16[c]
Total score>70%, n (%)	10 (53)	11 (58)	.68[a]

[a]McNemar's test for categorical variables.

[b]SD, standard deviation.

[c]dependent *t*-tests for continuous variables, and.

[d]IQR, interquartile range.

[e]Wilcoxon signed-rank test for ordinal variables.

Discussion

Principal Findings

The use of a mobile app provides a more engaging and accessible approach to clinical practice at home care, fitting well with the evolving demands of the digital ages. The development of a PI prevention support mobile app and its evaluation as an educational tool has already been reported; however, studies approaching a mobile app in improving knowledge, practice, and behavioral change within social welfare workers who work at home-care facilities are lacking. In this study, we used Pips-Map in an actual clinical setting, and the effect of improving general knowledge, heightened practice, and positively changing behavior among CMs in the prevention of PI development was investigated in a pretest-posttest study design. As the incidence of PI in the home-care setting impacts the quality of care provided, the adoption of good practice with adequate knowledge through improvement in PI prevention is necessary. Adequate knowledge is essential for assessing behavioral change. The key finding of this study is that the use of a mobile app in daily practice can increase knowledge about PI prevention among Japanese CMs. However, our results could not demonstrate that increased knowledge indicated improved practice level.

The most significant finding of this study is that Pips-Map is effective in improving the total score of knowledge level (Table 2). Consistently, a 6-week interventional study based on a mobile app focused on PI prevention education for informal carers of individuals at risk of PI [21]. Using a self-administered PI knowledge assessment questionnaire consisting of 4 categories (support surfaces, nutrition and hydration, keep moving, and skin care and inspection), they reported that the use of the app effectively increased knowledge level regarding PI prevention. CMs need to manage their care work with other responsibilities, and their busy schedule makes learning about PI prevention challenging. Active methodology with digital technologies enables flexibility for conducting these activities once the barrier of time and space has been overcome. Compared with conventional on-site education, such as through lectures or handouts, Pips-Map is more advantageous as a training tool in terms of accessibility, leading to the prevention of common PIs in the home-care setting.

However, changes in the participants' PI prevention practices were not observed (Table 3). This might be explained by our results to understand the behavioral change potential using ABACUS (Multimedia Appendix 5). A high score was obtained for item 2, "providing information consistent with national guidelines," which directly affects the increase in the knowledge of PI prevention. By contrast, lower scores were observed for item 13, "ability to export data from app," item 14, "providing a financial incentive," and item 19, "allowing for practice or rehearsal," which are significantly important items for CMs. In this current form, we believe that Pips-Map is didactic and inadequate in terms of customized content and usability.

As shown in the lowest mean score in ABACUS, the participants did not perceive receiving social or financial incentive by using the app. Although financial rewards are a powerful extrinsic motivation for app use, studies suggest that the most powerful plans for behavioral change include both extrinsic and intrinsic factors. For example, rankings, social recognition, and gaming are reported to be strong intrinsic factors relating to long-term app use [36]. In addition, the opinions raised from the behavioral change questionnaire regarding customized content and usability could not allow for adaptation of the app in the real-world

XSL•FO
RenderX

setting. Hence, we will provide feedback to the industry on the need to create a modified version of the app to address these issues for the development phase. In addition, we will investigate the impact of the modified version on behavioral change of the user and health care outcomes of individuals who need assistance or nursing care at home. Evidence on knowledge and practice of wound care is updated daily, and all the researchers have to keep up with the fast-moving pace of app technology.

Limitations

Our study has several limitations. First, our tests for knowledge and practice of PI prevention have not yet been validated. It would be better to use validated tests simultaneously; however, tests targeting Japanese CMs are lacking. The face validity, construct validity, and reliability of the knowledge and practice test questions used in this study were not considered, because we prioritized evaluating the effect of a mobile app prototype on knowledge, practice, and behavior change. As a future work, determining the reliability and validity of the newly developed test questions is a key imperative, which ultimately could benefit individuals who receive preventive care as well as social welfare professionals such as CMs who provide it.

Next, this single-arm, pre-post study aimed to preliminarily assess the effectiveness and feasibility of Pips-Map. Determining whether the value of the outcome observed upon intervention is better than the value without intervention is challenging. Finally, we did not perform power analysis to estimate the ideal sample size because an accurately estimated population could not be set before the start of this study. The sample size is relatively small, but it should be acceptable considering the design of a pilot study. The authors were unable to determine the cause for the dropout of study participants because it was difficult to chase them. For the effectiveness and usability of modified Pips-Map, comparative studies between intervention and nonintervention groups with appropriate sample sizes are required in the future.

Conclusions

Our results revealed that the use of a mobile app prototyping model for 6 months in actual practice increased the knowledge level of Japanese CMs in preventing PI development, but it did not affect the level of practice between the pretest and posttest groups. Considering the need for updating apps that aim to promote behavioral change, this study identified limitations associated with the app. A higher quality app must be developed that is suitable for the home-care setting.

Acknowledgments

We thank Eisaku Oba for his insightful advice on study procedure. We wish to thank ALCARE for their generous financial assistance. No generative AI and AI-assisted technologies were used in any portion of the manuscript writing process. We also thank Enago (http://www.enago.jp) for its English language editing services.

Data Availability

The datasets generated during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Concept of the pressure injury prevention support mobile application prototyping model (Pips-Map).
[PNG File, 182 KB - formative_v9i1e57768_app1.png]

Multimedia Appendix 2
CONSORT 2010 checklist of information to include when reporting a pilot or feasibility trial.
[DOC File, 233 KB - formative_v9i1e57768_app2.doc]

Multimedia Appendix 3
Pressure injury prevention knowledge test questionnaire.
[DOC File, 88 KB - formative_v9i1e57768_app3.doc]

Multimedia Appendix 4
Pressure injury prevention practice test questionnaire.
[DOC File, 80 KB - formative_v9i1e57768_app4.doc]

Multimedia Appendix 5
Results of app behavioral change scales.
[DOC File, 84 KB - formative_v9i1e57768_app5.doc]

XSL•FO
RenderX

References

1. Iwagami M, Tamiya N. The long-term care insurance system in Japan: past, present, and future. JMA J 2019 Mar 4;2(1):67-69. [doi: 10.31662/jmaj.2018-0015] [Medline: 33681515]
2. Goto Y, Miura H, Ito N. Comparison between the chief care manager and the normal care manager on hospitalization and discharge coordination activities in Japan: an online cross-sectional study of care managers in Aichi prefecture. Int J Environ Res Public Health 2022 Sep 25;19(19):12122. [doi: 10.3390/ijerph191912122] [Medline: 36231421]
3. Wang Z, Fukayama K, Niu B. Does long-term care literacy matter in evaluating older care recipients' satisfaction with care managers? Empirical evidence from Japanese survey data. Int J Environ Res Public Health 2023 Jan 30;20(3):2456. [doi: 10.3390/ijerph20032456] [Medline: 36767824]
4. Elli C, Novella A, Nobili A, Ianes A, Pasina L. Factors associated with a high-risk profile for developing pressure injuries in long-term residents of nursing homes. Med Princ Pract 2022;31(5):433-438. [doi: 10.1159/000527063] [Medline: 36122563]
5. Sebba Tosta de Souza DM, Veiga DF, Santos IDDAO, Abla LEF, Juliano Y, Ferreira LM. Health-related quality of life in elderly patients with pressure ulcers in different care settings. J Wound Ostomy Continence Nurs 2015;42(4):352-359. [doi: 10.1097/WON.0000000000000142]
6. Guest JF, Fuller GW, Vowden P, Vowden KR. Cohort study evaluating pressure ulcer management in clinical practice in the UK following initial presentation in the community: costs and outcomes. BMJ Open 2018 Jul 25;8(7):e021769. [doi: 10.1136/bmjopen-2018-021769] [Medline: 30049697]
7. Sari SP, Everink IH, Sari EA, et al. The prevalence of pressure ulcers in community-dwelling older adults: a study in an Indonesian city. Int Wound J 2019 Apr;16(2):534-541. [doi: 10.1111/iwj.13081] [Medline: 30768769]
8. Konya C, Takeuchi Y, Nakagami G, et al. Nationwide time-series surveys of pressure ulcer prevalence in Japan. In: Takeuchi Y, editor. J Wound Care 2022 Dec 1;31(Sup12):S40-S47. [doi: 10.12968/jowc.2022.31.Sup12.S40]
9. Bottega M, Tempesta M, Piovesan C, et al. The risk and prevalence of pressure injuries in older people in the home care service: a cross-sectional study. J Wound Care 2023 Oct 1;32(Sup10):ccxi-ccxviii. [doi: 10.12968/jowc.2023.32.Sup10.ccxi] [Medline: 37830802]
10. Schmitt S, Andries MK, Ashmore PM, Brunette G, Judge K, Bonham PA. WOCN Society position paper: avoidable versus unavoidable pressure ulcers/injuries. J Wound Ostomy Continence Nurs 2017;44(5):458-468. [doi: 10.1097/WON.0000000000000361] [Medline: 28877112]
11. Kohta M, Ohura T, Tsukada K, et al. Inter-rater reliability of a pressure injury risk assessment scale for home care: a multicenter cross-sectional study. J Multidiscip Healthc 2020;13:2031-2041. [doi: 10.2147/JMDH.S291162] [Medline: 33376343]
12. Kohta M, Ohura T, Okada K, et al. Convergent validity of three pressure injury risk assessment scales: comparing the PPRA-Home (Pressure Injury Primary Risk Assessment Scale for Home Care) to two traditional scales. J Multidiscip Healthc 2021;14:207-217. [doi: 10.2147/JMDH.S294734] [Medline: 33564237]
13. Nakamura Y, Kohta M. Predictive validity of a risk assessment scale for preventing pressure injuries in home care. Tenri Med Bull 2024;27(1):11-18. [doi: 10.12936/tenrikiyo.27-001]
14. Mosa ASM, Yoo I, Sheets L. A systematic review of healthcare applications for smartphones. BMC Med Inform Decis Mak 2012 Jul 10;12:67. [doi: 10.1186/1472-6947-12-67] [Medline: 22781312]
15. O'Leary KJ, Lohman ME, Culver E, Killarney A, Randy Smith G Jr, Liebovitz DM. The effect of tablet computers with a mobile patient portal application on hospitalized patients' knowledge and activation. J Am Med Inform Assoc 2016 Jan;23(1):159-165. [doi: 10.1093/jamia/ocv058] [Medline: 26078412]
16. Iribarren SJ, Akande TO, Kamp KJ, Barry D, Kader YG, Suelzer E. Effectiveness of mobile apps to promote health and manage disease: systematic review and meta-analysis of randomized controlled trials. JMIR Mhealth Uhealth 2021 Jan 11;9(1):e21563. [doi: 10.2196/21563] [Medline: 33427672]
17. Moore ZEH, Aynge GE, Carr CG, et al. A clinical support app for routine wound management: reducing practice variation, improving clinician confidence and increasing formulary compliance. Int Wound J 2022 Aug;19(5):1263-1275. [doi: 10.1111/iwj.13868] [Medline: 35793908]
18. Do Khac A, Jourdan C, Fazilleau S, et al. mHealth app for pressure ulcer wound assessment in patients with spinal cord injury: clinical validation study. JMIR Mhealth Uhealth 2021 Feb 23;9(2):e26443. [doi: 10.2196/26443] [Medline: 33620327]
19. Kitamura A, Nakagami G, Watanabe C, et al. Efficacy of teleconsultations with wound, ostomy, and continence nurses for patients with pressure ulcers in the community setting. J Jpn WOCM 2021;25(3):654-660. [doi: 10.32201/jpnwocm.25.3_654]
20. Kohta M, Sugiyama A, Hachimori A, Sugama J, Fukuro S. User acceptance of a mobile application prototyping model for supporting pressure injury treatment at home care. J Nurs Sci Eng 2023;11:28-36. [doi: 10.24462/jnse.11.0_28]
21. McKeown E, McGraw C, Holder P, Shand J, Hirani SP. Acceptability and impact of an educational app (iCare) for informal carers looking after people at risk of pressure ulceration: mixed methods pilot study. JMIR Form Res 2022 Sep 16;6(9):e36517. [doi: 10.2196/36517] [Medline: 36112413]
22. Cuddigan J, Haesler E, Moore Z, Carville K, Kottner J. Development, dissemination and evaluation of a smartphone-based app for pressure ulcer/injury prevention and treatment for use at the bedside. J Wound Care 2022 Dec 1;31(Sup12):S29-S39. [doi: 10.12968/jowc.2022.31.Sup12.S29] [Medline: 36475841]

XSL•FO
RenderX

23. Amann J, Fiordelli M, Scheel-Sailer A, Brach M, Rubinelli S. Opportunities and challenges of a self-management app to support people with spinal cord injury in the prevention of pressure injuries: qualitative study. JMIR Mhealth Uhealth 2020 Dec 9;8(12):e22452. [doi: 10.2196/22452] [Medline: 33295876]
24. Eldridge SM, Chan CL, Campbell MJ, et al. CONSORT 2010 statement: extension to randomised pilot and feasibility trials. Pilot Feasibility Stud 2016;2:64. [doi: 10.1186/s40814-016-0105-8] [Medline: 27965879]
25. Lancaster GA, Thabane L. Guidelines for reporting non-randomised pilot and feasibility studies. Pilot Feasibility Stud 2019;5:114. [doi: 10.1186/s40814-019-0499-1] [Medline: 31608150]
26. Japanese Ministry of Health, Labor and Welfare. URL: www.mhlw.go.jp/toukei/list/24-22-2.html [accessed 2024-08-19]
27. In J. Introduction of a pilot study. Korean J Anesthesiol 2017;70(6):601. [doi: 10.4097/kjae.2017.70.6.601]
28. Beeckman D, Vanderwee K, Demarré L, Paquay L, Van Hecke A, Defloor T. Pressure ulcer prevention: development and psychometric validation of a knowledge assessment instrument. Int J Nurs Stud 2010 Apr;47(4):399-410. [doi: 10.1016/j.ijnurstu.2009.08.010] [Medline: 19781701]
29. Delmore B, Ayello EA, Smart H, Sibbald RG. Assessing pressure injury knowledge using the Pieper-Zulkowski pressure ulcer knowledge test. Adv Skin Wound Care 2018 Sep;31(9):406-412. [doi: 10.1097/01.ASW.0000540071.45158.29] [Medline: 30134276]
30. Kohta M, Kameda Y, Morita S. Knowledge and practice for pressure injury prevention among care managers in a home care setting: a cross-sectional study. Chron Wound Care Manag Res 2017;4:99-105. [doi: 10.2147/cwcmr.s140673]
31. The Japanese Society of Pressure Ulcers. JSPU Guideline for Prevention and Management of Pressure Ulcers, 5th edition: Shorinsha Inc; 2022:1-112. [doi: 10.32249/9784796525534]
32. McKay FH, Slykerman S, Dunn M. The app behavior change scale: creation of a scale to assess the potential of apps to promote behavior change. JMIR Mhealth Uhealth 2019 Jan 25;7(1):e11130. [doi: 10.2196/11130] [Medline: 30681967]
33. Bricca A, Pellegrini A, Zangger G, Ahler J, Jäger M, Skou ST. The quality of health apps and their potential to promote behavior change in patients with a chronic condition or multimorbidity: systematic search in App Store and Google Play. JMIR Mhealth Uhealth 2022 Feb 4;10(2):e33168. [doi: 10.2196/33168] [Medline: 35119367]
34. Merolli M, Francis JJ, Vallance P, Bennell KL, Malliaras P, Hinman RS. Patient-facing mobile apps to support physiotherapy care: protocol for a systematic review of apps within app stores. JMIR Res Protoc 2021 Dec 9;10(12):e29047. [doi: 10.2196/29047] [Medline: 34889767]
35. Fulbrook P, Lawrence P, Miles S. Australian nurses' knowledge of pressure injury prevention and management: a cross-sectional survey. J Wound Ostomy Continence Nurs 2019;46(2):106-112. [doi: 10.1097/WON.0000000000000508] [Medline: 30801563]
36. Patel MS, Small DS, Harrison JD, et al. Effectiveness of behaviorally designed gamification interventions with social incentives for increasing physical activity among overweight and obese adults across the United States: the STEP UP randomized clinical trial. JAMA Intern Med 2019 Dec 1;179(12):1624-1632. [doi: 10.1001/jamainternmed.2019.3505] [Medline: 31498375]

Abbreviations

ABACUS: App Behavior Change Scale
CM: care manager
CONSORT: Consolidated Standards of Reporting Trials
PI: pressure injury
Pips-Map: pressure injury prevention support mobile app prototyping model

Edited by A Mavragani; submitted 26.02.24; peer-reviewed by Y Chen; revised version received 14.12.24; accepted 15.12.24; published 07.02.25.

Please cite as:
Kohta M, Takahashi M, Koyanagi H, Sugama J
Evaluating the Knowledge Level, Practice, and Behavioral Change Potential of Care Managers in Pressure Injury Prevention Using a Mobile App Prototyping Model in the Home-Care Setting: Single-Arm, Pre-Post Pilot Study
JMIR Form Res 2025;9:e57768
URL: https://formative.jmir.org/2025/1/e57768
doi:10.2196/57768

© Masushi Kohta, Mayumi Takahashi, Hiroe Koyanagi, Junko Sugama. Originally published in JMIR Formative Research (https://formative.jmir.org), 7.2.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction

XSL•FO
RenderX

in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Assessing Dynamic Cognitive Function in the Daily Lives of Youths With and Without Type 1 Diabetes: Usability Study

Mary Katherine Ray[1,2], PhD; Jorie Fleming[1], BA; Andrew Aschenbrenner[3], PhD; Jason Hassenstab[3], PhD; Brooke Redwine[1], BA; Carissa Burns[1], BA; Ana Maria Arbelaez[4], MD; Mary Ellen Vajravelu[5], MD; Tamara Hershey[1,2], PhD

[1]Department of Psychiatry, Washington University in St. Louis, 660 S. Euclid Avenue, St. Louis, MO, United States

[2]Department of Radiology, Washington University in St. Louis, St. Louis, MO, United States

[3]Department of Neurology, Washington University in St. Louis, St. Louis, MO, United States

[4]Department of Pediatrics, Washington University in St. Louis, St. Louis, MO, United States

[5]Division of Pediatric Endocrinology, Diabetes, and Metabolism, UPMC Children's Hospital of Pittsburgh, University of Pittsburgh School of Medicine, Pittsburgh, PA, United States

Corresponding Author:
Mary Katherine Ray, PhD
Department of Psychiatry, Washington University in St. Louis, 660 S. Euclid Avenue, St. Louis, MO, United States

Abstract

Background: Studies have shown a relationship between worse glycemic control and lower cognitive scores in youths with type 1 diabetes (T1D). However, most studies assess long-term glucose control (eg, years-decades) and cognition at a single time point. Understanding this relationship at a higher temporal resolution (eg, minutes-hours) and in naturalistic settings has potential clinical implications. Newer technology (eg, continuous glucose monitoring [CGM] and ecological momentary assessment) provides a unique opportunity to explore the glucose dynamics that influence dynamic cognition; that is, cognitive functions that fluctuate short-term and are influenced by environmental factors.

Objective: Before we can assess this relationship, we need to determine the feasibility of measuring cognition in youths in daily life and determine the plausibility of obtaining glucose variation with CGM to be integrated with real-time cognition measures. This study's purpose was to assess the acceptability of measuring dynamic cognition using a smartphone app and adherence to cognitive testing in daily life in youths with and without T1D. Further, we assessed CGM-derived glucose measures at temporally related timeframes to cognitive testing in naturalistic settings.

Methods: Data were obtained from 3 studies including one in-laboratory study and 2 remote studies. For all studies, youths were asked to complete cognitive tests on the Ambulatory Research in Cognition (ARC) smartphone app that measured processing speed, associative memory, and working memory. For the in-laboratory study, youths completed testing 4 times during 1 session. For the remote studies, youths were asked to complete cognitive tests 5 times per day for either 10 or 14 consecutive days in daily life. Youths were asked to rate their impressions of the app. Youths with T1D wore a CGM.

Results: 74 youths (n=53 control; n=21 T1D) aged 4 - 16 years participated. Youths generally reported liking or understanding the ARC app tasks in a laboratory and remote setting. Youths had high testing adherence in daily life (2350/3080 to 721/900, 76.3% - 80.2%) and none dropped out. The percentage of measurements within each glycemic range taken immediately before the app's cognitive testing was 3% (28/942) low glucose, 51% (484/942) euglycemia, 23% (221/942) high glucose, and 22% (210/942) very high glucose. In the 2-hour window before each cognitive task, mean glucose was 182.5 (SD 76.2) mg/dL, SD in glucose was 27.1 mg/dL (SD 18.7), and the mean maximum difference between the highest and lowest glucose was 85.5 (SD 53.7) mg/dL.

Conclusions: The results suggest that using the ARC smartphone app to assess dynamic cognitive functions in youths with and without T1D is feasible. Further, we showed CGM-derived glycemic variability at temporally associated timeframes of dynamic cognitive assessments. The next steps include using ecological momentary assessment in a fully powered study to determine the relationship between short-term glycemic control and cognition in youths with T1D.

(JMIR Form Res 2025;9:e60275) doi:10.2196/60275

KEYWORDS

ecological momentary assessment; EMA; ambulatory; smartphone; continuous glucose monitoring; CGM; assessment; daily lives; youth; type 1 diabetes; diabetes; feasibility study; pilot study; glycemic control; environmental factor; phone; acceptability; young; cognitive test; app; application; mobile phone

XSL•FO
RenderX

Introduction

Research has shown that youths with type 1 diabetes (T1D) often have slightly lower cognitive scores compared to their peers without T1D, and a relationship between worse glycemic control and lower cognitive scores [1-11]. These results highlight an important relationship between glucose and cognition in T1D. However, they are based on measures of long-term glucose control over long periods (eg, years-decades) and cognitive function tested in a laboratory setting at a single time point. Understanding this relationship at a higher temporal resolution (eg, minutes-hours) and in naturalistic settings has potential clinical implications.

Newer technology (eg, continuous glucose monitoring [CGM] and ecological momentary assessment [EMA]) provides a unique opportunity to explore the specific glucose patterns that influence dynamic cognition; that is, cognitive functions that fluctuate in the short-term and are easily influenced by environmental factors [12,13]. Using these approaches in adults with T1D, a few studies have shown an association between short-term glycemic control and cognitive functioning. Namely, studies have shown an association between significant glucose fluctuations and slower objective measures of processing speed at the moment [14], person-reported hypoglycemia and worse subjective measures of cognitive functioning later in the day [15], and increases in nocturnal hypoglycemia with slower processing speed the following day [16].

Whether this relationship or others are seen in youths with T1D is unknown. Understanding this relationship is significant given the importance of optimal cognitive functioning in academic settings in youths. However, before we can assess this relationship, we need to determine the feasibility and practicality of measuring dynamic cognition in youths in daily life. Although EMA is a feasible methodology for better understanding daily functioning in youths [17], including cognitive functioning [18], to our knowledge, no published study has assessed the feasibility of using a smartphone app to obtain EMAs of cognition in youths in the context of T1D.

This study aimed to test the feasibility of using a smartphone app called the Ambulatory Research in Cognition (ARC) app in youths. The ARC app, developed at Washington University in St. Louis, was originally designed to assess cognitive function in everyday environments in adults at risk of developing dementia [19]. Performance on ARC app cognitive tasks is sensitive to clinical status and genetic risk for Alzheimer disease [20,21] and can capture circadian fluctuations in cognition in adults at risk of Alzheimer disease [22]. The ability to capture variability in cognition in tandem with fluctuations in glucose in youths with T1D will be essential for determining the effects of glycemic variability on cognition in the daily lives of youths with T1D in future studies. Thus, in addition to testing the feasibility of the ARC app in youths, we also sought to determine the plausibility of obtaining glucose variation with CGM and integrating it with real-time measures of cognitive function using our protocol. This is important because EMA schedule selection has shown to be essential for capturing enough glycemic events throughout the day to integrate with cognitive data in adults with T1D [23].

Methods

Participants and Procedures

Recruitment

Study flyers were distributed throughout St. Louis Children's Hospital clinics, Midwest diabetes support groups, and the Washington University Volunteers for Health Research Registry program. The flyers included a short description of this study and the research team's contact information; interested parents were asked to contact the research team. St. Louis Children's Hospital health care workers also provided names of clinic patients who may be interested in this study to the research team. Participants were also recruited from word of mouth approaches with enrolled families telling friends about this study. After making contact with the families, the research team conducted a phone screen meeting to determine eligibility based on the inclusion and exclusion criteria outlined in Table 1. Data were obtained from 3 separate studies in youths with and without T1D. Study 1 was conducted in the laboratory setting, and studies 2 and 3 were conducted in remote settings. Of note, the study 2 inclusion criteria included owning an iPhone (6s or newer as this was the minimum required to support the ARC app) and established use of a Dexcom G5 or G6 CGM. For study 3, we obtained additional funding that allowed us to provide iPhones to youths who did not have their own and Dexcom G6 PRO CGMs to youths who did not already use a CGM. Given that we could provide CGMs to youths with T1D, we changed this study's protocol for study 3 from 14 days to 10 days to align with the life of a provided CGM (described in more detail below).

Figures and tables created with bioRender and REDCap (Research Electronic Data Capture) was used to support data collection [24-27].

XSL•FO
RenderX

Table . Inclusion and exclusion criteria for the pilot and feasibility studies that aimed at assessing the feasibility of using the Ambulatory Research in Cognition smartphone app to measure cognitive function in youths with and without type 1 diabetes in an in-laboratory setting (study 1), for 14 days in naturalistic settings (study 2), and for 10 days in naturalistic settings (study 3).

Study	Inclusion criteria	Exclusion criteria
Study 1: in laboratory (1 session)	• Age 4 - 16 years	• No English fluency • Unable to use a smartphone • Severe mental, neurological, or medical condition that would make them unable to complete tasks • Use of medications with known effects on the central nervous system (except ADHD[a] medicines)
Study 2: real world (14 days)	• Age 9 - 16 years • Owning an iPhone (6s or newer) • Established use of Dexcom G5/G6 continuous glucose monitoring	• No English fluency • Unable to use a smartphone • Severe mental, neurological, or medical condition that would make them unable to complete tasks • Use of medications with known effects on the central nervous system (except ADHD medicines) • No parent-approved screen time for participation • No tablet or computer with reliable internet connection • Severe phone screen cracks that would affect testing
Study 3: real world (10 days)	• Age 9 - 16 years	• No English fluency • Unable to use a smartphone • Severe mental, neurological, or medical condition that would make them unable to complete tasks • Use of medications with known effects on the central nervous system (except ADHD medicines) • No parent-approved screen time for participation • No tablet or computer with reliable internet connection • Severe phone screen cracks that would affect testing

[a]ADHD: attention-deficit/hyperactivity disorder.

Study 1

Participants completed a tutorial of all the app tasks and were able to ask the research team any clarifying questions. They completed app testing 4 times with 5-minute breaks and then completed an experiential interview to assess their impressions of the app in the laboratory setting.

Study 2

Participants had a video call with researchers to practice with the ARC app tasks and ask the research team any clarifying questions. The research team established CGM data sharing privileges through the Dexcom Clarity app in youths with T1D. The following day, all youths were asked to complete cognitive testing in daily life 5 times per day for 14 consecutive days. They had a video call with the research team 1 and 2 weeks later to complete an experiential interview.

Study 3

Participants had a video call or visited the laboratory to practice the app tasks and establish CGM data sharing privileges (T1D group). The following day, all youths were asked to cognitive complete testing in daily life 5 times per day for 10 consecutive days. For both remote studies, monetary incentives were provided to encourage testing compliance as outlined in the Ethical Considerations section below.

Ethical Considerations

The Washington University Human Research Protection Office (IRB00009237) approved all study procedures. Parents provided consent for them and their child to participate and youths provided assent. For privacy and confidentiality protection, all study data were deidentified. Compensation was provided to participants. For study 1, youths were compensated US $15 per hour (total of US $30), for their participation, paid via check. For study 2, youths could earn up to US $125 in Amazon gift cards for their participation (US $20 for participation in video

XSL•FO
RenderX

calls; US $70 for being in this study for 2 weeks—US $5 per day when they completed at least 2 of the 5 cognitive task sessions; US $35 for completing tests—US $0.50 per test). For study 3, youths could earn up to US $115 in Amazon gift cards for their participation (US $25 for an introductory session via Zoom [Zoom Communications, Qumu Corporation], US $40 for an introductory session at Washington University; US $50 for being in this study for 10 days—US $5 per day if they completed at least 3 of the 5 cognitive tasks; US $25 for completing tests—US $0.50 per test). For study 2 and 3, youths who were aged ≥13 years were also given a Fitbit to track sleep and activity (data not shown) and kept the tracker as compensation.

Measures

ARC App

The ARC app is comprised of 3 tasks. The symbols task assesses processing speed by asking participants to compare visual stimuli. Participants are shown 3 cards with 2 images at the top of a screen and 2 cards with 2 images at the bottom of a screen and asked to determine which card on the bottom matches a card on the top as quickly as possible (Figure 1A). There are 12 trials. The performance score is median reaction time, calculated for trials with correct responses and reaction times above 150 ms to avoid anticipatory responses. The prices task assesses associative memory. Participants are given a household item and an associated price. They are shown 10 pairs and asked to recall the price (Figure 1B). The performance score is percent error (0%=all correct, 100%=all incorrect, and 50%=chance performance). The grids task assesses working memory. Participants are shown a grid with 3 items and asked to remember their location (Figure 1C). They are presented with a distraction task where they are shown a screen of “E”s with sporadic “F”s and asked to select the “F”s. Then they are presented with a blank grid and asked to choose where the items were located. Participants are presented with this task twice per session. The performance score is the average Euclidean distance from the selected response to the correct placement (score range: 0 - 5.65; 0=perfect placement of all items and trials). There was a “tutorial” button on each task interface throughout this study so youths could reorient themselves to the tasks before completing the actual tasks being recorded for this study if needed.

XSL•FO
RenderX

Figure 1. The Ambulatory Research in Cognition app being tested in this pilot and feasibility study. (**A**) The symbols task assesses processing speed. (**B**) The prices task assesses associative memory. (**C**) The grids task assesses working memory.

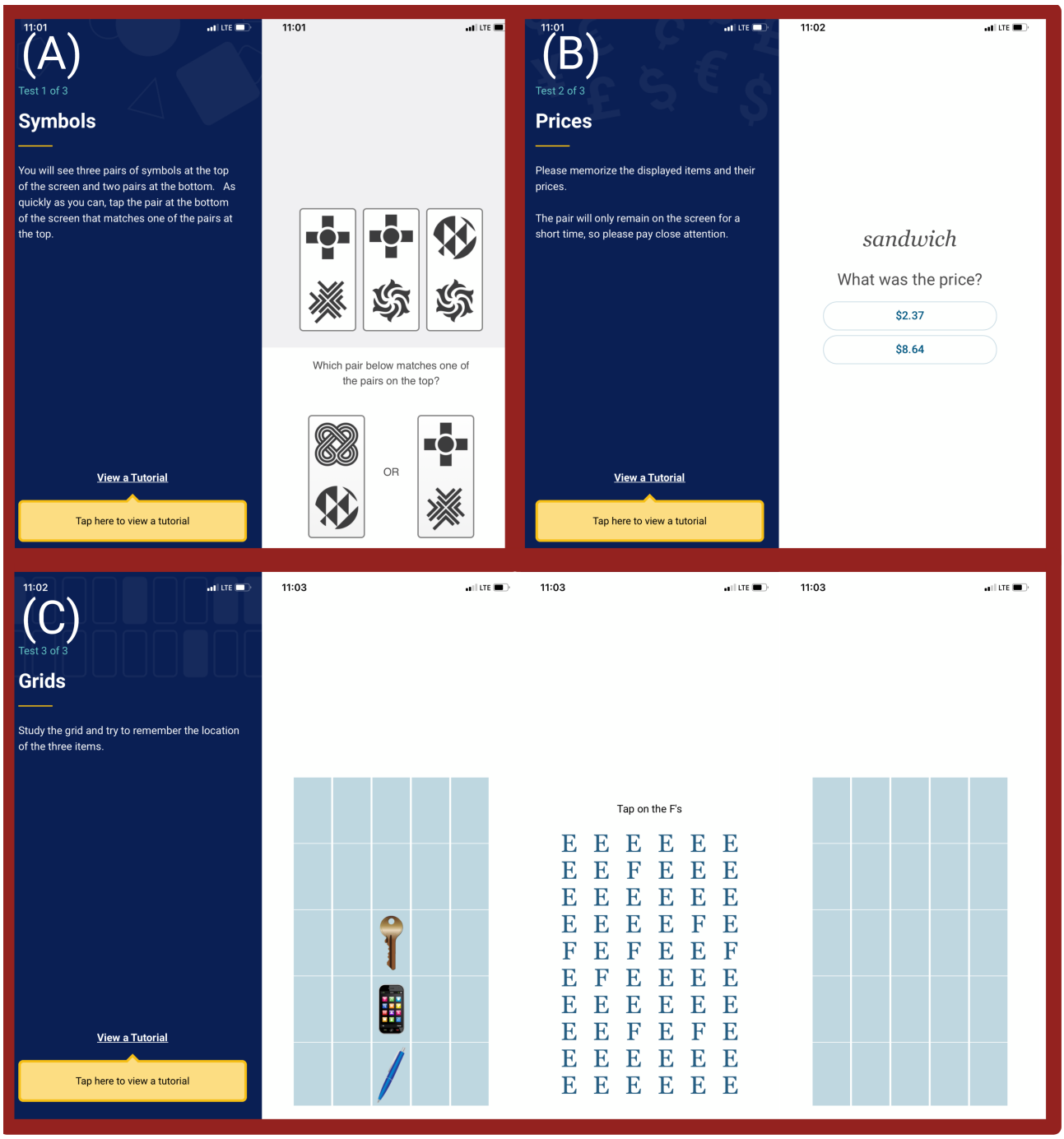

Each session, including all 3 tasks, only takes about 3 minutes to complete. For both remote studies, youths received notifications from the app to take the tests randomly and had 2 hours to complete them. This sampling method was chosen to capture variability in cognitive function and glycemic control throughout the entire day when youths were in different contexts in daily life. Youths and their parents were able to choose the timeframes when they wanted to receive notifications and could change these times through the app to reduce the burden and reduce the likelihood of receiving a notification during times that would affect daily functioning (eg, overnight while sleeping, sleeping in on the weekend, or in class at school). Parents were not explicitly told to encourage youths to complete the sessions because the goal of using EMA methodology in this context is for youths to be able to complete testing unsupervised. Relatedly, youths were told to answer questions on the ARC app tests without help from others. All participants completed the tasks using an iPhone (6s or newer). Compliance to app testing was automatically logged electronically through an ARC app dashboard. Although the app was developed for an adult population [19], the app was used as is in our study of youths; if deemed acceptable for youths in its original state, it would be a cost-effective method that could be easily implemented into study protocols.

Experiential Interview

For each task in study 1 and study 2, youths were asked to rate on a scale of 1 - 5 how challenging or easy the task was, how

XSL•FO
RenderX

confusing or clear the instructions were, how confusing or clear the task was, and how boring or fun the task was. The Likert scale had an anchor of 3 to designate a neutral appraisal for each question. For example, “How confusing were the instructions for the symbols game?” Option 3 was “not confusing or clear.”

About CGM

For studies 2 and 3, participants with T1D wore a Dexcom G5/G6 CGM to measure glucose every 5 minutes. Youths who already used a Dexcom CGM changed their devices as they normally would. For study 3, if youths with T1D did not already use a Dexcom CGM, they were provided with a Dexcom G6 PRO in blinded mode to wear for the duration of this study.

Statistical Analysis

Data were analyzed with R (R Foundation), Python (Python Software Foundation), and SPSS (IBM Corp). ARC app performance: Mean and SD were calculated for performance on each ARC app task. For study 1, Pearson correlations also assessed the relationship between age and ARC task performance (Kendall τ was used for nonnormal data), and scatterplots were created to visually appraise the age at which youths started performing better on the ARC app tasks. Further, 2-tailed *t* tests determined differences in performance between age groups. Experiential interview ratings: mean and SD were calculated for each question assessing impressions of the app tasks; a mean score greater than 3 was determined to be a positive impression of the task. ARC app adherence: Adherence was calculated as the number of tests completed divided by the number of tests offered.

CGM: Glucose variable descriptives were calculated from the raw Dexcom Clarity exports for all participants with T1D across the duration of this study, including mean, SD, percent time in range (TIR; 70 - 180 mg/dL), percent time below range (<70 mg/dL), percent time high (≥180 mg/dL), and percent time very high (≥250 mg/dL). The glucose measure taken immediately before each cognitive task for each individual with T1D was extracted. Glucose obtained from the CGM were matched to the corresponding cognitive test by timestamps within each system, both set to CST. Given that glucose was assessed every 5 minutes, the closest glucose measure from CGM was taken within 5 minutes of each cognitive session. Lastly, the following glycemic variables were extracted for the 2-hour window immediately before each cognitive task for each individual with T1D: mean glucose, SD of glucose, and maximum difference in glucose (ie, highest glucose within a 2 h block minus lowest glucose within the 2 h block). Mean sensor usage during participation was calculated; Dexcom Clarity defines sensor usage as the number of days during the reporting period with at least 50% CGM readings [28]. Deletion by list was used for missing CGM data.

Results

Participants

Details for eligibility and consent rate are outlined in Figure 2.

XSL•FO
RenderX

Figure 2. Eligibility flowchart for the pilot and feasibility studies aimed at assessing the feasibility of using the Ambulatory Research in Cognition smartphone app to assess cognitive function in the daily lives of youths with and without T1D. Study 1 was conducted in a single session in the laboratory setting in youths without T1D. Studies 2 and 3 were conducted in naturalistic settings for 14 and 10 days, respectively, in youths with and without T1D. T1D: type 1 diabetes.

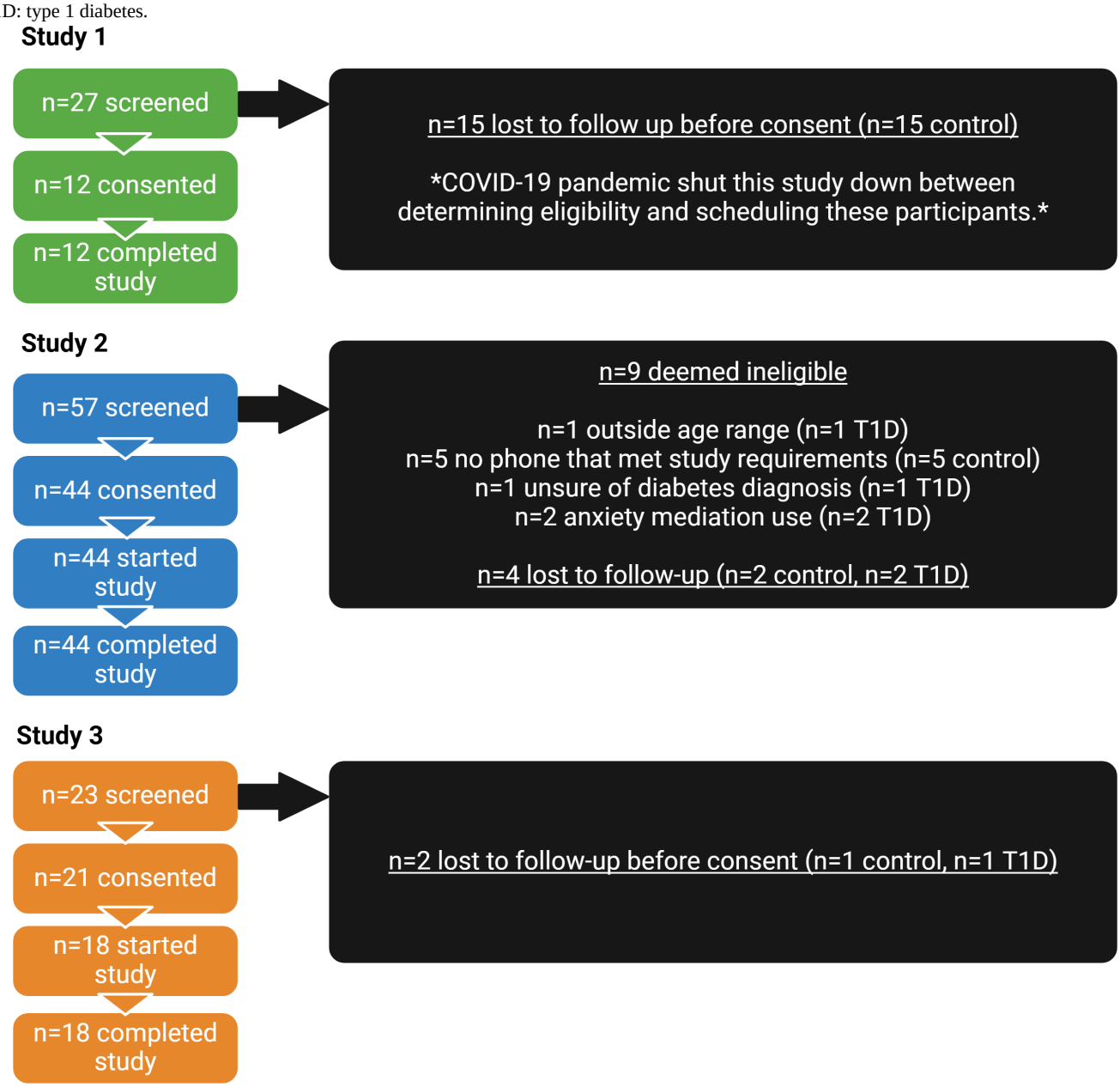

Ultimately, a total of 74 participants aged 4‑16 years completed study procedures (study 1: n=12 control; study 2: n=44; n=30 control, n=14 T1D; study 3: n=18; n=11 control, n=7 T1D). No participants were voluntarily or involuntarily withdrawn from the studies. Demographic data are shown in Table 2. This study's samples were relatively homogenous, with most participants being non-Hispanic White. For study 1, there was a glitch with the ARC app for 1 participant; thus, this participant did not have cognitive function data. For study 3, a total of 8 participants were provided with an iPhone, and all iPhones were returned to this study's team after participation. The total annual household income for participants enrolling in this study with their own iPhone was US $131,111, and the annual household income for participants who were provided with an iPhone was US $68,750.

XSL•FO
RenderX

Table . Demographic data for the pilot and feasibility studies that aimed to assess the feasibility of using the Ambulatory Research in Cognition smartphone app to measure cognitive function in youths with and without type 1 diabetes in an in-laboratory setting (study 1), for 14 days in naturalistic settings (study 2), and for 10 days in naturalistic settings (study 3).

			Group 1	Group 2
Study 1, n			12[a]	7[b]
	Age (years), mean (SD)		10.6 (3.6)	13.1 (2.3)
	Sex, n			
		Female	7	5
		Male	5	2
	Race or ethnicity, n (%)			
		Black	3 (25)	2 (29)
		Hispanic White	1 (8)	1 (14)
		Non-Hispanic White	8 (67)	4 (57)
Study 2, n			30[c]	14[d]
	Age (years), mean (SD)		13.0 (1.7)	13.4 (2)
	Sex, n			
		Female	17	7
		Male	13	7
	Race or ethnicity, n (%)			
		Black	2 (7)	0 (0)
		Hispanic	1 (3)	0 (0)
		Non-Hispanic White	25 (83)	14 (100)
		Mixed race	2 (7)	0 (0)
	Annual household income (US $), mean (SD)		138,907 (71,860)[e]	125,231 (44,129)[f]
Study 3, n			11[c]	7[d]
	Age (years), mean (SD)		13 (2.3)	12.4 (2.2)
	Sex, n			
		Female	5	1
		Male	6	6
	Race or ethnicity, n (%)			
		Black	0 (0)	0 (0)
		Hispanic	0 (0)	0 (0)
		White	11 (100)	7 (100)
		Mixed race	0 (0)	0 (0)
	Annual household income (US $), mean (SD)		91,818 (37,031)	120,000 (47,010)[g]

[a]Overall
[b]≥9 years old
[c]Control.
[d]Type 1 diabetes.
[e]n=27
[f]n=13
[g]n=6

Although the protocol for study 3 offered CGMs to use for the duration of this study for youths who did not use them, all participants enrolled already used a Dexcom CGM and used their own for this study. For study 3, one participant with T1D and his family could not sync their CGM data to Dexcom Clarity. Mean sensor usage was 97%. Average glycemic

XSL•FO
RenderX

measures for youths with T1D for the entire study duration are shown in Table 3. On average, youths with T1D in our sample did not meet the recommended guidelines for percent TIR or percent time in hyperglycemia (59% vs recommended 70% TIR; 39% vs recommended 25% time in hyperglycemia) [29].

Table . Glycemic measures for youths with type 1 diabetes as measured by continuous glucose monitoring across studies 2 and 3 when youths were completing cognitive tasks on the Ambulatory Research in Cognition app in naturalistic settings for 14 and 10 days, respectively. Data are presented as mean (SD) across all days of study participation (n=20). We were unable to collect continuous glucose monitoring data in 1 participant with T1D.

	Values, mean (SD)
Mean glucose (mg/dL)	176 (46)
SD glucose (mg/dL)	63.9 (17.7)
Time in range (%, 70 - 180 mg/dL)	58.7 (20.6)
Time below range (%, <70 mg/dL)	2.5 (3.3)
Time high (%, ≥180 mg/dL)	38.8 (22.3)
Time very high (%, ≥250 mg/dL)	17.2 (19.1)

Feasibility and Acceptability of Using a Smartphone App to Assess Dynamic Cognitive Function in Youths

Study 1: Youths reported liking and understanding the grids and symbols tasks but not the prices task (Table 4).

XSL•FO
RenderX

Table . Youths' impression of and performance on the Ambulatory Research in Cognition smartphone app cognitive tasks being tested in these pilot and feasibility studies. Study 1 was conducted in the laboratory setting in youths without type 1 diabetes. Studies 2 and 3 were conducted in naturalistic settings in youths with and without type 1 diabetes for 14 or 10 consecutive days, respectively. There was an app malfunction with 1 participant from study 1, so there are only data for 11 participants.

	Symbols (processing speed), median reaction time (seconds)		Prices (associative memory), error score (%)		Grids (working memory), Euclidean distance	
	Mean (SD)	Range	Mean (SD)	Range	Mean (SD)	Range
Performance (lower score=better)						
Study 1 (n=11)	2.4 (0.6)	1.7-4	28.6 (14.5)	6-46	0.7 (0.4)	0.1-1.7
Study 2 (n=44)	1.8 (0.3)	1-2.7	40.6 (7.7)	10.3-50.8	0.5 (0.2)	0.1-1.3
Study 3 (n=18)	1.9 (0.5)	1.3-2.8	40.1 (6.7)	27.1-50.8	0.5 (0.2)	0.2-0.9
Impressions						
How hard or easy was the task? (1=very hard and 5=very easy)						
Study 1 (n=12)	4.2 (0.7)	3 - 5	2.3 (0.8)	1 - 4	3.3 (1.4)	1 - 5
Study 2 week 1 (n=42)	4.3 (0.8)	2 - 5	2.2 (1.1)	1 - 4	4 (1)	2 - 5
Study 2 week 2 (n=43)	4.5 (0.8)	2 - 5	2.6 (1.2)	1 - 5	4.1 (0.9)	2 - 5
How confusing or clear were the task instructions? (1=very confusing and 5=very clear)						
Study 1 (n=12)	4.4 (1.1)	2 - 5	4.4 (1)	2 - 5	4.5 (0.9)	3 - 5
Study 2 week 1 (n=42)	4.9 (0.3)	4 - 5	4.8 (0.4)	4 - 5	4.9 (0.4)	3 - 5
Study 2 week 2 (n=43)	4.9 (0.4)	3 - 5	4.9 (0.4)	3 - 5	4.9 (0.4)	3 - 5
How confusing or clear was the task itself? (1=very confusing and 5=very clear)						
Study 1 (n=12)	4.7 (0.7)	3 - 5	3.8 (1.4)	2 - 5	4.3 (1.4)	1 - 5
Study 2 week 1 (n=42)	4.7 (0.7)	2 - 5	4.3 (1.1)	2 - 5	4.7 (0.7)	2 - 5
Study 2 week 2 (n=43)	4.8 (0.4)	4 - 5	4.4 (0.9)	2 - 5	4.7 (0.6)	2 - 5
How boring or fun was the task? (1=very boring and 5=very fun)						
Study 1 (n=12)	3.6 (1)	2 - 5	3.8 (0.8)	3 - 5	4.5 (0.7)	3 - 5
Study 2 week 1 (n=42)	3.8 (0.9)	2 - 5	2.8 (1.1)	1 - 5	4.1 (0.8)	2 - 5
Study 2 week 2 (n=43)	3.8 (1)	1 - 5	2.7 (1)	1 - 5	4 (0.6)	3 - 5

Performance on the tasks correlated with age (grids: $R=-0.77$, $P=.006$, n=11; prices: $R=-0.77$, $P=.005$, n=11; symbols: $\tau=-0.42$, $P=.07$, n=11), such that younger youths performed worse than older youths. Visually, scatterplots showed that around the age of 9 years, youths started performing better (Figure 3).

XSL•FO
RenderX

Figure 3. Scatterplots representing performance on each of the Ambulatory Research in Cognition smartphone app tasks and age during study 1 that aimed to test the Ambulatory Research in Cognition app in youths without type 1 diabetes at a single time point in the laboratory setting. Each dot represents an individual participant. Lower scores on each task represent better cognitive performance. The (**A**) symbols task assesses processing speed, (**B**) prices task assesses associative memory, and (**C**) grids task assesses working memory.

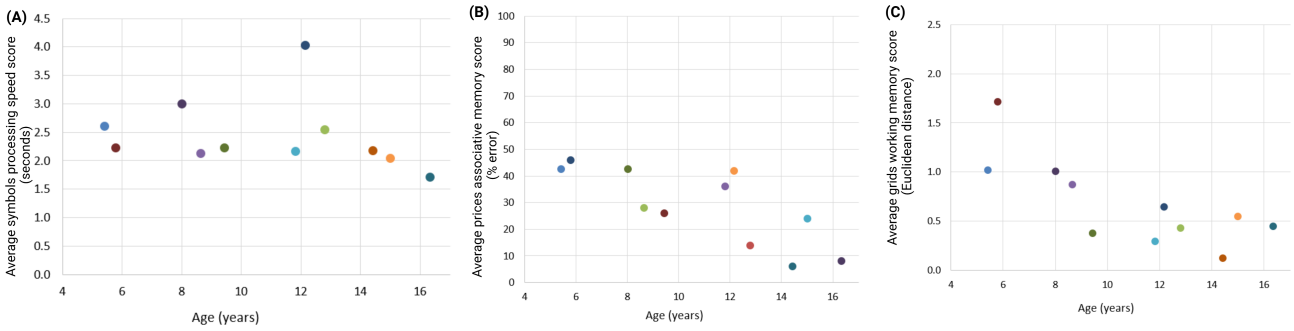

Based on these visual representations, we assessed performance in youths aged between <9 years and ≥9 years. There were statistically significant differences in performance such that youths aged <9 years had worse performance than youths aged ≥9 years on the grids task (<9 years: mean 1.2, SE 0.19; ≥9 years: mean 0.41, SE 0.10; *P*=.001) and prices task (<9 years: mean 39.8, SE 4.00; ≥9 years: mean 22.3, SE 5.19; *P*=.047). These data guided our decision to include ages 9 - 16 years in study 2 and study 3. For study 2, youths completed 76.3% (SD 19) of offered tasks, and for study 3, youths completed 80.2% (SD 13) of offered tasks. Across both studies, youths reported liking and understanding the grids and symbols tasks but not the prices task (Table 4). Figure 4 shows histograms for the distribution of task rankings for the control and T1D groups.

XSL•FO
RenderX

Figure 4. Histograms displaying ratings of the Ambulatory Research in Cognition app symbols, prices, and grids tasks from the experiential interview in study 2 which was conducted in naturalistic settings for 14 consecutive days in youths with and without type 1 diabetes. Ratings for the symbols task, which assesses processing speed, is shown in blue. Ratings for the prices task, which assesses associative memory is shown in green. Ratings for the grids task, which assesses working memory, is shown in red.

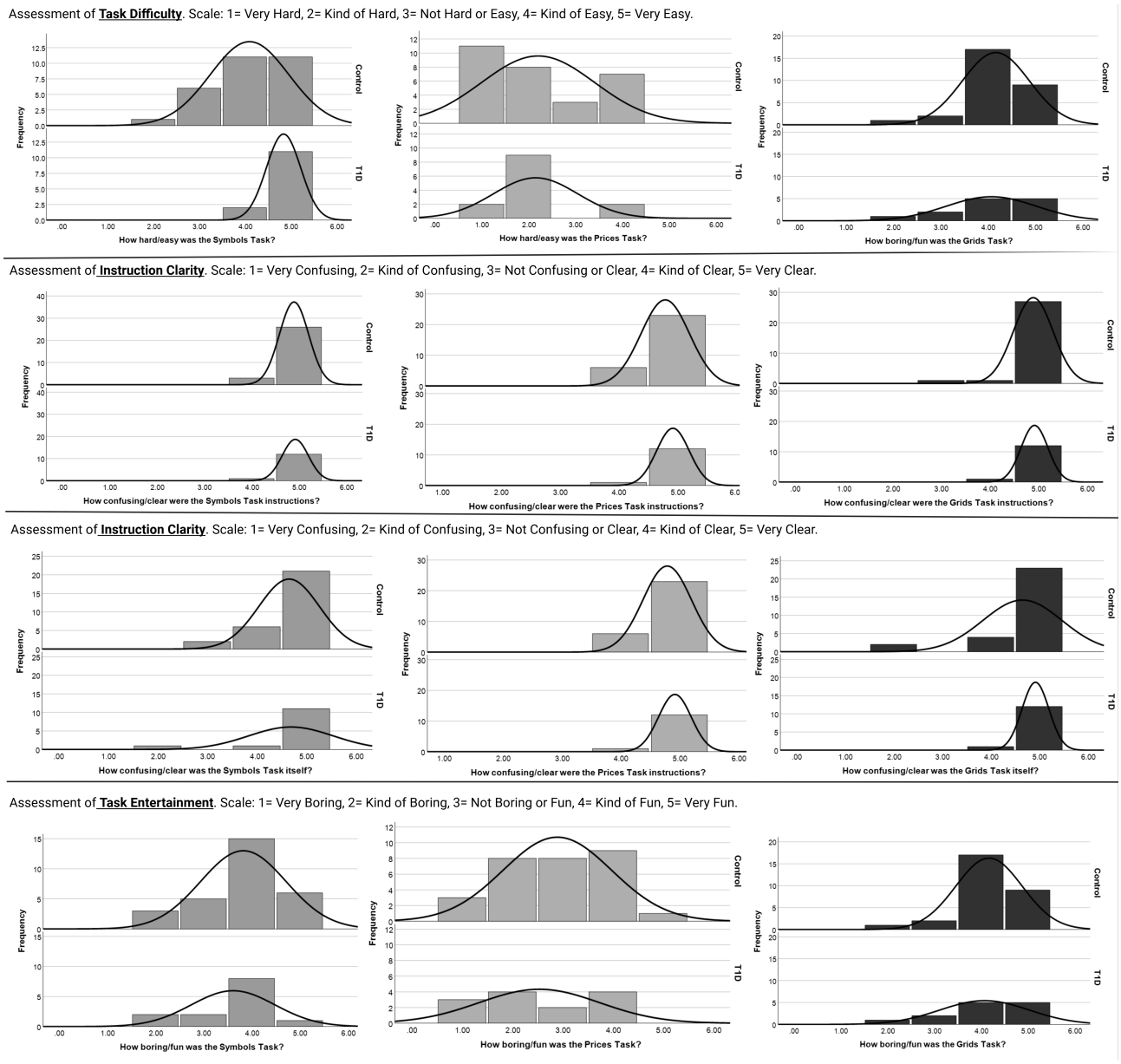

Glycemic Variability and Cognitive Function Data Integration

Across all participants with T1D from studies 2 and 3, the percentage of measurements within each glycemic range for the glucose measure taken immediately before cognitive testing was as follows: 3% (28/942) low glucose, 51% (484/942) euglycemia, 23% (221/942) high glucose, and 22% (210/942) very high glucose. In the 2-hour window immediately before each cognitive task, mean glucose was 182.5 (SD 76.2) mg/dL, SD in glucose was 27.1 (SD 18.7) mg/dL, and the mean maximum difference between the highest and lowest glucose was 85.5 (SD 53.7) mg/dL.

Discussion

Principal Findings

This study aimed to assess the feasibility of using a smartphone app to measure dynamic cognitive function in youths in daily life and to integrate this information with data obtained from CGM in youths with T1D. Our results suggest that using the ARC smartphone app to assess dynamic cognitive functions in youths with and without T1D is feasible. Further, we showed glycemic variability from CGM measures at temporally associated timeframes of dynamic cognitive function measures using our EMA schedule.

XSL•FO
RenderX

Feasibility and Acceptability of Using a Smartphone App to Assess Dynamic Cognitive Function in Youths

Testing of the app in an in-laboratory setting revealed that youths aged younger than 9 years performed worse than youths aged ≥9 years. Differences in performance on the ARC app in the laboratory setting guided our decision to include youths aged ≥9 years in the remote testing phases of the research. Although we expected older children to perform better than younger youths simply because of normal cognitive development progression, we wanted to be cautious of including youths in remote settings who may not understand the tasks enough to be able to complete them independently. Thus, for remote testing, we chose to focus on older children whose performance was uniformly better in our in-laboratory data. Future research is needed to determine if the ARC app could be used in younger youths in remote settings, and if not, what stimuli would better engage and test cognition in younger children.

Testing the app in naturalistic settings in youths, we found that youths had high testing adherence in daily life (76% - 80%). Although we expected lower adherence in youths given the amount of time spent at school and participating in extracurricular activities, compliance in our study was similar to that of adults from a previous study using the ARC app (~80%) [19]. Further, none of all the 62 youths who started with the remote study procedures dropped out. Furthermore, youths had high appraisal for 2/3 cognitive tasks, including the grids working memory task and the symbols processing speed task. It is also important to note that providing iPhones to youths who did not have their own for participation (study 3) helped expand our reach to include youths from lower income communities compared to this study's protocol when youths were required to have their own iPhones (study 2). Of note, all iPhones provided to participants for this study were returned to the research team which shows the logistical feasibility of this provision. Given that worse T1D outcomes have been associated with lower income [30-32], reaching youths from lower SES communities in future studies will be crucial to understand the relationship between short-term glycemic control and dynamic cognitive function. Taken together, our data support the use of the ARC smartphone app to measure dynamic cognition in youths aged 9 - 16 years, with and without T1D in daily life, and highlight the logistical feasibility of providing devices to youths who do not have their own.

When integrating glucose data and cognitive function data, we found that there was variation in glucose at the time of cognitive testing and during a short timeframe immediately before cognitive testing (eg, 2 h before). Specifically, youths with T1D had dysglycemia at the time of almost half of cognitive tests (49%) and on average, experienced 86 mg/dL swings in glucose in the 2 hours immediately before cognitive testing. The ability to capture glucose at various ranges (eg, low glucose or very high glucose) and variability in glucose (eg, amount of change in glucose during short periods before cognitive testing) using this methodology will be necessary for future studies aimed at determining the relationship between short-term glycemic control and dynamic cognitive function in youths with T1D.

Of note, only 3% of cognitive tasks were completed when glucose was in the hypoglycemia range in youths with T1D. Although this small percentage does not provide a significant amount of data for assessing the effects of low glucose on cognitive function, many other glycemic variables may be important for cognitive functioning that were captured frequently in our sample, including very high glucose at the time of testing and amount of change in glucose before cognitive testing. For example, using single measures of self-monitored blood glucose in the home, Gonder-Frederick et al [33] found that severe hyperglycemia (>400 mg/dl) was associated with equally substantial deteriorations in cognition in youths with T1D than severe hypoglycemia [33]. This study highlights the need to collect measures of hyperglycemia in addition to hypoglycemia in this population. Further, the use of CGM may allow future studies to capture glycemic patterns that may predict cognitive functioning outside of standard hypoglycemia and hyperglycemia cutoffs (eg, swings in glucose).

There are study limitations to consider. First, our sample was relatively homogenous. Although we were able to reach youths from lower SES communities in study 3 when we provided iPhones for this study, our sample with T1D lacked racial and ethnic diversity. Practices are needed to better engage youths from all communities in future work. Additionally, there is a documented lag between interstitial glucose obtained from CGM and blood glucose in past literature [34,35]. Given that lag time is patient-specific [36], we chose to use the glucose measurement obtained closest to each individual cognitive test to illustrate the integration of glucose measured via CGM and cognition measured via the smartphone app; future studies assessing the relationship between short-term glucose control and dynamic cognitive function need to consider multiple glycemic features that account for lag. In the same vein, we did not ask youths to report the cause for CGM sensor errors during their participation. Thus, if youths had missing CGM data, it was unclear why (eg, intrinsic sensor failure or intentional removal of the sensor). Future EMA studies should ask participants to report the cause of missing glucose data to help account for missingness when assessing the relationship between dynamic cognitive function and short-term glucose control.

In conclusion, we found that it is feasible to obtain measures of dynamic cognitive function in youths in daily life using a smartphone app and that we could capture glucose variability for integration with measures of glucose in the daily lives of youths with T1D. Using this methodology in fully powered studies assessing the relationship between short-term glycemic control in real-time cognition would move the field toward a fuller understanding of the impacts of T1D on cognitive function in youths.

XSL•FO
RenderX

Acknowledgments

This work was made possible with support from WU-CDTR (Washington University in St. Louis-Center for Diabetes Translation Research; P30DK092950), Washington University in St. Louis DRC (Diabetes Research Center; P30DK020579), and Washington University in St. Louis and BJC HealthCare Innovation Lab Big Ideas grant. MKR was supported by National Center for Advancing Translational Sciences of the National Institutes of Health Career Development Award (KL2TR002346), National Institute of Diabetes and Digestive and Kidney Diseases Career Development Award (NIDDK; 1K01DK131339), Washington University in St. Louis Biomedical Research Training in Drug Abuse training grant (T32DA007261), and Washington University in St. Louis Transdisciplinary Postdoctoral Training Program in Obesity and Cardiovascular Disease (T32HL130357). AA was supported by National Institute of Aging (NIA; K01AG071847). MEV was supported by National Institute of Diabetes and Digestive and Kidney Diseases (K23DK125719). The content is solely the responsibility of the authors and does not necessarily represent the official views of the WU-CDTR, DRC, NIDDK, NIA, or Healthcare Innovation Lab.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

References

1. Shalimova A, Graff B, Gąsecki D, et al. Cognitive dysfunction in type 1 diabetes mellitus. J Clin Endocrinol Metab 2019 Jun 1;104(6):2239-2249. [doi: 10.1210/jc.2018-01315] [Medline: 30657922]
2. Cato A, Hershey T. Cognition and type 1 diabetes in children and adolescents. Diabetes Spectr 2016 Nov;29(4):197-202. [doi: 10.2337/ds16-0036] [Medline: 27899870]
3. Cato MA, Mauras N, Mazaika P, et al. Longitudinal evaluation of cognitive functioning in young children with type 1 diabetes over 18 months. J Int Neuropsychol Soc 2016 Mar;22(3):293-302. [doi: 10.1017/S1355617715001289] [Medline: 26786245]
4. Semenkovich K, Bischoff A, Doty T, et al. Clinical presentation and memory function in youth with type 1 diabetes. Pediatr Diabetes 2016 Nov;17(7):492-499. [doi: 10.1111/pedi.12314] [Medline: 26377697]
5. Blasetti A, Chiuri RM, Tocco AM, et al. The effect of recurrent severe hypoglycemia on cognitive performance in children with type 1 diabetes: a meta-analysis. J Child Neurol 2011 Nov;26(11):1383-1391. [doi: 10.1177/0883073811406730] [Medline: 21572053]
6. Ghetti S, Lee JK, Sims CE, Demaster DM, Glaser NS. Diabetic ketoacidosis and memory dysfunction in children with type 1 diabetes. J Pediatr 2010 Jan;156(1):109-114. [doi: 10.1016/j.jpeds.2009.07.054] [Medline: 19833353]
7. Kirchhoff BA, Jundt DK, Doty T, Hershey T. A longitudinal investigation of cognitive function in children and adolescents with type 1 diabetes mellitus. Pediatr Diabetes 2017 Sep;18(6):443-449. [doi: 10.1111/pedi.12414] [Medline: 27444539]
8. Semenkovich K, Patel PP, Pollock AB, et al. Academic abilities and glycaemic control in children and young people with type 1 diabetes mellitus. Diabet Med 2016 May;33(5):668-673. [doi: 10.1111/dme.12854] [Medline: 26173465]
9. He J, Ryder AG, Li S, Liu W, Zhu X. Glycemic extremes are related to cognitive dysfunction in children with type 1 diabetes: a meta-analysis. J Diabetes Investig 2018 Nov;9(6):1342-1353. [doi: 10.1111/jdi.12840] [Medline: 29573221]
10. Gaudieri PA, Chen R, Greer TF, Holmes CS. Cognitive function in children with type 1 diabetes: a meta-analysis. Diabetes Care 2008 Sep;31(9):1892-1897. [doi: 10.2337/dc07-2132] [Medline: 18753668]
11. Naguib JM, Kulinskaya E, Lomax CL, Garralda ME. Neuro-cognitive performance in children with type 1 diabetes--a meta-analysis. J Pediatr Psychol 2009 Apr;34(3):271-282. [doi: 10.1093/jpepsy/jsn074] [Medline: 18635605]
12. Gamaldo AA, Allaire JC. Daily fluctuations in everyday cognition: is it meaningful? J Aging Health 2016 Aug;28(5):834-849. [doi: 10.1177/0898264315611669] [Medline: 26538267]
13. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol 2008;4:1-32. [doi: 10.1146/annurev.clinpsy.3.022806.091415] [Medline: 18509902]
14. Hawks ZW, Beck ED, Jung L, et al. Dynamic associations between glucose and ecological momentary cognition in type 1 diabetes. NPJ Digit Med 2024 Mar 18;7(1):59. [doi: 10.1038/s41746-024-01036-5] [Medline: 38499605]
15. Søholm U, Broadley M, Zaremba N, et al. The impact of hypoglycaemia on daily functioning among adults with diabetes: a prospective observational study using the Hypo-METRICS app. Diabetologia 2024 Oct;67(10):2160-2174. [doi: 10.1007/s00125-024-06233-1] [Medline: 39080044]
16. Zuniga-Kennedy M, Wang OH, Fonseca LM, et al. Nocturnal hypoglycemia is associated with next day cognitive performance in adults with type 1 diabetes: pilot data from the GluCog study. Clin Neuropsychol 2024 Oct;38(7):1627-1646. [doi: 10.1080/13854046.2024.2315749] [Medline: 38380810]
17. Russell MA, Gajos JM. Annual research review: ecological momentary assessment studies in child psychology and psychiatry. J Child Psychol Psychiatry 2020 Mar;61(3):376-394. [doi: 10.1111/jcpp.13204] [Medline: 31997358]

XSL•FO
RenderX

18. Könen T, Dirk J, Schmiedek F. Cognitive benefits of last night's sleep: daily variations in children's sleep behavior are related to working memory fluctuations. J Child Psychol Psychiatry 2015 Feb;56(2):171-182. [doi: 10.1111/jcpp.12296] [Medline: 25052368]
19. Nicosia J, Aschenbrenner AJ, Balota DA, et al. Unsupervised high-frequency smartphone-based cognitive assessments are reliable, valid, and feasible in older adults at risk for Alzheimer's disease. J Int Neuropsychol Soc 2023 Jun;29(5):459-471. [doi: 10.1017/S135561772200042X] [Medline: 36062528]
20. Welhaf MS, Wilks H, Aschenbrenner AJ, et al. Naturalistic assessment of reaction time variability in older adults at risk for Alzheimer's disease. J Int Neuropsychol Soc 2024 Jun;30(5):428-438. [doi: 10.1017/S1355617723011475] [Medline: 38282413]
21. Aschenbrenner AJ, Hassenstab J, Morris JC, Cruchaga C, Jackson JJ. Relationships between hourly cognitive variability and risk of Alzheimer's disease revealed with mixed-effects location scale models. Neuropsychology 2024 Jan;38(1):69-80. [doi: 10.1037/neu0000905] [Medline: 37079810]
22. Wilks H, Aschenbrenner AJ, Gordon BA, et al. Sharper in the morning: cognitive time of day effects revealed with high-frequency smartphone testing. J Clin Exp Neuropsychol 2021 Oct;43(8):825-837. [doi: 10.1080/13803395.2021.2009447] [Medline: 35037593]
23. Mascarenhas Fonseca L, Strong RW, Singh S, et al. Glycemic Variability and Fluctuations in Cognitive Status in Adults With Type 1 Diabetes (GluCog): observational study using ecological momentary assessment of cognition. JMIR Diabetes 2023 Jan 5;8:e39750. [doi: 10.2196/39750] [Medline: 36602848]
24. Create professional science figures in minutes. bioRender. URL: https://www.biorender.com/ [accessed 2025-01-23]
25. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)-a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381. [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
26. Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: building an international community of software platform partners. J Biomed Inform 2019 Jul;95:103208. [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]
27. Lawrence CE, Dunkel L, McEver M, et al. A REDCap-based model for electronic consent (eConsent): moving toward a more personalized consent. J Clin Transl Sci 2020 Apr 3;4(4):345-353. [doi: 10.1017/cts.2020.30] [Medline: 33244416]
28. Glossary. Dexcom Clarity. URL: https://clarity.dexcom.com/glossary#eA1cTopic [accessed 2025-01-23]
29. de Bock M, Codner E, Craig ME, et al. ISPAD Clinical Practice Consensus Guidelines 2022: glycemic targets and glucose monitoring for children, adolescents, and young people with diabetes. Pediatr Diabetes 2022 Dec;23(8):1270-1276. [doi: 10.1111/pedi.13455] [Medline: 36537523]
30. Secrest AM, Costacou T, Gutelius B, Miller RG, Songer TJ, Orchard TJ. Associations between socioeconomic status and major complications in type 1 diabetes: the Pittsburgh Epidemiology of Diabetes Complication (EDC) Study. Ann Epidemiol 2011 May;21(5):374-381. [doi: 10.1016/j.annepidem.2011.02.007] [Medline: 21458731]
31. Simba S, Von Oettingen JE, Rahme E, Ladd JM, Nakhla M, Li P. Socioeconomic disparities in glycemic management in children and youth with type 1 diabetes: a retrospective cohort study. Can J Diabetes 2023 Dec;47(8):658-664. [doi: 10.1016/j.jcjd.2023.07.005] [Medline: 37481125]
32. Mönkemöller K, Müller-Godeffroy E, Lilienthal E, et al. The association between socio-economic status and diabetes care and outcome in children with diabetes type 1 in Germany: the DIAS study (diabetes and social disparities). Pediatr Diabetes 2019 Aug;20(5):637-644. [doi: 10.1111/pedi.12847] [Medline: 30912245]
33. Gonder-Frederick LA, Zrebiec JF, Bauchowitz AU, et al. Cognitive function is disrupted by both hypo- and hyperglycemia in school-aged children with type 1 diabetes: a field study. Diabetes Care 2009 Jun;32(6):1001-1006. [doi: 10.2337/dc08-1722] [Medline: 19324943]
34. Wadwa RP, Laffel LM, Shah VN, Garg SK. Accuracy of a factory-calibrated, real-time continuous glucose monitoring system during 10 days of use in youth and adults with diabetes. Diabetes Technol Ther 2018 Jun;20(6):395-402. [doi: 10.1089/dia.2018.0150] [Medline: 29901421]
35. Sinha M, McKeon KM, Parker S, et al. A comparison of time delay in three continuous glucose monitors for adolescents and adults. J Diabetes Sci Technol 2017 Nov;11(6):1132-1137. [doi: 10.1177/1932296817704443] [Medline: 28459159]
36. Schmelzeisen-Redeker G, Schoemaker M, Kirchsteiger H, Freckmann G, Heinemann L, Del Re L. Time delay of CGM sensors: relevance, causes, and countermeasures. J Diabetes Sci Technol 2015 Aug 4;9(5):1006-1015. [doi: 10.1177/1932296815590154] [Medline: 26243773]

Abbreviations

ARC: Ambulatory Research in Cognition
CGM: continuous glucose monitoring
EMA: ecological momentary assessment
REDCap: Research Electronic Data Capture
T1D: type 1 diabetes
TIR: time in range

XSL•FO
RenderX

Edited by A Mavragani; submitted 06.05.24; peer-reviewed by L Germine, NS Chaytor; revised version received 20.12.24; accepted 21.12.24; published 11.02.25.

Please cite as:
Ray MK, Fleming J, Aschenbrenner A, Hassenstab J, Redwine B, Burns C, Arbelaez AM, Vajravelu ME, Hershey T
Assessing Dynamic Cognitive Function in the Daily Lives of Youths With and Without Type 1 Diabetes: Usability Study
JMIR Form Res 2025;9:e60275
URL: https://formative.jmir.org/2025/1/e60275
doi:10.2196/60275

© Mary Katherine Ray, Jorie Fleming, Andrew Aschenbrenner, Jason Hassenstab, Brooke Redwine, Carissa Burns, Ana Maria Arbelaez, Mary Ellen Vajravelu, Tamara Hershey. Originally published in JMIR Formative Research (https://formative.jmir.org), 11.2.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Use of a Technology-Based Fall Prevention Program With Visual Feedback in the Setting of Early Geriatric Rehabilitation: Controlled and Nonrandomized Study

Carolin Steinmetz[1], PhD; Christina Stenzel[2], BSc; Maj Sylvester[2], BSc; Denis Glage[2], PhD; Anne Linke[1], MSc; Monika Sadlonova[1,3,4,5], MD; Christine A F von Arnim[1,5], MD; Marlena Schnieder[1,6], MD; Miroslava Valentová[1,5], MD; Stephanie Heinemann[1], PhD

[1]Department of Geriatrics, University Medical Center Göttingen, Robert-Koch-Straße 40, Göttingen, Germany

[2]Institute of Sports Science, Georg-August-University Göttingen, Göttingen, Germany

[3]Department of Psychosomatic Medicine and Psychotherapy, University Medical Center Göttingen, Göttingen, Germany

[4]Department of Cardiovascular and Thoracic Surgery, University Medical Center Göttingen, Göttingen, Germany

[5]German Center for Cardiovascular Research (DZHK), partner site Lower Saxony, Göttingen, Germany

[6]Department of Neurology, University Medical Center Göttingen, Göttingen, Germany

Corresponding Author:
Carolin Steinmetz, PhD
Department of Geriatrics, University Medical Center Göttingen, Robert-Koch-Straße 40, Göttingen, Germany

Abstract

Background: The Otago program (OP) is evidence-based and focuses on fall prevention in older people. The feasibility and usability of a short-term digital program modeled after the principles of the OP in the setting of early geriatric rehabilitation (EGR) are unclear.

Objective: This study investigated the feasibility and usability of an additional technology-based fall prevention program (FPP) in the setting of EGR.

Methods: We performed a feasibility study in the setting of EGR. A sample of 30 patients (mobility at least by walker; mini-mental status test score >17) was recruited between March and June 2024 and compared with a retrospective cohort (n=30, former EGR patients). All patients in the intervention group (IG) received a supervised, OP-modified FPP thrice/week for 20 minutes using a technology-based platform called “Pixformance.” The device is a digital trainer and enables real-time corrections. The primary end point was the feasibility (given when 80% of the IG participated in 6 trainings within 2 weeks). Secondary outcomes were usability (patients’ and facilitators’ perspective; ≥75%), risk of falls (Berg Balance Scale), mobility (Timed Up and Go Test), functional independence (Functional Independence Measure), and activities of daily living (Barthel Index). Several further exploratory end points were analyzed including anxiety and depression (Four-Item Patient Health Questionnaire; PH-Q4). Data were accessed at entry to EGR and after 2 weeks prior to discharge. To analyze the pre-posttest results, the dependent Student *t* test and the Wilcoxon test were applied. A mixed ANOVA with repeated measurements was used for statistical analyses of time-, group-, and interaction-related changes.

Results: A cohort of 60 patients (mean 80.2, SD 6.1 y; 58% females, 35/60) was analyzed. The main indication for EGR was stroke (9/60, 15%). Patients were recruited into a prospective IG (n=30) and a retrospective control group (n=30). Of the 30 patients in the prospective IG, 11 patients (37%) completed 6 training sessions within 2 weeks. Reasons why participants did not complete 6 training sessions were diagnostic appointments (33%), pain/discomfort (33%), or fatigue (17%). EGR patients rated FPP usability at 84% and facilitators at 65% out of 100%. Pre-posttest analysis of the standard assessments showed a significant interaction in Berg Balance Scale (<.01). In both groups, a significant improvement over time was found in the Timed Up and Go Test (<.01), Barthel Index (<.01), and Functional Independence Measure (<.01). Likewise, in the IG, the PH-Q4 score (.02) improved.

Conclusions: While the technology-based FPP in the EGR setting was generally well-accepted by patients, with high usability ratings, its feasibility was limited. Only 37% of participants completed the required additional training sessions. Further studies should test the technology-based FPP as an integrated part of the EGR complex therapy concept. Our findings suggest potential benefits of incorporating technology-based FPPs in EGR, but further refinement is needed to enhance participation and feasibility.

(JMIR Form Res 2025;9:e66692) doi:10.2196/66692

XSL•FO
RenderX

KEYWORDS

fall prevention; fall prevention program; early geriatric rehabilitation; gerontology; older adult; elder; aging; digital exercise intervention; digital activity; physical exercise; functional capacity; new technology; technology-based; digital intervention; feasibility

Introduction

Hospitalization in Old Age and Its Consequences

In 2050, around 22.4 million people in Germany will be ≥65 years old. It is predicted that the number of hospital stays in this cohort will increase by around 30% over the next 6 years [1].

In the case of acute medical illness with hospitalization, old and multimorbid patients have a high prevalence of hospital-associated disability [2] with impairment of activities of daily living, cognition, malnutrition, and sensory deficits [3]. Functional impairments such as muscle atrophy of the lower extremities and balance disorders following hospitalization are very common (30% - 80%) [4,5]. The consequences of this deterioration include rehospitalization, nursing home admissions [6], an increased number of falls, a poorer quality of life (QoL), increased consumption of health-related resources [7], and an increase in mortality [8]. Preventing hospital-associated disability therefore is crucial because it significantly impacts patient outcomes and overall health care systems.

Early Geriatric Rehabilitation

Early geriatric rehabilitation (EGR) is a specialized multidisciplinary and multiprofessional approach for older patients in Germany who need rehabilitation treatment due to acute health events such as injuries or surgeries that require hospitalization [9]. In general, the duration of an EGR is 2 weeks and comprises 20 training sessions in acute medical care. The main aim of EGR is to improve functional independence as well as QoL so that patients can be released to their homes, an inpatient rehabilitation program, or lower-level care such as long-term nursing care [10].

Fall Prevention Program

Fall prevention programs (FPPs) are often used to optimize functional limitations in older people. The Otago program (OP) is one of these programs; it is evidence-based and was explicitly developed for older people. Long-term studies have shown significant results with regard to the reduction of falls and fall-related injuries [11,12], an increase in strength and balance as well as an increase in confidence in being able to continue with everyday activities [13]. The OP showed the greatest effectiveness in the high-risk groups of over 80-year-olds and people with a history of falls [13,14]. We hypothesized that such prevention programs are not only effective as long-term interventions but can also be effective when the duration of the intervention is short. Previous studies in the area of prehabilitation have shown that older, multimorbid patients benefit from short-term interventions of 5 - 14 days [15,16].

Likewise, Martínez-Velilla et al [17] conducted a randomized, controlled intervention study in 370 older patients (age 87.3, SD 4.9 y) who were hospitalized for acute treatment. The intervention group (IG) included individualized moderate strength training, balance exercises, and gait training. The control group (CG) received standard treatment and interventions when needed. The training was safe and 2 sessions daily were sufficient to significantly improve mobility measured by short physical performance battery (SPPB) and the Barthel Index (BI) during an average hospital stay of 5 days, whereas the CG showed a deterioration in mobility [17]. In particular, assessments for measuring a patient's mobility (eg, Timed Up and Go Test [TUG] or SPPB) have the highest sensitivity in the EGR setting compared with other methods and are suitable to determine the effects of an FPP [17,18].

Digital Interventions

The number of digital interventions in geriatrics is increasing and may help to improve outcomes such as physical activity or balance. A systematic review of eHealth applications in geriatric rehabilitation showed that digital interventions are promising and have the potential to improve rehabilitation outcomes. In 16 out of 17 studies, eHealth applications were feasible in this setting [19].

The technology-based device "Pixformance" is a digital trainer that uses an integrated 2D full high-definition camera to scan 25 body points, enabling real-time corrections. The screen is divided into 2 parts, one part acts like a mirror and the other shows a person who is demonstrating the exercise. It has been shown that videos that show real people in action are effective in achieving long-term behavioral change [20].

However, acceptance of digital interventions in older persons is often limited by factors such as unfamiliarity with technology, cognitive decline, physical impairments, and concerns about privacy or security. On the other hand, there might be several challenges when it comes to implementing digital interventions for older adults from a facilitators' perspective. Facilitators are individuals who assist patients in initiating the digital FPP and remain present while the patient interacts with the device. This study is the first to use "Pixformance" in the EGR setting as part of an evidence-based FPP to investigate feasibility and usability.

The purpose of this study is to evaluate the feasibility and usability (patients' and facilitators' perspectives) of a short-term technology-based FPP in the setting of EGR, performed in addition to the standard complex therapy concept. Furthermore, the effect of the FPP on the risk of falls, functional independence, activities of daily living, mobility, and cognitive performance compared with a retrospective CG was analyzed in order to determine which assessments can be used to sensitively detect changes due to an additional FPP. This analysis will be used to make informed decisions about the study design and instruments used in a follow-up project. The pre-posttest of muscle status, QoL, anxiety and depression as well as frailty status, which are assessed only in the IG, will also provide information about the sensitivity of these instruments for detecting change in this older patient cohort.

XSL•FO
RenderX

Methods

Patient Population

Patient recruitment was carried out in the Department of Geriatrics at the University Medical Center Göttingen, Germany, between March and June 2024.

All patients undergoing an EGR complex therapy, at least able to walk with a walker, and cognitively fit to slightly cognitively impaired (mini mental status test score >17) were eligible for inclusion.

Patients were excluded if they were unable to understand the study information and give written consent due to poor German language skills, or had cognitive or visual impairments.

Study Setting

This is a prospective 2-arm feasibility study in the setting of EGR. The IG was recruited prospectively and the CG retrospectively. Eligible patients for the IG were asked to participate in the feasibility study by physicians at the beginning of EGR during regular clinical consultations. All patients were provided study information and gave their written informed consent. Subsequently, the baseline assessment and follow-up assessment took place (Figure 1).

Figure 1. Research design and conducted assessments. CG: control group; EGR: early geriatric rehabilitation; IG: intervention group.

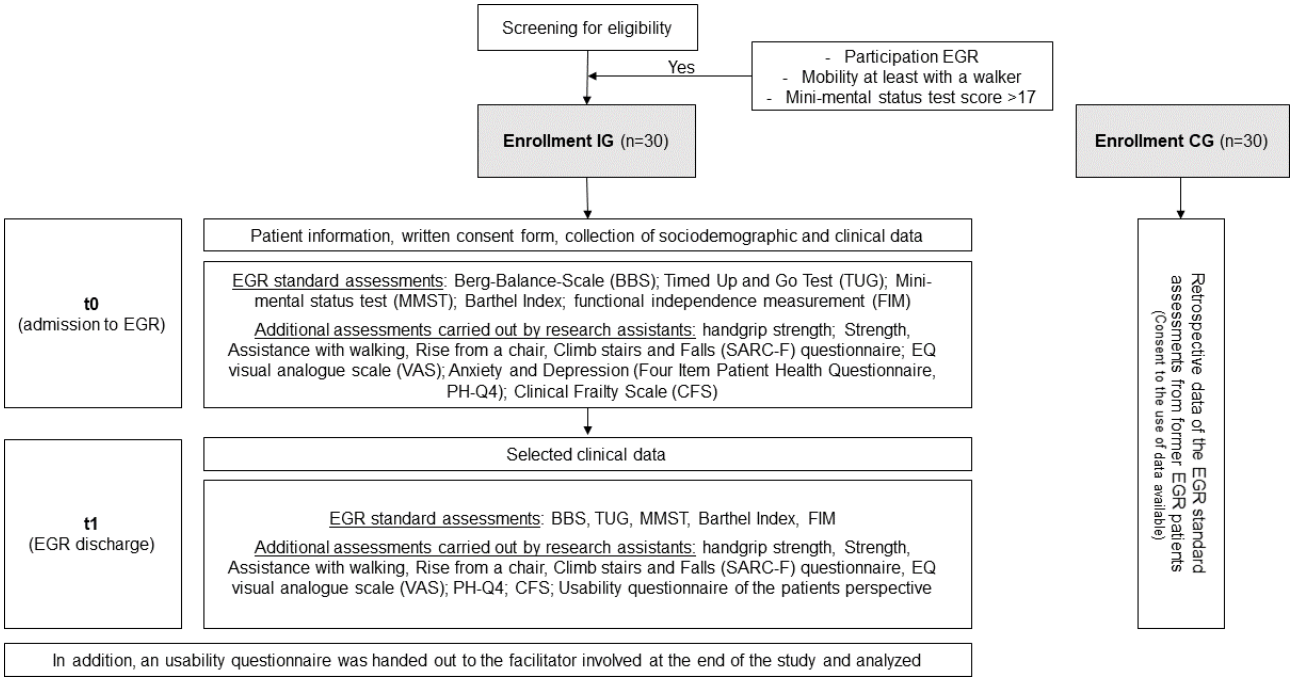

A retrospective CG consists of former patients who have undergone EGR complex therapy at the same department in the last 3 months prior to the recruitment of the prospective group and fulfill the inclusion criteria specified in this feasibility study. All included patients gave written consent for the use of their pseudonymized data for scientific purposes upon admission to the University Medical Center Göttingen. All CG patients were screened and included in the CG according to their order of admission.

Sample Size

A sample size of 60 patients (IG: n=30; CG: n=30) has 80% power to detect a standardized difference in means of 0.74 with 80% power, using a 2-group *t* test with a 5% 2-sided significance level. This corresponds to comparably large effects, which are likely to be clinically relevant. The sample size was seen as feasible, taking into account the number of potentially eligible patients undergoing EGR within 3 months, as well as the inclusion and exclusion criteria. The study design is comparable to a recently conducted pilot study by Fränzel et al [21] which includes a FPP with 58 participants (IG: n=29; CG: n=29) in the same setting. This study found significant developments in patient mobility measured by TUG, SPPB, and gait speed [21].

Procedure

All patients of the IG received the regular EGR complex therapy concept, which consists of 20 sessions, including regular physiotherapy, occupational and speech therapy as well as psychological support. In addition, 6 sessions of a supervised FPP modified according to the OP [22] using the technology-based platform called “Pixformance” (Figure 2) were offered to IG participants 3 times per week.

XSL•FO
RenderX

Figure 2. The pixformance station (picture provided by pixformance sports GmbH).

The content of the 2-week training program is summarized in Table 1.

XSL•FO
RenderX

Table . Overview of the content of the fall prevention program.

Category	Details
Frequency	Thrice/week; 10 repetitions, 1 set
Intensity	Body weight, Rating of Perceived Exertion scale 13 - 15 (medium)
Time	20 min
Type	mobilization, coordination, strengthening exercises focusing on the lower extremities, relaxation/stretching
Exercises	Examples include shoulder and torso rotation, calf raises, lateral leg lift, standing hamstring curl, foot lift and grinding heel in sitting position, and standing up and sitting down from a chair

Before being discharged from the EGR, all participants in the IG received an Otago booklet to continue the exercises at home.

Outcomes

The primary end point was the feasibility of the technology-based FPP in the setting of EGR. We defined feasibility as at least 80% of the patients completing 6 training sessions within 2 weeks in addition to the regular 20 therapeutic sessions in the EGR. For the statistical analysis, we report how many participants completed the training as well as the reasons for not completing the training in absolute numbers and percentages.

User Experience

Usability was assessed from the patients' and facilitators' perspectives. Facilitators did not provide therapy directly; instead, they addressed technical issues, such as starting, operating, or stopping the device. All training instructions and feedback were delivered by the device itself, not by the facilitators. In this study, facilitators included both geriatric staff (2 physical therapists and 2 occupational therapists) and study personnel (3 sport science students and 1 medical student), all of whom received training in device operation. All facilitators provided written informed consent.

At the end of the study, all patients and facilitators filled out a usability questionnaire. The results of the structured questionnaire were quantitatively analyzed. The device was considered to be usable if the "Overall rating for the training with the device" and "Overall rating for using the device" was ≥75%, modified according to Rabinovich et al [23]. For the statistical analysis, we report the results of the usability questionnaire as the percent of the 2 highest values on the 5-point Likert scale.

Secondary Outcomes

In addition to feasibility and user experience data, further end points are detailed in Figure 1. Other secondary outcomes included the risk of falls assessed by Berg Balance Scale [24], mobility assessed with the TUG [25], functional independence carried out with the help of the Functional Independence Measure (FIM) [26], and activity of daily living collected by BI [27,28].

Data were assessed at the start of the EGR and after 2 weeks. How the additional technology-based FPP influences the effect of EGR complex therapy was examined using standard assessment data from the IG and a retrospective CG.

Statistical Analysis

We used the Kolmogorov–Smirnov test for the assessment of normal distributions. Continuous and categorical variables are presented by mean (SD) with absolute and relative frequencies, respectively. Two-group comparisons of baseline variables were performed using Student's *t* test and chi-square test of independence for continuous and categorical variables. To analyze the results of the pre-posttest, dependent Student *t* test and the Wilcoxon signed rank test were applied. A 2-way mixed ANOVA with repeated measurements was used for statistical analyses of time- (t0/t1), group- (IG or CG), and interaction-related (between time and group) changes. In all analyses, a $P<.05$ was considered statistically significant. The results of the usability questionnaire include the percent of the 2 highest values on the 5-point Likert scale. Analyses were performed using the IBM Statistical Package for the Social Sciences (SPSS version 29.0; IBM Co).

Ethical Considerations

This feasibility study was approved by the local ethics committee of the University Medical Center Göttingen (application number: 22/1/24; January 23, 2024) and conforms to the Declaration of Helsinki. All participants signed informed consent forms.

Results

Study Population

A total of 70 patients were screened for eligibility for the prospective IG. Of these, 40 were excluded because of "no interest" (n=13), "poor mobility" (n=10), "hospital stay too short" (n=5), "low cognition" (n=4), "low cognition and poor mobility" (n=3), "visual impairment" (n=2), and "others" (n=3). In the end, 30 patients from the IG participated in the FPP.

A sample of 60 patients (mean age 80.2 years, SD 6.1 years; 58% females, 35/60; IG: n=30; CG: n=30) were included in the analysis. Patients of the prospective IG were consecutively recruited (n=30; mean age 79.0 years, SD 6.0 years; 53% female, 16/30) at admission to EGR. The retrospective CG (mean age 85.0 years, SD 4.6 years; 63% female, 19/30) is made up of former EGR patients who were treated in the 3 months prior to the "Pixformance" intervention and fulfilled all inclusion and exclusion criteria.

The baseline characteristics of the IG showed the cohort to be significantly younger and taller than the CG. With regards to

XSL•FO
RenderX

clinical and medical history characteristics, patients in the IG were more likely to be taking a higher number of medications and had a degree of disability, a post-Covid condition, and depressive symptoms in comparison to the retrospective CG (Table 2).

XSL•FO
RenderX

Table . Baseline characteristics of the cohort.

Characteristics	All (n=60)	IG[a] (n=30)	CG[b] (n=30)	*P* value
Age (years), mean (SD)	82.0 (6.1)	79.0 (6.0)	85.0 (4.6)	<.01[c]
Sex, n (%)				
Male	24 (40)	14 (47)	10 (33)	.34[d]
Female	35 (58)	16 (53)	19 (63)	.34[d]
Unknown	1 (2)	0 (0)	1 (2)	.31[d]
Height (cm), mean (SD)	166.4 (8.4)	168.6 (8.0)	164.1 (8.4)	.04[c]
Weight (kg), mean (SD)	71.4 (16.6)	73.8 (18.3)	68.9 (14.6)	.26[c]
BMI (kg/m²), mean (SD)	25.8 (5.2)	26.1 (5.9)	25.5 (4.5)	.67[c]
Blood pressure (mmHg[e]), mean (SD)				
Systolic	N/A[f]	131.7 (23.1)	N/A	N/A
Diastolic	N/A	73.1 (12.0)	N/A	N/A
Heart rate (beats/min), mean (SD)	N/A	71.4 (11.3)	N/A	N/A
Number of medications (regularly), mean (SD)	7.6 (4.1)	9.1 (3.9)	6.1 (3.7)	.01[c]
Need for nursing assistance, n (%)	30 (50)	16 (53)	14 (47)	.87[d]
Any degree of disability, n (%)	9 (15)	8 (27)	1 (3)	<.01[d]
Situation of living, n (%)				
Living alone	N/A	11 (18)	N/A	N/A
Visual aid, n (%)	41(68)	24 (80)	17 (57)	.07[d]
Hearing aid device, n (%)	16 (27)	7 (23)	9 (30)	.08[d]
Walking aid, n (%)	34 (57)	15 (50)	19 (63)	.29[d]
Walker	26 (43)	12 (40)	14 (47)	.60[d]
Cane	4 (7)	2 (7)	2 (7)	1.0[d]
Forearm crutch	5 (8)	1 (3)	4 (13)	.16[d]
Main diagnosis for EGR[g] admission, n (%)				
Poststroke or suspected stroke	9 (15)	5 (17)	4 (13)	.72[d]
Cardiac decompensation	8 (13)	4 (13)	4 (13)	1.0[d]
Acute kidney failure	4 (7)	2 (7)	2 (7)	>.99[d]
Vertigo including syncope	4 (7)	3 (10)	1 (3)	.30[d]
Concomitant diseases, n (%)				
Diabetes mellitus	23 (38)	12 (40)	11 (37)	.12[d]
Hypertension	35 (58)	16 (53)	19 (63)	.06[d]
Dyslipidemia	15 (25)	6 (20)	9 (30)	.39[d]
Stroke or TIA[h]	12 (20)	6 (20)	6 (20)	.37[d]

XSL•FO
RenderX

Characteristics	All (n=60)	IG[a] (n=30)	CG[b] (n=30)	*P* value
Dementia	7 (12)	3 (10)	4 (13)	.41[d]
Kidney failure	13 (42)	5 (17)	8 (27)	.19[d]
Cancer disease<5 years	5 (8)	3 (10)	2 (7)	.54[d]
Cancer disease ≥5 years	5 (8)	2 (7)	3 (10)	.50[d]
Thyroid disorder	8 (13)	6 (20)	2 (7)	.21[d]
Rheumatism	10 (17)	8 (27)	2 (7)	.08[d]
Condition after COVID-19 infection	20 (33)	17 (57)	3 (10)	<.01[d]
Depression	5 (8)	1 (3)	4 (13)	.04[d]
Anxiety	5 (8)	3 (10)	2 (7)	.19[d]

[a]IG: intervention group.
[b]CG: control group.
[c]Independent *t* test.
[d]Chi-square test.
[e]mmHg: millimeters of mercury.
[f]N/A: not available.
[g]EGR: early geriatric rehabilitation.
[h]TIA: transient ischemic attack.

Feasibility of the FPP

Out of a total of 30 patients in the IG, 11 (37%) participated in 6 trainings within 2 weeks. Of these 11 patients, 7 (64%) completed the FPP including all strengthening/balance exercises. The remaining 4 patients ended their training session early or skipped one or more important strengthening or balance exercises.

The main reasons why participants did not complete the amount of training were the inability to attend the FPP due to diagnostic appointments (33%), pain/discomfort (33%), or fatigue (17%).

User Experience

The analysis of user experiences from the patients' and facilitators' perspectives is summarized in Table 3. The 8 facilitators were on average 35.1 (SD 12.3) years old. A quarter of the facilitators were male (2/8, 25%) and 3 quarters were female (6/8, 75%). The group members had an average work experience of 20.0 (SD 8.8) years.

XSL•FO
RenderX

Table . Overview of the results of the user experience questionnaire from the perspective of patients and facilitators.

Usability items	Rating	Patients' perspective (n=26), (%)	Facilitators' perspective (n=8), (%)
Instructions by facilitators for using the device were clear	Agreeing	96	__[a]
There were technical problems	Never	85	—
Training with the device was comfortable	Often	92	—
Training with the device was embarrassing	Never	100	—
Corrections during the training by the device were clear	Agreeing	89	—
Willingness to continue the learned exercises at home	≥3 months	54	—
Overall patient usability rating (%)(100%=verygood experience), mean (SD)	—	84.0 (15.90)	—
Instructions by the manufacturer for using the device were clear	Agreeing	—	50
Difficulties when commissioning the device	No starting problems	—	63
Regular use of the device was easy	Agreeing	—	50
The exercise selection of the device was easy to handle	Agreeing	—	63
Selected exercises could be easily adapted during the training	Agreeing	—	25
The device was easy to switch on and off	That worked very well	—	75
I am in favor of the future use of the device	Yes, very	—	50
Overall facilitator usability rating (%) (100%=verygood experience), mean (SD)	—	—	65.4 (20.0)

[a]Not applicable.

The overall usability rating for the device from the patients' perspective was 84.0% (SD 15.9%), where 0% means a bad experience and 100% a very good experience. The facilitators rated the usability of the device at 65.4% (SD 20.0%).

Secondary Outcomes

The results of the EGR standard assessments are indicated in Table 4. For 22 patients in the IG and 29 patients in the CG, complete measurements at both time periods were recorded. The 2-way mixed ANOVA showed a significant interaction between groups over time in the Berg Balance Scale (<.01). Likewise, a significant positive development over time was seen in both groups in TUG time (<.01), FIM motor subscale (<.01) score, and BI (<.01).

XSL•FO
RenderX

Table . Descriptive and 2-way mixed ANOVA results for the early geriatric rehabilitation routine assessments at baseline (t0, admission) and after 2 weeks (t1, discharge) differentiated by group.

Parameters	CG[a], mean (SD)	IG[b], mean (SD)	*P* value		
			Time	Group	Interaction
Berg Balance Scale (score)			.54	.02	<.01
t0[c]	31.3 (12.5)	33.6 (10.0)			
t1[d]	38.4 (8.7)	24.5 (7.5)			
Timed Up and Go Test (s)			<.01	.40	.82
t0	22.1 (9.0)	20.2 (7.6)			
t1	17.5 (7.5)	16.1 (6.7)			
Minimental status test (score)			.07	.21	.21
t0	24.2 (3.3)	24.6 (4.3)			
t1	24.7 (3.0)	27.2 (3.9)			
Barthel Index			<.01	.68	.78
t0	61.6 (17.5)	63.7 (18.2)			
t1	76.2 (12.9)	76.8 (15.9)			
FIM[e] motor subscale (score)			<.01	.35	.64
t0	53.8 (12.2)	51.0 (10.6)			
t1	67.0 (10.9)	65.3 (6.6)			

[a]CG: control group.

[b]IG: intervention group.

[c]t0: baseline (early geriatric rehabilitation admission).

[d]t1: after 2 weeks (early geriatric rehabilitation discharge).

[e]FIM: Functional Independence Measure.

The Wilcoxon signed rank test confirmed the significant development of the mixed ANOVA in Berg Balance Scale (BBS) in IG (.01; $n_{paired_results}$ =22) and CG (<.01; $n_{paired_results}$ =29). In the IG, 5 patients improved and 17 showed poorer results in BBS between baseline and EGR discharge, whereas in the CG, 17 participants improved, 2 showed poorer results and 10 did not change in BBS.

Table 5 summarizes the results of muscle status, QoL, anxiety and depression, and frailty status, which were additionally assessed by study personnel in the IG. A significant improvement in Four-Item Patient Health Questionnaire (PH-Q4) score (.024) from baseline to EGR discharge was observed. Furthermore, the frailty status and the positive screen for sarcopenia each decreased by 27%.

XSL•FO
RenderX

Table . Pre-posttest and descriptive results for the assessments at baseline (t0) and after 2 weeks of early geriatric rehabilitation (t1) of the intervention group (IG).

Parameters	IG		*P* value
	t0 (n=30)	t1 (n=26)	
Handgrip strength (kg), mean (SD)	24.9 (7.6)	24.7 (7.3)	.93[a]
Female	26.2 (6.4)	20.0 (6.2)	.40[a]
Male	23.9 (8.6)	28.4 (6.1)	.18[a]
CFS[b] (score), mean (SD)	4.3 (1.5)	3.8 (1.6)	.27[a]
Frailty (CFS ≥4), n (%)	21 (70)	13 (43)	>.99[c]
SARC-F[d] score, mean (SD)	4.0 (2.4)	3.4 (1.9)	.28[a]
Sarcopenia (positive screen SARC-F ≥4), n (%)	18 (60)	10 (33)	.13[c]
EQ VAS[e], mean (SD)	52.1 (15.4)	60.4 (13.0)	.06[a]
PH-Q4[f], mean (SD)	3.3 (2.4)	1.6 (1.9)	.02[a]

[a]Dependent *t* test.

[b]CFS: Clinical Frailty Scale.

[c]Chi-square test.

[d]SARC-F: Strength, Assistance with walking, Rise from a chair, Climb stairs and Falls questionnaire.

[e]EQ VAS: European Quality of Life Visual Analogue Scale.

[f]PH-Q4: Four-Item Patient Health Questionnaire.

Discussion

Summary of Target and Findings

In this study, we aimed to evaluate the feasibility and usability of a short-term technology-based FPP in the EGR setting from the patients' and facilitators' perspectives. The FPP was conducted in addition to the conventional complex therapy concept. Furthermore, the effect on the risk of falls, daily function, mobility, and cognitive performance compared with a retrospective CG was analyzed. In the IG, the pre-posttest of muscle status, QoL, anxiety and depression as well as frailty status were additionally examined. Regarding the primary end point, we observed the limited feasibility of the FPP in the EGR setting, while usability and acceptance from the patients' perspective were proven. The results of the BBS showed a significant decrease in the IG and a significant improvement in the CG. Further significant improvements from EGR admission to discharge were observed in TUG, FIM, BI, and PH-Q4.

This study is the first to use the "Pixformance" device in the EGR setting as part of an evidence-based FPP.

Study Cohort

The number of older patients receiving complex therapy in EGR in Germany (211,270 participants in 2007 to 274,926 in 2020) is increasing steadily due to demographic change [29]. The baseline characteristic of our cohort is similar to others who performed studies in the EGR setting [21,30,31]. The other studies reported an average age of 78 to 87 years, the admission of more women than men, and strokes, cardiac decompensation, and hip fractures as the most common admission diagnoses [21,30,31]. In our cohort, significant differences between the groups were observed in age, height, number of regularly taken medications, degree of disability, condition after COVID-19 infection, and depression. Our CG is an unmatched-pairing retrospective comparison cohort of former patients from the time point directly before the intervention. The patients were consecutively included in the CG, beginning at the start of the intervention and going backward in time until 30 patients were identified. In this feasibility study, a CG was included in order to determine which assessments might be used to sensitively detect changes due to an additional FPP in order to design a follow-up project.

Participants of the IG were younger and had more often severe concomitant diseases. The high number of patients after COVID-19 infection in the IG may explain why only 11 out of 30 patients completed 6 training sessions within 2 weeks. Fatigue is one of the most common reasons why the FPP was not completed. Especially older adults aged 65 years or older are at greater risk of persisting symptoms associated with COVID-19; fatigue is 1 of the 5 common long-term symptoms [32].

eHealth in Geriatrics

Kraaijkamp et al [19] published a systematic review of eHealth in the setting of geriatric rehabilitation. eHealth is defined as "the use of digital information and communication to support and/or improve health and healthcare" [19,33]. The authors summarized that eHealth is often feasible and can potentially improve geriatric rehabilitation outcomes. Simple eHealth interventions are more likely to be feasible for older patients receiving geriatric rehabilitation [19]. "Pixformance" can be seen as such a "simple" eHealth intervention. Likewise, the

XSL•FO
RenderX

authors concluded that there is a lack of evidence on usability which might hamper the implementation [19].

Feasibility of the FPP

In our study, feasibility was defined as participation in 6 FPP training sessions throughout the EGR. The most common reason for missing a training session was due to medical examinations outside the ward. Sometimes, there were delays in the morning so patients could not participate in the FPP. The clinical routine always took priority. The low participation rate was therefore due to logistical problems that the study team was unable to solve. An additional training session in the EGR setting therefore seems to be too much. Similarly, Fränzen et al [21] conducted a pilot study including a FPP in the form of a square-stepping exercise (SSE) in the EGR setting and investigated its feasibility. The required minimum number of SSE units, each lasting 30 minutes, was 6 during the EGR stay. The amount of exercise units is in line with our target and definition of feasibility. The SSE program proved to be feasible. The main difference between the 2 interventions was that the SSE was performed as a part of conventional physiotherapy included in the complex therapy concept [21]. In our feasibility study, the FPP took place in the afternoons in addition to the complex therapy concept. In a future study, we will include the study intervention as part of conventional physiotherapy in the complex therapy concept so that logistical problems do not affect the intervention.

User Experience

We have investigated the usability of the technology-based FPP and followed the advice from Kraaijkamp et al [19] to pave the way for a possible implementation [19]. User experience was assessed independently from the facilitators' and patients' perspectives. Usability from the patients' perspective was achieved, while the facilitators' did not reach the target of ≥75%. The results are surprising in that eHealth is often not sufficiently tailored to age-related barriers such as cognition, physical ability, perception, and motivation [34,35]. "Pixformance" as a "simple" eHealth intervention seems to be attractive to patients. Similarly, a usability study of technology-based exercise games in older patients following hip replacement surgery confirmed that the cohort enjoyed performing exercises under digital guidance [36]. In contrast to this, usability issues were pointed out by facilitators.' Only every second respondent was in favor of using the device in the future in the EGR setting. The item "selected exercises could be easily adapted during the training" was only agreed on by 25% of the facilitators. The low rating could be because the exercises had to be changed via a separate computer platform. Facilitators' explained that there was not enough time to learn how to adapt the exercises and navigate the training platform in detail. This may be due to the nature of the study, which was conducted in addition to the staff's normal daily activities [37].

Secondary Outcomes

The BBS is widely used in clinical settings to assess the risk of falls, particularly in older individuals with balance impairments or in geriatric wards within general hospitals [38,39]. A score of less than 45 on the BBS is considered the threshold for increased fall risk. A change of 7 points is usually considered clinically relevant. When the score lies between 35 and 44 points, an increased risk of falls is indicated [40]. At baseline, the mean BBS scores of our cohort ranged from 31 to 34 points, indicating an increased risk of falls. Interestingly, the results of the BBS showed a significant decrease in the IG (−9.1 points) while the CG showed a significant improvement (+7.1). Even so, both groups remained in the range of increased fall risk. One reason for this unexpected outcome could be that the BBS may not be the most appropriate assessment tool to capture changes resulting from a FPP. The BBS primarily includes tasks that were not trained in the FPP, such as "Retrieving object from floor" or "turning to look behind." For a follow-up project, the SPPB appears to be a more suitable assessment tool. As mentioned in the introduction, measuring a patient's mobility using tools like the SPPB has shown the highest sensitivity in EGR settings compared with other methods [17,18]. The SPPB consists of 3 parts: "gait speed test," "chair rising test," and "balance test." These tests are well suited for measuring changes resulting from the FPP [41]. Moreover, the SPPB is frequently used in studies conducted in cohorts similar to ours [21,42]. It should be noted, however, that there was some missing data in the IG, resulting in only 22 paired results being included in the analysis. Therefore, these findings should be interpreted with caution.

Further significant positive developments (ie, the patients' condition improved after 2 weeks) based on the multimodal and interdisciplinary EGR complex therapy concept [9,10] were found in functional independence (FIM), activity of daily living (BI), mobility (TUG), and anxiety and depression (PH-Q4). These instruments, therefore, seem appropriate for measuring differences due to digital FPP in further studies.

Limitations

This feasibility study has a relatively small cohort and includes an unmatched-pairing retrospective comparison cohort. Further studies with larger cohorts and a randomized controlled design with blinded assessors are needed to validate the results presented. Key factors such as disability (yes or no), sex (male or female), and age (younger than 81 years vs 81 years and older) should be considered in the 1:1 block randomization process with random block lengths to ensure a homogeneous cohort at baseline. Additionally, the training intensity of the FPP might be increased by incorporating a digital balance pad or components of the SSE unit intervention [21] to improve outcomes.

Conclusion

Due to the high percentages of frail and sarcopenic patients admitted to EGR with an elevated fall risk, there is a significant unmet need for a FPP. However, the feasibility of the technology-based FPP was not achieved, as only 37% of participants completed 6 additional FPP sessions within 2 weeks, primarily due to diagnostic appointments, pain/discomfort, or fatigue. Usability analysis indicates that EGR patients accept the technology-based FPP, while staff members are less convinced. Significant positive changes from EGR entry to discharge were observed in TUG, FIM motor subscale, and BI, independent of the FPP. A statistically significant, but not clinically relevant improvement in the BBS was noted in the

XSL•FO
RenderX

CG, while the IG showed a significant positive change in the PHQ-4 score, indicating fewer symptoms of anxiety and depression. Further studies should explore the technology-based FPP as an integrated part of the EGR complex therapy concept, increase the FPP intensity by incorporating a digital balance pad or components of the SSE, and assess progress from admission to discharge using the SPPB instead of the BBS.

Acknowledgments

We want to thank all patients as well as the staff who were involved in this study. Our special thanks go to Niklas Meyer and Mira Fleur Seitz, who supported the team during training sessions and assessments. We also thank Pixformance Sports GmbH, which generously loaned the "Pixformance" device for 3 months during the feasibility study. This cohort study was financially supported by the University Medical Center Göttingen, Department of Geriatrics.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

CAFvA is the principal investigator. CStei and SH coordinated the study as well as prepared the ethics documents. The protocol was developed by CStei, SH, MSch, MV, CAFvA, and MSad. The content of the modified Otago program was created by CStei, CSten, MSyl, and DG. MSch und MV screened patients for the inclusion and exclusion criteria and informed patients about the study. CSten and MSyl conducted the assessments and served as facilitators during the exercise training sessions. CStei, SH, CSten, and MSyl were responsible for study protocol implementation. CStei was responsible for statistical analysis and writing. CStei, SH, MSch, MV, CAFvA, MSad, and DG contributed to the data interpretation. CStei and SH prepared the initial draft of the manuscript. All authors reviewed and commented upon the initial draft and approved the final manuscript.

Conflicts of Interest

CAFvA received honoraria from serving on the scientific advisory board of Biogen, Roche, Novo Nordisk, and Dr Willmar Schwabe GmbH & Co. KG and has received funding for travel as well as speaker honoraria from Biogen, Roche Diagnostics AG, Medical Tribune Verlagsgesellschaft mbH, Landesvereinigung für Gesundheit und Akademie für Sozialmedizin Niedersachsen e. V., and Dr Willmar Schwabe GmbH & Co. KG. CAFvA has received research support from Roche Diagnostics AG and research funding from the Innovationsfond (Fund of the Federal Joint Committee, Gemeinsamer Bundesausschuss, G-BA grant numbers VF1_2016-201; 01NVF21010; 01VSF21019). The remaining authors have nothing to disclose.

References

1. bevölkerungsvorausberechnung für deutschland [population forecast for Germany]. Statistisches bundesamt [Federal Statistical Office]. URL: https://service.destatis.de/bevoelkerungspyramide/#!y=2050&a=20,65&g [accessed 2024-08-26]
2. Skains RM, Zhang Y, Osborne JD, et al. Hospital-associated disability due to avoidable hospitalizations among older adults. J Am Geriatr Soc 2023 May;71(5):1395-1405. [doi: 10.1111/jgs.18238] [Medline: 36661192]
3. Buurman BM, Hoogerduijn JG, de Haan RJ, et al. Geriatric conditions in acutely hospitalized older patients: prevalence and one-year survival and functional decline. PLoS ONE 2011;6(11):e26951. [doi: 10.1371/journal.pone.0026951] [Medline: 22110598]
4. Zisberg A, Shadmi E, Sinoff G, Gur-Yaish N, Srulovici E, Admi H. Low mobility during hospitalization and functional decline in older adults. J Am Geriatr Soc 2011 Feb;59(2):266-273. [doi: 10.1111/j.1532-5415.2010.03276.x] [Medline: 21314647]
5. Martínez-Velilla N, Cadore EL, Casas-Herrero Á, Idoate-Saralegui F, Izquierdo M. Physical activity and early rehabilitation in hospitalized elderly medical patients: systematic review of randomized clinical trials. J Nutr Health Aging 2016;20(7):738-751. [doi: 10.1007/s12603-016-0683-4] [Medline: 27499308]
6. Fortinsky RH, Covinsky KE, Palmer RM, Landefeld CS. Effects of functional status changes before and during hospitalization on nursing home admission of older adults. J Gerontol A Biol Sci Med Sci 1999 Oct;54(10):M521-M526. [doi: 10.1093/gerona/54.10.m521] [Medline: 10568535]
7. Covinsky KE, King JT Jr, Quinn LM, et al. Do acute care for elders units increase hospital costs? a cost analysis using the hospital perspective. J Am Geriatr Soc 1997 Jun;45(6):729-734. [doi: 10.1111/j.1532-5415.1997.tb01478.x] [Medline: 9180668]
8. Brown CJ, Friedkin RJ, Inouye SK. Prevalence and outcomes of low mobility in hospitalized older patients. J Am Geriatr Soc 2004 Aug;52(8):1263-1270. [doi: 10.1111/j.1532-5415.2004.52354.x] [Medline: 15271112]
9. Stier-Jarmer M, Pientka L, Stucki G. Frührehabilitation in der geriatrie [early geriatric rehabilitation]. Phys Rehab Kur Med 2002 Aug;12(4):190-202. [doi: 10.1055/s-2002-33901]

XSL•FO
RenderX

10. Palzer M, Meyer U, Abderhalden LA, Gazzotti A, Hierholzer C, Bischoff-Ferrari HA. Geriatrische komplexbehandlung bei alterstraumatologischen patienten: einfluss auf mobilität, selbsthilfefähigkeit und austrittsdestination. [acute geriatric treatment of older trauma patients: influence on mobility, autonomy and postdischarge destination]. Z Gerontol Geriatr 2021;54:816-822. [doi: 10.1007/s00391-020-01812-4] [Medline: 33201306]
11. Campbell AJ, Robertson MC, Gardner MM, Norton RN, Tilyard MW, Buchner DM. Randomised controlled trial of a general practice programme of home based exercise to prevent falls in elderly women. BMJ 1997 Oct 25;315(7115):1065-1069. [doi: 10.1136/bmj.315.7115.1065] [Medline: 9366737]
12. Campbell AJ, Robertson MC, Gardner MM, Norton RN, Buchner DM. Falls prevention over 2 years: a randomized controlled trial in women 80 years and older. Age Ageing 1999 Oct;28(6):513-518. [doi: 10.1093/ageing/28.6.513] [Medline: 10604501]
13. Robertson MC, Campbell AJ, Gardner MM, Devlin N. Preventing injuries in older people by preventing falls: a meta-analysis of individual-level data. J Am Geriatr Soc 2002 May;50(5):905-911. [doi: 10.1046/j.1532-5415.2002.50218.x] [Medline: 12028179]
14. Robertson MC, Devlin N, Gardner MM, Campbell AJ. Effectiveness and economic evaluation of a nurse delivered home exercise programme to prevent falls. 1: randomised controlled trial. BMJ 2001 Mar 24;322(7288):697-701. [doi: 10.1136/bmj.322.7288.697] [Medline: 11264206]
15. Herdy AH, Marcchi PLB, Vila A, et al. Pre- and postoperative cardiopulmonary rehabilitation in hospitalized patients undergoing coronary artery bypass surgery: a randomized controlled trial. Am J Phys Med Rehabil 2008 Sep;87(9):714-719. [doi: 10.1097/PHM.0b013e3181839152] [Medline: 18716482]
16. Steinmetz C, Bjarnason-Wehrens B, Baumgarten H, Walther T, Mengden T, Walther C. Prehabilitation in patients awaiting elective coronary artery bypass graft surgery - effects on functional capacity and quality of life: a randomized controlled trial. Clin Rehabil 2020 Oct;34(10):1256-1267. [doi: 10.1177/0269215520933950] [Medline: 32546065]
17. Martínez-Velilla N, Casas-Herrero A, Zambom-Ferraresi F, et al. Effect of exercise intervention on functional decline in very elderly patients during acute hospitalization: a randomized clinical trial. JAMA Intern Med 2019 Jan 1;179(1):28-36. [doi: 10.1001/jamainternmed.2018.4869] [Medline: 30419096]
18. Heldmann P, Werner C, Belala N, Bauer JM, Hauer K. Early inpatient rehabilitation for acutely hospitalized older patients: a systematic review of outcome measures. BMC Geriatr 2019 Jul 9;19(1):189. [doi: 10.1186/s12877-019-1201-4] [Medline: 31288750]
19. Kraaijkamp JJM, van Dam van Isselt EF, Persoon A, Versluis A, Chavannes NH, Achterberg WP. eHealth in geriatric rehabilitation: systematic review of effectiveness, feasibility, and usability. J Med Internet Res 2021 Aug 19;23(8):e24015. [doi: 10.2196/24015] [Medline: 34420918]
20. Abu Abed M, Himmel W, Vormfelde S, Koschack J. Video-assisted patient education to modify behavior: a systematic review. Patient Educ Couns 2014 Oct;97(1):16-22. [doi: 10.1016/j.pec.2014.06.015] [Medline: 25043785]
21. Fränzel K, Koschate J, Freiberger E, Shigematsu R, Zieschang T, Tietgen S. Square-stepping exercise in older inpatients in early geriatric rehabilitation. a randomized controlled pilot study. BMC Geriatr 2024 Apr 10;24(1):326. [doi: 10.1186/s12877-024-04932-3] [Medline: 38600478]
22. Campbell AJ, Robertson MC. Otago exercise programme to prevent falls in older adults. Live Stronger for Longer. URL: https://www.livestronger.org.nz/assets/Uploads/acc1162-otago-exercise-manual.pdf [accessed 2024-08-22]
23. Rabinovich RA, Louvaris Z, Raste Y, et al. Validity of physical activity monitors during daily life in patients with COPD. Eur Respir J 2013 Nov;42(5):1205-1215. [doi: 10.1183/09031936.00134312] [Medline: 23397303]
24. Scherfer E, Bohls C, Freiberger E, Heise KF, Hogan D. Berg-Balance-Scale - deutsche version. Physioscience 2006 Jun;2(2):59-66. [doi: 10.1055/s-2006-926833]
25. Podsiadlo D, Richardson S. The timed “Up & Go”: a test of basic functional mobility for frail elderly persons. J Am Geriatr Soc 1991 Feb;39(2):142-148. [doi: 10.1111/j.1532-5415.1991.tb01616.x] [Medline: 1991946]
26. Lüthi H. Alltagsfähigkeiten zuverlässig messen [reliably measuring everyday skills]. Thieme. URL: https://www.thieme-connect.de/products/ejournals/pdf/10.1055/s-0032-1307959.pdf [accessed 2024-08-27]
27. Heuschmann PU, Kolominsky-Rabas PL, Nolte CH, et al. Untersuchung der reliabilität der deutschen version des Barthel-Index sowie entwicklung einer postalischen und telefonischen fassung für den einsatz bei schlaganfall-patienten [investigation of the reliability of the German version of the Barthel Index and development of a postal and telephone version for use with stroke patients]. Fortschr Neurol Psychiatr 2005 Feb;73(2):74-82. [doi: 10.1055/s-2004-830172]
28. Mahoney FI, Barthel DW. Functional evaluation: the Barthel index. Md State Med J 1965 Feb;14:61-65. [Medline: 14258950]
29. Kolb G, Breuninger K, Gronemeyer S, et al. 10 Jahre geriatrische frührehabilitative komplexbehandlung im DRG-system [10 years of early geriatric rehabilitation complex therapy in the DRG system]. Z Gerontol Geriat 2014 Jan;47(1):6-12. [doi: 10.1007/s00391-013-0592-7]
30. Aizen E, Lutsyk G, Wainer L, Carmeli S. Effectiveness of individualized fall prevention program in geriatric rehabilitation hospital setting: a cluster randomized trial. Aging Clin Exp Res 2015 Oct;27(5):681-688. [doi: 10.1007/s40520-015-0330-7] [Medline: 25697080]
31. Quiaios L, Krief H, Lamy O, da Rocha CG, Siegle-Authemayou ML, Lang PO. Impact of an early geriatric rehabilitation program in acutely hospitalised vulnerable patients: a real-life study in an ACE unit in Switzerland. Geriatr Psychol Neuropsychiatr Vieil 2023 Mar 1;21(1):51-62. [doi: 10.1684/pnv.2023.1089] [Medline: 37115679]

XSL•FO
RenderX

32. Mansell V, Hall Dykgraaf S, Kidd M, Goodyear-Smith F. Long COVID and older people. Lancet Healthy Longev 2022 Dec;3(12):e849-e854. [doi: 10.1016/S2666-7568(22)00245-8] [Medline: 36480981]
33. Lettow B, Wouters M, Sinnige J. E-health, wat is dat? Nictiz. URL: https://nictiz.nl/app/uploads/2022/08/E-health-Wat-is-dat.pdf [accessed 2024-08-28]
34. Scott AR, Alore EA, Naik AD, Berger DH, Suliburk JW. Mixed-methods analysis of factors impacting use of a postoperative mHealth app. JMIR Mhealth Uhealth 2017 Feb 8;5(2):e11. [doi: 10.2196/mhealth.6728] [Medline: 28179215]
35. Jensen CM, Overgaard S, Wiil UK, Smith AC, Clemensen J. Bridging the gap: a user-driven study on new ways to support self-care and empowerment for patients with hip fracture. SAGE Open Med 2018;6:2050312118799121. [doi: 10.1177/2050312118799121] [Medline: 30210796]
36. Ling Y, Ter Meer LP, Yumak Z, Veltkamp RC. Usability test of exercise games designed for rehabilitation of elderly patients after hip replacement surgery: pilot study. JMIR Serious Games 2017 Oct 12;5(4):e19. [doi: 10.2196/games.7969] [Medline: 29025696]
37. Li AC, Kannry JL, Kushniruk A, et al. Integrating usability testing and think-aloud protocol analysis with “near-live” clinical simulations in evaluating clinical decision support. Int J Med Inform 2012 Nov;81(11):761-772. [doi: 10.1016/j.ijmedinf.2012.02.009] [Medline: 22456088]
38. Berg K. Measuring balance in the elderly: preliminary development of an instrument. Physiother Can 1989 Nov;41(6):304-311. [doi: 10.3138/ptc.41.6.304]
39. Steffen TM, Hacker TA, Mollinger L. Age- and gender-related test performance in community-dwelling elderly people: six-minute walk test, berg balance scale, timed up & go test, and gait speeds. Phys Ther 2002 Feb;82(2):128-137. [doi: 10.1093/ptj/82.2.128] [Medline: 11856064]
40. Donoghue D, Physiotherapy Research and Older People (PROP) group, Stokes EK. How much change is true change? the minimum detectable change of the Berg Balance Scale in elderly people. J Rehabil Med 2009 Apr;41(5):343-346. [doi: 10.2340/16501977-0337] [Medline: 19363567]
41. Guralnik JM, Ferrucci L, Pieper CF, et al. Lower extremity function and subsequent disability: consistency across studies, predictive models, and value of gait speed alone compared with the short physical performance battery. J Gerontol A Biol Sci Med Sci 2000 Apr;55(4):M221-M231. [doi: 10.1093/gerona/55.4.m221] [Medline: 10811152]
42. Abel B, Bongartz M, Rapp K, et al. Multimodal home-based rehabilitation intervention after discharge from inpatient geriatric rehabilitation (GeRas): study protocol for a multicenter randomized controlled trial. BMC Geriatr 2024 Jan 17;24(1):69. [doi: 10.1186/s12877-023-04634-2] [Medline: 38233746]

Abbreviations

BBS: Berg Balance Scale
BI: Barthel Index
CG: control group
EGR: early geriatric rehabilitation
FIM: Functional Independence Measure
FPP: fall prevention program
IG: intervention group
OP: Otago program
PH-Q4: Four-Item Patient Health Questionnaire
QoL: quality of life
SPPB: short physical performance battery
SSE: square-stepping exercise
TUG: Timed Up and Go Test

Edited by A Mavragani; submitted 21.09.24; peer-reviewed by PE Roos; revised version received 21.11.24; accepted 09.12.24; published 11.02.25.

Please cite as:
Steinmetz C, Stenzel C, Sylvester M, Glage D, Linke A, Sadlonova M, von Arnim CAF, Schnieder M, Valentová M, Heinemann S
Use of a Technology-Based Fall Prevention Program With Visual Feedback in the Setting of Early Geriatric Rehabilitation: Controlled and Nonrandomized Study
JMIR Form Res 2025;9:e66692
URL: https://formative.jmir.org/2025/1/e66692
doi:10.2196/66692

XSL•FO
RenderX

© Carolin Steinmetz, Christina Stenzel, Maj Sylvester, Denis Glage, Anne Linke, Monika Sadlonova, Christine A F von Arnim, Marlena Schnieder, Miroslava Valentová, Stephanie Heinemann. Originally published in JMIR Formative Research (https://formative.jmir.org), 11.2.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Acceptance of Unsupervised App-Based Cognitive Assessment in Outpatient Care: An Implementation Study

Iris Blotenberg[1], Dr rer nat; Melanie Boekholt[1], MA; Nils Lieberknecht[2], MSc; Paula Säring[2], Dr rer nat; Jochen René Thyrian[1,3], Prof Dr

[1]German Center for Neurodegenerative Diseases (DZNE), Interventional Health Care Research, Ellernholzstr. 1-2, Greifswald, Germany

[2]neotiv GmbH, Magdeburg, Germany

[3]Institute for Community Medicine, University Medicine Greifswald, Greifswald, Germany

Corresponding Author:
Iris Blotenberg, Dr rer nat
German Center for Neurodegenerative Diseases (DZNE), Interventional Health Care Research, Ellernholzstr. 1-2, Greifswald, Germany

Abstract

Background: The use of unsupervised digital cognitive assessments provides considerable opportunities for early and comprehensive testing for Alzheimer disease, minimizing the demand on time and personnel resources in medical practices. However, the acceptance within health care has yet to be assessed.

Objective: In this implementation study, the acceptance of an app-based, repeated cognitive assessment for early symptoms of Alzheimer disease in the outpatient care setting from both physicians' and patients' perspectives was examined.

Methods: In total, 15 primary care practices participated, where patients with self- or relative-reported memory problems could be prescribed an app (neotivCare app [neotiv GmbH]) for comprehensive cognitive testing. Patients used the app to test their episodic memory function weekly for 12 weeks at home. After the testing period and the final consultation, physicians and patients received questionnaires to assess the app's acceptance.

Results: We received completed questionnaires from physicians for 45 patients. In addition, we received 45 completed questionnaires from the patients themselves. The physicians reported that, for most patients, the app supported their decision-making in the diagnostic process (26/45, 58%). In addition, most physicians found the app's information dependable (34/45, 76%) and felt more certain in their decisions (38/45, 84%). From the patients' perspective, a majority felt thoroughly tested (34/45, 76%), and only a few considered the time commitment for the cognitive tests to be too burdensome (7/45, 16%). Furthermore, despite the weekly cognitive testing and the lengthy 12-week testing period, a majority of patients participated in all tests (39/54, 72%).

Conclusions: Our results indicate a high level of acceptance by physicians and patients, suggesting significant potential for the implementation of unsupervised digital cognitive assessments into routine health care. In the future, acceptance should be assessed in large-scale studies, with a particular focus on the impact on health care delivery and patient outcomes.

(JMIR Form Res 2025;9:e62706) doi:10.2196/62706

KEYWORDS

mild cognitive impairment; Alzheimer disease; dementia; cognition; computerized assessment; digital assessment; digital cognitive biomarkers; home-based assessment; digital platform; mobile phone

Introduction

In many countries, the number of people with Alzheimer disease and related dementias is rapidly increasing; today, more than 50 million people are living with dementia, and this number is expected to triple by 2050 [1,2]. Timely diagnosis of Alzheimer disease is a key objective, as interventions and novel pharmacological treatment approaches shift to the disease's early stages [3,4]. It is the primary care physicians who often have the closest contact with their patients and who serve as the patient's initial contact point when they seek clarification for perceived changes in cognitive abilities. However, conventional neuropsychological screening tests lack sensitivity for early stages of Alzheimer disease [5]. The drawback of more comprehensive cognitive test batteries is that these tests are time-consuming and that they need to be administered by specialized personnel, thus they are quite resource-intensive [6]. This makes it challenging to integrate early diagnosis into primary care and outpatient practices, impacting access, especially in rural areas where the nearest memory clinic is often more distant. Another disadvantage is the 1-time testing in the setting of a medical practice, which has been shown to be influenced by daily variations (eg, lack of sleep) [7] and even

XSL•FO
RenderX

the stereotype threat effect [8], reducing reliability and validity of the assessment.

A promising solution lies in digital cognitive assessments at home, which can save time and resources. Furthermore, they offer the possibility for repeated testing, reducing the impact of daily variations and allowing for the assessment of symptom progression over time. In recent years, several mobile apps for cognitive testing for early Alzheimer symptoms have been developed [9,10]. Among them is the neotivCare app (neotiv GmbH), which has been developed based on current insights into the functional anatomy of episodic memory [11-13]. The cognitive tests have already been assessed for their psychometric quality and feasibility in cognitively healthy older adults and in a memory clinic sample [6,14,15]. For other unsupervised mobile apps, feasibility and acceptance have been explored as well [e.g. 16,17].

What is currently lacking and of particular importance is testing the acceptance in the actual health care setting, for which this and other apps have been developed. In this study, participating physicians in outpatient care had the opportunity to prescribe the neotivCare app to patients who consulted them for memory problems. After the patients had used the app for a duration of 3 months, and once the physicians had received and analyzed the test results, the physicians then evaluated the usefulness of the app in the diagnostic process, and the patients provided their feedback on their experience with the app-based testing.

The objective of this study was to evaluate the acceptance of unsupervised app-based cognitive testing within the realm of care from both the physicians' and the patients' perspective. This study adheres to the Guidelines and Checklist for the Reporting on Digital Health Implementations (iCHECK-DH) [18].

Methods

Study Procedure and Questionnaires

The project took place in collaboration between the physicians' network in a neighbourhood in the north-east of Germany (Magdeburg-Schönebeck), a major statutory health insurance in Germany (AOK Saxony-Anhalt), and neotiv GmbH. The study was conducted to test improved diagnostic methods for dementia care according to §140a "Special Care" of the German Social Code Book V. The study was conducted between September 2021 and December 2022 in Magdeburg-Schönebeck. This period coincided with social distancing measures due to the SARS-CoV-2 pandemic.

Participating practices were recruited from the members of the physicians' network Magdeburg-Schönebeck. Patients consulting their physician for memory problems and who fulfilled further inclusion criteria were informed about the opportunity to test their cognition with the neotivCare app for 3 months. Inclusion criteria were self- or relative-reported memory problems persisting for at least 6 months and perceived as progressive, as well as owning a smartphone and being able to use it. In addition, patients had to be insured with a specific major statutory health insurance in Germany (AOK Saxony-Anhalt). The exclusion criterion was a clear indication of dementia, as it requires timely optimal care. After providing informed consent, patients received an activation code to install and use the app. The testing period spanned 12 weeks with weekly assessments. An automated report was generated from the test results and made available to the treating physician.

After the physician and patient had discussed the results, both the physician and the patient received a short questionnaire, the items are displayed in Table 1. The physician answered, for each patient individually, (1) to what extent the test results from the app had supported decision-making in the diagnostic process, (2) how they assessed the dependability of the app information for deciding on the further diagnostic process, and (3) how certain they were, based on the app information, in having made the correct decision about the subsequent process.

Patients completed a questionnaire providing sociodemographic information (age, sex, and educational attainment). In addition, they answered (1) to what extent they felt their cognitive abilities were thoroughly assessed through the app-based assessment and (2) to what extent they felt the assessments were too time-consuming for them.

XSL•FO
RenderX

Table . Description of the items answered by physicians and patients regarding app-based cognitive testing for early symptoms of Alzheimer disease in this implementation study in outpatient care.

Item		Response options
Physician questionnaire		
	How much did the app information provided about the patient support your decision regarding the further diagnostic process?	• Strongly • Rather strongly • Rather limited • Not at all
	How do you assess the dependability of the app information provided for deciding on the further diagnostic process?	• Dependable • Rather dependable • Rather undependable • Undependable
	How certain were you that, based on the app information provided about the patient, you made the right decision for the further process?	• Certain • Rather certain • Rather uncertain • Uncertain
Patient questionnaire		
	The regular memory assessments with the neotivCare app were, in terms of timing, for me:	• Very burdensome • Quite burdensome • Minimally burdensome • Not burdensome
	Through the regular memory assessments with the neotivCare app, I felt that I was thoroughly examined.	• Fully applies • Applies more • Applies less • Does not apply

Ethical Considerations

Before the commencement of the study, the study was first ethically reviewed and approved by the ethical board of the physicians' network Netz Magdeburg Schönebeck (Network Magdeburg Schönebeck), an association of resident physicians that participated in this study. Ethics approval was granted (no reference number) and a written statement confirming the approval is available. However, this statement is dated after the study's commencement, as the authors requested documentation of the prior ethical review. The patients were thoroughly informed about the study by their physician and through a patient information sheet. They provided informed consent through a participation declaration (Multimedia Appendix 1) before receiving an activation code to install and access the neotivCare app. The study reports results based on returned questionnaires regarding app evaluation; all questionnaires were anonymous. Linkage between the questionnaires, with medical records or with the results from the neotivCare app, was not possible. The participating physicians were compensated by the collaborating health insurance for their additional efforts, such as informing and enrolling patients in the study. The patients did not receive any compensation for their participation in the app-based cognitive testing. The results, figures, or supplements reported in the manuscript do not allow any identification of individual study participants.

Sample

Medical practices

A total of 19 physicians from 15 medical practices participated in the study, including 2 (11%) specializing in internal medicine and 2 (11%) in neurology, with the remaining participating physicians being general practitioners (15/19, 79%). Physicians provided anonymized questionnaire responses on their acceptance of the app in the diagnostic process for 45 patients.

Patients

A total of 54 patients were included in the study. Of these, 45 individuals completed a questionnaire on sociodemographic factors and app usage. Table 2 presents the descriptive statistics for the study participants. The average age was 66 (SD 7.2, age range 52-80) years, 56% (25/45) were female. The majority had a moderate level of education (23/45, 51%).

XSL•FO
RenderX

Table . Description of the sociodemographic characteristics of the patients who participated in app-based cognitive testing for early symptoms of Alzheimer disease in this implementation study in outpatient care.

Variable		Value (N=45)
Age (years), mean (SD)		66.4 (7.2)
Sex, n (%)		
	Female	25 (56)
Education, n (%)		
	Low (<10 years of schooling)	11 (24)
	Medium (10 - 11 years of schooling)	23 (51)
	High (≥12 years of schooling)	9 (20)
	Did not disclose	2 (4)

The neotivCare App

Cognitive testing using the neotivCare app was conducted remotely, without supervision, and in a home setting. The patients used their own mobile phone or tablet. The tests were administered weekly over a period of 12 weeks. To remind patients to participate in the weekly test, they received a push notification when a test was available. If a patient did not start the test, they were reminded up to 5 more times via push notification (once a day for 5 days). To boost motivation for regular participation, study teams were trained to inform patients at enrollment that correct test execution and adherence to the study schedule are essential for generating valid data about their cognitive status.

A total of 3 different tasks were presented to assess episodic memory function. Each task was presented 4 times, resulting in the repetition of each specific task every 3 weeks. In the first task, patients were asked to indicate whether an image was an identical repetition or a modified version of a previous image and tap on the location of change (Mnemonic discrimination for objects and scenes, Figure 1A). In the second task, participants had to memorize objects and their positions within a room. Subsequently, they were shown an empty room with a blue circle highlighting the previous position of one of the objects and they had to select the correct object from a set of options (Object-in-room recall, Figure 1B). In the third task, patients were shown photographic images and asked to determine whether they depicted indoor or outdoor scenes. After 60 - 70 minutes, images were presented again, and patients had to decide whether they were new or familiar (Complex scene recognition, Figure 1C).

XSL•FO
RenderX

Figure 1. Illustration of the tasks used in app-based cognitive testing for early symptoms of Alzheimer disease in this implementation study in outpatient care (used with permission from neotiv GmbH).

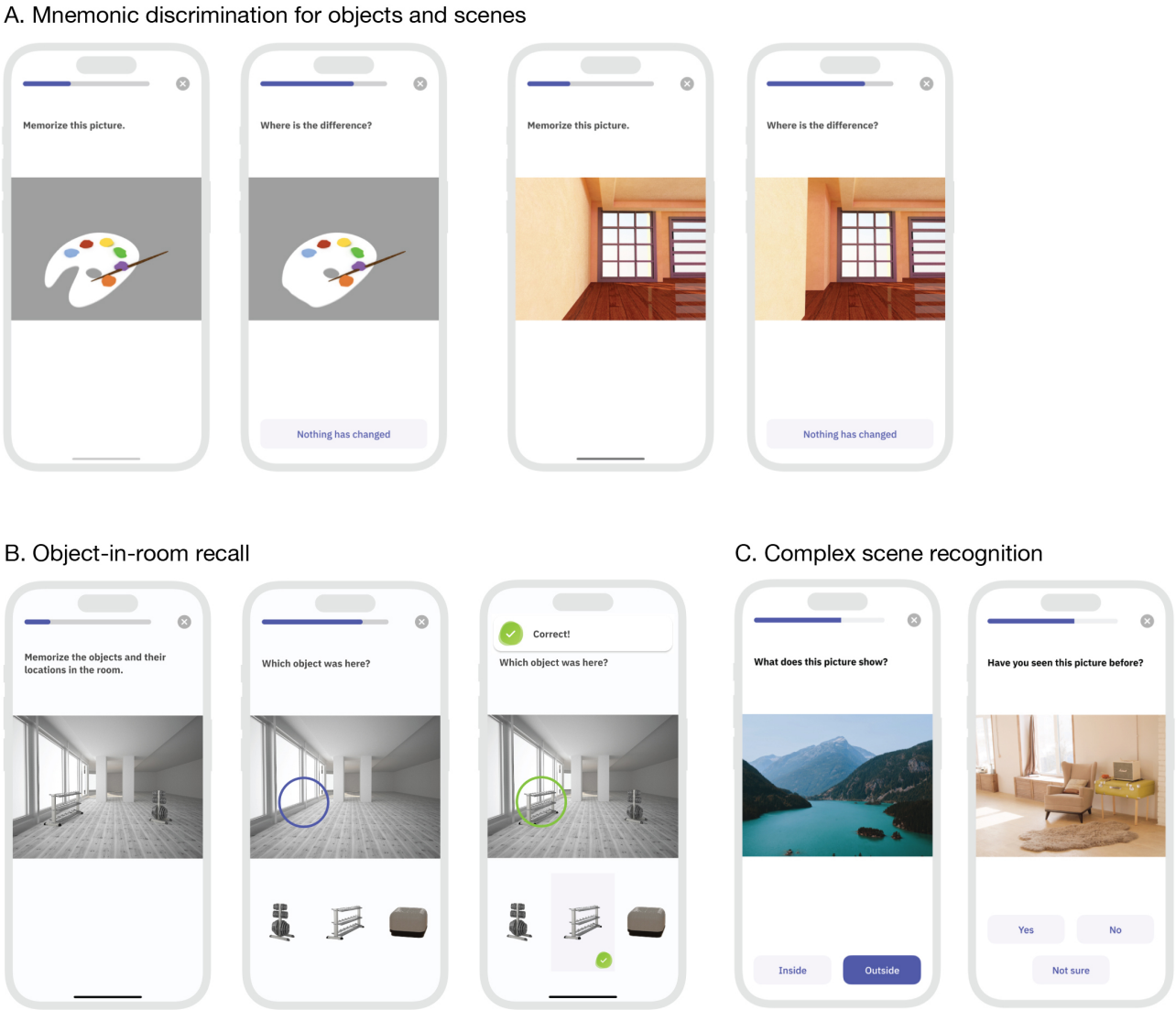

Statistical Analysis

The processing and analysis of the data were outsourced as an external contract and conducted by an independent research group. The analysis of responses to the items was done descriptively (relative frequencies).

Comparison With Previous Work

A literature search was conducted to identify studies on the feasibility and acceptance of unsupervised digital cognitive assessments in health care and to compare them with our findings.

Results

Acceptance Among Outpatient Physicians

Figure 2 shows the distribution of responses from the participating physicians. For the majority of patients, the physicians indicated that the information from the app supported their decision-making in the diagnostic process (strongly: 12/45, 27%; rather strongly: 14/45, 31%), with only a few cases where the app was perceived by physicians as not supporting the diagnostic process at all (4/45, 9%).

For a clear majority of patients, the physicians assessed the app information as dependable for making decisions about the further diagnostic process (dependable: 14/45, 31%; rather dependable: 20/45, 44%), with only a few patients where the app was considered somewhat (6/45, 13%) or not dependable at all (2/45, 4%).

In the clear majority of cases, the treating physicians reported feeling certain (23/45, 51%) or rather certain (15/45, 33%) in having made the correct decision for the further diagnostic process based on the app information. Only in a small number of cases did the physicians indicate that they were somewhat uncertain (4/45, 9%) or uncertain (1/45, 2%).

XSL•FO
RenderX

Figure 2. Distribution of responses from the physicians concerning acceptance of app-based cognitive testing for early symptoms of Alzheimer disease in this implementation study in outpatient care (data based on questionnaires from the 15 participating practices regarding 45 patients). N/A: not available.

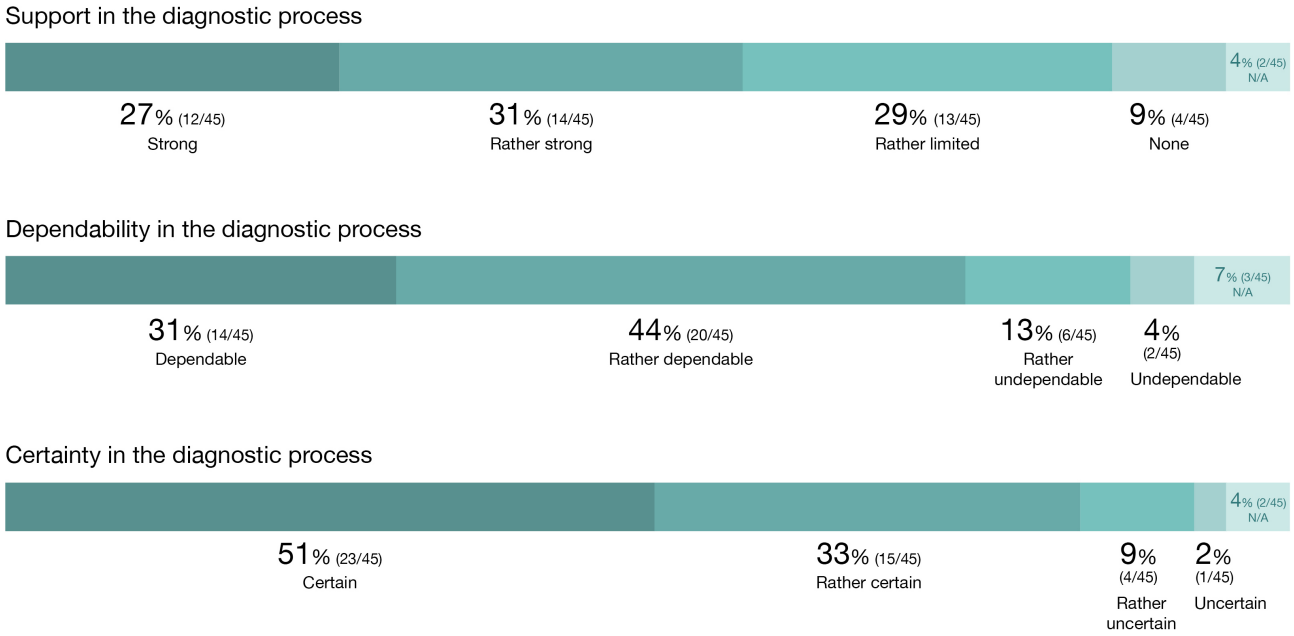

Acceptance Among Patients

Figure 3 shows the distribution of responses from the participating patients. Of the patients who had started participating in the study, 72% (39/54) completed all weekly tests over 12 weeks, indicating high adherence throughout the study period.

When asked about the extent to which using the app was time-consuming, a clear majority responded that they felt either not at all (12/45, 27%) or only slightly (24/45, 53%) time-burdened by using the app. Feeling quite or very burdened was reported by 13% (6/45) and 2% (1/45) of the patients.

A large majority of patients stated that they felt thoroughly examined through the app; 38% (17/45) agreed and another 38% (17/45) somewhat agreed. Slightly more than one-fifth disagreed to some extent (8/45, 18%) or completely (1/45, 2%).

Figure 3. Distribution of responses from the patients concerning acceptance of app-based cognitive testing for early symptoms of Alzheimer disease in this implementation study in outpatient care (data based on responses from 45 of the 54 included patients). N/A: not available.

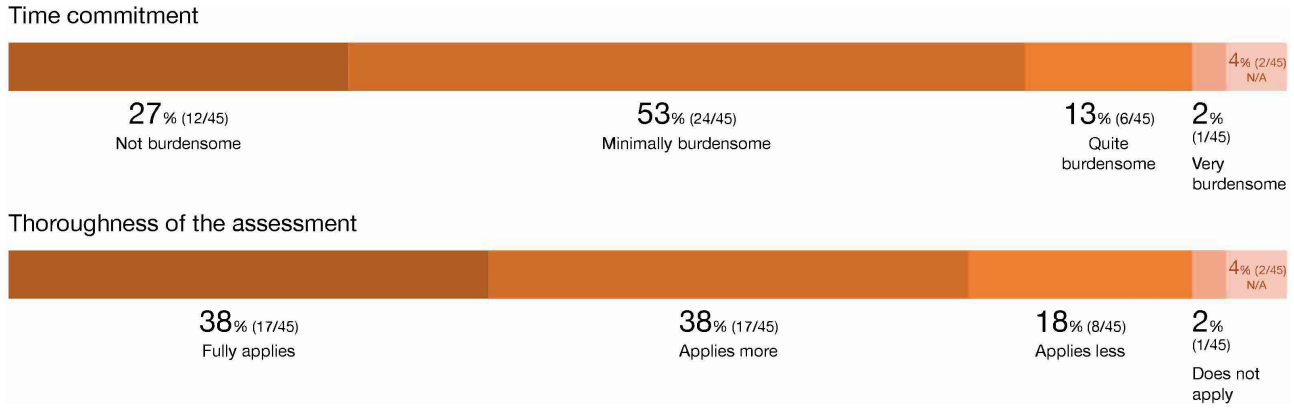

Discussion

Principal Results

In this implementation study of unsupervised app-based cognitive assessment in outpatient care, both physicians and patients showed high levels of acceptance. The physicians reported, for the majority of patients, that the app had supported the diagnostic process. The majority of patients found the repeated assessments to be only minimally burdensome and reported feeling thoroughly assessed. Most patients participated in all the weekly assessments over the 12-week testing period.

Comparison With Previous Work

Previous studies have reported high feasibility for unsupervised digital cognitive assessments [9]. Several studies have documented high consent rates during recruitment; however, these were from ongoing studies, with the majority of participants approached agreeing to additional digital cognitive testing [19-21]. Adherence rates in previous studies were also high, with most participants completing multiple assessments, ranging from several days [22] to several months [23]). Unsupervised digital cognitive assessments have been tested both in the general population [14] and among older adults [6].

XSL•FO
RenderX

What is currently lacking and urgently needed are studies that test unsupervised digital cognitive assessments within the realm of health care [24,25]. Particularly important in this context is the question of acceptance by physicians and patients, as physicians are responsible for prescribing, interpreting, and acting on these assessments, while patients are required to thoroughly complete the cognitive tests. To our knowledge, there are no published results on this topic and this study is the first to report on the acceptance of such assessments by physicians and patients.

Physician and Patient Perspectives Regarding App-Based Cognitive Assessment

Physicians in our study indicated that, for most patients, the app supported them in making decisions about the diagnostic process. Furthermore, they indicated that they perceived the app information as dependable and that they felt more certain in their decisions. These initial findings suggest that participating physicians have shown acceptance toward the app for evaluating cognitive performance and were using the results to inform their next steps. This is promising, as acceptance from outpatient care physicians is crucial for integrating such tools into routine care and enhancing early testing for Alzheimer disease, as they often serve as the first point of contact for individuals experiencing memory problems.

The participating patients also gave largely positive feedback on the app; a majority felt thoroughly tested, and only a few considered the time commitment for the cognitive tests as too burdensome. Furthermore, despite weekly cognitive testing and the lengthy testing period of 12 weeks, a large majority of patients participated in all tests, indicating high adherence. Reasons for this likely included the generally high motivation of patients in the context of assessing a potential Alzheimer disease diagnosis. In addition, the study design and app programming were aimed at enhancing adherence—study teams were trained to inform patients at enrollment about the necessity of correctly performing tests and adhering to the study plan to generate valid data for the diagnostic report. Furthermore, the app sent weekly push notifications (and reminders) when a new test was available.

These initial results indicate a high level of acceptance and adherence from the patients. This is an important finding for the implementation of app-based cognitive assessment into routine health care, given that it is ultimately the patients who need to integrate the tests into their daily lives and perform them without supervision over an extended time period.

Limitations

To ensure data protection, it was not possible to link the physicians' and patients' questionnaires with each other, nor was it possible to link them with results from the app or additional patient data. The results presented here are solely derived from the returned questionnaires, which were anonymized and provided to participating physicians and patients. The overall sample size was relatively small, and the inclusion criteria of the study may have introduced a potential selection bias in the recruitment process. For example, only patients who owned and could use a smartphone or tablet were included. However, currently, a majority of people aged 55-65 years already own a smartphone, and this number is expected to rise in the future [26]. Furthermore, only patients insured with a specific major public health insurance (AOK Saxony-Anhalt) were eligible to participate. However, as the largest statutory health insurance in Saxony-Anhalt, it covers a cross-section of the population.

Conclusion

The high acceptance and adherence observed in our study is a promising finding for the implementation of unsupervised digital cognitive assessments within the realm of care. Integrating these assessments into routine health care has the potential to enhance early detection of Alzheimer disease, improve access to diagnostics in rural areas, and reduce the burden on medical practices. Future research should focus on larger-scale studies to confirm the high acceptance rates and evaluate the benefits of unsupervised digital cognitive assessments in health care delivery and their impact on patient outcomes.

Acknowledgments

We thank the participants of the study, patients, and physicians in the eastern German area of Magdeburg-Schönebeck. The project was conducted in collaboration of 3 parties: AOK Saxony-Anhalt (a major statutory health insurance in Germany), the physicians' network in Magdeburg-Schönebeck, and neotiv GmbH. The study was funded by AOK Saxony-Anhalt and neotiv GmbH. MB received funding from the Hans and Ilse Breuer Foundation.

Data Availability

The datasets generated or analyzed during this study are not publicly available since the study was conducted under a contract with a major statutory health insurance (AOK Saxony-Anhalt) in accordance with §140a German Social Code Book V (SGB V), which also provided the funding.

Authors' Contributions

IB, JRT, MB, NL, and PS designed the current study. IB and MB analyzed the data. IB, MB, and JRT interpreted the data. IB and JRT wrote the manuscript. All authors critically reviewed the manuscript, provided feedback, and approved it for submission.

XSL•FO
RenderX

Conflicts of Interest

IB, MB, and JRT declare no conflicts of interest. NL and PS are employees of neotiv GmbH; they were involved in conducting the study including data collection, but did not influence preparation, analysis and interpretation of data, which was carried out by IB, MB, and JRT.

Multimedia Appendix 1
Participant declaration.
[PDF File, 147 KB - formative_v9i1e62706_app1.pdf]

References

1. GBD 2019 Dementia Forecasting Collaborators. Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the Global Burden of Disease Study 2019. Lancet Public Health 2022 Feb;7(2):e105-e125. [doi: 10.1016/S2468-2667(21)00249-8] [Medline: 34998485]
2. Blotenberg I, Hoffmann W, Thyrian JR. Dementia in Germany: epidemiology and prevention potential. Dtsch Arztebl Int 2023 Jul 10;120(27-28):470-476. [doi: 10.3238/arztebl.m2023.0100] [Medline: 37226316]
3. National plan to address Alzheimer's disease: 2022 update. Assistant Secretary for Planning and Evaluation. 2022. URL: https://aspe.hhs.gov/reports/national-plan-2022-update#main-content [accessed 2025-01-30]
4. Jessen F, Bohr L, Kruse C, Dodel R. Die S3-Leitlinien Demenzen [Article in German]. Nervenarzt 2023 Jul;94(7):609-613. [doi: 10.1007/s00115-023-01492-6]
5. Tsoi KKF, Chan JYC, Hirai HW, Wong SYS, Kwok TCY. Cognitive tests to detect dementia: a systematic review and meta-analysis. JAMA Intern Med 2015 Sep;175(9):1450-1458. [doi: 10.1001/jamainternmed.2015.2152] [Medline: 26052687]
6. Öhman F, Berron D, Papp KV, et al. Unsupervised mobile app-based cognitive testing in a population-based study of older adults born 1944. Front Digit Health 2022;4:933265. [doi: 10.3389/fdgth.2022.933265] [Medline: 36426215]
7. Sliwinski MJ, Mogle JA, Hyun J, Munoz E, Smyth JM, Lipton RB. Reliability and validity of ambulatory cognitive assessments. Assessment 2018 Jan;25(1):14-30. [doi: 10.1177/1073191116643164] [Medline: 27084835]
8. Fresson M, Dardenne B, Geurten M, Meulemans T. The effect of stereotype threat on older people's clinical cognitive outcomes: investigating the moderating role of dementia worry. Clin Neuropsychol 2017 Nov;31(8):1306-1328. [doi: 10.1080/13854046.2017.1307456] [Medline: 28349725]
9. Polk SE, Öhman F, Hassenstab J, et al. A scoping review of remote and unsupervised digital cognitive assessments in preclinical Alzheimer's disease. medRxiv 2024 Sep 27. [doi: 10.1101/2024.09.25.24314349] [Medline: 39399008]
10. Öhman F, Hassenstab J, Berron D, Schöll M, Papp KV. Current advances in digital cognitive assessment for preclinical Alzheimer's disease. Alzheimers Dement (Amst) 2021;13(1):e12217. [doi: 10.1002/dad2.12217] [Medline: 34295959]
11. Düzel E, Berron D, Schütze H, et al. CSF total tau levels are associated with hippocampal novelty irrespective of hippocampal volume. Alzheimers Dement (Amst) 2018;10(1):782-790. [doi: 10.1016/j.dadm.2018.10.003] [Medline: 30555890]
12. Maass A, Berron D, Harrison TM, et al. Alzheimer's pathology targets distinct memory networks in the ageing brain. Brain (Bacau) 2019 Aug 1;142(8):2492-2509. [doi: 10.1093/brain/awz154] [Medline: 31199481]
13. Berron D, Cardenas-Blanco A, Bittner D, et al. Higher CSF tau levels are related to hippocampal hyperactivity and object mnemonic discrimination in older adults. J Neurosci 2019 Oct 30;39(44):8788-8797. [doi: 10.1523/JNEUROSCI.1279-19.2019] [Medline: 31541019]
14. Berron D, Ziegler G, Vieweg P, et al. Feasibility of digital memory assessments in an unsupervised and remote study setting. Front Digit Health 2022;4:892997. [doi: 10.3389/fdgth.2022.892997] [Medline: 35721797]
15. Berron D, Glanz W, Clark L, et al. A remote digital memory composite to detect cognitive impairment in memory clinic samples in unsupervised settings using mobile devices. NPJ Digit Med 2024 Mar 26;7(1):79. [doi: 10.1038/s41746-024-00999-9] [Medline: 38532080]
16. Lancaster C, Koychev I, Blane J, Chinner A, Wolters L, Hinds C. Evaluating the feasibility of frequent cognitive assessment using the Mezurio smartphone app: observational and interview study in adults with elevated dementia risk. JMIR Mhealth Uhealth 2020 Apr 2;8(4):e16142. [doi: 10.2196/16142] [Medline: 32238339]
17. Thompson LI, Harrington KD, Roque N, et al. A highly feasible, reliable, and fully remote protocol for mobile app-based cognitive assessment in cognitively healthy older adults. Alzheimers Dement (Amst) 2022;14(1):e12283. [doi: 10.1002/dad2.12283] [Medline: 35415201]
18. Perrin Franck C, Babington-Ashaye A, Dietrich D, et al. iCHECK-DH: guidelines and checklist for the reporting on digital health implementations. J Med Internet Res 2023 May 10;25:e46694. [doi: 10.2196/46694] [Medline: 37163336]
19. Samaroo A, Amariglio RE, Burnham S, et al. Diminished Learning Over Repeated Exposures (LORE) in preclinical Alzheimer's disease. Alzheimers Dement (Amst) 2020;12(1):e12132. [doi: 10.1002/dad2.12132] [Medline: 33426266]
20. Berron D, Olsson E, Andersson F, et al. Remote and unsupervised digital memory assessments can reliably detect cognitive impairment in Alzheimer's disease. Alzheimers Dement 2024 Jul;20(7):4775-4791. [doi: 10.1002/alz.13919] [Medline: 38867417]

XSL•FO
RenderX

21. Nicosia J, Aschenbrenner AJ, Adams SL, et al. Bridging the technological divide: stigmas and challenges with technology in digital brain health studies of older adults. Front Digit Health 2022;4:880055. [doi: 10.3389/fdgth.2022.880055] [Medline: 35574256]
22. Nicosia J, Aschenbrenner AJ, Balota DA, et al. Unsupervised high-frequency smartphone-based cognitive assessments are reliable, valid, and feasible in older adults at risk for Alzheimer's disease. J Int Neuropsychol Soc 2023 Jun;29(5):459-471. [doi: 10.1017/S135561772200042X] [Medline: 36062528]
23. Jutten RJ, Rentz DM, Fu JF, et al. Monthly at-home computerized cognitive testing to detect diminished practice effects in preclinical Alzheimer's disease. Front Aging Neurosci 2021;13:800126. [doi: 10.3389/fnagi.2021.800126] [Medline: 35095476]
24. Perumal TM, Wolfer A, Lipsmeier F, et al. Quantification of cognitive impairments in preclinical and early Alzheimer's disease: a proof of concept study to investigate the feasibility, adherence and preliminary clinical validity of remote smartphone based self-assessments of cognition, function and behavior (P14-3.002). Neurology (ECronicon) 2022 May 3;98(18_supplement). [doi: 10.1212/WNL.98.18_supplement.1574]
25. Duzel E, Schöttler M, Sommer H, Griebe M. Protocol: prospective evaluation of feasibility, added value and satisfaction of remote digital self-assessment for mild cognitive impairment in routine care with the neotivCare app. BMJ Open 2024 Mar 12;14(3):e081159. [doi: 10.1136/bmjopen-2023-081159] [Medline: 38479732]
26. Berenguer A, Goncalves J, Hosio S, Ferreira D, Anagnostopoulos T, Kostakos V. Are smartphones ubiquitous? An in-depth survey of smartphone adoption by seniors. IEEE Consumer Electron Mag 2016;6(1):104-110. [doi: 10.1109/MCE.2016.2614524]

Abbreviations

iCHECK-DH: Guidelines and Checklist for the Reporting on Digital Health Implementations

Edited by A Mavragani; submitted 30.05.24; peer-reviewed by JL Agne, M Butryn; revised version received 29.11.24; accepted 01.12.24; published 13.02.25.

Please cite as:
Blotenberg I, Boekholt M, Lieberknecht N, Säring P, Thyrian JR
Acceptance of Unsupervised App-Based Cognitive Assessment in Outpatient Care: An Implementation Study
JMIR Form Res 2025;9:e62706
URL: https://formative.jmir.org/2025/1/e62706
doi:10.2196/62706

© Iris Blotenberg, Melanie Boekholt, Nils Lieberknecht, Paula Säring, Jochen René Thyrian. Originally published in JMIR Formative Research (https://formative.jmir.org), 13.2.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Users' Perspectives of Direct-to-Consumer Telemedicine Services: Survey Study

Kate Churruca[1], PhD; Darran Foo[1,2], MD; Andrew Turner[1], BClinSci; Emily Crameri[1], MD; Maree Saba[1], BClinSci; Samantha Spanos[1], PhD; Matthew Vickers[3], MBBS; Jeffrey Braithwaite[1], PhD; Louise A Ellis[1], PhD

[1]Australian Institute of Health Innovation, Macquarie University, Macquarie Park, Australia
[2]Healthdirect Australia Limited, Sydney, Australia
[3]Eucalyptus Pty Ltd

Corresponding Author:
Kate Churruca, PhD
Australian Institute of Health Innovation, Macquarie University, Macquarie Park, Australia

Abstract

Background: Commercially run direct-to-consumer (DTC) telemedicine services are on the rise in countries such as Australia and the United States. These include DTC services that are web-based, largely asynchronous, and offer targeted treatment pathways for specific health issues (eg, weight loss or sexual function). It has been argued that DTC telemedicine improves access to health care and promotes patient empowerment. Despite research examining quality and safety issues, little is known about users' reasons for accessing DTC telemedicine services or their perceptions of them.

Objective: In this study, we aimed to examine the perspectives of Australian users accessing DTC telemedicine services, including the reasons for use, perceived benefits, and concerns, in addition to their usage and interaction with traditional general practice services.

Methods: A web-based cross-sectional survey including questions on demographics, published and validated scales, and author-developed closed- and open-response questions was administered via REDCap in 2023 to Australian adults accessing DTC telemedicine services.

Results: Among the 151 respondents, most (136/151, 90.1%) had seen a general practitioner (GP) in the previous 12 months and were somewhat or very satisfied (118/136, 86.8%) with the care, just over half found it easy to get an appointment with their GP (76/151, 50.3%), and a quarter found it difficult (38/151, 25.2%). Among the 136 respondents who had seen a GP, more than half either "never" (55/136, 40.4%) or "rarely" (23/136, 16.9%) discussed the information and treatment received from DTC telemedicine service with their GP. The majority of respondents were using a DTC telemedicine service offering prescription skin care (92/151, 60.9%), had received treatment in the previous 6 months (123/151, 81.5%), and had self-initiated care (128/151, 84.8%). The most frequently cited reasons for using DTC telemedicine were related to convenience (97/121, 80.2%) and flexibility (71/121, 58.7%), while approximately a third of the sample selected that it was difficult to see traditional health care provider in their preferred time frame (44/121, 36.4%) and that the use of DTC telemedicine allowed them to gain access to services otherwise unavailable through traditional health care (39/121, 32.2%). Most participants felt "more in control" (106/128, 82.9%) and "in charge" of their health concern (102/130, 78.5%) when using DTC telemedicine services, in addition to having "more correct knowledge" (92/128, 71.9%) and "feeling better informed as a patient" (94/131, 71.8%). "Costs of services" (40/115, 34.8%) and "privacy" (31/115, 27%) were the most frequently reported concerns with using digital health care technologies such as DTC telemedicine.

Conclusions: We report that most users perceive DTC telemedicine services as offering ease of access and convenience, and that their use contributes to a greater sense of empowerment over their health. Concerns were related to data privacy and the costs of utilizing the services. Responses suggest that DTC telemedicine may be tapping into a previously unmet need, rather than complementing traditional health care provided by a GP.

(JMIR Form Res 2025;9:e68619) doi:10.2196/68619

KEYWORDS

telemedicine; digital health technology; direct-to-consumer; digital tools; telehealth; consumer experience

XSL•FO
RenderX

Introduction

Telemedicine services involve the provision of health care, including the assessment of conditions and the delivery of treatments, through telephone and other digital channels [1,2]. In the past decade, direct-to-consumer (DTC) telemedicine services provided by private companies have become increasingly commonplace; prominent early examples include *Teladoc*, which had 6 million members in the United States by 2013 [3], and now operates in 16 countries including Australia [4]. Consumers using DTC services are able to initiate access to care through the internet, without referral, and often with no prior relationship with the consulting clinician, an approach also sometimes labeled as "commercial virtual visits" [5,6]. As a business model, users are either charged directly, or visits are paid for by contracts between the DTC telemedicine service and health care insurers [7].

The benefits of DTC telemedicine services include convenience and efficiency [8,9]. Patients have reported relatively high levels of satisfaction in using these services [10] and, particularly in the United States, these services have been suggested to increase access to health care for those who do not have insurance coverage [3,5]. By providing consumers with ready access to information, advice, diagnoses, and treatments, DTC telemedicine arguably contributes to patient empowerment through greater knowledge or a sense of control over their health [11,12]. However, issues have been raised about the lack of integration with traditional health care services [3,8] and concerns about the quality of care provided. For example, Schoenfeld et al [5] conducted an audit of 8 of the most widely used DTC companies in the United States, finding that complete history-taking occurred only in about 70% of visits and that wrong or no diagnosis was provided in almost a quarter, with considerable variability identified between conditions (eg, urinary tract infection or low back pain) and companies. The appropriate use of antibiotics for pediatric patients was also found to be lower in DTC telemedicine service, compared with urgent care and primary care providers [13]. However, other recent studies have suggested comparable quality and safety to traditional services in the rate of side effects from medication to treat erectile dysfunction [14] and in the management of respiratory infections [15].

In Australia, the prevalence of DTC telemedicine services has grown rapidly since the COVID-19 pandemic [16]. This includes a new model of DTC telemedicine service, one that is largely asynchronous, may be paid for by consumers on a subscription basis, and is organized around providing medication or treatments for a particular condition or type of issue (eg, men's health, sexual and reproductive health, skin conditions, and weight loss) [8,9,16]. In 2023, Foo and colleagues [16] identified more than 20 DTC telemedicine services in Australia that offer these "targeted treatment pathways for a specific issue or condition," including *Moshy* [17], which markets skin care and weight loss services to women, and *Pilot* [18], which provides a range of men-focused treatments for conditions including weight loss, hair and skin issues, and sexual health issues. As the delivery of health care via DTC telemedicine platforms frequently emphasizes discrete medical conditions, it precludes the development of sustained therapeutic relationships between clinicians and patients. This modality stands in marked contrast to conventional primary care practice, wherein the longitudinal physician-patient relationship and continuity of care constitute fundamental pillars of clinical excellence. Traditional primary care practitioners employ a holistic approach to patient management, rather than compartmentalizing health concerns into isolated conditions—a paradigm that enables comprehensive health oversight and integrated care delivery.

Despite the rise of DTC telemedicine services in Australia, and in countries such as the United States, the viewpoints of consumers have been underexamined [16,19], and this is particularly true for new models of DTC telemedicine that offer targeted treatment pathways for specific health issues. A recently published scoping review identified a clear gap in understanding user perspectives, including the reasons patients may choose to use these services [20]; little is known about the benefits they perceive, and any reservations they have in accessing this type of service [8]. This study aimed to examine the perspectives of consumers accessing DTC telemedicine for specific issues, including interactions with traditional health care services, the types of DTC telemedicine services used and reasons for using, and perceived benefits and concerns.

Methods

A web-based cross-sectional survey that included questions on demographics, published and validated scales, and author-developed closed- and open-response questions was administered to Australian adults accessing DTC telemedicine services.

Participants and Recruitment

Convenience sampling of users of DTC telemedicine services was used. Participants were recruited through an Australian digital health care company offering multiple DTC telemedicine services organized around targeted treatment pathways for specific conditions [16]. The company is a leading provider of DTC telemedicine services in Australia, with recent expansion internationally. At the time of data collection, the 5 different services it ran broadly dealt with the following areas: (1) Women's health care focusing largely on weight management and menopause, (2) Men's health care, (3) Skin care, (4) Women's fertility, and (5) Sexual well-being. The company moderates closed Facebook groups for users of these different services to interact with one another, share their journeys, and ask questions (medical advice is not permitted in the Facebook groups). An advertisement to take part in the survey was posted on the company's closed Facebook communities in November 2023 seeking the perspectives of Australian adults on their use of DTC services.

Materials

The web-based survey was conducted via REDCap hosted by Macquarie University. The survey included questions on traditional health care utilization, followed by a series of closed- and open-response questions about their use of DTC services, including reasons for using and their perceived benefits and concerns (see Multimedia Appendix 1). Finally, the survey

XSL•FO
RenderX

asked for demographic information. All questions were optional to complete.

To assess reasons for using DTC telemedicine services, respondents were asked a series of closed-response questions about experiences with traditional health care services, including whether they had seen a general practitioner (GP) over the preceding 12 months, reasons for accessing traditional health care, satisfaction with this care, and ease of getting an appointment. They were also asked about how they had come to initiate the use of the DTC telemedicine service (ie, recommended by a health care provider, recommended by a family member or friend, or self-initiated) and their main reason(s) for using the service from a list of possible reasons derived from a literature review [3,5-9,16], which included, for example, challenges with accessing traditional health care and convenience.

To assess the perceived benefits of DTC services, we adapted the empowering outcomes scales from van Uden-Kraan et al [12] All 4 items from "being better informed" were used (eg, "I feel better informed as a patient"). From their "increased optimism and control" subscale, 5 of the 8 original items were used (eg, "I feel more in charge of the course of my health concern/s"). Minor modifications were made to wording, such as changing "illness" to "health concern/s." Items for these subscales are answered on a 5-point Likert (1=completely disagree to 5=completely agree). To assess concerns about using digital health care technologies, participants were asked to select as many as apply from a range of options, derived from the literature [3,5-9,16], including safety and quality and the lack of regulation. Participants were also asked 2 open response questions, one about what they liked about using the DTC telemedicine service (ie, benefits), and one about their concerns with using digital health care technologies.

Analysis

Closed-ended data were analyzed using descriptive statistics in IBM SPSS Statistics (version 29; IBM Corp.) to summarize patterns in experiences with traditional health care services, DTC telemedicine services accessed, reasons for using, perceived benefits, and concerns with using digital health care technologies. For location, postcode data were mapped to the Australian Statistical Geography Standard and classified as major city, regional area, or remote area. For the 2 open response questions, inductive content analysis [21] was conducted by KC in Microsoft Excel. This involved initial open coding to write short summaries (codes) of meaning units within each response; this was followed by identifying recurrent meaning units across responses to identify categories. Multiple categories could be present in the response for a single participant and were summarized by frequency and with example extracts.

Ethical Considerations

The study received ethics approval from the Macquarie University Faculty of Medicine, Health and Human Sciences Low-Risk Ethics Subcommittee (No. 520221256844148). All participants opted in to participate and provided voluntary informed consent prior to completing the survey. The survey was designed to be anonymous, with no identifiable data collected. Those who took part in the survey had the option of going into the draw to win a $250 gift card.

Results

Responses were received from 204 people; Table 1 summarizes the characteristics of the participants. A total of 151 (74.0%) of these responses included sufficient information for inclusion in the analysis; 113 respondents completed the demographic information. Approximately two-thirds of those responding to demographic questions (75/113, 66.4%) were female, most were between the ages of 30 - 39 years (30/113, 26.8%) or 40 - 49 years (32/113, 28.6%) and most were employed in full-time paid work (62/113, 55.4%).

Most participants rated their health as "good" (36/115, 31.3%) or "very good" (38/115, 33.0%). Approximately half (58/115, 50.4%) reported having at least one chronic condition, such as a mental or behavioral health condition, back problems, or asthma.

The DTC telemedicine services most frequently used were those for prescription skincare (92/151, 60.9%) and the men's health care service (50/151, 33.1%); the women's health care (4/151, 2.6%), fertility (6/151, 4%), and sexual well-being services (4/151, 2.6%) were each used by only a small number of participants. Some participants (13/151, 8.6%) reported accessing multiple services. Most respondents (123/151, 81.5%) had been prescribed treatment within the previous 6 months.

XSL•FO
RenderX

Table . Demographic characteristics of direct-to-consumer telemedicine users responding to the survey.

Characteristics	N (%)[a]
Sex	
Female	75 (66.4)
Male	38 (33.6)
Age, years	
18 - 29	22 (19.6)
30 - 39	30 (26.8)
40 - 49	32 (28.6)
50 - 59	13 (11.6)
>60	15 (13.4)
Highest level of education	
Postgraduate degree	21 (18.3)
Graduate diploma/certificate	18 (15.7)
Bachelor degree	26 (22.6)
Certificate level	27 (23.5)
Secondary School (year 12 or less)	21 (18.3)
Other	2 (1.3)
Employment status	
Full-time paid work	62 (55.4)
Part-time paid work	34 (30.4)
Unemployed	6 (5.4)
Retired	4 (3.6)
Other	6 (5.4)
Location	
Major city	71 (64.5)
Regional area	35 (31.8)
Remote area	4 (3.6)

[a]Valid percentage used. Note. Responses may not equal 151 owing to missing data.

Experience With Traditional Health Care Services

Most participants (136/151, 90.1%) had seen a GP in the previous 12 months. Of the 15 respondents who had not, most (11/15, 73.3%) had not needed to; only 2 indicated it was too difficult to get an appointment, and 1 suggested it was too expensive. The most common reasons for seeing a GP were to get a prescription (67/136, 49.3%), management of an ongoing condition (64/136, 47.1%), a general or preventive health check-up (49/136, 35.0%), and for information/treatment of a new condition (46/136, 33.8%). A large majority of participants (118/136, 86.8%) were somewhat or very satisfied with the quality of care they received from their GP in the preceding 12 months. Approximately half found it easy or very easy to get an appointment with their GP (76/151, 50.3%), while a quarter found it difficult or very difficult (38/151, 25.2%), and the rest were neutral in their response. A majority of respondents either “never” (55/136, 40.4%) or “rarely” (23/136, 16.9%) discussed the information, advice, support, or treatment they received from the DTC telemedicine service with their GP; of the 136 respondents, 20 (14.7%) “sometimes” did and 22 (16.2%) either “often” or “always” talked to their GP.

Reasons for Using DTC Telemedicine Services

Over four-fifths of respondents (128/151, 84.8%) indicated that they had self-initiated use of the DTC telemedicine service, while 16 of 151 (10.6%) reported that they had received a recommendation from a friend or family member; only 2 of 151 (1.3%) responded that their health care provider had recommended using the service. Reasons for accessing DTC telemedicine services are displayed in Table 2. The most frequently cited reasons were “convenience in accessing information, support, and treatment at a time and place of my choosing” (97/121, 80.2%), “the flexibility and choice available online” (71/121, 58.7%), and “difficulty in seeing a traditional health care provider in the timeframe I would like (eg, appointments booked out weeks in advance or times not suitable for me)” (44/121, 36.4%). Thirty-nine participants among 121

XSL•FO
RenderX

(32.2%) selected one of their main reasons as "gaining access to services or medications not available to me through traditional healthcare services."

Table . Frequency of reported reasons for accessing direct-to-consumer (DTC) telemedicine services among users responding to the survey.

Reasons for accessing DTC telemedicine services	Number of respondents
Convenience in access at a time and place of my choosing	97
Flexibility and choice available online	71
Difficulty accessing traditional health care provider in the timeframe I would like	44
Gaining access to services/medications not available to me through traditional health care services	39
Costs associated with accessing traditional health care services (e.g., not eligible for Medicare, out-of-pocket fees)	39
Preference for accessing health care remotely at this time	31
Check whether my symptoms require attention by a traditional health care provider	16
Better understand the advice or treatment recommended by my traditional health care provider	10
Seeking a second opinion/verifying my traditional health care provider's advice	10
Other	6

Perceived Benefits of Using DTC Telemedicine Services

Figure 1 displays the frequency of participants' responses to items on "being better informed" and having "increased optimism and control" from using the DTC telemedicine service. Most participants agreed that the information, advice, and support they received from the service had made them feel more in control over what was happening to them (106/128, 82.9%) and feel more in charge of the course of their health concern/s (102/130, 78.5%). Most disagreed that the use of the DTC telemedicine service had led them to feel less in control over what was happening to them (98/128, 76.6%). The majority of respondents agreed that they had more correct knowledge at their disposal to manage their health concern/s (92/128, 71.9%) and felt better informed as a patient (94/131, 71.8%).

Seventy-four participants (74/151; 49%) provided an open response for what they liked about using the DTC telemedicine service(s). Eight categories of benefits were identified (Table 3), with "ease and convenience" (n=46) being the most frequently cited.

XSL•FO
RenderX

Figure 1. Perceptions of the impact of DTC telemedicine service use on feeling informed and having more control over health concern among users responding to the survey. * Item is negatively worded.

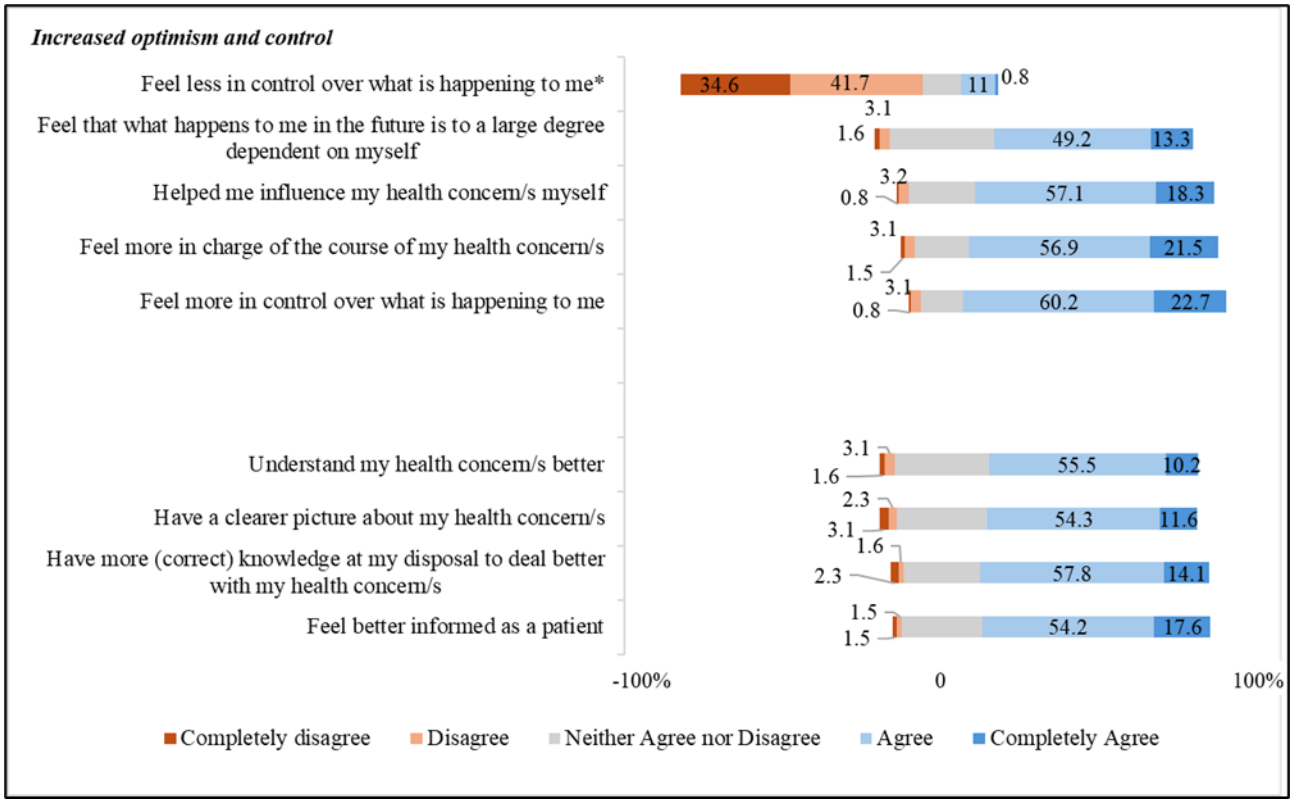

Table . Categories of benefits of DTC telemedicine services identified from open response of users responding to the survey.

Category	Count	Exemplar quote(s)
Ease and convenience	46	"Ease of use, convenience" (ID132) "Convenient access at a time of my choosing" (ID188)
Quality of service	12	"Quality professional advice" (ID86) "Great communication from the staff (doctors)" (ID206)
Efficiency	11	"it means I don't need to wait for an appointment to see my Dr" (ID119) "Ease of accessing treatments and to have it delivered with minimal downtime of face to face appointments" (ID81)
Benefits of the interface/platform	9	"It can be done from home and I am able to type messages in order to be better understood hopefully." (ID145)
Privacy	8	"Some face-to-face appointments can be daunting either because of the cost associated or the topic of conversation itself. I like that it's affordable, easily accessible and discreet" (ID110)
Cost savings	6	"It's expensive and time consuming seeing a doctor face-to-face. (Service) works, and with little effort" (ID146)
Enabling access to medication/treatment	4	"It is a reliable way to get custom skin care with medical grade actives to help with acne control." (ID197)
Greater control over health	3	"My ability to manage my needs myself to a degree." (ID148)

Concerns With Using Digital Health Care Technologies

Table 4 shows the frequency of participants' concerns with using digital health care technologies. The most frequently selected concern was "the costs associated with some of these services" (40/115, 34.8%), followed by "concerns over privacy and what the information I provide will be used for" (31/115, 27%).

XSL•FO
RenderX

Thirty-four respondents (34/151, 22.5%) provided open response answers for their concerns about using digital health care technologies, which were coded into 6 categories. The most frequent category of concern was related to "privacy and data security" (n=13). Table 5 summarizes the categories of concerns with using digital health care technologies.

Table . Frequency of reported concerns with using digital health care technologies among direct-to-consumer telemedicine users responding to the survey.

Concerns with using digital health care technologies among DTC telemedicine users	Number of respondents
Costs associated with services	40
Concerns over privacy and what the information I provide will be used for	31
Lack of knowledge of services	27
Lack of integration	26
Concerns about safety, quality, or accuracy of the advice and services	25
Lack of regulation	17
Other	6

Table . Categories of concerns with using digital health care technologies identified from open response of direct-to-consumer telemedicine users responding to the survey.

Category	Count	Exemplar quote(s)
Privacy and data security	13	"Concerned about my data leaking (because I am getting treatment for Erectile Dysfunction)" (ID178) "Data being collected and sold" (ID38)
Lack of face-to-face interaction	6	"Them not being able to physically assess me" (ID71)
Expensive	6	"More so the costing as it can get expensive" (ID46)
Need to share with GP	3	"If I had a problem with treatment I would be best to see my GP about it" (ID36)
Lack of regulation	3	"Might be a scam" (ID164)
Quality issues	2	"Authenticity of the providers and the information given" (ID181)
Other	5	For example, difficulty accessing medication, digital literacy.

Discussion

Principal Results

This study of Australian users of new DTC telemedicine services, organized around specific health issues (eg, weight loss and skin concerns), demonstrated that users largely seek out these services of their own volition for reasons of convenience and flexibility. As a result of using these services, most users feel they are better informed and have greater control over their health concerns; they perceive further benefits in the ease of use and, to a lesser extent, the quality of the services. Users were less likely to indicate concerns with using digital health care technologies and, where they did, these mostly related to personal factors such as the costs of services or reservations about the privacy of personal health information.

Comparison With Prior Work

Commentaries and empirical research have previously indicated that the use of DTC telemedicine services may support access to health care [3,10]. However, results from the present study do not indicate difficulty in accessing health care is a great driver of use, with most of the sample participants having seen a GP in the past year, reporting satisfaction with the service, and the ability to get an appointment. Prior research largely comes from the United States, which has a more pluralistic health care system with greater inequity in access than Australia. Moreover, these studies have predominately focused on DTC services providing synchronous "virtual visits" for urgent, but not life-threatening, conditions [3,5,10], rather than the new forms of DTC that offer "lifestyle medications" for the management of often ongoing health issues (eg, weight management, acne, and erectile dysfunction) [8,16]. For the latter service type, our results indicate that the benefit may be less in accessing health care than in accessing specific treatments that are otherwise inconvenient, difficult, or impossible to access through one's traditional health care provider, even despite the increased adoption of telehealth in Australian general practice since the COVID-19 pandemic [22].

XSL•FO
RenderX

Very few respondents indicated that they used DTC telemedicine to complement traditional health care (such as getting second opinions or checking symptoms). This suggests DTC telemedicine may be tapping into previously unmet health care needs, rather than simply replacing conventional GP visits. Most respondents also rarely if ever discussed using these services with their GP, further implying the provision of care for their specific health condition remained separate to their use of traditional health care services. What is apparent, though, is that use of these services contributed to a sense of empowerment among participants, who indicated feeling better informed and able to manage their health condition, a finding also identified by Kyweluk [23] in a study of users of DTC ovarian reserve testing services.

Compared with these benefits, participants in this sample far less frequently highlighted concerns with using digital health care technologies, such as DTC telemedicine services. The most common apprehensions were related to personal issues of privacy and cost; few respondents indicated feeling concerned about more systemic factors such as lack of regulation, integration, and quality and safety, despite these being frequently raised in the literature [3,8,9]. The perception of these services as costly may be more pertinent to Australian consumers, many of whom have historically been able to visit a GP without incurring out-of-pocket costs [24], and then access heavily subsidized medicines through the Pharmaceutical Benefits Scheme [25].

Strengths and Limitations

The use of existing subscales on patient empowerment [12], and the triangulation of qualitative and quantitative data analysis, are strengths of the study. Limitations include potential self-selection bias in responses to the survey. Data were missing for some demographic characteristics and other variables due to attrition and the fact that all responses were optional; we lack information as to the profile of users of these DTC telemedicine services for specific health conditions to understand how representative the sample may have been.

Conclusions

This study provides insights into perspectives of patients who use DTC telemedicine services for specific health concerns, which is an emergent area of the health care system and one with limited research to date. Through a survey of users, benefits of these services were identified, including ease of access, convenience, and providing a sense of empowerment. Some concerns were identified related to data privacy and the costs of utilizing the services. Further research is needed to understand how the use of DTC telemedicine fits into the existing landscape of primary health care and the degree to which the use of these services provides a complement, supplement, or alternative to traditional general practice.

Data Availability

The data reported on in this manuscript are not publicly available but may be available from the corresponding author upon reasonable request and with appropriate ethics approval.

Authors' Contributions

KC and LAE conceptualized and designed the study with input from DF, SS, and JB. MS contributed to the ethics and survey administration. MV facilitated recruitment. KC led the data analysis and interpretation, supported by EC and AT. KC wrote the original manuscript draft, which was critically revised by DF, AT, EC, MS, SS, MV, JB, and LAE. All authors approved the final manuscript.

Conflicts of Interest

MV works for and owns shares in the digital health care company whose users participated in this research. At the time of this research, DF provided clinical services for the digital health care company whose users participated in this research. DF is no longer affiliated with this company. All other authors declared no conflicts of interest.

Multimedia Appendix 1
Survey items.
[DOCX File, 21 KB - formative_v9i1e68619_app1.docx]

References

1. Bocking H, Russell-Bennett R, Letheren K. I'm no expert, but …? Consumer use of supportive digital tools in health services. J Serv Theory Pract 2022 Mar 9;32(2):105-131. [doi: 10.1108/JSTP-09-2020-0225]
2. Li KY, Zhu Z, Ng S, Ellimoottil C. Direct-to-consumer telemedicine visits for acute respiratory infections linked to more downstream visits. Health Aff (Millwood) 2021 Apr;40(4):596-602. [doi: 10.1377/hlthaff.2020.01741] [Medline: 33819099]
3. Uscher-Pines L, Mehrotra A. Analysis of Teladoc use seems to indicate expanded access to care for patients without prior connection to a provider. Health Aff (Millwood) 2014 Feb;33(2):258-264. [doi: 10.1377/hlthaff.2013.0989] [Medline: 24493769]
4. Teladoc Health. Teladoc Health. 2024. URL: https://teladochealth.com.au [accessed 2025-01-22]

XSL•FO
RenderX

5. Schoenfeld AJ, Davies JM, Marafino BJ, et al. Variation in quality of urgent health care provided during commercial virtual visits. JAMA Intern Med 2016 May 1;176(5):635-642. [doi: 10.1001/jamainternmed.2015.8248] [Medline: 27042813]
6. Cohen AB, Mathews SC, Dorsey ER, Bates DW, Safavi K. Direct-to-consumer digital health. Lancet Digit Health 2020 Apr;2(4):e163-e165. [doi: 10.1016/S2589-7500(20)30057-1] [Medline: 33328077]
7. Welch BM, Harvey J, O'Connell NS, McElligott JT. Patient preferences for direct-to-consumer telemedicine services: a nationwide survey. BMC Health Serv Res 2017 Nov 28;17(1):784. [doi: 10.1186/s12913-017-2744-8] [Medline: 29183372]
8. Elliott T, Yopes MC. Direct-to-consumer telemedicine. J Allergy Clin Immunol Pract 2019;7(8):2546-2552. [doi: 10.1016/j.jaip.2019.06.027] [Medline: 31706486]
9. Jain T, Lu RJ, Mehrotra A. Prescriptions on demand: the growth of direct-to-consumer telemedicine companies. JAMA 2019 Sep 10;322(10):925-926. [doi: 10.1001/jama.2019.9889] [Medline: 31348489]
10. Martinez KA, Rood M, Jhangiani N, et al. Patterns of use and correlates of patient satisfaction with a large nationwide direct to consumer telemedicine service. J Gen Intern Med 2018 Oct;33(10):1768-1773. [doi: 10.1007/s11606-018-4621-5] [Medline: 30112737]
11. Chung R. Telehealth physician oversight over direct to consumer testing: doctors working with patients towards patient empowerment. Telemed J E Health 2012 Jun;18(5):354-359. [doi: 10.1089/tmj.2011.0196] [Medline: 22545797]
12. van Uden-Kraan CF, Drossaert CHC, Taal E, Seydel ER, van de Laar M. Participation in online patient support groups endorses patients' empowerment. Patient Educ Couns 2009 Jan;74(1):61-69. [doi: 10.1016/j.pec.2008.07.044] [Medline: 18778909]
13. Ray KN, Shi Z, Gidengil CA, Poon SJ, Uscher-Pines L, Mehrotra A. Antibiotic prescribing during pediatric direct-to-consumer telemedicine visits. Pediatrics 2019 May;143(5):e20182491. [doi: 10.1542/peds.2018-2491] [Medline: 30962253]
14. Broffman L, Barnes M, Stern K, Westergren A. Evaluating the quality of asynchronous versus synchronous virtual care in patients with erectile dysfunction: retrospective cohort study. JMIR Form Res 2022 Jan 13;6(1):e32126. [doi: 10.2196/32126] [Medline: 34905499]
15. Shi Z, Mehrotra A, Gidengil CA, Poon SJ, Uscher-Pines L, Ray KN. Quality of care for acute respiratory infections during direct-to-consumer telemedicine visits for adults. Health Aff (Millwood) 2018 Dec;37(12):2014-2023. [doi: 10.1377/hlthaff.2018.05091] [Medline: 30633682]
16. Foo D, Spanos S, Dammery G, Ellis LA, Willcock SM, Braithwaite J. The rise of direct-to-consumer telemedicine services in Australia: implications for primary care and future research. Med J Aust 2023 Oct 16;219(8):344-347. [doi: 10.5694/mja2.52097] [Medline: 37666506]
17. Moshy. 2024. URL: https://www.getmoshy.com.au [accessed 2025-01-22]
18. Pilot. 2024. URL: https://pilot.com.au [accessed 2025-01-22]
19. Bollmeier SG, Stevenson E, Finnegan P, Griggs SK. Direct to consumer telemedicine: is healthcare from home best? Mo Med 2020;117(4):303-309. [Medline: 32848261]
20. Foo D, Spanos S, Dammery G, Ellis LA, Willcock S, Braithwaite J. Direct-to-consumer telemedicine in primary care settings: A scoping review of contemporary empirical literature. Aust J Gen Pract 2024;53(11 Suppl):S107-S114. [doi: 10.31128/AJGP-12-23-7070] [Medline: 39542684]
21. Bosco A, Jones KA, Di Lorito C, Stephan BCM, Orrell M, Oliveira D. Changing lifestyle for dementia risk reduction: Inductive content analysis of a national UK survey. PLoS ONE 2020;15(5):e0233039. [doi: 10.1371/journal.pone.0233039] [Medline: 32413085]
22. Ashley C, Williams A, Dennis S, et al. Telehealth's future in Australian primary health care: a qualitative study exploring lessons learnt from the COVID-19 pandemic. BJGP Open 2023 Jun;7(2):BJGPO.2022.0117. [doi: 10.3399/BJGPO.2022.0117] [Medline: 36750374]
23. Kyweluk MA. Quantifying fertility? Direct-to-consumer ovarian reserve testing and the new (in)fertility pipeline. Soc Sci Med 2020 Jan;245:112697. [doi: 10.1016/j.socscimed.2019.112697] [Medline: 31786460]
24. De Abreu Lourenco R, Kenny P, Haas MR, Hall JP. Factors affecting general practitioner charges and Medicare bulk-billing: results of a survey of Australians. Med J Aust 2015 Feb 2;202(2):87-90. [doi: 10.5694/mja14.00697] [Medline: 25627740]
25. Willis E, Reynolds L, Rudge T, editors. The Pharmaceutical Benefits Scheme. In: Understanding the Australian Health Care System, 4th edition: Elsevier Australia; 2020:100-120.

Abbreviations

DTC: direct-to-consumer
GP: general practitioner

XSL•FO
RenderX

Edited by A Mavragani; submitted 11.11.24; peer-reviewed by EVD Velde, K Ufholz, S Bollmeier; revised version received 18.12.24; accepted 19.12.24; published 03.02.25.

Please cite as:
Churruca K, Foo D, Turner A, Crameri E, Saba M, Spanos S, Vickers M, Braithwaite J, Ellis LA
Users' Perspectives of Direct-to-Consumer Telemedicine Services: Survey Study
JMIR Form Res 2025;9:e68619
URL: https://formative.jmir.org/2025/1/e68619
doi:10.2196/68619

© Kate Churruca, Darran Foo, Andrew Turner, Emily Crameri, Maree Saba, Samantha Spanos, Matthew Vickers, Jeffrey Braithwaite, Louise A Ellis. Originally published in JMIR Formative Research (https://formative.jmir.org), 3.2.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Feasibility, User Acceptance, and Outcomes of Using a Cancer Prehabilitation App for Exercise: Pilot Cohort Study

Fuquan Zhang[1], MD; Deepali Bang[1], MD, PhD, MRCP; Christine Alejandro Visperas[1], MD; Mon Hnin Tun[2], PhD; San San Tay[1], MBBS, MRCP

[1]Department of Rehabilitation Medicine, Changi General Hospital, 2 Simei St 3, Singapore, Singapore

[2]Department of Health Services Research, Changi General Hospital, Singapore

Corresponding Author:
San San Tay, MBBS, MRCP
Department of Rehabilitation Medicine, Changi General Hospital, 2 Simei St 3, Singapore, Singapore

Abstract

Background: The efficacy of cancer prehabilitation programs is supported by international reviews and meta-analyses. Technology has been deployed in cancer prehabilitation to address challenges such as access or limited resources. This study evaluated the feasibility, user acceptance, safety, and program outcomes of a newly developed mobile app for cancer prehabilitation. The app integrates with Singapore's existing health care mobile app, Health Buddy, and provides instructional videos for prescribed exercises.

Objective: The objectives of this study were to investigate the feasibility, user experience, safety, and outcomes of a mobile app for cancer prehabilitation within a hospital-associated, home-based, multimodal cancer prehabilitation program.

Methods: This retrospective study analyzed the records of patients enrolled in the cancer prehabilitation program from September 1, 2022, to March 30, 2023. Patients who participated in the prehabilitation program (n=63) were categorized into 2 groups: those prescribed the app (n=41) and those who were not (n=22). There was further subgroup analysis of those who were prescribed: app users (n=25) versus those who were non-app users (n=16). Demographics, Fried Frailty Phenotype, prehabilitation duration, app use, and functional outcome measures (6-minute walk test [6MWT], 30-second sit-to-stand test [STS], timed up and go test [TUG], and Hospital Anxiety and Depression Scale [HADS]) were collected. Compliance was determined by the completion of prescribed exercises and the accuracy of executing these exercises, with a high compliance rate considered to be at 80% or more. Baseline characteristics and preoperative outcomes were compared between the groups. User satisfaction was assessed through surveys among app users (n=25).

Results: Among 63 patients, 41 (65.1%) patients were prescribed the app, of which 22 (34.9%) patients were users. No significant differences in preoperative functional improvements were observed between app users and nonusers (6MWT: P=.60; STS: P=.81; TUG: P=.53; HADS: P=.36), or between those prescribed and not prescribed the app (6MWT: P=.94; STS: P=.26; TUG: P=.39; HADS: P=.62). However, high compliance rates (80%) were observed among app users. Patient satisfaction with the app was high (>90%), with positive feedback on ease of use and technical reliability. Baseline measures revealed significantly lower functional scores and higher mean frailty scores in the nonprescribed group.

Conclusions: This preliminary study demonstrates the acceptability, feasibility, and safety of Singapore's first smartphone app for exercise prescription in cancer prehabilitation. Lower baseline functional outcome measures and a higher mean frailty score in the unprescribed group have implications for the selection process and patient participation. Further studies should include strategies to enhance patients' readiness for technology, sustainability, and effectiveness in older patients.

(JMIR Form Res 2025;9:e64427) doi:10.2196/64427

KEYWORDS

cancer prehabilitation; mobile app; technology; feasibility; acceptance; cancer; prehabilitation; mHealth; exercise; application; app; mobile application; reliability; smartphone app; sustainability; effectiveness; older patients; older adults; elderly

Introduction

Cancer prehabilitation is defined as a set of interventions delivered within the cancer care continuum, occurring during the period between diagnosis and the initiation of definitive treatment [1]. It encompasses baseline assessments of physical function, nutritional status, and psychological well-being. Prehabilitation also involves the identification of potential impairments and the implementation of targeted interventions aimed at optimizing health, thereby reducing the risk and severity of future treatment-related complications. The efficacy of cancer prehabilitation programs is supported by findings

XSL•FO
RenderX

from several international reviews and meta-analyses [2-4]. However, ensuring equitable access to prehabilitation services remains a challenge, particularly in settings with limited health care resources or geographical barriers to care [5]. In another study, the preferred method for exercise program delivery was home-based, with 1 supervised session per week [6]. Technology-enabled prehabilitation programs may be able to support home-based programs, which is important to improve adherence and maximize program effectiveness within a short prehabilitation window [5].

A systematic review of cancer prehabilitation programs delivered through technological enablers found the studies to be feasible. While the effectiveness of these programs varied based on the intervention design and the specific technologies used, many studies reported improvements in physical function, activity levels, and overall patient satisfaction. It was recommended that future efforts focus on implementing multimodal programs and using intervention-specific outcome measures to enhance the overall effectiveness of cancer prehabilitation [7].

Our institution has established a hospital-associated, home-based prehabilitation program targeting patients with newly diagnosed gastrointestinal and urological cancers awaiting surgery [8,9]. This multimodal program incorporates medical optimization strategies, individualized exercise prescriptions, and nutritional and mental health interventions. This program was facilitated by a small core team consisting of a coordinator who collects the outcome measures and a physiatrist who sees the patients on the same day as the surgeon visits, providing a 1-stop service. The compliance of the patients to the home exercise program was tracked by phone calls made by the coordinator on a weekly to fortnightly basis. The program has demonstrated significant improvements in functional capacity, as evidenced by metrics such as the 6-minute walk test (6MWT), 30-second sit-to-stand test (STS), and timed up and go test (TUG). There were significant improvements in the Hospital Anxiety and Depression Scale (HADS) as well.

In November 2021, the program integrated a cancer prehabilitation exercise diary into our existing network's widely used regional health care mobile app, Health Buddy (Singapore Health Services [Singhealth]) [10]. The Health Buddy app allows the user to manage appointments, make payments, and have access to health tools and education.

Despite the wide range of digital health apps [11], no cancer prehabilitation apps are currently available in Singapore. This study aims to evaluate the feasibility, user acceptance, safety, and program outcomes associated with using this mobile app within our multimodal cancer prehabilitation program.

Methods

The Cancer Prehabilitation Exercise Diary

The cancer prehabilitation exercise diary incorporates a library of instructional videos demonstrating commonly prescribed exercises. These exercises are selected by a physiatrist and are customized for individual patients to ensure patient safety and efficacy. Patients can use the videos for guidance during their daily exercise routines. Additionally, the diary features functionalities such as customized exercise reminders, exercise logs for tracking progress, and an achievement summary to motivate program adherence (Figure 1). Some patients are sedentary and older patients may not be exercising regularly or making the right movements. The intention of having this exercise diary was to facilitate patients in their exercises as they would be able to view the exercises and follow along. The intention was that it would be easier to follow than figures on paper and it may encourage increased participation. Patients who decline the app will be provided with paper exercise prescriptions whereas those who accepted the exercise prescription on the exercise diary will follow the videos for strengthening exercises in addition to aerobic exercise prescriptions. Both groups of patients were prescribed both aerobic and strengthening exercises.

XSL•FO
RenderX

Figure 1. Health Buddy app exercise diary.

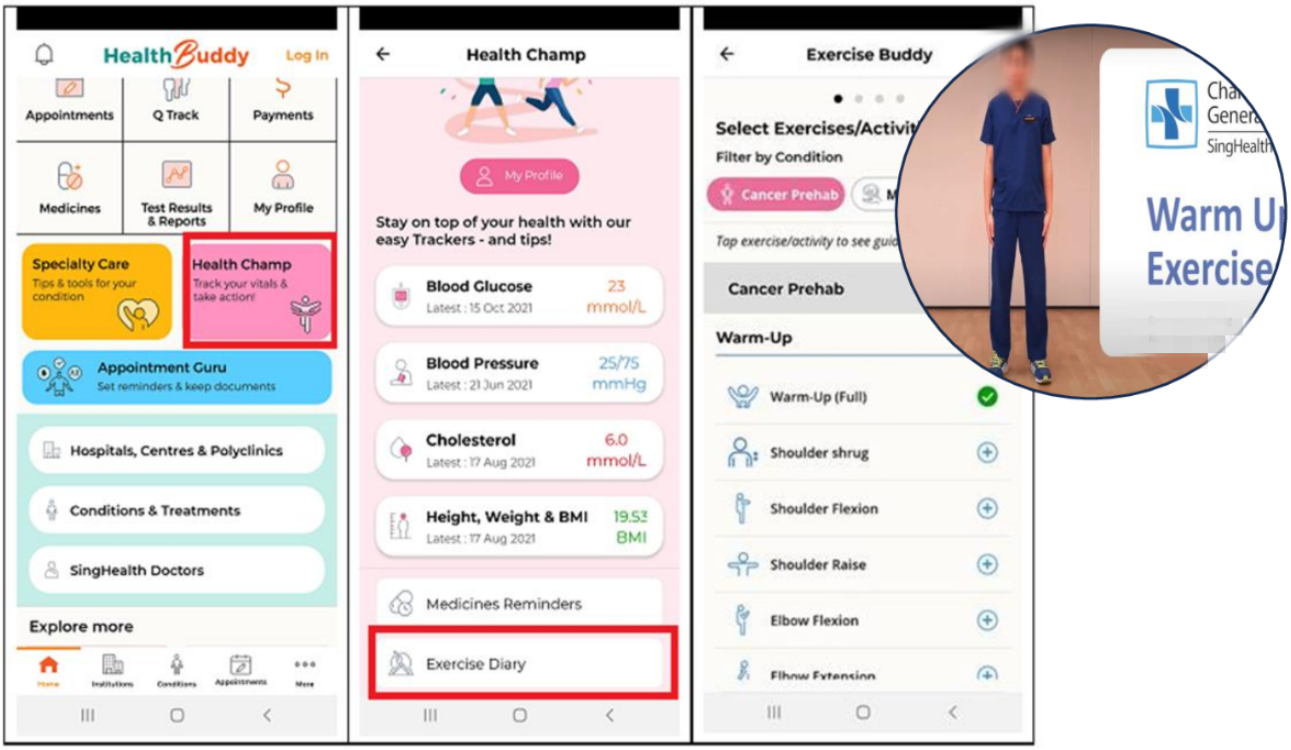

Aims and Hypothesis

This study aimed to evaluate the feasibility, user acceptance, safety, and program outcomes associated with using the cancer prehabilitation exercise diary within our multimodal cancer prehabilitation program. The hypothesis was that it is feasible and safe, and having the app will facilitate greater compliance and accuracy of the exercises. Patient satisfaction with the app would be sought. The physical and mental health improvements should be no worse off than the conventional group.

Study Design

We examined the records of patients who participated in the cancer prehabilitation program from September 1, 2022, to March 30, 2023. The cancer diagnostic groups included colorectal, hepatobiliary, upper gastrointestinal, and urological cancers planned for surgery at our hospital. They were referred by the surgeons at the point of diagnosis to the prehabilitation coordinator.

The inclusion criteria were cancer patients older than 18 years who attended the surgical clinics and were listed for surgery and who were enrolled in the cancer prehabilitation program. Consecutive patients who were enrolled were studied. Patients who declined surgical intervention were excluded from the study. Outcome measures at baseline and the preoperative period were routinely collected, as well as the patient satisfaction surveys for audit purposes.

Participants who received exercise interventions through the cancer prehabilitation exercise diary were categorized into 2 primary groups: those who were prescribed the app and those who were not. Among those prescribed the app, a further subdivision was made into app users and nonusers. The categorization into app prescribed or not prescribed was not decided a priori but based on the physician's behavior. The categorization of the app used or nonused was also a result of the patient's autonomous decision. As this is a retrospective study, the participants were nonrandomized. Patients prescribed exercises via the app were invited to complete a patient satisfaction survey upon program completion.

Demographics

Demographic data, including age, gender, and disease diagnosis, were obtained from patient medical records. Additionally, medical charts were reviewed for data pertaining to the Fried Frailty Phenotype [12] as well as the duration of the prehabilitation program. Information regarding if the patients were prescribed the app, and the reason if not, as well as the actual use of the app were collected.

The Fried Frailty Phenotype [13] assesses physical frailty through 5 criteria: unintentional weight loss; weakness or poor handgrip strength; self-reported exhaustion; slow walking speed; and low physical activity. Patients would be considered frail if they met the score criteria of at least 2 items, prefrail if they met 1 criteria, and nonfrail if they met none. The duration of prehabilitation is the number of days from enrollment to the preoperative visit (typically held on the day before the surgery).

Baseline and Preoperative Outcome Measures, Compliance, and Adverse Events

Baseline and preoperative outcome measures were collected at the program's initiation and completion (preoperative visit), respectively. Physical function was assessed using the 6MWT, STS, and TUG. Mental health outcomes were evaluated using the HADS.

Compliance data were also captured, with full compliance defined as completing all prescribed exercises and demonstrating

XSL•FO
RenderX

them correctly at the preoperative visit. Partial compliance was defined as having some inaccuracy in exercise performance or not completing the entire set. Although there is no universal agreed-upon threshold, we considered a compliance rate of 80% to be high. Reasons for not prescribing the app or for nonuse of the app were extracted from electronic medical records.

For patients who used the app, a patient satisfaction survey was administered upon program completion. Questions pertaining to the awareness of the Health Buddy app and the use of the smartphone for features within Health Buddy were asked.

Statistical Analysis

Categorical variables were presented as proportions or percentages, while continuous variables were described using means and SDs or medians and IQRs. A linear mixed-effects model for repeated measures was used to assess differences in outcome measures between groups (prescribed app vs nonprescribed app and app users vs nonusers). Two-tailed statistical tests with a significance level (α) of 0.05 were used. All statistical analyses were performed using Stata 16 (StataCorp LP) software.

Ethical Considerations

This study was reviewed and approved by our institutional review board (2021/2751). As all patient information in the dataset was obtained from retrospective records and deidentified, informed consent was waived and allowed for secondary analysis without additional consent. No financial compensation was offered to participants of this study.

Results

Demographics and Baseline Outcome Measures

Between September 1, 2022, and March 30, 2023, a total of 66 patients were enrolled in the cancer prehabilitation program. A total of 3 patients dropped out from surgery, leaving 63 patients to be included in the study (Table 1). Of these, 41 (65.1%) patients were prescribed exercise interventions through the cancer prehabilitation app. The remaining 22 patients (34.9%) were not prescribed the app for various reasons, including lack of a smartphone or data plan (n=5), undocumented physician rationale (n=9), patient refusal due to a preference for established methods (n=6), short duration to the operation (n=1), and hearing impairment (n=1). Details regarding the distribution of these reasons are provided in Table S1 in Multimedia Appendix 1.

The majority of participants were of Chinese ethnicity, reflecting the demographics of Singapore. No statistically significant differences were observed in age or gender distribution between the 2 groups (app-prescribed vs nonprescribed). Baseline physical function assessments revealed significantly lower scores in the nonprescribed group on the 6MWT, STS, and TUG (Table 1). The nonprescribed group also demonstrated a higher baseline depression score on the HADS. Patients who were not prescribed the app exhibited significantly weaker hand grip strength, slower gait speed, and a higher mean frailty score.

Among the 41 patients prescribed exercises via the cancer prehabilitation app, 25 (58.1%) were identified as users, while 16 (41.9%) were nonusers. Baseline outcome measures did not reveal significant differences between these groups (Table 2).

Patients were categorized under “undefined” compliance, as there were no data available for collection in view of patients not being contactable/no-show for appointments. Patients who had their operation canceled were not included in the analysis (reasons for cancellation of operation: medically unfit due to new medical condition [n=2], transferred to another hospital for a second opinion [n=1]).

XSL•FO
RenderX

Table . Baseline characteristics of app-prescribed patients versus nonprescribed patients in outpatient cohort.

	Outpatient (app prescribed)		*P* value
	No (n=22, 34.9%)	Yes (n=41, 65.1%)	
Sex, n (%)			.27
Female	10.0 (45.5)	12.0 (29.3)	
Male	12.0 (54.5)	29.0 (70.7)	
Race, n (%)			.62
Chinese	19.0 (86.4)	35.0 (85.4)	
Malay	2.0 (9.1)	2.0 (4.9)	
Indian	0.0 (0.0)	3.0 (7.3)	
Others	1.0 (4.5)	1.0 (2.4)	
Age (years)			
Mean (SD)	72.1 (8.8)	68.2 (8.9)	.11
Cancer diagnostic groups			.59
Colorectal cancer	7.0 (31.8)	13.0 (31.7)	
Hepatobiliary cancer	9.0 (40.9)	21.0 (51.2)	
Upper gastrointestinal cancer	4.0 (18.2)	3.0 (7.3)	
Urological cancer	2.0 (9.1)	4.0 (9.8)	
Duration of prehabilitation			
Mean (SD)	20.8 (8.9)	20.3 (9.3)	.84
Baseline			
Frailty score			
Mean (SD)	2.3 (1.5)	1.3 (1.2)	0.002[a]
Median (IQR)	2.0 (1.0-3.0)	1.0 (0.0-2.0)	0.003[a]
6MWT[b]			
Mean (SD)	267.6 (89.7)	348.0 (88.3)	0.002[a]
30-second sit-to-stand test			
Mean (SD)	8.5 (3.7)	10.7 (3.9)	.06
Time up and go			
Mean (SD)	16.3 (10.5)	10.5 (3.9)	0.003[a]
Anxiety			
Mean (SD)	2.1 (2.0)	2.3 (2.3)	.71
Depression			
Mean (SD)	2.5 (4.2)	1.2 (1.8)	.08
EQ-5D[c]			
Mean (SD)	68.0 (14.9)	70.2 (20.4)	.66
Anxiety and depression base			
Mean (SD)	4.6 (4.9)	3.5 (3.5)	.31
Maximal hand grip strength (kg)			
Mean (SD)	20.3 (11.8)	27.1 (6.9)	0.001[a]
Gait speed less than 0.67 m/s, n (%)			0.002[a]
No	13.0 (59.1)	38.0 (92.7)	

XSL•FO
RenderX

	Outpatient (app prescribed)		*P* value
	No (n=22, 34.9%)	Yes (n=41, 65.1%)	
Yes	9.0 (40.9)	3.0 (7.3)	
Reduced physical activity, n (%)			>.99
No	15.0 (68.2)	27.0 (65.9)	
Yes	7.0 (31.8)	14.0 (34.2)	
Exhaustion, n (%)			.36
No	15.0 (68.2)	33.0 (80.5)	
Yes	7.0 (31.8)	8.0 (19.5)	
Unintentional weight loss, n (%)			.05
No	11.0 (50.0)	31.0 (75.6)	
Yes	11.0 (50.0)	10.0 (24.4)	
Compliance, n (%)			
Undefined	11.0 (50.0)	8.0 (19.5)	__[d]
Full	8.0 (36.4)	24.0 (58.5)	.04[a]
Partial	3.0 (13.6)	9.0 (21.9)	.14

[a]Statistically significant values.
[b]6MWT: 6-minute walk test.
[c]EQ-5D: EuroQol group 5-dimensions questionnaire.
[d]Not applicable.

XSL•FO
RenderX

Table . Baseline characteristics of users and nonusers of the app in outpatient cohort.

	Outpatient (app used)		*P* value
	No (n=16, 41.9%)	Yes (n=25, 58.1%)	
Sex, n (%)			.08
Female	2.0 (12.5)	10.0 (40.0)	
Male	14.0 (87.5)	15.0 (60.0)	
Race, n (%)			.59
Chinese	14.0 (87.5)	21.0 (84.0)	
Malay	0.0 (0.0)	2.0 (8.0)	
Indian	1.0 (6.3)	2.0 (8.0)	
Others	1.0 (6.3)	0.0 (0.0)	
Age (years)			
Mean (SD)	67.0 (10.5)	69.0 (7.9)	.48
Cancer diagnostic groups			.28
Colorectal cancer	4.0 (25.0)	9.0 (36.0)	
Hepatobiliary cancer	11.0 (68.8)	10.0 (40.0)	
Upper gastrointestinal cancer	0.0 (0.0)	3.0 (12.0)	
Urological cancer	1.0 (6.2)	3.0 (12.0)	
Duration of prehabilitation			.77
Mean (SD)	19.8 (9.2)	20.7 (9.5)	
Baseline			
Frailty score			
Mean (SD)	1.1 (0.9)	1.3 (1.3)	.63
Median (IQR)	1.0 (0.0-2.0)	1.0 (0.0-2.0)	.81
6MWT[a]			
Mean (SD)	362.7 (98.4)	338.6 (81.9)	.40
30-second sit-to-stand test			
Mean (SD)	10.9 (3.3)	10.5 (4.3)	.73
Time up and go			
Mean (SD)	10.6 (5.2)	10.5 (3.1)	.94
Anxiety			
Mean (SD)	1.8 (2.0)	2.6 (2.4)	.26
Depression			
Mean (SD)	0.8 (1.4)	1.4 (2.0)	.37
EQ-5D[b]			
Mean (SD)	72.8 (18.1)	68.6 (22.0)	.53
Anxiety and depression base			
Mean (SD)	2.7 (2.9)	4.0 (3.7)	.23
Maximal hand grip strength (kg)			
Mean (SD)	28.7 (7.8)	26.08 (6.4)	.26
Gait speed less than 0.67 m/s, n (%)			.10
No	15.0 (93.4)	23.0 (92.0)	
Yes	1.0 (6.3)	2.0 (8.0)	

XSL•FO
RenderX

	Outpatient (app used)		*P* value
	No (n=16, 41.9%)	Yes (n=25, 58.1%)	
Reduced physical activity, n (%)			.50
No	12.0 (75.0)	15.0 (60.0)	
Yes	4.0 (25.0)	10.0 (40.0)	
Exhaustion, n (%)			>.99
No	13.0 (81.3)	20.0 (80.0)	
Yes	3.0 (18.8)	5.0 (20.0)	
Unintentional weight loss, n (%)			.47
No	11.0 (68.8)	20.0 (80.0)	
Yes	5.0 (31.3)	5.0 (20.0)	
Compliance, n (%)			
Undefined	8.0 (50.0)	0.0 (0.0)	—[c]
Full	4.0 (25.0)	20.0 (80.0)	.001[d]
Partial	4.0 (25.0)	5.0 (20.0)	.005[d]

[a]6MWT: 6-minute walk test.

[b]EQ-5D: EuroQol group 5-dimensions questionnaire.

[c]Not applicable.

[d]Statistically significant values.

Duration of Prehabilitation

Prehabilitation duration was comparable between the 2 groups (mean of 20.3 d for app-prescribed exercise, mean of 20.8 d for nonprescribed exercise) (Table 1). Among patients prescribed the app, the mean prehabilitation duration was slightly longer for app users (20.7 d) compared with nonusers (19.8 d) (Table 2).

Preoperative Outcome Measures, Compliance, and Adverse Events

Preoperative outcome measures were compared between patients prescribed the app and those who were not prescribed the app. No statistically significant differences in functional improvements were observed between the groups (Table 3). This finding is consistent with the observation that both groups experienced improvements in functional outcomes and mental health, with no documented adverse events.

XSL•FO
RenderX

Table . Changes in outcome measures in outpatient cohort between app-prescribed and nonprescribed patients.

Mean changes in outcome	Outpatient (app prescribed)		*P* value
	No (n=22, 34.9%)	Yes (n=41, 65.1%)	
6MWT[a]			
Baseline (mean)	267.6	348.0	__[b]
Pre-op (mean, 95% CI)	24.2 (3.8 to 44.6)[c]	23.3 (9.9 to 36.5)[c]	.94
30-second sit-to-stand test			
Baseline (mean)	8.5	10.7	—
Pre-op (mean, 95% CI)	2.3 (0.6 to 3.9)[c]	1.1 (0.1 to 2.2)[c]	.26
Time up and go			
Baseline (mean)	16.3	10.5	—
Pre-op (mean, 95% CI)	−1.3 (−2.3 to −0.3)[c]	−0.8 (−1.4 to −0.1)[c]	.39
Anxiety			
Baseline (mean)	2.1	2.3	—
Pre-op (mean, 95% CI)	−0.5 (−1.7 to 0.6)	−0.3 (−1.0 to 0.5)	.68
Depression			
Baseline (mean)	2.5	1.2	—
Pre-op (mean, 95% CI)	−0.3 (−1.1 to 0.5)	−0.2 (−0.7 to 0.3)	.84
Anxiety and depression base			
Baseline (mean)	4.6	3.5	—
Pre-op (mean, 95% CI)	−0.9 (−2.6 to 0.6)	−0.5 (−1.6 to 0.5)	.62
EQ-5D[d]			
Baseline (mean)	68.0	70.2	—
Post-op 3 month (mean, 95% CI)	8.8 (−2.3 to 19.8)	−1.6 (−7.9 to 4.7)	.11

[a]6MWT: 6-minute walk test.

[b]Not applicable.

[c]Statistically significant drop or increase from the baseline.

[d]EQ-5D: EuroQol group 5-dimensions questionnaire.

Of all patients who were prescribed the app, 24 patients (58.5%) exhibited full compliance, 9 (21.9%) patients exhibited partial compliance, and 8 (19.5%) patients were lost to follow-up. Among those who were not prescribed the app, 8 (36.4%) patients were fully compliant, 3 (13.6%) patients were partially compliant, and 11 patients were lost to follow up (Table 1). The higher incidence of full compliance in the app-prescribed group as compared with the nonprescribed group was statistically significant (*P*=.04).

Among those who were prescribed the app, there was no significant difference in the functional and mental health outcome improvements during the preoperative period between the app users and the nonusers (Table 4).

XSL•FO
RenderX

Table . Changes in outcome measures between users and nonusers of the app.

Mean changes in outcome	App used		*P* value
	No (n=16, 41.9%)	Yes (n=25, 58.1%)	
6MWT[a]			
Baseline (mean)	362.7	338.6	__[b]
Pre-op (mean, 95% CI)	28.8 (4.8 to 52.8)[c]	21.0 (4.9 to 37.1)[c]	.60
30-second sit-to-stand test			
Baseline (mean)	10.9	10.5	—
Pre-op (mean, 95% CI)	0.9 (−0.9 to 2.9)[c]	1.2 (0.01 to 2.4)	.81
Time up and go			
Baseline (mean)	10.6	10.5	—
Pre-op (mean, 95% CI)	−0.5 (−1.6 to 0.6)[c]	−0.9 (−1.6 to −0.2)[c]	.53
Anxiety			
Baseline (mean)	1.8	2.6	—
Pre-op (mean, 95% CI)	0.4 (−1.0 to 1.8)	−0.6 (−1.6 to 0.4)	.27
Depression			
Baseline (mean)	0.8	1.4	—
Pre-op (mean, 95% CI)	−0.1 (−0.9 to 0.7)	−0.3 (−0.9 to 0.3)	.74
Anxiety and depression base			
Baseline (mean)	2.7	4.0	—
Pre-op (mean, 95% CI)	0.2 (−1.7 to 2.2)	−0.9 (−2.3 to 0.5)	.36
EQ-5D[d]			
Baseline (mean)	72.8	68.6	—
Post-op 3 month (mean, 95% CI)	−5.8 (−17.9 to 6.2)	0.3 (−7.6 to 8.2)	.41

[a]6MWT: 6-minute walk test.

[b]Not applicable.

[c]Statistically significant drop or increase from the baseline.

[d]EQ-5D: EuroQol group 5-dimensions questionnaire.

High compliance rates (80%) were observed among app users. Conversely, a high proportion of nonusers (50%) did not return for the preoperative visit or were not contactable. For the remaining nonusers who returned for the pre-op visit and underwent surgery, half (25%) demonstrated full compliance with the exercise program, while the other half (25%) exhibited partial compliance (Table 2). The higher full compliance rate in the app users as compared with the nonusers was statistically significant (*P*=.001).

User Satisfaction Survey

A 76% (19/25) response rate was observed for the patient satisfaction survey among app users. The survey revealed high overall satisfaction with the app, ease of use, and technical reliability, with over 90% (23/25) of respondents reporting these aspects favorably. Furthermore, a high proportion of users indicated a likelihood to recommend the app to others. Patients also reported generally positive experiences with the clarity of exercise instructions and their ability to perform the exercises as instructed (Figure 2).

XSL•FO
RenderX

Figure 2. Impression of the prehabilitation smartphone app.

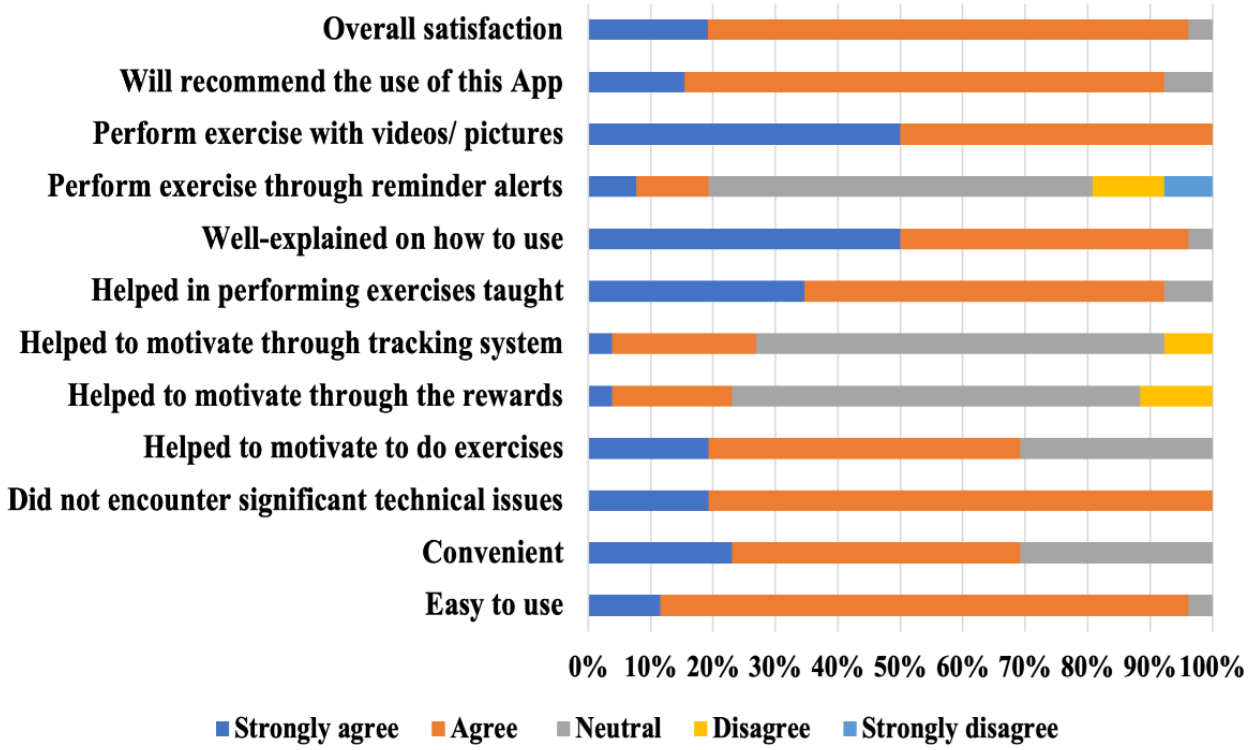

However, the survey identified the app's motivational features, particularly reminders and rewards, as an area for improvement. Interestingly, a significant portion of app users reported prior familiarity with the Health Buddy app and regular smartphone use for various activities. A total of 100% (25/25) of responders were aware of the Health Buddy app; 100% (25/25) of responders were aware of the Health Buddy functions; and 80% (20/25) , 64% (16/25), and 60% (15/25) of the responders had used the app for checking or scheduling appointments, receiving health education and making payments respectively. Features such as ordering medication and tracking of clinic queues were less used, at 44% (11/25) and 32% (8/25) respectively.

Discussion

Principal Findings

A recent systematic review [13] identified technical barriers, lack of direct supervision, and noncompliance as key challenges in implementing technology-supported cancer prehabilitation programs. Our findings suggest these concerns may not be universally applicable.

Patient satisfaction surveys revealed that app users in our study reported high technical reliability, ease of use, and the ability to follow exercise instructions (Figure 2). This suggests that technical issues were not a significant barrier for our user population. This could likely be attributed to the fact that the exercise diary was incorporated into a well-established health care app in Singapore's largest health care system [10,11]. As such, familiarity with an existing app might have an important role in facilitating ease of use, making this a more user-friendly experience. Additionally, a curated selection of suitable exercise videos will be made available in a patient's individualized cancer prehabilitation exercise diary within the Health Buddy app. This user-centric approach streamlines the process, enabling patients to commence their program effortlessly by accessing the designated videos.

Furthermore, a high compliance rate was observed among app users, with 80% achieving full compliance and 20% achieving partial compliance. These findings indicate that a lack of direct supervision did not hinder functional improvement in our program for the population that used the app. Conversely, those who were not prescribed had a higher noncompliance rate and loss of follow-up. That suggests compliance may be an issue with home-based programs itself.

The development of cancer prehabilitation applications is a growing field. For instance, a prototype app designed to support blood flow restriction training and sports nutrition demonstrated high user satisfaction regarding ease of use, utility, and overall function [14]. Our study adds to this emerging body of evidence by investigating the feasibility and user acceptance of a mobile app for cancer prehabilitation program delivery.

Feasibility and Acceptance of App-Based Prehabilitation

Among the 63 patients in the study, only 41 (65.1%) patients were prescribed exercise interventions through the app. There is borderline acceptance by physicians and patients. Given the full compliance rate of 80% among app users, app-based prehabilitation is feasible.

The main reasons for not prescribing the app included (1) patient limitations: lack of technological knowledge, smartphone/data plan limitations, hearing impairment, or short prehabilitation windows; (2) patient preferences: unwillingness to participate;

XSL•FO
RenderX

and (3) undocumented rationale: in 10 cases, the physician's rationale for not prescribing the app was not documented.

The study highlights the potential impact of socioeconomic factors on technology readiness for cancer prehabilitation. Unlike studies in Western countries with established social acceptance of telehealth [15], a prior local study suggests lower acceptance among diabetic patients [16]. This underscores the need to address potential disparities in access and use. Similar to interventions promoting telehealth for diabetic patients, considerations for app-based prehabilitation might include (1) tailoring support to specific socioeconomic groups less likely to use technology, (2) providing trained staff to address patient concerns and guide app engagement, and (3) exploring the feasibility of loaner devices for the prehabilitation period.

A recent scoping review of reviews on digital health and telehealth in cancer care identified gaps in the literature [12]. These gaps include (1) studies focusing on older adults or those experiencing bereavement, (2) research on the long-term sustainability of telehealth interventions, and (3) comparative studies that directly evaluate telehealth versus in-person interventions.

Similarly, in our patient population, addressing these knowledge gaps is crucial for overcoming challenges associated with nonreadiness and nonacceptance of technology-enabled prehabilitation programs.

Baseline Differences and Preoperative Outcomes

Patients who were not prescribed the app exhibited significantly lower baseline scores on functional performance measures, and a higher baseline depression score on the HADS. Furthermore, the Fried Frailty Phenotype assessment revealed significantly weaker handgrip strength, slower gait speed, a higher incidence of unintentional weight loss, and a higher mean frailty score in the nonapp group [12].

The Fried Frailty Phenotype focuses on physical frailty [12]. Cognitive frailty, defined by the International Consensus Group organized by the International Academy on Nutrition and Aging and the International Association of Gerontology and Geriatrics, refers to a syndrome in older adults characterized by both physical frailty and cognitive impairment, excluding dementia diagnoses [17]. Our study design did not assess cognitive function, leaving it unclear whether cognitive frailty was present in our cohort and potentially influenced physician or patient selection regarding app prescription.

The significantly higher Fried frailty score in the nonprescribed population suggested that being generally frailer and having lower baseline outcome measures may impair program participation. This is in concurrence with a study demonstrating a negative moderate significant correlation found between frailty and participation [18]. This may be even more so for technology-based programs in the face of a challenging new diagnosis. Participation in this study was a broad concept with several aspects, such as preferences, enjoyment, and satisfaction from activities [19].

Both app users and nonusers who were prescribed the app demonstrated improvements in functional measures (6MWT, STS, and TUG) at the preoperative assessment, with no statistically significant differences observed between these groups. Similarly, no significant differences in preoperative outcomes were found when comparing patients prescribed the app to those who were not. While high compliance was observed among app users who were able to follow the exercise videos, these findings suggest no preoperative outcome advantage associated with app use in this preliminary study. Compliance is defined by the completion of the prescribed set of exercises and the accuracy of the exercises. Using the app may have resulted in greater accuracy of the exercise performed, as was intended.

Patient Satisfaction and Future Considerations

A high level of user satisfaction was observed, with 90% of postintervention survey respondents reporting satisfaction with the app and a willingness to recommend it to others. This finding suggests promise for the app's utility among patients receptive to technology-enabled health care interventions.

In a mixed methods systematic review of access, acceptance, and adherence to cancer prehabilitation, there was a perceived value of home-based prehabilitation and the acceptability of tele-prehabilitation [13]. The availability and extent of integrated health care professional supervision and support were perceived to enable intervention access and adherence, especially if this was personalized. Other studies found prehabilitation enabled through technology to be acceptable and perceived to be accessible, especially during the pandemic [6,13,20,21]. Thus, there is value in refining our app-enabled prehabilitation program in the longer term.

Future research efforts should include potential pilot trials among older adults as well as diverse socioeconomic groups, with an aim to adopt strategies tailoring specific support for them while overcoming nonadherence challenges. To promote scalability and integration within our existing health care systems, prioritizing strategies to enhance patient readiness and adoption of technology-based prehabilitation programs should also be considered. Additionally, investigating the long-term sustainability of the intervention within the health care system and staff fidelity to program implementation are crucial areas for further exploration.

Limitations of Study

One limitation of the study is that we have a substantial number of participants who were eventually uncontactable or lost to follow-up, contributing to the missing data pool. We were unable to determine the reason for this, whether this was due to participant disengagement versus an inability to successfully operate the app. Nonetheless, efforts should be made in future research to implement strategies for real-time monitoring of a patient's progress and proactive intervention to address potential issues promptly, such as troubleshooting technical problems early or initiating motivational interviewing to address concerns or disinterest early to help maximize retention.

The participating groups were nonrandomized and heterogeneous. As such, selection bias can also be a potential concern in our study.

XSL•FO
RenderX

Finally, given the small sample size, our findings are preliminary and might be limited to acceptability and usability. Further research with a larger sample is needed to confirm and generalize these results.

Conclusion

In conclusion, this preliminary study demonstrates the acceptability, feasibility, and safety of Singapore's first smartphone app for exercise prescription in cancer prehabilitation. Future research directions include investigating the program's effectiveness among older adults, exploring strategies to enhance patient readiness for technology-based interventions, evaluating the program's long-term sustainability within the health care system, and assessing staff fidelity to program implementation.

Acknowledgments

We would like to express our sincere appreciation to Dr Kwok Kah Meng and Mr Christopher Richard Mattew De Conceicao Paul for their contributions.

Data Availability

The anonymized datasets generated and analyzed in this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Reasons for not prescribing the app.
[DOCX File, 14 KB - formative_v9i1e64427_app1.docx]

References

1. Silver JK, Baima J, Mayer RS. Impairment-driven cancer rehabilitation: an essential component of quality care and survivorship. CA Cancer J Clin 2013 Sep;63(5):295-317. [doi: 10.3322/caac.21186] [Medline: 23856764]
2. Treanor C, Kyaw T, Donnelly M. An international review and meta-analysis of prehabilitation compared to usual care for cancer patients. J Cancer Surviv 2018 Feb;12(1):64-73. [doi: 10.1007/s11764-017-0645-9] [Medline: 28900822]
3. Tay SS. Perspectives on the direction of cancer prehabilitation in the pandemic and beyond. Arch Rehabil Res Clin Transl 2022 Dec;4(4):100236. [doi: 10.1016/j.arrct.2022.100236] [Medline: 36277731]
4. Lambert JE, Hayes LD, Keegan TJ, Subar DA, Gaffney CJ. The impact of prehabilitation on patient outcomes in hepatobiliary, colorectal, and upper gastrointestinal cancer surgery: a PRISMA-accordant meta-analysis. Ann Surg 2021 Jul 1;274(1):70-77. [doi: 10.1097/SLA.0000000000004527] [Medline: 33201129]
5. Ferreira V, Agnihotram RV, Bergdahl A, et al. Maximizing patient adherence to prehabilitation: what do the patients say? Support Care Cancer 2018 Aug;26(8):2717-2723. [doi: 10.1007/s00520-018-4109-1] [Medline: 29478189]
6. Waterland JL, Chahal R, Ismail H, et al. Implementing a telehealth prehabilitation education session for patients preparing for major cancer surgery. BMC Health Serv Res 2021 May 10;21(1):443. [doi: 10.1186/s12913-021-06437-w] [Medline: 33971869]
7. Tay SS, Zhang F, Neo EJR. The use of technology in cancer prehabilitation: a systematic review. Front Oncol 2024;14:1321493. [doi: 10.3389/fonc.2024.1321493] [Medline: 38706603]
8. Tay SS, Kwok KM. Setting up a cancer prehabilitation framework in Singapore. J Cancer Rehabil 2021;4:1-6. [doi: 10.48252/JCR5]
9. Kwok KM, Tay SS. Outcomes of a multi-modal hospital-associated home-based cancer prehabilitation program. Ann Rehabil Med 2023 Feb;47(1):52-67. [doi: 10.5535/arm.22126] [Medline: 36710586]
10. Introducing a new & improved health buddy app interface!. SingHealth. URL: https://www.singhealth.com.sg/patient-care/patient-visitor-info/health-buddy-app [accessed 2024-12-04]
11. Perdana A, Mokhtar IA. Leveraging digital technologies for information technology-enabled healthcare transformation at SingHealth. J Inf Technol Teach Cases 2023 May;13(1):97-103. [doi: 10.1177/20438869221091582]
12. Fried LP, Tangen CM, Walston J, et al. Frailty in older adults: evidence for a phenotype. J Gerontol A Biol Sci Med Sci 2001 Mar;56(3):M146-M156. [doi: 10.1093/gerona/56.3.m146] [Medline: 11253156]
13. Watts T, Courtier N, Fry S, et al. Access, acceptance and adherence to cancer prehabilitation: a mixed-methods systematic review. J Cancer Surviv 2024 May 6. [doi: 10.1007/s11764-024-01605-3] [Medline: 38709465]
14. Wang T, Stanforth PR, Fleming RYD, Wolf JS Jr, Stanforth D, Tanaka H. A mobile app with multimodality prehabilitation programs for patients awaiting elective surgery: development and usability study. JMIR Perioper Med 2021 Dec 30;4(2):e32575. [doi: 10.2196/32575] [Medline: 34967752]

XSL•FO
RenderX

15. Poder TG, Bellemare CA, Bédard SK, Lemieux R. Social acceptance and population confidence in telehealth in Quebec. BMC Health Serv Res 2015 Feb 21;15:72. [doi: 10.1186/s12913-015-0727-1] [Medline: 25889230]
16. AshaRani PV, Jue Hua L, Roystonn K, et al. Readiness and acceptance of eHealth services for diabetes care in the general population: cross-sectional study. J Med Internet Res 2021 Sep 2;23(9):e26881. [doi: 10.2196/26881] [Medline: 34473062]
17. Kelaiditi E, Cesari M, Canevelli M, et al. Cognitive frailty: rational and definition from an (I.A.N.A./I.A.G.G.) international consensus group. J Nutr Health Aging 2013 Nov;17(9):726-734. [doi: 10.1007/s12603-013-0367-2]
18. Rand D, Sternberg SA, Gasner Winograd R, Buckman Z, Bentur N. The contribution of frailty to participation of older adults. Int J Environ Res Public Health 2022 Jan 30;19(3):1616. [doi: 10.3390/ijerph19031616] [Medline: 35162637]
19. Law M. Participation in the occupations of everyday life. Am J Occup Ther 2002 Nov 1;56(6):640-649. [doi: 10.5014/ajot.56.6.640]
20. Drummond K, Lambert G, Tahasildar B, Carli F. Successes and challenges of implementing teleprehabilitation for onco-surgical candidates and patients' experience: a retrospective pilot-cohort study. Sci Rep 2022;12(1):6775. [doi: 10.1038/s41598-022-10810-y]
21. Wu F, Laza-Cagigas R, Rampal T. Understanding patients' experiences and perspectives of tele-prehabilitation: a qualitative study to inform service design and delivery. Clin Pract 2022 Aug 16;12(4):640-652. [doi: 10.3390/clinpract12040067] [Medline: 36005070]

Abbreviations

6MWT: 6-minute walk test
HADS: Hospital Anxiety and Depression Scale
STS: 30-second sit-to-stand test
TUG: timed up and go test

Edited by A Mavragani; submitted 17.07.24; peer-reviewed by J Khan, M Gasmi; revised version received 18.11.24; accepted 18.11.24; published 20.01.25.

Please cite as:
Zhang F, Bang D, Visperas CA, Tun MH, Tay SS
Feasibility, User Acceptance, and Outcomes of Using a Cancer Prehabilitation App for Exercise: Pilot Cohort Study
JMIR Form Res 2025;9:e64427
URL: https://formative.jmir.org/2025/1/e64427
doi:10.2196/64427

© Fuquan Zhang, Deepali Bang, Christine Alejandro Visperas, Mon Hnin Tun, San San Tay. Originally published in JMIR Formative Research (https://formative.jmir.org), 20.1.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Longitudinal Changes in Pitch-Related Acoustic Characteristics of the Voice Throughout the Menstrual Cycle: Observational Study

Jaycee Kaufman[1], MS; Jouhyun Jeon[1], PhD; Jessica Oreskovic[1], MASc; Anirudh Thommandram[1], MASc; Yan Fossat[1,2], MS

[1]Klick Applied Sciences, Klick Health, Toronto, ON, Canada

[2]Faculty of Science, University of Ontario Institute of Technology, Oshawa, ON, Canada

Corresponding Author:
Yan Fossat, MS
Klick Applied Sciences, Klick Health, Toronto, ON, Canada

Abstract

Background: Identifying subtle changes in the menstrual cycle is crucial for effective fertility tracking and understanding reproductive health.

Objective: The aim of the study is to explore how fundamental frequency features vary between menstrual phases using daily voice recordings.

Methods: This study analyzed smartphone-collected voice recordings from 16 naturally cycling female participants, collected every day for 1 full menstrual cycle. Fundamental frequency features (mean, SD, 5th percentile, and 95th percentile) were extracted from each voice recording. Ovulation was estimated using luteinizing hormone urine tests taken every morning. The analysis included comparisons of these features between the follicular and luteal phases and the application of changepoint detection algorithms to assess changes and pinpoint the day in which the shifts in vocal pitch occur.

Results: The fundamental frequency SD was 9.0% (SD 2.9%) lower in the luteal phase compared to the follicular phase (95% CI 3.4% - 14.7%; *P*=.002), and the 5th percentile of the fundamental frequency was 8.8% (SD 3.6%) higher (95% CI 1.7% - 16.0%; *P*=.01). No significant differences were found between phases in mean fundamental frequency or the 95th percentile of the fundamental frequency (*P*=.65 and *P*=.07). Changepoint detection, applied separately to each feature, identified the point in time when vocal frequency behaviors shifted. For the fundamental frequency SD and 5th percentile, 81% (n=13) of participants exhibited shifts within the fertile window (*P*=.03). In comparison, only 63% (n=10; *P*=.24) and 50% (n=8; *P*=.50) of participants had shifts in the fertile window for the mean and 95th percentile of the fundamental frequency, respectively.

Conclusions: These findings indicate that subtle variations in vocal pitch may reflect changes associated with the menstrual cycle, suggesting the potential for developing a noninvasive and convenient method for monitoring reproductive health. Changepoint detection may provide a promising avenue for future work in longitudinal fertility analysis.

(JMIR Form Res 2025;9:e65448) doi:10.2196/65448

KEYWORDS

menstrual cycle; women's health; voice; acoustic analysis; longitudinal observational study; fertility tracking; fertility; reproductive health; feasibility; voice recording; vocal pitch; follicular; luteal phase; fertility status; mobile phone

Introduction

The menstrual cycle is a recurring sequence of physiological changes regulated by hormonal fluctuations that prepare the body for potential pregnancy. This cycle is crucial for women's health, as it influences reproductive function, affects various aspects of physical and emotional well-being, and serves as an indicator of overall health. Irregularities or disruptions in the menstrual cycle can signal underlying health issues, making its regular monitoring and understanding essential for women's health care. The female menstrual cycle is broken into 2 distinct phases: the follicular phase, in which hormonal estrogen levels rise, and the luteal phase, in which estrogen levels begin to decrease, and progesterone is elevated. An important step in the evaluation and management of female fertility and pregnancy is awareness of the timing of ovulation, the release of an egg from the ovary for conception. Ovulation occurs between the follicular and luteal phases and is initiated by high estrogen and luteinizing hormone (LH). The fertile window, or the time during the menstrual cycle when conception is most likely, encompasses the days just before and during ovulation. Follicle tracking through ultrasound is a common fertility procedure in a clinical setting and is considered an accurate method to evaluate the true ovulation timing. However, it comes with significant patient burden and cost. Currently, ovulation detection can be estimated at home through tracking the length

XSL•FO
RenderX

of time between periods (calendar method) and assessing the day-to-day changes of cervical mucus, basal body temperature (BBT), and estrogen, progesterone and LH measurements [1].

The current at-home ovulation detection methods come with their own drawbacks. Hormone measurements and BBT recordings require the purchase of additional devices and materials. Cervical mucus monitoring requires proper education on the different appearances and consistencies throughout the menstrual cycle. Perhaps, the easiest and most accessible method, the calendar method, is widely available through various menstrual tracking apps [2]. Both the predicted day of ovulation and the estimated fertile window can be provided; however, the variation in ovulatory days is distinct. It was shown that the most common predicted fertile window days only included the day of ovulation itself for 65% of female participants [2], yielding some uncertainty in the reliability of this method.

An interesting supplement to the calendar method may be achieved via the assessment of voice throughout the menstrual cycle. Increased estrogen, such as what is seen immediately preceding ovulation, leads to increased mucus secretion by the granular glands in the respiratory tract and increased cell permeability. Conversely, laryngeal tissues begin to absorb water, as estrogen levels decrease in the luteal phase, ultimately causing mucosal edema, vascular congestion, and increased vocal fold mass [3,4]. These changes have been associated with decreased vocal range, vocal and pitch instability, vocal fatigue, and reduced power and efficiency in the voice [5,6]. As such, voice-based fertility analysis may capture physiological changes throughout the menstrual cycle, with the added benefit of convenient smartphone-based data collection, similar to the accessibility offered by the calendar method.

While assessing prior work, it is important to note the variation in methodology and subsequent results. An early study evaluated speech recordings of the vowel sounds /a/, /i/, and /u/ from 20 undergraduate female participants. Voice was recorded twice throughout the menstrual—once at ovulation and once 2 - 3 days prior to the onset of menses. There was no difference in hoarseness or fundamental frequency (F0) between the 2 recordings, a result that was confirmed with a validation sample of 27 additional undergraduate female participants [7]. Similar methodology (recording once during menstruation, immediately following menstruation, midcycle, and premenstruation) with vowel sounds [8,9], counting from 1 to 10 [10], and free speech [11] also yielded no significant changes in F0. Conversely, a study conducted with 30 participants recording the vowel sound /a/ once a week for 4 weeks reported a significant increase in F0 in the ovulatory period and a decrease in F0 in the luteal period compared to the other recording time points [12], and high-fertility recordings of a spoken introductory sentence from 69 participants were significantly higher in F0 than in recordings obtained in a low-fertility period [8]. Looking at other acoustic parameters, counting from 1 to 10 yielded significant differences in shimmer (sound wave amplitude variation) and frequency variation between periods of high fertility and low fertility [10,13], and the vowel sounds /a/, /e/, /i/, /o/, and /u/ collected once each within the fertile window, luteal phase, and during menstruation from 44 participants indicated an increased minimum F0 in the fertile window [14]. In addition, a study conducted on 23 female participants recording 1 minute of free speech every day for 1 menstrual cycle showed marginally significant F0 variation with an increase prior to and a distinct drop during ovulation [15]. Furthermore, a study conducted on 17 naturally cycling female participants reported that female voices were rated more attractive in periods of high fertility when tasked with a counting task at 4 equally spaced time points throughout their menstrual cycle [16].

One of the drawbacks of these studies is that, apart from a single study [15], voice recordings were collected at a single time point within a menstrual phase; thus, day-to-day changes within the voice were notably not assessed. Ovulation itself is a result of longitudinal changes in hormone values compared to the previous days. Therefore, it is essential to trace day-to-day changes of acoustic characteristics to understand the effect of ovulation on voice signals.

In this study, we examined longitudinal changes in pitch-related acoustic characteristics using daily smartphone recordings of spoken sentences from 16 healthy female participants. Our overall goal was to determine how these vocal features change in relation to the LH surge and ovulation. We hypothesize that there are temporal changes in the F0 features associated with the menstrual cycle. The analysis is thus split into two sections:

1. Replication of the day-to-day voice change analysis conducted by Fischer et al [15] using smartphone-recorded (real-world) voice segments. The feature set will be expanded from F0 and F0 variation to include metrics denoting the minimum and maximum F0 values within a recording.
2. Employment of changepoint detection, which will be used to understand the statistical moment in which the participant voices change from one behavior to another. We hypothesize that for features significantly different between the follicular and luteal phases, this day will align with the switch between menstrual phases (ovulation). It may also provide insight into how voice may be used as a fertility indicator in future work.

Methods

Study Design and Participants

In total, 19 English-speaking, cis-gender female participants were recruited from Klick Health, a technology, media, and research company in the health care sector based in Toronto, Canada. Recruitment information was posted on the company's internal communication platform, and interested parties contacted our research team privately for more information and to sign the informed consent. After the informed consent was signed, participants were onboarded through either a video call or in person. As part of the onboarding, participants were provided instructions on how to conduct the urine tests, take their BBT, and record their voice through the research app.

All participants reported a natural menstrual cycle without taking hormonal birth control or medication that would affect the menstrual cycle and reported their menstrual cycle as consistent (1 - 2 days variation per cycle) or somewhat consistent (3 - 4

XSL•FO
RenderX

days variation per cycle). Participants were instructed to take their BBT, self-administer a consumer-grade LH urine test (Easy@Home Ovulation Tests; Easy@Home Fertility), and record themselves saying "Hello, how are you?" every morning immediately upon waking for 1 full menstrual cycle. The decision to use a fixed phrase was made based on the success of previous studies in using fixed sentences and conversational speech to identify significant features between menstrual phases compared to vowel sounds [3,8,15]. Participants were instructed to record their voices in a quiet environment with no other voices present in the recording and took pictures of the LH test strips each day to record hormone test results.

The data collection started the day after the end of menstruation, and recordings were collected for 1 full menstrual cycle, ending on the last day of menstruation. Participants did not perform the hormone test on days they were menstruating but did record their voice and BBT. Data (LH test results, BBT, and voice recordings) were collected using a custom mobile app developed by Klick Applied Science that participants installed on their personal cellphones following the procedure below:

1. Participants entered their participant ID (a deidentified alphanumeric code) upon their first login to the mobile app.
2. When the app was opened, the participant selected "Record a New Entry" to input the study data for the day.
3. They then recorded the results of their BBT (to 2 decimal places), their LH ovulation test results, and their voice while holding their phone.
4. After all fields were completed, the participants could submit the data recording. Submitted data were uploaded to a secure Google Firestore database, where it could be accessed by our researchers via a private application programming interface key.

Participants were excluded if the LH peak could not be detected, either through multiple consecutive missed data recordings (n=2) or if no discernable LH rise was detected through hormone tests (n=1). In total, 16 participants were included with a total of 320 voice recordings.

Ethical Considerations

Ethics was obtained from the Canadian SHIELD Ethics Review Board (REB tracking 2022-02-004). Informed consent was collected for each participant, and all procedures corresponded to local guidelines. Participant data were deidentified during collection. Participants were compensated US $77 for participating in the study and received an additional US $38.50 if they recorded all the voice recordings.

Longitudinal Landscape of Voice Signals

Individual participants recorded their voices once a day, every day for 1 menstrual cycle. The average menstrual cycle length was 28.81 (SD 2.90) days. The average length of the follicular phase (including menstruation and the day of the LH surge and the day after) was 15.92 (SD 3.30) days, and the average length of the luteal phase was 12.75 (SD 1.44) days. On average, menstruation was 5.54 (SD 1.66) days and ended 9.44 (SD 3.12) days prior to the LH surge. In total, 320 voice recordings were collected. Study participants had an average age of 30.6 (SD 4.5) years and an average BMI of 22.7 (SD 2.8) kg/m^2 and included various ethnic backgrounds (East or Southeast Asian: n=4, 25%; Jewish: n=1, 6%; Middle Eastern: n=2, 13%; South Asian: n=6, 37%; White: n=3, 19%; self-identification).

Voice Feature Extraction and Analysis

The LH surge was defined to be when the hormone strip test line was at its maximum darkness out of all the days in the menstrual cycle and approximately equal to or greater than the darkness of the control line. The ovulation window was defined as the day of the LH peak and the day after (to cover the 24 - 36 hour likelihood for ovulation). Days prior to this were classified as "follicular," corresponding to the follicular phase in the menstrual cycle, and days following ovulation classification and occurring prior to menstruation were classified as "luteal." Recordings collected while participants were menstruating were labeled "menstrual." BBT was used to verify the likelihood that ovulation occurred through a sustained 0.3 °C increase in the days following the suspected ovulation day.

Each audio recording was assessed by researchers to ensure adequate recording quality and the absence of background noise. In total, 4 vocal features associated with the recording F0 were extracted from each voice recording: the mean (F0AV), SD (F0SD), 5th percentile (F0MIN), and 95th percentile (F0MAX). The 5th and 95th percentiles were chosen as robust analogs to the minimum and maximum F0 values. Voice features were extracted using librosa [17], a Python library widely used for music and audio analysis, providing tools for audio feature extraction and processing tasks. F0 values for the recording were extracted using the *pyin* function, with the specified frequency minimum of C2 (~65 Hz) and maximum of C7 (~2093 Hz). These values were chosen to encompass a wide range of vocal frequencies and are the recommended minimum and maximum frequencies for analysis [17].

Each voice feature was scaled independently for each participant. Each voice feature was minimum-maximum scaled according to the following formula:

(1)Xsc=X-XminXmax-Xmin

Where X_{sc} is the scaled feature value, X is the original feature value, X_{min} is the minimum daily value of the feature throughout the participant's menstrual cycle, and X_{max} is the maximum daily value of the feature throughout the participant's menstrual cycle. This scaling was conducted such that the maximum scaled feature value throughout the recording period was 1 and the minimum value was 0 (Figure 1). This was repeated for each feature, for each participant.

XSL•FO
RenderX

Figure 1. Generated mean fundamental frequency (F0AV) data to demonstrate the feature scaling process. Data are scaled for each feature within each participant, such that the daily values of a participant's feature fall between 0 and 1.

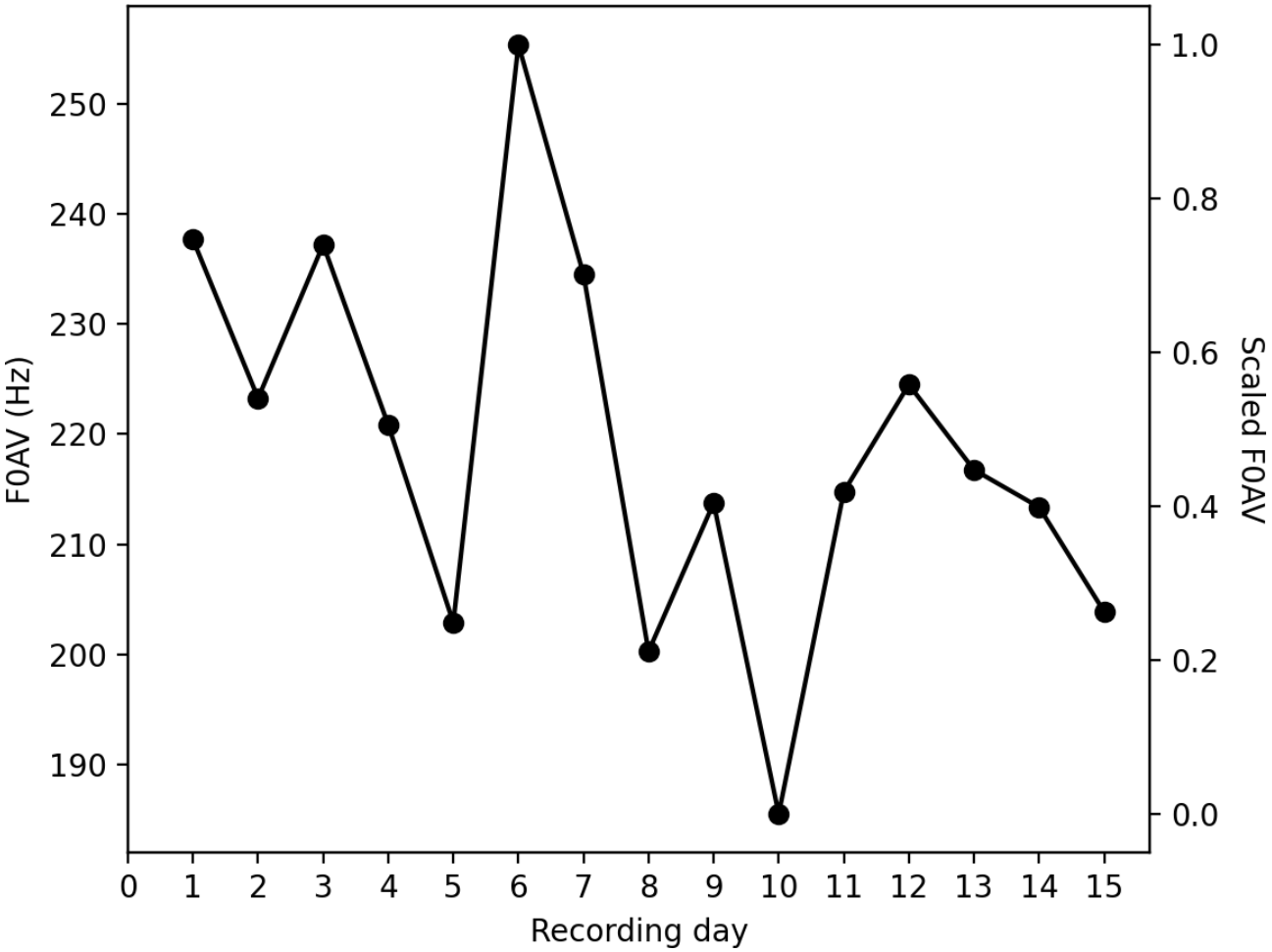

Changepoint Detection

To assess the change in voice features throughout the menstrual cycle, we used changepoint detection. Changepoint detection algorithms are tools used to identify points in time where the statistical properties of a data sequence change. These changes can signify shifts in mean, variance, distribution, or other characteristics of the data and can be used to detect shifts in patient health status or the effectiveness of treatments over time [18-23]. Changepoint detection algorithms are especially valuable for identifying transitions between distinct stages [23], making them a promising tool for research on physiological changes throughout the menstrual cycle. In this study, we used the *Ruptures* package, a Python library designed for offline changepoint detection, which considers all data points within a time series to identify underlying statistical changes in the data [24]. This unsupervised algorithm does not use prior knowledge of ovulation days and is not influenced by other participants' data, ensuring that the detection process only identifies changes in the voice recording F0 data themselves.

We applied 3 changepoint detection algorithms: dynamic programming, binary segmentation, and bottom-up segmentation. Dynamic programming explores all possible partitions of the time series to identify the optimal segmentation, constrained to detect only a single changepoint. Binary segmentation iteratively identifies the most significant changepoint in each segment but, for this study, was configured to stop after detecting the first changepoint. Bottom-up segmentation begins with many small segments and merges the segments that are most similar until only 1 changepoint remains, meeting the defined stopping criterion.

To quantify changes, we used 3 cost functions: least absolute deviation loss function (L1), least squared deviation loss function (L2), and radial basis function. The least absolute deviation cost function measures absolute differences, making it robust to outliers. The least squared deviation cost function emphasizes larger deviations by using squared differences. The radial basis function cost function captures nonlinear changes by comparing segments with a radial basis kernel. Detailed descriptions of these algorithms and cost functions are available in the *Ruptures* documentation [24].

We were not aiming to identify changes associated with menstruation so we specified the detection of a single changepoint within a window spanning from 9 days before to 9 days after the LH surge. These time points correspond to the average number of days we observed occurring prior to the LH surge following menstruation (9.4, SD 3.1 days) and the postovulatory day that the corpus luteum begins to disintegrate and hormone levels fall [25]. The algorithm was configured with a minimum distance of 3 days between any 2 breakpoints or from the start of the time series, with a jump parameter set

XSL•FO
RenderX

to 1 to ensure that each day was evaluated for potential changepoints of the vocal features between menstrual phases.

Statistical Analysis

To analyze differences between the follicular and luteal phases, 4 generalized linear mixed models were applied for the 4 F0 features. The voice feature was the dependent variable, the menstrual phase (follicular vs luteal) was the independent variable, and participant ID was a random effect to account for intraindividual variability. To avoid data that may occur between the follicular and luteal phases and ensure the clarity of phase-specific effects, days identified as ovulatory (the day of LH surge and the following day) and menstrual days were excluded from the analysis. All other data points were used. The models were estimated using restricted maximum likelihood.

To assess the significance of our changepoint findings against a null hypothesis of random results, we conducted 1-sided 2-proportion *z* tests. We compared the observed proportion of participants with a changepoint occurring before ovulation in the fertile window (days −5 to +1 relative to LH surge) to a random expectation, corresponding to 50% (n=8) of participants with a changepoint occurring in the fertile window. This proportion is based on 7 days in the fertile window of an average of 14 days in the possible prediction window (6 days prior to the LH surge maximum, the LH surge maximum, and 7 days following the LH surge maximum). The alternate hypothesis posited that the observed proportion would be higher than the random expectation due to the hormonal changes associated with the fertile window. All statistical analysis was performed in Python (version 3.12.2; Python Software Foundation) using the *statsmodels* package (*mixedlm* and *proportions_ztest* functions) [26].

Results

Results 1: F0 Behavior Surrounding LH Surge

To assess vocal dynamics consistent across participants, generalized linear mixed models were fit for each of the F0 features (F0AV, F0SD, F0MIN, and F0MAX). Overall, there was a significant difference in the F0SD and F0MIN between the follicular and luteal recording days for each participant, such that F0SD was, on average, 9.0% (SD 2.9%) lower in the luteal phase compared to the follicular phase (95% CI 3.4% - 14.7%; P=.002) and F0MIN was 8.8% (SD 3.6%) higher (95% CI 1.7% - 16.0%; P=.01). There were no significant differences between phases in F0AV or F0MAX (P=.65 and P=.07). Average daily trends of the participant-scaled F0 values are displayed in Figure 2.

XSL•FO
RenderX

Figure 2. Daily patterns of fundamental frequency (**F0**) voice features—F0 average (F0AV), F0 standard deviation (F0SD), F0 5th percentile (F0MIN), and F0 95th percentile (F0MAX)—in the 9 days preceding and following the luteinizing hormone (LH) surge for all participants. Data points represent the average values of these voice features across 16 healthy female participants with regular menstrual cycles. Error bars indicate the SE of the mean.

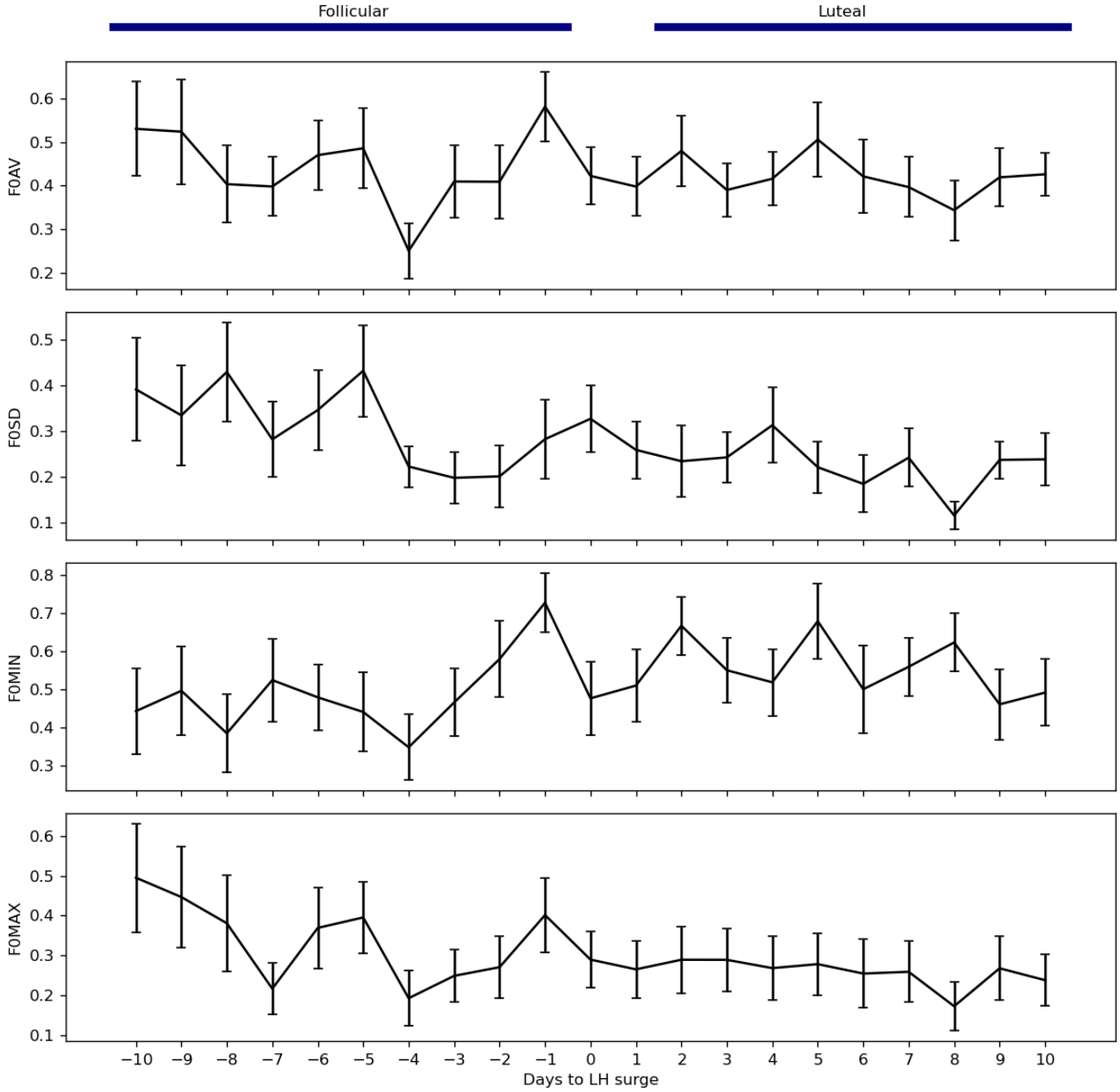

Results 2: F0 Changepoint Detection

F0 feature data for each participant was fed into the changepoint detection algorithms, and the predicted changepoints are displayed in Figure 3. Menstrual phases were not added as information in the algorithms at any point, and any changepoints detected solely reflect changes in the vocal features themselves. Additionally, data from each participant were used independently for changepoint identification, ensuring that data were not shared or mixed between participants. Overall, the average (SD) detected changepoints for all algorithms were 0.20 (4.13) days following ovulation in F0AV, 1.42 (3.52) days prior to ovulation in F0SD, 0.88 (3.27) days prior to ovulation in F0MIN, and 1.31 (4.05) days following ovulation in F0MAX. As the variation in the predicted changepoints was large, we also looked where the predicted changepoints occurred specifically. We observed that up to 81% (n=13) of participants had a predicted changepoint occurring within the fertile window, or the 5 days preceding the LH surge to the 1 day following the LH surge, for both F0SD and F0MIN (P=.03). Meanwhile, a maximum of 63% (n=10; P=.24) and 50% (n=8; P=.50) of participants had changepoints detected in the fertile window for F0AV and F0MAX, respectively (Figure 3).

XSL•FO
RenderX

Figure 3. The percentage of participants with detected changepoints in the fertile window is presented for the mean, SD, 5th percentile, and 95th percentile of the fundamental frequency (F0AV, F0SD, F0MIN, and F0MAX, respectively). Changepoints were identified using 3 different detection algorithms—binary segmentation (BS), bottom-up segmentation (BU), and dynamic programming (DP)—applied to fundamental frequency features extracted from daily smartphone voice recordings of 16 naturally cycling female participants. The fertile window is defined as the 5 days leading up to and including 1 day after the luteinizing hormone surge. L1: least absolute deviation loss function; L2: least squared deviation loss function; RBF: radial basis function.

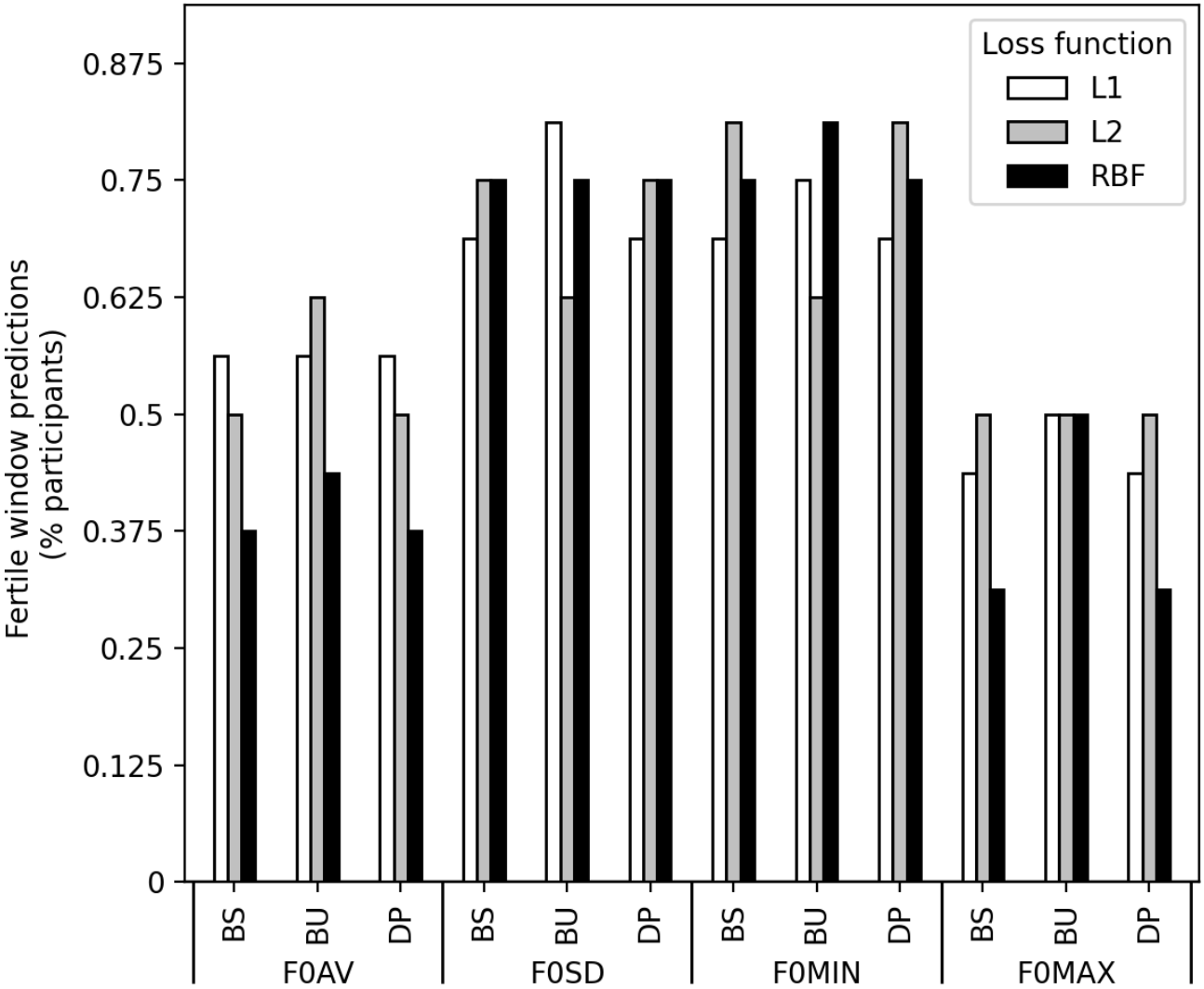

Discussion

Primary Findings

This study explored how vocal pitch features fluctuate across the menstrual cycle, with a focus on vocal patterns surrounding phase transitions and ovulation. Building on prior research [15], we expanded the analysis by examining a broader set of F0-related features, including minimum and maximum pitch values, using real-world voice recordings collected daily via smartphones. With more frequent recordings, we investigated whether these vocal features varied between the follicular and luteal phases, providing a more comprehensive perspective on daily vocal behavior. We further hypothesized that changepoint detection could pinpoint the transition between these phases, offering insight into the potential for voice-based markers of fertility in future applications.

Physiologically, the differences in voice features between menstrual phases may be attributed to differences in hormone concentrations. In the follicular phase, estrogen levels rise, whereas the luteal phase is characterized by increased progesterone levels. The larynx itself is a sexual organ and can be influenced by reproductive hormones, as estrogen stimulates the mucosal cells. Mucus on the vocal cords has been shown to be similar in consistency and makeup as cervical mucus throughout the menstrual cycle, displaying increased thin, wet mucus in the follicular phase and thick, dry mucus in the luteal phase [27]. A thin, wet mucus layer is necessary for proper vocal cord vibration and range of pitch [28], and decreased vocal range has been associated with thicker, dry laryngeal mucus [27,29]. Our results align well with these depictions, as we observed a decrease in the F0SD and range (observed through an increase in the minimum F0) in the luteal phase. An alternative explanation for vocal changes may reside in more nuanced apps. Fatigue and stress may be exacerbated in the luteal phase [30], and increased fatigue has been associated with increased F0SD [31]. Regardless, changes in mental and hormonal states associated with the changing menstrual phases provide evidence for the observed vocal changes.

XSL•FO
RenderX

We found that individually scaled F0 voice features had distinct differences between the follicular and luteal phases. Similar to what was observed in previous studies, F0SD decreased and F0MIN increased in the luteal phase [14,15]. Meanwhile, the association between the F0AV and the menstrual phase was insignificant. F0AV throughout the menstrual cycle has traditionally reported mixed results, in which some authors found that F0 increased in periods of high fertility [8,32], decreased in high fertility [15,33,34], or did not change [6,9,14,35-39]. This difference may be attributed to different study methodologies, such as recording a vowel instead of a fixed phrase or only recording 1 voice sample per menstrual phase.

Changes in the voice surrounding the fertile period have been hypothesized to be the result of evolutionary processes. More attractive female participants have an increased likelihood of finding a suitable mate, and mating during the fertile window has the highest likelihood of conception [1]. Research has shown that variations in female voice characteristics throughout the menstrual cycle can influence perceived attractiveness. In the high fertility days prior to ovulation, female voices have typically been rated “more attractive” by heterosexual male participants [15,16]. It has been shown that 44% - 55% of the variance in the desirability scores for female participants was explained by their minimum F0 in speech. This may indicate that female voices with lower pitch minima, opposed to a change in the mean pitch, are more desirable [40]. Our findings align nicely with this result, as the 5th percentile of the pitch was an average of 8.8% lower in the follicular phase (a phase corresponding to higher fertility) compared to the luteal phase for all our participants. Furthermore, increased pitch variability, similar to what was seen in the F0SD in the follicular phase, has been associated with increased social attractiveness of the speaker [41,42].

One of the primary distinctions of our work is the employment of a changepoint detection algorithm to identify the day of the switch in vocal features within the menstrual cycle. Although the average predicted changepoints fell within 2 days of ovulation in all participants, there was a large variation in the values that would not permit accurate ovulation prediction. However, when assessing the changepoints in the broader sense of fertility identification, we note that changepoints are identified within the fertile window in a high proportion of participants. The fertile window is associated with an increased likelihood of conception, making it a meaningful target for detection, even if exact ovulation timing is not achieved [1]. In particular, we observed that changepoints are detected within the fertile window in up to 81% (n=13) of participants for F0SD and F0MIN. However, this result was not observed for F0AV and F0MAX.

Offline changepoint detection is a useful method for fertility monitoring, even without the real-time benefits of online detection (in which changepoints can be detected as the data are collected). Current fertility monitoring methods can also require daily use or cycle-long data collection to accurately pinpoint the timing of ovulation. For example, BBT is used as an ovulation confirmation, in which the BBT must be measured every day for at least 1 full menstrual cycle to identify ovulation [1]. After further refinement, offline detection used in a consumer product would have similar confirmatory utility.

Unlike previous studies that had limited voice data collection methodologies, our approach reflects real-world situations by using at-home smartphone recordings, allowing for a more comprehensive assessment of voice parameters throughout the menstrual cycle. This opens up avenues for exploring the feasibility of using voice parameters as indicators of high fertility and ovulation. Moreover, the simplicity and noninvasiveness of smartphone-based recording make it suitable for widespread use in both clinical settings and personal health monitoring. With just a voice and a smartphone app, users do not require any additional devices, highlighting the convenience and accessibility of this method.

Limitations

Our study’s insights into vocal pitch changes across the menstrual cycle are limited by a small sample size (n=16 participants), which may not capture the full spectrum of variability across a wider population. Additionally, analyzing recordings from only 1 menstrual cycle per participant neglects potential intraindividual variability across different cycles. Finally, our examination of a limited set of pitch features restricts our understanding of the nuanced relationship between vocal characteristics and menstrual phases, and there may be more information located in other pitch-related features. With further refinement through larger sample sizes, multicycle analyses, and a broader examination of vocal features, we could gain better insights into how vocal pitch correlates with menstrual cycle changes. Importantly, using larger, multicycle datasets could enable the use of online changepoint detection algorithms, which have the advantage of identifying phase transitions in real-time without requiring access to the entire dataset beforehand, or could be used to train a machine learning model for enhanced performance. With these advancements, this method could contribute to the early detection of fertility issues, personalized fertility tracking, and natural family planning.

Conclusions

Our study contributes to the evolving understanding of how menstrual cycle phases impact vocal pitch, underscoring the potential of voice as a noninvasive marker for fertility and ovulation. Using methods that use real-world data captured through smartphone recordings, our study analyzed these data with changepoint detection algorithms to identify transitions between menstrual phases. This approach not only confirms the relationship between vocal pitch variations and the menstrual cycle phase but also demonstrates the feasibility of applying such findings in practical, real-world settings. Our work could have implications for clinical and personal health monitoring, offering potential improvements in fertility tracking and natural family planning methods. Future efforts to integrate more sophisticated machine learning techniques and real-time analysis could further refine the accuracy and applicability of vocal pitch as a tool for women’s health management.

XSL•FO
RenderX

Acknowledgments

The authors thank all participants, project members, and researchers at Klick Health for the successful development, implementation, and evaluation of this research. This study was funded internally by Klick Health. The funder played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript.

Data Availability

The datasets generated and analyzed during this study are not publicly available to protect participant privacy. They may be available from the corresponding author on reasonable request, subject to ethics approval and compliance with data protection regulations.

Authors' Contributions

JK designed experiments and performed analyses. JK and JJ developed the method and prepared figures. JK, JJ, JO, and YF wrote the initial draft and edited the manuscript. JK and AT developed a custom mobile app and collected data. JJ and YF supervised the project. All authors reviewed the manuscript and approved it for submission.

Conflicts of Interest

All authors are employed by Klick Health, a technology, media, and research company in the health care sector based in Toronto, Canada. Authors JK and YF currently hold a provisional patent on the use of voice-based technology in the prediction of ovulation (Systems, Devices and Methods for Fertility Analysis Using Voice, United States Provisional Application 63/416,819, filed October 17, 2022). All authors declare no nonfinancial conflicts of interest.

References

1. Su HW, Yi YC, Wei TY, Chang TC, Cheng CM. Detection of ovulation, a review of currently available methods. Bioeng Transl Med 2017 Sep;2(3):238-246. [doi: 10.1002/btm2.10058] [Medline: 29313033]
2. Johnson S, Marriott L, Zinaman M. Can apps and calendar methods predict ovulation with accuracy? Curr Med Res Opin 2018 Sep;34(9):1587-1594. [doi: 10.1080/03007995.2018.1475348] [Medline: 29749274]
3. Haselton MG, Gildersleeve K. Human ovulation cues. Curr Opin Psychol 2016 Feb;7:120-125. [doi: 10.1016/j.copsyc.2015.08.020]
4. Haselton MG, Gildersleeve K. Can men detect ovulation? Curr Dir Psychol Sci 2011 Apr;20(2):87-92. [doi: 10.1177/0963721411402668]
5. Zamponi V, Mazzilli R, Mazzilli F, Fantini M. Effect of sex hormones on human voice physiology: from childhood to senescence. Hormones (Athens) 2021 Dec;20(4):691-696. [doi: 10.1007/s42000-021-00298-y] [Medline: 34046877]
6. Amir O, Biron-Shental T. The impact of hormonal fluctuations on female vocal folds. Curr Opin Otolaryngol Head Neck Surg 2004 Jun;12(3):180-184. [doi: 10.1097/01.moo.0000120304.58882.94] [Medline: 15167026]
7. Silverman EM, Zimmer CH. Effect of the menstrual cycle on voice quality. Arch Otolaryngol Head Neck Surg 1978 Jan 1;104(1):7-10. [doi: 10.1001/archotol.1978.00790010011003]
8. Bryant GA, Haselton MG. Vocal cues of ovulation in human females. Biol Lett 2009 Feb 23;5(1):12-15. [doi: 10.1098/rsbl.2008.0507]
9. Çelik Ö, Çelik A, Ateşpare A, et al. Voice and speech changes in various phases of menstrual cycle. J Voice 2013 Sep;27(5):622-626. [doi: 10.1016/j.jvoice.2013.02.006] [Medline: 23497797]
10. Shoup-Knox ML, Ostrander GM, Reimann GE, Pipitone RN. Fertility-dependent acoustic variation in women's voices previously shown to affect listener physiology and perception. Evol Psychol 2019;17(2):1474704919843103. [doi: 10.1177/1474704919843103] [Medline: 31023082]
11. Pavela Banai I, Burriss RP, Šimić N. Voice changes across the menstrual cycle in response to masculinized and feminized man and woman. Adapt Human Behav Physiol 2022 Jun;8(2):238-262. [doi: 10.1007/s40750-022-00190-y]
12. Narasimhan SV, Pooja M. Vocal changes in different phases of menstrual cycle: an evidence from the acoustic, cepstral, and spectral analysis. J Indian Speech Lang Hear Assoc 2022;36(1):1. [doi: 10.4103/jisha.jisha_5_22]
13. Santos SMS. Is there any correlation between voice and human ovulation? [Doctoral Dissertation]. : Universidade da Beira Interior, ProQuest; 2018.
14. Pavela Banai I. Voice in different phases of menstrual cycle among naturally cycling women and users of hormonal contraceptives. PLoS One 2017;12(8):e0183462. [doi: 10.1371/journal.pone.0183462] [Medline: 28829842]
15. Fischer J, Semple S, Fickenscher G, et al. Do women's voices provide cues of the likelihood of ovulation? The importance of sampling regime. PLoS One 2011;6(9):e24490. [doi: 10.1371/journal.pone.0024490] [Medline: 21957453]
16. Pipitone RN, Gallup GG Jr. Women's voice attractiveness varies across the menstrual cycle. Evol Hum Behav 2008 Jul;29(4):268-274. [doi: 10.1016/j.evolhumbehav.2008.02.001]

XSL•FO
RenderX

17. McFee B, McVicar M, Faronbi D, et al. librosa. Zenodo. URL: https://zenodo.org/records/11192913 [accessed 2024-11-30]
18. Yang P, Dumont G, Ansermino JM. Adaptive change detection in heart rate trend monitoring in anesthetized children. IEEE Trans Biomed Eng 2006 Nov;53(11):2211-2219. [doi: 10.1109/TBME.2006.877107] [Medline: 17073326]
19. Malladi R, Kalamangalam GP, Aazhang B. Online bayesian change point detection algorithms for segmentation of epileptic activity. Presented at: Asilomar Conference on Signals, Systems and Computers; Nov 3-6, 2013; Pacific Grove, CA, United States p. 1833-1837. [doi: 10.1109/ACSSC.2013.6810619]
20. Staudacher M, Telser S, Amann A, Hinterhuber H, Ritsch-Marte M. A new method for change-point detection developed for on-line analysis of the heart beat variability during sleep. Physica A Stat Mech Appl 2005 Apr;349(3-4):582-596. [doi: 10.1016/j.physa.2004.10.026]
21. Bosc M, Heitz F, Armspach JP, Namer I, Gounot D, Rumbach L. Automatic change detection in multimodal serial MRI: application to multiple sclerosis lesion evolution. Neuroimage 2003 Oct;20(2):643-656. [doi: 10.1016/S1053-8119(03)00406-3] [Medline: 14568441]
22. Kass-Hout TA, Xu Z, McMurray P, et al. Application of change point analysis to daily influenza-like illness emergency department visits. J Am Med Inform Assoc 2012;19(6):1075-1081. [doi: 10.1136/amiajnl-2011-000793] [Medline: 22759619]
23. Aminikhanghahi S, Cook DJ. A survey of methods for time series change point detection. Knowl Inf Syst 2017 May;51(2):339-367. [doi: 10.1007/s10115-016-0987-z] [Medline: 28603327]
24. Truong C, Oudre L, Vayatis N. Selective review of offline change point detection methods. Signal Processing 2020 Feb;167:107299. [doi: 10.1016/j.sigpro.2019.107299]
25. McCracken JA, Custer EE, Lamsa JC. Luteolysis: a neuroendocrine-mediated event. Physiol Rev 1999 Apr;79(2):263-323. [doi: 10.1152/physrev.1999.79.2.263] [Medline: 10221982]
26. Seabold S, Perktold J. Statsmodels: econometric and statistical modeling with python. Presented at: Proceedings of the 9th Python in Science Conference; Jun 28 to Jul 3, 2010; Austin, TX, United States. [doi: 10.25080/Majora-92bf1922-011]
27. Abitbol J, Abitbol P, Abitbol B. Sex hormones and the female voice. J Voice 1999 Sep;13(3):424-446. [doi: 10.1016/s0892-1997(99)80048-4] [Medline: 10498059]
28. Levendoski EE, Leydon C, Thibeault SL. Vocal fold epithelial barrier in health and injury: a research review. J Speech Lang Hear Res 2014 Oct;57(5):1679-1691. [doi: 10.1044/2014_JSLHR-S-13-0283] [Medline: 24686981]
29. Bonilha HS, White L, Kuckhahn K, Gerlach TT, Deliyski DD. Vocal fold mucus aggregation in persons with voice disorders. J Commun Disord 2012;45(4):304-311. [doi: 10.1016/j.jcomdis.2012.03.001] [Medline: 22510352]
30. Li SH, Lloyd AR, Graham BM. Physical and mental fatigue across the menstrual cycle in women with and without generalised anxiety disorder. Horm Behav 2020 Feb;118:104667. [doi: 10.1016/j.yhbeh.2019.104667] [Medline: 31899259]
31. Vogel AP, Fletcher J, Maruff P. Acoustic analysis of the effects of sustained wakefulness on speech. J Acoust Soc Am 2010 Dec;128(6):3747-3756. [doi: 10.1121/1.3506349] [Medline: 21218906]
32. Tatar EC, Sahin M, Demiral D, et al. Normative values of voice analysis parameters with respect to menstrual cycle in healthy adult Turkish women. J Voice 2016 May;30(3):322-328. [doi: 10.1016/j.jvoice.2015.04.014] [Medline: 26003887]
33. Amir O, Kishon-Rabin L, Muchnik C. The effect of oral contraceptives on voice: preliminary observations. J Voice 2002 Jun;16(2):267-273. [doi: 10.1016/s0892-1997(02)00096-6] [Medline: 12150379]
34. Karthikeyan S, Locke JL. Men's evaluation of women's speech in a simulated dating context: effects of female fertility on vocal pitch and attractiveness. Evol Behav Sci 2015;9(1):55-67. [doi: 10.1037/ebs0000014]
35. Barnes L, Latman N. Acoustic measure of hormone affect on female voice during menstruation. Int J Humanit Soc Sci 2011;1(3):5-10 [FREE Full text]
36. Chae SW, Choi G, Kang HJ, Choi JO, Jin SM. Clinical analysis of voice change as a parameter of premenstrual syndrome. J Voice 2001 Jun;15(2):278-283. [doi: 10.1016/S0892-1997(01)00028-5] [Medline: 11411481]
37. Figueiredo LCD, Gonçalves MIR, Pontes A, Pontes P. Vocal behavior during menstrual cycle: perceptual-auditory, acoustic and self-perception analysis. Rev Bras Otorrinolaringol 2004;70:331-339. [doi: 10.1590/S0034-72992004000300008]
38. Meurer EM, Garcez V, von Eye Corleta H, Capp E. Menstrual cycle influences on voice and speech in adolescent females. J Voice 2009 Jan;23(1):109-113. [doi: 10.1016/j.jvoice.2007.03.001] [Medline: 17981011]
39. Raj A, Gupta B, Chowdhury A, Chadha S. A study of voice changes in various phases of menstrual cycle and in postmenopausal women. J Voice 2010 May;24(3):363-368. [doi: 10.1016/j.jvoice.2008.10.005]
40. Pisanski K, Oleszkiewicz A, Plachetka J, Gmiterek M, Reby D. Voice pitch modulation in human mate choice. Proc Biol Sci 2018 Dec 19;285(1893):20181634. [doi: 10.1098/rspb.2018.1634] [Medline: 30963886]
41. Ray GB, Ray EB, Zahn CJ. Speech behavior and social evaluation: an examination of medical messages. Commun Q 1991 Mar;39(2):119-129. [doi: 10.1080/01463379109369790]
42. Zuckerman M, Miyake K. The attractive voice: what makes it so? J Nonverbal Behav 1993;17(2):119-135. [doi: 10.1007/BF01001960]

Abbreviations

BBT: basal body temperature
F0: fundamental frequency

XSL•FO
RenderX

F0AV: mean fundamental frequency
F0MAX: 95th percentile of the fundamental frequency
F0MIN: 5th percentile of the fundamental frequency
F0SD: fundamental frequency SD
LH: luteinizing hormone

Edited by A Mavragani; submitted 16.08.24; peer-reviewed by J Grieger, V Martin; revised version received 22.10.24; accepted 13.11.24; published 09.01.25.

Please cite as:
Kaufman J, Jeon J, Oreskovic J, Thommandram A, Fossat Y
Longitudinal Changes in Pitch-Related Acoustic Characteristics of the Voice Throughout the Menstrual Cycle: Observational Study
JMIR Form Res 2025;9:e65448
URL: https://formative.jmir.org/2025/1/e65448
doi:10.2196/65448

© Jaycee Kaufman, Jouhyun Jeon, Jessica Oreskovic, Anirudh Thommandram, Yan Fossat. Originally published in JMIR Formative Research (https://formative.jmir.org), 9.1.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Postpartum Remote Health Coaching Intervention for Individuals With a Hypertensive Disorder of Pregnancy: Proof-of-Concept Study

Jaclyn D Borrowman[1,2], PhD; Lucas J Carr[2], PhD; Gary L Pierce[2], PhD; William T Story[3], MPH, PhD; Bethany Barone Gibbs[4], PhD; Kara M Whitaker[2,5], MPH, PhD

[1]Department of Preventative Medicine, Feinberg School of Medicine, Northwestern University, Chicago, IL, United States
[2]Department of Health and Human Physiology, University of Iowa, Iowa City, IA, United States
[3]Department of Community and Behavioral Health, University of Iowa, Iowa City, IA, United States
[4]Department of Epidemiology and Biostatistics, West Virginia University, Morgantown, WV, United States
[5]Department of Epidemiology, University of Iowa, Iowa City, IA, United States

Corresponding Author:
Jaclyn D Borrowman, PhD
Department of Preventative Medicine, Feinberg School of Medicine, Northwestern University, Chicago, IL, United States

Abstract

Background: Cardiovascular disease (CVD) is the leading cause of death among women in America. Hypertensive disorders of pregnancy (HDP) negatively impact acute and long-term cardiovascular health, with approximately 16% of all pregnancies affected. With CVD 2‑4 times more likely after HDP compared to normotensive pregnancies, effective interventions to promote cardiovascular health are imperative.

Objective: With postpartum physical activity (PA) interventions after HDP as an underexplored preventative strategy, we aimed in this study to assess (1) the feasibility and acceptability of a remotely delivered PA intervention for individuals with HDP 3‑6 months postpartum and (2) changes in average steps per day, skills related to PA behavior, and postpartum blood pressure (BP).

Methods: A remotely delivered 14-week health coaching intervention was designed based on prior formative work. The health coaching intervention called the Hypertensive Disorders of Pregnancy Postpartum Exercise (HyPE) intervention was tested for feasibility and acceptability with a single-arm proof-of-concept study design. A total of 19 women who were 3‑6 months postpartum HDP; currently inactive; 18 years of age or older; resided in Iowa; and without diabetes, kidney disease, and CVD were enrolled. Feasibility was assessed by the number of sessions attended and acceptability by self-reported satisfaction with the program. Changes in steps achieved per day were measured with an activPAL4 micro, PA behavior skills via validated surveys online, and BP was assessed remotely with a research-grade Omron Series 5 (Omron Corporation) BP monitor.

Results: Participants at enrollment were on average 30.3 years of age, 4.1 months postpartum, self-identified as non-Hispanic White (14/17, 82%), in a committed relationship (16/17, 94%), and had a bachelor's degree (9/17, 53%). A total of 140 of 152 possible health coaching sessions were attended by those who started the intervention (n=19, 92%). Intervention completers (n=17) indicated they were satisfied with the program (n=17, 100%) and would recommend it to others (n=17, 100%). No significant changes in activPAL measured steps were observed from pre- to posttesting (mean 138.40, SD 129.40 steps/day; P=.75). Significant improvements were observed in PA behavior skills including planning (mean 5.35, SD 4.97 vs mean 15.06, SD 3.09; P<.001) and monitoring of PA levels (mean 7.29, SD 3.44 vs mean 13.00, SD 2.45; P<.001). No significant decreases were observed for systolic (mean –1.28, SD 3.59 mm Hg; Hedges g=–0.26; P=.16) and diastolic BP (mean –1.80, SD 5.03 mm Hg; Hedges g=–0.44; P=.12).

Conclusions: While PA behaviors did not change, the intervention was found to be feasible and acceptable among this sample of at-risk women. After additional refinement, the intervention should be retested among a larger, more diverse, and less physically active sample.

Trial Registration: ClinicalTrials.gov NCT06019715; https://clinicaltrials.gov/study/NCT06019715

(JMIR Form Res 2025;9:e65611) doi:10.2196/65611

KEYWORDS

cardiovascular disease; postpartum; hypertensive disorders of pregnancy; intervention; physical activity intervention; proof-of-concept; cardiovascular health; CVD risk; cardiovascular disease risk; feasibility; acceptability; health coaching; women's health; postnatal

XSL•FO
RenderX

Introduction

Cardiovascular disease (CVD) is the leading cause of death among women in America; with 1 in 4 deaths being preventable [1]. Hypertensive disorders of pregnancy (HDP) are associated with a 2 - 4 times greater risk of developing CVD later in life [1,2]. Despite this increased risk for CVD and with 16% of all pregnancies affected, few studies have explored potential interventions for this population after delivery [3]. Consequently, it is imperative that preventative strategies are developed and started as early as possible after delivery to mitigate the risk of future CVD and improve maternal morbidity and mortality.

Cardiovascular changes during pregnancy act as a natural stress test and may uncover an individual's potential risk of future CVD [4]. Specifically, individuals with HDP are at an immediately increased risk for hypertension (HTN) after delivery [4]. Long-term cardiometabolic health is also impacted by HDP, with an increased risk of future CVD, morbidity, and mortality compared to normotensive pregnancies [5]. The odds of developing HTN at 1 year postpartum are 5 to 7 times greater in hypertensive pregnancies compared to normotensive pregnancies [2,6]. The 5-year likelihood of developing HTN for pregnancies complicated by HDP is 7.1 times greater than for normotensive pregnancies [4]. Even in those not diagnosed with HTN, average blood pressure (BP) is greater in HDP women than in women with normotensive pregnancies.

Low physical activity (PA) is a known modifiable risk factor for CVD, and increasing PA has the potential to improve BP [7]. PA, specifically aerobic exercise, has beneficial effects on mean, systolic BP (SBP), and diastolic BP (DBP) in adults [8,9]. This effect has been observed in pregnant populations as well, with the greatest improvements seen among inactive individuals [10,11]. Despite the known benefits of PA, recent studies suggest moderate to vigorous PA decreases after pregnancy compared to during pregnancy activity levels [12]. The perinatal period, therefore, presents a critical opportunity for behavior change [13-15]. Interventions promoting PA in the postpartum period after HDP are an underexplored primary prevention strategy warranting further investigation.

The main objective of our study was to assess the feasibility and acceptability of a remotely delivered PA intervention for individuals with HDP during the postpartum period. We hypothesized the intervention would be highly feasible and acceptable. Our secondary objectives included assessing changes in average steps per day, skills related to PA behavior, and postpartum BP. We hypothesized that we would observe significant increases in steps, improvements in PA skills, and decreases in postpartum SBP and DBP.

Methods

Study Design

A remotely delivered PA intervention for individuals with HDP called the Hypertensive Disorders of Pregnancy Postpartum Exercise (HyPE) intervention, was tested with a single-arm quasi-experimental study design. The Obesity-Related Behavioral Intervention Trials (ORBIT) model was the conceptual framework that was leveraged in the development and subsequent testing of the HyPE intervention [16]. The ORBIT model guides the development of behavioral treatments to prevent or manage chronic disease. This model describes a flexible and iterative process that includes prespecified clinically significant milestones necessary for forward movement through the developmental pathway.

Phase I of the ORBIT model was previously completed by our group, where the design of this intervention was informed by an assessment of the PA determinants, needs, and desired components of a PA intervention of postpartum individuals after HDP [17]. Given phase I was completed successfully, our group subsequently developed the HyPE intervention for proof-of-concept testing (phase IIa of the ORBIT model). Study methods appropriate at this stage include a quasi-experimental, treatment-only design to determine whether the intervention is feasible, acceptable, and can achieve a clinically significant signal on the behavioral risk factor of interest (ie, PA).

Ethical Considerations

All methods and procedures were approved by an institutional review board prior to the start of the study (IRB #202302470; ClinicalTrials: NCT06019715). Potential participants had a call with a research staff member to review study details, ask questions, and provide informed consent if still interested in enrolling. All data were deidentified and secured. For compensation, participants were provided a Fitbit Inspire 2 (Google) and Omron Series 5 BP monitor. Participants received a US $20 Amazon gift card after completion of half of the study visits and a US $25 Amazon gift card after completion of the entire study.

Recruitment

Recruitment of currently inactive individuals (defined in this study as achieving <9000 steps on average per day) occurred via mass emails to faculty and staff members of a large midwestern university, targeted emails to participants in the Pregnancy 24/7 Cohort Study, and postpartum medical record searches of individuals who received care at local hospitals and clinics. If participants were interested in the study, the target email contained information, as well as a link to a REDcap (Vanderbilt University) survey for initial screening. If eligible, interested individuals were contacted via phone at their preferred time by a research staff member to describe the study in further detail and answer questions. Subsequently, potential participants were provided the opportunity for participants to provide informed consent if still interested in the study.

Following informed consent, participants started baseline screen procedures involving the assessment of their current PA levels. The threshold of inactivity was set at <9000 steps per day based on a prior study conducted by our group and concerns regarding recruitment within a restricted timeframe [18]. Screening of baseline PA levels was conducted with an activPAL4 micro prior to participation in the first study visit.

Eligibility

To be included, participants were 3 - 6 months postpartum after a pregnancy complicated by HDP; 18 years of age or older; able

XSL•FO
RenderX

to speak, read, and write English; lived in Iowa; and owned a smartphone. Exclusions included current enrollment in a PA intervention; achieving >9000 steps at pretesting; inability to walk half block or 2 flights of stairs; physician's recommendation to limit PA; and current diagnosis and treatment of diabetes, kidney disease, and CVD. All self-report diagnoses of HDP were verified by medical records.

Intervention

The overall goal of the intervention was to increase the average number of steps participants achieved per day, with the specific objective to increase steps by approximately 10% between each session. The intervention consisted of a total of 8 online health coaching sessions, with sessions 1 - 5 occurring weekly (ie, the intensive intervention) and subsequent visits 6 - 8 occurring with reduced frequency (Figure 1). Participants were provided a Fitbit Inspire 2 to track their progress toward the step goals set with their health coach. Formative work conducted by our group informed the creation of tailored educational postpartum PA intervention, including the use of a health coach, remote delivery approach, selected frequency of study visits, and topics offered to study participants (eg, breastfeeding and PA).

Health coaching was informed by motivational interviewing techniques and conducted by a trained interventionist (principle investigator: JDB). This approach was developed to promote behavioral change and work through ambivalence by supporting an individual's commitment to change using "change talk" (ie, eliciting the client to express desires, ability, reasons, and need for a change of behavior) [19,20].

Figure 1. The Hypertensive Disorders of Pregnancy Postpartum Exercise (HyPE) intervention health coaching study timeline. The HyPE intervention was a 14-week health coaching intervention for postpartum individuals after having a pregnancy complicated by a hypertensive disorder. This intervention included 8 total health coaching session aiming to promote physical activity.

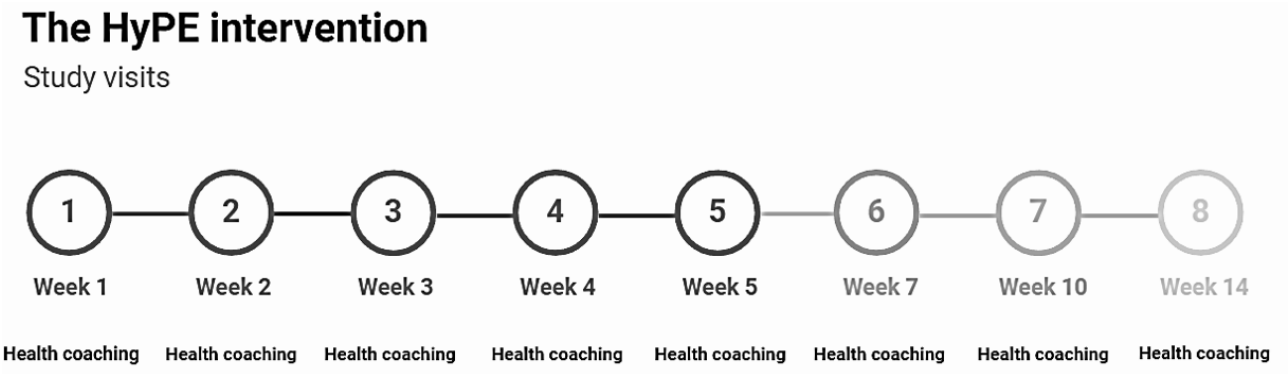

Measures

REDCap was used to build and manage survey assessments [21]. At pre- and posttesting, the following information was collected.

Demographics

Questions assessed the individual's age, race, ethnicity, marital status, education, annual household income, insurance coverage, parity, and employment status. Those currently employed were asked questions assessing job type, hours worked per week, if parental leave was provided, and weeks of parental leave provided.

PA Identity

PA identity was assessed with the validated exercise self-identity scale [22]. The 9-item survey assessed various components for how strongly individuals felt exercise was an important component of self. Individuals rated each item on a 7-point Likert scale from "strongly disagree" (0) to "strongly agree" (6) which were summed for a total score (higher scores indicated exercise was a greater component of self).

PA Planning and Monitoring

PA planning and self-monitoring behaviors were assessed with the validated 8-item self-monitoring and 6-item action planning questionnaires [23]. Questions were assessed on a 4-point Likert scale from "completely disagree (0)" to "totally agree (3)." Scores were calculated by summing responses; higher scores indicated greater self-monitoring and planning of PA.

Program Evaluation

At the conclusion of the intervention, participants responded to a series of closed-ended and open-ended questions assessing their acceptance of the overall program and different components of the intervention including health coaching sessions, perception of PA changes, quality of life, usefulness of tailored intervention materials, the Fitbit, activPAL4 micro use, and home BP monitoring with a 5-point Likert scale from "strongly disagree" (0) to "strongly agree" (4). These questions were adapted from a previously tested intervention [18].

PA Behavior

Device-based measurement of PA was conducted with both the activPAL4 micro and the Fitbit Inspire 2. ActivPAL measurement gave research grade assessments of PA at pre- and posttesting while the Fitbit allowed for steps measurement to occur at the end of the intensive intervention (week 5). The activPAL4 micro assessed changes in PA between pre- and posttesting and data from the Fitbit Inspire 2 were used to assess changes in steps during the intervention. The week prior to the first and last health coaching sessions, participants were instructed to wear an activPAL4 micro for 9 consecutive days (2 partial days and 7 full days) that they received via mail along with instructions for the device. They were also provided an activity log to self-report sleep, naps, and nonwear times. The activPAL4 micro was wrapped with retention tape and then

XSL•FO
RenderX

adhered to the anterior thigh with transparent dressing (3M Tegaderm Film) for 24 hours a day, to be removed only when swimming to prevent monitor loss.

For assessment of the preliminary efficacy of the intervention, our primary outcome was a change in steps per day measured with the activPAL4 micro. Data collected from the activPAL4 micro were exported as events using PAL technologies software using the VANE algorithm (PAL Technologies). Participants' self-reported sleep and nonwear times in their activity logs were removed using R (R Core Team) [24]. The processing procedures for this approach are available via the pregnancy247 package located on GitHub [25]. This code generates a figure depicting sitting, standing, or stepping activities, as well as sleep or nonwear periods. Visual inspection of these graphs was conducted to verify the accuracy of reported sleep onset, offset, and naps. Movement behaviors during waking hours on valid days were summarized using the activPAL processing package created by Lynden et al [26]. Total wear time and valid days were defined by at least 5 days of total wear with greater than 10 hours of wear time each day [27].

For the Fitbit Inspire 2, participants were instructed to wear the device every day, only removing it at night to charge. At the end of study participation, all step data collected from the Fitbit were retrieved by the study staff for data analysis. A valid day was defined as achieving >1000 steps [28].

Blood Pressure

The protocol for this study's BP assessment followed the American Heart Association's monitoring home BP recommendations [26,29]. Participants were provided the validated Omron Series 5 BP monitor at the start of the study [30,31], and received verbal and visual instructions on how to take an accurate reading during visit 1, visit 4, and visit 7. Participants were asked to measure their BP at pretesting, at the end of the intensive intervention (between visits 4 and 5), and a week prior to the final study visit which was at the same time as PA assessments that occurred with both the activPAL4 micro and Fitbit. During each of these assessments, participants were asked to measure their BP twice in the morning and twice at night for 7 consecutive days (ie, 28 total readings). These readings were synced with the Omron Connect app on the participant's cellular device and shared with the research staff via email. Assessment of appropriate BP technique occurred remotely via Zoom (Zoom Communications) at each study visit with corrections of technique provided when needed.

Statistical Analysis

All statistical analyses were determined a priori for main objectives and SAS (version 9.4; SAS Institute) was used for analysis. Demographic characteristics were assessed using frequencies and averages. Feasibility was measured by the number of completed sessions out of the total sessions possible. Acceptability was assessed by participants' self-reported satisfaction and likelihood of recommending the program to others. High feasibility was defined as 75% of all sessions attended. High acceptability was defined by at least 75% of participants responding that they were either satisfied or very satisfied with the program.

To assess changes in PA and BP between timepoints, data from the activPAL4 micro, Fitbit, and Omron series 5 were analyzed using paired *t* tests, Wilcoxon signed-rank, and Fisher exact tests. To estimate the effect size, Hedges *g* statistics were used. Thresholds for small, medium, and large effect sizes were set at 0.2, 0.5, and 0.8, respectively [32]. Linear regression models were used to assess the association of change in PA (average steps per day) and change in BP across the intervention [33]. SBP and DBP were treated as continuous variables and included in separate models. Covariates included in the final model were the participant's age, annual household income, activPAL wear time, and pretest steps.

Although a proof-of-concept study, a sample size calculation was conducted with a 0.50 effect size, based on preliminary evidence concerning peripartum PA interventions, using 80% power and at an α level of .05. From this, it was estimated that 15 individuals would be needed to detect significant changes in PA. Allowing for 25% drop out, the goal was to recruit 20 participants.

Results

Recruitment

Of the 94 individuals who completed the eligibility survey, 38 were found eligible to participate in the intervention. Those who were ineligible were excluded most often due to lack of HDP diagnosis (n=13), being greater than 6 months postpartum (n=24), or self-reporting current CVD or taking antihypertensive medications (n=17). Of the 38 eligible, 23 consented to the intervention. Only 19 of the 23 consented to enroll in the study with 3 individuals no longer interested and 1 excluded due to activity (achieving >9000 steps/day). Two participants dropped out (n=1 moved out of the area and n=1 lack of time; attrition=11%) during the intervention, leaving 17 participants who successfully completed the intervention.

Participants

Participants were on average 30.3 years of age (Table 1), and more often self-identified as White (14/17, 82%), were in a committed relationship (16/17, 94%), had earned at least a bachelor's degree (9/17, 54%), and had an annual household income over US $100,000 (9/17, 54%). At the start of participation in the study, no individuals were on parental leave and were on an average of 4.1 (SD 0.81) months postpartum.

XSL•FO
RenderX

Table . Demographic characteristics of Hypertensive Disorders of Pregnancy Postpartum Exercise intervention participants[a].

Participant characteristics (N=17)	Values
Age (years), mean (SD)	30.3 (3.6)
Race, n (%)	
White	14 (82)
Hispanic or Latina	1 (6)
Black	2 (12)
Marital status, n (%)	
Married or in a committed relationship	16 (94)
Single	1 (6)
Education, n (%)	
High school graduate or equivalent	4 (24)
Associate's degree	3 (18)
Bachelor's degree	4 (24)
Master's or doctoral degree	5 (30)
Annual household income (US $), n (%)	
<$25,000	2 (12)
$25,000 - $99,999	6 (35)
$100,000 - $149,999	6 (35)
>$150,000	3 (18)
Earned wages outside of home, n (%)	14 (82)
Parity, n (%)	
1 child	6 (35)
2 children	7 (41)
3 or more children	4 (24)
Hypertensive disorder of pregnancy, n (%)	
Gestational hypertension	11 (65)
Preeclampsia	6 (35)
Months postpartum at start of intervention, mean (SD)	4.1 (0.81)

[a]Prior to the start of the intervention, participants completed an online survey to self- report their demographic characteristics. Continuous metrics were reported as mean (SD) and categorical variables as frequencies, n (%).

Feasibility

Of all the participants who completed pretesting (n=19), 140 of the 152 (92%) sessions were completed. Of participants who completed the intervention (N=17), all 136 (100%) possible health coaching sessions were completed. Completers wore the Fitbit on 96% of all possible days. Almost all BP measures were obtained from completers with 90% of the pretest, 85% of week 5, and 85% of the posttest measures taken.

Acceptability

Participants evaluated the intervention and components of the program including health coaching sessions, observed changes in their PA behaviors, changes in health and well-being, use of information sheets, Fitbit, activPAL, and BP monitors (Table 2). All completers (N=17) indicated they were satisfied with the program (100%) and would recommend this program to others (100%).

XSL•FO
RenderX

Table . Program evaluation of the Hypertensive Disorders of Pregnancy Postpartum Exercise intervention (N=17)[a].

Item	Agree, n (%)	Neutral, n (%)	Disagree, n (%)
Health coaching			
The health coaching sessions helped me increase my physical activity	17 (100)	0 (0)	0 (0)
The health coaching sessions helped me decrease the amount of time I spend sitting	16 (94)	1 (6)	0 (0)
I felt supported by my health coach	17 (100)	0 (0)	0 (0)
My health coach was able to answer my questions	17 (100)	0 (0)	0 (0)
My health coach helped me understand how to perform physical activity safely	17 (100)	0 (0)	0 (0)
My health coach helped me create solutions to problems I faced	16 (94)	1 (6)	0 (0)
I felt like it was easy to talk to my health coach	17 (100)	0 (0)	0 (0)
I had enough time for my health coaching sessions	16 (94)	1 (6)	0 (0)
I wish I had more sessions with my health coach	8 (47)	8 (47)	1 (6)
I wish I had fewer sessions with my health coach	11 (65)	6 (35)	0 (0)
The length of time between sessions (weekly, every other week, and monthly) was appropriate	17 (100)	0 (0)	0 (0)
I enjoyed the ability to have health coaching sessions virtually from the comfort of my own home	17 (100)	0 (0)	0 (0)
Movement behaviors			
I feel like I have become more physically active	17 (100)	0 (0)	0 (0)
I know how to be physically active	17 (100)	0 (0)	0 (0)
I sit less during the day	17 (100)	0 (0)	0 (0)
I think that physical activity is an important part of living a healthy lifestyle	17 (100)	0 (0)	0 (0)
I am confident in my ability to stay physically active	17 (100)	0 (0)	0 (0)
Health and well-being			
I feel healthier	15 (88)	2 (12)	0 (0)
I have more energy during the day	13 (77)	2 (12)	2 (12)
I feel supported to live an active lifestyle	15 (88)	2 (12)	0(0)
I have noticed an improvement in my sleep	9 (53)	7 (41)	1 (6)
I have noticed an improvement in my mood	15 (88)	2 (12)	0 (0)

XSL•FO
RenderX

Item	Agree, n (%)	Neutral, n (%)	Disagree, n (%)
Information sheets			
I referenced the information sheets during the intervention	14 (82)	1 (6)	2 (12)
The information sheets helped me increase my physical activity	10 (59)	4 (24)	3 (18)
The information sheets were difficult to understand	3 (18)	7 (41)	7 (41)
Fitbit wear			
It was easy to set up my Fitbit (downloading app and linking device)	17 (100)	0 (0)	0 (0)
My Fitbit was easy to use	16 (94)	1 (6)	0 (0)
I checked the step count on my Fitbit daily	15 (88)	1 (6)	1 (5.9)
My Fitbit was uncomfortable to wear	4 (24)	2 (12)	11 (65)
The Fitbit app was easy to use	17 (100)	0(0)	0(0)
I checked my Fitbit app daily	9 (53)	5 (29)	3 (18)
The instructions given to me on how to use my Fitbit were helpful	15 (88)	1 (6)	0 (0)
It was difficult to remember to wear my Fitbit daily	8 (47)	9 (53)	0 (0)
It was difficult to remember to sync my Fitbit daily	2 (12)	4 (24)	11 (65)
I remembered to charge my Fitbit regularly	13 (77)	3 (18)	1 (6)
My Fitbit helped me achieve my step goals	15 (88)	2 (12)	0 (0)
ActivPAL wear			
The thigh activity monitor was uncomfortable to wear	3 (18)	5 (29)	9 (53)
It was easy to remember to complete my sleep log daily	8 (47)	5 (29)	4 (24)
I always completed my sleep log at the same time of day	6 (35)	4 (24)	7 (41)
Blood pressure measurements			
It was easy to remember to take my blood pressure	8 (47)	4 (24)	5 (29)
It was easy connecting my monitor to my phone	14 (82)	2 (13)	1 (6)
I know how to take a blood pressure correctly	17 (100)	0 (0)	0 (0)
It was useful knowing my blood pressure	16 (94)	0 (0)	1 (6)
It was hard to find time to take my blood pressure	2 (13)	7 (44)	7 (44)
Overall program rating			
I was satisfied with the program	17 (100)	0 (0)	0 (0)

XSL•FO
RenderX

Item	Agree, n (%)	Neutral, n (%)	Disagree, n (%)
I would recommend this program to others	17 (100)	0 (0)	0 (0)

[a]Participants ranked intervention features on a 5-point Likert scale from (1) “strongly agree” to (5) “strongly disagree.” Items were assessed at the end of the health coaching intervention via an online survey. To improve readability, (1) “strongly agree” and (2) “agree” were combined, as well as (4) “disagree” and (5) “strongly disagree”.

Concerning specific components of the intervention, all agreed that health coaching sessions helped increase their PA (n=17, 100%) and most agreed these sessions helped decrease their sitting time (n=16, 94%). Most participants said they referenced the information sheets during the program (n=14, 82%) and that this information helped them increase their PA (n=10, 59%). The Fitbit overall was rated as easy to use (n=16, 94%) and set up (n=17, 100%). Most said they checked their Fitbit step count daily (n=15, 88%). Most said Fitbit instructions were easy to understand (n=15, 88%), Fitbit was easy to remember to charge (n=13, 77%), and it helped them achieve their step goals (n=15, 88%).

All participants indicated they knew how to take a BP correctly after starting the program (n=17, 100%), with a majority indicating it was useful knowing their BP (n=16, 94%) and that it was easy to connect their monitor to their phone (n=14, 82%). Less than half of the participants agreed it was easy to remember to take their BP (n=8, 47%) and find time to do so (n=7, 44%).

Preliminary Efficacy

Participants on average wore the activPAL for 7.39 (SD 0.51) days at pretesting and 6.55 (SD 1.46) days at posttesting. As seen in Table 3, activPAL measures did not significantly change from pre- to posttesting, including average steps achieved per week (*P*=.75), standing time (*P*=.78), stepping time (*P*=.49), and sitting time (*P*=.38). Fitbit data further contextualized the changes observed in steps during the intervention period. A statistically significant increase in Fitbit steps was only observed from pretesting to the end of the intensive intervention (*P*=.01).

Table . Physical activity and blood pressure changes of Hypertensive Disorders of Pregnancy Postpartum Exercise intervention participants (N=17)[a].

Variable	Pretest	Week 5	Posttest	Hedges *g*
ActivPAL measures				
Steps, mean (SD)	6505.17 (2133.40)	—[b]	6643.60 (1997.68)	0.75
Standing time, mean (SD)	249.57 (65.40)	—	269.52 (113.51)	0.78
Stepping time, mean (SD)	88.57 (27.07)	—	91.93 (21.53)	0.49
Sitting time, mean (SD)	868.14 (189.03)	—	899.85 (255.90)	0.38
Fitbit measures				
Steps, mean (SD)	6584.47 (2705.65)	7898.27 (2359.32)[c]	7031.67 (2288.79)	0.48
BP[d] measures				
Systolic BP, mean (SD)	122.12 (8.04)	120.67 (8.33)	120.83 (8.10)	–0.26
Diastolic BP, mean (SD)	82.95 (6.02)	81.77 (6.31)	81.15 (5.47)	–0.44
Hypertensive, n (%)	11 (64.7)	10 (58.8)	7 (41.2)[c]	–0.51

[a]All steps reported in units of average steps per week, time variables in average minutes per day, and blood pressures in average mm Hg. The activPAL4 micro and Fitbit Inspire 2 assessed changes in participants’ physical activity. The Omron Series 5 blood pressure monitor assessed changes in participants’ blood pressure. Reported Hedges *g* are effect size estimates between pretest and posttest. Given all variables were nonparametric in nature, Wilcoxon signed-rank test was used to detect significant differences between 2 timepoints. Statistical significance defined as *P*<.05.

[b]Not applicable.

[c]Significant difference from baseline.

[d]BP: blood pressure.

Changes in PA determinants, self-efficacy, planning, monitoring, and identity from pre- to posttesting are shown in Table 4. Overall, PA planning behaviors increased significantly (mean 5.35, SD 4.97 to mean 15.06, SD 3.09; *P*<.001), with increases in participant’s ability to plan when to be active, where to be active, who to be active with, how to overcome disruptions of plans and plan around difficult situations. Behaviors associated with monitoring of PA also significantly increased (mean 7.29, SD 3.44 to mean 13.00, SD 2.45; *P*<.001), with participants indicating increases in monitoring whether they are active enough, that they walk enough steps per day, and the awareness of PA goals. Components of PA identity significantly increased, including being more likely to identify as an active person, describing self to others as being active, needing PA to feel good, and indicating that giving up exercise would cause a sense of loss. No other statistically significant changes were observed.

XSL•FO
RenderX

Table . Change in physical activity determinants, self-efficacy, planning, monitoring, and identity of participants in the Hypertensive Disorders of Pregnancy Postpartum Exercise intervention (N=17)[a].

Items	Pretest	Posttest	*P* value
Planning physical activity, mean (SD)			
I can plan when to be active	0.77 (0.75)	2.18 (0.39)	<.01
I can plan where to be active	0.82 (0.95)	2.12 (0.49)	.01
I can plan how to be active	0.65 (0.79)	2.29 (0.47)	.06
I can plan how often to be active	0.65 (0.70)	2.17 (0.64)	.06
I can plan who to be active with	0.65 (0.86)	1.65 (0.79)	.01
I know what to do if my plans are disrupted	0.65 (0.61)	1.41 (0.80)	.02
I can cope with setbacks	0.59 (0.51)	1.59 (0.62)	.10
I can plan around difficult situations	0.59 (0.51)	1.65 (0.79)	.04
Total planning	5.35 (4.97)	15.06 (3.09)	<.001
Monitoring physical activity, mean (SD)			
I monitor whether I am active enough	1.00 (0.71)	2.12 (0.60)	.01
I monitor that I walk enough steps per day	0.76 (0.56)	2.00 (0.87)	.02
I have physical activity intentions on my mind	1.82 (0.64)	2.41 (0.51)	.09
I am aware of my physical activity goals	1.12 (0.86)	2.12 (0.70)	.02
I try to be physically active regularly	1.35 (0.93)	2.18 (0.39)	.17
I try to achieve my physical activity goals	1.24 (0.90)	2.18 (0.39)	.21
Total score	7.29 (3.44)	13.00 (2.45)	<.001
Physical activity identity, mean (SD)			
I am an active person	1.88 (0.86)	2.18 (0.95)	.02
I describe myself to others as active	0.88 (0.70)	1.53 (1.23)	.04
I have goals related to physical activity	2.00 (1.28)	2.77 (0.66)	.94
Physical activity is central to my self-concept	1.53 (1.07)	1.82 (1.07)	.19
I need physical activity to feel good about myself	2.47 (1.01)	2.35 (1.06)	.04
Others see me as an active person	1.52 (1.01)	1.76 (1.15)	.47
Being active is more than just exercise	2.71 (1.21)	2.71 (0.92)	.32
I would feel loss if I had to give up exercising	1.94 (0.97)	2.29 (1.21)	.001
I think about physical activity often	2.29 (1.10)	2.29 (0.99)	.13
Total score	17.24 (6.88)	19.71 (6.55)	.09

[a]Participants completed online surveys prior to the start and at the conclusion of the intervention. Validated surveys assessed participants' ability to plan for physical activity, monitor physical activity levels, and physical activity related identities. Differences assess in continuous measures with paired *t* tests and categorical with Fisher exact tests. All Likert scales ranged from "strongly disagree" (0) to "strongly agree" (5). Statistical significance

XSL•FO
RenderX

defined as $P<.05$.

At pretesting, the average SBP was 122.12 (SD 8.04) mm Hg and DBP was 82.95 (SD 6.02) mm Hg (Table 3). Of the pretest readings, 11 (65%) participants had average readings that would be classified as HTN based on American Heart Association guidelines. Significant decreases in the number of participants with hypertensive readings were observed between pre- and posttesting (11 vs 7 participants; $P=.01$; Hedges $g=-0.51$). No statistically significant changes in SBP (mean –1.44, SD 4.68 mm Hg; $P=.22$) and DBP (mean –1.19, SD 4.76 mm Hg; $P=.32$) were observed from pretesting to the end of the intensive intervention (week 5) or from pretesting to posttesting (SBP: mean –1.28, SD 3.59; Hedges $g=-0.26$; $P=.16$ and DBP: mean –1.80, SD 5.03; Hedges $g=-0.44$; $P=.12$). When assessing the association between an increase in steps by 1000 and BP, no findings were statistically significant (SBP: mean –0.30, SD 1.11 mm Hg; $P=.79$ and DBP: mean –1.52, SD 1.46 mm Hg; $P=.32$).

Discussion

Principal Findings

The HyPE intervention included 17 participants whose prior pregnancies were complicated by HDP. Overall, the intervention was found to be feasible and acceptable as we hypothesized. Preliminary efficacy assessments did not support our hypothesis as participants did not significantly increase the number of activPAL measured steps achieved from pre- to posttesting. Indicating that we failed to achieve a clinically significant impact on the behavioral target, a critical milestone in phase IIa of the ORBIT model [16]. Factors promoting PA (eg, PA monitoring, planning, and identity) were observed to significantly improve. HTN prevalence was the only BP metric observed to significantly decrease from pre- to posttesting.

Contextualizing Findings

CVD has previously been described as a “man’s disease,” but the acknowledged importance of CVD in women is increasing. Reviews of CVD-related clinical trials have highlighted an underrepresentation of women in research [34], subsequently attributing to sex-based disparities in the clinical guidelines, recommendations, treatments, and health outcomes related to CVD. Potential downstream effects of these disparities were reflected in our findings, with 94% (n=16) of individuals reporting at no time did a health care provider discuss the increased risk of CVD associated with HDP diagnosis. With HDP an important risk factor for CVD, it is critical as a field to find effective strategies to promote cardiovascular health in this high-risk population.

To the authors’ knowledge, only 1 randomized clinical trial has assessed the efficacy of a PA intervention after HDP. Participants of that study increased their steps more than the control (mean difference: 647 steps per day, 95% CI 169 - 1124; $P=.009$) [35]. A total of 65% of participants completed the study with 94% indicating that they would recommend the program to others [35]. The HyPE intervention had similar adherence and acceptability results, with 90% of participants completing the intervention and 100% satisfied with the program. Intervening on PA this early in postpartum (3 - 6 months postpartum) is novel and our results indicate that interventions during this timeframe may be feasible and acceptable. Although we did not observe significant increases in device-measured PA, participants did significantly increase their self-reported abilities to plan and monitor PA. This discrepancy may be a result of seasonality, given that the winter season is associated with a decrease of up to 2500 steps compared to summer months [36]. Those completing their posttesting in the winter (ie, December or January; n=9, 53%), consequently may have seen a decrease in their steps due to the colder temperatures. While participants in the HyPE intervention did not increase their steps, the intervention may have prevented the typical decline observed when transitioning from summer to winter months. Future interventions should consider the impacts seasonality may have on their population and develop strategies to combat potential effects colder weather may bring. Finally, we included the target population in the development of the intervention in previous formative work assessing the unique behavioral determinants, needs, and desires of postpartum individuals with HDP [17]. Formative work has been called for in postpartum PA intervention development with inclusion of the target population in this process essential to creating equitable and culturally appropriate solutions [37,38].

Limitations

There are several limitations of this study to consider when interpreting results. There was no control group due to this being a proof-of-concept study, limiting the ability to assess whether the changes observed were a direct effect of the intervention itself. This decision was intentional, however, as very few PA interventions have targeted the early postpartum period, and demonstrating feasibility and acceptability was our primary objective. This also follows the ORBIT model, representing phase IIa in the behavioral treatment developmental pathway [16]. Given that the milestone for successful completion of phase IIa was not reached with no clinically significant changes in behavior detected, further refinement and optimization of the intervention is warranted. These refinements should include lowering the <9000 step inclusion criteria threshold, as well as the duration or frequency of study visits. In addition, there was no activPAL4 micro assessment of PA at the end of the intensive intervention (week 5), which limited our ability to determine the immediate effects of the intervention on PA behavior. It is also possible that our PA eligibility criteria (must achieve <9000 steps/day) may have biased our results by not recruiting a truly inactive population. It is estimated that achieving Americans 7000 - 8000 steps is equivalent to meeting the PA guidelines for Americans [39]. Further, the significant decrease in hypertensive individuals should be interpreted with caution, as this may reflect that most individuals who were classified as hypertensive were only slightly above the American Heart Association defined BP threshold for HTN. Finally, the population recruited was not racially, ethnically, or socioeconomically diverse, with a majority of individuals being White and of high socioeconomic status. Future studies will require a larger, more diverse sample and should consider recruiting individuals achieving <7000 steps per day. Future

XSL•FO
RenderX

studies are needed to contextualize the needs and desires of diverse populations. Given the importance intersectional identities may have in a birthing person's health and experiences with the medical system, the inclusion of diverse cohorts is essential to the equitable promotion of female cardiovascular health after HDP.

Conclusions

HDP incidence is increasing and is associated with an increased risk of CVD later in life. Our study suggests that intervening in PA behaviors early after HDP is feasible and acceptable. Future studies should further refine intervention strategies for this population and test its efficacy in a larger, more diverse, and less active sample.

Acknowledgments

The authors would like to thank the support of the American Heart Association for funding this project (23PRE1018306). The authors would also like to thank all the participants who made this study possible.

Data Availability

The datasets generated and analyzed during this study are not publicly available due to participants not providing prior consent to their data being shared.

Conflicts of Interest

None declared.

References

1. Khosla K, Heimberger S, Nieman KM, et al. Long-term cardiovascular disease risk in women after hypertensive disorders of pregnancy: recent advances in hypertension. Hypertension 2021 Sep;78(4):927-935. [doi: 10.1161/HYPERTENSIONAHA.121.16506] [Medline: 34397272]
2. Melchiorre K, Thilaganathan B, Giorgione V, Ridder A, Memmo A, Khalil A. Hypertensive disorders of pregnancy and future cardiovascular health. Front Cardiovasc Med 2020 Apr;7:59. [doi: 10.3389/FCVM.2020.00059/BIBTEX]
3. Benschop L, Duvekot JJ, Roeters van Lennep JE. Future risk of cardiovascular disease risk factors and events in women after a hypertensive disorder of pregnancy. Heart 2019 Aug;105(16):1273-1278. [doi: 10.1136/heartjnl-2018-313453] [Medline: 31175138]
4. Bushnell C, McCullough LD, Awad IA, et al. Guidelines for the prevention of stroke in women. Stroke 2014 May;45(5):1545-1588. [doi: 10.1161/01.str.0000442009.06663.48]
5. Tooher J, Thornton C, Makris A, Ogle R, Korda A, Hennessy A. All hypertensive disorders of pregnancy increase the risk of future cardiovascular disease. Hypertension 2017 Oct;70(4):798-803. [doi: 10.1161/HYPERTENSIONAHA.117.09246] [Medline: 28893895]
6. Lewey J, Levine LD, Yang L, Triebwasser JE, Groeneveld PW. Patterns of postpartum ambulatory care follow-up care among women with hypertensive disorders of pregnancy. J Am Heart Assoc 2020 Sep;9(17):e016357. [doi: 10.1161/JAHA.120.016357] [Medline: 32851901]
7. Pouliou T, Ki M, Law C, Li L, Power C. Physical activity and sedentary behaviour at different life stages and adult blood pressure in the 1958 British cohort. J Hypertens 2012 Feb;30(2):275-283. [doi: 10.1097/HJH.0b013e32834f1915] [Medline: 22179092]
8. Whelton SP, Chin A, Xin X, He J. Effect of aerobic exercise on blood pressure: a meta-analysis of randomized, controlled trials. Ann Intern Med 2002 Apr 2;136(7):493-503. [doi: 10.7326/0003-4819-136-7-200204020-00006] [Medline: 11926784]
9. Muntner P, Jaeger BC, Hardy ST, et al. Age-specific prevalence and factors associated with normal blood pressure among US adults. Am J Hypertens 2022 Apr 2;35(4):319-327. [doi: 10.1093/ajh/hpab154] [Medline: 34599824]
10. Sobierajski FM, Purdy GM, Usselman CW, et al. Maternal physical activity is associated with improved blood pressure regulation during late pregnancy. Can J Cardiol 2018 Apr;34(4):485-491. [doi: 10.1016/j.cjca.2018.01.021] [Medline: 29571429]
11. Haakstad LAH, Edvardsen E, Bø K. Effect of regular exercise on blood pressure in normotensive pregnant women. A randomized controlled trial. Hypertens Pregnancy 2016 Apr;35(2):170-180. [doi: 10.3109/10641955.2015.1122036]
12. Borodulin K, Evenson KR, Herring AH. Physical activity patterns during pregnancy through postpartum. BMC Womens Health 2009 Nov 19;9(1):32. [doi: 10.1186/1472-6874-9-32] [Medline: 19925650]
13. Parikh NI, Gonzalez JM, Anderson CAM, et al. Adverse pregnancy outcomes and cardiovascular disease risk: unique opportunities for cardiovascular disease prevention in women: a scientific statement from the American Heart Association. Circulation 2021 May 4;143(18):e902-e916. [doi: 10.1161/CIR.0000000000000961] [Medline: 33779213]
14. Mehta LS, Warnes CA, Bradley E, et al. Cardiovascular considerations in caring for pregnant patients: a scientific statement from the American Heart Association. Circulation 2020 Jun 9;141(23):e884-e903. [doi: 10.1161/CIR.0000000000000772] [Medline: 32362133]

XSL•FO
RenderX

15. Mehta LS, Sharma G, Creanga AA, et al. Call to action: maternal health and saving mothers: a policy statement from the American Heart Association. Circulation 2021 Oct 12;144(15):e251-e269. [doi: 10.1161/CIR.0000000000001000] [Medline: 34493059]
16. Czajkowski SM, Powell LH, Adler N, et al. From ideas to efficacy: the ORBIT model for developing behavioral treatments for chronic diseases. Health Psychol 2015 Oct;34(10):971-982. [doi: 10.1037/hea0000161] [Medline: 25642841]
17. EPI: abstract detail. EPI Lifestyle. URL: https://epi.apprisor.org/epsAbstractEPI.cfm?id=4 [accessed 2024-03-22]
18. Whitaker KM, Jones MA, Dziewior J, et al. Feasibility, acceptability, and preliminary efficacy of a single-arm, remotely-delivered health coaching intervention to increase physical activity and reduce sedentary behavior during pregnancy. Research Square. Preprint posted online on Jul 19, 2022. [doi: 10.21203/rs.3.rs-1755320/v1]
19. Hettema J, Steele J, Miller WR. Motivational interviewing. Annu Rev Clin Psychol 2005;1(1):91-111. [doi: 10.1146/annurev.clinpsy.1.102803.143833] [Medline: 17716083]
20. Bundy C, Cpsychol P. Changing behaviour: using motivational interviewing techniques. J R Soc Med 2004;97 Suppl 44(Suppl 44):43-47. [Medline: 15239293]
21. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381. [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
22. Biber DD. Exercise identity, self-regulatory efficacy, and self-compassion prepared for psychological studies. Psychol Stud 2020 Sep;65(3):261-269. [doi: 10.1007/s12646-020-00556-w]
23. Sniehotta FF, Scholz U, Schwarzer R. Bridging the intention–behaviour gap: planning, self-efficacy, and action control in the adoption and maintenance of physical exercise. Psychol Health 2005 Apr;20(2):143-160. [doi: 10.1080/08870440512331317670]
24. The R Project for Statistical Computing. URL: https://www.r-project.org/ [accessed 2024-12-20]
25. Deboonstra/pregnancy247. GitHub. URL: https://github.com/deboonstra/pregnancy247 [accessed 2024-11-28]
26. Lyden K, Keadle SK, Staudenmayer J, Freedson PS. The activPALTM accurately classifies activity intensity categories in healthy adults. Med Sci Sports Exerc 2017 May;49(5):1022-1028. [doi: 10.1249/MSS.0000000000001177] [Medline: 28410327]
27. Kozey-Keadle S, Libertine A, Lyden K, Staudenmayer J, Freedson PS. Validation of wearable monitors for assessing sedentary behavior. Med Sci Sports Exerc 2011 Aug;43(8):1561-1567. [doi: 10.1249/MSS.0b013e31820ce174] [Medline: 21233777]
28. Hardcastle SJ, Jiménez-Castuera R, Maxwell-Smith C, Bulsara MK, Hince D. Fitbit wear-time and patterns of activity in cancer survivors throughout a physical activity intervention and follow-up: exploratory analysis from a randomised controlled trial. PLOS ONE 2020;15(10):e0240967. [doi: 10.1371/journal.pone.0240967] [Medline: 33075100]
29. Shimbo D, Artinian NT, Basile JN, et al. Self-measured blood pressure monitoring at home: a joint policy statement from the American Heart Association and American Medical Association. Circulation 2020 Jul 28;142(4):e42-e63. [doi: 10.1161/CIR.0000000000000803] [Medline: 32567342]
30. Grover-Páez F, Cardona-Muñoz EG, Cardona-Müller D, et al. Validation of the Omron HEM-7320-LA, upper arm blood pressure monitor with Intelli Wrap Technology Cuff HEM-FL1 for self-measurement and clinic use according to the European Society of Hypertension International Protocol revision 2010 in the Mexican population. Blood Press Monit 2017 Dec;22(6):375-378. [doi: 10.1097/MBP.0000000000000290] [Medline: 28945665]
31. Suva M. Validation of BP monitoring devices. J Clin Hypertens 2020 Feb;3:15-19.
32. Lakens D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Front Psychol 2013 Nov 26;4:863. [doi: 10.3389/fpsyg.2013.00863] [Medline: 24324449]
33. Kaneko H, Yano Y, Itoh H, et al. Association of blood pressure classification using the 2017 American College of Cardiology/American Heart Association blood pressure guideline with risk of heart failure and atrial fibrillation. Circulation 2021 Jun 8;143(23):2244-2253. [doi: 10.1161/CIRCULATIONAHA.120.052624] [Medline: 33886370]
34. Melloni C, Berger JS, Wang TY, et al. Representation of women in randomized clinical trials of cardiovascular disease prevention. Circ Cardiovasc Qual Outcomes 2010 Mar;3(2):135-142. [doi: 10.1161/CIRCOUTCOMES.110.868307] [Medline: 20160159]
35. Lewey J, Murphy S, Zhang D, et al. Effectiveness of a text-based gamification intervention to improve physical activity among postpartum individuals with hypertensive disorders of pregnancy: a randomized clinical trial. JAMA Cardiol 2022 Jun 1;7(6):591-599. [doi: 10.1001/jamacardio.2022.0553] [Medline: 35442393]
36. Hamilton SL, Clemes SA, Griffiths PL. UK adults exhibit higher step counts in summer compared to winter months. Ann Hum Biol 2008;35(2):154-169. [doi: 10.1080/03014460801908058] [Medline: 18428010]
37. Peralta LR, Cotton WG, Dudley DA, Hardy LL, Yager Z, Prichard I. Group-based physical activity interventions for postpartum women with children aged 0-5 years old: a systematic review of randomized controlled trials. BMC Womens Health 2021 Dec 28;21(1):435. [doi: 10.1186/s12905-021-01581-1] [Medline: 34963456]
38. Vastine A, Gittelsohn J, Ethelbah B, Anliker J, Caballero B. Formative research and stakeholder participation in intervention development. Am J Health Behav 2005;29(1):57-69. [doi: 10.5993/ajhb.29.1.5] [Medline: 15604050]

XSL•FO
RenderX

39. Tudor-Locke C, Craig CL, Brown WJ, et al. How many steps/day are enough? For adults. Int J Behav Nutr Phys Act 2011 Jul 28;8(1):79. [doi: 10.1186/1479-5868-8-79] [Medline: 21798015]

Abbreviations

BP: blood pressure
CVD: cardiovascular disease
DBP: diastolic blood pressure
HDP: hypertensive disorders of pregnancy
HTN: hypertension
HyPE: Hypertensive Disorders of Pregnancy Postpartum Exercise
ORBIT: Obesity-Related Behavioral Intervention Trials
PA: physical activity
SBP: systolic blood pressure

Edited by A Mavragani; submitted 23.08.24; peer-reviewed by H Liu, J Kitt; revised version received 01.11.24; accepted 05.11.24; published 08.01.25.

Please cite as:
Borrowman JD, Carr LJ, Pierce GL, Story WT, Gibbs BB, Whitaker KM
Postpartum Remote Health Coaching Intervention for Individuals With a Hypertensive Disorder of Pregnancy: Proof-of-Concept Study
JMIR Form Res 2025;9:e65611
URL: https://formative.jmir.org/2025/1/e65611
doi:10.2196/65611

© Jaclyn D Borrowman, Lucas J Carr, Gary L Pierce, William T Story, Bethany Barone Gibbs, Kara M Whitaker. Originally published in JMIR Formative Research (https://formative.jmir.org), 8.1.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Testing the Recruitment Frequency, Implementation Fidelity, and Feasibility of Outcomes of the Heart Failure Activity Coach Study (HEALTHY): Pilot Randomized Controlled Trial

Andreas Blomqvist[1], PhD; Maria Bäck[1,2], PhD; Leonie Klompstra[1], PhD; Anna Strömberg[1,3], PhD; Tiny Jaarsma[1], PhD

[1]Department of Health, Medicine and Caring Sciences, Linköping University, Linköping, Sweden

[2]Department of Occupational Therapy and Physiotherapy, Sahlgrenska University Hospital, Gothenburg, Sweden

[3]Department of Cardiology, Linköping University, Linköping, Sweden

Corresponding Author:
Andreas Blomqvist, PhD
Department of Health, Medicine and Caring Sciences
Linköping University
Sandbäcksgatan 7
Linköping, SE-581 83
Sweden
Phone: 46 739617729
Email: andreas.blomqvist@liu.se

Abstract

Background: Heart failure (HF) is a common and deadly disease, precipitated by physical inactivity and sedentary behavior. Although the 1-year survival rate after the first diagnosis is high, physical inactivity and sedentary behavior are associated with increased mortality and negatively impact the health-related quality of life (HR-QoL).

Objective: We tested the recruitment frequency, implementation fidelity, and feasibility of outcomes of the Activity Coach app that was developed using an existing mobile health (mHealth) tool, Optilogg, to support older adults with HF to be more physically active and less sedentary.

Methods: In this pilot clinical randomized controlled trial (RCT), patients with HF who were already using Optilogg to enhance self-care behavior were recruited from 5 primary care health centers in Sweden. Participants were randomized to either have their mHealth tool updated with the Activity Coach app (intervention group) or a sham version (control group). The intervention duration was 12 weeks, and in weeks 1 and 12, the participants wore an accelerometer daily to objectively measure their physical activity. The HR-QoL was measured with the Kansas City Cardiomyopathy Questionnaire (KCCQ), and subjective goal attainment was assessed using goal attainment scaling. Baseline data were collected from the participants' electronic health records (EHRs).

Results: We found 67 eligible people using the mHealth tool, of which 30 (45%) initially agreed to participate, with 20 (30%) successfully enrolled and randomized to the control and intervention groups in a ratio of 1:1. The participants' daily adherence to registering physical activity in the Activity Coach app was 69% (range 24%-97%), and their weekly adherence was 88% (range 58%-100%). The mean goal attainment score was –1.0 (SD 1.1) for the control group versus 0.6 (SD 0.6) for the intervention group (P=.001). The mean change in the overall HR-QoL summary score was –9 (SD 10) for the control group versus 3 (SD 13) in the intervention group (P=.027). There was a significant difference in the physical limitation scores between the control (mean 45, SD 27) and intervention (mean 71, SD 20) groups (P=.04). The average length of sedentary bouts increased by 27 minutes to 458 (SD 84) in the control group minutes and decreased by 0.70 minutes to 391 (SD 117) in the intervention group (P=.22). There was a nonsignificant increase in the mean light physical activity (LPA): 146 (SD 46) versus 207 (SD 80) minutes in the control and intervention groups, respectively (P=.07).

Conclusions: The recruitment rate was lower than anticipated. An active recruitment process is advised if a future efficacy study is to be conducted. Adherence to the Activity Coach app was high, and it may be able to support older adults with HF in being physically active.

Trial Registration: ClinicalTrials.gov NCT05235763; https://clinicaltrials.gov/study/NCT05235763

(JMIR Form Res 2025;9:e62910) doi:10.2196/62910

XSL•FO
RenderX

KEYWORDS

heart failure; disease management; physical activity; sedentary; older adults; aging; mobile health; mHealth; feasibility; quality of life; digital health; smartphone

Introduction

Background

Heart failure (HF) is a clinical syndrome characterized by a reduced cardiac output leading to a deficit in meeting the metabolic needs of the cells [1]. It is a major cause of morbidity and mortality worldwide [2]. Physical inactivity and sedentary behavior adversely affect outcomes in patients with HF. Sedentary behavior is a state where no activities performed increase energy expenditure substantially above the resting level, typically sitting or lying down [3]. Physical inactivity can be defined as the nonachievement of targets according to the physical activity guidelines that are recommended for an individual [4]. Although the 1-year survival rate for patients after the first diagnosis of HF is high [5], physical inactivity is associated with additionally increased mortality them [6]. Similarly, sedentary behavior is also associated with increased mortality [7] and negatively impacts the health-related quality of life (HR-QoL) [8]. It is, therefore, desirable to support patients with HF to be more physically active and break their sedentary behavior. A tool aimed at doing this, called "Activity Coach," has been developed as a separate module on an existing mobile health (mHealth) tool [9].

The Activity Coach app is aimed at supporting everyday physical activity and reducing sedentary behavior and should not be viewed as an exercise tool. Furthermore, it is primarily intended for the physically inactive population with HF and is planned to be tested for efficacy in that population [9].

The Medical Research Council (MRC) suggests that before conducting an appropriately powered efficacy study, enough piloting and feasibility work has to be performed to be confident that the intervention can be delivered as intended (fidelity) and that safe assumptions can be made about effect sizes and variability and about rates of recruitment and retention [10]. A pilot randomized controlled trial (RCT) is the best way to assess the feasibility of a large, expensive full-scale study [11], and hence, the objectives of a pilot RCT need to be different from those of a future efficacy study, including comments on the uncertainties to be investigated [12]. In accordance with the MRC [10] and the CONSORT (Consolidated Standards of Reporting Trials) guidelines for pilot RCTs [13] (Multimedia Appendix 1), we conducted a pilot RCT to evaluate a study design, with the main focus being the rate of recruitment. In addition, we sought to assess the fidelity of the intervention, as well as the plausibility of the proposed outcome measures, in terms of data quality and effect sizes.

Objective

The objective of this pilot RCT was to test the study protocol of an intervention in terms of patient recruitment, intervention fidelity, and plausibility of outcomes. The research questions were as follows:

- Recruitment: What was the recruitment frequency and attrition rate in the target population?
- Implementation fidelity: To what extent was the intervention delivered as intended?
- Outcomes: How was the plausibility of outcomes in terms of data integrity and quality, and what limited efficacy was seen regarding the included outcome measures?

Methods

Trial Design and Participants

The Heart Failure Activity Coach Study (HEALTHY) was a single-blinded, parallel pilot clinical RCT. Since the Activity Coach app was developed on the platform of an already existing mHealth tool called Optilogg, only people equipped with a home-based self-care support device called Optilogg were eligible to participate. As all study participants regardless of study arm allocation would have Optilogg since before, this also eliminated the risk of the effects of Optilogg per se confounding the results. Participants also needed to be listed at a primary health care center.

The eligibility criteria for participation were age≥18 years, a confirmed HF diagnosis, using Optilogg (obtained from a participating HF clinic), being physically inactive based on a screening question, and having the ability to consent to participation. Eligible people were excluded if they did not consent to a home visit or to using an accelerometer, if they had a life expectancy shorter than 6 months, or if they were already participating in another study on physical activity.

Change to Trial Design After Commencement

After the initial slow recruitment of just 2 randomized participants after 3 months, an amendment was submitted to the ethical review board, the inclusion criteria were modified, and it was decided that all people willing to participate would be included regardless of their answer to the screening question on physical inactivity. This change was justified based on the fact that the Activity Coach app might have benefits for a more active population with HF, too. Furthermore, potential participants were now contacted by telephone instead of regular mail. In the *Ancillary Analyses* section, potential differences between participants as stratified by the screening question are reported.

Study Procedure

Eligible people were first contacted by regular mail, and those willing to participate would be asked to answer a single-item self-report question. This question has previously been evaluated as a screening tool for classifying respondents as being physically active or physically inactive, and the same methodology was used in this study to identify physically inactive people [14]. The question is item 9 in the validated the European Heart Failure Self-care Behaviour Scale, and the question is in the shape of a statement ("I exercise regularly"), which the respondent answers using an ordinal scale from 1 to

XSL•FO
RenderX

5, with 1 being “I completely agree” to 5 being “I don’t agree at all” [15]. In this study, responses from 3 to 5 were coded as physically inactive [14]. If a participant met the eligibility criteria, a home visit was scheduled. During the home visit, the participant signed an informed consent form, was randomized to the intervention or the control group, filled out baseline questionnaires, and received an accelerometer. Lastly, the Activity Coach app or the sham version was installed on their Optilogg mHealth tool.

The intervention period was 12 weeks, and during weeks 1 and 12 of the study, participants wore the accelerometer every day. At the end of the study the same questionnaires that were completed at the start of the study were filled out again. Clinical data, such as demographics, etiology, comorbidities, and pharmacological treatments, were collected from the participants’ electronic health records (EHRs) by the HF nurse at each respective primary health care center.

Intervention and Control

The mHealth Tool Optilogg

Optilogg consists of a tablet computer wirelessly connected to a weight scale, and the user is encouraged to interact with it daily. Without the addition of the Activity Coach app, Optilogg has 3 main features: subjective and objective symptom tracking, interactive education about living with HF, and an automated flexible loop-diuretics regime. It has been shown to improve self-care behavior in people with HF [16,17].

The Activity Coach App Intervention

The Activity Coach app was added as a novel feature to the existing Optilogg, with the purpose of supporting people with HF to be physically active, and was developed using principles described by the MRC [9,10]. Therefore, in addition to existing Optilogg features, the mHealth tool included Activity Coach features as well. Every day, the user can, in an intuitive graphical user interface (GUI), enter how much time they have been physically active in increments of 10 minutes regardless of what physical activity means for them. This allows for a one-size-fits-all approach to accommodate heterogenous populations. The accumulated physical activity is logged and presented as trends in a separate tab. Each week, there is an on-screen summary of the past week’s logged activity, and the user can, if desired, set a goal for the upcoming week, although they can at any time turn off the goal-setting feature. The progress toward reaching the goal is illustrated in the trends tab. The Activity Coach app was designed to first create motivation to be physically active and then provide a means of tracking physical activity. Motivation is achieved by providing an educational tip a day for 1 week, aimed at establishing 3 beliefs: physical activity leads to certain outcomes (positive outcome expectancy), the person can perform physical activity (self-efficacy), and the outcomes are desirable to the person (goal congruence). More details on the underlying intervention theory and theoretical model, as well as the development process, have been published elsewhere [9]. The participants in the intervention group were instructed to try to track their physical activity in the Activity Coach app daily.

Control Version

As previously mentioned, the original Optilogg contains an education module that, in part, deals with physical activity and exercise, and this education module has been shown to increase knowledge about HF [17]. This could have been a possible confounder when trying to analyze the additional effects of the Activity Coach app. To mitigate this, all participants were recruited from existing Optilogg users. The intervention group had the Activity Coach app installed on their existing Optilogg system, and the control group had a sham version of the Activity Coach app installed instead. This slightly changed the appearance of the GUI compared to the original Optilogg prior to the study, such that instead of starting each day on the regular home-screen, the modified system started in the section of the education module dealing with physical activity, but none of the actual Activity Coach features were enabled. This led to an apparent change, also relating to physical activity, which would be immediately noted by the user.

Research Question 1: Recruitment

The recruitment procedure metric to be evaluated was defined as the number of people who met the eligibility criteria who needed to be contacted and screened to reach a certain number of recruited and randomized participants. Monitoring attrition rates was also part of evaluating the study design. As described earlier, the study design was changed to also include participants who originally were intended to not be included, since they were defined as physically active. Hence, this outcome was reported as 2 numbers: one where the original screening was still used and the other where physically active people were also included.

Research Question 2: Implementation Fidelity

Implementation fidelity was evaluated by analyzing how much the participants randomized to the intervention group actually used the Activity Coach app (eg, adherence, which is a common means of assessing intervention fidelity in physical activity interventions) [18]. Adherence was defined as the ratio of actual use to intended use [19], and since, no clear recommendation exists on how often the Activity Coach app or similar behavior change interventions relating to physical activity should be used, we reported daily and weekly adherence. Daily adherence was calculated as the number of days that the participant performed physical activity divided by the study duration in days (ie, 84). Weekly adherence was calculated as the number of weeks that the participant performed physical activity in 1 day or more divided by the number of weeks in the study (ie, 12). Furthermore, the Activity Coach app offers a goal-setting feature on a weekly basis. The usage of this feature was also examined.

Research Question 3: Outcomes

The outcome measures studied were the subjective goal attainment score; the HR-QoL, measured with the Kansas City Cardiomyopathy Questionnaire (KCCQ); and objective physical activity, measured with an accelerometer.

Subjective Goal Attainment

For goal attainment, we relied on the methods proposed by Urach et al [20]. Upon enrollment, participants selected 2 goals

XSL•FO
RenderX

from a list of 10 physical activity–related goals they wanted to achieve during the 12-week study. The list was based on a literature review of the effects of increased physical activity on people with HF. The 10 goals are listed in Table 1. Upon completing their participation in the study, each participant rated the 2 selected goals during follow-up, comparing the current status with the status when the study started. The rating was on a scale from –2 to 2, where –2 indicated worsening and 2 indicated improvement. A mean value of the scores of the 2 goals was then calculated and constituted the reported value, referred to as the goal attainment score.

Table 1. Prespecified goals for the subjective goal attainment outcome.

Goal number	Goal
1	Better sleep
2	Boosted daytime energy
3	Able to walk with ease without frequent pauses
4	Less physical pain
5	Feel better
6	Reduced shortness of breath
7	Less daytime fatigue
8	More energy to do things around the house
9	Increased desire and energy for social activities
10	Improved appetite

Health-Related Quality of Life

The Kansas City Cardiomyopathy Questionnaire (KCCQ) short version is a 12-item questionnaire used to assess the HR-QoL in people with cardiomyopathies, which outputs 4 subscale summary scores (physical limitation, symptom frequency, quality of life [QoL], and social limitation), as well as a total summary score [21]. The scores range from 0 to 100, where 100 represents the best-possible value for any given score. The questionnaire has been validated in the current population [22].

Objective Physical Activity

Sedentary behavior and physical inactivity were assessed objectively using an accelerometer [23]. The accelerometer used was the validated ActiGraph GT9X [24], and participants were instructed to wear it around the hip for 7 consecutive days. The accelerometer data were reported as the percentage sedentary time, the average length of sedentary bouts, and steps per day. Data were sampled as raw data (30 Hz), converted to 60-second time intervals (or epochs) in counts per minute for the vector magnitude, and analyzed using ActiLife software (v.6.13.4).

Certain criteria were applied to the accelerometer data acquired for it to be included in the analyses. The accelerometer had to be used during at least 4 of the 7 days, with a minimum wear time of 540 minutes/day. Periods with consecutive 0s for 60 minutes or more (with 2-minute tolerance) were interpreted as time of nonuse and excluded from further analyses.

Sample Size

An appropriate sample size for a pilot should be based on the estimated effect size, where a large effect size warrants 15-30 participants, an intermediate effect size requires 25-35 participants [25], and a more general recommendation is a sample size of 12 per group [26]. In this study, the target sample size was set to 40, and enrollment was to stop either when 40 study participants had been randomized or when all eligible people had been contacted (whichever came first).

Randomization

Sheets of paper with an “I” for intervention or a “C” for control were printed and put in sealed envelopes. After an eligible participant had agreed to participate in the study, a home visit was scheduled. The participant signed the informed consent form and filled out the baseline questionnaire. The researcher selected an envelope from the batch and opened it and subsequently install the Activity Coach app or the sham version onto the participant’s Optilogg without informing them about the allocation.

Blinding

This was a single-blinded, parallel pilot RCT. Every participant already had the Optilogg system at home and at the start of the study and was randomized to either the control or the intervention group at the beginning, at which point the sham version or the actual Activity Coach app was installed on their Optilogg without them knowing which version they received. The participants were not informed of what arm they were randomized to until the study was over and all data collected. The person who entered and registered all the data was also blinded to the allocation, but the investigators knew which group each participant was allocated to.

Ethical Considerations

Ethics approval was obtained from Sweden’s Ethical Review Board (Dnr 2021-05366-01; approval date: November 17, 2021), with the amendment regarding the recruitment procedure also being approved (Dnr 2023-00799-02; approval date: February 19, 2023). All recruited participants signed and submitted their informed consent to participate in writing.

XSL•FO
RenderX

Statistical Analysis

Continuous variables, such as age and steps, were analyzed using the Student *t* test, and categorical variables were analyzed using the chi-square test. For subjectively graded goals, a mean value was computed from the 2 answers, and the Student *t* test was performed to test the differences between groups.

To test the differences in ordinal variables, such as self-reported physical activity and exercise, as well as nonnormally distributed data, such as the scaled outputs of the KCCQ, the Mann-Whitney U nonparametric test was performed. $P<.05$ was considered statistically significant.

All statistical calculations were computed in SPSS Statistics v.29 (IBM Corporation). All statistical tests performed on outcome data were performed without power calculations and were interpreted accordingly.

Results

Population

The participant selection flowchart, along with the study design, is illustrated in Figure 1.

Figure 1. Participant flow and study design of the pilot RCT. KCCQ - Kansas City Cardiomyopathy Questionnaire, EHR – electronic health records.

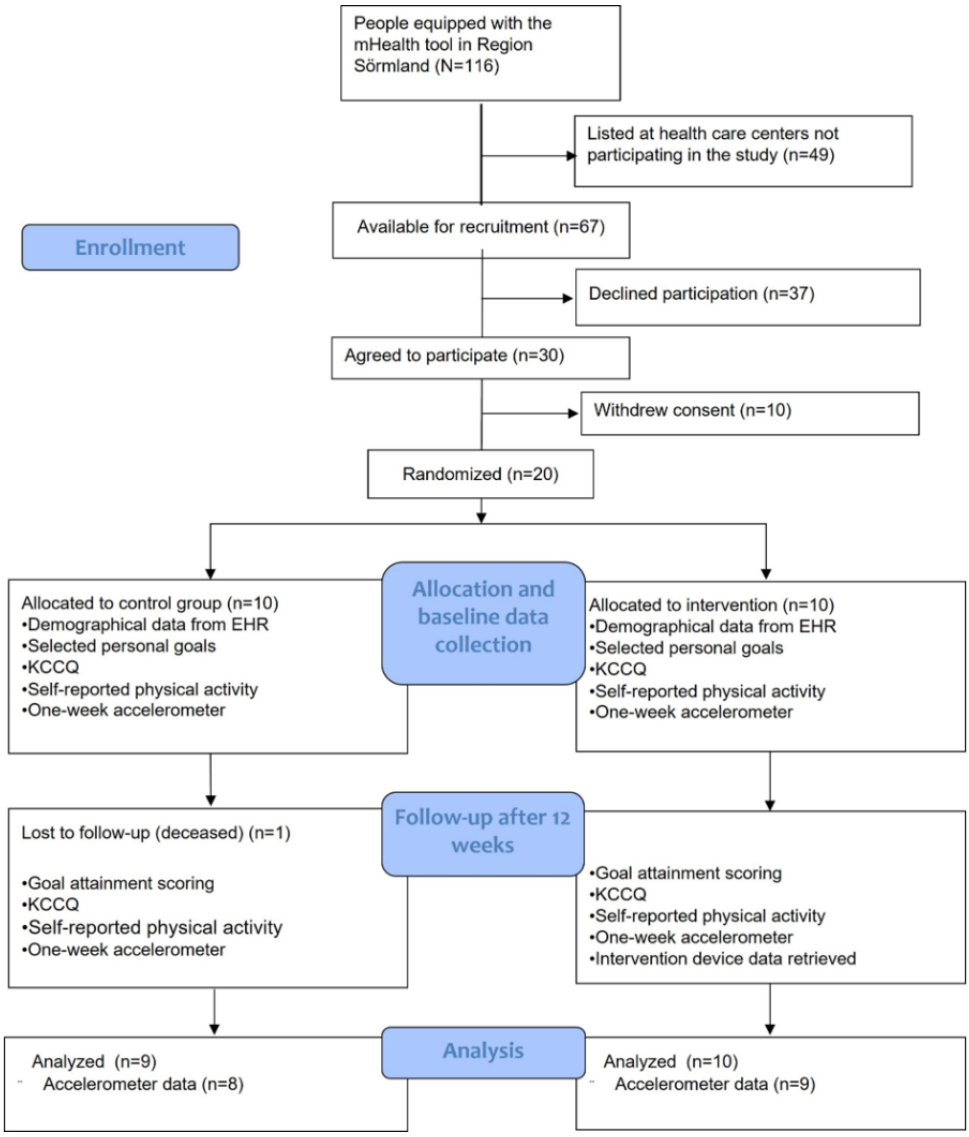

In total, 116 people were using the mHealth tool as part of their regular self-care at the start of this study. These 116 potential participants were listed at 11 different primary health care centers, 5 (45%) of which agreed to participate in the study. At these 5 health care centers, 67 (58%) of the 116 people were listed and were contacted and asked to participate. Participants were recruited between September 13, 2022, and March 25, 2023.

Baseline Data

The participants' baseline demographical data are summarized in Table 2. No significant differences between the 2 study arms were detected.

XSL•FO
RenderX

Table 2. Baseline data, including demographics, etiology, comorbidities, and pharmacological treatment, for study participants.

Characteristics	Control group (n=10)	Intervention group (n=10)	Total (N=20)
Age (years), mean (SD)	77 (5)	78 (9)	78 (7)
Females, n (%)	4 (40)	4 (40)	8 (40)
HF[a] type, n (%)			
HFrEF[b]	4 (40)	3 (30)	7 (35)
HFmrEF[c]	1 (10)	4 (40)	5 (25)
HFpEF[d]	5 (50)	3 (30)	8 (40)
NYHA[e] class, n (%)			
NYHA II	4 (40)	6 (60)	10 (50)
NYHA III	6 (60)	4 (40)	10 (50)
Comorbidities, n (%)			
Diabetes	3 (30)	2 (20)	5 (25)
COPD[f]	3 (30)	2 (20)	5 (25)
Hypertension	4 (40)	6 (60)	10 (50)
Atrial fibrillation	3 (30)	5 (50)	8 (40)
Kidney disease	2 (20)	3 (30)	5 (25)
ACEi[g]/ARB[h]/ARNI[i]	8 (80)	10 (100)	18 (90)
Beta blocker	7 (70)	10 (100)	17 (85)
MRA[j]	1 (10)	2 (20)	3 (15)
SGLT2i[k]	2 (20)	5 (50)	7 (35)
Physically inactive	7 (70)	8 (80)	15 (75)

[a]HF: heart failure.
[b]HFrEF: heart failure with reduced ejection fraction.
[c]HFmrEF: heart failure with mildly reduced ejection fraction.
[d]HFpEF: heart failure with preserved ejection fraction.
[e]NYHA: New York Heart Association.
[f]COPD: chronic obstructive pulmonary disease.
[g]ACEi: angiotensin-converting enzyme inhibitor.
[h]ARB: angiotensin receptor blocker.
[i]ARNI: angiotensin receptor blocker neprilysin inhibitor.
[j]MRA: mineralocorticoid receptor antagonist.
[k]SGLT2i: sodium glucose cotransporter 2 inhibitor.

Numbers Analyzed

Of the 20 participants recruited, 1 (5%) participant in the control group died during the study. Another participant in the control group and 1 (5%) in the intervention group did not meet the requirements of wear time for their accelerometer data to be included in the analyses. Hence, accelerometer-based outputs were available for 8 (80%) participants in the control group and 9 (90%) participants in the intervention group. All questionnaire-based data were available for 9 (90%) participants in the control group and 10 (100%) participants in the intervention group.

Research Question 1: Recruitment

Of the 67 participants with HF whom we could contact, 30 (45%) agreed to participate, but 10 (33%) withdrew their consent before being formally recruited and signing the informed consent form. Most did not provide a reason, but 1 (10%) reported a spouse who had gotten sick, and another had hurt himself and did not wish to participate any longer.

Finally, 20 (30%) of the 67 potential participants were recruited and could be randomized to the control group or the intervention group. However, had the amendment to also include people who were assessed to be physically active not been made, only 15 (22%) people would have been recruited and randomized.

XSL•FO
RenderX

Research Question 2: Implementation Fidelity

The daily adherence to registering physical activity in the Activity Coach app was 69% (range 24%-97%), and weekly adherence was 88% (range 58%-100%). On average, participants in the intervention group registered 323 (range 66-422) minutes per week. The lowest weekly adherence was 50%. In addition, 40% (n=8) of the participants had a daily adherence exceeding 90%, while 40% (n=8) had a weekly adherence of 100%. Furthermore, 9 (90%) of the 10 participants equipped with the Activity Coach app chose to enable the goal-setting functionality. The median weekly goal achievement was 83% (range 61%-96%).

Research Question 3: Outcomes

Subjective Goal Attainment

All 10 prespecified goals were selected by at least 1 (5%) study participant. The first goal selected was distributed among the first 5 goals (see Table 1), whereas the choices for the second goal were distributed over 9 of the 10 options, with the most frequently selected goa being reduced shortness of breath. In the control group, 8 (80%) participants experienced overall worsening and 1 (10%) reported improvement, whereas in the intervention group, 7 (70%) participants reported improvement, 2 (20%) reported no difference, and 1 (10%) reported overall worsening.

The mean goal attainment score for the 2 groups was –1.0 (SD 1.1) for the control group versus 0.6 (SD 0.6) for the intervention group (*P*=.001). This is illustrated in Figure 2.

Figure 2. Goal attainment score by group, calculated as the mean (SD) value for the 2 selected goals.

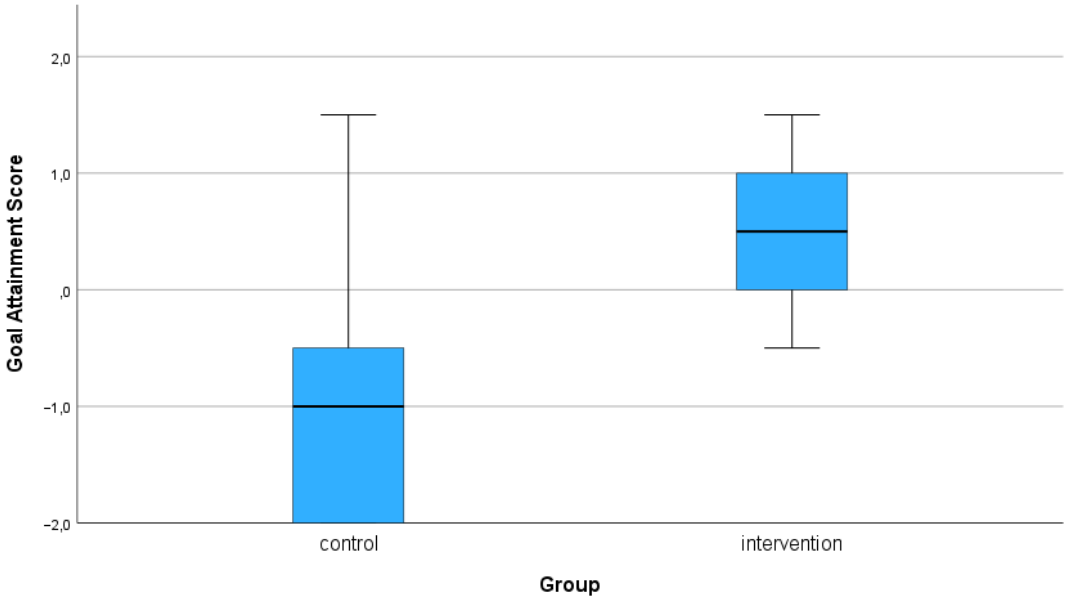

Health-Related Quality of Life

In the control group the mean KCCQ summary score was 59 (SD 25) at baseline compared to 67 (SD 22) in the intervention group (*P*=0.6). At follow-up, the scores were 54 (SD 25) and 70 (SD 21) for the control and the intervention group, respectively (*P*=.18). ANCOVA to control for baseline differences yielded *P*=.04. The mean change in the overall summary score was –9 (SD 10) in the control group versus 3 (SD 13) in the intervention group (*P*=.03) and is illustrated in Figure 3, while the other KCCQ subscales are presented in Table 3.

XSL•FO
RenderX

Figure 3. Change in the KCCQ overall summary score by group at follow-up after 12 weeks. KCCQ: Kansas City Cardiomyopathy Questionnaire.

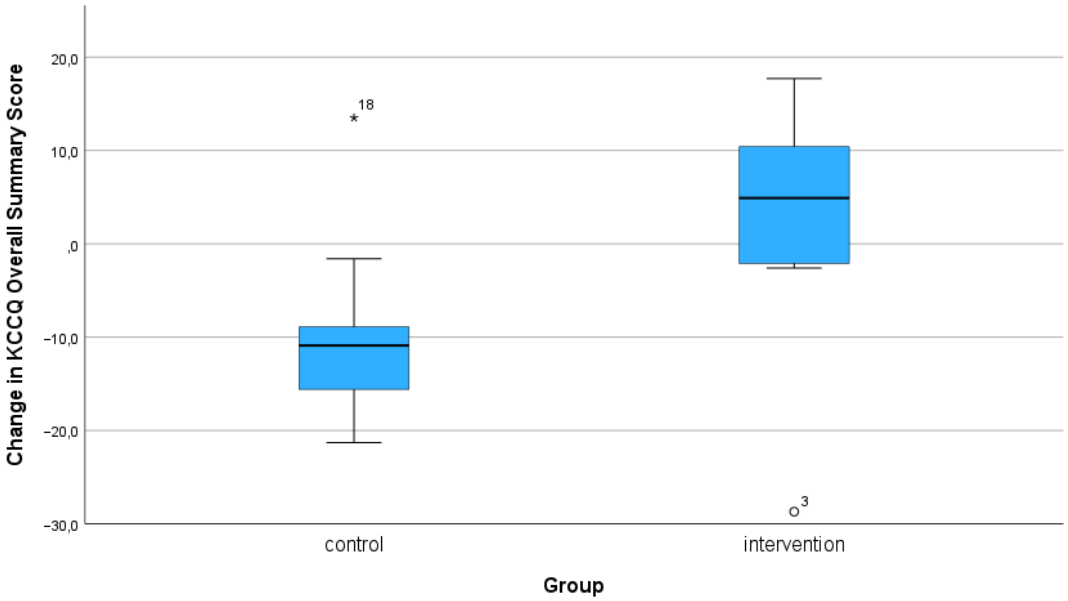

Table 3. KCCQ[a] output (summary score and subdimensions) at baseline and follow-up.

Subdimensions and time	Control group score, mean (SD)	Intervention group score, mean (SD)	*P* value
Physical limitation			
Baseline	57 (26)	65 (28)	.46
Follow-up	45 (27)	71 (20)	.04
Symptom frequency			
Baseline	57 (33)	67 (26)	.38
Follow-up	56 (26)	66 (30)	.29
QoL[b]			
Baseline	60 (19)	67 (19)	.62
Follow-up	63 (26)	68 (21)	.74
Social limitation			
Baseline	61 (33)	69 (31)	.62
Follow-up	54 (30)	74 (26)	.15
Summary score			
Baseline	59 (25)	67 (22)	.55
Follow-up	54 (25)	70 (21)	.18

[a]KCCQ: Kansas City Cardiomyopathy Questionnaire.

[b]QoL: quality of life.

Objective Physical Activity

The accelerometer-derived data are summarized in Table 4. Participants spent, on average, 76% (SD 7%) of the time in a sedentary state at baseline (control group: mean 79%, SD 6%; intervention group: mean 74%, SD 8%; *P*=.20), which at follow-up was 74% (SD 10%) of the time (control group: mean 78%, SD 8%; intervention group: mean 71%, SD 12%; *P*=.22). The decrease was mean –2.6% (SD 7%) in the intervention group versus –0.91% (SD 5%) in the control group (*P*=.59).

The average length of sedentary bouts at baseline was 408 (SD 87) minutes (control group: mean 431, SD 72 minutes; intervention group: mean 392, SD 96 minutes; *P*=.38) based on all the data recorded during the week of wearing the accelerometer. Over the course of the study, the average length of sedentary bouts increased by 27 minutes to 458 (SD 84) minutes in the control group and decreased by 0.70 minutes to 391 (SD 117) minutes in the intervention group (*P*=.22).

On average, participants walked 2735 (SD 2481) steps per day at baseline (control group: mean 2058, SD 1356 steps; intervention group: mean 3210, SD 3019 steps; *P*=.36). The

XSL•FO
RenderX

average number of steps per day increased by 861 to 2919 (SD 2303) steps in the control group and by 322 to 3532 (SD 3208) steps in the intervention group (*P*=.67). The time spent in light physical activity (LPA) increased by 8 minutes in the control group and by 42 minutes in the intervention group (*P*=.07). Moderate-to-vigorous physical activity (MVPA) and LPA data are included in Table 4.

Table 4. Overview of objectively measured physical activity obtained from the accelerometer at baseline and 12-week follow-up.

Data items and time	Control group, mean (SD)	Intervention group, mean (SD)	*P* value
Sedentary behavior (%)			
Baseline	79 (6)	74 (8)	.33
Follow-up	78 (8)	71 (12)	.33
Average length of sedentary bouts (minutes)			
Baseline	431 (72)	392 (96)	.56
Follow-up	458 (84)	391 (117)	.14
Steps per day, n			
Baseline	2058 (1356)	3210 (3019)	.33
Follow-up	2919 (2303)	3532 (3208)	.70
LPA[a] (minutes)			
Baseline	138 (47)	169 (52)	.27
Follow-up	146 (46)	207 (80)	.07
MVPA[b] (minutes)			
Baseline	15 (11)	31 (35)	.54
Follow-up	21 (24)	33 (42)	.96

[a]LPA: light physical activity.

[b]MVPA: moderate-to-vigorous physical activity.

Ancillary Analyses

Since the original trial design was modified to also include people who were not physically inactive, ancillary analyses were performed to study what impact, if any, that had on the data. No significant differences appeared. The differences are listed in Table 5.

A comparison of the steps per day for the physically active versus physically inactive participants is shown in Figure 4.

Table 5. Comparison of objectively measured physical activity between groups with the original data set and the data set including only participants classified as physically inactive (n=15, 75%).

Data items and time	Control group, mean (SD)	Intervention group, mean (SD)
Steps per day at baseline, n		
Original data set	2058 (1356)	3210 (3019)
Physically inactive	1459 (814)	1952 (856)
Steps per day at follow-up, n		
Original data set	2919 (2303)	3532 (3208)
Physically inactive	2194 (1943)	2218 (1139)
Average length of sedentary bouts at baseline (minutes)		
Original data set	431 (72)	392 (96)
Physically inactive	455 (71)	406 (99)
Average length of sedentary bouts at follow-up (minutes)		
Original data set	458 (84)	391 (117)
Physically inactive	495 (55)	409 (119)

XSL•FO
RenderX

Figure 4. Accelerometer-measured steps per day for physically active versus physically inactive participants.

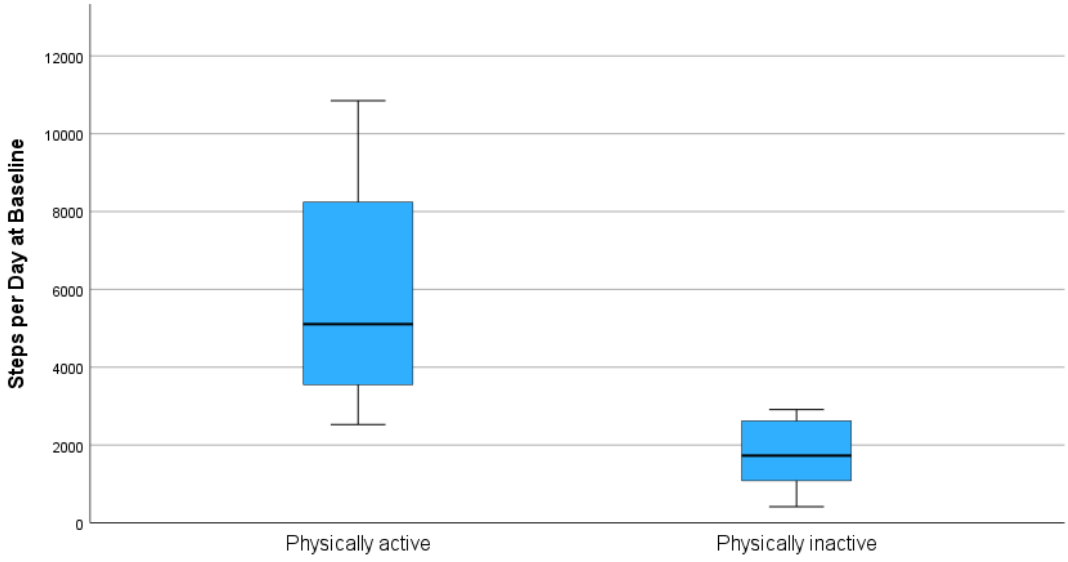

Discussion

Principal Findings

The results of this study show that it may be challenging to recruit physically inactive older adults with HF to participate in an mHealth study, but there might be benefits in using mHealth to support this population in achieving goals relating to physical activity and reduce sedentary behavior.

Recruitment

It is known to be challenging to recruit older people with severe illness for studies evaluating digital health technologies [27]. It seems reasonable that seeking to recruit physically inactive people makes it even more likely that they decline participation. Only 22% of the people contacted for this study were successfully recruited and randomized, although 45% initially consented to participate, but every third person withdrew consent prior to starting the trial. It might, therefore, be appropriate to investigate whether something in the recruitment process could be improved (eg, the information material or striving to initiate the trial at the same time as collecting the consent). Reasons given for withdrawal were typically negative life events (eg, having been hospitalized, having received a new diagnosis, or having a spouse whose health had deteriorated). The withdrawal rate was lower in patients who were recruited via telephone versus via mail.

The recruitment rate in this study of 22% was low compared to other studies, where the corresponding rates ranged from 70% to 90%, although these numbers come from studies in which participants were recruited in the in-patient setting [28,29]. It should also be noted that aside from the 1 participant who died, there was no attrition during the study, and all participants who started also finished the study.

Based on our experiences from this study, we conclude that researchers must be active and may have to call up each possible participant and that merely sending out letters via regular mail is likely insufficient.

Implementation Fidelity

To be able to judge the quality of a study and to properly replicate it, some reporting of intervention fidelity is required, and examples of aspects of intervention fidelity to report are intervention design and intervention delivery or receipt of the intervention [30]. The intervention design has been published elsewhere [9], and a qualitative study to investigate user experiences was beyond the scope of this study. An acceptable intervention delivery ensures that results from the intervention are generalizable [31]. Delivery of the intervention can be reported as intervention adherence [18], which is what we reported in this manuscript. Other mHealth studies have used 60% adherence as a cutoff point for acceptable level of use [17,32], and that level was also adopted in the development of the Activity Coach app [9]. In this study, the median adherence was 69%, which is higher than what is normally reported for mHealth interventions [19] and also higher than that in the Activity Coach feasibility study, suggesting an improvement in usability following the final adjustments made to the Activity Coach app after the feasibility study [9]. The 2022 systematic review by Jakob et al [19] concluded that factors that drive adherence are tailoring of content to the user and gamification, both of which are present in the Activity Coach app and may explain the better-than-average adherence. What level of adherence to an intervention is acceptable should depend on which intervention is being studied. The current guidelines on physical activity from the World Health Organization (WHO) for the general population [33], as well as guidelines for people with HF [34], make recommendations on a weekly basis, so it may be reasonable to also assess adherence to a tool supporting physical activity on a weekly basis; in this study, weekly adherence to the Activity Coach app was 88%, which is similar to other rates in the published literature, which ranged from 55% to 80%, depending on the definition of adherence [29,35-39].

XSL•FO
RenderX

Outcomes

Plausibility of Outcomes

Questionnaires were completed by 19 of the 20 participants, and system (Activity Coach) data from all 10 participants in the intervention group and accelerometer data from 17 of the 20 participants passed wear time scrutiny. The use of a prespecified list of personal goals to choose from, which was a learning from a feasibility study [9], worked well, and each participant successfully selected 2 goals and could grade them at follow-up. All data were analyzable and of acceptable quality.

Limited Efficacy Testing

Goal attainment is an important part of person-centered care [40,41] and is also more responsive to change than other measures [42]. The 5-point scale we used is considered the norm when using goal attainment as an outcome [43,44]. There was a significant difference in goal attainment between the 2 groups, and although there was a subtle improvement in the intervention group, there was a more pronounced worsening in the control group.

It is difficult to assess the clinical importance of a goal attainment scaling outcome, but a 25% shift on an ordinal scale has previously been used as a clinically meaningful change [41,45]. The shift in this study was of a magnitude of 32% of the used scale. In the light of improvements in the HR-QoL, this might imply that there is a clinically meaningful effect.

The KCCQ results mirrored the subjective goal attainment results, in that the KCCQ summary score showed a 12-point difference between the 2 groups, expressed as a 5% improvement in the intervention group and a 15% worsening in the control group. This difference is of a similar size as effects reported from another mHealth intervention aimed at increasing physical activity, which reported a 6-point difference, also at 12-week follow-up [46]. Other telehealth interventions, albeit not strict mHealth, have reported smaller or no improvements in the KCCQ score [47-50]. The difference between the 2 groups at baseline in the summary score was large, albeit nonsignificant, and when that difference was corrected for, the difference between the 2 groups at follow-up also became significant.

Looking at the subscales, the physical limitation dimension was the subdomain with the largest effect, where the intervention group at follow-up had a 37% higher score. A change in that subscale exceeding 8.33 points has been identified as the minimal clinically important difference for improvement and a decrease of 4.17 or more indicates deterioration [51]. In this study, the score of the intervention group increased by 6 points and that of the control group decreased by 12 points, which suggests a clinically meaningful deterioration in the control group.

There was a significant difference of 4135 in the steps per day between the physically inactive versus active participants, which needs to be considered when interpreting the results from the objective measurements of physical activity. The change in the time spent in sedentary behavior was of a magnitude similar to that in other reports on physical activity interventions [52]. The differences between the 2 study arms were more pronounced after removing the physically active participants, with an increase of 40 minutes in the control group compared to an increase of 3 minutes in the intervention group.

Longer sedentary bouts are associated with poor health outcomes [53-55]. For these effects to manifest into outcomes, the reduction in sedentary time needs to be shifted to MVPA according to Rossen et al [56], which was, however, disputed by Diaz et al [57]. There is still a knowledge gap regarding the impact of the sedentary bout length and the effects of changing it [58].

There was no significant difference in the steps per day between the 2 groups, neither at baseline nor at follow-up, although the number of steps per day increased more in the control group (861 steps) compared to the intervention group (322 steps). The increase in steps in the control group is large enough to potentially be clinically meaningful [59], although successful interventions aimed at increasing the step count in chronically ill older adults usually report larger effects (eg, mean ~2000, SD 500, more steps per day or above) [28,60,61]. It is not possible to say whether this effect on the control group was a consequence of the small sample size or external factors. Perhaps there is a connection between more steps per day and longer sedentary bouts. Other research has also proven the positive effects of mHealth interventions on physical activity outcomes [62].

Overall, it appears that the Activity Coach app does not improve the outcomes studied, but it may offer protection against deterioration, as the control group experienced worsened situations relating to their individual goals, a significantly worse HR-QoL, and a trend toward increased length of sedentary bouts. The Activity Coach app does not necessary tell the user to go out and walk, but it focuses more on subtle suggestions, such as standing up a couple of times or doing leg kicks during television commercials, shifting the balance between the feet, doing calf raises while doing the dishes, or using an exercise band. Assuming this is adhered to, it would explain a perceived increase in exercise and improvements in the HR-QoL and goal attainment, without necessarily showing up on the step counts. Possibly, it is manifested in the shorter average lengths of sedentary bouts.

Combining the recruitment rate with the effect sizes on the outcomes measured showed that a future efficacy study designed in the same way, with the primary endpoint being either the KCCQ summary score or sedentary time, needs to contact approximately 250-350 people to be sufficiently powerful.

Limitations

The original study design was modified, and as seen in the ancillary analyses, there were significant differences in daily physical activity between physically active and physically inactive participants, which makes the results more difficult to interpret.

Fewer participants than desired were included, and even though the study was not designed based on specific power calculations, a larger included population would have provided better grounds for designing an efficacy trial, a more well-grounded sample size estimation, and improved generalizability.

XSL•FO
RenderX

Conclusion

The recruitment rate of this study, where older adults with HF were to evaluate an mHealth tool, was 22%, which is a low rate. An active recruitment process is advised if a future efficacy study is to be performed. Adherence to the Activity Coach intervention was better than average at 69%. Deterioration in health and the QoL may be attenuated by the Activity Coach app, possibly by increasing certain aspects of physical activity and by reducing the length of sedentary bouts.

Acknowledgments

We would like to acknowledge the contribution of heart failure nurses and medical doctors working at the participating primary health care centers, who were helpful in reviewing the electronic health records of recruited patients. This study received no external funding.

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

AB, MB, AS, and TJ handled the conceptualization of the study. AB, MB, LK, AS, and TJ managed the methodology. AB, LK, and TJ performed validation. AB was responsible for the software, formal analysis, investigation, writing the original draft of the manuscript, visualization, and project administration. LK and TJ handled the resources. MB, LK, AS, and TJ managed the review and editing of the manuscript and study supervision. All authors have read and agreed to the published version of the manuscript. In addition, TJ assumed responsibility for the paper, including for the accuracy and appropriateness of the reference list.

Conflicts of Interest

AB is an employee of CareLigo AB, the company behind Optilogg. An independent researcher at Linköping University entered all the data.

Multimedia Appendix 1
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 1165 KB - formative_v9i1e62910_app1.pdf]

References

1. Savarese G, Becher P, Lund L, Seferovic P, Rosano G, Coats A. Global burden of heart failure: a comprehensive and updated review of epidemiology. Cardiovasc Res 2023 Jan 18;118(17):3272-3287. [doi: 10.1093/cvr/cvac013] [Medline: 35150240]
2. Inamdar A, Inamdar A. Heart failure: diagnosis, management and utilization. J Clin Med 2016 Jun 29;5(7):62 [FREE Full text] [doi: 10.3390/jcm5070062] [Medline: 27367736]
3. Tremblay MS, Aubert S, Barnes JD, Saunders TJ, Carson V, Latimer-Cheung AE, SBRN Terminology Consensus Project Participants. Sedentary behavior research network (SBRN)–terminology consensus project process and outcome. Int J Behav Nutr Phys Act 2017 Jun 10;14(1):1-17 [FREE Full text] [doi: 10.1186/s12966-017-0525-8] [Medline: 28599680]
4. Thivel D, Tremblay A, Genin PM, Panahi S, Rivière D, Duclos M. Physical activity, inactivity, and sedentary behaviors: definitions and implications in occupational health. Front Public Health 2018;6:288 [FREE Full text] [doi: 10.3389/fpubh.2018.00288] [Medline: 30345266]
5. Jones NR, Roalfe AK, Adoki I, Hobbs FR, Taylor CJ. Survival of patients with chronic heart failure in the community: a systematic review and meta-analysis. Eur J Heart Fail 2019 Nov 16;21(11):1306-1325 [FREE Full text] [doi: 10.1002/ejhf.1594] [Medline: 31523902]
6. Doukky R, Mangla A, Ibrahim Z, Poulin M, Avery E, Collado FM, et al. Impact of physical inactivity on mortality in patients with heart failure. Am J Cardiol 2016 Apr 01;117(7):1135-1143 [FREE Full text] [doi: 10.1016/j.amjcard.2015.12.060] [Medline: 26853954]
7. Park LG, Dracup K, Whooley MA, McCulloch C, Lai S, Howie-Esquivel J. Sedentary lifestyle associated with mortality in rural patients with heart failure. Eur J Cardiovasc Nurs 2019 Apr 21;18(4):318-324. [doi: 10.1177/1474515118822967] [Medline: 30663898]
8. Zhao Q, Chen C, Zhang J, Ye Y, Fan X. Sedentary behavior and health outcomes in patients with heart failure: a systematic review and meta-analysis. Heart Fail Rev 2022 Jul 22;27(4):1017-1028. [doi: 10.1007/s10741-021-10132-7] [Medline: 34159521]

XSL•FO
RenderX

9. Blomqvist A, Bäck M, Klompstra L, Strömberg A, Jaarsma T. Usability and feasibility analysis of an mHealth-tool for supporting physical activity in people with heart failure. BMC Med Inform Decis Mak 2024 Feb 12;24(1):44 [FREE Full text] [doi: 10.1186/s12911-024-02452-z] [Medline: 38347499]
10. Skivington K, Matthews L, Simpson SA, Craig P, Baird J, Blazeby JM, et al. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. BMJ 2021 Sep 30;374:n2061. [doi: 10.1136/bmj.n2061] [Medline: 34593508]
11. Thabane L, Ma J, Chu R, Cheng J, Ismaila A, Rios LP, et al. A tutorial on pilot studies: the what, why and how. BMC Med Res Methodol 2010 Jan 06;10:1 [FREE Full text] [doi: 10.1186/1471-2288-10-1] [Medline: 20053272]
12. Lancaster GA. Pilot and feasibility studies come of age!. Pilot Feasibility Stud 2015;1(1):1 [FREE Full text] [doi: 10.1186/2055-5784-1-1] [Medline: 29611687]
13. Eldridge SM, Chan CL, Campbell MJ, Bond CM, Hopewell S, Thabane L, PAFS consensus group. CONSORT 2010 statement: extension to randomised pilot and feasibility trials. BMJ 2016 Oct 24;355:i5239 [FREE Full text] [doi: 10.1136/bmj.i5239] [Medline: 27777223]
14. Blomqvist A, Bäck M, Klompstra L, Strömberg A, Jaarsma T. Utility of single-item questions to assess physical inactivity in patients with chronic heart failure. ESC Heart Fail 2020 Aug 06;7(4):1467-1476 [FREE Full text] [doi: 10.1002/ehf2.12709] [Medline: 32372549]
15. Jaarsma T, Strömberg A, Mårtensson J, Dracup K. Development and testing of the European Heart Failure Self-Care Behaviour Scale. Eur J Heart Fail 2003 Jun 09;5(3):363-370 [FREE Full text] [doi: 10.1016/s1388-9842(02)00253-2] [Medline: 12798836]
16. Sahlin D, Rezanezad B, Edvinsson ML, Bachus E, Melander O, Gerward S. Self-care management intervention in heart failure (SMART-HF)-a multicentre randomized controlled trial. J Card Fail 2022 Jan;28(1):3-12 [FREE Full text] [doi: 10.1016/j.cardfail.2021.06.009] [Medline: 34161807]
17. Melin M, Hägglund E, Ullman B, Persson H, Hagerman I. Effects of a tablet computer on self-care, quality of life, and knowledge: a randomized clinical trial. J Cardiovasc Nurs 2018;33(4):336-343. [doi: 10.1097/JCN.0000000000000462] [Medline: 29369123]
18. Lambert JD, Greaves CJ, Farrand P, Cross R, Haase AM, Taylor AH. Assessment of fidelity in individual level behaviour change interventions promoting physical activity among adults: a systematic review. BMC Public Health 2017 Oct 02;17(1):765 [FREE Full text] [doi: 10.1186/s12889-017-4778-6] [Medline: 28969669]
19. Jakob R, Harperink S, Rudolf AM, Fleisch E, Haug S, Mair JL, et al. Factors influencing adherence to mHealth apps for prevention or management of noncommunicable diseases: systematic review. J Med Internet Res 2022 May 25;24(5):e35371. [doi: 10.2196/35371] [Medline: 35612886]
20. Urach S, Gaasterland C, Posch M, Jilma B, Roes K, Rosenkranz G, et al. Statistical analysis of goal attainment scaling endpoints in randomised trials. Stat Methods Med Res 2019 Jun;28(6):1893-1910 [FREE Full text] [doi: 10.1177/0962280218777896] [Medline: 29921167]
21. Spertus JA, Jones PG. Development and validation of a short version of the Kansas City Cardiomyopathy Questionnaire. Circ: Cardiovasc Qual Outcomes 2015 Sep;8(5):469-476. [doi: 10.1161/circoutcomes.115.001958]
22. Patel H, Ekman I, Spertus JA, Wasserman SM, Persson L. Psychometric properties of a Swedish version of the Kansas City Cardiomyopathy Questionnaire in a chronic heart failure population. Eur J Cardiovasc Nurs 2008 Sep;7(3):214-221. [doi: 10.1016/j.ejcnurse.2007.08.005] [Medline: 18060843]
23. Loyen A, Clarke-Cornwell AM, Anderssen SA, Hagströmer M, Sardinha LB, Sundquist K, et al. Sedentary time and physical activity surveillance through accelerometer pooling in four European countries. Sports Med 2017 Jul 10;47(7):1421-1435 [FREE Full text] [doi: 10.1007/s40279-016-0658-y] [Medline: 27943147]
24. Sasaki JE, John D, Freedson PS. Validation and comparison of ActiGraph activity monitors. J Sci Med Sport 2011 Sep;14(5):411-416. [doi: 10.1016/j.jsams.2011.04.003] [Medline: 21616714]
25. Sim J, Lewis M. The size of a pilot study for a clinical trial should be calculated in relation to considerations of precision and efficiency. J Clin Epidemiol 2012 Mar;65(3):301-308. [doi: 10.1016/j.jclinepi.2011.07.011] [Medline: 22169081]
26. Julious SA. Sample size of 12 per group rule of thumb for a pilot study. Pharm Stat 2005 Nov 24;4(4):287-291. [doi: 10.1002/pst.185]
27. Poli A, Kelfve S, Motel-Klingebiel A. A research tool for measuring non-participation of older people in research on digital health. BMC Public Health 2019 Nov 08;19(1):1487 [FREE Full text] [doi: 10.1186/s12889-019-7830-x] [Medline: 31703655]
28. Saleh ZT, Elshatarat RA, Elhefnawy KA, Helmi Elneblawi N, Abu Raddaha AH, Al-Za'areer MS, et al. Effect of a home-based mobile health app intervention on physical activity levels in patients with heart failure: a randomized controlled trial. J Cardiovasc Nurs 2022 Apr 6;38(2):128-139. [doi: 10.1097/jcn.0000000000000911] [Medline: 00005082-202303000-00008]
29. Indraratna P, Biswas U, McVeigh J, Mamo A, Magdy J, Vickers D, et al. A smartphone-based model of care to support patients with cardiac disease transitioning from hospital to the community (TeleClinical care): pilot randomized controlled trial. JMIR Mhealth Uhealth 2022 Feb 28;10(2):e32554 [FREE Full text] [doi: 10.2196/32554] [Medline: 35225819]

XSL•FO
RenderX

30. Murphy SL, Gutman SA. Intervention fidelity: a necessary aspect of intervention effectiveness studies. Am J Occup Ther 2012;66(4):387-388. [doi: 10.5014/ajot.2010.005405] [Medline: 22742685]
31. Evans E, Zengul A, Subhash Chilke T, Knight A, Willig A, Cherrington A, et al. Telehealth behavioral intervention for chronic disease self-management in adults with physical disabilities (My Health, My Life, My Way): protocol for intervention fidelity and dashboard design for a randomized controlled trial. JMIR Res Protoc 2024 Feb 12;13:e53410 [FREE Full text] [doi: 10.2196/53410] [Medline: 38345845]
32. Lyngå P, Persson H, Hägg-Martinell A, Hägglund E, Hagerman I, Langius-Eklöf A, et al. Weight monitoring in patients with severe heart failure (WISH). A randomized controlled trial. Eur J Heart Fail 2012 Apr 18;14(4):438-444 [FREE Full text] [doi: 10.1093/eurjhf/hfs023] [Medline: 22371525]
33. Bull FC, Al-Ansari SS, Biddle S, Borodulin K, Buman MP, Cardon G, et al. World Health Organization 2020 guidelines on physical activity and sedentary behaviour. Br J Sports Med 2020 Dec 25;54(24):1451-1462 [FREE Full text] [doi: 10.1136/bjsports-2020-102955] [Medline: 33239350]
34. McDonagh T, Metra M, Adamo M, Gardner R, Baumbach A, Böhm M. Correction to: 2023 focused update of the 2021 ESC guidelines for the diagnosis and treatment of acute and chronic heart failure: developed by the task force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC) with the special contribution of the Heart Failure Association (HFA) of the ESC. Eur Heart J 2024 Jan 01;45(1):53-726. [doi: 10.1093/eurheartj/ehad613] [Medline: 37995350]
35. Vuorinen A, Leppänen J, Kaijanranta H, Kulju M, Heliö T, van Gils M, et al. Use of home telemonitoring to support multidisciplinary care of heart failure patients in Finland: randomized controlled trial. J Med Internet Res 2014 Dec 11;16(12):e282 [FREE Full text] [doi: 10.2196/jmir.3651] [Medline: 25498992]
36. Choi E, Park J, Min D, Ahn S, Ahn J. Heart Failure-Smart Life: a randomized controlled trial of a mobile app for self-management in patients with heart failure. BMC Cardiovasc Disord 2023 Jan 09;23(1):7 [FREE Full text] [doi: 10.1186/s12872-023-03039-8] [Medline: 36624373]
37. Chaudhry SI, Mattera JA, Curtis JP, Spertus JA, Herrin J, Lin Z, et al. Telemonitoring in patients with heart failure. N Engl J Med 2010 Dec 09;363(24):2301-2309 [FREE Full text] [doi: 10.1056/NEJMoa1010029] [Medline: 21080835]
38. Koehler F, Winkler S, Schieber M, Sechtem U, Stangl K, Böhm M, TIM-HF Investigators. Telemedical interventional monitoring in heart failure (TIM-HF), a randomized, controlled intervention trial investigating the impact of telemedicine on mortality in ambulatory patients with heart failure: study design. Eur J Heart Fail 2010 Dec;12(12):1354-1362 [FREE Full text] [doi: 10.1093/eurjhf/hfq199] [Medline: 21098580]
39. Seto E, Leonard KJ, Cafazzo JA, Barnsley J, Masino C, Ross HJ. Mobile phone-based telemonitoring for heart failure management: a randomized controlled trial. J Med Internet Res 2012 Feb 16;14(1):e31 [FREE Full text] [doi: 10.2196/jmir.1909] [Medline: 22356799]
40. Giovannetti ER, Clair CA, Jennings LA, Sandberg SF, Bowman A, Reuben DB, et al. Standardised approach to measuring goal-based outcomes among older disabled adults: results from a multisite pilot. BMJ Qual Saf 2021 Feb 09;30(2):157-166. [doi: 10.1136/bmjqs-2019-010742] [Medline: 33037142]
41. Rietkerk W, Uittenbroek RJ, Gerritsen DL, Slaets JPJ, Zuidema SU, Wynia K. Goal planning in person-centred care supports older adults receiving case management to attain their health-related goals. Disabil Rehabil 2021 Jun 07;43(12):1682-1691. [doi: 10.1080/09638288.2019.1672813] [Medline: 31589075]
42. Rockwood K, Joyce B, Stolee P. Use of goal attainment scaling in measuring clinically important change in cognitive rehabilitation patients. J Clin Epidemiol 1997 May;50(5):581-588. [doi: 10.1016/s0895-4356(97)00014-0] [Medline: 9180650]
43. Logan B, Jegatheesan D, Viecelli A, Pascoe E, Hubbard R. Goal attainment scaling as an outcome measure for randomised controlled trials: a scoping review. BMJ Open 2022 Jul 22;12(7):e063061 [FREE Full text] [doi: 10.1136/bmjopen-2022-063061] [Medline: 35868829]
44. Roberts MA, Abery BH. A person-centered approach to home and community-based services outcome measurement. Front Rehabil Sci 2023 Feb 1;4:1056530 [FREE Full text] [doi: 10.3389/fresc.2023.1056530] [Medline: 36817716]
45. Farrar JT, Young JP, LaMoreaux L, Werth JL, Poole RM. Clinical importance of changes in chronic pain intensity measured on an 11-point numerical pain rating scale. Pain 2001 Nov;94(2):149-158. [doi: 10.1016/S0304-3959(01)00349-9] [Medline: 11690728]
46. Felker GM, Sharma A, Mentz RJ, She L, Green CL, Granger BB, et al. A randomized controlled trial of mobile health intervention in patients with heart failure and diabetes. J Card Fail 2022 Nov;28(11):1575-1583. [doi: 10.1016/j.cardfail.2022.07.048] [Medline: 35882260]
47. Bakitas MA, Dionne-Odom JN, Ejem DB, Wells R, Azuero A, Stockdill ML, et al. Effect of an early palliative care telehealth intervention vs usual care on patients with heart failure: the ENABLE CHF-PC randomized clinical trial. JAMA Intern Med 2020 Sep 01;180(9):1203-1213 [FREE Full text] [doi: 10.1001/jamainternmed.2020.2861] [Medline: 32730613]
48. Collins SP, Liu D, Jenkins CA, Storrow AB, Levy PD, Pang PS, et al. Effect of a self-care intervention on 90-day outcomes in patients with acute heart failure discharged from the emergency department: a randomized clinical trial. JAMA Cardiol 2021 Feb 01;6(2):200-208 [FREE Full text] [doi: 10.1001/jamacardio.2020.5763] [Medline: 33206126]

XSL•FO
RenderX

49. Nolan RP, Ross HJ, Farkouh ME, Huszti E, Chan S, Toma M, et al. Automated e-counseling for chronic heart failure: CHF-CePPORT trial. Circ Heart Fail 2021 Jan;14(1):e007073. [doi: 10.1161/CIRCHEARTFAILURE.120.007073] [Medline: 33464959]
50. Johnson AE, Routh S, Taylor CN, Leopold M, Beatty K, McNamara DM, et al. Developing and implementing an mHealth heart failure self-care program to reduce readmissions: randomized controlled trial. JMIR Cardio 2022 Mar 21;6(1):e33286 [FREE Full text] [doi: 10.2196/33286] [Medline: 35311679]
51. Butler J, Spertus J, Bamber L, Khan M, Roessig L, Vlajnic V, VITALITY-HFpEF Study Group. Defining changes in physical limitation from the patient perspective: insights from the VITALITY-HFpEF randomized trial. Eur J Heart Fail 2022 May;24(5):843-850 [FREE Full text] [doi: 10.1002/ejhf.2481] [Medline: 35274420]
52. Swartz AM, Cho CC, Welch WA, Widlansky ME, Maeda H, Strath SJ. Pattern analysis of sedentary behavior change after a walking intervention. Am J Health Behav 2018 May 01;42(3):90-101 [FREE Full text] [doi: 10.5993/AJHB.42.3.9] [Medline: 29663984]
53. Diaz KM, Duran AT, Colabianchi N, Judd SE, Howard VJ, Hooker SP. Potential effects on mortality of replacing sedentary time with short sedentary bouts or physical activity: a national cohort study. Am J Epidemiol 2019 Mar 01;188(3):537-544 [FREE Full text] [doi: 10.1093/aje/kwy271] [Medline: 30551177]
54. Healy GN, Dunstan DW, Salmon J, Cerin E, Shaw JE, Zimmet PZ, et al. Breaks in sedentary time: beneficial associations with metabolic risk. Diabetes Care 2008 Apr;31(4):661-666. [doi: 10.2337/dc07-2046] [Medline: 18252901]
55. Duran AT, Friel CP, Serafini MA, Ensari I, Cheung YK, Diaz KM. Breaking up prolonged sitting to improve cardiometabolic risk: dose-response analysis of a randomized crossover trial. Med Sci Sports Exerc 2023 May 01;55(5):847-855. [doi: 10.1249/MSS.0000000000003109] [Medline: 36728338]
56. Rossen J, Buman MP, Johansson U, Yngve A, Ainsworth B, Brismar K, et al. Reallocating bouted sedentary time to non-bouted sedentary time, light activity and moderate-vigorous physical activity in adults with prediabetes and type 2 diabetes. PLoS One 2017 Jul 28;12(7):e0181053 [FREE Full text] [doi: 10.1371/journal.pone.0181053] [Medline: 28753626]
57. Diaz K, Duran A, Colabianchi N, Judd S, Howard V, Hooker S. Effects of replacing prolonged sedentary bouts with short sedentary bouts or physical activity on mortality: 2552 June 1 2. Med Sci Sports Exerc 2018;50(5S):634. [doi: 10.1249/01.mss.0000537180.26323.85]
58. Byrom B, Stratton G, Mc Carthy M, Muehlhausen W. Objective measurement of sedentary behaviour using accelerometers. Int J Obes (Lond) 2016 Nov;40(11):1809-1812 [FREE Full text] [doi: 10.1038/ijo.2016.136] [Medline: 27478922]
59. Demeyer H, Burtin C, Hornikx M, Camillo CA, Van Remoortel H, Langer D, et al. The minimal important difference in physical activity in patients with COPD. PLoS One 2016 Apr 28;11(4):e0154587 [FREE Full text] [doi: 10.1371/journal.pone.0154587] [Medline: 27124297]
60. Martin SS, Feldman DI, Blumenthal RS, Jones SR, Post WS, McKibben RA, et al. mActive: a randomized clinical trial of an automated mhealth intervention for physical activity promotion. JAHA 2015 Oct 29;4(11):e002239. [doi: 10.1161/jaha.115.002239]
61. Hemnes AR, Silverman-Lloyd LG, Huang S, MacKinnon G, Annis J, Whitmore CS, et al. A mobile health intervention to increase physical activity in pulmonary arterial hypertension. Chest 2021 Sep;160(3):1042-1052 [FREE Full text] [doi: 10.1016/j.chest.2021.04.012] [Medline: 33878341]
62. Tong HL, Alnasser A, Alshahrani NZ, Bawaked RA, AlAhmed R, Alsukait RF, et al. The use of mobile technologies to promote physical activity and reduce sedentary behaviors in the Middle East and North Africa region: systematic review and meta-analysis. J Med Internet Res 2024 Mar 19;26:e53651 [FREE Full text] [doi: 10.2196/53651] [Medline: 38502160]

Abbreviations

EHR: electronic health record
GUI: graphical user interface
HF: heart failure
HR-QoL: health-related quality of life
KCCQ: Kansas City Cardiomyopathy Questionnaire
LPA: light physical activity
MRC: Medical Research Council
MVPA: moderate-to-vigorous physical activity
QoL: quality of life
RCT: randomized controlled trial

XSL•FO
RenderX

Edited by A Mavragani; submitted 04.06.24; peer-reviewed by P Dunn, A Zelko, R Austin; comments to author 22.10.24; revised version received 06.11.24; accepted 29.11.24; published 08.01.25.

Please cite as:
Blomqvist A, Bäck M, Klompstra L, Strömberg A, Jaarsma T
Testing the Recruitment Frequency, Implementation Fidelity, and Feasibility of Outcomes of the Heart Failure Activity Coach Study (HEALTHY): Pilot Randomized Controlled Trial
JMIR Form Res 2025;9:e62910
URL: https://formative.jmir.org/2025/1/e62910
doi:10.2196/62910
PMID:

©Andreas Blomqvist, Maria Bäck, Leonie Klompstra, Anna Strömberg, Tiny Jaarsma. Originally published in JMIR Formative Research (https://formative.jmir.org), 08.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effectiveness of a Videoconference-Based Cognitive Behavioral Therapy Program for Patients with Schizophrenia: Pilot Randomized Controlled Trial

Masayuki Katsushima[1,2], PhD; Hideki Nakamura[3,4], MSc; Yuki Shiko[5,6], PhD; Hideki Hanaoka[6], MD, PhD; Eiji Shimizu[2,4,7], MD, PhD

[1]Department of Rehabilitation, Faculty of Health Care and Medical Sports, Teikyo Heisei University, Ichihara, Japan
[2]Research Center for Child Mental Development, Chiba University, Chiba, Japan
[3]Department of Nursing, Faculty of Medicine, The Jikei University School of Medicine, Tokyo, Japan
[4]Department of Cognitive Behavioral Physiology, Chiba University Graduate School of Medicine, Chiba, Japan
[5]Department of Biostatistics, Graduate School of Medicine, Saitama Medical University, Saitama, Japan
[6]Clinical Research Center, Chiba University Hospital, Chiba, Japan
[7]Cognitive Behavioral Therapy Center, Chiba University Hospital, Chiba, Japan

Corresponding Author:
Masayuki Katsushima, PhD
Department of Rehabilitation, Faculty of Health Care and Medical Sports
Teikyo Heisei University
4-1 UruidoMinami
Ichihara, 290-0193
Japan
Phone: 81 436 74 6915
Fax: 81 436 74 6915
Email: m.katsushima@thu.ac.jp

Abstract

Background: Cognitive behavioral therapy for psychosis (CBTp) is not sufficiently widespread in clinical practice, although evidence has been presented.

Objective: The purpose of this study was to explore whether one-on-one videoconference-based CBTp (vCBTp) is more effective than usual care (UC) alone for improving psychiatric symptoms in patients with schizophrenia attending outpatient clinics.

Methods: In this exploratory randomized controlled trial, patients with schizophrenia and schizoaffective disorders who were still taking medication in an outpatient clinic were randomly assigned to either the vCBTp plus UC group (n=12) or the UC group (n=12). The vCBTp program was conducted once a week, with each session lasting for 50 minutes, for a total of 7 sessions conducted in real-time and in a one-on-one format remotely using a loaned tablet computer (iPad). The primary outcome was the Positive and Negative Syndrome Scale (PANSS) total score, which measures the difference in the mean change from baseline (week 0) to posttest (week 8).

Results: The study included 24 participants. There were no significant differences between the 2 groups at baseline. With regard to significant differences between the 2 groups in terms of the primary outcome, the mean change in the PANSS total score from baseline to week 8 in the vCBTp plus UC group was –9.5 (95% CI –12.09 to –6.91) and the mean change in the UC alone group was 6.9 (95% CI 1.54-12.30). The difference between the 2 groups was significant (P<.001). In addition, significant improvements were observed in the subscales of positive (P<.001) and negative (P=.004) symptoms and general psychopathology (P<.001). Significant differences were also observed in the secondary outcomes of the General Anxiety Disorder-7 (GAD-7; P=.04) and EQ-5D-5L (P=.005). There were no dropouts and no serious adverse events in this study.

Conclusions: A total of 7 remote vCBTp sessions conducted in the vCBTp plus UC group could be safely administered to patients with schizophrenia. They were also observed to be effective for psychiatric symptoms, general anxiety, and quality of life. However, because of the observed worsening of scores in the UC group, caution is required in interpreting significant differences between the 2 groups. This approach is expected to improve accessibility to CBTp for outpatients with schizophrenia

XSL•FO
RenderX

and social anxiety regarding transportation use and financial and physical burdens related to transportation, and to contribute to promoting CBTp acceptability by compensating for the shortage of implementers.

Trial Registration: University Hospital Medical Information Network Clinical Trials Registry UMIN000043396; https://center6.umin.ac.jp/cgi-open-bin/ctr/ctr_view.cgi?recptno=R000049544

International Registered Report Identifier (IRRID): RR2-10.1136/bmjopen-2022-069734

(JMIR Form Res 2025;9:e59540) doi:10.2196/59540

KEYWORDS

schizophrenia; randomized controlled trial; cognitive behavioral therapy; videoconference; remote therapy

Introduction

Background

Schizophrenia is a psychiatric disorder that affects approximately 1% of the total population and is typically prevalent between the ages of 16 and 30 years [1,2]. Pharmacotherapy with antipsychotic drugs is the standard treatment for schizophrenia. Moreover, cognitive behavioral therapy, a psychological intervention, is recommended by the National Institute of Health and Clinical Excellence (NICE) [3]. In the West, cognitive behavioral therapy for psychosis (CBTp) has shown promise in improving positive symptoms and acquiring coping skills [4]. A meta-analysis of CBTp reported a moderate effect of 0.35-0.44 for positive and generalized symptoms [5,6]. Despite the effectiveness of traditional face-to-face CBTp in treating schizophrenia, progress in training therapists to conduct CBTp has been slow [7,8]. The Schizophrenia Commission in the United Kingdom reported that only 10% of patients with schizophrenia receive CBTp due to limited access [9]. Therefore, despite supporting evidence, it is infrequently implemented in clinical practice owing to a shortage of practitioners, making it difficult for patients to receive CBTp [10]. It is thus desirable to create an environment in which effective CBTp can be implemented in the future and increase the number of clinicians who can practice. Evidence of low-intensity CBTp has also been collected, suggesting its efficacy [11,12].

Previous studies have found that community psychiatric nurses effectively increased their awareness of illnesses by conducting as few as 6 sessions of CBTp [13,14]. The 6 CBTp sessions improved their understanding of illness, general health, and depression. NICE recommends a minimum of 16 sessions of CBTp [3]. However, a meta-analysis by Hazell et al [11] reported an effect size of 0.46 for CBTp conducted in 6-15 sessions.

In addition, some studies have reported the incorporation of new theories, such as mindfulness, acceptance and commitment therapy, compassion mind training, and metacognitive therapy, into CBTp [15]. Thus, in recent years, studies have been conducted that have not limited themselves to formulated CBTp but have adjusted the frequency and content of CBTp.

In this context, studies reported that patients with schizophrenia have traumatic memories and may be treated with CBT [16]. However, CBTp does not focus directly on the sequelae of trauma, and therapists are concerned that experiential reprocessing of trauma may exacerbate psychotic symptoms; therefore, the treatment of posttraumatic disorder (PTSD) symptoms often involves reluctance [17]. Nevertheless, 3 reviews [18-20] suggested that trauma-focused psychotherapy is safe and effective for psychotic patients.

In recent years, research on internet-based CBTp has increased [21-25], and a growing interest exists for CBT delivered through a videoconference system called vCBT [26,27]. This method enhances access to CBT for patients in remote areas and enables teletherapy through 2-way real-time communication between the therapist and patient. Therapist-led internet-based CBT has been demonstrated to be equally effective as face-to-face CBT in treating anxiety, depression, insomnia, and physical disorders [28].

There have also been several previous studies targeting schizophrenia. A CBTp effort in a group therapy format using a videoconference method for early psychosis reported positive results regarding efficacy [29]. Moreover, a study of remote sessions based on acceptance commitment therapy in a group therapy format mentioned the potential for both accessibility and recovery effects [30]. In addition, there have been reports of neurocognitive and metacognitive training delivered remotely as group therapy [31]. However, there is still little to no evidence regarding randomized controlled trials (RCTs) of CBTp interventions in a one-on-one videoconference format.

Therefore, we developed a real-time and one-on-one videoconference-based CBTp (vCBTp) program for schizophrenia through an RCT design [32]. The vCBTp program included 2 sessions that addressed intrusive memories of stressful events related to the onset of psychosis.

Objective

This study examined the results and efficacy of an exploratory RCT designed to compare the clinical efficacy of 7 sessions of vCBTp as an adjunct to usual care (UC) versus UC alone. The study included patients with persistent positive symptoms of schizophrenia despite pharmacotherapy.

Methods

Study Design and Participants

This study was designed as a single-center, 2-arm, parallel, rater-blinded, prospective RCT with a 7-week treatment regimen. Study participants were assigned to either the vCBTp plus UC or UC alone groups. The recruitment of participants was posted on the Chiba University Graduate School of Medicine's Cognitive Behavioral Physiology subject recruitment

XSL•FO
RenderX

website. Recruitment posters were displayed in the outpatient waiting room of Chiba University Hospital. Recruitment posters and leaflets were also mailed to medical institutions and employment support facilities in Chiba Prefecture.

Ethical Considerations

All procedures were approved by the Institutional Review Board of Chiba University Hospital in January 2021 (reference number: G2020031). All participants received a thorough verbal and written explanation of the study, and verbal and written consent was obtained from those who were confirmed to have the capacity to consent to the study and who wished to participate. Participation in the clinical trial was voluntary on the part of the subjects and their surrogates, and they were fully informed that they could refuse or withdraw from participation in the clinical trial at any time. The consent obtained also covered that no additional consent would be required for secondary analyses.

Data were anonymized so that they could be identified only by a unique identification number and were maintained at the data center. Consent documents and other materials related to the study were strictly maintained in a locked archive.

Study participants were paid a burden-reduction fee (10,000 JPY [US $65] gift card at baseline and 10,000 JPY [US $65] gift card at the endpoint for a total of 20,000 JPY [US $130]) for assessments at 0 and 8 weeks. This burden-reduction fee was paid equitably to both the vCBTp plus UC group and the UC group.

Inclusion Criteria

The criteria for those eligible to participate in the study were as follows: age 16-65 years; primary diagnosis of schizophrenia, schizoaffective disorder, or delusional disorder according to the Diagnostic and Statistical Manual of Mental Disorders, 5th edition; and satisfactory assessment of competence to provide informed consent using the MacArthur Competence Assessment Tool for Clinical Research (MacCAT-CR) [33,34]. The MacCAT-CR assessed participants' ability to "understand," "recognize," "think validly," and "express choice" about the study. Participants who responded with a score of 0 on any of the items were judged to have poor capacity to consent and were eligible for exclusion. In addition, participants needed to have at least 3 points on any one of the positive symptoms of the Positive and Negative Syndrome Scale (PANSS) [35]. In addition, participants had to be attending their primary psychiatrist, be taking antipsychotic medication for at least 3 months continuously, and not intend to change their medication within the next 3 months. Participants were also required to obtain consent from their primary psychiatrist to participate in the study. An internet environment with wireless fidelity (Wi-Fi) was needed for videoconference using a tablet computer (iPad).

Exclusion Criteria

Those with a diagnosis of alcoholism, drug dependence, intellectual disability, or dementia from their primary psychiatrist were excluded from participation in the study. Those who were predicted to be at risk of self-harm or harm as a result of the clinical trial were also excluded. Risk of self-harm or harm was assessed by responses to outcome items and observation of adverse events at the time of consent and at 0- and 8-week assessments. In addition, patients were screened for suicidal ideation with item 9 of the Patient Health Questionnaire-9 (PHQ-9) [36] at the 0- and 8-week assessments; those scoring 1 or more points were considered positive for suicidal ideation [37]. Patients were assessed by the investigator and a psychiatrist, and those who were assessed as being at imminent risk of suicide and thus unable to complete the study were considered for exclusion. Patients who were hospitalized were excluded from the study.

Sample Size

The main objective of this study was to assess the superiority of vCBTp for symptom improvement in a group of patients with schizophrenia compared with a UC group, referencing a previous study by Morrison et al [38], which compared CBT and antipsychotic groups, where the mean difference between the vCBTp plus UC group and UC group after the intervention was 12 points, with the SD of the pooled UC group assumed to be 8 points. The significance level for both groups was set at 5%, and the power was set at 90%. The number of patients in each group was calculated to be 11. Assuming a participant dropout rate of 10%, the target number of patients was set at 12 in each group.

Eligibility Assessment

Two researchers (a psychiatrist [ES] and a researcher [MK]) conducted the initial consultation. It was conducted face-to-face at Chiba University Hospital. At the hospital, ES, as a psychiatrist, assessed and confirmed eligibility, including the participant's history of treatment by their family physician, severity of symptoms, and ability to consent to participate in the study. MK assessed the PANSS for allocation adjustment factors and the MacCAT-CR for competency to consent to this study. In addition, the researchers ascertained the participants' clinical characteristics, including gender, age, marital status, age at onset, duration of illness, employment status, and treatment history (eg, names of antipsychotic medications, drug titers at baseline, and changes in treatment during the study period).

Randomization and Masking

After ES or MK examined applicants for suitability prior to study enrollment, those who were identified as meeting the criteria and whose consent was obtained were enrolled as cases by the Chiba University Hospital Data Center. After registration, all patient names were key-coded, anonymized, and kept confidential. The Chiba University Hospital Data Center then conducted dynamic allocation of participants using a computer program. Because this was a pilot study and the number of patients was small, 2 allocation adjustment factors were used. Participants were randomized in a 1:1 ratio to the vCBTp plus UC group or UC group using a minimization method with 2 factors as allocation adjustment factors: total baseline PANSS score (PANSS ≥ 51) and gender. The PANSS values were calculated based on the participants' baseline values in the study by Naeem et al [39]. Each participant was randomly assigned to 1 of the 2 treatments. In addition, the data center was strictly controlled to ensure that the allocation program was not divulged

XSL•FO
RenderX

to outside parties. After study entry, the author HN, as an evaluator, was blinded for the duration of the study, and only HN who was an expert in assessing the PANSS at 0 and 8 weeks assessed the PANSS in a semistructured interview format via videoconference. HN was not informed which group participants were assigned to until study completion to ensure fairness. All secondary assessments other than the PANSS involved self-administered questionnaires.

Procedures

Patients who wished to participate in the study had an initial consultation with MK and ES at Chiba University Hospital, where eligibility assessment and consent to participate in the study were explained. At that time, they received an explanation and lecture on the basic operation of the loaned iPad and videoconference connection for vCBTp. With regard to vCBTp, MK participated from Chiba University Hospital, and patients attended sessions from home using the loaned iPad. Participants reported their email address and cell phone number to the study office for contact, but were also required to provide emergency contact information for a family member or support person in case of emergency. Videoconference was used for 0- and 8-week assessments and all vCBTp sessions. The PANSS assessments at weeks 0 and 8 were conducted via videoconference with only HN and the participants. MK did not participate in the videoconference.

All vCBTp sessions were conducted by MK. ES appropriately supervised MK after each session. In addition to research experience in CBT interventions for schizophrenia [40], MK had conducted high-intensity CBT sessions for depression and anxiety at the Cognitive Behavioral Therapy Center at Chiba University Hospital, and had the qualities and fidelity of a therapist. ES had experience in creating remote CBT programs for patients with panic disorder, obsessive-compulsive disorder, social anxiety [26], eating disorders [41], chronic pain [42], and other disorders. The vCBTp program for patients with schizophrenia was created by 2 researchers (ES and MK). The number of sessions in this study was set based on a meta-analysis of low-intensity CBTp [11]. Since the median number of sessions in the 10 previous studies in the meta-analysis was 7.5, the program was created with a total of 7 sessions of vCBTp. The duration of treatment was 7 weeks (50 minutes per session, once a week). Seven sessions were printed as text in a paper booklet and mailed to the vCBTp plus UC group participants after the 0-week assessment. vCBTp sessions were conducted in a one-on-one format between the therapist and the participant. The therapist wrote the session content on slides in real-time while the camera was turned on and a PowerPoint file was shared on the screen in the messaging application software (Microsoft Teams) for videoconference. After each homework session, participants wrote on the text file that had been sent to them in advance and submitted the images to the therapist as email attachments with photos.

In schizophrenia, depression and anxiety are significantly related to positive symptoms [43]. We developed the vCBTp sessions based on "cognitive restructuring," a basic technique incorporated into CBT for depression and anxiety. vCBTp consists of the following 7 sessions: (1) assessment and goal setting, (2) case formulation using cognitive-behavioral models, (3) cognitive restructuring, (4) coping and relaxation, (5) externalization of intrusive memories about stressful events related to the onset of psychosis, (6) re-evaluation of intrusive memories about stressful events related to the onset of psychosis, and (7) relapse prevention (Table 1).

Table 1. Composition and contents of videoconference-based cognitive behavioral therapy for psychosis.

Session	Contents and goals	Homework
Session 1	Assessment and goal setting; identify symptoms of distress and set goals	Record stressful situations felt during the week
Session 2	Case formulation of current stress; use the cognitive-behavioral model to objectively understand events and thoughts/anxiety	Cognitive-behavioral model to record stressful situations felt during the week
Session 3	Cognitive restructuring of current stress; psychological flexibility by mentioning "other ideas;" improvement of confidence	Use the column method to derive and record another way of thinking
Session 4	Coping and relaxation; engage in fun and calming activities and activate behaviors	Record coping done in a week
Session 5	Externalize intrusive memories about stressful events related to the onset of psychosis; use cognitive-behavioral models to capture past stressful experiences and understand core beliefs that have persisted since that time; moreover, engage in self-compassion at that time	Record words of self-compassion about your symptoms and stress felt during the week
Session 6	Reassess intrusive memories about stressful events related to the onset of psychosis; reflect on how you have coped with past stressful experiences and recognize that you have coping skills	Record your coping behaviors about symptoms and stress felt during the week
Session 7	Prevention of recurrence, review of the past session theme, and setting of future goals	Fill in what you learned or felt during the vCBTp[a] session

[a]vCBTp: videoconference-based cognitive behavioral therapy for psychosis.

We also focused on intrusive thoughts and subjective distress caused by broadly defined traumatic stress experiences in patients with schizophrenia [44-46]. The Graduate School of Medicine, Chiba University, to which the researcher belongs, also has a track record of research on imagery rescripting of traumatic memories [47]. However, in this study, instead of

XSL•FO
RenderX

dealing with "traumatic memories" in consideration of the burden on participants, the aim was to help participants become aware of their core beliefs and work on themselves by reflecting on stressful events related to their past onset of psychosis. The fact that the research sessions would include an important session in which participants would reflect on painful events in their past was fully explained and agreed upon during the study description. It was also decided that sessions would focus on episodes that participants felt comfortable talking about. The order of sessions was flexible, with sessions being changed or repeated depending on the participants' levels of understanding and their reactions at the time.

There was no staff involvement in this study with regard to UC, which was defined as participants' own use of outpatient care and rehabilitation that they were engaged in prior to participation in the study and until its conclusion.

While use anxiety and connection problems are possible with the use of digital devices, the study addressed these by providing the following support: vCBTp patients practiced connecting to MK for participation through videoconference from home prior to the 0-week baseline assessment. vCBTp patients were also provided with an iPad charger and a microphone with an earpiece for consideration of their family living together. Patients with no or unstable home Wi-Fi access were loaned mobile Wi-Fi at no charge. While participating in the study, they could be contacted at any time via email. Participants were able to contact the researcher via email if they caught a cold or had other urgent appointments. When connected, participants were also called as needed to provide support for iPad use and remote connection anxiety, as appropriate. In addition, after each session was completed for the vCBTp plus UC group, the materials from that session were sent in PDF format via email. After the research trial was completed, the PowerPoint presentation materials for the participants, which had been included during 7 sessions, were printed as a paper-based booklet for the participants' own use and mailed to them by the researcher.

Outcomes

Primary Outcome

The primary outcome was the PANSS total score. The change from baseline to week 8 was compared. The PANSS is an assessment consisting of 7 positive symptoms, 7 negative symptoms, and 16 general psychopathology items. Each item is scored on a scale of 1-7 points. The scores range from 30 (no symptoms) to 210 (the most severe symptoms). The PANSS is an objective instrument for assessing schizophrenia symptoms in clinical and experimental studies and is the global standard for reliability and validity. Assessors blinded to the treatment assignment assessed the PANSS via a videoconference system at baseline and 8 weeks after the test.

Secondary Outcomes

Subtotal scores were assessed for positive symptoms, negative symptoms, and general psychopathology extracted from the PANSS subscale scores. Total scores for the 7 positive symptom subscales of the PANSS ranged from 7 (no symptoms) to 49 (severe positive symptoms). Total scores for the 7 negative symptom subscales ranged from 7 (no symptoms) to 49 (severe negative symptoms). Total scores for the 16 items in the general psychopathology subscale ranged from 16 (no symptoms) to 112 (severe psychopathology).

The Japanese version of the Beck Cognitive Insight Scale (BCIS-J) [48] was also used to assess cognitive insight. All 15 items of the BCIS-J were divided into 9 items of self-reflection (0-36) and 6 items of confidence (0-24). A lower subtotal of the 9 self-reflection items minus the 6 confidence items is associated with a lower cognitive pathology score.

In addition, the Impact of Event Scale-Revised (IES-R) [49] was used to assess subjective distress over specific stressful life events. The IES-R consists of 22 items (0-88) that are divided into 8 intrusive symptoms, 8 avoidance symptoms, and 6 hyperarousal symptoms, measuring the symptoms of persons exposed to traumatic experiences.

The PHQ-9 was used to assess depressive symptoms. It consists of 9 items ranging from 0 (no depressive symptoms) to 27 (severe depressive symptoms).

The Generalized Anxiety Disorder-7 (GAD-7) [50] was used to assess anxiety. The GAD-7 consists of 7 items ranging from 0 (no anxiety symptoms) to 21 (severe anxiety symptoms).

The EQ-5D-5L [51] was used to assess quality of life (QoL). It consists of 5 items and assesses QoL on a 5-point Likert scale ranging from 1 (not severe) to 5 (severe). The EQ-5D-5L is the most commonly used measure of QoL worldwide.

Other Outcomes

Other assessments included chlorpromazine (CPZ) [52], which indicates (equivalent to) the daily prescribed dose of antipsychotic medication. In addition, the therapists inquired about participants' experiences of adverse events during each assessment. All measures were assessed at week 0 (baseline) and week 8 (after treatment), and the results were analyzed according to set principles.

Statistical Analyses

This study followed the Consolidated Standards of Reporting Trials (CONSORT) guidelines for statistical analyses and reporting. The primary analysis was based on the intention-to-treat principle. We calculated the PANSS total score change at week 8 for both the vCBTp and UC groups and compared the mean differences between the groups to assess the primary outcome. Comparisons between the 2 groups were performed using an unpaired *t* test. Secondary endpoints were analyzed using the same approach as the primary endpoint to understand the primary endpoint further. In addition, we considered participants in each group with a 20% improvement in the total PANSS score as treatment responders. We compared the proportion of participants in each group. The frequency of adverse events was used as a safety endpoint. For percentage estimates, the exact 2-sided 95% CIs of the binomial distribution were calculated for each group. Statistical analyses were performed using SAS version 9.4 (SAS Institute Inc). Statistical significance was set at a *P*-value of <.05. The Cohen *d* (M baseline–M posttest)/SD pooled formula was used to calculate

XSL•FO
RenderX

the effect size (Cohen *d*) from the difference in the means of each group.

Results

Recruitment

Figure 1 shows the flow diagram of patient recruitment based on the CONSORT guidelines. A total of 33 patients applied to participate. Applicants included those who applied through the website (n=5), those who saw the poster in the hospital outpatient waiting room and applied on their own (n=4), and those who were referred by their doctor or daycare facility staff (n=14). Three of these patients were excluded. Of the 3 excluded patients, 1 was excluded because of hospitalization. The other 2 patients were excluded because they had multiple zeros on the MacCAT-CR, did not meet the criteria for understanding the trial, and were found not to have adequate capacity to consent. These 2 individuals were referred by a daycare facility staff member, who did not have a good understanding of the clinical trial and was primarily interested in obtaining the burden-reduction funding. In addition, 6 patients voluntarily declined to participate in the study after being informed about the study, citing concerns about the burden of the study and fatigue. Ultimately, 24 patients underwent in-person eligibility examinations at Chiba University Hospital, and all 24 were enrolled in the study. Study participants had schizophrenia (n=18) and schizoaffective disorder (n=4) and were assigned in equal numbers to the 2 groups. All 24 participants in the assigned vCBTp plus UC and UC groups completed the clinical trial. They provided informal feedback on the vCBTp plus UC intervention. Nearly all the study participants who completed the study reported positive feedback, indicating that the intervention was appropriate and easy to engage in (Figure 1).

Figure 1. Flow diagram of patient recruitment. DSM-5: Diagnostic and Statistical Manual of Mental Disorders, 5th edition; MacCAT-CR: MacArthur Competence Assessment Tool for Clinical Research; UC: usual care; vCBTp: videoconference-based cognitive behavioral therapy for psychosis.

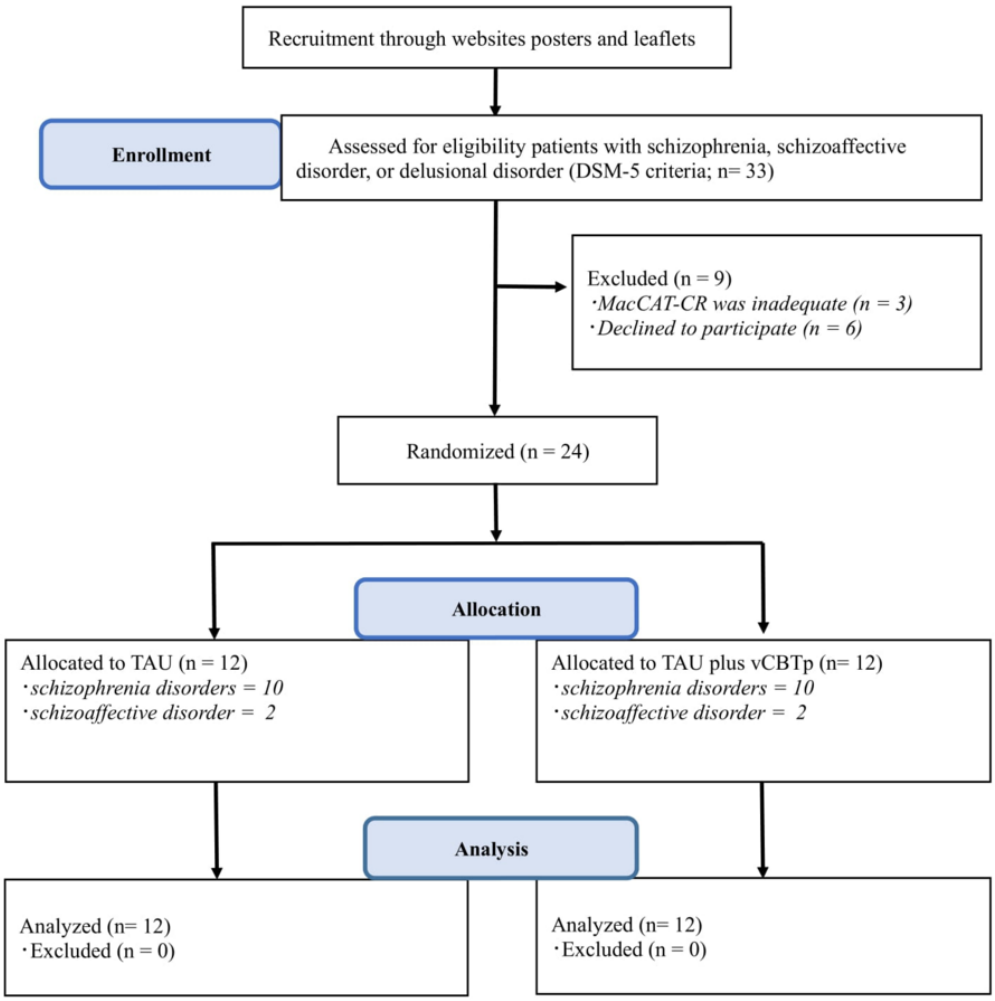

Participants' Characteristics

Table 2 summarizes the clinical characteristics of the participants at the time of assignment. There were no significant differences between the 2 groups in terms of age, gender, marital status, employment status, number of people living together, number of psychiatric hospitalizations, age of onset, duration of illness, total MacArthur comprehension score, and PANSS participant proportion score below 50 (Table 2).

XSL•FO
RenderX

Table 2. Baseline characteristics of the study participants (N=24).

Variable	vCBTp[a] + UC[b] group (n=12)	UC group (n=12)
Female, n (%)	7 (58)	7 (58)
Unmarried, n (%)	8 (89)	9 (75)
Unemployed, n (%)	10 (83)	9 (75)
Age (years), mean (SD)	34.7 (8.98)	32.3 (12.26)
Hospitalization (times), mean (SD)	1.6 (1.44)	1.3 (1.87)
Age of onset (years), mean (SD)	22.8 (6.70)	24.4 (8.26)
Duration of an illness (years), mean (SD)	11.6 (7.44)	7.6 (5.74)
MacCAT-CR[c] (total score), mean (SD)	41.1 (1.11)	40.3 (1.84)
Number of patients with a PANSS[d] score less than 51, n (%)	3 (25)	2 (17)

[a]vCBTp: videoconference-based cognitive behavioral therapy for psychosis.

[b]UC: usual care.

[c]MacCAT-CR: MacArthur Competence Assessment Tool for Clinical Research.

[d]PANSS: Positive and Negative Syndrome Scale.

Participants were mostly single (n=17) and unemployed (n=19). They were in their early 30s on average, and the average number of hospitalizations was less than two, although some had never been hospitalized (n=5). Some respondents were living alone in apartments (n=1) or in assisted living (n=1), but the rest were living with family members (n=22). An equal number of patients with schizophrenia (n=10) and schizoaffective disorder (n=2) participated in each group. These ambulatory variables were not used as randomization factors.

In addition, all participants had a certain affinity for digital devices, as they usually used smartphones and PCs to watch videos. Both groups were able to maintain a remote connection environment without major problems by exchanging emails and receiving advice over the phone in case of connection problems.

Primary Outcome

Table 3 displays the means and SDs of the participants' PANSS total scores from baseline to posttest. The mean change in PANSS total scores from baseline to posttest was –9.5 (95% CI –12.09 to –6.91) for the vCBTp plus UC group and 6.9 (95% CI 1.54-12.30) for the UC group (Table 4). The vCBTp plus UC group demonstrated a significant improvement compared to the UC group, with a between-group difference of 16.41 (95% CI 10.79-22.04; $P<.001$).

XSL•FO
RenderX

Table 3. Raw data of the primary and secondary outcomes (N=24).

Variable	vCBTp[a] + UC[b] group (n=12)			UC group (n=12)		
	Pre score, mean (SD)	Post score, mean (SD)	Cohen *d*	Pre score, mean (SD)	Post score, mean (SD)	Cohen *d*
Primary outcome						
PANSS[c] (total score)	52.3 (7.02)	42.8 (6.02)	1.45	54.2 (11.07)	61.1 (10.72)	–0.61
Secondary outcomes						
PANSS (positive)	14.8 (2.77)	11.4 (1.88)	1.43	13.7 (3.77)	14.6 (3.50)	–0.24
Delusion	3.4 (1.31)	2.7 (0.78)	0.65	3.5 (1.45)	3.6 (1.31)	–0.07
Conceptual disorganization	1.8 (0.75)	1.4 (0.51)	0.62	1.6 (0.90)	1.8 (0.83)	–0.23
Hallucinatory behavior	1.8 (1.11)	1.6 (0.67)	0.22	2.3 (1.30)	2.8 (1.42)	–0.37
Excitement	1.8 (1.29)	1.3 (0.45)	0.52	1.2 (0.39)	1.2 (0.39)	0.00
Grandiosity	1.8 (1.54)	1.3 (0.45)	0.35	1.5 (1.00)	1.4 (0.79)	0.11
Suspiciousness	3.2 (1.53)	2.3 (0.97)	0.70	2.6 (1.16)	2.8 (0.87)	–0.20
Hostility	1.1 (0.29)	1.0 (0.00)	0.14	1.0 (0.00)	1.1 (0.29)	–0.13
PANSS (negative)	10.8 (2.63)	9.6 (1.78)	0.53	13.7 (6.87)	16.3 (5.03)	–0.43
PANSS (general psychopathology)	26.8 (4.18)	21.8 (2.96)	1.38	26.8 (4.57)	30.2 (5.65)	–0.66
Anxiety	2.8 (1.11)	2.0 (0.60)	0.90	2.8 (1.19)	2.8 (1.19)	0.00
Guilt feeling	2.9 (1.38)	2.1 (0.90)	0.69	2.1 (0.67)	2.3 (0.75)	–0.28
Depression	2.1 (1.00)	1.6 (0.67)	0.59	2.0 (1.04)	2.8 (1.40)	–0.42
Lack of judgment and insight	1.9 (0.90)	1.4 (0.51)	0.68	2.1 (0.79)	2.3 (0.87)	–0.24
BCIS-J[d] (self-certainty)	4.7 (2.27)	4.2 (2.41)	0.21	7.0 (3.28)	7.9 (5.05)	–0.21
BCIS-J (self-reflectiveness)	14.5 (3.12)	16.7 (3.85)	0.63	11.6 (1.19)	12.8 (4.67)	0.35
IES-R[e] (total score)	37.1 (13.43)	31.8 (14.68)	0.36	30.3 (16.55)	32.1 (16.08)	–0.11
Intrusion	13.3 (5.38)	10.0 (4.88)	0.64	11.8 (9.36)	11.3 (7.81)	0.06
Avoidance	15.3 (4.59)	14.2 (6.28)	0.20	12.2 (7.90)	12.8 (5.72)	–0.09
Hyperarousal	8.6 (5.92)	7.6 (5.30)	0.18	6.3 (4.89)	8.0 (5.29)	–0.33
PHQ-9[f]	10.7 (6.05)	8.0 (0.18)	0.47	8.6 (4.72)	9.4 (4.44)	–0.17
GAD-7[g]	7.7 (3.58)	5.7 (4.21)	0.51	5.5 (3.18)	7.2 (4.45)	–0.44
EQ-5D-5L (index value)	0.7 (0.19)	0.8 (0.18)	0.54	0.8 (0.13)	0.7 (0.08)	–0.93
Mobility	1.7 (1.07)	1.3 (0.63)	0.45	1.3 (0.45)	1.1 (0.29)	0.53
Self-care	1.4 (0.67)	1.3 (0.65)	0.15	1.1 (0.29)	1.3 (0.45)	–0.53
Usual activities	2.3 (1.22)	1.9 (1.08)	0.35	1.8 (0.87)	2.0 (1.04)	–0.21
Pain/discomfort	2.2 (0.94)	1.8 (1.22)	0.36	1.4 (0.67)	2.3 (1.06)	–0.73
Anxiety/depression	2.7 (1.15)	2.3 (1.22)	0.34	1.8 (0.72)	2.3 (0.65)	–0.73
CPZ[h] equivalent (mg/day)	488.5 (340.91)	462.3 (336.90)	0.08	413.9 (252.29)	386.8 (229.34)	0.11

[a]vCBTp: videoconference-based cognitive behavioral therapy for psychosis.
[b]UC: usual care.
[c]PANSS: Positive and Negative Syndrome Scale.
[d]BCI-J: Japanese version of the Beck Cognitive Insight Scale
[e]IES-R: Impact of Event Scale-Revised.
[f]PHQ-9: Patient Health Questionnaire-9.
[g]GAD-7: Generalized Anxiety Disorder-7.
[h]CPZ: chlorpromazine.

XSL•FO
RenderX

Table 4. Amount of change from baseline to posttest (N=24).

Variable	vCBTp[a] + UC[b] group (n=12), mean (SD)	UC group (n=12), mean (SD)	*t* test	*P* value
Primary outcome				
PANSS[c] (total score)	–9.5 (4.08)	6.9 (8.47)	6.05	<.001
Secondary outcomes				
PANSS (positive)	–3.3 (2.99)	0.9 (1.98)	4.10	<.001
Delusion	–0.8 (0.75)	0.1 (1.00)	2.31	.03
Conceptual disorganization	–0.3 (0.49)	0.3 (0.62)	0.56	.02
Hallucinatory behavior	–0.3 (0.75)	0.4 (0.79)	2.11	.046
Excitement	–0.5 (1.17)	0.0 (0.00)	1.48	.15
Grandiosity	–0.5 (1.17)	–0.1 (0.51)	1.13	.27
Suspiciousness	–0.9 (1.24)	0.2 (1.11)	2.25	.04
Hostility	–0.1 (0.29)	0.1 (0.29)	1.41	.17
PANSS (negative)	–1.2 (1.40)	2.7 (3.89)	3.21	.004
PANSS (general psychopathology)	–5.0 (2.80)	3.3 (4.79)	5.20	<.001
Anxiety	–0.8 (0.94)	0.0 (1.41)	1.70	.10
Guilt feeling	–0.8 (1.03)	0.2 (0.83)	2.61	.02
Depression	–0.5 (0.67)	0.8 (1.47)	2.86	.009
Lack of judgment and insight	–0.5 (0.52)	0.2 (0.58)	2.97	.007
BCIS-J[d] (self-certainty)	–0.5 (1.83)	0.9 (3.37)	1.28	.21
BCIS-J (self-reflectiveness)	2.2 (4.63)	1.3 (3.33)	–0.56	.58
IES-R[e] (total score)	–5.3 (14.87)	1.8 (16.65)	1.10	.28
Intrusion	–3.3 (5.55)	–0.5 (6.16)	1.15	.26
Avoidance	–1.1 (7.03)	0.6 (9.00)	0.51	.62
Hyperarousal	–1.0 (4.57)	1.7 (5.00)	1.36	.19
PHQ-9[f]	–2.7 (5.33)	0.8 (4.22)	1.78	.09
GAD-7[g]	–2.0 (4.71)	1.7 (3.52)	2.16	.04
EQ-5D-5L (index value)	0.1 (0.15)	–0.1 (0.13)	–3.14	.005
Mobility	–0.3 (0.89)	–0.2 (0.39)	0.60	.56
Self-care	–0.1 (0.79)	0.2 (0.58)	0.88	.39
Usual activities	–0.3 (0.89)	0.3 (0.87)	1.63	.12
Pain/discomfort	–0.4 (1.24)	0.8 (1.03)	2.69	.01
Anxiety/depression	–0.4 (0.79)	0.5 (0.67)	3.05	.006
CPZ[h] equivalent (mg/day)	6.8 (22.61)	–27.1 (51.63)	–2.01	.006

[a]vCBTp: videoconference-based cognitive behavioral therapy for psychosis.
[b]UC: usual care.
[c]PANSS: Positive and Negative Syndrome Scale.
[d]BCI-J: Japanese version of the Beck Cognitive Insight Scale.
[e]IES-R: Impact of Event Scale-Revised.
[f]PHQ-9: Patient Health Questionnaire-9.
[g]GAD-7: Generalized Anxiety Disorder-7.
[h]CPZ: chlorpromazine.

XSL•FO
RenderX

Secondary Outcomes

The mean change at posttest for the 3 PANSS subscales was as follows. The mean change in positive symptoms from baseline was –3.33 (95% CI –5.24 to –1.43) in the vCBTp plus UC group and 0.92 (95% CI –0.34 to 2.17) in the UC group. The difference between groups was significant at 4.25 (95% CI 2.10-6.40; *P*<.001). The mean change in negative symptoms was 2.67 (95% CI 0.19-5.14) in the UC group and –1.17 (95% CI –2.06 to –0.28) in the vCBTp plus UC group. The difference between groups was 3.83 (95% CI 1.36-6.31; *P*=.004). The mean change in overall psychopathology was 3.33 (95% CI 0.29-6.38) in the UC group and –5.00 (95% CI –6.78 to –3.22) in the vCBTp plus UC group. The difference between groups was 8.33 (95% CI 5.01-11.66; *P*<.001), with all 3 subscales showing significant differences.

The means and SDs of the secondary endpoints other than the PANSS and the mean changes are presented in Tables 3 and 4, respectively. There were no statistically significant differences between the 2 groups, including the subitems for BCIS-J, IES-R, PHQ-9, and CPZ. However, the vCBTp plus UC group had improved mean values for all change measures.

Other Outcomes

When analyzing participants who showed a 20% or greater improvement in the PANSS total score as treatment responders, we observed that 4 participants (33%) in the vCBTp plus UC group responded positively to the treatment. In comparison, none (0%) in the UC group exhibited a 20% or greater improvement. When comparing the effect size of the mean change between the 2 groups in Table 3, the PANSS total score for the vCBTp plus UC group had Cohen *d*=1.45 (95% CI 3.96-15.04). Furthermore, the subscale effect sizes for the vCBTp plus UC group were as follows: PANSS positive symptoms, Cohen *d*=1.44 (95% CI 1.38-5.42); PANSS negative symptoms, Cohen *d*=0.53 (95% CI –0.72 to 3.12); and PANSS overall psychopathology, Cohen *d*=1.38 (95% CI 1.91-8.09). Notably, a significant effect was observed outside of negative symptoms.

At 0 weeks, the scores on the PHQ-9 [36] screening for suicidal ideation (item 9) were 3 points in 0 participants, 2 in 2 participants, 1 in 5 participants, and 0 in 17 participants. At 8 weeks, the scores were 3 points in 0 participants, 2 in 1 participant, 1 in 7 participants, and 0 in 16 participants. The patients were not excluded from the study because they were not at risk for imminent suicide as assessed by the researcher's interviews and psychiatrist's ratings.

During the session on trauma in the broadest sense, several participants began to cry as they recalled their painful past. However, none declined to participate in the study because of this.

As for safety evaluation, no serious adverse events occurred among the participants, including hospitalization. There were 2 adverse events in the UC group and 4 in the vCBTp plus UC group, none of which were attributable to vCBTp.

The total PANSS scores in the UC group worsened. The self-administered questionnaire, a secondary assessment, also showed worsening scores. The 8-week assessment of the UC group revealed changes in the participants' life circumstances, including plans to retire due to poor physical and mental health, impending separation from a partner, difficulties in making friends at school, progressive illness in the family, and impending return to school after a leave of absence. The participants talked about the environmental changes in their lives.

Discussion

Principal Findings

This is the first pilot RCT to examine the efficacy of providing real-time and one-on-one vCBTp for patients with schizophrenia over 7 sessions. The study revealed that vCBTp was effective as an adjunct to UC with regard to the PANSS total score and each subscale at 8 weeks after the intervention.

Comparison With Prior Work

In a study by Morrison et al [38], the PANSS total score improved in the group provided with pharmacotherapy and CBTp (compared with baseline) at 6 weeks, with an effect size of 0.66 (mean change of 6.06 points), and at 12 weeks, with an effect size of 1.04 (mean change of 12.36 points). In addition, although treatment guidelines recommend 16 or more sessions, a meta-analysis by Hazell et al [11] revealed that CBTp performed with 6 to 15 sessions still yielded a notable effect size of 0.46, indicating that a smaller number of sessions than the recommended 16 can be effective. Furthermore, a previous study of an intervention for self-esteem reported an effect size as large as 1.2 (n=18) for 7 sessions in an unformulated configuration [53]. This suggests that even fewer sessions may produce effects by clarifying target symptoms.

However, our study requires careful discussion because the significant difference in the PANSS score observed in our study involved not only an improvement in the vCBTp plus UC group but also a worsening in the UC group. Therefore, the effect size was not a comparison of the change in the 2 groups, but rather the change between week 0 and week 8 in the vCBTp plus UC group. The vCBTp plus UC group showed a significant effect size of 1.45 (mean change 9.5) in the total PANSS score at 8 weeks, exceeding that observed in previous studies. This exceeded the results observed in a previous study.

However, the scores themselves are not remarkable improvements, since the study by Morrison et al [38] showed a 12.36-point improvement in the PANSS total score at 12 weeks, whereas this study showed a 9.5-point improvement. The smaller SD of the vCBTp plus UC group in this study may have been one of the reasons for the large effect size.

It has been reported that the addition of CBT to pharmacotherapy resulted in an improvement of more than 20% in overall symptoms in 44.5% of participants [54]. In this study, 33% of participants in the vCBTp plus UC group had an improvement of 20% or more, resulting in a generally similar rate of improvement, although not as high as in previous studies.

The reasons for the improvement are discussed below. In the study by Naeem et al [39], which we referred to in setting our

XSL•FO
RenderX

allocation adjustment factors, the PANSS score was 51.6 at baseline in the control group and 49.05 in the intervention group, which indicates mild symptoms. Again, the mean baseline scores for the vCBTp plus UC and UC groups were 54.2 and 52.3, respectively, indicating that participants had relatively mild to moderate symptom severity. Patients suitable for CBTp had mild to moderate psychiatric symptoms, and had a high level of insight [55], motivation to participate, cognitive insight [55], willingness to participate, and cognitive flexibility [56]. The following factors may have contributed to the improvement in the vCBTp plus UC group in this study. Participants with the ability to consent appropriately were included through the MacCAT-CR. This suggested that participants retained a certain level of comprehension and cognitive ability. Cognitive functioning is said to facilitate adaptation to CBTp [57]. It has also been reported that insights into the disease and mild symptoms predict good outcomes even after short-term CBTp sessions [58]. The participants in this study had mild symptoms and clear research consent, which may have contributed to the success of vCBTp.

Some participants had self-odor delusions, some were restricted from using transportation due to stress diarrhea symptoms, and some were not able to go outside due to a sense of forethought or paranoia. In addition, some patients struggled with the burden of transportation costs for travel. The ability to receive vCBTp at home seemed to reduce the labor burden associated with transportation.

After each session, a PDF summary of the session was sent to each participant, and the sessions were individualized to the case. The subjective assistance needs of patients with schizophrenia include the desire to be listened to by a professional [59]. There was a lot of positive feedback about the experience of being listened to in a one-on-one setting. Sessions included positive rapport building through listening and individual case formulation through collaborative demonstration work to develop strategies for coping with stress. The participants seemed to find it refreshing to discuss coping strategies with a specialist. After forming rapport in the first half of the session, the participants shared their painful experiences with the therapist in the second half of the session. These experiences included intrusive memories about stressful events related to the onset of psychosis, while taking into consideration their own distress. The participants then shared their experiences of working on themselves, which they said helped them to accept themselves as they are now. There were also reports of the effects of working on oneself [60]. For most of the participants, this was the first experience of talking to a therapist about a painful event in their past.

Traumatic stress symptoms have been reported in approximately 12.4% of patients with schizophrenia [61], suggesting an integrated model of trauma and schizophrenia [62]. Intervention attempts have also been reported [63]. In this study, the 5th and 6th sessions dealt with intrusive memories of stressful events related to the onset of psychosis, and were designed to help participants understand core beliefs, work on themselves, and reflect on coping strategies that have helped them overcome their painful past. Although there were no significant differences between groups on the IES-R, the mean change scores for the vCBTp + UC group, including the subscales, showed a trend toward improvement.

Significant improvements were observed in the PANSS subscale scores for positive symptoms ($P<.001$), negative symptoms ($P=.009$), and general psychopathology ($P<.001$). Improvements were also observed in the self-administered GAD-7 and EQ-5D-5L and the EQ-5D-5L subcomponents of pain/discomfort and anxiety/depression. In recent years, nearly half of patients with schizophrenia have comorbid anxiety disorders [64]. In this study, it is possible that vCBTp was effective for those with subjective anxiety and discomfort.

We next discuss exacerbations in the UC group. Schizophrenia has been noted for its disease characteristics, including symptom instability and agitation due to stress in daily life [65,66], and several patients in the UC group reported anger, hopelessness, depression, and anxiety at assessment due to daily work, friends and family, and other life circumstances during the study period. There have been reports of a lack of coping skills [67] and high sensitivity to stress and symptoms [68] with regard to daily life stress in schizophrenia. In the UC group, stress related to coping with interpersonal relationships and worries in daily life may have also affected symptoms and contributed to worsening outcomes [69]. Schizophrenia carries a high risk of relapse, with 30% of patients being readmitted within 6 months to 1 year [70]. This may indicate that patients with schizophrenia have difficulty with coping skills and stress management [71]. In this study, most of the secondary endpoints of self-administered questionnaires also showed worsening values. The worsening of anxiety/pain on the GAD-7 and EQ-5D-5L was particularly noticeable, suggesting that there are anxiety-based effects in daily life events.

We initially estimated a dropout rate of 10%. This was considered an appropriate estimate considering that a previous study reported an average dropout rate of 13% [72] and another study reported a dropout rate of 11% [40]. However, there were no dropouts in this study. There are several possible reasons for this. First, this study consisted of fewer sessions than the formulated number of CBTp sessions. Some meta-analyses have shown that the dropout rate from research trials among patients with schizophrenia increases as the number of intervention sessions increases [73]. In addition, the implementation of MacCAT-CR with the inclusion criteria and the inclusion of participants who had firm research consent and understanding may have been some of the reasons for the lack of dropouts. Furthermore, many participants were motivated to participate in the study because they commented that they wanted to see more treatment options for schizophrenia and they wanted to help others who were suffering like they were. It has been reported that careful explanation of the methods and transparency of the study can increase motivation to participate and decrease the dropout rate [74]. As the participants had never received CBT before, there may have been some novelty to the study. In any case, it was inferred that they were highly motivated to participate. These factors could be the reasons for the absence of dropouts.

PHQ-9 item 9 was screened and used as a safety measure. Attention to suicidal ideation is important when the subject has

XSL•FO
RenderX

schizophrenia, as a study [75] validating the risk of suicidal behavior in psychosis with responses to item 9 of the PHQ-9 stated that the risk is higher for psychosis than depression. In this study, the number of participants who scored more than 1 point was also higher. Some participants in this study also scored 1 or more, but when interviewed afterward, they said that they scored based on their mood of anxiety and discouragement, saying that they feel sad or painful enough to want to die, and no one showed real suicidal thoughts. None of the adverse events were attributable to the vCBTp intervention (eg, cold and reflux esophagitis), and there were no serious adverse events, confirming a certain level of safety.

The use of videoconference systems can reduce the burden on the patient's hospital visit effort, that is, the financial burden of transportation costs, weather effects, and social fears associated with transportation use. It can also reduce the burden on CBTp providers. As a result, accessibility to CBTp could be improved.

Regarding remote interventions, it has been reported that simply providing devices is not sufficient to address digital exclusion, and training in using the devices is likely to be required [76]. However, with a simple structure, pretraining, and a support system, such as in this study, which can allow patients to open an email and tap a URL to participate in a videoconference, it may be possible to implement psychosocial interventions remotely using digital technology for patients with schizophrenia.

The major originality of this study is that participants could receive CBTp in the comfort of their own homes. Currently, self-help approaches using digital media, such as the internet and cell phones, are being actively developed [77,78]. As a result, many patients are interested in using digital technology to improve their mental health [25]. Low-intensity interventions may benefit many patients with schizophrenia without increasing costs [79]. Furthermore, studies have demonstrated the effectiveness of targeted CBTp with reduced session frequency [11,12,14]. Furthermore, an online intervention via videoconferencing has been reported to be feasible for patients with schizophrenia [80]. These considerations suggest that the provision of vCBTp using a remote format with reduced session frequency may be feasible enough for implementation in practice, especially in terms of a stepped care model to deliver CBTp [8,81].

Limitations

This exploratory study had some limitations. First, the UC settings varied, and these were not subject to comparison or analysis between the 2 groups. The frequency of outpatient visits varied from once a week to once every 2 months. The frequency and duration of use of psychiatric rehabilitation facilities also varied. These may have influenced outcomes as confounding variables. Prescription medications and methods of administration also varied. The study did not analyze the influence of external variables such as the participants' living environments; thus, future research on correlations with external variables would be desirable.

Second, it was not clear what targets the vCBTp program influenced. Although various points were considered, the primary mechanism of change was unclear because the study incorporated a variety of techniques.

Third, due to the absence of a follow-up assessment, the determination of efficacy remained unclear. Therefore, a 1-year follow-up is being conducted for vCBTp participants as an additional study.

Fourth, the EQ-5D-5L tool was used as a subjective measure of quality of life but was not validated for quality-adjusted useful life or cost-effectiveness. Future studies should examine the contribution of the vCBTp program to health care costs.

Fifth, the sample size in our pilot study was relatively small. Since improvements were observed in this study, a larger multi-center study is desired.

Finally, since this trial was conducted with only 1 therapist and 1 assessor, we are yet to examine how to share the experience of this vCBTp intervention with staff at other facilities. It is important to consider the possibility of social implementation in the future.

Conclusions

The findings of this RCT indicate that vCBTp is effective in treating symptoms over 7 sessions. The findings show that vCBTp can be effective at a distance and with a reduced number of sessions, which can reduce the burden on the practitioner and improve the accessibility of CBTp for patients. It may also contribute to relapse prevention and psychosocial support during stepped care. Further investigations are required to determine the durability of this effect.

Acknowledgments

We are grateful to all participants and all those involved in this study. We thank all the staff and people involved at the Center for Cognitive Behavioral Therapy, Chiba University Hospital. This work was supported by a Grant-in-Aid for Scientific Research (grant number 21K11197; Japan Society for the Promotion of Science).

Data Availability

The data are available upon reasonable request.

XSL•FO
RenderX

Authors' Contributions

MK and ES designed the study and managed the research methods and project. HH collected and organized the data. YS performed the formal analyses. MK, ES, and HN recruited patients. MK and ES created the 7-session videoconference-based cognitive behavioral therapy for psychosis component. MK performed all sessions of the intervention. HN performed all Positive and Negative Syndrome Scale assessments (0 and 8 weeks) in this study and participated in patient enrollment. The manuscript was written by MK and ES. Study funding was secured by MK through a grant. MK wrote the manuscript. All authors, including ES, contributed to and approved the final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
CONSORT eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 869 KB - formative_v9i1e59540_app1.pdf]

References

1. Mueser KT, McGurk SR. Schizophrenia. Lancet 2004 Jun 19;363(9426):2063-2072. [doi: 10.1016/S0140-6736(04)16458-1] [Medline: 15207959]
2. Saha S, Chant D, Welham J, McGrath J. A systematic review of the prevalence of schizophrenia. PLoS Med 2005 May 31;2(5):e141 [FREE Full text] [doi: 10.1371/journal.pmed.0020141] [Medline: 15916472]
3. Psychosis and schizophrenia in adults: prevention and management. London: National Institute for Health and Care Excellence (NICE); 2014.
4. Mankiewicz PD, Turner C. Cognitive restructuring and graded behavioural exposure for delusional appraisals of auditory hallucinations and comorbid anxiety in paranoid schizophrenia. Case Rep Psychiatry 2014;2014:124564-124568 [FREE Full text] [doi: 10.1155/2014/124564] [Medline: 25302129]
5. Jauhar S, McKenna PJ, Radua J, Fung E, Salvador R, Laws KR. Cognitive-behavioural therapy for the symptoms of schizophrenia: systematic review and meta-analysis with examination of potential bias. Br J Psychiatry 2014 Jan 02;204(1):20-29. [doi: 10.1192/bjp.bp.112.116285] [Medline: 24385461]
6. Wykes T, Steel C, Everitt B, Tarrier N. Cognitive behavior therapy for schizophrenia: effect sizes, clinical models, and methodological rigor. Schizophr Bull 2008 May 04;34(3):523-537 [FREE Full text] [doi: 10.1093/schbul/sbm114] [Medline: 17962231]
7. Farhall J, Freeman NC, Shawyer F, Trauer T. An effectiveness trial of cognitive behaviour therapy in a representative sample of outpatients with psychosis. Br J Clin Psychol 2009 Mar 31;48(Pt 1):47-62. [doi: 10.1111/j.2044-8260.2009.tb00456.x] [Medline: 18851771]
8. Kopelovich SL, Nutting E, Blank J, Buckland HT, Spigner C. Preliminary point prevalence of Cognitive Behavioral Therapy for psychosis (CBTp) training in the U.S. and Canada. Psychosis 2022 Jan 14;14(4):344-354. [doi: 10.1080/17522439.2021.1971744]
9. The abandoned illness: a report by the Schizophrenia Commission. Rethink. URL: https://www.rethink.org/media/2637/the-abandoned-illness-final.pdf [accessed 2024-12-11]
10. Shafran R, Clark D, Fairburn C, Arntz A, Barlow D, Ehlers A, et al. Mind the gap: Improving the dissemination of CBT. Behav Res Ther 2009 Nov;47(11):902-909. [doi: 10.1016/j.brat.2009.07.003] [Medline: 19664756]
11. Hazell CM, Hayward M, Cavanagh K, Strauss C. A systematic review and meta-analysis of low intensity CBT for psychosis. Clin Psychol Rev 2016 Apr;45:183-192 [FREE Full text] [doi: 10.1016/j.cpr.2016.03.004] [Medline: 27048980]
12. Malik N, Kingdon D, Pelton J, Mehta R, Turkington D. Effectiveness of brief cognitive-behavioral therapy for schizophrenia delivered by mental health nurses: relapse and recovery at 24 months. J Clin Psychiatry 2009 Feb 10;70(2):201-207. [doi: 10.4088/jcp.07m03990] [Medline: 19210949]
13. Brabban A, Tai S, Turkington D. Predictors of outcome in brief cognitive behavior therapy for schizophrenia. Schizophr Bull 2009 Sep 01;35(5):859-864 [FREE Full text] [doi: 10.1093/schbul/sbp065] [Medline: 19571248]
14. Turkington D, Kingdon D, Turner T, Insight into Schizophrenia Research Group. Effectiveness of a brief cognitive-behavioural therapy intervention in the treatment of schizophrenia. Br J Psychiatry 2002 Jun 2;180(06):523-527. [doi: 10.1192/bjp.180.6.523] [Medline: 12042231]
15. Tai S, Turkington D. The evolution of cognitive behavior therapy for schizophrenia: current practice and recent developments. Schizophr Bull 2009 Sep;35(5):865-873 [FREE Full text] [doi: 10.1093/schbul/sbp080] [Medline: 19661198]
16. Keen N, Hunter ECM, Peters E. Integrated Trauma-Focused Cognitive-Behavioural Therapy for Post-traumatic Stress and Psychotic Symptoms: A Case-Series Study Using Imaginal Reprocessing Strategies. Front Psychiatry 2017 Jun 01;8:92 [FREE Full text] [doi: 10.3389/fpsyt.2017.00092] [Medline: 28620323]

XSL•FO
RenderX

17. Gairns S, Alvarez-Jimenez M, Hulbert C, McGorry P, Bendall S. Perceptions of clinicians treating young people with first-episode psychosis for post-traumatic stress disorder. Early Interv Psychiatry 2015 Feb 27;9(1):12-20. [doi: 10.1111/eip.12065] [Medline: 23802596]
18. Sin J, Spain D, Furuta M, Murrells T, Norman I. Psychological interventions for post-traumatic stress disorder (PTSD) in people with severe mental illness. Cochrane Database Syst Rev 2017 Jan 24;1(1):CD011464. [doi: 10.1002/14651858.CD011464.pub2] [Medline: 28116752]
19. Sin J, Spain D. Psychological interventions for trauma in individuals who have psychosis: A systematic review and meta-analysis. Psychosis 2016 May 23;9(1):67-81. [doi: 10.1080/17522439.2016.1167946]
20. Swan S, Keen N, Reynolds N, Onwumere J. Psychological interventions for post-traumatic stress symptoms in psychosis: a systematic review of outcomes. Front Psychol 2017 Mar 14;8:341 [FREE Full text] [doi: 10.3389/fpsyg.2017.00341] [Medline: 28352239]
21. Gottlieb JD, Romeo KH, Penn DL, Mueser KT, Chiko BP. Web-based cognitive-behavioral therapy for auditory hallucinations in persons with psychosis: a pilot study. Schizophr Res 2013 Apr;145(1-3):82-87. [doi: 10.1016/j.schres.2013.01.002] [Medline: 23410709]
22. Gottlieb JD, Gidugu V, Maru M, Tepper MC, Davis MJ, Greenwold J, et al. Randomized controlled trial of an internet cognitive behavioral skills-based program for auditory hallucinations in persons with psychosis. Psychiatr Rehabil J 2017 Sep;40(3):283-292. [doi: 10.1037/prj0000258] [Medline: 28517948]
23. Rüegg N, Moritz S, Berger T, Lüdtke T, Westermann S. An internet-based intervention for people with psychosis (EviBaS): study protocol for a randomized controlled trial. BMC Psychiatry 2018 Apr 13;18(1):102 [FREE Full text] [doi: 10.1186/s12888-018-1644-8] [Medline: 29653532]
24. Westermann S, Rüegg N, Lüdtke T, Moritz S, Berger T. Internet-based self-help for psychosis: Findings from a randomized controlled trial. J Consult Clin Psychol 2020 Oct;88(10):937-950. [doi: 10.1037/ccp0000602] [Medline: 32790453]
25. Arnold C, Farhall J, Villagonzalo K, Sharma K, Thomas N. Engagement with online psychosocial interventions for psychosis: A review and synthesis of relevant factors. Internet Interv 2021 Sep;25:100411 [FREE Full text] [doi: 10.1016/j.invent.2021.100411] [Medline: 34401370]
26. Matsumoto K, Sutoh C, Asano K, Seki Y, Urao Y, Yokoo M, et al. Internet-based cognitive behavioral therapy with real-time therapist support via videoconference for patients with obsessive-compulsive disorder, panic disorder, and social anxiety disorder: pilot single-arm trial. J Med Internet Res 2018 Dec 17;20(12):e12091-e12890 [FREE Full text] [doi: 10.2196/12091] [Medline: 30559094]
27. Horita H, Seki Y, Yamaguchi T, Shiko Y, Kawasaki Y, Shimizu E. Videoconference-delivered cognitive behavioral therapy for parents of adolescents with internet addiction: pilot randomized controlled trial. JMIR Pediatr Parent 2024 Oct 03;7:e60604 [FREE Full text] [doi: 10.2196/60604] [Medline: 39361415]
28. Carlbring P, Andersson G, Cuijpers P, Riper H, Hedman-Lagerlöf E. Internet-based vs. face-to-face cognitive behavior therapy for psychiatric and somatic disorders: an updated systematic review and meta-analysis. Cogn Behav Ther 2018 Jan;47(1):1-18. [doi: 10.1080/16506073.2017.1401115] [Medline: 29215315]
29. Lecomte T, Abdel-Baki A, Francoeur A, Cloutier B, Leboeuf A, Abadie P, et al. Group therapy via videoconferencing for individuals with early psychosis: A pilot study. Early Interv Psychiatry 2021 Dec 23;15(6):1595-1601. [doi: 10.1111/eip.13099] [Medline: 33354926]
30. Wood HJ, Gannon JM, Chengappa KR, Sarpal DK. Group teletherapy for first-episode psychosis: Piloting its integration with coordinated specialty care during the COVID-19 pandemic. Psychol Psychother 2021 Jun 20;94(2):382-389. [doi: 10.1111/papt.12310] [Medline: 33078496]
31. Mendelson D, Thibaudeau É, Sauvé G, Lavigne KM, Bowie CR, Menon M, et al. Remote group therapies for cognitive health in schizophrenia-spectrum disorders: Feasible, acceptable, engaging. Schizophr Res Cogn 2022 Jun;28:100230 [FREE Full text] [doi: 10.1016/j.scog.2021.100230] [Medline: 35242604]
32. Katsushima M, Nakamura H, Hanaoka H, Shiko Y, Komatsu H, Shimizu E. Randomised controlled trial on the effect of video-conference cognitive behavioural therapy for patients with schizophrenia: a study protocol. BMJ Open 2023 Sep 11;13(9):e069734 [FREE Full text] [doi: 10.1136/bmjopen-2022-069734] [Medline: 37696635]
33. Fovet T, Amad A, Thomas P, Jardri R. [Schizophrenia and informed consent to research]. Encephale 2015 Oct;41(5):394-402. [doi: 10.1016/j.encep.2014.07.004] [Medline: 25127894]
34. Hostiuc S, Rusu MC, Negoi I, Drima E. Testing decision-making competency of schizophrenia participants in clinical trials. A meta-analysis and meta-regression. BMC Psychiatry 2018 Jan 05;18(1):2 [FREE Full text] [doi: 10.1186/s12888-017-1580-z] [Medline: 29304845]
35. Kay SR, Fiszbein A, Opler LA. The positive and negative syndrome scale (PANSS) for schizophrenia. Schizophr Bull 1987;13(2):261-276. [doi: 10.1093/schbul/13.2.261] [Medline: 3616518]
36. Muramatsu K, Miyaoka H, Kamijima K, Muramatsu Y, Yoshida M, Otsubo T, et al. The patient health questionnaire, Japanese version: validity according to the mini-international neuropsychiatric interview-plus. Psychol Rep 2007 Dec 01;101(3 Pt 1):952-960. [doi: 10.2466/pr0.101.3.952-960] [Medline: 18232454]

XSL•FO
RenderX

37. Na PJ, Yaramala SR, Kim JA, Kim H, Goes FS, Zandi PP, et al. The PHQ-9 Item 9 based screening for suicide risk: a validation study of the Patient Health Questionnaire (PHQ)-9 Item 9 with the Columbia Suicide Severity Rating Scale (C-SSRS). J Affect Disord 2018 May;232:34-40. [doi: 10.1016/j.jad.2018.02.045] [Medline: 29477096]
38. Morrison AP, Law H, Carter L, Sellers R, Emsley R, Pyle M, et al. Antipsychotic drugs versus cognitive behavioural therapy versus a combination of both in people with psychosis: a randomised controlled pilot and feasibility study. The Lancet Psychiatry 2018 May;5(5):411-423. [doi: 10.1016/s2215-0366(18)30096-8]
39. Naeem F, Johal R, McKenna C, Rathod S, Ayub M, Lecomte T, et al. Cognitive Behavior Therapy for psychosis based Guided Self-help (CBTp-GSH) delivered by frontline mental health professionals: Results of a feasibility study. Schizophr Res 2016 May;173(1-2):69-74. [doi: 10.1016/j.schres.2016.03.003] [Medline: 26971071]
40. Katsushima M, Shimizu E. Brief cognitive behavioral therapy for depression and anxiety in patients with schizophrenia in psychiatric home nursing service: pilot randomized controlled trial. Behav Sci (Basel) 2024 Aug 05;14(8):680 [FREE Full text] [doi: 10.3390/bs14080680] [Medline: 39199076]
41. Hamatani S, Numata N, Matsumoto K, Sutoh C, Ibuki H, Oshiro K, et al. Internet-based cognitive behavioral therapy via videoconference for patients with bulimia nervosa and binge-eating disorder: pilot prospective single-arm feasibility trial. JMIR Form Res 2019 Oct 23;3(4):e15738 [FREE Full text] [doi: 10.2196/15738] [Medline: 31647472]
42. Taguchi K, Numata N, Takanashi R, Takemura R, Yoshida T, Kutsuzawa K, et al. Clinical effectiveness and cost-effectiveness of videoconference-based integrated cognitive behavioral therapy for chronic pain: randomized controlled trial. J Med Internet Res 2021 Nov 22;23(11):e30690 [FREE Full text] [doi: 10.2196/30690] [Medline: 34813489]
43. Hartley S, Barrowclough C, Haddock G. Anxiety and depression in psychosis: a systematic review of associations with positive psychotic symptoms. Acta Psychiatr Scand 2013 Nov 04;128(5):327-346. [doi: 10.1111/acps.12080] [Medline: 23379898]
44. Malcolm CP, Picchioni MM, Ellett L. Intrusive prospective imagery, posttraumatic intrusions and anxiety in schizophrenia. Psychiatry Res 2015 Dec 30;230(3):899-904. [doi: 10.1016/j.psychres.2015.11.029] [Medline: 26619917]
45. Morrison AP, Baker CA. Intrusive thoughts and auditory hallucinations: a comparative study of intrusions in psychosis. Behav Res Ther 2000 Nov;38(11):1097-1106. [doi: 10.1016/s0005-7967(99)00143-6] [Medline: 11060938]
46. O'Driscoll C, Mason O, Brady F, Smith B, Steel C. Process analysis of trauma-focused cognitive behavioural therapy for individuals with schizophrenia. Psychol Psychother 2016 Jun 23;89(2):117-132. [doi: 10.1111/papt.12072] [Medline: 26202461]
47. Shibuya T, Seki Y, Nagata S, Murata T, Hiramatsu Y, Yamada F, et al. Imagery rescripting of traumatic memories for panic disorder: an exploratory study. tCBT 2018 Jun 06;11:1-21. [doi: 10.1017/s1754470x18000028]
48. Uchida T, Matsumoto K, Kikuchi A, Miyakoshi T, Ito F, Ueno T, et al. Psychometric properties of the Japanese version of the Beck Cognitive Insight Scale: relation of cognitive insight to clinical insight. Psychiatry Clin Neurosci 2009 Jun 14;63(3):291-297 [FREE Full text] [doi: 10.1111/j.1440-1819.2009.01946.x] [Medline: 19566759]
49. Asukai N, Kato H, Kawamura N, Kim Y, Yamamoto K, Kishimoto J, et al. Reliability and validity of the Japanese-language version of the impact of event scale-revised (IES-R-J): four studies of different traumatic events. J Nerv Ment Dis 2002 Mar;190(3):175-182. [doi: 10.1097/00005053-200203000-00006] [Medline: 11923652]
50. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
51. Feng Y, Kohlmann T, Janssen MF, Buchholz I. Psychometric properties of the EQ-5D-5L: a systematic review of the literature. Qual Life Res 2021 Mar 07;30(3):647-673 [FREE Full text] [doi: 10.1007/s11136-020-02688-y] [Medline: 33284428]
52. Woods SW. Chlorpromazine equivalent doses for the newer atypical antipsychotics. J Clin Psychiatry 2003 Jun 15;64(6):663-667. [doi: 10.4088/jcp.v64n0607] [Medline: 12823080]
53. Hall PL, Tarrier N. The cognitive-behavioural treatment of low self-esteem in psychotic patients: a pilot study. Behav Res Ther 2003 Mar;41(3):317-332. [doi: 10.1016/s0005-7967(02)00013-x] [Medline: 12600402]
54. Bighelli I, Huhn M, Schneider-Thoma J, Krause M, Reitmeir C, Wallis S, et al. Response rates in patients with schizophrenia and positive symptoms receiving cognitive behavioural therapy: a systematic review and single-group meta-analysis. BMC Psychiatry 2018 Dec 04;18(1):380 [FREE Full text] [doi: 10.1186/s12888-018-1964-8] [Medline: 30514268]
55. Jung E, Wiesjahn M, Rief W, Lincoln TM. Perceived therapist genuineness predicts therapeutic alliance in cognitive behavioural therapy for psychosis. Br J Clin Psychol 2015 Mar 15;54(1):34-48. [doi: 10.1111/bjc.12059] [Medline: 25040363]
56. Currell S, Christodoulides T, Siitarinen J, Dudley R. Patient factors that impact upon cognitive behavioural therapy for psychosis: therapists' perspectives. Behav. Cogn. Psychother 2015 Jun 25;44(4):493-498. [doi: 10.1017/s1352465815000260]
57. Menon M, Andersen DR, Quilty LC, Woodward TS. Individual factors predicted to influence outcome in group CBT for psychosis (CBTp) and related therapies. Front Psychol 2015 Oct 28;6:1563 [FREE Full text] [doi: 10.3389/fpsyg.2015.01563] [Medline: 26578995]
58. Naeem F, Kingdon D, Turkington D. Predictors of response to cognitive behaviour therapy in the treatment of schizophrenia: a comparison of brief and standard interventions. Cogn Ther Res 2008 Mar 11;32(5):651-656. [doi: 10.1007/s10608-008-9186-x]

XSL•FO
RenderX

59. Gunnmo P, Bergman H. What do individuals with schizophrenia need to increase their well-being. Int J Qual Stud Health Well-being 2011 Feb 14;6(1):5412 [FREE Full text] [doi: 10.3402/qhw.v6i1.5412] [Medline: 21562628]
60. Maisey S, Correia H, Paulik G. The role of self-compassion on the relationship between trauma and hearing voices. Clin Psychol Psychother 2022 Mar 02;29(2):698-705. [doi: 10.1002/cpp.2663] [Medline: 34476866]
61. Achim AM, Maziade M, Raymond E, Olivier D, Mérette C, Roy M. How prevalent are anxiety disorders in schizophrenia? A meta-analysis and critical review on a significant association. Schizophr Bull 2011 Jul 03;37(4):811-821 [FREE Full text] [doi: 10.1093/schbul/sbp148] [Medline: 19959704]
62. Morrison AP, Frame L, Larkin W. Relationships between trauma and psychosis: a review and integration. Br J Clin Psychol 2003 Nov 24;42(Pt 4):331-353. [doi: 10.1348/014466503322528892] [Medline: 14633411]
63. McCartney L, Douglas M, Varese F, Turkington D, Morrison AP, Dudley R. Cognitive behavioural therapy for psychosis targeting trauma, voices and dissociation: a case report. tCBT 2019 Mar 01;12:e18. [doi: 10.1017/S1754470X19000035]
64. Nemoto T, Uchino T, Aikawa S, Matsuo S, Mamiya N, Shibasaki Y, et al. Impact of changes in social anxiety on social functioning and quality of life in outpatients with schizophrenia: A naturalistic longitudinal study. J Psychiatr Res 2020 Dec;131:15-21 [FREE Full text] [doi: 10.1016/j.jpsychires.2020.08.007] [Medline: 32911206]
65. Halverson TF, Meyer-Kalos PS, Perkins DO, Gaylord SA, Palsson OS, Nye L, et al. Enhancing stress reactivity and wellbeing in early schizophrenia: A randomized controlled trial of Integrated Coping Awareness Therapy (I-CAT). Schizophr Res 2021 Sep;235:91-101. [doi: 10.1016/j.schres.2021.07.022] [Medline: 34332429]
66. Emsley R, Chiliza B, Asmal L, Harvey BH. The nature of relapse in schizophrenia. BMC Psychiatry 2013 Feb 8;13(1):50. [doi: 10.1186/1471-244x-13-50]
67. Lukoff D, Snyder K, Ventura J, Nuechterlein KH. Life events, familial stress, and coping in the developmental course of schizophrenia. Schizophr Bull 1984 Jan 01;10(2):258-292. [doi: 10.1093/schbul/10.2.258] [Medline: 6729412]
68. Myin-Germeys I, Delespaul P, van Os J. Behavioural sensitization to daily life stress in psychosis. Psychol Med 2005 May;35(5):733-741. [doi: 10.1017/s0033291704004179] [Medline: 15918350]
69. Ventura J, Nuechterlein KH, Lukoff D, Hardesty JP. A prospective study of stressful life events and schizophrenic relapse. J Abnorm Psychol 1989 Nov;98(4):407-411. [doi: 10.1037//0021-843x.98.4.407] [Medline: 2592674]
70. Maki S, Nagai K, Ando S, Tamakoshi K. Structure and predictors of in-hospital nursing care leading to reduction in early readmission among patients with schizophrenia in Japan: A cross-sectional study. PLoS One 2021 Apr 30;16(4):e0250771 [FREE Full text] [doi: 10.1371/journal.pone.0250771] [Medline: 33930056]
71. Wang Q, Zhu X, Jiang X, Li M, Chang R, Chen B, et al. Relationship between stressful life events, coping styles, and schizophrenia relapse. Int J Ment Health Nurs 2021 Oct 06;30(5):1149-1159 [FREE Full text] [doi: 10.1111/inm.12865] [Medline: 33960095]
72. Villeneuve K, Potvin S, Lesage A, Nicole L. Meta-analysis of rates of drop-out from psychosocial treatment among persons with schizophrenia spectrum disorder. Schizophr Res 2010 Aug;121(1-3):266-270. [doi: 10.1016/j.schres.2010.04.003] [Medline: 20452749]
73. Szymczynska P, Walsh S, Greenberg L, Priebe S. Attrition in trials evaluating complex interventions for schizophrenia: Systematic review and meta-analysis. J Psychiatr Res 2017 Jul;90:67-77. [doi: 10.1016/j.jpsychires.2017.02.009] [Medline: 28231496]
74. Jørgensen R, Munk-Jørgensen P, Lysaker PH, Buck KD, Hansson L, Zoffmann V. Overcoming recruitment barriers revealed high readiness to participate and low dropout rate among people with schizophrenia in a randomized controlled trial testing the effect of a Guided Self-Determination intervention. BMC Psychiatry 2014 Feb 3;14(1):28. [doi: 10.1186/1471-244x-14-28]
75. Simon GE, Yarborough BJ, Rossom RC, Lawrence JM, Lynch FL, Waitzfelder BE, et al. Self-reported suicidal ideation as a predictor of suicidal behavior among outpatients with diagnoses of psychotic disorders. Psychiatr Serv 2019 Mar 01;70(3):176-183 [FREE Full text] [doi: 10.1176/appi.ps.201800381] [Medline: 30526341]
76. Watson A, Mellotte H, Hardy A, Peters E, Keen N, Kane F. The digital divide: factors impacting on uptake of remote therapy in a South London psychological therapy service for people with psychosis. J Ment Health 2022 Dec 28;31(6):825-832. [doi: 10.1080/09638237.2021.1952955] [Medline: 34319202]
77. Donker T, Petrie K, Proudfoot J, Clarke J, Birch M, Christensen H. Smartphones for smarter delivery of mental health programs: a systematic review. J Med Internet Res 2013 Nov 15;15(11):e247 [FREE Full text] [doi: 10.2196/jmir.2791] [Medline: 24240579]
78. Hedman E, Ljótsson B, Lindefors N. Cognitive behavior therapy via the Internet: a systematic review of applications, clinical efficacy and cost-effectiveness. Expert Rev Pharmacoecon Outcomes Res 2012 Dec;12(6):745-764. [doi: 10.1586/erp.12.67] [Medline: 23252357]
79. Kopelovich SL, Turkington D. Remote CBT for psychosis during the COVID-19 pandemic: challenges and opportunities. Community Ment Health J 2021 Jan;57(1):30-34 [FREE Full text] [doi: 10.1007/s10597-020-00718-0] [Medline: 33001323]
80. Santesteban-Echarri O, Piskulic D, Nyman RK, Addington J. Telehealth interventions for schizophrenia-spectrum disorders and clinical high-risk for psychosis individuals: A scoping review. J Telemed Telecare 2018 Aug 22;26(1-2):14-20. [doi: 10.1177/1357633x18794100]

XSL•FO
RenderX

81. Kopelovich SL, Strachan E, Sivec H, Kreider V. Stepped care as an implementation and service delivery model for cognitive behavioral therapy for psychosis. Community Ment Health J 2019 Jul 8;55(5):755-767. [doi: 10.1007/s10597-018-00365-6] [Medline: 30623294]

Abbreviations

BCIS-J: Japanese version of the Beck Cognitive Insight Scale
CBTp: cognitive behavioral therapy for psychosis
CONSORT: Consolidated Standards of Reporting Trials
CPZ: chlorpromazine
GAD-7: Generalized Anxiety Disorder-7
IES-R: Impact of Event Scale-Revised
MacCAT-CR: MacArthur Competence Assessment Tool for Clinical Research
NICE: National Institute of Health and Clinical Excellence
PANSS: Positive and Negative Syndrome Scale
PHQ-9: Patient Health Questionnaire-9
RCT: randomized controlled trial
UC: usual care
vCBTp: videoconference-based cognitive behavioral therapy for psychosis

Edited by A Mavragani; submitted 16.04.24; peer-reviewed by E Granholm, M Kobayashi, S Zaman; comments to author 17.06.24; revised version received 07.11.24; accepted 28.11.24; published 14.01.25.

Please cite as:
Katsushima M, Nakamura H, Shiko Y, Hanaoka H, Shimizu E
Effectiveness of a Videoconference-Based Cognitive Behavioral Therapy Program for Patients with Schizophrenia: Pilot Randomized Controlled Trial
JMIR Form Res 2025;9:e59540
URL: https://formative.jmir.org/2025/1/e59540
doi:10.2196/59540
PMID:39610049

©Masayuki Katsushima, Hideki Nakamura, Yuki Shiko, Hideki Hanaoka, Eiji Shimizu. Originally published in JMIR Formative Research (https://formative.jmir.org), 14.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Testing a Web-Based Interactive Comic Tool to Decrease Obesity Risk Among Racial and Ethnic Minority Preadolescents: Randomized Controlled Trial

May May Leung[1], RDN, PhD; Katrina F Mateo[2], MPH; Marlo Dublin[3], RDN, MS; Laura Harrison[2], MPH; Sandra Verdaguer[2], MPH; Katarzyna Wyka[2], PhD

[1]Friedman School of Nutrition Science and Policy, Tufts University, Boston, MA, United States
[2]Graduate School of Public Health Policy, City University of New York, New York, NY, United States
[3]School of Urban Public Health, Hunter College, City University of New York, New York, NY, United States

Corresponding Author:
May May Leung, RDN, PhD
Friedman School of Nutrition Science and Policy
Tufts University
150 Harrison Avenue
Boston, MA, 02111
United States
Phone: 1 6176363676
Email: maymay.leung@tufts.edu

Abstract

Background: Childhood obesity prevalence remains high, especially in racial and ethnic minority populations with low incomes. This epidemic is attributed to various dietary behaviors, including increased consumption of energy-dense foods and sugary beverages and decreased intake of fruits and vegetables. Interactive, technology-based approaches are emerging as promising tools to support health behavior changes.

Objective: This study aimed to assess the feasibility and acceptability of Intervention INC (Interactive Nutrition Comics for Urban, Minority Preadolescents), a 6-chapter web-based interactive nutrition comic tool. Its preliminary effectiveness on diet-related psychosocial variables and behaviors was also explored.

Methods: A total of 89 Black or African American and Hispanic preadolescents with a mean age of 10.4 (SD 1.0) years from New York City participated in a pilot 2-group randomized study, comprising a 6-week intervention and a 3-month follow-up (T4) period. Of the 89 participants, 61% were female, 62% were Black, 42% were Hispanic, 53% were overweight or obese, and 34% had an annual household income of <US $20,000. Participants were randomly assigned to the experimental group (45/89, 50% received the web-based comic tool), or the comparison group (44/89, 50% received web-based nutrition newsletters). Primary measures included feasibility and usability at intervention midpoint (T2) and intervention end (T3). Semistructured interviews were conducted at the same time to assess acceptability and satisfaction. Secondary measures, collected at baseline (T1), T2, T3, and at T4, included attitudes, beliefs, and behaviors related to fruit, vegetable, water, sugar, and junk food intake. Descriptive analyses were conducted for use and usability data. Interviews were systematically analyzed to facilitate identification of patterns and themes. Secondary data were analyzed using descriptive statistics. Within- and between-group effect sizes were reported.

Results: In total, 72% (33/45) and 60% (27/44) of the experimental and comparison groups, respectively, accessed their tool weekly. The mean total usability score was high and moderately high for the experimental and comparison groups, respectively (mean 4.01, SD 0.37 and mean 3.81, SD 0.51; P=.048), based on a 5-point Likert scale). Children in both groups found the tool acceptable, and few reported difficulties logging in or accessing content. Between-group effect sizes for beliefs and attitudes related to dietary intake, while favoring the experimental group at T3, were in the small range. These improvements in both groups were largely diminished by T4. However, between-group effect sizes for behaviors related to fruit, vegetable, and water intake, favoring the experimental group, were medium to large and were maintained at T4.

Conclusions: This pilot feasibility study suggests that an interactive comic tool may be an appealing and useful format to promote positive dietary behaviors in racial and ethnic minority preadolescents. However, further research, including a full-scale randomized controlled trial, is warranted to determine the effectiveness of Intervention INC.

XSL•FO
RenderX

Trial Registration: ClinicalTrials.gov NCT03165474; https://clinicaltrials.gov/study/NCT03165474
International Registered Report Identifier (IRRID): RR2-10.2196/10682

(JMIR Form Res 2025;9:e58460) doi:10.2196/58460

KEYWORDS
childhood obesity; preadolescents; racial and ethnic minority populations; dietary behaviors; BMI; digital health

Introduction

Background

The prevalence of childhood obesity in the United States, especially among racial and ethnic minority communities with low incomes, remains consistently high and shows no signs of stabilizing. Specifically, Hispanic and Black or African American children have markedly higher rates of obesity (25.8% and 22%, respectively) compared to the national prevalence of 18.5% [1], with rates rising as age progresses, especially in African American and Hispanic children [2]. Compared to school-age children (18.4%) and preschoolers (13.9%), adolescents (20.6%) have a higher prevalence of obesity [3,4]. In the short term, obesity in youth may lead to issues such as an increased risk of elevated blood pressure and lipid concentrations, anxiety, and depression [5,6]. In the long term, adverse health conditions, including cardiovascular disease, diabetes, and disabilities, can manifest in youth who are obese and be maintained into adulthood [7-10].

Childhood obesity has been linked to several behavioral determinants, namely, increased consumption of highly processed, energy-dense foods and sugary beverages and a decreased intake of fruits and vegetables [11-13]. With the pervasiveness of daily technology and new media use in youth, particularly within the Hispanic and African American population [14], digital platforms are increasingly being used to deliver creative messaging around health and nutrition to children and their caregivers [6]. These tools, which include mobile apps and interactive websites, are emerging as promising approaches to engage and deliver tailored interventions to support health behavior change [15]. However, knowledge gaps in the usability, engagement, and usefulness of such applications still exist, limiting their potential effectiveness [16]. Furthermore, there is a shortage of effective health promotion interventions that are culturally tailored to address the needs and preferences of populations with disproportionately high rates of chronic disease. Even fewer interventions have been designed specifically for children, particularly those transitioning from childhood to adolescence [17]. It is crucial to intervene during this development stage as food preferences established during this period often persist into adulthood [18-20]. In addition, children at this age are gaining autonomy and developing decision-making skills related to dietary behaviors [21], which highlights the importance of intervening at this critical period.

Study Purpose

The primary purpose of this pilot study was to assess the feasibility and participant acceptability of Intervention INC (Interactive Nutrition Comics for Urban, Minority Preadolescents), an interactive, web-based health promotion tool. In addition, this study aimed to explore tool effectiveness by determining if it improved attitudes, beliefs, and behaviors related to obesity. Taken together, these data will provide insight for further improvements to the tool and reveal user preferences, which is important feedback for developing culturally tailored, effective, and engaging digital tools, particularly for populations at risk for childhood obesity.

Methods

Study Design

The Intervention INC study was a pilot, 2-group randomized controlled trial (RCT) that evaluated a 6-week intervention promoting fruit, vegetable and water consumption, with a 3-month follow-up (T4) period. A protocol for this feasibility trial was published and registered with the Clinical Trials Registry (NCT03165474) [22]. Parent-child dyads were randomized into either the experimental or comparison group, with both child and parent participants blinded to their assigned group. Details regarding the randomization process are provided in the protocol paper [22].

Participants

Children residing in New York City (NYC) were recruited based on the following inclusion and exclusion criteria: self-identified as Black, African American, or Latino; aged between 9 and 12 years (preadolescents) at the time of scheduled baseline (T1) visit; read and spoke in English; had a BMI percentile ≥5% at T1 (categorized as healthy, overweight, or obese); had regular internet access via a tablet device, mobile phone, or computer or laptop; had regular access to a phone with texting capability; was comfortable reading or viewing material on electronic devices; was comfortable speaking with study staff about thoughts or experiences while participating in the study; had no allergies, food aversions, food disorders, or medications with side effects that may impact participation in the study; did not have a pacemaker or heart condition; and had a legal parent or guardian willing to participate in the study. While parents or guardians were a part of this study, the focus of this paper is to present primary outcome data for the child participants.

Multiple recruitment approaches were implemented between August and November 2017. Initially, mailings were sent to parents or guardians of potentially eligible child patients of a community-based clinic (partnering organization) in Harlem, NYC. Recruitment efforts were expanded to include community flyers in East Harlem and Upper Manhattan as well as posting notices in and around local businesses, housing complexes, community centers, schools, and churches. Through partnerships with local schools and community initiatives, recruitment efforts

XSL•FO
RenderX

also occurred via tabling at community and school events. Further details regarding study recruitment and sample size considerations are included in other publications [22,23].

Intervention

Content for the child participants, regardless of group assignment, was tailored (to include more information related to either fruit and vegetable or water) based on responses to initial screening questions, at T1, related to child fruit, vegetable and water intake, child self-efficacy to increase fruit, vegetable and water intake, and parent self-efficacy to support the child in increasing fruit, vegetable and water intake.

Experimental Group

Children in the experimental group received access to Intervention INC, an interactive web-based comic tool informed by the Health Belief Model [24], Social Cognitive Theory [25], and the Narrative Transportation Theory [26], which contained health messages focused on fruit, vegetable and water consumption. It comprises a 6-chapter comic with pop-up windows (highlighting health facts, food-related fun facts, or character information) and prompted audio and visual effects to enhance tool engagement. It also includes a weekly goal setting and assessment feature and texting or email messages reminding children to read the comic and work on their selected health behavior goals. Most of the health messages are delivered by the comic characters themselves in a narrative form, but some messages were delivered didactically (such as in the pop-up windows), as entertainment education research suggests a combination of narrative and nonnarrative information may be the most effective in improving health beliefs and behaviors [27]. Multiple behavior change techniques (BCTs) were used to promote healthy dietary behaviors. Table 1 lists all the BCTs that were incorporated into the comic and examples of how each of them were used. The comic tool was designed and developed by a collaborative research team, consisting of scientists, digital health experts, content and system developers, and intended users (Black or African American and Hispanic preadolescents and their parents). Extensive formative research, including usability testing, was conducted during the development process [28]. Additional details of the intervention, including its development and specific components, have been described in previous papers [22,28].

XSL•FO
RenderX

Table 1. BCTs[a] incorporated into the web-based comic tool, Intervention INC (Interactive Nutrition Comics for Urban, Minority Preadolescents).

BCT	Action	Example
Goal setting (behavior)	At the end of each comic chapter, a character prompts the child to select a weekly goal to work on.	• End of chapter message includes a picture of a character along with a prompt to set a goal: "Let's pick a goal to work on for the week!" • Sample goal: "I will eat fruits I like (such as bananas or apples) at breakfast."
Self-monitoring of behavior	The following week, the child is prompted to assess how they did with the weekly goal they selected to work on.	• End of chapter message: "Your goal for last week was (selected goal)." • "How often did you do this in the last week?" • Response options: never, sometimes, most of the time, and all of the time
Feedback on behavior	On the basis of the child's self-assessment of the weekly goal, feedback is provided on the performance of the behavior.	• Sample message if a child achieves the goal: "Congrats! Keep up the good work!" • Sample message if a child does not achieve the goal: "That's ok! Michael Jordan was cut from his high school basketball team, and he didn't give up!"
Social support	Children can select weekly goals related to practical social support from their parents.	• Sample goal related to practical parental social support: "I will ask my parent to pack a water bottle for me when I leave the house this week."
Information about health consequences	Information about the consequences of eating healthfully or unhealthfully is provided in the comic in didactic and narrative forms, such as pop-up messages, character dialogue, and end of chapter messages from characters.	• Sample pop-up message (in a panel where one character tells another to stay hydrated): "Did you know that not having enough water can make you feel dehydrated? Grab a water if you're feeling tired or you're having a hard time concentrating."
Salience of consequences	Scenes are included in the comic to specifically emphasize the consequences of performing both positive and negative dietary behaviors, with the aim of making them more memorable.	• A meter highlights the energy levels of the comic characters based on the nutritional quality of the foods they consume. For example, a character eats a banana, and his energy meter rises. He also acknowledges how "great" he feels.
Demonstration of behavior	Comic characters model positive dietary behaviors.	• The father of the main character is seen drinking water and eating fruits and vegetables in multiple scenes.
Prompts or cues	Messages from the comic characters are delivered weekly (via texting, email, or both) to remind children to work on the goal they selected that week.	• Sample text or email message from a character: "Hey, [child name]—We have a mission for you! Be sure to log in and choose a goal if you haven't already. Then you'll get tips to train and get stronger! See you soon!"
Social reward	Positive feedback is provided if the child achieves their weekly goal.	• Bonus content related to the characters and comic story is released if the child achieves their weekly goal.

[a]BCT: behavior changes technique.

Comparison Group

Children in the comparison group received access to 6 web-based newsletters. Health-related content for the comparison group was similar to that for the experimental group but presented in a newsletter (didactic) format, which includes tips, recipes, diet-related knowledge and facts, and web-based games. Similar to the experimental group, the comparison group also had a weekly goal setting and assessment component and received reminder messages via texting, email or both.

Data Collection and Measures

Overview

Data were collected at 4 different time points: T1, intervention midpoint or 3 weeks after T1 (T2), intervention end or 6 weeks after T1 (T3), and T4. Brief descriptions of our primary and secondary outcome measures are provided in the following sections, while details of all measures are described elsewhere [22]. T1 and T4 data collection was conducted in person at the study site, while T2 and T3 data collection was conducted via phone. At T2 and T3, links to any questionnaires were sent to the participants ahead of time to complete, and then, responses were confirmed during the phone call. If the questionnaire was not completed before the call, children completed it during the call with a trained research assistant. Table 2 lists the measures collected and the time points at which they were collected.

XSL•FO
RenderX

Table 2. Key primary and secondary measures, data sources, and time points at which the measures were assessed.

Measures	Data source	Time points				
		T1[a]	T2[b]	T3[c]	T4[d]	O[e]
Feasibility and acceptability measures (primary measures)						
Use of the web-based tool	Tracking system (internally created)					✓
Usability of the web-based tool	Interview and questionnaire items[f]		✓	✓	✓	
Acceptance and satisfaction	Interview				✓	
Outcome measures (secondary measures)						
Dietary beliefs and attitudes	Questionnaire items[f]	✓	✓	✓	✓	
Dietary intake	Questionnaire items[f]	✓	✓	✓	✓	
BMI-for-age percentile	Digital stadiometer	✓			✓	

[a]T1: baseline.
[b]T2: midpoint (3 weeks after baseline).
[c]T3: intervention end (6 weeks after baseline).
[d]T4: follow-up (3 months after intervention).
[e]O: ongoing throughout the intervention period.
[f]All questionnaire items were either taken from or directly informed by validated questionnaires.

Feasibility and Acceptability Measures (Primary Measures)

Use of the Web-Based Comic Tool

Use data, such as the number of weeks logged in, were captured throughout the intervention on a custom-built platform. Additional use data related to goal setting were collected, including the percentage of participants who selected, assessed (ie, "How often did you do this [behavior] in the last week?"), and achieved (defined as responses of "Most of the Time" or "All the time") weekly dietary goals.

Usability of the Web-Based Comic Tool

Five domains of usability (usability, usefulness, ease of use, ease of learning, and satisfaction) were assessed with a 30-item questionnaire on a 5-point Likert scale (ranging from strongly disagree to strongly agree). This questionnaire was adapted from the System Usability Scale [29]; Usefulness, Satisfaction, and Ease-of-use questionnaire [30]; and an acceptability or usability measure by Ben-Zeev et al [31]. Minor modifications were made to tailor the questionnaire to the literacy levels of our intended population based on pilot testing with them. Specific details related to the modification process are described elsewhere [22]. The modified questionnaire was not validated for web use. A total score of 4.0 out of 5.0 was considered high usability.

Acceptance and Satisfaction

Semistructured interviews were conducted at T2 and T3 where open-ended questions were asked to both experimental and comparison group participants about the experience of engaging with their web-based tool. Specifically, interview questions aimed to assess participant acceptability and satisfaction with the overall tool and its specific components.

Outcome Measures (Secondary Measures)

Dietary Beliefs and Attitudes

Outcome expectations (OEs), self-efficacy, behavioral intention (BI), and attitudes related to fruit, vegetable, water, junk food, and sugary drinks consumption were assessed at all 4 time points with an 84-item questionnaire, which was informed by and modified from the validated ProChildren Questionnaire and the validated Reynolds Questionnaire [32,33]. The OE, self-efficacy, BI, and attitudes questionnaires had a Cronbach α of 0.87, 0.74, 0.52, and 0.70, respectively.

Dietary Intake

Frequency of consumption of fruit, vegetable, water, junk food and sugary drinks during the past 7 days were assessed at all 4 time points with a 17-item questionnaire comprising questions from and informed by the validated 2017 Youth Risk Behavior Surveillance System questionnaire and the validated Beverage and Snack Questionnaire [34,35]. Three items assessing the intake of different types of water were internally created. The dietary intake questionnaire had a Cronbach α of 0.79.

BMI-for-Age Percentile

The height and weight of child participants were measured at T1 and T4 using standardized methods [36]. A SECA 264 digital stadiometer and Tanita MC-780U body composition monitor were used to collect anthropometric measures. The Centers for Disease Control and Prevention BMI percentile calculator was used to determine BMI-for-age percentile [37].

Data Analysis

Feasibility and Acceptability Measures (Primary Measures)

Use and Usability of the Web-Based Comic Tool

Use of the web-based tool for each group was assessed by first aggregating the "click-tracking" data by participant, URL, study

XSL•FO
RenderX

week, and across the intervention period. Descriptive analyses, including means, SD, and ranges (minimum-maximum) were performed. Similar descriptive analyses were also conducted to measure the overall score for the usability data of the web-based tools across both groups.

Acceptance and Satisfaction

Interviews from T2 and T3 were systematically analyzed by 3 trained coders through a streamlined process of directly listening to individual audio files and then transferring relevant information into a matrix sheet to produce a summary of overall impressions of usability, acceptability, and feasibility of the tool and its specific components (ie, reminders and goal setting). A rapid memo technique, informed by the Rapid Evaluation and Assessment Method [38] and the Rapid Identification of Themes From Audio Recordings [38] analysis method, was used to summarize interview responses as well as to note key quotations. These techniques have been used as a way to preclude the traditional steps of transcribing qualitative data and coding transcripts when resources are limited and quick reporting time is needed [38,39].

Audio files were randomly assigned to each coder. In addition to the primary coder, a secondary coder reviewed the audio files for accuracy, noting any additional information, until very few additions were being made or discrepancies were noted. This process was used for approximately the first 10 interviews. However, when a coder had a subsequent audio file with inaudible sections, another coder conducted a secondary review and added any missing information to the matrix. Coders met frequently throughout the analysis process to compare their results; discuss overall matrix content and trends; identify, clarify, and understand emergent findings; and highlight representative quotations.

Outcome Measures (Secondary Measures)

Child and parent sociodemographic characteristics were compared between conditions (experimental or comparison) using 2-tailed independent samples *t* tests and chi-square tests. To assess improvements in outcome measures (child dietary beliefs and attitudes, dietary intake, and BMI-for-age percentile), change scores (relative to T1) were computed and summarized using descriptive statistics. To estimate the magnitude of the effects at T3 and T4, both within- and between-group effect sizes were computed on change scores. Analyses were completed using SPSS statistical software (version 26; IBM Corp) [40]. Of note, the data analyst was blinded to the intervention assignment to minimize bias.

Ethical Considerations

All study activities were approved by the Institutional Review Board at Hunter College in New York, New York (2015-0547). Participation was entirely voluntary, and study participants could choose to withdraw from the study, without any reason, at any point in time during the study. Informed consent, child assent and parental permission were obtained before study participation. Both child and parent participants were compensated. Children received up to US $70 and parents or guardians received up to US $65 (in the form of gift cards) for completing data collection. To maintain participant privacy and confidentiality, all study data have been deidentified.

Results

Participants

A total of 89 child-parent dyads were recruited and completed T1 measures. There were no differences in characteristics between the 2 groups. The mean age of child participants was 10.38 (SD 1.03) years; the majority were female (54/89, 61%), Black or African American (42/89, 47%) or Hispanic (29/89, 33%), and overweight or obese (47/89, 53%). The mean age of parent participants was 40.83 (SD 8.87) years; nearly all were female (84/89, 94%), born in the United States (63/89, 71%), with some college education or more (61/89, 68%), and single (40/89, 45%). One-third (30/89, 34%) of the dyads reported an annual household income of <US $20,000, and nearly two-thirds (54/89, 61%) reported receiving Supplemental Nutrition Assistance Program benefits. Refer to Table 3 for additional demographic details. The CONSORT (Consolidated Standards of Reporting Trials) diagram (Figure 1) shows the randomization and retention progress throughout the study. Retention rates were high: T2, 87% (77/89); T3, 89% (79/89); and T4, 84% (75/89) [23].

XSL•FO
RenderX

Table 3. Baseline child and parent demographic characteristics (N=89).

Characteristics	Overall (N=89)	E[a] (n=45)	C[b] (n=44)	*P* value[c]
Child				
Age (y), mean (SD)	10.38 (1.03)	10.38 (1.06)	10.39 (1.01)	.97
Sex n (%)				
Male	35 (39)	17 (38)	18 (41)	.76
Female	54 (61)	28 (62)	26 (59)	__[d]
Race or ethnicity, n (%)				
Hispanic	37 (42)	20 (51)	17 (49)	.58
Black or AA[e]	55 (62)	28 (54)	27 (46)	.93
Multiracial or multiethnic	18 (20)	10 (56)	8 (44)	.64
BMI-for-age percentile, mean (SD)	74 (26.5)	75 (26.4)	73 (26.8)	.66
BMI category, n (%)				
Normal	42 (47)	21 (47)	21 (48)	.91
Overweight	19 (21)	9 (20)	10 (23)	—
Obese	28 (32)	15 (33)	13 (30)	—
Parent				
Age (y), mean (SD)	40.83 (8.87)	40.07 (7.67)	41.63 (10.01)	.41
Sex, n (%)				
Male	5 (6)	2 (4)	3 (7)	.63
Female	84 (94)	43 (96)	41 (93)	—
Race or ethnicity, n (%)				
Black or AA	49 (55)	25 (56)	24 (55)	.93
Hispanic	40 (45)	22 (49)	18 (51)	.45
Multiracial or multiethnic	9 (20)	6 (13)	3 (7)	.31
Country of birth, n (%)				
United States	63 (71)	31 (69)	32 (73)	.69
Other	26 (29)	14 (31)	12 (27)	—
Bilingual, n (%)				
No	62 (70)	31 (69)	31 (71)	.87
Yes	27 (30)	14 (31)	13 (30)	—
Education, n (%)				
Less than high school	18 (21)	11 (25)	7 (16)	.45
Finished high school or GED[f]	10 (11)	6 (14)	4 (9)	—
Some college	26 (230)	10 (23)	16 (36)	—
Finished college	34 (39)	17 (39)	17 (39)	—
Marital status, n (%)				
Single	40 (45)	22 (49)	18 (41)	.70
Married or in a marriage-like relationship	35 (39)	17 (38)	18 (41)	—
Separated, divorced, or widowed	14 (16)	6 (13)	8 (18)	—
Annual household income (US $), n (%)				
<20,000	30 (34)	15 (33)	15 (34)	.66
20,000-39,999	30 (34)	17 (38)	13 (30)	—

XSL•FO
RenderX

Characteristics	Overall (N=89)	E[a] (n=45)	C[b] (n=44)	*P* value[c]
>40,000	29 (33)	13 (29)	16 (36)	—
SNAP[g] benefits, n (%)				
Yes	54 (61)	27 (60)	27 (61)	.90
No	35 (39)	18 (40)	17 (39)	—

[a]E: experimental group.

[b]C: comparison group.

[c]*P* values are based on independent sample *t* tests (continuous variables) or chi-square tests (categorical variables). Conclusions are based on one *P* value, apart from race or ethnicity categories, which are not mutually exclusive.

[d]Not applicable.

[e]AA: African American.

[f]GED: General Educational Development.

[g]SNAP: Supplemental Nutrition Assistance Program.

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) diagram of the randomization and retention progress throughout the study. T1: baseline; T2: midpoint (3 weeks after baseline); T3: intervention end (6 weeks after baseline); T4: follow-up (3 months after intervention).

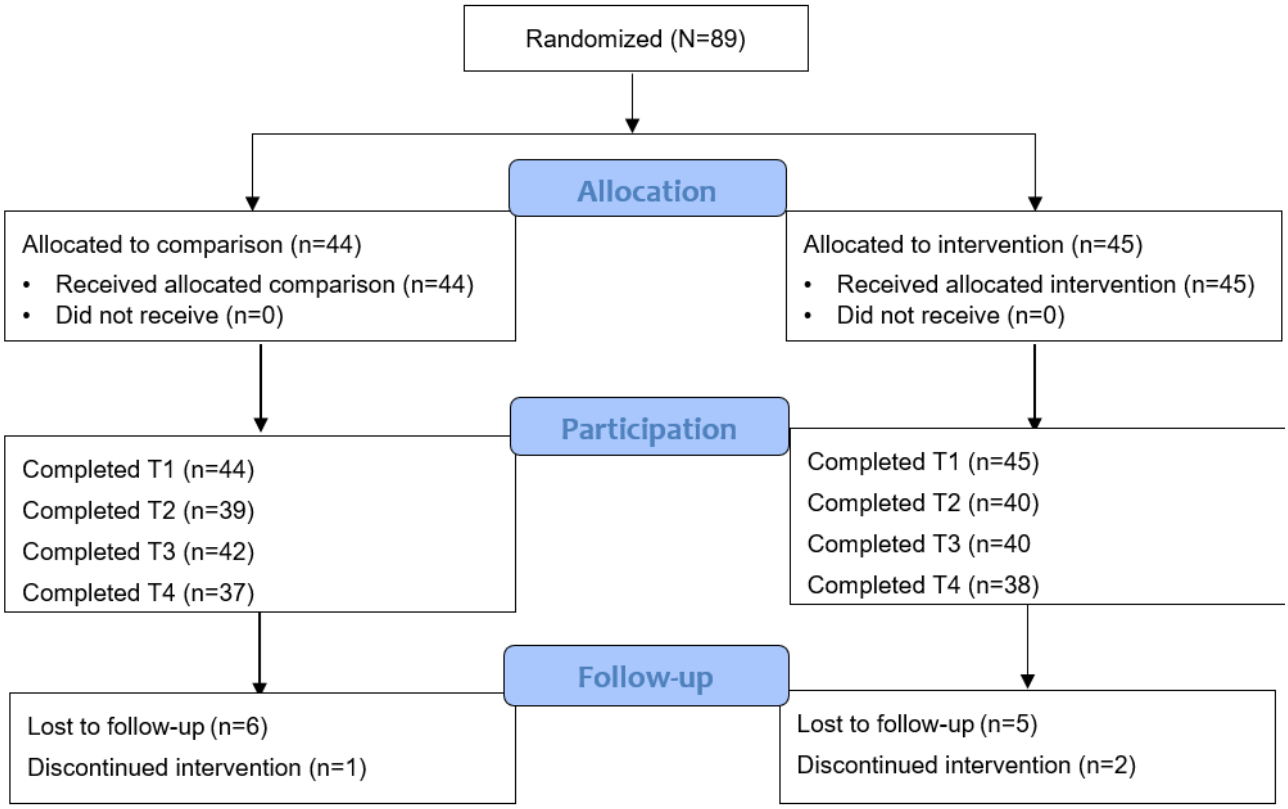

Feasibility and Acceptability Measures (Primary Measures)

Use of the Web-Based Comic Tool

Overview

During the 6-week intervention, 73% (33/45) of children in the experimental group accessed their tool weekly. Percentage of those who accessed the tool each week were as follows: week 1, 100% (45/45); week 2, 76% (34/45); week 3, 71% (32/45); week 4, 64% (29/45); week 5, 62% (28/45); and week 6, 56% (25/45). The experimental group accessed the web tool an average of 7.84 (SD 5.09) days, out of 42 possible days. Figure 2 highlights the percentage of children who accessed each of the 6 comic chapters. Of those who accessed the comic chapters, an average of 65% (SD 16.8) of each comic chapter was read, while most of the children (24/44, 53%) read the chapters in full.

XSL•FO
RenderX

Figure 2. Percentage of the 6 comic chapters that were read by children in the experimental group.

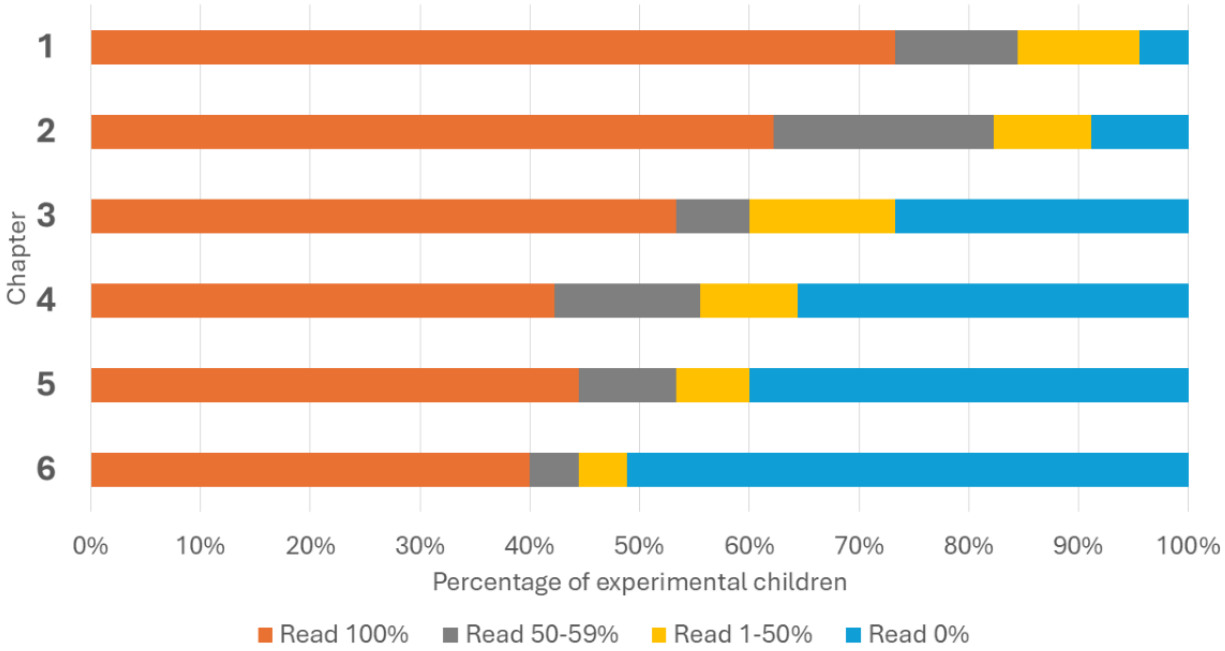

In the comparison group, 60% (27/44) accessed their newsletter weekly. Percentage of those who accessed the tool each week were as follows: week 1, 100% (44/44); week 2, 50% (22/44); week 3, 52% (23/44); week 4, 59% (26/44); week 5, 59% (26/44); and week 6, 41% (18/44). Furthermore, the comparison group accessed the web tool an average of 5.82 (SD 4.20) days, out of 42 possible days.

Goal Setting Feature

On average, 49% (22/45) of children in the experimental group selected a goal each week. Of all the weekly goals that were selected, 67.9% (74/109) were assessed, while 40% (30/74) of those goals were reported to have been achieved. For the comparison group, on average, 51% (23/44) of the children selected a goal. Of all the weekly goals selected, 61.7% (71/115) were assessed. Of the assessed goals, 37% (26/71) were reported to have been achieved. Figure 3 presents the percentage of those in the experimental and comparison groups who selected, assessed, and achieved a goal during each week of the intervention. There was a general decline in goal selection over the weeks in both experimental and comparison groups—in week 1, 69% (31/45) and 82% (36/44) selected goals; by week 5, 42% (19/45) and 36% (16/44) selected goals, respectively. In terms of goal assessment, trends were observed between the experimental and comparison groups in week 1 (P=.06) and week 4 (P=.07), however, these results were not statistically significant. In week 1, 74% (23/31) and 39% (14/36) assessed goals they selected, while in week 4, 88% (14/16) and 57% (12/21) assessed goals for experimental and comparison groups, respectively (P=.07 and P=.10, respectively). No other differences were observed across the other weeks. In terms of achieving the weekly goals, the experimental group slightly declined over the weeks (week 1; 11/23, 46% to week 5; 5/12, 42%), while there appeared to be an increase in the comparison group (week 1; 5/14, 36% to week 5; 5/12, 42%).

Figure 3. Percentage of experimental and comparison groups who selected, assessed, and reported achieving weekly goals across the 6-week intervention.

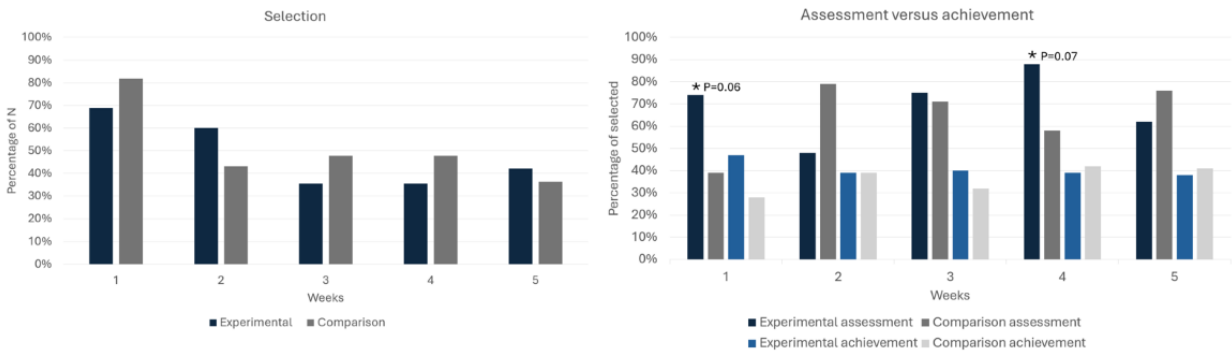

XSL•FO
RenderX

Usability of the Web-Based Comic Tool

Table 4 highlights the mean scores of the 5 usability domains (usability, usefulness, ease of use, ease of learning, and satisfaction) and total for both the experimental and comparison groups. The experimental group reported moderate to high scores in each of the domains. Specifically, ease of learning and ease of use had the highest scores relative to their maximum range (4.58, SD 0.72 and 4.47, SD 0.46, respectively). The mean total score of perceived usability was high and moderately high for the experimental and comparison groups, respectively (4.01, SD 0.37 and 3.81, SD 0.51; *P*=.048).

Table 4. Perceived usability of the assigned tool at T3[a], by group[b].

Domain	E[c] (n=43), mean (SD)	C[d] (n=41), mean (SD)	*P* value
Usability (30 items)	3.14 (0.44)	3.05 (0.40)	__[e]
Usefulness (10 items)	4.35 (0.68)	4.02 (1.11)	—
Ease of use (10 items)	4.47 (0.46)	4.25 (0.66)	—
Ease of learning (2 items)	4.58 (0.72)	4.40 (0.82)	—
Satisfaction (6 items)	4.36 (0.62)	4.07 (0.82)	—
Total	4.01 (0.37)	3.81 (0.51)	.048

[a]T3: intervention end point or 6 weeks postbaseline.

[b]Assessment questionnaire was developed using a total of 30 items from the System Usability Scale [41]; Usefulness, Satisfaction, and Ease-of-use [42] questionnaire; and acceptability or usability measure [43]. Response options ranged from 1 (strongly disagree) to 5 (strongly agree).

[c]E: experimental group.

[d]C: comparison group.

[e]*P* values for individual usability domains have been removed as they were determined to not be applicable.

Acceptance and Satisfaction

Overall, children in both arms found their tool to be acceptable, and very few reported any difficulties logging in or accessing content. Specific findings for different components of the comic tool are noted in the following sections.

Comic

Most children shared that the comic is a "fun way to read" as it's exciting, adventurous, motivating, and "helpful like a video game." Some children explained that the storyline teaches new things about fruits, vegetables, and water and how to stay healthy and take care of oneself. One male child participant described it as "a superhero story where being nutritious is the only way to save the world and junk food slowly starts depleting your energy and, in the end, makes you an awful person." Most children also reported engaging a parent or family member in their experience of reading the comic.

The comic included a variety of elements aimed at engaging children, including unique comic characters and audio-visual effects. Most children liked the comic characters because of their personalities, which were described as "cool," "different," "mysterious," "admirable" and "brave," among other adjectives. Some children reported that the characters gave them a good feeling or that they felt a connection to them as they reminded them of someone in their life or shared a favorite hobby. Moreover, nearly all children reported liking the audio or visual effects. Reasons included that the effects made the story more imaginable, "fun to read" and it "gives you a feeling that you are there...in the story." In addition, some children noted that the audio effects allowed the tool to read for them "if you don't want to (or can't) read" and "it's a pronunciation helper."

Pop-ups were also used to present additional information related to healthy foods and the characters. Nearly all children liked the pop-ups, describing them as "interesting," informative," "surprising," "cool," and "pretty funny." Some children shared that the information pop-ups taught them something they never knew. A male child liked the information pop-ups "because they're telling stuff that I didn't really know about; stuff I didn't know about carrots, I know now that's why I like it." Some of the reasons shared for liking the character pop-ups included that they helped with developing a deeper understanding of the characters and their role in the story. As 1 male child shared, he "liked [them] because they tell [me] where are they from, what do they like...basically their ideals. It's like their ID."

Goal Setting

Most children in both groups reported selecting weekly goals both at T2 and T3 time points and acknowledged the "motivation" and "confidence" generated by the process. Some children who did not set weekly goals cited time constraints and involvement in after-school activities as their primary reasons. Most children found the work around goal selection and achievement to be manageable and also easier if the items needed (such as fruit, vegetables, and water) were already available at home and if parental help was accessible to obtain them. Many children also acknowledged that the reminders to work on goals were helpful.

Reminders

Most children in both arms preferred text message reminders over email reminders, saying that they were enjoyable to receive, helpful, easier to understand, more convenient, and could be checked more often. In relation to dose and frequency of messaging, across both arms, most children reported receiving "enough messages," with only a few concluding that they

XSL•FO
RenderX

received too many, citing an annoyance at being interrupted or inundated, and preferring a frequency of 2 messages per week.

Outcome Measures (Secondary Measures)

Dietary Beliefs and Attitudes

At T3, both experimental and comparison groups reported a substantial and consistent increase in OE, self-efficacy, and BI related to fruit and vegetable consumption. In terms of water consumption, both groups reported a substantial increase in BI. The experimental group also reported an increase in self-efficacy for water intake, which was maintained at T4, and the comparison group showed an increase in attitudes toward water consumption. Changes related to OE, self-efficacy, BI, and attitudes toward sugary drinks and junk food consumption were generally small and mixed. In summary, the within-group effect sizes for child dietary beliefs and attitudes were generally larger in the experimental group at T3; however, the majority diminished by T4. The between-group effect sizes, while favoring the experimental group, were in the small range. Refer to Table 5 for the change scores for all psychosocial variables for the experimental and comparison groups and Figure 4 for the effect sizes comparing the experimental and comparison groups. Descriptive statistics by group and child gender are presented in Multimedia Appendix 1.

Table 5. Change scores for psychosocial variables related to dietary behaviors from T1[a] to T3[b] and T1 to T4[c], by group.

	Outcome expectations				Self-efficacy				Behavioral intention				Attitudes			
	E[d] (n=37), mean change (SD)	C[e] (n=39), mean change (SD)	ES[f,g] within E	ES within C	E (n=37), mean change (SD)	C (n=39), mean change (SD)	ES within E	ES within C	E (n=37), mean change (SD)	C (n=39), mean change (SD)	ES within E	ES within C	E (n=37), mean change (SD)	C (n=39), mean change (SD)	ES within E	ES within C
Fruit and 100% fruit juice																
T1 to T3	2.17 (3.43)	2.05 (3.70)	0.63	0.55	0.64 (2.41)	0.59 (1.91)	0.27	0.31	0.45 (1.80)	0.61 (1.61)	0.25	0.38	0.45 (2.49)	–0.13 (2.49)	0.18	–0.05
T1 to T4	1.00 (3.34)	–0.08 (3.79)	0.30	–0.02	–0.22 (2.78)	–0.74 (2.49)	–0.08	–0.30	0.22 (2.00)	–0.10 (2.19)	0.11	–0.05	–0.36 (2.77)	–0.69 (2.27)	–0.13	–0.30
Vegetable																
T1 to T3	1.69 (3.21)	1.29 (3.76)	0.53	0.34	1.50 (3.09)	1.02 (2.68)	0.48	0.38	0.60 (2.40)	0.80 (1.96)	0.25	0.41	0.98 (3.20)	0.54 (3.51)	0.30	0.15
T1 to T4	0.16 (3.96)	0.03 (4.63)	0.04	0.01	0.65 (3.51)	–0.41 (3.30)	0.19	–0.12	0 (1.90)	–0.08 (2.16)	0	–0.04	0.50 (3.01)	0.38 (3.48)	0.17	0.11
Water																
T1 to T3	1.00 (3.56)	0.78 (5.00)	0.28	0.16	0.31 (1.00)	0.12 (0.95)	0.31	0.13	0.52 (1.80)	0.93 (2.54)	0.29	0.36	0.50 (2.24)	1.17 (3.19)	0.22	0.37
T1 to T4	0.57 (3.09)	–0.67 (3.91)	0.18	–0.17	0.32 (0.94)	0 (0.83)	0.34	0	–0.16 (1.28)	0.62 (2.62)	–0.13	0.23	–0.33 (2.07)	0.56 (3.26)	–0.16	0.17
Sugary drinks																
T1 to T3	–0.10 (4.01)	–0.07 (3.29)	–0.02	–0.02	0.21 (2.09)	–0.30 (1.73)	0.10	–0.17	0.55 (2.44)	–0.37 (3.02)	0.22	–0.12	0.10 (3.14)	–1.44 (3.24)	0.03	–0.44
T1 to T4	–0.22 (3.31)	–0.38 (2.98)	–0.07	–0.13	–0.19 (2.01)	0.08 (2.07)	–0.10	0.04	0.11 (2.17)	0.49 (2.79)	0.05	0.17	–0.27 (2.93)	–0.51 (3.11)	–0.09	0.16
Junk food																
T1 to T3	0.26 (2.56)	0.32 (3.09)	0.10	0.10	–0.14 (2.42)	0.46 (1.64)	–0.06	0.28	0.40 (2.58)	0.55 (3.31)	0.16	0.17	1.17 (3.01)	–0.63 (2.84)	0.39	–0.22
T1 to T4	0.43 (2.24)	0.41 (2.60)	0.19	0.16	–0.24 (2.02)	0.18 (1.35)	–0.12	0.13	0.06 (2.62)	1.21 (2.56)	0.02	0.47	0.41 (3.13)	0.77 (2.16)	0.13	0.36

[a]T1: baseline.

[b]T3: intervention end or 6 weeks postbaseline.

[c]T4: 3-month follow-up postintervention.

[d]E: experimental group.

[e]C: comparison group.

[f]ES: effect size.

[g]"ES within" represents changes in score divided by the SD of change within the condition (E and C, respectively); positive ES indicates improvement, except for junk food and sugary drinks, where negative ES indicates improvement.

XSL•FO
RenderX

Figure 4. Effect sizes at intervention end or 6 weeks after baseline (T3) and 3-month follow-up after the intervention (T4), comparing experimental (E) and comparison (C) groups on psychosocial variables and self-reported diet-related behaviors. Positive effect size indicates a larger improvement in the E compared to the C group.

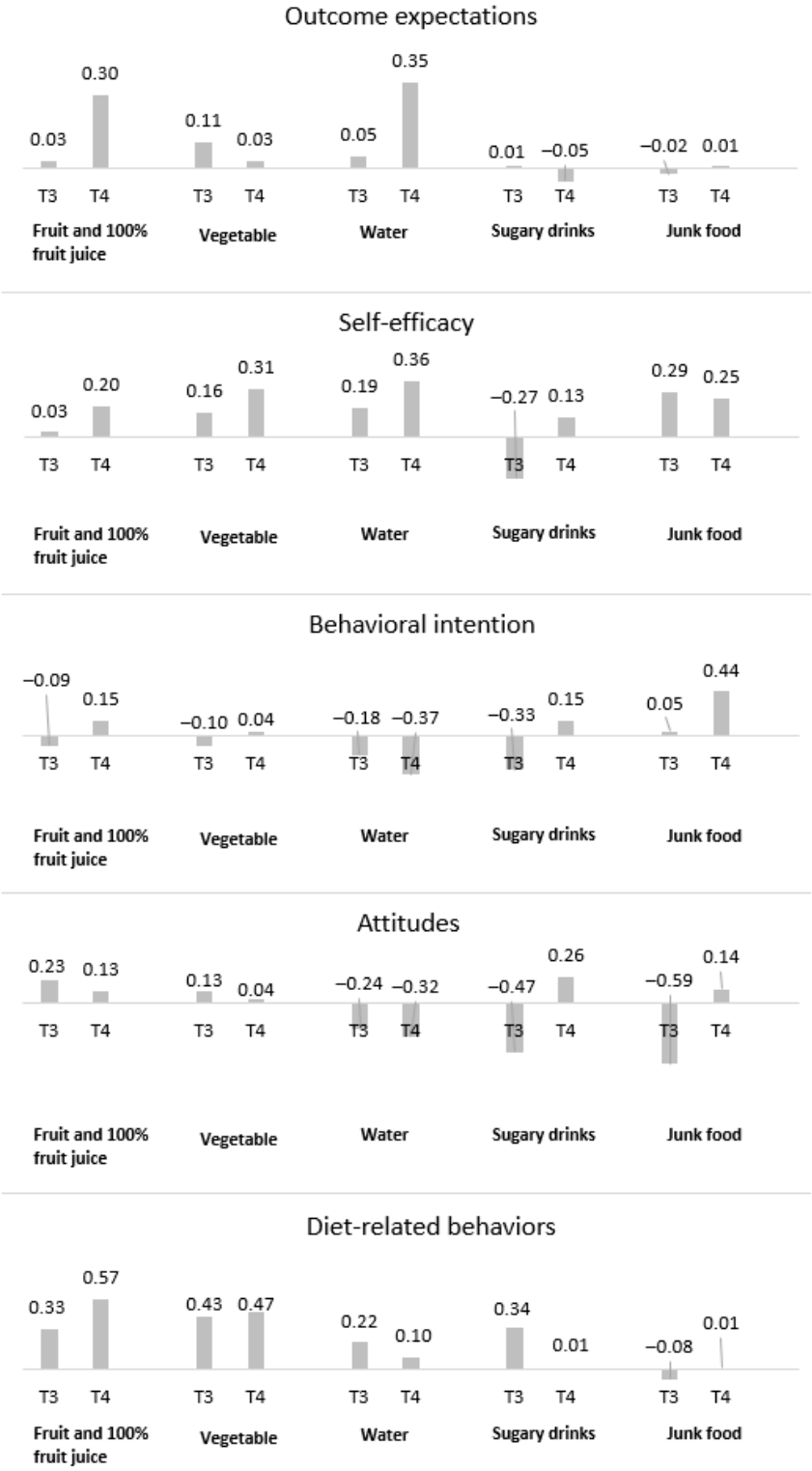

Dietary Behaviors

The experimental group reported a substantial increase in fruit, vegetable, and water intake, which were maintained at T4. Changes in the comparison group were small. Both groups reported a decrease in junk food intake at T3, and the experimental group additionally reported a decrease in sugary drinks intake from T1 to T3; however, these improvements were diminished at T4. The between-group effect sizes for the child dietary intake, favoring the experimental group, were medium to large. Refer to Table 6 for the changes in self-reported diet-related behaviors and Figure 4 for the related effect sizes at T3 and T4 comparing the experimental and comparison groups. Descriptive statistics by child gender are presented in Multimedia Appendix 1.

XSL•FO
RenderX

Table 6. Change in self-reported diet-related behaviors from T1[a] to T3[b] and T1 to T4[c], by group.

	E[d] (n=37), mean change (SD)	C[e] (n=39), mean change (SD)	ES[f,g] within E	ES within C
Fruit and 100% fruit juice				
T1 to T3	0.71 (5.12)	–0.80 (4.04)	0.14	–0.20
T1 to T4	1.30 (3.39)	–0.79 (3.71)	0.38	–0.21
Vegetable				
T1 to T3	3.10 (7.27)	0.22 (5.82)	0.43	0.04
T1 to T4	3.08 (4.58)	0.51 (5.99)	0.67	0.09
Water				
T1 to T3	2.52 (6.85)	1.00 (6.76)	0.37	0.15
T1 to T4	2.22 (4.84)	1.64 (6.41)	0.46	0.26
Sugary drinks				
T1 to T3	–1.80 (5.31)	0 (5.34)	–0.34	0
T1 to T4	–0.03 (7.06)	0.05 (5.50)	0	0.01
Junk food				
T1 to T3	–1.55 (4.77)	–1.95 (5.63)	–0.32	–0.35
T1 to T4	–1.08 (6.40)	–1.03 (4.91)	–0.17	–0.21

[a]T1: baseline.

[b]T3: intervention end or 6 weeks postbaseline.

[c]T4: 3-month follow-up postintervention.

[d]E: experimental group.

[e]C: comparison group.

[f]ES: effect size.

[g]"ES within" represents changes in score divided by the SD of change within the condition (E and C); positive ES indicates improvement, except for junk food and sugary drinks, where negative ES indicates improvement.

BMI-for-Age Percentile

The mean change between T1 and T4 in BMI-for-age percentile was small and comparable in both experimental and comparison groups (mean –0.58, SD 5.91 and mean –0.50, SD 11.49, respectively). A post hoc exploration by child's sex showed that BMI-for-age percentile tended to decrease in the experimental group for male participants (mean change –3.26, SD 4.67) and the comparison group for female participants (mean change –3.30, SD 9.77). Refer to Table 7 for BMI-for-age percentile for the experimental and comparison groups, overall and by sex, at T1 and T4.

Table 7. BMI-for-age percentile at T1[a] and T4[b] for E[c] and C[d] groups, overall and among male and female participants.

	Overall		Male participants		Female participants	
	E (n=36)	C (n=36)	E (n=14)	C (n=15)	E (n=22)	C (n=21)
T1, mean (SD)	74.94 (26.01)	70.29 (27.29)	73.74 (29.53)	70.13 (28.82)	75.67 (24.29)	70.40 (26.84)
T4, mean (SD)	73.70 (26.77)	70.57 (29.44)	70.48 (32.12)	73.55 (31.07)	75.76 (23.34)	66.44 (28.81)
T1 to T4, mean change (SD)	–0.58 (5.91)	–0.50 (11.41)	–3.26 (4.67)	3.42 (12.68)	–1.12 (6.09)	–3.30 (9.77)

[a]T1: baseline.

[b]T4: 3-month follow-up postintervention.

[c]E: experimental group.

[d]C: comparison group.

Discussion

Principal Findings

The primary purpose of this study was to assess the feasibility and participant acceptability of an interactive, web-based comic tool. In addition, this study explored the tool's potential impact on dietary behaviors, related psychosocial variables, and BMI-for-age percentile.

There was a decline in use of the digital tool (in both groups) over the 6-week intervention period. While direct comparisons are difficult due to the lack of standardization in how engagement has been defined and measured in previous research

XSL•FO
RenderX

[44], low to moderate engagement and a decline in use as an intervention progresses are commonly reported [45-47]. This is concerning given that engagement is critical to the effectiveness of digital health interventions [48]. This was a key consideration during the design process for our tool as we used user-centered design principles throughout the development process [49], with the goal of ultimately developing a culturally relevant and meaningful tool. To enhance potential engagement, we incorporated interactive features into the tool, such as tap or click icons that opened pop-up windows (highlighting health facts, food-related fun facts, or character information, such as a character's favorite healthy food and beverage).

While engagement did decline in both groups, the use data appear to support the feasibility and acceptability of the comic tool as children in the experimental group were more engaged; an average of 72% (33/45) of the experimental group accessed their tool weekly compared to 60% (27/44) of the comparison group. Furthermore, most of the children in the experimental group, who accessed the comic, actually read the complete chapter (which ranged from 18 to 34 pages) each week. In addition to a higher level of engagement, the experimental group also reported higher usability of their tool, which is an indicator of tool acceptability and satisfaction [50,51].

Other interactive features included embedded audio recordings for certain character dialogue, which also presented opportunities to deliver health messages in another format. Such a format can help with the understanding of the health messages promoted within the comic [52], particularly as low literacy levels continue to exist in our intended population [53-55]. Research has found that Black and Hispanic students begin high school with literacy skills 3 years behind those of White and Asian students [53]. In addition, students from families with low incomes enter high school with literacy skills that are 5 years behind those of students from families with high incomes [55]. Feedback from participants acknowledged the benefits of the embedded interactive features as 1 child noted that the audio effects allow the tool to read for the children, "if you don't want to or can't read." Such features can also add to the immersion of the reader into the story, as studies have found that audio features can improve the manga reading experience provided it connects the reader with the objects represented in the scene [56]. This could lead to enhanced persuasion of the story's health messages and, thus, motivate positive behavior change, as suggested by the Narrative Transportation Theory [57,58].

While several BCTs were incorporated into the digital comic tool, a key component of the tool was the goal selection and assessment activity. Research has shown that the BCT of goal setting is associated with positive health outcomes, although primarily related to BMI measures in children above a healthy weight [59]. Children in our study acknowledged that goal achievement was easier if the appropriate support was in place, which included fruit, vegetables, and water being available at home and if parental help was accessible. This highlights the importance of ensuring that healthy foods are accessible to allow the children to be empowered to make positive choices. It also reinforces the importance of engaging parents to not only be positive role models but to also provide a supportive home environment as parent-directed interventions can improve the self-efficacy of their children [60]. Prompts and cues were also included in the comic tool in the form of 4 reminder messages each week from the characters themselves. The frequency of the reminder messages was informed by formative research with our intended population, as the literature is quite limited in this area [61]. While a few children did note that they received too many, overall, the children acknowledged that the reminder function, comprising 4 texting or email messages each week was "enough" and an important feature to promote sustained engagement [61].

Secondary outcomes of this study included changes in psychosocial variables related to specific diet-related behaviors, dietary intake, and BMI-for-age percentile. Results showed that while changes in psychosocial variables for certain diet-related behaviors were observed in both the experimental and comparison groups from T1 to T3, most improvements were diminished by T4. However, improvements in intake related to fruit, vegetable, and water behaviors observed in the experimental group were maintained at T4. Furthermore, while no differences were observed in BMI-for-age percentile change between both groups from T1 to T4, a post hoc analysis found that BMI percentile tended to decrease in the experimental group for male participants and in the comparison group for female participants. The improvement in fruit, vegetable, and water intake that was maintained through the T4 period in the experimental group may partially explain the anthropometric findings observed in the male participants, as such behaviors are recommended to reduce obesity risk [62,63]. Further research is warranted to better understand the mechanisms that may have resulted in the observed trend of the decreased BMI percentile in both groups, particularly as no sustained changes were found in related psychosocial variables and no significant improvements in dietary behaviors were observed in the comparison group.

Sex differences in the BMI percentile outcome are interesting to note. Health-promoting comics have been used effectively to reach children of both genders [41,64,65]; however, in our study, it appears that male participants may have responded more positively to the comic tool. While extensive formative research was conducted with both genders from our intended population, with the goal of developing a broad-reaching, meaningful, and engaging narrative, the main character of the story was male. This may have led to unintended differences in how male and female participants responded to the health messages. Interestingly, prior research by the team found that both male and female children often selected a male character when asked who their favorite comic character was [42]; however, other research has shown that gender does play a role in how people relate and respond to media characters [43,66]. Furthermore, while we aimed to address a broad range of genres in the comic to appeal to both genders, differences in genre preferences do exist between the sexes as females tend to prefer romance, fantasy, and action, while males tend to prefer humor and action when reading comics [67].

Implications for Future Research

Further research is needed to not only understand the elements that influence tool engagement but to also gain a better

XSL•FO
RenderX

understanding of the comic tool's key components that influenced the observed health-related outcomes. This pilot and feasibility study suggests that such a comic tool could be an appealing format to promote positive beliefs and behaviors related to a healthier diet in racial and ethnic minority preadolescents (particularly in male children). In addition, our study had successful recruitment and high retention with our participants [23]. Furthermore, 81% (61/75) of the children reported at T4 that they were very satisfied or extremely satisfied with their interactions with the study staff [23]. Therefore, a full-scale RCT with a longer follow-up period is warranted to determine the effectiveness of Intervention INC. However, given potential gender differences in character relatability and genre preferences, the development of future health-promoting comic tools should more carefully consider how to frame storylines and messaging to effectively impact both genders. Moreover, as newly adopted behaviors often diminish once an intervention ends [68], a maintenance intervention should be considered following the initial intervention period, as it could provide critical continued support and reinforce newly adopted behaviors. In addition, this tool's messages mainly focused on obesity-related dietary behaviors. As physical activity is a key contributing factor to reducing obesity risk, the inclusion of a physical activity component should be considered to produce a more comprehensive tool focused on additional influences related to childhood obesity.

Limitations

It is important to acknowledge that this study is not without its limitations. We were unable to capture more granular use data, such as the minutes participants used the tool itself. Self-report data are always a challenge, in particular with dietary intake data. This limitation is further exacerbated by the fact that our questionnaires, while informed by validated questionnaires commonly used in the literature [25-30], were not validated tools and were completed by the children themselves. However, it should be acknowledged that T1 and T4 data were collected in person where a research team member was present and available to answer any questions and assist the children in completing the questionnaire. Seasonality issues may be a concern with the outcome measures, as recruitment was conducted on a rolling basis over a 3-month period. Thus, some participants completed the study during the fall season, while others completed it during the spring season. Another limitation is the small sample size, thus limiting statistical power. The generalizability of these findings is also limited as the data are unique to racial and ethnic minority preadolescent children residing in underserved NYC communities. In addition, all study participants had to have ongoing access to the internet, further limiting the study's generalizability.

Conclusions

This pilot and feasibility study suggests that the multimedia platform of comics could be an appealing and engaging format to promote healthy diet-related beliefs and behaviors in racial and ethnic minority, urban preadolescents at risk of childhood obesity. Furthermore, as comics are popular among US youth and can be easily disseminated, particularly in digital format, the graphics and minimal text make it a promising format for low-literacy populations. The sex differences observed in BMI percentile suggest that boys may respond more positively to health messages delivered in such a format; however, the comic's main character (who is male) may have been more relatable to our male participants. Thus, future iterations of the comic should carefully consider the main character to ensure they are relatable to all genders. While the results are promising, further research is needed, including a full-scale RCT, to determine the effectiveness of Intervention INC.

Acknowledgments

This research has been supported by the Agency for Healthcare Research and Quality (grant R21H5024117). The content is solely the responsibility of the authors and does not necessarily represent the official views of the Agency for Healthcare Research and Quality. The authors wish to acknowledge the research team as well as their community partners for their contribution to this study. The authors want to acknowledge Kris Hoyt (lead developer of the intervention platform and interface) and the research assistants who helped to implement this study.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

MML conceived the intervention concept and study design. KFM drafted the study protocol, with critical inputs from MML. KFM coordinated the trial, with the assistance of SV. KFM, LH, and SV were involved in the recruitment process. KW and LH contributed to the statistical analysis and interpretation, and MD contributed to the qualitative data analysis and interpretation. MML led the writing of the manuscript, with KW, KFM, and MD assisting with the process. All authors read and approved the final manuscript.

Conflicts of Interest

MML, KFM, and SV were involved in the development of the Intervention INC (Interactive Nutrition Comics for Urban, Minority Preadolescents) tool. The authors declare that they have no other conflicts of interest related to this study.

XSL•FO
RenderX

Multimedia Appendix 1
Descriptive statistics for psychosocial variables and self-reported diet related behaviors across timepoints, by group and by sex.
[DOCX File , 425 KB - formative_v9i1e58460_app1.docx]

Multimedia Appendix 2
CONSORT-eHEALTH checklist (V 1.6.2).
[PDF File (Adobe PDF File), 95 KB - formative_v9i1e58460_app2.pdf]

References

1. Hales CM, Carroll MD, Fryar CD, Ogden CL. Prevalence of obesity among adults and youth: United States, 2015-2016. NCHS Data Brief 2017 Oct(288):1-8 [FREE Full text] [Medline: 29155689]
2. Ogden CL, Martin CB, Freedman DS, Hales CM. Trends in obesity disparities during childhood. Pediatrics 2022 Aug 01;150(2):e2022056547 [FREE Full text] [doi: 10.1542/peds.2022-056547] [Medline: 35788862]
3. Hales CM, Fryar CD, Carroll MD, Freedman DS, Ogden CL. Trends in obesity and severe obesity prevalence in US youth and adults by sex and age, 2007-2008 to 2015-2016. JAMA 2018 Apr 24;319(16):1723-1725 [FREE Full text] [doi: 10.1001/jama.2018.3060] [Medline: 29570750]
4. Skinner AC, Ravanbakht SN, Skelton JA, Perrin EM, Armstrong SC. Prevalence of obesity and severe obesity in US children, 1999-2016. Pediatrics 2018 Mar;141(3):1-18 [FREE Full text] [doi: 10.1542/peds.2017-3459] [Medline: 29483202]
5. Harbin MM, Hultgren NE, Kelly AS, Dengel DR, Evanoff NG, Ryder JR. Measurement of central aortic blood pressure in youth: role of obesity and sex. Am J Hypertens 2018 Nov 13;31(12):1286-1292 [FREE Full text] [doi: 10.1093/ajh/hpy128] [Medline: 30107492]
6. Tate EB, Spruijt-Metz D, O'Reilly G, Jordan-Marsh M, Gotsis M, Pentz MA, et al. mHealth approaches to child obesity prevention: successes, unique challenges, and next directions. Transl Behav Med 2013 Dec;3(4):406-415 [FREE Full text] [doi: 10.1007/s13142-013-0222-3] [Medline: 24294329]
7. Horesh A, Tsur AM, Bardugo A, Twig G. Adolescent and childhood obesity and excess morbidity and mortality in young adulthood-a systematic review. Curr Obes Rep 2021 Sep 05;10(3):301-310. [doi: 10.1007/s13679-021-00439-9] [Medline: 33950400]
8. Skinner AC, Perrin EM, Moss LA, Skelton JA. Cardiometabolic risks and severity of obesity in children and young adults. N Engl J Med 2015 Oct;373(14):1307-1317. [doi: 10.1056/NEJMoa1502821] [Medline: 26422721]
9. Antwi FA, Fazylova N, Garcon MC, Lopez L, Rubiano R, Slyer JT. Effectiveness of web-based programs on the reduction of childhood obesity in school-aged children: a systematic review. JBI Database Syst Rev Implement Rep 2013 Jun;11(6):1-44. [doi: 10.11124/jbisrir-2013-459]
10. Reilly JJ, Kelly J. Long-term impact of overweight and obesity in childhood and adolescence on morbidity and premature mortality in adulthood: systematic review. Int J Obes (Lond) 2011 Jul;35(7):891-898. [doi: 10.1038/ijo.2010.222] [Medline: 20975725]
11. Costa CS, Del-Ponte B, Assunção MC, Santos IS. Consumption of ultra-processed foods and body fat during childhood and adolescence: a systematic review. Public Health Nutr 2018 Jan;21(1):148-159 [FREE Full text] [doi: 10.1017/S1368980017001331] [Medline: 28676132]
12. Campbell MK. Biological, environmental, and social influences on childhood obesity. Pediatr Res 2016 Jan 20;79(1-2):205-211. [doi: 10.1038/pr.2015.208] [Medline: 26484623]
13. SM M. Childhood obesity: epidemiology, determinants, and prevention. J Nutr Disorders Ther 2015;5(2). [doi: 10.4172/2161-0509.1000156]
14. The common sense census: media use by tweens and teens, 2021. Common Sense. 2022 Mar 9. URL: https://www.commonsensemedia.org/research/the-common-sense-census-media-use-by-tweens-and-teens-2021 [accessed 2023-04-22]
15. Delamater AM, Pulgaron ER, Rarback S, Hernandez J, Carrillo A, Christiansen S, et al. Web-based family intervention for overweight children: a pilot study. Child Obes 2013 Feb;9(1):57-63 [FREE Full text] [doi: 10.1089/chi.2011.0126] [Medline: 23308372]
16. Kracht CL, Hutchesson M, Ahmed M, Müller AM, Ashton LM, Brown HM, et al. E- and mHealth interventions targeting nutrition, physical activity, sedentary behavior, and/or obesity among children: a scoping review of systematic reviews and meta-analyses. Obes Rev 2021 Dec;22(12):e13331 [FREE Full text] [doi: 10.1111/obr.13331] [Medline: 34476890]
17. Nguyen B, Kornman KP, Baur LA. A review of electronic interventions for prevention and treatment of overweight and obesity in young people. Obes Rev 2011 May;12(5):e298-e314. [doi: 10.1111/j.1467-789X.2010.00830.x] [Medline: 21348921]
18. Demory-Luce D, Morales M, Nicklas T, Baranowski T, Zakeri I, Berenson G. Changes in food group consumption patterns from childhood to young adulthood: the Bogalusa Heart Study. J Am Diet Assoc 2004 Nov;104(11):1684-1691. [doi: 10.1016/j.jada.2004.07.026] [Medline: 15499355]

XSL•FO
RenderX

19. Mikkilä V, Räsänen L, Raitakari OT, Pietinen P, Viikari J. Consistent dietary patterns identified from childhood to adulthood: the cardiovascular risk in Young Finns Study. Br J Nutr 2005 Jun 08;93(6):923-931. [doi: 10.1079/bjn20051418] [Medline: 16022763]
20. Kelder SH, Perry CL, Klepp KI, Lytle LL. Longitudinal tracking of adolescent smoking, physical activity, and food choice behaviors. Am J Public Health 1994 Jul;84(7):1121-1126. [doi: 10.2105/ajph.84.7.1121] [Medline: 8017536]
21. Golan M, Crow S. Parents are key players in the prevention and treatment of weight-related problems. Nutr Rev 2004 Jan;62(1):39-50. [doi: 10.1111/j.1753-4887.2004.tb00005.x] [Medline: 14995056]
22. Leung MM, Mateo KF, Verdaguer S, Wyka K. Testing a web-based interactive comic tool to decrease obesity risk among minority preadolescents: protocol for a pilot randomized control trial. JMIR Res Protoc 2018 Nov 09;7(11):e10682 [FREE Full text] [doi: 10.2196/10682] [Medline: 30413399]
23. DeFrank G, Singh S, Mateo KF, Harrison L, Rosenthal A, Gorman A, et al. Key recruitment and retention strategies for a pilot web-based intervention to decrease obesity risk among minority youth. Pilot Feasibility Stud 2019 Sep 05;5(1):109 [FREE Full text] [doi: 10.1186/s40814-019-0492-8] [Medline: 31516726]
24. Champion VL, Skinner CS. The health belief model. In: Glanz K, Rimer BK, Viswanath K, editors. Health Behavior and Health Education: Theory, Research, and Practice. Hoboken, NJ: Jossey-Bass Publishers; 2008:45-65.
25. McAlister AL, Perry CL, Parcel GS. How individuals, environments, and health behaviors interact: social cognitive theory. In: Glanz K, Rimer BK, Viswanath K, editors. Health Behavior and Health Education: Theory, Research, and Practice. Hoboken, NJ: Jossey-Bass Publishers; 2008:169-188.
26. Green MC, Clark JL. Transportation into narrative worlds: implications for entertainment media influences on tobacco use. Addiction 2013 Mar;108(3):477-484. [doi: 10.1111/j.1360-0443.2012.04088.x] [Medline: 22994374]
27. Slater MD, Rouner D. Entertainment—education and elaboration likelihood: understanding the processing of narrative persuasion. Commun Theory 2002 May;12(2):173-191. [doi: 10.1111/j.1468-2885.2002.tb00265.x]
28. Verdaguer S, Mateo KF, Wyka K, Dennis-Tiwary TA, Leung MM. A web-based interactive tool to reduce childhood obesity risk in urban minority youth: usability testing study. JMIR Form Res 2018 Nov 01;2(2):e21 [FREE Full text] [doi: 10.2196/formative.9747] [Medline: 30684417]
29. Brooke J. SUS: a 'quick and dirty' usability scale. In: Usability Evaluation in Industry. Hoboken, NJ: CRC Press; 1996.
30. Lund AM. Measuring usability with the USE Questionnaire PDF. Scribd. 2001 Jan. URL: https://www.scribd.com/document/313132052/Measuring-Usability-with-the-USE-Questionnaire-pdf [accessed 2023-05-26]
31. Ben-Zeev D, Brenner CJ, Begale M, Duffecy J, Mohr DC, Mueser KT. Feasibility, acceptability, and preliminary efficacy of a smartphone intervention for schizophrenia. Schizophr Bull 2014 Nov;40(6):1244-1253 [FREE Full text] [doi: 10.1093/schbul/sbu033] [Medline: 24609454]
32. De Bourdeaudhuij I, Klepp KI, Due P, Rodrigo CP, de Almeida M, Wind M, et al. Reliability and validity of a questionnaire to measure personal, social and environmental correlates of fruit and vegetable intake in 10-11-year-old children in five European countries. Public Health Nutr 2005 Apr 02;8(2):189-200. [doi: 10.1079/phn2004673] [Medline: 15877912]
33. Reynolds KD, Yaroch AL, Franklin FA, Maloy J. Testing mediating variables in a school-based nutrition intervention program. Health Psychol 2002;21(1):51-60. [doi: 10.1037//0278-6133.21.1.51]
34. Youth risk behavior surveillance system (YRBSS). Centers for Disease Control and Prevention. URL: https://www.cdc.gov/healthyyouth/data/yrbs/index.htm [accessed 2023-05-26]
35. Neuhouser ML, Lilley S, Lund A, Johnson DB. Development and validation of a beverage and snack questionnaire for use in evaluation of school nutrition policies. J Am Diet Assoc 2009 Sep;109(9):1587-1592. [doi: 10.1016/j.jada.2009.06.365] [Medline: 19699839]
36. NHANES anthropometry procedures manual. Centers for Disease Control and Prevention. 2021 May. URL: https://wwwn.cdc.gov/nchs/data/nhanes/2021-2023/manuals/2021-Anthropometry-Procedures-Manual-508.pdf [accessed 2023-12-12]
37. Child and teen BMI calculator. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/bmi/child-teen-calculator/index.html [accessed 2023-05-26]
38. Neal JW, Neal ZP, VanDyke E, Kornbluh M. Expediting the analysis of qualitative data in evaluation: a procedure for the rapid identification of themes from audio recordings (RITA). Am J Eval 2014 May 27;36(1):118-132 [FREE Full text] [doi: 10.1177/1098214014536601]
39. Birks M, Chapman Y, Francis K. Memoing in qualitative research: probing data and processes. J Res Nurs 2008 Jan 01;13(1):68-75 [FREE Full text] [doi: 10.1177/1744987107081254]
40. SPSS statistics_29.0.x. IBM. URL: https://www.ibm.com/support/pages/ibm-spss-statistics290x [accessed 2023-05-31]
41. Nascimento GB, de Maio Nascimento M, de Araújo LM, Gouveia É, Ihle A. Comics as a physical education tool for health promotion in Brazilian primary education, based on Paulo Freire's principles of empowerment. Children (Basel) 2023 Sep 19;10(9):1575 [FREE Full text] [doi: 10.3390/children10091575] [Medline: 37761536]
42. Leung MM, Green MC, Cai J, Gaba A, Tate DF, Ammerman A. Fight for your right to fruit: development of a manga comic promoting fruit consumption in youth. Open Nutr J 2015 Feb;9(1):82-90. [doi: 10.2174/1876396001509010082]
43. Steinke J, Applegate B, Lapinski M, Ryan L, Long M. Gender differences in adolescents' wishful identification with scientist characters on television. Sci Commun 2011 Jun 28;34(2):163-199. [doi: 10.1177/1075547011410250]

XSL•FO
RenderX

44. Georgeson AR, Highlander A, Loiselle R, Zachary C, Jones DJ. Engagement in technology-enhanced interventions for children and adolescents: current status and recommendations for moving forward. Clin Psychol Rev 2020 Jun;78:101858 [FREE Full text] [doi: 10.1016/j.cpr.2020.101858] [Medline: 32413679]
45. Wrobel J, Silvasstar J, Peterson R, Sumbundu K, Kelley A, Stephens D, et al. Text messaging intervention for mental wellness in American Indian and Alaska native teens and young adults (BRAVE study): analysis of user engagement patterns. JMIR Form Res 2022 Feb 25;6(2):e32138 [FREE Full text] [doi: 10.2196/32138] [Medline: 35212633]
46. Thornton L, Brown HM, Osman B, Stewart C, Whife J, Champion KE, et al. Factors associated with adolescents' engagement with a Healthy Lifestyles app. Procedia Comput Sci 2022;206:56-67. [doi: 10.1016/j.procs.2022.09.085]
47. Schwarz A, Cardon G, Chastin S, Stragier J, De Marez L, Consortium SmartLife, et al. Does dynamic tailoring of a narrative-driven exergame result in higher user engagement among adolescents? Results from a cluster-randomized controlled trial. Int J Environ Res Public Health 2021 Jul 12;18(14):7444 [FREE Full text] [doi: 10.3390/ijerph18147444] [Medline: 34299895]
48. Thies K, Anderson D, Cramer B. Lack of adoption of a mobile app to support patient self-management of diabetes and hypertension in a federally qualified health center: interview analysis of staff and patients in a failed randomized trial. JMIR Hum Factors 2017 Oct 03;4(4):e24 [FREE Full text] [doi: 10.2196/humanfactors.7709] [Medline: 28974481]
49. Youth-centred digital health interventions: a framework for planning, developing and implementing solutions with and for young people. World Health Organization. 2021 Jul 19. URL: https://www.who.int/publications/i/item/9789240011717 [accessed 2024-12-20]
50. Jake-Schoffman DE, Silfee VJ, Waring ME, Boudreaux ED, Sadasivam RS, Mullen SP, et al. Methods for evaluating the content, usability, and efficacy of commercial mobile health apps. JMIR Mhealth Uhealth 2017 Dec 18;5(12):e190 [FREE Full text] [doi: 10.2196/mhealth.8758] [Medline: 29254914]
51. Stinson J, McGrath P, Hodnett E, Feldman B, Duffy C, Huber A, et al. Usability testing of an online self-management program for adolescents with juvenile idiopathic arthritis. J Med Internet Res 2010 Jul 29;12(3):e30 [FREE Full text] [doi: 10.2196/jmir.1349] [Medline: 20675293]
52. Skouge JR, Rao K, Boisvert PC. Promoting early literacy for diverse learners using audio and video technology. Early Childhood Educ J 2007 May 22;35(1):5-11. [doi: 10.1007/s10643-007-0170-y]
53. Murnane R, Sawhill I, Snow C. Literacy challenges for the twenty-first century: introducing the issue. Future Child 2012 Sep;22(2):3-15. [doi: 10.1353/foc.2012.0013] [Medline: 23057128]
54. Racial/ethnic score gap tool. The Nation's Report Card. URL: https://www.nationsreportcard.gov/dashboards/regression/ [accessed 2023-11-21]
55. Reardon SF, Valentino RA, Shores KA. Patterns of literacy among U.S. students. Future Child 2012 Sep;22(2):17-37. [doi: 10.1353/foc.2012.0015] [Medline: 23057129]
56. Issaev S, Zahiri M, Cao M, LuL. Applying audio cues to increase manga immersion. The University of British Columbia. 2019. URL: https://courses.ece.ubc.ca/518/previous/hit2019F/papers/IssaevZahiriCaoLu2-2019.pdf [accessed 2024-12-20]
57. Green MC. Narratives and cancer communication. J Commun 2006 Aug 04;56(suppl_1):S163-S183. [doi: 10.1111/j.1460-2466.2006.00288.x]
58. Kreuter MW, Green MC, Cappella JN, Slater MD, Wise ME, Storey D, et al. Narrative communication in cancer prevention and control: a framework to guide research and application. Ann Behav Med 2007 Jun;33(3):221-235. [doi: 10.1007/BF02879904] [Medline: 17600449]
59. Family weight management: changing behaviour techniques. United Kingdom government. 2020 Mar 20. URL: https://www.gov.uk/government/publications/family-weight-management-changing-behaviour-techniques [accessed 2023-05-29]
60. Swindle T, Poosala AB, Zeng N, Børsheim E, Andres A, Bellows LL. Digital intervention strategies for increasing physical activity among preschoolers: systematic review. J Med Internet Res 2022 Jan 11;24(1):e28230 [FREE Full text] [doi: 10.2196/28230] [Medline: 35014962]
61. Wang Y, Min J, Khuri J, Xue H, Xie B, A Kaminsky L, et al. Effectiveness of mobile health interventions on diabetes and obesity treatment and management: systematic review of systematic reviews. JMIR Mhealth Uhealth 2020 Apr 28;8(4):e15400 [FREE Full text] [doi: 10.2196/15400] [Medline: 32343253]
62. Hampl SE, Hassink SG, Skinner AC, Armstrong SC, Barlow SE, Bolling CF, et al. Clinical practice guideline for the evaluation and treatment of children and adolescents with obesity. Pediatrics 2023 Feb 01;151(2):e2022060640. [doi: 10.1542/peds.2022-060640] [Medline: 36622115]
63. Hargreaves D, Mates E, Menon P, Alderman H, Devakumar D, Fawzi W, et al. Strategies and interventions for healthy adolescent growth, nutrition, and development. Lancet 2022 Jan 08;399(10320):198-210. [doi: 10.1016/S0140-6736(21)01593-2] [Medline: 34856192]
64. Branscum P, Sharma M, Wang LL, Wilson BR, Rojas-Guyler L. A true challenge for any superhero: an evaluation of a comic book obesity prevention program. Fam Community Health 2013;36(1):63-76. [doi: 10.1097/FCH.0b013e31826d7607] [Medline: 23168347]
65. Leung MM, Tripicchio G, Agaronov A, Hou N. Manga comic influences snack selection in Black and Hispanic New York City youth. J Nutr Educ Behav 2014;46(2):142-147. [doi: 10.1016/j.jneb.2013.11.004] [Medline: 24433817]

XSL•FO
RenderX

66. Hoffner CA. Same-gender characters: appeal and identification. In: The International Encyclopedia of Media Psychology. Hoboken, NJ: John Wiley & Sons; Sep 08, 2020.
67. Madeley JM. Girly girls and pretty boys: gender and audience reception of English-translated manga. Faculty of Arts, University of Regina. 2010 Dec. URL: https://ourspace.uregina.ca/items/be77fbc3-4335-4e0e-aaac-44637e9065f4 [accessed 2024-12-20]
68. Jones A, Armstrong B, Weaver RG, Parker H, von Klinggraeff L, Beets MW. Identifying effective intervention strategies to reduce children's screen time: a systematic review and meta-analysis. Int J Behav Nutr Phys Act 2021 Sep 16;18(1):126-146 [FREE Full text] [doi: 10.1186/s12966-021-01189-6] [Medline: 34530867]

Abbreviations

BCT: behavior change technique
BI: behavioral intention
CONSORT: Consolidated Standards of Reporting Trials
INC: Interactive Nutrition Comics for Urban, Minority Preadolescents
NYC: New York City
OE: outcome expectation
RCT: randomized controlled trial
T1: baseline
T2: intervention midpoint or 3 weeks after baseline
T3: intervention end or 6 weeks after baseline
T4: 3-month follow-up

Edited by A Mavragani; submitted 17.03.24; peer-reviewed by T Baranowski; comments to author 25.06.24; revised version received 18.10.24; accepted 04.11.24; published 15.01.25.

Please cite as:
Leung MM, Mateo KF, Dublin M, Harrison L, Verdaguer S, Wyka K
Testing a Web-Based Interactive Comic Tool to Decrease Obesity Risk Among Racial and Ethnic Minority Preadolescents: Randomized Controlled Trial
JMIR Form Res 2025;9:e58460
URL: https://formative.jmir.org/2025/1/e58460
doi:10.2196/58460
PMID:39813083

©May May Leung, Katrina F Mateo, Marlo Dublin, Laura Harrison, Sandra Verdaguer, Katarzyna Wyka. Originally published in JMIR Formative Research (https://formative.jmir.org), 15.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Exploring Web-Based Support for Suicidal Ideation in the Scottish Population: Usability Study

Heather McClelland[1], BSc, MRes, PhD; Rory C O'Connor[1], BSc, PhD; Laura Gibson[2]; Donald J MacIntyre[3], MD

[1]School of Health and Wellbeing, University of Glasgow, Glasgow, United Kingdom

[2]Police Scotland, Glasgow, United Kingdom

[3]NHS Lothian, Edinburgh, United Kingdom

Corresponding Author:
Heather McClelland, BSc, MRes, PhD
School of Health and Wellbeing
University of Glasgow
90 Byres Road, Clarice Pears Building
Glasgow, G12 8TB
United Kingdom
Phone: 44 141 330 3299
Email: heather.mcclelland@glasgow.ac.uk

Abstract

Background: Suicide is a global health concern. In the United Kingdom, Scotland has the highest suicide rate. Lived experience and suicide prevention stakeholders in Scotland have identified a key gap in suicide prevention activities: the lack of 24-hour peer-driven web-based support for people who are suicidal.

Objective: This usability study aimed to evaluate the feasibility, acceptability, utility, and reach of a suicide prevention website (Surviving Suicidal Thoughts) specifically designed to support residents in Scotland who are experiencing suicidal thoughts themselves or suspect or know someone who is experiencing suicidal thoughts. Intended support was delivered through the provision of personal testimony videos of individuals with lived experience.

Methods: A peer-driven website was developed specifically to support residents of Scotland experiencing suicidal thoughts. The website included resources (eg, videos from lived experience and written guidance about how to respond to someone who may be experiencing suicidal thoughts) to help reduce distress, normalize experiences, and challenge distressing thoughts. The website was promoted via leading web-based social media channels and Google Ads. Evaluation of the website was based on website engagement, marketing strategy, and direct web user feedback via a cross-sectional survey.

Results: Data were collected for 41 weeks (June 2022 to February 2023) spanning the launch of the website and the conclusion of the second marketing campaign. On average, the website received 99.9 visitors per day. A total of 56% (n=14,439) of visitors were female, ages ranged from younger than 18 years to older than 70 years (commonly between 25 and 34 years) and originated from all regions of Scotland. According to Google Search terms of Scottish residents, of the individuals indicated to be experiencing suicidal thoughts but not looking for help, 5.3% (n=920) engaged with the website compared to 10.5% (n=2898) who were indicated to be looking for help for themselves. Based on participant responses to the evaluation survey (n=101), the website was associated with a significant reduction in suicidal thoughts (P=.03). Reasons for visiting the website varied. Marketing data implied that people were more likely to engage with advertisements, which they felt were more personal, and visitors to the website were more likely to engage with videos, which corresponded to their age.

Conclusions: A peer-led website may help residents of Scotland who are experiencing suicidal thoughts. Web-based interventions may have considerable reach in Scotland both in terms of age and geographic area. Engagement with the website was similar to other self-help websites for suicidal ideation; however, more nuanced methods of analyzing website engagement for help-seeking behavior are recommended. Future work would benefit from exploring the effectiveness of this website based on a larger participant sample with website modifications guided by the principles of social learning theory.

(JMIR Form Res 2025;9:e55932) doi:10.2196/55932

KEYWORDS

suicide prevention; Scotland; suicidal thoughts; digital intervention; internet; self-help; crisis intervention

XSL•FO
RenderX

Introduction

Background

Suicide is a leading global health concern. The World Health Organization [1] estimates that 1 person dies by suicide every 40 seconds. In the United Kingdom, Scotland continues to have the highest suicide rate across its 4 nations [2-4]. Most people with thoughts of suicide do not seek professional help, and for those who do, too often they struggle to access timely care [5]. Therefore, improving timely and easy-to-access care for those experiencing suicidal ideation is a public health priority in Scotland.

In 2022, the Scottish government published a 10-year, suicide prevention strategy that aims to reduce suicide on a national level [6]. The strategy drew on evidence from over 200 Scottish suicide prevention stakeholders (eg, lived experience, mental health professionals, and charity organization professionals) [7], which indicated that digital peer support was among the most sought-after resource needed. Subsequently, the current strategy includes the following goal: "anyone in suicidal crisis should be able to access support remotely by phone, text or online" [6,7].

National strategies for suicide prevention are underpinned by the drive to understand human behavior within a psychosocial context [8]. The integrated motivational-volitional model of suicidal behavior [9,10] is a leading theory that maps the emergence of suicidal thoughts and behaviors. Specifically, it posits that feelings of defeat (in response to stress) can develop into entrapment, which, in the absence of factors such as social support and belongingness, can give rise to suicidal ideation and potentially suicidal behaviors.

However, other theories of human behavior can also be applied to understand suicide risk. Social learning theory [11], for example, posits that people are more likely to model their behavior on those with whom they identify [12]. Indeed, videos of individuals describing their own experience of, and surviving, thoughts of suicide have been shown to be effective in improving the life satisfaction of those who are experiencing suicidal thoughts themselves [13]. What is more, according to group identity theory [14], identifying with others with shared thoughts or experiences can facilitate feelings of belongingness, the absence of which is consistently associated with suicide risk [13,14].

Systematic reviews by both Lai et al [15] and Robinson et al [16] indicate that a range of web-based suicide prevention strategies (including self-guided, digital cognitive behavioral therapy, web chat, and signposting to resources) are effective in reducing thoughts of suicide, including among those in a suicidal crisis. Although the number of studies included in these reviews is small, the evidence also points toward the potential utility of testimonies from others with similar experiences as a helpful suicide prevention strategy. In the United States, NowMattersNow [17]—a website designed specifically for US residents—provides evidence-based resources including videos, blogs, and suicide prevention strategies (eg, emotion self-regulation coping strategies) to promote the self-management of suicidal thoughts. The website is available 24 hours per day and is freely accessible to the public. Visitors to this website were found to stay on the website for an average of 1 minute 31 seconds, with a bounce rate (percentage of visitors who left after viewing 1 page) of 61% [18]. To put this in context, a bounce rate of 60%-90% is average for landing pages. Based on an embedded anonymous web user survey, Whiteside et al [18] found that engagement with the NowMattersNow website was associated with a self-reported reduction in feelings of suicidal ideation and distress. Given the cultural similarities between the United States and Scotland and the perceived demand of lived experience support from people who are suicidal [6,19], such a resource may be effective within the Scottish context.

Aims

This pilot phase usability study is part of a wider, ongoing digital project to reduce suicide across Scotland. The website was co-designed and delivered by individuals with lived experiences of suicidal thoughts and behaviors. The overall aims of the project are listed in Multimedia Appendix 1. To help guide these aims, theory-driven outcomes for this study were determined in advance of website development using a logic model (Multimedia Appendix 2). Given the early stage of the development of this suicide prevention website, this study aimed to address the following short-term intended outcomes of the logic model. Specifically, the development of a suicide prevention website aimed to (1) support the Scottish population when experiencing a suicidal crisis that is codeveloped with the target population, (2) investigate which users of the website feel that they have received a compassionate response, (3) investigate which users experience a reduction in feelings of entrapment during engagement with the website, (4) investigate which users experience a reduction in suicidal thoughts during engagement with the website, and (5) investigate which users of the website find it accessible and relatable.

Methods

Study Design

The website, Surviving Suicidal Thoughts (SST) [20], was developed using an iterative approach involving a multidisciplinary website development team (comprised of digital experts, health professionals, and academic researchers) and direct feedback from the Scottish government's Living and Lived Experience Panel (n=5; Scottish residents with lived experience of suicidal thoughts and behaviors).

Ethical Considerations

Research permission was granted by the local National Health Service (NHS) Research and Development Committee in 2022. A Data Protection Impact Assessment was completed and approved prior to the launch of the web user survey. However, following NHS guidance, ethics approval is not required for service development and evaluation studies such as this (section 3.1 of the UK Policy Framework for Health and Social Care Research) [21]. Informed consent to participating in the anonymized feedback survey was inferred through participants opting in to the survey following the survey invitation appearing

XSL•FO
RenderX

on the web page, and a privacy notice was available on the invitation page of the survey. Anonymous web user metrics and engagement with the website were collected only from users who had consented to allow tracking cookies on their internet-enabled device prior to "landing" on any NHS24 web page. Due to the anonymity of participation, participants could not be compensated.

Data Collection

Overview

Data collection for this pilot phase spanned 41 weeks (see Multimedia Appendix 3 for the timeline). Originally, the timeline was expected to span 6 months; however, due to low participant engagement with the web user survey, data collection was extended to include the second marketing campaign spanning from October 2022 to February 2023. Visitors' engagement with the website was monitored during development with adjustments made based on user behavior (Multimedia Appendix 3). Such behavior was monitored before and during visiting the website, which allowed SST visitors to be categorized into 1 of 5 user groups (defined in Table 1). The way in which the 5 visitor groups were expected to navigate to, and use, the SST website is depicted in Multimedia Appendix 4. Further details of the development and evaluation of the website are summarized under the following subheadings: "Website Content," "Web User Evaluation Survey," and "Marketing."

Table 1. Surviving Suicidal Thoughts website visitor groups[a].

User type	Behavior	Search term examples
1. Web-based suicide methods seeking	Individuals not looking for support for their suicidal thoughts	"Want to die" "End my life"
2. Web-based support seeker	Individuals actively looking for support for their suicidal thoughts	"Suicide support"
3. Told friends or family	Individuals who are looking for support for a friend or relative who they know is experiencing suicidal thoughts	"Suicidal relative," "Suicidal friend"
4. Suspecting friends or family	Individuals who are looking for support for a friend or relative who they suspect may be experiencing suicidal thoughts; however, this has not been confirmed	"Is my friend suicidal?"
5. Generic curious or professional	Individuals not help-seeking for themselves or others and not in suicidal crisis	"Suicide prevention day"

[a]The 5 user groups identified to visit the Surviving Suicidal Thoughts website between June 2022 and February 2023 based on search terms entered into Google by Scottish residents.

Website Content

SST was hosted on the NHS Inform internet platform, Scotland's national health care website (Figure 1). The website included personal testimony videos of 5 people (Neil, Steph, Denise, Linda, and Kirsty) with lived experience of suicide and 1 person bereaved by suicide (Jenn). These videos were developed specifically for the website. In these videos, the speakers shared their broad experiences of low mood and suicidal thoughts and behaviors, rather than discussing specific details related to their exact experiences of suicidal crisis. Each video was cut into between 5 and 10 chapters, with each chapter being approximately 90 seconds each. The chapters were titled, which allowed visitors to jump between individual sections. Early in the first marketing campaign, it was observed that visitors of the website only explored the top 2 videos that appeared on their computer or phone screen. In response, the videos were rotated each month to minimize any order effects.

The website also featured emergency service contact information, signposting to third-sector mental health organizations, and how to respond to someone who discloses or may be experiencing suicidal thoughts. To directly evaluate the experiences of web users, the website included a pop-up survey modeled on NowMattersNow [17]. Engagement with the website was captured by the intervention delivery team at NHS Inform using Google Analytics [22].

XSL•FO
RenderX

Figure 1. Image of the Surviving Suicidal Thoughts landing page.

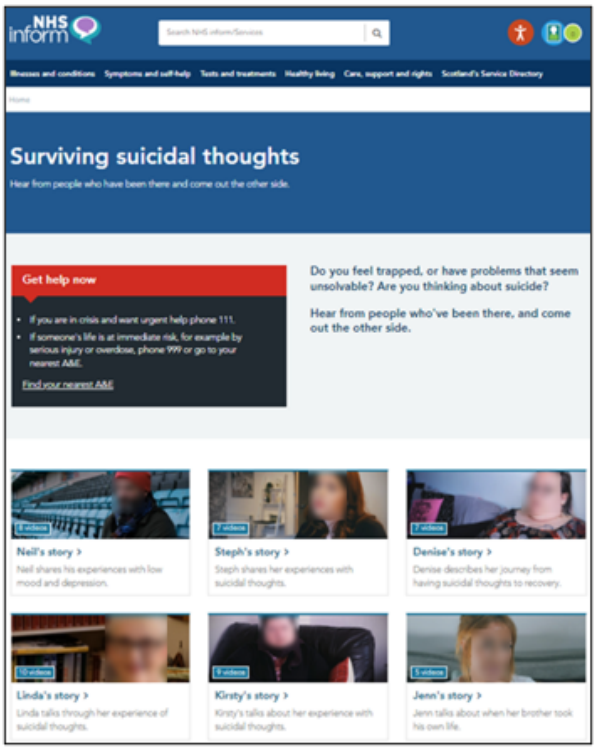

Web User Evaluation Survey

The SST survey was developed using the HotJar survey platform [23]. Initial survey content and methodology were collated by HM (study author) and presented to a focus group of individuals with lived experience of suicidal thoughts and behaviors for their feedback (see Multimedia Appendix 5 for focus group feedback). To minimize overburdening SST website visitors, the Lived Experience Panel recommended reducing the survey items from 12 to 4 questions. As a result, it was intended that a series of "blocks" of survey items would be displayed during the pilot phase to capture information on a range of demographic and psychological factors over time. However, due to the limited engagement with the survey (see Overview in the Results section), it was decided that only the first block of questions would be used in the pilot phase to improve statistical power for analysis. The final survey items and measurement scales included in the pilot are presented in Multimedia Appendix 6; however, in brief: 2 items assessed participants' intensity of suicidal ideation (taken from NowMattersNow [17]; range: 1=not at all and 5=completely overwhelming), 1 demographic variable explored participants' age group (10-year increments from "younger than 18 years" to "70+ years"), and 1 item captured geographic location (based on NHS health board). A fifth item (reason for visiting the website) was added midway through survey recruitment. All survey items were asked at a single time point. The suicidal ideation items asked web users to recall the level of intensity of their suicidal thoughts at the time of first landing on the SST website as well as the current intensity of their suicidal ideation at the time of participating in the survey. These survey items were anonymous, and participants could discontinue the survey and return to the SST website at any time.

To be consistent with the format of NowMattersNow [17], the SST survey invitation initially appeared 3 minutes after landing on the SST website. However, after 2 days, it became evident that the average web user duration was less than 3 minutes (with only 4 visitors seeing the invitation); therefore, we adjusted the survey invitation to appear after 90 seconds (Multimedia Appendix 3) to match 50% of user traffic.

Marketing

Digital marketing was used to promote the website resources on social media platforms (ie, Meta [Facebook and Instagram], X [formerly known as Twitter], and Spotify). Live targeting platforms (Nano and Google Ads) were used to advertise the website to relevant web users. Marketing data were collected from social media platforms, from Nano and Spotify by a third party (Republic of Media) using a pay-per-click marketing campaign. Advertisements promoting the website were designed to convey empathy and hope. Following some prosuicide comments from the public, advertisements for the SST website on Meta platforms were discontinued 1 week after launch to protect the public. This funding was reallocated to Google Ads. Additionally, advertising on X was discontinued after 23 weeks (November 27, 2022) of this 41-week pilot phase. This was due to changes in X's ownership and uncertainty around its operating procedures. Thereafter, funding for the promotion of the SST website on X was reallocated to Nano.

Data Analysis

Website Content Engagement

Analysis was modeled on those used by Whiteside et al [18]. Website engagement was inferred using bounce rates (the percentage of web visitors who leave or "bounce" from a web

XSL•FO
RenderX

page without taking follow-up action). Web users' device used, date, time, and journey to the website were also captured. Video engagement (which videos and duration) was provided by Nano.

Web User Evaluation Survey

Data relating to the date, time, and type of internet-enabled device (eg, laptop and mobile phone) were summarized using frequencies and percentages collated by HotJar survey systems. Using G*Power [24], a priori power analysis determined that a minimum of 102 participants were needed for pairwise comparisons between baseline and follow-up data (effect size=0.03; α=.05; power=0.9). All statistical analyses were conducted using SPSS (version 28; IBM Corp). The distribution of the data was inspected visually and then tested using Kolmogorov-Smirnov tests, which indicated that the data were not normally distributed ($P<0.001$). Subsequently, a nonparametric approach was deemed appropriate to analyze the survey responses, thereby avoiding a type I error. Subsequently, a Wilcoxon 1-sampled, 2-tailed *t* test was used to statistically test the change in participants' intensity of suicidal ideation. A statistical subanalysis of other outcomes (eg, age and geographic location) was not possible due to the small group sizes. Participants who did not answer both suicidal ideation questions were not included in the statistical analysis; all remaining missing data were grouped with "prefer not to say."

Marketing

Information captured included engagement with advertisements (click-through rates [CTRs]) and videos (view-through rates) as well as demographic information (sex: male, female, or unknown and age group: 18-24, 25-34, 35-44, 45-54, 55-64, 65+ years, "unknown"). These data were reported in relation to advertising trends commonly observed in similar advertising strategies [25,26]. Additionally, Nano provided geographical information regarding advertising exposure and engagement based on the number of people who saw an advertisement (impressions).

Results

Overview

SST [20] was developed using an iterative approach between the web page development team and Scotland's Living and Lived Experience Panel and remained live throughout the 41-week pilot phase (from June 28, 2022, to February 20, 2023). The SST website received 25,784 unique visitors equating to an average of 99.9 page views per 24-hour day. Daily traffic to the website peaked at 10 PM, with the greatest amount of traffic on Sundays.

Most visitors accessed the website via mobile phone (n=21,402, 83%), followed by desktop computer (n=3068, 11.9%). When compared to commercial metrics [22], social media marketing CTRs for SST were considered "good" on all platforms. For videos, the view-through rate was 80.2% (n=892,943), which was considered "high." Journeys of SST users were most commonly via Google Search (n=6314, 24.3%), followed by Meta (n=4705, 17.1%). For more details of marketing expenses, marketing strategy, and user engagement per platform, please see Multimedia Appendix 7.

More "web-based suicide methods seeking" than "web-based support seeker" user journey members clicked on testimonial advertisements (advertisements that included names) or engaged with advertisements titled "hear from one whose been there." However, sample sizes were too small to explore this statistically.

The bounce rate to the website was 76.9%, suggesting that almost a quarter of website visitors clicked on a subpage of the SST website (eg, a video link). Of the web users who were invited to participate in the evaluation (n=282, 1.1%), 42.2% (n=119) opened the survey link. In total, 84.9% of those who opened the survey answered at least 1 question (Figure 2).

XSL•FO
RenderX

Figure 2. Participant flowchart of participant engagement with the Surviving Suicidal Thoughts website and evaluation survey between July 2022 and February 2023. *Answered at least 1 survey question.

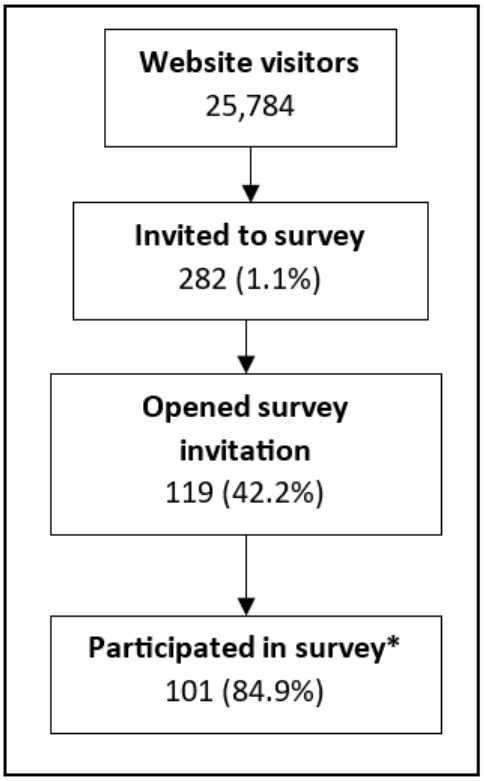

Demographic Characteristics

Google Analytics indicated that 56% (n=14,439) of SST visitors were female, and most "web-based nonsupport seeking" groups were also indicated to be female. Web users who accessed the website through Google Ads were most likely to be aged between 25 and 34 years. In contrast, half (n=47, 49.6%) of the respondents to the evaluation survey were between 40 and 59 years (Table 2). Most (n=69, 66.4%) survey participants chose not to indicate which health board they lived in.

XSL•FO
RenderX

Table 2. Demographic and geographic summary of participant characteristics of the survey evaluation.

Variable	Values, n (%)
Age (years)	
<18	6 (5)
19-29	11 (9.2)
30-39	11 (11.8)
40-49	14 (27.7)
50-59	33 (21.9)
60-69	26 (6.7)
70+	8 (0.8)
Prefer not to say or not applicable	7 (5.9)
Unanswered	13 (10.9)
Health board	
Ayrshire and Arran	5 (4.2)
Borders	1 (0.8)
Fife	2 (1.7)
Fourth Valley	2 (1.7)
Grampian	3 (2.5)
Greater Glasgow and Clyde	7 (5.9)
Highland	1 (0.8)
Lanarkshire	5 (4.2)
Lothian	9 (7.6)
Shetland	1 (0.8)
Tayside	3 (2.5)
Western Isles	1 (0.8)
Prefer not to say or not applicable	10 (8.4)
Unanswered	69 (58)
Reason for visit[a]	
I am looking for resources to help me cope if I feel suicidal again	9 (7.6)
I am thinking about suicide	14 (11.8)
I do not think I can keep myself safe	4 (3.4)
I have a plan for suicide	6 (5)
I am looking for resources to support someone else	8 (6.7)
Other or prefer not to say	11 (9.2)
Unanswered	67 (56.3)
Suicide intensity	
Intensity of suicidal thoughts upon first visiting the site (mean 3.06, SD 1.58)	
1=not at all	30 (25.2)
2	12 (10.1)
3	19 (16)
4	19 (16)
5=completely overwhelming	30 (25.2)
Unanswered	9 (7.6)
Intensity of suicidal thoughts at time of survey ("now"; mean 2.87, SD 1.59)	

XSL•FO
RenderX

Variable	Values, n (%)
1=not at all	33 (27.7)
2	11 (9.2)
3	17 (14.3)
4	16 (13.5)
5=completely overwhelming	24 (20.2)
Unanswered	18 (15.1)

[a]Item added 14 weeks into data collection phase.

Video Engagement

The first 2 videos that appear on a web visitor's browser commonly received the most visitor engagement. "Neil's story: Experiences with low mood and depression" and "Steph's story: Experiences with suicidal thoughts" were the most viewed chapters. On average, web users watched 86.64% (SD 5.91%) of the video chapters. Chapters most likely to be viewed in full were titled "The anxiety underneath was so much to bear" (Kirsty, chapter 3) and "There is light at the end of the tunnel" (Neil, chapter 4).

Suicidal Thoughts

Most (n=5290, 83.8%) CTRs via Google Ads were attributed to the "web-based support seeker group," "told friends or family," or "suspecting friends or family" SST visitor journey groups (Table 3). In contrast, 86.5% of CTRs from Google Search terms related to concern for others experiencing suicidal thoughts (50% CTR for "suicidal relative" and 36.5% CTR for "suicidal friend"), which represent the "told friends or family" and "suspecting friends or family" user journey groups (Table 1). However, 65.5% (n=67) of respondents to the evaluation survey did not indicate their reason for visiting the website (9.2% responded "other or prefer not to say" and 56.3% of participants did not answer).

Table 3. Total number of impressions and clicks to the Surviving Suicidal Thoughts website via adverts on social media and Google.

	Impressions, n	Clicks to website, n	CTR[a] (%)
SST user journey group			
Web-based suicide methods seeking[b]	17,282	920	5.32
Web-based support seeker[c]	27,638	2898	10.49
Told friends or family[d]	14,565	2172	14.91
Suspecting friends or family[e]	1730	220	12.72
Generic curious or professional[f]	1496	104	6.95
Media platform			
X	1,209,432	12,847[g]	1.06
Nano	1,124,972	3471	0.31
Spotify	326,359	__[h]	0.26
Facebook	202,326	4705	2.33
Total	2,925,800	22,637	N/A[i]

[a]CTR: click-through rate.
[b]Individuals not looking for support for their suicidal thoughts.
[c]Individuals actively looking for support to relieve their suicidal thoughts.
[d]Friend or relative who they know to be experiencing suicidal thoughts.
[e]Friend or relative who they suspect may be experiencing suicidal thoughts.
[f]Those looking for general information on suicide prevention who does not otherwise fit in the other web user groups.
[g]Based on engagement with social media post, not clicks to website.
[h]Data not available as these were audio advertisements; therefore, clicking to the website was not possible.
[i]N/A: not applicable.

Based on survey responses, 25.2% (n=30) of visitors retrospectively described the intensity of their suicidal thoughts when first visiting the website as "completely overwhelming" compared to 20.1% (n=24) of participants at the time of survey

XSL•FO
RenderX

engagement (Table 2). Overall, there was a significant reduction in the intensity of suicidal thoughts from the recollection of intensity when first landing on the web page to participating in the survey (t_{100}=–1.89; P=.03). No statistically significant differences were observed between the intensity of suicidal thoughts as a function of age (≤49 years vs ≥50 years).

Discussion

This study aimed to evaluate a pilot web-based self-help resource to help support residents of Scotland who have suicidal ideation. The findings of this study are discussed below in relation to the 5 short-term intended outcomes of the logic model (Multimedia Appendix 2).

Principal Findings

The design and content of SST [20] was developed through consultation with the Living and Lived Experience Panel in Scotland. This input included direct contributions to the website videos—the defining feature of the SST website. Together with other stakeholders (digital specialists and health professionals), this led to a suicide prevention website launched in June 2022 specifically designed to support residents of Scotland. Therefore, outcome 1 (support the Scottish population when experiencing a suicidal crisis that is codeveloped with the target population) was achieved.

Due to low survey engagement, the survey blocks were not rotated as initially intended. Therefore, this pilot phase was unable to obtain visitor views on receiving a compassionate response from the SST website (eg, helpful videos and guidance). Despite this, amendments to the website and survey user evaluation (Multimedia Appendix 3) were conducted in response to user engagement. Thus, a compassionate approach was exhibited by the website developer team to meet the changing needs of the website visitors. Additionally, involvement from the Lived Experience Panel was intended to best serve the needs of the Scottish population. Their input included guidance on sensitive advertising of the website to best avoid triggering people who were not in distress. Additionally, on Meta, comments encouraging suicide prompted the team to discontinue the marketing campaign on this platform for the well-being of social media users. Subsequently, outcome 2 (users of the website feel that they have received a compassionate response) was only partially addressed.

Due to low survey engagement, questions about entrapment were excluded from the web user survey; therefore, outcome 3 (users experience a reduction in feelings of entrapment during engagement with the website) was not addressed. Despite this, the Lived Experience Panel expressed that both internal and external entrapment survey questions were acceptable items to be included within the SST website survey.

The survey data indicated that the intensity of suicidal thoughts had significantly reduced between the time participants first arrived on the SST website and the time they engaged with the survey. However, as the P value just met the traditional level of significance, this needs to be replicated in a large sample. Nonetheless, outcome 4 (users experience a reduction in suicidal thoughts during engagement with the website) was achieved.

Data from Google Analytics indicated that the website was accessed by all geographic regions of Scotland and people up to 70 years of age. This geographic spread of engagement was further reflected by data collected by Nano. Additionally, the website was used at all times of day and night and from a range of devices (eg, computers, phones, and tablets). These findings suggest that the website was both accessible and relatable to residents of Scotland; therefore, outcome 5 (users of the website find it accessible and relatable) was achieved.

Comparison With Prior Work

Consistent with Lai et al [15] and Robinson et al [16], engagement with the SST [20] seemed to significantly reduce the intensity of suicidal thoughts of visitors to the website. The findings of this pilot study are similar to those of the NowMattersNow evaluation [18]. Specifically, the prevalence of web users who described the intensity of their suicidal thoughts as “completely overwhelming” when first visiting the site was similar across the 2 websites (25% of SST web users compared to 28% of NowMattersNow) [18]. Equally, both websites’ bounce rate was within the expected range according to web-based marketing norms of 60%-90% [26] (NowMattersNow [17]: 61% and SST [20]: 76.91%). Both websites reported an absence of participants aged 70+ years. Conversely, Scotland’s SST website was indicated to engage younger people (<18 years: 10.15% of SST visitors), whereas this age group was not captured by NowMattersNow [17].

The pattern of engagement with the SST websites is consistent with leading theories of learned behavior such as social learning theory [11]. Social learning theory posits that people are more likely to model their own behavior on those with whom they identify. Williams et al [19], based on a cross-sectional sample of self-harm web-based communities, found that such groups are more likely to seek informal help from perceived peers than professional services, and advice on self-care was reportedly more effective when provided by these peer supports. Social identification was evident in this study, for example, where web users were more likely to engage with advertisements or SST videos, which corresponded with their own age or sex.

Most visitors were aged between 25 and 34 years, consistent with “Steph,” whose video was viewed the most compared to all other videos on the SST website. Additionally, most visitors accessing the SST website via “X” (formerly known as Twitter) were typically male participants and landed on the SST website via advertisements about “Neil’s story.” Similarly, the marketing company used in this study (Republic of Media) reported that, in general, female participants were more likely than male participants to engage with Google Ads. In this study, advertisements relating to “Linda’s story” were most likely to be engaged with on Google platforms. Although these findings are preliminary, these trends would suggest that it may be beneficial to consider principles of social learning theory [11] when developing suicide prevention websites.

This pilot phase usability study was developed to inform a larger study aimed to reduce suicide in Scotland via timely peer-led support. Subsequently, recommendations based on the findings here include (1) updating marketing strategies to meet the needs of those seeking and not seeking help for their suicidal thoughts,

XSL•FO
RenderX

(2) using pop-up invitations for the user evaluation to improve survey engagement, and (3) increasing the demographic diversity of populations represented in the survivor videos. As illustrated by the logic model (Multimedia Appendix 2), the long-term goals aim to integrate the website into the health care landscape, which clinical and nonclinical populations can use as an established resource to support well-being. The current evaluation also indicates that future work should consider the role of the learned behavior (eg, social learning theory) [11] in relation to engagement with suicide prevention resources. Additionally, to explore the extent to which web visitors felt that they received compassionate support, qualitative interviews with individuals who have used the website would be beneficial.

Limitations

The study findings need to be interpreted in the context of the limitations related to the content of the website, the web user survey, and how the website was marketed. First, most web users usually only engaged with the first 2 videos that appeared on their browser; however, this trend was only identified after the website was live. Although Neil's story and Steph's story were the most engaged with by web users, this may have reflected their position on the website, as opposed to the web users' video preference overall. Additionally, it is unknown whether engagement with the website provided a compassionate response to its visitors.

Second, due to constraints of the survey platform used in this intervention, not all preferred approaches could be implemented into the survey methods. Namely, the survey could only appear once to web users, thereby limiting the survey to a cross-sectional design, where the intensity of suicidal ideation was explored both retrospectively and currently. Additionally, some of the recommendations made by the Lived Experience Panel could not be incorporated into the survey (eg, having the survey invitation being a permanent feature on the screen so website visitors could participate in the survey whenever they felt ready). This may account for the low engagement with the SST survey. Additionally, although the study originally followed the Lived Experience Panel's recommendation of a 3-minute delay before the survey invitation pop-up, due to the average duration of visits to the SST website being shorter than anticipated, the invitation was reduced to 90 seconds, consistent with NowMattersNow [18]. Although survey participation rates were good, the number of people who saw the invitation was lower than expected, limited to those who stayed on the website long enough to see the invitation, and engaged by those who chose to participate in the survey. Subsequently, the findings are unlikely to be generalizable to the wider population of Scotland. The low number of survey invitations was likely due to web users having disabled cookies across the NHS platform. This lower recruitment subsequently led to the blocks of survey items not being rotated; therefore, additional information (eg, sex) not being captured in the data. The small participant sample also limited subgroup analyses to age comparisons only as well as preventing further survey items (eg, entrapment) from being substituted into the survey over time. Given that most participants did not answer the self-reported question relating to the health board, it is unclear whether the participant sample was reflective of the overall SST web user population. Additionally, the final survey item included responses for individuals who were help-seeking for themselves, while those who were seeking help for others were grouped together with those who responded with "prefer not to say." Therefore, it was not possible to make comparisons between reasons for visits in relation to demographic characteristics or suicidal distress. Finally, most respondents to the survey were aged between 40 and 59 years, whereas Google Analytics indicated that most web users were aged between 25 and 34 years. This indicates that pop-up surveys may not be the most effective way to capture web user satisfaction on self-help platforms such as SST and that the findings from the survey may not be generalizable to all visitors to the website.

Third, the marketing approach of the website must be considered. Word limits on Google Ads may have impacted on the success of engaging certain demographics, namely male participants. This is reflected in the alternative advertising platforms, such as Spotify, where proportionately, non–help-seeking male participants were more likely to engage with SST advertisements than non–help-seeking female participants. Challenges were encountered with advertising, with advertising being discontinued on some platforms. This highlights a challenge between balancing a soft-marketing strategy against the need for ethical considerations such as safeguarding to ensure public well-being. Solutions to achieving this balance may include the use of trigger warnings and further developments in social media algorithms.

Finally, the bounce rate may not have provided a true representation of the SST websites' utility by the user. Specifically, the bounce rate of the SST was higher than NowMattersNow [18], reflective of web users leaving the website after arriving on the landing page. However, it is unclear whether the bounce rate was due to visitors benefiting from the website approach to making information salient (eg, crisis card information on the landing page) or because the website did not meet the individual's expectations during their visit, thereby prompting them to leave. Therefore, the bounce rate (ie, engagement) may not necessarily be reflective of the website failing to meet the need, but that the landing page provided sufficient information for the web user.

Conclusions

A codeveloped website for suicide prevention, designed for the Scottish population, may help to reduce the intensity of one's suicidal thoughts. Marketing of the website across various social media, entertainment, and internet search platforms is feasible. The website was particularly popular with those who were looking for help for themselves, and principles of social learning theory may play a role in optimizing the utility of such resources. A larger study is required to explore the effectiveness of such a website and determine whether it works for different subgroups of the Scottish population.

XSL•FO
RenderX

Acknowledgments

The authors would like to thank the Living and Lived Experience Panel of the National Suicide Prevention Leadership Group and members of the NHS24 digital team (Mr Gary Black, Mr Rory Bone, Mr Chris Begg, Ms Lucy Hartely, Ms Gail Lumsden, Ms Carole McCamberly, and Ms Lindsey Thomson) for their expertise in the development of the Surviving Suicidal Thoughts website. No generative artificial intelligence tools (eg, ChatGPT) were used at any stage of this project. This project was funded by the Scottish government.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

DJM is employed by NHS 24 which receives funding for Surviving Suicidal Thoughts from the Scottish Government.

Multimedia Appendix 1
Project aims of the Surviving Suicidal Thoughts website.
[DOCX File , 17 KB - formative_v9i1e55932_app1.docx]

Multimedia Appendix 2
Logic model summarizing the anticipated inputs, outputs, and outcomes of the Surviving Suicidal Thoughts website.
[DOCX File , 19 KB - formative_v9i1e55932_app2.docx]

Multimedia Appendix 3
Timeline illustrating developmental milestones of the Surviving Suicidal Thoughts website between 2020 and 2023.
[DOCX File , 18 KB - formative_v9i1e55932_app3.docx]

Multimedia Appendix 4
The 5 user journeys anticipated visitors to the Surviving Suicidal Thoughts (SST) website will take from using a search engine to landing on the SST website.
[DOCX File , 714 KB - formative_v9i1e55932_app4.docx]

Multimedia Appendix 5
Collated feedback form a Lived Experience Panel in February 2022 regarding the potential survey items to be included in the Surviving Suicidal Thoughts website.
[DOCX File , 16 KB - formative_v9i1e55932_app5.docx]

Multimedia Appendix 6
Survey items included in the Surviving Suicidal Thought website cross-sectional user evaluation survey from July 2022 to February 2023.
[DOCX File , 16 KB - formative_v9i1e55932_app6.docx]

Multimedia Appendix 7
Summary of social media costs for the Surviving Suicidal Thoughts website first marketing campaign from July to September 2022.
[DOCX File , 16 KB - formative_v9i1e55932_app7.docx]

References

1. World Health Organization. Suicide prevention. 2024. URL: https://www.who.int/health-topics/suicide#tab=tab_1 [accessed 2024-11-19]
2. Probable suicides. National Records of Scotland. 2022. URL: https://www.nrscotland.gov.uk/statistics-and-data/statistics/statistics-by-theme/vital-events/deaths/suicides#:~:text=Probable%20Suicides.%20Information%20on%20the [accessed 2024-11-19]
3. Review of suicide statistics in Northern Ireland 2015 - 2020. Northern Ireland Statistcs and Research Agency. 2022. URL: https://www.nisra.gov.uk/system/files/statistics/Suicide_Review_Report.pdf [accessed 2024-11-19]

XSL•FO
RenderX

4. Suicides in England and Wales: 2021 registrations. Office for National Statistics. 2021. URL: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/bulletins/suicidesintheunitedkingdom/2021registrations [accessed 2024-11-19]
5. Michelmore L, Hindley P. Help-seeking for suicidal thoughts and self-harm in young people: a systematic review. Suicide Life Threat Behav 2012;42(5):507-524. [doi: 10.1111/j.1943-278X.2012.00108.x] [Medline: 22889130]
6. Creating hope together: suicide prevention strategy 2022 to 2032. Scottish Government. 2022. URL: https://www.gov.scot/publications/creating-hope-together-scotlands-suicide-prevention-strategy-2022-2032/ [accessed 2024-11-19]
7. Time, space, compassion three simple words, one big difference: recommendations for improvements in suicidal crisis response. Scottish Government. 2021. URL: https://www.gov.scot/publications/time-space-compassion-three-simple-words-one-big-difference-recommendations-improvements-suicidal-crisis-response/ [accessed 2024-11-19]
8. Mann JJ, Apter A, Bertolote J, Beautrais A, Currier D, Haas A, et al. Suicide prevention strategies: a systematic review. JAMA 2005;294(16):2064-2074. [doi: 10.1001/jama.294.16.2064] [Medline: 16249421]
9. O'Connor RC, Kirtley OJ. The integrated motivational-volitional model of suicidal behaviour. Philos Trans R Soc Lond B Biol Sci 2018;373(1754):20170268 [FREE Full text] [doi: 10.1098/rstb.2017.0268] [Medline: 30012735]
10. O'connor RC. Towards an integrated motivational–volitional model of suicidal behaviour. In: O'connor RC, Platt S, Gordon J, editors. International Handbook of Suicide Prevention: Research, Policy and Practice. Chichester: John Wiley & Sons, Ltd; 2011.
11. Bandura A. Model of causality in social learning theory. In: Cognition and Psychotherapy. Englewood Cliffs, NJ: Prentice-Hall; 1977:81-99.
12. Coussi-Korbel S, Fragaszy DM. On the relation between social dynamics and social learning. Anim Behav 1995;50(6):1441-1453. [doi: 10.1016/0003-3472(95)80001-8]
13. Till B, Strauss M, Sonneck G, Niederkrotenthaler T. Determining the effects of films with suicidal content: a laboratory experiment. Br J Psychiatry 2015;207(1):72-78. [doi: 10.1192/bjp.bp.114.152827] [Medline: 25906792]
14. Tajfel H, Turner J. An integrative theory of intergroup conflict. In: The Social Psychology of Intergroup Relations. Monterey, CA: Brooks/Cole Publishing Company; 1979:56-65.
15. Lai MH, Maniam T, Chan LF, Ravindran AV. Caught in the web: a review of web-based suicide prevention. J Med Internet Res 2014;16(1):e30 [FREE Full text] [doi: 10.2196/jmir.2973] [Medline: 24472876]
16. Robinson J, Cox G, Bailey E, Hetrick S, Rodrigues M, Fisher S, et al. Social media and suicide prevention: a systematic review. Early Interv Psychiatry 2016;10(2):103-121. [doi: 10.1111/eip.12229] [Medline: 25702826]
17. Whiteside URJEM. NowMattersNow. 2014. URL: https://nowmattersnow.org/ [accessed 2024-10-04]
18. Whiteside U, Richards J, Huh D, Hidalgo R, Nordhauser R, Wong AJ, et al. Development and evaluation of a web-based resource for suicidal thoughts: NowMattersNow.org. J Med Internet Res 2019;21(5):e13183 [FREE Full text] [doi: 10.2196/13183] [Medline: 31045498]
19. Williams AJ, Nielsen E, Coulson NS. "They aren't all like that": perceptions of clinical services, as told by self-harm online communities. J Health Psychol 2020;25(13-14):2164-2177 [FREE Full text] [doi: 10.1177/1359105318788403] [Medline: 30024273]
20. MacIntyre DJG, McClelland H, O'Connor RC. Surviving suicidal thoughts. NHS Inform. 2022. URL: https://www.nhsinform.scot/surviving-suicidal-thoughts [accessed 2024-10-04]
21. UK policy framework for health and social care research. Health Research Authority. URL: https://s3.eu-west-2.amazonaws.com/www.hra.nhs.uk/media/documents/uk-policy-framework-health-social-care-research.pdf [accessed 2025-01-09]
22. Google Analytics: analytics tools & solutions for your business. Google Marketing Platform. 2023. URL: https://marketingplatform.google.com/about/analytics/ [accessed 2024-10-04]
23. What is Hotjar? HotJar. 2014. URL: https://www.hotjar.com/blog/what-is-hotjar/ [accessed 2024-10-04]
24. Faul F, Erdfelder E, Buchner A, Lang AG. Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. Behav Res Methods 2009;41(4):1149-1160. [doi: 10.3758/BRM.41.4.1149] [Medline: 19897823]
25. About demographic targeting. Google Ads Help. 2023. URL: https://support.google.com/google-ads/answer/2580383 [accessed 2024-09-28]
26. Bounce rates: what's good, what's bad, and why you should give a damn. CustomMedia. URL: https://www.customedialabs.com/blog/bounce-rates/ [accessed 2024-10-04]

Abbreviations

CTR: click-through rate
NHS: National Health Service
SST: Surviving Suicidal Thoughts

XSL•FO
RenderX

Edited by A Mavragani; submitted 02.02.24; peer-reviewed by IT Hwang, C Hughes , A AL-Asadi; comments to author 06.06.24; revised version received 30.07.24; accepted 13.10.24; published 24.01.25.

Please cite as:
McClelland H, O'Connor RC, Gibson L, MacIntyre DJ
Exploring Web-Based Support for Suicidal Ideation in the Scottish Population: Usability Study
JMIR Form Res 2025;9:e55932
URL: https://formative.jmir.org/2025/1/e55932
doi:10.2196/55932
PMID:

©Heather McClelland, Rory C O'Connor, Laura Gibson, Donald J MacIntyre. Originally published in JMIR Formative Research (https://formative.jmir.org), 24.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Engagement With an Internet-Administered, Guided, Low-Intensity Cognitive Behavioral Therapy Intervention for Parents of Children Treated for Cancer: Analysis of Log-Data From the ENGAGE Feasibility Trial

Christina Reuther[1], MSc; Louise von Essen[1], PhD; Mudassir Imran Mustafa[1], PhD; Markus Saarijärvi[2*], PhD; Joanne Woodford[1*], PhD

[1]CIRCLE - Complex Intervention Research in Health and Care, Department of Women's and Children's Health, Uppsala University, Uppsala, Sweden
[2]Department of Neurobiology, Care Sciences and Society, Karolinska Institute, Stockholm, Sweden
[*]these authors contributed equally

Corresponding Author:
Louise von Essen, PhD
CIRCLE - Complex Intervention Research in Health and Care
Department of Women's and Children's Health
Uppsala University
Dag Hammarskjölds 14B
Uppsala, 75237
Sweden
Phone: 46 70 425 07 14
Email: louise-von.essen@uu.se

Abstract

Background: Parents of children treated for cancer may experience psychological difficulties including depression, anxiety, and posttraumatic stress. Digital interventions, such as internet-administered cognitive behavioral therapy, offer an accessible and flexible means to support parents. However, engagement with and adherence to digital interventions remain a significant challenge, potentially limiting efficacy. Understanding factors influencing user engagement and adherence is crucial for enhancing the acceptability, feasibility, and efficacy of these interventions. We developed an internet-administered, guided, low-intensity cognitive behavioral therapy (LICBT)–based self-help intervention for parents of children treated for cancer, (EJDeR [internetbaserad självhjälp för föräldrar till barn som avslutat en behandling mot cancer or internet-based self-help for parents of children who have completed cancer treatment]). EJDeR included 2 LICBT techniques—behavioral activation and worry management. Subsequently, we conducted the ENGAGE feasibility trial and EJDeR was found to be acceptable and feasible. However, intervention adherence rates were marginally under progression criteria.

Objective: This study aimed to (1) describe user engagement with the EJDeR intervention and examine whether (2) sociodemographic characteristics differed between adherers and nonadherers, (3) depression and anxiety scores differed between adherers and nonadherers at baseline, (4) user engagement differed between adherers and nonadherers, and (5) user engagement differed between fathers and mothers.

Methods: We performed a secondary analysis of ENGAGE data, including 71 participants. User engagement data were collected through log-data tracking, for example, communication with e-therapists, homework submissions, log-ins, minutes working with EJDeR, and modules completed. Chi-square tests examined differences between adherers and nonadherers and fathers and mothers concerning categorical data. Independent-samples *t* tests examined differences regarding continuous variables.

Results: Module completion rates were higher among those who worked with behavioral activation as their first LICBT module versus worry management. Of the 20 nonadherers who opened the first LICBT module allocated, 30% (n=6) opened behavioral activation and 70% (n=14) opened worry management. No significant differences in sociodemographic characteristics were found. Nonadherers who opened behavioral activation as the first LICBT module allocated had a significantly higher level of depression symptoms at baseline than adherers. No other differences in depression and anxiety scores between adherers and nonadherers were found. Minutes working with EJDeR, number of log-ins, days using EJDeR, number of written messages sent to e-therapists, number of written messages sent to participants, and total number of homework exercises submitted were

XSL•FO
RenderX

significantly higher among adherers than among nonadherers. There were no significant differences between fathers and mothers regarding user engagement variables.

Conclusions: Straightforward techniques, such as behavioral activation, may be well-suited for digital delivery, and more complex techniques, such as worry management, may require modifications to improve user engagement. User engagement was measured behaviorally, for example, through log-data tracking, and future research should measure emotional and cognitive components of engagement.

Trial Registration: ISRCTN Registry 57233429; https://doi.org/10.1186/ISRCTN57233429

(JMIR Form Res 2025;9:e67171) doi:10.2196/67171

KEYWORDS

childhood cancer survivor; cognitive behavioral therapy; engagement; internet-administered intervention; log-data; parents

Introduction

Background

The digital delivery of psychological interventions, such as cognitive behavioral therapy (CBT), particularly internet-administered cognitive behavioral therapy (iCBT), has rapidly grown over the past 2 decades [1]. iCBT interventions have shown to be effective in reducing symptoms of depression [2] and anxiety [3,4], with outcomes comparable with traditional face-to-face CBT when professional guidance is included as a component [5,6]. Flexibility, convenience [7], and the ability to overcome logistical and financial barriers [8], alongside the potential to reduce the stigma associated with seeking mental health treatment [8,9], are significant advantages of iCBT. In addition, given the lack of available therapists, iCBT interventions represent a potentially low-cost alternative to disseminating psychological interventions to a wider audience [10].

Despite these advantages, how users engage with digital interventions, including iCBT, is uncertain [11,12]. User engagement, which can be defined as how and to what extent users actively use digital interventions, is associated with the efficacy of iCBT interventions [13,14]. Low or inconsistent user engagement with digital interventions is closely associated with common issues such as low uptake (ie, a smaller-than-expected number of users participating in and benefiting from the intervention), variable usage (ie, users engaging with the intervention inconsistently or unevenly), and high attrition rates (ie, users drop out of the intervention) [15,16].

A related concept to user engagement is “adherence,” referring to whether an intervention is used in the way it was designed to maximize efficacy [17]. Adherence may be based on a minimal level of user engagement considered necessary to result in changes to intended outcomes, that is, a minimum treatment dose [14]. Although adherence to digital interventions is an evolving concept [18], the total number of completed sessions has been found to predict treatment response in iCBT interventions [19]. However, knowledge concerning the causal pathways between user engagement, adherence, and efficacy remains limited [20]. Furthermore, there is insufficient understanding of the potential association between sociodemographic factors, for example, age, gender, sex, and clinical characteristics (ie, depression or anxiety symptoms), and user engagement and adherence [19]. Adherence rates also vary between studies, with a review of adherence to iCBT interventions reporting adherence rates ranging from 6% to 100% [16]. Furthermore, high attrition rates, particularly in unguided iCBT interventions, present additional challenges [21], potentially resulting in biased treatment effects [22] and limiting generalizability [23].

Although user engagement is associated with treatment response [18], there remains a lack of clarity on how to conceptualize, define, and measure user engagement within the context of digital interventions [11,12,20]. Engagement is considered multidimensional [24,25] and can include to what extent an intervention is used, for example, by log-data tracking to quantitatively capture digital intervention usage [20], as well as the user’s subjective experience, for example, affect, attention, and interest [25]. Therefore, one way to explore user engagement is by analyzing log-data (ie, automatically collected digital records of a participant’s use of an intervention such as session duration, completion rates, and interaction patterns) [26,27] and examining potential relationships between user engagement and participant’s sociodemographic and clinical characteristics [28]. There is also a need to investigate the use of digital interventions to explore how certain aspects of user engagement may relate to adherence [20] and inform the minimal usage required to establish adherence [29].

This Study

We conducted a single-arm feasibility trial (ENGAGE) [30-32] to examine the acceptability and feasibility of an internet-administered, guided, low-intensity cognitive behavioral therapy (LICBT) intervention for parents of children treated for cancer: EJDeR (Swedish acronym: intErnetbaserad sJälvhjälp för förälDrar till barn som avslutat en behandling mot canceR [Internet-based self-help for parents of children who have completed cancer treatment]) [33-38]. EJDeR was developed alongside parent research partners [38] and is designed for parents of children treated for cancer experiencing symptoms of depression and/or generalized anxiety disorder (GAD) related to their child’s cancer treatment [33-38]. EJDeR is delivered through the Uppsala University Psychosocial Care Programme (U-CARE) portal (Portal), an in-house web-based platform, designed to deliver digital interventions and support the execution of study procedures (eg, online consent and data collection) [38]. EJDeR includes four modules: (1) introduction and psychoeducation (IPE), (2) behavioral activation (BA), (3)

XSL•FO
RenderX

worry management (WM), and (4) relapse prevention (RP), and is guided by an e-therapist. Findings from ENGAGE show EJDeR is an acceptable and feasible intervention; therefore, progression to an external pilot randomized controlled trial (RCT) to prepare for the design and conduct of a future superiority RCT is warranted. However, adherence to the minimal treatment dose was 47.9%, marginally under the progression criteria of 50% [31]. In addition, visual inspection of adherence data indicated possible differences in adherence rates by module worked with and between fathers and mothers [31]. Previous research indicates fathers and mothers may engage differently with psychological interventions [39], emphasizing the need to consider gender-specific engagement patterns in digital intervention design. For example, fathers have been found to prefer solution-focused, practical, and time-efficient formats that offer structured guidance and actionable strategies [40]. In contrast, mothers have been found to prefer more emotional and process-oriented intervention content, valuing opportunities for self-reflection and emotional processing [41].

Given these findings, there is a need to conduct a secondary analysis of data from ENGAGE to better understand how participants engaged with the intervention. Results may help inform ways to adapt and modify the intervention to improve adherence rates before proceeding to an external pilot RCT. Further modification of the intervention to improve user engagement and thus adherence may also improve treatment outcomes in a future superiority RCT [31]. Findings may also have wider applicability to other iCBT interventions.

Study Aims

This study aimed to (1) describe user engagement with the EJDeR intervention, (2) examine whether sociodemographic and baseline clinical characteristics differed between adherers and nonadherers, (3) examine whether depression and anxiety scores differed between adherers and nonadherers at baseline, (4) examine whether user engagement differed between adherers and nonadherers, and (5) examine whether user engagement differed between fathers and mothers.

Methods

Study Design

This study is a secondary analysis of data from ENGAGE. The trial protocol [32] and main findings from ENGAGE are reported elsewhere [31].

Participant Inclusion Criteria

Eligibility criteria included being (1) a parent of a child diagnosed with cancer during childhood (0-18 years) who completed cancer treatment 3 months to 5 years previously (timespan informed by our previous longitudinal research that has identified this as a period of vulnerability for parents [42,43], (2) resident in Sweden, (3) able to read and understand Swedish, (4) able to access email, internet, and BankID (a citizen authentication system used in Sweden), and (5) seeking psychological support related to the child's cancer. Exclusion criteria included (1) severe and enduring mental health problems, for example, bipolar disorder and psychosis; (2) acute suicidality; (3) misuse of alcohol; street drugs or prescription medication; and (4) currently attending psychological treatment. Participants excluded due to criteria 1-3 were directed to appropriate health care services.

Participant Recruitment

Parents were recruited into ENGAGE using 2 main approaches. First, personal identification numbers of children were obtained from the Swedish Childhood Cancer Registry (National Quality Registry) and linked to parents' names and addresses through NAVET, a registry held by the Swedish Tax Agency. Parents were invited to participate by postal invitations using random blocks of 100 until the target number of 50 was reached. Second, advertisements were placed on relevant social media sites and websites of patient organizations and interest groups [31].

An adapted CONSORT (Consolidated Standards of Reporting Trials) diagram for ENGAGE has been published [31]. Recruitment took place over 5 months (July 3, 2020, to November 30, 2020). A total of 72 participants gained access to EJDeR, with 1 excluded shortly after access (severe and enduring mental health difficulty), resulting in a sample of 71.

Intervention

EJDeR was delivered via the Portal, incorporating text, homework exercises, films, illustrations, and audio content. A detailed description of EJDeR has been published [38], and an overview of the intervention is presented in Figure 1. EJDeR was delivered over 12 weeks and consisted of 4 modules: (1) IPE, (2) BA, (3) WM, and (4) RP. It includes 2 LICBT techniques—BA for depression [44] and WM for GAD [45]. The LICBT clinical protocol for BA has been published elsewhere [44]. Participants are supported to re-engage with pleasurable, necessary, and routine activities they have stopped doing in a gradual and structured way. The full LICBT clinical protocol for WM has also been published elsewhere [45,46]. Participants record worries and categorize worries into 2 types—practical (eg, important and can be solved) and hypothetical (eg, important but have no way of being solved). Participants work with problem-solving for practical worries and worry time for hypothetical worries.

Guidance was provided by e-therapists, who were trained in the competencies required to support LICBT. After being provided with access to EJDeR, participants were invited to attend an initial assessment session (telephone or videoconferencing) with their e-therapist. During the initial assessment, e-therapists used patient-centered interviewing techniques to understand the participant's main presenting difficulties, provide psychoeducation, introduce the LICBT techniques used in EJDeR, and come to a collaborative decision to allocate 1 of the 2 LICBT modules to first work with, based on the participant's main difficulty, for example, BA for depression or WM for GAD. E-therapists did not have access to participants' baseline assessment scores (ie, depression and anxiety symptoms) and therefore allocation to the first LICBT module to work with was not informed by symptom severity at baseline. Subsequently, the e-therapist provided participants with access to the LICBT module. E-therapists provided weekly written messages via the Portal to guide participants in using

XSL•FO
RenderX

the intervention [47]. To facilitate engagement with the LICBT modules, weekly homework exercises were either completed on the Portal or printed as a PDF and completed offline. E-therapists also invited participants to attend a midintervention booster session (telephone or videoconferencing). After completing the first LICBT module allocated (BA or WM), participants could choose to work with the remaining LICBT module. After completing either BA or WM, participants were provided with access to RP by their e-therapist.

Persuasive system design elements [48,49] are built into EJDeR to improve user engagement, including tunneling, for example, content is delivered in a predefined step-wise order to guide participants through using EJDeR, tailoring (eg, EJDeR content is personalized to user needs [ie, depression or GAD]), rehearsal (eg, homework exercises are repeated), and liking (eg, use of professional illustrations).

Figure 1. An overview of the structure of EJDeR. Figure from Woodford et al [38] which is published under Creative Commons Attribution 4.0 International License [50].

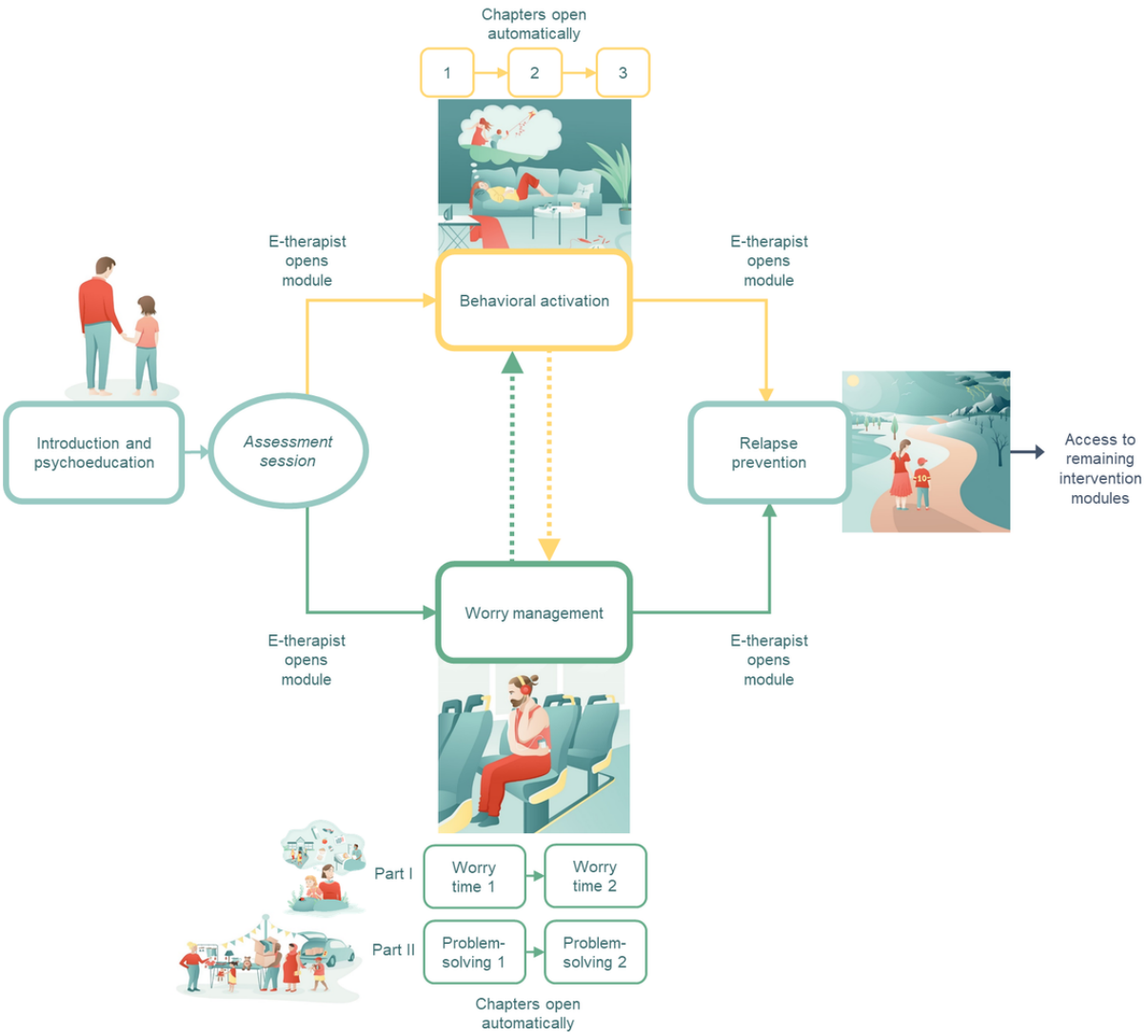

Intervention Adherence

Adherence to EJDeR was conceptualized as participants engaging with the intervention following a priori–defined minimum treatment dose. Intervention adherers were defined as those (1) attending the initial assessment session with an e-therapist (telephone or videoconferencing), (2) completing the IPE module (2 chapters) and 1 LICBT treatment module (ie, BA [3 chapters] or WM defined as completing an introduction [2 chapters] and completing either problem solving [2 chapters] or worry time [2 chapters]), and (3) attending the midintervention booster session with an e-therapist (telephone or videoconferencing). For a module to be defined as being completed, participants were required to submit all chapters included within each module to their e-therapist on the Portal. Our minimum treatment dose was informed by research suggesting that guidance from a trained professional is associated with larger treatment effects than self-administered LICBT [51,52]. In addition, we considered it important for participants to complete the IPE module to gain an understanding of the CBT rationale and complete all chapters for 1 LICBT technique to enable them to understand and apply

XSL•FO
RenderX

the technique to manage their symptoms of depression and/or anxiety. We did not include the submission of weekly homework exercises in our definition of adherence given participants were able to complete on the Portal or print as a PDF and complete offline.

Measures

Sociodemographic Characteristics

Sociodemographic characteristics were collected during a telephone eligibility interview with a licensed psychologist and included age, sex (male or female), relationship status (partner or single), highest level of education (lower secondary, upper secondary, postsecondary nontertiary, tertiary, or PhD), employment status (employed or unemployed), number of children, housing situation (rental, apartment ownership, house ownership, or other), region of birth (Nordic countries, Asia, Europe [excluding Nordic countries], or Africa), previous psychological treatment (yes or no), physical health problem (yes or no), and previous traumatic or difficult life events (yes or no).

Baseline Clinical Characteristics

Clinical characteristics were collected via a Portal assessment at baseline. Symptoms of depression were assessed using the 9-item Patient Health Questionnaire (PHQ-9) [53]. Symptoms of anxiety were assessed using the 7-item Generalized Anxiety Disorder scale (GAD-7) [54].

User Engagement

Log-Data Tracking on the Portal

User action logging was enabled through action metadata management to allow user behavior analysis, including (1) minutes working with EJDeR, (2) number of log-ins, (3) number of days using EJDeR, (4) number of written messages sent to e-therapists, (5) number of written messages sent to participants, (6) total number of modules opened, (7) total number of modules completed, (8) opened modules (IPE and RP), (9) completed modules (IPE and RP), (10) opened first LICBT module allocated (BA and WM), (11) completed first LICBT module allocated (BA and WM), and (12) submitted homework exercises.

Data Collected Outside the Portal

Intervention activities outside of the Portal were collected by research team members and entered into a Microsoft Excel spreadsheet, including (1) attending the initial assessment session (yes or no), and (2) attending the midintervention booster session (yes or no).

Statistical Analysis

All statistical analyses were performed in the SPSS (version 28.0.1; IBM Corp). Given multiple statistical comparisons were conducted, a Bonferroni correction was applied to adjust for type I error inflation [55]. Missing data (% of missing items <0.001) were handled using single-value imputation. Sociodemographic and baseline clinical characteristics and user engagement data for all participants are reported using descriptive statistics. Absolute proportions are reported for categorical variables and means and SDs are reported for continuous variables. Chi-square tests were used to examine differences between adherers and nonadherers and fathers and mothers concerning categorical data. Effect sizes (Cohen w) were calculated to determine the magnitude of difference between the groups for categorical variables. Cohen w rule of thumb cutoffs was adopted: 0.1-0.3, indicating a small difference; 0.3-0.5, a moderate difference; and >0.5, a large difference [56]. Independent-sample *t* tests were used to examine differences regarding continuous variables.

Ethical Considerations

This study was approved by the Regional Ethical Review Board in Uppsala, Sweden (Dnr: 2017/527), with an amendment obtained from the Swedish Ethical Review Authority (2019-03083). The research was conducted following the Helsinki Declaration and relevant Swedish regulations for human subject research. Parents interested in participating provided informed consent via the Portal. Data collected for the study were pseudonymized to ensure participant confidentiality. Participants were not compensated for their participation in the study. Participation was voluntary, and participants could withdraw at any time.

Results

Intervention Flow

The flow of participants through the intervention is illustrated in Figure 2. In total, 77% (20/26) of participants starting with BA as the first LICBT module allocated completed the module versus 50% (14/28) starting with WM.

XSL•FO
RenderX

Figure 2. Flow of participants through the EJDeR intervention. BA: behavioral activation; EJDeR: intErnetbaserad sJälvhjälp för förälDrar till barn som avslutat en behandling mot canceR; IPE: introduction and psychoeducation; LICBT: low-intensity cognitive behavioral therapy; PE: psychoeducation; RP: relapse prevention; WM: worry management. Two participants were provided access to the RP module without completing a LICBT module, due to e-therapist error.

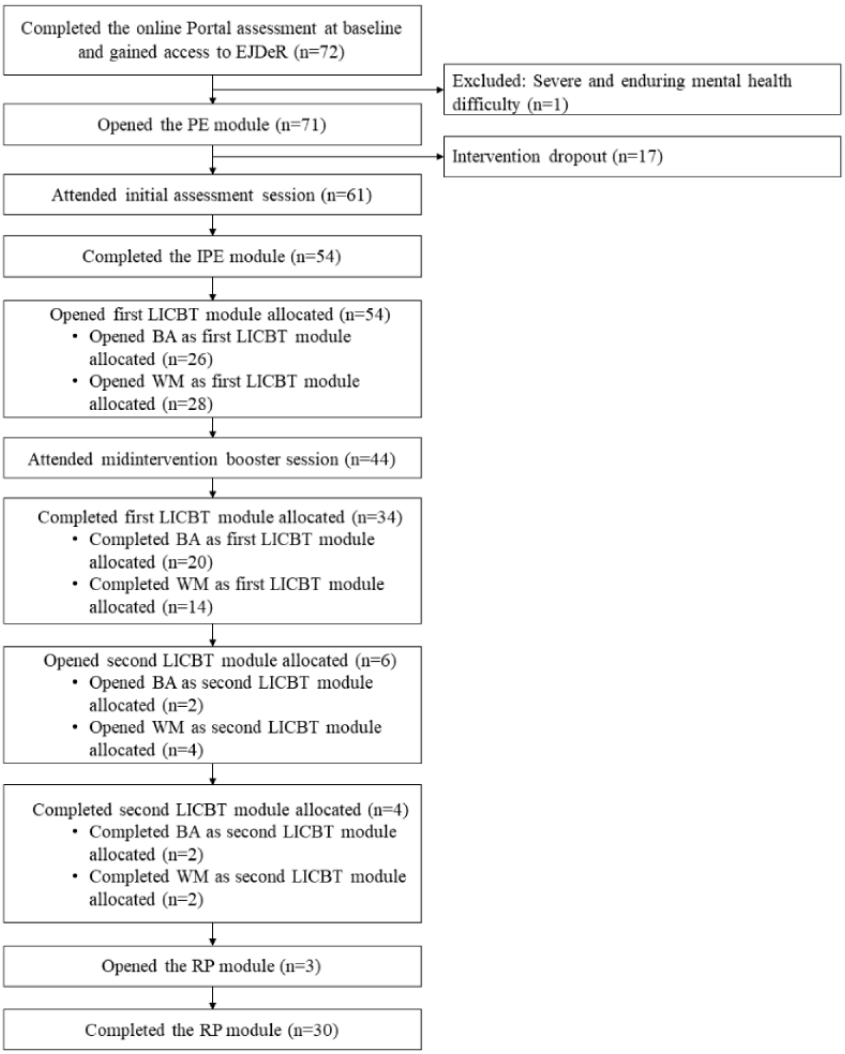

User Engagement With the EJDeR Intervention

Participants spent an average of 205.36 (SD 154.64) minutes working with EJDeR, logging in an average of 19.99 (SD 14.95) times over 76.31 (SD 31.49) days. Participants opened a mean number of 2.3 (SD 0.92) modules and completed an average of 1.65 (SD 1.31) modules (Table 1).

Table 1. User engagement with the EJDeR intervention (N=71).

User engagement (log-data on the Portal)	Mean (SD)	Range
Minutes working with EJDeR[a]	205.36 (154.64)	8.59-611.74
Number of log-ins	19.99 (14.95)	1-72
Number of days using EJDeR	76.31 (31.49)	0-126
Number of written messages sent to e-therapists	8.52 (7.56)	0-33
Number of written messages sent to participants	28.79 (16.26)	0-74
Total number of modules opened	2.30 (0.92)	1-4
Total number of modules completed	1.65 (1.31)	0-4
Total homework exercises submitted	2.70 (2.78)	0-11

[a]EJDeR: intErnetbaserad sJälvhjälp för förälDrar till barn som avslutat en behandling mot cancer.

XSL•FO
RenderX

Sociodemographic Between Intervention Adherers and Nonadherers

There were no significant differences between intervention adherers and nonadherers by sociodemographic characteristics (Table 2).

XSL•FO
RenderX

Table 2. Sociodemographic characteristics for total sample (N=71) and by adherers (n=34) and nonadherers (n=37) to the minimum treatment dose for the EJDeR intervention[a].

Sociodemographic characteristics	Total study sample (N=71)		Adherers (n=34)		Nonadherers (n=37)		*P* value (effect size)[b]
	n (%)[c]	Range	n (%)	Range	n (%)	Range	
Age (years), mean (SD)	42.73 (7.02)	26-62	41.71 (5.30)	33-59	43.68 (8.25)	26-62	.24 (*d*=0.28)
Sex							.99 (w=0.12)
Female	46 (65)	__[d]	22 (65)	—	24 (65)	—	
Male	25 (35)	—	12 (35)	—	13 (35)	—	
Relationship status							.41 (w=0.05)
Partner	60 (85)	—	30 (88)	—	30 (81)	—	
Single	11 (15)	—	4 (12)	—	7 (19)	—	
Highest level of education							.74 (w=0.09)
Lower secondary	1 (1)	—	0 (0)	—	1 (3)	—	
Upper secondary	13 (18)	—	5 (15)	—	8 (22)	—	
Postsecondary nontertiary	3 (4)	—	1 (3)	—	2 (5)	—	
Tertiary	52 (73)	—	27 (79)	—	25 (68)	—	
PhD	2 (3)	—	1 (3)	—	1 (3)	—	
Employment status							.90 (w=0.11)
Employed	63 (89)	—	30 (88)	—	33 (89)	—	
Unemployed	8 (11)	—	4 (12)	—	4 (11)	—	
Number of children, mean (SD)	2.28 (0.78)	1-5	2.12 (0.69)	1-5	2.43 (0.84)	1-4	.09 (*d*=0.40)
Housing situation							.07 (w=0.01)
Rental	7 (10)	—	2 (6)	—	5 (14)	—	
Apartment ownership	17 (24)	—	6 (18)	—	11 (30)	—	
House ownership	44 (62)	—	26 (76)	—	18 (49)	—	
Other	3 (4)	—	0 (0)	—	3 (8)	—	
Region of birth							.18 (w=0.02)
Nordic countries[e]	59 (83)	—	31 (91)	—	28 (76)	—	
Asia	6 (9)	—	3 (9)	—	3 (8)	—	
Europe (excluding Nordic countries)	5 (7)	—	0 (0)	—	5 (14)	—	
Africa	1 (1)	—	0 (0)	—	1 (3)	—	
Previous psychological treatment							.57 (w=0.07)
Yes	38 (54)	—	17 (50)	—	21 (57)	—	
No	33 (46)	—	17 (50)	—	16 (43)	—	
Physical health problems							.61(w=0.07)
Yes	23 (32)	—	10 (29)	—	13 (35)	—	
No	48 (68)	—	24 (71)	—	24 (65)	—	
Previous traumatic or difficult life event							.67 (w=0.08)
Yes	57 (80)	—	28 (82)	—	29 (78)	—	
No	14 (20)	—	6 (18)	—	8 (22)	—	

[a]Adherence to EJDeR was conceptualized as users engaging with the intervention in accordance with a priori–defined minimum treatment dose.

XSL•FO
RenderX

Intervention adherers are therefore defined as those (1) attending the initial assessment session with an e-therapist, (2) completing the psychoeducation module and one low-intensity treatment module (ie, behavioral activation or worry management), and (3) attending the midintervention booster session with an e-therapist.

[b]A Bonferroni correction was applied based on 11 tests resulting in an α of .005.

[c]Percentages may not add to 100 due to rounding.

[d]Not applicable.

[e]Nordic countries represented in the study sample include Denmark, Finland, Norway, and Sweden.

Depression and Anxiety Scores Between Intervention Adherers and Nonadherers at Baseline

Nonadherers who opened BA as the first LICBT module allocated had a significantly higher level of depression symptoms at baseline than adherers (*P*=.04); however, this finding was no longer significant after applying a Bonferroni correction (α=.008). There were no other significant differences between intervention adherers and nonadherers by depression and anxiety score at baseline (Table 3).

Table 3. Depression and anxiety scores at baseline for total sample (N=71) and by adherers (n=34) and nonadherers (n=37) to the minimum treatment dose for the EJDeR intervention[a].

Depression and anxiety scores at baseline	Total study sample (N=71)		Adherers (n=34)		Nonadherers (n=37)		*P* value (effect size)[b]
	n	Mean (SD)	n	Mean (SD)	n	Mean (SD)	
PHQ-9[c] for all participants	71	6.55 (5.05)	34	6.24 (4.65)	37	6.84 (5.43)	.62 (d=0.12)
PHQ-9 for participants who opened BA[d] as first LICBT[e] module[f]	26	8.31 (6.03)	20	7.00 (5.49)	6	12.67 (6.15)	.04 (d=0.97)
PHQ-9 for participants who opened WM[g] as first LICBT module[f]	28	4.46 (3.01)	14	5.14 (2.96)	14	3.79 (3.02)	.24 (d=–0.45)
GAD-7[h] for all participants	71	5.92 (4.53)	34	6.06 (4.25)	37	5.78 (4.83)	.80 (d=–0.06)
GAD-7 for participants who opened BA as first LICBT module[f]	26	6.85 (5.09)	20	5.90 (4.62)	6	10.0 (5.73)	.08 (d=0.79)
GAD-7 for participants who opened WM as first LICBT module[f]	28	5.11 (3.75)	14	6.29 (3.81)	14	3.93 (3.41)	.10 (d=–0.65)

[a]Adherence to EJDeR was conceptualized as users engaging with the intervention in accordance with a priori–defined minimum treatment dose. Intervention adherers are therefore defined as those (1) attending the initial assessment session with an e-therapist, (2) completing the psychoeducation module and 1 low-intensity treatment module (ie, behavioral activation or worry management), and (3) attending the midintervention booster session with an e-therapist.

[b]A Bonferroni correction was applied based on 6 tests resulting in an α of .008.

[c]PHQ-9: 9-item Patient Health Questionnaire.

[d]BA: behavioral activation.

[e]LICBT: low-intensity cognitive behavioral therapy.

[f]Total study sample based on those gaining access to a low-intensity cognitive behavioral therapy module.

[g]WM: worry management.

[h]GAD-7: 7-item Generalized Anxiety Disorder scale.

User Engagement Between Intervention Adherers and Nonadherers

There was a significant difference between adherers and nonadherers by first LICBT module allocated (*P*=.04); however, this finding was no longer significant after applying a Bonferroni correction (α=.007). Of the 34 adherers who opened the first LICBT module allocated, 59% (n=20) opened BA as the first LICBT module allocated and 41% (n=14) opened WM. Of the 20 nonadherers who opened the first LICBT module allocated, 30% (6) opened BA as their first LICBT module allocated and 70% (14) opened WM.

The number of minutes working with EJDeR (*P*≤.001), number of log-ins (*P*≤.001), number of days using EJDeR (*P*≤.001), number of written messages sent to e-therapists (*P*≤.001), number of written messages sent to participants (*P*≤.001), and total number of homework exercises submitted (*P*≤.001) were significantly higher among intervention adherers compared with nonadherers. There were no significant differences between adherers and nonadherers related to the other user engagement variables (Table 4).

XSL•FO
RenderX

Table 4. User engagement with the EJDeR intervention by adherers (n=34) and non-adherers (n=37) to the minimum treatment dose for the EJDeR intervention[a].

Measure of user engagement (log-data on the Portal)	Adherers (n=34)		Nonadherers (n=37)		*P* value (effect size)[b]
	Mean (SD)	Range	Mean (SD)	Range	
Minutes working with EJDeR[c]	325.82 (126.61)	142.41-611.74	94.67 (74.21)	8.59-282.95	<.001 (d=2.23)
Number of log-ins	31.32 (12.64)	5-72	9.57 (7.42)	1-29	<.001 (d=2.10)
Number of days using EJDeR	91.47 (13.76)	57-126	62.38 (36.58)	0-126	<.001 (d=1.05)
Number of written messages sent to e-therapists	12.88 (7.64)	2-33	4.51 (4.79)	0-17	<.001 (d=1.31)
Number of written messages sent to participants	40.21 (12.09)	24-74	18.30 (11.98)	0-47	<.001 (d=1.82)
First LICBT[d] module allocated opened					.04 (w=0.01)
Opened BA[e], n (%)	20[f] (59)	—[g]	6[h] (30)	—	
Opened WM[i], n (%)	14[f] (41)	—	14[h] (70)	—	
Total homework exercises submitted	4.79 (2.53)	0-11	0.78 (1.13)	0-4	<.001 (d=2.05)

[a]Adherence to EJDeR was conceptualized as users engaging with the intervention in accordance with a priori defined minimum treatment dose. Not all user engagement variables listed in Table 1 are included in this analysis as some measures of user engagement, that is, number of modules opened, attended initial assessment session, attended mid-intervention booster session, inform the minimum treatment dose.

[b]A Bonferroni correction was applied based on 7 tests resulting in an α of .007.

[c]EJDeR: intErnetbaserad sJälvhjälp för förälDrar till barn som avslutat en behandling mot canceR.

[d]LICBT: low-intensity cognitive behavioral therapy.

[e]BA: behavioral activation.

[f]number of adherers who opened the first low-intensity cognitive behavioral therapy module allocated (n=34).

[g]Not applicable.

[h]number of nonadherers who opened the first low-intensity cognitive behavioral therapy module allocated (n=20).

[i]WM: worry management.

User Engagement Between Fathers and Mothers

There were no significant differences between fathers and mothers related to user engagement variables (Table 5).

XSL•FO
RenderX

Table 5. User engagement with the EJDeR intervention[a] by fathers (n=25) and mothers (n=46).

Measure of user engagement	Fathers		Mothers		*P* value (effect size)[b]
	Statistical value	Range	Statistical value	Range	
Log-data on the Portal					
Minutes working with EJDeR[c], mean (SD)	185.53 (135.71)	18.67-546.51	216.14 (164.44)	8.59-611.74	.43 (*d*=–0.20)
Number of log-ins, mean (SD)	18.08 (14.05)	1-56	21.02 (15.46)	1-72	.43 (*d*=–0.20)
Number of days using EJDeR, mean (SD)	73.04 (33.79)	0-126	78.09 (30.41)	12-126	.52 (*d*=–0.16)
Number of written messages sent to e-therapists, mean (SD)	6.72 (5.93)	0-22	9.50 (8.20)	0-33	.14 (*d*=–0.39)
Number of written messages sent to participants, mean (SD)	26.52 (15.82)	0-66	30.02 (16.53)	7-74	.39 (*d*=–0.22)
Total number of modules opened, mean (SD)	2.28 (0.94)	1-4	2.30 (0.92)	1-4	.92 (*d*=–0.02)
Opened IPE[d] module, n/N (%)	25/25 (100)	—[e]	46/46 (100)	—	—
Completed IPE module, n/N (%)					.57 (w=0.07)
Yes	19/25 (76)	—	32/46 (70)	—	
No	6/25 (24)	—	14/46 (30)	—	
First LICBT[f] module allocated opened, n/N (%)					.36 (w=0.04)
Opened BA[g]	8/20[h] (40)	—	18/34[i] (53)	—	
Opened WM[j]	12/20[h] (60)	—	16/34[i] (47)	—	
Completed first LICBT module allocated					
Completed BA, n/N (%)					.39 (w=0.08)
Yes	7/8 (88)	—	13/18 (72)	—	
No	1/8 (13)	—	5/18 (28)	—	
Completed WM, n/N (%)					.45 (w=0.08)
Yes	5/12 (42)	—	9/16 (56)	—	
No	7/12 (58)	—	7/16 (44)	—	
Opened RP[k] module[l], n/N (%)					.03 (w=0.00)
Yes	8/20 (40)	—	24/34 (71)	—	
No	12/20 (60)	—	10/34 (29)	—	
Completed RP module, n/N (%)					.40 (w=0.07)
Yes	8/8 (100)	—	22/24 (92)	—	
No	0/8 (0)	—	2/24 (8)	—	
Total homework exercises submitted, mean (SD)	2.52 (2.52)	0-9	2.80 (2.94)	0-11	.68 (*d*=–0.10)
Data collected outside the Portal					
Attended initial assessment session, n/N (%)					.73 (w=0.09)
Yes	21/25 (84)	—	40/46 (87)	—	
No	4/25 (16)	—	6/46 (13)	—	
Attended midintervention booster session, n/N (%)					.80 (w=0.09)
Yes	16/25 (64)	—	28/46 (61)	—	

XSL•FO
RenderX

Measure of user engagement	Fathers		Mothers		*P* value (effect size)[b]
	Statistical value	Range	Statistical value	Range	
No	9/25 (36)	—	18/46 (39)	—	

[a]Adherence to EJDeR was conceptualized as users engaging with the intervention in accordance with a priori–defined minimum treatment dose. Not all user engagement variables listed in Table 1 are included in this analysis as some measures of user engagement, that is, number of modules opened, attended initial assessment session, attended midintervention booster session, and inform the minimum treatment dose.

[b]A Bonferroni correction was applied based on 15 tests resulting in an α of .003.

[c]EJDeR: intErnetbaserad sJälvhjälp för förälDrar till barn som avslutat en behandling mot canceR.

[d]IPE: introduction and psychoeducation.

[e]Not applicable.

[f]LICBT: low-intensity cognitive behavioral therapy.

[g]BA: behavioral activation.

[h]Number fathers who opened the first low-intensity cognitive behavioral therapy module allocated (n=20).

[i]Number mothers who opened the first low-intensity cognitive behavioral therapy module allocated (n=34).

[j]WM: worry management.

[k]RP: relapse prevention.

[l]Two participants were provided access to the RP module without completing a low-intensity cognitive behavioral therapy module due to e-therapist error.

Discussion

Principal Findings

This study aimed to (1) describe user engagement with the EJDeR intervention, (2) examine whether sociodemographic and baseline clinical characteristics differed between adherers and nonadherers, (3) examine whether depression and anxiety scores differed between adherers and nonadherers at baseline, (4) examine whether user engagement differed between adherers and nonadherers, and (5) examine whether user engagement differed between fathers and mothers. Overall, 48% (34/71) of participants adhered to the minimum treatment dose. In total, 77% (20/26) of participants starting with BA as the first LICBT module allocated completed the module versus 50% (14/28) starting with WM. There were no significant differences between intervention adherers and nonadherers related to sociodemographic characteristics. Nonadherers who opened BA as the first LICBT module allocated had a significantly higher level of depression symptoms at baseline than adherers (*P*=.04); however, after applying a Bonferroni correction (α=.008), this finding was no longer significant. There were no other significant differences between intervention adherers and nonadherers by depression and anxiety score at baseline (Table 3). There was a significant difference between adherers and nonadherers by first LICBT module allocated (*P*=.04) however this finding was no longer significant after applying a Bonferroni correction (α=.007). Of the 20 nonadherers who opened the first LICBT module allocated, 30% (n=6) opened BA, and 70% (n=14) opened WM as their first LICBT module allocated. The number of log-ins, written messages sent to e-therapists, written messages sent to participants, and homework exercises submitted were significantly higher among intervention adherers than nonadherers. There were no significant differences between fathers and mothers related to user engagement variables.

User Engagement and Adherence

Importantly, adherence differed by the first LICBT module allocated, with 30% (6/20) of nonadherers opening BA versus 70% (14/20) opening WM. One potential explanation for this finding is that BA interventions are simple, straightforward, and easy to comprehend [57,58] and, therefore, may be particularly well-suited to digital delivery. Conversely, the WM module included 2 treatment components [45,46] (problem-solving and worry time), which may have been perceived as more complex and challenging for participants. A recent study examining engagement with a mobile phone app to deliver WM found that 84% (803/956) of users did not achieve the minimum treatment dose [59]. In addition, the LICBT module BA is designed to overcome sources of negative reinforcement and increase engagement with pleasurable, routine, and necessary activities in a structured and graded way. BA may therefore result in more immediate and tangible rewards (eg, improved mood and increased engagement in valued activities) than WM. Indeed, BA is associated with sudden therapeutic gains [60], and a recent component network meta-analysis on iCBT for depression found stronger evidence for BA as an effective intervention component compared with other techniques, for example, cognitive restructuring and problem-solving [61].

However, differences in adherence by the first LICBT module allocated may also reflect the nonrandomized nature of module allocation. Module allocation was informed by the participant's main presenting difficulties as assessed by the e-therapist, with a collaborative decision made with the participant to first start with BA for depression or WM for GAD. As such, the participant's presenting difficulties, rather than module content alone, may have contributed to differences in adherence [62,63]. Overall, baseline clinical characteristics (depression and anxiety scores) did not differ between intervention adherers and nonadherers, similar to recent studies in other populations [64,65]. However, nonadherers who opened BA as the first

XSL•FO
RenderX

LICBT module had a higher level of depression symptoms at baseline compared with adherers, although this finding was no longer significant after applying a Bonferroni correction. There were no other significant differences between intervention adherers and nonadherers by baseline clinical characteristics. The finding that nonadherers who opened BA as the first LICBT module had a higher level of depression symptoms at baseline is somewhat counterintuitive, given the effects of iCBT interventions are larger in those with moderate to severe depression [5]. However, other research has found higher levels of depressive symptoms at baseline to be associated with lower iCBT adherence [66,67] and higher rates of dropout from blended iCBT interventions [68]. Other studies have also found baseline anxiety symptoms to be an important predictor of adherence [69]. While this finding should be treated with caution, given small sample sizes and lack of power, symptoms at baseline may have influenced participants' engagement with and adherence to specific modules, highlighting a potential need to consider baseline characteristics when allocating digital intervention components, for example, specific therapeutic techniques [63,70]. Future research with large sample sizes is warranted to explore the association between baseline clinical characteristics and intervention engagement and adherence.

Importantly, adherers had good levels of engagement with intervention components that were not related to the minimum treatment dose, for example, minutes working with the intervention, number of log-ins, number of written messages sent to e-therapists, number of written messages sent to participants, and the total number of homework exercises submitted. Examining user engagement variables not included in the minimum treatment dose by adherers and nonadherers helped provide a more nuanced understanding of how participants engaged (or not) with the intervention. Defining and measuring adherence to digital interventions in a way that accurately reflects user engagement is challenging, and there is a lack of universally accepted definitions of adherence [23]. Definitions typically vary depending on the intervention's structure, ranging from module completion to log-in frequency to provide evidence of meaningful engagement with therapeutic components [71].

Indeed, defining user engagement and adherence in digital interventions is complex [72] and log-data tracking alone may not indicate a lack of intervention engagement, as participants may engage in intervention techniques offline, leading to higher levels of engagement than recorded by user action logging [64]. Furthermore, participants may have received an active treatment dose before reaching the a priori–determined minimum treatment dose, underscoring the complexities of conceptualizing adherence to digital interventions [18]. There have been calls to move beyond measuring user engagement using only behavioral tools (ie, system usage) and to measure emotional and cognitive engagement components of engagement [26]. For example, user engagement may include the extent of usage (eg, measured by user action logging) and subjective experience (eg, affect, attention, and interest) [20, 25]. In this study, we focused on measures of behavior within the Portal and did not measure subjective experience. Consequently, our findings may not fully capture the multifaceted nature of how participants may have engaged in the intervention and its associated factors. In our planned external pilot RCT, we will include additional measurements of user engagement beyond system usage data, such as validated self-report questionnaires [73,74] that measure affect, behavior, and cognition. However, there remains a need to develop ways of better defining, measuring, and differentiating between the different components of engagement in digital interventions [75].

The provision of support and guidance is another important factor potentially related to intervention adherence [28]. In this study, e-therapists sent an average of 40.21 (SD 12.09) written messages to adheres versus 18.30 (SD 11.98) to nonadherers. Personalized support and guidance are associated with improved retention [61] and effectiveness [5,6] in iCBT interventions. Our findings suggest additional consideration may be needed to train e-therapists to provide personalized support and motivate and encourage nonadherers to improve future intervention adherence. Future research could also look at providing an automated guidance function, for example, through a chatbot, to enhance adherence further [76].

No sociodemographic characteristics differed between intervention adherers and nonadherers. Furthermore, there were no significant differences between fathers and mothers regarding user engagement variables. The wider iCBT literature suggests mixed findings regarding the association between sociodemographic characteristics and adherence [77,78]. Some studies suggest that women and individuals with higher education levels are more likely to adhere to iCBT interventions compared with men and those with lower education levels [21]. However, conflicting results and lack of consistent associations limit the ability to conclude sociodemographic predictors of adherence [77].

Implications

Findings build upon our previously published results from ENGAGE [31]. Analysis of usage data provides a more detailed and nuanced picture of how participants interacted with the intervention. This analysis also identifies key factors associated with engagement and adherence. One important finding is the difference in adherence based on the first LICBT module allocated, with 70% (14/20) of nonadherers who opened their first LICBT module opening WM versus 30% (6/20) of nonadherers opening BA. This suggests early exposure to potentially challenging and complex content may cause users to disengage. While developed using persuasive systems design elements [48,49], for example, tunneling, tailoring, rehearsal, and liking, future adaptations to the EJDeR intervention will need to focus on improving navigation and usability, especially for WM. Flexibility and ease of use of digital interventions for mental health have been identified as effective engagement strategies [75]. Engaging in user-centered design processes, that is, involving users in future system development, may represent a way to improve intervention engagement and adherence [59].

In addition, the WM module's dual-component structure (ie, including 2 LICBT techniques—problem solving for practical worries and worry time for hypothetical worries) may have resulted in higher cognitive load, which could be a barrier to engagement and adherence. Participants first allocated to the

XSL•FO
RenderX

WM module may benefit from additional support and guidance from an e-therapist. For example, e-therapists could be trained to provide more personalized support and feedback [59], which may help minimize challenges and help users understand how to engage with each module effectively [68]. Using visual aids, videos, and immediate feedback can also make content more accessible and motivating, improving adherence and overall user experience [75,76]. Furthermore, acknowledging that module length and complexity can impact adherence, and intervention modifications could also include efforts to streamline module design to reduce cognitive load without compromising therapeutic content. Although parents and e-therapists came to a collaborative decision to allocate the LICBT module to first work with, further intervention modifications could include providing more module-specific information, for example, a module preview, which may also facilitate starting with a module that meets their needs and preferences [79].

Finally, issues with engagement and adherence can arise due to digital interventions adopting a "one-size-fits-all" approach. Instead, interventions require careful tailoring to meet participants' specific needs and context [80]. This is especially important when considering users of digital health interventions engage with these interventions in their own environment, whereby they may be experiencing multiple other demands [81]. iCBT interventions include user demands, such as time to engage in intervention (eg, reading the content, completing homework exercises, and attending support sessions). Lack of time is a common barrier to accessing psychological support reported by parents [82] and for parents of children treated for cancer, particularly [83]. Time was also the most frequently reported burden limiting participation in a group-based CBT videoconferencing program to support parents of children treated for cancer [84]. Parents may also experience reduced opportunities for privacy, further interfering with their ability to engage with the intervention. Other research exploring engagement with digital interventions found users with no children report higher intervention usage [85]. Other factors, including socioeconomic status, may influence how participants engage with interventions [14]. Furthermore, individual contextual factors such as digital literacy and access to technology can influence user engagement [86-88]. Despite the EJDeR intervention being developed alongside parent research partners [38], additional intervention tailoring may be required to further adapt the intervention to the specific needs and context of the population.

Strengths and Limitations

Given that this study is a secondary analysis of data from a single-arm feasibility trial, the sample size is small, and analyses may be underpowered to detect statistically significant differences. Due to the small sample size, we could not conduct a regression or correlation analysis. Multiple statistical comparisons were conducted, increasing the risk of Type I error. However, this risk was minimized by applying a Bonferroni correction to reduce the likelihood of spurious results [55]. Most participants were female, had an education level higher than upper secondary school, and were born in a Nordic country. This may limit the generalizability of findings, especially considering Sweden is a diverse and multicultural society, with 21% (2,170,627/10,551,700) of the population being foreign-born [89], and there is a need to develop culturally responsive digital interventions to enhance acceptability and engagement in ethnic minority populations [90]. The research team defined intervention adherence a priori following a minimum treatment dose. This definition fails to consider that participants may have conducted intervention activities outside the Portal, that is, offline [59]. Some participants may have achieved their personal intervention goals early and stopped engaging with the intervention, and yet were classified as intervention nonadherers [12].

We also adopted an exploratory approach that aimed to examine a broad range of variables potentially associated with user engagement and adherence. We recognize that the exploratory analysis of log-data has significant limitations, for example, lack of transparency, difficulties with replication, and the risk of "data dredging" [29]. While we set specific research aims to inform our approach, we did not use a framework to underpin our analysis. Adopting the framework for Analyzing and Measuring Usage and Engagement Data (AMUsED) [29] to define meaningful usage variables and generate research questions related to factors such as the interventions' design and theoretical underpinning would have strengthened our approach. Recognizing this limitation, we aim to adopt the AMUsED framework to inform usage analysis in our external pilot RCT to prepare for the design and conduct of a future superior RCT. Our exploratory approach also risked circular analysis, for example, examining user engagement variables associated with adherence may be considered circular given our definition of adherence is partially defined by intervention usage. However, as adherence (ie, the minimum treatment dose) was defined as attending sessions with the e-therapist and module completion, this does tell us much about how adherers versus nonadherers engaged with the Platform beyond these 2 metrics. For example, user engagement variables associated with frequency of use (eg, length of time working with EJDeR and number of log-ins) or engagement in intervention content such as homework exercises (eg, number of written homework exercises submitted) provided more nuanced information concerning how adherers versus nonadherers interacted with the intervention beyond the minimum treatment dose.

As a feasibility study, ENGAGE was not designed to examine the effectiveness of the EJDeR intervention, and we did not examine the association between intervention usage and clinical outcomes. This association will be examined in the planned future superiority RCT, should progression to a superiority RCT be warranted after conducting the external pilot RCT. The association between intervention usage and clinical outcomes will be used to further inform our definition of the minimum treatment dose by helping us to understand what optimal level of intervention usage can affect clinical outcomes [28].

Despite these limitations, our study has some important strengths. We examined the relationship between engagement and adherence, considering the therapeutic content of modules (eg, BA or WM) [59]. In addition, we used objective data (eg, log-data tracking on the Portal and data collected outside the Portal by members of the research team) to examine

XSL•FO
RenderX

engagement, overcoming biases associated with self-reported usage data [91].

Conclusions

Our findings provide important information concerning modifications to the EJDeR intervention to enhance user engagement and adherence, which may be related to intervention efficacy. Findings may also have broader applicability to improving user engagement in other iCBT interventions. Our findings suggest simple and straightforward LICBT techniques, such as BA, are well-suited to digital delivery. However, more complex LICBT techniques, such as WM, may need modification to improve ease of use and thus increase user engagement and adherence. Participants using WM or other more complex therapeutic techniques and those with a higher level of depression symptoms at baseline may need additional support from an e-therapist. Finally, engaging in user-centered design processes (ie, involving users in future system development [92]) may represent a way to improve intervention engagement and adherence [59] by improving navigation as well as further tailoring the intervention to the needs and context of parents of children treated for cancer.

Acknowledgments

We wish to thank Ian Horne (Portal Team Member) for performing all data extractions on the Portal.

Data Availability

The data that support the findings of this study are available from the corresponding author (LvE) upon reasonable request.

Conflicts of Interest

None declared.

References

1. Andersson G, Titov N, Dear BF, Rozental A, Carlbring P. Internet-delivered psychological treatments: from innovation to implementation. World Psychiatry 2019;18(1):20-28 [FREE Full text] [doi: 10.1002/wps.20610] [Medline: 30600624]
2. Mamukashvili-Delau M, Koburger N, Dietrich S, Rummel-Kluge C. Efficacy of computer- and/or internet-based cognitive-behavioral guided self-management for depression in adults: a systematic review and meta-analysis of randomized controlled trials. BMC Psychiatry 2022;22(1):730 [FREE Full text] [doi: 10.1186/s12888-022-04325-z] [Medline: 36424570]
3. Eilert N, Enrique A, Wogan R, Mooney O, Timulak L, Richards D. The effectiveness of Internet-delivered treatment for generalized anxiety disorder: an updated systematic review and meta-analysis. Depress Anxiety 2021;38(2):196-219 [FREE Full text] [doi: 10.1002/da.23115] [Medline: 33225589]
4. Pauley D, Cuijpers P, Papola D, Miguel C, Karyotaki E. Two decades of digital interventions for anxiety disorders: a systematic review and meta-analysis of treatment effectiveness. Psychol Med 2023;53(2):567-579 [FREE Full text] [doi: 10.1017/S0033291721001999] [Medline: 34047264]
5. Karyotaki E, Efthimiou O, Miguel C, Bermpohl FMG, Furukawa TA, Cuijpers P, Individual Patient Data Meta-Analyses for Depression (IPDMA-DE) Collaboration, et al. Internet-based cognitive behavioral therapy for depression: a systematic review and individual patient data network meta-analysis. JAMA Psychiatry 2021;78(4):361-371 [FREE Full text] [doi: 10.1001/jamapsychiatry.2020.4364] [Medline: 33471111]
6. Hedman-Lagerlöf E, Carlbring P, Svärdman F, Riper H, Cuijpers P, Andersson G. Therapist-supported Internet-based cognitive behaviour therapy yields similar effects as face-to-face therapy for psychiatric and somatic disorders: an updated systematic review and meta-analysis. World Psychiatry 2023;22(2):305-314 [FREE Full text] [doi: 10.1002/wps.21088] [Medline: 37159350]
7. Johnsen HM, Haddeland K. User evaluation of a therapist-guided internet-delivered treatment program for anxiety disorders: a qualitative study. Internet Interv 2021;25:100389 [FREE Full text] [doi: 10.1016/j.invent.2021.100389] [Medline: 33996506]
8. Thapar S, Nguyen M, Khan BN, Fanaieyan R, Kishimoto V, Liu R, et al. Patient and therapist perceptions of a publicly funded internet-based cognitive behavioral therapy (iCBT) program for ontario adults during the covid-19 pandemic: qualitative study. JMIR Form Res 2024;8:e50113 [FREE Full text] [doi: 10.2196/50113] [Medline: 38373027]
9. Doukani A, Cerga Pashoja A, Fanaj N, Qirjako G, Meksi A, Mustafa S, et al. Organizational readiness for implementing an internet-based cognitive behavioral therapy intervention for depression across community mental health services in Albania and Kosovo: directed qualitative content analysis. JMIR Form Res 2021;5(11):e29280 [FREE Full text] [doi: 10.2196/29280] [Medline: 34723822]
10. Svärdman F, Sjöwall D, Lindsäter E. Internet-delivered cognitive behavioral interventions to reduce elevated stress: a systematic review and meta-analysis. Internet Interv 2022;29:100553 [FREE Full text] [doi: 10.1016/j.invent.2022.100553] [Medline: 35781929]

XSL•FO
RenderX

11. Baltierra NB, Muessig KE, Pike EC, LeGrand S, Bull SS, Hightow-Weidman LB. More than just tracking time: complex measures of user engagement with an internet-based health promotion intervention. J Biomed Inform 2016;59:299-307 [FREE Full text] [doi: 10.1016/j.jbi.2015.12.015] [Medline: 26732997]
12. Sanatkar S, Baldwin PA, Huckvale K, Clarke J, Christensen H, Harvey S, et al. Using cluster analysis to explore engagement and e-attainment as emergent behavior in electronic mental health. J Med Internet Res 2019;21(11):e14728 [FREE Full text] [doi: 10.2196/14728] [Medline: 31778115]
13. O'Brien HL, Roll I, Kampen A, Davoudi N. Rethinking (Dis)engagement in human-computer interaction. Comput. Hum. Behav 2022;128(3):107109. [doi: 10.1016/j.chb.2021.107109]
14. Yardley L, Spring BJ, Riper H, Morrison LG, Crane DH, Curtis K, et al. Understanding and promoting effective engagement with digital behavior change interventions. Am J Prev Med 2016;51(5):833-842. [doi: 10.1016/j.amepre.2016.06.015] [Medline: 27745683]
15. Yeager CM, Benight CC. If we build it, will they come? Issues of engagement with digital health interventions for trauma recovery. Mhealth 2018;4:37 [FREE Full text] [doi: 10.21037/mhealth.2018.08.04] [Medline: 30363749]
16. Borghouts J, Eikey E, Mark G, De Leon C, Schueller SM, Schneider M, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. J Med Internet Res 2021;23(3):e24387 [FREE Full text] [doi: 10.2196/24387] [Medline: 33759801]
17. Flett JAM, Fletcher BD, Riordan BC, Patterson T, Hayne H, Conner TS. The peril of self-reported adherence in digital interventions: a brief example. Internet Interv 2019;18:100267 [FREE Full text] [doi: 10.1016/j.invent.2019.100267] [Medline: 31890620]
18. Sieverink F, Kelders SM, van Gemert-Pijnen JE. Clarifying the concept of adherence to eHealth technology: systematic review on when usage becomes adherence. J Med Internet Res 2017;19(12):e402. [doi: 10.2196/jmir.8578] [Medline: 29212630]
19. Karyotaki E, Riper H, Twisk J, Hoogendoorn A, Kleiboer A, Mira A, et al. Efficacy of self-guided internet-based cognitive behavioral therapy in the treatment of depressive symptoms: a meta-analysis of individual participant data. JAMA Psychiatry 2017;74(4):351-359 [FREE Full text] [doi: 10.1001/jamapsychiatry.2017.0044] [Medline: 28241179]
20. Short CE, DeSmet A, Woods C, Williams SL, Maher C, Middelweerd A, et al. Measuring engagement in eHealth and mHealth behavior change interventions: viewpoint of methodologies. J Med Internet Res 2018;20(11):e292 [FREE Full text] [doi: 10.2196/jmir.9397] [Medline: 30446482]
21. Andersson G, Cuijpers P, Carlbring P, Riper H, Hedman E. Guided Internet-based vs. face-to-face cognitive behavior therapy for psychiatric and somatic disorders: a systematic review and meta-analysis. World Psychiatry 2014;13(3):288-295 [FREE Full text] [doi: 10.1002/wps.20151] [Medline: 25273302]
22. Karin E, Dear BF, Heller GZ, Crane MF, Titov N. "Wish You Were Here": examining characteristics, outcomes, and statistical solutions for missing cases in web-based psychotherapeutic trials. JMIR Ment Health 2018;5(2):e22 [FREE Full text] [doi: 10.2196/mental.8363] [Medline: 29674311]
23. Forbes A, Keleher MR, Venditto M, DiBiasi F. Assessing patient adherence to and engagement with digital interventions for depression in clinical trials: systematic literature review. J Med Internet Res 2023;25:e43727 [FREE Full text] [doi: 10.2196/43727] [Medline: 37566447]
24. Kelders SM, van Zyl LE, Ludden GDS. The concept and components of engagement in different domains applied to eHealth: a systematic scoping review. Front Psychol 2020;11:926 [FREE Full text] [doi: 10.3389/fpsyg.2020.00926] [Medline: 32536888]
25. Perski O, Blandford A, West R, Michie S. Conceptualising engagement with digital behaviour change interventions: a systematic review using principles from critical interpretive synthesis. Transl Behav Med 2017;7(2):254-267 [FREE Full text] [doi: 10.1007/s13142-016-0453-1] [Medline: 27966189]
26. Miller S, Yardley L, Smith P, Weal M, Anderson A, Stuart B, et al. A digital intervention for respiratory tract infections (Internet Dr): process evaluation to understand how to support self-care for minor ailments. JMIR Form Res 2022;6(1):e24239 [FREE Full text] [doi: 10.2196/24239] [Medline: 35044317]
27. Morrison LG, Geraghty AW, Lloyd S, Goodman N, Michaelides DT, Hargood C, et al. Comparing usage of a web and app stress management intervention: an observational study. Internet Interv 2018;12:74-82 [FREE Full text] [doi: 10.1016/j.invent.2018.03.006] [Medline: 30135771]
28. Enrique A, Palacios JE, Ryan H, Richards D. Exploring the relationship between usage and outcomes of an internet-based intervention for individuals with depressive symptoms: secondary analysis of data from a randomized controlled trial. J Med Internet Res 2019;21(8):e12775 [FREE Full text] [doi: 10.2196/12775] [Medline: 31373272]
29. Miller S, Ainsworth B, Yardley L, Milton A, Weal M, Smith P, et al. A framework for analyzing and measuring usage and engagement data (AMUsED) in digital interventions: viewpoint. J Med Internet Res 2019;21(2):e10966 [FREE Full text] [doi: 10.2196/10966] [Medline: 30767905]
30. Eldridge SM, Lancaster GA, Campbell MJ, Thabane L, Hopewell S, Coleman CL, et al. Defining feasibility and pilot studies in preparation for randomised controlled trials: development of a conceptual framework. PLoS One 2016;11(3):e0150205 [FREE Full text] [doi: 10.1371/journal.pone.0150205] [Medline: 26978655]

XSL•FO
RenderX

31. Thiblin E, Woodford J, Reuther C, Lundgren J, Lutvica N, von Essen L. Internet-administered, low-intensity cognitive behavioral therapy for parents of children treated for cancer: a feasibility trial (ENGAGE). Cancer Med 2023;12(5):6225-6243 [FREE Full text] [doi: 10.1002/cam4.5377] [Medline: 36404407]
32. Woodford J, Wikman A, Cernvall M, Ljungman G, Romppala A, Grönqvist H, et al. Study protocol for a feasibility study of an internet-administered, guided, CBT-based, self-help intervention (ENGAGE) for parents of children previously treated for cancer. BMJ Open 2018;8(6):e023708 [FREE Full text] [doi: 10.1136/bmjopen-2018-023708] [Medline: 29903802]
33. Ljungman L, Boger M, Ander M, Ljótsson B, Cernvall M, von Essen L, et al. Impressions that last: particularly negative and positive experiences reported by parents five years after the end of a child's successful cancer treatment or death. PLoS One 2016;11(6):e0157076 [FREE Full text] [doi: 10.1371/journal.pone.0157076] [Medline: 27272318]
34. Ljungman L, Cernvall M, Ghaderi A, Ljungman G, von Essen L, Ljótsson B. An open trial of individualized face-to-face cognitive behavior therapy for psychological distress in parents of children after end of treatment for childhood cancer including a cognitive behavioral conceptualization. PeerJ 2018;6:e4570 [FREE Full text] [doi: 10.7717/peerj.4570] [Medline: 29666751]
35. Wikman A, Kukkola L, Börjesson H, Cernvall M, Woodford J, Grönqvist H, et al. Development of an internet-administered cognitive behavior therapy program (ENGAGE) for parents of children previously treated for cancer: participatory action research approach. J Med Internet Res 2018;20(4):e133 [FREE Full text] [doi: 10.2196/jmir.9457] [Medline: 29669710]
36. Woodford J, Norbäck K, Hagström J, Grönqvist H, Parker A, Arundel C, et al. Study within a trial (SWAT) protocol. Investigating the effect of personalised versus non-personalised study invitations on recruitment: an embedded randomised controlled recruitment trial. Contemp Clin Trials Commun 2020;18:100572 [FREE Full text] [doi: 10.1016/j.conctc.2020.100572] [Medline: 32420511]
37. Woodford J, Wikman A, Einhorn K, Cernvall M, Grönqvist H, Romppala A, et al. Attitudes and preferences toward a hypothetical trial of an internet-administered psychological intervention for parents of children treated for cancer: web-based survey. JMIR Ment Health 2018;5(4):e10085 [FREE Full text] [doi: 10.2196/10085] [Medline: 30563814]
38. Woodford J, Farrand P, Hagström J, Hedenmalm L, von Essen L. Internet-administered cognitive behavioral therapy for common mental health difficulties in parents of children treated for cancer: intervention development and description study. JMIR Form Res 2021;5(7):e22709 [FREE Full text] [doi: 10.2196/22709] [Medline: 34142662]
39. Boucher EM, Raiker JS. Engagement and retention in digital mental health interventions: a narrative review. BMC Digit Health 2024;2(1):1-21. [doi: 10.1186/s44247-024-00105-9]
40. Seidler ZE, Rice SM, Ogrodniczuk JS, Oliffe JL, Dhillon HM. Engaging men in psychological treatment: a scoping review. Am J Mens Health 2018;12(6):1882-1900 [FREE Full text] [doi: 10.1177/1557988318792157] [Medline: 30103643]
41. Zimmer-Gembeck MJ, Rudolph J, Kerin J, Bohadana-Brown G. Parent emotional regulation: a meta-analytic review of its association with parenting and child adjustment. Int. J. Behav. Dev 2021;46(1):63-82. [doi: 10.1177/01650254211051086]
42. Ljungman L, Cernvall M, Grönqvist H, Ljótsson B, Ljungman G, von Essen L. Long-term positive and negative psychological late effects for parents of childhood cancer survivors: a systematic review. PLoS One 2014;9(7):e103340 [FREE Full text] [doi: 10.1371/journal.pone.0103340] [Medline: 25058607]
43. Ljungman L, Hovén E, Ljungman G, Cernvall M, von Essen L. Does time heal all wounds? A longitudinal study of the development of posttraumatic stress symptoms in parents of survivors of childhood cancer and bereaved parents. Psychooncology 2015;24(12):1792-1798 [FREE Full text] [doi: 10.1002/pon.3856] [Medline: 26042579]
44. Farrand P. Behavioural activation: working outside in, rather than inside out. In: Farrand P, editor. Low-Intensity CBT Skills and Interventions. Thousand Oaks, California, United States: SAGE Publications Ltd; 2020:1-408.
45. Small F, Lockwood K. Worry management: a practical solution to a problem of hypotheticals. In: Farrand P, editor. Low-Intensity CBT Skills and Interventions. Thousand Oaks, California, United States: SAGE Publications Ltd; 2020:1-408.
46. Miles G. Problem solving: doing what it says on the tin. In: Farrand P, editor. Low-Intensity CBT Skills and Interventions: A Practitioner's Manual. Thousand Oaks, California, United States: SAGE Publications Ltd; 2020:1-408.
47. Cuijpers P, Huibers M, Ebert DD, Koole SL, Andersson G. How much psychotherapy is needed to treat depression? A metaregression analysis. J Affect Disord 2013;149(1-3):1-13. [doi: 10.1016/j.jad.2013.02.030] [Medline: 23528438]
48. Kelders SM, Kok RN, Ossebaard HC, Van Gemert-Pijnen JEWC. Persuasive system design does matter: a systematic review of adherence to web-based interventions. J Med Internet Res 2012;14(6):e152 [FREE Full text] [doi: 10.2196/jmir.2104] [Medline: 23151820]
49. Oinas-Kukkonen H, Harjumaa M. Persuasive systems design: key issues, process model, and system features. Commun Assoc Inf Syst 2009;24(1):2428.
50. Attribution 4.0 International (CC BY 4.0). Creative Commons. URL: https://creativecommons.org/licenses/by/4.0/ [accessed 2025-01-24]
51. Farrand P, Woodford J. Impact of support on the effectiveness of written cognitive behavioural self-help: a systematic review and meta-analysis of randomised controlled trials. Clin Psychol Rev 2013;33(1):182-195. [doi: 10.1016/j.cpr.2012.11.001] [Medline: 23238023]
52. Werntz A, Amado S, Jasman M, Ervin A, Rhodes JE. Providing human support for the use of digital mental health interventions: systematic meta-review. J Med Internet Res 2023;25:e42864 [FREE Full text] [doi: 10.2196/42864] [Medline: 36745497]

XSL•FO
RenderX

53. Kroenke K, Spitzer RL, Williams JBW. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
54. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
55. Armstrong RA. When to use the bonferroni correction. Ophthalmic Physiol Opt 2014;34(5):502-508. [doi: 10.1111/opo.12131] [Medline: 24697967]
56. Cohen J. A power primer. Psychol Bull 1992;112(1):155-159. [doi: 10.1037//0033-2909.112.1.155] [Medline: 19565683]
57. Ekers D, Webster L, Van Straten A, Cuijpers P, Richards D, Gilbody S. Behavioural activation for depression; an update of meta-analysis of effectiveness and sub group analysis. PLoS One 2014;9(6):e100100 [FREE Full text] [doi: 10.1371/journal.pone.0100100] [Medline: 24936656]
58. Wang X, Feng Z. A narrative review of empirical literature of behavioral activation treatment for depression. Front Psychiatry 2022;13:845138 [FREE Full text] [doi: 10.3389/fpsyt.2022.845138] [Medline: 35546948]
59. Farrand P, Raue PJ, Ward E, Repper D, Areán P. Use and engagement with low-intensity cognitive behavioral therapy techniques used within an app to support worry management: content analysis of log data. JMIR Mhealth Uhealth 2024;12:e47321 [FREE Full text] [doi: 10.2196/47321] [Medline: 38029300]
60. Masterson C, Ekers D, Gilbody S, Richards D, Toner-Clewes B, McMillan D. Sudden gains in behavioural activation for depression. Behav Res Ther 2014;60:34-38. [doi: 10.1016/j.brat.2014.06.008] [Medline: 25049140]
61. Furukawa TA, Suganuma A, Ostinelli EG, Andersson G, Beevers CG, Shumake J, et al. Dismantling, optimising, and personalising internet cognitive behavioural therapy for depression: a systematic review and component network meta-analysis using individual participant data. Lancet Psychiatry 2021;8(6):500-511 [FREE Full text] [doi: 10.1016/S2215-0366(21)00077-8] [Medline: 33957075]
62. Andrade AQ, Beleigoli A, Diniz MDF, Ribeiro AL. Influence of baseline user characteristics and early use patterns (24-Hour) on long-term adherence and effectiveness of a web-based weight loss randomized controlled trial: latent profile analysis. J Med Internet Res 2021;23(6):e26421 [FREE Full text] [doi: 10.2196/26421] [Medline: 34081012]
63. Beintner I, Görlich D, Berger T, Ebert DD, Zeiler M, Herrero Camarano R, ICare Consortium. Interrelations between participant and intervention characteristics, process variables and outcomes in online interventions: a protocol for overarching analyses within and across seven clinical trials in ICare. Internet Interv 2019;16:86-97 [FREE Full text] [doi: 10.1016/j.invent.2018.05.001] [Medline: 30775268]
64. Packer T, Austin N, Lehman M, Douglas SL, Plow M. Factors influencing how informal caregivers of people with multiple sclerosis access and use a curated intervention website: analysis from an RCT. Digit Health 2024;10:20552076241228403 [FREE Full text] [doi: 10.1177/20552076241228403] [Medline: 38343587]
65. Solar C, Halat AM, MacLean RS, Rajeevan H, Williams DA, Krein SL, et al. Predictors of engagement in an internet-based cognitive behavioral therapy program for veterans with chronic low back pain. Transl Behav Med 2021;11(6):1274-1282. [doi: 10.1093/tbm/ibaa098] [Medline: 33098304]
66. Gasslander N, Alfonsson S, Jackalin A, Tengberg C, Håkansson J, Huotari L, et al. Predictors of adherence to an internet-based cognitive behavioral therapy program for individuals with chronic pain and comorbid psychological distress. BMC Psychol 2021;9(1):156 [FREE Full text] [doi: 10.1186/s40359-021-00663-x] [Medline: 34641946]
67. Wenger F, Allenhof C, Schreynemackers S, Hegerl U, Reich H. Use of Random Forest to Predict Adherence in an Online Intervention for Depression Using Baseline and Early Usage Data: Model Development and Validation on Retrospective Routine Care Log Data. JMIR Form Res 2024 Nov 15;8:e53768 [FREE Full text] [doi: 10.2196/53768] [Medline: 39546342]
68. Wu MS, Chen SY, Wickham RE, Leykin Y, Varra A, Chen C, et al. Predicting non-initiation of care and dropout in a blended care CBT intervention: Impact of early digital engagement, sociodemographic, and clinical factors. Digit Health 2022;8:20552076221133760 [FREE Full text] [doi: 10.1177/20552076221133760] [Medline: 36312847]
69. Wallert J, Gustafson E, Held C, Madison G, Norlund F, von Essen L, et al. Predicting adherence to internet-delivered psychotherapy for symptoms of depression and anxiety after myocardial infarction: machine learning insights from the u-care heart randomized controlled trial. J Med Internet Res 2018;20(10):e10754 [FREE Full text] [doi: 10.2196/10754] [Medline: 30305255]
70. Hoorelbeke K, Vervaeke J, Siegle GJ, Baeken C, Koster EH. Individual differences associated with treatment adherence and transfer effects following gamified web-based cognitive control training for repetitive negative thinking. Internet Interv 2022;27:100507 [FREE Full text] [doi: 10.1016/j.invent.2022.100507] [Medline: 35242588]
71. Lipschitz JM, Pike CK, Hogan TP, Murphy SA, Burdick KE. The engagement problem: a review of engagement with digital mental health interventions and recommendations for a path forward. Curr Treat Options Psychiatry 2023;10(3):119-135 [FREE Full text] [doi: 10.1007/s40501-023-00297-3] [Medline: 38390026]
72. Torous J, Michalak EE, O'Brien HL. Digital health and engagement-looking behind the measures and methods. JAMA Netw Open 2020;3(7):e2010918 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.10918] [Medline: 32678446]
73. Kelders SM, Kip H, Greeff J. Psychometric Evaluation of the TWente Engagement with Ehealth Technologies Scale (TWEETS): Evaluation Study. J Med Internet Res 2020 Oct 09;22(10):e17757 [FREE Full text] [doi: 10.2196/17757] [Medline: 33021487]

XSL•FO
RenderX

74. Perski O, Lumsden J, Garnett C, Blandford A, West R, Michie S. Assessing the Psychometric Properties of the Digital Behavior Change Intervention Engagement Scale in Users of an App for Reducing Alcohol Consumption: Evaluation Study. J Med Internet Res 2019 Nov 20;21(11):e16197 [FREE Full text] [doi: 10.2196/16197] [Medline: 31746771]
75. Saleem M, Kühne L, De Santis KK, Christianson L, Brand T, Busse H. Understanding engagement strategies in digital interventions for mental health promotion: scoping review. JMIR Ment Health 2021;8(12):e30000 [FREE Full text] [doi: 10.2196/30000] [Medline: 34931995]
76. Yasukawa S, Tanaka T, Yamane K, Kano R, Sakata M, Noma H, et al. A chatbot to improve adherence to internet-based cognitive-behavioural therapy among workers with subthreshold depression: a randomised controlled trial. BMJ Ment Health 2024;27(1):e300881 [FREE Full text] [doi: 10.1136/bmjment-2023-300881] [Medline: 38199786]
77. Beatty L, Binnion C. A systematic review of predictors of, and reasons for, adherence to online psychological interventions. Int J Behav Med 2016;23(6):776-794. [doi: 10.1007/s12529-016-9556-9] [Medline: 26957109]
78. Christensen H, Griffiths KM, Farrer L. Adherence in internet interventions for anxiety and depression. J Med Internet Res 2009;11(2):e13 [FREE Full text] [doi: 10.2196/jmir.1194] [Medline: 19403466]
79. Radomski AD, Bagnell A, Curtis S, Hartling L, Newton AS. Examining the usage, user experience, and perceived impact of an internet-based cognitive behavioral therapy program for adolescents with anxiety: randomized controlled trial. JMIR Ment Health 2020;7(2):e15795 [FREE Full text] [doi: 10.2196/15795] [Medline: 32022692]
80. Soobiah C, Cooper M, Kishimoto V, Bhatia RS, Scott T, Maloney S, et al. Identifying optimal frameworks to implement or evaluate digital health interventions: a scoping review protocol. BMJ Open 2020;10(8):e037643 [FREE Full text] [doi: 10.1136/bmjopen-2020-037643] [Medline: 32792444]
81. Nahum-Shani I, Shaw SD, Carpenter SM, Murphy SA, Yoon C. Engagement in digital interventions. Am Psychol 2022;77(7):836-852 [FREE Full text] [doi: 10.1037/amp0000983] [Medline: 35298199]
82. Tapp B, Gandy M, Fogliati VJ, Karin E, Fogliati RJ, Newall C, et al. Psychological distress, help‐seeking, and perceived barriers to psychological treatment among Australian parents. Aust. J. Psychol 2020;70(2):113-121. [doi: 10.1111/ajpy.12170]
83. Carlsson T, Kukkola L, Ljungman L, Hovén E, von Essen L. Psychological distress in parents of children treated for cancer: An explorative study. PLoS One 2019;14(6):e0218860 [FREE Full text] [doi: 10.1371/journal.pone.0218860] [Medline: 31226159]
84. Wakefield CE, Sansom-Daly UM, McGill BC, Hetherington K, Ellis SJ, Robertson EG, et al. Providing psychological support to parents of childhood cancer survivors: 'Cascade' intervention trial results and lessons for the future. Cancers (Basel) 2021;13(22):5597 [FREE Full text] [doi: 10.3390/cancers13225597] [Medline: 34830752]
85. Nordholt PU, Christalle E, Zill JM, Dirmaier J. Engagement with a web-based intervention to reduce harmful drinking: secondary analysis of a randomized controlled trial. J Med Internet Res 2020;22(11):e18826 [FREE Full text] [doi: 10.2196/18826] [Medline: 33216008]
86. O'Connor S, Hanlon P, O'Donnell CA, Garcia S, Glanville J, Mair FS. Understanding factors affecting patient and public engagement and recruitment to digital health interventions: a systematic review of qualitative studies. BMC Med Inform Decis Mak 2016 Sep 15;16(1):120 [FREE Full text] [doi: 10.1186/s12911-016-0359-3] [Medline: 27630020]
87. Chen AT, Wu S, Tomasino KN, Lattie EG, Mohr DC. A multi-faceted approach to characterizing user behavior and experience in a digital mental health intervention. J Biomed Inform 2019;94:103187 [FREE Full text] [doi: 10.1016/j.jbi.2019.103187] [Medline: 31026595]
88. O'Brien H, Chen AT, Kaneshiro J, Zaslavsky O. User engagement in an online digital health intervention to promote problem solving. Interac Comput 2024;34(5):355-369.
89. SCB. Utrikes födda i Sverige. 2023. URL: https://www.scb.se/hitta-statistik/sverige-i-siffror/manniskorna-i-sverige/utrikes-fodda-i-sverige/ [accessed 2024-10-07]
90. Eustis EH, LoPresti J, Aguilera A, Schueller SM. Cultural responsivity in technology-enabled services: integrating culture into technology and service components. J Med Internet Res 2023;25:e45409 [FREE Full text] [doi: 10.2196/45409] [Medline: 37788050]
91. Sieverink F, Köhle N, Cheung K, Roefs A, Trompetter HR, Keizer J, et al. Evaluating eHealth. In: Gemert-Pijnen LV, Kelders SM, Kip H, Sanderman R, editors. eHealth research, theory and development: a multi-disciplinary approach. Abingdon: Routlegde; 2018:290-318.
92. Stawarz K, Preist C, Tallon D, Wiles N, Kessler D, Turner K, et al. Design considerations for the integrated delivery of cognitive behavioral therapy for depression: user-centered design study. JMIR Ment Health 2020;7(9):e15972 [FREE Full text] [doi: 10.2196/15972] [Medline: 32880580]

Abbreviations

AMUsED: Analyzing and Measuring Usage and Engagement Data
BA: behavioral activation
CBT: cognitive behavioral therapy
CONSORT: Consolidated Standards of Reporting Trials

XSL•FO
RenderX

EJDeR: internetbaserad självhjälp för föräldrar till barn som avslutat en behandling mot cancer or internet-based self-help for parents of children who have completed cancer treatment
GAD: generalized anxiety disorder
GAD-7: 7-item Generalized Anxiety Disorder scale
iCBT: internet-administered cognitive behavioral therapy
IPE: introduction and psychoeducation
LICBT: low-intensity cognitive behavioral therapy
PHQ-9: 9-item Patient Health Questionnaire
RCT: randomized controlled trial
RP: relapse prevention
U-CARE: Uppsala University Psychosocial Care Programme
WM: worry management

Edited by A Mavragani; submitted 07.10.24; peer-reviewed by A Anderson; comments to author 05.11.24; revised version received 26.11.24; accepted 21.12.24; published 28.01.25.

Please cite as:
Reuther C, von Essen L, Mustafa MI, Saarijärvi M, Woodford J
Engagement With an Internet-Administered, Guided, Low-Intensity Cognitive Behavioral Therapy Intervention for Parents of Children Treated for Cancer: Analysis of Log-Data From the ENGAGE Feasibility Trial
JMIR Form Res 2025;9:e67171
URL: https://formative.jmir.org/2025/1/e67171
doi:10.2196/67171
PMID:

©Christina Reuther, Louise von Essen, Mudassir Imran Mustafa, Markus Saarijärvi, Joanne Woodford. Originally published in JMIR Formative Research (https://formative.jmir.org), 28.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Women Empowered to Connect With Addiction Resources and Engage in Evidence-Based Treatment (WE-CARE)—an mHealth Application for the Universal Screening of Alcohol, Substance Use, Depression, and Anxiety: Usability and Feasibility Study

Krystyna Isaacs[1], PhD; Autumn Shifflett[1], MS; Kajal Patel[1], BDS, MPH; Lacey Karpisek[2], LCSW, MSW, CPH, MPH; Yi Cui[1], MS; Maayan Lawental[2], MSW, PhD; Golfo Tzilos Wernette[3], PhD; Brian Borsari[4,5], PhD; Katie Chang[1], MS; Tony Ma[1], MS

[1]Benten Technologies, Manassas, VA, United States

[2]College of Behavior and Community Sciences, School of Social Work, University of South Florida, Tampa, FL, United States

[3]University of Michigan, Department of Family Medicine, Ann Arbor, MI, United States

[4]Center for Data to Discovery and Delivery Innovation, San Francisco Veterans Affairs Health Care System, San Francisco, CA, United States

[5]Department of Psychiatry and Behavioral Sciences, UCSF Weill Institute for Neurosciences, University of California, San Francisco, San Francisco, CA, United States

Corresponding Author:
Krystyna Isaacs, PhD
Benten Technologies
9408 Grant Ave
Unit 206
Manassas, VA, 20110
United States
Phone: 1 7036625858
Email: krysisaacs@gmail.com

Abstract

Background: Women of childbearing age (aged 18-44 years) face multiple barriers to receiving screening and treatment for unhealthy alcohol and substance use, depression, and anxiety, including lack of screening in the primary care setting and lack of support in accessing care. The Women Empowered to Connect with Addiction Resources and Engage in Evidence-based Treatment (WE-CARE) mobile app was developed to test universal screening with women of childbearing age and linkage to care after an anonymous assessment.

Objective: In this study, we aimed to investigate the feasibility and acceptability of providing anonymous screening instruments through mobile phones for alcohol and substance use, as well as depression and anxiety, for women of childbearing age.

Methods: We used agile development principles based on previous formative research to test WE-CARE mobile health app with women of childbearing age (N=30) who resided in 1 of 6 counties in central Florida. WE-CARE included screening instruments (for alcohol, substance use, depression, and anxiety), a moderated discussion forum, educational microlearning videos, a frequently asked questions section, and resources for linkage to treatment. Individuals were recruited using flyers, academic listserves, and a commercial human subject recruiting company (Prolific). Upon completion of the screening instruments, women explored the educational and linkage to care features of the app and filled out a System Usability Scale to evaluate the mobile health app's usability and acceptability. Postpilot semistructured interviews (n=4) were conducted to further explore the women's reactions to the app.

Results: A total of 77 women downloaded the application and 30 completed testing. Women of childbearing age gave the WE-CARE app an excellent System Usability Scale score of 86.7 (SD 12.43). Our results indicate elevated risk for substance use in 18 of the 30 (60%) participants, 9/18 (50%) also had an elevated risk for anxiety or depression, and 11/18 (61%) had an elevated risk for substance use, anxiety, or depression. Participants reported that WE-CARE was easy to navigate and use but they would have liked to see more screening questions and more educational content. Linkage to care was an issue; however, as none of the women identified as "at-risk" for substance use disorders contacted the free treatment clinic for further evaluation.

XSL•FO
RenderX

Conclusions: The mobile health app was highly rated for acceptability and usability, but participants were not receptive to seeking help at a treatment center after only a few brief encounters with the app. The linkage to care design features was likely insufficient to encourage them to seek treatment. The next version of WE-CARE will include normative scores for participants to self-evaluate their screening status compared with their age- and gender-matched peers and enhanced linkages to care features. Future development will focus on enhancing engagement to improve change behaviors and assess readiness for change.

(JMIR Form Res 2025;9:e62915) doi:10.2196/62915

KEYWORDS

service linkage; digital health; education; mental health; substance use disorder; SUD; alcohol use disorder; chatbot; childbearing women; women; alcohol; substance use; empowerment; evidence-based treatment; usability; feasibility; mobile health; mhealth; app; depression; anxiety; screening; e-screening

Introduction

Alcohol and substance use can have devastating health consequences for women. The physical health risks to women who engage in heavy alcohol and substance use include increased likelihood of injury, overdose, organ damage, and STDs due to unsafe sex practices [1]. In addition, binge drinking (defined as 4 or more drinks on one occasion in women) has reached epic proportions (44%) in young women in the United States [2,3], and the rate is increasing rapidly in women over the age of 35 [2,3]. The number of women dying of opioid overdoses has also increased by 260% between 1999 and 2017 [4]. Substance use during pregnancy can also elevate the risk of miscarriage, fetal alcohol syndrome, and birth defects [5]. The US Preventative Task Force (USPTF) reported that about 50% of US pregnancies are unplanned or unintended, with women between the ages of 18-24 facing a higher risk of unplanned pregnancies [6]. As such, alcohol and substance use during pregnancy can pose a combination of risks and potential consequences for both the mother and baby, including fetal alcohol spectrum disorders [7,8].

While the physical health risks of alcohol and substance use are well-documented, it is equally important to recognize the associated risks of mental health illness. A systematic review revealed that 20%-40% of women with an alcohol use disorder also had an anxiety disorder and showed a strong association with major depressive disorder [9,10]. In addition, women who enter treatment for substance use disorder (SUD) frequently have a personal history of domestic violence, trauma, or sexual assault [11-13]. Domestic violence, trauma, or sexual assault can all have physical consequences, on an individual's physical health which can further impact mental health.

Universal screening for alcohol and substance use in health care settings is an effective way to identify individuals in need of help and support but is not widely implemented. Universal screening is defined as providing the same set of questions regarding alcohol and substance use to all individuals, regardless of past medical history or how the patient presents in a health care setting [14]. By implementing universal screening for all of their patients, including women of childbearing age, health care providers (HCPs) minimize the stigma associated with the inquiries. In 2016, the implementation of universal screening in primary care practices was less than 3% and less than 1% of the HCP subsequently engaged the patient in counseling [15]. Since then, USPTF has recommended universal screening for unhealthy alcohol use for all adults aged 18 years or older within primary care settings, followed by a brief conversation discussing the results [5]. This has resulted in more awareness and use of alcohol and drug screens in private practice settings, but as of 2020, private practice settings have still not met the goals set by the USPTF in 2018 [5,16]. USPTF specified the goals of their recommendation were to reduce overall rates of unhealthy alcohol use and improve individual health and social outcomes resulting from risky use through early intervention [17].

Challenges exist for implementing universal screening and supporting women of childbearing age who are at risk or with alcohol or SUD. For women of childbearing age, one such challenge included a lack of awareness of their risk of misuse as noted in the 2017 Surgeon General's *Facing Addiction Report* [18]. Women aware of personal risk may be reluctant to speak with their HCPs out of fear induced by stigma [19,20]. In some states, women are imprisoned for substance use during pregnancy or their children can be taken away [21]. Even if the women had concerns about personal risk and were looking to make an appointment, many were faced with long wait times to receive care [22]. The National Council for Mental Health Wellbeing revealed that 43% of adults were not able to engage in substance use treatment with 28% of those individuals not being able to get an appointment shortly after requesting care [23]. Rural patients may perceive additional barriers related to living in small, close-knit communities where the confidentiality of screener results is perceived as not being as secure as in an urban setting [24-27]. Anonymous self-screening solutions can overcome many of these barriers.

Challenges also exist for HCPs to implement universal screening and support those at the various stages of risk for alcohol or SUD. One common obstacle is a lack of awareness of recommended screening tools and training on tools and substance use. While 84% of HCPs reported asking young adults about their alcohol use, and 80% of obstetrician-gynecologists asked about alcohol or substance use as part of a visit for women of childbearing age, the majority of HCPs did not use a validated screening instrument [28] such as the Tobacco, Alcohol, Prescription medication, and Other Substance Use Tool (TAPS-1 [29] and the Substance Use Risk Profile–Pregnancy (SURP-P [30]), which are valid and effective in accurately identifying those at risk for substance use [31]. As a result, it is unclear if the risk for alcohol or substance use is being identified in these encounters. In addition, the majority of HCPs are not adequately

XSL•FO
RenderX

trained to administer, interpret results, and take appropriate action based on the results of the screeners [25,32]. In addition, providers in rural communities with these skills often have a heavier patient load than urban doctors and have less time for discussing the screener results, risks, and appropriate next steps [25].

Screening, Brief Intervention, and Referral to Treatment (SBIRT) can be an effective approach in the screening process and referral to treatment for individuals with alcohol or SUD [33]. SBIRT involves screening for alcohol and substance use for severity, followed by a brief intervention (eg, 10-15 minutes) and, if indicated, referral to treatment [33]. SBIRT's effectiveness in reducing alcohol use is well supported across a variety of settings [34]. For example, when SBIRT was implemented in a trauma outpatient clinic, 59% of the patients were identified as being at risk for substance use [33]. SBIRT promotes shared decision-making [35], hinging on clinicians building trust [36], using patient-centered language to reduce stigma [37], and exploring treatment options together. Within primary care settings, SBIRT can help providers identify those suffering from substance use or mental health challenges [38], despite the benefits challenges still exist in implementing SBIRT or other screenings, referrals, and linkage to care in all care settings [39] (eg, gynecologist offices) for women of childbearing age. However, the efficacy of SBIRT's "brief intervention" portion for drug use in primary care and emergency departments is still under consideration [40] with 1 systematic review revealing only moderate supporting evidence that interventions in the emergency department reduced alcohol-related injuries [41], another showing a reduction in visits to emergency departments following the SBIRT intervention [42], and another study showed mixed evidence surrounding the effectiveness of non–face-to-face computerized screening in the emergency department [43].

The brief intervention offered in SBIRT often incorporates motivational interviewing (MI) [44]. MI is a collaborative communication style that evokes and seeks to resolve personal ambivalence and strengthen personal commitment and motivation to change the behavior of interest [45]. The 2021 SAMHSA (Substance Abuse and Mental Health Services Administration) Advisory, *Using Motivational Interviewing in Substance Use Disorder Treatment*, emphasizes the need to discuss each individual's readiness to change and personal history [45]. Engaging clients through the use of MI-consistent skills (eg, reflections and affirmations) is a way to evoke personal goals and values consistent with behavior change, thus increasing intrinsic motivation to achieve a healthier lifestyle. However, learning high-quality MI skills can be labor-intensive [46,47], and as such MI is difficult to provide at scale in primary care settings.

Recently, MI has been used in multiple settings and with multiple populations, including inpatient, emergency rooms, rural settings [48], and other ethnicities to reduce alcohol or substance use [49-51] and MI through text messages has been well received by adults [52,53]. MI has also been applied through digital format using text messaging to address a variety of health behaviors including tuberculosis [54], smoking [55], and alcohol use [56]. Most recently, chatbots, which are conversational tools that frequently use artificial intelligence to provide a more engaging approach, have been used in apps (SoberGrid and CHESS) to support motivational interviewing to encourage action toward recovering from substance use [57-59], but as yet no apps exist that provide screening for SUD and alcohol and related risk factors to encourage change in risky behaviors. The use of electronic screeners (eg, assessments completed on a tablet) combined with motivational interviewing, can yield more accurate self-reports [60-63].

Digital health technology may be able to address many of these challenges. Smartphone ownership rates are high among women and individuals from diverse backgrounds, with 90% of women and 91%-97% of individuals who identify as White, Black, Asian or Hispanic backgrounds having access to a smartphone [64]. Mobile apps are therefore well suited to be used to screen for risk factors in pregnancy [65] and substance use [66] and they can play a crucial role in addressing stigma for those afraid of in-person treatment [67]. With issues related to stigma and fear addressed by the anonymous status, mobile apps can encourage more women to undergo screenings, leading to early identification of at-risk individuals, improved education, and increased treatment options for HCPs [68]. Novel digital strategies to improve mental health screenings, patient engagement, and referral to treatment show preliminary effectiveness [69]. A recent report found that using tablets in the waiting room resulted in nearly twice as many individuals screening for depression than when interviewed by nursing staff [70]. Treloar and colleagues highlight that universal screenings can help reduce stigma by reducing discrimination against certain people and groups [71]. As such, mobile health apps have the potential to support universal screening by alleviating stigma, increasing awareness for change through on-demand education, and reducing fear of seeking support and care.

This study was designed to determine whether women of childbearing age will complete anonymous screenings for substance use, alcohol, depression, and anxiety from their mobile phones. We built on previous formative research and online recruitment approaches, explored the benefits of educational and evidence-based materials and the use of linkage to care options, requested feedback on the usability and acceptability of the app, and used qualitative interviews to determine what future modifications need to be implemented before commercial release.

Methods

Digital Health Tool Design

The Women Empowered to Connect with Addiction Resources and Engage in Evidence-based Treatment (WE-CARE) app was designed to comprise a cross-platform (iOS or Android) mobile app where the participants were asked to complete multiple mental health screeners and then invited to explore each of the accompanying app features. All participants who were assessed as being "at-risk" for either alcohol use disorder or SUD were advised they could contact a treatment center anonymously or to call the center directly. An initial prototype was tested with 30 participants drawn from the central Florida region, recruited through listserves, a company that provides research participants

XSL•FO
RenderX

for a small fee, flyers, and community outreach events. Details of the final design of each of the features to be tested in the pilot are described below.

After extensive formative research [72], the following features were included for pilot testing in the mobile app: e-screenings for substance use, depression, and anxiety; anonymous self-referral to a local treatment center for further evaluation; a moderated forum for women to ask questions; a list of frequently asked questions (FAQs) to get evidence-based answers; educational videos on substance use and treatment-related information; national and local hotlines for those suffering from anxiety, depression, or a combination of the two; a chatbot to determine whether the women at risk had made an appointment with the partnered treatment center to receive a substance use evaluation; and follow up to the chatbot responses with motivational interviewing through a licensed social worker (Figure 1).

Figure 1. Screenshots of WE-CARE features tested in the pilot study. WE-CARE: Women Empowered to Connect with Addiction Resources and Engage in Evidence-based Treatment.

Features Developed

Screening Instruments

Multiple screening instruments to be incorporated in WE-CARE were reviewed, and subject matter experts were consulted to determine the most appropriate ones for testing. In total, 4 screeners were selected to be used within the app. Immediately after downloading the app, all participants had to take all the screeners: (1) SURP-P (if the woman self-identified as pregnant); (2) TAPS-1; (3) Generalized Anxiety Disorder-2 Scale (GAD-2) for anxiety; and (4) Patient Health Questionnaire-2 (PHQ-2) for depression (Textbox 1). A binary risk assessment (elevated risk or low risk) was immediately provided to the participants who could then store the results within the app or in the phone memory (refer to Multimedia Appendix 1). The women were reminded that all their responses would be kept anonymous.

Textbox 1. Tools used to screen for alcohol and substance use risk, as well as anxiety and depression.

Substance Use Risk Profile-Pregnancy [73]

- A 3-question tool designed for use in prenatal clinics with pregnant women to screen for alcohol, marijuana, illegal substances, and nonprescribed use of prescription medications.

Tobacco, Alcohol, Prescription Medication, and Other Substance Use Tool [74]

- Consists of a combined screening tool that assesses users for commonly used substances, eliminating the need for multiple screening and lengthy assessment tools.

Generalized Anxiety Disorder-2 Scale [75]

- A brief questionnaire to determine if someone needs additional screening for generalized anxiety.

Patient Health Questionnaire-2 [76]

- A simple 2-question screen to determine if someone should be considered for additional screening for depression.

If a participant was assessed as having elevated risk for alcohol or substance use, she was offered 4 choices: she could leave their phone number, but no personal information, so the treatment center could call back within 24-48 hours to provide a standardized intake, she could call a dedicated phone line and reach someone at the treatment center directly, she could call a national or local crisis hotline, or she could do nothing. If the woman was not at an elevated risk for substance or alcohol misuse but scored high for risk for depression or anxiety, she was presented with a series of crisis hotline numbers to call, as well as contact information for mental health treatment centers in one of the 6 counties of her choice.

Educational Microlearning Video Production

Short (1-5 min) animated cartoon videos were produced in Vyond [77] to provide information identified in the focus groups as pertinent to the women (Textbox 2). Each video topic was researched using websites at National Institute on Alcohol Abuse and Alcoholism (NIAAA), National Institute on Drug Abuse (NIDA), Substance Abuse and Mental Health Services

XSL•FO
RenderX

Administration (SAMHSA), and Centers for Disease Control and Prevention (CDC). A script involving 2 or more characters was then written to provide a more naturalistic setting and conversational tone and reviewed by subject matter experts for accuracy. A Vyond editor then worked with the producer to provide costumes and a setting tailored to Florida state demographics. To add the personalization desired by the women, the videos were presented in the mobile app based on how an individual answered the screeners, (eg, if women indicated they were at elevated risk for substance use, videos about the treatment process and general information on substance use were pushed to the top of the playlist). Several new videos were created for this specific project and additional videos were repurposed from other projects related to breastfeeding for women with SUD (Textbox 2).

Textbox 2. Eleven short, animated videos created for educational purposes.

Video topics

- What is universal screening for alcohol and drug use?
- Drug treatment center referral process.
- Information for women about marijuana use.
- Self-advocacy in a health care setting.
- Family planning.
- Breastfeeding-how to pump.
- Breastfeeding-opioid use disorder.
- Breastfeeding-marijuana use.
- Breastfeeding-tobacco cigarettes.
- Breastfeeding-prescription medications.
- Breastfeeding-alcohol use.

Moderated Forum

To ensure the app was providing evidence-based answers, a licensed social worker (LSW) was assigned to the moderated forum. Any participant with access to the app could make a post anonymously, but the forum was "moderated" in the sense that the LSW was scheduled to review posts each morning and would only publish ones that were relevant to the group. The LSW also could post responses to questions in the forum.

Frequently Asked Questions

The most popular FAQ topics and subtopics for women wanting information on substance and alcohol use were identified by querying ChatGPT 3 [78]. In the formative research stage [72], the women and HCPs expressed concern over the credibility of the information in both the moderated forum and the FAQs. As such, answers to the questions were manually compiled using source materials published at NIH, CDC, and SAMHSA, which were then reviewed by subject matter experts before being published in the FAQ section.

Follow-up by a Chatbot and Licensed Social Worker

A chatbot was built to engage with any individual evaluated as at-risk within 3-4 days of completing the screeners. The chatbot would ask if the person had made an appointment yet and if so, when? After discussion with the chatbot, if an appointment had not been made, the chatbot would encourage the participant to reach out to the treatment center and then a licensed social worker would follow up with the participant. The social worker would engage the participant in a manner consistent with motivational interviewing, to encourage her to make an appointment with the treatment center for further evaluation.

Additional Resources: National and Local Hotlines and Mental Health Facilities

Links to several national crisis lines were provided in the app such that the participant could click on the button and be immediately connected to the crisis line through her mobile phone. In addition, for those who self-identified as having either high anxiety or high levels of depression, a list of mental health centers, sorted by county, was provided within the app. All of the mental health centers were confirmed to be taking new patients before including them in the list.

Recruitment

Upon thorough in-house testing of all of these app features, participants were recruited to test the application over a 3-month period. Participants for the pilot study were eligible to participate in the study if they were women aged 18-44 years old, able to read English at the 8th-grade level, located within central Florida (1 of 6 counties served by the treatment center), and owned a smartphone device. In addition to recruitment through listserves and contacts in central Florida, a research assistant posted flyers within the community at health care agencies, barber shops, public libraries, and recreation centers. The third approach to recruitment included inviting individuals through Prolific [79], a commercial company that recruits individuals for research projects and surveys.

Demographic information was collected through an online survey instrument embedded in the app and all participants were asked to complete a checklist of the features to test within the app before receiving compensation for participation in the research study. At the end of the app testing, all participants

XSL•FO
RenderX

completed an online version of the System Usability Scale (SUS) and several open-ended questions.

Qualitative Feedback

Upon completion of the pilot study, participants were recruited by email or through direct messaging using their Prolific IDs to participate in a virtual 1:1 semistructured interview to gather more feedback on their experience with the mobile app. Interviewers were aware of whether the woman had given the app a low or high SUS score.

Ethical Considerations

The study was approved by the ADVARRA institutional review board (approval 00054640). All participants signed electronic informed consent documents before enrollment in the study. Participants were compensated for their time.

Results

Demographics

A total of 77 women downloaded the app and a total of 30 women aged 18-44 years old tested the WE-CARE mobile app. The majority of the participants were white (47%) and not of Hispanic or Latino descent (56.7%). The participant racial composition was as follows: 46.7% White, 33.3% unknown, 16.7% mixed race, and 3.3% African American. In total, 20% were of Hispanic ethnicity., 56.7% were not Hispanic or Latino, and the remaining 23% were of unknown ethnicity.

Usability and Acceptability Score for Features Developed

WE-CARE was well received by participants who scored it as EXCELLENT (an average SUS score of 86.7, SD 12.43). Of the individuals who scored "at-risk" for substance use, the average SUS score was slightly lower, at 84.69 (SD 12.3) while the individuals who scored "not at risk" for substance use gave the app an average SUS score of 92.73 (SD 9.14), which would have been considered "BEST IMAGINABLE" (Figure 2).

Figure 2. System Usability Score results for WE-CARE pilot. WE-CARE: Women Empowered to Connect with Addiction Resources and Engage in Evidence-based Treatment.

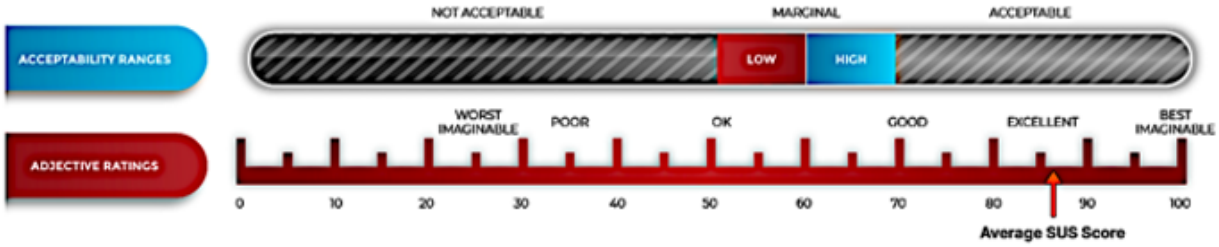

Screening Results

While this was a usability and feasibility study, not a study designed or powered to identify risk for SUDs or mental health illness, we did examine the screening results collected in the pilot study. Out of the 30 individuals who completed the screeners, (18/30, 60%) scored at elevated risk for substance use or alcohol use behaviors, (9/18, 50%) scored at elevated risk for substance use as well as generalized anxiety, (9/18, 50%) scored elevated risk for substance use and depression, and (11/18, 61%) scored at elevated risk for substance use and either high risk for generalized anxiety or depression.

Digital Health Tool Usage Data

Microlearning Video Usage

All of the women had to watch the first video (Universal Screening for Alcohol and Drug Use) but reviewing the remaining videos was optional. Using the frequency of views, it is possible to rank the videos in terms of interest, with Self-Advocacy in the Healthcare Setting being the second most popular video with 6 views. As far as the videos were concerned, participants wanted the videos to be less robotic, shorter in length, and to have more videos relating to coping with stress. Participants were asked what additional information they would like to see included in the app, and 1 participant reported they wanted to learn: "Strategies and interventions that can be used at home …. (Breathing, mindfulness exercises)". Additional suggestions included more information on mental health issues, including "What can I do for fun or to relax instead of drinking or doing drugs? How do I tell if I have an anxiety disorder? What are some ways to reduce anxiety? How do I tell if I have a depressive disorder? What are some ways to alleviate depression?" as well as more information on self-advocacy in the doctor's office.

Moderated Forum Usage

All participants were required to engage with the forum (eg, write a post or give a reaction). The forum was preseeded with some sample questions and answers, and LIKES were scattered over a variety of posts, indicating the women were reviewing the other posts. One person made an anonymous post to which the licensed social worker then responded.

Frequently Asked Questions Usage

The women made extensive use of the FAQs, with the majority of the main topics being explored relating to depression, anxiety, or alcohol use in both the main topics. The most popular subtopics were "How do you define a standard drink" and "How to tell if I might have an anxiety disorder? (Refer to Multimedia Appendix 2: "Number of viewings for FAQ main topics" and Multimedia Appendix 3: "Number of viewings for FAQ

XSL•FO
RenderX

Subtopics" for a complete list of main topics and subtopics and their viewing frequencies).

Chatbot and LSW Follow-Up

Although half the individuals scored elevated risk for substance use, none of the women called or left their number with the treatment center or responded to the chatbot when it reached out to inquire whether they had made an appointment with someone at the treatment center. As such, the licensed social worker at the substance use treatment center was not able to follow up with any of the participants. During the postpilot focus group discussions, we learned one of the 4 individuals interviewed had reached out to a mental health treatment center within their area. As we were not able to track participant interactions with the mental health centers, we were not able to gauge how many of the women with at-risk scores for anxiety or depression sought additional help.

Qualitative Feedback

A total of 4 women completed the postpilot interviews. In total, 50% of the participants were White and 50% were of more than one race. 25% of the participants were Hispanic.

In addition to providing an objective rating of the app's usability and acceptability using the SUS scoring instrument, we asked several open-ended questions to get qualitative feedback on specific features of the application and conducted postpilot interviews. In general, the participants were complimentary about the app. One noted, "I think it's a very simplified app which I actually enjoy. It was a little nostalgic and brought me back to the earlier days of the Internet, especially with the forum included. I did think the screening tests were a little alarming because I wasn't aware I was so depressed. Overall, I enjoyed the app."

Several participants questioned whether asking just 2 questions on either anxiety or depression was sufficient. One participant commented, "I think there needs to be additional screening questions. The issues the app wants to tackle require more than the handful of questions asked." New features suggested by the participants included the ability to add individuals as friends through the app, making the FAQ interactive instead of a list of questions to pick from, and a person-to-person chat option.

When asked whether they had answered honestly when completing the screeners or had been "playing," all 4 participants in the postpilot focus group said they had answered based on their own behaviors. A majority of the participants were not surprised by their scores or alarmed and felt the questions were able to accurately assess their risk for substance or alcohol use, with 1 participant reporting, "would have preferred 4-5 questions and would have been good to ask follow up – for instance, in what circumstances are you using (socially, mental coping skills)…" Another suggestion made by one of the participants involved switching the order of the questions to start with screening for depression and anxiety.

The microlearning videos received more criticism with all participants noting a mix of live-action and cartoon videos would benefit more individuals. The topics of the videos were well received with none of the individuals noting a missed topic. One participant reported, "topics were really important (esp. drinking a glass of wine"). A couple of participants felt the length of the videos was sufficient but could be shortened further to 2-3 minutes.

The FAQs did not receive much interaction from the participants, although the participants did feel having the option to look through the FAQ was beneficial to help sort through a lot of information. On the other hand, participants who reported interacting with the moderated forum felt the feature would be helpful with some additional changes such as having a live professional to contact and adding visuals to the page to make the page more "homey."

To wrap up the discussion, questions were asked of the participants on how they viewed the motivational interviewing chatbot feature and their overall experience. The participants reported they didn't mind the feature being a chatbot, but if a phone call needed to be made or a person was experiencing an emergency, they believed a human should step in at that point. They also desired the chatbot to provide free follow-up services and to reiterate the individual's information will be kept confidential. Overall, participants felt they would recommend this app to another individual, one person reported, "Yes … It was marketed to the right population of people; low stakes in terms of divulging deepest darkest truths-allows a space for help that you might need."

Discussion

Principal Findings

The participants scored the app very highly in terms of usability and acceptability, and enthusiasm was high for the anonymous e-screenings to be offered through a mobile phone app. SAMHSA recommends the use of the SBIRT method to combat the problem of lack of universal screening [80] and this mobile health application successfully provided the screening component of SBIRT. The data collected during the formative research period indicates that WE-CARE can support the SBIRT method by providing universal screening to women aged 18-44 years old and educational materials and referrals to partnering treatment centers, but the prototype is not yet capable of replacing the "brief intervention" component. For example, despite providing access to a free licensed counselor at the treatment center, none of the women who participated in this study were ready to seek changes in their behavior as measured by calling the treatment center for further evaluation. One participant did mention in the postpilot focus group that they did reach out to a mental health facility to address their stress or anxiety, which shows the potential of the app to help women act. The digital handshake, or linkage to care, component of the app, therefore, was not used by the participants in the pilot study. After consultation with subject matter experts, it would appear that expecting the women to seek help after completing 5 minutes' worth of screeners on their phone was likely too high a "bar." Instead, these results suggest that additional engagement and support would be needed to encourage the women to seek more information or treatment. Plans are underway to redesign the prototype to offer more interactions with both a chatbot and

XSL•FO
RenderX

a human counselor to encourage women at risk to engage in the next steps of care.

The educational components of the app (videos, FAQs, and moderated forum) were well received. In the SUS open-ended questions and the postpilot group interviews, there was significant interest in expanding the educational materials to include additional information on coping skills and mental health (with videos, FAQs, PDFS, etc.). Other recommended additional features included creating a function to invite friends to download the app, adding a person-to-person chat option in the moderated forum, and making it possible to write in their questions for a chatbot to respond.

Given the women's willingness to complete the screeners on their mobile phones with guaranteed anonymity, it is possible that the women would be willing to complete the screeners in a physician's waiting room, if they could be assured of confidentiality [28,81-83]. Once a screener indicates a woman is at risk for an alcohol or SUD, the next hurdle is for the HCP to offer resources, make proper referrals, or provide suitable treatment options [84]. HCPs require training on how to effectively care for patients with alcohol or SUDs and need to understand the underlying reasons why women may engage in such behavior, such as past trauma [85]. Upon review of the screening results, the HCPs could create care plans and provide accessible mental health resources to patients based on their specific needs or refer their patients to a current list of clinics accepting new patients. If the latter option, HCPs then need to share with the patient what her journey with a care plan will look like. However, medical school training delivers minimal education in the area of substance use diagnosis and treatment [32] and HCPs have repeatedly noted they feel unprepared to offer SUD treatment or counseling [25,32]. Nonphysician or alternate health care models may be viable approaches to screening [86] and treatment and digital health technologies such as WE-CARE can help to fill in the time and knowledge gaps in the health care workforce.

In this study, over half of the women who completed the screeners in the WE-CARE app scored "at-risk" for SUD or alcohol use disorder and rates for risk for depression and anxiety in the respondents were 50% for each. The average rate for women of childbearing age meeting the *DSM-IV* (*Diagnostic and Statistical Manual of Mental Disorders* [Fourth Edition]) criteria of SUD or being in treatment in the past year for a SUD in the US is 3% for illicit drug use, 1.1% for opioid use, and 7.4% for alcohol use [87]. Similarly, in a study using the same screeners used in WE-CARE (namely the PHQ-2 for depression and the GAD-2 for anxiety), rates of depression in women of childbearing age have been reported to vary from 10% to 27% in the United States (depending on geographical location, race, and ethnicity [88,89]) and to fluctuate based on COVID prevalence [90]. The rates for anxiety in adults over 18 years of age are equally variable, peaking in 2021 at the height of the COVID-19 pandemic, at 35% [91] but before the pandemic, typically averaging around 19% [92]. Rates for anxiety and depression would be expected to be significantly higher in women of childbearing age than in the general population but gender was not reported in this study [92].

The screeners in WE-CARE were selected to identify women who are at risk for risky substance use or mental illness, not to identify substance use or mental health disorders using the criteria of a *DSM-IV* diagnosis per se. But there is a chance these rates may not be inflated. This pilot study was completed in a community that had recently experienced a major hurricane that severely impacted the citizens in 6 counties just 9 months before the study. Increased rates of alcohol and drug use [93] as well as mental illness are not uncommon after natural disasters [94]. Finally, individuals completing the WE-CARE screeners were assured of complete anonymity and therefore the screener results were not verified by individual in-person assessments. When an individual completes an assessment to report risky behavior, it has been demonstrated that the use of assessments completed on a tablet can yield more accurate self-reports [60-63]. However, self-reporting of alcohol or drug use, or concerns with mental illness, may differ in accuracy compared with when a screener is completed in front of a medical professional. The respondent may be confused about the question wording or definitions of terms, or because it was a research study, they may not have given accurate responses. Given the excellent score for usability and acceptability of the app, plans are ongoing to conduct clinical trials where the anonymous screener results will be compared with results from screeners administered by medical professionals to confirm the validity of this approach.

Mobile health apps can facilitate referrals, deliver educational materials to patients and providers, and enhance communication between patients and providers. Universal e-screenings are suitable for delivery before, during, and after an appointment and could remove the barriers associated with HCP time, stigma, willingness, or ability [95] and therefore are well-suited to the mobile health application environment. Other tailored apps exist to assist HCPs in delivering health interventions [96-103] related to substance use [96,104-106] and mental health [97,99,102] that are feasible and acceptable, but more research must be conducted to determine if the screeners are valid when delivered by a mobile health app, and also to ensure that after screening, the woman receives sufficient guidance and support to seek treatment if needed [107,108].

Limitations

A key limitation of this study was the self-selection recruitment method used, which may have impacted the generalizability of the findings. Future studies will include samples from a larger, more diverse population. Furthermore, it is possible that only women who had concerns about their mental health opted to participate in the research study. Also, as discussed above, responses generated through self-report may be flawed if the respondent misunderstands the questions or doesn't answer truthfully. Future versions of this tool will be tested in general clinics, where all patients will be asked to download the app as part of their annual physical and to talk with their health care provider after completing the screeners.

Conclusions

Preliminary results on the usability and feasibility of the WE-CARE mobile health app in providing anonymous screenings for drug and alcohol use, as well as anxiety and

XSL•FO
RenderX

depression, are excellent. Participants readily completed the screeners and were enthused about the opportunity to learn more about a variety of topics tailored for women in their childbearing years. Further research is necessary to validate this approach and to determine how to encourage women who score at risk for significant substance use or mental illness to seek treatment.

Acknowledgments

This research was funded by the National Institute on Alcohol Abuse and Alcoholism (NIAAA; 1R43AA029646-01).

Conflicts of Interest

TM and KC are co-owners of Benten Technologies, the company designing this platform, and will eventually market the Women Empowered to Connect with Addiction Resources and Engage in Evidence-based Treatment (WE-CARE) mobile app. The remaining authors declare that the research was conducted without any commercial or financial relationships that could be construed as a potential conflict of interest. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health, the Department of Veterans Affairs, or the United States Government.

Multimedia Appendix 1
Detailed review of screeners and how they were scored in the WE-CARE app. WE-CARE: Women Empowered to Connect with Addiction Resources and Engage in Evidence-based Treatment.
[DOCX File , 20 KB - formative_v9i1e62915_app1.docx]

Multimedia Appendix 2
Number of viewings for FAQ (frequently asked question) main topics.
[DOCX File , 16 KB - formative_v9i1e62915_app2.docx]

Multimedia Appendix 3
Number of viewings of FAQ (frequently asked question) subtopics.
[DOCX File , 15 KB - formative_v9i1e62915_app3.docx]

References

1. Schulte MT, Hser YI. Substance use and associated health conditions throughout the lifespan. Public Health Rev 2014;35(2). [doi: 10.1007/BF03391702] [Medline: 28366975]
2. McCaul ME, Roach D, Hasin DS, Weisner C, Chang G, Sinha R. Alcohol and women: a brief overview. Alcohol Clin Exp Res 2019;43(5):774-779 [FREE Full text] [doi: 10.1111/acer.13985] [Medline: 30779446]
3. Adams RS, McKetta SC, Jager J, Stewart MT, Keyes KM. Cohort effects of women's mid-life binge drinking and alcohol use disorder symptoms in the United States: impacts of changes in timing of parenthood. Addiction 2023;118(10):1932-1941. [doi: 10.1111/add.16262] [Medline: 37338343]
4. VanHouten JP, Rudd RA, Ballesteros MF, Mack KA. Drug overdose deaths among women aged 30-64 years - United States, 1999-2017. MMWR Morb Mortal Wkly Rep 2019;68(1):1-5 [FREE Full text] [doi: 10.15585/mmwr.mm6801a1] [Medline: 30629574]
5. Curry SJ, Krist AH, Owens DK, Barry MJ, Caughey AB, Davidson KW, et al. Screening and behavioral counseling interventions to reduce unhealthy alcohol use in adolescents and adults: US preventive services task force recommendation statement. JAMA 2018;320(18):1899-1909. [doi: 10.1001/jama.2018.16789] [Medline: 30422199]
6. Unintended pregnancy.: Centers for Disease Control and Prevention Reproductive Health; 2019. URL: https://www.cdc.gov/reproductivehealth/contraception/unintendedpregnancy/index.htm [accessed 2019-12-10]
7. Jacobson JL, Jacobson SW, Sokol RJ. Effects of prenatal exposure to alcohol, smoking, and illicit drugs on postpartum somatic growth. Alcohol Clin Exp Res 1994;18(2):317-323. [doi: 10.1111/j.1530-0277.1994.tb00020.x] [Medline: 8048733]
8. Sokol RJ, Delaney-Black V, Nordstrom B. Fetal alcohol spectrum disorder. JAMA 2003;290(22):2996-2999. [doi: 10.1001/jama.290.22.2996] [Medline: 14665662]
9. Smith JP, Randall CL. Anxiety and alcohol use disorders: comorbidity and treatment considerations. Alcohol Res 2012;34(4):414-431 [FREE Full text] [Medline: 23584108]
10. Kathryn Mchugh R, Weiss RD. Alcohol use disorder and depressive disorders. Alcohol Res 2019;40(1):arcr.v40.1.01 [FREE Full text] [doi: 10.35946/arcr.v40.1.01] [Medline: 31649834]
11. Covington SS. Women and addiction: a trauma-informed approach. J Psychoactive Drugs 2008;Suppl 5:377-385. [doi: 10.1080/02791072.2008.10400665] [Medline: 19248395]
12. ACOG Committee on Health Care for Underserved Women, American Society of Addiction Medicine. ACOG committee opinion No. 524: opioid abuse, dependence, and addiction in pregnancy. Obstet Gynecol 2012;119(5):1070-1076. [doi: 10.1097/AOG.0b013e318256496e] [Medline: 22525931]

XSL•FO
RenderX

13. Alcohol abuse and other sbustance use disorders: ethical issues in obstetric and gynecologic practice.: American College of Obstetricians and Gynecologists; 2015. URL: https://www.acog.org/Clinical-Guidance-and-Publications/Committee-Opinions/Committee-on-Ethics/Alcohol-Abuse-and-Other-Substance-Use-Disorders-Ethical-Issues-in-Obstetric-and-Gynecologic-Practice [accessed 2024-12-30]
14. Screening.: Columbia University Department of Psychiatry URL: https://www.columbiapsychiatry.org/research/research-areas/substance-use-disorders/education/sbirt/about/screening [accessed 2024-12-30]
15. Physical health and mental health.: Mental Health Foundation URL: https://www.mentalhealth.org.uk/explore-mental-health/a-z-topics/physical-health-and-mental-health [accessed 2024-05-23]
16. Woodward D, Wilens TE, Glantz M, Rao V, Burke C, Yule AM. A systematic review of substance use screening in outpatient behavioral health settings. Addict Sci Clin Pract 2023;18(1):18 [FREE Full text] [doi: 10.1186/s13722-023-00376-z] [Medline: 36967381]
17. Curry SJ, Krist AH, Owens DK, Barry MJ, Caughey AB, Davidson KW, et al. Screening and behavioral counseling interventions to reduce unhealthy alcohol use in adolescents and adults: US preventive services task force recommendation statement. JAMA 2018;320(18):1899-1909. [doi: 10.1001/jama.2018.16789] [Medline: 30422199]
18. Department of Health and Human Services Office of the Surgeon General.. Facing addiction in America. The Surgeon General?s Report on Alcohol, Drugs, and Health 2016. [Medline: 28252892]
19. Stone R. Pregnant women and substance use: fear, stigma, and barriers to care. Health Justice 2015;3(1). [doi: 10.1186/s40352-015-0015-5]
20. Muncan B, Walters SM, Ezell J, Ompad DC. "They look at us like junkies": influences of drug use stigma on the healthcare engagement of people who inject drugs in New York City. Harm Reduct J 2020;17(1):53 [FREE Full text] [doi: 10.1186/s12954-020-00399-8] [Medline: 32736624]
21. Substance abuse reporting and pregnancy: the role of the obstetrician-gynecologist.: American College of Obstetricians and Gynecologists; 2011. URL: https://www.acog.org/clinical/clinical-guidance/committee-opinion/articles/2011/01/substance-abuse-reporting-and-pregnancy-the-role-of-the-obstetrician-gynecologist [accessed 2024-12-30]
22. Redko C, Rapp RC, Carlson RG. Waiting time as a barrier to treatment entry: perceptions of substance users. J Drug Issues 2006;36(4):831-852 [FREE Full text] [doi: 10.1177/002204260603600404] [Medline: 18509514]
23. More than 4 in 10 U.S. adults who needed substance use and mental health care did not get treatment.: National Council for Mental Wellbeing; 2022. URL: https://www.thenationalcouncil.org/news/more-than-4-in-10-us-adults-who-needed-substance-use-and-mental-health-care-did-not-get-treatment/ [accessed 2024-12-30]
24. Saunders EC, Moore SK, Gardner T, Farkas S, Marsch LA, McLeman B, et al. Screening for substance use in rural primary care: a qualitative study of providers and patients. J Gen Intern Med 2019;34(12):2824-2832 [FREE Full text] [doi: 10.1007/s11606-019-05232-y] [Medline: 31414355]
25. Brems C, Johnson ME, Warner TD, Roberts LW. Barriers to healthcare as reported by rural and urban interprofessional providers. J Interprof Care 2006;20(2):105-118. [doi: 10.1080/13561820600622208] [Medline: 16608714]
26. Fortney J, Booth BM. Access to substance abuse services in rural areas. Recent Dev Alcohol 2001;15:177-197. [doi: 10.1007/978-0-306-47193-3_10] [Medline: 11449741]
27. Desjarlais-deKlerk K, Wallace JE. Instrumental and socioemotional communications in doctor-patient interactions in urban and rural clinics. BMC Health Serv Res 2013;13:261 [FREE Full text] [doi: 10.1186/1472-6963-13-261] [Medline: 23835062]
28. Anderson BL, Dang EP, Floyd RL, Sokol R, Mahoney J, Schulkin J. Knowledge, opinions, and practice patterns of obstetrician-gynecologists regarding their patients' use of alcohol. J Addict Med 2010;4(2):114-121. [doi: 10.1097/ADM.0b013e3181b95015] [Medline: 21769028]
29. TAPS: tobacco, alcohol, prescription medication, and other substance use tool.: National Institutes of Health URL: https://www.drugabuse.gov/taps/#/ [accessed 2024-12-30]
30. Coleman-Cowger VH, Oga EA, Peters EN, Trocin KE, Koszowski B, Mark K. Accuracy of three screening tools for prenatal substance use. Obstet Gynecol 2019;133(5):952-961 [FREE Full text] [doi: 10.1097/AOG.0000000000003230] [Medline: 30969217]
31. Using standardized tools.: IntellectAbility URL: https://replacingrisk.com/using-standardized-tools/ [accessed 2024-12-30]
32. Recommendations for curricular elements in substance use disorders training.: Substance Abuse and Mental Health Services Administration URL: https://www.samhsa.gov/medications-substance-use-disorders/provider-support-services/recommendations-curricular-elements-substance-use-disorders-training [accessed 2024-12-30]
33. TAP 33: systems-level implementation of screening, brief intervention, and referral to treatment (SBIRT).: Substance Abuse and Mental Health Services Administration; 2013. URL: https://library.samhsa.gov/product/tap-33-systems-level-implementation-screening-brief-intervention-and-referral-treatment [accessed 2024-12-30]
34. Solberg LI, Maciosek MV, Edwards NM. Primary care intervention to reduce alcohol misuse ranking its health impact and cost effectiveness. Am J Prev Med 2008;34(2):143-152. [doi: 10.1016/j.amepre.2007.09.035] [Medline: 18201645]
35. Implementing care for alcohol & other drug use in medical settings: an extension of SBIRT.: National Council for Behavioral Health; 2018. URL: https://www.thenationalcouncil.org/wp-content/uploads/2021/04/

XSL•FO
RenderX

Implementing_Care_for_Alcohol_and_Other_Drug_Use_In_Medical_Settings_-_An_Extension_of_SBIRT.pdf [accessed 2024-12-30]
36. Matthews EB, Savoy M, Paranjape A, Washington D, Hackney T, Galis D, et al. Shared decision making in primary care based depression treatment: communication and decision-making preferences among an underserved patient population. Front Psychiatry 2021;12:681165 [FREE Full text] [doi: 10.3389/fpsyt.2021.681165] [Medline: 34322040]
37. Phelan SM, Salinas M, Pankey T, Cummings G, Allen JP, Waniger A, et al. Patient and health care professional perspectives on stigma in integrated behavioral health: barriers and recommendations. Ann Fam Med 2023;21(Suppl 2):S56-S60 [FREE Full text] [doi: 10.1370/afm.2924] [Medline: 36849477]
38. Hargraves D, White C, Frederick R, Cinibulk M, Peters M, Young A, et al. Implementing SBIRT (Screening, brief intervention and referral to treatment) in primary care: lessons learned from a multi-practice evaluation portfolio. Public Health Rev 2017;38:31 [FREE Full text] [doi: 10.1186/s40985-017-0077-0] [Medline: 29450101]
39. Gomez E, Gyger M, Borene S, Klein-Cox A, Denby R, Hunt S, et al. Using SBIRT (Screen, Brief Intervention, and Referral Treatment) training to reduce the stigmatization of substance use disorders among students and practitioners. Subst Abuse 2023;17:11782218221146391 [FREE Full text] [doi: 10.1177/11782218221146391] [Medline: 36685721]
40. Simioni N, Rolland B, Cottencin O. Interventions for increasing alcohol treatment utilization among patients with alcohol use disorders from emergency departments: a systematic review. J Subst Abuse Treat 2015;58:6-15. [doi: 10.1016/j.jsat.2015.06.003] [Medline: 26206477]
41. Havard A, Shakeshaft A, Sanson-Fisher R. Systematic review and meta-analyses of strategies targeting alcohol problems in emergency departments: interventions reduce alcohol-related injuries. Addiction 2008;103(3):368-377. [doi: 10.1111/j.1360-0443.2007.02072.x] [Medline: 18190671]
42. Pringle JL, Kelley DK, Kearney SM, Aldridge A, Dowd W, Johnjulio W, et al. Screening, brief intervention, and referral to treatment in the emergency department: an examination of health care utilization and costs. Med Care 2018;56(2):146-152. [doi: 10.1097/MLR.0000000000000859] [Medline: 29256973]
43. Biroscak BJ, Pantalon MV, Dziura JD, Hersey DP, Vaca FE. Use of non-face-to-face modalities for emergency department screening, brief intervention, and referral to treatment (ED-SBIRT) for high-risk alcohol use: a scoping review. Subst Abus 2019;40(1):20-32 [FREE Full text] [doi: 10.1080/08897077.2018.1550465] [Medline: 30829126]
44. Miller W, Rollnick S. Motivational Interviewing: Helping People Change and Grow. Fourth. New York: The Guilford Press; 2024.
45. Advisory: using motivational interviewing in substance use disorder treatment (based on TIP 35).: Substance Abuse and Mental Health Services Administration; 2021. URL: https://store.samhsa.gov/product/advisory-using-motivational-interviewing-substance-use-disorder-treatment-based-tip-35 [accessed 2024-12-30]
46. Madson MB, Loignon AC, Lane C. Training in motivational interviewing: a systematic review. J Subst Abuse Treat 2009;36(1):101-109. [doi: 10.1016/j.jsat.2008.05.005] [Medline: 18657936]
47. Kaltman S, Tankersley A. Teaching motivational interviewing to medical students: a systematic review. Acad Med 2020;95(3):458-469. [doi: 10.1097/ACM.0000000000003011] [Medline: 31577585]
48. Dickerson D, Moore LA, Rieckmann T, Croy CD, Venner K, Moghaddam J, et al. Correlates of motivational interviewing use among substance use treatment programs serving American indians/Alaska Natives. J Behav Health Serv Res 2018;45(1):31-45 [FREE Full text] [doi: 10.1007/s11414-016-9549-0] [Medline: 28236017]
49. Martino S, Zimbrean P, Forray A, Kaufman JS, Desan PH, Olmstead TA, et al. Implementing motivational interviewing for substance misuse on medical inpatient units: a randomized controlled trial. J Gen Intern Med 2019;34(11):2520-2529 [FREE Full text] [doi: 10.1007/s11606-019-05257-3] [Medline: 31468342]
50. Ewing SF, Bryan AD, Dash GF, Lovejoy TI, Borsari B, Schmiege SJ. Randomized controlled trial of motivational interviewing for alcohol and cannabis use within a predominantly Hispanic adolescent sample. Exp Clin Psychopharmacol 2022;30(3):287-299 [FREE Full text] [doi: 10.1037/pha0000445] [Medline: 33749294]
51. Balbi AM, Gak AE, Kim ES, Park TD, Quinn JF, Colon MF, et al. Brief motivational interviewing for substance use by medical students is effective in the emergency department. J Emerg Med 2019;57(1):114-117. [doi: 10.1016/j.jemermed.2019.03.032] [Medline: 31027991]
52. Louwagie G, Kanaan M, Morojele NK, Van Zyl A, Moriarty AS, Li J, et al. Effect of a brief motivational interview and text message intervention targeting tobacco smoking, alcohol use and medication adherence to improve tuberculosis treatment outcomes in adult patients with tuberculosis: a multicentre, randomised controlled trial of the ProLife programme in South Africa. BMJ Open 2022;12(2):e056496 [FREE Full text] [doi: 10.1136/bmjopen-2021-056496] [Medline: 35165113]
53. Gex KS, Mun EY, Barnett NP, McDevitt-Murphy ME, Ruggiero KJ, Thurston IB, et al. A randomized pilot trial of a mobile delivered brief motivational interviewing and behavioral economic alcohol intervention for emerging adults. Psychol Addict Behav 2023;37(3):462-474 [FREE Full text] [doi: 10.1037/adb0000838] [Medline: 35482647]
54. Moriarty AS, Louwagie GM, Mdege ND, Morojele N, Tumbo J, Omole OB, et al. ImPROving TB outcomes by modifying LIFE-style behaviours through a brief motivational intervention followed by short text messages (ProLife): study protocol for a randomised controlled trial. Trials 2019;20(1):457 [FREE Full text] [doi: 10.1186/s13063-019-3551-9] [Medline: 31349850]

XSL•FO
RenderX

55. Saiyed A, Layton J, Borsari B, Cheng J, Kanzaveli T, Tsvetovat M, et al. Technology-assisted motivational interviewing: developing a scalable framework for promoting engagement with tobacco cessation using NLP and machine learning. Procedia Computer Science 2022;206:121-131. [doi: 10.1016/j.procs.2022.09.091]
56. Kazemi DM, Borsari B, Levine MJ, Li S, Lamberson KA, Matta LA. A Systematic review of the mHealth interventions to prevent alcohol and substance abuse. J Health Commun 2017;22(5):413-432 [FREE Full text] [doi: 10.1080/10810730.2017.1303556] [Medline: 28394729]
57. Aggarwal A, Tam CC, Wu D, Li X, Qiao S. Artificial intelligence-based chatbots for promoting health behavioral changes: systematic review. J Med Internet Res 2023;25:e40789 [FREE Full text] [doi: 10.2196/40789] [Medline: 36826990]
58. Dulin P, Mertz R, Edwards A, King D. Contrasting a mobile app with a conversational chatbot for reducing alcohol consumption: randomized controlled pilot trial. JMIR Form Res 2022;6(5):e33037 [FREE Full text] [doi: 10.2196/33037] [Medline: 35576569]
59. Almusharraf F, Rose J, Selby P. Engaging unmotivated smokers to move toward quitting: design of motivational interviewing-based chatbot through iterative interactions. J Med Internet Res 2020;22(11):e20251 [FREE Full text] [doi: 10.2196/20251] [Medline: 33141095]
60. Marsch LA, Campbell A, Campbell C, Chen CH, Ertin E, Ghitza U, et al. The application of digital health to the assessment and treatment of substance use disorders: the past, current, and future role of the National drug abuse treatment clinical trials network. J Subst Abuse Treat 2020;112S:4-11 [FREE Full text] [doi: 10.1016/j.jsat.2020.02.005] [Medline: 32220409]
61. Frost H, Campbell P, Maxwell M, O'Carroll RE, Dombrowski SU, Williams B, et al. Effectiveness of motivational interviewing on adult behaviour change in health and social care settings: a systematic review of reviews. PLoS One 2018;13(10):e0204890 [FREE Full text] [doi: 10.1371/journal.pone.0204890] [Medline: 30335780]
62. DiClemente CC, Corno CM, Graydon MM, Wiprovnick AE, Knoblach DJ. Motivational interviewing, enhancement, and brief interventions over the last decade: a review of reviews of efficacy and effectiveness. Psychol Addict Behav 2017;31(8):862-887. [doi: 10.1037/adb0000318] [Medline: 29199843]
63. Lundahl BW, Burke BL. The effectiveness and applicability of motivational interviewing: a practice-friendly review of four meta-analyses. J Clin Psychol 2009;65(11):1232-1245. [doi: 10.1002/jclp.20638] [Medline: 19739205]
64. Mobile fact sheet.: Pew Research Center; 2024. URL: https://www.pewresearch.org/internet/fact-sheet/mobile/ [accessed 2024-12-30]
65. Gance-Cleveland B, Leiferman J, Aldrich H, Nodine P, Anderson J, Nacht A, et al. Using the technology acceptance model to develop StartSmart: mHealth for screening, brief intervention, and referral for risk and protective factors in pregnancy. J Midwifery Womens Health 2019;64(5):630-640. [doi: 10.1111/jmwh.13009] [Medline: 31347784]
66. Glass JE, McKay JR, Gustafson DH, Kornfield R, Rathouz PJ, McTavish FM, et al. Treatment seeking as a mechanism of change in a randomized controlled trial of a mobile health intervention to support recovery from alcohol use disorders. J Subst Abuse Treat 2017;77:57-66 [FREE Full text] [doi: 10.1016/j.jsat.2017.03.011] [Medline: 28476273]
67. Kim HM, Xu Y, Wang Y. Overcoming the mental health stigma through m-Health apps: results from the healthy minds study. Telemed J E Health 2022;28(10):1534-1540. [doi: 10.1089/tmj.2021.0418] [Medline: 35254886]
68. Renker PR. Breaking the barriers: the promise of computer-assisted screening for intimate partner violence. J Midwifery Womens Health 2008;53(6):496-503. [doi: 10.1016/j.jmwh.2008.07.017] [Medline: 18984505]
69. Jardine J, Bowman R, Doherty G. Digital interventions to enhance readiness for psychological therapy: scoping review. J Med Internet Res 2022;24(8):e37851 [FREE Full text] [doi: 10.2196/37851] [Medline: 36040782]
70. Miller DP, Foley KL, Bundy R, Dharod A, Wright E, Dignan M, et al. Universal screening in primary care practices by self-administered tablet vs nursing staff. JAMA Netw Open 2022;5(3):e221480 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.1480] [Medline: 35258581]
71. Treloar C, Cama E, Lancaster K, Brener L, Broady TR, Cogle A, et al. A universal precautions approach to reducing stigma in health care: getting beyond HIV-specific stigma. Harm Reduct J 2022;19(1):74 [FREE Full text] [doi: 10.1186/s12954-022-00658-w] [Medline: 35799296]
72. Shifflett A, Karpisek L, Patel K, Cui Y, Lawental M, Tzilos Wernette G, et al. Women Empowered to Connect with Addiction Resources and Engage in Evidence-based Treatment (WE-CARE): A Usability and Feasibility Study of a Mobile Health Application for the Universal Screening of Alcohol, Substance Use, Depression and Anxiety. JMIR Formative Research. Preprint posted online on December 09, 2024 2024 [FREE Full text] [doi: 10.2196/62915]
73. Yonkers KA, Gotman N, Kershaw T, Forray A, Howell HB, Rounsaville BJ. Screening for prenatal substance use: development of the substance use risk profile-pregnancy scale. Obstet Gynecol 2010;116(4):827-833 [FREE Full text] [doi: 10.1097/AOG.0b013e3181ed8290] [Medline: 20859145]
74. McNeely J, Wu LT, Subramaniam G, Sharma G, Cathers LA, Svikis D, et al. Performance of the tobacco, alcohol, prescription medication, and other substance use (TAPS) tool for substance use screening in primary care patients. Ann Intern Med 2016;165(10):690-699 [FREE Full text] [doi: 10.7326/M16-0317] [Medline: 27595276]
75. Hughes AJ, Dunn KM, Chaffee T, Bhattarai JJ, Beier M. Diagnostic and clinical utility of the GAD-2 for screening anxiety symptoms in individuals with multiple sclerosis. Arch Phys Med Rehabil 2018;99(10):2045-2049 [FREE Full text] [doi: 10.1016/j.apmr.2018.05.029] [Medline: 29964000]

XSL•FO
RenderX

76. Instrument: Patient Health Questionnaire-2 (PHQ-2).: NIDA CTN Common Data Elements URL: https://cde.nida.nih.gov/instrument/fc216f70-be8e-ac44-e040-bb89ad433387 [accessed 2024-12-30]
77. Vyond products: the effortless, all-in-one AI video creation platform.: Vyond URL: https://www.vyond.com/product/ [accessed 2024-12-30]
78. ChatGPT 4o mini.: ChatGPT URL: https://chat.openai.com [accessed 2024-12-30]
79. Quality data. From real people. For faster breakthroughs.: Prolific URL: https://www.prolific.com/ [accessed 2024-12-30]
80. Screening, Brief Intervention, and Referral to Treatment (SBIRT).: Substance Abuse and Mental Health Services Administration URL: https://www.samhsa.gov/sbirt [accessed 2024-12-30]
81. Barry KL, Blow FC, Willenbring ML, McCormick R, Brockmann LM, Visnic S. Use of alcohol screening and brief interventions
in primary care settings: implementation and barriers. Subst Abus 2004;25(1):27-36. [doi: 10.1300/J465v25n01_05] [Medline: 15201109]
82. Kokole D, Mercken L, Jané-Llopis E, Natera Rey G, Arroyo M, Medina P, et al. Perceived appropriateness of alcohol screening and brief advice programmes in colombia, mexico and peru and barriers to their implementation in primary health care - a cross-sectional survey. Prim Health Care Res Dev 2021;22:e4 [FREE Full text] [doi: 10.1017/S1463423620000675] [Medline: 33504413]
83. Im GY, Mellinger JL, Winters A, Aby ES, Lominadze Z, Rice J, et al. Provider attitudes and practices for alcohol screening, treatment, and education in patients with liver disease: A Survey From the American association for the study of liver diseases alcohol-associated liver disease special interest group. Clin Gastroenterol Hepatol 2021;19(11):2407-2416.e8 [FREE Full text] [doi: 10.1016/j.cgh.2020.10.026] [Medline: 33069880]
84. What is SBIRT?.: SBIRT URL: https://sites.education.miami.edu/sbirt/what-is-sbirt/ [accessed 2024-12-30]
85. Tsogia D, Copello A, Orford J. Entering treatment for substance misuse: a review of the literature. J. Mental Health 2001;10(5):481-499. [doi: 10.1080/09638230120041254]
86. Mertens JR, Chi FW, Weisner CM, Satre DD, Ross TB, Allen S, et al. Physician versus non-physician delivery of alcohol screening, brief intervention and referral to treatment in adult primary care: the ADVISe cluster randomized controlled implementation trial. Addict Sci Clin Pract 2015;10:26 [FREE Full text] [doi: 10.1186/s13722-015-0047-0] [Medline: 26585638]
87. Martin CE, Scialli A, Terplan M. Unmet substance use disorder treatment need among reproductive age women. Drug Alcohol Depend 2020;206:107679. [doi: 10.1016/j.drugalcdep.2019.107679] [Medline: 31740208]
88. Mental Health Disorder Statistics.: Johns Hopkins Medicine URL: https://www.hopkinsmedicine.org/health/wellness-and-prevention/mental-health-disorder-statistics [accessed 2024-12-30]
89. Lee B, Wang Y, Carlson SA, Greenlund KJ, Lu H, Liu Y, et al. National, state-level, and county-level prevalence estimates of adults aged ≥18 Years self-reporting a lifetime diagnosis of depression - United States, 2020. MMWR Morb Mortal Wkly Rep 2023;72(24):644-650 [FREE Full text] [doi: 10.15585/mmwr.mm7224a1] [Medline: 37318995]
90. Jia H, Guerin RJ, Barile JP, Okun AH, McKnight-Eily L, Blumberg SJ, et al. National and state trends in anxiety and depression severity scores among adults during the COVID-19 pandemic - United States, 2020-2021. MMWR Morb Mortal Wkly Rep 2021;70(40):1427-1432 [FREE Full text] [doi: 10.15585/mmwr.mm7040e3] [Medline: 34618798]
91. Vahratian A, Blumberg SJ, Terlizzi EP, Schiller JS. Symptoms of anxiety or depressive disorder and use of mental health care among adults during the COVID-19 pandemic - United States, August 2020-February 2021. MMWR Morb Mortal Wkly Rep 2021;70(13):490-494 [FREE Full text] [doi: 10.15585/mmwr.mm7013e2] [Medline: 33793459]
92. Any Anxiety Disorder.: National Institute of Mental Health URL: https://www.nimh.nih.gov/health/statistics/any-anxiety-disorder [accessed 2024-12-30]
93. Alexander AC, Ward KD. Understanding postdisaster substance use and psychological distress using concepts from the self-medication hypothesis and social cognitive theory. J Psychoactive Drugs 2018;50(2):177-186 [FREE Full text] [doi: 10.1080/02791072.2017.1397304] [Medline: 29125424]
94. Newnham EA, Mergelsberg ELP, Chen Y, Kim Y, Gibbs L, Dzidic PL, et al. Long term mental health trajectories after disasters and pandemics: a multilingual systematic review of prevalence, risk and protective factors. Clin Psychol Rev 2022;97:102203. [doi: 10.1016/j.cpr.2022.102203] [Medline: 36162175]
95. Morse BA, Hutchins E. Reducing complications from alcohol use during pregnancy through screening. J Am Med Womens Assoc (1972) 2000;55(4):225-240. [Medline: 10935357]
96. Gordon MS, Carswell SB, Schadegg M, Mangen K, Merkel K, Tangires S, et al. Avatar-assisted therapy: a proof-of-concept pilot study of a novel technology-based intervention to treat substance use disorders. Am J Drug Alcohol Abuse 2017;43(5):518-524. [doi: 10.1080/00952990.2017.1280816] [Medline: 28635345]
97. Gaggioli A, Mantovani F, Castelnuovo G, Wiederhold B, Riva G. Avatars in clinical psychology: a framework for the clinical use of virtual humans. Cyberpsychol Behav 2003;6(2):117-125. [doi: 10.1089/109493103321640301] [Medline: 12804023]
98. Galvão Gomes da Silva J, Kavanagh DJ, Belpaeme T, Taylor L, Beeson K, Andrade J. Experiences of a motivational interview delivered by a robot: qualitative study. J Med Internet Res 2018;20(5):e116 [FREE Full text] [doi: 10.2196/jmir.7737] [Medline: 29724701]

XSL•FO
RenderX

99. Baylor AL. Promoting motivation with virtual agents and avatars: role of visual presence and appearance. Philos Trans R Soc Lond B Biol Sci 2009;364(1535):3559-3565 [FREE Full text] [doi: 10.1098/rstb.2009.0148] [Medline: 19884150]
100. Baccon LA, Chiarovano E, MacDougall HG. Virtual reality for teletherapy: avatars may combine the benefits of face-to-face communication with the anonymity of online text-based communication. Cyberpsychol Behav Soc Netw 2019;22(2):158-165. [doi: 10.1089/cyber.2018.0247] [Medline: 30540490]
101. Dworkin MS, Lee S, Chakraborty A, Monahan C, Hightow-Weidman L, Garofalo R, et al. Acceptability, feasibility, and preliminary efficacy of a theory-based relational embodied conversational agent mobile phone intervention to promote HIV medication adherence in young HIV-positive African American MSM. AIDS Educ Prev 2019;31(1):17-37. [doi: 10.1521/aeap.2019.31.1.17] [Medline: 30742481]
102. Rehm IC, Foenander E, Wallace K, Abbott JM, Kyrios M, Thomas N. What role can avatars play in e-mental health interventions? Exploring new models of client-therapist interaction. Front Psychiatry 2016;7:186 [FREE Full text] [doi: 10.3389/fpsyt.2016.00186] [Medline: 27917128]
103. Canidate S, Hart M. The use of avatar counseling for HIV/AIDS health education: the examination of self-identity in avatar preferences. J Med Internet Res 2017;19(12):e365 [FREE Full text] [doi: 10.2196/jmir.6740] [Medline: 29196281]
104. Tzilos GK, Sokol RJ, Ondersma SJ. A randomized phase I trial of a brief computer-delivered intervention for alcohol use during pregnancy. J Womens Health (Larchmt) 2011;20(10):1517-1524 [FREE Full text] [doi: 10.1089/jwh.2011.2732] [Medline: 21823917]
105. Martino S, Ondersma SJ, Forray A, Olmstead TA, Gilstad-Hayden K, Howell HB, et al. A randomized controlled trial of screening and brief interventions for substance misuse in reproductive health. Am J Obstet Gynecol 2018;218(3):322.e1 [FREE Full text] [doi: 10.1016/j.ajog.2017.12.005] [Medline: 29247636]
106. Ondersma SJ, Beatty JR, Svikis DS, Strickler RC, Tzilos GK, Chang G, et al. Computer-delivered screening and brief intervention for alcohol use in pregnancy: a pilot randomized trial. Alcohol Clin Exp Res 2015;39(7):1219-1226 [FREE Full text] [doi: 10.1111/acer.12747] [Medline: 26010235]
107. Terplan M. Commentary on ondersma et al. (2012): beyond the quest for the perfect test-drug use screening in pregnancy. Addiction 2012;107(12):2107-2108. [doi: 10.1111/j.1360-0443.2012.04094.x] [Medline: 23167646]
108. Pace CA, Gergen-Barnett K, Veidis A, D'Afflitti J, Worcester J, Fernandez P, et al. Warm handoffs and attendance at initial integrated behavioral health appointments. Ann Fam Med 2018;16(4):346-348 [FREE Full text] [doi: 10.1370/afm.2263] [Medline: 29987084]

Abbreviations

AUD: alcohol use disorder
CDC: Centers for Disease Control and Prevention
DSM-IV: Diagnostic and Statistical Manual of Mental Disorders (Fourth Edition)
FAQ: frequently asked question
GAD-2: Generalized Anxiety Disorder-2 Scale
HCP: health care provider
LSW: licensed social worker
MI: motivational interviewing
NIAAA: National Institute on Alcohol Abuse and Alcoholism
NIDA: National Institute on Drug Abuse
PHQ-2: Patient Health Questionnaire-2
SAMHSA: Substance Abuse and Mental Health Services Administration
SBIRT: Screening, Brief Intervention, and Referral to Treatment
SUD: substance use disorder
SURP-P: Substance Use Risk Profile–Pregnancy
SUS: System Usability Scale
TAPS-1: Tobacco, Alcohol, Prescription medication, and Other Substance Use Tool
USPTF: United States Preventative Task Force
WE-CARE: Women Empowered to Connect with Addiction Resources and Engage in Evidence-based Treatment

XSL•FO
RenderX

Edited by A Mavragani; submitted 11.06.24; peer-reviewed by R Yin; comments to author 18.11.24; revised version received 07.12.24; accepted 09.12.24; published 07.02.25.

Please cite as:
Isaacs K, Shifflett A, Patel K, Karpisek L, Cui Y, Lawental M, Tzilos Wernette G, Borsari B, Chang K, Ma T
Women Empowered to Connect With Addiction Resources and Engage in Evidence-Based Treatment (WE-CARE)—an mHealth Application for the Universal Screening of Alcohol, Substance Use, Depression, and Anxiety: Usability and Feasibility Study
JMIR Form Res 2025;9:e62915
URL: https://formative.jmir.org/2025/1/e62915
doi:10.2196/62915
PMID:39918861

©Krystyna Isaacs, Autumn Shifflett, Kajal Patel, Lacey Karpisek, Yi Cui, Maayan Lawental, Golfo Tzilos Wernette, Brian Borsari, Katie Chang, Tony Ma. Originally published in JMIR Formative Research (https://formative.jmir.org), 07.02.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Digital Solution to Support Medication Adherence and Self-Management in Patients with Cancer (SAMSON): Pilot Randomized Controlled Trial

Thu Ha Dang[1,2,3], MIPH, MD; Nilmini Wickramasinghe[4,5,6], MBA, PhD; Prem Prakash Jayaraman[7], PhD; Kate Burbury[8,9,10], PhD; Marliese Alexander[10,11], PhD; Ashley Whitechurch[10,12], MNS; Mitchell Dyer[11], MClinPharm; Stephen Quinn[13], PhD; Abdur Rahim Mohammad Forkan[14], PhD; Penelope Schofield[10,15,16], BSc(Hons), PhD

[1]Department of Psychological Sciences, School of Health Sciences, Swinburne University of Technology, Melbourne, Australia

[2]Health Services Research and Implementation Sciences, Peter MacCallum Cancer Centre, Melbourne, Australia

[3]Digital Health Cooperative Research Centre, Sydney, Australia

[4]Optus Chair Digital Health, La Trobe University, Melbourne, Australia

[5]Department Health and Bio Statistics, School of Health Sciences and Iverson Health Innovation Research Institute, Swinburne University of Technology, Melbourne, Australia

[6]Epworth Healthcare, Melbourne, Australia

[7]Factory of the Future and Digital Innovation Lab, Department of Computer Science and Software Engineering, School Software and Electrical Engineering, Swinburne University of Technology, Melbourne, Australia

[8]Tasmanian Health Services, Department of Health, Hobart, Australia

[9]Digital and Healthcare Innovation, Peter McCallum Cancer Centre, Melbourne, Australia

[10]Sir Peter MacCallum Department of Oncology, The University of Melbourne, Melbourne, Australia

[11]Pharmacy Department, Peter MacCallum Cancer Centre, Melbourne, Australia

[12]Department of Clinical Haematology, Peter MacCallum Cancer Centre, Melbourne, Australia

[13]Department of Health Science and Biostatistics, Swinburne University of Technology, Melbourne, Australia

[14]Digital Innovation Lab, Department of Computer Science and Software Engineering, School Software and Electrical Engineering, Swinburne University of Technology, Melbourne, Australia

[15]School of Computing, Engineering and Mathematical Sciences, La Trobe University, Melbourne, Australia

[16]Department of Psychological Sciences and Iverson Health Innovation Research Institute, Swinburne University of Technology, Melbourne, Australia

Corresponding Author:
Thu Ha Dang, MIPH, MD
Department of Psychological Sciences
School of Health Sciences
Swinburne University of Technology
John Street
Hawthorn
Melbourne, 3122
Australia
Phone: 61 422703347
Email: thuhadang@swin.edu.au

Abstract

Background: Medication nonadherence is a serious problem in cancer, potentially impacts patients' health outcomes and health care costs. Although technology-based medication adherence (MA) interventions have emerged, evidence supporting their quality and effectiveness remains limited.

Objective: This study tested the acceptability, feasibility, and potential effects of Safety and Adherence to Medications and Self-care Advice in Oncology (SAMSON), a digital solution designed to support MA and self-management in cancer.

Methods: A 12-week, 2-arm, unblinded, pragmatic pilot randomized controlled trial was conducted. Adults with hematological malignancies who started oral cancer medicines within the last 12 months were recruited from a metropolitan specialized hospital and randomized 1:1 to SAMSON or control (usual care). The SAMSON solution included a smartphone app with tailored alerts

XSL•FO
RenderX

and real-time self-care advice, a web-based dashboard for health care professionals (HCPs) to monitor patients' adherence and symptoms, and motivational interviewing (MI) teleconsultations delivered by oncology nurses and pharmacists at baseline and weeks 1, 4, 8, and 12. Primary outcomes were the patients' acceptance of SAMSON, measured by the Unified Theory of Acceptance and Use of Technology at 12 weeks, and study feasibility, measured by predefined rates of recruitment, randomization, retention, intervention adherence, and outcome assessment completion. Secondary outcomes were comparison of MA and clinical self-assessments through online questionnaires, including adherence, toxicity self-management, anxiety and depression symptoms, and quality of life, measured at baseline and 12 weeks between the 2 arms. Data retrieved from the SAMSON app (Swinburne University of Technology) was analysed for task completion.

Results: A total of 33 patients (79% of those who were approached) consented to participate in the trial. Of those, 31/33 (94%) completed baseline surveys and were randomized to SAMSON (15/31) and control arms (16/31). Of 31 patients, 28 (90%) completed the 12-week surveys (12 SAMSON and 16 control). Overall, patients rated the SAMSON solution as highly acceptable (13/15, 87% app usage; 14/15, 93% MI teleconsultation delivery). They reported that SAMSON was easy to use (10/12, 83%) and helpful in improving their MA (6/12, 50%). All study HCPs reported the SAMSON solution was helpful in supporting patients' MA. Patients completed an average of 99 tasks over the 12-week study period (71% of scheduled tasks). Most patients (10/12, 83%) completed all 5 scheduled consultations. All study feasibility measures were higher than the predefined upper thresholds, except the rate of patients' responses to medication reminders.

Conclusions: The results demonstrated that the SAMSON solution is acceptable, usable, and useful for oncology HCPs and patients with cancer. The SAMSON solution is feasible in real-life oncology settings. Our next steps involve refining the SAMSON solution based on participants' feedback, conducting a large-scale randomized controlled trial to evaluate its clinical and economic effectiveness, and exploring potential commercialization.

Trial Registration: Australian New Zealand Clinical Trials Registry ACTRN12623000472673; https://www.anzctr.org.au/Trial/Registration/TrialReview.aspx?id=385728

International Registered Report Identifier (IRRID): RR2-10.1136/bmjopen-2023-079122

(JMIR Form Res 2025;9:e65302) doi:10.2196/65302

KEYWORDS

home-based cancer treatment; smartphone app; oral chemotherapy; patient safety; SAMSON; mobile phone; digital solution; medication adherence; self-management; cancer; randomized controlled trial; RCT; pilot study; oncology; mobile health; mHealth; quality of life; eHealth

Introduction

Increasingly, cancer is being treated with self-administered medications [1,2], often involving long and complex treatment regimens [2-4]. A patient's ability to adhere to medications throughout the treatment period is central to the successful delivery of self-administered anticancer regimens [5,6]. However, medication adherence (MA) in cancer is low [7] and often decreases over time [8]. The MA rate is particularly low in hematological cancers, with a variation between 6% [9] and 53% [10]. Improving MA in patients with cancer is crucial, as evidence shows that nonadherence is associated with low survival rates, disease progression, as well as increased health care use and costs [6,11,12].

Given the importance of MA in cancer, there has been an increase in the number of MA interventions in recent times, especially digital interventions [13,14]. With the boom of technology, for instance, smartphones [15], digital health solutions promise more advantages in terms of improved clinical outcomes, and cost efficiency, and they are increasingly accepted by patients [16]. However, evidence on the quality and effectiveness of available MA interventions in cancer remains sparse [17,18]. Our most recent systematic review showed that MA interventions that have multicomponents that are theory- and evidence-based and rigorously designed and evaluated are more likely to be effective [18].

Based on findings from the literature review and the need to address the medication nonadherence problem in cancer, we developed Safety and Adherence to Medications and Self-care Advice in Oncology (SAMSON), a multicomponent digital solution to improve MA. Individual components of the SAMSON solution were co-designed and rigorously developed based on evidence and largely informed by behavioral science research and design science research methodology. The solution comprises 2 components. The first is a smartphone app (SAMSON app) involving individually tailored phone alerts and real-time advice for side effect self-management [19] and a web-based dashboard where patients can track their adherence performance and health care professionals (HCPs) can manage patients' profile, view their adherence performance and survey responses, and manage their medication schedules and related side effects. The second is a motivational interviewing training platform (MITP) to train HCPs in motivational interviewing (MI) techniques to support patient adherence and side effect self-management [20]. After being co-designed and developed, the SAMSON app was tested on end users, specifically patients with hematological cancer [19], and the MITP was tested on HCPs [20]. We hypothesized that the combination of these 2 components (the SAMSON solution) would be broadly acceptable to patients, practically feasible in a busy clinical practice, and potentially effective in improving MA for patients with cancer.

XSL•FO
RenderX

Methods

The CONSORT (Consolidated Standards of Reporting Trials) EHEALTH checklist V 1.6.1 [21] was used to report the present study (Multimedia Appendix 1).

Study Design

We aimed to test the acceptability, feasibility, and potential effects of the SAMSON solution on 30-50 patients with hematological cancer through a 2-armed, unblinded, 12-week, pragmatic pilot randomized controlled trial (RCT) [22]. After providing written consent (Multimedia Appendix 2) and completing baseline questionnaires (Multimedia Appendix 3), participants were randomized to either the intervention group (SAMSON solution) or the control group (usual care) at a 1:1 ratio. After 12 weeks, they completed end-of-study surveys (Multimedia Appendix 3).

Randomization

Using a computer-generated randomization chart, a permuted block randomization of size 4 was used to ensure an even balance of patients in each group throughout the study period. The allocation schedule was generated by a statistician who was blinded to the participants to prevent any predictability when randomizing participants to intervention or control [23].

Patients and Eligibility

Patients were recruited from a metropolitan specialized cancer hospital in Melbourne, Australia, between August 2023 and February 2024. Eligible patients were adults (more than 18 years old), diagnosed with hematological cancer, scheduled to commence oral anticancer medicines (OAMs) or commenced the medication for less than 12 months, willing to have OAMs dispensed at the hospital for the duration of the trial, able to communicate in English, and had access to the internet, a smartphone or computer, and telehealth.

Recruitment

Patients were identified by study site nurses, pharmacists, or treating consultants, who were informed about the study and eligible criteria. At their scheduled consultation, they were asked if they would like their details passed on for contact by the research team. If the patient agreed, they were referred to the study RC. Then, the RC contacts the patient for screening and a comprehensive informed consent process, either in person or online.

Intervention

SAMSON Solution

Patients allocated to the intervention group received the SAMSON solution in addition to their usual care at Peter MacCallum Cancer Centre (PMCC). They were instructed by the RC on how to install the SAMSON app on their smartphone and received login details with a protected password as well as the SAMSON app user manual. Patients' personal and clinical information, such as diagnosis and treatment, extracted from the hospital's electronic medical record system (EMR), was entered in the SAMSON web-based dashboard by the RC and made available in the smartphone app.

Patients were informed that they would receive a teleconsultation (either via phone or telehealth) from a hospital clinical pharmacist in the first 3 days after enrolling in the study, and a maximum of 4 follow-up teleconsultations from a hospital clinical nurse on weeks 1, 4, 8, and 12 of the study. These teleconsultations were on top of the usual care at the hospital. All teleconsultations followed predefined structures and were delivered by hospital clinical nurses and pharmacists who were previously trained in MI using the developed MITP. The initial consultation took 30-60 minutes, aiming to provide education on the OAM(s) that the patient received and the importance of adherence, support the patient in making decisions regarding their medication-taking schedule, and identify and document possible risks and barriers to MA.

Based on the agreed medication-taking schedule, individualized daily medication reminders and weekly side effect surveys were set up in the SAMSON backend platform using the web-based dashboard user interface, so patients can receive them in the installed smartphone app. Medicine information and side effect self-care advice, developed by experienced oncology pharmacists based on available reliable resources and clinician review, and approved by PMCC's Human Research Ethics Committees (HREC), were populated in the SAMSON backend platform using the web-based dashboard. Patients were asked to respond to daily medication reminders and weekly side effect surveys, as well as review self-care advice in the smartphone app. Data on patients' adherence and drug toxicity collected through the SAMSON solution were stored centrally on a secured server, then aggregated, analyzed, and uploaded onto the web-based dashboard so that study nurses and pharmacists could monitor patients' adherence and symptoms. These data were also used by HCPs to tailor their teleconsultations with patients. Patients used the SAMSON app throughout the 12-week period of the study.

The follow-up structured teleconsultations (15-30 minutes in length) aimed to check the patient's understanding of diagnosis, symptoms, self-care strategy, and medications; further explore the patient's facilitators and barriers to MA; motivate the patient's adherence, strengthen their medication self-management skills, and change patient's nonadherence behavior by using MI skills. The quantity and length of these consultations were tailored to the individual patient's need and adherence status.

Intervention nurses and pharmacists had more than 5 years of clinical experience in providing oncology care and successfully completed MI training via the MITP. They were also equipped with instruction manuals on how to conduct teleconsultation and use the SAMSON web-based dashboard. Brief notes were produced and recorded in hospital electronic medical records, as well as sent to the patient at the end of the consultation session.

Usual Care

Patients who were allocated to the control group received usual care. The usual care at the hospital consisted of a clinician consultation, an initial in-person pharmacist consultation (often 5-10 minutes in length), and a phone call follow-up from a clinical nurse within 1-2 weeks after commencing medication.

XSL•FO
RenderX

Measures

Demographic was completed at baseline (t_0). Patients' personal and clinical information was collected from the hospital's electronic medical record system. The study's primary outcomes included the patients' acceptance of SAMSON, measured by the Unified Theory of Acceptance and Use of Technology (UTAUT) [24,25] at 12 weeks (t_1), and study feasibility, measured by predefined rates of recruitment, randomization, retention, intervention adherence, and outcome assessment completion [26,27]. Secondary outcomes included MA, self-reported adherence, toxicity self-management, anxiety, depression and symptoms, and quality of life. Outcomes were assessed at baseline (t_0) and at the end of week 12 (t_1), except MA measured in week 16. All survey data collection (Multimedia Appendix 3) was done through (Research Electronic Data Capture) REDCap (Vanderbilt University) [28].

Primary Outcome Measures

Acceptability

The UTAUT questionnaire was adapted to assess determinants of HCPs' and patients' acceptance and use of the SAMSON solution, including 5 dimensions: performance expectancy, effort expectancy, social influence, facilitating conditions, and behavioral intention [25]. Patients were asked to rate their satisfaction with the SAMSON solution on a 5-point Likert scale. Participants were also invited to provide free-text feedback and suggestions on the solution.

Feasibility

A traffic light approach [29] (Table S1 in Multimedia Appendix 4) was used to determine the feasibility success at 3 levels: (1) feasible (above the upper threshold), (2) infeasible (below the lower threshold), and (3) protocol revision (between the 2 thresholds). The thresholds were defined in consultation with the Steering Committee and based on the literature of similar studies. Informal discussions with study personnel were attempted to obtain feedback on the feasibility.

Secondary Outcome Measures

MA was measured by medication refill adherence (MRA) [30] collected from pharmacy dispensing data. MRA was defined as a percentage calculated from the total days' supply divided by the number of days of study participation and multiplied by 100. In this study, the patient was considered as optimal adherence if their MRA was≥90%.

Self-reported adherence (Adherence Starts with Knowledge 12 [ASK-12]) was measured [31]; this measure included 12 items in 3 subscales: adherence behavior, health beliefs, and inconvenience or forgetfulness.

Toxicity self-management was measured with the Patient Activation Measure-Short Form [32,33], a 13-item self-report measure assessing the patient's knowledge, skills, and confidence in the self-management of their disease and related symptoms.

Anxiety, depression, and symptoms were measured by the Patient-Reported Outcomes Measurement Information System (PROMIS) [34] to assess depression, anxiety, pain interference, fatigue, sleep disturbance, and physical function.

Quality of Life was measured by Functional Assessment of Cancer Therapy-General (FACT-G) [35], which is a 27-item self-report scale measuring the quality of life of patients currently undergoing cancer treatment.

Analysis

Descriptive statistics were used to summarize participant characteristics across study arms, and differences in baseline attributes were assessed using *t* tests or chi-square tests as appropriate.

SAMSON acceptability was analyzed thematically. The UTAUT aims to examine individual quality dimensions, which means it is a suite of scales rather than one quality measure; therefore, adding up the overall scale of the questionnaire is not suitable. Results of UTAUT surveys were summarized for each of the 5 dimensions with the percentage of patients endorsing Likert scale ratings of 3 (disagree, neither agree nor disagree, and agree). Free text answers to UTAUT questionnaires were narratively summarized to gain further insight into acceptability.

Feasibility was determined based on the traffic light approach. Recruitment feasibility was assessed by the number of patients recruited (consented) divided by the number of patients who were approached to join the study. Randomization feasibility was assessed by the number of patients who were randomized divided by the number of patients who consented. Retention in both arms of the study was assessed by the number of patients who remained at the end of the study divided by the total number of patients who consented to join the study. We also tracked intervention adherence, for example, the percentage of patients who completed tasks on the SAMSON smartphone app and received MI teleconsultations. Compliance data for survey completion was calculated as the percentage of patients who completed surveys at t_0 and $t_1 t_1$ out of the total patients in the study at these time points.

Secondary outcomes were analyzed using linear regression. The dependent variable was the outcome at follow-up and the independent variables were arm assignment and the outcome at baseline. Standard checks for normality and homoscedasticity of residuals were conducted for each outcome. All analyses were conducted using StataNow 18 (StataCorp) [36].

Ethical Considerations

A steering committee—which included chief investigators; experts in the fields of digital health, information technology, nursing, pharmacy, psychology, and oncology; and a patients' representative—was formed to provide support for the study. This study was approved by the HRECs of PMCC (number HREC/95332/PMCC) and Swinburne University of Technology (number 20237273-15836) (Multimedia Appendices 5 and 6). Patients reviewed study details and indicated their consent using e-consent forms. Patients were encouraged to contact the study team if they had any questions or concerns. Patients' personal and health information in the SAMSON app was encrypted in transit and stored in a secure server at Swinburne. Study team access to patient's data on the SAMSON web-based dashboard

XSL•FO
RenderX

was password protected and limited to the research coordinator (RC) and 4 study nurses and pharmacists. All data analyses were conducted on deidentified data. Patients received a US $31 gift voucher if they completed all surveys in the study.

Results

Patient Characteristics

Among 42 patients who were approached, 33/42 (79%) consented to participate, and 31/33 (94%) completed baseline (t_0) surveys. Of those who completed baseline surveys, 3/31 patients (all from the intervention arm) withdrew from the study (10 %) due to the burden of the disease. All the remaining 28 patients (100%) completed week-12 surveys (12 intervention and 16 control) (Figure 1).

Demographics of randomized patients are provided in Table S2 in Multimedia Appendix 7. There were no significant differences in participants' demographics between the 2 arms.

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) diagram of Safety and Adherence to Medications and Self-care Advice in Oncology (SAMSON) pilot randomized controlled trial.

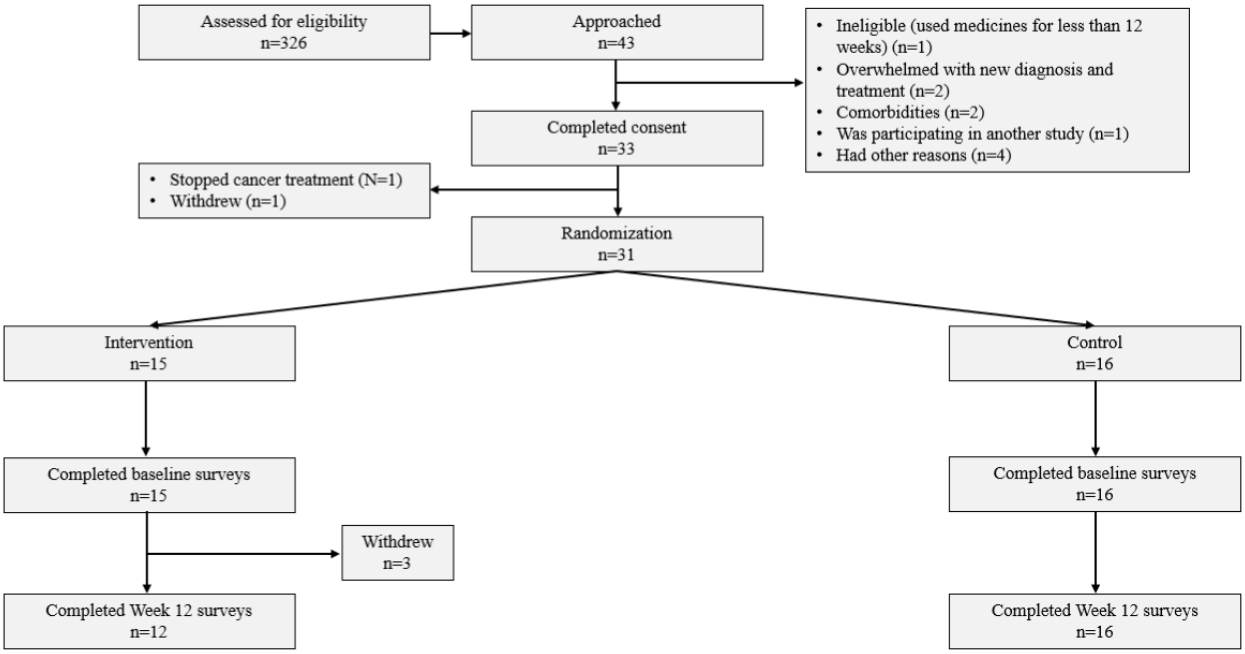

Acceptability Results (Primary Outcome 1)

Overview

Out of 15 intervention arm patients, 13/15 (87%) installed the SAMSON app on their phone for use (2 patients withdrew). All 15 patients in the intervention arm received the initial pharmacy teleconsultation. Most patients (14/15, 93%) received at least 1 nurse teleconsultation, and 10 of them (10/15, 67%) completed all 4 scheduled nurse teleconsultations.

Data retrieved from the SAMSON app showed moderate engagement among participants. Patients completed an average of 99 app tasks (including responses to daily medication reminders and weekly side effect surveys) over the 12-week study period, which accounted for 99/140, 71% of scheduled tasks. A total of 36 severe episodes of symptoms were reported.

Out of 15 patients, 12/15 (80%) and 3 study HCPs (100%) completed the UTAUT surveys. A summary of patients' and HCPs' responses to the UTAUT questionnaire is presented below, with full details in Multimedia Appendices 8 and 9. Figure 2 presents patients' and HCPs' key opinions about the SAMSON solution.

XSL•FO
RenderX

Figure 2. Patients' and health care professionals' opinions about Safety and Adherence to Medications and Self-care Advice in Oncology (SAMSON) solution (n=12 and n=3, respectively).

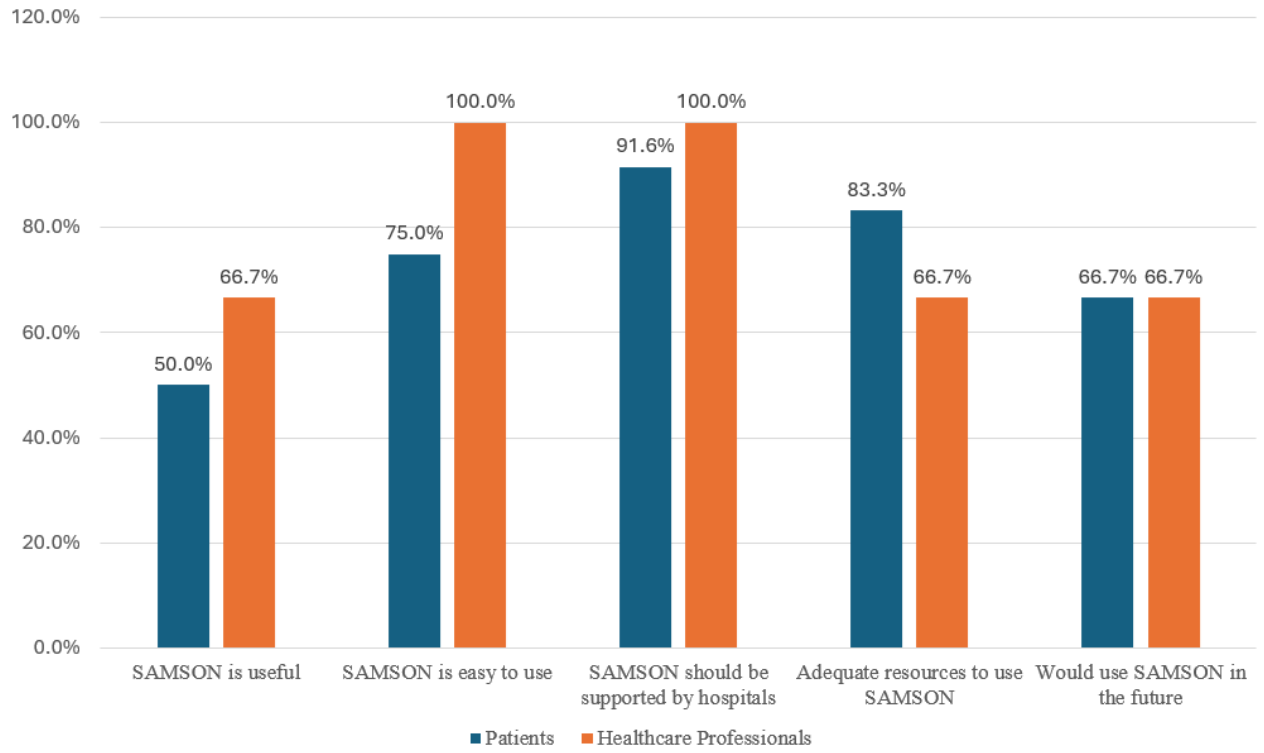

Performance Expectancy

Half of the patients (6/12) in the intervention arm reported that their MA could improve with the help of the SAMSON solution. Patients used different MA-supporting mechanisms provided by SAMSON solution, including prompting reminders (7/12, 58%), disease and treatment education (10/12, 83%), improving confidence in treatment (6/12, 50%), and improving side effect self-management skills (9/12, 75%). Despite this, 3/12 patients (25%) found the SAMSON app component was not that helpful. Of those, 2 were using another commercial MA app with greater functionality. Some patients (4/12, 33%) reported functional issues of the app occurring in a short period of time (1-3 days), for example, medication reminders were not delivered or responses to reminders were not saved properly.

Of the 3 HCPs participating in the trial, 2 found the SAMSON solution useful in their job. All HCPs agreed that SAMSON could enable their 2-way communication with patients which would increase their ability in supporting patient treatment adherence. However, 1 HCP suggested that the SAMSON dashboard's visual design could be improved to be more appealing. HCPs' comments on the time required for MI teleconsultations were controversial: some suggested that the scheduled time was good, while it was reported as quite long for another, which might be a constraint in clinical practice.

Effort Expectancy

The SAMSON app was found to be easy to use by most patients (10/12, 83%). Its presentation was clear, and the content was easy to understand (11/12, 92%). Most participants could easily navigate the app (10/12, 83%), yet 1 (1/12, 8%) struggled in responding to surveys in the app. Some issues with the app's functionality were reported, for example, the medication reminders' time was not automatically updated after the daylight saving time changed, and the app was frozen sometimes, which affected patients' responses to reminders.

All HCPs reported that SAMSON was quite easy to use. However, some issues with the SAMSON web-based dashboard were reported, for example, it did not allow more than 1 HCP to monitor the patient, so all study HCPs had to share 1 account. Two (2/3, 67%), HCPs found implementing MI consultations quite challenging in practice due to time constraints.

Social Influence

Out of 12 patients, 5 (41.6%) thought that their family and friends would support their use of SAMSON, while most of them (10/12, 83%) thought that other patients with cancer would find SAMSON valuable. Almost all patients (11/12, 92%) desired SAMSON to be available for use in cancer hospitals.

All HCPs commented that their colleagues would find the SAMSON valuable and desire to receive the hospital's support in implementing SAMSON in daily practice.

Facilitating Conditions

Most participants were confident that they had adequate knowledge and resources (10/12, 83%) to use SAMSON and could access support when needed (9/12, 75%). Nevertheless, 2/12 patients (17%) reported their difficulties when dealing with some technical issues when using the app.

Despite all HCPs reporting that they had adequate knowledge to apply MI techniques in teleconsultations, 1/3 (33%) wished to have more training resources on the SAMSON app.

XSL•FO
RenderX

Behavioral Intention

Most patients (9/12, 75%) felt confident using the SAMSON app. They would want to continue using SAMSON after the study finishes (8/12, 67%) or recommend it to their peers (7/12, 58%). The 3 patients (25%) who were using another commercial MA app with more functionalities would prefer to receive MI teleconsultations alone. Cost was reported as an important factor influencing the intention to use SAMSON by the majority of patients (8/12, 67%).

Regarding the open-ended questions in the UTAUT survey on experience and perception of the SAMSON solution, patients valued the SAMSON smartphone app, because it was "easy to use and prompted reminders" (P31), "placed all medication resources in one place" (P16), "informative and helpful" (P25), and provided "help quickly if [the patient has] any queries" (P30). The "side-effects" tab within the app, providing self-care advice, was reported to make the patient "feel more secure in managing medication regimens" (P2), which could result in high adherence—as one patient commented, "I didn't miss a dose" (P7). The number of teleconsultations, duration, and quality were reported by most patients as good or "perfect" (P3) and "at the right length of time and allowed me [the patient] to ask all questions I need while allowing the nurse to gather all information" (P2); however, 1 mentioned it was "too long" (P31). Patients provided some helpful suggestions to improve the convenience of using the SAMSON smartphone app, including fixing glitches, combining all medications scheduled at the same time in 1 reminder (rather than separate reminders), having more options for responding to reminders, and more attractive presentation.

Over 60% of study HCPs felt confident in delivering SAMSON to patients in the trial and would like to continue using it in the future. All of them would introduce the solution to their peers to use.

Feasibility Results (Primary Outcome 2)

The study recruitment rate was 33/42 (79%). Most participants went through the information and consent process via phone, except 4 were recruited on-site. The randomization rate was 31 of 33 (94%). The retention rate was 28 of 33 (85%). The RC conducted phone calls or in-person check-ups at least 2 times during the 12-week study. All recruitment, randomization, and retention rates were higher than the upper threshold.

A total of 13 intervention arm patients (13/15, 87%) used the SAMSON app. The proportions of responses to medication reminders and side effect surveys compared to the scheduled tasks during the 12-week study period were 1061 of 1541 (68.9%) and 128 of 141 (90.8%), respectively. The unmet threshold of responses to medication reminders could be due to the app's functional issues as reported earlier. In total, 65 MI teleconsultations were delivered from August 2023 to June 2024, of which 35 of 65 (54%) were conducted on time as scheduled, 26 (40%) were later than scheduled, and 4 (6%) were missed. The average length of initial pharmacy consultations was 40 minutes, while nurse follow-up consultations were 40 minutes on average. Reasons for delayed and missed consultation sessions were mostly on the patient side, including not showing up at the appointment (25/30, 83%), not responding to HCP phone calls (4/30, 13%), or international travel (1/30, 3%). Only 2 sessions (7%) were delayed, because the HCP could not match their work schedule. Of the 12 intervention arm patients who were retained until the end of the study, 10 of 12 (83%) received all 5 scheduled teleconsultations, 1 (8%) received 4 teleconsultations and 1 (8%) received 2 teleconsultations.

Baseline surveys were completed by 31/33 (94%) patients. All who stayed until the end of the study (28/28, 100%) completed week-12 surveys. Participants completed all surveys online without any need for support. An alert email was sent by RC to all participants a few days before the survey due date. However, over one-third of participants (10/28, 36%) only completed surveys after being reminded. Details of feasibility results are presented in Table S3 in Multimedia Appendix 10.

Preliminary Efficacy Results (Secondary Outcomes)

Medication Refill Adherence

Of 28 patients who completed the study, 25/28 (89%) had a 100% adherence rate and 3/28 (11%) had over 95% adherence rate. There is no difference in the proportion of patients who had optimal adherence (MRA ≥90%) between the intervention and control groups (100% for both groups).

The mean and 95% CI of ASK-12, Patient Activation Measure-Short Form, PROMIS, and Functional Assessment of Cancer Therapy-General at baseline and week 12 are shown in Table S4 in Multimedia Appendix 11.

Discussion

Principal Findings

In this study, we aimed to test the acceptability, feasibility, and preliminary efficacy of a multicomponent MA solution, SAMSON, to help patients with cancer improve their adherence to OAMs and self-manage their physical and emotional symptoms. Overall, patients and oncology HCPs rated SAMSON as highly acceptable, usable, and useful. These high levels of user satisfaction evidence that the solution meets the various needs of support among patients with cancer to medically adhere to and manage side effects at home [37], as well as the needs of oncology HCPs for a practical and tailored MI training, and a means to regularly monitor and provide ongoing support to patients' MA [20]. Moreover, qualitative findings suggest that SAMSON has the potential to help patients in MA and symptom self-management, as well as to assist HCPs in monitoring and supporting patients' adherence. The results of the study help to address the gap of knowledge and the need in oncology practice [17,18] by providing evidence of the high quality and potential effect of a digital multicomponent MA solution.

Regarding SAMSON's feasibility, we noted a high level of engagement with the SAMSON solution in terms of the overall tasks completed by patients on the SAMSON smartphone app and the MI teleconsultations completed by both HCPs and patients. The high acceptability and feasibility of SAMSON is a result of several factors. First, co-design and rigorous design framework were applied to develop SAMSON [22]. By

XSL•FO
RenderX

involving end users and stakeholders throughout all design stages of the intervention, co-designing helped to improve the solution's acceptability, desirability, and usability [38,39]. The use of design frameworks, for example, design science research methodology in this study, could enhance the artifact's design, which is crucial in digital intervention development, and improve its acceptance, usage, and efficacy [40]. Furthermore, the design framework usage can improve the rigor and translatability of research, which is currently limited or poorly reported in the development of available MA interventions in cancer [18]. Second, after development, individual components of SAMSON were successfully tested by end users on their acceptability, usability, and usefulness [19,20]. To the best of our knowledge, SAMSON is the first comprehensive digital MA solution in cancer that is co-designed, theory-based, evidence-based, and rigorously developed and tested.

Although overall acceptance and feasibility were relatively high, some HCPs and patients in the trial reported several technical issues with the SAMSON app and desired better visualization, more functionality, and further instructions and training. This feedback will be used to further improve the solution in the future. A couple of study HCPs were concerned about the time required for MI consultations. Time constraints have been reported as one of the barriers to MI implementation in clinical practice [41]. It is noted that in this study, SAMSON was delivered as an add-on service on top of the hospital's usual care, which required extra time and effort from busy HCPs. In addition, delivering MI consultations to promote adherence was a newly acquired skill by the study HCPs, thus, they required more time to master this approach. A holistic approach, including the hospital's mechanisms to provide continuous MI monitoring and training, as well as support (both in terms of skills and resources) to facilitate and maintain HCPs' confidence and motivation in using this skill set can be a solution to tackle MI implementation barriers [41,42].

The overall recruitment, randomization, retention, and data collection compliance rates were higher than the preset upper thresholds. A high attrition rate is generally one of the major concerns in digital health RCTs compared to RCTs testing more "traditional" interventions [43]. However, the high recruitment rate and the low rate of loss to follow-up in this trial indicate that the SAMSON solution and online pragmatic RCT design are feasible in busy oncology clinical settings, as long as appropriate methodological strategies are applied. Specifically, in this study, clinical nurses and pharmacists from the Hematology Department were recruited and funded to support study recruitment and deliver teleconsultations. The RC dedicated additional time to building rapport with participants and consistently assisted them through all processes of the trial (consenting, baseline assessment, randomization, app installation, technical training and support, and outcome assessments). In addition, different channels of communication, for example, emails, phone calls, and SMSs were used to follow-up and motivate participants throughout the study. The importance of check-ups and follow-ups postrandomization in trials of digital interventions, where participants are required to have certain levels of digital literacy to perform tasks of the intervention [44], has been emphasized in several studies [45]. Future digital trials may benefit from these recruitment and retention strategies.

In this trial, a combination of techniques was used to measure MA, including self-report measures and prescription refill reports. Although triangulation of measurements could improve the accuracy of adherence [46], its interpretation needs to be cautious. The MRA using pharmacy computer records, which is objective [47], does not guarantee that all dispensed medications were consumed by patients. The self-reported adherence survey (ASK-12) is simple, but often subjective and more about barriers to adherence than actual adherence status [31]. Future studies should consider using high-accuracy methods, such as the Medication Event Monitoring System [5], to measure MA.

This study has some limitations. Like many digital health studies, our sample was predominantly Caucasian and highly educated [48,49]. Consequently, the findings may not be representative of patients who are non-Caucasian or have lower socioeconomic status, who might face higher barriers to adherence and could potentially benefit more from the SAMSON solution than their Caucasian or higher socioeconomic status peers [50,51]. Participants were recruited from the Hematology department at PMCC, one of Australia's leading oncology hospitals, and were prescribed only 1 oral anticancer regimen. Therefore, the results may not be generalizable to those who use multiple anticancer medications or receive care in low-resource oncology settings. A more targeted recruitment strategy focusing on underserved cancer patient populations with other types of cancer in various levels of oncology care institutions is warranted.

Conclusion

Before undertaking this pilot trial, both components of the SAMSON solution were co-designed and developed based on evidence and theory, and then individually tested on target users. The results of this study confirmed that SAMSON is acceptable, usable, and useful for both HCPs and patients with cancer. Both the SAMSON solution and the pragmatic RCT design are feasible in real-life oncology settings. These findings are very encouraging, given the numerous challenges in applying RCT as an evaluation design for digital health interventions [52]. Our next steps will involve refining the SAMSON solution based on participants' feedback from this study and conducting a full RCT to evaluate its clinical and economic effectiveness.

Acknowledgments

We thank all Peter MacCallum Cancer Centre HCPs, especially Ms Megan Palmer (nurse), and Ms Jenny Rogers (research pharmacist and trial coordinator) for their support in participant recruitment and teleconsultation delivery; Ms Gail Rowan and Ms Sally L Brooks (senior pharmacists) for developing medicine information and side effect self-care advice; Professor Michael

XSL•FO
RenderX

Jefford (Director of the Australian Cancer Survivorship Centre), Professor Ben Solomon, Associate Professor Shahneen Sandhu (consultant medical oncologists), Mr Alan White, and Mrs Fiona White (consumer representatives) for contributing to the study protocol development and being part of the study's steering committee. We thank all patients for their dedicated time and participation in this study.

This work was supported by the Digital Health CRC Limited (DHCRC), Swinburne University of Technology, and Peter MacCallum Cancer Centre (project DHCRC-0043). DHCRC is funded under the Australian Commonwealth's Cooperative Research Centers (CRC) program. THD is supported by the Australian Government Research Training Program Scholarship. The sponsor had no influence on the study design or the collection, analysis, and interpretation of data. The final decision to include the comments and submit the manuscript for publication was made only by the authors.

Data Availability

The datasets generated and analyzed during this study are not publicly available but are available from the corresponding author on reasonable request.

Authors' Contributions

THD is the submitting and corresponding author. THD conceived of the study, its design, coordination, and implementation and drafted the manuscript. THD, AW, KB, MA, NW, PPJ, PS, and SQ were involved in study design and protocol development. THD and PS were involved in the literature review and developing study instruments and materials. AW and MD supported patient recruitment and delivered MI teleconsultations. MA assisted in collecting medication refill data. THD, PS, and SQ contributed to the design of the statistical analysis approach. THD analyzed the data in consultation with SQ and PS. All authors were involved in providing a critical review of the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 1135 KB - formative_v9i1e65302_app1.pdf]

Multimedia Appendix 2
Participant information statement and consent form.
[DOCX File , 292 KB - formative_v9i1e65302_app2.docx]

Multimedia Appendix 3
Survey booklet.
[DOCX File , 314 KB - formative_v9i1e65302_app3.docx]

Multimedia Appendix 4
Table S1. Thresholds for traffic light approach to feasibility.
[DOCX File , 15 KB - formative_v9i1e65302_app4.docx]

Multimedia Appendix 5
Ethics approvals.
[PDF File (Adobe PDF File), 247 KB - formative_v9i1e65302_app5.pdf]

Multimedia Appendix 6
Ethics approvals.
[PDF File (Adobe PDF File), 148 KB - formative_v9i1e65302_app6.pdf]

Multimedia Appendix 7
Table S2. Participant demographics by arm.
[DOC File , 70 KB - formative_v9i1e65302_app7.doc]

Multimedia Appendix 8
Unified Theory of Acceptance and Use of Technology results of patient participants.

XSL•FO
RenderX

[DOCX File , 19 KB - formative_v9i1e65302_app8.docx]

Multimedia Appendix 9
Unified Theory of Acceptance and Use of Technology results of health care professional participants.
[DOCX File , 20 KB - formative_v9i1e65302_app9.docx]

Multimedia Appendix 10
Table S3. Study feasibility results.
[DOCX File , 18 KB - formative_v9i1e65302_app10.docx]

Multimedia Appendix 11
Table S4. Differences and change from baseline to week 12 between arms, analysis of covariance test.
[DOCX File , 18 KB - formative_v9i1e65302_app11.docx]

References

1. Weingart SN, Brown E, Bach PB, Eng K, Johnson SA, Kuzel TM, et al. NCCN Task Force Report: oral chemotherapy. J Natl Compr Canc Netw 2008 Mar;6 Suppl 3:S1-14. [Medline: 18377852]
2. NICE. Improving Outcomes in Hematological Cancers: The Manual. London: National Institute for Clinical Excellence; 2003.
3. Agrawal M, Garg RJ, Cortes J, Quintás-Cardama A. Tyrosine kinase inhibitors: the first decade. Curr Hematol Malig Rep 2010;5(2):70-80. [doi: 10.1007/s11899-010-0045-y] [Medline: 20425399]
4. Alshawa A, Gong J, Marcott V, Khan R, Boving V, Brown L, et al. Strategic improvement of oral antineoplastic investigational agents compliance. Global Journal on Quality and Safety in Healthcare 2019;2(1):5-10. [doi: 10.4103/jqsh.Jqsh_23_18]
5. Sabate E. Adherence to Long-term Therapies: Evidence for Action. Geneva, Switzerland: World Health Organization; 2003.
6. Ruddy K, Mayer E, Partridge A. Patient adherence and persistence with oral anticancer treatment. CA Cancer J Clin 2009;59(1):56-66 [FREE Full text] [doi: 10.3322/caac.20004] [Medline: 19147869]
7. Bouwman L, Eeltink CM, Visser O, Janssen J, Maaskant JM. Prevalence and associated factors of medication non-adherence in hematological-oncological patients in their home situation. BMC Cancer 2017;17(1):739. [doi: 10.1186/s12885-017-3735-1] [Medline: 29121889]
8. Gater A, Heron L, Abetz-Webb L, Coombs J, Simmons J, Guilhot F, et al. Adherence to oral tyrosine kinase inhibitor therapies in chronic myeloid leukemia. Leuk Res 2012;36(7):817-825. [doi: 10.1016/j.leukres.2012.01.021] [Medline: 22364811]
9. Jacobsen P, Sweet KL, Lee YH, Tinsley S, Lancet JE, Komrokji RS, et al. Adherence to Tyrosine Kinase Inhibitor (TKI) Therapy in Patients with Chronic Myeloid Leukemia (CML). Blood 2011;118(21):4431 [FREE Full text]
10. Guilhot F, Coombs J, Zernovak O, Szczudlo T, Rosti G. A global retrospective and physician-based analysis of adherence to tyrosine kinase inhibitor (TKI) therapies for chronic myeloid leukemia (CML). Blood 2010;116(21):1514 [FREE Full text]
11. Jabbour EJ, Kantarjian H, Eliasson L, Cornelison AM, Marin D. Patient adherence to tyrosine kinase inhibitor therapy in chronic myeloid leukemia. Am J Hematol 2012;87(7):687-691 [FREE Full text] [doi: 10.1002/ajh.23180] [Medline: 22473898]
12. Wu EQ, Johnson S, Beaulieu N, Arana M, Bollu V, Guo A, et al. Healthcare resource utilization and costs associated with non-adherence to imatinib treatment in chronic myeloid leukemia patients. Curr Med Res Opin 2010;26(1):61-69. [doi: 10.1185/03007990903396469] [Medline: 19905880]
13. Ekinci E, Nathoo S, Korattyil T, Vadhariya A, Zaghloul HA, Niravath PA, et al. Interventions to improve endocrine therapy adherence in breast cancer survivors: what is the evidence? J Cancer Surviv 2018;12(3):348-356. [doi: 10.1007/s11764-017-0674-4] [Medline: 29396760]
14. Kavookjian J, Wittayanukorn S. Interventions for adherence with oral chemotherapy in hematological malignancies: a systematic review. Res Social Adm Pharm 2015;11(3):303-314. [doi: 10.1016/j.sapharm.2014.08.006] [Medline: 25262600]
15. Bank T. World bank group. Mobile phone access reaches three quarters of planet's population. 2012. URL: https://tinyurl.com/46yjd99k [accessed 2012-07-17]
16. Kay M, Santos J, Takane M. mHealth: new horizons for health through mobile technologies. World Health Organization Contract No 2011;9(1):66-71.
17. Greer JA, Amoyal N, Nisotel L, Fishbein JN, MacDonald J, Stagl J, et al. A systematic review of adherence to oral antineoplastic therapies. Oncologist 2016;21(3):354-376 [FREE Full text] [doi: 10.1634/theoncologist.2015-0405] [Medline: 26921292]

XSL•FO
RenderX

18. Dang TH, Forkan ARM, Wickramasinghe N, Jayaraman PP, Alexander M, Burbury K, et al. Investigation of intervention solutions to enhance adherence to oral anticancer medicines in adults: overview of reviews. JMIR Cancer 2022;8(2):e34833 [FREE Full text] [doi: 10.2196/34833] [Medline: 35475978]
19. Dang TH, Wickramasinghe N, Forkan ARM, Jayaraman PP, Burbury K, O'Callaghan C, et al. Co-design, development, and evaluation of a mobile solution to improve medication adherence in cancer: design science research approach. JMIR Cancer 2024;10:e46979 [FREE Full text] [doi: 10.2196/46979] [Medline: 38569178]
20. Dang TH, Ludlow C, Borle H, Alexander M, Wickramasinghe N, Burbury K, et al. Co-designing a motivational interviewing training platform to enhance oncology healthcare professional communication. PEC Innov 2024;5:100335 [FREE Full text] [doi: 10.1016/j.pecinn.2024.100335] [Medline: 39257627]
21. Eysenbach G, CONSORT-EHEALTH Group. Improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res 2011;13(4):e126 [FREE Full text] [doi: 10.2196/jmir.1923] [Medline: 22209829]
22. Dang TH, Wickramasinghe N, Jayaraman PP, Burbury K, Alexander M, Whitechurch A, et al. Safety and adherence to medications and self-care advice in oncology (SAMSON): pilot randomised controlled trial protocol. BMJ Open 2024;14(7):e079122 [FREE Full text] [doi: 10.1136/bmjopen-2023-079122] [Medline: 39043598]
23. Broglio K. Randomization in clinical trials: permuted blocks and stratification. JAMA 2018;319(21):2223-2224. [doi: 10.1001/jama.2018.6360] [Medline: 29872845]
24. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view. MIS Quarterly 2003;27(3):425. [doi: 10.2307/30036540]
25. Venkatesh V, Thong JYL, Xu X. Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology. MIS Quarterly 2012;36(1):157. [doi: 10.2307/41410412]
26. Leon AC, Davis LL, Kraemer HC. The role and interpretation of pilot studies in clinical research. J Psychiatr Res 2011;45(5):626-629 [FREE Full text] [doi: 10.1016/j.jpsychires.2010.10.008] [Medline: 21035130]
27. Abbott JH. The distinction between randomized clinical trials (RCTs) and preliminary feasibility and pilot studies: what they are and are not. J Orthop Sports Phys Ther 2014;44(8):555-558. [doi: 10.2519/jospt.2014.0110] [Medline: 25082389]
28. Vanderbilt University. Tennessee, USA: Vanderbilt University. REDCap (Research Electronic Data Capture). URL: https://www.project-redcap.org/ [accessed 2025-09-25]
29. Eldridge SM, Chan CL, Campbell MJ, Bond CM, Hopewell S, Thabane L, PAFS consensus group. CONSORT 2010 statement: extension to randomised pilot and feasibility trials. BMJ 2016;355:i5239 [FREE Full text] [doi: 10.1136/bmj.i5239] [Medline: 27777223]
30. Hess LM, Raebel MA, Conner DA, Malone DC. Measurement of adherence in pharmacy administrative databases: a proposal for standard definitions and preferred measures. Ann Pharmacother 2006;40(7-8):1280-1288. [doi: 10.1345/aph.1h018]
31. Matza LS, Park J, Coyne KS, Skinner EP, Malley KG, Wolever RQ. Derivation and validation of the ASK-12 adherence barrier survey. Ann Pharmacother 2009;43(10):1621-1630. [doi: 10.1345/aph.1M174] [Medline: 19776298]
32. Hibbard JH, Mahoney ER, Stockard J, Tusler M. Development and testing of a short form of the patient activation measure. Health Serv Res 2005;40(6 Pt 1):1918-1930 [FREE Full text] [doi: 10.1111/j.1475-6773.2005.00438.x] [Medline: 16336556]
33. Bahrom NH, Ramli AS, Isa MR, Baharudin N, Badlishah-Sham SF, Mohamed-Yassin MS, et al. Validity and reliability of the patient activation measure (PAM)-13 Malay version among patients with metabolic syndrome in primary care. Malays Fam Physician 2020;15(3):22-34 [FREE Full text] [Medline: 33329860]
34. Quach CW, Langer MM, Chen RC, Thissen D, Usinger DS, Emerson MA, et al. Reliability and validity of PROMIS measures administered by telephone interview in a longitudinal localized prostate cancer study. Qual Life Res 2016;25(11):2811-2823 [FREE Full text] [doi: 10.1007/s11136-016-1325-3] [Medline: 27240448]
35. Cella DF, Tulsky DS, Gray G, Sarafian B, Linn E, Bonomi A, et al. The functional assessment of cancer therapy scale: development and validation of the general measure. J Clin Oncol 1993;11(3):570-579. [doi: 10.1200/JCO.1993.11.3.570] [Medline: 8445433]
36. StataCorp. Stata Statistical Software: Release 18. URL: https://www.businesswire.com/news/home/20230426005929/en/New-Stata-18-Statistical-Software-Empowers-Researchers-to-Make-the-Most-Out-of-Their-Analyses [accessed 2023-04-26]
37. Dang TH, O'Callaghan C, Alexander M, Burbury K, Jayaraman PP, Wickramasinghe N, et al. 'Take the tablet or don't take the tablet?' A qualitative study of patients' experiences of self-administering anti-cancer medications related to adherence and managing side effects. Support Care Cancer 2023;31(12):680 [FREE Full text] [doi: 10.1007/s00520-023-08122-6] [Medline: 37934298]
38. Borgstrom E, Barclay S. Experience-based design, co-design and experience-based co-design in palliative and end-of-life care. BMJ Support Palliat Care 2019;9(1):60-66. [doi: 10.1136/bmjspcare-2016-001117] [Medline: 28209568]
39. Slattery P, Saeri AK, Bragge P. Research co-design in health: a rapid overview of reviews. Health Res Policy Syst 2020;18(1):17 [FREE Full text] [doi: 10.1186/s12961-020-0528-9] [Medline: 32046728]
40. Hevner A, Wickramasinghe N. Design science research opportunities in health care. In: Wickramasinghe NS, Schaffer JL, editors. Theories to Inform Superior Health Informatics Research and Practice. 1st ed. Cham: Springer; 2018.
41. Bell DL, Roomaney R. Exploring the barriers that prevent practitioners from implementing motivational interviewing in their work with clients. Social Work 2020;56(4):416-429. [doi: 10.15270/56-4-884]

XSL•FO
RenderX

42. Nagy SM, Butame SA, Todd L, Sheffler JL, Budhwani H, Fernandez MI, et al. Barriers and facilitators to implementing a motivational interviewing-based intervention: a multi-site study of organizations caring for youth living with HIV. AIDS Care 2022;34(4):486-491 [FREE Full text] [doi: 10.1080/09540121.2021.1950604] [Medline: 34251935]
43. Mathieu E, McGeechan K, Barratt A, Herbert R. Internet-based randomized controlled trials: a systematic review. J Am Med Inform Assoc 2013;20(3):568-576 [FREE Full text] [doi: 10.1136/amiajnl-2012-001175] [Medline: 23065196]
44. Daniore P, Nittas V, Gille F, von Wyl V. Promoting participation in remote digital health studies: an expert interview study. Digit Health 2023;9:20552076231212063 [FREE Full text] [doi: 10.1177/20552076231212063] [Medline: 38025101]
45. Linardon J, Fuller-Tyszkiewicz M. Attrition and adherence in smartphone-delivered interventions for mental health problems: a systematic and meta-analytic review. J Consult Clin Psychol 2020;88(1):1-13. [doi: 10.1037/ccp0000459] [Medline: 31697093]
46. Stewart SJF, Moon Z, Horne R. Medication nonadherence: health impact, prevalence, correlates and interventions. Psychol Health 2023;38(6):726-765 [FREE Full text] [doi: 10.1080/08870446.2022.2144923] [Medline: 36448201]
47. Brízido C, Ferreira AM, Lopes P, Strong C, Sá Mendes G, Fernandes Gama F, et al. Medication adherence to direct anticoagulants in patients with non-valvular atrial fibrillation - a real world analysis. Rev Port Cardiol (Engl Ed) 2021;40(9):669-675 [FREE Full text] [doi: 10.1016/j.repce.2021.07.002] [Medline: 34503705]
48. Smith SK, Kuhn E, O'Donnell J, Koontz BF, Nelson N, Molloy K, et al. Cancer distress coach: pilot study of a mobile app for managing posttraumatic stress. Psychooncology 2018;27(1):350-353 [FREE Full text] [doi: 10.1002/pon.4363] [Medline: 28032921]
49. Kelley MM, Kue J, Brophy L, Peabody AL, Foraker RE, Yen PY, et al. Mobile health applications, cancer survivors, and lifestyle modification: an integrative review. Comput Inform Nurs 2021;39(11):755-763 [FREE Full text] [doi: 10.1097/CIN.0000000000000781] [Medline: 34074873]
50. Smith SK, Zimmerman S, Williams CS, Benecha H, Abernethy AP, Mayer DK, et al. Post-traumatic stress symptoms in long-term non-Hodgkin's lymphoma survivors: does time heal? JCO 2011;29(34):4526-4533. [doi: 10.1200/jco.2011.37.2631]
51. Mollica M, Nemeth L, Newman SD, Mueller M. Quality of life in African American breast cancer survivors: an integrative literature review. Cancer Nurs 2015;38(3):194-204. [doi: 10.1097/NCC.0000000000000160] [Medline: 24945269]
52. Hrynyschyn R, Prediger C, Stock C, Helmer SM. Evaluation methods applied to digital health interventions: what is being used beyond randomised controlled trials? A scoping review. Int J Environ Res Public Health 2022;19(9):5221 [FREE Full text] [doi: 10.3390/ijerph19095221] [Medline: 35564616]

Abbreviations

ASK-12: Adherence Starts with Knowledge 12
HCP: health care professional
HREC: Human Research Ethics Committee
MA: medication adherence
MI: motivational interviewing
MITP: motivational interviewing training platform
MRA: medication refill adherence
OAM: oral anticancer medicine
PMCC: Peter MacCallum Cancer Centre
PROMIS: Patient-Reported Outcomes Measurement Information System
RC: research coordinator
RCT: randomized controlled trial
SAMSON: Safety and Adherence to Medications and Self-care Advice in Oncology
UTAUT: Unified Theory of Acceptance and Use of Technology

Edited by A Mavragani; submitted 12.08.24; peer-reviewed by O Santin, M Östbring; comments to author 05.12.24; revised version received 12.12.24; accepted 12.12.24; published 19.02.25.

Please cite as:
Dang TH, Wickramasinghe N, Jayaraman PP, Burbury K, Alexander M, Whitechurch A, Dyer M, Quinn S, Forkan ARM, Schofield P
Digital Solution to Support Medication Adherence and Self-Management in Patients with Cancer (SAMSON): Pilot Randomized Controlled Trial
JMIR Form Res 2025;9:e65302
URL: https://formative.jmir.org/2025/1/e65302
doi:10.2196/65302
PMID:39969972

XSL•FO
RenderX

©Thu Ha Dang, Nilmini Wickramasinghe, Prem Prakash Jayaraman, Kate Burbury, Marliese Alexander, Ashley Whitechurch, Mitchell Dyer, Stephen Quinn, Abdur Rahim Mohammad Forkan, Penelope Schofield. Originally published in JMIR Formative Research (https://formative.jmir.org), 19.02.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Assessing the Feasibility and Utility of Patient-Specific 3D Advanced Visualization Modeling in Cerebrovascular Disease: Retrospective Analysis and Prospective Survey Pilot Study

Korak Sarkar[1,2,3,4], MD, MHDS; Vishal Bhimarasetty[1], BS, MS; Abdul Rahim[4], BE, MS; Colin Curtis[5], BSE; Kimberly Hughes[1], BA, BS

[1]Department of Academics, Ochsner BioDesign Lab, Ochsner Health, New Orleans, LA, United States
[2]Ochsner Neuroscience Institute, Ochsner Health, New Orleans, LA, United States
[3]Physical Medicine and Rehabilitation Service, Southeast Louisiana Veterans Health Care System, New Orleans, LA, United States
[4]University of Queensland, School of Medicine, Queensland, Australia
[5]Ochsner Information Services, Ochsner Health, New Orleans, LA, United States

Corresponding Author:
Korak Sarkar, MD, MHDS
Department of Academics
Ochsner BioDesign Lab
Ochsner Health
1514 Jefferson Highway
New Orleans, LA, 70119
United States
Phone: 1 504 894 2799
Email: korak.sarkar.md@gmail.com

Abstract

Background: The prevalence, clinical burden, and health care costs (>US $100 billion) associated with cerebrovascular disease (CVD) will increase significantly as the US population grows and ages over the next 25 years. Existing 2D imaging modalities have inherent limitations in visualizing complex CVD, which may be mitigated with the use of patient-specific 3D advanced visualization (AV) technologies. There remain gaps in knowledge, however, regarding how and with what impact these technologies are being used in CVD.

Objective: The aim of this study was to characterize the clinical attributes and reported utility associated with the use of 3D AV modeling in CVDs, specifically intracerebral arteriovenous malformations.

Methods: This pilot study employs a combination of retrospective analysis and prospective surveys to describe the utilization and utility of patient-specific AV models at a single high-volume certified comprehensive stroke center.

Results: From July 2017 to February 2023, 25 AV models were created for 4 different clinicians. The average patient age was 37.4 years; 44% (11/25) of the patients were African Americans, 52% (13/25) were on public insurance, and 56% (14/25) were associated with a neurovascular procedure. In this study, 18 clinicians with diverse experience responded to AV model surveys, with a 92.2% (166/180) completion rate. There was an average reported utility of 8.0 on a 0-10 scale, with higher scores reflecting increased utility. Compared to 2D viewing, AV models allowed staff to appreciate novel abnormal anatomy, and therefore, they would have changed their therapeutic approach in 45% (23/51) of the cases.

Conclusions: AV models were used in complex CVDs associated with young, publicly insured individuals requiring resource-intensive interventions. There was strong and diverse clinician engagement with overall report of substantial utility of AV models. Staff clinicians frequently reported novel anatomical and therapeutic insights based on AV models compared to traditional 2D viewing. This study establishes the infrastructure for future larger randomized studies that can be repeated for CVDs or other disease states and incorporate assessments of other AV modalities such as 3D printing and medical extended reality.

(JMIR Form Res 2025;9:e51939) doi:10.2196/51939

XSL•FO
RenderX

KEYWORDS

cerebrovascular disease; advanced visualization; 3D modeling; cerebrovascular; intracerebral arteriovenous malformations; artery; vein; vessel; medical extended reality; 3D printing; medical simulation; virtual reality; augmented reality; usability; survey; stroke; brain; cerebral

Introduction

Background

Cerebrovascular disease (CVD) is a common cause of stroke, aneurysms, and thromboembolic disease, resulting in significant neurologically based disability and death [1,2]. Although CVD can be treated with medications, many disease presentations require procedural interventions. It is estimated that by 2030, the annual costs for cerebrovascular accidents alone will exceed US $140 billion [3]. Improved visualization, greater repetition, and more frequent training experiences are factors known to improve outcomes in CVD treatments requiring complex procedures [4]. Treatment of patients with difficult or otherwise hostile anatomies carries more risk when operators have less general experience navigating unusual or atypical structures [5]. The lack of anatomical and procedural familiarity can lead to lengthier procedures, longer anesthetic duration, increased radiation exposure, and poorer outcomes [5,6]. Access to personalized 3D models designed for patient-specific planning can provide higher fidelity visualization and simulation than traditional 2D viewing, thereby reducing the prevalence, associated costs, and morbidity from complex CVD procedures [7].

The rapid adoption of advanced visualization (AV) tools such as 3D rendering, 3D printing, and medical extended reality has created novel opportunities for medical visualization and simulation. Patient-specific 3D digital and physical anatomical models based on clinical radiographic data can now be rendered and fabricated for patient education, clinical training, and procedural planning [8]. This approach holds great promise as a tool in visualizing complex neurovascular anatomy and simulating complex surgical and endovascular procedures [9]. The increasing adoption of advanced modeling has prompted organizations such as the Radiological Society of North America to issue guidelines for implementing medical 3D printing as training models for highly complex clinical scenarios [10]. The Radiological Society of North America guidelines discuss the use of patient-specific 3D models for a variety of specialties, including congenital heart disease, craniomaxillofacial, breast, and musculoskeletal pathologies. There, however, remains limited guidance on the optimal use of 3D modeling in CVD.

CVD interventions are an ideal application of AV modalities due to the idiosyncratic and patient-specific 3D nature of intracerebral vascular pathology. Current evidence suggests that AV technologies such as 3D rendering, medical extended reality, and 3D printing are an appropriate platform for studying and rehearsing plans for treating complex CVD [11]. Several studies have reported the growing utilization of AV modeling, visualization, and procedural training in CVD [12]. The current state of research, however, remains limited and warrants more thorough investigation [12,13]. To address this gap, the field would benefit from the development of a scalable modeling and assessment platform, which could describe and evaluate the use of patient-specific AV anatomical modeling. Such an infrastructure could address the clear, critical, and well-identified need to characterize the use and utility of AV models in CVD.

Objectives

This pilot study was designed and executed to describe the utilization and utility associated with the use of patient-specific anatomical models in the management of intracerebral arteriovenous malformations (iAVMs). This exploratory investigation involves the implementation of a scalable digital fabrication infrastructure to produce patient-specific anatomical AV models, characterize their use, and quantify their potential impact in the management of iAVMs. The aims of this study include describing the clinical, anatomical, and demographic features associated with the use of patient-specific AV models in the management of iAVMs. Additionally, this study aims to quantify the utility of patient-specific AV anatomical models when compared to traditional 2D viewing across different clinical experience levels. We hypothesize there is a negative correlation between experience and reported utility of AV patient-specific anatomical models in CVD. We specifically predict that when compared to standard 2D viewing, clinicians with less experience will report higher utility of AV patient-specific anatomical models than more experienced staff clinicians.

Methods

Digital Fabrication Process

Clinicians request AV modeling based on routine clinical neuroimaging imaging, for example, computed tomography, magnetic resonance, or rotational angiography via an electronic medical record order set created to request patient-specific 3D AV models. Anatomical models are based on the digital imaging and communications in medicine (DICOM) datasets used in clinical imaging. DICOM data are accessed by a staff biomedical engineer via a compliant and secure enterprise picture archiving and communications system. The biomedical engineer segments these DICOM datasets by using computer-aided design software to produce 3D mesh files, specifically standard tessellated language files. Segmentation is the process by which regions of interest are identified through automatic methods (eg, thresholding, edge detection, region growing) and are then refined by manual selection and separation of anatomical structures. This step is completed by the biomedical engineer in conjunction with a radiologist, neurologist, or neurosurgeon to ensure an accurate representation of the region of interest. Once created, these 3D models are rendered and manipulated digitally via an internally developed web-based viewer. These digital models can be further postprocessed for 3D printing to create physical models (Figure 1).

XSL•FO
RenderX

Figure 1. Ochsner BioDesign Digital Request and Fabrication Process. An electronic medical record–based order request is placed by a clinician. Source imaging DICOM (Digital Imaging and Communications in Medicine) datasets are accessed via a secure enterprise PACS (Picture Archiving and Communications System). Biomedical engineers utilize computer-aided design applications to transform DICOM datasets into 3D file types. These 3D files are then either visualized using a web-based browser or extended reality modalities such as virtual or augmented reality. These 3D files can be further processed to produce physical models using 3D printing. DICOM: Digital Imaging and Communications in Medicine; EMR: electronic medical record.

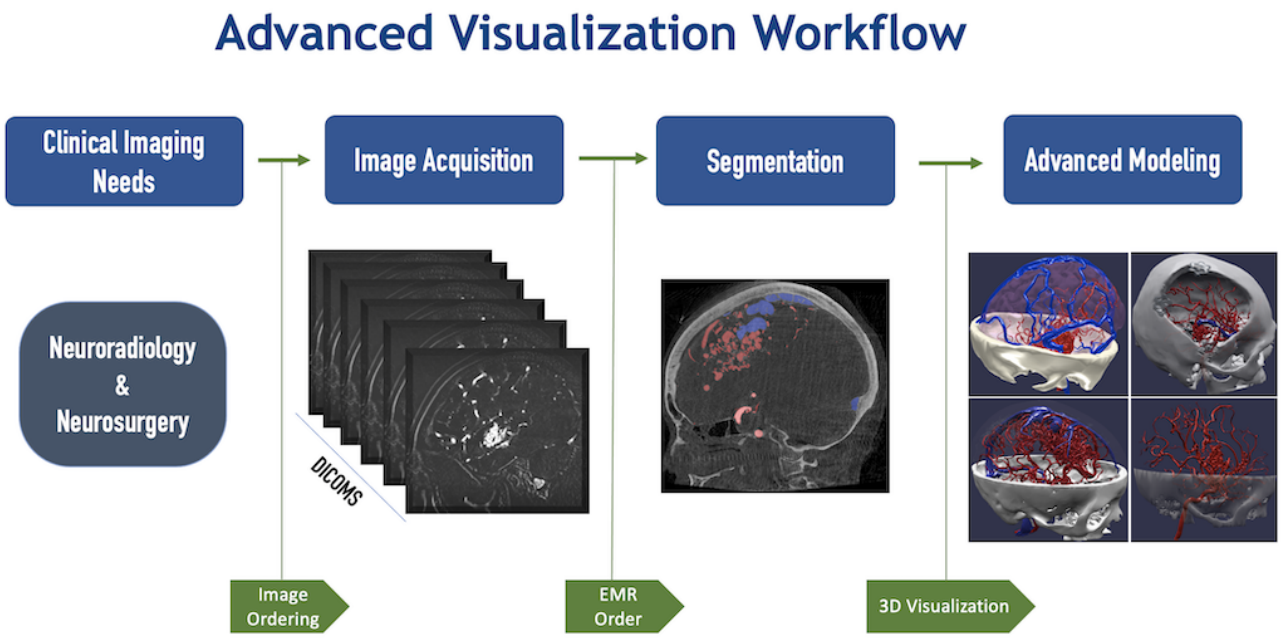

Epidemiological Data Collection

A CVD AV registry is used to collect data on clinical and utilization metrics associated with AV model use via a compliant research data capture application. Data are collected from the electronic medical record and include patient age at the time of modeling, sex, race, insurance status, imaging modality, comorbidities (hypertension, diabetes, epilepsy, tobacco use), size/scale/location of iAVM, associated current procedural terminology (CPT) codes, presence and type of any associated and relevant procedure within 9 months of model request, the presence and duration of any associated intensive care unit stay, as well as a Spetzler-Martin Grade (SMG). SMG is a clinical tool to assess the procedural risk of iAVMs that incorporates the size, location, and type of venous drainage.

Clinician and Model Survey Data Collection

Twenty-four clinicians were invited to participate in this study from the organization's neuroscience institute via clinical and educational conferences, 18 of whom responded. Each clinician assessed a set of 10 sequential iAVM cases, which had previously been requested for modeling. Each participant received a secure email, which included a link to the model utility survey as well as the following information to review: (1) a traditional 2D viewer with the axial, sagittal, and coronal source images available in serial via an embedded digital multimedia container format file, specifically an mp4; (2) information to access and view the source imaging via the standard picture archiving and communications system viewing approach; (3) the associated anatomical 3D AV model accessible via an internally developed web-based viewer; and (4) the radiographic report of the imaging from which the model was created.

Drawing from similar instruments used in the literature, a survey is used to assess (1) demographics and level of experience, (2) clinician-reported complexity of each case on 1-10 Likert Scale, and (3) utility of the patient-specific AV models using a 1-10 Likert scale, with higher scores representing increased complexity and utility (Multimedia Appendices 1-3) [10,12-16]. The survey includes questions regarding the potential diagnostic and therapeutic impact of 3D AV models compared to standard 2D representations. The research data capture application was used to digitally distribute and collect survey data in an automated and compliant way.

Analysis Plan

Descriptive statistics (averages, range, frequency) were reported for the epidemiological and utilization data for the AV models as well as the demographic and experience data of the survey respondents. Model survey responses were treated as ordinal. Each respondent's complexity and utility response was averaged to mitigate the limitations associated with the statistical analysis of the dependent and repeated measurements. The independent variable is experience level, and 2 approaches were used to classify experience groups. One approach categorizes experience by training level: (1) trainee defined as medical student, resident, or fellow; (2) staff as defined by those who have completed their terminal clinical training; and (3) advanced practice provider (APP). An alternative approach categorizes experience based on the reported number of iAVM evaluations: low (0-10), medium (11-100), and high (>101). The dependent variables are reported complexity and utility of the AV model compared

XSL•FO
RenderX

to traditional 2D viewing. The nonparametric Kruskal-Wallis H test was used to assess the null hypothesis that there is no difference in reported complexity or utility of an AV model among clinicians in these 3 groups. A significance level or α of .05 was used for statistical testing. A Dunn multiple comparison test was used for post hoc analysis to compare specific experience category combinations. A significance level or α of .05 was used for statistical testing.

Ethics Approval

This study was reviewed and approved by the Ochsner institutional review board (approval 2019.089) and was determined to be exempt and granted a waiver of informed consent, as it posed minimal risk to participants. All data collected for this study were fully deidentified prior to analysis to ensure participant privacy and confidentiality. Study models and data were deidentified and anonymized. Secure storage protocols were implemented in compliance with institutional and federal data protection standards. Access to the data was restricted to authorized study personnel directly involved in data collection and analysis. No compensation was provided to study participants. All procedures were conducted in accordance with appropriate ethical guidelines and applicable regulations.

Results

The aim of this study was to deploy a digital fabrication process that can (1) describe the clinical attributes associated with the use of AV models and (2) assess the utility of AV models across clinical experience levels when compared to standard 2D viewing.

Epidemiology and Utilization

AV models for 26 patients with iAVMs were requested from July 2017 to February 2023. Of the 26 models requested, 25 AV models were created. One request was not modeled because the primary source image was ordered by the clinician but never completed by the patient. Of note, 6 out of the 25 cases were also 3D printed to create physical patient-specific models (Multimedia Appendix 4). Of the AV models, 60% (15/25) were based on rotational angiography, 32% (8/25) on computed tomography angiography, and 8% (2/25) on magnetic resonance angiography. Patient age ranged from 13 to 71 years (mean age 37.4 years, SD 16.4 years) with various comorbidities and diverse demographics (Table 1).

Table 1. Demographics and comorbidities of the participants for whom 3D models were created.

	Values, n (%)
Smoking (n=25)	
Nonsmoker	4 (16)
Smoker	21 (84)
Comorbidities (n=10)	
Hypertension	8 (80)
Diabetes mellitus	2 (20)
Hyperlipidemia	0 (0)
Seizure/epilepsy	0 (0)
Other	0 (0)
Total comorbidities (n=21)	
1	8 (38)
2	4 (19)
3	6 (29)
4	1 (5)
5	2 (10)
Other epidemiology metrics (n=25)	
Gender	
Female	8 (32)
Male	17 (68)
Race	
Caucasian	14 (56)
African American	11 (44)

With regards to pathophysiology, iAVM diameter ranged from 4.5 mm to 62 mm, with an average diameter of 29.2 mm.

XSL•FO
RenderX

Intracerebral hemorrhage was associated with 28% (7/25) of the cases. SMG scores were obtained and ranged from 1 to 5 with a median score of 2 (Figure 2).

The most frequent iAVM location was the frontal lobe (9/25, 36%), and the most frequent venous draining was the superior sagittal sinus (9/25, 36%) (Table 2). Clinically, 56% (14/25) of the models were associated with some form of open or endovascular intervention within 9 months of the AV model request.

With regard to utilization, 9 unique CPT codes were found to be associated with these models. The most frequent CPT code associated with AV models was for selective unilateral catheter access of the common carotid artery (36223) and was billed in 84% (21/25) of the cases. Of the patients who had models created, 64% (16/25) had public insurance, with most patients utilizing Medicaid (14/25, 56%). Intensive care unit stays were associated with 56% (14/25) of the cases with a mean length of 6.9 days and a maximum length of 34 days. The 25 AV models were created for 4 unique clinicians at the following frequencies: interventional neuroradiologist (1/25, 4%), pediatric neurosurgeon (2/25, 8%), interventional neurologist (5/25, 20%), and a vascular neurosurgeon (17/25, 68%).

Figure 2. (A) Spetzler-Martin Grading scoring approach and (B) dot plot of Spetzler-Martin Grading scores, showing the distributions for values 1 (6/25, 24%), 2 (8/25, 32%), 3 (5/25, 20%), 4 (5/25, 20%), and 5 (1/25, 4%). SMG: Spetzler-Martin Grading.

(A)

Characteristics		Points assigned
Size		
	Small (<3 cm)	1
	Medium (3-6 cm)	2
	Large (>6 cm)	3
Eloquence		
	No	0
	Yes	1
Venous drainage		
	Superficial only	0
	Any deep	1

(B)

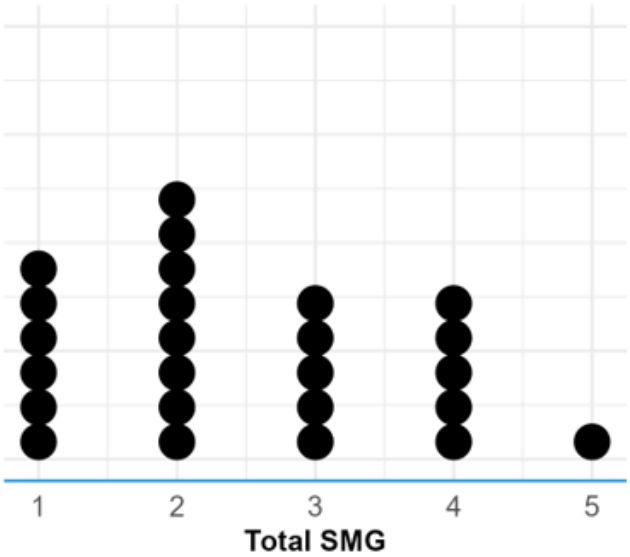

XSL•FO
RenderX

Table 2. Frequency distribution of intracerebral arteriovenous malformation locations and venous drainage categories (n=25).

Location/category	Values, n (%)
Intracerebral arteriovenous malformation locations	
Frontal lobe	9 (36)
Frontoparietal lobe	5 (20)
Occipital lobe	3 (12)
Other	8 (32)
Venous drainage category	
Superior sagittal sinus	9 (36)
Cortical vein	5 (20)
Transverse sinus	3 (12)
Sigmoid sinus	2 (8)
Other	6 (24)

Descriptive Statistics for Clinicians

Of the 24 clinicians who completed the initial recruitment survey between November 2022 to January 2023, 18 completed at least one model assessment survey. Six clinicians were lost to follow-up, that is, 1 medical student and 5 APPs. The descriptive statistics of the demographics and experience were calculated for the 18 clinicians who responded (Table 3). Clinicians (mean age 36.5 years; median age 34.5 years, range 26-56 years) identified their specialties as neurosurgery (1/18, 6%), neurointerventional radiology (3/18, 17%), and neurology (12/18, 67%). Two respondents (11%) were medical students and did not identify with a specialty. The other respondents identified as residents (4/18, 22%), APPs (5/18, 28%), fellow (1/18, 6%), junior attendings with <5 years of experience from terminal training (3/18, 17%), and senior attendings with >5 years of experience after terminal training (3/18, 17%). Self-reported iAVM evaluations was reported as 0 (2/18, 11%), 1-10 (4/18, 22%), 11-50 (6/18, 33%), 51-100 (3/18, 17%), 101-200 (2/18, 11%), and >201 (1/18, 6%).

Table 3. Summary statistics of the specialties and level of training for clinician survey respondents (n=18).

	Values, n (%)
Specialties/level of training	
Neurosurgery	1 (6)
Neurointerventional radiology	3 (17)
Neurology	12 (67)
Medical student	2 (11)
Level of training	
Trainees	5 (28)
Staff	6 (34)
Advanced practice provider	5 (28)
Number of arteriovenous malformation evaluations	
Low (<10)	6 (34)
Medium (51-100)	9 (50)
High (>101)	3 (17)
Clinician gender distribution	
Male	10 (56)
Female	8 (44)

Descriptive Statistics of Model Surveys

Each of the 18 clinicians were provided 10 model surveys for a total of 180 possible responses. Approximately 95.6% (172/180) of the surveys were initiated, and 92.2% (166/180) were completed. The possible complexity scores ranged from 1 to 10, with 10 being the most complex. The median of the average complexity reported was 6.95 with a minimum average score of 3.5 and a maximum average score of 8.1 across all respondents. Descriptive statistics, including median and IQR,

XSL•FO
RenderX

were calculated for utility, complexity, and improvement attributed to access to AV models for each expertise level (Figure 3) based on training level and the number of iAVMs evaluated (Figure 4). Clinicians responded affirmatively in 99.4% (171/172) of the cases that they would like to share AV models with their patients.

Figure 3. Boxplots for complexity, utility, and improvement derived from 3D model based on expertise: advanced practice provider (n=5), staff (n=6), and trainee (n=7). APP: advanced practice provider.

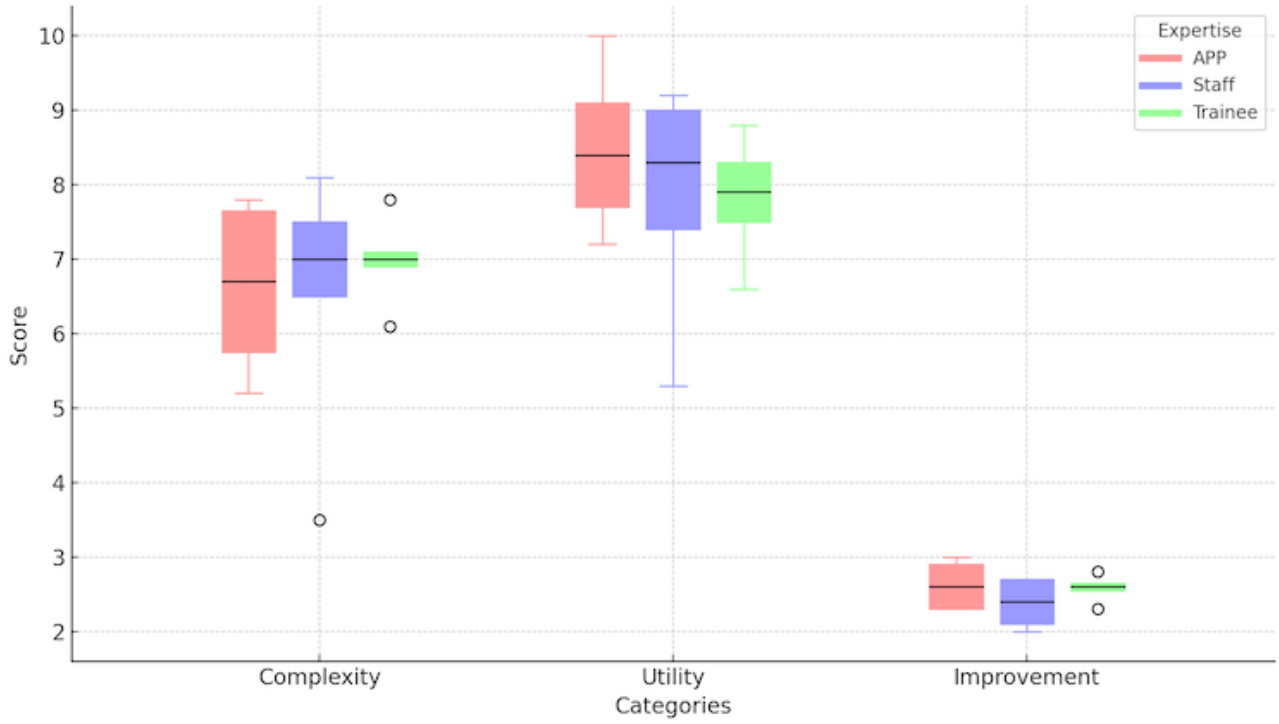

Figure 4. Boxplots for complexity, utility, and improvement derived from 3D model based on the number of arteriovenous malformations evaluated: low (0-10, n=6), intermediate (11-100, n=9), and high (>101, n=3). AVM: arteriovenous malformation.

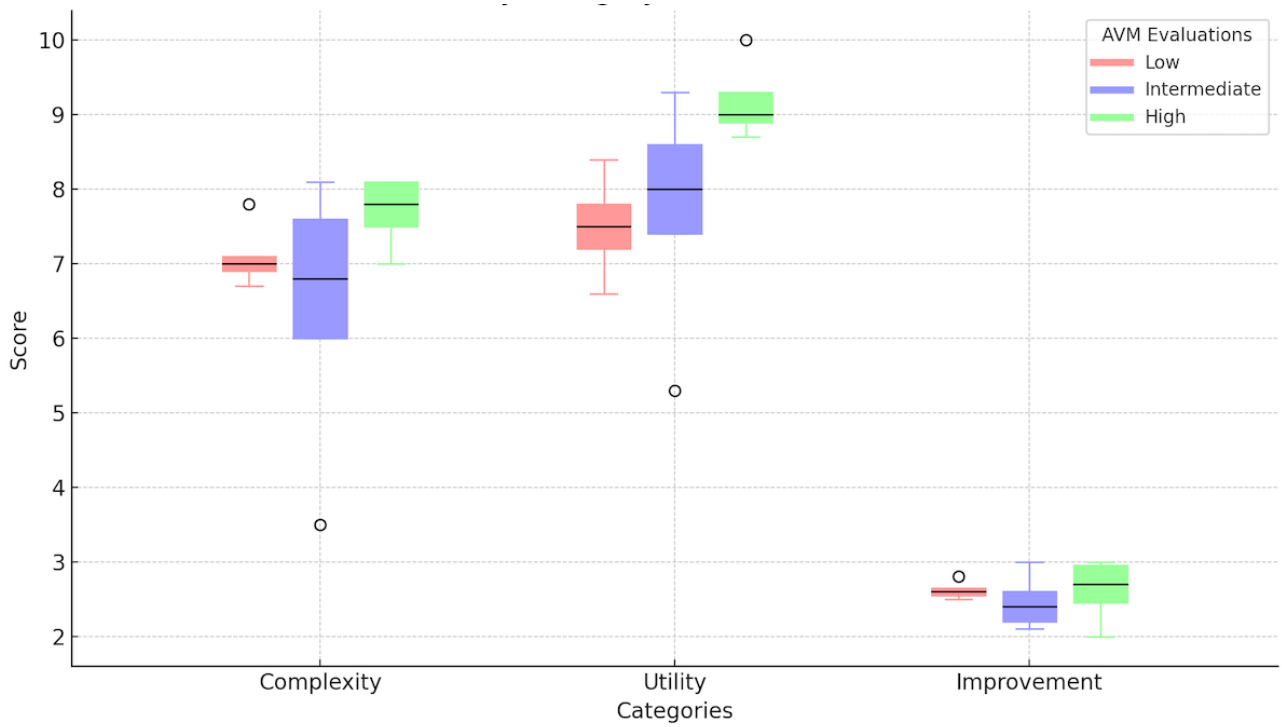

There was a total of 6 physicians who had completed their terminal training with a cumulative survey response rate of 85% (51/60). When compared to standard 2D models, these 6 staff clinicians reported that AV models provided some improvement in visualization in 49% (25/51) of the cases and substantial improvement in 41% (21/51) of the cases. Staff level physicians reported that in 49% (25/51) of the cases, they were able to appreciate normal anatomy in the AV model that they could

XSL•FO
RenderX

not in standard 2D imaging. Similarly, staff reported that in 45% (23/51) of the cases, they were able to appreciate abnormal anatomy in the AV model that they could not in standard 2D imaging. Staff answered affirmatively in 14% (7/51) of the cases that the AV model would change their diagnosis while reporting affirmatively in 45% (23/51) of the cases that the AV model would change their therapeutic approach (Table 4).

Table 4. Diagnostic and management impact of advanced visualization compared to traditional 2D viewing on staff level clinicians (n=51).

	Values, n (%)
Change in diagnosis	
No	39 (76)
Yes	7 (14)
No response	5 (10)
Change in therapeutic procedures	
Yes	23 (45)
No	21 (41)
No response	7 (14)
Change in appreciation of normal anatomy in 3D	
Yes	25 (49)
No	20 (39)
No response	6 (12)
Change in appreciation of pathologic defects in 3D	
Yes	23 (45)
No	22 (43)
No response	6 (12)

Inferential Analytics of Model Surveys

When categorized by training level, the 3 groups were distributed between APPs (5/18, 28%), trainees (7/18, 39%), and staff (6/18, 33%). The Kruskal-Wallis H nonparametric test was deployed with nonsignificant *P* values >.05 for the survey questions relevant to model complexity (*P*=.72) and utility (*P*=.48). Thus, we fail to reject the null hypothesis. Rather, we conclude that there is no statistically significant difference in reported complexity nor utility between clinician types. When categorized by self-reported experience, the 3 groups were distributed between low (6/18, 33%), intermediate (9/18, 50%), and high (3/18, 17%). The Kruskal-Wallis H nonparametric test was deployed with *P*=.12 for the survey questions relevant to model complexity. Thus, we fail to reject the null hypothesis for complexity and conclude that there is no statistically significant difference in reported complexity between experience levels. The *P* value for utility, however, did achieve significance with *P*=.03. Thus, for utility, we can reject the null hypothesis and conclude that there is a statistically significant difference in the reported utility metric between self-reported experience groups. Post hoc analysis with Dunn multiple comparison test demonstrated that the comparison between the low and high group showed a significant difference in the utility metric. There was not a significant difference in the utility metric between the comparisons of the other groupings: high-intermediate or intermediate-low (Figure 5).

XSL•FO
RenderX

Figure 5. Boxplot of utility across self-reported experience. The Kruskal Wallis H test demonstrated a statistical difference between the low and high experience groups (*P*=.03). The analysis reveals that the median utility scores for low (n=6), intermediate (n=9), and high (n=3) categories of self-reported experience are 7.45, 8.00, and 9.00, respectively.

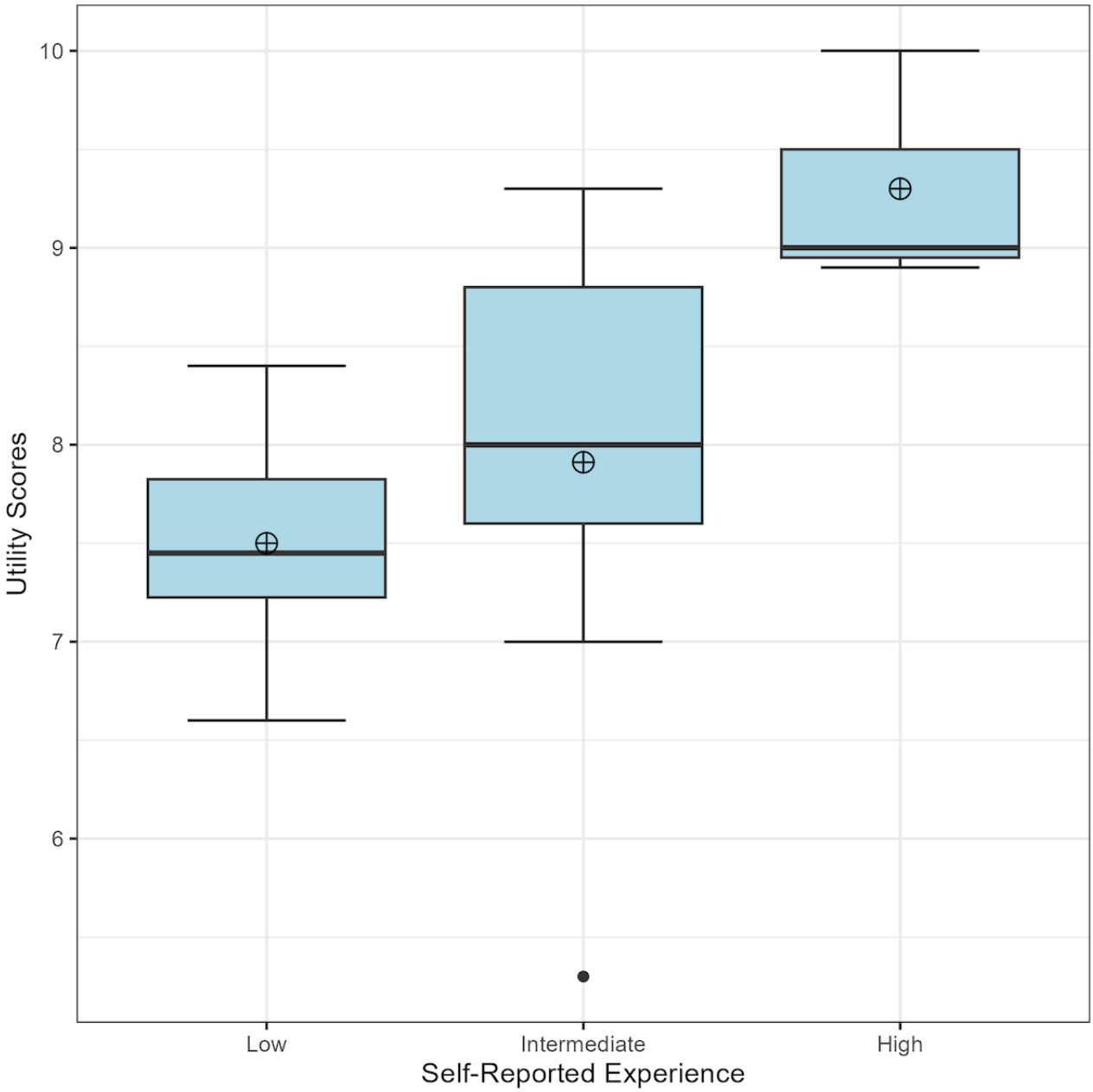

Discussion

Principal Findings

This study demonstrates the feasibility of creating and deploying a digital fabrication and assessment infrastructure for AV models in complex presentations of CVDs, such as iAVMs. The scalable and secure request, creation, distribution, and evaluation of AV models was confirmed by the implementation of 25 AV models for iAVMs from July 2017 to February 2023 at a single high-volume certified stroke center. This retrospective analysis confirmed the utilization of AV models in CVDs, specifically iAVMs. Multiple clinicians requested AV models that were created from a variety of imaging modalities, most commonly from rotational angiography. Of note, requests for AV models persisted despite the decrease in elective procedures during the COVID-19 pandemic and a nationwide iodinated contrast shortage in 2022, which severely impacted imaging and procedure volumes and is indicative of the utility clinicians derive from AV models.

This analysis demonstrates associations between AV model use in CVD and a diverse set of clinical attributes and utilization metrics. Consistent with the complex and diverse population in Louisiana and the surrounding Gulf Coast region, a substantial number of models were created for individuals younger than 40 years, African Americans, and patients on Medicaid. Many of these patients were not only young but otherwise healthy with few if any other comorbidities, besides tobacco use. The feasibility to collect health care finance and utilization metrics such as CPT codes and intensive care unit lengths of stay associated with the request for AV models in complex CVD

XSL•FO
RenderX

presentations was confirmed. The analysis revealed the association of AV models with resource-intensive health resources such as intensive care unit stays, rotational angiography, and neurovascular procedures.

A survey instrument grounded in the literature was successfully created and implemented to assess the impact of AV models in complex CVD at scale. There was strong engagement across a variety of clinical experience levels, particularly APPs, neurological residents, and neurovascular staff, with an overall 92.2% (166/180) survey completion rate. On a set of iAVMs of varying complexity, a diverse array of clinicians reported substantial utility of AV. Staff-level clinicians reported that AV models would have rarely changed their diagnosis compared to traditional 2D viewing. Staff, however, reported that in nearly half of cases, the patient-specific AV models allowed them to appreciate normal and abnormal anatomy that they could not with traditional 2D imaging. Moreover, in 45% (23/51) of the cases, access to the AV model would have changed their therapeutic approach compared to their assessment based on traditional 2D viewing. Staff physicians reported that AV models were much more likely to change their therapeutic approach than their diagnosis. There was no significant difference in reported complexity or utility when clinicians were categorized by training: APP, trainee, staff. There was, however, a statistically significant difference (P=.03) in reported utility when clinicians were categorized by self-reported experience. Interestingly, the difference in reported utility was most pronounced between the high and low experience group, with the most experienced group reporting more utility than those with less experience. This was contrary to our hypothesis and suggests that the ability to understand the idiosyncratic and complex pathologies and extract relevant information from AV models may require a certain level of pre-existing knowledge and clinical expertise. Such a prerequisite appreciation of clinical anatomy may help clinicians derive diagnostic and therapeutic insights from AV models. Alternatively, this observation may be attributable to the psychological phenomena known as the Dunning-Kruger effect in which people with low ability, expertise, or experience regarding a certain type of task or area of knowledge tend to overestimate their ability or knowledge [17].

Strengths and Limitations

This study has multiple strengths attributable to a multidisciplinary approach that focused on clinical operational integration as a foundation for clinical research. Evaluation of an AV intervention required coordinating a cross-functional team consisting of biomedical engineers, research staff, clinical administrators, and clinicians. This study was the impetus to expand and ruggedize a scalable digital fabrication and assessment process for AV models. The survey response rates indicate not only enthusiasm for the field of inquiry but the willingness of these busy and increasingly overextended clinicians to engage in clinical research on 3D AV [18,19]. This points to the importance of continuing to develop and validate AV technologies not just for CVDs but for other complex disease states, including but not limited to spine deformity, genitourinary pathologies, orthopedic trauma, craniomaxillofacial abnormalities, and congenital heart disease.

This study has several limitations. This was a nonrandomized study, which could only establish associations and not causative impacts of AV modeling. The ability to extrapolate these findings is also limited by its small sample size and focus on a single center. Clinician respondents were basing feedback on hypothetical use cases and not the actual implementation of the 10 models in diagnosis or treatment planning. Moreover, the clinicians were all colleagues who may have influenced each other's responses, which may have introduced a diffusion of treatment effect as described by Urban and van Eeden-Moorefield [20]. Finally, the results of the statistical testing may be due to the small sample size rather than the absence or presence of differential utility among varying experience levels.

Challenges and Future Directions

The creation of AV models remains resource-intensive in terms of expertise, time, and infrastructure. Complex disease states where AV models are likely to be the most impactful are often low in prevalence requiring longer periods of study and collaboration between multiple sites, engendering a gap in the literature. These barriers have limited the number of conclusive studies validating the clinical and utilization impact of 3D AV models. Approaches to increase the efficiency of AV model creation, for example, machine-learning enhanced segmentation, may mitigate these obstacles to validation and warrant further research and development.

This study expands and utilizes a scalable digital fabrication infrastructure for future larger multisite randomized studies to assess the efficacy of AV modeling. Such future trials could examine the utility of AV models in clinical care delivery by incorporating technical resource requirements such as segmentation time, utilization measures such as operative time and radiation exposure, as well as clinical outcome metrics such as functional status and hospital stay. This study establishes the feasibility of collecting medical billing and utilization metrics associated with 3D AV models that can be used in future cost-effective analyses. This infrastructure could also be applied to understand the use and impact of AV models in other disease states that would benefit from AV techniques. Future studies could incorporate quantitative and qualitative assessments of other AV modalities such as 3D printing or extended reality to investigate the different applications and utility of digital versus physical models in visualization and simulation. Finally, this infrastructure could be utilized to study the impact AV models have on health literacy and patient experience. The use of AV is quickly being adopted in clinical training and care delivery but requires further investigation and validation.

Conclusion

This pilot feasibility study establishes the infrastructure to create and assess patient-specific AV models in a digital and scalable way. This study confirms the use of AV models in complex, resource-intensive disease states such as iAVMs. This study demonstrates enthusiasm for the use of AV across different specialties and experience levels involved in the management of CVD. A heterogenous cohort of clinicians reported substantial utility of AV modeling when compared to standard 2D imaging. Counter to our hypothesis, those clinicians with the most

XSL•FO
RenderX

experience found AV models the most useful. Finally, staff level physicians reported at high rates that AV models modified therapeutic approaches and provided anatomical insights not appreciated in standard 2D imaging. Almost universally, clinicians across different specialties and experience levels affirmed that that their patients would benefit from the ability to view their pathology-specific 3D AV models. Collectively, these findings support the idea that compared to traditional 2D viewing, patient-specific AV models are useful to a diverse array of clinicians involved in the treatment of complex disease states, such as iAVMs.

Acknowledgments

We thank our colleagues at Ochsner Health, specifically the academics, philanthropy, and neurosciences departments who supported the development and implementation of this work. The authors would like to thank the faculty and staff of Cedars Sinai Master's Health Delivery Sciences as well as the 2023 cohort for their guidance, support, and mentorship. Research reported in this publication was supported by the Terence D D'Souza MD Fellowship in Neuroscience and the Ochsner Louisiana State University Collaborative Intramural Research program. The funders had no role in the study design, data collection, analysis, or interpretation of results. Generative artificial intelligence, specifically ChatGPT 4o V 1.2024.346, was used to assist in literature review and improve manuscript content.

Data Availability

The datasets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

KS conceptualized, designed, and supported the implementation of this study as well as led the writing of this manuscript. VB supported the data collection, cleaning, and overall implementation of this study. AR supported the biostatistical analysis of the data. CC developed the web-based viewer, performed the segmentation, and supported study implementation. KH designed and deployed the digital request and assessment infrastructure. KS, VB, AR, CC, and KH wrote, edited, and approved the final manuscript.

Conflicts of Interest

KS has past but no current roles as an advisor to digital health companies, none of which have any financial or intellectual property rights to this research. He also engages in medical-legal consulting. During the period when this research was conducted, VB was affiliated with Ochsner Health. VB currently works in the digital health industry. VB has no financial, consulting, patent, or intellectual property interests related to this research, and declares no conflicts of interest that could have influenced this work. AR, CC, and KH have no disclosures to report.

Multimedia Appendix 1
Survey template: clinician, on-screen model survey, and survey email template.
[DOCX File , 18 KB - formative_v9i1e51939_app1.docx]

Multimedia Appendix 2
Representative 2D viewer included in surveys.
[MP4 File (MP4 Video), 11859 KB - formative_v9i1e51939_app2.mp4]

Multimedia Appendix 3
Representative web-based 3D viewer included in surveys.
[ZIP File (Zip Archive), 9237 KB - formative_v9i1e51939_app3.zip]

Multimedia Appendix 4
Photograph of 3D printing of an intracerebral arteriovenous malformation example.
[PDF File (Adobe PDF File), 1722 KB - formative_v9i1e51939_app4.pdf]

References

1. Portegies M, Koudstaal PJ, Ikram MA. Cerebrovascular disease. Handb Clin Neurol 2016;138:239-261. [doi: 10.1016/B978-0-12-802973-2.00014-8] [Medline: 27637962]
2. Ovbiagele B, Goldstein LB, Higashida RT, Howard VJ, Johnston SC, Khavjou OA, et al. Forecasting the future of stroke in the United States. Stroke 2013 Aug;44(8):2361-2375. [doi: 10.1161/str.0b013e31829734f2]

XSL•FO
RenderX

3. Colby S, Ortman JM. Projections of the size and composition of the US population: 2014 to 2060. US Census Bureau. 2015. URL: https://www.census.gov/content/dam/Census/library/publications/2015/demo/p25-1143.pdf [accessed 2023-04-16]
4. Kim BM, Baek J, Heo JH, Kim DJ, Nam HS, Kim YD. Effect of cumulative case volume on procedural and clinical outcomes in endovascular thrombectomy. Stroke 2019 May;50(5):1178-1183. [doi: 10.1161/strokeaha.119.024986]
5. Perri JL, Nolan BW, Goodney PP, DeMartino RR, Brooke BS, Arya S, et al. Factors affecting operative time and outcome of carotid endarterectomy in the Vascular Quality Initiative. J Vasc Surg 2017 Oct;66(4):1100-1108 [FREE Full text] [doi: 10.1016/j.jvs.2017.03.426] [Medline: 28712813]
6. Ghosn Y, Khdhir M, Jabbour Y, Dushfunian D, Kobeissi I, Abbas N, et al. Factors affecting radiation dose, radiation exposure time and procedural time in arterial embolization for active hemorrhage. Emerg Radiol 2024 Oct;31(5):641-652. [doi: 10.1007/s10140-024-02262-w] [Medline: 38955874]
7. Branson T, Shapiro L, Venter RG. Observation of patients' 3D printed anatomical features and 3D visualization technologies improve spatial awareness for surgical planning and in-theatre performance. Adv Exp Med Biol 2021;1334:23-37. [doi: 10.1007/978-3-030-76951-2_2] [Medline: 34476743]
8. Popescu D, Marinescu R, Laptoiu D, Deac GC, Cotet CE. DICOM 3D viewers, virtual reality or 3D printing - a pilot usability study for assessing the preference of orthopedic surgeons. Proc Inst Mech Eng H 2021 Sep;235(9):1014-1024. [doi: 10.1177/09544119211020148] [Medline: 34176364]
9. Liang F, Yang Y, Luo L, Liao B, Zhang G, Ou S, et al. Endovascular treatment of complex middle cerebral artery aneurysms using TuBridge flow diverters. Interv Neuroradiol 2020 Oct;26(5):539-546 [FREE Full text] [doi: 10.1177/1591019920946216] [Medline: 32722987]
10. Chepelev L, Wake N, Ryan J, Althobaity W, Gupta A, Arribas E, RSNA Special Interest Group for 3D Printing. Radiological Society of North America (RSNA) 3D printing Special Interest Group (SIG): guidelines for medical 3D printing and appropriateness for clinical scenarios. 3D Print Med 2018 Nov 21;4(1):11 [FREE Full text] [doi: 10.1186/s41205-018-0030-y] [Medline: 30649688]
11. Chae MP, Lin F, Spychal RT, Hunter-Smith DJ, Rozen WM. 3D-printed haptic "reverse" models for preoperative planning in soft tissue reconstruction: a case report. Microsurgery 2015 Feb;35(2):148-153. [doi: 10.1002/micr.22293] [Medline: 25046728]
12. Sun Z. A systematic review of 3-D printing in cardiovascular and cerebrovascular diseases. Anatol J Cardiol 2017:423-435. [doi: 10.14744/anatoljcardiol.2017.7464]
13. McGuire LS, Fuentes A, Alaraj A. Three-dimensional modeling in training, simulation, and surgical planning in open vascular and endovascular neurosurgery: a systematic review of the literature. World Neurosurg 2021 Oct;154:53-63. [doi: 10.1016/j.wneu.2021.07.057] [Medline: 34293525]
14. Rybicki FJ. 3D Printing in Medicine: an introductory message from the Editor-in-Chief. 3D Print Med 2015;1(1):1 [FREE Full text] [doi: 10.1186/s41205-015-0001-5] [Medline: 30050970]
15. Ryan J, Plasencia J, Richardson R, Velez D, Nigro JJ, Pophal S, et al. 3D printing for congenital heart disease: a single site's initial three-yearexperience. 3D Print Med 2018 Nov 08;4(1):10 [FREE Full text] [doi: 10.1186/s41205-018-0033-8] [Medline: 30649650]
16. Sebök M, Dufour JP, Cenzato M, Kaku Y, Tanaka M, Tsukahara T, et al. When is diagnostic subtraction angiography indicated before clipping of unruptured and ruptured intracranial aneurysms? an international survey of current practice. Acta Neurochir Suppl 2021;132:9-17 [FREE Full text] [doi: 10.1007/978-3-030-63453-7_2] [Medline: 33973023]
17. Kruger J, Dunning D. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. J Pers Soc Psychol 1999 Dec;77(6):1121-1134. [doi: 10.1037//0022-3514.77.6.1121] [Medline: 10626367]
18. West CP, Dyrbye LN, Shanafelt TD. Physician burnout: contributors, consequences and solutions. J Intern Med 2018 Jun;283(6):516-529 [FREE Full text] [doi: 10.1111/joim.12752] [Medline: 29505159]
19. Rodrigues H, Cobucci R, Oliveira A, Cabral JV, Medeiros L, Gurgel K, et al. Burnout syndrome among medical residents: A systematic review and meta-analysis. PLoS One 2018;13(11):e0206840 [FREE Full text] [doi: 10.1371/journal.pone.0206840] [Medline: 30418984]
20. Urban J, van Eeden-Moorefield BM. Designing and proposing your research project. In: Concise Guides to Conducting Behavioral, Health, and Social Science Research Series. Washington DC: American Psychological Association; 2018:119-127.

Abbreviations

APP: advanced practice provider
AV: advanced visualization
CPT: current procedural terminology
CVD: cerebrovascular disease
DICOM: Digital Imaging and Communications in Medicine
iAVM: intracerebral arteriovenous malformation
SMG: Spetzler-Martin Grade

XSL•FO
RenderX

Edited by A Mavragani; submitted 20.08.23; peer-reviewed by M Waleed, M Aksoy; comments to author 06.09.23; revised version received 13.01.25; accepted 22.01.25; published 21.02.25.

Please cite as:
Sarkar K, Bhimarasetty V, Rahim A, Curtis C, Hughes K
Assessing the Feasibility and Utility of Patient-Specific 3D Advanced Visualization Modeling in Cerebrovascular Disease: Retrospective Analysis and Prospective Survey Pilot Study
JMIR Form Res 2025;9:e51939
URL: https://formative.jmir.org/2025/1/e51939
doi:10.2196/51939
PMID:

©Korak Sarkar, Vishal Bhimarasetty, Abdul Rahim, Colin Curtis, Kimberly Hughes. Originally published in JMIR Formative Research (https://formative.jmir.org), 21.02.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effect on Response Rates of Adding a QR Code to Patient Consent Forms for Qualitative Research in Patients With Cancer: Pilot Randomized Controlled Trial

Rebecca Wyse[1,2], PhD; Erin Forbes[1,2], PhD; Grace Norton[1], BSc; Priscilla Viana Da Silva[1,2], PhD; Kristy Fakes[1,2], PhD; Sally Ann Johnston[3], BSc; Stephen R Smith[1,2,3], PhD; Alison Zucca[1,2], PhD

[1]School of Medicine and Public Health, The University of Newcastle, Callaghan, NSW, Australia

[2]Hunter Medical Research Institute, New Lambton Heights, NSW, Australia

[3]Hunter New England Local Health District, John Hunter Hospital, New Lambton Heights, NSW, Australia

Corresponding Author:
Rebecca Wyse, PhD
School of Medicine and Public Health
The University of Newcastle
University Dr
Callaghan, NSW, 2308
Australia
Phone: 61 02 40420272
Email: rebecca.wyse@health.nsw.gov.au

Abstract

Background: The successful conduct of health and medical research is largely dependent on participant recruitment. Effective, yet inexpensive methods of increasing response rates for all types of research are required. QR codes are now commonplace, and despite having been extensively used to recruit study participants, a search of the literature failed to reveal any randomized trial investigating the effect of adding a QR code on qualitative research response rates.

Objective: This study aimed to collect data on rates of response, consent, and decline among patients with cancer, and the average time taken to respond following randomization to receive either a QR code or no QR code on the patient consent form for a qualitative research study.

Methods: This was a pilot randomized controlled trial (RCT) embedded within a qualitative research study. In total, 40 eligible patients received a recruitment pack for the qualitative study, which included an information statement, a consent form, and an addressed, stamped envelope to return their consent form. Patients were randomized 1:1 to the control (standard recruitment pack only) or intervention group (standard recruitment pack including modified consent form with a QR code).

Results: In total, 27 out of 40 patients (age: mean 63.0, SD 14.8 years; 45% female) responded to the consent form. A lower proportion of the QR code group (60%) responded (odd ratio [OR] 0.57, 95% CI 0.14-2.37; *P*=.44), compared to 75% of the standard recruitment group. However, a higher proportion of the QR group (35%) consented (OR 1.84, 95% CI 0.41-8.29; *P*=.43), compared to the standard recruitment group (20%). A lower proportion of the QR group (25%) declined (OR 0.34, 95% CI 0.09-1.38; *P*=.13) relative to the standard recruitment group (55%). The mean response time of the QR code group was 16 days (rate ratio [RR] 0.79, 95% CI 0.47-1.35; *P*=.39) compared to 19 days for the standard recruitment group. None of the age-adjusted analyses were statistically significant.

Conclusions: This underpowered pilot study did not find any evidence that offering an option to respond through a QR code on a patient consent form for a qualitative study increased the overall patient response rate (combined rate of consent and decline). However, there was a nonsignificant trend, indicating that more patients who received the QR code consented compared to those who did not receive the QR code. This study provides useful preliminary data on the potential impact of QR codes on patient response rates to invitations to participate in qualitative research and can be used to inform fully powered RCTs.

Trial Registration: OSF Registries 10.17605/OSF.IO/PJ25X; https://doi.org/10.17605/OSF.IO/PJ25X

(JMIR Form Res 2025;9:e66681) doi:10.2196/66681

XSL•FO
RenderX

KEYWORDS

QR code; qualitative research; cancer; randomized controlled trial; RCT; patient recruitment; consent forms; response rates

Introduction

The successful conduct of health and medical research depends on effective recruitment strategies. Low recruitment rates affect both quantitative and qualitative research designs [1-3], and there is a need to identify effective and affordable methods that are minimally burdensome. For example, a 2019 study reported the costs of 14 recruitment strategies among patients with chronic diseases and reported that 20% of the study budget was spent on recruitment [4]. Given the recent widespread adoption of QR codes [5-8], they offer potential as a valuable and cost-effective recruitment tool. A number of randomized controlled trials (RCTs) have investigated the use of QR codes to recruit participants for quantitative surveys [9]. However, the evidence of the effectiveness of QR codes in increasing survey recruitment is equivocal, with earlier studies [10,11], tending to find no impact of QR codes on response rates, and later studies tending to find a small but significant increases [9,12]. Authors of these studies note that earlier research should be interpreted with caution, given the widespread use and familiarity with QR codes, which is relatively recent [9], and has increased rapidly following the COVID-19 epidemic [13]. A 2024 large RCT (N=5550) that evaluated the effect of including a QR code in a printed and mailed recruitment to a web-based survey found a small but significant increase in survey participation (+1.31%) [12]. It also found that including QR codes resulted in higher participation rates from some harder-to-reach groups, including younger people and single people [12].

However, there is little research about the effect of QR codes on consent rates to other research forms, such as recruitment to research trials or qualitative research, which often require ongoing or more maintained or intense commitment than a one-off survey. A 2024 study randomized recruitment methods ([Twitter rebrand as] X, Facebook [Meta] or QR code displayed on a poster) to a clinical trial and found little to no engagement with the QR code relative to the other recruitment strategies; however, this may have been due to the format (ie, poster) rather than the technology [8].

A literature search failed to identify any study that empirically tests the use of QR codes among people invited to participate in qualitative health research. Given the dearth of studies of this nature, the primary objective of this pilot RCT was to determine whether adding a QR code to a participant consent form for a qualitative research study increased the participant consent rate (ie, the proportion of patients who responded and consented to participate). Secondary objectives were to collect data regarding differences in (1) the proportion of patients who responded (either to consent or decline; this outcome was registered in the Open Science Framework as the proportion of patients who did not respond, but for ease of reporting and understanding, the inverse is reported instead [ie, the proportion who did respond])), (2) the proportion of patients who responded and declined to participate, and (3) the average time (in days) taken to respond measured from the date the recruitment packs were sent.

Methods

Study Design

This was a pilot RCT embedded within a qualitative study with patients with colorectal cancer [14]. The purpose of the qualitative study was to determine the views of patients with colorectal cancer about a Digital Health Intervention (“RecoverEsupport”) [15] to enhance their recovery from surgery. The qualitative study required participants to review a newly developed web-based support program and to provide feedback in a recorded telephone-based interview (approximately 20 minutes) that would be scheduled at a time convenient to them. The interview asked them to provide feedback about the relevance and ease of understanding of the program, as well as the information they could recall and the likely impact of the program.

Participants

Participants were recruited from a Surgical Department at a public hospital in a noncapital metropolitan region of Australia. The Colorectal Cancer Liaison Nurse identified eligible patients from medical records. The eligibility criteria for this pilot RCT were consistent with the eligibility criteria for the qualitative study. Patients were invited to participate in the qualitative study through a recruitment pack containing a printed consent form and a prepaid return envelope (mailed on 1 November 2021). Reminder packs were sent to all nonresponders 3 weeks later (22 November 2021).

For the qualitative study in which the QR code study was embedded, the patient eligibility criteria were as follows: (1) a diagnosis of colorectal cancer, (2) being 18-90 years old, and (3) having undergone a bowel resection for colorectal cancer within the last 6 months (excluding stoma reversals and secondary bowel surgeries).

Allocation

An independent statistician randomly allocated the patients in a 1:1 ratio using a random number function in Microsoft Excel to the intervention or control group (Figure 1).

XSL•FO
RenderX

Figure 1. CONSORT Flowchart.

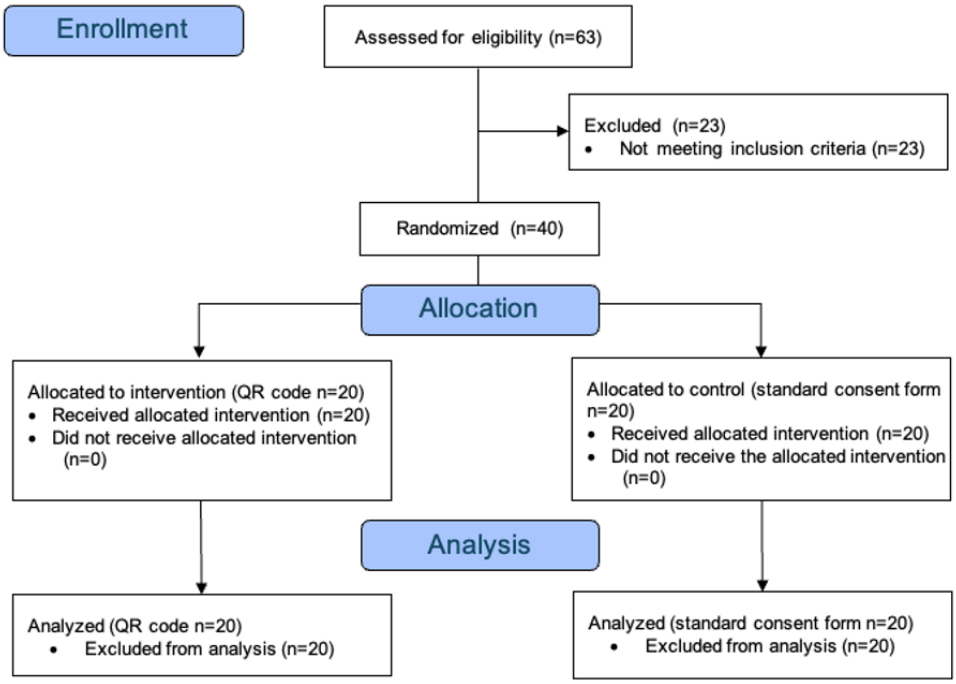

Standard Recruitment (Control)

These patients received the standard consent form (without the QR code). The form included 2 tick boxes for patients to indicate their decision to either consent or decline to participate in the qualitative study.

QR Code Group (Intervention)

These patients received an identical consent form to the control group, except this version included a QR code and instructions for accessing the web-based consent form to provide their consent. Scanning the QR code directed the participant to a web-consent-form, identical to the standard consent form, hosted within a REDCap (Research Electronic Data Capture; Vanderbilt University) database. REDCap automatically recorded the date of completion once the patient responded the web-based consent form.

Outcomes

The primary outcome was the proportion of patients who consented to participate in the qualitative study (through any channel, QR code, or mail). Secondary outcomes included (1) the proportion of patients who responded (either to consent or decline), (2) the proportion of patients who responded and declined to participate in the qualitative study, and (3) the average time in days taken to respond (measured from 1 November 2021 until either the postmark date on the return envelope or the date the web-based consent form was completed).

Statistical Analysis

Descriptive statistics are presented as counts (%) for categorical variables or mean (SD) for continuous variables. An intention-to-treat approach was used. Unadjusted and age-adjusted logistic regression models were fitted to estimate the odds ratio (OR) and 95% CIs. The between-group difference for the time taken to respond was analyzed using negative binomial instead of Poisson regression modeling (due to overdispersion) and unadjusted and age-adjusted rate ratios (RRs) are presented. Assumptions for the logistic regression model and negative binomial regression model were checked and found to be reasonable.

Ethical Considerations

The pilot was approved by the Ethics Committees of the Hunter New England Local Health District (2019/ETH00869) and the University of Newcastle (H-2015-0364) and was prospectively registered [16] and is reported in accordance with CONSORT (Consolidated Standards of Reporting Trials) guidelines [17]. Each invited patient received a patient information sheet inviting them to participate in the qualitative research, assuring them that their participation was entirely voluntary and that any information they provided would remain confidential. No incentive was offered for patients to participate. All project staff members were bound by confidentiality agreements. Collected data are stored in a deidentified format.

Results

Overview

Of the 63 patients assessed for eligibility, 23 were ineligible, most commonly because they did not have internet access (7/23 patients). The mean age was 58.60 (SD 17.38) years for the QR code, 67.40 (SD 10.35) years for the standard recruitment, totaling 63 (SD 14.8) years for both groups. In total, 45% (n=18) of the sample was female; 40% (n=8) for the QR code and 50% (n=10) for the standard recruitment groups were female. The outcomes are presented below and in Table 1 in sequential order (ie, overall response rate, consent rate, and decline rate).

XSL•FO
RenderX

Table 1. Pilot outcomes response rates and time to respond to the invitation to participate in a qualitative research study for participants randomized to receive a quick response code (intervention) or no quick response code (standard recruitment control).

	QR[a] code (Intervention) (n=20), (%)	Standard recruitment (Control) (n=20) (%)	Unadjusted effect estimate (95% CI)	*P* value	Adjusted for age effect estimate (95% CI)	*P* value
Responded (overall response rate)[a]	12 (60)	15 (75)	0.50 (0.13-1.93)	0.3145	0.57 (0.14-2.37)	.44
Responded and consented (primary outcome)[a]	7 (35)	4 (20)	2.15 (0.52-9.00)	0.2930	1.84 (0.41-8.29)	.43
Responded and declined (secondary outcome)[a]	5 (25)	11 (55)	0.27 (0.07-1.04)	0.0577	0.34 (0.09-1.38)	.13
Response time in days (Secondary outcome)[b]	16.33 (11.12)	19.27 (12.40)	0.85 (0.50-1.43)	0.5361	0.79 (0.47-1.35)	.39

[b]Odds ratio.

[c]Rate ratio.

Overall Response Rate

A total of 27 patients responded to the consent form, either to consent or decline; 12/20 (60%) from the QR code group and 15/20 (75%) from the standard recruitment group. The unadjusted effect estimate was 0.50 (95% CI 0.13 to 1.93, *P*=.31), indicating that the odds of the intervention group responding were 50% lower compared to the control group. After adjusting for age, the OR was 0.57 (95% CI 0.14 to 2.37, *P*=.44). Although the results suggest a lower overall response rate in the QR code group relative to the standard recruitment group, neither analysis was significant.

Proportion of Patients Who Responded and Consented (Primary Outcome)

Overall, 11/40 (27.5%) patients consented to participate in the qualitative study. Seven out of 40 consenting patients were from the QR code group (representing 35% of this group), and 4 were from the standard recruitment group (representing 20% of this group). The unadjusted OR was 2.15 (95% CI 0.52-9.00, *P*=.29), indicating that the odds of patients in the intervention group consenting was 2.15 times the odds among control patients, although this was not statistically significant. After adjusting for age, the OR was 1.84 (95% CI 0.41-8.29). However, neither was significant.

Proportion of Patients Who Responded and Declined to Participate

A total of 16 out of the 40 patients declined to participate in the qualitative study; 5 declining patients were in the QR code group (representing 25% of this group), and 11 were in the standard recruitment group (representing 55% of this group). The unadjusted OR of the QR code group declining to participate was 0.27 (95% CI 0.07-1.04), which was borderline significant (*P*=.06). The age-adjusted OR was 0.34 (95% CI 0.06-1.87, *P*=.21) and was not statistically significant.

Average Time (in Days) Taken to Respond

The average response time was 16 days (SD 11) days for the QR code group and 19 (SD 12) days for the standard recruitment group. The unadjusted incident RR for the QR code group compared to the standard recruitment group was 0.85 (95% CI 0.5-1.43, *P*=.54), indicating that the response time of the QR code group was 15% lower than the response time for the standard recruitment group. After adjusting for age, the RR was 0.79 (95% CI 0.47-1.35), indicating the QR code group response time was 21% lower than the control group. However, neither was statistically significant.

Discussion

Principal Findings

Given the dearth of studies of this nature, the aim of this pilot RCT was to determine the effect on response rates when a QR code was added to a participant consent form for a qualitative research study. Specifically, this pilot investigated the participant response rate, consent rate (primary outcome), rate of decline, as well as the average time taken to respond. Out of 40 patients invited to participate in a qualitative research study were randomized to receive either a standard consent form (control) or a consent form that included a QR code (intervention). The overall response rate (ie, either to consent or decline) was lower in the intervention group relative to the control group, however, the intervention consent rate was higher, the decline rate was lower, and the average time to respond to the invitation to participate was shorter in the intervention group (16 vs 19 days). None of the adjusted estimates of between-group differences were statistically significant. However, our results suggest a trend of patients in the QR code group having higher odds of consenting and lower odds of declining (relative to the control group), but lower odds of responding at all (ie, either to consent or decline). These results should be investigated further in a fully powered RCT.

The overall pattern of results is interesting, despite being nonsignificant. It suggests that the addition of a QR code to a consent form for qualitative research resulted in higher consent rates and lower rates of decline. Both outcomes are desirable for researchers who often spend a large portion of their research budget on recruitment. The shorter response time in the intervention group also suggests that the use of a QR code in recruitment materials could confer benefits to researchers in

XSL•FO
RenderX

terms of more timely information about likely recruitment outcomes.

This is the first pilot RCT to investigate the effect of adding a QR code to a consent form within the context of recruitment of patients with cancer to a qualitative study. It is important to consider the rapidly changing context in which QR code research has taken place when comparing these results to the existing literature. Over the past 5 years, responses to the COVID-19 pandemic have made QR code levels of awareness, familiarity, and skill in the community is now much higher. The more recent RCTs testing the addition of QR codes to surveys, showed small but significant increases, of approximately 1-2% in consent rates [9,12]. The magnitude of the changes observed in the current pilot are much greater (15% increase in consent rates), despite being nonsignificant due to under-powering. This may be due to the targeted nature of the study (ie, people who had previously had colorectal cancer surgery, compared with a general population statewide public health survey) [12] and due to differences in the specific nature of what they were being asked to do (ie, review an intervention designed to support patients with colorectal cancer going through surgery and provide feedback in a 1:1 interview, vs complete a web-based public health survey).

The randomized design is a strength; however, a few limitations should be considered when interpreting the findings of this term. Despite the randomization, due to the small sample size, the groups were not evenly distributed in age, which was subsequently adjusted for. The low sample size is another limitation, although it is still within the recommended range for pilot studies [18]. Therefore, as a next step, a fully powered RCT is recommended. Furthermore, the study did not collect data about whether the participants actually used the QR to respond (only overall response rates) or the characteristics of those users who responded using this channel. This would be valuable information to collect and analyze in future studies to better understand the impact of QR codes on the diversity of recruited participants. Given the context in which this research occurred, it is unclear the extent to which these findings would generalize to the general population, younger people or both groups. Future RCTs should also allow for subgroup analysis to determine the specific effects of QR codes on recruiting different age groups.

Conclusion

Despite the nonsignificance of the findings in this underpowered pilot trial, this research question warrants further investigation. Given their ease of use and inexpensive application, using QR codes on consent forms may hold promise for improving recruitment to qualitative research studies. As recruitment often consumes a large proportion of research budgets, even a small percentage of change could have important benefits for researchers and research programs. QR codes may be a low-risk strategy to complement other attempts to maximize research response rates, although a fully powered RCT is required to confirm this. Further research using larger and more diverse samples is warranted. This pilot study provides important preliminary data on which such studies could be based.

Acknowledgments

The authors acknowledge the patients included in this research, the research assistants of the Health Behavior Research Collaborative (HBRC) and the University of Newcastle (UoN); and Hunter Medical Research Institute (HMRI) staff members who supported this research. The research was supported by NHMRC, UoN, and HMRI grants. The content is solely the responsibility of the authors.

This research was funded as part of a National Health and Medical Research Council Partnership Grant (NHMRC APP1169822). RW is supported by a Heart Foundation postdoctoral fellowship (ID: 102156) and a Cancer Institute NSW Early Career Fellowship (ECF1355).

Data Availability

The datasets generated during this study or analysis of the data are available from the corresponding author on reasonable request and pending ethical approval.

Conflicts of Interest

None declared.

Multimedia Appendix 1
CONSORT (Consolidated Standards of Reporting Trials) extension for pilot and feasibility studies checklist.
[PDF File (Adobe PDF File), 91 KB - formative_v9i1e66681_app1.pdf]

References

1. Kasenda B, von Elm E, You J, Blümle A, Tomonaga Y, Saccilotto R, et al. Prevalence, characteristics, and publication of discontinued randomized trials. JAMA 2014 Mar 12;311(10):1045-1051 [FREE Full text] [doi: 10.1001/jama.2014.1361] [Medline: 24618966]

XSL•FO
RenderX

2. Bernardez-Pereira S, Lopes RD, Carrion MJM, Santucci EV, Soares RM, de Oliveira Abreu M, Methodological Evaluation of clinical TriAls Study Group. Prevalence, characteristics, and predictors of early termination of cardiovascular clinical trials due to low recruitment: insights from the ClinicalTrials.gov registry. Am Heart J 2014 Aug;168(2):213-9.e1. [doi: 10.1016/j.ahj.2014.04.013] [Medline: 25066561]
3. Spratling R. Recruitment of medically fragile children and adolescents: lessons learned from qualitative research. J Pediatr Health Care 2013 Jan;27(1):62-65. [doi: 10.1016/j.pedhc.2012.08.001] [Medline: 23084469]
4. Kakumanu S, Manns BJ, Tran S, Saunders-Smith T, Hemmelgarn BR, Tonelli M, et al. Cost analysis and efficacy of recruitment strategies used in a large pragmatic community-based clinical trial targeting low-income seniors: a comparative descriptive analysis. Trials 2019 Oct 07;20(1):577 [FREE Full text] [doi: 10.1186/s13063-019-3652-5] [Medline: 31590686]
5. Biddle N, Gray M, Sollis K. The ANU centre for social research and methods. The use of QR codes to identify COVID-19 contacts and the role of data trust and data privacy. 2021 Nov 5. URL: https://tinyurl.com/438cey75 [accessed 2024-12-01]
6. Cohney S, Cheong M. COVID Down Under: where did Australia?s pandemic apps go wrong? arXiv:2212.12178 2023 Jun 23 [FREE Full text] [doi: 10.1109/ethics57328.2023.10154912]
7. Grekousis G, Liu Y. Digital contact tracing, community uptake, and proximity awareness technology to fight COVID-19: a systematic review. Sustain Cities Soc 2021 Aug;71:102995 [FREE Full text] [doi: 10.1016/j.scs.2021.102995] [Medline: 34002124]
8. Brodie K, Madden LL, Rosen CA. Applications of quick response (QR) codes in medical education. J Grad Med Educ 2020;12(2):138-140 [FREE Full text] [doi: 10.4300/JGME-D-19-00516.1] [Medline: 32322342]
9. Endres K, Heiden EO, Park K. Surv methods insights field SMIF. The Effect of quick response (QR) codes on response rates and respondent composition: results from a statewide experiment. 2024 Aug 05. URL: https://surveyinsights.org/?p=19389 [accessed 2024-12-01]
10. Marlar J. Do Quick Response Codes Enhance or Hinder Surveys?.: Gallup; 2018 Aug 30. URL: https://news.gallup.com/opinion/methodology/241808/quick-response-codes-enhance-hinder-surveys.aspx [accessed 2024-12-02]
11. Lugtig P, Luiten A. Do shorter stated survey length and inclusion of a QR code in an invitation letter lead to better response rates? Surv Methods Insights Field. 2021 Feb 12. URL: https://surveyinsights.org/?p=14216 [accessed 2021-02-12]
12. Endres K, Heiden EO, Park K, Losch ME, Harland KK, Abbott A. Experimenting with QR codes and envelope size in push-to-web surveys. J Surv Stat Methodol 2023 Apr 26;12(4):893-905 [FREE Full text] [doi: 10.1093/jssam/smad008]
13. Gostin I. Forbes Communications Council. How the pandemic saved the QR code from extinction. 2021 May 25. URL: https://tinyurl.com/mka68j87 [accessed 2024-12-02]
14. Zucca A, Mansfield E, Sanson-Fisher R, Wyse R, Johnston SA, Fakes K, et al. Perceived provision of perioperative information and care by patients who have undergone surgery for colorectal cancer: a cross-sectional study. Int J Environ Res Public Health 2022 Nov 18;19(22):15249 [FREE Full text] [doi: 10.3390/ijerph192215249] [Medline: 36429966]
15. Wyse R, Smith S, Zucca A, Fakes K, Mansfield E, Johnston SA, et al. Effectiveness and cost-effectiveness of a digital health intervention to support patients with colorectal cancer prepare for and recover from surgery: study protocol of the RecoverEsupport randomised controlled trial. BMJ Open 2023 Mar 06;13(3):e067150 [FREE Full text] [doi: 10.1136/bmjopen-2022-067150] [Medline: 36878662]
16. Wyse R, Zucca A, Norton G, Johnston SA. The Effect on study consent rates of adding a QR code to patient consent forms: a randomised controlled Pilot Study. 2021 Sep 23. URL: https://doi.org/10.17605/OSF.IO/PJ25X [accessed 2021-09-23]
17. Schulz KF, Altman DG, Moher D. CONSORT 2010 Statement: Updated guidelines for reporting parallel group randomised trials. J Clin Epidemiol 2010 Aug;63(8):834-840 [FREE Full text] [doi: 10.1016/j.jclinepi.2010.02.005] [Medline: 20346629]
18. Kunselman AR. A brief overview of pilot studies and their sample size justification. Fertil Steril 2024 Jun;121(6):899-901. [doi: 10.1016/j.fertnstert.2024.01.040] [Medline: 38331310]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
OR: odds ratio
RCT: randomized controlled trial
REDCap: Research Electronic Data Capture
RR: rate ratio

XSL•FO
RenderX

Edited by A Mavragani; submitted 19.09.24; peer-reviewed by F Temidayo, M Tandon; comments to author 08.11.24; revised version received 01.12.24; accepted 09.12.24; published 21.02.25.

Please cite as:
Wyse R, Forbes E, Norton G, Viana Da Silva P, Fakes K, Johnston SA, Smith SR, Zucca A
Effect on Response Rates of Adding a QR Code to Patient Consent Forms for Qualitative Research in Patients With Cancer: Pilot Randomized Controlled Trial
JMIR Form Res 2025;9:e66681
URL: https://formative.jmir.org/2025/1/e66681
doi:10.2196/66681
PMID:

©Rebecca Wyse, Erin Forbes, Grace Norton, Priscilla Viana Da Silva, Kristy Fakes, Sally Ann Johnston, Stephen R Smith, Alison Zucca. Originally published in JMIR Formative Research (https://formative.jmir.org), 21.02.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Testing and Iterative Improvement of the CEN ISO/TS 82304-2 Health App Quality Assessment: Pilot Interrater Reliability Study

Anna-Lena Frey[1], PhD; Diana Matei[1], MEng; Ben Phillips[1], BSc; Adam McCabe[1], BSc; Rachel Fuller[1], BSc; Begoña Laibarra[2], MSc; Laura Alonso[2], BSc; Victor de la Hoz[2], MEng; Carme Pratdepadua Bufill[3], MSc; Berta Llebot Casajuana[3], BSc; Giuseppe D'Avenio[4], MSc; Pier Angelo Sottile[5], MSc; Laura Melania Rocchi[6], BA; Matteo Errera[6]; Yasmine Laaissaoui[7], MSc; Michael Cardinal[7], PharmD; Menno Kok[8], PhD; Petra Hoogendoorn[9], MSc

[1]Organisation for the Review of Care and Health Apps (ORCHA), Daresbury, United Kingdom
[2]Software Quality Systems, Getxo, Spain
[3]TIC Social Health Foundation, Ministry of Health, Government of Catalonia, Barcelona, Spain
[4]National Center for Innovative Technologies in Public Health, National Institute of Health, Rome, Italy
[5]BI Health Srl, Rome, Italy
[6]BaasBox Srl, Rome, Italy
[7]TherAppX, Granby, QC, Canada
[8]EIT Health Belgium and The Netherlands, Rotterdam, The Netherlands
[9]National eHealth Living Lab, Department of Public Health and Primary Care, Leiden University Medical Center, Leiden, The Netherlands

Corresponding Author:
Petra Hoogendoorn, MSc
National eHealth Living Lab
Department of Public Health and Primary Care
Leiden University Medical Center
Albinusdreef 2
Leiden, 2333 ZA
The Netherlands
Phone: 31 654341785
Email: a.p.y.hoogendoorn@lumc.nl

Abstract

Background: With the increasing use of health apps and ongoing concerns regarding their safety, effectiveness, and data privacy, numerous health app quality assessment frameworks have emerged. However, assessment initiatives experience difficulties scaling, and there is currently no comprehensive, consistent, internationally recognized assessment framework. Therefore, health apps often need to undergo several quality evaluations to enter different markets, leading to duplication of work. The CEN ISO/TS 82304-2 health app assessment seeks to address this issue, aiming to provide an internationally accepted quality evaluation through a network of assessment organizations located in different countries.

Objective: This study aimed to develop and evolve the draft CEN ISO/TS 82304-2 assessment handbook and developer guidance by testing them across organizations in several countries.

Methods: Assessment organizations from 5 countries were engaged to evaluate 24 health apps using the evolving CEN ISO/TS 82304-2 assessment across 3 evaluation rounds. The information submitted by a given health app developer was evaluated by 2 assessment organizations, and interrater reliability was examined. In addition, app developers and assessors were asked to report how much time they spent on information collation or evaluation and to rate the clarity of the developer guidance or assessor handbook, respectively. The collected data were used to iteratively improve the handbook and guidance between rounds.

Results: The interrater reliability between assessment organizations improved from round 1 to round 2 and stayed relatively stable between rounds 2 and 3, with 80% (55/69) of assessment questions demonstrating moderate or better (Gwet AC1>0.41) agreement in round 3. The median time required by developers to prepare the assessment information was 8 hours and 59 minutes (IQR 5.7-27.1 hours) in round 3, whereas assessors reported a median evaluation time of 8 hours and 46 minutes (IQR 7.1-11.0 hours). The draft guidance and handbook were generally perceived as clear, with a median round-3 clarity rating of 1.73 (IQR 1.64-1.90) for developers and 1.78 (IQR 1.71-1.89) for assessors (0=“very unclear”, 1=“somewhat unclear”, and 2=“completely clear”).

XSL•FO
RenderX

Conclusions: To our knowledge, this is the first study to examine the consistency of health app evaluations across organizations located in different countries. Given that the CEN ISO/TS 82304-2 guidance and handbook are still under development, the interrater reliability findings observed at this early stage are promising, and this study provided valuable information for further refinement of the assessment. This study marks an important first step toward establishing the CEN ISO/TS 82304-2 assessment as a consistent, cross-national health app evaluation. It is envisioned that the assessment will ultimately help avoid duplication of work, prevent inequities by facilitating access to smaller markets for developers, and build trust among users, thereby increasing the adoption of high-quality health apps.

(JMIR Form Res 2025;9:e64565) doi:10.2196/64565

KEYWORDS

health apps; mobile health; digital health; quality evaluation; assessment framework; interrater reliability

Introduction

Background

Health apps, which include both medical and wellness apps, have the potential to greatly benefit patients, the public, health care professionals, and the broader health care system [1,2]. They can increase the effectiveness of health care services by supporting behavior change and improving clinical outcomes, enhance efficiency by preventing hospitalization and freeing up resources, and allow for remote monitoring and treatment and, as a result, can have a large positive socioeconomic impact [1,2].

However, alongside these potential benefits, there are several risks associated with the use of low-quality health apps. Previous research has found that many health apps demonstrate data privacy or security risks [3-6], raise clinical safety concerns [7], provide incorrect medical information [8,9], or are not supported by efficacy or effectiveness evidence [4,5,10].

With >350,000 health apps on the market [11], it is difficult for health care professionals and the public to distinguish between beneficial and potentially harmful or ineffective apps. This issue is exacerbated by the fact that readily available metrics such as user ratings and download numbers are not correlated with health app quality [12]. Moreover, concerns regarding effectiveness, safety, usability, and data privacy can be a barrier to the widespread adoption of health apps [13-15], thereby preventing the realization of their potential benefits.

Health app assessments can help address this issue by providing transparent quality information to the public, health care professionals, payers, medical societies, and health care organizations. However, there is currently no comprehensive and consistent assessment that is internationally recognized. While the certification provided under the European Union (EU) Medical Device Regulation (MDR) is valid across EU member states, a large proportion of health apps are not independently evaluated by notified bodies under the MDR. This is because many apps are not classified as medical devices or fall into risk class I, for which app developers only need to declare conformity based on a self-assessment that is not externally validated [16,17]. Furthermore, there are important aspects, such as data privacy, that are not usually independently examined as part of the MDR certification, and research shows that many health apps on the European market do not meet relevant data protection requirements [18]. Finally, the only output provided by the MDR assessment is a CE (European conformity) mark. Consequently, even for health apps that have been examined under the MDR, stakeholders need to carry out additional evaluations to gather more detailed quality information that is required to inform decision-making.

Partly as a result of these limitations of the MDR and similar regulations in other regions, a range of organizations have developed their own assessment frameworks to guide health app–related decision-making. This includes health authorities, charitable and academic institutions, patient and medical professional associations, and commercial companies [19,20]. The resulting fragmented assessment landscape leads to duplicated work for both health systems and app developers if a given app needs to undergo several assessments to evidence its quality for different markets and varying types of organizations.

The CEN ISO/TS 82304-2 health app assessment, which was developed with funding from the European Commission, seeks to address this issue. It aims to provide a comprehensive and consistent quality evaluation that is recognized across Europe and beyond without duplicating the legal tasks of notified bodies under the MDR [21,22].

In this context, a notable challenge is scaling the assessment to potentially cover a substantial proportion of the >350,000 health apps on the market [11]. Notably, a recent review reported that only a median of 45 evaluated digital health technologies, comprising mostly health apps, were available in the assessment libraries of 24 digital health assessment organizations [20]. This limited availability of evaluated apps is likely due to several factors. First, some assessments that require developers to actively submit their apps for evaluation may not offer sufficient incentives to lead to a substantial number of app submissions.For example, incentives may be insufficient if the assessments are not associated with a large enough increase in app users, are not integrated into health care system decision-making (eg, for procurement or reimbursement), or provide reimbursement payment conditions or postprescription procedures that are unappealing (eg, causing delays in patient access to the app) [23-27]. Second, app developers may not be able (due to limited funds and staff in start-ups and small companies) or willing (in light of lacking incentives) to invest the time and money required to meet the quality requirements of individual schemes, such as those requiring new clinical studies, which may affect the number of apps submitted to and passing particular assessments [26,28]. Third, difficulties in scaling the assessment

XSL•FO
RenderX

process may play a role. A review focusing specifically on national health app evaluation initiatives found that even the front-runners in this area, such as Germany, Belgium, and the United Kingdom, are struggling with efficient assessment implementation. This may partly be due to the use of centralized approaches to app evaluation, which can create bottlenecks [29]. It has been proposed that international collaboration and the use of distributed evaluation bodies is needed to address this issue [29]. Following this approach, the CEN ISO/TS 82304-2 assessment is planned to be implemented across a network of assessment organizations in Europe and beyond to achieve efficient scaling while ensuring consistent assessment results across countries (and providing sufficient incentives, as explored in our recent discrete choice experiment [25]).

Objectives

This study aimed to develop and evolve the CEN ISO/TS 82304-2 assessment handbook and developer guidance by testing the assessment across organizations located in different countries. For this purpose, assessment organizations from 5 countries were engaged to evaluate 24 diverse health apps using information submitted by the app developers. In total, 2 organizations assessed each app, and their interrater reliability was examined. In addition, app developers and assessors were asked to report how much time they spent collating or evaluating the assessment information, respectively, and to rate the clarity of the assessment instructions. The collected interrater reliability, time, and clarity rating data were used to iteratively improve the CEN ISO/TS 82304-2 assessment handbook and guidance across 3 evaluation rounds.

Methods

CEN ISO/TS 82304-2 Assessment Framework

The CEN ISO/TS 82304-2 assessment framework is intended to provide an independent, evidence-based, international evaluation of health app quality for a range of health care stakeholders. The framework was created using a Delphi methodology with input from 83 experts from 8 stakeholder groups, including medical organizations, health care authorities, and health app developers across 6 continents [21].

It should be noted that, while the CEN ISO/TS 82304-2 quality requirements are applicable worldwide, some requirements are influenced by local regulations (which is the case for other digital health assessment frameworks as well [29]). This necessitates minor amendments to the assessment across jurisdictions. Given that the project as part of which this study was conducted was funded by the European Commission, the initial CEN ISO/TS 82304-2 testing and implementation efforts were focused on Europe. Therefore, steps were taken to ensure that the CEN ISO/TS 82304-2 criteria are in line with European regulations and standards (as described further in the Discussion section). Future work will aim to support the alignment of the assessment with regulations in other regions.

The CEN ISO/TS 82304-2 framework consists of 5 subsections containing between 7 and 28 questions each: “Product information,” “Healthy and safe,” “Easy to use,” “Secure data,” and “Robust build.” In total, the assessment comprises 81 questions, 12 of which serve to collect background data about the app, mostly to enable the display of this information in a related health app quality label. As the latter questions are answered in free-text or multiple-selection format and do not affect the quality rating of the app, they were not considered in the interrater reliability analysis reported in this paper. The remaining 69 questions (quality requirements) have either 2 (“yes” or “no”) or 3 (“yes,” “no,” or “not applicable”) response options. A list of all assessment questions can be found in the supplement of the CEN ISO/TS 82304-2 framework development paper [21].

Final responses to the assessment questions are determined by an assessment organization based on documentation and other information that is provided by health app developers as part of the assessment process. Examples of the documentation that developers are asked to submit alongside providing access to their apps include evidence of the societal and health benefits of the app, an analysis of health risks, standard operating procedures (or relevant excerpts thereof), and privacy policies.

To facilitate the assessment process, detailed instructions were developed during this study to indicate what documentation is acceptable and what subrequirements need to be met for a given assessment question to be answered with “yes.” These instructions were outlined in a draft developer guidance document and a draft app assessor handbook and were tested and evolved across evaluation rounds (see the App Assessment Procedure section for details). As the handbook is still under development, it is not presented in this paper; however, it will likely be published once the first operational version is finalized.

The final output of the CEN ISO/TS 82304-2 assessment is a quality label [30] that summarizes the assessment results for each quality domain and for the app overall using letter grading (E to A) and color coding (red, orange, yellow, light green, and dark green), similar to the EU energy label. This allows health care professionals and the public to quickly determine whether a given app meets their needs and preferences. For more detailed information, a full assessment report is also available [22].

It should be noted that, as part of this study, no quality labels were issued given that the CEN ISO/TS 82304-2 assessment guidance and handbook were under development.

App Selection

As mentioned previously, conducting the CEN ISO/TS 82304-2 assessment requires health app developers to submit documentation that is not publicly available. Therefore, developers were actively involved in this study, as is described further in the App Assessment Procedure section. This also allowed for the collection of feedback on the assessment process from the app developer perspective.

To thoroughly test the evolving assessor handbook, developer guidance, and assessment process, a variety of health apps were sought to be included in this study, mimicking real-life app diversity. For this purpose, the assessors involved in this study compiled a list of health app development companies to contact, aiming to cover a wide range of company sizes, countries of origin, condition areas, and app functionalities. In addition, this study was advertised on a dedicated project website where

XSL•FO
RenderX

developers had the opportunity to get in touch to express their interest in participating.

To ensure that a given app could be evaluated by any assessor (see the App Assessment Procedure section for details), only apps available in English were eligible for inclusion in the study. However, in 2 cases, it was discovered, after the developers had collated all the assessment information, that their apps did not meet this criterion. Given that considerable effort had already been exerted at this stage, these apps were included, allocating them to assessors with relevant language knowledge or using automated translation tools where needed. No other eligibility criteria were applied apart from a willingness of the app developers to provide the necessary assessment information as well as access to their apps (or, as an exception, relevant screenshots or video demonstrations if access could not be granted, eg, due to geographical restrictions).

The assessment took place across 3 rounds, and the initial aim was to evaluate 8 health apps per round, with funding being available for the evaluation of 24 apps in total. However, due to developers' competing commitments, the amount of effort involved, and the relatively short project timelines, a number of app development companies dropped out of the study after initially agreeing to take part. In round 1, timelines did not permit the recruitment of replacements. In subsequent rounds, more companies than needed were initially recruited to account for potential dropouts. The final sample consisted of the intended total of 24 health apps, which were distributed across the assessment rounds as follows: 3 (12%) apps in round 1 (3 dropouts), 7 (29%) apps in round 2 (7 dropouts and 1 moved to round 3), and 14 (58%) apps in round 3 (10 dropouts).

As mentioned previously, no quality labels were provided for the evaluated apps as part of this study given that the aim was to develop, test, and evolve the CEN ISO/TS 82304-2 assessment guidance and handbook to arrive at an operational version that is suitable for use in quality certification, with sufficiently consistent assessment results. Therefore, no letter grades were calculated or shared. The main benefits offered to participating health app development companies were the receipt of feedback for the individual assessment questions (based on the respective draft versions of the handbook), which could be used internally for product improvements; a chance to familiarize themselves with the CEN ISO/TS 82304-2 assessment; and the offer of a reduced fee if they decided to submit their apps to the operational version of the CEN ISO/TS 82304-2 certification once the latter is rolled out.

App Assessment Procedure

The app evaluations were conducted between November 2022 and April 2024 by assessment organizations with diverse backgrounds from different countries to approximate the full scope of real-life assessment conditions. Specifically, participating organizations were located in the United Kingdom, Italy, Spain, Canada, and France and consisted of 3 private assessment organizations and 2 health authorities, the latter of which each used a subcontractor. In the final round, 1 assessment organization had to drop out as the assessors involved in the study had left the company. Other than that, the same organizations were involved in all 3 evaluation rounds.

Across all organizations, a total of 12 individual assessors were involved in the study, for 10 (83%) of whom background information was available (the remaining n=2, 17%, were based at the organization that had to drop out in the last round). The median previous experience that assessors had with evaluating the quality of health apps using any assessment framework was 19.5 (IQR 0.0-48.0) months. With regard to previous exposure to the CEN ISO/TS 82304-2 assessment in particular, 40% (4/10) of the assessors reported having previously used the CEN ISO/TS 82304-2 assessment to evaluate health apps, and 20% (2/10) of the assessors had been involved in the development of the framework. The remaining 40% (4/10) of the assessors had no previous familiarity with the CEN ISO/TS 82304-2 assessment.

The app assessment procedure across all 3 evaluation rounds was as follows. At the start of each round, the app developers who had agreed to participate in the study were supplied with the draft CEN ISO/TS 82304-2 assessment guidance for round 1, 2, or 3, as applicable. They were asked to collate and submit the required information, including relevant documentation, and grant the assessors access to their health app. In addition, developers were asked to report, for each question, how long it took them to gather the relevant information and how clear they found the associated assessment guidance (red="very unclear," yellow="somewhat unclear," and green="completely clear," coded as 0, 1, and 2, respectively, for the data analysis). Moreover, additional feedback was collected from app developers via individual interviews that were conducted via video calls after the submission of the assessment information but before the receipt of the evaluation results (see the Feedback From Developers section for details).

The assessment information submitted by a given app developer was passed on to 2 of the 5 assessment organizations, which were pseudorandomly allocated in such a way as to achieve diverse organization pairings across apps. A range of expertise was needed to complete the evaluations as the CEN ISO/TS 82304-2 assessment covers clinical, usability, privacy, and technical domains in depth. Therefore, some assessment organizations engaged different assessors to evaluate each of the assessment subsections, whereas others appointed a lead assessor who consulted subject matter experts when needed. The median number of apps evaluated by individual assessors across all 3 rounds with an evolving handbook was 9.5 (IQR 7.0-11.0).

As part of the app evaluation, assessors examined the submitted evidence, which included documentation and screenshots as well as relevant screens within the app itself based on the CEN ISO/TS 82304-2 draft handbook for round 1, 2, or 3 and recorded their responses ("yes," "no," or "not applicable") for each question. Evaluations were conducted independently (ie, different assessment organizations did not have access to each other's responses). Assessors also provided evaluation times and handbook clarity ratings for each of the 81 assessment questions.

After the app evaluations within a given round, meetings were held among the assessors to discuss any disagreements. For each app, a consensus decision on a final response for each

XSL•FO
RenderX

assessment question was reached, which was passed on to the respective app developers. The information gathered during these discussions, together with the interrater reliability results, time and clarity ratings, and the feedback collected from developers, was used to amend and evolve the assessor handbook and developer guidance between rounds (and after the final round). Due to these iterative changes, the results are reported separately for each round.

Data Analysis

Interrater reliability was determined for the agreement between the 2 assessment organizations based on their independent responses recorded before any discussion. For this purpose, Gwet AC1 values [31,32] were calculated for each assessment question in MATLAB (version 9.10.0.1710957 [R2021a]; The MathWorks, Inc). The Gwet AC1 was used as it has been shown to provide stable interrater reliability coefficients without being substantially affected by response prevalence or marginal probability (as is, eg, the case for the Cohen κ). In addition, it is suitable for designs that are not fully crossed over, such as the one used in this study, in which *different pairs* of raters evaluated the varying apps [33].

Gwet AC1 values can be interpreted as follows using Altman benchmark ranges [33]: <0.20="poor agreement," 0.21-0.40= "fair agreement," 0.41-0.60="moderate agreement," and >0.60="good agreement." The percentage of questions falling into each of these categories within each subsection of the CEN ISO/TS 82304-2, as well as across the entire assessment, was reported.

There were some instances in which assessors did not provide evaluation responses for individual questions, possibly because they were uncertain about which option to select (where coinciding clarity ratings were low). These instances were omitted from the analysis, as were a small number of cases in which an assessor responded with “not applicable” to a particular question even though this was not a valid response option. In addition, for 1 app in round 3, the evaluation responses from 1 assessment organization were not available due to technical issues. Therefore, this app was excluded from the interrater reliability analysis. Moreover, due to a miscommunication, 1 app in round 1 was evaluated by 3 assessment organizations instead of 2. As the assessors of one of these organizations were able to see the evaluation answers of another organization, the former’s responses were excluded from the interrater reliability (but not the time or clarity rating) analysis as a precaution to maintain full independence of responses. In subsequent rounds, additional measures were taken to ensure that access to other assessors’ responses was not possible.

For the time and clarity ratings collected from app developers and assessors, summary values were computed in MATLAB. Specifically, for each rating set (from a given rater for a particular app), the sum of the time ratings was calculated across all *questions* within a given section and across the entire assessment. The medians and IQRs of these values were reported as an indication of the central tendency and dispersion of the data *across rating sets*. A similar calculation was applied to the clarity ratings except that, for the latter, the average rather than the sum was computed across the questions within a given section before examining medians and IQRs across rating sets.

It should be noted that some assessors and developers provided time or clarity ratings for <75% of the questions within a particular section or across the entire assessment. These cases were excluded from the relevant section summary or the total assessment summary, respectively, as not enough data were available to arrive at a summary value that was likely to represent the rater’s time investment or perceived clarity (sample sizes are reported in the Results section). Where ratings were missing for <75% of questions, median imputation was applied using the available ratings for a given *question* before calculating the sum or average across all questions for a particular rating set. When calculating the summary values across the entire assessment (for cases in which <75% of all ratings were missing), median imputation was applied to *all* missing values, including to those within the sections that were excluded from the section summary (based on the aforementioned criterion). This was necessary for the calculation of the total amount of time spent on the assessment, for which time ratings were *summed* across *all* questions. The same methodology was applied to the clarity ratings for consistency.

Ethical Considerations

Approval by an ethics or scientific research committee was deemed unnecessary according to Dutch national regulations, since the study was not subject to the Medical Research Involving Human Subjects Act, and according to the guidelines of the Central Committee on Research Involving Human Subjects. The study did not require ethics approval as no personal or sensitive data were collected and all activities performed by assessors and developers lay strictly within their professional competence and job responsibilities. Moreover, no identifiable information is disclosed about individual health apps in this paper.

Results

App Characteristics

Of the 24 apps included across the 3 assessment rounds, 8 (33%) were developed by micro-sized start-ups with <10 employees, 12 (50%) were developed by small businesses with 10 to 49 employees, 2 (8%) were developed by medium-sized companies with 50 to 249 employees, and 2 (8%) were developed by large organizations with ≥250 employees. Most companies (22/24, 92%) were based in Europe, that is, in the Netherlands (6/22, 27%), Belgium (4/22, 18%), and France (3/22, 14%), as well as in Bulgaria, Denmark, Finland, Greece, Italy, Portugal, Spain, Switzerland, and the United Kingdom (1/22, 5% each). The remaining companies (2/24, 8%) had headquarters in the United States.

The sample included both disease-specific and more general health or health system service support apps, with the former targeting a wide range of condition areas such as diabetes, asthma, cardiovascular disease, mental health, chronic pain, nicotine dependence, cognitive impairment, hepatitis C, and colorectal conditions. Most apps had more than one intended use, including health system service support (3/24, 12%),

XSL•FO
RenderX

information provision (12/24, 50%), disease prevention (eg, through behavior change; 7/24, 29%), self-management (16/24, 67%), monitoring (15/24, 62%), treatment (4/24, 17%), communication with health care professionals (14/24, 58%), calculation of values that are likely to affect health care decisions (6/24, 25%), and research support (3/24, 12%). Overall, 46% (11/24) of the apps were classified as medical devices under the MDR.

Interrater Reliability Between Assessment Organizations

For round 1, a median Gwet AC1 value of 0.54 (IQR 0.14-1) was observed across all 69 CEN ISO/TS 82304-2 assessment questions included in the analysis, indicating moderate overall agreement between assessment organizations. Breaking this finding down further revealed that the interrater reliability was poor (<0.20) for 32% (22/69) of the questions, fair (0.21-0.40) for 4% (3/69) of the questions, moderate (0.41-0.60) for 29% (20/69) of the questions, and good (>0.60) for 35% (24/69) of the questions.

In round 2, the overall interrater reliability between assessment organizations was good, with a median Gwet AC1 value of 0.63 (IQR 0.48-0.84). Agreement was poor (<0.20) for 7% (5/69) of the questions, fair (0.21-0.40) for 4% (3/69) of the questions, moderate (0.41-0.60) for 26% (18/69) of the questions, and good (>0.60) for 62% (43/69) of the questions.

In round 3, a median Gwet AC1 value of 0.67 (IQR 0.42-0.85) indicated good overall agreement between organizations. Interrater reliability was poor (<0.20) for 6% (4/69) of the questions, fair (0.21-0.40) for 14% (10/69) of the questions, moderate (0.41-0.60) for 20% (14/69) of the questions, and good (>0.60) for 59% (41/69) of the questions.

A visual representation of the interrater reliability findings by CEN ISO/TS 82304-2 subsection can be found in Figure 1.

Figure 1. Percentage of questions falling into each interrater reliability category for the agreement between assessment organizations (Gwet AC1≤0.20="poor agreement," 0.21-0.40="fair agreement," 0.41-0.60="moderate agreement," and >0.6="good agreement"). Results are shown for the following assessment sections (A) "Healthy and safe" (23 questions), (B) "Easy to use" (13 questions), (C) "Secure data" (21 questions), and (D) "Robust build" (12 questions), as well as for (E) the Entire assessment (69 questions). Note that the "Product information" section only contained free-text and multiple-response options and, therefore, was not included in the interrater reliability analysis.

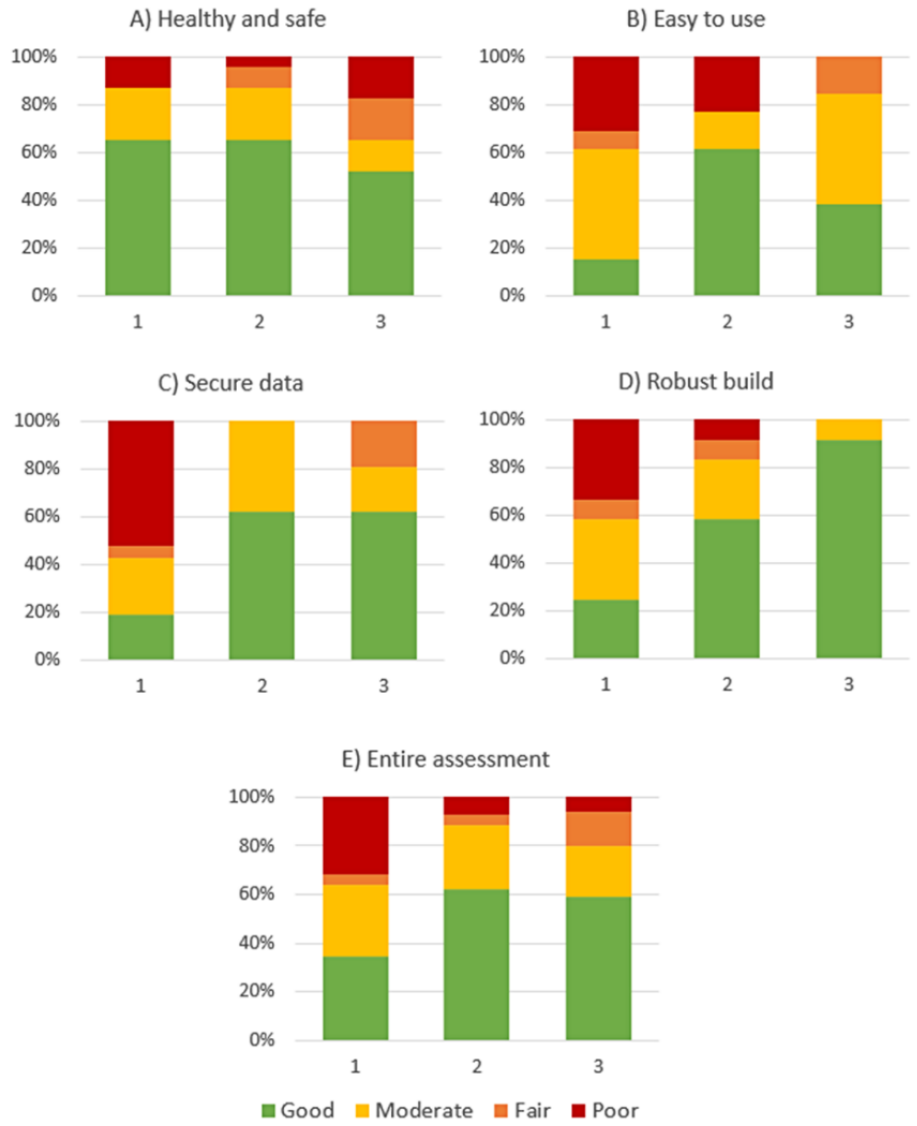

XSL•FO
RenderX

Time and Clarity Ratings

Assessor Ratings

The median evaluation time across all CEN ISO/TS 82304-2 questions reported by the assessors was 10 hours and 34 minutes (IQR 9.5-13.1 hours) in round 1 (n=5 rating sets), 8 hours and 33 minutes (IQR 4.9-9.0 hours) in round 2 (n=11 rating sets), and 8 hours and 46 minutes (IQR 7.1-11.0 hours) in round 3 (n=19 rating sets). The amount of time spent on each CEN ISO/TS 82304-2 subsection is presented in Table 1. It should be noted that these ratings include only the time it took to arrive at a response to the assessment questions based on the submitted evidence but not the time dedicated to other activities, such as meetings.

The median assessor handbook clarity rating across all questions was 1.46 (IQR 1.44-1.65) in round 1 (n=6 rating sets), 1.72 (IQR 1.53-1.88) in round 2 (n=10 rating sets), and 1.78 (IQR 1.71-1.89) in round 3 (n=12 rating sets), where 0=“very unclear”, 1=“somewhat unclear”, and 2=“completely clear”. Clarity ratings broken down by subsection can be found in Table 2.

Table 1. Evaluation time reported by the assessors across rounds for each subsection of the CEN ISO/TS 82304-2 assessment.

Section and round	Number of rating sets	Evaluation time (hours), median (IQR)
Product information		
1	5	0.6 (0.5-0.7)
2	11	0.5 (0.2-0.5)
3	19	0.4 (0.3-0.7)
Healthy and safe		
1	5	3.6 (3.2-4.3)
2	11	3.2 (1.7-3.4)
3	19	3.1 (2.1-3.7)
Easy to use		
1	5	1.9 (1.7-3.1)
2	11	1.5 (0.8-1.8)
3	19	1.8 (1.3-2.3)
Secure data		
1	5	2.9 (2.6-3.4)
2	11	2.0 (1.2-2.3)
3	19	2.6 (1.0-3.3)
Robust build		
1	4	1.5 (1.2-1.9)
2	10	1.1 (0.7-1.4)
3	18	1.1 (0.1-1.7)

XSL•FO
RenderX

Table 2. Handbook clarity ratings reported by the assessors across evaluation rounds for each subsection of the CEN ISO/TS 82304-2 assessment.

Section and round	Number of rating sets	Assessor clarity rating (0=very unclear, 1=somewhat unclear, and 2=completely clear), median (IQR)
Product information		
1	6	2 (2-2)
2	9	2 (2-2)
3	14	2 (2-2)
Healthy and safe		
1	6	1.54 (1.5-1.57)
2	10	1.84 (1.29-1.96)
3	12	1.88 (1.7-1.97)
Easy to use		
1	6	1.54 (1.31-1.54)
2	10	1.62 (0.62-1.77)
3	9	1.54 (1.17-1.94)
Secure data		
1	6	1.5 (1.24-1.67)
2	10	1.9 (1.62-2)
3	8	1.74 (1.52-2)
Robust build		
1	6	1.21 (1.08-1.83)
2	10	2 (1.67-2)
3	13	1.92 (1.67-2)

App Developer Ratings

The median time spent by app developers on preparing the assessment information using the evolving guidance was 2 hours in round 1 (n=1 rating set), 7 hours and 3 minutes (IQR 3.8-8.8 hours) in round 2 (n=5 rating sets), and 8 hours and 59 minutes (IQR 5.7-27.1 hours) in round 3 (n=10 rating sets). The information collation time for each CEN ISO/TS 82304-2 subsection can be found in Table 3.

The median guidance clarity rating indicated by app developers across all questions was 1.84 (IQR 1.69-1.99) in round 1 (n=2 rating sets), 1.80 (IQR 1.74-1.87) in round 2 (n=5 rating sets), and 1.73 (IQR 1.64-1.90) in round 3 (n=12 rating sets). Clarity ratings broken down by subsection are presented in Table 4.

XSL•FO
RenderX

Table 3. Information collation time reported by app developers across rounds for each subsection of the CEN ISO/TS 82304-2 assessment.

Section and round	Number of rating sets	Information collation time (hours), median (IQR)
Product information		
1	1	0.06 (—[a])
2	5	0.22 (0.16-0.31)
3	10	0.48 (0.27-0.82)
Healthy and safe		
1	1	0.5 (—)
2	5	2.14 (1.32-3.4)
3	10	4.11 (2.63-7.65)
Easy to use		
1	1	0.26 (—)
2	5	1.53 (0.8-2.22)
3	8	1.8 (0.96-3.75)
Secure data		
1	1	0.92 (—)
2	5	1.32 (1.1-2.13)
3	10	1.56 (0.99-8.5)
Robust build		
1	1	0.28 (—)
2	5	0.77 (0.38-1.46)
3	9	0.9 (0.63-3.45)

[a]Not applicable.

XSL•FO
RenderX

Table 4. Guidance clarity ratings reported by app developers across evaluation rounds for each subsection of the CEN ISO/TS 82304-2 assessment.

Section and round	Number of rating sets	Developer clarity ratings (0=very unclear, 1=somewhat unclear, and 2=completely clear), median (IQR)
Product information		
1	3	2 (1.36-2)
2	5	2 (1.93-2)
3	13	2 (1.96-2)
Healthy and safe		
1	2	1.82 (1.64-2)
2	5	1.93 (1.56-1.97)
3	12	1.87 (1.65-1.92)
Easy to use		
1	2	1.81 (1.62-2)
2	5	1.92 (1.65-2)
3	12	1.71 (1.65-1.88)
Secure data		
1	2	1.98 (1.95-2)
2	5	1.95 (1.6-1.96)
3	12	1.81 (1.67-1.95)
Robust build		
1	2	1.58 (1.25-1.92)
2	5	1.83 (1.71-1.94)
3	12	1.67 (1.33-1.88)

Feedback From App Developers

All 24 app developers took part in individual follow-up interviews, which were conducted vial video calls after the developers had submitted their assessment information but before they received the evaluation responses from the assessors. The interviews were led by MK and took between 25 and 75 minutes. On the developer side, between 1 and 5 team members attended the interviews, which in all cases included individuals who had been involved in collating the assessment information and in 92% (22/24) of the cases included the person leading this process.

When asked what improvements they would recommend for the draft version of the assessment guidance and the evolving evaluation process, developers provided several suggestions. First, in rounds 1 and 2, it was noted that some questions were mainly applicable to medical apps, causing uncertainties about whether and, if so, how they could be answered for wellness apps. It was suggested to include stratification points in the assessment to determine which questions need to be answered for which type of app (which is part of the CEN ISO/TS 82304-2 framework but had not yet been applied in the draft guidance). Second, uncertainties were reported regarding what evidence was required for some of the questions. It was indicated that the provision of example answers and document templates, as well as the opportunity to interact with assessors as needed (which was always intended for the real-world assessment implementation), would be helpful to clarify requirements. Third, it was noted, especially by companies that outsourced their software development, that it would be useful to receive more explicit information on which expertise (eg, clinical, data privacy, cybersecurity, and software development) was needed to answer each individual question to allow for a more efficient distribution of the questions across the team and subcontractors. Fourth, developers asked for guidance on how to efficiently integrate CEN ISO/TS 82304-2 into their internal processes and strategy alongside other required assessments. In this context, it was noted that there is overlap between MDR and CEN ISO/TS 82304-2 requirements (which was deliberate and necessary to cover relevant quality criteria for apps that are not classified as medical devices or, given their class-I classification, not assessed by notified bodies). It was suggested to allow for the reuse of documentation or the automatic passing of certain CEN ISO/TS 82304-2 questions if an app had already been assessed by a notified body (which was indeed always intended but requires further consideration for effective implementation).

Developers were further asked which, if any, benefits they thought the CEN ISO/TS 82304-2 assessment and the associated quality label could provide for their company. In response to this question, 67% (16/24) of the developers indicated that they expected the quality label to be an asset for marketing purposes, in addition to creating confidence and trust among users, payers, and health care providers. A further 25% (6/24) of the developers primarily mentioned receiving an independent quality

XSL•FO
RenderX

certification as a benefit in and of itself. Moreover, 21% (5/24) of the developers saw the assessment process as a valuable learning experience for the company, offering guidance on how to measure and improve the quality of their app, with 1 company reporting that they created an internal quality management system as a result of going through the assessment process.

Finally, when asked whether, knowing what they did now about the time and effort that was required, they would still have chosen to submit their app for CEN ISO/TS 82304-2 assessment, 88% (21/24) of the developers replied in the affirmative, with 8% (2/24) being unsure and 4% (1/24) indicating that they would not have submitted their app.

Discussion

Principal Findings

This study aimed to develop and evolve the CEN ISO/TS 82304-2 assessment handbook and developer guidance by testing them across a diverse range of apps and assessment organizations from different countries. Time spent on information collation or evaluation, instruction clarity, and assessment consistency were examined to inform the next steps toward establishing a robust certification process. To our knowledge, this is the first cross-national, cross-stakeholder collaboration in the field of health app assessment of this extent, diversity, and rigor.

This study revealed that the interrater reliability between assessment organizations improved from round 1 to round 2 for all assessment sections, likely reflecting both an initial practice effect and clarifications made to the assessor handbook. Between rounds 2 and 3, interrater agreement improved or stayed relatively stable for the “Robust build” and “Secure data” sections, respectively. However, unexpectedly, the “Healthy and safe” and “Easy to use” sections demonstrated somewhat lower interrater reliability in round 3 than in round 2 (which likely also drove the very slight decrease in interrater reliability observed across the entire assessment between these rounds). The evolving rigor of the handbook between rounds likely contributed to this finding, which may have affected some sections more than others depending on the changes that were made. Moreover, differences in app characteristics may have also played a role. For instance, in round 2, only 29% (2/7) of the evaluated apps were classified as medical devices, whereas in round 3, this classification applied to 43% (6/14) of the apps. The greater complexity of medical device app functionalities may have made the latter more difficult to evaluate in some (but not all) domains compared to apps that are not medical devices. In the future, to the extent possible, the assessment outcomes of notified bodies will be used to inform those CEN ISO/TS 82304-2 quality requirements that overlap with the MDR for medical device apps, which will increase the consistency of the latter’s assessment. In addition, the improvements discussed in the Learnings and Planned Improvements to the Assessment section aim to ensure consistent evaluation across all assessment sections and app types. Still, even with the round-3 draft version of the assessment handbook, 59% (41/69) of the questions across the entire assessment demonstrated good interrater reliability, with another 20% (14/69) of the questions showing moderate agreement. These results are promising, indicating that, even at this early stage, the CEN ISO/TS 82304-2 assessment already results in largely consistent outcomes across organizations located in different countries. The implications of this finding are discussed further in the Envisioned Impact of the CEN ISO/TS 82304-2 Assessment section.

The median handbook clarity ratings reported by assessors increased steadily across rounds, as was expected due to increasing familiarity with the handbook and the improvements made to the latter between rounds based on the collected data. Somewhat surprisingly, the median guidance clarity ratings collected from app developers decreased across rounds. This may have been an unintended consequence of the iterative changes (increasing rigor) that were made to the guidance. For instance, in round 3, the subrequirements that were used in the assessor handbook were added to the developer guidance with the aim of clarifying evidence needs and assessor expectations. However, this may inadvertently have caused uncertainty among developers about how to answer the associated assessment questions, with seemingly only a minority of developers having taken the individual subrequirements into account. Alternatively, especially given the small sample size, the observed differences across rounds may have been a result of variations in the included apps and development companies, which were likely to affect how easily certain assessment questions could be answered. In either case, clarification of the written guidance, giving developers the opportunity to contact the assessment organizations with clarification questions, and other improvements discussed in the Learnings and Planned Improvements to the Assessment section will likely help enhance developers’ understanding of the guidance in the future.

With regard to the time spent on the assessment process, in round 3, developers reported taking a median of 8 hours and 59 minutes (IQR 5.7-27.1 hours) to gather the assessment information, and assessors indicated spending a median of 8 hours and 46 minutes (IQR 7.1-1.0 hours) evaluating the apps. While this is a substantial time investment, it reflects the rigor of the CEN ISO/TS 82304-2 assessment, which is intended to ensure that different stakeholder groups internationally are provided with sufficient information to support their health app–related decision-making, significantly reducing the need for further evidence gathering and avoiding similar evaluations. Importantly, it should be noted that, despite the effort involved, 88% (21/24) of the app developers reported that they would submit their app for CEN ISO/TS 82304-2 assessment again, indicating that the perceived benefits of the assessment outweigh the time cost. In addition, several steps are being considered to reduce the required time for both developers and assessors in the future implementation of the assessment, as discussed in the Learnings and Planned Improvements to the Assessment section.

All in all, this study provided promising results regarding the interrater reliability between assessors at this early stage of development and testing, as well as valuable insights to support the future implementation of the CEN ISO/TS 82304-2 assessment. In the following sections, it is discussed how these results compare to previous findings; which improvements were made based on this study; what next steps are planned for the

XSL•FO
RenderX

refinement of the CEN ISO/TS 82304-2 assessment handbook, developer guidance, and process; and what benefits the widespread use of the assessment is expected to provide. Strengths and limitations of the study are also considered.

Comparison to Previous Health App Assessment Interrater Reliability Studies

Previous studies that have examined the interrater reliability of health app assessments have yielded mixed results. For several assessments, <15% of the questions demonstrated good agreement between assessors (eg, Fleiss κ>0.60 or Krippendorff α>0.667) [34-36]. Our results compare favorably to these findings, with approximately 60% of CEN ISO/TS 82304-2 questions in rounds 2 and 3 (43/69, 62%, and 41/69, 59%, respectively) showing good assessor agreement (Gwet AC1>0.60), already at this early stage of development. However, it should be noted that higher interrater reliability has been reported for several health app assessments, such as outstanding agreement (κ>0.80) for 86% of the Enlight questions [37], substantial or good agreement (κ>0.60) for 79% of the questions within the American Psychiatric Association's app evaluation framework [38], and satisfactory intraclass correlations (>0.75) for all sections of the Mobile App Rating Scale [39].

In this context, the complexity of the assessments should be considered. The aforementioned frameworks are less comprehensive in scope, rely only on publicly available information (mostly the app itself), and include less complex and often more descriptive questions than those in the CEN ISO/TS 82304-2 assessment. Therefore, it is unsurprising that good interrater reliability could be achieved relatively easily for these frameworks. In contrast, the in-depth evaluation of clinical and technical aspects provided as part of the CEN ISO/TS 82304-2 assessment requires expert assessment knowledge and careful examination of elaborate documentation submitted by developers. Consequently, it is to be expected that repeated testing and iterative improvement are required to achieve a clear handbook and good assessor agreement for all CEN ISO/TS 82304-2 questions. In line with this, a recent qualitative study found that, even for the MDR, there are significant variations in the interpretation of the included quality requirements both within and between notified bodies, which can result in inconsistent assessment outcomes [40]. This highlights the difficulty of achieving consistency for more complex digital health technology quality assessments. In this context, it is worth noting that, unlike some other quality or efficiency assessments (eg, those carried out for the EU energy label), digital health technology assessments involve evaluations carried out by assessors rather than relying only on automated measurements (eg, of energy consumption). This introduces variability, which needs to be minimized to ensure that national authorities, citizens, and health care professionals can trust that health apps are evaluated consistently. The refinements discussed in the Learnings and Planned Improvements to the Assessment section aim to further improve the CEN ISO/TS 82304-2 assessment consistency. This includes amendment of the handbook and guidance as well as considerations related to the expertise, training, and support from subject matter experts and automated tooling that can assist assessors in achieving consistent ratings.

With regard to the time required for app evaluation, the few previous interrater reliability studies that have examined this measure have reported assessment times ranging from 15 minutes to 1 hour [34,41]. These values are considerably lower than those observed in this study, with a median evaluation time of 8 hours and 46 minutes (IQR 7.1-11.0 hours) reported in round 3. Again, this difference is likely due to the reliance only on publicly available information and the lower complexity of the former assessments compared to that of the CEN ISO/TS 82304-2. Pilot studies more similar to this one, as part of which information was requested from health app developers, have reported app evaluation times more comparable to those in our study. Specifically, one study observed assessment times of 10 to 30 work hours per app (including writing the recommendation, meetings, and other communications [42]), and another study (including 2 apps) reported indicative app evaluation times of up to 6 hours (if uncertainties were encountered [43]). The time information gathered during this study will be used to inform developers of how long it is expected to take to collate the assessment information (developer workload), determine assessment pricing, and guide scalability efforts (assessor workload).

Learnings and Planned Improvements to the Assessment

On the basis of the information collected during this study and complementary studies conducted as part of the initiative that this study is a part of, improvements were made to further increase the consistency, clarity, efficiency, and value of the assessment. Several of these learnings may also be applicable to other health app evaluation initiatives.

First, to increase consistency, the handbook was further evolved to include separate sections with standardized, simplified wording. These sections provide additional guidance as to (1) the documentation needed from app developers; (2) which subrequirements the assessors need to evaluate with the provided documentation and, if applicable, with information from the app itself; and (3) which subrequirements need to be met for the assessment question to be answered with a "yes." In addition, it is being explored whether automation can be introduced to evaluate or support the evaluation of some requirements, such as testing accessibility features, examining data flow, and scanning privacy policies for relevant content. In addition to improving consistency, automation can help reduce the required assessment time and may enable the evaluation of additional quality aspects that cannot be examined manually. Moreover, in the future, the outcomes of trusted existing assessments, in particular those provided by notified bodies as part of their legal task, will, to the extent possible, be used to address the related CEN ISO/TS 82304-2 requirements, which is expected to improve consistency as well as avoid duplication of effort. Finally, consideration is being given to how best to train new assessors (eg, using practice evaluations on already assessed apps) to ensure that they demonstrate a high level of consistency before starting to conduct "real" app assessments.

Second, to increase clarity and efficiency, a fit-for-purpose platform, including subsection headings, stratification, and documentation templates, is intended to be developed to collect

XSL•FO
RenderX

assessment information from developers. Subsection headings will make the expertise needed to answer different questions (eg, medical, data privacy, or cybersecurity knowledge) more apparent. This can help developers distribute the assessment questions more effectively across the company and potential suppliers, asking relevant experts to answer the questions most suited to their expertise, thereby saving time in collating the assessment information. Stratification, carried out automatically via the platform, will more efficiently determine which assessment questions need to be answered based on the intended use and characteristics of the app, such as whether it provides health information, has already been assessed by a notified body, processes personal data, or incorporates application programming interfaces. Document templates, which interviewed developers suggested would be helpful, will further clarify the evidence requirements and allow developers to identify more quickly and effectively which information they need to provide. This, in turn, will ensure that assessors receive all and only the information needed for the evaluation in a standard format, speeding up the evaluation process and making it more robust.

Efficiency is also expected to be increased by giving developers and assessors the opportunity to interact as part of the assessment process (as was always intended). This was not the case during the study to enable the examination of developers' understanding of the guidance in the absence of outside advice and to ensure that assessors' evaluation responses were independent and based on the same information. Due to the lack of interaction during this study, developers were occasionally uncertain about how to respond to particular questions or what documentation to provide, and in some instances, assessors spent considerable amounts of time searching for relevant information, including within the apps. In the real-world implementation of the assessment, developers will have the chance to ask clarification questions, and assessors will be able to contact developers to request additional information if needed. It is expected that such interactions will reduce uncertainty and expedite the evaluation process. Future consideration will additionally be given to how to efficiently handle reassessments where changes to an app have been made that affect the responses to individual quality requirement questions and how to guide developers in efficiently integrating CEN ISO/TS 82304-2 into their internal processes and strategy.

Third, to increase the value and uptake potential of the CEN ISO/TS 82304-2 assessment, steps were taken to evolve and align it more specifically with the European context, where the initial implementation efforts are intended to take place. This included incorporation of findings of a comparative analysis with (mostly) European health technology assessment frameworks, EU-level legislation and values, international and European standards, advice from subject matter experts, literature on health app quality, and information needs reported by (mostly) European stakeholders for health app–related decision-making. Consultation of related European authorities such as the Medical Device Coordination Group, the European Data Protection Supervisor, and the EU Agency for Cybersecurity is proposed for the future to obtain feedback and ensure sustained alignment. In addition, future work will aim to align the assessment with regulations and standards in other regions as well. Notably, it has been observed that many regulations are promulgated from the EU to other regions, where they become entrenched in relevant legal frameworks (which has been termed the "Brussels effect" [44]). Therefore, by aligning the handbook with EU regulations such as the MDR and General Data Protection Regulation, many (often less rigorous) legal requirements in other regions are likely already covered, and future efforts will likely only require relatively small additions (or removal) of requirements to align with local legislation. Once the assessment is taken up in different regions, a flag displayed at the top of the CEN ISO/TS 82304-2 quality label can indicate in which region the assessment was conducted (or recognized). This is similar to the EU energy label, which is used by many countries outside the EU with some, strong, or full alignment and displays the region or country flag at the top of the label [45].

Finally, it should be noted that the initial focus of the development and testing effort was on smaller organizations and start-ups, assuming that, if the assessment process was optimized for the latter, it would also be suitable for larger companies that tend to employ more subject matter experts. However, this study revealed that larger organizations may require somewhat different guidance from that required by small or medium-sized businesses due to differing organizational structures and interrelations between products and associated documentation. For instance, in large companies, a given health app may be a supportive tool for a medical intervention that is not delivered through the app itself, and documentation for these connected products may not be easily separable. Relatedly, certain policies in place at large companies, such as data privacy provisions, may be applicable to many offered products rather than being health app specific. In addition, in some cases, larger companies may have (recently) acquired smaller app development organizations, with only some but not all documentation already having been updated to align with the larger company's standard procedures. Such scenarios may cause uncertainties within large companies regarding what documentation is suitable and relevant for the app-specific CEN ISO/TS 82304-2 assessment if the evaluation guidance is mainly aligned with the needs of smaller organizations. As only 2 large companies were involved in this pilot study, additional testing is needed and planned to calibrate the assessment guidance for use in large organizations.

Envisioned Impact of the CEN ISO/TS 82304-2 Assessment

As mentioned in the Introduction section, the need for a health app quality certification beyond that provided under the MDR seems to be widely recognized [15,19,20], but current health app assessment initiatives, including national programs, are struggling to achieve efficient, appealing, and scalable assessment implementation. This may partly be due to the use of centralized national approaches to app evaluation, which can create bottlenecks [29], as well as due to insufficient incentives for developers. It has been proposed that international collaboration and the use of distributed evaluation bodies is needed to address these issues [29]. As such, the envisioned establishment of the CEN ISO/TS 82304-2 assessment as an

XSL•FO
RenderX

efficient, effective, consistent, and scalable health app quality certification scheme that is delivered through a network of assessment organizations and is recognized across Europe and beyond, requiring only limited context-specific requirements to be added, would be a major step forward.

It is expected that the widespread availability of the CEN ISO/TS 82304-2 quality label will minimize risk and build trust among health care professionals and the public, thereby increasing the adoption of high-quality health apps. In line with this expectation, our recent research has shown that health care professionals' willingness to recommend health apps and users' intention to download them is positively influenced by the CEN ISO/TS 82304-2 quality label ([46]; Adriaanse, unpublished data).

In addition, the cross-national recognition of the CEN ISO/TS 82304-2 assessment would help address several issues associated with the current fragmented assessment landscape in which numerous organizations apply differing assessment frameworks for overlapping quality aspects [19,20]. This fragmentation leads to duplicated work both for health systems and app developers if a given app needs to be submitted to several full assessments to evidence its quality to different markets and varying types of organizations. In this regard, it is specifically problematic that the quality criteria used can differ widely across frameworks, with a recent review finding that, among 12 health system- and government-led digital health quality assessments, the criteria overlap in some domains was <50% [20]. Such inconsistencies in quality standards can cause additional work (if varying evidence needs to be prepared and submitted) and confusion among both developers and users, potentially jeopardizing trust in quality evaluations if the same health app may fail one assessment but pass another. In addition, the need for an app to undergo several assessments to enter various markets may lead to inequities. While developers may regard it as worthwhile to invest time and money to receive quality certifications in countries with large populations, this is less likely to be the case for smaller markets, for which the cost-benefit ratio of undergoing an additional assessment and addressing local evidence needs, such as a study conducted within the related country, may not be appealing. This may lead to a situation in which some high-quality apps are not available in certain regions or languages.

The CEN ISO/TS 82304-2 assessment is intended to address these issues, avoiding duplication of work, preventing inequities from arising, and building trust by serving as a "gold standard" assessment that is recognized across a wide range of countries and types of organizations. A recent paper with the European Society of Cardiology [47] and preliminary results from a comparative analysis with European health technology assessment frameworks (Hoogendoorn, unpublished data) indicate that the CEN ISO/TS 82304-2 assessment covers the vast majority of criteria that are regarded as important to evidence health app quality. By primarily relying on the CEN ISO/TS 82304-2 quality label and only carrying out additional, context-specific evaluations if the label indicates that a given app is of acceptable quality, health systems and other stakeholders can obtain substantial time and resource savings during health app–related decision-making and create streamlined processes for developers.

Strengths and Limitations

This study took an innovative and ambitious approach to health app quality assessment by engaging organizations from 5 countries in the evaluation process and examining their interrater reliability. In addition, 24 health app development companies were involved in this study, and their feedback was sought. To our knowledge, this is the first such collaboration, which lays the foundation for the consistency and scaling of the CEN ISO/TS 82304-2 assessment across Europe and beyond.

However, some limitations should be noted. First, the sample size of included apps was small (given that each app was rated by only 2 assessment organizations), limiting the generalizability of the findings. This was partly due to the funding and time constraints of the project, as well as due to developer dropouts, especially in the first assessment round. Given the small sample involved in this pilot study and the evolving nature of the guidance and handbook throughout the study and beyond, future studies with a larger sample and using the operational version of the guidance and handbook are required to provide conclusive evidence for the interrater reliability of the CEN ISO/TS 82304-2 assessment.

Second, as mentioned previously, relatively high developer dropout rates were observed, which may have distorted the findings if developers who found the guidance more difficult to understand were more likely to drop out. However, this is unlikely to have been the case as many of the dropouts occurred before developers had even received the guidance, and across all rounds, only 3 developers dropped out after they had started to collate the assessment information. In this regard, it should also be emphasized that developers were not provided with a quality label as part of this study given that the aim was to develop and refine the CEN ISO/TS 82304-2 assessment handbook and guidance before arriving at an operational version that is suitable for use in quality certification. This likely contributed to the dropouts given that, as indicated during the developer interviews, receiving a quality label to build trust and support marketing efforts is the main benefit of the assessment for developers. It is expected that dropouts will be less common in the future implementation of the assessment once a quality label is issued and developers actively seek out the assessment of their apps to gain the benefits of the label.

Third, given that this study aimed to approximate real-world conditions with a focus on app, developer, and assessor diversity, some aspects were not as tightly controlled as they could have been, such as the application of eligibility criteria or the number of individual assessors involved in the evaluation of a given app across organizations. This was deemed necessary to allow for the detection and solution of potential issues that may arise during the real-life assessment process. Moreover, this approach ensured that the gathered consistency, time, and clarity data were representative of what would be expected in the real world (with the draft handbook and guidance) without being (potentially positively) distorted by artificially imposed conditions that will not be present once the assessment is rolled out.

XSL•FO
RenderX

Conclusions

This study involved an unprecedented collaboration across organizations in 5 countries to develop, test, and evolve the CEN ISO/TS 82304-2 health app quality assessment handbook and guidance. Promising findings were obtained, especially given the early stage of development, with regard to the interrater reliability between assessors, and valuable information was gathered to support the future evolution and implementation of the assessment. These are important steps toward establishing the CEN ISO/TS 82304-2 assessment as a commonly recognized health app quality evaluation across Europe and beyond. The widespread use of the CEN ISO/TS 82304-2 assessment across countries is expected to have a substantial positive impact, avoiding duplication of work, preventing inequities by facilitating access to smaller markets for developers, enhancing the quality of apps, and building trust in health apps among health care professionals and the public. This can ultimately help increase the adoption of high-quality health apps that have the potential to improve health outcomes and health care system efficiency.

Acknowledgments

The authors are grateful to the health app developers and assessment organizations that contributed to the testing of the CEN ISO/TS 82304-2 assessment. This study was funded by the European Union. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or of the European Health and Digital Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

ALF programmed the MATLAB (The MathWorks, Inc) script, analyzed the data, and wrote the manuscript draft. DM and RF led the project administration, and ALF, DM, BP, and RF curated the data. DM, BP, AM, RF, BL, LA, VdlH, CPB, BLC, GD, PAS, LMR, ME, YL, and MC played key roles in the assessment data acquisition, and MK led the interviews with app developers. PH conceptualized and designed the study and provided supervision. All authors reviewed and edited the manuscript and provided final approval for its publication.

Conflicts of Interest

All authors are involved in the Label2Enable project, which supports the implementation of the CEN ISO/TS 82304-2 health app quality assessment framework and related label across Europe. In addition, at the time of this study, all authors except PAS and MK were employed by organizations that provide digital health technology assessments.

References

1. Socio-economic impact of mHealth: an assessment report for the European Union. GSMA. URL: https://www.gsma.com/iot/wp-content/uploads/2013/06/Socio-economic_impact-of-mHealth_EU_14062013V2.pdf [accessed 2023-10-20]
2. Butcher CJ, Hussain W. Digital healthcare: the future. Future Healthc J 2022 Jul 28;9(2):113-117 [FREE Full text] [doi: 10.7861/fhj.2022-0046] [Medline: 35928188]
3. Alfawzan N, Christen M, Spitale G, Biller-Andorno N. Privacy, data sharing, and data security policies of women's mHealth apps: scoping review and content analysis. JMIR Mhealth Uhealth 2022 May 06;10(5):e33735 [FREE Full text] [doi: 10.2196/33735] [Medline: 35522465]
4. Melcher J, Torous J. Smartphone apps for college mental health: a concern for privacy and quality of current offerings. Psychiatr Serv 2020 Nov 01;71(11):1114-1119. [doi: 10.1176/appi.ps.202000098] [Medline: 32664822]
5. Sander LB, Schorndanner J, Terhorst Y, Spanhel K, Pryss R, Baumeister H, et al. 'Help for trauma from the app stores?' A systematic review and standardised rating of apps for Post-Traumatic Stress Disorder (PTSD). Eur J Psychotraumatol 2020 Jan 09;11(1):1701788 [FREE Full text] [doi: 10.1080/20008198.2019.1701788] [Medline: 32002136]
6. Tangari G, Ikram M, Ijaz K, Kaafar MA, Berkovsky S. Mobile health and privacy: cross sectional study. BMJ 2021 Jun 16;373:n1248 [FREE Full text] [doi: 10.1136/bmj.n1248] [Medline: 34135009]
7. Akbar S, Coiera E, Magrabi F. Safety concerns with consumer-facing mobile health applications and their consequences: a scoping review. J Am Med Inform Assoc 2020 Feb 01;27(2):330-340 [FREE Full text] [doi: 10.1093/jamia/ocz175] [Medline: 31599936]
8. Faessen JP, Lucassen DA, Buso ME, Camps G, Feskens EJ, Brouwer-Brolsma EM. Eating for 2: a systematic review of dutch app stores for apps promoting a healthy diet during pregnancy. Curr Dev Nutr 2022 Jun;6(6):nzac087 [FREE Full text] [doi: 10.1093/cdn/nzac087] [Medline: 35711572]
9. van Galen LS, Xu X, Koh MJ, Thng S, Car J. Eczema apps conformance with clinical guidelines: a systematic assessment of functions, tools and content. Br J Dermatol 2020 Feb;182(2):444-453 [FREE Full text] [doi: 10.1111/bjd.18152] [Medline: 31179535]

XSL•FO
RenderX

10. Simon L, Reimann J, Steubl LS, Stach M, Spiegelhalder K, Sander LB, et al. Help for insomnia from the app store? A standardized rating of mobile health applications claiming to target insomnia. J Sleep Res 2023 Feb 27;32(1):e13642. [doi: 10.1111/jsr.13642] [Medline: 35624078]
11. Consumer health apps and digital health tools proliferate, improving quality and health outcomes for patients. IQVIA. URL: https://tinyurl.com/4haescn6 [accessed 2023-10-20]
12. Hyzy M, Bond R, Mulvenna M, Bai L, Frey A, Carracedo JM, et al. Don't judge a book or health app by its cover: user ratings and downloads are not linked to quality. PLoS One 2024 Mar 4;19(3):e0298977 [FREE Full text] [doi: 10.1371/journal.pone.0298977] [Medline: 38437233]
13. Jacob C, Sanchez-Vazquez A, Ivory C. Understanding clinicians' adoption of mobile health tools: a qualitative review of the most used frameworks. JMIR Mhealth Uhealth 2020 Jul 06;8(7):e18072 [FREE Full text] [doi: 10.2196/18072] [Medline: 32442132]
14. Jacob C, Sezgin E, Sanchez-Vazquez A, Ivory C. Sociotechnical factors affecting patients' adoption of mobile health tools: systematic literature review and narrative synthesis. JMIR Mhealth Uhealth 2022 May 05;10(5):e36284 [FREE Full text] [doi: 10.2196/36284] [Medline: 35318189]
15. Digital Health in the WHO European region: the ongoing journey to commitment and transformation. World Health Organization. URL: https://tinyurl.com/yhs47c65 [accessed 2024-04-29]
16. van Drongelen A, de Bruijn A, Roszek B, Vonk R. Apps under the medical devices legislation. Rijksinstituut voor Volksgezondheid en Milieu. URL: https://tinyurl.com/2f4ytyre [accessed 2023-12-16]
17. Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC. The European Parliament and The Council of the European Union. 2017. URL: https://eur-lex.europa.eu/eli/reg/2017/745/oj/eng [accessed 2024-04-29]
18. Fan M, Yu L, Chen S, Zhou H, Luo X, Li S. An empirical evaluation of GDPR compliance violations in android mHealth apps. In: Proceedings of the 31st International Symposium on Software Reliability Engineering. 2020 Presented at: ISSRE '20; October 12-15, 2020; Coimbra, Portugal p. 253-264 URL: https://ieeexplore.ieee.org/document/9251060 [doi: 10.1109/issre5003.2020.00032]
19. D2.1 knowledge tool 1: health apps assessment frameworks. European mHealth Hub. URL: https://mhealth-hub.org/download/d2-1-knowledge-tool-1-health-apps-assessment-frameworks [accessed 2023-10-20]
20. Frey AL, Phillips B, Baines R, McCabe A, Elmes E, Yeardsley-Pierce E, et al. Domain coverage and criteria overlap across digital health technology quality assessments: a systematic review. Health Technol 2024 Dec 20;15(1):15-27. [doi: 10.1007/S12553-024-00934-8]
21. Hoogendoorn P, Versluis A, van Kampen S, McCay C, Leahy M, Bijlsma M, et al. What makes a quality health app-developing a global research-based health app quality assessment framework for CEN-ISO/TS 82304-2: Delphi study. JMIR Form Res 2023 Jan 23;7:e43905 [FREE Full text] [doi: 10.2196/43905] [Medline: 36538379]
22. ISO/TS 82304-2:2021: health software: part 2: health and wellness apps — quality and reliability. International Organization for Standardization. URL: https://www.iso.org/standard/78182.html [accessed 2023-07-14]
23. Freitag B, Fehring L, Uncovska M, Olsacher A, Meister S. Negotiating pricing and payment terms for insurance covered mHealth apps: a qualitative content analysis and taxonomy development based on a German experience. Health Econ Rev 2024 Oct 04;14(1):81. [doi: 10.1186/s13561-024-00558-8] [Medline: 39365482]
24. Goeldner M, Gehder S. Digital health applications (DiGAs) on a fast track: insights from a data-driven analysis of prescribable digital therapeutics in Germany from 2020 to mid-2024. J Med Internet Res 2024 Aug 29;26:e59013 [FREE Full text] [doi: 10.2196/59013] [Medline: 39208415]
25. Frey AL, Leigh S, Toro C, Bufill CP, McCay C, Trupec TP, et al. Preferences and willingness to pay for health app assessments among healthcare stakeholders: a discrete choice experiment. JMIRP Preprint posted online February 17, 2024 [FREE Full text] [doi: 10.2196/preprints.57474]
26. Ataiy O, Burniki M, Sohrabi K, Groß V, Scholtes M, Schudt F. Digital health applications in Germany: explorative pilot study on the challenges of approval from the perspective of manufacturers. Digit Health 2024 Dec 19;10:20552076241302676 [FREE Full text] [doi: 10.1177/20552076241302676] [Medline: 39711743]
27. Schmidt L, Pawlitzki M, Renard BY, Meuth SG, Masanneck L. The three-year evolution of Germany's digital therapeutics reimbursement program and its path forward. NPJ Digit Med 2024 May 24;7(1):139 [FREE Full text] [doi: 10.1038/s41746-024-01137-1] [Medline: 38789620]
28. Mäder M, Timpel P, Schönfelder T, Militzer-Horstmann C, Scheibe S, Heinrich R, et al. Evidence requirements of permanently listed digital health applications (DiGA) and their implementation in the German DiGA directory: an analysis. BMC Health Serv Res 2023 Apr 17;23(1):369 [FREE Full text] [doi: 10.1186/s12913-023-09287-w] [Medline: 37069592]
29. Essén A, Stern AD, Haase CB, Car J, Greaves F, Paparova D, et al. Health app policy: international comparison of nine countries' approaches. NPJ Digit Med 2022 Mar 18;5(1):527 [FREE Full text] [doi: 10.1038/S41746-022-00573-1]
30. About CEN-ISO/TS 82304-2. Label2Enable. URL: https://label2enable.eu/about-cen-iso-ts-82304-2 [accessed 2025-01-04]
31. Haley DT, Thomas P, Petre M, De Roeck A. Using a new inter-rater reliability statistic. Faculty of Mathematics, Computing and Technology, The Open University. URL: https://oro.open.ac.uk/90234/1/TR2008_15.pdf [accessed 2024-05-30]

XSL•FO
RenderX

32. Gwet K. Kappa statistic is not satisfactory for assessing the extent of agreement between raters. Statistical Methods For Inter-Rater Reliability Assessment. 2002. URL: https://agreestat.com/papers/kappa_statistic_is_not_satisfactory.pdf [accessed 2024-04-29]
33. Wongpakaran N, Wongpakaran T, Wedding D, Gwet KL. A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. BMC Med Res Methodol 2013 Apr 29;13(1):61 [FREE Full text] [doi: 10.1186/1471-2288-13-61] [Medline: 23627889]
34. Dona SW, Nguyen D, Angeles MR, Cooper P, Winter N, Chatterton ML, et al. Developing and testing a health app evaluation framework for organisations to recommend the best health apps for consumers in the Australian setting. Deakin University. URL: https://iht.deakin.edu.au/wp-content/uploads/sites/153/2022/07/D-HEAL-Final-Report.pdf [accessed 2023-07-14]
35. Powell AC, Torous J, Chan S, Raynor GS, Shwarts E, Shanahan M, et al. Interrater reliability of mHealth app rating measures: analysis of top depression and smoking cessation apps. JMIR Mhealth Uhealth 2016 Feb 10;4(1):e15 [FREE Full text] [doi: 10.2196/mhealth.5176] [Medline: 26863986]
36. Miró J, Llorens-Vernet P. Assessing the quality of mobile health-related apps: interrater reliability study of two guides. JMIR Mhealth Uhealth 2021 Apr 19;9(4):e26471 [FREE Full text] [doi: 10.2196/26471] [Medline: 33871376]
37. Baumel A, Faber K, Mathur N, Kane JM, Muench F. Enlight: a comprehensive quality and therapeutic potential evaluation tool for mobile and web-based eHealth interventions. J Med Internet Res 2017 Mar 21;19(3):e82 [FREE Full text] [doi: 10.2196/jmir.7270] [Medline: 28325712]
38. Lagan S, Aquino P, Emerson MR, Fortuna K, Walker R, Torous J. Actionable health app evaluation: translating expert frameworks into objective metrics. NPJ Digit Med 2020;3:100 [FREE Full text] [doi: 10.1038/s41746-020-00312-4] [Medline: 32821855]
39. Terhorst Y, Philippi P, Sander LB, Schultchen D, Paganini S, Bardus M, et al. Validation of the Mobile Application Rating Scale (MARS). PLoS One 2020 Nov 2;15(11):e0241480 [FREE Full text] [doi: 10.1371/journal.pone.0241480] [Medline: 33137123]
40. Baines R, Hoogendoorn P, Stevens S, Chatterjee A, Ashall-Payne L, Andrews T, et al. Navigating medical device certification: a qualitative exploration of barriers and enablers amongst innovators, notified bodies and other stakeholders. Ther Innov Regul Sci 2023 Mar 04;57(2):238-250 [FREE Full text] [doi: 10.1007/s43441-022-00463-4] [Medline: 36194368]
41. Mackey R, Gleason A, Ciulla R. A novel method for evaluating mobile apps (app rating inventory): development study. JMIR Mhealth Uhealth 2022 Apr 15;10(4):e32643 [FREE Full text] [doi: 10.2196/32643] [Medline: 35436227]
42. Jääskelä J, Haverinen J, Kaksonen R, Reponen J, Halunen K, Tokola T, et al. Digi-HTA, assessment framework for digital healthcare services: information security and data protection in health technology – initial experiences. FinJeHeW 2022 Apr 14;14(1):19-30. [doi: 10.23996/fjhw.111776]
43. Scherer J, Youssef Y, Dittrich F, Albrecht U, Tsitsilonis S, Jung J, et al. Proposal of a new rating concept for digital health applications in orthopedics and traumatology. Int J Environ Res Public Health 2022 Nov 13;19(22):14952 [FREE Full text] [doi: 10.3390/ijerph192214952] [Medline: 36429670]
44. Bradford A. The Brussels Effect: How the European Union Rules the World. Oxford, UK: Oxford Academic Press; 2012.
45. New energy efficiency labels explained. European Commission. 2019. URL: https://ec.europa.eu/commission/presscorner/detail/en/memo_19_1596 [accessed 2024-04-29]
46. Biliunaite I, van Gestel L, Hoogendoorn P, Adriaanse M. Value of a quality label and European healthcare professionals' willingness to recommend health apps: an experimental vignette study. J Health Psychol 2024 Aug 02:13591053241258205 [FREE Full text] [doi: 10.1177/13591053241258205] [Medline: 39096027]
47. Caiani EG, Kemps H, Hoogendoorn P, Asteggiano R, Böhm A, Borregaard B, et al. Standardised assessment of evidence supporting the adoption of mobile health solutions: a clinical consensus statement of the ESC regulatory affairs committee. Eur Heart J Digit Health 2024 Sep;5(5):509-523. [doi: 10.1093/ehjdh/ztae042] [Medline: 39318699]

Abbreviations

EU: European Union
MDR: Medical Device Regulation

XSL•FO
RenderX

Edited by A Mavragani; submitted 20.07.24; peer-reviewed by E Baker, J Haverinen; comments to author 19.12.24; revised version received 09.01.25; accepted 12.01.25; published 10.03.25.

Please cite as:
Frey AL, Matei D, Phillips B, McCabe A, Fuller R, Laibarra B, Alonso L, de la Hoz V, Pratdepadua Bufill C, Llebot Casajuana B, D'Avenio G, Sottile PA, Rocchi LM, Errera M, Laaissaoui Y, Cardinal M, Kok M, Hoogendoorn P
Testing and Iterative Improvement of the CEN ISO/TS 82304-2 Health App Quality Assessment: Pilot Interrater Reliability Study
JMIR Form Res 2025;9:e64565
URL: https://formative.jmir.org/2025/1/e64565
doi:10.2196/64565
PMID:

©Anna-Lena Frey, Diana Matei, Ben Phillips, Adam McCabe, Rachel Fuller, Begoña Laibarra, Laura Alonso, Victor de la Hoz, Carme Pratdepadua Bufill, Berta Llebot Casajuana, Giuseppe D'Avenio, Pier Angelo Sottile, Laura Melania Rocchi, Matteo Errera, Yasmine Laaissaoui, Michael Cardinal, Menno Kok, Petra Hoogendoorn. Originally published in JMIR Formative Research (https://formative.jmir.org), 10.03.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Digital Translation Platform (Translatly) to Overcome Communication Barriers in Clinical Care: Pilot Study

Victor Olsavszky[1], PD, Dr med; Mutaz Bazari[2]; Taieb Ben Dai[2]; Ana Olsavszky[3]; Fabian Finkelmeier[4], Prof Dr Med; Mireen Friedrich-Rust[5], Prof Dr Med; Stefan Zeuzem[5], Prof Dr Med; Eva Herrmann[6], Prof Dr Med; Jan Leipe[7], Prof Dr Med; Florian Alexander Michael[5], Dr med; Hans von Westernhagen[2]; Olivier Ballo[4], PD, Dr med

[1]Department of Dermatology, Venereology and Allergy, University Medical Center and Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany

[2]Translatly UG, Offenbach, Germany

[3]Clinic for General and Visceral Surgery, Vivantes Klinikum Kaulsdorf, Berlin, Germany

[4]Department of Internal Medicine 2, Hematology/Oncology, Goethe University Hospital, Frankfurt am Main, Germany

[5]Department of Internal Medicine 1, Gastroenterology, Hepatology and Endocrinology, Goethe University Hospital, Frankfurt am Main, Germany

[6]Institute of Biostatistics and Mathematical Modeling, Goethe University Hospital, Frankfurt am Main, Germany

[7]Department of Medicine V, Division of Rheumatology, University Medical Center and Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany

Corresponding Author:
Victor Olsavszky, PD, Dr med
Department of Dermatology, Venereology and Allergy
University Medical Center and Medical Faculty Mannheim
Heidelberg University
Theodor-Kutzer-Ufer 1-3
Mannheim, 68167
Germany
Phone: 49 621 383 2280
Fax: 49 621 383 3815
Email: victor.olsavszky@medma.uni-heidelberg.de

Abstract

Background: Language barriers in health care can lead to misdiagnosis, inappropriate treatment, and increased medical errors. Efforts to mitigate these include using interpreters and translation tools, but these measures often fall short, particularly when cultural nuances are overlooked. Consequently, medical professionals may have to rely on their staff or patients' relatives for interpretation, compromising the quality of care.

Objective: This formative pilot study aims to assess the feasibility of Translatly, a digital translation platform, in clinical practice. Specifically, the study focuses on evaluating (1) how health care professionals overcome language barriers and their acceptance of an on-demand video telephony platform, (2) the feasibility of the platform during medical consultations, and (3) identifying potential challenges for future development.

Methods: The study included ethnographic interviews with health care professionals and an observational pilot to assess the use of the Translatly platform in clinical practice. Translatly was developed to make real-time translation easy and accessible on both Android and iOS devices. The system's backend architecture uses Java-based services hosted on DigitalOcean. The app securely exchanges data between mobile devices and servers, with user information and call records stored in a MySQL database. An admin panel helps manage the system, and Firebase integration enables fast push notifications to ensure that health care professionals can connect with translators whenever they need to. The platform was piloted in a German university hospital with 170 volunteer nonprofessional translators, mainly medical students, supporting translation in over 20 languages, including Farsi, Dari, and Arabic.

Results: Ethnographic research conducted by interviewing health care professionals in Frankfurt am Main and other German cities revealed that current practices for overcoming language barriers often rely on family members or digital tools such as Google Translate, raising concerns about accuracy and emotional distress. Respondents preferred an on-demand translation service staffed by medically experienced translators, such as medical students, who understand medical terminology and can empathize

XSL•FO
RenderX

with patients. The observational pilot study recorded 39 requests for translation services, 16 (41%) of which were successfully completed. The translations covered 6 different languages and were carried out by a team of 10 translators. Most requests came from departments such as infectious diseases (5/16, 31%) and emergency (4/16, 25%). Challenges were identified around translator availability, with 23 (59%) total requests (N=39) going unanswered, which was further evidenced by user feedback.

Conclusions: This pilot study demonstrates the feasibility of the Translatly platform in real-world health care settings. It shows the potential to improve communication and patient outcomes by addressing language barriers. Despite its potential, challenges such as translator availability highlight the need for further development.

(JMIR Form Res 2025;9:e63095) doi:10.2196/63095

KEYWORDS

language barriers; health care communication; medical app; real-time translation; medical translation

Introduction

Language barriers in health care can have a profound impact on patient management. Patients who are not fluent in the primary language of the health care facility are at higher risk of misdiagnosis due to miscommunication during history taking or description of symptoms [1]. This can lead to inappropriate treatments and interventions, worsening the patient's condition [2]. In addition, there may be an increased risk of medical errors or even an increased risk of a serious medical event [3]. Patients may feel ignored, excluded, or misunderstood, leading to reduced satisfaction with care. Language barriers can also hinder the development of trust between the patient and the health care provider [4]. Clinic staff, by contrast, complain of frustration due to suboptimal treatment and the additional time spent with patients as a result of miscommunication [5]. To address these challenges, many health care facilities employ interpreters or multilingual staff and use written materials, simplified language, or even online translation tools [6-8]. Difficulties in communication and misunderstandings can result in additional costs [9], ultimately borne by hospitals, statutory health insurance funds, and the health care system as a whole.

In Germany, there are more than 6 million people with limited knowledge of German. More importantly, there are no or only inadequate translation services available in medical facilities. A study commissioned by the Federal Anti-Discrimination Agency summarizes the role of communication barriers as a risk of discrimination in health care, based on data from 2010 to 2020 in Germany [10]. One of the main issues raised is the lack of efficient access to professional translators for German health services. However, given that Germany is currently home to nearly 22 million people with a migrant background, and that an estimated 25% of German citizens speak 2 foreign languages in addition to German [11], this hidden potential of language skills can be reflected in health care professionals such as doctors, medical students, nurses, and medical technicians. At present, this diversity of foreign language skills is either not used at all or only used in exceptional situations that occur by chance to overcome communication obstacles in the medical sector. In addition, the shortage of health care professionals [12] will lead to a further increase in the number of people with foreign language skills in German society due to immigration in the coming years [13].

The problem of language barriers in health care is not only confined to Germany. As the world's population becomes increasingly diverse and migratory, the need to ensure effective patient-provider communication in a multilingual context is growing. Other countries with a long history of immigration have faced similar challenges in their health care systems. In Canada, for instance, the Winnipeg Regional Health Authority, a local health care organization, employs trained health interpreting staff for in-person and telephone interpreting [14]. While such language services are useful proactive policies, they are not available in all Canadian communities [15] and cannot meet the most common ad hoc interpreting needs [16]. In the United States, state laws require health care facilities to provide interpretation services for patients with limited English proficiency, but these requirements cannot be adequately met [17]. Finally, in Australia, almost half of all clinical practices make use of the translation and interpreting service, which is specifically funded by the government to address language barriers in health care [18]. While these examples provide valuable insights and potential solutions that Germany could consider in its efforts to address its own unique challenges related to language barriers, none of these possible approaches take into account the nuances of dialect or the importance of cultural context in communication. This is where technology-enabled solutions, such as Translatly (Translatly UG), play a critical role in addressing these barriers more effectively. Translatly harnesses the existing, underutilized linguistic potential within the health care workforce, offering an innovative and scalable solution to overcome communication obstacles. With more than €400 (US $436) billion spent on health care in Germany each year [19], digitally targeting these resources to overcome translation barriers could have a positive impact not only on health care delivery, but also on health care expenditure.

With this formative pilot study, we aim to assess the feasibility of Translatly, a digital eHealth platform designed to address language barriers in everyday clinical practice in Germany. Our approach focuses on implementing this digital on-demand translation platform for clinics. The main research questions of the study are as follows: (1) What are the current practices of health care professionals in overcoming language barriers and do they accept an on-demand medical translation platform? (2) Is the Translatly platform feasible for use in clinical settings to overcome language barriers? (3) What other challenges may arise in the use of Translatly that can be addressed in future studies? Translatly is a mobile and tablet app developed in Germany by a founding team of doctors, technicians, and

XSL•FO
RenderX

entrepreneurs to facilitate live translation in everyday medical practice by virtually connecting qualified language mediators with doctors and their patients. This platform allows in-house health care professionals with foreign language skills to register as translators who can be requested and engaged via video telephony when communication barriers arise during doctor-patient interactions. By providing an accessible, scalable, and efficient solution for translation, Translatly has the potential to not only improve the care and management of nonnative patients but also reduce health care costs, making it a sustainable solution for the future.

Methods

Study Design

This formative pilot study used a mixed methods approach combining qualitative ethnographic interviews and observational data collection methods to assess the feasibility of the digital translation platform Translatly in a clinical setting. To explore the extent of language barriers in health care and how they are managed in daily practice, an extensive opinion survey was conducted using an ethnographic questionnaire at the University Hospital of the Goethe University in Frankfurt am Main, as well as in private practices in the Frankfurt area and other selected cities.

The findings from these interviews were used in the design and development of the platform. An observational pilot study was then conducted to test the feasibility of the platform in real medical settings. The use of the Translatly platform was monitored and recorded without manipulating any variables, focusing on how it was integrated into the routine clinical management of patients by health care providers. The pilot study took place over 2 months, from December 1, 2022, to January 31, 2023, at the University Hospital of the Goethe University in Frankfurt am Main.

Study Participants

Participants were recruited using convenience sampling. Health care professionals, including doctors, physiotherapists, and psychotherapists, participated in interviews using an ethnographic research questionnaire. They were identified and contacted through 2 primary methods: first, practitioners at the University Hospital of the Goethe University in Frankfurt am Main were approached directly; second, health care professionals from private practices in the Frankfurt and Offenbach areas were contacted by email, followed by telephone calls to those who expressed interest. Contact details for these private practices were obtained from public directories maintained by the University Hospital of Goethe University. Additionally, practitioners in Düsseldorf and the United States were contacted through recommendations from respondents.

Locating volunteer translators within the health care workforce and matching them with medical staff presented a technical challenge. To ensure the platform's optimal function, a substantial pool of volunteer translators was needed. Recruitment focused on medical students, as they possess the medical knowledge to handle complex translations [20]. They were enrolled through a clinical elective course titled "Translatly and Digitalization in Healthcare," which also provided them with community service certification. Volunteer translators were briefed on data protection regulations, including General Data Protection Regulation (GDPR) compliance.

By the last quarter of 2022, a total of 170 volunteer translators had been registered, 90% (n=153) of whom were medical students from the University Hospital. The remaining 10% (n=17) comprised medical staff. For the pilot study, translators were recruited for the following languages: Farsi, Dari, Arabic, Italian, Urdu, Portuguese, Spanish, Turkish, Russian, Albanian, Vietnamese, Romanian, Slovenian, Pashto, Macedonian, French, Croatian, Hindi, Bosnian, Azerbaijani, Korean, and Berber.

Patients were recruited through direct contact with doctors at the University Hospital, who identified cases needing translation services. Doctors interested in participating were provided instructions on how to use the Translatly app, and patients were also supported in using the app if needed.

Data Collection

Overview

Data were collected through ethnographic interviews with health care professionals using a questionnaire, as well as observational monitoring of the Translatly platform in clinical practice during the pilot study. The interviews focused on current practices for overcoming language barriers, while the observational data recorded translation requests, languages, and involved medical departments.

The ethnographic questions were designed specifically for this study and followed the format of semistructured interviews (Multimedia Appendix 1). This approach provided a mix of open-ended questions and specific information requests, allowing for both structured inquiry and the flexibility for interviewees to express their own experiences and opinions freely. Specifically, practitioners were asked about their personal experiences, their challenges in dealing with non–German-speaking patients, and their thoughts on how to improve communication. Questions about the use of software and technology, and specific instances where translation services have been required, provide deeper insights into everyday practice and the issues involved. Typical of ethnographic interviews, the questionnaire can be used in a context where interviews are more informal and conversational. The interviews were designed to last approximately 50 minutes, with time allocated for each section of the interview guide.

The pilot study collected data on the frequency of requests for translation services, the response rate to these requests, the duration and languages of completed translations, and the hospital departments making these requests. User feedback was also collected to assess problems and challenges during the pilot phase. This information allows for a comprehensive assessment of the use of the Translatly platform in clinical practice by providing data on the demand for translation services, the availability of the translator network, and the different medical specialties requiring translations. As this feasibility study only aims to measure and observe whether Translatly can be used as a translation platform in everyday clinical practice, a comparison

XSL•FO
RenderX

group (eg, patients without a translation platform) was not required.

Instruments

Data from ethnographic interviews were collected through written notes and supplemented by audio recordings to ensure accuracy and completeness. The interviews were then transcribed into words for analysis.

The tools required for the observational pilot study were the Translatly app and a mobile device or tablet. While some medical departments at the University Hospital of the Goethe University in Frankfurt am Main provided their own tablets, this was not a requirement as the app could be used by doctors, patients, and translators on their own smartphones. Translatly was preinstalled on all devices used in the pilot study with the help of the study team. Network connectivity was provided by the university hospital's wireless network or by the mobile phone's network provider.

Analytical Techniques and Software

Qualitative Content Analysis

A qualitative content analysis was conducted to systematically evaluate the data from the ethnographic interviews. This analysis provided insights into clinical standards, patient documentation, and the use of translation services.

The analysis of the pilot study data was performed using Excel (version 2408 build 16.0.17928.20114, 64-bit; Microsoft Corp.), with raw data provided by the Translatly digital platform, which is methodologically described in detail below.

Operating Mode of the Translatly Digital Translation Platform

Translatly's digital translation platform is designed to provide medical facilities, such as hospitals or private practices, with an app that can be used on smartphones or tablets within the facility. With the app, doctors or other medical staff treating patients can access a pool of translators who can provide language assistance during patient consultations. The medical practitioner chooses the language to be translated and the Translatly algorithm selects and contacts the appropriate translators from the Translatly pool (Figure 1). The first translator to accept the contact request is connected directly to the treating team. The medical staff, patient, and translator communicate via Translatly's internal audio or video interface.

Figure 1. Flowchart describing the operational process of the Translatly digital translation platform. It starts with a medical practitioner (doctor) identifying the need for a translator during a patient consultation and selecting the desired language for translation using the Translatly app on a smartphone or tablet. Once the patient selects their preferred mode of communication, either audio or video, the Translatly algorithm is activated and searches its database for available translators who have registered as proficient in the required language. The algorithm sends a contact request to these translators. The translators receive the request on their devices and the first one to accept the request is connected to the doctor and the patient. If the translator is accepted, an audio or video call is initiated. If all the translators reject the request or do not respond within a certain time frame, the algorithm notifies the doctor and the patient that the search was unsuccessful. At the end of the call, the translator's screen returns to the list of recent calls and the doctor's and patient's screen returns to language selection. Doctor and patient process operations are shown in purple, while translator operations are shown in green.

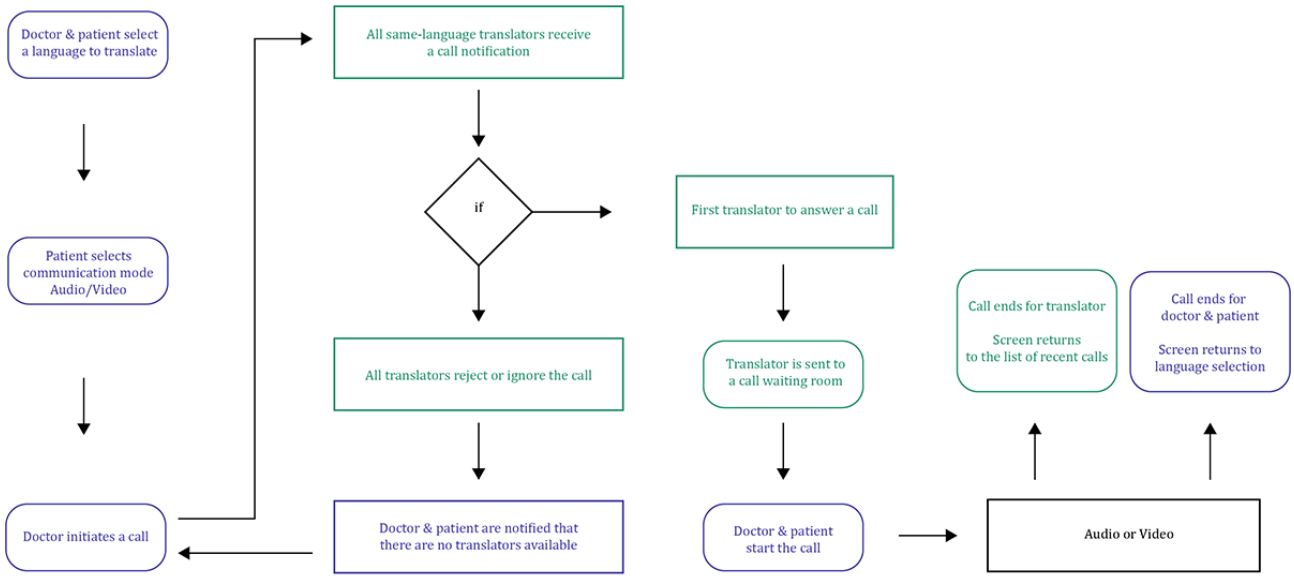

Translatly's Digital Platform Development Framework

The initial steps to ensure the successful implementation of the digital platform before the full implementation of the pilot phase consisted of 3 main pillars: (1) evaluation of the feasibility of a digital interaction-translation platform in a German university hospital setting; (2) regulatory compliance research; and (3) app design and development.

Following an extensive review of the existing medical literature, the feasibility assessment (pillar 1) consisted of developing the ethnographic research questionnaire (Multimedia Appendix 1) to determine the impact of language barriers in health care and how they are managed in daily practice.

Regulatory compliance research (pillar 2) involved an in-depth assessment of relevant medical regulations and standards in Germany. In brief, our team ensured that the app complies with

XSL•FO
RenderX

regulatory requirements for data security and privacy. In addition, we ensured that our research project was in line with the purpose of the Innovation Fund, which was established within the German Federal Joint Committee to promote new forms of health care that go beyond the existing standard of care covered by statutory health insurance in Germany [21].

Design thinking workshops, cofacilitated by members of the clinical research team and software engineering specialists, marked the start of the design and development phase (pillar 3). These sessions set the stage for a subsequent design sprint phase. During this phase, the software development team built and evaluated prototypes and integrated and validated the selected ideas for the app. The research team consisted of a multidisciplinary collective, including 2 software developers with computer science expertise, 2 clinical oncologists to provide medical insight, a clinical trial investigator to oversee research integrity, and a business engineer to manage regulatory issues.

Translatly Service System Architecture

To ensure the desired functionality of the Translatly digital interaction platform, a high-level software, database, and server architecture have been defined and implemented by our software development team (Figure 2). To achieve cross-platform compatibility on both Android (Google LLC/Alphabet Inc.) and iOS (Apple Inc.) devices, the architecture consists of 2 main mobile apps developed using the Flutter framework [22]. One app is designed specifically for medical staff, acting as a practitioner interface, while the other acts as a translator interface, supporting real-time communication via audio or video telephony.

Figure 2. Architecture and data flow of the Translatly Digital Interaction Platform. This diagram illustrates the systematic process from the initiation of a translation request to the establishment of a video telephony session between medical staff and translators. Each step (0-12) marks a critical interaction between components of the Translatly platform, ensuring an efficient and fast response to overcome language barriers in real-time health care scenarios. “0. Administration panel”: Initially developed with AngularJS for the front end, Java Spring for the backend, and integrated with a MySQL database, the administration panel is used by administrators to create, edit, and delete accounts for medical staff and translators. “1. Language request”: The operator uses the mobile app to send a Representational State Transfer (REST) request for translation services to the backend hosted on DigitalOcean. ”2. Translator list retrieval”: Upon receiving the request, the backend accesses the MySQL database to retrieve a list of available translators who speak the required language. “3. Translator notification request”: The backend sends a request to Firebase, the cloud service, to initiate a push notification to the appropriate translators. “4. Push notification to translators”: Firebase processes the request and sends a push notification to the translators' mobile devices. “5. Translator availability confirmation”: The first translator to respond sends an availability confirmation back to the backend via the REST application programming interface (API). “6. Notification of translator confirmation”: Upon confirmation, the backend updates the Firebase cloud service with the translator's availability. “7. Cancellation notification to translators”: Firebase will then send a push notification to the remaining translators to cancel the call as a translator has been found. “8. Notification of attending staff”: Firebase also notifies the medical staff that a translator has been secured for the session. “9. Meeting link to medical staff”: The medical staff member receives the meeting link from the backend after joining the call. “10. Meeting link to the translator”: Similarly, the translator receives the meeting link via a push notification from Firebase. “11. Medical staff meeting join”: The medical staff member joins the video chat meeting using the link provided. “12. Translator meeting join”: The translator also joins the video chat meeting so that the translation session can begin.

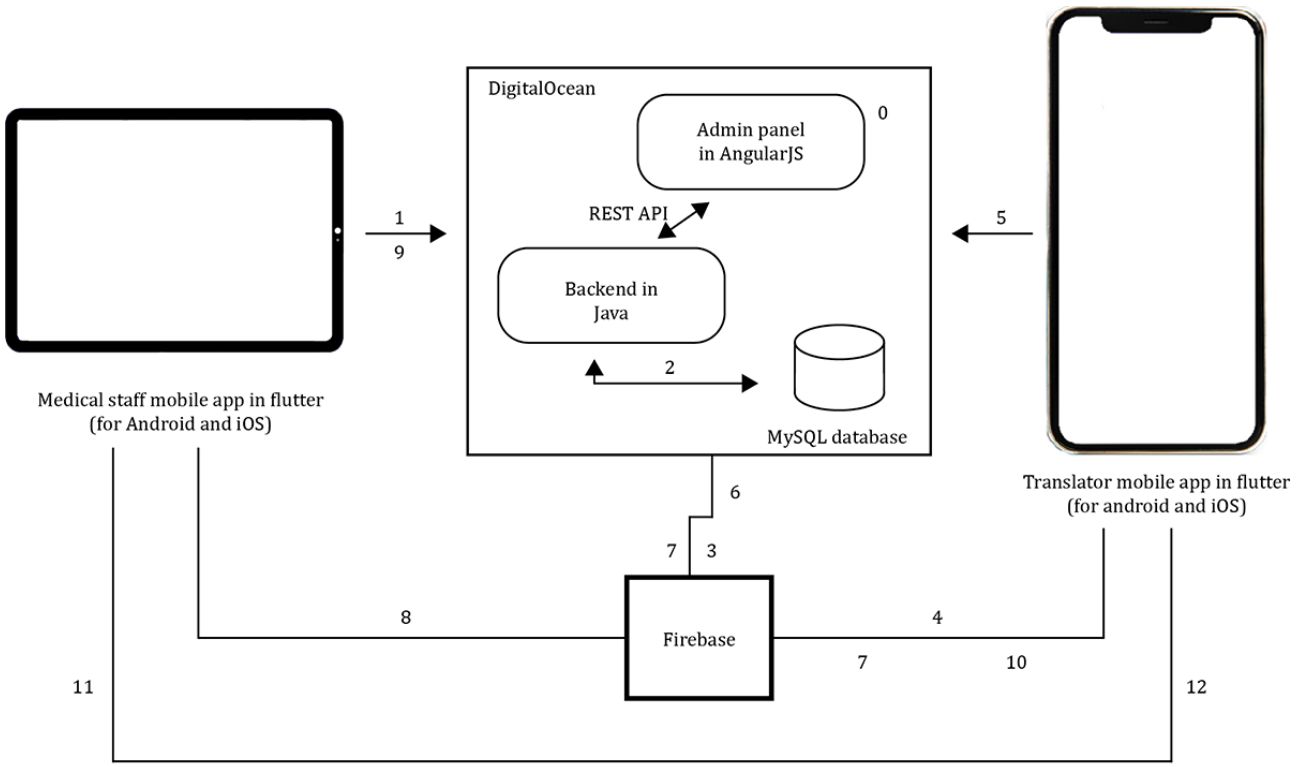

The platform’s backend is built on Java-based services, hosted on DigitalOcean (DigitalOcean Holdings, Inc.) for a stable and scalable cloud infrastructure [23]. The services can be accessed through a REST API (Representational State Transfer Application Programming Interface) [24] protected by an OAuth (Open Authorization) algorithm [25] that secures the connection

XSL•FO
RenderX

between the mobile apps and the server, ensuring efficient data transfer and real-time interactions. Moreover, a comprehensive administrative panel, originally developed with AngularJS (Angular Java Script) [26] and later migrated to ReactJS (React Java Script) [27], provides full system management and is responsible for maintaining user profiles, language preferences, and service availability of registered translators.

For data persistence, a MySQL (Oracle Corporation) database is used to store structured data, handling user information, interaction records, and call history [28]. To facilitate push notifications, we have integrated Firebase (Google LLC/Alphabet Inc.) into our system architecture [29]. This integration enables in-house health care professionals to be promptly notified and instantly connected to ongoing consultations. Real-time database and user authentication are implemented internally. Finally, Twilio (Twilio Inc.) is used as the video communication tool, known for enabling developers to build custom communication solutions [30].

Translatly User Experience Design and Interface Functionality

To ensure the intended mode of operation, we have opted for an intuitive user interface, specifically designed to guarantee and immediately establish an optimal interaction between medical professionals and translators. Upon opening the app, medical professionals must enter their credentials provided by our team to securely access Translatly's services (Multimedia Appendix 2). After logging in, they are presented with the main app screen, which prominently features an on-demand translation tab with an extensive language selection menu. Next, the medical staff user can choose between audio and video call modes, conceivably in the presence of and in agreement with the patient. The interface then displays the status of the translator call process and notifies the medical staff when a translator has accepted the call, providing an easy transition into the call.

On the other side, translators access their dedicated portal through a similar secure log-in interface (Figure 3A). Their main screen displays recent and missed calls, with relevant details such as the caller's name and the languages translated during the respective calls (Figure 3B). An incoming call is displayed as a slide-up in-app notification with relevant information (Figure 3C). Upon acceptance, the translator is temporarily held in a virtual waiting room until the calling user agrees to start the call (Figure 3D). After the call (Figure 3E), a postcall notification provides a summary of the call (Figure 3F).

XSL•FO
RenderX

Figure 3. Selection of Translatly mobile frontend screenshots showing different stages of the user experience for translators. (A) Log-in screen with header and logo, where translators enter their credentials to access the platform. (B) The main screen displays the last calls and missed calls of the translator, including the language that was interpreted during the call and the name of the doctor. (C) Notification when a translator receives a video call request from a health care professional, indicating the languages to be translated. (D) Upon acceptance of the call, the translator is directed to a waiting room and is awaiting entry into the consultation. (E) Active video call interface used by the translator to communicate with medical staff and patient. (F) The postcall screen provides a summary of the call duration and a thank you message with an option to end the session.

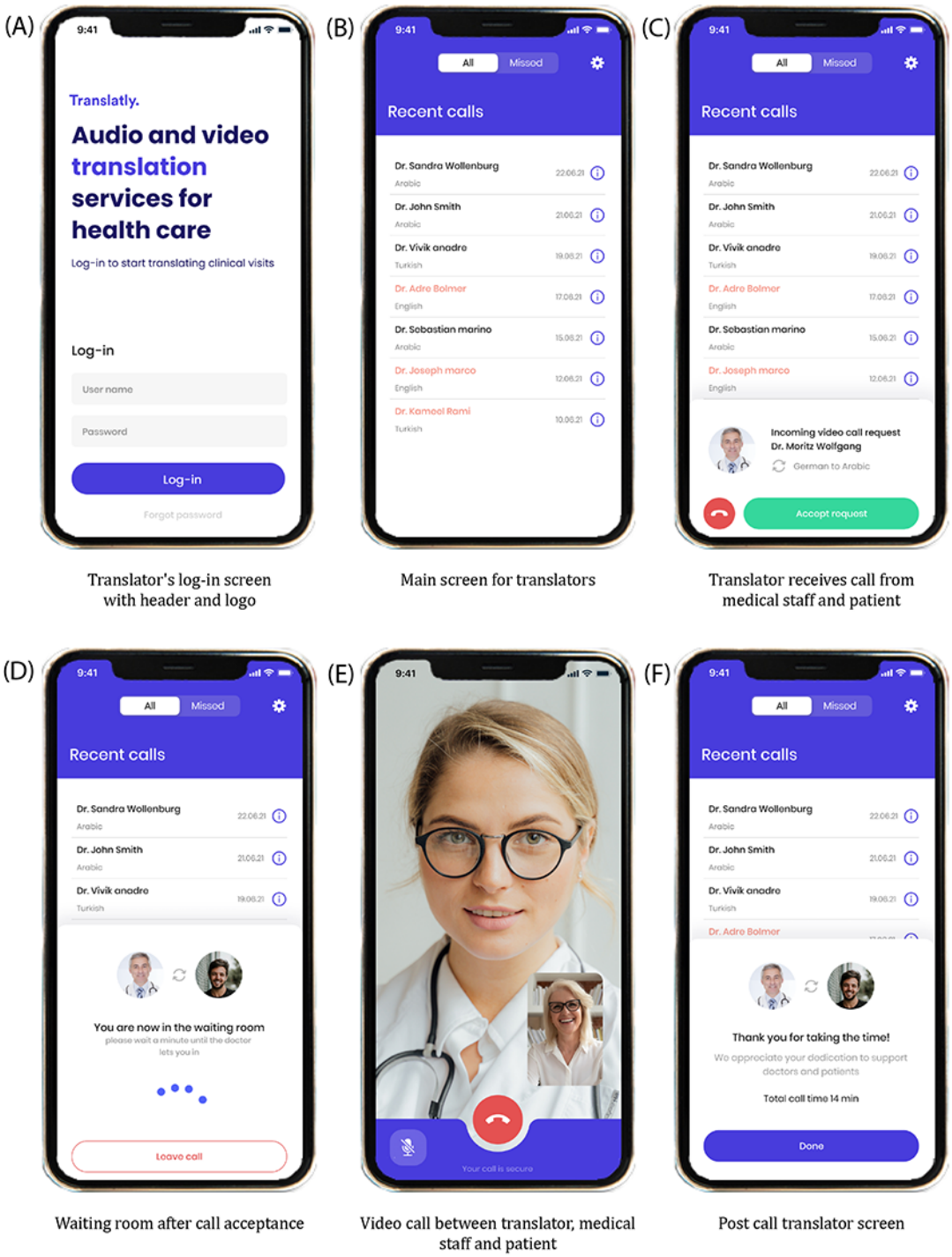

Ethical Considerations

This study involving human participants has been reviewed and approved by the Medical Ethics Committee of the Goethe University, Frankfurt am Main, Germany (ethics approval 2024-1776). The study was conducted in accordance with the principles of the Declaration of Helsinki. All participants provided informed consent for their involvement in the study. Additionally, the Medical Ethics Committee has approved the secondary analysis of the pilot phase data without the need for further consent. We confirm that the Translatly team adheres to regulations, including compliance with the GDPR, the Federal Data Protection Act, the Hessian Data Protection Act, and confidentiality in accordance with § 203 of the German Criminal Code. In addition, all data collected are anonymized. No compensation was provided to participants in this study. All images used in this study are open source and publicly available

XSL•FO
RenderX

on the official Translatly website [31], with no identifiable individual participants or users depicted.

Results

Summary of Responses to Ethnographic Research Questionnaires

The ethnographic research was conducted to assess the impact of language barriers in health care, and to gain a holistic view of the challenges health care professionals face in their clinical routine when dealing with nonnative speaking patients. Respondents included specialists and therapists from maximum care hospitals or university hospitals (6/10, 60%), from private practice (3/10, 30%), or from both the private and public sectors (1/10, 10%; Multimedia Appendix 3). The majority of respondents were oncologists (4/10, 40%) from major German cities such as Offenbach am Main (2/10, 20%), Frankfurt am Main (1/10, 10%), Düsseldorf (1/10, 10%), and even from the United States (1/10, 10%). We were able to identify a number of key findings from the health care professionals we interviewed, which were later of great importance for the design and development of the app. Most respondents used both medical staff and family members (5/10, 50%), but no other formal medical translation services such as third parties or software (9/10, 90%), when facing language barriers in their workplace (Multimedia Appendix 3). Common foreign languages encountered in medical settings in Frankfurt am Main and other selected German cities were Russian, Kurdish, Arabic, French, Turkish, Afghan, Italian, Portuguese, Eritrean, Serbian, Croatian, Bulgarian, Spanish, Greek, and Chinese.

Most respondents raised concerns about accuracy and the emotional burden on relatives, with some giving specific examples of relatives hiding the truth from patients to appease them. Some professionals used digital tools such as Google Translate (Google LLC/Alphabet Inc.), but with dissatisfaction due to the lack of confidence in the accuracy of the translation and the lack of empathy associated with using a device, particularly when making difficult diagnoses. The availability of hospital interpreter pools in larger hospitals was noted but underutilized, mostly due to complex technicalities. Notably, some doctors working in private practice stated that, in the case of non–German-speaking patients, they would offer the patient a new appointment and ask them to bring an interpreter with them.

When considering either a scheduled appointment or an on-demand call, health care professionals expressed a preference for an on-demand translation service such as Translatly (8/10, 80%; Multimedia Appendix 3). The ideal translation solution was identified as a pool of medically experienced translators, suggesting that medical staff, including medical students, would be most effective due to their knowledge of medical terminology and empathy. Estimated usage of a translation app varied, with most professionals predicting usage of at least once a week. Finally, concerns about bureaucratic and logistical hurdles, the efficiency of nonmedical translators, and the overall feasibility of integrating such a system into daily medical practice were raised, despite the enthusiasm for a digital translation platform.

Evaluation of Pilot Study

The pilot study recorded 39 translation requests over a 2-month period, of which 16 (41%) were successfully completed (Table 1).

The successfully completed translations accounted for a cumulative duration of 209 minutes across 16 distinct conversations. These translation sessions were carried out by a team of 10 translators, who contributed to facilitating communication across language barriers within the hospital setting. The breakdown of languages translated through the platform included Russian, Ukrainian, Serbian/Croatian/Bosnian, Arabic, Greek, and Urdu, reflecting a wide range of linguistic needs among the patient population.

The origin of translation requests spanned several medical departments, with the Department of Infectious Diseases submitting the highest number of requests (n=5), followed by the Emergency Department with 4 requests. Internal medicine departments submitted 3 requests, while pediatrics and psychiatry departments each submitted 2 requests for translation services.

XSL•FO
RenderX

Table 1. Overview of translation requests during the pilot phase of the implementation of the digital translation platform Translatly at the Goethe University Hospital in Frankfurt am Main (December 2022-January 2023). The table summarizes the total number of on-demand translation requests, the outcome of these requests (whether completed or unanswered), the languages translated, and the medical departments from which the requests originated. The study involved health care professionals from different departments and focused on the feasibility of the Translatly platform in a clinical setting.

Description	Values (n=39), n (%)
Translation requests	
Remained unanswered	23 (59)
Successfully completed	16 (41)
Started, but not completed	0 (0)
Translated languages (n=16)	
Russian	6 (38)
Ukrainian	3 (19)
Serbian/Croatian/Bosnian	2 (13)
Arabic	2 (13)
Greek	2 (13)
Urdu	1 (6)
Medical departments of translation requests (n=16)	
Infectious diseases	5 (31)
Emergency department	4 (25)
Internal medicine departments	3 (19)
Pediatrics	2 (13)
Psychiatry	2 (13)

Challenges of Using the Translatly Platform

Several challenges were identified during the pilot study. A key issue was translator availability, with 23 out of 39 (59%) translation requests going unanswered, mainly due to translators being unavailable at the time of the request. In addition, while feedback from both medical staff and translators was generally positive—particularly regarding the simple and intuitive design of the platform—some users expressed concern about the delayed responses. Notably, some translators indicated that they had been contacted via instant messaging beforehand, to ensure availability when a translation request was made.

Taken together, the results of this study addressed the 3 primary research questions: (1) how health care professionals overcome language barriers and their acceptance of Translatly's on-demand medical translation platform, (2) the feasibility of implementing Translatly in clinical settings to overcome language barriers, and (3) the identification of potential challenges that could inform future studies.

Discussion

Principal Findings

With this formative research study, we present the first successful implementation of the digital translation platform "Translatly" in a medical setting to tackle language barriers. Ethnographic research revealed that health care professionals currently rely on family members or online tools to overcome language barriers. There was a strong preference among health care professionals for an on-demand translation service staffed by medically experienced professionals with foreign language skills. The pilot study, conducted at the Goethe University Hospital in Frankfurt am Main, Germany, demonstrated the feasibility of using Translatly in a real-world clinical setting. Translatly's solution is to virtually enable patient-doctor communication by leveraging the language skills of health care professionals. Facilitating real-time translation services during patient consultations directly addresses the problems of miscommunication in health care by providing immediate access to translators, who in this study are either medical students or trained medical professionals. Our translator pool, which is predominantly made up of migrant health professionals, not only ensures that complex medical terminology is translated properly, but can also address cultural language differences [32], thereby reducing the likelihood of misunderstandings and subsequent medical errors.

To the best of our knowledge, our digital approach to addressing the problems of miscommunication due to language barriers in health care is unique in that it seeks to combine human language skills with the capabilities of modern eHealth technology. Currently, approaches to the problem of language barriers focus heavily on so-called machine translation (MT) software, which is specifically designed to translate text or speech from one language to another without human intervention [33,34]. In fact, medical translation apps have been published or are already available in app stores. However, they either rely on MT or contain predefined health phrases [35]. The modern focus on automated translation alternatives is mainly motivated by the

XSL•FO
RenderX

fact that human translators can be costly and time-consuming [36]. This is evidenced by studies that have shown, for example, that the use of human interpreters has decreased in several health centers in the United States from 2009 to 2019 [37]. Therefore, the Translatly platform represents a novel solution that does not require translators to be tied to a specific hospital, and moreover, the intensive search for a specific translator can be quickly overcome by accessing the translator pool, which can be done by medical staff directly from the patient's side of the bed.

One of the key findings from the responses of the medical staff interviewed using our ethnographic research questionnaires was that the ideal solution proposed by the majority of interviewees would be to actually use medical staff or medical students with foreign language skills, and this was a key influence in prompting our team to look for volunteer translators to join our translator pool. However, we acknowledge that this could be a major limitation of our pilot study as the volunteer translators we recruited are not specialist interpreters but rather medical practitioners or students. Therefore, the issue of variability in the medical knowledge and language skills of the translators recruited for this study is a critical concern that could affect the quality and accuracy of translations in clinical settings. Regarding the foreign language skills of volunteer translators with a migrant background, it has been estimated that in 2023 in Germany, approximately 64.5% of people with a migrant background were first-generation migrants, and therefore native speakers of their own language [38]. Native speakers do not normally need language certificates, as they are assumed to have an innate and comprehensive understanding. A native speaker is defined as an individual who typically acquires the language as a young child and maintains it into adulthood [39]. Notably, there are linguists who distinguish between the basic and extended proficiency of native speakers [40]. With regard to the German language skills of translators from migrant backgrounds, there are data showing that in Germany 66% of migrants consider themselves to have a good or professional level of German [41] and strong language skills are required for enrollment in medical faculties. Although these statistics do not resolve the issue of variability in the medical translation skills of the recruited translators, they do support the fact that translation errors via the Translatly platform may indeed be minimal. Contrary to our findings from the ethnographic research questionnaires, other similar interviews showed that hospital staff were strongly opposed to the use of bilingual medical staff services [42]. It should be noted, however, that these interviews were conducted to assess language services for written materials and not for direct dialogue translation. By contrast, we strongly believe that while the Translatly concept does not provide an overarching solution, the use of bilingual medical staff for the Translatly platform represents a significant improvement over the already precarious situation in daily clinical routine, as evidenced by our questionnaire results and others [43], of relying on family members or untrained interpreters. Furthermore, even when official translators are present, there is still potential for mistranslation [44], making other translation alternatives necessary in some cases to improve communication with patients.

To ensure that the Translatly digital platform is user-centric, the needs of the medical staff that emerged from the questionnaires were carefully considered during the design of the operating mode and later during the development of the app. However, as the call is always initiated by the doctor in the presence of the patient, it may be unannounced and unexpected for the translator. We have therefore taken into account the possibility of some translators being unavailable during ad hoc calls and have specifically designed the operating mode and system architecture so that all translators in the pool will receive the calls if they are listed for the selected language to be interpreted. The integration of the Translatly app with Firebase [45], a widely used service for sending push notifications to a large number of users quickly and reliably, helps to facilitate the calling of all required translators from the pool. Moreover, real-time video telephony in clinical settings can be disrupted by connectivity problems due to network instability. Solutions to this issue could include switching to an audio call, which requires less data and can tolerate latency, or reconnecting with another translator via push notifications provided by the robust backend architecture with Firebase. However, these measures may not fully prevent disruptions, so offline text-based translation or rescheduling during outages would be helpful. Further time-saving is achieved through Translatly's simple, user-friendly app design. Knowing that effective and prompt communication is paramount in the medical setting, we focused on designing our app for the pilot study with simple accessibility, such as few visual elements and features, intuitive navigation, legible fonts, and a few color palettes. As our pilot study aimed to test Translatly's usability in a real-world medical setting, we had not yet included other app features such as scheduling, an integrated calendar, or chat services.

The results of the pilot study demonstrated a significant need for effective translation services in the medical setting, as evidenced by the Translatly app being used for a total of almost 3.5 hours. The data showed that over 40% (16/39) of translation requests were successfully fulfilled, highlighting the potential of the app to improve communication between health care providers and non–German-speaking patients. Another notable limitation observed during the pilot study was that 23 out of 39 (59%) translation requests went unanswered. This highlights the challenges associated with translator availability and responsiveness, particularly at peak times, which could affect the scalability of such a digital solution across larger or more diverse linguistic landscapes. As a consequence, our pilot study led us to add an optional scheduling feature and calendar to the app. To further mitigate availability issues, we are exploring the possibility of allowing translators to indicate their availability in advance. This system would prioritize periods of high demand and ensure that enough translators are on call when they are most needed. Another option would be a real-time monitoring system that tracks translator availability and dynamically adjusts call routing to match supply with demand. This would help to minimize delays and ensure that translation services are readily available when they are needed. Finally, expanding the pool of translators to include the official translators employed by hospitals and incorporating a translator selection algorithm to improve response rates or using MT-based translation options would ensure broader and more professional

XSL•FO
RenderX

coverage. Nonetheless, the diversity of languages requested, from Russian to Urdu, across a range of departments from infectious diseases to psychiatry, underlines the widespread need for effective translation services in medical settings and the successful initial broad implementation of Translatly in a variety of medical departments for patients from diverse linguistic backgrounds.

While Translatly offers many benefits, there are further inherent challenges and limitations that need to be considered. First, to address the issue of variability in the professional translation of medical conversations, more rigorous selection criteria and training protocols for translators are needed. These protocols should include targeted medical terminology training focusing on the most common conditions and procedures in the hospital departments where translation services are critical. In addition, the inclusion of role-playing scenarios with simulated doctor-patient interactions could be implemented to allow students to practice translation in a controlled environment and receive feedback on real-life challenges. Further cultural competency training would also be beneficial, with modules designed to deepen understanding of cultural nuances that may impact on communication, ensuring that translations are both linguistically accurate and culturally sensitive. Next, relying on health care professionals to perform dual roles as medical providers and translators could potentially lead to work overload and burnout. Effective workload management is essential to prevent fatigue and ensure the sustainability of translation services [46]. Therefore, we will consider assessing workload reduction in future studies on the implementation of Translatly.

Collecting robust data on user usage and satisfaction, as well as the impact on treatment outcomes will provide insights that could guide future improvements and adaptations to the platform. In addition, expanding the range of languages offered and scaling the solution to other regions and health care settings could increase its impact, making inclusive health care more accessible nationally and potentially globally. Next, Translatly's mode of operation does not address the issue of face-to-face interpreting by existing hospital interpreters. While official hospital interpreters can easily be added to Translatly's pool of translators, we continue to advocate for video interpreting, as recently a systemic review has shown that patient satisfaction with video interpreting is the same as with in-person communication [47].

Conclusions

This formative pilot study assessed the feasibility of overcoming language barriers in clinical settings through the use of the Translatly digital translation platform. The study aimed to determine how health care professionals overcome language barriers, whether they prefer and accept this on-demand translation solution, and whether it can be effectively integrated into routine clinical workflows. The platform leverages the language skills of medical students and health care professionals, enabling them to serve as on-demand volunteer translators during medical consultations via video telephony. The results showed that Translatly has significant potential to improve communication between health care providers and nonnative speaking patients, thereby improving patient outcomes and optimizing health care resources. Although this early-stage study focused primarily on feasibility, it also identified key challenges, such as translator availability, that can inform future refinements to the platform. The study did not evaluate the quality of translations, nor did it focus on the platform's architecture or mechanisms for attracting users. Rather, it laid the groundwork for future work by exploring the operational feasibility of Translatly in a real-world health care setting.

Translatly's alignment with statutory health insurance mandates in Germany positions it as a potential eHealth solution for widespread adoption [48]. By reducing language barriers, it can support better health outcomes and promote health education, particularly in diverse patient populations. A subscription-based model could address the platform's cost implications, allowing health care facilities to scale the service according to their needs. Initial costs related to technology infrastructure, training, and ongoing maintenance would be manageable, particularly in larger hospitals with robust IT systems, and would be justified by the long-term benefits of the platform, including improved workflow, improved patient outcomes, and reduced overall health care costs. Projected revenue growth of €16,852 (US $18,334) in the first year of implementation, rising to €5,933,455 (US $6,455,302) in the fourth year, supported by strategic expansion and scaling efforts, further underscores the platform's potential for sustainable financial viability within the health care sector. With further development and strategic implementation, platforms such as Translatly could become an integral part of health care systems, improving communication, patient satisfaction, and reducing health care costs.

Ultimately, Translatly has the potential to set a new standard for overcoming language barriers in health care. By enabling health care professionals to communicate effectively with nonnative speakers, the platform strengthens trust between patients and providers, reduces the emotional burden on families, and supports the delivery of high-quality, culturally competent care.

Acknowledgments

JL and VO received a grant from the Ministry for Sciences, Research and Arts, Baden-Wuerttemberg, supporting the Clinical Cooperation Unit Digital Spaces in Inflammatory Diseases at the Centre for Preventive Medicine and Digital Health, Medical Faculty of Mannheim, Heidelberg University. We thank all the registered interpreters for their valuable assistance.

Data Availability

The anonymized data supporting the findings of this study are included as Multimedia Appendix 4.

XSL•FO
RenderX

Conflicts of Interest

MB was the project manager for app design and development and TBD was responsible for software architecture and development. Both MB and TBD are working for CodeCooperation GmbH and were financially enlisted by Translatly UG. HvW and OB are owners of Translatly UG, registered at the address Frankfurter Strasse 134, 63067 Offenbach, Germany. SZ has provided consultancy and participated in speakers' bureaus for Abbvie. All other authors have no conflicts of interest.

Multimedia Appendix 1
Ethnographic research questionnaire used for the feasibility study, translated into English.
[PDF File (Adobe PDF File), 131 KB - formative_v9i1e63095_app1.pdf]

Multimedia Appendix 2
Examples of screenshots of the Translatly app user interface for medical professionals on a tablet device. (A) The log-in screen with header and logo where medical staff enter their credentials. (B) The main app screen for medical staff, showing an on-demand translation tab with a drop-down menu for language selection. (C) After selecting the language to be translated, the medical staff user can choose between audio or video call mode to initiate a translation session. (D) Screen showing the process of calling available translators once a translation request has been made. (E) Notification screen informing medical staff when a translator has accepted the call, with options to continue or start again. (F) An illustrative example of videoconferencing with a translator during medical consultation.
[PNG File , 1556 KB - formative_v9i1e63095_app2.png]

Multimedia Appendix 3
Clinical profile and characteristics of health care professionals interviewed for the feasibility assessment of the Translatly digital translation platform, conducted through an ethnographic research questionnaire. This appendix provides a comprehensive statistical summary of respondents' medical specialties, clinic types, and work locations. It also details their current practices for managing language barriers in clinical settings, including the types of translation services they use and their preferred mode of access to online translation options (on-demand or appointment-based).
[PDF File (Adobe PDF File), 97 KB - formative_v9i1e63095_app3.pdf]

Multimedia Appendix 4
Anonymized data set of translation requests from the Translatly pilot phase at the University Hospital of the Goethe University in Frankfurt am Main, Germany, covering the period from December 1, 2022, to January 31, 2023. The data include the total number of translation requests (n=39), the number of completed translations (n=16), and the number of unanswered requests (n=23). The appendix also provides detailed information on the duration of the translations, the languages translated, and the medical departments from which the requests originated. This data set is fully anonymized.
[PDF File (Adobe PDF File), 74 KB - formative_v9i1e63095_app4.pdf]

References

1. Slade S, Sergent S. Language Barrier. Treasure Island, FL: StatPearls; Apr 10, 2023:1.
2. Divi C, Koss RG, Schmaltz SP, Loeb JM. Language proficiency and adverse events in US hospitals: a pilot study. Int J Qual Health Care 2007 Apr;19(2):60-67. [doi: 10.1093/intqhc/mzl069] [Medline: 17277013]
3. Cohen AL, Rivara F, Marcuse EK, McPhillips H, Davis R. Are language barriers associated with serious medical events in hospitalized pediatric patients? Pediatrics 2005 Sep;116(3):575-579. [doi: 10.1542/peds.2005-0521] [Medline: 16140695]
4. Mustafa R, Mahboob U, Khan RA, Anjum A. Impact of language barriers in doctor - patient relationship: a qualitative study. Pak J Med Sci 2023;39(1):41-45 [FREE Full text] [doi: 10.12669/pjms.39.1.5805] [Medline: 36694787]
5. Wune G, Ayalew Y, Hailu A, Gebretensaye T. Nurses to patients communication and barriers perceived by nurses at Tikur Anbessa Specilized Hospital, Addis Ababa, Ethiopia 2018. International Journal of Africa Nursing Sciences 2020;12:100197 [FREE Full text] [doi: 10.1016/j.ijans.2020.100197]
6. Al Shamsi H, Almutairi AG, Al Mashrafi S, Al Kalbani T. Implications of language barriers for healthcare: a systematic review. Oman Med J 2020 Mar;35(2):e122 [FREE Full text] [doi: 10.5001/omj.2020.40] [Medline: 32411417]
7. Diamond LC, Schenker Y, Curry L, Bradley EH, Fernandez A. Getting by: underuse of interpreters by resident physicians. J Gen Intern Med 2009 Feb;24(2):256-262 [FREE Full text] [doi: 10.1007/s11606-008-0875-7] [Medline: 19089503]
8. Dsouza C. The power of words: how language impacts medical communication. medium. URL: https://medium.com/@kefitup/the-power-of-words-how-language-impacts-medical-communication-aa1ff2947fee [accessed 2023-10-27]
9. Hurtig RR, Alper RM, Berkowitz B. The cost of not addressing the communication barriers faced by hospitalized patients. Perspect ASHA Spec Interest Groups 2018 Jan;3(12):99-112 [FREE Full text] [doi: 10.1044/persp3.SIG12.99] [Medline: 30854467]

XSL•FO
RenderX

10. Bartig S, Kalkum D, Le HM, Lewicki A. Diskriminierungsrisiken und diskriminierungsschutz im gesundheitswesen – wissensstand und forschungsbedarf für die antidiskriminierungsforschung. Deutsches Zentrum fu¨r Integrations- und Migrationsforschung (DeZIM). 2021 Jun 01. URL: https://tinyurl.com/52pa5djj [accessed 2024-11-11]
11. Kommission E. Eurobarometer spezial 243: Die Europäer und ihre sprachen. Europäische Kommission. 2021. URL: https://www.bpb.de/system/files/pdf/1ZSNUZ.pdf [accessed 2023-10-08]
12. Reiff E, Gade C, Böhlich S. Handling the shortage of nurses in Germany: opportunities and challenges of recruiting nursing staff from abroad (IUBH Discussion Papers - Human Resources No. 3/2020). EconStar. 2022 Jul 01. URL: https://www.econstor.eu/handle/10419/222921 [accessed 2024-11-11]
13. Angenendt S, Knapp N, Kipp D. Germany is looking for foreign labour: how to make recruitment development-orientated, sustainable and fair (SWP Research Paper 3). Stiftung Wissenschaft und Politik/German Institute for International and Security Affairs. 2023 Mar 01. URL: https://www.swp-berlin.org/publications/products/research_papers/2023RP03_LookingforForeignLabour.pdf [accessed 2024-11-11]
14. Ariste R, di Matteo L. Non-official language concordance in urban Canadian medical practice: implications for care during the COVID-19 pandemic. Healthc Policy 2021 May;16(4):84-96 [FREE Full text] [doi: 10.12927/hcpol.2021.26497] [Medline: 34129480]
15. Ahmed S, Shommu NS, Rumana N, Barron GRS, Wicklum S, Turin TC. Barriers to access of primary healthcare by immigrant populations in Canada: a literature review. J Immigr Minor Health 2016 Dec;18(6):1522-1540. [doi: 10.1007/s10903-015-0276-z] [Medline: 26364053]
16. Bowen S. Language Barriers in Access to Health Care. Ottawa, ON, Canada: Health Canada; Nov 01, 2001:1-134.
17. Espinoza J, Derrington S. How should clinicians respond to language barriers that exacerbate health inequity? AMA J Ethics 2021 Feb 01;23(2):E109-E116 [FREE Full text] [doi: 10.1001/amajethics.2021.109] [Medline: 33635190]
18. Saito S, Harris MF, Long KM, Lewis V, Casey S, Hogg W, et al. Response to language barriers with patients from refugee background in general practice in Australia: findings from the OPTIMISE study. BMC Health Serv Res 2021 Sep 06;21(1):921 [FREE Full text] [doi: 10.1186/s12913-021-06884-5] [Medline: 34488719]
19. Statistisches Bundesamt. Gesundheitsausgaben im Jahr 2019 auf über 400 milliarden Euro gestiegen. Statistisches Bundesamt (Destatis). 2023. URL: https://www.destatis.de/DE/Presse/Pressemitteilungen/2021/04/PD21_167_236.html [accessed 2024-11-09]
20. Aitken G. Medical students as certified interpreters. AMA J Ethics 2019 Mar 01;21(3):E232-E238 [FREE Full text] [doi: 10.1001/amajethics.2019.232] [Medline: 30893036]
21. Bundesministerium für Gesundheit. Innovationsfonds. Bundesministerium für Gesundheit. 2023. URL: https://www.bundesgesundheitsministerium.de/service/begriffe-von-a-z/i/innovationsfonds [accessed 2023-11-06]
22. Stošović S, Stefanović D, Bogdanović M, Vukotić N. The use of the flutter framework in the development process of hybrid mobile applications. kij 2022 Sep 30;54(3):477-483. [doi: 10.35120/kij5403477s]
23. DigitalOcean. DigitalOcean® cloud computing. DigitalOcean. URL: https://www.digitalocean.com/ [accessed 2023-11-08]
24. IBM. What is a REST API? IBM. 2023. URL: https://tinyurl.com/yz5pe45h [accessed 2023-11-08]
25. Solomon S. What is OAuth? Definition, how it works, and tips for success. Frontegg. 2023 Jul 10. URL: https://frontegg.com/blog/oauth [accessed 2024-02-01]
26. AngularJS. AngularJS developer guide. AngularJS. 2022. URL: https://docs.angularjs.org/guide [accessed 2023-11-08]
27. ReactJS. React - the library for web and native user interfaces. ReactJS. 2024. URL: https://react.dev/ [accessed 2024-04-06]
28. Oracle. MySQL 8.2 reference manual. Oracle. 2023. URL: https://dev.mysql.com/doc/refman/8.2/en/ [accessed 2023-08-11]
29. Firebase. Firebase/Google. 2023. URL: https://firebase.google.com/ [accessed 2023-11-08]
30. Twilio. Twilio Inc. 2024. URL: https://www.twilio.com/en-us [accessed 2024-06-04]
31. Translatly. Translatly UG. 2024 Oct 01. URL: https://translatly.de/ [accessed 2024-11-04]
32. Bader M, Jobst LJ, Zettler I, Hilbig BE, Moshagen M. Disentangling the effects of culture and language on measurement noninvariance in cross-cultural research: The culture, comprehension, and translation bias (CCT) procedure. Psychol Assess 2021 May;33(5):375-384. [doi: 10.1037/pas0000989] [Medline: 33734754]
33. Noll R, Frischen LS, Boeker M, Storf H, Schaaf J. Machine translation of standardised medical terminology using natural language processing: a scoping review. N Biotechnol 2023 Nov 25;77:120-129 [FREE Full text] [doi: 10.1016/j.nbt.2023.08.004] [Medline: 37652265]
34. Urlaub P, Dessein E. Machine translation and foreign language education. Front Artif Intell 2022;5:936111 [FREE Full text] [doi: 10.3389/frai.2022.936111] [Medline: 35937139]
35. Panayiotou A, Gardner A, Williams S, Zucchi E, Mascitti-Meuter M, Goh AM, et al. Language translation apps in health care settings: expert opinion. JMIR Mhealth Uhealth 2019 Apr 09;7(4):e11316 [FREE Full text] [doi: 10.2196/11316] [Medline: 30964446]
36. Herrmann-Werner A, Loda T, Zipfel S, Holderried M, Holderried F, Erschens R. Evaluation of a language translation app in an undergraduate medical communication course: proof-of-concept and usability study. JMIR Mhealth Uhealth 2021 Dec 02;9(12):e31559 [FREE Full text] [doi: 10.2196/31559] [Medline: 34860678]
37. Regenstein M, Park YH, Krips M. The use of interpreters in health centers: a mixed-methods analysis. J Health Care Poor Underserved 2023;34(1):224-245. [doi: 10.1353/hpu.2023.0015] [Medline: 37464491]

XSL•FO
RenderX

38. Statistisches Bundesamt. Bevölkerung mit migrationshintergrund. Bundeszentrale für politische Bildung. 2023. URL: https://www.bpb.de/kurz-knapp/zahlen-und-fakten/soziale-situation-in-deutschland/61646/bevoelkerung-mit-migrationshintergrund/ [accessed 2024-08-17]
39. Lee J. The native speaker: an achievable model? Asian EFL Journal 2005 Jun 01;7(2):152-163 [FREE Full text]
40. Hulstijn J. In: Deusen-Scholl NSNV, editor. Language Proficiency in Native and Non-Native Speakers. Amsterdam, The Netherlands: John Benjamins Publishing Company; 2015.
41. Gazzola M. Language skills and employment status of adult migrants in Europe. In: Jean-Claude B, Hans-Jürgen K, David K, Philia T, editors. The linguistic integration of adult migrants/L'intégration linguistique des migrants adultes. Berlin, Germany: De Gruyter Mouton; 2017:297-302.
42. Regenstein M, Andres E. Hospital language service programs: a closer look at translation practices. J Health Care Poor Underserved 2014 Nov;25(4):2003-2018. [doi: 10.1353/hpu.2014.0162] [Medline: 25418255]
43. Kuo DZ, O'Connor KG, Flores G, Minkovitz CS. Pediatricians' use of language services for families with limited English proficiency. Pediatrics 2007 Apr;119(4):e920-e927. [doi: 10.1542/peds.2006-1508] [Medline: 17371933]
44. Bravo M, Lim R, Poonai N, Chen B. Clinical suspicion and language translation in the pediatric emergency department. Pediatr Emerg Care 2021 May 01;37(5):e272-e274. [doi: 10.1097/PEC.0000000000001606] [Medline: 30130342]
45. Di Filippo M, Avellone A, Belingheri M, Paladino ME, Riva MA, Zambon A, et al. A mobile app leveraging citizenship engagement to perform anonymized longitudinal studies in the context of COVID-19 adverse drug reaction monitoring: development and usability study. JMIR Hum Factors 2022 Nov 04;9(4):e38701 [FREE Full text] [doi: 10.2196/38701] [Medline: 35930561]
46. Montgomery A, Panagopoulou E, Esmail A, Richards T, Maslach C. Burnout in healthcare: the case for organisational change. BMJ 2019 Jul 30;366:l4774. [doi: 10.1136/bmj.l4774] [Medline: 31362957]
47. Joseph C, Garruba M, Melder A. Patient satisfaction of telephone or video interpreter services compared with in-person services: a systematic review. Aust Health Rev 2018 Apr;42(2):168-177. [doi: 10.1071/AH16195] [Medline: 30021688]
48. Gerke S, Stern AD, Minssen T. Germany's digital health reforms in the COVID-19 era: lessons and opportunities for other countries. NPJ Digit Med 2020;3:94 [FREE Full text] [doi: 10.1038/s41746-020-0306-7] [Medline: 32685700]

Abbreviations

AngularJS: Angular Java Script
GDPR: General Data Protection Regulation
MT: machine translation
OAuth: Open Authorization
ReactJS: React Java Script
REST API: Representational State Transfer Application Programming Interface

Edited by A Mavragani; submitted 14.06.24; peer-reviewed by A Islam, K Tekwa; comments to author 19.08.24; revised version received 24.08.24; accepted 16.10.24; published 14.03.25.

Please cite as:
Olsavszky V, Bazari M, Dai TB, Olsavszky A, Finkelmeier F, Friedrich-Rust M, Zeuzem S, Herrmann E, Leipe J, Michael FA, Westernhagen HV, Ballo O
Digital Translation Platform (Translatly) to Overcome Communication Barriers in Clinical Care: Pilot Study
JMIR Form Res 2025;9:e63095
URL: https://formative.jmir.org/2025/1/e63095
doi:10.2196/63095
PMID:

©Victor Olsavszky, Mutaz Bazari, Taieb Ben Dai, Ana Olsavszky, Fabian Finkelmeier, Mireen Friedrich-Rust, Stefan Zeuzem, Eva Herrmann, Jan Leipe, Florian Alexander Michael, Hans von Westernhagen, Olivier Ballo. Originally published in JMIR Formative Research (https://formative.jmir.org), 14.03.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Multilevel Intervention to Increase Patient Portal Use in Adults With Type 2 Diabetes Who Access Health Care at Community Health Centers: Single Arm, Pre-Post Pilot Study

Robin Whittemore[1*], PhD; Sangchoon Jeon[1*], PhD; Samuel Akyirem[1], RN, MRES; Helen N C Chen[2], PhD; Joanna Lipson[1], MS, MPH; Maritza Minchala[3], MPH; Julie Wagner[4], PhD

[1]School of Nursing, Yale University, Orange, CT, United States

[2]Department of Epidemiology, Richard M. Fairbanks School of Public Health, Indiana University, Indianapolis, IN, United States

[3]School of Public Health, Yale University, New Haven, CT, United States

[4]Department of Behavioral Sciences and Community Health, University of Connecticut School of Dental Medicine, Farmington, CT, United States

[*]these authors contributed equally

Corresponding Author:
Robin Whittemore, PhD
School of Nursing
Yale University
400 West Campus Drive
Orange, CT, 06477
United States
Phone: 1 860 420 9623
Email: robin.whittemore@yale.edu

Abstract

Background: Diabetes self-management education and support (DSMS) delivered via patient portals significantly improves glycemic control. Yet, disparities in patient portal use persist. Community health centers (CHCs) deliver care to anyone who needs it, regardless of income or insurance status.

Objective: This study aimed to evaluate the feasibility, acceptability, and preliminary efficacy of a multilevel intervention to increase access and use of portals (MAP) among people with type 2 diabetes (T2D) receiving health care at CHCs.

Methods: A within-subjects, pre-post design was used. Adults with T2D who were portal naive were recruited from 2 CHCs. After informed consent, participants met with a community health worker for referrals for social determinants of health, provision of a tablet with cell service, and individualized training on use of the tablet and portal. Next, a nurse met individually with participants to develop a DSMS plan and then communicated with patients via the portal at least twice weekly during the first 3 months and weekly for the latter 3 months. Data were collected at baseline, 3 months and 6 months. The primary outcome was patient activation and engagement with the portal. Secondary outcomes included technology attitudes, digital health literacy, health-related outcomes and psychosocial function.

Results: In total, 26 patients were eligible, 23 received the intervention, and one was lost to follow up. The sample was predominately Latino or Hispanic (17/22, 77%) and reported low income (19/22, 86%< US $40,000/year), low education (13/22, 59% <high school), and no health insurance (12/22, 55%). All participants had access to a Smartphone, but 91% (20/22) had never accessed a health app. The baseline hemoglobin A_{1c} level was 8.31%. Portal activation was high; 100% (22/22) of participants created a portal account and logged in within the first month. Mean participant logins per week over the first 3 months was 3.16 (SD 1.55) and 1.45 (SD 0.93) over the final 3 months; mean logins per month over the first 3 months was 12.65 (SD 6.21) and 5.79 (SD 3.74) over last 3 months. Engagement was high; 96% (20/21) logged in at least twice per month in the first 3 months and 76% (16/21) between 3 and 6 months. At 6 months, improvements were seen in technology confidence, digital health literacy, diabetes self-efficacy, and diabetes distress. Participant satisfaction with MAP was high as was intention to continue portal use. Barriers to clinical integration and recommendations for portal development were identified.

Conclusions: MAP shows promise for improving health equity in portal use for T2D. Larger, controlled studies are needed to determine how best to implement MAP in complex clinical settings and to evaluate efficacy over time.

Trial Registration: ClinicalTrials.gov NCT05180721; https://clinicaltrials.gov/study/NCT05180721

XSL•FO
RenderX

(JMIR Form Res 2025;9:e67293) doi:10.2196/67293

KEYWORDS

patient portal; mobile phone; diabetes; community health center; adults; diabetic; DM; diabetes mellitus; Type 2 diabetes; T2D; community health centers; CHCs; pilot study; feasibility; self-management; glycemic control; patient portals; social determinants of health; primary outcome; digital health; digital health literacy; health technology; health technologies; psychosocial; efficacy

Introduction

There are significant racial and ethnic disparities in the prevalence, morbidity, and mortality of type 2 diabetes (T2D). Despite medical advances and increased access to medical care, these disparities persist. Racial or ethnic minorities are more likely to have poor glycemic control [1], develop diabetes-related complications [2], and are 1.5-2.3 times more likely to die from diabetes than White individuals [3]. Further, racial and ethnic minorities were particularly affected by COVID-19, with increased risk for infection, morbidity, hospitalization, and mortality [4]. Preexisting conditions, including and especially diabetes, increase the risk for poor COVID-19 outcomes [5,6]. Thus, innovative approaches are urgently needed to address health inequities in T2D.

Patient portals provide secure web-based access to medical records with the capability of messaging providers, filling prescriptions, viewing educational materials, and accessing clinic services. Increased patient portal use has the potential to increase engagement with care and improve diabetes health outcomes. In the general population, patient portal use has been shown to increase office visits while decreasing emergency room visits and hospitalizations [7]. Patient portal use also increases patient knowledge, self-efficacy, decision-making, medication use, and preventive screening [8]. In adults with diabetes, greater portal use of secure messaging with providers led to improved glycemic control (measured based on hemoglobin A_{1c} [HbA_{1c}] levels) compared to nonusers [9-12]. Other trials have demonstrated significant reductions in HbA_{1c} through diabetes self-management education or support (DSMS) via the portal [13-15]. Thus, fostering communication with providers and diabetes self-management support are promising features of portals for adults with diabetes.

There is considerable evidence documenting disparities in patient portal use. Adults who are older, Black, Latinx, and those with low socioeconomic status and low health literacy are less likely to use patient portals as an adjunct to clinical care [16,17]. Despite increased access and popularity of patient portal use in the United States with over 92% of health care organizations offering portals in 2015 [18,19], disparities in portal use continue. More than 100 studies have demonstrated substantial disparities in portal use [17]. Concern has been raised that this well-intentioned solution for patient-centered care may actually worsen health inequities unless portal use among the underserved is increased [17,20,21]. Portal adoption interventions can successfully increase portal use. In a systematic review of interventions to increase portal use in vulnerable populations, 67% (12/18) of studies showed a significant increase in portal use, predictors of use, or reduced disparities. Free or low-cost internet access, technical training and assistance, and proactive outreach from the health care team through the portal were reported to have the strongest evidence for improving health equities in portal use and outcomes [21]. Technical training was the most effective strategy in improving patient portal logins, use of features, and secure messaging [17]. To date, no interventions on portal adoption have targeted adults with T2D of diverse races or ethnicities with limited resources, who have unique structural and social barriers to portal access, use, and diabetes self-management.

Community health centers (CHCs) play a critical role in addressing health inequities in T2D, providing care to over 27 million people in the United States [22]. The aim of CHCs is to provide affordable, high-quality, comprehensive primary care to medically underserved populations, regardless of insurance status or ability to pay for care. Most CHC patients (92%) who live in poverty or near poverty, as defined by the Federal Poverty Level, are disproportionately from racial or ethnic minority groups (total 63%: 32% Hispanic, 22% Black, 9% other minorities), and have high rates of chronic conditions compared to the general population. In 2018, 21% of adults seen at CHCs had T2D compared to 11% in the general population [23].

Social determinants of health (SDoH) are important considerations in developing interventions for adults with T2D who access care at CHCs. The SDoH Equity framework [24], developed by the World Health Organization (WHO), posits that structural determinants of health operate through intermediary and SDoH to shape health outcomes. Intermediary and social determinants include material circumstances (eg, access to tablets or the internet and food security), behavioral, and biological factors (eg, taking medications), psychosocial factors (eg, technology literacy), and the health care system (access to care). To increase patient portal use in adults with T2D accessing care at CHCs, our intervention goals were (1) to address the intermediary and SDoH by providing tablets, home internet, and technology support and (2) to improve engagement with health care by personalizing care and DSMS through known community health workers and nurses employed at the CHC.

The purpose of this study was to pilot-test a multilevel intervention to increase access and use of portals (MAP). We sought to determine MAP's feasibility, acceptability, and potential to improve outcomes over 6 months among adults with T2D who access health care at 2 CHCs. Outcomes included portal-related outcomes (portal activation [logins during the first month], portal engagement [logins over 6 months], digital health literacy, technology acceptability), health-related outcomes (HbA_{1c}, T2D self-management [medication, blood glucose monitoring, healthy eating, and physical activity], and psychosocial function [diabetes self-efficacy, autonomy support, and diabetes distress]).

XSL•FO
RenderX

Methods

Overview

A within-subjects, pre-post design was used to pilot MAP in 22 adults with T2D who were portal naïve. We developed the intervention protocol after seeking feedback from stakeholders on barriers and facilitators to patient portal use and logistics to consider optimizing the implementation of MAP in CHCs [25].

Recruitment

Participants were recruited from 2 CHCs which have been previously described [25]. The clinics are located in Connecticut which is a small and densely populated state with prominent health disparities. In Connecticut, Black residents are nearly 4 times more likely than white residents to have a diabetes-related lower extremity amputation, and among Latino individuals, the rate is nearly 3 times higher than for non-Latino White individuals [26]. Inclusion criteria for this study were as follows: (1) established patient at 1 of the 2 CHCs, (2) age 21-65 years, (3) diagnosed with T2D >6 months, (4) most recent HbA_{1c} measure >7.5%, (5) no use of the patient portal in the past year, (6) no intention of moving or changing clinic within 6 months, and (7) self-reported ability to read in English or Spanish. Exclusion criteria included cognitive impairment (≥3 incorrect answers on the Six Item Screener) [27] or current gestational diabetes.

Participants were recruited from select primary care provider panels at each clinic. A designated clinic staff member reviewed the weekly schedule for potentially eligible recruits and introduced the study to recruits in order to determine preliminary interest. If interested, a trained research assistant explained the study and determined eligibility with a screening questionnaire. If eligible, an appointment was scheduled in person or via Zoom (Zoom Video Communications) for informed consent, baseline data collection, and study enrollment. Written informed consent and data collection were completed in the language preference of the participant (English or Spanish).

Upon completion of baseline data collection, participants were scheduled for the first intervention session. All participants received standard T2D care at the CHC, which followed the guidelines for T2D management as recommended by the American Diabetes Association [28] (eg, quarterly appointments with primary care providers, medical management, and referrals to specialists as indicated). At the study clinics, trained nurses provided diabetes education individually during clinic appointments, as needed. All participants received the MAP intervention from community health workers (CHWs) and nurses employed at the clinics. Clinics were compensated for the salary of the nurses and CHWs for training, delivering the intervention, and completing study tasks.

Training and Supervising Interventionists

Before delivering MAP to study participants, CHWs and nurses were provided a 1-day, in-person, interactive training on the study protocol. They were also provided supportive supervision throughout the study (weekly to biweekly). Training covered orientation to research and goals of the study; human subjects protection; protocol and documentation; and team roles, responsibilities, and supervision. Training also covered details of the intervention content through a study manual and specific strategies for working with low-literacy or low-numeracy individuals with diabetes [29]. Such strategies include using the teach-back method; asking open-ended questions; avoiding unclear statements ("your test was positive"); keeping sessions brief; presenting small chunks of information; encouraging practice between sessions; using nonnumerical measures (eg, 1 serving butter=size of the tip of the thumb); using pictures when possible; using plain language (no jargon or acronyms); reducing the reading level of written materials; and, using a friendly tone. In addition, they were trained in principles of autonomy support. Autonomy support refers to a patient's perception that their health care provider recognizes the person's personal agency, encourages self-efficacy, and supports their self-care choices [30].

Intervention

The MAP intervention intentionally and directly intervened on the 4 intermediary determinants of disparities as outlined by the WHO Health Equity framework (material circumstances, psychosocial factors, behavioral and biological factors, and the health care system (Multimedia Appendix 1). All study participants received a tablet (which they were allowed to keep at the end of the study) along with internet access for the 6 months of the study.

The intervention was sequenced to first have CHWs assess SDoH needs using a questionnaire specific to each clinic (each clinic had a slightly different form already in use) and connect the participant to relevant community resources (eg, SNAP benefits). CHWs then provided training on how to gain access to the patient portal and on to mastery of the tablet and portal functionality. In addition to lack of training and lack of encouragement to use the portal, other barriers to patient use of portals addressed by CHWs included doubt about portal usefulness, lost passwords, anxiety about viewing medical information, and privacy concerns [31]. Once the technology training and social determinant referrals were completed, participants were referred to the clinic nurse (Multimedia Appendix 2).

Next, the clinic nurse contacted the participant via the portal to provide DSMS. According to the American Diabetes Association, DSMS is "the support that is required for implementing and sustaining coping skills and behaviors needed to self-manage on an ongoing basis" [32,33]. The nurse initially met individually with study participants to establish rapport, assess diabetes self-management (DSM) behaviors, and develop a DSMS plan collaboratively with the participant. Nurses were asked to communicate with patients via the portal at least twice weekly during the first 3 months and weekly for the latter 3 months, individualizing interactions based on participant needs. In these interactions, nurses assessed challenges and successes with the DSM plan of care and provided education, support, and encouragement. Nurses were also provided with a variety of electronic health education resources that could be sent to participants via the portal. Resources were assembled by the researchers from a thorough review of written and video

XSL•FO
RenderX

materials available in English and Spanish for adults with low health literacy.

Characteristics of each CHC required that MAP be integrated into clinic operations in a tailored fashion. First, one CHC did not employ CHWs and their care coordinators (who usually addressed patient social needs) were not available to deliver the MAP intervention. Therefore, at that clinic, the study nurse was trained to complete both CHW and nursing components of MAP. Second, the 2 clinics used different portal platforms. Therefore, training procedures and written materials were designed to be equivalent between, but tailored to, each clinic and its platform.

Data Collection

Data were collected and managed using REDCap (Research Electronic Data Capture; Vanderbilt University) an National Institutes of Health supported, Food and Drug Administration compliant electronic data capture application for data collection and storage. Data were collected from participants at baseline, 3 months, and 6 months including point-of-care HbA_{1c} values. Research assistants entered data at the time of data collection into the database via tablets or a computer. All participants received a gift card after each wave of data collection—US $40 at baseline, US $40 at 3 months, and US $60 at 6 months. Nurses at the clinics extracted clinical data and information technology specialists at the clinics extracted patient portal data from the electronic health record.

Measures

Overview

Demographics were collected at baseline with questions on age, sex, race or ethnicity, marital status, insurance status, and preferred language (English or Spanish). Participants reported educational attainment and indicated whether they require assistance with reading written health information (never or rarely sometimes or often or always). Participants reported annual household income and also rated their financial strain on a 4-point scale from “We have enough and can save” to “We don’t have enough and we have great difficulties” [34].

Clinical Characteristics

BMI data were extracted from the electronic health record with the value most proximal to the date of the baseline assessment. Participants self-reported smoking status, and duration of T2D and completed the Charlson comorbidity index [35] which assesses the presence or absence of 20 common comorbidities.

Portal-Related Outcomes

Portal activation and engagement were recorded as the number of portal logins per participant per month. Portal activation was defined as the creation of a patient portal account and the use of the portal in the first month. The portal engagement was defined as ongoing use of the portal over the 6-month study duration. We calculated logins per week and per month. Based on previous studies, we defined portal engagement (consistent portal use) as two or more patient logins per month [36] and calculated the percentage of participants meeting this benchmark. Digital health literacy was measured with 4 items (eg, “I can use apps [like Zoom] on my cell phone, tablet or computer on my own without asking for help from someone else”) on a scale from 1=strongly disagree to 5=strongly agree. Technology acceptability was measured via self-report at baseline, 3 months and 6 months with subscales including ease of use (eg, “I think it will be easy [is easy] to send a message to my provider in the patient portal,” perceived usefulness (“I think that using the patient portal will help me [helps me] understand my diabetes care”) and confidence (eg, “I am confident in my ability to review my health records on the patient portal”) on a scale from 1=strongly disagree to 5=strongly agree [37]. Items were tailored to reference each participant’s respective patient portal (MyChart or Healow) [38,39].

Health-Related Outcomes

HbA_{1c} was measured using a fingerstick point of care A1CNow+ (PTS Diagnostics). The validity of this point-of-care assessment of HbA_{1c} has been confirmed through comparisons with clinical laboratory measurements of HbA_{1c} [40,41]. Self-report data included self-management, self-efficacy, perceived autonomy support, and diabetes distress. To assess the reliability of self-report scales, α coefficients were calculated. Self-management was measured by the summary of diabetes self-care activities (eg, diet, medication adherence, blood glucose monitoring, and physical activity). Participants are asked in the last 7 days, how many days did you follow your T2D recommendations for each health behavior (from 0 to 7). Reliability and validity of English and Spanish versions have been established [42,43]. Self-efficacy was measured by the Stanford Diabetes Self-Efficacy Scale, an 8-item scale specific to T2D self-management self-efficacy. Items include confidence in exercise, interpreting blood glucose levels, and following dietary recommendations, with response options from 1=not at all confident to 5=totally confident. Reliability and validity have been established in Spanish-speaking adults [44,45], with an α coefficient of 0.92 in our sample. Perceived autonomy support was measured with the 15-item Health-Care Climate Questionnaire [46], which assesses patients’ perceptions of the degree to which they experience the nurse from the CHC to be autonomy supportive versus controlling (eg, “my diabetes provider encourages me to ask questions”). Response options were on a 5-point Likert scale from 1=strongly disagree to 5=strongly agree and the α coefficient was 0.91 in our sample. Diabetes distress was assessed with the 20-item Problem Areas in Diabetes scale (PAID) [47,48] (eg, “feeling overwhelmed by your diabetes”) with response options on a 5-point Likert scale from 0=not a problem to 4=serious problem. This widely used scale has evidence of strong reliability and validity; the α coefficient was 0.96 in our sample.

Feasibility and acceptability of data collection were determined as a priori based on an established framework [49]. Feasibility data included recruitment (recruits invited vs consented), attrition (number of participants who withdrew from the study or were lost to follow-up), and participant technology access and use (smartphone, computer, or tablet). Treatment fidelity was assessed by calculating percent adherence to session checklists created for the CHWs and nurses respectively. Barriers to clinical integration were documented by the

XSL•FO
RenderX

researchers during regular supervision with study nurses and research assistants.

Acceptability data included participant surveys at 3 months on satisfaction with the intervention [50] (eg, “Were the topics in the program important to you?”) and participant-reported therapeutic alliance with the nurse (“How much do you like or trust or have confidence in your nurse?”) [51,52] both rated from 1=not at all to 5=extremely. These questionnaires were modified from published versions to be specific to the MAP intervention protocol. At 6 months, we also asked participants about their intention to continue portal use after study completion (yes or no) and whether they would recommend MAP to a friend (yes or no).

Ethical Considerations

This study (NCT05180721) was approved by the institutional review board at Yale University (IRB# 2000031753; approved on December 21, 2022). All patients provided informed consent before participating in this study.

Data Analysis

All data were downloaded from REDCap onto a secure server. Descriptive analyses were performed to assess demographic and clinical characteristics of the sample. Distributions of outcome variables were examined for central tendency and dispersion. We estimated the Cohen *d* effect size of MAP on A1C and tested the statistical significance of the change from baseline using longitudinal models, including generalized linear mixed model (GLMM), a logistic model with random intercept (which incorporates the correlation within repeated measures), and a negative binomial model with random intercept. The coefficients of the categorial time variable (ie, baseline, 3 months, and 6 months) represent the average change of HbA_{1c} at 3 and 6 months from baseline. The GLMM included all participants with data at baseline and at least one postintervention value and missing data was handled using the maximum likelihood approach. The repeatedly measured secondary outcomes were analyzed with the same approach using GLMM, with each outcome analyzed individually. We used the GLMM identify link function for continuous variables, the logit link function for binary variables, and the log link function for count variables. Residuals were assessed for normality assumption and variables were transformed with an appropriate form when the normality assumption did not hold. Log-in data from one participant was excluded due to errors encountered in downloading it from the portal that our IT expert could not resolve.

Results

Sample

Recruitment and data collection took place from May 2023 to July 2024. A total of 47 participants were approached about the study, 26 were eligible and provided informed consent, 23 received the intervention and one was lost to follow up (refer to Figure 1 for CONSORT [Consolidated Standards of Reporting Trials] diagram). The sample of completers (n=22) was recruited from 2 clinics, 68% (n=15) from 1 clinic, 32% (n=7) from the other clinic. Refer to Table 1 for demographic characteristics. The sample was predominately Latino or Hispanic (7/22, 77%) had a mean age of 56.32 (SD 10.93) years, 73% (16/22) were female, and 55% (12/22) were married or partnered. The majority reported low-income (86% <US $40,000/year, 19/22), low educational attainment (59%, 13/22 less than high school graduates), and no health insurance (55%, 12/22). All participants had access to a smartphone, but 91% (20/22) had never accessed a health app. Yet, they reported some confidence in using apps, setting up video chats, solving basic technical issues, and using a tablet (mean score 3.1, SD 1.0 with a scale ranging from 1 to 5). They also reported positive perceptions of portal ease of use, usefulness, and confidence in using the portal. The mean duration of diabetes was 11.75 (SD 9.09) years, 36% (8/22) were on insulin at baseline, 14% (3/22) were current smokers, 32% (7/22) reported a history of depression, 59% (13/22) reported severe diabetes distress, 68% (15/22) reported a history of hypertension, and baseline HbA_{1c} was 8.31% (1.05%; measured using the A1CNow system). In total, 73% (16/22) of participants had a baseline HbA_{1c} level greater than the recommended 7.0% for adults with T2D and 73% (16/22) rated their health as fair or poor while 27% (6/22) rated it as good or very good. Mean BMI was 32.21 (SD 5.98), with 23% (5/22) of the sample overweight and 68% (15/22) were categorized as obese. Other baseline clinical data are reported in Table 2 (Data extracted from medical records except for HbA_{1c}, which was collected by researchers for study assessment).

XSL•FO
RenderX

Figure 1. Consolidated Standards of Reporting Trials (CONSORT) diagram.

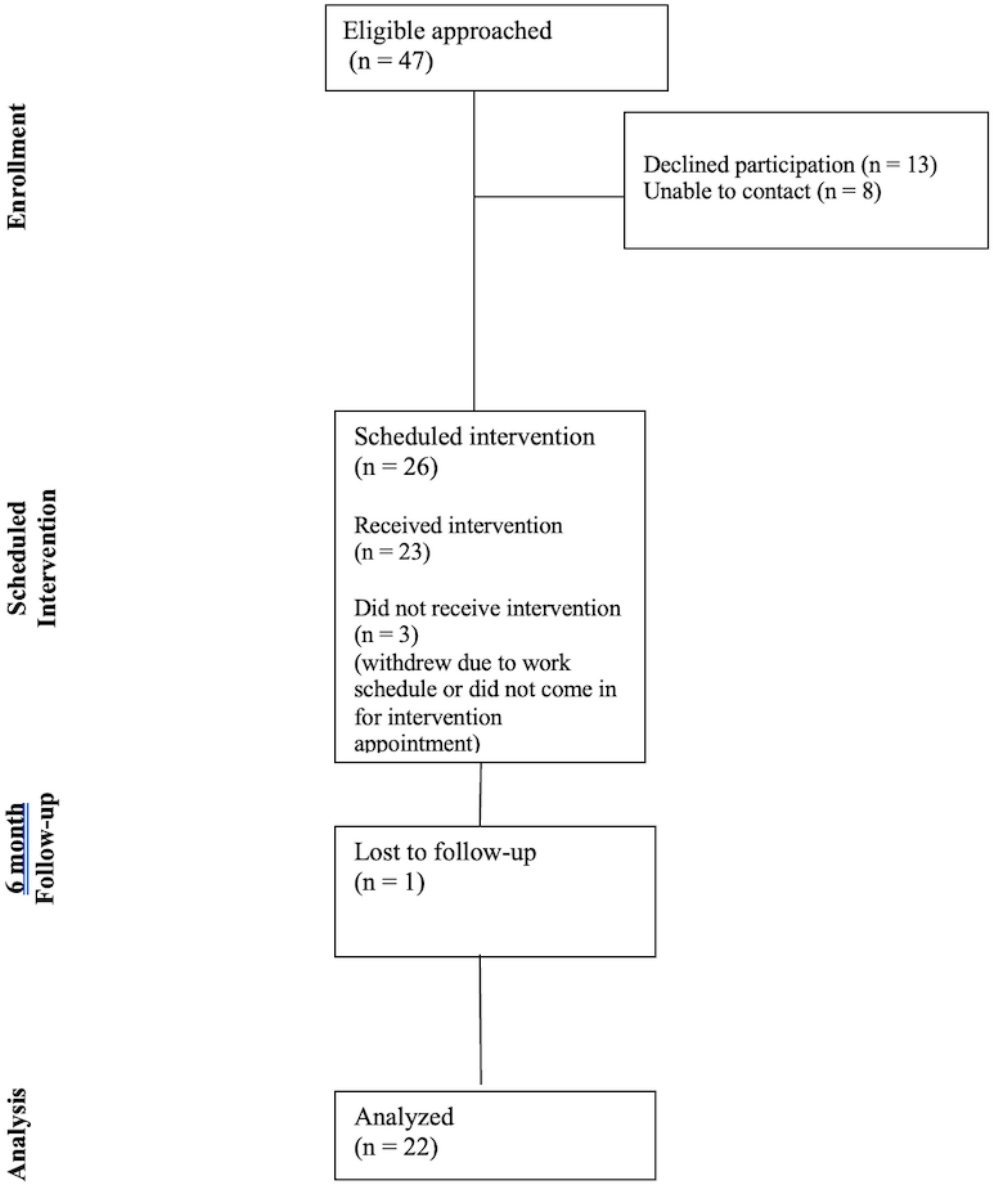

XSL•FO
RenderX

Table 1. Participant demographics (n=22).

Variables	Values
Site	
Percentage of site 1, n (%)	15 (68)
Age in years, mean (SD)	56.32 (10.93)
Sex	
Percentage of females, n (%)	16 (73)
Duration diabetes in years, mean (SD)	11.75 (9.09)
Charleston comorbidity index, mean (SD)	3.41 (2.32)
Married or partnered (vs not), n (%)	12 (55)
Ethnicity	
Percentage of Hispanic, n (%)	17 (77)
Race, n (%)	
Percentage of White	10 (45)
Percentage of Black	4 (18)
Percentage of other	8; of them, 6 reported Latino/Hispanic (36)
Employment, n (%)	
Percentage of working part/full time (vs others)	8 (36)
Education, n (%)	
Percentage of high school graduates or more	9 (39)
Annual income, n (%)	
Less than US $40,000	19 (86)
Financial difficulties, n (%)	
Percentage of difficulties or great difficulties	15 (68)
Preferred language, n (%)	
Spanish	14 (64)
Spanish–not able to converse in English	12 (55)
Assistance with reading health information	
Sometimes or often or always	10 (46)
Health insurance, n (%)	
No insurance	12 (55)
Self-reported use of community resources, n (%)	
Food assistance	7 (32)
Medication access	6 (27)
Housing	4 (18)
Health insurance	4 (18)
Transportation	2 (9)
Utilities	2 (9)
Childcare or employment	0 (0)
Home internet stability, n (%)	
No home internet or unsure	3 (15)
Fair	6 (27)
Good	7 (32)
Very good	6 (27)

XSL•FO
RenderX

Variables	Values
Home cell phone data stability, n (%)	
Unsure	3 (15)
Fair	6 (27)
Good	7 (32)
Very good	6 (27)

Table 2. Participant clinical characteristics at baseline (n=22).

Variables	Mean (SD)	Recommended values	Meeting recommendations, n (%)
Hemoglobin A_{1c}, %	8.31 (1.50)	<7	6 (28)
BMI	32.31 (5.98)	18.5-24.9	2 (9)
Systolic BP[a], mm Hg	124.57 (14.53)	<130	13 (62)
Diastolic BP, mm Hg	74.48 (7.90)	<80	16 (76)
Cholesterol, mg/dL	161.33 (44.59)	<200	17 (81)
LDL[b], mg/dL	76.57 (35.73)	<100	16 (78)
HDL[c], md/dL	53.00 (26.73)	>40	10 (63)
Triglycerides, mg/dL	164.48 (119.24)	<150	15 (71)

[a]BP: blood pressure.

[b]LDL: low-density lipoprotein.

[c]HDL: high-density lipoprotein.

Protocol Implementation

SDOH needs were assessed before beginning the technology training with 36% (8/22) of participants requiring a referral. Referrals needed were for assistance with utilities (n=4), transportation (n=3), food (n=3), and insurance (n=2). Training on the use of the portal averaged 84.55 (SD 49.32) minutes, with a range between 30-180 minutes. The portal training protocol implementation was high at 85% (524/616 tasks) fidelity across all protocol items and participants. Participants reported moderate confidence in using the portal after the training session with a mean confidence of 2.3 (SD 0.93) on a 3 point-scale (low, moderate, and high confidence). The majority of participants completed the portal training in one session; however, 4 participants (4/22, 18%) required technology support or additional training due to technical challenges using the portal. Nurse protocol implementation was high at 84% (222/264; 12 weeks × 22 participants) at 3 months; but decreased to 52% (137/264) at 6 months with fewer messages sent by nurses over time.

Barriers to integrating MAP into routine clinical care that was noted during regular supervision were varied and included the following: (1) turnover of CHWs, (2) MAP nurses not embedded in the participant's clinical team, (3) portal or clinic constraints on messaging from patient to PCP, (4) difficulty tracking those portal messages from nurses to patients that were not opened and acknowledged, (5) cumbersome and error-prone portal features for transmitting glucose data from patient to clinician, and (6) few high-quality diabetes education videos suitable for low-literacy, low-numeracy, and Spanish-speaking individuals that could be delivered via the portal.

Acceptability

Program satisfaction was high at 3 and 6 months (mean 4.03, SD 0.81 and mean 4.14, SD 0.49 on a 5-point scale respectively). All participants indicated that they would recommend the MAP intervention to a friend. Participants also reported high autonomy support from nurses at 3 months which increased at 6 months. At 6 months, 100% (22/22) of participants reported that they would continue to use the portal for diabetes care and 100% (22/22) felt like the portal was helpful for their diabetes care.

Portal Outcomes

Refer to Table 3 for outcomes over time. Portal activation was high with 100% (22/22) of participants creating a portal account and logging into the portal in the first month (n=21). Mean logins per week was 3.16 (SD 1.55) over the first 3 months and 1.45 (SD 0.93) between 3 and 6 months (Figure 2). Mean portal logins per month were 12.65 (SD 6.21) for the first 3 months and 5.79 (SD 3.74) for the last 3 months. Participant engagement (as defined by the literature) was high, with 96% (20/21) of participants logging into the portal at least twice per month in the first 3 months and 76% (16/21) of participants meeting this benchmark between 3 and 6 months. At baseline, participants perceived that the portal would be easy to use (mean score 4.0, SD 0.55 on a 5-point scale) and useful (mean score 4.25, SD 0.44) with no significant change over time. There was a significant increase in technology confidence over 6 months ($P<.05$), with a trend for increased digital health literacy at 6 months ($P=.08$).

XSL•FO
RenderX

Table 3. Change over time for portal and health outcomes (n=22).

	Possible Range	Baseline	3 months	6 months
Portal-related outcomes				
Portal logins/week[a]	—[b]	—	3.16 (3.18)	1.45 (1.91)
Portal logins/month[a]	—	—	12.65 (6.21)	5.79 (3.74)
Portal engagement[a]	—	—	96% met benchmark	76% met benchmark
Technology confidence	1-5	3.81 (0.86)	4.00 (0.73)	4.38 (0.56)[c]
Digital health literacy	1-5	3.15 (1.03)	3.23 (0.93)	3.59 (0.87)[c]
Portal perceived ease of use	1-5	4.00 (0.55)	3.89 (0.78)	4.08 (0.67)
Portal perceived usefulness	1-5	4.25 (0.44)	4.15 (0.660	4.31 (0.51)
Health-related outcomes				
Hemoglobin A_{1c}, %	—	8.31 (1.65)	8.09 (1.64)	8.23 (1.35)
PAID[d]	0-100	43.58 (30.51)	29.09 (23.03)[c]	25.45 (24.88)[c]
DSM[e] Self-Efficacy	1-5	3.14 (1.01)	3.44 (1.17)	3.68 (1.03)[c]
Healthcare Climate Questionnaire	1-5	4.15 (0.78)	4.14 (0.83)	3.95 (1.07)
SDSCA[f]	0-7	—	—	—
Diet	—	2.82 (2.91)	4.00 (2.54)	3.68 (2.80)
Exercise	—	2.61 (2.23)	3.23 (2.11)	3.41 (2.29)
Glucose checking	—	4.66 (2.49)	4.95 (2.44)	5.41 (2.21)
Footcare	—	5.30 (1.01)	1. 1.06)	1.79 (.94)
Medication	—	6.65 (1.00)	6.90 (0.31)	6.71 (0.73)
Use of community resources	0-9	1.32 (1.52)	1.32 (1.55)	1.36 (1.65)

[a]n=21.

[b]Not applicable.

[c]significant change from baseline at a 5% significance level.

[d]PAID: Problem Areas in Diabetes (diabetes distress).

[e]DSM: diabetes self-management.

[f]SDSCA: Summary of Diabetes Self-Care Behaviors; the Healthcare Climate questionnaire measures autonomy support.

XSL•FO
RenderX

Figure 2. Portal logins over time.

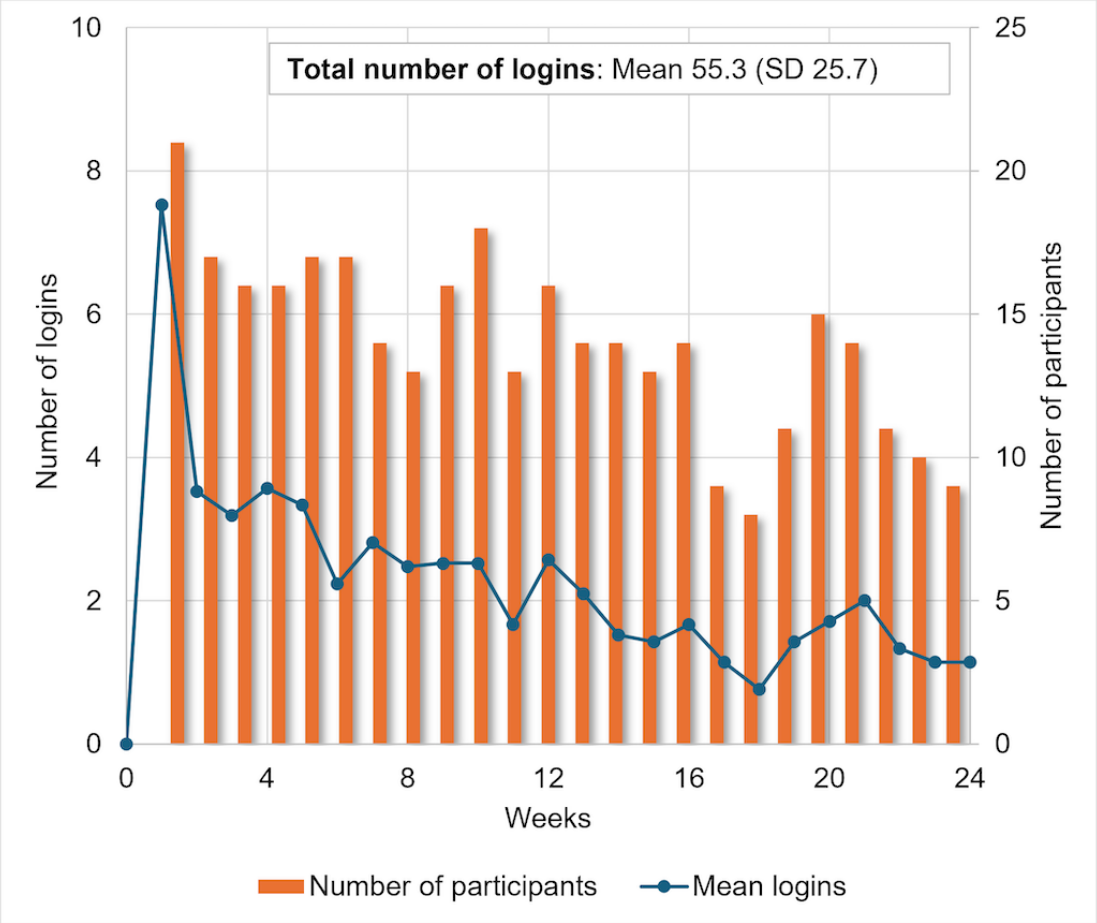

Health Outcomes

In the mixed effect model, diabetes distress (PAID) significantly decreased from baseline to 3 months; it continued to decrease significantly from 3 months to 6 months ($P<.01$; Table 3). The proportion of clinically elevated diabetes distress (ie, PAID>40) was 50% (11/22) at baseline, and it decreased to 36.4% (8/22) at 3 months ($P=.20$) and 22.7% (5/22) at 6 months ($P=.03$). Diabetes self-efficacy increased from baseline to 3 months and continued to increase until 6 months when it became statistically significant. A similar pattern of significant change from baseline to 6 months was found for confidence in using technology and digital health literacy. HbA_{1c} was slightly decreased at 3 months with a small effect size (Cohen $d=-0.17$) but returned to baseline level at 6 months (Cohen $d=-0.07$). There was no significant change in perceived autonomy support or diabetes self-care behaviors.

Discussion

Principal Findings

This pilot study evaluated an intervention to increase patient portal use in adults with T2D who access health care at CHCs. We reached a sample of adults of diverse races or ethnicities, low income, and low educational attainment with low use of the patient portal. The MAP intervention was specifically designed to address multiple levels of challenges identified by CHC patients and providers to patient portal usage. Thus, to enhance successful implementation, participants received a tablet and data plan, SDOH screening, training in patient portal navigation with ongoing support, and diabetes education and support by nurses. Overall, feasibility, acceptability, and improvement in portal and health outcomes of clinical significance were demonstrated yet integration into clinical care was challenging.

First, we demonstrated high feasibility by tailoring implementation strategies to meet the unique needs of each clinic. We also demonstrated low attrition, high patient portal activation (22/22, 100%), and high patient portal engagement over 6 months. High participant activation is not surprising as we had dedicated staff available to train participants in portal use and assist with any challenges. Although the measurement of patient portal engagement is not standardized in the literature, a common metric is the frequency of logins per month. Consistent portal use has been defined as two or more patient logins per month according to a recent systematic review of the measurement of patient portal use [36]. In another study, super-users were defined as those logging in twice or more per month [53]. In our study, consistent portal use was demonstrated in 96% of participants over 3 months, 76% of participants between 3 and 6 months, and 87% over a 6-month period among 21 patients. Thus, the results of our portal activation and engagement are encouraging. Factors that likely contributed to these positive outcomes were technology training or support for the duration of the study, participants' development of a relationship with a clinic nurse, and consistent outreach by the

XSL•FO
RenderX

nurse over 6 months via the portal to provide DSMS. While the provision of a tablet and a data plan may have also contributed to portal engagement over time in this study, we judge that increasing access to smartphones combined with free internet sites in community settings (eg, libraries and coffee shops) may be sufficient for future implementation efforts. Decline over time in consistent use may be partially explained by less consistent protocol implementation by nurses from 3 to 6 months and lack of full integration of MAP into clinic workflow.

Engagement over time was supported by improvements in other portal-related outcomes. Whereas there was no change in how participants viewed the portal per se (perceived usefulness and perceived ease of use), participant perceptions of their own interactions with the technology did improve. Increases were observed in both technology confidence ("I am confident in my ability to use MyChart") and digital health literacy ("I can solve, or figure out how to solve, basic technical issues on my cell phone, computer, or other device"). Those improvements reached statistical significance at 6 months suggesting that individuals may need more than 3 months of supported use of the tablet and the portal to create change in confidence and digital health literacy. Also encouraging was the high level of acceptability including participant satisfaction with the MAP intervention and high intention to continue use of the portal after the MAP study ended.

Second, some clinically important health-related outcomes significantly improved in our small sample. The decrease in diabetes distress at 3 months, with a further decrease at 6 months, is important, given the body of research showing that diabetes distress is robustly associated with lower self-management and may be associated with higher HbA_{1c} [54,55]. The MAP intervention's reduction in diabetes distress is consistent with findings from other interventions. A meta-analysis of 41 studies testing eHealth interventions for diabetes self-management found that such interventions are effective in reducing diabetes distress [56]. Our rate of elevated diabetes distress (59%) is higher than a meta-analysis of 55 studies, which showed an overall prevalence of 36% of diabetes distress in people with type 2 diabetes, with higher rates in predominantly female samples and those with more comorbidities [57]. Our small sample size, comprised of participants with elevated HbA_{1c} and who enrolled in a treatment study may account for this high rate of distress. Alternatively, diabetes self-management may be exceptionally distressing in the context of high SDoH needs, as is common among patients at CHCs. In this way, MAP may be particularly beneficial for adults with T2D who access health care at CHCs. One study in T2D found that patients who had high diabetes distress at baseline had a greater increase in self-management and a decrease in HbA_{1c} from mHealth DSMS compared to their low-distress counterparts [58].

Diabetes self-efficacy increased at 3 months but only became statistically significant at 6 months, which may reflect our small sample size since larger studies have shown increases by 3 months [59]. Alternatively, this sample with elevated HbA_{1c}, high diabetes distress, and high social needs may require 6 months of intervention to increase self-efficacy. Diabetes self-efficacy has been shown in a systematic review to be an important predictor of self-care behaviors [60] including for example medication adherence [61]. The improvement in diabetes self-efficacy observed in our study may prove important for downstream clinical outcomes since interventions that improve self-efficacy have been shown to directly and indirectly improve glycemic control [62].

We observed that HbA_{1c} and DSM improved at 3 and 6 months relative to baseline values but these improvements were not statistically significant in our small sample. Research on the effectiveness of patient portals on glycemic outcomes has produced inconsistent results [13,14,63-65]; yet, observational studies with adequate sample sizes do in fact provide consistent evidence that patient portal use has a beneficial impact on HbA_{1c} [66-68]. The consistency of portal use is also associated with benefits for HbA_{1c}. In a study that examined the consistency of patient portal use among adults with T2D (N=95,043), an increasing number of calendar months of patient portal use was associated with a significant decrease in HbA_{1c} levels [69]. In addition, several studies have highlighted the role of patient portals in improving diabetes self-management behaviors. A retrospective cohort study involving over 100,000 adults with diabetes found that when patients are given access to the patient portal via a mobile device, their adherence to oral antihyperglycemic medications improves significantly [66].

In addition to the small sample size, our study may not have produced significant improvements in HbA_{1c} and DSM because of the low degree of integration of MAP into clinical care. We ensured that MAP was delivered in a clinical setting by embedded clinicians, yet we encountered a number of barriers to a deeper level of clinical integration. For example, in one clinic, participants were not able to send portal messages directly to their PCPs because that portal functionality had not yet been activated clinic-wide. As another example, at the other clinic, the MAP nurse was not part of the patient's own clinical team. Whereas MAP nurses did send monthly progress notes to the patients' respective PCPs, it is unknown how the PCPs made use of the reports. In our study, one nurse was bilingual in English or Spanish; however, not all were and one participant had Polish as a first language. In these situations, translation of messages was required. Thus, MAP's good metrics of implementation were only made possible with creative approaches and "work arounds" to clinic and portal idiosyncrasies. It remains untested how effectively MAP could improve glycemic control and DSM when fully integrated into clinical care.

Health care providers and adults with diabetes in CHCs continue to face barriers that limit their use of portals. Well-documented challenges include low health- and technical literacy, lack of regular access to internet-connected mobile phones, limited language concordance between providers and patients, and other SDoH-related issues [25,70]. CHCs also face the challenge of lacking standardized implementation strategies for rolling out portals and supporting patients to use portals. The findings from the current pilot study provide preliminary evidence of what may work in CHC settings to mitigate these barriers and improve portal use among adults with T2D.

XSL•FO
RenderX

Clinical Implications

Our findings have several clinical implications. First, having clinicians recommend the use of the patient portal and having dedicated staff to provide training and technical support to patients in CHCs can be useful in improving portal use. Racial and ethnic differences in portal use attenuate when clinicians offer portals to all patients irrespective of their racial and ethnic background [71]. In addition, a recent systematic review to summarize research priorities and best practices in patient portals found that studies that report high rates of portal use often have dedicated staff to enroll and assist patients [72]. Evidence-based techniques to educate socially vulnerable patients, such as screening for health literacy, using the teach-back method, and using bilingual messaging, were used in our intervention and can also be adopted by clinicians in CHCs. Many CHCs incorporate nursing case management for patients with chronic conditions. Nurse case management communication could also be expanded to use the patient portal for ongoing check-ins, support, and patient education, which may enhance portal engagement and health outcomes. In a systematic review of the effectiveness of patient education through patient portals, significant improvements in knowledge, self-management, health behaviors, mental health, and select health outcomes were demonstrated [73].

Lastly, clinic administrators at CHCs could work with their electronic health record support team to eliminate the need for email accounts to activate the patient portal. Patients who access care at CHCs may not have email accounts, which creates an additional barrier to portal use among CHC patients [74]. Alternative approaches such as the use of phone numbers to authenticate users should be encouraged, as the majority of patients in CHCs have access to mobile phones.

Limitations and Future Research

Findings should be interpreted with caution given study limitations, the primary of which is a design without a comparison group or randomization to treatment. Future research should test MAP using a more rigorous study design. Implementation research should identify strategies to more fully integrate interventions like MAP into complex clinical settings. Common metrics of portal use should be used as outcomes, in addition to study-specific indicators, in order to facilitate comparison across studies. Our sample size at each clinic was insufficient to test for differences between clinics in sample characteristics or differences in outcomes. Study samples should also be large enough to be powered to test for clinically meaningful improvements in HbA_{1c} and to include adequate representation of medically underserved individuals so that intervention effects on disparities in portal use—both ameliorating and exacerbating—can be examined. A longer duration of follow-up would allow for investigation of the durability of any treatment effect.

There is also a need for research to develop portal platforms and functionality that better meet the needs of patients and clinicians. This could include, for example, graphical and pictorial data displays, audiovisual capabilities, and libraries of patient education materials. Beyond the patient portal, our preparation for the intervention implementation revealed a paucity of high-quality, patient-friendly diabetes educational materials (pictorial, video, interactive, and gaming) that are designed for low-literacy, low-numeracy, and Spanish-speaking individuals. Whereas a wide variety of materials exist, few if any meet all the above criteria, and those that try tend to be extremely cursory.

Conclusions

In this pilot study among people with diabetes receiving care at CHCs, the MAP intervention produced high activation and engagement in portal use as well as meaningful improvement in psychosocial outcomes and promising changes in clinical outcomes. Numerous challenges were also identified that can be addressed in future research. Increasing portal use specifically in health disparities populations may be one key to ameliorating disparities in diabetes outcomes.

Acknowledgments

National Institute on Minority Health and Health Disparities (R21MD016949).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Multilevel intervention to increase access and use of portals (MAP) components aligned with the WHO (World Health Organization) Health Equity framework.
[DOCX File , 15 KB - formative_v9i1e67293_app1.docx]

Multimedia Appendix 2
Sequence and timing of multilevel intervention to increase access and use of portals (MAP).
[DOCX File , 14 KB - formative_v9i1e67293_app2.docx]

XSL•FO
RenderX

References

1. Kirk J, D'Agostino RBJ, Bell RA, Passmore LV, Bonds DE, Karter AJ, et al. Disparities in HbA1c levels between African-American and non-hispanic white adults with diabetes: a meta-analysis. Diabetes Care 2006;29(9):2130-2136 [FREE Full text] [doi: 10.2337/dc05-1973] [Medline: 16936167]
2. Osborn CY, de Groot M, Wagner JA. Racial and ethnic disparities in diabetes complications in the northeastern United States: the role of socioeconomic status. J Natl Med Assoc 2013;105(1):51-58 [FREE Full text] [doi: 10.1016/s0027-9684(15)30085-7] [Medline: 23862296]
3. Spanakis EK, Golden SH. Race/ethnic difference in diabetes and diabetic complications. Curr Diab Rep 2013;13(6):814-823 [FREE Full text] [doi: 10.1007/s11892-013-0421-9] [Medline: 24037313]
4. Yancy CW. COVID-19 and African Americans. JAMA 2020;323(19):1891-1892. [doi: 10.1001/jama.2020.6548] [Medline: 32293639]
5. Yan Y, Yang Y, Wang F, Ren H, Zhang S, Shi X, et al. Clinical characteristics and outcomes of patients with severe covid-19 with diabetes. BMJ Open Diabetes Res Care 2020;8(1):e001343 [FREE Full text] [doi: 10.1136/bmjdrc-2020-001343] [Medline: 32345579]
6. Richardson S, Hirsch JS, Narasimhan M, Crawford JM, McGinn T, Davidson KW, the Northwell COVID-19 Research Consortium, et al. Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with COVID-19 in the New York City area. JAMA 2020;323(20):2052-2059 [FREE Full text] [doi: 10.1001/jama.2020.6775] [Medline: 32320003]
7. Reed ME, Huang J, Brand RJ, Neugebauer R, Graetz I, Hsu J, et al. Patients with complex chronic conditions: health care use and clinical events associated with access to a patient portal. PLoS One 2019;14(6):e0217636 [FREE Full text] [doi: 10.1371/journal.pone.0217636] [Medline: 31216295]
8. Han HR, Gleason KT, Sun CA, Miller HN, Kang SJ, Chow S, et al. Using patient portals to improve patient outcomes: systematic review. JMIR Hum Factors 2019;6(4):e15038 [FREE Full text] [doi: 10.2196/15038] [Medline: 31855187]
9. Devkota B, Salas J, Sayavong S, Scherrer JF. Use of an online patient portal and glucose control in primary care patients with diabetes. Popul Health Manag 2016;19(2):125-131. [doi: 10.1089/pop.2015.0034] [Medline: 26237200]
10. Chung S, Panattoni L, Chi J, Palaniappan L. Can secure patient-provider messaging improve diabetes care? Diabetes Care 2017;40(10):1342-1348. [doi: 10.2337/dc17-0140] [Medline: 28807977]
11. Harris SM, Joyce H, Miller A, Connor C, Amiel SA, Mulnier H. The attitude of healthcare professionals plays an important role in the uptake of diabetes self-management education: analysis of the barriers to uptake of type 1 diabetes education (BUD1E) study survey. Diabet Med 2018. [doi: 10.1111/dme.13704] [Medline: 29882990]
12. Kuo A, Dang S. Secure messaging in electronic health records and its impact on diabetes clinical outcomes: a systematic review. Telemed J E Health 2016;22(9):769-777. [doi: 10.1089/tmj.2015.0207] [Medline: 27027337]
13. Tang PC, Overhage JM, Chan AS, Brown NL, Aghighi B, Entwistle MP, et al. Online disease management of diabetes: engaging and motivating patients online with enhanced resources-diabetes (EMPOWER-D), a randomized controlled trial. J Am Med Inform Assoc 2013;20(3):526-534 [FREE Full text] [doi: 10.1136/amiajnl-2012-001263] [Medline: 23171659]
14. Ralston JD, Hirsch IB, Hoath J, Mullen M, Cheadle A, Goldberg HI. Web-based collaborative care for type 2 diabetes: a pilot randomized trial. Diabetes Care 2009;32(2):234-239 [FREE Full text] [doi: 10.2337/dc08-1220] [Medline: 19017773]
15. Alturkistani A, Qavi A, Anyanwu PE, Greenfield G, Greaves F, Costelloe C. Patient portal functionalities and patient outcomes among patients with diabetes: systematic review. J Med Internet Res 2020;22(9):e18976 [FREE Full text] [doi: 10.2196/18976] [Medline: 32960182]
16. Anthony DL, Campos-Castillo C, Lim PS. Who Isn't using patient portals and why? Evidence and implications from a national sample Of US adults. Health Aff (Millwood) 2018;37(12):1948-1954. [doi: 10.1377/hlthaff.2018.05117] [Medline: 30633673]
17. Grossman L, Masterson Creber RM, Benda N, Wright D, Vawdrey D, Ancker J. Interventions to increase patient portal use in vulnerable populations: a systematic review. J Am Med Inform Assoc 2019;26(8-9):855-870 [FREE Full text] [doi: 10.1093/jamia/ocz023] [Medline: 30958532]
18. American Heart Association. Individuals' Ability to Electronically Access Their Hospital Medical Records, Perform Key Tasks Is Growing. URL: https://www.aha.org/system/files/research/reports/tw/16jul-tw-healthIT.pdf [accessed 2021-02-04]
19. Adler-Milstein J, DesRoches CM, Kralovec P, Foster G, Worzala C, Charles D, et al. Electronic health record adoption in US hospitals: progress continues, but challenges persist. Health Aff (Millwood) 2015;34(12):2174-2180. [doi: 10.1377/hlthaff.2015.0992] [Medline: 26561387]
20. Veinot TC, Mitchell H, Ancker JS. Good intentions are not enough: how informatics interventions can worsen inequality. J Am Med Inform Assoc 2018;25(8):1080-1088 [FREE Full text] [doi: 10.1093/jamia/ocy052] [Medline: 29788380]
21. Lorenc T, Petticrew M, Welch V, Tugwell P. What types of interventions generate inequalities? Evidence from systematic reviews. J Epidemiol Community Health 2013;67(2):190-193. [doi: 10.1136/jech-2012-201257] [Medline: 22875078]
22. The number of patients continues to grow at community health centers. National Association of Community Health Centers. 2017. URL: https://www.nachc.org/the-number-of-patients-continues-to-grow-at-community-health-centers/ [accessed 2020-10-01]

XSL•FO
RenderX

23. Community health center chart book. National Association of Community Health Centers. 2019. URL: http://www.nachc.org/wp-content/uploads/2019/01/Community-Health-Center-Chartbook-FINAL-1.28.19.pdf [accessed 2020-10-01]
24. Solar O, Irwin AA. A conceptual framework for action on the social determinants of health. Social Determinants of Health Discussion Paper.: World Health Organiation; 2010. URL: https://www.who.int/sdhconference/resources/ConceptualframeworkforactiononSDH_eng.pdf [accessed 2021-02-01]
25. Akyirem S, Wagner J, Chen HN, Lipson J, Minchala M, Cortez K, et al. Recommendations to address barriers to patient portal use among persons with diabetes seeking care at community health centers: interview study with patients and health care providers. JMIR Diabetes 2024;9:e58526 [FREE Full text] [doi: 10.2196/58526] [Medline: 39284181]
26. Becker AL. Health Disparities in Connecticut: Causes, Effects, and What We Can Do. 2020. URL: https://www.cthealth.org/wp-content/uploads/2020/01/Health-disparities-in-Connecticut.pdf [accessed 2025-02-28]
27. Callahan CM, Unverzagt FW, Hui SL, Perkins AJ, Hendrie HC. Six-item screener to identify cognitive impairment among potential subjects for clinical research. Med Care 2002;40(9):771-781. [doi: 10.1097/00005650-200209000-00007] [Medline: 12218768]
28. American Diabetes Association Professional Practice Committee. Introduction and methodology: standards of care in diabetes-2024. Diabetes Care 2024;47(Supplement_1):S1-S4. [doi: 10.2337/dc24-SINT] [Medline: 38078587]
29. Osborn CY, Cavanaugh K, Kripalani S. Strategies to address low health literacy and numeracy in diabetes. Clinical Diabetes 2010;28(4):171-175. [doi: 10.2337/diaclin.28.4.171]
30. Robinson SA, Zocchi MS, Netherton D, Ash A, Purington CM, Connolly SL, et al. Secure messaging, diabetes self-management, and the importance of patient autonomy: a mixed methods study. J Gen Intern Med 2020;35(10):2955-2962 [FREE Full text] [doi: 10.1007/s11606-020-05834-x] [Medline: 32440998]
31. Dendere R, Slade C, Burton-Jones A, Sullivan C, Staib A, Janda M. Patient portals facilitating engagement with inpatient electronic medical records: a systematic review. J Med Internet Res 2019;21(4):e12779 [FREE Full text] [doi: 10.2196/12779] [Medline: 30973347]
32. Diabetes self-management education and support toolkit. Centers for Disease and Control. 2020. URL: https://www.cdc.gov/diabetes-toolkit/php/about-dsmes/index.html [accessed 2025-02-28]
33. Powers MA, Bardsley J, Cypress M, Duker P, Funnell MM, Fischl AH, et al. Diabetes self-management education and support in type 2 diabetes. Diabetes Educ 2017;43(1):40-53. [doi: 10.1177/0145721716689694] [Medline: 28118121]
34. Paras PP-E PR. El rostro de la pobreza: la inseguridad alimentaria en el Distrito Federal. Rev Este Pais 2004;158:45-50.
35. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J Chronic Dis 1987;40(5):373-383. [doi: 10.1016/0021-9681(87)90171-8] [Medline: 3558716]
36. Beal LL, Kolman JM, Jones SL, Khleif A, Menser T. Quantifying patient portal use: systematic review of utilization metrics. J Med Internet Res 2021;23(2):e23493 [FREE Full text] [doi: 10.2196/23493] [Medline: 33629962]
37. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly 1989;13(3):319-340. [doi: 10.2307/249008]
38. Whittemore R, Vilar-Compte M, De La Cerda S, Delvy R, Jeon S, Burrola-Méndez S, et al. A self-management randomized controlled pilot trial for low-income adults with type 2 diabetes in Mexico City. Curr Dev Nutr 2020;4(5):nzaa074 [FREE Full text] [doi: 10.1093/cdn/nzaa074] [Medline: 32368713]
39. Wagner J, Berthold SM, Buckley T, Bermudez-Millan A, Ha T, Scully M, et al. Remote peer learning between US and cambodian lay health workers to improve outcomes for cambodians with type 2 diabetes: a pilot study. Int J Behav Med 2020;27(5):609-614. [doi: 10.1007/s12529-020-09896-3] [Medline: 32435878]
40. Moskowitz KW, Mercer K, Suscha E, Shafai M, Lindke C, Xie C, et al. Multi-Center Accuracy Assessment of A1CNow® + : A Disposable System for Monitoring Hemoglobin A1c. 2017 Presented at: American Diabetes Association; June 2017; 1701 N BEAUREGARD ST, ALEXANDRIA, VA 22311-1717, United States p. A-231 URL: https://ptsdiagnostics.com/wp-content/uploads/2018/08/lit_001815_r0_abstract_multi_center_accuracy_a1cnow_system.pdf
41. Nathan DM, Griffin A, Perez FM, Basque E, Do L, Steiner B. Accuracy of a point-of-care hemoglobin A1c assay. J Diabetes Sci Technol 2019;13(6):1149-1153 [FREE Full text] [doi: 10.1177/1932296819836101] [Medline: 30943787]
42. Toobert DJ, Hampson SE, Glasgow RE. The summary of diabetes self-care activities measure: results from 7 studies and a revised scale. Diabetes Care 2000;23(7):943-950. [doi: 10.2337/diacare.23.7.943] [Medline: 10895844]
43. Vincent D, McEwen MM, Pasvogel A. The validity and reliability of a Spanish version of the summary of diabetes self-care activities questionnaire. Nurs Res 2008;57(2):101-106. [doi: 10.1097/01.NNR.0000313484.18670.ab] [Medline: 18347481]
44. Ritter PL, Lorig K, Laurent DD. Characteristics of the Spanish- and English-Language self-efficacy to manage diabetes scales. Diabetes Educ 2016;42(2):167-177. [doi: 10.1177/0145721716628648] [Medline: 26846336]
45. Victorson D, Banas J, Smith J, Languido L, Shen E, Gutierrez S, et al. eSalud: designing and implementing culturally competent ehealth research with latino patient populations. Am J Public Health 2014;104(12):2259-2265. [doi: 10.2105/AJPH.2014.302187] [Medline: 25320901]
46. Williams GC, Freedman ZR, Deci EL. Supporting autonomy to motivate patients with diabetes for glucose control. Diabetes Care 1998;21(10):1644-1651. [doi: 10.2337/diacare.21.10.1644] [Medline: 9773724]

XSL•FO
RenderX

47. Beléndez M, Hernández-Mijares A, Marco J, Domínguez JR, Pomares FJ. Validation of the Spanish version of the problem areas in diabetes (PAID-SP) scale. Diabetes Res Clin Pract 2014;106(3):e93-e95. [doi: 10.1016/j.diabres.2014.09.012] [Medline: 25451898]
48. Polonsky WH, Anderson BJ, Lohrer PA, Welch G, Jacobson AM, Aponte JE, et al. Assessment of diabetes-related distress. Diabetes Care 1995;18(6):754-760. [doi: 10.2337/diacare.18.6.754] [Medline: 7555499]
49. Bowen DJ, Kreuter M, Spring B, Cofta-Woerpel L, Linnan L, Weiner D, et al. How we design feasibility studies. Am J Prev Med 2009;36(5):452-457 [FREE Full text] [doi: 10.1016/j.amepre.2009.02.002] [Medline: 19362699]
50. Wagner J, Bermudez-Millan A, Damio G, Segura-Perez S, Chhabra J, Vergara C, et al. Community health workers assisting Latinos manage stress and diabetes (CALMS-D): rationale, intervention design, implementation, and process outcomes. Transl Behav Med 2015;5(4):415-424 [FREE Full text] [doi: 10.1007/s13142-015-0332-1] [Medline: 26622914]
51. Budman SH, Soldz S, Demby A, Feldstein M, Springer T, Davis MS. Cohesion, alliance and outcome in group psychotherapy. Psychiatry 1989;52(3):339-350. [doi: 10.1080/00332747.1989.11024456] [Medline: 2772092]
52. Wagner J, Bermudez-Millan A, Buckley T, Buxton OM, Feinn R, Kong S, et al. A randomized trial to decrease risk for diabetes among cambodian Americans with depression: intervention development, baseline characteristics and process outcomes. Contemp Clin Trials 2021;106:106427 [FREE Full text] [doi: 10.1016/j.cct.2021.106427] [Medline: 33957272]
53. Toscos T, Daley C, Heral L, Doshi R, Chen YC, Eckert GJ, et al. Impact of electronic personal health record use on engagement and intermediate health outcomes among cardiac patients: a quasi-experimental study. J Am Med Inform Assoc 2016;23(1):119-128 [FREE Full text] [doi: 10.1093/jamia/ocv164] [Medline: 26912538]
54. Gonzalez JS, Krause-Steinrauf H, Bebu I, Crespo-Ramos G, Hoogendoorn CJ, Naik AD, GRADE Research Group. Emotional distress, self-management, and glycemic control among participants enrolled in the glycemia reduction approaches in diabetes: a comparative effectiveness (GRADE) study. Diabetes Res Clin Pract 2023;196:110229 [FREE Full text] [doi: 10.1016/j.diabres.2022.110229] [Medline: 36549506]
55. Skinner TC, Joensen L, Parkin T. Twenty-five years of diabetes distress research. Diabet Med 2020;37(3):393-400. [doi: 10.1111/dme.14157] [Medline: 31638279]
56. Fernández-Rodríguez R, Zhao L, Bizzozero-Peroni B, Martínez-Vizcaíno V, Mesas AE, Wittert G, et al. Are e-health interventions effective in reducing diabetes-related distress and depression in patients with type 2 diabetes? A systematic review with meta-analysis. Telemed J E Health 2024;30(4):919-939. [doi: 10.1089/tmj.2023.0374] [Medline: 38010739]
57. Perrin NE, Davies MJ, Robertson N, Snoek FJ, Khunti K. The prevalence of diabetes-specific emotional distress in people with type 2 diabetes: a systematic review and meta-analysis. Diabet Med 2017;34(11):1508-1520. [doi: 10.1111/dme.13448] [Medline: 28799294]
58. Clark TL, Gallo L, Euyoque JA, Philis-Tsimikas A, Fortmann A. Does diabetes distress influence clinical response to an mHealth diabetes self-management education and support intervention? Diabetes Educ 2020;46(3):289-296 [FREE Full text] [doi: 10.1177/0145721720913276] [Medline: 32228288]
59. Young HM, Miyamoto S, Dharmar M, Tang-Feldman Y. Nurse coaching and mobile health compared with usual care to improve diabetes self-efficacy for persons with type 2 diabetes: randomized controlled trial. JMIR Mhealth Uhealth 2020;8(3):e16665 [FREE Full text] [doi: 10.2196/16665] [Medline: 32130184]
60. Qin W, Blanchette JE, Yoon M. Self-efficacy and diabetes self-management in middle-aged and older adults in the United States: a systematic review. Diabetes Spectr 2020;33(4):315-323 [FREE Full text] [doi: 10.2337/ds19-0051] [Medline: 33223769]
61. Vluggen S, Hoving C, Schaper NC, De Vries H. Psychological predictors of adherence to oral hypoglycaemic agents: an application of the ProMAS questionnaire. Psychol Health 2020;35(4):387-404 [FREE Full text] [doi: 10.1080/08870446.2019.1672873] [Medline: 31588778]
62. Trief PM, Teresi JA, Eimicke JP, Shea S, Weinstock RS. Improvement in diabetes self-efficacy and glycaemic control using telemedicine in a sample of older, ethnically diverse individuals who have diabetes: the IDEATel project. Age Ageing 2009;38(2):219-225. [doi: 10.1093/ageing/afn299] [Medline: 19171951]
63. Grant RW, Wald JS, Schnipper JL, Gandhi TK, Poon EG, Orav EJ, et al. Practice-linked online personal health records for type 2 diabetes mellitus: a randomized controlled trial. Arch Intern Med 2008;168(16):1776-1782 [FREE Full text] [doi: 10.1001/archinte.168.16.1776] [Medline: 18779465]
64. McCarrier KP, Ralston JD, Hirsch IB, Lewis G, Martin DP, Zimmerman FJ, et al. Web-based collaborative care for type 1 diabetes: a pilot randomized trial. Diabetes Technol Ther 2009;11(4):211-217 [FREE Full text] [doi: 10.1089/dia.2008.0063] [Medline: 19344195]
65. van Vugt M, de Wit M, Sieverink F, Roelofsen Y, Hendriks SH, Bilo HJG, et al. Uptake and effects of the e-vita personal health record with self-management support and coaching, for type 2 diabetes patients treated in primary care. J Diabetes Res 2016;2016:5027356 [FREE Full text] [doi: 10.1155/2016/5027356] [Medline: 26955640]
66. Graetz I, Huang J, Muelly ER, Fireman B, Hsu J, Reed ME. Association of mobile patient portal access with diabetes medication adherence and glycemic levels among adults with diabetes. JAMA Netw Open 2020;3(2):e1921429 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.21429] [Medline: 32074289]
67. Lau M, Campbell H, Tang T, Thompson DJS, Elliott T. Impact of patient use of an online patient portal on diabetes outcomes. Can J Diabetes 2014;38(1):17-21. [doi: 10.1016/j.jcjd.2013.10.005] [Medline: 24485208]

XSL•FO
RenderX

68. Reed M, Huang J, Graetz I, Brand R, Hsu J, Fireman B, et al. Outpatient electronic health records and the clinical care and outcomes of patients with diabetes mellitus. Ann Intern Med 2012;157(7):482-489 [FREE Full text] [doi: 10.7326/0003-4819-157-7-201210020-00004] [Medline: 23027319]
69. Zocchi M, Robinson S, Ash A, Vimalananda VG, Wolfe HL, Hogan TP, et al. Patient portal engagement and diabetes management among new portal users in the veterans health administration. J Am Med Inform Assoc 2021;28(10):2176-2183 [FREE Full text] [doi: 10.1093/jamia/ocab115] [Medline: 34339500]
70. Tieu L, Sarkar U, Schillinger D, Ralston JD, Ratanawongsa N, Pasick R, et al. Barriers and facilitators to online portal use among patients and caregivers in a safety net health care system: a qualitative study. J Med Internet Res 2015;17(12):e275 [FREE Full text] [doi: 10.2196/jmir.4847] [Medline: 26681155]
71. Richwine C. Progress and persistent disparities in patient access to electronic health information. JAMA Health Forum 2023;4(11):e233883 [FREE Full text] [doi: 10.1001/jamahealthforum.2023.3883] [Medline: 37948063]
72. Lyles CR, Nelson EC, Frampton S, Dykes PC, Cemballi AG, Sarkar U. Using electronic health record portals to improve patient engagement: research priorities and best practices. Ann Intern Med 2020;172(11 Suppl):S123-S129 [FREE Full text] [doi: 10.7326/M19-0876] [Medline: 32479176]
73. Johnson AM, Brimhall AS, Johnson ET, Hodgson J, Didericksen K, Pye J, et al. A systematic review of the effectiveness of patient education through patient portals. JAMIA Open 2023;6(1):ooac085 [FREE Full text] [doi: 10.1093/jamiaopen/ooac085] [Medline: 36686972]
74. Casillas A, Abhat A, Mahajan A, Moreno G, Brown AF, Simmons S, et al. Portals of change: how patient portals will ultimately work for safety net populations. J Med Internet Res 2020;22(10):e16835 [FREE Full text] [doi: 10.2196/16835] [Medline: 33094732]

Abbreviations

CHC: community health center
CHW: community health worker
CONSORT: Consolidated Standards of Reporting Trials
DSM: diabetes self-management
DSMS: diabetes self-management education and support
GLMM: generalized linear mixed model
HbA_{1c}: hemoglobin A_{1c}
MAP: multilevel intervention to increase access and use of portals
PAID: Problem Areas in Diabetes scale
REDCap: Research Electronic Data Capture
SDoH: social determinants of health
T2D: type 2 diabetes
WHO: World Health Organization

Edited by A Mavragani; submitted 07.10.24; peer-reviewed by Y Lu, J Njeru, AN Ali; comments to author 14.02.25; revised version received 20.02.25; accepted 21.02.25; published 25.03.25.

Please cite as:
Whittemore R, Jeon S, Akyirem S, Chen HNC, Lipson J, Minchala M, Wagner J
Multilevel Intervention to Increase Patient Portal Use in Adults With Type 2 Diabetes Who Access Health Care at Community Health Centers: Single Arm, Pre-Post Pilot Study
JMIR Form Res 2025;9:e67293
URL: https://formative.jmir.org/2025/1/e67293
doi:10.2196/67293
PMID:40131327

©Robin Whittemore, Sangchoon Jeon, Samuel Akyirem, Helen N C Chen, Joanna Lipson, Maritza Minchala, Julie Wagner. Originally published in JMIR Formative Research (https://formative.jmir.org), 25.03.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Patients' and Health Care Professionals' Expectations of Virtual Therapeutic Agents in Outpatient Aftercare: Qualitative Survey Study

Diana Immel[1], DipFH; Bernhard Hilpert[2,3,4], MSc; Patricia Schwarz[5], MSc; Andreas Hein[5], PhD; Patrick Gebhard[2], PhD; Simon Barton[1*], PhD; René Hurlemann[1*], MSc, MD, PhD

[1]Department of Psychiatry, School of Medicine and Health Sciences, Carl von Ossietzky Universität Oldenburg, Oldenburg, Germany

[2]Affective Computing Group, Deutsches Forschungszentrum für Künstliche Intelligenz GmbH, German Research Center for Artificial Intelligence, Kaiserslautern, Germany

[3]Leiden Institute of Advanced Computer Science, Leiden University, Snellius Gebouw, Leiden, The Netherlands

[4]Department of Computer Science, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

[5]Assistance Systems and Medical Device Technology, Department for Health Services Research, School of Medicine & Health Sciences, Carl von Ossietzky Universität Oldenburg, Oldenburg, Germany

[*]these authors contributed equally

Corresponding Author:
Diana Immel, DipFH
Department of Psychiatry
School of Medicine and Health Sciences
Carl von Ossietzky Universität Oldenburg
Hermann-Ehlers-Str. 7
Oldenburg, 26160
Germany
Phone: 49 441961517
Fax: 49 44196151599
Email: diana.immel@uni-oldenburg.de

Abstract

Background: Depression is a serious mental health condition that can have a profound impact on the individual experiencing the disorder and those providing care. While psychotherapy and medication can be effective, there are gaps in current approaches, particularly in outpatient care. This phase is often associated with a high risk of relapse and readmission, and patients often report a lack of support. Socially interactive agents represent an innovative approach to the provision of assistance. Often powered by artificial intelligence, these virtual agents can interact socially and elicit humanlike emotions. In health care, they are used as virtual therapeutic assistants to fill gaps in outpatient aftercare.

Objective: We aimed to explore the expectations of patients with depression and health care professionals by conducting a qualitative survey. Our analysis focused on research questions related to the appearance and role of the assistant, the assistant-patient interaction (time of interaction, skills and abilities of the assistant, and modes of interaction) and the therapist-assistant interaction.

Methods: A 2-part qualitative study was conducted to explore the perspectives of the 2 groups (patients and care providers). In the first step, care providers (n=30) were recruited during a regional offline meeting. After a short presentation, they were given a link and were asked to complete a semistructured web-based questionnaire. Next, patients (n=20) were recruited from a clinic and were interviewed in a semistructured face-to-face interview.

Results: The survey findings suggested that the assistant should be a multimodal communicator (voice, facial expressions, and gestures) and counteract negative self-evaluation. Most participants preferred a female assistant or wanted the option to choose the gender. In total, 24 (80%) health care professionals wanted a selectable option, while patients exhibited a marked preference for a female or diverse assistant. Regrading patient-assistant interaction, the assistant was seen as a proactive recipient of information, and the patient as a passive one. Gaps in aftercare could be filled by the unlimited availability of the assistant. However, patients should retain their autonomy to avoid dependency. The monitoring of health status was viewed positively by both groups. A biofeedback function was desired to detect early warning signs of disease. When appropriate to the situation, a sense of humor in the assistant was desirable. The desired skills of the assistant can be summarized as providing structure and

XSL•FO
RenderX

emotional support, especially warmth and competence to build trust. Consistency was important for the caregiver to appear authentic. Regarding the assistant–care provider interaction, 3 key areas were identified: objective patient status measurement, emergency suicide prevention, and an information tool and decision support system for health care professionals.

Conclusions: Overall, the survey conducted provides innovative guidelines for the development of virtual therapeutic assistants to fill the gaps in patient aftercare.

(JMIR Form Res 2025;9:e59527) doi:10.2196/59527

KEYWORDS

socially interactive agent; e-mental health; mental illness; mental disorder; depression; major depressive disorder; suicide prevention; suicidal ideation; outpatient aftercare; artificial intelligence; virtual therapeutic assistant; public health; digital technology; digital intervention; digital health care

Introduction

Background

As stated by the World Health Organization, depressive disorders are among the most serious mental illnesses. They represent a serious loss of societal potential through premature mortality. It is estimated that 4.4% of the global population experiences depressive disorders. This is an increase of 18% compared to more than a decade ago [1]. The lifetime prevalence of depressive disorders is estimated to be between 16% and 20% [2]. Depressive disorders represent the most common mental health condition associated with suicide. It is estimated that 60% to 70% of individuals experience suicidal ideation during depressive episodes. The risk of suicide is 30 times higher than that observed in the general population. The mortality rate within the first year is approximately 3.4 times that of the general population. Furthermore, the probability of a depressive episode recurring within the first year is approximately 70%. In addition, approximately 21% of patients are readmitted to the hospital within 1 year of discharge [3]. The initial 6-month period following discharge is of particular significance with regards to the potential for relapse and suicide [4,5]. In contrast with the recommendations set forth in the German S3 guideline for unipolar depression, it is observed that 90% of patients diagnosed with major depression do not receive sufficient outpatient aftercare following psychiatric hospitalization [2]. This elucidates the elevated risk of relapse and suicide during the initial year [3]. Patients with restricted access to social, economic, or structural resources are particularly susceptible to these risks [2]. This conspicuous deficiency in outpatient care results in the forfeiture of potential therapeutic benefits that could otherwise be achieved through inpatient care. Conversely, enhancing outpatient care for patients with depression has been demonstrated to result in improved survival rates. In the context of this study, the term health care professionals refers to psychiatrists, psychotherapists, nurses and social workers who treat patients after they have been discharged from the hospital. The responsibility for patients with suicidal tendencies represents the aspect of working with depressed individuals that evokes the strongest emotional responses from this group of health care professionals and those who have previously cared for such patients. The experience of losing of a patient to suicide can have a substantial and lasting negative emotional impact. The risk of suicide among practitioners increases substantially following the suicide of a patient. In one study, 10% of participants, who were health care professionals, reported having suicidal thoughts themselves [6,7]. Thus, solutions that enhance the security of particularly susceptible patients, such as effective posthospital psychiatric follow-up, can consequently enhance the lives of patients and their families, as well as health care professionals, by minimizing occupational stress.

Before discharge, health care professionals need to assess the patients' resilience in this regard and share responsibility. The decision-making process is often influenced by other factors, such as insurance coverage. Similarly, the factors that may lead to individual crisis situations are difficult to predict; therefore, postdischarge risk is also a professional burden for caregivers [8,9]. Although some patients continue to receive outpatient care after discharge, the sudden change in their situation poses several problems. One of the most prevalent problems seems to be the lack of a daily structure in time and tasks throughout the day, as provided by the hospital during inpatient care. Instead, patients are mostly responsible for their own daily structure, a task that many struggle with as their goal setting and persistence are not necessarily stable enough to overcome these hurdles. Conversely, in moments of self-doubt and struggle, they now have limited access to timely support. In the hospital, a safe environment with a wide range of support is available, especially in critical situations. However, this changes after discharge [10]. Outpatient psychiatric aftercare, when available, is limited and does not cover all critical situations that prevail in suicidal moments [11]. The limited number of health care professionals, together with time and emotional resource constraints and unclear responsibilities, pose a serious problem for patients. In addition, another source of support regarding social companionship, such as friends or family [12], may not always be available. More importantly, relatives are not professionally trained to deal with critical and potentially suicidal behavior. Some patients may even feel ashamed. In general, patients are left alone after discharge.

It appears that a gap exists in the system for outpatient care in critical situations, particularly with regards to the availability of health care professionals, the definition of responsibilities, and the lack of qualifications in many cases [2]. In the absence of in-person assistance, face-to-face support has declined and an increased demand for online support has been observed. This indicates a necessity for the development of secure, safe, and effective digital tools. Nevertheless, research indicates that internet-based cognitive behavioral therapies have a relatively modest impact on patients with suicidal tendencies [13]. Patients

XSL•FO
RenderX

who continue to experience suicidal thoughts and behaviors also present a unique challenge for electronic mental health (e-mental health) services. Therefore, this group is explicitly excluded from e-mental health services [14]. A variety of interventions are available, including web-based psychoeducation, online counseling, SMS text messaging aftercare, email therapy, and computer-assisted cognitive behavioral therapy [15,16]. However, these interventions fail to address patients' need for attachment, or for the formation of a social, empathic relationship. This has been described as the most important variable in determining the effectiveness of psychotherapy [17,18]. Consequently, the lack of a therapeutic relationship is a substantial criticism expressed by patients when evaluating internet-based cognitive behavioral therapy self-help methods [19].

It is thus evident that an intermediate professional solution is required to address the gap in care for patients with depression and suicidality, with the objective of providing support to both this susceptible patient population and their health care professionals. The solution should be readily accessible without overwhelming amounts of information, focus on depression-related cognitive and emotional impairments, and include suicide prevention. It should also be noted that using such solution, people in underserved rural or economically less developed regions could be better reached, overcoming socio-spatial barriers. Furthermore, the stress on the therapists in charge could be greatly reduced, thus contributing to staff stability in psychiatric institutions.

Using Socially Interactive Agents for Outpatient Aftercare

Socially interactive agents (SIAs) can engage in verbal and nonverbal interactions as dialogue partners, exhibiting emotional and social capabilities and performing a diverse social support and monitoring functions [20-23], including provision of emotional support to users [24], monitoring of patient safety [25] and simulation of social skills to emulate social relationships, such as active listening, mimicry, gestures and emotion models [26-29]. Given the crucial role of social relationships in coping with psychological distress [30], there is the potential for SIAs to enhance technology-assisted outpatient care by emulating human interactions and tasks [31].

To date, SIAs have been used in a variety of settings as virtual coaches and assistants. Unified platforms for SIAs as coaches have been developed [32-34] and used for stress management purposes [35]. However, they have also been used in applied contexts as training partners for social skills [36] and empathy training [37]. In clinical research, initial approaches to the use of empathic conversational agents in the context of mental health have been proposed [38]. Furthermore, the use of relational agents has been demonstrated to have a positive impact on hospitalized patients with depressive disorders [39]. These findings suggest that these approaches may offer promising possibilities for the use of a virtual therapeutic assistant (VTA) as an assistant in a supportive, gap filling role in outpatient aftercare. It is unclear which settings a VTA could be used in and what role it should play. There is also a lack of clarity regarding the design of the interface for interaction with patients or health care professionals.

In SIAs, embodiment is of substantial consequence with regard to the manner in which they are perceived. While technology-mediated solutions can have a supportive effect, as previously outlined, communicating with a disembodied, faceless machine can impede positive outcomes, even when a human is on the other end. Conversely, embodiment in SIAs has been demonstrated to influence perceptions [40-42], increase trust [43,44], facilitate the formation of an emotional connection, and enhance overall relationship quality [45].

With regards to interpersonal interaction, a substantial body of research has examined timing, skills, and behavior of SIAs. It has been demonstrated that SIAs can evoke emotional responses that are pertinent to interpersonal interaction and attachment. The timing of SIAs behavior has been shown to be of critical importance in the formation of human-agent relationships [46,47]. In a job interview setting, SIAs were found to elicit feelings of shame to the same extent as human interviewers [48]. Moreover, analogous effects have been demonstrated with respect to trust [49] and rapport through backchanneling behavior [22,23,50], the elicitation of empathetic behavior in human physicians when dealing with virtual patients [51], and the elicitation of prosocial behavior in interpersonal interaction settings [52-54].

An embodied, affective SIA has the potential to facilitate novel avenues for relationship-based, technology-mediated outpatient aftercare. Nevertheless, its application to date has been confined to the context of scientific studies or hospitalization. The context of outpatient aftercare, which is particularly sensitive due to its applied nature, requires a meticulous and precise implementation of the VTA for it to have the intended supportive effect.

Research Questions

Overview

This study presents a comprehensive analysis of the expectations held by both patients and health care professionals regarding the design of a proposed tool. The objective of the survey was to ascertain both expectations regarding the structural framework of how a VTA might interact with patients and regarding the interaction itself.

Role

The initial stage of the process entailed an evaluation of the potential integration of a VTA within the existing framework of outpatient care. It is necessary to establish the role that both groups anticipate a VTA will play in the context of outpatient care.

- Research question (RQ) 1: What is the role of a VTA in the context of outpatient care?

Embodiment

A substantial body of research and practical applications outlined previously indicates that the favorable outcomes of SIA interaction are contingent upon the manner of embodiment within the interaction interface.

XSL•FO
RenderX

- RQ 2: What are the embodiment preferences for a VTA?

Interaction Timing

The objective of this study was to address the expectations of patients and health care professionals regarding the interaction between a VTA and a patient. As no previous testing of this framework has been conducted, it is unclear what frequency and timing of interaction might prove beneficial for the patient.

- RQ 3: When should the VTA interact with the patient?

Interaction Abilities

Although a VTA offers numerous advantages to the patient, the specific interaction capabilities that the VTA should possess to be beneficial for treatment remain uncertain.

- RQ 4: What interaction capabilities should the VTA have to assist patients in their treatment?

Interaction Behavior

The following section will present the observed interaction behaviors and behaviors associated with the use of the aforementioned skills during the interaction.

- RQ 5: What interaction behaviors should the VTA use to facilitate patient treatment?

Interaction Between VTAs and Health Care Professionals

For the VTA to be considered an integral part of the treatment framework, it would be beneficial for it to provide useful information to the health care professionals themselves.

- RQ 6: How might the framework be designed in a way that would allow health care professionals at all levels (ie, psychiatrists, psychotherapists, nurses, and social workers) to benefit from the implementation of a VTA?

Methods

Overview

To investigate these questions, a 2-part qualitative study was conducted. The objective was to explore the perspectives of the 2 groups that were most directly involved (health care professionals and patients) and to assess their expectations of a VTA in the context of treatment. Initially, a group of health care professionals were queried about their expectations regarding the potential role of the assistant (RQ 1), assistant-patient interface (RQ 2), timing of the interaction (RQ 3), interaction skills (RQ 4), interaction behavior (RQ 5), and assistant–health care interaction (RQ 6). Following a comprehensive analysis of the responses, a qualitative interview was conducted with a group of patients undergoing treatment. This will provide a 360-degree perspective of the requirements that a socially interactive VTA should fulfill in a specific clinical setting to address the identified gap in outpatient aftercare. The methods and results are presented in parallel for the sake of comparability.

Participants

In the initial phase of the study, data were gathered from a sample of health care professionals (n=30; 6 male and 24 female participants) employed in various outpatient settings within the field of outpatient care in social psychiatry in Oldenburg (Germany). All health care professionals participated voluntarily, which was ensured by providing them detailed information about the objectives and implementation of the study. No data were excluded. The age of the therapists ranged from 18 to 24 years (n=1), 25 to 34 years (n=13), 35 to 44 years (n=6), 45 to 54 years (n=4), 55 to 64 years (n=5), and ≥65 years (n=1), and their professions included psychiatrist (n=3), psychotherapist (n=10), social worker (n=15), psychiatric nurse (n=1), and physician (n=1).

The patient sample consisted of individuals (n=20; 11 male and 9 female participants) who were inpatients at the Karl-Jaspers Hospital in Oldenburg and were diagnosed with depression. All patients participated on a voluntary basis, which was ensured by providing detailed information about the objectives and implementation of the study. No data were excluded from the analysis. The age of the participants ranged from 19 to 56 years (mean 40.85, SD 12.46 years). Patients were administered the Beck Depression Inventory II (mean 24.3, SD 13.45; Beck et al [55]) to obtain an overview of their current depression status. To assess trust, the Yamagishi and Yamagishi General Trust Scale (mean 3.19, SD 0.72; [56]) was used, which comprises 7 statements, including “Most people are basically honest” with a rating scale from 1 (low) to 5 (high agreement), Cronbach α=0.83.

Ethics Approval

Before data collection, approval was obtained from the Medical Ethics Committee of the University of Oldenburg (2022-005). Participants provided written informed consent to participate in this study. Data were anonymized. The participants received no compensation for this study.

Procedure

At an informative meeting of health care professionals in our clinic, general information about the purpose of the study and details of the survey were delivered. Following the provision of a link via email, they were invited to complete a qualitative web-based questionnaire (implemented in Qualtrics) comprising a combination of multiple-choice and open questions (described in detail in subsequent sections).

The patient cohort was recruited through the clinical routine of their daily inpatient therapeutic examinations at the same psychiatric hospital. A semistructured qualitative interview was conducted in a confidential session. Following the introduction, a trained interviewer proceeded to elucidate the objective and structure of the forthcoming interview. It was emphasized that the responses provided by the participants would not be shared with any of the other patients and therapeutic staff.

Materials

To ensure that all participants had the same level of knowledge about VTAs, a standardized description was first provided, which was adapted from the description set out by Hilpert et al [42,44]. The questionnaire was constructed in accordance with the 6 distinct expectations (corresponding to the 6 RQs) pertaining to a VTA in outpatient aftercare. The questions

XSL•FO
RenderX

themselves were divided into 5 blocks, each addressing a specific area of inquiry: general expectations, specific situations, usefulness, communication and data handling, and demographics. The web-based questionnaire and the interview were comparably structured and comprised the same set of questions. However, the terminology and specific questions were tailored to align with the perspective of each respective group.

Health Care Questionnaire

The questionnaire for health care professionals consisted of several sections. The initial section of the questionnaire addressed general inquiries pertaining to the health care professionals' anticipated interface, capabilities, interaction behavior, and timing (eg, how and when to contact and how and when to support) of the VTA. In addition, it encompassed broader expectations related to the role of the VTA in the therapeutic process. Second, 2 particular scenarios in which patients may find themselves were delineated, accompanied by corresponding inquiries. In the initial scenario, health care professionals were instructed to envision a typical day in the life of an outpatient, beginning to end. Regarding competencies, the inquiries centered on the diagnostic abilities and subject matter that the VTA should possess to obtain a comprehensive understanding of the patient's condition. The question pertaining to interaction timing addressed the optimal timing for the VTA to interact with the patient, while the questions concerning interaction behavior sought to ascertain the types of behaviors that health care professionals deemed beneficial. The second scenario pertains to an emergency situation that may arise during a day after discharge. The inquiry commenced with queries pertaining to the VTA's capacity to procure external assistance for the patient and how the VTA could facilitate this. Subsequently, the participants were asked to indicate the circumstances under which they would interact with the VTA and to describe the helpful types of behavior. Third, the participants were requested to evaluate the utility of a VTA in relation to a series of therapeutic functions, including medication planning and crisis intervention. Fourth, we investigated their perceptions of their own interaction with the VTA, specifically regarding the type of information they would find beneficial if the VTA were to provide it to health care professionals involved in outpatient care. Finally, the participants were asked to provide demographic information.

Patient Interview

The introductory information and briefing remained consistent throughout the process. As the interview with the patient sample was conducted after the analysis of the health care professionals responses, some of the questions were adapted based on the findings from this initial sample. This was done to elicit more detailed and nuanced responses and to avoid potentially vague responses. The overall structure of the questionnaire remained consistent. The questionnaire used in the interview encompassed the elements of RQ 1 to 5, excluding the interaction between the assistant and the therapist (as this did not impact the patients). The questionnaire was divided into 6 blocks: general expectations, assistant-patient-interaction, assistant-health care professional-interaction, feelings of the patient and demographics.

The interview commenced with a series of general inquiries pertaining to patients' expectations of a VTA's competencies, interaction style (including the modalities of contact, forms of support, and the VTA's presence in the patient's daily life), and a few more detailed questions concerning appearance. Subsequently, a description of 2 particular scenarios in which patients might find themselves was provided, accompanied by corresponding questions. Each specific situation was introduced with a subjective assessment of how the patients would feel in that particular situation. The questions were largely similar to those posed to the health care professionals, but the patients were also asked to provide an explanation for their responses. The section on usefulness was identical to that used with the health care professional sample.

In accordance with the recommendations for open research practices [57], all materials can be found on the OSF website [58].

Data Analysis

To extract meaningful insights for each of the RQs in this qualitative dataset, a 4-step process based on grounded theory was used to analyze the responses given by both groups of interest [59]. This was executed following the guidelines described by Hansen and Grimshaw-Aagaard [60]. In the initial phase, the responses from each group of interest were transcribed (for the patient interview data) and organized by question in 2 distinct datasets, one for each group. Second, 3 independent raters classified the responses according to content and relevance to the RQs. Third, the raw data and extracted categories were compared for each group of interest (health care professionals and patients), identifying both similarities and differences between the 2 groups. Finally, this comparative dataset was classified to extract clusters and subclusters of explanatory value for each RQ.

Results

This section presents the findings of the analysis of the health care professional and patient surveys. A joint analysis of both health care professional and patient responses is presented for each RQ.

RQ 1: Role of the VTA

When queried about the role of the VTA, health care professionals primarily referenced the role of a "good friend." It was emphasized that the VTA should be "designed in such a way that you enjoy the interaction, otherwise it's probably not going to be used." A similar argument was put forth with regard to the role of a social worker, which was described as "something between a good friend (feeling of closeness) and a social worker (knowledge about the disease and how to help)." Conversely, the explicit role of a VTA as a "therapist" was reported, with the addition that the VTA should be a figure with "distance to the family" or "not a substitute for family or friends," stating that "these should be present in real life or newly built." The VTA was recommended as an adjunct to existing therapy, with "more of an additional supportive role,"

XSL•FO
RenderX

with an emphasis on supporting the patient "in phasing out therapy." In a less role-specific description, the VTA was referred to as a "companion" or "extended inner support, more like a part of oneself that is currently too weak to provide self-support." Health care professionals also described the VTA as a "coach" or "someone the patient can confide in, but who remains independent." A more abstract idea was a "good spirit that no one else sees, who sees the positive and can sometimes be very direct but warm."

Patients referenced the role of a therapist, "a known person," as well as frequent references to a "good friend," "being there for me like a best friend," "asking questions about everyday life (instead of 'psychological' questions)." The wish for additional support in outpatient care is clear: "to work out goals and important topics with therapists, which the assistant knows about."

In general, both groups anticipated a system that integrates expertise and represents a figure with whom they can establish a relationship. In conjunction with the primary recommendation of the therapist's good friend, the role of a peer counselor would also be conceivable.

RQ 2: Embodiment of the Assistant-Patient Contact

Regarding the topic of embodiment, we inquired about the participants' expectations regarding general appearance characteristics of VTA. As previous research [40] has shown a clear pattern of human embodiment, we also asked about preferences for the assistant's age and gender. The question regarding general appearance was posed in an open-ended format, with the objective of obtaining an unbiased report from the participants. Most of the patients (14/20, 70%) related their responses to the human appearance of the assistant. Some described nonhuman appearance and characteristics, to which age and gender do not apply. Regarding general appearance, the responses of both health care professionals and patients who endorsed a humanlike appearance were classified and summarized into 2 relevant clusters that play a role in the perception of the assistant: *physical appearance* and *expressiveness*.

In *physical appearance*, the health care professionals' interview focused on 3 features: personal characteristics, clothing, and animation style. For physical characteristics, they tended to be the least risky, such as "medium height," "normal weight," "friendly face," and "calm speech." The same pattern was observed for clothing, as shown by responses such as "[the assistant] wears clothes in pleasant pastel shades, nothing too dark or too colorful." The third feature could be defined as animation style, with health care professionals agreeing that the assistant should appear "as realistic as possible" with "no cartoonish features." Regarding gender, 80% (24/30) of the health care professionals wanted a selectable option. The remaining participants recommended either a female assistant (3/30, 10%) or a diverse (3/30, 10%) assistant, while male gender was not explicitly recommended. Regarding age, a "middle-aged" assistant was recommended, as it should "exude life experience," "not too young for older people." Other recommendations were "age depending on the patient, maybe a bit older" or that the age should be "chosen by the patient."

For the *expressiveness* cluster, they focused on facial expressions and gestures that should be convincing enough for the assistant to be able to "respond to feelings and show empathy" as well as "natural movements."

A different set of priorities for the *physical appearance* of a VTA was observed by the patients. While they wanted the assistant to have "an upper body and hands, but no legs," it should "not have a perfect figure," with the same mentioned for the face. Clothing was not mentioned. For animation style, they reported a tendency toward a more animated style "like in a computer game" or "like in science fiction movies." However, in contrast to health care professionals, patients expressed a desire for familiarity, as the assistant should look like a known person, for example, the caregiver. Regarding gender, there was a clear and predominant preference of all participants for a female or diverse character. The reasons provided for this preference were "I am a woman myself," "I can talk better to women," "Navi, Siri are also women" and "it is easier to build trust." In cases where patients did not explicitly prefer a female character, responses such as "voice is more important (soft, calm voice)" were given. It should be emphasized that none of the patients explicitly asked for a male embodiment for their VTA. Regarding the age of the VTA, "middle-aged," "older," but also contradictory answers, such as "between teenager and young adult" were provided. However, there was a tendency to favor a middle-aged figure to older figure. For *expressiveness*, it was desired that the assistant should be able to use facial expressions and gestures "so that emotions can also be clearly expressed" (clear facial expressions). Overall, they wanted "friendly" and "laughing" expressions.

Finally, there were some health care professionals who preferred a nonhuman appearance, citing concerns about "negative transference" or being "triggered" by a human figure. One participant illustrated:

> *I would like it if the character didn't necessarily try to look like a human or even a therapist. I would be afraid that the difference would be too great and the character would look unbelievable or even ridiculous. I would therefore suggest that anthropomorphized, sympathetic avatars be created that patients could easily cast in a positive light.* [Psychotherapist #35-44, male]

Suggestions for this alternative, nonhuman appearance included a talking smiley instead of a real human face, or a "wise but humorous owl." A few patients had different nonhuman suggestions, such as "blue water on the beach," "self-selectable (robot, human, animal...)," "between emoji and human" and "can give itself a shape."

These findings indicate that both groups anticipated an imperfect or neutral-looking but emotionally expressive human embodiment for their VTA, with a tendency to avoid male characters.

RQ 3: Interaction Timing

The third RQ pertained to the timing of the interaction: when should the assistant interact with the patient? The joint analysis of the responses from both groups revealed 2 main clusters of

XSL•FO
RenderX

situations in which the VTA could potentially interact with patients.

The first cluster was designated *situation-independent or proactive* and refers to behavior that does not depend on the socioemotional context of the patient at the time of interaction. Situation-independent or proactive interaction timing does not necessitate a request for assistance from the patient or the automatic recognition of an ongoing crisis situation by the assistant. This encompasses any form of interaction that could be proactively initiated by the assistant, including both scheduled contact and unscheduled check-in messages. For example, health care professionals proposed that the assistant could contact patients in the morning to "help start the day" and in the evening to "reflect on the day." In between, at "agreed times" and "at random times." Patients corroborated this assertion by requesting regular contact "maybe 1-2 times a day," "daily contact," "regularly in the morning," "check in 3-4 times a day."

The second cluster was designated *situation-dependent* and represents a complementary counterpart to the first. It includes all instances wherein the interaction is triggered by a response to the patient's current therapeutic or socioemotional circumstances. This could manifest as the patient requesting contact, for example, to talk about a feeling at any time of the day, or even in crisis situations, such as a panic attack. It could also be triggered by monitoring the patient's current stress indicators (eg, heart rate and lack of interaction) to detect an emergency situation and trigger a *situation-dependent* interaction in the assistant. Responses that were clustered as *situation-dependent* differed between health care professionals in the frequency with which this should occur. For example, patients mentioned "more often immediately after hospitalization" and that the assistant should contact them "on holidays/at weekends." At this point, patients would like to be able to personalize the assistant's presence according to their own needs. Patients would also like to be notified "as soon as he notices that I am not feeling well" or "if I have been inactive for a long time."

Overall, health care professionals' responses to the "when" question exhibited considerable variability. Regular or daily contact, fixed times, and contact when the patient is inactive were all represented to a similar extent. Although patients emphasized "unstructured" times when they expected to feel alone, there were no substantial differences between the responses of patients and health care professionals. These numerous variabilities may indicate the importance of an individual rhythm of contact.

RQ 4: Interaction Abilities

This RQ was concerned with interaction skills: what competencies should the VTA possess to support patients in their treatment? The analysis revealed 2 main clusters of interaction skills that a VTA should possess in general: the capacity to furnish structural support and the capacity to furnish emotional support. Given the considerable overlap between the clusters extracted from the joint analysis of the responses of both groups of interest, it was determined that there was no additional benefit in reporting the responses of health care professionals and patients separately in this subsection.

The first cluster was defined as *structure-providing* capabilities, encompassing all the interaction capabilities of the VTA that facilitate structure for the patient regarding behavior (proposing action strategies or step-by-step plans for upcoming situations, monitoring the patient's emotional state, and providing feedback) and time management (eg, structuring the day and reminder functions) to enable coping with daily challenges or tasks.

The second cluster was described as *emotional-supportive*, comprising skills that assist patients in managing their emotions in a relational context. This cluster comprised 3 subclusters: *supportive-trusting,supportive-motivational,* and *supportive-resource-promoting.* The capacity to be *supportive-trusting* enables the assistant to facilitate trusting conversations by being honest, authentic, interested, sensitive, nonintrusive, and reassuring. The capacity to be *supportive-motivational* enables the assistant to promote a positive attitude and increase self-esteem by acknowledging courage, praising achievements, and generally providing positive messages. The capacity to be supportive-resource-promoting enables the VTA to expand knowledge, increase self-efficacy and self-reflection, and develop and apply skills. Rather than giving instructions, this cluster emphasizes the promotion of patient autonomy through the development of self-management skills.

The results of RQs 3 and 4 provide a 2-dimensional conceptual space for interpreting participants' responses. The first dimension characterizes the situational aspect of interaction timing and encompasses the clusters of *situation-independent or proactive* and *situation-dependent* expectations regarding the VTA. The second dimension pertains to the function of the interaction and encompasses the 2 clusters of expectations: one pertaining to *structure-providing* skills and the other to *emotional-supportive* skills. The *emotional-supportive* cluster may be further subdivided into 3 facets: *supportive-trusting*, *supportive-motivational*, and *supportive-resource-promoting*. This resulted in a total of 8 response subclusters, which can be used to categorize the interaction and the interactional behavior displayed. In addition, a further category of responses was identified as *taboo topics* representing a set of issues that should not be discussed. Figure 1 provides an overview of the dimensions and clusters with their subclusters, including individual definitions.

XSL•FO
RenderX

Figure 1. An overview of a 2D concept with clusters and subclusters in the interaction abilities of a virtual therapeutic assistant (VTA) in outpatient care. The definitions used to interpret the responses of participants (N=50, including health care professionals and patients with depression) in a semistructured qualitative study.

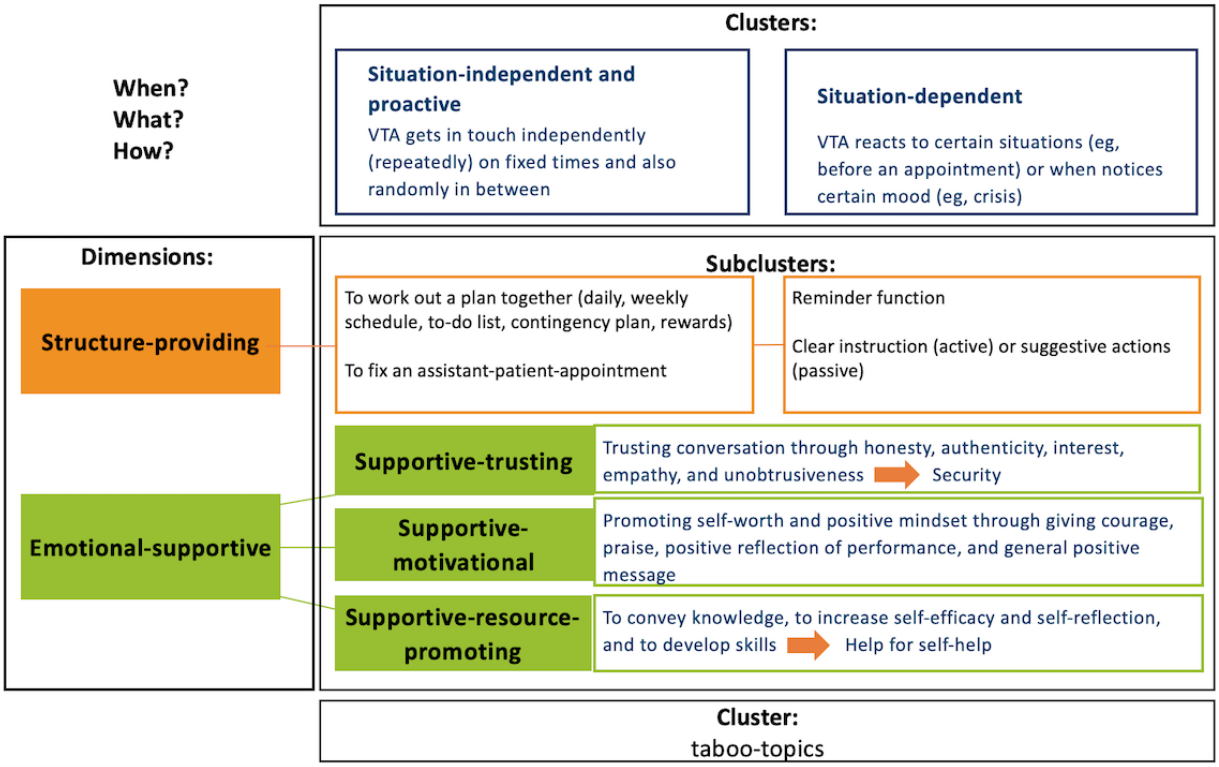

RQ 5: Interaction Behavior

Overview

This RQ was concerned with interaction behaviors that the VTA should display to provide effective support to patients undergoing treatment. The responses are presented separately for each of the 8 subclusters of the conceptual space as described in the RQ 4: Interaction Abilities section. However, the questionnaire included a specific section of questions pertaining to an emergency situation or breakdown, the responses to which differed considerably from those provided in the general sections. Therefore, crisis behavior is reported in a separate section below.

Regarding *situation-independent* or *proactive structure-providing* behavior, health care professionals responded that the assistant could collaborate with the patient to develop a plan for the structure of the day, that is, “suggestions for organizing the day,” “checking off to-do lists” as well as “validating mood, suggesting daily activities.” Patients indicated a preference for regular contact, including “daily contact” and “welcoming words.” In addition, patients expressed a desire for a mood or habit tracker as well as daily advice.

For *situation-dependent structure-providing* behavior (ie, in situations of distress or instability), health care professionals recommended reminders, concrete instructions, or suggestions for action. These included “behavioral training,” “reminder of positive activities,” and “taking medication.” In addition, patients requested task reminders, eg, “appointments,” instructions on specific markers (eg, “sleep quality”) and “psychoeducation.”

Overall, health care professionals overwhelmingly stated that the assistant should remind patients of the strategies they have learned and provide them with positive support. Patients indicated a preference for the assistant to provide guidance and support through the provision of helpful suggestions and emotional support. There were no major differences between the responses of the 2 groups. However, health care professionals repeatedly emphasized the importance of providing positive patient-emancipating support, whereas patients increasingly indicated that they would prefer the assistant to inquire about their condition.

Regarding *situation-independent* or *proactive supportive-trusting* behavior, the health care professionals indicated that authenticity was a crucial element: “no good coaxing,” “moderate criticism,” and “avoid empty phrases.” Moreover, the VTA should monitor the patient’s emotional state and inquire about their well-being, while avoiding excessive intrusion (“don’t ask the same probing questions every day” and “don’t lecture, don’t put pressure”). The patients requested a personal connection and a relational memory to “refer to past episodes.” In addition, patients requested displays of empathy and humor, as well as affirmations of positivity and emotional support “as empathetic as possible” or “help to keep my head above water.”

In regard to *situation-dependentsupportive-trusting* behavior, health care professionals emphasized the importance of honesty. A “perceived deterioration” in mood, “early warning signs” or “exceeding critical levels” should be communicated to the patient first. Moreover, the VTA should demonstrate interest by “listening when the patient tells you something on his/her

XSL•FO
RenderX

own" and "keep the conversation going." In addition, the assistant may proactively "engage the patient in conversation." After specific incidents, the VTA may inquire about the patient's experience by asking, "How did it go?" The assistant should demonstrate empathy and understanding, taking "negative or sad feelings seriously" and refraining from trivializing them. Unsolicited "funny" advice should be avoided. The patient should be queried as to whether they would prefer to discuss a "particular topic" in more detail or "at another time." The patient should be "relieved" and "reassured" as much as possible. Patients in this cluster expressed a desire for honesty, interest, "being there for me like a best friend," "empathy/understanding" and a sense of security "trusting that he will respond when I contact him."

Overall, the responses were not markedly divergent, with both groups viewing the support worker as a primary source of assistance. Health care professionals had a strong focus on "avoiding harm" in anything the assistant would do, while patients' responses reflected a fear of being left alone.

For the subsequent cluster, "situation-independent or proactive supportive-motivational behavior, health care professionals advised the presentation of positive or "thought-provoking" messages, "positive images," and "music." In addition, it was recommended that the daily routine be reflected upon by asking positive questions, such as "What are five things that have been good in the last week?" after which the VTA could summarize the successes to the patient. On the patient side, this cluster included promoting a positive attitude, "being supportive and encouraging," self-esteem-boosting content, such as "responding to the patient's stress and needs," "not being annoying or condescending," and general positive messages.

In the *situation-dependent* part of the *supportive-motivational* cluster, health care professionals mentioned "encouraging," "promoting positive thoughts" and "praising" in relation to upcoming activities or occasions. In addition, patients mentioned "encouraging, praising" and "reminding of learned structures" to reflect positively on their performance.

Overall, the responses converged on the request for positive supportive behaviors, with the difference that health care professionals asked for proactive and patients for reactive behaviors.

In the *supportive-resource-promoting* cluster, health care professionals most frequently mentioned "psychoeducation" as a situation-independent behavior, with the addition that "reference to online forums should be avoided." In addition, they suggested that calming techniques, such as "breathing exercises, mindfulness exercises" could be included. At this juncture, the transition to situation-dependent recommendations is fluid, as *supportive-resource-promoting* behaviors frequently prove effective for both types of timing. The health care professionals suggested that the assistant should help to "focus on the positive," "address brooding thoughts," and "motivate, show understanding." As an example, for panic attacks, breathing exercises could be used to help. For behaviors that are exclusively mentioned as situation-dependent, health care professionals mentioned providing explanations, for example, "in bad phases, the changeability of feelings and mood phases" could be explained with reference to individual examples from the patient's past. As this type of self-help promotion was considered particularly important by the health care professionals, an exemplary list is provided in Textbox 1.

Textbox 1. Examples of how a virtual therapeutic assistants can be used to promote self-help for patients with depression in the outpatient setting, which is considered particularly important by health care professionals (n=30) in a semistructured qualitative study.

Examples for self-help promotion:

- "Where could your bad mood come from?"
- "What has helped with low mood so far?"
- "What self-stabilization options do you have?"
- "Who could you call?"

Situation-independent reminders should also include the following:

- Knowing about or instructing, if not already available, the patient's skills, strategies, and stability factors
- Methods for dealing with circles of thought and rumination
- Defining emergency contacts together (making them quickly accessible)
- Informing about outpatient help offers (self-help groups, outpatient treatment, telephone counseling, specialists, and social psychiatric service)
- Building up activities (eg, maintaining social contacts) and providing inspiration for activities
- Supporting eating behavior (only if desired) and sleep hygiene
- Homework support

Patients in the *supportive-resource-promoting* cluster requested specific forms of assistance and guidance, including "offering autogenic training," psychoeducation, and dialogues—enabling conversations between the assistant and the patient as a situation-independent behavior. However, monologues and input from the assistant, such as "nice topics" and "providing facts and clear lines" were also requested. In the situational dimension, patients expressed a desire for biofeedback, skills and competencies—showing ways to solve problems," relational

XSL•FO
RenderX

memory—“referring to past episodes,” and active dialogue—“talking to me about what is bothering me.”

Overall, regarding *emotional-supportive* dimension and its 3 subclusters, practitioners indicated that they would use specific strategies and interactional behaviors to facilitate conversations through honesty, authenticity, interest, empathy, and unobtrusiveness to build trust and convey safety. To motivate, positive self-esteem and minimum standards should be achieved through encouragement, praise, positive reflection on achievements, and general positive messages. The objective is to provide knowledge, enhance self-efficacy and self-reflection, and cultivate skills. Patients tended to mention more functional characteristics at this point, whereas practitioners were more likely to cite the assistant’s capacity for empathic understanding.

A further category of responses was identified, namely a cluster of *tabootopics* that should not be discussed “because the questions might trigger and the patient would be left alone with negative feelings.” These included suicidal thoughts, weight gain, and possible crises that could not be detected, “risk of not being able to detect this in an individual way, or inadequate response to suicidality.” Nevertheless, some health care professionals advocated for the inclusion of all topics and the elimination of taboos, stating “I think there should be no taboos. Social interaction should play a big role, sexuality can also be mentioned.” Additional responses on this page indicated that: “In general, the character should address many topics so that the patient thinks about his condition” and has to “show his colors.”

A similar list of topics was identified in the patient responses. Topics that are deemed irrelevant or potentially “thought triggering” should be excluded from discussion. Furthermore, certain intimate topics should be avoided, as respondents indicated: “For many topics I would need a person.”

In general, when queried about the appropriate conduct of the assistant, both respondents concurred that the assistant should refrain from exerting undue pressure or authority and should avoid engaging in repetitive or critical interactions. There is a distinction in the manner of presentation, with patients describing their preference for an assistant who is not “annoying,” “pointing out mistakes,” “using empty phrases,” and “scolding.” Health care professionals have indicated that the assistant should not be “pushy,” “lecturing,” “trivializing,” and “admonishing.”

Crisis Behavior

When queried as to how the assistant could facilitate a secure treatment environment during a crisis, health care professionals posited that the assistant could facilitate the establishment of contacts (personal and professional), “motivate [the] patient to turn to friends/family/therapists,” continue communication by “maintaining dialogue with the person.” Additional strategies included providing reassurance to the patient “Keep calm! Remind them of what has been agreed in advance,” assuring them of support, “reassuring them” by “announcing supportive next steps and being as transparent as possible” and providing positive support: “Be transparent, do not deny that it is only digital support, only do what has been discussed with the patient beforehand and what they have agreed to.”

A noteworthy response tendency emerged from the analysis of the practitioner questionnaire: “motivate the patient to seek help.” In response to this, we explicitly asked the patients how the assistant could most likely motivate them to seek help. The respondents indicated that the assistant should refrain from calling for help prematurely and should instead afford the opportunity to ask for help first. However, half of the respondents indicated that the assistant should directly request assistance, indicating a preference for contacting the therapists rather than friends or family members. In the context of a crisis situation, patients also expressed a desire for the assistant to maintain objectivity, to be “factual/informative about the situation (what benefit of seeking help)” and to “communicate mood clearly (visually).” Patients also wanted the assistant to be authoritative: “Just say: It’s time to get help” or “I’m convinced that help is appropriate now (as an announcement).” However, they also wanted to be motivated by the assistant’s reassuring behavior “by using a reassuring explanatory address (not ‘You have to do it now,’ but ‘Maybe you should...’).” In addition, the concept of a “panic button” was proposed, along with the display of assistance contacts, the presentation of support alternatives, and the provision of psychoeducation.

Overall, a wide range of behavioral options for the crisis situation was desired, especially by patients. In contrast to the patients, authoritarian behavior was not desired by the health care professionals. However, there was a tendency toward a preference for the system to be able to call for help without the patient having to give consent in a crisis, showing affinity toward decision-making authority. This may reflect the need for patients in crisis to be able to relinquish responsibility or to find guidance in the often-difficult decision-making process about whether and when a call for help is appropriate and necessary.

RQ 6: Interaction Between VTA and Health Care Professionals

This RQ aimed to investigate the relationship between the VTA and the health care professionals themselves. What modifications could be made to the framework to ensure that health care professionals at all levels could benefit from the implementation of a VTA? When queried on this matter, the health care professionals indicated that this should be done at the patient’s request, if the patient is able to provide consent. It is recommended that a preestablished “limit” be set regarding when the practitioner should be contacted. Moreover, the assistant should “obtain the patient’s permission beforehand.” In such instances, the assistant is permitted to “report suicidal thoughts, progressive deterioration of mood, or documented deterioration of mood over a period of 1-2 weeks,” “if a constructive discussion is not possible.” A further cluster of considerations pertains to instances of potential danger to self and others, “if the virtual character cannot assess the further danger with certainty,” “with threat or lack of dissociation of self—or harm to others,” or with “report of suicidal thoughts (even once).” In particular, in the absence of clear dissociation of suicidal thoughts, it was proposed that the assistant could

XSL•FO
RenderX

potentially dial the emergency number in the event of inactivity and a lack of response from the patient for a predefined period, in the case of an unannounced exit from the program, or in the event of inactivity for a period exceeding 48 hours. The final 2 clusters were identified as imminent when the "SOS button" was pressed or in the case of "external stressful events and crisis situations." At this juncture, the patients casted their votes on the question of whether the therapist should be informed without the patient's consent. The votes were one in favor and one against. A single individual posited that the assistant should be permitted to establish contact without delay, if the patient has consented. In total, 3 patients indicated a preference for multiple inquiries before action. This reinforces the necessity for individualized approaches and step-by-step plans that can be triggered according to the circumstances.

Discussion

Principal Findings

In the first part of the research findings, we examined how a VTA could be used as part of ongoing outpatient care. This includes the role (RQ1) of a VTA, particularly in the potential of filling the treatment gap, after discharge from the hospital. The respondents anticipated a system that would integrate expertise and facilitate the establishment of relationships between individuals. In conjunction with a primary recommendation from a trusted confidant of the therapist, the role of a peer counselor was also considered. Results on VTA embodiment (RQ2) indicated the need for such a counselor and provided practical implications, such as an imperfect or neutral-looking but emotionally expressive human embodiment, with a tendency to avoid male characters. In the second part of the findings, we evaluated patient and health care professionals' expectations regarding the interaction between a VTA and a patient. The results on interaction timing (RQ3) were considerable variables, reflecting the importance of an individual rhythm. The results obtained were delineated into 2 main clusters: *situation-independent* and *situation-dependent*. Regarding interaction abilities (RQ4) that the VTA should offer to support patients during treatment, 2 main clusters were also identified. *Structure-providing* included behavior for structuring and time management. *Emotional-supportive* included 3 subclusters *supportive-trusting,supportive-motivational*, and *supportive-resource-promoting.* The results regarding interactive behavior (RQ5) are presented separately for each of the subclusters identified in RQ3 and RQ4 and include explicit, competence-related, or emotion-stabilizing behavioral instructions. In addition, another cluster was identified as *taboo topics*, which includes content that should not be discussed.

The results pertaining to the interface between the assistant and the treating professionals (RQ6) are particularly relevant in determining the content that should be shared, the frequency, the timing, and patients consent.

Role

Expectations of the role of the assistant strongly reflect the treatment gap—both groups recognized the potential of a VTA as a supportive anchor following hospital discharge, not constrained by availability of health care professionals, but still equipped with professional techniques to navigate challenging circumstances (that relatives may not possess), while simultaneously serving as a projection screen for personal attachment. This leads us to conclude that a VTA has the potential to address the treatment gap, at least in part, and thus enhance outpatient aftercare.

Embodiment

This survey replicates previous findings [40] showing the need for an embodiment of the VTA and provides further insight into the practical implementation of this perspective. At first glance, the patients' expressed desire for an imperfect figure that could be interpreted as an indication of the "uncanny Valley effect" [61]. This phenomenon describes the abrupt decline in the acceptance of robots when they become too similar to humans. Nevertheless, the statements explicitly refer to aesthetic features, such as the (imperfect) figure and suggest that the assistant should not evoke negative social comparisons or negative self-evaluations, which can be a symptom in patients with depression. In this context, patients express fears of doubting their own self-esteem [62]. A common clinical presentation among individuals with depressive disorders is a perceived lack of self-esteem [63]. It would be beneficial to consider whether the VTA could be helpful in demonstrating how to make self-evaluation more positive, to show solution strategies. Notable examples include the chat robot Webot, which uses self-learning mechanisms through the presentation of short narratives about itself [64], or the home robot Jibo, which uses positive psychology techniques to assist individuals through the day [65].

A striking tendency in our data was the recommendation regarding the gender of the virtual assistant. Here, neither health care professionals nor patients explicitly recommended a male character; rather, a female character or a character with selectable gender was the preferred option. In other words, when explicitly asked to select a gender for a VTA, both health care professionals and patients avoided choosing a male character. Considering the highly sensitive target group of inpatient depressive patients on which our data were based, this recommendation suggests that the figure of the assistant should not convey any sense of threat. Other indications from the results regarding appearance included a "soft, calm voice" and "pleasant pastel shades, nothing too dark..." It can be concluded that men are perceived as more threatening according to the specific needs of the target group with depression. Statistical evidence indicates that instances of domestic violence against males and same-sex domestic violence are perceived as less threatening to society than instances of domestic male-on-female violence [66]. Research indicates that domestic violence has a profound impact on individuals and families and it is a substantial risk factor for mental illness [67]. In light of the observed tendency to perceive women as more empathetic, a phenomenon that has been documented in previous research on VTAs [68], it is plausible that the gender recommendation may have a dual impact here. It can be posited that both patients and health care professionals subconsciously had a preference for a female figure due to the aforementioned associations and the underlying motive to establish trust and security. Conversely, research on stereotypes in the form of dominance and gender in the VTA

XSL•FO
RenderX

indicates that male-associated behavior in female VTAs that does not align with the stereotype is met with rejection [41]. This indicates that it is crucial for gender-related conduct to align with the character's physical appearance. With respect to the specific target groups under consideration in this study, the impact of human-agent interaction in this regard should be examined, particularly in light of the notable gender-related outcomes observed here. Furthermore, these findings could be extended to other applications of SIAs where gender is a pivotal factor.

Some patients expressed a desire for the VTA to be modeled on their personal therapist. In light of the potential for technical advancement to enable the realistic modeling of avatars based on specific individuals [69], this option may be regarded as a realistic prospect. The therapeutic advantages (eg, availability of assistance) or disadvantages (eg, increased dependence on the therapist) would be highly specific to the particular therapeutic function or therapeutic relationship in question. As robots and intelligent systems begin to blur the boundaries between physical and virtual reality, there is an increased risk of negative emotions and thoughts being transferred to an assistant that has been modeled on someone from the patient's life. It remains to be seen whether intelligent systems will develop the capacity to counteract such effects in the future. This would require careful research [70]. However, the desire to recreate a known person as their VTA reflects patients' fear of being left alone and without trusted support after leaving the hospital. This was a common theme throughout this survey.

Regarding external expressiveness, patients and health care professionals indicated that the assistant should be capable of communicating through facial expressions and gestures. It is recommended that the face, upper body, arms, and hands should be visible. In addition to voice and content, body language is also of importance. This indicates that the affective functions of the assistant should be multimodal to be able to act emphatically. In crisis situations, patients may engage with such a system with their undivided attention. In contrast to other contextual studies, such as those examining a driving assistant where users have only a divided attention available for interaction with the assistant, it is desired for VTAs to be less distracting (only one visible head; [41). In addition, it is demonstrated that participants want to allocate more attention to the assistant and want it to be expressive and empathetic. The research on nonverbal synchrony in therapeutic sessions indicates that high levels of nonverbal synchrony are associated with a higher quality of the therapeutic relationship and greater symptom reduction. Consequently, the design of an assistant capable of nonverbal synchrony with the patient serves as an additional indicator of therapeutic efficacy [71]. Inferentially, through an assumption of an outward appearance with corresponding expressiveness, participants in both survey groups expressed a desire for this function. However, it is of particular importance, particularly in the context of mental health, to design expressive VTA behavior in accordance with the specific requirements of the health care target group [72].

In conclusion, the findings indicate that embodiment is a substantial factor, but its implementation must be approached with caution to ensure a beneficial impact.

Timing

In response to the question of timing, the 2 main clusters were identified as *situation-independent* and *situation-dependent*. The results of the *situation-independent* cluster may also be summarized as proactive behavior on the part of the assistant. Proactive behavior on the part of the patient was not mentioned by any of the respondents. In general, this demonstrates that the assistant is regarded as an activating medium and the patients are perceived as passive recipients of data by the therapists and by themselves, except in situations where requests are situation related. Here, the particular requirements of the specific depressive group are evident. This indicates that the initial 6-month period following discharge is a critical period for the risk of relapse and suicide [4]. The reintegration into to the home environment frequently results in incomplete remission, and the persistence of symptoms further complicates the ability to independently coordinate the various stakeholders in sociopsychiatric aftercare. The accompanying psychosocial restrictions, such as a lack of daily structure or cognitive limitations, make the necessity for a situation-independent, activating figure apparent. The almost unlimited availability of activating resources of such a system appears to be of great functional advantage at this point. Therefore, the unconstrained accessibility of activating resources within an attachment-oriented system gives rise to concerns that are not seen in the natural boundaries of human-human interactions. In light of the aforementioned limitations, a situation-independent activating assistant appears to be a logical solution for patients with depression during aftercare. However, there is a risk that unrestricted availability of the attachment-oriented assistant may result in dependency among patients. For example, a previous study [73] on attachment-oriented robots demonstrated that a system that provides emotional attachment in addition to factual competencies and addresses loneliness may potentially foster dependency in patients. It is yet to be determined whether a system designed to counteract this can be developed in a way that incorporates factual competencies at all levels into a training mode that is inherently geared toward resource building, thereby creating space for development and independence.

The frequency-related results, that is, the question of when the assistant and patient should interact, demonstrate the necessity for a more individualized system. The need for individualization can be regulated in 5 different ways: at the request of the patient, according to content, according to time, when the patient's condition changes and before outpatient therapy or appointments.

Regarding frequency of interactions, these may occur daily at designated times, on a weekly or monthly basis, or at other intervals. Interactions surrounding particular events may also be structured temporally in accordance with scheduled appointments. Human-interaction therapy sessions in a social psychiatric context are typically conducted at varying frequencies. At the outset of a treatment agreement, a potential frequency is typically established, contingent on the patient's needs and the therapist's availability. Research has shown that participants, especially those with lower levels of education, preferred a VTA to a human after discharge from hospital because they could control the pace of information [74].

XSL•FO
RenderX

Furthermore, the frequency of therapy appointments tends to increase gradually, contingent upon the efficacy of the therapy and the availability of approved cost coverage. In this context, the frequency of therapy has an impact on the success of therapy. Consequently, researchers advise therapists to treat patients as often as possible, particularly during the initial phase of therapy [75]. At this juncture, a virtual system can bridge gaps in care due to its unlimited availability and supplementary therapies that are more limited in scope. Conversely, it is of the utmost importance to prevent the patient from becoming dependent on the therapist or on a digital system. It is therefore imperative that a digital system, which is capable of maintaining unlimited availability, simultaneously instills a sense of autonomy and independence to prevent dependency. Should the patient wish to interact, this would necessitate the individual activation of the system by the patient. Content-dependent interactions refer to the dialogue or content that the patient shares with the system, either verbally or by answering questions. The individualization of the content of the dialogue represents a key aspect of this process. Interactions that occur when the patient's emotional state changes can be proactively initiated by the patient. It is similarly conceivable that this could form the basis of a digital phenotyping monitoring system. The concept of being "seen" by a system that responds to crises, provides assistance, or alerts to early warning symptoms was generally viewed positively. New interventions that take advantage of the quantification of clinical markers through condition monitoring can contribute to clinical improvement [76,77]. Our results indicate user acceptance of such monitoring at this point, as well as needs and preferences.

Interaction Abilities

Regarding the interaction skills that the assistant should be able to offer to provide support to patients undergoing treatment, 2 broad clusters were identified. On the one hand, the skills that could be classified as structured assistance were designated as the *structure-providing* cluster. In contrast, the skills that could be assigned to emotional support were classified as the *emotional-supportive* cluster. Both of these main clusters could be defined in a cross-dimensional manner, thereby allowing for the identification of a higher-order meaning. In this context, we focused on 2 sociocognitive dimensions that are particularly relevant for the description of individuals: warmth and competence [78,79]. These 2 variables have been identified by Fiske et al [78] as the 2 most important factors by which people evaluate encounters. The character traits that individuals form about others when they spontaneously gather impressions can be summarized in these 2 basic dimensions [80]. The dimension of warmth is defined as a set of traits that are perceived to indicate intentions of friendliness, helpfulness, sincerity, trustworthiness, and morality. In contrast, the competence dimension traits are associated with perceptions of abilities, including intelligence, skill, creativity, and efficacy [78]. Our data point to these sociocognitive dimensions within the 2 main clusters.

In addition, the data delineates the interaction skills, which include the virtual assistant's emotions, personality, and social skills. In conjunction with the character's appearance, these skills influence the user's assessment of the character's believability. The term "believability" in the context of VTAs is defined as the capacity to be perceived as a genuine entity [81]. For this to occur, the character must act in a manner that is consistent with its stated goals, personality, and state of mind [82]. In this context, consistency is defined as the coherence between the characters' verbal and nonverbal behaviors as well as its physical appearance. In particular, our data substantiate the hypothesis that users of virtual assistants within the subcluster *supportive-trusting* desire consistency in the assistant's behavior. This desire was described as "authentic" as well as socially interactive and competent in the sense that the assistant's emotional expressions and statements should be appropriate to the situation. It is therefore evident that the VTA must be capable of modifying its social conduct in accordance with the circumstances to appear credible and pertinent. Our findings corroborate those of previous research, which have highlighted the significance of believability in virtual assistants [81]. In addition, our data indicate that believability can be classified as a subcluster of the warmth dimension, and can be seen as a subconstruct of the 2 sociocognitive dimensions of warmth and competence.

Interaction Behavior

In contrast to human therapeutic interactions, which always entail an exchange about a momentary, subjective state and require the patient to disclose personal content, a digital monitoring system maintains a pervasive and comprehensive impression that may not necessitate an active decision by the patient to communicate. At this juncture, users of a digital monitoring system are disclosing a considerable amount of personal information by sharing data that is not actively shared, thereby relinquishing both responsibility and control over the sharing process. The data indicates that, regarding the variety of tools and forms of intervention, there is no substantial bias toward data sharing among the 2 groups of respondents. The primary objective is to identify a figure who can effectively combine competence with warmth, thereby fostering the most robust possible trust between the patient and the figure. As evidenced by psychotherapy research, trust and shame play a major impact in the efficacy of therapeutic interventions [83,84]. Assisting patients in modifying their behaviors also necessitates transparency regarding their shortcomings. Concurrently, individuals divulge a considerably more intimate array of information to computers than to humans [85,86]. Despite the well-established fact that data is identifiable, indelible, and, in certain circumstances, obtainable [87], computers do not elicit the same social inhibitions as humans. In the context of human–computer interaction, particularly in the context of digital assistants, it is pertinent to inquire into the underlying mechanisms that facilitate this phenomenon. Against this background, it seems plausible that crucial psychometric inquiries in psychiatric diagnostics may be more effectively elucidated through human-machine interaction. It is beyond dispute that the protection of patient data must be accorded with the highest priority. In its social function, shame serves as a reflection of the affected person's behavior and has developmental potential. Conversely, shame entails the participation of at least 2 individuals, thereby indicating that the observed transgression was inadvertent. In this way, shame

XSL•FO
RenderX

functions as a feedback mechanism, protecting against extremes or loss of control and leading to inclusive behavior in interaction partners. In social interactions, the successful overcoming of shame can facilitate the formation of attachment on both sides. However, existential shame, which is defined as the perception of oneself as generally flawed, can result in social withdrawal and, ultimately, the onset of depressive symptomatology [88,89]. It is, therefore, possible to hope that deficiencies in care in human-machine interaction can also be addressed at the level of communication. A solution to social psychiatric diagnostics that is free from shame can help to prevent crises, particularly in situations of critical transition, such as after hospitalization. At the same time, it should be noted that a shame-free interaction with a computer will not replace the challenges that people experience in social encounters.

Assistant-Health Care Professional Interaction

This final inquiry relates to the interface between the assistant and the health care professionals, that is, modifications to the framework to ensure that health care professionals benefit from the implementation of a VTA. Moreover, this could be regarded as a control function to guarantee that the assistant does not unintentionally impede the ongoing treatment.

The results from all cluster dimensions demonstrate substantial overlap across the full spectrum of social psychiatric care, which raises the question of how such a support system could potentially transform the landscape of social psychiatric care. For example, it could result in a shift in social values within the domain of social psychiatry or contribute to an increased reliance on digital technologies for the provision of care. Accordingly, the societal impact of an intelligent, attachment-oriented assistance system on both human-human and human-robot relationships must be explored and designed to be participatory at all levels of use [90]. Evaluations need to take place at all levels of development in this process, so that all targeted professional groups, as well as patient advocacy organizations, can provide input [76]. In particular, possible short- and long-term effects on the individual patient's identity, freedom of choice, and self-perception need to be explored.

Two potential outcomes emerge at the interaction interface between therapist and assistant, as evidenced by the overall results. On the one hand, there is the possibility of receiving information about the patient's health status via a ubiquitous computer, which the patient would be disinclined to share with the perhaps new, still unknown therapist for reasons already discussed. The primary reasons for this were insufficient time or a lack of confidence on the part of the patient. Conversely, the alarm system serves to alert the therapist or the rescue system in the event of indications of suicidal ideation. These 2 points present the potential for addressing existing care and enhancing the safety of patient follow-up for those with depression [91]. Concurrently, the frequently arduous period of follow-up care will be less onerous for treatment providers. The enhanced objectivity of the assessment of patients with depression who are to be monitored provides a basis for decision-making in the context of treatment.

It would be beneficial to ascertain which groups of patients are particularly at risk of relapse following hospitalization. Those from socioeconomically disadvantaged backgrounds are frequently inadequately served. Psychiatric care is underrepresented in rural, economically underdeveloped regions [1]. Furthermore, the risk of relapse is elevated in individuals who are young at the onset of illness, female, older, have somatic comorbidity, are single, and have a general lack of psychosocial support determine a high susceptibility to relapse [2]. During the course of the pandemic, the rise in the incidence of mental health disorders occurred concurrently with substantial disruptions to the provision of mental health services [92]. A deficiency in the provision of essential health services was observed, including those pertaining to mental health and suicide prevention, as reported by the World Health Organization. While the situation showed some improvement by the end of the pandemic, it still demonstrates the vulnerability of a health care system to unforeseen crises. Health care professionals caring for these susceptible people could benefit professionally from a VTA system that helps to make outpatient care safer. The effect sizes, considering the different groups of people affected, remain an open RQ.

Limitations and Directions for Future Research

This study has several limitations. First, the recruitment and questioning of the 2 groups of participants was carried out in different ways, with the questioning being adapted both to the specific groups and to the specific situations. This could have led to distortion. Second, the results showed that the answer option "diverse" to the gender question was misunderstood. Future studies should pay attention to clearer formulation. In addition, psychometric data were used solely for descriptive purposes, despite the potential for exploring the impact of individual factors on outcomes, a direction that merits attention in future research. The study's scope is further limited by its inclusion of health care professionals and patients from Oldenburg, Germany, precluding the generalizability of its results. To enhance the relevance and external validity of the findings, subsequent studies should aim to encompass more representative samples of the general population. Furthermore, this study exclusively examined participants' attitudes toward an SIA, neglecting to investigate its potential effects on patients. Consequently, further research is necessary to investigate the impact of such a system and to ensure the benefits for those affected at all levels of the system.

Conclusions

Main conclusions can be derived from the study related to the role of VTAs, their embodiment, the timing of interaction, interaction skills and behavior, and interaction between the assistant and health professionals.

Role

The VTA was identified as a valuable supplementary component within the outpatient aftercare system. The VTA is able to address deficiencies in care due to its extensive repertoire of therapeutic skills and its capacity to establish a rapport with patients where social workers, therapists, and family members are unable to do so.

XSL•FO
RenderX

Embodiment

A multimodal function for empathic communication (ie, voice, facial expressions, and gestures) is expected. Patients may be more likely to engage with an assistant who appears imperfect, as this may help to prevent negative self-evaluation, particularly in patients with depression who lack self-esteem. The assistant may be able to assist in the development of solution strategies to enhance a more positive self-image. The gender of the assistant should be either female or selectable, but not male. This preference is thought to be based on negative associations with the male gender, such as domestic violence, and the perception of the female gender as more empathetic.

Timing of Interaction

The findings indicated that both patients and health care professionals anticipate the assistant to engage in proactive behavior. Therefore, the assistant is regarded as an activator of information, whereas patients are viewed as passive recipients. The expectations regarding the timing of specific behaviors can be differentiated into 2 categories: situation-independent and situation-dependent behaviors.

Interaction Skills and Behavior

Two core skills were identified as being essential for both health care professionals and patients: the capacity to provide structural support and the capacity to provide emotional support. These abilities correspond to the dimensions of interpersonal social perception, namely competence and warmth. An agent who exhibits these 2 abilities through appropriate behavior at the appropriate time could serve as a valuable addition to outpatient aftercare. For each subcluster and appropriate timing, the results provide an extensive list of suggestions for appropriate behavior.

Interaction Between Assistant and Health Professional

The results demonstrate the potential for therapists to obtain otherwise unavailable information regarding patients' health status, as well as the feasibility of an emergency system in the event of suicidal risk. These innovations have the potential to address existing gaps in aftercare, enhance safety during this critical period, reduce the burden on health care professionals, and, by providing an objective measure of the patient's health status, facilitate health care professional's decision-making.

Summary of Conclusions

In summary, our findings offer a promising framework for the development of a VTA to support patients with depression during aftercare and support the work of health care professionals.

Acknowledgments

This research was funded by the UBIDENZ project (13GW0568A) and partly sponsored by the Hybrid Intelligence project (024.004.022). The artificial intelligence–supported DeepL translation tool was used for translation of certain expressions used in the final version of this paper. The authors would like to thank Dr Klaus Bajorat, Maira Keller, Anna Lea Reinwarth, Daniel Faber, Madina Khalillaeva, David Mietzner, and Rachana Mahadevan for their help with creating the experimental setup, data collection, and analysis.

The authors would like to thank their collaborators for their contributions: Janet Wessler, PhD, Tanja Schneeberger, PhD, Jan Alexanderson, PhD, Nicklas Linz, PhD, Norbert Helff, Dipl, Arno Elmer, PhD, and Elisabeth André, PhD.

Data Availability

Following the recommendations for open research practices [50], questionnaires and experimenter instruction can be found on the OSF website [51]. The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

References

1. Depression and other common mental disorders: global health estimates. World Health Organization. URL: https://policycommons.net/artifacts/546082/depression-and-other-common-mental-disorders/1523689/ [accessed 2025-09-03]
2. Unipolare depression. Nationale Versorgungs Leitlinie. 2015. URL: https://register.awmf.org/assets/guidelines/nvl-005l_S3_Unipolare-Depression_2023-07.pdf [accessed 2024-04-29]
3. Wiegand HF, Saam J, Marschall U, Chmitorz A, Kriston L, Berger M, et al. Challenges in the transition from in-patient to out-patient treatment in depression. Dtsch Arztebl Int 2020 Jul 06;117(27-28):472-479 [FREE Full text] [doi: 10.3238/arztebl.2020.0472] [Medline: 33050996]
4. Baldessarini RJ, Tondo L, Pinna M, Nuñez N, Vázquez GH. Suicidal risk factors in major affective disorders. Br J Psychiatry 2019 Jul 11;215(4):1-6. [doi: 10.1192/bjp.2019.167] [Medline: 31292010]

XSL•FO
RenderX

5. Chung DT, Ryan CJ, Hadzi-Pavlovic D, Singh SP, Stanton C, Large MM. Suicide rates after discharge from psychiatric facilities: a systematic review and meta-analysis. JAMA Psychiatry 2017 Jul 01;74(7):694-702 [FREE Full text] [doi: 10.1001/jamapsychiatry.2017.1044] [Medline: 28564699]
6. Castelli Dransart DA, Heeb J, Gulfi A, Gutjahr EM. Stress reactions after a patient suicide and their relations to the profile of mental health professionals. BMC Psychiatry 2015 Oct 28;15(1):265 [FREE Full text] [doi: 10.1186/s12888-015-0655-y] [Medline: 26511910]
7. Kleespies PM, Penk WE, Forsyth JP. The stress of patient suicidal behavior during clinical training: incidence, impact, and recovery. Prof Psychol Res Pr 1993;24(3):293-303. [doi: 10.1037//0735-7028.24.3.293]
8. Cushway D, Tyler P. Stress in clinical psychologists. Int J Soc Psychiatry 1996 Jun 01;42(2):141-149. [doi: 10.1177/002076409604200208] [Medline: 8811398]
9. Deutsch CJ. Self-reported sources of stress among psychotherapists. Prof Psychol Res Pr 1984;15(6):833-845. [doi: 10.1037//0735-7028.15.6.833]
10. Sfetcu R, Musat S, Haaramo P, Ciutan M, Scintee G, Vladescu C, et al. Overview of post-discharge predictors for psychiatric re-hospitalisations: a systematic review of the literature. BMC Psychiatry 2017 Jun 24;17(1):227 [FREE Full text] [doi: 10.1186/s12888-017-1386-z] [Medline: 28646857]
11. Ho TP. The suicide risk of discharged psychiatric patients. J Clin Psychiatry 2003 Jun 15;64(6):702-707. [doi: 10.4088/jcp.v64n0613] [Medline: 12823086]
12. Mahmoud AS, Elsayed Berma A, Saleh Gabal SA. Relationship between social support and the quality of life among psychiatric patients. J Psychiatry Psychiatric Disord 2017;01(02):57-75. [doi: 10.26502/jppd.2572-519x008]
13. Büscher R, Torok M, Terhorst Y, Sander L. Internet-based cognitive behavioral therapy to reduce suicidal ideation: a systematic review and meta-analysis. JAMA Netw Open 2020 Apr 01;3(4):e203933 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.3933] [Medline: 32343354]
14. Eichenberg C. E-Mental Health Anwendungen für depressive und suizidale Menschen: eine Übersicht. Psychother Forum 2019 Nov 14;23(3-4):111-119. [doi: 10.1007/S00729-019-00123-3]
15. Klein JP, Knaevelsrud C, Bohus M, Ebert DD, Gerlinger G, Günther K, et al. [Internet-based self-management interventions : quality criteria for their use in prevention and treatment of mental disorders]. Nervenarzt 2018 Nov 20;89(11):1277-1286. [doi: 10.1007/s00115-018-0591-4] [Medline: 30128736]
16. Miller I, Peake E, Strauss G, Vierra E, Koepsell X, Shalchi B, et al. Self-guided digital intervention for depression in adolescents: feasibility and preliminary efficacy study. JMIR Form Res 2023 Nov 22;7:e43260 [FREE Full text] [doi: 10.2196/43260] [Medline: 37991839]
17. Bowlby J. Attachment theory and its therapeutic implications. Adolesc Psychiatry 1978;6:5-33. [Medline: 742687]
18. Grawe K, Donati R, Bernauer F. Psychotherapie im Wandel: Von der Konfession zur Profession. Göttingen, Germany: Hogrefe Publishing; 1994.
19. Urech A, Krieger T, Möseneder L, Biaggi A, Vincent A, Poppe C, et al. A patient post hoc perspective on advantages and disadvantages of blended cognitive behaviour therapy for depression: a qualitative content analysis. Psychother Res 2019 Nov 31;29(8):986-998. [doi: 10.1080/10503307.2018.1430910] [Medline: 29385964]
20. Bickmore T, Gruber A, Picard R. Establishing the computer-patient working alliance in automated health behavior change interventions. Patient Educ Couns 2005 Oct;59(1):21-30. [doi: 10.1016/j.pec.2004.09.008] [Medline: 16198215]
21. Gebhard P, Schneeberger T, Andre E, Baur T, Damian I, Mehlmann G, et al. Serious games for training social skills in job interviews. IEEE Trans Games 2019 Dec;11(4):340-351. [doi: 10.1109/tg.2018.2808525]
22. Gratch J, Wang N, Gerten J, Fast E, Duffy R. Creating rapport with virtual agents. In: Proceedings of the 7th International Working Conference on Intelligent Virtual Agents. 2007 Presented at: IVA '07; September 17-19, 2007; Paris, France p. 125-138 URL: https://link.springer.com/chapter/10.1007/978-3-540-74997-4_12 [doi: 10.1007/978-3-540-74997-4_12]
23. Gratch J, Kang SH, Wang N. Using social agents to explore theories of rapport and emotional resonance. In: Gratch J, Marsella S, editors. Social Emotions in Nature and Artifact. Oxford, UK: Oxford Academic Press; 2013:181-197.
24. Darcy A, Daniels J, Salinger D, Wicks P, Robinson A. Evidence of human-level bonds established with a digital conversational agent: cross-sectional, retrospective observational study. JMIR Form Res 2021 May 11;5(5):e27868 [FREE Full text] [doi: 10.2196/27868] [Medline: 33973854]
25. Dias LP, Barbosa JL, Feijó LP, Vianna HD. Development and testing of iAware model for ubiquitous care of patients with symptoms of stress, anxiety and depression. Comput Methods Programs Biomed 2020 Apr;187:105113. [doi: 10.1016/j.cmpb.2019.105113] [Medline: 31607411]
26. Schneeberger T, Reinwarth AL, Wensky R, Anglet MS, Gebhard P, Wessler J. Fast friends: generating interpersonal closeness between humans and socially interactive agents. In: Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents. 2023 Presented at: IVA '23; September 19-22, 2023; Würzburg, Germany p. 1-8 URL: https://dl.acm.org/doi/10.1145/3570945.3607302 [doi: 10.1145/3570945.3607302]
27. Gebhard P. ALMA: a layered model of affect. In: Proceedings of the 4th International Joint Conference on Autonomous Agents and Multiagent Systems. 2005 Presented at: AAMAS '05; July 25-29, 2005; Amsterdam, The Netherlands p. 29-36 URL: https://dl.acm.org/doi/10.1145/1082473.1082478 [doi: 10.1145/1082473.1082478]

XSL•FO
RenderX

28. Gebhard P, Schneeberger T, Baur T, André E. MARSSI: model of appraisal, regulation, and social signal interpretation. In: Proceedings of the 17th International Conference on Autonomous Agents and Multi Agent Systems. 2018 Presented at: AAMAS '18; July 10-15, 2018; Stockholm-Uppsala, Sweden p. 497-506 URL: https://dl.acm.org/doi/10.5555/3237383.3237458 [doi: 10.5555/3237383.3237458]
29. Zhang JE, Hilpert B, Broekens J, Jokinen JP. Simulating emotions with an integrated computational model of appraisal and reinforcement learning. In: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. 2020 Presented at: CHI '24; May 11-16, 2024; Honolulu, HI p. 1-12 URL: https://dl.acm.org/doi/full/10.1145/3613904.3641908
30. Davis M, Hadiks D. Nonverbal aspects of therapist attunement. J Clin Psychol 1994 May;50(3):393-405. [doi: 10.1002/1097-4679(199405)50:3<393::aid-jclp2270500311>3.0.co;2-t] [Medline: 8071445]
31. Lugrin B. Introduction to socially interactive agents. In: Lugrin B, Pelachaud C, Traum D, editors. The Handbook on Socially Interactive Agents: 20 Years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics. New York, NY: Association for Computing Machinery; 2021:1-20.
32. Schneeberger T, Hirsch A, König C, Gebhard P. Impact of virtual environment design on the assessment of virtual agents. In: Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents. 2019 Presented at: IVA '19; July 2-5, 2019; Paris, France p. 148-150 URL: https://dl.acm.org/doi/10.1145/3308532.3329455 [doi: 10.1145/3308532.3329455]
33. Nicora ML, André E, Berkmans D, Carissoli C, D'Orazio T, Fave AD. A human-driven control architecture for promoting good mental health in collaborative robot scenarios. In: Proceedings of the 30th IEEE International Conference on Robot & Human Interactive Communication. 2021 Presented at: RO-MAN '21; August 8-12, 2021; Vancouver, BC p. 285-291 URL: https://ieeexplore.ieee.org/document/9515315 [doi: 10.1109/ro-man50785.2021.9515315]
34. Bhuvaneshwara C, Anglet M, Hilpert B, Chehayeb L, Meyer AK, Don DW, et al. MITHOS-mixed reality interactive teacher training system for conflict situations at school. In: Proceedings for the 2023 Conference on International Society of the Learning Sciences. 2023 Presented at: ISLS '23; June 10-15, 2023; Montreal, QC p. 43 URL: https://tinyurl.com/vv9e8a5y
35. Schneeberger T, Sauerwein N, Anglet MS, Gebhard P. Stress management training using biofeedback guided by social agents. In: Proceedings of the 26th International Conference on Intelligent User Interfaces. 2021 Presented at: IUI '21; April 14-17, 2021; College Station, TX p. 564-574 URL: https://dl.acm.org/doi/10.1145/3397481.3450683 [doi: 10.1145/3397481.3450683]
36. Ali MR, Van Orden K, Parkhurst K, Liu S, Nguyen VD, Duberstein P, et al. Aging and engaging: a social conversational skills training program for older adults. In: Proceedings of the 23rd International Conference on Intelligent User Interfaces. 2018 Presented at: IUI '18; March 7-11, 2018; Tokyo, Japan p. 55-66 URL: https://dl.acm.org/doi/10.1145/3172944.3172958 [doi: 10.1145/3172944.3172958]
37. Aylett RS, Louchart S, Dias J, Paiva A, Vala M. FearNot! – an experiment in emergent narrative. In: Proceedings of the 5th International Working Conference on Intelligent Virtual Agents. 2005 Presented at: IVA '05; September 12-14, 2005; Kos, Greece p. 305-316 URL: https://link.springer.com/chapter/10.1007/11550617_26 [doi: 10.1007/11550617_26]
38. Morris RR, Kouddous K, Kshirsagar R, Schueller SM. Towards an artificially empathic conversational agent for mental health applications: system design and user perceptions. J Med Internet Res 2018 Jun 26;20(6):e10148 [FREE Full text] [doi: 10.2196/10148] [Medline: 29945856]
39. Bickmore TW, Mitchell SE, Jack BW, Paasche-Orlow MK, Pfeifer LM, Odonnell J. Response to a relational agent by hospital patients with depressive symptoms. Interact Comput 2010 Jul 01;22(4):289-298 [FREE Full text] [doi: 10.1016/j.intcom.2009.12.001] [Medline: 20628581]
40. Kim K, Boelling L, Haesler S, Bailenson J, Bruder G, Welch GF. Does a digital assistant need a body? The influence of visual embodiment and social behavior on the perception of intelligent virtual agents in AR. In: Proceedings of the 2018 IEEE International Symposium on Mixed and Augmented Reality. 2018 Presented at: ISMAR '18; October 16-20, 2018; Munich, Germany p. 105-114 URL: https://ieeexplore.ieee.org/document/8613756 [doi: 10.1109/ismar.2018.00039]
41. Wessler J, Schneeberger T, Christidis L, Gebhard P. Virtual backlash: nonverbal expression of dominance leads to less liking of dominant female versus male agents. In: Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents. 2022 Presented at: IVA '22; September 6-9, 2022; Faro, Portugal p. 1-8 URL: https://dl.acm.org/doi/10.1145/3514197.3549682 [doi: 10.1145/3514197]
42. Hilpert B, da Silva CA, Christidis L, Bhuvaneshwara C, Gebhard P, Nunnari F, et al. Avatar visual similarity for social HCI: increasing self-awareness. arXiv Preprint posted online August 23, 2024 [FREE Full text] [doi: 10.2139/ssrn.4797708]
43. Weitz K, Schiller D, Schlagowski R, Huber T, André E. "Do you trust me?": increasing user-trust by integrating virtual agents in explainable AI interaction design. In: Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents. 2019 Presented at: IVA '19; July 2-5, 2019; Paris, France p. 7-9 URL: https://dl.acm.org/doi/10.1145/3308532.3329441 [doi: 10.1145/3308532.3329441]
44. Hilpert B, Schneeberger T, Gebhard P. Employing virtual agents for building trust in driving automation: a qualitative pilot study. In: Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction. 2023 Presented at: FAccT '23; June 12-15, 2023; Honolulu, HI URL: https://www.researchgate.net/publication/349924456_Employing_Virtual_Agents_for_Building_Trust_in_Driving_Automation_A_Qualitative_Pilot_Study

XSL•FO
RenderX

45. Loveys K, Sebaratnam G, Sagar M, Broadbent E. The effect of design features on relationship quality with embodied conversational agents: a systematic review. Int J of Soc Robotics 2020 Sep 24;12(6):1293-1312. [doi: 10.1007/s12369-020-00680-7]
46. Gebhard P, Schneeberger T, Mehlmann G, Baur T, André E. Designing the impression of social agents' real-time interruption handling. In: Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents. 2019 Presented at: IVA '19; July 2-5, 2019; Paris, France p. 19-21 URL: https://dl.acm.org/doi/10.1145/3308532.3329435 [doi: 10.1145/3308532.3329435]
47. Zhao R, Sinha T, Black AW, Cassell J. Socially-aware virtual agents: automatically assessing dyadic rapport from temporal patterns of behavior. In: Proceedings of the 16th International Conference on Intelligent Virtual Agents. 2016 Presented at: IVA '16; September 20-23, 2016; Los Angeles, CA p. 218-233 URL: https://link.springer.com/chapter/10.1007/978-3-319-47665-0_20 [doi: 10.1007/978-3-319-47665-0_20]
48. Schneeberger T, Scholtes M, Hilpert B, Langer M, Gebhard P. Can social agents elicit shame as humans do? In: Proceedings of the 8th International Conference on Affective Computing and Intelligent Interaction. 2019 Presented at: ACII '19; September 3-6, 2019; Cambridge, UK p. 164-170 URL: https://ieeexplore.ieee.org/document/8925481 [doi: 10.1109/acii.2019.8925481]
49. Rheu M, Shin JY, Peng W, Huh-Yoo J. Systematic review: trust-building factors and implications for conversational agent design. Int J Hum Comput Interact 2020 Sep 02;37(1):81-96. [doi: 10.1080/10447318.2020.1807710]
50. DeVault D, Artstein R, Benn G, Dey T, Fast E, Gainer A. SimSensei kiosk: a virtual human interviewer for healthcare decision support. In: Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems. 2014 Presented at: AAMAS '14; May 5-9, 2014; Paris, France p. 1061-1068 URL: https://dl.acm.org/doi/10.5555/2615731.2617415
51. Deladisma AM, Cohen M, Stevens A, Wagner P, Lok B, Bernard T, Association for Surgical Education. Do medical students respond empathetically to a virtual patient? Am J Surg 2007 Jun;193(6):756-760. [doi: 10.1016/j.amjsurg.2007.01.021] [Medline: 17512291]
52. Kramer P. Frames of Reference for Pediatric Occupational Therapy. Thousand Oaks, CA: Lippincott Williams & Wilkins; 2018.
53. Paiva A, Leite I, Boukricha H, Wachsmuth I. Empathy in virtual agents and robots: a survey. ACM Trans Interact Intell Syst 2017 Sep 19;7(3):1-40. [doi: 10.1145/2912150]
54. Chehayeb L, Bhuvaneshwara C, Anglet M, Hilpert B, Meyer AK. MITHOS: interactive mixed reality training to support professional socio-emotional interactions at schools. arXiv Preprint posted online September 2, 2024 [FREE Full text] [doi: 10.48550/arXiv.2409.12968]
55. Beck Depression Inventory II (BDI-II). APA PsycTests. URL: https://psycnet.apa.org/doiLanding?doi=10.1037%2Ft00742-000 [accessed 2024-04-29]
56. Yamagishi T, Yamagishi M. Trust and commitment in the United States and Japan. Motiv Emot 1994 Jun;18(2):129-166. [doi: 10.1007/bf02249397]
57. Wessler J, Schneeberger T, Hilpert B, Alles A, Gebhard P. Empirical research in affective computing: an analysis of research practices and recommendations. In: Proceedings of the 9th International Conference on Affective Computing and Intelligent Interaction. 2021 Presented at: ACII '21; September 28-October 1, 2021; Nara, Japan p. 1-8 URL: https://ieeexplore.ieee.org/document/9597418 [doi: 10.1109/acii52823.2021.9597418]
58. Immel D. Materials to a survey about virtual therapeutic assistants in outpatient aftercare - expectations of patients and health care professionals. Open Science Framework. URL: https://osf.io/umwkf/ [accessed 2025-04-03]
59. Glaser BG, Strauss AL. Grounded Theory: Strategien Qualitativer Forschung. Berlin, Germany: Verlag Hans Huber; 2010.
60. Hansen C, Grimshaw-Aagaard S. Developing a guideline for reporting and evaluating grounded theory research studies. Int J Health Sci 2018:64-76. [doi: 10.15640/ijhs.v6n1a8]
61. Mori M. Bukimi no tani [the uncanny valley]. CiNii. URL: https://cir.nii.ac.jp/crid/1370013168736887425 [accessed 2025-09-03]
62. Battle J. Relationship between self-esteem and depression. Psychol Rep 1978 Jun 01;42(3 PT 1):745-746. [doi: 10.2466/pr0.1978.42.3.745] [Medline: 674498]
63. Sowislo JF, Orth U. Does low self-esteem predict depression and anxiety? A meta-analysis of longitudinal studies. Psychol Bull 2013 Jan;139(1):213-240. [doi: 10.1037/a0028931] [Medline: 22730921]
64. Fitzpatrick KK, Darcy A, Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. JMIR Ment Health 2017 Jun 06;4(2):e19 [FREE Full text] [doi: 10.2196/mental.7785] [Medline: 28588005]
65. Jeong S, Alghowinem S, Aymerich-Franch L, Arias K, Lapedriza A, Picard R. A robotic positive psychology coach to improve college students' wellbeing. In: Proceedings of the 29th IEEE International Conference on Robot and Human Interactive Communication. 2020 Presented at: RO-MAN '20; August 31-September 4, 2020; Naples, Italy p. 187-194 URL: https://ieeexplore.ieee.org/document/9223588 [doi: 10.1109/ro-man47096.2020.9223588]
66. Maurya BR. An overview on domestic violence. Asian J Multidiscip Res 2021;10:12. [doi: 10.5958/2278-4853.2021.01139.3]

XSL•FO
RenderX

67. Devakumar D, Palfreyman A, Uthayakumar-Cumarasamy A, Ullah N, Ranasinghe C, Minckas N, et al. Mental health of women and children experiencing family violence in conflict settings: a mixed methods systematic review. Confl Health 2021 Oct 15;15(1):74 [FREE Full text] [doi: 10.1186/s13031-021-00410-4] [Medline: 34654456]
68. Tanaka H, Nakamura S. The acceptability of virtual characters as social skills trainers: usability study. JMIR Hum Factors 2022 Mar 29;9(1):e35358 [FREE Full text] [doi: 10.2196/35358] [Medline: 35348468]
69. da Silva CA, Hilpert B, Bhuvaneshwara C, Gebhard P, Nunnari F, Tsovaltzi D. Visual similarity for socially interactive agents that support self-awareness. In: Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents. 2023 Presented at: IVA '23; September 19-22, 2023; Würzburg, Germany p. 1-3 URL: https://dl.acm.org/doi/10.1145/3570945.3607329 [doi: 10.1145/3570945.3607329]
70. Fiske A, Henningsen P, Buyx A. Your robot therapist will see you now: ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. J Med Internet Res 2019 May 09;21(5):e13216 [FREE Full text] [doi: 10.2196/13216] [Medline: 31094356]
71. Ramseyer F, Tschacher W. Nonverbal synchrony in psychotherapy: coordinated body movement reflects relationship quality and outcome. J Consult Clin Psychol 2011 Jun;79(3):284-295. [doi: 10.1037/a0023419] [Medline: 21639608]
72. Reinwarth AL, Schneeberger T, Nunnari F, Gebhard P, Altmann U, Wessler J. Look what I Made it do - the ModelIT method for manually modeling nonverbal behavior of socially interactive agents. In: Proceedings of the 25th International Conference on Multimodal Interaction. 2023 Presented at: ICMI '23; October 9-13, 2023; Paris, France p. 200-204 URL: https://tinyurl.com/nhbysh68 [doi: 10.1145/3610661.3616549]
73. Cresswell K, Cunningham-Burley S, Sheikh A. Health care robotics: qualitative exploration of key challenges and future directions. J Med Internet Res 2018 Jul 04;20(7):e10410 [FREE Full text] [doi: 10.2196/10410] [Medline: 29973336]
74. Bickmore TW, Pfeifer LM, Jack BW. Taking the time to care: empowering low health literacy hospital patients with virtual nurse agents. In: Proceedings of the 2009 SIGCHI Conference on Human Factors in Computing Systems. 2009 Presented at: CHI '09; April 4-9, 2009; Boston, MA p. 1265-1274 URL: https://dl.acm.org/doi/10.1145/1518701.1518891 [doi: 10.1145/1518701.1518891]
75. Erekson DM, Lambert MJ, Eggett DL. The relationship between session frequency and psychotherapy outcome in a naturalistic setting. J Consult Clin Psychol 2015 Dec;83(6):1097-1107. [doi: 10.1037/a0039774] [Medline: 26436645]
76. Huckvale K, Venkatesh S, Christensen H. Toward clinical digital phenotyping: a timely opportunity to consider purpose, quality, and safety. NPJ Digit Med 2019 Sep 6;2(1):88 [FREE Full text] [doi: 10.1038/s41746-019-0166-1] [Medline: 31508498]
77. Sachan D. Self-help robots drive blues away. Lancet Psychiatry 2018 Jul;5(7):547. [doi: 10.1016/s2215-0366(18)30230-x]
78. Fiske ST, Cuddy AJ, Glick P. Universal dimensions of social cognition: warmth and competence. Trends Cogn Sci 2007 Feb;11(2):77-83. [doi: 10.1016/j.tics.2006.11.005] [Medline: 17188552]
79. Harris LT, Fiske ST. Dehumanizing the lowest of the low: neuroimaging responses to extreme out-groups. Psychol Sci 2006 Oct 01;17(10):847-853. [doi: 10.1111/j.1467-9280.2006.01793.x] [Medline: 17100784]
80. Fischer K, Jungermann H. Rarely occurring headaches and rarely occurring blindness: is rarely=rarely? The meaning of verbal frequentistic labels in specific medical contexts. J Behav Decis Making 1996;9(3):153-172. [doi: 10.1002/(sici)1099-0771(199609)9:3<153::aid-bdm222>3.3.co;2-n]
81. Niewiadomski R, Demeure V, Pelachaud C. Warmth, competence, believability and virtual agents. In: Proceedings of the 10th International Conference on Intelligent Virtual Agents. 2010 Presented at: IVA '10; July 12-16, 2010; Philadelphia, PA p. 285 URL: https://link.springer.com/chapter/10.1007/978-3-642-15892-6_29 [doi: 10.1007/978-3-642-15892-6_29]
82. Rosis FD, Pelachaud C, Poggi I, Carofiglio V, Carolis BD. From Greta's mind to her face: modelling the dynamics of affective states in a conversational embodied agent. Int J Hum Comput Interact 2003 Jul;59(1-2):81-118. [doi: 10.1016/s1071-5819(03)00020-x]
83. Black RS, Curran D, Dyer KF. The impact of shame on the therapeutic alliance and intimate relationships. J Clin Psychol 2013 Jun 04;69(6):646-654. [doi: 10.1002/jclp.21959] [Medline: 23382093]
84. Leach MJ. Rapport: a key to treatment success. Complement Ther Clin Pract 2005 Nov;11(4):262-265. [doi: 10.1016/j.ctcp.2005.05.005] [Medline: 16290897]
85. Joinson AN, Woodley A, Reips U. Personalization, authentication and self-disclosure in self-administered internet surveys. Comput Human Behav 2004 Nov 18;23(1):35. [doi: 10.1016/s0747-5632(04)00168-2]
86. Morris GE, Kennedy C. Personal finance questions elicit slightly different answers in phone surveys than online. Pew Rearch Centre. URL: https://www.pewresearch.org/short-reads/2017/08/04/personal-finance-questions-elicit-slightly-different-answers-in-phone-surveys-than-online/ [accessed 2025-09-03]
87. Chen A. Cambridge Analytica and our lives inside the surveillance machine. The New Yorker. 2018. URL: https://www.newyorker.com/tech/annals-of-technology/cambridge-analytica-and-our-lives-inside-the-surveillance-machine [accessed 2025-09-03]
88. Gilbert P. The relationship of shame, social anxiety and depression: the role of the evaluation of social rank. Clin Psychol Psychother 2000 Jul;7(3):174-189. [doi: 10.1002/1099-0879(200007)7:3<174::aid-cpp236>3.0.co;2-u]
89. Lutwak N, Ferrari JR. Shame-related social anxiety: replicating a link with various social interaction measures. Anxiety Stress Coping 2007 May 29;10(4):335-340. [doi: 10.1080/10615809708249307]

XSL•FO
RenderX

90. Rheinberger D, Baffsky R, McGillivray L, Zbukvic I, Dadich A, Larsen ME, et al. Examining the feasibility of implementing digital mental health innovations into hospitals to support youth in suicide crisis: interview study with young people and health professionals. JMIR Form Res 2023 Nov 16;7:e51398. [doi: 10.2196/51398] [Medline: 37971790]
91. Lv N, Kannampallil T, Xiao L, Ronneberg CR, Kumar V, Wittels NE, et al. Association between user interaction and treatment response of a voice-based coach for treating depression and anxiety: secondary analysis of a pilot randomized controlled trial. JMIR Hum Factors 2023 Nov 06;10:e49715 [FREE Full text] [doi: 10.2196/49715] [Medline: 37930781]
92. Lee E, Agustines D, Woo BK. Selection bias in digital conversations on depression before and during COVID-19. JMIR Form Res 2023 Nov 20;7:e42545 [FREE Full text] [doi: 10.2196/42545] [Medline: 37983077]

Abbreviations

RQ: research question
SIA: socially interactive agent
VTA: virtual therapeutic assistant

Edited by A Mavragani; submitted 15.04.24; peer-reviewed by A AL-Asadi; comments to author 07.11.24; revised version received 25.12.24; accepted 03.02.25; published 26.03.25.

Please cite as:
Immel D, Hilpert B, Schwarz P, Hein A, Gebhard P, Barton S, Hurlemann R
Patients' and Health Care Professionals' Expectations of Virtual Therapeutic Agents in Outpatient Aftercare: Qualitative Survey Study
JMIR Form Res 2025;9:e59527
URL: https://formative.jmir.org/2025/1/e59527
doi:10.2196/59527
PMID:

©Diana Immel, Bernhard Hilpert, Patricia Schwarz, Andreas Hein, Patrick Gebhard, Simon Barton, René Hurlemann. Originally published in JMIR Formative Research (https://formative.jmir.org), 26.03.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

AI-Enhanced Virtual Reality Self-Talk for Psychological Counseling: Formative Qualitative Study

Moreah Zisquit[1], MA; Alon Shoa[2], MSc; Ramon Oliva[3], PhD; Stav Perry[1], BA; Bernhard Spanlang[4], PhD; Anat Brunstein Klomek[1], Prof Dr; Mel Slater[3], Prof Dr; Doron Friedman[2], Prof Dr

[1]Baruch Ivcher School of Psychology, Reichman University, Herzliya, Israel
[2]Sammy Ofer School of Communications, Reichman University, Herzliya, Israel
[3]Event Lab, University of Barcelona, Barcelona, Spain
[4]Kiin S.L., Barcelona, Spain

Corresponding Author:
Moreah Zisquit, MA
Baruch Ivcher School of Psychology
Reichman University
Baruch Ivcher School of Psychology
Ha'universita 8
Herzliya
Israel
Phone: 972 523091406
Email: Zisquit.moreah@post.runi.ac.il

Abstract

Background: Access to mental health services continues to pose a global challenge, with current services often unable to meet the growing demand. This has sparked interest in conversational artificial intelligence (AI) agents as potential solutions. Despite this, the development of a reliable virtual therapist remains challenging, and the feasibility of AI fulfilling this sensitive role is still uncertain. One promising approach involves using AI agents for psychological self-talk, particularly within virtual reality (VR) environments. Self-talk in VR allows externalizing self-conversation by enabling individuals to embody avatars representing themselves as both patient and counselor, thus enhancing cognitive flexibility and problem-solving abilities. However, participants sometimes experience difficulties progressing in sessions, which is where AI could offer guidance and support.

Objective: This formative study aims to assess the challenges and advantages of integrating an AI agent into self-talk in VR for psychological counseling, focusing on user experience and the potential role of AI in supporting self-reflection, problem-solving, and positive behavioral change.

Methods: We carried out an iterative design and development of a system and protocol integrating large language models (LLMs) within VR self-talk during the first two and a half years. The design process addressed user interface, speech-to-text functionalities, fine-tuning the LLMs, and prompt engineering. Upon completion of the design process, we conducted a 3-month long exploratory qualitative study in which 11 healthy participants completed a session that included identifying a problem they wanted to address, attempting to address this problem using self-talk in VR, and then continuing self-talk in VR but this time with the assistance of an LLM-based virtual human. The sessions were carried out with a trained clinical psychologist and followed by semistructured interviews. We used applied thematic analysis after the interviews to code and develop key themes for the participants that addressed our research objective.

Results: In total, 4 themes were identified regarding the quality of advice, the potential advantages of human-AI collaboration in self-help, the believability of the virtual human, and user preferences for avatars in the scenario. The participants rated their desire to engage in additional such sessions at 8.3 out of 10, and more than half of the respondents indicated that they preferred using VR self-talk with AI rather than without it. On average, the usefulness of the session was rated 6.9 (SD 0.54), and the degree to which it helped solve their problem was rated 6.1 (SD 1.58). Participants specifically noted that human-AI collaboration led to improved outcomes and facilitated more positive thought processes, thereby enhancing self-reflection and problem-solving abilities.

Conclusions: This exploratory study suggests that the VR self-talk paradigm can be enhanced by LLM-based agents and presents the ways to achieve this, potential pitfalls, and additional insights.

XSL•FO
RenderX

(JMIR Form Res 2025;9:e67782) doi:10.2196/67782

KEYWORDS

virtual human; large language model; virtual reality; self-talk; psychotherapy; artificial intelligence; AI

Introduction

Background

Overview

Access to mental health services is a critical issue on a global scale. Major depression is the leading cause of living with disability for many years and the fourth leading cause of disability-adjusted life years worldwide [1]. Furthermore, research indicates that >20% of people will experience mental illness in their lifetime [2]. Unfortunately, the current clinical workforce is insufficient, with approximately 9 psychiatrists per 100,000 people in high-income countries [3]. This shortfall in care has prompted interest in technology as a potential solution to bridge the gap between the need for treatment and the capacity to deliver it, particularly through the development of conversational agents or multipurpose virtual assistants [4]. These artificial intelligence (AI) agents can be integrated into virtual reality (VR) to simulate a personal, realistic therapeutic environment to reduce feelings of unnaturalness for the patient.

However, creating a virtual therapist presents significant challenges. Despite advancements in generative AI, the reliability of AI in sensitive therapeutic roles remains uncertain [4]. Some researchers argue that AI-based therapists should be used as tools to support both patients and therapists rather than serving as direct replacements for human therapists [5]. Building on this perspective, we propose the integration of an AI agent into the VR counseling experience as explored by Osimo et al [6] and Slater et al [7]. In this framework, participants embody 2 avatars: one representing themselves as a patient and the other as a counselor, allowing an “outside-looking-in” perspective that facilitates self-directed advice and reflection. However, sometimes self-conversation can become stuck, and the patient needs outside help. We contend that large language models (LLMs), while not yet fully capable of replacing human therapists, can function in this more limited role as an effective aid to help the conversation along, thereby enriching the self-talk process. This paper details the integration of an LLM-based agent into the VR self-talk paradigm. We outline the iterative design and development process, share the lessons learned, and present the first successful version, culminating in a small qualitative study in the context of motivational interviewing [8].

Self-Talk in VR

In times of crisis, individuals may find that their capacity to analyze issues and help themselves is limited. Their ability to gain insight is constrained, and it often feels like there is only one solution or, in more dire circumstances, “no way out.” Interestingly, contemplating the problem as an onlooker or from a friend’s perspective can improve problem-solving capabilities, a phenomenon known as Solomon’s paradox [9]. The paradox states that people reason more wisely regarding other people’s problems than they do about their own. Furthermore, a straightforward linguistic change from using “me or I” to “you” has also been shown to increase the psychological distance from personal problems, consequently alleviating the distress they cause and fostering cognitive flexibility [10]. Hence, a relatively straightforward shift in language and perception can enhance problem-solving capabilities.

Both findings are at the base of the VR self-conversation paradigm, first introduced by Osimo et al [6] (Figure 1). The virtual environment comprises a consultation room and 2 avatars. One avatar represents the participants as themselves, and the other represents the participants as the counselor. The participant is able to swap between the 2 avatars, thus enabling a unique experience of talking to a representation of oneself. The VR self-talk experience is based on the ability of participants to switch in and out of virtual bodies in VR. Multisensory integration between the virtual body and real body yields a strong illusion of ownership of the virtual body [7]. The virtual body moves with the participant’s movements, and the participants can see their virtual body from the first-person perspective, and in a virtual mirror. The participants (as well as their avatars) are seated, and the illusion is based on the upper body—hand and head tracking (head is visible only in the mirror; Figure 2, top).

XSL•FO
RenderX

Figure 1. Screenshot of the counseling scenario interface. Top right inset shows a woman wearing a virtual reality headset (not a study participant). She is embodied as a female avatar reflected in a virtual mirror, with a same-sex counselor avatar seated across from her. A virtual control panel in front of her gives her instructions regarding how to proceed with the self-counseling experience and features an artificial intelligence help button.

The initial study compared 2 counselor avatars; 1 group interacted with an avatar resembling Sigmund Freud, while the other engaged with an additional avatar resembling themselves. The results indicated that using the self-talk paradigm proved beneficial in finding more satisfying solutions and improved mood, with the Freud group exhibiting better outcomes. Slater et al [7] showed that body swapping led to superior results compared to a prescripted and animated Freud virtual body and that body ownership of the Freud body was an important contributor.

Since then, these results have been replicated in multiple domains. VR self-talk has been shown to improve motivation for weight loss in individuals with obesity when they conversed with a future self who had lost and maintained weight over a hypothetical 5 years [11]. Convicted offenders spoke to their future selves, reducing self-defeating behaviors, such as alcohol consumption and violent behavior toward others [12]. An additional study showed that using VR self-talk when the "counselor" is an avatar of a leading athlete, such as Serena Williams or LeBron James, could improve adherence to an exercise regime (D Levy, unpublished data, June 2021). All these examples imply that using VR self-talk enhances cognitive flexibility, subsequently facilitating behavioral changes that previously appeared unattainable. Importantly, the design of such paradigms requires care and attention; a study using stereoscopic video revealed that besides the potential benefits of self-talk in VR, under some circumstances, the experience can also "backfire" for some specific populations [13].

Despite the success of VR self-talk, Doron Friedman, Prof Dr, and Mel Slater, Prof Dr, report (email, March 20, 2019) that

XSL•FO
RenderX

people occasionally tend "to get stuck" in the conversation and cannot provide a suitable solution or advice. Sometimes, participants want to end the session or do not know how to continue after a certain point. Often, people run out of advice from the counselor's point of view. This is where AI can play an important role. Using LLMs, we can provide individuals with the external help and ideas they need to continue making progress in their self-talk sessions. This use of AI can also be considered an intermediate step toward automated therapists.

Automated Dialogue

With advancements in machine learning, natural language processing, and deep neural networks, AI has become more capable of performing complex tasks and understanding human language [14]. One of the most notable developments is the emergence of conversational agents, such as ChatGPT, based on LLMs [15]. LLMs have seen tremendous advances in recent years, thanks to increased computational power and the availability of vast datasets for training. Models today, with many billions of parameters, can generate remarkably humanlike text and engage in dialogue while demonstrating some reasoning capabilities [16]. However, significant challenges remain in making these models more aligned with human values, interpreting instructions correctly, and generating factually accurate statements [17]. Current models may generate plausible but incorrect or nonsensical text. Thus, while LLMs today are impressive and valid for specific applications, they require close monitoring and oversight before being deployed into sensitive real-world settings, such as the clinical psychology domain [18].

Embodied AI applications have increasing relevance in mental health applications ranging from social robots to artificially intelligent virtual agents [19]. These applications aim to improve the quality of care and control expenditures [20]. In addition, they are also an important pathway through which the availability of therapeutic treatments may expand [21]. While AI mental health technologies continue to advance rapidly, a significant challenge remains in their successful implementation within clinical settings, where both practitioners and patients face barriers to adoption [22].

The psychological implications of the representation of the virtual therapist have been studied for decades [23], showing mixed results; sometimes, a human appearance enhanced the effectiveness of an application, while at other times, it did not. A meta-analysis by Weber et al [24] demonstrated that adding a face (as opposed to just voice or text) was more significant than the effect of realism; that is, there was greater gain in impact from having a face than from making that face more photographically or behaviorally realistic. Demeure et al [25] evaluated the impact of an agent's representation (in a nonimmersive environment) and perceived emotion over the perceived social believability in the agent. They found that appropriate emotions conveyed through the agent's body, mainly related to the sense of competence and warmth, could lead to higher believability.

LLMs can be the missing piece to the puzzle of VR in therapeutic mental health treatment. In this study, we examined whether LLMs can support participants in their self-talk when they run out of advice and understand how we can use an AI character to help people feel safe and not judged. We describe lessons learned during the development of this unique protocol, which involves a range of advanced techniques such as body ownership illusions [6], self-talk in VR, and LLM-based virtual humans [26]. Finally, we evaluate the final version of the system with a qualitative study.

Iterative Design and Pilot Studies: Lessons Learned

Integrating LLM into VR self-talk requires text-to-speech and speech-to-text functionalities. To address this, we used MILO [27]. The system transcribes participant voice responses from within the VR session to text, which then serves as input for the LLM. Once the LLM generates a response, the text is converted to speech and played through the Unity application.

Initially, the AI was represented as a "help" button with voice functionality (Figure 1). Participants were instructed that a virtual counselor was listening to their conversation, and they could receive its input if they felt stuck and did not know how to continue the self-conversation. We have evaluated this as part of a study comparing VR self-talk with a physical-world equivalent setup, the empty chair technique from Gestalt therapy [28]. However, of the 11 participants in this pilot study, only 1 (9%) engaged with AI. It turned out that a short psycho-education tutorial regarding basic emotion regulation skills followed by VR self-talk resulted in an overwhelming experience; all participants were highly engaged in the session and did not remember they could ask AI for "help." The minimal AI representation within the virtual environment, as only a button, can explain why it was ignored. In additional pilot evaluations, the combination of less-than-perfect speech recognition with less-than-perfect dialogue capabilities (refer to the information presented subsequently) often resulted in low-quality generated responses from the agent.

To make the AI agent more salient in the experience, we replaced the clickable button with a virtual human (Figure 2, bottom). This required addressing the repetition of texts; a by-product of the VR self-talk paradigm is that each text is repeated twice. First, it is spoken by the participant live and recorded, and next, it is played back by the corresponding avatar to be experienced by the participant after the body switch. When there was a third character in the scene, these repetitions became confusing. Consequently, we restricted access to the AI only from the "counselor" avatar and instructed participants that interaction with the AI avatar was recommended only between speech turns.

XSL•FO
RenderX

Figure 2. Screenshots from the counseling scenario. (A and B) Self-talk virtual reality (VR) as seen through the virtual mirror, featuring male and female avatars seated across from Barack Obama and Michelle Obama avatars, accordingly (both depicting the first VR experience without artificial intelligence). (C) Male participant's view, seated across from Barack Obama, with Einstein avatar representing artificial intelligence present, second VR experience. (D) Virtual mirror view showing male avatar, Barack Obama avatar, and Einstein avatar.

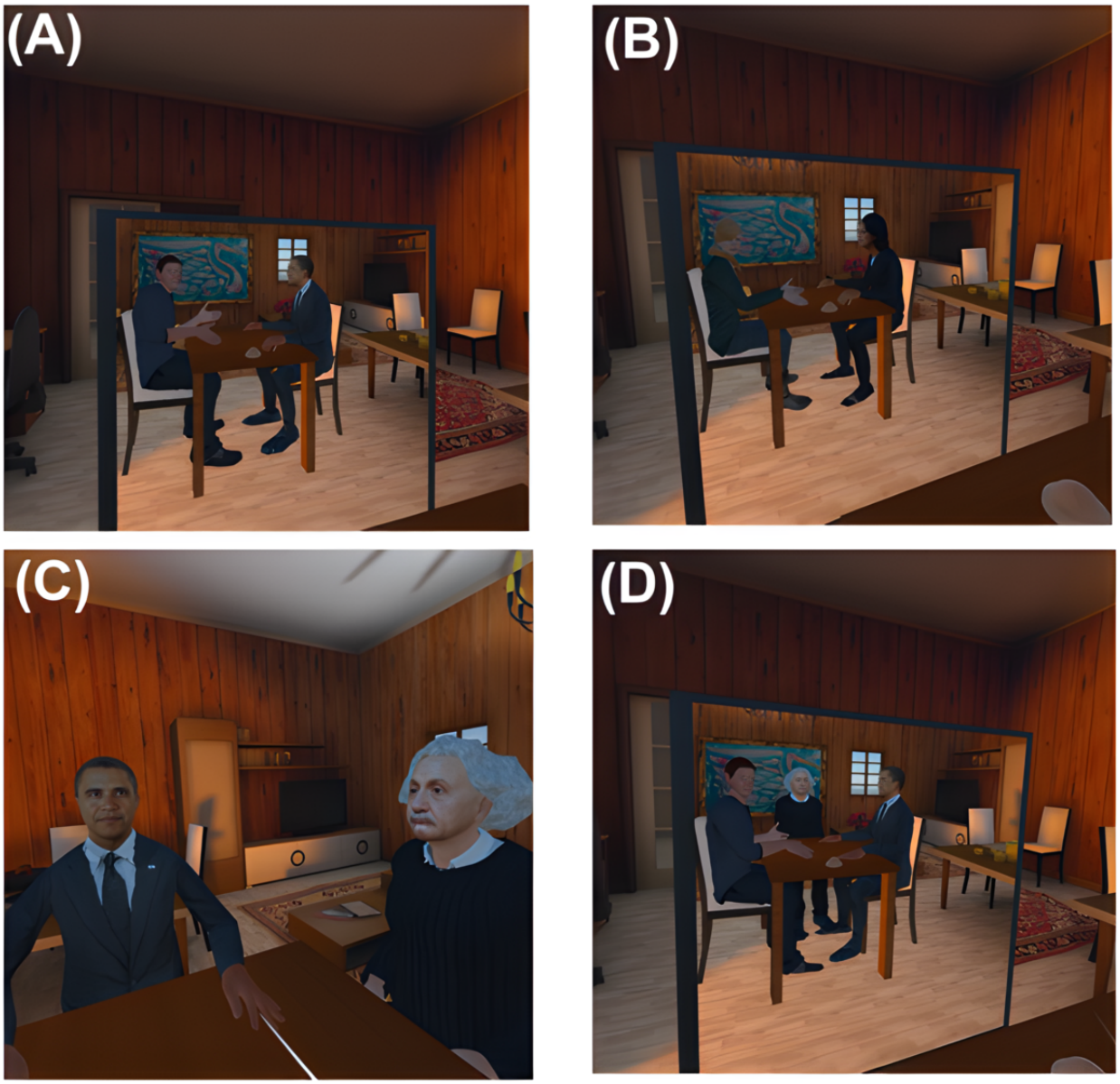

Furthermore, significant progress was made in LLMs over the 2 years of development. Initially, we used a specific model developed in the laboratory. This was based on the 7B parameters version of the GPT-J model [29], fine-tuned on the 2 volumes of published counseling and psychotherapy data from Alexander Street Press. The volumes are searchable collections of transcripts containing real counseling and therapy sessions and first-person narratives illuminating the experience of mental illness and treatment. The 2 volumes contain 3500 session transcripts and >700,000 utterances between a counselor and a patient. We fine-tuned the model with an 80% to 20% train-test split. While this model was state of the art at the time of the early sessions, rapid developments in LLMs rendered it obsolete; therefore, we replaced it with a pretrained model by OpenAI (described in Materials and Equipment in Methods section).

Our LLM interface provides an interface for a human operator, which can be used by the experimenters (refer to the study by Shoa et al [27] for full details). The ongoing automatically transcribed conversation appears in a text window. The operator can select parts of the text or even modify the text and send specific parts to the voice playback in the application. We have also explored the possibility of allowing the experimenter to decide when to intervene in the conversation.

To allow for a realistic experience, we designed and implemented the AI avatar to be gaze activated; that is, the virtual human representing the AI played an idle and silent animation loop and only spoke after the participant, while in the body of the counselor, stared at it for a duration of 1 second. Such gaze activation requires careful tuning—if the gaze duration is too short or the gaze area too broad, false positives

XSL•FO
RenderX

might occur, whereas otherwise, the activation becomes unnatural. In our postexperimental interviews, none of the participants complained about this type of activation.

Another challenge was response latency. The delay includes the following: (1) waiting for the participant to finish talking (indicated by clicking a controller button), (2) gaze activation of the AI avatar, (3) speech recognition in the cloud, (4) latency of one LLM predication, mostly depending on server configuration and context length, and (5) cloud-based text-to-speech conversion. To mitigate latency, the system can be used in a continuous mode; that is, the system is prompted to keep generating responses after each counselor utterance based on the ongoing conversation, regardless of whether the AI agent was activated or not.

Therefore, the consequence is that latency mostly depends on LLM prediction call latency. When using laboratory-based LLMs running from our own server, we could control the latency. The overall round trip would take several seconds, and this was acceptable in the context of a counseling experience. Using OpenAI resulted in depending on OpenAIs' response time. In our study, there were no major delays, and none of the participants in the postexperiment interview (as described subsequently) complained about latency.

Finally, we note that due to limitations in AI tools in many languages, we had to carry out the sessions in English rather than the local native language. Hence, the participants were selected to participate in the study based on their level of English speaking, ranging between conversational level English and native. Language support limitations include all parts of the pipeline: recognition, dialogue, and generation quality.

Methods

Overview

The system was iteratively refined and tested as described before until pilot studies indicated that it operated smoothly. Consequently, we carried out a qualitative explorative study.

Because the overall experience was very rich and overwhelming (as described subsequently), we opted for an exploratory study, carrying out careful sessions, including the presence of a trained clinical psychologist. Although it was made clear to the participants that the system was experimental and this was not a real counseling session, it was clear from the responses that participants behaved almost like in a "real" counseling session, and most participants went through a psychologically meaningful session, as evident from in-depth postexperiment interviews (refer to the Results section).

Participants

The study population comprised 14 anonymous university students (female participants: n=5, 36%) aged 19 to 26 years (mean 23.43, SD 2.21 years). Participants who self-declared as English speakers, had no history of epilepsy, and were not pregnant at the time of the study were invited to participate in the study as part of their credit requirements for their BA psychology degree. In total, 3 (21%) of the participants were not included in the final sample due to miscellaneous technical problems that came up during their session. Given the limited sample size, data saturation was not reached.

Materials and Equipment

Overview

The VR environment featured a consultation room. The VR simulation was developed with the Unity game engine (version 2020.3.20f; Unity Technologies). The VR headset Quest 2 (Meta Platforms, Inc), including its handheld controllers for embodiment (upper body tracking), was used. Quest 2 has a single liquid crystal display panel for each eye, with a display resolution of 1832×1920 pixels. The refresh rate of the panel is 120 Hz. Its weight is approximately 500 g, and it has a head strap that ensures comfort during prolonged use. In addition, Quest 2 delivers a comprehensive 6 dfs, providing participants with both rotational and positional tracking capabilities.

The software includes a VR self-talk experience and integration of an LLM-based agent based on MILO. In the exploratory study, we used Barack Obama and Michelle Obama, sex matched, as the counselor avatars. The AI avatar was Albert Einstein. The participants were initially embodied in a generic sex-matched avatar (Figure 2). The first part of the VR experience was intended to strengthen the virtual body ownership illusion. The participants engaged in a short "embodiment" exercise, alternately moving both hands, looking around, and seeing themselves in the mirror. Following the embodiment exercise, the participants were asked to describe a current life issue. When they finished, the participants pressed a button, transitioning them to the counselor avatar. Previous studies found that it is best to model the counselor avatar based on famous inspiring persons; we selected Barack Obama in our studies. Next, the participants pressed a button and listened to a replay of what they just said, spoken by the avatar representing themselves as patients. A prompt instructed them to respond "like a counselor," and their responses were recorded. The participants' voice pitches were changed in the recording to avoid having it sound exactly like the participant, as described in the study by Osimo et al [6]. Once completed, participants returned to their look-alike avatar to continue the conversation in this iterative manner.

We used 2 VR self-talk implementations; one was a commercial product called ConVRself (Kiin Tech) used during the iterative design, and the other was a research version with very similar design and functionality used during the study.

For LLM, we used GPT-3.5 using the following prompt: "You will now act as a motivational interviewer with a lot of experience in interpersonal psychotherapy. I am a counselor and will turn to you for advice while speaking to my patient. Sometimes, I will get stuck and address you, and in this case, you should try to act as an expert counselor and say something that would help me progress the session as best as possible. Please do not mention that you are an artificial intelligence in your replies. Act as a real person."

Semistructured Interviews

Participants' experiences and thoughts regarding the AI were assessed using semistructured interviews that lasted

XSL•FO
RenderX

approximately 15 minutes, and the audio of the interview was recorded. Due to the exploratory nature of this complex experience, we opted for the flexibility offered by semistructured interviews. We constructed 14 questions in advance, such as, "Please describe this experience as if you were describing it to a friend who was not here," (Multimedia Appendix 1). The questions were presented individually to the participants, and additional questions were added based on the participants' responses. The interviews were conducted by one of the main authors (MZ), a female clinical psychologist and PhD candidate, with appropriate training. There was no relationship between the interviewer and the participants before the study commencement, and the participants were informed of her professional expertise.

The actual conversation of the participants was logged by default as part of the system, including the AI comments. However, we opted not to analyze the conversations due to the privacy of the participants, and the content was removed.

Applied thematic analysis identifies and describes implicit and explicit ideas within the data, consequently linking them into the theoretical framework of the study [30]. The analysis was thus conducted using a bottom-up approach in which the answers provided by the participants determined the main themes rather than the questions asked. In addition, at the start of the interview, participants were asked to provide a description of the experience, to elicit unconstrained participant impressions.

Ethical Considerations

The ethical considerations regarding implementing AI in psychological settings are paramount [31]. The development process and exploratory study outlined in this paper aimed to address preliminary questions concerning effectiveness and validation. In addition, in our study, a clinical psychologist provided professional oversight, which is typically recommended for AI implementation. The study received ethics approval from the institutional review board of Reichman university (P_2022145). Participants received course credit for their participation in the study and gave written informed consent. All data protection procedures were adhered to. All participants provided informed consent, which explicitly outlined that their responses would be recorded and kept as part of the study data but would be saved anonymously and that they would interact with AI during the study, receiving advice from it regarding a personal problem.

Procedure

Upon arrival at the laboratory, the participants were given instructions about the experiment and completed a consent form. They were told that they would have the opportunity to talk about a personal problem that caused them an average level of distress (on a scale of 1-10, a problem rated between 4 and 7). Then, they wrote a short sentence describing the problem in their words. Importantly, although it was made clear to the participants that the session was experimental and should not be considered psychological counseling, our sessions were carried out in the presence of a trained clinical psychologist in a private laboratory setting in which only the participant and the psychologist were present. Next, the participants donned the VR headset and performed the VR experience in which they conversed freely between their sex-matched avatars and matching Obama characters (Michelle for female participants and Barack for male participants). After a back-and-forth conversation between the participant and Obama avatar, the participants chose when to end the session.

After the first session, the participants answered two questions regarding their experience as follows: (1) Please rate this experience on a scale of 1 to 10 (1=not useful at all and 10=very useful) and (2) On a scale of 1 to 10, how helpful was this self-conversation for solving your problem?

Next, it was explained to them that they would perform a similar session, only this time, another third avatar, depicted as Albert Einstein, would be seated in the virtual consultation room. They were told that it would be listening to their conversation and that they should turn to him during the conversation to receive his input but only from the embodied counselor's perspective. The gaze activation was explained, and they were instructed to turn to Einstein only from the Obama avatar. The participants proceeded to conduct the AI-enhanced iterative VR self-conversation, and when they were done, the semistructured interview was conducted.

Results

Overview

A total of 5 (45%) of the 11 participants chose to speak about a problem related to work or school. Another 5 (45%) participants chose to speak about relationship problems, and 1 (9%) participant chose to speak about difficulty managing stress.

The semistructured interviews were transcribed and analyzed using the thematic coding method by 2 of the coauthors. The transcripts were not reviewed by the participants. A total of 5 (45%) participants preferred the session with the AI, 4 (36%) preferred the session without it, and the remaining 2 (18%) did not address this. In total, 7 (64%) participants said that the AI's attitude toward them was positive, 1 (9%) said it was neutral, and the remaining 3 (27%) did not address this.

The participants were asked 2 questions in the break between the first session (without AI) and the second (with AI). On average, the usefulness of the session was rated 6.9 (SD 0.54); the median value was 7 (IQR 7-7). On average, the degree to which it helped solve their problem was rated 6.1 (SD 1.58), with a median value of 6 (IQR 6-7). At the end of the interview, the participants were asked to rate the extent to which they would like to come back for another session on a scale of 1 to 10; the mean response was 8.3 (SD 1.55), and the median was 8.5 (IQR 7-9), indicating a very high level of satisfaction from the experience.

The applied thematic analysis qualitative analysis revealed 4 high-level themes, which were derived from the data. These 4 themes are outlined subsequently. The participants did not give feedback on the findings outlined below.

Quality of Advice

From the thematic analysis of the participants' responses, a layered or complex pattern emerged regarding the perceived

XSL•FO
RenderX

quality of AI-generated advice. On the one hand, participants acknowledged the AI's ability to provide insightful, well-structured guidance:

> *...Einstein brought a lot of emotional considerations to the two-way problem I didn't think about. It was interesting...It sounds like if I would listen to him like everything would work out.* [S6]
>
> *For my questions it was very, very beneficial. It was. It was able to provide very thorough and clear and insightful points and information regarding specifically related to my problem related to the career path and choosing a career path. And it was really cool to see how the interaction with the computer algorithm was able to really contribute to my own self-reflection.* [S8]

They viewed the advice as emotional while sustaining a thorough analysis of the situation. However, sometimes this very sophistication created a feeling of uncanniness, with participants noting that it was "inhumanly smart." Some participants pointed toward the notable tension between the logical precision of AI and the inherently "messy" nature of emotional problems, suggesting that the AI's responses tended to be overly orderly or systematic:

> *I think that on the one hand it was good, and he was saying smart things, but on the other hand it was a little bit too logical and technical, and sometimes emotional problems are a little more complex.* [S4]
>
> *...Good as I said, but it was a little bit surface level.* [S7]
>
> *Like robotic smart. Like human, but inhuman smart. You know what I'm saying? Like he was too good of a psychologist. I don't know how to say it. It felt like a nice thing to say, but feels like I don't know how to describe it too, like, kitsch...The AI surprised me...Yeah, the AI was kind of freaking yeah. Like it felt. It felt like too good. I don't know how even to describe it...Yeah, but also, I mean it was like in between. It was like really compassionate and understanding and like what the perfect answer from a psychologist should be. But the perfect answer is like not, not human to some degree. It also felt like he just blabbered a bunch of information, like instantly to me, like, Oh yeah, it's OK to be compassionate and it's OK to be. You shouldn't be perfect all the time, like, in one sentence.* [S13]

There was also evidence pointing to potential overtrust in AI responses, which has been mentioned before [32]. This might suggest that careful consideration of AI authority is necessary in therapeutic settings due to increased credibility attributed to AI-reorganized versions of the participants' own thoughts:

> *I think he just had a way of wording the same ideas that I had in my head, but sort of organizing them to a more sophisticated or organized manner which made them more believable or reliable, seem more reliable.* [S11]

In addition, some of the "complaints" could also be a reasonable approach to a human counselor:

> *S6: ...that it was not really practical, but it was like another part that you need to think about.*
>
> *Question: What wasn't practical about it?*
>
> *S6: Like for my problem. I need an answer, it's like a yes or no question and he really talked about like what it made you feel and maybe rephrase the problem, and there's no way to really rephrase it.*

Self-AI Collaboration

In some cases, participants highlighted the complementary nature of the combination between the AI advice and their own input, suggesting a unique form of alliance where AI and human intelligence united and enhanced each other's strengths. The AI's tendency to a more structured approach was balanced effectively with the more emotionally attuned self-counsel, creating what one participant even termed as a "perfect" mixture. This combination led to a more comprehensive experience. As mentioned, rather than viewing AI as a replacement for traditional therapy, its optimal role, as the participants experienced, was as a tool that enhanced their self-reflection and personal insight:

> *...I liked the mix of the AI and myself as the counselor...I think the advice I gave myself was taking into account more like an emotional feeling and maybe even a little bit too much. And he was a little too technical. So maybe a mixture of them? Yeah, a mixture of them would be perfect.* [S4]
>
> *I think in terms of advice, it was similar to what I said, although the AI was able to provide a more detailed and well-structured answer. The AI's examples and ideas were a bit more beneficial.* [S8]
>
> *And I liked the mix of him and Michelle together, that I, as Michelle gave a bit of a softer input and he was like, do this and this and this and this and that together was like, they filled each other...No, I think they were really complementary to each other. I benefited very much from both of them. I'm really grateful. I got to talk to myself...OK, more minds better.* [S9]

Believability and Attitude Toward AI

The analysis suggested a notable dichotomy between the cognitive aspect and the perceptual issues of the experience. While the content and reasoning capabilities of the AI were valuable and convincingly humanlike, the nonverbal and visual elements created "breaks in presence" [33]. These elements were mainly manifested in voice quality, robotic movements, and lack of natural pauses in speech turns:

> *...yeah, the sound and obviously the, the graphics obviously. But in terms of the syntax and the structure of how the answers were provided and delivered then it was fairly good. Fairly satisfying.* [S8]
>
> *I feel like a mixture, like the words that he was saying felt like they made sense, they were human-like, but*

XSL•FO
RenderX

> *the way that he was telling them and like it was very technical...* [S4]
>
> *The conversation felt like very human and Einstein Avatar was a little bit robotic and felt like scripted and just a little bit like not related...No, the look was OK.* [S5]
>
> *I'm asking a question and then we'll just speak like we talk and talk and talk and talk...Humans usually wait a little bit.* [S7]
>
> *Technical stuff, which is like the character will move a little bit differently than human, and that's pretty much.* [S7]

However, a phenomenon of gradual acceptance regarding the participants' ability to adapt to the limitations of technology aligned with established research on place illusion in immersive VR environments [34]. In other words, the participants wanted to believe what they were experiencing in such a way that they were willing to ignore the technical imperfections that created disbelief. This process of adaptation was evident in the participants' ability to shift their focus from the robotic characteristics to the contextual value of the dialogue. In this way, the content served as a compensatory factor, allowing participants to overcome initial reservations regarding the artificial nature of the interaction. While technical refinements and improvements in nonverbal behavior are called for, the therapeutic value of the AI-advice might be more dependent on the quality of the dialogue rather than the behavioral human mimicry:

> *I think it's still not 100% human like but it did feel like I was talking to a very insightful person because at the end of the day I was I was focusing more on the information that was given to me by the algorithm rather than the actual feeling and and the the sound of the the interaction itself.* [S8]
>
> *...at once I got over the metallic voice, yeah, it felt a bit more natural...It's like a fake character. It's an avatar, but I think when you get in, in into it, it's become less and less weird. Less and less, Yeah.* [S9]

The Use of Celebrity Avatars

The data suggest that the selection of well-known figures as therapeutic avatars creates complex psychological dynamics that can either enhance or impede the experience. Participants' reactions demonstrated that the personal associations with these celebrity avatars substantially influenced their engagement with the AI and with their own counselor persona. For instance, responses to the Obama avatar highlighted the importance of perceived authority and strength, while reactions to Einstein often centered on his scientific reputation and intellectual credibility. However, these preexisting associations could also act as barriers, with some participants expressing a preference for avatars they could "relate to and love more." This finding suggests that the effectiveness of celebrity avatars in therapeutic settings may be highly individualized, with the same figure potentially eliciting different responses based on personal preconceptions. The data particularly emphasized how the avatar's identity can affect the reception of therapeutic advice, with some participants explicitly attributing the perceived value of the guidance to the celebrity's reputation rather than the content itself:

> *And I think Obama is quite a good character to use, actually. I feel that she's confident. She's a woman, I guess like a strong woman. She's very powerful. It seems like it's very fun to sit around her and like look at her and talk to her.* [S5]
>
> *That's the point, Talking to Obama, even if, like, I don't agree with him that much. But like, yeah, it's a cool person to talk to.* [S7]
>
> *I think I would change Barack Obama's character to someone that I love. As I said, one of the reasons I felt more comfortable with Einstein is because that's a character I can relate to and love more.* [S10]
>
> *Again, maybe it's just because it's Einstein. I mean if like a human being would say something like that to me but would phrase it differently and, you know, have different facial expressions like not like the computer Einstein, then I would feel differently about the answer.* [S13]
>
> *...Maybe because it was Einstein. Like maybe if it was someone else maybe I wouldn't think about like it was but just because it was him and I know that he was like really big scientist so he might have like great things to say.* [S14]

Discussion

Principal Findings

Mental health and well-being are major challenges worldwide, and the demand for counseling is much greater than the supply. Mental health systems worldwide are under immense strain, and in the face of growing need, many underqualified or pseudoprofessional services have appeared. As a result, an increasing number of stakeholders are turning to AI-based solutions, including LLM-driven tools, as a potential quick fix (this includes Israel's minister of science [January 3, 2025, personal communication]). However, we advocate a more measured approach, emphasizing rigorous evaluation and clear delineation of clinical versus nonclinical use. Our intervention explicitly targets nonclinical populations (ie, focusing on mental health improvement rather than formal therapy) and relies on AI only as an adjunct to self-talk rather than a stand-alone therapeutic agent. Such caution is vital given the nascent state of these technologies and the ethical implications tied to safety, efficacy, and user well-being. Accordingly, we strongly support further research and stepwise validation, coupled with the ongoing involvement of trained professionals before any large-scale implementation.

Technologies, such as VR and AI, may be a part of the solution. VR self-talk is unique in suggesting counseling without the need of either a human in the loop or AI and has been shown in the past to be beneficial and effective. Adding an LLM assistant is a natural next step; nevertheless, introducing an LLM-based AI counselor assistant into VR is not trivial, and

XSL•FO
RenderX

our iterative design process has revealed several important lessons.

Integrating multiple challenging technologies, such as VR, body tracking, dialogue, voice recognition, and speech generation into a meaningful psychological experience proved arduous. While each of these technologies separately has made impressive progress recently, accumulated problems or errors easily resulted in an unusable experience. In the case described here, it seems that during 2023, the underlying AI technologies, specifically speech recognition and LLM-based dialogue, passed an essential threshold, increasing the probability of a successful and meaningful experience regardless of the complex nature of the technological design.

While both VR and AI, individually and in combination, present significant opportunities for counseling, their widespread deployment in real-world settings remains challenging [35]. The contribution of each of these components as well as their combined contribution needs to be carefully evaluated [36]. Industry trends indicate that integrating AI models into a wide range of applications is becoming viable, for example, applications such as Tess and Woebot [37]. VR is gradually becoming more accessible, though its availability remains somewhat limited. We propose that the experience of "interacting with yourself in the third person," facilitated by VR and virtual body–ownership illusions, offers unique benefits that are unlikely to be replicated on 2D screens. In addition, the integration of LLMs into text and voice-based interactions is expected to become more widely accessible. Therefore, we hope that this study, despite its challenges and complexities, can serve as a foundation for further advancements in this field.

On the basis of the themes that emerged from the semistructured interviews, we can conclude that participants generally responded positively to the AI advice, with reactions ranging from mixed to favorable. None of the participants evaluated it as poor or as obstructive to the self-talk process. Considering the complexity of the experience, this result is very encouraging. VR self-talk, when performed correctly, is a very powerful experience, though it can be confusing even without the addition of AI. To address this, we implemented a gradual protocol, allowing the participants to accustom themselves to the VR self-talk before introducing the AI agent in a subsequent session. This protocol was necessary for our preliminary research. However, future studies could benefit from a research design such as randomized controlled trials that could compare self-talk with and without AI, providing a deeper understanding of the intricate nature of this experience and the added value of AI to it.

The overarching goal of this endeavor was to design and implement an AI agent within a self-talk experience. We envisioned the AI advice as complementing the participants' own self-guidance rather than replacing their need to think, reflect, and reason with themselves regarding the issue at hand. This vision was in alignment with the participants' responses. They found it straightforward to implement the AI advice into their self-dialogue and quickly adapted to the flow of the conversation. Our results further support the perspective that the integration of technology into counseling has the potential to enhance skills and abilities that humans already possess [16].

In our case, an "AI" button was not enough, and we opted to integrate a gaze-activated avatar to embody the AI. While participants commented that the voice and animation of the avatar were not completely realistic, several of the participants indicated that they were able to overcome such limitations and focus on the social interaction as well as the content of the conversation.

Finally, most of the participants pointed to the need to improve design elements, such as voice, body language, avatar appearance, and character selection. There was a relative consensus among participants regarding the need for improvements in this area to enhance the overall experience. This is not surprising as the level of realism has been shown to have a strong impact on affective responses of participants [26,27]. Now that we have successfully integrated multiple technologies and created a seamless psychological experience, future development and testing should focus on refining and improving these design aspects.

Limitations

This study has several limitations that should be addressed. Although multiple pilot studies were conducted throughout the iterative design process, including approximately 30 additional participants, the final sample consisted of only 11 participants. The relatively small sample size underscores the necessity for further research with larger and more diverse populations that could confirm these initial results. However, as emphasized throughout this paper, this is a novel and promising application of technology, and every exploration must begin somewhere. Due to the psychological focus of this study, further research should explore specific participant characteristics that could influence the effectiveness of self-talk in VR. Furthermore, research could target particular symptoms of psychopathology or specific diagnoses to improve the understanding of potential clinical applications.

Another limitation is the bias of novelty. The use of VR and AI in a psychological setting could have potentially fascinated the participants, leading to an inflated sense of efficacy and distorting the true impact of the experience. In addition, semistructured interviews could have contributed to social desirability and acquiescence bias, as participants might have felt inclined to provide responses they expected would be favorable to the researcher. Further research could explore ways to mitigate these identified biases.

Technological limitations are the lack of systematic analysis of the model's responses and their adherence to the therapeutic principles outlined in the prompt. The quality of the model's response is influenced by both the prompt and the internal architectures of the LLM itself, in our case GPT-3.5, which has proven highly capable but is proprietary and continuously evolving. The initial study used a relatively simple prompt, instructing the AI to adopt the role of an experienced motivational or interpersonal psychotherapist. While this approach was based on established research [38], it did not undergo formal pilot testing specifically targeting the fidelity

XSL•FO
RenderX

of the model's alignment with the therapeutic technique. This decision was made, in part, due to the exploratory nature of this study, which aimed to assess the potential use of LLMs in VR-based self-counseling rather than conduct an in-depth evaluation of their adherence to specific therapeutic techniques. Nonetheless, we acknowledge that future work would benefit from piloting more refined prompts, including detailed instructions for eliciting key therapeutic elements.

The field faces significant challenges in ensuring replicability and consistency, particularly when relying on such models. As a result, the research community must consider how to establish more robust methods for quality control and model evaluation to ensure the ongoing utility and ethical deployment of LLMs in therapeutic settings. Recommendations include clinical evaluation of the LLMs developed and their output; interdisciplinary collaboration; and attending to risk assessment, transparency, and bias [36]. We were able to implement some of the recommendations in our initial study, such as clinical evaluation of output and interdisciplinary collaboration, and further studies could improve on others.

Conclusions

The introduction of new technologies could potentially transform psychotherapy, giving rise to numerous potential challenges, limitations, and ethical considerations that should be addressed. Currently, clinical expertise is necessary for best practice; however, that may evolve in the future, and even now, monitoring and validating can be more cost-effective for certain patients compared to traditional psychotherapy. Given our encouraging results, we suggest that the paradigm of AI-enhanced VR self-talk may be ready for further research on a larger sample from the general population and studies with clinical populations. In addition, our work suggests there are numerous potential opportunities for integrating AI into VR wellness, extending beyond just "automated therapist agents."

Acknowledgments

The authors thank Peleg Shinar, bachelor of arts, humanities and social sciences research assistant, who assisted during the iterative design process. This research was supported by the Horizon 2020 Information and Communication Technologies (ICT) project - Leadership in Enabling and Industrial Technologies - SOCRATES (#951930) and European Union's Horizon 2020 program H2020-FETPROACT-2020-2 Research and Innovation - GuestXR (#101017884). MS is supported by the ERC Advanced Grant MoTIVE (#742989).

Data Availability

The datasets generated or analyzed during this study are not publicly available due to the confidential nature of the content, but numerical results will be made available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Complete interview protocol used in this study.
[DOCX File , 17 KB - formative_v9i1e67782_app1.docx]

References

1. Murray CJ, Vos T, Lozano R, Naghavi M, Flaxman AD, Michaud C, et al. Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet 2012 Dec 15;380(9859):2197-2223 [FREE Full text] [doi: 10.1016/S0140-6736(12)61689-4] [Medline: 23245608]
2. Trautmann S, Rehm J, Wittchen HU. The economic costs of mental disorders: do our societies react appropriately to the burden of mental disorders? EMBO Rep 2016 Sep;17(9):1245-1249 [FREE Full text] [doi: 10.15252/embr.201642951] [Medline: 27491723]
3. Oladeji BD, Gureje O. Brain drain: a challenge to global mental health. BJPsych Int 2016 Aug;13(3):61-63 [FREE Full text] [doi: 10.1192/s2056474000001240] [Medline: 29093905]
4. Vaidyam AN, Wisniewski H, Halamka JD, Kashavan MS, Torous JB. Chatbots and conversational agents in mental health: a review of the psychiatric landscape. Can J Psychiatry 2019 Jul;64(7):456-464 [FREE Full text] [doi: 10.1177/0706743719828977] [Medline: 30897957]
5. Torous J, Bucci S, Bell IH, Kessing LV, Faurholt-Jepsen M, Whelan P, et al. The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. World Psychiatry 2021 Oct 09;20(3):318-335 [FREE Full text] [doi: 10.1002/wps.20883] [Medline: 34505369]
6. Osimo SA, Pizarro R, Spanlang B, Slater M. Conversations between self and self as Sigmund Freud--a virtual body ownership paradigm for self counselling. Sci Rep 2015 Sep 10;5:13899 [FREE Full text] [doi: 10.1038/srep13899] [Medline: 26354311]

XSL•FO
RenderX

7. Slater M, Neyret S, Johnston T, Iruretagoyena G, Crespo M, Alabèrnia-Segura M, et al. An experimental study of a virtual reality counselling paradigm using embodied self-dialogue. Sci Rep 2019 Jul 29;9(1):10903 [FREE Full text] [doi: 10.1038/s41598-019-46877-3] [Medline: 31358846]
8. Hettema J, Steele J, Miller WR. Motivational interviewing. Annu Rev Clin Psychol 2005;1:91-111 [FREE Full text] [doi: 10.1146/annurev.clinpsy.1.102803.143833] [Medline: 17716083]
9. Grossmann IE, Kross E. Exploring Solomon's paradox: self-distancing eliminates the self-other asymmetry in wise reasoning about close relationships in younger and older adults. Psychol Sci 2014 Aug;25(8):1571-1580 [FREE Full text] [doi: 10.1177/0956797614535400] [Medline: 24916084]
10. Kross E, Ayduk O. Self-distancing: theory, research, and current directions. In: Advances in Experimental Social Psychology. New York, NY: Academic Press; 2017.
11. Anastasiadou D, Herrero P, Vázquez-De Sebastián J, Garcia-Royo P, Spanlang B, Álvarez de la Campa E, et al. Virtual self-conversation using motivational interviewing techniques to promote healthy eating and physical activity: a usability study. Front Psychiatry 2023;14:999656 [FREE Full text] [doi: 10.3389/fpsyt.2023.999656] [Medline: 37151962]
12. van Gelder JL, Cornet LJ, Zwalua NP, Mertens EC, van der Schalk J. Interaction with the future self in virtual reality reduces self-defeating behavior in a sample of convicted offenders. Sci Rep 2022 Feb 10;12(1):2254 [FREE Full text] [doi: 10.1038/s41598-022-06305-5] [Medline: 35145177]
13. Landau D, Hasler B, Golland Y, Huebbe B, Idan O, Magnat M, et al. A self-compassion experience in immersive video: opportunities and pitfalls. PsyArXiv Preprint posted online on November 28, 2022 [FREE Full text] [doi: 10.31234/osf.io/z4q9f]
14. Wei J, Tay Y, Bommasani R, Raffel C, Zoph B, Borgeaud S, et al. Emergent abilities of large language models. arXiv Preprint posted online on June 15, 2022 [FREE Full text]
15. Black S, Biderman S, Hallahan E, Anthony Q, Gao L, Golding L, et al. GPT-NeoX-20B: an open-source autoregressive language model. arXiv Preprint posted online on April 14, 2022 [FREE Full text]
16. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, et al. Chain-of-thought prompting elicits reasoning in large language models. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. 2022 Presented at: NIPS'22; November 28-December 9, 2022; New Orleans, LA.
17. Menick J, Trebacz M, Mikulik V, Aslanides J, Song F, Chadwick M, et al. Teaching language models to support answers with verified quotes. arXiv Preprint posted online on March 21, 2022 [FREE Full text]
18. Hagendorff T. Machine psychology: Investigating emergent capabilities and behavior in large language models using psychological methods. arXiv Preprint posted online on March 24, 2023 [FREE Full text]
19. Fiske A, Henningsen P, Buyx A. Your robot therapist will see you now: ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. J Med Internet Res 2019 May 09;21(5):e13216 [FREE Full text] [doi: 10.2196/13216] [Medline: 31094356]
20. Abd-Alrazaq AA, Rababeh A, Alajlani M, Bewick BM, Househ M. Effectiveness and safety of using chatbots to improve mental health: systematic review and meta-analysis. J Med Internet Res 2020 Jul 13;22(7):e16021. [doi: 10.2196/16021] [Medline: 32673216]
21. Holohan M, Buyx A, Fiske A. Staying curious with conversational AI in psychotherapy. Am J Bioeth 2023 May;23(5):14-16 [FREE Full text] [doi: 10.1080/15265161.2023.2191059] [Medline: 37130403]
22. Holmes EA, Ghaderi A, Harmer CJ, Ramchandani PG, Cuijpers P, Morrison AP, et al. The Lancet Psychiatry Commission on psychological treatments research in tomorrow's science. Lancet Psychiatry 2018 Mar;5(3):237-286 [FREE Full text] [doi: 10.1016/s2215-0366(17)30513-8]
23. Cresswell K, Cunningham-Burley S, Sheikh A. Health care robotics: qualitative exploration of key challenges and future directions. J Med Internet Res 2018 Jul 04;20(7):e10410 [FREE Full text] [doi: 10.2196/10410] [Medline: 29973336]
24. Weber S, Weibel D, Mast FW. How to get there when you are there already? Defining presence in virtual reality and the importance of perceived realism. Front Psychol 2021;12:628298 [FREE Full text] [doi: 10.3389/fpsyg.2021.628298] [Medline: 34025504]
25. Demeure V, Niewiadomski R, Pelachaud C. How is believability of a virtual agent related to warmth, competence, personification, and embodiment? Presence Teleoperators Virtual Environ 2011;20(5):431-448 [FREE Full text] [doi: 10.1162/PRES_a_00065]
26. Slater M. Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. Philos Trans R Soc Lond B Biol Sci 2009 Dec 12;364(1535):3549-3557 [FREE Full text] [doi: 10.1098/rstb.2009.0138] [Medline: 19884149]
27. Shoa A, Oliva R, Slater M, Friedman D. Sushi with Einstein: enhancing hybrid live events with LLM-based virtual humans. In: Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents. 2023 Presented at: IVA '23; September 19-22, 2023; Würzburg, Germany URL: https://doi.org/10.1145/3570945.3607317 [doi: 10.1145/3570945.3607317]
28. Zisquit M, Klomek AB, Shoa A, Slater M, Spanlang B, Friedman D. Intrapersonal emotion regulation training in virtual reality: embodying self-talk. Virtual Real 2025 (forthcoming).

XSL•FO
RenderX

29. Bell IH, Nicholas J, Alvarez-Jimenez M, Thompson A, Valmaggia L. Virtual reality as a clinical tool in mental health research and practice . Dialogues Clin Neurosci 2022 Apr 01;22(2):169-177 [FREE Full text] [doi: 10.31887/dcns.2020.22.2/lvalmaggia]
30. Guest G, MacQueen KM, Namey EE. Applied Thematic Analysis. Thousand Oaks, CA: SAGE Publications; 2012.
31. Imel ZE, Caperton DD, Tanana M, Atkins DC. Technology-enhanced human interaction in psychotherapy. J Couns Psychol 2017 Jul;64(4):385-393 [FREE Full text] [doi: 10.1037/cou0000213] [Medline: 28318277]
32. Hitron T, Morag NM, Erel H. Implications of AI bias in HRI: risks (and opportunities) when interacting with a biased robot. In: Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction. 2023 Presented at: HRI '23; March 13-16, 2023; Stockholm, Sweden URL: https://doi.org/10.1145/3568162.3576977 [doi: 10.1145/3568162.3576977]
33. Garau M, Widenfeld HR, Antley A, Friedman D, Brogni A, Slater M. Temporal and spatial variations in presence: a qualitative analysis. In: Proceedings of the 7th Annual International Workshop on Presence. 2004 Presented at: PRESENCE 2004; October 13-15, 2004; Valencia, Spain. [doi: 10.1162/pres.17.3.293]
34. Spanlang B, Normand J, Borland D, Kilteni K, Giannopoulos E, PomÃ©s A, et al. How to build an embodiment lab: achieving body representation illusions in virtual reality. Front Robot AI 2014 Nov 27;1 [FREE Full text] [doi: 10.3389/frobt.2014.00009]
35. Bell IH, Pot-Kolder R, Rizzo A, Rus-Calafell M, Cardi V, Cella M, et al. Advances in the use of virtual reality to treat mental health conditions. Nat Rev Psychol 2024 Jul 10;3(8):552-567 [FREE Full text] [doi: 10.1038/s44159-024-00334-9]
36. Stade EC, Stirman SW, Ungar LH, Boland CL, Schwartz HA, Yaden DB, et al. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. Npj Ment Health Res 2024 Apr 02;3(1):12 [FREE Full text] [doi: 10.1038/s44184-024-00056-z] [Medline: 38609507]
37. Stephens TN, Joerin A, Rauws M, Werk LN. Feasibility of pediatric obesity and prediabetes treatment support through Tess, the AI behavioral coaching chatbot. Transl Behav Med 2019 May 16;9(3):440-447 [FREE Full text] [doi: 10.1093/tbm/ibz043] [Medline: 31094445]
38. Reynolds L, McDonell K. Prompt programming for large language models: beyond the few-shot paradigm. arXiv Preprint posted online on February 15, 2021 [FREE Full text] [doi: 10.1145/3411763.3451760]

Abbreviations

AI: artificial intelligence
LLM: large language model
VR: virtual reality

Edited by A Mavragani; submitted 21.10.24; peer-reviewed by F Shojaei, B Meyer; comments to author 14.12.24; revised version received 18.01.25; accepted 10.02.25; published 02.04.25.

Please cite as:
Zisquit , Shoa A, Oliva R, Perry S, Spanlang B, Brunstein Klomek A, Slater M, Friedman D
AI-Enhanced Virtual Reality Self-Talk for Psychological Counseling: Formative Qualitative Study
JMIR Form Res 2025;9:e67782
URL: https://formative.jmir.org/2025/1/e67782
doi:10.2196/67782
PMID:

© Moreah Zisquit, Alon Shoa, Ramon Oliva, Stav Perry, Bernhard Spanlang, Anat Brunstein Klomek, Mel Slater, Doron Friedman. Originally published in JMIR Formative Research (https://formative.jmir.org), 02.04.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Factors Impacting Mobile Health Adoption for Depression Care and Support by Adolescent Mothers in Nigeria: Preliminary Focus Group Study

Lola Kola[1,2], PhD; Tobi Fatodu[1], MPH; Manasseh Kola[3], BEng; Bisola A Olayemi[1], MBBS; Adeyinka O Adefolarin[1,4], PhD; Simpa Dania[5], MBBS; Manasi Kumar[6], PhD; Dror Ben-Zeev[7], PhD

[1]WHO Collaborating Centre for Research and Training in Mental Health, Neurosciences and Drug and Alcohol Abuse, Department of Psychiatry, College of Medicine, University of Ibadan, Ibadan, Nigeria

[2]Centre for Global Mental Health, Kings College London, London, United Kingdom

[3]School of Computer Science, University of Leicester, Leicester, United Kingdom

[4]Department of Health Promotion and Education, College of Medicine, Faculty of Public Health, University of Ibadan, Ibadan, Nigeria

[5]Association for Reproductive and Family Health (ARFH), Ibadan, Nigeria

[6]Department of Population Health, School of Medicine, Institute for Excellence in Health Equity, New York University, New York, NY, United States

[7]Department of Psychiatry and Behavioral Sciences, BRiTE Center, University of Washington, Seattle, WA, United States

Corresponding Author:
Lola Kola, PhD
WHO Collaborating Centre for Research and Training in Mental Health, Neurosciences and Drug and Alcohol Abuse, Department of Psychiatry
College of Medicine
University of Ibadan
P.M.B 3017 G.P.O
Ibadan, 23401
Nigeria
Phone: 234 (2) 2411768
Email: lola_kola2004@yahoo.com

Abstract

Background: Mobile health (mHealth), the use of mobile technology in health care, is increasingly being used for mental health service delivery even in low- and middle-income countries to scale up treatment, and a variety of evidence supports their potential in different populations.

Objective: This study aims to use the Social Cognitive Theory (SCT) as a lens to explain knowledge of mHealth use for mental health care, personal behavioral capabilities, and the external social contexts that can impact the adoption of an mHealth app for depression care among perinatal adolescents in Nigeria.

Methods: At the preliminary stage of a user-centered design (UCD), 4 focus group discussions were conducted among 39 participants: 19 perinatal adolescents with a history of depression and 20 primary care providers. Guided by the SCT, a popular model used for predicting and explaining health behaviors, we documented participants' knowledge of mHealth use for health purposes, advantages, and challenges to the adoption of an mHealth app by young mothers, and approaches to mitigate challenges. Data collection and analysis was an iterative process until saturation of all topic areas was reached.

Results: The mean age for young mothers was 17.3 (SD 0.9) years and 48 (SD 5.8) years for care providers. Mistrust from relatives on mobile phone use for therapeutic purposes, avoidance of clinic appointments, and sharing of application contents with friends were some challenges to adoption identified in the study population. Supportive personal factors and expressions of self-efficacy on mobile app use were found to be insufficient for adoption. This is because there are social complications and disapprovals that come along with getting pregnant at a young age. Adequate engagement of parents, guardians, and partners on mHealth solutions by care providers was identified as necessary to the uptake of digital tools for mental health care in this population.

Conclusions: The SCT guided the interpretations of the study findings. Young mothers expressed excitement at the use of mHealth technology to manage perinatal depression. Real-life challenges, however, need to be attended to for successful

XSL•FO
RenderX

implementation of such interventions. Communications between care providers and patients' relatives on the therapeutic use of mHealth are vital to the success of a mHealth mental health management plan for depression in young mothers in Nigeria.

(JMIR Form Res 2025;9:e42406) doi:10.2196/42406

KEYWORDS

adolescent perinatal depression; primary care; mHealth app; user centered design; smartphone; human-centered design; HCD; depression; postpartum; perinatal; postnatal; teenage; adolescent; youth; low-middle-income countries; LMIC; middle income; adoption; acceptability; mobile health; mHealth; mobile app; women's health; mental health; depressive

Introduction

The World Health Organization (WHO) strongly recommends the rapid widespread use of technology as a possible solution for transforming pressing public health services [1]. Mobile Health (mHealth), the use of mobile technology in health care, is increasingly being used for mental health service delivery even in low- and middle-income countries to scale up treatment, and a variety of evidence supports their potential in different populations [2]. Mobile phones are ubiquitous globally and in low-income countries with evidence of use in different population groups [3]. Recent reports indicate their high use among the youth perinatal population in Nigeria [4]. In Nigeria, digital research has been successfully conducted for perinatal mental care within the WHO Mental Health Global Action Programme (mhGAP) task-shifting initiative in primary care [4-6] to address barriers to care in patients. Task shifting, which is a process of delegating tasks to less specialized health care workers where appropriate, is recommended by the WHO to bridge the treatment gap in global mental health [7]. Digital health care presents avenues to further scale-up care in the process of task shifting [1]. Technology-based studies for perinatal care in Nigeria have been implemented to increase pregnant women's access to care, but several knowledge gaps exist on factors that influence the use and adoption of mHealth in adolescent mothers [8,9]. Addressing these gaps is crucial for increasing access in a vulnerable population. Perinatal adolescents, often unmarried, might have their pregnancies disputed by the father of their child and could also have little or no social support from family members [8,9] necessitating the need for age and context-specific health care for this vulnerable population.

Perinatal depression, a condition common in pregnant women, is higher among adolescent mothers than older mothers [10,11] occurring during pregnancy and up to one year after childbirth. Untreated perinatal depression is a risk for negative health outcomes for mothers and their infants [12]. There are limited resources for managing mental disorders in primary care in low-income countries [13] and access for young mothers is even more limited by multiple age-related access barriers, which often translate to self and social stigma [8]. The complexities of hindrances to care experienced by this population indicate the need for practical and flexible interventions tailored to meet their unique needs [14].

Mobile phone apps, which are software programs designed to work on mobile devices, can serve as mHealth tools used in real-time for mental health care [15,16]. Mobile phone apps offer easy and affordable digital access to mental health information and solutions and can be used to increase internet-based access to mental health services. Evidence supports the involvement of users in the designing and developing of technology-based interventions to increase their potential for success and uptake [17]. User-centered designs [18] allow users to participate in iterative developments and design processes of mHealth tools, for reflection of users' needs in treatment designs. The use of mHealth even in low-resource settings [2] and the willingness of young mothers to engage in mobile phone use for health care services [4] present several opportunities for the design of technology-based interventions that attend to their unique needs [19].

In this study, we explored the benefits, barriers, and mitigating factors to mHealth adoption for depression by adolescent mothers in Nigeria through a theoretical lens. Theories play a crucial role in explaining health behaviors, and in digital health care they have provided frameworks for predicting and influencing human behaviors in the context of health and well-being. [20,21]. In this study, however, we used the Social Cognitive Theory (SCT), a widely recognized theory of behavior change [22] in explaining and understanding participants' knowledge of mHealth use for mental health care (personal experiences), personal behavioral capabilities of adolescent mothers to engage with a mobile phone app for depression care (behavioral factors), and external social context (environmental factors) that may impact their expected health outcomes [23].

Focus group discussions (FGDs) were carried out with adolescent mothers as part of the second stage of a user-centered design (UCD) of a mHealth app that was designed to complement face-to-face care for perinatal depression within routine primary care. The question guide explored participants' knowledge of mHealth use, the perceived advantages and challenges of application use, and approaches to mitigate identified challenges. The research questions were as follows:

1. What is the knowledge of mHealth use for mental health care among young mothers and care providers in primary care?
2. What are the potential barriers and facilitators of mobile app adoption in this population? What challenges could limit its use?
3. What are the possible strategies to mitigate the challenges to mHealth apps adoption by young mothers?

Methods

Participants and Study Setting

The study was conducted in Ibadan, southwest Nigeria using FGDs: 2 FGDs of young mothers (n=19), and 2 separate ones

XSL•FO
RenderX

among care providers (n=20). The venue was a seminar room at the Department of Psychiatry, University of Ibadan that was chosen because it was private and easily accessible using public transportation. All 4 FGDs took place in a 2-week period in February 2020. Participants were recruited from the database of an ongoing trial on perinatal depression in primary care [24]. A total of 30 participants aged between 16 and 19 years had completed treatment in an ongoing trial (Responding to the challenge of Adolescent Perinatal Depression [RAPiD]) [24], and those in remission from clinical depression as indicated by their hospital records were selected. Only 22 of them had phone numbers that rang through on the mobile network. The phone numbers of the remaining 8 were no longer available on the mobile networks. Even though all 22 young mothers who were contacted by phone indicated their availability for the FGDs, 3 participants later called to indicate their unavailability on the days chosen for the discussions, leaving 19 participants who attended the FGDs. On the part of the care providers, 20 primary care providers from the same trial's intervention arm were invited, and all accepted and participated in the FGDs. All participants were briefed on the purpose of the FGDs when they were contacted on the phone. The young mothers and the clinicians were selected from clinics in the 11 local government areas in Ibadan. On the days of the FGDs, the facilitator gave additional detailed information on the purpose of the FGDs. All participants signed an informed consent form after giving verbal consent to participate in the study. In total, 3 groups had 10 participants each, while the fourth (a young mother's group) had 9 participants.

FGD Procedure

The average duration for the FGDs was 90 minutes. Author LK, a medical sociologist, moderated all 4 groups using predetermined open-ended questions. Author TF was the timekeeper and note-taker in all the discussions. Both LK and DA had no previous contact or interactions with study participants. Following the introductions, the facilitator asked the initial question on knowledge of mobile phone use for health purposes. Participants who indicated such knowledge were asked to describe the interventions they were aware of. The facilitator probed on the mode of delivery of such interventions identified, the information content, and the disease conditions they were used for. Participants were then asked about their familiarity with mobile phone apps, and irrespective of the answer given, the facilitator showed all participants a video-based mobile phone app on an Infinix Hot 7 Pro Android phone. This demonstration ensured that participants understood what an mHealth app was before asking them related questions. All FGDs were audiotaped (for research purposes only), transcribed verbatim, and analyzed shortly after they were completed. Therefore, data collection and analysis became an iterative process until saturation of all topic areas was reached. Responses of participants were deidentified, and the transcribed data was not shared with anyone outside the investigating team.

Data Analysis

The data were transcribed by 2 experienced transcriptionists TA (a research assistant) and DA (co-author). Authors LK and DA conducted thematic coding of the transcripts manually along with the areas of interest. Specifically, LK created the initial code, which the authors reviewed. Identical subcodes were merged while redundant ones were removed. The interview transcripts were then read to understand and familiarize by LK and DA and independently coded to highlight texts relevant to the derived themes. The process led to a descriptive account focused on identifying critical variables for each heading [25].

SCT was used to explain the interpretations of positive supports for behavior changes in areas of personal, behavioral, and environmental factors [26]. In this investigation, personal factors (eg, previous knowledge of mHealth risks and benefits) were identified as prerequisites for behavior change but not the sole factor for the needed changes. Also, given the current dearth of mHealth use in mental health care in Nigeria and in the study population of young mothers, the point of inquiry on previous knowledge of use focused on all known areas of use within general health care. Second, the behavioral capability of adolescent mothers, their beliefs in their ability (self-efficacy) to engage in mobile phone app use for health purposes to carry out desired actions and produce desired effects, was assessed within the present social context on the roles and autonomy of adolescent girls as foundational to their likelihood of adopting and maintaining mHealth use [27,28]. Third, we documented outcome expectancies of engaging in mobile app use through participants' perceptions of the advantages and disadvantages of such action with the aim of identifying motivating and mitigating factors to assist in achieving desired results [29].

Ethical Considerations

The University College Hospital, University of Ibadan, ethical review board approved the study in accordance with the 1964 Declaration of Helsinki and its later amendments (UI/EC/19/0165). In this study, only perinatal adolescents aged 16-19 years who could give consent to participate, and primary health care providers over the age of 18 years were recruited. Study data from observations and audio recordings during the FGDs were stored as an mp4 file on a password-protected encrypted computer at the Department of Psychiatry, University of Ibadan. Group facilitators took notes that strictly contained only deidentified information. All the participants approached were given full information on study purposes and procedures and told that their participation was voluntary, before they were invited to participate. The informed consent form had two original copies for those who agreed to participate: one of these was stored securely in the research file and a copy was given to the participant. All participants gave voluntary responses to interview questions and were told they could decline to answer any questions that they chose. The adolescent girls received a total sitting and transport allowance of the equivalent of US $10, while the primary care providers received US $15.

Results

The mean age for young mothers was 17.3 (SD 0.9) years; they were predominantly single (14/19, 73.7%) with a few full cohabiting (4/19, 21.1%). The age of mothers who had some high school education was 10.6 (SD 1.9) years: 8 out of 19 (42.1%) of them were still students, 2 out of 19 (10.5%) of these were first-year university undergraduates); and 11 out of 19

XSL•FO
RenderX

(57.8%) others had dropped out of school and engaged in petty trading. The care providers were middle-aged (mean 48, SD 5.8 years) majorly female (18/19, 95%), and with significant (mean 22.2, SD 4.7 years) work experience. A summary of the demographic information of participants is presented in Tables 1 and 2. Our open-ended question guide unveiled several responses according to the themes presented in Tables 1 and 2 below.

Table 1. Demographic information of women participants.

Variables	Values
Marital status, n (%)	
Single	14 (74)
Married	1 (5)
Cohabiting	44 (21)
Occupation, n (%)	
Students	8 (42)
Petty trading	11 (58)
Age of perinatal adolescent participants (year), mean (SD)	17.3 (0.9)
Years of education of adolescent participants, mean (SD)	10.6 (1.9)

Table 2. Demographic information of care provider participants.

Variable	Values
Sex of care providers, n (%)	
Male	1 (5)
Female	19 (95)
Total	20 (100)
Marital status, n (%)	
Married	20 (100)
Age of care provider participants (year), mean (SD)	48 (5.8)
Years of work experience, mean (SD)	22.2 (4.7)

Factors Related to mHealth Use

The perinatal adolescent groups had knowledge of mHealth use for health interventions in reproductive health care and mental health care. The latter knowledge was related to their participation in the ongoing RAPID study [24] and its mHealth component. Group members commonly reported receiving SMS on reproductive health promotion and prevention programs:

> *I receive regular text messages telling me about contraceptive use and different types*
>
> *I know about text messages advising people to space their children*
>
> *I don't know who sends it but I get regular messages in Yoruba (a language in Nigeria) telling me to visit my nearest clinic to ask about family planning methods.*

However, the adolescent mothers were of two minds about the usefulness of health information (SMS) in reproductive care. On the one hand, they saw it as not useful since stigma would prevent them from taking advantage of such a program because of their age:

> *I would be too self-conscious to go and ask for such services, also I don't think health care providers will attend to people unmarried especially young girls who want contraceptives.*

On the other hand, some agreed that the SMS provided useful information that was relevant for the future.

The majority in the perinatal adolescent groups identified their sources of knowledge of mHealth use for mental health care as through mobile phone calls or SMS for follow-up from care providers during their participation in the ongoing RAPID project [24]:

> *Matron would call me to remind me about clinic appointments for depression care on the day before the appointment dates.*
>
> *The nurse sent me a message to remind me to take my baby for immunisation because when she called the day before then I had complained of not having much energy.*
>
> *I got a call from the nurse…she asked how I was doing, the calls from her always lifts my mood.*
>
> *Matron called me the first evening I visited the clinic..she said she wanted to be sure I was okay…she*

XSL•FO
RenderX

also told me not to forget the homework she gave me as part of the treatment for my depression.

All the care providers reported their use of mobile phone apps and SMS in various ways in their practice in maternal and child health care:

…we use a mobile app given to use by the WHO for child immunisation data cordination and we also have other one from a project on family planning which helps us in self learning.

All of the care providers had wide experiences in using phone-to-phone calls for service provision and the WhatsApp (Meta) platform for supervision and support of health interventions.

We call doctors and matrons who are our clinical supervisors on the mobile phone when needed and staff members also chat on group Whatsapp platforms for work purposes.

Perinatal Adolescents and Mobile App Use

Both groups of participants (perinatal adolescents and primary health care providers) had positive comments on the appropriateness of mobile app use in the perinatal adolescent population for depression care, and the abilities of this population to engage its use for depression care. One care provider commented:

…you know that young people like to explore phones so delivering health services to them through this avenue increases their likelihood of engaging in such care. More than nine out of ten of these girls can read and write so accessing health care on the mobile phone is what they will be able to do.

Many of the young mothers noted it could help them to reduce the times they might have to go to the clinics because of the long waiting times they often experience:

Many times at the clinic, I get tired and just want to go home because the crowd can be many…so this kind of app, will help someone like me since it could reduce clinic visits.

A few saw it as a way to overcome access stigma related to being pregnant at a young age and cost-related barriers to health care.

Going to the clinic when I was pregnant was always troubling for me because I felt so self-conscious of my being young compared to adult mothers… so I skipped many clinic days.

Majority of the young mothers were excited about the thoughts of a mobile phone app with short videos for depression self-care, and a few also believed that it could be used by young mothers to improve users' moods and to prevent boredom. They generally found the use of video for treatment on the application very appealing, noting that it will help them recollect their discussions with the nurses at the clinic, which would help consolidate the actions they need to improve mental health:

It will keep me from being bored (shows excitement)…I like the idea because videos will be remembered easily…and If I have such app, I can also use it to remember anything I forget that the nurse said

I watch short videos on face book, so such an app like this will be easy for me.

Many young mothers reported watching short video clips on Facebook (Meta) for leisure. Care providers also agreed that

…a health app would help reinforce treatment from the clinics and help young mothers with self-care.

Reports of both groups on mobile app use within the social context of perinatal adolescents gave very strong indications of the need to get parents, guardians, and partners to be on board and be supportive of its use. The care providers said:

Some parents might think the young mothers are using their phone as an excuse not to help with housework rather than for accessing their depression treatment.

In the same vein, young mothers' perceptions were similar:

If my mother thinks I am staying on the phone too much and texting my boyfriend, she controls the use of my phone

…if I have an app for treatment, my mother would need to be told from the clinic for her because if she sees me staying long on the phone, she could think that I am just playing games and she can seize the phone for that reason.

One care provider had concerns about cases of severe depression that could result in functional limitations, thereby impacting the ability to use the application:

…a few might be too depressed to even use the app or even spend time looking at the treatment on it.

Outcome Expectations of Application Adoption Among Young Mothers

In general, the two participant groups made constant reference to the positive outcome expectancies associated with engaging in mobile app use for the management of depression in this population. Some members in the perinatal adolescent groups showed visible excitement at the possibility of accessing health care through audiovisual materials on a mobile phone app by talking in high-pitched voices:

if I has an app like that when I was sick, I would have gotten better quicker because this would have increase the treatment I received.

An app with videos as treatment will be good… it could help with feelings of sadness and how to take any needed action.

I like watching videos on Facebook…they helped me when I felt down, so this is good because it will help to prevent boredom and sadness.

Although the care providers were generally positive on the outcome expectation of application use on the mental health of perinatal adolescents, 2 of the older nurses were skeptical, given the possibility of increased treatment default rates [11] among

XSL•FO
RenderX

young mothers who replace clinic appointments with application use:

> *…if the app treatment is not properly presented to patients as what they should use along with the face-to-face clinic visits, some might not come as they should.*

This will be more so if a treatment plan for depression modify clinic visits inconsideration for frequency of application use.

> *You know adolescents don't like going to the clinic for appointments, my fear is that they will now not come at all if they have an app for self care.*

Also, another care provider pointed out the possibility of worsening symptoms if patients are unable to use the application, for example, as a result of severe depression. Some care providers have negative consequences of application misuse if patients share treatment contents with people with similar symptoms:

> *The patients would need to be warned against sharing the app with friends or relatives with similar symptoms to avoid severe consequences.*

Mitigating Challenges to mHealth Adoption in Young Mothers

Participants identified 5 action plans to mitigate challenges to young mothers' mHealth use. Both groups of participants agreed on the need for care providers to educate patient's relatives on its usefulness in young mothers' depression care:

> *If the nurse tells my mother that a mobile phone app is for treatment purposes, she will even be the one to encourage me to use it more so I can get well quickly.*

> *My grandmother that I live with will cooperate with whatever they say from the clinic regarding my health.*

The care providers generally agreed that mHealth use should be included as part of the total management plan for perinatal depression to remove possible misconceptions from patients who many see as an alternative to face-to-face care. Care providers and young mothers also suggested the use of phone calls from care providers to encourage and remind young mothers of application use. The majority of the care providers identified the need to warn adolescent mothers on a regular basis about the harms that could result from sharing the mHealth app with relatives and friends with similar symptoms of depression. A few of the care providers pointed out it would be necessary to block young mothers' access to mHealth apps after treatment to guide against misuse. Both caregiver and young mother FGD groups wanted the application attractive and interactive in order to engage young mothers and encourage usage.

Discussion

Principal Findings

Interventions based on behavioral theory use have proven to be more effective than theoretical approaches [30,31] allowing for a deeper understanding of the subjective dimension of the process [32].

To our knowledge, this is the first UCD approach assessing the feasibility of mobile app-based depression intervention for adolescent mothers, while also identifying factors to mitigate possible challenges to use by young mothers in Africa. Our study population (mean 17.3, SD 0.9 years) was adequately educated (mean 10.6, SD 1.9) contrary to findings in previous studies where perinatal adolescents were mostly uneducated [33,34].

On the whole, young mothers' and health workers responded positively to a video-based mHealth app used to manage adolescent perinatal depression [35]. This response could be an inference that there may be a high rate of adoption of such interventions in the future. A significant strength of the study is the participation of key stakeholders who were part of a recent trial. Contrary to a previous report on insufficient knowledge of mHealth use among clinicians in Nigeria [36], this study's primary health care providers were knowledgeable and currently engaged in mHealth use for different health purposes.

Perinatal adolescents had a sense of personal agency (self-efficacy) with respect to using the mHealth app, which was indicative of them having the skills and abilities necessary for performing the behavior under a variety of circumstances. This self-efficacy in young mothers is ultimately linked to their general level of education in young people in Nigeria, which is a result of basic universal education in the country [37].

Young mothers and care providers noted several positive outcome expectations as advantages of application use. Both groups agreed that this mode of digital health service provision could keep young mothers in contact with care as they navigate personal- and clinical-level barriers to care [38]. The general response supported the enthusiasm of young people for technology, which increases the likelihood of their engagement with it and its usefulness for depression care in the group. This finding is supported by a recent survey of the current use of mobile phones among young people in Nigeria, which reported the broad willingness of young mothers to engage in technology use for mental health care [4]. Based on the values and expectations of the young mothers' who reported watching videos on Facebook to entertain themselves, the idea of a video-based application therefore presented something familiar with their current pattern of entertainment, thereby increasing the likelihood of future use.

The young mothers identified the influence of social support in mediating behavior change [39,40] identifying both parents, partners, and health care providers as key stakeholders [24] in a mHealth intervention design for this population with depression. The importance of warning young mothers of the dangers of application sharing and uninstalling the mobile app after completion of treatment were potential ways to prevent misuse of mobile apps by patients. As in previous findings, the risks associated with patients sharing electronic health information with other users were an identified challenge [41].

Study Finding Through an SCT Lens

Unlike other cognitive development theories, SCT incorporates a 3-way relationship between individuals, their self-regulating process of efficacy, and their social circumstances. In this study,

XSL•FO
RenderX

we understand that supportive personal factors and the ability to engage in application use for depression care are insufficient for adoption given the limitations that can occur on the social level. Misunderstanding from parents and partners of its purposes for entertainment rather than for therapy is a main hindrance. This might be more so the case if adolescent mothers spend time away from participating in household activities due to mHealth use, or if parents suspect it is being used for other reasons such as "texting of boyfriend" or "playing games." This necessitated the need for care providers to adequately engage parents, guardians, and partners on the health benefits of mHealth solutions for mental health care and to prepare them for assisting young mothers with future challenges that they might encounter in its use. Such applications in the future also can entail tools for health care providers, parents, and partners, which can help mitigate clinical- and family-level stresses. A combined intervention strategy of mHealth and face-to-face care is a model that provides an opportunity to reduce parents' misconceptions about application use through interactions between care providers and patients' relatives.

Limitations

Data were collected from participants who were part of an ongoing clinical trial on perinatal depression that used mHealth and their current knowledge might have influenced their responses. For this reason, this study result might not be generalizable to views of all care providers or perinatal adolescents. Also, the transferability of the study finding is limited by the fact that young mothers in low-resource settings are majorly students or low-income earners and therefore might not all be able to afford smartphones if after a trial stage the intervention is found to be successful. Despite these limitations, this study provides useful information on how to mitigate identified barriers to adopting an mHealth app for adolescent perinatal depression in primary care.

Conclusions

Young mothers expressed excitement at the use of mHealth technology to manage perinatal depression. Real-life challenges, however, need to be attended to for successful implementation of such interventions. We found that care providers and relatives of young mothers are vital stakeholders in the success of a mHealth mental health management plan to manage depression in young mothers. Our results reinforce the need for management plans that combine face-to-face care in adopting mHealth in young mothers' mental health interventions.

Conflicts of Interest

DB-Z has financial interests in Merlin LLC and FOCUS technology and has provided consultation services to Trusst Health, K Health, Boehringer Ingelheim, eQuility, Deep Valley Labs, Butler Hospital, and Otsuka Pharmaceuticals.

References

1. Recommendations on digital interventions for health system strengthening. World Health Organization. URL: https://www.who.int/publications/i/item/9789241550505 [accessed 2016-06-06]
2. Naslund JA, Aschbrenner KA, Araya R, Marsch LA, Unützer J, Patel V, et al. Digital technology for treating and preventing mental disorders in low-income and middle-income countries: a narrative review of the literature. Lancet Psychiatry 2017;4(6):486-500 [FREE Full text] [doi: 10.1016/S2215-0366(17)30096-2] [Medline: 28433615]
3. Statistics ITUs reports on mobile cellular subscriptions. Sustainable Development Goals. 2019. URL: https://www.itu.int/en/ITU-D/Statistics/pages/stat/default.aspx [accessed 2025-02-05]
4. Kola L, Abiona D, Adefolarin AO, Ben-Zeev D. Mobile phone use and acceptability for the delivery of mental health information among perinatal adolescents in Nigeria: survey study. JMIR Ment Health 2021;8(1):e20314 [FREE Full text] [doi: 10.2196/20314] [Medline: 33496678]
5. Gureje O, Oladeji BD, Araya R, Montgomery AA, Kola L, Kirmayer L, et al. Expanding care for perinatal women with depression (EXPONATE): study protocol for a randomized controlled trial of an intervention package for perinatal depression in primary care. BMC Psychiatry 2015;15:136 [FREE Full text] [doi: 10.1186/s12888-015-0537-3] [Medline: 26122982]
6. Kola L, Abiona D, Oladeji BD, Ayinde O, Bello T, Gureje O. Theory-driven development of a mobile phone supported intervention for adolescents with perinatal depression. Soc Psychiatry Psychiatr Epidemiol 2022;57(6):1201-1210. [doi: 10.1007/s00127-021-02198-3] [Medline: 34779878]
7. Patel V, Saxena S, Lund C, Thornicroft G, Baingana F, Bolton P, et al. The lancet commission on global mental health and sustainable development. Lancet 2018;392(10157):1553-1598. [doi: 10.1016/S0140-6736(18)31612-X] [Medline: 30314863]
8. Kola L, Bennett IM, Bhat A, Ayinde OO, Oladeji BD, Abiona D, et al. Stigma and utilization of treatment for adolescent perinatal depression in Ibadan Nigeria. BMC Pregnancy Childbirth 2020;20(1):294 [FREE Full text] [doi: 10.1186/s12884-020-02970-4] [Medline: 32410586]
9. Kumar M, Huang KY. Impact of being an adolescent mother on subsequent maternal health, parenting, and child development in Kenyan low-income and high adversity informal settlement context. PLoS One 2021;16(4):e0248836 [FREE Full text] [doi: 10.1371/journal.pone.0248836] [Medline: 33793592]
10. Kingston D, Heaman M, Fell D, Chalmers B, Maternity Experiences Study Group of the Canadian Perinatal Surveillance System, Public Health Agency of Canada. Comparison of adolescent, young adult, and adult women's maternity experiences and practices. Pediatrics 2012;129(5):e1228-e1237. [doi: 10.1542/peds.2011-1447] [Medline: 22529278]

XSL•FO
RenderX

11. Gureje O, Oladeji BD, Montgomery AA, Araya R, Bello T, Chisholm D, et al. High- versus low-intensity interventions for perinatal depression delivered by non-specialist primary maternal care providers in Nigeria: cluster randomised controlled trial (the EXPONATE trial). Br J Psychiatry 2019;215(3):528-535. [doi: 10.1192/bjp.2019.4] [Medline: 30767826]
12. Nove A, Matthews Z, Neal S, Camacho AV. Maternal mortality in adolescents compared with women of other ages: evidence from 144 countries. Lancet Glob Health 2014;2(3):e155-e164 [FREE Full text] [doi: 10.1016/S2214-109X(13)70179-7] [Medline: 25102848]
13. Gureje O, Lasebikan VO. Use of mental health services in a developing country. Results from the Nigerian survey of mental health and well-being. Soc Psychiatry Psychiatr Epidemiol 2006;41(1):44-49. [doi: 10.1007/s00127-005-0001-7] [Medline: 16341828]
14. Leplatte D, Rosenblum KL, Stanton E, Miller N, Muzik M. Mental health in primary care for adolescent parents. Ment Health Fam Med 2012;9(1):39-45. [Medline: 23277797]
15. Ben-Zeev D, Kaiser SM, Brenner CJ, Begale M, Duffecy J, Mohr DC. Development and usability testing of FOCUS: a smartphone system for self-management of schizophrenia. Psychiatr Rehabil J 2013;36(4):289-296 [FREE Full text] [doi: 10.1037/prj0000019] [Medline: 24015913]
16. Ventola CL. Mobile devices and apps for health care professionals: uses and benefits. P T 2014;39(5):356-364 [FREE Full text] [Medline: 24883008]
17. Gudmundsdottir GB, Vasbø KB. Methodological Challenges When Exploring Digital Learning Spaces in Education. Cham: Springer; 2014.
18. Abras CM, Preece J. User-centered design. In: Encyclopedia of Human-Computer Interaction. Thousand Oaks, CA: Sage Publications; 2004:445-456.
19. Hussain-Shamsy N, Shah A, Vigod SN, Zaheer J, Seto E. Mobile health for perinatal depression and anxiety: scoping review. J Med Internet Res 2020;22(4):e17011 [FREE Full text] [doi: 10.2196/17011] [Medline: 32281939]
20. Riley WT, Rivera DE, Atienza AA, Nilsen W, Allison SM, Mermelstein R. Health behavior models in the age of mobile interventions: are our theories up to the task? Transl Behav Med 2011;1(1):53-71 [FREE Full text] [doi: 10.1007/s13142-011-0021-7] [Medline: 21796270]
21. Naslund JA, Shidhaye R, Patel V. Digital technology for building capacity of nonspecialist health workers for task sharing and scaling up mental health care globally. Harv Rev Psychiatry 2019;27(3):181-192 [FREE Full text] [doi: 10.1097/HRP.0000000000000217] [Medline: 30958400]
22. Bandura A. Social cognitive theory of self-regulation. Organ Behav Hum Decis Process 1991;50(2):248-287. [doi: 10.1016/0749-5978(91)90022-l]
23. McAlister AL, Perry CL, Parcel GS. How individuals, environments, and health behaviors interact: social cognitive theory. In: Glanz K, Rimer BK, Viswanath K, editors. Health Behavior and Health Education: Theory, Research, and Practice. Hoboken, NJ: Jossey-Bass; 2008:169-188.
24. Gureje O, Kola L, Oladeji BD, Abdulmalik J, Ayinde O, Zelkowitz P, et al. Responding to the challenge of adolescent perinatal depression (RAPiD): protocol for a cluster randomized hybrid trial of psychosocial intervention in primary maternal care. Trials 2020;21(1):231 [FREE Full text] [doi: 10.1186/s13063-020-4086-9] [Medline: 32106885]
25. Ritchie J, Lewis J, Nicholls CM. Qualitative Research Practice: A Guide for Social Science Students and Researchers. Washington, DC: SAGE Publications; 2003.
26. Wong C, Monaghan M. Behavior change techniques for diabetes technologies. In: Diabetes Digital Health. Amsterdam, Netherlands: Elsevier; 2020:65-75.
27. Maddux JE, Kleiman E, Gosselin JT. Self-Efficacy. New York, NY: Guilford Press; 2012.
28. Meltzer-Brody S, Bledsoe-Mansori SE, Johnson N, Killian C, Hamer RM, Jackson C, et al. A prospective study of perinatal depression and trauma history in pregnant minority adolescents. Am J Obstet Gynecol 2013;208(3):211.e1-211.e7 [FREE Full text] [doi: 10.1016/j.ajog.2012.12.020] [Medline: 23246315]
29. Reesor L, Vaughan EM, Hernandez DC, Johnston CA. Addressing outcomes expectancies in behavior change. Am J Lifestyle Med 2017;11(6):430-432 [FREE Full text] [doi: 10.1177/1559827617722504] [Medline: 29302249]
30. Noar SM, Benac CN, Harris MS. Does tailoring matter? Meta-analytic review of tailored print health behavior change interventions. Psychol Bull 2007;133(4):673-693. [doi: 10.1037/0033-2909.133.4.673] [Medline: 17592961]
31. Glanz K, Bishop DB. The role of behavioral science theory in development and implementation of public health interventions. Annu Rev Public Health 2010;31:399-418. [doi: 10.1146/annurev.publhealth.012809.103604] [Medline: 20070207]
32. Stacey FG, James EL, Chapman K, Courneya KS, Lubans DR. A systematic review and meta-analysis of social cognitive theory-based physical activity and/or nutrition behavior change interventions for cancer survivors. J Cancer Surviv 2015;9(2):305-338 [FREE Full text] [doi: 10.1007/s11764-014-0413-z] [Medline: 25432633]
33. Chiazor IA, Ozoya MI, Idowu AE, Udume M, Osagide M. Teenage pregnancy: the female adolescent dilemma. Int J Sci Commer Humanit 2017;5(1):70-82 [FREE Full text]
34. Fleming N, O'Driscoll T, Becker G, Spitzer RF, CANPAGO COMMITTEE. Adolescent pregnancy guidelines. J Obstet Gynaecol Can 2015;37(8):740-756. [doi: 10.1016/S1701-2163(15)30180-8] [Medline: 26474231]
35. Kenny R, Dooley B, Fitzgerald A. Developing mental health mobile apps: exploring adolescents' perspectives. Health Informatics J 2016;22(2):265-275 [FREE Full text] [doi: 10.1177/1460458214555041] [Medline: 25385165]

XSL•FO
RenderX

36. Owolabi BS, Odugbemi TO, Odeyemi KA, Onigbogi OO. mHealth: knowledge and use among doctors and nurses in public secondary health-care facilities of Lagos, Nigeria. J Clin Sci 2018;15(1):27. [doi: 10.4103/jcls.jcls_41_17]
37. Uko-Aviomoh E, Okoh E, Omatseye BOJ. Universal basic education (UBE) in Nigeria: an appraisal. Education 2007;127(4):558-569 [FREE Full text]
38. Kruse C, Betancourt J, Ortiz S, Valdes Luna SM, Bamrah IK, Segovia N. Barriers to the use of mobile health in improving health outcomes in developing countries: systematic review. J Med Internet Res 2019;21(10):e13263 [FREE Full text] [doi: 10.2196/13263] [Medline: 31593543]
39. Luszczynska A, Cieslak R. Mediated effects of social support for healthy nutrition: fruit and vegetable intake across 8 months after myocardial infarction. Behav Med 2009;35(1):30-38. [doi: 10.3200/BMED.35.1.30-38] [Medline: 19297302]
40. Skarin F, Olsson LE, Friman M, Wästlund E. Importance of motives, self-efficacy, social support and satisfaction with travel for behavior change during travel intervention programs. Transp Res F: Traffic Psychol Behav 2019;62:451-458. [doi: 10.1016/j.trf.2019.02.002]
41. Hao J, Gao B. Advantages and disadvantages for nurses of using social media. Journal of Primary Health Care and General Practice 2017;1(1):2 [FREE Full text]

Abbreviations

FGD: focus group discussions
mHealth: mobile health
mhGAP: Mental Health Global Action Programme
RAPiD: Responding to the challenge of Adolescent Perinatal Depression
SCT: Social Cognitive Theory
UCD: user-centered design
WHO: World Health Organization

Edited by A Mavragani; submitted 06.09.22; peer-reviewed by K Atmore; comments to author 01.11.23; revised version received 16.11.23; accepted 20.11.23; published 09.04.25.

Please cite as:
Kola L, Fatodu T, Kola M, Olayemi BA, Adefolarin AO, Dania S, Kumar M, Ben-Zeev D
Factors Impacting Mobile Health Adoption for Depression Care and Support by Adolescent Mothers in Nigeria: Preliminary Focus Group Study
JMIR Form Res 2025;9:e42406
URL: https://formative.jmir.org/2025/1/e42406
doi:10.2196/42406
PMID:

©Lola Kola, Tobi Fatodu, Manasseh Kola, Bisola A Olayemi, Adeyinka O Adefolarin, Simpa Dania, Manasi Kumar, Dror Ben-Zeev. Originally published in JMIR Formative Research (https://formative.jmir.org), 09.04.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

An Internet-Based and Mobile Family Management Intervention for Mothers of Very Preterm Infants Hospitalized in the Neonatal Intensive Care Unit (the Preemie Progress Program): Pilot Randomized Controlled Trial

Ashley Weber[1], RNC-NIC, RN, PhD; Tamilyn Bakas[1], RN, PhD; Qutaibah Oudat[1], RN, MPH, PhD; Nehal A Parikh[2,3], MS, DO; Joshua Lambert[1], PhD; Heather L Tubbs-Cooley[4], RN, PhD; Jared Rice[1], RRT-NPS, NNP-BC, MSN, MBA, DHSc; Kristin Voos[5], MD; Matthew Rota[6], PhD; Heather C Kaplan[3,7,8], MSCE, MD

[1]Department of Population Health, College of Nursing, University of Cincinnati, Cincinnati, OH, United States
[2]Neurodevelopmental Disorders Prevention Center, Perinatal Institute, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, United States
[3]Department of Pediatrics, College of Medicine, University of Cincinnati, Cincinnati, OH, United States
[4]Pitzer Center for Women, Children, and Youth, College of Nursing, The Ohio State University, Columbus, OH, United States
[5]Department of Pediatrics, School of Medicine, Case Western Reserve University, Cleveland, OH, United States
[6]Office of Learning Design and Digital Innovation, College of Nursing, University of Cincinnati, Cincinnati, OH, United States
[7]Perinatal Institute, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, United States
[8]James M. Anderson Center for Health Systems Excellence, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, United States

Corresponding Author:
Ashley Weber, RNC-NIC, RN, PhD
Department of Population Health
College of Nursing
University of Cincinnati
310 Procter Hall
3110 Vine St
Cincinnati, OH, 45221
United States
Phone: 1 5135580132
Email: weber2ay@ucmail.uc.edu

Abstract

Background: Flexible approaches to parenting training interventions in the neonatal intensive care unit (NICU), including family integrated care (FICare) models, are urgently needed across the globe. Many FICare trials inadvertently exclude parents with low resources who cannot commit to daily infant care (eg, 4-8 hours/day). Preemie Progress (PP) is a fully automated, video-based training program that allows parents to choose when and where they learn, without requiring parent bedside presence.

Objective: This study aims to examine the feasibility of recruitment, retention, fidelity, and changes in outcomes during a pilot randomized controlled trial of PP, a video-based intervention aimed at training mothers of very preterm infants in evidence-based family management skills in the NICU.

Methods: Mothers of infants born between 25 weeks and 0 days to 31 weeks and 6 days of gestation were enrolled in an NICU in the Midwestern United States. Electronic surveys were sent to collect maternal outcomes (Patient-Reported Outcomes Measurement Information System [PROMIS] 8a depression and anxiety scales) at baseline (T1), 14 days (T2) and 28 days (T3) after T1, and 30 days after NICU discharge (T4). Infant electronic health records were extracted to collect infant (ie, weight gain velocity at 36 weeks and receipt of mother's milk) and health care outcomes (ie, NICU length of stay as well as readmissions and emergency department visits within 30 days of discharge).

Results: Of 123 eligible mothers, 64 (52%) were randomly assigned to 1 of 2 arms (PP: n=33, 52%; attention control [AC]: n=31, 48%). Loss to follow-up was 30% (10/33) in the PP arm and 13% (4/31) in the AC arm. PP mothers watched a mean 17.8 (SD 18.9) of 49 videos. PP retention was linked to higher fidelity. PP mothers showed trends toward greater reductions in anxiety 30 days after discharge (mean −7.54, SD 1.93; 95% CI −11.32 to −3.76) compared to AC mothers (mean −4.67, SD 1.59; 95%

XSL•FO
RenderX

CI −7.80 to −1.55). PP infants trended toward greater receipt of exclusively mother's milk 28 days after baseline (PP: 14/26, 54%; AC: 10/28, 36%) and decreased NICU stay (PP: 57.2 days; AC: 68.3 days) but higher readmissions (PP: 4/33, 12%; AC: 2/31, 6%).

Conclusions: We were able to recruit a diverse sample of mothers from a range of socioeconomic backgrounds, including mothers experiencing barriers to bedside presence. Recruitment goals were met. PP showed promising trends in improving maternal, infant, and health care outcomes. Additional studies are needed to optimize PP and study procedures to improve retention and fidelity. PP has the potential to support parent training outside of traditional FICare models or serve as a complement to structure the parent education pillar of adapted FICare models.

Trial Registration: ClinicalTrials.gov NCT04638127; https://www.clinicaltrials.gov/study/NCT04638127

(JMIR Form Res 2025;9:e66073) doi:10.2196/66073

KEYWORDS

neonatal intensive care units; preterm infant; family nursing; parents; self-management; family integrated care

Introduction

Background

Increasing numbers of very preterm infants (born at less than 32 weeks of gestation) survive and have chronic, complex health care needs because of prematurity [1]. These infants require long neonatal intensive care unit (NICU) stays and have high health care use, with approximately 50% being readmitted within 90 days of discharge [2-4]. The NICU stay is a unique opportunity for nurses to train families to effectively manage care at the beginning of their infant's illness trajectory [5]. The duration of hospitalization for these infants is often months, providing parents with extended time to learn about their infant's care. Nevertheless, parents in the United States and globally typically receive limited structured training to effectively understand, monitor, and manage complex infant care within the chaotic NICU environment [6]. As a result, parents become overwhelmed, disengage early from infant care during NICU hospitalization, and develop depressive and anxious symptoms [7,8].

Flexible approaches to training parents in managing complex infant care in the NICU are urgently needed worldwide. Family integrated care (FICare) is a promising model of care in which parents are integrated into the care team as equal members and are trained by nurses to deliver the majority of infant care in the NICU. FICare has been shown to reduce maternal anxiety and depression [9], improve parenting self-efficacy in the NICU [6], increase infant weight gain and receipt of mother's milk [10], and decrease NICU length of stay [11]. Various FICare programs have been developed and tested in countries around the world, including Canada, Australia, and New Zealand [10]; China [11]; countries across western Europe [12-14]; and, most recently, the United States [6,15]. Traditional FICare models use an intensive apprenticeship type model, requiring parents to commit to 6 to 8 hours of daily infant care with their NICU nurses, weekly in-person parent education, and presence during medical rounds [14]. In the United States, paid parental leave is not universal, and barriers such as lack of transportation, childcare, and work release make traditional models of FICare untenable for many families. Researchers have recognized that traditional FICare models rely on resource-intensive practices and are designed within specific sociocultural contexts, namely, high-income countries with universal health care and extensive paid parental leave [16]. Thus, adaptations are likely needed for successful implementation in other countries, such as the United States [6]. Very few FICare studies have been conducted in the United States [6,15], and these studies highlight the importance of providing more adaptable approaches toward the FICare pillar of parent education. In sum, adaptations and additional frameworks are needed to structure parent training outside of or within adapted FICare programs to ensure that all parents gain the complex knowledge, skills, and confidence required to care for their infants in the NICU.

Prior Work

With a clinical academic partnership and deep engagement from key stakeholders [17], we developed an innovative intervention called Preemie Progress (PP) to address the critical need for a feasible parent training program in the United States. PP is a video-based program that teaches evidence-based family management skills informed by the Self- and Family Management Framework (ie, focusing on infant illness needs, activating resources, and living with infant illness) that have been identified by parents of hospitalized preterm infants as being essential in helping them manage care in the NICU [18]. The PP program allows parents to choose when and where they learn, without requiring extensive presence in the NICU. By moving beyond basic information and tasks (eg, diaper changing and feeding) and instead building the parent's capacity to self-manage the infant's condition, PP is designed to better prepare parents to execute chronic, complex care in the NICU and at home. Finally, PP was flexibly designed so that this promising intervention could be used as a stand-alone parent training program or as a complement to structure the parent education pillar of adapted FICare programs.

Goal of the Study

As the initial weeks after birth of a very preterm infant are stressful and chaotic for parents in the NICU, our team needed to understand the feasibility of our intervention and study procedures before a future definitive trial could commence with testing the efficacy of PP.

Objectives

The aim of our pilot randomized controlled trial (RCT), which has been registered with ClinicalTrials.gov (NCT04638127), was to examine the feasibility of the PP intervention and study

XSL•FO
RenderX

procedures within four domains: (1) recruitment (goal: recruit >50% of eligible mothers), (2) retention (goal: achieve <20% loss to follow-up in each trial arm), (3) fidelity to the PP intervention (examine the number and percentage of videos watched by PP mothers), and (4) potential for clinically meaningful changes in our trial outcomes over time (ie, sensitivity to change) throughout the NICU and postdischarge periods.

Methods

Design

This pilot study was a prospective, 2-parallel-arm randomized clinical trial with a 1:1 allocation ratio. Mothers of hospitalized preterm infants were randomly assigned to the PP intervention or an attention control (AC) condition within the first 5 weeks of infant life. Mothers, data collectors, and outcome assessors were blinded to trial arm, and data were collected longitudinally across multiple time points. The study was conducted according to the CONSORT (Consolidated Standards of Reporting Trials) guidelines for pilot and feasibility studies (Multimedia Appendix 1) and the TIDieR (Template for Intervention Description and Replication) checklist for interventions (Multimedia Appendix 2 [17-19]). The results are presented according to the CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) checklist (Multimedia Appendix 3) [19].

Conceptual Model

Figure 1 presents our conceptual model for the study. We believed that the PP intervention would decrease maternal depression and anxiety, increase infant weight gain and receipt of mother's milk, and decrease neonatal health care use, including NICU length of stay as well as readmissions and emergency department visits within 30 days of discharge. We theorized that this would occur through improvements in maternal parenting self-efficacy, NICU-related stress, postpartum bonding, and maternal ability to manage the infant's condition and family life difficulties surrounding a NICU hospitalization (Figure 1). As the focus of this pilot trial was feasibility, formal mediation analyses were not conducted because this pilot trial was not powered or intended for such analyses.

Figure 1. Theoretical model of the Preemie Progress program and its potential impact on outcomes. ED: emergency department; NICU: neonatal intensive care unit.

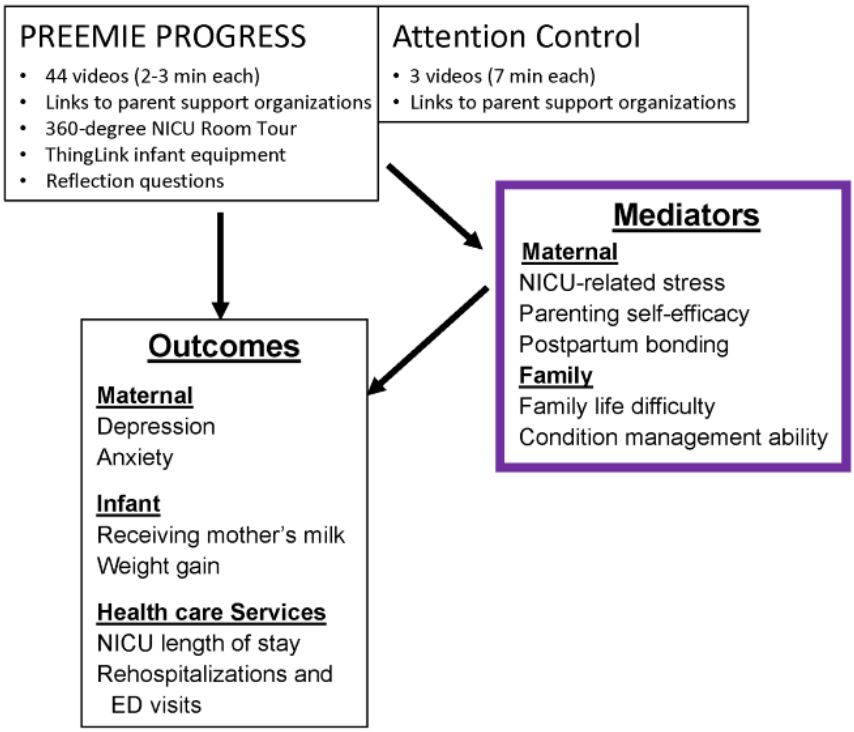

Inclusion and Exclusion Criteria

Mothers were eligible for enrollment if they were English speaking, aged >18 years, and gave birth to infants between 25 weeks and 0 days to 31 weeks and 6 days of gestational age (GA). This GA range provided adequate variability to meet enrollment goals within the study time frame and minimized attrition due to death from extreme prematurity [1]. Due to the pragmatic nature of this study, we included mothers of infants with comorbid conditions (eg, neonatal abstinence syndrome, congenital anomalies, brain injury, and surgical conditions). Mothers were excluded if they did not have custody of their infant, or their infant had imminent or probable risk of death based on the health care team's judgment.

Recruitment

Study recruitment and enrollment were conducted during initial infant hospitalization in a level IV NICU in the Midwestern United States. This NICU housed 82 beds and admitted approximately 500 preterm infants per year. All infant rooms are individual or semiprivate family rooms, equipped with free Wi-Fi and unit iPads for education.

XSL•FO
RenderX

Each weekday, study staff screened the NICU electronic health record (EHR) for eligible mothers. Study staff approached eligible mothers during the first 5 weeks of infant life for participation. Mothers were approached either in person or by telephone if in-person contact was not possible after 7 days. After explaining all aspects of the study, study staff sent interested mothers an electronic consent form through REDCap (Research Electronic Data Capture; Vanderbilt University) software.

Ethical Considerations

Written informed consent was obtained electronically through REDCap from all mothers before participation in this minimal-risk study. This study received ethics approval from the University of Cincinnati Institutional Review Board (2019-0475; Federalwide Assurance: 00003152), and all procedures were conducted in accordance with the Declaration of Helsinki ethical principles and Good Clinical Practice guidelines. Participant privacy and confidentiality were rigorously maintained throughout the study with strict adherence to a data safety and monitoring plan. The principal investigator and her research team held data safety and monitoring meetings quarterly to review protocol deviations, adverse events, study recruitment and retention, data management procedures designed to protect the privacy and confidentiality of participants, relevant literature and conference information, and protocol amendments. All data were deidentified and stored on secure, password-protected servers with access restricted to authorized study personnel. For participation in this study, mothers received "milestone cards" to celebrate steps made toward coming home at the first baseline visit. The cards were valued at approximately US $25 each. Each mother also received a US $25 electronic Amazon gift card for each completed visit after baseline (T1; 14 days [T2] and 28 days [T3] after T1, and 30 days after NICU discharge [T4]) for a total of US $75. No identifiable information is presented in this manuscript or associated supplementary materials.

Study Interventions and Trial Arms

Overview

Mothers were randomly assigned to either the AC arm or the PP arm. Both trial arms viewed their arm's assigned videos within the trial's website on their personal devices (ie, mobile phones, computers, or tablets). Both trial arms also received the same number and length of data collection assessments.

AC Arm

Mothers allocated to the AC arm received usual care, in which NICU nurses provide verbal education when available at the bedside. To maintain their attention, AC mothers were encouraged to view the welcome videos on the study website that covered topics such as hand hygiene, visitor badges, parking, and the NICU's unit design.

PP Arm

Mothers allocated to the PP arm received usual care and additionally watched PP videos on their personal devices. The PP multimedia content is hosted on the study website and includes 49 videos (each 2-4 minutes in length), review questions, a 360-degree interaction-enabled NICU room tour, and an interaction-enabled ThingLink infant anatomy interactive model that explains equipment commonly used with NICU infants. By accessing short content on their mobile devices, mothers were able to choose when, where, how much, how often, and how quickly they engaged with the material, personalizing the learning experience. The development process for the PP intervention has been previously described [17], and details of the PP program are provided in Multimedia Appendix 4.

Fidelity to PP and AC Arms

We applied the National Institutes of Health Behavioral Change Consortium's framework to address the 5 domains of treatment fidelity for both the AC and PP arms [20,21]. Study staff provided a standardized introduction on video access to all parents after baseline visits (training). Mothers within each group viewed standardized video content (design), which was hosted on the university's secure website portal with advanced analytics capabilities that enabled tracking each mother's fidelity to their assigned trial arm, defined as the percentage of videos watched over those assigned to the mother (delivery, receipt, and enactment). All mothers received 2 fidelity visits by telephone (approximately 15 minutes each) to answer questions, troubleshoot technology issues, and obtain feedback on the program. We did not set a specific fidelity goal a priori because we wanted to explore the number and percentage of PP videos that were feasible for mothers to watch.

Instruments With Validity and Reliability

We collected data at T1, T2, T3, and T4. The time points allowed us to assess immediate (T2) and long-term (T3 and T4) sensitivity to changes in outcomes related to intervention implementation. Table 1 lists the outcomes, mediators, measures, and data collection time points. Outcomes were chosen based on demonstrated relationships with family involvement in preterm infant care [22], importance to key stakeholders, and best practices for assessing family involvement with infant care in the NICU [23]. Infant outcomes (ie, weight gain at 36 weeks corrected GA and receipt of mother's milk) and health care use outcomes (NICU length of stay as well as readmissions and emergency department visits within 30 days of discharge) were collected directly from the infant's EHR and entered into the trial's REDCap database.

Maternal outcomes were collected through electronic REDCap surveys sent to mothers at baseline and then 4 days before each postbaseline assessment (ie, T2, T3, and T4) for completion at their leisure. Standardized survey measures were used to assess maternal outcomes, that is, self-reported depressive (Patient-Reported Outcomes Measurement Information System [PROMIS] 8a depression scale) and anxious (PROMIS 8a anxiety scale) symptoms, along with the mediators of condition management ability and family life difficulty (Family Management Measure), NICU-related stress (Parental Stressor Scale: NICU), parenting self-efficacy (Perceived Maternal Parenting Self-Efficacy Scale), and maternal-infant bonding (Postpartum Bonding Questionnaire). These measures have been thoroughly validated, shown to be reliable, and tested with mothers of preterm infants (Table 1).

XSL•FO
RenderX

Table 1. Primary outcomes, mediators, measures, and data collection at each study time point in the Preemie Progress trial.

Variables and domains and measures	T1[a]	T2[b]	T3[c]	T4[d]
Primary outcomes				
Maternal				
Depression: PROMIS[e] 8a T score [24], 8 items; T score standardized to mean 50 (SD 10); higher score indicates greater depressive symptoms; content validity, precision, and reliability established	✓	✓	✓	✓
Anxiety: PROMIS 8a T score [24], 8 items; T score standardized to mean 50 (SD 10); higher score indicates greater anxious symptoms	✓	✓	✓	✓
Infant				
Receipt of mother's milk (exclusive, partial, or none)	✓	✓	✓	
Weight gain at 36 weeks corrected gestational age; *z* score method [10,25,26]			✓	
Health care use				
NICU[f] length of stay (days) [27,28]				✓
Hospital readmissions and ED[g] visits within 30 days of discharge [4,29]				✓
Mediators				
Maternal				
Self-efficacy: Perceived Maternal Parenting Self-Efficacy Scale [30], 20 items (80-point maximum, with higher scores indicating greater self-efficacy); Cronbach α=0.91 and test-retest reliability=0.96; content, convergent, and discriminant validity established	✓	✓	✓	
NICU-related stress: Parental Stressor Scale: NICU [31], 50 items; 5-point Likert scale ranging from "not at all stressful" (0) to "extremely stressful" (5), with higher scores indicating more stress; Cronbach α=0.87 and test-retest reliability=0.60; content, convergent, and discriminant validity established	✓	✓	✓	
Maternal-infant bonding: Postpartum Bonding Questionnaire [32], 25 items (125-point maximum, higher scores indicating more impaired bonding); Cronbach α=0.76; content validity established	✓	✓	✓	
Family management				
Family Management Measure [33] condition management ability scale, 12 items; higher scores indicate greater condition management ability; internal consistency reliability=0.72 and test-retest reliability=0.79; content, convergent, and discriminant validity established	✓	✓	✓	✓
Family Management Measure [33] family life difficulty scale, 14 items; higher scores indicate greater family life difficulty; internal consistency reliability=0.90 and test-retest reliability=0.94; content, convergent, and discriminant validity established	✓	✓	✓	✓

[a]T1: baseline.

[b]T2: 14 days after T1.

[c]T3: 28 days after T1.

[d]T4: 30 days after discharge.

[e]PROMIS: Patient-Reported Outcomes Measurement Information System.

[f]NICU: neonatal intensive care unit.

[g]ED: emergency department.

Data Collection

After mothers signed the electronic consent form, study staff sent them baseline surveys electronically through REDCap. Mothers without an electronic device were provided with a study tablet to complete surveys. After the baseline surveys were completed, study staff met with mothers in person (or by telephone if unavailable in person) to help them access video content.

Randomization

Mothers were randomly assigned to 1 of the 2 arms by study staff after baseline data collection to minimize dropout effects on randomization. Using R software (R Foundation for Statistical Computing), the statistician (JL) generated a random allocation sequence stratified by GA at birth (GA <28 weeks and >28 weeks) to balance the representation of the youngest and sickest infants. We used permuted block randomization for each stratum, with block sizes of 2, 4, and 6 and a 1:1 treatment allocation. The statistician (JL) uploaded the randomization sequence into REDCap, had no contact with participants, and was independent of all assessment procedures.

Blinding

REDCap's secure randomization feature prevented the study team from viewing the treatment allocation sequence, ensuring

XSL•FO
RenderX

allocation concealment. We blinded mothers to trial arm. After randomization, study staff provided mothers with a registration code that would load their assigned video content (ie, AC or PP content) into the study website. Study staff who assessed trial outcomes were blinded through the use of REDCap's restricted user rights and did not participate in fidelity visits.

Data Analysis

Power

We aimed to enroll at least 60 mothers (n=30, 50% per arm), with planned overrecruitment to account for up to 20% loss to follow-up (ie, 72mothers) during the pilot trial. As we did not have reliable estimates for the CIs, SEs, or mean differences for our outcomes a priori, our sample size was based on the recommendations of Whitehead et al [34], who suggest enrolling at least 20 persons per treatment arm in a pilot trial when expecting a small standardized difference (δ=0.2) to achieve at least 80% power in the main trial. Our recruitment and retention goals, set a priori, were to achieve a recruitment rate of at least 50% among eligible mothers and a loss to follow-up rate of <20% for each arm. As our emphasis was on feasibility, this pilot trial was not powered to detect statistically significant changes in outcome or mediating variables between arms.

Analysis Goals and Plan

The primary emphasis of this pilot RCT was to explore data trends surrounding the feasibility domains of (1) recruitment, (2) retention, (3) fidelity to the PP intervention, and (4) sensitivity to changes in our trial outcomes over time, rather than test for statistical significance. Maternal surveys, EHR data, and website analytic data were entered and stored in REDCap and then exported to Stata 17 (StataCorp LLC) software. Descriptive and summary statistics were calculated for all data. Descriptive and summary statistics were used to assess demographic characteristics within each arm, baseline equivalence between arms, and data surrounding the feasibility domains of recruitment (aim to recruit >50% of eligible mothers), retention (achieve <20% loss to follow-up in each arm), fidelity (number and percentage of PP videos watched), and changes in outcomes over time. To understand the impact and potential mechanisms of retention and attrition in the trial, we conducted a post hoc power and statistical analysis to examine whether mothers and their infants differed on demographic characteristics, outcome measures, and fidelity to trial arm. Spaghetti plots with 95% CIs were also used to explore outcome and mediator trends by trial arm. We computed means and 95% CIs for changes from baseline (from T1 to T2-T4) for maternal-reported outcomes by arm. At each follow-up (T2-T4), we also computed mean differences and 95% CIs between arms to evaluate trends in outcomes. To explore trends in NICU length of stay between arms, we computed the mean number of hospitalized days, performed survival analyses, and plotted Kaplan-Meier survival curves for each trial arm.

Results

Participant Flow and CONSORT Diagram

A total of 140 mothers were screened over a period of 14 months (from May 17, 2022, to July 27, 2023), of whom 123 (87.9%) met the eligibility criteria (Figure 2). Of these 123 eligible mothers, 20 (16.3%) declined to participate, with the most frequently cited reasons for declining being not having time and being stressed and overwhelmed. Importantly, study staff were not able to make any contact (eg, telephone call, SMS text message, or in person) with 17 (13.8%) of the 123 eligible mothers, and, after successful initial contact, could not make subsequent contact for a final decision with 17 (13.8%) mothers. These mothers spent limited time in the NICU and did not answer recruitment telephone calls or SMS text messages. Ultimately, of the 123 eligible mothers, 64 (52%) were randomly assigned to the PP arm or the AC arm during the 14-month recruitment period.

Of the 64 mothers, 31 (48%) were allocated to the AC arm. Of these 31 mothers, 4 (13%) were lost to follow-up; and for 1 (3%), the infant was still hospitalized at the end of the study period. Thus, 16% (5/31) of the data for maternal-reported outcomes in the AC arm were missing at T4, and the analyzable AC sample for maternal-reported outcomes at T4 comprised 26 (84%) of the 31 mothers (Figure 2). Of the 33 mothers allocated to the PP arm, 10 (30%) were lost to follow-up; and for 2 (6%), the infants were still hospitalized at the end of the study period; moreover, 3 (9%) withdrew; 1 (3%) was withdrawn by the principal investigator due to severe health issues requiring multiple hospitalizations; and for 1 (3%), the infant died. Thus, "52% (17/33) of the data for maternal-reported outcomes in the PP arm were missing at T4," and the analyzable PP sample for maternal-reported outcomes at T4 was 16 (48%) of the 33 mothers (Figure 2). Interestingly, mothers who did not complete the study in both arms did not differ on any demographic characteristics or on baseline outcome measures from those who completed the study (Table 2).

XSL•FO
RenderX

Figure 2. CONSORT (Consolidated Standards of Reporting Trials) diagram for the Preemie Progress pilot randomized controlled trial. PI: principal investigator; T2: 14 days after baseline; T3: 28 days after baseline; T4: 30 days after discharge from the neonatal intensive care unit. Note: mother-infant dyads can be ineligible for more than one reason.

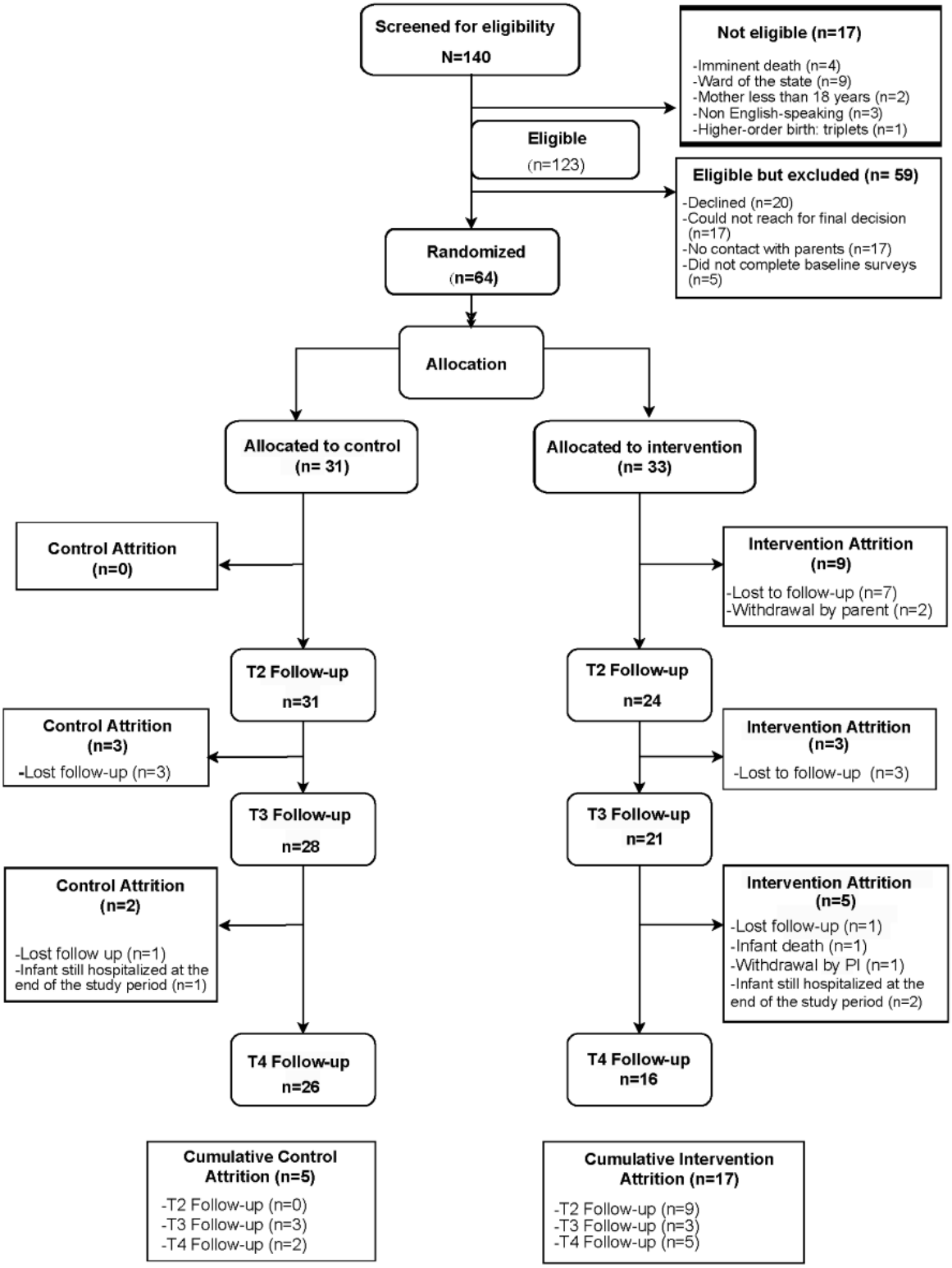

XSL•FO
RenderX

Table 2. Maternal and infant demographics, baseline outcomes, primary outcomes, and fidelity by study attrition status in the Preemie Progress trial (n=61)[a].

Variables	Did not complete study (n=19)	Completed study (n=42)	*P* value	Chi-square (*df*) or Mann *U*
Maternal demographics				
Age (y), mean (SD)	27.95 (6.78)	30.52 (5.18)	.17	3.5 (2)
Education (y), n (%); mean (SD)	17 (89); 13.82 (2.40)	41 (98); 15.24 (2.69)	.16	3.6 (2)
Income (in US $10,000), n (%); mean (SD)	18 (95); 3.22 (3.98)	41 (98); 5.10 (3.85)	.17	3.6 (2)
Social determinants of health	17 (89)	41 (98)		
No social determinants of health needs, n (%)	10 (59)	32 (78)	.18	3.4 (2)
Gravida, n (%); mean (SD)	19 (100); 2.79 (1.72)	41 (98); 2.66 (1.85)	.83	0.38 (2)
Para, mean (SD)	2.00 (1.25)	2.00 (1.23)	.90	0.21 (2)
Duke Social Support Index score, mean (SD)	18.16 (5.70)	20.60 (5.06)	.20	3.25 (2)
Marital status: married, n (%)	6 (32)	17 (41)	.82	4.36 (8)
Race: White, n (%)	8 (42)	26 (62)	.51	9.3 (10)
Employment status, n (%)				
Employed	17 (89)	41 (98)	.85	0.3 (2)
Not employed in past 12 months	6 (35)	13 (32)	.85	0.3 (2)
Infant demographics				
Gestational age at birth (weeks), mean (SD)	29.56 (1.82)	29.27 (1.94)	.10	4.5 (2)
Birth weight (g), mean (SD)	1298.84 (398.67)	1292.69 (345.52)	.10	4.7 (2)
Apgar score (1 min), n (%); mean (SD)	19 (100); 4.74 (2.70)	41 (98); 5.17 (2.62)	.54	1.25 (2)
Apgar score (5 min), n (%); mean (SD)	19 (100); 6.95 (1.99)	41 (98); 7.76 (1.26)	.24	2.8 (2)
Day of life at enrollment, mean (SD)	14.37 (9.23)	10.36 (5.91)	.12	4.3 (2)
Corrected gestational age at enrollment (weeks), mean (SD)	31.61 (1.92)	30.76 (2.03)	.12	1.6 (2)
Maternal outcomes at baseline				
PROMIS[b] 8a depression scale, mean (SD)	48.64 (8.46)	48.60 (8.23)	.86	0.29 (2)
PROMIS 8a anxiety scale, mean (SD)	54.12 (7.58)	55.76 (10.58)	.66	0.84 (2)
Condition (ie, prematurity) management ability, n (%); mean (SD)	18 (95); 47.44 (6.57)	41 (98); 46.61 (6.07)	.85	0.3 (2)
Family life difficulty, n (%); mean (SD)	18 (95); 27.72 (9.40)	41 (98); 30.95 (11.04)	.73	0.6 (2)
NICU[c]-related stress: Parental Stressor Scale: NICU, mean (SD)	1.54 (1.04)	1.82 (0.81)	.27	2.6 (2)
Perceived Maternal Parenting Self-Efficacy Scale, mean (SD)	59.68 (19.91)	52.69 (18.82)	.27	2.59 (2)
Postpartum Bonding Questionnaire, mean (SD)	4.95 (2.99)	6.02 (7.01)	.76	0.56 (2)
Infant health care use				
NICU length of stay (median time: 50% [days]), median (IQR)	46 (40-80)	49 (44-77)	.44	–0.8
Hospital readmission or ED[d] visit within 30 days	4 (21)[e]	2 (5)[e]	*.001*[f]	52.2 (2)
z score change in weight from birth to 36 weeks corrected gestational age, mean (SD)	−0.94 (0.42)	−0.80 (0.69)	.0.3	–1.04
Receipt of mother's milk (baseline)	19 (100)	41 (98)	.48	4.17 (4)
Infants receiving exclusively mother's milk, n (%)	8 (42)	24 (59)	.48	4.17 (4)
Fidelity to PP intervention				
Number of videos watched (PP group only), n (%); mean (SD); median (IQR)	15 (79); 4.13 (5.11); 2 (0-8)	17 (40); 28.88 (18.54); 32 (13-49)	*.001*	14.04 (2)

[a]For 3 (5%) of the 64 mothers, their infants were still hospitalized at the end of the study period; thus, they did not have the opportunity to complete or

XSL•FO
RenderX

not complete the study. These analyses do not include these mother-infant dyads. Chi-square analyses were used to examine differences between categorical variables (eg, maternal race, marital status, and employment status, as well as infant receipt of exclusive mother's milk) based on study attrition status. Mann Whitney-*U* nonparametric rank sum tests were used to examine differences between continuous variables based on study attrition status because these variables did not have a normal distribution.

[b]PROMIS: Patient-Reported Outcomes Measurement Information System.

[c]NICU: neonatal intensive care unit.

[d]ED: emergency department.

[e]All 4 mothers who did not complete the study and whose infants were readmitted were in the Preemie Progress (PP) arm of the trial. The 2 mothers who completed the study and had infants who were readmitted were all in the attention control arm of the trial.

[f]Italicization indicates significant *P* values.

Characteristics of the Sample

Tables 3 and 4 describe baseline demographic and clinical characteristics by trial arm for enrolled mothers and their infants, respectively. As a whole, the mothers (n=64) had a mean age of 29.8 (SD 5.7) years, a mean of 14.8 (SD 2.7) years of education, and a mean income of US $45,400 (SD US $3,900). Nearly 31% (20/64) of the mothers identified at least 1 concern related to the social determinants of health. Of the mothers, 24 (38%) were married. The racial and ethnic composition of our sample reflected the local population, with 36 (56%) of the 64 mothers identifying as White, 23 (36%) identifying as Black or African American, and 62 (97%) identifying as non-Hispanic. Infants were born at a mean of 29.2 (SD 1.9) weeks of gestation with a mean birth weight of 1269 (SD 374) g. Infants were enrolled on average at 11.8 (SD 7.2) days of life, with a corrected GA of 30.9 (2.0) weeks at enrollment. Maternal and infant characteristics were similar between trial arms.

XSL•FO
RenderX

Table 3. Demographic, outcome, and mediator measures at baseline for mothers by trial arm in the Preemie Progress trial (n=64)[a].

Variables	Attention control (n=31)	Preemie Progress (n=33)	*P* value	*t* score (*df*) or Mann *U* or chi-square (*df*)
Demographic characteristics				
Maternal age (y), mean (SD; range)	30.5 (4.3; 20-41)	29.1 (6.7; 18-46)	.31	1.03 (62)
Income (in US $10,000), n (%); mean (SD; range)	30 (97); 44,000 (39,000; 0-100,000)	31 (91); 46,700 (39,000; 0-100,000)	.78	−0.28 (59)
Social determinants of health, n (%)	31(100)	33(100)	.36	0.83
No social determinants of health needs, n (%)	23 (74)	21 (63)	.36	0.83 (1)
Years of education, n (%); mean (SD; range)	30 (97); 15 (2.6; 11-19)	30 (91); 14.6 (2.8; 11-21)	.46	0.74
Gravida[b], n (%); mean (SD; range)	31 (100); 2.8 (1.9; 1-7)	32 (97); 2.6 (1.7; 1-9)	.82	0.23
Para, mean (SD; range)	2.0 (1.2; 1-5)	2.0 (1.3; 0-6)	.85	−0.18
Duke Social Support Index score, mean (SD; range)	20.7 (4.4; 11-30)	19.3 (5.9; 6-30)	.11	1.1 (62)
Marital status, n (%)			.92	1.55 (4)
Single, not partnered	9 (29)	9 (27)		
Married	11 (35)	13 (39)		
Divorced	0 (0)	1 (3)		
Partnered, not living together	4 (13)	5 (17)		
Partnered, living together	7 (23)	5 (17)		
Race, n (%)	30 (97)	33 (100)	.85	3.28 (5)
Asian	1 (3)	0 (0)		
Black or African American	10 (33)	13 (39)		
White	17 (57)	18 (56)		
Multiracial	2 (7)	2 (6)		
Ethnicity, n (%)	30 (97)	30 (91)	.48	1.12 (1)
Hispanic or Latinx	1 (3)	0 (0)		
Not Hispanic or Latinx	29 (97)	30 (100)		
Employment in past 12 months, n (%)	30 (97)	30 (91)	.41	1.20 (1)
Yes	22 (73)	18 (60)		
No	8 (27)	12 (40)		
Baseline maternal outcomes, mean (SD; range)				
PROMIS[c] 8a anxiety scale	57.3 (10.4; 37.1-76.9)	53.4 (8.8; 37.1-67.8)	.08	1.76
PROMIS 8a depression scale	49.3 (8.0; 38.2-60.3)	48.1 (8.6; 38.2-65.9)	.43	0.79
Baseline maternal mediators, n (%); mean (SD; range)				
Family life difficulty	30 (97); 30.1 (10.3; 15-53)	31 (94); 29.7 (10.9; 15-54)	.95	0.96
Condition management ability	30 (97) 46.5 (5.7; 33-60)	31 (94); 47.2 (6.7; 34-60)	.89	−0.15
Maternal parenting self-efficacy	31 (100); 51.5 (18.6; 13-80)	33 (100); 58.1 (19.8; 8-80)	.11	−1.61
Maternal postpartum bonding	31 (100); 5.3 (5.7; 0-23)	32 (97); 5.9 (6.4; 0-30)	.1	-1.66 (61)
NICU[d]-related stress	31 (100); 1.7 (0.8; 0.4-3.3)	33 (100); 1.7 (1.0; 0.2-4.3)	.68	0.42

[a]None of the maternal demographic characteristics, outcome measures, or mediator measures were statistically different between trial arms at baseline. Mothers were not required to self-report any questions they did not wish to answer. Thus, some data are missing.

[b]Gravida information is missing for 1 mother, whose gravida status was unknown in the electronic health record because she had a series of miscarriages.

[c]PROMIS: Patient-Reported Outcomes Measurement Information System.

XSL•FO
RenderX

[d]NICU: neonatal intensive care unit.

Table 4. Demographic characteristics of very preterm infants by trial arm (n=64)[a].

Variables	Attention control (n=31)	Preemie Progress (n=33)	*t* test (*df*) or Mann *U* or chi-square (*df*)	*P* value
Gestational age at birth (weeks), mean (SD; range)	29.3 (2.1; 25.4-31.9)	29.2 (1.8; 25-31.6)	0.3	.77
Birth weight (g), mean (SD; range)	1286.2 (359.2; 670-2130)	1253.9 (392.4; 480-2050)	0.3 (62)	.73
Apgar[b] score (1 min), n (%); mean (SD; range)	30 (97); 5.3 (2.5; 1-9)	33 (100); 4.6 (2.8; 0-9)	0.8	.40
Apgar[b] score (5 min), n (%); mean (SD; range)	30 (97); 7.7 (1.6; 2-9)	33 (100); 7.1 (1.8; 1-9)	1.6	.12
Day of life at enrollment, mean (SD; range)	10.2 (5.6; 2-23)	13.3 (8.1; 1-37)	−1.6	.11
Corrected gestational age at enrollment (weeks), mean (SD; range)	30.7 (2.3; 25.9-35.1)	31.1 (1.9; 26.9-34.2)	−0.8 (62)	.44
Sex: male, n (%)	23 (74)	17 (52)	3.5	.06
Race, n (%)			2.2	.75 (3)
Asian	0 (0)	1 (3)		
Asian, Black	0 (0)	1 (3)		
Black	12 (39)	14 (42)		
White	19 (61)	17 (52)		
Ethnicity, n (%)			1.1	.48 (1)
Not Hispanic or Latinx	1 (3)	0 (0)		
Hispanic or Latinx	30 (97)	33 (100)		
Infant and mother with positive toxicology screen: no, n (%)	31 (100)	33 (100)	1.0	.99 (1)
Infant required surgery, n (%)			1.0	.32 (1)
No	31 (100)	32 (97)		
Yes	0 (0)	1 (3)		
Grade III or IV IVH [c] or brain injury, n (%)			0.002	.99 (1)
No	30 (97)	32 (97)		
Yes	1 (3)	1 (3)		

[a]None of the infant demographic characteristics were statistically different between arms.

[b]Apgar score for 1 infant was not documented in the electronic health record.

[c]IVH: intraventricular hemorrhage.

Feasibility Domain: Recruitment and Retention

We met our recruitment goal of recruiting >50% of eligible mothers (64/123, 52%) during our 14-month recruitment period. However, we did not meet our retention goal in the PP arm, which experienced a 30% (10/33) loss to follow-up rate versus the anticipated 20%. This resulted in an analyzable sample size of 16 mothers for PP maternal-reported outcomes at T4. To understand the impact that attrition might have on initial estimates for our outcomes, we conducted a post hoc power analysis using the methods recommended by Whitehead et al [34], discussed earlier. Even when using the smallest (ie, the most conservative) standardized difference between trial arms at any time point for the primary outcomes of anxiety and depression (ie, T3 PROMIS 8a depression scale=0.29), we achieved a standardized difference much higher than expected (δ=1.06 calculated as [(0.29) / (2.18 / sqrt (64)]), indicating that a final sample size of 10 mothers per arm would have been sufficient for this trial and future trials.

Feasibility Domain: Fidelity to PP Intervention

Mothers in the PP arm showed large variation in their fidelity to watching PP videos. They watched a mean of 17.8 (SD 18.9; range 0-49) of 49 videos, with a median of 9 (IQR 2-39) videos, and they logged into the study website a mean of 6 (SD 7; range 0-30) times, with a median of 4 (IQR 1-11) times. They spent a mean of 95 (SD 104; range 0-310) minutes, with a median of 57 (IQR 5-165) minutes, on the website. PP mothers who did not complete the study watched significantly fewer videos (P=.0009; Table 2). PP mothers who completed the trial had higher PP fidelity (median 32, IQR 13-49, of 49 videos watched; Table 2).

XSL•FO
RenderX

Feasibility Domain: Changes in Outcome Variables by Trial Arm

Descriptive and summary statistics are presented in Table 5 for maternal, infant, and health care use outcomes. PP mothers showed trends toward greater reductions in anxiety (mean −7.54, SD 1.93; 95% CI −11.32 to −3.76) versus AC mothers (mean −4.67, SD 1.59; 95% CI −7.80 to −1.55) 30 days after NICU discharge (Table 6; Figure 3, upper right corner). Both arms showed significant decreases in depressive and anxious symptoms from baseline to 30 days after NICU discharge (Table 6; Figure 3, top row). PP infants showed greater receipt of exclusive mother's milk (14/26, 54%) versus AC infants (10/28, 36%) 28 days after baseline (Table 5). PP infants showed trends toward decreased NICU length of stay (30/34, 88%; 57.2 days; Table 5) versus AC infants (30/31, 96%; 68.3 days; Table 5). When examined using survival analysis (Figure 3), the median NICU length of stay was similar in both arms, but the PP arm began to show a trend toward decreased length of stay after 51 days (the median). In the face of lower length of stay, the PP arm had higher readmission rates (PP arm: 4/33, 12%; AC arm: 2/31, 6.5%; Table 5). The mothers (4/19, 21%) who did not complete the study and whose infants were readmitted were in the PP arm of the trial (Table 2). Of these 4 mothers, 2 (50%) did not watch any PP videos, and 1 (25%) watched only 2 PP videos. The AC mothers (2/19, 11%) with infants who were readmitted completed the study (Table 2).

XSL•FO
RenderX

Table 5. Descriptive and summary statistics for primary outcome measures by trial arm in the Preemie Progress trial (n=64)[a].

Domains, measures, and time periods	Attention control (n=31)	Preemie Progress (n=33)
Health care use outcomes		
NICU length of stay (days)[b], n (%); mean (SD; range)	30 (97); 68.3 (36.3; 26 to 171)	30 (91); 57.2 (23.4; 26 to 114)
Hospital readmission or ED visit within 30 days of discharge[c], n (%)	2 (6)	4 (12)
Infant outcomes		
z score Δ weight (g) from birth to 36 weeks CGA[d], mean (SD; range); median (IQR)	−0.67 (0.64; −1.6 to 1.8); −0.72 (−1.11 to −0.39)	−0.92 (0.65; −2.6 to 0.95); −0.80 (−1.36 to −0.51)
Receipt of mother's milk[e], n (%)		
T1		
Infants on enteral feeds	30 (97)	33 (100)
Infants receiving exclusively mother's milk	18 (51.5)	17 (48.6)
T2		
Infants on enteral feeds	29 (94)	31 (94)
Infants receiving exclusively mother's milk	15 (53.6)	13 (46.4)
T3		
Infants on enteral feeds	28 (90)	26 (79)
Infants receiving exclusively mother's milk	10 (41.6)	14 (58.3)
Maternal outcomes, n (%); mean (SD; range)		
PROMIS 8a anxiety scale		
T1	31 (100); 57.3 (10.4; 37.1 to 76.9)	33 (100); 53.4 (8.8; 37.1 to 67.8)
T2	30 (97); 55.2 (8.8; 37.1 to 69)	25 (76); 51.2 (9.4; 37.1 to 68.8)
T3	28 (90); 53.8 (9.1; 37.1 to 68.9)	22 (67); 51.5 (8.5 37.1 to 66.6)
T4	25 (81); 52.7 (9.5; 37.1 to 66.9)	16 (48); 47.2 (9.0; 37.1 to 61.5)
PROMIS 8a depression scale		
T1	31 (100); 49.3 (8.0; 38.2 to 60.3)	33 (100); 48.1 (8.6; 38.2 to 65.9)
T2	30 (97); 47.8 (7.6; 38.2 to 58.5)	26 (79); 46.5 (7.6; 38.2 to 67.6)
T3	28 (90); 46.7 (7.9; 38.2 to 62.0)	21 (64); 45.0 (8.3; 38.2 to 63.1)
T4	24 (77); 45.6 (7.6; 38.2 to 59.3)	16 (48); 43.3 (6.7; 38.2 to 56.2)

[a]Health care use outcomes (ie, neonatal intensive care unit [NICU] length of stay as well as readmissions and emergency department [ED] visits within 30 days of discharge) and infant outcomes (ie, weight gain velocity at 36 w and receipt of mother's milk) were extracted from the infant's electronic health record 30 days after infant discharge from the NICU. Maternal outcomes (Patient-Reported Outcomes Measurement Information System [PROMIS] 8a depression and anxiety scales) were reported by mothers through REDCap (Research Electronic Data Capture) electronic surveys, which were completed upon enrollment (baseline, T1), 14 days after T1 (T2), 28 days after T1 (T3), and 30 days after infant discharge from the NICU (T4). For continuous variables, we reported the mean, SD, and minimum and maximum. For highly skewed data, we included the median and IQR.

[b]We calculated the mean NICU length of stay in days. We excluded 1 infant in the attention control arm who was still hospitalized at the end of the study period. We excluded 3 infants in the Preemie Progress group: 2 (67%) were still hospitalized at the end of the study period, and 1 (33%) died. Survival analyses that include the censored length of stay for the 4 infants still hospitalized at the end of the study period are presented in Figure 4.

[c]Outcome was coded as "yes" if the infant was rehospitalized or had an ED visit. In this sample, all infants with an ED visit were rehospitalized.

[d]CGA: corrected gestational age.

[e]At each time point, some infants were on intravenous fluids only (ie, nothing by mouth) and were not included in the denominator for this measure.

XSL•FO
RenderX

Table 6. Estimated mean changes from baseline (T1) in mother-reported outcomes and mediators by trial arm[a].

Domains and measures	Attention control (n=31), mean change (SD; 95% CI)	Preemie Progress (n=33), mean change (SD; 95% CI)	Difference in mean change (SD; 95% CI)
PROMIS[b] 8a anxiety scale			
T2[c]–T1	−1.96 (1.52; −4.95 to 1.02)	−3.08 (1.74; −6.49 to 0.33)	−1.12 (2.47; −5.96 to 3.72)
T3[d]–T1	−3.91 (1.53; −6.90 to −0.91)	2.87 (1.77; −6.34 to 0.60)	1.04 (2.46; −3.78 to 5.86)
T4[e]–T1	−4.67 (1.59; −7.80 to −1.55)	−7.54 (1.93; −11.32 to −3.76)	−2.87 (2.61; −7.99 to 2.25)
PROMIS 8a depression scale			
T2−T1	−0.77 (1.34; −3.40 to 1.86)	−1.33 (1.51; −4.29 to 1.64)	−0.56 (2.16; −4.79 to 3.67)
T3−T1	−2.50 (1.35; −5.14 to 0.14)	−2.21 (1.58; −5.31 to 0.89)	0.29 (2.18; −3.99 to 4.57)
T4−T1	−4.08 (1.42; −6.86 to −1.29)	−3.75 (1.70; −7.07 to −0.42)	0.33 (2.31; −4.19 to 4.85)
Condition management ability			
T2−T1	0.59 (1.15; −1.66 to 2.83)	0.00 (1.38; −2.70 to 2.70)	−0.59 (1.89; −4.29 to 3.12)
T3−T1	3.22 (1.15; 0.96 to 5.47)	3.18 (1.42; 0.40 to 5.97)	−0.04 (1.91; −3.78 to 3.71)
T4−T1	4.73 (1.19; 2.40 to 7.06)	4.39 (1.48; 1.50 to 7.28)	−1.54 (1.94; −5.34 to 2.25)
Family life difficulty			
T2−T1	−0.18 (1.51; −3.14 to 2.78)	−1.12 (1.81; −4.68 to 2.43)	−0.94 (2.49; −5.82 to 3.94)
T3−T1	−1.07 (1.52; −4.05 to 1.90)	−3.32 (1.87; −7.00 to 0.35)	−2.25 (2.51; −7.18 to 2.68)
T4−T1	−6.08 (1.57; −9.15 to −3.01)	−7.90 (1.95; −11.72 to −4.08)	−1.82 (2.60; −6.92 to 3.28)
NICU[f]-related stress			
T2−T1	−0.13 (0.14; −0.41 to 0.15)	−0.06 (0.17; −0.39 to 0.27)	0.07 (0.24; −0.40 to 0.53)
T3−T1	−0.43 (0.15; −0.73 to −0.14)	−0.39 (0.18; −0.74 to −0.04)	0.05 (0.24; −0.42 to 0.53)
Parenting self-efficacy			
T2−T1	2.36 (2.80; −3.13 to 7.85)	4.30 (3.25; −2.08 to 10.68)	1.94 (4.59; −7.06 to 10.95)
T3−T1	7.76 (2.80; 2.27 to 13.26)	12.23 (3.37; 5.62 to 18.84)	4.46 (4.64; −4.63 to 13.56)
Postpartum bonding			
T2−T1	0.57 (0.94; −1.28 to 2.42)	1.20 (1.09; −0.94 to 3.34)	0.63 (1.54; −2.40 to 3.65)
T3−T1	1.07 (0.95; −0.80 to 2.93)	1.56 (1.13; −0.66 to 3.78)	0.49 (1.56; −2.57 to 3.56)

[a]Estimates were obtained while controlling for treatment dose (ie, the percentage of videos watched by the mothers).

[b]PROMIS: Patient-Reported Outcomes Measurement Information System.

[c]T2: 14 days after T1.

[d]T3: 28 days after T1.

[e]T4: 30 days after discharge.

[f]NICU: neonatal intensive care unit.

XSL•FO
RenderX

Figure 3. Note: Mothers in the Attention-Control (AC, blue) and Preemie Progress (PP, red) trial arms showed significant decreases from baseline in their PROMIS8a T-scores for depressive (A) and anxious (B) symptoms at 30 days post NICU discharge. Mothers in the PP arm showed a trend towards greater reductions in anxiety (Right). Mothers in both arms reported significant increases in condition (i.e.) management ability at 30 days after hospital discharge (C). Mothers in both arms reported significant decreases in family life difficulty 28 days after baseline and at 30 days post NICU discharge (D). Mothers in both arms reported statistically significant decreases in NICU-related stress (E) and statistically significant increases in parenting self-efficacy (Right) at 28 days post-baseline. Mothers in the PP arm (Red) showed trends towards greater improvements in parenting self-efficacy 28 days after baseline (F).

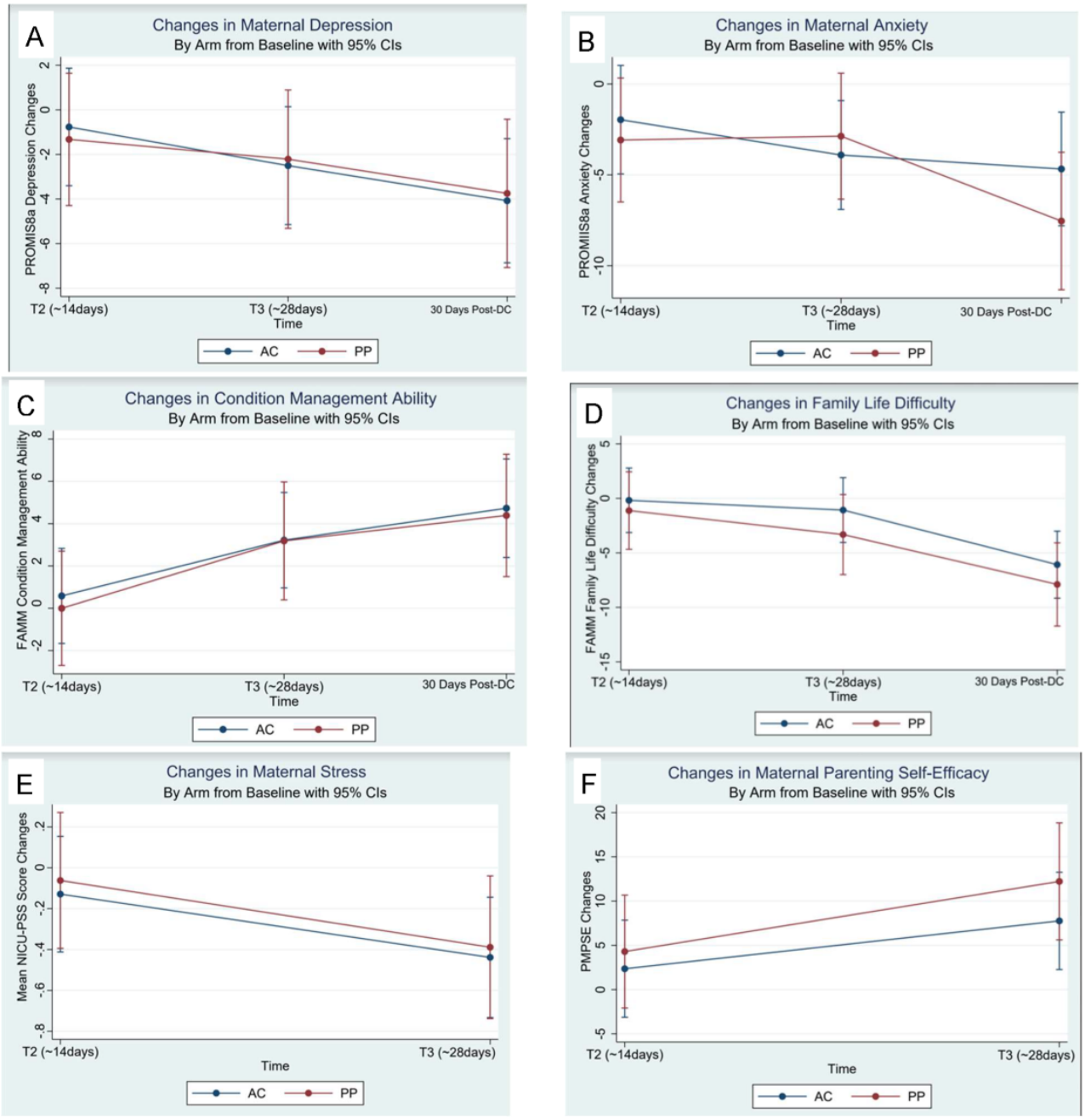

XSL•FO
RenderX

Figure 4. We excluded 1 infant who died in the Preemie Progress (PP) arm. The 4 infants who were still hospitalized at the conclusion of the study were included in this survival analysis and right-censored at the end of the study period. While the median length of neonatal intensive care unit (NICU) stay was similar in both arms (0.5=50%=51 days), the PP arm started to show trends towards decreased length of stay after the 50% median time to NICU discharge. The interquartile lengths of stay at 75% were 44 days for the attention-control (AC) arm and 40 days for the PP arm, and at 25% were 85 days for the AC arm and 75 days for the PP arm.

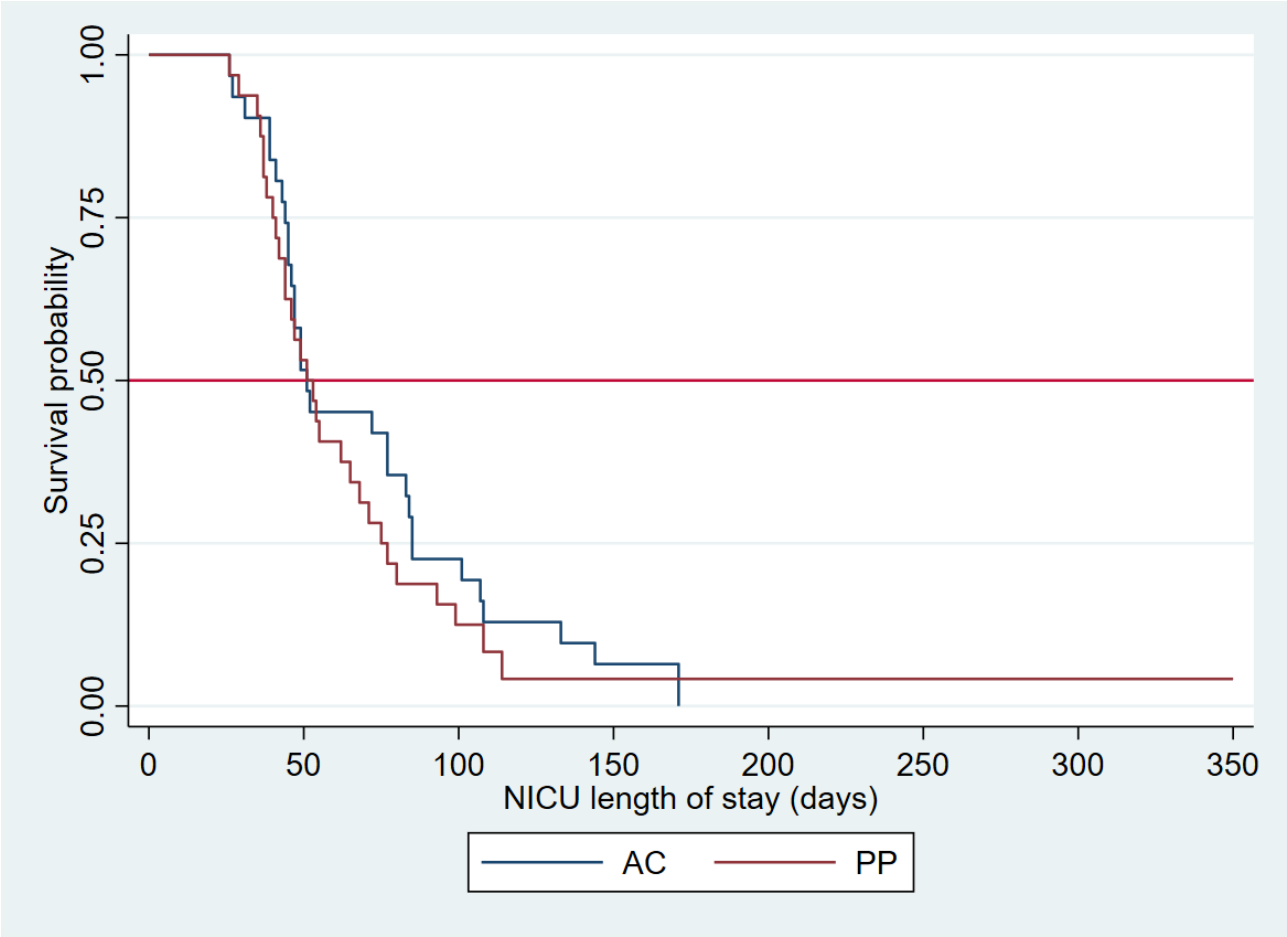

Mediator Variables by Trial Arm

Table 7 provides descriptive and summary statistics by arm for our hypothesized mediating variables. Mothers in both arms reported significant increases in condition (ie, prematurity) management ability at T3 and T4 (Table 6; Figure 3, middle left). Mothers in both arms reported significant decreases in family life difficulty at 30 days after NICU discharge (Table 6; Figure 3, middle right). Mothers in both arms reported significant decreases in NICU-related stress and significant increases in parenting self-efficacy at 28 days after baseline (Table 6; Figure 3, bottom row). Mothers in the PP arm showed promising trends for greater improvements in parenting self-efficacy (mean change 12.23, SD 3.37; 95% CI 5.62-18.84) versus AC mothers (mean change 7.76, SD 2.80; 95% CI 2.27, 13.26) 28 days after baseline (Table 6; Figure 3, lower right corner). Mothers in both arms did not have significant or clinically meaningful changes from baseline in postpartum bonding (Table 6).

XSL•FO
RenderX

Table 7. Descriptive and summary statistics for mother-reported mediators by trial arm (n=64).

Domains and time periods	Attention control (n=31), n (%); mean (SD; range)	Preemie Progress (n=33), n (%); mean (SD; range)
Family life difficulty		
T1[a]	30 (97); 30.1 (10.3; 15-53)	31 (94); 29.7 (10.9; 15-54)
T2[b]	30 (97); 28.7 (10.0; 15-55)	24 (73); 29.1 (11.8; 15-56)
T3[c]	28 (90); 28.8 (10.0; 15-54)	19 (58); 26.3 (12.3; 15-55)
T4[d]	24 (77); 24.8 (7.3; 15-40)	16 (48); 23.8 (9.6; 15-47)
Condition (ie, prematurity) management ability		
T1	30 (97); 46.5 (5.7; 33-60)	31 (94); 47.2 (6.7; 34-60)
T2	30 (97); 47.3 (6.5; 29-60)	24 (73); 47.8 (6.0; 35-58)
T3	28 (90); 50.0 (4.8; 40-60)	19 (58); 51.2 (7.6; 34-60)
T4	24 (77); 51.1 (7.1; 39-60)	16 (48); 51.6 (5.8; 39-60)
Perceived maternal parenting self-efficacy		
T1	31 (100); 51.5 (18.6; 13-80)	33 (100); 58.1 (19.8; 8-80)
T2	30 (97); 54.4 (19.2; 0-80)	24 (73); 62.0 (11.9; 33-80)
T3	28 (90); 59.0 (19.4; 0-80)	19 (58); 69.6 (10.4; 47-80)
Maternal postpartum bonding		
T1	31 (100); 5.3 (5.7; 0-23)	33 (100); 5.9 (6.4; 0-30)
T2	30 (97); 4.7 (4.3; 0-6)	24 (73); 5.1 (5.1; 0-6)
T3	27 (87); 4.1 (4.5; 0-15)	19 (58); 3.9 (5.4; 0-6)
NICU[e]-related stress		
T1	31 (100); 1.7 (0.8; 0.4-3.3)	33 (100); 1.7 (1.0; 0.2-4.3)
T2	30 (97); 1.5 (0.8; 0.0-3.4)	23 (70); 1.6 (1.0; 0-4)
T3	26 (84); 1.3 (0.7; 0.3-2.9)	19 (58); 1.3 (0.7; 0.4-2.9)

[a]T1: baseline.

[b]T2: 14 days after T1.

[c]T3: 28 days after T1.

[d]T4: 30 days after discharge.

[e]NICU: neonatal intensive care unit.

Adverse Events and Harms

There were no adverse events attributed to this study. One infant died during the study period due to a rare and lethal genetic condition, which was diagnosed after several months of infant life. In our data safety monitoring plan, we anticipated that mothers might experience clinically significant symptoms of anxiety, depression, and psychological distress due to high rates of anxiety and depression in the general population of mothers with infants in the NICU. Three mothers (all in the AC arm) experienced significant self-reported psychological distress or clinically significant symptoms of anxiety or depression (ie, T score >75.0 on the PROMIS 8a anxiety or depression scales), which resulted in the NICU team being notified and referral to mental health support. Symptoms in these mothers improved after referral.

Discussion

Overview

This pilot trial is one of the very few studies in the United States [6,15] investigating parent training designed to help parents comanage complex infant care within the first few weeks of a very preterm birth, which can be a traumatic and chaotic time for parents. This discussion is organized around the trial results and recommendations to optimize the feasibility domains of recruitment, retention, fidelity, and sensitivity to changes in outcomes. In addition, the results in this pilot trial are compared to those previously reported by researchers conducting FICare studies in the United States and around the world.

Feasibility of Recruitment and Retention

It is critical to note that in many FICare trials, parents are either required or highly encouraged to be present at the infant's bedside 6 to 8 hours a day. As a result, FICare trials may inadvertently include only parents who have the resources to

XSL•FO
RenderX

commit to such prolonged, focused time at their infant's bedside. Unlike our study, in which we could not make contact with approximately 28% (34/123) of the eligible mothers for participation, the FICare study by Franck et al [15] found that approximately 63% of the mothers in their FICare cohort from the United States could not be approached (and therefore recruited) for study participation, and the authors did not distinguish between mothers who could not be approached and those who were ineligible. However, similar to our study, another study by Franck et al [9] reported that 28% and 16% of the enrolled mothers did not complete surveys at infant discharge in their control and mobile FICare cohorts, respectively. The case-control study by Kubicka et al [6] achieved a 90% response rate for maternal surveys, which included instruments used in our study, such as the Parental Stressor Scale: NICU, and the Perceived Maternal Parenting Self-Efficacy Scale; however, the measures were only sent once to mothers at 30 days of infant life; and mothers in the post-FICare cohort were predominantly White (80%), married (96.7%), and highly educated (70% held a bachelor's degree or higher).

In their FICare trial, van Veenendaal et al [14] found that mothers in the Netherlands who did not complete study surveys were less likely to hold a university degree and be employed and more likely to have infants with longer NICU stays and infants born at less than 32 weeks GA, which are key demographic characteristics of the sample in our study. The high-risk nature of our patient population may explain the higher-than-anticipated loss to follow-up rate in the PP arm. We were surprised to find that mothers who did not complete our study did not differ on any baseline characteristics or on any baseline outcome measures such as family life difficulty and condition management ability. However, mothers in both arms who did not complete the study watched significantly fewer videos. It may be that for these mothers, their life was chaotic in multiple domains, and their ability to complete study procedures and manage their infant's illness was overwhelming. Very few FICare trials have been conducted in the United States [6,15], and among those conducted, maternal survey measures were typically only assessed at baseline and around discharge [6]. Our design assessed the impact of the PP intervention throughout NICU hospitalization, which may be overwhelming for mothers in the beginning weeks of infant life and may have led to higher loss to follow-up rates.

Feasibility of Fidelity to PP

In several FICare trials [9,12,14], researchers have found that only mothers who participated in education sessions and medical rounds demonstrated improvements in outcomes such as decreased anxiety. Thus, PP treatment dose and fidelity are likely important for achieving the intended outcomes. A significant strength of our study is that we were able to obtain the number of videos watched for each mother in our study. Mothers in the PP arm watched a median of 9 (IQR 2-39) videos, which suggests that watching all PP videos could be unrealistic for some mothers with infants in the NICU. Similar to our results, Franck et al [15] found in their US trial of a mobile FICare intervention that mothers logged into their mobile education app a median of 10 times during NICU hospitalization, with highly variable maternal use of the We3health app and no effect of the app on infant outcomes. Less than 30% of FICare parents were paired with a parent mentor, and approximately half of the parents attended ≥3 FICare parent group education classes during their infant's entire NICU hospitalization [15], which we believe further emphasizes the need for parent training interventions that do not rely on in-person methods.

It is imperative that future work around adapted FICare programs optimize the packaging, sequencing, and delivery of web-based parent training. Future trials of the PP intervention could prioritize and segment smaller portions of PP videos to make the intervention feasible. Having mothers select up front what videos they want to focus on may make the PP intervention easier to complete. At the same time, watching only part of the PP program may have implications for PP efficacy. Future trials should also examine how to optimize feasibility with fidelity while preserving the efficacy of the PP intervention.

Feasibility of Sensitivity to Changes in Outcomes

The results from our pilot trial were consistent with our belief that PP mothers would experience lower anxious symptoms and increased parenting self-efficacy and that their infants would have decreased length of stay in the NICU and increased receipt of mother's milk. Our trial is also consistent with the systematic review and meta-analysis conducted by North et al [22], which demonstrated that interventions designed to increase family involvement in preterm infant care significantly decrease maternal anxiety, reduce NICU length of stay, and increase infant receipt of mother's milk. We were surprised that in our sample, PP mothers did not show trends toward greater improvements in maternal depression, NICU-related stress, and postpartum bonding. However, maternal levels of depression, NICU-related stress, and impaired bonding in our sample were considerably lower than those reported by other trials [9,14]. In FICare trials, other researchers have noted that associations between the intervention and maternal mental health symptoms can often only be detected among mothers with initial high burden of mental health symptoms [9]. Thus, additional analyses with subpopulations of mothers with high symptom burden will be warranted in a larger PP trial.

Many of our maternal-reported outcomes only started to improve at 28 days after baseline, which has important implications for future trials of PP. Future trials of PP could consider reducing the number of maternal-reported surveys at 14 days after baseline to decrease study burden because maternal-reported outcome measures did not demonstrate clinically meaningful change at 14 days after baseline. The impact of loss to follow-up on our maternal-reported outcomes could also mean that mothers who were more resilient in multiple areas of life were able to complete survey measures at 28 days after baseline and 30 days after NICU discharge. In many FICare trials, maternal mental health symptoms (eg, stress and anxiety) are only assessed at enrollment and at NICU discharge [12,14,35]. Thus, the impact of FICare on the trajectory of maternal mental health symptoms during NICU hospitalization is not fully known. A significant strength of our study is that our data collection time points at 14 days and 28 days after baseline were standardized for all

XSL•FO
RenderX

infants and did not rely on time to infant discharge, which can be highly variable based on infant illness.

Strengths and Limitations

There are many strengths associated with our pilot trial, including the strong study design; our diverse sample of mothers; a high recruitment rate (64/123, 52%); and the use of blinding, randomization, and validated survey measures. Our use of website analytics allowed us to capture the number of videos watched for each mother, including those mothers who dropped out of the study, which would not have been possible with self-report measures of fidelity. Our use of both in-person and telephone methods of recruitment allowed us to reach a diverse group of mothers, including mothers from a wide range of socioeconomic backgrounds and those who experienced barriers to parent presence in the NICU. Finally, our outcome data showed promising trends for lower anxiety symptoms in the PP arm.

A higher-than-anticipated loss to follow-up rate in the PP arm and large variation in PP fidelity (mean 17.8, SD 18.9 of 49 videos watched) are limitations of this work. The small sample size, while appropriate for a pilot RCT, meant that this study was not powered to detect significant differences in mediator and outcome variables; therefore, the data trends must be interpreted with caution. Our results need to be replicated and confirmed with larger sample sizes.

Recommendations for Further Research

While we met our recruitment goals, we experienced a higher-than-anticipated loss to follow-up rate in the PP arm. Additional analyses are needed to understand how the PP intervention and study procedures can be refined to optimize PP fidelity and retention rates. Thus, further studies are warranted before a future full-scale efficacy trial can commence. We are currently conducting a companion study analyzing qualitative feedback from mothers in this trial to better understand their experiences with the PP intervention and with contextual factors that influenced their ability to adhere to PP, complete the study, and manage care for their infant. Future studies using mixed methods analyses have the potential to enhance our understanding of the efficacy of PP and offer additional strategies to increase the effectiveness, adoption, and implementation of this intervention for real-world clinical settings.

Conclusions

Training mothers in evidence-based family management skills has the potential to improve maternal, infant, and health care use outcomes in the NICU and beyond. Mothers in the PP arm showed trends toward greater improvements in anxious symptoms and parenting self-efficacy, and their infants showed trends toward decreased length of stay in the NICU and increased receipt of mother's milk. While we met our recruitment goals, additional studies are needed to optimize PP and trial procedures so that retention and fidelity goals can be met. However, the PP intervention shows great promise in supporting parent training as a stand-alone program outside of traditional FICare models or as a complement to adapted FICare models in the United States.

Acknowledgments

The authors would like to thank Shriya Parthasarathi for her contributions in creating the tables, figures, and supplementary files for this manuscript. Research reported in this publication was also supported by the National Institute of Nursing Research of the National Institutes of Health under Award Number K23NR019081. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Data Availability

The datasets generated and analyzed during this study are not publicly available due to participant confidentiality and ethical restrictions but are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
CONSORT (Consolidated Standards of Reporting Trials) 2010 checklist.
[PDF File (Adobe PDF File), 155 KB - formative_v9i1e66073_app1.pdf]

Multimedia Appendix 2
TIDieR (Template for Intervention Description and Replication) checklist.
[PDF File (Adobe PDF File), 124 KB - formative_v9i1e66073_app2.pdf]

Multimedia Appendix 3
CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) checklist (V 1.6.1).

XSL•FO
RenderX

[PDF File (Adobe PDF File), 54200 KB - formative_v9i1e66073_app3.pdf]

Multimedia Appendix 4
Preemie Progress modules.
[DOCX File , 18 KB - formative_v9i1e66073_app4.docx]

References

1. Horbar JD, Greenberg LT, Buzas JS, Ehret DE, Soll RF, Edwards EM. Trends in mortality and morbidities for infants born 24 to 28 weeks in the US: 1997-2021. Pediatrics 2024 Jan 01;153(1):e2023064153. [doi: 10.1542/peds.2023-064153] [Medline: 38053449]
2. Bell EF, Hintz SR, Hansen NI, Bann CM, Wyckoff MH, DeMauro SB, Eunice Kennedy Shriver National Institute of Child HealthHuman Development Neonatal Research Network. Mortality, in-hospital morbidity, care practices, and 2-year outcomes for extremely preterm infants in the US, 2013-2018. JAMA 2022 Jan 18;327(3):248-263 [FREE Full text] [doi: 10.1001/jama.2021.23580] [Medline: 35040888]
3. Walsh EM, Li SX, Black LK, Kuzniewicz M. Incremental cost of prematurity by week of gestational age. AJP Rep 2019 Jan;9(1):e76-e83 [FREE Full text] [doi: 10.1055/s-0039-1683934] [Medline: 31041115]
4. Vohr B, McGowan E, Keszler L, O'Donnell M, Hawes K, Tucker R. Effects of a transition home program on preterm infant emergency room visits within 90 days of discharge. J Perinatol 2018 Feb;38(2):185-190. [doi: 10.1038/jp.2017.136] [Medline: 28906495]
5. Jiang S, Warre R, Qiu X, O'Brien K, Lee SK. Parents as practitioners in preterm care. Early Hum Dev 2014 Nov;90(11):781-785. [doi: 10.1016/j.earlhumdev.2014.08.019] [Medline: 25246323]
6. Kubicka Z, Fiascone J, Williams D, Zahr E, Ditzel A, Perry D, et al. Implementing modified family integrated care in a U.S. neonatal intensive care unit: nursing perspectives and effects on parents. J Perinatol 2023 Apr;43(4):503-509 [FREE Full text] [doi: 10.1038/s41372-023-01601-y] [Medline: 36627393]
7. Sanders MR, Hall SL. Trauma-informed care in the newborn intensive care unit: promoting safety, security and connectedness. J Perinatol 2018 Jan;38(1):3-10 [FREE Full text] [doi: 10.1038/jp.2017.124] [Medline: 28817114]
8. Beck DC, Tabb KM, Tilea A, Hall SV, Vance A, Patrick SW, et al. The association between NICU admission and mental health diagnoses among commercially insured postpartum women in the US, 2010-2018. Children (Basel) 2022 Oct 13;9(10):33 [FREE Full text] [doi: 10.3390/children9101550] [Medline: 36291486]
9. Franck LS, Gay CL, Hoffmann TJ, Kriz RM, Bisgaard R, Cormier DM, et al. Maternal mental health after infant discharge: a quasi-experimental clinical trial of family integrated care versus family-centered care for preterm infants in U.S. NICUs. BMC Pediatr 2023 Aug 10;23(1):396 [FREE Full text] [doi: 10.1186/s12887-023-04211-x] [Medline: 37563722]
10. O'Brien K, Robson K, Bracht M, Cruz M, Lui K, Alvaro R, FICare Study Group and FICare Parent Advisory Board. Effectiveness of family integrated care in neonatal intensive care units on infant and parent outcomes: a multicentre, multinational, cluster-randomised controlled trial. Lancet Child Adolesc Health 2018 Apr;2(4):245-254. [doi: 10.1016/S2352-4642(18)30039-7] [Medline: 30169298]
11. Hei M, Gao X, Li Y, Gao X, Li Z, Xia S, et al. Family integrated care for preterm infants in China: a cluster randomized controlled trial. J Pediatr 2021 Jan;228:36-43.e2. [doi: 10.1016/j.jpeds.2020.09.006] [Medline: 32898578]
12. Ansari NS, Franck LS, Tomlinson C, Colucci A, O'Brien K. A pilot study of family-integrated care (FICare) in critically ill preterm and term infants in the NICU: FICare plus. Children (Basel) 2023 Aug 02;10(8):1337. [doi: 10.3390/children10081337] [Medline: 37628336]
13. Banerjee J, Aloysius A, Mitchell K, Silva I, Rallis D, Godambe SV, et al. Improving infant outcomes through implementation of a family integrated care bundle including a parent supporting mobile application. Arch Dis Child Fetal Neonatal Ed 2020 Mar 21;105(2):172-177. [doi: 10.1136/archdischild-2018-316435] [Medline: 31227521]
14. van Veenendaal NR, van Kempen AA, Broekman BF, de Groof F, van Laerhoven H, van den Heuvel ME, et al. Association of a zero-separation neonatal care model with stress in mothers of preterm infants. JAMA Netw Open 2022 Mar 01;5(3):e224514 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.4514] [Medline: 35344044]
15. Franck LS, Gay CL, Hoffmann TJ, Kriz RM, Bisgaard R, Cormier DM, et al. Neonatal outcomes from a quasi-experimental clinical trial of family integrated care versus family-centered care for preterm infants in U.S. NICUs. BMC Pediatr 2022 Nov 22;22(1):674 [FREE Full text] [doi: 10.1186/s12887-022-03732-1] [Medline: 36418988]
16. Alferink MT, Moreno-Sanz B, Cabrera-Lafuente M, Ergenekon E, de Haan TR, van Kempen AA, RISEinFAMILY Consortium (European Union, MSCA-RISE-H2020). RISEinFAMILY project: the integration of families at neonatal intensive care units (NICUs) to empower them as primary caregivers: study protocol for a stepped wedge cluster controlled trial. Trials 2024 Apr 10;25(1):248 [FREE Full text] [doi: 10.1186/s13063-024-08043-7] [Medline: 38594733]
17. Weber AM, Voos KC, Bakas TM, Rice JB, Blatz MA, Ribeiro AP, et al. A clinical-academic partnership to develop a family management intervention for parents of preterm infants. J Clin Nurs 2022 Feb;31(3-4):390-405 [FREE Full text] [doi: 10.1111/jocn.15929] [Medline: 34219302]

XSL•FO
RenderX

18. Weber AM, Bakas T, Schulman-Green D, Voos KC, Rice JB, Bailey R, et al. Family management skills reported by parents of preterm infants in the NICU using the self- and family management framework (SFMF). Adv Neonatal Care 2024 Apr 01;24(2):119-131. [doi: 10.1097/ANC.0000000000001140] [Medline: 38127650]
19. Home page. Preemie Progress. URL: https://www.preemieprogress.org/ [accessed 2024-04-29]
20. Borrelli B. The assessment, monitoring, and enhancement of treatment fidelity in public health clinical trials. J Public Health Dent 2011;71(s1):S52-S63 [FREE Full text] [doi: 10.1111/j.1752-7325.2011.00233.x] [Medline: 21499543]
21. Borrelli B, Sepinwall D, Ernst D, Bellg AJ, Czajkowski S, Breger R, et al. A new tool to assess treatment fidelity and evaluation of treatment fidelity across 10 years of health behavior research. J Consult Clin Psychol 2005 Oct;73(5):852-860. [doi: 10.1037/0022-006X.73.5.852] [Medline: 16287385]
22. North K, Whelan R, Folger LV, Lawford H, Olson I, Driker S, et al. Family involvement in the routine care of hospitalized preterm or low birth weight infants: a systematic review and meta-analysis. Pediatrics 2022 Aug 01;150(Suppl 1):e2022057092O. [doi: 10.1542/peds.2022-057092O] [Medline: 35921672]
23. Jaeger CB, Altimier L. NICU Couplet Care: metrics to guide an evolving model of care. J Perinatol 2023 Dec;43(Suppl 1):30-34. [doi: 10.1038/s41372-023-01783-5] [Medline: 38086964]
24. Pilkonis PA, Choi SW, Reise SP, Stover AM, Riley WT, Cella D, PROMIS Cooperative Group. Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS®): depression, anxiety, and anger. Assessment 2011 Sep;18(3):263-283 [FREE Full text] [doi: 10.1177/1073191111411667] [Medline: 21697139]
25. O'Brien K, Bracht M, Robson K, Ye XY, Mirea L, Cruz M, et al. Evaluation of the family integrated care model of neonatal intensive care: a cluster randomized controlled trial in Canada and Australia. BMC Pediatr 2015 Dec 15;15(1):210 [FREE Full text] [doi: 10.1186/s12887-015-0527-0] [Medline: 26671340]
26. Fenton TR, Sauve RS. Using the LMS method to calculate z-scores for the Fenton preterm infant growth chart. Eur J Clin Nutr 2007 Dec;61(12):1380-1385. [doi: 10.1038/sj.ejcn.1602667] [Medline: 17299469]
27. White-Traut RC, Rankin KM, Yoder JC, Liu L, Vasa R, Geraldo V, et al. Influence of H-HOPE intervention for premature infants on growth, feeding progression and length of stay during initial hospitalization. J Perinatol 2015 Aug;35(8):636-641 [FREE Full text] [doi: 10.1038/jp.2015.11] [Medline: 25742287]
28. Welch MG, Hofer MA, Stark RI, Andrews HF, Austin J, Glickstein SB, et al. Randomized controlled trial of family nurture intervention in the NICU: assessments of length of stay, feasibility and safety. BMC Pediatr 2013 Sep 24;13:148 [FREE Full text] [doi: 10.1186/1471-2431-13-148] [Medline: 24063360]
29. Vohr B, McGowan E, Keszler L, Alksninis B, O'Donnell M, Hawes K, et al. Impact of a transition home program on rehospitalization rates of preterm infants. J Pediatr 2017 Feb;181:86-92.e1. [doi: 10.1016/j.jpeds.2016.10.025] [Medline: 27817878]
30. Pitetti K, Smith B, Hsiao R. Application of Perceived Maternal Parenting Self-Efficacy (PMP S-E) Questionnaire in a Mid-West Community Medical Center NICU in the United States. Neonat Pediatr Med 2016;02(01):e365 [FREE Full text] [doi: 10.4172/2572-4983.1000106]
31. Miles MS, Brunssen SH. Psychometric properties of the parental stressor scale: infant hospitalization. Adv Neonatal Care 2003 Aug;3(4):189-196. [doi: 10.1016/s1536-0903(03)00138-3] [Medline: 14502526]
32. Ionio C, Ciuffo G, Colombo C, Melani O, Figlino MF, Landoni M, et al. Preterm birth and maternal mood states: what is the impact on bonding? Pediatr Rep 2024 Jan 05;16(1):35-45 [FREE Full text] [doi: 10.3390/pediatric16010004] [Medline: 38251313]
33. Knafl K, Deatrick JA, Gallo A, Dixon J, Grey M, Knafl G, et al. Assessment of the psychometric properties of the family management measure. J Pediatr Psychol 2011 Jun;36(5):494-505 [FREE Full text] [doi: 10.1093/jpepsy/jsp034] [Medline: 19451173]
34. Whitehead AL, Julious SA, Cooper CL, Campbell MJ. Estimating the sample size for a pilot randomised trial to minimise the overall trial sample size for the external pilot and main trial for a continuous outcome variable. Stat Methods Med Res 2016 Jun;25(3):1057-1073 [FREE Full text] [doi: 10.1177/0962280215588241] [Medline: 26092476]
35. Zhang Y, Jiang M, Xiang X, He W, Du J, Hei M. Effect of family integrated care on maternal stress in preterm infants in the neonatal intensive care unit: a bi-center experience. Asia Pac Psychiatry 2023 Mar;15(1):e12521. [doi: 10.1111/appy.12521] [Medline: 36502821]

Abbreviations

AC: attention control
CONSORT: Consolidated Standards of Reporting Trials
CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
EHR: electronic health record
FICare: family integrated care
NICU: neonatal intensive care unit
PP: Preemie Progress

XSL•FO
RenderX

PROMIS: Patient-Reported Outcomes Measurement Information System
RCT: randomized controlled trial
REDCap: Research Electronic Data Capture
TIDieR: Template for Intervention Description and Replication

Edited by A Mavragani; submitted 04.11.24; peer-reviewed by R Heyne; comments to author 22.12.24; revised version received 12.03.25; accepted 16.03.25; published 21.05.25.

Please cite as:
Weber A, Bakas T, Oudat Q, Parikh NA, Lambert J, Tubbs-Cooley HL, Rice J, Voos K, Rota M, Kaplan HC
An Internet-Based and Mobile Family Management Intervention for Mothers of Very Preterm Infants Hospitalized in the Neonatal Intensive Care Unit (the Preemie Progress Program): Pilot Randomized Controlled Trial
JMIR Form Res 2025;9:e66073
URL: https://formative.jmir.org/2025/1/e66073
doi:10.2196/66073
PMID:

©Ashley Weber, Tamilyn Bakas, Qutaibah Oudat, Nehal A Parikh, Joshua Lambert, Heather L Tubbs-Cooley, Jared Rice, Kristin Voos, Matthew Rota, Heather C Kaplan. Originally published in JMIR Formative Research (https://formative.jmir.org), 21.05.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Detection of Depressive Symptoms in College Students Using Multimodal Passive Sensing Data and Light Gradient Boosting Machine: Longitudinal Pilot Study

Jessica L Borelli[1], PhD; Yuning Wang[2], PhD; Frances Haofei Li[1], MA; Lyric N Russo[1], PhD; Marta Tironi[3], PhD; Ken Yamashita[1], BA; Elayne Zhou[4], MS; Jocelyn Lai[1], PhD; Brenda Nguyen[1], BA; Iman Azimi[1], PhD; Christopher Marcotullio[1], BA; Sina Labbaf[1], PhD; Salar Jafarlou[1], PhD; Nikil Dutt[1], PhD; Amir Rahmani[1], PhD

[1]Department of Psychological Science, University of California, Irvine, Irvine, CA, United States

[2]Department of Computing, University of Turku, Turku, Finland

[3]Department of Educational Sciences, University of Genoa, Genoa, Italy

[4]Department of Psychology, University of Southern California, Los Angeles, CA, United States

Corresponding Author:
Jessica L Borelli, PhD
Department of Psychological Science
University of California, Irvine
4201 Social and Behavioral Sciences Gateway
Irvine, CA, 92617
United States
Phone: 1 203 887 8857
Email: jessica.borelli@uci.edu

Abstract

Background: Depression is the top contributor to global disability. Early detection of depression and depressive symptoms enables timely intervention and reduces their physical and social consequences. Prevalence estimates of depression approach 30% among college students. Passive, device-based sensing further enables detection of depressive symptoms at a low burden to the individual.

Objective: We leveraged an ensemble machine learning method (light gradient boosting machine) to detect depressive symptoms entirely through passive sensing.

Methods: A diverse sample of undergraduate students (N=28; mean age 19.96, SD 1.23 y; 15/28, 54% women; 13/28, 46% Latine; 10/28, 36% Asian; 4/28, 14% non-Latine White; 11/28, 4% other) participated in an intensive longitudinal study. Participants wore 2 devices (an Oura ring for sleep and physiology data, and a Samsung smartwatch for physiology and movement data) and installed the AWARE software on their mobile devices, which collects passive sensing data such as screen time. Participants were derived from a randomized controlled trial of a positive psychology mobile health intervention. They completed a self-report measure of depressive symptoms administered weekly over a 19- to 22-week period.

Results: The light gradient boosting machine model achieved an F_1-score of 0.744 and a Cohen κ coefficient of 0.474, indicating moderate agreement between the predicted labels and the ground truth. The most predictive features of depressive symptoms were sleep quality and missed mobile interactions.

Conclusions: Findings suggest that data collected from passive sensing devices may provide real-time, low-cost insight into the detection of depressive symptoms in college students and may present an opportunity for future prevention and perhaps intervention.

(JMIR Form Res 2025;9:e67964) doi:10.2196/67964

KEYWORDS

depression; college students; emerging adulthood; machine learning; passive sensing

XSL•FO
RenderX

Introduction

Depression in College Students

Depression is the foremost contributor to global disability [1]. Longitudinal studies show that symptoms of depression typically begin in a person's 20s to early 30s [2]. In recent years, college students' mental health has worsened, with major depression rising disproportionately within this population [3-5]. College students assessed during the COVID-19 pandemic witnessed a 300% increase in the risk of developing depressive disorders as compared to the previous 8 years [6]. Given the cascading mental health and cognitive consequences of depressive symptoms [7,8], there is an urgent need for early detection of depression in college students to inform prevention and intervention efforts. The presence of depressive symptoms is a key indicator of risk for full-blown clinical depressive syndromes [9]. Consistent with a developmental psychopathology framework [10], this study examined the presence of depressive symptoms within a community sample of college students to develop and test an algorithm to predict depressive symptoms, and hence, risk for depressive disorders.

The Biopsychosocial Model of Depression Risk

The biopsychosocial model provides an integrative perspective that underscores the roles of different risk factors in predicting depression risk [11]. This model captures the complexity of individuals' environments, acknowledging that everyone's health is affected by a variety of biological, psychological, and social or interpersonal factors that coexist (Figure 1). Examining these factors in isolation provides an incomplete view of mental health. Thus, the central goal of this study was to use multiple indicators of functioning, including biological, psychological, and social factors, to predict depressive symptoms.

Figure 1. Conceptual model underpinning the use of multimodal real-world monitoring data to predict depressive symptoms.

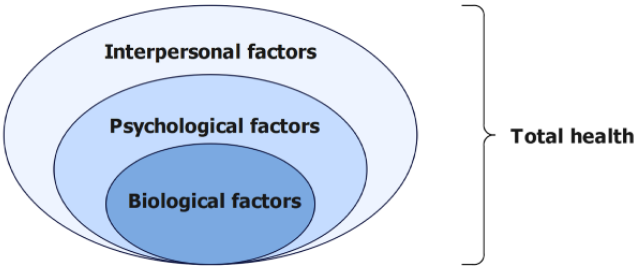

In support of the key tenets of this model, certain psychophysiological habits (eg, sleep and physical activity), along with specific behavioral factors associated with the lifestyle of emerging adults (eg, frequent use of electronic devices and mobile phones), substantially impact well-being both physically and mentally. These habits, particularly the use of electronic devices, are closely linked with depression and may act as predisposing factors [12], particularly among college students. In the sections that follow, we review 3 factors that have strong evidence for their association with depression within college students.

Sleep and Depression Risk

Sleep quality has emerged as a reliable predictor of mental health among college students [13-15]. Individuals who fall asleep easily and maintain uninterrupted sleep exhibit lower levels of depression [16]. Conversely, poor sleep quality has emerged not only as a strong predictor of depression but also as a consequence of it [14,17], revealing a cyclic interplay between sleep patterns and mental health. Greater depressive symptoms in adolescence are associated with subjective and objective sleep difficulties in emerging adulthood; furthermore, these sleep difficulties are related to more depressive symptoms in emerging adulthood [18]. The importance of high-quality sleep for mental health was made particularly salient during the COVID-19 pandemic, when sleep disturbances increased among college students due to heightened stress, uncertainty, and disruptions to daily routines [12]. This reciprocal relationship highlights the importance of using objective measures to assess sleep quality and its effects on mental health, as these measures can provide valuable insights into sleep duration, efficiency, and fragmentation [19]. Considering these findings, investigating objective sleep quality as a predictor of depression is critical.

Smartphone Use and Depression Risk

The vast majority (97%) of college students own a smartphone and use it daily [20]. Studies identify a bidirectional and reciprocal relationship between the quality and quantity of smartphone use in college students with mood [21,22]. Considering the function and motivation behind smartphone use can be useful in making sense of these connections [23]. Smartphones can be used in ways that promote mental health, but they can also be used in ways that are destructive or harmful. When smartphone use becomes intense, compulsive, or indiscriminate, such as when smartphones are used to disengage from or avoid aversive emotions [19,24], depressive symptoms may arise. Although research has predominantly focused on the tolls of smartphone use on health [21,22], the prevalence of smartphones and their integration into daily social life have also motivated research into the potential harms of disengagement from smartphones. For instance, Elhai et al [12] found that higher levels of depressive symptoms were associated with lower social smartphone use in adults, such as lower use of social media for fostering relationships or staying in contact with people. To the best of our knowledge, the links between socially disengaged smartphone use and depressive symptoms have not been explored in college students. Furthermore, we contend that social isolation may also manifest through other dimensions of smartphone use (eg, beyond social media use), such as not answering calls or texts, and that this behavior may be bidirectionally connected with depressive symptoms. The more depressed students feel, the less they use their phone, and the more socially disengaged they feel, the less they have the

XSL•FO
RenderX

urge to use it to interpersonally interact. Moving forward, it is necessary to expand our knowledge regarding the frequency of social interactions, assessing the amount of engagement between college students and their smartphones in their daily life (ie, the rate of messages and calls) to better understand the connection with depressive symptoms. This study builds upon existing research by exploring these questions in college students and focusing on additional indicators of smartphone engagement or disengagement.

Physical Activity and Depression Risk

Physical activity is a robust predictor of mental health and well-being in college students [25,26]. More physical activity in college students predicts lower anxiety, depression, and stress, as well as higher self-esteem, and overall improved mental health [24,27-31]. Physical activity decreases across the transition from adolescence to young adulthood, particularly for female individuals [27,32], which is concerning given its link with mental health. Importantly, physical activity is a modifiable behavior [33], making it an ideal target for intervention [24,34]. Particularly during the COVID-19 pandemic, physical inactivity emerged as a powerful predictor of depression, when options for remote learning reduced the need for movement and promoted a sedentary lifestyle [35,36]. Objective assessments of physical activity have the advantage of providing accurate documentation of behavior that may be difficult to remember or report. A recent meta-analysis found that objectively measured indices of physical activity are associated with depression risk [37]. Multiple indices of objectively-assessed physical activity, such as time spent being stationary [13,38], energy consumed [39], and number of steps taken [40], can be unobtrusively obtained from devices, providing a source of data that avoids some of the pitfalls of self-report measures of physical activity.

Leveraging Machine Learning Methods

The use of machine learning (ML) methods with sleep, smartphone use, and physical activity data may offer unique advantages and insights in the detection of psychopathology. Among the array of ML approaches, classification tasks stand out for their capacity to map attributes (representing features of data instances) to a designated label, such as the presence of depression. Classification tasks remain the most used ML method in the diagnosis and detection of mental illnesses. A light gradient boosting machine (LightGBM) [41] is a form of supervised learning algorithm that offers state-of-the-art performance while remaining highly efficient. Built on a decision-tree architecture, LightGBM uniquely incorporates gradient-based information, allowing it to handle large-scale datasets while remaining model-efficient and using little memory.

While research applying and advancing ML methods to detect psychopathology is still in its early stages, there has been a surge of such research in recent years. Studies have applied various ML algorithms, encompassing different classes (ie, traditional ML and deep learning) and types (ie, supervised and unsupervised learning) of ML approaches. In addition, existing studies have also leveraged diverse types (ie, physiological, semantic, acoustic, or facial feature) of data [42-44]. These studies suggest some common features associated with depression, such as acoustic volumes (eg, lower pitch associated with more severe depression) [45], social media use (higher problematic social media use associated with more severe depression) [46], combined facial features (eg, the movement of pupil and mouth) [13]. Notably, 1 limitation of many existing studies is that data were collected from highly controlled laboratory settings and may not accurately reflect individuals' day-to-day behaviors [47]. Combining LightGBM with wearable device data collection can offer precise and individualized detections of depression from data gathered in naturalistic settings. In addition, explainable artificial intelligence methods can be used to analyze the relative impact of each variable within the collected data on depression detections.

This Study

Building upon previous work [48-51], this study uses digital detection processes as an indicator of depressive symptoms risk. Depressive symptoms can fluctuate over time, and accurately capturing depressive states may necessitate a fine-grained assessment methodology. We used 3 months of data from a randomized clinical trial conducted with a community sample of college students. We assessed depressive symptoms using a commonly used and reliable measure, the Patient Health Questionnaire-9 (PHQ-9) [52], to provide the most comprehensive assessment of depressive symptoms.

We pursued 3 aims. First, we sought to classify samples using a binary system, which we refer to as depressed or nondepressed (above or below PHQ-9 scores of 4, or mild depression) based on our data using ML techniques (aim 1). This aim involved examining the overall performance of the depression detection system. Meta-analyses on the efficacy of these models found that they predict the risk of depression based on both psychological and behavioral factors [40,53,54]. Using a confusion matrix to analyze the performance of the ML model, we hypothesized that the model would be able to differentiate between depressed and nondepressed participants. It is worth noting that although we use the terms *depressed* or *nondepressed* for ease of communication, it is important to recall that we examine these research questions in a community sample. Furthermore, we use a self-report questionnaire to assess depressive symptoms rather than clinician interviews of depressive disorders. Therefore, our use of these terms is shorthand, used for ease of communication, but should not be viewed as indicative of a clinical diagnosis.

Second, we examined depressive symptom detection across intervention versus control groups within our study (aim 2). The data used in this study were part of a randomized controlled trial examining a just-in-time mobile health intervention (ie, relational savoring) [55,56] designed to prevent loneliness (the study by Nguyen et al [unpublished data, 2024] provides an overview of the design of the intervention component of the study). Given that inclusion in the intervention condition could impact depressive symptoms as well as the relationship between contextual factors (ie, sleep, physiology, and behavior) with depressive symptoms, aim 2 involved examining depression detection separately across groups.

XSL•FO
RenderX

Third, we examined the features that contributed to the prediction of depressive symptoms (aim 3). We focused on the features that were important for the ML model to accurately identify indicators of depression. On the basis of previous literature, sleep quality is a variable that has repeatedly been found to be linked to depression [17,57,58]. From this, we developed our hypothesis that sleep quality measured through breathing patterns and movement would be the greatest indicator of depressive symptoms among physiological and behavioral measurements. Growing research has also identified earlier problematic mobile phone use as a predictor of depression later in development [59] and that call activity has been traced to indicate depression [60-62]. Therefore, we hypothesized that messaging and call frequency would be significant indicators among mobile activities when measuring depression. Finally, based on research linking physical activity to depression risk [37], we anticipated that low physical activity would be associated with depression risk. Exploratory analyses were conducted as a bottom-up approach where all potential passive sensing features were included in the same model to see which features emerged as most predictive.

The pursuit of these 3 aims combined to allow us to examine the utility of predicting depressive symptoms within a sample of college students using the LightGBM method, contributing to the knowledge base regarding ML and depression detection.

Methods

Participants

Undergraduate students between 18 and 22 years old were recruited through flyers and campus announcements at a large West Coast university in the United States. Participants met the eligibility criteria if they were fluent in English and used an Android smartphone with an operating system of 6.0 or higher (the study by Borelli [55] provides full details on participation eligibility). Exclusion criteria included anyone who is a parent, married, coming back to school after some time, aged >22 years, unable to speak or write English fluently, or currently meets criteria for depression. These exclusion criteria were included due to our desire to generalize our findings to the general population of college students. In total, 37 participants enrolled, with 10 withdrawing their participation over the course of the study (N=28; mean age 19.96, SD 1.23 years; 15/28, 54% women). Out of 28 participants, 13 (46%) identified as Latine, 10 (36%) as Asian, 3 (14%) as non-Latine White, and 1 (4%) participant as other. Participants' demographic information is provided in Table 1.

Table 1. Participants' demographic information (N=28).

Parameter	Participants
Age (y), mean (SD)	19.96 (1.23)
Gender, n (%)	
Women	15 (54)
Men	13 (46)
Year level in college, n (%)	
First	4 (14)
Second	4 (14)
Third	12 (43)
Fourth	8 (29)
Race and ethnicity, n (%)	
Asian	10 (36)
Latine	13 (46)
White	4 (14)
Other	1 (4)

Ethical Considerations

The study was approved by the principal investigator's institutional review board (HS 2019-5153) and informed consent was obtained from all participants before the inception of data collection. Depending on the completion of study components, participants were compensated between US $30 to US $660. We structured data handling into 3 phases: acquisition, transmission, and storage. Sensors were registered via media access control addresses, with secure and dynamic matching between users and devices. Access was restricted based on user level and duration, with University of California Irvine Office of Information Technology (OIT) conducting security assessments. For transmission, secure sockets layer, firewalls, and multifactor authentication secure communications were used, while OIT provided intrusion detection and monitoring. Local storage used encryption and key management, while cloud data were encrypted, with identifiers stored securely for report generation. OIT ensured secure transmission, backup, recovery, and disposal. We followed legal standards to protect personal data. The Anonymization and Reidentification Exchange Data Anonymization Tool removed identity traces, replacing them with local IDs. Sensitive details were abstracted, and personal data were locally processed, not stored in the cloud. Automated

XSL•FO
RenderX

tools handle anonymized cloud data, with access limited to authorized Institutional Review Board–approved researchers.

Procedures

All 28 participants began enrollment in the study between January and February of 2022 and participated for at least 4.5 months (19 weeks). After being deemed eligible, participants were scheduled for an in-person session where they were provided with wearable smart devices (Oura ring [Oura Health Ltd] and Samsung watch [Samsung Electronics Co, Ltd]) and directed to install study-related phone apps (AWARE) to record passive sensing data, as well as a different app to record daily and weekly surveys. After the research staff provided instructions on how to set up and use their devices, participants were informed on how to complete the questionnaires administered on a daily and weekly basis. Research assistants asked participants to wear or keep their devices with them as much as possible unless charging or engaging in any intense activity (eg, sports) that might damage the devices.

For subsequent procedures, we focus on those relevant to the study aims. Following the in-person laboratory session, participants completed a 6-week monitoring period, during which they wore the smart devices. On each Sunday during their participation in the study, participants completed weekly depression questionnaires. Passive sensing data (eg, geographic location, physiology, actigraphy, sleep, among others) were collected through the smart devices. After the 6-week monitoring period, participants were randomly assigned to 1 of 2 conditions—monitoring only, that is, the control condition in which participants continued as during the monitoring phase, and intervention condition, that is, the experimental condition in which participants received a mobile health intervention named relational savoring [53,54] delivered via their smartphone based on an algorithm. The central aim of the intervention was to examine whether delivery of this intervention reduced participant loneliness. Although the results of this intervention are outside the scope of this investigation, we refer interested readers to the reported findings present in the study by Borelli [55]. This randomized monitoring or intervention period phase lasted 4 weeks, followed by ≥9 weeks of monitoring only.

Measures

Assessment of Depressive Symptoms

PHQ-9 [50] is a 9-item questionnaire and a standardized tool designed to screen, diagnose, monitor, and measure the severity of depressive symptoms, including mood, anhedonia, sleep, appetite, concentration, and suicidal ideation. The traditional questionnaire, which asks about depressive symptoms experienced over the past 2 weeks, was modified for use in this weekly context, for example, “over the last week, how often have you been bothered by any of the following problems: little interest or pleasure in doing things?” The PHQ-9 has robustly demonstrated validity and reliability across diverse populations, including US college students, which was the targeted population in this study [63]. While this measure has not been specifically validated with weekly administration, it has been widely used in published ecological momentary assessment–related studies for weekly administration, for example, in the study by Nickels et al [64] and with varying frequencies (eg, 3 times per day) [65], demonstrating appropriateness to be used in repeated measurements.

Items were rated on a 4-point Likert scale (from 0=not at all to 3=nearly every day) and summed, where higher scores indicated higher levels of depression. As expected for the current sample, depressive symptoms were relatively low across the study period (mean 4.90, SD 4.25). Thus, for this study, to address our aims of predicting depressive symptomatology and identifying the features most predictive of these symptoms, we created a dichotomous variable using a clinical cutoff: participants with scores >4 were identified as being in the “follow-up needed” group, whereas participants with scores of ≤4 were identified as being in the “none-minimal” group. Internal consistency in this sample was good (α=0.83). As described earlier, participants completed this measure weekly; participants were monitored throughout the assessment with supervision from the first author, who has clinical training and expertise. After 3 months of monitoring, there were a total of 355 valid PHQ-9 submissions and 4 missed submissions. The valid submissions were used as the depression labels in the subsequent analysis.

Passive Sensing of Health, Sleep, and Behavior

To capture an accurate depiction of participants’ daily physical habits, sleep, and health, we fitted participants with the Oura ring and Samsung Gear Sport smartwatch and downloaded the corresponding Oura and Samsung Android mobile apps. The Oura ring assesses sleep quality by measuring sleep duration, average heart rate (HR) during sleep, and HR variability (HRV) during sleep [66]. It also collects physical activity parameters, including energy consumption and moving steps. In addition to sleep and physical activity data, the Samsung Gear Sport smartwatch collected raw photoplethysmogram (PPG) signal using a green light LED sensor for 12 minutes every 2 hours, from which daily HR and daily HRV were extracted [67].

The watch feature engineering process involved extracting HR and HRV measures from PPG data. To extract the HR and HRV features, the PPG signals were divided into nonoverlapping 5-minute segments. Each segment was processed individually using the pipeline described in the Data Preprocessing section. The features were extracted from the detected systolic peaks. The full list of extracted watch features is included in Table S1 in Multimedia Appendix 1.

The data collected from the smartwatch were preprocessed with a developed PPG preprocessing pipeline [68]. The pipeline consisted of 3 stages: signal quality assessment (SQA), signal reconstruction, and PPG peak detection. The SQA stage involved the classification of PPG signals into 2 categories, namely “clean” and “noisy” [69]. After the SQA stage, short-term “noisy” segments, which lasted for less than 15 seconds, were reconstructed using a generative adversarial network model [70]. The generative adversarial network–based model consists of a generator and a discriminator. The generator learned the representative features of clean PPG signals, while the discriminator distinguished between generated and original signals. This adversarial process improved the model’s performance. The trained generator was used to reconstruct the distorted PPG signals. Subsequently, a trained dilated

XSL•FO
RenderX

convolution neural network was used to detect systolic peaks [71] and interbeat intervals. Finally, HR and HRV-related features were extracted from the detected interbeat intervals information.

The Oura ring feature engineering process involved 2 categories: daily features and nondaily features (the full list is provided in Table S2 in Multimedia Appendix 1). The daily features were provided directly from the Oura ring with 1 value per day. The nondaily features collected 1 value every 5 minutes. The daily slope, intercept, SD, and mean from the 5-minute measures were extracted to obtain a daily representation for these nondaily features. The Oura ring's PPG sensors captured the characterization of daily sleep and activity patterns.

Mobile Activity

The AWARE app passively collected biometric data and logged daily routines [72]. Collected data included movement (ie, steps and exercises), social relationship information (ie, amount of time spent with other people based on audio detection and proximity), daily rhythms (ie, routines such as going out vs staying in one's home based on location), and phone interactions (ie, texting, calling, and app browsing). The connection between the AWARE app and research servers were encrypted to ensure the privacy of participants' data. The AWARE feature engineering process involved the feature extraction from calls, messages, notifications, screen activities, and locations from the participants' smartphones. The full list of AWARE-related features is demonstrated in Table S3 in Multimedia Appendix 1.

The phone call–related features in AWARE were obtained from 4 types of call events: incoming calls, outgoing calls, missed calls, and voicemail. Within a specified time window of 5 days, the total duration and counts of each call event type were calculated. For the message and notification features in AWARE, 2 types of event notifications (ie, received messages and sent messages) were calculated; each event type was summed during the 5-day time window.

To obtain the participants' mobile phone use patterns, 4 types of screen use events were analyzed: screen on, screen off, screen lock, and screen unlock. From these features, screen activity use was calculated through the sum of these events over the preceding 5 days.

For location data, the latitude and longitude data of meaningful addresses from the Google Map application programming interface [73] were extracted in terms of variance of latitude, variance of speed, mean speed, number of places visited, home-stay duration, outdoor-stay duration, mean outdoor-stay duration, SD of outdoor-stay duration, the type of place with the longest duration (excluding home), and total travel distance over a 5-day window.

Data Analytic Plan

The overall data analytic process is illustrated in Figure 2. Given the study design, the data analysis was conducted separately for the control group, the savor group, and the combined dataset (control+savor). The data analysis process includes data processing, model selection, evaluation, and model transparency.

Figure 2. Data analytic flowchart. LightGBM: light gradient boosting machine; SHAP: Shapley Additive Explanations; SMOTE: synthetic minority oversampling technique.

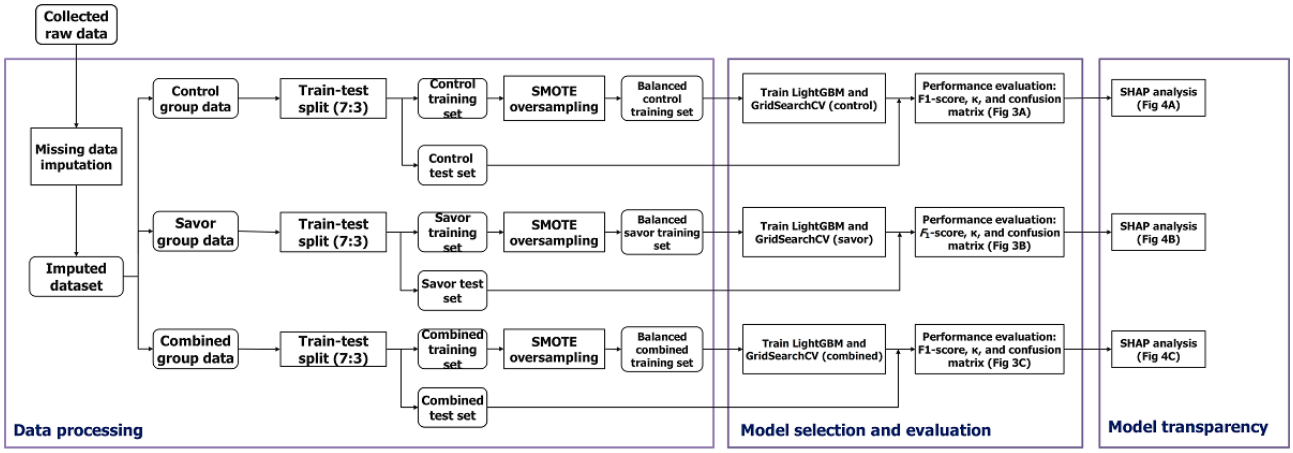

Data Preprocessing

Before building the ML model, several data preprocessing steps were applied to ensure data quality and suitability for model training. The aforementioned PPG preprocessing pipeline [64] was applied to reconstruct the poor-quality PPG signals with short duration and remove the unreliable PPG signals with long duration. The missing values of the dataset were then imputed with the nearest previous value of the time series. The synthetic minority oversampling technique (SMOTE) was used to preprocess the resulting imbalanced training dataset and will be described subsequently.

Missing Value Imputation

In the longitudinal monitoring study, multiple factors resulted in missing values, including human-related issues (ie, forgetting to charge the devices or removal of the wearables due to daily activities) and technical issues (ie, interruption of data collection due to server congestion and permission allowance on the phone). Such missing data were classified as missing completely

XSL•FO
RenderX

at random, as its occurrence was independent of the depression scale [74].

To quantify the extent of missing data, we computed the missing data percentages across participants. The overall missing rate varied across data sources, with smartphone-based passive sensing data showing an average missing rate of 16% and wearable-derived physiological data missing 11% of values.

Missing values were handled using the nearest-previous imputation method. This technique replaces the missing values with the nearest previous nonmissing value in the same feature column. We selected this method due to its ability to preserve temporal continuity in time series data, which is crucial for longitudinal analysis. Alternative imputation methods such as mean or median imputation were considered but deemed unsuitable as they might distort temporal structure of the data. Advanced ML-based imputations (eg, k-nearest neighbors or multiple imputation) were not applied to avoid introducing artificial patterns or biases in the data [75]. This method was applied to continuous physiological and smartphone use data, as these signals exhibit temporal dependencies where the previous observation is often the best estimate in short-term missing data scenarios. For ecological momentary assessment–related responses, missing values were not imputed to maintain the integrity of self-reported measures.

SMOTE Assessment

The dataset used for depression detection was imbalanced, with a larger proportion of nondepressed participants compared to depressed cases. To address this issue, the SMOTE was used to generate synthetic training samples for the minority class. The SMOTE technique created new synthetic samples by oversampling a random sample in the minority class and its k-nearest neighbors in the feature space [76]. In the original training set, there were 126 and 109 instances in “nonminimal” class and “follow-up needed” class, respectively. After performing the SMOTE technique, there were 126 samples in both the “nonminimal” class and the “follow-up needed” class.

Model

LightGBM was chosen as the ML algorithm to detect the individuals as depressed or nondepressed, considering the high dimensionality of the dataset comprising 104 features. LightGBM is a gradient boosting framework that uses tree-based learning algorithms to build predictive models. By using Exclusive Feature Bundling technique, LightGBM is best suited to effectively handle large numbers of features [39].

Performance Evaluation

The F_1-score is the harmonic mean of precision and recall, providing a balanced measure of a model’s performance [77]. The F_1-score is calculated where the precision is the proportion of true positives among the sum of true positives and false positives. The recall is the ratio of true positives to the sum of true positives and false negatives as defined in equation 1.

Cohen κ measures the agreement between the model’s predictions and the ground truth [78]. The variable p_0 is the observed agreement between the model’s predictions and the ground truth, while the variable p_e is the expected agreement by chance, which is the marginal probabilities of agreement for each class. This relationship is calculated through equation 2.

Results

Descriptive Statistics and Correlations

We conducted separate analyses for the control group, savor group, and the combined dataset (control+savor). For each group, the dataset was split into 70% training and 30% testing, with 3-fold cross-validation applied during training for hyperparameter tuning. The final test dataset contained 27 depressed and 23 nondepressed instances for the control group; 22 depressed and 30 nondepressed instances for the savor group, and 41 depressed and 60 nondepressed instances for the combined dataset. These instances were derived from 28 unique participants.

Aim 1: Overall Performance of Depressive Symptom Detection

The overall performance of the depression detection model with the observations from 28 participants (ie, combined group) is shown in Table 2. Given that this study was conducted within the context of an intervention study, we present our findings separately by group (control and intervention) as well as combined across groups. To evaluate model fit, we compared the detected labels (ie, the label predicted by the detection model) and the ground truth (ie, the depression severity reported on weekly PHQ-9 questionnaires). Our model achieved an F_1-score of 0.744 and a Cohen κ coefficient of 0.474. A higher F_1-score generally represents better detection capability, and our model achieved a moderate score, and the Cohen κ coefficient [79] indicated a moderate agreement between the predicted labels and the ground truth.

Table 2. Overall performance of the depressive symptom detection for control, intervention, and combined models.

Group	True positive	True negative	False positive	False negative	Accuracy	Precision	Recall	F_1-score	Cohen κ
Control	18	14	9	9	0.64	0.667	0.667	0.667	0.275
Intervention	14	23	7	8	0.712	0.667	0.636	0.651	0.405
Combined	30	45	15	11	0.743	0.667	0.732	0.698	0.474

XSL•FO
RenderX

The confusion matrix of the detection results is shown in Figure 3. Our model successfully detected the 45 out of 60 “none-minimal” participants (ie, those who scored between 0-4 on the PHQ-9) and 30 out of 41 “follow-up needed” participants (ie, those who scored between 5 and 9 on the PHQ-9). These results indicate that our model was able to accurately assess the depression level of the individuals based on their physiological and behavioral data, achieving an F_1-score of 0.70, a sensitivity of 0.73, a precision of 0.67, and an accuracy of 0.74.

Figure 3. Confusion matrix of depressive symptom detection model performance for (A) control, (B) savor (intervention), and (C) combined (ie, control and intervention) groups.

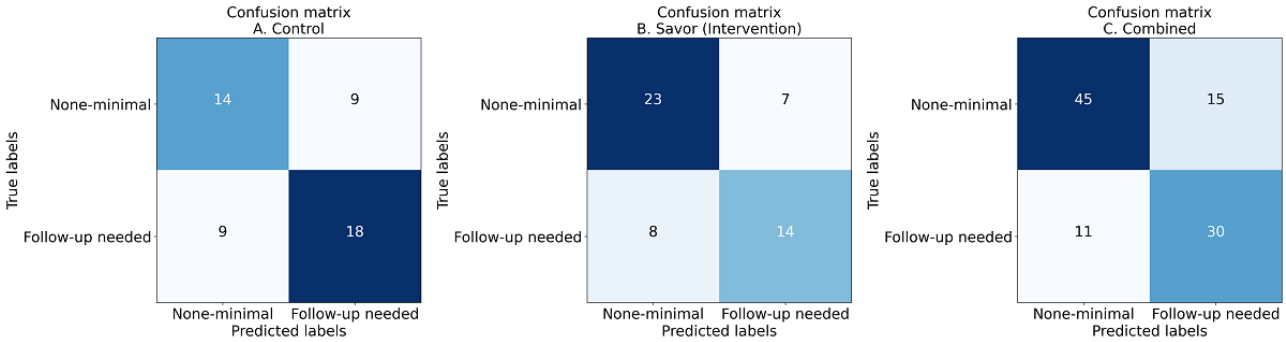

The full confusion matrices for the combined, control, and intervention groups are summarized in Table 2, and the detailed confusion matrices are provided in Figure S1 in Multimedia Appendix 2.

Aim 2: Detect Depressive Symptoms Across Control, Intervention, and Combined Participant Groups

In this study, we evaluated the efficacy of depression detection models across different participant groups, specifically the control group, the intervention group, and the combined dataset encompassing both groups. In total, 3 LightGBM models were trained separately for the control group, the intervention group, and the combined group. The performance metrics for these models were assessed using the F_1-score and Cohen κ. The control group model yielded an F_1-score of 0.640 and a Cohen κ of 0.275, indicating moderate detection performance and agreement. In contrast, the intervention group, which received targeted just-in-time interventions during the second month, demonstrated improved detection with an F_1-score of 0.711 and a Cohen κ of 0.405, reflecting better model accuracy and substantial agreement. The combined model, integrating both control and intervention group data, achieved the highest performance, with an F_1-score of 0.744 and a Cohen κ of 0.474. The most plausible explanation for these differential findings is that the combined model provided more sample size, enabling greater predictive capacity. Differences between the relational group and the control group may be explained by the relational intervention reducing variability among participants—in other words, all participants received an intervention, enabling greater predictability of symptoms.

Aim 3: Explainability and Feature Importance Analysis

To enhance the interpretability of our depression detection model, we used Shapley Additive Explanations (SHAP) to analyze feature importance and their effects on the model’s outputs [80]. SHAP values provide a unified measure of feature contribution by calculating the change in the model’s prediction when a feature is added or removed. This approach ensures a consistent and reliable determination of each feature’s impact. By calculating SHAP values, we can ascertain how each feature influences the prediction. The x-axis in our SHAP summary plots (Figure 4) represents the SHAP value, indicating the impact of each feature on the model’s output: positive values suggest a feature pushes the prediction toward depression, while negative values push it away. Meanwhile, the y-axis lists the feature values, with each point color-coded to show the actual feature value (ie, low to high).

XSL•FO
RenderX

Figure 4. Shapley Additive Explanations (SHAP)–based explainers of top 30 features for (A) control, (B) intervention, and (C) combined (ie, control and intervention) groups' depressive symptom detection. HR: heart rate; HRV: heart rate variability; HTI: heart rate triangular index; IQRNN: IQR of normal-to-normal intervals; LnHF: ratio of low frequency to high frequency; MCVNN: mean consecutive variation of normal-to-normal intervals; RMSSD: root mean square of successive differences; TINN: triangular interpolation of normal-to-normal histogram.

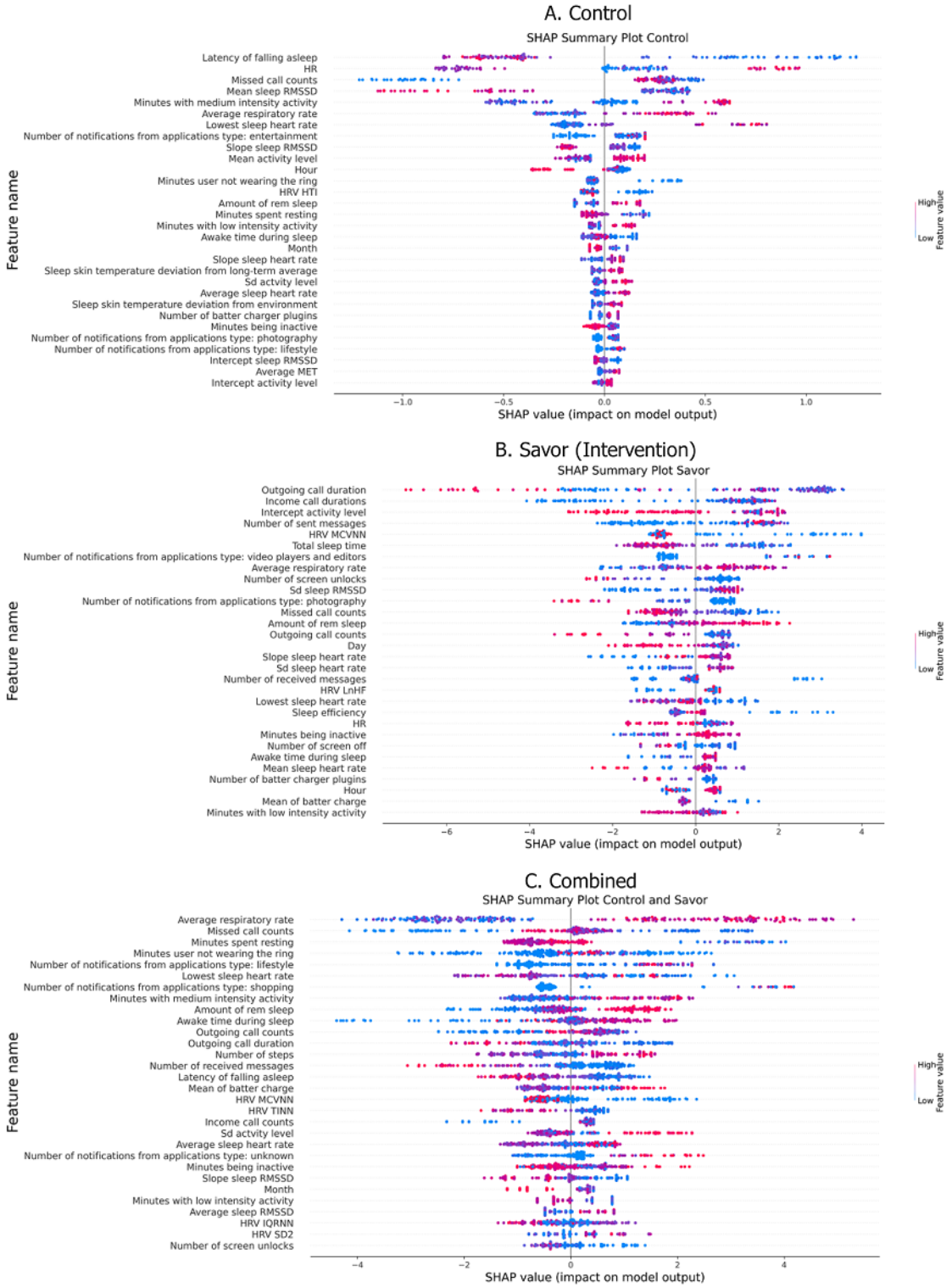

Figure 4 presents the top-30 features obtained by SHAP analysis for the control group, intervention group, and the combined dataset. For the control group, the 5 most important features for the detection were sleep latency, average HR, missed call counts, average sleep root mean square of successive differences, and minutes with medium activity intensity. Specifically, for the most predictive feature, sleep latency, lower values were associated with a higher likelihood of depression detection.

In the intervention group, the 5 key features driving the model's predictions were outgoing call duration, incoming call duration, the intercept of daily activity level regression (indicating an individual's initial activity level upon waking), number of sent

XSL•FO
RenderX

messages, and mean consecutive variation of normal-to-normal intervals. For the most predictive feature, outgoing call duration, both extreme high and low values predicted a lower likelihood of depression detection.

For the combined group, integrating data from both control and intervention participants, the most influential features were average sleep breathing rate, missed call counts, duration of resting minutes throughout the day, nonwear time of wearable devices, and the number of notifications from applications in the "lifestyle" category [81] (eg, Samsung Wallet/Pay and SmartThings). In this group, the most predictive feature, average sleep respiratory rate, demonstrated that higher values were associated with a greater likelihood of depressed detection.

Discussion

Main Findings

Our goal was to build upon previous work and investigate the utility of an ML approach for integrating mobile sensing data (ie, sleep, physiology, physical activity, and smartphone data) in the prediction of depressive symptoms as measured by weekly PHQ-9 questionnaires within a nonclinical sample of college students during the COVID-19 pandemic. Furthermore, we were interested in isolating the features that were most highly predictive of depressive symptoms to identify potential targets of prevention and intervention efforts. Our findings revealed that the ML model (LightGBM) was successful in differentiating participants in a low depressed and high depressed group, as indicated by 2 labels ("none-minimal" and "follow-up needed"), with a satisfactory F_1-score. Furthermore, certain features were most strongly predictive of membership in the high depressed group, such as the average sleep breathing rate. We discuss these findings subsequently.

First, the data derived from the passive sensing devices and LightGBM yielded sufficient information to predict participants' depression scores, measured weekly, with sufficient accuracy. The model in our study had an F_1 statistic of 0.74, and a Cohen κ of 0.474, which indicates moderate agreement between the prediction and actual observations adjusted for random chance [73]. In interpreting these findings, it is important to note that we predicted group membership using a dichotomous grouping ("none-minimal" and "follow-up needed"). Group membership was based on a clinical cutoff that is meaningful [82,83]; people with scores >4 (minimal depression or greater) are more likely to experience consistent symptoms of depression, influenced by both length of symptoms as well as severity [84]. We created these groupings to increase the predictive power and clinical significance of our findings in the context of a small sample study. However, as we dichotomized the variable, we invited the trade-off that the ML findings do not reflect the various degrees of depressive symptoms.

Second, our study revealed that features differed in the degree to which they predicted depressive symptoms. We analyzed the features separately within the control and the intervention groups. Within the control (monitoring only) group, the strongest predictors of depression risk were sleep latency, average HR, missed call counts, average sleep root mean square of successive differences, and minutes with medium activity intensity. Within the intervention (relational savoring) group, the stronger predictors were outgoing call duration, incoming call duration, the interception of daily activity level regression, number of sent messages and HRV mean consecutive variation of NN intervals. When the groups were combined, the strongest predictors were average sleep breathing rate, number of missed calls, duration of resting minutes throughout the day, nonwear time of wearable devices, and number of notifications from apps in the "lifestyle" category (eg, Samsung Wallet or Pay). Notably, of the 5 most highly predictive features, 2 pertained to sleep (ie, average sleep breathing rate and total amount of wake time recorded during the sleep period), 2 pertained to smartphone use (ie, on-wear time of wearable devices and number of missed calls), and 1 pertained to physical activity (ie, duration of resting minutes). In other words, all 3 areas of inquiry were represented in the top predictors. In general, they were associated with depression in theoretically predictable directions—falling asleep quickly, breathing faster during sleep, and making many and very few calls were associated with a greater likelihood of depression. The findings are consistent with previous studies showing that objective measures of sleep [16], smartphone use and misuse [85,86], and physical activity [24] are all associated with depressive symptoms in college students. However, our study extends the existing literature by using wearable devices in everyday settings under free-living conditions. Most studies have relied on controlled laboratory environments, which may not capture the nuanced, real-world interactions of physiological and behavioral factors. Moreover, the long-term and fine-grained quantitative measurements obtained from wearable and mobile devices allow for a responsive monitoring of depression symptoms. By revealing the contributing factors related to depression for both the control group and the savor group, our study highlights the differential impacts of these variables depending on the intervention status.

Our findings suggest that it may not be possible to use a single source of information, such as a smartphone or wearable device, to predict depressive symptoms on college campuses. This underscores the need for personalized assessment of mental health, making the argument that many different indicators are important in predicting mental health outcomes [87-92]. Although this is exciting from a research perspective, it reduces the clinical utility of this work at the current time. A top predictor of depressive symptoms in the combined group, as well as in each subgroup, was derived from the Oura ring, which may suggest that of the data analyzed in this study, these may be the most useful or important in predicting depressive states in this population.

Strengths and Limitations

The study is characterized by several strengths, including the use of a longitudinal design, a racially and ethnically diverse sample, and the incorporation of repeated depressive symptoms assessments. The inclusion of multiple passive sensing measures in the study is an additional strength that increases the generalizability of the findings and moves beyond previous studies on this topic. The fine-grained approach to understanding people's experiences and behaviors helps to address the heterogeneity of symptomatology and depression to allow for

XSL•FO
RenderX

better assessment of depressive symptoms. Furthermore, the study used a nonnested cross-validation procedure, which enhances the robustness and generalizability of the findings by making efficient use of the available data and reducing overfitting. Finally, participants in this study had high rates of adherence to study protocols—although some participants dropped out at the outset of the study, participants who continued with the study provided consistent data streams, increasing confidence in the conclusions we could draw from the study.

There are several limitations worth considering as well. This study was conducted with a community sample of college students rather than a clinical sample. Our goal was preventive in nature—to be able to detect depression risk before full-blown depression has manifested. On the basis of the assumption that subclinical levels of depressive symptoms confer risk for full-blown clinical depressive syndromes, we use depressive symptoms as a proxy here for risk for future depression. We urge readers to keep this in mind when interpreting these findings. To test whether the findings of this algorithm have predictive validity for the prediction of clinical levels of depression, a follow-up study would need to be conducted in which the study uses a clinical sample of participants who met criteria for major depressive disorder. The study would also benefit from the inclusion of diagnostic interviews at baseline and follow-up to assess for the presence of depressive disorders (rather than depressive symptoms).

Furthermore, the sample size was small, which considerably limits the generalizability of the findings. This study was conducted using a pilot study grant; given the technology, equipment, and staffing involved in this study, more funding would be required to support a study with a larger sample. The findings of the study indicate the importance of a personalized approach to assessment, which argues against reducing the number of data points, at least initially. However, it may be possible to increase the sample and reduce the cost by minimizing the measurement window. The creation of data repositories across research teams could facilitate larger-scale investigations with fewer resources.

An additional limitation involves our use of college students as participants. College students are an important population to target due to their high risk for mental health issues such as depression. However, their familiarity and comfort with technology and mobile devices distinguish them from other groups [93]. This restricts the generalizability of the findings. To have a better sense of the feasibility of this study's design, it would be important to repeat this study among young adults who are not in college, as well as with older adult samples (eg, middle-aged and older adults) less familiar with technology.

Furthermore, despite being longitudinal, the study design is correlational, limiting our ability to make causal inferences. Our study used features collected 5 days before the PHQ-9 assessment, supporting the notion that passive sensing features indicate subsequent depression. However, it is possible that depressive symptoms in the previous week may influence the passive sensing indicators observed in the current week; therefore, confounding the direction of this relationship. As a result, the key indicators identified in this study cannot necessarily be considered predictors of depression. While separating the cause and effect in this relationship is extremely difficult, given the reciprocal and mutually reinforcing links between these indicators and depressive symptoms, identifying predictors of depressive symptoms is essential to pinpoint targets for prevention efforts. Furthermore, the current model did not consider interrelations between different contributing factors, for instance, sleep may influence exercise and exercise may influence sleep, both of which can in turn influence depression. In other words, our model assumed that each of these factors has independent rather than intersecting influences on depression, when in reality the interrelation between these factors may be much more complex and reciprocal. Furthermore, while SHAP-based explanations enhanced our understanding of the model's internal decision process, these should be interpreted with caution. SHAP values reflect the contribution of each feature to the prediction within the model, and do not imply a causal relationship with depression symptoms. The SHAP framework is model-dependent and can be influenced by feature correlations, sampling variability, and model complexity. Future studies using larger and more diverse samples, along with causal inference techniques, are needed to validate the observed patterns. In this study, we use SHAP to enhance transparency and to generate hypotheses, rather than to make strong causal claims. In addition, analyses did not control for covariates that could be relevant to the questions of interest, such as BMI, socioeconomic status (which could be related to the type of mobile device used, namely Android versus Apple), travel (sleep routine), or psychological comorbidities, such as anxiety, which is highly related to depression in college students [94].

Given the difficulty in assessing depressed mood states, as well as the degree to which college students increasingly engage in life in an internet-based environment [95,96], the ability to predict mental health from passive sensing devices is of great utility. This investigation builds upon other studies that have adopted a similar approach of using mobile devices to collect real-world data in the service of predicting important outcomes [55,80,97], such as mental health. Although the current protocol as it was conducted does not have a high potential for widespread accessibility in that it involved multiple devices, multiple staff monitoring data collection, a multi-person computer science team to analyze the data, and a licensed psychologist to assess and provide mental health resources when participants reported elevated mental health symptoms, the ultimate vision for this work involves increasing access to care. Over the long term, we envision a scenario in which remote monitoring could reduce the need for mental health interventions or facilitate timely interventions by a mental health professional, thereby reducing costs and improving quality of care [98].

Conclusions

This study offered an exciting glimpse into the potential of ML for predicting depressive symptoms among college students. Limitations notwithstanding, our study suggested the promise of multimodal passive sensing and ML to identify and respond to episodes of depressed mood. With the high global disease burden of depression and the widespread adoption of mobile

XSL•FO
RenderX

and wearable devices, we hope this work can serve as a foundation for future studies that refine, streamline, and personalize our sensing and prediction paradigm, and that explore the utility of similar approaches with regard to other disorders in addition to depression. Ultimately, we hope this work culminates in the development of a monitoring system that can be implemented at scale. Such a system would offer an efficient, data-driven solution to mental health care and could improve depressive outcomes by helping connect individuals in need of mental health resources in a timely fashion, empowering them to proactively respond to challenges and maintain well-being.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Full list of features.
[DOCX File , 62 KB - formative_v9i1e67964_app1.docx]

Multimedia Appendix 2
Confusion matrix for model performance.
[DOCX File , 55 KB - formative_v9i1e67964_app2.docx]

References

1. Friedrich MJ. Depression is the leading cause of disability around the world. JAMA 2017 Apr 18;317(15):1517. [doi: 10.1001/jama.2017.3826] [Medline: 28418490]
2. Olson JS, Crosnoe R. Are you still bringing me down?: romantic involvement and depressive symptoms from adolescence to young adulthood. J Health Soc Behav 2017 Mar;58(1):102-115 [FREE Full text] [doi: 10.1177/0022146516684536] [Medline: 28661766]
3. Haikalis M, Doucette H, Meisel MK, Birch K, Barnett NP. Changes in college student anxiety and depression from pre- to during-COVID-19: perceived stress, academic challenges, loneliness, and positive perceptions. Emerg Adulthood 2022 Apr;10(2):534-545 [FREE Full text] [doi: 10.1177/21676968211058516] [Medline: 35382515]
4. COVID-19 Mental Disorders Collaborators. Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. Lancet 2021 Nov 06;398(10312):1700-1712 [FREE Full text] [doi: 10.1016/S0140-6736(21)02143-7] [Medline: 34634250]
5. Wang X, Hegde S, Son C, Keller B, Smith A, Sasangohar F. Investigating mental health of US college students during the COVID-19 pandemic: cross-sectional survey study. J Med Internet Res 2020 Sep 17;22(9):e22817 [FREE Full text] [doi: 10.2196/22817] [Medline: 32897868]
6. Alzueta E, Podhajsky S, Zhao Q, Tapert SF, Thompson WK, de Zambotti M, et al. Risk for depression tripled during the COVID-19 pandemic in emerging adults followed for the last 8 years. Psychol Med 2023 Apr;53(5):2156-2163 [FREE Full text] [doi: 10.1017/S0033291721004062] [Medline: 34726149]
7. Fermani A, Bongelli R, Canestrari C, Muzi M, Riccioni I, Burro R. "Old wine in a new bottle". Depression and romantic relationships in Italian emerging adulthood: the moderating effect of gender. Int J Environ Res Public Health 2020 Jun 09;17(11):4121 [FREE Full text] [doi: 10.3390/ijerph17114121] [Medline: 32527021]
8. Hodgetts S, Gallagher P, Stow D, Ferrier IN, O'Brien JT. The impact and measurement of social dysfunction in late-life depression: an evaluation of current methods with a focus on wearable technology. Int J Geriatr Psychiatry 2017 Mar;32(3):247-255. [doi: 10.1002/gps.4632] [Medline: 27911019]
9. Ayuso-Mateos JL, Nuevo R, Verdes E, Naidoo N, Chatterji S. From depressive symptoms to depressive disorders: the relevance of thresholds. Br J Psychiatry 2010 May;196(5):365-371. [doi: 10.1192/bjp.bp.109.071191] [Medline: 20435961]
10. Cicchetti D, Rogosch FA. A developmental psychopathology perspective on adolescence. J Consult Clin Psychol 2002 Feb;70(1):6-20. [doi: 10.1037//0022-006x.70.1.6] [Medline: 11860057]
11. Jeon SW, Amidfar M, Kim YK. Bio-psycho-social risk factors for depression. In: Kim YK, editor. Major Depressive Disorder: Risk Factors, Characteristics and Treatment Options. Hauppauge, NY: Nova Science; 2017:71.
12. Elhai JD, Levine JC, Dvorak RD, Hall BJ. Non-social features of smartphone use are most related to depression, anxiety and problematic smartphone use. Comput Human Behav 2017 Apr;69:75-82. [doi: 10.1016/j.chb.2016.12.023]
13. Brooks PR, Girgenti AA, Mills MJ. Sleep patterns and symptoms of depression in college students. Coll Stud J 2009;43(2):464-473 [FREE Full text]
14. Deng J, Zhou F, Hou W, Silver Z, Wong CY, Chang O, et al. The prevalence of depressive symptoms, anxiety symptoms and sleep disturbance in higher education students during the COVID-19 pandemic: a systematic review and meta-analysis. Psychiatry Res 2021 Jul;301:113863. [doi: 10.1016/j.psychres.2021.113863] [Medline: 33984824]

XSL•FO
RenderX

15. Wang R, Wang W, Dasilva A, Huckins JF, Kelley WM, Heatherton TF, et al. Tracking depression dynamics in college students using mobile phone and wearable sensing. Proc ACM Interact Mob Wearable Ubiquitous Technol 2018 Mar;2(1):1-26. [doi: 10.1145/3191775] [Medline: 39449996]
16. Dinis J, Bragança M. Quality of sleep and depression in college students: a systematic review. Sleep Sci 2018;11(4):290-301 [FREE Full text] [doi: 10.5935/1984-0063.20180045] [Medline: 30746048]
17. Nutt D, Wilson S, Paterson L. Sleep disorders as core symptoms of depression. Dialogues Clin Neurosci 2008;10(3):329-336 [FREE Full text] [doi: 10.31887/DCNS.2008.10.3/dnutt] [Medline: 18979946]
18. Regestein Q, Natarajan V, Pavlova M, Kawasaki S, Gleason R, Koff E. Sleep debt and depression in female college students. Psychiatry Res 2010 Mar 30;176(1):34-39. [doi: 10.1016/j.psychres.2008.11.006] [Medline: 20079935]
19. Carter JR, Gervais BM, Adomeit JL, Greenlund IM. Subjective and objective sleep differ in male and female collegiate athletes. Sleep Health 2020 Oct;6(5):623-628. [doi: 10.1016/j.sleh.2020.01.016] [Medline: 32147360]
20. Brooks DC, Pomerantz J. ECAR study of undergraduate students and information technology, 2019. Educause Center for Analysis and Research. 2017. URL: https://library.educause.edu/-/media/files/library/2019/10/studentstudy2019.pdf?la=en&hash=25FBB396AE482FAC3B765862BA6B197DBC98B42C [accessed 2024-05-29]
21. Shi X, Wang A, Zhu Y. Longitudinal associations among smartphone addiction, loneliness, and depressive symptoms in college students: Disentangling between- And within-person associations. Addict Behav 2023 Jul;142:107676. [doi: 10.1016/j.addbeh.2023.107676] [Medline: 36878182]
22. Zhang HX. Smartphone addiction among university students' during the post-COVID-19 era: the role of emotional intelligence and future anxiety. Psychiatry Investig 2023 Oct;20(10):951-961 [FREE Full text] [doi: 10.30773/pi.2023.0021] [Medline: 37899219]
23. Park N, Lee H. Social implications of smartphone use: Korean college students' smartphone use and psychological well-being. Cyberpsychol Behav Soc Netw 2012 Sep;15(9):491-497. [doi: 10.1089/cyber.2011.0580] [Medline: 22817650]
24. Elhai JD, Tiamiyu MF, Weeks JW, Levine JC, Picard KJ, Hall BJ. Depression and emotion regulation predict objective smartphone use measured over one week. Pers Individ Dif 2018 Oct;133:21-28 [FREE Full text] [doi: 10.1016/j.paid.2017.04.051]
25. Tyson P, Wilson K, Crone D, Brailsford R, Laws K. Physical activity and mental health in a student population. J Ment Health 2010 Dec;19(6):492-499. [doi: 10.3109/09638230902968308] [Medline: 20812852]
26. Vankim NA, Nelson TF. Vigorous physical activity, mental health, perceived stress, and socializing among college students. Am J Health Promot 2013;28(1):7-15 [FREE Full text] [doi: 10.4278/ajhp.111101-QUAN-395] [Medline: 23470187]
27. Ghrouz AK, Noohu MM, Dilshad Manzar M, Warren Spence D, BaHammam AS, Pandi-Perumal SR. Physical activity and sleep quality in relation to mental health among college students. Sleep Breath 2019 Jun;23(2):627-634. [doi: 10.1007/s11325-019-01780-z] [Medline: 30685851]
28. Kimm SY, Glynn NW, Kriska AM, Barton BA, Kronsberg SS, Daniels SR, et al. Decline in physical activity in black girls and white girls during adolescence. N Engl J Med 2002 Sep 05;347(10):709-715. [doi: 10.1056/NEJMoa003277] [Medline: 12213941]
29. Nelson TF, Gortmaker SL, Subramanian SV, Wechsler H. Vigorous physical activity among college students in the United States. J Phys Act Health 2007 Oct;4(4):495-508. [Medline: 18209239]
30. Troiano RP, Berrigan D, Dodd KW, Mâsse LC, Tilert T, McDowell M. Physical activity in the United States measured by accelerometer. Med Sci Sports Exerc 2008 Jan;40(1):181-188. [doi: 10.1249/mss.0b013e31815a51b3] [Medline: 18091006]
31. Trost SG, Pate RR, Ward DS, Saunders R, Riner W. Correlates of objectively measured physical activity in preadolescent youth. Am J Prev Med 1999 Aug;17(2):120-126. [doi: 10.1016/s0749-3797(99)00056-2] [Medline: 10490054]
32. Butler SM, Black DR, Blue CL, Gretebeck RJ. Change in diet, physical activity, and body weight in female college freshman. Am J Health Behav 2004;28(1):24-32. [doi: 10.5993/ajhb.28.1.3] [Medline: 14977156]
33. Mokdad AH, Marks JS, Stroup DF, Gerberding JL. Actual causes of death in the United States, 2000. JAMA 2004 Mar 10;291(10):1238-1245. [doi: 10.1001/jama.291.10.1238] [Medline: 15010446]
34. Farren GL, Zhang T, Martin SB, Thomas KT. Factors related to meeting physical activity guidelines in active college students: a social cognitive perspective. J Am Coll Health 2017 Jan;65(1):10-21. [doi: 10.1080/07448481.2016.1229320] [Medline: 27593500]
35. Dziewior J, Carr LJ, Pierce GL, Whitaker K. College students report less physical activity and more sedentary behavior during the COVID-19 pandemic. J Am Coll Health 2024 Oct;72(7):2022-2030. [doi: 10.1080/07448481.2022.2100708] [Medline: 35881778]
36. Romero-Blanco C, Rodríguez-Almagro J, Onieva-Zafra MD, Parra-Fernández ML, Prado-Laguna MD, Hernández-Martínez A. Physical activity and sedentary lifestyle in university students: changes during confinement due to the COVID-19 pandemic. Int J Environ Res Public Health 2020 Sep 09;17(18):6567 [FREE Full text] [doi: 10.3390/ijerph17186567] [Medline: 32916972]
37. Gianfredi V, Blandi L, Cacitti S, Minelli M, Signorelli C, Amerio A, et al. Depression and objectively measured physical activity: a systematic review and meta-analysis. Int J Environ Res Public Health 2020 May 25;17(10):3738 [FREE Full text] [doi: 10.3390/ijerph17103738] [Medline: 32466242]

XSL•FO
RenderX

38. Zhang X, Gu X. Adherence to the 24-hour movement behavior guidelines and associations with depressive symptoms among college students. Int J Kinesiol High Educ 2021 May 28;6(4):225-237 [FREE Full text] [doi: 10.1080/24711616.2021.1921637]
39. Hills AP, Mokhtar N, Byrne NM. Assessment of physical activity and energy expenditure: an overview of objective measures. Front Nutr 2014;1:5 [FREE Full text] [doi: 10.3389/fnut.2014.00005] [Medline: 25988109]
40. Tao K, Liu W, Xiong S, Ken L, Zeng N, Peng Q, et al. Associations between self-determined motivation, accelerometer-determined physical activity, and quality of life in Chinese college students. Int J Environ Res Public Health 2019 Aug 16;16(16):2941 [FREE Full text] [doi: 10.3390/ijerph16162941] [Medline: 31426280]
41. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: a highly efficient gradient boosting decision tree. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017 Presented at: NIPS '17; December 4-9, 2017; Long Beach, CA p. 3149-3157 URL: https://dl.acm.org/doi/10.5555/3294996.3295074
42. Aleem S, Huda NU, Amin R, Khalid S, Alshamrani SS, Alshehri A. Machine learning algorithms for depression: diagnosis, insights, and research directions. Electronics 2022 Mar 31;11(7):1111 [FREE Full text] [doi: 10.3390/electronics11071111]
43. Joshi ML, Kanoongo N. Depression detection using emotional artificial intelligence and machine learning: a closer review. Mater Today Proc 2022;58:217-226 [FREE Full text] [doi: 10.1016/j.matpr.2022.01.467]
44. Mao K, Wu Y, Chen J. A systematic review on automated clinical depression diagnosis. Npj Ment Health Res 2023 Nov 20;2(1):20 [FREE Full text] [doi: 10.1038/s44184-023-00040-z] [Medline: 38609509]
45. Cummins N, Sethu V, Epps J, Schnieder S, Krajewski J. Analysis of acoustic space variability in speech affected by depression. Speech Commun 2015 Dec;75:27-49 [FREE Full text] [doi: 10.1016/j.specom.2015.09.003]
46. Cunningham S, Hudson CC, Harkness K. Social media and depression symptoms: a meta-analysis. Res Child Adolesc Psychopathol 2021 Feb;49(2):241-253. [doi: 10.1007/s10802-020-00715-7] [Medline: 33404948]
47. Richter T, Fishbain B, Richter-Levin G, Okon-Singer H. Machine learning-based behavioral diagnostic tools for depression: advances, challenges, and future directions. J Pers Med 2021 Sep 26;11(10):957 [FREE Full text] [doi: 10.3390/jpm11100957] [Medline: 34683098]
48. Ben-Zeev D, Scherer EA, Wang R, Xie H, Campbell AT. Next-generation psychiatric assessment: using smartphone sensors to monitor behavior and mental health. Psychiatr Rehabil J 2015 Sep;38(3):218-226 [FREE Full text] [doi: 10.1037/prj0000130] [Medline: 25844912]
49. Callan JA, Wright J, Siegle GJ, Howland RH, Kepler BB. Use of computer and mobile technologies in the treatment of depression. Arch Psychiatr Nurs 2017 Jun;31(3):311-318. [doi: 10.1016/j.apnu.2016.10.002] [Medline: 28499574]
50. Chan LL, Brodie MA, Lord SR. Prediction of incident depression in middle-aged and older adults using digital gait biomarkers extracted from large-scale wrist sensor data. J Am Med Dir Assoc 2023 Aug;24(8):1106-13.e11. [doi: 10.1016/j.jamda.2023.04.008] [Medline: 37236263]
51. Kumar P, Chauhan R, Stephan T, Shankar A, Thakur S. A machine learning implementation for mental health care. Application: smart watch for depression detection. In: Proceedings of the 11th International Conference on Cloud Computing, Data Science & Engineering. 2021 Presented at: Confluence '21; January 28-29, 2021; Noida, India p. 568-574 URL: https://ieeexplore.ieee.org/document/9377199 [doi: 10.1109/confluence51648.2021.9377199]
52. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
53. Lee Y, Ragguett RM, Mansur RB, Boutilier JJ, Rosenblat JD, Trevizol A, et al. Applications of machine learning algorithms to predict therapeutic outcomes in depression: a meta-analysis and systematic review. J Affect Disord 2018 Dec 01;241:519-532. [doi: 10.1016/j.jad.2018.08.073] [Medline: 30153635]
54. Narayanrao PV, Lalitha Surya Kumari P. Analysis of machine learning algorithms for predicting depression. In: Proceedings of the 2020 International Conference on Computer Science, Engineering and Applications. 2020 Presented at: ICCSEA '20; March 13-14, 2020; Gunupur, India p. 1-4 URL: https://ieeexplore.ieee.org/document/9132963 [doi: 10.1109/iccsea49143.2020.9132963]
55. Borelli JL. Relational Savoring: Using Guided Reflection to Strengthen Relationships and Improve Outcomes in Therapy. Washington, DC: American Psychological Association; 2024.
56. Borelli JL, Smiley PA, Kerr ML, Hong K, Hecht HK, Blackard MB, et al. Relational savoring: an attachment-based approach to promoting interpersonal flourishing. Psychotherapy (Chic) 2020 Sep;57(3):340-351. [doi: 10.1037/pst0000284] [Medline: 31999188]
57. Adrien J. Neurobiological bases for the relation between sleep and depression. Sleep Med Rev 2002 Oct;6(5):341-351. [Medline: 12531125]
58. Niu X, Snyder HR. The role of maladaptive emotion regulation in the bidirectional relation between sleep and depression in college students during the COVID-19 pandemic. Anxiety Stress Coping 2023 Jan;36(1):83-96 [FREE Full text] [doi: 10.1080/10615806.2022.2073441] [Medline: 35536737]
59. Coyne SM, Stockdale L, Summers K. Problematic cell phone use, depression, anxiety, and self-regulation: evidence from a three year longitudinal study from adolescence to emerging adulthood. Comput Human Behav 2019 Jul;96:78-84 [FREE Full text] [doi: 10.1016/j.chb.2019.02.014]

XSL•FO
RenderX

60. Aubourg T, Demongeot J, Renard F, Provost H, Vuillerme N. Association between social asymmetry and depression in older adults: a phone Call Detail Records analysis. Sci Rep 2019 Sep 18;9(1):13524 [FREE Full text] [doi: 10.1038/s41598-019-49723-8] [Medline: 31534178]
61. Muñoz RF, Leykin Y, Barrera AZ, Brown CH, Bunge EL. The impact of phone calls on follow-up rates in an online depression prevention study. Internet Interv 2017 Jun;8:10-14 [FREE Full text] [doi: 10.1016/j.invent.2017.02.001] [Medline: 30135824]
62. Thomée S, Härenstam A, Hagberg M. Mobile phone use and stress, sleep disturbances, and symptoms of depression among young adults--a prospective cohort study. BMC Public Health 2011 Jan 31;11(1):66 [FREE Full text] [doi: 10.1186/1471-2458-11-66] [Medline: 21281471]
63. Keum BT, Miller MJ, Inkelas KK. Testing the factor structure and measurement invariance of the PHQ-9 across racially diverse U.S. college students. Psychol Assess 2018 Aug;30(8):1096-1106. [doi: 10.1037/pas0000550] [Medline: 29565614]
64. Nickels S, Edwards MD, Poole SF, Winter D, Gronsbell J, Rozenkrants B, et al. Toward a mobile platform for real-world digital measurement of depression: user-centered design, data quality, and behavioral and clinical modeling. JMIR Ment Health 2021 Aug 10;8(8):e27589 [FREE Full text] [doi: 10.2196/27589] [Medline: 34383685]
65. Torous J, Staples P, Shanahan M, Lin C, Peck P, Keshavan M, et al. Utilizing a personal smartphone custom app to assess the Patient Health Questionnaire-9 (PHQ-9) depressive symptoms in patients with major depressive disorder. JMIR Ment Health 2015;2(1):e8 [FREE Full text] [doi: 10.2196/mental.3889] [Medline: 26543914]
66. Oura ring (generation 2). Oura Health Ltd. 2018. URL: https://ouraring.com [accessed 2024-05-29]
67. Samsung Gear sport smartwatch. Samsung Electronics. URL: https://www.samsung.com/us/mobile/wearables/smartwatches/gear-sport-blue-sm-r600nzbaxar/ [accessed 2024-05-29]
68. Feli M, Kazemi K, Azimi I, Wang Y, Rahmani AM, Liljeberg P. End-to-end PPG processing pipeline for wearables: from quality assessment and motion artifacts removal to HR/HRV feature extraction. In: Proceedings of the 2023 IEEE International Conference on Bioinformatics and Biomedicine. 2023 Presented at: BIBM '23; December 5-8, 2023; Istanbul, Turkiye p. 1895-1900 URL: https://ieeexplore.ieee.org/document/10385998 [doi: 10.1109/bibm58861.2023.10385998]
69. Feli M, Azimi I, Anzanpour A, Rahmani AM, Liljeberg P. An energy-efficient semi-supervised approach for on-device photoplethysmogram signal quality assessment. Smart Health 2023 Jun;28:100390 [FREE Full text] [doi: 10.1016/j.smhl.2023.100390]
70. Wang Y, Azimi I, Kazemi K, Rahmani AM, Liljeberg P. PPG signal reconstruction using deep convolutional generative adversarial network. In: Proceedings of the 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society. 2022 Presented at: EMBC '22; July 11-15, 2022; Glasgow, UK p. 3387-3391 URL: https://ieeexplore.ieee.org/document/9871678 [doi: 10.1109/embc48229.2022.9871678]
71. Kazemi K, Laitala J, Azimi I, Liljeberg P, Rahmani AM. Robust PPG peak detection using dilated convolutional neural networks. Sensors (Basel) 2022 Aug 13;22(16):6054 [FREE Full text] [doi: 10.3390/s22166054] [Medline: 36015816]
72. Ferreira D, Kostakos V, Dey A. AWARE: mobile context instrumentation framework. Front ICT 2015 Apr 20;2:e43296 [FREE Full text] [doi: 10.3389/fict.2015.00006]
73. Svennerberg G. Beginning Google Maps API 3. Cham, Switzerland: Springer; 2010.
74. Mack C, Su Z, Westreich D. Managing Missing Data in Patient Registries: Addendum to Registries for Evaluating Patient Outcomes: A User's Guide. Rockville, MD: Agency for Healthcare Research and Quality (US); 2018.
75. Kaiser J. Dealing with missing values in data. J Syst Integr 2014;5(1):42-51. [doi: 10.20470/jsi.v5i1.178]
76. Chawla NV, Bowyer KW, Hall L, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002;16:321-357. [doi: 10.1613/jair.953]
77. Sokolova M, Japkowicz N, Szpakowicz S. Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. In: Proceedings of the 19th Australian Joint Conference on Artificial Intelligence. 2006 Presented at: AI '06; December 4-8, 2006; Hobart, Australia p. 1015-1021 URL: https://link.springer.com/chapter/10.1007/11941439_114 [doi: 10.1007/11941439_114]
78. Vieira SM, Kaymak U, Sousa JM. Cohen's kappa coefficient as a performance measure for feature selection. In: proceedings of the 2010 International Conference on Fuzzy Systems. 2010 Presented at: FUZZY '10; July 18-23, 2010; Barcelona, Spain p. 1-8 URL: https://ieeexplore.ieee.org/document/5584447 [doi: 10.1109/fuzzy.2010.5584447]
79. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960 Apr 01;20(1):37-46 [FREE Full text] [doi: 10.1177/001316446002000104]
80. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. arXiv. Preprint posted online May 22, 2017 2017 [FREE Full text]
81. Android Apps on Google Play. Google. URL: https://play.google.com/web/store/apps?hl=en_US [accessed 2024-06-19]
82. Kocalevent RD, Hinz A, Brähler E. Standardization of the depression screener patient health questionnaire (PHQ-9) in the general population. Gen Hosp Psychiatry 2013 Jun 15;35(5):551-555 [FREE Full text] [doi: 10.1016/j.genhosppsych.2013.04.006] [Medline: 23664569]
83. McMillan D, Gilbody S, Richards D. Defining successful treatment outcome in depression using the PHQ-9: a comparison of methods. J Affect Disord 2010 Dec;127(1-3):122-129 [FREE Full text] [doi: 10.1016/j.jad.2010.04.030] [Medline: 20569992]

XSL•FO
RenderX

84. Kroenke K, Spitzer RL. The PHQ-9: a new depression diagnostic and severity measure. Psychiatric Annals 2002 Sep;32(9):509-515. [doi: 10.3928/0048-5713-20020901-06]
85. Rosenthal SR, Zhou J, Booth ST. Association between mobile phone screen time and depressive symptoms among college students: A threshold effect. Hum Behav Emerg Technol 2021 Feb 18;3(3):432-440 [FREE Full text] [doi: 10.1002/hbe2.256]
86. Zhang K, Guo H, Wang T, Zhang J, Yuan G, Ren J, et al. A bidirectional association between smartphone addiction and depression among college students: a cross-lagged panel model. Front Public Health 2023;11:1083856 [FREE Full text] [doi: 10.3389/fpubh.2023.1083856] [Medline: 36761134]
87. Jafarlou S, Lai J, Azimi I, Mousavi Z, Labbaf S, Jain RC, et al. Objective prediction of next-day's affect using multimodal physiological and behavioral data: algorithm development and validation study. JMIR Form Res 2023 Mar 15;7:e39425 [FREE Full text] [doi: 10.2196/39425] [Medline: 36920456]
88. Jafarlou S, Azimi I, Lai J, Wang Y, Labbaf S, Nguyen B, et al. Objective monitoring of loneliness levels using smart devices: a multi-device approach for mental health applications. PLoS One 2024 Mar 15;19(6):e0298949 [FREE Full text] [doi: 10.1371/journal.pone.0298949] [Medline: 38900745]
89. Labbaf S, Abbasian M, Nguyen B, Lucero M, Ahmed MS, Yunusova A, et al. Physiological and emotional assessment of college students using wearable and mobile devices during the 2020 COVID-19 lockdown: an intensive, longitudinal dataset. Data Brief 2024 Jun;54:110228 [FREE Full text] [doi: 10.1016/j.dib.2024.110228] [Medline: 39669644]
90. Lai J, Rahmani A, Yunusova A, Rivera AP, Labbaf S, Hu S, et al. Using multimodal assessments to capture personalized contexts of college student well-being in 2020: case study. JMIR Form Res 2021 May 11;5(5):e26186 [FREE Full text] [doi: 10.2196/26186] [Medline: 33882022]
91. Yang Z, Wang Y, Yamashita KS, Khatibi E, Azimi I, Dutt N, et al. Integrating wearable sensor data and self-reported diaries for personalized affect forecasting. Smart Health 2024 Jun;32:100464 [FREE Full text] [doi: 10.1016/j.smhl.2024.100464]
92. Yunusova A, Lai J, Rivera AP, Hu S, Labbaf S, Rahmani AM, et al. Assessing the mental health of emerging adults through a mental health app: protocol for a prospective pilot study. JMIR Res Protoc 2021 Mar 02;10(3):e25775 [FREE Full text] [doi: 10.2196/25775] [Medline: 33513124]
93. Jones S, Madden MA, Clarke LN. The internet goes to college: how students are living in the future with today?s technology. Pew Research Center. 2002. URL: https://www.pewresearch.org/internet/2011/07/19/college-students-and-technology/ [accessed 2024-05-29]
94. Li W, Zhao Z, Chen D, Peng Y, Lu Z. Prevalence and associated factors of depression and anxiety symptoms among college students: a systematic review and meta-analysis. J Child Psychol Psychiatry 2022 Nov;63(11):1222-1230. [doi: 10.1111/jcpp.13606] [Medline: 35297041]
95. Davis L, Sun Q, Lone T, Levi A, Xu P. In the storm of COVID-19: college students' perceived challenges with virtual learning. J High Educ Theory Pract 2022 Feb 9;22(1):66-82 [FREE Full text] [doi: 10.33423/jhetp.v22i1.4964]
96. Sage K, Jackson S, Fox E, Mauer L. The virtual COVID-19 classroom: surveying outcomes, individual differences, and technology use in college students. Smart Learn Environ 2021 Nov 01;8(1):9 [FREE Full text] [doi: 10.1186/s40561-021-00174-7]
97. Moshe I, Terhorst Y, Opoku Asare K, Sander LB, Ferreira D, Baumeister H, et al. Predicting symptoms of depression and anxiety using smartphone and wearable data. Front Psychiatry 2021;12:625247 [FREE Full text] [doi: 10.3389/fpsyt.2021.625247] [Medline: 33584388]
98. Rahmani AM, Lai J, Jafarlou S, Azimi I, Yunusova A, Rivera AP, et al. Personal mental health navigator: harnessing the power of data, personal models, and health cybernetics to promote psychological well-being. Front Digit Health 2022;4:933587 [FREE Full text] [doi: 10.3389/fdgth.2022.933587] [Medline: 36213523]

Abbreviations

HR: heart rate
HRV: heart rate variability
LightGBM: light gradient boosting machine
ML: machine learning
OIT: Office of Information Technology
PHQ-9: Patient Health Questionnaire-9
PPG: photoplethysmogram
SHAP: Shapley Additive Explanations
SMOTE: synthetic minority oversampling technique
SQA: signal quality assessment

XSL•FO
RenderX

Edited by A Mavragani; submitted 24.10.24; peer-reviewed by S Ruiz-Correa; comments to author 25.02.25; revised version received 11.03.25; accepted 08.04.25; published 03.06.25.

Please cite as:
Borelli JL, Wang Y, Li FH, Russo LN, Tironi M, Yamashita K, Zhou E, Lai J, Nguyen B, Azimi I, Marcotullio C, Labbaf S, Jafarlou S, Dutt N, Rahmani A
Detection of Depressive Symptoms in College Students Using Multimodal Passive Sensing Data and Light Gradient Boosting Machine: Longitudinal Pilot Study
JMIR Form Res 2025;9:e67964
URL: https://formative.jmir.org/2025/1/e67964
doi:10.2196/67964
PMID:

©Jessica L Borelli, Yuning Wang, Frances Haofei Li, Lyric N Russo, Marta Tironi, Ken Yamashita, Elayne Zhou, Jocelyn Lai, Brenda Nguyen, Iman Azimi, Christopher Marcotullio, Sina Labbaf, Salar Jafarlou, Nikil Dutt, Amir Rahmani. Originally published in JMIR Formative Research (https://formative.jmir.org), 03.06.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Participatory Intervention Development of a Peer-Guided Self-Help App for Anxiety Disorders: Mixed Methods Study

Laura Duddeck[1,2], MSc; Timo Stolz[3], MSc; Christian Zottl[4], Dipl-Soz päd; Thomas Berger[5*], PhD; Johanna Boettcher[1*], PhD

[1]Department of Psychology and Psychotherapy, Psychologische Hochschule Berlin, Berlin, Germany
[2]Department of Education and Psychology, Freie Universität Berlin, Berlin, Germany
[3]Nullachtvierzehn Unternehmergesellschaft (Haftungsbeschränkt), Berlin, Germany
[4]Deutsche-Angsthilfe Eingetragener Verein, Munich, Germany
[5]Department of Psychology and Psychotherapy, University of Bern, Bern, Switzerland
[*]these authors contributed equally

Corresponding Author:
Laura Duddeck, MSc
Department of Psychology and Psychotherapy
Psychologische Hochschule Berlin
Am Köllnischen Park 2
Berlin, 10179
Germany
Phone: 49 1637398073
Email: l.duddeck@phb.de

Abstract

Background: Anxiety disorders affect approximately 27% of the global population, posing a major mental health challenge. Limited access to treatment due to resource constraints highlights the need for scalable solutions. Web-based self-help programs provide low-threshold access to evidence-based strategies. When guided by peers, these programs enhance engagement and acceptability by merging autonomy with support. Peer-guided self-help apps offer a cost-effective alternative to traditional care, reaching those who might otherwise remain untreated.

Objective: This study aims to describe the development of a peer-guided self-help app for anxiety, incorporating input from individuals with lived experience. It assesses user feedback on usability and helpfulness during the development process.

Methods: The intervention was developed in 3 iterative stages using the integrate, design, assess, and share framework. In stage 1, a prototype was cocreated by employees of a German self-help organization with lived experience, software engineers, and psychologists. In stage 2, qualitative feedback was collected from a focus group (n=5) and interviews (n=4), with participants recruited through group leaders of the organization. The research team directly contacted the participants. Qualitative data were analyzed with inductive and deductive content analysis (interrater reliability Cohen κ=0.88), which informed the minimum viable product (MVP) development. In stage 3, the MVP was pilot-tested with a larger online sample (N=126) recruited via the organization's website, accessible to all. Anxiety (Generalized Anxiety Disorder-7) and well-being (the World Health Organization-Five Well-Being Index) were assessed at baseline, 4, 8, and 12 weeks. Use metrics (eg, log-ins, time spent, and feature use) were recorded automatically. Quantitative data were analyzed descriptively.

Results: Stage 1 produced no data. In stage 2, feedback revealed unclear functionality, confusion in peer interaction, and safety concerns, leading to MVP revisions. In stage 3 (N=126), engagement was low—average log-ins were 3.15 (SD 14.37), with only 20 (SD 15.9) participants completing follow-ups. While many joined exposure (79/126, 62.7%) or activity scheduling groups (104/126, 82.5%), 123 (98.4%) did not send messages, undermining peer support goals. Baseline scores showed moderate anxiety (Generalized Anxiety Disorder-7: mean 10.52, SD 5.15), low well-being (World Health Organization-Five Well-Being Index: mean 15.80, SD 6.17), and low social support (Oslo Social Support Scale-3: mean 7.25, SD 2.68), consistent with the target group. Low engagement and high attrition indicated usability problems and limited perceived value.

Conclusions: Despite rapid sign-ups, user engagement was low and dropout rates high, indicating poor acceptance. Key barriers included user confusion, underused peer features, and technical issues. Future development should include structured onboarding for better clarity. Peer engagement be improved with prompts and enhanced safety perception. The participatory approach was

XSL•FO
RenderX

challenging and fell short of expectations. Smaller testing phases with regular user feedback will ensure user-centered refinement. Insights from successful peer communities can inform a more intuitive, engaging design.

(JMIR Form Res 2025;9:e62781) doi:10.2196/62781

KEYWORDS

peer-guided self-help; participatory design; integrate, design, assess, and share framework; digital mental health intervention; anxiety self-management; user engagement; qualitative usability study

Introduction

Background

Anxiety disorders are among the most prevalent mental health conditions globally, representing 26.9% of all mental health issues worldwide [1], putting considerable strain on mental health services. In Germany, individuals seeking psychotherapy frequently face waiting periods of 3 to 9 months [2]. These delays are linked to negative impacts on individual well-being and rising health care costs [3]. Given these challenges, scalable and accessible intervention formats are receiving more attention in mental health research.

One approach explored in this context is self-help, which supports individuals in managing their mental health challenges independently and may reduce reliance on professional care [4]. Earlier studies indicate a general preference for self-help among certain populations [5]. More recent findings support this trend; for instance, Samuel and Kamenetsky [6] found that students were more likely to seek help from friends, family, or the internet than the university's health and counseling center.

With internet-based interventions (IBIs) growing in availability over the past 2 decades, self-help has become increasingly accessible. IBIs have demonstrated effectiveness across various mental health conditions, particularly anxiety disorders [7], and have been considered a potentially suitable option for individuals who are hesitant to access traditional mental health services [8]. These interventions are commonly based on cognitive behavioral therapy and vary in their level of support; *guided interventions* offer (professional) support throughout the treatment process, while *unguided interventions* rely entirely on self-directed engagement, independent of human support. Although meta-analyses often show advantages of guided programs (eg, improved outcomes [9] and adherence [10]), randomized controlled trials (RCTs) do not consistently confirm superior outcomes for guided interventions (eg, concerning anxiety [11-13] and depression [14,15]). Different forms of guidance have been studied, including nonprofessional guidance, technical support, or diagnostic interviews [16-20], with promising results.

These findings have led to increased interest in how digital interventions might be structured to support adherence and outcomes in the absence of professional guidance. One potential form of guidance is peer support. Peer support as a key concept in traditional self-help refers to mutual support between individuals who share experiences in overcoming common challenges [21,22]. According to this definition, peers offer relatable guidance and emotional and motivational support and serve as role models for coping behaviors [23,24]. Peer support offers a balanced approach between fully guided and completely unguided interventions. It can strengthen adherence [25] and social support [18] and help address common barriers to the uptake of IBIs, such as comprehension issues, lack of time, or insufficient personalization [26,27]. In line with social identity theory [28], peer support can activate a sense of belonging and cohesion and reduce mental health stigma. It also provides an opportunity for social learning [29], with peers demonstrating adaptive coping strategies and providing behavioral models for users to emulate.

Despite increasing evidence for the efficacy of IBIs in treating anxiety, little is known about how peer-guided online formats can enhance engagement and outcomes. Most existing IBIs rely on professional guidance or are entirely self-guided, leading to limited research on intermediate models involving peers with lived experience. A recent study addressing this gap examined an online written exposure therapy program designed for veterans with posttraumatic stress disorder. Supported by peer coaches, the program demonstrated promising improvements in symptoms [30]. While the intervention successfully decreased posttraumatic stress disorder and depressive symptoms, it also revealed substantial challenges related to engagement and dropout rates. Over 50% of the participants did not start the program, and nearly half of those who did ultimately dropped out. These results underscore the need for careful design in digital mental health interventions, specifically targeting technical barriers and improving user engagement.

Peer support may benefit from being integrated into interventions that align with the needs and expectations of target users to enhance relevance and usability. Participatory intervention development (PD) offers such a pathway by actively involving end users in the design process. Instead of positioning users as passive recipients, PD encourages cocreation and continuous feedback to improve acceptance and effectiveness [25,31-33]. A recent study by Gonsalves et al [34] examines the co-design of "Baatcheet," a peer-supported, web-based storytelling intervention aimed at young individuals facing mental health issues in India. The project used iterative co-design workshops to develop a culturally relevant platform that integrates user-generated stories, reflection activities, and peer support, aiming to reduce stigma and foster connections. This research highlights the significance of incorporating peer components and using participatory design to enhance the authenticity, relevance, and engagement of digital interventions, providing valuable insights for similar app-based initiatives. However, despite its advantages, PD also entails challenges, such as recruiting representative participants [35], managing stakeholder conflicts [36], and time-intensive processes in gathering and analyzing user feedback [37]. Therefore, the evidence from PD regarding the challenges of PD's impact on

XSL•FO
RenderX

intervention efficacy is limited due to insufficient outcome data and inconsistent results across specific contexts [32,35,38]. Nevertheless, interventions seem to benefit from PD, as demonstrated in various contexts, including somatic complaints, youth mental health, and outpatient psychotherapy. Examples include a pharmacy system intervention involving pharmacy staff and older adults [37], a mental health e-Clinic developed in collaboration with young people and professionals [33], and a blended outpatient psychotherapy program incorporating input from psychotherapists and individuals with lived experience [39]. In addition, the value of PD extends beyond developing a final product. Participating in the PD process is associated with increased knowledge, enhanced mental health literacy, and opportunities for peer exchange [35]. So far, no study has applied a participatory approach to the development of a peer-guided self-help app for anxiety.

Objectives

This study addresses this gap by (1) developing a peer-guided self-help app for anxiety through a participatory approach using the integrate, design, assess, and share (IDEAS) framework and (2) evaluating its uptake, adherence, and preliminary effects on anxiety and well-being. The aim is to explore the potential of peer support in digital self-help and contribute to more user-centered, scalable mental health care solutions.

Methods

Overview

This study is based on the recurring phases 1 to 8 of the IDEAS framework [31], which served as a flexible orientation rather than a rigid sequence: (1) empathize with target users; (2) specify target behavior; (3) ground in behavioral theory; (4) ideate creative implementation strategies; (5) prototype potential products; (6) gather user feedback; (7) build a minimum viable product (MVP), defined as the first fully functioning version of the program that includes all core features; and (8) pilot-test.

Intervention Development With Individuals With Lived Experience (Phases 1-7)

Participants

The development process began in August 2020 with a team of 3 individuals who were affected by anxiety disorders and were also active members of a German self-help association. This association aimed to create a self-help app and secured funding for the project in collaboration with 2 software development company employees, including a programmer (TS) and a project manager. The funding enabled the initiation of the development process. After 19 months, 2 psychologists (JB and LD) joined the team. The participants in these phases were not specifically recruited. They applied collaboratively for further funding to support the app's development. The selection criteria for the software company included previous experience in developing health apps, while the psychologists were required to have expertise in online interventions.

Procedure

Weekly meetings involving all team members were convened to discuss target behaviors, core interventions, implementation strategies, linguistic expression, usability, and app design. Software engineers commenced programming while representatives from the self-help organization offered feedback on the technical prototypes. Psychologists contributed clinical insights and developed a research plan for the systematic piloting of the app. All team members conducted systematic testing of the prototype, adhering to a predefined testing schema.

Analysis

No data were systematically evaluated or analyzed.

Testing the Prototype: Focus Group and Individual Interviews (Phases 6 and 7)

By September 2022, a prototype incorporating fundamental functionalities had been prepared for the initial phase of external user testing. The objective was to collect external feedback (phase 6) and subsequently enhance the prototype into an MVP (phase 7).

Participants

Leaders of face-to-face self-help groups were informed about the study by the organizing committee of the self-help organization. They were asked to either participate themselves or invite their self-help group members to join. Detailed information was sent to those who expressed interest via email (n=9), including study information, the app link, and a date for the online focus group. For those who were unable to attend the focus group, individual interviews were arranged. While the sample size (n=9) may appear modest, it is consistent with qualitative research standards for early-stage, exploratory studies aimed at identifying key themes and informing intervention design. According to Guest et al [40], thematic saturation often occurs within the first 6 to 12 interviews, especially when the research question is narrow and the participant group is relatively homogeneous. In this study, the combination of a focus group and individual interviews provided complementary perspectives. Core themes were consistently observed across both formats, indicating that the most salient user needs, preferences, and barriers were captured. No specific inclusion or exclusion criteria were defined; the only requirement was that participants had to be members of self-help groups within the German self-help organization.

Measures

A focus group (n=5) and individual interviews (n=4) were conducted, moderated by LD (a PhD student with advanced training in psychotherapy) and Pauline Becker (an academic assistant). The video-based focus group lasted approximately 90 minutes, while each individual interview lasted around 60 minutes. Participants in the focus groups and individual interviews were asked to share their expectations regarding the app's effects and content (eg, "What did you expect from the app when you started using it?") and identify the strengths and weaknesses of the app, focusing on its general effects and benefits (eg, "What do you think are the app's biggest strengths and weaknesses?"). In addition, participants were encouraged to share their experiences regarding communication with other test users (eg, "How did you experience the exchange with others in the app?"). Finally, they were asked to describe the barriers and facilitators they encountered in the different

XSL•FO
RenderX

modules (eg, "What was difficult or disruptive? What keeps you from using this feature?"). Both the focus group and the individual interviews used a semistructured interview format (Multimedia Appendices 1 and 2). The format allowed for deviations based on participants' perspectives, opinions, and experiences. Topics discussed in the focus group were documented in an online tool, which was visible to all group members during the interview and shared afterward so that participants could provide additional feedback and comment on any misunderstandings or missing aspects.

Qualitative Analysis and Interrater Reliability

The focus group and interview recordings were anonymized and transcribed based on semantic content transcription guidelines by Dresing and Pehl [41], using MAXQDA (version 20; VERBI Software). Transcriptions were largely verbatim, with adaptations to standard written German where appropriate. Dialects were standardized, syntactic errors were retained, stuttering was smoothed, incomplete sentences were marked with interruption symbols, and paralinguistic elements (eg, laughter, pauses, and emphasis) were systematically noted. Each speaker's turn, including brief interjections, was formatted into its paragraph. Participants and interviewers were pseudonymized using numerical codes.

The qualitative analysis followed a structured, multistep coding process integrating both deductive and inductive approaches. First, level 1 categories were developed deductively from the interview guide (Multimedia Appendices 1 and 2), aligning with overarching research aims (eg, expectations, interaction, modules, which encompassed activity scheduling, exposure, and psychoeducation), user experience, and functionality. Two raters independently analyzed the same 27% of the data using these preliminary categories. Discrepancies in coding were resolved through consensus discussion between the raters, using specific excerpts to refine category definitions and improve clarity. Thereafter, the 2 raters developed level 2 categories inductively from the material (eg, community, interpersonal exchange, suggestions for improvement, and system bugs). Again, discrepancies in coding were resolved through consensus discussion. The process was rehearsed for level 3 categories (eg, time management, direct messaging, and upload features). No third party was required. Once sufficient category coverage and definition precision were achieved, the finalized framework was applied to the full dataset by both raters (Multimedia Appendix 3). To ensure consistency, interrater reliability was calculated using MAXQDA 2022. The degree of code overlaps across various text segments served as the reliability criterion. Interrater reliability was quantified using the Cohen κ statistic, following the method established by Brennan and Prediger [42]. The resulting Cohen κ value of 0.88 indicated a high level of agreement among the independent raters. Although member checking was not conducted, the refinement of categories was guided by direct alignment with participant expressions, thereby supporting interpretative validity.

Additional Testing

The app was tested by a group of psychology students (n=9) to evaluate its user-friendliness in greater detail. They used the app for 3 weeks. Usability feedback was collected during 2 face-to-face group discussions led by the psychologists (JB and LD) and the software developer (TS). This feedback was incorporated into the development of the MVP but was not further analyzed.

Testing the MVP: Questionnaires and Log Data (Phase 8)

Overview

Using feedback from the focus groups and interviews, the prototype was improved to develop an MVP. This version was tested with a larger group of users to evaluate its acceptability and gain an initial impression of the intervention's potential effectiveness. The pilot trial was preregistered (DRKS00030781).

Participants

Test users were recruited from the website of a German self-help association. The study website offered comprehensive details about the app, its content, and the study's objectives and procedures. Interested individuals could register using the link to the app. No inclusion or exclusion criteria were applied; any interested individuals could register for the app, regardless of symptom severity or membership in the German self-help association.

Measures

Quantitative Measures

Quantitative data were collected using validated self-report instruments at 4 time points: baseline time point (T0), 4-week time point (T1), 8-week time point (T2), and 12-week time point (T3). These assessments aimed to capture changes in anxiety, well-being, life satisfaction, self-efficacy, and social support, which were core constructs aligned with the app's objective to reduce anxiety symptoms and promote psychological empowerment through peer-guided self-help. Participants received automated email reminders to complete questionnaires at various time points. Demographic information was collected at T0 to describe the sample composition. Anxiety symptoms were assessed using the Generalized Anxiety Disorder-7 (GAD-7) [43], a widely used 7-item measure specifically developed for screening and monitoring generalized anxiety in clinical and community samples over the previous 2 weeks (eg, "nervousness, anxiety, or tension"). It was chosen for its brevity, high sensitivity to change, and excellent internal consistency (Cronbach α=0.85) [44]. GAD-7 aligned directly with the app's primary goal of anxiety reduction. Well-being was measured using the World Health Organization-Five Well-Being Index (WHO-5) [45], a 5-item scale assessing positive mood, vitality, and general interest in daily life (eg, "I have felt calm and relaxed"). It complemented the GAD-7 by focusing on positive indicators of psychological health, relevant for tracking subjective improvement during self-help interventions. The WHO-5 has shown high internal consistency (Cronbach α>0.92) [46]. Life satisfaction was captured via the General Life Satisfaction Short Scale-1 (L-1) [47], a single-item measure validated as a global index of subjective well-being ("How satisfied are you overall with your life at the moment?"). Its inclusion allowed efficient assessment of broader quality-of-life

XSL•FO
RenderX

outcomes that may improve through symptom reduction and increased social connectedness. The L-1 shows good retest reliability (r=0.67-0.70) [47] and is especially suited for low-burden measurement in digital interventions. General self-efficacy was measured using the Short Scale for General Self-efficacy Beliefs (ASKU) [48], a 3-item instrument evaluating perceived coping ability. It reflects users' belief in their capacity to manage challenges (eg, "In difficult situations, I can rely on my abilities."). The ASKU demonstrates good internal consistency (McDonald ω=0.81-0.86) [48] and construct validity in both clinical and nonclinical samples. To assess mental health-specific self-efficacy, the Mental Health Self-Efficacy Scale (MHSE) [49] was included. This 6-item measure evaluates confidence in handling emotional distress over the next month (eg, "How confident are you that on an average day over the next month...you will be able to effectively handle stress, anxiety, and depression?") and was specifically chosen for its relevance to expected app outcomes, including anxiety and stress management. The MHSE has shown high internal consistency (Cronbach α=0.89) [49]. Perceived social support was evaluated using the Oslo Social Support Scale [50], a brief 3-item measure assessing structural and functional support (eg, "How many people are close enough for you to count on during serious personal problems?"). As peer connection was central to the intervention, this measure determined whether app use enhanced users' support perception. Despite moderate internal consistency (Cronbach α=0.64) [50], it has been widely used in population-based studies due to its brevity and adequate construct validity.

Log Data

The log data delineated the use behavior pattern of each user. The app automatically recorded various use metrics. Our examination encompassed the number of visits to the app, the volume of messages sent, the average number of page hits per visit, the frequency of visits over time, the duration of active engagement within different modules, and whether users participated in or initiated exposure or behavioral activation groups (refer to the description of the app in the subsequent sections).

Analysis

The questionnaire and log data underwent descriptive analysis, which involved calculating means and SDs using SPSS (version 29; IBM Corp). This analysis provided insights into the characteristics of the study sample and app use. The log data captured participants' adherence and engagement with the app, thereby offering insight into the app's acceptability. Because only a very small proportion of individuals completed the questionnaires (refer to the subsequent sections), we forewent applying inferential statistics.

Deviations From the Trial Registration

Due to low user engagement with the app, data collection ceased at T3. No follow-up measurements were conducted at the 6-month time point, and no client change interviews occurred at the 8-week time point. The planned analyses were limited to a descriptive evaluation in SPSS (Figure 1).

Figure 1. Overview of the development process for a self-help app addressing anxiety, based on the integrate, design, assess, and share framework. MVP: minimum viable product.

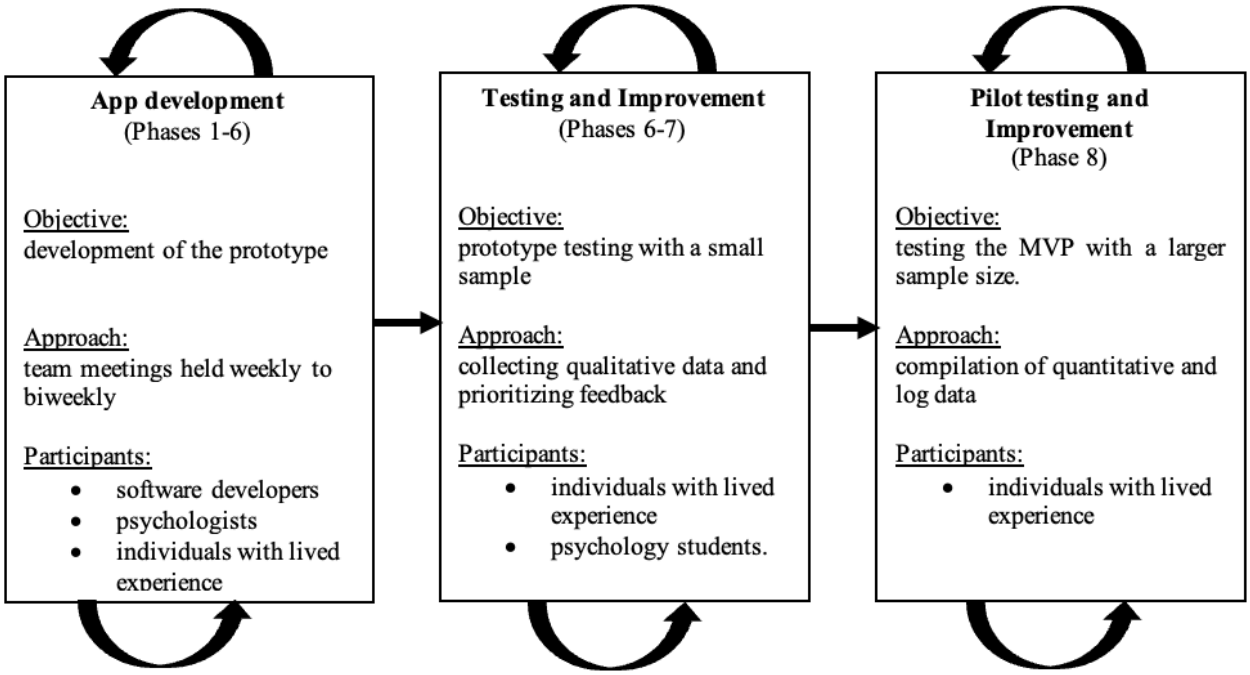

Ethical Considerations

Overview

This study received approval from the research ethics committee of the Psychologische Hochschule Berlin (EK202206). Informed consent outlined the study's objectives and procedures, potential benefits, and risks for participants and included data protection information, such as pseudonymization, anonymization, data processing, and the right to withdraw consent at any time.

XSL•FO
RenderX

Informed Consent in Prototype Development (Phases 1-6)

No explicit informed consent was obtained from individuals during this phase of the study, as no external participants were involved in the collaborative intervention development; no data were systematically assessed. The participants received an expense allowance of US $58. No compensations were paid in the other phases.

Informed Consent in Prototype Testing With Users Having Lived Experience (Phases 6 and 7)

Informed consent was obtained before participation. Participants received study information via email and confirmed their agreement in the app using checkboxes. In addition, verbal informed consent was secured before the focus groups and individual interviews.

Informed Consent in MVP Testing (Phase 8)

Informed consent was obtained through mandatory checkboxes during app registration, and a downloadable PDF version of the study information was provided for reference. For risk mitigation, the informed consent explained that the app did not replace psychological, psychiatric, or medical treatment, and the app contained a page with telephone numbers for current psychological emergencies.

Quantitative questionnaire data were pseudonymized during collection and anonymized before analysis.

Data Security and Privacy

In terms of data security, the app complied with the General Data Protection Regulation and the International Organization for Standardization 27001 standards. During prototype testing, audio recordings from Zoom (Zoom Communications, Inc) sessions were stored locally, transcribed without personal identifiers, and deleted following transcription. In MVP testing, participant data collected during registration were securely stored and removed upon permanent account deactivation or at the participant's request. Log data were pseudonymized during data collection, and fully anonymized data were archived for research purposes after the study. Encrypted backups were stored in European Union–based data centers managed by Hetzner, with password-protected access limited to project leads. Team members signed confidentiality agreements and adhered to strict data protection protocols, ensuring no identifiable participant information was included in publications, images, or supplementary materials. Moreover, users maintained full control over their privacy and anonymity in the app. During registration, participants selected a username without any mandatory disclosure of personal information, such as age, gender, or potential diagnoses. Users decided how much they wanted to share. For example, when creating groups, they could choose whether to keep them private or accessible to others. Furthermore, the preparation and follow-up of exposure exercises could be kept private or shared with the group. This approach ensured that participants could interact in a way aligned with their comfort and well-being while safeguarding their data.

Results

The qualitative and quantitative results of the PD study are presented subsequently in chronological order and sorted by development stages (initial app development, prototype testing, and MVP testing). The results from each stage informed the further development of the app.

Stage 1 (Phases 1-6): Development of the App Prototype

Overview

On the basis of iterative co-design across phases 1 to 6, which included qualitative input from individuals with lived experience and usability feedback from students, the app was developed into three key modules: (1) psychoeducation, (2) exposure, and (3) activity scheduling. Users can access all modules simultaneously without a prescribed order of engagement, allowing for flexible navigation and personalized content selection.

Psychoeducation Module

The module offers concise texts and graphics as microlearning units [51], providing essential information on anxiety and therapeutic strategies such as activity scheduling and exposure. This format was designed to improve comprehension and retention [51], especially for users who reported limited time.

Exposure and Activity Scheduling Module

These modules are interactive and group based, directly addressing users' requests for social support and shared accountability, enabling users to join or create groups based on shared goals. Using a private-public slider, users can name their group, add a description, upload an image, and adjust its visibility. Public groups appear in a list within the respective module, where users can preview group details and join. Upon joining, users can engage in ongoing discussions, post comments, share experiences, and provide feedback, such as fostering a collaborative and supportive community. The *activity scheduling module* helps users discover and incorporate positive activities into their daily routines. For instance, a user struggling to practice progressive muscle relaxation might join a "relaxation" group to connect with others who share practical tips, such as starting with brief sessions or using audio guides. Alternatively, users can create their own groups, such as a "yoga" group, to encourage participation in weekly sessions and track and share progress. Activity scheduling, originally derived from depression treatment [52], aims to increase engagement in pleasurable activities and promote positive environmental interactions. Research by Taylor et al [53] suggested that activity scheduling may also reduce anxiety by enhancing positive affect, cognitions, and behaviors. The peer-driven dynamic within the activity scheduling module is expected to encourage healthy behaviors, inspire users through shared experiences, and increase the likelihood of sustained engagement. Similarly, the *exposure module* supports users in confronting anxiety-inducing situations. Users can join groups with shared challenges or create their own, such as a "bus riding" group for individuals experiencing anxiety related to

XSL•FO
RenderX

public transportation. Within this module, users document their experiences, for example, riding a bus, and rate their anxiety before and after completing the task. Entries can be kept private or shared with the group, allowing peer feedback and support. Group members can discuss setbacks, share successes, and offer practical advice, such as starting with short rides during off-peak hours. The exposure module is designed to systematically encourage users to confront their fears and test the validity of their anxious thoughts through behavioral experiments. According to treatment guidelines, exposure exercises are a cornerstone of cognitive behavioral therapy (for reducing anxiety symptoms and decreasing avoidance behavior in anxiety disorders [54]. Peer-to-peer interaction within this module facilitates the planning and implementation of exposure exercises, motivating users to persist through mutual accountability and emotional support. This approach empowers users, reduces avoidance behaviors, and fosters adherence to anxiety-reducing strategies.

Senior Role

To further enhance engagement, the app introduces a *senior role* inspired by group leaders in traditional self-help settings. Selection criteria for senior members included their personal experience with anxiety disorders, familiarity with self-help group practices, and affiliation with the German self-help organization. In addition, candidates were required to be willing to volunteer regularly. During prototype testing, participants were asked whether they could imagine stepping into a senior role, what challenges they anticipated in that position, and the support they would require. They recommended enabling direct messaging with other senior team members and emphasized the need for intervision and supervision. Senior members who supported MVP testing were test users during the prototype testing. They were familiar with the app and its functions and had contributed to refining the role of senior members. Therefore, no formal training was necessary. Once MVP testing began, senior members participated in regular supervision meetings with CZ.

The senior role was designed to minimize disruption to group dynamics while positively influencing user outcomes. Senior members were expected to intervene only in specific situations, such as providing advice upon request, assisting with task completion, or mediating conflicts. However, their role did not include offering regular feedback or one-on-one support, ensuring that the group maintained its peer-to-peer structure. By serving as accessible and equal-status points of contact, senior members modeled positive behaviors and contributed to fostering a supportive and empowering community environment.

Stage 2 (Phases 6 and 7): Qualitative User Feedback on the Prototype

Overview

During this phase, the prototype of the app was tested with 9 users who were unfamiliar with the app. Through qualitative analyses of focus groups and interviews, key barriers and facilitators were identified (Textbox 1), which directly informed the refinement of app features and structure for the MVP. The first 2 main categories primarily reflected participants' expectations and attitudes toward anticipated app use. In contrast, the remaining 4 main categories pertained to actual user experiences during the prototype testing period. Multimedia Appendix 3 provides a detailed coding system, including anchor statements.

XSL•FO
RenderX

Textbox 1. Outcomes from the qualitative analysis conducted during the prototype testing phase of the self-help app for anxiety. Participants (n=9) were recruited through a German self-help organization. The deductive main categories and inductive subcategories outline user feedback regarding the app's features, design, and functionality.

Outcome expectation

- Specific guidance
- Community
- Safety experience

Interaction

- Role of an experienced peer supporter: time management, supervision, responsibility, and psychological strain
- Interpersonal exchange: no exchange, communication appeal, and direct messaging

Activity scheduling

- Inspiration
- Improvement suggestions

Exposure

- Inspiration
- Improvement suggestions

Psychoeducation

- Linguistic expression
- Knowledge sharing
- Improvement suggestions: references and indexing

User experience and functionality

- User-friendliness
- Bugs: save and upload, edit, and navigation
- Lack of clarity
- Proposal for optimization: menu, module name, group content, optical presentation, guidance, visibility and privacy, and information flow

Outcome Expectations

Overview

Users were asked about their expectations for the app and the changes they hoped to achieve by using it. With this question, we aimed to assess users' needs for a self-help app and how well those needs align with the app's specific setup and content. Three main themes emerged from this.

Specific Guidance

Users anticipated that the app would offer structured knowledge and practical tools for managing anxiety, including step-by-step instructions for completing exercises. One participant expressed the following expectations:

> *[I]t was an idea to expect to...get some hints: how do I deal with anxiety in general?... Maybe skills for panic attacks are also listed, in a list or ideas or something.* [B4:22]

These expectations influenced updates to the psychoeducation module and exposure preparation features.

Community

Participants anticipated that the app would encourage mutual exchange, social support, group cohesion, and a sense of belonging among individuals experiencing anxiety. One participant shared the following:

> *So actually, you meet people who maybe have the same problem, and then you don't feel so alone anymore.* [B5:16]

This feedback highlighted the significance of peer interactions and validated the app's development goals in creating a supportive user community.

Safety Experience

Users emphasized the significance of moderated group discussions and crisis management strategies to reduce risks such as inappropriate comments or distressing situations. One participant shared the following concerns:

> *[B]ut in terms of content, I fear there are obscure stories in it. Sometimes, even in the group...anything like "do not put up with anything, hit back immediately!" and so on.... Then I almost fall off my*

XSL•FO
RenderX

> *chair...because I think...that's not a positive thing.... And these are my fears, if this is not moderated, that this will come up in there...* [Interview number 2:102]

In response, safety features were integrated into the MVP. First, in the psychoeducation module, we included information on managing personal crises and provided a link to national crisis hotlines. Second, we established group rules, such as navigating differing opinions, offering constructive feedback, practicing self-kindness, avoiding the glorification of dysfunctional behaviors, and avoiding self-promotion and spam. Third, we introduced the role of senior members, who could be contacted for support in using the app and moderating discussions within groups. They could be reached if group rules were violated or when conflicts arose. Fourth, we developed a reporting tool that enabled every user to flag inappropriate content. Flagged content became invisible until a senior peer decided to address or remove it. Senior members could consult one another before acting. This system was designed to ensure user safety and distribute responsibility among group members, thereby reducing the burden on individual senior members (refer to the subsections Responsibility and Psychological Strain).

Interaction

We asked about the app's specific functions and modules to identify what additions or changes might be beneficial as well as which features were viewed as especially useful. In addition, we explored the role of senior members and any challenges related to their position.

Role of Experienced Peer Supporters

Participants were asked about their perceptions of a senior peer as a guide within the app as well as any challenges they might face.

Time Management

Users emphasized the necessity for clearly defined response times to manage expectations about senior availability:

> *[S]o I would say that there should be the possibility that you can...write in [the app] at any time of day or night, but the seniors' reaction...that maybe needs to be set to a time frame.* [Interview number 3:50]

Supervision and Guidance

Users highlighted the importance of providing senior members with continuous feedback and support to manage complex situations.

Responsibility

Participants raised concerns regarding senior members' responsibility for app-generated content. In response, the MVP clarified that senior members were not required to monitor discussions continuously and were supported through intervision and supervision (Stage 1 [Phases 1-6]: Development of the App Prototype section). As the user base grows, additional senior members will be recruited. Prospective candidates must meet specific eligibility criteria, including a minimum registration period, and complete a structured training program before assuming the senior role.

Psychological Strain

Participants expressed concerns about the potential stress linked to being a senior member, fearing it could lead to pressure and anxiety:

> *Do I feel responsible for what comes up in the chat room or in the exchange at all?... Do I FEEL responsible there? And this can...create pressure, stress, and anxiety...* [Focus group 1:162]

To address this matter, explicit role definitions and supportive mechanisms, namely intervision and supervision, were established.

Interpersonal Exchange

Although the prototype did not fully integrate various communication modes, participants shared their thoughts on in-app communication, emphasizing its benefits and challenges. Users expressed a substantial interest in direct communication options, highlighting the need for effective messaging features. These findings reinforced the ongoing efforts to develop a consistent messaging system.

Feedback on App Modules

Overview

Participants assessed the modules centered on activity scheduling, exposure, and psychoeducation. We aimed to understand how participants, who were experts with lived experience but were not part of the initial app development, perceived the importance and practical application of the modules. Key findings are given in the subsequent sections.

Activity Scheduling and Exposure

Users appreciated the ability to draw inspiration from their peers' exercises and valued the feature that allowed them to rate their anxiety levels both before and after exposure. This functionality helped contextualize their personal experiences and normalize anxiety:

> *Yes...I think it makes sense. I think that simply recording fear triggers [helps],...it is often the case that I have a fear and think to myself, "I am weird, no one else has that"...* [Interview number 4:48]

On the basis of the feedback in this category, the preparation and follow-up sections of exposure exercises were refined to improve clarity and understanding.

Psychoeducation

Although users generally appreciated the psychoeducational content, they suggested simplifying the texts for better understanding.

Usability, Navigation, and Function

From the feedback gathered, 3 main themes emerged.

Technical Issues

The most frequently reported issue was fixing bugs, especially those related to saving posts, uploading photos, and editing user comments.

XSL•FO
RenderX

Navigation

Users expressed a desire for improved navigation within the app. They often found themselves seeking clearer guidance on the actions to take and the appropriate moments to undertake those actions. They suggested integrating a well-structured user journey in the MVP to address this confusion and promote greater compliance. A recurring theme in their feedback was the arrangement of the app's modules. Users believed that a more logical flow would substantially enhance their experience. For instance, they recommended moving the exposure module to precede the activity scheduling module, enabling users to transition smoothly between tasks. In addition to reorganizing the modules, users expressed a preference for renaming them to make them more relatable. They proposed changing "activity scheduling" to something such as "what helps me" and suggested replacing "psychoeducation" with a simpler term such as "wiki." They felt these changes would resonate better with users and make navigation more intuitive. Furthermore, they suggested adding a sorting function to simplify the activity scheduling process. This feature would help users easily identify groups that match their personal interests and preferences, ultimately fostering a more personalized and engaging experience with the app.

Information

Participants emphasized the importance of information buttons, pop-ups, reminder functions, and a progress-tracking page to display exercises, along with detailed descriptions of the purposes and benefits of modules and groups. These features were deemed essential for enhancing task comprehension and potentially increasing user engagement and participation. A commonly suggested improvement was the ability to preview group descriptions before joining, enabling users to make more informed decisions. Participants also highlighted the need for better visualization of exposure exercises, such as graphical representations that showed how their experienced anxiety was lower than expected. In addition, they requested a broader variety of profile pictures for groups, possibly by incorporating stock images.

Additional Steps

Qualitative feedback informed further development of the prototype before it was tested in 2 focus groups with 9 students to explicitly test and enhance user-friendliness.

Key Areas for Improvement

Students and test users highlighted the necessity for clearer explanations of the app's core content and features to improve usability. Their recommendations centered on refining the navigation bar layout, providing concise instructions within modules to clarify their purposes, and adding notifications to keep users updated about new messages and events. However, opinions differed regarding the effectiveness of notifications as motivational tools, with some participants expressing concerns that these might create unnecessary pressure.

Usability and Technical Improvement

Students suggested creating preexisting example groups to illustrate proper use of exposure and activity scheduling modules, implementing a messenger-style interface that allows direct replies to specific messages for improved communication, streamlining the psychoeducation module with graphics and expandable sections to support the microlearning approach, enabling progress tracking for completed tutorials and beneficial activities, and developing user profiles to foster more interactions about personal experience.

Adjustment of App Development Process

Considering the extensive feedback and challenges faced during prototype testing, we decided to modify the development process by narrowing the scope of future iterations to prevent unnecessary programming, collecting feedback at shorter intervals to ensure it meets user needs, and adding a reporting feature for users to directly report technical issues.

Significance of Student Testing

The dual feedback approach provided essential insights from various perspectives. Test users with lived experience evaluated whether the app's content met their expectations and identified potential barriers to adapting the self-help model to an online format. Students primarily assessed usability and technical execution, confirming that previously reported issues were resolved and identifying new technical and comprehension challenges. By incorporating feedback from both groups, we ensured a broader range of perspectives, including those from both technologically proficient users and less experienced individuals. In addition, students served as a readily accessible and impartial test group compared to members of the German self-help organization.

Stage 3 (Phase 8): Quantitative Results of MVP Testing

Overview

Following the qualitative feedback obtained from individuals with lived experience and students (as described in the earlier phases), the app prototype was revised to address specific usability concerns. Improvements were made to simplify navigation, clarify module instructions, particularly for exposure exercises, and include supportive group functions. These changes were integrated into the MVP, which was made publicly accessible via the German self-help organization's website for broader testing.

Demographic Characteristics

Baseline data were collected from a sample of 20 users who completed pretreatment questionnaires. Table 1 presents the demographic characteristics of the sample.

XSL•FO
RenderX

Table 1. Overview of the demographic characteristics of test users who provided baseline data for testing the minimum viable product (n=20).

Variable	Users, n (%)
Age (y)	
18-24	0 (0)
25-34	6 (30)
35-50	10 (50)
>50	4 (20)
Gender	
Women	15 (75)
Men	5 (25)
Marital status	
Married	10 (50)
Divorced	3 (15)
Single	7 (35)
Self-reported mental disorder diagnoses	
Generalized anxiety disorder	9 (45)
Panic disorder	2 (10)
Remaining diagnoses (depressive, alcoholic, or other disorder)	9 (45)
Treatment history	
Outpatient psychotherapy	11 (55)
Inpatient or partial inpatient therapy	2 (10)
Self-help groups	1 (5)
No previous treatment	5 (25)
Other treatment form	1 (5)

MVP Use and Engagement Patterns; Participant Flow and Dropout

Despite targeted design improvements, log data indicated persistent engagement issues. Of the 126 registered users, only 20 (15.9%) completed the baseline questionnaires, and substantial dropout occurred over time, with 75% (15/20) of the participants failing to respond at T3 (Figure 2).

This limited engagement reflected the qualitative findings where users reported confusion, emotional overload, and a lack of structured support, particularly in the exposure module. For example, several interviewees shared uncertainty about how to perform exposure exercises. This aligned with the observation that only 25% (5/20) of the survey participants engaged with the exposure module, and only 2 participants from the entire sample sent any messages via the platform.

XSL•FO
RenderX

Figure 2. Flowchart illustrating the quantitative data collection process during the minimum viable product (MVP) testing phase for the self-help app aimed at alleviating anxiety. A total of 126 participants were involved. Measurement points included baseline (T0) and 4 (T1), 8 (T2), and 12 weeks (T3), with questionnaires evaluating app use, adherence, and user experience. Dropout rates steadily increased across the measurement points.

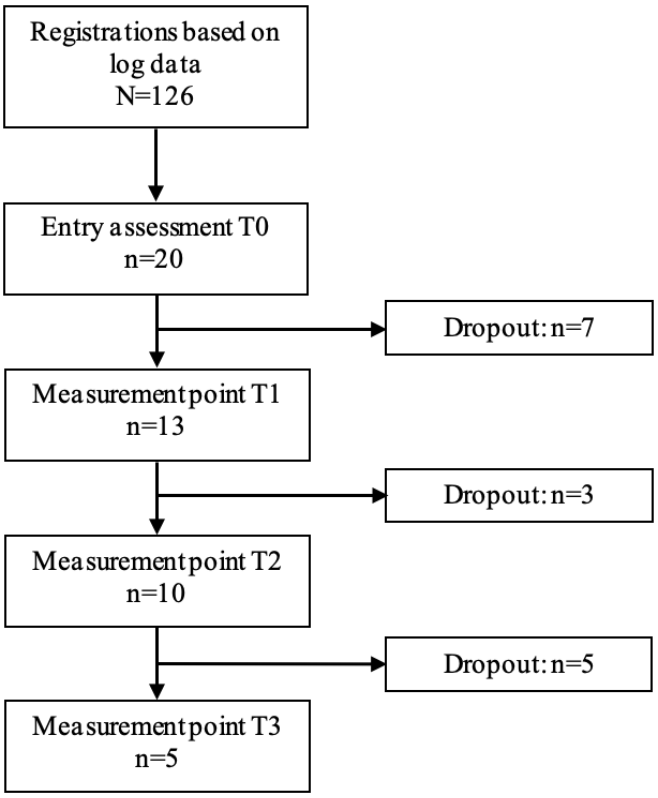

Outcome Questionnaires and Relation to App Use

The mean anxiety scores at T0 (GAD-7: mean 10.52, SD 5.15) reflected moderate impairment, consistent with the target population. According to established clinical cutoffs [43], 9% (2/21) of the participants were classified as experiencing minimal strain, 43% (9/21) as experiencing mild strain, 24% (5/21) as experiencing moderate strain, and 24% (5/21) as experiencing severe strain. The MHSE scores seemed comparable to normative data in clinical samples [49] and improved over time (from a mean of 32.85, SD 12.94 to a mean of 43.83, SD 12.50), suggesting a possible link between active use, especially of the activity scheduling module, and perceived coping ability. This module also had the longest average use time of 8.34 minutes, indicating that users found it both accessible and helpful. Qualitative feedback corroborated this, as users highlighted that “concrete activity suggestions felt more motivating than educational texts.” The brief measure of general self-efficacy (ASKU) showed that participants felt moderately self-efficacious (mean 3.10, SD 0.90), with a nonclinical sample reporting slightly higher average scores. Conversely, general well-being (WHO-5) slightly decreased at T3 (from a mean of 15.80, SD 6.17 to a mean of 12.70, SD 5.85), which may reflect reduced overall engagement or increasing dropout over time. Using nonclinical cutoffs [46], 10% (2/20) of the participants were classified as having substantially reduced well-being, 20% (4/20) had reduced well-being, 25% (5/20) reported satisfactory well-being, and 45% (9/20) reported very good well-being. Participants reported low perceived social support (Oslo Social Support Scale-3: mean 7.25 SD 2.67), aligning with both quantitative results and qualitative concerns about the usefulness of peer group features. Although many participants joined groups for activity scheduling (9/20, 45% among survey participants), almost none initiated or actively communicated within these groups, highlighting a persistent challenge in fostering interactive support environments. Among participants, 0% (0/10) were classified as having strong social support, 30% (3/10) had moderate social support, and 70% (7/10) had poor social support. Life satisfaction, as measured by the L-1, was low on average (mean 4.15, SD 2.08; Table 2).

XSL•FO
RenderX

Table 2. Summary of the quantitative results for the minimum viable product. It displays survey data collected at various time points (baseline time point [T0], 4-week time point [T1], 8-week time point [T2], and 12-week time point [T3]) for different mental health and well-being indicators. Classification scales are used to categorize the results for each indicator.

Scale and comparative scores	T0, mean (SD)	T1 (SD)	T2, mean (SD)	T3, mean (SD)	Scores, mean (SD)
GAD-7[a] (0-4: minimal anxiety, 5-9: mild anxiety, 10-14: moderate anxiety, 15-21: severe anxiety)	10.52 (5.15)[b]	13.30 (4.35)[c]	11.45 (5.99)[d]	—[e]	2.95 (3.41)[f]
WHO-5[g] (19-25: very good, 13-18: overall satisfactory, 12-10: reduced, 10>: significantly reduced)	15.80 (6.17)[h]	18.70 (3.74)[c]	17.18 (5.17)[d]	12.70 (5.85)[i]	17.58 (4.97)[j]
MHSE[k]	32.85 (12.94)[h]	35.11 (8.57)[l]	37.27 (14.56)[d]	43.83 (12.50)[i]	33.23 (11.45)[m]
ASKU[n]	3.10 (0.90)[h]	2.96 (0.48)[l]	3.50 (0.96)[d]	3.83 (0.86)[i]	4.00 (0.74)[o]
OSSS-3[p] (3-8: poor support, 9-11: moderate support, 12-14: strong support)	7.25 (2.67)[h]	7.78 (2.17)[l]	9.00 (2.49)[d]	8.83 (2.40)[i]	10.16 (2.07)[q]
L-1[r] (0: not at all satisfied, 10: completely satisfied)	4.15 (2.08)[c]	3.70 (1.25)[c]	4.27 (3.07)[d]	5.83 (1.47)[i]	7.18 (2.07)[s]

[a]GAD-7: Generalized Anxiety Disorder-7.
[b]n=21.
[c]n=10.
[d]n=11.
[e]Not available; no calculations could be performed due to missing data (dropout).
[f]Löwe et al [44] and Spitzer et al [43].
[g]WHO-5: World Health Organization-Five Well-Being Index.
[h]n=20.
[i]n=6.
[j]Brähler et al [46].
[k]MHSE: Mental Health Self-Efficacy Scale.
[l]n=9.
[m]Clarke et al [49].
[n]ASKU: Short Scale for General Self-efficacy Beliefs.
[o]Beierlein et al [47].
[p]OSSS-3: Oslo Social Support Scale-3.
[q]Kocalevent et al [50].
[r]L-1=General Life Satisfaction Short Scale-1.
[s]Beierlein et al [48].

System Use: Full Sample Versus Survey Participants

A total of 125 participants (excluding senior members and 1 individual whose log data were not saved) logged into the app, with an average of 3.15 (SD 14.37) visits per participant. Among these, 86 participants (68.8%) accessed the app only once after registration, 18 (14.4%) visited twice, 13 (10.4%) visited 3 times, and 8 (6.4%) visited ≥5 times.

Regarding messaging activity, 123 (98.4%) of the 125 participants sent no messages, while 2 (1.6%) sent ≥2 messages. The average page hits per visit were 32.39 (SD 28.49). On average, participants engaged in activities on 2.02 (SD 4.94) days. The average active time spent was 6.02 (SD 10.57) minutes in the psychoeducation module, 1.55 (SD 2.43) minutes in the exposure module, and 8.34 (SD 32.28) minutes in the activity scheduling module.

Regarding group participation, 79 (62.7%) of the 125 individuals joined ≥1 exposure group, while 2 (1.6%) created a new exposure group. In total, 104 (82.5%) participants joined ≥1 group for activity scheduling, and no (0%) participant created any activity-scheduling groups.

The average number of visits among the 20 survey participants was 2.60 (SD 2.19). Specifically, 7 (35%) participants visited the app only once after registration, 7 (35%) visited twice, 3 (15%) visited 3 times, and 3 (15%) visited ≥6 times. None of the participants sent messages. The average number of page hits per visit was 31.95 (SD 21.37). On average, participants engaged in activities on 2.30 (SD 1.81) days. The average active time spent in the psychoeducation module for 16 participants was 7.42 (SD 10.51) minutes; in the exposure module for 15 participants it was 1.34 (SD 1.29) minutes; in the activity scheduling module for all 20 participants it was 9.27 (SD 12.21) minutes. Regarding group participation, 5 (25%) participants joined ≥1 exposure group, and none created a new exposure group. For activity scheduling, 9 (45%) participants joined ≥1 group, and no participants created any activity scheduling groups.

XSL•FO
RenderX

Integrated Interpretation and Implications for Further Development

The MVP of the self-help app was refined based on qualitative feedback from individuals with anxiety and students, addressing issues such as unclear structure, a lack of motivation, and confusing modules. These changes aimed at enhancing usability and engagement. Despite these improvements, quantitative data indicated that user engagement remained low: only 15.9% (20/126) of the registered users completed the baseline survey, and 75% (15/20) dropped out by T3. Most (7/20, 35%) users logged in only once, and very few interacted with group features and none sent messages, mirroring earlier qualitative reports of confusion and insufficient guidance, particularly regarding the exposure module. Outcome measures revealed moderate anxiety and self-efficacy at T0. While self-efficacy increased over time, especially in modules such as activity scheduling, overall well-being and life satisfaction remained low, likely due to limited engagement. Qualitative and quantitative findings consistently highlighted the need for better onboarding, clearer structure, and more support to improve adherence and therapeutic effectiveness.

Discussion

Principal Findings

This study aimed to investigate the development of a self-help app using the IDEAS framework as a guiding structure rather than as a strictly linear road map. Overall, the findings indicate that, despite the extensive involvement of experts with lived experience, a stepwise and framework-informed development approach, and repeated mixed methods data assessments, the app was not acceptable to individuals seeking self-help for anxiety. Efficacy could not be estimated due to low use. Nevertheless, this study provides valuable insights into user engagement, adherence, and the challenges of creating a peer-supported digital intervention for anxiety.

Participatory Intervention Development

The app's development actively engaged software engineers, members of a self-help organization, and psychologists. This multidisciplinary collaboration aimed to ensure that the intervention was both technically feasible and responsive to the needs of individuals with anxiety. Although stakeholders' involvement enriched the content and functionality, several challenges emerged, particularly in aligning diverging priorities within a limited time frame, which often affects the thoroughness of analysis and preparation between user feedback sessions [37]. Such difficulties are common, as varied professional backgrounds can lead to task conflicts. Remarkably, these conflicts often yield more innovative and effective outcomes [55] when managed effectively.

In this study, the limitations of time led to the initial prototype featuring only a small set of functions and several unresolved usability issues. External test users in prototype testing (phases 6 and 7) reported fundamental difficulties in understanding and navigating the app. These late-stage insights were highly relevant but arrived too late to inform earlier design decisions meaningfully. Consequently, the MVP of the app included only a limited number of features and retained unresolved barriers to user engagement. This led to minimal messaging and low overall activity within the app, findings that directly relate back to these development gaps. Future iterations should integrate usability testing, including participants with minimal previous experience in digital self-help, to ensure that app structures are intuitive and actionable for the intended audience.

Adherence and User Engagement

As demonstrated by baseline data, the MVP successfully attracted its intended target group, individuals experiencing anxiety and low social support. However, true engagement with the intervention remained low. Log data revealed that many users accessed the app only once, logged in infrequently, rarely communicated, and did not actively participate in scheduling activities or exposure-related group tasks. This lack of adherence raises substantial concerns about the feasibility of a fully self-guided peer-support approach without professional facilitation.

A particularly striking result was the near-complete absence of messaging activity; although 63.2% (79/125) of the participants joined an exposure group and 83.2% (104/125) joined an activity scheduling group, 98.4% (123/125) did not send a single message. This suggests that participants were curious enough to explore group features but lacked the confidence, clarity, or motivation to engage in communication, an essential mechanism of peer support. Qualitative feedback further highlighted barriers such as uncertainty about message visibility, fear of doing something wrong, and the absence of initial engagement from others. One interviewee explained the following:

> *[T]he problem was probably that there were simply no comments or feedback, so I had to make the first step, and then I realized...this is visible to all. What if I do something wrong?* [Interview number 1:27]

These findings suggest that joining a group does not equate to meaningful participation. While similar issues have been reported in previous studies [56-58], the level of engagement observed in this study was comparably lower. For instance, Hanano et al [57] conducted a secondary analysis of an RCT on a self-guided intervention for anxiety and depression and found an average of 2.14 logins per module across the 8 modules available, with 48% to 63% of the participants not logging in at all. Similarly, another RCT evaluating an online intervention for treating depression, which included interactive features, such as animation, video, and narration, recorded an average of 18.7 logins per participant [56]. By contrast, in this study, the mean number of logins was only 3.15 (SD 14.37), with many (125/125, 100%) users logging in just once.

In terms of page views, our findings (mean 32.39, SD 28.49) were comparable to those of another self-guided intervention aimed at reducing social anxiety symptoms, which reported an average of 37.6 (SD 41.3) page views [58]. However, this apparent similarity requires critical interpretation, as the intervention followed a structured, sequential module design, offering clearer user pathways and guidance. In contrast, our app featured an open structure without step-by-step progression or automated guidance, potentially limiting user orientation and

XSL•FO
RenderX

reducing sustained engagement. The comparison underscores a critical insight; that is, while peer support is theoretically promising, it may not emerge spontaneously in an unmoderated digital environment. This app was designed with an open structure that does not dictate a specific sequence of interactions or instructions on use. Without consistent user engagement, the core principle of peer support cannot be realized, leading to reduced interaction, less communication, and a weakened sense of community. In turn, this leads to user attrition as group cohesion fails to develop.

Technical challenges and usability issues likely compounded these difficulties and contributed to dropout [27]. Despite efforts to improve navigation and clarify app functions during development, user feedback indicated ongoing confusion and anxiety about how the app was supposed to work. This is consistent with the systematic review by Borghouts et al [59] of user engagement with digital mental health interventions, which found that technical difficulties and privacy concerns are key predictors of disengagement. This aligns with feedback from our test users, who reported difficulties with the app's complexity and a lack of transparency regarding its functions and tasks, which hindered their engagement. To address these issues, efforts were made to improve the app's intuitiveness by introducing a clearer structure and a user journey to enhance navigation.

Overall, our findings indicate that an open, nonsequential peer support model may create entry barriers too high for meaningful user participation. Consequently, the app did not create an environment conducive to users learning from one another [60,61], experiencing empowerment, and enhancing their self-efficacy in anxiety management [4]. Unlike previous research [62,63], our results did not confirm that peer support can be a substitute for professional guidance or substantially enhance adherence. Therefore, future versions should include structured onboarding and behavioral guidance, such as short interactive tutorials, step-by-step introductions to key features, or even temporary facilitation (eg, peer moderators or bots), to encourage interaction and support early-stage engagement while building initial trust, particularly for users with low digital confidence.

Anxiety and Self-Efficacy

Due to low adherence, inferential statistics were not used. However, descriptive data revealed only minimal changes in anxiety levels (GAD-7), suggesting that the intervention had limited impact on symptom reduction. Similarly, general self-efficacy scores showed negligible changes in mean values. However, mental health–specific self-efficacy showed a somewhat more pronounced improvement (MHSE: mean 42.83, SD 12.50) compared to clinical reference values (mean 33.23, SD 11.45) [49]), indicating that even minimal peer-based digital interactions might offer motivational effects. While this increase is promising, it should be interpreted with caution due to limited use and the lack of inferential testing; these factors highlight the limited effects in the context of low user engagement. Because the app relied heavily on peer interaction and user engagement, insufficient communication combined with minimal repeated use likely hindered the intended changes from occurring. Specifically, features such as exposure group participation and activity scheduling were seldom used effectively, often lacking peer dialogue or follow-up actions. In the absence of active involvement or feedback from peers, chances for collaborative learning, emotional support, and reinforcement of coping strategies were mostly missing. The slight increase in mental health–specific self-efficacy may reflect a general sense of motivation or positive expectation following onboarding. However, this optimism did not translate into sustained interaction with the app, as shown by use patterns. One possible interpretation is that some users initially believed the app could help but became disengaged when faced with unclear instructions, a lack of guidance, or low peer activity, dampening the motivational effect over time.

Future research should focus on strategies to enhance the app's attractiveness and usability to bridge the effectiveness gap. Potential upgrades may involve setting up exposure exercises in advance to clarify their application, encouraging active engagement among senior members, or creating a buddy system to boost user interaction.

Sample Characteristics

The sample predominantly consisted of individuals aged >35 years (14/20, 70%), many (19/20, 95%) of whom reported no previous experience with self-help. Although not statistically tested in this study, age and previous experience may have influenced digital engagement. Older users, particularly those unfamiliar with digital platforms, may have required more structured onboarding and clearer guidance to navigate the app effectively. This was also reflected in usability feedback during prototype testing, where several participants expressed confusion regarding the app's functions, even though they were experienced with self-help. These factors likely contributed to the app's limited use, as shown by log data.

Limitations and Future Directions

This study has several limitations that should be acknowledged when interpreting the results and considering their generalizability. First, all participants for prototype testing were recruited from a specific German self-help association and had previous experience with face-to-face self-help. In contrast, MVP participants were recruited via the association's website without specific inclusion criteria. This recruitment strategy likely introduced a selection bias favoring individuals with a positive predisposition toward self-help and a higher digital affinity. Therefore, the findings may not represent individuals who are less motivated, less digitally literate, or skeptical of peer-based approaches. Future studies should aim to diversify recruitment by engaging participants from clinical settings, community organizations, and social media platforms to include a broader spectrum of digital experiences and attitudes toward self-help. Second, the regional focus on German-speaking users limits the generalizability of the findings across various cultural and linguistic contexts. Attitudes toward mental health, help-seeking, digital interventions, and peer support may vary substantially among countries, potentially influencing engagement and effectiveness. Therefore, future research might include culturally adapted versions of the app and recruit participants from different linguistic and cultural backgrounds.

XSL•FO
RenderX

Third, while 126 individuals registered for MVP testing, only 86 (68.3%) logged in at least once. Moreover, a substantial proportion did not complete the baseline assessments, and approximately 75% (15/20) of the active users dropped out by T3. This high attrition rate resulted in substantial missing data and reduced the ability to assess long-term usability, engagement, or symptom changes. In addition, the limited follow-up undermined statistical power and internal validity. Researchers should implement targeted retention strategies (eg, regular reminders) to reduce attrition in future studies and ensure good usability. Furthermore, qualitative check-ins or exit interviews could also help capture the reasons for dropout. Fourth, although 63.2% (79/125) of the users joined peer groups for exposure exercises and 83% (104/1125) joined peer groups for activity scheduling, few actively engaged with the group features. This raises concerns about the distinction between initial interest and sustained behavioral engagement. Future studies should explore mechanisms to enhance group involvement, such as onboarding tutorials, group moderators, gamification elements, and tailored content recommendations. Collecting early qualitative feedback on perceived barriers to group participation would help improve peer interaction features and inform design changes that foster stronger community ties. Ultimately, the absence of a control group or randomization prevents causal conclusions regarding the app's effects. Without comparative data, it remains uncertain whether the observed patterns stem from the intervention itself or from external factors. To address this, future research should use RCT designs and consider using mixed methods approaches to better grasp both quantitative and qualitative outcomes.

Conclusions and Implications

This study contributes to the field of digital mental health research by evaluating the strengths and weaknesses of participatory methods in creating a peer-guided self-help app for managing anxiety. Involving users with personal experience in anxiety was beneficial, but the results challenge the assumptions of social identity theory. They reveal that shared digital environments do not automatically foster a sense of community or user engagement without adequate support and moderation. Furthermore, the research emphasizes the importance of consistently comparing user feedback with actual use data and technical capabilities during the app's development. Issues such as high dropout rates, limited understanding of why users leave, and low interaction with features highlight fundamental design and implementation struggles digital self-help tools face. These findings stress the need for well-organized onboarding processes, clear interaction prompts, and transparent moderation policies to minimize user uncertainty. By pinpointing key challenges in process and design, this study offers valuable insights for creating user-centered digital interventions not just for anxiety self-help but also for other peer-supported or community-driven mental health initiatives. Looking ahead, future studies should focus on refining technology, broadening recruitment efforts, and conducting early user testing to improve overall effectiveness.

Acknowledgments

The authors used the generative artificial intelligence tool ChatGPT (OpenAI) for the linguistic refinement of this paper. The authors are solely responsible for the perspectives presented in this paper, which do not necessarily align with their affiliated institutions' views, decisions, or policies. The authors would like to express their gratitude to Pauline Becker, Luisa Bührig, Jana Walzog, Sandra Ewald, and all the test users for their invaluable contributions to shaping this intervention and making this study possible.

Data Availability

The datasets generated or analyzed during this study are available in the Open Science Framework repository [64].

Authors' Contributions

LD wrote the manuscript, TS supervised the application's technical development, and CZ contributed expertise in self-help. TB and JB provided supervisory guidance throughout the study.

Conflicts of Interest

TS owns the software development company responsible for the app's technical development. He acknowledges a potential conflict of interest due to personal and financial stakes in the product's commercialization and sale. To ensure the integrity and objectivity of the research, TS did not participate in evaluating the study results. This measure was taken to minimize potential biases and maintain the scientific validity of the findings. The development and analysis phases were conducted with a commitment to transparency, aiming to disclose any factors that could influence the interpretation of the data or the recommendations provided. While future commercialization remains a possibility, the primary motivation behind this work is to enhance scientific understanding and improve mental health treatment options. TS has voluntarily refrained from any commercial use of the researched minimum viable product. All other authors declare no conflicts of interest.

Multimedia Appendix 1
Semistructured focus group interviews.
[DOCX File , 17 KB - formative_v9i1e62781_app1.docx]

XSL•FO
RenderX

Multimedia Appendix 2
Semistructured individual interviews.
[DOCX File , 18 KB - formative_v9i1e62781_app2.docx]

Multimedia Appendix 3
Coding system.
[DOCX File , 25 KB - formative_v9i1e62781_app3.docx]

References

1. Nochaiwong S, Ruengorn C, Thavorn K, Hutton B, Awiphan R, Phosuya C, et al. Global prevalence of mental health issues among the general population during the coronavirus disease-2019 pandemic: a systematic review and meta-analysis. Scientific Reports 2021 May 13;11(1):10173 [FREE Full text] [doi: 10.1038/s41598-021-89700-8] [Medline: 33986414]
2. BPtK - Newsletter. Bundes Psychotherapeuten Kammer. 2022 Feb. URL: https://api.bptk.de/uploads/B_Pt_K_Newsletter_02_2022_615a6efb95.pdf [accessed 2022-10-05]
3. Konnopka A, König H. Economic burden of anxiety disorders: a systematic review and meta-analysis. Pharmacoeconomics 2020 Jan 24;38(1):25-37. [doi: 10.1007/s40273-019-00849-7] [Medline: 31646432]
4. Burke E, Pyle M, Machin K, Varese F, Morrison AP. The effects of peer support on empowerment, self-efficacy, and internalized stigma: a narrative synthesis and meta-analysis. Stigma Health 2019 Aug;4(3):337-356. [doi: 10.1037/sah0000148]
5. Issakidis C, Andrews G. Service utilisation for anxiety in an Australian community sample. Soc Psychiatry Psychiatr Epidemiol 2002 Apr 18;37(4):153-163. [doi: 10.1007/s001270200009] [Medline: 12027241]
6. Samuel R, Kamenetsky SB. Help-seeking preferences and factors associated with attitudes toward seeking mental health services among first-year undergraduates. Can J High Educ 2022;52(1):30-50. [doi: 10.47678/cjhe.v52i1.189245]
7. Zhang M, Fan C, Ma L, Wang H, Zu Z, Yang L, et al. Assessing the effectiveness of internet-based interventions for mental health outcomes: an umbrella review. Gen Psychiatr 2024 Jul 18;37(4):e101355. [doi: 10.1136/gpsych-2023-101355] [Medline: 39040128]
8. Heinig I, Wittchen HU, Knappe S. Help-seeking behavior and treatment barriers in anxiety disorders: results from a representative German community survey. Community Ment Health J 2021 Nov 20;57(8):1505-1517 [FREE Full text] [doi: 10.1007/s10597-020-00767-5] [Medline: 33471256]
9. Karyotaki E, Efthimiou O, Miguel C, Bermpohl FM, Furukawa TA, Cuijpers P, Individual Patient Data Meta-Analyses for Depression (IPDMA-DE) Collaboration, et al. Internet-based cognitive behavioral therapy for depression: a systematic review and individual patient data network meta-analysis. JAMA Psychiatry 2021 Apr 01;78(4):361-371 [FREE Full text] [doi: 10.1001/jamapsychiatry.2020.4364] [Medline: 33471111]
10. Baumeister H, Reichler L, Munzinger M, Lin J. The impact of guidance on internet-based mental health interventions — a systematic review. Internet Interv 2014 Oct;1(4):205-215. [doi: 10.1016/j.invent.2014.08.003]
11. Berger T, Caspar F, Richardson R, Kneubühler B, Sutter D, Andersson G. Internet-based treatment of social phobia: a randomized controlled trial comparing unguided with two types of guided self-help. Behav Res Ther 2011 Mar;49(3):158-169. [doi: 10.1016/j.brat.2010.12.007] [Medline: 21255767]
12. Furmark T, Carlbring P, Hedman E, Sonnenstein A, Clevberger P, Bohman B, et al. Guided and unguided self-help for social anxiety disorder: randomised controlled trial. Br J Psychiatry 2009 Nov;195(5):440-447. [doi: 10.1192/bjp.bp.108.060996] [Medline: 19880935]
13. Ivanova E, Lindner P, Ly KH, Dahlin M, Vernmark K, Andersson G, et al. Guided and unguided Acceptance and Commitment Therapy for social anxiety disorder and/or panic disorder provided via the internet and a smartphone application: a randomized controlled trial. J Anxiety Disord 2016 Dec;44:27-35. [doi: 10.1016/j.janxdis.2016.09.012] [Medline: 27721123]
14. Krämer R, Köhne-Volland L, Schumacher A, Köhler S. Efficacy of a web-based intervention for depressive disorders: three-arm randomized controlled trial comparing guided and unguided self-help with waitlist control. JMIR Form Res 2022 Apr 04;6(4):e34330 [FREE Full text] [doi: 10.2196/34330] [Medline: 35105536]
15. Zagorscak P, Heinrich M, Sommer D, Wagner B, Knaevelsrud C. Benefits of individualized feedback in internet-based interventions for depression: a randomized controlled trial. Psychother Psychosom 2018 Jan 6;87(1):32-45. [doi: 10.1159/000481515] [Medline: 29306945]
16. Andersson G, Carlbring P, Furmark T, S. O. F. I. E. Research Group. Therapist experience and knowledge acquisition in internet-delivered CBT for social anxiety disorder: a randomized controlled trial. PLoS One 2012;7(5):e37411 [FREE Full text] [doi: 10.1371/journal.pone.0037411] [Medline: 22649526]
17. Boettcher J, Berger T, Renneberg B. Does a pre-treatment diagnostic interview affect the outcome of internet-based self-help for social anxiety disorder? A randomized controlled trial. Behav Cogn Psychother 2012 Oct 17;40(5):513-528 [FREE Full text] [doi: 10.1017/S1352465812000501] [Medline: 22800984]

XSL•FO
RenderX

18. Domhardt M, Geßlein H, von Rezori RE, Baumeister H. Internet- and mobile-based interventions for anxiety disorders: a meta-analytic review of intervention components. Depress Anxiety 2019 Mar 19;36(3):213-224. [doi: 10.1002/da.22860] [Medline: 30450811]
19. Krieger T, Bur OT, Weber L, Wolf M, Berger T, Watzke B, et al. Human contact in internet-based interventions for depression: a pre-registered replication and meta-analysis of randomized trials. Internet Interv 2023 Apr;32:100617 [FREE Full text] [doi: 10.1016/j.invent.2023.100617] [Medline: 37273939]
20. Titov N, Andrews G, Davies M, McIntyre K, Robinson E, Solley K. Internet treatment for depression: a randomized controlled trial comparing clinician vs. technician assistance. PLoS One 2010 Jun 08;5(6):e10939 [FREE Full text] [doi: 10.1371/journal.pone.0010939] [Medline: 20544030]
21. Borkman T, Munn-Giddings C, Boyce M. Self-help/mutual aid groups and peer support. BRP Voluntaristics Rev 2020 Dec 29;5(2-3):1-219. [doi: 10.1163/24054933-12340033]
22. Humphreys K, Rappaport J. Researching self-help/mutual aid groups and organizations: many roads, one journey. Appl Prev Psychol 1994;3(4):217-231. [doi: 10.1016/s0962-1849(05)80096-4]
23. Davidson L, Chinman M, Sells D, Rowe M. Peer support among adults with serious mental illness: a report from the field. Schizophr Bull 2006 Jul;32(3):443-450 [FREE Full text] [doi: 10.1093/schbul/sbj043] [Medline: 16461576]
24. Dennis CL. Peer support within a health care context: a concept analysis. Int J Nurs Stud 2003 Mar;40(3):321-332. [doi: 10.1016/S0020-7489(02)00092-5]
25. Biagianti B, Quraishi SH, Schlosser DA. Potential benefits of incorporating peer-to-peer interactions into digital interventions for psychotic disorders: a systematic review. Psychiatr Serv 2018 Apr 01;69(4):377-388 [FREE Full text] [doi: 10.1176/appi.ps.201700283] [Medline: 29241435]
26. Fenski F, Rozental A, Heinrich M, Knaevelsrud C, Zagorscak P, Boettcher J. Negative effects in internet-based interventions for depression: a qualitative content analysis. Internet Interv 2021 Dec 14;26:100469 [FREE Full text] [doi: 10.1016/j.invent.2021.100469] [Medline: 34754755]
27. Titzler I, Saruhanjan K, Berking M, Riper H, Ebert DD. Barriers and facilitators for the implementation of blended psychotherapy for depression: a qualitative pilot study of therapists' perspective. Internet Interv 2018 Jan 16;12:150-164 [FREE Full text] [doi: 10.1016/j.invent.2018.01.002] [Medline: 30135779]
28. Tajfel H, Turner JC. An integrative theory of intergroup conflict. In: Austin WG, Worchel S, editors. The Social Psychology of Intergroup Relations. Monterey, CA: Brooks/Cole Publishing Company; 1979:33-47.
29. Bandura A, Walters RH. Social Learning Theory (Volume 1). Engelwood Cliffs, NJ: Prentice Hall; 1977.
30. McLean CP, Malek N, Straud CL. A pilot randomized controlled trial of online written exposure therapy delivered by peer coaches to veterans with posttraumatic stress disorder. J Trauma Stress 2024 Jun;37(3):471-482. [doi: 10.1002/jts.23020] [Medline: 38348490]
31. Mummah SA, Robinson TN, King AC, Gardner CD, Sutton S. IDEAS (integrate, design, assess, and share): a framework and toolkit of strategies for the development of more effective digital interventions to change health behavior. J Med Internet Res 2016 Dec 16;18(12):e317 [FREE Full text] [doi: 10.2196/jmir.5927] [Medline: 27986647]
32. Orlowski SK, Lawn S, Venning A, Winsall M, Jones GM, Wyld K, et al. Participatory research as one piece of the puzzle: a systematic review of consumer involvement in design of technology-based youth mental health and well-being interventions. JMIR Hum Factors 2015 Jul 09;2(2):e12 [FREE Full text] [doi: 10.2196/humanfactors.4361] [Medline: 27025279]
33. Ospina-Pinillos L, Davenport TA, Ricci CS, Milton AC, Scott EM, Hickie IB. Developing a mental health eClinic to improve access to and quality of mental health care for young people: using participatory design as research methodologies. J Med Internet Res 2018 May 28;20(5):e188 [FREE Full text] [doi: 10.2196/jmir.9716] [Medline: 29807878]
34. Gonsalves PP, Mittal D, Aluria S, Khan A, Razdan E, Kashyap P, et al. Co-design of "Baatcheet," a peer-supported, web-based storytelling intervention for young people with common mental health problems in India. Glob Ment Health (Camb) 2024 Dec 16;11:e128. [doi: 10.1017/gmh.2024.148] [Medline: 39776987]
35. Bevan Jones R, Stallard P, Agha SS, Rice S, Werner-Seidler A, Stasiak K, et al. Practitioner review: co-design of digital mental health technologies with children and young people. J Child Psychol Psychiatry 2020 Aug;61(8):928-940 [FREE Full text] [doi: 10.1111/jcpp.13258] [Medline: 32572961]
36. Greenhalgh T, Jackson C, Shaw S, Janamian T. Achieving research impact through co-creation in community-based health services: literature review and case study. Milbank Q 2016 Jun;94(2):392-429 [FREE Full text] [doi: 10.1111/1468-0009.12197] [Medline: 27265562]
37. Reddy A, Lester CA, Stone JA, Holden RJ, Phelan CH, Chui MA. Applying participatory design to a pharmacy system intervention. Res Social Adm Pharm 2019 Nov;15(11):1358-1367 [FREE Full text] [doi: 10.1016/j.sapharm.2018.11.012] [Medline: 30509852]
38. Rifkin SB. Examining the links between community participation and health outcomes: a review of the literature. Health Policy Plan 2014 Sep 01;29 Suppl 2(Suppl 2):98-106 [FREE Full text] [doi: 10.1093/heapol/czu076] [Medline: 25274645]
39. Behr S, Fenski F, Boettcher J, Knaevelsrud C, Hammelrath L, Kovacs G, et al. TONI - one for all? Participatory development of a transtheoretic and transdiagnostic online intervention for blended care. Internet Interv 2024 Feb 07;35:100723 [FREE Full text] [doi: 10.1016/j.invent.2024.100723] [Medline: 38370289]

XSL•FO
RenderX

40. Guest G, Namey E, Chen M. A simple method to assess and report thematic saturation in qualitative research. PLoS One 2020 May 5;15(5):e0232076 [FREE Full text] [doi: 10.1371/journal.pone.0232076] [Medline: 32369511]
41. Dresing T, Pehl T. Praxisbuch interview, transkription and analyse: anleitungen und regelsysteme für qualitativ forschende. Deutsche Nationalbibliothek. Marburg: Eigenverlag; 2015. URL: https://d-nb.info/1077320221/34 [accessed 2022-10-10]
42. Brennan RL, Prediger DJ. Coefficient kappa: some uses, misuses, and alternatives. Educ Psychol Meas 1981 Oct;41(3):687-699. [doi: 10.1177/001316448104100307]
43. Spitzer RL, Kroenke K, Williams JB, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
44. Löwe B, Decker O, Müller S, Brähler E, Schellberg D, Herzog W, et al. Validation and standardization of the Generalized Anxiety Disorder Screener (GAD-7) in the general population. Med Care 2008 Mar;46(3):266-274. [doi: 10.1097/MLR.0b013e318160d093] [Medline: 18388841]
45. Bech P, Olsen LR, Kjoller M, Rasmussen NK. Measuring well-being rather than the absence of distress symptoms: a comparison of the SF-36 Mental Health subscale and the WHO-Five Well-Being Scale. Int J Methods Psychiatr Res 2003;12(2):85-91 [FREE Full text] [doi: 10.1002/mpr.145] [Medline: 12830302]
46. Brähler E, Mühlan H, Albani C, Schmidt S. Teststatistische prüfung und normierung der deutschen versionen des EUROHIS-QOL lebensqualität-index und des WHO-5 wohlbefindens-index. Diagnostica 2007 Apr;53(2):83-96. [doi: 10.1026/0012-1924.53.2.83]
47. Beierlein C, Kovaleva A, László Z, Kemper CJ, Rammstedt B. Kurzskala zur erfassung der allgemeinen lebenszufriedenheit (L-1). GESIS. 2015. URL: https://tinyurl.com/5cu6p888 [accessed 2022-09-21]
48. Beierlein C, Kemper CJ, Kovaleva A, Rammstedt B. Short scale for measuring general self-efficacy beliefs (ASKU). Methods Data Anal 2013 Feb 06;7(2):251-278 [FREE Full text] [doi: 10.12758/mda.2013.014]
49. Clarke J, Proudfoot J, Birch MR, Whitton AE, Parker G, Manicavasagar V, et al. Effects of mental health self-efficacy on outcomes of a mobile phone and web intervention for mild-to-moderate depression, anxiety and stress: secondary analysis of a randomised controlled trial. BMC Psychiatry 2014 Sep 26;14:272 [FREE Full text] [doi: 10.1186/s12888-014-0272-1] [Medline: 25252853]
50. Kocalevent RD, Berg L, Beutel ME, Hinz A, Zenger M, Härter M, et al. Social support in the general population: standardization of the Oslo Social Support Scale (OSSS-3). BMC Psychol 2018 Jul 17;6(1):31 [FREE Full text] [doi: 10.1186/s40359-018-0249-9] [Medline: 30016997]
51. Kossen C, Ooi CY. Trialling micro-learning design to increase engagement in online courses. Asian Assoc Open Univ J 2021 Nov 22;16(3):299-310. [doi: 10.1108/aaouj-09-2021-0107]
52. Cuijpers P, van Straten A, Warmerdam L. Behavioral activation treatments of depression: a meta-analysis. Clin Psychol Rev 2007 Apr;27(3):318-326. [doi: 10.1016/J.CPR.2006.11.001]
53. Taylor CT, Lyubomirsky S, Stein MB. Upregulating the positive affect system in anxiety and depression: outcomes of a positive activity intervention. Depress Anxiety 2017 Mar;34(3):267-280 [FREE Full text] [doi: 10.1002/da.22593] [Medline: 28060463]
54. Bandelow B, Wiltink J, Alpers GW, Benecke C, Deckert J, Eckhardt-Henn A, et al. S3-leitlinie behandlung von angststörungen. AWMF Leitlinienregister. 2014. URL: https://register.awmf.org/assets/guidelines/051-028l_S3_Angstst%C3%B6rungen_2014-05_1.pdf [accessed 2022-09-26]
55. Pelled LH, Eisenhardt KM, Xin KR. Exploring the black box: an analysis of work group diversity, conflict and performance. Adm Sci Q 1999;44(1):1-28. [doi: 10.2307/2667029]
56. Donkin L, Hickie IB, Christensen H, Naismith SL, Neal B, Cockayne NL, et al. Rethinking the dose-response relationship between usage and outcome in an online intervention for depression: randomized controlled trial. J Med Internet Res 2013 Oct 17;15(10):e231 [FREE Full text] [doi: 10.2196/jmir.2771] [Medline: 24135213]
57. Hanano M, Rith-Najarian L, Boyd M, Chavira D. Measuring adherence within a self-guided online intervention for depression and anxiety: secondary analyses of a randomized controlled trial. JMIR Ment Health 2022 Mar 28;9(3):e30754 [FREE Full text] [doi: 10.2196/30754] [Medline: 35343901]
58. Powell J, Williams V, Atherton H, Bennett K, Yang Y, Davoudianfar M, et al. Effectiveness and cost-effectiveness of a self-guided internet intervention for social anxiety symptoms in a general population sample: randomized controlled trial. J Med Internet Res 2020 Jan 10;22(1):e16804 [FREE Full text] [doi: 10.2196/16804] [Medline: 31821151]
59. Borghouts J, Eikey E, Mark G, De Leon C, Schueller SM, Schneider M, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. J Med Internet Res 2021 Mar 24;23(3):e24387 [FREE Full text] [doi: 10.2196/24387] [Medline: 33759801]
60. Droge D, Arntson P, Norton R. The social support function in epilepsy self-help groups. Small Group Behav 1986 May;17(2):139-163. [doi: 10.1177/104649648601700202]
61. Jay S, Winterburn M, Choudhary R, Jha K, Sah AK, O'Connell BH, et al. From social curse to social cure: a self‐help group community intervention for people affected by leprosy in Nepal. Community Appl Soc Psychol 2021 Feb 22;31(3):276-287. [doi: 10.1002/casp.2510]

XSL•FO
RenderX

62. Proudfoot J, Parker G, Manicavasagar V, Hadzi-Pavlovic D, Whitton A, Nicholas J, et al. Effects of adjunctive peer support on perceptions of illness control and understanding in an online psychoeducation program for bipolar disorder: a randomised controlled trial. J Affect Disord 2012 Dec 15;142(1-3):98-105. [doi: 10.1016/j.jad.2012.04.007] [Medline: 22858215]
63. Schulz A, Stolz T, Vincent A, Krieger T, Andersson G, Berger T. A sorrow shared is a sorrow halved? A three-arm randomized controlled trial comparing internet-based clinician-guided individual versus group treatment for social anxiety disorder. Behav Res Ther 2016 Sep;84:14-26. [doi: 10.1016/j.brat.2016.07.001] [Medline: 27423374]
64. Duddeck L, Stolz T, Zottl C, Berger T, Boettcher J. With a little help from a-friend: participatory intervention development of a peer-guided self-help app for anxiety disorders. OSF Home. 2024 May 31. URL: https://osf.io/zmj86/ [accessed 2025-06-11]

Abbreviations

ASKU: Short Scale for General Self-efficacy Beliefs
GAD-7: Generalized Anxiety Disorder-7
IBI: internet-based intervention
IDEAS: integrate, design, assess, and share
L-1: General Life Satisfaction Short Scale-1
MHSE: Mental Health Self-Efficacy Scale
MVP: minimum viable product
PD: participatory intervention development
RCT: randomized controlled trial
T0: baseline time point
T1: 4-week time point
T2: 8-week time point
T3: 12-week time point
WHO-5: World Health Organization-Five Well-Being Index

Edited by A Mavragani; submitted 06.06.24; peer-reviewed by R Nooripour; comments to author 28.11.24; revised version received 16.03.25; accepted 20.05.25; published 20.06.25.

Please cite as:
Duddeck L, Stolz T, Zottl C, Berger T, Boettcher J
Participatory Intervention Development of a Peer-Guided Self-Help App for Anxiety Disorders: Mixed Methods Study
JMIR Form Res 2025;9:e62781
URL: https://formative.jmir.org/2025/1/e62781
doi:10.2196/62781
PMID:40540739

©Laura Duddeck, Timo Stolz, Christian Zottl, Thomas Berger, Johanna Boettcher. Originally published in JMIR Formative Research (https://formative.jmir.org), 20.06.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Assessing the Impact of Home Environmental Exposures on Allergic Rhinitis Using Real-Time Air Quality Monitoring and Symptom Assessment: Observational Feasibility Study

Aero Cavalier[1], MA; Anthony I Dick[2], MBBS, MPH; Vickie Johnson II[3], MPH; Emily Cramer[4], PhD; Kamal Eldeirawi[5], RN, PhD; Jayant Pinto[6], MD; Sharmilee M Nyenhuis[1], MD; Victoria S Lee[2], MD

[1]Section of Allergy, Immunology, and Pulmonology, Department of Pediatrics, University of Chicago, Chicago, IL, United States
[2]Department of Otolaryngology, University of Illinois Chicago, Chicago, IL, United States
[3]University of Illinois College of Medicine, Chicago, IL, United States
[4]Department of Biostatistics and Epidemiology, Children's Mercy Hospital, Kansas City, United States
[5]Department of Population Health Nursing Science, University of Illinois Chicago, Chicago, IL, United States
[6]Department of Otolaryngology, University of Chicago, Chicago, United States

Corresponding Author:
Aero Cavalier, MA
Section of Allergy, Immunology, and Pulmonology
Department of Pediatrics
University of Chicago
5841 S Maryland Avenue
Chicago, IL, 60637
United States
Phone: 1 7737950426
Email: aero.cavalier@bsd.uchicago.edu

Abstract

Background: Rhinitis is the most common sinonasal condition and poses a significant cost burden. Rhinitis symptom control is associated with exposure to environmental triggers (eg, aeroallergens, pollutants, and irritants). While people spend much of their time at home, studies examining the association of rhinitis symptoms with home environmental exposures, especially in low-income, urban, and racial or ethnic minorities, are limited. Frequently, 3 types of surveys are used in ecological momentary assessment (EMA): a survey conducted at a predetermined rate, an event-triggered survey, and a follow-up survey to gauge behavioral changes in response to the event.

Objective: This study aims to determine the feasibility and usability of daily and triggered EMA paired with an indoor air quality monitor to collect exposure and rhinitis symptom data.

Methods: Participants were recruited from the Allergy and Ear, Nose, and Throat clinics at 2 academic centers. Participants had to have a rhinitis diagnosis with active symptoms, be 18 years of age or older, self-identify as a racial or ethnic minority, live in the city of Chicago, be able to read and speak English, and have a smartphone. Participants received the Awair Omni air quality monitor to measure volatile organic compounds, particulate matter, and humidity. EMA data were collected using a personal smartphone using the PiLR Health app. Participants were sent daily scheduled surveys, random check-in surveys, and air quality event-triggered survey EMA notifications to assess rhinitis symptoms, environmental exposures, and mitigation strategies for 14 days. After the 14-day data collection period, participants completed the acceptability, appropriateness, and feasibility survey items. Feasibility metrics captured included recruitment and retention, demographics, rhinitis symptoms, and the usability of the PiLR Health App and Awair Omni. Barriers and challenges were identified and captured by the study staff. Descriptive statistics were performed using Excel (Microsoft Corp).

Results: A total of 24 participants were approached, 15 participants consented and 12 participants completed the study. Participants received an average of 62.42 (SD 14.26) total surveys during their study period, and of those surveys, an average of 36.83 (SD 22.18; 59%) surveys were completed. All 12 participants met the threshold for successful home air monitoring (11 days of continuous environmental data assessment). The usability of study components and integration into the overall study was high (usability scale≥68), indicating participants considered each of the devices to be usable. Participant feedback on the study was

XSL•FO
RenderX

positive; yet, they did identify areas for improvement including getting air quality data in real time, providing more detailed instructions for device setup, and doing more check-ins.

Conclusions: A real-time assessment of home environmental exposures and subjective rhinitis symptoms was feasible to conduct. This study will support the development of targeted interventions to address disparities in sinonasal disease care and outcomes.

(JMIR Form Res 2025;9:e73215) doi:10.2196/73215

KEYWORDS

rhinitis; allergic rhinitis; chronic rhinitis; ecological momentary assessment; EMA; air quality; home air quality; mobile phone

Introduction

Background

Allergic rhinitis (AR) is the most common allergic airway disease and is characterized by nasal congestion, runny nose, sneezing, and a reduced sense of smell in response to environmental allergen triggers [1]. The prevalence of this disorder has been rising globally, reaching a range of 12%-20% in high-income countries [2]. About 70% of pharmacy customers purchasing nasal treatments were self-medicating for AR without a diagnosis, and only 44.3% of those with AR symptoms had received a clinical diagnosis [3].

AR has a significant impact on quality of life and is associated with other comorbidities such as asthma, cognitive dysfunction, and depression [4]. Economically, AR poses a significant burden on individuals, health care systems, and society. The direct medical expenditure for AR was estimated to be over US $3.4 billion, mostly from clinic visits and prescription medications [5]. AR incurred the highest indirect costs due to absenteeism or the inability of the individual to function at full capacity in the workplace, with an annual average loss of US $593 per employee, surpassing conditions such as arthritis, migraine, diabetes, and depression [1].

AR is triggered by a range of allergens, most commonly pollen (primarily outdoor), mammal or arthropod-derived allergens, and environmental pollutants, which can be indoor or outdoor [6]. Several studies have shown that indoor air quality (IAQ) in the home is directly associated with the development and severity of AR symptoms [7,8]. The IAQ is determined by factors such as outdoor air infiltration, ventilation systems, and the amount of pollutants generated indoors [9].

The most common indoor pollutants that trigger AR include volatile organic compounds (VOCs), particulate matter (PM), passive tobacco smoke, carbon monoxide, formaldehyde, and greenhouse gases like ozone, sulfur dioxide, and nitrous oxide [10]. These are mostly generated from daily household activities including sanitizing, heating, cooking, burning candles, and smoking [11-13].

The US Environmental Protection Agency estimates that US residents aged 18-65 years spend an average of 80% of the day indoors. A large portion of this indoor time is spent at home [14]. To develop targeted strategies and therapeutic measures to control these environmental factors, it will be vital to analyze them, and how they influence the development and exacerbation of AR.

The predominant methods for the assessment of the relationship between home IAQ data and sinonasal symptoms include periodic patient self-reports on home characteristics and the collection of samples from homes. These are subjective, prone to recall bias, and often capital-intensive [15]. More so, they have no room for real-time evaluation of adjustable home environmental factors and provision of interventions that may reduce exposure to triggers.

Ecological momentary assessment (EMA), also called experience sampling, is a data collection technique where people provide real-time reports on their experience, including their symptoms, actions, and environment [16]. This method usually involves multiple, frequent measurements across a period of time, offering an opportunity for real-time collection of longitudinal data [16]. EMA can inform understanding of the real-time relationship between the environment and symptoms of sinonasal disease. EMA initially used handwritten logs but has since evolved with technology to use automated logs in mobile devices leading to an increased use of EMA in the study of diseases affected by environmental exposures [17,18]. Our prior study, however, showed that its adoption in rhinology studies remains limited [19]. Furthermore, the advent of commercial air monitoring devices, such as the Awair system, presents an opportunity to supplement the EMA symptoms data with continuous home IAQ data [20]. Despite the availability of these tools for real-time assessment of home environmental exposure, no studies have applied them to assess the relationship between home environmental exposures and AR symptoms.

Objectives and Aims

This study aimed to determine the feasibility and usability of (1) EMA to assess self-reported home environmental exposures (eg, aeroallergens or irritants) and rhinitis symptoms (sneezing, itching, and nasal congestion or drainage), and (2) real-time objective measures of home environmental exposures (VOCs, $PM_{2.5}$ [particulate matter with a diameter of ≤2.5 µm], and humidity) using a home air quality monitor (Awair) over a 2-week study period. We assessed the feasibility of using the Awair Omni monitors to collect measures of IAQ such as humidity, VOCs, and $PM_{2.5}$ in real time. This multifaceted approach enables us to notify participants of elevated indoor environmental pollutants while simultaneously gathering real-time data on IAQ and AR symptoms. This paper outlines the feasibility and usability results of this study.

XSL•FO
RenderX

Methods

Ethical Considerations

All studies were conducted in accordance with the Declaration of Helsinki and the International Conference on the Harmonization of Good Clinical Practice guidelines and approved by the relevant institutional review boards at the University of Chicago (IRB #22-1469) and the University of Illinois Chicago (IRB #2022-0361). All participants were required to provide written consent to participate prior to any study procedures. All study data were deidentified before being shared. Participants were given a study ID and only site-specific research teams had access to those codes. Participants who completed all study activities and returned the device were given US $60 via physical or electronic gift card.

Study Components

The Awair Omni IAQ Monitor continuously monitors about 1000 square feet of indoor air for 7 air quality indicators: total VOC (TVOC), $PM_{2.5}$, temperature, humidity, carbon dioxide, ambient light, and ambient noise. This study only focused on humidity, VOCs, and $PM_{2.5}$ levels because of their known associations with rhinitis symptoms. The detectable ranges for the Awair Omni were 0-60,000 ppb/±10% for TVOCs, 0-1000 $\mu g/m^3$/±15 $\mu g/m^3$ or 15% for $PM_{2.5}$, and 0% to 100%/±2% RH for humidity. The acceptable ranges were <333 ppb for TVOCs, <15 $\mu g/m^3$ for $PM_{2.5}$, and >30% to <60% for humidity. The Awair Omni device measures about 4×4×1.3 inches, is Wi-Fi– and Bluetooth-enabled, plugs into a standard alternating current outlet, and includes an 8-hour battery backup. Participants were asked to place their device in either their bedroom or living room in their home. Participants who did not have access to stable Wi-Fi in their homes were ineligible to participate. An air quality reading was taken every 10 seconds and real-time data were uploaded to a dashboard accessed only by research personnel. The Awair Omni proprietary dashboard provided secure communication between the Awair Omni and its dashboard. Data were exported from the dashboard as a .csv file.

EMA data were collected electronically using the PiLR EMA Health software platform, which was installed on the participant's personal smartphone (iOS or Android). Data captured were transmitted to the PiLR EMA platform and stored in a secure database. Participants received three scheduled surveys on their smartphones each day. The first scheduled survey was the Daily Rhinitis Survey that was sent once daily at 9 AM. The next scheduled survey was the Rhinitis Check-In Survey, which was sent two times per day, once in the midlate morning and once in the afternoon. In addition to the scheduled surveys, participants received at least 2 triggered surveys when the air quality monitor values would go outside of acceptable thresholds. The first was the Triggered Air Quality Survey that was sent at the time of the event, and the second survey was the air quality follow-up survey that was sent 45 minutes after the event. The surveys were available for a limited period of time, with the Daily Rhinitis Survey available for 120 minutes, the random Rhinitis Check-In Surveys available for 60 minutes, the initial Triggered Air Quality Survey available for 45 minutes, and the air quality follow-up survey for 60 minutes (Table 1). To reduce the burden of surveys sent, all components of air quality had to return to the recommended thresholds before another air quality survey was triggered. Surveys were also disabled between 10 PM and 5 AM.

Table 1. The types of ecological momentary assessment surveys sent to study participants to assess rhinitis symptoms.

Survey title	Per day	Duration of availability
Daily Rhinitis Survey	1	120 minutes
Rhinitis Check-In Survey	2	60 minutes each
Triggered Air Quality Survey	Per event	45 minutes each
Follow-Up Air Quality Survey	Per event	60 minutes each

Researchers were able to access the PiLR dashboard to monitor participants' use of the EMA app and review their data. Data were exported from the dashboard as a .csv file.

After the 14-day study period, participants were asked to complete an adapted System Usability Scale (SUS) on the Awair Omni, the PiLR EMA Health app, and integration of the 2 systems. Due to an error, the usability of the PiLR EMA Health app and integration of the 2 systems were collected in only half of the participants. The SUS scores were recoded and then analyzed by calculating the mean scores for the individual scales where the scheme assumed 1=strongly agree and 5=strongly disagree. These scores were then corrected so the highest potential score would be 100 and the lowest potential score would be 0.

Participants

Participants were recruited from a convenience sample of patients attending a clinician encounter in the Allergy or Ear, Nose, and Throat (ENT) clinics at the University of Chicago Medical Center or the University of Illinois Hospital and Health System. Participants had to have a diagnosis of rhinitis with active symptoms (sneezing, itching, nasal congestion, and nasal drainage), be 18 years of age or older, self-identified as being from racial or ethnic minority groups (Hispanic, Black, American Indian, Alaskan Native, Asian, Pacific Islander, Native Hawaiian, Arab, or mixed race), live in the city limits of Chicago, be able to read and speak English, and have a smartphone. Participants who had comorbid sinonasal conditions, a sinus tumor, or a terminal illness or immune dysfunction were excluded.

XSL•FO
RenderX

Recruitment

Potential participants were prescreened by looking through electronic health records and looking for age, race, diagnosis of rhinitis, and no history of comorbid sinonasal conditions. Study staff approached a patient, after receiving approval from the clinician, and gave a brief overview of the study. Patients who expressed interest completed a REDCap (Research Electronic Data Capture; Vanderbilt University) survey that collected current contact information and verified eligibility. In the eligibility screener, patients had to confirm that they had their own stable Wi-Fi signal that they could use 24/7 in their homes. Patients were informed that they would receive an electronic gift card and information about their air quality upon completion of the study. Eligible patients were consented onsite after completing the eligibility survey.

Feasibility Measures

To assess the feasibility of this study, we collected measures related to recruitment and retention (Textbox 1). Recruitment and retention were determined by the number of participants recruited and the percentage of participants who completed the study after consenting, respectively. Sociodemographic information, including race and ethnicity, education level, occupation, yearly household income, and health insurance, was collected at baseline using a REDCap form that the participant completed after consent was completed. A 6-item survey called the Rhinitis Control Assessment Test and a 22-item survey called the Sino-Nasal Outcome Test were collected to evaluate different aspects of rhinitis control [21]. These were completed at baseline and the end of the study period in REDCap.

Textbox 1. Feasibility measures assessed in Home Air and Rhinitis study.

Secondary outcomes

- Age, race, sex, primary language, household income (REDCap [Research Electronic Data Capture])
- Rhinitis control assessment and sinonasal outcome test (baseline, day 14; REDCap)

Recruitment rate and eligibility

- Number of people approached, screened, eligible, and ineligible (REDCap)

Participant retention

- Number of participants recruited (REDCap)
- Number of participants who completed 14 days of data collection (REDCap)

Fidelity

- Mean (SD) number of ecological momentary assessment (EMA) surveys sent and completed (PiLR Health Dashboard)
- Mean (SD) number of Awair Omni days with data (Awair Dashboard)

Overall usability

- Day 14 Awair Omni System Usability Scale (SUS; REDCap)
- Day 14 EMA SUS (REDCap)
- Day 14 platform integration (REDCap)

Overall acceptability

- Day 14 interview data

Challenges

- Day 14 interview, study staff notes (REDCap)

Fidelity

Reports were downloaded from the PiLR EMA Health app to determine how many surveys were sent to participants and how many surveys were completed. Additionally, reports were downloaded from the Awair dashboard to assess the number of days the devices were online.

Usability Assessment

After the 14-day data collection period, participants were asked to complete a short REDCap survey assessing usability (SUS) [22]. The SUS asked about participants' experience with the Awair Omni IAQ monitor (10 items), the PiLR EMA Health app the Awair, and the integration of the two types of technology (see SUS in Multimedia Appendix 1). The responses ranged from 1=strongly agree to 5=strongly disagree. After each section, there was an open-ended item for participants to comment on what they liked or did not like about both study tools.

Acceptability Assessment

Adherence to the EMA prompts was measured by the frequency of responses to EMA prompts over the entire study period. This

XSL•FO
RenderX

includes the percentage of complete and incomplete responses to prompts out of the total number of prompts. Adherence to the Awair Omni air monitor was considered successful if participants achieved a minimum of 11 days of continued environment data assessment through Awair by keeping the device on and connected to Wi-Fi.

At the end of the study period, study staff scheduled a time to contact participants via phone or Zoom (Zoom Video Communications) to discuss their experience with the study and go over the close-out procedures. Participants were asked open-ended questions about what they liked about the study, what they disliked about the study, and suggestions for future studies. The study staff member would probe if the participant did not answer a question.

Barriers and Challenges

During follow-up calls between participants and study staff, participants were asked questions about difficulties they have been experiencing with the air monitor or the PiLR EMA Health app. These difficulties were documented by study staff and acknowledged or addressed during larger team meetings. Study staff also encountered difficulties when setting up the device or app or while attempting to collect data. These were also documented.

Results

Participant Characteristics

Participants were on average 37.2 (SD 10.7) years of age (range 20-57 years) and predominantly female (n=8, 67%; Table 2). Participants self-identified as Black (n=7, 47%), Hispanic White (n=3, 20%), Asian (n=2, 13%), Native American (n=1, 7%), or belonging to more than one race (n=2, 13%). At baseline, almost all participants reported uncontrolled rhinitis symptoms (n=14, 92%) or severe sinonasal disease (n=14, 92%) based on their Rhinitis Control Assessment Test (<21) and Sino-Nasal Outcome Test scores (), respectively.

Table 2. Participant characteristics.

Characteristics	Value
Age (years), mean (SD)	37.2 (10.7)
Sex (female), n (%)	8 (67)
Race and ethnicity, n (%)	
American Indian or Alaskan Native	1 (7)
Asian	2 (13)
Black or African American	7 (47)
Hispanic White	3 (20)
More than one race	2 (13)
Language more comfortable speaking, n (%)	
English	14 (93)
Spanish	1 (7)
Yearly household income (US $), n (%)	
<30,000/year	3 (20)
30,000-90,000/year	4 (26)
>90,000/year	3 (20)
Did not answer	5 (33)
RCAT[a] score, mean (SD)	16.6 (4.7)
SNOT[b] score, mean (SD)	51.8 (21.6)

[a]Rhinitis Control Assessment Test.

[b]Sino-Nasal Outcome Test.

Participant Recruitment and Retention

Figure 1 diagrams the study participant flow. A total of 24 patients were approached in the Allergy and ENT clinics, 18 patients were identified through prescreening of clinic schedules and 6 patients were identified through referral from one of the clinic physicians. Of the 24 patients approached, 21 patients expressed interest and were screened for eligibility by study staff. In total, 16 of the screened patients were eligible for the study yet 1 patient declined to consent due to being too busy to participate. A total of 15 participants consented, 3 participants withdrew from one site (reasons not provided), and 12 participants (6 from each site) completed the study.

XSL•FO
RenderX

Figure 1. Recruitment (CONSORT) diagram for home air and rhinitis study. CONSORT: Consolidated Standards of Reporting Trials.

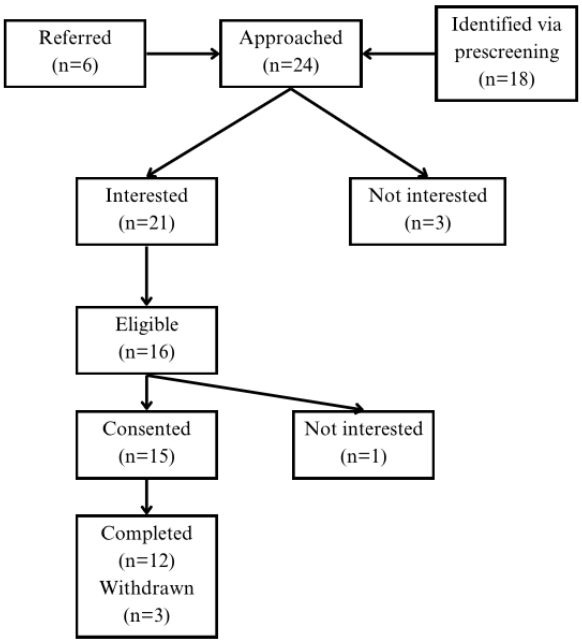

Fidelity

A total of 749 EMA surveys were sent to all participants for the duration of the study (Table 3). Individually, participants received an average of 62.42 (SD 14.26) total surveys during their study period, approximately 4.0 (SD 0.8) per day, and of those surveys, an average of 36.83 (SD 22.18; 59%) surveys were completed over the study period. Daily rhinitis (25.1%) and random rhinitis check-in (50.9%) surveys accounted for the majority of surveys sent to the participants. The triggered air quality and air quality follow-up surveys accounted for 11.9% and 12.2% of surveys sent, respectively. The air quality follow-up survey was the most likely to be completed (67%). Two participants did not receive any air quality surveys due to user issues with the Awair Omni air monitor. These issues included turning the Awair Omni device off or removing the device from Wi-Fi. One participant received surveys but did not complete any surveys.

Table 3. Number of ecological momentary assessment surveys sent to study participants.

	Daily Rhinitis Survey	Random Rhinitis Check-In Survey	Triggered Air Quality Survey	Air Quality Follow-Up Survey	All Surveys
Total surveys, n (%)	188 (25.1)	381 (50.9)	89 (11.9)	91 (12.1)	749 (100)
Number of surveys sent per participant, mean (SD)	15 (3)	31 (6)	7 (6)	7 (6)	62 (14)
Number of surveys started per participant, mean (SD)	11 (5)	18 (10)	5 (6)	6 (6)	42 (20)
Percentage of surveys started per participant (%), mean (SD)	67 (25)	56 (25)	66 (34)	74 (35)	66 (20)
Number of surveys completed per participant, mean (SD)	9 (6)	17 (12)	5 (5)	5 (5)	37 (22)
Percentage of surveys completed per participant (%), mean (SD)	59 (28)	52 (30)	53 (33)	62 (39)	58 (26)

All 12 participants met our threshold for successful home air monitoring, predetermined as a minimum of 11 days of continuous environmental data assessment. In total, 11 (92%) participants had their air monitors operational throughout the entire 14 days of their study period without significant disruption (defined as ≥2 hours of interruption) to continuous assessment of home environment data. One participant, however, had their monitor disconnected from Wi-Fi for 3 days during their study period. The quantitative data for all participants was successfully downloaded for analysis.

Overall Usability

All 12 participants received and completed the SUS for the Awair. In total, 6 out of the 12 participants received and

XSL•FO
RenderX

completed the SUS for the PiLR Health App and the overall integration of study components due to site coordination differences. The mean SUS score for Awair, PiLR Health App, and overall integration was high (≥68; Figure 2), indicating participants considered each of the devices to be usable.

Figure 2. Participant reported usability of the Awair Omni air monitor and the PiLR Health app. Usability was measured by a modified SUS. SUS: System Usability Scale.

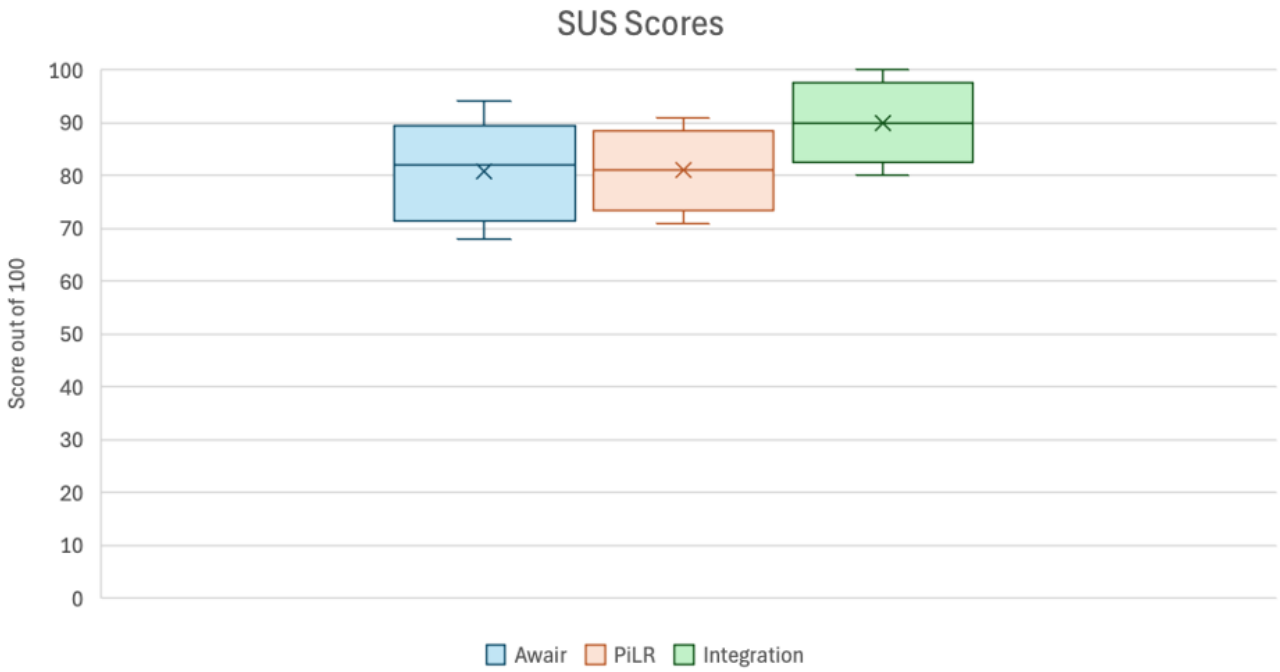

Overall Acceptability

Study staff notes from the 14-day interview recorded what participants liked and disliked about the study (Textbox 2). The most common things participants liked about the study was that the device was easy to use and the surveys were short and easy to complete. The most noted dislike by participants was that the PiLR Health App would send "ghost notifications" where the participant would receive a notification or sometimes the survey would not allow them to move forward, followed by participants saying the study or the instructions for setting up the air monitor were too complex for people who were less tech-savvy. Suggestions for improvement included getting their air quality data in real-time, providing video instructions for the Awair Omni device setup, and doing more frequent check-ins.

XSL•FO
RenderX

Textbox 2. Participant reported feasibility of study components collected via telephone at end of study.

Likes (n)

- The device was easy to use (3)
- Surveys were short, easy to complete (3)
- The study was interesting (2)
- The consistency of the surveys (2)
- The device collecting air quality data (2)
- Liked the survey answers (2)
- Receiving notifications when the air quality changed (1)
- App was simple (1)

Dislikes (n)

- The app was "buggy" (5)
- The study or instructions were complex (4)
- None (3)
- The surveys came at an inconvenient time (3)
- Wording of questions was challenging (2)
- The Awair Omni did not have enough battery time (1)
- The device took up too much space (1)
- Surveys were too repetitive (1)
- Survey answers were not detailed enough (1)

Suggestions for improvement (n)

- Seeing the data in real-time or getting results faster (2)
- More frequent check-ins (2)
- Provide video instructions (1)

Study Challenges

Study challenges were divided into the PiLR EMA Health app and the Awair Omni air monitor. The most common challenge identified was with the PiLR EMA Health App surveys. This primarily included issues with the device assignment logs, which are used to pair the PiLR EMA account to its respective Awair Omni air monitor. This resulted in a participant not receiving triggered surveys when there was a change in air quality noted by the Awair Omni. These issues were resolved by contacting the vendor and collecting information that had been overlooked previously. For the Awair Omni air monitor, the main challenge was in regard to the device shutting down temporarily or disconnecting from Wi-Fi. In one instance, a participant had changed their Wi-Fi password and had made their device disconnect from Wi-Fi. Most situations were resolved by troubleshooting with participants to determine why their devices went offline.

To address these challenges, study staff worked with vendors to troubleshoot issues, and then relayed fixes to participants. Study staff contacted and helped participants correct issues with the devices and provided information to prevent future issues. If there was a significant time of data being lost, determined by the study staff, the study period for the participant was extended to ensure that 14 days of data were collected. These barriers were recorded and modifications to our manual of operations were made to be used in future, adequately powered studies.

Discussion

Principal Findings

This study aimed to assess the feasibility of an air quality monitor and EMA app used simultaneously in a study exploring real-time residential Total VOC and $PM_{2.5}$ exposures and real-time rhinitis symptoms in minoritized adults. We were able to recruit and retain participants in the study and found both the air quality monitor and EMA surveys feasible to implement with all participants keeping the air quality monitor active during the study period and approximately two-thirds of participants completing the surveys. Additionally, while we did identify some technical challenges in the pilot study, we were able to develop solutions to address the challenges and are prepared to conduct a larger, appropriately powered observational study to collect further information on rhinitis symptoms.

The field of measuring symptoms and air quality in real time in rhinitis is a nascent area, especially in minoritized populations that often have higher rates of exposure to rhinitis triggers [23].

XSL•FO
RenderX

Prior studies in real-time monitoring and rhinitis reviewed the existing research on EMA as a health monitoring tool and the use of ecological monitoring in understanding rhinitis symptoms and air quality both indoors and outdoors for individuals with asthma, but none focused on minoritized populations [19,24]. This is the only study to our knowledge that used EMA for real-time rhinitis symptoms and residential air quality monitoring in minoritized adults.

We did encounter technical challenges in regard to the PiLR EMA Health app and Awair Omni air monitor. Common challenges were participants not receiving Triggered Air Quality Surveys or the air monitor disconnecting from Wi-Fi. These challenges were technical in nature and were anticipated as this was a system never tested before in this patient population. This may have occurred at a higher rate in our participant population as Black (36%) and Hispanic (30%) households are less likely to have broadband internet than White households (21%) [25].

As research and medical care are leaning into technology more, it is important to acknowledge that there may be challenges in certain populations that could impact their ability to successfully participate in research or get medical care. Although our usability scale showed the Awair Omni air monitor and the PiLR EMA Health app to be acceptable for our participants, future research should be done that includes older participants, low-income populations, and people in digital deserts. Digital deserts are areas where there is a lack of broadband Wi-Fi, a lack of electronic devices (eg, tablets, smartphones, and personal computers), or low technological literacy. Some studies suggest that these digital deserts are primarily due to low digital literacy, language barriers, and affordability [26]. Some studies also suggest that there may be discrimination in broadband service delivery based on race and geographic location [27]. This implies that there may be a relationship between racially minoritized people having access to broadband and their ability to treat asthma symptoms. Future studies may want to consider providing participants compensation for their Wi-Fi use or providing a hotspot for the duration of their study participation.

Limitations

Limitations to our study include a lack of broad generalizability as participants were recruited from a convenience sample of patients from Allergy and ENT clinics in urban academic centers. Further, older adults were not represented in this study sample which may have been due to the heavy technology component of our study and the requirement for Wi-Fi and a smartphone. Future studies should offer Wi-Fi or smartphones for those who do not have access to these and provide additional technology training to those requesting assistance.

Conclusions

This was the first study to use EMA and air quality monitors to monitor rhinitis symptoms in an urban and racially minoritized population. The findings from this study will provide insight into developing research using EMA to monitor rhinitis symptoms and using EMA for studies pertaining to low-income and racially minoritized populations, or other groups, with low technological literacy.

Acknowledgments

VSL is supported by the National Institutes of Health (grant K12AR084225). This study was supported by a Triological Society Research Career Development Award (Principal Investigator: VSL) and the Chicago Center for Health and Environment CACHET Pilot grant (P30 ES027792; principal investigator: SMN).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

JP is part of the speaker's board/advisory board for Sanofi/Regeneron, Optinose, InsMed, and is the site investigator for Sanofi, Regeneron, and Lyra. The other authors declare no conflict of interest.

Multimedia Appendix 1
System Usability Scales.
[DOCX File , 22 KB - formative_v9i1e73215_app1.docx]

References

1. Lamb CE, Ratner PH, Johnson CE, Ambegaonkar AJ, Joshi AV, Day D, et al. Economic impact of workplace productivity losses due to allergic rhinitis compared with select medical conditions in the United States from an employer perspective. Curr Med Res Opin 2006;22(6):1203-1210. [doi: 10.1185/030079906X112552] [Medline: 16846553]
2. Bauchau V, Durham S. Prevalence and rate of diagnosis of allergic rhinitis in Europe. Eur Respir J 2004;24(5):758-764 [FREE Full text] [doi: 10.1183/09031936.04.00013904] [Medline: 15516669]
3. Tan R, Cvetkovski B, Kritikos V, Price D, Yan K, Smith P, et al. Identifying the hidden burden of allergic rhinitis (AR) in community pharmacy: a global phenomenon. Asthma Res Pract 2017;3:8 [FREE Full text] [doi: 10.1186/s40733-017-0036-z] [Medline: 29201385]

XSL•FO
RenderX

4. Bernstein J, Bernstein J, Makol R, Ward S. Allergic rhinitis: a review. JAMA 2024;331(10):866-877. [doi: 10.1001/jama.2024.0530] [Medline: 38470381]
5. Stempel DA, Woolf R. The cost of treating allergic rhinitis. Curr Allergy Asthma Rep 2002;2(3):223-230. [doi: 10.1007/s11882-002-0023-0] [Medline: 11918864]
6. Naclerio R, Ansotegui IJ, Bousquet J, Canonica GW, D'Amato G, Rosario N, et al. International expert consensus on the management of allergic rhinitis (AR) aggravated by air pollutants: impact of air pollution on patients with AR: current knowledge and future strategies. World Allergy Organ J 2020;13(3):100106 [FREE Full text] [doi: 10.1016/j.waojou.2020.100106] [Medline: 32256939]
7. Wu AC, Dahlin A, Wang AL. The role of environmental risk factors on the development of childhood allergic rhinitis. Children (Basel) 2021;8(8):708 [FREE Full text] [doi: 10.3390/children8080708] [Medline: 34438599]
8. Blanc PD, Eisner MD, Katz PP, Yen IH, Archea C, Earnest G, et al. Impact of the home indoor environment on adult asthma and rhinitis. J Occup Environ Med 2005;47(4):362-372. [doi: 10.1097/01.jom.0000158708.32491.9d] [Medline: 15824627]
9. Cecchi L, D'Amato G, Annesi-Maesano I. External exposome and allergic respiratory and skin diseases. J Allergy Clin Immunol 2018;141(3):846-857. [doi: 10.1016/j.jaci.2018.01.016] [Medline: 29519451]
10. Vijayan V, Paramesh H, Salvi S, Dalal AK. Enhancing indoor air quality—the air filter advantage. Lung India 2015;32(5):473-479 [FREE Full text] [doi: 10.4103/0970-2113.164174] [Medline: 26628762]
11. Ambient (outdoor) air pollution. World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health [accessed 2024-10-24]
12. Particulate matter (PM) basics. US Environmental Protection Agency. URL: https://www.epa.gov/pm-pollution/particulate-matter-pm-basics [accessed 2022-04-25]
13. Korpi A, Järnberg J, Pasanen A. Microbial volatile organic compounds. Crit Rev Toxicol 2009;39(2):139-193. [doi: 10.1080/10408440802291497] [Medline: 19204852]
14. US Environmental Protection Agency. Exposure factors handbook. URL: https://www.epa.gov/expobox/about-exposure-factors-handbook [accessed 2024-06-12]
15. Chew GL, Burge HA, Dockery DW, Muilenberg ML, Weiss ST, Gold DR. Limitations of a home characteristics questionnaire as a predictor of indoor allergen levels. Am J Respir Crit Care Med 1998;157(5 Pt 1):1536-1541. [doi: 10.1164/ajrccm.157.5.9708011] [Medline: 9603135]
16. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol 2008;4:1-32. [doi: 10.1146/annurev.clinpsy.3.022806.091415] [Medline: 18509902]
17. Williams MT, Lewthwaite H, Fraysse F, Gajewska A, Ignatavicius J, Ferrar K. Compliance with mobile ecological momentary assessment of self-reported health-related behaviors and psychological constructs in adults: systematic review and meta-analysis. J Med Internet Res 2021;23(3):e17023 [FREE Full text] [doi: 10.2196/17023] [Medline: 33656451]
18. Burton C, Weller D, Sharpe M. Are electronic diaries useful for symptoms research? A systematic review. J Psychosom Res 2007;62(5):553-561. [doi: 10.1016/j.jpsychores.2006.12.022] [Medline: 17467410]
19. Schusteff RA, Chervinko MA, Nyenhuis SM, Lee VS. Ecological momentary assessment of outcomes in allergic rhinitis and chronic rhinosinusitis: a review. Int Forum Allergy Rhinol 2022;12(10):1282-1290. [doi: 10.1002/alr.22982] [Medline: 35092173]
20. Polivka B, Krueger K, Bimbi O, Huntington-Moskos L, Nyenhuis S, Cramer E, et al. Integrating real-time air quality monitoring, ecological momentary assessment, and spirometry to evaluate asthma symptoms: usability study. JMIR Form Res 2024;8:e60147 [FREE Full text] [doi: 10.2196/60147] [Medline: 39388233]
21. Hopkins C, Gillett S, Slack R, Lund VJ, Browne JP. Psychometric validity of the 22-item sinonasal outcome test. Clin Otolaryngol 2009;34(5):447-454. [doi: 10.1111/j.1749-4486.2009.01995.x] [Medline: 19793277]
22. Brooke J. SUS: a quick and dirty usability scale. ResearchGate. 1995. URL: https://www.researchgate.net/publication/228593520_SUS_A_quick_and_dirty_usability_scale [accessed 2021-11-21]
23. Burbank AJ, Hernandez ML, Jefferson A, Perry TT, Phipatanakul W, Poole J, et al. Environmental justice and allergic disease: a work group report of the AAAAI environmental exposure and respiratory health committee and the diversity, equity and inclusion committee. J Allergy Clin Immunol 2023;151(3):656-670 [FREE Full text] [doi: 10.1016/j.jaci.2022.11.025] [Medline: 36584926]
24. Nyenhuis S, Cramer E, Grande M, Huntington-Moskos L, Krueger K, Bimbi O, et al. Utilizing real-time technology to assess the impact of home environmental exposures on asthma symptoms: protocol for an observational pilot study. JMIR Res Protoc 2022;11(8):e39887 [FREE Full text] [doi: 10.2196/39887] [Medline: 35916686]
25. Sanders CK, Scanlon E. The digital divide is a human rights issue: advancing social inclusion through social work advocacy. J Hum Rights Soc Work 2021;6(2):130-143 [FREE Full text] [doi: 10.1007/s41134-020-00147-9] [Medline: 33758780]
26. Li Y, Spoer BR, Lampe TM, Hsieh PY, Nelson IS, Vierse A, et al. Racial/ethnic and income disparities in neighborhood-level broadband access in 905 US cities, 2017-2021. Public Health 2023;217:205-211 [FREE Full text] [doi: 10.1016/j.puhe.2023.02.001] [Medline: 36917875]
27. Reddick CG, Enriquez R, Harris RJ, Sharma B. Determinants of broadband access and affordability: an analysis of a community survey on the digital divide. Cities 2020;106:102904 [FREE Full text] [doi: 10.1016/j.cities.2020.102904] [Medline: 32921864]

XSL•FO
RenderX

Abbreviations

AR: allergic rhinitis
EMA: ecological momentary assessment
ENT: Ear, Nose, and Throat
IAQ: indoor air quality
PM: particulate matter
PM2.5: particulate matter with a diameter of ≤2.5 µm
REDCap: Research Electronic Data Capture
SUS: System Usability Scale
TVOC: total volatile organic compound
VOC: volatile organic compound

Edited by A Mavragani; submitted 27.02.25; peer-reviewed by L Huntington-Moskos; comments to author 31.03.25; revised version received 21.04.25; accepted 20.05.25; published 23.06.25.

Please cite as:
Cavalier A, Dick AI, Johnson II V, Cramer E, Eldeirawi K, Pinto J, Nyenhuis SM, Lee VS
Assessing the Impact of Home Environmental Exposures on Allergic Rhinitis Using Real-Time Air Quality Monitoring and Symptom Assessment: Observational Feasibility Study
JMIR Form Res 2025;9:e73215
URL: https://formative.jmir.org/2025/1/e73215
doi:10.2196/73215
PMID:40550125

©Aero Cavalier, Anthony I Dick, Vickie Johnson II, Emily Cramer, Kamal Eldeirawi, Jayant Pinto, Sharmilee M Nyenhuis, Victoria S Lee. Originally published in JMIR Formative Research (https://formative.jmir.org), 23.06.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Evaluation of a Novel eHealth Tool for Pulmonary Rehabilitation in People With Chronic Obstructive Pulmonary Disease: Randomized Controlled Pilot and Feasibility Trial

Åsa Karlsson[1], PhD; Pernilla Sönnerfors[2,3], PhD; Sara Lundell[1], PhD; Annika Toots[1], PhD; Karin Wadell[1], PhD

[1]Department of Community Medicine and Rehabilitation, Umeå University, Umeå, Sweden

[2]Department of Medicine Solna, Division of Immunology and Respiratory Medicine, Center for Molecular Medicine, Karolinska Institutet, Stockholm, Sweden

[3]Medical Unit Occupational therapy and Physiotherapy, Womens´s Health and Allied Health Professionals Theme, Karolinska University Hospital, Stockholm, Sweden

Corresponding Author:
Åsa Karlsson, PhD
Department of Community Medicine and Rehabilitation
Umeå University
Linnaeusväg 15
Umeå, 901 87
Sweden
Phone: 46 907869887
Email: asa.karlsson@umu.se

Abstract

Background: There is a growing interest in eHealth solutions to enhance access to and use of pulmonary rehabilitation for people with chronic obstructive pulmonary disease (COPD).

Objective: This study aims to evaluate the feasibility of a novel eHealth tool (Me&COPD) to support pulmonary rehabilitation concerning usability, exercise adherence, intensity, progression, and adverse events. Moreover, this study aims to evaluate clinical outcome measures to prepare for a future larger trial.

Methods: A multicenter, parallel-group randomized controlled pilot and feasibility trial was conducted in 6 primary health care centers. People with mild to severe COPD were recruited by physiotherapists at the included health care centers and randomized either to the intervention group with access to Me&COPD for 3 months or to the control group receiving usual care. The Me&COPD tool comprised audio-visual and written self-management strategies, including an individually tailored home-based exercise program and interaction with a physiotherapist. The exercise program was prescribed in a face-to-face meeting with a physiotherapist, and thereafter it was regularly reviewed and adjusted through the eHealth tool. The primary outcome, usability, was self-assessed at intervention completion in the intervention group and among participating physiotherapists (n=7) using the Swedish version of the Mobile Health App Usability Questionnaire (S-MAUQ). In addition, use data on exercise adherence, intensity, and progression and adverse events were exported from the eHealth tool. Clinical outcomes, assessed by blinded assessors at baseline and 3 months in the intervention and control groups, included exercise capacity, balance, physical activity level, COPD-related symptoms, and health-related quality of life. Descriptive statistics were used for analysis.

Results: In total, 22 participants (women: n=12, 55%), aged 72.3 (SD 8.4) years on average, were included in the intervention (n=15) and control (n=7) groups. The mean overall S-MAUQ scores out of 7 (highest possible usability) were 4.4 (SD 1.5) for participants and 4.5 (SD 1.2) for physiotherapists. Among the subscales, the highest score was assigned to *usefulness* among participants (S-MAUQ: mean 4.9, SD 1.3) and physiotherapists (S-MAUQ: mean 5.1, SD 1.7). No severe adverse events were registered, although exercise adherence, intensity, and progression evaluation were limited by incomplete exercise session registration. The test procedures and the clinical outcome measures used were found to be feasible for the participants and the assessors.

Conclusions: The novel eHealth tool, Me&COPD, seemed feasible in terms of safety and had acceptable usability among people with COPD and participating physiotherapists. Usability may be improved by better organization of the information and simplification of the exercise diary to enable collection of data on exercise adherence, intensity, and progression through the

XSL•FO
RenderX

eHealth tool. The test procedures seemed feasible, although the recruitment process needs further consideration. The effectiveness of the intervention remains to be evaluated in a future larger trial.

Trial Registration: ClinicalTrials.gov NCT05086341; https://clinicaltrials.gov/study/NCT05086341

(JMIR Form Res 2025;9:e68195) doi:10.2196/68195

KEYWORDS

chronic obstructive pulmonary disease; COPD; telerehabilitation; eHealth; usability testing; user-centered design; pulmonary rehabilitation; eHealth tool; feasibility trial; randomized controlled trial; RCT

Introduction

Background

Chronic obstructive pulmonary disease (COPD) is a complex and heterogeneous disease that not only involves airflow limitation but also multiple extrapulmonary manifestations, such as muscle dysfunction, decreased exercise capacity and balance, cardiovascular disease, fatigue, and depression [1]. Fortunately, there is an established nonpharmacological intervention, pulmonary rehabilitation (PR), that can address these systemic consequences and improve the health of people with COPD [2]. PR is defined as a comprehensive, individualized intervention including but not limited to physical exercise, education, and behavioral change that aims to improve the individual's physical and psychological condition and promote long-term adherence to health-enhancing behaviors [3]. There is conclusive evidence of the positive effects of PR on exercise capacity, symptoms, health-related quality of life, and reduced health care use [2-5]. The intervention has proven to be cost-effective in different settings [6]. Despite strong recommendations in national as well as international guidelines [3,4,7-9], the accessibility of PR for people with COPD is limited [10]. According to Desveaux et al [10], the available PR programs in 7 different countries could only serve less than 1.2% of the population with COPD in these countries, with Sweden having the lowest annual capacity of 0.2%. In addition, central barriers to participation in PR are distance to PR centers, limited knowledge of benefits, and fluctuating health status [11,12], emphasizing a need for novel strategies to provide person-centered care and improve access to PR among people with COPD [13].

eHealth solutions, that is, the use of information and communication technologies in health services [14], such as telerehabilitation, represent a promising alternative strategy to improve access to and uptake of PR [15,16]. The COVID-19 pandemic has further highlighted the need for rehabilitation services provided at a distance and accelerated their development. Telerehabilitation for people with COPD, provided at home or at an outpatient health care facility, has been shown to be safe with comparable effects on exercise capacity, quality of life, and symptoms as in-person, center-based PR [15,17,18]. Furthermore, home-based telerehabilitation interventions have been reported to increase the likelihood of program completion [15] and be more effective than usual care in terms of improvements in dyspnea [18] and health status [18,19]. However, to be able to implement an eHealth solution in clinical practice, it is crucial that the intervention is accepted by those who are going to use it and the intervention fits in the context where it will be implemented [20]. A previous study reported that the users, that is, people with COPD, relatives, health care providers (HCPs), and a patient organization representative, had a positive attitude toward using an eHealth tool to support self-management of physical activity and exercise training [21]. However, they expressed that eHealth solutions should be seen as a complement to already existing rehabilitation options. Furthermore, Tsai et al [22] reported that a home-based exercise program using real-time videoconferencing technology was well accepted by the study participants, and the informants taking part in a supervised telerehabilitation program expressed that the intervention provided autonomy support [23]. Moreover, eHealth tools have been voiced to support and enhance self-management of the disease [24,25]. However, it has also been found that using digital technology is not suitable for everyone, such as for individuals who lack motivation, are fearful of new information, and are not comfortable with IT [26]. This was emphasized in the study by Cox and Holland [27] who concluded that one size does not fit all, and more research is needed on different models of telerehabilitation in people with COPD to reach evidence-based consensus and increase possibilities for individualization [16,28].

To promote health, provide individualized treatment, and increase access to evidence-based treatment for people with COPD in Sweden, we developed an eHealth tool for PR, Me&COPD (Swedish: Min KOL). The tool was developed in a structured cocreation process with intended end users [29], which was not common in the development of previous eHealth tools [28]. The use of the cocreation process may be beneficial for the implementation of the intervention [30] and result in more problem-based research, increased contextual relevance, and greater impact [31]. The content of Me&COPD aligns with the international guidelines for PR in COPD [32], the national guidelines for COPD care in Sweden [4], and with reported requirements and needs of people with COPD and HCPs [21,33,34]. The Me&COPD tool was built in collaboration with health care region Västerbotten on 1177 Care guide—Support and Treatment (Swedish: 1177 Vårdguiden—Stöd och Behandling) platform, an established, national medical information system that provides a platform for various internet-based support and treatment programs. However, none of the platform's existing programs have targeted PR in people with COPD.

Aims

The aim of this trial was to evaluate the feasibility of Me&COPD in terms of usability, exercise adherence, intensity, progression, and adverse events. An additional aim was to

XSL•FO
RenderX

evaluate clinical outcome measures to prepare for a future larger trial.

Methods

Trial Design

We conducted a multicenter, parallel-group randomized controlled pilot and feasibility trial, which was reported according to the CONSORT (Consolidated Standards of Reporting Trials) statement for pilot- and feasibility trials [35], the eHealth checklist [36], and the TIDieR (Template for Intervention Description and Replication) checklist [37]. The trial was prospectively registered at ClinicalTrials.gov (ID NCT05086341).

Ethical Considerations

The trial received ethics approval from the Swedish Ethical Review Authority (Dnr: 2020-01693 and 2021-03538) and was conducted according to the Declaration of Helsinki. Participants were informed about the trial by telephone. If oral consent was obtained, a package encompassing the details of the trial procedures, their right to withdraw from the study without prejudice, a consent form, and baseline questionnaires was posted to the participants and collected at the baseline assessment. All participants provided written informed consent to participate in this trial. To ensure confidentiality of the participants, data were stored encrypted at a secure server at the Umeå University with access restricted to the research team. The security solution, Cryptshare (Pointsharp Secure Information Exchange solution), was used for secure data transfer between the participating health care regions and the researchers at the Umeå University. Participants were offered reimbursement for their travel expenses, but no other monetary compensation was offered.

Settings, Recruitment, and Participants

The trial was conducted between November 2021 and June 2023 at 6 primary health care centers (PHCs). PHCs are the first point of access to health care for many, especially those with chronic diseases such as COPD. PHCs offer services, including assessment, treatment, prevention, and rehabilitation without the need for a referral. Most PHCs in Sweden are multiprofessional, including experts such as general practitioners, nurses, occupational therapists, and physiotherapists. The PHCs in this trial were situated in big and medium-sized cities in central and northern parts of Sweden. Publicly funded as well as private alternatives were included. PHCs in 2 health care regions were included at trial commencement. Due to the low recruitment rate, a third health care region was included, and the trial period was extended by 6 months. To facilitate recruitment of PHCs into the trial, 1 member of the research group worked exclusively with approaching staff at eligible PHCs, that is, physiotherapists experienced in COPD, their managers, and COPD nurses using personal letters, emails, and telephone. Furthermore, the research group spread information about the trial by participating in digital meetings with the regional administrations and COPD-related education events for HCPs. Potential participants, that is, adults with COPD who were in a stable condition [32], were identified by physiotherapists at the participating PHCs using their patient lists. The physiotherapists asked for their approval to let a researcher contact them for trial information and delivered the intervention for those who were randomized to the intervention group.

Inclusion criteria were (1) a confirmed diagnosis of COPD, (2) the ability to read and understand Swedish, and (3) the absence of severe comorbidity that could be considered as the main contributing factor for limitation in physical activity. The trial had no definite inclusion criteria for computer literacy, and if a participant did not have access to a computer, tablet, or smartphone, there were tablets available to borrow from the research group. In case of a COPD exacerbation, the participant had to wait 6 weeks from the start of pharmacological treatment before being eligible for the trial.

Randomization

Following baseline assessment, participants were randomized either to the intervention (Me&COPD) or the control group (usual care) in a 2:1 allocation ratio with stratification for sex. A computer-generated randomization list was used, and the results were stored in sealed envelopes. A researcher, not involved in data collection, administered the randomization and assigned participants to interventions.

Intervention

In addition to usual care, participants in the intervention group had access to the Me&COPD tool during the 3-month intervention period. The intervention was provided on the 1177 Support and Treatment platform, which was accessed by the participants through a secure log-in using a national electronic identification system. The intervention included components such as a physical activity plan, an individualized exercise program, and educational texts and films to support self-management strategies. Components important for behavioral change, such as personal goal setting and feedback, were also included [38]. The exercise program comprised warm-up exercises and 5 muscle strength or endurance exercises, targeting the shoulders, thighs, and calves, as well as balance and aerobic exercises, all with an individually adapted degree of difficulty and intensity and emphasis on safe performance. The exercise program was to be performed at least two times per week across the intervention period, that is, a minimum of 24 sessions. To monitor exercise intensity, the participants were instructed to report their exercise experience and symptoms (dyspnea and muscle fatigue ratings) for each exercise on the Borg CR10 scale [39] in an exercise diary after each session. If the participants had not accessed Me&COPD in the last 7 days, they received a reminder (email or SMS text message) to log in to the eHealth tool. An overview of the content and features of the intervention is presented in Figure 1.

XSL•FO
RenderX

Figure 1. An overview of the content and features of the eHealth tool Me&COPD. COPD: chronic obstructive pulmonary disease.

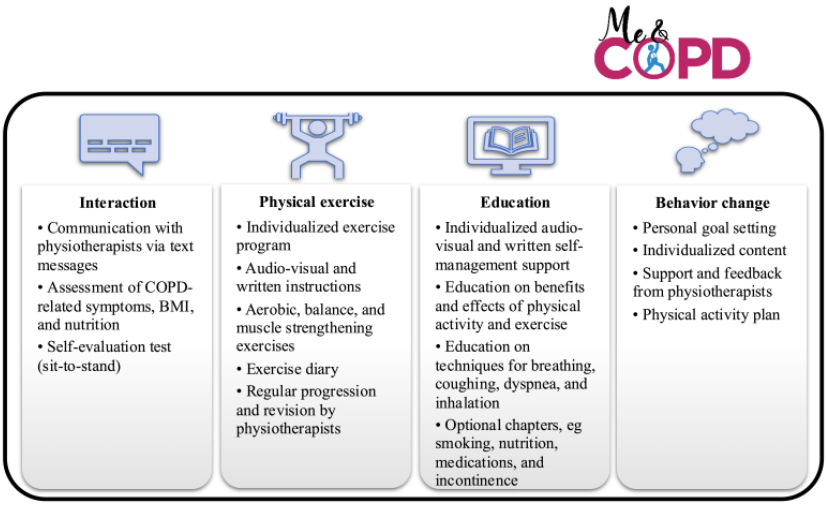

During an initial in-person meeting at the PHC with their physiotherapist, participants received detailed information on how to use Me&COPD as well as a booklet about the log-in procedure of the 1177 Support and Treatment platform. In addition, the individualized exercise program and physical activity plan were set up. The remaining part of the intervention was performed in the participants' homes with the possibility to communicate with the physiotherapists through the eHealth tool. During the intervention period, the physiotherapists were advised to log in to Me&COPD regularly, that is, at least once a week to answer any messages from the participants and every other week to review and adjust or further develop the exercise program when needed, based on the information provided from participants' symptom assessments in the exercise diary. The physiotherapists were recommended to shift from continuous- to interval-based aerobic training after 4 weeks and prescribe muscle strength exercises no longer than 4 weeks before shifting to muscle endurance exercises for the remainder of the intervention, if possible [40,41]. The participants were instructed to use ordinary contact paths for health care in case of acute illness. If a COPD exacerbation occurred during the intervention, access to Me&COPD could be extended by up to 2 weeks. The participants were allowed to take part in other physical activities or interventions involving physical activity during the trial period.

Technical support was available throughout the intervention period for both participants and physiotherapists via the 1177 Support and Treatment platform and the study coordinator. Before study commencement, the physiotherapists also received a digital training session with the study coordinator and a representative from the 1177 Support and Treatment platform as well as a detailed, written user manual and practical exercises to learn how to use Me&COPD. The content of the intervention was not changed during this trial.

Control

All participants in the control group received the usual care alone. Usual care is recommended to include (but is not restricted to) the use of long-acting anticholinergics and long-acting β2-agonists with 24-hour duration and support for smoking cessation, physical activity and exercise, self-management, and nutrition [4,7]. There were no restrictions on those allocated to the control group to participate in other physical activities or physical activity interventions during the trial period.

Outcomes

All participants were invited to in-person test visits at baseline and at the end of the intervention, that is, at 3 months, at a university or a health care facility. Two researchers (physiotherapists) blinded to group allocation performed the assessments. Usability was the primary outcome of the trial, measured at intervention completion in the intervention group and in participating physiotherapists at the included PHCs, using the Mobile Health App Usability Questionnaire (MAUQ; the interactive version for patients and HCPs) [42]. The patient versions of the questionnaire have been found to have strong validity and reliability [42]. The questionnaire comprises 21 statements on a 7-point Likert scale ranging from 1 (strongly disagree) to 7 (strongly agree), with an additional "not applicable" option available for all items (maximum 147 points). The scale can be divided into 3 subscales: *ease of use and satisfaction* (items 1-8), *system information arrangement* (items 9-14), and *usefulness* (items 15-21). The average of the responses to all statements is calculated for the entire scale as well as for each subscale where a mean score of 7 indicates the best possible usability. In this trial, we used a Swedish version of the MAUQ (S-MAUQ) where the patient version had space in the end for comments. The questionnaire was translated into Swedish using a forward-backward translation procedure by native speakers in the respective language [43]. Two of the researchers translated the English version to Swedish, and 1 independent person without a medical background performed a back translation to English. The research group reached consensus on the final version after discussions and review by a representative at the unit for IT support and system development at the Umeå University. To evaluate exercise adherence, intensity, progression, and adverse events, use data were exported from the eHealth tool.

Clinical outcomes, assessed using valid and reliable tests and questionnaires for people with COPD, were collected to prepare for a future larger trial. Exercise capacity was measured by the

XSL•FO
RenderX

6-minute walk test [44], the 1-minute sit-to-stand test [45], and the unsupported upper limb exercise test [46]. Balance was measured by the timed up and go test [47]. Physical activity level was self-reported using a questionnaire from the Swedish National Board of Health and Welfare [48]. The mean number of steps per day was objectively measured over 7 consecutive days using accelerometers validated in people with COPD (DynaPort; McRoberts BV). The assessors attached the accelerometer at the end of the in-person test visit at baseline and at the 3-month follow-up, and the measurement started later the same day. Measurements with ≥4 valid weekdays were included, that is, weekends and weekdays with <8 hours of daytime wear time of the accelerometer were excluded [49]. In addition, COPD-related symptoms were measured using the Modified Medical Research Council dyspnea scale [50] and the COPD Assessment Test [51]. Health-related quality of life was measured by the St George's Respiratory Questionnaire [52] and health status by EQ-5D [53]. The frequency of COPD-related health care contacts during the trial was reported by the participants.

Baseline descriptive assessments included a lung function test (spirometry) according to guidelines [54]. Demographic and anthropometric measurements included age, sex, height, weight, and BMI. All participants were asked to report on living conditions, educational level, employment, smoking status, medications, comorbidities, and COPD-related health care contacts in the past year. Descriptive characteristics of the participating physiotherapists included age, number of years in the profession, and educational level.

Sample Size and Data Analysis

Because this was a pilot and feasibility trial, no formal sample size calculation was performed. The sample size estimation was based on an expected medium standardized mean difference (medium target effect size of $0.3 \leq d < 0.7$) between the intervention and control groups on the 6-minute walk test in an 80% powered future definitive randomized controlled trial, suggesting that about 10 participants per group were needed [55,56]. However, to gain more knowledge about the feasibility of the intervention, we aimed to allocate 20 participants to the intervention group and 10 to the control group (30 participants in total). Data were analyzed according to the intention-to-treat principle using available data from all participants, according to their original allocation and regardless of level of attendance. Descriptive statistics with estimates were used for analysis [35]. Data are reported as means, SD, or frequencies (percentages) in this paper. For clinical outcomes with continuous data, a paired 2-tailed *t* test was used to describe within-group changes (mean and SD) between baseline and 3 months. An independent sample *t* test was used to determine if the mean changes were different between the groups (mean and 95% CI). Statistical analyses were conducted using SPSS Statistics (version 29.0.1.0; IBM Corp). The outcome usability was also reported narratively based on written comments from the participants.

Results

Participant Characteristics

The recruitment rate was slow, and despite a 6-month trial extension and inclusion of an additional health care region, the intended sample size was not reached. Of the 39 approached eligible participants, 22 (56%) were included and randomized to either the intervention (n=15, 68%) or control (n=7, 32%) group (Figure 2).

The participants lived in central (16/22, 73%) and northern (6/22, 27%) parts of Sweden. One participant in the intervention group was introduced to Me&COPD by the physiotherapist but did not use the tool during the intervention period, although the participant was included in the analysis and assessed at the 3-month follow-up in accordance with the intention-to-treat approach. Participants' characteristics are shown in Table 1. On average, they were aged 72.3 (SD 8.4) years; 55% (12/22) were women. The disease severity ranged from mild to severe according to the Global Initiative for Chronic Obstructive Lung Disease [32].

Over 70 PHC managers in the included health care regions were personally contacted, of whom 6 agreed to participate in the trial. Reasons for refusal included an insufficient number of patients with COPD, high workload, or not being connected to the 1177 Support and Treatment platform. In total, 7 physiotherapists, 5 (71%) women and 2 (29%) men, working at the participating PHCs, were included in the trial and answered the S-MAUQ questionnaire for HCPs. On average, they were aged 47.7 (SD 17.6) years and had worked for 17.6 (SD 10) years as a physiotherapist.

XSL•FO
RenderX

Figure 2. CONSORT (Consolidated Standards of Reporting Trials) participant flow diagram.

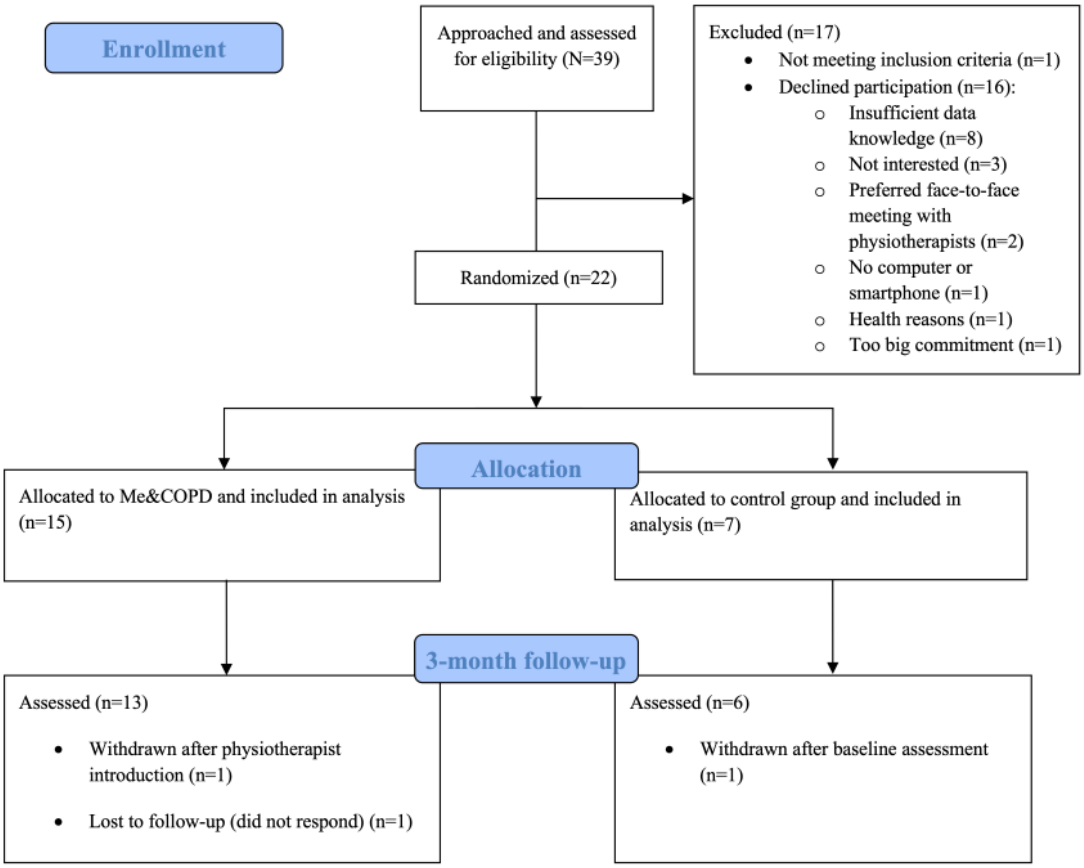

XSL•FO
RenderX

Table 1. Baseline characteristics of the included adults with chronic obstructive pulmonary disease in the randomized controlled pilot and feasibility trial (N=22).

Characteristic	Total	Intervention (n=15)	Control (n=7)
Age (y), mean (SD)	72.3 (8.4)	74.5 (7.4)	67.6 (8.9)
Sex, n (%)			
Female	12 (55)	9 (60)	3 (43)
Male	10 (45)	6 (40)	4 (57)
Living alone, n (%)	10 (45)	8 (53)	2 (29)
Employment status, n (%)			
Employed	4 (18)	2 (13)	2 (29)
Unemployed	1 (5)	0 (0)	1 (14)
Retired	17 (77)	13 (87)	4 (57)
BMI (kg/m^2), mean (SD)	25.6 (4.6)	25.7 (5)	25.2 (4)
FEV_1[a] (L), mean (SD)	1.7 (0.6)	1.7 (0.5)	1.8 (0.8)
FEV_1, % predicted, mean (SD)	65.7 (20.9)	66.8 (20.4)	63.3 (23.4)
FEV_1/FVC[b], mean (SD)	0.6 (0.1)	0.6 (0.1)	0.6 (0.2)
GOLD[c] grades, n (%)			
Grade 1	6 (27)	4 (27)	2 (29)
Grade 2	10 (45)	7 (47)	3 (43)
Grade 3	6 (27)	4 (27)	2 (29)
Grade 4	0 (0)	0 (0)	0 (0)
GOLD[d] groups, n (%)			
Grade A	5 (23)	5 (33)	0 (0)
Grade B	11 (50)	5 (33)	6 (86)
Grade E	6 (27)	5 (33)	1 (14)
Exacerbations in the previous year, n (%)			
1	3 (14)	3 (20)	0 (0)
2-3	4 (18)	3 (20)	1 (14)
≥4	1 (5)	1 (7)	0 (0)
Smoking status, n (%)			
Former	21 (95)	15 (100)	6 (86)
Current	2 (9)	0 (0)	2 (29)
Pack year[e], mean (SD)	20.7 (10.4)	23.3 (9.3)	14.2 (10.7)
Diagnoses (current or former), n (%)			
Diabetes	1 (5)	1 (7)	0 (0)
Hypertension	10 (45)	8 (53)	2 (29)
Stroke	2 (9)	1 (7)	1 (14)
Heart disease[f]	5 (23)	5 (33)	0 (0)
Osteoporosis	4 (18)	2 (13)	2 (29)
Depression and anxiety	2 (9)	1 (7)	1 (14)
Sleep apnea	6 (27)	4 (27)	2 (29)
Cancer	7 (32)	5 (33)	2 (29)

XSL•FO
RenderX

Characteristic	Total	Intervention (n=15)	Control (n=7)
Daily steps[g] (n=21), mean (SD)	6704 (2512)	6525 (2130)	7151 (3494)
Daily steps, all valid days[h] (n=21), mean (SD)	6645 (2830)	6513 (2544)	6975 (3707)

[a]FEV_1: predicted forced expiratory volume in 1 second.

[b]FEV_1/FVC: ratio between FEV_1 and forced vital capacity.

[c]GOLD: Global Initiative for Chronic Obstructive Lung Disease.

[d]Values are calculated according to the Chronic Obstructive Pulmonary Disease Assessment test score.

[e]Pack year is the equivalent of smoking 1 pack of cigarettes per day for 1 year.

[f]Diseases include myocardial infarction, angina pectoris, or heart failure.

[g]Measurements of ≥4 valid weekdays with at least 8 hours of daytime wearing time.

[h]For comparison between studies according to the international task force on physical activity [49].

Usability

A total of 80% (12/15) of the participants in the intervention group answered the S-MAUQ questionnaire at the 3-month follow-up. The mean overall score was 4.4 (SD 1.5) out of 7; that is, responses were overall neutral, with scores ranging between 1.9 and 6.5. The average overall score and the scores for the subscales are illustrated in Figure 3.

Among the subscales, the highest score was assigned to *usefulness* (Table 2). The statements that received the highest scores were that the app had been useful for participants' health and well-being and improved interaction with the physiotherapist and that participants felt confident to communicate via the eHealth tool. The mean score for each statement is presented in Table 2.

Some (7/12, 58%) participants made comments on the S-MAUQ questionnaire regarding their experience of using the Me&COPD eHealth tool. The positive aspects were related to the subscale *usefulness* where participants appreciated having a flexible exercise schedule and felt confident about safety. The negative aspects were related to the subscales *ease of use and satisfaction* and *system information arrangement* and were concerned with how information was organized and also that some functionalities of the eHealth tool were unsatisfactory. A summary of the participants' comments is shown in Textbox 1.

Among physiotherapists, the mean overall score on the S-MAUQ was 4.5 (SD 1.2) out of 7; that is, responses were overall neutral, with scores ranging between 2.8 and 6.3. The average overall score and the scores for the subscales are illustrated in Figure 4.

Figure 3. Boxplots showing the intervention groups' overall score on the Mobile Health App Usability Questionnaire and scores of the subscales ease of use and satisfaction, system information arrangement, and usefulness at the end of the trial. The boxplot whiskers indicate the highest and lowest values, respectively.

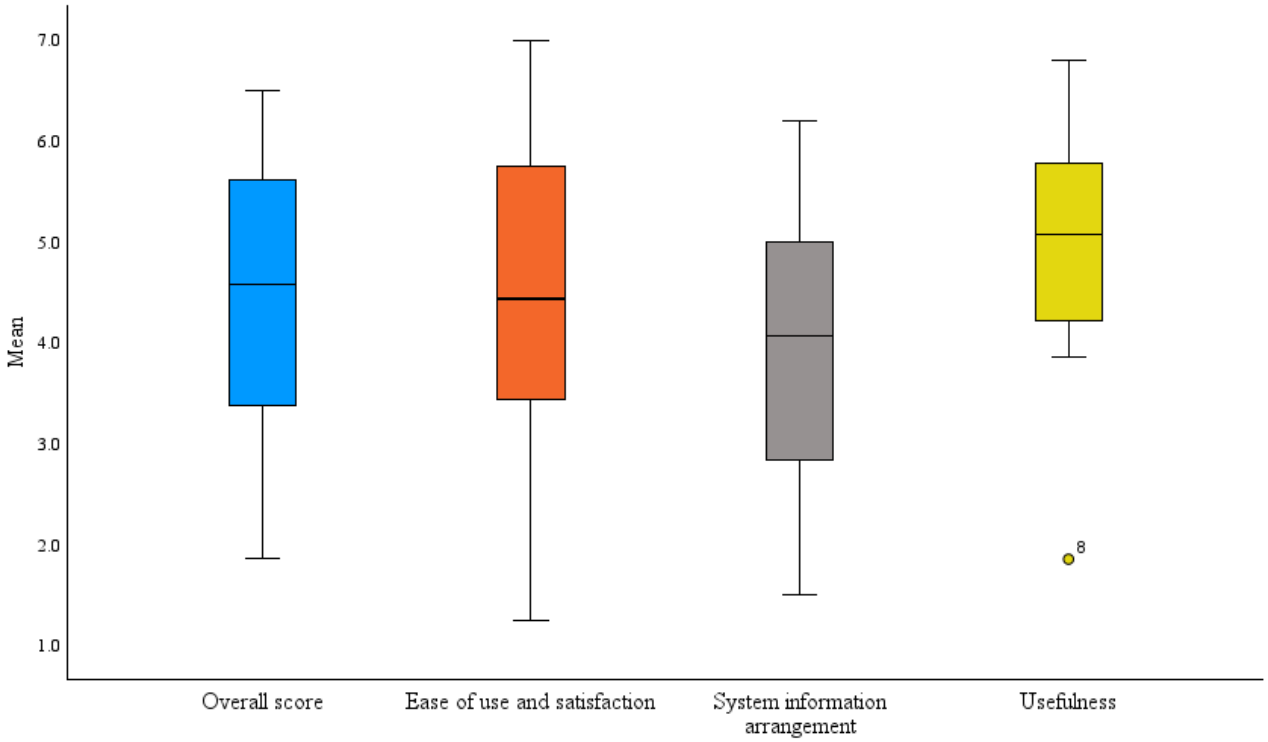

XSL•FO
RenderX

Table 2. The intervention groups' scores on the Swedish version of the Mobile Health App Usability Questionnaire (S-MAUQ) for the subscales and individual statements at the end of the trial (n=12).

S-MAUQ statements	Scores, mean (SD)
Ease of use and satisfaction	4.4 (1.7)
The app was easy to use.	4.1 (2.2)
It was easy for me to learn to use the app.	4.3 (2.1)
I like the interface of the app.	4.2 (2.1)
The information in the app was well organized so I could easily find the information I needed.	4.2 (1.6)
I feel comfortable using this app in social settings (n=10).	4.1 (2.6)
The amount of time involved in using this app has been fitting for me.	4.8 (2)
I would use this app again.	4.8 (2.3)
Overall, I am satisfied with this app.	4.6 (1.8)
System information arrangement	3.9 (1.5)
Whenever I made a mistake using the app, I could recover easily and quickly (n=9).	2.6 (1.7)
This mHealth app provides an acceptable way to receive health care services (n=11).	4.4 (1.8)
The app adequately acknowledged and provided information to let me know the progress of my action.	3.5 (2.2)
The navigation was consistent when moving between screens.	3.6 (1.9)
The interface of the app allowed me to use all the functions (such as entering information, responding to reminders, viewing information) offered by the app.	4.4 (1.4)
This app has all the functions and capabilities I expected it to have.	4.4 (1.9)
Usefulness	4.9 (1.3)
The app would be useful for my health and well-being.	5.2 (1.4)
The app improved my access to health care services (n=10).	4.2 (1.8)
The app helped me manage my health effectively (n=10).	4.3 (1.6)
The app made it convenient for me to communicate with my healthcare provider. (n=11).	5.2 (1.9)
Using the app, I had many more opportunities to interact with my healthcare provider (n=11).	5.2 (1.5)
I felt confident that any information I sent to my provider using the app would be received (n=11).	5.4 (1.5)
I felt comfortable communicating with my healthcare provider using the app (n=11).	4.9 (1.9)

Textbox 1. A summary of user experiences in the intervention group at the end of the trial.

- An advantage to have the opportunity to choose when and where the exercise program should be performed (n=1)
- Felt safe to use (n=1)
- Required a certain degree of computer skills (n=2)
- Unintuitive interface, making it difficult to find the desired content (n=5)
- The exercise diary was not satisfactory, not clear enough, and lacked the possibility to retrospectively register exercises (n=3)
- Unhandy to have to log in to the 1177 platform to access the tool (n=2)
- The in-person meeting with the physiotherapist, where the eHealth tool was introduced, was not properly structured (n=2)

Similar to the participants, the physiotherapists assigned the highest score to the *usefulness* subscale (Table 3). Statements that received the highest scores were regarding more opportunities to interact with the patients, improved access to and delivery of health care services, and the app being useful for their health care practice (Table 3). Some statements had a low response rate, particularly statement 5, because 86% (6/7) of the physiotherapists had chosen the "not applicable" option. In sensitivity analyses where the questions with the lowest response rate (questions 5 and 9) were removed, the results remained essentially the same.

XSL•FO
RenderX

Figure 4. Boxplots showing the physiotherapists' overall score on the Mobile Health App Usability Questionnaire and scores of the subscales ease of use and satisfaction, system information arrangement, and usefulness at the end of the trial. The boxplot whiskers indicate the highest and lowest values, respectively.

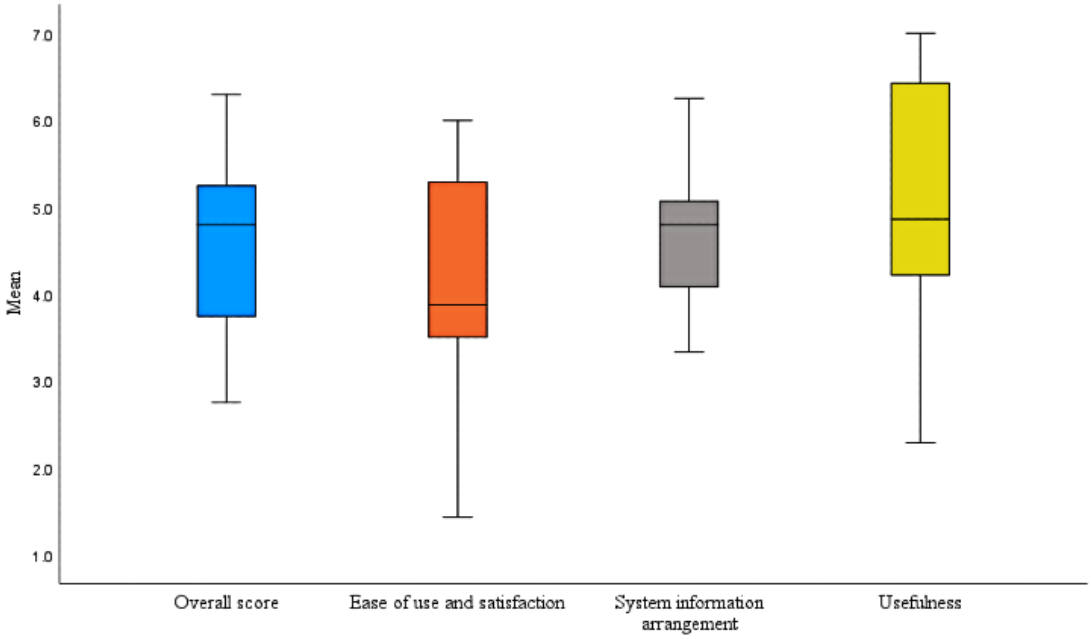

XSL•FO
RenderX

Table 3. The physiotherapists' scores on the Swedish version of the Mobile Health App Usability Questionnaire (S-MAUQ) for the subscales, and individual statements at the end of the trial (n=7).

S-MAUQ statements	Scores, mean (SD)
Ease of use and satisfaction	4.1 (1.5)
The app was easy to use.	4.1 (1.7)
It was easy for me to learn to use the app.	4.6 (1)
I like the interface of the app (n=6).	4 (2.3)
The information in the app was well organized so I could easily find the information I needed.	4.3 (1.8)
I felt comfortable using this app in social settings (n=1).	3[a]
The amount of time involved in using this app has been fitting for me (n=6).	3.8 (1.7)
I would use this app again.	4.4 (2.6)
Overall, I am satisfied with this app.	3.9 (2)
System information arrangement	4.7 (1)
Whenever I made a mistake using the app, I could recover easily and quickly (n=4).	4 (0.8)
This mHealth app provides an acceptable way to deliver health care services (n=6).	5.2 (2.2)
The app adequately acknowledged and provided information to let me know the progress of my action (n=6).	4.8 (1.3)
The navigation was consistent when moving between screens.	4 (1.4)
The interface of the app allowed me to use all the functions (such as entering information, responding to reminders, viewing information) offered by the app (n=5).	3.8 (0.8)
This app has all the functions and capabilities I expected it to have.	5 (1.3)
Usefulness	5.1 (1.7)
The app would be useful for my health care practice.	5 (2.8)
The app improved my access to delivering health care services.	5.4 (2.2)
The app helped me manage my patients' health effectively.	4 (2.4)
The app made it convenient for me to communicate with my patients (n=6).	4.8 (1.2)
Using the app, I had many more opportunities to interact with my patients. (n=6)	5.8 (1.2)
I felt confident that any information I sent to my patients using the app would be received (n=6).	4.5 (1.6)
I felt comfortable communicating with my patients using the app (n=6).	4.7 (1.9)

[a]No SD is reported as only 1 physiotherapist responded to this item.

Exercise Adherence, Intensity, Progression, and Adverse Events

To evaluate exercise adherence, intensity, progression, and adverse events, data were exported from the eHealth tool. In total, data were exported for 12 participants from 4 of the 6 PHCs. Data showed that the participants had a mean of 11.2 (SD 10.3) registered exercise sessions of the intended 24 sessions, with a range between 1 and 28 registered sessions. During the trial, the physiotherapists reported difficulties faced by the participants in registering their exercise sessions in the exercise diary. The limited number of registered exercise sessions meant that evaluation of exercise intensity and progression was not adequate to perform. No falls or other severe adverse events were reported in relation to the exercise program. A total of 5 (42%) participants described single, minor, or transient adverse events, such as joint-related pain (hip, knee, ankle, and wrist), cough, pain from the torso, and breathlessness more than usual in connection with exercise. In addition, 2 (17%) participants commented on previous shoulder and hand impairments being an obstacle to the upper limb exercises. During the baseline assessment, 1 (8%) participant with a previous hip injury experienced increased hip pain. The pain negatively impacted the participant's daily life, decreased the walking distance, and limited leg and cardio training during the trial period.

Clinical Outcome Measures

The test procedures and the clinical outcome measures used were found to be feasible for the participants as well as the assessors. The clinical outcomes for the groups are presented in Table 4. In total, 3 (16%) of 19 participants, 2 (33%) in the control and 1 (8%) in the intervention group, had a COPD exacerbation between baseline and the 3-month follow-up, although hospitalization was not required neither was the intervention extended for the participant in the intervention group. The participants reported few additional COPD-related health care contacts during the trial. In total, 23% (3/13) of the participants in the intervention group received a follow-up from a COPD team, a physician, or a nurse specialized in COPD.

XSL•FO
RenderX

Moreover, 17% (1/6) of the participants in the control group participated in structured COPD-related training sessions.

Table 4. Clinical outcomes at baseline and at the 3-month follow-up for the intervention and control groups.

Clinical outcomes	Me&COPD, mean (SD)			Control, mean (SD)			Between-group difference, mean (95% CI)
	Baseline (n=15)	3-month follow-up (n=13)	Within-group difference (n=13)	Baseline (n=7)	3-month follow-up (n=6)	Within-group difference (n=6)	
Tests, mean (SD)							
6MWT[a] (m)	450.3 (82.7)	465.3 (88)[b]	23.2 (62.9)[b]	502.9 (107.6)	491.2 (125.5)	–10.2 (23.2)	–33.4 (–90.4 to 23.5)[b]
60-second STS[c]	19.3 (7.1)	21 (6.8)[b]	2.1 (4.6)[b]	20.1 (2.4)	20.5 (3.1)	0.0 (2.8)	–2.1 (–6.4 to 2.3)[b]
TUG[d], (s)	10.1 (1.8)	9.8 (2.5)[b]	–0.1 (1.5)[b]	10.4 (2.8)	8.5 (2.9)	–2.2 (1)	–2.1 (–3.5 to –0.6)[b]
UULEX[e] (s)	501.7 (153.6)	453.3 (197)[b]	–43.8 (231.6)[b]	487.4 (122.4)	518.8 (162.4)	25.5 (88.9)	69.3 (–140.9 to 279.5)[b]
Daily steps[f]	6525.1 (2129.8)	5497.3 (2507.6)	–874.1 (2488.8)	7150.7 (3493.6)[b]	7226.4 (2215.5)	75.8 (3080.9)[b]	949.9 (–1837.2 to 3737.1)
Questionnaires							
PA[g] level, mean (SD)	10.7 (3.6)	11 (4.3)	0.1 (4.3)	9.3 (5.2)	12 (4)	2.5 (6.7)	2.4 (–2.9 to 7.8)
≥150 minutes of PA per week, n (%)	6 (40)	6 (46)	—[h]	2 (29)	3 (50)	—	—
CAT[i], mean (SD)	12.6 (5.3)	13.2 (4.7)	0.0 (4.5)	17 (4.5)	19.5 (8.5)	2.3 (6.8)	2.3 (–3.2 to 7.8)
CAT ≥10, n (%)	9 (60)	10 (77)	—	7 (100)	6 (100)	—	—
mMRC[j], mean (SD)	1.4 (0.6)	1.5 (0.8)	0.0 (0.6)	1.29 (0.8)	1.83 (0.8)	0.7 (0.8)	0.7 (–0.0 to 1.4)
mMRC ≥2, n (%)	5 (33.3)	4 (30.8)	—	3 (42.9)	4 (66.7)	—	—
SGRQ[k], mean (SD)	25.3 (10.4)	26.6 (11.8)	0.6 (4.5)	32.8 (13.8)	37.3 (10)	5.5 (10.6)	4.9 (–6.3 to 16)
EQ-5D, health status, mean (SD)	61.7 (17)	68.8 (13)[b]	10 (15.4)[b]	63.3 (12.6)	68.7 (13.4)	6.5 (8.7)	–3.5 (–18 to 11)[b]

[a]6MWT: 6-minute walk test.

[b]One participant was missing.

[c]STS: sit-to-stand test.

[d]TUG: timed up and go test (higher values indicate worse balance).

[e]UULEX: unsupported upper limb exercise test (higher values indicate better upper limb function).

[f]Measurements with ≥4 valid weekdays and with at least 8 hours of daytime wearing time were included.

[g]PA: physical activity (physical activity level according to the Swedish National Board of Health and Welfare indicator questions; values of ≥11 correspond to ≥150 minutes of at least moderate physical activity per week.

[h]Not applicable.

[i]CAT: Chronic Obstructive Pulmonary Disease Assessment Test (higher values indicate greater impact of chronic obstructive pulmonary disease).

[j]mMRC: Modified Medical Research Council dyspnea scale (higher values indicate more dyspnea).

[k]SGRQ: St George's Respiratory Questionnaire (higher values indicate a worse state of health).

Discussion

Principal Findings

This trial is the first to evaluate the feasibility of a novel eHealth tool for home-based PR called Me&COPD that has been developed in a structured cocreation process with end users [29]. The primary finding concerned the usability of the tool, where people with COPD and physiotherapists in PHCs rated overall usability as acceptable. The subscale *usefulness* was assigned the highest scores by both people with COPD and

XSL•FO
RenderX

physiotherapists, supporting the notion that Me&COPD may be a suitable alternative strategy to provide PR in people with COPD. However, the trial results did indicate that the exercise diary was underused and needs further development in order to collect data on exercise adherence, intensity, progression, and adverse events. The clinical outcome measures used were found to be feasible.

Interpretation of the Findings

Providing alternative models for PR goes hand-in-hand with the principle of person-centered care [13]. In the latest clinical practice guideline for PR, the recommendation is to offer either center-based PR or solutions for telerehabilitation [8]. Me&COPD includes the core components of PR—individualized exercise training, education, and behavior change strategies [13]. In addition, the results of the S-MAUQ indicated that the eHealth tool was beneficial for the participants' health and well-being, physiotherapists' delivery of health care to their patients, and interaction between them. Moreover, the use of cocreation with end users in the entire development process, which has been sparsely used in this field [28], further enhances the likelihood of it being a feasible eHealth tool for people with COPD. With this feasibility trial, we continue the cocreation process by letting the participants answer structured questions regarding the functionality, implications, and potential areas of improvement, which is a recommended strategy in cocreation research [31].

Previous home-based telerehabilitation programs for PR have been heterogeneous in design and content but have often included components of education and exercise [18]. However, interaction with HCPs via a web-based platform or smartphone app, as in Me&COPD, is not commonly used because communication has often been performed by videoconferencing or telephone [18]. The interactive web-based PR intervention reported by Chaplin et al [57], which was found to be a feasible alternative to conventional PR, shares some of the features of Me&COPD, that is, it was home-based and included components of individually adapted education and an exercise program with regular HCP support to progress the exercise. In PR, it is essential that the exercise program is individually prescribed and progressed [13]. In this trial, we used an exercise diary for participants to register their exercise experience and symptoms, including dyspnea and muscle fatigue ratings, after every exercise session. The purpose of the diary was to monitor exercise adherence and provide comprehensive information to the physiotherapists, enabling adjustment and progression of the participants' exercise programs accordingly. Unfortunately, only a few sessions were registered. Participants and participating physiotherapists reported that the registration was too challenging, leading to underuse. Nevertheless, we received indications that participants exercised more frequently than the registered average of 11 (SD 10) sessions. Because of the low number of exercise registrations, the information provided to the physiotherapists to guide exercise progression was limited, and we did not find it useful to evaluate exercise intensity and progression. We believe that this was a very important result, which highlights that the design of the exercise diary was not optimal and needs further development before the eHealth tool can be implemented in clinical practice.

Me&COPD was built on the 1177 Support and Treatment platform, which is associated with advantages and disadvantages. Positive aspects are that the platform is well-known by citizens in Sweden and that several health care regions already use the platform to provide various digital health care interventions, demonstrating continuity. Furthermore, the platform is considered to be credible with high data security, which has previously been voiced to be an important criterion for an eHealth tool by prospective users [21] and by participants in this study. One of the disadvantages that was reported by participants in this study was the inconvenience of having to log in to the platform to access Me&COPD and to be able to read eventual messages from their physiotherapist. Another disadvantage was that the eHealth tool had to be adapted to the platform's preset features and functionalities, which limited the possibility of making changes to the interface. Consequently, during its development, some functionalities, such as exercise reminders, could not be designed as intended. These concerns were validated by the participants' feedback, which frequently highlighted difficulties in navigating the interface and locating the desired content. This aligns with previous research reporting that people with COPD prefer eHealth tools that are not overly technical and are easy to navigate [21].

The recruitment of physiotherapists and people with COPD through PHCs was challenging. We prolonged the trial, included an additional health care region, and hired a person to work specifically with the recruitment. Despite that, the intended sample size of 30 participants was not reached. Possible reasons for the low recruitment of physiotherapists might be the heavy workload in Swedish primary health care, partly related to the COVID-19 pandemic. The pandemic had a huge negative effect on the health care given to people with asthma and COPD in Sweden [58]. Therefore, people with COPD were not referred as frequently to annual follow-ups, which are important for identifying their rehabilitation needs. This was confirmed by the PHC managers and the physiotherapists during the study period, as they reported fewer referrals for PR than before the COVID-19 pandemic. Another contributing factor to the difficulty of recruiting PHCs was the demand to have the 1177 Support and Treatment platform implemented in their work organization to be included in the trial, which turned out to be a barrier in one of the health care regions. Furthermore, the views and beliefs of the HCPs can significantly influence the implementation of an innovation [20]. For example, Slevin et al [59] reported that HCPs may be unwilling to adopt eHealth solutions because there is no strong evidence base for their effectiveness. Due to this, they may not recommend such treatment options to their patients, which could have affected the recruitment in our study as well. Other known barriers from the literature regarding the implementation of eHealth tools are a lack of organizational support to change the work procedures and insufficient digital literacy among the HCPs [59,60]. A possible barrier to the recruitment of people with COPD might have been that some were skeptical to replace a face-to-face contact with an HCP with an eHealth solution, as previously described in the literature [21,61] and by physiotherapists within this trial. By contrast, only 2 (5%) of the 39 eligible participants for this trial indicated that their main reason for declining participation was a preference for face-to-face treatment.

XSL•FO
RenderX

Nevertheless, although our sample size fairly aligned with the recommendations for pilot trials [56], we have realized that the procedure of recruitment needs some consideration to be a feasible alternative in a larger trial.

Clearly, eHealth solutions are not universally suitable, despite the high level of internet access and digital engagement among people with COPD in Sweden [62]. Participants in this trial expressed that some degree of computer skills was required to manage the eHealth tool. To provide additional support and the possibility of repeated information, a written user manual, similar to the one that the physiotherapists received, could be considered, along with incorporating a film within the eHealth tool describing the features of the intervention. Consequently, digital literacy should be considered in clinical practice when choosing the best treatment for the patient [57]. Applying a patient-centered approach where aspects such as perceived usefulness, digital literacy, self-efficacy, and social context are considered has been suggested to be important when it comes to the individual's readiness to adopt a digital health intervention [61]. However, in the end, the individual's treatment preference may be the most important for the completion of an intervention.

Strengths of this trial include its design as a multicenter randomized controlled trial and the use of cocreation in both the development process and throughout the study. The eHealth tool seemed safe to use, and the exercises included in the program were both feasible and suitable for a home environment, requiring no equipment besides resistance bands. Furthermore, the pragmatic design, that is, the trial being conducted as part of ordinary clinical practice, increases the chances of a successful future implementation.

Limitations

To evaluate usability, we used a Swedish version of the MAUQ questionnaire. The original questionnaire was developed for evaluating mobile health (mHealth) apps [42]. However, it is important to note that our eHealth tool is not solely an mHealth app, as the 1177 Support and Treatment platform, through which Me&COPD is accessed, is also available via a website. The original questionnaire has been found to have good validity and reliability, although the HCP versions need further evaluation [42]. We used a structured process when translating the questionnaire into Swedish, but no psychometric analysis of the Swedish version has been performed, which may be acknowledged as a limitation. The MAUQ has no established cutoff or reference values for what is considered as good usability. A possible reason for this may be that the questionnaire has not often been used in mHealth studies aiming to evaluate usability, despite that it was specifically designed to evaluate that outcome [63]. Nevertheless, this makes it a bit difficult to interpret the results. A previous study using the stand-alone version of MAUQ for patients suggested that an average score <4 would indicate that the usability is not good [64]. Comparable to the approach in this study, previous studies using the interactive versions of the questionnaire have chosen to report the average overall score and the scores for respective subscales with reference to the descriptions of each number on the 7-point Likert scale [65,66]. A couple of the statements in the health care version received a low response rate because the physiotherapists had chosen the "not applicable" option. The first statement concerned the use of the eHealth tool in social settings, which was not relevant for them, and the other question was concerned with whether it had been easy to recover from making a mistake in the app. Choosing the "not applicable" option in this case could be considered positive if it meant that they had not made any mistakes when using the eHealth tool.

One limitation was that we did not measure physiotherapist adherence, which means that we do not know if they complied with the intervention's recommended work procedures. Furthermore, due to the nature of this feasibility trial, no statistical between-group comparisons were performed [35], and no clear trends in the clinical outcomes were seen. The intervention was delivered as a complement to the health care provided for people with COPD, primarily within primary care. The participants in both groups could take part in other interventions during the trial, which could hypothetically make it harder to evaluate the effect of a specific intervention. However, only 21% (4/19) of the participants, 23% (3/13) in the intervention group and 17% (1/6) in the control group, declared that they had received additional COPD-related health care contacts, of which most of them were single follow-ups. Only 17% (1/6) of the participants in the control group participated in a PR program at an outpatient health care facility, which supports the notion that access to PR is limited in people with COPD.

Conclusions

The novel eHealth tool, Me&COPD, seems to be a safe and feasible alternative strategy for PR because the tool was found to have acceptable usability by people with COPD as well as participating physiotherapists. Usability may be improved by better organization of the information and simplification of the exercise diary to enable collection of data on exercise adherence, intensity, and progression via the eHealth tool. The test procedures and outcome measures used in the trial worked satisfactorily, but the recruitment process needs further consideration. The results from this trial can be used to inform the design of future research studies with similar interventions, although the effectiveness of the intervention remains to be evaluated in a future larger trial. When the tool has been revised and is ready to be implemented in clinical practice, the intervention can be offered to people with COPD across the country, thanks to the established nationwide digital health care platform.

Acknowledgments

The authors would like to express their sincere gratitude to the participants and physiotherapists who were part of the trial. They would also like to thank Karin Gebart-Hedman at Region Västerbotten for designing Me&COPD on the 1177 Support and

XSL•FO
RenderX

Treatment platform, Dycke Forslund Stiby for the work with the recruitment of primary health care centers, and Andreas Petersson for being involved as an assessor.

The trial was supported by the Swedish Heart and Lung Foundation (grant 20190406), the Swedish Research Council, and the medical faculty at the Umeå University. The funders had no role in the conception and design of the study, collection and interpretation of the data, or writing of the manuscript.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

All authors have made a direct and significant contribution to this work. KW was the principal investigator of the trial. PS, SL, AT, and KW contributed to the concept of the study, and all authors contributed to the design. ÅK and PS performed the data collection. ÅK was primarily responsible for the analysis and writing the original manuscript draft. The other authors (PS, SL, AT, and KW) critically reviewed and revised the manuscript. All authors have approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) checklist (V 1.6).
[PDF File (Adobe PDF File), 1097 KB - formative_v9i1e68195_app1.pdf]

References

1. Machado A, Marques A, Burtin C. Extra-pulmonary manifestations of COPD and the role of pulmonary rehabilitation: a symptom-centered approach. Expert Rev Respir Med 2021 Jan;15(1):131-142. [doi: 10.1080/17476348.2021.1854737] [Medline: 33225762]
2. McCarthy B, Casey D, Devane D, Murphy K, Murphy E, Lacasse Y. Pulmonary rehabilitation for chronic obstructive pulmonary disease. Cochrane Database Syst Rev 2015 Feb 23;2015(2):CD003793 [FREE Full text] [doi: 10.1002/14651858.CD003793.pub3] [Medline: 25705944]
3. Spruit MA, Singh SJ, Garvey C, ZuWallack R, Nici L, Rochester C, et al. An official American Thoracic Society/European Respiratory Society statement: key concepts and advances in pulmonary rehabilitation. Am J Respir Crit Care Med 2013 Oct 15;188(8):e13-e64. [doi: 10.1164/rccm.201309-1634ST] [Medline: 24127811]
4. Nationella riktlinjer för vård vid astma och KOL: stöd för styrning och ledning. Socialstyrelsen. 2020 Dec. URL: https://www.socialstyrelsen.se/globalassets/sharepoint-dokument/artikelkatalog/nationella-riktlinjer/2020-12-7135.pdf [accessed 2025-06-05]
5. Rochester CL, Vogiatzis I, Holland AE, Lareau SC, Marciniuk DD, Puhan MA, et al. An official American Thoracic Society/European Respiratory Society policy statement: enhancing implementation, use, and delivery of pulmonary rehabilitation. Am J Respir Crit Care Med 2015 Dec 01;192(11):1373-1386. [doi: 10.1164/rccm.201510-1966ST] [Medline: 26623686]
6. Liu S, Zhao Q, Li W, Zhao X, Li K. The cost-effectiveness of pulmonary rehabilitation for COPD in different settings: a systematic review. Appl Health Econ Health Policy 2021 May 20;19(3):313-324. [doi: 10.1007/s40258-020-00613-5] [Medline: 33079374]
7. Kroniskt obstruktiv lungsjukdom (KOL) – behandlingsrekommendation. Läkemedelsverket. 2023 Mar 27. URL: https://www.lakemedelsverket.se/sv/behandling-och-forskrivning/behandlingsrekommendationer/sok-behandlingsrekommendationer/kroniskt-obstruktiv-lungsjukdom-kol--behandlingsrekommendation#hmainbody1 [accessed 2025-06-05]
8. Rochester CL, Alison JA, Carlin B, Jenkins AR, Cox NS, Bauldoff G, et al. Pulmonary rehabilitation for adults with chronic respiratory disease: an official American Thoracic Society clinical practice guideline. Am J Respir Crit Care Med 2023 Aug 15;208(4):e7-26 [FREE Full text] [doi: 10.1164/rccm.202306-1066ST] [Medline: 37581410]
9. Man W, Chaplin E, Daynes E, Drummond A, Evans RA, Greening NJ, et al. British Thoracic Society Clinical Statement on pulmonary rehabilitation. Thorax 2023 Oct 08;78(Suppl 4):s2-15. [doi: 10.1136/thorax-2023-220439] [Medline: 37770084]
10. Desveaux L, Janaudis-Ferreira T, Goldstein R, Brooks D. An international comparison of pulmonary rehabilitation: a systematic review. COPD 2015 Apr;12(2):144-153 [FREE Full text] [doi: 10.3109/15412555.2014.922066] [Medline: 24984085]
11. Wadell K, Janaudis Ferreira T, Arne M, Lisspers K, Ställberg B, Emtner M. Hospital-based pulmonary rehabilitation in patients with COPD in Sweden--a national survey. Respir Med 2013 Aug;107(8):1195-1200 [FREE Full text] [doi: 10.1016/j.rmed.2013.04.019] [Medline: 23702089]

XSL•FO
RenderX

12. Milner SC, Boruff JT, Beaurepaire C, Ahmed S, Janaudis-Ferreira T. Rate of, and barriers and enablers to, pulmonary rehabilitation referral in COPD: a systematic scoping review. Respir Med 2018 Apr;137:103-114 [FREE Full text] [doi: 10.1016/j.rmed.2018.02.021] [Medline: 29605192]
13. Holland AE, Cox NS, Houchen-Wolloff L, Rochester CL, Garvey C, ZuWallack R, et al. Defining modern pulmonary rehabilitation. An official American Thoracic Society workshop report. Ann Am Thorac Soc 2021 May;18(5):e12-e29 [FREE Full text] [doi: 10.1513/AnnalsATS.202102-146ST] [Medline: 33929307]
14. Guiding optimal development and use of digital health towards improved health outcomes. World Health Organization. URL: https://www.who.int/westernpacific/activities/guiding-optimal-development-and-use-of-digital-health-towards-improved-health-outcomes [accessed 2025-06-05]
15. Cox NS, Dal Corso S, Hansen H, McDonald CF, Hill CJ, Zanaboni P, et al. Telerehabilitation for chronic respiratory disease. Cochrane Database Syst Rev 2021 Jan 29;1(1):CD013040 [FREE Full text] [doi: 10.1002/14651858.CD013040.pub2] [Medline: 33511633]
16. Cox NS, Khor YH. Telerehabilitation in pulmonary diseases. Curr Opin Pulm Med 2023 Jul 01;29(4):313-321. [doi: 10.1097/MCP.0000000000000962] [Medline: 37132293]
17. Sakunrag I, Boontha N, Boonpattharatthiti K, Dhippayom T. Impact of tele-pulmonary rehabilitation in patients with chronic obstructive disease: a systematic review and network meta-analysis. Telemed J E Health 2025 Apr;31(4):441-450. [doi: 10.1089/tmj.2024.0476] [Medline: 39630505]
18. Michaelchuk W, Oliveira A, Marzolini S, Nonoyama M, Maybank A, Goldstein R, et al. Design and delivery of home-based telehealth pulmonary rehabilitation programs in COPD: a systematic review and meta-analysis. Int J Med Inform 2022 Mar 31;162:104754. [doi: 10.1016/j.ijmedinf.2022.104754] [Medline: 35395474]
19. Chung C, Lee JW, Lee SW, Jo MW. Clinical efficacy of mobile app-based, self-directed pulmonary rehabilitation for patients with chronic obstructive pulmonary disease: systematic review and meta-analysis. JMIR Mhealth Uhealth 2024 Jan 04;12:e41753 [FREE Full text] [doi: 10.2196/41753] [Medline: 38179689]
20. Harvey G, Kitson A. PARIHS revisited: from heuristic to integrated framework for the successful implementation of knowledge into practice. Implement Sci 2016 Mar 10;11:33 [FREE Full text] [doi: 10.1186/s13012-016-0398-2] [Medline: 27013464]
21. Sönnerfors P, Skavberg Roaldsen K, Lundell S, Toots A, Wadell K, Halvarsson A. Preferences for an eHealth tool to support physical activity and exercise training in COPD: a qualitative study from the viewpoint of prospective users. BMC Pulm Med 2023 Feb 13;23(1):65 [FREE Full text] [doi: 10.1186/s12890-023-02353-3] [Medline: 36782155]
22. Tsai LL, McNamara RJ, Dennis SM, Moddel C, Alison JA, McKenzie DK, et al. Satisfaction and experience with a supervised home-based real-time videoconferencing telerehabilitation exercise program in people with chronic obstructive pulmonary disease (COPD). Int J Telerehabil 2016;8(2):27-38 [FREE Full text] [doi: 10.5195/ijt.2016.6213] [Medline: 28775799]
23. Cox NS, Lee JY, McDonald CF, Mahal A, Alison JA, Wootton R, et al. Perceived autonomy support in telerehabilitation by people with chronic respiratory disease: a mixed methods study. Chest 2023 Jun;163(6):1410-1424. [doi: 10.1016/j.chest.2022.12.023] [Medline: 36574926]
24. Slevin P, Kessie T, Cullen J, Butler MW, Donnelly SC, Caulfield B. Exploring the potential benefits of digital health technology for the management of COPD: a qualitative study of patient perceptions. ERJ Open Res 2019 Apr;5(2):00239-02018 [FREE Full text] [doi: 10.1183/23120541.00239-2018] [Medline: 31111039]
25. Barenfeld E, Ali L, Wallström S, Fors A, Ekman I. Becoming more of an insider: a grounded theory study on patients' experience of a person-centred e-health intervention. PLoS One 2020 Nov 23;15(11):e0241801 [FREE Full text] [doi: 10.1371/journal.pone.0241801] [Medline: 33226986]
26. Marklund S, Tistad M, Lundell S, Östrand L, Sörlin A, Boström C, et al. Experiences and factors affecting usage of an eHealth tool for self-management among people with chronic obstructive pulmonary disease: qualitative study. J Med Internet Res 2021 Apr 30;23(4):e25672 [FREE Full text] [doi: 10.2196/25672] [Medline: 33929327]
27. Cox NS, Holland AE. Pulmonary rehabilitation: one size does not fit all. Thorax 2025 Mar 18;80(4):189-190. [doi: 10.1136/thorax-2024-222680] [Medline: 39900490]
28. Pimenta S, Hansen H, Demeyer H, Slevin P, Cruz J. Role of digital health in pulmonary rehabilitation and beyond: shaping the future. ERJ Open Res 2023 Mar 08;9(2):00212-2022 [FREE Full text] [doi: 10.1183/23120541.00212-2022] [Medline: 36923569]
29. Lundell S, Toots A, Sönnerfors P, Halvarsson A, Wadell K. Participatory methods in a digital setting: experiences from the co-creation of an eHealth tool for people with chronic obstructive pulmonary disease. BMC Med Inform Decis Mak 2022 Mar 18;22(1):68 [FREE Full text] [doi: 10.1186/s12911-022-01806-9] [Medline: 35303895]
30. Halvorsrud K, Kucharska J, Adlington K, Rüdell K, Brown Hajdukova E, Nazroo J, et al. Identifying evidence of effectiveness in the co-creation of research: a systematic review and meta-analysis of the international healthcare literature. J Public Health (Oxf) 2021 Apr 12;43(1):197-208 [FREE Full text] [doi: 10.1093/pubmed/fdz126] [Medline: 31608396]
31. Leask CF, Sandlund M, Skelton DA, Altenburg TM, Cardon G, Chinapaw MJ, et al. Framework, principles and recommendations for utilising participatory methodologies in the co-creation and evaluation of public health interventions. Res Involv Engagem 2019 Jan 9;5(1):2 [FREE Full text] [doi: 10.1186/s40900-018-0136-9] [Medline: 30652027]

XSL•FO
RenderX

32. Global strategy for prevention, diagnosis and management of COPD: 2023 report. Global Initiative for Chronic Obstructive Lung Disease. 2023. URL: https://goldcopd.org/2023-gold-report-2/ [accessed 2025-06-05]
33. Lundell S, Tistad M, Rehn B, Wiklund M, Holmner Å, Wadell K. Building COPD care on shaky ground: a mixed methods study from Swedish primary care professional perspective. BMC Health Serv Res 2017 Jul 10;17(1):467 [FREE Full text] [doi: 10.1186/s12913-017-2393-y] [Medline: 28693473]
34. Tistad M, Lundell S, Wiklund M, Nyberg A, Holmner Å, Wadell K. Usefulness and relevance of an ehealth tool in supporting the self-management of chronic obstructive pulmonary disease: explorative qualitative study of a cocreative process. JMIR Hum Factors 2018 Oct 26;5(4):e10801 [FREE Full text] [doi: 10.2196/10801] [Medline: 30368440]
35. Eldridge SM, Chan CL, Campbell MJ, Bond CM, Hopewell S, Thabane L, et al. CONSORT 2010 statement: extension to randomised pilot and feasibility trials. BMJ 2016 Oct 24;355:i5239 [FREE Full text] [doi: 10.1136/bmj.i5239] [Medline: 27777223]
36. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126 [FREE Full text] [doi: 10.2196/jmir.1923] [Medline: 22209829]
37. Hoffmann TC, Glasziou PP, Boutron I, Milne R, Perera R, Moher D, et al. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. BMJ 2014 Mar 07;348(mar07 3):g1687 [FREE Full text] [doi: 10.1136/bmj.g1687] [Medline: 24609605]
38. Effing TW, Vercoulen JH, Bourbeau J, Trappenburg J, Lenferink A, Cafarella P, et al. Definition of a COPD self-management intervention: International Expert Group consensus. Eur Respir J 2016 Jul;48(1):46-54 [FREE Full text] [doi: 10.1183/13993003.00025-2016] [Medline: 27076595]
39. Borg GA. Psychophysical bases of perceived exertion. Med Sci Sports Exerc 1982;14(5):377-381. [Medline: 7154893]
40. American College of Sports Medicine. American College of Sports Medicine position stand. Progression models in resistance training for healthy adults. Med Sci Sports Exerc 2009 Mar;41(3):687-708. [doi: 10.1249/MSS.0b013e3181915670] [Medline: 19204579]
41. Gloeckl R, Zwick RH, Fürlinger U, Jarosch I, Schneeberger T, Leitl D, et al. Prescribing and adjusting exercise training in chronic respiratory diseases - expert-based practical recommendations. Pulmonology 2023 Jul;29(4):306-314 [FREE Full text] [doi: 10.1016/j.pulmoe.2022.09.004] [Medline: 36272962]
42. Zhou L, Bao J, Setiawan IM, Saptono A, Parmanto B. The mHealth App Usability Questionnaire (MAUQ): development and validation study. JMIR Mhealth Uhealth 2019 Apr 11;7(4):e11500 [FREE Full text] [doi: 10.2196/11500] [Medline: 30973342]
43. Sousa VD, Rojjanasrirat W. Translation, adaptation and validation of instruments or scales for use in cross-cultural health care research: a clear and user-friendly guideline. J Eval Clin Pract 2011 Apr;17(2):268-274. [doi: 10.1111/j.1365-2753.2010.01434.x] [Medline: 20874835]
44. Holland AE, Spruit MA, Troosters T, Puhan MA, Pepin V, Saey D, et al. An official European Respiratory Society/American Thoracic Society technical standard: field walking tests in chronic respiratory disease. Eur Respir J 2014 Dec;44(6):1428-1446 [FREE Full text] [doi: 10.1183/09031936.00150314] [Medline: 25359355]
45. Crook S, Büsching G, Schultz K, Lehbert N, Jelusic D, Keusch S, et al. A multicentre validation of the 1-min sit-to-stand test in patients with COPD. Eur Respir J 2017 Mar 02;49(3):1601871 [FREE Full text] [doi: 10.1183/13993003.01871-2016] [Medline: 28254766]
46. Takahashi T, Jenkins SC, Strauss GR, Watson CP, Lake FR. A new unsupported upper limb exercise test for patients with chronic obstructive pulmonary disease. J Cardiopulm Rehabil 2003;23(6):430-437. [doi: 10.1097/00008483-200311000-00007] [Medline: 14646791]
47. Mesquita R, Wilke S, Smid DE, Janssen DJ, Franssen FM, Probst VS, et al. Measurement properties of the Timed Up and Go test in patients with COPD. Chron Respir Dis 2016 Nov 08;13(4):344-352 [FREE Full text] [doi: 10.1177/1479972316647178] [Medline: 27165963]
48. Olsson SJ, Ekblom Ö, Andersson E, Börjesson M, Kallings LV. Categorical answer modes provide superior validity to open answers when asking for level of physical activity: a cross-sectional study. Scand J Public Health 2016 Feb 21;44(1):70-76. [doi: 10.1177/1403494815602830] [Medline: 26392418]
49. Demeyer H, Mohan D, Burtin C, Vaes AW, Heasley M, Bowler RP, et al. Objectively measured physical activity in patients with COPD: recommendations from an international task force on physical activity. J COPD Found 2021;8(4):528-550. [doi: 10.15326/jcopdf.2021.0213]
50. Mahler DA, Wells CK. Evaluation of clinical methods for rating dyspnea. Chest 1988 Mar;93(3):580-586 [FREE Full text] [doi: 10.1378/chest.93.3.580] [Medline: 3342669]
51. Jones PW. COPD assessment test --rationale, development, validation and performance. COPD 2013 Apr 02;10(2):269-271 [FREE Full text] [doi: 10.3109/15412555.2013.776920] [Medline: 23547637]
52. Jones P, Lareau S, Mahler DA. Measuring the effects of COPD on the patient. Respir Med 2005 Dec;99 Suppl B:S11-S18 [FREE Full text] [doi: 10.1016/j.rmed.2005.09.011] [Medline: 16236492]
53. EuroQol Group. EuroQol--a new facility for the measurement of health-related quality of life. Health Policy 1990 Dec;16(3):199-208. [doi: 10.1016/0168-8510(90)90421-9] [Medline: 10109801]

XSL•FO
RenderX

54. Agustí A, Celli BR, Criner GJ, Halpin D, Anzueto A, Barnes P, et al. Global initiative for chronic obstructive lung disease 2023 report: GOLD executive summary. Eur Respir J 2023 Apr;61(4):2300239 [FREE Full text] [doi: 10.1183/13993003.00239-2023] [Medline: 36858443]
55. Tsai LL, McNamara RJ, Moddel C, Alison JA, McKenzie DK, McKeough ZJ. Home-based telerehabilitation via real-time videoconferencing improves endurance exercise capacity in patients with COPD: the randomized controlled TeleR Study. Respirology 2017 May 19;22(4):699-707 [FREE Full text] [doi: 10.1111/resp.12966] [Medline: 27992099]
56. Bell ML, Whitehead AL, Julious SA. Guidance for using pilot studies to inform the design of intervention trials with continuous outcomes. Clin Epidemiol 2018 Jan;Volume 10:153-157. [doi: 10.2147/clep.s146397]
57. Chaplin E, Hewitt S, Apps L, Bankart J, Pulikottil-Jacob R, Boyce S, et al. Interactive web-based pulmonary rehabilitation programme: a randomised controlled feasibility trial. BMJ Open 2017 Mar 31;7(3):e013682 [FREE Full text] [doi: 10.1136/bmjopen-2016-013682] [Medline: 28363923]
58. Stridsman C, Konradsen JR, Hesselmar B, Lindberg A, Vanfleteren LE. Impact of the COVID-19 pandemic on Swedish asthma and COPD care. Eur Respir J 2022;60(suppl 66):2461 [FREE Full text] [doi: 10.1183/13993003.congress-2022.2461]
59. Slevin P, Kessie T, Cullen J, Butler MW, Donnelly SC, Caulfield B. Exploring the barriers and facilitators for the use of digital health technologies for the management of COPD: a qualitative study of clinician perceptions. QJM 2020 Mar 01;113(3):163-172. [doi: 10.1093/qjmed/hcz241] [Medline: 31545374]
60. Nyberg A, Lundell S, Pesola UM, Audulv Å, Wadell K. Evaluation of a digital COPD education program for healthcare professionals in long-term care - a mixed methods study. Int J Chron Obstruct Pulmon Dis 2022 Apr;17:905-918 [FREE Full text] [doi: 10.2147/COPD.S353187] [Medline: 35497374]
61. Slevin P, Kessie T, Cullen J, Butler MW, Donnelly SC, Caulfield B. A qualitative study of chronic obstructive pulmonary disease patient perceptions of the barriers and facilitators to adopting digital health technology. Digit Health 2019 Aug 25;5:2055207619871729 [FREE Full text] [doi: 10.1177/2055207619871729] [Medline: 31489206]
62. Sönnerfors P, Skavberg Roaldsen K, Ståhle A, Wadell K, Halvarsson A. Access to, use, knowledge, and preferences for information technology and technical equipment among people with chronic obstructive pulmonary disease (COPD) in Sweden. A cross-sectional survey study. BMC Med Inform Decis Mak 2021 Jun 10;21(1):185 [FREE Full text] [doi: 10.1186/s12911-021-01544-4] [Medline: 34112150]
63. Hajesmaeel-Gohari S, Khordastan F, Fatehi F, Samzadeh H, Bahaadinbeigy K. The most used questionnaires for evaluating satisfaction, usability, acceptance, and quality outcomes of mobile health. BMC Med Inform Decis Mak 2022 Jan 27;22(1):22 [FREE Full text] [doi: 10.1186/s12911-022-01764-2] [Medline: 35081953]
64. Mustafa N, Safii NS, Jaffar A, Sani NS, Mohamad MI, Abd Rahman AH, et al. Malay version of the mHealth app usability questionnaire (M-MAUQ): translation, adaptation, and validation study. JMIR Mhealth Uhealth 2021 Feb 04;9(2):e24457 [FREE Full text] [doi: 10.2196/24457] [Medline: 33538704]
65. Amin R, Qazi AM, Ambreen M, Naushad N, Pizzuti R, Buchanan F, et al. Opening the digital front door for individuals using long-term in-home ventilation (LIVE) during a pandemic- implementation, feasibility and acceptability. Digit Health 2023;9:20552076231213447 [FREE Full text] [doi: 10.1177/20552076231213447] [Medline: 38025099]
66. Angelucci A, Pongiglione B, Bernasconi S, Carrone F, Mazziotti G, Costantino ML, et al. A participatory process to design an app to improve adherence to anti-osteoporotic therapies: a development and usability study. Digit Health 2023 Dec 11;9:20552076231218858 [FREE Full text] [doi: 10.1177/20552076231218858] [Medline: 38107981]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
COPD: chronic obstructive pulmonary disease
HCP: health care provider
MAUQ: Mobile Health App Usability Questionnaire
mHealth: mobile health
PHC: primary health care center
PR: pulmonary rehabilitation
S-MAUQ: Swedish version of the Mobile Health App Usability Questionnaire
TIDieR: Template for Intervention Description and Replication

XSL•FO
RenderX

Edited by A Mavragani; submitted 31.10.24; peer-reviewed by V Binson, J Edwards, B Lai; comments to author 09.04.25; revised version received 28.04.25; accepted 15.05.25; published 23.06.25.

Please cite as:
Karlsson Å, Sönnerfors P, Lundell S, Toots A, Wadell K
Evaluation of a Novel eHealth Tool for Pulmonary Rehabilitation in People With Chronic Obstructive Pulmonary Disease: Randomized Controlled Pilot and Feasibility Trial
JMIR Form Res 2025;9:e68195
URL: https://formative.jmir.org/2025/1/e68195
doi:10.2196/68195
PMID:40550122

©Åsa Karlsson, Pernilla Sönnerfors, Sara Lundell, Annika Toots, Karin Wadell. Originally published in JMIR Formative Research (https://formative.jmir.org), 23.06.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Evaluating the Impact on Pain Perceptions, Pain Intensity, and Physical Activity of a Mobile App to Empower Employees With Musculoskeletal Pain: Mixed Methods Pilot Study

Stijn Keyaerts[1,2], PhD; Maxwell Szymanski[3], PhD; Lode Godderis[1,2], Prof Dr; Vero Vanden Abeele[3], Prof Dr; Liesbeth Daenen[1,2,4], PhD

[1]Knowledge, Information and Research Center, Group IDEWE, Leuven, Belgium
[2]Department of Public Health and Primary Care, KU Leuven, Leuven, Belgium
[3]Department of Computer Science, KU Leuven, Leuven, Belgium
[4]Department of Physiotherapy, Human Physiology and Anatomy, VU Brussels, Brussels, Belgium

Corresponding Author:
Stijn Keyaerts, PhD
Knowledge, Information and Research Center
Group IDEWE
Interleuvenlaan 58
Leuven, 3001
Belgium
Phone: 32 16390558
Email: stijn.keyaerts@idewe.be

Abstract

Background: Mobile apps present opportunities to empower employees with musculoskeletal pain and reduce long-term absenteeism. However, adoption remains limited because of a lack of empirical evidence and challenges in user-friendly design.

Objective: This pilot study aimed to evaluate the potential effects of a fully automated, app-based pain management intervention tailored for employees. Specifically, the study aimed to (1) assess the effect of the intervention on maladaptive pain perceptions, pain intensity, and physical activity and (2) identify factors influencing its effectiveness.

Methods: A total of 66 employees from a Belgian university hospital who had been experiencing musculoskeletal pain for at least 6 weeks participated in a 24-week intervention. The app-based intervention focused on reducing maladaptive pain perceptions, providing work-related guidance, and promoting healthy activity habits through interactive modules, real-time recommendations, and goal-setting features. Every 6 weeks, participants completed a questionnaire measuring maladaptive pain perceptions (pain catastrophizing and fear-avoidance beliefs). Pain intensity was recorded daily using a visual analog scale, and step count was tracked daily using an activity tracker. In addition, semistructured interviews were conducted with 12 participants to explore how they engaged with the intervention and perceived its impact.

Results: Quantitative analysis showed a significant reduction in pain catastrophizing (B=−0.83, $P<.001$, $d=-0.27$), with greater decreases observed among participants with higher baseline scores ($\sigma=-0.38$; $P=.09$). No significant overall change was found in fear-avoidance beliefs (B=−0.35; $P=.15$), although individual trajectories varied ($\sigma^2=1.34$; $P=.04$). Pain intensity also showed significant variability across participants ($\sigma^2=17.29$; $P=.03$) despite no overall effect (B=−0.37; $P=.67$). No significant change was observed in the daily step count (B=107.50; $P=.23$). Qualitative analysis revealed that the effectiveness of the intervention was hindered by content and design choices that did not adequately account for diverse work settings and the busy lives of employees. Cognitive biases and nonsupportive work environments further complicated the successful implementation of the intervention in the workplace.

Conclusions: This pilot study demonstrates the potential of an app-based intervention to support employees with musculoskeletal pain by reducing pain-related fear and promoting active coping strategies. While promising for some, digital interventions alone may be insufficient for employees with more complex needs. Blended approaches and integration within supportive workplace environments are likely essential to enhance effectiveness and promote sustainable work participation.

(JMIR Form Res 2025;9:e67886) doi:10.2196/67886

XSL•FO
RenderX

KEYWORDS

mobile health; mHealth; occupational health; musculoskeletal disorders; mixed methods; pain perceptions; pain management; digital health; human-centered design

Introduction

Background

The European Agency for Safety and Health at Work has highlighted that 58% of the European working population is affected by musculoskeletal pain, which is significantly higher than the global average of 34% [1,2]. Musculoskeletal conditions are a major cause of long-term absenteeism, responsible for up to 60% of cases in Europe [3]. In Belgium, these concerns accounted for 32% of long-term sickness absence in 2022, second only to mental health conditions [4]. These statistics underscore the urgent need for effective interventions to reduce the burden of musculoskeletal pain on employees and the workforce.

However, despite the pressing need, access to appropriate interventions remains a substantial challenge. European legislation mandates periodic health examinations based on occupational risk exposure, leaving approximately 30% of employees without proactive occupational health support [5,6]. Furthermore, the growing shortage of occupational physicians, combined with the limited capacity of specialized pain centers, exacerbates delays in care [7-9]. In Belgium, waiting times for treatment at a specialized center range from 1 to 3 months [10]. By the time employees receive treatment, many have already experienced prolonged sickness absence, making return to work more difficult [11]. These barriers highlight the urgent need for scalable workplace health strategies that facilitate timely interventions to prevent long-term absenteeism.

Mobile health (mHealth) technologies use mobile devices to provide information, deliver treatments, and support self-monitoring, allowing individuals to manage their conditions more conveniently [12,13]. The growing popularity of mHealth has led to the rapid expansion of pain management apps, with 279 apps already available by 2014 [14]. However, most of these apps have substantial limitations. Nearly 60% provide only basic self-care information, lacking essential behavioral components such as goal setting. Moreover, 92% of these apps were developed without input from end users or domain experts, and only 0.4% had undergone clinical validation, raising concerns about their effectiveness.

More recently, efforts to translate validated pain management interventions (eg, mindfulness, education, and guided exercise) into digital formats have shown promising results [15-18]. However, these pain management apps have not been designed to address the unique challenges of employees managing musculoskeletal pain in the workplace. A systematic review by Grant et al [19] identified symptom management, work relations, and workplace adjustments as major barriers to returning to work for employees with musculoskeletal pain. In addition, pain-related fear was highlighted as a major challenge. According to the fear-avoidance model, maladaptive pain perceptions (eg, pain catastrophizing and fear-avoidance beliefs) contribute to hypervigilance and activity avoidance, increasing the risk of persistent pain and disability [20-24]. Research further suggests that pain perceptions and work-related factors interact, influencing the risk of pain and related absenteeism [25,26]. Notably, Cullen et al [27] found that return-to-work interventions that integrate health improvement, work relations, and modified working conditions are more effective than single domain approaches. Therefore, mHealth interventions for employees with musculoskeletal pain should integrate evidence-based pain management strategies within the broader occupational context.

Objectives

To address this gap, we developed an app-based intervention to provide actionable pain management strategies within the occupational context. The intervention focused on reducing maladaptive pain perceptions, providing work-related guidance, and promoting healthy activity habits through interactive modules, real-time recommendations, and goal-setting features. The app was developed using a human-centered design approach, involving 320 employees with musculoskeletal pain and 8 health care professionals [28,29]. This iterative process ensured that the final prototype was user-friendly and aligned with employees' needs.

Establishing the clinical effectiveness of this intervention is essential for its integration into workplace health strategies. This pilot study evaluated its impact on maladaptive pain perceptions. Given the role of maladaptive pain perceptions in pain persistence and avoidance behavior, this pilot study also assessed whether the intervention contributed to reductions in pain intensity and improvements in physical activity [20]. Finally, factors influencing the intervention's effectiveness were explored, as understanding these factors is critical for optimizing the intervention in the future [30].

Methods

Study Design

Participants were recruited between August and October 2021 using a convenience sampling approach. Employees from a Belgian university hospital were informed about the project through various channels, including mailings and brochures. Those interested were asked to contact the researchers for additional information. Eligible participants were aged at least 18 years, had been experiencing musculoskeletal pain for at least 6 weeks, were proficient in Dutch, had access to a smartphone (Android or iOS), and provided written or electronic informed consent. Exclusion criteria included pregnancy or planned retirement between November 2021 and May 2022.

In November 2021, the participants received a Fitbit Inspire activity tracker (Fitbit Inc) and were granted access to the pain management app, Health Empower (available on the App Store and Play Store). Upon logging in and securing their accounts, participants were asked to complete demographic and background information along with the baseline questionnaire.

XSL•FO
RenderX

They were encouraged to use the Health Empower app and Fitbit for 24 weeks and complete a diary and 6-weekly follow-up questionnaires.

After 24 weeks, participants were invited to participate in a semistructured interview. These web-based interviews, conducted via Microsoft Team, lasted approximately 60 minutes. At the start of the interviews, participants were briefed on the procedure and purpose. The interviews were recorded, transcribed verbatim, and the recordings were subsequently deleted.

Ethical Considerations

This pilot study was approved by the Ethics Committee Research UZ/KU Leuven (S-65610) and adhered to Belgian and international privacy and ethical regulations. All participants provided written informed consent before participation. The consent process ensured that participants were fully informed about the study's purpose, procedures, potential risks, and benefits before agreeing to participate.

Participation in the study was voluntary, and participants could withdraw at any time without any consequences. There was no cost or fee associated with participation, and no financial compensation was offered.

To maintain participant confidentiality, all collected data were stored in secure, access-controlled databases and were deidentified before analysis. All data presented in this manuscript and in any supplementary materials are fully anonymous and cannot be traced back to individual participants.

Intervention

The app-based intervention consisted of 6 modules, each addressing distinct aspects of pain management through information, exercises, and actionable recommendations (Figure 1). Modules 1 and 2 focused on improving cognitive and affective responses to pain. Module 1 included pain neuroscience education, explaining the neurophysiology of pain and its contributing factors, aiming to reconceptualize pain and reduce its threat value [31,32]. Participants had to complete this module before gaining access to the other modules. The second module expanded on recognizing and addressing unhelpful thoughts and emotions through cognitive behavioral therapy, including mindfulness techniques [33]. Modules 3, 4, and 5 focused on the work context. In the third module, participants learned about ergonomics and activity management [34,35]. The information was tailored to sedentary and physically demanding jobs, including guidelines for workstation organization, manual handling of loads, work postures, work-rest cycles, and activity pacing. Ergonomic advice was aligned with pain neuroscience education to avoid conflicting messages [31,36,37]. The fourth module focused on stress management and resilience training, including self-regulation, problem-solving, and goal-setting techniques [33,38,39]. The fifth module promoted social support through empathic communication and conflict resolution skills [38]. The final module encouraged healthy lifestyle habits, offering advice on physical activity, sleep, and nutrition [40,41].

Figure 1. Example screens of the Health Empower app: (A) coaching modules; (B) pain reaction record; (C) activity log; and (D) goal setting.

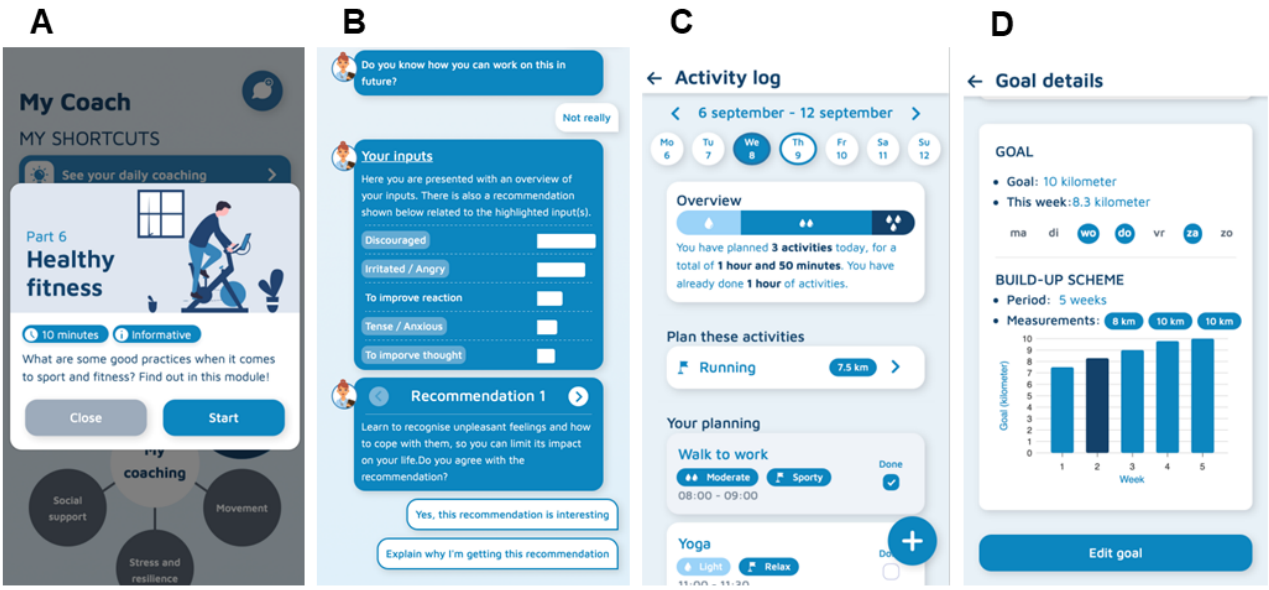

The app also featured a pain reaction record and an activity log (Figure 1) [31,42]. The pain reaction record allowed users to select statements reflecting their thoughts and coping strategies during pain flare-ups, with tailored recommendations provided by a hierarchical algorithm to improve their response in future situations. The development of the pain reaction record and recommender system is discussed elsewhere [43,44]. The activity log included tools for planning activities, setting personal goals, and monitoring progress [45,46]. To support goal setting and behavioral change, the app was integrated with the Fitbit Inspire activity tracker, allowing participants to synchronize their daily step count. If a participant set a daily step goal in the app, Fitbit data were used to track progress within their graded activity scheme, helping them gradually increase physical activity in a structured manner.

To support sustained engagement with the app and study procedures, several built-in features were included. Participants

XSL•FO
RenderX

were limited to completing 1 new module chapter per day, encouraging spaced learning and promoting daily app interaction. All previously completed content remained accessible for review at any time. To enhance adherence to the study protocol, a visual timer was displayed for both the daily (24-h window) and 6-weekly (7-d window) questionnaires, indicating how much time remained for completion. Participants could also enable push notifications to receive alerts when a new coaching chapter or questionnaire became available. In addition, they had the option to set up custom notifications at preferred times and days of the week. These reminders were triggered if a participant had not yet accessed the coaching content that day or if there was still an open questionnaire requiring completion.

Measurements

Sociodemographic and Background Information

Musculoskeletal pain concerns were assessed using the Dutch Musculoskeletal Questionnaire, where participants reported pain or discomfort in specific body regions over the past week (ie, the neck, shoulders, back, elbows, wrists and hands, hips, knees, or ankles and feet) [47]. Multisite pain was considered if participants reported pain in more than 1 body part. Participants also provided information on the duration of concerns and the frequency of absenteeism because of pain in the previous year. Additional sociodemographic variables collected included age, sex, level of education, profession, work arrangement, and tenure.

Work-Related Factors

Work-related factors measured at baseline included physical job demands, workload, job autonomy, and social support. Physical job demands were evaluated using 4 of 12 items from the Dutch Musculoskeletal Questionnaire [47]. Participants rated the frequency of engaging in various physical tasks, such as standing, sitting, or handling loads, on a scale from (almost) never to (almost) always. Exploratory factor analysis identified 2 factors with eigenvalues >1.00 (Multimedia Appendix 1). One factor, representing physically demanding work, included 4 items and demonstrated good internal consistency (Cronbach α=0.86). The composite score for this factor ranged from 4 to 16, with higher scores indicating greater physical demands at work. Workload and social support were evaluated using subscales of the Short Inventory to Monitor Psychosocial Hazards, a validated questionnaire that assesses major psychosocial hazards at work [48]. The workload subscale, consisting of 3 items (scores ranging from 3 to 15), measures the perceived volume of tasks and associated time pressures. The social support subscale, consisting of 4 items (scores ranging from 4 to 20), measures the perceived availability and appreciation from colleagues and supervisors. Both the workload (Cronbach α=0.83) and social support (Cronbach α=0.76) subscales showed good reliability, with higher scores indicating higher workload and greater social support. Job autonomy was measured using a subscale of the Questionnaire on the Experience and Evaluation of Work (version 2.0), a validated questionnaire measuring psychosocial well-being and work-related stress [49]. The autonomy subscale, consisting of 4 items (scores ranging from 4 to 20), evaluates the degree of control and discretion individuals have over various aspects of their work environment, including task planning and choice of work methods. This scale demonstrated good reliability (Cronbach α=0.80), with higher scores indicating higher levels of autonomy.

Maladaptive Pain Perceptions

Pain-related perceptions were assessed through a 6-weekly questionnaire. Catastrophizing was measured using the validated Pain Catastrophizing Scale (PCS), which consists of 13 items evaluating rumination, magnification, and helplessness [50]. The scale had excellent reliability (Cronbach α=0.90) with a total score ranging from 0 to 52 and higher scores indicating greater degrees of catastrophizing. Fear-avoidance beliefs were assessed using the validated Fear-Avoidance Beliefs Questionnaire [51]. This questionnaire includes two subscales: (1) the FABQA work subscale (FABQW), comprising 7 items (scores ranging from 0 to 42), which measures attitudes about work and its relationship to musculoskeletal concerns, and (2) the FABQA physical activity subscale (FABQA), comprising 4 items (scores ranging from 0 to 24), which evaluates perceptions of movement and physical activities. The FABQW demonstrated good reliability (Cronbach α=0.82), whereas the FABQA demonstrated poor reliability (Cronbach α=0.42). Higher scores on each subscale indicated higher levels of fear-avoidance beliefs.

Pain Intensity

Participants indicated their pain intensity on a visual analog scale (VAS), ranging from 0 (“no pain”) to 100 (“worst pain”) daily. The adaption of the VAS to digital platforms, including smartphones, has been validated in previous studies [52].

Physical Activity

Daily step count, measured by the Fitbit Inspire, was used to estimate physical activity. Research has demonstrated the accuracy of Fitbit activity trackers in measuring step count in both laboratory and free-living settings [53]. Previous studies have associated step count with physical activity, indicating individuals with <5000 steps per day as sedentary and those with at least 10,000 steps per day as active [54].

Personal Experience With the Intervention and Perceived Impact on Health

After the 24-week intervention, semistructured interviews were conducted in 2 parts. The first part explored participants’ perceived impact of the intervention through open-ended questions, while the second part presented illustrations of various app features, prompting participants to discuss their experiences with each one. The interview guide was developed following the qualitative research framework outlined by Braun and Clarke [55]. Two researchers compiled an initial set of questions relevant to the study’s objectives. The questions were structured, progressing from general to more specific topics, and were iteratively refined to ensure clarity and appropriateness. To encourage deeper reflection and uncover underlying thought processes, probing follow-up questions were incorporated, guided by the laddering theory [56]. The final interview guide was piloted with a third researcher.

XSL•FO
RenderX

Statistical Analysis

Sample Size

An a priori power analysis was conducted using G*Power [57] to determine the required sample size. Kristjánsdóttir et al [58] reported a moderate improvement in pain catastrophizing (Cohen *d*=0.74) after a 5-month intervention using a smartphone app for individuals with pain. To detect a moderate effect size at a 5% significance level with a power of 80%, a minimum group of 28 participants was required. Considering a 50% dropout rate, enrolling at least 56 participants was recommended [59].

Quantitative Analysis

Quantitative data were analyzed using the SPSS statistical software package (version 28.0; IBM Corp). Univariate descriptive statistics were conducted to summarize sociodemographic and background variables. Linear multilevel models were used to evaluate changes in pain perceptions (PCS, FABQA, and FABQW), pain intensity (VAS), and physical activity (daily step count) [60]. Given the hierarchical structure of the data, with repeated measurements nested within individuals, this approach accounts for both within-individual and between-individual variation. Random intercepts were included to allow each participant to have a unique baseline value, while random slopes accounted for individual differences in the rate of change over time. In addition, the model estimated the covariance between intercepts and slopes, providing insight into whether baseline values influenced subsequent trajectories. To assess pain intensity and physical activity, the average scores were calculated across 5 periods: 4 periods of 5 consecutive weeks and 1 period of 4 consecutive weeks. Periods with fewer than 5 valid measurements were excluded, and days with fewer than 500 recorded steps were also considered invalid [61,62]. To maximize data inclusion, multilevel models use a full-information maximum likelihood approach, which estimates parameters based on all available data points rather than excluding participants with incomplete data [60].

Qualitative Analysis

Qualitative data from the semistructured interviews were analyzed using the Qualitative Analysis Guide of Leuven [63]. This 10-step approach included a preparation phase and a coding phase. First, interviews were transcribed and summarized into brief abstracts. Concrete experiences were transformed into concepts, and these were compared across interviews. Common concepts served as initial codes for the coding process and underwent refinement over several iterations. The final codes were integrated into a meaningful conceptual framework in relation to the research questions. Two researchers conducted the analysis, resolving disagreements through consensus or consultation with a third researcher. To ensure anonymity and maintain neutrality, all participant quotes are presented using the pronoun “he” throughout the paper, regardless of the participant’s sex.

Integration of Quantitative and Qualitative Results

Two researchers evaluated the alignment between the quantitative and qualitative results, encompassing confirmation, expansion, or discordance [64]. Disagreements were resolved through consensus or consultation with a third researcher. The metainferences derived from this analysis were presented in a side-by-side joint display for each research question.

Results

Participant Characteristics

A total of 96 employees expressed interest in participating in the study (Figure 2). Overall, 27 were excluded for not meeting the inclusion criteria or, failing to provide informed consent and 3 never logging into the app. The remaining 66 employees (n=57, 86% female) were included in this study. An overview of sociodemographic and background information is provided in Table 1. Baseline characteristics were comparable between participants who completed the study (ie, completed the 24-week follow-up questionnaire) and those who dropped out (Multimedia Appendix 2).

XSL•FO
RenderX

Figure 2. Flowchart of the study.

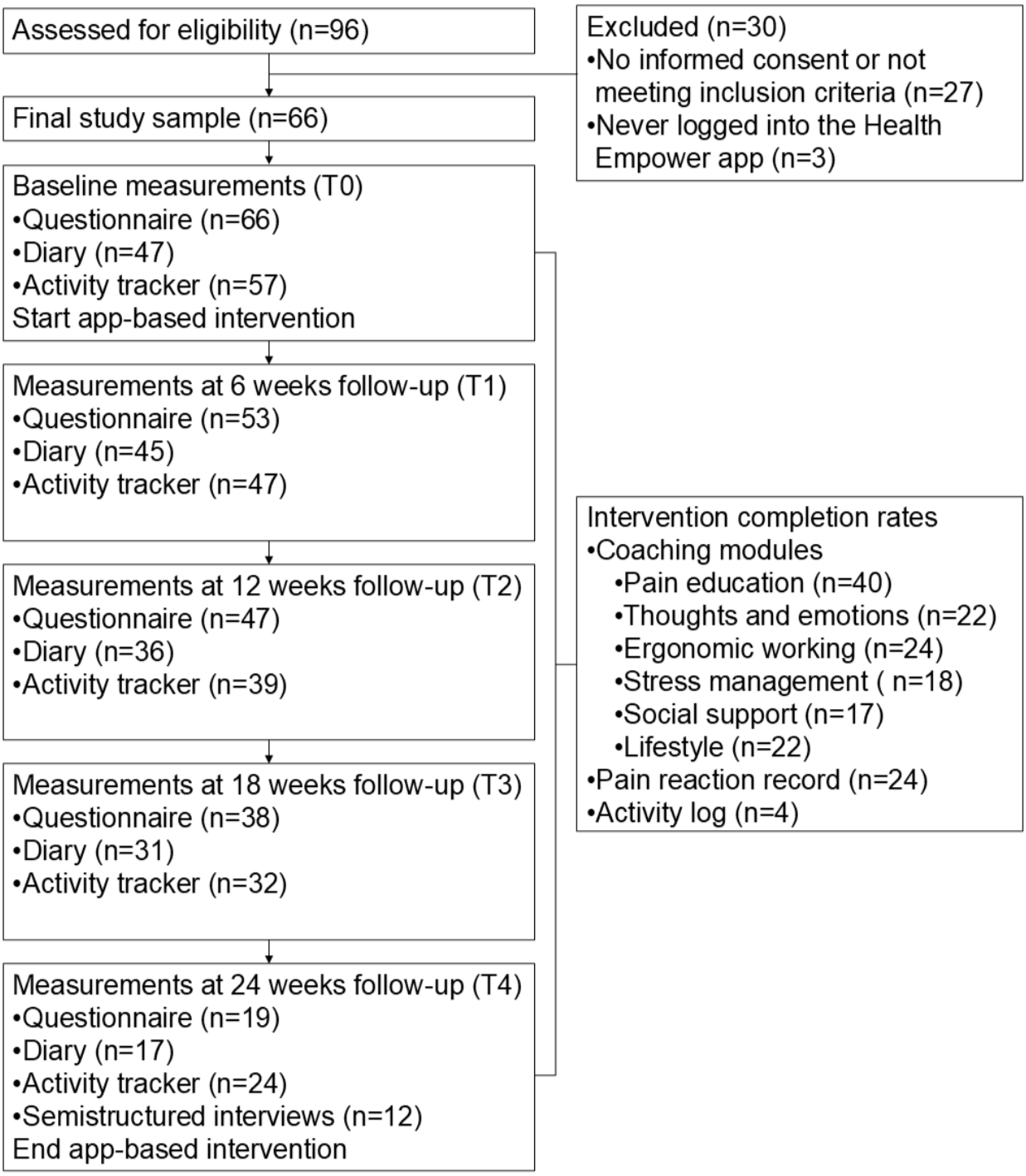

XSL•FO
RenderX

Table 1. Sociodemographic and baseline information of study participants (n=66).

Characteristics	Values
Age (y), mean (SD)	42 (10.8)
Sex (female), n (%)	57 (86)
Degree (bachelor or higher), n (%)	55 (83)
Profession (health care), n (%)	60 (91)
Tenure (≥10 y), n (%)	37 (56)
Work arrangement (full time), n (%)	35 (53)
Multisite pain (yes), n (%)	59 (89)
Pain duration (≥3 mo), n (%)	61 (92)
Pain-related sick leave (≥1 mo), n (%)	3 (5)
Work-related factors, mean (SD)	
Physical job demand	6.2 (2.7)
Workload	9.5 (2.5)
Job autonomy	12.9 (3.8)
Social support	14.2 (3.5)
Maladaptive pain perceptions, mean (SD)	
Pain catastrophizing	13.3 (7.7)
Fear-avoidance beliefs (activities)	9.3 (4.3)
Fear-avoidance beliefs (work)	15.2 (7.4)
Pain intensity (VAS[a]), mean (SD)	33.0 (18.4)
Physical activity (daily step count), mean (SD)	9564 (2579.3)

[a]VAS: visual analog scale.

Effect of the App-Based Pain Management Intervention

Metainferences of quantitative and qualitative results on the potential effect of the intervention are presented in Table 2. Overall, there was alignment between the quantitative and qualitative results, with qualitative insights explaining observations in the quantitative data.

XSL•FO
RenderX

Table 2. Joint display of quantitative and qualitative findings on the intervention effect.

Theme	Quantitative findings	Qualitative findings	Metainferences
Maladaptive pain perceptions	• There was a small but significant decrease in PCS[a]. • Variance in regression slopes for FABQA[b] was significant.	• Participants felt reassured by the information. • Some adopted a biopsychosocial mindset, while others reinforced a biomedical perspective.	• The reduction in PCS aligns with feelings of reassurance. • Varying regression slopes for FABQA reflect shifts toward a biopsychosocial or biomedical perspective.
Pain intensity	• Variance in regression slopes for VASc was significant.	• Participants had varied experiences regarding pain evolution. • Pain reduction was often attributed to other interventions.	• Varying regression slopes for VAS reflect diverse experiences regarding pain evolution. • Qualitative findings expand on the intervention effect by attributing pain reduction to other interventions.
Physical activity	• There was no significant change in daily step count.	• Most participants reported minimal lifestyle changes. • Participants reported increased awareness of activity levels.	• No change in step count is consistent with reports of minimal lifestyle changes. • Qualitative findings expand on intervention effect by highlighting increased awareness of activity levels.

[a]PCS: Pain Catastrophizing Scale.

[b]FABQA: Fear-Avoidance Beliefs Questionnaire (physical activity subscale).

[c]VAS: visual analog scale.

Pain Perceptions

Quantitative data analysis showed a decrease in pain catastrophizing (B=−0.83, P<.001, d=−0.27), reflecting reduced pain-related fear (Table 3; Multimedia Appendix 3). The marginally significant variance in individual slopes for pain catastrophizing (σ^2=1.00; P=.09) and the negative intercept-slope covariance (σ=−0.38; P=.09) suggest that participants with higher baseline catastrophizing tended to show a more pronounced decrease over time. Furthermore, the significant variance in individual slopes for fear-avoidance beliefs regarding physical activities (σ^2=1.34; P=.04) indicates increasing fear-avoidance beliefs in some participants and decreasing beliefs in others (Table 3; Multimedia Appendix 3).

Qualitative data analysis revealed two subthemes regarding the impact of the intervention on pain perceptions: (1) worries and insecurities and (2) biopsychosocial mindset. Consistent with the observed decrease in pain catastrophizing scores, many participants found reassurance in the information provided by the intervention, feeling that reducing the impact of pain on their lives was possible. This reassurance inspired them to optimize their management strategies. The recognition of their existing coping efforts also provided validation and motivation to continue these strategies. For instance, participant 2 explained how the intervention confirmed that his current pain management approaches were effective, reinforcing his confidence:

> *When I read the information in the app, I often thought, "Well, yes, I can actually manage my pain this way too." Sometimes I was already doing it, but it was nice to have confirmation that I wasn't doing anything wrong and that I should keep going. That information really helped me.* [Participant 2]

The variance in regression slopes for fear-avoidance beliefs was reflected in the diverse perspectives on the factors contributing to pain and strategies for coping after the intervention. For example, participant 3 continued to emphasize physical factors as the primary contributors to his pain. This participant attributed his pain to past injuries or surgeries and viewed his work environment as an important aggravator:

> *I had surgery for a cartilage injury, and my joints inflame easily. I don't believe my pain is psychological. Stress might affect muscle tension, but it's not as detrimental as wear and tear...My posture isn't great when I'm working at the computer. I think this has caused muscle imbalances that led to my neck pain.* [Participant 3]

This focus on physical contributors led participants, such as participant 5, to prioritize strategies aimed at reducing physical strain. For example, this participant adopted avoidance behaviors to prevent further pain:

> *I'm much more conscious of my movements and posture now. I don't lift things by bending over anymore. I bend my knees or get down on one knee to pick something up. I learned that from the app.* [Participant 5]

In contrast, several participants recognized the role of psychosocial factors in their pain, acknowledging that cognitive and emotional processes can influence their pain experience. For instance, participant 7 reflected on how the intervention shifted his perspective from viewing pain primarily as a physical issue to a biopsychosocial phenomenon, helping him recognize the drawbacks of avoidance and the benefits of staying active. Consequently, this participant became less concerned with ergonomics to reduce physical strain:

XSL•FO
RenderX

> *I didn't realize how convinced your brain can be that you'll feel pain during certain activities...One of the things I learned was that staying still is actually bad. But what do many people with pain do? They just sit there, and it only gets worse....I found the idea of stopping complaining about the pain and just getting moving really useful...I didn't spend much time on the information about lifting, I don't think it's that important in my situation.* [Participant 7]

Table 3. Multilevel analysis of pain perceptions, pain intensity, and physical activity.

Parameter	Estimate (SE)	*P* value
Pain catastrophizing		
Fixed effect	−0.83 (0.23)	<.001[a]
Random intercepts	44.15 (9.09)	<.001
Random slopes	1.00 (0.59)	.09
Covariance (intercept-slope)	−0.38 (0.22)	.09
Fear-avoidance beliefs (activities)		
Fixed effect	−0.35 (0.24)	.15
Random intercepts	8.40 (2.74)	.002
Random slopes	1.34 (0.64)	.04
Covariance (intercept-slope)	−0.12 (0.29)	.67
Fear-avoidance beliefs (work)		
Fixed effect	−0.25 (0.26)	.34
Random intercepts	44.16 (9.70)	<.001
Random slopes	0.81 (0.71)	.23
Covariance (intercept-slope)	−0.11 (0.32)	.72
Pain intensity (VAS[b])		
Fixed effect	−0.37 (0.84)	.67
Random intercepts	292.60 (66.88)	<.001
Random slopes	17.29 (8.07)	.03
Covariance (intercept-slope)	−0.26 (0.19)	.17
Physical activity (daily step count)		
Fixed effect	107.50 (87.48)	.23
Random intercepts	5,667,049.59 (1,256,486.10)	<.001
Random slopes	108,345.09 (73,646.87)	.14
Covariance (intercept-slope)	0.34 (0.33)	.30

[a]Small effect size (Cohen d=−0.27).

[b]VAS: visual analog scale.

Pain Intensity

Quantitative data analysis showed a significant variance in regression slopes for pain intensity scores (σ^2=17.29; P=.03; Table 3; Multimedia Appendix 3). While pain intensity increased for some participants, it decreased for others.

The qualitative data revealed two subthemes related to pain intensity: (1) evolution of pain and (2) reasons for pain reduction. Consistent with the quantitative data, participants provided diverse perspectives on the evolution of their pain. Many participants expressed disappointment that their pain had not decreased after the intervention. For example, participant 9 was frustrated with the limited impact of the intervention on his pain, noting that it did not provide the actionable tips he had expected:

> *A lot of people ask me, "Has it helped with the intensity of the pain?" and I have to answer, "No," because I was expecting more specific tips that I could try out* [Participant 9]

Despite not experiencing a general improvement in their pain concerns, some participants recognized that the intervention helped reduce the burden of pain in specific situations. For instance, participant 12 shared how the intervention helped him communicate his needs more effectively, thereby reducing stress and pain:

XSL•FO
RenderX

> *I've learned to inform others about my pain problem without complaining. Just a brief "look, I can't do this alone, I need help because of my neck," and that reduces my stress and pain...But it's not gone, of course.* [Participant 12]

Participants who experienced improvements in their pain often viewed the intervention as secondary to other factors. They believed that changes might have occurred over time or were attributable to other interventions. Participant 6 conveyed this ambiguity, questioning whether the intervention had any real impact on his pain:

> *I have to say that my pain has decreased...But is it because of the app? I don't really feel that way. It might just be coincidental since my pain was already starting to decrease, so it could be a combination of both.* [Participant 6]

Physical Activity

Quantitative data analysis showed no significant effect of the intervention on daily step count (Table 3; Multimedia Appendix 3).

Two subthemes emerged from the qualitative data regarding the effect of the intervention on physical activity: (1) awareness and (2) physical activity habits. Many participants reported that the intervention heightened their awareness of daily activities, leading them to reflect on their daily routines. For instance, participant 2 described how the daily questionnaires encouraged him to think critically about his activity level:

> *You had to fill out your questionnaire every evening and reflect on your day...I might have sat down too much or walked a lot...It made me aware that maybe I had done too little or too much. I found that really useful.* [Participant 2]

This awareness, combined with Fitbit prompts, served as a motivator for some participants to move more and rethink their sedentary habits. Participant 1 explained how tracking his activities encouraged him to be more active:

> *Both the Fitbit and the app motivate you to think more about your health and to sit less. Since you have to record how long you've been sitting and walking, it makes you realize. I never really thought about it before.... Now, I'm more likely to go for a walk or do something else instead of lying on the couch. When you see that you've only taken a few steps today, you make sure to walk more the next day.* [Participant 1]

However, overall changes in physical activity habits were minimal, which is consistent with the quantitative findings. Participants with physically demanding jobs, such as participant 9, reported that the intervention had little impact on their routines:

> *I have a demanding job and my household responsibilities...I do stand still a lot at work, but I also move around a lot. By the evening, I just don't feel like doing more. I do check the Fitbit, but it doesn't motivate me to go outside in the evening to reach ten thousand steps. My life hasn't changed because of the Fitbit.* [Participant 9]

Factors Influencing the Effectiveness of the App-Based Pain Management Intervention

An analysis of the semistructured interviews identified three themes related to factors influencing the effectiveness of the intervention: (1) the intervention quality, (2) the participant's personal characteristics, and (3) the work context. The metainferences of the quantitative and qualitative results on these factors are presented in Table 4. Overall, there was alignment between the quantitative and qualitative results, with the qualitative insights explaining observations in the quantitative data.

XSL•FO
RenderX

Table 4. Joint display of quantitative and qualitative findings on factors influencing effectiveness.

Theme	Quantitative findings	Qualitative findings	Metainferences
Intervention quality	• In total, 36% of participants completed each module, 36% used the pain reaction record, and 6% used the activity log. • Final follow-up questionnaire was completed by 29% of the participants.	• Some found the content clear and useful, while others saw it as too basic or impractical. • Interactive elements were well-received, but limited flexibility and integration options hindered engagement.	• Nonuse and dropout rates align with mixed perceptions of content relevance and practicality. • Nonuse and dropout rates reflect engagement barriers related to intervention design.
Personal characteristics	• Little sick leave and favorable baseline scores were reported for pain perceptions, pain intensity, and physical activity. • Completion rates were higher for physical-focused modules (35%) compared with psychosocial-focused ones (29%).	• Pain concerns were often perceived as manageable, leading to casual engagement with the app. • Cognitive biases, including perceived expertise and skepticism, influenced engagement with the app.	• Limited impact of pain on life, as reflected by favorable baseline scores, corresponds to high nonuse. • Cognitive biases reflect high nonuse and lower completion rates for psychosocial-focused modules compared with physical-focused ones.
Work context	• Favorable baseline scores were reported for physical job demands, workload, social support, and job autonomy.	• Limited access to adjustable equipment and shared workstations made it difficult to apply ergonomic recommendations. • Lack of understanding and support at work made it difficult to implement work modifications and coping strategies.	• Despite favorable work-related scores, qualitative findings reveal structural barriers hindering ergonomic adjustments. • Despite favorable social support ratings, qualitative findings reveal workplace dynamics hindering coping and modifications.

Intervention Quality

Follow-up data showed relatively high nonuse and dropout rates (Figure 2). On average, each module was completed by only 24 (36%) participants. The pain reaction record was used by 24 (36%), while the activity planner was used by only 4 (6%) participants. Regarding dropout, 47 participants completed the 12-week follow-up questionnaire 28 (71%), but only 19 (29%) completed the 24-week follow-up.

Two subthemes emerged from the qualitative data regarding the quality of the intervention: (1) content and (2) design. Consistent with the high nonuse and dropout rates, mixed reactions to content and design suggested that the intervention struggled to maintain long-term engagement and meet the diverse needs of its users. Some found the information provided clear, actionable, and accessible. For instance, participant 7 appreciated the straightforward, nonmedical language:

> *I found the content good and accessible for a broad audience ... not too medical, just very basic but with enough information. And the videos in between were well done.* [Participant 7]

However, some participants felt the advice was too basic or impractical for their specific situations. Participant 1 highlighted the difficulty of applying certain ergonomic tips because of his use of bifocals:

> *Tips like "your screen should be at eye level".... Sorry, that doesn't work for me because I have bifocals, so my screen needs to be lower to read properly...Also the tips on how to sit properly, I can't keep that up for two hours.* [Participant 1]

In addition, some participants, such as participant 6, found that the intervention focused too much on pain, particularly through repetitive questionnaires. This emphasis was counterproductive, making participants more fixated on their pain rather than helping them manage it:

> *I am not someone who is fixated on pain all the time. Reflecting on pain in the questionnaires was a drawback of this study for me. I didn't want to always think about pain.* [Participant 6]

The design of the app, including its usability and visual appeal, also impacted participant engagement. Many participants appreciated the interactive elements, such as videos and quizzes. Familiar design choices, such as chat-style questionnaires, were particularly well-received, as noted by participant 12:

> *It wasn't just text, but also videos and quizzes. I enjoyed testing myself to see if I understood everything...The chat-style questionnaire was very intuitive. I buy a lot online so it was recognizable...The colors were neutral, which was easy on the eyes.* [Participant 12]

Despite these positives, several design elements did not align well with the busy lives of the participants. A lack of flexibility was a common concern. Participant 9 expressed frustration with not being able to advance through the content when he had time available:

> *It would have been better if there was more flexibility.… It was disappointing that I couldn't move on when I had some time. Don't block people, let them explore further. People have limited time because of work. When they find time, let them continue with the modules.* [Participant 9]

XSL•FO
RenderX

Time-consuming actions and poor integration with other tools also detracted from the experience. Some participants found the app redundant, particularly if they were already using more comprehensive tools. For example, participant 5 preferred his own activity tracker because it featured automated tracking instead of questionnaires:

> *I have another activity tracker with more options. It automatically tracks heart rate and activities, so I always checked that app.... I mostly used your app to go through the modules, but I didn't really use the other features because I had something else, and maybe even better.* [Participant 5]

Furthermore, notifications, while helpful, were often insufficient to maintain regular engagement with the app. Participant 11 suggested that more prominent alerts, similar to those used in social media apps, might have encouraged more consistent use:

> *What would help me is something like a red dot or a number, like on social media apps...That would remind me that there is something to do. I got the notifications for the questionnaires, but that wasn't enough to keep me actively engaged.* [Participant 11]

Personal Characteristics

Completion rates varied considerably between coaching modules (Figure 2). While 61% (n=40) of participants completed the first module on pain education, the modules focusing on physical topics (ie, modules 3 and 6) had an average completion rate of 35% (n=23), and those on psychosocial topics (ie, modules 2, 4, and 5) had a lower average completion rate of 29% (n=19).

Two subthemes emerged from the qualitative data regarding personal characteristics: (1) the impact of pain on daily life and (2) cognitive biases. The differences in completion rates of coaching modules highlighted how participants' beliefs and attitudes influenced their engagement with various aspects of the intervention. For many participants, pain had a manageable impact on their daily lives, leading to a more casual engagement with the app. For instance, participant 11, who described his pain as an inconvenience rather than debilitating, participated out of curiosity:

> *I didn't have specific expectations about the study. I thought, "Well, if it helps, great; if not, no harm done".... I have to say, I didn't complete or use everything. I often thought, "It's not bad, I'm doing well." My back pain is unpleasant, but manageable. I live my life; I don't stay home just because my back hurts.* [Participant 11]

This mindset aligns with favorable baseline scores on pain perceptions and pain intensity, as well as the relatively high daily step count at baseline and low self-reported sick leave (Table 1).

Furthermore, many participants felt they were already well-versed in pain management, whether through personal experience, previous treatments, or a professional background. This existing knowledge often led to limited engagement with the app, as they did not expect to learn new information. Participant 8 reflected on how much content was a review of what he had already learned:

> *I found it valuable to check if there were things I didn't know or could improve...But a lot of it was repetition, as I had already gone through a similar program at a specialized center. So, I skimmed through some parts fairly quickly.* [Participant 8]

Participants also had different perspectives on the biopsychosocial approach of the intervention. For example, participant 4 appreciated the cognitive approach as it aligned with his ongoing treatment:

> *I found it interesting because I'm seeing a new physiotherapist who also focuses a lot on how thoughts influence pain and how it's processed in the brain...So, this aligned with that approach, especially since no clear cause has been found for my pain. I have a muscle condition, but I'm not convinced that it explains all my symptoms.* [Participant 4]

In contrast, others were skeptical of the cognitive and psychosocial elements, linking their pain to biomedical factors. Consequently, these participants favored physical approaches to pain management. For example, participant 3 expressed doubt about the effectiveness of cognitive strategies for his pain, limiting his engagement with those aspects of the content:

> *With my background and education, I'm skeptical about the cognitive approach to pain...I didn't have high expectations for improvement through this method, as I usually know what's causing my pain, muscle imbalance or a damaged joint. I'm more inclined to address pain by exercising.... I still read everything in the coaching, but I paid more attention to some things than others.* [Participant 3]

Work Context

Overall, participants reported favorable scores for physical job demands, workload, social support, and job autonomy at baseline, indicating a good balance between job demands and job resources (Table 1).

Two subthemes emerged from the qualitative data regarding work context: (1) ergonomic equipment and (2) social support. Despite the favorable scores on work-related factors, several participants encountered barriers when trying to implement advice from the app. Some participants, such as participant 5, were unable to customize their workspaces because of shared workstations and a lack of adjustable equipment, which often resulted in discomfort or exacerbated pain:

> *We have five different workstations, and at stations three and four, the tables are lower, which is harder for me because I have to bend over and lift patients from a lower position. I notice that on those stations, I have more complaints.... I can't do much about it because those tables have standard heights.* [Participant 5]

Many participants reported that persistent pain often received less recognition compared with more visible health conditions. This lack of understanding and empathy made it difficult for

XSL•FO
RenderX

them to openly discuss their pain and request necessary accommodations. Participant 11 expressed frustration with the dismissive attitudes of some colleagues:

> *Most people don't take it seriously when you say you're in pain. They might listen, but they quickly forget. It's the usual "Just take a pill and keep going."…. There's just a lack of understanding, and I've tried to accept it because they don't get it. If you don't have pain yourself, it's hard to understand.* [Participant 11]

In addition, resistance from colleagues and employers toward modifying work conditions was a recurring issue. This resistance often stemmed from concerns about creating distinctions between employees or fears of reduced productivity. Participant 8 noted the difficulties he faced when returning to work after a long absence because of pain:

> *They don't adjust anything; it's really hard to deal with...I was off work for a year, and when I returned, the first thing they said was that I had to do all the work - all kinds of patients, all shifts...I think just an extra ten-minute break would help, but it's really hard to get that. They don't like it because then the other colleagues have to work harder or cover for me.* [Participant 8]

However, efforts made by colleagues and supervisors to optimize the work environment and reduce physical demands were much appreciated by the participants. These efforts often included providing ergonomic equipment or adjusting work schedules and tasks. For instance, participant 2 highlighted the accommodations made to help manage his chronic pain:

> *I feel like they make an effort to keep people with chronic pain employed...They created a schedule for me where I work three half-days, one full day, and have one day off. They've also tried to ensure I have a variety of tasks, so I can't say they aren't taking it into account.* [Participant 2]

Discussion

Effect of the App-Based Pain Management Intervention

This study evaluated an app-based intervention designed to support employees with musculoskeletal pain by addressing maladaptive pain perceptions, providing work-related guidance, and promoting healthy activity habits.

Consistent with previous research demonstrating that app-based pain management can effectively reduce pain catastrophizing in patient populations, this study suggests that such an approach can also benefit employees [16-18]. The significant reduction in pain catastrophizing scores aligned with participants' reports of feeling reassured. By focusing on pain neuroscience education to shift perceptions of pain from a structural or injury-based model toward a broader understanding of pain mechanisms, participants may have gained a greater sense of control, thereby reducing the perceived threat of pain [31,65]. However, the intervention did not result in significant overall improvements in fear-avoidance beliefs, diverging from previous findings [16]. The significant variability in regression slopes for fear-avoidance beliefs suggests that, for some participants, the intervention reinforced their focus on physical contributors to pain.

Similarly, while prior studies have shown that mHealth interventions can effectively reduce pain intensity, our intervention did not yield a significant overall effect [15]. Moreover, the significant variability in pain trajectories suggests that some participants experienced increased discomfort.

Despite incorporating strategies for graded activity, personal goal setting, and healthy lifestyle habits, the intervention did not lead to significant increases in daily step count [36,41,45,46]. While many participants reported greater awareness of their movement habits, this did not translate into behavioral change. This contrasts with prior studies reporting small to moderate increases in physical activity following mHealth interventions, even without direct human support [66].

Factors Influencing the Effectiveness of the App-Based Pain Management Intervention

The content of the intervention likely influenced its effectiveness. Unlike app-based pain management interventions that integrate structured exercise programs, our intervention primarily focused on education [16]. While modifying cognitive representations of pain can positively impact pain experiences, research suggests that education should be combined with exercise therapy to achieve meaningful reductions in pain and disability [65,67,68]. Moreover, exercise programs also provide positive movement experiences, which can help reframe fearful cognitive representations of pain [36,67]. Combining education with progressive exercise may explain why Sitges et al [16] observed significant improvements in fear-avoidance beliefs following app-based pain management. In addition, our intervention provided general ergonomic recommendations, which may not have adequately addressed the specific challenges faced by some employees in physically demanding jobs [69,70]. Prior research has highlighted that physically demanding jobs exhibit greater variability in workstation design compared with office-based roles, making broad ergonomic guidance less effective [71].

Beyond content, design-related factors might have influenced the intervention's effectiveness. Some participants reported that self-monitoring heightened vigilance toward pain. Research has shown that heightened attention to pain can reinforce avoidance behavior and amplify pain perception, potentially explaining the variability in regression slopes for fear-avoidance beliefs and pain intensity [20,32]. In addition, the absence of gamification elements, such as leaderboards and rewards, could explain the limited effect of the intervention on physical activity [72]. A systematic review found that social features, in particular, are effective in promoting physical activity [73]. However, individuals with chronic musculoskeletal conditions often prefer privacy in self-management apps and may be reluctant to engage with social features [74]. Furthermore, several design choices hindered engagement with our intervention, potentially hindering its effectiveness. Participants expressed frustration with restrictive features such as time locks, which prevented them from progressing through content at their own pace. There was also a clear demand for more seamless

XSL•FO
RenderX

and automated functionality, particularly for activity tracking. Given the rapid evolution of digital technologies, users have increasingly high expectations for convenience and personalization [75]; when interventions are perceived as cumbersome or misaligned with individual preferences, the likelihood of disengagement increases.

Individual differences also shaped engagement with the intervention. Many participants reported that pain had a limited impact on their daily lives, which may have contributed to high nonuse rates. In addition, some employees selectively attended to information that aligned with their preexisting beliefs while disregarding content that challenged them, limiting the potential for meaningful change [76,77]. This confirmation bias may be particularly relevant for those with strong fear-avoidance beliefs, who may have used the intervention to validate rather than modify their perceptions of pain. Cognitive biases frequently contribute to errors in clinical decision-making among health care professionals [78]. Moreover, research has shown that biomedical perspectives remain dominant in health care despite the growing emphasis on psychosocial factors in medical education [79-81]. Raising awareness of cognitive biases, combined with simple communication and counterexplanation, can foster critical thinking and openness to new perspectives [82,83]. While pain neuroscience education already integrates some of these strategies, individuals in the precontemplation stage of behavior change (eg, participants who consider themselves experts on their condition) may require additional guidance [31,84].

Finally, the broader workplace context also played a role in the intervention's effectiveness. Participants highlighted several workplace challenges, including a lack of ergonomic equipment and resistance from colleagues and supervisors to modifying work conditions. These challenges may have contributed to the significant variability in regression slopes of fear-avoidance beliefs and pain intensity, as participants who perceived a conflict between the intervention's recommendations and their work environment might have experienced worsening scores. Prior research suggests that nonsupportive work environments can undermine pain management interventions, particularly when colleagues or supervisors question the legitimacy of work modifications or coping strategies [19,69,85,86].

Study Contributions and Future Directions

This was the first study to evaluate an app-based pain management intervention specifically designed for employees. Our findings indicate that mHealth can support the multidimensional treatment of musculoskeletal pain in the workplace by reducing the perceived threat of pain, an important barrier to work participation [19,20,22-24]. Moreover, addressing maladaptive pain perceptions and increasing awareness of activity habits are essential steps in initiating active coping strategies, such as physical exercise, which can improve physical functioning [34,36,87]. These findings suggest that app-based pain management interventions may offer accessible and scalable support for some employees, reducing reliance on intensive face-to-face care. This, in turn, could enable health care providers to focus time and resources on employees at higher risk of long-term disability.

Our results also highlight the value of self-monitoring in capturing the evolution of risk factors over time. In line with prior research, our results suggest that maladaptive pain perceptions may emerge over time as employees continue to struggle with managing pain in the workplace [88]. Therefore, conducting assessments only at the onset of pain concerns may overlook employees whose risk of disability develops gradually over time. Continuous self-monitoring with mobile apps could facilitate the early identification of these at-risk employees. However, self-monitoring tools should be carefully designed to avoid reinforcing hypervigilance to pain signals. Future studies should explore optimal monitoring frequency or investigate automated tracking of physiological parameters as a less intrusive alternative. For instance, a systematic review has demonstrated significant associations between self-reported pain intensity and physiological markers, such as heart rate variability and skin conductance [89].

Despite the potential benefits of mHealth, our findings suggest that digital interventions alone are unlikely to meet the needs of all employees with musculoskeletal pain. Many participants required more tailored advice and additional support to initiate and sustain new coping strategies. This highlights the need for blended interventions that combine digital tools with professional guidance [90]. Research shows that face-to-face ergonomic consultations are more effective than generic self-help materials in reducing musculoskeletal symptoms [91]. Likewise, health care providers can enhance the effectiveness of pain neuroscience education by tailoring it to the individual's context and using empathic communication [65,67,92]. Previous studies have successfully combined digital learning modules with in-person consultations, enabling participants to review educational materials at their own pace while using face-to-face sessions to clarify misconceptions and personalize treatment strategies [93,94].

Finally, our results underscore the importance of embedding pain management interventions within workplace policies. Several participants encountered barriers that limited the implementation of the app's recommendations, including a lack of ergonomic equipment and resistance from colleagues or supervisors. Research suggests that successful return-to-work interventions foster a supportive workplace culture by educating teams about the challenges of chronic pain and involving them in defining appropriate work modifications [27,69,85,95,96]. Occupational services for prevention and protection at work could play a key role by using mHealth data to provide targeted guidance to managers, increasing their willingness and ability to support health-related interventions [70].

Study Limitations

Several limitations of this study should be acknowledged. First, the 1-group pilot design limits causal inference regarding the effectiveness of the app-based pain management intervention. Nevertheless, the integration of quantitative and qualitative data provided a rich understanding of how participants engaged with the intervention and perceived its impact [97]. These insights are valuable for refining the intervention and informing the design of a future large-scale clinical trial [30].

XSL•FO
RenderX

Second, the study sample consisted of employees with persistent musculoskeletal pain who remained in active employment. These participants reported relatively favorable baseline scores for pain intensity, pain perceptions, and physical activity, which may have reduced the observed effectiveness of the app-based intervention. However, the significant variance in individual slopes for pain intensity suggests that the intervention may have been beneficial for a subset of participants. Furthermore, our results also revealed a negative intercept-slope covariance for pain catastrophizing, indicating that participants with higher baseline levels of catastrophizing experienced more pronounced improvements over time. This is consistent with findings from the study by Sitges et al [16], which observed greater improvements among participants with higher baseline fear-avoidance beliefs, and with the study by Mönninghoff et al [66], which reported stronger effects of mHealth interventions in individuals with lower initial physical activity levels. Recruitment through the employer may partly explain the favorable baseline characteristics. Employees with more severe musculoskeletal pain may have been less inclined to participate because of privacy concerns despite IDEWE's strict data protection policies. In addition, recruitment via work-related channels may have inadvertently excluded employees on sick leave, who may report more pronounced maladaptive pain perceptions and less favorable work conditions.

Third, the overrepresentation of health care workers in our sample limits the generalizability of our findings to other sectors [98]. Future studies should explore the effectiveness of mHealth for musculoskeletal pain in more diverse workplace populations, particularly among employees with lower digital health literacy, who may require different communication strategies and additional support [92,99,100].

High dropout and nonuse attrition rates were also a limitation of this study. Although multilevel analysis was used to address missing data, attrition may have influenced the findings [60]. Importantly, discontinuation of app use does not necessarily indicate intervention failure. Some participants may have stopped engaging after they had adopted new coping strategies or reached their personal goals [75]. Future studies should aim to determine the minimum exposure needed to achieve meaningful benefits and establish clearer criteria for evaluating dropout and nonuse attrition.

Finally, the measurement of physical activity warrants consideration. While daily step count is a widely used metric, it does not capture activity intensity [54]. In addition, wrist-worn trackers, such as the Fitbit Inspire, may not accurately record activities that do not involve vertical arm movements, such as cycling [101]. This may have obscured changes in physical activity related to the intervention. Future studies should consider alternative methods for quantifying physical activity levels, such as measuring minutes of moderate and vigorous activity or total energy expenditure [66,101].

Conclusions

This pilot study demonstrates the potential of an app-based intervention to support employees with musculoskeletal pain by reducing maladaptive pain perceptions and promoting active coping strategies. While the intervention showed promising effects, particularly in reducing pain catastrophizing, its overall impact on fear-avoidance beliefs, pain intensity, and physical activity was limited. The findings underscore the value of mHealth as a low-threshold solution for some employees while also highlighting the need for blended approaches that incorporate professional guidance for those requiring additional support. Moreover, the results emphasize the importance of embedding pain management strategies within workplace policies, as nonsupportive work environments might hinder treatment efficacy. Together, these insights offer valuable guidance for the development of more personalized, scalable, and context-sensitive interventions aimed at improving self-management and work participation among employees with musculoskeletal pain.

Acknowledgments

This pilot study is a part of a broader research project titled Personal Health Empowerment, which aimed to develop and evaluate innovative monitoring and coaching solutions for employees with pain concerns. This project is funded by ITEA (grant 16040), a European financier of innovative projects that have social added value and make use of technological innovations, and the Flemish Agency for Innovation & Entrepreneurship (grant HBC.2018.2012), which provides grants to companies for the development or strengthening of their research and development activities.

Authors' Contributions

SK contributed to the design of the study, conducted data collection, performed the quantitative and qualitative analyses, contributed to data interpretation, and was responsible for drafting of the manuscript. LD contributed to the design of the study, conducted qualitative analysis, contributed to data interpretation, and provided a critical review of the manuscript. MS was responsible for programming the mobile app and recommender system, contributed to data interpretation, and provided a critical review of the manuscript. LG and VVA contributed to data interpretation and provided a critical review of the manuscript. SK serves as the guarantor for the overall content of the manuscript. All authors have approved the publication of the final version of the manuscript.

Conflicts of Interest

LG, LD, and SK are associated with IDEWE, which owns the intellectual property of the Health Empower app.

XSL•FO
RenderX

Multimedia Appendix 1
Exploratory factor analysis of the Dutch Musculoskeletal Questionnaire.
[DOCX File , 28 KB - formative_v9i1e67886_app1.docx]

Multimedia Appendix 2
Baseline comparison between study completers and dropouts.
[DOCX File , 29 KB - formative_v9i1e67886_app2.docx]

Multimedia Appendix 3
Regression slopes of outcome variables.
[DOCX File , 192 KB - formative_v9i1e67886_app3.docx]

References

1. de Kok J, Vroonhof P, Snijders J, Roullis G, Clarke M, Peereboom K, et al. Work-related musculoskeletal disorders: prevalence, costs and demographics in the EU. European Agency for Safety and Health at Work. 2019. URL: https://osha.europa.eu/sites/default/files/Work_related_MSDs_prevalence_costs_and_demographics_in_EU_summary.pdf [accessed 2024-10-24]
2. Macchia L, Delaney L, Daly M. Global pain levels before and during the COVID-19 pandemic. Econ Hum Biol 2024 Jan;52:101337 [FREE Full text] [doi: 10.1016/j.ehb.2023.101337] [Medline: 38104357]
3. Cammarota A. The commissions initiative on WRMSDs second phase consultation of European social partners and future developments. Communication and Information Resource Centre for Administrations, Businesses and Citizens. 2007. URL: https://circabc.europa.eu/sd/a/5fc5910b-f393-4695-b594-35ad0a335d93/mines_070903_the_ec_initiative_on_wrmsds_cammarota.pdf [accessed 2024-10-24]
4. Statistieken van de uitkeringen over 2022. Rijksinstituut voor ziekte- en invaliditeitsverzekering. URL: https://www.riziv.fgov.be/nl/statistieken/statistieken-uitkeringen/statistieken-2022#invaliditeit-in-2022 [accessed 2024-10-24]
5. Gonzales CM. Council Directive 89/391/EEC of 12 June 1989 on the introduction of measures to encourage improvements in the safety and health of workers at work. The Council of the European Communities. URL: http://data.europa.eu/eli/dir/1989/391/oj [accessed 2024-10-24]
6. Colosio C, Mandic-Rajcevic S, Godderis L, van der Laan G, Hulshof C, van Dijk F. Workers' health surveillance: implementation of the Directive 89/391/EEC in Europe. Occup Med (Lond) 2017 Oct 01;67(7):574-578. [doi: 10.1093/occmed/kqx113] [Medline: 29016828]
7. Lambrechts MC, Ketterer F, Symons L, Mairiaux P, Peremans L, Remmen R, et al. The approach taken to substance abuse by occupational physicians: a qualitative study on influencing factors. J Occup Environ Med 2015 Nov;57(11):1228-1235 [FREE Full text] [doi: 10.1097/JOM.0000000000000549] [Medline: 26539772]
8. Walters D, Johnstone R, Bluff E, Jørgen Limborg H, Gensby U. Prevention services for occupational safety and health in the European Union: anachronisms or supports for better practice? Saf Sci 2022 Aug;152:105793. [doi: 10.1016/j.ssci.2022.105793]
9. Choinière M, Peng P, Gilron I, Buckley N, Williamson O, Janelle-Montcalm A, et al. Accessing care in multidisciplinary pain treatment facilities continues to be a challenge in Canada. Reg Anesth Pain Med 2020 Dec 06;45(12):943-948. [doi: 10.1136/rapm-2020-101935] [Medline: 33024007]
10. Benahmed N, Jonckheer P, Zeevaert R, Vos B, Kohn L. Assessing and monitoring waiting times in healthcare: how to proceed in Belgium? Belgian Health Care Knowledge Centre (KCE). 2024. URL: https://kce.fgov.be/en/assessing-and-monitoring-waiting-times-in-healthcare-how-to-proceed-in-belgium [accessed 2024-10-24]
11. Lynch ME, Campbell F, Clark AJ, Dunbar MJ, Goldstein D, Peng P, et al. A systematic review of the effect of waiting for treatment for chronic pain. Pain 2008 May;136(1-2):97-116. [doi: 10.1016/j.pain.2007.06.018] [Medline: 17707589]
12. Steinhubl SR, Muse ED, Topol EJ. The emerging field of mobile health. Sci Transl Med 2015 Apr 15;7(283):283rv3 [FREE Full text] [doi: 10.1126/scitranslmed.aaa3487] [Medline: 25877894]
13. Rowland SP, Fitzgerald JE, Holme T, Powell J, McGregor A. What is the clinical value of mHealth for patients? NPJ Digit Med 2020 Jan 13;3(1):4 [FREE Full text] [doi: 10.1038/s41746-019-0206-x] [Medline: 31970289]
14. Lalloo C, Jibb LA, Rivera J, Agarwal A, Stinson JN. "There's a pain app for that": review of patient-targeted smartphone applications for pain management. Clin J Pain 2015 Jun;31(6):557-563. [doi: 10.1097/AJP.0000000000000171] [Medline: 25370138]
15. Moreno-Ligero M, Moral-Munoz JA, Salazar A, Failde I. mHealth intervention for improving pain, quality of life, and functional disability in patients with chronic pain: systematic review. JMIR Mhealth Uhealth 2023 Feb 02;11:e40844 [FREE Full text] [doi: 10.2196/40844] [Medline: 36729570]
16. Sitges C, Terrasa JL, García-Dopico N, Segur-Ferrer J, Velasco-Roldán O, Crespí-Palmer J, et al. An educational and exercise mobile phone-based intervention to elicit electrophysiological changes and to improve psychological functioning

XSL•FO
RenderX

in adults with nonspecific chronic low back pain (BackFit App): nonrandomized clinical trial. JMIR Mhealth Uhealth 2022 Mar 15;10(3):e29171 [FREE Full text] [doi: 10.2196/29171] [Medline: 35289758]
17. Morcillo-Muñoz Y, Sánchez-Guarnido AJ, Calzón-Fernández S, Baena-Parejo I. Multimodal chronic pain therapy for adults via smartphone: randomized controlled clinical trial. J Med Internet Res 2022 May 11;24(5):e36114 [FREE Full text] [doi: 10.2196/36114] [Medline: 35373776]
18. Mascaro JS, Singh V, Wehrmeyer K, Scott B, Juan J, McKenzie-Brown AM, et al. Randomized, wait-list-controlled pilot study of app-delivered mindfulness for patients reporting chronic pain. Pain Rep 2021;6(1):e924 [FREE Full text] [doi: 10.1097/PR9.0000000000000924] [Medline: 34778688]
19. Grant M, O-Beirne-Elliman J, Froud R, Underwood M, Seers K. The work of return to work. Challenges of returning to work when you have chronic pain: a meta-ethnography. BMJ Open 2019 Jun 20;9(6):e025743 [FREE Full text] [doi: 10.1136/bmjopen-2018-025743] [Medline: 31227529]
20. Vlaeyen JW, Linton SJ. Fear-avoidance and its consequences in chronic musculoskeletal pain: a state of the art. Pain 2000 Apr;85(3):317-332 [FREE Full text] [doi: 10.1016/S0304-3959(99)00242-0] [Medline: 10781906]
21. Picavet HS, Vlaeyen JW, Schouten JS. Pain catastrophizing and kinesiophobia: predictors of chronic low back pain. Am J Epidemiol 2002 Dec 01;156(11):1028-1034. [doi: 10.1093/aje/kwf136] [Medline: 12446259]
22. Carlesso LC, Raja Rampersaud Y, Davis AM. Clinical classes of injured workers with chronic low back pain: a latent class analysis with relationship to working status. Eur Spine J 2018 Jan 31;27(1):117-124. [doi: 10.1007/s00586-017-4966-1] [Medline: 28138782]
23. Macías-Toronjo I, Rojas-Ocaña MJ, Sánchez-Ramos JL, García-Navarro EB. Pain catastrophizing, kinesiophobia and fear-avoidance in non-specific work-related low-back pain as predictors of sickness absence. PLoS One 2020 Dec 10;15(12):e0242994 [FREE Full text] [doi: 10.1371/journal.pone.0242994] [Medline: 33301458]
24. Steenstra IA, Munhall C, Irvin E, Oranye N, Passmore S, Van Eerd D, et al. Systematic review of prognostic factors for return to work in workers with sub acute and chronic low back pain. J Occup Rehabil 2017 Sep 19;27(3):369-381 [FREE Full text] [doi: 10.1007/s10926-016-9666-x] [Medline: 27647141]
25. Keyaerts S, Godderis L, Delvaux E, Daenen L. The association between work-related physical and psychosocial factors and musculoskeletal disorders in healthcare workers: moderating role of fear of movement. J Occup Health 2022 Jan;64(1):e12314 [FREE Full text] [doi: 10.1002/1348-9585.12314] [Medline: 35043512]
26. Keyaerts S, Godderis L, Vanden Abeele V, Daenen L. Identifying pain profiles in employees including work-related factors and pain perceptions: a cross-sectional study in Belgian companies. BMJ Open 2024 Aug 12;14(8):e082804 [FREE Full text] [doi: 10.1136/bmjopen-2023-082804] [Medline: 39134443]
27. Cullen KL, Irvin E, Collie A, Clay F, Gensby U, Jennings PA, et al. Effectiveness of workplace interventions in return-to-work for musculoskeletal, pain-related and mental health conditions: an update of the evidence and messages for practitioners. J Occup Rehabil 2018 Mar 21;28(1):1-15 [FREE Full text] [doi: 10.1007/s10926-016-9690-x] [Medline: 28224415]
28. Duffy A, Christie GJ, Moreno S. The challenges toward real-world implementation of digital health design approaches: narrative review. JMIR Hum Factors 2022 Sep 09;9(3):e35693 [FREE Full text] [doi: 10.2196/35693] [Medline: 36083628]
29. Voorheis P, Zhao A, Kuluski K, Pham Q, Scott T, Sztur P, et al. Integrating behavioral science and design thinking to develop mobile health interventions: systematic scoping review. JMIR Mhealth Uhealth 2022 Mar 16;10(3):e35799 [FREE Full text] [doi: 10.2196/35799] [Medline: 35293871]
30. Kistin C, Silverstein M. Pilot studies: a critical but potentially misused component of interventional research. JAMA 2015 Oct 20;314(15):1561-1562 [FREE Full text] [doi: 10.1001/jama.2015.10962] [Medline: 26501530]
31. Nijs J, Paul van Wilgen C, Van Oosterwijck J, van Ittersum M, Meeus M. How to explain central sensitization to patients with 'unexplained' chronic musculoskeletal pain: practice guidelines. Man Ther 2011 Oct;16(5):413-418 [FREE Full text] [doi: 10.1016/j.math.2011.04.005] [Medline: 21632273]
32. Garland EL. Pain processing in the human nervous system: a selective review of nociceptive and biobehavioral pathways. Prim Care 2012 Sep;39(3):561-571 [FREE Full text] [doi: 10.1016/j.pop.2012.06.013] [Medline: 22958566]
33. Vitoula KV, Venneri A, Varrassi G, Paladini A, Sykioti P, Adewusi J, et al. Behavioral therapy approaches for the management of low back pain: an up-to-date systematic review. Pain Ther 2018 Jun 16;7(1):1-12 [FREE Full text] [doi: 10.1007/s40122-018-0099-4] [Medline: 29767395]
34. Meeus M, Nijs J, Van Wilgen P, Noten S, Goubert D, Huijnen I. Moving on to movement in patients with chronic joint pain. Pain 2016;24:1-8.
35. Stock SR, Nicolakakis N, Vézina N, Vézina M, Gilbert L, Turcot A, et al. Are work organization interventions effective in preventing or reducing work-related musculoskeletal disorders? A systematic review of the literature. Scand J Work Environ Health 2018 Mar 01;44(2):113-133 [FREE Full text] [doi: 10.5271/sjweh.3696] [Medline: 29188299]
36. Blickenstaff C, Pearson N. Reconciling movement and exercise with pain neuroscience education: a case for consistent education. Physiother Theory Pract 2016 Jul 29;32(5):396-407. [doi: 10.1080/09593985.2016.1194653] [Medline: 27356079]
37. Ryan CG, Gray HG, Newton M, Granat MH. Pain biology education and exercise classes compared to pain biology education alone for individuals with chronic low back pain: a pilot randomised controlled trial. Man Ther 2010 Aug;15(4):382-387. [doi: 10.1016/j.math.2010.03.003] [Medline: 20359937]

XSL•FO
RenderX

38. Robertson IT, Cooper CL, Sarkar M, Curran T. Resilience training in the workplace from 2003 to 2014: a systematic review. J Occup Organ Psychol 2015 Apr 25;88(3):533-562. [doi: 10.1111/joop.12120]
39. Joyce S, Shand F, Tighe J, Laurent SJ, Bryant RA, Harvey SB. Road to resilience: a systematic review and meta-analysis of resilience training programmes and interventions. BMJ Open 2018 Jun 14;8(6):e017858 [FREE Full text] [doi: 10.1136/bmjopen-2017-017858] [Medline: 29903782]
40. Nijs J, D'Hondt E, Clarys P, Deliens T, Polli A, Malfliet A, et al. Lifestyle and chronic pain across the lifespan: an inconvenient truth? PM R 2020 Apr 21;12(4):410-419. [doi: 10.1002/pmrj.12244] [Medline: 31437355]
41. Bull FC, Al-Ansari SS, Biddle S, Borodulin K, Buman MP, Cardon G, et al. World Health Organization 2020 guidelines on physical activity and sedentary behaviour. Br J Sports Med 2020 Dec 25;54(24):1451-1462 [FREE Full text] [doi: 10.1136/bjsports-2020-102955] [Medline: 33239350]
42. Sullivan MJ, Adams H, Rhodenizer T, Stanish WD. A psychosocial risk factor--targeted intervention for the prevention of chronic pain and disability following whiplash injury. Phys Ther 2006 Jan;86(1):8-18. [doi: 10.1093/ptj/86.1.8] [Medline: 16386058]
43. Szymanski M, Abeele VV, Verbert K. Explaining health recommendations to lay users: The dos and dont's. In: Proceedings of the ACM IUI Workshops 2022. 2022 Presented at: IUI '22; March 21-22, 2022; Helsinki, Finland p. 1-10 URL: https://ceur-ws.org/Vol-3124/paper1.pdf
44. Szymanski M, Conati C, Abeele VV, Verbert K. Designing and personalising hybrid multi-modal health explanations for lay users. In: Proceedings of the 10th Joint Workshop on Interfaces and Human Decision Making for Recommender. 2023 Presented at: IntRS '23; September 18, 2023; Singapore, Singapore p. 35-52 URL: https://tinyurl.com/djxw8zf5
45. Lindström I, Ohlund C, Eek C, Wallin L, Peterson LE, Fordyce WE, et al. The effect of graded activity on patients with subacute low back pain: a randomized prospective clinical study with an operant-conditioning behavioral approach. Phys Ther 1992 Apr;72(4):279-291. [doi: 10.1093/ptj/72.4.279] [Medline: 1533941]
46. Swann C, Jackman PC, Lawrence A, Hawkins RM, Goddard SG, Williamson O, et al. The (over)use of SMART goals for physical activity promotion: a narrative review and critique. Health Psychol Rev 2023 Jun;17(2):211-226 [FREE Full text] [doi: 10.1080/17437199.2021.2023608] [Medline: 35094640]
47. Hildebrandt VH, Bongers PM, van Dijk FJ, Kemper HC, Dul J. Dutch Musculoskeletal Questionnaire: description and basic qualities. Ergonomics 2001 Oct 10;44(12):1038-1055. [doi: 10.1080/00140130110087437] [Medline: 11780727]
48. Notelaers G, de Witte H, van Veldhoven M, Vermunt J. Construction and validation of the short inventory to monitor psychosocial hazards. Médecine Travail Ergonom 2007;44(1):11-18.
49. van Veldhoven MJ, Prins J, van der Laken PA, Dijkstra L. QEEW2.0: 42 Short Scales for Survey Research on Work, Well-Being and Performance. Amsterdam, the Netherlands: SKB; 2015.
50. Sullivan MJ, Bishop SR, Pivik J. The pain catastrophizing scale: development and validation. Psychol Assess 1995 Dec;7(4):524-532. [doi: 10.1037/1040-3590.7.4.524]
51. Waddell G, Newton M, Henderson I, Somerville D, Main C. A Fear-Avoidance Beliefs Questionnaire (FABQ) and the role of fear-avoidance beliefs in chronic low back pain and disability. Pain 1993 Feb;52(2):157-168. [doi: 10.1016/0304-3959(93)90127-B] [Medline: 8455963]
52. Delgado DA, Lambert BS, Boutris N, McCulloch PC, Robbins AB, Moreno MR, et al. Validation of digital visual analog scale pain scoring with a traditional paper-based visual analog scale in adults. JAAOS Glob Res Rev 2018;2(3):e088. [doi: 10.5435/jaaosglobal-d-17-00088]
53. Fuller D, Colwell E, Low J, Orychock K, Tobin MA, Simango B, et al. Reliability and validity of commercially available wearable devices for measuring steps, energy expenditure, and heart rate: systematic review. JMIR Mhealth Uhealth 2020 Sep 08;8(9):e18694 [FREE Full text] [doi: 10.2196/18694] [Medline: 32897239]
54. Tudor-Locke C. Steps to better cardiovascular health: how many steps does it take to achieve good health and how confident are we in this number? Curr Cardiovasc Risk Rep 2010 Jul;4(4):271-276 [FREE Full text] [doi: 10.1007/s12170-010-0109-5] [Medline: 20672110]
55. Braun V, Clarke V. Successful Qualitative Research: A Practical Guide for Beginners. Thousand Oaks, CA: Sage Publication; 2016.
56. Reynolds TJ, Gutman J. Laddering theory, method, analysis, and interpretation. J Advert Res 1988;28:11-31 [FREE Full text]
57. Faul F, Erdfelder E, Lang A, Buchner A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods 2007 May;39(2):175-191. [doi: 10.3758/BF03193146]
58. Kristjánsdóttir Ó, Fors EA, Eide E, Finset A, Stensrud TL, van Dulmen S, et al. A smartphone-based intervention with diaries and therapist-feedback to reduce catastrophizing and increase functioning in women with chronic widespread pain: randomized controlled trial. J Med Internet Res 2013 Jan 07;15(1):e5 [FREE Full text] [doi: 10.2196/jmir.2249] [Medline: 23291270]
59. Meyerowitz-Katz G, Ravi S, Arnolda L, Feng X, Maberly G, Astell-Burt T. Rates of attrition and dropout in app-based interventions for chronic disease: systematic review and meta-analysis. J Med Internet Res 2020 Sep 29;22(9):e20283 [FREE Full text] [doi: 10.2196/20283] [Medline: 32990635]
60. Field A. Discovering Statistics Using IBM SPSS Statistics. 5th edition. Thousand Oaks, CA: Sage Publication; 2018.

XSL•FO
RenderX

61. Yao J, Tan CS, Lim N, Tan J, Chen C, Müller-Riemenschneider F. Number of daily measurements needed to estimate habitual step count levels using wrist-worn trackers and smartphones in 212,048 adults. Sci Rep 2021 May 05;11(1):9633 [FREE Full text] [doi: 10.1038/s41598-021-89141-3] [Medline: 33953288]
62. Tang LM, Meyer J, Epstein DA, Bragg K, Engelen L, Bauman A, et al. Defining adherence: making sense of physical activity tracker data. Proc ACM Interact Mob Wearable Ubiquitous Technol 2018 Mar 26;2(1):1-22. [doi: 10.1145/3191769]
63. Dierckx de Casterlé B, Gastmans C, Bryon E, Denier Y. QUAGOL: a guide for qualitative data analysis. Int J Nurs Stud 2012 Mar;49(3):360-371 [FREE Full text] [doi: 10.1016/j.ijnurstu.2011.09.012] [Medline: 21996649]
64. Guetterman TC, Fetters MD, Creswell JW. Integrating quantitative and qualitative results in health science mixed methods research through joint displays. Ann Fam Med 2015 Nov 09;13(6):554-561 [FREE Full text] [doi: 10.1370/afm.1865] [Medline: 26553895]
65. Watson JA, Ryan CG, Cooper L, Ellington D, Whittle R, Lavender M, et al. Pain neuroscience education for adults with chronic musculoskeletal pain: a mixed-methods systematic review and meta-analysis. J Pain 2019 Oct;20(10):1140.e1-1140.22 [FREE Full text] [doi: 10.1016/j.jpain.2019.02.011] [Medline: 30831273]
66. Mönninghoff A, Kramer JN, Hess AJ, Ismailova K, Teepe GW, Tudor Car L, et al. Long-term effectiveness of mHealth physical activity interventions: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2021 Apr 30;23(4):e26699 [FREE Full text] [doi: 10.2196/26699] [Medline: 33811021]
67. Malfliet A, Ickmans K, Huysmans E, Coppieters I, Willaert W, Bogaert WV, et al. Best evidence rehabilitation for chronic pain part 3: low back pain. J Clin Med 2019 Jul 19;8(7):1063 [FREE Full text] [doi: 10.3390/jcm8071063] [Medline: 31331087]
68. Siddall B, Ram A, Jones MD, Booth J, Perriman D, Summers SJ. Short-term impact of combining pain neuroscience education with exercise for chronic musculoskeletal pain: a systematic review and meta-analysis. Pain 2022 Jan 01;163(1):e20-e30. [doi: 10.1097/j.pain.0000000000002308] [Medline: 33863860]
69. Driessen MT, Groenewoud K, Proper KI, Anema JR, Bongers PM, van der Beek AJ. What are possible barriers and facilitators to implementation of a Participatory Ergonomics programme? Implement Sci 2010 Aug 24;5(1):64 [FREE Full text] [doi: 10.1186/1748-5908-5-64] [Medline: 20735822]
70. Rothmore P, Aylward P, Karnon J. The implementation of ergonomics advice and the stage of change approach. Appl Ergon 2015 Nov;51:370-376. [doi: 10.1016/j.apergo.2015.06.013] [Medline: 26154235]
71. Sundstrup E, Seeberg KG, Bengtsen E, Andersen LL. A systematic review of workplace interventions to rehabilitate musculoskeletal disorders among employees with physical demanding work. J Occup Rehabil 2020 Dec 26;30(4):588-612 [FREE Full text] [doi: 10.1007/s10926-020-09879-x] [Medline: 32219688]
72. Xu L, Shi H, Shen M, Ni Y, Zhang X, Pang Y, et al. The effects of mHealth-based gamification interventions on participation in physical activity: systematic review. JMIR Mhealth Uhealth 2022 Feb 03;10(2):e27794 [FREE Full text] [doi: 10.2196/27794] [Medline: 35113034]
73. Tu R, Hsieh P, Feng W. Walking for fun or for “likes”? The impacts of different gamification orientations of fitness apps on consumers’ physical activities. Sport Manag Rev 2021 Feb 03;22(5):682-693. [doi: 10.1016/j.smr.2018.10.005]
74. Geuens J, Geurts L, Swinnen TW, Westhovens R, Vanden Abeele V. Mobile health features supporting self-management behavior in patients with chronic arthritis: mixed-methods approach on patient preferences. JMIR Mhealth Uhealth 2019 Mar 25;7(3):e12535 [FREE Full text] [doi: 10.2196/12535] [Medline: 30907737]
75. Eysenbach G. The law of attrition. J Med Internet Res 2005 Mar 31;7(1):e11 [FREE Full text] [doi: 10.2196/jmir.7.1.e11] [Medline: 15829473]
76. Balakrishnan K, Arjmand EM. The impact of cognitive and implicit bias on patient safety and quality. Otolaryngol Clin North Am 2019 Feb;52(1):35-46. [doi: 10.1016/j.otc.2018.08.016] [Medline: 30245040]
77. Anderson CA, Lepper MR, Ross L. Perseverance of social theories: the role of explanation in the persistence of discredited information. J Pers Soc Psychol 1980 Dec;39(6):1037-1049. [doi: 10.1037/H0077720]
78. Kunitomo K, Harada T, Watari T. Cognitive biases encountered by physicians in the emergency room. BMC Emerg Med 2022 Aug 26;22(1):148 [FREE Full text] [doi: 10.1186/s12873-022-00708-3] [Medline: 36028810]
79. Jaini PA, Lee JS. A review of 21st century utility of a biopsychosocial model in United States medical school education. J Lifestyle Med 2015 Sep 30;5(2):49-59 [FREE Full text] [doi: 10.15280/jlm.2015.5.2.49] [Medline: 26770891]
80. Setchell J, Costa N, Ferreira M, Makovey J, Nielsen M, Hodges PW. Individuals' explanations for their persistent or recurrent low back pain: a cross-sectional survey. BMC Musculoskelet Disord 2017 Nov 17;18(1):466 [FREE Full text] [doi: 10.1186/s12891-017-1831-7] [Medline: 29149847]
81. van Dijk H, Köke AJ, Elbers S, Mollema J, Smeets RJ, Wittink H. Physiotherapists using the biopsychosocial model for chronic pain: barriers and facilitators-a scoping review. Int J Environ Res Public Health 2023 Jan 16;20(2):1634 [FREE Full text] [doi: 10.3390/ijerph20021634] [Medline: 36674387]
82. Siebert J, Siebert JU. Effective mitigation of the belief perseverance bias after the retraction of misinformation: awareness training and counter-speech. PLoS One 2023 Mar 8;18(3):e0282202 [FREE Full text] [doi: 10.1371/journal.pone.0282202] [Medline: 36888583]
83. Kountz DS. Strategies for improving low health literacy. Postgrad Med 2009 Sep;121(5):171-177. [doi: 10.3810/pgm.2009.09.2065] [Medline: 19820287]

XSL•FO
RenderX

84. Prochaska JO, Velicer WF. The transtheoretical model of health behavior change. Am J Health Promot 1997 Sep;12(1):38-48. [doi: 10.4278/0890-1171-12.1.38] [Medline: 10170434]
85. Yarker J, Lewis R, Sinclair A, Michlig G, Munir F. Meta-synthesis of qualitative research on the barriers and facilitators to implementing workplace mental health interventions. SSM Ment Health 2022 Dec;2:100148. [doi: 10.1016/j.ssmmh.2022.100148]
86. Coole C, Watson PJ, Drummond A. Low back pain patients' experiences of work modifications; a qualitative study. BMC Musculoskelet Disord 2010 Dec 06;11(1):277 [FREE Full text] [doi: 10.1186/1471-2474-11-277] [Medline: 21134248]
87. Nijs J, Lluch Girbés E, Lundberg M, Malfliet A, Sterling M. Exercise therapy for chronic musculoskeletal pain: innovation by altering pain memories. Man Ther 2015 Feb;20(1):216-220. [doi: 10.1016/j.math.2014.07.004] [Medline: 25090974]
88. Bunzli S, Smith A, Schütze R, Lin I, O'Sullivan P. Making sense of low back pain and pain-related fear. J Orthop Sports Phys Ther 2017 Sep;47(9):628-636. [doi: 10.2519/jospt.2017.7434] [Medline: 28704621]
89. Avila FR, McLeod CJ, Huayllani MT, Boczar D, Giardi D, Bruce CJ, et al. Wearable electronic devices for chronic pain intensity assessment: a systematic review. Pain Pract 2021 Nov 26;21(8):955-965. [doi: 10.1111/papr.13047] [Medline: 34080306]
90. Van Daele T, Gerard S, Belmont A. Digital interventions and apps for mental health. Superior Health Council. 2024. URL: https://tinyurl.com/296ezu8k [accessed 2024-10-24]
91. Pillastrini P, Mugnai R, Farneti C, Bertozzi L, Bonfiglioli R, Curti S, et al. Evaluation of two preventive interventions for reducing musculoskeletal complaints in operators of video display terminals. Phys Ther 2007 May;87(5):536-544 [FREE Full text] [doi: 10.2522/ptj.20060092] [Medline: 17405805]
92. Nijs J, Wijma AJ, Willaert W, Huysmans E, Mintken P, Smeets R, et al. Integrating motivational interviewing in pain neuroscience education for people with chronic pain: a practical guide for clinicians. Phys Ther 2020 May 18;100(5):846-859. [doi: 10.1093/ptj/pzaa021] [Medline: 31995191]
93. Malfliet A, Kregel J, Meeus M, Roussel N, Danneels L, Cagnie B, et al. Blended-learning pain neuroscience education for people with chronic spinal pain: randomized controlled multicenter trial. Phys Ther 2018 May 01;98(5):357-368. [doi: 10.1093/ptj/pzx092] [Medline: 29669079]
94. Toonders SA, van Westrienen PE, Konings S, Nieboer ME, Veenhof C, Pisters MF. Patients' perspectives on the usability of a blended approach to an integrated intervention for patients with medically unexplained physical symptoms: mixed methods study. J Med Internet Res 2021 Sep 28;23(9):e19794 [FREE Full text] [doi: 10.2196/19794] [Medline: 34581674]
95. Fassier JB, Durand MJ, Caillard JF, Roquelaure Y, Loisel P. Results of a feasibility study: barriers and facilitators in implementing the Sherbrooke model in France. Scand J Work Environ Health 2015 May 01;41(3):223-233 [FREE Full text] [doi: 10.5271/sjweh.3489] [Medline: 25743881]
96. Main CJ, Shaw WS, Nicholas MK, Linton SJ. System-level efforts to address pain-related workplace challenges. Pain 2022 Aug 01;163(8):1425-1431 [FREE Full text] [doi: 10.1097/j.pain.0000000000002548] [Medline: 35195550]
97. Aschbrenner KA, Kruse G, Gallo JJ, Plano Clark VL. Applying mixed methods to pilot feasibility studies to inform intervention trials. Pilot Feasibility Stud 2022 Sep 26;8(1):217 [FREE Full text] [doi: 10.1186/s40814-022-01178-x] [Medline: 36163045]
98. Tewerkstelling per sector. Statistiek Vlaanderen. URL: https://www.vlaanderen.be/statistiek-vlaanderen/arbeid/tewerkstelling-per-sector [accessed 2024-10-24]
99. Svendsen MT, Bak CK, Sørensen K, Pelikan J, Riddersholm SJ, Skals RK, et al. Associations of health literacy with socioeconomic position, health risk behavior, and health status: a large national population-based survey among Danish adults. BMC Public Health 2020 Apr 28;20(1):565 [FREE Full text] [doi: 10.1186/s12889-020-08498-8] [Medline: 32345275]
100. Arias López MD, Ong BA, Borrat Frigola X, Fernández AL, Hicklent RS, Obeles AJ, et al. Digital literacy as a new determinant of health: a scoping review. PLOS Digit Health 2023 Oct;2(10):e0000279 [FREE Full text] [doi: 10.1371/journal.pdig.0000279] [Medline: 37824584]
101. Sylvia LG, Bernstein EE, Hubbard JL, Keating L, Anderson EJ. Practical guide to measuring physical activity. J Acad Nutr Diet 2014 Feb;114(2):199-208 [FREE Full text] [doi: 10.1016/j.jand.2013.09.018] [Medline: 24290836]

Abbreviations

FABQA: Fear-Avoidance Beliefs Questionnaire (physical activity subscale)
FABQW: Fear-Avoidance Beliefs Questionnaire (work subscale)
mHealth: mobile health
PCS: Pain Catastrophizing Scale
VAS: visual analog scale

XSL•FO
RenderX

Edited by J Sarvestan; submitted 24.10.24; peer-reviewed by KL Mauco, H Hah; comments to author 07.03.25; revised version received 01.04.25; accepted 21.05.25; published 27.06.25.

Please cite as:
Keyaerts S, Szymanski M, Godderis L, Vanden Abeele V, Daenen L
Evaluating the Impact on Pain Perceptions, Pain Intensity, and Physical Activity of a Mobile App to Empower Employees With Musculoskeletal Pain: Mixed Methods Pilot Study
JMIR Form Res 2025;9:e67886
URL: https://formative.jmir.org/2025/1/e67886
doi:10.2196/67886
PMID:

©Stijn Keyaerts, Maxwell Szymanski, Lode Godderis, Vero Vanden Abeele, Liesbeth Daenen. Originally published in JMIR Formative Research (https://formative.jmir.org), 27.06.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Conversational Agent (PracticePal) to Support the Delivery of a Brief Behavioral Activation Treatment for Depression in Rural India: Development and Pilot-Testing Study

Ravindra Agrawal[1,2,3], MBBS, MPM, MRCPsych, DPM; Kimberley Monteiro[1], MA; Nityasri Sankha Narasimhamurti[1], MPH; Shreya Sharma[1], MA; Amruta Suryawanshi[1], BDS; Aman Bariya[1], MA; Shravani Narvekar[1], MA; Lilianna Bagnoli[4], BA; Mohit Saxena[5], BBA; Lauren Magoun[6], MPH; Shradha S Parsekar[1], PhD; Julia R Pozuelo[7], PhD; Neal Lesh[6], PhD; Mohit Sood[8], MPH; Tanushri Sharma[1], MA; Harshita Yadav[8], MA; Anant Bhan[8], MBBS; Abhijit Nadkarni[9,10], MBBS, MRCPsych, PhD, DPM; Vikram Patel[1,7,11], MBBS, MRCPsych, PhD

[1]Sangath, Panaji, India
[2]Department of Psychiatry, Manipal Hospital, Panaji, India
[3]Antarman Centre, Panaji, India
[4]Impact Lab, Dimagi India, New Delhi, India
[5]Dimagi India, New Delhi, India
[6]Dimagi United States, Cambridge, MA, United States
[7]Department of Global Health and Social Medicine, Harvard Medical School, Boston, MD, United States
[8]Sangath, Bhopal, India
[9]Department of Population Health, London School of Hygiene & Tropical Medicine, London, United Kingdom
[10]Addictions & Related Research Group, Sangath, Panaji, India
[11]Department of Global Health and Population, Harvard TH Chan School of Public Health, Boston, MD, United States

Corresponding Author:
Ravindra Agrawal, MBBS, MPM, MRCPsych, DPM
Sangath
Socorro, House No. 451 (168), Bhatkar Waddo
Porvorim
Panaji, 403501
India
Phone: 1 9322809431
Email: ravigoa@gmail.com

Abstract

Background: Brief psychosocial interventions, such as the Healthy Activity Program (HAP), which are based on behavioral activation and delivered by nonspecialist providers (NSPs), have emerged as cost-effective solutions for the treatment of depression. HAP treatment outcomes are improved by the engagement of patients in activation-focused homework assignments and their adherence to these assignments during therapy. Currently, patients are expected to complete these homework assignments using a paper workbook.

Objective: The aim of this study was to describe the user-centered development process of PracticePal, a chatbot designed to enhance patient engagement and homework adherence, and to evaluate its feasibility and acceptability as a therapy aid in India.

Methods: We used a user-centered approach to co-develop PracticePal, incorporating conversational flows and video scripts in Hindi. The chatbot was piloted with 30 participants having depression who were receiving the HAP from 15 nonspecialist counselors in primary care in rural Madhya Pradesh, India. The feasibility and acceptability of PracticePal were assessed through engagement data, in-depth interviews with a subset of 6 participants, and focus group discussions with 11 counselors. Treatment completion rates and changes in depressive symptoms were explored as secondary outcomes.

Results: Average patient engagement spanned 29 days (95% CI 24-34) during the 60-day treatment period. The engagement of patients with PracticePal increased as their treatment progressed, particularly after the third HAP session. Of the 30 patients, 20 (67%) accessed more than half of the multimedia content available on the chatbot. On average, there was a greater frequency

XSL•FO
RenderX

of self-initiated engagement (1558 out of a total of 1835 times, 84.9%) across all sessions compared with reminder prompts (277 out of 1835 times, 15.1%). All 30 patients completed treatment and experienced a reduction in the mean Patient Health Questionnaire-9 score from 13 (95% CI 12.6-13.6; signifying moderate severity) to 4 (95% CI 2.9-4.7; signifying none/minimal severity). Patients found the chatbot's reminders for activities, mood tracking, and video messages helpful and observed that it could help others in their social network. NSPs also reported improved participation of patients in the homework tasks compared with the paper workbook. A few patients faced challenges with low internet bandwidth, and those with limited literacy suggested increasing the amount of video content for easier accessibility.

Conclusions: The PracticePal chatbot is a feasible and acceptable therapy aid to complement a psychological treatment, with promising potential to enhance the effectiveness of NSP-delivered psychosocial interventions in low-resource settings. Future steps include conducting a fully powered randomized controlled trial to assess its effectiveness in improving mental health outcomes.

(JMIR Form Res 2025;9:e73563) doi:10.2196/73563

KEYWORDS

conversational agent; chatbot; digital tool; homework assignment; behavior activation; treatment adherence; depression; psychological treatment

Introduction

Depression is a common mental health issue, with symptoms, such as low mood, fatigue, loss of interest, and insomnia, as well as unexplained somatic symptoms. According to the World Health Organization, 5% of adults currently experience depression [1]. As of 2023, an estimated 280 million individuals worldwide had depression, including 56 million living in India (approximately 4.5% of the country's population) [2]. These numbers are compounded by a large treatment gap [3], especially in India [4], where 85% of people with depression do not receive an evidence-based intervention for their illness. This is largely due to the severe shortage of mental health professionals, as there are only 0.03 psychologists and 0.20 psychiatrists per 10,000 individuals [5]. Task sharing, which involves training nonspecialist providers (NSPs) to provide low-intensity mental health treatments, has been found to be effective and cost-effective for the treatment of depression [6-9].

The Healthy Activity Program (HAP) is one such evidence-based psychological intervention delivered by NSPs in routine primary care settings [10], and it is already being scaled up in partnership with health care systems [11]. The core active ingredient of HAP is behavioral activation, which is based on the principle that engaging in specific activities can help improve mood and counteract the patterns of avoidance, withdrawal, and inactivity that often accompany depression [12,13]. The therapy focuses on identifying and scheduling meaningful and enjoyable activities that align with an individual's values and interests, with the goal of increasing positive reinforcement from the environment. The HAP, which is delivered in 6 to 8 sessions, includes the assignment of homework to keep clients engaged between sessions and offers them opportunities to monitor their mood and to practice and ultimately master the skills learnt during treatment. Typically, homework assignments are completed using paper-based worksheets or pamphlets. However, in rural and low-resource settings, distributing and completing printed materials pose logistical and literacy challenges. In addition, these printed materials lack interactivity and tailored reminders, and require privacy, which may be difficult for individuals living in crowded family environments. Therefore, we aimed to explore whether the deployment of digital technologies might help address these challenges, after being inspired by research showing that such tools are effective in improving health outcomes in low-resource settings [14-16].

Conversational agents (CAs) are digital tools that interact with users through text or voice (eg, Siri and Alexa). CAs have the potential to increase the availability of mental health support and reduce its costs [17]. One type of CA is a rule-based chatbot, a software designed to simulate a human conversation through a scripted conversation workflow and text-based communication. Chatbots provide a promising solution for increasing patient engagement, offering personalized and interactive support, and filling gaps in traditional mental health care systems. With their ability to deliver tailored content and reminders, chatbots can enhance the effectiveness of interventions [7,14].

However, most existing CAs operate in English and have been primarily tested in high-income and resource-rich countries [10,17]. To address this gap, this study deployed user-centered and participatory action research methods to design "PracticePal," a rule-based chatbot intended to serve as a therapy aid in the delivery of the HAP by NSPs, such as accredited social health activists (ASHAs), who are on-ground community workers of India's National Health Mission. The chatbot was designed to support engagement between therapy sessions by prompting patients to complete their homework assignments and mood ratings, and reinforce therapeutic messages via engaging multimedia content. In this paper, we describe the development process of PracticePal and a pilot study evaluating its feasibility and acceptability among patients receiving a brief psychological treatment for depression in a rural, underserved setting of central India.

Methods

Ethical Considerations

This study was approved by the Institutional Review Board at Sangath (approval letter number: RA_2023_89) and the Indian Council of Medical Research. The approval provided by the institute's Institutional Review Board and by the Indian Council

XSL•FO
RenderX

of Medical Research is for study implementation, data collection, and data analysis. Informed consent was obtained from all study participants (ie, ASHAs and participants with depression) to be part of the study and to allow the collection of quantitative and qualitative data for further analysis. All participants were assured that their data would be deidentified and that their identities would be kept confidential. The patient participants were provided smartphones by the study team for the duration of the study for two main reasons: (1) to ensure that their engagement was not influenced by restricted access (as many participants had only 1 smartphone per household) and (2) to collect research data in anonymized form. Many of the participants who owned a smartphone also had the Telegram messaging app (Telegram LLC) already downloaded and registered in their own name. This would have meant that their identity would be visible at the time of collecting research data from the TextIt backend (a cloud-based platform that allows users to create and manage interactive messaging workflows). The study-provided phone ensured registration on the Telegram app with an anonymized participant ID. No monetary compensation was provided to the participants for taking part in the study, although they were provided with light refreshments, such as tea and snacks, following the interviews. The manuscript does not include any identifiable images of individual participants.

As this was an early-stage pilot study focused on assessing the feasibility and acceptability of a digital intervention in a real-world context, which had a small sample size and was nonrandomized, it was not prospectively registered.

Setting

The study was conducted in the Raisen district of Madhya Pradesh, India. Madhya Pradesh is one of the largest states in India, and is classified as one of its least developed states, with a predominantly rural population [18-21]. The lack of access to mental health services in the state has made it a key priority for the health system [22,23]. However, smartphone penetration is expanding at a rapid pace, with the state having added nearly 2.2 million new subscribers between May 2023 and February 2024 [24,25].

Sample

In the pilot study, the HAP was delivered by ASHAs as part of a community mental health program led by Sangath, in collaboration with the local government. ASHAs are female community health workers who serve as incentivized volunteers in the government health system. Typically, each ASHA serves a population of around 1000, primarily in rural areas, with key responsibilities focused on maternal and child health, and serves as a vital link between the community and the health system [26-28]. The ASHAs were trained in delivering the HAP by Sangath's trainers. Recruitment was conducted by the ASHAs at the residences of the study participants. The study participants were conveniently sampled during routine home visits by the ASHAs from eligible adult residents of the study sites. Eligible study participants were above 18 years of age, were fluent in Hindi, screened positive for depression on the Patient Health Questionnaire-9 (PHQ-9; score of ≥10), and provided written informed consent to be a study participant. We aimed to achieve maximum variability in gender, age, and education level in our sample. Additionally, a subsample of ASHAs who provided the counseling participated in 2 focus group discussions (FGDs) to share their perspectives on using the chatbot as a therapy aid.

A wider range of stakeholders was recruited during the development phase, including other ASHAs, a range of HAP practitioners, and members of the community. The objective of involving multiple stakeholders was to gather diverse perspectives, triangulate the information collected from these diverse stakeholders, and ensure that the chatbot was culturally appropriate and acceptable within the community.

Study Design

We developed PracticePal in partnership with Dimagi, a software social enterprise specializing in scalable digital solutions for frontline workers in low-resource settings. We used an iterative approach to design and develop the PracticePal chatbot, guided by the ADDIE (Analysis, Design, Development, Implementation, and Evaluation) framework [29] and by previous digital intervention development projects in mental health [30-32]. User-centered development was carried out in 5 phases: analysis, design, development, implementation, and evaluation. The phases of development are outlined in Figure 1 and described in the sections below, with each phase building upon the findings of the previous one.

XSL•FO
RenderX

Figure 1. Steps in the development of the PracticePal chatbot.

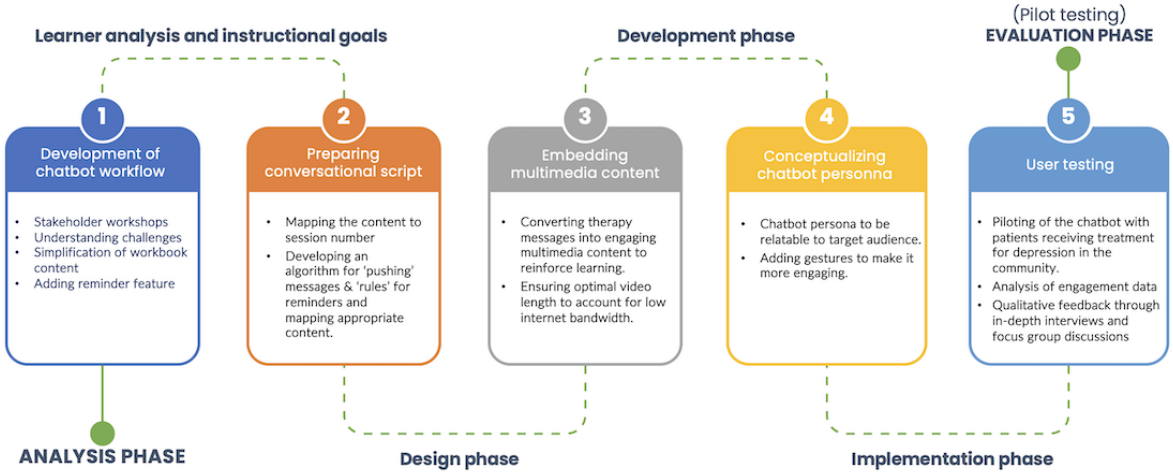

Analysis

The objective of this phase was to identify the challenges with the paper workbook, assess the feasibility of the chatbot, and define its key features. We conducted 3 stakeholder workshops, involving HAP practitioners (n=5), ASHAs (n=17), and chatbot developers (n=3). During the workshops, we identified several key challenges associated with using the paper workbook, including the need for careful storage, its lack of privacy, and its tendency to be lost or damaged. For patients with depression, completing the workbook felt burdensome due to low energy and motivation. To simplify the content, the workbook had to be less detailed, limiting the depth of therapy messages. Importantly, there were no reminders to do the homework tasks. Providers also expressed a need for real-time updates on their patients' homework progress and mood ratings, and a way to reinforce therapy messages between sessions. These findings shaped the design of the chatbot, focusing on simplifying patient engagement, enhancing privacy, and integrating reminders and multimedia content to support therapy.

Design

The objective of this phase was to use the findings from the analysis phase to design the chatbot features and components, with a focus on developing a conversational flow that aligned with the therapeutic progression of HAP sessions. Following the first therapy session, users gained access to psychoeducational content related to that session's topics. After the second session, users could access all chatbot features, including activity scheduling, rescheduling, and additional multimedia content. The multimedia content was designed to reinforce the core messages of the HAP, and it was designed in consultation with stakeholders. The videos were scripted and recorded in Hindi, were deliberately kept brief (under 5 minutes) based on suggestions from experienced HAP practitioners, and were kept accessible multiple times (a detailed description of each video can be found in Multimedia Appendix 1). Since the HAP is designed to be delivered by NSPs (ie, ASHAs in this study context), it was essential for the chatbot to maintain fidelity to the HAP counseling manual. This required the chatbot's conversational messages, the multimedia content, and the features, such as scheduling activities, mood ratings, and homework tracking, to be carefully structured to mirror the patients' stages in therapy. All patient participants received the full course of psychological treatment, consisting of 6 sessions. At any given point during the study period, participants could be at different stages of therapy, depending on their individual scheduling preferences. However, all participants ultimately completed all 6 sessions. Participants were encouraged to enter their session number while engaging with the chatbot during the intersession period, which allowed the chatbot to provide access to content and features relevant to that session (managed via a backend algorithm, as explained in Multimedia Appendix 2) and helped the study team to map the frequency and duration of chatbot engagement to the corresponding therapy sessions when collecting data from the backend. Another key aspect was to design an avatar that would be relatable to users, be more engaging, and improve trustworthiness through humanization [33]. Accordingly, the chatbot's avatar was called *Mann-mitra* (which, in Hindi, means a friend of the mind or mood). NSPs and members of the community gave suggestions on its characteristics, recommending that the avatar look like a married woman dressed in a saree, like ASHAs. To increase user engagement, the avatar was also given various gestures and expressions depending on the status of task completion (see Multimedia Appendix 3).

During this phase, we also determined other key components of the chatbot, such as its overall esthetics, its technical specifications, and the frequency of reminders sent to participants. All participants received activity-specific reminders at the beginning of the day (ie, 10 AM) on which the activity was scheduled. Based on feedback received during stakeholder meetings, additional reminders were sent at 5 PM, 7 PM, and 9 PM.

XSL•FO
RenderX

Development

Over a period of 8 months, from October 2022 to May 2023, we added features allowing clients to schedule their activities, set reminders, log activity (homework) completion, record their mood, and access multimedia content. The therapy messages were packaged into engaging multimedia content, while, at the same time, we were mindful to keep the length of the videos between 2 and 5 minutes to ensure optimum attention span and reduce file size. We adopted the principles of the agile project management approach [34] by developing a session-wise workflow and adding content incrementally, with features added based on the feedback received from stakeholders. Multiple versions of the chatbot were tested to identify errors in the algorithm, and the nodes were pruned to enable triggering of an appropriate "decision point."

Screengrabs of the chatbot interface are provided in Multimedia Appendix 4. PracticePal was developed on the TextIt platform and launched on Telegram, a messaging platform. We used Telegram as a channel for this chatbot due to the lower cost associated with it (compared with other messaging platforms). After development, the chatbot underwent 2 months of rigorous testing by the project team to identify and resolve bugs.

The PracticePal chatbot was piloted in a single-arm, nonrandomized study in August 2023 in the Raisen district of Madhya Pradesh. Trained ASHAs, who were already delivering the HAP to patients with depression, introduced the chatbot during the first counseling session, guiding patients on how to use its features. All subsequent counseling sessions were conducted at the residences of the study participants. ASHAs received weekly summaries on each patient's engagement, homework completion, and mood ratings, which helped them prepare for subsequent sessions. They also addressed any technical or content-related issues during the therapy, ensuring patients fully benefited from the digital tool. This summary was shared either as a printout or as a PDF sent to their smartphones. The process of integrating the chatbot during the therapy has been summarized in Figure 2.

Figure 2. Phase-wise outline describing the PracticePal chatbot features for patients. S: counseling session.

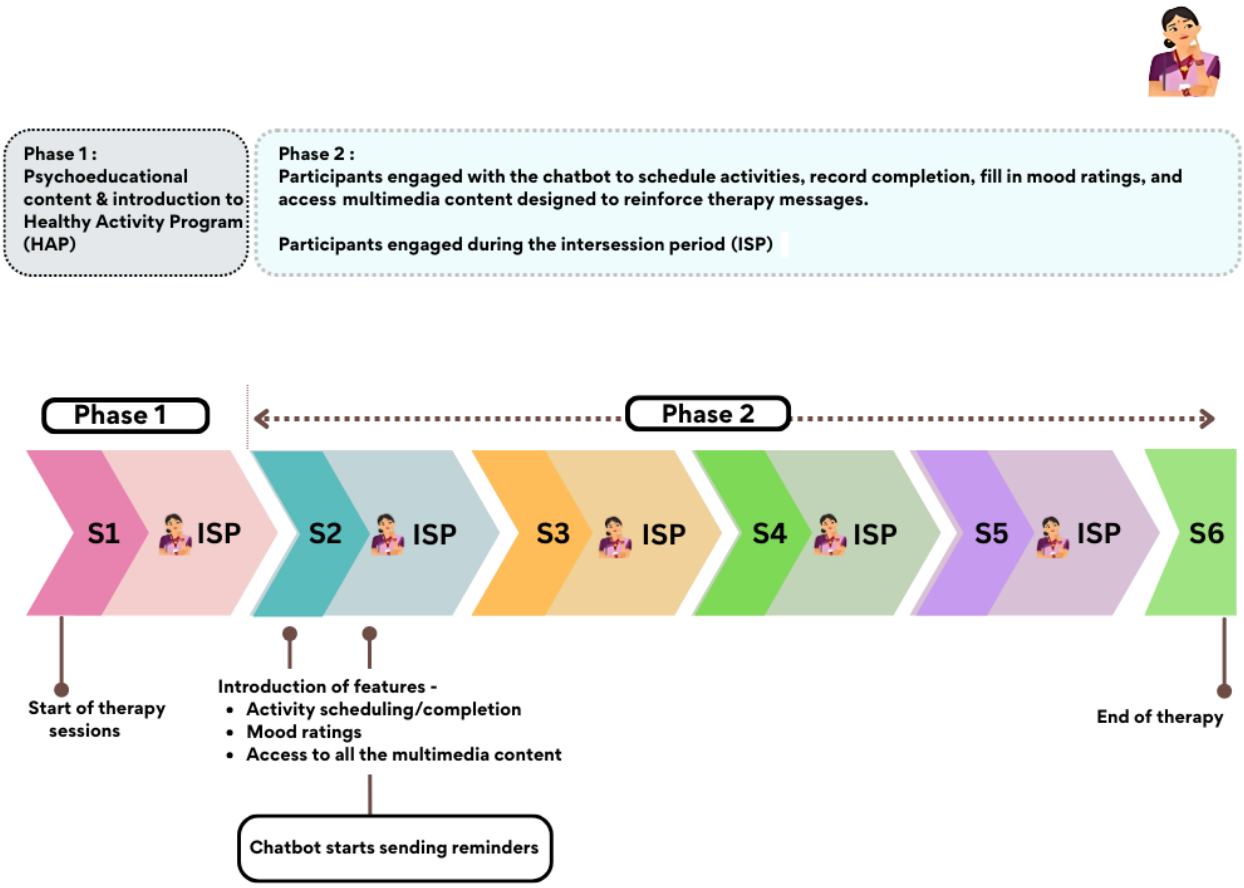

Evaluation

To evaluate the usability, feasibility, and acceptability of PracticePal, we collected both quantitative and qualitative data. Quantitative data on engagement were extracted from the backend of the chatbot platform, which tracked the frequency of access, total time spent engaging with the chatbot (including watching the multimedia content), and the depth of engagement (measured in terms of completion of homework within the workflow). The patient baseline PHQ-9 and endline PHQ-9 assessments were conducted by an independent researcher. Endline assessments were carried out through a phone survey. We also carried out purposive sampling to select patients and ASHAs, with whom we conducted postintervention in-depth

XSL•FO
RenderX

interviews (IDIs) and FGDs, respectively, to gather qualitative data on user experience, perceptions of the chatbot, and suggestions for improvement. The interview guide was designed to achieve the study objectives related to feasibility (the extent to which this digital innovation could be successfully used or implemented within a given setting) and acceptability (perception among stakeholders that the use of the chatbot is agreeable, palatable, or satisfactory) [35]. Patients were asked about their experiences using the chatbot, including ease of engagement, scheduling homework tasks, completing mood ratings, helpfulness of reminders, and understanding the therapy messages. ASHAs were asked about their experiences delivering therapy to patients who used the chatbot and their perceptions of its impact on homework completion rates. The IDIs and FGDs were conducted by Kavita Mandhare, an independent female researcher with a master's degree in clinical psychology and more than 4 years of experience in qualitative research methods. She was assisted by the project staff in identifying the participants and ensuring their availability for the IDIs or FGDs. First contact with the participants was over a phone call, following which the IDIs and FGDs were conducted in person in the local vernacular language (Hindi). Six IDIs were conducted with patients aged 21-43 years (5 female patients and 1 male patient; 5 married and 1 unmarried) either at each patient's house or at the house of a trusted friend. Four of these patients were homemakers, and 1 was a student. The education level ranged from 10th grade to graduate. Two FGDs (FGD1: n=6; FGD2: n=5) were conducted at the nearest health centers. In both scenarios, every effort was made to ensure the privacy of the participants, and they were informed that the researcher was independent from the intervention delivery team and was collecting feedback for improvement purposes. No participant refused to participate or backed out. All interviews were audio-recorded, lasted between 45 and 60 minutes, and were translated and transcribed into English. Field notes were taken during and after the interviews, and they guided the analyses.

Analysis

The quantitative data of all study participants were extracted by 2 study team members (AS and KM) and analyzed using descriptive analysis. The evaluation focused on several key metrics, including (1) the sociodemographic profile of the participants; (2) engagement with the chatbot, through the following key metrics: (a) frequency of interactions, (b) duration of interactions, (c) task completion (eg, scheduling activities, homework completion, and submitting mood ratings), (d) viewing of multimedia content, (e) user retention, and (f) qualitative feedback; and (3) patient outcomes as measured through treatment completion rates and changes in depressive symptoms (measured by the PHQ-9 score). We conducted an independent *t* test to assess for significant changes in PHQ-9 scores and multiple regression analysis to examine the relationship of gender and age with chatbot engagement. Qualitative data were analyzed using the thematic analysis approach with the NVivo software (version 14; Lumivero). A deductive approach was employed to explore key aspects of acceptability and feasibility based on participants' experiences and reports. The researchers first created initial codes from the transcripts, which were then grouped into themes and subthemes according to the original research questions and the interview guide. The reliability of the coding system was tested by carrying out blinded, double coding of the first 2 transcripts by SS and KM to reach a consensus on the preliminary codes and themes. While the transcripts were not returned to participants for comment or correction, the analysis team held regular peer discussions and reviewed the codes and themes in an ongoing way. Formal participant validation of the findings was not done, but care was taken to include different and contrasting views, such as minority or dissenting opinions, in the final analysis. The meta categories for coding were (1) acceptability of the chatbot, (2) feasibility of the chatbot as a therapy aid, and (3) challenges. The researchers developed a comprehensive codebook that was reviewed and finalized by RA and NSN. SS and KM subsequently coded the remaining transcripts and generated higher-order themes and subthemes. Finally, a narrative interpretation of the themes was developed, and illustrative quotes were selected. Neither of the coders was directly involved with the intervention delivery at the study site.

Results

Participant Characteristics

Our study sample included 30 patient participants. The majority of the participants (23/30, 77%) were married, and 57% (17/30) were women. Most participants (23/30, 77%) were between 18 and 35 years of age (median 27.4 years). The educational backgrounds of these 30 participants showed considerable variation. Of the 30 participants, 7 (23%) had completed primary education (up to class 4), 15 (50%) had studied up to high school (up to class 10), 3 (10%) had completed higher secondary education (up to class 12), 3 (10%) were graduates, and 2 (7%) held postgraduate degrees. All participants were familiar with the use of a smartphone and were able to operate the study-provided mobile phone.

Participant Engagement

All 30 participants interacted with the chatbot between sessions 1 and 6 over the 60-day treatment period, exhibiting varying levels of engagement.

Frequency and Duration of Use

We found that a participant engaged on average 53 times (95% CI 38.44-67.69) with the chatbot and spent an average of 233.5 minutes (ie, 3 hours and 53.5 minutes; 95% CI 119.74-347.33) engaging with the chatbot during the entire 60-day treatment period. In between any 2 therapy sessions, the average frequency of engagement was 10.61 times (95% CI 7.69-13.54), and the time spent was 46.7 minutes (95% CI 23.95-69.47). Figure 3 graphically represents the engagement frequency and duration. Analysis of chatbot engagement revealed a robust interaction immediately following session 1, when the chatbot was introduced. To account for varying participant engagement, median values were used. Between sessions 1 and 2, participants engaged with the chatbot on average 7.5 times for 28 minutes. As expected, engagement increased after session 2, with the introduction of features like activity scheduling and mood ratings. However, engagement gradually decreased to a median of 4.5 times and 16.5 minutes between sessions 5 and 6. Notably,

XSL•FO
RenderX

all 30 (100%) participants interacted with the chatbot during sessions 1-5, with a slight decrease to 93% (28/30) after session 5. During the intersession periods, a high proportion of participants (29/30, 97%) engaged with the chatbot for over 10 minutes, with more than half (18/30, 60%) interacting for 20 minutes or longer.

Figure 3. Overall frequency and duration of engagement during therapy with the chatbot.

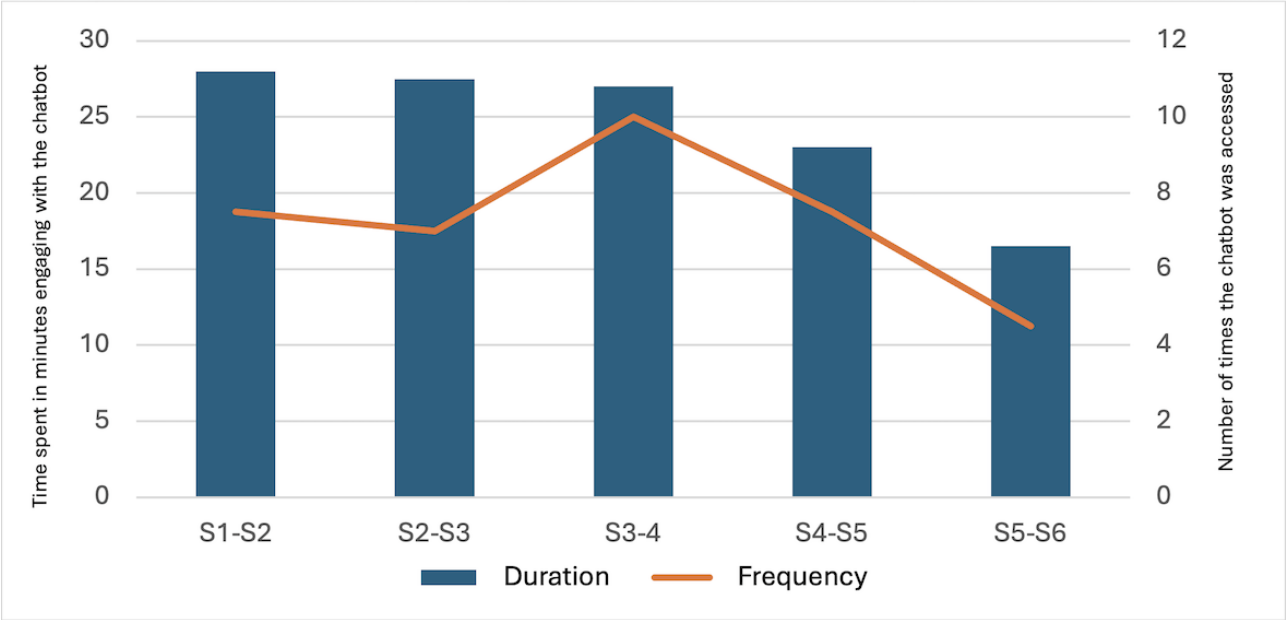

Participant Use of the Chatbot's Features

A key function of the chatbot was to encourage participants to schedule and engage in behavior activation tasks (homework activity) during the intersession period. Participants received reminders to schedule and complete the homework, record mood ratings, and access multimedia content. These reminders were sent at 10 AM, 5 PM, 7 PM, and 9 PM on the day an activity had been scheduled, irrespective of whether the individual had been active on the chatbot that day or not. All participants scheduled at least one homework activity and recorded mood ratings across sessions. Nearly two-thirds (22/30, 73%) of participants accessed more than half of the multimedia content presented by the chatbot. Figures 4A, 4B, and 4C depict the average engagement of the user with the chatbot's features of scheduling homework activities, reporting mood, and accessing multimedia content during the intersession period, respectively.

Figure 4. Participant engagement with PracticePal's features: (A) activity scheduling feature; (B) mood ratings; and (C) accessing multimedia. S: counseling session.

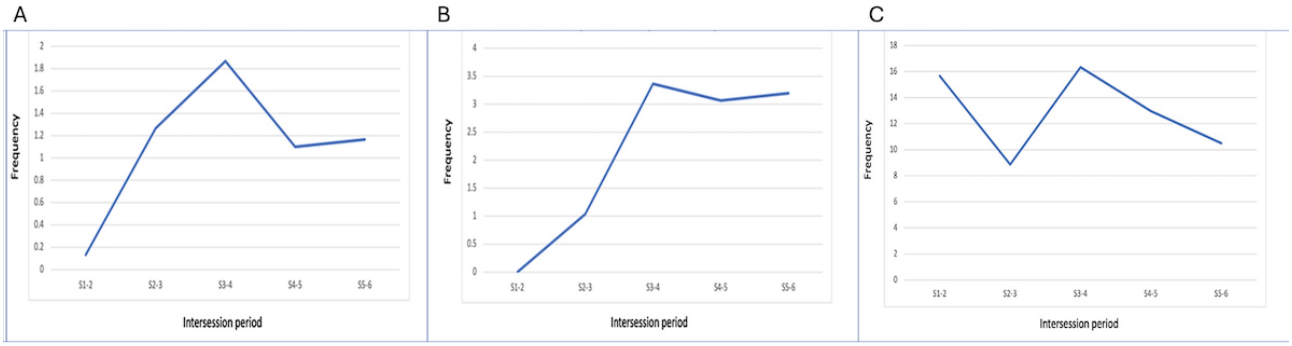

While the multimedia content was accessed at a high rate throughout the therapy duration, the use of features, such as activity scheduling and reporting of mood, picked up as the therapy progressed. This trend may reflect delayed understanding and orientation to the chatbot's features and was also reported during the IDIs.

Self-Initiated Versus Prompted Engagement

Participants engaged with the chatbot either through self-initiated interaction (unprompted) or in response to chatbot reminders (prompted). The chatbot reminders started only after session 2 and were restricted to 4 per day. We defined self-initiated engagement as the use of the chatbot by the patient either without any reminder or more than 1 hour after a reminder had been sent. Engagement within 1 hour of receiving a reminder was considered prompted. Self-initiated access (1558 out of a total of 1835 times, 84.9%) was substantially higher than prompted access (277 out of 1835 times, 15.1%) across all participants, indicating that most participants engaged with the chatbot voluntarily, beyond reminders (Figure 5).

XSL•FO
RenderX

Figure 5. Percentage of overall self-initiated (unprompted by chatbot) and chatbot-initiated (prompted by chatbot) engagements with the PracticePal chatbot between each therapy session. S: counseling session.

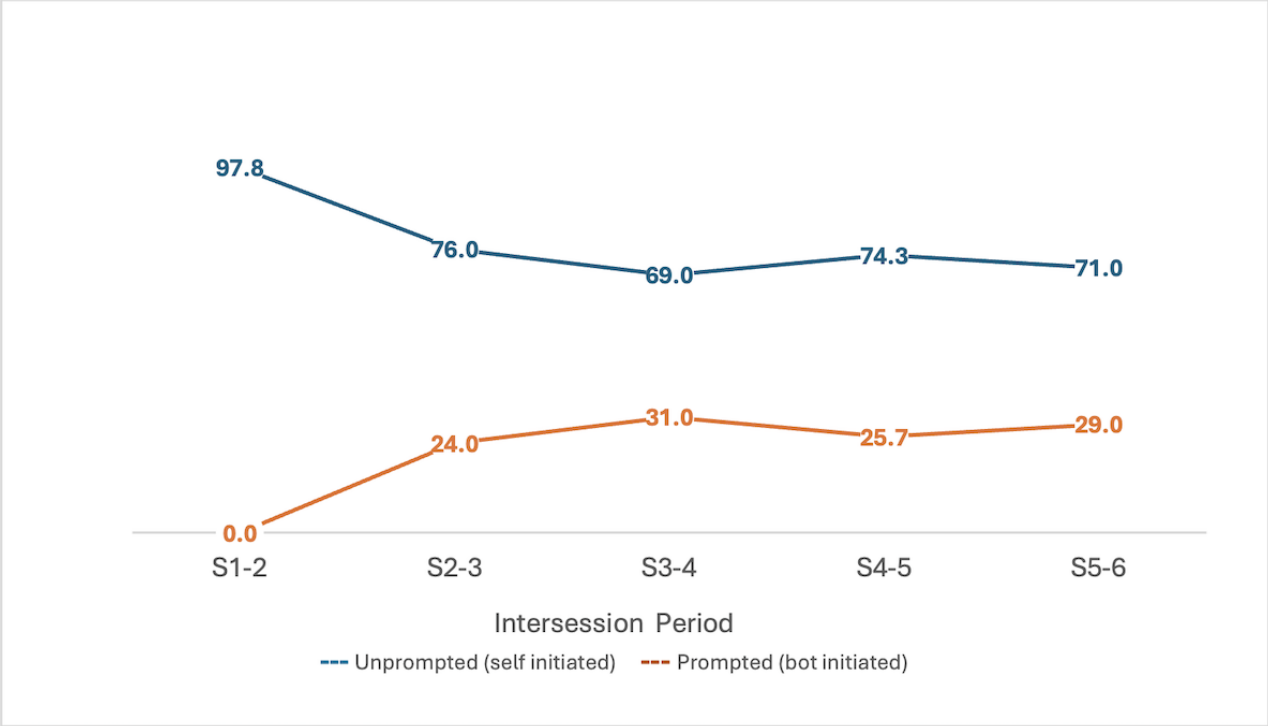

Pattern of Engagement Based on Sociodemographics

On average, women accessed the chatbot with greater frequency and tended to spend twice the amount of time engaging with it compared to men. Similarly, younger individuals between the ages of 18 and 25 years had frequencies and durations of engagement that were 2-3 times those for patients aged above 35 years. Table 1 lists the various engagement metrics in terms of gender and age.

Table 1. Sociodemographic differences in the pattern of engagement with the chatbot.

Sociodemographics	Frequency of engagement (times), value (95% CI)	Time spent engaging (minutes), value (95% CI)
Gender		
Female	61.88 (36.57-87.20)	62.24 (22.76-101.71)
Male	41.54 (32.57-50.51)	26.4 (15.40-37.40)
Age (years)[a]		
18-25	71.78 (27.72-115.84)	66.58 (10.54-122.62)
26-35	50.57 (33.35-67.79)	42.78 (26.61-58.94)
>35	34.00 (22.95-45.05)	22.2 (19.53-25.26)

[a]Regarding age, the *P* values were .15 for frequency of engagement and .32 for time spent engaging.

Retention Rate

Despite decreases in the frequency and duration of interactions over time, participant engagement with the chatbot remained remarkably high throughout the treatment period. All 30 (100%) participants actively engaged with the chatbot from session 1 to session 5, and there was a slight decline in engagement observed after session 5, with 28 (93%) participants still engaging with the chatbot at the end of treatment.

Participant Outcomes

All the participants in the pilot study completed the treatment. The average PHQ-9 score at baseline was 13.2 (95% CI 12.53-13.87) and decreased significantly to 3.8 (95% CI 2.77-4.83) upon treatment completion ($P<.001$). This reflects a reduction from a clinically moderate range of severity to a clinically nonsignificant range. Multiple regression analysis examined whether gender and age had an effect on the frequency of engagement. While the results showed that females engaged with the chatbot at a higher frequency, it was statistically insignificant ($P=.19$). A negative correlation was seen between age and the frequency of engagement, that is, as participant age increased, their engagement frequency decreased by 1.85 units, a relationship that was marginally significant ($P=.07$) (see Table 2).

XSL•FO
RenderX

Table 2. Multiple regression analysis to examine the relationship of the gender and age of patients with chatbot engagement.

Independent variable[a]	Regression coefficient	Standard error	*t* test (*df*)	*P* value
Gender (male)	–18.26	13.61	–1.34 (27)	.19
Age (years)	–1.85	0.98	–1.89 (27)	.07

[a]R^2=0.1768 (model explains 17.7% of variance in engagement); adjusted R^2=0.1159; residual standard error=36.83; *F*-statistic=2.9 on 2 and 27 degrees of freedom; model *P* value=.07 (not statistically significant at the 5% level).

Acceptability, Feasibility, and Effectiveness of PracticePal: Qualitative Findings

What Participants Liked About PracticePal

Participants generally found the PracticePal chatbot to be a valuable tool. They appreciated how the chatbot simplified concepts and made it easier for them to understand the principles of behavior activation, as explained by the ASHAs during the HAP therapy sessions. It reinforced the therapy messages and helped in the application of behavior activation strategies.

> *She (ASHA counsellor) told me, I will get a lot of help through this mobile if I go through it. She asked me to go through it and I did. Madam (ASHA counsellor) told me about social interaction, and I saw that (on the chatbot); it helped me a lot.* [42 years, male, married]

> *Because if you keep it (mobile) then you also get entertainment also. It (chatbot) guides us on everything, such as, if we have tension then take our mind off it, meet people, talk to them. It tells about whatever problems we have. I used to listen to and watch the videos every day.* [30 years, female, married]

The multimedia content was a well-received feature. Participants found the videos relatable and reported that they summarized and reiterated the session learnings effectively.

> *In the video they used to explain it so nicely that if we are in depression then we should walk outside so that our mind will be light, all this they explained properly. So, in that I understood a little that I should not be at home, I should go outside, roam, see others a little, understand everything.* [22 years, female, married]

> *That video resonated with me the most. Just like ASHA Didi (sister) explained to me, in that video, she (the lady in the video) explained to another person who used to be troubled by her husband's problems, such as drinking and not saving money. She explained how to manage your life, like how to save, to do something and not stay idle.* [21 years, female, married]

Another theme some participants reported was that they could use the chatbot at times convenient to them, for example, during their free time, while others specifically made time for it and integrated it into their daily routines.

> *When I had free time, I would use it. So, whenever I was bored, I used to open (the video on) social interaction and watch that again. So, I used to watch as much as I liked and then close it later.* [42 years, male, married]

> *...I had to cook food, my elder daughter would go to school, and I would go to the shop. I would watch (the videos) there. I would (also) watch while sitting at home in the evening.* [30 years, female, married]

Many participants mentioned that the chatbot could also benefit other people with stress and related health issues, and some had even shared it with their family and friends.

> *It (Chatbot) should also be given to those who cannot sleep, it should also be given to those who live in a lot of tension, whose mind will not be able to function, and it must be given to those suffering from depression.* [22 years, female, married]

> *I showed it to my friends and others, and they said it (chatbot) is good. They said if they know anyone going through these problems, they will tell them about this (chatbot) so they can fix their problems.* [21 years, female, married]

How PracticePal Complemented the Delivery of the HAP by ASHAs

ASHAs reported that PracticePal improved their ability to communicate and explain therapy messages clearly, reducing their anxiety about delivering incorrect information. They reported that the videos on the chatbot aided them in communicating messages more efficiently.

> *The program (use of PracticePal) is really good. We have got a lot of help from this for ourselves, for counselling and for the patients as well. The videos worked better for places where we weren't able to tell the patient. People usually don't talk so much; sometimes we also used to hesitate because we feared we would say something bad.* [FGD with ASHAs]

> *The books we gave them to read sometimes she (patient) used to read them, and sometimes she didn't. Yes, the mobile was better… She said reading was difficult for her, but she understood better by watching videos. It refreshed her mind.* [FGD with ASHAs]

ASHAs also appreciated some of the key features of PracticePal. For example, they found that the PracticePal avatar made it easier for patients to relate to the chatbot and improved their engagement in therapy. Some appreciated the feature that provided a summary of patient engagement, as it helped them feel more prepared for upcoming sessions and facilitated discussions with patients during the sessions.

XSL•FO
RenderX

> *She (chatbot) used to come wearing a saree; they (patients) used to say, "Didi, she looks like you, and she is like you.* [FGD with ASHAs]

> *We used to get a summary sheet that would come to us. Through that, we could come to know whether the patient is really doing any activity or not. If we asked the people using the booklet, they would tell us that they did, but how do we know whether they did it or not.* [FGD with ASHAs]

ASHAs reported encountering resistance from patients' family members when visiting homes to deliver counseling sessions. These family members often questioned the purpose of the visits and the need for counseling, with some even implying that the ASHAs might think their relatives were "crazy." In such cases, the chatbot allowed patients to receive therapy content on their mobile phones, as family members had no problems with the patients' phone use. Thus, the chatbot was instrumental in bypassing the stigma and resistance associated with in-person counseling sessions.

How PracticePal Helped the Recovery Process

Many participants reported that the chatbot's content, particularly the videos, motivated them to engage in behavior activation, interact more, and solve problems, and that these activities led to improvements in their mood.

> *It (chatbot) said that if you are living under stress, then go out for a walk, look at others, try to understand their point of view very well, just do not stay alone, do not feel stressed. The videos helped me understand. I felt better after seeing it. Earlier I did not do anything, after seeing it I started doing everything. I sit (with others), I work, I feel good in everything.* [22 years, female, married]

> *Earlier, I used to not talk about my problems, I just used to say that there is tension, there is tension, but used to not say that I have debt, I don't have anyone (family). So, I used to not tell it to anyone, it was like a secret… But they (videos) said that if there is any problem then we have to share it. Have to meet people, have to talk, and have to sit. So, I liked this that slowly I told it to everyone then all of them motivated me, that everything will be all right by the time, it just takes some time. So, through this I got that I used to talk, before I used to hesitate.* [30 years, female, married]

Participants also mentioned that other chatbot features, such as reminders and mood charts, were useful in completing the homework tasks given by ASHAs during the counseling sessions. These tools gave them timely reminders, helped them track their mood, made it easier to stay on top of the exercises, and enabled them to fit the tasks into their daily routines.

> *It (chatbot) used to remind me every morning and evening of what I had to do that day. It helped a lot; complete plans used to come. The messages (reminders) told me to do things, like watching videos. If we get daily messages, it uplifts the mood. Everyone should do these activities like socializing, making plans. I really liked it.* [21 years, female, married]

> *There was a mood scale in it. We had to rate our mood, like how our mood was. So, in the beginning, my mood was like, I would give it a 5 or 6, but gradually, my mood improved. I started giving it a 10.* [21 years, female, married]

Challenges Faced With Using PracticePal

Although there was almost full internet coverage in the study districts, participants sometimes faced issues with slow bandwidth and weak cellular network indoors. While most features of the chatbot worked over low bandwidth, accessing multimedia content faced difficulties. Poor weather conditions, such as rain and cloudy skies, further affected internet connectivity. Some participants had difficulty understanding the flow of PracticePal and choosing from the various menu options provided during the chatbot's conversation. Anticipating this, the study team had provided them with user manuals, which had screengrabs from the chatbot and clear instructions on the options to choose. Yet, ASHAs had to initially guide the participants in using and navigating the chatbot features during their sessions.

> *Yes, in the beginning, there was difficulty in using it (chatbot). But when the ASHA would visit for each session, she would start it (chatbot on the mobile) and ask me to read the book (user manual). I read it and understood everything.* [22 years, female, married]

> *Initially I had trouble understanding it (chatbot), understanding what to see. Then Didi (ASHA) helped me, but I was not able to understand it right away.* [42 years, male, married]

Some participants felt that people from different age groups might find it difficult to use the chatbot due to their unfamiliarity with technology. They felt that older individuals or those with less exposure to digital tools might especially struggle to engage with it. It was also pointed out that the chatbot may not be suitable for those who cannot read, limiting its benefits for people with low literacy. It was suggested that the text content be replaced with explanatory videos.

> *I mean, I am young, so I can do it (use the mobile phone), but there are elders who can't. ASHA Didi should explain to them.* [21 years, female, married]

> *Everything is written in it (chatbot). I can read it. But for someone who cannot read, for them it will be easier to follow a video on how to do it (use the chatbot).* [22 years, female, married]

Some participants felt that the videos provided through the chatbot were repetitive, which led to boredom over time. They appreciated the usefulness of the videos but expressed a strong preference for new content to maintain their interest.

> *…when something new came up, they would watch it with interest. But it was repeating the same thing, so they were losing interest.* [FGD with ASHAs]

XSL•FO
RenderX

> *...they used to say it's the same thing, it's the same thing. So, something new should be added, a new story or improvement from a patient's story. That would benefit them more. [FGD with ASHAs]*

Discussion

This study described the development of PracticePal, a chatbot designed to support the delivery of the HAP, which is a frontline worker–delivered behavioral activation-based psychological treatment for depression, and assessed its feasibility and acceptability in a rural Indian community. The treatment was delivered by ASHAs, community health workers in rural India, to members of their community who had been screened for depression. The chatbot performed multiple roles, particularly reinforcing the knowledge and skills learned in therapy sessions and promoting patient engagement between sessions. Our findings indicate that patient engagement with the chatbot was high and remained consistently strong throughout the duration of the psychological treatment. Participants especially appreciated the reminder feature, which prompted them to complete their planned activities; the multimedia content, which reinforced the key messages from therapy sessions; and the mood monitoring feature. The use of the chatbot improved the overall treatment experience for patients, and the findings are consistent with those of other chatbot studies [36,37].

Our findings of high levels of engagement, acceptability, and feasibility, and the perceived benefits of the chatbot are likely the result of the study design incorporating feedback from both frontline providers and patients with depression. As our results indicate, there was a higher frequency of engagement among female participants than male participants. One of the reasons for this could simply be that there were more female participants than male participants. There was no significant difference observed in the times of the day male and female participants interacted with the chatbot. In addition, it is possible that the higher engagement with the chatbot among female participants than male participants was influenced by differences in smartphone ownership. Women were less likely to own personal smartphones and thus may have used the study-provided phones more extensively, whereas men who owned smartphones may have divided their attention between their personal phones and the study device. We intend to examine this gender-specific engagement pattern further in future studies through a real-world and adequately powered design, wherein participants will access the chatbot using their own smartphones. We note that our results differ from those of some prior studies, which observed that digital tool usage significantly declined over time, typically due to lack of personalization, privacy concerns, trustworthiness, and user interface issues [38,39]. Specifically, we found that our study demonstrated a remarkably high retention rate, with 28 of 30 (93%) participants continuing to engage with the chatbot throughout the average 60-day therapy period. This exceptional engagement can likely be attributed to the chatbot's role as a complementary therapy aid to in-person NSP-delivered therapy. We suggest 4 key factors that may have contributed to the sustained engagement with PracticePal. First, PracticePal was not designed to replace human interaction but to complement the therapeutic experience without replacing the human connection. The interactive multimedia content further enhanced the experience, reinforcing therapeutic messages in a way that was both engaging and enjoyable for patients. Second, the provider (ie, ASHAs) played a crucial role in reinforcing the use of the chatbot, regularly reminding patients to track their activities and access the content. This was facilitated by a summary sheet that was provided to the ASHAs prior to each session. This sheet detailed the participant's frequency of engagement, activity completion, and mood reports. The regular reinforcement of participants by the ASHAs helped establish the chatbot as a natural extension of the therapy, making it more acceptable and valuable to patients. It became part of a patient's routine rather than an external or unfamiliar tool. In comparison to using a paper workbook, which was the original format of the HAP, the chatbot was more readily accessible due to the ubiquitous presence of smartphones, which enabled engagement on demand, and the ease of navigation through its features. Third, the chatbot's avatar was designed based on local stakeholder input, resembling ASHAs themselves, which helped create a sense of trust and relatability among patients. Finally, the chatbot was embedded in a widely used messaging platform, as opposed to a stand-alone app that has no other purpose.

One of the key strengths of our study was that the chatbot was co-designed with stakeholders and was promoted by frontline health workers who were already embedded in the community, enhancing its credibility and reach. Rather than functioning as a standalone digital tool, the chatbot acted as an aid for frontline health workers, improving the quality of therapy provided. The use of multimedia content increased engagement and buy-in from users, who appreciated the clear and effective messaging. Our study has several limitations. First, the small sample size of the pilot limits the generalizability of the findings. Second, participants in the study were provided with smartphones and data recharges to engage with the chatbot. This raises concerns about accessibility, as individuals without smartphones would be unable to use the chatbot during therapy, potentially limiting its reach to those who already have access to such technology. However, the penetration of smartphones is already very high and rapidly rising across India. Third, the lack of a comparator arm makes it impossible to assess the incremental efficacy of PracticePal added to the HAP compared to the HAP alone. Both the baseline and endline PHQ-9 assessments were conducted by an independent researcher. Baseline assessments were conducted face-to-face, while endline assessments were conducted via phone survey. The use of independent assessors helped minimize the risk of provider bias. However, we acknowledge that the face-to-face format of the baseline assessment may have introduced a degree of social desirability bias, as participants may have felt less comfortable disclosing symptoms in person. In contrast, the phone-based endline assessment may have offered a greater sense of privacy, thereby reducing, though not entirely eliminating, the likelihood of socially desirable responses. Although the HAP is an evidence-based treatment with a robust effect size, future research should involve adequately powered, multi-arm randomized controlled trials to compare the outcomes of the HAP alone, the HAP with the chatbot as a therapy aid, and the chatbot as a standalone intervention.

XSL•FO
RenderX

We also note that the PracticePal chatbot itself may benefit from further enhancements. First, while Telegram was used during the pilot phase due to lower costs, we recognize that WhatsApp is more widely used in India. For future scaling efforts, deploying PracticePal on WhatsApp could enhance accessibility and reach. Second, this chatbot was engineered before the widespread availability of generative artificial intelligence (AI) technologies. The chatbot used in this study was rule-based, and while it was effective, the integration of more advanced AI technologies could offer greater personalization and engagement for users. Future versions of PracticePal could integrate more AI-powered conversational tools to further enhance patient outcomes. Third, PracticePal's reliance on text-based interactions presented challenges for participants with limited literacy skills. This highlights the need for incorporating audio-to-text capabilities, voice commands, or more video content to enhance accessibility for users with varying literacy levels. Future iterations of the chatbot should incorporate these features to ensure broader usability and inclusivity, particularly in low-literacy settings. Lastly, this study's findings should be interpreted with caution due to the small sample size, which may limit the generalizability of the results and reduce the statistical power to detect differences in chatbot engagement metrics.

Although there is substantial evidence supporting the effectiveness of mobile platform–delivered digital interventions for improving health outcomes and service utilization in high-income countries, most of these interventions seek to replace a human provider, and much of the evidence is from high-income settings [40]. Our experience reported in a rural Indian community shows that CAs like the PracticePal chatbot, which complement rather than replace human interactions, are highly acceptable to patients with little prior experience of mental health care and can facilitate the delivery of psychosocial interventions by frontline workers. The rapid increase in smartphone penetration across all regions, not least in India, makes the deployment of such chatbots highly feasible.

The PracticePal chatbot is feasible and acceptable as a therapy aid to enhance the quality of NSP-delivered psychosocial interventions in low-resource settings. Its promising potential to improve mental health outcomes underscores the importance of further research to refine and expand the application of chatbot interventions, especially after the advent of generative AI.

Looking ahead, it would be valuable to assess the effectiveness of such chatbots in definitive trials and to assess the efficacy, effectiveness, and cost-effectiveness of chatbots, such as PracticePal, in comparison with traditional treatment options, including as a standalone intervention. Moreover, research should explore the optimal integration of chatbots into broader mental health programs across different settings, addressing potential implementation challenges and scalability.

Acknowledgments

This project was made possible through the generous support of the Agency Fund, whose grant to Sangath and Dimagi was instrumental in the development and testing of the conversational agent. We would like to extend our sincere gratitude to the dedicated developers from Dimagi, particularly Akash Lohi and Jaideep Singh, who were receptive to repeated requests for iterations based on stakeholder inputs. Additionally, we are deeply appreciative of the efforts of the field staff from the EMPOWER project, including Balkrishna Tripathi, Prashant Sharma, Vandana Shukla, Mahendra Singh, Ramkumar Sapre, Pratiksha, Pragya Sharma, Gorishankar Dwivedi, Ritesh Dubey, Brajesh Parashar, and Shweta Dubey, who worked tirelessly to ensure the effective implementation of the project activities on the ground.

We are grateful to the National Health Mission, Madhya Pradesh, for granting the necessary permissions and providing us access to stakeholders, without which this study would not have been possible. Our heartfelt thanks also go to the accredited social health activists and patients in the community, whose participation and support were vital to the success of this project.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

NL is the Chief Strategy Officer at Dimagi, a social enterprise with offerings for digital solutions for mental health, including those assessed in this paper.

Multimedia Appendix 1
Multimedia content made available through the chatbot.
[PDF File (Adobe PDF File), 52 KB - formative_v9i1e73563_app1.pdf]

Multimedia Appendix 2
PracticePal conversational workflow.
[PDF File (Adobe PDF File), 890 KB - formative_v9i1e73563_app2.pdf]

XSL•FO
RenderX

Multimedia Appendix 3
PracticePal avatar and expressions.
[PDF File (Adobe PDF File), 738 KB - formative_v9i1e73563_app3.pdf]

Multimedia Appendix 4
Screengrabs of the PracticePal chatbot interface.
[PDF File (Adobe PDF File), 292 KB - formative_v9i1e73563_app4.pdf]

References

1. Herrman H, Patel V, Kieling C, Berk M, Buchweitz C, Cuijpers P, et al. Time for united action on depression: a Lancet–World Psychiatric Association Commission. The Lancet 2022 Mar;399(10328):957-1022. [doi: 10.1016/s0140-6736(21)02141-3]
2. Bashar M, Mehra A, Aggarwal A. Integrating mental health into primary care for addressing depression in a rural population: An experience from North India. Indian J Psychiatry 2019;61(3):319. [doi: 10.4103/psychiatry.indianjpsychiatry_374_18]
3. Thirunavukarasu M, Thirunavukarasu P. Training and National deficit of psychiatrists in India - A critical analysis. Indian J Psychiatry 2010 Jan;52(Suppl 1):S83-S88 [FREE Full text] [doi: 10.4103/0019-5545.69218] [Medline: 21836723]
4. Weaver LJ, Karasz A, Muralidhar K, Jaykrishna P, Krupp K, Madhivanan P. Will increasing access to mental health treatment close India's mental health gap? SSM - Mental Health 2023 Dec;3:100184. [doi: 10.1016/j.ssmmh.2022.100184]
5. Thirunavukarasu M. Closing the treatment gap. Indian J Psychiatry 2011 Jul;53(3):199-201 [FREE Full text] [doi: 10.4103/0019-5545.86803] [Medline: 22135436]
6. Muke SS, Tugnawat D, Joshi U, Anand A, Khan A, Shrivastava R, et al. Digital training for non-specialist health workers to deliver a brief psychological treatment for depression in primary care in India: findings from a randomized pilot study. Int J Environ Res Public Health 2020 Sep 01;17(17):6368 [FREE Full text] [doi: 10.3390/ijerph17176368] [Medline: 32883018]
7. Singla DR, Kohrt BA, Murray LK, Anand A, Chorpita BF, Patel V. Psychological treatments for the world: lessons from low- and middle-income countries. Annu Rev Clin Psychol 2017 May 08;13(1):149-181 [FREE Full text] [doi: 10.1146/annurev-clinpsy-032816-045217] [Medline: 28482687]
8. Chatterton ML, Belay YB. Cost-effectiveness of task-shifting in resource-constrained settings. The Lancet Global Health 2024 Apr;12(4):e546-e547. [doi: 10.1016/s2214-109x(24)00038-x]
9. Singla DR, Schleider JL, Patel V. Democratizing access to psychological therapies: Innovations and the role of psychologists. J Consult Clin Psychol 2023 Nov;91(11):623-625. [doi: 10.1037/ccp0000850] [Medline: 37782248]
10. Patel V, Xiao S, Chen H, Hanna F, Jotheeswaran AT, Luo D, et al. The magnitude of and health system responses to the mental health treatment gap in adults in India and China. The Lancet 2016 Dec;388(10063):3074-3084. [doi: 10.1016/s0140-6736(16)00160-4]
11. Agrawal R, Sood M, Patel A, Sharma T, Yadav H, Kaur J, et al. Scaling up the task-sharing of an evidence-based psychological treatment for depression in rural India. medRxiv. 2024. URL: https://www.medrxiv.org/content/10.1101/2024.10.23.24315962v1 [accessed 2025-07-10]
12. Malik K, Ibrahim M, Bernstein A, Venkatesh RK, Rai T, Chorpita B, et al. Behavioral Activation as an 'active ingredient' of interventions addressing depression and anxiety among young people: a systematic review and evidence synthesis. BMC Psychol 2021 Oct 07;9(1):150 [FREE Full text] [doi: 10.1186/s40359-021-00655-x] [Medline: 34615559]
13. Patel V, Weobong B, Weiss HA, Anand A, Bhat B, Katti B, et al. The Healthy Activity Program (HAP), a lay counsellor-delivered brief psychological treatment for severe depression, in primary care in India: a randomised controlled trial. The Lancet 2017 Jan;389(10065):176-185. [doi: 10.1016/s0140-6736(16)31589-6]
14. Noble JM, Vuong W, Surood S, Urichuk L, Greenshaw AJ, Agyapong VIO. Text4Support mobile-based programming for individuals accessing addictions and mental health services-retroactive program analysis at baseline, 12 weeks, and 6 months. Front Psychiatry 2021 May 28;12:640795 [FREE Full text] [doi: 10.3389/fpsyt.2021.640795] [Medline: 34122173]
15. Beratarrechea A, Lee AG, Willner JM, Jahangir E, Ciapponi A, Rubinstein A. The impact of mobile health interventions on chronic disease outcomes in developing countries: a systematic review. Telemed J E Health 2014 Jan;20(1):75-82 [FREE Full text] [doi: 10.1089/tmj.2012.0328] [Medline: 24205809]
16. Jain M, Duvendack M, Shisler S, Parsekar SS, Leon MDA. Effective interventions for improving routine childhood immunisation in low and middle-income countries: a systematic review of systematic reviews. BMJ Open 2024 Feb 15;14(2):e074370 [FREE Full text] [doi: 10.1136/bmjopen-2023-074370] [Medline: 38365291]
17. Sabour S, Zhang W, Xiao X, Zhang Y, Zheng Y, Wen J, et al. A chatbot for mental health support: exploring the impact of Emohaa on reducing mental distress in China. Front Digit Health 2023 May 4;5:1133987 [FREE Full text] [doi: 10.3389/fdgth.2023.1133987] [Medline: 37214342]
18. Arunachal Pradesh Population 2025. Population Census. URL: https://www.census2011.co.in/census/state/arunachal+pradesh.html [accessed 2024-09-26]

XSL•FO
RenderX

19. Menon P, Deolalikar A, Bhaskar A. India State Hunger Index - Comparisons of Hunger Across States. International Food Policy Research Institute. 2009. URL: https://cdm15738.contentdm.oclc.org/utils/getfile/collection/p15738coll2/id/13891/filename/13892.pdf [accessed 2024-09-26]
20. Annual Plan 2010-11 Volume-I. Government of Madhya Pradesh. 2010. URL: https://cag.gov.in/uploads/download_audit_report/2011/Madhya_Pradesh_SF_2011_App.pdf [accessed 2024-09-26]
21. Suryanarayana MH, Agrawal A, Prabhu KS. Inequality-adjusted human development index: states in India. Indian Journal of Human Development 2016 Dec 08;10(2):157-175. [doi: 10.1177/0973703016675793]
22. Kokane A, Pakhare A, Gururaj G, Varghese M, Benegal V, Rao GN, et al. Mental health issues in Madhya Pradesh: insights from national mental health survey of India 2016. Healthcare (Basel) 2019 Mar 31;7(2):53 [FREE Full text] [doi: 10.3390/healthcare7020053] [Medline: 30935112]
23. Shidhaye R, Lyngdoh T, Murhar V, Samudre S, Krafft T. Predictors, help-seeking behaviour and treatment coverage for depression in adults in Sehore district, India. BJPsych Open 2017 Sep 02;3(5):212-222 [FREE Full text] [doi: 10.1192/bjpo.bp.116.004648] [Medline: 28904815]
24. Highlights of Telecom Subscription Data as on 31st March, 2023. Telecom Regulatory Authority of India. 2023. URL: https://www.trai.gov.in/sites/default/files/2024-08/PR_No.46of2023.pdf [accessed 2024-09-26]
25. Highlights of Telecom Subscription Data as on 31st May, 2024. Telecom Regulatory Authority of India. 2024. URL: https://www.trai.gov.in/sites/default/files/2024-08/PR_No.40of2024.pdf [accessed 2024-09-26]
26. Scott K, George AS, Ved RR. Taking stock of 10 years of published research on the ASHA programme: examining India's national community health worker programme from a health systems perspective. Health Res Policy Syst 2019 Mar 25;17(1):29 [FREE Full text] [doi: 10.1186/s12961-019-0427-0] [Medline: 30909926]
27. Ved R, Scott K, Gupta G, Ummer O, Singh S, Srivastava A, et al. How are gender inequalities facing India's one million ASHAs being addressed? Policy origins and adaptations for the world's largest all-female community health worker programme. Hum Resour Health 2019 Jan 08;17(1):3 [FREE Full text] [doi: 10.1186/s12960-018-0338-0] [Medline: 30616656]
28. Waskel B, Dixit S, Singodia R, D.K. P, Toppo M, SC T, et al. Evaluation of Asha Programme in selected block of Raisen district of Madhya Pradesh under the national rural health mission. J Evol Med Dent Sci 2014 Jan 16;03(03):689-694. [doi: 10.14260/jemds/2014/1887]
29. Kruse K. Introduction to Instructional Design and the ADDIE Model. Connecticut State. 2002. URL: https://portal.ct.gov/-/media/ctdn/ttt2015/ttt2015module5introinstdesignaddiepdf.pdf [accessed 2024-09-16]
30. Gonsalves PP, Hodgson ES, Kumar A, Aurora T, Chandak Y, Sharma R, et al. Design and development of the "POD Adventures" smartphone game: a blended problem-solving intervention for adolescent mental health in India. Front Public Health 2019 Aug 23;7:238 [FREE Full text] [doi: 10.3389/fpubh.2019.00238] [Medline: 31508404]
31. Khan A, Shrivastava R, Tugnawat D, Singh A, Dimidjian S, Patel V, et al. Design and development of a digital program for training non-specialist health workers to deliver an evidence-based psychological treatment for depression in primary care in India. J Technol Behav Sci 2020 Dec;5(4):402-415 [FREE Full text] [doi: 10.1007/s41347-020-00154-7] [Medline: 34150992]
32. Nadkarni A, Costa S, Gupta D, Fernandes D, Catalano A, Velleman R, et al. The systematic development of a mobile phone-delivered brief intervention for hazardous drinking in India. J Subst Abuse Treat 2021 Jul;126:108331. [doi: 10.1016/j.jsat.2021.108331] [Medline: 34116809]
33. Fraser AD, Branson I, Hollett RC, Speelman CP, Rogers SL. Do realistic avatars make virtual reality better? Examining human-like avatars for VR social interactions. Computers in Human Behavior: Artificial Humans 2024 Aug;2(2):100082. [doi: 10.1016/j.chbah.2024.100082]
34. What is agile project management? Association for Project Management. URL: https://www.apm.org.uk/resources/find-a-resource/agile-project-management/ [accessed 2024-10-24]
35. Acceptability/feasibility. Dissemination & Implementation Models in Health. URL: https://dissemination-implementation.org/constructs/acceptability-feasibility/ [accessed 2023-05-14]
36. Daley K, Hungerbuehler I, Cavanagh K, Claro HG, Swinton PA, Kapps M. Preliminary evaluation of the engagement and effectiveness of a mental health chatbot. Front Digit Health 2020 Nov 30;2:576361 [FREE Full text] [doi: 10.3389/fdgth.2020.576361] [Medline: 34713049]
37. Vaidyam AN, Wisniewski H, Halamka JD, Kashavan MS, Torous JB. Chatbots and conversational agents in mental health: a review of the psychiatric landscape. Can J Psychiatry 2019 Jul;64(7):456-464 [FREE Full text] [doi: 10.1177/0706743719828977] [Medline: 30897957]
38. Balcombe L, De Leo D. Human-computer interaction in digital mental health. Informatics 2022 Feb 22;9(1):14. [doi: 10.3390/informatics9010014]
39. Torous J, Nicholas J, Larsen ME, Firth J, Christensen H. Clinical review of user engagement with mental health smartphone apps: evidence, theory and improvements. Evid Based Ment Health 2018 Aug 05;21(3):116-119 [FREE Full text] [doi: 10.1136/eb-2018-102891] [Medline: 29871870]

XSL•FO
RenderX

40. Marcolino MS, Oliveira JAQ, D'Agostino M, Ribeiro AL, Alkmim MBM, Novillo-Ortiz D. The impact of mHealth interventions: systematic review of systematic reviews. JMIR Mhealth Uhealth 2018 Jan 17;6(1):e23 [FREE Full text] [doi: 10.2196/mhealth.8873] [Medline: 29343463]

Abbreviations

AI: artificial intelligence
ASHA: accredited social health activist
CA: conversational agent
FGD: focus group discussion
HAP: Healthy Activity Program
IDI: in-depth interview
NSP: nonspecialist provider
PHQ-9: Patient Health Questionnaire-9

Edited by A Mavragani; submitted 06.03.25; peer-reviewed by B Thies; comments to author 09.04.25; revised version received 30.04.25; accepted 11.06.25; published 29.08.25.

Please cite as:
Agrawal R, Monteiro K, Narasimhamurti NS, Sharma S, Suryawanshi A, Bariya A, Narvekar S, Bagnoli L, Saxena M, Magoun L, Parsekar SS, Pozuelo JR, Lesh N, Sood M, Sharma T, Yadav H, Bhan A, Nadkarni A, Patel V
A Conversational Agent (PracticePal) to Support the Delivery of a Brief Behavioral Activation Treatment for Depression in Rural India: Development and Pilot-Testing Study
JMIR Form Res 2025;9:e73563
URL: https://formative.jmir.org/2025/1/e73563
doi:10.2196/73563
PMID:

©Ravindra Agrawal, Kimberley Monteiro, Nityasri Sankha Narasimhamurti, Shreya Sharma, Amruta Suryawanshi, Aman Bariya, Shravani Narvekar, Lilianna Bagnoli, Mohit Saxena, Lauren Magoun, Shradha S Parsekar, Julia R Pozuelo, Neal Lesh, Mohit Sood, Tanushri Sharma, Harshita Yadav, Anant Bhan, Abhijit Nadkarni, Vikram Patel. Originally published in JMIR Formative Research (https://formative.jmir.org), 29.08.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Machine Learning–Enhanced Surveillance for Surgical Site Infections in Patients Undergoing Colon Surgery: Model Development and Evaluation Study

Ugur Celik[1*], MSc; Feifan Liu[2*], PhD; Kimiyoshi Kobayashi[2,3,4*], MD, MBA; Richard T Ellison III[3,5*], MD; Yurima Guilarte-Walker[2*], MSc; Deborah Ann Mack[5*], RN; Qiming Shi[1*], MSc; Adrian Zai[2*], MD, MPH, PhD

[1]Center for Clinical and Translational Sciences, University of Massachusetts Chan Medical School, Worcester, MA, United States

[2]Department of Population and Quantitative Health Sciences, Division of Health Informatics and Implementation Science, University of Massachusetts Chan Medical School, Worcester, MA, United States

[3]Department of Medicine, University of Massachusetts Chan Medical School, Worcester, MA, United States

[4]UMass Memorial Medical Center, Worcester, United States

[5]Infection Control Department, UMass Memorial Medical Center, Worcester, MA, United States

[*]all authors contributed equally

Corresponding Author:
Ugur Celik, MSc
Center for Clinical and Translational Sciences
University of Massachusetts Chan Medical School
55 Lake Avenue North
Worcester, MA, 01655
United States
Phone: 1 7744554679
Email: ugur_celik@student.uml.edu

Abstract

Background: Surgical site infections (SSIs) are one of the most common health care–associated infections, accounting for nearly 20% of all health care–associated infections in hospitalized patients. SSIs are associated with longer hospital stays, increased readmission rates, higher health care costs, and a mortality rate twice that of patients without infections.

Objective: This study aimed to develop and evaluate machine learning (ML) models for augmenting SSI surveillance after colon surgery with the goal of improving the efficiency of infection control practices by prioritizing patients at high risk.

Methods: We conducted a retrospective study using data from 1508 patients undergoing colon surgery treated between 2018 and 2023 at a single academic medical center. Of these 1508 patients, 66 (4.4%) developed SSIs as adjudicated by infection control practitioners following Centers for Disease Control and Prevention National Healthcare Safety Network criteria. Data included 78 structured variables (eg, demographics, comorbidities, vital signs, laboratory tests, medications, and operative details) and 2 features derived from unstructured clinical notes using natural language processing. ML models—logistic regression, random forest, and Extreme Gradient Boosting (XGBoost)—were trained using stratified 80/20 train-test splits. Class imbalance was addressed using cost-sensitive learning and the synthetic minority oversampling technique. Model performance was evaluated using precision, recall, F_1-score, area under the receiver operating characteristic curve, and Brier scores for calibration.

Results: Of the 1508 patients, those who developed SSIs had longer hospital stays (mean 8.1, SD 6.8 days vs mean 6.3, SD 10.5 days; $P<.001$), higher rates of an American Society of Anesthesiologists score of 3 (52/66, 79% vs 653/1442, 45.3%; $P<.001$), and elevated white blood cell counts (51/66, 77% vs 734/1442, 50.9%; $P<.001$). XGBoost achieved the best overall performance with an area under the receiver operating characteristic curve of 0.788, precision of 50%, recall of 38%, and Brier score of 0.035. Random forest yielded perfect precision (100%) but lower recall (23%), with a Brier score of 0.034. Logistic regression showed the highest recall (46%) but the lowest precision (10%), with a Brier score of 0.139. Feature importance analysis using Shapley additive explanations (SHAP) values revealed that the top predictors included recovery duration (SHAP=1.18), SSI keyword frequency (SHAP=1.12), patient age (SHAP=1.12), and American Society of Anesthesiologists score (SHAP=0.94), with natural language processing–derived features ranking among the top 10.

XSL•FO
RenderX

Conclusions: ML models can augment traditional SSI surveillance by improving early identification of patients at high risk. The XGBoost model offered the best trade-off between discrimination and calibration, suggesting its utility in clinical workflows. Incorporating structured and unstructured electronic health record data enhances model accuracy and clinical relevance, supporting scalable and efficient infection control practices.

(JMIR Form Res 2025;9:e75121) doi:10.2196/75121

KEYWORDS

surgical site infection; machine learning; surveillance; electronic health records; natural language processing; colon surgery; risk prediction; Extreme Gradient Boosting; XGBoost; random forest; artificial intelligence; AI

Introduction

Background

Surgical site infections (SSIs) are a significant category of health care–associated infections, representing a serious challenge to health care systems worldwide. SSIs are estimated to account for nearly 20% of all health care–associated infections among hospitalized patients [1]. In the United States, SSIs occur in approximately 2% to 4% of patients undergoing inpatient surgical procedures [2]. SSIs can lead to severe complications, including increased morbidity, extended hospital stays, higher readmission rates, and increased health care costs, ultimately impacting patient outcomes and putting a strain on health care resources [3-5]. The financial burden of SSIs is substantial, with associated costs reaching billions of dollars annually. SSIs are the third most costly type of health care–acquired infection, with an estimated cost of US $20,785 per patient case [6]. Patients with SSIs have a mortality rate twice that of patients without infections [7].

The current surveillance process for SSIs is resource intensive, requiring manual chart reviews by infection control practitioners to monitor surgical procedures and screen for potential infections. This manual surveillance is time-consuming and labor intensive and detracts from direct patient care [8]. The Centers for Disease Control and Prevention (CDC) National Healthcare Safety Network (NHSN) has established comprehensive guidelines to support the systematic tracking and identification of SSIs; yet, the manual workload remains significant [9].

Several studies have applied machine learning (ML) and natural language processing (NLP) to improve the detection and prediction of SSIs. These approaches have shown promise in automating surveillance and improving accuracy by using structured and unstructured clinical data. However, previous research has largely focused on retrospective detection rather than prospective prediction, relied on limited data sources, or failed to integrate both structured and unstructured data [10-13].

Objectives

To address these limitations, our study introduces an ML-based surveillance tool designed to enhance SSI monitoring. By integrating structured data and clinical notes, our approach improves detection efficiency and accuracy. Automating surveillance through ML reduces the burden on infection control practitioners, allowing them to focus on patients at high risk and critical tasks, ultimately improving patient outcomes and optimizing resource allocation.

Methods

Ethical Considerations

This study was reviewed by the University of Massachusetts Chan Medical School (UMass Chan) Institutional Review Board and deemed exempt under institutional guidelines for quality improvement. As a retrospective analysis of existing electronic health record (EHR) data with no patient contact, informed consent was not required. The institutional review board confirmed that secondary use of clinical data did not necessitate additional consent. All data were deidentified in accordance with HIPAA (Health Insurance Portability and Accountability Act) and analyzed within a secure, access-controlled environment at UMass Chan. Only authorized personnel had access to the data. No identifiable patient information appears in images or supplementary materials; all figures are fully anonymized.

Data Extraction and Preparation

Data for this study were extracted from the UMass Chan Data Lake, which is a copy of the Epic Clarity EHR system at the hospital and refreshes weekly. This data repository comprises 2.5 million unique patients since November 2017. We focused on patients who underwent colon surgeries between 2018 and 2023, identified using Current Procedural Terminology (CPT) codes according to the NHSN guidelines. Our cohort included 1508 procedures, with 66 (4.4%) confirmed SSI cases.

Cohort Inclusion and Exclusion Criteria

We identified all adult patients (aged ≥18 years) who underwent colon surgery at the University of Massachusetts Memorial Medical Center between January 1, 2018, and December 31, 2023, using CPT procedure codes per NHSN guidelines. Patients were included if they were aged ≥18 years at the time of surgery, underwent elective or urgent colon resection procedures (CPT codes 44140-44160 and 44204-44208), and had available postoperative follow-up data for at least 30 days. We excluded patients with noncolon procedures or combined multiorgan resections, those with missing critical EHR data (eg, American Society of Anesthesiologists [ASA] score, surgery date, or outcome label), and patients who were deceased before completion of the 30-day postoperative surveillance window. After applying these criteria, 1508 unique procedures remained for analysis.

Outcome Labeling and Gold Standard

All SSI labels were assigned by the hospital's infection prevention and control (IPC) team according to CDC NHSN

XSL•FO
RenderX

criteria. Each postoperative patient who underwent colon surgery is manually reviewed daily by IPC specialists, who screen charts, microbiology reports, wound assessments, and nursing notes against the NHSN definitions. Ambiguous cases are flagged for discussion at a weekly consensus meeting with at least 2 IPC specialists and a supervising epidemiologist, and final SSI determinations are made through consensus. The IPC department also conducts monthly peer audit exercises to ensure consistency with NHSN standards.

To ensure clinical relevance, infection control nurses provided input during the term selection process for clinical note analysis and validated the model's outputs against real-world clinical scenarios. The dataset included 78 structured variables, such as demographics, medications, laboratory test results, and medical histories (Table 1). In addition, 2 variables—SSI keyword count and SSI negation term count—were derived from unstructured clinical notes using NLP techniques.

Table 1. Candidate predictors used to train machine learning models in a retrospective cohort study of postoperative surgical site infection (SSI) after colon surgery at the University of Massachusetts Memorial Medical Center (Worcester, Massachusetts, United States) from January 1, 2018, to December 31, 2023.

Data domain	Data points
Demographics	Gender, race, ethnicity, age, alcohol usage and smoking status
Comorbidities	Obesity, cancer, diabetes, immunological disease, depression, dementia, anemia, heart failure, AIDS, and alcohol
Encounters	Total stay days and inpatient stay
Laboratory tests	Hemoglobin, culture, white blood cell count, and C-reactive protein
Medications	Antibiotics, immunosuppressants, and steroids
Patient	Patient ID and patient encounter ID
Surgery details	Surgery class, surgery procedure code, physician ID, number of procedures, surgery duration, surgery recovery duration, department, room number, wound status, anesthesia type, ASA[a] score, incision closure, and SSI
Vital signs	BMI
NLP[b]-derived features	SSI_Keyword and SSI_Negation
SSI keywords	"Fever," "nausea," "vomiting," "pain," "tenderness," "odynophagia," "dysphagia," "hypotension," "jaundice," "dysuria," "abscess," and "infection"

[a]ASA: American Society of Anesthesiologists.

[b]NLP: natural language processing.

Unstructured Data Processing

Specifically, we used the spaCy library (Explosion AI) [14] for tokenization and named entity recognition to identify relevant keywords associated with SSIs, such as "redness," "swelling," "drainage," and "purulent discharge." Negation detection was conducted using the NegEx algorithm [15], which allowed us to determine whether an identified keyword was negated in the context of the clinical note. For example, "no signs of infection" or "denies fever" would be identified as negated terms. These NLP tools enabled us to accurately quantify the occurrence of SSI-related terms and their context within the notes.

Infection-related terms were first extracted from the CDC NHSN SSI surveillance criteria and then reviewed and finalized by IPC nurses—who conduct daily chart surveillance—to ensure relevance to our colon surgery population. No formal statistical testing was conducted on the term list. All SSI related keywords were curated and validated by the IPC nurses based on daily practice. While transformer-based embeddings (eg, BioBERT) may capture richer linguistic patterns, we deferred their use because our IT department could not approve deployment of large language models in the current infrastructure. These contextual methods will be explored in follow-up work.

Data Processing and Feature Engineering

Data preprocessing involved imputing missing values in numerical columns using column-wise means, one-hot encoding categorical variables, and validating data types for compatibility with ML algorithms. The dataset was processed using the Python pandas library [16] and then deployed to a secure workspace, the Platform for Learning Health System environment at UMass Chan, for further analysis [17].

After median imputation of numerical variables and one-hot encoding of categorical variables, the dataset expanded to 150 features in total: the original 78 numerical inputs, 2 NLP-derived counts, and 70 dummy variables from one-hot encoding. For the logistic regression model, we then generated all second-degree polynomial interaction terms, yielding 11,325 features in the final matrix used for model fitting.

To better understand the relative impact of different imbalance-handling techniques, we conducted an ablation study on all 3 models (logistic regression, random forest, and Extreme Gradient Boosting [XGBoost]). Three strategies were compared: cost-sensitive approach only by applying class weights (class_weight=*balanced* for tree and linear models and scale_pos_weight for XGBoost), oversampling only via 1:1 random upsampling of the minority (SSI) class in the training set, and oversampling combined with cost-sensitive approaches. Pipelines for each strategy were retrained and evaluated on the

XSL•FO
RenderX

same held-out test set (n=302), and metrics including precision, recall, F_1-score (for the SSI class), area under the receiver operating characteristic curve (AUC-ROC), and Brier score were recorded.

Model Development

The dataset was split into training (80%) and validation (20%) sets using stratified sampling to maintain the class distribution. To address class imbalance, we applied cost-sensitive learning [18] and the synthetic minority oversampling technique (SMOTE) [19]. SMOTE works by creating synthetic examples of the minority class by interpolating between existing instances, thereby balancing the class distribution in the training set.

We developed 3 ML models: logistic regression [20], random forest [21], and XGBoost [22]. Each model was chosen based on specific strengths—logistic regression for interpretability, random forest for handling complex interactions, and XGBoost for efficient structured data analysis.

Logistic regression was chosen due to its simplicity and interpretability, which allows health care professionals to understand the relationships between features and outcomes, making it useful for clinical decision-making. Random forest was selected for its ability to handle complex feature interactions and its robustness in managing missing data, making it effective for capturing nonlinear patterns in the dataset [23]. XGBoost was included for its high predictive performance and efficiency, particularly with structured datasets, and its ability to handle imbalanced data effectively through boosting techniques.

For logistic regression, we used L2 regularization [24] to prevent overfitting and applied polynomial features (adding interaction terms between features) to capture potential nonlinear relationships in the data. For random forest, we optimized the number of trees, maximum depth, and minimum samples per leaf to ensure robustness. We used the same feature set for all models, including the polynomial features, to ensure a fair comparison. For XGBoost, we tuned hyperparameters such as learning rate, maximum depth, and regularization parameters through grid search using cross-validation [25].

Model Calibration

To assess the reliability of each model's predicted probabilities in a clinical context, we conducted both quantitative and visual calibration analyses on the held-out test set. We computed the Brier score (using scikit-learn's [Google Summer of Code] brier_score_loss function), which measures the mean squared difference between predicted probabilities and observed outcomes (lower=better calibration). We also generated calibration (reliability) plots via scikit-learn's calibration_curve function, dividing predictions into 10 equal-width bins, and for each bin, we plotted the mean predicted probability against the observed SSI rate, overlaying the 45° line to represent perfect calibration. All calibration calculations were conducted in Python (version 3.8; Python Software Foundation) using scikit-learn (version 1.0).

Model Evaluation

Because SSIs represent only approximately 4% of cases, overall accuracy can be misleading. Therefore, we report precision, recall (sensitivity), F_1-score, and AUC-ROC as our primary performance metrics for the minority class, with accuracy included only for completeness.

To assess performance stability, we performed 5-fold stratified cross-validation on the training set and report each metric's mean and SD.

Performance metrics, including precision, sensitivity (recall), specificity, F_1-score, and AUC-ROC, were calculated to comprehensively evaluate model performance, especially given the imbalanced nature of the dataset. Sensitivity and specificity help assess the model's ability to correctly identify positive and negative cases, respectively, whereas the AUC-ROC offers an overall measure of discriminative performance. The final models were validated on the reserved 20% of the dataset to assess their generalizability.

Results

Study Population and Characteristics

A total of 1508 patients met the inclusion criteria (n=66, 4.4% with SSIs and n=1442, 95.6% without SSIs). Table 2 presents the comprehensive characteristics of the study population comparing patients with and without SSIs. The table is organized into demographic, clinical, surgical, and medication-related variables to facilitate interpretation (Multimedia Appendix 1).

Demographically, patients with SSIs were significantly older (mean age 61.1, SD 8.9 years, *P*=.01) than those without SSIs (mean age 58.5, SD 15.4 years), with most SSI cases (38/66, 58%) occurring in the category of 61 to 80 years. A higher proportion of female individuals developed SSIs (43/66, 65% vs 765/1442, 53.1% in the non-SSI group), although this difference did not reach statistical significance (*P*=.07).

The ASA score [26] differed significantly between groups (*P*<.001), with a substantially higher percentage of patients with SSIs having an ASA score of 3 (52/66, 79% vs 653/1442, 45.3%), indicating greater preoperative risk and comorbidity burden. Patients who developed SSIs showed a higher prevalence of several comorbidities, including diabetes (17/66, 26% vs 265/1442, 18.4%; *P*=.18), depression (17/66, 26% vs 296/1442, 20.5%; *P*=.18), anemia (24/66, 36% vs 423/1442, 29.3%; *P*=.28), hypertension (36/66, 55% vs 655/1442, 45.4%; *P*=.18), and chronic kidney disease (9/66, 14% vs 123/1442, 8.5%; *P*=.23), although these differences individually did not reach statistical significance.

Surgical and procedural characteristics revealed important differences. Patients who developed SSIs underwent longer surgeries (mean duration 262.5, SD 137.3 min vs 229.9, SD 103.6 min), had a significantly higher number of procedures (mean 1.7, SD 1.1 vs 1.4, SD 0.7; *P*=.02), and experienced substantially longer hospital stays (mean 8.1, SD 6.8 days vs 6.3, SD 10.5 days; *P*<.001). Wound classification showed significant variation, with contaminated wounds being more common in the SSI group (27/66, 41% vs 397/1442, 27.5%; *P*=.03) and clean-contaminated wounds being less common (24/66, 36% vs 799/1442, 55.4%; *P*=.004).

XSL•FO
RenderX

Laboratory and medication factors also demonstrated significant associations with SSI development. Abnormal white blood cell (WBC) counts were significantly more common in patients with SSIs (51/66, 77% vs 734/1442, 50.9%; $P<.001$), defined as WBC counts of >11 before surgery. Steroid use, categorized as receiving any steroid within 12 months before surgery, was significantly higher in the SSI group (55/66, 83% vs 846/1442, 58.7%; $P<.001$). All patients who developed SSIs had received antibiotics.

Table 2. Bivariate comparisons of patient, operative, laboratory, and medication characteristics by 30-day National Healthcare Safety Network–defined surgical site infection (SSI) status in a retrospective single-center cohort of patients undergoing colon surgery at the University of Massachusetts Memorial Medical Center (Worcester, Massachusetts, United States) from 2018 to 2023. Only variables with $P<.05$ are shown.

Variable	No SSI (n=1442)	SSI (n=66)	*P* value
Age (y), mean (SD)	58.5 (15.4)	61.1 (8.9)	.01
ASA[a] score of 3, n (%)	653 (45.3)	52 (78.8)	<.001
Contaminated wound, n (%)	397 (27.5)	27 (40.9)	.03
Length of stay (d), mean (SD)	6.3 (10.5)	8.1 (6.8)	<.001
WBC[b] flag—high, n (%)	734 (50.9)	51 (77.3)	<.001
Steroid use, n (%)	846 (58.7)	55 (83.3)	<.001

[a]ASA: American Society of Anesthesiologists.

[b]WBC: white blood cell.

Imbalance Handling and Ablation Study Results

To evaluate the effectiveness of different approaches for handling class imbalance in our dataset, we performed an ablation study that compares three strategies: cost-sensitive learning, random oversampling, and a combination of both techniques. Each strategy was applied to all three ML models (logistic regression, random forest, and XGBoost) and evaluated on the same held-out set to ensure fair comparison. The results of this ablation study are presented in Table 3, which shows the performance metrics for each model and strategy combination.

Table 3. Ablation study comparing class imbalance–handling strategies—cost-sensitive learning, random oversampling, and their combination—for predicting 30-day National Healthcare Safety Network–defined surgical site infections after colon surgery in a retrospective single-center cohort (University of Massachusetts Memorial Medical Center, Worcester, Massachusetts, United States; 2018-2023).

Model and strategy	Precision (%)	Recall (%)	F_1-score	AUC-ROC[a]	Brier score
Logistic regression					
Cost-sensitive learning	9.84	46.15	0.162	0.709	0.143
Oversampling only	10.34	46.15	0.169	0.704	0.140
Oversampling+cost-sensitive learning	10.34	46.15	0.169	0.704	0.140
Random forest					
Cost-sensitive learning	100	15.38	0.267	0.776	0.032
Oversampling only	100	15.38	0.267	0.777	0.031
Oversampling+cost-sensitive learning	100	15.38	0.267	0.777	0.031
XGBoost[b]					
Cost-sensitive learning	57.14	30.77	0.400	0.719	0.036
Oversampling only	41.67	38.46	0.400	0.758	0.043
Oversampling+cost-sensitive learning	41.67	38.46	0.400	0.758	0.043

[a]AUC-ROC: area under the receiver operating characteristic curve.

[b]XGBoost: Extreme Gradient Boosting.

Model Performance and Calibration

Table 4 summarizes each model's precision, recall, F_1-score, and AUC-ROC for SSI detection as means and SDs from 5-fold cross-validation alongside Brier scores on the held-out test set. Accuracy was high for all models but less informative given the 4.4% (66/1508) SSI rate. Confusion matirces for all three models on the held-out test are shown in Figure 1.

In terms of discrimination, XGBoost achieved the highest AUC-ROC (0.788), followed by random forest (0.778) and logistic regression (0.706; Figure 2). Regarding calibration, random forest and XGBoost both exhibited low Brier scores (0.034 and 0.035, respectively), indicating well-calibrated

XSL•FO
RenderX

probability estimates, whereas logistic regression's higher Brier score (0.139) reflects moderate miscalibration (Figure 3).

Among the 3 models, XGBoost demonstrated the highest AUC-ROC score (0.788) and F_1-score (0.43), suggesting the best overall discriminative ability and balance between precision and recall. The random forest model showed perfect precision but lower recall, indicating that it was highly conservative in predicting SSIs. Logistic regression had the highest recall but the lowest precision, suggesting that it was more liberal in flagging potential SSI cases but at the cost of many false positives.

Table 4. Model performance for predicting 30-day National Healthcare Safety Network–defined surgical site infections after colon surgery in a retrospective single-center cohort (University of Massachusetts Memorial Medical Center, Worcester, Massachusetts, United States; 2018-2023)[a].

Model	Precision (%), mean (SD)	Recall (%), mean (SD)	F_1-score, mean (SD)	AUC-ROC[b], mean (SD)	Brier score
Logistic regression	10.7 (3.9)	56.0 (18.5)	0.18 (0.06)	0.775 (0.059)	0.139
Random forest	80.0 (40.0)	11.1 (6.8)	0.19 (0.11)	0.756 (0.046)	0.034
XGBoost[c]	40.0 (17.0)	20.5 (8.6)	0.27 (0.11)	0.735 (0.059)	0.035

[a]Five-fold stratified cross-validation results are shown; Brier scores were computed on the 20% held-out test set.

[b]AUC-ROC: area under the receiver operating characteristic curve.

[c]XGBoost: Extreme Gradient Boosting.

Figure 1. Confusion matrices on the 20% held-out test set for the logistic regression, random forest, and Extreme Gradient Boosting (XGBoost) models predicting 30 day National Healthcare Safety Network–defined surgical site infections (SSIs) after colon surgery in a retrospective single-center cohort (University of Massachusetts Memorial Medical Center, Worcester, Massachusetts, United States; 2018-2023; N=1508 with 66/1508, 4.4% SSIs).

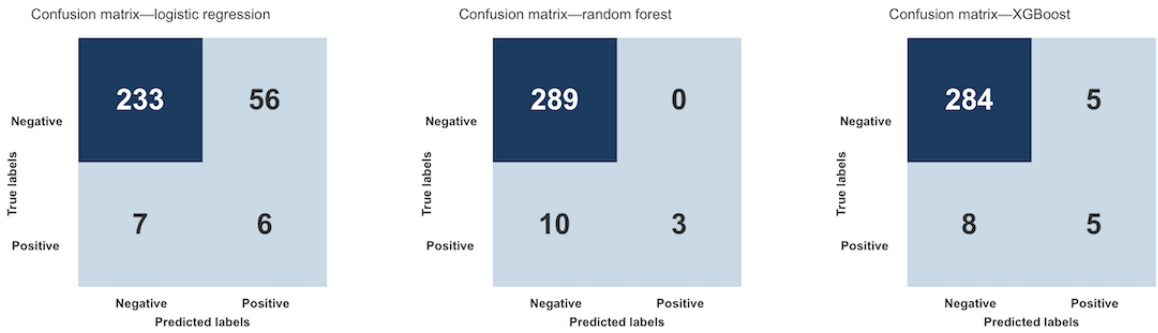

XSL•FO
RenderX

Figure 2. Receiver operating characteristic (ROC) curves on the 20% held-out test set for the logistic regression, random forest, and Extreme Gradient Boosting (XGBoost) models trained to predict 30-day National Healthcare Safety Network–defined surgical site infections following colon surgery in a retrospective single center cohort (University of Massachusetts Memorial Medical Center, Worcester, Massachusetts, United States; 2018-2023). Area under the ROC curve values summarize discrimination. AUC: area under the curve.

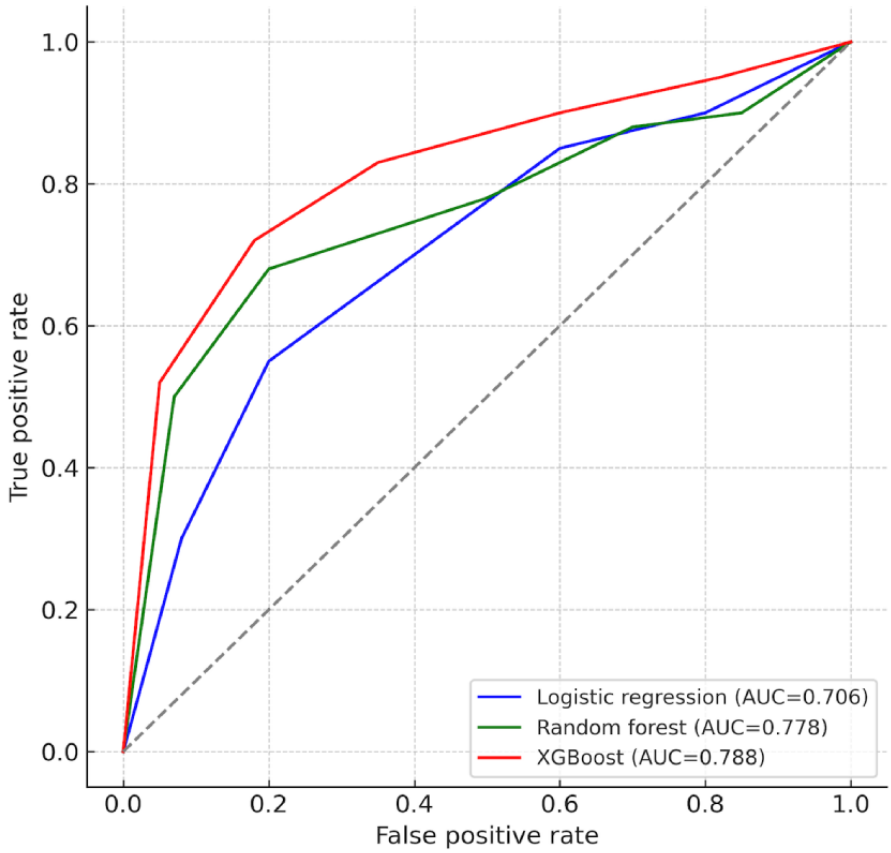

Figure 3. Calibration (reliability) curves on the 20% held-out test set for logistic regression, random forest, and Extreme Gradient Boosting (XGBoost) predicting 30-day National Healthcare Safety Network–defined surgical site infections after colon surgery in a retrospective single-center cohort (University of Massachusetts Memorial Medical Center, Worcester, Massachusetts, United States; 2018-2023). Predictions were binned into 10 equal-width bins; the dashed 45° line indicates perfect calibration.

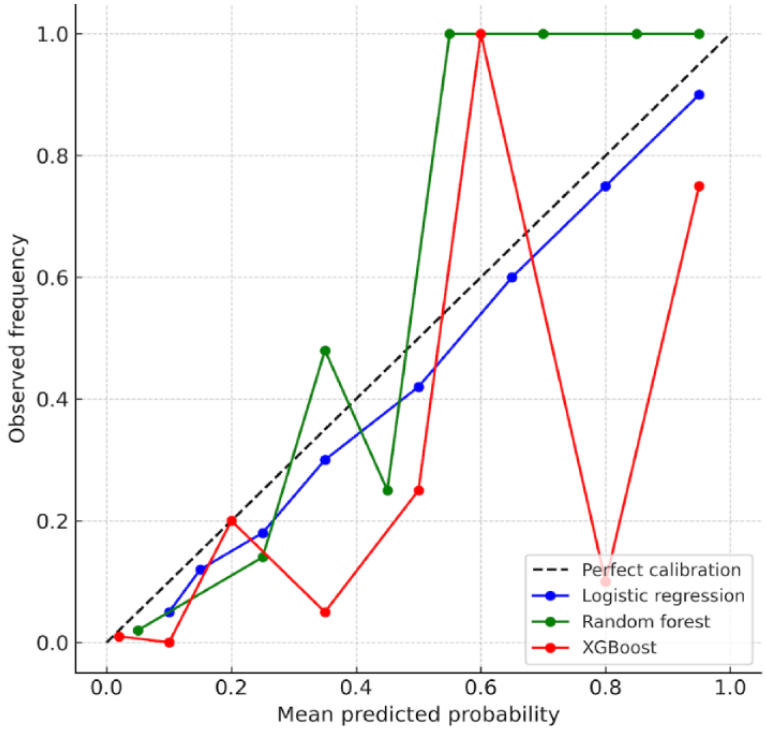

XSL•FO
RenderX

Statistical Comparison of Area Under the Curve

To formally assess whether observed area under the curve (AUC) differences were significant, we conducted a paired bootstrap analysis with 1000 resamples of the test set. The 95% CIs and *P* values are shown in Table 5. All intervals overlapped, and both pairwise *P* values exceeded .05, indicating no statistically significant differences between XGBoost and the other models.

Because all intervals overlapped substantially and both *P* values exceeded .05, we conclude that the small observed AUC differences were not statistically significant. These results support our interpretation that XGBoost and random forest performed equivalently in discriminating SSI risk and that XGBoost's edge over logistic regression did not reach significance under bootstrap testing.

Table 5. Statistical comparison of model areas under the curve (AUCs) via paired bootstrap (1000 resamples) on the 20% held-out test set in a retrospective single-center study predicting 30-day National Healthcare Safety Network–defined surgical site infections after colon surgery (University of Massachusetts Memorial Medical Center, Worcester, Massachusetts, United States; 2018-2023).

	AUC (95% CI)	*P* value
LR[a]	0.706 (0.538-0.850)	—[b]
RF[c]	0.778 (0.588-0.935)	—
XGB[d]	0.788 (0.607-0.933)	—
XGB vs LR	—	.50
XGB vs RF	—	.79

[a]LR: logistic regression.

[b]Not applicable.

[c]RF: random forest.

[d]XGB: Extreme Gradient Boosting.

Feature Importance (Shapley Additive Explanations Analysis)

Table 6 lists the top 10 features by mean absolute Shapley additive explanations (SHAP) value for the XGBoost model, demonstrating the influence of both structured EHR variables and our NLP-derived metrics on SSI risk prediction.

Feature importance analysis revealed that the most predictive factors for SSI risk included ASA score, patient age, wound classification (particularly contaminated wounds), steroid use, and laboratory indicators such as WBC count. The NLP-derived features (SSI keyword count and negation term count) also contributed significantly to the models' predictive performance, highlighting the value of incorporating unstructured clinical notes in SSI risk assessment. Notably, ssi_keyword (rank 2) and ssi_negated (rank 7) were among the top predictors—on par with established clinical features such as age and ASA score.

Table 6. Top 10 predictors of 30-day National Healthcare Safety Network–defined surgical site infections (SSIs) after colon surgery based on mean absolute Shapley additive explanations (SHAP) values from the final Extreme Gradient Boosting model trained on the retrospective single-center cohort (University of Massachusetts Memorial Medical Center, Worcester, Massachusetts, United States; 2018-2023)a.

Rank	Feature	SHAP value, mean (SD)
1	surgery_minutes_in_recovery	1.1761
2	ssi_keyword	1.1217
3	surgery_patient_age	1.1188
4	asa_score	0.9354
5	surgery_minutes_in_or	0.7716
6	average_bmi	0.7397
7	ssi_negated	0.7311
8	surgery_physician_id_E8779	0.4223
9	length_of_stay	0.3408
10	patient_gender_female	0.3100

[a]Predictors include structured electronic health record features (eg, American Society of Anesthesiologists score, operative times, BMI, and length of stay) and natural language processing–derived features from clinical notes (SSI keyword and negation counts).

XSL•FO
RenderX

Discussion

Principal Findings

This study aimed to develop and evaluate ML models for the early prediction of SSIs following colon surgery using both structured EHR data and unstructured clinical notes. Our main findings show that ML models, particularly XGBoost, can effectively augment traditional surveillance practices by providing well-calibrated, discriminative risk predictions that prioritize patients at high risk. Among the models tested, XGBoost demonstrated the best balance between precision and recall (AUC-ROC=0.788; precision=50%; recall=38%), whereas random forest achieved perfect precision at the cost of low sensitivity (recall=23%).

A more detailed analysis reveals that each model has specific strengths that may suit different clinical priorities. The XGBoost model provides a practical compromise between sensitivity and specificity, which is ideal for resource-limited infection control teams aiming to balance workload and risk. Random forest's high precision makes it suitable for contexts in which false positives must be minimized, whereas logistic regression offers higher recall but risks overburdening staff due to its lower precision. Our findings are consistent with those of previous literature in emphasizing the importance of model calibration for real-world implementation; the low Brier scores for both XGBoost and random forest suggest reliable probability estimates that can guide clinical triage. Furthermore, the integration of NLP-derived features—such as SSI keyword frequency and negation detection—significantly improved predictive performance, with these variables ranking among the top 10 predictors by SHAP value. This supports previous work [27,28] on the added value of unstructured clinical data in infection surveillance, but our study differs by using these data prospectively—leveraging information available before infection onset—rather than relying on retrospective documentation after SSIs have already occurred.

Previous studies have also demonstrated strong performance in SSI identification using ML but with important methodological differences. One study using NLP achieved a sensitivity and positive predictive value of 97% for SSI detection [29] but relied on postinfection clinical notes, limiting its utility for early intervention. Another study applied logistic regression and tree-based models to preoperative blood test results, achieving an AUC of 86% [30], although it used a more balanced dataset and focused narrowly on laboratory values. In contrast, our approach emphasizes prospective prediction using only preinfection data and incorporates a broader set of features—including operative and demographic characteristics and unstructured clinical text—to support earlier, more comprehensive risk assessment and real-time clinical deployment. This multimodal integration increases the generalizability and interpretability of our model, making it better suited for prospective deployment within existing clinical workflows.

To support clinical adoption, we propose embedding the model's SSI risk scores directly into the EHR's surveillance dashboard. Each morning, the model generates risk scores for all patients undergoing colon surgery, automatically flagging those above a configurable threshold (eg, the top 10%). Flagged patients appear on an IPC nurse worklist for prioritized chart review, focusing on vital signs, wound assessments, and microbiology. The threshold can be adjusted based on operational capacity—for example, flagging the top 5% during busy periods or expanding to 15% during lower volume. Nurse adjudications (SSI vs no SSI) are fed back into the system to support model retraining and recalibration over time. This workflow focuses human effort on high-risk cases, enhances surveillance efficiency, and remains adaptable to fluctuating staffing or patient volumes.

Limitations

Our evaluation showed that SMOTE was more effective than cost-sensitive learning in improving model performance for the minority class, increasing recall by approximately 15% while maintaining similar precision. However, both approaches still exhibited some bias toward the majority class, particularly the random forest model. The severe class imbalance in our dataset (only 66/1508, 4.4% were SSI cases) presents a significant challenge for model development and evaluation. Although we used techniques to address this imbalance, the models' performance in identifying positive cases remained suboptimal, as evidenced by the modest recall values. Future work could explore advanced synthetic generation methods [31] to address severe class imbalance.

Due to the severe class imbalance, feature importance scores—particularly from tree-based models—may be driven primarily by patterns in the majority (non-SSI) class. Although we applied both SMOTE and cost-sensitive learning to mitigate this imbalance before model training, readers should interpret importance rankings with caution. In future work, we plan to explore class-specific importance measures (eg, SHAP values stratified by outcome) to obtain a more balanced view of predictors for the minority class.

In addition, the use of retrospective data from a single health care center may introduce biases related to specific patient populations and clinical practices, potentially limiting the generalizability of the results. The predictive factors identified in our specific setting might not hold the same importance in other health care environments with different patient demographics, surgical practices, or infection control protocols.

Furthermore, our approach to processing unstructured clinical notes was limited to keyword counting and negation detection. More sophisticated NLP techniques such as embedding-based methods or transformer models might yield better feature extraction and, ultimately, enhance predictive performance [32-34].

Conclusions

This study demonstrates that ML models can enhance SSI surveillance by helping clinicians prioritize patients at high risk. Our ML-based tool, which integrates structured EHR data and unstructured clinical notes, offers a scalable approach to improve monitoring efficiency.

XSL•FO
RenderX

Among the models evaluated, XGBoost provided the best balance of precision, recall, and calibration, although each model presents unique strengths suited to different clinical needs. By triaging patients based on predicted risk, the tool can reduce manual workload and support more timely, targeted interventions to improve patient outcomes.

While this pilot study shows proof of concept, broader validation is needed to ensure generalizability and clinical utility. We plan to partner with other institutions using the Observational Medical Outcomes Partnership Common Data Model and federated learning to enable privacy-preserving, cross-site model training. Performance will be assessed through calibration, discrimination, and operational impact metrics.

Beyond technical refinement, this work underscores the potential of artificial intelligence–driven tools to transform infection surveillance from reactive monitoring to proactive, risk-based care. Future efforts should focus on handling class imbalance, improving NLP feature extraction, and ensuring model reliability through post hoc calibration and multisite validation.

Ultimately, this research advances the responsible and scalable integration of artificial intelligence into clinical workflows to support more targeted, efficient infection prevention.

Acknowledgments

The authors thank the infection prevention and control staff at the University of Massachusetts Memorial Medical Center (Melissa Dabkowski, and Maria Del Pilar Diaz Padro) for their invaluable support and collaboration, as well as the University of Massachusetts Chan Medical School IT department for their essential technical assistance. Their contributions were vital to the success of this project. The grant is funded by The Doctors Company Foundation and the University of Massachusetts Center for Clinical and Translational Science Pilot Project Program (UL1-TR001453).

Data Availability

The datasets generated or analyzed during this study are not publicly available due to usage of individual-level electronic health record data that includes protected health information such as structured tables and free-text clinical notes and, therefore, cannot be publicly shared under US HIPAA (Health Insurance Portability and Accountability Act) regulations and the terms of our institutional review board approval. Deidentified, aggregate summary outputs underlying the manuscript (eg, model performance tables) are provided in Multimedia Appendix 1.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed cohort characteristics and model performance metrics, including confusion matrices and receiver operating characteristic curve plots for each machine learning model.
[DOCX File , 21 KB - formative_v9i1e75121_app1.docx]

References

1. Battles JB, Farr SL, Weinberg DA. From research to nationwide implementation: the impact of AHRQ's HAI prevention program. Med Care 2014 Feb;52(2 Suppl 1):S91-S96. [doi: 10.1097/MLR.0000000000000037] [Medline: 24430273]
2. Awad SS. Adherence to surgical care improvement project measures and post-operative surgical site infections. Surg Infect (Larchmt) 2012 Aug;13(4):234-237. [doi: 10.1089/sur.2012.131] [Medline: 22913334]
3. Merkow RP, Ju MH, Chung JW, Hall BL, Cohen ME, Williams MV, et al. Underlying reasons associated with hospital readmission following surgery in the United States. JAMA 2015 Feb 03;313(5):483-495. [doi: 10.1001/jama.2014.18614] [Medline: 25647204]
4. de Lissovoy G, Fraeman K, Hutchins V, Murphy D, Song D, Vaughn BB. Surgical site infection: incidence and impact on hospital utilization and treatment costs. Am J Infect Control 2009 Jun;37(5):387-397. [doi: 10.1016/j.ajic.2008.12.010] [Medline: 19398246]
5. Sparling KW, Ryckman FC, Schoettker PJ, Byczkowski TL, Helpling A, Mandel K, et al. Financial impact of failing to prevent surgical site infections. Qual Manag Health Care 2007;16(3):219-225. [doi: 10.1097/01.QMH.0000281058.99929.ea] [Medline: 17627217]
6. Magill SS, Edwards JR, Bamberg W, Beldavs ZG, Dumyati G, Kainer MA, et al. Multistate point-prevalence survey of health care-associated infections. N Engl J Med 2014 Mar 27;370(13):1198-1208 [FREE Full text] [doi: 10.1056/NEJMoa1306801] [Medline: 24670166]
7. Anderson DJ, Podgorny K, Berríos-Torres SI, Bratzler DW, Dellinger EP, Greene L, et al. Strategies to prevent surgical site infections in acute care hospitals: 2014 update. Infect Control Hosp Epidemiol 2014 Jun;35(6):605-627 [FREE Full text] [doi: 10.1086/676022] [Medline: 24799638]

XSL•FO
RenderX

8. Mangram AJ, Horan TC, Pearson ML, Silver LC, Jarvis WR. Guideline for prevention of surgical site infection, 1999. Hospital Infection Control Practices Advisory Committee. Infect Control Hosp Epidemiol 1999 Apr;20(4):250-78; quiz 279-80. [doi: 10.1086/501620] [Medline: 10219875]
9. Surgical site infection event (SSI). Centers for Disease Control and Prevention. URL: https://www.cdc.gov/nhsn/pdfs/pscmanual/9pscssicurrent.pdf [accessed 2024-01-15]
10. Sanger PC, van Ramshorst GH, Mercan E, Huang S, Hartzler AL, Armstrong CA, et al. A prognostic model of surgical site infection using daily clinical wound assessment. J Am Coll Surg 2016 Aug;223(2):259-70.e2 [FREE Full text] [doi: 10.1016/j.jamcollsurg.2016.04.046] [Medline: 27188832]
11. Meeks DW, Lally KP, Carrick MM, Lew DF, Thomas EJ, Doyle PD, et al. Compliance with guidelines to prevent surgical site infections: as simple as 1-2-3? Am J Surg 2011 Jan;201(1):76-83. [doi: 10.1016/j.amjsurg.2009.07.050] [Medline: 20573335]
12. Ridgeway S, Wilson J, Charlet A, Kafatos G, Pearson A, Coello R. Infection of the surgical site after arthroplasty of the hip. J Bone Joint Surg Br 2005 Jun;87(6):844-850. [doi: 10.1302/0301-620X.87B6.15121] [Medline: 15911671]
13. Koek MB, Hopmans TE, Soetens LC, Wille JC, Geerlings SE, Vos MC, et al. Adhering to a national surgical care bundle reduces the risk of surgical site infections. PLoS One 2017 Sep 06;12(9):e0184200 [FREE Full text] [doi: 10.1371/journal.pone.0184200] [Medline: 28877223]
14. Honnibal M. spaCy 2: natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. Sentometrics Research. 2017 Jan. URL: https://sentometrics-research.com/publication/72/ [accessed 2025-08-29]
15. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform 2001 Oct;34(5):301-310 [FREE Full text] [doi: 10.1006/jbin.2001.1029] [Medline: 12123149]
16. McKinney W. Data structures for statistical computing in python. In: Proceedings of the 9th Python in Science Conference. 2010 Presented at: SciPy 2010; June 28-July 3, 2010; Austin, TX. [doi: 10.25080/majora-92bf1922-00a]
17. Zai AH, Wong S, Guilarte-Walker Y, Langlois P, Coleman B, Soni A, et al. Implementing virtual desktops for clinical research at an academic health center: a case report. JAMIA Open 2024 Aug 28;7(3):ooae083. [doi: 10.1093/jamiaopen/ooae083] [Medline: 39206281]
18. Elkan C. The foundations of cost-sensitive learning. In: Proceedings of the 17th International Joint Conference on Artificial Intelligence. 2001 Presented at: IJCAI'01; August 4-10, 2001; Seattle, WA.
19. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002 Jun 01;16:321-357. [doi: 10.1613/JAIR.953]
20. Hosmer DW, Lemeshow S, Sturdivant RX. Applied Logistic Regression. Hoboken, NJ: Wiley; 2013.
21. Breiman L. Random forests. Mach Learn 2001 Oct;45:5-32. [doi: 10.1023/a:1010933404324]
22. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016 Presented at: KDD '16; August 13-17, 2016; San Francisco, CA. [doi: 10.1145/2939672.2939785]
23. Dietterich TG. Ensemble methods in machine learning. In: Proceedings of the First International Workshop on Multiple Classifier Systems. 2000 Presented at: MCS 2000; June 21-23, 2000; Cagliari, Italy. [doi: 10.1007/3-540-45014-9_1]
24. Ng AY. Feature selection, L1 vs. L2 regularization, and rotational invariance. In: Proceedings of the Twenty-First International Conference on Machine Learning. 2004 Presented at: ICML '04; July 4-8, 2004; Banff, AB.
25. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. J Mach Learn Res 2012;13:281-305.
26. Doyle D, Garmon E. American Society of Anesthesiologists Classification (ASA Class). StatPearls. Treasure Island (FL): StatPearls Publishing; 2025. URL: https://www.ncbi.nlm.nih.gov/books/NBK441940/ [accessed 2025-05-29]
27. Shen F, Larson DW, Naessens JM, Habermann EB, Liu H, Sohn S. Detection of surgical site infection utilizing automated feature generation in clinical notes. J Healthc Inform Res 2019 Sep;3(3):267-282 [FREE Full text] [doi: 10.1007/s41666-018-0042-9] [Medline: 31728432]
28. da Silva DA, Ten Caten CS, Dos Santos RP, Fogliatto FS, Hsuan J. Predicting the occurrence of surgical site infections using text mining and machine learning. PLoS One 2019 Dec 13;14(12):e0226272 [FREE Full text] [doi: 10.1371/journal.pone.0226272] [Medline: 31834905]
29. Thirukumaran CP, Zaman A, Rubery PT, Calabria C, Li Y, Ricciardi BF, et al. Natural language processing for the identification of surgical site infections in orthopaedics. J Bone Joint Surg Am 2019 Dec 18;101(24):2167-2174 [FREE Full text] [doi: 10.2106/JBJS.19.00661] [Medline: 31596819]
30. Mandagani P, Coleman S, Zahid A, Pugel Ehlers A, Basu Roy S, De Cock M. Machine learning models for surgical site infection prediction. In: Proceedings of the American Medical Informatics Association Knowledge Discovery and Data Mining Working Group Symposium. 2016 Presented at: AMIA KDDMWG 2016; November 12-16, 2016; Chicago, IL.
31. Park N, Mohammadi M, Gorde K, Jajodia S, Park H, Kim Y. Data synthesis based on generative adversarial networks. Proc VLDB Endow 2018 Jun;11(10):1071-1083. [doi: 10.14778/3231751.3231757]
32. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. ArXiv Preprint posted online on October 11, 2018. [doi: 10.48550/arXiv.1810.04805]

XSL•FO
RenderX

33. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb 15;36(4):1234-1240 [FREE Full text] [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
34. Alsentzer E, Murphy J, Boag W, Weng WH, Jindi D, Naumann T, et al. Publicly available clinical BERT embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop. 2019 Presented at: ClinicalNLP 2019; June 7, 2019; Minneapolis, MN. [doi: 10.18653/v1/w19-1909]

Abbreviations

ASA: American Society of Anesthesiologists
AUC: area under the curve
AUC-ROC: area under the receiver operating characteristic curve
CDC: Centers for Disease Control and Prevention
CPT: Current Procedural Terminology
EHR: electronic health record
HIPAA: Health Insurance Portability and Accountability Act
IPC: infection prevention and control
ML: machine learning
NHSN: National Healthcare Safety Network
NLP: natural language processing
SHAP: Shapley additive explanations
SMOTE: synthetic minority oversampling technique
SSI: surgical site infection
UMass Chan: University of Massachusetts Chan Medical School
WBC: white blood cell
XGBoost: Extreme Gradient Boosting

Edited by A Mavragani; submitted 01.04.25; peer-reviewed by RT Potla, D Pant; comments to author 28.05.25; revised version received 18.06.25; accepted 19.08.25; published 01.10.25.

Please cite as:
Celik U, Liu F, Kobayashi K, Ellison III RT, Guilarte-Walker Y, Mack DA, Shi Q, Zai A
Machine Learning–Enhanced Surveillance for Surgical Site Infections in Patients Undergoing Colon Surgery: Model Development and Evaluation Study
JMIR Form Res 2025;9:e75121
URL: https://formative.jmir.org/2025/1/e75121
doi:10.2196/75121
PMID:41032359

©Ugur Celik, Feifan Liu, Kimiyoshi Kobayashi, Richard T Ellison III, Yurima Guilarte-Walker, Deborah Ann Mack, Qiming Shi, Adrian Zai. Originally published in JMIR Formative Research (https://formative.jmir.org), 01.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Applying the Human-Centered Innovation Biodesign Framework to the Development and Piloting of a Program to Mitigate Risk for Cognitive Decline Among Historically Underrepresented Individuals: Case Study

Rebecca Lassell[1,2], OTR, PhD; Ada Metaxas[3]; Katherine Wang[4], BA; Sara Hantgan[5]; Prabhat Gottipati[6], BA; Sarah Zwerling[7]; Triana Pena[1], BSN; Chava Pollak[8], PhD, RN; Laura Gitlin[9], PhD; Sunit Jariwala[10,11], MD

[1]Department of Health and Wellness Design, School of Public Health, Indiana University, Bloomington, IN, United States
[2]Regenstrief Institute, Indiana University Center for Aging Research, Indianapolis, IN, United States
[3]Department of Chemistry, Princeton University, Princeton, NJ, United States
[4]Albert Einstein College of Medicine, Bronx, NY, United States
[5]Michigan Medicine Department of Computational Medicine and Bioinformatics, University of Michigan School of Literature, Science, and the Arts, University of Michigan School of Information, Ann Arbor, MI, United States
[6]Lewis Katz School of Medicine, Temple University, Philadelphia, PA, United States
[7]College of Human Ecology and majoring in Human Development, Cornell University, Ithaca, NY, United States
[8]Department of Neurology, Stony Brook Medicine, Long Island, NY, United States
[9]College of Nursing and Health Professions, Drexel University, Philadelphia, PA, United States
[10]Department of Medicine, Albert Einstein College of Medicine, Bronx, NY, United States
[11]Division of Allergy/Immunology, Albert Einstein College of Medicine, Bronx, NY, United States

Corresponding Author:
Rebecca Lassell, OTR, PhD
Department of Health and Wellness Design
School of Public Health
Indiana University
1025 East Seventh Street
Bloomington, IN, 47405-7000
United States
Phone: 1 8128552395
Email: blassell@iu.edu

Abstract

Background: Physical inactivity is a modifiable risk factor for dementia. Past physical activity interventions often overlook the voices of the end user in the design process, particularly minoritized groups living with dementia or memory challenges. To develop physical activity interventions, we use the principles of human-centered design.

Objective: We applied human-centered design using the innovation biodesign framework to develop a physical activity intervention, Nurturing Aging Through Uplifting Activities in a Restorative Environment (NATURE) program for minoritized individuals as a use case.

Methods: The innovation biodesign framework has three domains: (1) problem space, (2) invention, and (3) solution space. Each domain includes several activities. The problem space involves a needs assessment, needs screening, evidence-based literature review, review of existing models of programs, and iterative feedback from partners, leading to an invention. The solution space encompasses the implementation and validation of the invention and outcomes. We applied this framework in 3 steps: (1) identifying the problem: we used data points from multiple sources to identify needs and mapped them onto the problem space. These sources included reviews of the literature to identify existing interventions, findings from other nature programs to surmise gaps, and focus groups to iteratively identify unmet needs. (2) Designing the invention: we developed NATURE with Hispanic or Latino people with memory challenges and identified their preferred outcomes. (3) Mapping the pilot study. We added the study protocol and planned outcomes to the solution space.

XSL•FO
RenderX

Results: In step 1, three evidence-based programs guided the development of NATURE to address physical inactivity and related risks of decreased well-being and dementia. We received 50 referrals for focus group participants, 22 were eligible and completed consent, and 21 (n=6 Hispanic or Latino people with memory challenges and care partners, n=8 outdoor professionals, and n=7 health care providers) participants completed the focus groups. We received feedback from participants on local nature activities, program frequency, duration, and delivery mode, a referral pathway, and outcomes using 5 focus groups and 2 interviews. In step 2, the 12-week NATURE program was developed to promote an active lifestyle and well-being, using nature activities that a person enjoys. NATURE accounts for a person's preferences, needs, and daily situation and includes 4-6 sessions with 2 phone check-ins. Preferred outcomes were well-being, sleep, and social connections. In step 3, we mapped the plan to pilot NATURE using activity tracker technology to measure sleep, heart rate, and activity (well-being), and validated questionnaires.

Conclusions: The framework provided a systematic approach for mapping the development of NATURE to address the needs of Hispanic or Latino people with memory challenges, using human-centered design principles. Application of the framework can be a helpful tool to map the development of other interventions for minoritized populations.

Trial Registration: ClinicalTrials.gov NCT06403345; https://clinicaltrials.gov/study/NCT06403345

(JMIR Form Res 2025;9:e64930) doi:10.2196/64930

KEYWORDS

user-centered design; medical innovation; Alzheimer's disease; dementia; lifestyle intervention; nature-based; innovation; human-centered; biodesign; cognitive; cognition; cognitive decline; Alzheimer's; gaps; human-centered design; physical activity

Introduction

Background

Reducing sedentary behavior and physical inactivity has been identified as an important modifiable approach to reduce risk for dementia [1,2]. Physical inactivity is 30% higher for older adults, people with disabilities, including people living with mild cognitive impairment (MCI) and dementia. Additionally, 31% of individuals who identify as Hispanic or Latino do not meet physical activity guidelines [3] and are 1.5 times more likely than non-Hispanic White individuals to develop dementia [4]. Hispanic or Latino individuals may face barriers to physical activity, such as a lack of sidewalks in their neighborhoods, unsafe streets and parks, and a lack of time and social support [3].

Although an active lifestyle is a key modifiable factor to lessen risk for developing dementia [5], Hispanic or Latino individuals are underrepresented in research to design and test physical activity interventions [6,7]. A promising approach is developing interventions that tap into the goals, preferences, and values of underrepresented populations is the use of physical activity in the context of nature.

A concern with existing physical activity interventions is that they are difficult to sustain [8,9], and often lack input from the end user in the development process. Most of these interventions have not been rooted in a human-centered approach that can iteratively account for the values and preferences of end users [10]. This is particularly the case for underrepresented populations such as Hispanic or Latino groups, who are estimated to rapidly increase in dementia incidence from 430,000 in 2012 to 2,383,000 in 2050 [4]. The purpose of this paper is to show the utility of the human-centered innovation biodesign (IB) framework [11] to develop and test interventions. Specifically, we apply the framework to develop and pilot the Nurturing Aging Through Uplifting Activities in a Restorative Environment (NATURE) program for Hispanic or Latino groups.

NATURE Program

The NATURE program, formerly called the Green Activity Program, was co-designed with individuals who identified as Hispanic or Latino with MCI as a precursor to Alzheimer disease (AD) and early-stage dementia, called memory challenges herein, based on participants' preferences [12]. NATURE promotes an active lifestyle in a social context through the skillful tailoring of local nature activities that a person enjoys and is delivered by occupational therapists. Furthermore, the intervention was designed with input from multiple partners, including local health care providers and nature organizations. Mapping the development of NATURE can provide a clear roadmap for incorporating human-centered design principles in the ongoing pilot study and provide a framework to depict additional refinements for future efficacy and effectiveness testing, and address challenges with sustainability faced by other physical activity interventions [8,9].

The IB Framework

The IB framework uses a human-centered design approach. Human-centered design is an umbrella term for a flexible and replicable approach to innovation that prioritizes human values and experiences when creating and implementing complex systems, services, or products [10]. While there is disagreement about what "human" means, key design principles include placing the human experience at the center of the design, understanding the demands on users when planning tasks, and considering the surrounding environment. Human-centered design involves the end user in every step of the process and applies user-centered evaluation to iteratively refine the design [10,13]. These principles are applied across disciplines and sectors, from developing technology and mobile health to health care programming to promote behavior change. One critique of human-centered design is that it uses broad methods of engagement and can lack systematic processes to ensure rigor and quality. Particularly, end user engagement can occur through observations or more structured co-design with the end user's input through participatory design sessions, town halls, focus

XSL•FO
RenderX

groups, or individual interviews [14]. Using a framework can help mitigate this with a systematic process to map the application of human-centered design principles across the development and testing of an intervention.

Human-centered design principles within the IB framework enable researchers to identify and solve the unmet needs of underrepresented populations [11]. The framework has been applied to unmet clinical needs for asthma in minoritized populations [11] and can help map the development and testing of health care innovations, technologies, and interventions through collaborative partnerships. Collaborations between researchers, health care professionals, and those living with dementia in the design and implementation of an intervention may promote acceptability and applicability in addressing a population's needs [14]. Here, we apply the framework to the development of an intervention to promote physical activity for Hispanic or Latino individuals with low resources.

The IB framework has three domains the (1) problem space, (2) invention, and (3) solution space [11]. Each domain includes activities that may not always occur in a linear process. The problem space involves a needs assessment, needs screening, evidence-based literature review, review of existing models of programs, and iterative feedback from partners, leading to an invention. The solution space comprises of implementation and validation of the invention and assessment of outcomes (planned or actual). The framework enables researchers to identify unmet needs through clinical observation and patient feedback and iteratively design, refine, and test user-centered solutions from development to dissemination.

Accordingly, we sought to apply the IB framework to map the development and piloting of the NATURE program to address the unmet needs of Hispanic or Latino individuals with memory challenges. This paper focuses on mapping the development of NATURE onto the problem space of the framework and uses the solution space to map a plan for an ongoing pilot study.

Methods

Study Design

In this use case, we applied the IB framework to map the development of the NATURE program and ongoing pilot study. Use-case findings are reported in adherence with the guidance for the reporting of intervention development checklist for reporting intervention development studies (Multimedia Appendix 1) [15].

This use case took place in the Bronx, New York; Indianapolis, Indiana; and Burlington, Vermont. NATURE, formerly the Green Activity Program, was initially designed for the Bronx [12]. The intervention was expanded with input from Chinese Americans and Black individuals. The focus of this use case is on the initial development in the Bronx, with Hispanic or Latino groups.

Recruitment

We used convenience sampling to recruit participants with memory challenges from local aging organizations in the Bronx, Indianapolis, and Burlington. We also used direct referrals from 2 geriatric clinics in the Bronx, and Indianapolis. Inclusion criteria for participants from secondary data were previously reported [12]. Due to the current political climate in the United States, we expanded our population to include all races and ethnicities due to safety concerns for Hispanic or Latino individuals. Recruitment efforts were focused on areas with lower resources, with strategies of recruiting through in-person events, newsletters through local aging organizations, and clinical referrals.

Participants in the ongoing pilot study were (1) 45 years or older, (2) had subjective cognitive decline, and (3) had access and ability to respond to the telephone (mobile or landline). Subjective cognitive decline was determined by two questions: (1) Have you noticed difficulty with your memory? (2) Do you think that your memory is worse than 5 years ago? [16] Subjective cognitive decline was also characterized after enrollment with the Telephone Montreal Cognitive Assessment.

Participants were excluded if they had greater than 3 hospitalizations in the past year, or if they indicated any of the following on the 2023 Physical Activity Readiness Questionnaire +: (1) heart failure or difficulty controlling coronary artery disease, diagnosed abnormality of heart rhythm, or other cardiovascular condition; (2) cancer and are actively receiving therapies, (3) experienced a black out, fainted, or lost consciousness as a result of a head injury in the past 12 months, or (4) often experienced signs and symptoms of low blood glucose (hypoglycemia) following exercise or daily activities.

Analysis: Steps to Apply the IB Framework to NATURE

We used a 3-step approach to apply the IB framework to NATURE. Step 1 involved identifying the problem to map onto the problem space. These sources included findings from other nature programs, reviews of the literature to identify existing interventions and surmise gaps, as well as focus groups to identify unmet needs. Focus groups occurred between March 2023 to August 2023. Focus groups included iterative co-design to design programmatic components, such as identifying local nature activities, program frequency, duration, and preferred outcomes, and identifying potential challenges and solutions for Hispanic or Latino individuals with memory challenges. Research design included identifying the preferred language for recruitment materials and referral pathways within a local health care system. In step 2, we describe the developed invention, the NATURE program. We used information from the pilot study protocol and intervention manual to describe NATURE. Step 3 encompassed mapping the ongoing pilot study. We used the protocol and planned data collection, including activity tracker technology, in the solution space to create a plan for an ongoing feasibility pilot study (ClinicalTrials.gov NCT06403345).

Ethical Considerations

Secondary data from the initial study of NATURE (formerly the Green Activity Program) was approved by NYU (New York University) Langone (s22-01409) and Indiana University-Bloomington's Institutional Review Boards (20660). Primary data collection for the feasibility study that is underway

XSL•FO
RenderX

was approved by Indiana University-Bloomington's Institutional Review Board (22206). Participants in both studies provided either written or verbal consent to participate. To protect confidentiality, study data were deidentified in both studies. Participants were compensated US $50 to participate in co-design sessions (secondary data) and US $225 to participate in the feasibility study (primary data study protocol). Applying human-centered design principles enabled participation from participants with varying abilities. We maintained confidentiality at every stage of intervention development and elicited feedback and participants at each stage to improve the intervention. No artificial intelligence technology was used in the writing or preparation of this paper.

Results

We applied the IB framework in 3 steps (Figure 1).

Figure 1. Applying the Innovation Biodesign Framework to map the development and piloting of the NATURE Intervention. The pilot study is underway in the Bronx, NY; Indianapolis, IN; and Burlington, VT. ADRD: Alzheimer disease-related dementias; HR: heart rate; NATURE: Nurturing Aging Through Uplifting Activities in a Restorative Environment; SUS: System Usability Scale.

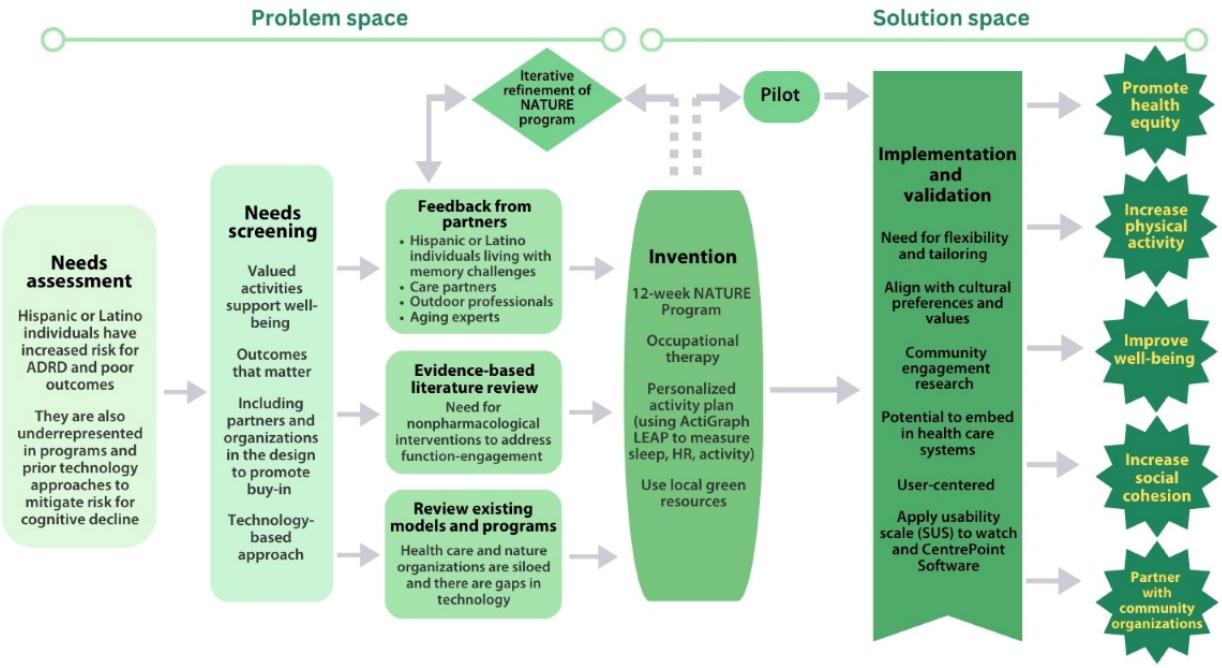

Step 1: Identifying the Problem (Problem Space)

Needs Assessment: Review of the Literature and Existing Programs

We identified a need for nature programs to meet the unmet need of physical inactivity and related risks of decreased well-being, dementia, and subsequent cognitive decline for Hispanic or Latino individuals living with memory challenges.

Physical Inactivity, Sedentary Behavior, and Dementia Risk

Physical inactivity is defined as failing to meet physical activity guidelines [17] and is attributed with 2% of the risk for developing dementia [1]. However, physical activity can also impact other modifiable risk factors, such as decreasing depression (3% risk reduction), hypertension (2%), and high low-density lipoprotein (7%) in midlife, and social isolation (5%) later in life [1], which collectively can contribute to a 19% reduction in the risk for developing dementia. In fact, the American Heart Association endorses physical activity as a critical component of a lifestyle approach as a first line of treatment for hypertension and high low-density lipoprotein [18]. Additionally, a recent meta-analysis with 191,130 participants found that if adults met physical activity guidelines, 11.5% of depression cases could be prevented [19].

Yet, physical activity alone is not enough to reduce risk for dementia and cognitive decline. A longitudinal study of 404 community-dwelling older adults who were free from dementia at baseline found that despite having high levels of physical activity, individuals with high sedentary behavior had neurodegeneration and worse cognition over 7 years [2]. Individuals with apolipoprotein E ε4, a well-established genetic risk factor for AD [20], were associated with high sedentary time and demonstrated faster hippocampal atrophy and decline in naming and information processing speed despite being physically active [2].

Sedentary behavior is an important risk factor for dementia [2]. Sedentary behavior is defined as exerting 1.5 metabolic equivalents in sitting, reclining, or lying down [17]. The World Health Organization recommends replacing 30 minutes of sedentary behavior a day with any kind of physical activity [21]. Replacing 30 minutes/day of leisure-time sedentary behavior with an equal amount of time with different types of physical activity was associated with a 7%-18% decreased risk of dementia incidence and a 12%-21% risk reduction in mortality from dementia for adults with a diagnosis [22].

XSL•FO
RenderX

Nature and Physical Activity

One promising approach to reduce sedentary time and promote physical activity is using nature activities that a person enjoys and are meaningful to them. Nature activity is physical activity in contact with nature. Nature activity can lower anxiety more than physical activity without contact with nature (indoors or outdoors) [23,24]. Simply visiting a park or natural space for 30 minutes at least once a week can reduce hypertension and depression by 7%-9% [25], which are risk factors for dementia [1]. Moreover, visiting parks or green spaces is also associated with higher physical activity and social cohesion [25].

Review of Existing Nature Programs

Nature or "green" prescriptions are integrated into health care systems in New Zealand, Australia, and the United Kingdom for middle-aged and older adults to reduce risk for cardiovascular disease, diabetes, and improve mental health [15,26,27]. Green prescriptions describe the process of prescribing nature experiences or activities to improve a person's health and well-being. Green prescribing provides an enjoyable means to engage in physical activity at a low cost and is rapidly growing [15,26-31]. These interventions are particularly beneficial for people who lack access to nature [29] or experience lower resources [32]. However, there are many different types of nature prescriptions that are implemented in varied settings and with different populations. Components of these prescriptions show promise for people at risk for or living with dementia (Table 1).

Table 1. Examples of nature prescription programs and models.

Type	Setting or delivery	Population	Benefits	Challenges
Green exercise: New Zealand [15,28,33,34]; Australia [27]	Primary care, nurse, and exercise specialist	Inactive middle-aged, older adults [33]	Physical activity, cognition, QoL[a] [34], safe, effective [33], and sustained change [15]	Time, referrals [35], and adherence are impacted by institutional policies and local resources [36]
Green social prescribing [26,31,37,38]: United Kingdom	Primary care and community worker	Mild to moderate mental health problems, community-dwelling adults	Health, well-being, equity, cost-effective, and sustainable [38]	Infrastructure [31], low referrals, appropriateness of referral, and terminology [26]
Gardening [39-42]	Community, long-term care, recreation staff (eg, horticulture therapist), and OT[b]	AD/ADRD[c], MCI[d], care partners	Well-being, engagement, social, cognitive function [39], decreased BPSD[e] [39-42]	Falls, wheel-chair accessibility, and tailoring activities
Animal-assisted interventions [43-46] and farms	Community, long-term care, recreation staff, OT's, physical therapists, and nurses	AD/ADRD, MCI, care partners	Engagement, well-being, cognitive function, social interaction [43], and decreased BPSD [44-46]	Tailoring activities and environmental modifications to avoid safety risks (eg, falls)

[a]QoL: quality of life.

[b]OT: occupational therapist.

[c]AD/ADRD: Alzheimer disease and Alzheimer disease disease-related dementias.

[d]MCI: mild cognitive impairment.

[e]BPSD: biopsychosocial symptoms of dementia.

Outcomes of Nature Prescription Programs

As depicted in Table 1, nature prescriptions for other populations come with potential benefits and challenges. Benefits include increased physical activity, health, cognition, and well-being with sustained behavior change for middle-aged and older adults. While nature prescriptions have not been studied for populations with dementia, nature activities involving gardens and animals add potential benefits of reduced behavioral symptoms (eg, anxiety or depression) and improve well-being [44-46]. Moreover, nature prescriptions contain beneficial components of social engagement, nature, and meaningful activities, which can support function and stave off decline in persons living with dementia [47,48]. Plausible explanations for positive outcomes include (1) regulation of immunological and physiological (stress) responses, (2) enhancement of psychological states (mood and attention), and (3) facilitation of health-promoting behaviors (exercise and social contact) [36], with a central biological pathway of immune functioning by Kuo [49] encompassing all. Despite the potential benefits, a formal nature activity prescription program for this population has yet to be developed and tested.

Identification of Gaps: Challenges, and Critiques of Nature Programming

Despite the benefits, there are challenges to implementing nature programs. Challenges include appropriateness of the activities to meet the person's needs, institutional policies, and a lack of available natural (green) resources, as contributing to low referral rates in primary care [36]. Difficulties with participant engagement also present a challenge to implementing nature prescriptions, as many of these programs grew from a bottom-up approach in the United Kingdom, resulting in challenges with buy-in [15,50].

Critiques of current nature prescriptions include challenges with referrals in primary care settings, varying terminology [36], and overlooking groups who have historically been excluded from

XSL•FO
RenderX

research. Limitations of nature prescriptions include a lack of testing in persons living with dementia and a lack of culturally congruent programs for individuals who have been minoritized in the United States. A recommended strategy to engage underrepresented groups in physical activity interventions is to promote social cohesion and use local resources to increase participation and sustain behavior change [51].

Identification of the Problem

Despite the potential benefits of nature activities to improve well-being and reduce risk for dementia and cognitive decline after a dementia diagnosis, a comprehensive nature activity program for this population has not been developed. Prior models of nature programs have often overlooked communities that are minoritized in the United States, who are at higher risk for developing dementia, such as Hispanic or Latino populations [4]. Additionally, most nature prescribing models are commonly implemented with community workers, who are often unequipped to tailor activities to a person's fluctuating levels of function and decline associated with dementia [52]. Occupational therapists are uniquely positioned to address this need as tailoring and matching valued activities to a person's function, preferences, and needs are central to their scope of practice [53]. There is a need to design an occupational therapist-led comprehensive nature activity program to safely tailor and implement nature activities for people with MCI as a precursor to AD and those living with dementia.

A nature program has not been developed for Hispanic or Latino individuals with memory challenges in the United States, and culturally adapted programs for this population in general are scarce [54]. There is a need to develop a comprehensive nature program that prioritizes the values and human experiences of Hispanic or Latino people living with memory challenges. In line with human-centered design principles, key Hispanic or Latino values that impact the use of supportive programming include familismo (responsibility to family), personalismo (relationships with others), respeto (respect), and dignidad (dignity and self-respect) [55].

Additionally, there is a need to use a partner-driven approach to develop a nature-prescribing program. Particularly, prior nature prescribing programs have lacked input from Hispanic or Latino individuals with memory challenges, care partners, local health care providers, and outdoor organizations. Using a partner-driven approach to develop and implement a nature intervention could mitigate challenges with buy-in and cohesion [36]. To address these gaps, the NATURE program was designed with input from multiple partners, including Hispanic and Latino individuals with memory challenges, care partners, outdoor professionals, and health care providers.

Need Screening: Feedback From Partners and Iterative Refinement

Feedback from partners was elicited during 5 focus groups and 2 interviews to co-design the NATURE intervention. End user engagement through iterative co-design was as described in Lassell et al [12] to inform key elements of the program and research design of the ongoing pilot study. Co-design occurred with Hispanic or Latino individuals living with memory challenges, care partners, outdoor professionals, and interdisciplinary health care providers using a community-based participatory action research [56] approach with human-centered design principles. A community-based participatory action research approach involves an iterative cycle of acting, assessing, reflecting, and refining. People living with memory challenges and their care partners were empowered to contribute to each phase of the research process. Community-based participatory action research principles of empowerment, capacity building, and resource use were applied throughout the development and implementation process. Actions supporting these principles occurred with identifying outcomes relevant to community partners and end users (all groups) with program co-design, implementing the protocol, assessing its acceptability and feasibility, and refining based on their input.

Co-design enabled key elements of the program to align with participant preferences and needs. Co-designed elements of the research design were recruitment strategies, including preferences for terminology and language from the end users (eg, people living with memory challenges vs dementia and Hispanic or Latino vs Latinx). The NATURE program was designed to align with the cultural preferences and values of the Hispanic or Latino community. Particularly, the values of the design of the program also incorporated Hispanic or Latino values of respeto and dignidad were built into the occupational therapy assessment by asking what matters most to the person. Additionally, the values of personalismo and familismo were incorporated with options to include family or friends in the Nature Activity Plan and intervention sessions [12].

End users of Hispanic or Latino people living with memory challenges identified a range of nature activities they enjoyed (eg, walking, dancing, or feeding squirrels) and preferences for program frequency, duration, and delivery-based on prior evidence-based programs [15,25]. They underscored the need to offer a range of options for flexible participation that could be tailored according to individual needs and preferences (eg, 4-8 sessions between 30-90 minutes long). Outdoor professionals provided input on local nature activities and how they would prefer to be supported by the occupational therapist to participate in the program. Health care providers advised on clinical referral pathways, including primary care and community-based organizations. See Lassell et al [12] for iterations of the program before and after the co-design process.

Solution Space

Step 2: Developing the Invention NATURE

The NATURE intervention was informed by the Theory of Cognitive Reserve [57], the Attention Restoration Theory [58], and the environmental press model by Lawton [59], with the assumption that tailoring nature activities that are meaningful to a person (matching the activity demands to their context, preferences, and needs) can support a person's participation in nature activities and well-being. The co-design process incorporated options for session duration and frequencies based on previously successful programs such as the Tailored Activity Program for people living with dementia [60-62] and 2 green prescribing models [15,63]. Co-design resulted in the creation

XSL•FO
RenderX

of a 12-week NATURE intervention that incorporated the preferences and needs of Hispanic or Latino and care partners, local outdoor professionals, and interdisciplinary health care providers [12].

The goal of the NATURE intervention is to promote an active lifestyle and well-being by mitigating risk for dementia and cognitive decline. The intervention advances the US Department of Health and Human Services "National Plan to Address Alzheimer's Disease" Goal 3 [64] as a support for people living with AD and their families and is driven by outcomes that mattered to Hispanic or Latino individuals living with memory challenges [12]. The program involved an occupational therapy assessment, goal-setting, and a tailored green activity plan that could be built into a person's daily routine to address physical inactivity and sedentary time throughout the day. The program emphasizes using local community resources to increase access for homebound participants and to promote sustainability (Table 2).

Table 2. NATURE[a] program for people living with memory challenges.

Time	Description
Week 1: visit 1	• OT[b] evaluation (90 minutes): clinical interview, daily habits and routines, function, fall risk, nature activity checklist, and goal attainment scaling • Education: depression and local resources (eg, food, housing, and transportation). Benefits of nature activities. Participants received a nature activity program binder with educational materials, study contact information, weekly planning behavioral activation worksheets, and a whiteboard weekly calendar
Week 2: visit 2	• Personalized nature activity plan (30-90 minutes) • Finish assessments from the evaluation. Observe or test 1 green activity together. Update plan as needed • Education: SMART[c] goal setting, strategy training, and behavioral activation
Week 3	• Option for additional visit; follow the nature activity plan
Week 4: visit 3	• Test group green activity (30-90 minutes) • Adjust green activity plan as needed • Education: fall prevention, strategy training, and green activity plan
Week 5	• Option for additional visit; follow the nature activity plan
Week 6: visit 4	• Test third green activity (30-90 minutes) • Adjust plan as needed • Education: fall prevention and strategy training
Week 7	• Option for additional visit; follow the nature activity plan
Week 8	• Support from OT: phone or Zoom (Zoom Communications, Inc.) check-in (20-30 minutes)
Week 9	• Option for additional visit; follow the nature activity plan
Week 10	• Support from OT: phone or Zoom (Zoom Communications, Inc.) check-in (20-30 minutes)
Week 11-12	• Follow the nature activity plan

[a]NATURE: Nurturing Aging Through Uplifting Activities in a Restorative Environment.

[b]OT: occupational therapy.

[c]SMART: Specific, Measurable, Achievable, Relevant, and Time-Bound.

Notes

Function was assessed with the Functional Activities Questionnaire, fall risk. The Fall Risk Assessment Tool and the World Health Organization fall risk clinical algorithm were conducted by a trained research assistant before the program. Individuals who were at high fall risk were required to have a care partner present for program sessions and green activities.

NATURE was implemented by 2 licensed occupational therapists with a minimum of 4 years of experience in geriatrics and dementia care. The occupational therapists were trained using a training manual of NATURE with templates for session notes and evaluation forms. The occupational therapists delivered NATURE either virtually using a secure Zoom platform or in-person, based on participant need and preference, in alignment with co-design [12]. NATURE involves 3 phases: assessment, coaching, and self- or care partner–monitoring and sustained behavior change.

The assessment phase informed how the nature activities were tailored in a participant's NATURE plan (Table 3). The assessment phase involved 2 sessions, ranging from 60-90 minutes. Based on the assessment which included clinical interviews, functional assessments, and validated questionnaires, tailoring occurred on 3 levels: the person and family, the nature

XSL•FO
RenderX

activity, and their daily contexts (environmental modifications based on accessibility, safety, weather, equipment needed, level of social support, socioeconomic resources, and the nature resources available to the person in their home, neighborhood, and community). As part of the assessment, participants engaged in collaborative goal-setting using Goal-Attainment Scaling to develop 3 goals for 3 separate individual nature activities that the person wanted to do. Local nature activities were selected from a list compiled through co-design [12]. The local nature list included cost and whether there were virtual options. The participant worked with the occupational therapist to select 3 nature activities they would prefer to do. Goals to engage in the 3 selected nature activities were based on the person's level of function and incorporated the World Health Organization's guidelines for physical activity. Goals could be made to engage in a nature activity or to safely access a nature activity (eg, learning a new bus route to a park) and had options for good days and hard days, and virtual options for inclement weather.

XSL•FO
RenderX

Table 3. Areas of assessment and tailoring in the NATURE[a] program.

Areas	Description
Assessment	
Personalized delivery	• PROMIS[b] health, mental health, and social participation • UCLA[c] Loneliness Scale • Geriatric depression screening • Neuropsychiatric Symptom Questionnaire • Behavioral Activation Scale • Nature-Connectedness Scale (also relevant to context) • Clinical interview A: background • Clinical interview B: routines and preferred nature activities (also context) • Comportment Scale
Activity modification	• Activity levels (Quick Physical Activity Questionnaire) • Physical Activity Readiness Questionnaire + (includes relevant comorbid conditions) • Telephone Montreal Cognitive Assessment • Functional Activities Questionnaire • Function executive function screen
Context	• Health-related social needs • Fall risk assessment (also relevant to personalized delivery) • Environmental assessment: modified Wemstead home assessment • Clinical interview B: routines and preferred nature activities • Nature-Connectedness Scale (also relevant to context)
Tailoring	
Personalized delivery	• Incorporated the person and family's strengths, interests, and values, and honored dislikes. • Options for adaptation based on the person's level of function • Personalized with the level of supervision and cueing (verbal, visual, and physical) needed for safety • How the nature activity was delivered: individually or with family, or with friends
Activity modification	• Options and strategies to make an activity harder or easier for good and challenging days • Activity adaptations included: the number of steps in directions, frequency, and duration of the activity, and the number of rest breaks; included options for backward chaining and forward chaining • Backward chaining • Forward chaining
Context	• Environmental modifications (eg, raised garden beds or tabletop gardening activities) • Level of safety (eg, neighborhood crime rates, environmental hazards, and accessibility) • Weather • Accessibility of sidewalks, parks, and natural spaces • Toxic land use and pollution • Social support • Socioeconomic resources • Available natural resources in a person's home, neighborhood, and community

[a]NATURE: Nurturing Aging Through Uplifting Activities in a Restorative Environment.

[b]PROMIS: Patient-Reported Outcomes Measurement Information System.

[c]UCLA: University of California Los Angelos.

The coaching phase involved either 1:1 virtual, phone, or in-person visits with a trained and licensed occupational therapist lasting between 30 and 90 minutes. The occupational therapist developed a personalized NATURE plan for each participant. Care partners were invited to join but were not required unless a participant was at high fall risk or if they required a legally authorized representative to attend. Each session involved the following components: greeting or introduction, reviewing progress toward each goal, testing at least 1 nature activity related to their goal, making adjustments as needed, reviewing the weekly behavioral activation worksheets, planning their nature activities for the next week,

XSL•FO
RenderX

educational materials (nature activities, depression, reduction of fall risk during plan, and participant-specific education), reviewing session points and scheduling the next session. Behavioral activation worksheets involved weekly goal-setting using SMART (Specific, Measurable, Achievable, Relevant, and Time-Bound) goals, planning their activity, telling their team (occupational therapist, family, or friend), reflecting on how the nature activity went, and problem-solving with the occupational therapist afterward. Care partners completed the behavioral activation form for individuals who had greater cognitive impairment. Participants also mapped out their nature activities for each week using a dry-erase calendar. They planned their nature activities embedded into their daily schedules while considering the weather and air quality. There was an option to include an additional 3 in-person or virtual visits as needed, based on participant need.

The self-monitoring phase involved 2 phone check-ins that were at least 15 minutes with the occupational therapist. Check-in sessions involved a greeting, discussing progress toward goals, and problem-solving any challenges that arose with their NATURE plan, with additional adjustments as needed. Additional education was also provided in relation to each participant's NATURE plan.

Step 3: Mapping the Pilot Study and its Implementation and Validation

NATURE sessions are being implemented virtually or in-person with 1:1 sessions by a licensed occupational therapist. Two occupational therapists are implementing NATURE and received 4 hours of training guided by the NATURE training manual.

The pilot study is being implemented through continued community engagement with local aging organizations that provide services in the Bronx; Indianapolis; and Burlington. Participant feedback regarding program feasibility (retention, feasibility of outcome measures, or fidelity) and acceptability (questionnaires and interviews) is being gathered. Planned outcomes of the program include promoting health equity and an active lifestyle, mitigating risk factors for cognitive decline related to physical inactivity and social isolation. Additionally, health outcomes related to self-reported physical activity using the Quick Physical Activity Rating [65] scale and well-being, such as the PROMIS measures of Social Participation, Mental Health, and Physical Health [66,67], are also being collected among other health outcomes (Table 4).

Table 4. Health outcomes.

Primary goal	Goal attainment scaling (primary goal met, yes or no)
Function	Telephone Montreal Cognitive Assessment, Quick Physical Activity Rating Scale, Functional Activities Questionnaire, Physical Activity (ActiGraph)
Health	Heart rate, sleep (ActiGraph), PROMIS[a] (mental health, physical health, and social participation), Geriatric Depression Scale Short form, UCLA[b] Loneliness Scale, Neuropsychiatric Symptoms Questionnaire
Nature	Nature Connectedness Scale
Usability	Holden Systems' Usability Scale [68]
Mechanism of change	Behavioral Activation Short-Form

[a]PROMIS: Patient-Reported Outcomes Measurement Information System.

[b]UCLA: University of California Los Angeles.

Using human-centered design principles, we are incorporating multiple methods to measure preferred outcomes of well-being, sleep, and social connection. We are using the ActiGraph LEAP to measure physiological data related to well-being (heart rate and activity), sleep, and validated questionnaires. We are also piloting the CentrePoint Connect mobile app (ActiGraph, LLC) sync ActiGraph data in real-time and assess the usability of these technologies using Holden Systems' Usability Scale [68]. Additionally, qualitative data will be gathered to learn about each participant's experience during the program to iteratively refine the intervention and measurement of outcomes before efficacy testing.

The NATURE intervention is designed to be sustainable through its use of local resources that are accessible by participants (eg, ongoing nature activities in their community, recommending nature activities they can still access after the intervention) with tailored strategies to adapt and integrate nature activities into the person's daily routine. Using local natural resources may also contribute to social cohesion for participants and their communities. This can be achieved through continued partnerships with local outdoor organizations and our local health care partners.

If determined effective in the future, the NATURE intervention has the potential to be embedded into current health care systems, with potential reimbursement under Medicare with skilled home health care (part A) and home-based outpatient services covered under part B [69]. However, before being implemented in other contexts with other groups, context-specific adaptations will need to be made by identifying local nature activities and referral pathways for local community-based organizations and hospital systems. Additional refinements may also be needed for language preferences to describe AD and related dementias to successfully recruit participants from other cultural groups, and have members of the group assist in implementation. Additional adaptations may also be needed to align with cultural values.

XSL•FO
RenderX

Discussion

Overview

The main finding of this use case is that the IB framework provides a systematic approach to apply human-centered design principles to design a program to address well-being and physical activity in Hispanic or Latino individuals with a range of cognitive abilities. The application of the IB framework [11] use case of the NATURE [12] intervention highlights the utility in visualizing the application of human-centered design principles throughout NATURE's progression from development to piloting. The framework will continue to guide human-centered refinements through efficacy and effectiveness testing and dissemination as the program develops and can serve as a useful tool for other researchers seeking to develop human-centered programs for underrepresented populations.

Principal Results

The IB framework [11] enabled us to map human-centered design principles in end user engagement through the application of co-design in creating the NATURE [12] program to be responsive to the preferences and needs of Hispanic or Latino people living with memory challenges. It also provided a map to guide piloting an activity tracker with a mobile app as a means to measure sleep and outcomes related to well-being (heart rate and activity) and assess the usability of these technologies. Use of the IB framework and other human design frameworks can advance health equity by prioritizing diverse participants' needs and experiences [10,11].

The current study underscores the benefits of using the IB framework to develop and test interventions and programs with input from multiple partners, as well as the end user. This aligns with the human-centered design principle of involving the user and multiple partners in every step of the design process and understanding the context in which the intervention or program will occur [10,13]. The IB framework provided guiding principles for moving the challenge of addressing physical inactivity in Hispanic or Latino older adults with memory challenges from a problem space to a solution space informed by the needs and preferences of partners and participants [11]. The inclusion of multi-partner perspectives is an inclusive innovation process that values input from all involved, including older adults living with memory challenges.

Rapid-cycle iterations within the IB framework enable researchers to map human-centered involvement and adapt a program or innovation based on participants' dynamic needs [11]. For example, the rapid-cycle iterations that occurred during the co-design of NATURE enabled researchers to update the program in response to participant feedback and insights. Rapid-cycle iterations can help NATURE and other health innovations successfully evolve based on the end users' dynamic preferences and needs, enhancing the program or products' accessibility, impact, and reach. This aligns with human-centered design principles of rapid-cycle iterations and refinements based on the end user's evaluation [10,13]. Next steps for NATURE are to finish piloting the intervention and receive input from the end users (participants and partners) on refinements. Once proof of concept and feasibility have been established, the next steps include effectiveness testing and maximizing the outreach of the program.

Comparison With Prior Work

While there are other helpful human-centered design frameworks, they are not specifically designed to incorporate the voices of underrepresented populations. For example, the Approach to Human-Centered, Evidence-Driven Adaptive Design for health care interventions focuses on an empathy-driven approach to rapidly design health care innovations [70] and does not focus on the inclusion of minoritized groups in the design process. Additionally, the discover, design or build, and test framework draws from other evidence-based human-centered design frameworks to modify and evaluate the usability of evidence-based psychosocial interventions and implementation strategies [71]. A benefit of using the IB framework was that it allowed us to incorporate the voices of Hispanic or Latino individuals to design NATURE and assess the usability of an activity tracker and mobile app technology to measure and monitor outcomes related to sleep, heart rate, and activity.

Benefits of using the IB framework from prior research include improved health and digital health literacy for patients with asthma [11]. Through its collaborative approach, this framework has improved clinical outcomes for individuals with asthma but also enabled health care providers to offer more personalized care plans. Patients were better equipped to understand and manage their health conditions through education provided by an app. The framework fosters stronger patient-provider relationships, bolstering trust and communication. For individuals with memory challenges, active involvement in this framework could enhance their grasp of health issues, treatment choices, and self-care strategies. This empowerment may lead to better adherence to treatment and overall well-being, emphasizing a patient's proactive role in health management and aligning with human-centered design principles of prioritizing the human experience [13]. Empowering patients as the end user can be critical when designing health interventions and innovations that seek to promote health equity [10].

The framework can also be helpful for health care providers as it fosters the development of cutting-edge medical technologies and solutions tailored to address unmet clinical needs [11]. By integrating engineering, business, and medical disciplines, the framework equips providers with the tools to innovate and improve patient care outcomes effectively. This aligns with the human-centered design principle of using a multidisciplinary team [13]. The framework is also highly relevant to multidisciplinary medical and clinical informatics professionals because it focuses on collaborative and inclusive health care innovation with a systematic process to map human-centered design. Public health professionals may also find this framework helpful to develop policies and public health initiatives by using human-centered design principles, such as the end users' evaluation to iteratively refine a policy for better uptake [10]. This process can guide the development of programs, interventions, and innovations that are specifically designed to

XSL•FO
RenderX

meet the needs of diverse populations, which can increase their effectiveness and acceptance.

Application of the IB framework provided a structure for applying human-centered design principles to design and pilot the NATURE program for Hispanic or Latino individuals with memory challenges. The framework can be applied to the development and implementation of other interventions to incorporate the human experiences and values of the end user and may support the applicability and sustainability of interventions for underrepresented populations.

Limitations

A limitation of applying the IB framework in the NATURE use case is that the framework spans development through the dissemination phases of an intervention or program. Our use case is focused on a program that is in the early piloting phase of development and does not provide the full range of the framework through efficacy testing and dissemination. Yet, a strength of our use case is that it shows the utility of using the framework in the early stages of intervention development, which can guide feasibility and piloting through efficacy testing and dissemination. A limitation of using secondary data was that it was initially collected for another purpose (designing NATURE) and was retroactively fitted to the IB framework. Yet, a strength of our secondary data was that it matched the user-centered design process of the IB framework.

Acknowledgments

This research was conducted while RL was a Clinician-Scientists Transdisciplinary Aging Research Pilot Grantee. No federal funds were awarded in support of this project. The project was entirely funded through private funds provided by the American Federation for Aging Research (A23-0142-001). The investigators retain full independence in the conduct of this research. The funder had no involvement in this study's design, data collection, analysis, interpretation, or the writing of this paper. ChatGPT (OpenAI) was not used in the writing ideation, writing, or the generation of figures in this paper.

Data Availability

Data sharing does not apply to this paper as no datasets were generated or analyzed during this study.

Authors' Contributions

RL handled the funding, conceptualization, data collection, data analysis, writing, and revision of this paper. AM did the conceptualization, preparation of tables, writing, and revision of this paper. KW worked on the conceptualization, figure preparation, and writing of this paper. SH and PG conceptualized, wrote, and revised this paper. SZ prepared the figure and wrote this paper. TP collected the data and wrote this paper. CP wrote and revised this paper. LG carried out the funding, conceptualization, supervision, and revision. SJ conceptualized and revised this paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
GUIDED checklist.
[PDF File (Adobe PDF File), 588 KB - formative_v9i1e64930_app1.pdf]

References

1. Livingston G, Huntley J, Liu KY, Costafreda SG, Selbæk G, Alladi S, et al. Dementia prevention, intervention, and care: 2024 report of the Lancet standing Commission. Lancet 2024 Aug 10;404(10452):572-628. [doi: 10.1016/S0140-6736(24)01296-0] [Medline: 39096926]
2. Gogniat MA, Khan OA, Li J, Park C, Hudson Robb W, Zhang P, et al. Increased sedentary behavior is associated with neurodegeneration and worse cognition in older adults over a 7-year period despite high levels of physical activity. Alzheimers Dement 2025 May;21(5):e70157. [doi: 10.1002/alz.70157] [Medline: 40357887]
3. Center for disease control and prevention. Adult Physical Inactivity Outside of Work. 2025. URL: https://www.cdc.gov/physical-activity/php/data/inactivity-maps.html [accessed 2025-09-06]
4. Matthews KA, Xu W, Gaglioti AH, Holt JB, Croft JB, Mack D, et al. Racial and ethnic estimates of Alzheimer's disease and related dementias in the United States (2015-2060) in adults aged ≥65 years. Alzheimers Dement 2019 Jan;15(1):17-24 [FREE Full text] [doi: 10.1016/j.jalz.2018.06.3063] [Medline: 30243772]
5. Livingston G, Huntley J, Sommerlad A, Ames D, Ballard C, Banerjee S, et al. Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. The Lancet 2020 Aug;396(10248):413-446. [doi: 10.1016/s0140-6736(20)30367-6]
6. Shaw AR, Perales-Puchalt J, Johnson E. Representation of racial and ethnic minority populations in dementia prevention trials: a systematic review. J Prev Alzheimers Dis 2022;9:113-118. [doi: 10.1101/2021.05.18.21257411]

XSL•FO
RenderX

7. Quiroz YT, Solis M, Aranda MP, Arbaje AI, Arroyo-Miranda M, Cabrera LY, et al. Alzheimers Dement 2022 Sep;18(9):1677-1686 [FREE Full text] [doi: 10.1002/alz.12589] [Medline: 35199931]
8. Willinger N, Steele J, Atkinson L, Liguori G, Jimenez A, Mann S, et al. Effectiveness of Structured Physical Activity Interventions Through the Evaluation of Physical Activity Levels, Adoption, Retention, Maintenance, and Adherence Rates: A Systematic Review and Meta-Analysis. J Phys Act Health 2021 Jan 01;18(1):116-129. [doi: 10.1123/jpah.2019-0349] [Medline: 33383569]
9. Sansano-Nadal O, Giné-Garriga M, Brach JS, Wert DM, Jerez-Roig J, Guerra-Balic M, et al. Exercise-Based Interventions to Enhance Long-Term Sustainability of Physical Activity in Older Adults: A Systematic Review and Meta-Analysis of Randomized Clinical Trials. Int J Environ Res Public Health 2019 Jul 15;16(14):2527 [FREE Full text] [doi: 10.3390/ijerph16142527] [Medline: 31311165]
10. Holeman I, Kane D. Human-centered design for global health equity. Inf Technol Dev 2019 Sep 29;26(3):477-505 [FREE Full text] [doi: 10.1080/02681102.2019.1667289] [Medline: 32982007]
11. Metaxas A, Hantgan S, Wang KW, Desai J, Zwerling S, Jariwala SP. A Framework for Social Needs-Based Medical Biodesign Innovation. Appl Clin Inform 2024 May;15(3):456-459 [FREE Full text] [doi: 10.1055/a-2312-8621] [Medline: 38657678]
12. Lassell RKF, Tamayo V, Pena TA, Kishi M, Zwerling J, Gitlin LN, et al. "When she goes out, she feels better:" co-designing a Green Activity Program with Hispanic/Latino people living with memory challenges and care partners. Front Aging Neurosci 2024;16:1401255 [FREE Full text] [doi: 10.3389/fnagi.2024.1401255] [Medline: 38957542]
13. ISO 9241-210: Human-centered design for interactive systems. Vernier, Geneva: International Organization for Standardization; 2019:1-40.
14. Slattery P, Saeri AK, Bragge P. Research co-design in health: a rapid overview of reviews. Health Res Policy Syst 2020 Feb 11;18(1):17 [FREE Full text] [doi: 10.1186/s12961-020-0528-9] [Medline: 32046728]
15. Gifford KA, Liu D, Romano R, Jones RN, Jefferson AL. Development of a subjective cognitive decline questionnaire using item response theory: a pilot study. Alzheimers Dement (Amst) 2015 Dec 01;1(4):429-439 [FREE Full text] [doi: 10.1016/j.dadm.2015.09.004] [Medline: 26878034]
16. Duncan E, O'Cathain A, Rousseau N, Croot L, Sworn K, Turner KM, et al. Guidance for reporting intervention development studies in health research (GUIDED): an evidence-based consensus study. BMJ Open 2020 Apr 08;10(4):e033516 [FREE Full text] [doi: 10.1136/bmjopen-2019-033516] [Medline: 32273313]
17. Tremblay MS, Aubert S, Barnes JD, Saunders TJ, Carson V, Latimer-Cheung AE, SBRN Terminology Consensus Project Participants. Sedentary Behavior Research Network (SBRN) - Terminology Consensus Project process and outcome. Int J Behav Nutr Phys Act 2017 Jun 10;14(1):75 [FREE Full text] [doi: 10.1186/s12966-017-0525-8] [Medline: 28599680]
18. Barone Gibbs B, Hivert MF, Jerome GJ, Kraus WE, Rosenkranz SK, Schorr EN, American Heart Association Council on LifestyleCardiometabolic Health; Council on CardiovascularStroke Nursing;Council on Clinical Cardiology. Physical Activity as a Critical Component of First-Line Treatment for Elevated Blood Pressure or Cholesterol: Who, What, and How?: A Scientific Statement From the American Heart Association. Hypertension 2021 Aug;78(2):e26-e37 [FREE Full text] [doi: 10.1161/HYP.0000000000000196] [Medline: 34074137]
19. Pearce M, Garcia L, Abbas A, Strain T, Schuch FB, Golubic R, et al. Association Between Physical Activity and Risk of Depression: A Systematic Review and Meta-analysis. JAMA Psychiatry 2022 Jun 01;79(6):550-559 [FREE Full text] [doi: 10.1001/jamapsychiatry.2022.0609] [Medline: 35416941]
20. Dumuid D, Mellow ML, Olds T, Tregoweth E, Greaves D, Keage H, et al. Does APOE 4 Status Change How 24-Hour Time-Use Composition Is Associated with Cognitive Function? An Exploratory Analysis Among Middle-to-Older Adults. JAD 2022 Aug 02;88(3):1157-1165. [doi: 10.3233/jad-220181]
21. Risk Reduction of Cognitive Decline and Dementia: WHO Guidelines. Geneva, Switzerland: World Health Organization; 2019.
22. Sun Y, Chen C, Yu Y, Zhang H, Tan X, Zhang J, et al. Replacement of leisure-time sedentary behavior with various physical activities and the risk of dementia incidence and mortality: A prospective cohort study. J Sport Health Sci 2023 May;12(3):287-294 [FREE Full text] [doi: 10.1016/j.jshs.2022.11.005] [Medline: 36379419]
23. de Brito JN, Pope ZC, Mitchell NR, Schneider IE, Larson JM, Horton TH, et al. Changes in Psychological and Cognitive Outcomes after Green versus Suburban Walking: A Pilot Crossover Study. Int J Environ Res Public Health 2019 Aug 13;16(16):2894 [FREE Full text] [doi: 10.3390/ijerph16162894] [Medline: 31412602]
24. Thompson Coon J, Boddy K, Stein K, Whear R, Barton J, Depledge MH. Does participating in physical activity in outdoor natural environments have a greater effect on physical and mental wellbeing than physical activity indoors? A systematic review. Environ Sci Technol 2011 Mar 01;45(5):1761-1772. [doi: 10.1021/es102947t] [Medline: 21291246]
25. Shanahan DF, Bush R, Gaston KJ, Lin BB, Dean J, Barber E, et al. Health Benefits from Nature Experiences Depend on Dose. Sci Rep 2016 Jun 23;6(1):28551 [FREE Full text] [doi: 10.1038/srep28551] [Medline: 27334040]
26. Bragg R, Leck C. Good Practice in Social Prescribing for Mental Health: The Role of Nature-Based Interventions. London: Natural England; 2017.
27. Annear M, Lucas P, Wilkinson T, Shimizu Y. Prescribing physical activity as a preventive measure for middle-aged Australians with dementia risk factors. Aust. J. Prim. Health 2019;25(2):108-112. [doi: 10.1071/py18171]

XSL•FO
RenderX

28. Elliot CA, Hamlin MJ. Combined diet and physical activity is better than diet or physical activity alone at improving health outcomes for patients in New Zealand's primary care intervention. BMC Public Health 2018 Feb 08;18(1):230 [FREE Full text] [doi: 10.1186/s12889-018-5152-z] [Medline: 29422040]
29. Allen J, Balfour R. Natural Solutions for Tackling Health Inequalities. England: Realising the Opportunity; 2012.
30. Kruize H, van der Vliet N, Staatsen B, Bell R, Chiabai A, Muiños G, et al. Urban Green Space: Creating a Triple Win for Environmental Sustainability, Health, and Health Equity through Behavior Change. Int J Environ Res Public Health 2019 Nov 11;16(22):4403 [FREE Full text] [doi: 10.3390/ijerph16224403] [Medline: 31717956]
31. Polley M, Bertotti M, Kimberlee R. A Review of the Evidence Assessing Impact of Social Prescribing on Healthcare Demand and Cost Implications. London: University of Westminster; 2017.
32. Rigolon A, Browning MHEM, McAnirlin O, Yoon H. Green Space and Health Equity: A Systematic Review on the Potential of Green Space to Reduce Health Disparities. Int J Environ Res Public Health 2021 Mar 04;18(5):2563 [FREE Full text] [doi: 10.3390/ijerph18052563] [Medline: 33806546]
33. Kerse N, Elley CR, Robinson E, Arroll B. Is physical activity counseling effective for older people? A cluster randomized, controlled trial in primary care. J Am Geriatr Soc 2005 Nov;53(11):1951-1956. [doi: 10.1111/j.1532-5415.2005.00466.x] [Medline: 16274377]
34. Lawton BA, Rose SB, Elley CR, Dowell AC, Fenton A, Moyes SA. Exercise on prescription for women aged 40-74 recruited through primary care: two year randomised controlled trial. BMJ 2008 Dec 11;337(dec11 3):a2509 [FREE Full text] [doi: 10.1136/bmj.a2509] [Medline: 19074218]
35. Van den Berg AE. From Green Space to Green Prescriptions: Challenges and Opportunities for Research and Practice. Front Psychol 2017;8:268 [FREE Full text] [doi: 10.3389/fpsyg.2017.00268] [Medline: 28289398]
36. Van den Berg AE. From Green Space to Green Prescriptions: Challenges and Opportunities for Research and Practice. Front Psychol 2017;8:268 [FREE Full text] [doi: 10.3389/fpsyg.2017.00268] [Medline: 28289398]
37. Howarth M, Griffiths A, da Silva A, Green R. Social prescribing: a 'natural' community-based solution. Br J Community Nurs 2020 Jun 02;25(6):294-298 [FREE Full text] [doi: 10.12968/bjcn.2020.25.6.294] [Medline: 32496851]
38. Polley M, Fleming J, Anfilogoff T. Making sense of social prescribing. In: Westminster, University. London: SAGE Publications Inc; 2017:1-79.
39. Zhao Y, Liu Y, Wang Z. Effectiveness of horticultural therapy in people with dementia: A quantitative systematic review. J Clin Nurs 2022 Jul;31(13-14):1983-1997. [doi: 10.1111/jocn.15204] [Medline: 32017241]
40. Smith-Carrier TA, Béres L, Johnson K, Blake C, Howard J. Digging into the experiences of therapeutic gardening for people with dementia: An interpretative phenomenological analysis. Dementia (London) 2021 Jan;20(1):130-147. [doi: 10.1177/1471301219869121] [Medline: 31426675]
41. Lassell R, Wood W, Schmid AA, Cross JE. A comparison of quality of life indicators during two complementary interventions: adaptive gardening and adaptive riding for people with dementia. Complement Ther Med 2021 Mar;57:102658 [FREE Full text] [doi: 10.1016/j.ctim.2020.102658] [Medline: 33429038]
42. Noone S, Jenkins N. Digging for Dementia: Exploring the experience of community gardening from the perspectives of people with dementia. Aging Ment Health 2018 Jul;22(7):881-888. [doi: 10.1080/13607863.2017.1393793] [Medline: 29068703]
43. Wesenberg S, Mueller C, Nestmann F, Holthoff-Detto V. Effects of an animal-assisted intervention on social behaviour, emotions, and behavioural and psychological symptoms in nursing home residents with dementia. Psychogeriatrics 2019 May;19(3):219-227. [doi: 10.1111/psyg.12385] [Medline: 30393989]
44. Yakimicki ML, Edwards NE, Richards E, Beck AM. Animal-Assisted Intervention and Dementia: A Systematic Review. Clin Nurs Res 2019 Jan;28(1):9-29. [doi: 10.1177/1054773818756987] [Medline: 29441797]
45. Nordgren L, Engström G. Effects of animal-assisted therapy on behavioral and/or psychological symptoms in dementia: a case report. Am J Alzheimers Dis Other Demen 2012 Dec;27(8):625-632 [FREE Full text] [doi: 10.1177/1533317512464117] [Medline: 23100630]
46. Nordgren L, Engström G. Animal-assisted intervention in dementia: effects on quality of life. Clin Nurs Res 2014 Feb;23(1):7-19. [doi: 10.1177/1054773813492546] [Medline: 23800890]
47. Regier NG, Parisi JM, Perrin N, Gitlin LN. Engagement in Favorite Activity and Implications for Cognition, Mental Health, and Function in Persons Living With and Without Dementia. J Appl Gerontol 2022 Feb;41(2):441-449 [FREE Full text] [doi: 10.1177/0733464821999199] [Medline: 33733908]
48. Oh A, Gan S, Boscardin WJ, Allison TA, Barnes DE, Covinsky KE, et al. Engagement in Meaningful Activities Among Older Adults With Disability, Dementia, and Depression. JAMA Intern Med 2021 Apr 01;181(4):560-562 [FREE Full text] [doi: 10.1001/jamainternmed.2020.7492] [Medline: 33492334]
49. Kuo M. How might contact with nature promote human health? Promising mechanisms and a possible central pathway. Front Psychol 2015;6:1093 [FREE Full text] [doi: 10.3389/fpsyg.2015.01093] [Medline: 26379564]
50. Drinkwater C, Wildman J, Moffatt S. Social prescribing. BMJ 2019 Mar 28;364:l1285. [doi: 10.1136/bmj.l1285] [Medline: 30923039]

XSL•FO
RenderX

51. Yi SS, Kanaya AM, Wen M, Russo R, Kandula N. Associations of Neighborhood Factors and Activity Behaviors: The Mediators of Atherosclerosis in South Asians Living in America (MASALA) Study. J Immigr Minor Health 2021 Feb;23(1):54-61 [FREE Full text] [doi: 10.1007/s10903-020-01021-6] [Medline: 32418001]
52. Matar E, Shine JM, Halliday GM, Lewis SJG. Cognitive fluctuations in Lewy body dementia: towards a pathophysiological framework. Brain 2020 Jan 01;143(1):31-46. [doi: 10.1093/brain/awz311] [Medline: 31612904]
53. Dorsey J. Occupational therapy scope of practice. Am J Occup Ther 2021;75(Supplement_3):7513410020. [doi: 10.5014/ajot.2021.75s3005]
54. Vila-Castelar C, Fox-Fuller JT, Guzmán-Vélez E, Schoemaker D, Quiroz YT. A cultural approach to dementia - insights from US Latino and other minoritized groups. Nat Rev Neurol 2022 May;18(5):307-314 [FREE Full text] [doi: 10.1038/s41582-022-00630-z] [Medline: 35260817]
55. C. Arévalo-Flechas L, Acton G, I. Escamilla M, N. Bonner P, L. Lewis S. Latino Alzheimer's caregivers: what is important to them? Journal of Managerial Psych 2014 Aug 05;29(6):661-684. [doi: 10.1108/jmp-11-2012-0357]
56. Kwon S, Tandon S, Islam N, Riley L, Trinh-Shevrin C. Applying a community-based participatory research framework to patient and family engagement in the development of patient-centered outcomes research and practice. Transl Behav Med 2018 Sep 08;8(5):683-691 [FREE Full text] [doi: 10.1093/tbm/ibx026] [Medline: 30202926]
57. STERN Y. What is cognitive reserve? Theory and research application of the reserve concept. J Int Neuropsychol Soc 2002 May 16;8(3):448-460. [doi: 10.1017/s1355617702813248]
58. Kaplan S. The restorative benefits of nature: toward an integrative framework. J Environ Psychol 1995;15(3):169-182. [doi: 10.1016/0272-4944(95)90001-2]
59. Lawton MP. The physical environment of the person with Alzheimer's disease. Aging & Mental Health 2010 Jun 09;5(sup1):56-64. [doi: 10.1080/713650004]
60. Gitlin LN, Winter L, Vause Earland T, Adel Herge E, Chernett NL, Piersol CV, et al. The Tailored Activity Program to reduce behavioral symptoms in individuals with dementia: feasibility, acceptability, and replication potential. Gerontologist 2009 Jun;49(3):428-439 [FREE Full text] [doi: 10.1093/geront/gnp087] [Medline: 19420314]
61. Bennett S, Travers C, Liddle J, Smith S, Clemson L, O'Reilly M, et al. Acceptability and willingness to participate in the Tailored Activity Program: perceptions of people living with dementia, their care partners and health professionals. Brain Impairment 2020 Oct 28;21(3):259-273. [doi: 10.1017/brimp.2020.15]
62. Gitlin LN, Marx K, Piersol CV, Hodgson NA, Huang J, Roth DL, et al. Effects of the tailored activity program (TAP) on dementia-related symptoms, health events and caregiver wellbeing: a randomized controlled trial. BMC Geriatr 2021 Oct 20;21(1):581 [FREE Full text] [doi: 10.1186/s12877-021-02511-4] [Medline: 34670502]
63. Green social prescribing. National Health Services England. 2021. URL: https://www.england.nhs.uk/personalisedcare/social-prescribing/green-social-prescribing/ [accessed 2025-09-06]
64. National plan to address Alzheimer's disease 2025 update. U.S. Department of Health and Human Services. 2025. URL: https://aspe.hhs.gov/national-plan-address-alzheimers-disease [accessed 2025-09-13]
65. Galvin JE, Tolea MI, Rosenfeld A, Chrisphonte S. The Quick Physical Activity Rating (QPAR) scale: A brief assessment of physical activity in older adults with and without cognitive impairment. PLoS One 2020;15(10):e0241641 [FREE Full text] [doi: 10.1371/journal.pone.0241641] [Medline: 33125429]
66. Cook KF, Jensen SE, Schalet BD, Beaumont JL, Amtmann D, Czajkowski S, et al. PROMIS measures of pain, fatigue, negative affect, physical function, and social function demonstrated clinical validity across a range of chronic conditions. J Clin Epidemiol 2016 May;73:89-102 [FREE Full text] [doi: 10.1016/j.jclinepi.2015.08.038] [Medline: 26952842]
67. Cook KF, Bamer AM, Amtmann D, Molton IR, Jensen MP. Six patient-reported outcome measurement information system short form measures have negligible age- or diagnosis-related differential item functioning in individuals with disabilities. Arch Phys Med Rehabil 2012 Jul;93(7):1289-1291. [doi: 10.1016/j.apmr.2011.11.022] [Medline: 22386213]
68. Holden R. A simplified system usability scale (SUS) for cognitively impaired and older adults. Proc Int Symp Hum Factors Ergon Health Care 2020 Sep 16;9(1):180-182. [doi: 10.1177/2327857920091021]
69. Medicare & Home Health Care. 2023. URL: https://www.cms.gov/ [accessed 2025-09-06]
70. Fischer M, Safaeinili N, Haverfield MC, Brown-Johnson CG, Zionts D, Zulman DM. Approach to Human-Centered, Evidence-Driven Adaptive Design (AHEAD) for Health Care Interventions: a Proposed Framework. J Gen Intern Med 2021 Apr;36(4):1041-1048 [FREE Full text] [doi: 10.1007/s11606-020-06451-4] [Medline: 33537952]
71. Lyon AR, Munson SA, Renn BN, Atkins DC, Pullmann MD, Friedman E, et al. Use of Human-Centered Design to Improve Implementation of Evidence-Based Psychotherapies in Low-Resource Communities: Protocol for Studies Applying a Framework to Assess Usability . JMIR Res Protoc 2019 Oct 09;8(10):e14990 [FREE Full text] [doi: 10.2196/14990] [Medline: 31599736]

Abbreviations

AD: Alzheimer disease
IB: innovation biodesign
MCI: mild cognitive impairment

XSL•FO
RenderX

NATURE: Nurturing Aging Through Uplifting Activities in a Restorative Environment
SMART: Specific, Measurable, Achievable, Relevant, and Time-Bound

Edited by A Mavragani; submitted 31.07.24; peer-reviewed by B Tosti; comments to author 28.05.25; revised version received 31.07.25; accepted 30.08.25; published 15.10.25.

Please cite as:
Lassell R, Metaxas A, Wang K, Hantgan S, Gottipati P, Zwerling S, Pena T, Pollak C, Gitlin L, Jariwala S
Applying the Human-Centered Innovation Biodesign Framework to the Development and Piloting of a Program to Mitigate Risk for Cognitive Decline Among Historically Underrepresented Individuals: Case Study
JMIR Form Res 2025;9:e64930
URL: https://formative.jmir.org/2025/1/e64930
doi:10.2196/64930
PMID:

©Rebecca Lassell, Ada Metaxas, Katherine Wang, Sara Hantgan, Prabhat Gottipati, Sarah Zwerling, Triana Pena, Chava Pollak, Laura Gitlin, Sunit Jariwala. Originally published in JMIR Formative Research (https://formative.jmir.org), 15.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Promoting Walking in Cardiopulmonary Disease With Mindful Steps: Pilot Feasibility Randomized Controlled Trial of a Web-Based, Pedometer-Mediated Mind-Body Intervention

Kristen M Kraemer[1], PhD; Daniel Litrownik[2], BA; Peter M Wayne[2], PhD; Caroline R Richardson[3], MD; Neha Bhomia[4], BDS, MHI; Reema Kadri[4], MLIS; Pamela M Rist[2], ScD; Long Ngo[1], PhD; Marilyn L Moy[5,6], MD; Gloria Y Yeh[1,2], MD, MPH

[1]Division of General Medicine, Department of Medicine, Beth Israel Deaconess Medical Center/Harvard Medical School, Boston, MA, United States
[2]Osher Center for Integrative Health, Harvard Medical School and Brigham and Women's Hospital, Boston, MA, United States
[3]Warren Alpert Medical School, Brown University, Providence, RI, United States
[4]Department of Family Medicine, University of Michigan, Ann Arbor, MI, United States
[5]Veterans Administration Boston Healthcare System, Boston, MA, United States
[6]Harvard Medical School, Boston, MA, United States

Corresponding Author:
Kristen M Kraemer, PhD
Division of General Medicine, Department of Medicine, Beth Israel Deaconess Medical Center/Harvard Medical School
330 Brookline Ave
Boston, MA, 02215
United States
Phone: 1 617 754 1443
Email: kkraemer@bidmc.harvard.edu

Abstract

Background: Physical inactivity is highly prevalent in heart failure (HF) and chronic obstructive pulmonary disease (COPD) and is associated with poor outcomes, including worsened quality of life, increased hospitalizations, readmissions, and mortality. Accessible interventions that improve physical activity are needed. Mind-body strategies are well-suited for promoting physical activity; they show promise for targeting key health behavior change processes.

Objective: The aim of this study was to examine the feasibility and acceptability of a web-based pedometer-mediated mind-body intervention (Mindful Steps) for promoting walking among individuals with HF and COPD.

Methods: In this pilot randomized controlled trial, participants with chronic, stable HF and COPD were randomized to Mindful Steps or usual care in a 2:1 ratio. Mindful Steps is a 12-month multimodal intervention that includes a pedometer with individualized step-count goals, live mind-body exercise (MBE) classes, and a web platform with mind-body videos, motivation messages, and educational tips. Feasibility (recruitment rate, retention), intervention acceptability, and intervention adherence were the primary outcomes. Exploratory outcomes assessed at baseline, 3-, 6-, 9-, and 12-months included daily step counts, cognitive-behavioral/psychosocial measures, health-related quality of life, and self-reported physical function. Participants were enrolled in the study from April 2019 to July 2021. The study was converted to all-digital during the pandemic after March 2020.

Results: Forty-one participants were randomized to Mindful Steps (n=26) or usual care (n=15). The recruitment rate was 3% (43/178), and overall study retention was 76% (31/41). In the intervention group, over 12 months, 58% (15/26) met a predefined benchmark for MBE class adherence (attending >70% of classes). Participants engaged most consistently with the MBE classes, the pedometer, and mind-body videos. There was a positive signal regarding group differences in the change in daily step counts from baseline, favoring intervention at 3 months (estimate=2038.77 steps per day between groups, 95% CI 289.76-3788.77), 6 months (estimate=3031.45, 95% CI 1261.15-4801.74), and 9 months (estimate=2703.80, 95% CI 862.97-4544.62). There were also positive signals regarding group differences in the change from baseline favoring intervention in the following outcomes: emotional awareness (estimate=0.88, 95% CI 0.15-1.61) and body listening (estimate=1.16, 95% CI 0.25-2.07) at 3-months; internal motivation (estimate=1.03, 95% CI 0.01-2.04) and pressure/tension at 6-months (estimate=–1.59, 95% CI –2.55 to –0.63); and exercise self-efficacy at 12 months (estimate=1.77, 95% CI 0.20-3.33).

XSL•FO
RenderX

Conclusions: Mindful Steps was largely feasible, acceptable, and had adequate intervention engagement. There were positive signals favoring the multimodal web intervention for daily step counts, interoceptive awareness, internal motivation, and exercise self-efficacy that will inform hypotheses in future studies. A pivot to fully remote conduct during the pandemic was successful. A larger trial examining the efficacy of Mindful Steps for promoting physical activity is warranted.

Trial Registration: ClinicalTrials.gov NCT03003780; https://clinicaltrials.gov/study/NCT03003780

(JMIR Form Res 2025;9:e74118) doi:10.2196/74118

KEYWORDS

heart failure; chronic obstructive pulmonary disease; physical activity; mind-body; step counts

Introduction

Chronic obstructive pulmonary disease (COPD) and heart failure (HF) are chronic cardiopulmonary conditions that share similar morbidity and clinically relevant decreases in physical activity. Physical inactivity is highly prevalent among individuals with COPD and HF [1-5] and is associated with lower health-related quality of life (HRQL) [6,7], increased exacerbations [8], more frequent hospitalizations and readmissions [8-11], and increased mortality [9,12-16]. In COPD, specifically, physical activity was shown to be the strongest predictor of mortality [15]. Thus, physical activity monitoring and promotion are key treatment goals for both COPD and HF [17,18]. Conventional pulmonary and cardiac rehabilitation programs include supervised aerobic exercise training, often targeting moderate intensity exercise on treadmill or bicycle ergometer, that have been shown to reduce symptoms, improve HRQL, and increase exercise capacity [19-21] However, these programs are vastly underused [22,23] among individuals with COPD or HF and may not lead to significant improvements in long-term habitual physical activity in one's daily life outside of the formal rehabilitation setting [24-26]. While increasing regular physical activity is recommended, there is no consensus on optimal exercise modalities or regimens in COPD or HF, and studies support the continued adoption of general physical activity guidelines (including daily walking) in these populations [27,28]. There remains a critical need for highly accessible interventions that target processes relevant to sustaining daily physical activity for individuals with COPD and HF [29].

Moy et al [30,31] developed an accessible, web-based, pedometer-mediated intervention to promote walking behaviors among individuals with COPD called Every Steps Count (ESC) [30-37]. The 6-month intervention, based on self-regulation theories of health behavior change [38], included a web platform coupled with a pedometer, which provided individualized step-count goals, iterative feedback graphically and on the face of the pedometer, educational tips for disease self-management, motivational messages to help sustain walking, and a web-based community forum for social support. The intervention was safe, engaging, and associated with improvements in daily step counts and HRQL at 4 or 6 months compared with controls [30-33]. However, daily step count and HRQL improvements were not maintained at the 12-month follow-up [32]. Mindful Steps was thus designed to incorporate mind-body training as an adaptation of this original web-based intervention to further bolster and better sustain effects on walking.

Rationale, design, and conceptual model of the Mindful Steps intervention are described elsewhere [39]. In brief, Mindful Steps was developed to promote long-term walking behaviors by incorporating mind-body principles and content within the web-based, pedometer-mediated intervention. Mind-body components, such as mindful attention, breathing awareness/regulation, body awareness, and mindful movement, may support long-term physical activity by targeting key processes relevant to the initiation and maintenance of health behavior change. Indeed, mindfulness interventions may impact health behavior initiation and maintenance via changes in self-regulatory processes, such as emotion regulation, attention/cognitive control, self-related processes (eg, self-efficacy and interoceptive awareness), and motivation [40]. Mindful movement interventions, in particular (eg, tai chi and yoga), may serve as a gentle introduction to physical activity and simultaneously target processes needed to maintain physical activity, including improved physical function [41], exercise self-efficacy [42], enhanced body and self-awareness [43], and positive affective experiences during exercise [44-46]. Indeed, compared with nonmindful exercise, mindful exercise is associated with greater positive affect and enjoyment during exercise [44,47], which may promote continued physical activity engagement [48]. With the goal of targeting multiple pathways for behavior change, Mindful Steps combined adapted components from the ESC intervention (iterative feedback and goal-setting, educational and motivational messages, social support) together with new mind-body exercise (MBE) classes and a video library focused on mindful body awareness and movement, and cultivating intrinsic motivation and self-efficacy for walking. Moreover, Mindful Steps was expanded with additional educational content to be relevant for individuals with HF, in addition to COPD, given the similarities in patient phenotype, shared pathophysiology and symptoms, and need for physical activity intervention. While the ESC intervention did not include live content (ie, all components were asynchronous), we included live MBE classes to foster modeling; individuals may develop confidence and motivation from seeing others succeed (Litrownik et al [39] provides details of our theoretical model).

The aim of this pilot randomized controlled trial was to examine the feasibility and acceptability of Mindful Steps among individuals with COPD and HF and explore preliminary effect estimates on physical activity objectively measured as daily step counts, disease-specific HRQL, cognitive and psychosocial outcomes, and self-reported physical function compared with usual care. In line with guidelines for pilot trials [49], findings

XSL•FO
RenderX

from this study will inform the design of a fully-powered RCT to examine the efficacy of Mindful Steps for promoting and sustaining long-term adherence to walking in COPD and HF.

Methods

Study Design

Individuals with COPD or HF were randomized in a 2:1 ratio to a 12-month Mindful Steps intervention or usual care. Following the CONSORT (Consolidated Standards of Reporting Trials) guidelines for pilot trials [50], main outcomes included study feasibility and acceptability. Exploratory outcomes were included to inform hypotheses and the design of a future larger-scale RCT. Participants were assessed at baseline, 3, 6, 9, and 12 months. Study staff completing assessments were blinded to group assignment. Details of the study protocol are published [39] and briefly summarized below.

Ethical Considerations

The study was approved by the Beth Israel Deaconess Medical Center Institutional Review Board (IRB # 2016P000368), and all participants provided written informed consent. Study data were stored confidentially using REDCap (Research Electronic Data Capture; Vanderbilt University), which is Health Insurance Portability and Accountability Act–compliant, password-protected, and managed by Beth Israel Deaconess Medical Center, or stored on secure institutional servers. Participant privacy was maintained at all times. All participants were assigned a study ID. Participants received up to US $200 via check for study participation and kept their Fitbit devices (approximately US $80 in value).

Recruitment, Eligibility, and Randomization

Potential participants, identified through hospital registries, physician clinic schedules (cardiology and pulmonology), and specialist and primary care physician referrals, were sent an opt-out mailing and contacted by telephone by research assistants from February 2019 to May 2020. Participants were also recruited via flyers, digital screen ads in specialist and primary care clinic waiting rooms, and in-person presentations at local cardiac or pulmonary rehabilitation programs (for those who completed or no longer engaged in the program). Interested individuals completed a phone screen and scheduled an in-person consent visit.

Individuals were eligible if they met the following criteria: (1) older than 40 years; (2) clinical diagnosis of COPD, defined as either a ratio of forced expiratory volume in one second to forced vital capacity <0.70 or chest computed tomography evidence of emphysema, or clinical diagnosis of HF syndrome (with left ventricular systolic dysfunction or preserved ejection fraction, and New York Heart Association Class 1-3); (3) medical clearance from a provider to participate in an exercise program; (4) an active email account with the ability to check email at least weekly; and (5) access to a computer with an internet connection, USB port, and Windows. Individuals were excluded if they met any of the following criteria: (1) self-reported COPD or HF exacerbation in the previous 2 weeks, (2) inability to ambulate, (3) clinical signs of unstable cardiovascular disease, (4) hypoxemia during 6-minute walk test (oxygen saturation <85% with supplemental oxygen), (5) inability to collect at least 7 of 14 days of baseline step counts, and (6) current participation in a cardiac or pulmonary rehabilitation program.

Participants completed a 14-day run-in period where they were asked to wear the pedometer (Fitbit Alta HR; Fitbit Inc) all waking hours. Seven days of valid data (>200 steps per day) were required for randomization. Participants were then randomized in a 2:1 ratio to Mindful Steps or usual care using block randomization with varying/unpredictable block sizes. The random numbers were computer-generated and concealed by sealed and numbered opaque envelopes. After eligibility was confirmed, research assistants opened the next envelope in sequence and informed the participant of their assignment. The study was conducted in 4 waves (cohorts of ≈10 participants each). During the COVID-19 pandemic, all study activities, including recruitment, consent, and testing, were transitioned to all-remote procedures in March 2020.

Interventions

All participants continued to receive pharmacological treatment and usual care from their health care providers. In addition, all participants received an education booklet from Harvard Health Publishing titled “Walking for Health,” which describes the health benefits of walking, strategies for starting a walking program, and specific walking workouts [51].

Mindful Steps

This was a 12-month multicomponent web-based, pedometer-mediated intervention that included: (1) a pedometer (Fitbit Alta HR); (2) the Mindful Steps web platform with mind-body videos, motivational messages, educational content, a web-based community forum; and (3) live MBE classes.

The pedometer is paired with the Mindful Steps web platform to visually display the participant’s daily/weekly steps and individualized step goals. Participants were asked to upload their step counts to the study server at least once a week. The initial step-count goal was calculated using an algorithm designed to promote slow and safe increases, where the most recent 7 consecutive days of synced data were averaged and 400 steps were added. Previous research among individuals with type 2 diabetes used an 800-step increment based on participant feedback [52]. Subsequent research selected a 400-step increment in COPD as a more attainable goal in this population [31]. Subsequent goals could increase, decrease, or stay the same, but not go up by more than 400 steps over the previous goal and not exceed 10,000 steps for safety reasons. In the absence of new data, the prior week’s goal was reissued. Participants were asked to wear the wrist-worn pedometer daily during waking hours. Motivational messages and educational tips focused on self-managing HF and COPD and walking-related topics (eg, “Managing Stress and Anxiety,” “Handling Setbacks,” and “Preparing for Your Walk”). A new motivational message appeared on the web homepage weekly throughout the 12-month intervention (a total of 52) and a new educational tip every other day for the first 6 months (a total of 88; repeated in the final 6 months). Education content also embedded additional disease-specific resource links (eg, American Heart Association, Heart Failure Society of America,

XSL•FO
RenderX

and COPD Foundation). The web-based community forum allowed participants to post comments or respond to seeded posts from study staff to stimulate discussion (eg, “Share something new you noticed on your walk today”); participants were able to respond to other participants’ posts in these threads. To help promote interaction with the study website, the web platform included a visual star reward system. Participants would earn up to 3 stars daily for reaching their step-count goal, watching videos, and viewing the motivational messages and educational tips. Stars were retroactively awarded for meeting step-count goals even if participants did not upload their steps daily.

The Mindful Steps web platform also included 2 sets of mind-body videos that participants could access, which were designed to support and integrate with the live MBE classes (described below). The first set was a themed mindful walking video curriculum that included 26 videos mirroring the themes from the live MBE classes. A new mindful walking video was posted on the web homepage weekly for the first 6 months (repeated in the second 6 months). Each video included a brief didactic introduction to the theme and a guided MBE related to the theme. The second set was an MBE video library, including 13 videos demonstrating specific exercises taught in classes to aid in home practice. Participants were encouraged to use the website daily and watch the weekly mindful walking video.

Live MBE classes were 75 minutes and held weekly during the first 6 weeks of the intervention, biweekly for the next 34 weeks, and monthly for the last 12 weeks. Each class focused on a theme for promoting walking and included a guided meditative walk and a menu of MBEs used in our previous trials [53-56]. MBEs included mindful warm-up and stretching exercises, and traditional qigong movement and breathing exercises [39]. In-person MBE classes were transitioned to live videoconferencing via Starleaf (the hospital’s preferred meeting platform in 2020) during the COVID-19 pandemic in March 2020. Classes were scheduled at a time that was convenient for the group during regular business hours. There were opportunities for real-time interaction with instructors and other participants during the classes on videoconference.

The themes of the mind-body exercise classes and mindful walking video curriculum in the Mindful Steps intervention were (1) introduction: body, mind, and breath; (2) mindful warm-ups: lower body; (3) motivation to move; (4) putting joy back into exercise; (5) rewarding yourself with the gift of walking; (6) mindfulness in motion; (7) mindful warm-ups: upper body; (8) renew your body with your breath; (9) every step counts; (10) self-kindness; (11) overcoming barriers and challenges; (12) walking for your mind and spirit; (13) pain management; (14) preventing falls; (15) strength and flexibility; (16) breath awareness; (17) your body affects your mind; (18) stop and smell the roses; (19) relaxation and stress management; (20) exploring balance in walking; (21) belly breathing; (22) rest and recharge along the way; (23) importance of posture; (24) leg strength; (25) moderated effort: the 70% rule; and (26) mechanics of walking.

Outcome Measures

Recruitment and Retention

Study feasibility was assessed via recruitment and retention rates. The a priori benchmark for successful retention was 80%, or less than 20% dropouts.

Intervention Acceptability and Adherence

Acceptability was primarily assessed via qualitative exit interviews at 6 and 12 months. The qualitative results are published elsewhere [57]. In addition, participants were asked to rate weekly, using a sliding scale, the likeability (0=dislike to 100=really like) and helpfulness with walking (0=not at all to 100=very much) of each of the 26 videos in the mindful walking video curriculum. Average likeability and helpfulness ratings were calculated using the ratings obtained from the first 26 weeks.

Intervention adherence included MBE class adherence, engagement with the web platform, and use of the pedometer. The a priori definition of adherence to MBE classes was 70% attendance (17 of 25 classes). Home practice included practicing exercises from the video library or live classes and was tracked via weekly self-report surveys, including the number of minutes of practice, if any. Web platform use was automatically tracked and included the total number of pages visited within the website (ie, total of each individual click on a page on the website), total number of stars earned, and total/unique clicks on pages with educational tips and motivational messages. Video watch time was assessed via the number of minutes each participant watched any of the videos. Pedometer use was assessed via the total number of days with any step counts greater than 0.

Safety

Adverse events were systematically assessed at 3-, 6-, 9-, and 12-month testing visits. Participants were asked about specific symptoms, including muscle strain/soreness, dizziness, episodes of shortness of breath, falls, chest discomfort, and any other adverse events not specifically queried. Participants were also asked about emergency room visits and hospitalizations. In addition, participants were sent weekly surveys that asked them to report any symptoms or adverse events.

Exploratory Outcomes

Physical Activity

Daily Step Counts

Average daily step counts were measured by a Fitbit Alta HR monitor worn on the wrist. At each study visit, participants in usual care were asked to wear the pedometer for 14 days during waking hours except when showering/bathing. Similarly, those in the intervention group continued to wear the pedometer as usual. At each time point, data from 14 days starting at the testing visit were extracted for analysis. A wear day was considered valid if ≥200 steps were recorded. An average daily step count variable was created by averaging daily step counts across all valid wear days.

XSL•FO
RenderX

Community Healthy Activities Model Program for Seniors Physical Activity Questionnaire

The Community Healthy Activities Model Program for Seniors Physical Activity Questionnaire [58] is a 41-item self-report measure that assesses the frequency and duration of multiple activities across differing exercise intensities among older adults. The Community Healthy Activities Model Program for Seniors Physical Activity Questionnaire allows the calculation of weekly frequency and estimation of weekly caloric expenditure, which we calculated for both "all activity" and "moderate intensity activity."

Disease-Specific HRQL

Minnesota Living With Heart Failure Questionnaire

The Minnesota Living With Heart Failure Questionnaire [59] is a 21-item self-report measure assessing HRQL among individuals with HF. Items are rated on a 6-point Likert scale (0=no to 5=very much). This instrument was used only in study participants with HF.

St George's Respiratory Questionnaire

The St George's Respiratory Questionnaire [60] is a 50-item measure that assesses HRQL among individuals with respiratory illnesses. Total scores range from 0 to 100, with lower scores reflecting better HRQL. This instrument was used only in study participants with COPD.

Cognitive-Behavioral and Psychosocial

Self-Efficacy for Exercise Scale

The Self-Efficacy for Exercise Scale [61] is a 9-item self-report measure that assesses one's confidence for exercise in the face of barriers (eg, when feeling tired). Items are rated on a scale from 0 (not at all confident) to 10 (very confident).

Self-Efficacy for Managing Chronic Disease Scale

The Self-Efficacy for Managing Chronic Disease Scale [62] is a 6-item self-report measure that assesses one's confidence in managing their chronic illness. Items are rated on a scale from 1 (not at all confident) to 10 (totally confident).

Intrinsic Motivation Inventory

The Intrinsic Motivation Inventory (IMI) [63] is a self-report measure that assesses motivation for a specific activity. The 22-item version was used in this study, and the items were modified to assess the activity of walking. The 22-item version has 4 subscales: interest/enjoyment, perceived competence, perceived choice, and pressure/tension. Items are rated on a 7-point Likert scale (1=not at all to 7=very true).

Patient Activation Measure

The PAM [64] is a 13-item self-report measure that assesses one's skills, knowledge, and confidence for managing their chronic illness. Items are rated on a scale from 1 (strongly disagree) to 4 (agree strongly), and raw scores are converted into an activation score ranging from 0 to 100, with higher scores reflecting greater activation.

Medical Outcomes Study Social Support Survey

The Medical Outcomes Study Social Support Survey [65] is a 19-item self-report measure that assesses social support among individuals with chronic illnesses. Items are rated on 5-point Likert scale (1= none of the time to 5= all of the time). The scale includes a total score and 4 subscales assessing specific types of support: emotional/informational support, tangible support, affectionate support, and positive social interaction.

Multidimensional Assessment of Interoceptive Awareness

The Multidimensional Assessment of Interoceptive Awareness (MAIA) [66] is a 32-item self-report measure that assesses interoceptive body awareness. Items are rated 6-point Likert scale (0=never to 5=always). The 8 separately scored scales of the MAIA include: noticing, not-distracting, not-worrying, attention regulation, emotional awareness, self-regulation, body listening, and body trusting.

Center for Epidemiologic Studies Depression Scale

The Center for Epidemiologic Studies Depression Scale [67] is a 20-item self-report measure assessing symptoms of depression. Items are rated on a scale from 0 (rarely or none of the time) to 3 (most or almost all the time).

Physical Function

Physical function (10-item measure), fatigue (7-item measure), and pain interference (8-item measure) were assessed using the National Institutes of Health PROMIS (Patient-Reported Outcomes Measurement Information System) scales [68-71]. Higher scores reflect better physical function, greater fatigue, and more pain interference. Scores are calculated using a standardized T-score system with a population mean of 50 (SD 10).

All survey instruments were initially administered at in-person testing visits; subsequently, they were administered over Zoom or phone once the study transitioned to remote study procedures during the COVID-19 pandemic.

Statistical Analysis

Feasibility, Acceptability, Intervention Adherence, and Safety

Feasibility and adherence, including website engagement, were assessed using descriptive statistics, including means, SDs, medians, IQRs, and frequencies. For adherence, descriptive statistics were calculated for the first 6 months, the final 6 months, and the total 12 months of the intervention to assess patterns over time. Given that website usage data were objectively tracked (eg, clicks and video watch time), missing data are an absence of website engagement and therefore were entered as zero. Acceptability of the study videos was assessed by calculating average likeability and helpfulness ratings for each of the 26 videos and reporting the range of these ratings. Safety was assessed via the number of adverse events according to expectedness, relatedness to the study, and severity.

Preliminary Effect Estimates

In line with guidelines for pilot trials [72,73], we assessed changes in patient-centered outcomes, including

XSL•FO
RenderX

cognitive-emotional constructs from our theoretical model [39], to explore which measures may be sensitive to change and inform variability within and between groups to inform the design of future studies. For each patient-centered outcome, we computed the difference (between treatment arms) of differences (from baseline). We performed repeated measures analyses using the linear mixed effects models (SAS PROC MIXED) to evaluate trajectories over the course of the 4 time points for each group [74-76]. The estimated mean differences between the groups in these trajectories at each time point were obtained via interaction terms (treatment group × time) and linear contrasts. Cohen *d* effect sizes were computed to evaluate the magnitude of group differences in change from baseline.

Results

Recruitment and Retention

Recruitment and retention are detailed in the CONSORT diagram (Figure 1). Of 1444 individuals (reached out to via letter or in-person at rehabilitation program), we made further contact with 178 to assess interest and eligibility. Of those, 43 consented to participate, for an overall recruitment rate of 3%. Forty-one participants were randomized, and 31 completed the 12-month study (retention rate=76%). Of those randomized who did not complete the study, 1 dropped out due to time/scheduling, 2 were no longer interested in the study, 2 were deceased (1 due to COVID, 1 due to COVID-19 and metastatic cancer; both unrelated to the study), and 5 dropped out due to "other reasons."

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) flow diagram of recruitment, randomization, and retention in the Mindful Steps pilot randomized controlled trial among individuals with chronic obstructive pulmonary disease or heart failure.

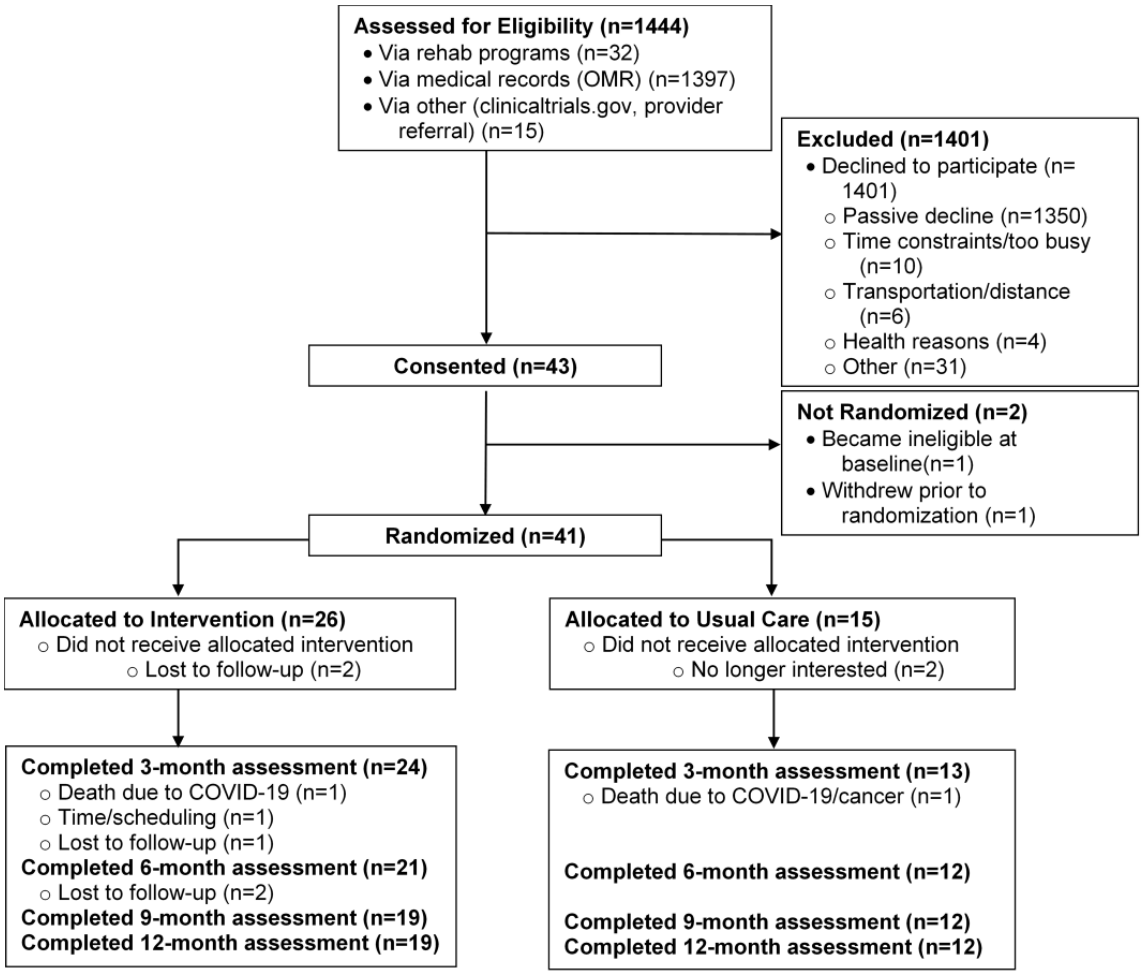

Baseline Characteristics

Baseline characteristics are summarized in Table 1. Participants were 56% (23/41) female with a mean age of 69.99 (SD 8.0) years. For qualifying diagnoses, 56% (23/41) of participants had COPD, 34% (14/41) had HF, and 10% (4/41) had a diagnosis of both COPD and HF. Most participants (36/41, 88%) self-identified as White, 7% (3/41) Black, 2% (1/41) Asian, and 2% (1/41) other. Regarding rehabilitation status, 22% (9/41) had previously completed pulmonary rehabilitation, and 7% (3/41) cardiac rehabilitation. The mean Charlson Comorbidity Index was 4.95 (SD 1.75), indicating fairly high morbidity in our sample (Charlson Comorbidity Index score 3-4=moderate, ≥5=severe [77]). The mean tobacco pack-years was 34.09 (SD 28.56).

XSL•FO
RenderX

Table 1. Baseline sociodemographic and clinical characteristics of study participants with chronic obstructive pulmonary disease and heart failure in the Mindful Steps pilot randomized controlled trial.

Characteristic	Intervention (n=26)	Control (n=15)	Total sample (N=41)
Age (years), mean (SD)	69.88 (8.1)	70.17 (8.1)	69.99 (8.0)
Sex (female), n (%)	14 (54)	9 (60)	23 (56)
Race, n (%)			
White	22 (85)	14 (93)	36 (88)
Black	2 (8)	1 (7)	3 (7)
Asian	1 (4)	0 (0)	1 (2)
Other	1 (4)	0 (0)	1 (2)
Annual income <US $35,000, n (%)	4 (15)	2 (13)	6 (15)
Married, n (%)	18 (69)	11 (73)	29 (71)
Unemployed, retired, or disabled, n (%)	22 (85)	7 (47)	29 (71)
Qualifying diagnosis, n (%)			
COPD	14 (54)	9 (60)	23 (56)
Chronic heart failure	8 (31)	6 (40)	14 (34)
Both	4 (15)	0 (0)	4 (10)
Charlson Comorbidity Index, mean (SD)	4.92 (1.81)	5.00 (1.69)	4.95 (1.75)
Regular oxygen use, n (%)	2 (8)	1 (7)	3 (7)
Smoking pack years, mean (SD)	32.06 (30.98)	37.96 (24.34)	34.09 (28.56)
Completed pulmonary rehabilitation, n (%)	8 (31)	1 (7)	9 (22)
Completed cardiac rehabilitation	2 (7.7)	1 (7)	3 (7)
Comorbidities, n (%)			
CVD[a] (CAD[b], Angina, Angioplasty, and MI[c])	3 (12)	2 (13)	5 (12)
Heart failure	12 (46)	6 (40)	18 (44)
Cancer	4 (15)	2 (13)	6 (15)
Hypertension	18 (69)	11 (73)	29 (71)
Limitation of limb (weakness, paralysis)	2 (8)	2 (13)	4 (10)
OA[d], sciatica, chronic back pain	16 (62)	9 (60)	25 (61)
Peripheral vascular disease	0 (0)	1 (7)	1 (2)
Stroke or cerebrovascular disease	1 (4)	0 (0)	1 (2)
Connective tissue disease	1 (4)	1 (7)	2 (5)
Obstructive sleep apnea	8 (31)	4 (27)	12 (29)
Kidney disease	3 (12)	4 (27)	7 (10)

[a]CVD: cardiovascular disease.
[b]CAD: coronary artery disease.
[c]MI: myocardial infarction.
[d]OA: osteoarthritis.

Intervention Acceptability and Adherence

Among participants who completed ratings (n range=15-22), average likeability ratings for the 26 weekly videos of the mindful walking curriculum ranged from 68.59 (SD 23.14) to 80.19 (SD 8.42), and average helpfulness for daily walking ratings ranged from 58.23 (SD 29.52) to 77.06 (SD 16.08).

In terms of intervention adherence, among those randomized to Mindful Steps, 58% of participants (15/26) attended 17 or more of the live MBE classes. Of those who completed the study 79% (15/19) attended 17 or more of the live MBE classes. The first 2 cohorts of intervention participants (cohort 1 n=4, cohort 2 n=7) had a “hybrid” experience with in-person classes early during the 12-month study period before converting to digital classes after the start of the COVID-19 pandemic (first

XSL•FO
RenderX

cohort=92%, 23/25 in-person classes; second cohort=52%, 13/25 in-person classes). The third and fourth cohorts (cohort 3: n=8; cohort 4: n=7) had all MBE classes conducted via videoconferencing. There were no significant differences in class attendance in-person or via web.

Table 2 presents home practice adherence in the first 6 months, the second 6 months, and the full 12 months of the study. Across 12 months, 22 participants completed home practice forms and reported "yes" to home practice in 32.05 (SD 19.92) weeks (range 1-52 weeks) and completed an average of 48.54 (SD 30.81) minutes of practice per week. The mean number of Fitbit wear days in the Mindful Steps group was 294.65 (SD 109.71) days of an expected maximum of 365 days.

Table 2. Home practice of mind-body exercise completion of participants with chronic obstructive pulmonary disease and heart failure, randomized to the Mindful Steps intervention.

	First 6 months (n=22)		Second 6 months (n=17)		Full 12 months (n=22)	
	Mean (SD)	Range	Mean (SD)	Range	Mean (SD)	Range
Total weeks reporting home practice	17.77 (9.29)	1-26	18.47 (9.17)	0-26	32.05 (19.92)	1-52
Home practice time (mins/week)	45.45 (27.43)	12.75-127.50	54.39 (39.92)	11.96-129.56	48.54 (30.81)	13-127.50

Website usage for the first 6 months, the second 6 months, and the total 12 months of the intervention is summarized in Table 3. For the total 12 months, participants (n=25) watched a mean 20.28 (SD 14.30; median 18.00, IQR 5.00-34.50) unique videos, with a mean 335.76 minutes (SD 515.35; median 103.40, IQR 27.31-413.55) of video watch time. Participants earned a mean of 295.64 (SD 281.45; median 214.00, IQR 104.50-411.00) stars throughout the 12 months. On average, 69% of stars were earned through reaching step goals, 17% from watching videos, and 14% from other content clicks on the website. The mean number of website pages visited throughout the 12 months was 1220.68 (SD 1400.53; median 563.00, IQR 203.50-2193.50). Throughout the 12-month intervention, participants visited a unique tip website page a mean of 22.80 (SD 28.21; median 10.00, IQR 0.50-39.00) times and a unique motivational message page a mean of 19.36 (SD 19.59; median 12.00, IQR 2.50-40.00) times. Ten participants (38%) made at least one comment in the web-based forum.

Table 3. Engagement with the Mindful Steps website components in a pilot randomized controlled trial among individuals with chronic obstructive pulmonary disease or heart failure.

	First 6 months (n=25)[a]			Second 6 months (n=25)[a]			Full 12 months (n=25)[a]		
	Mean (SD)	Median (IQR)	Range	Mean (SD)	Median (IQR)	Range	Mean (SD)	Median (IQR)	Range
Unique videos watched	18.48 (13.58)	15.00 (5.00-30.50)	0-39	11.04 (11.36)	5.00 (0.00-23.50)	0-30	20.28 (14.30)	18.00 (5.00-24.50)	0-40
Video watch time (mins)	200.30 (274.01)	91.45 (22.89-239.80)	0-1140	134.69 (250.66)	21.90 (0.00-138.51)	0-1103	335.76 (515.35)	103.40 (27.31-413.55)	0-2244
Stars earned	163.76 (138.28)	113.00 (70.00-231.00)	3-489	131.88 (148.23)	88.00 (24.50-173.50)	0-545	295.64 (281.45)	214.00 (104.50-411.00)	3-979
Nonunique clicks: any website page	726.96 (790.32)	376.00 (16.00-1464.50)	5-2484	488.40 (679.11)	215.00 (0.00-738.50)	0-2696	1220.68 (1400.53)	563.00 (203.50-2193.50)	5-4769
Unique clicks: educational tips	21.56 (27.85)	9.00 (0.50-36.50)	0-85	2.84 (4.06)	0.00 (0.00-5.00)	0-13	22.80 (28.21)	10.00 (0.50-39.00)	0-85
Unique clicks: motivational messages	10.80 (10.83)	6.00 (2.50-18.50)	0-35	10.32 (13.01)	2.00 (0.00-24.50)	0-41	19.36 (19.59)	12.00 (2.50-40.00)	0-52
Nonunique clicks: educational tips	38.52 (55.39)	12.00 (0.50-64.00)	0-159	3.60 (5.34)	0.00 (0.00-6.50)	0-15	42.48 (59.86)	13.00 (0.50-73.50)	0-177
Nonunique clicks: motivational messages	29.20 (36.90)	11.00 (2.50-63.00)	0-109	47.04 (78.74)	9.00 (0.00-58.00)	0-239	76.76 (110.67)	20.00 (2.50-108.50)	0-317

[a]Missing data are the absence of website engagement and therefore were entered as 0.

Adverse Events

One adverse event was considered reportable per the BIDMC Institutional Review Board, which was deemed serious and possibly related and included an episode of noncardiac chest pain in an intervention participant with HF upon arrival for 6-month testing. The patient was evaluated in the emergency department and discharged with possible musculoskeletal pain

XSL•FO
RenderX

and anxiety. Of 85 nonreportable events (55 in intervention, 30 in usual care), 3 were deemed expected and possibly related, including mild-moderate musculoskeletal pain during walking or MBE classes in the intervention group.

Preliminary Effect Estimates

Results on the effects of Mindful Steps and usual care on daily step counts are presented in Table 4. Figure 2 provides a graphical display of mean daily step counts at each timepoint. In mixed effect models, there were greater increases in daily step counts from baseline to 3 months (estimate=2038.77 steps per day between groups, 95% CI 288.76-3788.78), 6 months (estimate=3031.45, 95% CI 1261.15-4801.74), and 9 months (estimate=2703.80, 95% CI 862.97-4544.62) in the Mindful Steps group compared with usual care. Effect sizes for group differences in the change from baseline to each timepoint were medium-to-large (Cohen *d* range 0.73-1.18). The difference in the change in daily step counts between groups from baseline to 12 months was not significant, although the effect size is medium-to-large (Cohen *d*=0.73).

Table 4. Comparison of daily step counts between groups (Mindful Steps intervention vs usual care) among individuals with chronic obstructive pulmonary disease or heart failure.

	Mindful Steps (n=26)		Usual care (n=15)		Difference in differences (95% CI)[a,b]	*P* value[b]	Cohen *d* of difference in differences
	Mean (SD)[c]	Difference from baseline, mean (SD)	Mean (SD)[c]	Difference from baseline, mean (SD)			
Baseline	5180.74 (4013.46)	0 (0)	5710.44 (4130.39)	0 (0)	__[d]	—	—
3-month	5910.11 (4049.34)	1066.90 (2128.21)	5489.02 (3699.05)	–1074.49 (950.19)	2038.77 (288.76 to 3788.78)	.02	1.18
6-month	6628.06 (4325.51)	1320.68 (3351.20)	4537.91 (3648.83)	–1821.20 (1508.98)	3031.45 (1261.15 to 4801.74)	<.01	1.11
9-month	6139.62 (4162.30)	882.71 (3457.38)	4650.36 (3864.16)	–2033.88 (1999.38)	2703.80 (862.97 to 4544.62)	<.01	0.96
12-month	6149.94 (3739.57)	714.62 (2842.02)	5485.72 (4257.90)	–1333.01 (2789.37)	1427.94 (–454.45 to 3309.33)	.13	0.73

[a]Difference from baseline of Mindful Steps minus the difference from baseline of usual care.

[b]Linear Mixed Effects Model testing the difference of differences from baseline between Mindful Steps and usual care.

[c]Average daily step counts within a 14-day window around each timepoint.

[d]Not applicable.

Figure 2. Mean daily step counts with 95% CI at each assessment period in a pilot randomized controlled trial comparing Mindful Steps to usual care among individuals with chronic obstructive pulmonary disease or heart failure.

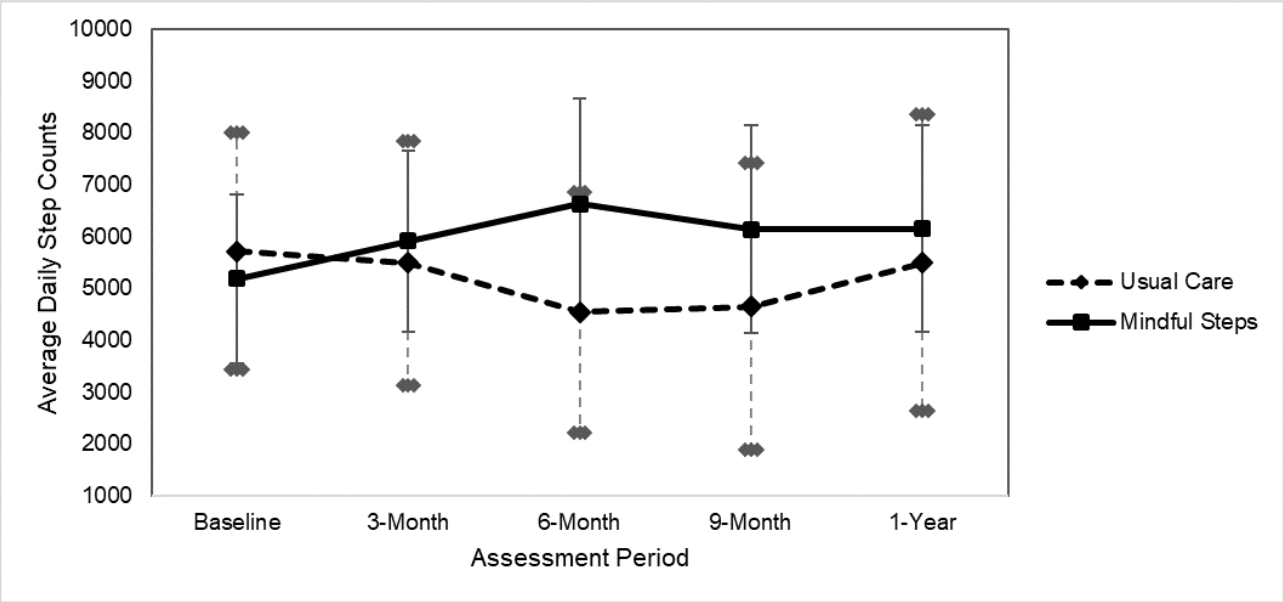

Results on the effects of Mindful Steps and usual care on all other exploratory outcomes are presented in Table S1 in Multimedia Appendix 1. In mixed effect models, there were greater improvements in MAIA body listening (estimate=1.16, 95% CI 0.25-2.07) and MAIA emotional awareness (estimate=0.88, 95% CI 0.15-1.61) from baseline to 3 months

XSL•FO
RenderX

in Mindful Steps compared with usual care. There were greater improvements in IMI-Interest (estimate=1.03, 95% CI 0.01-2.04) and IMI-Pressure (estimate=–1.59, 95% CI –2.55 to –0.63) from baseline to 6 months in Mindful Steps compared with usual care. There were greater improvements in exercise self-efficacy from baseline to 12 months in Mindful Steps compared with usual care (estimate=1.77, 95% CI 0.20-3.33). There were no significant group differences in the changes from baseline to any timepoint for all other outcomes.

Discussion

The aim of this study was to assess the feasibility and acceptability of a multicomponent web-based, pedometer-mediated, mind-body intervention to improve and sustain walking among individuals with COPD or HF. Findings suggest that the study was feasible. While the recruitment rate was slower than expected, the rate is similar to other reports in this population and reflects commonly cited barriers of time and transportation [78-80]. We also noted a significant number of passive declines, which largely reflects our recruitment process involving mass mailings from a hospital database. Once randomized, more than three-quarters of participants were retained in the study at 12 months, slightly below our predefined benchmark of 80%. Recruitment and retention were additionally hindered by the COVID-19 pandemic, which occurred midway through the study. There were no significant, unexpected safety concerns.

Adherence to the 12-month Mindful Steps intervention was adequate. Participants engaged most with the MBE classes, pedometer, and mind-body videos. Over half of all participants randomized to Mindful Steps achieved our predefined benchmark of 70% class attendance; as expected, the proportion of participants achieving this benchmark was much higher when excluding those lost to follow-up. On average, participants watched more than half of the available videos on the website, with moderate to good scores for likeability and helpfulness. Participants also demonstrated good compliance with the pedometer, wearing the Fitbit on most (≈295) days in a year. Engagement with motivational messages, tips, and the web-based forum on the website was less robust. Of note, for all website engagement metrics, the mean was higher than the median, suggesting that a proportion of high website engagers may have impacted overall averages. This intervention was designed to better support walking up to 12 months, and results may suggest modest maintenance of step gains, but in this pilot feasibility study, we are unable to compare with the ESC intervention with pedometer and web-platform alone without mind-body content [32]. Nonetheless, there were still notable declines in website engagement in the second 6 months. The exception to this was clicks on motivational messages, which were similar or higher in the second compared with the first 6 months. Overall, these engagement results highlight the known challenges and overall difficulty engaging patients with COPD and HF long-term in behavioral interventions [32,81]. Future research could examine additional strategies to enhance website engagement (eg, leaderboards and the ability to comment on others' achievements).

The intervention engagement data are consistent with our published qualitative data [57]. In individual qualitative interviews at 6 months, participants endorsed the pedometer with feedback, MBE classes, and mind-body videos as the most helpful components of the intervention. In developing this multimodal intervention, we purposefully included multiple components, which provided a multipronged approach to behavior change that targeted multiple cognitive-behavioral constructs based on our conceptual model [39]. In addition, on the web-platform alone, there were several elements to engage with, as we recognized that participant engagement with the varied content might differ based on individual preferences and needs. To further optimize the intervention, future work might use innovative designs (eg, using the Multiphase Optimization Strategy framework) to better understand the contribution of individual components, and examine whether engagement with different components predicts physical activity outcomes [82]. Factorial designs could also inform whether the mind-body content in Mindful Steps added benefit beyond the ESCs intervention. Further research might also explore the declines in website engagement in the later periods of the intervention, and also the relationship between declines in website engagement and the rate of daily step count change. It is possible that participants have built confidence in the first 6 months to practice/implement mind-body strategies on their own, only needing to return to the website intermittently to review or reinforce a skill, as step-count data support a positive effect on physical activity that is strongest at 6 months but still durable at 12 months. Alternatively, it is possible that more consistent website engagement may maximally target physical activity outcomes, but website fatigue or other barriers interfere with engagement, thus limiting the overall intervention effects. While our intention was to include a live component (ie, MBE classes) to foster modeling and connectedness, future research might examine the impact of this live component on scalability and explore additional avenues for targeting these processes.

Consistent with guidelines for pilot trials [72,73], the aim of this study was to identify outcomes sensitive to change to inform the design of a future larger trial. Importantly, daily step counts were sensitive to change in this study, and there were indeed medium-to-large between-group differences in change in step counts from baseline to 3-, 6-, and 9-months, favoring Mindful Steps. While some of the initial improvements waned in the final 6 months, there was still a relative increase from baseline to 12 months, both within the intervention group and compared with usual care. Given the pilot nature of this study, future larger trials are needed to examine the efficacy of Mindful Steps for improving daily step counts.

In addition, we found that important constructs from our theoretical model [39], including indices of interoceptive awareness, internal motivation, and exercise self-efficacy, were sensitive to change. We saw medium-to-large changes in interoceptive awareness and internal motivation in the first several months of the intervention, which were explicit intervention targets in our conceptual model. Interoceptive and body awareness are inherently cultivated in all of the mindful movements presented in Mindful Steps. One of the main recurring themes of the mindful walking curriculum and MBE

XSL•FO
RenderX

classes emphasized internal motivation (eg, "Motivation to Move," "Putting Joy Back Into Exercise," and "Rewarding Yourself with the Gift of Walking"), which occurred early in the intervention. Thus, as expected, participants in Mindful Steps reported more motivation due to interest or enjoyment in the activity and less due to external pressures within the first 6 months. These findings are also consistent with our published qualitative data [57], where participants noted body awareness as a strategy that helped with their walking and a shift in walking enjoyment and internal motivations for walking. Interestingly, differences between groups in self-efficacy did not emerge until 12 months. It is possible that longer-term practice with mind-body strategies and walking is needed to move the needle on self-efficacy, particularly in HF and COPD populations that have longstanding fears of physical activity due to breathlessness [83].

Several study limitations are worth noting. First, as this was a small pilot study, results should be used to inform the design and hypotheses for a future larger study and should not be interpreted regarding efficacy. Second, while we intended a more even mix, we enrolled more patients with COPD than HF; however, we do not believe this impacts the overall generalizability to a chronic cardiopulmonary disease population. Given the sample size and scope of this trial, we are not able to do subgroup analyses to determine if intervention effects are different among the 2 conditions. Similarly, we did not collect information on time from cardiac or pulmonary rehabilitation for those with prior exposure, and recognize that some subgroups may be differentially motivated to exercise. Our study also included patients before and after the start of the COVID-19 pandemic. Third, while we gathered substantial data on the use of individual components, our study design and nature of the intervention ("multicomponent package") do not allow us to more definitively tease apart the effects of individual components. Fourth, we did not include eligibility criteria related to baseline physical activity levels. Future work should explore who benefits most from this intervention (eg, those with lower baseline physical activity levels) and whether baseline physical activity moderates one's response to the intervention. Fifth, given that Mindful Steps did not focus on achieving a particular walking intensity, we did not collect reliable information on objective or subjective walking intensity throughout the intervention. Future research might examine the impact of Mindful Steps on other indices of physical activity, such as moderate-to-vigorous physical activity or sedentary time. Last, due to the COVID-19 pandemic, our trial procedures were necessarily changed midstudy to accommodate remote assessments and classes via videoconference. This change, however, provided valuable experience with remote study conduct in this population and strongly supported feasibility, increased access, and patient convenience (previously discussed in qualitative analysis) [57]. For example, participants emphasized the convenience of digital MBE classes, and also highlighted opportunities to foster more connection and discussion with peers and instructors [57].

Overall, the 12-month Mindful Steps intervention was feasible and acceptable, adherence was adequate, and a transition from in-person classes and testing visits to fully remote was feasible without significant challenges. We also found positive signals and sensitivity to change with outcomes of step counts, internal motivation, interoceptive awareness, and exercise self-efficacy. These findings suggest that a larger efficacy trial to examine the effects of Mindful Steps on promoting physical activity engagement and walking in patients with COPD and HF is warranted, and that fully remote conduct could be considered.

Acknowledgments

This study was supported by an award from the National Center for Complementary and Integrative Health (NCCIH) at the National Institutes of Health (R34AT009354). KMK was supported by a grant (K23AT011043). GYY was supported by a grant (K24AT009465). PMW was supported by a grant (K24AT009282).

Data Availability

The datasets from this study are not publicly available due to concerns of participant confidentiality, but can be made available upon reasonable request to the principal investigator.

Authors' Contributions

GYY and MLM contributed to the conceptualization of the study. RK, NB, and PMR were responsible for data curation, while LN, KMK, and PMR performed the formal analysis. GYY acquired funding for the project and, together with DL, conducted the investigation. GYY, PMW, CRR, and MLM developed the methodology. Project administration was carried out by DL and GYY, who also provided supervision. KMK, GYY, and DL contributed to visualization. The original draft was written by KMK and GYY, and all authors—KMK, GYY, DL, PMW, CRR, NB, RK, PMR, LN, and MLM—contributed to the review and editing of the manuscript.

Conflicts of Interest

PMW is the founder and sole owner of the Tree of Life Tai Chi Center. PMW's interests were reviewed and managed by Mass General Brigham in accordance with their conflict of interest policies. The remaining authors declared no conflicts of interest.

XSL•FO
RenderX

Multimedia Appendix 1
Difference in exploratory outcomes between Mindful Steps and usual care.
[DOCX File , 65 KB - formative_v9i1e74118_app1.docx]

Multimedia Appendix 2
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 1133 KB - formative_v9i1e74118_app2.pdf]

References

1. Dontje ML, van der Wal MHL, Stolk RP, Brügemann J, Jaarsma T, Wijtvliet PEPJ, et al. Daily physical activity in stable heart failure patients. J Cardiovasc Nurs 2014;29(3):218-226 [FREE Full text] [doi: 10.1097/JCN.0b013e318283ba14] [Medline: 23416939]
2. O'Donnell J, Smith-Byrne K, Velardo C, Conrad N, Salimi-Khorshidi G, Doherty A, et al. Self-reported and objectively measured physical activity in people with and without chronic heart failure: UK Biobank analysis. Open Heart 2020;7(1):e001099 [FREE Full text] [doi: 10.1136/openhrt-2019-001099] [Medline: 32153787]
3. Moy ML, Danilack VA, Weston NA, Garshick E. Daily step counts in a US cohort with COPD. Respir Med 2012;106(7):962-969 [FREE Full text] [doi: 10.1016/j.rmed.2012.03.016] [Medline: 22521225]
4. Waschki B, Kirsten AM, Holz O, Mueller K, Schaper M, Sack A, et al. Disease progression and changes in physical activity in patients with chronic obstructive pulmonary disease. Am J Respir Crit Care Med 2015;192(3):295-306. [doi: 10.1164/rccm.201501-0081OC] [Medline: 26020495]
5. Van Remoortel H, Hornikx M, Demeyer H, Langer D, Burtin C, Decramer M, et al. Daily physical activity in subjects with newly diagnosed COPD. Thorax 2013;68(10):962-963 [FREE Full text] [doi: 10.1136/thoraxjnl-2013-203534] [Medline: 23604460]
6. Bunyamin V, Spaderna H, Weidner G. Health behaviors contribute to quality of life in patients with advanced heart failure independent of psychological and medical patient characteristics. Qual Life Res 2013;22(7):1603-1611. [doi: 10.1007/s11136-012-0312-6] [Medline: 23161327]
7. Esteban C, Quintana JM, Aburto M, Moraza J, Egurrola M, Pérez-Izquierdo J, et al. Impact of changes in physical activity on health-related quality of life among patients with COPD. Eur Respir J 2010;36(2):292-300 [FREE Full text] [doi: 10.1183/09031936.00021409] [Medline: 20075059]
8. Moy ML, Teylan M, Weston NA, Gagnon DR, Garshick E. Daily step count predicts acute exacerbations in a US cohort with COPD. PLoS One 2013;8(4):e60400 [FREE Full text] [doi: 10.1371/journal.pone.0060400] [Medline: 23593211]
9. Hegde SM, Claggett B, Shah AM, Lewis EF, Anand I, Shah SJ, et al. Physical activity and prognosis in the TOPCAT trial (Treatment of Preserved Cardiac Function Heart Failure With an Aldosterone Antagonist). Circulation 2017;136(11):982-992 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.117.028002] [Medline: 28637881]
10. Nguyen HQ, Chu L, Amy Liu IL, Lee JS, Suh D, Korotzer B, et al. Associations between physical activity and 30-day readmission risk in chronic obstructive pulmonary disease. Ann Am Thorac Soc 2014;11(5):695-705. [doi: 10.1513/AnnalsATS.201401-017OC] [Medline: 24713094]
11. Waring T, Gross K, Soucier R, ZuWallack R. Measured physical activity and 30-day rehospitalization in heart failure patients. J Cardiopulm Rehabil Prev 2017;37(2):124-129. [doi: 10.1097/HCR.0000000000000204] [Medline: 27798506]
12. Cacciatore F, Amarelli C, Ferrara N, Della Valle E, Curcio F, Liguori I, et al. Protective effect of physical activity on mortality in older adults with advanced chronic heart failure: a prospective observational study. Eur J Prev Cardiol 2019;26(5):481-488. [doi: 10.1177/2047487318790822] [Medline: 30066588]
13. Moy ML, Gould MK, Liu IA, Lee JS, Nguyen HQ. Physical activity assessed in routine care predicts mortality after a COPD hospitalisation. ERJ Open Res 2016;2(1) [FREE Full text] [doi: 10.1183/23120541.00062-2015] [Medline: 27730174]
14. Stewart RAH, Held C, Hadziosmanovic N, Armstrong PW, Cannon CP, Granger CB, STABILITY Investigators. Physical activity and mortality in patients with stable coronary heart disease. J Am Coll Cardiol 2017;70(14):1689-1700 [FREE Full text] [doi: 10.1016/j.jacc.2017.08.017] [Medline: 28958324]
15. Waschki B, Kirsten A, Holz O, Müller KC, Meyer T, Watz H, et al. Physical activity is the strongest predictor of all-cause mortality in patients with COPD: a prospective cohort study. Chest 2011;140(2):331-342. [doi: 10.1378/chest.10-2521] [Medline: 21273294]
16. Doukky R, Mangla A, Ibrahim Z, Poulin M, Avery E, Collado FM, et al. Impact of physical inactivity on mortality in patients with heart failure. Am J Cardiol 2016;117(7):1135-1143 [FREE Full text] [doi: 10.1016/j.amjcard.2015.12.060] [Medline: 26853954]
17. Vestbo J, Hurd SS, Agustí AG, Jones PW, Vogelmeier C, Anzueto A, et al. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: GOLD executive summary. Am J Respir Crit Care Med 2013;187(4):347-365. [doi: 10.1164/rccm.201204-0596PP] [Medline: 22878278]
18. Authors/Task Force Members:, McDonagh TA, Metra M, Adamo M, Gardner RS, Baumbach A, ESC Scientific Document Group. 2021 ESC guidelines for the diagnosis and treatment of acute and chronic heart failure: developed by the task force

XSL•FO
RenderX

for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC). with the special contribution of the Heart Failure Association (HFA) of the ESC. Eur J Heart Fail 2022;24(1):4-131 [FREE Full text] [doi: 10.1002/ejhf.2333] [Medline: 35083827]
19. Long L, Mordi IR, Bridges C, Sagar VA, Davies EJ, Coats AJ, et al. Exercise-based cardiac rehabilitation for adults with heart failure. Cochrane Database Syst Rev 2019;1(1):CD003331 [FREE Full text] [doi: 10.1002/14651858.CD003331.pub5] [Medline: 30695817]
20. Bozkurt B, Fonarow GC, Goldberg LR, Guglin M, Josephson RA, Forman DE, ACC's Heart FailureTransplant SectionLeadership Council. Cardiac rehabilitation for patients with heart failure: JACC expert panel. J Am Coll Cardiol 2021;77(11):1454-1469 [FREE Full text] [doi: 10.1016/j.jacc.2021.01.030] [Medline: 33736829]
21. McCarthy B, Casey D, Devane D, Murphy K, Murphy E, Lacasse Y. Pulmonary rehabilitation for chronic obstructive pulmonary disease. Cochrane Database Syst Rev 2015;2015(2):CD003793 [FREE Full text] [doi: 10.1002/14651858.CD003793.pub3] [Medline: 25705944]
22. Ritchey MD, Maresh S, McNeely J, Shaffer T, Jackson SL, Keteyian SJ, et al. Tracking cardiac rehabilitation participation and completion among medicare beneficiaries to inform the efforts of a national initiative. Circ Cardiovasc Qual Outcomes 2020;13(1):e005902 [FREE Full text] [doi: 10.1161/CIRCOUTCOMES.119.005902] [Medline: 31931615]
23. Spitzer KA, Stefan MS, Priya A, Pack QR, Pekow PS, Lagu T, et al. Participation in pulmonary rehabilitation after hospitalization for chronic obstructive pulmonary disease among medicare beneficiaries. Ann Am Thorac Soc 2019;16(1):99-106 [FREE Full text] [doi: 10.1513/AnnalsATS.201805-332OC] [Medline: 30417670]
24. Cindy Ng LW, Mackney J, Jenkins S, Hill K. Does exercise training change physical activity in people with COPD? a systematic review and meta-analysis. Chron Respir Dis 2012;9(1):17-26 [FREE Full text] [doi: 10.1177/1479972311430335] [Medline: 22194629]
25. Dibben GO, Dalal HM, Taylor RS, Doherty P, Tang LH, Hillsdon M. Cardiac rehabilitation and physical activity: systematic review and meta-analysis. Heart 2018;104(17):1394-1402 [FREE Full text] [doi: 10.1136/heartjnl-2017-312832] [Medline: 29654095]
26. Egan C, Deering BM, Blake C, Fullen BM, McCormack NM, Spruit MA, et al. Short term and long term effects of pulmonary rehabilitation on physical activity in COPD. Respir Med 2012;106(12):1671-1679 [FREE Full text] [doi: 10.1016/j.rmed.2012.08.016] [Medline: 23063203]
27. Mueller S, Winzer EB, Duvinage A, Gevaert AB, Edelmann F, Haller B, OptimEx-Clin Study Group. Effect of high-intensity interval training, moderate continuous training, or guideline-based physical activity advice on peak oxygen consumption in patients with heart failure with preserved ejection fraction: a randomized clinical trial. JAMA 2021;325(6):542-551 [FREE Full text] [doi: 10.1001/jama.2020.26812] [Medline: 33560320]
28. F Piepoli M. 2016 European guidelines on cardiovascular disease prevention in clinical practice : the sixth joint task force of the European society of cardiology and other societies on cardiovascular disease prevention in clinical practice (constituted by representatives of 10 societies and by invited experts). Int J Behav Med 2017;24(3):321-419. [doi: 10.1007/s12529-016-9583-6] [Medline: 28290077]
29. Vitacca M, Paneroni M. Rehabilitation of patients with coexisting COPD and heart failure. COPD 2018;15(3):231-237 [FREE Full text] [doi: 10.1080/15412555.2018.1468427] [Medline: 29799285]
30. Moy ML, Collins RJ, Martinez CH, Kadri R, Roman P, Holleman RG, et al. An internet-mediated pedometer-based program improves health-related quality-of-life domains and daily step counts in COPD: a randomized controlled trial. Chest 2015;148(1):128-137 [FREE Full text] [doi: 10.1378/chest.14-1466] [Medline: 25811395]
31. Moy ML, Weston NA, Wilson EJ, Hess ML, Richardson CR. A pilot study of an internet walking program and pedometer in COPD. Respir Med 2012;106(9):1342-1350 [FREE Full text] [doi: 10.1016/j.rmed.2012.06.013] [Medline: 22795984]
32. Moy ML, Martinez CH, Kadri R, Roman P, Holleman RG, Kim HM, et al. Long-term effects of an internet-mediated pedometer-based walking program for chronic obstructive pulmonary disease: randomized controlled trial. J Med Internet Res 2016;18(8):e215 [FREE Full text] [doi: 10.2196/jmir.5622] [Medline: 27502583]
33. Robinson SA, Cooper JA, Goldstein RL, Polak M, Cruz Rivera PN, Gagnon DR, et al. A randomised trial of a web-based physical activity self-management intervention in COPD. ERJ Open Res 2021;7(3) [FREE Full text] [doi: 10.1183/23120541.00158-2021] [Medline: 34476247]
34. Wan ES, Kantorowski A, Homsy D, Teylan M, Kadri R, Richardson CR, et al. Promoting physical activity in COPD: insights from a randomized trial of a web-based intervention and pedometer use. Respir Med 2017;130:102-110 [FREE Full text] [doi: 10.1016/j.rmed.2017.07.057] [Medline: 29206627]
35. Wan ES, Kantorowski A, Polak M, Kadri R, Richardson CR, Gagnon DR, et al. Long-term effects of web-based pedometer-mediated intervention on COPD exacerbations. Respir Med 2020;162:105878 [FREE Full text] [doi: 10.1016/j.rmed.2020.105878] [Medline: 32056676]
36. Bamonti PM, Perndorfer C, Robinson SA, Mongiardo MA, Wan ES, Moy ML. Depression symptoms and physical activity in veterans with COPD: insights from a web-based, pedometer-mediated physical activity intervention. Ann Behav Med 2023;57(10):855-865. [doi: 10.1093/abm/kaad026] [Medline: 37260290]

XSL•FO
RenderX

37. Robinson SA, Shimada SL, Quigley KS, Moy ML. A web-based physical activity intervention benefits persons with low self-efficacy in COPD: results from a randomized controlled trial. J Behav Med 2019;42(6):1082-1090 [FREE Full text] [doi: 10.1007/s10865-019-00042-3] [Medline: 30980223]
38. Cameron L, Leventhal H. The Self-Regulation of Health and Illness Behaviour. London: Routledge; 2003.
39. Litrownik D, Gilliam EA, Wayne PM, Richardson CR, Kadri R, Rist PM, et al. Development of a novel intervention (Mindful Steps) to promote long-term walking behavior in chronic cardiopulmonary disease: protocol for a randomized controlled trial. JMIR Res Protoc 2021;10(4):e27826 [FREE Full text] [doi: 10.2196/27826] [Medline: 33913819]
40. Schuman-Olivier Z, Trombka M, Lovas DA, Brewer JA, Vago DR, Gawande R, et al. Mindfulness and behavior change. Harv Rev Psychiatry 2020;28(6):371-394 [FREE Full text] [doi: 10.1097/HRP.0000000000000277] [Medline: 33156156]
41. Chen Y, Hunt MA, Campbell KL, Peill K, Reid WD. The effect of Tai Chi on four chronic conditions-cancer, osteoarthritis, heart failure and chronic obstructive pulmonary disease: a systematic review and meta-analyses. Br J Sports Med 2016;50(7):397-407. [doi: 10.1136/bjsports-2014-094388] [Medline: 26383108]
42. Li F, McAuley E, Harmer P, Duncan T, Chaumeton N. Tai Chi enhances self-efficacy and exercise behavior in older adults. J Aging Phys Act 2001;9(2):161-171 [FREE Full text] [doi: 10.1123/japa.9.2.161]
43. Rivest-Gadbois E, Boudrias M. What are the known effects of yoga on the brain in relation to motor performances, body awareness and pain? a narrative review. Complement Ther Med 2019;44:129-142. [doi: 10.1016/j.ctim.2019.03.021] [Medline: 31126545]
44. Bigliassi M, Galano BM, Lima-Silva AE, Bertuzzi R. Effects of mindfulness on psychological and psychophysiological responses during self-paced walking. Psychophysiology 2020;57(4):e13529. [doi: 10.1111/psyp.13529] [Medline: 31953844]
45. Cox A, Ullrich-French S, Hargreaves E, McMahon A. The effects of mindfulness and music on affective responses to self-paced treadmill walking. Sport Exerc Perform Psychol 2020;9(4):571-584 [FREE Full text] [doi: 10.1037/spy0000192]
46. Gillman AS, Bryan AD. Mindfulness versus distraction to improve affective response and promote cardiovascular exercise behavior. Ann Behav Med 2020;54(6):423-435 [FREE Full text] [doi: 10.1093/abm/kaz059] [Medline: 31859347]
47. Cox AE, Roberts MA, Cates HL, McMahon AK. Mindfulness and affective responses to treadmill walking in individuals with low intrinsic motivation to exercise. Int J Exerc Sci 2018;11(5):609-624 [FREE Full text] [doi: 10.70252/AEVJ8865] [Medline: 29541336]
48. Rhodes RE, Kates A. Can the affective response to exercise predict future motives and physical activity behavior? a systematic review of published evidence. Ann Behav Med 2015;49(5):715-731. [doi: 10.1007/s12160-015-9704-5] [Medline: 25921307]
49. Leon AC, Davis LL, Kraemer HC. The role and interpretation of pilot studies in clinical research. J Psychiatr Res 2011;45(5):626-629 [FREE Full text] [doi: 10.1016/j.jpsychires.2010.10.008] [Medline: 21035130]
50. Eldridge SM, Chan CL, Campbell MJ, Bond CM, Hopewell S, Thabane L, PAFS consensus group. CONSORT 2010 statement: extension to randomised pilot and feasibility trials. BMJ 2016;355:i5239 [FREE Full text] [doi: 10.1136/bmj.i5239] [Medline: 27777223]
51. Harvard HP. Walking for health. Harvard Health Publishing. 2015. URL: https://www.health.harvard.edu/exercise-and-fitness/walking-for-health [accessed 2025-06-02]
52. Richardson C, Mehari K, McIntyre L, Janney A, Fortlage L, Sen A, et al. A randomized trial comparing structured and lifestyle goals in an internet-mediated walking program for people with type 2 diabetes. Int J Behav Nutr Phys Act 2007;4:59 [FREE Full text] [doi: 10.1186/1479-5868-4-59] [Medline: 18021411]
53. Yeh GY, McCarthy EP, Wayne PM, Stevenson LW, Wood MJ, Forman D, et al. Tai chi exercise in patients with chronic heart failure: a randomized clinical trial. Arch Intern Med 2011;171(8):750-757 [FREE Full text] [doi: 10.1001/archinternmed.2011.150] [Medline: 21518942]
54. Yeh GY, Roberts DH, Wayne PM, Davis RB, Quilty MT, Phillips RS. Tai chi exercise for patients with chronic obstructive pulmonary disease: a pilot study. Respir Care 2010;55(11):1475-1482 [FREE Full text] [Medline: 20979675]
55. Moy ML, Wayne PM, Litrownik D, Beach D, Klings ES, Davis RB, et al. Long-term exercise after pulmonary rehabilitation (LEAP): a pilot randomised controlled trial of Tai Chi in COPD. ERJ Open Res 2021;7(3) [FREE Full text] [doi: 10.1183/23120541.00025-2021] [Medline: 34262967]
56. Yeh GY, Litrownik D, Wayne PM, Beach D, Klings ES, Reyes Nieva H, et al. BEAM study (breathing, education, awareness, movement): a randomised controlled feasibility trial of tai chi exercise in patients with COPD. BMJ Open Respir Res 2020;7(1):e000697 [FREE Full text] [doi: 10.1136/bmjresp-2020-000697] [Medline: 33219007]
57. Kraemer KM, Kilgore K, Litrownik D, Jean-Laurent B, Wayne PM, Richardson CR, et al. A web-based mind-body intervention (Mindful Steps) for promoting walking in chronic cardiopulmonary disease: insights from a qualitative study. Glob Adv Integr Med Health 2023;12:27536130231212169 [FREE Full text] [doi: 10.1177/27536130231212169] [Medline: 38050584]
58. Stewart AL, Mills KM, King AC, Haskell WL, Gillis D, Ritter PL. CHAMPS physical activity questionnaire for older adults: outcomes for interventions. Med Sci Sports Exerc 2001;33(7):1126-1141. [doi: 10.1097/00005768-200107000-00010] [Medline: 11445760]

XSL•FO
RenderX

59. Rector TS, Cohn JN. Assessment of patient outcome with the Minnesota living with heart failure questionnaire: reliability and validity during a randomized, double-blind, placebo-controlled trial of pimobendan. pimobendan multicenter research group. Am Heart J 1992;124(4):1017-1025. [doi: 10.1016/0002-8703(92)90986-6] [Medline: 1529875]
60. Jones PW, Quirk FH, Baveystock CM, Littlejohns P. A self-complete measure of health status for chronic airflow limitation. the St George's respiratory questionnaire. Am Rev Respir Dis 1992;145(6):1321-1327. [doi: 10.1164/ajrccm/145.6.1321] [Medline: 1595997]
61. Resnick B, Jenkins LS. Testing the reliability and validity of the self-efficacy for exercise scale. Nurs Res 2000;49(3):154-159. [doi: 10.1097/00006199-200005000-00007] [Medline: 10882320]
62. Lorig KR, Sobel DS, Ritter PL, Laurent D, Hobbs M. Effect of a self-management program on patients with chronic disease. Eff Clin Pract 2001;4(6):256-262. [Medline: 11769298]
63. McAuley E, Duncan T, Tammen VV. Psychometric properties of the intrinsic motivation inventory in a competitive sport setting: a confirmatory factor analysis. Res Q Exerc Sport 1989;60(1):48-58. [doi: 10.1080/02701367.1989.10607413] [Medline: 2489825]
64. Hibbard JH, Stockard J, Mahoney ER, Tusler M. Development of the patient activation measure (PAM): conceptualizing and measuring activation in patients and consumers. Health Serv Res 2004;39(4 Pt 1):1005-1026 [FREE Full text] [doi: 10.1111/j.1475-6773.2004.00269.x] [Medline: 15230939]
65. Sherbourne CD, Stewart AL. The MOS social support survey. Soc Sci Med 1991;32(6):705-714. [doi: 10.1016/0277-9536(91)90150-b] [Medline: 2035047]
66. Mehling W, Price C, Daubenmier J, Acree M, Bartmess E, Stewart A. The multidimensional assessment of interoceptive awareness (MAIA). PLoS One 2012;7(11):e48230 [FREE Full text] [doi: 10.1371/journal.pone.0048230] [Medline: 23133619]
67. Radloff LS. The CES-D scale: A self-report depression scale for research in the general population. Appl Psychol Meas 1977;1(3):385-3401 [FREE Full text] [doi: 10.1177/014662167700100]
68. Rose M, Bjorner JB, Becker J, Fries JF, Ware JE. Evaluation of a preliminary physical function item bank supported the expected advantages of the patient-reported outcomes measurement information system (PROMIS). J Clin Epidemiol 2008;61(1):17-33. [doi: 10.1016/j.jclinepi.2006.06.025] [Medline: 18083459]
69. Rose M, Bjorner JB, Gandek B, Bruce B, Fries JF, Ware JE. The PROMIS physical function item bank was calibrated to a standardized metric and shown to improve measurement efficiency. J Clin Epidemiol 2014;67(5):516-526 [FREE Full text] [doi: 10.1016/j.jclinepi.2013.10.024] [Medline: 24698295]
70. Amtmann D, Cook KF, Jensen MP, Chen W, Choi S, Revicki D, et al. Development of a PROMIS item bank to measure pain interference. Pain 2010;150(1):173-182 [FREE Full text] [doi: 10.1016/j.pain.2010.04.025] [Medline: 20554116]
71. Lai J, Cella D, Choi S, Junghaenel DU, Christodoulou C, Gershon R, et al. How item banks and their application can influence measurement practice in rehabilitation medicine: a PROMIS fatigue item bank example. Arch Phys Med Rehabil 2011;92(10 Suppl):S20-S27 [FREE Full text] [doi: 10.1016/j.apmr.2010.08.033] [Medline: 21958919]
72. Teresi JA, Yu X, Stewart AL, Hays RD. Guidelines for designing and evaluating feasibility pilot studies. Med Care 2022;60(1):95-103 [FREE Full text] [doi: 10.1097/MLR.0000000000001664] [Medline: 34812790]
73. Kistin C, Silverstein M. Pilot studies: A critical but potentially misused component of interventional research. JAMA 2015;314(15):1561-1562 [FREE Full text] [doi: 10.1001/jama.2015.10962] [Medline: 26501530]
74. Lindstrom M, Bates D. Newton—raphson and EM algorithms for linear mixed-effects models for repeated-measures data. Journal of the American Statistical Association 1988;83(404):1014-1022 [FREE Full text] [doi: 10.1080/01621459.1988.10478693]
75. Littell RC, Milliken GA, Stroup WW, Wolfinger RD, Schabenberger O. SAS for Mixed Models. Cary, NC: SAS Institute Inc; 2006.
76. Liang K, Zeger S. Longitudinal data analysis using generalized linear models. Biometrika 1986;73(1):13-22 [FREE Full text] [doi: 10.1093/biomet/73.1.13]
77. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J Chronic Dis 1987;40(5):373-383. [doi: 10.1016/0021-9681(87)90171-8] [Medline: 3558716]
78. Ezekowitz J, Mentz RJ, Westerhout CM, Sweitzer NK, Givertz MM, Piña IL, et al. Participation in a heart failure clinical trial: perspectives and opportunities from the VICTORIA trial and VICTORIA simultaneous registry. Circ Heart Fail 2021;14(9):e008242. [doi: 10.1161/CIRCHEARTFAILURE.120.008242] [Medline: 34407626]
79. Starbird LE, DiMaina C, Sun C, Han H. A systematic review of interventions to minimize transportation barriers among people with chronic diseases. J Community Health 2019;44(2):400-411 [FREE Full text] [doi: 10.1007/s10900-018-0572-3] [Medline: 30206755]
80. Broendum E, Ulrik CS, Gregersen T, Hansen EF, Green A, Ringbaek T. Barriers for recruitment of patients with chronic obstructive pulmonary disease to a controlled telemedicine trial. Health Informatics J 2018;24(2):216-224 [FREE Full text] [doi: 10.1177/1460458216667166] [Medline: 27638452]

XSL•FO
RenderX

81. Amirova A, Taylor L, Volkmer B, Ahmed N, Chater AM, Fteropoulli T. Informing behaviour change intervention design using systematic review with Bayesian meta-analysis: physical activity in heart failure. Health Psychol Rev 2023;17(3):456-484 [FREE Full text] [doi: 10.1080/17437199.2022.2090411] [Medline: 35701235]
82. Collins LM, Trail JB, Kugler KC, Baker TB, Piper ME, Mermelstein RJ. Evaluating individual intervention components: making decisions based on the results of a factorial screening experiment. Transl Behav Med 2014;4(3):238-251 [FREE Full text] [doi: 10.1007/s13142-013-0239-7] [Medline: 25264464]
83. Gilliam EA, Kilgore KL, Liu Y, Bernier L, Criscitiello S, Litrownik D, et al. Managing the experience of breathlessness with Tai Chi: a qualitative analysis from a randomized controlled trial in COPD. Respir Med 2021;184:106463 [FREE Full text] [doi: 10.1016/j.rmed.2021.106463] [Medline: 34023739]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
COPD: chronic obstructive pulmonary disease
ESC: Every Steps Count
HF: heart failure
HRQL: health-related quality of life
IMI: Intrinsic Motivation Inventory
MAIA: Multidimensional Assessment of Interoceptive Awareness
MBE: mind-body exercise
PROMIS: Patient-Reported Outcomes Measurement Information System
REDCap: Research Electronic Data Capture

Edited by A Mavragani; submitted 18.03.25; peer-reviewed by T Wojckci; comments to author 28.05.25; revised version received 17.06.25; accepted 18.06.25; published 28.10.25.

Please cite as:
Kraemer KM, Litrownik D, Wayne PM, Richardson CR, Bhomia N, Kadri R, Rist PM, Ngo L, Moy ML, Yeh GY
Promoting Walking in Cardiopulmonary Disease With Mindful Steps: Pilot Feasibility Randomized Controlled Trial of a Web-Based, Pedometer-Mediated Mind-Body Intervention
JMIR Form Res 2025;9:e74118
URL: https://formative.jmir.org/2025/1/e74118
doi:10.2196/74118
PMID:

©Kristen M Kraemer, Daniel Litrownik, Peter M Wayne, Caroline R Richardson, Neha Bhomia, Reema Kadri, Pamela M Rist, Long Ngo, Marilyn L Moy, Gloria Y Yeh. Originally published in JMIR Formative Research (https://formative.jmir.org), 28.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Evaluation of a Novel Goals-of-Care Discussion Priming Tool (MyCare) in Inpatient General Internal Medicine Ward Settings: Feasibility, Acceptability, and Usability Study

Maren Kimura[1*], MPH; Sydney Ruller[1*], MSc; Alan Forster[1*], MD; James Downar[2,3*], MD; Sarina R Isenberg[3,4*], PhD; Leah Steinberg[5], MD; Kumanan Wilson[1,3,6*], MD; Victoria Shaffer[7*], PhD; Pete Wegier[5*], PhD; Jeff Myers[5*], MD; Russell Goldman[5*], MD; Carolyn Steele Gray[8,9*], PhD; Daniel Kobewka[1,3,4,6*], MD

[1]Ottawa Hospital Research Institute, Ottawa, ON, Canada
[2]Division of Palliative Care, The Ottawa Hospital, Ottawa, ON, Canada
[3]Bruyère Research Institute, Ottawa, ON, Canada
[4]Department of Medicine, University of Ottawa, Ottawa, ON, Canada
[5]Temmy Latner Centre for Palliative Care, Sinai Health System, Toronto, ON, Canada
[6]Division of General Internal Medicine, Department of Medicine, The Ottawa Hospital, Ottawa, ON, Canada
[7]Department of Psychological Sciences, University of Missouri, Columbia, MO, United States
[8]Science of Care Institute, Lunenfeld-Tanenbaum Research Institute, Sinai Health, Toronto, ON, Canada
[9]Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, ON, Canada
[*]these authors contributed equally

Corresponding Author:
Daniel Kobewka, MD
Ottawa Hospital Research Institute
1053 Carling Ave
Ottawa, ON, K1Y4E9
Canada
Phone: 1 613 761 4395
Email: dkobewka@toh.ca

Abstract

Background: For patients who are seriously ill, conversations about prognosis, values, and goals can improve person-centered care; however, these conversations do not occur consistently.

Objective: We aimed to develop an online, interactive goals-of-care discussion priming tool called MyCare to facilitate conversations about patient goals among physicians, patients, and family members. Our objective was to determine the feasibility, acceptability, and usability of MyCare for patients who are seriously ill in hospitals.

Methods: We conducted a mixed methods study on internal medicine wards with patients who were seriously ill at 2 hospitals. Participants completed MyCare with a research assistant in their hospital room, and their responses were sent to their attending physician team. Patients, family members, and physicians participated in semistructured interviews to assess advance care planning engagement, usability, and acceptability. Patients also completed the System Usability Scale, the eHealth literacy scale, and survey questions about acceptability.

Results: Patients took a median time of 32 (IQR 21-43) minutes to complete MyCare, with a range of 12 to 101 minutes. The mean eHealth literacy score was 56 out of 100. In total, 19 (76%) out of 25 patients indicated that MyCare helped them think about and understand what was most important to them, and 16 (64%) stated that MyCare helped them think about the type of care they would want if their illness progressed. The mean advance care planning engagement scores on a 5-point scale before and after MyCare were 4.11 (SD 0.77) and 4.20 (SD 0.74), respectively (mean difference 0.09, 95% CI –0.08 to 0.34; P=.22). After completing MyCare, 16% (4/25) of the patients reported feeling more prepared to talk to their medical decision-maker about the type of care they would want if they were very sick or near the end of life, and none felt less prepared. The most useful elements for patients were (1) clarifying goals and wishes for care, (2) identifying people who supported them, and (3) having a resource to facilitate conversations with their families. The most useful elements for physicians were patients' prioritized values and goals.

XSL•FO
RenderX

Conclusions: MyCare and similar tools will be more usable and acceptable if they are short and simple, focus on shared physician-patient priorities, and empower patients to advocate for their own goals.

(JMIR Form Res 2025;9:e66932) doi:10.2196/66932

KEYWORDS

goals of care; web-based tool; patient conversations; mixed methods; inpatient; medicine ward; medicine; ward; prognosis; person-centered care; pilot study; feasibility; acceptability; usability; ill patient; semistructured interviews; communication; web-based; tool; doctor-patient communication

Introduction

Background

During goals-of-care discussions (GoCDs) in hospitals, patients express their goals, fears, and worries, while health care providers share information about prognosis and treatment options [1]. The exchange of information allows clinicians and patients to create a plan of care that aligns with the patient's wishes and is consistent with their goals [2]. Information sharing during GoCDs can play a critical role in preparing substitute decision-makers for future decision-making if a person is no longer capable [3]. Many patients lack documented GoCDs in their medical records, with recent studies showing documentation rates of only 21% to 36% [4,5].

These conversations do not occur due to several factors, including physicians' fear that patients will have difficulty accepting poor prognoses, concern about conflict between family members, and physicians placing low priority on these conversations because of competing demands [4,6,7]. For patients, barriers include lack of knowledge, limited time during physician visits, and the expectation that physicians will raise the topic when necessary [8,9]. When GoCDs do not take place, patients report receiving treatments that are either more or less intensive than they would want and do not feel heard and understood by clinicians, resulting in increased anxiety and depression [10,11]. To address these barriers, we developed an interactive GoCD priming tool called MyCare with a team of patient partners, clinicians, researchers, and software designers. MyCare was created to (1) help patients who are seriously ill in hospitals express and communicate their fears, goals, and wishes for their health within their circle of care and (2) empower patients to initiate GoCDs with their family members and clinicians.

Objectives

We aimed to conduct a pilot study of MyCare on internal medicine wards with patients who are seriously ill and physicians to assess its acceptability, usability, and feasibility. A secondary objective was to gather efficacy data to support a future randomized controlled trial.

Methods

Tool Development and Description

MyCare is an interactive web-based application designed to help patients express their goals, values, and wishes related to their health care. The tool was designed to (1) educate and empower patients to participate in their own care, addressing the patient barrier of lack of knowledge; (2) elicit information from patients about their informational needs, prognostic understanding, and need for support so that these needs can be communicated to their care team; and (3) give patients an opportunity to reflect, clarify, and express their goals as they cope with serious illness, targeting the barrier of perceived irrelevance of GoCDs [9,12]. Screenshots of the tool and a detailed description of the design, development, and testing of MyCare can be found in Multimedia Appendices 1 and 2. In brief, we first reviewed the literature on existing advance care planning tools. Next, we convened a group of key stakeholders, including palliative care and internal medicine clinicians, patients who have experienced serious illnesses, web designers, and graphic designers, to design MyCare. The key elements for MyCare were determined by input from stakeholders and the literature on existing tools [13,14]. Our initial draft was adapted and modified through 6 rounds of feedback with stakeholders. The final design was presented to the Equity-Mobilizing Partnerships in Community group to determine how MyCare meets the needs of diverse populations [15].

The final version of MyCare included questions about the understanding of the illness, informational needs, support networks, preferred language, level of independence before hospital admission, current symptoms, and a question from the dignity-preserving care movement [16,17]. MyCare also includes 3 fictional narratives of people living with serious illness who have different goals for their health. In one narrative, the person values living as long as possible more than comfort; in the second, the person values comfort more than living longer; and in the third, the person values living longer only if certain abilities are maintained and distress is limited. Users were asked to select the narrative where the person's goals were most similar to their own. The final questions asked patients to select and rank their most important priorities, worries, and goals with both closed and open-ended questions. The final section presented a summary of the patient's responses and instructions on how to share this information with family and physicians. Screenshots of the tool are provided in Multimedia Appendix 2.

Study Design

We used a convergent parallel mixed methods design to evaluate the acceptability, usability, and feasibility of MyCare for inpatients and their attending physicians at an academic teaching hospital in Ottawa, Ontario, between May 2022 and November 2022 [18]. A convergent mixed methods design allows us to collect qualitative and quantitative data simultaneously, analyze them separately, and subsequently meld the results together to provide a more in-depth answer to our research question.

XSL•FO
RenderX

Ethical Considerations

Our study was approved by the Ottawa Health Science Network Research Ethics Board (approval code OSHN-REB#20210787-01H), which approved all study material, including consent forms. All participants provided informed consent before participation and were not compensated. Study data were securely stored on hospital servers. Study data have been anonymized to protect participant identity.

Participants and Recruitment

Study participants were recruited from the internal medicine wards at 2 tertiary care teaching hospitals in Ottawa, Ontario. Admitted patients were eligible if were comfortable communicating in English and met one or more of the following criteria: (1) aged ≥70 years; (2) 1-year risk of death >20% calculated by an automated, validated predictive model run within the electronic health record [19]; or (3) recommended for inclusion by their attending physician based on the potential benefit to the patient. Patients were excluded if they could not participate in a discussion about their care due to cognitive issues. If patients could explain the study elements and risks back to the research assistant during the consent process, they were deemed to have adequate cognition to participate. Consenting participants were asked if they would like to complete the study with a support person. After patient recruitment was complete, internal medicine physicians who worked at the 2 study hospitals during the study period were contacted via email and asked to participate in semistructured interviews.

Intervention

MyCare was administered on an Apple iPad with the support of a caregiver or the research assistant. With consent, a copy of MyCare responses was sent to the attending physician in the hospital and uploaded into the electronic medical record.

Data Collection, Outcomes, and Evaluation Framework

Fit Between Individuals, Tasks, and Technology Framework

We used the fit between individuals, tasks, and technology framework to guide our evaluation of the feasibility (the ability for a technology to be adopted into a setting), usability (how well a technology meets user needs), and acceptability (user-task fit) of MyCare [20,21]. The framework posits that adoption of a technology requires a fit between the user, the technology, and the task. We measured qualitative and quantitative outcomes and compared or combined the matching categories where appropriate. For example, the quantitative and qualitative usability findings were compared to provide a richer understanding of usability for our study population.

Outcomes and Data Collection

We measured usability (fit between user and technology) using 3 different methods. The eHealth literacy assessment tool measures user confidence with technology across 3 domains: using a computer, using a touchscreen, and finding information online [22]. We measured usability with the System Usability Scale (SUS) and semistructured interviews [23]. The SUS consists of 10 statements scored on a 5-point Likert scale and is normalized on a 100-point scale, with higher scores indicating greater usability. In addition, semistructured interview questions further explored the experience of using the MyCare interface (user-task fit).

We assessed acceptability through semistructured interviews with patients and physicians focusing on aspects of the tool they found useful, what could be improved and added to the tool, and what should or could be removed from the tool. Ethnographic field observations were collected during patient encounters to supplement data on acceptability.

We assessed feasibility by using MyCare with the target patients and physicians for whom it was designed, measuring the time required for patients to complete MyCare and semistructured interview questions and researcher memoing.

We measured advance care planning engagement as an exploratory efficacy outcome that could be used for a future randomized trial. The advance care planning engagement survey is a validated measure of behavior change in 4 advance care planning domains [24].

The semistructured interview guide is provided in Multimedia Appendix 3. Interviews were conducted in person and audio recorded by MK, and SR, both early-career, female research assistants with training in qualitative methods. Participants had no relationship with or knowledge of the interviewers before study commencement.

Analysis: Quantitative and Qualitative

We calculated Pearson correlation coefficients between the SUS and age, sex, and eHealth literacy scores. The advance care planning engagement survey was analyzed using two methods as suggested by the tool's creators: (1) mean difference and (2) percentage of participants who were more ready to engage in advance care planning behaviors [25,26].

We used inductive qualitative thematic analysis to analyze our semistructured interviews [27]. One member of the research team transcribed interviews, field notes, and postinterview memos, while a second member reviewed transcripts for accuracy. Inductive thematic coding was used to depict themes from interview transcripts related to acceptability, usability, and feasibility of our tool for patients and physicians. Coding of transcripts was completed by 2 coders (MK and DK) in Microsoft Excel. DK is a midcareer, male, internal medicine physician scientist. One coder (MK) read multiple transcripts to establish a draft codebook, which was reviewed through iterative discussion with a second coder (DK) until consensus was met and a final codebook was created. We then tested the codebook with other transcripts to ensure no new codes were identified. The final codebook was uploaded to Excel to facilitate coding of transcripts. To ensure consistency, the first transcript was done via consensus coding (MK and DK), and then the remaining transcripts were coded (MK) alongside field notes and postinterview memos. Codes were organized into themes and subthemes according to acceptability, usability, and feasibility for patients and physicians. Both coders identified themes and subthemes and discussed them with a third team member (SR). Themes were then defined and described.

XSL•FO
RenderX

Results

Overview

There were 28 patients who consented, 25 (89%) completed the study, and 8 internal medicine physicians participated. In total, 3 (11%) patients withdrew due to declines in health, interruptions in medical procedures, fatigue, and discharge from the hospital before study completion. In total, 4 (16%) patients had a support person involved; it was up to the participant if they wanted to have a support person with them during the research encounter. Through memoing by the research team, we know that these encounters were comparable to those without a support person present. The mean age of participants was 76 (range 70-88) years, and 64% (16/25) were women. Most patient participants were White, English-speaking, Catholic, and lived at home before admission (Table 1).

XSL•FO
RenderX

Table 1. Characteristics of participants using the MyCare study tool (N=25).

Participant characteristics	Values
Age (y), mean (SD)	76.24 (6.46)
Sex, n (%)	
Female	16 (64)
Male	9 (36)
Place of residence, n (%)	
At your home (apartment, townhouse, or bungalow)	20 (80)
Hospital	1 (4)
Retirement residence	3 (12)
Other	1 (4)
Ethnicity, n (%)	
East Indian	1 (4)
Middle Eastern	1 (4)
White	22 (88)
Missing	1 (4)
Primary language, n (%)	
English	21 (84)
French	3 (12)
Bilingual	1 (4)
Religion, n (%)	
Catholic	12 (48)
Atheist or agnostic	7 (28)
Protestant	5 (20)
Muslim	1 (4)
Education level, n (%)	
Did not complete high school	1 (4)
Completed high school	8 (32)
Had university or postsecondary education (trade, technical, or vocational school)	8 (32)
University degree	5 (20)
Graduate degree	3 (12)
Annual household income (CAD $[a]), n (%)	
<10,000	1 (4)
10,000-19,999	1 (4)
20,000-34,999	3 (12)
35,000-49,999	3 (12)
50,000-74,999	4 (16)
75,000-99,999	6 (24)
>100,000	2 (8)
Missing	5 (20)

[a]CAD $1 was US $1.30 at the time of the study.

Patients took a median time of 32 (IQR 21-43) minutes to complete MyCare, with a range of 12 to 101 minutes. The mean eHealth literacy score was 56 out of 100. In total, 19 (76%) of the 25 patients indicated that MyCare helped them think about and understand what is most important to them, and 16 (64%)

XSL•FO
RenderX

stated MyCare helped them think about the type of care they would want if their illness progressed.

The mean advance care planning engagement scores on a 5-point scale before and after MyCare were 4.11 (SD 0.77) and 4.20 (SD 0.74), respectively (mean difference 0.09, 95% CI –0.08 to 0.34; P=.22). After completing MyCare, 16% (4/25) of the patients reported feeling more prepared to talk to their medical decision-maker about the type of care they would want if they were very sick or near the end of life, and none felt less prepared. Moreover, 12% (3/25) of the patients felt more ready to ask their physician questions, while 4% (1/25) felt less ready. In addition, 12% (4/25) of the patients felt more ready and 12% (4/25) felt less ready to talk to their physician about the kind of medical care they would want if they were very sick or near the end of life.

Usability

The mean SUS score was 64.7 out of 100 (range 42.5-80.0) or "okay" using the adjective rating scale presented by Bangor et al [28]. The mean score for internet-based webpages and applications is 68.1 [23]. Patients' technology confidence was positively correlated with SUS scores (r=0.49; P<.05). Age and sex had no correlation with the SUS score. Consistent with the "okay" score, some patients required assistance from research staff to navigate MyCare, read questions, and select answers.

Semistructured Interviews

Interviews with patients and physicians uncovered themes related to usability, feasibility, and acceptability. Thematic codes, definitions, and exemplary quotes from patient and health care provider interviews are provided in Multimedia Appendices 4 and 5.

Usability

Within the usability theme, we found facilitators and barriers. Patients were motivated to use the tool because it was easy and enjoyable and empowered them to initiate and participate in conversations about their care. Barriers for patients included needing physical and technological support to use the iPad and navigate the MyCare interface, repetitive questions, and difficulty knowing how to answer some questions. Patients commented that prespecified response options for questions did not always capture the complexity of their responses or allow for flexibility in responses that may change depending on their state of health. Patients asked for free text fields to express themselves:

> *A lot of the questions are yes or no answers but a lot of them are not yes or no answers, they are in between. Or there is an explanation to the yes and the no and that is not provided. I think it should say yes and qualify, or no and qualify.*

Being able to provide context or additional information that is not captured within a yes or no answer was important. Some patients commented that they had difficulty using technology, including touch screens.

Physicians thought that having the patient's wishes documented in the MyCare output would help them keep family meetings focused on the patient's values. A physician shared the following:

> *Goals always get made. Just it's the confidence of which you've made them right. Whether you've talked to enough family? Or whether or not it's the first time you talk to Mr. Smith? Was he truly with it or not so much. Right? And then the family come and they doubt your word.... So having extra information is always very helpful.*

Physicians expressed that sharing patients' responses with the health care team would promote shared decision-making, as the MyCare results are uploaded to a patient's electronic medical record. Physicians thought that MyCare could serve as a framework for junior staff or trainees who have less experience facilitating GoCDs. In addition to serving as a framework, the tool could help patients and physicians reach a common understanding. Discussing questions and responses from MyCare could help physicians initiate GoCDs around specific topics that may be difficult to introduce.

Acceptability

Patients and physicians provided insights about the perceived utility of the tool and made suggestions for improvement, an aspect of user-task fitness and acceptability (Multimedia Appendix 6). Patients said the most useful aspects of the tool were questions about goals after hospital discharge, the type of care they would want near the end of life, and the people who are most important to them. Patients thought that questions were repetitive, which they found frustrating. Patients suggested adding questions about mental health, specific preferences for end-of-life care, and community care would better reflect the needs of the target patient population. Patients self-identified 2 functions of MyCare: personal reflection and supporting shared communication of their goals, wishes, and values. The patient stated that MyCare was informative and allowed them to think about next steps as they age:

> *[T]he part about getting older and close to death and prioritizing and reflecting on the things, that made me aware to do the things.*

Patients appreciated that the tool helped them articulate what is most important to them and think about future decisions they may face. Furthermore, patients planned to keep their responses for their own reflection and to ensure their wishes were respected.

Some patients felt they already knew what they wanted and did not need the tool to guide them; these patients still planned to use the tool to initiate and guide communication with their families and health care team. One patient said that although the tool would not be useful to them, it "might be [helpful] to the family to show it to the kids and give them an understanding of how I feel without having to say."

Another patient shared that by using the tool "the doctor can clarify what lies ahead for [me] that I don't realize right now...and to help me understand why I'm going through certain steps."

XSL•FO
RenderX

For some patients, completing MyCare reinforced that they had been talking to their physicians about the right things, which increased their confidence in their health care team and their own ability to engage in conversations about their care.

However, the value of MyCare was limited for some patients who shared that the tool content did not relate to their experiences or that they would prefer to talk to their physician about their goals of care instead of using an online tool. Several patients voiced concerns that the tool would only be useful if "the doctors are able and willing to listen and would take action to what I'm saying."

Some patients doubted the tool could improve their health care experiences related to GoCDs.

Physicians reported that the most useful aspects were asking patients to prioritize their most important goals and to share their illness understanding, while the least useful aspect was the symptom scale because this information is already obtained during routine clinical assessment. Physicians suggested more ranking tasks could be integrated for goals, wishes, fears, and symptoms to identify what is the most important to the patient. There was mixed feedback about the description of patients' activities of daily living. Some physicians found it helpful to know what activities of daily living patients could complete before coming to the hospital, while others indicated it was less important than knowing what they want to be able to do when they leave.

Physicians said they would trust and value the patient's responses to MyCare questions if they knew responses were up-to-date and received the results promptly. One physician shared that they would be motivated to use the tool with their patients if they knew it was supported at an institutional level by the hospital. Others suggested that integration of MyCare into the hospital workflows and explicit links to legal frameworks for decision-making would allow physicians to feel confident and supported in abiding by their patients' wishes documented through MyCare.

Physicians found MyCare valuable to support reflection and shared decision-making. Physicians thought MyCare would help patients reflect on what was most important to them and help physicians to get to know the patient before meeting with them. By reviewing patients' responses ahead of meeting with them, physicians suggested that the tool could promote shared decision-making between patients and their health care team:

> *I think what [the tool] does is it really sets the stage for understanding what [the patient's] values are. So I definitely think you know, getting it before I had a goals of care discussion with the patient would be really helpful. I think for me to understand where that coming from, but also even for them to start getting them thinking about these things.*

Physicians shared that MyCare results would improve their confidence that decisions align with the patient's values, especially in cases where patients later lost the ability to participate in decision-making and the family's wishes do not align with the patient's.

Emotional responses to MyCare, captured in researcher memoing, were common. Some patients seemed empowered and confident in using the tool, making comments about how the questions were great and improved their confidence in telling others what is most important to them. Other patients expressed feelings of fatigue due to the length of the tool. Researchers' memoing highlights that patients struggled with how long it took to complete MyCare in addition to studying questionnaires. Some patients expressed confusion and worry because the tool talked about living with serious illnesses, which they did not consider themselves to have or had not given much prior thought. Conversely, some patients did not seem to express positive or negative emotions while using MyCare.

Feasibility

Physicians were concerned that symptomatic illness would be a barrier for some patients, preventing them from engaging with MyCare. The concern was validated by memoing of fatigue for patients completing MyCare. In addition, physicians shared mixed perspectives on how time constraints would impact their ability to review MyCare results. Some felt MyCare would help facilitate GoCD's under time constraints, stating the following:

> *[P]art of the challenge while on service is not enough time to do the necessary background, you know, getting to know the patient. So this helps you get a step ahead and gives you the information available to then have the conversations.*

While others indicated that they would not have time to review the information from MyCare after the patient completed it.

Discussion

Principal Findings

We designed and pilot-tested an interactive online GoCD priming tool with patients who were seriously ill and admitted to the hospital's general internal medicine ward. The piloted version of MyCare had a below-average SUS score [28]. Length of the tool and familiarity with touch screen technology were the primary usability issues and could be improved by removing the symptom assessment and activities of daily living questions that patients and physicians thought were unnecessary. MyCare had moderate to poor user-technology fit, but we think this can be improved by pushing further toward minimalist design with succinct content as suggested by the study participants and guidelines on user interface design for older patients [29,30]. Patient concerns with inflexible response options could also be improved with optional free text fields to allow for in-depth answers while maintaining prespecified options.

Overall physicians found the MyCare results highly usable, allowing them to quickly absorb key information about patients' goals and serve as a guide for the patient, physician, and junior trainees learning to have GoCDs.

MyCare had moderate acceptability. Most acceptability data from patients was related to perceived utility, the fit between task and individual. Many advance care planning interventions have poor acceptability because patients perceive them as irrelevant because they do not see themselves as having a high

XSL•FO
RenderX

risk of death or decline even when they have a high risk of death [31]. We attempted to address this barrier by testing MyCare with people who are hospitalized for acute illness and have a high risk of death. Some patients commented that this context made GoCD more relevant, motivating them to use MyCare and share results with loved ones. MyCare also used narratives to demonstrate how real people have benefited from GoCD. Despite these attempts to target the barrier of "perceived irrelevance," we were only partially successful. GoCD and advance care planning are complex behaviors with multiple barriers. Any future implementation of MyCare needs to target these multiple barriers using implementation science approaches to maximize uptake [32,33].

Future versions of MyCare and related tools can improve the task-individual fit by ensuring that tools meet the needs of patients and physicians. Patients told us that identifying goals after discharge, care preferences for the end of life, and the people who support them were the most helpful parts of MyCare. Physicians focused on prioritizing patients' values, wishes, and goals for care to support medical decision-making. Although patients and physicians had different priorities, both agreed that values and goals for care were important and useful. Future versions of MyCare and similar tools should focus on shared patient-physician priorities, such as eliciting value-based information about what patients want. Values-based information is a shared priority and is more useful to substitute decision-makers than treatment-based information that only applies to a single decision [34]. Patients and physicians in our study said that MyCare helped to clarify values and could improve communication about these values between substitute decision-makers and medical providers.

We found several feasibility challenges. Patients took considerable time to complete the tool, required assistance, and became fatigued. Physicians thought that MyCare results would improve their care by allowing them to quickly understand what is most important to their patient, but the physicians we interviewed had not all received MyCare results when they were on busy clinical service. We do not know if MyCare will truly address the barrier of "not enough time" commonly reported by physicians [8]. Implementation of MyCare will require institution-level support to integrate MyCare into workflows, electronic medical records, and institutional documentation standards [35]. Stripping MyCare and the response summary to essential elements will further improve usability and feasibility by minimizing time investment for physicians and patients.

Given that GoCDs are not prioritized by physicians, tools must empower patients to initiate conversations or prompt physicians to do so. Patients need to know that information about their goals and wishes will have direct benefits for them, and they need tangible steps they can take to initiate these conversations. Patients can be further empowered by being given explicit control over who sees their responses. Finally, MyCare will be more effective if it is introduced by the patient's physician, who commits to discussing the results. If patients share sensitive information using MyCare only for it to be disregarded by their team, it could break patient-physician trust and discourage future GoCDs.

Strengths and Limitations

Our study has several strengths. Our study is grounded and organized around the fit between individuals, tasks, and the technology framework. Our convergent parallel mixed method design allowed us to collect and analyze both objective and subjective data around the use of MyCare, increasing reliability and depth. Another strength is that we implemented our tool in a real-world setting, an academic hospital with patients who are seriously ill and who benefit from GoCDs.

Our study had several limitations. First, our sample was younger, healthier, and less diverse than the tool's intended target population. Some age-eligible patients were unable to consent or complete the study because of symptoms such as fatigue and cognitive impairment. For those who did complete MyCare, data collection was time-consuming. Participant fatigue contributed to missing data from the most unwell patients and may have biased results. Limited sociodemographic diversity may reduce generalizability, and therefore, testing with diverse populations will be important as we move forward. The highlights of this study are provided in Textbox 1.

Textbox 1. Highlights of the study.

- We co-designed an online survey tool to help patients who are frail reflect on and share their goals, wishes, and fears with their care team and people important to them.
- Patients and physicians found the tool helpful to clarify patient goals, identify people who support the patient, and facilitate conversations.
- Our tool would be more usable if it were shorter and focused only on shared physician-patient priorities.

Conclusions

Patients and physicians found the MyCare tool valuable for promoting conversations about goals of care. To improve usability and utility, MyCare and future tools must empower patients, prioritize simplicity, and limit content to areas that meet the shared patient's and physician's priorities.

Acknowledgments

This work was supported by The Ottawa Hospital Academic Medical Organization and the Temmy Latner Centre for Palliative Care.

XSL•FO
RenderX

Data Availability

Data are available upon reasonable request to the corresponding author.

Conflicts of Interest

None declared.

Multimedia Appendix 1
MyCare tool development.
[DOCX File , 29 KB - formative_v9i1e66932_app1.docx]

Multimedia Appendix 2
Screenshots from the MyCare tool.
[DOCX File , 2139 KB - formative_v9i1e66932_app2.docx]

Multimedia Appendix 3
Semistructured interview guide.
[DOCX File , 20 KB - formative_v9i1e66932_app3.docx]

Multimedia Appendix 4
Patient interview themes.
[DOCX File , 28 KB - formative_v9i1e66932_app4.docx]

Multimedia Appendix 5
Physician interview themes.
[DOCX File , 28 KB - formative_v9i1e66932_app5.docx]

Multimedia Appendix 6
Suggestions for improvements.
[DOCX File , 28 KB - formative_v9i1e66932_app6.docx]

References

1. Edmonds KP, Ajayi TA. Do we know what we mean? An examination of the use of the phrase "goals of care" in the literature. J Palliat Med 2019 Dec 01;22(12):1546-1552. [doi: 10.1089/jpm.2019.0059] [Medline: 31215827]
2. Secunda K, Wirpsa MJ, Neely KJ, Szmuilowicz E, Wood GJ, Panozzo E, et al. Use and meaning of "goals of care" in the healthcare literature: a systematic review and qualitative discourse analysis. J Gen Intern Med 2020 May;35(5):1559-1566 [FREE Full text] [doi: 10.1007/s11606-019-05446-0] [Medline: 31637653]
3. McMahan RD, Hickman SE, Sudore RL. What clinicians and researchers should know about the evolving field of advance care planning: a narrative review. J Gen Intern Med 2024 Mar 02;39(4):652-660 [FREE Full text] [doi: 10.1007/s11606-023-08579-5] [Medline: 38169025]
4. Wheless M, Lee JJ, Domenico HJ, Martin BJ, Bennett ML, Martin SF, et al. Factors and barriers to goals-of-care conversations for patients with cancer and inpatient mortality. JCO Oncol Pract 2023 Sep;19(9):767-776. [doi: 10.1200/OP.23.00095] [Medline: 37390380]
5. Uyeda AM, Lee RY, Pollack LR, Paul SR, Downey L, Brumback LC, et al. Predictors of documented goals-of-care discussion for hospitalized patients with chronic illness. J Pain Symptom Manage 2023 Mar;65(3):233-241 [FREE Full text] [doi: 10.1016/j.jpainsymman.2022.11.012] [Medline: 36423800]
6. You JJ, Downar J, Fowler RA, Lamontagne F, Ma IW, Jayaraman D, et al. Barriers to goals of care discussions with seriously ill hospitalized patients and their families: a multicenter survey of clinicians. JAMA Intern Med 2015 Apr;175(4):549-556. [doi: 10.1001/jamainternmed.2014.7732] [Medline: 25642797]
7. Zannella V, Vaillancourt S, Argintaru N, McGowan M, Ailon J, Quinn K. Perceived barriers to goals of care discussions in the emergency department: a multi-center analysis of hospitalist physicians. Univ Toronto Med J 2022;99(3) [FREE Full text] [doi: 10.21203/rs.3.rs-137260/v1]
8. Bernard C, Tan A, Slaven M, Elston D, Heyland DK, Howard M. Exploring patient-reported barriers to advance care planning in family practice. BMC Fam Pract 2020 May 25;21(1):94 [FREE Full text] [doi: 10.1186/s12875-020-01167-0] [Medline: 32450812]

XSL•FO
RenderX

9. Simon J, Porterfield P, Bouchal SR, Heyland D. 'Not yet' and 'just ask': barriers and facilitators to advance care planning--a qualitative descriptive study of the perspectives of seriously ill, older patients and their families. BMJ Support Palliat Care 2015 Mar;5(1):54-62. [doi: 10.1136/bmjspcare-2013-000487] [Medline: 24644192]
10. Wright AA, Zhang B, Ray A, Mack JW, Trice E, Balboni T, et al. Associations between end-of-life discussions, patient mental health, medical care near death, and caregiver bereavement adjustment. JAMA 2008 Oct 08;300(14):1665-1673 [FREE Full text] [doi: 10.1001/jama.300.14.1665] [Medline: 18840840]
11. Zhang B, Wright AA, Huskamp HA, Nilsson ME, Maciejewski ML, Earle CC, et al. Health care costs in the last week of life: associations with end-of-life conversations. Arch Intern Med 2009 Mar 09;169(5):480-488 [FREE Full text] [doi: 10.1001/archinternmed.2008.587] [Medline: 19273778]
12. Peck V, Valiani S, Tanuseputro P, Mulpuru S, Kyeremanteng K, Fitzgibbon E, et al. Advance care planning after hospital discharge: qualitative analysis of facilitators and barriers from patient interviews. BMC Palliat Care 2018 Dec 05;17(1):127 [FREE Full text] [doi: 10.1186/s12904-018-0379-0] [Medline: 30518345]
13. Hui D, Bruera E. The Edmonton Symptom Assessment System 25 years later: past, present, and future developments. J Pain Symptom Manage 2017 Mar;53(3):630-643 [FREE Full text] [doi: 10.1016/j.jpainsymman.2016.10.370] [Medline: 28042071]
14. Paladino J, Bernacki R, Neville BA, Kavanagh J, Miranda SP, Palmor M, et al. Evaluating an intervention to improve communication between oncology clinicians and patients with life-limiting cancer: a cluster randomized clinical trial of the serious illness care program. JAMA Oncol 2019 Jun 01;5(6):801-809. [doi: 10.1001/jamaoncol.2019.0292] [Medline: 30870556]
15. Sayani A, Maybee A, Manthorne J, Nicholson E, Bloch G, Parsons JA, et al. Equity-mobilizing partnerships in community (EMPaCT): co-designing patient engagement to promote health equity. Healthc Q 2022 Apr;24(SP):86-92. [doi: 10.12927/hcq.2022.26768] [Medline: 35467517]
16. Chochinov HM. Dignity-conserving care--a new model for palliative care: helping the patient feel valued. JAMA 2002 May 01;287(17):2253-2260. [doi: 10.1001/jama.287.17.2253] [Medline: 11980525]
17. Heuzenroeder L, Ibrahim F, Khadka J, Woodman R, Kitson A. A Delphi study to identify content for a new questionnaire based on the 10 Principles of Dignity in Care. J Clin Nurs 2022 Jul;31(13-14):1960-1971. [doi: 10.1111/jocn.15462] [Medline: 32799400]
18. Creswell JW, Creswell JD. Research Design: Qualitative, Quantitative, and Mixed Methods Approaches. Thousand Oaks, CA: SAGE Publications; 2017.
19. van Walraven C, McAlister FA, Bakal JA, Hawken S, Donzé J. External validation of the Hospital-patient One-year Mortality Risk (HOMR) model for predicting death within 1 year after hospital admission. CMAJ 2015 Jul 14;187(10):725-733 [FREE Full text] [doi: 10.1503/cmaj.150209] [Medline: 26054605]
20. Tsiknakis M, Kouroubali A. Organizational factors affecting successful adoption of innovative eHealth services: a case study employing the FITT framework. Int J Med Inform 2009 Jan;78(1):39-52. [doi: 10.1016/j.ijmedinf.2008.07.001] [Medline: 18723389]
21. Sheehan B, Lee Y, Rodriguez M, Tiase V, Schnall R. A comparison of usability factors of four mobile devices for accessing healthcare information by adolescents. Appl Clin Inform 2012;3(4):356-366 [FREE Full text] [doi: 10.4338/ACI-2012-06-RA-0021] [Medline: 23227134]
22. Karnoe A, Furstrand D, Christensen KB, Norgaard O, Kayser L. Assessing competencies needed to engage with digital health services: development of the eHealth literacy assessment toolkit. J Med Internet Res 2018 May 10;20(5):e178 [FREE Full text] [doi: 10.2196/jmir.8347] [Medline: 29748163]
23. Bangor A, Kortum PT, Miller JT. An empirical evaluation of the system usability scale. Int J Hum Comput Interact 2008 Jul 30;24(6):574-594. [doi: 10.1080/10447310802205776]
24. Sudore RL, Heyland DK, Barnes DE, Howard M, Fassbender K, Robinson CA, et al. Measuring advance care planning: optimizing the advance care planning engagement survey. J Pain Symptom Manage 2017 Apr;53(4):669-81.e8 [FREE Full text] [doi: 10.1016/j.jpainsymman.2016.10.367] [Medline: 28042072]
25. Advance care planning engagement survey. PREPARE. URL: https://prepareforyourcare.org/en/prepare/research/acp-engagement-survey [accessed 2024-08-12]
26. Fried TR, Redding CA, Robbins ML, Paiva A, O'Leary JR, Iannone L. Stages of change for the component behaviors of advance care planning. J Am Geriatr Soc 2010 Dec;58(12):2329-2336 [FREE Full text] [doi: 10.1111/j.1532-5415.2010.03184.x] [Medline: 21143441]
27. Guest G, Macqueen KM, Namey EE. Applied Thematic Analysis. Thousand Oaks, CA: SAGE Publications; 2012.
28. Bangor A, Kortum P, Miller J. Determining what individual SUS scores mean: adding an adjective rating scale. J Usability Stud 2009;4(3):114-123 [FREE Full text]
29. Srivastava A, Kapania S, Tuli A, Singh P. Actionable UI design guidelines for smartphone applications inclusive of low-literate users. Proc ACM Hum Comput Interact 2021 Apr 22;5(CSCW1):1-30. [doi: 10.1145/3449210]
30. Morris JM. User interface design for older adults. Interact Comput 1994 Dec;6(4):373-393. [doi: 10.1016/0953-5438(94)90009-4]

XSL•FO
RenderX

31. Schickedanz AD, Schillinger D, Landefeld CS, Knight SJ, Williams BA, Sudore RL. A clinical framework for improving the advance care planning process: start with patients' self-identified barriers. J Am Geriatr Soc 2009 Jan 31;57(1):31-39 [FREE Full text] [doi: 10.1111/j.1532-5415.2008.02093.x] [Medline: 19170789]
32. Michie S, van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. Implement Sci 2011 Apr 23;6:42 [FREE Full text] [doi: 10.1186/1748-5908-6-42] [Medline: 21513547]
33. Powell BJ, Waltz TJ, Chinman MJ, Damschroder LJ, Smith JL, Matthieu MM, et al. A refined compilation of implementation strategies: results from the Expert Recommendations for Implementing Change (ERIC) project. Implement Sci 2015 Feb 12;10:21 [FREE Full text] [doi: 10.1186/s13012-015-0209-1] [Medline: 25889199]
34. Myers J, Cosby R, Gzik D, Harle I, Harrold D, Incardona N, et al. Provider tools for advance care planning and goals of care discussions: a systematic review. Am J Hosp Palliat Care 2018 Aug 12;35(8):1123-1132. [doi: 10.1177/1049909118760303] [Medline: 29529884]
35. Ouchi K, George N, Schuur JD, Aaronson EL, Lindvall C, Bernstein E, et al. Goals-of-care conversations for older adults with serious illness in the emergency department: challenges and opportunities. Ann Emerg Med 2019 Aug;74(2):276-284 [FREE Full text] [doi: 10.1016/j.annemergmed.2019.01.003] [Medline: 30770207]

Abbreviations

GoCD: goals-of-care discussion
SUS: System Usability Scale

Edited by A Mavragani; submitted 26.09.24; peer-reviewed by D Grossman, K Tenzek, J Lyu; comments to author 09.04.25; revised version received 30.07.25; accepted 05.08.25; published 28.10.25.

Please cite as:
Kimura M, Ruller S, Forster A, Downar J, Isenberg SR, Steinberg L, Wilson K, Shaffer V, Wegier P, Myers J, Goldman R, Steele Gray C, Kobewka D
Evaluation of a Novel Goals-of-Care Discussion Priming Tool (MyCare) in Inpatient General Internal Medicine Ward Settings: Feasibility, Acceptability, and Usability Study
JMIR Form Res 2025;9:e66932
URL: https://formative.jmir.org/2025/1/e66932
doi:10.2196/66932
PMID:

©Maren Kimura, Sydney Ruller, Alan Forster, James Downar, Sarina R Isenberg, Leah Steinberg, Kumanan Wilson, Victoria Shaffer, Pete Wegier, Jeff Myers, Russell Goldman, Carolyn Steele Gray, Daniel Kobewka. Originally published in JMIR Formative Research (https://formative.jmir.org), 28.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Detecting Perceived Unfair Treatment Among US College Students Using Mobile Sensing: Pilot Machine Learning Study

Yiyi Ren[1], MS; Raghu Mulukutla[2], MS; Jennifer Mankoff[3], PhD; Anind K Dey[1], PhD

[1]Information School, University of Washington, Seattle, WA, United States

[2]College of Computing, Georgia Institute of Technology, Atlanta, GA, United States

[3]Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, WA, United States

Corresponding Author:
Yiyi Ren, MS
Information School
University of Washington
1851 NE Grant Ln
Seattle, WA, 98105
United States
Phone: 1 4123308417
Email: yren98@uw.edu

Abstract

Background: Experiences of unfair treatment on college campuses are linked to adverse mental and physical health outcomes, highlighting the need for interventions. However, detecting such experiences relies mainly on self-reports. No prior research has examined the feasibility of using mobile sensing via smartphones and wearables for the passive detection of these experiences.

Objective: This pilot study explores the potential of using passive sensing to detect daily experiences of perceived unfair treatment (PUT) after they occur. It aims to develop and evaluate machine learning models against naive baselines and establish a benchmark for future research.

Methods: We analyzed data from 201 undergraduate students collected over two 10-week academic terms in 2018. PUT was self-reported at the daily level via ecological momentary assessment (EMA) surveys, with 413 of 9629 (4.3%) total responses indicating unfair treatment. We implemented two modeling approaches with distinct training schemes: (1) supervised classification models trained in a user-independent manner using data from different individuals, and (2) anomaly detection models trained in a user-dependent manner using historical data from the same individuals. Classification performance was assessed using stratified group 5-fold cross-validation for user-independent models and a chronological train-test split for user-dependent models.

Results: Of the 201 study participants, 110 reported experiencing unfair treatment at least once. On average, participants reported unfair treatment in 4.66% of their EMA responses (95% CI 3.13% to 6.19%). User-independent classification models showed mixed performance (AUC-ROC [area under the receiver operating characteristic curve]: 0.546-0.640, AUC-PR [area under the precision-recall curve]: 0.047-0.093, F_1-score: 0.070-0.121). Tree-based models, particularly light gradient boosting machine (LightGBM) and Random Forest, outperformed all 3 baselines in AUC-ROC and AUC-PR; LightGBM also improved the F_1-score. In comparison, user-dependent anomaly detection models performed better, with the multiday long short-term memory-AE model (50 features, 7-day window) achieving the highest recall (0.830, +73.3%, $P<.001$) and F_1-score (0.391, +24.9%, $P<.001$) without reducing precision (0.256), and improving AUC-PR by 45.9% and AUC-ROC by 21.6% relative to naive baselines ($P<.001$). Feature importance analysis identified key behavioral patterns for population-level detection, including increased time spent off campus, elevated evening and nighttime activity, reduced indoor mobility on campus, prolonged screen use, delayed sleep onset, and shorter sleep duration.

Conclusions: Mobile sensing shows promise for detecting daily experiences of PUT in college students and identifying associated behavioral patterns. Our findings highlight opportunities for timely interventions through mobile technology to mitigate the impact of these experiences on students' mental health and well-being.

(JMIR Form Res 2025;9:e78657) doi:10.2196/78657

KEYWORDS

passive sensing; perceived discrimination; anomaly detection; digital phenotyping; mental health; mobile health

XSL•FO
RenderX

Introduction

Unfair treatment refers to the act of denying individuals equal and just consideration based on characteristics such as race, gender, age, or disability [1]. In US college environments, perceived unfair treatment (PUT) remains a persistent issue with significant impacts on students' lives [2-5]. While the literature often uses discrimination interchangeably with unfair treatment [6,7], our study adopts the broader construct of PUT, which includes not only overt acts of discrimination but also subtle, everyday indignities known as microaggressions [8]. Drawing on past research on perceived discrimination to contextualize our work, we define PUT as an individual's subjective perception of being treated unjustly based on group characteristics. Within the university setting, this can manifest in various ways. For example, students being stereotyped by faculty, perceiving bias in academic evaluation, being unfairly blamed for dorm noise, or encountering classmates who express surprise at a minority student's success [9].

These experiences of PUT can induce acute physiological and emotional distress [10-13], contribute to increased suicidality [14-16], substance use [3,17,18], poor academic performance [19,20], and have long-lasting effects on social well-being and mental health, such as disrupted personality development [21], hindered career growth [22,23], eating disorders [24,25], and impaired social integration [26,27]. Despite their prevalence, many incidents go unreported [28-30], resulting in limited institutional awareness and response. Developing reliable methods to detect these experiences soon after they occur is crucial for enabling timely interventions [31,32] and providing social support [33,34] for at-risk students.

Currently, PUT is primarily studied and detected based on self-reports, either via standard questionnaires such as the Major Experiences of Discrimination and the Everyday Discrimination Scale [35-37], or through the Experience Sampling Method, also known as Ecological Momentary Assessment (EMA) [38-40]. While these self-reported measures provide valuable insights into individuals' experiences, they are subject to recall and nonresponse biases, inconsistent reporting, and significant participant burden [41-43], making them challenging to scale for continuous or longitudinal monitoring and detection. To the best of our knowledge, no framework or system exists that can automatically or passively detect PUT after it happens.

The health care landscape is undergoing a notable transformation, shifting towards noninvasive and accessible methods for early detection [44-46]. This shift is being largely fueled by advancements in mobile sensing technology and the growing interest in machine learning, which together offer unprecedented opportunities [47-50]. Numerous studies have highlighted the effectiveness of these technologies in addressing mental health and well-being tasks, such as depression screening and detection [51-55]. Concurrently, emerging research has begun to uncover short-term behavioral correlates of discrimination experiences [39,56,57], including changes in physical activity, sleep patterns, phone use, and social interactions—behaviors that can be objectively measured through smartphone and wearable sensors. However, most previous studies have focused on uncovering health and behavioral associations with perceived discrimination. While these analytical approaches provide valuable insights, they do not directly address the challenge of detection. Therefore, this study aims to fill this gap by developing and evaluating models that detect PUT after it occurs based on behavioral changes. To our knowledge, this is the first work to explore passive detection using mobile sensing data. Our goal is to establish a benchmark that can inform and advance future research in this emerging field.

The combination of high-dimensional mobile sensing features and the flexibility of machine learning techniques allows for a data-driven, scalable, and personalized approach compared with traditional statistical methods [58-62]. Moreover, mobile sensing via smartphones or wearables offers a ubiquitous, continuous, and nonintrusive means of data collection, making it a powerful tool for capturing momentary experiences more effectively than traditional survey-based approaches [63-66]. Importantly, our goal is not to replace human expertise, but to augment it using technology for early detection and intervention at scale [67-69].

Passive detection of PUT presents unique challenges. Such events are often sporadic, vary greatly in form [70,71], and are perceived subjectively across individuals [72-74], leading to significant variability in experiences and reporting. This makes it difficult for longitudinal studies like ours to collect sufficient samples of day-to-day self-reported incidents, with adequate variance in the ground truth. Nonetheless, traditional instruments for measuring PUT, such as self-reports, have well-documented limitations [28], including biases [75,76] and limited ability to account for confounding factors such as physical or mental stressors. These tools also rely on repeated measures [77], which can lead to participant fatigue and reduced compliance over time. Importantly, self-reporting is inherently episodic, making it less suited for continuous, population-level screening across a campus setting. In contrast, passive sensing offers the ability to continuously and unobtrusively monitor behavioral and physiological signals over time, rendering it a promising tool to complement traditional methods in detecting and understanding these complex psychosocial phenomena.

Recent advancements in machine learning–based rare event detection (RED) have shown promising results across a range of domains, including health care and mobile sensing [78,79]. While traditional ensemble methods such as random forests and gradient boosting have been used [80-82], deep learning architectures such as autoencoders (AEs) and long short-term memory networks (LSTMs) are gaining momentum [83-88]. This is particularly relevant in multivariate time series settings [88,89], where smartphones and wearable devices continuously generate multiple streams of time-stamped data (eg, location, activity, and phone use) that capture complex behavioral patterns over time [90-93]. Despite these advancements, RED remains a challenging task [78,79], especially given the infrequent nature of the events. This often leads to reduced quantitative performance, particularly in metrics such as precision, recall, and F_1-score, as studies across various domains frequently report only modest improvements over baseline methods [79]. For instance, Coley et al [94] reported a precision of 0.09, a recall

XSL•FO
RenderX

of 0.53, and an F_1-score of 0.16 using a random forest model for suicide risk detection on a health care dataset with a rarity of 0.2%. Closer to our context, Pillai et al [95] proposed a multitask learning framework in the Tesserae study [92] that combined an unsupervised AE with an auxiliary sequence prediction task to detect rare life events (<2%) using mobile sensing data. This approach improved performance compared with several baselines, achieving an F_1-score of 0.29. Such approaches have not yet been explored for detecting PUT experiences in social contexts like college environments. This gap presents an important opportunity to investigate the potential of mobile sensing-based RED methods in screening behavioral anomalies that may signal experiences of PUT.

Building on prior work [57,95], we developed and evaluated machine learning models leveraging mobile sensing data collected in 2018 as part of a multiyear study of undergraduate students [93,96]. We focused on two modeling approaches: (1) user-independent classification models, trained to identify behavioral patterns that are indicative of PUT across individuals; and (2) user-dependent anomaly detection models, personalized to detect short-term deviations in individual behavior that may signal responses to such experiences. Our objective is to assess the feasibility of using exclusively mobile sensing data, collected up until the time students wake the next morning, to detect experiences of PUT from the previous day at both the population and individual levels. We evaluated model performance against naive baselines using metrics commonly used in RED [97], including area under the receiver operating characteristic curve (AUC-ROC), area under the precision-recall curve (AUC-PR), precision, recall, and F_1-score.

Methods

Ethical Considerations

This work was approved by the University of Washington Institutional Review Board and was assigned the ID Study00003324. All participants in the study provided their informed consent in person. All data collected in the study were kept separate from participants' personal identifiers, to provide anonymity and protect privacy. Participants could receive compensation up to US $245/quarter in gift cards, depending on the completeness of their data collection, both passively sensed data and EMAs.

Data Collection

To be eligible for the study, participants were required to be over 18 years old, enrolled as first-year full-time undergraduates, and own an Android or iOS smartphone with an active data plan. The data collection period spanned 2 academic terms (January to June 2018), lasting approximately 20 weeks. At the beginning of the study, participants installed a smartphone app built using the AWARE framework [98] and wore a Fitbit tracker continuously throughout the study period, enabling passive data collection. This setup captured a wide range of mobile contextual information, including location, activity recognition, battery status, phone calls, screen use, and Bluetooth and Wi-Fi scans, while the Fitbit tracker provided steps and sleep data.

During the study, participants received regular EMA surveys [99] to report PUT experiences (Multimedia Appendix 1 contains survey questions). They completed EMAs twice weekly (on Sundays and Wednesdays) over an 8-week period, reporting on events from the previous day, and completed daily evening EMAs for an additional 2 weeks each academic term, reflecting on the same day's experiences. To identify which days participants experienced PUT, each EMA asked whether they had experienced unfair treatment on the reported day (today or yesterday).

Feature Extraction

In this study, we computed behavioral features from 7 smartphone data streams (activity recognition, battery, Bluetooth, call, location, screen, and Wi-Fi) and 2 wearable data streams (sleep and steps). We extended a behavioral feature extraction library [100] and followed a similar approach to prior work [90,101] by aggregating mobile sensing data into statistical summaries across various epochs of the day, including night (12 AM to 6 AM), morning (6 AM to 12 PM), afternoon (12 PM to 6 PM), and evening (6 PM to 12 AM), as well as the entire day (12 AM to 12 AM). This approach allowed us to characterize human behavior patterns within the day, providing a structured representation of daily behavior. Additionally, aligning our feature extraction process with prior work ensured methodological consistency, enabling direct comparisons and potential cross-study generalization. Notably, sleep features were computed only on a daily basis, as most student participants typically experienced a single major sleep episode each night. Some features were stream-specific, such as location variance derived solely from GPS data or the frequency of screen unlocks from screen events. Others involved multiple streams, for example, estimating indoor mobility duration by fusing location and activity recognition data. We describe the extraction of each set of behavioral features in Multimedia Appendix 2. In addition to these features, we computed the number of data samples collected for each data stream as an additional feature. This allowed us to assess data compliance and gain insights into event-based data streams such as calls, screen use, activity, and sleep, where the number of records reflects the frequency of these events.

Data Availability Analysis

Out of the 201 study participants, 110 individuals reported experiencing PUT at least once during the study period. To estimate how frequently PUT was reported overall, we calculated the proportion of positive EMA responses (indicating PUT) for each participant (ie, the number of positive responses divided by the total number of EMA responses they submitted). We then averaged these individual-level proportions across all participants, resulting in a mean reporting rate of 4.66% (95% CI 3.13% to 6.19%). The distribution of the EMA survey responses, with a total of 9629 submissions, included 413 (4.3%) positive responses and 9216 (95.7%) negative responses.

We observed a significant level of missing values in the behavioral features. Issues related to data collection, such as poor study compliance, phone battery depletion, or app crashes, directly contributed to the lack of raw sensing data. Additionally, event-based data streams, such as call logs, only record data

XSL•FO
RenderX

when specific events occur (eg, when a call is made), making it challenging to determine whether the absence of data was due to the absence of such events (eg, no calls made) or due to issues in data collection. Insufficient volume of raw sensing data per time period can also result in missing feature values. This is especially relevant for statistical features that require a reasonable number of samples for aggregated calculations. Similarly, features like Bluetooth and location often rely on a sufficient number of raw data records for effective data clustering. Last, the limited diversity of data streams per sample can affect the computation of fused features, as they depend on nonmissing values from multiple data streams. For instance, both location and activity recognition data are required for extracting features such as study duration and indoor mobility.

In Table 1, we report the availability of each data stream in the raw dataset, which serves as the foundation for feature calculation. We computed data availability as the percentage of daily samples where each data stream was available, relative to the total samples across all participants (total days × total participants). A higher percentage indicates broader availability.

Table 1. Availability of each data stream in the raw dataset.

Data stream	Availability (%)
Activity	49.93
Battery	43.61
Bluetooth	48.65
Calls	35.05
Locations	51.08
Screen	51.85
Wi-Fi	52.44
Sleep	65.17
Steps	70.52

Modeling

In this study, we focus on 2 modeling approaches (user-independent and -dependent) to retrospectively detect PUT using daily inference windows (Figure 1), implemented through 3 specific model types. The user-independent classification models leverage labeled data from training participants to detect PUT in new or unseen participants. In contrast, the user-dependent anomaly detection models learn from participants' historical data to differentiate between normal and rare patterns in their own behavior. These approaches are driven by practical considerations. First, the scarcity of labeled data makes it difficult to build a model that generalizes well to unseen users [61,93,96,102-105], though identifying broad patterns and key features remains essential in imbalanced settings [106]. Second, individuals' behavioral patterns before and after experiencing PUT can vary greatly [21,107-110] and are often moderated by various factors [111-115], highlighting the need for personalized training strategies. Third, given the rarity of target events, modeling them as anomalies is a commonly adopted approach in a wide variety of studies [78,79]. Furthermore, we believe these dual modeling approaches enhance the robustness of our models across diverse real-world deployment scenarios.

XSL•FO
RenderX

Figure 1. Illustration of detection time windows and inference times for three model types: (A) user-independent classification model, (B) intraday long short-term memory-autoencoder user-dependent model, and (C) multiday long short-term memory-autoencoder user-dependent model.

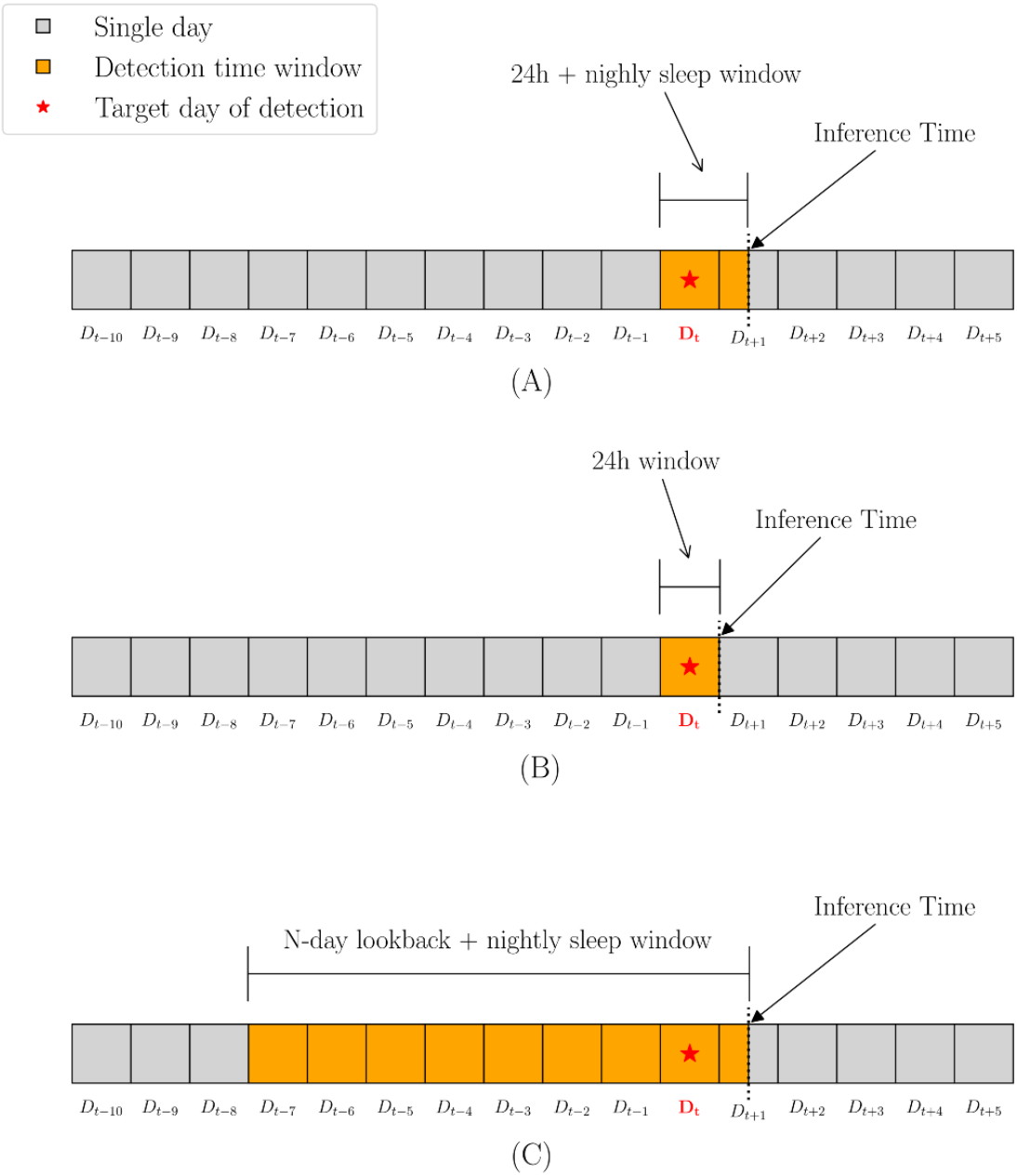

User-Independent Modeling

We selected light gradient boosting machine (LightGBM) [116] as our primary algorithm for user-independent modeling, for its ability to handle high-dimensional mobile sensing features, capture nonlinearity, mitigate overfitting through its ensemble mechanism, and natively manage missing data. LightGBM is a gradient boosting framework that builds decision trees efficiently, using a histogram-based approach to speed up training while maintaining high accuracy. Its built-in feature importance calculation enhances model interpretability and makes it an ideal choice for gaining population-wide insights in imbalanced settings. We benchmarked LightGBM with 4 classic machine learning algorithms: k-nearest neighbors (KNN), logistic regression, support vector machine (SVM), and random forest. All models were implemented using the scikit-learn Python library [117], with binary cross-entropy loss as the objective function. For comparison, we also implemented 3 baseline classifiers. The first uses only static demographic information (Multimedia Appendix 3) without incorporating any behavioral data. The other two are naive baselines, which make random predictions without considering input features: (1) a uniform classifier that assigns labels randomly with equal probability, ignoring class distribution; and (2) a stratified classifier that assigns labels randomly while preserving the class distribution observed in the data.

Prior to training, we applied a filtering step that required each sample to contain at least 7 available data streams, resulting in a final dataset of 4720 person-day records. Of these, 167 (3.5%) were labeled positive and 4553 (96.5%) as negative. Participants contributed an average of 24 days of usable data (SD 9.6). LightGBM natively handled missing values, while for the classic

XSL•FO
RenderX

models, we applied median imputation via scikit-learn's SimpleImputer. To address class imbalance, we experimented with class weight adjustments as well as 2 widely used oversampling approaches: the synthetic minority oversampling technique and SVM-based synthetic minority oversampling technique [118,119].

We chose the best-performing population-level model for subsequent feature analysis and selection, which was then used to build user-dependent models. As illustrated in Figure 1A, all models were trained on features extracted from a full 24-hour period on the target day, combined with sleep features from the nightly sleep window ending the following morning, after which inference was performed.

User-Dependent Modeling

For user-dependent modeling, we focused on anomaly detection with LSTM-AEs. LSTM-AEs are a type of neural network designed to learn temporal patterns in sequential data by encoding and reconstructing time-series inputs, making them effective for detecting deviations from typical behavioral patterns. Since human behavior typically follows daily routines and longer-term cycles (eg, weekly), we believe this approach helps detect anomalies in individual trajectories that may signal PUT. Our LSTM-AE architecture (Figure 2) consists of 2 stacked LSTM layers (encoder) that compress input sequences into a fixed-size representation, which is then replicated and decoded by another 2 LSTM layers (decoder). A final dense layer reconstructs the original sequence. The model learns normal behavioral patterns by minimizing reconstruction loss (mean squared error). We implemented the model architecture using Keras and TensorFlow libraries. We explored both intraday (1-day lookback) and multiday (N-day lookback) input constructions to assess whether behavioral changes associated with PUT are better detected through short-term or longer-term temporal context. The intraday model focuses on capturing short-term daily patterns by dividing each 24-hour period into 4 fixed 6-hour epochs: night (12 AM to 6 AM), morning (6 AM to 12 PM), afternoon (12 PM to 6 PM), and evening (6 PM to 12 AM), resulting in sequences with 4 aggregated data points. In contrast, the multiday model captures longer-term trends using an N-day lookback window, where each sequence includes N days of daily features, ending with the target day. For both intraday and multiday models, we selected features based on the top behavioral features identified by the best-performing user-independent model. For the intraday model, inference occurs daily at the end of the day (Figure 1B); whereas for the multiday model, inference is performed the following morning after sleep-related features are available (Figure 1C). During inference, anomaly scores for intraday sequences were derived from the reconstruction loss of the entire sequence, whereas for multiday sequences, only the loss of the target (last) day was used. Anomalies were then defined using a threshold set at the 75th percentile of training reconstruction errors, providing a conservative and data-driven decision rule.

Figure 2. LSTM autoencoder architecture. LSTM: long short-term memory.

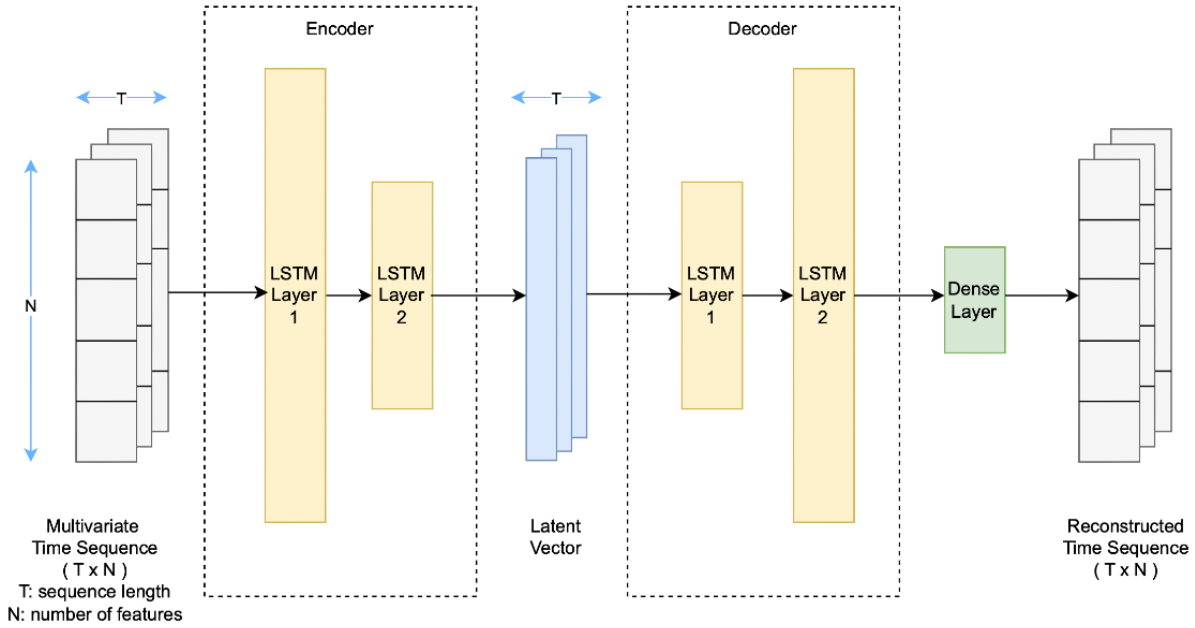

The models were trained exclusively on negative samples, with all positive samples reserved for testing, which is a common practice in RED [120-122]. Negative samples were split chronologically (90:10) into training and test sets. The training and test sample sizes are shown in Table 2. To prevent information leakage, multiday training samples overlapping with positive test samples were excluded. To ensure data quality, input sequences with more than 80% missing data were removed, and for multiday models, sequences with more than 50% missingness on the target day were also excluded. Reconstruction loss was computed only for observed (non-NaN) positions to minimize bias.

XSL•FO
RenderX

Table 2. Training and test sample sizes for user-dependent modeling with long short-term memory autoencoders.

Model	Train (Neg)	Test		Per user	
		Pos	Neg	Train, mean (SD)	Test, mean (SD)
Intraday models					
50 features	6595	311	833	33 (12)	6 (3)
100 features	6558	309	827	33 (12)	6 (3)
150 features	4659	199	618	23 (8)	4 (3)
Multiday models					
50 features, 3-day window	4428	191	598	22 (9)	4 (3)
50 features, 5-day window	4341	191	598	22 (9)	4 (3)
50 features, 7-day window	4299	190	596	21 (9)	4 (3)
50 features, 9-day window	4248	189	596	21 (9)	4 (3)

Additionally, for comparison, we implemented two naive baseline models: (1) a uniform classifier that assigns labels randomly with equal probability, ignoring class distribution; and (2) a stratified classifier that assigns labels randomly while preserving the class distribution observed in the data.

Model Evaluation

For the user-independent model, we implemented nested stratified group K-fold cross-validation to address the dataset's limited size and class imbalance. The outer loop (K=5) split the data into training and test sets, ensuring class balance and participant-level separation. The inner loop (K=4) further splits the training data for feature selection and hyperparameter tuning, following the same stratified, participant-dependent structure. We evaluated performance using threshold-independent metrics (AUC-PR and AUC-ROC) and threshold-dependent metrics (precision, recall, and F_1-score), reporting the mean and standard deviation across 25 independent performance estimates obtained from 5 repeats of 5-fold cross-validation with different random seeds. Threshold-dependent metrics are sensitive to both the chosen cutoff and outcome prevalence. To address this, we set the classification threshold to the empirical positive class probability from the training data, providing a consistent and data-driven operating point. Threshold-independent metrics complement this choice by characterizing performance across all thresholds, with AUC-PR being particularly informative under class imbalance [123,124]. Together, these metrics allow both robust model comparison and an understanding of the practical trade-offs between capturing rare events and limiting false alarms.

For the user-dependent anomaly detection model, we likewise evaluated performance using threshold-independent metrics (AUC-PR and AUC-ROC) and threshold-dependent metrics (precision, recall, and F_1-score), reporting the mean and standard deviation across 10 randomized runs with different model initializations. During inference, the anomaly threshold was conservatively set at the 75th percentile of training reconstruction errors, as the limited sample size precluded further fine-tuning. Threshold-independent metrics were computed from anomaly scores across test samples, while threshold-dependent metrics reflected performance at the selected threshold.

Results

Participants' Characteristics

The study focused on 201 full-time undergraduate students from the cohort enrolled in 2018 at the University of Washington. The mean age of the sample was 18.4 (SD 0.56) years. Female students comprised 64.7% (130/201 students) of the sample. Academically, students were drawn from a variety of departments, with approximately half majoring in engineering. To ensure a diverse sample, recruitment strategies included the oversampling of students from underrepresented backgrounds, specifically those with disabilities, first-generation students, and gender minorities.

User-Independent Modeling Results

Table 3 reports the performance metrics of all user-independent classification models alongside the baseline classifiers. To assess whether benchmarked models significantly outperformed the baselines, we conducted pairwise comparisons using the Wilcoxon signed-rank test (paired, 1-sided). Table 4 presents the P values from these comparisons, adjusted using the Benjamini-Hochberg false discovery rate procedure. The classification models showed mixed performance, with AUC-ROC ranging from 0.546 to 0.640, AUC-PR from 0.047 to 0.093, and F_1-scores from 0.070 to 0.121. Compared with the baselines, some benchmarked models showed modest improvements. Both Random Forest and LightGBM achieved higher AUC-ROC and AUC-PR scores compared with KNN, Logistic Regression, SVM, and the baseline classifiers. Among all models, LightGBM achieved the highest F_1-score. The observed variability across the cross-validation folds highlights the challenge of between-individual generalizability.

XSL•FO
RenderX

Table 3. Performance of user-independent classification models compared with baseline classifiers. Metrics are reported as mean (SD) across 25 estimates from repeated 5-fold cross-validation (5 repeats, different random seeds).

Model	AUC-ROC[a], mean (SD)	AUC-PR[b], mean (SD)	Precision, mean (SD)	Recall, mean (SD)	F_1-score, mean (SD)
Baseline models					
Uniform	0.500 (0.000)	0.036 (0.018)	0.034 (0.018)	0.492 (0.121)	0.063 (0.031)
Stratified	0.501 (0.022)	0.037 (0.018)	0.046 (0.048)	0.041 (0.040)	0.039 (0.036)
Demographic	0.523 (0.126)	0.057 (0.043)	0.051 (0.044)	0.267 (0.143)	0.083 (0.065)
User-independent models					
KNN[c]	0.561 (0.062)	0.049 (0.025)	0.037 (0.020)	0.836 (0.074)	0.070 (0.036)
Logistic Regression	0.546 (0.082)	0.047 (0.026)	0.041 (0.023)	0.579 (0.145)	0.075 (0.040)
SVM[d]	0.567 (0.095)	0.065 (0.045)	0.053 (0.039)	0.268 (0.139)	0.089 (0.056)
Random Forest	0.634 (0.086)	0.093 (0.094)	0.045 (0.022)	0.709 (0.143)	0.084 (0.038)
LightGBM[e]	0.640 (0.065)	0.077 (0.043)	0.083 (0.050)	0.275 (0.118)	0.121 (0.064)

[a]AUC-ROC: area under the receiver operating characteristic curve.

[b]AUC-PR: area under the precision-recall curve.

[c]KNN: k-nearest neighbors.

[d]SVM: support vector machine.

[e]LightGBM: light gradient boosting machine.

XSL•FO
RenderX

Table 4. Results of Wilcoxon signed-rank tests comparing user-independent models with baseline models. Reported *P* values were obtained from paired, one-sided Wilcoxon signed-rank tests and adjusted for multiple comparisons using the Benjamini-Hochberg false discovery rate (FDR) correction.

Baseline models	*P* values				
	AUC-ROC[a]	AUC-PR[b]	Precision	Recall	F_1-score
KNN[c]					
Uniform	<.001	<.001	.03	<.001	.004
Stratified	<.001	<.001	.72	<.001	.002
Demographic	.16	.88	.97	<.001	.79
Logistic regression					
Uniform	.007	<.001	.004	.006	.003
Stratified	.007	<.001	.57	<.001	<.001
Demographic	.33	.91	.96	<.001	.65
SVM[d]					
Uniform	.002	<.001	.002	>.99	.004
Stratified	.003	<.001	.16	<.001	<.001
Demographic	.16	.34	.56	.67	.54
Random forest					
Uniform	<.001	<.001	<.001	<.001	<.001
Stratified	<.001	<.001	.41	<.001	<.001
Demographic	.002	.03	.72	<.001	.23
LightGBM[e]					
Uniform	<.001	<.001	<.001	>.99	<.001
Stratified	<.001	<.001	.002	<.001	<.001
Demographic	<.001	.03	.003	.47	.013

[a]AUC-ROC: area under the receiver operating characteristic curve.

[b]AUC-PR: area under the precision-recall curve.

[c]KNN: k-nearest neighbors.

[d]SVM: support vector machine.

[e]LightGBM: Light Gradient Boosting Machine.

The pairwise comparisons indicated that both LightGBM and Random Forest outperformed all 3 baselines in terms of AUC-ROC and AUC-PR. LightGBM improved AUC-ROC by 22.4% (0.640 vs 0.523, $P<.001$) and Random Forest improved AUC-PR by 63.2% (0.093 vs 0.057, $P=.03$), both relative to the demographic baseline. While KNN, Logistic Regression, and Random Forest achieved significantly higher recall than the baselines, their precision values were significantly lower. LightGBM showed significant improvements in both precision and F_1-score compared with all 3 baselines, albeit at the cost of lower recall. Other benchmarked models showed smaller or nonsignificant differences, particularly compared with the demographic baseline. Overall, these findings suggest that tree-based models can improve the detection of PUT events, which we explore further in the Discussion.

User-Dependent Modeling Results

Intraday Models

Figure 3 plots the performance of the intraday LSTM-AE model as the number of input features increases, compared with 2 naive baselines. Models trained with 50 and 100 features did not yield improvements over the baselines except for recall, while the model trained with 150 features showed improvements in both recall and F_1-score. In general, models with more input features achieved higher recall.

XSL•FO
RenderX

Figure 3. Performance of the intraday anomaly detection model using varying numbers of input features, compared against naive baseline models. LSTM-AE: long short-term memory-autoencoder.

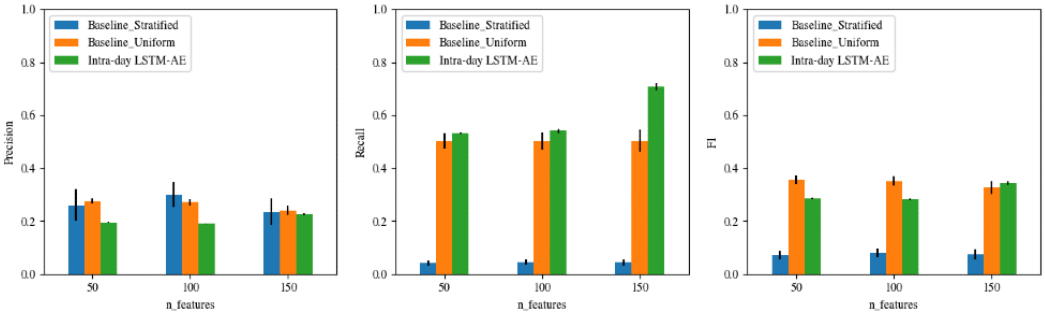

Multiday Models

Figure 4 plots the performance of the multiday LSTM-AE model trained on input sequences with 50 features across varying window sizes (3, 5, 7, and 9 days), compared with 2 naive baselines. All models consistently outperformed both the stratified baseline and the uniform baseline with respect to recall and F_1-score (P<.001 for both). The model with a 7-day window achieved the highest recall (0.836), while the 3-day model achieved the highest precision (0.279). All multiday LSTM-AE models had comparable F_1-scores (0.397-0.405). The 9-day window model showed a slight decline relative to the 7-day model in both recall (0.807 vs 0.836) and precision (0.397 vs 0.405), suggesting that extending the temporal window beyond 7 days yields diminishing returns.

Figure 4. Performance of the multiday anomaly detection model across different window sizes, compared against naive baseline models. LSTM-AE: long short-term memory-autoencoder.

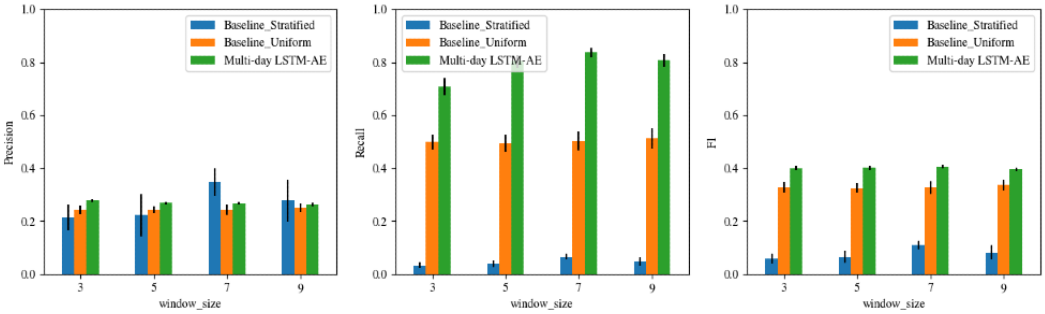

Table 5 summarizes the performance of the best-performing intraday and multiday models, alongside their respective naive baselines for reference. To assess whether benchmarked models significantly outperformed the baselines, we conducted pairwise comparisons using the Wilcoxon signed-rank test (paired, one-sided). Table 6 presents the P values from these comparisons, adjusted using the Benjamini-Hochberg false discovery rate procedure. As shown in Tables 5 and 6, the intraday model (150 features) did not show significant improvements over baselines except for recall (P<.001) and F_1-score (P=.012 vs uniform; P<.001 vs stratified). In contrast, the multiday model (50 features, 7-day window) significantly outperformed both baselines on nearly all evaluation metrics, except for precision against the stratified baseline. It achieved significantly higher recall (0.830 vs 0.479 for the uniform baseline, +73.3%, P<.001) and F_1-score (0.391 vs 0.313 for the uniform baseline, +24.9%, P<.001) with comparable precision (0.256), while improving AUC-PR by 45.9% (0.353 vs 0.242, P=.002) and AUC-ROC by 21.6% (0.605 vs 0.500, P=.002) relative to the baselines.

XSL•FO
RenderX

Table 5. Performance of user-dependent models compared with baseline classifiers. Metrics are reported as mean (SD) across 10 randomized runs with different model initializations.

Model	AUC-ROC[a], mean (SD)	AUC-PR[b], mean (SD)	Precision, mean (SD)	Recall, mean (SD)	F_1-score, mean (SD)
Intraday (150 features)					
Baseline (“uniform”)	0.500 (0.000)	0.244 (0.000)	0.243 (0.013)	0.495 (0.030)	0.326 (0.017)
Baseline (“stratified”)	0.500 (0.000)	0.244 (0.000)	0.263 (0.069)	0.046 (0.016)	0.078 (0.027)
LSTM-AE[c]	0.447 (0.004)	0.240 (0.002)	0.227 (0.003)	0.703 (0.010)	0.343 (0.005)
Multiday (50 features, 7-day window)					
Baseline (“uniform”)	0.500 (0.000)	0.242 (0.000)	0.233 (0.015)	0.479 (0.034)	0.313 (0.020)
Baseline (“stratified”)	0.500 (0.000)	0.242 (0.000)	0.250 (0.079)	0.044 (0.014)	0.074 (0.024)
LSTM-AE	0.608 (0.008)	0.353 (0.008)	0.256 (0.005)	0.830 (0.027)	0.391 (0.007)

[a]AUC-ROC: area under the receiver operating characteristic curve.

[b]AUC-PR: area under the precision-recall curve.

[c]LSTM-AE: long short-term memory-autoencoder.

Table 6. Results of the Wilcoxon signed-rank tests comparing user-dependent models with baseline models. Reported *P* values were obtained from paired, one-sided Wilcoxon signed-rank tests and adjusted for multiple comparisons using the Benjamini-Hochberg false discovery rate (FDR) correction.

Baseline	*P* values				
	AUC-ROC[a]	AUC-PR[b]	Precision	Recall	F_1-score
Intraday LSTM-AE[c] (150 features)					
Uniform	>.99	>.99	>.99	<.001	.012
Stratified	>.99	>.99	>.99	<.001	<.001
Multiday LSTM-AE (50 features, 7-day window)					
Uniform	.002	.002	<.001	<.001	<.001
Stratified	.002	.002	.99	<.001	<.001

[a]AUC-ROC: area under the receiver operating characteristic curve.

[b]AUC-PR: area under the precision-recall curve.

[c]LSTM-AE: long short-term memory-autoencoder.

Feature Importance Analysis

Analyzing feature importance in user-independent classification models is essential for identifying key behavioral features associated with reported PUT at the population level. In this study, we examined feature importance scores from the best-performing LightGBM model computed using information gain and averaged across all cross-validation folds. Higher scores indicate greater influence on the model’s decision. However, feature importance in tree-based models can be sensitive to correlated features [125] or sampling imbalance [126] and should be interpreted with caution. Future work could incorporate Shapley additive explanations [127] values to provide more robust insights into feature contributions and interactions in model decisions.

Table 7 summarizes the top 30 daily behavioral features ranked by their average importance scores. To better understand their associations with the outcome, we analyzed value distributions across positive and negative daily samples and computed standardized mean differences (SMDs) as effect sizes. Larger SMDs indicate stronger class separation. All top-ranked features had |SMD| values below 0.5, with 17 falling in the small-to-moderate range (|SMD| between 0.2 and 0.5) and 13 in the minimal separation range (|SMD| less than 0.2). This indicates that while these features may not independently distinguish between the 2 outcomes, they likely contribute to the model's decision through complex interactions or nonlinear associations.

Interestingly, the top 10 features are evenly distributed across a variety of sensing modalities, including 2 from steps, one fused feature derived from both activity and location, and one feature each from the remaining 6 data streams: location, activity, screen, Bluetooth, Wi-Fi, and call. This balanced distribution suggests that diverse behavioral signals, rather than any single dominant source, collectively contribute to the model’s detection ability. Across the top 30 features, the most frequently represented time epochs were the all-day window (n=9) and the evening period (n=7), indicating that both cumulative and evening-specific behavioral patterns are particularly informative for detection. As summarized in Textbox 1, these selected features touch upon key aspects of students’ daily lives, reflecting mobility, activity, phone use, social interactions, and sleep patterns.

XSL•FO
RenderX

Table 7. Top 30 daily behavioral features identified by the best-performing light gradient boosting machine model.

Feature	Data stream	Epoch	Positive, mean (SD)	Negative, mean (SD)	Importance score	Effect size (SMD[a])
Total time spent off-campus (minutes)	Location	All day	729.5 (505.8)	535.3 (435.3)	892.1	0.44
Total step count	Steps	Evening	4140.6 (2878.8)	3106.0 (2726.1)	628.5	0.38
Activity sample count	Activity	Night	171.7 (102.4)	139.6 (96.2)	595.8	0.33
Total screen time (minutes)	Screen	Afternoon	116.7 (70.1)	90.1 (64.9)	447.9	0.41
Bluetooth sample count	Bluetooth	Morning	25.2 (76.1)	45.9 (106.0)	415.3	–0.20
Number of unique Wi-Fi access points	Wi-Fi	All day	17.8 (16.9)	17.7 (24.0)	380.1	0.01
Number of missed calls	Call	Evening	1.2 (1.9)	0.8 (1.6)	290.2	0.25
Indoor mobility duration (minutes)	Activity/location	All day	48.3 (64.0)	65.2 (72.1)	289.2	–0.24
Sleep sample count	Sleep	All day	456.6 (128.9)	469.4 (112.0)	273.5	–0.11
Average duration of sedentary bouts (minutes)	Steps	Evening	37.8 (74.8)	37.2 (69.9)	257.2	0.01
Total number of active bouts	Steps	All day	56.9 (29.4)	52.3 (21.1)	192.2	0.21
Longest stay duration at study places (minutes)	Location	All day	50.6 (98.6)	45.4 (79.6)	175.5	0.06
Average stay duration off-campus (minutes)	Location	All day	137.6 (207.9)	91.1 (160.8)	175.4	0.29
Shortest phone interaction bout (minutes)	Screen	All day	10.5 (105.4)	3.5 (46.4)	173.2	0.14
First unlock time (seconds since midnight)	Screen	Night	1244.6 (2611.8)	1436.3 (3400.3)	159.2	–0.06
Average stay duration in green spaces (minutes)	Location	All day	27.4 (57.1)	26.1 (68.1)	146.4	0.02
Total sedentary time (minutes)	Steps	Evening	289.1 (48.9)	306.0 (40.4)	146.2	–0.42
Percentage of time spent off-campus	Location	All day	0.6 (0.4)	0.4 (0.3)	141.9	0.32
Shortest phone interaction bout (minutes)	Screen	Night	14.1 (44.2)	6.9 (34.3)	140.7	0.21
Activity sample count	Activity	Morning	203.2 (94.6)	211.0 (138.1)	135.1	–0.06
Sedentary bout duration variation (minutes)	Steps	Evening	28.5 (25.6)	29.8 (24.2)	132.8	–0.05
Circadian movement	Location	Night	2.3 (0.0)	2.3 (0.0)	132.7	0.02
Duration of physical activities (minutes)	Activity	Evening	84.8 (64.3)	64.9 (53.9)	131.9	0.37
Number of unique Wi-Fi access points	Wi-Fi	Afternoon	8.8 (8.6)	9.7 (38.8)	130.6	–0.03
Total time spent off-campus (minutes)	Location	Morning	171.1 (146.6)	117.8 (125.5)	130.0	0.42
Average steps per active bout	Steps	Evening	209.7 (137.0)	169.8 (141.0)	128.8	0.28
Last active bout end time (seconds since midnight)	Steps	Morning	40,457.5 (4216.2)	41,539.9 (2925.4)	127.9	–0.36
Variation in time spent in green spaces (minutes)	Location	Morning	3.2 (11.3)	1.3 (7.6)	120.9	0.23
Start time of Nightly sleep (seconds since midnight)	Sleep	All day	34,585.9 (7774.7)	33,569.8 (8917.1)	120.5	0.11

XSL•FO
RenderX

Feature	Data stream	Epoch	Positive, mean (SD)	Negative, mean (SD)	Importance score	Effect size (SMD[a])
Shortest phone interaction bout (minutes)	Screen	Afternoon	1.6 (18.9)	2.3 (19.8)	118.8	–0.03

[a]SMD: standardized mean difference.

Textbox 1. Examples of selected sensing features and their relation to students' daily behaviors.

- Campus-map features: Features such as time spent off-campus and indoor mobility show that students who experience perceived unfair treatment (PUT) tend to spend more time away from campus and exhibit reduced movement within campus buildings.
- Physical activity features: Features such as step count and activity duration show that these students tend to be more active during the evening and night hours.
- Screen use features: Longer afternoon screen time, longer minimum interaction durations, and earlier phone unlocks at night among students who report experiences of PUT. These features provide insights into phone use patterns and digital engagement.
- Bluetooth features: Lower sample counts in the morning for students who experience PUT. These features may serve as indicators of social exposure and proximity to others.
- Call features: A higher number of missed calls during evening hours among students reporting PUT. These features may reflect phone availability and social responsiveness.
- Sleep features: Shorter sleep durations and later sleep onset times on days marked by PUT.

Discussion

Principal Findings

In this study, we reported results from both classification and anomaly detection machine learning models evaluated under user-independent and user-dependent settings for detecting PUT among college students. A novel aspect of our work is the exclusive use of passively collected mobile sensing data for training and inference, offering a nonintrusive and low-burden alternative to traditional self-reports. As shown in Figure 1, our models were able to detect past events within a day of occurrence (including the subsequent nightly sleep window), indicating that mobile sensing may offer timely detection of PUT experiences and potential practical applications in future interventions. The best-performing classification model (LightGBM) significantly outperformed all 3 baseline classifiers, including the demographic baseline, in AUC-ROC, AUC-PR, precision, and F_1-score. The top-performing anomaly detection model, which incorporated a 7-day temporal context and 50 features, significantly outperformed the baselines across most metrics, achieving notably higher recall while maintaining comparable precision. These results suggest that mobile sensing features may serve as behavioral indicators for detecting experiences of PUT. While not intended to replace traditional assessments, mobile sensing could potentially complement existing methods, especially for large-scale or continuous monitoring.

Among the user-independent classification models, ensemble tree-based classifiers such as random forest and LightGBM consistently outperformed traditional machine learning algorithms like logistic regression and SVM, achieving higher overall performance. This result is consistent with previous work from various disciplines, highlighting the robustness of ensemble methods for handling imbalanced datasets [128-132] and the advantage of nonlinear models in capturing the complex behavioral patterns typically observed in mobile sensing data [90,93,133,134]. However, the overall limited performance across all models, along with the observed high variability across different user-independent data splits (eg, higher standard deviations in performance metrics), underscores the challenges of between-individual generalizability posed by the infrequent and subjective nature of the detection task. While user-independent models remain valuable for identifying globally informative features, future work should focus on personalized or semipersonalized modeling approaches that can better accommodate individual variability. Balancing scalability with personalization, for instance, by deploying personalized models for individuals identified as high-risk by a global model, may guide a more effective and targeted detection framework.

Our preliminary results suggest that daily behavioral signals alone, whether modeled in a user-independent setting (eg, classification models using daily features) or a user-dependent setting (eg, the intraday LSTM-AE model), were insufficient to reliably capture behavioral shifts associated with PUT. In contrast, models that incorporated multiday temporal windows demonstrated notable performance improvements. For example, all multiday anomaly detection models achieved statistically significant gains in recall and F_1-score compared with the corresponding baselines. This suggests that a longer temporal context may help detect more gradual or subtle behavioral deviations that may not be evident within a single day. Among the multiday models, the 7-day model achieved the best overall performance, while longer windows (eg, 9-day) showed reduced effectiveness. This suggests that the 7-day window likely strikes a balance between capturing sufficient behavioral context and remaining short enough to detect localized anomalies. It may also align with natural weekly rhythms that help reveal meaningful patterns. Taken together, these findings highlight the potential value of modeling temporal dynamics across multiple days. We recommend that future work identify and apply an optimal window length that balances contextual

XSL•FO
RenderX

richness with anomaly detectability to improve sensitivity without compromising precision.

In user-dependent anomaly detection, intraday models with 150 features outperformed those with 50 or 100 features, underscoring the value of a rich feature set for capturing behavioral anomalies. Feature importance analysis further indicated that performance gains were not driven by any single type of sensor data. Instead, a diverse set of behavioral indicators spanning multiple sensing modalities, including physical activity, phone use, mobility, and sleep, all contribute meaningfully to model performance. The absence of a dominant modality or feature suggests that effective detection likely relies on multimodal inputs and potentially their interactions to capture the complex, context-dependent nature of behavioral patterns rather than on isolated signals. These findings highlight the multifaceted nature of behavioral responses to PUT, though larger studies are needed to confirm these patterns. In the following section, we further contextualize these results by relating them to prior work.

Connections to Existing Literature

Our location-based features reveal that on days with reported PUT, students spend more time off-campus, which may reflect disengagement or withdrawal from campus life. Indoor mobility within campus buildings is also lower (eg, fewer transitions between classrooms, libraries, or other study areas), potentially indicating reduced academic or social participation. Moreover, Bluetooth sample counts are lower on these days, revealing fewer nearby Bluetooth-enabled devices, which may imply decreased social exposure. These findings support prior research showing that discrimination can undermine students' sense of belonging and social well-being, often resulting in increased feelings of isolation, social withdrawal, decreased academic and campus engagement, and even truancy [27,135-139]. Regarding physical activity, our features show that evening and nighttime activity tends to be higher on days with reported PUT, with greater step counts and longer durations of physical activity. These findings align with previous studies linking higher physical activity to perceived discrimination [109,140,141], possibly reflecting altered routines or coping behaviors. Our phone-use related features show longer afternoon screen time, earlier first unlocks, and longer shortest nightly phone interactions on days with reported PUT compared with days without. These patterns align with prior findings linking perceived discrimination to problematic phone use among students [142-144]. Finally, our sleep features show that students reporting PUT tend to have a later sleep onset and shorter sleep duration on the same day. This is consistent with previous research linking perceived discrimination to reduced sleep duration and poorer sleep quality [145-150].

While closely aligning with prior research, our study extends the literature by offering a unique short-term behavioral lens on the impact of PUT, complementing the longer-term patterns typically emphasized in the field. Many of the sensed patterns we identify likely reflect students' immediate behavioral and physiological responses to PUT. These patterns are not only informative for detection but can also guide interventions in campus environments, enabling continuous monitoring and timely, targeted support. While our models cannot prevent the initial event, they could facilitate timely interventions to mitigate its adverse effects and promote student well-being. For instance, our models could trigger personalized microinterventions delivered via push notifications [151-154], such as encouraging on-campus engagement, prompting positive social interaction, suggesting phone-use breaks, offering early sleep reminders and sleep hygiene education. These data-driven strategies can be integrated with traditional approaches, including support groups, self-care practices, and professional mental health services.

Cross-Domain Reflection: Rare Event Detection Challenges

We explored applying RED methodologies to the domain of detecting daily PUT. While we were able to connect some of our findings to existing literature on discrimination, most studies focused on retrospective self-reports, prevalence, and associated health outcomes. Relatively little attention has been given to detecting these events as they occur in daily life. To better situate our methodology and evaluate our model's performance, we conducted a comparative review of established RED approaches across diverse domains, including health care, crowd behavior, and mechanical systems. As summarized in Multimedia Appendix 4, these studies [94,95,155-158] use a wide range of methods and demonstrate anomaly detection performance broadly comparable to ours.

Although our research is situated in a different domain, we find significant value in such cross-domain reflection. First, it allows us to assess whether our progress in capturing PUT experiences aligns with the advancements achieved in other novel contexts. Second, it helps validate our modeling approaches. Given the novelty of our application, we view these comparisons as reflective benchmarks rather than direct performance evaluations.

In our work, we encountered several challenges inherent to RED, including a highly skewed class distribution, difficulties in capturing correlations due to data sparsity, and similarities between rare and nonrare events. These challenges are consistent with those faced by other RED studies across various domains [79]. A further commonality is the inherent trade-off between recall and precision. High recall indicates a model's ability to capture rare instances, but it is often accompanied by lower precision, suggesting that many predicted events may not be actual occurrences. This pattern is evident not only in our results but also across other domains. For instance, Pillai et al [95], in a context most similar to ours, achieved a recall of 0.21 and a precision of 0.47 using a dataset where rare events constituted approximately 1.9% of the data. In health care, however, models often deliberately favor recall over precision to ensure that critical cases are not missed. Inspired by this perspective, our study advocates for a recall-oriented approach that prioritizes identifying as many PUT experiences as possible, thereby maximizing opportunities to support students at risk.

A distinguishing aspect of our work is the relatively small dataset compared with those used in other studies, such as Coley et al [94] in suicide risk assessment. Despite this limitation, our

XSL•FO
RenderX

models achieved recall and F_1-scores that are comparable to, or in some cases exceed, those reported in similar domains, although direct comparisons are limited by dataset differences.

Limitations

We recognize several limitations in this pilot study. First, responses to PUT were collected at the daily level in a binary format (yes or no), without capturing the exact timing of each incident. This limited our ability to precisely align intraday sensor data patterns with specific experiences. For this reason, we cannot rule out that the intraday model's performance was affected by this limitation. Future work could benefit from more granular reporting to enable more accurate temporal analyses.

Second, our dataset was relatively small. While adequate for this exploratory pilot study, larger and more diverse datasets will be needed to validate and generalize these findings.

Third, this study focused exclusively on mobile sensing features. While this approach allowed us to isolate the predictive utility of passive sensing data, we acknowledge that the integration of additional domain knowledge, such as demographic and socioeconomic variables, or information about the type or reason for each incident, may improve model performance, interpretability, and fairness. Future research could reintroduce demographic and contextual information to account for heterogeneity across individuals and potentially enhance both predictive accuracy and fairness.

Finally, while our study EMAs were designed to directly ask about the target event and we aimed to detect PUT based on participants' behavioral responses, we acknowledge that many of the identified behavioral patterns (eg, social withdrawal, changes in phone use, sleep, and physical activity) could result from other negative experiences or mental health conditions such as depression, representing potential confounding factors. In addition, we cannot rule out the potential moderating influence of other variables on the association between PUT and the sensed features. Fully disentangling PUT-related signals from overlapping influences remains an open challenge and an important direction for future research.

Conclusions

This pilot study demonstrates the feasibility of using mobile sensing to screen for instances of PUT (4.3% or 413 of 9629 responses) among college students, providing a promising alternative to traditional self-report methods. Our machine learning models, leveraging diverse mobile sensing features and multiday temporal context, show strong potential for capturing short-term behavioral changes indicative of these infrequent experiences. We envision that future personalized and context-aware ML approaches, enhanced by larger datasets and deeper domain knowledge, will further improve detection accuracy, ultimately enabling timely interventions and support for at-risk students.

Acknowledgments

This material is based upon work supported by the National Science Foundation (grant numbers IIS1816687 and IIS7974751), and the University of Washington College of Engineering and the Paul G. Allen School of Computer Science and Engineering.

Data Availability

The datasets generated or analyzed during this study are available in the GLOBEM repository [159].

Authors' Contributions

YR and AD conceptualized the project. YR led the data curation, decisions about methodology, and the formal analysis, with support from RM and AD. AD provided the resources for the project. YR and RM led the writing of the original draft. All authors contributed to the reviewing and editing of the manuscript. AD and JM provided supervision for this project.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Ecological momentary assessment questions.
[PDF File (Adobe PDF File), 51 KB - formative_v9i1e78657_app1.pdf]

Multimedia Appendix 2
Behavioral feature extraction details.
[PDF File (Adobe PDF File), 49 KB - formative_v9i1e78657_app2.pdf]

Multimedia Appendix 3
Demographic variables used in baseline classifier.
[PDF File (Adobe PDF File), 54 KB - formative_v9i1e78657_app3.pdf]

XSL•FO
RenderX

Multimedia Appendix 4
Comparison of rare event detection studies.
[PDF File (Adobe PDF File), 60 KB - formative_v9i1e78657_app4.pdf]

References

1. Stuber J, Meyer I, Link B. Stigma, prejudice, discrimination and health. Soc Sci Med 2008;67(3):351-357 [FREE Full text] [doi: 10.1016/j.socscimed.2008.03.023] [Medline: 18440687]
2. Bravo AJ, Wedell E, Villarosa-Hurlocker MC, Looby A, Dickter CL, Schepis TS, Stimulant NormsPrevalence (SNAP) Study Team. Perceived racial/ethnic discrimination among young adult college students: prevalence rates and associations with mental health. J Am Coll Health 2023;71(7):2062-2073. [doi: 10.1080/07448481.2021.1954012] [Medline: 34398695]
3. Qeadan F, Azagba S, Barbeau WA, Gu LY, Mensah NA, Komaromy M, et al. Associations between discrimination and substance use among college students in the United States from 2015 to 2019. Addict Behav 2022;125:107164. [doi: 10.1016/j.addbeh.2021.107164] [Medline: 34735979]
4. Qeadan F, Madden EF, Barbeau WA, Mensah NA, Azagba S, English K. Associations between discrimination and adverse mental health symptoms and disorder diagnoses among college students in the United States. J Affect Disord 2022;310:249-257. [doi: 10.1016/j.jad.2022.05.026] [Medline: 35561881]
5. 2023 annual report. Center for Collegiate Mental Health (CCMH). URL: https://ccmh.psu.edu/assets/docs/2023_Annual%20Report.pdf [accessed 2025-05-08]
6. Sovern J. Is discrimination unfair? Ga State Univ Law Rev 2025;41(3):631 [FREE Full text]
7. Denise EJ, Hagiwara N. “Discrimination” versus “Unfair Treatment”: measuring differential treatment and its association with health. Sociol Inq 2019;89(4):645-676. [doi: 10.1111/soin.12277]
8. Sue DW, Capodilupo CM, Torino GC, Bucceri JM, Holder AMB, Nadal KL, et al. Racial microaggressions in everyday life: implications for clinical practice. Am Psychol 2007;62(4):271-286. [doi: 10.1037/0003-066x.62.4.271]
9. Suarez-Balcazar Y, Orellana-Damacela L, Portillo N, Rowan JM, Andrews-Guillen C. Experiences of differential treatment among college students of color. J High Educ 2016;74(4):428-444. [doi: 10.1080/00221546.2003.11780855]
10. Emmer C, Dorn J, Mata J. The immediate effect of discrimination on mental health: a meta-analytic review of the causal evidence. Psychol Bull 2024;150(3):215-252. [doi: 10.1037/bul0000419] [Medline: 38330346]
11. Thayer JF, Carnevali L, Sgiofo A, Williams DP. Angry in America: psychophysiological responses to unfair treatment. Ann Behav Med 2020;54(12):924-931. [doi: 10.1093/abm/kaaa094] [Medline: 33416840]
12. Ong AD, Deshpande S, Williams DR. Biological consequences of unfair treatment: a theoretical and empirical review. Handb Cult Biol 2017:279-315. [doi: 10.1002/9781119181361.ch12]
13. Harrell JP, Hall S, Taliaferro J. Physiological responses to racism and discrimination: an assessment of the evidence. Am J Public Health 2003;93(2):243-248. [doi: 10.2105/AJPH.93.2.243]
14. Gomez J, Miranda R, Polanco L. Acculturative stress, perceived discrimination, and vulnerability to suicide attempts among emerging adults. J Youth Adolesc 2011;40(11):1465-1476 [FREE Full text] [doi: 10.1007/s10964-011-9688-9] [Medline: 21717234]
15. Busby DR, Horwitz AG, Zheng K, Eisenberg D, Harper GW, Albucher RC, et al. Suicide risk among gender and sexual minority college students: the roles of victimization, discrimination, connectedness, and identity affirmation. J Psychiatr Res 2020;121:182-188 [FREE Full text] [doi: 10.1016/j.jpsychires.2019.11.013] [Medline: 31837538]
16. Mao Y, Liu L, Lu Z, Wang W. Relationships between perceived discrimination and suicidal ideation among impoverished Chinese college students: the mediating roles of social support and loneliness. Int J Environ Res Public Health 2022;19(12):7290 [FREE Full text] [doi: 10.3390/ijerph19127290] [Medline: 35742540]
17. Le TP, Iwamoto DK. A longitudinal investigation of racial discrimination, drinking to cope, and alcohol-related problems among underage Asian American college students. Psychol Addict Behav 2019;33(6):520-528. [doi: 10.1037/adb0000501] [Medline: 31414850]
18. Fahey MC, Morris JD, Robinson LA, Pebley K. Association between perceived discrimination and vaping among college students. Subst Use Misuse 2021;56(5):738-741. [doi: 10.1080/10826084.2021.1887250] [Medline: 33682601]
19. Steele R, Rosado A, Hernandez N, Brondolo E. Discrimination, acculturative stress, and academic achievement in emerging adults. J Vincent Soc Action 2022;6(1) [FREE Full text]
20. Stevens C, Liu CH, Chen JA. Racial/ethnic disparities in US college students' experience: discrimination as an impediment to academic performance. J Am Coll Health 2018;66(7):665-673. [doi: 10.1080/07448481.2018.1452745] [Medline: 29565755]
21. Kim J, Song K, Sutin AR. Gender differences in the relationship between perceived discrimination and personality traits in young adulthood: evidence using sibling fixed effects. Soc Sci Med 2021;286:114329. [doi: 10.1016/j.socscimed.2021.114329] [Medline: 34428601]
22. Schmidt CK, Miles JR, Welsh AC. Perceived discrimination and social support: the influences on career development and college adjustment of LGBT college students. J Career Dev 2010;38(4):293-309. [doi: 10.1177/0894845310372615]
23. Liu X, Sun X, Hao Q. Influence of discrimination perception on career exploration of higher vocational students: chain mediating effect test. Front Psychol 2022;13:968032 [FREE Full text] [doi: 10.3389/fpsyg.2022.968032] [Medline: 35967637]

XSL•FO
RenderX

24. Kwan MY, Gordon KH, Minnich AM. An examination of the relationships between acculturative stress, perceived discrimination, and eating disorder symptoms among ethnic minority college students. Eat Behav 2018;28:25-31. [doi: 10.1016/j.eatbeh.2017.12.003] [Medline: 29306178]
25. Harris CL, Haack S, Miao Z. Everyday discrimination is a stronger predictor of eating competence than food insecurity or perceived stress in college students amidst COVID-19. Appetite 2022;179:106300 [FREE Full text] [doi: 10.1016/j.appet.2022.106300] [Medline: 36075312]
26. Chaw A. The correlation between perceived discrimination and social anxiety in college students who identify as LGBTQ. Undergrad Res J 2023;3(2). [doi: 10.58361/2766-3590.1063]
27. Villegas-Gold R, Yoo HC. Coping with discrimination among Mexican American college students. J Couns Psychol 2014;61(3):404-413. [doi: 10.1037/a0036591] [Medline: 25019544]
28. Lewis TT, Cogburn CD, Williams DR. Self-reported experiences of discrimination and health: scientific advances, ongoing controversies, and emerging issues. Annu Rev Clin Psychol 2015;11(1):407-440. [doi: 10.1146/annurev-clinpsy-032814-112728] [Medline: 25581238]
29. Shammas D. Underreporting discrimination among Arab American and Muslim American community college students. J Mix Methods Res 2016;11(1):99-123. [doi: 10.1177/1558689815599467]
30. Irby-Shasanmi A, Leech TGJ. 'Because I Don't know': uncertainty and ambiguity in closed-ended reports of perceived discrimination in US health care. Ethn Health 2017;22(5):458-479. [doi: 10.1080/13557858.2016.1244659] [Medline: 27741709]
31. Sue DW, Alsaidi S, Awad MN, Glaeser E, Calle CZ, Mendez N. Disarming racial microaggressions: microintervention strategies for targets, White allies, and bystanders. Am Psychol 2019;74(1):128-142. [doi: 10.1037/amp0000296] [Medline: 30652905]
32. Awad MN, Connors EH. Active bystandership by youth in the digital era: microintervention strategies for responding to social media-based microaggressions and cyberbullying. Psychol Serv 2023;20(3):423-434. [doi: 10.1037/ser0000749] [Medline: 36951730]
33. Morton SCM, Everhart R, Dautovich N, Chukmaitov A. Perceived discrimination and mental health outcomes in college students: the mediating effect of preventive health behaviors and social support. J Am Coll Health 2025;73(6):2380-2389. [doi: 10.1080/07448481.2023.2286462] [Medline: 38010405]
34. Ajrouch KJ, Reisine S, Lim S, Sohn W, Ismail A. Perceived everyday discrimination and psychological distress: does social support matter? Ethn Health 2010;15(4):417-434 [FREE Full text] [doi: 10.1080/13557858.2010.484050] [Medline: 20582775]
35. Williams DR, Gonzalez HM, Williams S, Mohammed SA, Moomal H, Stein DJ. Perceived discrimination, race and health in South Africa. Soc Sci Med 2008;67(3):441-452 [FREE Full text] [doi: 10.1016/j.socscimed.2008.03.021] [Medline: 18486292]
36. Williams DR, Yan Yu, Jackson JS, Anderson NB. Racial differences in physical and mental health: socio-economic status, stress and discrimination. J Health Psychol 1997;2(3):335-351 [FREE Full text] [doi: 10.1177/135910539700200305] [Medline: 22013026]
37. Williams DR. Measuring discrimination resource. Psychology 1997;2(3):335-351. [doi: 10.1093/acprof:oso/9780199732166.003.0003]
38. Civitillo S, Jugert P. Zooming in on everyday ethnic-racial discrimination: a review of experiencing sampling methodology studies in adolescence. Eur J Dev Psychol Routledge 2024;21(4):592-611. [doi: 10.31234/osf.io/chxed]
39. Nam S, Jeon S, Ash G, Whittemore R, Vlahov D. Racial discrimination, sedentary time, and physical activity in African Americans: quantitative study combining ecological momentary assessment and accelerometers. JMIR Form Res 2021;5(6):e25687 [FREE Full text] [doi: 10.2196/25687] [Medline: 34096870]
40. Livingston NA, Flentje A, Heck NC, Szalda-Petree A, Cochran BN. Ecological momentary assessment of daily discrimination experiences and nicotine, alcohol, and drug use among sexual and gender minority individuals. J Consult Clin Psychol 2017 Dec;85(12):1131-1143 [FREE Full text] [doi: 10.1037/ccp0000252] [Medline: 29189029]
41. Tarrant MA, Manfredo MJ, Bayley PB, Hess R. Effects of recall bias and nonresponse bias on self-report estimates of angling participation. North Am J Fish Manag 1993;13(2):217-222. [doi: 10.1577/1548-8675(1993)013<0217:eorban>2.3.co;2]
42. Napa SC, Prieto CK, Diener E. In: Diener E, editor. Experience Sampling: Promises and Pitfalls, Strength and Weaknesses. Netherlands: Assess Well- Collect Works Ed Diener Dordrecht Springer; 2009:157-180.
43. Van Dyke ME, Kramer MR, Kershaw KN, Vaccarino V, Crawford ND, Lewis T. Inconsistent reporting of discrimination over time using the experiences of discrimination scale: potential underestimation of lifetime burden. Am J Epidemiol 2022;191(3):370-378 [FREE Full text] [doi: 10.1093/aje/kwab151] [Medline: 34017974]
44. Abernethy A, Adams L, Barrett M, Bechtel C, Brennan P, Butte A, et al. The promise of digital health: then, now, and the future. NAM Perspect 2022 [FREE Full text] [doi: 10.31478/202206e] [Medline: 36177208]
45. Naik N, Hameed BMZ, Sooriyaperakasam N, Vinayahalingam S, Patil V, Smriti K, et al. Transforming healthcare through a digital revolution: a review of digital healthcare technologies and solutions. Front Digit Health 2022;4:919985 [FREE Full text] [doi: 10.3389/fdgth.2022.919985] [Medline: 35990014]

XSL•FO
RenderX

46. The promise of digital healthcare technologies. Frontiers. URL: https://www.frontiersin.org/journals/public-health/articles/10.3389/fpubh.2023.1196596/full [accessed 2025-05-10]
47. Sameh A, Rostami M, Oussalah M, Korpelainen R, Farrahi V. Digital phenotypes and digital biomarkers for health and diseases: a systematic review of machine learning approaches utilizing passive non-invasive signals collected via wearable devices and smartphones. Artif Intell Rev 2024;58(2):66. [doi: 10.1007/s10462-024-11009-5]
48. Bhatt P, Liu J, Gong Y, Wang J, Guo Y. Emerging artificial intelligence-empowered mHealth: scoping review. JMIR Mhealth Uhealth 2022;10(6):e35053 [FREE Full text] [doi: 10.2196/35053] [Medline: 35679107]
49. Boukhechba M, Baglione AN, Barnes LE. Leveraging mobile sensing and machine learning for personalized mental health care. Ergon Des 2020;28(4):18-23. [doi: 10.1177/1064804620920494]
50. Mohr DC, Zhang M, Schueller SM. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. Annu Rev Clin Psychol 2017;13:23-47 [FREE Full text] [doi: 10.1146/annurev-clinpsy-032816-044949] [Medline: 28375728]
51. Khoo LS, Lim MK, Chong CY, McNaney R. Machine learning for multimodal mental health detection: a systematic review of passive sensing approaches. Sensors (Basel) 2024;24(2):348 [FREE Full text] [doi: 10.3390/s24020348] [Medline: 38257440]
52. Wang R, Wang W, Dasilva A, Huckins JF, Kelley WM, Heatherton TF, et al. Tracking depression dynamics in college students using mobile phone and wearable sensing. Proc ACM Interact Mob Wearable Ubiquitous Technol 2018;2(1):1-26. [doi: 10.1145/3191775] [Medline: 39449996]
53. Opoku AK, Visuri A, Ferreira DST. Towards early detection of depression through smartphone sensing. : Association for Computing Machinery; 2019 Presented at: Adjun Proc 2019 ACM Int Jt Conf Pervasive Ubiquitous Comput Proc 2019 ACM Int Symp Wearable Comput; 2019 May 10; New York, NY, USA p. 1158-1161. [doi: 10.1145/3341162.3347075]
54. Levine LM, Gwak M, Kärkkäinen K, Fazeli S, Zadeh B, Peris T, et al. In: Alam MM, Hämäläinen M, Mucchi L, Niazi IK, Le Moullec Y, editors. Anxiety Detection Leveraging Mobile Passive Sensing. Cham: Springer International Publishing; 2020:212-225.
55. Acikmese Y, Alptekin SE. Prediction of stress levels with LSTM and passive mobile sensors. Procedia Comput Sci 2019;159:658-667. [doi: 10.1016/j.procs.2019.09.221]
56. Tapia AL, Wallace ML, Hasler BP, Holmes J, Pedersen SL. Effect of daily discrimination on naturalistic sleep health features in young adults. Health Psychol 2024;43(4):298-309. [doi: 10.1037/hea0001359] [Medline: 38190204]
57. Sefidgar YS, Seo W, Kuehn KS, Althoff T, Browning A, Riskin E, et al. Passively-sensed behavioral correlates of discrimination events in college students. Proc ACM Hum Comput Interact 2019;3(CSCW):1-29 [FREE Full text] [doi: 10.1145/3359216] [Medline: 34322658]
58. Rajula HSR, Verlato G, Manchia M, Antonucci N, Fanos V. Comparison of conventional statistical methods with machine learning in medicine: diagnosis, drug development, and treatment. Medicina (Kaunas) 2020;56(9):455 [FREE Full text] [doi: 10.3390/medicina56090455] [Medline: 32911665]
59. Bennett M, Kleczyk EJ, Hayes K, Mehta R, Bennett M, Kleczyk EJ, et al. Evaluating similarities and differences between machine learning and traditional statistical modeling in healthcare analytics. In: Artificial Intelligence Annual Volume 2022. London, UK: IntechOpen; 2022.
60. Mann J, Lyons M, O'Rourke J, Davies S. Machine learning or traditional statistical methods for predictive modelling in perioperative medicine: a narrative review. J Clin Anesth 2025;102:111782. [doi: 10.1016/j.jclinane.2025.111782] [Medline: 39977974]
61. Adler DA, Wang F, Mohr DC, Choudhury T. Machine learning for passive mental health symptom prediction: generalization across different longitudinal mobile sensing studies. PLoS One 2022;17(4):e0266516 [FREE Full text] [doi: 10.1371/journal.pone.0266516] [Medline: 35476787]
62. Cho H, She J, Marchi DD, El-Zaatari H, Barnes EL, Kahkoska AR, et al. Machine learning and health science research: tutorial. J Med Internet Res 2024;26:e50890 [FREE Full text] [doi: 10.2196/50890] [Medline: 38289657]
63. Ságvári B, Gulyás A, Koltai J. Attitudes towards participation in a passive data collection experiment. Sensors (Basel) 2021;21(18):6085 [FREE Full text] [doi: 10.3390/s21186085] [Medline: 34577291]
64. Cornet VP, Holden RJ. Systematic review of smartphone-based passive sensing for health and wellbeing. J Biomed Inform 2018;77:120-132 [FREE Full text] [doi: 10.1016/j.jbi.2017.12.008] [Medline: 29248628]
65. Zhang H, Ibrahim A, Parsia B, Poliakoff E, Harper S. Passive social sensing with smartphones: a systematic review. Computing 2022;105(1):29-51. [doi: 10.1007/s00607-022-01112-2]
66. Maharjan SM, Poudyal A, van Heerden A, Byanjankar P, Thapa A, Islam C, et al. Passive sensing on mobile devices to improve mental health services with adolescent and young mothers in low-resource settings: the role of families in feasibility and acceptability. BMC Med Inform Decis Mak 2021;21(1):117 [FREE Full text] [doi: 10.1186/s12911-021-01473-2] [Medline: 33827552]
67. Singh S, Melendez K, Sezginis N. Examining the effect of discrimination and stigma on utilization of mental health services among college students. J Am Coll Health 2023;71(8):2398-2405. [doi: 10.1080/07448481.2021.1970561] [Medline: 34519631]

XSL•FO
RenderX

68. Madrid-Cagigal A, Kealy C, Potts C, Mulvenna MD, Byrne M, Barry MM, et al. Digital mental health interventions for university students with mental health difficulties: a systematic review and meta-analysis. Early Interv Psychiatry 2025;19(3):e70017. [doi: 10.1111/eip.70017] [Medline: 40033658]
69. Topooco N, Fowler LA, Fitzsimmons-Craft EE, DePietro B, Vázquez MM, Firebaugh M, et al. Digital interventions to address mental health needs in colleges: perspectives of student stakeholders. Internet Interv 2022;28:100528 [FREE Full text] [doi: 10.1016/j.invent.2022.100528] [Medline: 35378846]
70. Mikula G, Petri B, Tanzer N. What people regard as unjust: types and structures of everyday experiences of injustice. Eur J Soc Psychol 2006;20(2):133-149. [doi: 10.1002/ejsp.2420200205]
71. Vargas SM, Huey SJ, Miranda J. A critical review of current evidence on multiple types of discrimination and mental health. Am J Orthopsychiatry 2020;90(3):374-390. [doi: 10.1037/ort0000441] [Medline: 31999138]
72. Gong F, Xu J, Takeuchi DT. Racial and ethnic differences in perceptions of everyday discrimination. Sociol Race Ethn 2016;3(4):506-521. [doi: 10.1177/2332649216681587]
73. Gonzalez D, McDaniel M, Kenney GM, Skopec L. Perceptions of unfair treatment or judgment due to race or ethnicity in five settings. Robert Wood Johnson Foundation. URL: https://www.rwjf.org/en/insights/our-research/2021/07/perceptions-of-discriminatory-experiences-in-health-care-and-other-settings.html [accessed 2021-08-05]
74. Yang P, Henderson S. Race, gender, class, and perceived everyday discrimination. J Ethn Cult Stud 2024 Jul 30;11(3):51-66 [FREE Full text] [doi: 10.29333/ejecs/1801]
75. Kaiser CR, Major B. A social psychological perspective on perceiving and reporting discrimination. Law Soc Inq 2006;31(4):801-830. [doi: 10.1111/j.1747-4469.2006.00036.x]
76. Foster MD, Dion KL. Dispositional hardiness and women's well-being relating to gender discrimination: the role of minimization. Psychol Women Q 2003;27(3):197-208. [doi: 10.1111/1471-6402.00099]
77. Gaston SA, Jackson CL. Invited commentary: the need for repeated measures and other methodological considerations when investigating discrimination as a contributor to health. Am J Epidemiol 2022;191(3):379-383 [FREE Full text] [doi: 10.1093/aje/kwab223] [Medline: 34431505]
78. Abubakar YI, Othmani A, Siarry P, Sabri AQM. A systematic review of rare events detection across modalities using machine learning and deep learning. IEEE Access 2024;12:47091-47109. [doi: 10.1109/access.2024.3382140]
79. Shyalika C, Wickramarachchi R, Sheth AP. A comprehensive survey on rare event prediction. ACM Comput Surv 2024;57(3):1-39. [doi: 10.1145/3699955]
80. Bai Y, Huang Z, Lam H, Zhao D. Rare-event simulation for neural network and random forest predictors. ACM Trans Model Comput Simul 2022;32(3):1-33. [doi: 10.1145/3519385]
81. Islam MDK, Hridi P, Hossain MDS, Narman HS. Network anomaly detection using LightGBM: a gradient boosting classifier. 2020 Presented at: 2020 30th Int Telecommun Netw Appl Conf ITNAC; 2020 Nov 25; Melbourne, Australia p. 1-7. [doi: 10.1109/itnac50341.2020.9315049]
82. Blagus R, Lusa L. Gradient boosting for high-dimensional prediction of rare events. Comput Stat Data Anal 2017;113:19-37. [doi: 10.1016/j.csda.2016.07.016]
83. Provotar O, Linder Y, Veres M. Unsupervised anomaly detection in time series using LSTM-based autoencoders. 2019 Presented at: 2019 IEEE International Conference on Advanced Trends in Information Theory; 2019 Dec 18; Kyiv, Ukraine p. 513-517. [doi: 10.1109/atit49449.2019.9030505]
84. Said EM, Le-Khac NA, Dev S, Jurcut AD. Network anomaly detection using LSTM based autoencoder. : Association for Computing Machinery; 2020 Presented at: Proceedings of the 16th ACM Symposium on QoS and Security for Wireless and Mobile Networks; 2020 Nov 06; New York, NY, USA p. 37-45. [doi: 10.1145/3416013.3426457]
85. Wei Y, Jang-Jaccard J, Xu W, Sabrina F, Camtepe S, Boulic M. LSTM-autoencoder-based anomaly detection for indoor air quality time-series data. IEEE Sensors J 2023;23(4):3787-3800. [doi: 10.1109/jsen.2022.3230361]
86. Wang Z, Kasongo Dahouda M, Hwang H, Joe I. Explanatory LSTM-AE-based anomaly detection for time series data in marine transportation. IEEE Access 2025;13:23195-23208. [doi: 10.1109/access.2025.3535695]
87. Malhotra P, Ramakrishnan A, Anand G, Vig L, Agarwal P, Shroff G. LSTM-based encoder-decoder for multi-sensor anomaly detection. arXiv 2016. [doi: 10.48550/arXiv.1607.00148]
88. Pang G, Shen C, Cao L, Hengel AVD. Deep learning for anomaly detection. ACM Comput Surv 2021;54(2):1-38. [doi: 10.1145/3439950]
89. Li G, Jung JJ. Deep learning for anomaly detection in multivariate time series: approaches, applications, and challenges. Inf Fusion 2023;91:93-102. [doi: 10.1016/j.inffus.2022.10.008]
90. Wang R, Chen F, Chen Z, Li T, Harari G, Tignor S, et al. StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. : Association for Computing Machinery; 2014 Presented at: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing; 2014 Jun 10; New York, NY, USA p. 3-14. [doi: 10.1145/2632048.2632054]
91. Yao S, Hu S, Zhao Y, Zhang A, Abdelzaher T. DeepSense: A Unified Deep Learning Framework for Time-Series Mobile Sensing Data Processing. : International World Wide Web Conferences Steering Committee; 2017 Presented at: Proceedings of the 26th International Conference on World Wide Web; 2017 Apr 03; Geneva p. 351-360. [doi: 10.1145/3038912.3052577]

XSL•FO
RenderX

92. Mattingly SM, Gregg JM, Audia P, Bayraktaroglu AE, Campbell A, Chawla N, et al. The tesserae project: large-scale, longitudinal, in situ, multimodal sensing of information workers. : Association for Computing Machinery; 2019 Presented at: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems; 2019 May 03; New York, NY, USA p. 1-8. [doi: 10.1145/3290607.3299041]
93. Xu X, Liu X, Zhang H, Wang W, Nepal S, Sefidgar Y, et al. GLOBEM: Cross-dataset generalization of longitudinal human behavior modeling. : Association for Computing Machinery; 2023 Presented at: Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies; December 2022; New York, NY, United States p. 1-34. [doi: 10.1145/3569485]
94. Coley RY, Liao Q, Simon N, Shortreed SM. Empirical evaluation of internal validation methods for prediction in large-scale clinical data with rare-event outcomes: a case study in suicide risk prediction. BMC Med Res Methodol 2023;23(1):33. [doi: 10.1186/s12874-023-01844-5] [Medline: 36721082]
95. Pillai A, Nepal S, Campbell A. Rare life event detection via mobile sensing using multi-task learning. arXiv 2023. [doi: 10.48550/arXiv.2305.20056]
96. Xu X, Zhang H, Sefidgar Y, Ren Y, Liu X, Seo W, et al. GLOBEM dataset: multi-year datasets for longitudinal human behavior modeling generalization. arXiv 2023. [doi: 10.1145/3686138.3686147]
97. Adhikari S, Normand SL, Bloom J, Shahian D, Rose S. Revisiting performance metrics for prediction with rare outcomes. Stat Methods Med Res 2021;30(10):2352-2366 [FREE Full text] [doi: 10.1177/09622802211038754] [Medline: 34468239]
98. Ferreira D, Kostakos V, Dey AK. AWARE: mobile context instrumentation framework. Front ICT 2015;2:2. [doi: 10.3389/fict.2015.00006]
99. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol 2008;4:1-32. [doi: 10.1146/annurev.clinpsy.3.022806.091415] [Medline: 18509902]
100. Doryab A, Chikarsel P, Liu X, Dey AK. Extraction of behavioral features from smartphone and wearable data. arXiv 2019;21(2):8-10. [doi: 10.5260/chara.21.2.8]
101. Wang R, Harari G, Hao P, Zhou X, Campbell A. SmartGPA: how smartphones can assess and predict academic performance of college students. : Association for Computing Machinery; 2015 Presented at: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing; 2015 Sep 11; New York, NY, USA p. 295-306. [doi: 10.1145/2750858.2804251]
102. Saeb S, Lonini L, Jayaraman A, Mohr DC, Kording KP. The need to approximate the use-case in clinical machine learning. Gigascience 2017;6(5):1-9 [FREE Full text] [doi: 10.1093/gigascience/gix019] [Medline: 28327985]
103. Gu X, Deligianni F, Han J, Liu X, Chen W, Yang GZ, et al. Beyond supervised learning for pervasive healthcare. IEEE Rev Biomed Eng 2024;17:42-62. [doi: 10.1109/rbme.2023.3296938]
104. Meegahapola L, Droz W, Kun P, de Götzen A, Nutakki C, Diwakar S, et al. Generalization and personalization of mobile sensing-based mood inference models. : Association for Computing Machinery; 2023 Presented at: Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies; December 2022; New York, NY, United States p. 1-32. [doi: 10.1145/3569483]
105. Woll S, Birkenmaier D, Biri G, Nissen R, Lutz L, Schroth M, et al. Applying AI in the context of the association between device-based assessment of physical activity and mental health: systematic review. JMIR Mhealth Uhealth 2025;13:e59660. [doi: 10.2196/59660] [Medline: 40053765]
106. Japkowicz N, Stephen S. The class imbalance problem: a systematic study. Intell Data Anal 2002;6(5):429-449. [doi: 10.3233/ida-2002-6504]
107. Kim Y, Jung J, Na J. Socioeconomic status differences in psychological responses to unfair treatments: behavioral evidence of a vicious cycle. PLoS One 2022;17(6):e0268286 [FREE Full text] [doi: 10.1371/journal.pone.0268286] [Medline: 35687607]
108. Barclay LJ, Kiefer T. In the aftermath of unfair events: understanding the differential effects of anxiety and anger. J Manag 2017;45(5):1802-1829. [doi: 10.1177/0149206317739107]
109. Pascoe EA, Smart Richman L. Perceived discrimination and health: a meta-analytic review. Psychol Bull 2009;135(4):531-554 [FREE Full text] [doi: 10.1037/a0016059] [Medline: 19586161]
110. Williams DR, Mohammed SA. Discrimination and racial disparities in health: evidence and needed research. J Behav Med 2009;32(1):20-47 [FREE Full text] [doi: 10.1007/s10865-008-9185-0] [Medline: 19030981]
111. Sloan MM. Unfair treatment in the workplace and worker well-being. Work Occup 2011;39(1):3-34. [doi: 10.1177/0730888411406555]
112. Cariello AN, Perrin PB, Williams CD, Espinoza GA, Paredes AM, Moreno OA. Moderating influence of social support on the relations between discrimination and health via depression in Latinx immigrants. J Lat Psychol 2022;10(2):98-111 [FREE Full text] [doi: 10.1037/lat0000200] [Medline: 35434535]
113. Moore J, Tilki M, Clarke L, Waters E. The moderating effect of functional social support on the association between unfair treatment and self-rated health: a study of the resilience of a community-based sample of Irish migrants in London. Ir J Sociol 2018;26(3):267-288. [doi: 10.1177/0791603518811775]
114. Xu YE, Chopik WJ. Identifying moderators in the link between workplace discrimination and health/well-being. Front Psychol 2020;11:458 [FREE Full text] [doi: 10.3389/fpsyg.2020.00458] [Medline: 32256433]

XSL•FO
RenderX

115. Meier LL, Semmer NK, Hupfeld J. The impact of unfair treatment on depressive mood: the moderating role of self-esteem level and self-esteem instability. Pers Soc Psychol Bull 2009;35(5):643-655. [doi: 10.1177/0146167208331337] [Medline: 19244135]
116. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: a highly efficient gradient boosting decision tree. 2017 Presented at: 31st Conference on Neural Information Processing Systems; 04 December 2017; Long Beach, CA, USA URL: https://proceedings.neurips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html
117. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in python. J Mach Learn Res 2011;12(85):2825-2830 [FREE Full text] [doi: 10.5555/1953048.2078195]
118. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002;16:321-357. [doi: 10.1613/jair.953]
119. Nguyen HM, Cooper EW, Kamei K. Borderline over-sampling for imbalanced data classification. Int J Knowl Eng Soft Data Paradigm 2011;3(1):4. [doi: 10.1504/ijkesdp.2011.039875]
120. Chandola V, Banerjee A, Kumar V. Anomaly detection: s survey. ACM Comput Surv 2009;41(3):1-58. [doi: 10.1145/1541880.1541882]
121. Pimentel MAF, Clifton DA, Clifton L, Tarassenko L. A review of novelty detection. Signal Process 2014;99:215-249. [doi: 10.1016/j.sigpro.2013.12.026]
122. Schölkopf B, Platt JC, Shawe-Taylor J, Smola AJ, Williamson RC. Estimating the support of a high-dimensional distribution. Neural Comput 2001;13(7):1443-1471. [doi: 10.1162/089976601750264965] [Medline: 11440593]
123. Sofaer HR, Hoeting JA, Jarnevich CS. The area under the precision‐recall curve as a performance metric for rare binary events. Methods Ecol Evol 2019;10(4):565-577. [doi: 10.1111/2041-210x.13140]
124. Cook J, Ramadas V. When to consult precision-recall curves. Stata J 2020 Mar 24;20(1):131-148 [FREE Full text] [doi: 10.1177/1536867X20909693]
125. Salih AM. Explainable artificial intelligence and multicollinearity?: a mini review of current approaches. arXiv 2024. [doi: 10.48550/arXiv.2406.11524]
126. Dube L, Verster T. Interpretability of the random forest model under class imbalance. Data Sci Finance Econ 2024;4(3):446-468. [doi: 10.3934/dsfe.2024019]
127. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. : Curran Associates Inc; 2017 Presented at: Proceedings for the 31st International Conference on Neural Information Processing Systems; 2017 Dec 04; NY, USA p. 4768-4777.
128. Kaur H, Pannu HS, Malhi AK. A systematic review on imbalanced data challenges in machine learning. ACM Comput Surv 2019;52(4):1-36. [doi: 10.1145/3343440]
129. Haixiang G, Yijing L, Shang J, Mingyun G, Yuanyue H, Bing G. Learning from class-imbalanced data: review of methods and applications. Expert Syst Appl 2017;73:220-239. [doi: 10.1016/j.eswa.2016.12.035]
130. Tanha J, Abdi Y, Samadi N, Razzaghi N, Asadpour M. Boosting methods for multi-class imbalanced data classification: an experimental review. J Big Data 2020;7(1):70. [doi: 10.1186/s40537-020-00349-y]
131. Liu L, Wu X, Li S, Li Y, Tan S, Bai Y. Solving the class imbalance problem using ensemble algorithm: application of screening for aortic dissection. BMC Med Inform Decis Mak 2022;22(1):82. [doi: 10.1186/s12911-022-01821-w] [Medline: 35346181]
132. Khan AA, Chaudhari O, Chandra R. A review of ensemble learning and data augmentation models for class imbalanced problems: combination, implementation and evaluation. Expert Syst Appl 2024;244:122778. [doi: 10.1016/j.eswa.2023.122778]
133. Doryab A, Villalba DK, Chikersal P, Dutcher JM, Tumminia M, Liu X, et al. Identifying behavioral phenotypes of loneliness and social isolation with passive sensing: statistical analysis, data mining and machine learning of smartphone and fitbit data. JMIR Mhealth Uhealth 2019;7(7):e13209 [FREE Full text] [doi: 10.2196/13209] [Medline: 31342903]
134. Saeb S, Zhang M, Karr CJ, Schueller SM, Corden ME, Kording KP, et al. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. J Med Internet Res 2015;17(7):e175 [FREE Full text] [doi: 10.2196/jmir.4273] [Medline: 26180009]
135. Verkuyten M, Thijs J, Gharaei N. Discrimination and academic (dis)engagement of ethnic-racial minority students: a social identity threat perspective. Soc Psychol Educ 2019;22(2):267-290. [doi: 10.1007/s11218-018-09476-0]
136. Smith HJ, Jaurique A, Ryan D. The mistreatment of others: discrimination can undermine university identification, student health, and engagement. Soc Just Res 2016;29(4):355-374. [doi: 10.1007/s11211-016-0274-x]
137. Nakhaie R. Discrimination, psychological isolation, and flight from school. J Int Migr Integr 2021;23(3):1515-1541. [doi: 10.1007/s12134-021-00901-6]
138. Gyan C, Baskh B, Song W, Yeboah AS. “Withdrawal Syndrome”: the effects of acts of microaggression in the classroom on racialized students. Appl Res Qual Life 2024;19(6):3169-3187. [doi: 10.1007/s11482-024-10373-2]
139. Diehl D, Houseworth J, Grier-Reed T. Examining the variable relationship between race and considerations of campus withdrawal. Coll Stud J 2020;53(4):417-429 [FREE Full text]

XSL•FO
RenderX

140. Martins Neto C, Confortin SC, Lima ABS, Mouzinho LSN, Oliveira BLCAD. Association between perceived discrimination and physical activity among adolescents. Ciênc Saúde Coletiva 2022;27(10):4003-4013. [doi: 10.1590/1413-812320222710.05992022en]
141. Sims M, Diez-Roux AV, Gebreab SY, Brenner A, Dubbert P, Wyatt S, et al. Perceived discrimination is associated with health behaviours among African-Americans in the Jackson heart study. J Epidemiol Community Health 2016;70(2):187-194. [doi: 10.1136/jech-2015-206390] [Medline: 26417003]
142. Zhu J, Xie R, Chen Y, Zhang W. Relationship between parental rejection and problematic mobile phone use in Chinese university students: mediating roles of perceived discrimination and school engagement. Front Psychol 2019;10:428 [FREE Full text] [doi: 10.3389/fpsyg.2019.00428] [Medline: 30890982]
143. Guan W, Wang S, Liu C. Influence of perceived discrimination on problematic smartphone use among Chinese deaf and hard-of-hearing students: serial mediating effects of sense of security and social avoidance. Addict Behav 2023;136:107470. [doi: 10.1016/j.addbeh.2022.107470] [Medline: 36055057]
144. Li W, Xu T, Diao L, Wu Q. The impact of perceived discrimination on mobile phone addiction among Chinese higher vocational college students: a chain mediating role of negative emotions and learning burnout. Psychol Res Behav Manag 2024;17:401-411 [FREE Full text] [doi: 10.2147/PRBM.S440958] [Medline: 38343428]
145. Slopen N, Lewis TT, Williams DR. Discrimination and sleep: a systematic review. Sleep Med 2016;18:88-95 [FREE Full text] [doi: 10.1016/j.sleep.2015.01.012] [Medline: 25770043]
146. Park K, Kim J. Longitudinal association between perceived discrimination and sleep problems among young adults in the United States: tests of moderation by race/ethnicity and educational attainment. Soc Sci Med 2023;321:115773. [doi: 10.1016/j.socscimed.2023.115773] [Medline: 36801753]
147. Majeno A, Tsai KM, Huynh VW, McCreath H, Fuligni AJ. Discrimination and sleep difficulties during adolescence: the mediating roles of loneliness and perceived stress. J Youth Adolesc 2018;47(1):135-147 [FREE Full text] [doi: 10.1007/s10964-017-0755-8] [Medline: 29164378]
148. Johnson DA, Lewis TT, Guo N, Jackson CL, Sims M, Wilson JG, et al. Associations between everyday discrimination and sleep quality and duration among African-Americans over time in the Jackson heart study. Sleep 2021;44(12):zsab162 [FREE Full text] [doi: 10.1093/sleep/zsab162] [Medline: 34197610]
149. Gordon AM, Prather AA, Dover T, Espino-Pérez K, Small P, Major B. Anticipated and experienced ethnic/racial discrimination and sleep: a longitudinal study. Pers Soc Psychol Bull 2020;46(12):1724-1735. [doi: 10.1177/0146167220928859] [Medline: 32571161]
150. Fuller-Rowell TE, Nichols OI, Burrow AL, Ong AD, Chae DH, El-Sheikh M. Day-to-day fluctuations in experiences of discrimination: associations with sleep and the moderating role of internalized racism among African American college students. Cultur Divers Ethnic Minor Psychol 2021;27(1):107-117. [doi: 10.1037/cdp0000342] [Medline: 32309971]
151. Nahum-Shani I, Smith SN, Spring BJ, Collins LM, Witkiewitz K, Tewari A, et al. Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. Ann Behav Med 2018;52(6):446-462 [FREE Full text] [doi: 10.1007/s12160-016-9830-8] [Medline: 27663578]
152. Hsu TC, Whelan P, Gandrup J, Armitage CJ, Cordingley L, McBeth J. Personalized interventions for behaviour change: a scoping review of just-in-time adaptive interventions. Br J Health Psychol 2025;30(1):e12766. [doi: 10.1111/bjhp.12766] [Medline: 39542743]
153. Carpenter SM, Menictas M, Nahum-Shani I, Wetter DW, Murphy SA. Developments in mobile health just-in-time adaptive interventions for addiction science. Curr Addict Rep 2020;7(3):280-290 [FREE Full text] [doi: 10.1007/s40429-020-00322-y] [Medline: 33747711]
154. Improving engagement and efficacy of mHealth micro-interventions for stress coping: an in-the-wild study. arxiv. URL: https://arxiv.org/abs/2407.11612 [accessed 2025-05-10]
155. Dong Y, Pinelli F, Gkoufas Y, Nabi Z, Calabrese F, Chawla N. Inferring unusual crowd events from mobile phone call detail records. In: Machine Learning and Knowledge Discovery in Databases. Springer, Cham: Springer International Publishing; 2015:474-492.
156. Cheon SP, Kim S, Lee SY, Lee CB. Bayesian networks based rare event prediction with sensor data. Knowl-Based Syst 2009;22(5):336-343. [doi: 10.1016/j.knosys.2009.02.004]
157. Dangut MD, Jennions IK, King S, Skaf Z. Application of deep reinforcement learning for extremely rare failure prediction in aircraft maintenance. Mech Syst Signal Process 2022;171:108873. [doi: 10.1016/j.ymssp.2022.108873]
158. Wong ZSY. Statistical classification of drug incidents due to look-alike sound-alike mix-ups. Health Inf J 2016;22(2):276-292 [FREE Full text] [doi: 10.1177/1460458214555040] [Medline: 25391848]
159. Generalization of longitudinal behavior modeling. GLOBEM. URL: https://the-globem.github.io [accessed 2025-08-15]

Abbreviations

AE: autoencoder
AUC-PR: area under the precision-recall curve
AUC-ROC: area under the receiver operating characteristic curve

XSL•FO
RenderX

EMA: ecological momentary assessment
KNN: k-nearest neighbors
LightGBM: light gradient boosting machine
LSTM: long short-term memory
PUT: perceived unfair treatment
RED: rare event detection
SMD: standardized mean difference
SVM: support vector machine

Edited by A Mavragani; submitted 11.06.25; peer-reviewed by GE Fronk; comments to author 10.07.25; revised version received 03.09.25; accepted 21.09.25; published 31.10.25.

Please cite as:
Ren Y, Mulukutla R, Mankoff J, Dey AK
Detecting Perceived Unfair Treatment Among US College Students Using Mobile Sensing: Pilot Machine Learning Study
JMIR Form Res 2025;9:e78657
URL: https://formative.jmir.org/2025/1/e78657
doi:10.2196/78657
PMID:

©Yiyi Ren, Raghu Mulukutla, Jennifer Mankoff, Anind K Dey. Originally published in JMIR Formative Research (https://formative.jmir.org), 31.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

User Experiences of Limb-Worn Wearable Devices for Monitoring Parkinson Disease Motor Function and Blood Pressure: Usability Study

Lorna Kenny[1], BSocSc, MSocSc; Marco Sica[2], BSc, MSc; Colum Crowe[2], BE, ME; Lauren O'Mahony[1], BSc; Brendan O'Flynn[2], MEng, PhD, BEng; David Scott Mueller[3], BS; Salvatore Tedesco[2], BSc, MSc, PhD; John Barton[2], BE, MEng; Suzanne Timmons[1], MD

[1]Centre for Gerontology and Rehabilitation, School of Medicine, University College Cork, Cork, Ireland

[2]Tyndall National Institute, University College Cork, Cork, Ireland

[3]AbbVie (United States), North Chicago, IL, United States

Corresponding Author:
Lorna Kenny, BSocSc, MSocSc
Centre for Gerontology and Rehabilitation
School of Medicine
University College Cork
Centre for Gerontology and Rehabilitation, School of Medicine
University College Cork, St Finbarr's Hospital, Douglas Road
Cork, T12 XH60
Ireland
Phone: 353 851085070
Email: lorna.kenny@ucc.ie

Abstract

Background: Wearable devices have the potential to provide reliable and objective assessment and monitoring of Parkinson disease (PD). During design, there is often a focus on technical performance, accuracy, and reliability, with less emphasis on the user experience.

Objective: This study explored the user experience of a novel prototype wearable device (wrist- and ankle-worn) to record limb movements (accelerations and angular velocities), and physiological data (eg, photoplethysmography and heart rate information for estimation of blood pressure).

Methods: This qualitative study used internet-based semistructured interviews with people with PD, following wearing prototype devices for 24 hours at home. Interviews were audio-recorded and transcripts analyzed using a hybrid deductive and inductive approach.

Results: Six people with PD, 3 male and 3 female, aged 52-83 years, with mild-to-moderate PD (Hoehn and Yahr scale score ≤3; median MDS-UPDRS [Movement Disorder Society - Unified Parkinson's Disease Rating Scale] score of 72/199) and cognition within normal limits (6-item Cognitive Impairment Test median score 0), with an average disease duration of 8 years took part in the study. Participants were overall positive toward the device, finding it generally comfortable, light in weight and noninvasive. Five out of 6 participants reported minor problems related to strap adjustability and challenges specific to PD. The prototype was comfortable, but this was a lesser priority than robustness, the device not hindering their usual clothing choices (device size or outward projection, or both), and adjustability for fit, including the need to switch to an elasticated strap. In particular, a discreet design was important, as some individuals may feel self-conscious about wearing visible condition-specific products, due to stigma. The ankle-worn device was perceived as unfamiliar and nondiscreet, with some participants likening it to a prisoner tracking system and dressing specifically to conceal it. The wrist-worn device was considered more user-friendly, especially if designed discreetly and resembling more familiar devices. Five of the 6 participants believed their health care teams should have access to the data, particularly relating to their symptoms, fluctuations, and medication. Three also wished to access these data for self-management. One participant was hesitant regarding the potential benefits of technology to support PD management, preferring to use feedback data personally and relying on their health care team's usual assessment to guide decisions. Across the group, desirable device technical features included symptom prediction, reminder prompts, and support for medication

XSL•FO
RenderX

management. Despite concerns about stigma, most participants were willing to wear PD-specific devices, believing that they could aid better symptom management.

Conclusions: In summary, wearable devices must be discreet, robust, comfortable, and easily applied to promote adherence regardless of technical specifications.

(JMIR Form Res 2025;9:e73423) doi:10.2196/73423

KEYWORDS

Parkinson disease; wearable device; technology; experience; stigma

Introduction

Parkinson disease (PD) is the second most common chronic neurodegenerative disorder, with an increased incidence with aging [1]. PD is characterized by impaired motor function (ie, tremor, bradykinesia, and rigidity) [2], a wide range of nonmotor symptoms (eg, sleep disturbances and fatigue) [3], and by orthostatic hypotension, reflecting autonomic dysfunction and medication side effects [4].

Improving assessment is crucial for monitoring disease progression, adjusting medication doses, enhancing quality of life, and improving access to treatment services for people with PD [5]. Due to the heterogeneity and complexity of PD features, clinical assessment may be challenging and not always consistent as it relies on subjective clinician assessment (eg, using the Movement Disorder Society - Unified Parkinson's Disease Rating Scale [MDS-UPDRS]) [6] and self-evaluation by patients [7]. Intermittent clinical examinations offer limited understanding of daily symptom fluctuations [8], and formal assessment of PD can often be time-consuming and susceptible to interrater variability [9,10]. Clinical assessments may not always capture patients in their typical condition [11], while self-report can be burdensome and relies on individuals' recollections, where time elapsed between appointments can make it difficult to recall symptoms [12]. Accessing services also remains challenging due to long waiting lists, infrequent routine clinic visits, and travel for appointments, which can be fatiguing and costly [13].

Wearable devices for daily and remote monitoring of people with PD present opportunities, especially in a health service model that values person-centered and community-based approaches [14]. They can collect continuous, accurate, and objective data, and have been used to detect tremor [15], mobility [16], bradykinesia [17], dyskinesia [18], motor fluctuations [19], gait speed [20], sleep [21], falls [22], and physical activity [23].

Current wearable devices offer PD symptom detection and physiological monitoring (ie, blood pressure [BP]); however, few certified solutions exist for at-home motor symptom monitoring in people with PD. These include Parkinson's Kinetigraph [24], STAT-ON [25], Kinesia 360 [26], and the Apple watch with StrivePD [27] mobile app. MoveMonitor [28], Actiwatch [29], and ActiGraph [30] products, GT9X Link, and wGT3X-BT are devices that have also been used in validation studies for estimating physical activity, gait parameters, and sleep patterns in people with PD [31-33]. Given the prevalence of orthostatic hypotension, BP monitoring is another important aspect of PD management. The Samsung SM-R850 has been evaluated for BP measurements in people with PD and fulfilled the BP validation criterion of the International Organization for Standardization [34]. To the authors' knowledge, there are no off-the-shelf systems or studies that incorporate motor symptom and gait speed monitoring with photoplethysmography (PPG) and electrocardiogram (ECG) sensors for simultaneous BP calculation. To address this gap, a prototype wearable device system called the "Wearable Enabled Symptom Assessment Algorithm" (WESAA) has been proposed to collect concurrent inertial, PPG, and ECG raw signal waveforms. In particular, differentiating off-symptoms from hypotension could support clinicians in decision-making to improve quality of life.

Technology research typically focuses on the device's technical performance, with less evaluation of the user experience [35], which may result in slow adoption of wearable devices into routine clinical care pathways [36]. Mostly, the user experience is based on surveys [37-39], focuses on the clinician perspective [40], or includes mixed patient populations [41]. Rarely do people with PD report their preferences and expectations [42], despite a need to enhance the patient's engagement in their own care [43]. Wearable device designers who adopt a person- and user-centered design philosophy that takes into consideration users' involvement in the early stage of design are more likely to enable ongoing user engagement via a device that is suited to the needs of people with PD [35]. Therefore, this study aimed to apply a novel approach by including people with PD in the design process, concurrent with device development, to answer the question: "what are the user experiences of the novel prototype system?" Specific objectives were to:

- assess the usability and acceptability of the WESAA prototype system among people with PD and the practical implications of using it in daily life.
- explore user-interface preferences among people with PD using the WESAA system.
- examine broader attitudes toward wearable technology for PD management and desirable features for future wearable devices.

Methods

User-Centered Design

The authors engaged with different stakeholders, including clinicians, people with PD, caregivers, human factors engineering practitioners, and medical device designers to collect perspectives and feedback to ensure a comprehensive perspective and its significance to people with PD. To ensure

XSL•FO
RenderX

that the study was grounded in the perspectives of people with PD, there was active involvement of 3 members of the Parkinson Ireland branch, comprising 2 people with PD and a caregiver of a person with PD. This collaboration included: identifying and prioritizing the research questions; reviewing the data collection tools to ensure that these were relevant, sensitive, and user-friendly; determining the duration for which participants wore the device, taking into account factors such as comfort and feasibility; corecruiting participants for the study; and informing effective ways to disseminate the results to the Parkinson community. These collaborators did not participate in the study as participants.

System requirements for the prototype device were developed based on this guidance together with a comprehensive study of the current state-of-the-art and device market. The industrial design of the system was conducted by an external collaborator, Design Partners, Ireland [44], under the direction of the authors. After concept validation, the approach moved to the prototype and testing phase, where a prototype was created to test and refine the proposed solution, and then evaluated with the users in the test phase in order to collect feedback and identify opportunities for improvement.

Included Devices: The WESAA Prototype System

In the WESAA system, 2 identical units are worn, on the external side of the wrist and ankle of the most affected side (Figure 1). Additional information regarding the design of the WESAA system is available in the publication titled “Design of a Multi-Sensors Wearable System for Continuous Home Monitoring of Individuals with Parkinson” [45]. Each device embeds an inertial monitoring unit to record accelerations and angular velocities as well as PPG and ECG sensors. Each enclosure is 45.3 mm × 76.8 mm × 14 mm, latex-free, weighs approximately 41 g, and is fastened to provide secure contact to reduce artefact. Tri-glide slides, used as buckles, use friction to secure the enclosure in place. The strap is 30 mm wide, latex-free, and weighs approximately 22 g (Figure 2).

XSL•FO
RenderX

Figure 1. Wearable Enabled Symptom Assessment Algorithm system - wrist and ankle. A usability study of limb-worn wearable devices for monitoring motor function and blood pressure in people with Parkinson disease in Ireland in the period 2021-2023. Photograph of a participant's lower body with the Wearable Enabled Symptom Assessment Algorithm wearable device system attached to the left wrist and left ankle.

XSL•FO
RenderX

Figure 2. The Wearable Enabled Symptom Assessment Algorithm unit was used in the usability study of limb-worn wearable devices for monitoring motor function and blood pressure in people with Parkinson disease in Ireland in the period 2021-2023. Photograph showing the system from the front and underside. The front view features an orange stretchable band with a tri-glide slide, a white rectangular box containing a semiannular gold ring and a photoplethysmography lens. The underside view reveals the same orange band and white box, now displaying a gold circular electrode. PPG: photoplethysmography.

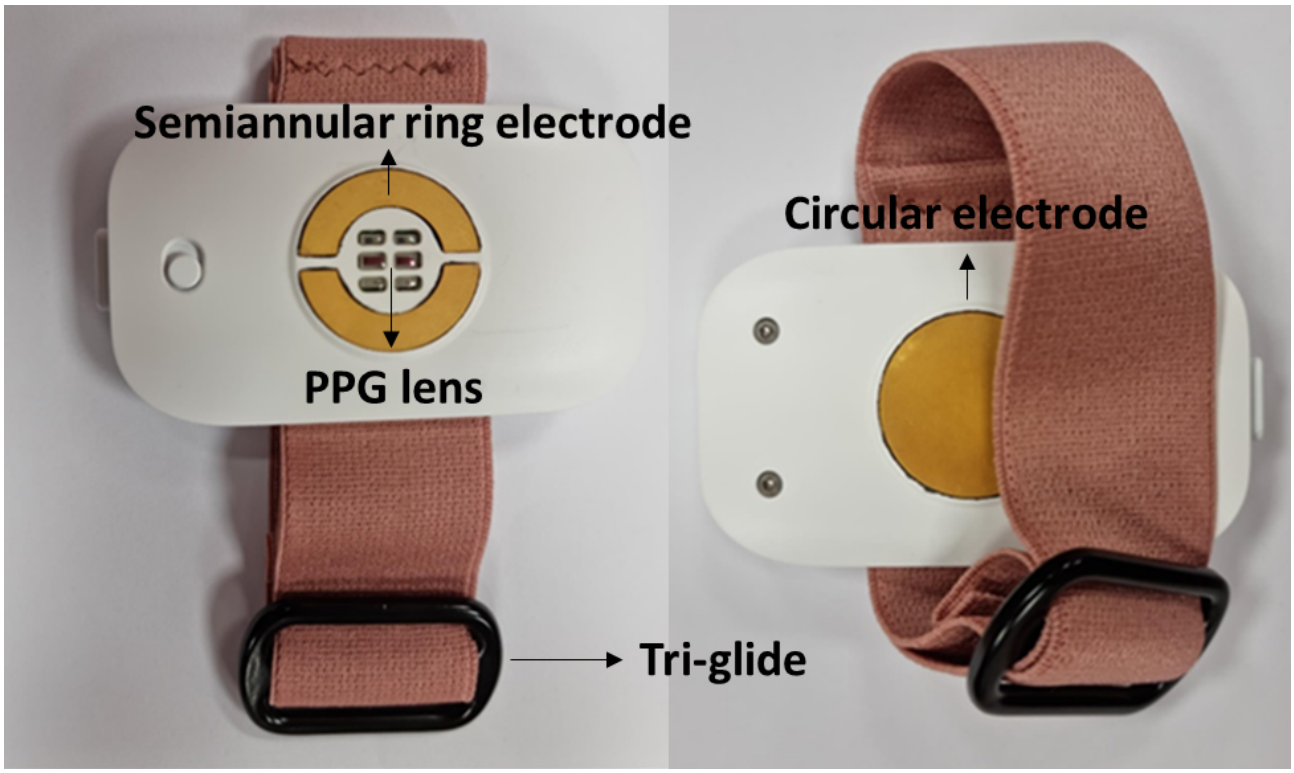

Participants wore the system at home, allowing a comprehensive approach where cultural and social activities are considered. Devices were removed during water activities, following guidance on how to remove and put them on. The WESAA system collects, contextualizes, and stores up to 10 days of data from the wrist and ankle using a low-power microcontroller unit within the device. These data, including raw inertial, PPG, and ECG signals, are transferred to a personal computer via a USB cable, where they can be further analyzed, processed, and visualized using sophisticated machine learning and data analytics methods. After a 24-hour period, researchers met with participants to collect the system. The decision to limit the wearing period was influenced by the prototype nature and input from the Parkinson advisors, who suggested that a 24-hour period would be sufficient to gather insights into the devices without being burdensome.

Recruitment Setting and Participants

Participants were recruited through local branches of Parkinson Ireland, a not-for-profit advocacy and support organization, and through a PD specialist clinic. The researcher (LK, female research assistant trained in qualitative methods, with a social science background) contacted the groups to share study information as appropriate (in-person, word of mouth, and posters in the clinic). Recruitment through Parkinson Ireland involved the circulation of the study details through online support group meetings (COVID-19 era). The researcher attended these meetings and gave an overview of what the study entailed. Recruitment through the PD specialist clinic involved the researcher attending in-person at a Health Service Executive clinical site where patients visit a consultant geriatrician with a special interest in PD. The consultant screened for suitability and initial interest and facilitated introductions between the researcher and patient, where once again an outline of the study was given. Those interested were invited to contact the researcher where the study was further discussed in detail, including details of the nature and duration of the study, and the optional poststudy internet-based interview, the potential risks and benefits involved, and their right to choose not to participate. LK had no prior relationship with study participants, who understood her as the study's research assistant. Participants satisfying the following criteria were eligible: diagnosis of PD, aged 50 years and older, ambulatory, and capable of fulfilling the study requirements. After wearing the WESAA prototype, a purposive sample of 6 participants who wore the device at home were asked to consider participating in an optional poststudy internet-based interview after they handed back the devices. They took the participant information leaflet and consent form for the interview home and had the period of wearing the device to decide if they wished to take part in the interview. The researcher contacted them the following morning while they were wearing the device to ensure all was well, and this provided an opportunity for further questions, and to arrange the internet-based interview from their home at a time of their convenience, if interested. The signed consent form was brought with them on the day they handed back the devices. Participants were selected using a criterion theoretical sampling strategy to satisfy the following criteria: inclusion of different age categories (50-60 years, 61-70 years, 71-80 years, and >80 years) and the inclusion of men and women. Participants were

XSL•FO
RenderX

excluded if they had significant communication deficits. Interviews took place as soon as possible, within a week of wearing the device.

Qualitative Research Methods

Due to the heterogeneity of PD, it was deemed more suitable to use qualitative research methods for this population, to potentially discover usability barriers that are overlooked by quantitative methods [46]. Internet-based semistructured interviews were chosen as the data collection method to reduce travel burden and costs and offer scheduling flexibility. Microsoft Teams hosted by the university was used, where only individuals with authorized access participated, and a "waiting room" was used, with a controlled experience (eg, attendees cannot change meeting settings or share content). A separate audio recording device was used, where audio files were password-protected in a secure database.

An interview topic guide was designed based on existing literature and the purpose of the study, reviewed by the stakeholders and researchers together in terms of the content, focus, and significance to people with PD. Questions explored the experience of using the wearables, perspectives on the sharing of health-related data, familiarity with wearable devices and technology for PD, perception of wearable devices for PD, and future capabilities of wearable devices for PD (Multimedia Appendix 1). Pilot testing of the interview topic guide was conducted to validate its effectiveness, and subsequently, interviews were conducted using the topic guide to stimulate general discussion. To support the interdisciplinary nature of the research, the interviews involved participation from a member of the research team (LK) and a member of the industrial design collaborators, Design Partners. Interviews lasted between 25 and 35 minutes. All interviews were audio-recorded, transcribed verbatim, and field notes were taken

Data Analysis

Transcripts were analyzed using NVivo (QSR International Pty Ltd) for common themes using a qualitative thematic approach [47]. The researcher (LK) initially used an inductive coding approach within an interpretivist paradigm, where data-derived codes emerged as the researcher navigated through the data. A second researcher (LOM, female researcher with a background in public health) reviewed a sample of transcripts to enhance intercoder reliability and ensure consistency in theme identification. Codes were grouped into provisional subthemes and themes, with key phrases assigned based on content relevance. The alignment of codes with research questions facilitated deductive analysis to ensure comprehensive coverage of study objectives.

Reflexivity was maintained through ongoing reflection on the researcher's assumptions, perceptions, and interactions with participants. Member checking and peer debriefing enhanced the credibility of the findings. After 4 interviews, a recurring pattern in the responses was observed which indicated that data saturation may have been reached. To ensure rigor, data collection continued until a total of 6 interviews were completed, at which point no new themes or perspectives emerged.

Ethical Considerations

This study was approved by the Clinical Research Ethics Committee at University College, Cork (ECM 4 (r) 11/10/2020) and consent was obtained prior to participant recruitment. The informed consent process provided information about rights as research participants, study objectives, potential risks and benefits, and confidentiality measures. Written informed consent was obtained from all participants. Participation was voluntary and no compensation was given. Any personal information was anonymized before analysis. While no identifiable facial features are presented, documented written consent was obtained from the participant depicted in the photograph in the publication.

Results

Participants

As per Table 1, people with PD, aged 52-83 years, with an average disease duration of 8 years, with mild-to-moderate PD (Hoehn and Yahr scale score ≤3; median MDS-UPDRS score of 72/199) and cognition within normal limits (6-item Cognitive Impairment Test [6-CIT] median score 0), took part in the internet-based interviews, which lasted between 26 and 50 minutes. No invited participant declined.

XSL•FO
RenderX

Table 1. Clinical and demographic data of the study participants enrolled in the usability study of limb-worn wearable devices for monitoring PD motor function and blood pressure in Ireland in the period 2021-2023 (N=6).

Clinical and demographic data	
Participants, n (%)	6 (100)
Sex, n (%)	
Male	3 (50)
Female	3 (50)
Age (years), median (IQR; range)	66 (12.5; 52-83)
BMI, median (IQR; range)	24.8 (3.14; 23.94-33.79)
Hoehn and Yahr stage, median (IQR; range)	2 (0; 2-3)
Years since diagnosis, median (IQR; range)	7 (4.5; 3-17)
MDS-UPDRS[a] (total score), median (IQR; range)	72 (18.5; 38-78)
MDS-UPDRS III (motor examination score), median (IQR; range)	36.5 (23.5; 7-45)
6-CIT[b], median (IQR; range)	0 (0; 0-4)

[a]MDS-UPDRS: Movement Disorder Society - Unified Parkinson's Disease Rating Scale.

[b]6-CIT: 6-Item Cognitive Impairment Test.

Five themes were identified in the analysis: (1) Life with Parkinson Disease, (2) Usability and Acceptability Factors of the WESAA System, (3) User interface and Data Access, (4) Attitudes toward Wearable Technology for PD, and (5) Additional Desirable Features of a Wearable Device. These are summarized in Table 2. To ensure anonymity while reporting, participants are labeled as P1, P2, P3, P4, P5, and P6.

Table 2. Summary of Thematic Topic Areas. A usability study of limb-worn wearable devices for monitoring motor function and blood pressure in people with PD in Ireland in the period 2021-2023.

Theme	Subtheme
Life with Parkinson Disease	• PD[a] is an impactful condition, with symptom variability and unpredictability, requiring adjustments of routine and pharmaceutical reliance. • Management techniques include adjusting medication and physical activity. • A team-based approach is essential. • Social challenges such as social isolation and stigma contribute to loneliness and employment difficulties.
Usability and Acceptability Factors of the WESAA[b] System	• Participants had a positive attitude toward the testing. • The device was comfortable and noninvasive. • Usability challenges included device size (with clothing), strap issues, and water resistance; the system was not discreet.
User Interface and Data Access	• Desire for feedback on system functioning and data upload status. • Suggestions for a user-friendly interface to enhance usability and provide real-time feedback. • Preferences varied regarding sharing data with health care teams, highlighting individual preferences; emphasis on health care professionals (not patients) making clinical decisions.
Attitudes toward Wearable Technology for PD	• Positive attitude toward wearable technology in general. • Perceived benefits such as accurate symptom recording and improved communication with health care teams. • Wearables were seen as a source of motivation for physical activity and a tool to enhance self-monitoring.
Additional Desirable Features of a Wearable Device	• Wearable technology should be personalized to individual needs. • Recommendations include personalized tracking of health metrics such as blood pressure, sleep habits, and medication levels. • Value in reminder prompts for medication, exercise, and symptom management. • Continuous monitoring is recommended for fine-tuning medication. • Suggestions to include additional features such as gait and balance tests, step counters, voice therapy applications, and cognitive training tools.

[a]PD: Parkinson disease.

[b]WESAA: Wearable Enabled Symptom Assessment Algorithm.

XSL•FO
RenderX

Topic Area 1: Life With Parkinson Disease

The first theme, "Life with Parkinson Disease," provides contextual background information that enriches the understanding of participants' experiences wearing the devices. This theme offers insights into the daily challenges, routines, and activities people with PD navigate, which can impact their perceptions of and interactions with wearable devices. Understanding the broader context of participants' lives is crucial for interpreting their responses and identifying potential barriers and facilitators to device use in real-world settings.

Participants expressed the profound impact of a PD diagnosis, requiring ongoing treatment and adjustments to daily routines. Common challenges included symptom fluctuations and unpredictability, with changes in both motor and nonmotor symptoms throughout the day, which might impact the ability to use a device.

> *You might get out of bed in the morning and say forget it, I'm not able to do what I want to do today. And then you might get out of bed in the morning and you might be absolutely fine and get much more done than you think; so it's you take each day as it comes.* [P4]

Many discussed strategies of adjusting medication timing and dosage, trialing new medications or therapies, and exercise and physical therapy in attempts to improve functioning. A device needs to support and not impede these goals.

Some felt that PD can be:

> *…very isolating when you're by yourself.* [P3]

Tremors, stiffness, and difficulty with movement make it challenging to perform everyday tasks, contributing to feelings of loneliness, isolation, and social stigma. PD symptoms together with social discomfort can create barriers to social interaction, leading to increased feelings of isolation in people's work and social lives:

> *The concern would be that people know I have Parkinson's, then they might not want me to do their work, because they might trust me to do it, even though I'm perfectly capable of doing it.* [P2]

Thus, a device needs to be discreet.

Topic Area 2: Usability and Acceptability Factors of the WESAA System

All participants expressed a positive attitude toward testing the prototype system for research purposes, eagerly volunteering to take part in the study:

> *I'm optimistic that this advancement today is going to bear fruit in time.* [P1]

While wearing the system, participants engaged in varied activities—some active, such as cycling and socializing, and others at home reading or watching TV.

The system was considered acceptable, generally comfortable, light in weight, and noninvasive. While generally satisfactory, significant practical implications were identified in terms of usability. The size of the current devices was found to be "bulky" especially in relation to their outward projection:

> *…the only thing that would strike me about it…is certainly is big. If you could reduce that in size again, I think it would help.* [P5]

Some felt that sleep was disrupted; one person experienced intermittent sleep with 1 or 2 awakenings, while another was aware of the devices initially but slept through without interruptions.

Devices were fastened using a latex-free stretchy band, and participants had different experiences with adjusting, putting on, and taking off. One person had no issues, but others reported minor problems related to strap adjustability and challenges specific to PD. In some instances, the strap became loose over time and it was challenging to keep the devices in place, affecting device stability. In contrast, some experienced a tightening of the strap, which made it difficult to reduce the tension of the strap when adjusting or removing.

The plastic adjuster slider used to adjust the tightness of the strap posed a challenge for some due to PD-related barriers such as tremors, slowness, and fluctuations, where

> *manipulating the strap... it was a little bit difficult.* [P6]

Future designs should incorporate a strap that is secure, less inclined to loosen, easier to adjust, and manipulate. Suggestions included features familiar to people with PD, such as Velcro or a watch-style strap as these could be used with one hand, without the assistance of others.

The prototype was not completely sealed and was not water-resistant, which impacted robustness:

> *I think the bottom line is that the devices can take a lot of punishment, but they can't take too much.* [P1]

Some found it challenging to be careful with movement for fear of damaging which in turn would affect the technical performance, "I was conscious all the time that I had it" (P3); thus, some participants altered their usual behaviors while wearing.

While generally comfortable, the challenge of making it work with clothing and adjusting it for a proper fit outweighed the comfort factor. The devices' size and outward projection meant, "*you'd have real problems getting a sleeve on over the [device]... or a tight pair of jeans if you wear on the leg*" (P4).

Moreover, devices were found to be nondiscreet, with the potential to draw attention to the wearer where "it was very obvious, because it is quite clunky and unusual looking" (P2).

There was a perceived social and emotional impact from this as not all people with PD openly communicated their condition, and there were concerns that a device could disclose a person's health status:

> *I'd be very conscious of having it. Particularly the arm one visible, because people are going to say, 'Oh, that's a strange looking watch,' and then you're trying to explain.* [P4]

Some had a negative perception of the ankle-worn device which was regarded as unfamiliar because "very few people wear something on their leg" (P6).

XSL•FO
RenderX

One person dressed in such a way as to conceal and another likened it to a prisoner tracking system saying:

> *I was afraid I'd be labelled some kind a sex offender or some kind of prisoner.* [P2]

The wrist-worn device was considered more user-friendly for long-term use, especially if it was made smaller, more discreet, and resembled more universally familiar devices.

Topic Area 3: User Interface and Data Access

There was a desire to receive information on the system's functioning status, namely whether it was recording or positioned correctly as participants never got feedback as to whether data successfully uploaded or not, and so a system that needed special attention had the potential to be troublesome:

> *There was a slight bit of anxiety on my part as to whether I'd screwed up the whole thing by nudging it off something.* [P2]

In addition, some participants were interested in accessing the health-related data collected, suggesting the inclusion of a user interface would enhance satisfaction.

Most believed that health care teams should have access to the collected data, but 1 participant expressed hesitation regarding the potential benefits of technology in managing their PD, preferring to keep information solely to themselves, believing that technology and health care should be kept as "two separate systems" and they "would go along with the (health) team as they are telling them" (P3).

Three participants wished for both themselves and their health care teams to receive the data, and were interested in sharing data relating to their symptoms, fluctuations, and medication as "they're capable bits of equipment that could collect information for my doctor' (P1).

Providing clinicians with access to data would aid in gaining a better understanding of their condition, with different methods suggested for achieving this goal. One saw information being given first to health care teams and then fed back to the people with PD:

> *…because then the consultant can say, well, we can see this is happening or that's happening. And then you understand where you're at more yourself.* [P4]

Another felt that health care teams would not always have the capacity to review data daily or weekly and therefore would like access themselves first, and then health care teams thereafter could view at intervals. Although having access to their own data was important, individuals still wanted input from their health care teams, as "there's only so much I can do with the information, but they could analyse it better" (P6).

Technology was seen as an additional tool for receiving adequate health care, but it should not replace the essential role of health care teams.

Two participants preferred to have their information relayed solely to their health care team. They explained that they did not want to be constantly reminded of their PD:

> *I'm a bit like the ostrich. I have this illness, I stick my head in the sand and I change the philosophy. Unless I need to know, I don't want to know.* [P5]

Topic 4: Attitudes Toward Wearable Technology for PD

Participants recognized the integration of wearable devices in the management of PD, as suitable technology could benefit their daily lives. Objective recording of disease symptoms was cited as one of the benefits, and predicting symptoms, particularly real-time ON and OFF states, could enhance medication management. Managing the condition can become overwhelming and that:

> *…it would be lovely to have something recording and because we've so much to think of, exercises, tablets, and keeping up your spirits. Each and every day, it would be great to have something working in the background to record what's actually happening.* [P6]

The system encouraged some to be more active where 1 participant mentioned that:

> *I suppose it could keep you active and keep you informed of what you're doing, and what improvements are there.* [P3]

As time with health care professionals was perceived to be limited, wearable devices were seen as a solution to reinforce communication with data and provide confidence when attending appointments. Recorded health parameters could be communicated to health care teams allowing them "to be more precise in their assessment of what is wrong and what can be done to lessen the impact of the illness" (P5), reducing the burden of people with PD self-recall. If health care teams had better access to the health data they would have a clearer understanding of the current and future projected state of people with PD:

> *…so I would be hopeful that it would bridge the unknown thing, where I'm at with my condition and where the clinician maybe doesn't quite have the full picture.* [P1]

There was potential to help regain confidence by providing clarity about their condition, which they can then act upon, as reflected here:

> *…but then you have confirmation that it is, you actually will get it treated.* [P6]

Although aspects of the current prototype may require refinement, participants were still willing to wear it if it meant that their health condition could be improved through its use:

> *I would be happy for the nuisance of it if it gave the consultant that type of information to help them determine how best you can treat my illness.* [P5]

Topic 5: Additional Desirable Features of a Wearable Device

Recommendations were made that wearable devices might need to be customized to the particular preferences of people with PD. A one-size-fits-all approach might not work due to the condition's diversity since there might be important variations

XSL•FO
RenderX

between those with early-stage disease and those with more advanced stages. An individualized system that can monitor specific health parameters, provide personalized feedback, and assist with the management of medical conditions would be highly valuable in promoting patient engagement and satisfaction.

Features to incorporate into future wearable devices included enhanced monitoring and therapy provision. One monitoring suggestion was the BP monitoring feature of the WESAA system. Reminder prompts could be advantageous, not only for medication intake, exercise, and movement but also for anticipating symptom fluctuations. The potential usefulness of the ability to fine-tune medication, for example, the possibility of including a dopamine level monitor, comparable to the continuous glucose monitoring used in diabetes, was proposed:

> *Is there some possibility that a device like that would help monitor your levels of dopamine?* [P2]

Monitoring sleep was another potentially useful feature, as sleep disturbances can be common in people with PD. Incorporating a gait and balance test application was recommended, especially for those prone to falling or who have balance issues. A step counter was proposed for those interested in physical activity levels, as it was felt that regular physical activity can help improve balance, mobility, mood, and overall physical well-being in PD. People with PD may experience challenges with speech, such as reduced volume and impaired articulation, and integrating voice therapy and brain training applications could provide beneficial therapy, offering a platform for voice maintenance and cognitive development.

Discussion

Principal Findings

The success of wearable devices is not only determined by their technological capabilities but also by their usability and wearability components. We examined the experience of people with PD in using the newly developed prototype system known as the WESAA. All participants reported a positive perception of the system, aligning with previous research where participants showed willingness and positivity to wearing complex wearable systems [48], and that any minor inconvenience in wearing the device was outweighed by potential value [41].

While the system was comfortable to wear, the size and outward projection of the individual devices were too large. Consistent with previous research [35,48], a device that was small, easy to fasten, noninvasive, waterproof, and durable was preferred.

Device fastenings used a latex-free stretchy band, and in some instances, the strap became loose and affected device stability. Similar to AlMahadin et al [12], Velcro straps were the most desired fastening mechanism. Ensuring a device does not have a negative impact on the user's daily routines is crucial [12]. When a user's routine is disturbed, particularly among older users, the device faces challenges in seamlessly becoming an essential part of the user's life and potentially resulting in a loss of its perceived value [49]. Incorporating wearable devices into the existing behavior of people with PD should be seamless, without requiring the learning of new practices or routines.

There is a recognized stigma associated with PD [50], highlighting the importance of discretion in future design. Our study revealed concerns about the potential to draw attention to the wearer, leading to feelings of embarrassment and self-consciousness, particularly if it revealed their condition. This could affect willingness to continue monitoring beyond the initial 24-hour period. Prioritizing ease of use and aesthetic appeal is important, as the latter appears to be closely linked to stigma concerns. Interestingly, previous research by our team [43], where people with PD considered the hypothetical wearing of various wearable devices, found that while the size and weight of the devices were crucial, aesthetics were deemed less significant (ie, "function over form"). Moreover, AlMahadin et al's study [12] showed that participants had no concerns about device visibility and actually preferred to be identified as a person with PD, signaling that they may require assistance. Our research reveals a contrasting preference for devices that were discreet, which perhaps reflects the reality of wearing a system for 24 hours in daily life. Previous research [37,41] demonstrated that people with PD can be uncomfortable with an ankle-worn device, a sentiment expressed by our participants, considering it unfamiliar, and even comparing it to a prisoner tracking system. Thus, the incorporation of an ankle device into a PD monitoring system would need consideration by designers if they want to ensure acceptance and compliance.

Responses revealed a consensus on the benefits of sharing health-related data with health care teams; however, a subset of individuals also expressed a desire for personal ownership of their information and the ability to access it themselves, while others preferred not to be constantly reminded of their PD condition. A combination of health care and self-care in managing any chronic condition has been highlighted [51], and for wearable systems in PD to achieve their maximum effectiveness, people with PD who wish should have regular access to their personalized health-related data and ideally, this access should be in real time. Hence, designers face a challenge in developing systems that can meet the diverse preferences regarding access to health care information on a wearable device. An additional task is to meet the stated diverse needs for additional features in a wearable system, extending from enhanced monitoring of several parameters to therapeutic applications, all in a smaller device.

Reliable, continuous, and objective recording of individual health parameters that can be shared with health care teams has been recognized by people with PD [12]. The majority expressed support for the exchange of information between people with PD and their health care team, via accurate and objective recording of disease symptoms. It was felt there was limited access to clinical consultations for people with PD; however, WESAA offers an opportunity to attend appointments confidently by reinforcing communication with reliable objective data. These data can be accessed by clinicians to support people with PD to accurately recall symptoms, improving the quality of communication during consultations and reducing the burden of self-recall.

A wearable system tailored to people with PD carries the hope of empowering patients by minimizing the disruption caused by "health care inconveniences" in their day-to-day lives and

XSL•FO
RenderX

potentially reducing the need to spend time recording symptoms and travel for appointments [52]. Furthermore, gathering data continuously offers a more comprehensive dataset compared to the snapshot readings obtained during visits to health care facilities [53]. Those with greater levels of health information are equipped to contribute more valuable inputs concerning their condition, therefore encouraging collaborative decision-making and facilitating adherence to agreed-upon plans [54].

Strength and Limitations

The research explored the practical application of wearable devices within a domestic 24-hour setting, as opposed to hypothetical perspectives or limited laboratory or clinical testing. It is important to acknowledge the limitations imposed by the small sample size and the recruitment from a PD support group, which may not represent the broader population of people with PD. While data saturation was reached, PD is a heterogeneous disease, and it is possible that these findings may have been subject to refinement if more perspectives were captured across different disease stages and phenotypes, and across more diverse age and cultural profiles. This study was in its infancy and research questions were still emerging. Future research should aim for a larger and more diverse sample through broader sampling, where new insights might emerge.

Acknowledging the limitations of the 24-hour wearing period is practical, largely in terms of capturing long-term or nuanced user experiences. As devices evolve, future studies may consider longitudinal study designs to gain deeper insights into users' experiences over time. Nonetheless, the chosen testing period was deemed appropriate for the goals of the study and the stage of prototype development.

As the interviews conducted were internet-based interviews—where there was limited visual interaction and nonverbal communication may have been missed. Participants had no technological engagement with the system; therefore, any bias regarding specific technical features was unlikely. Going forward, it might be useful to involve participants with varying levels of familiarity with technology to explore potential differences between those who are tech-savvy versus those less familiar with technology.

While no formal clinical safety testing (eg, dermatological assessment for skin irritation or pressure marks) was conducted, no user reported itch, redness, rash or other irritation. To ensure real-world adherence and safety, all future design developments should also include strap safety testing.

Conclusion

The study involved conducting research and progressively testing designs with people with PD. A deeper understanding of the requirements of people with PD, the problem domain, and the technological options available was obtained. This helps facilitate the future development of more sophisticated and targeted prototypes for people with PD. The study demonstrates that designers must take into consideration concerns of size, discretion, and strap design and be willing to include additional user interface elements in accordance with the user's preferences.

Acknowledgments

The authors would like to thank Design Partners for the support provided in the development of the industrial design concept and their involvement in the interviews, and all the participants who voluntarily took part in the study. This work was supported in part by Enterprise Ireland (EI) and AbbVie, Inc. under agreement IP 2017 0625.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to information that could compromise the privacy of research participants, but are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview guide.
[DOCX File , 15 KB - formative_v9i1e73423_app1.docx]

References

1. Maserejian N, Vinikoor-Imler L, Dilley A. Abstracts of the MDS Virtual Congress 2020. Mov Disord 2020 Sep;35(S1):S1-S702. [doi: 10.1002/mds.28268] [Medline: 32914891]
2. Jankovic J. Parkinson's disease: clinical features and diagnosis. J Neurol Neurosurg Psychiatry 2008;79(4):368-376. [doi: 10.1136/jnnp.2007.131045] [Medline: 18344392]
3. Postuma RB, Berg D, Stern M, et al. MDS clinical diagnostic criteria for Parkinson's disease. Mov Disord 2015;30(12):1591-1601 [FREE Full text] [doi: 10.1002/mds.26424] [Medline: 26474316]
4. Fanciulli A, Leys F, Falup-Pecurariu C, Thijs R, Wenning GK. Management of orthostatic hypotension in Parkinson's disease. J Parkinsons Dis 2020;10(s1):S57-S64 [FREE Full text] [doi: 10.3233/JPD-202036] [Medline: 32716319]

XSL•FO
RenderX

5. Dorsey ER, Bloem BR. The Parkinson pandemic-a call to action. JAMA Neurol 2018;75(1):9-10. [doi: 10.1001/jamaneurol.2017.3299] [Medline: 29131880]
6. Goetz CG, Tilley BC, Shaftman SR, et al, Movement Disorder Society UPDRS Revision Task Force. Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results. Mov Disord 2008;23(15):2129-2170. [doi: 10.1002/mds.22340] [Medline: 19025984]
7. Post B, Merkus MP, de Bie RM, de Haan RJ, Speelman JD. Unified Parkinson's disease rating scale motor examination: are ratings of nurses, residents in neurology, and movement disorders specialists interchangeable? Mov Disord 2005;20(12):1577-1584. [doi: 10.1002/mds.20640] [Medline: 16116612]
8. Kanellos FS, Tsamis KI, Rigas G, et al. Clinical evaluation in Parkinson's disease: is the golden standard shiny enough? Sensors (Basel) 2023;23(8):3807 [FREE Full text] [doi: 10.3390/s23083807] [Medline: 37112154]
9. Williams S, Wong D, Alty JE, Relton SD. Parkinsonian hand or clinician's eye? Finger tap bradykinesia interrater reliability for 21 movement disorder experts. J Parkinsons Dis 2023;13(4):525-536 [FREE Full text] [doi: 10.3233/JPD-223256] [Medline: 37092233]
10. Kenny L, Azizi Z, Moore K, et al. Inter-rater reliability of hand motor function assessment in Parkinson's disease: impact of clinician training. Clin Park Relat Disord 2024;11:100278 [FREE Full text] [doi: 10.1016/j.prdoa.2024.100278] [Medline: 39552791]
11. Perlmutter JS. Assessment of Parkinson disease manifestations. Curr Protoc Neurosci 2009;Chapter 10:Unit10.1 [FREE Full text] [doi: 10.1002/0471142301.ns1001s49] [Medline: 19802812]
12. AlMahadin G, Lotfi A, Zysk E, Siena FL, Carthy MM, Breedon P. Parkinson's disease: current assessment methods and wearable devices for evaluation of movement disorder motor symptoms - a patient and healthcare professional perspective. BMC Neurol 2020;20(1):419 [FREE Full text] [doi: 10.1186/s12883-020-01996-7] [Medline: 33208135]
13. O' Shea E, Rukundo A, Foley G, Wilkinson T, Timmons S. Experiences of health service access: a qualitative interview study of people living with Parkinson's disease in Ireland. Health Expect 2024;27(1):e13901 [FREE Full text] [doi: 10.1111/hex.13901] [Medline: 37926923]
14. Bloem BR, Henderson EJ, Dorsey ER, et al. Integrated and patient-centred management of Parkinson's disease: a network model for reshaping chronic neurological care. Lancet Neurol 2020;19(7):623-634 [FREE Full text] [doi: 10.1016/S1474-4422(20)30064-8] [Medline: 32464101]
15. López-Blanco R, Velasco MA, Méndez-Guerrero A, et al. Smartwatch for the analysis of rest tremor in patients with Parkinson's disease. J Neurol Sci 2019;401:37-42. [doi: 10.1016/j.jns.2019.04.011] [Medline: 31005763]
16. Del Din S, Godfrey A, Rochester L. Validation of an accelerometer to quantify a comprehensive battery of gait characteristics in healthy older adults and Parkinson's disease: toward clinical and at home use. IEEE J Biomed Health Inform 2016;20(3):838-847. [doi: 10.1109/JBHI.2015.2419317] [Medline: 25850097]
17. Memar S, Delrobaei M, Pieterman M, McIsaac K, Jog M. Quantification of whole-body bradykinesia in Parkinson's disease participants using multiple inertial sensors. J Neurol Sci 2018;387:157-165. [doi: 10.1016/j.jns.2018.02.001] [Medline: 29571855]
18. Ossig C, Antonini A, Buhmann C, et al. Wearable sensor-based objective assessment of motor symptoms in Parkinson's disease. J Neural Transm (Vienna) 2016;123(1):57-64. [doi: 10.1007/s00702-015-1439-8] [Medline: 26253901]
19. Rodríguez-Molinero A, Samà A, Pérez-Martínez DA, et al. Validation of a portable device for mapping motor and gait disturbances in Parkinson's disease. JMIR Mhealth Uhealth 2015;3(1):e9 [FREE Full text] [doi: 10.2196/mhealth.3321] [Medline: 25648406]
20. Del Din S, Elshehabi M, Galna B, et al. Gait analysis with wearables predicts conversion to Parkinson disease. Ann Neurol 2019;86(3):357-367 [FREE Full text] [doi: 10.1002/ana.25548] [Medline: 31294853]
21. Tedesco S, Crowe C, Sica M, et al. Sleep analysis via wearable sensors in people with Parkinson’s disease. Gait Posture 2023;106:S203-S204. [doi: 10.1016/j.gaitpost.2023.07.244]
22. Ramdhani RA, Khojandi A, Shylo O, Kopell BH. Optimizing clinical assessments in Parkinson's disease through the use of wearable sensors and data driven modeling. Front Comput Neurosci 2018;12:72 [FREE Full text] [doi: 10.3389/fncom.2018.00072] [Medline: 30254580]
23. Wendel N, Macpherson C, Webber K, Hendron K, DeAngelis T, Colon-Semenza C, et al. Accuracy of activity trackers in Parkinson disease: should we prescribe them? Phys Ther 2018;98(8):705-714. [doi: 10.1093/ptj/pzy054] [Medline: 29718452]
24. PKG device. PKG Health. URL: https://pkghealth.com/ [accessed 2023-12-05]
25. Stat-On. Sense4Care. URL: https://www.statonholter.com/ [accessed 2023-12-05]
26. Kinesia 360. Great Lakes NeuroTechnologies. URL: https://www.glneurotech.com/ [accessed 2023-12-05]
27. Apple Watch. Apple. URL: https://www.apple.com/watch/ [accessed 2023-12-05]
28. MoveMonitor | Ambulatory monitoring of Physical Activity for up to 14 days. McRoberts - Moving technology. URL: https://www.mcroberts.nl/products/movemonitor/ [accessed 2023-12-05]
29. Philips Respironics | Actigraphy monitoring devices. URL: http://www.actigraphy.respironics.com/solutions/actigraphy.aspx [accessed 2023-12-05]
30. Ametris (Formerly ActiGraph). URL: https://theactigraph.com/ [accessed 2023-12-05]

XSL•FO
RenderX

31. Prasad V, Brown C. A pilot study to determine the consistency of simultaneous sleep actigraphy measurements comparing all four limbs of patients with Parkinson disease. Geriatrics (Basel) 2017;3(1):1 [FREE Full text] [doi: 10.3390/geriatrics3010001] [Medline: 31011049]
32. Ginis P, Goris M, De Groef A, et al. Validation of commercial activity trackers in everyday life of people with Parkinson's disease. Sensors (Basel) 2023;23(8):4156 [FREE Full text] [doi: 10.3390/s23084156] [Medline: 37112496]
33. Cederberg KL, Jeng B, Sasaki JE, Lai B, Bamman M, Motl RW. Accuracy and precision of wrist-worn actigraphy for measuring steps taken during over-ground and treadmill walking in adults with Parkinson's disease. Parkinsonism Relat Disord 2021;88:102-107 [FREE Full text] [doi: 10.1016/j.parkreldis.2021.06.009] [Medline: 34171566]
34. Ahn JH, Song J, Choi I, Youn J, Cho JW. Validation of blood pressure measurement using a smartwatch in patients with Parkinson's disease. Front Neurol 2021;12:650929 [FREE Full text] [doi: 10.3389/fneur.2021.650929] [Medline: 34267718]
35. Shah SGS, Robinson I, AlShawi S. Developing medical device technologies from users' perspectives: a theoretical framework for involving users in the development process. Int J Technol Assess Health Care 2009;25(4):514-521. [doi: 10.1017/S0266462309990328] [Medline: 19845981]
36. Espay AJ, Hausdorff JM, Sánchez-Ferro Á, et al, Movement Disorder Society Task Force on Technology. A roadmap for implementation of patient-centered digital outcome measures in Parkinson's disease obtained using mobile health technologies. Mov Disord 2019;34(5):657-663 [FREE Full text] [doi: 10.1002/mds.27671] [Medline: 30901495]
37. Botros A, Schütz N, Camenzind M, et al. Long-term home-monitoring sensor technology in patients with Parkinson's disease-acceptance and adherence. Sensors (Basel) 2019;19(23):5169 [FREE Full text] [doi: 10.3390/s19235169] [Medline: 31779108]
38. Dominey T, Kehagia AA, Gorst T, et al. Introducing the Parkinson's kinetigraph into routine Parkinson's disease care: a 3-year single centre experience. J Parkinsons Dis 2020;10(4):1827-1832 [FREE Full text] [doi: 10.3233/JPD-202101] [Medline: 33016893]
39. Rodríguez-Martín D, Cabestany J, Pérez-López C, et al. A new paradigm in Parkinson's disease evaluation with wearable medical devices: a review of STAT-ON. Front Neurol 2022;13:912343 [FREE Full text] [doi: 10.3389/fneur.2022.912343] [Medline: 35720090]
40. Santos García D, López Ariztegui N, Cubo E, et al. Clinical utility of a personalized and long-term monitoring device for Parkinson's disease in a real clinical practice setting: an expert opinion survey on STAT-ON™. Neurologia 2023;38(5):326-333. [doi: 10.1016/j.nrl.2020.10.013]
41. Ozanne A, Johansson D, Hällgren Graneheim U, et al. Wearables in epilepsy and Parkinson's disease-a focus group study. Acta Neurol Scand 2018;137(2):188-194. [doi: 10.1111/ane.12798] [Medline: 28714112]
42. Fisher JM, Hammerla NY, Ploetz T, et al. Unsupervised home monitoring of Parkinson's disease motor symptoms using body-worn accelerometers. Parkinsonism Relat Disord 2016;33:44-50. [doi: 10.1016/j.parkreldis.2016.09.009] [Medline: 27637282]
43. Kenny L, Moore K, O' Riordan C, et al. The views and needs of people with parkinson disease regarding wearable devices for disease monitoring: mixed methods exploration. JMIR Form Res 2022;6(1):e27418 [FREE Full text] [doi: 10.2196/27418] [Medline: 34989693]
44. Design Partners. Homepage. URL: https://www.designpartners.com/ [accessed 2025-03-04]
45. Sica M, Varnosfaderani OT, Crowe C, et al. Design of a multi-sensors wearable system for continuous home monitoring of people with Parkinson's. IEEE Access 2024;12:38436-38455. [doi: 10.1109/access.2024.3375871]
46. Atkins S, Lewin S, Smith H, Engel M, Fretheim A, Volmink J. Conducting a meta-ethnography of qualitative literature: lessons learnt. BMC Med Res Methodol 2008;8:21 [FREE Full text] [doi: 10.1186/1471-2288-8-21] [Medline: 18416812]
47. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2008 Jul 21;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
48. Ferreira JJ, Godinho C, Santos AT, et al. Quantitative home-based assessment of Parkinson's symptoms: the SENSE-PARK feasibility and usability study. BMC Neurol 2015;15:89 [FREE Full text] [doi: 10.1186/s12883-015-0343-z] [Medline: 26059091]
49. Moore K, O'Shea E, Kenny L, et al. Older adults' experiences with using wearable devices: qualitative systematic review and meta-synthesis. JMIR Mhealth Uhealth 2021;9(6):e23832 [FREE Full text] [doi: 10.2196/23832] [Medline: 34081020]
50. Maffoni M, Giardini A, Pierobon A, Ferrazzoli D, Frazzitta G. Stigma experienced by Parkinson's disease patients: a descriptive review of qualitative studies. Parkinsons Dis 2017;2017:7203259 [FREE Full text] [doi: 10.1155/2017/7203259] [Medline: 28243481]
51. Riggare S, Stamford J, Hägglund M. A long way to go: patient perspectives on digital health for Parkinson's disease. J Parkinsons Dis 2021;11(s1):S5-S10 [FREE Full text] [doi: 10.3233/JPD-202408] [Medline: 33682728]
52. Awotunde JB, Jimoh RG, AbdulRaheem M, Oladipo ID, Folorunso SO, Ajamu GJ. IoT-based wearable body sensor network for COVID-19 pandemic. In: Hassanien AE, Elghamrawy SM, Zelinka I, editors. Advances in Data Science and Intelligent Data Communication Technologies for COVID-19. Studies in Systems, Decision and Control, vol 378. SWITZERLAND: Springer, Cham; 2022:253-275.

XSL•FO
RenderX

53. Keogh A, Taraldsen K, Caulfield B, Vereijken B. It's not about the capture, it's about what we can learn": a qualitative study of experts' opinions and experiences regarding the use of wearable sensors to measure gait and physical activity. J Neuroeng Rehabil 2021;18(1):78 [FREE Full text] [doi: 10.1186/s12984-021-00874-8] [Medline: 33975600]
54. Légaré F, Adekpedjou R, Stacey D, et al. Interventions for increasing the use of shared decision making by healthcare professionals. Cochrane Database Syst Rev 2018;7(7):CD006732 [FREE Full text] [doi: 10.1002/14651858.CD006732.pub4] [Medline: 30025154]

Abbreviations

6-CIT: 6-item Cognitive Impairment Test
ECG: electrocardiogram
MDS-UPDRS: Movement Disorder Society - Unified Parkinson's Disease Rating Scale
PD: Parkinson disease
PPG: photoplethysmography
WESAA: Wearable Enabled Symptom Assessment Algorithm

Edited by A Mavragani; submitted 18.03.25; peer-reviewed by B Cahill, V Manea; comments to author 08.07.25; revised version received 20.08.25; accepted 09.09.25; published 19.11.25.

Please cite as:
Kenny L, Sica M, Crowe C, O'Mahony L, O'Flynn B, Mueller DS, Tedesco S, Barton J, Timmons S
User Experiences of Limb-Worn Wearable Devices for Monitoring Parkinson Disease Motor Function and Blood Pressure: Usability Study
JMIR Form Res 2025;9:e73423
URL: https://formative.jmir.org/2025/1/e73423
doi:10.2196/73423
PMID:

©Lorna Kenny, Marco Sica, Colum Crowe, Lauren O'Mahony, Brendan O'Flynn, David Scott Mueller, Salvatore Tedesco, John Barton, Suzanne Timmons. Originally published in JMIR Formative Research (https://formative.jmir.org), 19.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Evaluating Causal and Noncausal Text Messages to Promote Physical Activity in Adults: Randomized Pilot Study

Elena Korshakova[1], PhD; Samantha Kleinberg[1], PhD

Stevens Institute of Technology, Hoboken, NJ, United States

Corresponding Author:
Samantha Kleinberg, PhD
Stevens Institute of Technology
1 Castle Point Terrace
Hoboken, NJ, 07030
United States
Phone: 1 2012165614
Email: samantha.kleinberg@stevens.edu

Abstract

Background: Physical inactivity increases the risk of chronic disease and reduces life expectancy, yet adherence to physical activity (PA) guidelines remains low. SMS text messages are promising for promoting PA, but it is not clear what type of messaging is most effective. Messages with causal information, which explain why a recommendation is being made, may be more persuasive than messages containing only recommendations.

Objective: This study aims to compare the effectiveness of causal versus noncausal SMS text messages for promoting PA in US adults.

Methods: In this pilot study, we randomized US adults (n=28 in the analytic sample) aged 18-64 years to receive causal or noncausal SMS text messages roughly every other day for 2 weeks, following a 1-week baseline. PA was measured using Empatica wristbands during intervention and baseline periods, and the International Physical Activity Questionnaire – Short Form (IPAQ-SF) at baseline, postintervention, and 4 weeks later. The primary outcome was the change in mean metabolic equivalent of tasks (METs) per minute from baseline to intervention. The secondary outcomes were (1) PA differences on intervention and nonintervention days (mean METs/min), (2) changes in self-reported METs per week between surveyed periods, and (3) participant satisfaction. We used a linear mixed model to analyze our primary outcome, the Mann-Whitney *U* test and the chi-square test of independence to analyze quantitative secondary outcomes, and qualitative coding to analyze survey data.

Results: The causal message group had a greater increase in mean METs per minute from baseline to intervention compared to the noncausal group with a moderate effect size (P=.01; Cohen d=0.54). In the causal group, PA was significantly higher on SMS text message days (mean 2.46, SD 0.12 METs/min) compared to nonmessage days (mean 2.25, SD 0.15 METs/min; P=.02), while there was no difference in the noncausal group (P=.54). No significant between-group difference was found in self-reported PA or satisfaction.

Conclusions: Causal information that links suggested PA to health outcomes can increase the effectiveness of SMS text messages promoting PA, indicating the value of incorporating causal information into intervention design. Our results provide further basis for just-in-time interventions, as activity was higher on message days. Further work is needed to better personalize message content and timing to maintain participant engagement.

(JMIR Form Res 2025;9:e80090) doi:10.2196/80090

KEYWORDS

causal information; text messaging; mHealth; physical activity; health communication

Introduction

Physical inactivity is associated with over 35 chronic medical conditions [1] and contributes to 8.3% of all-cause mortality in US adults, resulting in an estimated 320,000 annual deaths [2]. Despite widespread public health campaigns promoting regular physical activity (PA), adherence remains low. Sedentary behavior has increased among all age groups in the United States [3,4], a trend further exacerbated during the COVID-19 pandemic [5]. Thus, there is a need to address the increase in sedentary behavior [5,6].

XSL•FO
RenderX

To address the need to increase PA, many studies have explored mobile health interventions, particularly using SMS text messages. SMS text messages are cost-effective and can reach most of the US population, as 97% of US adults own a mobile phone capable of receiving SMS text messages [7]. Previous interventions have tested a variety of content types, such as motivating content about the benefits of PA [8] or providing reminders [9], with delivery varying in timing such as targeting morning due to higher motivation levels earlier in the day [10,11] or randomizing timing since responses may depend on individual timing patterns [12]. However, the effectiveness of SMS text message interventions in promoting PA has varied: 59% of studies reported no significant effects, and the remaining 41% demonstrated small to medium increases in PA [13]. The content of SMS text messages varied across studies, though the majority used motivational messages to encourage PA (76% of studies) or provided feedback (22% of studies) [13]. Given the mixed results of prior studies, there is a need to better understand what message characteristics have the largest influence on effectiveness.

One feature that has not been explored is the use of causal information, which explicitly links recommendations (stand up every 30 minutes) to outcomes (standing every 30 minutes can prevent chronic pain). Causal information could help individuals understand why a behavior matters. Recent research found that presenting causal information linking behavior to outcomes in a simple format, such as a 2-node causal diagram, can improve decision-making [14]. However, most experiments in this area have been conducted online using hypothetical decision-making scenarios, so it is unknown whether participants would replicate those choices in real-world situations. Relying on hypothetical scenarios can introduce an intention-behavior gap because intentions to take actions are not always translated into actual behavior [15]. Further, prior work has compared causal information to no information but has not compared causal versus noncausal messages.

To address this gap, we conducted a randomized pilot study to compare causal versus noncausal SMS text messages for increasing PA in real-world settings. Our primary outcome was the change in mean metabolic equivalent of tasks (METs) per minute from baseline to intervention. Secondary outcomes were (1) PA differences on intervention and nonintervention days (mean METs/min between message delivery and midnight), (2) changes in self-reported METs per week between surveyed periods, and (3) participant satisfaction.

Methods

Study Design

We conducted a block-randomized pilot study comparing causal versus noncausal SMS text messages for promoting PA in real-world settings. Participants were assigned in blocks of 4 to either causal or noncausal SMS text messages, with participants being evenly allocated to each group. Data collection for each participant took place over 3 weeks: the first week being a baseline to determine usual PA, and the second 2 weeks being the intervention period. During the intervention weeks, participants received SMS text messages on a set schedule at a set time roughly every other day. PA was measured using objective (wristband tracking METs) and subjective (survey) instruments. Participants received a follow-up survey 4 weeks after completing their active study participation. Data collection occurred from November 2024 to July 2025. The overall study timeline is shown in Figure 1.

Figure 1. Data collection timeline for each participant.

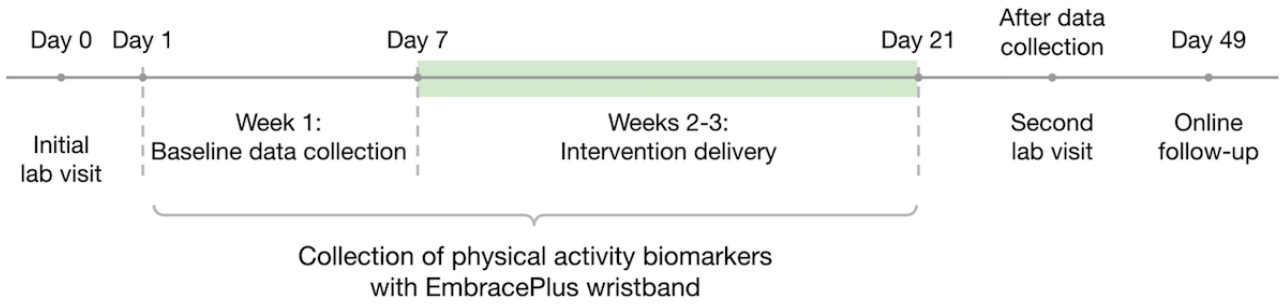

Participants

We recruited individuals aged 18-64 years through community flyers (in New York City, Hoboken, and Jersey City) and a Craigslist posting. Eligibility criteria were no restrictions preventing participation in PA. We screened participants using an online form where participants completed the Physical Activity Readiness Questionnaire (PAR-Q+), a validated tool designed to identify individuals who may require medical clearance before starting an exercise program [16]. The questionnaire includes 7 yes/no questions about the participant's health. If a participant answered "yes" to any of the questions, they were considered at risk and excluded from the study. Other exclusion criteria were inability to communicate in English, pregnancy or breastfeeding, plans to travel for an extended period during the study, or current participation in another PA intervention. Participants received a US $200 Amazon gift card after completing data collection and returning the devices, and a US $50 gift card after completing the follow-up survey.

Interventions

Participants received either causal or noncausal SMS text messages roughly every other day at 8 AM for 2 weeks. We selected an every other day schedule to maintain engagement without causing message fatigue, and selected the 2-week duration based on prior work and the pilot nature of the study [13]. Message timing was selected to increase the likelihood that participants would see the messages early in the day and

XSL•FO
RenderX

have sufficient time to include suggested activities into their daily routine. The message content was created based on PA guidelines developed by government agencies such as the US Centers for Disease Control and Prevention. Causal and noncausal texts were created in pairs such that each provided the same recommendation, with the difference being the causal messages included causal claims that linked recommendations to outcomes (eg, reducing risk of type 2 diabetes). Table 1 shows the full set of messages, delivery schedule, and source used to develop each message. Messages were automatically delivered according to our predetermined schedule. We used Textmagic to send and track messages, as it is a service that enables automated scheduling and delivery of texts, with confirmation of delivery. Participants were not required to reply to the messages, but during the consent meeting they were instructed to read them.

Table 1. Text message schedule and content by study day for each group, with sources used to develop each message.

Day	Causal group	Noncausal group
1 [17,18]	Engaging in moderate-intensity aerobic activities, like brisk walking, for at least 150 minutes every week helps reduce the risk of developing type 2 diabetes in adults. Let's get moving!	It is recommended that adults engage in moderate-intensity aerobic activities, like brisk walking, for at least 150 minutes every week. Let's get moving!
2 [19]	Taking a break from sitting every 30 minutes during your workday can help prevent chronic pain.	Take a break from sitting every 30 minutes during your workday.
4 [20]	Replacing sedentary time with physical activity of any intensity reduces the risk of developing chronic conditions such as arthritis.	It is recommended to replace sedentary time with physical activity of any intensity.
6 [18]	Incorporating moderate-intensity activities into your daily routine, such as brisk walking, biking on level ground or with minimal inclines, or pushing a lawn mower reduces the symptoms of depression and anxiety.	It is recommended to incorporate moderate-intensity activities into your daily routine, such as brisk walking, biking on level ground or with minimal inclines, or pushing a lawn mower.
8 [21]	Moving more and sitting less increases life expectancy. Stay active!	Health guidelines suggest moving more and sitting less. Stay active!
10 [17]	Adding physical activity to your workday by walking during your lunch break and taking the stairs when possible, boosts mental well-being by reducing stress and improving mood.	Add physical activity to your workday by walking during your lunch break and taking the stairs when possible.
12 [17]	By incorporating household chores like sweeping, vacuuming, and doing laundry into your daily routine, you can reduce the risk of heart disease.	Incorporate household chores like sweeping, vacuuming, and doing laundry into your daily routine.
14 [17]	Regular stair climbing improves heart health and helps lower blood pressure.	Take the stairs whenever you can.

Procedures

Interested participants completed 2 online screening questionnaires: a demographic form (to assess eligibility based on age) and the PAR-Q+ (to assess eligibility for a PA intervention). Eligible participants were contacted via email and scheduled for an in-person consent meeting. During the meeting, participants were informed about the study procedures, given an opportunity to ask questions, and provided informed consent to participate. After consenting to the study, participants received both verbal and printed instructions on how to use the devices (smartphone and wristband), completed the International Physical Activity Questionnaire – Short Form (IPAQ-SF) to report baseline PA levels, provided a phone number for receiving SMS text messages, and selected a start date. Many prior studies assessed PA using MET scores from questionnaires or wearable devices [22,23]; thus, we used both measures in this study. Participants were encouraged to contact the investigators with any questions or technical difficulties with devices. After consent, participants were assigned to the causal or noncausal group.

For 3 weeks (1-week baseline and 2-week intervention), participants wore an activity tracking wristband on their nondominant wrist. Initially, the Empatica E4 wristband was used for measuring PA, but the device was discontinued on February 15, 2025. All participants enrolled after that date used the Empatica EmbracePlus wristband for data collection. Both wristbands have accelerometers and additionally capture temperature, electrodermal activity, and heart rate. In addition to the raw data, the EmbracePlus output includes biomarkers such as METs. To collect data from the wristband, participants used a provided Google Pixel 8A smartphone with the Empatica Care Lab app (for the EmbracePlus) or the E4 Realtime App (for the E4). While the wristband records activity, participants were also asked to press a button on the wristband each time they engaged in purposeful PA (eg, walking for exercise rather than for errands).

After completing the 3-week data collection period, participants returned to the lab. During a 30-minute visit, they returned the devices, completed the IPAQ-SF again, and were interviewed about their satisfaction with the intervention through a series of multiple-choice and open-ended questions.

Finally, 4 weeks after completing data collection, participants completed an online follow-up questionnaire to assess the long-term effects of the intervention using the IPAQ-SF.

XSL•FO
RenderX

Outcomes

Objective PA Assessment

The primary study outcome and one secondary outcome focused on PA measured using METs. We selected METs because it is a widely accepted measure of PA intensity and captures energy expenditure across a range of activities, including both formal exercise and everyday tasks like household chores and stair climbing, as suggested in our SMS text messages. The wristband continuously recorded METs and sent data to the app via Bluetooth. When connected to Wi-Fi, the app uploaded data to the cloud, which enabled us to track compliance. Participants wore the wristband during all waking hours and were instructed to charge it overnight. Participants did not receive any feedback from the wristband, which only displayed time of day (EmbracePlus) or had no display (E4), and participants did not see any of the data being collected.

Self-Reported Measures

Our last two secondary outcomes focused on self-reported measures. First, we collected self-reported PA using the IPAQ-SF at 3 time points: before baseline data collection, after completion of the intervention period, and 4 weeks postcompletion. This questionnaire assesses PA categorically (low, medium, and high) and through a continuous score (calculated as a MET level multiplied by the number of minutes of activity and events per week). Due to the potential bias of self-reporting [24], this measure was used as a secondary outcome. Additionally, we surveyed participants about their satisfaction with the intervention using a combination of multiple-choice and open-ended questions. The full survey is in Multimedia Appendix 1.

Sample Size

Due to the preliminary nature of the study, we did not use an a priori power calculation. Based on sample sizes used in prior work to evaluate similar types of interventions [13], we set a sample size of 30 to assess feasibility and estimate effect sizes for future trials.

Randomization

We used block randomization with a block size of 4 to ensure balanced group sizes and distribution of assignment over time. The sequence was generated by the first author (EK).

Blinding

Our primary outcome used METs recorded automatically by the wristband. Participants did not have access to the recorded data, so assessment of outcomes was not influenced by expectations of participants or the investigators. Due to the nature of the intervention, it is not possible to conceal message content from participants or investigators, and data analysis was not blinded.

Analysis

To evaluate differences in demographics between groups, we used a chi-square test of independence, which is suitable for small sample sizes. We used a Mann-Whitney *U* test to compare baseline PA between groups in mean METs per minute, as the MET data were not normally distributed.

The primary outcome was defined as the change in mean METs/min between baseline and the 2-week intervention period. We used a linear mixed model (LMM) with fixed effects of group (causal and noncausal), age, and gender. Our target variable was calculated as the difference between participant-level mean METs/min during the intervention and baseline periods. We used METs/min due to varying durations of recording between and within participants (eg, differing number of waking hours) and incomplete hours (eg, starting recording at 8:56 AM means only 4 minutes of data would be available for that hour). We used random intercepts for participants to account for individual differences in baseline activity, and a random slope for week to allow variation in week-to-week change trajectories across individuals. Random effects were grouped by participant ID. The model was fit using restricted maximum likelihood estimation. We selected LMM as it is appropriate for analyzing repeated measurements taken over time under different conditions and was used in prior work [22]. To quantify the effect size of the between-group difference, we calculated the Cohen *d* score.

Our first secondary outcome focused on differences between message and nonmessage days between groups. Since messages were all delivered at 8 AM local time, we calculated METs/min between 8 AM and 12 AM for each day of the intervention period for each person, and took the mean of message and nonmessage days. We used the Mann-Whitney *U* test to compare mean METs/min on message and nonmessage days.

Self-reported PA was calculated as mean METs per week based on participants' responses to the IPAQ-SF. This measure reflects the frequency and duration of different intensities of activity reported over the past 7 days. The Mann-Whitney *U* test was used to compare the reported PA in mean METs per week between each survey.

Lastly, we analyzed participants' responses to the satisfaction questionnaire using the mean for each response option. We used a chi-square test of independence to compare whether the differences in responses were statistically significant. For qualitative data, we conducted a bottom-up (inductive) thematic analysis. The first author read all responses first, then manually assigned descriptive codes to specific comments (eg, "Something more personalized like facts about you and your physical activity" was coded as "lack of data integration"). Similar codes were then grouped into broader themes (eg, the theme "lack of personalization"), and we calculated the frequency of mentions for each subtheme by group.

Ethical Considerations

The study was approved by the institutional review board (IRB) of Stevens Institute of Technology (IRB protocol #2024-052). All participants provided informed consent before data collection. We maintain privacy of all participants and do not include any identifying information. Participants received a US $200 Amazon gift card after completing data collection and returning the devices, and a US $50 gift card after completing the follow-up survey.

XSL•FO
RenderX

Results

Participants

As shown in Figure 2, a total of 76 individuals completed the screening questionnaire and provided their email addresses for follow-up. Of these, 41 were consented and completed data collection. After examining the data collected with the E4, we observed significant missing data due to the devices frequently disconnecting from the app. Since each participant used a single device for all data collection (baseline and intervention), we thus excluded all participants who used the E4 from analysis. This decision was made prior to any data analysis and led to continuing enrollment until meeting our predetermined target of 30 participants with the EmbracePlus. Thus, 13 participants were excluded from all analyses (n=1 due to protocol deviation, n=1 due to device failure leading to over 5 days of missing data, and n=11 due to data collection using the E4), resulting in a final sample of 28 participants.

Figure 2. CONSORT (Consolidated Standards of Reporting Trials) diagram. Exclusions shown are for participants completely excluded from analysis. Only participants who used the EmbracePlus were included in the analysis. Partial exclusions are described in the text.

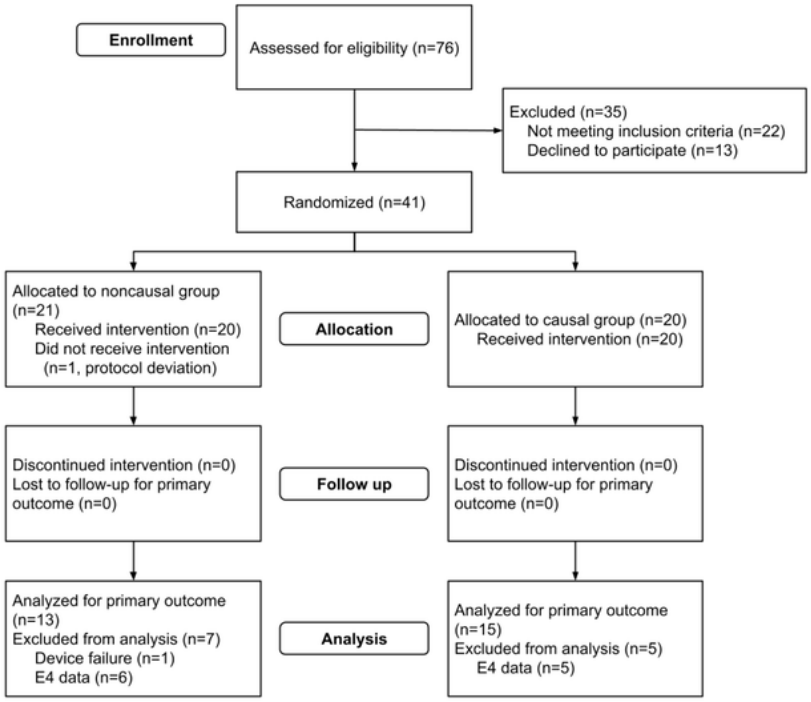

Participants' demographics and baseline activity by group are shown in Table 2, and there were no significant differences between groups. One participant did not complete the follow-up survey and was excluded from analysis of that data (and thus received partial payment for the study, US $200). A total of 588 days of data were collected and included in the analysis. Five days were excluded: 4 participants each had 1 missing day, with 1 day being during the baseline period and 3 days during the intervention period (1 on a day when a message was delivered, and the other 2 on a day without a message). Additionally, 1 participant reported completing an unplanned 32-mile walk during the baseline week. Since the baseline week was intended to represent usual behavior, the 32-mile walk was considered abnormal and that day was excluded, with means then calculated based on the present days.

XSL•FO
RenderX

Table 2. Participant demographics and baseline physical activity. Values are presented as n or mean (SD). *P*<.05 was considered statistically significant.

Variables	Causal (n=15)	Noncausal (n=13)	*P* value
Gender, n			.76
Male	10	7	
Female	5	6	
Age in years, mean (SD)	35 (10.49)	35.18 (10.67)	.89
Race, n			.31
White	4	5	
Asian	7	2	
Black or African American	2	2	
Hispanic, Latino/a, or Spanish Origin	2	2	
Asian and Black or African American	0	1	
Black or African American and Hispanic, Latino/a, or Spanish Origin	0	1	
Education, n			.38
High school	1	3	
Undergraduate degree	6	3	
Graduate degree	8	7	
Baseline physical activity (METs[a] per minute), mean (SD)	2.31 (0.47)	2.38 (0.51)	.61

[a]METs: metabolic equivalent of tasks.

Primary Outcome: Difference in PA Between the Baseline and Intervention Periods

Our primary outcome was the difference in mean METs/min between the baseline and intervention period. As shown in Table 3, there was a significant effect of group, indicating that the increase in the causal group was significantly different. The effect size was moderate (Cohen *d*=0.54), indicating a meaningful difference between the causal and noncausal groups. Additionally, there was a significant effect of gender, with women being more likely to increase PA from baseline to the intervention period compared to men. Table 4 reports the mean METs during the baseline and intervention periods. An increase was observed in the causal group and a decrease in the noncausal group; however, neither difference was statistically significant.

Table 3. Results for linear mixed model predicting change in mean metabolic equivalent of tasks (METs) per minute from baseline to intervention period. Coefficients show the change in METs for a one-unit increase in each predictor, with positive coefficients meaning an increase in METs. *P* values <.05 were considered statistically significant. N=28 for this analysis.

Predictor	β (95% CI)	SE	*z* score	*P* value
Intercept	–0.12 (–0.18 to –0.06)	0.03	–4.05	<.001
Group (causal)	0.13 (0.04 to 0.22)	0.05	2.74	.01
Gender (woman)	0.13 (0.01 to 0.26)	0.06	2.11	.04
Age	0.00 (–0.01 to 0.01)	0.00	0.34	.73

Table 4. Mean (SD) metabolic equivalent of tasks per minute during baseline and intervention periods (compared using the Mann-Whitney U test). N=28 for this analysis.

	Baseline	Intervention	*P* value
Causal	2.31 (0.47)	2.40 (0.47)	.51
Noncausal	2.38 (0.51)	2.37 (0.46)	.99

Comparing PA on Days With and Without Message Delivery by Group

Our primary outcome was overall activity change. We next examined differences in PA between days with and without SMS text message delivery. Participants in the causal group were significantly more active on days when they received SMS text messages (mean 2.46, SD 0.12) compared to days when they did not (mean 2.25, SD 0.15; *U*=6; *P*=.02). The difference was not significant in the noncausal group (days with texts: mean 2.40, SD 0.26; days without texts: mean 2.33, SD 0.19; *U*=19; *P*=.57).

XSL•FO
RenderX

Changes in Self-Reported PA

Based on self-reported data, baseline MET-minutes/week did not differ between participants in the causal group (mean 1661.82, SD 1026.94) and the noncausal group (mean 2316.75, SD 1562.08; *U*=48; *P*=.29; n=27). Both groups reported an increase in MET-minutes/week from baseline to intervention, though this increase was not significantly different between the causal (mean 1037.49, SD 1501.28) and noncausal groups (mean 369.45, SD 366.45; *U*=94; *P*=.15). At 4 weeks after the intervention, MET-minutes/week did not change significantly from baseline in either the causal (mean 372.12, SD 986.39; *U*=68; *P*=.36) or noncausal group (mean 83.21, SD 423.76; *U*=78; *P*=.42).

Participant Satisfaction

We first analyzed the quantitative survey data. As shown in Table 5, while there were some differences, none were significant between the two groups. Most participants in both groups reported that they always read the messages and found them easy or very easy to understand. While more participants reported satisfaction with text frequency in the causal compared to the noncausal group, the difference was not significant. Similarly, 8 participants in the causal group reported that the messages helped them better achieve their fitness goals, while only 2 participants in the noncausal group reported the same.

Lastly, we examined participants' answers to the open-ended questions using thematic analysis, as shown in Table 6. Across all participants (N=28), 15 were dissatisfied with the frequency of message delivery and expressed a preference for more messages. Six participants in the causal group indicated that references to specific health outcomes, such as diabetes and mental health, made the messages motivating.

> *The reminders of specific consequences like diabetes and chronic pain really stuck with me. Every message felt like a gentle but effective nudge toward being more active.* [Participant 7]

In the noncausal group, 4 participants mentioned that the messages were too generic and easy to ignore or forget.

> *The messages were too generic and failed to capture my attention.* [Participant 12]

Participants in both groups expressed positive views about the messages as helpful reminders. Three participants across both groups mentioned that they would prefer a friendlier, more conversational tone in the messages.

> *I think it's the copywriting that needs to be more conversational. It sounded too sterile and cold.* [Participant 9]

Participants in both groups mentioned a lack of personalization (n=22), expressing a preference for receiving messages specifically after periods of inactivity, when the reminder would be more relevant.

> *Messages could be more immediate when you don't move for a ling [sic] time.* [Participant 4]

XSL•FO
RenderX

Table 5. Participants' responses to the satisfaction questionnaire by group.

Question	Causal (n=15), n	Noncausal (n=13), n	*P* value
How often did you read the text messages?			.99
Always	13	12	
Often	2	1	
Sometimes	0	0	
Rarely	0	0	
Never	0	0	
Did the frequency of the text messages meet your expectations?			.24
Yes	9	4	
No	6	9	
Did the text messages help you better achieve your fitness goals?			.09
Yes	8	2	
No	7	11	
Did the text messages help you feel more motivated to make healthy lifestyle changes?			.28
Yes	11	6	
No	4	7	
How easy was it to understand the information provided in the text messages?			.83
Very easy	12	9	
Easy	3	4	
Neither easy nor difficult	0	0	
Difficult	0	0	
Very difficult	0	0	
Would you recommend this text message program to a friend?			.27
Yes	13	8	
No	2	5	
How likely are you to continue using the strategies or advice provided in the text messages after the study?			.30
Very likely	6	1	
Likely	7	8	
Neither likely nor unlikely	1	1	
Unlikely	1	2	
Very unlikely	0	1	
Overall, how satisfied were you with the convenience and accessibility of receiving health information via text messages?			.18
Very satisfied	10	6	
Satisfied	3	6	
Neither satisfied nor dissatisfied	2	0	
Dissatisfied	0	1	
Very dissatisfied	0	0	

XSL•FO
RenderX

Table 6. Themes and subthemes identified from participants' answers to open-ended questions. Numbers in the last two columns represent the number of mentions by group.

Theme and subtheme	Causal (n=15)	Noncausal (n=13)
Frequency		
Preference for more frequent messages	6	9
Message content		
Motivating messages	6	0
Generic content	0	4
Reminders	5	4
Tone of messages	2	1
Educational content	2	0
Already known information	0	2
Lack of personalization		
Lack of data integration	6	4
Inconvenient delivery timing	3	4
Lack of personalized feedback	2	3
Message features		
Not interactive	1	2
Not tied to real-time behavior data	2	1
Technology and usability		
App issues	1	0
Technical or connectivity issues	1	0
Recommendation logic		
Better for inactive users	2	3
Not helpful for already active users	1	2
Other		
No changes needed	3	2
No feedback	2	1

Discussion

Principal Results

We found that causal messages were more effective than noncausal ones, leading to a statistically significant increase in PA with a moderate effect size. Analysis of the secondary outcomes showed that participants in the causal group were more active on days when they received messages than on days when they did not. Potentially due to the small sample size, there were no significant differences in self-reported PA increase between groups or time periods, and no significant differences in satisfaction were found.

Comparison With Prior Work

We developed causal messages to provide a clear rationale for behavior change by linking PA to meaningful health outcomes, such as a reduced risk of chronic diseases like type 2 diabetes and improved mental health. This approach was effective for increasing PA, and our qualitative results suggest that the causal component may be especially motivating. Our findings advance prior work on SMS text messaging for increasing PA by identifying for the first time that causal messages may be more effective than statements that only recommend an activity. Our results align with previous research suggesting that individuals are more likely to engage in a behavior when they understand its relevance and how it impacts their overall well-being [25]. By explicitly connecting PA to potential health outcomes, causal messages may have increased perceived relevance and personal significance [26]. Similarly, Fogg's Behavior Model emphasizes that increasing the perceived importance of an action can serve as a trigger for behavior change [27].

Participants in the causal group were significantly more active on days when they received messages, and dissatisfaction with message frequency was due to a desire for more messages rather than burnout. Our messaging schedule was designed to avoid message fatigue, but our results show that more frequent messages could be both acceptable and effective. Prior research suggests that individualized or adaptive message schedules are often more effective than fixed-frequency approaches [28]. Our results suggest that consistent exposure to causal messages may

XSL•FO
RenderX

promote more sustainable engagement in PA. However, it is an open question whether these results would differ in a long-term intervention, as our study period was 2 weeks. Although more than half of the participants expressed a preference for more frequent messages, allowing people to customize message frequency according to their needs may increase the effectiveness of future interventions. Participants also mentioned a preference for receiving messages after prolonged periods of inactivity. This aligns with existing evidence that timely, context-sensitive prompts can positively influence behavior in real time [29]. Thus, it may be important to balance regular message delivery with personalized scheduling to optimize engagement and minimize fatigue in future PA interventions.

The increase in self-reported PA in our study was not significant from baseline to the intervention period and 4 weeks postintervention between groups, which may be due to the small sample size. Further, the self-reported values were inconsistent with the objective measures, where we found a significant increase in METs in the causal group and no increase in the noncausal group. This discrepancy is consistent with prior research showing that self-reported PA is often overestimated and may not accurately reflect actual behavior [30,31]. Our self-reported measures and METs measured with the EmbracePlus are not directly comparable, but the difference in trends highlights the importance of using objective measures when evaluating the effectiveness of behavioral interventions.

Our findings were supported by insights from the participant surveys. Participants reported that causal messages linking PA to health outcomes increased their motivation to be physically active. Additionally, we found that most participants wanted messages to be more personalized and timelier, based on their PA data. Since our work was a pilot, future work is needed to explore whether causal framing is still effective when combined with personalized PA data to deliver messages during periods of inactivity. Interventions could consider individuals' calendars, targeting moments when they are available to take a break and engage in PA.

Limitations

Our study had several limitations. The sample size for this pilot study was small and consisted of residents from New York and New Jersey, which may limit the generalizability of our findings to broader populations. The intervention period was only 2 weeks, and longer-term adherence and outcomes remain unknown. This is especially important given our finding that participants wanted more frequent messages. Since messages were on a set schedule, it is not known whether tailoring them to individual activity levels would have benefits beyond that of causal messaging, nor whether varying the schedule is important for long-term engagement. Future research is needed to explore the long-term effectiveness of causal messaging, test more personalized message delivery schedules, and examine the use of interventions based on real-time activity data. Finally, due to the nature of the intervention and pilot study, allocation could not be fully concealed. While our use of objective measures limits the potential for bias in outcome assessment, this is nevertheless a limitation.

Conclusions

We found that incorporating causal information into an SMS text message intervention for promoting PA can increase their effectiveness. Specifically, participants who received messages that included causal information linking recommended PA to health outcomes showed a greater increase in PA from baseline compared to those who received the messages with the same PA recommendations without causal information, and participants in the causal group had more PA on days with messages than on days without. These findings have important implications for the design of future behavioral interventions. We recommend incorporating causal information that explicitly connects PA recommendations to relevant health outcomes, such as reduced risk of chronic diseases. Moreover, tailoring message delivery to participants' preferences and real-time context may support sustained engagement and increase the intervention's effectiveness.

Funding

This work was supported by the National Science Foundation under award number 1907951.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

EK contributed to the conceptualization, data curation, formal analysis, investigation, methodology, and visualization of the study. She took the lead in writing the original draft and provided supporting contributions during the review and editing process. SK contributed to the conceptualization, funding acquisition, and methodology. She provided supporting contributions to writing the original draft and took the lead in the review and editing process.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Survey questionnaire.

XSL•FO
RenderX

[DOCX File , 17 KB - formative_v9i1e80090_app1.docx]

References

1. Booth FW, Roberts CK, Laye MJ. Lack of exercise is a major cause of chronic diseases. Compr Physiol 2012;2(2):1143-1211 [FREE Full text] [doi: 10.1002/cphy.c110025] [Medline: 23798298]
2. Carlson SA, Adams EK, Yang Z, Fulton JE. Percentage of deaths associated with inadequate physical activity in the United States. Prev Chronic Dis 2018;15:E38 [FREE Full text] [doi: 10.5888/pcd18.170354] [Medline: 29602315]
3. Matthews CE, Carlson SA, Saint-Maurice PF, Patel S, Salerno EA, Loftfield E, et al. Sedentary behavior in U.S. adults: fall 2019. Med Sci Sports Exerc 2021;53(12):2512-2519 [FREE Full text] [doi: 10.1249/MSS.0000000000002751] [Medline: 34310489]
4. Yang L, Cao C, Kantor ED, Nguyen LH, Zheng X, Park Y, et al. Trends in sedentary behavior among the US population, 2001-2016. JAMA 2019;321(16):1587-1597 [FREE Full text] [doi: 10.1001/jama.2019.3636] [Medline: 31012934]
5. Costa TG, Ramos-Urrea C, dos Santos DAT, Ostolin TLVDP, de Lira CAB. Strategies to combat the increase of sedentary behavior after COVID-19 pandemic: an opinion article. Res Soc Dev 2022;11(6):e9011628689. [doi: 10.33448/rsd-v11i6.28689]
6. Evenson KR, Alothman SA, Moore CC, Hamza MM, Rakic S, Alsukait RF, et al. A scoping review on the impact of the COVID-19 pandemic on physical activity and sedentary behavior in Saudi Arabia. BMC Public Health 2023;23(1):572 [FREE Full text] [doi: 10.1186/s12889-023-15422-3] [Medline: 36973687]
7. Mobile fact sheet. Pew Research Center. 2024 Apr 1. URL: https://www.pewresearch.org/internet/fact-sheet/mobile/ [accessed 2024-04-01]
8. St Quinton T, Morris B, Barwood MJ, Conner M. Promoting physical activity through text messages: the impact of attitude and goal priority messages. Health Psychol Behav Med 2021;9(1):165-181 [FREE Full text] [doi: 10.1080/21642850.2021.1891073] [Medline: 34104555]
9. Ibeneme SC, Ndukwu SC, Myezwa H, Irem FO, Ezenwankwo FE, Ajidahun AT, et al. Effectiveness of mobile text reminder in improving adherence to medication, physical exercise, and quality of life in patients living with HIV: a systematic review. BMC Infect Dis 2021;21(1):859 [FREE Full text] [doi: 10.1186/s12879-021-06563-0] [Medline: 34425789]
10. Grady CL, Murtagh E, Ng K, Bengoechea EG, Woods CB. Communicating physical activity messages with adolescents: what works? A scoping review with stakeholder consultation. Int J Behav Nutr Phys Act 2025;22(1):20 [FREE Full text] [doi: 10.1186/s12966-025-01717-8] [Medline: 39972320]
11. Huberty JL, Buman MP, Leiferman JA, Bushar J, Hekler EB, Adams MA. Dose and timing of text messages for increasing physical activity among pregnant women: a randomized controlled trial. Transl Behav Med 2017;7(2):212-223 [FREE Full text] [doi: 10.1007/s13142-016-0445-1] [Medline: 27800565]
12. Lee AM, Hojjatinia S, Courtney JB, Brunke-Reese D, Hojjatinia S, Lagoa CM, et al. Motivational message framing effects on physical activity dynamics in a digital messaging intervention: secondary analysis. JMIR Form Res 2023;7:e41414 [FREE Full text] [doi: 10.2196/41414] [Medline: 37083710]
13. Smith DM, Duque L, Huffman JC, Healy BC, Celano CM. Text message interventions for physical activity: a systematic review and meta-analysis. Am J Prev Med 2020;58(1):142-151 [FREE Full text] [doi: 10.1016/j.amepre.2019.08.014] [Medline: 31759805]
14. Kleinberg S, Marsh JK. Less is more: information needs, information wants, and what makes causal models useful. Cogn Res Princ Implic 2023;8(1):57 [FREE Full text] [doi: 10.1186/s41235-023-00509-7] [Medline: 37646868]
15. Sheeran P, Webb TL. The intention–behavior gap. Soc Personal Psychol Compass 2016;10(9):503-518. [doi: 10.1111/spc3.12265]
16. Bredin SSD, Gledhill N, Jamnik VK, Warburton DER. PAR-Q+ and ePARmed-X+: new risk stratification and physical activity clearance strategy for physicians and patients alike. Can Fam Physician 2013;59(3):273-277 [FREE Full text] [Medline: 23486800]
17. Steps for getting started with physical activity. U.S. Centers for Disease Control and Prevention. 2024 Apr 17. URL: https://www.cdc.gov/healthy-weight-growth/physical-activity/getting-started.html?CDC_AAref_Val=https://www.cdc.gov/healthyweight/physical_activity/getting_started.html [accessed 2024-04-17]
18. How much physical activity do adults need? U.S. Centers for Disease Control and Prevention. 2024 Apr. URL: https://www.cdc.gov/physicalactivity/basics/adults/index.htm [accessed 2024-04-17]
19. DeskFit. NASA Headquarters Fitness Center. 2020. URL: https://www.nasa.gov/wp-content/uploads/2020/03/hq_deskfit_booklet_6.10.2020.pdf [accessed 2024-04-17]
20. Physical activity. WHO. 2024 Jun 26. URL: https://www.who.int/news-room/fact-sheets/detail/physical-activity [accessed 2024-04-17]
21. Top 10 things to know about the second edition of the Physical Activity Guidelines for Americans. OASH. URL: https://odphp.health.gov/our-work/nutrition-physical-activity/physical-activity-guidelines/current-guidelines/top-10-things-know [accessed 2024-04-17]
22. Bootwong P, Intarut N. The effects of text messages for promoting physical activities in prediabetes: a randomized controlled trial. Telemed J E Health 2022;28(6):896-903. [doi: 10.1089/tmj.2021.0303] [Medline: 34619066]

XSL•FO
RenderX

23. Dietz de Loos A, Jiskoot G, van den Berg-Emons R, Louwers Y, Beerthuizen A, van Busschbach J, et al. The effect of tailored short message service (SMS) on physical activity: results from a three-component randomized controlled lifestyle intervention in women with PCOS. J Clin Med 2023;12(7):896-903 [FREE Full text] [doi: 10.3390/jcm12072466] [Medline: 37048550]
24. Meh K, Sember V, Sorić M, Vähä-Ypyä H, Rocha P, Jurak G. The dilemma of physical activity questionnaires: fitter people are less prone to over reporting. PLoS One 2023;18(8):e0285357 [FREE Full text] [doi: 10.1371/journal.pone.0285357] [Medline: 37647304]
25. Michaelsen MM, Esch T. Understanding health behavior change by motivation and reward mechanisms: a review of the literature. Front Behav Neurosci 2023;17:1151918 [FREE Full text] [doi: 10.3389/fnbeh.2023.1151918] [Medline: 37405131]
26. Van Cappellen P, Rice EL, Catalino LI, Fredrickson BL. Positive affective processes underlie positive health behaviour change. Psychol Health 2018;33(1):77-97 [FREE Full text] [doi: 10.1080/08870446.2017.1320798] [Medline: 28498722]
27. Fogg BJ. A behavior model for persuasive design. In: Persuasive '09: Proceedings of the 4th International Conference on Persuasive Technology. New York: Association for Computing Machinery; 2009:1-7.
28. Head KJ, Noar SM, Iannarino NT, Grant Harrington N. Efficacy of text messaging-based interventions for health promotion: a meta-analysis. Soc Sci Med 2013;97:41-48. [doi: 10.1016/j.socscimed.2013.08.003] [Medline: 24161087]
29. Nahum-Shani I, Smith SN, Spring BJ, Collins LM, Witkiewitz K, Tewari A, et al. Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. Ann Behav Med 2018;52(6):446-462 [FREE Full text] [doi: 10.1007/s12160-016-9830-8] [Medline: 27663578]
30. Schaller A, Rudolf K, Dejonghe L, Grieben C, Froboese I. Influencing factors on the overestimation of self-reported physical activity: a cross-sectional analysis of low back pain patients and healthy controls. Biomed Res Int 2016;2016:1497213 [FREE Full text] [doi: 10.1155/2016/1497213] [Medline: 27298820]
31. Wunsch K, Nigg CR, Weyland S, Jekauc D, Niessner C, Burchartz A, et al. The relationship of self-reported and device-based measures of physical activity and health-related quality of life in adolescents. Health Qual Life Outcomes 2021;19(1):67 [FREE Full text] [doi: 10.1186/s12955-021-01682-3] [Medline: 33648492]

Abbreviations

IPAQ-SF: International Physical Activity Questionnaire – Short Form
IRB: institutional review board
LMM: linear mixed model
MET: metabolic equivalent of task
PA: physical activity
PAR-Q+: Physical Activity Readiness Questionnaire

Edited by A Mavragani; submitted 30.07.25; peer-reviewed by A Singleton, X Su; comments to author 23.09.25; revised version received 03.11.25; accepted 03.11.25; published 24.11.25.

Please cite as:
Korshakova E, Kleinberg S
Evaluating Causal and Noncausal Text Messages to Promote Physical Activity in Adults: Randomized Pilot Study
JMIR Form Res 2025;9:e80090
URL: https://formative.jmir.org/2025/1/e80090
doi:10.2196/80090
PMID:

©Elena Korshakova, Samantha Kleinberg. Originally published in JMIR Formative Research (https://formative.jmir.org), 24.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Validity of Generative Artificial Intelligence in Evaluating Medical Students in Objective Structured Clinical Examination: Experimental Study

Masashi Yokose[1], MD, PhD; Takanobu Hirosawa[1], MD, PhD; Tetsu Sakamoto[1], MD, PhD; Ren Kawamura[1], MD, PhD; Yudai Suzuki[2], MD; Yukinori Harada[1], MD, PhD; Taro Shimizu[1], MD, MSc, MPH, MBA, FACP, PhD

[1]Department of Diagnostic and Generalist Medicine, Dokkyo Medical University, Tochigi, Japan

[2]Department of Internal Medicine, Yamagata Prefectural Kahoku Hospital, Yamagata, Japan

Corresponding Author:
Masashi Yokose, MD, PhD
Department of Diagnostic and Generalist Medicine
Dokkyo Medical University
880 Kitakobayashi
Mibu, Shimotsuga
Tochigi, 321-0293
Japan
Phone: 81 90 1663 2272
Email: myokose@dokkyomed.ac.jp

Abstract

Background: The Objective Structured Clinical Examination (OSCE) has been widely used to evaluate students in medical education. However, it is resource-intensive, presenting challenges in implementation. We hypothesized that generative artificial intelligence (AI) such as ChatGPT-4 could serve as a complementary assessor and alleviate the burden of physicians in evaluating OSCE.

Objective: By comparing the evaluation scores between generative AI and physicians, this study aims to evaluate the validity of generative AI as a complementary assessor for OSCE.

Methods: This experimental study was conducted at a medical university in Japan. We recruited 11 fifth-year medical students during the general internal medicine clerkship from April 2023 to December 2023. Participants conducted a mock medical interview with a patient experiencing abdominal pain and wrote patient notes. Four physicians independently evaluated the participants by reviewing medical interview videos and patient notes, while ChatGPT-4 was provided with interview transcripts and notes. Evaluations were conducted using the 6-domain rubric (patient care and communication, history taking, physical examination, patient notes, clinical reasoning, and management). Each domain was scored using a 6-point Likert scale, ranging from 1 (very poor) to 6 (excellent). Median scores were compared using the Wilcoxon signed-rank test, and the agreement between ChatGPT-4 and physicians was assessed using intraclass correlation coefficients (ICCs). All P values <.05 were considered statistically significant.

Results: Although ChatGPT-4 assigned higher scores than physicians in terms of physical examination (median 4.0, IQR 4.0-5.0 vs median 4.0, IQR 3.0-4.0; P=.02), patient notes (median 6.0, IQR 5.0-6.0 vs median 4.0, IQR 4.0-4.0; P=.002), clinical reasoning (median 5.0, IQR 5.0-5.0 vs median 4.0, IQR 3.0-4.0; P<.001), and management (median 6.0, IQR 5.0-6.0 vs median 4.0, IQR 2.5-4.5; P=.002), there were no significant differences in the scores of patient care and communication (median 5.0, IQR 5.0-5.0 vs median 5.0, IQR 4.0-5.0; P=.06) and history taking (median 5.0, IQR 4.0-5.0 vs median 5.0, IQR 4.0-5.0; P>.99), respectively. ICC values were low in all domains, except for history taking, where the agreement was still poor (ICC=0.36, 95% CI –0.32 to 0.78).

Conclusions: ChatGPT-4 produced higher evaluation scores than physicians in several OSCE domains, though the agreement between them was poor. Although these preliminary results suggest that generative AI may be able to support assessment in some domains of OSCE, further research is needed to establish its reproducibility and validity. Generative AI like ChatGPT-4 shows potential as a complementary assessor for OSCE.

XSL•FO
RenderX

Trial Registration: University Hospital Medical Information Network Clinical Trials Registry UMIN000050489; https://center6.umin.ac.jp/cgi-open-bin/ctr/ctr_his_list.cgi?recptno=R000057513

(JMIR Form Res 2025;9:e79465) doi:10.2196/79465

KEYWORDS

objective structured clinical examination; medical education; generative artificial intelligence; ChatGPT-4; OSCE; medical students

Introduction

Patient encounters, including history taking and physical examination, are fundamental skills in clinical medicine. High-quality patient encounters can yield numerous benefits, including timely and accurate diagnoses, a comprehensive understanding of a patient's status, the establishment of rapport between physicians and patients, and the development of optimal treatment strategies [1]. Given their importance, evaluating and ensuring the competency of health professionals in conducting these encounters is essential.

The Objective Structured Clinical Examination (OSCE) is one of the most widely accepted, objective, and standardized assessment tools in medical education [2]. It has been extensively employed to assess clinical competencies such as communication, physical examination, clinical reasoning, and procedural techniques [3]. OSCE typically consists of a series of time-limited stations where candidates perform specific clinical tasks under observation, often involving interactions with standardized patients (SPs). By simulating real-life clinical situations in a controlled environment, OSCE enables consistent assessment and provides the opportunity to evaluate a wide range of clinical skills [3]. As such, it contributes to improving the validity and reliability of clinical skills assessment, supporting the application of theoretical knowledge in practical, patient-centered care [4].

Despite its benefits, OSCE has some disadvantages. A major challenge is its high demand for human and logistical resources. Effective implementation requires trained examiners, SPs, administrative staff, and physical infrastructure [3,5,6]. Preparing SPs and ensuring objective scoring are both time- and labor-intensive. Financial costs are also substantial, with reports estimating up to €10,000 (about US $11,598) per exam and over €100 (about US $115) per examinee [7]. Furthermore, despite standardization efforts, interrater variability and scoring subjectivity can undermine assessment reliability [3,8]. These challenges have led to implementation difficulties in some settings [9].

In recent years, generative artificial intelligence (AI), particularly large language models, has shown potential in medicine [10]. In the context of medical education, such models may offer benefits including personalized feedback, enhanced simulation, and support for self-directed learning [11]. One emerging application is the use of AI to assist with or complement assessment processes in OSCEs. Although AI cannot currently replicate the full range of human evaluators' observational abilities, it may help reduce certain aspects of examiner workload such as providing structured feedback or scoring text-based materials [12]. However, the validity of AI-based assessment remains underexplored, particularly when compared with experienced human judgment.

This study aims to provide preliminary evidence of the validity of generative AI in evaluating the performance of medical students in OSCE. By comparing the evaluations between physicians and AI, we aim to explore the feasibility and potential limitations of generative AI as a supplemental tool for OSCE evaluation.

Methods

Overview

We conducted an experimental study to assess ChatGPT-4's evaluation of students' mock medical interviews compared to the evaluation by physicians. This research was conducted at the Department of Diagnostic and Generalist Medicine (General Internal Medicine), Dokkyo Medical University, Tochigi, Japan.

Ethical Considerations

This study followed the Declaration of Helsinki. The study protocol was approved by the institutional review board of Dokkyo Medical University (approval 2022-015). All participants provided written informed consent prior to participation. They were informed that their participation was voluntary and that they could withdraw from the study at any time without penalty. Participants received a 2,000-yen (about US $12.80) coffee ticket as compensation. All components, including the medical interviews, their transcriptions, the structured prompts, and outputs from ChatGPT-4, were in Japanese. To safeguard privacy, all data were anonymized before use. To prevent biases from previous interactions, we deactivated the settings related to ChatGPT-4's chat history and training controls. Moreover, each session was restarted before conducting a new evaluation. This approach ensured a single, unbiased output from ChatGPT-4 for each medical interview. Our research methods included recruiting participants, conducting mock medical interviews, transcribing and verifying the data, inputting the data into ChatGPT-4, and evaluation by physicians.

Participant Recruitment

We recruited fifth-year medical students who had rotated through the General Internal Medicine department at Dokkyo Medical University. Unlike the 4-year medical curriculum in many countries, Japan follows a 6-year medical curriculum. Typically, Japanese medical students learn basic medicine in the first two years and clinical medicine in the following two years. After passing the standard achievement tests in Japanese and OSCE, they attend clinical clerkships in the last two years.

XSL•FO
RenderX

Written informed consent was obtained from each medical student a week after the study's protocols were explained.

Conducting Mock Medical Interviews

Participants took a face-to-face mock medical interview with a patient experiencing intermittent abdominal pain in the primary care setting. The scenario was originally developed based on a textbook of the United States Medical Licensing Examination Step 2 Clinical Skills [13] and had been implemented in our general internal medicine clerkship for 2 years prior to the study with regular updates. At the beginning of the interview, participants had 1 minute to review the patient's basic information, including age, sex, chief complaints, and vital signs. Then, participants underwent a 12-minute history taking and physical examination. After the interview, participants completed patient notes on the laptop in 15 minutes. We prepared a template for the patient note, and participants filled in patient history, physical examination findings, assessments (including up to 3 differential diagnoses with supporting evidence), and management plan. We allocated the time based on the United States Medical Licensing Examination Step 2 Clinical Skills [13] and the OSCE that sixth-year medical students take after clinical clerkships in Japan [14]. We recorded each interview. Interviews were conducted in Japanese, and participants completed patient notes in Japanese.

Transcribing and Verifying the Data

The leading researcher (MY) reviewed medical interview videos and transcribed the conversation verbatim. Another researcher (TH) then reviewed the same interview videos to verify the accuracy of the transcripts and to finalize them. To clarify who was speaking in the conversation, we added the headers "Physician" or "Patient" to each conversation in the medical interview. The names of the medical students, as they introduced themselves in the medical interview, were changed to pseudonyms for privacy protection. Any discrepancies such as transcription differences between MY and TH were discussed and resolved. We present examples of the transcripts (Textbox 1) and patient notes (Textbox 2).

Textbox 1. An example of the transcripts of the medical interview. This is a part of the English translation of the original transcripts written in Japanese.

…

Physician: Thank you. So, what brings you in today?

Patient: Oh, today, I came because I have a bit of a stomachache.

Physician: A stomachache. I see. Could you tell me a little more about this stomach pain?

Patient: Please ask me anything, Doctor.

Physician: Okay, then. I assume there was a point when the pain started?

…

Textbox 2. An example of patient notes. This is a part of the English translation of the original patient notes written in Japanese.

(History)

Epigastric pain has been present for the past two weeks. The pain has been worsening and continues to persist.

…

(Physical examination findings)

Consciousness: Clear and coherent.

General appearance: Tends to be obese.

…

(Assessment)

Differential diagnosis #1: gastric ulcer

Supporting history: Epigastric pain that worsens after meals, particularly with pizza and hamburger intake

…

(Plan)

Blood tests (including C-Reactive Protein, WBC)

Upper gastrointestinal endoscopy

Inputting Data Into ChatGPT-4

We selected ChatGPT-4 with ChatGPT Plus plan for this study due to its advanced natural language processing capabilities and its potential in various fields, including health care. ChatGPT-4, developed by OpenAI, is recognized for its ability to understand and generate human-like text, making it a suitable tool for evaluating the complexities involved in medical interviews. Its multilingual support, including Japanese, further enabled us to conduct this research effectively within our local context [15].

XSL•FO
RenderX

We accessed ChatGPT-4 by using an individual ChatGPT Plus account. The access period was January 11, 2024. We input the transcribed medical interviews along with patient notes and structured prompts into ChatGPT-4. Patient notes included assessment and plan for the mock case. We created structured prompts based on a preliminary investigation, and each prompt combined the interview transcript with specific evaluation criteria. We adopted an evaluation format derived from the OSCE in Japan [16], which consisted of the following 6 items: patient care and communication, history taking, physical examination, patient notes, clinical reasoning, and management. For each aspect, ChatGPT-4 was asked to evaluate based on a scale of 1 to 6, with 1 being the lowest and 6 being the highest. The criteria for each score were clearly defined. To prevent biases from previous interactions, we deactivated the settings related to ChatGPT-4's chat history and training controls. Moreover, each session was restarted before conducting a new evaluation. This approach ensured a single, unbiased output from ChatGPT-4 for each medical interview. Details of the prompt are provided in Multimedia Appendix 1.

Evaluation by Physicians

We used a consensus-based approach in which 2 pairs of physicians (T Sakamoto and YS, RK and YH) were randomly allocated to the evaluation of 5 or 6 participants. Then, each physician independently reviewed medical interview videos and patient notes and evaluated the participants based on the same criteria as ChatGPT-4. Any disagreements within the pair regarding the evaluation were resolved through discussion.

Data Collection and Outcome

We collected data on participants' sex and prior experience with medical interviews. The primary outcome was the difference in the evaluation scores between ChatGPT-4 and physicians in each domain.

Statistical Analysis

The evaluation scores by physicians and ChatGPT-4 were presented as median with 25th and 75th percentiles and were compared using the Wilcoxon signed-rank test. We calculated intraclass correlation coefficients (ICCs) and their 95% CIs to evaluate the agreement between ChatGPT-4 and physician evaluations (ie, interrater reliability). A 2-way random-effects model with absolute agreement was used, based on a single measurement across 11 participants. Generally, the degree of reliability is estimated along with the following ICC values: <0.5=poor; 0.5-0.75=moderate; 0.75-0.9=good; >0.9=excellent [17]. All P values <.05 were considered statistically significant. We used R software (version 4.3.1; The R Foundation for Statistical Computing) for statistical analysis.

Results

Baseline Characteristics

From April 2023 to December 2023, 119 fifth-year medical students rotated through the General Internal Medicine department at Dokkyo Medical University, and 11 students participated in this study (Figure 1). All participants (11/11) had experienced medical interviews less than 10 times before this study, and 55% (6/11) of the participants were females.

Figure 1. Study flow diagram of the experimental evaluation of generative artificial intelligence in medical education. The study was conducted at Dokkyo Medical University, Japan, between April and December 2023. A total of 11 fifth-year medical students were recruited during their general internal medicine clerkship. Each participant completed a mock medical interview with a standardized patient. Physicians evaluated each student by using both the video recordings and written patient notes, while ChatGPT-4 conducted its evaluation based on interview transcripts and patient notes.

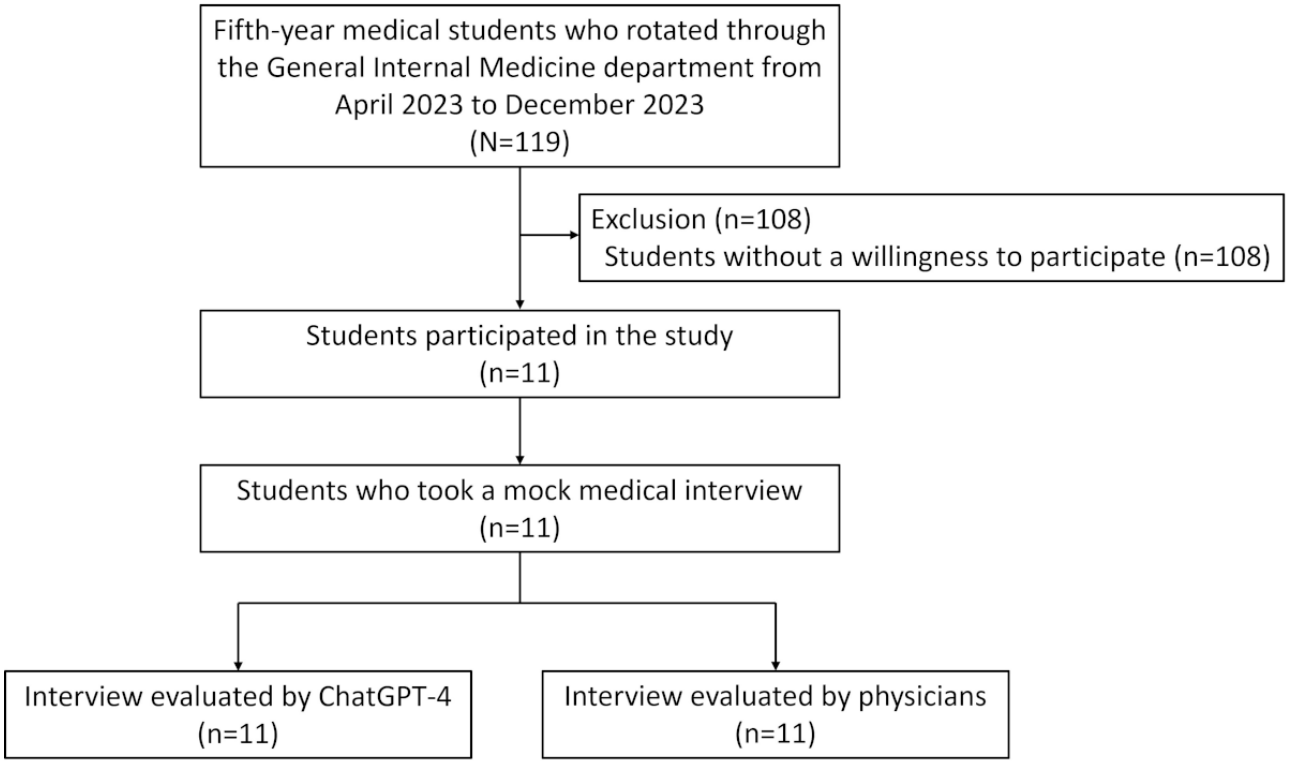

XSL•FO
RenderX

Evaluation Outcomes

Table 1 summarizes the comparison of the evaluation scores between ChatGPT-4 and physicians. With the 6-item evaluation format, ChatGPT-4 assigned higher scores than physicians in terms of physical examination (median 4.0, IQR 4.0-5.0 vs median 4.0, IQR 3.0-4.0; *P*=.02), patient notes (median 6.0, IQR 5.0-6.0 vs median 4.0, IQR 4.0-4.0; *P*=.002), clinical reasoning (median 5.0, IQR 5.0-5.0 vs median 4.0, IQR 3.0-4.0; *P*<.001), and management (median 6.0, IQR 5.0-6.0 vs median 4.0, IQR 2.5-4.5; *P*=.002). There were no significant differences in the scores for patient care and communication (median 5.0, IQR 5.0-5.0 vs median 5.0, IQR 4.0-5.0; *P*=.06) and history taking (median 5.0, IQR 4.0-5.0 vs median 5.0, IQR 4.0-5.0; *P*>.99), respectively. The ICCs and their 95% CIs were –4.27e-16 (–0.281 to 0.449) in patient care and communication, 0.359 (–0.318 to 0.780) in history taking, –0.029 (–0.258 to 0.386) in physical examination, –0.093 (–0.220 to 0.275) in patient notes, 0.033 (–0.038 to 0.232) in clinical reasoning, 0.120 (–0.105 to 0.502) in management, and 0.000 (–0.072 to 0.203) in the total score.

Table 1. Comparison of the evaluation scores between physicians and ChatGPT-4 across 6 domains in the Objective Structured Clinical Examination. Scores were assigned on a 6-point Likert scale, ranging from 1 (very poor) to 6 (excellent), for each domain. Total scores were also calculated. All *P* values <.05 were considered statistically significant.

	ChatGPT-4, median (IQR)	Physicians, median (IQR)	*P* value[a]	ICC[b] (95% CI)
Patient care and communication	5.0 (5.0-5.0)	5.0 (4.0-5.0)	.06	–4.27e-16 (–0.281 to 0.449)
History taking	5.0 (4.0-5.0)	5.0 (4.0-5.0)	>.99	0.359 (–0.318 to 0.780)
Physical examination	4.0 (4.0-5.0)	4.0 (3.0-4.0)	.02	–0.029 (–0.258 to 0.386)
Patient notes	6.0 (5.0-6.0)	4.0 (4.0-4.0)	.002	–0.093 (–0.220 to 0.275)
Clinical reasoning	5.0 (5.0-5.0)	4.0 (3.0-4.0)	<.001	0.033 (–0.038 to 0.232)
Management	6.0 (5.0-6.0)	4.0 (2.5-4.5)	.002	0.120 (–0.105 to 0.502)
Column total	30 (29-31)	24 (20-26)	<.001	0.000 (–0.072 to 0.203)

[a]Statistical comparisons were conducted using the Wilcoxon signed-rank test.

[b]ICC: intraclass correlation coefficient.

Discussion

Principal Results

This experimental study compares evaluation scores between ChatGPT-4 and physicians in OSCE by using 6-item evaluation criteria. ChatGPT-4 assigned higher scores in the domains of physical examination, patient notes, clinical reasoning, and management, while no significant differences were found for history taking and patient care and communication. The low ICCs indicated poor agreement between ChatGPT-4 and physician evaluations in most domains. These results suggest that although ChatGPT-4 can generate structured assessments, its agreement with human raters is limited.

The higher ratings given by ChatGPT-4, particularly for physical examination skills, may be attributed to differences in the input material. Physicians evaluated student performance by using both video recordings and patient notes, allowing them to assess the visual and behavioral aspects of physical examination. In contrast, ChatGPT-4 evaluated students solely on interview transcripts and written notes. The absence of visual data likely contributed to “coarse-grained” or inflated evaluations by ChatGPT-4. Similarly, ChatGPT-4 also gave high ratings to the domains of patient notes, clinical reasoning, and management. As shown in Multimedia Appendix 1, structured prompts led ChatGPT-4 to assess these skills based solely on information from patient notes. In contrast, physicians may have evaluated deeper clinical consistency across sources, which could have contributed to stricter scoring. These discrepancies highlight the importance of input parity in comparative assessment studies.

We evaluated the agreement between ChatGPT-4 and physician evaluations by using ICCs as a measure of interrater reliability. The point estimates of ICCs were extremely low across most domains except for history taking (0.359, 95% CI –0.318 to 0.780). These findings indicate poor agreement overall and raise concerns about the construct validity of AI-based scoring in this context. The observed discrepancies might be attributed to some reasons, including the small sample size, differences in input modality between ChatGPT-4 and physicians, and the inherent limitations of AI in evaluating nonverbal or visual cues. Despite such limitations, the slightly higher ICC in history taking suggests that generative AI might show promise in evaluating domains that rely primarily on text-based information. Nonetheless, our results highlight the need for further validation before generative AI can be considered as a supplemental assessor for OSCE.

While preliminary, this study implies that generative AI may be cautiously explored as a complementary tool in OSCE scoring. Evaluations by AI could provide an additional perspective or be used as a screening before evaluations by physicians. Supported by generative AI, we could develop hybrid evaluation systems where generative AI and human evaluators collaborate to enhance the overall quality and reliability of the assessments.

Limitations

This study has several limitations. First, it was conducted in Japanese, which may limit the generalizability of the findings to settings involving other languages. Second, the sample size of medical students was small, which restricts the statistical power and internal validity of the results. For example, the

XSL•FO
RenderX

absence of significant score differences in history taking and patient care and communication may be attributable to insufficient power. Third, we used a single OSCE scenario involving abdominal pain, which limits generalizability across other clinical contexts. Although this scenario had not undergone external validation, it had been used in our department, suggesting a degree of internal content validity. Fourth, participation was voluntary, raising the possibility of selection bias. Fifth, we used ChatGPT-4 with ChatGPT Plus plan, and the results may not extend to other generative AI models or account types with different data privacy guarantees. Sixth, ChatGPT-4 and physicians had access to different input materials: while physicians evaluated both medical interview videos and patient notes, ChatGPT-4 evaluated only transcripts and notes. At the time of data collection (January 11, 2024), ChatGPT-4 did not support video input, necessitating the use of text-based inputs. This discrepancy likely influenced evaluations, particularly in domains such as physical examination and communication, which rely heavily on visual and nonverbal cues. To improve future study designs, it is essential to standardize input materials across evaluators. One possible approach would be a 3-arm comparison: (1) AI evaluating transcripts, (2) humans evaluating transcripts, and (3) humans evaluating videos. Although human-video evaluation is commonly treated as the gold standard in OSCE contexts, it too has limitations such as difficulty capturing subtle verbal or behavioral nuances. Acknowledging these constraints encourages a more cautious interpretation of comparative findings. Seventh, certain competencies evaluated in OSCE, particularly those that require visual or nonverbal confirmation (eg, physical examination techniques or nonverbal communication skills), are inherently unevaluable by text-only AI models. In our study, ChatGPT-4 evaluated all domains by using only transcripts and patient notes. Therefore, it was structurally incapable of detecting inconsistencies between documented notes and observed behaviors. This introduces a critical limitation for AI-based scoring of multimodal clinical skills. To address this issue, future studies should consider excluding domains that require multimodal inputs from AI-based assessments. Ensuring alignment between competency type (eg, verbal vs nonverbal) and AI capability will be critical to the appropriate use of generative AI in medical assessment. Eighth, we did not assess the reproducibility of ChatGPT-4's evaluations. The specific version of ChatGPT-4 used is no longer available via the same interface, preventing retesting under consistent conditions. As a result, the stability of the AI assessments remains unclear. Ninth, physician ratings were determined through a consensus approach, which prevents evaluation of their reliability.

Comparison With Prior Work

While there has been little evidence investigating the possibility of generative AI as an assessor in OSCE, some literature has evaluated the role of generative AI in simulated patient encounters. One report from the United States [18] demonstrated the implementation of ChatGPT-3.5 in training for breaking bad news, in which ChatGPT-3.5 created a scenario, performed as an SP, and gave immediate feedback to the trainee along with the SPIKES model. Another report by Misra and Suresh [12] demonstrated that ChatGPT generated feedback to the trainee based on the checklist completed by SPs in the OSCE scenario. In our study, we evaluated the potential of generative AI as a complementary assessor of OSCE by comparing assessments made by generative AI with those made by humans. The results suggest that generative AI can complement traditional evaluation methods and address some of the limitations associated with human assessments.

Future Direction

Future research should expand both the sample size and the diversity of medical students and OSCE scenarios to enhance the generalizability of the findings. To determine whether observed score differences are attributable to evaluator type rather than input modality, it is essential to standardize input materials between human and AI evaluators. Moreover, aligning the types of competencies assessed in OSCE (eg, verbal vs nonverbal) with the capabilities of generative AI is critical for valid evaluation. The reproducibility of generative AI assessments should also be examined by performing repeated runs with the same inputs. These analyses will help clarify the stability and limitations of the evaluations by generative AI. In addition, future studies should explore hybrid assessment models that integrate AI with human evaluators, potentially combining the efficiency of automation with the nuance of expert judgment. Generative AI may also be valuable as a self-assessment tool for medical students during self-directed learning. Although generative AI shows promise in supporting medical student evaluation, it is crucial to recognize its limitations and ensure that it complements, rather than replaces, the essential role of human judgment in medical education.

Conclusions

This experimental study shows that ChatGPT-4 assigned significantly higher ratings than physicians in 4 OSCE domains: physical examination, patient notes, clinical reasoning, and management. No significant differences were observed in the domains of history taking and patient care and communication. Although the results suggest that ChatGPT-4 can generate structured assessments, the low ICC values indicate that its agreement with physicians is limited. Further research with larger, more diverse cohorts and standardized input conditions is essential to fully evaluate the reproducibility, consistency, and construct validity of generative AI in medical education. Rather than replacing human evaluators, generative AI could play a complementary role in hybrid evaluation models, supporting fairness, scalability, and feedback generation in OSCE.

XSL•FO
RenderX

Acknowledgments

The authors declare the use of generative artificial intelligence (AI) in the research and writing process. According to the Generative AI Delegation Taxonomy (2025), the following tasks were delegated to generative AI tools under full human supervision: (1) text generation and (2) proofreading and editing. The generative AI tool used was ChatGPT-4o. The responsibility for the final manuscript lies entirely with the authors. Generative AI tools are not listed as authors and do not bear responsibility for the final outcomes. All conversations with ChatGPT used in the preparation of the manuscript are made available in Multimedia Appendix 2.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Funding

This work was supported by Japan Society for the Promotion of Science KAKENHI grant JP22K10471. The funder was not involved in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Authors' Contributions

MY had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.
Concept and design: MY, TH
Acquisition, analysis, or interpretation of data: MY, TH, T Sakamoto, RK, YS, YH
Drafting of the manuscript: MY, TH
Critical review of the manuscript for intellectual content: All authors
Statistical analysis: MY
Obtained funding: MY

Conflicts of Interest

None declared.

Multimedia Appendix 1
English version of the structured prompt.
[DOCX File , 32 KB - formative_v9i1e79465_app1.docx]

Multimedia Appendix 2
All ChatGPT conversations used in the preparation of the manuscript.
[DOCX File , 85 KB - formative_v9i1e79465_app2.docx]

References

1. Lee G, Kathleen A C. Goldman-Cecil Medicine, 2-Volume Set, 27th Edition. Netherlands: Elsevier; Aug 31, 2023.
2. Khan KZ, Ramachandran S, Gaunt K, Pushkar P. The Objective Structured Clinical Examination (OSCE): AMEE Guide No. 81. Part I: an historical and theoretical perspective. Med Teach 2013 Sep;35(9):e1437-e1446. [doi: 10.3109/0142159X.2013.818634] [Medline: 23968323]
3. Harden RM, Gleeson FA. Assessment of clinical competence using an objective structured clinical examination (OSCE). Med Educ 1979 Jan;13(1):41-54. [Medline: 763183]
4. Newble D. Techniques for measuring clinical competence: objective structured clinical examinations. Med Educ 2004 Mar;38(2):199-203. [doi: 10.1111/j.1365-2923.2004.01755.x] [Medline: 14871390]
5. Hijazi M, Downing S. Objective structured clinical examinations as an assessment method in residency training: practical considerations. Ann Saudi Med 2008;28(3):192-199 [FREE Full text] [doi: 10.5144/0256-4947.2008.192] [Medline: 18500173]
6. Bdair IAA, Abuzaineh HF, Burqan HMR. Advantages and disadvantages of the Objective Structured Clinical Examination OSCE in nursing education: a literature review. IJTSRD 2019 Feb 28;Volume-3(Issue-2):270-274. [doi: 10.31142/ijtsrd20269]
7. Pereira DSM, Falcão F, Nunes A, Santos N, Costa P, Pêgo JM. Designing and building OSCEBot ® for virtual OSCE - Performance evaluation. Med Educ Online 2023 Dec;28(1):2228550 [FREE Full text] [doi: 10.1080/10872981.2023.2228550] [Medline: 37347808]
8. Weaver SB, Daftary M, Wingate L, Turner M. Evaluation of faculty inter-variability OSCE grade scoring on overall student performance in a laboratory course. Pharm Educ 2022 Jan 08;22(1):48-53. [doi: 10.46542/pe.2022.221.4853]

XSL•FO
RenderX

9. Ataro G, Worku S, Asaminew T. Experience and challenges of Objective Structured Clinical Examination (OSCE): perspective of students and examiners in a clinical department of Ethiopian university. Ethiop J Health Sci 2020 May;30(3):417-426 [FREE Full text] [doi: 10.4314/ejhs.v30i3.13] [Medline: 32874085]
10. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
11. Preiksaitis C, Rose C. Opportunities, challenges, and future directions of generative artificial intelligence in medical education: scoping review. JMIR Med Educ 2023 Oct 20;9:e48785 [FREE Full text] [doi: 10.2196/48785] [Medline: 37862079]
12. Misra SM, Suresh S. Artificial intelligence and Objective Structured Clinical Examinations: using ChatGPT to revolutionize clinical skills assessment in medical education. J Med Educ Curric Dev 2024;11:23821205241263475 [FREE Full text] [doi: 10.1177/23821205241263475] [Medline: 39070287]
13. Le T, Bhushan V, Sheikh-Ali M, Shahin F. First Aid for the USMLE Step 2 CS, Fourth Edition. New York: McGraw-Hill Medical; 2012.
14. Medical OSCE public release materials (in Japanese). The Common Achievement Tests Organization. URL: https://www.cato.or.jp/cbt/medical-osce/index.html [accessed 2024-12-01]
15. Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman F, et al. Gpt-4 technical report. ArXiv Preprint posted online on March 4, 2024. [doi: 10.48550/arXiv.2303.08774]
16. Learning and evaluation items on skills and attitudes required at the start of clinical training: version 1.1 (in Japanese). The Common Achievement Tests Organization. URL: https://www.cato.or.jp/pdf/kenshu_11.pdf [accessed 2024-12-01]
17. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016 Jun;15(2):155-163 [FREE Full text] [doi: 10.1016/j.jcm.2016.02.012] [Medline: 27330520]
18. Webb JJ. Proof of concept: using ChatGPT to teach emergency physicians how to break bad news. Cureus 2023 May;15(5):e38755 [FREE Full text] [doi: 10.7759/cureus.38755] [Medline: 37303324]

Abbreviations

AI: artificial intelligence
ICC: intraclass correlation coefficient
OSCE: objective structured clinical examination
SP: standardized patient

Edited by J Sarvestan; submitted 21.06.25; peer-reviewed by H Mihara, SK Nandipati; comments to author 27.10.25; accepted 25.11.25; published 04.12.25.

Please cite as:
Yokose M, Hirosawa T, Sakamoto T, Kawamura R, Suzuki Y, Harada Y, Shimizu T
The Validity of Generative Artificial Intelligence in Evaluating Medical Students in Objective Structured Clinical Examination: Experimental Study
JMIR Form Res 2025;9:e79465
URL: https://formative.jmir.org/2025/1/e79465
doi:10.2196/79465
PMID:

©Masashi Yokose, Takanobu Hirosawa, Tetsu Sakamoto, Ren Kawamura, Yudai Suzuki, Yukinori Harada, Taro Shimizu. Originally published in JMIR Formative Research (https://formative.jmir.org), 04.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Inhibitory Control, Attentional Bias, and Palatable Food Consumption in Adolescents: A Laboratory Feasibility Randomized Controlled Trial

Carolina M Bejarano[1,2], PhD; Adrian Ortega[3,4], PhD; Christopher C Cushing[5], PhD

[1]Division of Behavioral Medicine and Clinical Psychology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, United States

[2]Department of Pediatrics, College of Medicine, University of Cincinnati, Cincinnati, OH, United States

[3]Division of Developmental and Behavioral Health and Center for Children's Healthy Lifestyle & Nutrition, Children's Mercy Hospital, Kansas City, MO, United States

[4]Department of Pediatrics, University of Missouri–Kansas City School of Medicine, Kansas City, MO, United States

[5]Department of Clinical Child Psychology & Schiefelbusch Institute for Life Span Studies, University of Kansas, Lawrence, KS, United States

Corresponding Author:
Carolina M Bejarano, PhD
Division of Behavioral Medicine and Clinical Psychology
Cincinnati Children's Hospital Medical Center
3333 Burnet Ave S2.343
Cincinnati, OH, 45229
United States
Phone: 1 513 517 1557
Email: carolina.bejarano@cchmc.org

Abstract

Background: Understanding relationships among inhibitory control, attentional bias to food cues, and food consumption in nonclinical adolescent samples can inform preventive efforts for disordered eating and associated health risks.

Objective: This pilot study conducted preliminary testing of an inhibitory control training intervention on attentional bias toward food cues, hedonic appetite, and food consumption, as well as examined inhibitory control's relationship with hedonic appetite, food consumption, binge eating, and cognitive restraint.

Methods: Participants (N=43; mean age 15.1, SD 1.7 years; 31/43, 72% female) were randomized to a food cue–specific go/no go inhibitory control training intervention or control group, took part in a laboratory "taste test," and completed surveys and a dietary recall. Additional data were collected for subsamples through an eye tracking task (15/43, 35%) and cognitive tasks (19/43, 44%).

Results: The go/no go intervention showed preliminary associations with attentional bias toward food cues. There were preliminary associations between inhibitory control and sugar consumption measured via dietary recall, binge eating symptoms, and cognitive restraint, which should be examined in a fully powered study.

Conclusions: The role of inhibitory control should be examined in larger studies to inform strategies for promotion of healthy eating behavior and prevention of disordered eating in adolescents.

(JMIR Form Res 2025;9:e77579) doi:10.2196/77579

KEYWORDS

hedonic appetite; adolescent; inhibitory control; attentional bias; palatable food

Introduction

Reward sensitivity and inhibitory control have shown to be relevant factors in presentations of binge eating, although the most consistent findings come from adult samples in the context of studying obesity [1]. Few studies have examined these mechanisms simultaneously or in nonclinical adolescent samples to understand the risk of developing maladaptive eating patterns such as binge eating. There is existing evidence suggesting that difficulties with inhibitory control and higher reward sensitivity have both been associated with binge eating in youth [2], but further research is needed. The neurobehavioral inhibitory control model of feeding [3] proposes that inhibitory control, a

XSL•FO
RenderX

dimension of executive function, interacts with hedonic appetite to influence palatable food consumption. In the context of dietary behavior, inhibitory control refers to inhibiting attention or a response toward a food cue [3]. In turn, hedonic appetite is a psychological and motivational process that refers to the drive to consume palatable foods for pleasure rather than for physiological sustenance [4]. In this model, attention allocation is thought to play a role in the dynamics between inhibitory control and hedonic appetite, with attentional bias to food cues referring to an increase in attention to food-related stimuli over other information [3,5]. A construct related to inhibitory control is cognitive restraint, which, in the context of dietary behavior, refers to efforts to avoid food intake. Cognitive restraint has been associated with higher cravings and pleasure derived from food [6].

Existing research has connected inhibitory control with hedonic appetite, palatable food consumption [7,8], dysregulated eating, and poor psychosocial outcomes in adults [9,10]. In adolescents, the literature has connected deficits in inhibitory control with overeating, risk of binge eating [11], palatable food consumption [12], negative affect [13], and hedonic appetite [14]. For example, functional magnetic resonance imaging data have indicated that adolescents with higher hedonic appetite have lower neural activation in the prefrontal cortex, suggesting that experiencing higher hedonic appetite may be related to lower inhibition of impulses to consume energy-dense palatable food [14]. In general, youth in the adolescent age group (13-18 years) have a higher tendency than other age groups to be high in reward seeking, vulnerable to developing disordered or maladaptive eating patterns, and at high risk of metabolic health concerns related to diet [15,16]. Moreover, adolescence is a key developmental stage to study drivers of dietary behavior as youth in this age group have increasing independence with their dietary choices and often consume a diet high in palatable foods that are lower in nutritional quality [17-19]. Therefore, adolescence presents a unique and important opportunity to understand and intervene on dietary processes that are carried into adulthood and impact long-term health.

Evidence-based methods used to study cognitive processes in the context of dietary behavior include reaction time paradigms to assess and intervene on inhibitory control [20] and eye tracking tasks to measure attentional bias toward food cues [21]. Food-specific inhibitory control training (ICT) tasks have been used to influence food choices [22]. A go/no go (GNG) ICT task is a procedure in which 2 categories of visual stimuli are randomly presented and a respondent is required to respond to images from one stimuli category (*go*) while withholding a response to an alternative stimuli category (*no go*) [23]. There is a majority of evidence suggesting that GNG training creates automatic stop associations that can influence outcomes of food choice, portion size, and weight changes [24] by training an individual to inhibit a response to food cues [22]. The few studies on GNG training in youth have found relationships with adolescents' inhibitory control for intake of sugar-sweetened beverages [12] and reduction in consumption of energy-dense candy in children when measured in laboratory settings [25]. One pilot study found that a virtual reality ICT resulted in reduced binge eating up to the 2-week follow-up in adults [26]. While a GNG ICT task is often used as an intervention, a flanker task is an executive function measure that is used as an assessment of inhibitory control. The flanker task assesses a participant's ability to allocate cognitive resources to process stimuli while also ignoring extraneous stimuli [27]. Finally, gaze fixation tasks with an eye tracker are often used to measure attentional bias toward food cues by measuring overall gaze direction and duration bias [28]. In adolescents, studies using gaze fixation with eye tracking data have found greater gaze duration bias toward food cues in a group with binge eating as compared to the control group [29].

The literature to date indicates that the field lacks studies that examine associations of inhibitory control in the context of dietary behavior in nonclinical samples of adolescents through both objective laboratory-based and subjective self-report outcomes [2]. An enhanced understanding of processes impacting dietary behavior in adolescents can be harnessed to inform preventive interventions geared toward improving health and well-being in this age group as well as in the long term into adulthood. The limited previous research to date has found that inhibitory control has been inversely related to facets of binge eating in community samples of adolescents and reward sensitivity has been related to binge eating behaviors in youth [2]. Together with the theoretical underpinnings of the neurobehavioral inhibitory control model of feeding [3] and the background literature described previously, this study was designed to examine the preliminary effects and pilot methodology of a randomized laboratory study of ICT and examine mechanisms of inhibitory control, attentional bias to food cues, hedonic appetite, palatable food consumption, and binge eating in a nonclinical sample of adolescents. The following hypotheses were tested: (1) the food-specific GNG task will show preliminary associations with reductions in attention to palatable food cues, hedonic appetite, palatable food consumption in a laboratory-based eating paradigm, and food intake via dietary recall; and (2) inhibitory control will show preliminary inverse associations with hedonic appetite, palatable food consumption (in a laboratory-based paradigm and via dietary recall), and binge eating and preliminary positive associations with cognitive restraint.

Methods

Participants

A total of 43 adolescent participants aged 13 to 18 years were recruited through the local community for a study about adolescent health behaviors. To best control for the food environment across participants and study dietary behavior in typically functioning adolescents, the inclusion criteria were living in a home setting with their parents or caregivers, the ability to read at their grade level and in English, and absence of blindness or low vision. The exclusion criteria were presence of developmental delays, past or current self-reported eating disorder diagnoses, and any health conditions or dietary restrictions that significantly affected activity or eating habits (eg, celiac disease or vegan diet).

XSL•FO
RenderX

Procedure

Overview

The consent or assent process was completed at the laboratory visit, and it was communicated to parents or caregivers and adolescent participants that they were not fully informed of the processes being examined in the study but that they would receive more detailed information about the study after participating. Specifically, they were not informed about the study's focus on dietary behavior or covert measurement of consumption through the bogus taste test until they completed the study. Adolescent participants also completed a survey indicating their palatable food preferences for the laboratory visit (ie, choosing one of each of the following: gummy bears or M&Ms in the sweet category, rolls or cereal in the starchy category, fried chicken or hot dogs in the fatty category, and potato chips and cheese pizza in the fast food category [30]). They were instructed to fast for 3 hours before the laboratory visit. The procedures described in this paper focus on the laboratory visit, which was the main component of a larger study that encompassed a 10-day protocol with evening surveys to assess within-person dietary processes [31]. Relevant to this study, on the 10th day, participants also completed a dietary recall.

Laboratory Visit

Participants were provided a standard meal replacement shake (350 calories with composition of 15% protein, 57% carbohydrates, and 29% fats [32]) at the beginning of the laboratory procedures in efforts to ensure that observed eating behavior was due to hedonic appetite rather than physiological hunger. Participants were then randomized to the intervention (food-specific) or control GNG task designed on the PsychoPy software (Open Science Tools Ltd) [33] using a computerized random assignment. Following the GNG task, procedures took place in the following order at the laboratory visit: completion of hedonic appetite self-report measure, bogus taste test, self-report survey battery, eye tracking task, Automated Self-Administered 24-Hour Dietary Assessment Tool (ASA24) dietary recall, National Institutes of Health (NIH) Toolbox cognitive tasks, and anthropometric measurements. The research assistant responsible for assessing the outcome in the bogus taste test remained masked to the randomization condition so as not to introduce bias in outcome measurement. The research assistant administering the GNG task and eye tracking task was aware of the randomization group to administer the designated task. All study procedures were conducted with scripted instructions used by the research staff to administer measures and tasks. The approach of administering the eye tracking and NIH Toolbox tasks to only a subset of participants was taken to pilot the protocols within the laboratory visit procedures.

ICT Task

The GNG procedures were designed in accordance with previous studies [25,34], including a practice round of the task to ensure that participants could perform it correctly. The GNG task included instructions to press a button when a *go* cue was presented and to not press a button when a *no go* cue was presented. Participants in the food-specific GNG intervention condition were trained using pictures of palatable foods grouped as sweet, starchy, fatty, and fast food as *no go* cues, with 4 food stimuli pictures used for each category (16 in total) and 16 nonfood stimuli pictures of animals as *go* cues. Those in the control GNG condition completed the GNG task with all nonfood stimuli (eg, animals and shapes; Figure 1). The images in the GNG tasks were used with approval from the Food-pics image database designed to facilitate standardization and comparability across studies [35] (Multimedia Appendix 1). For both groups, each stimulus was presented for 1000 ms, and the intertrial interval was 500 ms. If the participant made an incorrect response or did not respond within 1000 ms, a red cross appeared on the screen for 500 ms (Figure 1). For both groups, 50% of the trials were *go* trials (ie, 16 images of animals presented 4 times each), and 50% were *no go* trials.

XSL•FO
RenderX

Figure 1. Examples of stimuli in the (A) food go/no go (GNG) task (intervention group) and (B) control GNG task.

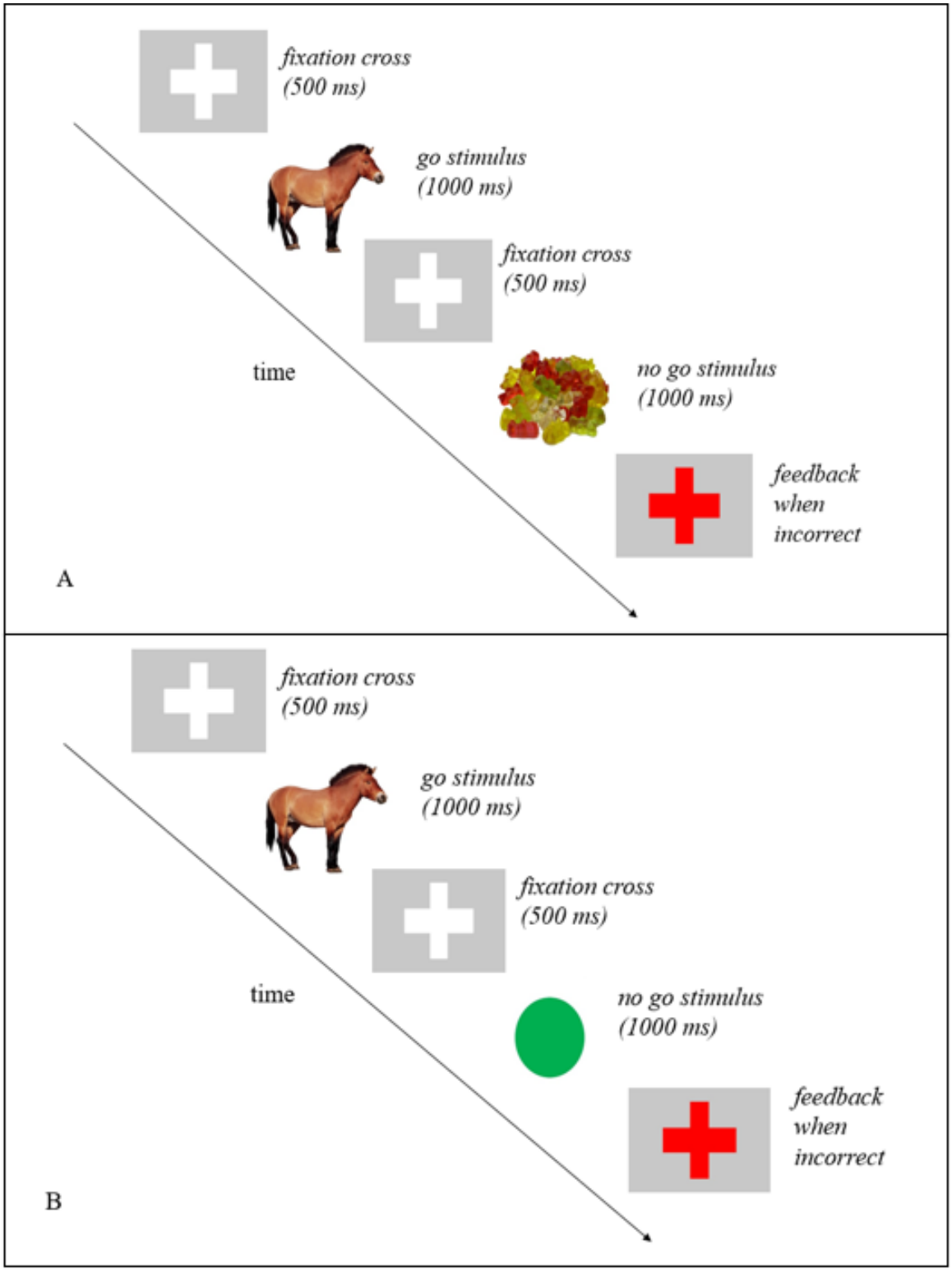

Bogus Taste Test

After completing the GNG training, participants completed a hedonic appetite survey measure and then were brought to a separate room with a comfortable chair, an iPad playing a television show, and large amounts of the preferred palatable foods. Participants were allowed 20 minutes to complete the taste test and instructed to eat as much as they liked. The amount of each type of food eaten by each participant was weighed covertly by the masked research assistant to calculate the amount of palatable food consumed.

Eye Tracker and Remaining Tasks

A gaze fixation task with the use of an eye tracker was administered on a desktop computer to a subset of the study participants to determine attentional bias for food versus nonfood images. Participants then each completed self-report questionnaires, a dietary recall, and cognitive tasks and had

XSL•FO
RenderX

their height and weight recorded. After the last day of study procedures, participants and parents were mailed their gift card compensation along with a debriefing letter outlining the research questions of interest.

Constructs and Measures

Demographics and Anthropometrics

Participants completed a demographic questionnaire at the laboratory visit. Each participant's height and weight were measured at the end of the laboratory visit following standard protocols. BMI percentile was calculated based on age and sex as indicated by the Centers for Disease Control and Prevention [36].

Food Preferences

The Food Preference Questionnaire was designed specifically for this study based on factor loadings from the Food Craving Inventory (ie, sweet, starchy, fatty, and fast food [30]) and on studies examining consumption of palatable food in a laboratory setting [7]. In this sample, the Food Craving Inventory demonstrated good internal consistency reliability (Cronbach α=0.93 for the total, 0.86 for sweet, 0.84 for starchy, 0.86 for fatty, and 0.76 for fast food). Participants completed this before the laboratory visit to determine which foods would be available during their bogus taste test.

Hedonic Appetite

The Power of Food Scale (PFS) is a 15-item measure that assesses the construct of hedonic appetite [37,38] using a 5-point Likert scale. Higher mean scores indicate higher hedonic appetite (range 1-5). The PFS has shown high internal consistency reliability in a previous sample of healthy adolescents (Cronbach α=0.94 [31]). The PFS was completed after completion of the GNG task.

Palatable Food Consumption

The amounts of the 4 foods offered in the bogus taste test were standardized based on caloric content. More specifically, research staff were trained to reliably measure and serve 500 to 600 calories of each food offered in the study. After the bogus taste test, the remaining portions of each food were weighed covertly on a food scale in grams, and the amount consumed was converted to calories.

Typical Dietary Intake

The ASA24 is a freely available web-based tool developed by the US National Cancer Institute [39] to obtain high-quality daily dietary data with minimal bias [40]. The 1-day dietary recall is considered acceptable as a valid assessment of typical dietary intake [40]. Two administrations of the ASA24 dietary recall took place: one at the laboratory visit and one 3 days after the laboratory visit.

Attentional Bias to Food Cues

Attentional bias was measured through administration of the gaze fixation task with the use of an eye tracker set up on a desktop computer. The amount of time participants spent looking at food versus nonfood images (referred to as dwell time bias or gaze duration bias) was used as a measure of attentional bias to food cues in the GNG intervention group as compared to the GNG control group. This measure of gaze fixation as attentional bias has shown excellent internal reliability and acceptable test-retest reliability in eye tracking procedures of this nature [41]. The eye tracking task was administered near the end of the laboratory visit procedures (ie, after the GNG task, bogus taste test, and laboratory visit surveys).

Eating Pathology

The Binge Eating and Cognitive Restraint subscales of the Eating Pathology Symptoms Inventory (EPSI) were used to measure disordered eating constructs of interest for this study. The EPSI is a well-validated 45-item self-report measure with 8 total subscales that has participants report on eating pathology in the previous 4 weeks [42] on a 5-point Likert scale. The binge eating subscale consists of 8 items, and scores range from 4 to 40. The cognitive restraint subscale measures cognitive efforts to limit or avoid eating and consists of 3 items, and scores range from 3 to 15. The binge eating subscale has shown excellent internal consistency in the sample of adolescents in this study [43], and the cognitive restraint subscale has shown good internal consistency reliability in a sample of adolescents previously [44].

Inhibitory Control

The NIH Toolbox iPad app provides access to the NIH Toolbox for Assessment of Neurological and Behavioral Function, a standard set of valid, reliable, and royalty-free tools for assessing cognitive, emotional, motor, and sensory function in participants aged 3 to 85 years. For this study, we used the NIH Toolbox Cognition: Executive Function model for assessing inhibitory control [45]. Participants completed a flanker task (Flanker Inhibitory Control and Attention Test for ages of ≥12 years; version 2.0) using the NIH Toolbox app on the iPad provided by the laboratory, which takes approximately 3 minutes to administer. The task has participants indicate the direction of an arrow that is flanked by 2 other arrows that point in either congruent or incongruent directions. The incongruent trials provide a measure of inhibitory control based on accuracy and speed, which is automatically calculated by the NIH program software. This task has shown to be a valid measure of inhibitory control and attention in youth [46].

Data Analysis

Preliminary Analyses

Bivariate correlations and independent-sample 2-tailed *t* tests were run to examine associations between potential covariates (eg, eating disorder symptomology and BMI percentile) and the dependent variables. Significant correlations and associations were then examined further using independent-sample *t* tests to examine potential differences between the participants randomized to the food GNG intervention group and the control GNG group, with each potential covariate as a dependent variable. In particular, the dietary intake data obtained through the dietary recall were used as a randomization check to determine whether the 2 groups ate significantly different amounts of total calories over the 24-hour period before the study visit. Information regarding participants' mental health and medication use that could potentially affect appetite and

XSL•FO
RenderX

eating behavior ere also examined as potential covariates. Information was coded as having a potential influence on dietary behavior or not, and chi-square analyses were run to determine any differences between the randomization groups.

Effect of GNG Intervention

A series of independent-sample *t* tests were run to determine the effects of the GNG intervention versus the GNG control on gaze fixation on food images as measured using the eye tracker. Effects on hedonic appetite, palatable food consumption in the laboratory bogus taste test, and food consumption via dietary recall (second administration 3 days following the laboratory visit) were also examined.

Associations Between Inhibitory Control and Dietary Constructs

Regression models were run to examine cross-sectional relationships between inhibitory control as measured using the NIH Toolbox and the dietary constructs of interest (hedonic appetite, palatable food consumption in the laboratory, food consumption via dietary recall [first administration at the laboratory visit], and EPSI binge eating and cognitive restraint subscales).

Ethical Considerations

All procedures followed in this study were in accordance with the ethical standards of the institutional review board, the 1964 Declaration of Helsinki and its later amendments or comparable ethical standards. This study was reviewed and approved by the institutional review board at the University of Kansas (STUDY00141427). Caregivers provided informed consent before data collection. No identifiable features of research participants are included in the data. Participants and caregivers were informed that participation was voluntary and they could discontinue at any time and were offered up to US $40 in compensation and snacks.

Results

Descriptive Information

A total of 43 participants were enrolled in the study. Table 1 shows demographic information and descriptive statistics, and Figure 2 shows a flowchart of participant inclusion. The subset of participants who completed the gaze fixation task consisted of 15 adolescents (n=6, 40% from the GNG intervention group and n=9, 60% from the control group), and the subset of participants who completed the NIH Toolbox task consisted of 19 adolescents (n=10, 53% from the intervention group and n=9, 47% from the control group). Participants had a mean age of 15.1 (SD 1.7) years, most were female (n=31, 72%) and White (n=25, 58%), and 49% (21/43) reported an annual household income of >US $60,000. The average BMI percentile was 50.2 (SD 28.7), and the average calories consumed the day before the study as measured through the dietary recall was 1798.7 (SD 672.3). Additional data about adolescent mental health and medication history were obtained from 81% (35/43) of the participant-parent dyads. Among them, 17% (6/35) reported current or past history of anxiety, 11% (4/35) reported current or past history of depression, 14% (5/35) reported current or past history of attention-deficit/hyperactivity disorder, 6% (2/35) reported current or past history of other concerns, and 0% reported current or past history of eating disorders. The same dyads reported current medications for the adolescents: 11% (4/35) reported taking medication for anxiety or depression, 14% (5/35) reported taking acne medication, 9% (3/35) reported taking attention-deficit/hyperactivity disorder medication, 3% (1/35) reported taking allergy medication, 3% (1/35) reported taking thyroid medication, and 9% (3/35) reported taking other unspecified medication. For the taste test, in the sweet category, 54% (23/43) of the participants chose M&Ms, and 47% (20/43) chose gummy bears; in the starchy category, 47% (20/43) chose cereal, and 54% (23/43) chose rolls; in the fatty category, 81% (35/43) chose fried chicken, and 19% (8/43) chose hot dogs; and, in the fast food category, 30% (13/43) chose potato chips, and 70% (30/43) chose pizza. The average consumption in the laboratory taste test was 91.9 (SD 72.1) sweet calories, 79.9 (SD 95.6) starchy calories, 140.3 (SD 99.5) fatty calories, 178.9 (SD 143.5) fast food calories, and 491.0 (SD 249.9) total calories. The results of the randomization check indicated that the control group consumed an average of 1742.9 (SD 724.2) calories and the intervention group consumed an average of 1831.2 (SD 633.8) calories in the first administration of the dietary recall (P=.79; nonsignificant difference). The results of the chi-square analyses indicated that there were no substantial differences between the randomized groups and mental health conditions or medication use that could influence dietary behavior. The results of covariate screening indicated that the randomization process functioned effectively in controlling for any demographic variables associated with the dependent variables.

XSL•FO
RenderX

Table 1. Demographic information and descriptive statistics (N=43)[a].

Demographic variable	Total	Intervention (n=21)	Control (n=22)
Sex, n/N (%)			
Female	31/43 (72)	15/21 (71)	16/22 (73)
Male	12/43 (28)	6/21 (29)	6/22 (27)
Race or ethnicity, n/N (%)			
Asian	8/43 (19)	5/21 (24)	3/22 (14)
Biracial or multiracial	4/43 (9)	3/21 (14)	1/22 (5)
Hispanic or Latino	4/43 (9)	1/21 (5)	3/22 (14)
White	25/43 (58)	11/21 (52)	14/22 (64)
Other or did not report	2/43 (5)	1/21 (5)	1/22 (5)
Approximate annual family income (US $), n/N (%)			
<30,000	11/39 (28)	6/19 (32)	5/20 (25)
30,000-60,000	7/39 (18)	2/19 (11)	5/20 (25)
>60,000	21/39 (54)	11/19 (58)	10/20 (50)
Age (y), mean (SD)	15.1 (2)	15.0 (1.6)	15.2 (2)
BMI percentile[b], mean (SD)	50.2 (29)	52.6 (30.8)	47.7 (27.0)
Calories consumed in dietary recall 1, mean (SD)	1798.7 (672.3)	1831.2 (633.8)	1742.9 (724.2)
Calories consumed in dietary recall 2, mean (SD)	1551.5 (697.9)	1650.9 (746.1)	1452.0 (652.2)
Sugar calories consumed in dietary recall 1, mean (SD)	99.8 (64.7)	108.9 (60.6)	91.2 (68.9)
Sugar calories consumed in dietary recall 2, mean (SD)	72.6 (39.0)	78.5 (41.6)	66.8 (36.5)
Hedonic appetite (PFS[c] score; 1-3.9), mean (SD)	1.6 (0.69)	1.6 (0.74)	1.6 (0.66)
Binge eating (EPSI[d] score; 4-40), mean (SD)	18.4 (8.0)	19.3 (8.8)	17.5 (7.3)
Cognitive restraint (EPSI score; 3-15), mean (SD)	7.0 (2.9)	6.9 (2.1)	7.1 (3.6)
Inhibitory control (NIH[e] Toolbox), mean (SD)	83.6 (10.7)	79.2 (8.8)	87.6 (11.1)
Sweet calories consumed (laboratory), mean (SD)	91.9 (72.1)	82.4 (79.2)	101.0 (65.2)
Starchy calories consumed (laboratory), mean (SD)	79.9 (95.6)	102.8 (110.4)	58.1 (75.1)
Fatty calories consumed (laboratory), mean (SD)	140.3 (99.5)	140.9 (88.8)	139.8 (110.8)
Fast food calories consumed (laboratory), mean (SD)	178.9 (143.5)	161.5 (132.5)	195.4 (154.5)
Total calories consumed (laboratory), mean (SD)	491.0 (249.9)	487.6 (247.8)	494.3 (257.8)

[a]A total of 84% (36/43) of the participants completed dietary recalls at both time points, and 9% (4/43) of the participants did not report approximate family income.

[b]In total, 85% (34/40) of the sample had a BMI percentile within the range categorized as normal weight, 10% (4/40) had a BMI percentile in the range categorized as overweight, and 5% (2/40) had a BMI percentile in the range categorized as obesity. BMI percentile data were missing for 3 participants.

[c]PFS: Power of Food Scale.

[d]EPSI: Eating Pathology Symptoms Inventory.

[e]NIH: National Institutes of Health. A total of 19 adolescents (n=9, 47% from the intervention group and n=10, 53% from the control group) completed the NIH Toolbox measurement of inhibitory control (age-corrected standard score for inhibitory control range 63-111).

XSL•FO
RenderX

Figure 2. Flowchart of participant inclusion.

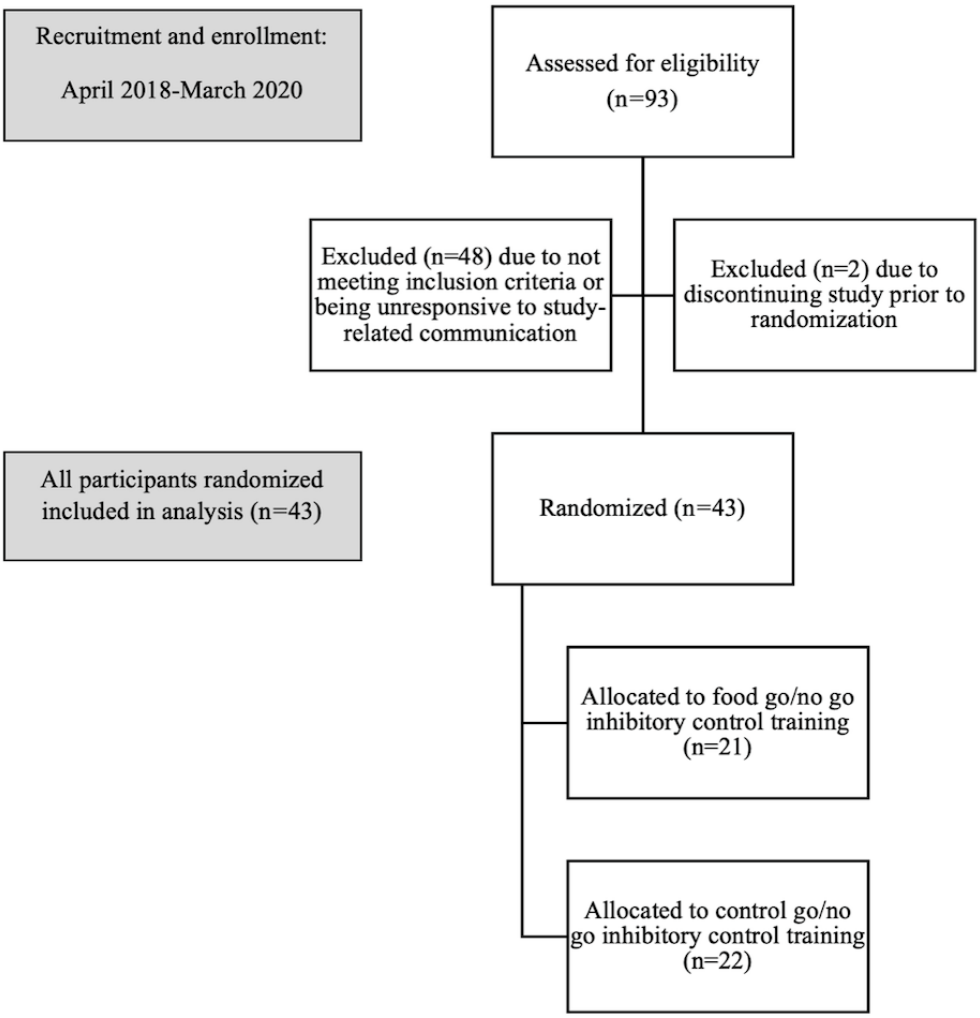

Results of the GNG Intervention

In the small subsample of participants who completed the gaze fixation task with eye tracking (15/43, 35%), participants from the intervention condition looked at food images for 20.85 (SD 10.25) seconds and participants from the control group looked at food images for 23.01 (SD 8.39) seconds. Therefore, the participants from the intervention condition looked at food images for 2.1 seconds less than control participants in absolute terms (95% CI −8.3 to 12.6). Expressed as an effect size, this is considered a small effect (Cohen d=0.23). There were no other significant differences in hedonic appetite or food consumption via the bogus taste test or dietary recall following the laboratory visit. Given the limited subsample, these are considered preliminary findings that should be examined in a fully powered study.

Associations Between Inhibitory Control and Dietary Constructs

In the small subsample of participants who completed the NIH Toolbox flanker task to measure inhibitory control (19/43, 44%), inhibitory control was significantly and positively associated with sugar intake as measured through the laboratory-based administration of the dietary recall (F_{10}=18.5; P=.006), with an effect size (η^2) of 0.979 (95% CI 0.43-0.97). Inhibitory control was also marginally related to binge eating (F_{13}=4.53; P=.05; η^2=0.922, 95% CI 0.00-0.89) and cognitive restraint (F_{13}=4.05; P=.06; η^2=0.913, 95% CI 0.00-0.87) as measured using the EPSI. There were no significant associations or trends between inhibitory control and hedonic appetite or food consumption at the laboratory visit. It must be noted that this is a limited subsample and the associations presented do not meet conventional significance thresholds. These findings represent preliminary patterns that should be investigated in fully powered studies.

Discussion

Principal Findings

This study found that the food-specific GNG ICT intervention had preliminary associations with attentional bias toward food cues as measured through an eye tracking gaze fixation task compared to a control GNG ICT task. Additionally, inhibitory control as measured using the NIH Toolbox was significantly associated with sugar intake as measured using the dietary recall completed at the laboratory visit and showed marginally significant cross-sectional associations with binge eating and cognitive restraint. Our findings contribute to the literature examining the mechanisms through which ICT functions. The understanding of these mechanisms is still evolving within the

XSL•FO
RenderX

field, with a few possible pathways, including the devaluation processes [47]. This means that attention is often allocated toward rewarding palatable food stimuli; therefore, decreasing the reward value of palatable food stimuli is a *possible* mechanism that could be at the core of these processes. Still, attentional bias has had mixed associations with palatable food intake thus far [5,48,49]. Further research is needed to examine the role of attention allocation in predicting eating behavior in adolescents and whether ICT is the most effective intervention.

Contrary to the hypotheses, the GNG intervention did not have any effect on hedonic appetite or food consumption measured either in the laboratory or via the dietary recall, and inhibitory control was not related to hedonic appetite or food consumption in the laboratory visit. While this study was indeed limited by the small sample size, we note that there are some mixed associations between inhibitory control and eating behavior in adolescents in the literature, with some studies showing no association [50] and others showing low inhibitory control as a risk factor for developing disordered eating [51]. While a large portion of the literature has focused on inhibitory control–binge eating relationships in the context of obesity, inhibitory control difficulties have been linked to loss-of-control eating in community samples of adolescents as well [2]. Interestingly, in our sample, inhibitory control was related to sugar consumption and binge eating in the opposite direction than hypothesized. In this small sample, higher inhibitory control was associated with higher sugar consumption and risk of binge eating. Considering this unexpected association and the fact that inhibitory control was also marginally associated with cognitive restraint, the mechanistic role of inhibitory control should be examined in fully powered studies aligned with the neurobehavioral inhibitory control model [7].

Study Strengths and Limitations

This study's strengths lie in the use of objective and gold-standard measurements, including the eye tracking gaze fixation task for attentional bias toward food cues, dietary recalls, objective measurement of palatable food consumption in the laboratory visit "taste test," and the NIH Toolbox, methods that could all be used in a future fully powered study. A main limitation of this study is the small sample size, which was an unanticipated change due to the COVID-19 pandemic that resulted in discontinuing data collection in March 2020. This study is also limited by administration of the eye tracking task and the NIH Toolbox in subsamples of our total sample. This was done to pilot each methodology with the adolescent age group and as part of the laboratory visit procedures overall. Because only a subsample completed the eye tracking and NIH Toolbox tasks, we cannot make any claims about external validity. It must be noted that these small subsamples limit power and generalizability and cannot be taken as definitive. The findings should be considered preliminary, and these procedures should be conducted in fully powered studies as a next step. Additionally, participants in this study were predominantly White, female, and from higher-income households. Recruiting more demographically diverse individuals is a necessary focus for future research in this area, especially as systemic inequities based on race and ethnicity, socioeconomic status (eg, food insecurity), and cultural factors influence eating behavior and overall health [52]. The lack of diversity in this study also limits the applicability of the results to broader adolescent populations, which can be improved upon in larger, fully powered, and representative studies.

Conclusions and Future Directions

Considering the need for real-time and objective measurement of appetite and eating, studies in which the data capture, intervention, and assessment occur in real time may be better suited for continued investigations. Moreover, theoretical models of dietary behavior can potentially be enhanced by expanding to include dynamic processes. For example, dynamically assessed motivational processes and behavioral outputs of the neurobehavioral inhibitory control model of feeding, such as inhibitory control, hedonic appetite, delay discounting, selective attention, and feeding behavior [3], could be included in a model as there is growing evidence that all of them can be measured successfully and show significant associations when measured via ecological momentary assessment (EMA). Knowing that individual differences (eg, motivation, weight status, and intentions to restrict intake) affect both dietary behavior [53] and the outcome of GNG interventions [54], capturing these processes more fully via EMA appears to be a promising avenue [55]. Studies focused on the adolescent age group specifically should also include adolescent-specific factors that impact dietary behavior, such as the social influences of peer groups and cognitive factors in this developmental period of increasing autonomy [53]. In conjunction with this, some studies have used internet platforms [34], video games [56], and virtual reality [26] as methods of administering GNG interventions, which may be promising for future EMA protocols especially targeting adolescents. Although our larger study included evening surveys to examine within-person processes [31], it was still not fully powered to examine the effects of the GNG intervention on fluctuating mechanisms. Rigorous and fully powered study designs that capture additional drivers of dietary behavior in real time as they fluctuate, such as access to food across environments, social cues, and attention toward food cues [3], will also improve the potential impact and application of the findings. Additionally, in this study, we piloted a single-session GNG ICT intervention, whereas other potential avenues for this research could include a longer duration or multiple sessions, alternate methods of capturing state change, and other conceptualizations of how to intervene on hedonic appetite examined longitudinally.

The role of inhibitory control in models of dysregulated eating and weight-related health continues to be a focus of investigation [13]. Although no claims about clinical utility can be made from our pilot study, recent work from other teams can inform next steps for research in this area. For example, Manasse et al [26] have found preliminary effects of a daily administered virtual reality ICT intervention on loss-of-control eating. Leaders in this literature acknowledge that clinical interventions for dysregulated eating conditions often result in limited improvement due to inadequately addressing inhibitory control and factors driving binge eating [57]. Given its evidence of reducing binge eating [58], continued application of ICT is warranted to determine effectiveness and understand mechanisms, but additional intervention avenues must also be

XSL•FO
RenderX

explored. A recent systematic review of inhibitory control, loss-of-control eating, and reward sensitivity in children and adolescents also called for more research examining cognitive mechanisms longitudinally in youth, especially through use of EMA [2]. There is potential for cognitive interventions used in the key developmental period of adolescence to prevent the maladaptive dietary processes and problematic eating behaviors that often underlie both obesity risk and eating psychopathology [59]. Additional interventions targeting intuitive eating [60] and mindful decision-making [61], improvement of state and trait mindfulness [62], and understanding of the role of affect regulation [63] to promote healthful eating behavior have also shown promise for prevention of disordered eating, and their comparative value should be examined in adolescent samples.

Acknowledgments

The authors thank Rathi Narayan for her assistance in the Teen HEALTH study. This study was funded by the Society of Pediatric Psychology Marion and Donald Routh Student Research Grant and the University of Kansas Office of Graduate Studies Doctoral Student Research Fund awarded to CMB under the mentorship of CCC. AO was supported by a grant from the National Institutes of Health (T32 MH115882). The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Data Availability

Deidentified data from this study are not available in a public archive and will be made available (as permissible according to institutional review board standards) upon reaching out to the corresponding author.

Authors' Contributions

CMB and CCC were responsible for the conceptualization and design of this study. CMB, AO, and CCC were responsible for the acquisition and analysis of the data. CMB drafted the manuscript. CMB, AO, and CCC critically revised the manuscript for intellectual content. CMB and AO provided administrative and technical support. CCC supervised the study.

Conflicts of Interest

CCC has received consulting funds and research support from PepsiCo.

Multimedia Appendix 1
Images used in go/no go task.
[DOCX File , 13 KB - formative_v9i1e77579_app1.docx]

Multimedia Appendix 2
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 1060 KB - formative_v9i1e77579_app2.pdf]

References

1. Pasquale EK, Boyar AM, Boutelle KN. Reward and inhibitory control as mechanisms and treatment targets for binge eating disorder. Curr Psychiatry Rep 2024 Nov;26(11):616-625 [FREE Full text] [doi: 10.1007/s11920-024-01534-z] [Medline: 39316228]
2. Ramalho SM, Conceição E, Tavares AC, Freitas AL, Machado BC, Gonçalves S. Loss of control over eating, inhibitory control, and reward sensitivity in children and adolescents: a systematic review. Nutrients 2023 Jun 08;15(12):2673 [FREE Full text] [doi: 10.3390/nu15122673] [Medline: 37375576]
3. Appelhans BM. Neurobehavioral inhibition of reward-driven feeding: implications for dieting and obesity. Obesity (Silver Spring) 2009 Apr;17(4):640-647 [FREE Full text] [doi: 10.1038/oby.2008.638] [Medline: 19165160]
4. Lowe MR, Butryn ML. Hedonic hunger: a new dimension of appetite? Physiol Behav 2007 Jul 24;91(4):432-439. [doi: 10.1016/j.physbeh.2007.04.006] [Medline: 17531274]
5. Hardman CA, Rogers PJ, Etchells KA, Houstoun KV, Munafò MR. The effects of food-related attentional bias training on appetite and food intake. Appetite 2013 Dec;71:295-300 [FREE Full text] [doi: 10.1016/j.appet.2013.08.021] [Medline: 24025548]
6. Racine SE. Emotional ratings of high- and low-calorie food are differentially associated with cognitive restraint and dietary restriction. Appetite 2018 Feb 01;121:302-308 [FREE Full text] [doi: 10.1016/j.appet.2017.11.104] [Medline: 29175441]
7. Appelhans BM, Woolf K, Pagoto SL, Schneider KL, Whited MC, Liebman R. Inhibiting food reward: delay discounting, food reward sensitivity, and palatable food intake in overweight and obese women. Obesity (Silver Spring) 2011 Nov;19(11):2175-2182 [FREE Full text] [doi: 10.1038/oby.2011.57] [Medline: 21475139]

XSL•FO
RenderX

8. Manasse SM, Espel HM, Forman EM, Ruocco AC, Juarascio AS, Butryn ML, et al. The independent and interacting effects of hedonic hunger and executive function on binge eating. Appetite 2015 Jun;89:16-21 [FREE Full text] [doi: 10.1016/j.appet.2015.01.013] [Medline: 25613129]
9. Lavagnino L, Arnone D, Cao B, Soares JC, Selvaraj S. Inhibitory control in obesity and binge eating disorder: a systematic review and meta-analysis of neurocognitive and neuroimaging studies. Neurosci Biobehav Rev 2016 Sep;68:714-726. [doi: 10.1016/j.neubiorev.2016.06.041] [Medline: 27381956]
10. Lowe MR, Arigo D, Butryn ML, Gilbert JR, Sarwer D, Stice E. Hedonic hunger prospectively predicts onset and maintenance of loss of control eating among college women. Health Psychol 2016 Mar;35(3):238-244 [FREE Full text] [doi: 10.1037/hea0000291] [Medline: 26690638]
11. Gowey MA, Lim CS, Dutton GR, Silverstein JH, Dumont-Driscoll MC, Janicke DM. Executive function and dysregulated eating behaviors in pediatric obesity. J Pediatr Psychol 2018 Sep 01;43(8):834-845 [FREE Full text] [doi: 10.1093/jpepsy/jsx091] [Medline: 28595362]
12. Ames SL, Kisbu-Sakarya Y, Reynolds KD, Boyle S, Cappelli C, Cox MG, et al. Inhibitory control effects in adolescent binge eating and consumption of sugar-sweetened beverages and snacks. Appetite 2014 Oct;81:180-192 [FREE Full text] [doi: 10.1016/j.appet.2014.06.013] [Medline: 24949566]
13. Byrne ME, Shank LM, Altman DR, Swanson TN, Ramirez E, Moore NA, et al. Inhibitory control and negative affect in relation to food intake among youth. Appetite 2021 Jan 01;156:104858 [FREE Full text] [doi: 10.1016/j.appet.2020.104858] [Medline: 32891676]
14. Jensen CD, Duraccio KM, Carbine KA, Barnett KA, Kirwan CB. Motivational impact of palatable food correlates with functional brain responses to food images in adolescents. J Pediatr Psychol 2017 Jun 01;42(5):578-587. [doi: 10.1093/jpepsy/jsw091] [Medline: 27780839]
15. Ferreira I, Twisk JW, van Mechelen W, Kemper HC, Stehouwer CD. Development of fatness, fitness, and lifestyle from adolescence to the age of 36 years: determinants of the metabolic syndrome in young adults: the Amsterdam Growth and Health Longitudinal Study. Arch Intern Med 2005 Jan 10;165(1):42-48. [doi: 10.1001/archinte.165.1.42] [Medline: 15642873]
16. Rohde P, Stice E, Marti CN. Development and predictive effects of eating disorder risk factors during adolescence: implications for prevention efforts. Int J Eat Disord 2015 Mar;48(2):187-198 [FREE Full text] [doi: 10.1002/eat.22270] [Medline: 24599841]
17. Banfield EC, Liu Y, Davis JS, Chang S, Frazier-Wood AC. Poor adherence to US dietary guidelines for children and adolescents in the national health and nutrition examination survey population. J Acad Nutr Diet 2016 Jan;116(1):21-27 [FREE Full text] [doi: 10.1016/j.jand.2015.08.010] [Medline: 26391469]
18. Reedy J, Krebs-Smith SM. Dietary sources of energy, solid fats, and added sugars among children and adolescents in the United States. J Am Diet Assoc 2010 Oct;110(10):1477-1484 [FREE Full text] [doi: 10.1016/j.jada.2010.07.010] [Medline: 20869486]
19. Stok FM, De Ridder DT, Adriaanse MA, De Wit JB. Looking cool or attaining self-rule. Different motives for autonomy and their effects on unhealthy snack purchase. Appetite 2010 Jun;54(3):607-610. [doi: 10.1016/j.appet.2010.02.017] [Medline: 20211669]
20. Iannazzo LH, Hayden MJ, Lawrence NS, Kakoschke N, Hughes LK, Van Egmond K, et al. Inhibitory control training to reduce appetitive behaviour: a meta-analytic investigation of effectiveness, potential moderators, and underlying mechanisms of change. Health Psychol Rev 2025 Mar;19(1):66-96 [FREE Full text] [doi: 10.1080/17437199.2024.2410018] [Medline: 39397386]
21. Hendrikse JJ, Cachia RL, Kothe EJ, McPhie S, Skouteris H, Hayden MJ. Attentional biases for food cues in overweight and individuals with obesity: a systematic review of the literature. Obes Rev 2015 May;16(5):424-432. [doi: 10.1111/obr.12265] [Medline: 25752592]
22. Allom V, Mullan B, Hagger M. Does inhibitory control training improve health behaviour? A meta-analysis. Health Psychol Rev 2016 Jun;10(2):168-186. [doi: 10.1080/17437199.2015.1051078] [Medline: 26058688]
23. Donders FC. On the speed of mental processes. Acta Psychol (Amst) 1969;30:412-431. [doi: 10.1016/0001-6918(69)90065-1] [Medline: 5811531]
24. Stice E, Lawrence NS, Kemps E, Veling H. Training motor responses to food: a novel treatment for obesity targeting implicit processes. Clin Psychol Rev 2016 Nov;49:16-27. [doi: 10.1016/j.cpr.2016.06.005] [Medline: 27498406]
25. Folkvord F, Veling H, Hoeken H. Targeting implicit approach reactions to snack food in children: effects on intake. Health Psychol 2016 Aug;35(8):919-922. [doi: 10.1037/hea0000365] [Medline: 27505216]
26. Manasse SM, Lampe EW, Juarascio AS, Zhu J, Forman EM. Using virtual reality to train inhibitory control and reduce binge eating: a proof-of-concept study. Appetite 2021 Feb 01;157:104988 [FREE Full text] [doi: 10.1016/j.appet.2020.104988] [Medline: 33049341]
27. Cole KR, Yen CL, Dudley-Javoroski S, Shields RK. NIH toolbox cognition battery in young and older adults: reliability and relationship to adiposity and physical activity. J Geriatr Phys Ther 2021;44(1):51-59 [FREE Full text] [doi: 10.1519/JPT.0000000000000244] [Medline: 31567883]

XSL•FO
RenderX

28. Werthmann J, Roefs A, Nederkoorn C, Mogg K, Bradley BP, Jansen A. Can(not) take my eyes off it: attention bias for food in overweight participants. Health Psychol 2011 Sep;30(5):561-569 [FREE Full text] [doi: 10.1037/a0024291] [Medline: 21767019]
29. Schmidt R, Lüthold P, Kittel R, Tetzlaff A, Hilbert A. Visual attentional bias for food in adolescents with binge-eating disorder. J Psychiatr Res 2016 Sep;80:22-29 [FREE Full text] [doi: 10.1016/j.jpsychires.2016.05.016] [Medline: 27267318]
30. White MA, Whisenhunt BL, Williamson DA, Greenway FL, Netemeyer RG. Development and validation of the food-craving inventory. Obes Res 2002 Feb;10(2):107-114 [FREE Full text] [doi: 10.1038/oby.2002.17] [Medline: 11836456]
31. Bejarano CM, Hesse DR, Cushing CC. Hedonic appetite, affect, and loss of control eating: macrotemporal and microtemporal associations in adolescents. J Pediatr Psychol 2023 May 20;48(5):448-457. [doi: 10.1093/jpepsy/jsad004] [Medline: 36763682]
32. Manore MM. Exercise and the Institute of Medicine recommendations for nutrition. Curr Sports Med Rep 2005 Aug;4(4):193-198. [doi: 10.1097/01.csmr.0000306206.72186.00] [Medline: 16004827]
33. Peirce JW. Generating stimuli for neuroscience using PsychoPy. Front Neuroinform 2009 Jan 15;2:10 [FREE Full text] [doi: 10.3389/neuro.11.010.2008] [Medline: 19198666]
34. Veling H, van Koningsbruggen GM, Aarts H, Stroebe W. Targeting impulsive processes of eating behavior via the internet. Effects on body weight. Appetite 2014 Jul;78:102-109. [doi: 10.1016/j.appet.2014.03.014] [Medline: 24675683]
35. Blechert J, Meule A, Busch NA, Ohla K. Food-pics: an image database for experimental research on eating and appetite. Front Psychol 2014 Jun 24;5:617 [FREE Full text] [doi: 10.3389/fpsyg.2014.00617] [Medline: 25009514]
36. SAS program for CDC growth charts. Centers for Disease Control and Prevention. 2024 Sep 25. URL: https://www.cdc.gov/growth-chart-training/hcp/computer-programs/sas.html [accessed 2024-12-02]
37. Cappelleri JC, Bushmakin AG, Gerber RA, Leidy NK, Sexton CC, Karlsson J, et al. Evaluating the Power of Food Scale in obese subjects and a general sample of individuals: development and measurement properties. Int J Obes (Lond) 2009 Aug 9;33(8):913-922. [doi: 10.1038/ijo.2009.107] [Medline: 19506564]
38. Lowe MR, Butryn ML, Didie ER, Annunziato RA, Thomas JG, Crerand CE, et al. The Power of Food Scale. A new measure of the psychological influence of the food environment. Appetite 2009 Aug;53(1):114-118. [doi: 10.1016/j.appet.2009.05.016] [Medline: 19500623]
39. Summary tables: recommendations on potential approaches to dietary assessment for different research objectives requiring group-level estimates. National Cancer Institute. URL: https://dietassessmentprimer.cancer.gov/approach/table.html [accessed 2020-04-07]
40. Subar AF, Kirkpatrick SI, Mittl B, Zimmerman TP, Thompson FE, Bingley C, et al. The Automated Self-Administered 24-hour dietary recall (ASA24): a resource for researchers, clinicians, and educators from the National Cancer Institute. J Acad Nutr Diet 2012 Aug;112(8):1134-1137 [FREE Full text] [doi: 10.1016/j.jand.2012.04.016] [Medline: 22704899]
41. van Ens W, Schmidt U, Campbell IC, Roefs A, Werthmann J. Test-retest reliability of attention bias for food: robust eye-tracking and reaction time indices. Appetite 2019 May 01;136:86-92. [doi: 10.1016/j.appet.2019.01.020] [Medline: 30682381]
42. Forbush KT, Wildes JE, Pollack LO, Dunbar D, Luo J, Patterson K, et al. Development and validation of the Eating Pathology Symptoms Inventory (EPSI). Psychol Assess 2013 Sep;25(3):859-878. [doi: 10.1037/a0032639] [Medline: 23815116]
43. Bejarano CM, Hesse DR, Cushing CC. Assessing drivers of eating behavior in a community sample of adolescents: implications for preventive interventions. Clin Pract Pediatr Psychol 2025 Mar 01;13(1):89-98. [doi: 10.1037/cpp0000553]
44. Richson BN, Forbush KT, Chapa DA, Gould SR, Perko VL, Johnson SN, et al. Measurement invariance of the Eating Pathology Symptoms Inventory (EPSI) in adolescents and adults. Eat Behav 2021 Aug;42:101538 [FREE Full text] [doi: 10.1016/j.eatbeh.2021.101538] [Medline: 34247036]
45. Zelazo PD, Anderson JE, Richler J, Wallner-Allen K, Beaumont JL, Weintraub S. II. NIH Toolbox Cognition Battery (CB): measuring executive function and attention. Monogr Soc Res Child Dev 2013 Aug;78(4):16-33. [doi: 10.1111/mono.12032] [Medline: 23952200]
46. Flanker inhibitory control and attention test. NIH Toolbox. URL: https://nihtoolbox.org/test/flanker-inhibitory-control-and-attention-test-age-12/ [accessed 2025-11-21]
47. Veling H, Lawrence NS, Chen Z, van Koningsbruggen GM, Holland RW. What is trained during food go/no-go training? A review focusing on mechanisms and a research agenda. Curr Addict Rep 2017;4(1):35-41 [FREE Full text] [doi: 10.1007/s40429-017-0131-5] [Medline: 28357193]
48. Kemps E, Tiggemann M, Orr J, Grear J. Attentional retraining can reduce chocolate consumption. J Exp Psychol Appl 2014 Mar;20(1):94-102. [doi: 10.1037/xap0000005] [Medline: 24079387]
49. Love H, Bhullar N, Schutte NS. Go/no-go for food: attention bias and intention to eat unhealthy food. Appetite 2020 Jul 01;150:104646. [doi: 10.1016/j.appet.2020.104646] [Medline: 32145372]
50. Van Malderen E, Goossens L, Verbeken S, Kemps E. Unravelling the association between inhibitory control and loss of control over eating among adolescents. Appetite 2018 Jun 01;125:401-409. [doi: 10.1016/j.appet.2018.02.019] [Medline: 29476801]

XSL•FO
RenderX

51. Bartholdy S, O'Daly OG, Campbell IC, Banaschewski T, Barker G, Bokde AL, IMAGEN Consortium. Neural correlates of failed inhibitory control as an early marker of disordered eating in adolescents. Biol Psychiatry 2019 Jun 01;85(11):956-965. [doi: 10.1016/j.biopsych.2019.01.027] [Medline: 31122340]
52. Eagleton SG, Na M, Savage JS. Food insecurity is associated with higher food responsiveness in low-income children: the moderating role of parent stress and family functioning. Pediatr Obes 2022 Jan;17(1):e12837 [FREE Full text] [doi: 10.1111/ijpo.12837] [Medline: 34402203]
53. Mason TB, Do B, Wang S, Dunton GF. Ecological momentary assessment of eating and dietary intake behaviors in children and adolescents: a systematic review of the literature. Appetite 2020 Jan 01;144:104465 [FREE Full text] [doi: 10.1016/j.appet.2019.104465] [Medline: 31541670]
54. Jones A, Di Lemma LC, Robinson E, Christiansen P, Nolan S, Tudur-Smith C, et al. Inhibitory control training for appetitive behaviour change: a meta-analytic investigation of mechanisms of action and moderators of effectiveness. Appetite 2016 Feb 01;97:16-28 [FREE Full text] [doi: 10.1016/j.appet.2015.11.013] [Medline: 26592707]
55. Smith KE, Mason TB, Schaefer LM, Juarascio A, Dvorak R, Weinbach N, et al. Examining intra-individual variability in food-related inhibitory control and negative affect as predictors of binge eating using ecological momentary assessment. J Psychiatr Res 2020 Jan;120:137-143 [FREE Full text] [doi: 10.1016/j.jpsychires.2019.10.017] [Medline: 31677551]
56. Poppelaars A, Scholten H, Granic I, Veling H, Johnson-Glenberg MC, Luijten M. When winning is losing: a randomized controlled trial testing a video game to train food-specific inhibitory control. Appetite 2018 Oct 01;129:143-154. [doi: 10.1016/j.appet.2018.06.039] [Medline: 30008434]
57. Manasse SM, Lampe EW, Gillikin L, Payne-Reichert A, Zhang F, Juarascio AS, et al. The project REBOOT protocol: evaluating a personalized inhibitory control training as an adjunct to cognitive behavioral therapy for bulimia nervosa and binge-eating disorder. Int J Eat Disord 2020 Jun;53(6):1007-1013 [FREE Full text] [doi: 10.1002/eat.23225] [Medline: 32221989]
58. Chami R, Cardi V, Lawrence N, MacDonald P, Rowlands K, Hodsoll J, et al. Targeting binge eating in bulimia nervosa and binge eating disorder using inhibitory control training and implementation intentions: a feasibility trial. Psychol Med 2022 Apr;52(5):874-883. [doi: 10.1017/S0033291720002494] [Medline: 32713405]
59. Neumark-Sztainer DR, Wall MM, Haines JI, Story MT, Sherwood NE, van den Berg PA. Shared risk and protective factors for overweight and disordered eating in adolescents. Am J Prev Med 2007 Nov;33(5):359-369 [FREE Full text] [doi: 10.1016/j.amepre.2007.07.031] [Medline: 17950400]
60. Denny KN, Loth K, Eisenberg ME, Neumark-Sztainer D. Intuitive eating in young adults. Who is doing it, and how is it related to disordered eating behaviors? Appetite 2013 Jan;60(1):13-19 [FREE Full text] [doi: 10.1016/j.appet.2012.09.029] [Medline: 23063606]
61. Forman EM, Shaw JA, Goldstein SP, Butryn ML, Martin LM, Meiran N, et al. Mindful decision making and inhibitory control training as complementary means to decrease snack consumption. Appetite 2016 Aug 01;103:176-183. [doi: 10.1016/j.appet.2016.04.014] [Medline: 27083129]
62. Kiken LG, Garland EL, Bluth K, Palsson OS, Gaylord SA. From a state to a trait: trajectories of state mindfulness in meditation during intervention predict changes in trait mindfulness. Pers Individ Dif 2015 Jul 01;81:41-46 [FREE Full text] [doi: 10.1016/j.paid.2014.12.044] [Medline: 25914434]
63. Hofmann W, Friese M, Roefs A. Three ways to resist temptation: the independent contributions of executive attention, inhibitory control, and affect regulation to the impulse control of eating behavior. J Exp Soc Psychol 2009 Feb;45(2):431-435 [FREE Full text] [doi: 10.1016/j.jesp.2008.09.013]

Abbreviations

ASA24: Automated Self-Administered 24-Hour Dietary Assessment Tool
EMA: ecological momentary assessment
EPSI: Eating Pathology Symptoms Inventory
GNG: go/no go
ICT: inhibitory control training
NIH: National Institutes of Health
PFS: Power of Food Scale

XSL•FO
RenderX

Edited by J Sarvestan; submitted 15.05.25; peer-reviewed by JA Kosmas, RN Sahlan; comments to author 24.06.25; revised version received 11.07.25; accepted 10.10.25; published 05.12.25.

Please cite as:
Bejarano CM, Ortega A, Cushing CC
Inhibitory Control, Attentional Bias, and Palatable Food Consumption in Adolescents: A Laboratory Feasibility Randomized Controlled Trial
JMIR Form Res 2025;9:e77579
URL: https://formative.jmir.org/2025/1/e77579
doi:10.2196/77579
PMID:

©Carolina M Bejarano, Adrian Ortega, Christopher C Cushing. Originally published in JMIR Formative Research (https://formative.jmir.org), 05.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Integrating a Brief Behavioral Intervention Into Case Management for Mothers With Perinatal Substance Use Disorder: Nonrandomized Pilot Feasibility Study

Amritha Bhat[1], MPH, MD; Mary C Curran[2], MSW, LICSW; Mary Jane Lohr[3], MS; Haley Smith[1], MPA; Brittany Blanchard[1], PhD; Susan A Stoner[1], PhD; Therese Grant[1], PhD; Nancy Grote[4], PhD

[1]Department of Psychiatry and Behavioral Sciences, University of Washington, 1959 NE Pacific Street, Box 356560, Seattle, WA, United States

[2]Department of Rehabilitation Medicine, University of Washington, Seattle, WA, United States

[3]School of Nursing, University of Washington, Seattle, WA, United States

[4]School of Social Work, University of Washington, Seattle, WA, United States

Corresponding Author:
Amritha Bhat, MPH, MD
Department of Psychiatry and Behavioral Sciences, University of Washington, 1959 NE Pacific Street, Box 356560, Seattle, WA, United States

Abstract

Background: Perinatal substance use disorders (SUD) are frequently comorbid with depression, anxiety, and posttraumatic stress disorder (PTSD), contributing to adverse maternal and child outcomes. Access to integrated mental health support within existing SUD service frameworks is limited, particularly for pregnant and parenting individuals facing socioeconomic and psychosocial instability. Promoting Healthy Families (PHF) is a brief behavioral intervention designed for delivery by case managers serving high-risk perinatal populations with substance use within programs such as Parent Child Assistance Program (PCAP).

Objective: This study aimed to evaluate the feasibility of integrating PHF into intensive case management for pregnant and postpartum clients with at-risk perinatal substance use, and to assess preliminary outcomes of measures of maternal depression, anxiety, and PTSD symptoms.

Methods: In this nonrandomized pilot study (April 2018-September 2021), eligible clients were allocated to either PCAP alone (control) or to PHF delivered within PCAP (intervention). Case managers completed an anonymous feasibility survey addressing ease of delivery and fit with their workflow. Participating clients completed the Patient Health Questionnaire-9 (PHQ-9), Generalized Anxiety Disorder-7 Scale, and PTSD (posttraumatic stress disorder) Checklist (PCL-6) at baseline and at 4, 6, and 12 months. Data collection overlapped with the COVID-19 pandemic, which affected service access and delivery.

Results: CMs and 1 program supervisor (n=10) reported that PHF was feasible to deliver within PCAP, and respondents indicated clients benefited somewhat (70%) or a lot (30%). Most (70%) noted an increase in workload and recommended additional supervision and training. The pilot study enrolled 58 clients (29 PHF+PCAP and 29 PCAP), with 60% (35/58) completing all follow-up assessments. While differences between groups over time were not statistically significant, changes were in the predicted direction for PHQ-9 and PCL-6 scores. Symptom improvement rates were high: In the PHF+PCAP group, 85% (25/29) showed ≥5-point decreases in PHQ-9 scores, 68% (20/29) had ≥6-point decreases in Generalized Anxiety Disorder-7 Scale scores, and 93% (27/29) had ≥5-point decreases in PCL-6 scores.

Conclusions: PHF can be feasibly delivered within an existing intensive case management program for perinatal SUD, with early signals of mental health improvement across both intervention and control groups. Future adequately powered randomized controlled trials should investigate the effectiveness of brief behavioral interventions within perinatal SUD case management programs, optimal delivery timing, and the potential to enhance mental health care integration for high-risk perinatal populations.

(JMIR Form Res 2025;9:e75132) doi:10.2196/75132

KEYWORDS

perinatal; mental health; substance use; case management; feasibility

XSL•FO
RenderX

Introduction

Substance use is prevalent among pregnant and parenting individuals in the United States. Recent research indicates that 13.5% of pregnant adults currently drink alcohol, and 5.2% have engaged in binge drinking in the past 30 days [1]. Additionally, recent findings from the National Survey on Drug Use and Health show that 5.8% of pregnant women have used an illicit substance in the past month [2], and 11.3% meet criteria for a perinatal substance use disorder (SUD). Substance use in pregnancy is associated with negative outcomes, including miscarriage, teratogenicity, preterm birth, low birth weight, neonatal withdrawal syndromes, and child mental health or behavioral conditions [3].

Individuals with SUD are at greater risk of experiencing mental health conditions and vice versa [4]. Up to 50% of women with SUD have depressive symptoms [5-7] and 24% - 58% have posttraumatic stress disorder (PTSD) [8-10]. Women with SUD who have comorbid mental health conditions tend to have more difficulty engaging with SUD treatment, more chronic SUD [11], and lower rates of regaining child custody [12]. Comorbidity also increases the risk for suicidal ideation and self-harm behaviors; up to 32% of women with comorbid alcohol and drug use and mental health conditions have attempted suicide or self-harm in the past 3 months [13], and mental health conditions and substance use together are the leading cause of maternal mortality in the United States [14].

Despite the high prevalence and additive negative impacts, treatment rates for comorbid SUD and mental health conditions remain concerningly low [15]. Treating comorbid mental health conditions can help improve outcomes of SUD treatment [11], and delivering mental health treatment concurrent with SUD treatment can reduce barriers to treatment access [16]. However, there are several barriers to delivering mental health and substance use treatment concurrently [17], including a shortage of providers trained to address both conditions.

The Parent Child Assistance Program (PCAP) is a 3-year intensive case - management program for pregnant and parenting individuals with at-risk perinatal substance use [12,18-20], first developed in Seattle in the early 1990s to reduce prenatal substance exposure. The program was built upon 3 theoretical foundations: relational theory, the transtheoretical model, and harm reduction. Relational theory underscores the importance of interpersonal relationships to clients as they grow, develop, and define themselves [21]. The PCAP model puts concepts of relational theory into practice by offering personalized, knowledgeable, and compassionate support from a single case manager (CM) who works consistently with the client for 3 years, a period long enough for the process of gradual and realistic change to occur. The transtheoretical model [22] is widely known for emphasizing that, with regard to behavior change, individuals ebb and flow in their readiness to change over time and that helping strategies will be most successful when they are tailored to the client's readiness to change, with strong attention to building intrinsic motivation and self-efficacy. In practice, the most important way in which a PCAP CM aims to exert a positive effect on a client's self-efficacy is by listening carefully to what is important to the client and how the client thinks about various problems, and valuing the client's perspective. CMs then promote self-directed action by helping clients define and accomplish explicit goals toward behavioral change. PCAP intervention strategies are based on harm-reduction principles with an understanding that alcohol and drug use patterns and practices can be placed along a continuum from minimally to extremely harmful [23]. PCAP seeks to reduce the risk of harmful consequences associated with substance use. Though abstinence from substance use is not required, PCAP does view sustained abstinence as the highest possible degree of harm reduction. PCAP aims to help clients pursue long-term SUD recovery through connection to community, social, health, treatment, and recovery services. PCAP CMs are highly familiar with and well-connected to the resources available in their local communities. They meet with clients at least twice per month, in clients' own homes whenever possible, to learn about clients' needs and goals and to gain a working understanding of the contexts of clients' lives. Together, PCAP CMs and clients identify goal-oriented action steps, and CMs leverage their community connections to support clients as they navigate recovery.

More than 70% of PCAP clients report having a diagnosed mental health disorder. Among these, 70% report a diagnosis of depression. More than half (58%) are involved with the child welfare system at enrollment [24]. High rates of psychiatric comorbidity in this population are especially concerning, as PCAP data indicate that those mothers with more serious psychiatric problems are less likely than other mothers to retain or regain child custody during the 3-year intervention [12]. Even with a PCAP CM, mothers' untreated mental health problems may limit their ability to access mental health services and use other available community services to build a stable home environment for themselves and their child or children.

Task-sharing models can be used to address the treatment gap for comorbid mental health and SUD. In these models, tasks are clearly defined and distributed across larger teams. For example, lay or community health workers are trained in mental health care delivery, and specialists provide ongoing supervision, support, and consultation to frontline nonspecialist health workers [25-27]. To address the unmet mental health needs of PCAP clients, we developed a task-shared, low-intensity intervention for depression that PCAP CMs could deliver to their clients. This intervention, Promoting Healthy Families (PHF), is adapted from MOMCare, an evidence-based model using Brief Interpersonal Psychotherapy [28] found to reduce depression and anxiety in pregnant women [29] that can be delivered effectively by nonspecialists [30]. Over the course of 3 years, we tested the integration of PHF into PCAP in a pilot quasi-experimental study to assess the feasibility of PHF delivered by PCAP CMs and the preliminary effectiveness of PHF on clients' mental health symptom severity. We hypothesized that PCAP CMs would find it feasible to deliver the PHF intervention to their clients.

XSL•FO
RenderX

Methods

Procedure

Participants were enrolled from April 2018 through October 2020 at 3 participating PCAP sites in Washington State (2 rural and 1 urban) [31]. PCAP eligibility criteria include age 18 years or older; pregnant or up to 24 months post partum; engaged in at-risk substance use during the most recent pregnancy; and inadequately connected to health, social, and community services. Eligibility for the study was assessed by PCAP supervisors at the time clients were enrolled in PCAP. Additional criteria to be enrolled in the study were current depressive symptomatology and either anxiety or PTSD, as evidenced by a Patient Health Questionnaire-9 (PHQ-9) score of ≥10; and a Generalized Anxiety Disorder-7 Scale (GAD-7) score of ≥10, or a PTSD (posttraumatic stress disorder) Checklist (PCL-6) score of ≥14. Clients were excluded from the study if they reported current psychotic symptoms or had permanently lost custody of the index child. Eligible clients were assigned by supervisors to either PCAP alone (control condition) or to PHF in PCAP (intervention condition) based on the availability of trained CMs to provide PHF. We continued enrollment until 60 women were enrolled. In total, 2 participants withdrew from the study before starting any study procedures, resulting in a final sample of 29 women in the PHF in the PCAP group and 29 women in the control PCAP group.

Interventions

PCAP Only

Participants in the PCAP-only condition received only the PCAP intervention, which is described in detail elsewhere [12,19]. In brief, each participant was assigned to a bachelor-level CM, whose charge was to meet with the client on approximately a biweekly basis, in the client's own home whenever possible. Together, the client and CM reviewed the client's goals and progress toward them every 4 months. Under the guidance of a PCAP clinical supervisor, CMs provided tailored care navigation, modeling, and support, including occasional transportation to appointments in support of the client's own goals, if needed.

PHF in PCAP

Participants in the PHF in the PCAP condition received both interventions. A total of 4 PCAP clinical supervisors and 8 PCAP CMs from 3 participating PCAP sites were trained by NG and MC, experts in interpersonal psychotherapy [32] and developers of the PHF intervention. Trainees received a training manual and participated in a 2-day training on the principles and methods of PHF, including strategies to help PCAP clients with symptoms of depression and anxiety, strengthen social supports, access resources to meet basic needs, and parent effectively. PHF is a culturally relevant brief behavioral intervention that teaches self-care for depression. The goals of PHF are to reduce depression symptom severity, increase and activate social support, and improve interpersonal functioning. CMs used "quick sheets" (detailed fidelity checklists based on methods established in MOMCare, a brief intervention based on Interpersonal Therapy) [33] with clients at each session to structure and guide them through the session and ensure delivery of the intervention's essential components. We modified the quick sheets based on feedback from PCAP CMs. First, we reduced the number of problem areas, limiting it to either managing a complicated pregnancy or role transition of parenting a child, specifically addressing issues related to substance use, such as the fear of loss of child custody at birth, family and cultural myths around seeking medication treatment, and coping with loss of custody. If the problem areas of complicated grief and role dispute came up, we recommended that they be a secondary focus or that they be addressed once the clients' depressive symptoms improved. We changed the name of the problem area role dispute to disagreements with a significant other. We focused education sessions on using 4 behavioral strategies: interpersonal behavioral activation, building effective communication skills, role coaching or role-playing, and problem solving. PHF CM used concrete tools to help build self-care skills, such as a mood chart, goal-setting worksheets, and weekly homework, focused on attaining pleasure or mastery. We facilitated access to support services, particularly around mental health counseling and medication treatment, and at times, the PHF CM would accompany the client to an initial appointment. During the intervention period, NG and MC provided virtual weekly, half-hour supervision and worked collaboratively with CMs to adapt PHF strategies to fit the context of the participants' lives.

Table 1 illustrates the mental health intervention components of the Parent Child Assistance Program and Promoting Healthy Families models.

XSL•FO
RenderX

Table . Key mental health intervention components of the Parent Child Assistance Program and Promoting Healthy Families.

Components	PCAP[a]	PHF[b]
SUD[c] treatment	Referral to SUD treatment in community	Referral to SUD treatment in community
Acute MH[d] intervention	Referral for mental health or depression care in community	CM[e] delivers PHF (weekly or every 2 weeks for 8 weeks or as needed up to 16 weeks). Assistance and encouragement with connecting with MH services and meds if desired
Maintenance MH intervention	None	Once a month for 1 year
MH outcomes tracking	PCAP intake and 3-year exit symptoms assessment using Addiction Severity Index	PHQ-9[f] assessments at each session, GAD-7[g] and PCL-6[h] each session (time-permitting, if relevant), as well as tracking medications, dosage, and adherence
Supervision	Twice a month individual supervision with PCAP clinical supervisor	Weekly PHF group supervision with CM, PCAP supervisor, and PHF trainers

[a]PCAP: Parent Child Assistance Program.
[b]PHF: promoting healthy families.
[c]SUD: substance use disorders.
[d]MH: mental health.
[e]CM: case manager.
[f]PHQ–9: Patient Health Questionnaire-9.
[g]GAD-7: Generalized Anxiety Disorder-7 Scale.
[h]PCL-6: PTSD (posttraumatic stress disorder) Checklist.

Measures

To assess feasibility, we used an 8-item anonymous online survey. Depression and anxiety symptom severity were assessed using the PHQ-9 [34] and GAD-7, respectively [35]. PTSD symptoms were assessed using the 6-item PCL-6, a brief screening tool [36] derived from items on the PTSD checklist–civilian version with evidence of good psychometric properties [37].

Data Collection and Analysis

We collected data on perceptions of the feasibility of delivering PHF from the intervention condition PCAP staff.

To assess preliminary effectiveness, standard measures of mental health symptoms (described above) were collected at baseline and at 3 follow-up time points: 4 - 5 months post baseline (mean 4.6, SD 0.4), 6 - 10 months post baseline (mean 8, SD 1.4), and 12 - 16 months post baseline (mean 13.8, SD 1.5). Baseline data were collected in person by a PCAP supervisor as part of PCAP intake and screening. A research assistant who was blind to the intervention assignment conducted the brief follow-up assessments by telephone. Attempts were made to collect data at each follow-up time point.

We compared means between groups, controlling for baseline scores and cumulative number of PHF sessions. In a secondary analysis, we categorized the sample into 3 groups based on how their PHQ-9, GAD-7, and PCL scores changed over the study period: improved, no change, and deteriorated. A 5-point change was used to determine a clinically significant change on the PHQ-9 [38] and PCL-6 [39], and 6 points for the GAD-7 [40]. We computed change scores from baseline through the fourth (final) timepoint. To be conservative, if a client's scores improved at some timepoints and worsened at others over the 4 timepoints, this would be scored as "deteriorated." Due to low base rates, we collapsed no change and deteriorated groups and conducted a Fisher exact test to assess differences between PHF in PCAP and the PCAP only. Analyses were conducted in IBM SPSS Statistics (version 19) [41].

Ethical Considerations

We obtained approval to conduct the pilot study from the Washington State Institutional Review Board [D-030310-S, "PCAP"]. Informed consent was obtained from all participants. Study data were deidentified for analysis. Participants received a US $5 gift card for a local retailer after each completed data collection (baseline and 4, 6, and 12 months post baseline).

Results

Survey Results

Overall, 9 CMs and 1 PCAP supervisor responded to the anonymous survey regarding their perceptions of the feasibility of delivery. All reported that their clients seemed to have benefited from PHF (3/10 reported that their clients benefited "a lot" and 7/10 that they benefited "somewhat"). CMs indicated that PHF had affected their work "a lot positively" (60% of respondents, 6/10) "somewhat positively" (2/10 respondents), and "somewhat negatively" (2/10). When asked about their work burden changes, they noted that their PHF-related duties had increased their workload "a lot" (2/10, 20%), "somewhat" (7/10, 70%), and "not at all" (1/10, 10%).

CMs were asked about ways to reduce the workload burden that came with adding PHF. They suggested hiring new CMs that would do PHF-specific work "rather than trying to carve out with existing CMs' caseload" or allowing for "smaller

XSL•FO
RenderX

caseloads." They also identified a need for more supervision, emotional support, and more streamlined program implementation to reduce administrative burdens associated with this added responsibility.

Survey participants suggested improving the capacity and skill levels of CMs (through more staff training in the PHF model and comorbid mental health and SUD) and increasing the amount of time available to incorporate PHF into their workload. Recommendations for improving PHF for clients included offering incentive funds and an official certificate of completion to clients. CMs also suggested that PHF enrollment should occur when clients are in active recovery, have their basic needs met (eg, housing), and when there has been enough time for relationship-building with the CM. One CM stated that:

> *I wonder if including PHF after the client's life has settled down a little would be more beneficial. This might also control for situational depression/anxiety due to life instability that can be solved by meeting [more foundational] needs.*

Pilot Trial Results

Only 1 participant who screened eligible for the study chose not to participate, 1 participant enrolled in PHF but dropped out of PCAP before starting PHF and so became ineligible, and 1 participant dropped out of PHF before starting (but was still in PCAP). A total of 29 clients were assigned to PCAP and 29 to PHF in PCAP. When it was not possible for the woman to participate by telephone, assessments were completed online; 29% (17/58) of assessments were completed in this way. The average number of completed follow-up assessments per participant was 2.3 out of 3 (SD 1.1), and 60% (35/58) of participants completed all assessments. There were no significant differences in the number of follow-up assessments between the 2 groups. The average number of PHF sessions over 1 year, including acute and maintenance sessions, was 13.3 (SD 8.9; range 0 - 34) sessions, and women received PHF services for an average of 8.5 (SD 4.7; range 0 - 14) months. The mean age of participants was 28 (SD 6) years, and on average, 52% (30/58) had completed 12 years of high school and 93% (54/58) were unmarried. With regard to race and ethnicity, 58% (33/58) were non-Hispanic White, 32% (18/58) were non-Hispanic American Indian/Alaska Native/Canadian First Nations, and 10% (6/58) were non-Hispanic Black, non-Hispanic Pacific Islander, or Hispanic American Indian/Alaska Native/Canadian First Nations. Baseline mental health data are shown in Table 2. At baseline, participants in the PHF in PCAP group had marginally higher depression scores than the control group (P=.06). While the means were not significantly different between groups at any of the follow-up assessments, the results were in the predicted direction for the PHQ-9 from T2-T4 and for the PCL-C from T2-T3 (Figure 1). Based on biannual CM reports, participants in the PHF in the PCAP group were significantly more likely to be receiving individual mental health counseling at 24 months (n=23; χ^2_1=4.48; P=.03) as compared to those in PCAP alone.

Table . Baseline mean symptom scores on mental health measures.

Mental health measures[a]	PHF[b] in PCAP[c] (n=29), mean (SD)	PCAP only (n=29), mean (SD)	Significance	
			t test (*df*)	*P* value
PHQ-9[d] Depression symptom scores	17.4 (5)	15.1 (4.4)	−1.9 (56)	.06
GAD-7[e] Anxiety symptom scores	15.1 (4.3)	15.3 (4.1)	0.2 (56)	.85
PCL-6[f] PTSD[g] symptom scores	21 (5.3)	19.8 (5.7)	−0.9 (56)	.39

[a]Cut-off scores that indicate need for further assessment and intervention: PHQ-9 ≥10, GAD-7 score ≥10, and PCL-6 ≥14.

[b]PHF: promoting healthy families.

[c]PCAP: Parent Child Assistance Program.

[d]PHQ-9: Patient Health Questionnaire-9.

[e]GAD-7: Generalized Anxiety Disorder-7.

[f]PCL-6: PTSD (posttraumatic stress disorder) Checklist.

[g]PTSD: posttraumatic stress disorder.

XSL•FO
RenderX

Figure 1. Comparison of Patient Health Questionnaire-9, Generalized Anxiety Disorder-7 Scale, and Posttraumatic Stress Disorder Checklist–Civilian mean scores between Parent Child Assistance Program and Promoting Healthy Families in Parent Child Assistance Program. GAD-7: Generalized Anxiety Disorder-7 Scale; PCAP: Parent Child Assistance Program; PCL-C: Posttraumatic Stress Disorder Checklist–Civilian; PHF: Promoting Healthy Families; PHQ-9: Patient Health Questionnaire-9.

XSL•FO
RenderX

Our final analysis was to categorize scores with respect to improved symptoms, stable symptoms, and worsening symptoms (Table 3). In our categorical analysis of depression, 85% (24/29) of the PHF in the PCAP group recorded a 5-point drop in PHQ depression score at some point from T1 during the intervention period, and none recorded a 5-point increase. In comparison, in the PCAP-only group, 71% (21/29) recorded a 5-point drop in PHQ-9 and 17% (5/29) recorded a 5-point increase. With regard to anxiety, 68% (20/29) of the PHF in the PCAP group reported a 6-point drop at some point from T1, and 7% (2/29) had a 6-point increase in scores at some point from T1. Among the PCAP-only group, 71% (21/29) reported a 6-point drop in GAD-7 scores, and 4% (1/29) of scores were worse. In our analysis of PTSD, 93% (27/29) of the PHF in the PCAP group recorded a 5-point drop in PTSD score at some point from T1, and none recorded a 5-point increase. In the PCAP-only group, 71% (21/29) recorded a 5-point drop in PTSD and 21% (6/29) recorded a 5-point increase.

Table . Clinical change in mental health through the study period by treatment group. Due to unbalanced cells and cell sizes smaller than 5, we report *P* values from the Fisher exact test.

Mental health measures and changes	PHF[a]/PCAP[b] (n=29), n (%)	PCAP (n=29), n (%)	Test statistic	
			OR[c] (95% CI)	*P* value
Depression			2.37 (0.60-9.40)	.31
Improvement	25 (85)	21 (71)		
No change or deterioration	4 (15)	8 (29)		
Anxiety			0.82 (0.25-2.71)	.77
Improvement	19 (67)	21 (71)		
No change or deterioration	10 (33)	8 (29)		
PTSD[d]			5.15 (0.95-27.84)	.07
Improvement	27 (93)	21 (71)		
No change or deterioration	2 (7)	8 (29)		

[a]PHF: promoting healthy families.

[b]PCAP: Parent-Child Assistance Program.

[c]OR: odds ratio.

[d]PTSD: post-traumatic stress disorder.

Discussion

Principal Findings

This pilot study examined the feasibility and potential impact of integrating the PHF brief behavioral intervention into PCAP for mothers with SUD. Findings suggest that CMs were able to deliver PHF within the existing PCAP framework. Although the small sample size and nonrandomized allocation limit conclusions, these initial results suggest important directions for service delivery and research in perinatal SUD care. Although no comparisons were statistically significant (likely due to insufficient power), the PTSD symptom improvement was notably higher in PHF in PCAP versus PCAP alone. The improvement in mental health symptoms across both study arms may reflect PCAP's emphasis on connecting mothers to community services. Prior work has demonstrated that service linkage is a key predictor of improved social-emotional functioning for mothers and children [20]. Our study did not collect information on service use during the intervention, limiting our interpretation of this finding; however, it is plausible that reducing unmet needs and enhancing social support through PCAP played a role in mitigating stress and improving mental health symptoms. To the extent that connecting mothers to community services improves mental health and social-emotional functioning, activating mothers who may feel stuck due to depression or anxiety and warm handoffs to beneficial services may have a synergistic impact.

These findings carry implications for case management programs and clinicians supporting individuals with perinatal SUD. Delivering targeted brief behavioral interventions alongside SUD case management appears feasible. Embedding brief behavioral interventions into programs such as PCAP may enhance program offerings without requiring major personnel or operational changes. However, the delivery of the brief behavioral intervention may be best timed to follow stabilization of housing, safety, and basic needs. For researchers, these results suggest the need for larger randomized control trials to disentangle the effects of behavioral interventions and PCAP, with careful measurement of service use, CM contact time, and contextual disruptions such as the COVID-19 pandemic. Policymakers and funders may view the combination of SUD case management and a brief behavioral intervention as a potentially cost-effective strategy with wide-reaching benefits, including reduced mental health burden, improved parenting, and greater family preservation. Finally, and most importantly, for clients and families, timely and concurrent access to a brief behavioral intervention within the trusted case management relationship may lead to improved mental health without increasing the burden of accessing mental health treatments.

XSL•FO
RenderX

Limitations

This study has several limitations, including the use of a nonvalidated feasibility survey, a nonrandomized design, and the absence of service use and CM contact time data. The COVID-19 pandemic disrupted core supports such as SUD treatment access, visitation with children, employment, and prenatal care, which may have reduced the measurable impact of both interventions. Staff turnover further delayed enrollment and affected continuity of care, potentially influencing participant outcomes. Future iterations of this approach would ideally take into consideration the feedback from CMs that delivering the intervention increased their workload. Despite these constraints, this pilot study offers foundational evidence that brief behavioral interventions can be delivered within PCAP and suggests potential benefits for clients with PTSD that merit further exploration. A fully powered randomized controlled trial is necessary to determine the effectiveness, optimal delivery timing, and sustainability of integrating brief behavioral interventions into SUD case management.

Conclusions

Our findings are broadly consistent with prior research on integrating behavioral health interventions into perinatal case management. Home visiting and case management programs have demonstrated that embedding mental health components can improve maternal engagement and symptom outcomes [42,43]. The observed trends toward improvement in depression and PTSD symptoms in our study align with evidence that structured, relationship-based interventions can reduce psychological distress among mothers with SUD [44]. .Our results did not reach statistical significance, likely due to limited sample size and pandemic-related disruptions.

Addressing the challenge of comorbid SUD and mental health in the critical perinatal period by integrating interventions into existing service frameworks is a pragmatic pathway to breaking intergenerational cycles of trauma. Although preliminary, our findings highlight the promise of a low-barrier integrated approach that meets parents where they are.

Acknowledgments

The authors would like to thank the Parent Child Assistance Program case managers, supervisors, and clients for their participation. The authors used ChatGPT for readability and word choice. The author has reviewed the content and takes full responsibility for the content of the publication.

Funding

Mark Torrance Foundation, Division of Behavioral Health and Recovery (DBHR) of the Washington State Health Care Authority, The funder was not involved in the study design, data collection, analysis, interpretation, or the writing of the manuscript. In the writing of this manuscript, the author used Chat-GPT for readability and word choice. The author has reviewed the content and takes full responsibility for the content of the publication.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization, investigation, writing – original draft, and writing – review and editing: AB
Conceptualization, investigation, and writing – review and editing: MC
Data curation, formal analysis, project administration, visualization, and writing – review and editing: MJL
Investigation, data collection, and writing – review and editing: HS
Data curation, formal analysis, and writing – review and editing: BB
Data curation, formal analysis, visualization, and writing – review and editing: SS
Conceptualization, investigation, resources, supervision, and writing – review and editing: TG
Conceptualization, funding acquisition, investigation, methodology, resources, and supervision: NG

Conflicts of Interest

None declared.

References

1. Gosdin LK, Deputy NP, Kim SY, Dang EP, Denny CH. Alcohol consumption and binge drinking during pregnancy among adults aged 18-49 years - United States, 2018-2020. MMWR Morb Mortal Wkly Rep 2022 Jan 7;71(1):10-13. [doi: 10.15585/mmwr.mm7101a2] [Medline: 34990444]
2. Substance abuse and mental health services administration. National Survey of Drug Use and Health (NSDUH) Releases. 2022.
3. Cook JL, Green CR, de la Ronde S, et al. Epidemiology and effects of substance use in pregnancy. J Obstet Gynaecol Can 2017 Oct;39(10):906-915. [doi: 10.1016/j.jogc.2017.07.005] [Medline: 28935056]

XSL•FO
RenderX

4. Pentecost R, Latendresse G, Smid M. Scoping review of the associations between perinatal substance use and perinatal depression and anxiety. J Obstet Gynecol Neonatal Nurs 2021 Jul;50(4):382-391. [doi: 10.1016/j.jogn.2021.02.008] [Medline: 33773955]
5. Burns K, Melamed J, Burns W, Chasnoff I, Hatcher R. Chemical dependence and clinical depression in pregnancy. J Clin Psychol 1985 Nov;41(6):851-854. [doi: 10.1002/1097-4679(198511)41:6<851::aid-jclp2270410621>3.0.co;2-4] [Medline: 4078013]
6. Ross LE, Dennis CL. The prevalence of postpartum depression among women with substance use, an abuse history, or chronic illness: a systematic review. J Womens Health (Larchmt) 2009 Apr;18(4):475-486. [doi: 10.1089/jwh.2008.0953] [Medline: 19361314]
7. Chapman SLC, Wu LT. Postpartum substance use and depressive symptoms: a review. Women Health 2013;53(5):479-503. [doi: 10.1080/03630242.2013.804025] [Medline: 23879459]
8. Najavits LM, Weiss RD, Shaw SR. The link between substance abuse and posttraumatic stress disorder in women. A research review. Am J Addict 1997;6(4):273-283. [Medline: 9398925]
9. Anderson RE, Hruska B, Boros AP, Richardson CJ, Delahanty DL. Patterns of co-occurring addictions, posttraumatic stress disorder, and major depressive disorder in detoxification treatment seekers: implications for improving detoxification treatment outcomes. J Subst Abuse Treat 2018 Mar;86:45-51. [doi: 10.1016/j.jsat.2017.12.009] [Medline: 29415850]
10. Shen Y, Lo-Ciganic WH, Segal R, Goodin AJ. Prevalence of substance use disorder and psychiatric comorbidity burden among pregnant women with opioid use disorder in a large administrative database, 2009–2014. J Psychosom Obstet Gynaecol 2020 Feb 18:1-7. [doi: 10.1080/0167482X.2020.1727882] [Medline: 32067526]
11. Worley MJ, Tate SR, Brown SA. Mediational relations between 12-Step attendance, depression and substance use in patients with comorbid substance dependence and major depression. Addiction 2012 Nov;107(11):1974-1983. [doi: 10.1111/j.1360-0443.2012.03943.x] [Medline: 22578037]
12. Grant T, Huggins J, Graham JC, Ernst C, Whitney N, Wilson D. Maternal substance abuse and disrupted parenting: Distinguishing mothers who keep their children from those who do not. Child Youth Serv Rev 2011 Nov;33(11):2176-2185. [doi: 10.1016/j.childyouth.2011.07.001]
13. Harned MS, Najavits LM, Weiss RD. Self‐harm and suicidal behavior in women with comorbid PTSD and substance dependence. American J Addict 2006 Sep 10;15(5):392-395. [doi: 10.1080/10550490600860387]
14. Creanga AA. Maternal mortality in the United States: a review of contemporary data and their limitations. Clin Obstet Gynecol 2018 Jun;61(2):296-306. [doi: 10.1097/GRF.0000000000000362] [Medline: 29561285]
15. Haller DL, Miles DR. Psychopathology is associated with completion of residential treatment in drug dependent women. J Addict Dis 2004;23(1):17-28. [doi: 10.1300/J069v23n01_02] [Medline: 15077837]
16. McKee SA. Concurrent substance use disorders and mental illness: Bridging the gap between research and treatment. Canadian Psychology / Psychologie canadienne 2017;58(1):50-57. [doi: 10.1037/cap0000093]
17. Priester MA, Browne T, Iachini A, Clone S, DeHart D, Seay KD. Treatment access barriers and disparities among individuals with co-occurring mental health and substance use disorders: an integrative literature review. J Subst Abuse Treat 2016 Feb;61:47-59. [doi: 10.1016/j.jsat.2015.09.006] [Medline: 26531892]
18. Ernst CC, Grant TM, Streissguth AP, Sampson PD. Intervention with high-risk alcohol and drug-abusing mothers: II. Three-year findings from the Seattle model of paraprofessional advocacy. J Community Psychol 1999 Jan;27(1):19-38. [doi: 10.1002/(SICI)1520-6629(199901)27:1<19::AID-JCOP2>3.0.CO;2-K]
19. Grant T, Christopher Graham J, Ernst CC, Michelle Peavy K, Brown NN. Improving pregnancy outcomes among high-risk mothers who abuse alcohol and drugs: factors associated with subsequent exposed births. Child Youth Serv Rev 2014 Nov;46:11-18. [doi: 10.1016/j.childyouth.2014.07.014]
20. Hildebrandt UC, Graham JC, Grant TM. Predictors and moderators of improved social-emotional functioning in mothers with substance use disorders and their young children enrolled in a relationship-based case management program. Infant Ment Health J 2020 Sep;41(5):677-696. [doi: 10.1002/imhj.21872] [Medline: 32578238]
21. Miller JB, Stiver IP. A relational approach to understanding women's lives and problems. Psychiatr Ann 1993 Aug;23(8):424-431. [doi: 10.3928/0048-5713-19930801-07]
22. Prochaska JO, Velicer WF. The transtheoretical model of health behavior change. Am J Health Promot 1997;12(1):38-48. [doi: 10.4278/0890-1171-12.1.38] [Medline: 10170434]
23. Marlatt GA, Somers JM, Tapert SF. Harm reduction: application to alcohol abuse problems. NIDA Res Monogr 1993;137(137):147-166. [Medline: 8289918]
24. Stoner S. University of Washington Alcohol and Drug Abuse Institute. Parent-Child Assistance Program (PCAP): prevention & intervention with at-risk mothers and their children. 2025. URL: https://pcap.psychiatry.uw.edu/what-is-pcap/publications/ [accessed 2025-12-22]
25. Magidson JF, Jack HE, Regenauer KS, Myers B. Applying lessons from task sharing in global mental health to the opioid crisis. J Consult Clin Psychol 2019 Oct;87(10):962-966. [doi: 10.1037/ccp0000434] [Medline: 31556672]
26. Tandon SD, Johnson JK, Diebold A, et al. Comparing the effectiveness of home visiting paraprofessionals and mental health professionals delivering a postpartum depression preventive intervention: a cluster-randomized non-inferiority clinical trial. Arch Womens Ment Health 2021 Aug;24(4):629-640. [doi: 10.1007/s00737-021-01112-9] [Medline: 33655429]

XSL•FO
RenderX

27. Patel V. Scale up task-sharing of psychological therapies. Lancet 2022 Jan 22;399(10322):343-345. [doi: 10.1016/S0140-6736(21)02736-7] [Medline: 34929197]
28. Swartz HA, Grote NK, Graham P. Brief interpersonal psychotherapy (IPT-B): overview and review of evidence. Am J Psychother 2014;68(4):443-462. [doi: 10.1176/appi.psychotherapy.2014.68.4.443] [Medline: 26453346]
29. Grote NK, Katon WJ, Russo JE, et al. Collaborative care for perinatal depression in socioeconomically disadvantaged women: a randomized trial. Depress Anxiety 2015 Nov;32(11):821-834. [doi: 10.1002/da.22405] [Medline: 26345179]
30. Meffert SM, Neylan TC, McCulloch CE, et al. Interpersonal psychotherapy delivered by nonspecialists for depression and posttraumatic stress disorder among Kenyan HIV-positive women affected by gender-based violence: randomized controlled trial. PLoS Med 2021 Jan;18(1):e1003468. [doi: 10.1371/journal.pmed.1003468] [Medline: 33428625]
31. Guidelines for using rural-urban classification systems for community health assessment. : Washington State Department of Health; 2016 URL: https://doh.wa.gov/sites/default/files/legacy/Documents/1500/RUCAGuide.pdf [accessed 2025-12-15]
32. Grote N, Swartz H, A. Z. Interpersonal psychotherapy for women with depression living on low incomes. In: Markowitz JC, Weissman MM, editors. Casebook of Interpersonal Psychotherapy 2012:296-320.
33. Grote NK, Katon WJ, Lohr MJ, et al. Culturally relevant treatment services for perinatal depression in socio-economically disadvantaged women: the design of the MOMCare study. Contemp Clin Trials 2014 Sep;39(1):34-49. [doi: 10.1016/j.cct.2014.07.001] [Medline: 25016216]
34. Kroenke K, Spitzer RL. The PHQ-9: a new depression diagnostic and severity measure. Psychiatr Ann 2002 Sep;32(9):509-515. [doi: 10.3928/0048-5713-20020901-06]
35. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
36. Lang AJ, Stein MB. An abbreviated PTSD checklist for use as a screening instrument in primary care. Behav Res Ther 2005 May;43(5):585-594. [doi: 10.1016/j.brat.2004.04.005] [Medline: 15865914]
37. Sudom K. Evaluation of three abbreviated versions of the PTSD Checklist in Canadian Armed Forces personnel. Journal of Military, Veteran and Family Health 2020 Aug 1;6(2):9-16. [doi: 10.3138/jmvfh-2019-0062]
38. Kroenke K. Enhancing the clinical utility of depression screening. CMAJ 2012 Feb 21;184(3):281-282. [doi: 10.1503/cmaj.112004] [Medline: 22231681]
39. Lang AJ, Wilkins K, Roy-Byrne PP, et al. Abbreviated PTSD Checklist (PCL) as a guide to clinical response. Gen Hosp Psychiatry 2012;34(4):332-338. [doi: 10.1016/j.genhosppsych.2012.02.003] [Medline: 22460001]
40. Bischoff T, Anderson SR, Heafner J, Tambling R. Establishment of a reliable change index for the GAD-7. Psychol Community Health 2020;8(1):176-187. [doi: 10.5964/pch.v8i1.309]
41. Meeting your challenges. IBM. URL: http://www-03.ibm.com/software/products/en/spss-statistics [accessed 2016-10-04]
42. Ammerman RT, Putnam FW, Bosse NR, Teeters AR, Van Ginkel JB. Maternal depression in home visitation: a systematic review. Aggress Violent Behav 2010 May;15(3):191-200. [doi: 10.1016/j.avb.2009.12.002] [Medline: 20401324]
43. Dugravier R, Tubach F, Saias T, et al. Impact of a manualized multifocal perinatal home-visiting program using psychologists on postnatal depression: the CAPEDP randomized controlled trial. PLoS ONE 2013;8(8):e72216. [doi: 10.1371/journal.pone.0072216] [Medline: 23977257]
44. Suchman NE, DeCoste CL, McMahon TJ, Dalton R, Mayes LC, Borelli J. Mothering from the inside out: results of a second randomized clinical trial testing a mentalization-based intervention for mothers in addiction treatment. Dev Psychopathol 2017 May;29(2):617-636. [doi: 10.1017/S0954579417000220] [Medline: 28401850]

Abbreviations

CM: case manager
GAD-7: Generalized Anxiety Disorder-7 Scale
PCAP: Parent Child Assistance Program
PCL-6: PTSD (posttraumatic stress disorder) Checklist
PHF: promoting healthy families
PHQ-9: Patient Health Questionnaire-9
PTSD: posttraumatic stress disorder
SUD: substance use disorder

XSL•FO
RenderX

Edited by A Stone; submitted 28.03.25; peer-reviewed by M Hirschi; revised version received 01.12.25; accepted 01.12.25; published 30.12.25.

Please cite as:
Bhat A, Curran MC, Lohr MJ, Smith H, Blanchard B, Stoner SA, Grant T, Grote N
Integrating a Brief Behavioral Intervention Into Case Management for Mothers With Perinatal Substance Use Disorder: Nonrandomized Pilot Feasibility Study
JMIR Form Res 2025;9:e75132
URL: https://formative.jmir.org/2025/1/e75132
doi:10.2196/75132

© Amritha Bhat, Mary C Curran, Mary Jane Lohr, Haley Smith, Brittany Blanchard, Susan A Stoner, Therese Grant, Nancy Grote. Originally published in JMIR Formative Research (https://formative.jmir.org), 30.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Evaluation of the 2020 American Urological Association Microscopic Hematuria Guidelines in Clinical Practice: Retrospective Chart Review Analysis

Dominique Munroe[1], MD, MPH; James O'Keefe[2], MD; Danyang Wang[2], MPH; Miranda A Moore[2,3], PhD

[1]Department of Family Medicine, Atrium Health Navicent, 777 Hemlock St, Macon, GA, United States

[2]Department of Medicine, Emory University, Atlanta, GA, United States

[3]Department of Family and Preventive Medicine, Emory University, Atlanta, GA, United States

Corresponding Author:
Dominique Munroe, MD, MPH
Department of Family Medicine, Atrium Health Navicent, 777 Hemlock St, Macon, GA, United States

Abstract

Background: Hematuria is one of the most common urologic diseases seen within clinical practice, with a prevalence range of 1.7% - 31.1%. In 2020, American Urological Association (AUA) guidelines were revised and recommend that following initial evaluation, clinicians should categorize patients into three tiers (low risk, intermediate risk, and high risk) based on various factors. Recent literature has shown that the AUA guidelines have high clinical utility when compared to other international guidelines such as those outlined by the Hematuria Risk Index, Canadian Urological Association, and Kaiser Permanente; however, this guideline remains unvalidated among the population of "well adults" within the United States.

Objective: We used a retrospective study design to evaluate data abstracted from the electronic medical records of patients seen in the Emory Healthcare Executive Health Clinic from September 29, 2017, to January 29, 2021, to investigate the utility of risk stratification as a tool for clinical decision-making.

Methods: According to the AUA risk stratification system, patients were stratified into low-risk and intermediate-risk/high-risk groups based on sex, age, smoking history, history of gross hematuria, and red blood cells/high-powered field. The frequencies and percentages of different causes of hematuria across the three risk strata were reported.

Results: Of the 882 instances of red blood cells in urine (URBC) ≥3, a total of 368 (41.72%) underwent a repeat analysis within a 6-month time span, 184 (20.86%) within a 12-month time span, and 330 (37.41%) at >12 months. Instances of a URBC <3 (N=1643) were more likely to have no urologic diagnosis—1503 (91.48%) in comparison to 633 (76.27%) for those instances with a URBC >3 (N=830). Ultimately, 23 (100%) participants in the low-risk group had no urologic diagnosis after urinalysis versus 608 (75.62%) in the intermediate-risk/high-risk group (N=804).

Conclusions: We found a need for a greater focus on monitoring elevated URBC counts, in accordance with clinical guidelines for managing hematuria in low-risk patients. Future research should examine the impact of risk stratification on clinical decisions and access to care, especially in underserved populations. It should also assess how the new AUA guidelines affect physician referral patterns and explore real-world implementation challenges and facilitators.

(JMIR Form Res 2025;9:e75929) doi:10.2196/75929

KEYWORDS

hematuria; cystoscopy; guideline; risk stratification; urinalysis

Introduction

The presence of blood in the urine (hematuria) may indicate the presence of underlying diseases such as infections, kidney stones, or cancer of the urinary system [1,2]. Hematuria is one of the most common urologic diseases, accounting for over 20% of urological evaluations [3]. Other studies have demonstrated a prevalence range between 2.4% and 3.1% among healthy volunteers [4]. The two categories of hematuria observed in the literature are gross (or visible) hematuria and microscopic hematuria (MH). There is general consensus that the presence of gross hematuria necessitates a complete evaluation for urologic cancers [4-6], whereas MH is often an incidental finding, and the medical community has not adhered to universal diagnostic evaluation recommendations [7,8].

In 2012, the American Urological Association (AUA) created a recommendation to guide the medical care of patients presenting with MH, recommending further medical testing for all patients over age 35 who do not have a benign cause identified (eg, menstrual period, vigorous exercise, or infection)

XSL•FO
RenderX

[4]. Contrary to AUA recommendations at that time, other professional organizations created guidelines that recommend against screening for adults at low cancer risk [2,9]. In 2020, these AUA guidelines were revised [5]. Among the revisions was the recommendation that following initial evaluation, clinicians should categorize patients into 3 tiers (low risk, intermediate risk, and high risk) based on various factors. Patients who were considered intermediate risk or high risk were recommended to undergo further evaluation, while those who were low risk were recommended to undergo a repeat urinalysis within 6 months or undergo other testing.

Use of the 2020 AUA guidelines has been shown to be a cost-effective alternative to more invasive testing through its unique ability to prioritize more costly and invasive workups for patients at greatest risk [8,10,11].

However, despite its utility, this guideline remains unvalidated. The goal of this study is to add to the growing literature surrounding the utility of the AUA index, further guiding clinical practice in the population of "well adults" while improving evaluation rates for high-risk patients and decreasing rates of unnecessary testing for low-risk patients.

Methods

Study Design

Data were abstracted from the electronic medical records of patients seen in the Emory Healthcare Executive Health Clinic from September 29, 2017, to January 29, 2021. The Executive Health Clinic's standard annual physical includes a routine urine collection, thus providing a population of patients with sufficient observations to adequately analyze the impact of the AUA guideline changes. We collected data on demographics, smoking history, results of urinalysis, instances of follow-up imaging or cystoscopy after urinalysis, history of gross hematuria, and urologic diagnoses of the patients.

The results of urinalysis were categorized into four groups according to degree of hematuria: 0 - 2, 3 - 10, 11 - 25, and >25 red blood cells (RBC)/high-powered field (HPF). For each group, the instances and percentages of initial repeat urinalysis, imaging, and cystoscopy within 6 months or 12 months or after 12 months were reported. The original four groups were further consolidated into two categories—URBC (red blood cells in urine) <3 and URBC ≥3—to allow more streamlined and interpretable evaluation. Additionally, this method also ensured sufficient sample size within each category to support a robust statistical analysis.

Although the AUA risk stratification system (Textbox 1) [5] stratifies patients into low-, intermediate-, and high-risk groups based on sex, age, smoking history, history of gross hematuria, and RBC/HPF, the main distinction between intermediate- and high-risk patients is the number of smoking pack-years. Unfortunately, our electronic health record does not provide adequate information to determine the number of pack years for our patients. Thus, we combined the intermediate- and high-risk groups for our analysis. Patients with any of the following were classified as being in the intermediate-/high-risk group: women age ≥50 years or men age ≥40 years; ≥10 pack-years smoking; >10 RBC/HPF on one urinalysis; history of gross hematuria; and previously low-risk with no prior evaluation and ≥3 RBC/HPF on repeat urinalysis. The causes of hematuria were categorized as stones, infections, glomerular noninfections, benign prostate diseases, other urologic benign diseases, malignant bladder diseases, malignant kidney diseases, malignant prostate diseases, malignant ureter diseases, other urologic malignant diseases, and nonurologic diseases. The frequencies and percentages of different causes of hematuria across the 4 hematuria degrees or risk strata were also reported. To calculate an effect size for final diagnosis by initial URBC value and risk stratification, we computed Cohen *h* to quantify the difference between two proportions.

A new urine collection technique (midstream urine collection) for males has been applied since November 21, 2019. The impact of this technique change on urinalysis results among males versus females was examined using 2-tailed χ^2 tests, 2, with $P<.05$ indicating statistical significance.

XSL•FO
RenderX

Textbox 1. American Urological Association risk stratification guidelines 2020.

Low (patient meets all criteria)

- Women age <50 years or men age <40 years
- Never smoker or <10 pack-years
- 3 - 10 red blood cells (RBC)/high-powered field (HPF) on a single urinalysis
- No risk factors for urothelial cancer

Intermediate (patient meets any one of these criteria)

- Women age 50 - 59 years or men age 40 - 59 years
- 10 - 30 pack-years
- 11 - 25 RBC/HPF on a single urinalysis
- Low-risk patient with no prior evaluation and 3 - 10 RBC/HPF on repeat urinalysis
- Additional risk factors for urothelial cancer

High (patient meets any one of these criteria)

- >30 pack-years
- Women or men age ≥60 years
- >25 RBC/HPF on a single urinalysis
- History of gross hematuria

Ethical Considerations

Emory University's institutional review board approved this retrospective study in January 2022 (protocol number 2022P004116). All data analysis was performed using StataSE 14 (StataCorp LLC) [12].

Results

A total of 2677 unique patients gave 5072 urine samples at the Emory Healthcare Executive Health Clinic from September 29, 2017, to January 29, 2021. The mean age of the cohort was 58.07 years, with 1770 (73.9%) participants identifying as male and 625 (26.1%) identifying as female.

Overall, 3700 (72.97%) instances of a URBC count of <3 were observed within this cohort compared to 1372 (27.05%) instances of URBC ≥3 (Table 1). Of the total 5072 instances observed, 2677 were unique patients, whereas the remaining instances were accounted for by patients who appeared multiple times within the data.

Table . Initial instances of red blood cells in urine. There were 2677 unique patients seen in the Emory Healthcare Executive Health Clinic from September 29, 2017, to January 29, 2021 (N=5072 urine samples).

Red blood cells in urine count	Value, n	Percentage
<3	3700	72.95
≥3	1372	27.05

Of the 1795 instances of URBC <3, there were 287 (15.97%) that were offered a repeat within 6 months (Table 2), 487 (27.13%) within 12 months, and 1021 (56.88%) more than 12 months since the initial urinalysis. In comparison, 882 of the instances of URBC ≥3 underwent a repeat analysis, with 368 (41.72%) being referred within a 6-month time span, 184 (20.86%) within 12 months, and 330 (37.41%) more than 12 months later.

Table . Initial urinalysis result and referral timeline. There were 2677 unique patients seen in the Emory Healthcare Executive Health Clinic from September 29, 2017, to January 29, 2021.

	URBC[a] <3 (n=1795)		URBC ≥3 (n=882)	
	Value, n	Mean %	Value, n	Mean %
Within 6 months	287	15.99	368	41.72
Within 12 months	487	27.12	184	20.86
Over 12 months	1021	56.88	330	37.41

[a]URBC: red blood cells in urine.

XSL•FO
RenderX

Regarding referral patterns following initial urinalysis results, those with URBC ≥3 were referred for follow-up imaging and procedures 40.21% (421/1047) of the time in comparison to 30.38% (371/1221) for those with URBC <3 (Table 3). Instances with URBC <3 were responsible for 20% (21/105) of cystoscopies and 52.78% (171/324) of any imaging within the 6-month time frame; 25% (32/130) of cystoscopies and 55% (219/396) of any imaging in the <12-month time frame; and 63% (51/81) of cystoscopies and 66% (144/219) of any imaging within the >12-month time frame (Table 4).

Table . Number (%) of initial screenings via imaging and procedures after urinalysis by timing and urinalysis results. There were 2677 unique patients seen in the Emory Healthcare Executive Health Clinic from September 29, 2017, to January 29, 2021.

	URBC[a] <3 (n=1221)		URBC ≥3 (n=1047)	
	Values, n	Percentage	Values, n	Percentage
Within 6 months				
Any imaging	171	14	153	14.61
CT[b]	89	7.29	97	9.26
MRI[c]	26	2.13	18	1.72
USG[d]	56	4.59	38	3.63
Cystoscopy	21	1.72	84	8.02
Any imaging + cystoscopy	8	0.66	31	2.96
Within 12 months				
Any imaging	219	17.94	177	16.91
CT	118	9.66	110	10.51
MRI	37	3.03	23	2.2
USG	64	5.24	44	4.2
Cystoscopy	32	2.62	98	9.36
Any imaging + cystoscopy	20	1.64	38	3.63
Over 12 months				
Any imaging	144	11.79	75	7.16
CT	72	5.90	39	3.72
MRI	32	2.62	15	1.43
USG	40	3.28	21	2.01
Cystoscopy	51	4.18	30	2.87
Any imaging + cystoscopy	21	1.72	13	1.24

[a]URBC: red blood cells in urine.

[b]CT: computed tomography.

[c]MRI: magnetic resonance imaging.

[d]USG: ultrasonography.

XSL•FO
RenderX

Table . Number (%) of diagnoses of patients after urinalysis by URBC[a] category. There were 2677 unique patients seen in the Emory Healthcare Executive Health Clinic from September 29, 2017, to January 29, 2021.[b]

Final diagnosis	URBC <3 (n=1643)		URBC ≥3 (n=830)	
	Values	Percentage	Values	Percentage
No urologic diagnosis	1503	91.48	633	76.27
Stones	24	1.46	42	5.06
Infection	15	0.91	28	3.37
Glomerular noninfection	1	0.06	2	0.24
Benign prostate	45	2.74	59	7.11
Benign urologic other[c]	32	1.95	40	4.82
Malignant bladder	1	0.06	2	0.24
Malignant kidney	4	0.24	5	0.60
Malignant prostate	18	1.10	15	1.81
Malignant ureter	0	0.00	2	0.24
Malignant urologic other[c]	0	0.00	1	0.12

[a]URBC: red blood cells in urine.

[b]Cohen *h*=0.425.

[c]Site not specified.

Instances of URBC <3 were more likely to have no urologic diagnosis (1503/1643, 91.48%) compared to those instances with a URBC >3 (633/830, 76.27%; Table 4). The Cohen *h* of 0.425 indicates that there is a medium-sized difference in the proportion of patients with no urologic diagnosis between those with URBC <3 and those with URBC >3. Of urologic problems identified for instances of URBC <3, the most common diagnoses were benign prostate, benign urologic (site not specified), and stones (2.74% vs 1.95% vs 1.46%, respectively; Table 4). Further stratification of instances with ≥3 RBC/HPF revealed that 23 instances would have been classified as low risk and 804 (97%) as intermediate/high risk (Table 5). Most notably, in the low-risk instances, 100% had no urologic diagnosis after urinalysis versus 75.62% in the intermediate-/high-risk group. The Cohen *h* of 1.033 indicates that there is a large-sized difference in the proportion of patients with no urologic diagnosis between those with low risk and those with high risk.

Table . Number (%) of diagnoses of patients after urinalysis by risk stratification[a].

	Low risk (n=23)		Intermediate/high risk (n=804)	
	Value, n	Percentage	Value, n	Percentage
No urologic diagnosis	23	100	608	75.62
Stones	0	0	42	5.22
Infection	0	0	27	3.36
Glomerular noninfection	0	0	2	0.25
Benign prostate	0	0	59	7.34
Benign urologic other[a]	0	0	40	4.98
Malignant bladder	0	0	3	0.37
Malignant kidney	0	0	5	0.62
Malignant prostate	0	0	15	1.87
Malignant ureter	0	0	2	0.25
Malignant urologic other[b]	0	0	1	0.12

[a]781 unique patients with ≥3 red blood cells/high-powered field on at least one urinalysis were stratified. Cohen *h*=1.033.

[b]Site not specified.

Significant differences were observed in the detection of URBC <3 RBC/HPF in males after implementation of the new urine collection technique (Table 6). We found that of 2763 male samples collected, 1812 (71.14%) males had a URBC <3 after

XSL•FO
RenderX

the new technique in comparison to 951 (92.98%) after the new technique (*P*<.001). Of the 930 female samples collected, 586 (65.4%) had a URBC <3 before the technique change versus 351 (72.67%) after the technique change (*P*=.006; Table 6).

Table . Differences in frequencies of red blood cells in urine ≥3 before versus after the new collection technique for males.

	Before technique change (n=3443)		After technique change (n=1629)		Before versus after difference *P* value
	Value, n	Percentage	Value, n	Percentage	
Male (n=2763)					
<3 RBC/HPF[a]	1812	71.14	951	82.98	<.001
≥3 RBC/HPF	735	28.86	195	17.02	
Female (n=930)					
<3 RBC/HPF	586	65.4	351	72.67	.006
≥3 RBC/HPF	310	34.6	132	27.33	

[a]RBC/HPF: red blood cells/high-powered field.

Discussion

Our results showing most patients with URBC ≥3 underwent repeat testing indicates a stronger emphasis on monitoring elevated URBC counts, which aligns with the clinical guideline for managing hematuria in low-risk patients [1]. However, the extended follow-up period observed raises questions concerning the adherence to management guidelines, which recommend earlier repeat testing to rule out significant urologic pathology [5]. These findings are consistent with research showing a lack of consistency with the evaluation of hematuria across multiple health care settings; it has been observed that physicians often do not refer for appropriate urologic investigation [7,13]. Furthermore, when looking at a large managed-care organization, out of the 456,674 patients diagnosed with MH, 1.7% were seen by a urologist [14]. Even more troubling, disparities surrounding urologic referral rates have been shown for age, sex, and race [13,15-18]. These disparities indicate the need for nationally recognized algorithms such as the AUA guidelines. When used correctly, these clinical guidelines have been shown to result in standardized referral patterns and timelines across diverse practice settings, reducing variability and promoting consistent patterns in patient care [19-22].

The use of risk factors for bladder cancer such as age, male, sex, smoking, and gross hematuria as stratification tools has been associated with increased incidence in various studies [14,23]. Our results, which showed that patients who were placed in the low-risk group with URBC ≥3 had no subsequent urologic diagnosis in comparison to the intermediate-risk and high-risk group, further provide support for the use of these factors as the foundation of clinical decision-making.

Our low incidence of cancer after urinalysis is consistent with previous research showing low rates of malignancy detection after evaluation of MH patients [23,24] and further strengthens the argument against unnecessary imaging and cystoscopies. Unnecessary invasive procedures can have potential side effects such as allergic reactions, increased cost, and radiation-induced malignancy [23,25,26]. Refining the MH guidelines reduces unnecessary invasive procedures in low-risk patients and avoids the potential harms of using unnecessary contrast-enhanced imaging.

The implementation of the midstream urine collection technique leading to a decrease of URBC >3 suggests that the technique change may be more effective in reducing false-positive findings of hematuria in males; however, as no adjustments for potential confounders were made, these results should be interpreted with caution. Although the improvement in females was statistically significant, the effect size was smaller, indicating that other factors may influence URBC levels in women or that the technique may be less effective within this population. This is consistent with research that showed that catheterization may be a better method to diagnose and screen for MH in women [27]. For example, a study showed that significantly fewer red blood and squamous epithelial cells were found in catheterized urinalysis samples. Furthermore, when applied as part of the criteria for a properly collected sample, the positive predictive value of referral urinalysis increased by 22.7% for true MH [27].

One of the primary limitations of our study was its retrospective design, which inherently restricts the ability to draw causal conclusions; its reliance on preexisting data may introduce biases [28].

Our patient population, consisting exclusively of individuals receiving care through the Executive Health Clinic, possesses substantial advantages such as reliable access to comprehensive preventive services, follow-up visits, high-quality private insurance, greater health literacy, and fewer barriers to care. In contrast, the broader and more diverse US patient population—many of whom may face significant health disparities and systemic barriers to care—may experience markedly different health trajectories. As a result, the associations observed in our cohort may be confounded by these advantages, potentially creating the illusion of stronger relationships between risk factors and outcomes that may be observed when examined in less advantaged populations.

As we needed an adequate number of urine samples, we used all data for these Executive Health Clinic patients; thus, our study is subject to selection and confounding bias from this

XSL•FO
RenderX

population. Specifically, as our population is comprised of individuals who are fully employed, highly insured, and of high socioeconomic status, our findings may not be generalizable to other populations. Another limitation was the lack of complete smoking data, which was unavailable to the research team. In the absence of a full smoking history, we used a proxy measure that potentially skewed results related to smoking-related risks and affected the study's reliability and accuracy. Furthermore, without accurate pack-year data, patients may have been incorrectly stratified into intermediate risk rather than high risk, thus diminishing observed differences in disease prevalence and health outcomes and minimizing the true predictive value of the AUA 3- tier module.

Combining intermediate- and high- risk patients into a single group may limit the comprehensiveness and granularity of the risk stratification analysis. Furthermore, our inclusion criteria were based on whether or not patients completed urine collection regardless of known or suspected urologic disease. This approach may have potentially influenced the risk stratification analysis. Finally, our study was primarily descriptive and limited in scope. As a result, we conducted no multivariable adjustments for potential confounders, thus limiting the ability to draw stronger inferences about differences across collection methods and risk groups.

The lack of referrals from primary care providers could have been associated with several external factors, such as the ongoing COVID-19 pandemic and the related health care protocols in place during the final year of the study [29,30]. Significant disruptions in health care services were observed during this period and resulted in a limited ability for providers to offer referrals as a part of routine care, thus leading to underreporting or delays in referral patterns. Finally, due to the relatively small sample size and specific demographic composition, the findings from our risk stratification may not be generalizable to the broader US population.

Future work should continue to investigate the effects of risk stratification on clinical decision-making and access to care, specifically in diverse and underserved populations. Additionally, it is crucial to identify the effect of the new AUA guidelines on physician referral patterns, particularly in relation to the aforementioned gender and race disparities. Investigating the real-world application of these guidelines would provide valuable insight into the barriers and facilitators of implementation on a broader scale.

Acknowledgments

The authors declare no use of generative artificial intelligence in the research and writing process. Responsibility for the final manuscript lies entirely with the authors.

Funding

This publication was supported by the Emory Academic Internal Medicine Center with donations from the O. Wayne Rollins Foundation.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

All authors agree to be accountable for all aspects of the work. Concept and design: JO. Acquisition, analysis, or interpretation of data: JO, MAM. Drafting of the manuscript: all authors. Critical revision of the manuscript for important intellectual content: all authors. Statistical analysis: DW, MAM. Obtained funding: JO. Review and approval of final submitted manuscript: all authors.

Conflicts of Interest

MAM reports receiving funding for other work from National Institutes of Health, Agency for Healthcare Research and Quality's, Health Resources and Services Administration, and Georgia State Department of Human Services. No conflict-of-interest disclosures were reported by the other authors of this paper.

References

1. Bolenz C, Schröppel B, Eisenhardt A, Schmitz-Dräger BJ, Grimm MO. The investigation of hematuria. Dtsch Arztebl Int 2018 Nov 30;115(48):801-807. [doi: 10.3238/arztebl.2018.0801] [Medline: 30642428]
2. Nielsen M, Qaseem A, High Value Care Task Force of the American College of Physicians. Hematuria as a marker of occult urinary tract cancer: advice for high-value care from the American College of Physicians. Ann Intern Med 2016 Apr 5;164(7):488-497. [doi: 10.7326/M15-1496] [Medline: 26810935]
3. Mariani AJ, Mariani MC, Macchioni C, Stams UK, Hariharan A, Moriera A. The significance of adult hematuria: 1,000 hematuria evaluations including a risk-benefit and cost-effectiveness analysis. J Urol 1989 Feb;141(2):350-355. [doi: 10.1016/s0022-5347(17)40763-4] [Medline: 2492350]
4. Davis R, Jones JS, Barocas DA, et al. Diagnosis, evaluation and follow-up of asymptomatic microhematuria (AMH) in adults: AUA guideline. J Urol 2012 Dec;188(6 Suppl):2473-2481. [doi: 10.1016/j.juro.2012.09.078] [Medline: 23098784]

XSL•FO
RenderX

5. Barocas DA, Boorjian SA, Alvarez RD, et al. Microhematuria: AUA/SUFU guideline. J Urol 2020 Oct;204(4):778-786. [doi: 10.1097/JU.0000000000001297] [Medline: 32698717]
6. Loo R, Whittaker J, Rabrenivich V. National practice recommendations for hematuria: how to evaluate in the absence of strong evidence? Perm J 2009;13(1):37-46. [doi: 10.7812/TPP/08-083] [Medline: 21373244]
7. Elias K, Svatek RS, Gupta S, Ho R, Lotan Y. High-risk patients with hematuria are not evaluated according to guideline recommendations. Cancer 2010 Jun 15;116(12):2954-2959. [doi: 10.1002/cncr.25048] [Medline: 20564400]
8. Georgieva MV, Wheeler SB, Erim D, et al. Comparison of the harms, advantages, and costs associated with alternative guidelines for the evaluation of hematuria. JAMA Intern Med 2019 Oct 1;179(10):1352-1362. [doi: 10.1001/jamainternmed.2019.2280] [Medline: 31355874]
9. Nager C, Sung V, Whitteside J. Committee opinion no.703. Asymptomatic microscopic hematuria in women. Obstet Gynecol 2017 Jun;129:6. [doi: 10.1097/AOG.0000000000002059]
10. Gold SA, Kenigsberg AP, Lotan Y. Diagnostic and cost implications of the 2020 AUA microhematuria guidelines: modeling impact in a large public health care system. J Urol 2022 Jan;207(1):52-60. [doi: 10.1097/JU.0000000000002175] [Medline: 34428924]
11. Halpern JA, Chughtai B, Ghomrawi H. Cost-effectiveness of common diagnostic approaches for evaluation of asymptomatic microscopic hematuria. JAMA Intern Med 2017 Jun 1;177(6):800-807. [doi: 10.1001/jamainternmed.2017.0739] [Medline: 28418451]
12. Stata. 2019. URL: https://www.stata.com [accessed 2025-11-24]
13. Johnson EK, Daignault S, Zhang Y, Lee CT. Patterns of hematuria referral to urologists: does a gender disparity exist? Urology 2008 Sep;72(3):498-502. [doi: 10.1016/j.urology.2008.01.086] [Medline: 18619657]
14. Loo RK, Lieberman SF, Slezak JM, et al. Stratifying risk of urinary tract malignant tumors in patients with asymptomatic microscopic hematuria. Mayo Clin Proc 2013 Feb;88(2):129-138. [doi: 10.1016/j.mayocp.2012.10.004] [Medline: 23312369]
15. Klaassen Z, DiBianco JM, Jen RP, et al. Female, Black, and unmarried patients are more likely to present with metastatic bladder urothelial carcinoma. Clin Genitourin Cancer 2016 Oct;14(5):e489-e492. [doi: 10.1016/j.clgc.2016.04.006] [Medline: 27212042]
16. Bassett JC, Alvarez J, Koyama T, et al. Gender, race, and variation in the evaluation of microscopic hematuria among Medicare beneficiaries. J Gen Intern Med 2015 Apr;30(4):440-447. [doi: 10.1007/s11606-014-3116-2] [Medline: 25451992]
17. Cohn JA, Vekhter B, Lyttle C, Steinberg GD, Large MC. Sex disparities in diagnosis of bladder cancer after initial presentation with hematuria: a nationwide claims-based investigation. Cancer 2014 Feb 15;120(4):555-561. [doi: 10.1002/cncr.28416] [Medline: 24496869]
18. Ark JT, Alvarez JR, Koyama T, et al. Variation in the diagnostic evaluation among persons with hematuria: influence of gender, race and risk factors for bladder cancer. J Urol 2017 Nov;198(5):1033-1038. [doi: 10.1016/j.juro.2017.06.083] [Medline: 28655530]
19. Clinical guidelines and standardization of practice to improve outcomes. ACOG. URL: https://www.acog.org/clinical/clinical-guidance/committee-opinion/articles/2019/10/clinical-guidelines-and-standardization-of-practice-to-improve-outcomes [accessed 2025-11-02]
20. Bolton Saghdaoui L, Lampridou S, Tavares S, et al. Interventions to improve referrals from primary care to outpatient specialist services for chronic conditions: a systematic review and framework synthesis update. Syst Rev 2025 May;14(1):103. [doi: 10.1186/s13643-025-02841-z]
21. Clarke A, Blundell N, Forde I, et al. Can guidelines improve referral to elective surgical specialties for adults? A systematic review. Quality and Safety in Health Care 2010 Jun 1;19(3):187-194. [doi: 10.1136/qshc.2008.029918]
22. McLachlan S, Kyrimi E, Dube K, Hitman G, Simmonds J, Fenton N. Towards standardisation of evidence-based clinical care process specifications. Health Informatics J 2020 Dec;26(4):2512-2537. [doi: 10.1177/1460458220906069] [Medline: 32186428]
23. Rabinowitz DE, Wood AM, Marziliano A, et al. Use of a risk stratification tool to guide evaluation of patients with asymptomatic microscopic hematuria. Urology 2020 Jul;141:27. [doi: 10.1016/j.urology.2020.03.045]
24. Committee opinion no.703: asymptomatic microscopic hematuria in women. Obstet Gynecol 2017;129(6):e168-e172. [doi: 10.1097/AOG.0000000000002059]
25. Aoki Y, Akemura T. Allergies correlated to adverse reactions induced by non-ionic monomeric and ionic dimeric contrast media for contrast enhanced CT examination(scientific paper) [Article in Japanese]. J-Stage 2002;58(9):1245-1251. [doi: 10.6009/jjrt.KJ00001364804]
26. Cochran ST, Bomyea K, Sayre JW. Trends in adverse events after IV administration of contrast media. AJR Am J Roentgenol 2001 Jun;176(6):1385-1388. [doi: 10.2214/ajr.176.6.1761385] [Medline: 11373197]
27. Chen A, Caron A, Jackson NJ, et al. Defining properly collected urine: thresholds to improve the accuracy of urinalysis for microscopic hematuria evaluation in women. J Urol 2022 Feb;207(2):385-391. [doi: 10.1097/JU.0000000000002200] [Medline: 34544262]
28. Talari K, Goyal M. Retrospective studies - utility and caveats. J R Coll Physicians Edinb 2020 Dec;50(4):398-402. [doi: 10.4997/JRCPE.2020.409] [Medline: 33469615]

XSL•FO
RenderX

29. Czeisler M, Marynak K, Clarke KEN, et al. Delay or avoidance of medical care because of COVID-19-related concerns - United States, June 2020. MMWR Morb Mortal Wkly Rep 2020 Sep 11;69(36):1250-1257. [doi: 10.15585/mmwr.mm6936a4] [Medline: 32915166]
30. Gertz AH, Pollack CC, Schultheiss MD, Brownstein JS. Delayed medical care and underlying health in the United States during the COVID-19 pandemic: a cross-sectional study. Prev Med Rep 2022 Aug;28:101882. [doi: 10.1016/j.pmedr.2022.101882] [Medline: 35813398]

Abbreviations

AUA: American Urological Association
HPF: high-powered field
MH: microscopic hematuria
RBC: red blood cell
URBC: red blood cells in urine

Edited by J Sarvestan; submitted 13.04.25; peer-reviewed by JD Souza, WW Hsu; revised version received 02.11.25; accepted 06.11.25; published 04.12.25.

Please cite as:
Munroe D, O'Keefe J, Wang D, Moore MA
Evaluation of the 2020 American Urological Association Microscopic Hematuria Guidelines in Clinical Practice: Retrospective Chart Review Analysis
JMIR Form Res 2025;9:e75929
URL: https://formative.jmir.org/2025/1/e75929
doi:10.2196/75929

© Dominique Munroe, James O'Keefe, Danyang Wang, Miranda A Moore. Originally published in JMIR Formative Research (https://formative.jmir.org), 4.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

“The Women’s PrEP Project,” a Clinic-Based, Socio-Structural Intervention to Improve the Provision of Preexposure Prophylaxis for Cisgender Women: Interrupted Time Series Pilot

Rachel K Scott[1,2,3], MD, MPH; Shawnika J Hull[4], PhD; Deanna Kerrigan[5], PhD; Yan Wang[5], MD, DrPH; Mandi Pratt-Chapman[6], PhD; Tara Mathias-Prabhu[1], MD; Naquia Unwala[1], BA; Marisa Sadauskas[7], BS; Bat-zion Hose[8], PhD; Marjanna Smith[9], MPH; Patricia Moriarty[2], MSHS; Tranessa Hanson[2], BA; Ariam Tedla[2], MHA; Pamela Lotke[1,3], MD; Peggy Ye[1,3], MD; Hannah Arem[9], PhD

[1]Department of Obstetrics and Gynecology, School of Medicine, Georgetown University, Washington, DC, United States

[2]Women's and Children's Research Network, MedStar Health Research Institute, MedStar Health, 110 Irving St NW, East Building 5108, Washington, DC, United States

[3]Women's and Infants' Services Department, MedStar Washington Hospital Center, Washington, DC, United States

[4]Department of Communication, School of Communication & Information, Rutgers, The State University of New Jersey, New Brunswick, NJ, United States

[5]Department of Prevention and Community Health, Milken Institute School of Public Health, George Washington University, Washington, DC, United States

[6]Center for Patient-Centered Initiatives and Health Equity, Cancer Center, George Washington University, Washington, DC, United States

[7]School of Medicine, Saint Louis University, Saint Louis, MO, United States

[8]MedStar Health National Center for Human Factors in Healthcare, Medstar Health Research Institute, Washington, DC, United States

[9]Implementation Science, Healthcare Delivery Research Program, MedStar Health Research Institute, Washington, DC, United States

Corresponding Author:
Rachel K Scott, MD, MPH
Department of Obstetrics and Gynecology, School of Medicine, Georgetown University, Washington, DC, United States

Abstract

Background: Cisgender women account for 23% of new HIV diagnoses in the United States, but there are significant socio-structural barriers to engagement and retention in the preexposure prophylaxis (PrEP) cascade, particularly for women of color.

Objective: In response to the lack of evidence-based interventions to improve PrEP initiation, adherence, and persistence among women in the United States, we developed and piloted a clinic-based, socio-structural intervention to measure (1) the feasibility of delivering the adapted intervention and (2) clinic team and patient perspectives on the intervention, in preparation for a future trial on engagement and retention in the PrEP cascade among women.

Methods: We previously applied the ADAPT-ITT (Assessment, Decision, Adaptation, Production, Topical experts, Integration, Training, Testing) model to develop a culturally appropriate, evidence-based intervention, responsive to Black women’s HIV prevention needs. In the present study, we set out to complete the *Training* and *Testing* phases: namely, to hire and train a PrEP navigator and to train clinic staff to deliver the adapted intervention. We completed a 4-month pilot to assess the feasibility of delivering the intervention and collecting outcomes of interest and initial outcome trends (compared to baseline). We further assessed the clinic team and patient perspectives on the intervention to understand the potential for future scale and delivery.

Results: The clinic team participants found the adapted Women’s PrEP project (W-PrEP) intervention both highly feasible and relevant to patients, with minimal impact on clinic flow. Patient participants reported that the W-PrEP intervention was highly relevant and appreciated the education and counseling from the PrEP navigator—many learning about PrEP for the first time from the navigator and health care provider. The outcome measures were both feasible to collect and appropriate to capture the primary outcomes of interest. Finally, the W-PrEP intervention increased the proportion of patients counseled about PrEP from 65% to 76% of patients seen ($P<.001$).

Conclusions: The intervention and outcome data collection was feasible, and open-ended clinic team and patient perspectives showed positive feedback about the intervention and relevance. The associated increase in PrEP counseling is promising but necessitates further evaluation of the effects of the W-PrEP intervention on the PrEP cascade (eg, initiation, persistence, adherence) in a larger randomized trial.

XSL•FO
RenderX

(JMIR Form Res 2025;9:e80653) doi:10.2196/80653

KEYWORDS

preexposure prophylaxis; female; primary prevention; education; HIV infections

Introduction

Background

Cisgender women (subsequently referred to as "women") account for approximately 6400 preventable new HIV diagnoses in the United States annually, underscoring the unrealized opportunity for HIV prevention programs specifically tailored for women [1]. The HIV epidemic in the United States disproportionately affects women of color and of reproductive age [2-4]. In 2019, despite only accounting for 12% of the population, Black women accounted for 54% of new HIV cases among women; reproductive-age women (25 - 44 y) comprised 51% [1]. HIV transmission occurs primarily through receptive vaginal heterosexual transmission (84%) among women [1,5] and is preventable through consistent use of barrier protection and HIV preexposure prophylaxis (PrEP) [6-12]. Barrier protection use, specifically male condoms, is limited by partner power discordance and is dependent upon partner cooperation [13], highlighting the importance of women-controlled HIV prevention options, such as PrEP.

PrEP with daily oral tenofovir disoproxil fumarate/emtricitabine, injectable cabotegravir bimonthly, or injectable lenacapavir biannually reduces HIV transmission by up to 100% in women [7-12]. Additional pharmacoprevention options are in the regulatory pipeline [14,15]. Both daily oral and longer-acting PrEP formulations offer women-controlled options for HIV prevention prior to risk exposure, circumventing the need for cooperation from sex partners. Despite the high efficacy, PrEP is underutilized by women, particularly women of color, in the United States [1,16-19]. The number of prescriptions of PrEP to women in the United States is dwarfed by the number of women with indications for PrEP; there is a considerable unmet need for pharmacoprevention among women [1,16-19]. While the absolute number of women initiating PrEP has increased from 3695 in 2012 to 28,870 in 2021, PrEP uptake among Black women lags far behind other groups [18-20]. Between 2014 and 2017, Black women comprised only 33% of women who initiated PrEP nationally; in contrast, they represented the majority (59%) of new HIV diagnoses among US women in 2017 [17]. Evidence-based interventions are urgently needed to improve PrEP utilization and HIV prevention among women.

Intervention Development

In response to the lack of evidence-based interventions to improve PrEP initiation, adherence, and persistence among women in the United States, we adapted 2 Centers for Disease Control and Prevention (CDC) Best Practices for HIV Prevention for women (HIV PrEP Services for Urban Women [21,22] and Project Shikamana [21,23,24]) and developed the Women's PrEP Project (W-PrEP), a clinic-based, socio-structural intervention [25]. The W-PrEP intervention was designed to provide education, resources, and support to empower and enable sexual and reproductive health clinics to provide high-quality, nonstigmatizing, universal PrEP services to women who may have increased exposure to HIV, using the ADAPT-ITT (Assessment, Decision, Adaptation, Production, Topical Experts, Integration, Training, Testing) model and guided by the Socioecological Framework [25-27]. We previously published on the formative work for the first 6 ADAPT-ITT steps of this project [25].

In this paper, we describe the last two phases of the ADAPT-ITT process, namely *Training* peer navigators and clinic staff and pilot *Testing* of the feasibility and acceptability of the intervention. The mixed methods pilot data collection reported in both manuscripts was guided by the Practical, Robust, Implementation, and Sustainability Model (PRISM) framework, including formative qualitative data collection on context and organizational and patient perspectives [28]. To evaluate the *Training* and pilot *Testing* phases, we focused on the impact of the PRISM context-informed approach on Reach, Effectiveness, Adoption, Implementation, Maintenance (RE-AIM) quantitative outcomes, specifically Reach, Effectiveness, Adoption, and more broadly capturing stakeholder perspectives and context by understanding perceptions of the intervention via qualitative interviews. In the following pages, we describe the pilot implementation and evaluation of the W-PrEP trainings and clinic-based, socio-structural intervention. Using the PRISM framework, we measure the feasibility of delivering the intervention and collecting outcomes of interest and initial outcome trends compared to baseline; we further assess the clinic team and patient perspectives on the intervention to understand the potential for future scale and delivery.

Methods

Ethical Considerations

We received approval from the MedStar Health Institutional Review Board (#0006908). Electronic informed consent was obtained prior to all research activities, namely questionnaires and semistructured informational interviews. All study data were deidentified prior to analysis and stored behind a firewall. We compensated participants US $10 for completing brief poststudy visit questionnaires and US $100 for completing 30 - to 60-minute semistructured interviews.

Evaluation of the W-PrEP Intervention

In this pilot study, we evaluated the feasibility, acceptability, and initial effects of the multifaceted, clinic-based, socio-structural intervention, which included (1) clinic-wide training and PrEP clinical navigator training, (2) toolkits for clinic roll out and toolkits for each clinical role, (3) increased availability and accessibility of PrEP follow-up options (eg, telehealth, off-site labs), (4) PrEP clinical navigation, and (5) patient-facing educational and support resources. Table 1 summarizes the socio-structural components of the adapted W-PrEP intervention [25]. Clinic-wide trainings and PrEP clinical navigation trainings for the pilot intervention were

XSL•FO
RenderX

completed during November and December of 2023. Toolkits and clinic-team facing resources were rolled out in tandem with the trainings. Pilot testing of the intervention ran from January to April 2024, including the assessment of the feasibility of collection of the proposed measures for a planned multicenter, cluster randomized controlled trial. July through October 2023 served as a historical control baseline.

XSL•FO
RenderX

Table . Adapted W-PrEP[a] intervention components.

Component	Description
Clinic-based PrEP[b] rollout toolkit	• W-PrEP[b] implementation checklist • Algorithm for universal PrEP counseling and services • Implementation algorithm(s) for roll out of new PrEP formulations (including prior authorization) • PrEP navigator job description • Identification of a PrEP champion
PrEP champion	• Medical provider to serve as a PrEP expert and site PrEP champion • Responsibilities include facilitation of trainings, consultation on clinic-based implementation of PrEP services, and advising on complex clinical cases
Clinic-wide training	Key training topics include the following: • HIV testing (which test at which time) and STI[f] testing • Taking a patient-centered sexual and social history • Positive reframing prevention communication/counseling around PrEP • How to incorporate patient-centered PrEP and nPEP[c] provision for women into routine care • PrEP initiation, continuation/persistence, and discontinuation • nPEP initiation and transition to PrEP • Implicit and explicit bias training • Trauma-informed care • Epidemiology of the US HIV epidemic • Social and structural determinants of health related to HIV Key features: • Synchronous, interactive PowerPoint presentations scripted for reproducibility and scalability • Asynchronous prerecorded modules to accommodate irregular schedules of medical providers and rotating learners, with activities to engage the asynchronous individual learner • Centering of the patient voice and experience throughout both modules
Increased availability and accessibility of PrEP follow-up	• Increased availability of telehealth/virtual visits for PrEP follow-up visits. These visits are designed as alternatives to in-person visits for women not ready to start PrEP but want to learn more about PrEP (from providers and PrEP navigator) or for women who have been prescribed/started PrEP to follow up regarding issues with prescription pick-up, medication side effects, adherence, or other concerns. • Off-site laboratory referrals: providers to offer referrals to off-site commercial laboratories for follow-up STI and HIV testing if more convenient • PrEP texting with PrEP navigator: email, text, and telephone follow-up options to address concerns/offer education
Clinic team toolkit	• EHR[d] PrEP order set (eg, STI screening lab orders, PrEP injection or prescription orders, HIV testing orders) • EHR PrEP quick-text to facilitate documentation of PrEP education/counseling and prescription or EHR PrEP prompts to remind and encourage providers to counsel all patients about PrEP (dependent upon EHR settings) • PrEP/nPEP badge buddy: an identification-badge card sized PrEP/nPEP reference “cheat” sheet with recommended STI and HIV screening, dosing information, and follow-up • PrEP/nPEP reference “cheat” sheet and checklists: full-size reference sheet and PrEP start/continuation checklists for quick reference during a busy clinic day (recommended to be posted at the provider work station) • Sample scripts for introduction to PrEP (for medical assistants/nurses) and sample scripts for introduction and counseling for prescribing providers • “Ask me about PrEP!” swag (eg, pins, lanyards)

XSL•FO
RenderX

Component	Description
PrEP navigator training	In addition to the clinic-wide PrEP training, through a combination of tailored one-on-one training from a PrEP education and navigation expert, independent learning, and online modules, the PrEP navigator training includes the following: • PrEP navigation through the "Please PrEP Me" Navigator training [29] • Implicit biases in the health care setting [30] • Trauma-informed care [31] • Introduction to reproductive/sexual health topics (led by the PrEP champion or site health care provider, topically tailored to the clinical context)
PrEP navigator toolkit	In addition to the above clinical tools, the PrEP navigator's toolkit includes the following: • PrEP quick-text to facilitate documentation of PrEP navigator education/counseling, including assessment of barriers to PrEP • Sample script for introduction to PrEP and counseling about PrEP • Patient-facing educational resources (see below)
PrEP navigation and support services	Informed by our formative work, PrEP navigators will be cisgender women, ideally from the same community the clinic serves and with first-hand experience with ARVs[e] through PrEP and/or HIV. PrEP navigators' job description includes the following services: • PrEP educational services: introducing PrEP to patients, combatting misperceptions/myths about PrEP • PrEP social and adherence support: breaking down social barriers to PrEP and mitigating the internalization of PrEP-related stigma; building/encouraging individual agency/self-efficacy to initiate and adhere to PrEP • PrEP navigation services: facilitating access to PrEP appointments; assisting with insurance authorization; referrals to indicated social services (eg, substance use, gender-based violence, mental health)
Patient-facing PrEP resources	• Videos on PrEP and women's health/sexual health topics (selected by the clinical site) spliced to play on loop in the clinic waiting room to increase awareness of PrEP • Inclusive pamphlets on PrEP available in the waiting room and exam rooms • Inclusive posters about PrEP and HIV prevention displayed in the waiting room and exam rooms • W-PrEP[b] educational website; QR code distributed by the PrEP navigator and available in the waiting area

[a]*W-PrEP*: women's PrEP project.

[b]PrEP: preexposure prophylaxis.

[c]nPEP: nonoccupational postexposure prophylaxis.

[d]EHR: electronic health record.

[e]ARV: antiretroviral

[f]STI: sexually transmitted infection

Study Setting

Pilot testing was conducted in a family planning and preventative care clinic housed in a large tertiary care center in Washington DC. Washington DC is an epicenter of the US HIV epidemic. At the time of the intervention development and pilot study, the prevalence of HIV among women in Washington DC was 1209 per 100,000, nearly 7-fold the national average (170 per 100,000) [32]; the overall prevalence among women was 0.9%, but nearly double (1.7%) among Black women [32]. At the time of this work, Black women represented 1 in 4 new HIV diagnoses in Washington, DC, and 90% of new diagnoses among women [32].

The selected clinic provides over 1300 family planning patient encounters annually, including pregnancy options counseling, pregnancy termination, contraceptive counseling and provision, and sterilization services. The clinic has 2 dedicated physicians, rotating residents, and students; 2 nurses; 2 medical assistants; and 1 dedicated front desk staff member. Since 2017, the clinic has offered PrEP services, initially as a research initiative (with provider and staff trainings completed in 2017) [33], and now as standard of care. We selected this site, given the demonstrated feasibility of studying PrEP interventions in this setting in order to show proof of concept prior to conducting a multisite study

XSL•FO
RenderX

[33]. We previously published that 98.4% of women seen in this clinic in 2017 met CDC eligibility criteria for PrEP [33].

Formative Work

As described previously [25], as part of our formative ADAPT-ITT process, we conducted 10 key informant interviews with community leaders, operations managers, social workers, clinicians, and health educators working in the PrEP field. These stakeholder interviews, in turn, informed a series of focus groups based on *PRISM* domains to gain insight on how best to adapt the two base cases (ie, HIV PrEP Services for Urban Women [21,22] and Project Shikamana [21,23,24]). We conducted focus groups with providers, PrEP educators and navigators, and patients. With the information gathered, we adapted the base cases into preliminary prototypes for the W-PrEP training and content delivery; these prototypes then underwent theater testing and evaluation by topical experts prior to finalization of the W-PrEP model to be pilot tested [25].

Pilot Implementation Testing Study Design

To assess the implementation of the W-PrEP intervention in the pilot *Testing* phase, we utilized a quasi-experimental, interrupted-time series (ITS) design, specifically a 4-month baseline period followed by 4 months of intervention delivery. An ITS design was chosen to allow internal comparison of the intervention's reach and effectiveness to the historical baseline period. ITS was additionally chosen for this pilot to assess the suitability of the design to assess within-site changes in the proportion of patients engaged/retained in the PrEP cascade associated with the intervention to be utilized in a subsequent cluster randomized controlled trial.

Outcome Assessment

In preparation for a future multicenter, cluster randomized controlled trial, outcomes were guided by the select elements of the PRISM framework [28], namely Reach, Effectiveness, Adoption, and Organizational and Patient Perspectives on the intervention (Table 2). Still, at this stage, we focused on the feasibility of measuring the quantitative outcomes of interest. We defined *Reach* as the proportions of eligible patients who received PrEP counseling by the PrEP navigator or a medical provider as documented in the electronic health record (EHR) or navigator log; eligibility was defined by the 2021 CDC criteria [34]. We defined *Effectiveness* as the proportion of eligible patients who initiated PrEP and as the measured change in perceived medical bias on the postvisit questionnaire. We collected this data primarily to assess the feasibility of the measures, as well as to conduct descriptive statistics and to calculate standard deviation to inform planning for the subsequent full-scale trial. We defined *Adoption* as the proportion of clinic team members who participated in the clinic-wide training and by the relative change in knowledge of PrEP and PrEP best practices on the pre- and posttraining evaluation. At this stage of the work, we did not quantitatively measure implementation fidelity, as we planned to have the pilot be pragmatic, allowing for further adaptation of W-PrEP as needed. From PRISM, we assessed the *organizational perspective* through interviews with the clinic team and PrEP navigators to evaluate perceived acceptability and feasibility of implementing the intervention and satisfaction with the training and the intervention. We also assessed the *patient perspective* via in-depth interviews focusing on the educational resources and counseling provided by the PrEP navigator and by medical providers.

XSL•FO
RenderX

Table . Evaluation of the pilot using the PRISM framework.

PRISM component	Measure
Reach	• Percentage of eligible patients who received introduction to PrEP[a] by navigators (by comparison of the navigators counseling log to the EHR[c]) • Percentage of eligible patients counseled about PrEP by providers (by EHR review)
Preliminary effectiveness[b]	• Percentage of eligible patients who initiate PrEP (by EHR) • Percentage of change in perceived medical bias (by online postvisit survey)
Adoption	• Percentage of clinic-team members who participated in training (by attendance log) • Percentage of change in PrEP knowledge and PrEP best practices (pre-/postclinic-wide training questionnaire)
Intervention characteristics: organizational and patient perspectives	• Perceived feasibility of implementing the intervention and identified barriers (by IDIs[d] with the clinic team and navigators) • Satisfaction with training and intervention usability/complexity among clinic team and navigators (by IDIs with the clinic team and navigators) • Acceptability of the provider or navigator PrEP education/counseling, focusing on burden and addressing patient barriers (by IDI with patients)

[a]PrEP: preexposure prophylaxis.

[b]For the feasibility of data collection and calculation of standard deviation, *not* for statistical testing in a pilot.

[c]EHR: electronic health record.

[d]IDI: in-depth interview.

Quantitative Tools and Outcome Data Collection

For patient-level characteristics and outcomes, the research team extracted patients' demographics, indications for PrEP by the 2021 CDC eligibility criteria, documented counseling about PrEP by PrEP clinical navigators and by providers, level of interest in PrEP (if recorded), and follow-up for subsequent PrEP visits (eg, for PrEP education, PrEP initiation, and PrEP continuation at 1 mo) from EHR. Additionally, in the baseline and intervention periods, we invited all patients to complete a 5-minute, smartphone-based, post-visit questionnaire, which queried sociodemographics, indications for PrEP, perceived medical discrimination [35], norms around and attitudes toward PrEP, intention to initiate PrEP, and interaction with the W-PrEP intervention. Given that many patients participating in routine care were offered analgesia and sedation for medical procedures, the postvisit questionnaire was offered via a flyer distributed by the front desk with a QR code for questionnaire completion at home and via email within a week of their appointment. Quantitative EHR and questionnaire data were entered into REDCap (Research Electronic Data Capture [36,37]) hosted at MedStar Health. Additionally, the PrEP clinical navigators kept a log of all patients seen, education provided, and planned follow-up as applicable. For clinic-level outcomes, we utilized attendance logs to capture attendance at the clinic-wide and PrEP clinical navigation trainings. At these trainings, participants completed a pre- and posttraining assessment of PrEP knowledge and understanding of best practices, as well as a postassessment evaluation of the training.

Quantitative Analysis

We calculated descriptive statistics, including frequencies and percentages, measures of central tendency, dispersion, and appropriate visualization approaches (eg, box plots and spaghetti plots) on each outcome variable; we used line graphs to illustrate the trends of the outcome over time. Bivariate analyses were conducted, including *t* tests for normally distributed continuous variables and χ^2 tests or Fisher Exact tests for categorical variables, as appropriate based on cell sizes, to compare sample characteristics between the intervention and historical-control time periods; nonparametric methods were used in the presence of nonnormally distributed variables. We conducted multilevel modeling (MLM) with implementation of intervention (post vs pre) to gage a preliminary assessment of the intervention effect, accounting for possible repeated measures within some participants over the period. Specifically, MLM assessed each outcome variable in a separate model, including the predictor as intervention status (post vs pre), with a random intercept at the patient level to account for the possible multiple visits (44% with 2 or more visits) within each patient, before and after adjusting for sample characteristics (eg, age, race, ethnicity). We conducted ITS analyses, including an interaction between time in month as a continuous variable, which is centered at the first month after intervention implementation, and a categorical variable indicating intervention (post vs pre) in MLM. A significant coefficient for intervention in the model indicates an immediate change after intervention, and a significant interaction indicates a change in the trend over time between pre- and postintervention periods. Of the 1437 patient encounters

XSL•FO
RenderX

in the dataset, 1 (0.1%) encounter was missing the month of the visit and the intervention status, 2 (0.1%) were missing age, 12 (0.8%) were missing race, and 27 (1.9%) were missing ethnicity. Eight were missing confirmation of the starting of PrEP, while 1 encounter was missing other outcomes. A higher percentage of missingness was found for the other race categories, 12.2% versus a range of 1% to 3% for Black, White, Asian or Pacific Islander, or Native American ($P<.001$). Those with any missingness or those with complete data did not differ in age, ethnicity, or any outcome. Listwise deletion was applied for the covariates, including age, race, and ethnicity, while maximum likelihood was used for the outcome if it was missing in MLM. A significance level was set at 0.05 for all the analyses.

Qualitative Tools and Data Collection

At the end of the *Testing* phase, we conducted semistructured in-depth interviews to assess *perspectives on the intervention* with clinic staff, PrEP navigators, and patients who participated in the pilot. See Multimedia Appendix 1 for interview guides. All clinic team members and PrEP navigators were invited to participate in interviews; interview questions focused on perceptions of the intervention, including trainings and the PrEP clinical navigator program, and potential improvements to be incorporated. We utilized purposive sampling to recruit a total of 20 patient-participants, specifically Black women of reproductive age, at varying stages in the PrEP cascade (ie, not interested in PrEP, considering initiating PrEP, initiated PrEP). These patient interview questions focused on the patient experience with the intervention and potential improvements to the intervention and provision of PrEP services. Interviews were conducted by experienced research scientist team members. Given the timing during the COVID-19 pandemic and to increase access for participants who would be unable to attend in person, all interviews were conducted remotely via a secure Microsoft Teams platform.

Qualitative Analysis

We applied rapid qualitative analysis (RQA) [38-40] to systematically translate the qualitative data from the interviews into actionable refinements. We audio-recorded interviews with the consent of participants, which were then transcribed verbatim. We developed summary templates to identify the coder, respondent role, and coding domain (ie, strengths, information gaps, actionable improvements). We conducted analyses of a subset of interviews (ie, one from each respondent type) in order to ensure adequate fit of the template to the data, then we coded the remaining transcripts. For the *Training* and *Testing* ADAPT-ITT phases, we developed and populated matrices to align with germane components of the PRISM implementation framework. We used the matrices to synthesize key themes during intervention development and to synthesize strengths, information gaps, and other actionable improvements (eg, unclear or confusing language, alternative or additional resources, and toolkit components), stratified by respondent type or role. These summaries guided team discussions and refinement decisions pertinent to the W-PrEP intervention. All RQA analyses were performed by two trained medical student research team members with experience in qualitative analyses. RQA were completed under the supervision of the study's primary investigator, the study implementation scientist, and the study health communications consultant.

Results

Training

Clinic-Wide Training

The adapted training was divided into two modules, each approximately 1 hour long. The first module focused on HIV prevention, offering a brief introduction to HIV and HIV transmission, HIV tests and timing of HIV testing, and HIV prevention, with a focus on PrEP and nonoccupational postexposure prophylaxis (nPEP). The training detailed the PrEP modalities and regimens approved for protection against vaginal exposure to HIV, indications for PrEP and nPEP, and the "PrEP Steps": taking a sexual and social history, education and counseling, laboratory testing, and management of starting, continuing, and stopping PrEP. The PrEP steps were interspersed with patient testimonials, group activities (eg, writing a script to introduce PrEP to all patients with nonstigmatizing language), an original video created by the research team modeling trauma-informed, PrEP-positive care, and sample cases for group discussion. The first module also focused on the incorporation of PrEP into clinical practice.

The second training module focused on the epidemiology of the HIV epidemic in the United States—specifically for women. The training reviewed the social and structural determinants of health as related to HIV and HIV prevention for women. It discussed trauma-informed care and incorporation of trauma-informed care into HIV prevention care. The second module also reviewed strategies to mitigate bias in clinical practice. Both training modules incorporated patient testimonials, group activities, and a component of the original video highlighting social and structural determinants of health barriers to care.

Adoption

The clinic team elected to complete both modules in single training days. Two training days were scheduled to accommodate training of clinic team members, including the PrEP navigators. The training was open to other medical providers within the department; all members of the clinic team, including the 2 PrEP navigators, completed the training. Clinic team members were additionally joined by several resident trainees and one certified nurse midwife who had expressed interest in completing the training. A total of 13 individuals completed the trainings, along with deidentified, linked pre- and posttests for each module. Attendance was monitored via sign-in sheet or attendance log. Combined knowledge test scores improved from a mean of 65% to 89% between the pre- and posttests. Participants favorably reviewed the training in the feedback section of the posttest, and 100% of participants reported "intend(ing) to apply what (they) learned in the training" and feeling "motivated to make changes in (their) work" Table 3 highlights exemplary quotations regarding the trainings from the feedback portion of the posttest.

XSL•FO
RenderX

Table . Exemplary quotations from the qualitative evaluation of the training and pilot testing.

Interview topic, theme	Quotations
Training	
Most useful component of training: content on guidelines	"(Review of) guidelines for prescribing and monitoring" -Training participant #11
Most useful component of training: content on patient-centered care	"Hearing patient voices (and) seeing examples of "bad" and "good" clinical scenarios. While we roll our eyes over shame tactics, mistreatment (by health care providers) it's always helpful to see it so clearly and remind us to be better" -Training participant #4
Main take-away: content on guidelines and universal approach to PrEP[a] education	"I can prescribe PrEP and should be talking to pretty much all of my patients about it" - Training participant #1
Main take-away: content on stigma	"(To) encourag(e) the topic of PrEP to become normal and less stigmatized in clinic settings" - Training participant #8
Most useful component of training: format of interactive exercise	"The elevator pitch and counseling for PrEP" (referring to the toolkit scripts and interactive exercise) - Training participant #10
Most useful component of training: content on clinical scenarios	"The training made sure to…go through…scenarios that we might encounter. Some of those scenarios did end up happening, so I felt like I was better equipped to handle certain conversations because (of) that training" - Training participant #2
Testing	
Perception of the intervention: PrEP navigation	"It's very helpful in terms of giving more information to our patients about PrEP, so that we're not…starting at Ground Zero to discuss PrEP, 'cause obviously …there's all these time crunches and … it's…one more thing to crunch into the time. And so while I review it with every patient, if I'm not starting from scratch, when I'm reviewing it, then it's a big help." - Clinic team member #2: medical provider, White cisgender female
Perception of the intervention: PrEP educational resources	It serves a very important purpose because…most of our patients are high risk for HIV and (they) hear about PrEP (in the educational videos) on (the) TV and actually having someone to talk to them more in depth…(and) we give them pamphlets. Some of them may say no initially, but when they go home they read the pamphlet." - Clinic team member # 4: Nurse, Black, cisgender female
Perception of the intervention: barriers to PrEP	"With the patients that I encountered a lot of them didn't know that (PrEP) was also…for women. A lot of them just assumed it was for males, gay males, …and they also appreciated the fact that this was something that was being provided to the women, especially the minorities in our community. …I think also being able to have the conversation with them maybe helped with the stigma part of it.…(A) lot of them became more comfortable in talking about HIV and HIV prevention because initially when I was talking to certain patients about it…they seemed like they were hesitant about talking about it, but the more I went into detail…they were …a bit more comfortable having the conversation." - PrEP Navigator #1: Black, cisgender female

XSL•FO
RenderX

Interview topic, theme	Quotations
Perception of the intervention: PrEP navigator	"She gave me an overview of what exactly PrEP is, how I can go about getting it, and if I'm interested, the steps that I could take to link up with her and move forward...I became informed of what PrEP was. …She also gave me the option of starting then or setting a date for myself, or even if I wanted to take a deeper dive and learn more. She provided those resources, and I was grateful for that…that was a generous gesture. I feel like being knowledgeable and aware of what's happening around you, especially for young women or men engaging in sexual activity, is important. Preventative care and hygienic care are essential. I appreciated having someone there to be informative, even if I couldn't proceed at that moment. It was very informative, and now I can assist others and have conversations." - Patient #18: Black, cisgender female, age 23, considering PrEP
Perception of the intervention: PrEP navigator	"My experience was great. Even though I might not remember all the details, the information was presented well. I'd like to have (the PrEP navigator) on my team to help me stay on track with PrEP. It's a process with appointments and doses, so I could use that reminder." - Patient #15: Black, cisgender female, 31, considering PrEP

[a]PrEP: preexposure prophylaxis.

PrEP Navigator Training

The PrEP clinical navigator training program consisted of the above clinic-wide training and a curated curriculum centered on the Please PrEP Me navigation manual [29]. PrEP navigators independently studied the Please PrEP Me navigation manual and then reviewed it with the PrEP navigation team from a local health center over two sessions, including talking through potential challenging scenarios and role plays. Additionally, the PrEP navigators independently completed online trainings on trauma-informed care [31] and implicit bias in the clinical setting [30], which were subsequently discussed as a small group. Lastly, navigators also completed tailored training on sexual and reproductive health, reviewing commonly used anatomical and medical terms in reproductive and sexual health care, and common conditions and medical indications for visits in the pilot clinic setting. The training was well received; the review of clinical scenarios and role plays was highlighted as particularly helpful. Table 3 highlights representative quotations regarding the trainings.

Testing

Instrumentation

We found the instrumentation and outcome measures to be both feasible to collect and appropriate to capture the primary outcomes of interest in the EHR. However, during the preimplementation period, only 54 respondents completed the survey due to competing priorities or sedation during the appointment, making it inappropriate and not feasible to continue survey data collection in clinic.

Interviews among clinic staff (n=5) and navigators (n=2) in the last weeks of the implementation period showed general perceptions of feasibility and acceptability of the training and intervention delivery, noting that the intervention minimally impacted clinic workflow.

Reach and Preliminary Effectiveness

We assessed the frequency of PrEP counseling by providers in the preimplementation period (July to October 2023) and compared the proportion of patients who received counseling on PrEP. We found that provider counseling was maintained for 6 years after the previous PrEP training, with counseling documented in 65% (350 of 537) of all patients seen during the 4-month period (controlling for patients with multiple follow-up visits). In the setting of already high provision of PrEP counseling, PrEP counseling *by providers* did not significantly increase during the intervention period; however, with the introduction of the PrEP clinical navigators, the reach of the W-PrEP intervention increased counseling to 76% of patients seen (controlling for patients with multiple follow-up visits). Based on MLM, the likelihood of PrEP counseling increased by 43% in posttest without covariates (odds ratio [OR]=1.43, 95% CI 1.16 - 1.77; $P<.001$) and after adjusting for covariates (adjusted OR [aOR]=1.51, 95% CI 1.21 - 1.88; $P<.001$). The ITS showed that there was a significant immediate change in the odds at the first month after the initiation of intervention implementation (aOR=1.82, 95% CI 1.02 - 3.24; $P=.04$), and there was also a significant increase in the trend in PrEP counseling (aOR=1.28, 95% CI 1.05 - 1.57; $P=.02$). We successfully collected data on preliminary *Effectiveness*, namely PrEP initiation, to demonstrate feasibility of data collection and calculation of standard deviation, *not* for statistical testing in a pilot. We found a statistically significant increase in interest in learning more about PrEP (n=39, 5.6% → n=93, 12.5%; $P<.001$); we did not see statistically significant increases in prescription of PrEP/initiation of a prior authorization nor differences in PrEP initiation but were not powered to do so. Table 4 describes the demographic characteristics, indications for PrEP, and PrEP-related outcome measures collected in the pilot. Figure 1 depicts the reach of counseling by providers or navigators by month in the baseline and intervention periods.

XSL•FO
RenderX

Table . Demographic characteristics, indications for PrEP[a], and PrEP-cascade outcomes by patient encounter during baseline and intervention periods of the pilot study[b].

Demographic characteristics, indications for PrEP, and PrEP-cascade outcomes	Baseline	Intervention	*P* value
Number of participants, n (%)	691 (48.1)	745 (51.9)	
Age (y), mean (SD)	29 (7.4)	30 (7.0)	.34
Race, n (%)			.06
Black	501 (73.8)	546 (73.3)	
White	102 (15)	85 (11.4)	
Asian/Pacific Islander	13 (1.9)	18 (2.4)	
Native American	2 (0.3)	1 (0.1)	
Other (including mixed race)	61 (9)	95 (12.8)	
Ethnicity, n (%)			.35
Hispanic	36 (5.3)	48 (6.5)	
Non-Hispanic	637 (94.7)	688 (93.5)	
Inconsistent condom use, n (%)			.02
No	348 (50.4)	420 (56.4)	
Yes	343 (49.6)	325 (43.6)	
STI[h] diagnosis within the past 6 months, n (%)			.46
No	683 (98.8)	733 (98.4)	
Yes	8 (1.2)	12 (1.6)	
IV drug use or needle sharing, n (%)			
No	691 (100)	745 (100)	
Partner(s) status[c], n (%)			<.001
Unknown HIV status	0 (0)	175 (23.5)	
Not documented	690 (100)	567 (76.2)	
Positive	0 (0)	2 (0.3)	
Requests PrEP, n (%)			.59
No	688 (99.6)	743 (99.7)	
Yes	3 (0.4)	2 (0.3)	
nPEP[d] use, n (%)			
No	691 (100)	745 (100)	
PrEP navigator introduced PrEP/counseled patient regarding PrEP, n (%)			<.001
No	N/A[e]	462 (62)	
Yes	N/A	283 (38)	
Provider introduced PrEP/counseled patient regarding PrEP, n (%)			.02
No	323 (46.8)	395 (53.2)	
Yes	367 (53.2)	347 (46.8)	
Either provider or navigator introduced PrEP/counseled patient regarding PrEP, n (%)			<.001

XSL•FO
RenderX

Demographic characteristics, indications for PrEP, and PrEP-cascade outcomes	Baseline	Intervention	*P* value
No	324 (46.9)	284 (38.1)	
Yes	367 (53.1)	461 (61.9)	
Patient expressed interest in learning more about PrEP, n (%)			<.001
No	652 (94.4)	651 (87.5)	
Yes	39 (5.6)	93 (12.5)	
Patient expressed interest in starting PrEP, n (%)			.32
No	680 (98.6)	729 (97.9)	
Yes	10 (1.4)	16 (2.1)	
Provider prescribed TDF/FTC[f] or initiated prior authorization for CAB[g], n (%)			.48
No	686 (99.3)	736 (98.9)	
Yes	5 (0.7)	8 (1.1)	
Verified PrEP initiation, n (%)			.66
No	685 (99.3)	734 (99.5)	
Yes	5 (0.7)	4 (0.5)	

[a]PrEP: preexposure prophylaxis.

[b]Fisher exact tests were conducted if any cell size is under 10.

[c]Partner HIV status was added to the EHR “quick text” for the W-PrEP Intervention.

[d]nPEP: nonoccupational postexposure prophylaxis.

[e]Role of PrEP navigator did not exist prior to W-PrEP Intervention.

[f]TDF/FTC: tenofovir disoproxil fumarate/emtricitabine.

[g]CAB: cabotegravir.

[h]STI: sexually transmitted infection

XSL•FO
RenderX

Figure 1. Proportion of patient visits with documented preexposure prophylaxis (PrEP) counsel by a health care provider or by a PrEP navigator by month (M) during the baseline and intervention periods (not controlling for multiple follow-up visits).

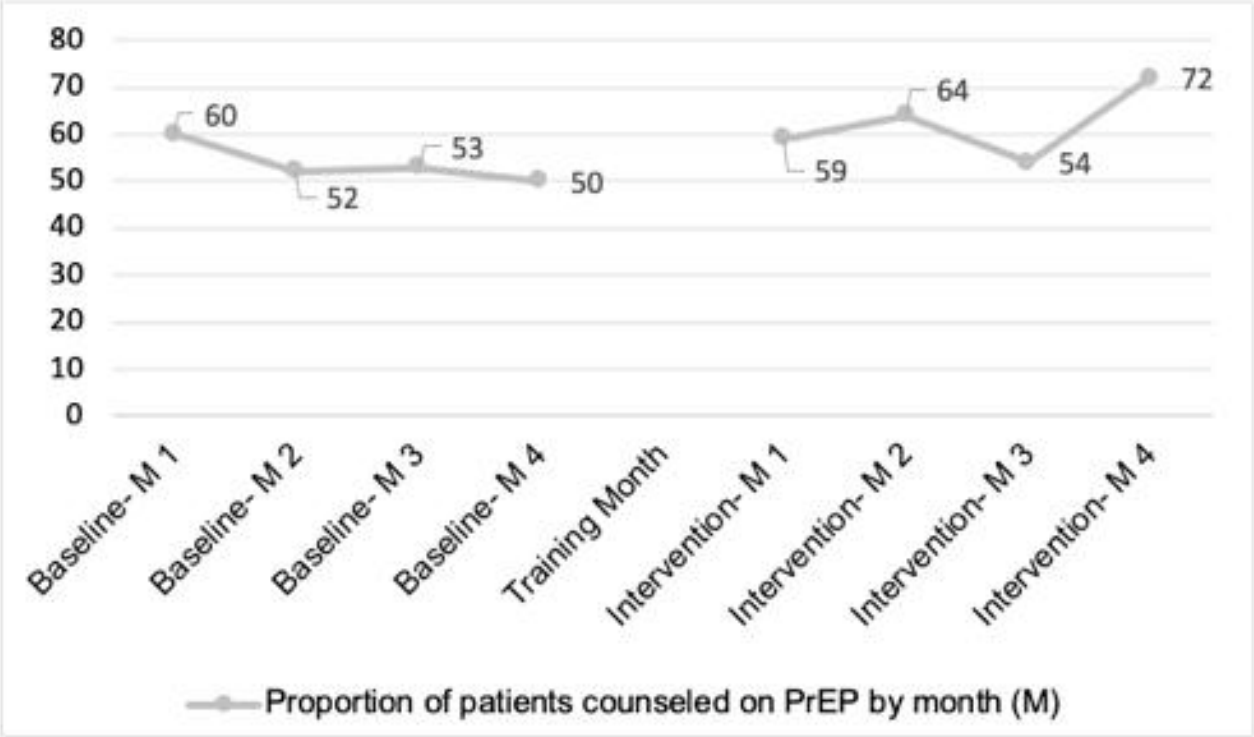

Organizational and Patient Perspectives on the Intervention

Multiple interviewees noted patient interest in PrEP education but challenges with continued low patient-perceived risk of HIV and subsequent low PrEP uptake. Multiple members of the clinic team mentioned that due to the time lag associated with the prior authorization for long-acting cabotegravir, it was challenging to comply with the CDC guidelines which recommend negative HIV testing within 7 days prior to initiation [34]. The clinic team universally praised having a PrEP clinical navigator to expand patient education, assist with prior authorizations, and orchestrate appointments for long-acting PrEP. They additionally lauded the expanded training, enhanced toolkit, and patient-facing resources. Patient-participants (n=20) interviewed likewise reported that their exposure to the W-PrEP intervention was highly acceptable and appreciated the education and counseling from the PrEP navigator—many learning about PrEP for the first time from the navigator and provider. A common sentiment was that although the information about HIV and PrEP was insightful, it did not apply to them. Table 3 highlights exemplary quotations regarding the W-PrEP intervention and implementation.

Discussion

Principal Results

The adapted W-PrEP intervention was responsive to the concerns and expressed needs of health care providers and clinic team members, PrEP navigators, and women of reproductive age, in particular Black women. There was high adoption (100%) of the training by clinic team members and significant improvement in clinic team knowledge of HIV prevention and PrEP. Though this pilot was not powered to assess increased PrEP uptake, W-PrEP demonstrated increased reach of PrEP counseling with the addition of the PrEP clinical navigator, dramatically increased patient interest in learning more about PrEP, and a nonsignificant trend toward increased prescription of oral PrEP/prior authorization request for injectable PrEP on the part of providers. Of note, we did observe a small decrease in provider-led PrEP counseling from baseline. In discussion with the clinic providers to understand this decrease, providers mentioned that if the PrEP navigator counseled a patient about PrEP and signaled a patient's disinterest in discussing PrEP further, the providers were then less likely to rebroach the topic in that visit. Clinic team members, PrEP navigators, and patients found W-PrEP highly acceptable. Clinic team members found the training and implementation of the intervention feasible, with minimal interruption of clinical flow. In terms of study instrumentation, we found that EHR review was highly feasible to assess our primary outcomes. In the context of the family planning clinic, the self-report survey tool was not feasible due to low response; in clinical settings without conscious sedation for procedures, this measure would have been more feasible as patients could complete the questionnaire on-site immediately postvisit.

Comparison to Prior and Contemporary Work

Despite increased public health marketing directed to women and direct-to-consumer advertising by industry targeting women, there remain limited CDC-designated best practices for HIV prevention or PrEP uptake inclusive of women, let alone focusing on women [41]. Coleman et al [42] describe a clinic-based integrated pharmacy and PrEP navigation program, inclusive of patients of all sexes/genders, designed to break

XSL•FO
RenderX

down barriers to PrEP utilization. Cisgender women comprised a tiny minority of the patients included, and, although not statistically significant, both the proportion and absolute number of cisgender women decreased from the baseline to the intervention period. Casey et al [43] describe a successful educational intervention across primary care and women's health clinics within the New York City public health care system, with significant increases in the numbers of primary care and women's health providers offering and prescribing PrEP, similar to the intervention that W-PrEP was adapted from [22]. The scope of the intervention was limited to educating and empowering medical providers to provide PrEP services and did not address the significant social and structural barriers to PrEP uptake and persistence.

W-PrEP expanded upon our solely educational intervention [22,33] to more fully address identified patient-, provider-, and clinic-level socio-structural barriers to PrEP and HIV prevention for women of reproductive age and, in particular, Black women. Specifically, W-PrEP was designed to combat low awareness and stigma about PrEP among patients, low knowledge/confidence of how to provide PrEP among providers, and clinic-level challenges of systems/logistics for obtaining prior authorization and tracking patient follow-up. This strategy demonstrated a significantly increased reach of PrEP counseling and education, resulting in increased interest in learning more about PrEP. Developed in parallel to W-PrEP and currently under evaluation, the POWER Up intervention similarly utilizes a clinic-based multifaceted approach consisting of routine PrEP education for patients, standardized provider training, EHR optimization, PrEP navigation, and PrEP clinical champions [44]. Both interventions appear to share a universal approach to patient PrEP education, a PrEP navigation program, utilization of PrEP clinical champions, and EHR toolkits. W-PrEP differs from POWER Up in several aspects: (1) W-PrEP's clinic-wide training program is inclusive not only of providers but of all clinic staff members, recognizing that patients often receive health information and messaging from both clinical and nonclinical members of the clinic staff. This feature leverages the potential of all members of the team, from the clinic receptionist to the nurse administrator, to provide PrEP resources, field basic questions about PrEP and HIV prevention, and direct patients to providers or PrEP navigators for additional information, as needed. (2) W-PrEP offers several toolkits with practical resources constructed to aid in provision of PrEP education and services. These include a clinic-level toolkit, including modifiable algorithms for incorporating PrEP into clinic flow and a clinic PrEP roll-out readiness checklist, a clinic team toolkit with quick reference "cheat sheets" and badge buddies and sample scripts for introducing PrEP and PrEP counseling, and a PrEP navigator toolkit with scripts and patient-facing educational resources. These toolkits were created in response to providers and administrators' plea for blueprints to integrate PrEP services into busy clinical practice. (3) Finally, W-PrEP offers expanded access through telehealth and web-based follow-up and off-site laboratory referrals.

Limitations

As this was a pilot, we limited both the number of sites and duration—thus we were not powered for statistical testing of the outcomes planned for a future trial of the intervention. This intervention was piloted in an urban hospital center affiliated with an academic institution; however, we posit that our findings are likely generalizable to sexual and reproductive health centers in both community and academic settings. Future work should incorporate additional geographic regions and clinical settings of varied sizes. It is also worth noting that the pilot was conducted in a setting where providers previously received informational training on PrEP in 2017 and continue to offer PrEP services to the majority of patients. We anticipate that in settings with lower baseline PrEP counseling, the W-PrEP intervention would have a more dramatic effect upon the outcome measures of interest.

Although we were successful in testing the use of EHR data collection for our primary outcomes, we were largely unsuccessful with the implementation of the patient questionnaire. We were unable to utilize the questionnaire immediately postvisit, as a significant proportion of patients received sedation for termination of pregnancy. We anticipate that in nonprocedural clinical settings (or settings that do not routinely utilize sedation), exit questionnaires could be administered by a coordinator after the clinical visit with a higher response rate. We are additionally exploring other data capture systems to simplify the process.

Conclusions

W-PrEP, a clinic-based, socio-structural intervention adapted from two CDC HIV Prevention Best Practices models, demonstrated increased reach of PrEP counseling, increased interest in PrEP among patients, and a nonsignificant trend toward increased prescription of oral PrEP/prior authorization request for injectable PrEP by providers. The intervention was both acceptable and feasible to the clinic team and the intended patient population. Further evaluation is needed to test the effects of the intervention on the PrEP cascade, namely PrEP initiation, adherence, and persistence, particularly among the intended population.

Acknowledgments

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors would like to thank the "Women's Preexposure Prophylaxis Project" joint community and scientific advisory board and the patient and clinic team participants in the pilot. The authors would additionally like to thank Whitman Walker Health for their participation in the preexposure prophylaxis Navigator Training. The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript. Generative AI was not used in any portion of the manuscript writing.

XSL•FO
RenderX

Funding

Research reported in this publication was supported by National Institute of Mental Health of the National Institutes of Health under award number R34MH128046.

Data Availability

Data are presented in the manuscript. Please contact the study authors to request any additional data (eg, qualitative data).

Conflicts of Interest

RS has received investigator-sponsored research funding (managed by MHRI) from ViiV Healthcare and participated in ViiV Healthcare and Gilead Science advisory meetings. DK has received investigator-sponsored research funding (managed by George Washington University) from ViiV Healthcare. SJH has participated in Gilead Science Advisory meetings. All other authors declare no conflicts of interest.

Multimedia Appendix 1
Interview guides.
[DOCX File, 22 KB - formative_v9i1e80653_app1.docx]

References

1. Fast facts: HIV and women. Centers for Disease Control and Prevention. 2024. URL: https://www.cdc.gov/hiv/data-research/facts-stats/women.html [accessed 2024-08-01]
2. AtlasPlus - charts. Centers for Disease Control and Prevention. URL: https://gis.cdc.gov/grasp/nchhstpatlas/charts.html [accessed 2024-12-27]
3. Hess KL, Hu X, Lansky A, Mermin J, Hall HI. Lifetime risk of a diagnosis of HIV infection in the United States. Ann Epidemiol 2017 Apr;27(4):238-243. [doi: 10.1016/j.annepidem.2017.02.003] [Medline: 28325538]
4. HIV diagnoses, deaths, and prevalence. Centers for Disease Control and Prevention. 2024. URL: https://www.cdc.gov/hiv-data/nhss/hiv-diagnoses-deaths-prevalence.html [accessed 2024-12-27]
5. Abdool Karim Q, Sibeko S, Baxter C. Preventing HIV infection in women: a global health imperative. CLIN INFECT DIS 2010 May 15;50 Suppl 3:S122-S129. [doi: 10.1086/651483] [Medline: 20397940]
6. Beksinska M, Wong R, Smit J. Male and female condoms: their key role in pregnancy and STI/HIV prevention. Best Pract Res Clin Obstet Gynaecol 2020 Jul;66:55-67. [doi: 10.1016/j.bpobgyn.2019.12.001] [Medline: 32007451]
7. Baeten JM, Donnell D, Ndase P, et al. Antiretroviral prophylaxis for HIV prevention in heterosexual men and women. N Engl J Med 2012 Aug 2;367(5):399-410. [doi: 10.1056/NEJMoa1108524] [Medline: 22784037]
8. Thigpen MC, Kebaabetswe PM, Paxton LA, et al. Antiretroviral preexposure prophylaxis for heterosexual HIV transmission in Botswana. N Engl J Med 2012 Aug 2;367(5):423-434. [doi: 10.1056/NEJMoa1110711] [Medline: 22784038]
9. Murnane PM, Celum C, Mugo N, et al. Efficacy of preexposure prophylaxis for HIV-1 prevention among high-risk heterosexuals: subgroup analyses from a randomized trial. AIDS 2013 Aug 24;27(13):2155-2160. [doi: 10.1097/QAD.0b013e3283629037] [Medline: 24384592]
10. Marrazzo J, Tao L, Becker M, et al. HIV preexposure prophylaxis with emtricitabine and tenofovir disoproxil fumarate among cisgender women. JAMA 2024 Mar 19;331(11):930-937. [doi: 10.1001/jama.2024.0464] [Medline: 38427359]
11. Delany-Moretlwe S, Hughes JP, Bock P, et al. Cabotegravir for the prevention of HIV-1 in women: results from HPTN 084, a phase 3, randomised clinical trial. Lancet 2022 May 7;399(10337):1779-1789. [doi: 10.1016/S0140-6736(22)00538-4] [Medline: 35378077]
12. Bekker LG, Das M, Abdool Karim Q, et al. Twice-yearly lenacapavir or daily F/TAF for HIV prevention in cisgender women. N Engl J Med 2024 Oct 3;391(13):1179-1192. [doi: 10.1056/NEJMoa2407001] [Medline: 39046157]
13. Sarkar NN. Barriers to condom use. Eur J Contracept Reprod Health Care 2008 Jun;13(2):114-122. [doi: 10.1080/13625180802011302] [Medline: 18465472]
14. The HIV prevention pipeline. AVAC. 2024. URL: https://avac.org/resource/infographic/the-hiv-prevention-pipeline/ [accessed 2024-08-01]
15. MATRIX: a USAID project to advance the research and development of innovative HIV prevention products for women. : MATRIX; 2024. URL: https://www.matrix4prevention.org [accessed 2024-08-01]
16. Smith DK, Van Handel M, Grey J. Estimates of adults with indications for HIV pre-exposure prophylaxis by jurisdiction, transmission risk group, and race/ethnicity, United States, 2015. Ann Epidemiol 2018 Dec;28(12):850-857. [doi: 10.1016/j.annepidem.2018.05.003] [Medline: 29941379]
17. Huang YLA, Zhu W, Smith DK, Harris N, Hoover KW. HIV preexposure prophylaxis, by race and ethnicity-United States, 2014-2016. MMWR Morb Mortal Wkly Rep 2018 Oct 19;67(41):1147-1150. [doi: 10.15585/mmwr.mm6741a3] [Medline: 30335734]

XSL•FO
RenderX

18. Bush S, Ng L, Magnuson D, Piontkowsky D, Giler RM. Significant uptake of truvada for preexposure prophylaxis (prep) utilization in the US in late 2014-1Q2015. 2015 URL: https://www.iapac.org/AdherenceConference/presentations/ADH10_OA74.pdf [accessed 2025-11-14]
19. Bush S, Magnuson D, Rawlings M, Hawkins T, McCallister S, Giler RM. ASM/ICAAC: racial characteristics of FTC/TDF for pre-exposure prophylaxis (prep) users in the US. 2016 URL: http://www.natap.org/2016/HIV/062216_02.htm [accessed 2025-11-14]
20. Sullivan PS, DuBose SN, Castel AD, et al. Equity of PrEP uptake by race, ethnicity, sex and region in the United States in the first decade of PrEP: a population-based analysis. Lancet Reg Health Am 2024 May;33:100738. [doi: 10.1016/j.lana.2024.100738] [Medline: 38659491]
21. Compendium | intervention research | research | HIV |. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/hivpartners/php/compendium/index.html [accessed 2024-08-01]
22. Scott RK, Deyarmond M, Marwitz S, et al. Implementation of an educational intervention to improve HIV pre-exposure prophylaxis services for women in an urban sexual health clinic. AIDS Patient Care STDS 2023 Sep;37(9):447-457. [doi: 10.1089/apc.2023.0107] [Medline: 37713289]
23. Kerrigan D, Mbwambo J, Likindikoki S, et al. Project Shikamana: baseline findings from a community empowerment-based combination HIV prevention trial among female sex workers in Iringa, Tanzania. J Acquir Immune Defic Syndr 2017 Jan 1;74 Suppl 1(Suppl 1):S60-S68. [doi: 10.1097/QAI.0000000000001203] [Medline: 27930613]
24. Kerrigan D, Mbwambo J, Likindikoki S, et al. Project Shikamana: community empowerment-based combination HIV prevention significantly impacts HIV incidence and care continuum outcomes among female sex workers in Iringa, Tanzania. J Acquir Immune Defic Syndr 2019 Oct 1;82(2):141-148. [doi: 10.1097/QAI.0000000000002123] [Medline: 31513552]
25. Scott RK, Hull SJ, Kerrigan D, et al. Development of a clinic-based, sociostructural intervention to improve the provision of pre-exposure prophylaxis for cisgender women: formative study using the assessment, decision, adaptation, production, topical experts, integration, training, and testing (ADAPT-ITT) framework. JMIR Form Res 2025 Aug 6;9(1):e75922. [doi: 10.2196/75922] [Medline: 40768750]
26. Wingood GM, DiClemente RJ. The ADAPT-ITT model: a novel method of adapting evidence-based HIV Interventions. J Acquir Immune Defic Syndr 2008 Mar 1;47 Suppl 1:S40-S46. [doi: 10.1097/QAI.0b013e3181605df1] [Medline: 18301133]
27. Bronfenbrenner U. Ecological models of human development. In: International Encyclopedia of Education, 2nd edition: Elsevier; 1994, Vol. 3. URL: https://www.ncj.nl/wp-content/uploads/media-import/docs/6a45c1a4-82ad-4f69-957e-1c76966678e2.pdf [accessed 2025-11-14]
28. Feldstein AC, Glasgow RE. A practical, robust implementation and sustainability model (PRISM) for integrating research findings into practice. Jt Comm J Qual Patient Saf 2008 Apr;34(4):228-243. [doi: 10.1016/s1553-7250(08)34030-6] [Medline: 18468362]
29. PrEP navigation manual. Please PrEP Me. URL: https://pleaseprepme.org/prepnavigatormanual/ [accessed 2025-01-22]
30. Implicit bias: a practical guide for healthcare settings. The George Washington University School of Medicine and Health Sciences. 2024. URL: https://cme.smhs.gwu.edu/www.dcrxce.com/content/implicit-bias-practical-guide-healthcare-settings#group-tabs-node-course-default1 [accessed 2025-11-14]
31. Intro to trauma informed care training modules. Trauma Informed Oregon. 2024. URL: https://traumainformedoregon.org/resources/training/tic-intro-training-modules/ [accessed 2024-08-02]
32. Young R, Bowser M, Nesbitt LS. HIV/AIDS and administration (HAHSTA), “Annual epidemiology & surveillance report: data through. District of Columbia Department of Health. 2019. URL: https://dchealth.dc.gov/node/1424881 [accessed 2025-11-10]
33. Brant AR, Dhillon P, Hull S, et al. Integrating HIV pre-exposure prophylaxis into family planning care: a RE-AIM framework evaluation. AIDS Patient Care STDS 2020 Jun;34(6):259-266. [doi: 10.1089/apc.2020.0004] [Medline: 32484743]
34. US public health service: preexposure prophylaxis for the prevention of HIV infection in the united states—2021 update: a clinical practice guideline. Centers for Disease Control and Prevention. 2021. URL: https://www.cdc.gov/hiv/pdf/risk/prep/cdc-hiv-prep-guidelines-2021.pdf [accessed 2025-11-10]
35. Peek ME, Nunez-Smith M, Drum M, Lewis TT. Adapting the everyday discrimination scale to medical settings: reliability and validity testing in a sample of African American patients. Ethn Dis 2011;21(4):502-509. [Medline: 22428358]
36. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381. [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
37. Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: building an international community of software platform partners. J Biomed Inform 2019 Jul;95:103208. [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]
38. Hamilton AB. Qualitative methods in rapid turn-around health services research. 2013 Presented at: URL: https://www.hsrd.research.va.gov/for_researchers/cyber_seminars/archives/video_archive.cfm?SessionID=780 [accessed 2025-11-14]
39. Palinkas LA, Mendon SJ, Hamilton AB. Innovations in mixed methods evaluations. Annu Rev Public Health 2019 Apr 1;40:423-442. [doi: 10.1146/annurev-publhealth-040218-044215] [Medline: 30633710]

XSL•FO
RenderX

40. Gale RC, Wu J, Erhardt T, et al. Comparison of rapid vs in-depth qualitative analytic methods from a process evaluation of academic detailing in the Veterans Health Administration. Implement Sci 2019 Feb 1;14(1):11. [doi: 10.1186/s13012-019-0853-y] [Medline: 30709368]
41. PRS HIV compendium of best practices. Prevention Research Synthesis. URL: https://wwwn.cdc.gov/HIVCompendium/SearchInterventions [accessed 2025-11-14]
42. Coleman M, Hodges A, Henn S, Lambert CC. Integrated pharmacy and PrEP navigation services to support PrEP uptake: a quality improvement project. J Assoc Nurses AIDS Care 2020;31(6):685-692. [doi: 10.1097/JNC.0000000000000182] [Medline: 32541194]
43. Casey E, Kaplan-Lewis E, Gala K, Lakew R. Successful integration of HIV PrEP in primary care and women's health clinical practice: a model for implementation. Viruses 2023 Jun 13;15(6):1365. [doi: 10.3390/v15061365] [Medline: 37376664]
44. Ridgway JP, Devlin SA, Friedman EE, et al. POWER Up—Improving pre-exposure prophylaxis (PrEP) uptake among Black cisgender women in the Southern United States: protocol for a stepped-wedge cluster randomized trial (SW-CRT). PLoS ONE 2023;18(5):e0285858. [doi: 10.1371/journal.pone.0285858] [Medline: 37196008]

Abreviations:

ADAPT-ITT: Assessment, Decision, Adaptation, Production, Topical Experts, Integration, Training, Testing
aOR: adjusted odds ratio
CDC: Centers for Disease Control and Prevention
EHR: electronic health record
ITS: interrupted-time series
MLM: multilevel modeling
OR: odds ratio
PrEP: pre-exposure prophylaxis
PRISM: Practical, Robust, Implementation, and Sustainability Model
RE-AIM: Reach, Effectiveness, Adoption, Implementation, Maintenance
RQA: rapid qualitative analysis
SSDOH: Social and structural determinants of health
W-PrEP: Women's PrEP Project

Edited by A Stone, A Mavragani; submitted 14.07.25; peer-reviewed by MC Pratt, M Pyra; accepted 20.10.25; published 03.12.25.

Please cite as:
Scott RK, Hull SJ, Kerrigan D, Wang Y, Pratt-Chapman M, Mathias-Prabhu T, Unwala N, Sadauskas M, Hose BZ, Smith M, Moriarty P, Hanson T, Tedla A, Lotke P, Ye P, Arem H
"The Women's PrEP Project," a Clinic-Based, Socio-Structural Intervention to Improve the Provision of Preexposure Prophylaxis for Cisgender Women: Interrupted Time Series Pilot
JMIR Form Res 2025;9:e80653
URL: https://formative.jmir.org/2025/1/e80653
doi:10.2196/80653

© Rachel K Scott, Shawnika J Hull, Deanna Kerrigan, Yan Wang, Mandi Pratt-Chapman, Tara Mathias-Prabhu, Naquia Unwala, Marisa Sadauskas, Bat-zion Hose, Marjanna Smith, Patricia Moriarty, Tranessa Hanson, Ariam Tedla, Pamela Lotke, Peggy Ye, Hannah Arem. Originally published in JMIR Formative Research (https://formative.jmir.org), 3.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Effectiveness of Live and Prerecorded Video Demonstrations in Teaching Restorative Dentistry to Undergraduate Students: Cohort Study

Rana Alkattan[1,2,3*], BDS, MSD, PhD; Lulwah Alreshaid[1,2,3*], BDS, PhD

[1]Department of Restorative and Prosthetic Dental Sciences, College of Dentistry, King Saud bin Abdulaziz University for Health Sciences, 7111-Al Rimayah District, Riyadh, Saudi Arabia

[2]Ministry of National Guard Health Affairs, King Abdullah International Medical Research Center, Riyadh, Saudi Arabia

[3]Ministry of the National Guard - Health Affairs, King Abdulaziz Medical City, Dental Services, Riyadh, Saudi Arabia

[*]all authors contributed equally

Corresponding Author:
Lulwah Alreshaid, BDS, PhD
Department of Restorative and Prosthetic Dental Sciences, College of Dentistry, King Saud bin Abdulaziz University for Health Sciences, 7111-Al Rimayah District, Riyadh, Saudi Arabia

Abstract

Background: Mastering complex psychomotor skills is essential in undergraduate dental education; however, traditional live demonstrations (LDs) face limitations such as high instructor-to-student ratios and restricted viewing angles. Prerecorded video demonstrations (VDs) offer scalable, repeatable instructions and the ability to integrate multimedia cues but may lack real-time interaction and immediate feedback. There is limited evidence comparing these teaching modalities, particularly regarding gender differences, in the acquisition of restorative dentistry skills.

Objective: This study aimed to (1) compare first-year dental students' knowledge acquisition and procedural performance following a LD versus a prerecorded VD of a class II amalgam restoration and (2) evaluate whether gender influences outcomes within each demonstration method.

Methods: A total of 51 students enrolled in an Introduction to Operative Dentistry course (2024 - 2025) participated in this cohort study. The students were randomized into 2 groups: LD (26/51, 51%) or VD (25/51, 49%). Both groups received identical lectures and demonstrations of a standardized class II cavity preparation and amalgam restoration. Knowledge was assessed via preprocedural and postprocedural multiple-choice questionnaires, and the procedural performance was graded by 2 blinded raters using a 10-point rubric. Student perceptions were measured with an 8-item Likert survey. Mixed ANOVA and independent and paired 2-tailed *t* tests evaluated between-group and within-group differences, while gender analyses used factorial ANOVA. Interrater reliability (interclass correlation coefficient=0.991) was confirmed.

Results: The baseline knowledge scores did not differ between the 2 groups. After the demonstration, knowledge was significantly higher with LD (mean 71.22, SD 17.3) than VD (mean 58.4, SD 21.7; P=.02; Cohen d=0.65). The LD method demonstrated significant within-group improvement (P<.001; Cohen d=0.83). Procedural grading favored LD (mean 8.3, SD 0.9 vs mean 7.9, SD 1.0); however, results were not statistically significant (P=.08; Cohen d=0.50). No significant differences were found in the student perception survey. Gender analysis revealed that male students in the LD group achieved higher postknowledge scores (mean 74.0, SD 12.3 vs mean 55.0, SD 24.3; P=.03), greater score improvements (P=.03), and higher grading scores (mean 8.5, SD 0.6 vs mean 7.6, SD 1.3; P=.03) compared to those in the VD group. No significant differences were observed among female students.

Conclusions: LDs yielded superior knowledge acquisition and better performance compared to VDs, particularly for male students. VDs remain a viable alternative when supplemented with interactive elements and instructor feedback. Blended teaching models integrating live and video methods may optimize the demonstration experience for the students, thus enhancing the learning outcomes.

(JMIR Form Res 2025;9:e74383) doi:10.2196/74383

KEYWORDS

dental education; prerecorded video; live demonstration; restorative dentistry; gender differences

XSL•FO
RenderX

Introduction

Background

Dental undergraduate education involves learning the science and practice of dentistry, culminating in the mastery of complex procedures that require refined manual skills. This is accomplished through didactic and laboratory-based training, emphasizing the development of intricate psychomotor skills for ultimately performing such procedures in a clinical environment [1]. Dental procedures can be complex and cannot be conveyed using only text and images in presentation formats. Hence, they are often taught by visual observation of the intended procedure in a simulated clinical environment. This typically involves a live, step-by-step demonstration by an instructor to a group of students; thereafter, the students are expected to perform the procedure independently [2].

Traditional live demonstrations (LDs) pose limitations in effectively conveying these skills. LDs offer real-time observation but face challenges such as a high instructor-to-student ratio limiting personalized attention and a limited view of the procedure from a student's perspective due to positioning within the simulated environment [3]. On the other hand, prerecorded video demonstrations (VDs) can present a potential alternative. Their advantages include scalability [4], wherein a wider student audience can access the content, and enhanced realism [5], as the procedures can be performed inside a mannequin head, replicating the clinical setting. However, the reduced interaction between the student and the instructor and the limited opportunity for student-instructor dialogue and real-time clarification and discussions may make students feel disconnected [6]. Furthermore, VDs are also typically filmed from an angle that may be different than the instructor's or students' perspective [7]. Hence, they require a trained instructor skilled in performing procedures while maintaining camera focus and addressing student concerns. Nonetheless, most studies do not describe sufficient information on how a video is presented.

Amid the growth of computer-based e-learning, traditional teaching methods alone may not meet the educational needs of the new generation of dental students [8]. With technological advancement, high-resolution videos showing each step of a procedure from different angles have become possible [9]. The ability to embed texts and images into the recording is also an advantage not offered by LDs [10]. Furthermore, the accessibility of these videos allows students to review specific steps repeatedly at their own pace without the need for instructor dependency. Students may replay any part of the video outside of scheduled classroom hours and have the video stream as they simultaneously perform the procedure, fostering learner autonomy [11]. This method may not only reduce educational costs but also enable teaching anytime and anywhere, shifting the process from teacher-centered to student-centered learning.

While e-learning technologies are increasingly used in undergraduate dental education, limited research exists comparing traditional and modern methods for practical skill development, particularly in restorative dentistry. In dental education, comparing LDs with VDs is important due to ongoing challenges in faculty availability and the need for efficient teaching methods. LDs, while valuable, demand significant faculty time and may vary in delivery between instructors, whereas VDs can help address the faculty shortage by allowing content to be reused across multiple cohorts. Videos also ensure standardization of instruction, providing every student with the same sequence and detail of a procedure. Furthermore, in large classes, videos enhance accessibility and scalability, ensuring all learners have equal opportunities to observe critical steps. By directly comparing both methods, educators can better evaluate learning outcomes, student performances, and cost-effectiveness, guiding decisions on the most effective blend of approaches for dental training. Comparing LDs with VDs among genders is also insightful in evaluating potential differences in learning style preferences and confidence or participation levels. Studies suggest that male and female students may engage differently with instructional formats, with some students benefiting more from interactive, real-time learning, while others prefer the ability to review material independently at their own pace [12]. In addition, LDs can sometimes feel intimidating, limiting participation among students who are less comfortable asking questions in front of peers, whereas VDs provide a private, repeatable way to observe and learn [13]. Examining these differences can help educators design more inclusive teaching strategies that support all students effectively.

Objectives

This research can potentially address the development of blended learning approaches, combining the advantages of LDs and prerecorded VDs for optimal learning outcomes in dental skills training [14,15]. Therefore, the aim of this study is (1) to compare students' performance following hands-on LD of a dental proximal restoration to a prerecorded VD of the same procedure and (2) to compare the knowledge and performance of male and female students within both demonstration groups. We hypothesize that (1) the prerecorded VD will be as effective as the LD and (2) there will be no difference in the knowledge and performance of students based on gender.

Methods

Overview

A total of 53 first-year dental students, enrolled in the Introduction to Operative Dentistry course (RSTO 311) at the College of Dentistry, King Saud bin Abdulaziz University for Health Sciences, Riyadh, Saudi Arabia, were invited to participate in this cohort study during the academic year 2024-2025. The sample size was based on the cohort size of the course. The collection and analysis of students' responses and grades were processed anonymously, and student results were treated confidentially. The students were then quasi-randomly divided into 2 groups based on their student serial number by an independent research assistant not involved in the course: those with odd serial numbers were allocated to LD, while those with even serial numbers were allocated to VD. Allocation was concealed from students and instructors until the start of the intervention. No specific measures were taken to balance gender during randomization.

XSL•FO
RenderX

Ethical Considerations

Institutional review board ethics approval was obtained from King Abdullah International Medical Research Center for this study (IRB/1044/25). The students were asked to participate voluntarily after a signed written consent form was obtained, with the option to withdraw from the study at any time. The collection and analysis of students' responses were hanlded securely. All responses and grades were processed anonymously and student results were treated with respect and confidentiality.

Study Design

The study was performed in the context of a preclinical course for restorative dentistry, Introduction to Operative Dentistry. This year-long course is held once a week and serves as students' first exposure to restorative procedures. It is given in a simulation laboratory where the students are asked to perform different procedures on plastic model teeth (Nissin Dental Products Inc) mounted in the mannequin heads of the simulation units (Planmeca Compact iSim, Planmeca). The students begin the course by learning basic theoretical and practical knowledge on cavity preparations for dental amalgam. Before this study was conducted, all students had previously performed 7 class I cavity preparations and amalgam restorations on upper and lower posterior teeth. This study entailed the performance of their first class II cavity preparation involving the occlusal and mesial surfaces of a mandibular right first molar, proper Tofflemire matrix placement, and amalgam restoration.

Each step of the planned procedure was demonstrated in either of the 2 methods: LD or VD. The dental simulation laboratory comprised 2 rooms, in which the demonstrations were delivered in 2 parallel groups simultaneously, that is, 1 group in each room. The LD included a 30-minute live hands-on demonstration by 1 of the 3 restorative dentists to a group of students, with an instructor-to-student ratio of 1:6, during which both the instructor and the students could comment on the procedure and ask questions. The demonstrations were standardized by having all demonstrators view the prerecorded video given to the VD group, with a checklist outlining the sequence of steps of the intended procedure to ensure that the demonstrators delivered the procedural steps in a uniform and comparable manner. The VD included a prerecorded video of the same procedure performed by a single demonstrator, with a total duration of approximately 30 minutes, shown on individual screens at each student's station. Both the LD and the VD described identical steps involved in preparing and restoring a class II amalgam.

The students participated in the study for 3 consecutive weeks, from February 2025 to March 2025, as part of their regular weekly schedule as per their RSTO 311 course curriculum. The course is given once a week, with 3 sessions correspondingly divided as follows: week 1 (class II amalgam cavity preparation lecture and demonstration), week 2 (class II matrix application lecture and demonstration), and week 3 (class II amalgam restoration lecture and demonstration). Each lecture was given by a single faculty member immediately before the assigned corresponding laboratory session. A flowchart of the study design is shown in Figure 1.

Figure 1. Workflow of the study comparing live demonstration (LD) with video demonstration (VD) in first-year dental students.

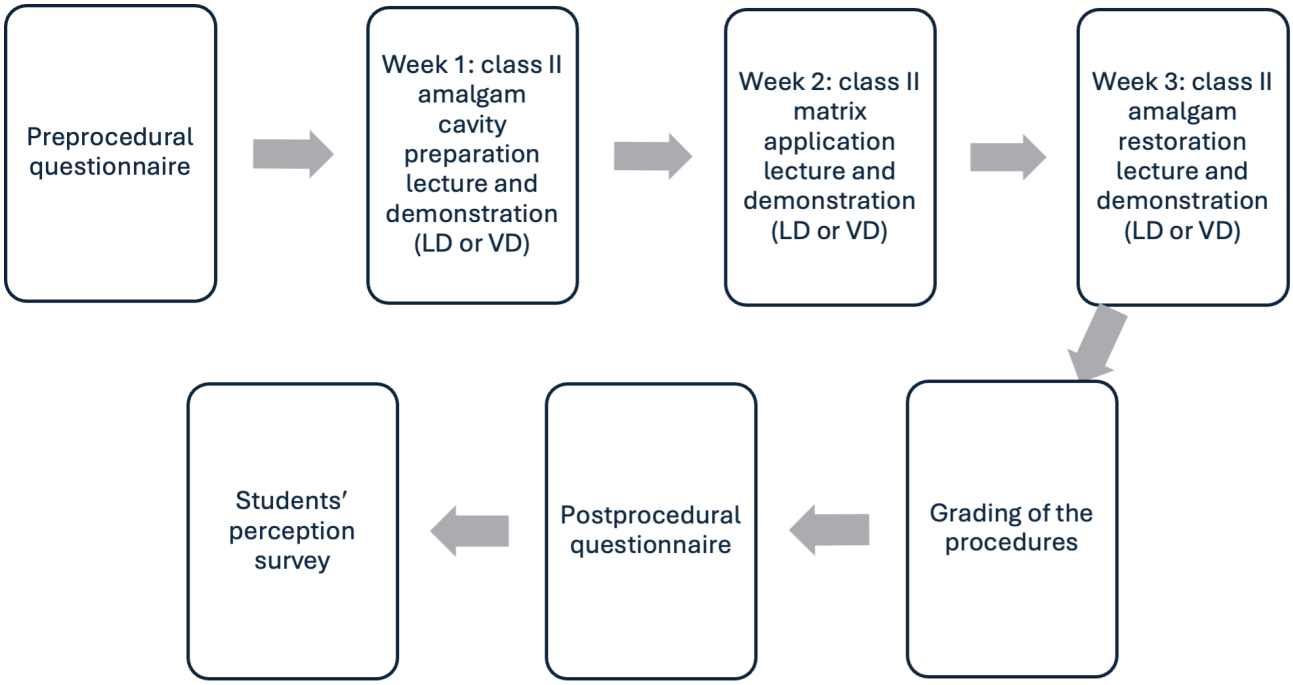

The students first received a preprocedural questionnaire to assess their knowledge before the lecture and demonstration, consisting of 10 four-choice questions about different procedural steps (Multimedia Appendix 1).

The students were then randomly divided into 2 parallel groups: LD or VD. Two blinded instructors not involved in the demonstration then graded the procedure for all students from both groups, and the average grade was recorded. A 10-point rubric was used to grade the procedural steps, with a maximum possible score of 1 for each criterion (Multimedia Appendix 2).

XSL•FO
RenderX

The students were expected to perform the procedure after applying the rubber dam; however, no grade was given for the rubber dam placement. Next, the students received a postprocedural questionnaire to assess their knowledge after the lecture and demonstration, consisting of the same 10 four-choice questions used in the preprocedural questionnaire (Multimedia Appendix 1).

Feasibility was evaluated based on the ability to deliver the intervention within regular course sessions. Students' screens in the VD group were tested before the sessions for any technical difficulties. Acceptability was assessed via an 8-item survey, adapted from a study by Alqahtani et al [15], which was used to measure the students' perception toward the 2 demonstration methods by capturing students' satisfaction, perceived clarity, and confidence in performing the procedure. The survey was based on a Likert scale with 5 response options (1=strongly disagree, 2=disagree, 3=uncertain, 4=agree, and 5=strongly agree).

Statistical Analyses

Mixed ANOVA was used to assess students' knowledge scores from the preprocedural and postprocedural questionnaires (within groups) as well as the difference between groups. Simple main effects with Bonferroni corrections were used to evaluate the differences between groups at each time point as well as changes within each group. Before and after the demonstration, students' knowledge was scored as follows: correct answers received a 1-point score, while erroneous answers received a 0-point score. The validity of the questionnaire was confirmed using the content validity method, which included the opinions of a 5-member expert panel. For the grades achieved for the procedure, mean differences between the groups were assessed using an independent samples *t* test. Two instructors independently used the 10-point evaluation form to assess students' performance. Interrater reliability was measured using the interclass correlation coefficient (ICC) with a 2-way random effects model for absolute agreement ICC (2,1 [indicating a 2-way random, single measure]), as described by McGraw and Wong [16]. The ICC value was 0.991 (95% CI 0.984 - 0.995), indicating excellent reliability and consistency in ratings, based on the benchmarks proposed by Cicchetti [17], where ICC values >0.75 are considered excellent. For analysis purposes, the average value of the 2 evaluation scores was used. Gender differences were assessed by factorial ANOVA. While gender stratification was not performed, the randomization process and the relatively balanced gender composition support the internal validity of the study. Dependent variables were assessed for normality using histograms and the Shapiro-Wilk test. A significance level of 0.05 was used for inferential analysis, with *P* values <.05 reported as statistically significant. Statistical analysis was conducted using the SPSS (version 29; IBM Corp).

Results

Student Enrollment

A total sample of 51 dental students (male students: 22/51, 43.1%; female students: 29/51, 56.9%) agreed to participate in the study. Two male students from the LD group were excluded because they missed at least 1 session of the procedural steps in which an LD was given following the CONSORT (Consolidated Standards of Reporting Trials) guidelines (Figure 2). The students were allocated into 2 groups: LD (26/51, 51%) or VD (25/51, 49%).

XSL•FO
RenderX

Figure 2. CONSORT (Consolidated Standards of Reporting Trials) flowchart illustrating the flow of students through each stage of the randomized study, including enrollment, allocation, follow-up, and analysis. LD: live demonstration; VD: video demonstration.

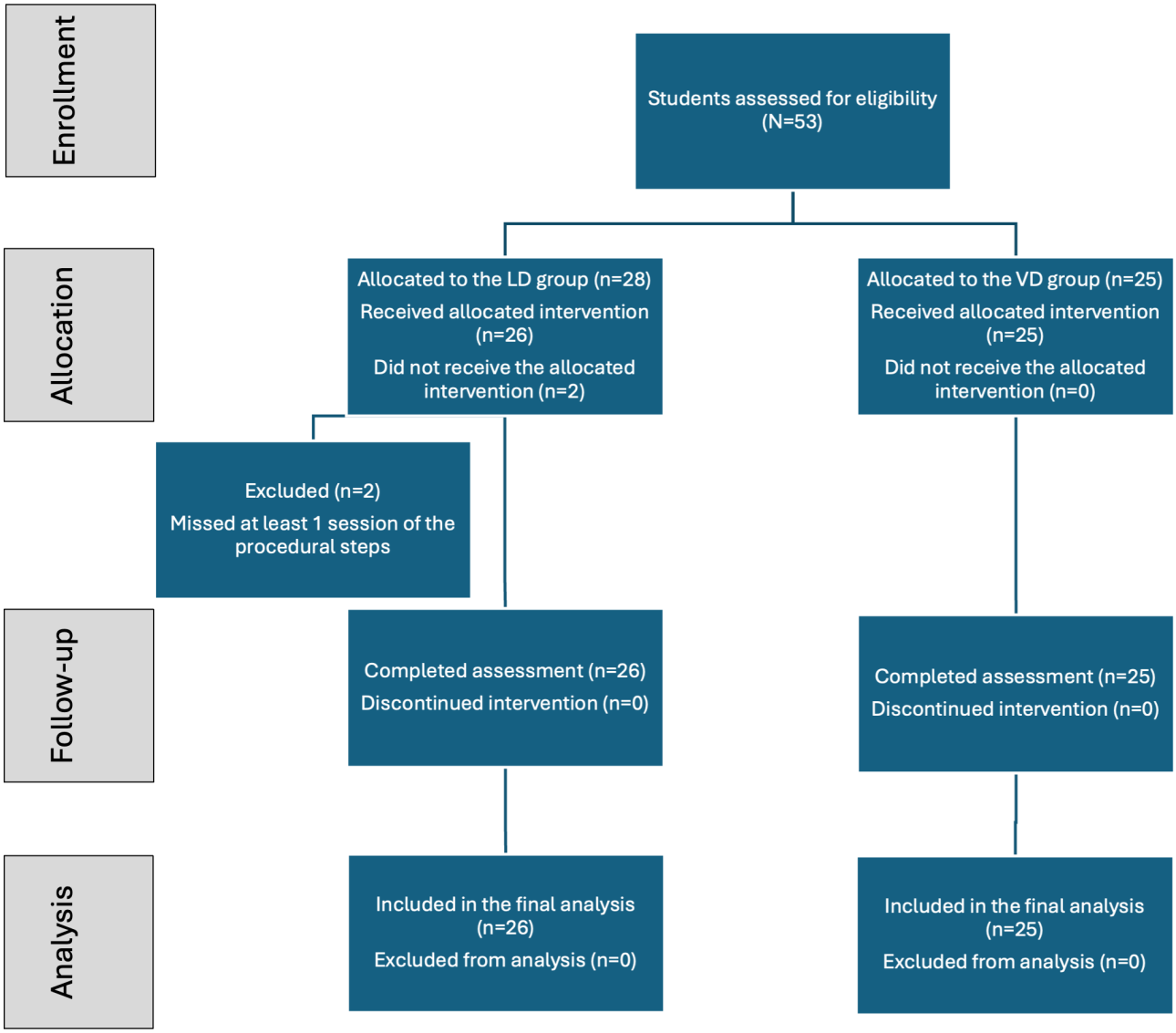

Predemonstration and Postdemonstration Questionnaires

While this quasi-randomization approach may have allowed for hidden biases related to admission order or other systematic factors, the relatively balanced gender composition and prior procedural experiences helped mitigate potential confounding effects. No significant difference was found between the 2 groups in terms of knowledge scores in the predemonstration questionnaire, suggesting that randomization was successful and the 2 study groups were similar in their preintervention characteristics. However, significantly higher postdemonstration knowledge scores were found with LD compared to VD (P=.02). The difference represents a medium-large effect size (Cohen d=0.65). Within-group analysis showed significant improvement in knowledge scores with LD (P<.001) with a large effect size (Cohen d=0.83; Table 1).

XSL•FO
RenderX

Table . Predemonstration and postdemonstration questionnaire scores and procedural grading scores between live demonstration (LD) and video demonstration (VD) groups.

	LD[a] (n=26)	VD[b] (n=25)	Comparison between groups		
			t test (*df*)	*P* value	Cohen *d*
Questionnaire scores					
Predemonstration questionnaire, mean (SD)	53.4 (21.9)	52.4 (20.1)	0.18 (49)	.86	0.05
Postdemonstration questionnaire, mean (SD)	71.2 (17.3)	58.4 (21.7)	2.32 (49)	.02	0.65
Grading scores, mean (SD)	8.3 (0.9)	7.9 (1.0)	1.80 (49)	.18	0.50

[a]Within-group comparison, paired samples *t* test: $F_{1,49}$=16.85; P<.001; Cohen d=0.83.

[b]Within-group comparison, paired samples *t* test: $F_{1,49}$=1.86; P=.18; Cohen d=0.27.

Procedural Grading Scores

Procedural grading scores were also found to be higher in the LD group compared to the VD group, but the difference was not statistically significant (P=.18) and showed a medium effect size (Cohen d=0.50). The results are shown in Table 1.

Survey Question Scores

Table 2 shows that no significant differences were found between the 2 groups in responses to any of the 8 Likert scale survey questions.

XSL•FO
RenderX

Table . Survey question scores between the live demonstration (LD) and video demonstration (VD) groups.

Survey questions	LD (n=26), mean (SD)	VD (n=25), mean (SD)	Comparison between groups		
			t test (*df*)	*P* value	Cohen *d*
Question 1: "I felt stressed when performing the procedure."	2.5 (1.3)	2.6 (1.4)	0.16 (49)	.87	0.05
Question 2: "It was easy to perform the procedure."	3.5 (1.1)	3.5 (1.0)	0.06 (49)	.95	0.02
Question 3: "I was satisfied with my performance in the procedure."	3.6 (1.1)	3.8 (1.1)	0.87 (49)	.39	0.24
Question 4: "The teaching method was helpful to perform the procedure."	4.1 (0.9)	3.9 (0.9)	0.59 (49)	.56	0.17
Question 5: "The teaching method was adequate for performing the procedure."	4.2 (0.8)	3.8 (1.1)	1.41 (49)	.16	0.40
Question 6: "The steps in the teaching method were presented in a clear fashion and easy to understand."	4.0 (1.0)	4.3 (0.8)	0.93 (49)	.36	0.26
Question 7: "I prefer the teaching method that I have received over the other teaching method."	3.3 (1.2)	3.8 (1.2)	1.75 (49)	.09	0.49
Question 8: "The teaching method needs further improvement to support my learning."	3.4 (1.2)	3.4 (1.2)	0.04 (49)	.97	0.01

Gender-Based Analysis

Table 3 demonstrates that no significant differences were found in the predemonstration knowledge scores between the LD and VD groups between the genders. However, significant differences between the LD and VD groups were found among male students regarding postdemonstration knowledge scores (*P*=.03). The LD group showed higher knowledge scores compared to the VD group (*P*=.03; Cohen *d*=0.93, a large effect size). Similarly, the LD group showed more improvement in knowledge scores from before the demonstration to after the demonstration compared to the VD group (*P*=.03; Cohen *d*=0.94, a large effect size). Male students also showed significantly higher grading scores with LD training compared to VD (*P*=.03; Cohen *d*=0.84, a large effect size). However, gender-based analyses were exploratory due to limited sample sizes; interaction terms were included to test for moderation effects, though power to detect small effects was constrained.

XSL•FO
RenderX

Table . Postdemonstration knowledge scores, improvement in knowledge scores, and procedural grading scores between live demonstration (LD) and video demonstration (VD) groups between genders.

	LD	VD	Comparison between groups		
			F test (*df*)	*P* value	Cohen *d*
Predemonstration knowledge scores[a]					
Female, mean (SD)	55.0 (23.1)	51.5 (22.7)	0.19 (1, 47)	.67	0.15
Male, mean (SD)	51.0 (20.8)	53.3 (17.8)	0.07 (1, 47)	.80	0.12
Postdemonstration knowledge scores[b]					
Female, mean (SD)	69.4 (19.1)	61.5 (19.5)	1.13 (1, 47)	.29	0.41
Male, mean (SD)	74.0 (12.3)	55.0 (24.3)	5.03 (1, 47)	.03	0.93
Knowledge scores improvement (change from predemonstration scores to postdemonstration scores)[c]					
Female, mean (SD)	14.4 (22.5)	10.0 (20.0)	0.28 (1, 47)	.60	0.20
Male, mean (SD)	23.0 (19.2)	1.7 (25.2)	5.12 (1, 47)	.03	0.94
Grading scores[d]					
Female, mean (SD)	8.2 (1.1)	8.1 (0.6)	0.20 (1, 47)	.66	0.18
Male, mean (SD)	8.5 (0.6)	7.6 (1.3)	4.76 (1, 47)	.03	0.84

[a]Comparison between genders for LD: $F_{1,47}$=0.22; *P*=.65; Cohen *d*=0.18. Comparison between genders for VD: $F_{1,47}$=0.04; *P*=.84; Cohen *d*=0.09. Gender×group interaction effect: $F_{1,47}$=0.23; *P*=.64.

[b]Comparison between genders for LD: $F_{1,47}$=0.34; *P*=.57; Cohen *d*=0.27. Comparison between genders for VD: $F_{1,47}$=0.68; *P*=.41; Cohen *d*=0.30. Gender×group interaction effect: $F_{1,47}$=0.99; *P*=.33.

[c]Comparison between genders for LD: $F_{1,47}$=0.95; *P*=.34; Cohen *d*=0.40. Comparison between genders for VD: $F_{1,47}$=0.89; *P*=.35; Cohen *d*=0.37. Gender×group interaction effect: $F_{1,47}$=1.84; *P*=.18.

[d]Comparison between genders for LD: $F_{1,47}$=0.67; *P*=.42; Cohen *d*=0.35. Comparison between genders for VD: $F_{1,47}$=1.19; *P*=.28; Cohen *d*=0.42. Gender×group interaction effect: $F_{1,47}$=1.82; *P*=.18.

Discussion

Principal Findings

The main objective of this study was to evaluate the effectiveness of prerecorded VD compared to LD in teaching class II amalgam to first-year dental students. The findings of this study revealed that while both methods improved student knowledge and performance, LD was associated with significantly higher postdemonstration knowledge scores and higher procedural grading scores (Table 1). Therefore, the first null hypothesis was rejected. The second objective of this study was to evaluate differences in the knowledge and performance of students based on gender. The results showed that although there were no statistically significant differences between genders, male students achieved significantly higher postdemonstration knowledge scores with LD (vs VD), as well as higher procedural grading scores (Table 3), rejecting the second null hypothesis.

The significantly higher postdemonstration knowledge scores in the LD group (Table 1) suggest that the interactive nature of LDs may enhance the understanding and retention of procedural steps. The large effect size (Cohen *d*=0.65) observed further highlights the impact of direct instructor-student engagement on knowledge acquisition. The interactive environment provided by LD allows students to ask questions, clarify doubts, and receive immediate corrections, reinforcing learning and reducing misconceptions. Additionally, observing the instructor's technique enhances visual learning and provides a deeper understanding of the procedural steps. This aligns with cognitive learning theories emphasizing active participation and feedback as critical components for knowledge retention [18]. Furthermore, LD fosters a collaborative learning atmosphere, where peer discussions and shared experiences can further solidify understanding.

In contrast, the VD group demonstrated less improvement in knowledge scores (Table 1). Although the prerecorded video allowed students to review the material at their own pace, the lack of in-person clarification and the potential for disengagement may have limited their learning outcomes. Moreover, VD students may struggle with passive learning, where attention can fade over time without active engagement. To tackle these challenges, incorporating interactive elements such as quizzes, guided reflections, and discussions could enhance the effectiveness of the VD method. In addition, providing access to instructors for follow-up questions or clarification can help bridge the gap between the benefits of live interaction and the flexibility of video-based learning.

The procedural grading scores showed higher performance in the LD group (Table 1), although the difference was not statistically significant. This finding may be attributed to the hands-on guidance during LDs, enabling students to receive immediate corrections and refine their techniques. Nonetheless,

XSL•FO
RenderX

the medium effect size (Cohen *d*=0.50) observed indicates that VD may still be a viable alternative for developing the needed skills, especially when combined with supplementary support, such as follow-up instructor feedback sessions. For instance, integrating video-based learning with practice sessions and virtual simulations can help reinforce muscle memory and improve technique [19]. Additionally, structured peer feedback and self-assessment tools can enhance learning outcomes and provide opportunities for skill refinement [20].

It is also worth noting that VD allows students to pause, rewind, and rewatch the demonstration, enabling repetitive practice and gradual skill acquisition. This flexibility is particularly beneficial for students who need extra time to master specific techniques, prefer self-paced learning, or were unable to attend the class.

The survey results indicated no significant differences between the groups in perceived ease of performing the procedure, performance satisfaction, and clarity of teaching methods. However, the VD group tended to prefer their assigned method over LD (Table 2). Individual learning styles and personal comfort with technology may influence the preference for the VD method. For instance, visual and auditory learners may find video-based content more appealing, and students who experience anxiety in live settings may feel more at ease [21].

While male students in the LD group showed higher postdemonstration scores and procedural grades compared to VD, the nonsignificant group×gender interactions indicated that these findings should be interpreted cautiously. Although the small subgroup sizes limit the generalizability of gender-specific conclusions, significant differences between the LD and VD groups were evident among male students, with the LD method leading to higher postdemonstration knowledge scores and more significant improvements from pre to postdemonstration scores. Male students also performed significantly better in the LD group, as demonstrated by their grading scores (Table 3). However, no significant differences were found among female students, indicating that both demonstration methods were equally effective for this group.

Comparison With Previous Work

The significantly higher postdemonstration knowledge scores in the LD group compared to the VD group suggest that LD allows the instructor to engage with the students through questions, answers, and discussions, ensuring they understood each step in the presented treatment. This finding is consistent with previous studies emphasizing the benefits of real-time interaction and immediate instructor feedback to aid in concept visualization and maintain student concentration [2,3,22]. Schlafer et al [3] compared hands-on LDs with video-supported demonstrations for teaching aesthetic composite restorations. Their method involved guiding one group of students through a live, hands-on demonstration, while the other group watched a prerecorded video of the same procedure, with the instructor pausing to ask and answer questions and initiate discussions in both groups. Although they did not assess the students' theoretical knowledge using any test or grade the restorations performed by the students, they found that the video group showed more engagement during in-class discussions and also discussed patient-related aspects more than the hands-on group. Another study by Ramlogan et al [14] similarly compared LD and VD for teaching clinical periodontology to undergraduate dental students. Their results showed that students achieved similar pretest knowledge scores in both groups; however, the mean posttest scores were significantly greater for the live lecture group. Surprisingly, more students favored VDs over live lectures, suggesting the potential of videos as an educational technology to complement live teaching and serve as a means for reaching teaching and learning goals.

In contrast, the VD group in this study demonstrated less improvement in knowledge scores. While VDs offer benefits like enhanced visualization and flexibility, they can also create a detachment from the clinical setting and reduce direct interaction between students and instructors, potentially hindering immediate feedback and personalized guidance. This finding aligns with published literature indicating that video-based learning can be effective but requires additional support mechanisms to maximize student engagement and comprehension [9,23,24]. Rystedt et al [9] highlighted the significance of the reciprocal relationship between visualization, instruction, and interaction in video-based demonstrations, which ultimately contributes to an effective learning experience for students. First, they explained that visualization refers to allowing students to witness treatments from the dentist's perspective, offering a detailed and magnified view of intricate procedures. Second, they emphasize that the teacher's instruction is key to guide students' attention to critical aspects of the treatment, explain the reasoning behind decisions, and connect the demonstration to theoretical concepts. Lastly, they emphasize the need to incorporate an interactive setting with video-based demonstrations, including discussions between teachers and students, fostering student engagement and facilitating a deeper understanding of the material. The shared visual field created by the video can allow for focused discussions and timely clarifications, which enhances the learning process.

Students' prior experience with multimedia learning, their familiarity with technology, and their baseline motivation can all influence how well they learn from VDs compared to LDs. These factors can affect their engagement; understanding; and, ultimately, their learning outcomes. Students who are already comfortable with learning through videos and other multimedia formats may adapt more quickly to VDs in dental education [25]. Furthermore, students who are already highly motivated to learn may be more likely to engage with both LD and VD. However, students who are less motivated might benefit more from the interactive aspects of LDs, which can provide a more engaging and stimulating learning experience [26]. These factors can influence a student's ability to understand and retain information as well as develop practical skills. For example, a student who is comfortable with multimedia learning might grasp complex dental procedures more quickly from a VD, while another student might need the hands-on experience of an LD to fully understand the nuances of the technique [27].

While there were no significant differences observed between the groups in how easy they found the procedure, how satisfied they were with their performance, or how clear the teaching methods were, survey results demonstrated that the VD group expressed a tendency to prefer their learning method over the

XSL•FO
RenderX

LD group. This finding suggests that some students may value the flexibility and accessibility of video-based learning, even if it does not lead to superior performance outcomes. This is supported by the self-determination theory, which highlights the importance of autonomy in enhancing motivation and learning outcomes [28]. Furthermore, a previous meta-analysis similarly indicated that students with control over their learning pace and environment report higher satisfaction and engagement [29]. A similar study by Alqahtani et al [15] found that while a procedural video for fabricating an orthodontic Adams clasp was as effective as an LD in terms of mean scores in a laboratory exercise, students rated the video higher in perceived clarity and ease of understanding. This may be because the procedural video offered better visualization of the laboratory steps and allowed students to review the procedure independently before, during, or after the laboratory sessions, free from crowding.

The study's findings also suggest that LDs' interactive and dynamic style may be particularly beneficial for male students. Published literature indicates that male students tend to prefer hands-on, experiential learning environments that provide immediate feedback and active engagement [30]. The opportunity to ask questions and receive in-person guidance during LD may cater to these preferences, enhancing their learning outcomes [31]. On the other hand, no significant differences were found among female students, indicating that they performed equally in both LD and VD groups. Previous studies demonstrate that female learners may excel in self-paced, independent learning environments, which could explain their equal performance in both groups when compared to male learners [32]. Moreover, female students may demonstrate higher levels of self-regulation and metacognitive strategies, allowing them to overcome the limitations of the VD method [33].

Al-Sbei et al [34] evaluated the impact of integrating a web-based course on the restorative dentistry knowledge of graduate students in Syria, which included theoretical and educational videos. The results showed a significant increase in the participants' levels of restorative dentistry knowledge after completing the web-based course. This demonstrates that video-based demonstrations may have implications beyond traditional classroom teaching, particularly in cases with difficulty in transportation, lack of qualified places to hold dental courses, and staff shortages. Another study by Kanzow et al [35] evaluated the restorative dentistry knowledge of undergraduate students using screen-captured presentations with narrated audio uploaded onto an online platform during the COVID-19 pandemic. Students viewed the presentations more frequently than they could have attended a conventional lecture and subsequently demonstrated better performance in theoretical knowledge tests in restorative dentistry–related topics compared to other dental specialties. Future research should explore the integration of blended learning approaches that combine the advantages of both LD and VD, such as incorporating interactive video modules and virtual simulations. The opportunity to engage with students under the supervision of an instructor also allows for in-person feedback and error correction, which are crucial for developing psychomotor skills. This aligns with previous findings that emphasize the importance of experiential learning in technical education [36,37].

In this study, gender-based differences warrant further investigation into the underlying factors influencing learning outcomes. Future studies could explore the role of learning preferences, cognitive strategies, and prior experiences, which may influence the effectiveness of different instructional methods and help explain potential gender-related differences. Lastly, it is worth exploring personalized learning styles tailored to individual preferences and learning needs.

Limitations

The study has several limitations. First, the relatively small sample size may have limited the generalizability of the findings, specifically related to gender analysis, and further research with larger samples is needed to explore potential moderating effects of gender. Second, the study focused on a single dental procedure, and the results may not apply to other aspects of dental education. Third, the study relied on self-reported surveys for student perceptions, which may introduce response bias. Fourth, factors such as prior experience with video-based learning were not controlled, which could influence the study's outcomes. Lastly, blinding instructors and students was not feasible due to the nature of LD versus VD; however, the raters assessing student performance were blinded to group allocation. This could introduce performance or expectation bias, as students' motivation or engagement may have been influenced by group assignment. However, to minimize potential bias, all demonstrations were standardized using the same procedural steps, and grading was performed by 2 independent, blinded instructors using a detailed rubric.

Conclusions

Within the limitations of this study, it can be concluded from the findings that the LD method was more effective in enhancing knowledge acquisition compared to the prerecorded VD method. However, the VD method remains a second-best alternative when needed, particularly when supplemented with interactive elements and instructor feedback. It was also evident that male students benefited more from the LD method than female students, an interesting finding that needs further exploration. Based on the results of this study, students' demonstration experience can be enhanced by incorporating hybrid learning models. Combining LD with interactive VD modules can improve knowledge acquisition and procedural performance. Furthermore, interactive elements may be added to make the VD method as effective as the LD, such as quizzes, guided reflections or discussions, and instructor feedback to improve engagement and comprehension.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Preprocedural and postprocedural knowledge questionnaires.
[DOCX File, 21 KB - formative_v9i1e74383_app1.docx]

Multimedia Appendix 2
Criteria for grading procedural steps.
[DOCX File, 16 KB - formative_v9i1e74383_app2.docx]

References

1. Hauser AM, Bowen DM. Primer on preclinical instruction and evaluation. J Dent Educ 2009 Mar;73(3):390-398. [Medline: 19289728]
2. Horst JA, Clark MD, Lee AH. Observation, assisting, apprenticeship: cycles of visual and kinesthetic learning in dental education. J Dent Educ 2009 Aug;73(8):919-933. [Medline: 19648563]
3. Schlafer S, Pedersen K, Jørgensen JN, Kruse C. Hands-on live demonstration vs. video-supported demonstration of an aesthetic composite restoration in undergraduate dental teaching. J Dent Educ 2021 Jun;85(6):802-811. [doi: 10.1002/jdd.12541] [Medline: 33527377]
4. Fakhry A, Dehkordi-Vakil FH. Video-Assisted Clinical Instruction in Dentistry (VACID) enhances real-time visualisation of clinical procedures. Eur J Dent Educ 2007 Nov;11(4):238-244. [doi: 10.1111/j.1600-0579.2007.00465.x] [Medline: 17935564]
5. Boucheix JM, Gauthier P, Fontaine JB, Jaffeux S. Mixed camera viewpoints improve learning medical hand procedure from video in nurse training? Comput Human Behav 2018 Dec;89:418-429. [doi: 10.1016/j.chb.2018.01.017]
6. Iwaki M, Kanazawa M, Sunaga M, Kinoshita A, Minakuchi S. Live broadcast lectures on complete denture prosthodontics at Tokyo Medical and Dental University: comparison of two years. J Dent Educ 2013 Mar;77(3):323-330. [doi: 10.1002/j.0022-0337.2013.77.3.tb05473.x] [Medline: 23486896]
7. Packer ME, Rogers JO, Coward TJ, Newman PS, Wakeley R. A comparison between videotaped and live demonstrations, for the teaching of removable partial denture procedures. Eur J Dent Educ 2001 Feb;5(1):17-22. [doi: 10.1034/j.1600-0579.2001.005001017.x] [Medline: 11168489]
8. Zitzmann NU, Matthisson L, Ohla H, Joda T. Digital undergraduate education in dentistry: a systematic review. Int J Environ Res Public Health 2020 May 7;17(9):3269. [doi: 10.3390/ijerph17093269] [Medline: 32392877]
9. Rystedt H, Reit C, Johansson E, Lindwall O. Seeing through the dentist's eyes: video-based clinical demonstrations in preclinical dental training. J Dent Educ 2013 Dec;77(12):1629-1638. [Medline: 24319135]
10. Alizadeh A, Sheikhi M, Vakili MM, Nourian A. Live demonstration versus multimedia: a comparison of two educational methods of orthodontic dental cast trimming. J Adv Med Educ Prof 2022 Apr;10(2):120-125. [doi: 10.30476/JAMP.2021.91561.1450] [Medline: 35434149]
11. Zhang D, Zhou L, Briggs RO, Nunamaker JF. Instructional video in e-learning: assessing the impact of interactive video on learning effectiveness. Information & Management 2006 Jan;43(1):15-27. [doi: 10.1016/j.im.2005.01.004]
12. Gatt G, Attard NJ. Multimodal teaching methods for students in dentistry: a replacement for traditional teaching or a valuable addition? A three-year prospective cohort study. BMC Med Educ 2023 Jun 2;23(1):401. [doi: 10.1186/s12909-023-04377-z] [Medline: 37268949]
13. Feng Q, Luo H, Li W, Chen T, Song N. Effects of gender diversity on college students' collaborative learning: from individual gender to gender pairing. Heliyon 2023 Jun;9(6):e16237. [doi: 10.1016/j.heliyon.2023.e16237] [Medline: 37265618]
14. Ramlogan S, Raman V, Sweet J. A comparison of two forms of teaching instruction: video vs. live lecture for education in clinical periodontology. Eur J Dent Educ 2014 Feb;18(1):31-38. [doi: 10.1111/eje.12053] [Medline: 24423173]
15. Alqahtani ND, Al-Jewair T, Al-Moammar K, Albarakati SF, ALkofide EA. Live demonstration versus procedural video: a comparison of two methods for teaching an orthodontic laboratory procedure. BMC Med Educ 2015 Nov 4;15(1):199. [doi: 10.1186/s12909-015-0479-y] [Medline: 26537393]
16. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychol Methods 1996;1(1):30-46. [doi: 10.1037//1082-989X.1.1.30]
17. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol Assess 1994;6(4):284-290. [doi: 10.1037//1040-3590.6.4.284]
18. Chi MTH, Wylie R. The ICAP framework: linking cognitive engagement to active learning outcomes. Educ Psychol 2014 Oct 2;49(4):219-243. [doi: 10.1080/00461520.2014.965823]

XSL•FO
RenderX

19. Cook DA, Levinson AJ, Garside S, Dupras DM, Erwin PJ, Montori VM. Instructional design variations in internet-based learning for health professions education: a systematic review and meta-analysis. Acad Med 2010 May;85(5):909-922. [doi: 10.1097/ACM.0b013e3181d6c319] [Medline: 20520049]
20. Yardley S, Teunissen PW, Dornan T. Experiential learning: transforming theory into practice. Med Teach 2012;34(2):161-164. [doi: 10.3109/0142159X.2012.643264] [Medline: 22288996]
21. Bishop J, Verleger M. The flipped classroom: a survey of the research. Presented at: 2013 ASEE Annual Conference & Exposition; Jun 23-26, 2013; Atlanta, Georgia. [doi: 10.18260/1-2--22585]
22. Zhang XY, Arata Found A, Butler S. Effects of distance-learning strategies in dental fixed prosthodontics amidst the COVID-19 pandemic: cross-sectional questionnaire study on preclinical dental students' perspective. JMIR Form Res 2023 Nov 8;7:e45311. [doi: 10.2196/45311] [Medline: 37938882]
23. Boucheix JM, Lowe RK, Putri DK, Groff J. Cueing animations: dynamic signaling aids information extraction and comprehension. Learn Instr 2013 Jun;25:71-84. [doi: 10.1016/j.learninstruc.2012.11.005]
24. Lee M, An SY, Ihm J. Dental students' satisfaction with web-based learning during the initial phase of the COVID-19 pandemic: mixed methods study. J Med Internet Res 2024 Mar 8;26:e50278. [doi: 10.2196/50278] [Medline: 38457808]
25. AlKahtani RN, Alnufaiy BM, Albaijan RS, Alnafaiy SM, Elfakhri FM, Aljudaibi SM. Comparing the efficacy of live vs. video instructional demonstrations in dental education: a systematic review and meta-analysis. BMC Med Educ 2025 Jan 22;25(1):108. [doi: 10.1186/s12909-025-06672-3] [Medline: 39844197]
26. French S, Ravn S, Balcaite E, Moore E. Understanding students' views on the efficacy of video technology to promote engagement in higher education. PJTEL 2023 Jul;1(2):1-14. [doi: 10.24135/pjtel.v1i2.172]
27. El Assraoui K, Al jalil Z, Cheikh Y, Kaoun K, Chaouir S, Bellemkhannate S. Perception of the contribution of video demonstrations in dental practical training. EJEDU 2024 Jan;5(1):1-6. [doi: 10.24018/ejedu.2024.5.1.780]
28. Deci EL, Ryan RM. The "What" and "Why" of goal pursuits: human needs and the self-determination of behavior. Psychol Inq 2000 Oct;11(4):227-268. [doi: 10.1207/S15327965PLI1104_01]
29. Means B, Toyama Y, Murphy R, Bakia M. Evaluation of evidence-based practices in online learning: a meta-analysis and review of online learning studies. US Department of Education. 2010. URL: https://www2.ed.gov/rschstat/eval/tech/evidence-based-practices/finalreport.pdf [accessed 2025-09-10]
30. Michael G, Stevens K. The Minds of Boys: Saving Our Sons from Falling Behind in School and Life: John Wiley & Sons; 2010.
31. O'Neill G, McMahon T. Student-centred learning: what does it mean for students and lecturers? In: O'Neill G, Moore S, McMullin B, editors. Emerging Issues in the Practice of University Learning and Teaching: All Ireland Society for Higher Education; 2005.
32. Belenky MF, Clinchy BM, Goldberger NR, Tarule JM. Women's Ways of Knowing: The Development of Self, Voice, and Mind: Basic Books; 1986.
33. Schunk DH, Zimmerman B. Handbook of Self-Regulation of Learning and Performance, 2nd edition: Taylor & Francis; 2011.
34. Al-Sbei R, Ataya J, Jamous I, Dashash M. The impact of a web-based restorative dentistry course on the learning outcomes of dental graduates: pre-experimental study. JMIR Form Res 2024 Mar 5;8:e51141. [doi: 10.2196/51141] [Medline: 38441921]
35. Kanzow P, Krantz-Schäfers C, Hülsmann M. Remote teaching in a preclinical phantom course in operative dentistry during the COVID-19 pandemic: observational case study. JMIR Med Educ 2021 May 14;7(2):e25506. [doi: 10.2196/25506] [Medline: 33941512]
36. Ericsson KA. Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. Acad Med 2004 Oct;79(10 Suppl):S70-S81. [doi: 10.1097/00001888-200410001-00022] [Medline: 15383395]
37. Shea CH, Wulf G, Whitacre CA, Park JH. Surfing the implicit wave. Q J Exp Psychol A 2001 Aug;54(3):841-862. [doi: 10.1080/713755993] [Medline: 11548037]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
ICC: intraclass correlation coefficient
LD: live demonstration
VD: video demonstration

XSL•FO
RenderX

Edited by J Sarvestan; submitted 23.03.25; peer-reviewed by A Hart, S Narasimhan; revised version received 20.08.25; accepted 26.08.25; published 25.09.25.

Please cite as:
Alkattan R, Alreshaid L
The Effectiveness of Live and Prerecorded Video Demonstrations in Teaching Restorative Dentistry to Undergraduate Students: Cohort Study
JMIR Form Res 2025;9:e74383
URL: https://formative.jmir.org/2025/1/e74383
doi:10.2196/74383

© Rana Alkattan, Lulwah Alreshaid. Originally published in JMIR Formative Research (https://formative.jmir.org), 25.9.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Changing Trends in Suicidal Ideation and Its Influencing Factors During the Transition From Quarantine to Post-Quarantine Among Chinese University Students During the COVID-19 Surge: Six-Wave Panel Study

Lijing Li[1], MD; Tingzhong Yang[2,3,4], MD; Sihui Peng[5], PhD; Randall R Cottrell[6], DEd

[1]Department of Obstetrics and Gynecology, Yongkang Women and Children's Health Hospital, Yongkang, China

[2]Maternal and Child Healthcare Service Department, Yongkang Women and Children's Health Hospital, Yongkang, China

[3]Research Center for Digital Health Behavior Theory and Management, Zhejiang University National Health Big Data Institute, Hangzhou, China

[4]Injury Control Research Center, West Virginia University, Morgantown, WV, United States

[5]Department of Public Health and Preventive Medicine, School of Medicine, Jinan University, No. 601 Huangpudadao West, Guangzhou, China

[6]Public Health Studies Program, School of Health and Applied Human Sciences, University of North Carolina Wilmington, Wilmington, NY, United States

Corresponding Author:
Sihui Peng, PhD
Department of Public Health and Preventive Medicine, School of Medicine, Jinan University, No. 601 Huangpudadao West, Guangzhou, China

Abstract

Background: To mitigate the rapid spread of COVID-19, numerous countries have adopted lockdowns and quarantine measures. Despite their public health benefits, the effects of these measures on suicidal ideation have not been well documented.

Objective: This study aims to examine the relationship among COVID-19 infection, perceived beliefs, uncertainty stress, and suicidal ideation during the transition from quarantine to post-quarantine periods amid China's COVID-19 surge.

Methods: A prospective longitudinal observational design was used. Changing trends across the 6 time points were assessed using the Mann-Kendall test and the Cochran-Armitage test. A generalized estimating equation was used to analyze the associations between independent variables and suicidal ideation.

Results: A total of 221 (96.5%) participants completed all 6 observation waves. The prevalence of suicidal ideation during the quarantine period was 16.7% (n=37), 14.5% (n=32), and 14.5% (n=32), while during the post-quarantine period, it was 13.8% (n=30), 10.9% (n=24), and 10.0% (n=22), respectively. A significant downward trend in suicidal ideation was observed. In contrast, perceived risk, perceived severity, and the number of new infections exhibited significant upward trends (z scores of 9.56, 7.13, and 3.69, respectively; $P<.001$, $P<.001$ and $P=.002$, respectively). However, uncertainty stress remained stable over time ($z=0.71$; $P=.48$). The generalized estimating equation indicated that perceived risk ($\beta=0.5482$; $P<.001$), perceived severity ($\beta=0.0817$; $P=.007$), and uncertainty stress ($\beta=0.1776$; $P<.001$) were positively associated with suicidal ideation. The number of new infections ($\beta=0.0041$; $P=.49$) was not significantly associated with suicidal ideation.

Conclusions: This study found that suicidal ideation gradually declined following the lifting of quarantine measures. Perceived risk, perceived severity, and uncertainty stress, rather than numbers of infected cases, were significantly associated with suicidal ideation. These findings highlight the importance of addressing individuals' perceptions with real-world context when developing effective strategies to manage COVID-19 and future infectious disease outbreaks.

(JMIR Form Res 2025;9:e74370) doi:10.2196/74370

KEYWORDS

COVID-19; lockdown; suicide; uncertainty; stress

Introduction

The COVID-19 pandemic is a global public health emergency that poses a great threat to people's health and causes large disruption to people's lives and work. In order to curb the rapid spread of COVID-19, many countries have implemented "lockdown" policies [1-4]. "Lockdown" is an emergency protocol that prevents the public from moving from one area to another. In China, complete and partial "lockdowns" were also implemented in some cities [5,6]. These public health emergency measures effectively prevented the rapid growth of the

XSL•FO
RenderX

COVID-19 infection and greatly limited the epidemic of the disease in China [6].

While lockdown measures are crucial in containing the highly infectious COVID-19 virus, they can also have a negative impact on mental health. Quarantine, regardless of the cause, has been associated with various mental and behavioral problems [3,4,7], including those related to the COVID-19 situation [1,3-5]. A review reported that quarantine revealed a positive association with posttraumatic stress disorder, anxiety, and depression [8]. Some studies have also found a link between quarantine and subsequent anxiety and depression [9,10]. A study conducted in China showed that personal quarantine increased anxiety, fear, and anger among individuals [5]. However, these studies primarily relied on cross-sectional observations, making it challenging to establish causal relationships. To address this gap in the literature, we investigate the changing trends of mental health issues and their associations during the transition from quarantine to post-quarantine periods in the COVID-19 epidemic in China.

According to the Stimulus-Cognition-Mental Health (SCM) theory, any stimulus may lead to an awareness of the serious threat posed by the disease (cognition), which in turn can trigger mental responses, ultimately resulting in mental and behavioral problems [11,12]. In the context of personal quarantine and its impact on mental and behavioral problems, the research framework should encompass stimuli, perceived beliefs, and mental health issues, facilitating the identification of causal relationships.

Obviously, the quarantine is a direct stimulus in the theory in this study, which may induce mental response, uncertainty stress, perceived risk, and perceived severity of COVID-19, ultimately impacting people's mental health issues [13]. Many studies found that quarantine is positively associated with mental and behavioral problems [4,5,8-10]. This study's focus is on how quarantine affects mental problems through mental response. Another stimulus is new COVID-19 infection in the period of observation. When people are infected during observation, it is a great shock to them, undoubtedly leading to a strong mental response, which can lead to mental disorder. It is inevitable that people would experience dual stimuli, quarantine, and COVID-19 infection in this study. Our task is to observe their respective roles in the SCM model.

This study hypothesizes that both quarantine and COVID-19 infection act as a direct stimuli, which may induce perceived risk and perceived severity of COVID-19, uncertainty stress, ultimately impacting people's suicidal ideation [13]. Many studies reported that quarantine and COVID-19 infection influenced mental and behavioral problems [1,3,5,7,9], but these studies only included stimuli (quarantine or COVID-19 infection) and outcome variables (mental problems) in their research framework. This study will include more aspects, stimuli, perceived beliefs and uncertainty stress, and mental problems.

According to the Health Belief Model, the health behavior of every individual is motivated and influenced by core belief variables [11,12]. Previous studies have demonstrated a close association between strong perceptions of risk and threat associated with COVID-19 and mental and behavioral outcomes [13-16]. In this study, we examine individuals' perceived risk and perceived severity of COVID-19 during quarantine.

Festinger [17] hypothesized that uncertainty can lead to cognitive dissonance. When individuals are confronted with intense and prolonged uncertainty stress, it can have detrimental effects on their mental and behavioral well-being [18]. The COVID-19 pandemic is characterized by a high degree of uncertainty, as it emerged suddenly, spread rapidly, and then declined abruptly. These factors contribute to a significant level of uncertainty stress. Numerous studies have found a positive association between uncertainty stress and mental and behavioral problems during the COVID-19 pandemic [18-21]. Some studies have specifically highlighted higher levels of mental stress among individuals who were under quarantine [22,23], although few have directly examined the role of uncertainty stress. In addition, for individuals who were not under quarantine, Vandoros and Kawachi [24] found a positive association between economic uncertainty and suicidal ideation. Another study demonstrated a link between uncertainty stress and suicidal thoughts [25].

The concentrated and prominent manifestation of mental problems is suicide, which is closely related to mental stress, anxiety, and depression across many populations [26-28]. It would be established that there is an association between quarantine and suicide, as well as other mental health problems [1,3,5]. However, so far, no studies have examined this relationship. Although suicidal ideation may have a different underlying mechanism than actualized suicide, it is the most proximate precursor and a necessary condition for suicide [27,29]. In this study, we aim to examine the status of uncertainty stress during quarantine in the COVID-19 surge in China and its influence on suicidal ideation. By investigating uncertainty stress as a specific dimension of the overall stress experienced during quarantine, we can gain a deeper understanding of its impact on individuals' mental well-being and the potential risk of suicidal ideation.

Methods

Study Design

A prospective longitudinal observation study was designed to examine temporal trends and changes in suicidal ideation over a 6-week period of time and their associations with selected stimulus and support variables from the quarantine to the post-quarantine period among university students during the COVID-19 surge in China.

Setting

This study was conducted during the COVID-19 lockdown period in Guangzhou, China, with a specific focus on university students who were under quarantine. These students were confined to their residential buildings and were not permitted to move freely in or out of these areas. Lockdown measures began on October 24, 2022, affecting all areas of Haizhu District and parts of Panyu and Tianhe Districts. Restrictions were gradually lifted starting on November 27, 2022, in Haizhu and Tianhe, with the entire city of Guangzhou fully released from

XSL•FO
RenderX

lockdown by November 30. Notably, university campuses did not immediately lift restrictions but instead gradually eased them over 1 week.

Our observation started in later stages after the lockdown was implemented, which may capture the cumulative mental effects of the lockdown. This panel study included 6 waves of data, which were collected over 6 weeks. Wave 1 (November 19, 2022), Wave 2 (November 24, 2022), and Wave 3 (December 1, 2022) are quarantine periods; Wave 4 (December 8, 2022), Wave 5 (December 16, 2022), and Wave 6 (December 23, 2022) are post-quarantine periods.

Participants

Participants were recruited through a survey advertisement on the campus Bulletin Board System, a social media platform at the university used for both public announcements and internet-based social interaction. Inclusion criteria included the following: (1) quarantine university students, (2) having access to a smartphone, (3) knowing the Chinese language, and (4) being willing to participate in the panel study and provide follow-up information at 6 scheduled observation points. Participants were excluded if they refused to provide this information or had a medical condition that could limit or preclude their participation.

Within the registration system, potential participants were screened to ascertain eligibility. Upon consent, participants received an electronic questionnaire and instructions on how to proceed. After reading the instruction, they were asked to provide an electronic consent by tapping the "Confirmation and Authorization" button to opt in. After authorizing the consent, they were directed to the questionnaire and began answering the survey.

A special administrative WeChat (Tencent) group was established to manage the follow-up data collection, using a unique QR code for each respondent. The QR code was the vehicle, not only for identifying unique participants but also for prohibiting nonparticipants from taking the survey. After scanning the QR code, survey participants could enter the investigation group without further preconditions. All responses were anonymous. The questionnaire took 10 minutes to complete, and the same survey protocol was used for each wave of the survey to ensure homogeneity of data administration and collection.

Measurement

Dependent Variable

The self-report suicidal ideation measure used in this study was based on the response to the following question: "Did you seriously consider suicide during the past year?" Options were dichotomous: "yes" or "no." "No" was coded as 1, and "yes" was coded as 2 [21,27].

Independent Variables

Demographics

Age, gender, ethnicity, major, parental occupation, and personal monthly spending.

Perceived Risk and Perceived Severity of COVID-19

They came from 2 key concepts in the Health Belief Model [11,12]. Perceived risk was measured by the question, "Do you feel you are always at risk of being infected with COVID-19?" Perceived severity was measured by a statement: "Infection with COVID-19 has serious health consequences." Responses were on a 5-point Likert-type scale ranging from "strongly disagree" to "strongly agree."

Uncertainty Stress

Uncertainty stress was measured using a scale designed by Yang et al [30], which has demonstrated acceptable validity and has since been used extensively in Chinese research [30,31]. It covered 4 items: current life uncertainty ("life is unstable and cannot be controlled"), social change uncertainty ("uncertain about what will happen in the future"), goals uncertainty ("uncertain about how to achieve goals"), and social values uncertainty ("cannot follow social values").

Respondents rated these items on a 5-point scale from feel no stress (0), a little stress (1), some stress (2), and considerable stress (3) to very strong stress (4). A total stress score was obtained by adding up the responses to the individual questions. The higher the total score, the greater the perceived level of the uncertainty stress [30,31]. In this study, we estimated the reliability and validity of uncertainty stress questionnaire at the 6 different observation time points. The Cronbach α coefficient was 0.85, 0.89, 0.88, 0.89, 0.90, and 0.90, respectively, suggesting acceptable reliability. Fact analysis was conducted in the following steps. For the Kaiser-Meyer-Olkin measure of sampling adequacy, the lowest of the values at each of the 5 survey waves was 0.85. The Bartlett test of sphericity for all waves was significant ($P<.001$), suggesting the samples were factorable. One factor was extracted at each wave, accounting for 68% (the lowest value in the 6 waves) of the variance. The lowest value of factor loading of these items was 0.68, 0.66, 0.76, and 0.62 at the 6 observation time points, indicating good construct validity.

Data Analysis

All data were entered into a database using Microsoft Excel. They were then imported into SAS (version 9.4; SAS Institute Inc) for the statistical analysis. As most of the continuous variables included in this study were not normally distributed, we used nonparametric testing methods to examine the changing trends. The Mann–Kendall test was used to assess changing trends across the 6 observation points, with the parameters for statistical testing being the z value [32,33]. For categorical variables, the Cochran-Armitage test was used to examine their changing trends; the parameters for statistical testing were also the z value [34]. The generalized estimating equation was used to examine the association among confirmed new COVID-19 infected persons, behavioral beliefs, uncertainty stress, and suicidal ideation [35].

Ethical Considerations

The study was conducted following the Declaration of Helsinki and was approved by the ethics committee of the Medical Center at Jinan University (JNUKY-2022‐047). All participants received an electronic instruction on how to proceed. After

XSL•FO
RenderX

reading the instructions, they were asked to provide electronic consent by tapping the "Confirmation and Authorization" button and then were directed to the questionnaire. Participation in the study was entirely voluntary, with participants having the right to withdraw or discontinue at any time without any penalties or loss of benefits. All information was kept strictly confidential and anonymized to ensure participants could not be identified. As appropriate, a token of appreciation, 35 RMB (approximately US $5.00) were given to those participants who completed all 6 questionnaires.

Results

We recruited 229 participants at baseline. The baseline data were linkable to all subsequent questionnaire administrations, 4 intermediate waves, and a final observation point. A total of 221 (96.5%) remained for all repeated measures. The majority of respondents were Han Chinese (n=215, 97.3%) and 76.5% (169/221) were female. In terms of age, 13.6% (30/221) were younger than 20 years old, 21.7% (48/221) were aged 22 or older, and the remaining respondents were between 20 and 22 years old. Of the participants, 62.3% (131/221) were majoring in science, engineering, and medicine while 37.7% (79/221) were majoring in the humanities and social sciences (Table 1).

XSL•FO
RenderX

Table . Sample characteristics.

Group	Values, n (%)
Age (year)	
<20	30 (13.6)
20	54 (24.4)
21	50 (22.6)
22	39 (17.6)
22 and older	48 (21.7)
Ethnicity	
Han	215 (97.3)
Minority	6 (2.7)
Sex	
Male	52 (23.5)
Female	169 (76.5)
Monthly expenditure (RMB)[a]	
<1500	61 (27.6)
1500-2000	67 (30.3)
2000-2500	47 (21.3)
2500 and more	46 (20.8)
Major	
Science and engineering, and medicine	131 (62.3)
Humanities and social sciences	79 (37.7)
Father's education level	
Primary school and low	33 (14.9)
Junior school	63 (28.5)
High school	51 (23.1)
Junior college	33 (14.9)
College and more	41 (18.6)
Father's occupation	
Operation	143 (64.7)
Administration, commercial, and service	50 (22.6)
Science, technology, and education	12 (5.4)
Others	16 (7.2)
Mother's education level	
Primary school and low	56 (25.3)
Junior school	59 (26.7)
High school	47 (21.3)
Junior college	26 (11.7)
College and more	33 (14.9)
Mother's occupation	
Operation and others	132 (59.7)
Administration, commercial, and service	39 (17.6)
Science, technology, and education	26 (11.8)
Others	24 (10.9)

XSL•FO
RenderX

[a]1 RMB=US $0.14.

The prevalence of suicidal ideation was 16.7% (95% CI 12.5%-20.9%), 14.5% (95% CI 10.0%-19.3%), and 14.5% (95% CI 9.8%-19.2%) in the quarantine period and 13.8% (95% CI 9.5%-18.3%), 10.9% (95% CI 6.8%-15.0%), and 10.0% (95% CI 6.0%-14.1%) in the post-quarantine period. The Mann-Kendall test estimating suicidal ideation found significant changes (z=−4.06; P<.001). Simultaneously, perceived risk and perceived severity showed a statistically significant upward trend across the total observation period, with a z value in the former of 9.56 and 7.13 for the latter, both at P<.001, and this was also true for new infectious persons (z=3.69; P=.002), with the prevalence being 1.7%, 3.5%, 3.5%, 4.6%, 17.2%, and 42.0% at the time of each wave, respectively. However, uncertainty stress did not change significantly over the observation period (z=0.71; P=.48; see Table 2 and Figure 1).

Table . Time change trend in suicide ideation, perceived risk, perceived severity, uncertainty stress, and new infectious persons.

	Suicidal ideation prevalence, proportions (95% CI)	Perceived risk, mean (95% CI)	Perceived severity, mean (95% CI)	Uncertainty stress, mean (95% CI)	The number of new infectious patients, proportions (95% CI)
Wave					
1	16.7 (12.5-20.9)	2.95 (2.82-3.06)	2.74 (2.62-2.86)	11.41 (10.95-11.87)	1.70 (0.50-4.57)
2	14.5 (10.0-19.3)	2.97 (2.85-3.09)	2.76 (2.65-2.88)	11.64 (10.88-11.82)	3.50 (1.58-7.01)
3	14.5 (9.8-19.2)	3.32 (3.20-3.43)	2.58 (2.47-2.68)	11.64 (11.18-12.10)	3.50 (1.58-7.01)
4	13.8 (9.5-18.3)	3.37 (3.26-3.49)	2.57 (2.46-2.69)	11.46 (10.98-11.94)	4.60 (2.19-8.16)
5	10.9 (6.8-15.0)	3.30 (3.18-3.41)	2.79 (2.67-2.90)	11.38 (10.90-11.87)	17.20 (12.46-22.83)
6	10.0 (6.0-14.1)	3.12 (3.00-3.24)	2.98 (2.86-3.11)	11.43 (10.94-11.90)	42.00 (35.49-48.89)
Overall prevalence or mean	13.6 (11.9-25.3)	3.14 (3.00-3.19)	2.74 (2.86-3.11)	11.49 (10.98-11.92)	12.08 (6.03-19.17)
Time trend test (z value)	−4.06[a]	9.56[a]	7.13[a]	0.71	3.69[a]

[a]P<.01.

XSL•FO
RenderX

Figure 1. Time change trend in all variables.

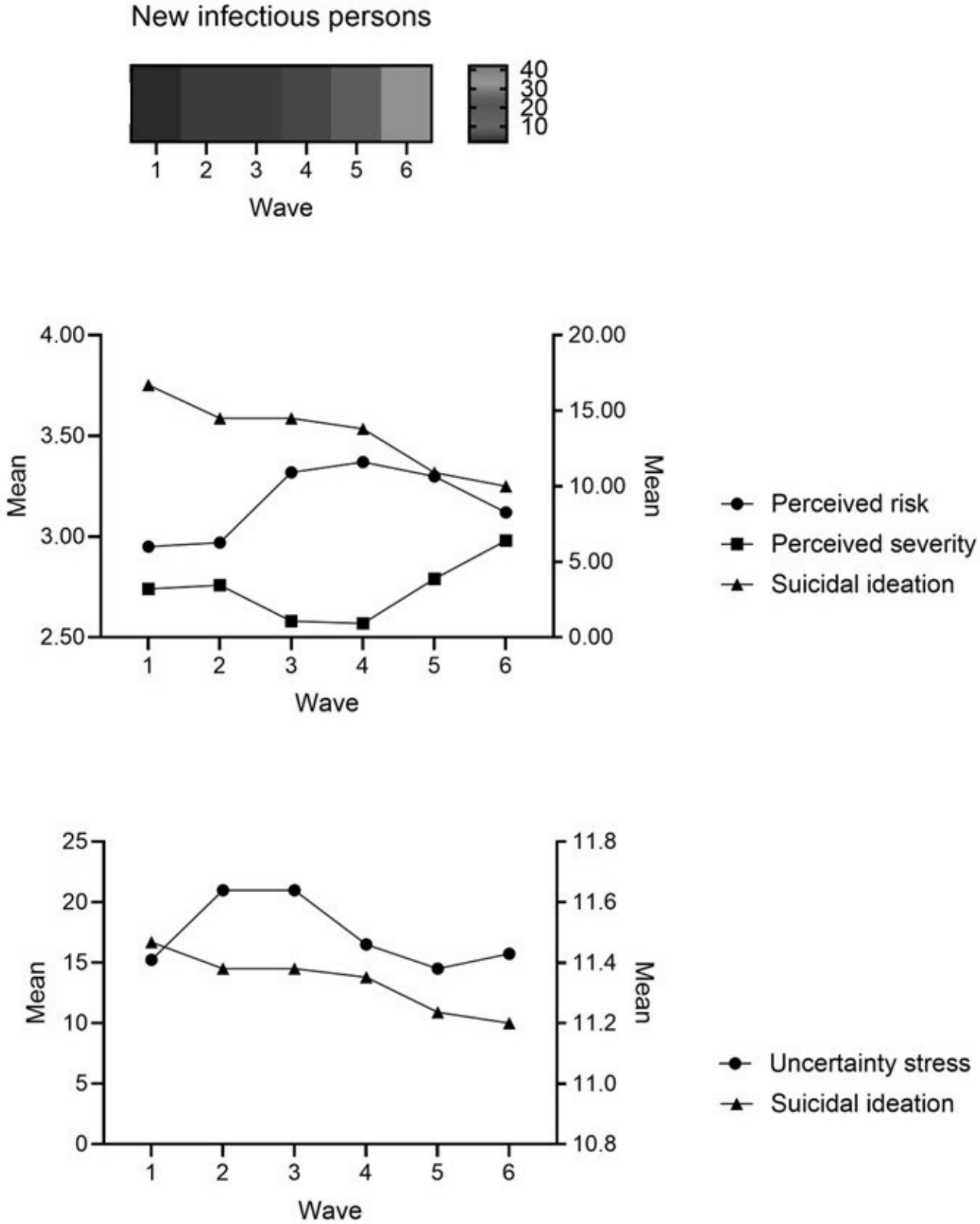

Table 3 shows the relationship among perceived risk, perceived severity, uncertainty stress, the number of new infectious patients, and suicidal ideation while accounting for repeat observations within individuals. It indicates that perceived risk and perceived severity were positively associated with suicidal ideation, with a β of 0.5482 (P<.001) and 0.0817 (P=.007) respectively. Higher uncertainty stress was associated with higher suicidal ideation (β=0.1776; P<.001). The number of new infectious patients did not have an association with suicidal ideation (β=0.0041; P=.49).

XSL•FO
RenderX

Table . Factors associated with suicidal ideation.

Group	β (SE)	*P* value
Perceived risk	0.5482 (0.0326)	<.001
Perceived severity	0.0817 (0.0303)	.007
Uncertainty stress	0.1776 (0.0207)	<.001
The number of new infectious patients	0.0041 (0.0059)	.49

Discussion

Contributions to the Literature

There is little evidence on changing trends in the impact of quarantine measures on suicidal ideation. In this study, the decline in suicidal ideation did not occur immediately after the lifting of quarantine measures; instead, a delayed effect or "grace period" was observed. Uncertainty stress was consistently high throughout the observation period and positively associated with suicidal ideation.

Principal Findings

This study aimed to fill a research gap by examining the changing trends of suicidal ideation and related factors during the transition from quarantine to post-quarantine periods in the COVID-19 surge in China. The prevalence of suicidal ideation was found to be 16.7%, 14.5%, and 14.5% during the quarantine period and 13.8%, 10.9%, and 10.0% during the post-quarantine period, showing a significant downward trend over the entire observation period. These findings indicate that quarantine has an impact on suicidal ideation. It is important to note that the decline in suicidal ideation did not occur immediately after the release of quarantine measures, but rather there was a grace period. Particularly in the early stages of release, the prevalence of suicidal ideation remained relatively high, with a larger decline observed in the later period. This suggests a time-lagged effect of lockdown on influencing suicidal ideation, a common phenomenon in public health policies. Understanding the duration of this lag can assist in predicting and addressing the consequences of implementing and lifting lockdown measures. These findings also highlight the necessity of expanding mental health services to individuals in quarantine.

Perceived risk and perceived severity showed a statistically significant upward trend across the total observation period. This means that the quarantine measure has decreased the role of perceived risk and perceived severity. This study adds new evidence on the influence of quarantine on the perceived risk of contracting COVID-19 and the perception of its severity. This phenomenon may be explained by SCM theory. Quarantine is a protective factor against stress stimuli, which can decrease high levels of perceived risk and perceived severity. When people are quarantined, they may think they are safe. However, when they are released, they may feel they have lost protection and perceive more risk and severity for COVID-19 infection.

Uncertainty stress is an important observational variable in this study. This study showed that uncertainty stress did not significantly change, remaining at a high level throughout the observation period. Extreme uncertainty is a key issue in the COVID-19 epidemic, which includes the uncertainty of COVID-19 itself and the suddenness of the lockdown and release policy; these factors may lead people to a high uncertainty stress level. Many studies found that uncertainty stress was positively associated with mental and behavioral problems during the COVID-19 period [18-21]. Due to uncertainty stress closely associated with social trust, it is possible that it influences the implementation of the policy during the COVID-19 pandemic [36-38]. We believe that sustainable high-level uncertainty stress poses a large challenge for people's compliance with COVID-19 control policies.

The Health Belief Model argues that individual behavior is determined by perceived risk and perceived severity [11,12]. The risk of disease and the severity of outcomes are crucial predictors of mental problems [11,12]. This study found that perceived risk and perceived severity of COVID-19 were positively associated with suicidal ideation. Many studies found there is a close association between strong risk and threat perceptions of COVID-19 and mental and behavioral outcomes [13-16]. At the same time, some studies also indicate that suicide is a prominent manifestation of mental problems [26-28]. Therefore, it is rational that perceived risk and perceived severity are related to suicidal ideation.

This study found that uncertainty stress was positively associated with suicidal ideation. A host of evidence supports the assertion that uncertainty constitutes a powerful stressor [12,39,40], significantly associated with many mental and behavioral problems related to COVID-19 [18,21]. An outstanding characteristic of the COVID-19 epidemic is uncertainty. It is a new virus, and much of the information about it is unknown or incomplete, with considerable ambiguity surrounding the disease, and there are frequent changes in pathogens. It should be emphasized that this characteristic has persisted so far. It should be emphasized that uncertainty is also closely related to the government's way of disease control, such as ambiguity in providing information, the suddenness of implementing lockdown and opening up, inconsistencies in policies issued, low professionalism and authority of experts invited, and other factors. A basic strategy to decrease uncertainty is to implement scientific decisions. All policies and interventions need to be based on scientific evidence when dealing with the COVID-19 epidemic. However, in some areas or regions, policies and interventions were not based on scientific evidence. Many of these policies and interventions were arbitrary and ineffective. Ineffective policies and interventions can cause people to question their government, which ultimately may be detrimental to disease control.

Suicidal ideation decreased over the study period, despite upward trends in perceived risk, perceived severity, and the number of new patients. However, perceived risk was positively

XSL•FO
RenderX

associated with suicidal ideation, and higher uncertainty stress contributed to higher suicidal ideation. This phenomenon may be difficult for some people to understand. It should be noted that trend analysis provides a description of the changes in variables during the observation period, whereas correlation analysis of the observed research variables examines the interactions between multiple factors, considering individual differences and temporal changes. Therefore, the results from trend analysis and correlation analysis should not be expected to align perfectly. This study also performed subgroup analysis based on different sample characteristics, which yielded results consistent with the previous ones. This phenomenon is observed in many other studies [41-43].

This study found a significant increase in the number of new infectious individuals throughout the observation period, which aligns with real-world observations. When the lockdown was lifted, people regained access to the outside world, increasing opportunities for infection, which led to a rise in the number of patients [8]. While quarantine measures were linked to suicidal ideation, the increase in new infections did not show a similar association with suicidal ideation. This highlights the importance of carefully considering the mental health impacts of quarantine when designing lockdown policies.

Study Limitations

There are several limitations to this study. First, our sample size appears to be small. Nevertheless, this is a prospective longitudinal panel study, and the variables included were repeatedly measured for each participant, which ensured statistical power for the tests used in this study was high. We estimated the power for repeated measures for each variable [44]. This analysis calculated statistical power (1-β) at given sample size (N=221), 5% significance level, and effect size used as a parameter in the model. A statistical power of 0.8 and higher was considered acceptable [45]. Since the effect size of various variables may differ, their statistical powers are also different. The results showed that statistical powers for all variables achieved accepted level (1-β=0.8 and higher). This indicates that the sample size in this study was large enough to make appropriate inferences. Second, sample attrition may introduce "cluster" bias because many longitudinal studies likely overrepresent some of these characteristics. A more sophisticated design and more representative sample would be necessary to resolve this problem. Another important limitation is that our participants were restricted to university students, who were selected due to their good cooperation in follow-up for a panel study. Thus, our results cannot be generalized to the wider Chinese population.

Conclusion

This study found that suicidal ideation among Chinese university students declined over time as quarantine measures were lifted, but the decrease was delayed, indicating a lagged mental health effect of the lockdown. Perceived risk and perceived severity of COVID-19 increased, suggesting that sense of safety during lockdown may have reduced risk perception, which then rebounds once restrictions are lifted. Uncertainty stress remained consistently high throughout, highlighting a persistent mental health risk burden during the COVID-19 surge. Moreover, higher perceived risk, higher perceived severity, and higher uncertainty stress were related to a higher likelihood of reporting suicidal ideation. However, the proportion of infected individuals was not linked to suicidal ideation. University students are more sensitive to perceptions influenced by the COVID-19 context rather than direct medical facts. This may primarily drive mental health problems. This study provides new information for the governments and policy makers to design effective intervention strategies targeted at reducing the COVID-19 and other infectious diseases.

Acknowledgments

We thank the contribution of Jinan University and Zhejiang University work teams. This study did not use generative artificial intelligence tools or other language models to design the proposal or assist with the manuscript. During the review process, we used ChatGPT to check the grammar. This study is partially funded by Guangdong Basic and Applied Basic Research Foundation (2023A1515110144) and Science and Technology Projects in Guangzhou (2023A04J1923).

Data Availability

The datasets used and analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

LL contributed to the resources, literature search, and primary writing. TY was involved in the conceptualizationn, methodology, resources, formal analysis and writing, and validation. SP handled the conceptualization, investigation, data curation, resources, and project administration. RRC performed the validation and review and editing.

Conflicts of Interest

None declared.

References

1. Atalan A. Is the lockdown important to prevent the COVID-19 pandemic? Effects on psychology, environment and economy-perspective. Ann Med Surg (Lond) 2020 Aug;56:38-42. [doi: 10.1016/j.amsu.2020.06.010] [Medline: 32562476]

XSL•FO
RenderX

2. Grover S, Sahoo S, Mehra A, et al. Psychological impact of COVID-19 lockdown: an online survey from India. Indian J Psychiatry 2020;62(4):354. [doi: 10.4103/psychiatry.IndianJPsychiatry_427_20]
3. Mohler-Kuo M, Dzemaili S, Foster S, Werlen L, Walitza S. Stress and mental health among children/adolescents, their parents, and young adults during the first COVID-19 lockdown in Switzerland. Int J Environ Res Public Health 2021 Apr 27;18(9):4668. [doi: 10.3390/ijerph18094668] [Medline: 33925743]
4. Uzobo E, Ayinmoro AD. Trapped between two pandemics: domestic violence cases under COVID-19 pandemic lockdown: a scoping review. Community Health Equity Res Policy 2023 Apr;43(3):319-328. [doi: 10.1177/0272684X211022121] [Medline: 34102910]
5. Gan Y, Ma J, Wu J, Chen Y, Zhu H, Hall BJ. Immediate and delayed psychological effects of province-wide lockdown and personal quarantine during the COVID-19 outbreak in China. Psychol Med 2022 May;52(7):1321-1332. [doi: 10.1017/S0033291720003116] [Medline: 32787981]
6. Tian H, Liu Y, Li Y, et al. An investigation of transmission control measures during the first 50 days of the COVID-19 epidemic in China. Science 2020 May 8;368(6491):638-642. [doi: 10.1126/science.abb6105] [Medline: 32234804]
7. Reynolds DL, Garay JR, Deamond SL, Moran MK, Gold W, Styra R. Understanding, compliance and psychological impact of the SARS quarantine experience. Epidemiol Infect 2008 Jul;136(7):997-1007. [doi: 10.1017/S0950268807009156]
8. Brooks SK, Webster RK, Smith LE, et al. The psychological impact of quarantine and how to reduce it: rapid review of the evidence. Lancet 2020 Mar 14;395(10227):912-920. [doi: 10.1016/S0140-6736(20)30460-8] [Medline: 32112714]
9. Charles ST, Piazza JR, Mogle J, Sliwinski MJ, Almeida DM. The wear and tear of daily stressors on mental health. Psychol Sci 2013 May;24(5):733-741. [doi: 10.1177/0956797612462222] [Medline: 23531486]
10. Parrish BP, Cohen LH, Laurenceau JP. Prospective relationship between negative affective reactivity to daily stress and depressive symptoms. J Soc Clin Psychol 2011 Mar;30(3):270-296. [doi: 10.1521/jscp.2011.30.3.270]
11. Cottrell R, McKenzie JF. Health Promotion & Education Research Methods: Using the Five Chapter Thesis/Dissertation Model: Jones & Bartlett Publishers; 2010.
12. Yang T. Health Research: Social and Behavioral Theory and Methods: People's Medical Publishing House; 2018.
13. Yang XY, Peng S, Yang T, Cottrell RR. Changing trends of mental and behavioral responses and associations during the COVID-19 epidemic in China: a panel study. Health Educ Res 2021 Apr 12;36(2):151-158. [doi: 10.1093/her/cyab012] [Medline: 33724361]
14. Mo PKH, Fong VWI, Song B, Di J, Wang Q, Wang L. Association of perceived threat, negative emotions, and self-efficacy with mental health and personal protective behavior among Chinese pregnant women during the COVID-19 pandemic: cross-sectional survey study. J Med Internet Res 2021 Apr 12;23(4):e24053. [doi: 10.2196/24053] [Medline: 33729983]
15. Li JB, Yang A, Dou K, Wang LX, Zhang MC, Lin XQ. Chinese public's knowledge, perceived severity, and perceived controllability of COVID-19 and their associations with emotional and behavioural reactions, social participation, and precautionary behaviour: a national survey. BMC Public Health 2020 Oct 21;20(1):1589. [doi: 10.1186/s12889-020-09695-1] [Medline: 33087109]
16. Li X, Lyu H. Epidemic risk perception, perceived stress, and mental health during COVID-19 pandemic: a moderated mediating model. Front Psychol 2021;11. [doi: 10.3389/fpsyg.2020.563741]
17. Festinger L. A Theory of Cognitive Dissonance: Stanford University Press; 1962, Vol. 2.
18. Peng S, Yang XY, Yang T, Zhang W, Cottrell RR. Uncertainty stress, and its impact on disease fear and prevention behavior during the COVID-19 epidemic in China: a panel study. Am J Health Behav 2021 Mar 1;45(2):334-341. [doi: 10.5993/AJHB.45.2.12] [Medline: 33888193]
19. Minervini G, Franco R, Marrapodi MM, et al. The association between COVID-19 related anxiety, stress, depression, temporomandibular disorders, and headaches from childhood to adulthood: a systematic review. Brain Sci 2023 Mar 12;13(3):481. [doi: 10.3390/brainsci13030481] [Medline: 36979291]
20. Seco Ferreira DC, Lisboa Oliveira W, Costa Delabrida ZN, et al. Intolerance of uncertainty and mental health in Brazil during the Covid-19 pandemic. Sumapsi 2020 Aug 20;27(1):62-69. [doi: 10.14349/sumapsi.2020.v27.n1.8]
21. Wu D, Rockett IRH, Yang T, Yang XY, Wang M, Jiao C. Perceived beliefs, uncertainty, and behavioral responses during the COVID-19 outbreak in China: findings from a convenience sample. Am J Health Promot 2021 Sep;35(7):977-983. [doi: 10.1177/08901171211004249] [Medline: 33787350]
22. Podder I, Agarwal K, Datta S. Comparative analysis of perceived stress in dermatologists and other physicians during national lock-down and COVID-19 pandemic with exploration of possible risk factors: a web-based cross-sectional study from Eastern India. Dermatol Ther 2020 Jul;33(4):e13788. [doi: 10.1111/dth.13788] [Medline: 32506785]
23. Zhou Y, Chen Z, Li W, Chen S, Xu H, Zhou Z. Impacting factors and sources of perceived stress by home-quarantined residents in Shanghai during COVID-19 epidemic. BMC Public Health 2023;23(1):780. [doi: 10.1186/s12889-023-15701-z]
24. Vandoros S, Kawachi I. Economic uncertainty and suicide in the United States. Eur J Epidemiol 2021 Jun;36(6):641-647. [doi: 10.1007/s10654-021-00770-4] [Medline: 34114185]
25. Wu D, Yang T, Rockett IR, Yu L, Peng S, Jiang S. Uncertainty stress, social capital, and suicidal ideation among Chinese medical students: findings from a 22-university survey. J Health Psychol 2021 Feb;26(2):214-225. [doi: 10.1177/1359105318805820] [Medline: 30318917]

XSL•FO
RenderX

26. Nock MK, Borges G, Bromet EJ, et al. Cross-national prevalence and risk factors for suicidal ideation, plans and attempts. Br J Psychiatry 2008 Feb;192(2):98-105. [doi: 10.1192/bjp.bp.107.040113] [Medline: 18245022]
27. Peng S, Yang XY, Rockett IRH. A typology of social capital and its mixed blessing for suicidal ideation: a multilevel study of college students. Soc Sci Med 2019 Dec;243:112640. [doi: 10.1016/j.socscimed.2019.112640] [Medline: 31665656]
28. Rockett IRH. Injury and violence: a public health perspective. Inj Prev 1999 Dec;5(4):311. [doi: 10.1136/ip.5.4.311]
29. Klonsky ED, May AM, Saffer BY. Suicide, suicide attempts, and suicidal ideation. Annu Rev Clin Psychol 2016;12:307-330. [doi: 10.1146/annurev-clinpsy-021815-093204] [Medline: 26772209]
30. Yang T, Barnett R, Fan Y, Li L. The effect of urban green space on uncertainty stress and life stress: a nationwide study of university students in China. Health Place 2019 Sep;59:102199. [doi: 10.1016/j.healthplace.2019.102199] [Medline: 31514059]
31. Wu D, Yu L, Yang T, et al. The impacts of uncertainty stress on mental disorders of Chinese college students: evidence from a nationwide study. Front Psychol 2020;11:243. [doi: 10.3389/fpsyg.2020.00243] [Medline: 32210868]
32. Hesamian G, Akbari MG, Shams M. A goodness-of-fit test based on fuzzy random variables. Fuzzy Inf Eng 2023 Mar;15(1):55-68. [doi: 10.26599/FIE.2023.9270005]
33. Kendall MG. Rank correlation methods. In: Public Program Analysis: Springer; 1948:146-163.
34. Rezaei B, Shahedi A, Keshtkaran Z, et al. Changes in incidence and mortality trend due to COVID-19 in southern Iran, from 2020 to 2021: based on Cochran–Armitage trend test. J Public Health (Berl) 2024 Jul;32(7):1153-1159. [doi: 10.1007/s10389-023-01902-8]
35. Zorn CJW. Generalized estimating equation models for correlated data: a review with applications. Am J Pol Sci 2001 Apr;45(2):470. [doi: 10.2307/2669353]
36. Cullen W, Gulati G, Kelly BD. Mental health in the COVID-19 pandemic. QJM 2020 May 1;113(5):311-312. [doi: 10.1093/qjmed/hcaa110]
37. Jiang N, Wu AM, Cheng EW. Social trust and stress symptoms among older adults during the COVID-19 pandemic: evidence from Asia. BMC Geriatr 2022 Apr 15;22(1):330. [doi: 10.1186/s12877-022-02847-5] [Medline: 35428191]
38. van Tilburg TG, Steinmetz S, Stolte E, van der Roest H, de Vries DH. Loneliness and mental health during the COVID-19 pandemic: a study among Dutch older adults. J Gerontol 2021 Aug 13;76(7):e249-e255. [doi: 10.1093/geronb/gbaa111]
39. Dar KA, Iqbal N, Mushtaq A. Intolerance of uncertainty, depression, and anxiety: examining the indirect and moderating effects of worry. Asian J Psychiatr 2017 Oct;29:129-133. [doi: 10.1016/j.ajp.2017.04.017] [Medline: 29061409]
40. Greco V, Roger D. Uncertainty, stress, and health. Pers Individ Dif 2003 Apr;34(6):1057-1068. [doi: 10.1016/S0191-8869(02)00091-0]
41. Rebar AL, Stanton R, Geard D, Short C, Duncan MJ, Vandelanotte C. A meta-meta-analysis of the effect of physical activity on depression and anxiety in non-clinical adult populations. Health Psychol Rev 2015;9(3):366-378. [doi: 10.1080/17437199.2015.1022901] [Medline: 25739893]
42. Primack BA, Shensa A, Sidani JE, et al. Social media use and perceived social isolation among young adults in the US. Am J Prev Med 2017 Jul;53(1):1-8. [doi: 10.1016/j.amepre.2017.01.010] [Medline: 28279545]
43. Chou HTG, Edge N. “They are happier and having better lives than I am”: the impact of using Facebook on perceptions of others’ lives. Cyberpsychol Behav Soc Netw 2012 Feb;15(2):117-121. [doi: 10.1089/cyber.2011.0324] [Medline: 22165917]
44. Rabe-Hesketh S, Everitt B. Handbook of Statistical Analyses Using Stata: CRC Press; 2003.
45. Peterman RM. Statistical power analysis can improve fisheries research and management. Can J Fish Aquat Sci 1990 Jan 1;47(1):2-15. [doi: 10.1139/f90-001]

Abbreviations

SCM: Stimulus-Cognition-Mental

Edited by A Mavragani; submitted 23.03.25; peer-reviewed by D Wu, MA Kato; revised version received 10.07.25; accepted 11.07.25; published 05.09.25.

Please cite as:
Li L, Yang T, Peng S, Cottrell RR
Changing Trends in Suicidal Ideation and Its Influencing Factors During the Transition From Quarantine to Post-Quarantine Among Chinese University Students During the COVID-19 Surge: Six-Wave Panel Study
JMIR Form Res 2025;9:e74370
URL: https://formative.jmir.org/2025/1/e74370
doi:10.2196/74370

XSL•FO
RenderX

© Lijing Li, Tingzhong Yang, Sihui Peng, Randall R Cottrell. Originally published in JMIR Formative Research (https://formative.jmir.org), 5.9.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Evaluation of Generative Artificial Intelligence Implementation Impacts in Social and Health Care Language Translation: Mixed Methods Case Study

Miia Martikainen[1*], MSc; Kari Smolander[1*], MSc, LicSc, PhD; Johan Sanmark[2*], MD; Enni Sanmark[3*], MD, PhD

[1]Department of Software Engineering, School of Engineering, Lappeenranta-Lahti University of Technology, Yliopistonkatu 34, Lappeenranta, Finland

[2]Department of Research, Development and Innovation, Western Uusimaa Wellbeing Services County, Espoo, Finland

[3]Department of Medicine, Faculty of Medicine, HUS Helsinki University Hospital and University of Helsinki, Helsinki, Finland

[*]all authors contributed equally

Corresponding Author:
Miia Martikainen, MSc
Department of Software Engineering, School of Engineering, Lappeenranta-Lahti University of Technology, Yliopistonkatu 34, Lappeenranta, Finland

Abstract

Background: Generative artificial intelligence (GAI) is expected to enhance the productivity of the public social and health care sector while maintaining, at minimum, current standards of quality and user experience. However, empirical evidence on GAI impacts in practical, real-life settings remains limited.

Objective: This study investigates productivity, machine translation quality, and user experience impacts of the GPT-4 language model in an in-house language translation services team of a large well-being services county in Finland.

Methods: A mixed methods study was conducted with 4 in-house translators between March and June 2024. Quantitative data of 908 translation segments were collected in real-life conditions using the computer-assisted language translation software Trados (RWS) to assess productivity differences between machine and human translation. Quality was measured using 4 automatic metrics (human-targeted translation edit rate, Bilingual Evaluation Understudy, Metric for Evaluation of Translation With Explicit Ordering, and Character n-gram F-score) applied to 1373 GAI-human segment pairs. User experience was investigated through 5 semistructured interviews, including the team supervisor.

Results: The findings indicate that, on average, postediting machine translation is 14% faster than translating texts from scratch (2.75 vs 2.40 characters per second, P=.03), and up to 37% faster when the number of segments is equalized across translators. However, productivity varied notably between individuals, with improvements ranging from −2% to 102%. Regarding translation quality, 11% (141/1261) of Finnish-Swedish and 16% (18/112) of Finnish-English GAI outputs were accepted without edits. Average human-targeted translation edit rate scores were 55 (Swedish) and 46 (English), indicating that approximately half of the words required editing. Bilingual Evaluation Understudy scores averaged 43 for Swedish and 38 for English, suggesting good translation quality. Metric for Evaluation of Translation With Explicit Ordering and Character n-gram F-scores reached 63 and 68 for Swedish and 59 and 57 for English, respectively. All metrics have been converted to an equivalent scale from 0 to 100, with 100 reflecting a perfect match. Interviewed translators expressed mixed reviews on productivity gains but generally perceived value in using GAI, especially for repetitive, generic content. Identified challenges included inconsistent or incorrect terminology, lack of document-level context, and limited system customization.

Conclusions: Based on this case study, GPT-4–based GAI shows measurable potential to enhance translation productivity and quality within an in-house translation team in the public social and health care sector. However, its effectiveness appears to be influenced by factors, such as translator postediting skills, workflow design, and organizational readiness. These findings suggest that, in similar contexts, public social and health care organizations could benefit from investing in translator training, optimizing technical integration, redesigning workflows, and implementing effective change management. Future research should examine larger translator teams to assess the generalizability of these results and further explore how translation quality and user experience can be improved through domain-specific customization.

(JMIR Form Res 2025;9:e73658) doi:10.2196/73658

KEYWORDS

generative artificial intelligence; large language model; ChatGPT; pretrained language model; language translation; machine translation evaluation; public social and health care

XSL•FO
RenderX

Introduction

Generative artificial intelligence (GAI) is being explored as one prospective innovation to address the pressing economic sustainability challenges of the Finnish public social and health care sector. GAI refers to digital technologies capable of producing original content, such as text, images, audio, and video by learning semantic patterns from vast amounts of pretraining data [1,2]. In the context of this study, GAI specifically refers to a large language model (LLM) engine called GPT-4 developed by a research organization called OpenAI.

In 2024, the Finnish Ministry of Social Affairs and Health established an artificial intelligence (AI) ecosystem for social and health care services to enhance care quality, service delivery productivity, and economic growth through AI-driven innovation [3]. Despite the growing strategic focus on GAI, operational AI maturity of the public sector remains at an early development stage [4,5]. Hjaltalin and Sigurdarson [6] consolidated 28 governmental AI strategies, including Finland's, into three national AI discourses: (1) empowerment through information, (2) enhanced administrative practices, and (3) improved service delivery. Implementing GAI in administrative practices presents a significantly lower barrier to entry due to lower compliance risk, as secondary use of social and health care data and application of GAI in service delivery are subject to stringent regulation. For instance, the European Union regulatory frameworks of the AI Act and the General Data Protection Regulation (GDPR) have a significant impact on the usage of GAI in public administration.

In Finland, 21 well-being services counties, alongside 3 other governance entities (HUS Group, the joint authority for Helsinki and Uusimaa), are responsible for public social, health, and rescue services [7]. Approximately 30% of these entities are legally required to provide services in both national languages, such as Finnish and Swedish [8]. The case organization of this study is Western Uusimaa, the third largest well-being services county in Finland. Western Uusimaa Wellbeing Services County (Länsi-Uudenmaan hyvinvointialue, LUVN) is not only a bilingual region but also a region where population growth is projected to be driven primarily by an increase in residents whose native language is neither Finnish nor Swedish [9]. A similar trend can be observed globally due to globalization, while the legislation of many countries still requires services to be provided in the official languages of the country. Translation capability is crucial in health care settings, because language barriers can lead to issues in patient safety and outcomes [10]. Consequently, language translation services are becoming an increasingly critical support function across the health care sector.

Machines have been used to support language translation from as early as the 1950s [11], but the more sophisticated, machine learning–based technologies have been available from approximately 2015 onwards [12]. Language translation is among the most widely recognized applications of LLMs in the health care sector [13]. Despite considerable technological progress in recent years, the quality of machine translation (MT) has not yet reached the level of experienced human translators [14,15]. Thus, postediting is an essential practice of modern translation work [16].

Earlier reviews, such as Wang et al [16], examined studies conducted between 2007 and 2011 and reported conflicting results regarding productivity gains associated with prior-generation MT. Several studies reported improvement in translator productivity, while others have found no significant increase or even a slight speed reduction. A study by Macken et al [14] found a 14% speed improvement with neural MT for the English-Finnish language pair and a 12% speed increase with statistical MT for English-French in real-world settings. A more recent study by Merali [17] evaluates the productivity effect of LLMs. The study found that a 10-fold increase in model training compute led to significant improvement in task completion speed (12.3% relative to baseline and grades by 0.18 SDs), translation quality, and translator earnings. Notably, the productivity benefits were reported higher among less experienced translators.

GAI is fundamentally transforming language translation. However, there is currently limited evidence on real-life impacts of GAI-based MT, particularly regarding specialized sectors, such as social and health care. This study aims to investigate the productivity impact, translation quality, and user experience of a GAI-based plugin for computer-assisted translation (CAT) software in social and health care services.

Methods

Overview

Language translation services of the case organization LUVN are coordinated and produced by an organizationally centralized in-house team of 4 translators. Translators primarily translate from Finnish to Swedish. A minor portion of translation tasks involves the language pairs, which are Swedish-Finnish and Finnish-English. Translators have translation experience ranging from 1 to 13 years and industry-specific experience of 0 - 2 years in social and health care sectors. Translators' educational backgrounds focus on literature, translation, and social sciences, but every one of them does have prior experience with MT aids.

The in-house translation department did not have a standardized practice of using MT aid prior to the GAI-based plugin. The GAI-based plugin was used independently and was not combined with any other MT plugins. The study was conducted during the early adoption stage, and no term bases were used for the GAI-based plugin. In addition, only the system prompt was used. The translators neither have visibility to the defined system prompt nor do they use any task-specific prompts.

Mixed methods were chosen to capture both objective and experiential aspects of GAI adoption. The quantitative and qualitative components were collected sequentially but analyzed in parallel to support triangulation of findings. Data collection was conducted from March until June 2024 and measured 3 dimensions, such as productivity, quality, and user experience. Translators processed usual translation materials, excluding any materials with sensitive or personal data.

XSL•FO
RenderX

Productivity

The productivity of the language translation was assessed as translation speed according to the earlier studies [18]. The data were extracted from the files in the SDL XML-based localization interchange file format in that Trados software (Trados Studio 2022, version SR1 - 17.1.6.16252, RWS Group), produced for each translation project. Translators were instructed to translate each text so that half of the segments are translated from scratch and the other half using GAI-based MT (AI Professional Plugin for Trados, RWS Group). The plugin is using the Microsoft Azure, OpenAI GPT-4 language model service. To reduce bias, translators were advised to randomly select each time whether they start with or without GAI. Translators retrieved GAI-based MT at a segment level only for segments without translation memory (TM) matches. Translators collected 66 documents in total, including 2946 segments. The data were filtered with four criteria:

1. Comparison between translation from scratch and postediting: Segments from TM (708), from other automated translations, such as auto-propagation and direct copies from source segments (257), as well as segments that were postedited based on the aforementioned other automated translation outputs (332) were excluded.
2. Translation time calculation: Trados does not record translation time directly, but 2 time stamps "Creation_date_and_time" and "Modified_on" were used to estimate translation time by calculating the time difference between the time stamps of consecutive segments. Only positive translation times were accepted. A negative time stamp is primarily caused by a translator revisiting the segment at a later stage, resulting in nonconsecutive segments.
3. Translation time maximum threshold: Segments with a translation time of more than 30 seconds per word were excluded, as this suggests the translator was not actively engaged in the task. This assumption is aligned with prior research [19].
4. Translation time minimum threshold: Segments with translation times under 0.5 seconds per word were excluded, as such results likely indicate typographical errors or software malfunctions, making the data unreliable. This assumption is aligned with prior research [19].

The final dataset for comparative analysis consisted of 908 segments. Of these, 47% (n=427) of segments were postedited while 53% (n=481) of segments were translated from scratch (Table 1). A total of 62% (427/692) of the GAI postedited segments and 50% (481/957) of the segments translated from scratch were incorporated into the analysis. The segments included in the analysis contained a higher average number of words and characters compared to those excluded.

Table . Summary of language translation segment data from an in-house translation team of a Finnish well-being services county.

Measure	Segments excluded from the analysis				Segments included in the analysis		
	Postediting GAI[a] output[b]	Human translation[b]	Other[c]	Total	Postediting GAI output	Human translation	Total
Number of segments	265	476	1297	2038	427	481	908
Number of words in a segment, mean (SD)	7.8 (5.9)	7.9 (5.5)	—[d]	—	9.7 (5.4)	8.9 (5.2)	9.3 (5.3)

[a] GAI: generative artificial intelligence.

[b]Segments excluded based on filtering criteria 2 - 4.

[c]Segments excluded based on filtering criterion 1.

[d]Not applicable.

Quality

Target quality is defined as quality accepted by a human translator [14,20,21]. The translation process began with the pretranslation of the entire text using GAI, incorporating TM. The file in the SDL XML-based localization interchange file format containing the machine-generated translations was saved, after which the translators performed postedits and saved the modified file under a different name. Differences between the original GAI output and the final human-reviewed translation were then compared. Similar to the collection of productivity data, quality data were collected during the translators' regular workflow.

Although data collection was initiated with 4 translators, only 3 provided datasets that could be included in the quality analysis. For 1 participant, no analyzable data were recorded due to a technical issue during the data capture process. Despite prior instructions and support, the system did not retain segment-level data for this individual, resulting in an empty export file. Consequently, their data could not be used in the final analysis.

The final dataset included 47 documents and 1766 segment pairs. TM was used for 20% (353/1766) of segments, with most matches being exact (254/353). To refine the dataset, TM segments and other automatically generated segments were excluded, resulting in a final dataset of 1373 segment pairs for analysis. Each segment pair consisted of a raw GAI output and its postedited version corresponding to a single source language segment.

The quality of MT was estimated using 4 established and widely adopted automatic metrics: human-targeted translation edit rate (HTER)–score, Bilingual Evaluation Understudy (BLEU) score, Metric for Evaluation of Translation With Explicit Ordering (METEOR)–score, and character-level metric Character n-gram F-score (ChrF). HTER aligns well with real-world use cases, as it evaluates the effort required to postedit MT output into an acceptable human translation, but it only calculates the minimum

XSL•FO
RenderX

amount of effort to finalize translation [14]. More sophisticated automatic measures BLEU and METEOR are both widely used methods for evaluating MT quality [22]. A BLEU score above 40 indicates high-quality translations while scores between 30 and 40 reflect translations that range from understandable to good [23]. Finnish is a morphologically rich language, which means it has a complex system of word formation and inflections. ChrF performs particularly well for such languages [24].

User Experience

A total of five 60-minute videoconference interviews were conducted in May 2024. The sessions were recorded and automatically transcribed for analysis. The aim was to explore the impact of GAI implementation on translation services, focusing on the perceived productivity, MT quality, and user experience. The interview questions are listed in the Multimedia Appendix 1.

Ethical Considerations

This study was reviewed and approved by Western Uusimaa Wellbeing Services County on March 6, 2024 (approval number 636/13.01.00/2024). All participants were informed about the purpose of the study and provided written informed consent prior to participation. Recruitment of participants was facilitated through managerial channels. All 4 in-house translators were invited to participate, and participation was voluntary. After initial coordination with the management, the researcher maintained regular contact with the in-house translators to (1) provide study information, (2) support data collection, and (3) address any questions related to the research process. The consent process included permission for collection, analysis, and publication of data in anonymized form. All analyzed and reported data were processed to ensure anonymization. No identifiable information is included in the published results or supplementary materials. Participants were not provided with monetary or material compensation, as their involvement in the study took place during regular work hours as part of their normal professional duties.

Results

Productivity

Human translation speed was 2.40 characters per second (total of 481 segments) and postediting a GAI output speed was 2.75 characters per second (total of 427 segments). Hence, the postediting GAI-generated segment was, on average, 14% faster than translating from scratch (characters per second). Due to the nonnormal distribution of the dataset, the Mann-Whitney *U* test was applied, indicating a statistically significant difference (P=.03).

Significant variation occurs among translators in terms of productivity differences and the volume of segments generated (Table 2). Translator A produced 61% (555/908) of the total segments, with a minor decrease in productivity when using GAI. In contrast, the productivity of the other 3 translators increased by 18%, 28%, and 102%, respectively, when using GAI. To provide additional context to the observed variation in productivity, a supplementary analysis was conducted in which the translated segments were evenly distributed across translators. While this approach is not a standard practice, the analysis is included as a complementary perspective to illustrate how differences in workload distribution may influence the overall productivity outcomes. When calculating productivity differences by equally distributing segments across translators, postediting GAI-generated output is, on average, 37% faster than translating from scratch.

Table . Comparative analysis of translation speed in a case study of an in-house translation team within a Finnish well-being services county. Segments are analyzed by translator.

Measure	Translator			
	A	B	C	D
Segments, n (%)	555 (61)	207 (23)	102 (11)	44 (5)
Average translation speed (characters per second)				
Human translation	2.4	1.7	3.5	2.2
Postediting GAI[a] output	2.3	2.1	7.2	2.6
Difference (%)	–2	28	102	18

[a]GAI: generative artificial intelligence.

Quality

In the context of this study, maximum quality is defined as MT output that requires no postediting by a human translator. Notably, 141/1261 (11%) Finnish-Swedish segment pairs were accepted by translators without edits, whereas 18/112 (16%) Finnish-English segment pairs required no human intervention. All other target segments underwent postediting by human translators. The quality of MT was evaluated using 4 automatic metrics (Table 3).

Based on the average BLEU scores, the MT output for English achieved a good quality level while the output for Swedish was of high quality. However, HTER scores indicate room for quality improvement, and scores of 55 for Swedish and 46 for English suggest that approximately half of the words required postediting. It is important to note, however, that “edit” does not necessarily imply that half of the words were incorrect; edits may involve insertions, deletions, substitutions, or shifts.

XSL•FO
RenderX

Table . Automatic quality metrics for GAI-based[a] machine translation output in a case study of an in-house translation team in a Finnish wellbeing services county.

Measure	Target segment language	
	Swedish	English
Total number of segments, n	1261	112
GAI output accepted without edits		
Segments, n (%)	141 (11)	18 (16)
Machine translation output quality metric (average), mean[b,c]		
HTER[d,e]	55	46
BLEU[f]	43	38
METEOR[gh]	63	59
ChrF[i]	68	57

[a]GAI: generative artificial intelligence.

[b]Scale 0-100, where 100=perfect match.

[c]The reported HTER, BLEU, METEOR, and ChrF values are aggregate machine translation evaluation metrics calculated over the entire dataset. Therefore, a standard deviation is not applicable in this context.

[d]HTER: human-targeted translation edit rate.

[e]Multiplied by 100, inverted and negative values changed to 0 to convert the metric to equivalent scale from 0 to 100.

[f]BLEU: Bilingual Evaluation Understudy.

[g]METEOR: Metric for Evaluation of Translation With Explicit Ordering.

[h]Multiplied by 100 to convert the metric to equivalent scale from 0 to 100.

[i]ChrF: Character n-gram F-score.

User Experience

Interviews revealed diverse user experiences on the impact of GAI. Qualitative findings supported the contextualization of the variability observed in quantitative productivity measures. One translator assessed that GAI accelerates their overall translation productivity, while another translator reported that it actually slowed down their work due to quality deficiencies and workflow inefficiencies. Two translators perceived the overall impact as neutral, with productivity varying based on text type, complexity, and terminology. Overall, 3 out of 4 translators expressed a desire to continue using a similar GAI tool in the future.

The translators acknowledged the potential of GAI to improve translation quality but noted its current limitations, including inconsistent outputs and challenges with industry-specific terminology, gender, and concord issues. In addition, variations in translation quality, even with identical prompts, and a lack of document-level context were identified as significant challenges. Despite these shortcomings, translators found that GAI is effective in improving clarity of vaguely written source texts and for longer, repetitive texts with generic terminology.

Suggestions for improving GAI included refining workflows within CAT tools, improving terminology management, and using context-specific fine-tuning. Translators expressed cautious optimism about GAI's potential to enhance both productivity and translation quality but emphasized the need for tailored development and ongoing support to address its limitations.

Discussion

Principal Findings

In this study, the impacts of implementing a GAI-based MT solution on productivity, quality, and user experience within the in-house language translation services of a large Finnish well-being services county were assessed. Findings from this case study indicate potential productivity benefits of GAI in public sector translation flows. Specifically, postediting GAI-generated outputs was on average 14% faster than translating segments from scratch.

Although GAI has already improved the productivity of translators in the case organization, the potential of GAI remains partially unrealized. Further productivity gains are more likely to be harnessed by prioritizing the enhancement of organizational GAI capabilities, including the development of human skills, context-specific configuration, and fine-tuning of GAI systems, as well as effective change management, rather than solely emphasizing technological development. In addition, the integration of GAI tools in social and health care requires addressing ethical and societal questions, particularly given the sensitive nature of information handled and potential consequences of mistranslations.

The results are consistent with the study conducted by Macken et al [14] that similarly reported a 14% increase in translation speed when using neural MT. The efficiency of postediting process is significantly influenced by output quality, source text complexity, and translators' technical ability [25,26]. The proportion of 100% accurate MT segments in Finnish-Swedish

XSL•FO
RenderX

language pairs was merely 11%, highlighting the need for improving the quality of GAI MT quality to enhance productivity. It is also important to note that postediting involves distinct skills compared to traditional translation [27,28]. For instance, translators may choose to integrate their stylistic preferences into the text rather than limiting the effort to minimal corrections, which affects overall productivity.

The automatic metrics calculated in the study indicate that the quality of MTs can still be significantly improved. For instance, HTER scores of 55 and 46 (Swedish and English, respectively) suggest that a postedit is needed for approximately every other word in a segment to finalize MT output. In the study, translators frequently encountered issues with industry- and organization-specific terminology, which mirrors findings by Turner et al [29] and Zappatore and Ruggieri [30]. They identified domain adaptation as a critical challenge for MT adoption, particularly in specialized sectors, such as health care. Another significant barrier to GAI adoption was the current integration with CAT software (Trados), which operates at the segment level rather than incorporating the broader document-level context. This limitation undermines one of the core strengths of GPT-based models, which are designed to account for long-range dependencies and contextual information [31].

Prior research on MT user experience remains very limited. The primary focus of the earlier research has been on the technical development of MT, with less emphasis on ensuring the usefulness and usability of MT tools for translation professionals [18]. In general, translators tend to dislike postediting and prefer translation from scratch, even if MT post-editing is faster [32]. This observation is based on statistical MT and neural MT aids and may not fully reflect translation technology preferences regarding more recent LLM-based solutions. Recent studies have highlighted that professional users express concerns regarding the quality and reliability of ChatGPT when used for translation tasks. However, they also recognize the potential of such tools, particularly for auxiliary tasks, such as drafting, summarizing, or generating translation suggestions [33,34]. In addition, translators with prior experience in postediting exhibited more favorable attitudes toward the method [26].

In this case study, translators generally expressed a positive attitude toward GAI, with most preferring its use in suitable contexts over traditional methods. Translators' own assessments on productivity impacts of the GAI solution in use were more conservative than the results of the quantitative analysis.

Beyond technical challenges, human and organizational factors seem to play a critical role in realizing the full potential of GAI. The transition from conventional translation to MT postediting redefines the role of the human translator, shifting it from producing and crafting text to reviewing and refining output [25]. Translators emphasized the need for stronger technical support and adequate time for skill development, underscoring that effective adoption requires both training and hands-on experience with the tool. Negative experiences early in the adoption process, exacerbated by high workloads as well as inadequate preparation and resourcing, illustrate the importance of effective change management when introducing new technologies. These findings are consistent with Mikalef et al [5], who emphasize the importance of basic resources and human skills in AI adoption, alongside technological infrastructure and data capabilities.

Translators acknowledged the future potential of GAI systems to deliver faster and more consistent translations, reduce reliance on outsourcing, and improve workflow efficiency. However, concerns regarding job security and the loss of personal translation styles underscore the need for a balanced approach that leverages GAI as a tool to augment, rather than replace, human expertise. As Sezgin [35] and Yang et al [36] argue, the goal of GAI in fields such as public social and health care should be to support human professionals, not to automate all tasks, given the critical and complex nature of the work involved.

Strengths and Limitations

This mixed methods case study combines quantitative and qualitative methods to holistically assess impacts of GAI in language translation services within a public social and health care organization. It is based on widely used metrics and research design from previous studies. Another strength of the study is the practical relevance as it reflects real-life conditions to measure impacts of GAI in the specialized domain of social and health care.

Given the limited sample size, the findings need to be interpreted with caution and are not directly generalizable. In general, language translation teams within public social and health care sectors are small, which limits the sample size. The primary aim of the study is to explore early-stage implementation and practical implications of a GAI tool in a real-world setting. While this approach limits statistical generalization potential, investigating actual deployment in an authentic work environment can offer valuable, practice-oriented insights in how such tools are adopted, used, and experienced by professional translators.

In addition, the study likely captures only part of GAI's potential, as its effectiveness is influenced by the organization's current capability to integrate and optimize the technology. The data derived from the CAT software Trados also present limitations in fully representing the complexity of translators' daily workflows. For instance, segment-level time stamps in Trados may not accurately account for time spent on tasks due to interruptions or unforeseen delays. As a result, several assumptions were made during data collection, analysis, and interpretation, as outlined in detail in the "Methods" description.

Although well-established and widely adopted automatic evaluation metrics (BLEU, METEOR, ChrF, and HTER) were applied in this study, they have known limitations particularly in capturing semantic adequacy and fluency. More recent neural-based metrics, such as Crosslingual Optimized Metric for Evaluation of Translation [37] and Bilingual Evaluation Understudy with Representations from Transformers [38] have been proposed to better correlate with human judgment. Although these metrics were not widely standardized at the time of conducting this study, their growing potential should be considered for future research.

XSL•FO
RenderX

Conclusion

This mixed methods study demonstrates promising productivity potential of using GAI in language translation within public social and health care sector. While quantitative data captured measurable productivity gains from postediting GAI outputs, translators expressed more reserved views on its overall impact, suggesting a gap between observed efficiency and perceived benefit. Realizing the potential of GAI depends not only on technological performance but also on organizational capability. Based on the insights gathered in this study, several practical strategies are suggested to support effective GAI adoption in public social and health care sector:

- Invest in upskilling: Provide training and technical support to help translators effectively postedit and work with GAI outputs.
- Redesign operational workflows: Collaborate with translators to update translation workflows for effective integration of GAI technology, and create best practices for, for example, prompt formulation.
- Optimize GAI performance: Continuously refine GAI systems to improve translation quality, which in turn enhances productivity and user experience.
- Address ethical considerations: Recognize and mitigate translators' concerns related to new technology. Ensure human oversight to secure quality control, especially in sensitive domains such as social and health care.
- Monitor and evaluate impact: Define clear operational objectives. Regularly assess the effectiveness of GAI tools in translation services to support continuous development and informed decision-making.

Future research should aim to expand the scope of investigation to include larger-scale studies, providing more robust evidence on the impacts of GAI. In addition, greater academic attention should focus on end-user perspectives and evaluations in real-world settings. Further exploration of the specific components of GAI capability and their influence on productivity, quality, and user experience could offer valuable guidance for optimizing the design, implementation, and operational use of GAI systems in health and social services sector.

Acknowledgments

This study would not have been possible without the support of the case organization, Western Uusimaa Wellbeing Service County, and specifically, the in-house translation team members.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to the fact that the research permit granted by the Western Uusimaa Wellbeing Services County explicitly prohibits the sharing of these data outside the research team; however, anonymized and aggregated data are available from the corresponding author on reasonable request.

Authors' Contributions

JS contributed to conceptualization, project administration, supervision, and resources. MM led the conceptualization, investigation, formal analysis, methodology, project administration, and data curation. KS supported the formal analysis, investigation, and methodology and supervised the project. ES contributed to the provision of resources.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview questions.
[DOCX File, 17 KB - formative_v9i1e73658_app1.docx]

References

1. Lv Z. Generative artificial intelligence in the metaverse era. Cogn Robot 2023;3:208-217. [doi: 10.1016/j.cogr.2023.06.001]
2. Banh L, Strobel G. Generative artificial intelligence. Electron Markets 2023 Dec;33(1):63. [doi: 10.1007/s12525-023-00680-1]
3. AI ecosystem in social and health services (SOTE). DigiFinland. 2024. URL: https://digifinland.fi/en/ai-ecosystem-in-social-and-health-services-sote/ [accessed 2024-10-27]
4. Kosmas I, Papadopoulos T, Dede G, Michalakelis C. The use of artificial neural networks in the public sector. FinTech 2023;2(1):138-152. [doi: 10.3390/fintech2010010]
5. Mikalef P, Lemmer K, Schaefer C, et al. Examining how AI capabilities can foster organizational performance in public organizations. Gov Inf Q 2023 Apr;40(2):101797. [doi: 10.1016/j.giq.2022.101797]
6. Hjaltalin IT, Sigurdarson HT. The strategic use of AI in the public sector: a public values analysis of national AI strategies. Gov Inf Q 2024 Mar;41(1):101914. [doi: 10.1016/j.giq.2024.101914]

XSL•FO
RenderX

7. Wellbeing services counties will be responsible for organising health, social and rescue services. Ministry of Social Affairs and Health. 2023. URL: https://stm.fi/en/wellbeing-services-counties [accessed 2025-08-05]
8. Kaksikielisyys (suomi-ruotsi) [Web page in Finnish]. Hyvil. 2024. URL: https://www.hyvil.fi/hallinto-ja-juridiikka/kaksikielisyys-suomi-ruotsi/ [accessed 2024-10-27]
9. Sosiaali- ja terveydenhuollon järjestäminen Länsi-Uudenmaan hyvinvointialueella : Alueellinen asiantuntija-arvio, syksy 2023 [Report in Finnish]. : The Finnish Institute for Health and Welfare; 2023 URL: https://www.julkari.fi/handle/10024/147643 [accessed 2024-10-27]
10. Olsavszky V, Bazari M, Dai TB, et al. Digital translation platform (Translatly) to overcome communication barriers in clinical care: pilot study. JMIR Form Res 2025 Mar 14;9:e63095. [doi: 10.2196/63095] [Medline: 39451122]
11. Hutchins J, Lovtskii E. Petr Petrovich Troyanskii (1894–1950): a forgotten pioneer of mechanical translation. Mach Transl 2000 Sep;15(3):187-221. [doi: 10.1023/A:1011653602669]
12. Moussallem D, Wauer M, Ngomo ACN. Machine translation using semantic web technologies: a survey. J Web Semant 2018 Aug;51:1-19. [doi: 10.1016/j.websem.2018.07.001]
13. Nagarajan R, Kondo M, Salas F, et al. Economics and equity of large language models: health care perspective. J Med Internet Res 2024 Nov 14;26:e64226. [doi: 10.2196/64226] [Medline: 39541580]
14. Macken L, Prou D, Tezcan A. Quantifying the effect of machine translation in a high-quality human translation production process. Informatics (MDPI) 2020;7(2):12. [doi: 10.3390/informatics7020012]
15. Yan J, Yan P, Chen Y, et al. GPT-4 vs human translators: a comprehensive evaluation of translation quality across languages, domains, and expertise levels. arXiv. Preprint posted online on Jul 4, 2024. [doi: 10.48550/arXiv.2407.03658]
16. Wang X, Wang T, Muñoz Martín R, Jia Y. Investigating usability in postediting neural machine translation: evidence from translation trainees' self-perception and performance. Acr 2021;22(1):100-123. [doi: 10.1556/084.2021.00006]
17. Merali A. Scaling laws for economic productivity: experimental evidence in LLM-assisted translation. arXiv. Preprint posted online on Sep 4, 2024. [doi: 10.48550/arXiv.2409.02391]
18. Ragni V, Nunes Vieira L. What has changed with neural machine translation? A critical review of human factors. Perspect (Am Probat Parole Assoc) 2022 Jan 2;30(1):137-158. [doi: 10.1080/0907676X.2021.1889005]
19. Federico M, Cattelan A, Trombetti M. Measuring user productivity in machine translation enhanced computer-assisted translation. Presented at: Proceedings of the 10th Conference of the Association for Machine Translation in the Americas: Research Papers; Oct 28 to Nov 1, 2012; San Diego, CA.
20. Way A. Quality expectations of machine translation. In: Moorkens J, Castilho S, Gaspari F, Doherty S, editors. Translation Quality Assessment—From Principles to Practice: Springer; 2018, Vol. 1:159-178. [doi: 10.1007/978-3-319-91241-7_8]
21. Depraetere I. Introduction. In: Perspectives on Translation Quality: De Gruyter Mouton; 2011, Vol. 9:1-6. [doi: 10.1515/9783110259889]
22. Hadla LS, Hailat TM, Al-Kabi MN. Comparative study between METEOR and BLEU methods of MT: Arabic into English translation as a case study. Int J Adv Comput Sci Appl 2015;6(11):215-223. [doi: 10.14569/IJACSA.2015.061128]
23. Evaluating models. Google. 2024. URL: https://cloud.google.com/translate/automl/docs/evaluate#:~:text=A%20value%20of%200%20means,human%20judgment%20of%20translation%20quality [accessed 2024-08-01]
24. Son J, Kim B. Translation performance from the user's perspective of large language models and neural machine translation systems. Information 2023;14(10):574. [doi: 10.3390/info14100574]
25. Jia Y, Sun S. Man or machine? Comparing the difficulty of human translation versus neural machine translation post-editing. Perspect (Am Probat Parole Assoc) 2023 Sep 3;31(5):950-968. [doi: 10.1080/0907676X.2022.2129028]
26. Guerberof Arenas A. Productivity and quality in the post-editing of outputs from translation memories and machine translation. Int J Localisation 2012;7(1):11-21 [FREE Full text]
27. Garcia I. Translating by post-editing: is it the way forward? Mach Transl 2011 Sep;25(3):217-237. [doi: 10.1007/s10590-011-9115-8]
28. do Carmo F. 'Time is money' and the value of translation. Transl Spaces 2020 Aug 17;9(1):35-57. [doi: 10.1075/ts.00020.car]
29. Turner AM, Brownstein MK, Cole K, Karasz H, Kirchhoff K. Modeling workflow to design machine translation applications for public health practice. J Biomed Inform 2015 Feb;53:136-146. [doi: 10.1016/j.jbi.2014.10.005] [Medline: 25445922]
30. Zappatore M, Ruggieri G. Adopting machine translation in the healthcare sector: a methodological multi-criteria review. Comput Speech Lang 2024 Mar;84:101582. [doi: 10.1016/j.csl.2023.101582]
31. Nerella S, Bandyopadhyay S, Zhang J, et al. Transformers and large language models in healthcare: a review. Artif Intell Med 2024 Aug;154:102900. [doi: 10.1016/j.artmed.2024.102900] [Medline: 38878555]
32. Moorkens J, Toral A, Castilho S, Way A. Translators' perceptions of literary post-editing using statistical and neural machine translation. Transl Spaces 2018 Nov 28;7(2):240-262. [doi: 10.1075/ts.18014.moo]
33. Sahari Y, Al-Kadi AMT, Ali JKM. A cross sectional study of ChatGPT in translation: magnitude of use, attitudes, and uncertainties. J Psycholinguist Res 2023 Dec;52(6):2937-2954. [doi: 10.1007/s10936-023-10031-y] [Medline: 37934302]
34. Rivas Ginel MI, Moorkens J. One year of ChatGPT: translators' attitudes and degree of adoption. tradumatica 2024(22):258-275. [doi: 10.5565/rev/tradumatica.369]

XSL•FO
RenderX

35. Sezgin E. Artificial intelligence in healthcare: complementing, not replacing, doctors and healthcare providers. Digit Health 2023;9:20552076231186520. [doi: 10.1177/20552076231186520] [Medline: 37426593]
36. Yang R, Tan TF, Lu W, Thirunavukarasu AJ, Ting DSW, Liu N. Large language models in health care: development, applications, and challenges. Health Care Sci 2023 Aug;2(4):255-263. [doi: 10.1002/hcs2.61] [Medline: 38939520]
37. Guerreiro NM, Fonseca E, Bawden R, et al. xcomet: transparent machine translation evaluation through fine-grained error detection. Trans Assoc Comput Linguist 2024;12:979-995. [doi: 10.1162/tacl_a_00683]
38. Sellam T, Das D, Parikh A. BLEURT: learning robust metrics for text generation. Presented at: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; Jul 1-5, 2020. [doi: 10.18653/v1/2020.acl-main.704]

Abbreviations

AI: artificial intelligence
BLEU: Bilingual Evaluation Understudy
CAT: computer-assisted translation
ChrF: Character n-gram F-score
GAI: generative artificial intelligence
GDPR: General Data Protection Regulation
HTER: human-targeted translation edit rate
LLM: large language model
LUVN: Länsi-Uudenmaan hyvinvointialue
METEOR: Metric for Evaluation of Translation With Explicit Ordering
MT: machine translation
TM: translation memory

Edited by A Mavragani; submitted 12.03.25; peer-reviewed by C Tavares, L Macken; revised version received 28.06.25; accepted 30.06.25; published 17.09.25.

Please cite as:
Martikainen M, Smolander K, Sanmark J, Sanmark E
Evaluation of Generative Artificial Intelligence Implementation Impacts in Social and Health Care Language Translation: Mixed Methods Case Study
JMIR Form Res 2025;9:e73658
URL: https://formative.jmir.org/2025/1/e73658
doi:10.2196/73658

© Miia Martikainen, Kari Smolander, Johan Sanmark, Enni Sanmark. Originally published in JMIR Formative Research (https://formative.jmir.org), 17.9.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Perceptions of Stigma Among Patients With Hepatitis B in Germany: Cross-Sectional Survey

Bettina M Zimmermann[1,2,3], PhD; Theresa Willem[1,2,4], MA; Michael Rost[5], PhD; Nina Matthes[1]; Alena Buyx[1], Prof Dr med

[1]Institute of History and Ethics in Medicine, TUM School of Medicine and Health, Technical University of Munich, Ismaninger Str. 22, Munich, Germany

[2]Institute for Molecular Immunology, TUM School of Medicine and Health, Klinikum Rechts der Isar, Technical University of Munich, Munich, Germany

[3]Institute of Philosophy, Multidisciplinary Center for Infectious Diseases, University of Bern, Bern, Switzerland

[4]Department of Science, Technology and Society (STS), School of Social Sciences and Technology, Technical University of Munich, Munich, Germany

[5]Institute for Biomedical Ethics, University of Basel, Basel, Switzerland

Corresponding Author:
Bettina M Zimmermann, PhD
Institute of History and Ethics in Medicine, TUM School of Medicine and Health, Technical University of Munich, Ismaninger Str. 22, Munich, Germany

Abstract

Background: Many studies find associations between hepatitis B and stigma, but studies from the Western European context are lacking. Based on available studies, we hypothesized that younger age, higher education, male gender, higher privacy needs, and non-German mother tongue were positively associated with perceived hepatitis B–related stigma.

Objective: This study aims to describe the prevalence of perceived social stigma among patients with hepatitis B in Germany and to assess what factors are associated with perceptions of hepatitis B–related stigma.

Methods: Applying the short version of the Berger stigma scale, we surveyed 195 patients with hepatitis B about their perceptions of hepatitis B–related stigma, privacy needs, and demographic variables through a paper-based questionnaire. Venue-based recruitment of adult patients diagnosed with acute or chronic hepatitis B was implemented at 3 clinical centers in Germany. Patients who could not read German were excluded from the study.

Results: From the 195 valid questionnaires, 45.1% (88/195) of participants identified as female, 36.6% (71/195) had a high school diploma, and 56.9% (111/195) reported a mother tongue other than German. The mean (SD) stigma score throughout the sample was 5.52 (6.02; range 0-24) and the median was 3.50 (IQR=9.75). Regression analysis revealed that non-German mother tongue, individual data privacy needs, and participants' secrecy regarding their hepatitis B diagnosis independently predicted perceived hepatitis B-related stigma. More precisely, the higher the data privacy need and the more secret the hepatitis B diagnosis, the higher the perceived stigma, and perceived stigma was higher for patients with a non-German mother tongue. Age, gender, and education were no predictors of perceived stigma.

Conclusions: The surveyed patients with hepatitis B in Germany reported lower levels of hepatitis B–related stigma than found in other studies conducted in Asian countries. The association with non-German mother tongue indicates an important cultural and social component in the perception of stigma. Community-based interventions and the sensibilization of health care professionals might help overcome perceptions of stigma among hepatitis B–affected populations.

(JMIR Form Res 2025;9:e66379) doi:10.2196/66379

KEYWORDS

infectious diseases; infection; contaminate; septic shock; hepatitis B; social stigma; stigmatize; discrimination; prejudice; social factor; social disparity; social inequality; social inequity; Germany

Introduction

Background

Viral hepatitis encompasses a spectrum of liver diseases resulting from hepatotropic viral infections affecting hepatocytes. The 5 hepatotropic viruses, hepatitis types A, B, C, D, and E, exhibit varied acute manifestations, with outcomes ranging from spontaneous resolution to chronic infection [1]. The hepatitis B virus (HBV) is primarily transmitted through infected blood or body fluids, including unprotected sexual intercourse, needle-sharing, or perinatal exposure [2]. Chronic HBV infections affect over 3% of the world's population and pose substantial health risks, including cirrhosis and liver cancer

XSL•FO
RenderX

[3]. While Germany is categorized as a low-prevalence country for HBV, with a prevalence of 0.3% in the general population [4], a disproportionately high burden of HBV exists among immigrants from high-endemic countries who are estimated to contribute 49% of chronic HBV cases in Germany [5]. Although prophylactic vaccinations against HBV are available, the absence of a cure for chronic infections underscores the importance of ethical considerations in clinical research involving patients affected by HBV infections [6].

One particularly relevant ethical and social aspect in the context of infectious diseases, such as hepatitis B, is stigma [7,8]. Stigma is a "social process, experienced or anticipated, characterized by exclusion, rejection, blame or devaluation" [7] where "some characteristic or behavior of others [are deemed] as falling short of communal norms of desirability" [9]. Stigma can be social (leading, for instance, to social isolation or loss of social status), internalized (referring to a person's perception or anticipation of being socially rejected), or structural (causing disadvantages in accessing high-quality social goods, such as access to health care or university admissions) [10]. People affected by infectious diseases (especially chronic forms) are prone to stigma because of the transmissibility of the disease: the fear of infecting others might lead to social isolation and internalized stigma [11]. However, even uninfected individuals who share characteristics associated with this disease tend to be stigmatized. For example, men who have sex with men and sex workers tend to be associated with sexually transmitted infections (ie, from HIV or HBV) and may suffer from the same stigma even if unaffected [12]. Stigma has been shown to negatively affect the quality of life [13] and mental health [14] of people with hepatitis B. Hepatitis B–related stigma also leads to reduced screening rates in at-risk populations [15,16], lower adherence to antiviral treatment [17] and presents a barrier to accessing health care [18]. Thus, stigma is a major barrier in hepatitis B treatment, prevention, and elimination [19,20].

Against this backdrop, many studies worldwide have assessed stigma related to infectious diseases. A study from Hong Kong reported that around one-third of HIV-affected participants experienced discrimination due to their condition, particularly by health care professionals [21]. Patients affected by chronic hepatitis C also experience social and internalized stigma [22], and some comparative studies report higher stigma scores among patients with hepatitis C as compared with hepatitis B [8,23]. Still, studies report considerable ratios of stigma among patients with hepatitis B, ranging from high perceptions of stigma in 47.4% of surveyed patients in Turkey [23] to over 60% in India [10]. Another Turkish survey found 19.5% of hepatitis B-affected participants reporting experienced stigma, and 27.4% were worried about stigma [24]. Toumi et al [11] illustrate self-stigma and report on experienced social stigma around people affected by hepatitis B in several world regions. Overall, hepatitis B–related stigma was reported to be insufficiently described in populations outside Asia [25].

Study Hypotheses

This study addresses this gap by reporting perceptions of stigma among patients with hepatitis B in Germany. It also aims to identify factors associated with perceptions of hepatitis B–related stigma. To this end, we developed the following hypotheses based on findings in other studies (referenced below) investigating factors associated with stigma related to hepatitis B or C:

H1: Men with hepatitis B report higher stigma than women with hepatitis B [23,26].

H2: Patients with hepatitis B with higher general education (high school degree) report higher stigma [27].

H3: Older age is associated with less perception of stigma [28].

H4: Patients with hepatitis B and a mother tongue other than German report higher stigma than those with solely a German mother tongue [29].

Further, qualitative interviews with patients with hepatitis B (results are reported elsewhere [30]) suggest a strong connection between privacy attitudes and stigma. Since we measured privacy attitudes in terms of the secrecy of HBV diagnosis and general data privacy needs, we further hypothesize:

H5a: Higher perceptions of stigma are associated with higher secrecy regarding HBV diagnosis.

H5b: Higher perceptions of stigma are associated with higher data privacy needs.

Methods

Overview

This study is part of the European Union–funded international research consortium "TherVacB–A Therapeutic Vaccine to Cure hepatitis B." As part of work package 6, our research team aimed to assess the ethical, legal, and social aspects of social media recruitment. Preceding the here-presented survey study, a qualitative multistakeholder interview study was conducted and informed survey construction. The primary aim of the subsequent quantitative survey study among patients with hepatitis B in Germany was to assess the acceptance of social media recruitment. We report the findings of this original survey study elsewhere [30]. This paper presents a secondary analysis from this survey study based on the same questionnaire, investigating patients with hepatitis B perceptions of stigma and factors associated with stigma perceptions. Reporting follows the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines [31].

Study nurses distributed a paper-based questionnaire to adults diagnosed with acute or chronic hepatitis B through 3 clinical centres in Germany (Munich, Hannover, and Leipzig). Even though study nurses were instructed to hand out the questionnaire to every incoming patient with hepatitis B (hospitals estimated n=939) and to avoid self-selection, only 285 questionnaires (22%) were distributed. The main reason was an overburden of workload from the study nurses in the aftermath of the COVID-19 pandemic. Therefore, the data collection period was extended from initially 7 to a total of 12 months (June 2022 to May 2023) until the goal of collecting 200 questionnaires was reached. The required sample size was calculated using G*Power, a free tool to compute statistical power analyses developed by researchers from the Heinrich

XSL•FO
RenderX

Heine University Dusseldorf [32]. Given an effect size Cohen f_2 of 0.15, an α error probability of .05, and 11 predictors, 178 participants would result in a power of 0.95. Data from the paper-based questionnaires were typed manually into an excel sheet. Data entry was double checked by a second researcher for correctness.

Ethical Considerations

The ethics committees from the Technical University of Munich (12/22-S-NP), Hannover Medical School (10368_BO_K_2022), and University Clinic Leipzig (189/22-lk) approved the study. Informed consent was obtained from all participants in this study. To protect their privacy, participants gave their consent by ticking a box at the beginning of the questionnaire rather than signing a consent form. For opting out, participants had to keep and state a unique identifier that they received on their study information leaflet. No names or other identifying information was collected. Participants did not receive any compensation for participating in this study.

Survey Construction

Survey construction involved several steps. First, based on the research team's expertise and findings of a preceding qualitative study, we discussed potential independent variables that possibly predict perceived stigma (see the hypotheses mentioned above). Then, we identified validated questionnaires that captured possible predictors or the dependent variable. If no validated tools existed for a variable of interest, we developed scales for the respective variables. The compiled survey was pretested with 6 patients with hepatitis B (3 of them with a non-German mother tongue), leading to minor changes.

The survey included questions about (1) participants' general use of social media, (2) self-reported digital literacy, (3) participants' use of social media in connection to their hepatitis B infection, (4) their interest in clinical studies, (5) trust in medical and non-medical information sources, (6) acceptance of social media as a recruitment tool for clinical studies, (7a) their attitudes regarding data privacy, (7b) perceived secrecy of their hepatitis B, and (8) perceived stigma related to hepatitis B. Because this is a secondary study, we included only variables with a theoretical or empirical connection to stigma in this analysis, excluding variables about social media (no. 1, 2, 3, 6), interest in clinical studies (no. 4) and trusted information sources (no. 5). Responses were measured with a 5-point Likert scale from 0=never or do not agree at all, to 4=always or fully agree. The full questionnaire is available in Multimedia Appendix 1.

Perceived hepatitis B–related stigma was measured using a 6-item version of the hepatitis C stigma scale [33]. This scale was modified from the validated Berger HIV stigma scale [34], and the short 6-item version showed comparable results to the original 40-item scale [33]. For our questionnaire, "hepatitis C" was replaced by "hepatitis B," and the original items were translated into German by back-and-forth translation by a professional translator. This scale showed excellent reliability in our sample (α=.931). Secrecy and data privacy needs were self-developed and included 2 items each. For secrecy, measuring the willingness of patients with hepatitis B to tell people they know about their condition, the items were "*My family and friends know about my Hepatitis B infection*" and "*My hepatitis B infection is a secret*" (Multimedia Appendix 1). For data privacy, the items were "*I am careful not to disclose anything about my hepatitis B infection on social media for fear that the platform will collect and store this information*" and "*I want my medical data in connection with my illness to be particularly well protected*." Both scales were developed based on findings with qualitative interviews including 6 patients with hepatitis B [30] and showed acceptable reliability (secrecy: α=.623; data privacy needs: α=.587). Besides these adapted and self-developed scales, we included 4 demographic variables in the analysis (age, gender, education, and mother tongue).

Statistical Analysis and Predictor Selection

We used SPSS (version 28.0 IBM) for data analysis, performing descriptive analyses and determining independent factors associated with participants' perceived stigma using multiple linear regression analysis. The statistical significance level was set at $P<.05$. For multiple linear regression analysis, assumption checks were performed before the interpretation of the model (Multimedia Appendix 2).

For our analysis, a literature review and hypotheses derived from our previous qualitative study determined predictor selection (hypotheses above). In addition, the sample-size and predictor ratio a priori determines variable selection for regression modeling. According to Harrell, a fitted regression model is likely to be reliable when $P<m/10$ or $P<m/20$ (average requirement: $P<m/15$), where P is the number of predictors and m is the sample size [35]. Applying this requirement to our sample size (N=195) and having missing data (80.6% of variables, 28.7% of cases, and 3.7% of values), the *a priori* limitation was 11 included predictors. The presented analysis included 6 predictors in the regression model for which we had hypothesized an association with hepatitis B–related stigma: age, gender (dichotomous), education (dichotomous), German or non-German mother tongue (dichotomous), secrecy of HBV diagnosis, and data privacy needs related to HBV diagnosis.

Results

Sociodemographic Characteristics of the Sample

From the 285 questionnaires distributed, 207 (72.6%) were returned. Because 12 questionnaires had to be excluded due to lack of consent, the final analysis is based on 195 questionnaires. Table 1 portrays the participants' demographic characteristics.

XSL•FO
RenderX

Table . Participant characteristics (N=195).

Characteristics	Values, n (%)
Gender	
Men	101 (51.8)
Women	88 (45.1)
Missing	6 (3.1)
Age (years)	
18 - 29	16 (8.2)
30 - 39	50 (25.6)
40 - 49	58 (29.7)
50 - 59	38 (19.5)
More than 60	24 (12.3)
Missing	9 (4.6)
Education: highschool diploma	
Yes	71 (36.4)
No	110 (56.4)
Missing	14 (7.2)
Mother tongue (multiple answers possible)	
German	101 (51.8)
Other	111 (56.9)
Missing	12 (6.2)
Total	195 (100)

Description of Scales

The perceived hepatitis B–related stigma score was calculated by building the sum of the 6 items (range 0-24). We excluded 15/195 (7.7%) questionnaires because, in these, not all items were answered. The mean score throughout the sample was 5.52 (SD 6.02), and the median was 3.50 (IQR 9.75. As Table 2 indicates as many as 50/180 (27.8%) participants who completed this score indicated zero perceived hepatitis B–related stigma, selecting “completely disagree” to all 6 items that asked them about various dimensions of hepatitis B–related stigma (Table 2).

XSL•FO
RenderX

Table . Distribution of perceived stigma score (N=180). The summative stigma score was calculated by building the sum of the 6 stigma-related items.

Summative stigma score	Case count
0	50
1	18
2	8
3	11
4	9
5	11
6	10
7	8
8	3
9	2
10	4
11	9
12	10
13	6
14	5
15	3
16	2
17	1
18	3
19	1
20	2
21	1
22	1
23	0
24	3

Regression Analysis

Using multiple linear regression analyses, we evaluated predictors' associations with participants' perceived hepatitis B–related stigma. Testing the statistical significance of the overall model fit, the *F* test indicated that the predictors included in the model significantly contributed to the explanation of perceived hepatitis B–related stigma as the dependent variable (Table 3). Regression analysis revealed that a non-German mother tongue, privacy, and the secrecy of the HBV diagnosis independently predicted perceived stigma. More precisely, the higher the data privacy needs and the more secret hepatitis B, the higher the perceived stigma, and perceived stigma was higher for patients with non-German mother tongue (Table 3, Figure 1).

XSL•FO
RenderX

Table . Multiple linear regression analysis (n=161, dependent variable: perceived hepatitis B–related stigma)[a].

	Unstandardized coefficients (B)	SE	β	*t* test (*df*)	*P* value	Tolerance	Variance inflation factor
Constant	−2.700	2.978	N/A[c]	−.907 (154)	.37	N/A[c]	
Age	.025	.040	.049	.623 (154)	.53	.882	1.134
Gender[b]	−.255	.893	−.021	−.286 (154)	.78	.972	1.029
Education[b]	−.321	.914	−.026	−.351 (154)	.72	.948	1.054
Non-German mother tongue[b]	3.318	.941	.272	3.525 (154)	<.001	.903	1.107
Secrecy	.522	.224	.182	2.333 (154)	.02	.882	1.134
Data privacy needs	.715	.236	.239	3.034 (154)	.003	.864	1.157

[a]Overall model fit: F_{6154}=5.293, P<.001; R^2=.171; n=161

[b]Dichotomous items.

[c]N/A = not applicable

XSL•FO
RenderX

Figure 1. Relationships between perceived stigma and (**A**) secrecy, (**B**) privacy, and (**C**) non-German mother tongue.

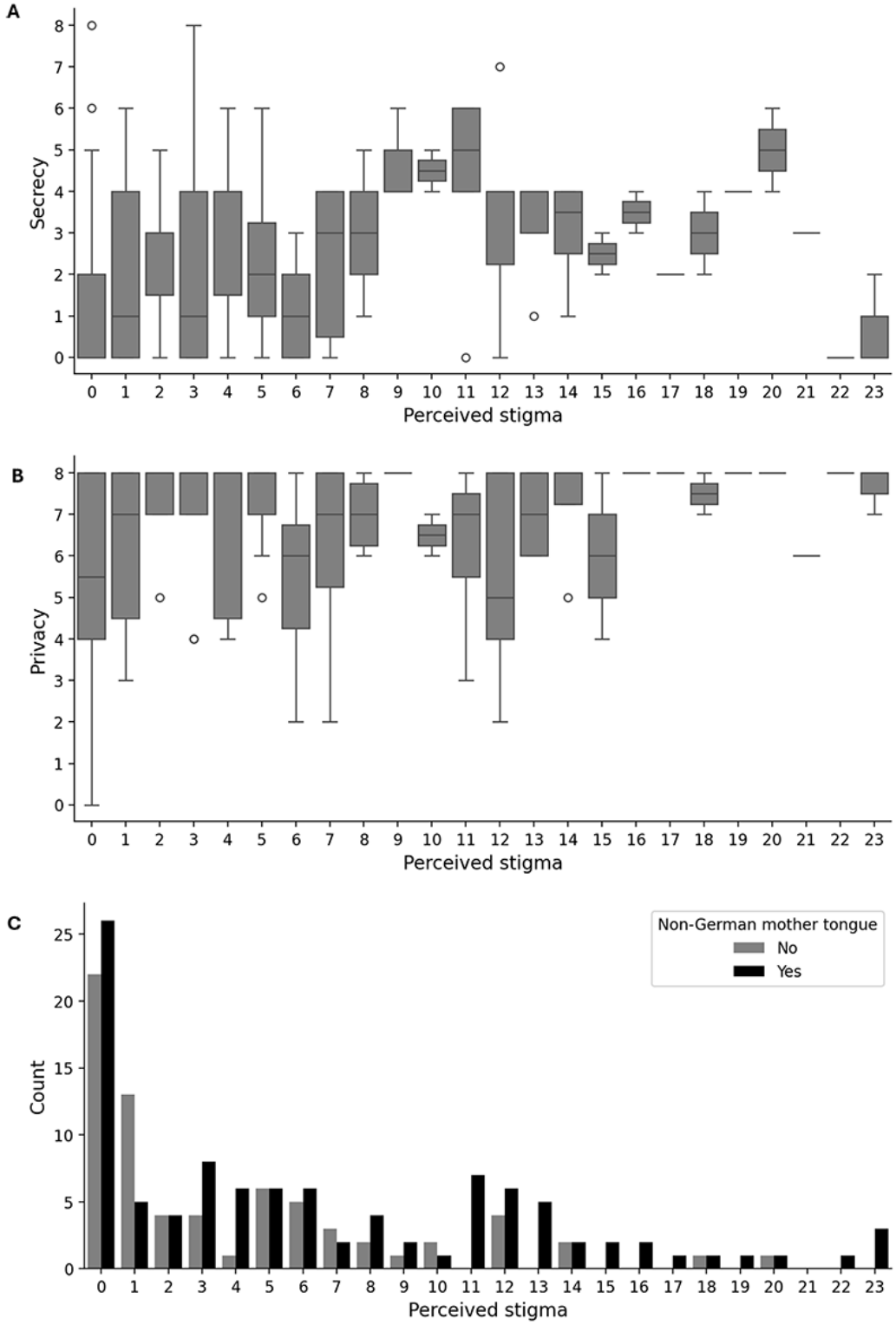

Discussion

Principal Results

The majority of study participants (patients with hepatitis B living in Germany) report moderate levels of perceived stigma. In their systematic review, Smith-Palmer et al [25] (2020) concluded that hepatitis B–related stigma was poorly characterized in non-Asian populations. Our findings suggest that perceived stigma might be lower among German-speaking patients with hepatitis B in Germany than among the other populations studied, where strong experiences or perceptions of stigma ranged from 20% - 60% of surveyed patients with hepatitis B [10,23,24]. However, quantitative comparisons between studies are difficult because stigma is measured in different ways in the scholarly literature. Comparative studies accounting for cultural dimensions of stigma and potential

XSL•FO
RenderX

language bias have to confirm whether perceived stigma is lower in Germany than in Asian populations.

The Role of Mother Tongue

Despite the above-mentioned under-representation of patients with hepatitis B with insufficient German language skills, the sample was large and diverse enough to show significantly higher perceptions of stigma in participants with a mother tongue other than German (confirming H4, Table 4). A study by Dam et al [36] juxtaposed stigma perception in Vietnamese patients with hepatitis B in Vietnam and Vietnamese Americans in the United States and found considerable differences in stigma perceptions, with the cohort in Vietnam being more progressive in their perspectives about hepatitis B and its attached stigma than the American immigrant cohort. This might represent an additional explanation for the relatively low overall levels of stigma measured in our cohort as compared with Asian cohorts.

Yet, in our study, patients with a non-German mother tongue perceived more hepatitis B–related stigma than people with a German mother tongue. Even though it was not possible in our study to analyze where participants with non-German mother tongue originated from, Dam et al [36] study pertains to the broader social context of stigma, indicating that immigration decontextualizes individuals from their social environments, which might change how they expect their environment to perceive them. Since participants with non-German mother tongue (or their families) most likely have an immigration background, our study speaks to the confounding impacts of social context changes due to immigration on immigrant patients' perceptions of illness and its attached stigma [37]. Indeed, a qualitative study among Chinese immigrants in Australia found that hepatitis B–related stigma mainly manifested in anticipated stigma and rooted in the ethnic and cultural backgrounds of the community [38]. Further studies are needed to assess perceptions of disease-related stigma in other populations of foreign origin.

Privacy Needs and Vulnerability

Patients reporting higher hepatitis B–related stigma also reported higher data privacy needs and more strongly perceived their diagnosis as a secret, confirming H5a and H5b (Table 4). This confirms findings from a qualitative interview study of our group that included 6 patients with hepatitis B patients from Germany [30,39]. Within this interview study, a predominant concern voiced by the interviewed patients revolved around the potential compromise of their privacy. Patients were worried that the information they shared on social media platforms could be easily disseminated or reposted without their consent. Interviewed patients expressed apprehension regarding the possible unauthorized dissemination or reposting of their health data on social media platforms, highlighting their fear of potential breaches of confidentiality. Their apprehension stems from the anticipated societal judgment and discrimination that could arise upon the disclosure of their HBV diagnosis.

Stigma (whether anticipated or experienced) adds an important layer of vulnerability to patients [40]. Thus, data privacy should be particularly strict for patients with stigmatized conditions. This affects, in particular, public outreach programs and online community work, for instance, on social media [39]. Hepatitis B–related health complications, as well as immigration status that comes with challenges related to poor language skills, assimilation, and cultural imprinting, may add additional layers of vulnerability, leading to cascading effects that exacerbate vulnerability [41]. Privacy-preserving public health programs targeted toward high-prevalence populations that aim to inform and prevent infection should also address these perceptions of stigma and the exacerbating vulnerabilities within those populations. Such community-based interventions and the sensibilization of health care professionals might help overcome remaining perceptions of stigma among hepatitis B–affected populations. They might contribute to increased testing rates in at-risk populations [42].

Contrary to what we had hypothesized, gender (H1), age (H3), and education (H2) were no predictors of stigma (Table 4). However, disease-related literacy, which is what is portrayed in the literature to be negatively associated with stigma [10], was not measured directly. As suggested in other studies 17], it might still be relevant that sociomedical education regarding local perceptions of stigma (particularly in countries with low stigma rates) could serve as a harm-mitigating factor in highly stigmatized patient communities.

Table . Outcomes of study hypotheses.

Hypotheses	Outcome
(H1) Men with hepatitis B report higher stigma than women with hepatitis B [23,26].	Rejected
(H2) Patients with hepatitis B with higher general education (high school degree) report higher stigma [27].	Rejected
(H3) Older age is associated with less perception of stigma [28].	Rejected
(H4) Patients with hepatitis B with a mother tongue other than German report higher stigma than those with a German mother tongue [29].	Confirmed
(H5a) Higher perceptions of stigma are associated with higher secrecy regarding HBV[a] diagnosis.	Confirmed
(H5b) Higher perceptions of stigma are associated with higher data privacy needs.	Confirmed

[a]HBV: hepatitis B virus.

XSL•FO
RenderX

Limitations

Even though the survey sample is representative of the hepatitis B population regarding age and gender when compared with the most recently measured prevalence of chronic hepatitis B in Germany [4], a relevant selection bias might lead to an underestimation of perceived stigma among patients with hepatitis B in Germany. First, the venue-based recruitment in 3 major clinical centres might have contributed to such a selection bias. Second, the questionnaire was only available in the German language, which led to an underrepresentation of patients who have insufficient German language skills to understand the survey. The questionnaire was also not designed with respect to cultural aspects of stigma from different immigrant populations.

Representing an additional study limitation, this paper reports a secondary analysis and stigma was not the primary dependent variable when the survey was designed [30]. Therefore, some variables that are potentially associated with stigma were not included in this analysis but might have increased the explanatory power of the regression model. Such variables that have been shown in other studies to affect perceptions of stigma are disease-related knowledge [10,43,44], patients' socioeconomic status or employment level [10], comorbidities [28], or family history of hepatitis B [27]. Other potentially relevant factors that were not measured include occupation, geographic location (urban or rural), duration of HBV diagnosis, or access to social support. Our regression model explained 17% of the variance of stigma perception, pointing to the multifaceted nature of stigma perceptions. We also used the short version of the stigma scale to increase the chance that patients would complete the survey and did not distinguish between perceived, enacted, and internalized stigma. While this short version showed excellent reliability and was validated [34], no in-depth analysis of stigma subscales was possible.

Because the original study focused on social media, there are potential framing effects that participants might target their answers to social media contexts. However, the acceptance of social media as a recruitment tool for hepatitis B–related clinical studies was low in this cohort and secrecy levels were high [30]. Since we would expect framing effects to increase stigma perceptions in this context, the findings of low stigma perceptions seem robust.

Conclusions

This is the first study in a Western European setting assessing perceived stigma among patients with hepatitis B. Despite its limitations related to the secondary analysis of this survey, the study adds to the evidence of perceived hepatitis B–related stigma, which has been predominantly conducted in Asian settings. The survey revealed that self-reported hepatitis B-related stigma was less pronounced among German-speaking patients with hepatitis B in Germany than reported in studies conducted in other world regions. Yet, participants with a non-German mother tongue perceived more hepatitis B–related stigma than those with a German mother tongue. While this study excludes patients with hepatitis B who do not speak German, these findings point to the need for further research into the context of stigma perceptions among patients with immigrant backgrounds in Germany and other understudied European contexts. High perceptions of stigma were also associated with higher secrecy around the HBV diagnosis and higher data privacy needs, indicating the need to use privacy-preserving strategies in hepatitis B–related public health programs.

Acknowledgments

We thank the staff at clinical centres in Munich, Hannover, and Leipzig who recruited study participants and returned the questionnaires. The research for this article received funding from the European Union's Horizon 2020 research and innovation program under grant agreement number 848223 (TherVacB).

Data Availability

The datasets generated or analyzed during this study are not publicly available because of the potential for reidentification and lack of participant consent but are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Questionnaire.
[DOCX File, 60 KB - formative_v9i1e66379_app1.docx]

Multimedia Appendix 2
Supplementary methods.
[DOCX File, 18 KB - formative_v9i1e66379_app2.docx]

References

XSL•FO
RenderX

1. Cheung A, Kwo P. Viral hepatitis other than A, B, and C: evaluation and management. Clin Liver Dis 2020 Aug;24(3):405-419. [doi: 10.1016/j.cld.2020.04.008] [Medline: 32620280]
2. Tu T, Douglas MW. Hepatitis B virus infection: from diagnostics to treatments. Viruses 2020 Nov 30;12(12):1366. [doi: 10.3390/v12121366] [Medline: 33265922]
3. Hsu YC, Huang DQ, Nguyen MH. Global burden of hepatitis B virus: current status, missed opportunities and a call for action. Nat Rev Gastroenterol Hepatol 2023 Aug;20(8):524-537. [doi: 10.1038/s41575-023-00760-9] [Medline: 37024566]
4. Poethko-Müller C, Zimmermann R, Hamouda O, et al. Die Seroepidemiologie der Hepatitis A, B und C in Deutschland [Article in Dutch]. Bundesgesundheitsbl 2013 May;56(5-6):707-715. [doi: 10.1007/s00103-013-1673-x]
5. Falla A, Veldhuijzen I, Ahmad A, European Centre for Disease Prevention and Control. Epidemiological assessment of hepatitis B and C among migrants in the EU/EEA. 2016 URL: https://www.ecdc.europa.eu/sites/default/files/media/en/publications/Publications/epidemiological-assessment-hepatitis-B-and-C-among-migrants-EU-EEA.pdf [accessed 2025-06-09]
6. Hudu SA, Jimoh AO, Ibrahim KG, Alshrari AS. Hepatitis B therapeutic vaccine: a patent review. Pharmaceuticals (Basel) 2022 Dec 12;15(12):1542. [doi: 10.3390/ph15121542] [Medline: 36558991]
7. Scambler G. Health-related stigma. Sociol Health Illn 2009 Apr;31(3):441-455. [doi: 10.1111/j.1467-9566.2009.01161.x] [Medline: 19366430]
8. Drazic YN, Caltabiano ML. Chronic hepatitis B and C: Exploring perceived stigma, disease information, and health-related quality of life. Nurs Health Sci 2013 Jun;15(2):172-178. [doi: 10.1111/nhs.12009] [Medline: 23171324]
9. Faden R, Bernstein J, Shebaya S. Public health ethics. In: Zalta EN, editor. The Stanford Encyclopedia of Philosophy: Metaphysics Research Lab, Stanford University; 2022. URL: https://plato.stanford.edu/archives/spr2025/entries/publichealth-ethics/ [accessed 2025-06-12]
10. Behera MK, Nath P, Behera SK, Padhi PK, Singh A, Singh SP. Unemployment and illiteracy are predictors of hepatitis B virus-related stigma and discrimination. J Clin Exp Hepatol 2022;12(3):767-773. [doi: 10.1016/j.jceh.2021.12.006] [Medline: 35677504]
11. Toumi M, Wallace J, Cohen C, et al. Experience and impact of stigma in people with chronic hepatitis B: a qualitative study in Asia, Europe, and the United States. BMC Public Health 2024 Feb 26;24(1):611. [doi: 10.1186/s12889-023-17263-6] [Medline: 38408941]
12. Fitzgerald-Husek A, Van Wert MJ, Ewing WF, et al. Measuring stigma affecting sex workers (SW) and men who have sex with men (MSM): A systematic review. PLoS ONE 2017;12(11):e0188393. [doi: 10.1371/journal.pone.0188393] [Medline: 29190642]
13. Ahmed S, Méndez RY, Naveed S, et al. Assessment of hepatitis-related knowledge, attitudes, and practices on quality of life with the moderating role of internalized stigma among hepatitis B-positive patients in Pakistan. Health Psychol Behav Med 2023;11(1):2192782. [doi: 10.1080/21642850.2023.2192782] [Medline: 37008419]
14. Katcher JG, Klassen AC, Hann HW, Chang M, Juon HS. Racial discrimination, knowledge, and health outcomes: the mediating role of hepatitis B-related stigma among patients with chronic hepatitis B. J Viral Hepat 2024 May;31(5):248-254. [doi: 10.1111/jvh.13932] [Medline: 38409935]
15. Adjei CA, Stutterheim SE, Bram F, Naab F, Ruiter RAC. Correlates of hepatitis B testing in Ghana: The role of knowledge, stigma endorsement and knowing someone with hepatitis B. Health Soc Care Community 2022 Nov;30(6):e4564-e4573. [doi: 10.1111/hsc.13860] [Medline: 35701984]
16. Li D, Tang T, Patterson M, Ho M, Heathcote J, Shah H. The impact of hepatitis B knowledge and stigma on screening in Canadian Chinese persons. Can J Gastroenterol 2012 Sep;26(9):597-602. [doi: 10.1155/2012/705094] [Medline: 22993729]
17. Juon HS, Yang D, Fang CX, et al. Perceived HBV-related stigma Is associated with lower antiviral medication adherence in patients with chronic hepatitis B. J Viral Hepat 2025 Mar;32(3):e70010. [doi: 10.1111/jvh.70010] [Medline: 39953814]
18. Cama E, Beadman M, Beadman K, Hopwood M, Treloar C. Health workers' perspectives of hepatitis B-related stigma among Aboriginal and Torres Strait Islander people in New South Wales, Australia. Harm Reduct J 2023 Aug 26;20(1):116. [doi: 10.1186/s12954-023-00844-4] [Medline: 37633903]
19. Tu T. Stigma: a major barrier to hepatitis B elimination. Nat Rev Gastroenterol Hepatol 2022 Oct;19(10):622. [doi: 10.1038/s41575-022-00670-2] [Medline: 35922484]
20. Freeland C, Wallace J, Wang S, et al. The urgent need to end hepatitis B stigma and discrimination. Lancet Gastroenterol Hepatol 2025 Feb;10(2):105-107. [doi: 10.1016/S2468-1253(24)00389-3] [Medline: 39706211]
21. Chan RCH, Mak WWS, Ma GYK, Cheung M. Interpersonal and intrapersonal manifestations of HIV stigma and their impacts on psychological distress and life satisfaction among people living with HIV: Toward a dual-process model. Qual Life Res 2021 Jan;30(1):145-156. [doi: 10.1007/s11136-020-02618-y] [Medline: 32909160]
22. Kostić M, Kocić B, Tiodorović B. Stigmatization and discrimination of patients with chronic hepatitis C. Vojnosanit Pregl 2016 Dec;73(12):1116-1124. [doi: 10.2298/VSP150511135K] [Medline: 29341568]
23. Yozgat A, Can G, Can H, et al. Social stigmatization in Turkish patients with chronic hepatitis B and C. Gastroenterol Hepatol 2021 May;44(5):330-336. [doi: 10.1016/j.gastrohep.2020.08.008] [Medline: 33213938]
24. Esen Yıldız İ, Bahçeci İ, Yılmaz Yavuz A, Kostakoğlu U, Ertürk A. Assessment of stigma exposure status of patients with hepatitis B infection. vhd 2021 Aug 1;27(2):80-88. [doi: 10.4274/vhd.galenos.2021.2021-5-1]

XSL•FO
RenderX

25. Smith-Palmer J, Cerri K, Sbarigia U, et al. Impact of stigma on people living with chronic hepatitis B. Patient Relat Outcome Meas 2020;11:95-107. [doi: 10.2147/PROM.S226936] [Medline: 32214859]
26. Adekanle O, Komolafe AO, Ijarotimi O, Olowookere AS, Ndububa DA. Gender disparity and stigma experience of patients with chronic hepatitis B virus infection: a prospective cross-sectional study from a hospital in Nigeria. J Infect Prev 2022 Nov;23(6):263-268. [doi: 10.1177/17571774221127546] [Medline: 36277856]
27. Marley G, Seto WK, Yan W, et al. What facilitates hepatitis B and hepatitis C testing and the role of stigma among primary care patients in China? J Viral Hepat 2022 Aug;29(8):637-645. [doi: 10.1111/jvh.13711] [Medline: 35633086]
28. Le TV, Vu TTM, Mai HT, et al. Social determinants of stigma and discrimination in Vietnamese patients with chronic hepatitis B. Int J Environ Res Public Health 2019 Jan 31;16(3):398. [doi: 10.3390/ijerph16030398] [Medline: 30708943]
29. Pachankis JE, Hatzenbuehler ML, Berg RC, et al. Anti-LGBT and anti-immigrant structural stigma: an intersectional analysis of sexual minority men's HIV risk when migrating to or within europe. J Acquir Immune Defic Syndr 2017 Dec 1;76(4):356-366. [doi: 10.1097/QAI.0000000000001519] [Medline: 28787329]
30. Willem T, Zimmermann BM, Matthes N, Rost M, Buyx A. Acceptance of social media recruitment for clinical studies among patients with hepatitis B: mixed methods study. J Med Internet Res 2024 Aug 26;26:e54034. [doi: 10.2196/54034] [Medline: 39186322]
31. von Elm E, Altman DG, Egger M, et al. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet 2007 Oct 20;370(9596):1453-1457. [doi: 10.1016/S0140-6736(07)61602-X] [Medline: 18064739]
32. Faul F, Erdfelder E, Buchner A, Lang AG. Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. Behav Res Methods 2009 Nov;41(4):1149-1160. [doi: 10.3758/BRM.41.4.1149] [Medline: 19897823]
33. Saine ME, Moore TM, Szymczak JE, et al. Validation of a modified Berger HIV stigma scale for use among patients with hepatitis C virus (HCV) infection. PLoS ONE 2020;15(2):e0228471. [doi: 10.1371/journal.pone.0228471] [Medline: 32023310]
34. Berger BE, Ferrans CE, Lashley FR. Measuring stigma in people with HIV: psychometric assessment of the HIV stigma scale. Res Nurs Health 2001 Dec;24(6):518-529. [doi: 10.1002/nur.10011] [Medline: 11746080]
35. Harrell FE. Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis, th edition: Springer; 2015. [doi: 10.1007/978-3-319-19425-7]
36. Dam L, Cheng A, Tran P, et al. Hepatitis B stigma and knowledge among vietnamese in Ho Chi Minh city and Chicago. Can J Gastroenterol Hepatol 2016;2016:1910292. [doi: 10.1155/2016/1910292] [Medline: 28101498]
37. Cabaniss ER, Cameron AE. Toward a social psychological understanding of migration and assimilation. Humanity Soc 2018 May;42(2):171-192. [doi: 10.1177/0160597617716963]
38. Jin D, Brener L, Treloar C. Hepatitis B-related stigma among Chinese immigrants living with hepatitis B virus in Australia: a qualitative study. Health Soc Care Community 2022 Nov;30(6):e5602-e5611. [doi: 10.1111/hsc.13986] [Medline: 36068665]
39. Goldman N, Willem T, Buyx A, Zimmermann BM. Practical benefits, challenges, and recommendations on social media recruitment: multi-stakeholder interview study. J Med Internet Res 2023 May 22;25:e44587. [doi: 10.2196/44587] [Medline: 37213177]
40. Luna F. Elucidating the concept of vulnerability: layers not labels. Int J Fem Approaches Bioeth 2009 Mar;2(1):121-139. [doi: 10.3138/ijfab.2.1.121]
41. Luna F. Identifying and evaluating layers of vulnerability - a way forward. Dev World Bioeth 2019 Jun;19(2):86-95. [doi: 10.1111/dewb.12206] [Medline: 30058768]
42. Lenton E, Johnson J, Brown G. Upscaling HIV and hepatitis C testing in primary healthcare settings: stigma-sensitive practice. Aust J Prim Health 2021 Aug;27(4):255-258. [doi: 10.1071/PY20176] [Medline: 34154704]
43. Valizadeh L, Zamanzadeh V, Bayani M, Zabihi A. The social stigma experience in patients with hepatitis B infection: a qualitative study. Gastroenterol Nurs 2017;40(2):143-150. [doi: 10.1097/SGA.0000000000000223] [Medline: 28362662]
44. Adjei CA, Stutterheim SE, Naab F, Ruiter RAC. Chronic Hepatitis B stigma in Ghana: a qualitative study with patients and providers. BMJ Open 2019 Jun 27;9(6):e025503. [doi: 10.1136/bmjopen-2018-025503] [Medline: 31248915]

Abbreviations

HBV: hepatitis B virus
STROBE: Strengthening the Reporting of Observational studies in Epidemiology

XSL•FO
RenderX

Edited by A Mavragani; submitted 11.09.24; peer-reviewed by A Girma, T Rana; revised version received 20.03.25; accepted 08.04.25; published 13.06.25.

Please cite as:
Zimmermann BM, Willem T, Rost M, Matthes N, Buyx A
Perceptions of Stigma Among Patients With Hepatitis B in Germany: Cross-Sectional Survey
JMIR Form Res 2025;9:e66379
URL: https://formative.jmir.org/2025/1/e66379
doi:10.2196/66379

© Bettina M Zimmermann, Theresa Willem, Michael Rost, Nina Matthes, Alena Buyx. Originally published in JMIR Formative Research (https://formative.jmir.org), 13.6.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Multimodal Noninvasive Assessment of C-Reactive Protein for Systemic Inflammation in Adults: Cross-Sectional Study

Jinjoo Shim[1,2*], PhD; Sinziana Muraru[3*], MD; Vanessa Löw[1], MS; Caroline Evers[3], MD; Sandro Riedo[3], MD; Rucsandra Dobrota[3], MD, PhD; Elgar Fleisch[1,4], PhD; Oliver Distler[3*], MD; Filipe Barata[1*], PhD

[1]Centre for Digital Health Interventions, ETH Zurich, Zurich, Switzerland

[2]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Harvard University, 677 Huntington Avenue, Boston, MA, United States

[3]Department of Rheumatology, University Hospital Zurich, University of Zurich, Zurich, Switzerland

[4]Centre for Digital Health Interventions, University of St Gallen, St Gallen, Switzerland

[*]these authors contributed equally

Corresponding Author:
Jinjoo Shim, PhD
Centre for Digital Health Interventions, ETH Zurich, Zurich, Switzerland

Abstract

Background: Accurate and accessible measurements of inflammatory biomarkers are crucial for the diagnosis and monitoring of inflammatory diseases. The gold-standard C-reactive protein (CRP) requires venipuncture, which, despite providing high-quality samples, can cause discomfort, anxiety, and pain, particularly in vulnerable populations such as older patients. It is also resource-intensive, is unsuitable for remote or at-home use, and lacks continuous monitoring capability. These limitations limit patient autonomy and self-management, potentially leading to poorer prognosis due to delays in assessment and medical treatments. As digital health technologies advance, there is increasing interest in leveraging digital biomarkers for remote and real-time monitoring of systemic inflammation. Digital biomarkers derived from noninvasive biofluids could provide a scalable solution for tracking inflammatory status, offering a patient-centered alternative to traditional blood-based assessments. To date, however, there is no consensus on the most suitable modality for assessment or its digitization potential. Therefore, a comprehensive evaluation of the feasibility, reliability, and patient acceptability toward noninvasive, digital inflammatory biomarkers is needed.

Objective: Our aim is to evaluate the feasibility of various noninvasive methods to assess inflammatory markers and identify the optimal modality for predicting serum CRP levels.

Methods: Inflammatory biomarkers were assessed in 20 participants (10 patients with systemic inflammation defined as a CRP level >5 mg/L and 10 controls) using 6 noninvasive samples (urine, sweat, saliva, exhaled breath, core body temperature, and stool samples) alongside serum samples. Patient preferences were retrieved via a questionnaire. Mann-Whitney *U* test, Spearman correlation, and all-subset regression were conducted to assess the relationships between serum and nonserum biomarkers and identify optimal predictive models for serum CRP levels.

Results: CRP levels were significantly elevated in the inflammation group compared to controls in urine (median 4.5, IQR 4.15 - 10.3 vs median 0.69, IQR 0.24 - 1.39 µg/mmol; P=.001) and saliva (median 4910, IQR 2735-13,275 vs median 473, IQR 309-700 pg/mL; P=.001). Urine and saliva CRP levels strongly correlated with serum CRP (r_{sp}=0.886; P<.001; r_{sp}=0.709; P<.001). The multimodal model using urine and saliva CRP predicted serum CRP levels with 76.1% outperforming single-modality models. Patients favored urine and saliva tests over blood tests.

Conclusions: Urine and saliva represent promising noninvasive alternatives to traditional blood tests for assessing CRP, enabling more accessible and less invasive diagnostic and monitoring approaches.

(JMIR Form Res 2025;9:e77108) doi:10.2196/77108

KEYWORDS

digital biomarker; noninvasive biomarker; C-reactive protein; CRP; systemic inflammation; remote monitoring; chronic inflammatory disease; biosensors; urine; sweat; saliva; exhaled breath; core body temperature; stool; interleukin; IL-1β; IL-6; IL-8; IL-10; tumor necrosis factor; TNF-α

Introduction

The diagnosis, assessment, and ongoing monitoring of systemic inflammatory diseases rely on accurate and readily available measurements of inflammatory biomarkers. Among these, the C-reactive protein (CRP) is extensively used to assess patients' inflammatory state and inform clinical decision-making due to its stability, long half-life, and lack of circadian variability [1].

XSL•FO
RenderX

In rheumatoid arthritis (RA), for example, disease activity and treatment response are assessed based on the levels of CRP (or erythrocyte sedimentation rate) in combination with physical assessment such as inflamed or tender joint counts [2,3].

In current practice, CRP levels are measured in blood samples. While providing high-quality results, venipunctures are associated with distress and anxiety and cause pain and discomfort for patients [4,5]. This issue is particularly pronounced in vulnerable populations, such as children or older people, who report significantly higher levels of distress during venipuncture procedures [6,7]. In individuals with difficult venous access, venipuncture may require multiple attempts, increasing the procedure risks [8].

Blood tests also incur high costs due to the need for trained personnel and specialized laboratory equipment for sample collection and analysis. Furthermore, repeated measurements are impractical in remote settings, limiting patient autonomy and posing substantial challenges in patient-centered care. The lack of timely, accessible monitoring may delay assessments and medical treatments, potentially leading to a poorer prognosis.

At present, consumer-accessible CRP measurement devices remain scarce. Point-of-care devices for the CRP test, priced between US $500 and US $2000 per unit, are designed for health care professionals and still require invasive blood samples. These devices are not only costly and bulky but also have limited sensitivity for CRP assessment, making them impractical for individual use in remote settings. As the global burden of systemic inflammatory diseases continues to rise [9], the development of noninvasive CRP monitoring is essential for improving disease monitoring and patient care.

As digital health technologies advance, there is increasing interest in leveraging digital biomarkers for remote and real-time monitoring of chronic conditions. To date, however, no noninvasive and digital biomarker for the assessment of CRP has been developed. Recent studies have demonstrated that noninvasive measurement of CRP and cytokines in nonblood biofluids such as saliva, urine, and sweat is generally feasible. While CRP has been quantified in urine, most research is focused on urologic conditions [10-13], and its relevance in systemic inflammatory diseases remains largely unexplored. Salivary CRP levels have been correlated with serum CRP in healthy adults and across cardiovascular conditions, chronic obstructive pulmonary disease, or depression [14-17]; however, data on salivary CRP in chronic inflammatory conditions are still limited [18,19]. Wireless wearable biosensors have also shown potential in detecting CRP and inflammatory cytokines in sweat [20,21].

For chronic inflammatory diseases with organ-specific manifestations such as asthma and inflammatory bowel disease, well-established biomarkers such as fractional exhaled nitric oxide (FeNO) and fecal calprotectin are available [22,23], but their relevance in assessing inflammation in other inflammatory diseases has yet to be thoroughly investigated. Moreover, recent studies have used continuous monitoring of core body temperature (CBT) via noninvasive digital sensors to explore circadian rhythmicity and thermoregulation, which might be disrupted in patients with systemic inflammation (SI) [24].

Despite this progress in developing noninvasive methods for the measurement of inflammatory markers, there is no consensus on the most suitable modality for their assessment and for the digitization potential. Existing studies have primarily focused on 1 or 2 biofluids without conducting multimodal measurements and a comprehensive comparison to identify the best-suited noninvasive modality for assessing inflammation. Furthermore, no study has examined patients' preferences for different measurement methods. Evidence suggests that preferences and adherence significantly impact health outcomes, treatment effectiveness, and patient well-being, which may critically impact the real-world applications [25].

Given the challenges associated with blood-based CRP measurement, the growing interest in noninvasive alternatives, and the potential of emerging digital solutions for remote monitoring, there is a need for a comprehensive evaluation of their feasibility, reliability, and patient acceptability.

To address this limitation, our study aims to investigate (1) whether CRP can be detected in noninvasive methods and (2) whether noninvasive CRP correlates significantly with serum CRP levels. To this end, we set out this proof-of-concept study to simultaneously collect and measure inflammatory biomarkers from 6 different noninvasive methods and evaluate their associations with serum biomarkers.

By identifying the optimal noninvasive surrogate markers for serum CRP, our study could guide further efforts in the development of scalable digital health monitoring solutions or digital inflammatory biomarkers, supporting an efficient and patient-centered personalized care.

Methods

Study Population and Recruitment

Our target population included patients with elevated inflammatory parameters (serum CRP >5 mg/L) due to flares across a variety of systemic inflammatory diseases. This approach enhances the generalizability of our findings, ensuring that the noninvasive biomarkers are applicable across different diseases. Patients admitted to the rheumatology ward of the University Hospital Zurich (USZ), Switzerland, for planned or emergency investigations were prescreened by a study physician for having CRP >5 mg/L. If deemed eligible and not needing immediate anti-inflammatory treatment, the patients were informed about our study. The control group was recruited the same way from the rheumatology ward from patients not having increased CRP levels (CRP≤5 mg/L). All participants were given 24 hours or overnight to consider study participation before providing informed consent. Both study visits occurred during the hospital stay, avoiding additional travel burden for the patients. The recruitment procedure is presented in Figure S1 in Multimedia Appendix 1 and further detailed in the study protocol [26].

We included 10 adult patients with SI and 10 control participants. Patients fulfilled the classification criteria of the

XSL•FO
RenderX

respective inflammatory disease. We excluded patients requiring initiation of a new treatment with glucocorticoids or immunosuppressive medication, patients diagnosed with active psychiatric disorders, or patients having language barriers or other factors impairing adherence to the study protocol. Moreover, samples with signs of urinary tract infection were excluded from the analysis to reduce the risk of false positive results related to local rather than SI and ensure the reliability of urinary CRP as a potential biomarker in this limited sample size. Detailed exclusion criteria are presented in Figure S1 in Multimedia Appendix 1.

Sample Procedures, Sample Collection, and Measurement

Figure S2 in Multimedia Appendix 1 describes the study procedure. Participants were asked to fast for at least 2 hours prior to the study visit and avoid drinking other liquids than water, using chewing gums, brushing teeth, mouthwash, or smoking.

At the first visit, we applied a questionnaire assessing demographic information and history regarding oral health as well as recent surgical procedures and hospitalizations (Methods S1 in Multimedia Appendix 1). Venous blood samples for biomarker analysis were collected (1 heparin gel tube for CRP and 1 native tube for the immunological parameters), followed by the collection of midstream urine samples. A dipstick test (ie, Roche Combur Test Strip) to screen for signs of urinary tract infection was performed immediately after urine collection. Saliva samples were obtained using an absorbent swab (Salimetrics SalivaBio Oral Swab), placed under the tongue for 3 minutes, and disposed directly into a collection tube. Two sweat patches (ie, PharmChek Sweat Patch) [21] were applied to participants' arms or abdomen for sweat collection.

A CBT sensor (greenTEG CaleraResearch CORE) was then placed on the upper body, approximately 20 cm below the armpit using a medical patch. Calera Research CORE is a heat flux-based thermometer calculating body temperature based on a clinically validated algorithm using built-in heat flux and skin temperature sensors [27]. We focused on nighttime hours to mitigate the potential impact of external factors such as participation in rehabilitation programs or physical activity on patients' core temperature rhythm during daytime.

FeNO measurements were conducted using the Bosch Vivatmo me device [28]. Finally, we provided instructions for stool sampling. To minimize potential confounding from the diurnal variation of CRP levels, all first visits were conducted within a restricted time window between 11 AM and 3 PM.

At the second visit after 48 hours, the researcher collected the CBT sensor, sweat patches, and stool sample. Participants were asked to rank their preference for sampling methods, including a blood test and 6 noninvasive tests, using an ordinal ranking method. They ranked each test on a scale from 1=most preferred to 7=least preferred. Questionnaires were applied in German or English as preferred by patients (Methods S2 and S3 in Multimedia Appendix 1 for English version; the German version is available upon request).

A part of the collected urine (10‑20 mL) was retained for analysis in the external laboratory and stored at −80 °C until analysis. The remaining urine as well as the blood samples were centrifuged at room temperature at 3710 rpm for 10 minutes (5 minutes for urine), aliquoted into cryovials, and stored at −80 °C until analysis. Stool, saliva, and sweat patches were frozen within 1 hour from collection and stored at −80 °C until analysis. In the external laboratory, the urine samples were thawed and centrifuged for 15 minutes at 1000× *g* at 2‑8 °C for preparation. Clinical data were gathered from electronic medical records, including anthropometric measurements (eg, weight, height, and BMI), laboratory results (eg, erythrocyte sedimentation rate), disease characteristics (eg, diagnosis of inflammatory disease, disease duration, and comorbidities), medication (eg, type, duration, and dose), and relevant measures of disease activity. Urine CRP was analyzed both as an absolute concentration and as a creatinine-normalized ratio to account for urine dilution. Specifically, we standardized urine CRP measurements according to urine creatinine levels, following established protocol [29]. Comorbidities were retrieved from the electronic records.

Immunoassays

Analysis of serum CRP, serum and urine total protein, albumin, creatinine, and stool calprotectin was conducted at the Institute of Clinical Chemistry of the USZ. Saliva, urine, and sweat CRP and cytokine analyses were performed at Swiss BioQuant AG. Serum cytokines were analyzed at the Department of Immunology of the USZ. Biomarkers with a detection rate below 70% were excluded from the analysis to ensure the reliability of the findings.

Serum CRP levels were assessed using an immunoturbidimetry assay conducted on a Cobas 8000 c702 system (Roche Diagnostics). Saliva, urine, and sweat CRP were determined via enzyme-linked immunosorbent assay (Salimetrics, Abcam, and Creative Diagnostics), following manufacturer protocols [30-32]. Readings for saliva, urine, and sweat CPR analyses were obtained with a SpectraMax i3 multimode microplate spectrophotometer (Molecular Devices), with a wavelength set to 450 nm.

Serum and urine concentrations of the total amount of protein, albumin, and creatinine were quantified with the immunoturbidimetry method (Cobas 8000). Stool calprotectin levels were determined using chemiluminescence immunoassay (Liaison XL, DiaSorin).

Serum levels of interleukin (IL)-1β, IL-6, and tumor necrosis factor (TNF)-α were quantified using a Luminex Magpix analyzer (Luminex Corporation). IL-8 in serum was analyzed using an enzyme-linked immunosorbent assay (R&D Systems). IL-10 in serum was assessed using a cytometric bead array (BD Biosciences). IL-6, IL-8, IL-10, IL-1β, and TNF-α in sweat and urine samples were analyzed using MSD V-PLEX plates and read with QuickPlex SQ 120 electrochemiluminescence reader (MSD, Meso Scale Discovery).

Statistical Analysis

In the descriptive analysis, we presented continuous variables with median and IQR and categorical variables with count and

XSL•FO
RenderX

percentage. We used the Mann-Whitney *U* test and the chi-square test (or Fisher exact test for expected counts<5) for univariable analysis. We tested the normality of distribution using the Shapiro-Wilk test. Comparisons of end points between SI and control groups were visualized using boxplots and assessed with the Mann-Whitney *U* test. Spearman correlation coefficient (r_{sp}) evaluated correlations between serum and nonserum biomarkers as well as between biomarkers and clinical variables.

We further assessed the predictive accuracy of multimodal, noninvasive biomarkers using all-subset regressions to identify the best model for predicting the serum CRP level among all possible combinations. This method systematically evaluates all possible predictor combinations for each full model and selects the fitted model with the optimal criterion value among candidate models based on the adjusted R^2 value. R^2 represents how much of the variation in the dependent variable of the model is explained by variation in its predictors [33]. An R^2 of 0 indicates the prediction is impossible based on input variables, whereas an R^2 of 1 represents a perfect prediction with no unexplained variability. Generally, an R^2 above 0.6 indicates an algorithm to be useful, and an R^2 above 0.7 would be considered a substantial predictive ability. All statistical tests were 2-sided, and a significance level of .05 was considered. Statistical analyses were performed using R (version 4.3.1; R Foundation for Statistical Computing). The study design and reporting followed the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) statement [34].

Sample Size Calculation

Because there is no prior evidence of the outcome measures in this context, we calculated the sample size based on studies of adult patients diagnosed with RA, which is one of the most prevalent systemic inflammatory diseases. To ensure the detection of even small differences in CRP within the lower range of the inflammatory spectrum, we adopted a conservative approach and based our sample size calculation on a population of patients with RA having low disease activity. For instance, Bechman et al [35] indicated serum CRP levels in patients with RA during flares ranging from 5 to 9 (SD 3.0) mg/L. Considering these results, it is plausible to expect a difference of 3 to 6 mg/L in serum CRP levels between SI versus non-SI states among patients with RA. On this basis, assuming 80% of power, 5% of type I error, and equal variance in groups, we determined that a sample size of 10 participants per group is required to detect a 4 mg/L difference in serum CRP levels with an SD of 3.0. Therefore, we aimed to complete data collection for a total of 20 participants.

Ethical Considerations

The study was conducted in accordance with the Declaration of Helsinki, the principles of Good Clinical Practice, the Human Research Act, and the Human Research Ordinance [36,37] and after receiving approval from the independent ethics committee of the Canton of Zurich (BASEC-Nr: 2022 - 01386) on October 28, 2022. Written patient information and consent documents were available in both English and German. All participants provided written informed consent before being included in the study. In addition to the primary data analysis, the consent included a clause allowing for the encrypted data and biological material collected during this project to be stored and reused for future research. They were also informed of their right to withdraw consent at any time without providing justification. All project data were stored and handled with strict confidentiality. At inclusion, participants were assigned a study-specific unique identifier to be used on case report forms and further data analysis. Clinical data were stored on a password-protected server at the USZ, accessible only through the unique identifier provided by the USZ research team. The participant identification list is securely stored at USZ and will be destroyed 10 years after project completion. FeNO and questionnaire data containing the unique identifier were saved on a password-protected network drive at ETH Zurich, with access limited to authorized personnel and monitored through access logs. No images or visual materials included in this paper contain identifiable participant information. Participants were compensated with a voucher for 30 CHF (US $37) for their involvement in the study.

Results

Characteristics of Patients With SI and Controls

In total, 20 participants were enrolled in the study, 1 of whom was excluded due to a urinary tract infection, resulting in 19 participants included in the analysis. Of these 19, a total of 4 had urine CRP concentrations below the assay's lower limit of quantification.

The baseline characteristics of 19 participants are presented in Table 1. Demographic and baseline characteristics were similar between SI and controls. The levels of creatinine, albumin, and total protein in serum and urine were also comparable.

XSL•FO
RenderX

Table . Baseline demographic, clinical, and laboratory characteristics of the 19 study participants.

Variables	SI[a] (n=10)	Control (n=9)	*P* value
Age at recruitment (years), median (IQR)	56 (38.0-63.0)	57 (47.0-59.0)	.68
Gender, n (%)			>.99
Women	7 (70)	6 (67)	
Men	3 (30)	3 (33)	
Employment status, n (%)			.72
Full-time	5 (50)	3 (33)	
Part-time	2 (20)	3 (33)	
Else	3 (30)	3 (33)	
Education, n (%)			.37
College or above	4 (40)	1 (11)	
Else	6 (60)	8 (89)	
BMI category, n (%)			.93
Underweight or normal (<25 kg/m^2)	2 (20)	2 (22)	
Overweight (25 - 29.9 kg/m^2)	3 (30)	2 (22)	
Obese (≥30 kg/m^2)	5 (50)	5 (56)	
Smoking, n (%)			>.99
Previous or current smoker	2 (20)	2 (22)	
No	8 (80)	7 (78)	
Comorbidities, n (%)			
Arterial hypertension	4 (40)	4 (44)	>.99
Chronic pain syndrome	4 (40)	8 (89)	.08
Depression	2 (20)	2 (22)	>.99
Prior history, n (%)			
Had hospitalization within 1 year	2 (20)	0 (0)	.50
Had operation within 1 year	2 (20)	0 (0)	.50
Oral health, n (%)			
Fasted more than 2 hours	4 (40)	7 (78)	.23
No brushing teeth within 2 hours	10 (100)	9 (100)	N/A[b]
No chewing gum within 2 hours	10 (100)	9 (100)	N/A
No smoking within 2 hours	8 (80)	9 (100)	.50
No drinking within 24 hours	9 (90)	9 (100)	>.99
No bleeding gums within 24 hours	10 (100)	9 (100)	N/A
No current oral diseases	7 (70)	8 (89)	.66
Laboratory parameters, median (IQR)			
Serum creatinine (mmol/L)	81.0 (74.3-96.3)	75.0 (67.0-89.0)	.46
Serum albumin (g/L)	39.5 (38.0-40.0)	44.0 (41.0-44.0)	.003
Serum total protein (g/L)	70.5 (67.3-74.0)	72.0 (69.0-74.0)	.62
Urine creatinine (mmol/L)	4.3 (3.4-5.5)	5.9 (4.5-11.4)	.12

XSL•FO
RenderX

Variables	SI[a] (n=10)	Control (n=9)	*P* value
Urine albumin/creatinine ratio (mg/mmol)	0.84 (0.58-1.53)	1.02 (0.67-2.45)	.41
Urine total protein/creatinine ratio (g/mmol)	10.74 (8.59-14.73)	8.91 (7.17-14.08)	.81

[a]SI: systemic inflammation.

[b]N/A: not applicable.

Patients with SI had various diagnoses, ranging from arthritis to adult-onset Still disease and giant cell arteritis (Table 2).

The median age at the onset of the inflammatory disease and disease duration were 55.5 (IQR 32.3-60.5) and 1.25 (IQR 0.5-5.3) years, respectively.

Control group patients were inpatients who attended a program for chronic musculoskeletal pain, including chronic back pain and fibromyalgia. Among them, 6 had no inflammatory disease. The most common comorbidity, like the SI group, was obesity. In total, 3 patients had an inflammatory disease in remission under treatment. Detailed diagnoses and treatment information for these 3 patients are provided in Table S5 in Multimedia Appendix 1.

Table . Inflammatory disease diagnosis, duration, and main anti-inflammatory medications at the time of inclusion for participants with systemic inflammation.

Inflammatory disease	Disease duration (years or diagnosed at inclusion)	Main medication for inflammatory diagnosis at inclusion
Mixed crystal-induced arthropathy (CPPD[a] and gout)	6	Prednisone
Transient polyarthritis	Diagnosed at inclusion	No DMARDs[b]
Seropositive erosive rheumatoid arthritis	20	Prednisone
Inflammatory syndrome likely linked to multiple sclerosis	1.5	Ocrelizumab
VEXAS[c] syndrome	0.5	Ruxolitinib, prednisone
Pigmented villonodular synovitis and metabolic syndrome	0.5	No DMARDs
Seronegative polyarthritis	3	NSAID[d]; no DMARDs
Adult-onset Still disease	0.3	NSAID; no DMARDs
Peripheral spondyloarthritis	9	Certolizumab pegol
Giant cell arteritis	1	Abatacept, methotrexate, prednisone

[a]CPPD: calcium pyrophosphate crystal deposition disease.

[b]DMARD: disease-modifying antirheumatic drug.

[c]VEXAS: vacuoles, E1 enzyme, X-linked, autoinflammatory, somatic.

[d]NSAID: nonsteroidal anti-inflammatory drug.

Analysis of the Primary End Point

The median serum CRP level in patients with SI and controls was 18 (IQR 9.3-68.8) and 1.3 (IQR 1.2-1.8) mg/L, respectively ($P<.001$). CRP detection rates were 100% (19/19) for serum and saliva, and 79% (15/19) for urine. Four urine samples (1 among the SI group and 3 controls) had CRP levels below the lower limit of quantification for urine (31.3 pg/mL), whereas CRP levels for these individuals were detectable in both serum and saliva (Table S4 in Multimedia Appendix 1). The lower detection rate in urine may be attributed to its higher limit of quantification (31.3 pg/mL), compared to 0.6 mg/L for serum and 25.0 pg/mL for saliva, indicating greater sensitivity in the latter 2 biofluids. CRP could not be detected in sweat samples.

CRP levels were significantly higher in the serum, urine, and saliva of patients with SI compared to controls (Figure 1). The median creatinine-adjusted urine CRP was significantly higher in patients with SI compared to controls: 4.5 (IQR 4.15-10.3) versus 0.69 (IQR 0.24-1.39) µg/mmol, respectively ($P=.001$). The absolute values are presented in Table S1 in Multimedia Appendix 1.

XSL•FO
RenderX

Figure 1. Distributions of CRP levels in serum and noninvasive biofluids among participants with SI and controls. (A) s-CRP, (B) u-CRP, (C) u-CRP/u-Cr, and (D) sa-CRP. Boxes indicate the IQR, and the horizontal line within each box represents the median, and whiskers denote 1.5×IQR. Dots represent outliers. The *P* values are based on the Wilcoxon test. CRP: C-reactive protein; s-CRP: serum C-reactive protein; sa-CRP: saliva C-reactive protein; SI: systemic inflammation; u-CRP: urine C-reactive protein; u-CRP/u-Cr: creatinine-normalized urine C-reactive protein.

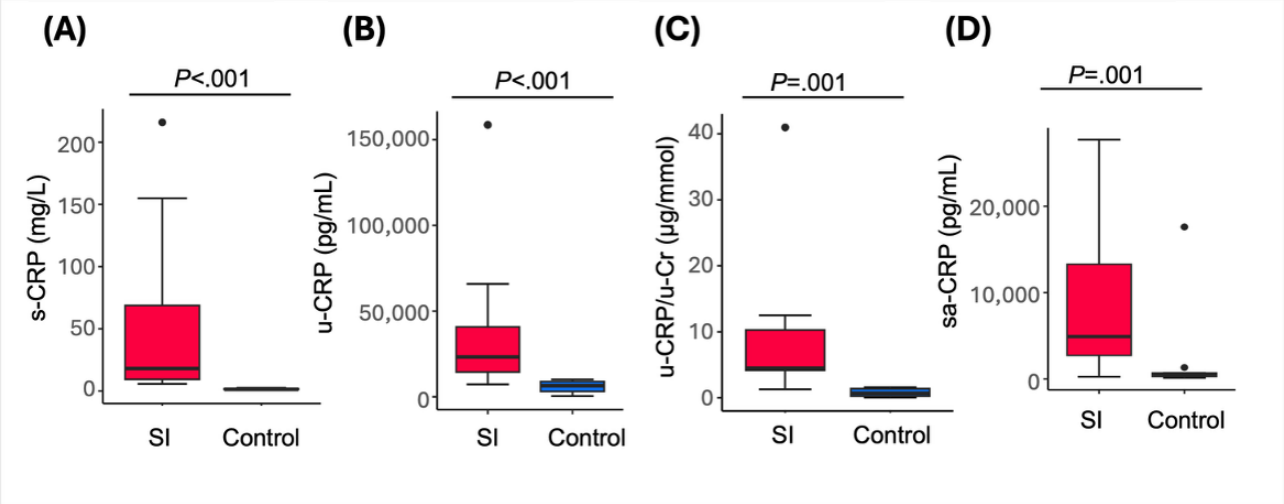

Similarly, median saliva CRP levels were substantially elevated in patients with SI compared to controls: 4910 (IQR 2735-13,275) versus 473 (IQR 309-700) pg/mL, respectively (*P*=.001).

We assessed the correlations between serum and nonserum CRP levels (Figure 2). Creatinine-normalized urine CRP showed the strongest correlation with serum CRP (r_{sp}=0.886; *P*<.001), followed by salivary CRP (r_{sp}=0.709; *P*<.001). These nonserum CRP values effectively discriminated between patients with SI and controls. Therefore, noninvasive CRP measurements accurately identified active inflammation irrespective of the underlying diagnosis.

Figure 2. Correlations between s-CRP and non–s-CRP levels. (A) s-CRP and u-CRP, (B) s-CRP and u-CRP/u-Cr, and (C) s-CRP and sa-CRP. Each top panel displays a scatterplot of log-transformed values with a linear regression line (red) and 95% confidence band (gray). Bottom panels show individual data points stratified by group: participants with SI (red squares) and controls (blue circles). CRP: C-reactive protein; r_{sp}: Spearman correlation coefficient; s-CRP: serum C-reactive protein; sa-CRP: saliva C-reactive protein; SI: systemic inflammation; u-CRP: urine C-reactive protein; u-CRP/u-Cr: creatinine-normalized urine C-reactivproteinn..

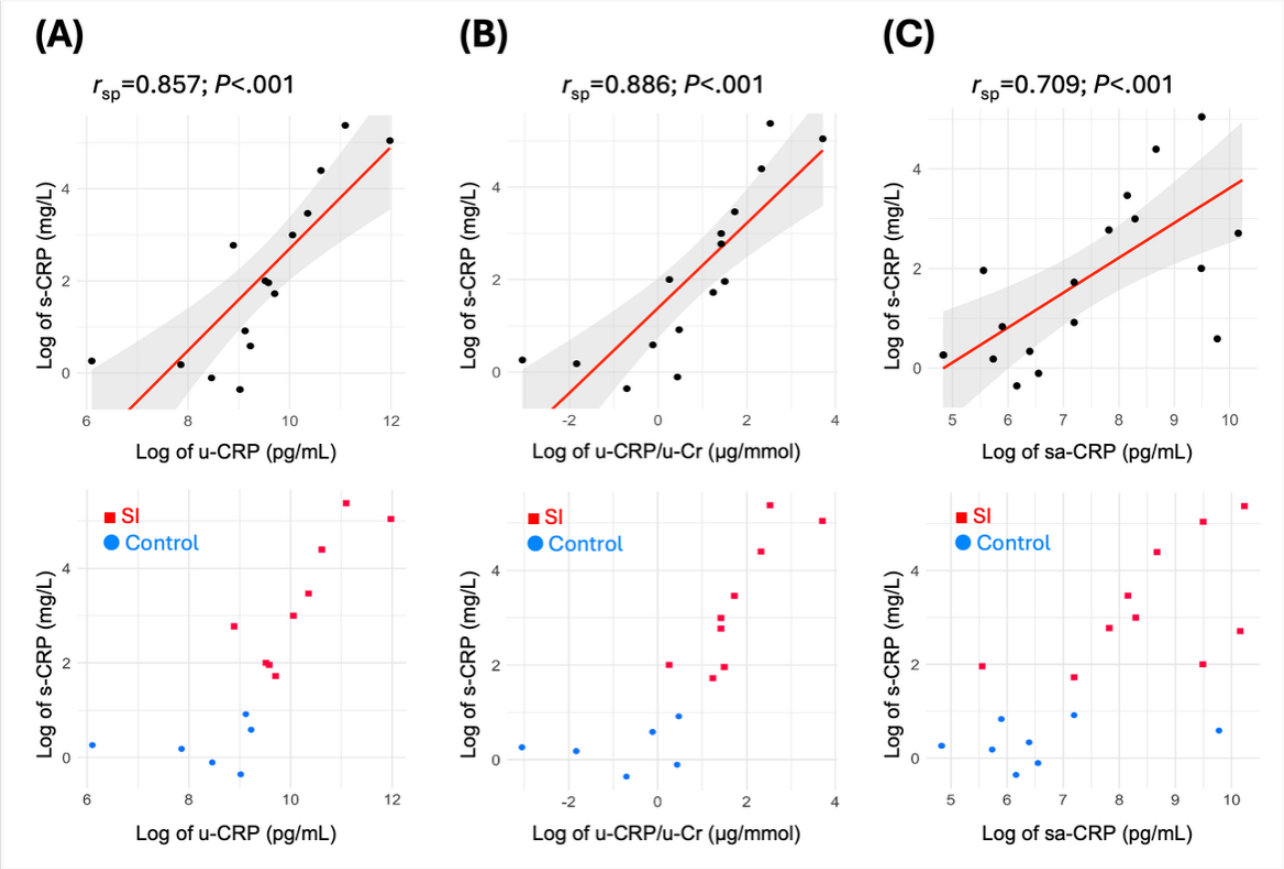

XSL•FO
RenderX

Analysis of the Secondary and Tertiary End Points

All cytokines were detected in serum samples. IL-8 was detected in 74% (14/19) of sweat samples and 100% (19/19) of urine samples. IL-6 was detected in 42% (8/19) of sweat samples and 74% (14/19) of the urine samples. IL-10, IL-1β, and TNF-α were detected in less than 50% of the sweat and urine samples.

The median serum IL-6 level in patients with SI and controls was 2.58 (IQR 1-10.1) and 1 (IQR 1-1) pg/mL, respectively (*P*=.02). The median serum TNF-α level in patients with SI and controls was 8.85 (IQR 6.72-9.44) and 5.83 (IQR 3.4-7.17) pg/mL, respectively (*P*=.02). Serum IL-6 and TNF-α demonstrate a moderate correlation with serum CRP (r_{sp}=0.68; *P*=.001 and r_{sp}=0.64; *P*=.003, respectively).

In the analysis of inflammatory cytokines, we observed significant differences in serum levels of IL-6 and TNF-α between groups (Figure 3). IL-1β, IL-10, and IL-8 levels were detectable in serum samples but did not show a statistically significant difference between groups.

Figure 3. Distributions of inflammatory cytokine levels in serum among participants with SI and controls. Biomarkers with a detectability rate above 70% are shown. Boxes indicate the IQR, and the horizontal line within each box represents the median, and whiskers denote 1.5×IQR. Dots represent outliers. The *P* values are based on the Wilcoxon test. IL-1β: interleukin-1 beta; IL-6: interleukin-6; IL-8: interleukin-8; SI: systemic inflammation; TNF-α: tumor necrosis factor-alpha.

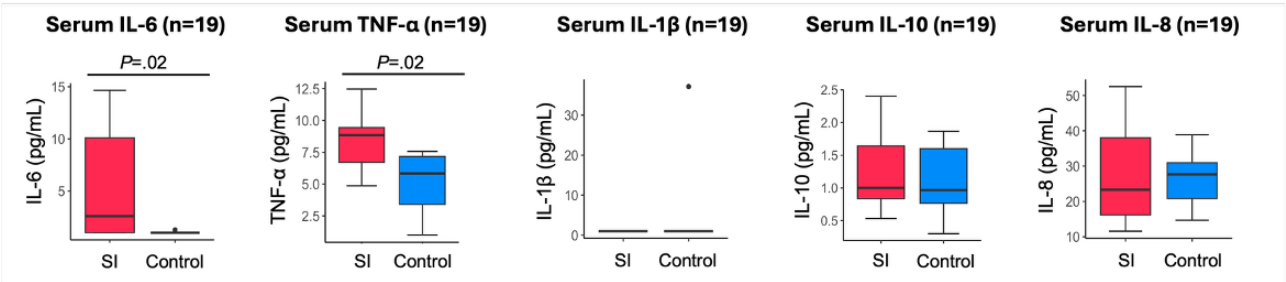

In noninvasive methods, sweat IL-8, urine IL-8, and urine IL-6 showed a sufficient detectability rate; however, they did not exhibit significant differences between groups (Figure S3 in Multimedia Appendix 1).

We observed a significant difference in the FeNO levels comparing SI and controls (21 vs 11 ppb; Table S2 in Multimedia Appendix 1). However, due to the low detection rate of FeNO (8/19, 42%), the sample size was small. Moreover, we found that 3 SI patients with measurable FeNO levels had a history of respiratory inflammation, including allergic asthma and viral airway infection. Thus, we assumed that the increased FeNO levels were more likely associated with the local inflammation rather than systemic disease.

No significant difference was observed in fecal calprotectin levels between groups. We analyzed continuous, longitudinal CBT data from 13 participants (6 SI patients and 7 controls; Figure S4 in Multimedia Appendix 1). The minimum (nadir), maximum (peak), mean, and SD of the core temperature values during night hours between 10 PM and 7 AM the next day are summarized in Table 3. Overall, we did not find significant differences in CBT characteristics between patients with SI and controls. When inspecting individual cases, a notable temperature pattern emerged in the patient with adult-onset Still disease, characterized by a significant increase in nocturnal CBT (Figure S4A in Multimedia Appendix 1). This is in line with the literature describing the highest temperature spikes in the evening in these patients [38].

XSL•FO
RenderX

Table . Core body temperature (CBT) metrics in participants with systemic inflammation (SI) and controls[a].

Diagnosis		Nadir CBT (°C)	Peak CBT (°C)	Mean CBT (SD) (°C)
Patients with SI				
1	Adult-onset Still disease	36.4	38.5	37.3 (0.708)
2	VEXAS[b] syndrome	36.4	37.0	36.6 (0.151)
3	Peripheral SpA[c]	36.7	37.2	36.9 (0.102)
4	Undifferentiated seronegative polyarthritis	36.3	36.8	36.6 (0.117)
5	Giant cell arteritis	36.3	36.8	36.5 (0.114)
6	Multiple sclerosis	36.8	37.1	36.9 (0.062)
	SI group average	36.5	37.2	36.8 (0.209)
Controls				
7	No inflammatory disease	36.7	37.5	37.0 (0.228)
8	No inflammatory disease	36.4	37.2	36.6 (0.148)
9	RA[d] in remission	36.3	36.9	36.5 (0.127)
10	No inflammatory disease	36.4	36.8	36.6 (0.094)
11	No inflammatory disease	36.8	37.3	37.0 (0.094)
12	No inflammatory disease	36.6	37.1	36.8 (0.133)
13	SpA in remission	36.3	36.8	36.6 (0.163)
	Control group average	36.5	37.1	36.7 (0.141)

[a]For each participant, minimum (nadir), maximum (peak), and mean (SD) of CBT were calculated based on the temperature measurements collected between 10 PM and 7 AM the following day.

[b]VEXAS: vacuoles, E1 enzyme, X-linked, autoinflammatory, somatic.

[c]SpA*:* spondyloarthritis.

[d]RA*:* rheumatoid arthritis.

Predictive Accuracy of Multimodal, Noninvasive Inflammatory Markers

To assess the predictive accuracy of the noninvasive methods, we identified the optimal subset of multimodal, noninvasive biomarkers for predicting serum CRP levels using the all-subset regression method (Table 4). We considered urine CRP, creatinine-normalized urine CRP, and saliva CRP measurements due to their robust correlations with serum CRP. To ensure analytical validity and consistency, models were developed using data from 15 participants with complete urine and saliva measurements, excluding 4 urine samples below the assay's lower limit of quantification. Among the possible covariates, the best subset of the model achieved an adjusted R^2 value of 0.761, indicating that 76.1% of the variance in the serum CRP is collectively explained by salivary CRP and absolute urine CRP, surpassing the performance of single modality approaches. Serum CRP and creatinine-normalized urine CRP yielded an adjusted R^2 value of 0.747. When urine CRP, creatinine-normalized urine CRP, and saliva CRP were modeled separately, the R^2 value ranged from 0.52 to 0.61.

XSL•FO
RenderX

Table . All-subset regression models evaluating combinations of multimodal, noninvasive inflammatory biomarkers for predicting serum C-reactive protein (CRP) levels[a].

Regression model		Adjusted R^2
Model 1	Serum CRP~ sa-CRP[b]	0.518
Model 2	Serum CRP~u-CRP[c]	0.608
Model 3	Serum CRP~u-CRP/u-Cr[d]	0.513
Model 4[e]	Serum CRP~saCRP+u-CRP	0.761
Model 5	Serum CRP~saCRP+u-CRP/u-Cr	0.747
Model 6	Serum CRP~uCRP+u-CRP/u-Cr	0.691
Model 7	Serum CRP~saCRP+uCRP+u-CRP/u-Cr	0.745

[a]Each model includes noninvasive markers as predictors: sa-CRP, u-CRP, and u-CRP/u-Cr.

[b]sa-CRP: saliva C-reactive protein.

[c]u-CRP: urine C-reactive protein.

[d]u-CRP/uCr: creatinine-normalized urine C-reactive protein.

[e]Model 4, which includes both sa-CRP and u-CRP, demonstrated the highest explanatory power.

Participant-level analysis revealed that this performance enhancement can be attributed to the complementary roles of saliva and urine, particularly in cases where 2 noninvasive methods display discrepancies in predicting accuracy (Table S3 in Multimedia Appendix 1). In the case of a patient with giant cell arteritis, for example, saliva CRP was more accurate than urine CRP in predicting serum CRP levels. In contrast, a patient with seropositive erosive RA showed that urine CRP levels predicted serum CRP levels more accurately than saliva. The multimodal approach was also effective in handling outliers, such as a control patient with adult-onset Still disease in remission, who exhibited markedly high saliva CRP values despite low serum and urine CRP levels. Integration of urine and saliva CRP measurements substantially mitigated discrepancies and significantly improved prediction accuracy, suggesting a potential for practical and noninvasive solutions for CRP detection.

Patient Preference

We obtained participants' preferences for various biospecimen sampling methods, ranked from most to least preferred. Among all participants (N=19), the sweat patch and urine test were rated as the most acceptable methods, with mean scores of 2.37 (SD 1.30) and 2.42 (SD 1.54), respectively. Preferences were consistent across groups. For the sweat patch, the mean scores were identical in the SI and control groups at 2.3 (SD 1.34 and 1.33, respectively); for the urine test, mean ratings were slightly higher in the SI group at 3.0 (SD 1.76) compared to 1.78 (SD 0.97) in controls.

Moderately preferred methods included FeNO test, saliva test, and CBT sensor, which yielded mean scores of 3.63 (SD 1.86), 3.68 (SD 1.25), and 4.08 (SD 1.38), respectively, across all participants. The FeNO test was rated similarly by both groups, with means of 3.9 (SD 1.52) in the SI group and 3.3 (SD 2.24) in controls. Saliva sampling received mean scores of 3.3 (SD 1.16) in the SI group and 4.11 (SD 1.27) among controls. CBT sensor was rated with means of 3.86 (SD 1.57) in the SI group and 4.33 (SD 1.21) in controls.

The least preferred methods were blood and stool tests. The blood test received a mean score of 4.47 (SD 2.04) overall, with the SI group rating it at 4.2 (SD 2.35) and the control group slightly higher at 4.78 (SD 1.72). The stool test was rated lowest across all sampling methods, with a mean of 5.84 (SD 1.46).

Discussion

Principal Findings

This is the first study to assess the accuracy of a multimodal, noninvasive assessment of inflammatory biomarkers compared to the gold-standard serum CRP in a cohort with and without SI. Our main findings revealed that CRP levels in urine and saliva were significantly elevated in patients with SI and exhibited strong correlations with serum CRP regardless of the underlying diagnosis. Notably, the combination of urine and saliva CRP achieved high predictive accuracy for serum CRP with improvements of 15% - 24% over single-modality models. The multimodal approach effectively addressed outliers and discrepancies occurring in the single modalities. The measurements in urine and saliva demonstrated high detection rates and were strongly favored over blood tests by participants. These results highlight urine and saliva as promising alternative biofluids for CRP assessments, with significant implications for digitization.

Establishing these relationships would provide the basis for digitizing measurement for serum biomarkers, such as for CRP, contributing to the creation of a digital surrogate biomarker. Once digitized, this approach could lead to the development of a scalable, cost-efficient sensor, allowing a timely and rapid monitoring of inflammatory activity. This advancement is particularly relevant, given the ever-rising incidence of systemic inflammatory conditions and the urgent need for precise and remote monitoring tools.

Comparison With Prior Research

Previous literature regarding noninvasive detection and digitization of inflammatory biomarkers is scarce. Few studies

XSL•FO
RenderX

showed that CRP might be detected in urine; however, they were mostly concerning urologic conditions and did not use the urine/creatinine ratio to overcome urine dilution variability [10,39]. An additional study suggested that CRP excretion in urine is dependent on the levels of proteinuria [40]. In our study, we found that patients in the SI and control groups had similar levels of proteinuria and albuminuria.

The first evidence regarding the detection of CRP in saliva, reported in veterinary medicine, demonstrated a correlation between serum and salivary CRP in dogs comparable to our findings. In humans, a wide range of correlation coefficients between serum and salivary CRP levels across various conditions was reported [17]. This variability may stem from several challenges, with the most relevant being the lack of standardized collection techniques, the impact of local microtrauma, variations in salivary flow rates, and dilution levels. We aimed to mitigate these issues in our study by implementing the same technique of collection and requirements regarding eating and drinking as well as collecting information about known gingival inflammatory diseases.

Strengths and Limitations

A key strength of this study lies in the innovative approach of a multimodal testing of noninvasive measurements in the context of SI irrespective of diagnosis. Moreover, the methods proposed are straightforward and easily adaptable for further testing and validation, ensuring their potential for broader application, integration into future research, clinical practice, and digitization for remote monitoring. Additionally, it prioritizes patient-centered care by incorporating patient preferences, making the findings highly relevant to clinical practice.

However, the generalizability of our findings is limited by the small size of the cohort and the lack of longitudinal data. Another limitation refers to the detection of inflammatory markers in sweat. Finally, our recruitment and study procedure were based on a single university hospital located in Zurich, Switzerland. These limitations do not compromise the interpretability of the main results regarding a high detection rate and good correlation of urine and saliva measurements with serum CRP.

Suggestions for Future Sensor Development

To achieve the digitization of CRP and efficient sensor development, we highlight several key considerations based on the findings of our study. First, sensor accuracy and reliability must be ensured, considering the extremely low concentration of CRP and its detection rate in biofluids at the picomolar level. While biofluids such as urine, saliva, and sweat contain a variety of biochemical molecules and clinically relevant biomarkers, the concentration of these markers is generally much lower in noninvasive biofluids than in blood. Due to the nature of the feasibility study, we used immunoassay analyses tailored to each biofluid, following methodologies established in previous studies. For the sensor to be scalable, developing cost-efficient yet sensitive analytic methods can be a major challenge.

Second, we need to consider interindividual variability and diurnal fluctuations of noninvasive biofluids [17,41]. The excretion of these biofluids is subject to influences from interpersonal or environmental factors. In our study, we used standardized protocols, requiring participants to abstain from food and substances prior to sample collection to minimize these variables. We also collected comprehensive demographic and health data through questionnaires. For future sensor developments, the collection of this contextual information is crucial, and its digitization could improve efficiency in real-world applications.

Third, patient preference and acceptability play a pivotal role in the practical implementation. In our qualitative results, participants ranked urine, saliva, and sweat as the top 3 most preferred methods due to their accessibility and the ability to perform stress-free, repeated sampling. This finding suggests that these methods could be successfully implemented with high acceptability in real-world settings, as they reduce the time and pain associated with traditional blood-based testing. Furthermore, participants did not find multimodal biofluid collection burdensome but easy to perform with minimal preparation required for testing, which aligns with a patient-centered approach that promotes autonomy.

Finally, the different biofluids may require different implementation strategies. In our study, the detection of inflammatory markers in sweat (passive sweat collected using sweat patches) was particularly challenging. Several factors may have contributed to this, including a short exposure window (48 hours as opposed to 72 hours in previous studies) and the relatively older age of study participants may have led to insufficient passive sweat excretion [42]. Additionally, the prolonged storage of sweat patches due to batch analysis design could have resulted in protein degradation. A recent study has demonstrated the real-time, noninvasive CRP monitoring using a wireless sweat sensor in both healthy individuals and patients with heart failure in a laboratory setting [20]. While our study did not yield similar results, the high preference for sweat collection among participants, easier accessibility, and reduced risk of contamination suggest that it still holds potential for the real-time assessment [43].

Conclusions

Our feasibility study demonstrates the potential of urine and saliva CRP as surrogate markers for SI and holds significant digitization potential. Beyond their immediate clinical relevance, this discovery underscores the potential of these biofluids as valuable targets for developing novel sensor technologies, which paves the way toward scalable health monitoring solutions and personalized patient-centered care. Digital biomarkers could enhance health equity and accessibility, especially in rural areas or underserved communities, thus reducing treatment gaps and fostering inclusive patient care. Future studies focusing on longitudinal follow-up and populations with systemic inflammatory disorders and comorbidities will be crucial for validating and generalizing our findings.

XSL•FO
RenderX

Acknowledgments

The authors would like to thank Lucia Fleckenstein and Sylvie Schiltz for their contribution to participant recruitment and data collection. The authors are also grateful to the participants who joined the study.

Authors' Contributions

JS, SM, EF, OD, and FB conceptualized the study design and developed the study methodology. JS, SM, VL, CE, and SR conducted participant recruitment, screening, data collection, and patient interviews. JS and SM performed the data analyses. JS, SM, VL, EF, OD, and FB contributed to the result interpretation and discussion. EF and OD provided resources for the study's design and execution. JS and SM drafted the original manuscript. All authors reviewed and revised the final version of the manuscript.

JS and SM share the first authorship. FB and OD share the last authorship.

Conflicts of Interest

OD has or had a consultancy relationship with or has received research funding from or has served as a speaker for the following companies: 4P-Pharma, AbbVie, Acceleron, Alcimed, Altavant, Amgen, AnaMar, Arxx, AstraZeneca, Blade, Bayer, Boehringer Ingelheim, Corbus, CSL Behring, Galderma, Galapagos, Glenmark, Gossamer, Horizon, Janssen, Kymera, Lupin, Medscape, Merck, Miltenyi Biotec, Mitsubishi Tanabe, Novartis, Prometheus, Redxpharma, Roivant, and Topadur. Patent issued "mir-29 for the treatment of systemic sclerosis" (US8247389 and EP2331143). RD has served as a speakers bureau for Actelion, an advisory board for Boehringer-Ingelheim, and has received grants or research support from Pfizer, Actelion, and Iten-Kohaut Foundation, and congress participation support from Amgen. SM has received congress participation support from AstraZeneca. JS was affiliated with ETH Zurich at the time of the study and is currently affiliated with Harvard University. SR was affiliated with the University Hospital of Zurich at the time of the study and is currently affiliated with Schulthess Klinik Zürich. All other authors have no conflicts of interest to declare.

Multimedia Appendix 1
Participant eligibility, study procedure, survey questionnaire, and additional analyses.
[DOCX File, 8893 KB - formative_v9i1e77108_app1.docx]

References

1. Gabay C, Kushner I. Acute-phase proteins and other systemic responses to inflammation. N Engl J Med 1999 Feb 11;340(6):448-454. [doi: 10.1056/NEJM199902113400607] [Medline: 9971870]
2. Orr CK, Najm A, Young F, et al. The utility and limitations of CRP, ESR and DAS28-CRP in appraising disease activity in rheumatoid arthritis. Front Med (Lausanne) 2018;5:185. [doi: 10.3389/fmed.2018.00185] [Medline: 30123796]
3. Smolen JS, Landewé RBM, Bergstra SA, et al. EULAR recommendations for the management of rheumatoid arthritis with synthetic and biological disease-modifying antirheumatic drugs: 2022 update. Ann Rheum Dis 2023 Jan;82(1):3-18. [doi: 10.1136/ard-2022-223356] [Medline: 36357155]
4. Ialongo C, Bernardini S. Phlebotomy, a bridge between laboratory and patient. Biochem Med (Zagreb) 2016;26(1):17-33. [doi: 10.11613/BM.2016.002] [Medline: 26981016]
5. Rao PB, Mohanty CR, Singh N, Mund M, Patel A, Sahoo AK. Effectiveness of different techniques of ethyl chloride spray for venepuncture-induced pain: a randomised controlled trial. Anesth Essays Res 2019;13(3):568-571. [doi: 10.4103/aer.AER_103_19] [Medline: 31602079]
6. Humphrey GB, Boon CM, van Linden van den Heuvell GF, van de Wiel HB. The occurrence of high levels of acute behavioral distress in children and adolescents undergoing routine venipunctures. Pediatrics 1992 Jul;90(1 Pt 1):87-91. [Medline: 1614786]
7. Mitra TP, Coulter-Nile S, Jegathees T, Luong J, Shetty A, Lai K. Spiced RCT: success and pain associated with intravenous cannulation in the emergency department randomized controlled trial. J Emerg Med 2024 Feb;66(2):57-63. [doi: 10.1016/j.jemermed.2023.10.008] [Medline: 38278677]
8. WHO guidelines on drawing blood: best practices in phlebotomy. World Health Organization. 2010. URL: https://www.who.int/publications/i/item/9789241599221 [accessed 2025-07-22]
9. Miller FW. The increasing prevalence of autoimmunity and autoimmune diseases: an urgent call to action for improved understanding, diagnosis, treatment, and prevention. Curr Opin Immunol 2023 Feb;80:102266. [doi: 10.1016/j.coi.2022.102266] [Medline: 36446151]
10. Asfuroglu A, Balci M, Aslan Y, et al. What is the interaction between urine C-reactive protein, prostatic inflammation, and doxazosin treatment in patients with benign prostatic hyperplasia? A pilot study. Urol Int 2022;106(10):992-996. [doi: 10.1159/000518474] [Medline: 34515255]

XSL•FO
RenderX

11. Khan A. Detection and quantitation of forty eight cytokines, chemokines, growth factors and nine acute phase proteins in healthy human plasma, saliva and urine. J Proteomics 2012 Aug 3;75(15):4802-4819. [doi: 10.1016/j.jprot.2012.05.018] [Medline: 22634087]
12. Ashkenazi-Hoffnung L, Livni G, Scheuerman O, et al. Differential serum and urine CRP, IP-10, and TRAIL levels in pediatric urinary tract infection. Front Pediatr 2021;9:771118. [doi: 10.3389/fped.2021.771118] [Medline: 34966702]
13. Chuang YC, Tyagi V, Liu RT, Chancellor MB, Tyagi P. Urine and serum C-reactive protein levels as potential biomarkers of lower urinary tract symptoms. Urol Sci 2010 Sep;21(3):132-136. [doi: 10.1016/S1879-5226(10)60028-0]
14. Punyadeera C, Dimeski G, Kostner K, Beyerlein P, Cooper-White J. One-step homogeneous C-reactive protein assay for saliva. J Immunol Methods 2011 Oct 28;373(1-2):19-25. [doi: 10.1016/j.jim.2011.07.013] [Medline: 21821037]
15. SaranyaDevi KS, Rekha BS, Thiyagarajan JV, Dhivya R, Mihiran S, Santhosh S. Comparative evaluation of salivary and serum high-sensitive C-reactive protein in acute myocardial infarction. J Pharm Bioallied Sci 2022 Jul;14(Suppl 1):S127-S130. [doi: 10.4103/jpbs.jpbs_845_21] [Medline: 36110764]
16. Dongiovanni P, Meroni M, Casati S, et al. Salivary biomarkers: novel noninvasive tools to diagnose chronic inflammation. Int J Oral Sci 2023 Jun 29;15(1):27. [doi: 10.1038/s41368-023-00231-6] [Medline: 37386003]
17. Pay JB, Shaw AM. Towards salivary C-reactive protein as a viable biomarker of systemic inflammation. Clin Biochem 2019 Jun;68:1-8. [doi: 10.1016/j.clinbiochem.2019.04.006] [Medline: 30995442]
18. Sikorska D, Orzechowska Z, Rutkowski R, et al. Diagnostic value of salivary CRP and IL-6 in patients undergoing anti-TNF-alpha therapy for rheumatic disease. Inflammopharmacology 2018 Oct;26(5):1183-1188. [doi: 10.1007/s10787-018-0515-8] [Medline: 30043213]
19. Smith M, Sharma P, Dahiya S. 255 The validation of salivary C-reactive protein as a marker of inflammation in humans. Rheumatology (Sunnyvale) 2018 Apr 1;57(Suppl 3):key075.479. [doi: 10.1093/rheumatology/key075.479]
20. Tu J, Min J, Song Y, et al. A wireless patch for the monitoring of C-reactive protein in sweat. Nat Biomed Eng 2023 Oct;7(10):1293-1306. [doi: 10.1038/s41551-023-01059-5] [Medline: 37349389]
21. Jagannath B, Lin KC, Pali M, Sankhala D, Muthukumar S, Prasad S. Temporal profiling of cytokines in passively expressed sweat for detection of infection using wearable device. Bioeng Transl Med 2021 Sep;6(3):e10220. [doi: 10.1002/btm2.10220] [Medline: 34589597]
22. Gemicioglu B, Musellim B, Dogan I, Guven K. Fractional exhaled nitric oxide (FeNo) in different asthma phenotypes. Allergy Rhinol (Providence) 2014 Jan;5(3):157-161. [doi: 10.2500/ar.2014.5.0099] [Medline: 25565052]
23. Walsham NE, Sherwood RA. Fecal calprotectin in inflammatory bowel disease. Clin Exp Gastroenterol 2016;9:21-29. [doi: 10.2147/CEG.S51902] [Medline: 26869808]
24. Ehlers UE, Ulmer J, Keller M, Klein C, Pietsch U. Comparison of continuous temperature measurement methods in the intensive care unit: standard bladder catheter measurements versus non-invasive transcutaneous sensors. J Clin Monit Comput 2025 Feb;39(1):193-203. [doi: 10.1007/s10877-024-01199-2] [Medline: 39066870]
25. Losi S, Berra CCF, Fornengo R, Pitocco D, Biricolti G, Orsini Federici M. The role of patient preferences in adherence to treatment in chronic disease: a narrative review. Drug Target Insights 2021;15:13-20. [doi: 10.33393/dti.2021.2342]
26. Shim J, Muraru S, Dobrota R, Fleisch E, Distler O, Barata F. Noninvasive, multimodal inflammatory biomarker discovery for systemic inflammation (NOVA Study): protocol for a cross-sectional study. JMIR Res Protoc 2024 Nov 5;13(1):e62877. [doi: 10.2196/62877] [Medline: 39499914]
27. Koenig C, Ammann RA, Schneider C, Wyss J, Roessler J, Brack E. Continuous timely monitoring of core temperature with two wearable devices in pediatric patients undergoing chemotherapy for cancer—a comparison study. Support Care Cancer 2024 Feb 24;32(3):188. [doi: 10.1007/s00520-024-08366-w] [Medline: 38400942]
28. Turner SW, Chang AB, Yang IA. Clinical utility of exhaled nitric oxide fraction in the management of asthma and COPD. Breathe (Sheff) 2019 Dec;15(4):306-316. [doi: 10.1183/20734735.0268-2019] [Medline: 31803265]
29. Tang Y, Ling N, Li S, et al. A panel of urine-derived biomarkers to identify sepsis and distinguish it from systemic inflammatory response syndrome. Sci Rep 2021;11(1):20794. [doi: 10.1038/s41598-021-99595-0]
30. Salivary C-reactive protein. Salimetrics. URL: https://salimetrics.com/wp-content/uploads/2017/05/c-reactive-protein-saliva-elisa-kit.pdf [accessed 2025-07-22]
31. Human CRP (C-reactive protein) ELISA kit (DEIA217). Creative Diagnostics. URL: https://www.creative-diagnostics.com/Human-CRP-C-Reactive-Protein-ELISA-Kit-3749-497.htm [accessed 2024-12-09]
32. Human C reactive protein ELISA kit (CRP) (ab99995). Abcam. URL: https://www.abcam.com/en-us/products/elisa-kits/human-c-reactive-protein-elisa-kit-crp-ab99995 [accessed 2024-12-09]
33. Scott I, Carter S, Coiera E. Clinician checklist for assessing suitability of machine learning applications in healthcare. BMJ Health Care Inform 2021 Feb;28(1):e100251. [doi: 10.1136/bmjhci-2020-100251] [Medline: 33547086]
34. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: guidelines for reporting observational studies. Ann Intern Med 2007 Oct 16;147(8):573-577. [doi: 10.7326/0003-4819-147-8-200710160-00010] [Medline: 17938396]
35. Bechman K, Tweehuysen L, Garrood T, et al. Flares in rheumatoid arthritis patients with low disease activity: predictability and association with worse clinical outcomes. J Rheumatol 2018 Nov;45(11):1515-1521. [doi: 10.3899/jrheum.171375] [Medline: 30173149]

XSL•FO
RenderX

36. 810.301: ordinance on human research with the exception of clinical trials (HRO). Swiss Federal Council. 2014. URL: https://www.fedlex.admin.ch/eli/cc/2013/642/en [accessed 2025-08-01]
37. World Medical Association. World Medical Association Declaration of Helsinki: Ethical principles for medical research involving human subjects. Bull World Health Organ 2001;79(4):373-374. [Medline: 11357217]
38. Ruscitti P, Cipriani P, Masedu F, et al. Adult-onset Still's disease: evaluation of prognostic tools and validation of the systemic score by analysis of 100 cases from three centers. BMC Med 2016 Dec 1;14(1):194. [doi: 10.1186/s12916-016-0738-8] [Medline: 27903264]
39. Steinhoff J, Einecke G, Niederstadt C, et al. Renal graft rejection or urinary tract infection? The value of myeloperoxidase, C-reactive protein, and alpha2-macroglobulin in the urine. Transplantation 1997 Aug 15;64(3):443-447. [doi: 10.1097/00007890-199708150-00013] [Medline: 9275111]
40. Laiho K, Tiitinen S, Teppo AM, Kauppi M, Kaarela K. Serum C-reactive protein is rarely lost into urine in patients with secondary amyloidosis and proteinuria. Clin Rheumatol 1998;17(3):234-235. [doi: 10.1007/BF01451054] [Medline: 9694059]
41. Ouellet-Morin I, Danese A, Williams B, Arseneault L. Validation of a high-sensitivity assay for C-reactive protein in human saliva. Brain Behav Immun 2011 May;25(4):640-646. [doi: 10.1016/j.bbi.2010.12.020] [Medline: 21236331]
42. Hladek MD, Szanton SL, Cho YE, et al. Using sweat to measure cytokines in older adults compared to younger adults: a pilot study. J Immunol Methods 2018 Mar;454:1-5. [doi: 10.1016/j.jim.2017.11.003] [Medline: 29128425]
43. Brasier N, Wang J, Gao W, et al. Applied body-fluid analysis by wearable devices. Nature New Biol 2024 Dec;636(8041):57-68. [doi: 10.1038/s41586-024-08249-4] [Medline: 39633192]

Abbreviations

CBT: core body temperature
CRP: C-reactive protein
FeNO: fractional exhaled nitric oxide
IL: interleukin
RA: rheumatoid arthritis
SI: systemic inflammation
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
TNF: tumor necrosis factor
USZ: University Hospital Zurich

Edited by A Mavragani; submitted 07.05.25; peer-reviewed by S Chowdhury, X Ding; revised version received 03.07.25; accepted 05.07.25; published 26.08.25.

Please cite as:
Shim J, Muraru S, Löw V, Evers C, Riedo S, Dobrota R, Fleisch E, Distler O, Barata F
Multimodal Noninvasive Assessment of C-Reactive Protein for Systemic Inflammation in Adults: Cross-Sectional Study
JMIR Form Res 2025;9:e77108
URL: https://formative.jmir.org/2025/1/e77108
doi:10.2196/77108

© Jinjoo Shim, Sinziana Muraru, Vanessa Löw, Caroline Evers, Sandro Riedo, Rucsandra Dobrota, Elgar Fleisch, Oliver Distler, Filipe Barata. Originally published in JMIR Formative Research (https://formative.jmir.org), 26.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Development and Testing of an Exergaming and Artmaking Program to Enhance Cognitive and Physical Function in Older Adults: Results of a Community-Based Pilot Study

Nikhil Satchidanand[1], PhD; Mia Lawrence[1]; Sameer Jhaveri[2], DO; Sara Opalka-Satchidanand[3], MFA

[1]Department of Internal Medicine, University at Buffalo, State University of New York, DK Miller Building Room C209, 462 Grider Street, Buffalo, NY, United States

[2]Department of Internal Medicine, University of Pittsburgh, Pittsburgh, PA, United States

[3]Fine Art Miracles, Pittsburgh, PA, United States

Corresponding Author:
Nikhil Satchidanand, PhD
Department of Internal Medicine, University at Buffalo, State University of New York, DK Miller Building Room C209, 462 Grider Street, Buffalo, NY, United States

Abstract

Background: Group-based recreational activities, including physical activity and art-making, can help older adults preserve or even improve key functional outcomes essential to successful aging, while also fostering social support. Community-university partnerships can promote successful aging by facilitating the design and implementation of evidence-based, relevant, and impactful programs that better meet the needs of older adults in our community.

Objective: This study aimed to leverage a community-integrated approach to develop and test a recreation-based exergaming and art-making program on improving cognitive and physical functional outcomes in older adults.

Methods: Working with the Buffalo-Niagara YMCA, we developed the "Movers and Makers Club," a group-based recreation program to improve cognitive and physical function in aging. We recruited adults, ≥65 years, to participate in weekly group-based SMARTfit exergaming and artmaking during the 12-week program. The program was divided into 2 phases. During the design phase, we sought direct feedback from older adults in the community regarding the appeal and usability of proposed exergaming activities and art projects. In the impact phase, we administered a 1-group, 12-week pilot study to test the effects of group-based SMARTfit exergaming and visual artmaking on performance on the Trail-Making Test and Stroop Color and Word Test. In addition, other functional assessments included a 4-Stage Balance Test, 30-second Sit-to-Stand test, and a 6-Minute Walk Test. Key participant-reported outcomes were also collected on the appeal, enjoyability, and usability of program activities.

Results: In all, 17 older adults (mean age 73.80, SD 4.10 y) enrolled in the program between 2 YMCA branches: one representing the City of Buffalo and one, a local suburb. At the end point, all 17 participants completed at least 9 out of 12 sessions. Participants reported high ratings of enjoyment and satisfaction with their experience. They also reported that their comfort level while exercising, maintaining their balance, and walking had improved. Participants also found their comfort level with doing art, and their interest in trying new artistic pursuits increased. In addition, performance on the Trail-Making Test and Stroop Color and Word Test improved from baseline to end point ($P<.001$). We also observed improvements in 1-leg balance, 30-second Sit-to-Stand test performance, and 6-Minute Walk Test distance ($P<.001$).

Conclusions: Our group-based recreation program, delivering SMARTfit exergaming and guided art-making to older adults, improved cognitive and physical functions important to successful aging. In addition, our approach of seeking direct feedback from older adults, in developing the program, produced a highly enjoyable and appealing experience that was age- and ability-appropriate. Our future programs will integrate formal assessments of key psychosocial factors that may influence both the implementation and functional impacts of the "Movers and Makers" curriculum.

(JMIR Form Res 2025;9:e73555) doi:10.2196/73555

KEYWORDS

exercise; aging; art; community; exergaming; older adults; social support

Introduction

Preserving cognitive and physical function is critical to successful aging [1-3]. Recreational activities, such as group exercise and visual artmaking, offer older adults an enjoyable experience while simultaneously helping them maintain function, independence, and quality of life [4-14]. These activities also often provide a social experience and have the

XSL•FO
RenderX

potential to enhance functional capacity while promoting social support [8,10,14,15]. In a review by Ali et al [6] that examined the effects of cognitive-motor exercise on cognition and physical function, they found small to medium effect sizes ranging from 0.19 to 0.56 for balance, dual-task cost, gait speed, attention, memory, executive function, and global cognition. Similarly, in our previous YMCA-based program, we found that 12 weeks of group-based cognitive-motor exercise, using the SMARTfit Cognitive-Motor Exercise System (Ventura, CA), was associated with significant improvements in attention, task-switching, and interference inhibition as assessed by the Trail Making Test (TMT) and Stroop Color and Word Test (SCWT) [12]. Overall, the evidence is encouraging that dual-task exercise can help older adults preserve key physical and cognitive functions critical to maintaining independence and quality of life, in a setting that fosters social support [6,16,17].

Recent studies have sought to characterize the potential of visual artmaking as a therapy to enhance functional capacity in aging. In a review by Masika et al [4], visual artmaking, including painting, drawing, and craft-based activities, was shown to improve cognitive function among older adults with mild cognitive impairment (MCI) [4]. In this review, visual artmaking was also associated with lower depression and anxiety scores among cognitively impaired and cognitively intact older adults. Programs that include socialization, art education, reminiscence, art processing and analysis, or art sharing were more likely to enhance cognitive function [4]. For instance, older adults in a nursing home setting who participated in a group-based ceramic painting program, which encouraged socialization, art sharing, and social support, demonstrated improved scores on the Mini-Mental State Examination from before to after the program [9].

Current evidence shows that separate, group-based, dual-task exercise and art-making can help maintain or even improve cognitive and physical function in aging [4,11-13,15,18-20]. A novel community-based recreation program that brings these activities together while encouraging socialization will promote participation and foster social support, while improving functions essential to successful aging. We established a collaboration with the Buffalo-Niagara YMCA to leverage the expertise and commitment of this community organization with demonstrated successes in serving the needs of older adults. This, in combination with our dedication to exploring evidence-based, relevant, and functional outcomes, allowed us to design and implement a series of novel, group-based programs to support the health and function of older adults in our community.

The Movers and Makers Club was developed from this community-university partnership. Our overarching aim was to identify ways in which our team can effectively integrate combined group exergaming and artmaking into diverse settings, which will improve outcomes in older adults. Our study was also designed to explore the enjoyability, appeal, and usability of program activities. The insights from this pilot have helped optimize program elements to create an enjoyable and impactful community-based program. To our knowledge, this is one of the first recreation-based programs that combines group dual-task exergaming with visual artmaking to explore their effects on cognition and physical function in older adults in a community setting.

Methods

Participants and Recruitment

Our approach to recruitment and participant inclusion was derived from our previous work [12]. Adults, 65 years or older, were recruited from nearby senior centers, libraries, and other community organizations using recruitment flyers. Older adults interested in participating contacted the YMCA by telephone to schedule a consent and screening session. Screening for eligibility included (1) cognitive assessment (Montreal Cognitive Assessment [MoCA]), (2) readiness (Physical Activity Readiness Questionnaire for Older Adults), and (3) capacity to provide consent (University of California Brief Assessment of Capacity to Consent). Participants were provided details about the study before enrollment, and informed consent was obtained.

Participants were included if they (1) were >65 years of age, (2) had no more than MCI (MoCA score ≥24 points), (3) had no contraindications to exercise guided by the American College of Sports Medicine, and (4) were fluent in English. Potential participants were not included if (1) they had a condition that would prevent safe participation in the exercise, as determined by the Physical Activity Readiness Questionnaire for Older Adults, (2) severe neurological disease, (3) severe psychiatric illness, (4) likelihood of withdrawing from the study due to severe illness or a life expectancy of <6 months, (5) they had a lower or upper limb amputation, and (6) had greater than MCI (MoCA score <24 points).

Study Phases

Design Phase

The design phase introduced community members to the SMARTfit training and art-making projects that were then integrated into the impact phase. The purpose was to garner direct feedback from older adults before testing the activities. This was to ensure an enjoyable, age- and ability-appropriate recreational experience.

The dual-task training program was delivered using the SMARTfit Cognitive-Motor Exercise System (SMARTfit). SMARTfit is an engaging exergaming platform used to facilitate dual-task training. This system has a series of touch-sensitive LED arrays attached to wall-mounted, high-strength composite panels. The LED arrays are programmed to deliver several cognitive games, including memory, alphabet, counting, and pattern recognition. Adding a simultaneous physical task, such as stepping in place, creates a cognitive-motor challenge, allowing hundreds of combined tasks to be executed. Participants can play various games by striking the individual targets with their hands or by hitting them with other objects such as “pool noodles,” drumsticks, or tennis balls. The SMARTfit system facilitates dual-task training while delivering an experience that is fun and engaging [12].

We conducted one 45-minute small group session at each of the 2 YMCA branches. In all, 12 participants completed several cognitive tasks, including a memory game, chase the target,

XSL•FO
RenderX

addition, and subtraction. These tasks were paired with stepping in place, walking to and from the SMARTfit board, striking the targets with "pool noodles," or lightly punching the targets (with boxing gloves on). Different game durations were tested, ranging from 30 seconds to 2 minutes.

Proposed art-making projects were also presented to older community members to gather their feedback and preferences regarding media type, project complexity, and the overall appeal of each project. A total of 11 older adults completed this phase at the Ken-Ton Family YMCA (KTY) branch. One 40-minute session was administered to introduce the art-making portion of the program. As shown in Table 1, example projects involving several media were presented along with a brief description of the project. This was meant to enhance participant understanding of the art projects and ensure a wide variety of projects were included, thus giving participants many opportunities to find enjoyable, rewarding projects at which they could succeed.

Table . Example art projects with corresponding inspiring artists used to deliver the art-making component of the Movers and Makers program.

Artist and inspiration	Project
David Hockney	Landscapes: oil pastel
Lascaux, France	Cave paintings: soft pastel
Terrance Osborne	New Orleans stylized painting: acrylic
Henri Matisse	Cut-paper collage
Yayoi Kusama	Mushroom art: acrylic
Georgia O'Keeffe	Landscape: watercolor

Impact Phase

The impact phase consisted of a 1-group intervention administered at 2 branches within the Buffalo-Niagara YMCA system. We assessed changes in cognition and physical function associated with 12 weeks of group-based SMARTfit training and guided artmaking. In all, 17 participants completed the program between the 2 YMCA branches. A total of 8 participants were enrolled at the William-Emslie Family YMCA (WEY), a city-based branch with a membership that is predominantly Black. In all, 9 participants were enrolled at the KTY, a branch that has a membership that is predominantly White. For both YMCA branches, the principal investigator (NS) monitored the trial conduct monthly and determined criteria for stopping the program early for any participant, which included participant request or self-disclosure of a new diagnosis that would meet exclusion criteria.

Program participants completed one 60-minute, group-based, exergaming session and one 60-minute guided group art-making session per week for 12 weeks. SMARTfit and art sessions were facilitated by trained and experienced instructors. The SMARTfit program was created by the principal investigator (NS), while the group art-making program was developed by the cocreator and art instructor (SO-S) for the "Movers and Makers Club."

Description of the SMARTfit Program

At each YMCA branch, participants completed the exergaming training in groups of 6-8 individuals. Sessions began with a 3-minute warm-up including standing stretching, walking in place, and light standing calisthenics. The SMARTfit instructor then introduced the dual-task training games for that session, guiding the group through each game. Participants completed the same training games of the same duration and order. As illustrated in Table 2, several individual and team (involving 2 participants working together) games were played.

XSL•FO
RenderX

Table . Example SMARTfit dual-task training games with focus areas, used to deliver the 12-week, 1-group, dual-task exercise program.

Name of game	Focus areas	Individual	Team
Subtraction with drumstick tap	Processing speed, eye-hand coordination	✓	—[a]
Pattern replication with medicine ball tap	Working memory, attention, strength	✓	—
Knock-out for speed with punching	Eye-hand coordination, processing speed	✓	—
Kicking with one-leg stand (assisted or unassisted)	Strength, attention	✓	—
Left, right, or both hands	Attention, processing speed, eye-hand coordination	—	✓
Addition with side-step (alternating)	Balance, processing speed	—	✓
Chase the color with ball toss and catch (alternating)	Eye-hand coordination, attention	—	✓
Counting with medicine ball carry (alternating)	Attention, strength, processing speed	—	✓
Trail-making with ball bounce pass (alternating)	Eye-hand coordination, attention, processing speed	—	✓

[a]Not applicable.

Games coincided with specific functional domains, including balance, muscle strength, hand-eye coordination, attention, and memory. These domains were identified and tested in the development of our previous trial [12]. A variety of physical movements added a dual-task dimension to the cognitive training (eg, stance changes, sit-to-stand, or patterned stepping). The games were 30 seconds to 2 minutes, and each training session allowed 4-7 different games to be completed. From our previous study, we estimate that each participant accumulated 8-10 minutes of sedentary time while transitioning from one game to the next. Each session concluded with 3 minutes of cool-down. During training, speed and accuracy of the response to each challenge were emphasized. In addition, participants were instructed not to help each other while performing the cognitive challenges. Instructors monitored participant safety throughout each training session.

Description of the Artmaking Program

Participants completed 1 hour of guided artmaking for 12 weeks, in small groups of up to 12 individuals, allowing for sufficient one-on-one instruction time. Art sessions were completed before or after the SMARTfit training, based on each participant's schedule.

Art sessions began with an introduction to the project, a brief description of the inspiring artist or art movement and history, and an explanation of the technique and media being used. Participants were provided with all the materials required, instructions for completing the project, and several complete examples of each project for reference and inspiration. In addition, templates, pre-made project elements (eg, pre-cut paper for collages), and other supportive materials were given to participants to promote success and ensure completion of each project in a timely manner.

A wide array of projects was selected to expose participants to many different media and techniques, while also ensuring participants completed several projects that were enjoyable and rewarding. Individual expression within each project was encouraged, as were artwork-sharing and discussion. The art instructor (SO-S) remained focused on supporting, guiding, and inspiring participants as they worked on their projects.

Outcomes Assessment

Participant-Reported Outcomes

A brief survey instrument was created for this project to explore participants' experiences in the program. This instrument was assessed for acceptability and ease of understanding in 10 older community members before using it for this program. Themes explored included participant interest, appreciation, comfort level, and confidence with each activity after completion of the Movers and Makers program. In addition, we asked participants to evaluate the quality of instruction, their overall enjoyment, and satisfaction with each program component. Participants were also asked to comment on any changes they thought would improve their experience. This information was used to design an expanded program involving several additional YMCAs and to recruit sufficient participants to more deeply assess changes in key functional outcomes.

Cognitive and Physical Function

Outcome assessment methods were adapted from our previously published trial [12]. Assessors administered the battery of instruments in an office or meeting room in private, to allow sufficient privacy to ensure participants were able to concentrate on the tasks at hand. The assessors were blinded to baseline performance at the time of completion of the end point assessment, having no access to baseline scores. The TMT and SCWT were administered at baseline and end point. The TMT assesses task-switching, visual attention, and memory. This test has 2 parts, TMT-A (rote memory) and TMT-B (executive functioning). The participant draws a line to connect 24

XSL•FO
RenderX

consecutive circles containing numbers and letters randomly arranged on a sheet of paper, without lifting the pen tip off the paper [21]. The TMT-A section has numbers, while the TMT-B has both numbers and letters, requiring the participant to alternate between a number and a letter in ascending order. The total score for each section is based on the time (in seconds) needed to complete each section. The TMT was administered on paper, once at baseline and once at end point, using standard procedures including 1 sample trial and 1 test trial [21].

The SCWT assesses interference inhibition and working memory. The participant being assessed is instructed to read aloud 3 different tables, one at a time, as quickly as possible. The first 2 tables are congruent conditions requiring the participant to read names of colors printed in black ink (word condition) and name different blocks of colors (color condition). The third condition is incongruent (color-word), containing color names printed in a different colored ink (eg, the word "red" printed in "blue" ink). The participant must name the color of the ink, not read the word. The score earned is based on the number of correct answers given in each condition within 45 seconds [22]. The SCWT was administered individually, with the participant accessing each table one at a time, at baseline and end point. A sample of each table was presented to anchor participants to the assessment.

Physical function was measured using a modified 4-Stage Balance Test (FSBT), a 6-Minute Walk Test (6MWT), and a 30-second Sit-to-Stand test (30STS). The FSBT included side-by-side, tandem, and left and right one-leg stands. These tests were administered individually in a private room by a trained assessor. The order of assessment was (1) FSBT: side-by-side, tandem, one-leg (timed for 10 seconds each); (2) 6MWT; and (3) 30STS. A higher score indicates better performance on each functional measure.

Data Management

Study data, both outcomes and safety, were collected by the principal investigator (NS) and project coordinator using collection forms and guided by the standard procedures for each assessment (eg, TMT and SCWT) that were described in the Methods section. Quantitative data were entered into a Microsoft Excel database weekly by a trained research assistant. Biweekly data quality checks were performed by the principal investigator (NS) by reviewing the paper data collection form and entered data. The password-protected Excel file was stored on a password-protected computer in the principal investigator's (NS) office. Safety data were reviewed by the principal investigator (NS) and project coordinator weekly, after completion of each training session.

Statistical Methods

First, measures of central tendency (mean, median, and mode) were calculated to assess continuous variables, while categorical and questionnaire data were examined using frequencies. To assess participants' experiences in the Movers and Makers program, responses to the survey items were examined as proportions of the highest score on each response. In addition, free-text responses were examined for commonly occurring themes. For both the quantitative and qualitative outcomes, we found no missing data.

A preliminary sample size calculation revealed that to detect a large effect size (0.50) in change in scores on the TMT, with a statistical power of 0.80, a total of 26 participants needed to be enrolled. Due to the low sample size and nonnormal distribution, median and IQR were calculated to explore pre- and postprogram performance on the TMT, SCWT, FSBT, 30STS, and 6MWT. Wilcoxon Signed Rank Tests were conducted to examine changes in performance on each test. In addition, the predicted color-word score (Pcw) was calculated based on scores on the 2 congruent conditions, using the following calculation:

$$Pcw=45/\{((45\times W)+(45\times C))/(W\times C)\}$$

The interference score (IS) was generated by subtracting the Pcw from the number of items correctly named in the Color-Word condition. As guided by Ivnik et al [23], a lower IS represents more difficulty inhibiting cognitive interference. Pre- to postprogram change in the IS was also assessed using a Wilcoxon Signed Rank test.

Ethical Considerations

The participants' protections in this study were reviewed and approved by the University at Buffalo institutional review board (STUDY00003525). All participants provided informed consent. Participants reviewed the consent form independently in a private office, allowing them sufficient time to examine all elements of the study thoroughly. Time was then provided for participants to ask questions or voice concerns to the researcher obtaining consent. Participants were compensated with a US $20 prepaid gift card to complete the baseline and end point assessments. In addition, the University of California Brief Assessment of Capacity to Consent was used to confirm that participants fully understood the consent form and the study procedures. Study data are being stored deidentified, with each participant assigned a unique identification number. The key that links participants to their data is being stored as a separate password-protected file on a password-protected computer in the principal investigator's (NS) office.

Results

Sample Characteristics

In all, 12 older adults completed the SMARTfit user groups in the design phase, consisting of 11 females and 1 male with a mean age of 71.28 (SD 2.54) years. A total of 11 older adults completed the introductory art session, including 10 females and 1 male. The mean age in this group equaled 74.43 (SD 2.67) years.

A total of 24 older adults volunteered to participate in the impact phase and were screened for eligibility, resulting in 17 being enrolled. Please refer to Figure 1 for participant flow through the research process. Exclusion was based on not meeting minimum age (n=3), a previous unresolved back injury (n=1), uncontrolled hypertension (n=1), and cognitive impairment beyond mild (n=2).

XSL•FO
RenderX

Figure 1. Participant flow diagram

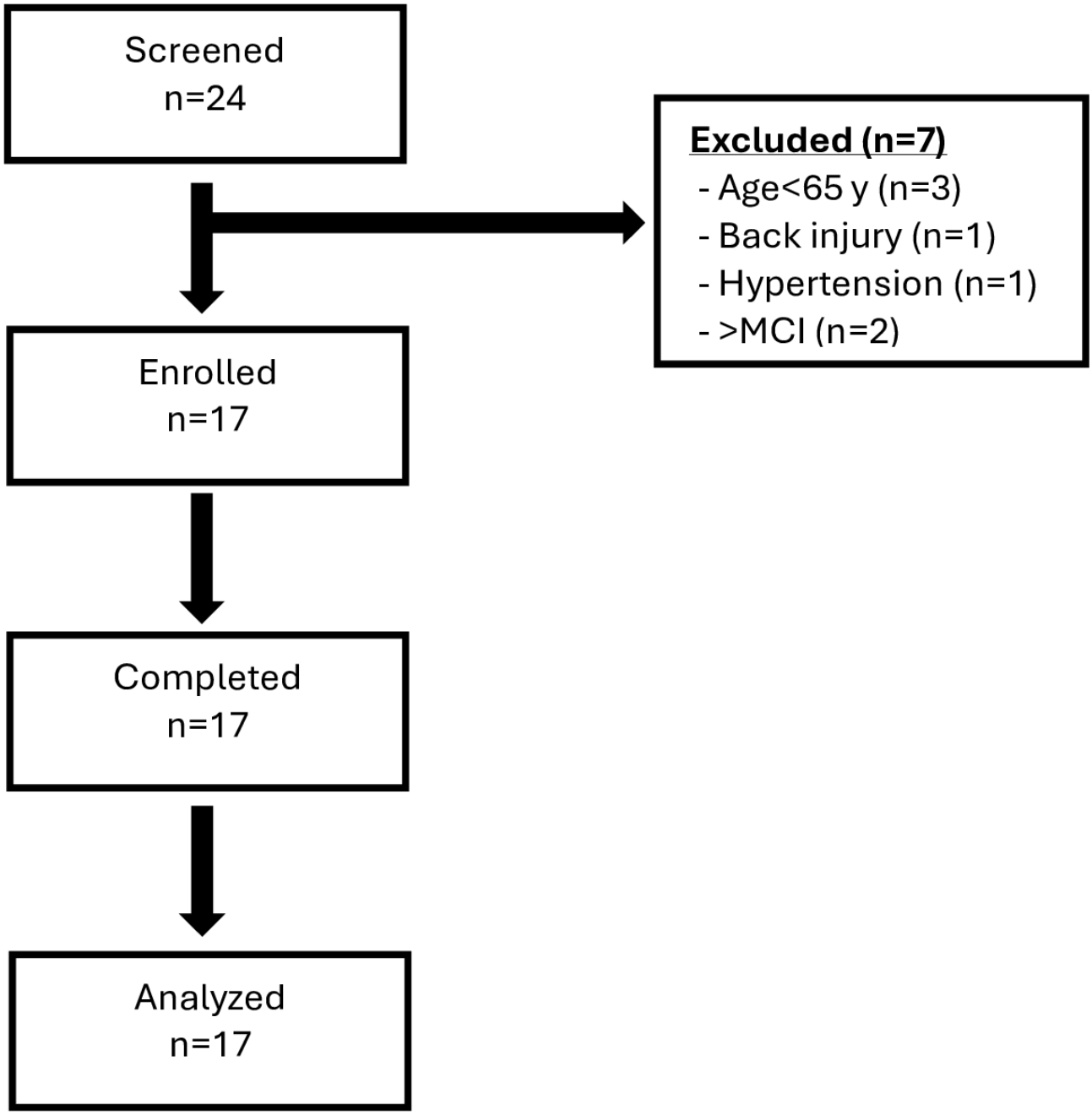

In all, 9 participants were enrolled at the KTY and 8 participants were enrolled at the WEY. At the end point, all 17 enrolled participants were "completers," defined as participating in at least 9 of the 12 sessions, and 5 participants completed all 12 sessions. Among the participants at the KTY, the mean age equaled 72.30 (SD 3.60) years and the mean age at the WEY equaled 74.13 (SD 4.28) years; these means were not found to be statistically significantly different (P=.34). In addition, all participants self-reported as female and 94% (16/17) reported being retired, while 6% (1/17) reported working part time. There were no differences between the 2 YMCA branches by employment status (P=.92).

Participant Experience

Overall, participants favorably rated their experiences in the Movers and Makers program. In all, 100% (N=17) of participants responded with "Very Enjoyable" when asked to rate their enjoyment on a scale of 1 - 5. In addition, 100% (N=17) reported they were "Highly Satisfied" with their overall experience. When they reflected on the social aspect of the program, 100% (N=17) of participants reported that it enhanced the program and their experience. One participant at the KTY stated that being in Movers and Makers helped her try new activities and meet new people. She said,

> *I thought the exercise and art would be fun, which is why I joined the program. The SMARTfit exercises were unique, and I enjoyed the games more than regular group exercises I tried before. I didn't realize how much fun we would have as a group. Playing SMARTfit together and sharing our art really helped me make some new friends that I don't think I would have met in another class.*

XSL•FO
RenderX

When asked about the perceived benefits of the SMARTfit training, 100% (N=17) of participants reported that their experience improved their comfort level with physical activity and made them feel more capable while walking, maintaining their balance, and avoiding falls. In addition, 100% (N=17) of participants reported their interest in trying new physical activities increased after the program. One participant at the WEY reported that being part of the Movers and Makers "Club" helped her realize her capabilities. She stated that,

> *Being on the M&M team made me see I am capable of much more than I realized. I used to be afraid of walking in large crowds, like at the mall. I thought I was walking so (slow) and would get walked right over. But now I see I can do it, even with a bad hip. I am definitely more comfortable and confident in public.*

Similarly, when asked about their experiences in the group artmaking, 100% (N=17) of participants had favorable responses. They reported that participation enhanced their appreciation for art, improved their comfort and confidence when doing art, and piqued their interest in trying new creative endeavors. One participant at the KTY reported that she has always found art to be intimidating. She said,

> *Art always intimidated me, even as a kid. Being in the Movers and Makers was a really positive experience for me. Sara took the fear and intimidation away and made it really fun. I realize I don't have to be perfect at it and I can improve even at my age.*

Participants were also asked to describe their experiences in 1 word, for both the SMARTfit sessions and the art classes. Several keywords were associated. The SMARTfit sessions were described as "enjoyable," "energizing," "fun," "engaging," and "beneficial." Similarly, the artmaking was described as "interesting," "relaxing," "engaging," "inspiring," and "relatable." Overall, all participants in the program indicated their experience was rewarding and enjoyable.

We also asked about the quality of instruction for both the SMARTfit training and art. The SMARTfit instructors were characterized as "supportive," "encouraging," and "well-prepared for each class." Participants reported feeling safe while exercising in their groups and that their instructor demonstrated enthusiasm for teaching. Similarly, participants reported that the art instructor was well-prepared for each lesson, she provided clear, easy-to-follow instructions, and was "supportive" and "encouraging." In addition, participants felt the art instructor demonstrated enthusiasm for teaching and made the lessons interesting. One participant at the WEY found that exercising with other women, with guidance from the instructor, was encouraging. She said,

> *I was scared to do the exercise part of the program because I had not exercised in many years and I can't move too (quick) now. But, on the first day I saw a few other women moving like I do. Our instructor was so encouraging and helped us stay positive.*

Participants were also asked to describe any challenges they had with the program or any changes they would recommend. Furthermore, 3 participants reported that they felt some of the physical movements during the SMARTfit sessions were too difficult for them and that they felt discouraged at first.

Performance Outcomes

Cognition

The MoCA was administered before beginning the program for screening purposes, to identify potential participants with cognitive impairment beyond mild. At baseline, the median MoCA score was 26.00 (IQR 25.00-27.00). For the cognitive assessments, the scores are reported as median (IQR). Median baseline performance on the TMT-A improved from 43.10 (IQR 42.14-45.02) seconds to 41.00 (IQR 39.56-41.99) seconds (W=1.00, $P<.001$, $r=-1.00$; Table 3). The score on the TMT-B also improved from baseline to end point, with a median baseline of 140.30 (IQR 137.27-145.29) seconds to an end point median of 137.80 (IQR 134.07-141.38) seconds (W=1.00, $P<.001$, $r=-1.02$; Table 3). In addition, performance on the color condition and word condition of the SCWT improved from baseline to the end point ($P<.05$; Table 3). Similarly, the incongruent (Color-Word) condition of the SCWT also improved. At baseline, the median equaled 32.50 (IQR 30.00-34.00) seconds, and at end point, the median equaled 35.50 (IQR 33.00-36.50) seconds (W=1.00, $P<.001$, $r=0.88$; Table 3). The median IS of the SCWT also improved significantly from −4.50 (IQR –5.70 to –3.53) to −1.90 (IQR –3.50 to –1.71; $P<.001$; Table 3).

Physical function also improved from baseline to end point. Median right leg balance at baseline was 7.39 (IQR 6.49-8.95) seconds and improved to 10.00 (IQR 8.98-10.00) seconds (W=0.00, $P<.001$, $r=1.00$) (Table 3). Median left leg balance improved from 7.21 (IQR 6.89-8.60) seconds to 9.25 (IQR 7.86-10.00) seconds (W=0.00, $P<.001$, $r=1.00$; Table 3). Performance on the side-by-side stand and tandem stand did not improve. Improvement was observed in the 30STS, with a baseline median score of 16.00 (IQR 15.00-17.25) completed stands increasing to an end point median of 19.00 (IQR 18.00-22.00) completed stands (W=0, $P<.001$, $r=0.90$; Table 3). Performance on the 6MWT also improved from a median of 547.50 (IQR 497.75-571.75) m to 566.70 (IQR 522.00-583.25) m (W=2.00, $P<.001$, $r=0.92$; Table 3).

XSL•FO
RenderX

Table . Quantitative results of the Wilcoxon signed-rank tests for baseline to end point comparison of median cognitive and physical functional performance, Wilcoxon W, *z* score, and effect size for the one-group dual-task exercise program.

Outcome	Median score (IQR)	96.91% CI	W	*P* value	*z* score	Effect size (r)
TMT A[a,b]			1.00	.02	–4.32	–1.00
Baseline	43.10 (42.14 to 45.02)	42.15 to 45.01				
End point	41.00 (39.56 to 41.99)	39.67 to 41.98				
TMT B[c,d]			1.00	<.001	–4.33	–1.02
Baseline	140.30 (137.22 to 145.29)	137.27 to 145.09				
End point	137.80 (133.86 to 141.51)	134.07 to 141.38				
Stroop Color[b,e]			0.00	.003	2.95	0.90
Baseline	73.50 (70.00 to 75.00)	71.00 to 75.00				
End point	75.00 (73.00 to 76.00)	73.00 to 76.00				
Stroop Word[b,e]			12.50	.022	2.31	0.64
Baseline	74.00 (72.00 to 76.00)	72.00 to 76.00				
End point	75.00 (73.00 to 77.00)	73.00 to 77.00				
Stroop Color – Word[e,d]			1.00	<.001	3.72	0.88
Baseline	32.50 (30.00 to 34.00)	30.00 to 34.00				
End point	35.50 (33.00 to 36.00)	33.00 to 36.50				
Stroop Interference[d,f]			0.00	<.001	4.48	1.05
Baseline	–4.50 (–5.70 to –3.53)	–5.75 to –3.25				
End point	–1.90 (–3.50 to –1.71)	–3.24 to –0.99				
FSBT Side-By-Side[g,j]			—	—	—	—
Baseline	10.00 (10.00 to 10.00)	10.00 to 10.00				
End point	10.00 (10.00 to 10.00)	10.00 to 10.00				
FSBT Tandem[g,j]			—	—	—	—
Baseline	10.00 (10.00 to 10.00)	10.00 to 10.00				
End point	10.00 (10.00 to 10.00)	10.00 to 10.00				
FSBT-Left[g,d]			0.00	<.001	4.17	1.00
Baseline	7.21 (6.89 to 8.60)	6.91 to 8.42				
End point	9.25 (7.86 to 10.00)	7.88 to 10.00				

XSL•FO
RenderX

Outcome	Median score (IQR)	96.91% CI	W	*P* value	*z* score	Effect size (r)
FSBT-Right[g,d]			0.00	<.001	4.01	1.04
Baseline	7.39 (6.49 to 8.95)	6.59 to 8.88				
End point	10.00 (8.98 to 10.00)	9.02 to 10.00				
30STS[h,d]			0.00	<.001	3.51	0.90
Baseline	16.00 (15.00 to 17.25)	15.00 to 17.00				
End point	19.00 (18.00 to 22.00)	18.00 to 22.00				
6MWT[i,d]			2.00	<.001	3.62	0.92
Baseline	547.50 (497.75 to 571.75)	502.00 to 571.00				
End point	563.50 (522.00 to 583.25)	523.00 to 582.00				

[a]TMT A: Trail-Making Test A.
[b]Difference is significant at the *P*<.05 level.
[c]TMT B: Trail-Making Test B.
[d]Difference is significant at the *P*<.001 level.
[e]Stroop Color and Word Test.
[f]Stroop Interference score.
[g]FSBT: 4-Stage Balance Test.
[h]30STS: 30-second Sit-to-Stand test.
[i]6MWT: 6-Minute Walk Test.
[j] No difference: not applicable.

Discussion

Principal Results

Overall, our participants reported high levels of enjoyment in the SMARTfit training and group artmaking. Previous studies demonstrate that positive affective states during exercise are associated with higher levels of adherence. In an umbrella review by Collado-Mateo et al [24], enjoyment is identified as an important predictor of exercise adherence. Enjoyable recreational experiences for older adults can substantially impact their willingness to maintain participation over time [24]. Woolley and Fishbach [25] describe enjoyment as an immediate reward that can lead to better persistence, versus the delayed reward of improved health or function. Enjoyment, then, is likely an important influence on both adoption of and adherence to health-related activities, including group exercise and art making. Our program focused on gathering input directly from older adults to identify the characteristics that are optimally appealing and enjoyable. While we could not find studies that directly investigated this phenomenon in group-based visual artmaking as an interventional therapy, it stands to reason that a similar underlying mechanism is at play. Simply, people are more likely to adopt activities they find enjoyable.

Participants reported several perceived benefits of the Movers and Makers program, informing the development of future, larger-scale community-based initiatives. Perceived benefits included increased comfort with exercise and artmaking, more physical and artistic confidence, and an increased interest in trying new activities. While we did not directly measure self-efficacy and self-competence during this small pilot, our findings are encouraging that participation in a community-based recreation program can enhance older adults' beliefs in their capacity to accomplish exercise- or art-based activities, thus encouraging future participation [26]. Our future studies will focus on formally assessing these key constructs in the context of delivering the Movers and Makers curriculum.

Insights from participants have helped us strengthen our methods as we expand our program. Our qualitative results demonstrate that specific attributes of our program may have influenced the overall positive attitudes reported by participants. In addition to creating an enjoyable experience and providing opportunities to enhance self-efficacy and self-competence, the group-based nature was positively received by all participants. This hints at the important role of social support in promoting long-term adherence to health-related activities and ultimately in improving functional capacity in aging. Exercise-based programs that promote social connectedness are more successful at maintaining participation in the longer term [27]. We observed this among participants where words of encouragement and positive peer feedback were commonplace, and 100% (N=17) of participants enrolled completed at least 9 of 12 sessions. It is reasonable to believe that this overall positive, supportive environment was associated with a high adherence rate and overall favorable attitudes about the Movers and Makers program.

XSL•FO
RenderX

Specific cognitive outcomes showed modest improvements with 12 weeks of participation. Specifically, the TMT-B and color-word conditions of the SCWT improved, with effect sizes ranging from 0.90 to 1.10. While the changes in baseline to end point medians are small, it is encouraging that specific cognitive domains, including attention, task-switching, and interference inhibition, can be enhanced in the context of a community-based group recreation program. Previous evidence shows that these specific cognitive domains often decline with advancing age [28,29]. Development and validation of practical and appealing strategies to preserve or enhance key cognitive functions is essential to maintaining independence. Community-based programming increases the reach of interventions and provides an opportunity to deliver evidence-based therapies to those who need them while promoting social support [30,31].

We found the median IS of the SCWT improved, a score that can be derived using different methods found in the literature. Among studies that have reported an IS, most have used the equation developed by Golden in 1978 [32]. One systematic review that compared studies that evaluated indirect scores found that of 27 studies, 16 used Golden's equation [32]. In addition, a review [22] found that 17 out of 21 studies relied on a 45-second time interval, as we did in this study. In all, 16 of these studies reported results from Golden's standardized formula. Given our previous experience using the Golden equation [12], we elected to report it as part of our findings.

Similar to our previous work [12], we found that the SCWT IS improved after 12 weeks. However, the score remained negative, showing that despite the significant improvement, there may be lingering impairments in interference inhibition. As we stated in our previous study, the persistence of a negative IS at the end point is reasonable when we consider a relatively short training duration of 12 weeks combined with a lower weekly training volume. We anticipate that a longer duration study in combination with higher training volume would result in greater improvements in interference inhibition [12]. Key physical functional outcomes improved with participation. One-leg balance improved, reaching the maximum score of 10 seconds for the right leg and 9.25 seconds for the left leg. Improvement in 1-leg balance is noteworthy given that the ability to stand on 1 leg for at least 10 seconds is independently associated with reduced all-cause mortality and adds prognostic value for mortality beyond age, sex, anthropometric, and clinical factors [33].

On the 6MWT, we observed that the end point median value of 563.50 (IQR 522.00-583.00) m is only slightly above the expected norm of similarly aged females (mean 560, SD 49 m) as revealed in a review by Bohannon [34]. These findings are promising that walking speed can be improved in older adults with just once weekly exposure to activity. However, the exact dose of this type of dual-task training that optimizes walking ability is still unclear.

The improvements we observed in balance, walking, and lower limb strength are also of note because each is an important predictor of fall risk in aging [35-37]. We have shown that a socially engaging, fun recreation program can help reduce fall risk by enhancing key physical functional outcomes. Importantly, the Movers and Makers program was not specifically designed to reduce the risk of falling in this population; however, our results are interesting. Our future efforts will focus more closely on testing the impact of the Movers and Makers curriculum on modifiable risk factors for falls in an aging population, and our assessment methods will be designed to reflect this important outcome.

Limitations

Our community-university partnership produced an appealing, enjoyable, and impactful recreation program. However, we must acknowledge some limitations. The small-scale pilot nature of our study and the lack of a control group prevent our ability to draw inferences of causation. Furthermore, several confounding factors in the intervention may have influenced our positive findings, including the socially engaging nature of the program alone and interaction with program instructors. It is also possible that increases in participants' leisure-time physical activity outside of the program could have contributed to our favorable results. While participants were instructed to maintain their same activity habits, leisure-time physical activity was not monitored during the program.

Also, regarding the training sessions, participants received 1 hour of exercise and 1 hour of art each week, which is a relatively low volume of exposure. While we are encouraged by our findings that even a low dose of training was associated with functional improvements, we do not know the impact that a higher dose of recreational activities would have on the same outcomes. Future studies might be designed with different doses of training to identify which is optimally beneficial for a variety of outcomes.

In addition, a limited sample size restricts us from developing predictive models or exploring the impacts of possible covariates. Future studies will focus on expanding the inclusion of more community-based sites to increase sample size. In addition, we were only able to enroll female participants, which is an ongoing challenge in this community setting. Future interventions will use more targeted recruitment strategies to improve diversity in gender and race or ethnicity. This includes enhancing awareness of our program in the community through a series of open-house events to showcase the program activities. We also propose working directly with older men in the community to develop a training program that is more appealing to them. In addition, while our instructors at each site received identical training, there may have been a small variation in content delivery between YMCA branches. However, we did not integrate any implementation outcomes, such as fidelity to track these factors.

We must also acknowledge that the combination of exercise and art prevents us from exploring the impact of the individual components on our program. This would require at least three groups: (1) combined exercise and art, (2) exercise only, and (3) art only. However, it must be noted that the purpose of our study was to elucidate the therapeutic benefits of the full recreation program and not of the individual components. However, we theorized that inclusion of the artmaking would enhance participant engagement, create more opportunities for mastery and skill-building, build self-confidence, and create a

XSL•FO
RenderX

more enjoyable overall experience. In addition, we did not include assessments of fine motor skills to better characterize the benefits of the art-making component. As stated above, parsing out the individual effects of the art classes would require a different study design, but would be valuable. Inclusion of a more diverse selection of functional outcomes could potentially answer additional questions about the real-world impact of our program.

Upon examination of our major quantitative findings in Table 3, we observe several effect sizes that might be unexpectedly large, given the small sample size. To derive the effect size, we relied on the "Exact" method given the sample size of 17, which is appropriate. However, we must emphasize the exploratory nature of our findings and encourage the reader to examine them with caution. Future studies will enroll a larger sample to more deeply explore these associations.

In addition, we observed similar changes to our previous study [12] in the Stroop IS. However, we must also interpret this with caution. The IS includes performance on both the word and color scores of the test. Improvement in either of these 2 conditions will produce a higher IS. In this study, we saw improved scores on the incongruent condition, and therefore, are hopeful that our program helped improve interference inhibition in this population.

The purpose of this pilot was to leverage the power of our partnership with the Buffalo-Niagara YMCA to implement a group recreation program that fosters social support and helps preserve cognitive and physical function. From this small pilot, we have developed a larger-scale program involving four YMCA branches and an attention-control group that we anticipate will allow us to more deeply explore the previously described associations and generate a more comprehensive explanatory model. This first pilot of the Movers and Makers program served as an effective stepping stone, allowing us to gather key insights from community members, develop and test our implementation methods, and assess limitations.

Comparison With Previous Work

In examining the existing literature, we have not found any studies that have specifically assessed the implementation and impacts of group-based, SMARTfit exergaming and art making on cognitive and physical function in older adults in a community setting. However, several randomized controlled trials have sought to characterize the effects of combined creative activities and exercise in general on cognition, well-being, quality of life, depressive symptoms, and subjective health outcomes [38-41].

Fundamentally, when we examine the different populations that have been studied in the existing literature, we observe important differences. Our study enrolled community-dwelling older adults, 65 years or older, with at most MCI. Kamegaya et al [38] recruited community-dwelling older adults at risk for cognitive decline without apparent dementia. In this study, a 12-week intervention consisting of aerobic exercise, stretching, and walking, along with games, crafts, and cooking, improved cognition and well-being [38]. In this study, the intervention duration was also 12 weeks, and we also found improvements in cognition, in a similar study population, that did not have overt dementia. Furthermore, Roswiyani et al [41] used a 4-group design (art-only, Qigong-only, art plus Qigong, and control) and 2 training sessions per week for 8 weeks. Results show improved well-being (social relations) among adults (age ≥50 y) assigned to the art-only group. Furthermore, depressive symptoms also declined in the art-only group and in the art plus Qigong group. The Qigong-only group did not demonstrate improvements in any functional outcomes [41]. While the populations among these studies are similar, important differences can be observed in the modes of exercise and art being delivered, with activities ranging from single-task training, like walking [38], to the more complicated dual-task challenges tested in our intervention. This high degree of variability has made it very difficult to synthesize the overall findings of similar studies across the literature.

Other investigators have examined these associations in individuals with dementia. Kang et al [39] explored the effects of a 9-week intervention that consisted of two 3-hour sessions each week, of cognitive stimulation, music, art, horticulture, and exercise. Participants received 30 minutes of cognitive stimulation, followed by 30 minutes of music therapy, a 10-minute break, then 30 minutes of art therapy, 30 minutes of a horticultural activity, and 30 minutes of memory training. Participation in this program among 20 older adults with dementia was associated with improved depression levels, mental-emotional health, and cognitive function. Similarly, Viola et al [40] implemented a 12-week intervention to assess the impact of memory training, computerized cognitive stimulation, writing, and painting, along with walking and stretching on patients with mild Alzheimer disease. These investigators found that participation was associated with improved quality of life and depressive symptoms. Cognitive function did not change in response to participation in the intervention [40]. Differences between these studies can be seen in their effect on cognition, with no improvements found in the study by Viola et al [40]. This highlights an important methodological difference that likely impacts the variability in outcomes we have observed. To date, the optimal dose of exercise and cognitively stimulating activities like creative pursuits that improve cognitive and physical functional outcomes has not been identified. Larger studies are still needed to fully explore these complex associations.

In comparison with previous studies that explore the effects of exercise and creative activities on health and function in aging, our study has some unique attributes. Our study appears to be the first with this focus to include key principles of community-based research to develop and test our intervention activities. Program design and implementation were directly influenced by feedback from older adult community members, creating a more appealing and relevant program that encouraged participation. Our study also integrated a qualitative assessment of the participants' experiences in the intervention program that is being used to design future larger-scale studies. The insights provided by participants have been invaluable in increasing the appeal, usability, and relevance of the program activities.

XSL•FO
RenderX

Conclusions

Our Movers and Makers program is among the first initiatives to combine group-based exergaming and guided artmaking in a community setting to improve key cognitive and physical functional outcomes in older adults. The YMCA's dedication to promoting successful aging makes it an ideal partner in developing and implementing this novel program. Our team was able to create a highly enjoyable and satisfying program, leading to a completion rate of 100%. The participants' positive experience in the program was impacted by the immersive, enriching nature of each activity (exergaming and art) and the high quality of instruction. Through this unique collaboration, we were able to assemble a highly effective interdisciplinary team that was genuinely dedicated to the success of the Movers and Makers program. Team members created a supportive and caring environment that was energizing and inspiring. Furthermore, our approach that emphasized creating a socially engaging atmosphere likely contributed to participants' willingness to complete the program.

We employed a community-integrated approach to designing and implementing our program, which produced an appealing and impactful training experience that is adaptable to meet the needs of different populations in different environments. Future efforts must focus on assessing key implementation outcomes in larger-scale interventions that integrate a control group. These changes will allow our team to generate a more comprehensive explanatory model and identify potential barriers to optimally delivering the Movers and Makers curriculum in diverse community settings with unique strengths and limitations.

Acknowledgments

Our sincerest gratitude to the stellar staff of the Ken-Ton Family YMCA and William-Emslie Family YMCA for their support. In addition, many thanks to Ms Cathi Lamberti and Mr TJ Fairchild for their extensive programmatic and technical support to ensure the Movers and Makers program was conducted effectively and efficiently. Research reported in this publication was supported through a grant from the Health Foundation for Western and Central New York. The research sponsor did not have any role in the design, conduct, analysis, and reporting of study findings.

Data Availability

Data generated from this study, the trial protocol, and analysis plans are available upon reasonable request from the principal investigator. Qualitative data collected is not publicly available to protect the privacy of participants.

Authors' Contributions

ML contributed to data analysis, summary of the results, and manuscript preparation. SO-S contributed to study design and manuscript preparation. SJ contributed to study design and manuscript preparation. NS contributed to study design, data analysis, and manuscript preparation.

Conflicts of Interest

None declared.

References

1. Gopinath B, Kifley A, Flood VM, Mitchell P. Physical Activity as a Determinant of Successful Aging over Ten Years. Sci Rep 2018 Jul 12;8(1):10522. [doi: 10.1038/s41598-018-28526-3] [Medline: 30002462]
2. Fiocco AJ, Yaffe K. Defining successful aging: the importance of including cognitive function over time. Arch Neurol 2010 Jul;67(7):876-880. [doi: 10.1001/archneurol.2010.130] [Medline: 20625097]
3. Depp C, Vahia IV, Jeste D. Successful aging: focus on cognitive and emotional health. Annu Rev Clin Psychol 2010;6(1):527-550. [doi: 10.1146/annurev.clinpsy.121208.131449] [Medline: 20192798]
4. Masika GM, Yu DSF, Li PWC. Visual art therapy as a treatment option for cognitive decline among older adults. A systematic review and meta‐analysis. J Adv Nurs 2020 Aug;76(8):1892-1910. [doi: 10.1111/jan.14362]
5. Cacciata M, Stromberg A, Lee JA, et al. Effect of exergaming on health-related quality of life in older adults: a systematic review. Int J Nurs Stud 2019 May;93:30-40. [doi: 10.1016/j.ijnurstu.2019.01.010] [Medline: 30861452]
6. Ali N, Tian H, Thabane L, et al. The effects of dual-task training on cognitive and physical functions in older adults with cognitive impairment; a systematic review and meta-analysis. J Prev Alzheimers Dis 2022;9(2):359-370. [doi: 10.14283/jpad.2022.16] [Medline: 35543010]
7. Bae S, Lee S, Lee S, et al. The effect of a multicomponent intervention to promote community activity on cognitive function in older adults with mild cognitive impairment: a randomized controlled trial. Complement Ther Med 2019 Feb;42:164-169. [doi: 10.1016/j.ctim.2018.11.011] [Medline: 30670238]
8. Camic PM, Tischler V, Pearman CH. Viewing and making art together: a multi-session art-gallery-based intervention for people with dementia and their carers. Aging Ment Health 2014 Feb 17;18(2):161-168. [doi: 10.1080/13607863.2013.818101]

XSL•FO
RenderX

9. Çetinkaya F, Duru Aşiret G, Direk F, Özkanlı NN. The effect of ceramic painting on the life satisfaction and cognitive status of older adults residing in a nursing home. Top Geriatr Rehabil 2019;35(2):108-112. [doi: 10.1097/TGR.0000000000000208]
10. Dadswell A, Bungay H, Wilson C, Munn-Giddings C. The impact of participatory arts in promoting social relationships for older people within care homes. Perspect Public Health 2020 Sep;140(5):286-293. [doi: 10.1177/1757913920921204] [Medline: 32508239]
11. Jenabi E, Bashirian S, Ayubi E, Rafiee M, Bashirian M. The effect of the art therapy interventions on depression symptoms among older adults: a meta-analysis of controlled clinical trials. J Geriatr Psychiatry Neurol 2023 May;36(3):185-192. [doi: 10.1177/08919887221130264] [Medline: 36165582]
12. Jhaveri S, Romanyk M, Glatt R, Satchidanand N. SMARTfit dual-task exercise improves cognition and physical function in older adults with mild cognitive impairment: results of a community-based pilot study. J Aging Phys Act 2023 Aug 1;31(4):621-632. [doi: 10.1123/japa.2022-0040] [Medline: 36716745]
13. Li Z, Wang T, Liu H, Jiang Y, Wang Z, Zhuang J. Dual-task training on gait, motor symptoms, and balance in patients with Parkinson's disease: a systematic review and meta-analysis. Clin Rehabil 2020 Nov;34(11):1355-1367. [doi: 10.1177/0269215520941142]
14. Sebastião E, Mirda D. Group-based physical activity as a means to reduce social isolation and loneliness among older adults. Aging Clin Exp Res 2021 Jul;33(7):2003-2006. [doi: 10.1007/s40520-020-01722-w]
15. Moody E, Phinney A. A community-engaged art program for older people: fostering social inclusion. Can J Aging 2012 Mar;31(1):55-64. [doi: 10.1017/S0714980811000596] [Medline: 22348701]
16. Xu L, Gu H, Cai X, et al. The effects of exercise for cognitive function in older adults: a systematic review and meta-analysis of randomized controlled trials. Int J Environ Res Public Health 2023 Jan 7;20(2):1088. [doi: 10.3390/ijerph20021088] [Medline: 36673844]
17. Law LLF, Barnett F, Yau MK, Gray MA. Effects of combined cognitive and exercise interventions on cognition in older adults with and without cognitive impairment: a systematic review. Ageing Res Rev 2014 May;15:61-75. [doi: 10.1016/j.arr.2014.02.008] [Medline: 24632497]
18. Jardim NYV, Bento-Torres NVO, Costa VO, et al. Dual-task exercise to improve cognition and functional capacity of healthy older adults. Front Aging Neurosci 2021;13:589299. [doi: 10.3389/fnagi.2021.589299] [Medline: 33679369]
19. Liao YY, Chen IH, Lin YJ, Chen Y, Hsu WC. Effects of virtual reality-based physical and cognitive training on executive function and dual-task gait performance in older adults with mild cognitive impairment: a randomized control trial. Front Aging Neurosci 2019;11:162. [doi: 10.3389/fnagi.2019.00162] [Medline: 31379553]
20. Park H, Park JH, Na HR, et al. Combined intervention of physical activity, aerobic exercise, and cognitive exercise intervention to prevent cognitive decline for patients with mild cognitive impairment: a randomized controlled clinical study. J Clin Med 2019 Jun 28;8(7):940. [doi: 10.3390/jcm8070940] [Medline: 31261796]
21. Bowie CR, Harvey PD. Administration and interpretation of the Trail Making Test. Nat Protoc 2006 Dec;1(5):2277-2281. [doi: 10.1038/nprot.2006.390]
22. Scarpina F, Tagini S. The stroop color and word test. Front Psychol 2017;8:557. [doi: 10.3389/fpsyg.2017.00557] [Medline: 28446889]
23. Ivnik RJ, Malec JF, Smith GE, Tangalos EG, Petersen RC. Neuropsychological tests' norms above age 55: COWAT, BNT, MAE token, WRAT-R reading, AMNART, STROOP, TMT, and JLO. Clin Neuropsychol 1996 Jul;10(3):262-278. [doi: 10.1080/13854049608406689]
24. Collado-Mateo D, Lavín-Pérez AM, Peñacoba C, et al. Key factors associated with adherence to physical exercise in patients with chronic diseases and older adults: an umbrella review. Int J Environ Res Public Health 2021 Feb 19;18(4):2023. [doi: 10.3390/ijerph18042023] [Medline: 33669679]
25. Woolley K, Fishbach A. Immediate rewards predict adherence to long-term goals. Pers Soc Psychol Bull 2017 Feb;43(2):151-162. [doi: 10.1177/0146167216676480] [Medline: 27899467]
26. O'Leary A. Self-efficacy and health. Behav Res Ther 1985;23(4):437-451. [doi: 10.1016/0005-7967(85)90172-X]
27. Farrance C, Tsofliou F, Clark C. Adherence to community based group exercise interventions for older people: a mixed-methods systematic review. Prev Med 2016 Jun;87:155-166. [doi: 10.1016/j.ypmed.2016.02.037] [Medline: 26921655]
28. Idowu MI, Szameitat AJ. Executive function abilities in cognitively healthy young and older adults-a cross-sectional study. Front Aging Neurosci 2023;15:976915. [doi: 10.3389/fnagi.2023.976915] [Medline: 36845657]
29. Lacreuse A, Raz N, Schmidtke D, Hopkins WD, Herndon JG. Age-related decline in executive function as a hallmark of cognitive ageing in primates: an overview of cognitive and neurobiological studies. Philos Trans R Soc Lond B Biol Sci 2020 Nov 9;375(1811):20190618. [doi: 10.1098/rstb.2019.0618] [Medline: 32951543]
30. Eng JJ, Chu KS, Kim CM, Dawson AS, Carswell A, Hepburn KE. A community-based group exercise program for persons with chronic stroke. Med Sci Sports Exerc 2003 Aug;35(8):1271-1278. [doi: 10.1249/01.MSS.0000079079.58477.0B] [Medline: 12900678]
31. Fien S, Linton C, Mitchell JS, et al. Characteristics of community-based exercise programs for community-dwelling older adults in rural/regional areas: a scoping review. Aging Clin Exp Res 2022 Jul;34(7):1511-1528. [doi: 10.1007/s40520-022-02079-y] [Medline: 35152393]

XSL•FO
RenderX

32. Golden CJ. Stroop color and word test: a manual for clinical and experimental uses. : Stoelting Company; 1978.
33. Araujo CG, de Souza e Silva CG, Laukkanen JA, et al. Successful 10-second one-legged stance performance predicts survival in middle-aged and older individuals. Br J Sports Med 2022 Sep;56(17):975-980. [doi: 10.1136/bjsports-2021-105360]
34. Bohannon RW. Six-Minute Walk Test: a meta-analysis of data from apparently healthy elders. Top Geriatr Rehabil 2007 Apr;23(2):155-160. [doi: 10.1097/01.TGR.0000270184.98402.ef]
35. Rapport LJ, Hanks RA, Millis SR, Deshpande SA. Executive functioning and predictors of falls in the rehabilitation setting. Arch Phys Med Rehabil 1998 Jun;79(6):629-633. [doi: 10.1016/S0003-9993(98)90035-1]
36. Tinetti ME, Speechley M, Ginter SF. Risk factors for falls among elderly persons living in the community. N Engl J Med 1988 Dec 29;319(26):1701-1707. [doi: 10.1056/NEJM198812293192604] [Medline: 3205267]
37. Adam CE, Fitzpatrick AL, Leary CS, et al. Change in gait speed and fall risk among community-dwelling older adults with and without mild cognitive impairment: a retrospective cohort analysis. BMC Geriatr 2023 May 25;23(1):328. [doi: 10.1186/s12877-023-03890-6] [Medline: 37231344]
38. Kamegaya T, Araki Y, Kigure H, Yamaguchi H, Long‐Term‐Care Prevention Team of Maebashi City. Twelve‐week physical and leisure activity programme improved cognitive function in community‐dwelling elderly subjects: a randomized controlled trial. Psychogeriatrics 2014 Mar;14(1):47-54. [doi: 10.1111/psyg.12038]
39. Kang HY, Bae YS, Kim EH, Lee KS, Chae MJ, Ju RA. An integrated dementia intervention for Korean older adults. J Psychosoc Nurs Ment Health Serv 2010 Dec;48(12):42-50. [doi: 10.3928/02793695-20100930-01] [Medline: 21053789]
40. Viola LF, Nunes PV, Yassuda MS, et al. Effects of a multidisciplinary cognitive rehabilitation program for patients with mild Alzheimer's disease. Clinics (Sao Paulo) 2011;66(8):1395-1400. [doi: 10.1590/s1807-59322011000800015] [Medline: 21915490]
41. Roswiyani R, Hiew CH, Witteman CLM, Satiadarma MP, Spijker J. Art activities and qigong exercise for the well-being of older adults in nursing homes in Indonesia: a randomized controlled trial. Aging Ment Health 2020 Oct 2;24(10):1569-1578. [doi: 10.1080/13607863.2019.1617239]

Abbreviations

30STS: 30-second Sit-to-Stand test
6MWT: 6-Minute Walk Test
FSBT: 4-Stage Balance Test
IS: interference score
KTY: Ken-Ton Family YMCA
MCI: mild cognitive impairment
MoCA: Montreal Cognitive Assessment
Pcw: predicted color-word score
SCWT: Stroop Color and Word Test
TMT: Trail-Making Test
WEY: William-Emslie Family YMCA

Edited by A Mavragani; submitted 06.03.25; peer-reviewed by RM Glatt; revised version received 11.06.25; accepted 12.06.25; published 29.08.25.

Please cite as:
Satchidanand N, Lawrence M, Jhaveri S, Opalka-Satchidanand S
Development and Testing of an Exergaming and Artmaking Program to Enhance Cognitive and Physical Function in Older Adults: Results of a Community-Based Pilot Study
JMIR Form Res 2025;9:e73555
URL: https://formative.jmir.org/2025/1/e73555
doi:10.2196/73555

© Nikhil Satchidanand, Mia Lawrence, Sameer Jhaveri, Sara Opalka-Satchidanand. Originally published in JMIR Formative Research (https://formative.jmir.org), 29.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Stepped Transition to Employment and Postsecondary Education Success (STEPS) for Adolescents and Adults with Autism: Community Implementation Pilot Trial

Susan W White[1], PhD; Alexis M Brewe[2], PhD; Nicole Powell[1], PhD

[1]Department of Psychology, Center for Youth Development and Intervention, University of Alabama, 200 Hackberry Lane Suite 101, Tuscaloosa, AL, United States

[2]TEACCH Autism Program, Department of Psychiatry, University of North Carolina, Chapel Hill, NC, United States

Corresponding Author:
Susan W White, PhD
Department of Psychology, Center for Youth Development and Intervention, University of Alabama, 200 Hackberry Lane Suite 101, Tuscaloosa, AL, United States

Abstract

Background: Programming to optimize successful transition into adulthood and build skills for independence is consistent with the goal of improving autonomous living among adults with autism, which is a top stakeholder-identified priority. There has been surprisingly little research; however, on structured curricula targeting transition into adulthood.

Objective: This formative community pilot trial of Stepped Transition to Employment and Postsecondary Education Success (STEPS) was designed to test feasibility and effectiveness as implemented by community-based providers and, secondarily, to identify factors that affect implementation.

Methods: This was a 2-phase study. Phase 1 involved engagement with a group of community stakeholders to identify factors likely to influence implementation of STEPS. Phase 2 involved an open pilot trial of STEPS. In the Hybrid Type 1 trial, 24 adolescents and young adults with autism received STEPS in their communities at a local agency unaffiliated with the research study.

Results: Based on stakeholder input (Phase 1), several adjustments were made to the program before implementation (eg, increased attention to building client motivation and clarification of the role of caregivers). Stakeholders and providers indicated that STEPS could be successfully delivered and adopted in the community. From the pilot (Phase 2), results indicate feasibility of study procedures and intervention implementation, supporting future larger-scale implementation. Satisfaction (eg, how helpful and beneficial) with the program was reported as moderate or higher by the participants, and the only 2 drops occurred before the program start. Caregiver-rated transition readiness significantly increased from baseline to end point ($P<.001$), as well as some domains of functional independence (finance management, self-care, and engagement in the community; all $P<.05$). Employment and education status at the end point did not yield a clear pattern indicating a positive or negative impact of the program.

Conclusions: This pilot study supports the feasibility, acceptability, and effectiveness of STEPS as delivered by community providers.

Trial Registration: ClinicalTrials.gov NCT05581069; https://clinicaltrials.gov/study/NCT05581069

(JMIR Form Res 2025;9:e70137) doi:10.2196/70137

KEYWORDS

autism; adult; transition; intervention; community

Introduction

Autism incurs an estimated lifetime per capita societal cost of US $3.2 million, and data suggest that lost productivity and adult care needs are the largest contributors [1]. As of 2020, adults with autism accounted for 35% of the overall autism service use sector; based on changes in rates of diagnosis during childhood, models project that adults with autism will constitute 71% of the cost of care by the year 2060 [2]. Consistent with these projections, stakeholder-identified research priorities place efforts to build skills necessary for adulthood, including seeking employment, and to improve services to support independent living among adults with autism as a top priority [3,4], although addressing lifespan issues like the transition to and supporting needs in adulthood has not historically been a highly active research area [5,6].

XSL•FO
RenderX

Adolescents and adults with autism experience lower subjective quality of life than do age- and cognitive ability-matched peers without autism [7,8], and they face challenges with living independently [9,10]. The majority of adults with autism are not consistently employed, and those who do secure employment are often underemployed, unable to support themselves financially or not working full-time [11,12]. In fact, paid employment has been linked to improved self-esteem and well-being for adults with autism [13]. Furthermore, young adults also experience elevated rates of dysphoric symptoms, including feelings that they do not belong or are a burden to their caregivers, depression, and suicidal ideation [14-16]. Transition-focused approaches serve to bridge the gap between childhood services and the services needed once an individual with autism leaves school, starts working, or requires different supports as they enter adulthood.

Transition planning for adolescents and adults with autism is frequently cited as lacking needed supports and skill-building [17,18], which is exacerbated by the "service cliff" of incredibly limited services and supports available to people with autism and their caregivers once adolescents with autism leave the K-12 school system [19]. The most commonly identified needs and challenges faced by emerging adults on the spectrum are in the areas of interpersonal competence, ability to manage competing demands, self-advocacy, and poor self- regulation [20,21]. Consistent with an experimental therapeutics framework [22], in which proximal processes that maintain the behavioral or clinical problem are targeted in order to bring about downstream changes in behavior or functioning, the Stepped Transition to Employment and Postsecondary Education Success (STEPS) program targets these domains in order to improve preparedness for adulthood among emerging adults with autism [23]. In a previous randomized controlled trial (RCT) with 59 adolescents and adults with autism, conducted in the lab where the program was developed, STEPS participants demonstrated significant improvement in transition readiness, compared with a "transition services as usual" control group, and this improvement was predicted by self-determination. Furthermore, STEPS was highly acceptable to families, feasible to implement with fidelity, and resulted in significantly increased readiness for transition, operationalized as age-typical attitudes and behaviors related to education, social interaction, and functional independence [24]. A recent small pilot of STEPS (n=12 individuals with autism) housed within a community-based nonprofit organization further illustrated high feasibility and acceptability and some individual-level improvement in transition readiness and self-efficacy [25]. Although more community-focused than the original RCT, this latest pilot study focused on additional "proof of concept" of STEPS to address individualized participant goals in employment, education, and daily living, as well as consideration of challenges to future implementation [25]. The important next step is to begin testing community implementation, while continuing to evaluate STEPS' feasibility and clinical impact.

The goal of the present, formative study is to demonstrate feasibility, acceptability, and preliminary impact of community-implemented STEPS on independence and readiness for adult transition in a sample of teens and young adults with autism. This is in line with the guidelines for reporting of nonrandomized pilot studies [26]. Using a hybrid type 1 open trial design [27], we primarily assessed the feasibility, acceptability, and effectiveness of STEPS, with the secondary goal to assess factors that influence implementation, consistent with the Consolidated Framework for Implementation Research (CFIR) [28]. The CFIR is a widely used implementation science framework that helps to elucidate barriers and facilitators to implementation across 5 domains: innovation (eg, the new program), the outer setting (eg, the state or broader system), the inner setting (eg, the delivery agency), individuals (eg, characteristics of providers), and the implementation process (eg, how to plan and deliver the program) [28]. Hybrid trials, which simultaneously measure impact and implementation processes in the target community (rather than in a more controlled lab environment), may reduce the research to practice gap and hasten dissemination.

Methods

Overview

This was a 2-phase study. The goal of Phase 1 was to identify factors that would likely influence effectiveness, adoption, implementation, and eventual sustained use of STEPS. An implementation workgroup comprising community partner agency representatives and other stakeholders in the region (families, educators, and mental health professionals) met 3 times, before the trial's launch, to elicit stakeholder thoughts and experiences relevant to the goal of optimizing program sustainability within community settings. Meetings were conducted over Zoom (Zoom Video Communications, Inc), each lasting 1 hour during which the research team presented informational slides, introduced the topics for discussion, and explained the relevance of the topics to the project. Meeting discussions included the full group and were led by the first author, covering implementation-related topics, guided by CFIR constructs. These included the "innovation" (eg, advantages and adaptability of STEPS), "outer setting" (eg, partnerships, policies, and financing), "inner setting" (compatibility and available resources), and "individuals" domains (eg, innovation deliverers). Meeting 1 focused on anticipated challenges in implementing a program targeting adult transition; meeting 2 addressed access to programming through schools and agencies, and ensuring program implementation fidelity; and meeting 3 covered engaging community champions, training needs for clinicians, and reaching underserved communities. Phase 2 involved the implementation pilot of STEPS in the community, which included 5 nodal assessments at eligibility, baseline, 7-week midpoint, end point at week 14, and follow-up 6 months after end point.

Sample

Phase 1 individuals were recruited from the partnering agency's network of professional connections, and the agency's director facilitated their contact with the research team. Phase 1 participants were drawn from the study's target location (predominantly rural area of the southeastern United States) and were recruited based on their knowledge and experience with regional autism services for adolescents and adults. Phase

XSL•FO
RenderX

1 participants provided informed consent at the time of enrollment.

Eligibility criteria for participation in Phase 2 included a previous autism diagnosis, confirmed by participants and their caregivers, and sufficient verbal skills to participate in STEPS, based on caregiver reports during the eligibility phone screening. There was no IQ cut-off to participate. Approximately 20% of the sample had a caregiver-reported previous diagnosis of intellectual disability. Participants were recruited from predominantly rural areas in the southern United States, using the partnering agency's website and client list (see below for more information on the agency partners), supplemented by local events (eg, autism walks) and social media, and participants included both existing and new clients. Participants received general information about the study through personal contact with the partnering agency's director or affiliated clinicians. If interested, the participant or caregiver completed a brief phone screen with the project coordinator. If eligible, the participant was assigned to one of the participating clinicians, according to a pre-established randomization process. When the provider had room in their caseload, they signaled the project coordinator, who facilitated scheduling. Phase 2 participants (TAY and caregivers) provided informed consent at the time of enrollment.

The partnering agency delivering STEPS was a community mental health clinic that accepted both public and private insurance. They did not specialize in autism, although some of the providers had previous clients with autism and served a diverse clientele. Participating providers were recruited from the partnering agency's affiliated clinicians. Agency providers all had doctoral degrees (PsyD-2; PhD-2) and included 2 males and 2 females. The agency director facilitated initial contact between the providers and the research team, and coordinated online and in-person meetings in which the details of the study were explained and training in STEPS was provided (see below). All providers completed informed consent before their participation.

STEPS Program

Described in detail elsewhere [23], STEPS is a modular cognitive-behavioral curriculum comprised of 12 content modules, meant to be delivered over a series of approximately 12 weekly sessions, one-on-one with a client. Informed by research and stakeholder input, STEPS was developed to facilitate the successful transition to adulthood among adolescents and young adults with autism by promoting essential skills for adult autonomy, including self-knowledge, self-determination, and self-regulation. In addition to the individual sessions, STEPS includes in vivo skills practice opportunities (termed "immersion"). Of note, in this trial, participants had one immersion during the program; based on consumer feedback, STEPS now includes 2 immersion sessions. The providers were trained during a full-day on-site workshop, including didactic presentation and role-play, before the start of the trial. They also participated in weekly supervision meetings with the lead author, via videoconference, throughout the trial. During these meetings, real-time program adaptations for clients were discussed, as well as troubleshooting to optimize applicability and personalization. This integrated training approach is similar to that used in other effectiveness trials with community agencies [29]. The qualitative feedback we will receive from providers will boost readiness for broad dissemination in the community.

Measures

The following measures were used in Phase 2 only. Consistent with previous research on metrics of feasibility and acceptability of interventions in clinical trials [30,31], feasibility of STEPS (ie, how doable the intervention was in the community) was operationalized as high fidelity to the STEPS manual, as reported by providers and confirmed by an independent rater; at least moderate provider-reported overall feasibility; and at least moderate client-reported therapeutic alliance. These aspects of treatment feasibility were selected to measure both providers' beliefs and behavior in delivering STEPS appropriately (ie, reported feasibility and fidelity in comparison to observer-rated fidelity), as well as whether delivery of all treatment objectives could still result in a good therapeutic relationship, given research demonstrating that feasibility consists of multiple components [30,32]. Acceptability of STEPS (ie, consumer and provider response to the intervention) was operationalized as low attrition, at least moderate provider-reported overall acceptability and intervention appropriateness, and at least moderate client- and caregiver-reported satisfaction ratings. These metrics were chosen to reflect clients', caregivers', and providers' beliefs about STEPS following the trial, as well as behavioral data on whether clients decided to complete STEPS [31]. Implementation was assessed via qualitative feedback from providers. Consistent with the conceptual framework of STEPS (ie, targeting self-knowledge, self-determination, and self-regulation to exert sustained change in readiness for transition to adulthood and independent living skills), our primary outcome measures assess transition readiness and functional adult living skills.

Client- and Caregiver-Completed Measures

American Institutes for Research–Self-Determination Scale (AIR-SD) assesses opportunity (ie, 12 items) and capacity (ie, 12 items) to choose activities and behave in goal-directed ways, key facets of self-determination. Participants completed the full self-report version, whereas caregivers completed the proxy report, which consisted of the full AIR-SD minus 6 capacity questions that measured their adult child's internal perceptions and knowledge. We adapted the opportunity questions about school to include school, work, and other environments participants are regularly in outside of their home to more accurately reflect the age and goals of our sample. The AIR-SD is strong psychometrically (α=.95) and the 2-factor structure has been replicated in autism [33,34]. It is sensitive to treatment change, including in autism and in the original STEPS trial [35-37]. In the current sample, internal consistency was good for self-report (α ranged from .94-.96). Caregiver-reported capacity at end point had lower internal consistency (α=.45); otherwise, internal consistency was also good for caregiver reports (α ranged from .81-.92). The AIR-SD Opportunity and Capacity scores were used as a measure of SD. The AIR-SD

XSL•FO
RenderX

was completed by participants and their caregivers at baseline, the 7-week midpoint, and the end point.

Emotion Dysregulation Inventory (EDI) [38] involves caregiver completion of the Reactivity-Short Form (7 items) and EDI Dysphoria (6 items), and participants completed an alpha version of the self-report EDI that consisted of the caregiver-report form reworded to person-first language. Both forms were used as a facet of self-regulation, in addition to the General Self-Efficacy Scale (GSES; described below), and assess the intensity and frequency of emotion dysregulation in an individual over the past 7 days [38]. The EDI captures 2 domains of dysregulation: reactivity, which is characterized by rapidly escalating negative emotions, and dysphoria, which is characterized by low positive affect or a motivation and a general sense of unease. The items are rated on a Likert scale, with higher scores indicating more emotion dysregulation. Participants and their caregivers completed the EDI at baseline, 7-week midpoint, and end point. The *t*-score conversions were completed based on norms generated from a sample of 1755 caregivers of youths with autism (ie, "autism norms"); self-report scores were also converted based on proxy report norms given the use of the alpha version of the self-report scale. The EDI has shown good construct validity and sensitivity to change [38]. Internal consistency ranged from .86-.97 for self-report and .59-.97 for caregiver-report.

General Self-Efficacy Scale (GSES) [39] is a 10-item form completed by participants and their caregivers at baseline, midpoint, and end point. It measures participants' self-efficacy, or their belief in their ability to deal with stressful situations effectively. Higher scores reflect increased self-efficacy. The GSES has shown strong internal consistency and convergent validity [40,41]. Internal consistency in the current study was similarly good for self- and caregiver-report (*a* ranged from .79 to .98). The GSES Total Score was used as our measure of self-regulation.

Integrative Self-Knowledge Scale (ISKS) [42] is a 13-item self-report measure of both experiential and reflective self-knowledge, or understanding one's experience over development (past, present, and future) in relation to one's goals and desired outcomes. Higher scores reflect increased self-knowledge. The ISKS has strong internal consistency and convergent and discriminant validity [42]. Internal consistency in the current sample was also good (α ranged from .82-.92). The ISKS total score was used to measure self-knowledge. The ISKS was completed by participants at baseline, the 7-week midpoint, and the end point.

Program Satisfaction Scale (PSS) is a self- and caregiver-report form adapted from the Consumer Satisfaction Survey [43] to measure satisfaction with STEPS (ie, intervention acceptability). The PSS assesses satisfaction on several metrics, including helpfulness, relevancy of content and materials, satisfaction with participants' progress, acceptability of format, and likelihood of referring another person to STEPS. Satisfaction with STEPS was defined as having an average PSS score of >3, which indicates moderate helpfulness, impact, and acceptability. On the PSS, scaling is as follows: 1=no or not at all; 3=moderate or somewhat; 5=very (helpful, beneficial, satisfied). The PSS was completed by participants and their caregivers at end point only.

Rehabilitation success is a measure that captures participants' engagement in employment and education. Caregivers completed the form at baseline, end point, and 6-month follow-up by indicating yes or no about whether their adult or adolescent with autism was employed and engaged in educational or training programs (including high school and postsecondary education) at each time point.

Relationships, Employment, Autonomy, and Life Satisfaction (REALS) [44] is a compilation of scales that measure multiple facets of independence and adult life for adults with autism and adults with intellectual disability (ID). It has a self-report and proxy-report form. The REALS measures support, or the level of support an adult with autism needs to perform certain tasks, and frequency, which is how often the participant engages in those tasks. The REALS provides theta scores for support and frequency in the domains of social activity, work readiness, work performance, home care, self-care, leisure, mobility, finances, and health (including sleep, diet, and exercise; proxy report only). The self-report form also includes three satisfaction scales on social activity satisfaction, autonomy satisfaction, and work or school satisfaction. REALS psychometric testing is currently ongoing; in the current study, internal consistency was adequate for self-report (*a* ranged from .70-.98) and ranged from weak to strong for caregiver-report (*a* ranged from .47-.95). The REALS was used as a primary outcome measure. The REALS was completed by participants and caregivers at baseline, end point, and 6-month follow-up.

Seven Component Self-Determination Skills Survey (SCSDSS) [45] is a self- and caregiver-report of self-determination. The form provides total scores for performance and perceived importance of 7 skills involved in self-determination including choice-making, decision-making, goal setting, problem solving, self-advocacy and leadership, self-awareness and self-knowledge, and self-management. The SCSDSS also includes a 6-item version of the AIR Self-Determination Scale [46], which measures overall self-determination. Internal consistency was adequate for ratings of importance (*a* ranged from .70-.90), performance (*a* ranged from .72-.90), and overall self-determination (*a* ranged from .69-.90) across both caregiver- and self-report. All 3 total scores (importance, performance, and overall self-determination) at baseline, 7-week midpoint, and end point were used to examine change in self-determination as a mechanism of STEPS.

Transition Readiness Scale (TRS) [47] is a 30-item self- and caregiver-report measure of transition readiness. It was initially developed for the original STEPS RCT to measure readiness to enter college; however, it was adapted to measure general readiness for the transition to adulthood, including meeting one's goals in postsecondary education or employment. Higher scores reflect more transition readiness. The TRS has demonstrated strong internal consistency, good item-level analyses, and discriminant and concurrent validity [47,48]. In the current study, internal consistency was good for both self- and caregiver-report (*a* ranged from .76-.91). The TRS Total Score was a primary outcome measure. The TRS was completed

XSL•FO
RenderX

by participants and caregivers at baseline, 7-week midpoint, and end point.

Working Alliance Inventory–Short Revised, client version (WAI-SR) is a 12-item self-report measure of participants' experience of the client-therapist alliance (ie, a metric of intervention feasibility). The items cover 3 aspects of working alliance, including goal (ie, agreement on clients' presenting problems and goals for therapy), task (ie, agreement on how to achieve goals and willingness to work together toward those goals), and bond (ie, the affective relationship between client and therapist). Higher scores reflect stronger alliance. The WAI-SR has strong internal consistency and construct validity with other alliance measures [48,49]. Internal consistency was high in the current sample (α=.96). The WAI-SR was completed by participants at the 7-week midpoint only.

Provider-Completed Measures

Session fidelity ratings were completed by providers following each STEPS session to capture providers' adherence to the manual in delivering session objectives (a metric of treatment feasibility). Therapists reported whether or not each session objective was met. Therapist-reported fidelity was also confirmed by independent raters (i.e, trained research assistants) who co-coded 10% of all STEPS sessions for fidelity to the manual. Independent raters were trained to rate session fidelity by receiving an overview of STEPS, including an explanation of treatment objectives and practice coding sessions. To establish reliability, independent raters coded randomly selected session videotapes until they achieved exact agreement across 3 consecutive videos with the study authors. After reliability was achieved, the independent raters co-coded a randomly selected 10% sample of the total videotapes for each participant.

Implementation interviews were completed with each provider at the end of the trial to collect their feedback on implementation barriers and facilitators, based on their experience in delivering STEPS. Each one-on-one interview was brief (< 1 hour) and consisted of the same set of 21 questions covering each of the CFIR domains: innovation, inner setting, outer setting, individual domain, and Implementation Process [28]. After introducing each domain, a series of questions related to both hindrances and facilitators to the implementation of the program within that domain were posed. Examples of interview questions include "To what degree did client motivation affect use?"; "What funding considerations need to be thought of?"; and "What provider characteristics or skills are important?".

Implementation surveys included the Feasibility of Implementation Measure (FIM), Acceptability of Implementation Measure (AIM), and Intervention Appropriateness Measure (IAM) [50]. These 3 measures, comprised of 4 items each, are rated by providers to assess their perceptions of acceptability, appropriateness, and feasibility of adopting STEPS. The implementation surveys all demonstrate strong internal consistency and sensitivity to change [50]. Providers completed the implementation surveys immediately following completion of STEPS training and following completion of all STEPS cases for the trial. A score of ≥3 reflects moderate acceptability, appropriateness, and feasibility.

Community Involvement

Phase 1 included several people with autism and advocates in the community (e.g., parents) who attended and contributed to the implementation workgroups (described above), sharing their personal experiences, thoughts, and opinions about services. In addition, the development of the STEPS program relied heavily on input from the autistic community [51]. To ensure scalability of the program and adoption in community agencies, providers and leadership from our partner agency were also included in Phase 1. This was critical for considerations such as training, billing, and provider qualifications.

Data Analysis

The Phase 1 workgroup meetings, designed to ensure successful delivery of the program in Phase 2, were all recorded and transcribed to allow for thorough and accurate review. Given the exploratory nature of Phase 1 and the implementation interviews, as well as the fairly small samples, a structured analytic approach [52] was not deemed necessary. The lead author, who has previously led such qualitative endeavors, led the workgroups and conducted the interviews. An informal qualitative content analysis approach was used for qualitative examination of content [53]. Specifically, for each meeting, participants' responses were individually examined, categorized according to the emergent topic areas (see results below), and considered collectively to summarize the results.

Phase 2 quantitative data were analyzed using SPSS (version 29.0; IBM Corp). Before any analyses, the data were cleaned, checked for missing data, and checked for assumptions. The amount of missing data varied across participants. Out of the 24 participants enrolled in the study, only 2 participants and 3 caregivers completed all data at every study time point. The highest missing data were found at the end point (23% of completers and 27% of their caregivers missing at least some data) and 6-month follow-up (59% of completers and 36% of their caregivers missing some or all their data). Preliminary analyses were conducted to examine if there were demographic differences in participants and caregivers who completed the end point versus not. There were no significant differences in demographics between those who did and those who did not complete the end point.

Descriptive analyses of all variables were run to characterize Phase 2 participants. Frequency and descriptive analyses were used to examine metrics of acceptability (ie, attrition rate, program satisfaction ratings, and providers' ratings of acceptability and appropriateness) and feasibility (ie, provider- and observer-rated fidelity to the manual, provider-reported overall feasibility, and client-reported therapeutic alliance). Qualitative responses from interviews with providers during Phase 2 were informally reviewed for common themes and grouped under the CFIR domains to consider implementation factors. Given there were only four providers, a more rigorous approach such as thematic analysis is not viable. Descriptive analyses were also used to characterize outcomes on the Rehabilitation Success form. Clinical impact was tested in two ways. Given our small sample size and the high amount of missing data for participants across time points, we used paired samples *t* tests to examine statistical change in main outcomes

XSL•FO
RenderX

(TRS, REALS) from (1) baseline to end point and (2) baseline to 6-month follow-up. Similarly, we examined change in mechanisms (ISKS, AIR-SD, SCSDSS, GSES, and EDI) from (1) baseline to 7-week midpoint and (2) baseline to end point. A significance threshold of P=.05 was used, without correction, given the pilot nature of the trial. We also examined clinically significant change, operationalized as movement of >0.5SD on the TRS from baseline to end point. Change was examined in both directions to identify positive and negative responders to STEPS.

Ethical Considerations

This study received full ethics review board approval from the University of Alabama Institutional Review Board (IRB Protocol numbers: 21-12-5245; 21-12-5231), and all participants and caregivers provided written informed consent, with ample time to ensure all questions were answered, before any data collection. All data were deidentified to protect confidentiality. Participants received US$25 per assessment (5 in total; up to US$125 for full participation in the assessments). There was no payment, nor cost incurred, for receiving the program.

Results

Phase 1

Phase 1 participants comprised 16 individuals (4 male, 12 female), including parents (2), a young adult with autism (1), administrators at agencies serving individuals with autism (5), school system administrators (3), a university support services representative (1), and mental health providers (4). Topics explored in the stakeholder meetings included identification of potential implementation challenges ("Some [autistic adolescents and adults with autism] may not see the need to do the program, even though we see they need it, and because they're adults, they choose to not want to"), reaching and retaining families in the program ("…many drop off our radar and we don't have the resources to follow up with them"), training issues ("One of the challenges is the training of the clinicians to make sure that you reach that reliability, which is important. So, you don't have one clinician doing one thing and another doing another"), reaching underserved and disadvantaged populations ("a number of these kids have Medicaid services…oftentimes those [care management] companies will send out bulletins to providers and to members so that might be something to consider"), and funding issues ("… in the adult service world, funding is so limited, and you have to be creative. It is a large barrier"). Based on stakeholder input, several adjustments were made to the STEPS program before implementation. The manual was edited to encourage incorporation of special interests to increase participant motivation, more visuals were incorporated into program materials and handouts, notes and tips were added to assist clinicians in anticipating common implementation issues, and more diverse examples were provided for immersion activities. An introductory section was also added to clarify the role of caregivers in the program (ie, that the "client" is the teen or adult with autism, with caregivers in a supportive, but peripheral, role). Please see Table 1 for additional quotations from Phase 1 meetings.

Table . Quotations from Phase 1 meetings.

Topic	Quotation
Potential implementation challenges	"... you would need less verbal and more visual hands-on opportunities to practice skills."
Reaching and retaining families in the program	"Once the clinicians have this tool and they are working through their regular client systems of referrals from pediatricians and family practice physicians, the awareness piece is going to be community based in the school system and then hopefully, it's self-sustaining."
Training issues	"Make sure that the providers feel supported, and that we have resources when we have questions in terms of the training. An easy route to get clarification, further training, or materials would be important as well."
Reaching underserved and disadvantaged populations	"I think we need to look at faith-based organizations ... we tend to have a significant network across faith-based organizations because there is nothing else up here at times, especially for the low-income (families)."
Funding issues	"Medicaid waivers. If that's (STEPS) something that we can get on the Medicaid Waiver list."

Phase 2

Phase 2 participants included 24 adolescents and young adults with autism aged 16 - 29 years old and their caregivers. Mean age of participants was 21.33 (SD 4.19) years. Frequency analyses of participants showed that the sample was predominantly male (79%), White (96%), and not Hispanic or Latino (96%). Approximately 20% of the sample had a caregiver-reported previous diagnosis of intellectual disability. About 33% of the sample had a job and 43% were engaged in education or training at baseline. See Table 2 for full Phase 2 participant demographics.

XSL•FO
RenderX

Table . Participant demographics and baseline data (N=24) for Phase 2 (Hybrid Type 1 open trial).

Cognitive level	Participants (N=24)
Intellectual disability diagnosis, n (%)	
Yes	5 (21)
No	13 (54)
Not Sure	6 (25)
Sex, n (%)	
Male	19 (79)
Female	4 (17)
Nonbinary	1 (4)
Race, n (%)	
Black or African American	1 (4)
White	23 (96)
Ethnicity, n (%)	
Hispanic or Latino	1 (4)
Not Hispanic or Latino	23 (96)
Annual household income (US$)	
25,000-49,999	3 (13)
50,000-99,999	8 (33)
Over 100,000	4 (17)
Prefer not to answer	9 (38)
Rehabilitation success, n (%)	
Had a job at baseline (yes)	7 (33)
Engaged in education or training at baseline (yes)	9 (43)
Participant age (years), mean (SD)	21.33 (4.19)
SCQ[a] total score, mean (SD)	
Participant	__[b]
Caregiver	16.04 (6)

[a]SCQ: Social Communication Questionnaire (total score possible range: 0 - 39).
[b]Not available.

Acceptability and Feasibility

Results supported acceptability of STEPS delivered in the community, evidenced by low program attrition and moderate to high program satisfaction ratings. Of the 24 participants enrolled in the STEPS trial, 22 completed the program (92% retention). Both “non-completers” dropped from the study before attending any STEPS sessions. One participant became unresponsive before scheduling sessions with their assigned therapist, and the other participant opted to withdraw before starting STEPS due to their discomfort with sessions being recorded for the study.

Participants provided ratings indicating moderate to high satisfaction on the PSS (ie, >3 on each item), suggesting that they found STEPS to be moderately helpful (mean 3.71, SD 1.26), beneficial (mean 3.47, SD 1.46), and acceptable to complete (mean 4.12, SD .99). Participants also reported moderate overall satisfaction with STEPS (mean 3.71, SD 1.21) and moderate likelihood to recommend STEPS to a peer with autism (mean 3.71, SD 1.40). Similarly, parents provided moderate to high satisfaction ratings, suggesting they found STEPS to be helpful (mean 4.19, SD .83), beneficial (mean 3.88, SD 1.03), and acceptable for their adolescent or young adult with autism to complete (mean 4.44, SD .81). They also reported high overall satisfaction with STEPS (mean 4.12, SD 1.09) and a high likelihood to recommend STEPS to other caregivers of adolescents and young adults with autism (mean 4.31, SD 1.08).

Both after they completed the initial training and after completing their STEPS cases, providers’ ratings on the AIM and IAM (ie, >3 or higher suggests moderate ratings) suggest they found STEPS acceptable as an intervention (posttraining mean 4.69, SD .63; post-STEPS cases mean 4.44, SD .66) and

XSL•FO
RenderX

appropriate to their clients (posttraining mean 4.56, SD .88; post-STEPS cases mean 4.88, SD .25).

Results indicate that STEPS was feasible to implement. Fidelity ratings completed by STEPS providers indicated that 96% of STEPS treatment objectives were delivered as intended across 264 sessions. Providers delivered all objectives designated for a specific session approximately 86% of the time. Approximately 10% of sessions (n=29) were double coded for fidelity to the manual by trained research assistants. These double-coded sessions indicated that 83.19% of STEPS treatment objectives were delivered as intended. Both after they completed the initial training and after completing their STEPS cases, providers' ratings on the FIM (ie, >3 or higher suggests moderate ratings) suggest they found STEPS feasible to implement (posttraining mean, 4.38, SD 0.32; post-STEPS cases mean, 4.94, SD 0.13). In addition, participants reported strong therapeutic alliance with their providers (mean, 50.21, SD 9.94, range 12 - 60).

Implementation

Qualitative data from the provider interviews is organized according to each CFIR domain [28]. Within Innovation (eg, the experimental program; STEPS), providers identified challenges associated with the time to prepare sessions and finding appropriate immersion activities for families in rural areas, and one provider acknowledged their own anxiety related to conducting immersion. For example, one provider commented, "It's hard to do this [immersion] as a clinician, but for the patients, it was such an impactful part." They also indicated they thought it could be readily used and billed for, given the CBT focus. One provider reported, "The program itself is very usable and approachable." And another stated, "The language is respectful of the autistic community, it fits a variety of clients." With respect to the outer setting domain (eg, the external systems), providers described the importance of family or caregiver engagement to success, and that a manualized program like STEPS could help make services more consistent for school districts. Considering the inner setting domain (eg, the setting wherein intervention is delivered), providers stated that STEPS could be used in an agency that did not specialize in autism, but that in-house expertise helped them, and that the setting must be flexible to address the diverse needs of the clients. With respect to the Individuals domain (eg, those people responsible for implementation), there was mention of helping parents find a good balance between "infantilizing" their grown child and not being involved enough; providers also indicated that the manual would be appropriate even for a clinician who does not normally work with patients with autism. For implementation (eg, those activities involved in actual implementation), several providers indicated that having strong partnerships (eg, with state vocational rehabilitation and local employers) would help, as well as ensuring that more people in the community were sensitive to their clients' needs. One provider commented, "Having partnerships is really helpful, for example, having an external person do a mock interview for immersion. Also, it's important to have employers on board—sensitive to clients' needs and strengths."

Primary Outcomes

Although not statistically significant at the P=.05 level, participants reported increased transition readiness (on TRS) at end point compared to baseline (t_{19}=−1.91; P=.07). Caregivers also reported significant increased transition readiness by end point for their adults with autism (t_{15}=−4.19; P<.001). Clinically significant changes, operationalized as movement of >0.5 SD in TRS scores from baseline to end point, were examined. Eleven participants (55% of participants who completed end point) demonstrated clinically significant improvement on the TRS and were considered "treatment responders." Four participants (20%) demonstrated clinically significant worsening on the TRS. Twelve caregivers (75% of caregivers who completed end point) reported that their adult or adolescent with autism had clinically significant improvement on the TRS from baseline to end point and were considered "treatment responders." No participants significantly worsened, based on caregiver report.

On the REALS, participants and caregivers only reported increases in functional independence in a few areas, and these varied by reporter. By end point, participants reported engaging in significantly more financial tasks (t_{14}=−2.21; P=.04) than at baseline. This improvement persisted six months afterward; at follow-up, participants were engaging in significantly more financial tasks (t_9=−2.47; P=.04) compared with baseline, as well as more social (t_9=−2.27; P=.0549) and self-care activities (t_9=−2.62; P=.03). At the end point, caregivers reported more engagement in the community (t_{15}=−3.31; P=.005) and needing less support to do so (t_{15}=−3.21; P=.006) compared with baseline. At follow-up, caregivers reported a decreased need for support to apply for and get a job (t_7=−12.31; P<.001), successfully perform at work or school (t_{11}=−7.22; P<.001), and complete household tasks (t_{13}=−4.90; P<.001). Caregivers also reported heightened engagement in tasks to succeed at work or school at follow-up compared with baseline (t_{11}=−4.27; P=.001).

Considering the Rehabilitation Success form, 7 participants were employed at baseline. At the end point, only 4 retained their employment, and 3 additional people gained employment (n=7 at end point). By follow-up, the 4 participants employed at baseline still retained their employment, 2 of the 3 additional participants employed by end point retained their employment, and one additional participant was employed (n=7 employed at follow-up). At baseline, 9 participants were engaged in education. At the end point, 7 were still involved in education, and 6 additional participants started educational activities (n=13 at end point). At follow-up, 8 of 13 participants from the end point were still involved in education, and 2 additional participants had started educational activities (n=10 at follow-up).

Secondary Outcomes

On the ISKS, participants did not report significantly more self-knowledge at midpoint or end point compared with baseline. On the AIR-SD, participants reported significantly increased capacity to be self-determined at both midpoint (t_{19}=−2.52;

XSL•FO
RenderX

P=.02) and end point (t_{15}=−3.41; P=.004) compared with baseline. They did not report significant increases in opportunities to be self-determined at midpoint or end point. Caregivers reported their adults with autism having significantly more capacity to be self-determined at end point compared with baseline (t_{15}=−3.42; P=.004). There were no significant differences between baseline and end point. Caregivers also reported increased opportunities for their young people with autism to be self-determined at both midpoint (t_{17}=−3.32; P=.004) and end point (t_{15}=−4.73; P<.001) compared with baseline.

On the SCSDSS, participants reported significantly increased beliefs that the 7 skills involved in self-determination were important for their lives by midpoint (t_{19}=−2.16; P=.04) and end point (t_{15}=−2.18; P=.0546) to baseline. They also reported significantly increased performance in the 7 skills involved in self-determination at both midpoint (t_{19}=−2.15; P=.04) and end point (t_{15}=−3.24; P=.006) compared to baseline, as well as increased overall self-determination at end point compared to baseline (t_{15}=−2.57; P=.02). Caregivers similarly reported significantly increased overall self-determination at both midpoint (t_{17}=−3.86; P=.001) and end point (t_{15}=−7.48; P<.001) compared with baseline. They also reported increased performance in the seven skills involved in self-determination at the end point (t_{15}=−4.64; P<.001). Caregivers did not report that their beliefs about the importance of the seven skills significantly changed from baseline.

On the GSES, participants reported feeling increased self-efficacy at both midpoint (t_{19}=−2.52; P=.02) and end point (t_{16}=−2.82; P=.01) compared with baseline. Caregivers similarly reported their young people with autism having significantly more self-efficacy at end point than baseline (t_{15}=−7.52; P<.001). They did not report significant change between baseline and midpoint. On the EDI, participants reported experiencing significantly less dysphoria at end point compared to baseline (t_{16}=3.19; P=.006). There was no change between participant-reported dysphoria at baseline and midpoint. Participants did not report any change in reactivity from baseline to midpoint and end point. Caregivers did not report any significant change in dysphoria or reactivity. See Table 3 for reported means of paired samples t tests for primary outcomes and mechanisms from baseline to end point.

XSL•FO
RenderX

Table . Phase 2 treatment outcomes at baseline to end point.

Primary outcomes	Baseline, mean (SD)	End point, mean (SD)
Participant-reported		
Transition readiness scale[a]	75.45 (15.87)	81.55 (10.18)
REALS[b]		
Social activities support	−.36 (.98)	.09 (1.14)
Social activities frequency	−.46 (.79)	−.13 (.96)
Work readiness support	−.03 (.53)	.06 (.97)
Work readiness frequency	−.29 (.50)	−.22 (.73)
Work performance support	−.27 (1.14)	−.18 (.72)
Work performance frequency	−.35 (.73)	−.10 (.99)
Self-care support	−.42 (.85)	−.18 (.89)
Self-care frequency	−.49 (.63)	−.20 (1.03)
Leisure support	.06 (1.00)	.04 (1.02)
Leisure frequency	.22 (.68)	.39 (.75)
Mobility support	−.64 (1.00)	−.40 (1.05)
Mobility frequency	−.73 (.93)	−.31 (.94)
Home care support	−.01 (.70)	−.07 (.87)
Home care frequency	−.10 (.83)	−.11 (.83)
Finances support	−.42 (1.01)	−.13 (1.16)
Finances frequency	−1.42 (.96)	−.51 (1.19)[c]
Social satisfaction	.10 (.92)	.59 (.87)
Autonomy satisfaction	.25 (.85)	.61 (.92)
Work/school satisfaction	−.43 (.83)	−.29 (.86)
Caregiver-reported		
Transition readiness scale	69.44 (11.79)	81.88 (8.29)[d]
REALS		
Social activities support	.29 (.85)	.43 (.81)
Social activities frequency	−.07 (.63)	.01 (.87)
Work readiness support	−.05 (.24)	.18 (.34)
Work readiness frequency	−.35 (.61)	−.32 (.61)
Work performance support	−1.20 (.51)	−.93 (.47)
Work performance frequency	−1.07 (.49)	−.71 (.39)
Self-care support	.13 (.54)	.26 (.48)
Sleep, diet, exercise support	−.07 (.72)	.03 (.93)
Leisure support	.47 (.57)	.55 (.60)
Leisure frequency	.03 (.75)	.32 (.80)
Mobility support	−.05 (.85)	.32 (.80)[c]
Mobility frequency	−.57 (.77)	−.17 (.87)[c]
Home care support	−.48 (.52)	−.29 (.39)
Home care frequency	−.80 (.72)	−.47 (.58)
Finances support	.26 (.71)	.32 (.61)
Finances frequency	−.53 (.42)	−.41 (.41)

XSL•FO
RenderX

Primary outcomes	Baseline, mean (SD)	End point, mean (SD)
Secondary outcomes		
Participant-Report		
ISKS[e]	27.35 (9.84)	29.53 (12.13)
AIR-SD[f] capacity	36.38 (13.06)	42.44 (9.95)[c]
AIR-SD opportunity	41.25 (10.54)	45.38 (10.66)
SCSDSS[g] importance	9.50 (4.16)	11.13 (3.22)[c]
SCSDSS performance	6.19 (3.83)	8.94 (4.19)[c]
SCSDSS self-determination	15.19 (6.42)	18.19 (4.65)[c]
GSES[h]	26.88 (11.45)	32.29 (13.07)[c]
EDI-dysphoria[i]	45.72 (9.27)	40.46 (6.23)
EDI-reactivity	37.81 (8.45)	35.27 (9.11)
Caregiver-Report		
AIR-SD capacity	16.19 (3.53)	19.63 (1.89)[c]
AIR-SD opportunity	42.00 (6.11)	47.69 (5.55)[d]
SCSDSS importance	12.81 (1.33)	12.44 (1.90)
SCSDSS performance	4.25 (2.14)	8.00 (3.18)[d]
SCSDSS self-determination	11.83 (2.94)	13.94 (2.01)[d]
GSES	26.50 (4.29)	32.19 (4.81)[d]
EDI dysphoria	44.48 (5.06)	41.61 (4.97)
EDI reactivity	37.44 (6.37)	34.81 (6.07)

[a]Transition Readiness Scale (total score; possible range: 30 - 150).

[b]REALS: Relationships, Employment, Autonomy, and Life Satisfaction (theta scores; theta ranges differ for each domain, but within the range: -2.83 - 2.40).

[c]*P<.05

[d]**P<.001

[e]ISKS: Integrative Self-Knowledge Scale (total score; possible range: 0 - 48).

[f]AIR-SD: American Institutes for Research–Self-Determination Scale (domain scores; possible range for each domain: 12 - 60 except for caregiver-report capacity, which is 6 - 30).

[g]SCSDSS: Seven Component Self-Determination Skills Survey (domain scores; possible range 6 - 18 for importance and performance, 7 - 35 for self-determination).

[h]GSES: General Self-Efficacy Scale (total score; possible range: 10 - 50).

[i]EDI: Emotion Dysregulation Inventory (T-scores; possible range: 26.10 - 79.20).

Discussion

Principal Findings

There is a large and growing population of adults with autism who could benefit from access to evidence-based services and supports. Programs that support transition to adulthood and that address issues related to quality of life and goal attainment could have considerable impact. STEPS is such a program, previously tested in both a preliminary open trial [54] and an RCT [24].

In this hybrid type 1 open trial of STEPS delivered in a community agency by clinicians uninvolved in research, we found strong evidence for acceptability to both consumers and providers and feasibility of delivery. In terms of clinical impact, significant improvements were seen in readiness for transition and some domains of functional independence. The findings with respect to the theoretical mechanism engagement are more nuanced, showing improvement in self-determination and self-efficacy, but no significant change in self-knowledge or emotion regulation capacity. Synthesizing the qualitative interview data from the 4 providers who were unaffiliated with the research, it seems that the program could conceivably be delivered by agencies and clinicians that do not necessarily specialize in autism treatment and that having the manual as a guide was appreciated. Caregiver and family engagement in the program seemed to help with implementation and impact. Providers also reported that strong partnerships within the implementation community (eg, potential employers, job

XSL•FO
RenderX

coaches, and education systems) and broader community acceptance of autism would facilitate ease of delivery.

Recent research suggests that professionals who work with transition-aged individuals with autism are concerned about their lack of knowledge about providing effective supports [55]. A readily available, manualized program could help providers target and address transition needs and provide quality services with minimal training. In fact, our findings suggest that providers found STEPS highly feasible and acceptable immediately after training, and their impressions were confirmed following delivering STEPS with high fidelity to the manual to participants with autism. STEPS, created to promote successful transition to adulthood, was initially developed with considerable input from the autism community and from experts in autism treatment [23]. In this study, we further refined the program based on community stakeholder feedback, ultimately leading to a transition support program that more readily addresses the needs of the consumers (ie, adolescents and adults with autism and their caregivers) as well as the providers. Information relevant to adoption and sustained use, across relevant implementation domains, was gleaned as well. With respect to the outer domain [28], for instance, the providers generally agreed that the program could be implemented in broad settings and by providers who do not often work with people on the spectrum.

Findings must be interpreted in the context of several considerable limitations, first and foremost being the lack of a control condition. Inferences about causation are not justified, nor can we make claims of mechanism engagement given the study design. Of course, an open trial is appropriate at this stage of research, but future implementation trials should include an active control condition and randomization. The design of the assessment battery was another limitation. One of our 2 primary outcomes, the TRS, was only collected at end point, not follow-up, which limited our ability to examine whether transition readiness persisted following completion of STEPS. It is important to note that the timing of the follow-up assessment may have impacted our objective indicators of outcome on the Rehabilitation Success measure. There were several participants who completed their follow-up over summer or holiday breaks and may have indicated they were no longer involved in educational activities despite just being on a regularly scheduled break. Related to this, STEPS is personalized and multifaceted, in that it does not target a single symptom or program. One client might be working toward obtaining a driver's permit, another on forming their first intimate relationship, and a third on securing a full-time job. The program's modular content is anchored to the client's goals, such that outcomes that focus strictly on vocational status likely do not fully capture change in functioning or program impact. We also had a relatively high degree of missing data at the end point and follow-up, which restricted our ability to use other statistical methods of assessing change. A final, yet substantial, limitation is the small sample and informal qualitative analytic approach. Though arguably appropriate for the pilot of community implementation, no firm conclusions regarding the readiness of STEPS for large-scale community implementation can be made. In terms of implementation evaluation, the small sample of providers (n=4) interviewed and the stakeholder group in Phase 1 must be noted; although appropriate for a pilot study, rigorous qualitative data analysis was not viable.

Despite these limitations, this is one of the first community implementation trials to evaluate a research-supported transition program for adolescents and young adults with autism. The results suggest that further research in this area is warranted to facilitate the community implementation of effective programming that is acceptable to consumers. Programming, such as STEPS, is both needed and promising for improving adult outcomes.

Conclusions

This study provides further support for STEPS, a program to support successful transition to adulthood for people who have autism. Prior research has shown feasibility, acceptability, and efficacy of STEPS. This community-based trial provides new evidence to support the program's ability to be delivered with integrity by practitioners, and its effectiveness clinically.

Acknowledgments

This trial was preregistered with ClinicalTrials.gov (Identifier NCT05581069). We received funding from the Organization for Autism Research (PI White). We gratefully acknowledge the many people within the autism community who provided input both on the program and its implementation in the community in preparation for this trial. We also are extremely grateful to Tripp Ritchie and the North Georgia Autism Foundation for helping to make this work possible, and for all that they do to help autistic people in rural communities. Generative artificial intelligence (AI) technology was not used in the preparation of any part of this manuscript.

Data Availability

All data generated or analyzed during this study are included in this published article and deidentified data and study protocols are available from the corresponding author on reasonable request.

Authors' Contributions

SW contributed to conceptualization, formal analysis, funding acquisition, investigation, methodology, project administration, resources, and writing of the original draft. AB contributed to data curation, formal analysis, and writing of the original draft. NP contributed to formal analysis, project administration, and writing original draft.

XSL•FO
RenderX

Conflicts of Interest

None declared.

References

1. Ganz ML. The lifetime distribution of the incremental societal costs of autism. Arch Pediatr Adolesc Med 2007 Apr;161(4):343-349. [doi: 10.1001/archpedi.161.4.343] [Medline: 17404130]
2. Blaxill M, Rogers T, Nevison C. Autism tsunami: the impact of rising prevalence on the societal cost of autism in the United States. J Autism Dev Disord 2022 Jun;52(6):2627-2643. [doi: 10.1007/s10803-021-05120-7] [Medline: 34278527]
3. Davies J, Romualdez AM, Malyan D, et al. Autistic adults' priorities for future autism employment research: perspectives from the United Kingdom. Autism Adulthood 2024 Mar 1;6(1):72-85. [doi: 10.1089/aut.2022.0087] [Medline: 38435326]
4. Roche L, Adams D, Clark M. Research priorities of the autism community: a systematic review of key stakeholder perspectives. Autism 2021 Feb;25(2):336-348. [doi: 10.1177/1362361320967790] [Medline: 33143455]
5. Harris L, Gilmore D, Longo A, Hand BN. Short report: patterns of US federal autism research funding during 2017-2019. Autism 2021 Oct;25(7):2135-2139. [doi: 10.1177/13623613211003430] [Medline: 33765838]
6. Pellicano E, Dinsmore A, Charman T. What should autism research focus upon? Community views and priorities from the United Kingdom. Autism 2014 Oct;18(7):756-770. [doi: 10.1177/1362361314529627] [Medline: 24789871]
7. Thiel T, Riedelbauch S, Gaigg S, Roessner V, Ring M. The impact of depressive and anxious symptoms on quality of life in adults on the autism spectrum. Autism Res 2024 Jun;17(6):1161-1174. [doi: 10.1002/aur.3144] [Medline: 38783834]
8. Bishop-Fitzpatrick L, Mazefsky CA, Eack SM. The combined impact of social support and perceived stress on quality of life in adults with autism spectrum disorder and without intellectual disability. Autism 2018 Aug;22(6):703-711. [doi: 10.1177/1362361317703090] [Medline: 28666391]
9. Ghanouni P, Quirke S, Blok J, Casey A. Independent living in adults with autism spectrum disorder: stakeholders' perspectives and experiences. Res Dev Disabil 2021 Dec;119:104085. [doi: 10.1016/j.ridd.2021.104085] [Medline: 34600352]
10. Steinhausen HC, Mohr Jensen C, Lauritsen MB. A systematic review and meta-analysis of the long-term overall outcome of autism spectrum disorders in adolescence and adulthood. Acta Psychiatr Scand 2016 Jun;133(6):445-452. [doi: 10.1111/acps.12559] [Medline: 26763353]
11. Cai RY, Hall G, Pellicano E. Predicting the financial wellbeing of autistic adults: Part I. Autism 2024 May;28(5):1203-1215. [doi: 10.1177/13623613231196085] [Medline: 37665058]
12. Engström I, Ekström L, Emilsson B. Psychosocial functioning in a group of Swedish adults with Asperger syndrome or high-functioning autism. Autism 2003 Mar;7(1):99-110. [doi: 10.1177/1362361303007001008] [Medline: 12638767]
13. Brouwers EPM, Bergijk M, van Weeghel J, Detaille S, Dewinter J. What is the meaning of paid employment for well-being? A focus group study on differences and similarities between autistic adults with and without employment. J Occup Rehabil 2024 Jun;34(2):415-424. [doi: 10.1007/s10926-023-10136-0] [Medline: 37749451]
14. Bentum JV, Sijbrandij M, Huibers M, Begeer S. Occurrence and predictors of lifetime suicidality and suicidal ideation in autistic adults. Autism 2024 Sep;28(9):2282-2294. [doi: 10.1177/13623613231225901]
15. Hymas R, Badcock JC, Milne E. Loneliness in autism and its association with anxiety and depression: a systematic review with meta-analyses. Rev J Autism Dev Disord 2024 Mar;11(1):121-156. [doi: 10.1007/s40489-022-00330-w]
16. Pelton MK, Cassidy SA. Are autistic traits associated with suicidality? A test of the interpersonal-psychological theory of suicide in a non-clinical young adult sample. Autism Res 2017 Nov;10(11):1891-1904. [doi: 10.1002/aur.1828] [Medline: 28685996]
17. Lee CE, Kim JG. Person-centered transition planning for youth on the autism spectrum: what are we still missing? Exceptionality 2022 May 27;30(3):173-186. [doi: 10.1080/09362835.2021.1938065]
18. Snell-Rood C, Ruble L, Kleinert H, et al. Stakeholder perspectives on transition planning, implementation, and outcomes for students with autism spectrum disorder. Autism 2020 Jul;24(5):1164-1176. [doi: 10.1177/1362361319894827] [Medline: 31957461]
19. Kirby AV, Bagatell N, Baranek GT. The formation of postsecondary expectations among parents of youth with autism spectrum disorder. Focus Autism Other Dev Disabl 2020 Jun;35(2):118-128. [doi: 10.1177/1088357619881221]
20. Sosnowy C, Silverman C, Shattuck P. Parents' and young adults' perspectives on transition outcomes for young adults with autism. Autism 2018 Jan;22(1):29-39. [doi: 10.1177/1362361317699585] [Medline: 29020791]
21. White SW, Elias R, Salinas CE, et al. Students with autism spectrum disorder in college: results from a preliminary mixed methods needs analysis. Res Dev Disabil 2016 Sep;56:29-40. [doi: 10.1016/j.ridd.2016.05.010] [Medline: 27262124]
22. White SW. Introduction to the special issue: experimental therapeutics in clinical child and adolescent psychology: identifying mechanisms and moving the needle. J Clin Child Adolesc Psychol 2022;51(3):261-263. [doi: 10.1080/15374416.2022.2051521] [Medline: 35671232]
23. White SW, Elias R, Capriola-Hall NN, et al. Development of a college transition and support program for students with autism spectrum disorder. J Autism Dev Disord 2017 Oct;47(10):3072-3078. [doi: 10.1007/s10803-017-3236-8] [Medline: 28685409]

XSL•FO
RenderX

24. White SW, Smith IC, Miyazaki Y, Conner CM, Elias R, Capriola-Hall NN. Improving transition to adulthood for students with autism: a randomized controlled trial of STEPS. J Clin Child Adolesc Psychol 2021;50(2):187-201. [doi: 10.1080/15374416.2019.1669157] [Medline: 31609666]
25. Brewe AM, White SW. Pilot trial of a community-based transition support program for autistic adults. Neurodiversity 2025 Jan;3. [doi: 10.1177/27546330251319884]
26. Lancaster GA, Thabane L. Guidelines for reporting non-randomised pilot and feasibility studies. Pilot Feasibility Stud 2019;5(1):114. [doi: 10.1186/s40814-019-0499-1] [Medline: 31608150]
27. Curran GM, Bauer M, Mittman B, Pyne JM, Stetler C. Effectiveness-implementation hybrid designs: combining elements of clinical effectiveness and implementation research to enhance public health impact. Med Care 2012 Mar;50(3):217-226. [doi: 10.1097/MLR.0b013e3182408812] [Medline: 22310560]
28. Damschroder LJ, Reardon CM, Widerquist MAO, Lowery J. The updated consolidated framework for implementation research based on user feedback. Implement Sci 2022 Oct 29;17(1):75. [doi: 10.1186/s13012-022-01245-0] [Medline: 36309746]
29. Chorpita BF, Daleiden EL, Park AL, et al. Child STEPs in California: a cluster randomized effectiveness trial comparing modular treatment with community implemented treatment for youth with anxiety, depression, conduct problems, or traumatic stress. J Consult Clin Psychol 2017 Jan;85(1):13-25. [doi: 10.1037/ccp0000133] [Medline: 27548030]
30. Aschbrenner KA, Kruse G, Gallo JJ, Plano Clark VL. Applying mixed methods to pilot feasibility studies to inform intervention trials. Pilot Feasibility Stud 2022 Sep 26;8(1):217. [doi: 10.1186/s40814-022-01178-x] [Medline: 36163045]
31. Sekhon M, Cartwright M, Francis JJ. Acceptability of healthcare interventions: an overview of reviews and development of a theoretical framework. BMC Health Serv Res 2017 Jan 26;17(1):88. [doi: 10.1186/s12913-017-2031-8] [Medline: 28126032]
32. Bowen DJ, Kreuter M, Spring B, et al. How we design feasibility studies. Am J Prev Med 2009 May;36(5):452-457. [doi: 10.1016/j.amepre.2009.02.002] [Medline: 19362699]
33. Chou YC, Wehmeyer ML, Shogren KA, Palmer SB, Lee J. Autism and self-determination: factor analysis of two measures of self-determination. Focus Autism Other Dev Disabl 2017 Sep;32(3):163-175. [doi: 10.1177/1088357615611391]
34. Mithaug DE, Campeau PL, Wolman JM. Assessing self-determination prospects among students with and without disabilities. In: Mithaug DE, Mithaug DK, Agran M, Martin JE, Wehmeyer ML, editors. Self-Determined Learning Theory: Construction, Verification, and Evaluation: Routledge; 2002:61-76. [doi: 10.4324/9781410606235]
35. Lee Y, Wehmeyer ML, Palmer SB, Williams-Diehm K, Davies DK, Stock SE. The effect of student-directed transition planning with a computer-based reading support program on the self-determination of students with disabilities. J Spec Educ 2011 Aug;45(2):104-117. [doi: 10.1177/0022466909358916]
36. Seong Y, Wehmeyer ML, Palmer SB, Little TD. Effects of the self-directed individualized education program on self-determination and transition of adolescents with disabilities. Career Dev Transit Except Individ 2015 Dec;38(3):132-141. [doi: 10.1177/2165143414544359]
37. White SW, Ollendick TH, Bray BC. College students on the autism spectrum: prevalence and associated problems. Autism 2011 Nov;15(6):683-701. [doi: 10.1177/1362361310393363] [Medline: 21610191]
38. Mazefsky CA, Yu L, White SW, Siegel M, Pilkonis PA. The emotion dysregulation inventory: psychometric properties and item response theory calibration in an autism spectrum disorder sample. Autism Res 2018 Jun;11(6):928-941. [doi: 10.1002/aur.1947] [Medline: 29624893]
39. Gershon RC, Wagster MV, Hendrie HC, Fox NA, Cook KF, Nowinski CJ. NIH toolbox for assessment of neurological and behavioral function. Neurology (ECronicon) 2013 Mar 12;80(11 Suppl 3):S2-S6. [doi: 10.1212/WNL.0b013e3182872e5f] [Medline: 23479538]
40. Kupst MJ, Butt Z, Stoney CM, et al. Assessment of stress and self-efficacy for the NIH Toolbox for Neurological and Behavioral Function. Anxiety Stress Coping 2015;28(5):531-544. [doi: 10.1080/10615806.2014.994204] [Medline: 25577948]
41. Salsman JM, Schalet BD, Merluzzi TV, et al. Calibration and initial validation of a general self-efficacy item bank and short form for the NIH PROMIS®. Qual Life Res 2019 Sep;28(9):2513-2523. [doi: 10.1007/s11136-019-02198-6] [Medline: 31140041]
42. Ghorbani N, Watson PJ, Hargis MB. Integrative Self-Knowledge Scale: correlations and incremental validity of a cross-cultural measure developed in Iran and the United States. J Psychol 2008 Jul;142(4):395-412. [doi: 10.3200/JRPL.142.4.395-412] [Medline: 18792651]
43. McMahon RJ, Forehand RL. Consumer satisfaction in behavioral treatment of children: types, issues, and recommendations. Behav Ther 1983 Mar;14(2):209-225. [doi: 10.1016/S0005-7894(83)80111-7]
44. MacKenzie KT, Beck KB, Eack SM, Zeglen KN, Conner CM, Mazefsky CA. Developing a measure of key adult outcomes in adults with developmental disabilities: conceptual model and item generation of the REALS (Relationships, Employment, Autonomy, and Life Satisfaction). J Autism Dev Disord 2025 Feb;55(2):428-439. [doi: 10.1007/s10803-023-06219-9] [Medline: 38291244]

XSL•FO
RenderX

45. Carter EW, Lane KL, Cooney M, Weir K, Moss CK, Machalicek W. Parent assessments of self-determination importance and performance for students with autism or intellectual disability. Am J Intellect Dev Disabil 2013 Jan;118(1):16-31. [doi: 10.1352/1944-7558-118.1.16] [Medline: 23301900]
46. Wolman J, Campeau P, Dubois P, Mithaug D, Stolarski V. AIR Self-Determination Scale and User Guide: Palo Alto, CA: American Institute for Research; 1994.
47. Elias R, Conner CM, White SW. Brief report: creation of a transition readiness scale for adolescents with ASD. J Autism Dev Disord 2023 Mar;53(3):1279-1284. [doi: 10.1007/s10803-022-05449-7] [Medline: 35138556]
48. Hatcher RL, Gillaspy JA. Development and validation of a revised short version of the working alliance inventory. Psychother Res 2006 Jan;16(1):12-25. [doi: 10.1080/10503300500352500]
49. Munder T, Wilmers F, Leonhart R, Linster HW, Barth J. Working Alliance Inventory-Short Revised (WAI-SR): psychometric properties in outpatients and inpatients. Clin Psychol Psychother 2010;17(3):231-239. [doi: 10.1002/cpp.658] [Medline: 20013760]
50. Weiner BJ, Lewis CC, Stanick C, et al. Psychometric assessment of three newly developed implementation outcome measures. Implement Sci 2017 Aug 29;12(1):108. [doi: 10.1186/s13012-017-0635-3] [Medline: 28851459]
51. Benevides TW, Shore SM, Palmer K, et al. Listening to the autistic voice: mental health priorities to guide research and practice in autism from a stakeholder-driven project. Autism 2020 May;24(4):822-833. [doi: 10.1177/1362361320908410] [Medline: 32429818]
52. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
53. Elo S, Kyngäs H. The qualitative content analysis process. J Adv Nurs 2008 Apr;62(1):107-115. [doi: 10.1111/j.1365-2648.2007.04569.x] [Medline: 18352969]
54. White SW, Richey JA, Gracanin D, et al. Psychosocial and computer-assisted intervention for college students with autism spectrum disorder: preliminary support for feasibility. Educ Train Autism Dev Disabil 2016 Sep;51(3):307-317. [Medline: 28111607]
55. LaPoint SC, Kim SY, Bottema-Beutel K. Barriers to providing transitional supports for autistic students: insights of school professionals. J Autism Dev Disord 2025 Aug;55(8):2807-2819. [doi: 10.1007/s10803-024-06375-6] [Medline: 38773033]

Abbreviations

AIM: Acceptability of Implementation Measure
AIR-SD: American Institutes for Research–Self-Determination Scale
CFIR: Consolidated Framework for Implementation Research
EDI: Emotion Dysregulation Inventory
FIM: Feasibility of Implementation Measure
GSES: General Self-Efficacy Scale
IAM: Intervention Appropriateness Measure
ISKS: Integrative Self-Knowledge Scale
PSS: Program Satisfaction Scale
RCT: randomized controlled trial
REALS: Relationships, Employment, Autonomy, and Life Satisfaction
SCSDSS: Seven Component Self-Determination Skills Survey
STEPS: Stepped Transition to Employment and Postsecondary Education Success
TRS: Transition Readiness Scale
WAI-SR: Working Alliance Inventory–Short Revised

Edited by A Mavragani; submitted 16.12.24; peer-reviewed by K Scott; revised version received 04.07.25; accepted 06.07.25; published 29.08.25.

Please cite as:
White SW, Brewe AM, Powell N
Stepped Transition to Employment and Postsecondary Education Success (STEPS) for Adolescents and Adults with Autism: Community Implementation Pilot Trial
JMIR Form Res 2025;9:e70137
URL: https://formative.jmir.org/2025/1/e70137
doi:10.2196/70137

XSL•FO
RenderX

© Susan W White, Alexis M Brewe, Nicole Powell. Originally published in JMIR Formative Research (https://formative.jmir.org), 29.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Feasibility and Acceptability of a Psychological Intervention for Internalized Health-Related Stigma Among Adults With Chronic Health Conditions: Preliminary Investigation

Rebecca L Pearl[1], PhD; Danielle Saunders[1], MPH; Laurie C Groshon[1], MS; Yulin Li[2], MS; Abigail Shonrock[1], PhD; Rebecca M Puhl[3], PhD; Kimberly A Driscoll[1], PhD; Preeti Manavalan[4], MD, MSc; Joel M Gelfand[5], MD, MSCE; Thomas A Wadden[6], PhD; Sarah C Westen[1], PhD; Marjorie Montanez-Wiscovich[7], MD, PhD; Xiang-Yang Lou[2], PhD

[1]Department of Clinical and Health Psychology, College of Public Health and Health Professions, University of Florida, P.O. Box 100165, Gainesville, FL, United States

[2]Department of Biostatistics, College of Medicine and College of Public Health and Health Professions, University of Florida, Gainesville, FL, United States

[3]Department of Human Development & Family Sciences, University of Connecticut, Hartford, CT, United States

[4]Department of Medicine, College of Medicine, University of Florida, Gainesville, FL, United States

[5]Department of Dermatology and Department of Biostatistics, Epidemiology and Informatics, Center for Clinical Sciences in Dermatology, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, United States

[6]Center for Weight and Eating Disorders, Department of Psychiatry, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, United States

[7]Department of Dermatology, College of Medicine, University of Florida, Gainesville, FL, United States

Corresponding Author:
Rebecca L Pearl, PhD
Department of Clinical and Health Psychology, College of Public Health and Health Professions, University of Florida, P.O. Box 100165, Gainesville, FL, United States

Abstract

Background: Health-related stigma is widely acknowledged as a threat to public health and a barrier to managing chronic health conditions. Internalized stigma is a particularly strong predictor of poor health outcomes across health conditions, yet few evidence-based interventions are available. Peer support and counseling have been investigated as interventions for reducing internalized stigma. Typically, these interventions are developed and tested in disease-specific research, focusing on one health condition in isolation from others. This approach may limit knowledge and dissemination of support for health-related stigma across health conditions. Recent work has highlighted the need for research that breaks down traditional silos by using cross-cutting approaches to understand and reduce stigma.

Objective: This study aimed to determine the feasibility and acceptability of a new group-based psychological intervention designed to reduce internalized health-related stigma among adults with different stigmatized chronic physical health conditions.

Methods: A group intervention that was initially designed to address internalized weight stigma was adapted to be generalizable to other forms of internalized health-related stigma. This was done with input from Advisory Board members living with different stigmatized chronic health conditions and health professionals who specialized in these conditions. Adults with obesity, diabetes, HIV, skin diseases, chronic pain, or cancers were recruited to attend 12 weekly online group meetings. The average session attendance rate was computed with and without makeup sessions. A treatment acceptability questionnaire was completed at week 12. Primarily for feasibility testing, participants completed pre- and post-treatment questionnaires that assessed internalized health-related stigma and other relevant aspects of mental health and health-related quality of life. At baseline, participants were also asked to report reasons for perceived discrimination. Data collection occurred from December 2023 through April 2024.

Results: In total, 10 adults were recruited within approximately 6 weeks, of whom 8 attended at least 1 treatment session and completed post-treatment questionnaires, with a 80% retention. The average session attendance rate was 95.8% with makeup sessions and 83.3% for those without makeup. Treatment acceptability ratings were high, with an overall acceptability rating of approximately 6.5 (SD 0.5) out of 7. Medium to large effect sizes were observed for changes in internalized stigma and some aspects of mental health. Almost all (n=7, 87.5%) of participants reported experiencing discrimination due to their health conditions, which accompanied a wide range of other reasons for perceived discrimination.

Conclusions: Results showed high feasibility and acceptability of a transdiagnostic, online group psychological intervention for internalized health-related stigma delivered to adults with different types of stigmatized chronic physical health conditions.

XSL•FO
RenderX

Given the small sample size and limited generalizability, testing in a large efficacy trial is needed to determine intervention benefits.

(JMIR Form Res 2025;9:e69548) doi:10.2196/69548

KEYWORDS

cancer; chronic disease; diabetes; HIV; obesity; pain; skin disease; stigma

Introduction

Health-related stigma involves negative judgment, blame, social rejection, and discrimination due to health conditions [1]. Stigma has been well documented across many types of chronic medical conditions that vary in key characteristics, such as their visibility (vs concealability), perceived risk of contagion, and perceived controllability [2,3]. For example, adults with obesity, type 1 or type 2 diabetes, skin diseases, HIV, chronic pain, and cancers face pervasive stigma for their health conditions [1,4-6].

In addition to experiencing stigma from other people, negative attitudes can be internalized by adults with these health conditions, contributing to self-blame and self-devaluation (ie, self-stigma) [7,8]. For example, individuals with obesity who internalize weight stigma often apply negative stereotypes to themselves (eg, believing that they are lazy or lack willpower) and blame themselves for their weight, leading them to view themselves as a failure [9]. The internalization of health-related stigma is associated with poor mental and physical health outcomes across different health conditions, including increased loneliness, depression, and anxiety, and reduced quality of life [1]. Internalized health-related stigma is also linked to reduced engagement in self-management behaviors for chronic health conditions—such as less healthy eating and physical activity, impaired medication adherence, and avoidance of medical appointments—which can further contribute to worse health outcomes [9-13]. Internalized stigma due to health conditions can interact with stigma related to other aspects of a person's identity, such as their age, gender, race, ethnicity, or socioeconomic status [14,15]. The intersection of multiple marginalized identities or characteristics can compound stigma and its adverse health consequences [16-20].

Despite documented harms of stigma and its internalization, few interventions exist to combat internalized health-related stigma. Intervention approaches that have been proposed include peer support and psychological counseling [21]. Some studies have found benefits of these interventions when tested to reduce internalized stigma due to mental illness, HIV, or obesity [22-26]. However, such interventions have been relatively unstudied for most health conditions (eg, cancers, skin diseases) [6,27]. Health-related stigma is typically studied in disease-specific silos, despite many similarities in the presentation and impacts of stigma across health conditions [21]. Given the lack of interventions for internalized stigma, an intervention that reduces internalized stigma across different health conditions—rather than a disease-specific intervention that benefits only 1 patient population—could have the potential to enhance dissemination of and increase access to psychosocial support for adults with chronic health conditions who have internalized stigma. A "transdiagnostic" intervention could also address the needs of adults with multiple chronic health conditions, as well as leave space to address other intersecting marginalized identities.

Transdiagnostic approaches are commonly used in mental health care [28]. For example, skills groups can teach evidence-based coping strategies (eg, mindfulness, cognitive restructuring) to individuals with different types of mental health concerns within the same group [29,30]. A transdiagnostic framework is less commonly applied to the treatment of chronic physical health conditions, due in part to the importance of understanding the specific etiology and pathophysiology of each health condition and the unique needs of each patient population. However, developing a psychological intervention that addresses stigma transdiagnostically, across different stigmatized health conditions, could have several benefits, including saving time and resources. Disease-specific interventions also require individual clinics or specialists to deliver treatment, while a transdiagnostic intervention could be delivered in a centralized manner (eg, through referrals from different specialties within a health system) and by a mental health provider who does not need to have expertise in any 1 particular health condition. Including individuals with different stigmatized health conditions within the same groups could also serve to further destigmatize their conditions by highlighting commonalities and shared experiences among them, thus emphasizing the universal nature of stigma and leading individuals with chronic health conditions to feel less isolated or "different" from others [8,31]. However, possible downsides of this approach include reduced capacity to tailor interventions to address the specific challenges faced by patients with certain health conditions, and the potential for individuals who have internalized stigma to feel uncomfortable discussing their health conditions with individuals who do not share their same diagnoses. Given the novelty of this proposed transdiagnostic approach to addressing health-related stigma, preliminary research to assess feasibility and acceptability is warranted.

The present study was an early-phase (stage 1 [32]) investigation of a transdiagnostic, online, group-based psychological intervention designed to reduce internalized health-related stigma across different stigmatized chronic physical health conditions. The primary aim of this study was to determine the feasibility and acceptability of the intervention. The study was conducted in preparation for a larger planned randomized controlled clinical trial that will test the efficacy of the intervention in comparison to general peer support and a waitlist control.

XSL•FO
RenderX

Methods

Participants and Procedures

Participants were eligible if they had one or more of the following health conditions: obesity; type 1 or type 2 diabetes; skin disease (eg, psoriasis, eczema, or vitiligo); HIV; chronic pain; or cancer (including in remission). These conditions represent a range of dimensions important to stigma (eg, perceived controllability, contagion, and concealability). Inclusion criteria also included self-reported perceptions and internalization of health-related stigma. The latter criterion was determined by a cutoff score of 3.4 or higher on the Internalized Health-Related Stigma (I-HEARTS) Scale [33] (described below), with confirmation by clinical interview that participants' health condition(s) negatively affected how they thought and felt about themselves. These criteria were included to ensure that the stigma intervention would be relevant to participants and that the study would include individuals who could benefit most from it. Study candidates were excluded if they currently or in the past 3 months received psychotherapy or psychosocial or peer support or if they had been hospitalized for psychiatric reasons in the past 6 months, due to potential confounding effects of treatment. Additional exclusion criteria were current, active suicidal thoughts or a reported suicide attempt within the past year; a current alcohol or substance use disorder that required immediate treatment; severe progression of disease (eg, end-of-life) or undergoing current acute, intensive treatment (eg, chemotherapy or radiation); or a current or past thought disorder, psychosis, or unmanaged bipolar disorder, as these conditions require a higher level of care and could interfere with the ability to participate in group meetings.

Participants were recruited primarily from the University of Florida (UF) Health system. Using a database of patients who agreed to be contacted for research studies, emails describing the study were sent to adults with eligible health conditions (in batches with equal numbers across the 6 health condition categories) with a link to a pre-screening questionnaire to determine initial eligibility. Additional efforts were made to recruit participants from community health organizations and through referrals from UF Health clinics. Study candidates who were initially eligible were contacted to conduct a phone screening interview with a research coordinator, followed by a virtual interview with the principal investigator to confirm eligibility and provide informed consent. During the virtual interview, the principal investigator discussed with participants their comfort sharing information about stigma in a group setting with others who may have different health conditions from them, as well as issues of privacy, confidentiality, and other concerns that may arise from participating in group treatment. Privacy, confidentiality, and group guidelines about treating others with respect were also reviewed during the first group meeting, with the goal of creating a safe and supportive environment.

In the week before starting the intervention, participants completed baseline questionnaires online. Participants received 12 weekly, 50-minute virtual treatment sessions led by the principal investigator (a licensed psychologist) and co-led by a doctoral student in clinical psychology. Participants who were absent from group sessions were offered brief make-up phone sessions with the group leader or co-leader within 1 week of the missed session. Participants completed questionnaires again during week 12. Study staff tracked adverse events reported by participants (none were determined to be related to the study).

Ethical Considerations

Study procedures were approved by the University of Florida Institutional Review Board (IRB202201862). Informed consent was obtained from all participants. Participants were assigned identification numbers so that study data could be analyzed in deidentified form, with only select study staff having access to identifying information. Participants were compensated with a US $50 electronic payment or debit card after completing week 12 questionnaires.

Intervention

The intervention was adapted from the Weight Bias Internalization and Stigma (Weight BIAS) Program, which was developed and tested for the purpose of reducing internalized weight stigma among adults with obesity [34]. The Weight BIAS program is a group intervention (incorporating peer support) that uses evidence-based principles and strategies from cognitive behavioral therapy and third-wave behavioral therapies, tailored specifically to address internalized stigma. To adapt the intervention, session topics and skills were largely retained, but session content was made generalizable to other health conditions beyond obesity by including examples relevant to other health conditions (eg, examples of negative stereotypes or self-critical thoughts that individuals with different health conditions may report). New topics and skills were also added to address facets of stigma that may not have been relevant to adults with obesity (eg, the topic of disclosure for individuals with concealable health conditions). All sessions were designed to be transdiagnostic—rather than having health condition-specific sessions or modules—so that individuals with any of the included health conditions could potentially benefit from each session. Some sessions also included prompts to discuss interactions between stigma due to health conditions and other identities.

Two Advisory Boards were formed to participate in the intervention adaptation: a Community Advisory Board comprised of adults representing each of the 6 categories of health conditions included in the study, and a Health Professional Advisory Board of researchers and clinicians who specialized in these health conditions. Advisory Board meetings were held online, in small groups, to discuss ideas for the intervention and provide feedback on a preliminary outline of session topics. Advisory Board members also provided feedback on the study design and structure (eg, including individuals with different stigmatized health conditions within the same treatment groups) in order to prevent the inadvertent perpetuation or exacerbation of stigma within the intervention or research study. Formal feedback on session topics and materials was obtained via a Qualtrics survey to determine acceptability and relevance of the content to each patient population. Changes were made based on this feedback (including incorporating specific examples of stigma provided by Advisory Board members to

XSL•FO
RenderX

include in the treatment manual), and final treatment materials (including the treatment manual and electronic participant handouts) were disseminated to Advisory Board members prior to starting the study.

Measures and Analyses

Feasibility was assessed for recruitment, retention, and adherence. The recruitment goal was 10 participants, with representation across the 6 categories of health conditions. This criterion was based on recommendations for optimal psychotherapy group sizes [31,35], and because the larger planned clinical trial will aim to recruit individuals for groups of 8 - 10 participants across these 6 types of health conditions. Other metrics of feasibility included achieving participant retention of at least 80% and an average session attendance rate of at least 70% (calculated by averaging the total number of sessions attended by participants, both inclusive and exclusive of makeup sessions) [36,37].

Treatment acceptability was assessed with a questionnaire completed by participants at week 12 [25]. In total, 4 items assessed from 1 (not at all) to 7 (extremely) how helpful and acceptable the program was, how much participants liked the program, and how satisfied they were with the program; these items were averaged to produce an overall acceptability score (Cronbach α=.93). Participants also rated on the same scale how likely they would be to recommend the program to others. In addition, they rated from 1 (not at all) to 7 (very much so) the extent to which they learned new things, changed their attitudes about themselves, and the program helped them manage their health conditions. Participants rated on the same scale how much they learned 10 specific skills in the program (eg, challenge myths and stereotypes about health conditions; see Table S1 in Multimedia Appendix 1 for a full list of skills; Cronbach α=.94), and rated from 1 (never) to 5 (frequently) how often they used these skills from the program during the prior 12 weeks (Cronbach α=.84). Participants were also asked a few open-ended questions to give brief, additional feedback about their favorite part of the program, what they did not find helpful, and what they believed they were taking away from the program. Given the brevity of responses elicited and limited qualitative data, a formal content analysis was not planned or conducted, but responses were reviewed and representative quotes were extracted.

As part of feasibility testing, participants completed a battery of pre- and post-treatment questionnaires, as they would in a larger efficacy trial. This battery included the I-HEARTS Scale to assess internalized health-related stigma, which was previously adapted from the Internalized Stigma of Mental Illness Scale [38] and validated in a large sample of adults with varying chronic health conditions, showing strong psychometric properties [33]. The cutoff score of 3.4 out of 7 was established in prior research to identify clinically significant levels of internalized stigma [33] and was used to determine eligibility in the present study. The scale produces a total score, as well as 3 subscale scores for Perceived and Anticipated Stigma, Stereotype Application and Self-Devaluation, and Stigma Resistance.

Other measures completed by participants included the Internalized Shame Scale (with subscales for shame and self-esteem [39]) and the UCLA Loneliness Scale (version 3 [40]), given that these constructs are highly relevant to internalized stigma. Mental health was assessed with the Patient Health Questionnaire-9 (to assess depression [41]); the Generalized Anxiety Disorder-7 [42]; and the 10-item Severity Measure for Social Anxiety Disorder [43]. The 4-item Perceived Stress Scale assessed general stress [44]; several subscales of the Revised Illness Perceptions Questionnaire assessed disease-related quality of life [45]; and the 12-item Short Form Health Survey (SF-12 [46]) and the CDC Healthy Days Core Measure [47] assessed general mental and physical health-related quality of life. These outcome measures will be relevant to include in a larger efficacy trial due to links between internalized stigma and poor mental and physical health outcomes. Although primarily included for feasibility testing, exploratory completer's analyses examined effect sizes from paired t-tests conducted between baseline and week 12.

For descriptive purposes, participants completed a questionnaire about their health conditions (eg, their diagnoses, visibility of the health conditions, and severity of symptoms) and were asked to report their demographic characteristics. They also completed the Everyday Discrimination Scale [48] to identify reasons for perceived discrimination.

Results

Figure S1 in Multimedia Appendix 1 displays the participant flow chart. One study candidate recruited from a community group completed a pre-screening survey in November 2023. Email invitations to patients in the UF Health database were first sent in early December 2023, and the first participant was consented in early January 2024, with recruitment ending in mid-January. Altogether, the goal of enrolling 10 participants was met within approximately 6 weeks. Two of these participants did not start the intervention due to a personal or family medical event, respectively.

Table S2 in Multimedia Appendix 1 presents characteristics of the 10 enrolled participants, and Table 1 presents characteristics of the 8 participants who received the intervention. Participant characteristics did not differ between those who did versus did not receive the intervention. Participants predominantly identified as non-Hispanic, White, female, heterosexual, married, and either retired or receiving disability benefits, with a mean age of 60 years. Participants reported some college education on average, and household income ranged from <US $10,000 to US $50,000 - 74,999. The most common health conditions reported were chronic pain and obesity, and most participants reported having 3 or more of the 6 different types of health conditions. Most participants indicated that their health conditions were visible and were moderate in severity overall. Apart from the 6 main categories of health conditions, other health conditions reported by participants included Hashimoto's disease, liver disease, heart disease, hypertension, autoimmune hepatitis, dystonia, dysphonia, and migraines.

XSL•FO
RenderX

Table . Characteristics of adults with stigmatized chronic health conditions who received the 12-week internalized health-related stigma intervention (*N*=8).

Variables	Participants (N=8), n (%)
Age (years), mean (SD); range	59.1 (8.9); 47 - 69
Sex, n (%)	
Female	6 (75.0)
Male	2 (25.0)
Race and ethnicity, n (%)	
Non-Hispanic White	6 (75.0)
Non-Hispanic Black	2 (25.0)
Sexual orientation, n (%)	
Heterosexual or straight	7 (87.5)
Asexual	1 (12.5)
Marital status, n (%)	
Single (never married)	2 (25.0)
Married	5 (62.6)
Divorced	1 (12.5)
Employment status, n (%)	
Retired	3 (37.5)
On disability	3 (37.5)
Employed part-time	1 (12.5)
Unemployed	1 (12.5)
Annual household income (US) $, n (%)	
Less than $10,000	1 (12.5)
$10,000-$24,999	1 (12.5)
$25,000-$34,999	1 (12.5)
$35,000-$49,999	3 (37.5)
$50,000-$74,999	1 (12.5)
Prefer not to answer	1 (12.5)
Education (years), mean (SD)	14.1 (2.5)
Health conditions, n (%)	
Obesity	5 (62.5)
Diabetes	
Type 1	0 (0)
Type 2	3 (37.5)
Skin disease	2 (20.0)
HIV	1 (10.0)
Chronic pain	7 (87.5)
Cancer (in remission)	2 (20.0)
One of the above health conditions	2 (20.0)
Two of the above health conditions	1 (10.0)
Three or more of the above health conditions	5 (62.5)
Participants who reported additional health conditions not listed above	6 (75.0)

XSL•FO
RenderX

All 8 of the participants who attended at least 1 session of the intervention completed the week 12 questionnaires in April 2024, thus meeting the overall retention goal of 80%. Among these 8 participants, the average attendance rate was 95.8% (11.5 sessions) when including makeup sessions, or 83.3% (10 sessions) when excluding makeup sessions—all exceeding the goal of an average 70% attendance rate.

Treatment acceptability ratings were high (Figure 1), with an average overall acceptability rating of 6.5 (SD 0.5) on a 1 - 7 scale. Open-ended responses were unilaterally positive and cited the following program benefits: connecting with others who had similar experiences, feeling less alone, learning coping strategies, and being less judgmental of themselves. Representative quotes are included in Table S3 in Multimedia Appendix 1.

Figure 1. Treatment acceptability ratings of the internalized health-related stigma intervention. *Ratings for "Used specific skills" were on a scale from 1 (never) to 5 (frequently). All other items were rated on a 1 - 7 scale.

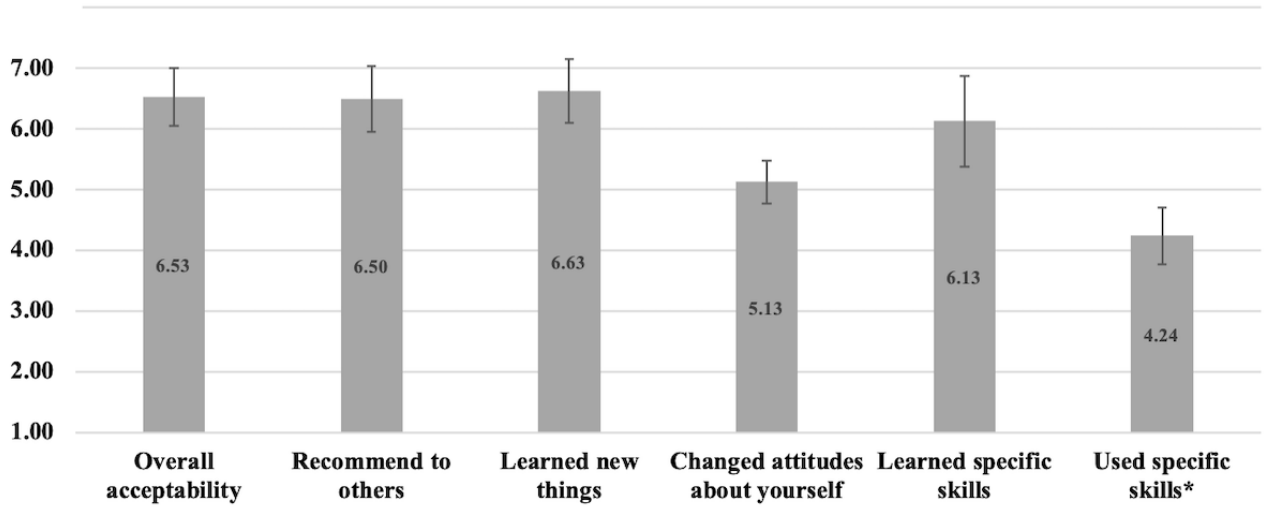

Table 2 summarizes changes in continuous measures. Effect sizes were medium to large for reductions in internalized stigma as measured by the I-HEARTS Scale and its subscales. Effect sizes were also medium to large for improvements in loneliness, stress, and some aspects of disease-related and general mental health-related quality of life. Effect sizes were smaller for other mental health changes and minimal for changes in physical health-related quality of life. Participants reported many reasons for discrimination, with their health conditions reported most frequently (Table 3).

XSL•FO
RenderX

Table . Baseline and week 12 scores on continuous measures assessed before and after the internalized health-related stigma intervention. Higher scores indicate worse outcomes on all measures, except for the Self-Esteem Subscale of the Internalized Shame Scale, the Personal Control Subscale of the Illness Perceptions Questionnaire-Revised, and the SF-12.

Measure	Scale range	Baseline scores, mean (SD)	Week 12 scores, mean (SD)	Change in scores, mean (SD), 95% CI	Cohen *d*
I-HEARTS[a] total score	1 - 7	4.49 (0.71)	3.84 (1.22)	−0.65 (1.45), −1.86 to 0.56	−0.65
Perceived and Anticipated Stigma Subscale	1 - 7	4.54 (0.74)	3.82 (1.33)	−0.72 (1.64), −2.09 to 0.65	−0.67
Stereotype Application and Self-Devaluation Subscale	1 - 7	4.53 (1.63)	3.31 (1.82)	−1.22 (1.96), −2.86 to 0.42	−0.71
Stigma Resistance Subscale (reverse-scored)	1 - 7	4.22 (1.29)	3.19 (1.25)	−1.03 (1.69), −2.44 to 0.38	−0.81
Internalized Shame Scale					
Shame	0 - 96	51.25 (18.01)	43.88 (18.23)	−7.38 (13.68), −18.81 to 4.06	−0.41
Self-esteem	0 - 24	12.88 (3.44)	13.50 (4.04)	0.63 (5.10), −3.64 to 4.89	0.17
UCLA[b] Loneliness Scale	20 - 80	53.75 (9.77)	46.25 (10.90)	−7.50 (7.35), −13.64 to −1.36	−0.73
Patient Health Questionnaire-9	0 - 27	11.63 (6.30)	10.25 (5.95)	−1.38 (3.02), −3.90 to 1.15	−0.23
Generalized Anxiety Disorder-7	0 - 21	9.25 (6.78)	7.00 (4.66)	−2.25 (3.99), −5.59 to 1.09	−0.39
Social Anxiety Disorder Questionnaire	0 - 40	11.50 (9.09)	9.75 (7.38)	−1.75 (7.31), −7.86 to 4.36	−0.21
Perceived Stress Scale	0 - 16	9.00 (2.56)	6.63 (2.77)	−2.38 (2.13), −4.16 to −0.59	−0.89
Illness Perceptions Questionnaire-Revised					
Consequences	1 - 5	4.33 (0.49)	4.00 (0.37)	−0.33 (0.35), −0.62 to −0.05	−0.77
Personal control	1 - 5	3.56 (0.98)	3.75 (0.46)	0.19 (1.30), −0.90 to 1.27	0.24
Illness coherence	1-5	3.55 (0.64)	3.43 (0.55)	0.13 (0.94) −0.91 to 0.66	−0.21
Emotional representations	1 - 5	3.73 (0.86)	3.46 (0.61)	−0.27 (0.66), −0.82 to 0.28	−0.36
SF-12[c]					
Mental health	0 - 100	37.63 (8.47)	43.02 (7.17)	5.40 (10.35), −3.25 to 14.05	0.69
Physical health	0 - 100	29.02 (9.20)	28.96 (10.82)	−0.06 (5.55), −4.70 to 4.58	−0.01
CDC[d] Unhealthy days index	0 - 30	26.50 (6.82)	25.00 (9.32)	−1.50 (7.76), −7.99 to 4.99	−0.18

[a]I-HEARTS: Internalized Health-Related Stigma.

[b]UCLA: University of California, Los Angeles.

XSL•FO
RenderX

[c]SF-12: 12-item Short Form Health Survey.

[d]CDC: Centers for Disease Control and Prevention.

Table . Reasons for discrimination identified by adults with stigmatized chronic health conditions with the Everyday Discrimination Scale (*N*=8). Frequencies reflect participants who endorsed these reasons in response to any item on the Everyday Discrimination Scale.

Reason for discrimination	Participants, n (%)
Your ancestry or national origins	1 (12.5)
Your race	2 (25)
Your gender	2 (25)
Your age	4 (50)
Your height	1 (12.5)
Your weight	5 (62.5)
Some other aspect of your physical appearance	4 (50)
Your education or income	2 (25)
Your health condition	7 (87.5)
Your physical disability	5 (62.5)
Your shade of skin color	1 (12.5)
Other	2 (25)

Discussion

This early-phase intervention study provides preliminary support for the feasibility and acceptability of a new group-based psychological intervention designed to reduce internalized health-related stigma among adults with different stigmatized chronic health conditions. Results showed successful recruitment, in a short period of time, of adults with obesity, diabetes, skin diseases, HIV, chronic pain, and cancers to participate in a 12-week online intervention study. Retention and session attendance rates and ratings of treatment acceptability were high. Altogether, study results support the potential to recruit participants in a similar manner for a larger trial.

The successful recruitment, retention, and treatment adherence suggest that individuals with chronic health conditions were open to engaging in group discussions of stigma with others who had different health conditions from their own. This was also reflected in open-ended treatment acceptability questionnaire responses and in Advisory Board meeting discussions, especially among Board members living with different kinds of health conditions. A prominent feature of internalized stigma is feeling “othered,” isolated from society, and alone [8]. Group treatment with individuals who vary in their health conditions may serve to highlight to individuals who have internalized stigma that they are not alone [31] and to facilitate finding commonalities with others who do not share their stigmatized health conditions. Notably, individuals who may not have felt comfortable participating in a group of individuals with differing health conditions likely would not have responded to recruitment advertisements, and thus may not have been represented in the sample. It is possible that individuals with the highest levels of internalized stigma may be most uncomfortable with this treatment format. Nevertheless, this preliminary study suggests that, for many individuals with chronic health conditions who have internalized health-related stigma, the treatment paradigm pilot-tested in this study was considered acceptable.

Measures of internalized stigma and relevant mental health and quality of life constructs were primarily included in the study for the purpose of feasibility testing. Pre-post treatment change scores indicated medium to large effects for many variables, including internalized stigma. This study was not designed to test the efficacy of the intervention, and these effect sizes should be interpreted with caution given the very small sample size and limited generalizability. Prior studies testing interventions for internalized health-related stigma have shown a wide range of effect sizes, some of which have been challenging to interpret due to small sample sizes, lack of control groups, and other design limitations [6,21,23,49-51]. Still, several promising interventions have been rigorously tested in recent years for specific health conditions [22-26,52], although such interventions are lacking for most physical health conditions [1,21,51,53,54]. Further testing of this transdiagnostic intervention in an adequately powered randomized controlled trial is needed in order to be able to compare its effect sizes to those of disease-specific stigma interventions.

Almost all participants reported experiencing discrimination due to their health conditions, which was not surprising given that participants were selected for having high levels of internalized stigma. However, participants reported many other reasons for perceived discrimination as well, including factors related to health (disability, weight) and other aspects of their identities (eg, age, race, and gender). These findings highlight the potential utility of applying an intersectional lens to interventions for health-related stigma, in order to acknowledge and address how interactions among different forms of stigma may affect adults with chronic health conditions [15].

XSL•FO
RenderX

The generalizability of the present work is limited in many respects. The study was conducted at a single site, so geographic location was limited, as was the age, gender, racial, and ethnic diversity of the small sample. No control group or randomization was used in the present research. Future testing in a larger efficacy trial is needed to determine the potential benefits of this transdiagnostic intervention for internalized stigma, mental health, and quality of life. This study also focused on stigmatized physical health conditions only; the stigma of mental health conditions was not an intended target for the intervention, in part due to the extensive research that has been conducted on addressing mental health stigma [55], in addition to the exclusion criterion of individuals currently or recently receiving psychotherapy (due to its potentially confounding effects with the treatment provided). Further research is needed to determine if one transdiagnostic stigma intervention could be applied to address both mental and physical health stigma, or if tailoring would be needed to address mental versus physical health concerns. More work is also needed to understand how to best balance the importance of meeting the unique needs of each patient population, while also leveraging knowledge and commonalities across health conditions to enhance dissemination of support to individuals who have internalized health-related stigma. Finally, in addition to interventions for intrapersonal-level interventions such as this, structural- and interpersonal-level interventions are needed to reduce public stigma and prevent the internalization of stigma and its negative associated health outcomes [56].

In conclusion, this preliminary study of a transdiagnostic, online, group-based psychological intervention for internalized health-related stigma showed strong evidence of feasibility and acceptability. This represents a first step in investigating this intervention, which will be followed by a randomized controlled trial to test its effects on internalized stigma. With further testing, this work could have the potential to shift the paradigm for how stigma is addressed and for how psychosocial support interventions may be more broadly disseminated to adults with chronic health conditions.

Acknowledgments

Research reported in this publication was supported by the National Institute of Mental Health and the National Institutes of Health (NIH) Common Fund, through the Office of Strategic Coordination/Office of the NIH Director, under Award Number DP2MH132938 (RLP). PM is supported by the NIMH (K23 MH131463), Fogarty International Center (R21 TW012650), and the Gilead Research Scholars Program in HIV. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. The authors would like to thank all participants for their engagement and for sharing their experiences. The authors also thank all Advisory Board members for providing input on the intervention content and materials.

Authors' Contributions

RLP involved in the conceptualization, methodology, investigation, resources, visualization, supervision, project administration, funding acquisition, and writing – original draft; DS involved in the investigation, project administration, and writing – review & editing; LCG involved in the investigation, resources, and writing – review & editing; YL contributed to software, formal analysis, data curation, visualization, and writing – review & editing; AS contributed to resources, writing – review & editing; RMP, KAD, PM, JMG, TAW, SCW, and MMW involved in the conceptualization, and writing – review & editing; XYL involved in the supervision, formal analysis, and writing – review & editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Additional information about: skills rated in treatment acceptability questionnaire (Supplemental Table S1); characteristics of all 10 enrolled participants (Table S2); representative quotes from open-ended responses in treatment acceptability questionnaire (Table S3); and flow chart of participants enrolled in the study (Figure S1).

[DOC File, 223 KB - formative_v9i1e69548_app1.doc]

References

1. Stangl AL, Earnshaw VA, Logie CH, et al. The Health Stigma and Discrimination Framework: a global, crosscutting framework to inform research, intervention development, and policy on health-related stigmas. BMC Med 2019 Feb 15;17(1):31. [doi: 10.1186/s12916-019-1271-3] [Medline: 30764826]
2. Jones EE, Farina A, Hastorf A, et al. Social Stigma: The Psychology of Marked Relationships: W H Freeman & Co; 1984.
3. Pachankis JE, Hatzenbuehler ML, Wang K, et al. The burden of stigma on health and well-being: a taxonomy of concealment, course, disruptiveness, aesthetics, origin, and peril across 93 stigmas. Pers Soc Psychol Bull 2018 Apr;44(4):451-474. [doi: 10.1177/0146167217741313] [Medline: 29290150]
4. Carr DB. President's message: patients with pain need less stigma, not more. Pain Med 2016;17(8):1391-1393. [doi: 10.1093/pm/pnw158] [Medline: 27418318]

XSL•FO
RenderX

5. Speight J, Holmes-Truscott E, Garza M, et al. Bringing an end to diabetes stigma and discrimination: an international consensus statement on evidence and recommendations. Lancet Diabetes Endocrinol 2024 Jan;12(1):61-82. [doi: 10.1016/S2213-8587(23)00347-9] [Medline: 38128969]
6. Topp J, Andrees V, Weinberger NA, et al. Strategies to reduce stigma related to visible chronic skin diseases: a systematic review. J Eur Acad Dermatol Venereol 2019 Nov;33(11):2029-2038. [doi: 10.1111/jdv.15734] [Medline: 31177601]
7. Corrigan PW, Rao D. On the self-stigma of mental illness: stages, disclosure, and strategies for change. Can J Psychiatry 2012 Aug;57(8):464-469. [doi: 10.1177/070674371205700804] [Medline: 22854028]
8. Goffman E. Stigma: Notes on the Management of a Spoiled Identity: Prentice Hall; 1963.
9. Pearl RL. Internalization of weight bias and stigma: scientific challenges and opportunities. Am Psychol 2024 Dec;79(9):1308-1319. [doi: 10.1037/amp0001455] [Medline: 39715377]
10. Bean DJ, Dryland A, Rashid U, Tuck NL. The determinants and effects of chronic pain stigma: a mixed methods study and the development of a model. J Pain 2022 Oct;23(10):1749-1764. [doi: 10.1016/j.jpain.2022.05.006] [Medline: 35700874]
11. Pearl RL, Sheynblyum M. How weight bias and stigma undermine healthcare access and utilization. Curr Obes Rep 2025 Jan 20;14(1):11. [doi: 10.1007/s13679-025-00605-3] [Medline: 39832116]
12. Puhl RM, Himmelstein MS, Hateley-Browne JL, Speight J. Weight stigma and diabetes stigma in U.S. adults with type 2 diabetes: associations with diabetes self-care behaviors and perceptions of health care. Diabetes Res Clin Pract 2020 Oct;168:108387. [doi: 10.1016/j.diabres.2020.108387] [Medline: 32858100]
13. van der Kooij YL, den Daas C, Bos AER, Willems RA, Stutterheim SE. Correlates of internalized HIV stigma: a comprehensive systematic review. AIDS Educ Prev 2023 Apr;35(2):158-172. [doi: 10.1521/aeap.2023.35.2.158] [Medline: 37129595]
14. Crenshaw K. Mapping the margins: intersectionality, identity politics, and violence against women of color. Stanford Law Rev 1991 Jul;43(6):1241. [doi: 10.2307/1229039]
15. Turan JM, Elafros MA, Logie CH, et al. Challenges and opportunities in examining and addressing intersectional stigma and health. BMC Med 2019 Feb 15;17(1):7. [doi: 10.1186/s12916-018-1246-9] [Medline: 30764816]
16. Fino E, Mazzetti M, Russo PM. Psoriasis-related stigma and its intersection with intergroup bias in medical students. J Am Acad Dermatol 2021 May;84(5):1432-1434. [doi: 10.1016/j.jaad.2020.06.060] [Medline: 32574586]
17. Goodin BR, Owens MA, White DM, et al. Intersectional health-related stigma in persons living with HIV and chronic pain: implications for depressive symptoms. AIDS Care 2018 Jun;30(sup2):66-73. [doi: 10.1080/09540121.2018.1468012] [Medline: 29848042]
18. Himmelstein MS, Puhl RM, Quinn DM. Intersectionality: an understudied framework for addressing weight stigma. Am J Prev Med 2017 Oct;53(4):421-431. [doi: 10.1016/j.amepre.2017.04.003] [Medline: 28579331]
19. Jackson-Best F, Edwards N. Stigma and intersectionality: a systematic review of systematic reviews across HIV/AIDS, mental illness, and physical disability. BMC Public Health 2018 Jul 27;18(1):919. [doi: 10.1186/s12889-018-5861-3] [Medline: 30049270]
20. Pearl RL, Donze LF, Rosas LG, et al. Ending weight stigma to advance health equity. Am J Prev Med 2024 Nov;67(5):785-791. [doi: 10.1016/j.amepre.2024.06.021] [Medline: 38945180]
21. van Brakel WH, Cataldo J, Grover S, et al. Out of the silos: identifying cross-cutting features of health-related stigma to advance measurement and intervention. BMC Med 2019 Feb 15;17(1):13. [doi: 10.1186/s12916-018-1245-x] [Medline: 30764817]
22. Drapalski AL, Lucksted A, Brown CH, Fang LJ. Outcomes of ending self-stigma, a group intervention to reduce internalized stigma, among individuals with serious mental illness. Psychiatr Serv 2021 Feb 1;72(2):136-142. [doi: 10.1176/appi.ps.201900296] [Medline: 33234053]
23. Ma PHX, Chan ZCY, Loke AY. Self-stigma reduction interventions for people living with HIV/AIDS and their families: a systematic review. AIDS Behav 2019 Mar;23(3):707-741. [doi: 10.1007/s10461-018-2304-1] [Medline: 30298241]
24. Mittal D, Sullivan G, Chekuri L, Allee E, Corrigan PW. Empirical studies of self-stigma reduction strategies: a critical review of the literature. Psychiatr Serv 2012 Oct;63(10):974-981. [doi: 10.1176/appi.ps.201100459] [Medline: 22855130]
25. Pearl RL, Wadden TA, Bach C, et al. Effects of a cognitive-behavioral intervention targeting weight stigma: a randomized controlled trial. J Consult Clin Psychol 2020 May;88(5):470-480. [doi: 10.1037/ccp0000480] [Medline: 31971410]
26. Pearl RL, Wadden TA, Bach C, et al. Long-term effects of an internalized weight stigma intervention: a randomized controlled trial. J Consult Clin Psychol 2023 Jul;91(7):398-410. [doi: 10.1037/ccp0000819] [Medline: 37155264]
27. Fujisawa D, Hagiwara N. Cancer stigma and its health consequences. Curr Breast Cancer Rep 2015 Sep;7(3):143-150. [doi: 10.1007/s12609-015-0185-0]
28. Barlow DH, Farchione TJ, Sauer-Zavala S, et al. Unified Protocol for Transdiagnostic Treatment of Emotional Disorders: Workbook, 2nd edition: Oxford University Press; 2017. [doi: 10.1093/med-psych/9780190686017.001.0001]
29. Conway CC, Snorrason I, Beard C, Forgeard M, Cuthbert K, Björgvinsson T. A higher order internalizing dimension predicts response to partial hospitalization treatment. Clin Psychol Sci 2021 May;9(3):373-384. [doi: 10.1177/2167702620959291]
30. Forgeard M, Beard C, Kirakosian N, Bjorgvinsson T. Research in partial hospital settings. In: Practice-Based Research: A Guide for Clinicians, 1st edition: Routledge; 2018:212-240. [doi: 10.4324/9781315524610-12]

XSL•FO
RenderX

31. Yalom ID, Lescze M. Specialized therapy groups. In: The Theory and Practice of Group Psychotherapy, 5th edition: Basic Books/Hachette Book Group; 2005:475-524.
32. Onken LS, Carroll KM, Shoham V, Cuthbert BN, Riddle M. Reenvisioning clinical science: unifying the discipline to improve the public health. Clin Psychol Sci 2014 Jan 1;2(1):22-34. [doi: 10.1177/2167702613497932] [Medline: 25821658]
33. Pearl RL, Li Y, Groshon LC, et al. Measuring internalized health-related stigma across health conditions: development and validation of the I-HEARTS Scale. BMC Med 2024 Oct 8;22(1):435. [doi: 10.1186/s12916-024-03661-z] [Medline: 39379928]
34. Pearl RL, Bach C, Wadden TA. Development of a cognitive-behavioral intervention for internalized weight stigma. J Contemp Psychother 2023 Jun;53(2):165-172. [doi: 10.1007/s10879-022-09543-w] [Medline: 37483684]
35. Biggs K, Hind D, Gossage-Worrall R, et al. Challenges in the design, planning and implementation of trials evaluating group interventions. Trials 2020 Jan 29;21(1):116. [doi: 10.1186/s13063-019-3807-4] [Medline: 31996259]
36. Freedland KE. Pilot trials in health-related behavioral intervention research: problems, solutions, and recommendations. Health Psychol 2020 Oct;39(10):851-862. [doi: 10.1037/hea0000946] [Medline: 32614197]
37. Teresi JA, Yu X, Stewart AL, Hays RD. Guidelines for designing and evaluating feasibility pilot studies. Med Care 2022 Jan 1;60(1):95-103. [doi: 10.1097/MLR.0000000000001664] [Medline: 34812790]
38. Boyd Ritsher J, Otilingam PG, Grajales M. Internalized stigma of mental illness: psychometric properties of a new measure. Psychiatry Res 2003 Nov;121(1):31-49. [doi: 10.1016/j.psychres.2003.08.008]
39. Rybak CJ, Brown B. Assessment of internalized shame: validity and reliability of the internalized shame scale. Alcohol Treat Quart 1996;14(1):71-83. [doi: 10.1300/J020V14N01_07]
40. Russell DW. UCLA Loneliness Scale (Version 3): reliability, validity, and factor structure. J Pers Assess 1996 Feb;66(1):20-40. [doi: 10.1207/s15327752jpa6601_2] [Medline: 8576833]
41. Kroenke K, Spitzer RL. The PHQ-9: a new depression diagnostic and severity measure. Psychiatr Ann 2002 Sep;32(9):509-515. [doi: 10.3928/0048-5713-20020901-06]
42. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
43. Craske M, Wittchen U, Bogels S, Stein M, Andrews G, Lebeu R. Severity measure for social anxiety disorder (social phobia) – adult. American Psychiatric Association. 2013. URL: https://www.psychiatry.org/psychiatrists/practice/dsm/educational-resources/assessment-measures [accessed 2025-07-22]
44. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. J Health Soc Behav 1983 Dec;24(4):385-396. [doi: 10.2307/2136404]
45. Moss-Morris R, Weinman J, Petrie K, Horne R, Cameron L, Buick D. The Revised Illness Perception Questionnaire (IPQ-R). Psychol Health 2002 Jan;17(1):1-16. [doi: 10.1080/08870440290001494]
46. Ware JE, Kosinski M, Keller SD. A 12-item short-form health survey: construction scales and preliminary tests of reliability and validity. Med Care 1996;34:220-233. [doi: 10.1097/00005650-199603000-00003]
47. Moriarty DG, Zack MM, Kobau R. The Centers for Disease Control and Prevention's Healthy Days Measures - population tracking of perceived physical and mental health over time. Health Qual Life Outcomes 2003 Sep 2;1(37):1-8. [doi: 10.1186/1477-7525-1-37] [Medline: 14498988]
48. Williams DR, Yu Y, Jackson JS, Anderson NB. Racial differences in physical and mental health: socio-economic status, stress and discrimination. J Health Psychol 1997 Jul;2(3):335-351. [doi: 10.1177/135910539700200305] [Medline: 22013026]
49. Andersson GZ, Reinius M, Eriksson LE, et al. Stigma reduction interventions in people living with HIV to improve health-related quality of life. Lancet HIV 2020 Feb;7(2):e129-e140. [doi: 10.1016/S2352-3018(19)30343-1] [Medline: 31776098]
50. D'Adamo L, Shonrock AT, Monocello L, et al. Psychological interventions for internalized weight stigma: a systematic scoping review of feasibility, acceptability, and preliminary efficacy. J Eat Disord 2024 Nov 29;12(1):197. [doi: 10.1186/s40337-024-01132-7] [Medline: 39614347]
51. Traxler J, Stuhlmann CFZ, Graf H, Rudnik M, Westphal L, Sommer R. Interventions to reduce skin-related self-stigma: a systematic review. Acta Derm Venereol 2024 Sep 10;104(104):adv40384. [doi: 10.2340/actadv.v104.40384] [Medline: 39254291]
52. Burke E, Pyle M, Machin K, Varese F, Morrison AP. The effects of peer support on empowerment, self-efficacy, and internalized stigma: a narrative synthesis and meta-analysis. Stigma Health 2019;4(3):337-356. [doi: 10.1037/sah0000148]
53. Akin-Odanye EO, Husman AJ. Impact of stigma and stigma-focused interventions on screening and treatment outcomes in cancer patients. Ecancermedicalscience 2021;15:1308. [doi: 10.3332/ecancer.2021.1308] [Medline: 34824631]
54. Ashton-James CE. Beyond future directions: what can we do to address the stigma of chronic pain today? Pain 2024;165(8):1657-1659. [doi: 10.1097/j.pain.0000000000003244]
55. Yanos PT, Lucksted A, Drapalski AL, Roe D, Lysaker P. Interventions targeting mental health self-stigma: a review and comparison. Psychiatr Rehabil J 2015 Jun;38(2):171-178. [doi: 10.1037/prj0000100] [Medline: 25313530]
56. Cook JE, Purdie-Vaughns V, Meyer IH, Busch JTA. Intervening within and across levels: a multilevel approach to stigma and public health. Soc Sci Med 2014 Feb;103:101-109. [doi: 10.1016/j.socscimed.2013.09.023] [Medline: 24513229]

XSL•FO
RenderX

Abbreviations

I-HEARTS: Internalized Health-Related Stigma
Weight BIAS: Weight Bias Internalization and Stigma

Edited by A Mavragani; submitted 02.12.24; peer-reviewed by E Toni, N Loecher; revised version received 02.04.25; accepted 08.04.25; published 29.07.25.

Please cite as:
Pearl RL, Saunders D, Groshon LC, Li Y, Shonrock A, Puhl RM, Driscoll KA, Manavalan P, Gelfand JM, Wadden TA, Westen SC, Montanez-Wiscovich M, Lou XY
Feasibility and Acceptability of a Psychological Intervention for Internalized Health-Related Stigma Among Adults With Chronic Health Conditions: Preliminary Investigation
JMIR Form Res 2025;9:e69548
URL: https://formative.jmir.org/2025/1/e69548
doi:10.2196/69548

© Rebecca L Pearl, Danielle Saunders, Laurie C Groshon, Yulin Li, Abigail Shonrock, Rebecca M Puhl, Kimberly A Driscoll, Preeti Manavalan, Joel M Gelfand, Thomas A Wadden, Sarah C Westen, Marjorie Montanez-Wiscovich, Xiang-Yang Lou. Originally published in JMIR Formative Research (https://formative.jmir.org), 29.7.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Prevalence of Multiple Chronic Conditions Among Adults in the All of Us Research Program: Exploratory Analysis

Xintong Li[1*]; Caitlin Dreisbach[1,2*], PhD, RN; Carolina M Gustafson[3*], PhD, RN, ACNP; Komal Patel Murali[4*], PhD, RN, ACNP; Theresa A Koleck[3*], PhD, RN

[1]Goergen Institute for Data Science and Artificial Intelligence, University of Rochester, Rochester, NY, United States
[2]School of Nursing, University of Rochester, Rochester, NY, United States
[3]School of Nursing, University of Pittsburgh, Pittsburgh, PA, United States
[4]Rory Meyers College of Nursing, New York University, New York, NY, United States
[*]all authors contributed equally

Corresponding Author:
Xintong Li
Goergen Institute for Data Science and Artificial Intelligence, University of Rochester, Rochester, NY, United States

Abstract

Background: The growing prevalence of multiple chronic conditions (MCC) has significant impacts on health care systems and quality of life. Understanding the prevalence of MCC throughout adulthood offers valuable insights into the evolving burden of chronic diseases and provides strategies for more effective health care outcomes.

Objective: This study estimated the prevalence and combinations of MCC among adult participants enrolled in the *All of Us (AoU)* Research Program, especially studying the variations across age categories.

Methods: We conducted an exploratory analysis using electronic health record (EHR) data from adult participants (N=242,828) in the version 7 Controlled Tier *AoU* Research Program data release. Data analysis was conducted using Python in a Jupyter notebook environment within the *AoU* Researcher Workbench. Descriptive statistics included condition frequencies, the number of chronic conditions per participant, and prevalence according to age categories. The presence of a chronic condition was determined by documentation of one or more *ICD-10* (*International Statistical Classification of Diseases, Tenth Revision*) codes for the respective condition. Age categories were established and aligned with diagnosis dates and stages of adulthood (early adulthood: 18-39 years; middle adulthood: 40-49 years; late middle adulthood: 50-64 years; late adulthood: 65-74 years; advanced old age: 75-89 years).

Results: Our findings demonstrated that approximately 76% (n=183,753) of *AoU* participants were diagnosed with MCC, with over 40% (n=98,885) having 6 or more conditions and prevalence increasing with age (from 33.78% in early adulthood to 68.04% in advanced old age). The most frequently occurring MCC combinations varied by age category. Participants aged 18-39 years primarily presented mental health–related MCC combinations (eg, anxiety and depressive disorders; n=845), whereas those aged 40-64 years frequently had combinations of physical conditions such as fibromyalgia, chronic pain, fatigue, and arthritis (204 in middle adulthood and 457 in late middle adulthood). In late adulthood and advanced old age, hyperlipidemia and hypertension were the most commonly occurring MCC combinations (n=200 and n=59, respectively).

Conclusions: We report notable prevalence of MCC throughout adulthood and variability in MCC combinations by age category in *AoU* participants. The significant prevalence of MCC underscores a considerable public health challenge, revealed by distinct condition combinations that shift across different life stages. Early adulthood is characterized predominantly by mental health conditions, transitioning to cardiometabolic and physical health conditions in middle, late, and advanced ages. These findings highlight the need for targeted, innovative care modalities and population health initiatives to address the burden of MCC throughout adulthood.

(JMIR Form Res 2025;9:e69138) doi:10.2196/69138

KEYWORDS

electronic health records; multiple chronic conditions; prevalence; secondary data analysis; United States; All of Us

Introduction

Multimorbidity is an increasingly prevalent global health challenge characterized by diverse, co-occurring chronic diseases, differing across life stages, demographics, and socioeconomic factors, necessitating shifts in health care practice and outcomes [1]. Multiple chronic conditions (MCC) is defined as the presence of two or more chronic conditions such as

XSL•FO
RenderX

Alzheimer disease, arthritis, asthma, cancer, chronic kidney disease, diabetes, heart failure, and hypertension [2]. Approximately 80% of adults aged 65 years and older in the United States are living with MCC [3]. Older adults diagnosed with MCC often face significant care management challenges, including unplanned hospital readmissions and unmet care needs, which indicates the urgent need for enhanced support and tailored interventions for this vulnerable population [4]. Being diagnosed with MCC, however, is not limited to older adults. One study estimated that 22% of young adults, aged 18-34 years, are diagnosed with MCC [5]. Moreover, research suggests that more contemporary generations of adults have a greater MCC burden and are diagnosed with MCC at earlier ages than previous generations [6]. Estimation of the prevalence of MCC throughout all stages of adulthood is a critical reflection of the MCC burden; it is important to examine the prevalence of MCC broadly using regularly updated data sources to inform targeted prevention and management strategies and resource prioritization. Thus, the purpose of this study is to estimate the prevalence of MCC for adult participants in the National Institutes of Health's *All of Us* (*AoU*) Research Program dataset—a curated disease and population-agnostic, longitudinal biomedical dataset of diverse individuals living in the United States—to gain a deeper understanding of MCC throughout adulthood (ie, early adulthood, middle adulthood, late middle adulthood, late adulthood, and advanced old age).

Methods

Participants

We conducted an exploratory cross-sectional analysis using the *AoU* Controlled Tier version 7 data release in the secure, cloud-based *AoU* Researcher Workbench platform (date of first access for research purposes: September 1, 2023; date of last access: June 16, 2024). Registered researchers can access the data through the *AoU* Research Hub [7]. This study was exempt from ethical approval for human subjects research as only deidentified data were analyzed.

Using the Cohort Builder tool, we identified all adult participants 18 years of age or older who consented to share electronic health record (EHR) data and had at least one *ICD-10* (*International Statistical Classification of Diseases, Tenth Revision*) code documented for any healthcare purpose (ie, not limited to a chronic condition-related code) to ensure the availability of EHR data (n=242,828). For all participants meeting these eligibility criteria, we extracted relevant *ICD-10* codes and the corresponding date of documentation for 58 chronic conditions (Multimedia Appendix 1) cataloged in the Centers for Medicare & Medicaid Services (CMS) Chronic Conditions Data Warehouse [8]. CMS supports a Chronic Conditions Data Warehouse that contains computable phenotypes for "common chronic conditions" plus "other chronic or potentially disabling conditions," which include additional chronic health, mental health, and substance abuse conditions [9]. We operationalized the presence of a chronic condition as documentation of one or more *ICD-10* codes for the respective condition (Multimedia Appendix 1). Due to reliance on *ICD-10* codes for both cohort identification and as the source of condition information, we limited participant visits to on or after October 1, 2015, that is, initiation of *ICD-10* implementation. Thus, data included in this analysis are from October 1, 2015, to July 1, 2022, that is, latest date available in the version 7 release.

Data Analysis

We used the Python programming language in Jupyter Notebook to chronologically order the chronic condition *ICD-10* codes by date of documentation for each participant. By distinguishing the onset date of the second unique chronic condition, we were able to identify participants with MCC. We calculated the number of participants with 0, 1, 2, 3, 4, 5, and 6+ chronic conditions in the overall cohort and stratified by age category, that is, early adulthood (18-39 years), middle adulthood (40-49 years), late middle adulthood (50-64 years), late adulthood (65-74 years), and advanced old age (75-89 years) [10,11]. Age categorization supports methodological requirements of analyses, promotes interpretability of results for clinicians and policy makers, and helps capture changes in physical health and social roles throughout adulthood. We used the age at first documentation of the second chronic condition for categorization. We used chi-square tests to assess differences in demographic variables between participants with 0 or 1 chronic condition and participants with MCC. This approach tested for differences in the frequencies of race, ethnicity, gender, marital status, education, health insurance status, and employment status. We used $P<.05$ to indicate statistical significance. Finally, we calculated the frequency of unique MCC combinations by the number of chronic conditions (ie, 2, 3, 4, 5, and 6+) and age category.

Ethical Considerations

This study analyzed deidentified data from participants enrolled in the *AoU* Research Program. Participants in the *AoU* Research Program were recruited either in-person at *AoU* enrollment sites or on the *AoU* web-based portal [12]. Participants gave informed consent when they joined the program, agreeing that their anonymized data could be used broadly for biomedical research [12]. Enrollment and data collection of participants in the *AoU* program is approved by the *AoU* institutional review board (IRB) [13].

We analyzed the EHR data in the secure, cloud-based *AoU* Researcher Workbench. All data used in this analysis were deidentified without personal identification including shifted dates for categorization of age. Researchers included in this analysis were registered within the *AoU* Research Program and the Researcher Workbench, completed the required training for data access, and signed the Data User Code of Conduct. All analyses were performed following *AoU* policies and guidelines to ensure confidentiality and responsible use of data [14]. Analyses for this study did not require subsequent IRB approval because the study does not constitute research involving human subjects according to the *AoU* Research Program IRB [12]. The data used in this analysis do not include the elements necessary to identify individuals so the study is not subject to the Common Rule protections 45 CFR part 46 [15].

XSL•FO
RenderX

Results

Participants Statistics

Approximately 76% (n=183,753) of participants were diagnosed with MCC (Figure 1), with over 40% of participants (n=98,885) diagnosed with 6 or more conditions (Table 1). Of the remaining participants, 11% (n=27,122) were diagnosed with a single chronic condition, and 13% (n=31,953) did not have a chronic condition documented. Overall, participants (Table 2) were on average 49.8 (SD 16.7) years of age and primarily women (n=147,991, 60.9%), White (n=137,039, 54.6%), not Hispanic or Latino (n=187,586, 77.3%), married or partnered (n=121,282, 49.9%), highly educated (n=166,815, 68.7% with some college or higher), insured (n=223,842, 92.2%), and not currently employed (n=111,229, 47.7%). Compared to participants diagnosed with 0 or 1 conditions, those diagnosed with MCC were older (age: mean 52.8 years vs 40.4 years). In addition, a higher percentage of participants with MCC were men (n=68,776, 37.4% vs n=19,956, 33.8%), White (n=107,725, 58.6% vs n=29,314, 49.6%), not Hispanic or Latino (n=145,393, 79.1% vs n=42,193, 71.4%), insured (n=170,985, 93.1% vs n=52,857, 89.5%), ever married (n=139,118, 75.7% vs n=39,264, 66.5%), and not currently employed (n=96,339, 52.4% vs n=19,484, 33.0%), with a much greater frequency of retired participants (n=48,565, 26.4% vs n=6025, 10.2%), compared to participants diagnosed with 0 or 1 conditions. The percentage of participants within an age category diagnosed with 6 or more chronic conditions increased with age (Table 1), from 33.78% in early adulthood to 68.04% in advanced old age.

Figure 1. Flow diagram.

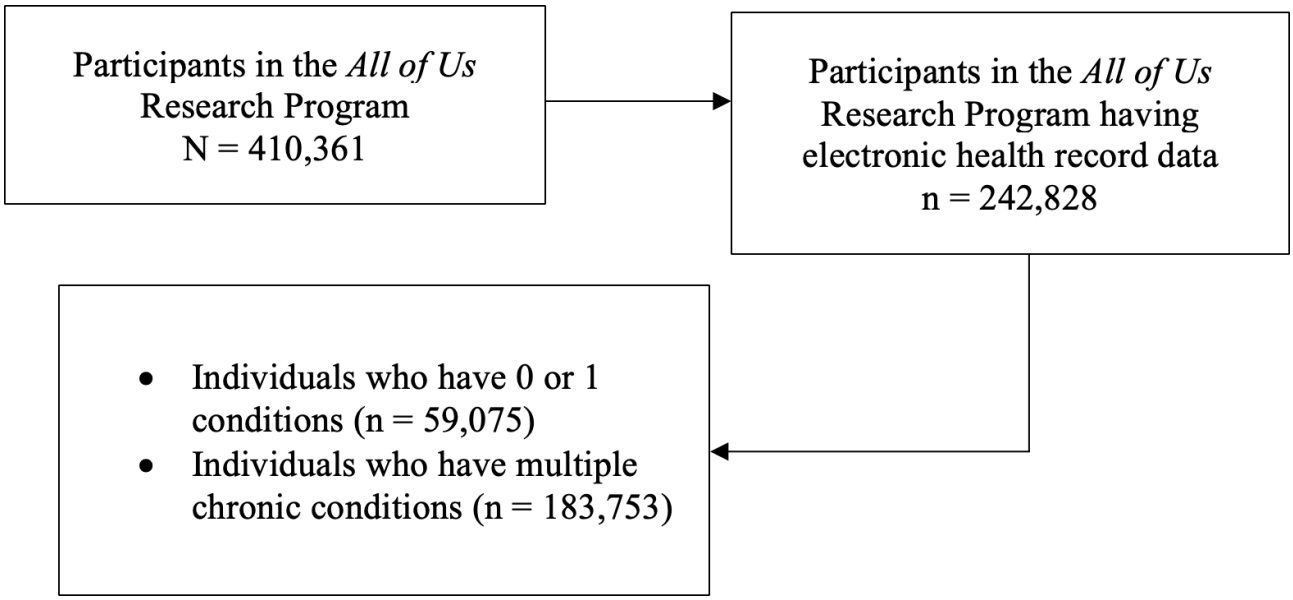

XSL•FO
RenderX

Table . Prevalence of multiple chronic conditions by age group among adults enrolled in the *All of Us* Research Program.

Conditions		Participants, n (%)
Overall (n=183,753)		
	2	24,039 (9.90)
	3	21,910 (9.02)
	4	20,075 (8.27)
	5	18,844 (7.76)
	6+	98,885 (40.72)
Early adulthood (18-39 years; n=40,237)		
	2	9327 (23.18)
	3	7235 (17.98)
	4	5648 (14.04)
	5	4435 (11.02)
	6+	13,592 (33.78)
Middle adulthood (40-49 years; n=28,906)		
	2	4139 (14.32)
	3	3689 (12.76)
	4	3244 (11.22)
	5	3052 (10.56)
	6+	14,782 (51.14)
Late middle adulthood (50-64 years; n=66,705)		
	2	6786 (10.17)
	3	6945 (10.41)
	4	6895 (10.34)
	5	6715 (10.07)
	6+	39,364 (59.01)
Late adulthood (65-74 years; n=35,549)		
	2	2778 (7.81)
	3	3019 (8.49)
	4	3227 (9.08)
	5	3462 (9.74)
	6+	23,063 (64.88)
Advanced old age (75-89 years; n=11,289)		
	2	811 (7.18)
	3	844 (7.48)
	4	919 (8.14)
	5	1034 (9.16)
	6+	7681 (68.04)

XSL•FO
RenderX

Table . Characteristics of participants in the *All of Us* Research Program with the presence of multiple chronic conditions (2 or more chronic conditions; N=242,828).

		Overall (N=242,828)	0 or 1 condition (n=59,075)	Multiple chronic conditions (n=183,753)	*P* value[a]
Age[b] (years), mean (SD)		49.8 (16.7)	40.4 (16.5)	52.8 (15.7)	<.001
Race, n (%)					<.001
	Asian	6514 (2.7)	2816 (4.8)	3698 (2.0)	
	Black	43,431 (17.9)	9991 (16.9)	33,440 (18.2)	
	Middle Eastern or North African	1413 (0.6)	462 (0.8)	951 (0.5)	
	Native Hawaiian or Pacific Islander	303 (0.1)	78 (0.1)	225 (0.1)	
	White	137,039 (56.4)	29,314 (49.6)	107,725 (58.6)	
	Multiracial	4233 (1.7)	1291 (2.2)	2942 (1.6)	
	Unknown	49,895 (20.5)	15,123 (25.6)	34,772 (18.9)	
Ethnicity, n (%)					<.001
	Not Hispanic or Latino	187,586 (77.3)	42,193 (71.4)	145,393 (79.1)	
	Hispanic or Latino	46,011 (18.9)	14,978 (25.4)	31,033 (16.9)	
	Unknown or not reported	9231 (3.8)	1904 (3.2)	7327 (4.0)	
Gender, n (%)					<.001
	Man	88,732 (36.5)	19,956 (33.8)	68,776 (37.4)	
	Woman	147,991 (60.9)	37,727 (63.9)	110,264 (60.0)	
	Nonbinary, transgender, or other gender	1622 (0.7)	479 (0.8)	1143 (0.6)	
	Unknown	4483 (1.8)	913 (1.5)	3570 (1.9)	
Health insurance status, n (%)					<.001
	Insured	223,842 (92.2)	52,857 (89.5)	170,985 (93.1)	
	Uninsured	11,268 (4.6)	4074 (6.9)	7194 (3.9)	
	Unknown	7718 (3.2)	2144 (3.6)	5574 (3.0)	
Employment status, n (%)					<.001
	Employed	99,435 (40.9)	31,910 (54.0)	67,525 (36.7)	
	Unemployed	20,477 (8.4)	5,634 (9.5)	14,843 (8.1)	
	Unable to work	27,250 (11.2)	2899 (4.9)	24,351 (13.3)	
	Retired	54,590 (22.5)	6025 (10.2)	48,565 (26.4)	
	Homemaker	8917 (3.7)	2704 (4.6)	6213 (3.4)	
	Student	4594 (1.9)	2515 (4.3)	2079 (1.1)	
	Multiple	18,653 (7.7)	4977 (8.4)	13,676 (7.4)	
	Unknown	8912 (3.7)	2411 (4.1)	6501 (3.5)	
Annual income (US $), n (%)					<.001
	Less than 10,000	31,274 (12.9)	7510 (12.7)	23,764 (12.9)	
	10,000 to 25,000	28,491 (11.7)	5702 (9.7)	22,789 (12.4)	
	25,000 to 35,000	17,271 (7.1)	4234 (7.2)	13,037 (7.1)	
	35,000 to 50,000	19,232 (7.9)	4769 (8.1)	14,463 (7.9)	
	50,000 to 75,000	25,428 (10.5)	6066 (10.3)	19,362 (10.5)	
	75,000 to 100,000	19,563 (8.1)	4752 (8.0)	14,811 (8.1)	

XSL•FO
RenderX

		Overall (N=242,828)	0 or 1 condition (n=59,075)	Multiple chronic conditions (n=183,753)	*P* value[a]
	100,000 to 150,000	23,765 (9.8)	6221 (10.5)	17,544 (9.5)	
	150,000 to 200,000	10,873 (4.5)	3020 (5.1)	7853 (4.3)	
	Greater than 200,000	15,186 (6.3)	4331 (7.3)	10,855 (5.9)	
	Unknown	51,745 (21.3)	12,470 (21.1)	39,275 (21.4)	
Highest level of education, n (%)					<.001
	Never attended	367 (0.2)	58 (0.1)	309 (0.2)	
	Up to high school	7710 (3.2)	1671 (2.8)	6039 (3.3)	
	Some high school	14,131 (5.8)	3351 (5.7)	10,780 (5.9)	
	High school graduate/GED	46,474 (19.1)	11,122 (18.8)	35,352 (19.2)	
	Some college	62,658 (25.8)	13,233 (22.4)	49,425 (26.9)	
	College graduate	53,993 (22.2)	15,107 (25.6)	38,886 (21.2)	
	Advanced degree	50,164 (20.7)	12,847 (21.7)	37,317 (20.3)	
	Unknown	7331 (3.0)	1686 (2.9)	5645 (3.1)	
Marital status, n (%)					<.001
	Married or living with partner	121,282 (49.9)	30,287 (51.3)	90,995 (49.5)	
	Divorced or separated	42,814 (17.6)	7351 (12.4)	35,463 (19.3)	
	Widowed	14,286 (5.9)	1626 (2.8)	12,660 (6.9)	
	Never married	56,246 (23.2)	17,710 (30.0)	38,536 (21.0)	
	Unknown	8200 (3.4)	2101 (3.6)	6099 (3.3)	

[a]*P* values calculated using chi-square tests indicate there is a significant difference in frequencies between participants with 0 or 1 chronic condition and those with multiple chronic conditions (2 or more conditions).

[b]Age (years) of participants on October 1, 2015, that is, the initiation of *ICD-10* (*International Statistical Classification of Diseases, Tenth Revision*) code implementation, for comparison between groups. Please note that age at first documentation of the second chronic condition, not age on October 1, 2015, was used for age categorization for participants diagnosed with multiple chronic conditions.

Conditions Across Age Distributions

The top 5 most frequently diagnosed MCC combinations in each age category are detailed in Multimedia Appendix 1. The most frequently occurring combination in each age categorization was as follows: early adulthood: anxiety disorders and depression, bipolar disorder, or other depressive mood disorders (n=845); middle and late middle adulthood: fibromyalgia, chronic pain and fatigue, and rheumatoid arthritis or osteoarthritis (n=204 and n=457, respectively); and late adulthood and advanced old age: hyperlipidemia and hypertension (n=200 and n=59, respectively).

Discussion

Principal Findings

In this study, we estimated MCC prevalence in adults living in the United States from early adulthood to advanced old age using *ICD-10* codes from participants enrolled in the *AoU* program. We found that approximately 76% of participants have MCC, with higher prevalence as age increases. The most frequent MCC combinations varied notably by age category. Mental health conditions (ie, anxiety disorders and depression, bipolar disorder, or other depressive mood disorders) were most common in early adulthood, whereas combinations of physical health conditions (eg, fibromyalgia, chronic pain, fatigue, arthritis, hyperlipidemia, and hypertension) predominated in middle and older adulthood. These findings emphasize the value of the *AoU* program in identifying the MCC burden throughout adulthood. We believe the percentage of adults with MCC in this study is higher than that reported previously due to our robust inclusion of 58 chronic conditions, selected based on the CMS Chronic Conditions Data Warehouse.

Comparison With Prior Work

Our results highlight distinct patterns in MCC combinations that vary by age category. Given the notable prevalence of MCC across adulthood, there is a continued need for innovative care modalities and public health initiatives to address the multifaceted nature of MCC [16]. MCC are associated with clinical complexity characterized by increased symptom burden, high health care usage and cost, and extensive caregiver strain [17]. The Agency for Healthcare Research and Quality states that care for people with MCC requires alignment among available resources, policies, and the workforce. For every dollar of health care spending in the United States, US $0.71 goes to treating individuals with MCC [3]. Subgroups at greater risk for MCC can benefit from early detection and prevention

XSL•FO
RenderX

strategies, which may reduce hospital admissions and associated health care costs [3]. Ongoing research efforts that seek to both reduce the burden of MCC on the health care system and the development of person- and family-centered care strategies across the health care continuum are needed [17]. Public health leaders should coordinate efforts with clinicians, administrators, and communities to support changes in care management and outcomes across the lifespan.

We found that the most frequent MCC combinations vary by age category. Notably, a combination of mental, rather than physical, health conditions (ie, anxiety disorders and depression, bipolar disorder, or other depressive mood disorders) was the most common MCC combination in early adulthood. Following early adulthood, combinations of physical health conditions were the most common (eg, hyperlipidemia and hypertension in late adulthood and advanced old age). For precision health, characterizing MCC profiles and the most frequently co-occurring conditions in different age categories can inform targeted prevention and treatment approaches and allow clinicians to better anticipate and manage comorbidities in their patients [18].

Limitations

We acknowledge, however, that this initial analysis is exploratory and descriptive in nature. The characteristics of adult participants who are enrolled in the *AoU* program and consented to share EHR data—ie, women, White individuals, non-Hispanic/Latino individuals, married/partnered individuals, highly educated individuals, insured individuals, and those not currently employed—must also be considered when interpreting the reported estimates and generalizing results to the broader US, and worldwide, population. Another limitation of our analysis is the reliance on *ICD-10* codes for identifying chronic conditions, which could introduce inaccuracies. Computable phenotyping challenges, such as accurate identification and measurement of chronic conditions, are not unique to this analysis but rather an inherent aspect of analysis using EHR data. Computable phenotypes for chronic conditions could be enhanced to reduce the risk of misdiagnosis and refine prevalence estimates. One approach to support the validity of chronic condition phenotypes would be comparison of CMS-derived diagnoses to EHR computable phenotypes from other repositories, such as the Observational Health Data Sciences and Informatics (OHDSI [19]) Phenotype Library and Phenotype KnowledgeBase (PheKB [20]). OHDSI and PheKB phenotypes use value sets including and beyond codes (eg, laboratory values and medications). Another approach to enhance the validity of chronic condition diagnoses would be to compare the computable phenotype–derived diagnoses to self-reported diagnoses on *AoU* surveys. Without a true gold standard, triangulation of EHR and survey sources may be an ideal strategy to achieve diagnosis accuracy in future work.

Conclusions

This brief report fills an important knowledge gap and provides a foundation for future studies on MCC correlates, predictors, and outcomes by communicating current MCC prevalence estimates, which capture a wider range of chronic conditions, across the adult lifespan. The high prevalence of MCC in adults from early adulthood to advanced old age underscores the need for resource prioritization, targeted prevention, and improved treatment and care plans in individuals with co-occurring conditions.

Acknowledgments

Research reported in this publication was supported by the National Institute of Nursing Research of the National Institutes of Health (awards K01NR020504, R00NR017651, and T32008857). The *All of Us* Research Program would not be possible without the partnership of its participants.

Data Availability

This study involved the analysis of existing data from the *All of Us* Research Program. No new datasets were generated during this analysis. Researchers interested in accessing the *All of Us* Research Program dataset can register and apply for access through the secure, cloud-based Researcher Workbench [7]. Access requires institutional agreement, identity verification, completion of training, and signing the data user code of conduct.

Conflicts of Interest

CD and TAK are *All of Us* Researcher Ambassadors through Pyxis Partners, which is funded by the Division of Engagement and Outreach, *All of Us* Research Program, National Institutes of Health (award OD028404, Pyxis Partners). CD also serves on the *All of Us* Research Program Community Advisory Board. The authors have no other conflicts of interest to disclose.

Multimedia Appendix 1
Additional tables.
[DOCX File, 54 KB - formative_v9i1e69138_app1.docx]

References

1. Langenberg C, Hingorani AD, Whitty CJM. Biological and functional multimorbidity-from mechanisms to management. Nat Med 2023 Jul;29(7):1649-1657. [doi: 10.1038/s41591-023-02420-6] [Medline: 37464031]

XSL•FO
RenderX

2. Whitson HE, Boyd CM. Multiple chronic conditions. In: Schmader KE, Givens J, editors. UpToDate. URL: https://www.uptodate.com/contents/multiple-chronic-conditions [accessed 2024-05-22]
3. Agency for Healthcare Research and Quality. The importance of investing in care for people with multiple chronic conditions. Agency for Healthcare Research and Quality. URL: https://www.ahrq.gov/sites/default/files/wysiwyg/professionals/prevention-chronic-care/decision/mcc/mcc_infographic.pdf [accessed 2024-05-22]
4. Coatsworth-Puspoky R, Dahlke S, Duggleby W, Hunter KF. Older persons with multiple chronic conditions' experiences of unplanned readmission: An integrative review. Int J Older People Nurs 2022 Nov;17(6):e12481. [doi: 10.1111/opn.12481] [Medline: 35621261]
5. Watson KB, Carlson SA, Loustalot F, et al. Chronic Conditions Among Adults Aged 18—34 Years - United States, 2019. MMWR Morb Mortal Wkly Rep 2022 Jul 29;71(30):964-970. [doi: 10.15585/mmwr.mm7130a3] [Medline: 35900929]
6. Bishop NJ, Haas SA, Quiñones AR. Cohort trends in the burden of multiple chronic conditions among aging U.S. adults. J Gerontol B Psychol Sci Soc Sci 2022 Oct 6;77(10):1867-1879. [doi: 10.1093/geronb/gbac070] [Medline: 35642746]
7. Welcome to the All of Us Research Hub. All of Us Research Hub. URL: https://researchallofus.org/ [accessed 2025-05-01]
8. Centers for medicare and medicaid services. Condition categories. Chronic conditions data warehouse. URL: https://www2.ccwdata.org/web/guest/condition-categories [accessed 2024-05-22]
9. Centers for Medicare and Medicaid Services. Other chronic health, mental health, and potentially disabling conditions. Chronic conditions data warehouse. URL: https://www2.ccwdata.org/web/guest/condition-categories-other [accessed 2024-01-28]
10. Stages of aging & ideas for successful aging in place. Live in Place Designs. 2023. URL: https://liveinplacedesigns.com/the-5-stages-of-aging-at-home/ [accessed 2024-05-22]
11. Stangor C, Walinga W. Introduction to Psychology – 1st Canadian Edition. BCcampus. 2014. URL: https://opentextbc.ca/introductiontopsychology/ [accessed 2024-05-22]
12. All of Us Research Program Investigators, Denny JC, Rutter JL, et al. The "All of Us" Research Program. N Engl J Med 2019 Aug 15;381(7):668-676. [doi: 10.1056/NEJMsr1809937] [Medline: 31412182]
13. All of Us Institutional Review Board (IRB). All of Us Research Program. 2021 Jul 16. URL: https://allofus.nih.gov/about/who-we-are/institutional-review-board-irb-of-all-of-us [accessed 2025-03-19]
14. Protecting data and privacy. All of Us Research Program. 2025. URL: https://allofus.nih.gov/article/protecting-data-and-privacy [accessed 2025-03-24]
15. Federal Policy for the Protection of Human Subjects ('Common Rule'). Department of Health and Human Services. 2024 Mar 27. URL: https://www.hhs.gov/ohrp/regulations-and-policy/regulations/common-rule/index.html [accessed 2025-03-19]
16. Bierman AS, Wang J, O'Malley PG, Moss DK. Transforming care for people with multiple chronic conditions: Agency for Healthcare Research and Quality's research agenda. Health Serv Res 2021 Oct;56 Suppl 1(Suppl 1):973-979. [doi: 10.1111/1475-6773.13863] [Medline: 34378192]
17. Murali KP, Merriman JD, Yu G, Vorderstrasse A, Kelley AS, Brody AA. Complex care needs at the end of life for seriously ill adults with multiple chronic conditions. J Hosp Palliat Nurs 2023 Jun 1;25(3):146-155. [doi: 10.1097/NJH.0000000000000946] [Medline: 37040386]
18. Dreisbach C, Grayson S, Leggio K, Conway A, Koleck T. Predictors of unrelieved symptoms in All of Us Research Program participants with chronic conditions. J Pain Symptom Manage 2022 Dec;64(6):555-566. [doi: 10.1016/j.jpainsymman.2022.08.018] [Medline: 36096320]
19. Observational Health Data Sciences and Informatics. URL: https://ohdsi.org/ [accessed 2025-05-02]
20. What is the phenotype knowledgebase? PheKB. URL: https://phekb.org/ [accessed 2025-05-02]

Abbreviations

AoU: *All of Us*
CMS: Centers for Medicare & Medicaid Services
EHR: electronic health record
ICD-10: *International Statistical Classification of Diseases, Tenth Revision*
IRB: institutional review board
MCC: multiple chronic conditions
OHDSI: Observational Health Data Sciences and Informatics
PheKB: Phenotype Library and Phenotype KnowledgeBase

XSL•FO
RenderX

Edited by A Mavragani; submitted 22.11.24; peer-reviewed by JE Linder, R korutla; revised version received 21.03.25; accepted 07.04.25; published 12.05.25.

Please cite as:
Li X, Dreisbach C, Gustafson CM, Murali KP, Koleck TA
Prevalence of Multiple Chronic Conditions Among Adults in the All of Us Research Program: Exploratory Analysis
JMIR Form Res 2025;9:e69138
URL: https://formative.jmir.org/2025/1/e69138
doi:10.2196/69138

© Xintong Li, Caitlin Dreisbach, Carolina M Gustafson, Komal Patel Murali, Theresa A Koleck. Originally published in JMIR Formative Research (https://formative.jmir.org), 12.5.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Impact of Acute Respiratory Infections on Medical Absenteeism Among Military Personnel: Retrospective Cohort Study

Premikha M[1], MPH, MBBS; Jit Khong Goh[1,2], MPH, MBBS; Jing Qiang Ng[1], MBBS; Adeliza Mutalib[1], MSc; Huai Yang Lim[1], MPH, MBBS

[1]Force Health Group, HQ Medical Corps, Singapore Armed Forces, 701 Transit Road, #02-03, Singapore, Singapore

[2]Saw Swee Hock School of Public Health, National University of Singapore, Singapore, Singapore

Corresponding Author:
Premikha M, MPH, MBBS
Force Health Group, HQ Medical Corps, Singapore Armed Forces, 701 Transit Road, #02-03, Singapore, Singapore

Abstract

Background: Acute respiratory infections (ARI) are a significant challenge in military settings due to close communal living, which facilitates the rapid transmission of pathogens. A variety of respiratory pathogens contribute to ARI, each varying in prevalence, severity, and impact on organizational productivity. Understanding and mitigating the impact of ARI is critical for optimizing the health of military personnel and maintaining organizational productivity.

Objective: This retrospective study of surveillance data aims to identify pathogens causing ARI among servicemen and determine which pathogens contribute most to medical absenteeism, defined as the combined duration of the issued medical certificate and light duty.

Methods: From September 2023 to August 2024, anonymous nasopharyngeal swabs (BioFire FilmArray Respiratory Panel) were collected from Singapore Armed Forces servicemen presenting with ARI symptoms after a doctor's consultation at a local military camp's medical centre. The presence of fever and duration of medical certificate and light duty were self-reported by Singapore Armed Forces servicemen.

Results: A total of 1095 nasopharyngeal swabs were collected, of which 608 (55.5%) tested positive. The most common respiratory pathogen was human rhinovirus/enterovirus (HRV/HEV) in 303 (27.7%) individuals. The highest proportions of fever were observed in servicemen with influenza (62.8%, 27/43), SARS-CoV-2 (34.3%, 12/35), and parainfluenza (31.6%, 12/38). The odds of patients with influenza that have fever was 5.8 times higher than those of patients infected with HRV/HEV (95% CI 2.95 - 11.40, $P<.001$). The median duration of medical certificate, light duty, and medical absenteeism were 0 (IQR 0), 2 (IQR 2) and 2 (IQR 0) days, respectively. The odds of patients with influenza having a medical certificate with duration ≥1 day was 5.34 times higher than those in patients with HRV/HEV (95% CI 2.63 - 10.88, $P<.001$). No significant differences in the duration of medical absenteeism were found between HRV/HEV and other pathogens.

Conclusions: Compared to HRV/HEV, influenza infections were significantly associated with longer medical certificate duration. Nonetheless, there were no significant differences in the overall duration of medical absenteeism across pathogens, as servicemen infected with other pathogens were given light duty instead. These findings emphasize the need for pathogen-agnostic ARI measures. While influenza vaccinations are already mandatory for servicemen in local military camps, encouraging additional public health measures (eg, mask-wearing among symptomatic servicemen, COVID-19 vaccinations, therapeutics) can further reduce ARI incidence, minimize the duration of medical absenteeism, and mitigate the impact on organizational productivity.

(JMIR Form Res 2025;9:e69113) doi:10.2196/69113

KEYWORDS

respiratory infections; military; epidemiology; public health; surveillance

Introduction

Background

Acute respiratory infections (ARI) pose a significant challenge in military settings like the Singapore Armed Forces (SAF), where close quarters living and shared communal spaces, equipment, and furniture could facilitate the rapid transmission of respiratory pathogens [1,2]. Singapore's high humidity and tropical climate further exacerbate these risks, promoting the survival and spread of respiratory viruses [3]. These conditions not only increase ARI risk and impact the health of military personnel but also reduce organizational productivity due to medical absenteeism [1]. Despite ongoing surveillance, gaps remain in understanding how specific ARI pathogens impact medical absenteeism and readiness in high-density military environments.

XSL•FO
RenderX

A comprehensive understanding of the full spectrum of ARI pathogens is critical. While influenza is known for causing significant morbidity and mortality, HRV is typically the most prevalent pathogen and can lead to chronic infections and secondary bacterial infections, particularly in immunocompromised individuals [4]. A population-wide study in the United States found that the hospitalization burden was higher for HRV/HEV compared to influenza among adults aged <65 years [5,6]. Additionally, bacterial pathogens such as *Streptococcus pneumoniae* and *Mycoplasma pneumoniae* can act as primary agents or secondary infections [7,8]. A 2016 study in Singapore military recruits further emphasized the diversity of ARI pathogens, with 41% of 2647 recruits presenting with febrile respiratory illness (FRI) were predominantly caused by adenovirus and influenza viruses [9]. Among the remaining nonfebrile ARI individuals, rhinovirus accounted for 47% [9]. These findings highlight the need for expanded surveillance to capture the full spectrum of pathogens, beyond influenza to assess their contribution to the ARI burden in military settings.

Furthermore, it is unclear whether current influenza vaccination efforts sufficiently address the impact of influenza in military settings. Historical and contemporary data underscore the disproportionate impact of influenza on military populations, with attack rates reaching as high as 37 - 45% during outbreaks [10]. During the 1918 influenza pandemic, military personnel living in confined spaces faced significantly higher morbidity and mortality rates compared to civilian populations [10]. Recent outbreaks further highlight the heightened vulnerability of military personnel; a 2013 influenza A (H1N1) outbreak in a Swiss military camp infected 31% of soldiers, while a 2009 influenza A outbreak in a Peruvian Navy ship affected 22% of crew members [11,12]. Recognizing influenza as a significant cause of ARI with substantial morbidity, the SAF has conducted annual Northern Hemisphere influenza vaccination drives since 2015, achieving over 97% coverage in 2023 [2]. Additionally, the SAF conducts Southern Hemisphere influenza vaccinations for high-risk servicemen (eg, health care workers, medically vulnerable individuals), and strongly encourages them to receive additional doses of the COVID-19 vaccines, aligning with national guidelines [13]. Despite these measures, there is limited understanding of whether SAF's influenza vaccination programme sufficiently mitigate influenza's prevalence and severity in the SAF. These efforts are also insufficient to address the broader ARI burden, as evidenced by the prevalence of other ARI pathogens, highlighting the need for pathogen-specific data to understand the impact of various ARI pathogens on medical absenteeism.

Another critical gap is the limited understanding of the differential impact of febrile and nonfebrile ARI on medical absenteeism in military settings. The FRI, characterized by the presence of fever (≥38°C), serves as a marker of severity and reflects a robust immune response to more virulent pathogens [14]. Influenza and SARS-CoV-2 often cause FRI, whereas HRVs typically cause nonfebrile illnesses [3,15,16]. This distinction is crucial not only for clinical management but also for informing public health strategies aimed at controlling the spread of more severe FRI. However, findings from the Singapore military recruit study in 2016 showed that ARI cases without fever constituted the majority (59%) of respiratory illnesses and contributed significantly to the overall burden, emphasizing the need to account for both FRI and nonfebrile ARI in public health strategies [9]. Clinical management for ARI is mostly supportive, making vaccines critical for certain pathogens such as influenza and SARS-CoV-2. Recent events such as the 2023 ARI outbreak in the Philippines military that affected 20% of military students due to low influenza vaccine coverage [17], highlight the importance of maintaining high vaccine uptake. However, they also underscore the need to explore additional public health measures to reduce the broader ARI burden in military settings.

Study Objectives

Therefore, this study aimed to address these gaps by reviewing the results of the respiratory pathogen surveillance study from a local military camp to answer the following questions:

1. What is the distribution of pathogens responsible for ARI among SAF servicemen?
2. Which pathogens had the greatest impact on duration of medical absenteeism and how can this inform public health interventions to mitigate the impact?

By addressing these questions, this study seeks to determine if additional public health measures need to be instituted at local camps, beyond the existing influenza vaccinations. Understanding and mitigating the impact of ARI is critical for reducing medical absenteeism and optimizing organizational productivity.

Methods

Study Design and Recruitment

This retrospective study of surveillance data analyzed ARI cases from September 2023 to August 2024 at the medical centre of a local military camp in Singapore, the primary health care facility for servicemen. The camp's population primarily consists of young adult males, both conscripted and regular military personnel, who live in high-density accommodations conducive to ARI transmission. The medical centre typically manages 100-150 servicemen per day with acute conditions. The study used a convenience sampling approach, with recruitment conducted solely at the medical centre. Doctors screened servicemen presenting with ARI symptoms (eg, cough, sore throat, rhinorrhea) during routine consultations. Eligible servicemen were provided a standardized explanation of the study and invited to participate voluntarily. Those who agreed to participate were directed to the study team for data collection. Inclusion criteria included servicemen who (1) presented with at least one ARI symptom and (2) be willing to provide nasopharyngeal swabs for diagnostic testing. Exclusion criteria included servicemen diagnosed with other acute concurrent illnesses or those unwilling to provide swabs.

Anonymized nasopharyngeal swabs were collected from all eligible participants by the study team and tested using the BioFire FilmArray Respiratory Panel (BioFire Diagnostics), a multiplex polymerase chain reaction assay capable of detecting 22 respiratory pathogens. These pathogens were grouped into:

XSL•FO
RenderX

(1) coronaviruses (coronavirus 229E, HKU1, NL63, OC43), (2) SARS-CoV-2, (3) human metapneumovirus, (4) HRV/HEV, (5) influenza (influenza A, A/H1, A/H3, A/H1-2009, B), (6) parainfluenza (parainfluenza 1, 2, 3, 4), (7) respiratory syncytial virus, (8) adenovirus, and (9) bacteria (ie, *Bordetella parapetussis*, *Bordetella pertussis*, *Chlamydia pneumoniae*, *Mycoplasma pneumoniae*) for analysis. Pathogen testing was conducted immediately after swab collection using our mobile BioFire systems, which were serviced annually to minimize testing errors. The presence of fever (≥38°C) and duration of medical certificate and light duty were recorded through self-reported surveys administered directly by the surveillance team to ensure consistency in data collection. Our outcome of interest was duration of medical absenteeism, defined as the combined duration of medical certificate and light duty, representing the total period during which servicemen were unable to perform their primary jobs at full capacity. Servicemen with medical certificates were fully exempted from their duties, while those on light duty were partially exempted.

Statistical Analysis

Descriptive statistics were presented as mean (SD) for continuous variables and distribution (n, %) for categorical variables. Invalid swabs, coinfections, and bacterial cases were excluded from the analysis. Coinfections were defined as the presence of more than one pathogen detected in a single swab. Excluding coinfections aimed to minimize potential bias and ensured clearer attribution of outcomes (eg, fever, duration of medical absenteeism) to individual pathogens. Bacterial cases, which included those positive for bacterial pathogens such as *Mycoplasma pneumoniae*, were excluded due to the small number of cases detected during the study period.

Comparative analyses were performed to examine the differences in the prevalence of fever, duration of medical certificate, light duty, and duration of medical absenteeism among different pathogens using Pearson's χ^2 test of independence. Logistic regression was then performed for the outcomes of interest: (1) medical certificate (medical certificate ≥1 day), (2) light duty (>1 day), and (3) medical absenteeism (>1 day) to quantify outcome differences across pathogens. Human rhinovirus/human enterovirus was selected as the reference group due to its high prevalence in the study population, contributing to the stability of regression estimates. While "negative" cases could serve as an alternative reference group, they introduce heterogeneity, as "negative" indicates the absence of detectable pathogens by the BioFire Respiratory Panel, rather than the absence of infection. The Hosmer-Lemeshow goodness-of-fit test was used to ensure that all models characterized the data distribution accurately. Data were analyzed using Stata (version 17.0, StatCorp LLC) at a 5% significance level.

Ethical Considerations

This study was granted exemption from ethical review by the institutional review board (IRB) of the Defence Science Organisation–SAF (DSO-SAF IRB; reference no.: 0006/2024), as it was retrospective in nature, with only nonidentifiable parameters extracted and all data anonymized prior to recording and analysis. The IRB exemption allowed for secondary analysis without additional consent. All data were deidentified to ensure privacy and confidentiality, and no personally identifiable information or images of participants were included in this manuscript. No compensation was provided to the study participants.

Results

ARI Pathogen Distribution

From September 2023 to August 2024, a total of 1095 servicemen with ARI symptoms were included in the study. The median age of participants was 20 (IQR 2.0, range 17-31) years, with 89.9% of participants aged 18 - 21 years, reflecting a relatively homogenous age distribution. Additionally, 99.8% of participants were male. Of the 1095 nasopharyngeal swabs collected, 608 (55.5%) tested positive for at least one respiratory pathogen. The distribution of pathogens is shown in (Table 1). The most common pathogen was HRV/HEV (n=303, 27.7%). There were 69 (6.3%) coinfections; specifically, HRV/HEV co-occurred with other pathogens in 50 cases (72.4%), influenza co-occurred in 6 cases (8.7%), and both HRV/HEV and influenza co-occurred in 7 cases (10.1%). Coinfections were excluded from further analysis to minimize bias and ensure a clearer attribution of the study outcomes to specific pathogens. Additionally, *Mycoplasma pneumoniae* (n=3) and adenovirus (n=1) were excluded, along with 31 (2.83%) invalid swabs.

XSL•FO
RenderX

Table . Nasopharyngeal swab test results among servicemen with acute respiratory infections in the Singapore Armed Forces from September 2023 to August 2024.

Nasopharyngeal swab test results	Number of swabs (N=1095), n (%)[a]
Positive	608 (55.5)[b]
Coronavirus[c]	68 (3.2)
SARS-CoV-2	35 (5.8)
Human metapneumovirus	37 (3.4)
Human rhinovirus/enterovirus	303 (27.7)
Influenza	43 (3.9)
Parainfluenza	38 (3.5)
Respiratory synctial virus	11 (1.0)
Adenovirus	1 (0.1)
Bacteria[d]	3 (0.3)
Coinfections	69 (6.3)
Negative	456 (41.6)
Invalid	31 (2.8)

[a]The percentages do not add up to 100% due to rounding.

[b]The percentages of all pathogens in the Positive test results category add up to 55.5%.

[c]Coronavirus includes seasonal coronaviruses (229E, HKU1, NL63, OC43) but does not include SARS-CoV-2, which is analyzed and reported as a separate pathogen group.

[d]*Bordetella parapetussis*, *Bordetella pertussis*, *Chlamydia pneumoniae*, *Mycoplasma pneumoniae.*

Comparing the Prevalence of Fever (Severity) Across Respiratory Pathogen Groups

Among the 991 servicemen included for further analysis, fever was observed among 212 (21.4%). The highest proportions of fever were observed in servicemen with influenza (27/43, 62.8%), SARS-CoV-2 (12/35, 34.3%) and parainfluenza (12/38, 31.6%) (Table 2). There were significant differences in the proportions of servicemen with fever across the various pathogens (χ^2_7=65.9, *P*<.001). The odds of patients with influenza having fever were 5.8 times higher than those in patients infected with HRV/HEV (95% CI 2.95 - 11.40; *P*<.001). There were no statistically significant associations between the other pathogen types and fever.

Table . Prevalence of fever among servicemen with specific respiratory pathogen groups from a retrospective study on acute respiratory infections among servicemen from the Singapore Armed Forces from September 2023 to August 2024.

Respiratory pathogen groups	Febrile cases/total infected (N=991), n (%)
Influenza	27/43 (62.8)
SARS-CoV-2	12/35 (34.3)
Parainfluenza	12/38 (31.6)
Human metapneumovirus	11/37 (29.7)
Human rhinovirus/enterovirus	70/303 (23.1)
Coronavirus[a]	13/68 (19.1)
Negative	66/456 (14.5)
Respiratory synctial virus	1/11 (9.1)

[a]Coronavirus includes seasonal coronaviruses (229E, HKU1, NL63, OC43) but does not include SARS-CoV-2, which is analyzed and reported as a separate pathogen group.

Impact of ARI on Organizational Productivity

The median durations of medical certificates, light duty and medical absenteeism were 0 (IQR 0), 2 (IQR 2) and 2 (IQR 0) days, respectively.

- Medical certificate: There were significant differences in the proportions of servicemen with medical certificates ≥1 day across the various pathogens (χ^2_7=29, *P*<.001). The odds of patients with influenza having medical certificates ≥1 day was 5.34 times that of HRV/HEV patients (95% CI 2.63 - 10.88; *P*<.001).

XSL•FO
RenderX

- Light duty: There were no significant differences in the proportion of servicemen with light duty >1 day across the various pathogens (χ^2_7=13.3, *P*=.07). However, patients with influenza had significantly lower odds of having light duty >1 day compared to that of HRV/HEV patients (odds ratio [OR] 0.33, 95% CI 0.17 - 0.64); *P*=.001).
- Medical absenteeism: No significant differences in medical absenteeism were found between HRV/HEV and other pathogens (Table 3).

Table . Results of logistic regression of medical certificate ≥1 day, light duty >1 day and medical absenteeism >1 day on pathogens (n=991).

Outcomes and pathogens	Odds ratio (95% CI)	*P* value
Medical certificate ≥1 day		
Human rhinovirus/enterovirus	Reference	_[a]
Coronavirus	1.57 (0.75 - 3.31)	.23
SARS-CoV-2	1.69 (0.65 - 4.38)	.28
Human metapneumovirus	1.28 (0.47 - 3.51)	.63
Influenza	5.34 (2.63 - 10.88)	*<.001*
Parainfluenza	1.24 (0.45 - 3.40)	.68
Respiratory syncytial virus	1.82 (0.38 - 8.78)	.46
Negative	1.10 (0.69 - 1.74)	.69
Light duty >1 day		
Human rhinovirus/enterovirus	Reference	_[a]
Coronavirus	0.93 (0.54 - 1.61)	.80
SARS-CoV-2	0.60 (0.30 - 1.22)	.16
Human metapneumovirus	0.83 (0.41 - 1.69)	.61
Influenza	0.33 (0.17 - 0.64)	*.001*
Parainfluenza	0.78 (0.39 - 1.56)	.48
Respiratory syncytial virus	0.89 (0.25 - 3.10)	.85
Negative	0.91 (0.67 - 1.24)	.56
Medical absenteeism >1 day		
Human rhinovirus/human enterovirus	Reference	_[a]
Coronavirus	1.31 (0.68 - 2.56)	.41
SARS-CoV-2	0.68 (0.32 - 1.46)	.32
Human metapneumovirus	0.97 (0.44 - 2.15)	.94
Influenza	1.03 (0.48 - 2.19)	.94
Parainfluenza	0.87 (0.40 - 1.88)	.73
Respiratory syncytial virus	1.30 (0.30 - 6.64)	.67
Negative	0.98 (0.70 - 1.38)	.91

[a]Not applicable.

Discussion

Principal Findings

This study provides critical insights into the epidemiology of ARI in a military setting, specifically among SAF servicemen in Singapore. It addresses a key gap in understanding the contribution of different ARI pathogens to medical absenteeism and highlights actionable strategies to mitigate ARI-related disruptions in military contexts. Over the study period, HRV/HEV was the most frequently detected pathogen (27.7%), aligning with previous research findings. Influenza was significantly associated with higher odds of fever, longer medical certificate duration, but the overall duration of medical absenteeism was similar to that of HRV/HEV. Even after adjusting for fever, influenza remained significantly associated with longer medical certificate duration (OR 2.81, 95% CI 1.28 - 6.13, *P*=.01) compared to HRV/HEV, suggesting that other ARI symptoms (eg, sore throat, rhinorrhea), contribute to its more severe clinical presentation. This distinction is important as it provides evidence that influenza leads to more severe illness overall, beyond fever, compared to other ARI pathogens and makes servicemen more likely to receive medical certificates than light duties. Conversely, servicemen infected

XSL•FO
RenderX

with other ARI pathogens—perceived as less severe and not significantly associated with fever—are more likely assigned to light duties instead. This reflects clinical practice, where servicemen with more severe ARI are allowed to recuperate at home for a longer duration, while those with milder ARI are required to return to work earlier but allowed to perform only light duties. These findings reinforce the importance of continuing robust influenza vaccination programmes in the SAF.

Despite these findings, our study demonstrated no significant differences in duration of medical absenteeism across pathogens, reflecting that ARIs, regardless of pathogen type, contribute substantially to absenteeism and operational disruptions. This highlights the need for pathogen-agnostic strategies targeting ARI prevention and management, rather than focusing solely on influenza in the SAF. The importance of nonpharmaceutical intervention (NPI) is further underscored by historical examples, such as the 1996 influenza A (H3N2) outbreak aboard a US Navy ship [18]. Despite 95% of the crew being vaccinated against influenza, the antigenic distinctness of the circulating strain led to an infection rate of 42%, demonstrating the limitations of vaccination alone and the critical need for complementary measures such as isolation protocols, improved ventilation, and rapid response strategies to mitigate influenza outbreaks in confined military settings [18].

Globally, ARIs beyond influenza have consistently disrupted organizational productivity in military settings, emphasising the universal challenges posed by communal living and intense operational demands. Among troops of the Polish Military Contingent deployed overseas from 2003 to 2005, ARIs were the leading cause of outpatient morbidity, with incidence rates ranging from 45.6-54.8 cases per 100 soldiers, significantly disrupting training schedules and productivity [19]. Similarly, respiratory illnesses affected 69.1% of US military personnel deployed to Iraq and Afghanistan from 2003 to 2004, which often resulted in missed patrols and transient decreases in operational efficiency [20]. Notably, the incidence of ARI doubled during the combat phase, reflecting the heightened susceptibility of military personnel under physically and mentally demanding conditions [20]. Additionally, a systematic review of 46 military studies in 17 countries observed a higher incidence of COVID-19 in military personnel on deployment and those who had overseas exposure, amplifying the risks faced by military personnel where operational stressors and novel environmental exposures exacerbate the burden of ARIs, including COVID-19 [21]. These findings align with our study results, which showed no significant differences in medical absenteeism across pathogens, underscoring the need to adopt prevention strategies targeting the broader spectrum of ARIs rather than focusing primarily on influenza.

A multipronged approach to ARI prevention, integrating vaccination programmes, targeted NPIs and the strategic use of therapeutics, is essential for military personnel. Vaccination has proven highly effective; for instance, the H1N1-2009 influenza vaccination programme in Singapore reduced the weekly cases by 54% among military personnel post vaccination [22]. However, vaccines alone are insufficient, as demonstrated by our study findings. The US military has successfully implemented cohorting strategies in basic training facilities, dividing personnel into smaller, stable groups to limit the exposure and spread of ARI [1]. Complementing these NPIs, personal protective measures such as mandatory mask-wearing during outbreaks or in enclosed, high-density environments, coupled with educational campaigns on respiratory hygiene, could further reduce the ARI transmission [1]. Similarly, Poland's military experienced peaks in ARI during troop rotations, highlighting the need for stricter isolation protocols and improved acclimatization strategies for personnel entering new environments [19]. Insights from the systematic review on COVID-19 in military settings also emphasised the utility of digital tools, such as contact tracing applications, in effectively managing outbreaks [21]. Therapeutics are another critical component of ARI management. A study in Singapore during the H1N1 influenza outbreak in 2009 demonstrated that early use of oseltamivir significantly reduced the risk of hospitalization and severe outcomes, even in young and healthy military populations [23]. Enhanced vaccination coverage for pathogens with available vaccines (ie, influenza, SARS-CoV-2), rapid isolation of infected personnel, and timely therapeutic interventions collectively contribute to reducing the overall ARI burden. By focusing on overall respiratory health rather than targeting specific pathogens, these strategies will help create a more resilient military workforce.

The findings of our study, while focused on a single SAF military camp, have broader implications for the military and other high-density living settings globally. The consistent alignment of our findings with other military studies, such as those conducted in US, Polish, and Singaporean military contexts, highlights the relevance of our findings across varied geographic and operational contexts. Our study findings are generalizable to other SAF camps or military units with similar profiles, where young men live and work together in close quarters, as these shared conditions contribute significantly to ARI transmission. Additionally, these findings may offer insights for surveillance and ARI prevention measures in other high-density living environments, such as dormitories, where communal living and close interactions similarly increase the risk of ARI outbreaks. Expanding surveillance efforts to multiple camps with different demographics and training intensities, as well as integrating findings with civilian studies, could further enhance the applicability of these findings to both military and nonmilitary populations.

Strengths

The study used data from comprehensive surveillance with the BioFire FilmArray Respiratory Panel, which enabled the detection of a wide range of respiratory pathogens. The large sample size allowed for accurate determination of pathogen prevalence in the military context. Additionally, the study's focus on medical absenteeism bridges the gap between clinical outcomes and organizational impact, making the results actionable for health policy and military planning.

Limitations

First, excluding coinfections from the analysis may have overlooked the complexity of pathogen interactions. Coinfections often exacerbate illness severity or prolong recovery time, and their exclusion may have led to an

XSL•FO
RenderX

underestimation of the duration of MA. However, the number of coinfections in our study was relatively small, and their exclusion is unlikely to have significantly impacted the overall results or conclusions. Second, the small number of bacterial ARI cases limited our understanding of their impact on organizational productivity. Bacterial pathogens, such as *Mycoplasma pneumoniae*, are known to cause severe symptoms and potentially longer MA. Without sufficient data, their role in contributing to productivity losses remains unclear. Third, the reliance on self-reported data for fever and medical certificates or light duty duration introduces the potential for recall bias. Lastly, while our study is generalizable to SAF units or camps with similar profiles (ie, young men in close quarters), its findings may not apply to camps with different demographics or training intensities.

Future studies should address these limitations by including coinfection cases to better understand the interplay between different pathogens and their combined impact on ARI severity and duration of medical absenteeism. Expanding surveillance to multiple military camps and similar high-density living settings would enhance generalizability of findings and account for variations in living conditions, vocations, and operational structures. This broader approach could help identify high-risk groups or settings and guide the implementation of targeted interventions. Additionally, longitudinal studies tracking the same cohort over multiple ARI seasons could provide insights into the effectiveness of preventive measures, such as influenza vaccinations, NPIs, and the impact of digital health tools such as the upcoming Data Analytics for Soldier Health (DASH 2.0) system in the SAF. The integration of iDASH2.0 into future electronic medical records could enable real-time monitoring of ARI trends, identification of seasonal trends in pathogen prevalence, and assessing waning immunity post vaccination. Finally, studies focusing on bacterial ARI cases should explore their clinical and operational impact using advanced diagnostic tools and longitudinal study designs, paving the way for antibiotic stewardship programmes.

Conclusion

This study highlights the significant burden of ARIs in military settings, with influenza contributing to longer medical leave compared to noninfluenza pathogens, yet no differences in the overall duration of medical absenteeism across pathogens. These findings underscore the need for a multipronged approach to ARI prevention that goes beyond vaccination programmes to include targeted nonpharmaceutical interventions and therapeutics. By addressing the identified gaps and building on the strengths of this study, future research should explore the impact of coinfections and bacterial ARIs, seasonal trends, and the long-term effectiveness of preventive measures across multiple military camps in Singapore. This will enhance generalizability and inform robust, evidence-based policies to optimize the health and productivity of servicemen. Ultimately, integrating vaccination programmes, NPIs, and therapeutics into a cohesive ARI management framework will strengthen the SAF's resilience against respiratory threats and ensure sustained organizational productivity.

Acknowledgments

No generative AI tools were used in the drafting, analysis, or writing of this manuscript.

Data Availability

The dataset analyzed during this study is not publicly available due to the sensitive nature of military-related health data but are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: HYL, PM
Data curation: PM, JQN
Formal analysis: PM
Investigation: JQN, AM
Methodology: HYL, PM, JKG
Project administration: AM, JQN
Resources: JKG, AM, JQN
Supervision: HYL, JKG
Writing – original draft: PM
Writing – review & editing: PM, JKG, JQN, AM, HYL

Conflicts of Interest

None declared.

References

1. Sanchez JL, Cooper MJ, Myers CA, et al. Respiratory infections in the U.S. military: recent experience and control. Clin Microbiol Rev 2015 Jul;28(3):743-800. [doi: 10.1128/CMR.00039-14] [Medline: 26085551]

XSL•FO
RenderX

2. Ho ZJM, Hwang YFJ, Lee JMV. Emerging and re-emerging infectious diseases: challenges and opportunities for militaries. Mil Med Res 2014;1(1):21. [doi: 10.1186/2054-9369-1-21] [Medline: 25722877]
3. Burbank AJ. Risk factors for respiratory viral infections: a spotlight on climate change and air pollution. J Asthma Allergy 2023;16:183-194. [doi: 10.2147/JAA.S364845] [Medline: 36721739]
4. Jacobs SE, Lamson DM, St George K, Walsh TJ. Human rhinoviruses. Clin Microbiol Rev 2013 Jan;26(1):135-162. [doi: 10.1128/CMR.00077-12] [Medline: 23297263]
5. Herold S, Becker C, Ridge KM, Budinger GRS. Influenza virus-induced lung injury: pathogenesis and implications for treatment. Eur Respir J 2015 May;45(5):1463-1478. [doi: 10.1183/09031936.00186214] [Medline: 25792631]
6. Zimmerman RK, Balasubramani GK, D'Agostino HEA, et al. Population-based hospitalization burden estimates for respiratory viruses, 2015-2019. Influenza Other Respir Viruses 2022 Nov;16(6):1133-1140. [doi: 10.1111/irv.13040] [Medline: 35996836]
7. Lalbiaktluangi C, Yadav MK, Singh PK, et al. A cooperativity between virus and bacteria during respiratory infections. Front Microbiol 2023;14:1279159. [doi: 10.3389/fmicb.2023.1279159] [Medline: 38098657]
8. Helou M, Mahdi A, Daoud Z, et al. Epidemiology of community-acquired respiratory tract infections in patients admitted at the emergency departments. Trop Med Infect Dis 2022 Sep 8;7(9):233. [doi: 10.3390/tropicalmed7090233] [Medline: 36136644]
9. Lau YF, Koh WHV, Kan C, et al. Epidemiologic analysis of respiratory viral infections among Singapore military servicemen in 2016. BMC Infect Dis 2018 Mar 12;18(1):123. [doi: 10.1186/s12879-018-3040-x] [Medline: 29529993]
10. Kiliç S, Gray GC. Nonpharmaceutical interventions for military populations during pandemic influenza. Turk Silahli Kuvvetleri Koruyucu Hekim Bul 2007;6(4):285-290 [FREE Full text] [Medline: 18516249]
11. Jeger V, Dünki A, Germann M, et al. H1N1 outbreak in a Swiss military boot camp--observations and suggestions. Swiss Med Wkly 2011;141:w13307. [doi: 10.4414/smw.2011.13307] [Medline: 22144395]
12. Centers for Disease Control and Prevention (CDC). Outbreak of 2009 pandemic influenza A (H1N1) on a Peruvian Navy ship - June-July 2009. MMWR Morb Mortal Wkly Rep 2010 Feb 19;59(6):162-165 [FREE Full text] [Medline: 20168295]
13. COVID-19 vaccination. Ministry of Health-Singapore. 2024 May 29. URL: https://www.moh.gov.sg/covid-19/vaccination [accessed 2024-08-05]
14. Puro V, Fusco FM, Lanini S, Nisii C, Ippolito G. Risk management of febrile respiratory illness in emergency departments. New Microbiol 2008 Apr;31(2):165-173 [FREE Full text] [Medline: 18623980]
15. Paules C, Subbarao K. Influenza. Lancet 2017 Aug 12;390(10095):697-708. [doi: 10.1016/S0140-6736(17)30129-0] [Medline: 28302313]
16. Guan WJ, Ni ZY, Hu Y, et al. Clinical characteristics of coronavirus disease 2019 in China. N Engl J Med 2020 Apr 30;382(18):1708-1720. [doi: 10.1056/NEJMoa2002032] [Medline: 32109013]
17. Velasco JM, Valderama MT, Diones PC, et al. Outbreak of influenza and SARS-CoV-2 at the Armed Forces of the Philippines Health Service Education and Training Center, September 25-October 10, 2023. MSMR 2024 May 20;31(5):9-15 [FREE Full text] [Medline: 38847656]
18. Earhart KC, Beadle C, Miller LK, et al. Outbreak of influenza in highly vaccinated crew of US Navy ship. Emerg Infect Dis 2001;7(3):463-465. [doi: 10.3201/eid0703.010320] [Medline: 11384530]
19. Korzeniewski K, Nitsch-Osuch A, Konior M, Lass A. Respiratory tract infections in the military environment. Respir Physiol Neurobiol 2015 Apr;209:76-80. [doi: 10.1016/j.resp.2014.09.016] [Medline: 25278277]
20. Sanders JW, Putnam SD, Frankart C, et al. Impact of illness and non-combat injury during Operations Iraqi Freedom and Enduring Freedom (Afghanistan). Am J Trop Med Hyg 2005 Oct;73(4):713-719 [FREE Full text] [Medline: 16222015]
21. Gwee SXW, Chua PEY, Pang J. SARS-CoV-2 transmission in the military during the early phase of the pandemic-a systematic analysis. Int J Environ Res Public Health 2022 Jun 16;19(12):7418. [doi: 10.3390/ijerph19127418] [Medline: 35742662]
22. Lee VJ, Tan CH, Yap J, et al. Effectiveness of pandemic H1N1-2009 vaccination in reducing laboratory confirmed influenza infections among military recruits in tropical Singapore. PLoS ONE 2011;6(10):e26572. [doi: 10.1371/journal.pone.0026572] [Medline: 22053196]
23. Lee VJ, Yap J, Cook AR, et al. Oseltamivir ring prophylaxis for containment of 2009 H1N1 influenza outbreaks. N Engl J Med 2010 Jun 10;362(23):2166-2174. [doi: 10.1056/NEJMoa0908482] [Medline: 20558367]

Abbreviations

ARI: acute respiratory infection
FRI: febrile respiratory illness
HEV: human enterovirus
HRV: human rhinovirus
HRV/HEV: human rhinovirus/human enterovirus
IRB: institutional review board
NPI: nonpharmaceutical intervention

XSL•FO
RenderX

OR: odds ratio
SAF: Singapore Armed Forces

Edited by A Mavragani; submitted 22.11.24; peer-reviewed by C Dalton, E Lau, TK Er; revised version received 04.02.25; accepted 05.02.25; published 18.04.25.

Please cite as:
M P, Goh JK, Ng JQ, Mutalib A, Lim HY
Impact of Acute Respiratory Infections on Medical Absenteeism Among Military Personnel: Retrospective Cohort Study
JMIR Form Res 2025;9:e69113
URL: https://formative.jmir.org/2025/1/e69113
doi:10.2196/69113

© Premikha M, Jit Khong Goh, Jing Qiang Ng, Adeliza Mutalib, Huai Yang Lim. Originally published in JMIR Formative Research (https://formative.jmir.org), 18.4.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Monthly Cycling of Food Insecurity in Latinas at Risk for Diabetes: Methods, Retention, and Sample Characteristics for a Microlongitudinal Design

Angela Bermúdez-Millán[1*], MPH, PhD; Rafael Pérez-Escamilla[2*], PhD; Sofia Segura-Pérez[3*], MS, RD; James Grady[1*], DrPH; Richard S Feinn VI[4*], PhD; Hanako Agresta[1*], BSc; Dean Kim[1*], BA; Julie Ann Wagner[5*], PhD

[1]Department of Public Health Sciences, School of Medicine, UConn Health, 263 Farmington Avenue, Farmington, CT, United States

[2]Department of Social and Behavioral Sciences, Yale School of Public Health, New Haven, CT, United States

[3]Hispanic Health Council, Hartford, CT, United States

[4]Frank H. Netter School of Medicine, Quinnipiac University, Hamden, United States

[5]Division of Behavioral Sciences and Community Health, School of Dental Medicine, UConn Health, Farmington, CT, United States

[*]all authors contributed equally

Corresponding Author:
Angela Bermúdez-Millán, MPH, PhD
Department of Public Health Sciences, School of Medicine, UConn Health, 263 Farmington Avenue, Farmington, CT, United States

Abstract

Background: Food insecurity (FI) is a risk factor for type 2 diabetes (T2D) that disproportionately affects Latinas. We conducted a microlongitudinal study to examine the relationship of monthly cycling of FI and diabetes risk factors.

Objective: This study aimed to determine the quantitative methodology, recruitment and retention strategies, predictors of retention across time, and baseline sample demographics.

Methods: Participants were adult Latinas living in Hartford, Connecticut who were recruited through a community agency, invited to participate if they were receiving Supplementary Nutrition Assistance Program (SNAP) benefits, screened positive for FI using the 2-item Hunger Vital Sign Screener, and had elevated risk factors for T2D using the American Diabetes Association risk factor test. Using a microlongitudinal design, we collected data twice per month for 3 months (week 2, which is a period of food budget adequacy; and week 4, which is a period of food budget inadequacy) to determine if the monthly cycling of FI was associated with near-term diabetes risk (fasting glucose, fructosamine, and glycosylated albumin) and long-term risk (BMI, waist circumference, and glycated hemoglobin) markers. We determined whether household food inventory, psychological distress, and binge eating mediated associations. We examined Health Action Process Approach model constructs. To assess the relationship between monthly cycling of FI with diabetes risk markers, we used repeated measures general linear mixed models. To assess the role of mediators, we performed a causal pathway analysis.

Results: Participant enrollment was from April 1, 2021 to February 21, 2023. A total of 87 participants completed 420 assessments or a mean of 4.83 (SD 2.02) assessments. About half (47/87, 54%) of the sample self-identified as Puerto Rican, mean age was 35.1 (SD 5.8) years, with 17.1 (SD 11.6) years in the mainland United States. Just under half (41/87, 47.1%) spoke Spanish only, 69% (60/87) had no formal schooling, and 31% (27/87) had less than eighth grade education. Modal household size was 4 including 2 children; 44.8% (39/87) were not living with a partner. About half (47/87, 54%) were unemployed, 63.2% (55/87) reported a monthly income <US $1000, and 63.2% (55/87) used food pantries. In total, 61 participants (70.1%) completed all 6 assessments. On Pearson correlation analysis, having internet at home and having a tablet at home were associated with a higher number of completed assessments.

Conclusions: This study demonstrated how FI cycles over the month and whether and to what degree the cycling itself is related to the risk for T2D development, as well as the evidence for some putative mechanisms of this association that can serve as future intervention targets including SNAP disbursement schedules.

(JMIR Form Res 2025;9:e66970) doi:10.2196/66970

KEYWORDS

food insecurity; monthly cycling; type 2 diabetes risk; quantitative methods; Latinas; endocrinology; nutrition; nutrition assistance; micro-longitudinal design

XSL•FO
RenderX

Introduction

Latinas have nearly double the rates of type 2 diabetes (T2D; 11.3%) compared with non-Hispanic white women (6.1%) [1]. This health inequity is not fully understood. Prediabetes, a condition in which a person's blood glucose levels are higher than normal but do not reach the cutoff for diabetes, is a significant risk factor for developing T2D. In the United States, more than one-third of Latinas meet at least 1 American Diabetes Association criterion of prediabetes, and those younger than 45 years merit immediate attention for the prevention of advancing to full diabetes [2].

A risk factor for T2D that disproportionately affects Latinas is food insecurity (FI). A novel risk factor is the monthly cycling of food insecurity, an issue that has been notably understudied. FI is the limited or uncertain availability of nutritionally adequate and safe foods or the ability to acquire acceptable foods in socially acceptable ways [3,4]. In 2023, 13.5% (18.0 million) of US households were food insecure at least some time during the year [5]. Rates of FI were higher than the national average for households with children and particularly households with children headed by single adults, Hispanic-headed households, and households in principal cities.

FI has been associated with overweight or obesity, cardiometabolic disease, and T2D [6-8]. FI may influence diabetes risk in several ways. First, it may affect diet by promoting dependence on inexpensive, highly palatable foods that are energy dense, which are associated with weight gain and promote the development of chronic conditions [9]. Second, women in food insecure households may have elevated mental distress as a result of their central role in food acquisition and preparation for their families. The rates of depression and anxiety are higher among food insecure mothers, and among Latinos living in poverty [10]. Third, FI may affect emotional eating [11,12], the tendency to eat or overeat in response to negative emotions. Emotional eating has repeatedly been linked to overweight and obesity due to cycles of restriction and binge eating [12], resulting in fluctuations of meal size and frequency that contribute to weight gain.

A useful theoretical framework for describing the sociocognitive processes involved in health behaviors, such as fruit and vegetable intake, is the Health Action Process Approach (HAPA) model [13]. HAPA constructs include risk awareness, outcome expectancy, self-efficacy, intention, action planning, and coping planning. Understanding these constructs among Latinas in relation to FI may inform behavioral approaches for this population.

Most studies of FI have used cross-sectional or multiwave longitudinal designs, referencing the past 30 days, 3, 6, or 12 months, and have not captured the weekly ebb and flow of access to food, which may be crucial for understanding the role of FI as a diabetes risk factor, and for designing interventions. Yet, it has been hypothesized that among households that receive Supplemental Nutrition Assistance Program (SNAP) benefits, FI is not stable but cycles over the course of the month [14]. SNAP benefits are distributed in the beginning of the month and may not be sufficient to attenuate or prevent FI by the end of the month [15]. A small number of studies have shown that SNAP benefits lasting less than a month is a predictor of poor dietary quality and health outcomes [16] but this has not been studied in our population of interest. Our hypothesis is that FI cycling over the course of a month is detrimental to women's health and is a risk factor for T2D.

Descriptive and formative research is crucial to address this potentially serious public health problem. Therefore, this innovative study aims to document FI occurrence over the month and test whether FI is associated with increases in diabetes risk markers in Latinas. Using a microlongitudinal design to capture the disbursement cycle of SNAP benefits, we collected data twice per month, the second week of the month which we expected to be a period of food budget adequacy and the fourth week of the month, which we expected to be a period of food budget inadequacy. We repeated this schedule for 3 months.

The first aim of the study was to determine if the monthly cycling of FI is associated with diabetes risk biomarkers. We hypothesized that compared with week 2 (food secure: period of food budget adequacy), week 4 (food insecure: period of food budget inadequacy) will be characterized by higher near-term diabetes risk biomarkers, that is, fasting glucose, fructosamine, and glycosylated albumin. We also hypothesized that compared with participants with little change in FI, participants with greater weekly changes in FI will have higher long-term metabolic risk markers for T2D (BMI, waist circumference, and glycated hemoglobin, HbA1c).

The second aim of the study was to describe the monthly cycling of household food inventory, eating behaviors, HAPA constructs, and distress. We hypothesized that compared with week 2 (food secure: period of food budget adequacy), week 4 (food insecure: period of food budget inadequacy) will be characterized by more eating disturbance (emotional eating and binge-eating), poorer diet quality (fat, sugar, fruit, and vegetables), and increased psychological distress (depression and anxiety symptoms; Figure 1). Changes in HAPA constructs over time are described. To our knowledge, this is the first microlongitudinal study to investigate FI cycling over time.

Longitudinal designs with multiple assessments can pose research burden on participants and are difficult to conduct in low-resourced populations. The aims of this paper are to report our quantitative methodology, to describe our recruitment and retention strategies, to examine predictors of retention across the 6 timepoints, and to report baseline demogrraphics for the sample.

XSL•FO
RenderX

Figure 1. Conceptual framework for the Food Insecurity Cycling study (Hartford, Connecticut, enrollment 2021 - 2023).

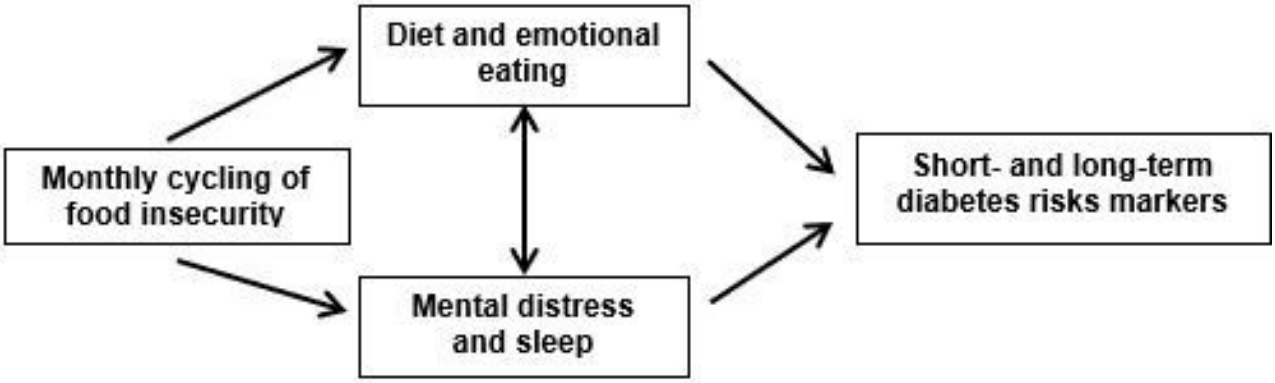

Methods

Design

This study used a microlongitudinal design with 6 observations across 3 months.

Participants

Eligibility criteria are shown in Textbox 1.

Textbox 1.

Inclusion criteria

- Age 18 - 44 years.
- Self-identify as Latina.
- Self-identify as a woman.
- Hartford area resident.
- Currently receiving Supplementary Nutrition Assistance Program (SNAP) benefits.
- Read and write in Spanish or English.
- Access to a smart phone to complete assessments.
- Screen positive for food insecurity per validated 2-item food insecurity screen [17,18] and at high risk for diabetes per standard criteria [19,20] (a score ≥4 indicates elevated risk for Hispanics). We studied adult Latinas of childbearing years because data suggest that they may prioritize nutrition for their children over their own nutrition [21].

Exclusion criteria

- Currently or recently pregnant or planning to become pregnant in the next 3 months (because of potential changes to nutrition benefits with pregnancy).
- Self-reported diabetes, kidney, or liver disease (because of nutritional requirements for management of those conditions).
- Positive screen for current substance abuse or dependence (to decrease attrition).
- Self-reported vasovagal syncope while undergoing phlebotomy (ie, history of feeling light-headed, dizziness, hot, sweaty, or fainting, to reduce medical urgencies associated with home phlebotomy).
- Living at a shelter (because women in a shelter do not have their own reliable food storage space).
- Those with previously undiagnosed diabetes who were identified at baseline as having unequivocal extant diabetes were excluded due to differences in cardiometabolic function and referred for treatment.

Setting

As a US state, Connecticut is predominantly white, but its capital city of Hartford is largely Latino. Consistent with national data, Latinos in Connecticut have a significantly higher rate of diabetes than their white counterparts [22]. The study was led and coordinated by UConn Health and implemented in close partnership with the Hispanic Health Council (HHC) that used the study's 2 bilingual and bicultural community health workers (CHWs). HHC is a community-based research, service, training, and advocacy organization headquartered in Hartford. This institutional partnership was built from a trusting relationship built over years of successful collaboration between UConn Health, Yale School of Public Health and HHC co-designing and implementing research studies using CHWs [23].

Recruitment

A comprehensive staff training program was developed and delivered for HHC recruiters and data collectors, including a core manual to support fidelity to the assessment procedures

XSL•FO
RenderX

and study protocol. CITI-trained, bilingual and bicultural recruiters distributed a recruitment flyer across HHC programs serving the study's target population. The recruitment flyer briefly described the study and informed intersted individuals to call the recruiters. Programs included the Santa Marquez Special Supplemental Nutrition Program for Women, Infant, and Children Program (WIC) Center, which serves to safeguard the health of low-income women, infants, and children up to age 5 who are at nutritional risk. The recruitment flyer was also posted in the HHC social media platforms (ie, Hispanic Health Council Facebook, Instagram, and Twitter pages). HHC staff also distributed flyers at health fairs, Hartford mobile markets, community health agencies, local convenience stores, and local daycare centers and it was mentioned on Spanish-language radio. Recruitment also took place at local food pantries that serve the target community. Foodshare is the largest antihunger organization in the Greater Hartford area, with a network of food pantries, meal programs, and mobile foodshare sites.

Screening and Enrollment

Eligibility screening was conducted by the CHW recruiters. Due to local rates of COVID-19 infection during the recruitment, most screening took place by phone, and some in a private room at the HHC. These sessions were scheduled through a virtual scheduling calendar accessible only to team personnel. The recruiters asked questions ascertaining the inclusion and exclusion criteria.

Screening Measures

Demographic and clinical inclusion and exclusion criteria were self-reported. The 2-item Hunger Vital Sign FI screener was used to detect likely FI [17]. The 2 questions were "In the past 3 months, have you or other adults in your household worried whether your food would run out before you got money to buy more?" and "In the past 3 months, has the food you or other adults in your household bought just didn't last and you didn't have money to get more?" Response options were "yes," or "no." A response of "yes" to either question met study inclusion criteria for food insecurity [17].

The American Diabetes Association risk test was used to determine elevated risk for developing T2D. It includes seven self-report questions (total score of 0 - 11) on age, gender, gestational diabetes mellitus, family history of diabetes, high blood pressure, physical activity, and obesity (based on BMI through a weight-height chart). The American Diabetes Association uses a cutoff of 5 or higher to indicate increased for diabetes. In this study, since all participants endorsed FI, which is itself a novel risk factor for diabetes, a score of ≥4 was required for inclusion [19,20].

The 4-item CAGE-AID (CAGE Adapted to Include Drugs) screener [24] was used to detect likely alcohol and substance abuse. A score of 2 is recommended as a positive screen for alcohol or substance abuse and indicates a potential problem that warrants further assessment by a health care professional; a score of 2 or higher excluded a recruit from participation.

Assessment Procedures and Incentives

After obtaining consent to participate and HIPAA (Health Insurance Portability and Accountability Act) authorization, physical assessments were conducted in the participant's preferred language (English or Spanish), at the HHC. Participants had their weight, height, waist, and hip circumferences measured and were asked to provide fasting blood samples. Assessments were performed early morning (ie, before 9 am) to obtain fasting blood samples (30 ml). Study phlebotomist from HHC took blood samples from an antecubital vein, using sterile technique. Samples were then promptly transported by trained personnel, compliant with HIPAA and biohazard regulations, in an ice chest to UConn Health Clinical Research Center for processing and analysis. There, blood samples were aliquoted. One aliquot went to UConn Health Hospital Clinical Laboratory for immediate assessment of glucose. The remaining aliquots went to the UConn Health Clinical Research Center where they were centrifuged and stored at −80°C until analyzed for all other assays described below.

Following the baseline assessment, participants provided data at 2 time points per month for 3 months, in order to capture the disbursement cycle of SNAP benefits. Participants received cash compensation for participation, US $15 at each timepoint for each blood draw and US $10 for surveys, totaling US $150.

Retention Efforts

Because the temporal cycling of food insecurity over time is central to our research question, retention of this difficult-to-reach population was of utmost importance. Several community-engaged efforts were made to retain participants including the following: (1) Socializing the participant to the research endeavor for the community and describing the importance of each data collection timepoint; (2) Maintaining updated contact information for participants and a designated secondary contact person; (3) Speaking in the participant's preferred language and fostering good rapport between CHWs and participants; (4) responding fully, genuinely, and in a timely manner to any participant concerns; (5) phone and text reminders for study appointments; and, (6) offering financial compensation for time spent on study activities.

Measures

Next, participants were scheduled for a baseline assessment. Due to COVID-19 social distancing mandates, the majority of the survey instruments were collected over the phone by trained bilingual interviewers who entered participant verbal responses into REDCap (Research Electronic Data Capture) [25] on large-screen electronic tablets. For all self-report surveys, response options referenced the preceding 7-days. Refer to Table 1 for data collection timeline.

XSL•FO
RenderX

Table . Measures and assessment schedule.

Predictor	Month 1[a]		Month 2[a]		Month 3[a]	
	Week 1	Week 4	Week 1	Week 4	Week 1	Week 4
Food security[b] (past 7 days timeframe)	✓	✓	✓	✓	✓	✓
Covariates						
Demographics	✓	N/A[c]	N/A	N/A	N/A	N/A
SNAP[d] benefits	✓	✓	✓	✓	✓	✓
Mediators (past 7 days timeframe)						
Nutrition[e]	✓	✓	✓	✓	✓	✓
Distress (symptoms of anxiety and depression)	✓	✓	✓	✓	✓	✓
Sleep quality	✓	✓	✓	✓	✓	✓
Outcomes (short-term)						
Fasting blood glucose and insulin	✓	✓	✓	✓	✓	✓
Fructosamine	✓	✓	✓	✓	✓	✓
Glycosylated albumin	✓	✓	✓	✓	✓	✓
Outcomes (long-term)						
BMI	✓	N/A	N/A	N/A	N/A	✓
Waist and hip circumference	✓	N/A	N/A	N/A	N/A	✓
HbA1c	✓	N/A	N/A	N/A	N/A	✓

[a]SNAP benefits are made available on the electronic benefits transfer card from the 1st to the 3rd of every month, based on the first letter of the client's last name. Data for week 1 were collected the week after food purchasing, during food abundancy; Week 4 data were collected before the next scheduled SNAP benefit disbursement, during food scarcity.

[b]15-item US Household Food Security Survey Module.

[c]N/A: not applicable.

[d]SNAP: Supplementary Nutrition Assistance Program.

[e]Nutrition measures included food frequency questionnaire, household food inventory, HAPA model constructs, emotional eating, and binge eating.

Participant Characteristics at Baseline

Demographics included age (18 - 24 years), language (English only, English and Spanish, or Spanish only), education (no formal schooling, eighth grade or less, some high school, high school graduate or GED [General Educational Development high school equivalency examination], some college, associates degree, or finished 4 years of college), employment (working full time or part time, unemployed, student, or homemaker), relationship status (living with a spouse or partner or not living with a spouse or partner), household composition (mother, father, mother-in-law, father-in-law, boyfriend/ or spouse, own children, sisters/, brothers, or others), and years living in the United States (free response). Financial strain was measured on a 5-point Likert scale [23] from 1= "We have enough and we can save" to 5 = "We don't have enough and we have great difficulties." Digital connectivity was measured with questions regarding whether the participant had internet and access to a computer and a tablet in the home, with response options yes or no.

Food Insecurity

Household food insecurity was assessed using a 15-item adapted and validated version of the original 18-item US Household Food Security Survey Module for pregnant Latinas [26,27]. The module captures the experience of anxiety regarding household food supplies, inadequate food quality and reduced food intake. This measure is validated in English and Spanish.

Nutrition Assistance

Participants reported how much in food stamps (SNAP) benefits they received per month, whether benefits were received once

XSL•FO
RenderX

or twice per month, how many weeks the benefits last, whether and how much others in the household receive SNAP, and whether and how much the participant received from WIC.

Food Frequency Questionnaire

Diet quality was measured with the Spanish and English versions of the Block Fat, Sugar, Fruitt, Vegetable Screener, a 55-item food frequency questionnaire shown to have good reliability and validity [28,29]. Participants were instructed how to access and enter data through a link sent to them through email. Visual aids were uploaded in REDCap as images to help participants with portion size estimations [30].

Emotional Eating

The Emotional Eating subscale of the Three Factor Eating Questionnaire R18-V2 [31,32] assesses emotional eating. This subscale was validated in English and Spanish among Latinos of various Latino heritages [33]. The Spanish-language version [34] of the Questionnaire on Eating and Weight Patterns - 5 is a screening instrument for binge eating. It assesses eating - disorder–behaviors and attitudes that align with DSM - 5 diagnostic criteria [35].

Psychological Distress

The validated Spanish version 8-item scale of the Patient-Reported Outcomes Measurement Information System emotional anxiety scale (Short Form 8a) assesses anxiety symptoms [36,37]. The validated Spanish version of the Personal Health Questionnaire Depression Scale (PHQ-8, ie, the PHQ-9 without the suicidality item) [38,39] assesses depressive symptoms. The suicidality item was omitted to avoid CHWs being required to respond to mental health urgencies in the field.

Sleep Quality

Sleep quality was measured with the Pittsburgh Sleep Quality Index (PSQI) [40]. The PSQI consists of 19 items that produce a global sleep quality score as well as individual items such as usual bedtime, wake-up time, actual hours slept, number of minutes to fall asleep, and nighttime awakenings. Global PSQI score was estimated for each participant as described previously, and as analyzed by our [41] team.

Health Action Process Approach Constructs

Questions were structured to assess motivational and volitional phases of the HAPA model as they pertain to eating behaviors. Specific items were adapted from previous studies of Latinas of childbearing age [42,43]. Four items tapped into risk awareness, for example, “If you continue with your current eating pattern, how likely is it that you will develop diabetes?” Response options were very unlikely, unlikely, likely, and very likely. Five items tapped into outcome expectancy, for example, “How true is it if you are able to improve and maintain healthier eating patterns, you will prevent diabetes?” Response options were not at all true, somewhat true, mostly true, and very true. Four items tapped into self-efficacy, for example, “I am certain or confident that I am able to improve my eating patterns, even if I have to make a detailed plan to have appropriate food available (on a food budget).” Response options were not at all true, somewhat true, mostly true, and very true. Three items tapped into intentions, for example, “During the next month, do you intend to increase the amount of vegetables you eat?” Response options were very unlikely, unlikely, likely, and very likely. Four items tapped into action planning, for example, “Do you have a plan of when you will go about eating more fruits and vegetables?” Response options were yes or no. Three items tapped into coping planning, for example, “Do you have a plan of how to incorporate more fruits and vegetables into your daily routine?” Response options were not at all, somewhat, or definitely.

Household Food Inventory

Participants completed a revised version of a home food inventory that has been validated in low-income Hispanic or Latinx households [44-46]. Participants completed the inventory themselves through REDCap [25]. Response options for each food were yes or no. Participants were instructed how to access and enter data through a link sent to them through email.

Anthropometrics

Weight was measured using a Seca Digital Scale and height was measured through a Seca Portable Stadiometer. Height was measured to the nearest 1 cm with the participant’s head positioned in the Frankfurt horizontal plane. Weight was measured to the nearest 0.1 kg. The average of 2 trials was used for both height and weight, unless measures were separated by a minimum of 1 cm or 0.5 kg, in which case, 2 more measures are taken. BMI (kg/m^2) was calculated from height and weight. Waist circumference was measured twice at the umbilicus and hip circumference at the buttocks (nearest 0.5 cm).

Glucometrics

Fasting glucose reflects the body’s ability to regulate glucose in the absence of any glycemic challenge (ie, food). It is a strong predictor of incident T2D. Fasting glucose was analyzed at the UConn Health clinical laboratory. Fasting insulin was analyzed at the UConn Health Clinical Research Center with a kit from Siemens Healthcare Diagnostics (Los Angeles, California). It has a reference range median of 8.9 uIU/mL, sensitivity of 2 uIU/mL, and inter- and intrarun coefficients of variation (CVs) of 2.5 and 3.6, respectively. Fructosamine reflects the physiology of glucose in the extracellular space and has a short half-life, reflecting glycemia over 2 - 4 weeks, ideal to address glycemia over intervals of several weeks. Fructosamine measures the fraction of total serum proteins that have undergone glycation. Assays were run at the UConn Health Clinical Research Center using a kit from ARUP laboratories (Salt Lake City, Utah) with an assay range 170 - 285 µmol/L. Glycosylated albumin is formed by glycation of serum albumin. Several large, prospective studies have found that fructosamine and glycated albumin, a protein formed when glucose (sugar) binds to albumin, is a test that reflects short-term glycemia, predict conversion to T2D [47] with some studies demonstrating this effect even after controlling for baseline fasting glucose and HbA1c [48]. Glycated albumin was analyzed in serum at the UConn Health Clinical Research Center using a kit from MyBioSource Inc, San Diego, California. The assay has a range of 0.626 - 20 mg/mL, with a sensitivity of 0.1 mg/mL. Intra- and interassay CVs are less than 15%.

XSL•FO
RenderX

Glycemia

HbA_{1c} was measured in the UConn clinical laboratory using high pressure liquid chromatography. In this laboratory, HbA_{1c} shows the following coefficients of variation for normal and high values: level 1 mean=5.51%, CV=3.3 based on n=320, and level 2 mean=9.01%, CV=3.2 based on n=304.

Insulin Resistance

Glucose was reported in mg/dL and insulin was reported in uiU/mL. Log transformed homeostatic model assessment of insulin resistance (logHOMA-IR, Homeostatic Model Assessment of Insulin Resistance) was calculated from fasting glucose and insulin values according to the standard formula: log(fasting glucose × fasting insulin)/405 [49].

Sample Size and Power Analysis

Considering insulin resistance as our primary outcome, a previous study by our team provided data for a power analysis based on insulin resistance among persons at risk for T2D [50]. In that study, the SD for logHOMA-IR was approximately 0.3 and we expected to see a difference in logHOMA-IR of 0.4 between FI and food security, and a difference of 0.5. Our targeted sample size of n=100 would have 99% power to detect a main effect of FI, 96% power for mental distress and 87% power for the (FI × distress) interaction in a regression model using a 2-sided α of 0.05.

Planned Data Analysis

Summary statistics will be used initially to present the results from the categorical and continuous variables. Skewness, kurtosis, and potential outliers will be examined and if needed transformations or categorical variables will be created. The internal consistency of self-report measures will be assessed by Cronbach α.

To assess the relationship between monthly cycling of FI with diabetes risk markers, we will use repeated measures general linear mixed models. With 3 months of data, collected at weeks 2 and 4, depending on the distribution of for each outcome variable, the most appropriate distribution and link function will be chosen. In separate analyses it will be possible to divide the sample by frequency of benefits (ie, once vs twice monthly benefits) and include a dummy variable to evaluate the effects of frequency of benefits in the statistical models. It will also be possible to divide the sample into food secure, intermittent and persistent FI based on longitudinal studies using this approach [51].

To assess the hypothesis that psychological distress, diet, and eating disturbance mediate the relationship between FI and diabetes risk, we will perform a causal pathway analysis. We are interested in examining whether there is a change in our mediators over time, and if so, whether this change significantly leads to a change in diabetes risk markers. For this we will conduct dynamic mediation using latent change (difference) score mediation analysis [52,53].

For food frequency measures, linear regression models will compare food security status on household food availability. Spearman rank-order correlations of total inventoried availability of fruit and vegetable servings with the food frequency questionnaire will be conducted. We plan to use “GENMOD” or “MIXED” procedures in SAS (SAS Institute), or similar, for regression models and MPLUS software (Muthén & Muthén), or similar, for the mediation causal pathway analysis.

Ethical Considerations

If a recruit was not excluded, then recruiters obtained verbal consent, collected contact information, and scheduled the first visit at HHC. During this first visit, the recruiters obtained written informed consent. Participants were given a signed, dated copy of the informed consent form for their records and the recruiter kept the original, signed, dated copy. All verbal and written study materials, including recruitment materials and informed consent forms, were provided to the participants in their preferred language (English or Spanish). Self-report measures and their response options were pilot tested for readability, understanding, logic, flow, cultural relevance, and face validity by 5 Spanish speakers and 5 English speakers at the Hispanic Health Council. The final assessment battery was deemed appropriate by the research team. This protocol was approved by the UConn Health Institutional Review Board Study 21-045OS-1 and HHC agreed for UConn’s Health Institutional Review Board to be the institutional review board of record. All procedures were conducted according to the Declaration of Helsinki. Incentives were calculated to compensate participants for the time spent in study-related activities but were not coercive in a sample that screens positive for food insecurity (refer to Assessment Procedures and Incentives).

Results

As can be seen in the diagram in Figure 2, 845 potential participants were screened. The majority (n=685) were not eligible, the most common exclusions being because they did not receive SNAP benefits (n=229), were outside of the age range (n=155), had extant diabetes (n=90), or did not have elevated risk for type 2 diabetes (n=83). Of the 160 participants who were eligible, 73 declined with the most common reasons for declining being lack of time, lack of transportation, and concerns about COVID-19 pandemic. A total of 89 participants were consented and completed baseline blood assays but 2 withdrew immediately afterwards; 87 of them went on to be enrolled.

As can be seen in Table 2, just over half (54%) of the sampled self-identified as Puerto Rican, mean age was 35.1 (5.8) years, with 17.1 (11.6) years in the mainland United States. Just under half (47.1%) spoke Spanish only, 69% had no formal schooling and 31% had less than eighth grade education. Modal household size was 4 including 2 children, and 44.8% were not living with a partner. About half (54%) were unemployed, 63.2% reported a monthly income <US $1000 and 63.2% used food pantries.

Overall, participants completed 420 assessments or mean=4.83 (SD=2.02) assessments. Of the 87, 70.1% (n=61) completed all 6 assessments. Results of Spearman’s correlations show that higher number of assessments completed was associated with

XSL•FO
RenderX

having internet in the home (r=0.27, p=.01) and there was a statistical trend for having a tablet in the home (r=-0.20, p=.07).

XSL•FO
RenderX

Figure 2. CONSORT diagram for the food insecurity cycling microlongitudinal study Hartford, Connecticut, enrollment 2021 - 2023.

XSL•FO
RenderX

Table . Demographic characteristics (n=87).

Characteristics	Values
Ethnicity n (%)	
Puerto Rican	47 (54)
Non-Puerto Rican	40 (46)
Age, mean (SD), range (min-max)	35.1 (5.8) (18-44)
People Living in Home, n (%)	
1	1 (1.1)
2	7 (8)
3	21 (24.1)
4	24 (27.6)
5	23 (26.4)
6+	11 (12.6)
Children in home, n (%)	
0	5 (5.7)
1	16 (18.4)
2	30 (34.5)
3	24 (27.6)
4+	12 (13.7)
Years in United States, mean (SD), range (min-max)	17.1 (11.6) (0.3-43.3)
Language spoken, n (%)	
English and Spanish	46 (52.9)
Spanish only	41 (47.1)
Education, n (%)	
No formal schooling	60 (69)
Eighth grade or less	27 (31)
Relationship status, n (%)	
Living with spouse or partner	48 (55.2)
Not living with partner	39 (44.8)
Employment status, n (%)	
Employed	39 (44.8)
Unemployed	47 (54)
No answer	1 (1.1)
Household income ($), n (%)	
0-1000	55 (63.2)
1000 or more	32 (36.8)
Financial strain, n (%)	
Have enough and save	13 (14.9)
Have just enough	48 (55.2)
Not enough, with difficulties	25 (28.7)
No answer	1 (1.1)
Financial assistance, n (%)	
Cash	3 (3.4)

XSL•FO
RenderX

Characteristics	Values
Food Pantries	55 (63.2)
Soup Kitchen	5 (5.7)
Supplemental security income	15 (17.2)
Title 19/Medicaid	54 (62.1)
Section 8 housing	16 (18.4)
Nutrition for women, infants, children and children (WIC)	35 (40.2)
Other (housing and diapers)	4 (4.6)
Digital connectivity, n (%)	
Internet in the home	82 (94.3)
Tablet in the home	39 (44.8)
Computer in the home	37 (42)

Discussion

Principal Findings

The main findings reported here are that the study successfully recruited and retained a sample of the target population. There was excellent participation over time with 71% of the sample completing all 6 assessment timepoints. Participants were able to complete assessments using only their smart phones by cell phone service. Nonetheless, higher completion of assessments was associated with having internet and a tablet at home. This may reflect ease of connectivity (rather than access to connectivity) or a familiarity with technology that was advantageous. Other indicators of socioeconomic status (income and educational attainment) were not associated with number of assessments completed. As in other parts of the United States, digital inequity in Connecticut persists across race and ethnicity, income and education [54]. The surprisingly high rate of internet observed in this sample of Latinas who screened positive for food insecurity was likely due to assistance during COVID-19 pandemic for internet connection in the home. Future studies may achieve even higher rates of participation by providing tablets or internet service to participants, but the expense and ethical dilemmas of terminating internet service at the end of the study makes this approach potentially problematic.

This study was conducted during the COVID-19 pandemic. Social distancing requirements presented numerous challenges especially for recruitment, retention, and data collection. The study team remained flexible regarding format for assessments (telephone, Cisco WebEx, or in person when it was safe to do so) and strived to balance standardization with participant preference and evolving public health guidelines.

The innovative study protocol presented here will provide a comprehensive description of how FI is related to SNAP distribution, as well as, importantly, whether and to what degree FI cycling itself is related to risk for T2D. Furthermore, the study will provide early evidence for some putative mechanisms of this association, including contextual mechanisms such as household food availability, psychosocial mechanisms such as symptoms of depression and anxiety, and behavioral mechanisms such as distress-related binge eating. Each of these findings may suggest potential intervention targets.

Novelty

Food insecurity is a known risk factor for onset of new T2D [55]. However, we postulate that food insecurity is not static; rather, it varies according to access to food which is largely driven by the reductions in the amount of money available from social assistance programs (including SNAP) over the course of the month in some instances with families finding themselves with hardly any economic resources left for food or other basic needs by the last week of the benefits monthly cycle. This cycling has been hypothesized in the literature [56] yet microlongitudinal data regarding FI cycling is not available, as far as we know (eg, [57]). Furthermore, we propose that this cycling is related not only to household food inventory, but also to distress and distress-related eating behaviors. Our group has documented that FI is associated cross-sectionally with higher symptoms of depression, anxiety, and stress hormones as well as lower sleep quality [41]. Yet it remains unknown whether FI and emotional distress covary over time.

Another novel contribution of this study is the procedure for the household food inventory. We taught participants to complete the inventory checklist themselves using REDCap [25]. Participants were instructed how to access and enter data through a link sent to them through email. Lessons learned regarding the feasibility of the method, and its fidelity, will be important to inform future studies using household food inventories with hard-to-reach populations, or studies when home visits by researchers are not possible due to pandemic, natural disaster, or other situation that disrupts home visits. We pilot tested items and omitted the items in the checklist that were not understood. However, were unable to conduct true fidelity checks to compare the participant responses with CHW responses, because the university protocols were not allowing home visits for research purposes. If found to be feasible and reliable, this method may also be more efficient for research teams and perhaps preferable for participants.

XSL•FO
RenderX

Limitations

There are several limitations to this study. There is no experimental manipulation, so, whereas the longitudinal design can establish temporal precedence, no conclusions about causation can be drawn. This study was conducted during the COVID-19 pandemic. SNAP benefits since early 2020 have been affected by temporary pandemic-related benefit increases; an adjustment to the Thrifty Food Plan, upon which SNAP benefits are based, and higher-than-normal cost-of-living adjustments to reflect high food price inflation. Thus, findings may be influenced by changes to SNAP benefits during COVID-19 pandemic. COVID-19 pandemic also delayed and then paused the study, thus hindering recruitment. Data collection therefore took longer than anticipated and we fell short of our enrollment target by 13 participants. Future research with larger sample sizes should explore interactions between the exposure (FI) and mediators (psychological distress, diet, and eating disturbance) on outcomes (biomarkers of diabetes risk). Finally, the generalizability of our findings will need to be tested in other populations (eg, men, non-Latinos, or other parts of the country).

Limitations are balanced by the strengths of this study. The microlongitudinal design allows investigation of the temporal unfolding of covariance between FI and outcomes of interest over time. Home-based methods, particularly household food inventories, greatly improve contextual validity of assessments. The study affords the opportunity to confirm or assess the validity of certain measures in a Latina population, such as the Block Fat, Sugar, Fruit, and Vegetable Screener food frequency questionnaire. The population is high-risk and difficult-to-reach. Our procedures leverage the rigor and resources of a university medical center with the experience, knowledge, and reach of a community-based organization. Use of CHWs for recruitment, retention, and data collection improves trust and legitimacy in the target population. Multilevel bio-behavioral-psychosocial-contextual assessments are a particular strength. If our hypotheses are not refuted and this line of research goes on to establish causal pathways, interventions to disrupt the cycling of FI may prove beneficial for Latina health.

Conclusions

Results of this study may have potential policy implications. In 2022, 375,400 Connecticut residents, or 10% of the state population, received SNAP benefits [58]. More than 50% of recipients were in households with children and more than 37% were working families. Findings of this study may have implications for SNAP policy, particularly amount and timing of SNAP benefit distribution. Qualitative findings show that SNAP benefits are often inadequate for participants to purchase nutrient-rich foods that are sufficient for 1 month of household intake, particularly for participants at the higher end of the SNAP income distribution. Two studies have found that participant responses to food security surveys depend upon when in the SNAP cycle they are queried, with higher reports of FI when SNAP benefits are most salient [56,59,60]. This suggests considerable variability in food security within the month.

Inadequate benefit levels may contribute to cyclical eating patterns, where participants increase food consumption when benefits are received and reduce food consumption when benefits are depleted [14]. It may also lead to the purchase of quite unhealthy ultraprocessed foods and sugar sweetened beverages that have a long shelf-life, rather than fresh fruits and vegetables that are more likely to spoil over the month. In another study, individuals were not receiving benefits at baseline and then received 3 months of benefits. At baseline, food expenditures were not cyclical and were not related to impulsivity. However, total food expenditures became cyclical after households started receiving benefits, with more impulsive shoppers demonstrating a more accentuated benefit cycle [15]. The current monthly distribution of benefits could be replaced with a biweekly distribution to address cyclical food security, food expenditures, and eating patterns. Such a disbursement would provide SNAP recipients with greater flexibility.

Acknowledgments

This study was funded by the US National Institute of Diabetes and Digestive and Kidney Diseases 1R21DK122312-01A1. We thank the staff from the Hispanic Health Council for their assistance implementing this study.

Data Availability

The datasets generated during and/or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

ABM contributed to conceptualization, funding acquisition, investigation, methodology, project administration, supervision, writing original, and writing review. JAW managed methodology, writing original, and writing review. RPE handled methodology, supervision, and writing review. JG and RF conducted formal analysis, resources, and writing review. HOA and DK assisted with formal analysis, visualization, and writing review. SSP handled project administration, resources, supervision, and writing review. All authors approve the final version. All authors agree to be accountable for all aspects of the work.

References

1. National diabetes statistics report centers for disease control and prevention. 2022. URL: https://www.cdc.gov/diabetes/data/statistics/statistics-report.html [accessed 2025-03-19]

XSL•FO
RenderX

2. Avilés-Santa ML, Pérez CM, Schneiderman N, et al. Detecting prediabetes among Hispanics/Latinos from diverse heritage groups: does the test matter? Findings from the Hispanic Community Health Study/Study of Latinos. Prev Med 2017 Feb;95:110-118. [doi: 10.1016/j.ypmed.2016.12.009] [Medline: 27956225]
3. Anderson SA. Core indicators of nutritional state for difficult-to-sample populations. J Nutr 1990 Nov;120(11):1555-1598. [doi: 10.1093/jn/120.suppl_11.1555]
4. Pérez-Escamilla R. Food and nutrition security definitions, constructs, frameworks, measurements, and applications: global lessons. Front Public Health 2024;12:1340149. [doi: 10.3389/fpubh.2024.1340149] [Medline: 38550323]
5. Rabbitt MP, Reed-Jones M, Hales LJ, Burke MP. Household food security in the united states in 2023. Economic Research Service, US Department of Agriculture. 2024 Sep ERR-337. [doi: 10.32747/2024.8583175.ers]
6. Martin MA, Lippert AM. Feeding her children, but risking her health: the intersection of gender, household food insecurity and obesity. Soc Sci Med 2012 Jun;74(11):1754-1764. [doi: 10.1016/j.socscimed.2011.11.013] [Medline: 22245381]
7. Leung CW, Epel ES, Ritchie LD, Crawford PB, Laraia BA. Food insecurity is inversely associated with diet quality of lower-income adults. J Acad Nutr Diet 2014 Dec;114(12):1943-1953. [doi: 10.1016/j.jand.2014.06.353] [Medline: 25091796]
8. Fitzgerald N, Hromi-Fiedler A, Segura-Pérez S, Pérez-Escamilla R. Food insecurity is related to increased risk of type 2 diabetes among Latinas. Ethn Dis 2011;21(3):328-334. [Medline: 21942166]
9. Laraia BA. Food insecurity and chronic disease. Adv Nutr 2013 Mar 1;4(2):203-212. [doi: 10.3945/an.112.003277] [Medline: 23493536]
10. Hromi-Fiedler A, Bermúdez-Millán A, Segura-Pérez S, Pérez-Escamilla R. Household food insecurity is associated with depressive symptoms among low-income pregnant Latinas. Matern Child Nutr 2011 Oct;7(4):421-430. [doi: 10.1111/j.1740-8709.2010.00266.x] [Medline: 20735732]
11. Nguyen-Rodriguez ST, Chou CP, Unger JB, Spruijt-Metz D. BMI as a moderator of perceived stress and emotional eating in adolescents. Eat Behav 2008 Apr;9(2):238-246. [doi: 10.1016/j.eatbeh.2007.09.001] [Medline: 18329603]
12. Rasmusson G, Lydecker JA, Coffino JA, White MA, Grilo CM. Household food insecurity is associated with binge-eating disorder and obesity. Int J Eat Disord 2018 Dec 19. [doi: 10.1002/eat.22990] [Medline: 30565270]
13. Schwarzer R. Self-regulatory processes in the adoption and maintenance of health behaviors. J Health Psychol 1999 Mar;4(2):115-127. [doi: 10.1177/135910539900400208] [Medline: 22021474]
14. Leung CW, Hoffnagle EE, Lindsay AC, et al. A qualitative study of diverse experts' views about barriers and strategies to improve the diets and health of Supplemental Nutrition Assistance Program (SNAP) beneficiaries. J Acad Nutr Diet 2013 Jan;113(1):70-76. [doi: 10.1016/j.jand.2012.09.018] [Medline: 23260725]
15. Valluri S, Mason SM, Peterson HH, French SA, Harnack LJ. Trends in cyclical food expenditures among low-income households receiving monthly nutrition assistance: results from a prospective study. Public Health Nutr 2021 Feb;24(3):536-543. [doi: 10.1017/S136898002000405X] [Medline: 33059779]
16. Whiteman ED, Chrisinger BW, Hillier A. Diet quality over the monthly supplemental nutrition assistance program cycle. Am J Prev Med 2018 Aug;55(2):205-212. [doi: 10.1016/j.amepre.2018.04.027] [Medline: 29935945]
17. Hager ER, Quigg AM, Black MM, et al. Development and validity of a 2-item screen to identify families at risk for food insecurity. Pediatrics 2010 Jul;126(1):e26-e32. [doi: 10.1542/peds.2009-3146] [Medline: 20595453]
18. Gundersen C, Engelhard EE, Crumbaugh AS, Seligman HK. Brief assessment of food insecurity accurately identifies high-risk US adults. Public Health Nutr 2017 Jun;20(8):1367-1371. [doi: 10.1017/S1368980017000180] [Medline: 28215190]
19. Ward ED, Hopkins WA, Shealy K. Evaluation of the use of a diabetes risk test to identify prediabetes in an employee wellness screening. J Pharm Pract 2022 Aug;35(4):524-527. [doi: 10.1177/0897190021997010] [Medline: 33622071]
20. Scanlan AB, Maia CM, Perez A, Homko CJ, O'Brien MJ. Diabetes risk assessment in Latinas: effectiveness of a brief diabetes risk questionnaire for detecting prediabetes in a community-based sample. Diabetes Spectr 2018 Feb;31(1):31-36. [doi: 10.2337/ds16-0051] [Medline: 29456424]
21. Olson CM. Food insecurity in women. Top Clin Nutr 2005 Oct;20(4):321-328. [doi: 10.1097/00008486-200510000-00005]
22. Mortality tables, statewide by race/ethnicity, 2005-2007 2010. 2011 Mar 3. URL: http://www.ct.gov/dph/lib/dph/hisr/hcqsar/mortality/xls/AAMR_State_RaceEthn_aggregate_2005-2007c.xls [accessed 2025-03-19]
23. Wagner J, Bermudez-Millan A, Damio G, et al. Community health workers assisting Latinos manage stress and diabetes (CALMS-D): rationale, intervention design, implementation, and process outcomes. Transl Behav Med 2015 Dec;5(4):415-424. [doi: 10.1007/s13142-015-0332-1] [Medline: 26622914]
24. Hinkin CH, Castellon SA, Dickson - Fuhrman E, Daum G, Jaffe J, Jarvik L. Screening for drug and alcohol abuse among older adults using a modified version of the CAGE. American J Addict 2001 Oct;10(4):319-326 [FREE Full text] [doi: 10.1111/j.1521-0391.2001.tb00521.x]
25. Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: Building an international community of software platform partners. J Biomed Inform 2019 Jul;95:103208. [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]
26. Hromi-Fiedler A, Bermúdez-Millán A, Melgar-Quiñonez H, Pérez-Escamilla R. Psychometric properties of an adapted version of the U.S. household food security survey module for assessing food insecurity among low-income pregnant Latinas. J Hunger Environ Nutr 2009 Jan 1;4(1):81-94. [doi: 10.1080/19320240802706866] [Medline: 20046913]

XSL•FO
RenderX

27. Hromi-Fiedler A, Bermúdez-Millán A, Segura-Pérez S, Damio G, Pérez-Escamilla R. Adaptation of the U.S. food security survey module for low-income pregnant Latinas: qualitative phase. J Hunger Environ Nutr 2009 Jan 1;4(1):62-80. [doi: 10.1080/19320240802706841] [Medline: 20046909]
28. Lalonde I, Graham M, Slovinec-D'Angelo M, Beaton L, Brown J, Block T. Validation of the block fat/sugar/fruit/vegetable screener in a cardiac rehabilitation setting. J Cardiopulm Rehabil Prev 2008;28(5):340. [doi: 10.1097/01.HCR.0000336162.26720.95]
29. Block G, Gillespie C, Rosenbaum EH, Jenson C. A rapid food screener to assess fat and fruit and vegetable intake. Am J Prev Med 2000 May;18(4):284-288. [doi: 10.1016/s0749-3797(00)00119-7] [Medline: 10788730]
30. Boushey CJ, Spoden M, Zhu FM, Delp EJ, Kerr DA. New mobile methods for dietary assessment: review of image-assisted and image-based dietary assessment methods. Proc Nutr Soc 2017 Aug;76(3):283-294. [doi: 10.1017/S0029665116002913] [Medline: 27938425]
31. Karlsson J, Persson LO, Sjöström L, Sullivan M. Psychometric properties and factor structure of the Three-Factor Eating Questionnaire (TFEQ) in obese men and women. Results from the Swedish Obese Subjects (SOS) study. Int J Obes Relat Metab Disord 2000 Dec;24(12):1715-1725. [doi: 10.1038/sj.ijo.0801442] [Medline: 11126230]
32. Stunkard AJ, Messick S. The three-factor eating questionnaire to measure dietary restraint, disinhibition and hunger. J Psychosom Res 1985;29(1):71-83. [doi: 10.1016/0022-3999(85)90010-8] [Medline: 3981480]
33. Jáuregui-Lobera I, García-Cruz P, Carbonero-Carreño R, Magallares A, Ruiz-Prieto I. Psychometric properties of Spanish version of the Three-Factor Eating Questionnaire-R18 (Tfeq-Sp) and its relationship with some eating- and body image-related variables. Nutrients 2014 Dec 4;6(12):5619-5635. [doi: 10.3390/nu6125619] [Medline: 25486370]
34. Toolkit P. Questionnaire on eating and weight patterns - adult. 2024 Apr 17. URL: https://www.phenxtoolkit.org/protocols/view/651201 [accessed 2025-03-19]
35. Yanovski SZ, Marcus MD, Wadden TA, Walsh BT. The questionnaire on eating and weight patterns-5: an updated screening instrument for binge eating disorder. Int J Eat Disord 2015 Apr;48(3):259-261. [doi: 10.1002/eat.22372] [Medline: 25545458]
36. Irwin DE, Stucky B, Langer MM, et al. An item response analysis of the pediatric PROMIS anxiety and depressive symptoms scales. Qual Life Res 2010 May;19(4):595-607. [doi: 10.1007/s11136-010-9619-3] [Medline: 20213516]
37. Craig BM, Reeve BB, Cella D, Hays RD, Pickard AS, Revicki DA. Demographic differences in health preferences in the United States. Med Care 2014 Apr;52(4):307-313. [doi: 10.1097/MLR.0000000000000066] [Medline: 24374420]
38. Gilmer TP, Walker C, Johnson ED, Philis-Tsimikas A, Unützer J. Improving treatment of depression among Latinos with diabetes using project Dulce and IMPACT. Diabetes Care 2008 Jul;31(7):1324-1326. [doi: 10.2337/dc08-0307] [Medline: 18356401]
39. Merz EL, Malcarne VL, Roesch SC, Riley N, Sadler GR. A multigroup confirmatory factor analysis of the Patient Health Questionnaire-9 among English- and Spanish-speaking Latinas. Cultur Divers Ethnic Minor Psychol 2011 Jul;17(3):309-316. [doi: 10.1037/a0023883] [Medline: 21787063]
40. Buysse DJ, Reynolds CF 3rd, Monk TH, Berman SR, Kupfer DJ. The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research. Psychiatry Res 1989 May;28(2):193-213. [doi: 10.1016/0165-1781(89)90047-4] [Medline: 2748771]
41. Bermúdez-Millán A, Pérez-Escamilla R, Segura-Pérez S, et al. Psychological distress mediates the association between food insecurity and suboptimal sleep quality in Latinos with type 2 diabetes mellitus. J Nutr 2016 Oct;146(10):2051-2057. [doi: 10.3945/jn.116.231365] [Medline: 27489004]
42. Hromi-Fiedler A, Chapman D, Segura-Pérez S, et al. Barriers and facilitators to improve fruit and vegetable intake among WIC-eligible pregnant Latinas: an application of the health action process approach framework. J Nutr Educ Behav 2016;48(7):468-477. [doi: 10.1016/j.jneb.2016.04.398] [Medline: 27373861]
43. Martinez-Brockman JL, Harari N, Segura-Pérez S, Goeschel L, Bozzi V, Pérez-Escamilla R. Impact of the lactation advice through texting can help (LATCH) trial on time to first contact and exclusive breastfeeding among WIC participants. J Nutr Educ Behav 2018 Jan;50(1):33-42. [doi: 10.1016/j.jneb.2017.09.001] [Medline: 29325660]
44. Hearst MO, Fulkerson JA, Parke M, Martin L. Validation of a home food inventory among low-income Spanish- and Somali-speaking families. Public Health Nutr 2013 Jul;16(7):1151-1158. [doi: 10.1017/S1368980012004508] [Medline: 23034097]
45. Fulkerson JA, Nelson MC, Lytle L, Moe S, Heitzler C, Pasch KE. The validation of a home food inventory. Int J Behav Nutr Phys Act 2008 Nov 4;5:55. [doi: 10.1186/1479-5868-5-55] [Medline: 18983668]
46. Kong A, Fitzgibbon M, Hubbard C, Campbell RT, Kessee N, Schiffer L. Validation of a self-report home food availability checklist against in-home food inventories conducted in low-income Black/African American and Hispanic/Latinx households with preschool-age children. Appetite 2022 May 1;172:105964. [doi: 10.1016/j.appet.2022.105964] [Medline: 35124159]
47. Zaccardi F, Kurl S, Pitocco D, Ronkainen K, Laukkanen JA. Serum fructosamine and risk of type 2 diabetes mellitus among middle-age Finnish men: a 23-year population-based prospective study. Acta Diabetol 2015 Feb;52(1):161-166. [doi: 10.1007/s00592-014-0625-8] [Medline: 25026924]
48. Juraschek SP, Steffes MW, Miller ER III, Selvin E. Alternative markers of hyperglycemia and risk of diabetes. Diabetes Care 2012 Nov;35(11):2265-2270. [doi: 10.2337/dc12-0787] [Medline: 22875225]

XSL•FO
RenderX

49. Matthews DR, Hosker JP, Rudenski AS, Naylor BA, Treacher DF, Turner RC. Homeostasis model assessment: insulin resistance and beta-cell function from fasting plasma glucose and insulin concentrations in man. Diabetologia 1985 Jul;28(7):412-419. [doi: 10.1007/BF00280883] [Medline: 3899825]
50. Wagner JA, Bermúdez-Millán A, Buckley TE, et al. Community-based diabetes prevention randomized controlled trial in refugees with depression: effects on metabolic outcomes and depression. Sci Rep 2023 May 30;13(1):8718. [doi: 10.1038/s41598-023-35738-9] [Medline: 37253820]
51. Garcia J, Hromi-Fiedler A, Mazur RE, et al. Persistent household food insecurity, HIV, and maternal stress in peri-urban Ghana. BMC Public Health 2013 Mar 11;13:215. [doi: 10.1186/1471-2458-13-215] [Medline: 23497026]
52. Preacher KJ. Advances in mediation analysis: a survey and synthesis of new developments. Annu Rev Psychol 2015 Jan 3;66:825-852. [doi: 10.1146/annurev-psych-010814-015258] [Medline: 25148853]
53. Selig JP, Preacher KJ. Mediation models for longitudinal data in developmental research. Res Hum Dev 2009 Jun 5;6(2-3):144-164. [doi: 10.1080/15427600902911247]
54. State of connecticut administrative services. State Digital Equity Plan. 2024. URL: https://portal.ct.gov/das/-/media/das/ctedtech/meetings/2024/9-9-24_de_updates.pdf [accessed 2025-03-19]
55. Seligman HK, Bindman AB, Vittinghoff E, Kanaya AM, Kushel MB. Food insecurity is associated with diabetes mellitus: results from the National Health Examination and Nutrition Examination Survey (NHANES) 1999-2002. J Gen Intern Med 2007 Jul;22(7):1018-1023. [doi: 10.1007/s11606-007-0192-6] [Medline: 17436030]
56. Gregory CA, Todd JE. SNAP timing and food insecurity. PLoS ONE 2021;16(2):e0246946. [doi: 10.1371/journal.pone.0246946] [Medline: 33630866]
57. Kinsey EW, Dupuis R, Oberle M, Cannuscio CC, Hillier A. Chronic disease self-management within the monthly benefit cycle of the Supplemental Nutrition Assistance Program. Public Health Nutr 2019 Aug;22(12):2248-2259. [doi: 10.1017/S1368980019001071] [Medline: 31104648]
58. Priorities CoBaP. A closer look at who benefits from SNAP: state-by-state fact sheets. 2023. URL: https://www.cbpp.org/research/a-closer-look-at-who-benefits-from-snap-state-by-state-fact-sheets#Connecticut [accessed 2025-03-19]
59. Gregory CA, Smith TA. Salience, food security, and SNAP receipt. J Policy Anal Manage 2019;38(1):124-154. [Medline: 30572412]
60. Pérez-Escamilla R, Ferris AM, Drake L, et al. Food stamps are associated with food security and dietary intake of inner-city preschoolers from Hartford, Connecticut. J Nutr 2000 Nov;130(11):2711-2717. [doi: 10.1093/jn/130.11.2711] [Medline: 11053511]

Abbreviations

CAGE-AID: CAGE Adapted to Include Drugs
CHW: community health workers
CV: coefficient of variation
FI: Food insecurity
GED: General Educational Development high school equivalency exam
HAPA: Health Action Process Approach
HbA1c: glycated hemoglobin
HHC: Hispanic Health Council
HIPAA: Health Insurance Portability and Accountability Act
logHOMA-IR: Log transformed homeostatic model assessment of IR
PHQ: Personal Health Questionnaire
PSQI: Pittsburgh Sleep Quality Index
REDCap: Research Electronic Data Capture
SNAP: Supplementary Nutrition Assistance Program
T2D: type 2 diabetes
WIC program: Women, Infants, and Children program

Edited by A Mavragani; submitted 27.09.24; peer-reviewed by MM Lee, R Bipat; revised version received 18.02.25; accepted 20.02.25; published 28.03.25.

Please cite as:
Bermúdez-Millán A, Pérez-Escamilla R, Segura-Pérez S, Grady J, Feinn VI RS, Agresta H, Kim D, Wagner JA
The Monthly Cycling of Food Insecurity in Latinas at Risk for Diabetes: Methods, Retention, and Sample Characteristics for a Microlongitudinal Design
JMIR Form Res 2025;9:e66970
URL: https://formative.jmir.org/2025/1/e66970
doi:10.2196/66970

XSL•FO
RenderX

© Angela Bermúdez-Millán, Rafael Pérez-Escamilla, Sofia Segura-Pérez, James Grady, Richard S Feinn VI, Hanako Agresta, Dean Kim, Julie Ann Wagner. Originally published in JMIR Formative Research (https://formative.jmir.org), 28.3.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Improving Health and Well-Being of People With Post–COVID-19 Consequences in South Africa: Situation Analysis and Pilot Intervention Design

Nicole Audrey Glover[1], MBChB; Farzana Sathar[1], MPH; Pride Mokome[1], BCur; Nkululeko Mathabela[1], BA; Sipokazi Taleni[1], PGDip; Sarah Alexandra van Blydenstein[2], CertPulm(SA)Phys; Anna-Maria Mekota[3], Dr rer nat; Salome Charalambous[1,4], PhD; Andrea Rachow[3,5,6], MD; Olena Ivanova[3,6], PhD

[1]Implementation Research Division, The Aurum Institute, 29 Queens Road, Parktown, Johannesburg, South Africa

[2]Department of Internal Medicine, Chris Hani Baragwanath Academic Hospital, University of the Witwatersrand, Johannesburg, South Africa

[3]Institute of Infectious Diseases and Tropical Medicine, Ludwig-Maximilians-Universität University Hospital, Munich, Germany

[4]School of Public Health, University of the Witwatersrand, Johannesburg, South Africa

[5]German Research Centre for Environmental Health, Helmholtz Zentrum München, Unit Global Health, Neuherberg, Germany

[6]German Center for Infection Research, Partner Site Munich, Munich, Germany

Corresponding Author:
Nicole Audrey Glover, MBChB
Implementation Research Division, The Aurum Institute, 29 Queens Road, Parktown, Johannesburg, South Africa

Abstract

Background: Multisystemic complications post–COVID-19 infection are increasingly described in the literature, yet guidance on the management remains limited.

Objectives: This study aimed to assess the needs, preferences, challenges, and existing interventions for individuals with post–COVID-19 symptoms. Based on this, we aimed to develop a context-adapted intervention to improve the overall health and well-being of individuals with post–COVID-19 complications.

Methods: We conducted a cross-sectional mixed-methods situation analysis assessing the needs, preferences, challenges, and existing interventions for patients with post–COVID-19 symptoms. We collected data through questionnaires, semistructured in-depth interviews, and focus group discussions (FGDs) from individuals diagnosed with COVID-19 within the previous 18-month period and health care providers who managed patients with COVID-19 in both inpatient and outpatient settings. Quantitative data were summarized using descriptive statistics, qualitative data were transcribed, and deductive analysis focused on suggestions for future interventions. Findings guided the development of a group intervention.

Results: We conducted 60 questionnaires, 13 interviews, and 3 FGDs. Questionnaires showed limited knowledge of post–COVID-19 complications at 26.7% (16/60). Of those who received any rehabilitation for COVID-19 (19/60, 31.7%), 94.7% (18/19) found it helpful for their recovery. Just over half (23/41, 56%) of those who did not receive rehabilitation reported that they would have liked to. The majority viewed rehabilitation as an important adjunct to post–COVID-19 care (56/60, 93.3%) and that support groups would be helpful (53/60, 88.3%). Qualitative results highlighted the need for mental health support, structured post–COVID-19 follow-up, and financial aid in post–COVID-19 care. Based on the insights from the situation analysis, the theory of change framework, and existing post–COVID-19 evidence, we designed and conducted a pilot support group and rehabilitation intervention for individuals with post–COVID-19 complications. Our main objective was to assess the change in physical and psychological well-being pre- and postintervention. The intervention included 8 weekly themed group sessions supplemented by home tasks. Effectiveness of the intervention was evaluated by questionnaires pre- and postintervention on post–COVID-19 symptoms, quality of life with the EuroQoL 5-Dimension 5-Level, short Warwick-Edinburgh Mental Wellbeing Scale, and physical function by spirometry and 1-minute sit-to-stand test. We also assessed the feasibility and acceptability of the intervention by questionnaires and semistructured in-depth interviews. The intervention outcome analysis is yet to be conducted.

Conclusions: Insights from patients and health care providers on the characteristics of post–COVID-19 complications helped guide the development of a context-adapted intervention program with potential to improve health and well-being post–COVID-19.

(JMIR Form Res 2025;9:e58436) doi:10.2196/58436

XSL•FO
RenderX

KEYWORDS

post–COVID-19; rehabilitation; support; quality of life; group care; well-being; South Africa; COVID-19; situation analysis; pilot; intervention; context-adapted; physical health; mental health; cross-sectional; mixed method; questionnaire; in-depth; interviews; survey; focus group; quantitative; qualitative; support group; hospital; patients; health care workers; health worker

Introduction

Most people who have had SARS-CoV-2 infection, causing COVID-19 disease, recover fully; however, a proportion experience longer-term complications resulting in persistent symptoms and functional impairment [1,2]. Terms used to describe this syndrome include long COVID-19, post-acute COVID-19 syndrome, or post–COVID-19 condition; however, the definitions remain heterogeneous [1-3]. Post–COVID-19 complications can affect multiple body systems including pulmonary, cardiovascular, neurological, and musculoskeletal systems, where over 50 long-term symptoms have been identified [1,4]. The prevalence of post–COVID-19 complications varies substantially, from 6% up to 80%, depending on definitions used, hospitalization status, and duration of follow-up [2,4,5]. Common symptoms include fatigue, headache, arthralgia and myalgia, dyspnea and notably neurocognitive, mental health, and sleep disorders [2,6,7]. In addition to the symptoms, objective assessments of lung, cardiac, and neurological function demonstrate long-term pathology [1,2,8]. These complications are reported to impact activities of daily living (ADL), quality of life, and ability to work [9,10].

Limited data exists in African countries on post-COVID-19 morbidity and its characterization [2,4,6]. A study conducted in South Africa demonstrated COVID-19-related symptoms in 66.7% of people 3 months posthospital discharge [11], which was supported in another study by Jassat et al [12]. Further descriptions of post–COVID-19 complications in the South African context have come from our previous study, demonstrating pathology in hospitalized post–COVID-19 individuals, where up to 1-month posthospital discharge 85.2% had abnormal chest–computed tomography, 49.5% had abnormal spirometry, 79% had abnormal diffusing capacity of the lung, and functional limitation was evidenced by reduced median 6-minute walking test (300, IQR 210-400 m) and elevated St George's Respiratory Questionnaire scores (median 21.6, IQR 9.2-48.3) [13]. Despite these findings, sample sizes are small and the breadth of data and characterization of post–COVID sequelae in these cohorts is limited.

In addition to the multisystem sequelae recognized post–COVID-19, recovery can be complicated by preexisting or newly developed chronic diseases such as cardiovascular disease, diabetes, HIV, and tuberculosis (TB) [14]. Understanding these multisystem complications is integral to directing health services effectively [9]. Due to the complexity of recovery post–COVID-19, a holistic approach should be taken for management. Furthermore, the burden of disease and local health resources differ across regions [2,6], demanding the evaluation of context-specific assessment tools and interventions, to improve long-term recovery post–COVID-19 [15-17].

The World Health Organization (WHO) and National Institute for Health and Care Excellence (NICE) guidelines recommend multidisciplinary and patient-centered care for post–COVID-19 complications [3,18]. In a randomized control trial conducted by Liu et al [19], a respiratory rehabilitation program was shown to improve respiratory function, quality of life, and anxiety. These findings are reflected in another study where post–COVID-19 rehabilitation showed improvement in lung function and a reduction in anxiety levels [20].

Considering this multidisciplinary approach in resource-constrained settings, such as low- and middle-income countries (LMICs), assisted self-management programs and peer support groups become pertinent. This was evidenced by Wright et al [21], where a self-management program guided by weekly educational support sessions demonstrated improvement in patient self-efficacy and mental well-being post–COVID-19 [21,22]. Self-management techniques and patient education with other chronic lung diseases, including post-TB lung disease and bronchiectasis, have been noted to empower patients, improve patient confidence, competence, and overall quality of life [23,24]. However, few studies are available on context-adapted and culturally appropriate tools and interventions for the management of post–COVID-19 in LMICs, the importance of which was highlighted by Jassat et al [25].

This paper aims to describe the findings from a situation analysis on post–COVID-19 rehabilitation services and the development of a context-adapted pilot intervention aimed to improve overall physical and mental health post–COVID-19 conducted in Johannesburg, South Africa.

Methods

Both the situation analysis and intervention have been conducted. Below we outline the methods used for the situation analysis and framework for intervention development.

Situation Analysis

We conducted a mixed-method situation analysis study with our main aim to explore the post–COVID-19 management landscape by assessing the needs, preferences, challenges, and existing interventions and strategies for patients with post–COVID-19 complications in Johannesburg, South Africa.

We conducted this study at the Chris Hani Baragwanath Academic Hospital (CHBAH) and Tembisa Provincial Tertiary Hospital (Tembisa), Johannesburg, South Africa. These health care facilities are located in townships in the City of Johannesburg and Ekurhuleni districts and cater to the general population. They are public health care facilities funded by the Gauteng Department of Health. Participants were recruited from outpatient medical and rehabilitation departments at the respective facilities, where patients who had COVID-19 disease were invited to participate, and health care workers who had managed patients with COVID-19 were invited to participate.

XSL•FO
RenderX

Potential participants were approached in the waiting areas of clinics, where prescreening procedures assessing COVID-19 history were done followed by full informed consent if potential participants were willing. The study was conducted between December 2022 and February 2023, which was 11-months after the last clear reported COVID-19 wave in South Africa [26].

We collected quantitative data through a cross-sectional standardized questionnaire-based interview (see Multimedia Appendix 1) administered by the research staff with individuals who were previously diagnosed with COVID-19. The questionnaire covered topics on post–COVID-19 symptoms, experience with rehabilitation programs, challenges, preferences, and suggestions for potential interventions. A total of 2 research assistants (male and female) and a research nurse (female) were trained and collected the survey data. We used descriptive statistics to summarize the data in frequencies, mean and median, using Microsoft Excel.

The qualitative study included semistructured in-depth interviews with health care workers who manage COVID-19 patients and focus group discussions (FGDs) with individuals who had post–COVID-19 symptoms. Separate interviews were used with health care workers to allow in-depth conversation on each discipline, and FGDs were used with patients to allow a broad conversation flow of COVID-19 reflection and to encourage group sharing (see Multimedia Appendices 2 and 3). Interviews and FGDs were facilitated by the research team (3 females and 1 male), in English, except 2 FGD's conducted in a combination of English, isiZulu, and Sepedi. The duration of the interviews and FGDs ranged from 10-minutes to 90-minutes depending on participant availability and willingness to engage in the content of the questions. A guide was used for the interviews focusing on health care worker reflection on managing COVID-19 patients, challenges and perceived need for rehabilitation; and a guide was used in the FGDs to discuss patient post–COVID-19 experiences and rehabilitation needs and challenges. Both guides allowed for flexibility of the discussion. Interviews and FGDs were recorded, transcribed, coded, and deductive thematic analysis was carried out focusing on main concepts for the intervention development.

Quantitative and qualitative data were triangulated through an iterative process to ensure the coherence of findings and obtain more insights on the desirable content and format of the intervention.

Intervention Development

The creation of the intervention followed the steps described in the Group Intervention framework [27]. We ensured a suitable context and setting for the intervention. Following this, we designed the group intervention addressing outcomes and measurements, sampling into groups, quantity of groups, and structure of the intervention. Finally, during the delivery of the intervention, we introduced team briefings and group notes to describe and reflect on the dynamics within each group and session, position of group participants and leaders, and communication.

The data collected in the situation analysis and available evidence in post–COVID-19 rehabilitation fed into the content and delivery format of the intervention. The design process, implementation, and evaluation plan for the intervention are described in more detail in the Results section. The outcomes of the intervention are being analyzed and will be presented in a separate paper.

Ethical Considerations

We obtained ethical approval for the situation analysis from Wits Human Research Ethics Committee (230405), and the 2 facilities, CHBAH and Tembisa Hospital. All participants gave written informed consent before participation. The study data are deidentified to protect participant confidentiality. Participants received R 300 (US $16.85) cash reimbursement for each study visit covering costs for travel and time.

Results

Situation Analysis and Needs Assessment

Quantitative Results

We conducted 60 questionnaires with individuals who had COVID-19; Table 1 shows the findings. We had 21.7% (13/60) male participants, with a median age of 49.5 (IQR 39.5-60) years. Knowledge of post–COVID-19 sequelae was limited, with only 26.7% (16/60) having heard of these complications. From the 60 participants, 31.7% (19/60) received any rehabilitation for COVID-19 in either inpatient or outpatient setting, where physiotherapy was the most common (17/19, 89.5%), followed by dietician (4/19, 21.1%), speech therapy and audiology (3/19, 15.8%), psychology (2/19, 10.5%), and occupational therapy (1/19, 5.3%). Of those who received rehabilitation, all received rehabilitation within 6 weeks of COVID-19 infection and 94.7% (18/19) found it helpful for recovery. Of those who did not receive any rehabilitation, 56.1% (23/41) reported that they would have liked to receive rehabilitation, notably psychological support and counseling (14/23, 60.9%) and physiotherapy (8/23, 34.8%). Majority of the participants felt that rehabilitation was an important adjunct to COVID-19 recovery (56/60, 93.3%) and said that they would find a support group helpful for their recovery post–COVID-19 (53/60, 88.3%).

XSL•FO
RenderX

Table . Characteristics and questionnaire responses from participants (N=60). Quantitative data collected in situation analysis at Chris Hani Baragwanath Academic Hospital and Tembisa Provincial Tertiary Hospital from December 2022 to February 2023.

Characteristic	Participants (N=60)
Sex (male), n (%)	13 (21.7)
Age (years), median (IQR)	49.5 (39.5-60)
Knowledge of post–COVID-19 complications[a], n (%)	16 (26.7)
Received rehabilitation for acute COVID-19 or post–COVID-19 complications, n (%)	19 (31.7)
Inpatient rehabilitation	14 (73.7)
Outpatient rehabilitation	5 (26.3)
Physiotherapy	17 (89.5)
Dietician	4 (21.1)
Speech therapy and audiology	3 (15.8)
Psychology	2 (10.5)
Occupational therapy	1 (5.3)
Received rehabilitation for acute COVID-19 or post–COVID-19 complications and found it helpful for recovery, n (%)	18 (30)
Did not receive rehabilitation and would have liked to, n (%)	23 (38.3)
Psychology or counseling requested	14 (23.3)
Physiotherapy requested	8 (13.3)
Main goals for recovery post–COVID-19, n (%)	
To carry out activities of daily living better	34 (56.7)
To be able to exercise or have better physical capacity	32 (53.3)
Improved mental health or coping strategies	25 (41.7)
Return to work	21 (35)
Connection with people and to be able to care for family	20 (33.3)
Feel rehabilitation is an important adjunct to care post–COVID-19, n (%)	56 (93.3)
Would find a support group helpful for recovery post–COVID-19, n (%)	53 (88.3)

[a]Knowledge of post–COVID-19 complications was assessed by asking “Do you know what long-COVID or post-COVID-19 condition is?” and allowing participants to answer yes or no and elaborate on their answer.

Qualitative Results

We conducted 13 interviews with health care providers and 3 FGDs with a total of 20 post–COVID-19 patients. The 3 main themes that emerged from the thematic analysis for potential interventions post–COVID-19 included the need for mental health support, structured post–COVID-19 follow-up, and the need for financial aid.

Need for mental health support was further explained by the request to focus on psychological well-being and support groups. Both health care providers and patients reported that psychological well-being was neglected during the pandemic and should be a focus in post–COVID-19 care. One participant reported:

> *I think we focus more on the physical healing than the psychological healing. So, most people who had Covid, they are not okay. They had post-Covid depression or post-Covid anxiety. So, I think that there is a lot of mental health [care] that needs to be done post-Covid.* [Female, health care provider]

Health care providers and patients who had opportunities to have debriefing sessions and support groups, further expanded on the importance of incorporating support groups into the mental health support post–COVID-19:

> *They really must get help. From our side… sitting together talking, [about our experience], it’s a good thing. It’s a helping thing, it’s a very, very good thing. Thank you.* [Male, patient in FGD]

This importance of sharing experiences and leaning on additional support from others in the group context during post–COVID-19 care was further supported by a health care provider, saying:

> *… looking at a rehab unit where… people feel comfortable with coming in… sharing their stories to each other, and their success stories, because that’s how you then motivate each other. It’s not coming*

XSL•FO
RenderX

> *from someone that may not understand it, you're hearing it from patient to patient. And that might be nice.* [Female, health care provider]

In addition to the mental health support, structured medical post–COVID-19 follow-up care was also reported as an important part of post–COVID-19 care, where both patients and health care providers should have a structured follow-up program, rehabilitation, and referral to specialists depending on their complications. The importance of recognizing each individual and the treatment that they require was emphasized:

> *We are not statistics, we are human beings, we are sick, and we are affected by COVID, and I think there must be a way to create a rehab for us [to have] proper check-up for us, because she is telling [us] about her lungs and I am telling you about my eyes and what are they doing? So, in six months or one year you must take this person to a specialist for check-ups.* [Female, health care provider]

Finally, financial aid was recognized as an important area for support post–COVID-19, largely expressed by patients and recognized by health care workers. Patients reported that their daily functioning is worse compared with their functioning before COVID-19. Financial assistance was requested for those who can no longer work due to their complications post–COVID-19:

> *… he said he is no longer going to work. His salary is no longer the same and also our sister here, she her health is no longer the same. I think they are supposed to approve her for grant.* [Female, patient in FGD]

Taking into account our findings from both the quantitative and qualitative data, there was a strong emphasis on mental health support, support groups, and structured medical and rehabilitation follow-up post–COVID-19.

Design and Implementation of the Intervention

Using the information obtained from the situation analysis and existing evidence on post–COVID-19 rehabilitation, we developed an in-person support group program for people with post–COVID-19 complications titled CoPilot. The main study objective was to assess the change in physical and psychological well-being in patients with post–COVID-19 sequelae before and after the pilot intervention. To guide the program and evaluation we present a theory of change (Figure 1). The main intended impact of the intervention is to improve the overall health and well-being of COVID-19 survivors.

The target population for the intervention was individuals who had been diagnosed with COVID-19 and were experiencing ongoing post–COVID-19 complications according to the WHO post–COVID-19 condition definition [3]. The participants were recruited from out-patient medical and rehabilitation departments at the health care facility, including both patients and health care providers, and further participants were recruited by snowball sampling. Eligibility criteria included: self-reported or diagnosed previous COVID-19; adults aged 18-years and older; reporting at least one symptom lasting for more than 2 months, at least 3 months since COVID-19, which could not otherwise be explained by another condition; and willingness to provide written informed consent and to follow study procedures. Reported symptoms could include, but were not limited to, fatigue, breathlessness, pain or discomfort, impaired sleep, or impaired usual activity [2]. Participants with known severe medical or psychiatric conditions which affected ability to consent or participate, or required a higher level of care were excluded.

Group formats in rehabilitation have been shown to strengthen curative factors like universality and altruism, to provide peer and social support, to share coping strategies, and to practice gratitude [28,29]. Taking this into account, with the strong findings for group support in post–COVID-19 rehabilitation from our situation analysis, we designed and conducted a group-based intervention. The program consisted of weekly in-person group sessions over eight-weeks complimented by home-tasks and self-management exercises, with the option to adapt tasks to each person's abilities. We had 7 groups where each group consisted of 8 to 10 participants. Our target sample size was 60, which was informed by evidence that sample sizes of 24 - 50 are sufficient to estimate the key parameters of an efficacy and feasibility trial [30]. Sessions were guided by the predesigned schedule and a duration of one-hour, but with flexibility for participant experiences and needs. The core of the sessions involved research staff facilitating discussion around participants' post–COVID-19 experiences, challenges and self-management strategies whilst encouraging group support and education on post–COVID-19 complications. Select sessions included invited rehabilitation specialists (eg, psychologist, physiotherapist, and dietician) for additional expertise. If any participant missed a session, they were contacted telephonically to explain the weekly tasks. If participants missed a session and were not contactable, they were regarded as lost-to-follow-up.

The content of the intervention was guided by the findings from our situation analysis, the WHO Brochure "Support for rehabilitation: self-management after COVID-19-related illness [31] and from the HOPE intervention for people with post–Covid-19 conducted by Wright et al [21]. The detailed description of the session's content, weekly themes, evidence that informed the themes are shown in Table 2 along with the planned duration of the session and home-tasks.

XSL•FO
RenderX

Figure 1. Theory of change for CoPilot, demonstrating the post–COVID-19 intervention development plan adapted to the theory of change framework for individuals with post–COVID-19 complications.

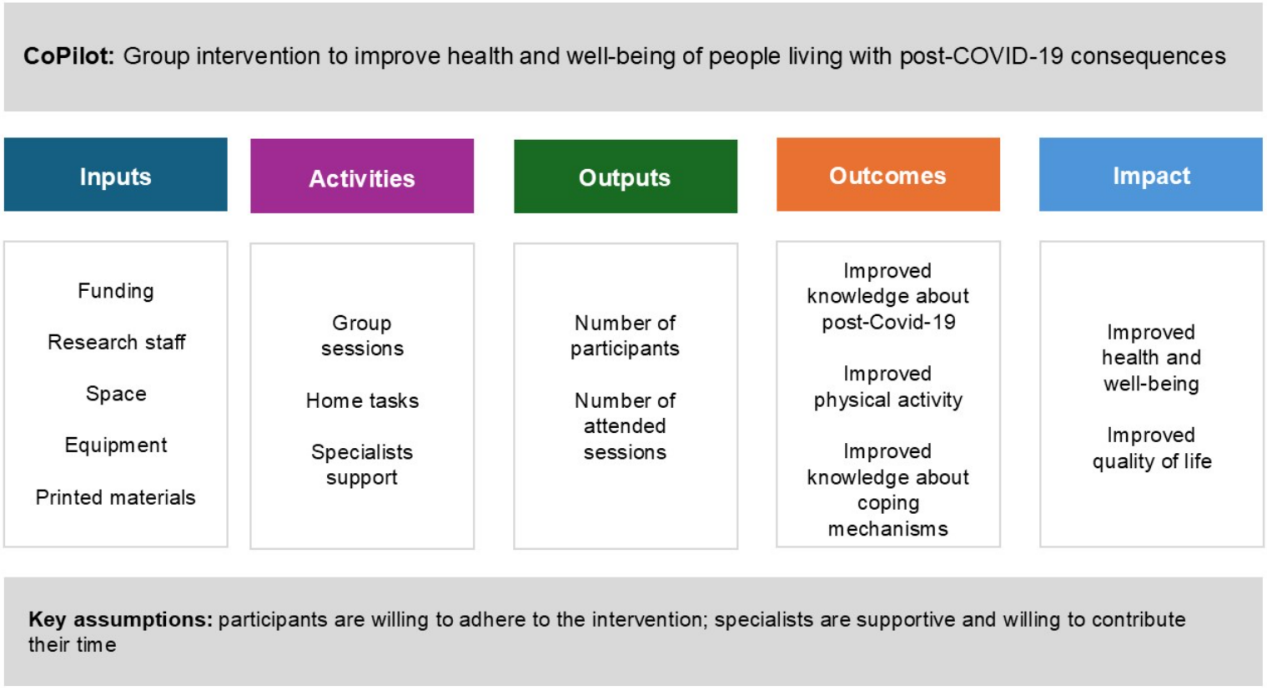

XSL•FO
RenderX

Table . Intervention modules and content.

Module	Content	Evidence-informing theme	Format: 1 hour	Home-task
Week 1: post–COVID-19 complications introduction	• Welcome and introductions. • Overview of the program. • Pacing and goal setting. • Breathing exercise introduction.	• Quantitative findings on lack of post–COVID-19 knowledge. HOPE intervention[a]. Supported by WHO[b] brochure.	• Group session facilitated by research staff – discussions, home tasks.	• Self-test: how are you feeling? • Goal setting: SMARTER tool (general)[c]. • Long–COVID-19 education. • Breathing exercise.
Week 2: symptoms and physical activity	• Main symptom complaints. • Exercise routine and demonstration. • Setting physical goals. • Share experience and symptom management. • Fatigue management.	• Quantitative findings on lack of post-COVID-19 knowledge, and goals for post-COVID-19 recovery to be able to carry out activities of daily living better and have better physical capacity. Qualitative finding on providing post-COVID-19 symptom follow-up. Supported by HOPE intervention and WHO brochure.	• Group session with physiotherapist – discussion, exercise demonstration, home tasks.	• Exercises self-moderated. • SMARTER goal setting (physical ability focus).
Week 3: coping strategies	• Prioritizing and taking breaks. • Relaxation. • Mindfulness practices to manage stress.	• Quantitative findings on request for psychological support, goal for post–COVID-19 recovery of better coping strategies. Qualitative findings on need for mental health support.	• Group session facilitated by research staff – discussions, guided meditation.	• SMARTER goal setting (relaxation focus). • Relaxation and mindfulness tasks. • Gratitude diary.
Week 4: support and grief	• Debriefing. • Grief process and tracker. • Finding support. • Self-compassion and prioritization. • Asking for and accepting help.	• Quantitative findings on request for psychological support, goal for post–COVID-19 recovery of better coping strategies, and better connection with people and family. Qualitative findings on need for mental health support.	• Group session with psychologist – discussion and home tasks.	• Grief tracker. • Speaking to family or a friend. • Self-compassion task. • Self-test: how are you feeling?
Week 5: sleep and mindfulness	• Sleep disorder information. • Tips for better sleep. • Mundane task focusing.	• Quantitative finding of goal for post–COVID-19 recovery for better coping strategies. Supported by HOPE intervention.	• Group session facilitated by research staff – discussions, home tasks.	• Sleep diary. • SMARTER goal setting (sleep). • Mundane task focusing.
Week 6: move and eat better	• Importance of keeping active. • Eating well for physical and mental health. • Continuing exercises.	• Quantitative findings on goals for post–COVID-19 recovery to be able to carry out activities of daily living better and have better physical capacity. Supported by HOPE intervention and WHO brochure.	• Group session with dietician – discussion and home tasks.	• SMARTER goal setting (nutrition). • Home exercises and stretching. • What makes you happy?

XSL•FO
RenderX

Module	Content	Evidence-informing theme	Format: 1 hour	Home-task
Week 7: happiness and strengths	• Long COVID-19 and mental health education. • Identifying strengths and how to use them. • Managing setbacks. • Reviewing, what makes you happy?	• Quantitative findings on request for psychological support, goal for post–COVID-19 recovery of better coping strategies. Qualitative findings on need for mental health support.	• Group session facilitated by research staff – discussions and home tasks.	• Gratitude diary. • SMARTER goal setting (strengths and happiness). • Long–COVID-19 and mental health education.
Week 8: looking forward	• Hopes and aspirations. • Keeping in touch with support. • Review of the program. • Planning pleasant activities.	• Quantitative findings on request for psychological support, goal for post–COVID-19 recovery of better coping strategies, and better connection with family. Qualitative findings on need for mental health support. Supported by HOPE intervention.	• Group session facilitated by research staff – discussions and home tasks.	• Self-test: how are you feeling? • SMARTER goal setting (hopes for the future). • Moving forward with hope.

[a]HOPE intervention refers to Hope Programme for Long COVID-19 [21].

[b]WHO: World Health Organization.

[c]The SMARTER goal-setting tool is used to define goals following the criteria: specific, measurable, achievable, relevant, time-bound, exciting, and reviewable.

Monitoring and Evaluation Plan

Monitoring of the intervention was done using a participant registry, discussions at each session to review understanding of the content and tasks, and completion of a questionnaire to assess adherence to tasks and subjective progress.

The evaluation of the intervention was performed by measuring effectiveness (pre-and postintervention assessment) and a process evaluation (including feasibility and acceptability) of the group intervention. The main effectiveness outcomes included: (1) lung function, (2) mental well-being, and (3) quality of life. Data collection included baseline information on demographics and knowledge of COVID-19 and sequelae. The pre-post-intervention assessments included questionnaires on COVID-19 symptoms, fatigue and sleepiness scales [21,32,33], quality of life using the EuroQoL 5-Dimension 5-Level (EQ-5D-5L) [34], short Warwick-Edinburgh Mental Wellbeing Scale to describe mental well-being [35], lung function testing by means of spirometry, and functional capacity by 1-minute sit-to-stand test [36]. The process evaluation was guided by the Medical Research Council framework to assess both the implementation process (how the project is delivered), what is delivered (fidelity, dose, adaptations, or reach) and contextual factors influencing the implementation and acceptability of the intervention [37]. We used mixed methods to assess feasibility and acceptability of the intervention through a questionnaire and semistructured interviews with participants to elaborate on the design of the intervention, the potential barriers, and facilitators. Data from the intervention will be analyzed using descriptive statistics producing frequencies, means, and proportions of the different variables of interest. Outcomes will be analyzed using paired sample *t* tests. Qualitative process evaluation data will be transcribed and analyzed by thematic analysis.

Discussion

Principal Findings

To the best of our knowledge, this is the first study to develop and implement a context-adapted group intervention for post–COVID-19 survivors in South Africa. It was aimed to address debilitating consequences of COVID-19 and provide people with the tools to manage their symptoms to improve every-day quality of life and potentially health outcomes including mental and lung health. This situation analysis and intervention development highlighted 2 main points: first, engaging with patients and health care workers on their perspectives on post–COVID-19 care is an integral component in developing management strategies for post–COVID-19 complications, and second, that mental health support should strongly be considered in rehabilitation guidelines post–COVID-19 while considering support groups as a component.

Engaging with health care providers and patients' perspectives on post–COVID-19 complications and the management thereof is a crucial step in guideline development. The cornerstone to this is understanding the patient experience of post–COVID-19 complications and leveraging on existing and innovative strategies for care [21,25]. Fatigue is a common symptom reported post–COVID-19; however, it has been described by patients to have physical, mental, and emotional components [6,21], thereby illustrating that understanding the problem on a deeper level can help direct health services toward patient needs, as well as develop suitable assessment measures for

XSL•FO
RenderX

patient-reported outcomes [9]. Understanding patient and health care provider perspectives and communication between them has been shown to be important in HIV care, specifically relating to social, mental health, and health care system challenges [38]. In addition, coworking and partnership with shared responsibility in care between health care provider and patient with a focus on self-management principles were recognized as key features in the management process of other chronic lung diseases [24]. Taking into account the health resource limitations in LMICs and the strain on insufficient rehabilitation services poses another challenge to the management of post–COVID-19 complications, emphasizing the need for context-adapted guidelines [39].

The importance of the need for mental health support was a major theme that came through during our situation analysis. Mental health symptoms have been reported in up to 21% of people post–COVID-19 and can include anxiety, depression, sleep disorders, and neurocognitive disorders [4,6]. In a study by Rofail et al [9], it was reported that despite no clear psychological diagnosis, most impacts on daily life were not attributed to any singular symptom, but rather to the holistic experience of COVID-19 infection, further supporting the need to support psychological processing in recovery. Findings from our situation analysis clearly showed the gaps in psychological support for patients and health care providers post–COVID-19 and that these services are sought after as an adjunct to post–COVID-19 care. More specifically, support groups were suggested as helpful platforms to engage in psychological support, facilitating connection and shared experiences. This focus in care has been shown to be valued and helpful in support of patients with other chronic lung conditions [24].

One limitation is that this was a narrow population, where participants were only recruited from 2 facilities; however, to overcome this we recruited from multiple departments and recruited both health care providers and patients in order to increase variation. Another limitation was the heterogeneity of the post–COVID-19 sequelae definition, where interpretation conducted by the study team was to meet the WHO post–COVID-19 condition definition. This could result in bias of either overestimating or underestimating post–COVID-19 consequences, depending on alternate definitions used in other contexts. Despite this, consistency was maintained throughout recruitment for the situation analysis and planned intervention, and symptoms were clearly documented.

The strengths of the study include the context and population of the study where post–COVID-19 rehabilitation has not been studied previously. Another strength includes the mixed-methods approach for both the situation analysis and intervention in order to obtain deeper insight into the rehabilitation post–COVID-19.

Conclusion

Understanding the characteristics and burden of post–COVID-19 complications, particularly from the patient and health care provider perspectives, helps to direct post–COVID-19 management principles to ensure a more holistic recovery. As more people recover from COVID-19, it becomes crucial to maximize on the opportunity to better understand postviral complications and adapted health service allocation, particularly in LMICs. We believe that our group intervention holds a potential to improve health and well-being of people living with post–COVID-19 complications in Johannesburg, and its effectiveness and feasibility will be analyzed in the next steps.

Acknowledgments

We would like to acknowledge Ency Mahlase, an additional member of the research team, who assisted with analyzing the data. We are grateful to the Bavarian State Ministry of Science and Arts for their funding of this project. We thank Chris Hani Baragwanath Academic Hospital and Tembisa Tertiary Hospital for welcoming us to work at their site and supporting us in our participant recruitment and follow-up. We thank all the participants for their time and support. Generative artificial intelligence was not used in any part of the development of this manuscript.

Data Availability

The datasets generated during or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

NAG and OI contributed to the literature search. OI, FS, AMM, SC, and AR contributed to the study funding acquisition. NAG, OI, FS, SC, SAB, and AR contributed to the conceptualization, methodology, and project administration. NAG, OI, AR, ST, PM, and NM contributed to the data analysis and interpretation. NAG, OI, and FS contributed to the intervention design. NAG and OI drafted the manuscript, and all authors reviewed the manuscript critically for intellectual content. All coauthors approved the final draft of the manuscript for publication.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Questionnaire guide.
[DOCX File, 35 KB - formative_v9i1e58436_app1.docx]

XSL•FO
RenderX

Multimedia Appendix 2
Focus group discussion guide.
[DOCX File, 17 KB - formative_v9i1e58436_app2.docx]

Multimedia Appendix 3
Clinical staff interview guide.
[DOCX File, 19 KB - formative_v9i1e58436_app3.docx]

References

1. Groff D, Sun A, Ssentongo AE, et al. Short-term and long-term rates of postacute sequelae of SARS-CoV-2 infection: a systematic review. JAMA Netw Open 2021 Oct 1;4(10):e2128568. [doi: 10.1001/jamanetworkopen.2021.28568] [Medline: 34643720]
2. O'Mahoney LL, Routen A, Gillies C, et al. The prevalence and long-term health effects of long COVID among hospitalised and non-hospitalised populations: a systematic review and meta-analysis. EClinicalMedicine 2023 Jan;55:101762. [doi: 10.1016/j.eclinm.2022.101762] [Medline: 36474804]
3. A clinical case definition of post COVID-19 condition by a Delphi consensus. World Health Organization. URL: https://iris.who.int/bitstream/handle/10665/345824/WHO-2019-nCoV-Post-COVID-19-condition-Clinical-case-definition-2021.1-eng.pdf?sequence=1&isAllowed=y [accessed 2023-08-22]
4. Lopez-Leon S, Wegman-Ostrosky T, Perelman C, et al. More than 50 long-term effects of COVID-19: a systematic review and meta-analysis. Sci Rep 2021 Aug 9;11(1):16144. [doi: 10.1038/s41598-021-95565-8] [Medline: 34373540]
5. Global Burden of Disease Long COVID Collaborators, Wulf Hanson S, Abbafati C, et al. Estimated global proportions of individuals with persistent fatigue, cognitive, and respiratory symptom clusters following symptomatic COVID-19 in 2020 and 2021. JAMA 2022 Oct 25;328(16):1604-1615. [doi: 10.1001/jama.2022.18931] [Medline: 36215063]
6. Natarajan A, Shetty A, Delanerolle G, et al. A systematic review and meta-analysis of long COVID symptoms. Syst Rev 2023 May 27;12(1):88. [doi: 10.1186/s13643-023-02250-0] [Medline: 37245047]
7. Anaya JM, Rojas M, Salinas ML, et al. Post-COVID syndrome. A case series and comprehensive review. Autoimmun Rev 2021 Nov;20(11):102947. [doi: 10.1016/j.autrev.2021.102947] [Medline: 34509649]
8. Parotto M, Gyöngyösi M, Howe K, et al. Post-acute sequelae of COVID-19: understanding and addressing the burden of multisystem manifestations. Lancet Respir Med 2023 Aug;11(8):739-754. [doi: 10.1016/S2213-2600(23)00239-4] [Medline: 37475125]
9. Rofail D, McGale N, Podolanczuk AJ, et al. Patient experience of symptoms and impacts of COVID-19: a qualitative investigation with symptomatic outpatients. BMJ Open 2022 May 2;12(5):e055989. [doi: 10.1136/bmjopen-2021-055989] [Medline: 35501077]
10. Tran VT, Riveros C, Clepier B, et al. Development and validation of the long coronavirus disease (COVID) symptom and impact tools: a set of patient-reported instruments constructed from patients' lived experience. Clin Infect Dis 2022 Jan 29;74(2):278-287. [doi: 10.1093/cid/ciab352] [Medline: 33912905]
11. Dryden M, Mudara C, Vika C, et al. Post-COVID-19 condition 3 months after hospitalisation with SARS-CoV-2 in South Africa: a prospective cohort study. Lancet Glob Health 2022 Sep;10(9):e1247-e1256. [doi: 10.1016/S2214-109X(22)00286-8] [Medline: 35961348]
12. Jassat W, Cohen C, Tempia S, et al. Risk factors for COVID-19-related in-hospital mortality in a high HIV and tuberculosis prevalence setting in South Africa: a cohort study. Lancet HIV 2021 Sep;8(9):e554-e567. [doi: 10.1016/S2352-3018(21)00151-X] [Medline: 34363789]
13. Glover N, Sathar F, Charalambous S, Ivanova O, Riess F, Shambhu RR, et al. Respiratory outcomes post-hospitalisation for symptomatic covid-19 disease in patients from a tertiary facility, South Africa. Presented at: Proceedings of the 7th South African TB Conference; Sep 13-16, 2022; Durban, South Africa.
14. Galal I, Hussein A, Amin MT, et al. Determinants of persistent post-COVID-19 symptoms: value of a novel COVID-19 symptom score. Egypt J Bronchol 2021 Dec;15(1). [doi: 10.1186/s43168-020-00049-4]
15. Munblit D, Nicholson T, Akrami A, et al. A core outcome set for post-COVID-19 condition in adults for use in clinical practice and research: an international Delphi consensus study. Lancet Respir Med 2022 Jul;10(7):715-724. [doi: 10.1016/S2213-2600(22)00169-2] [Medline: 35714658]
16. Alwan NA. Surveillance is underestimating the burden of the COVID-19 pandemic. Lancet 2020 Sep 5;396(10252):e24. [doi: 10.1016/S0140-6736(20)31823-7] [Medline: 32861312]
17. Demeco A, Marotta N, Barletta M, et al. Rehabilitation of patients post-COVID-19 infection: A literature review. J Int Med Res 2020 Aug;48(8):300060520948382. [doi: 10.1177/0300060520948382] [Medline: 32840156]
18. COVID-19 rapid guideline: managing the long-term effects of COVID-19. National Institute for Health and Care Excellence. 2024. URL: https://www.nice.org.uk/guidance/ng188/chapter/5-Management [accessed 2024-02-08]
19. Liu K, Zhang W, Yang Y, Zhang J, Li Y, Chen Y. Respiratory rehabilitation in elderly patients with COVID-19: a randomized controlled study. Complement Ther Clin Pract 2020 May;39:101166. [doi: 10.1016/j.ctcp.2020.101166] [Medline: 32379637]

XSL•FO
RenderX

20. Nambi G, Abdelbasset WK, Alrawaili SM, et al. Comparative effectiveness study of low versus high-intensity aerobic training with resistance training in community-dwelling older men with post-COVID 19 sarcopenia: a randomized controlled trial. Clin Rehabil 2022 Jan;36(1):59-68. [doi: 10.1177/02692155211036956] [Medline: 34344230]
21. Wright H, Turner A, Ennis S, et al. Digital peer-supported self-management intervention codesigned by people with long COVID: mixed methods proof-of-concept study. JMIR Form Res 2022 Oct 14;6(10):e41410. [doi: 10.2196/41410] [Medline: 36166651]
22. Oates GR, Niranjan SJ, Ott C, et al. Adherence to pulmonary rehabilitation in COPD: a qualitative exploration of patient perspectives on barriers and facilitators. J Cardiopulm Rehabil Prev 2019 Sep;39(5):344-349. [doi: 10.1097/HCR.0000000000000436] [Medline: 31348127]
23. Blackstock FC, Evans RA. Rehabilitation in lung diseases: “Education” component of pulmonary rehabilitation. Respirology 2019 Sep;24(9):863-870. [doi: 10.1111/resp.13582] [Medline: 31099164]
24. Kelly CA, Tsang A, Lynes D, Spencer S. “It’s not one size fits all”: a qualitative study of patients’ and healthcare professionals’ views of self-management for bronchiectasis. BMJ Open Respir Res 2021 Mar;8(1):e000862. [doi: 10.1136/bmjresp-2020-000862] [Medline: 33664124]
25. Jassat W, Reyes LF, Munblit D, et al. Long COVID in low-income and middle-income countries: the hidden public health crisis. Lancet 2023 Sep 30;402(10408):1115-1117. [doi: 10.1016/S0140-6736(23)01685-9] [Medline: 37652071]
26. Jassat W, Abdool Karim SS, Mudara C, et al. Clinical severity of COVID-19 in patients admitted to hospital during the omicron wave in South Africa: a retrospective observational study. Lancet Glob Health 2022 Jul;10(7):e961-e969. [doi: 10.1016/S2214-109X(22)00114-0] [Medline: 35597249]
27. Hoddinott P, Allan K, Avenell A, Britten J. Group interventions to improve health outcomes: a framework for their design and delivery. BMC Public Health 2010 Dec 31;10:800. [doi: 10.1186/1471-2458-10-800] [Medline: 21194466]
28. Wood AM, Maltby J, Gillett R, Linley PA, Joseph S. The role of gratitude in the development of social support, stress, and depression: two longitudinal studies. J Res Pers 2008 Aug;42(4):854-871. [doi: 10.1016/j.jrp.2007.11.003]
29. Larsen DJ, King RL, Stege R, Egeli NA. Hope in a strengths-based group activity for individuals with chronic pain. Couns Psychol Q 2015 Apr 3;28(2):175-199. [doi: 10.1080/09515070.2015.1007444]
30. Teresi JA, Yu X, Stewart AL, Hays RD. Guidelines for Designing and Evaluating Feasibility Pilot Studies. Med Care 2022 Jan 1;60(1):95-103. [doi: 10.1097/MLR.0000000000001664] [Medline: 34812790]
31. Support for rehabilitation self-management after COVID-19-related illness. : World Health Organization Regional Office for Europe; 2021 URL: https://www.who.int/publications/m/item/support-for-rehabilitation-self-management-after-covid-19-related-illness [accessed 2023-03-01]
32. Shahid A, Wilkinson K, Marcu S, Shapiro CM. Fatigue severity scale (FSS). In: In: STOP, THAT and One Hundred Other Sleep Scales: Springer; 2011:167-168. [doi: 10.1007/978-1-4419-9893-4_35]
33. Johns M. About the ESS. The Epworth Sleepiness Scale. URL: https://epworthsleepinessscale.com/about-the-ess/ [accessed 2023-11-09]
34. EQ-5D-5L. EuroQoL. 2024. URL: https://euroqol.org/information-and-support/euroqol-instruments/eq-5d-5l/ [accessed 2023-06-10]
35. The Warwick-Edinburgh Mental Wellbeing Scales - WEMWBS. Warwick Medical School. 2021. URL: https://warwick.ac.uk/fac/sci/med/research/platform/wemwbs [accessed 2023-06-20]
36. Vaidya T, Chambellan A, de Bisschop C. Sit-to-stand tests for COPD: a literature review. Respir Med 2017 Jul;128:70-77. [doi: 10.1016/j.rmed.2017.05.003] [Medline: 28610673]
37. Moore GF, Audrey S, Barker M, et al. Process evaluation of complex interventions: Medical Research Council guidance. BMJ 2015 Mar 19;350:h1258. [doi: 10.1136/bmj.h1258] [Medline: 25791983]
38. Bogart LM, Chetty S, Giddy J, et al. Barriers to care among people living with HIV in South Africa: contrasts between patient and healthcare provider perspectives. AIDS Care 2013;25(7):843-853. [doi: 10.1080/09540121.2012.729808] [Medline: 23061894]
39. Ngeh EN, Chigbo NN, Whitehouse Z, et al. A report on the development of COVID-19 guidelines for rehabilitation professionals in African settings. Pan Afr Med J 2021;38:129. [doi: 10.11604/pamj.2021.38.129.26311] [Medline: 33912299]

Abbreviations

ADL: activities of daily living
CHBAH: Chris Hani Baragwanath Academic Hospital
EQ-5D-5L: EuroQoL 5-Dimension 5-Level
FGD: focus group discussion
LMIC: low- and middle-income country
NICE: National Institute for Health and Care Excellence
TB: tuberculosis
WHO: World Health Organization

XSL•FO
RenderX

Edited by A Mavragani; submitted 16.03.24; peer-reviewed by A AL-Asadi, J Schulze, S Markham; revised version received 17.02.25; accepted 02.03.25; published 10.04.25.

Please cite as:
Glover NA, Sathar F, Mokome P, Mathabela N, Taleni S, van Blydenstein SA, Mekota AM, Charalambous S, Rachow A, Ivanova O Improving Health and Well-Being of People With Post–COVID-19 Consequences in South Africa: Situation Analysis and Pilot Intervention Design
JMIR Form Res 2025;9:e58436
URL: https://formative.jmir.org/2025/1/e58436
doi:10.2196/58436

© Nicole Audrey Glover, Farzana Sathar, Pride Mokome, Nkululeko Mathabela, Sipokazi Taleni, Sarah Alexandra van Blydenstein, Anna-Maria Mekota, Salome Charalambous, Andrea Rachow, Olena Ivanova. Originally published in JMIR Formative Research (https://formative.jmir.org), 10.4.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Designing and Evaluating a Prototype of a Trilingual Data Collection Tool for the Middle East and North Africa (MENA) Region to Collect Data About Violence Against Sex Workers: Multiple Methods Approach in User-Centered Design

Melissa Ditmore[1], PhD; Amal El Karouaoui[2], MPH; Jose Fernando Florez-Arango[3], PhD

[1]Clear-Eyed Consulting, P.O.Box 20853, New York, NY, United States

[2]Nawara Women's Network for the Middle East and North Africa, Casablanca, Morocco

[3]Department of Population Health Sciences, Weill Cornell Medicine, New York, NY, United States

Corresponding Author:
Melissa Ditmore, PhD
Clear-Eyed Consulting, P.O.Box 20853, New York, NY, United States

Abstract

Background: Sex workers face an epidemic of violence around the world. However, violence against sex workers (VASW) is underreported, and sex workers hesitate to report to the police because they are frequently punished; therefore, an alternative for reporting is needed. Sex workers also face stigmatization from health care workers, further discouraging help-seeking behavior or reporting in health care facilities. Sex workers have been recipients of services from nongovernmental organizations, typically related to HIV and sexual transmission of infections, but violence remains underaddressed.

Objective: This study aims to apply human-centered design methods to adapt ReportVASW for use in the Middle East and North Africa and to evaluate the usability of the prototype interface and identify opportunities for improvement.

Methods: Evaluation methods included cognitive walkthrough and System Usability Scale by 9 potential end users, and heuristic evaluation by 2 informatics professionals and 2 service providers.

Results: This study explores ways to improve the trilingual prototype of ReportVASW, with particular attention to ways to improve the data collection tool. Multiple methods identified multiple issues to address. Heuristic analysis revealed 2 serious issues to address, with scores over 2.5 out of 4 in the tool used. The most serious problems identified in heuristic analysis were related to language, particularly the Arabic version. Translation issues were addressed before end user testing. End users were enthusiastic about the idea of a mobile tool to document VASW, provided it led to change. They gave ReportVASW a System Usability Scale of 91.4, above the 68 considered good. Even as end users were enthusiastic, they offered suggestions for improvement.

Conclusions: Many opportunities to improve the interface were identified. Most changes are not overly complex, and the majority involve adapting the language used and improving the translation. Development of the trilingual ReportVASW data collection tool for the Middle East and North Africa region is worth pursuing.

(JMIR Form Res 2025;9:e65210) doi:10.2196/65210

KEYWORDS

mobile health; sex work; user-centered design methods; usability; heuristic analysis; cognitive walkthrough; mHealth; data collection tool; violence; sex worker; stigmatization; health care workers; help-seeking behavior; human-centered design methods; heuristic evaluation; System Usability Scale; abusive; sexual violence; Middle East and North Africa; MENA

Introduction

Sex workers are frequently victimized by a variety of perpetrators [1] and 32% to 55% report at least one violent experience in the previous year [2], with numbers rising to near-universal in some locations [3]. Sex workers often do not report violent incidents [2,4]. Sex workers in Morocco “fear police too much to report violence.” [5] In fact, widespread criminalization in the Middle East and North Africa (MENA) region means that police may arrest them for sex work [5] instead of investigating violent crimes against sex workers; sex workers have been victimized or harassed by police [4-7]. Sex workers who seek care after victimization may not reveal their status as sex workers because health care providers may stigmatize and discriminate against them [8,9]. These factors culminate in an unfortunate lack of knowledge in the domain, including in the MENA region. Documenting violence against sex workers (VASWs) will help develop more effective

XSL•FO
RenderX

responses to the health consequences of VASWs, including posttraumatic stress disorder [10] and HIV and other sexually transmitted infections [1,11]. Sex workers share information about perpetrators of violence among themselves [8], separate from reporting to law enforcement, in order to share information about violent experiences and to help others avoid their attackers.

While VASWs have been studied, few efforts to address this violence have been evaluated [12]. An exception is work by the Asia Pacific Network of Sex Workers, which commissioned an app called iMonitor+ that was designed for sex workers in Myanmar to help them report violence committed against them to an organization working with sex workers; the program was successful [13]. Other apps have been developed for sex workers in Cambodia [14] and South Africa [15]. No such app exists for use in the MENA region.

The lead author developed a prototype of a mobile health (mHealth) tool to collect data about incidents of VASW for use in the United States in line with the principles of user-centered design (UCD) for sex workers to report violence committed against them [16]. She then adapted this prototype for a trilingual data collection tool for the MENA region in partnership with Nawara Women's Network for the MENA. Nawara engaged in a strategic planning process in 2022 [17] during which gender-based violence (GBV) was identified as an area of focus.

UCD was implemented in the adaptation for the MENA region. UCD with multiple methods, such as considering both end user and design use context, has been successfully used in mHealth [18,19], though it presents some challenges [20,21]. Formative research using UCD can be time-consuming, and users are not always easy to engage, but the literature shows the value of engaging end users in formative processes [20,21].

Methods

Overview

This section of the paper details the evaluation methods undertaken in usability testing with a multidisciplinary team of informatics professionals, Nawara Secretariat, and 10 potential end users based on the known theory that this number will generally expose the majority of problems with usability [22], in our attempt to evaluate the usability of the trilingual prototype of ReportVASW.

Design Process and Prototype

Overview

The authors used UCD to design this prototype and its predecessor [16]. Ease of use is key because the user context is in the aftermath of a violent and possibly traumatic event. Ease of use and the lack of a learning curve are hallmarks of UCD. UCD is meant to improve usability by maximizing effectiveness, efficiency, and satisfaction of end users in the specific aim of the product in question. Early-stage implementation of UCD methods should reduce user error and limit the cost and time spent redesigning after developing software.

The specific principles of UCD include focusing on users and tasks, measuring usability empirically, and iterative testing of design and usability. UCD has been used with success to develop mHealth apps and health record systems [23,24]. These principles informed the specific methods used at each step. We followed the methodological steps of the UCD approach in our design process; these steps align with more specific methods: for example, contextualization using functional analysis and consulting potential end users; ideation through task analysis focused on the steps required for the end user to successfully complete the task; prototyping using representational analysis of the tool; and finally, usability testing using scenario-based testing and heuristic analysis.

Step 1: Contextualization

The authors seek to adapt a tool to respond to VASWs in MENA.

Contextualization included conducting a literature review about GBV in the MENA region, including the Nawara strategic plan [17]. The Steering Committee expressed interest in and support for adapting the existing data collection tool [16] for use in the MENA region.

A semistructured interview was conducted with 10 Nawara members to ask whether such a tool would be useful; members in Morocco and Sudan expressed great interest, and members from other countries were also interested, especially if the tool could be adapted for wider use after working with sex workers for GBV more generally.

Step 2: Ideation

We had an English-language prototype that was evaluated using task analysis and heuristic analysis [16,25]. Adaptation for the MENA region required assessing and refining the original language prototype, based on the increased cultural understanding gained from the literature review and consultation, followed by translation into the other 2 most common regional languages, French and Arabic. Members of the community from Sudan provided the initial Arabic translation, which was later reviewed by one member of the Nawara secretariat. French was initially undertaken by the lead author and revised by 2 members of the Nawara secretariat. The Nawara secretariat was particularly helpful because they communicate with members from across the entire Arab-speaking world, from as far east as the Persian Gulf and as far west as Mauritania and throughout the diaspora.

Step 3: Prototyping

The lead author developed the prototype in Qualtrics (Qualtrics International Inc; 29 New Mexico, 2009) web-based survey software, and created 3 versions—English, Arabic, and French—in March 2022, incorporating the Nawara logo in April 2022. Nawara members expressed a desire to have one prototype in all 3 languages together. The lead author then developed a fourth version, including all 3 languages, in April 2022.

Step 4: Representational Analysis

Representational analysis consisted of heuristic analysis undertaken by 3 authors and one colleague (2 informatics professionals and 2 colleagues from the Nawara Secretariat) in

XSL•FO
RenderX

April 2023. The form used is a spreadsheet developed by the final author [26], adapted from another spreadsheet tool [27], in which a scale of 0 - 4 is used to grade each issue, from minor (1) to catastrophic (4), and 0 is used to indicate disagreement that the point is an issue. The spreadsheet also contains a column labeled "proposed solution."

The Nawara Secretariat members identified multiple language problems, especially in the Arabic version, and refined the language to be simpler for people with little education to understand and for Arabic to be understood across the region, followed by adjustments to the English version to remain in line with the French and Arabic versions. The prototype was updated to reflect these changes before the cognitive walkthrough and usability testing.

Step 5: Cognitive Walkthrough

Additionally, the Nawara Secretariat recruited a convenience sample of 9 end users who self-identify as female, both assigned female at birth and transgender women, with experience in a variety of sex work venues (internet escorts, brothel workers, open-area sex workers—cafes, streets, parks, etc) to test the usability by entering data from scenarios provided (scenario based testing) in a cognitive walkthrough [22], using the proposed application to enter data from a scenario taken from interviews with sex workers; all 10 agreed to do the walkthrough remotely. Multiple meetings with stakeholders were held in order to reach a broad range of stakeholders by gender, education level, type of sex work, and venue.

All end users recruited did the cognitive walkthrough in April 2024, in 3 sessions. Three scenarios used were taken from the lead author's research [28-30], and 2 scenarios were written by Nawara based on MENA sex workers' web-based experiences; all are in the Multimedia Appendix 1.

Step 6: Usability Testing

The end users recruited for the scenario-based testing then rated the data collection tool using the System Usability Scale (SUS). The SUS is a 10-item Likert scale, with each item's score ranging from 0 to 4. The scale items were read aloud. Odd-numbered items are scored at the scale position minus 1, and even-numbered items are scored at 5 minus the scale position; the sum of the scores is then multiplied by 2.5 to obtain the overall score [31]. We searched the literature for an Arabic-language SUS but learned that translation had not been found to be acceptable despite translation of the SUS into other languages being acceptable and useful [32]. Considering this, Nawara translated the scale.

Ethical Considerations

This protocol was submitted first to the CUNY School of Public Health and Health Policy Human Research Protection Program in December 2022 and to the Weill Cornell School of Medicine Institutional Review Board (24-07027685) in March 2023 and was judged exempt at both institutions. Nawara Women's Network for the MENA recruited potential end users, specifically sex workers who identify as female, who were known to them and deemed unlikely to be adversely affected by the task of entering information about a scenario of VASWs. Informed consent was obtained verbally before beginning the cognitive walkthrough. Nawara was prepared to offer emotional support in the event of any adverse effects, and this was not necessary. Lunch and money for round-trip travel (averaging 100 Moroccan dirhams or approximately US $9) were provided for the 9 in-person testers. Generative artificial intelligence was not used in any portion of the manuscript writing.

Results

This section of this paper conveys the results of the evaluation and will be organized following the same 6 steps used in the methods section.

Step 1: Contextualization

The literature revealed that GBV is a problem throughout the MENA region at rates comparable to or higher than in other regions of the world [33,34], and that GBV and legal and social restrictions on women limit the economic development of countries in the MENA region [33]. Furthermore, economic constraints on women create vulnerability in other ways, including pushing women into informal sectors such as sex work. Finally, sex workers are singled out for interpersonal violence [5] by multiple actors, including law enforcement, family, and the public. All of these factors are compounded for sex workers [5,11], who typically come from the most vulnerable groups.

Nawara's strategic plan [17] recommended addressing GBV for the most vulnerable women, including sex workers, migrants, and transgender women.

During the consultation, members of Nawara's Steering Committee were enthusiastic about ReportVASW, and Nawara staff and leadership confirmed that information collected using ReportVASW could be used in multiple ways:

1. This site may facilitate information sharing in addition to documenting violence committed against sex workers and generate evidence to be used in reporting violent incidents in which sex workers are victimized.
2. Geolocation data about violent incidents can be used in the allocation of resources by organizations that work with sex workers, and in advocacy for additional resources.

Step 2: Ideation

After Nawara Steering Committee members confirmed that ReportVASW could be useful for people in the MENA region, Nawara undertook Arabic and French translations and edits of the original English and French text, which were then used in the creation of 2 more prototypes, in Arabic and French. The wording of the introductory screen was modified both to be understood throughout the Arabic-speaking world and also to cater to regional sensibilities in phrasing.

Potential end users and cybersecurity experts agreed that using a website to collect data was more desirable than an app because it would be less likely to be noticed by police if their mobile phones were under surveillance.

XSL•FO
RenderX

Step 3: Prototyping

The heuristic analysis and cognitive walkthrough using scenario-based testing methods generated similar assessments about the ease of use of the prototype and that the aesthetics were good; this overall agreement would seem to indicate that the findings reflect the actual usability of the prototype ReportVASW (Figure 1).

Figure 1. A prototype of ReportVASW presented separately in Arabic, English, and French.

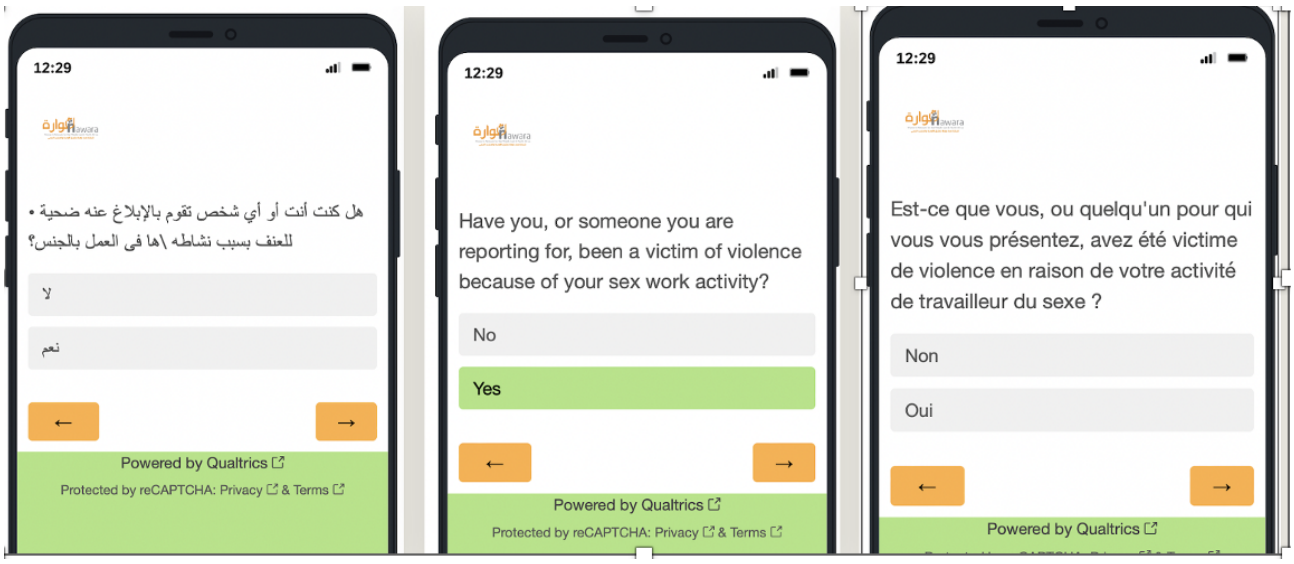

Step 4: Representational Analysis

Responses to the trilingual version (Figure 2) were compiled and average scores computed; heuristic analysis scores ranged from 0.5 to 3 and revealed that most problems were linguistic.

Two severe problems (with scores over 2.5) were identified. The first was to correct the name of the organization in Arabic on the first screen, violating the heuristic of visibility (average score 3). The second suggested problem and resolution was to add an Arabic label to the arrows because they are not the right direction for readers of Arabic (heuristic violation: Error prevention/Match/Language), with an average score of 2.5.

The number of violations of each heuristic listed in the Heuristic Analysis chart was counted. The most frequently violated heuristic was language (match the user's reality) (n=8), stemming from multiple translations. Language was fine-tuned in Arabic, and then English and French were adjusted to match. These corrections were completed before end user testing (step 5) began.

XSL•FO
RenderX

Figure 2. Final trilingual prototype of ReportVASW tested by end users (sex workers in Morocco).

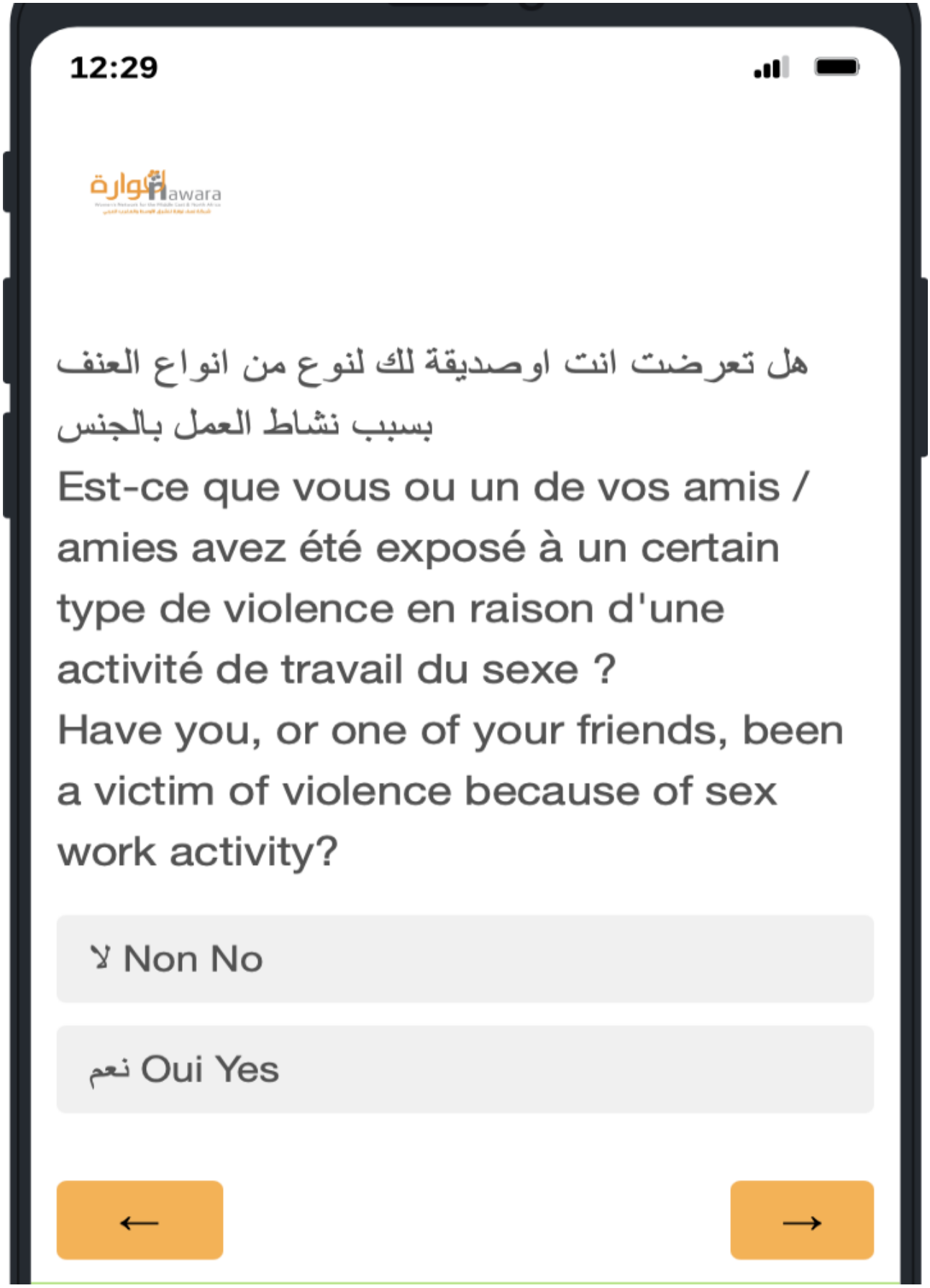

XSL•FO
RenderX

Step 5: Cognitive Walkthrough

Nine end users completed the evaluation, completing the task in under 2 minutes. End users confirmed the usefulness of an application like ReportVASW and said that it could be a viable tool for collecting information about violence committed against sex workers.

End users additionally confirmed that receiving a list of service providers and organizations offering services to sex workers would be helpful. This would need to be determined by location; Nawara has data for Morocco; this could be organized by city.

Some end users expressed a desire to share data about perpetrators, to prevent future exposure to violence, and one expressed a need to see the results of assistance, including help accessing services and possibly eventually seeking justice.

The cognitive walkthrough process delivered positive feedback, and users offered many ideas for improvements. Suggestions included the collection of additional information and comments about how they would benefit from this tool (eg, “If we had this five years ago, we could have prevented a lot of violence”), and what would be necessary to motivate them to use ReportVASW (eg, friendlier help from service providers in the aftermath of violent incidents).

End user testers of the prototype (Figure 1) recommended changes to data collection and follow-up as shown in Textbox 1.

Textbox 1. Changes to data collection and follow-up.

Changes to data collection:

- Add branching questions to follow negative responses to questions about reporting to the police to ask why people did not report to the police.
- Add branching questions to follow negative responses to questions about seeking health care to ask why people did not seek medical care, and include “stigma” among the choices.

Changes to follow-up:

- End users suggested following up to use the data collected for advocacy, including with police and health care personnel.
- End users saw the possibility of sharing information about violent incidents with sex workers in order to avoid perpetrators of violence.
- End users would like to see the ReportVASW function as a “warmline” with direct follow-up.

Step 6: Usability Testing

ReportVASW scored 91.4 on the system usability; SUS scores over 68 are considered good [31] (Table 1). The tool scored neutral on consistency inspection, and all other responses were positive toward the application, with most being strong.

XSL•FO
RenderX

Table . System Usability Scale chart with number of end users (sex workers in Morocco) reporting each response.

	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
I think that I would like to use this system frequently	7	2	0	0	0
I found the system unnecessarily complex	0	0	0	5	4
I thought the system was easy to use	4	5	0	0	0
I think that I would need the support of a technical person to be able to use this system	0	0	0	8	1
I found the various functions in this system were well integrated	0	9	0	0	0
I thought there was too much inconsistency in this system	0	0	0	0	9
I would imagine that most people would learn to use this system very quickly	6	2	1	0	0
I found the system very cumbersome to use	0	0	0	0	9
I felt very confident using the system	8	1	0	0	0
I needed to learn a lot of things before I could get going with this system	0	0	0	0	9

Discussion

Principal Findings

Using multiple methods (Figure 3 contains a flowchart of the methods used) to evaluate the prototype enabled the collection of new information, including phrasing for screening questions, and positive reception of ReportVASW. New knowledge was gained from the evaluation, particularly through engagement with end users. We believe we have sufficient input and information to proceed to a pilot study.

XSL•FO
RenderX

Figure 3. Flowchart of the process with input and output at each step. SUS: System Usability Scale.

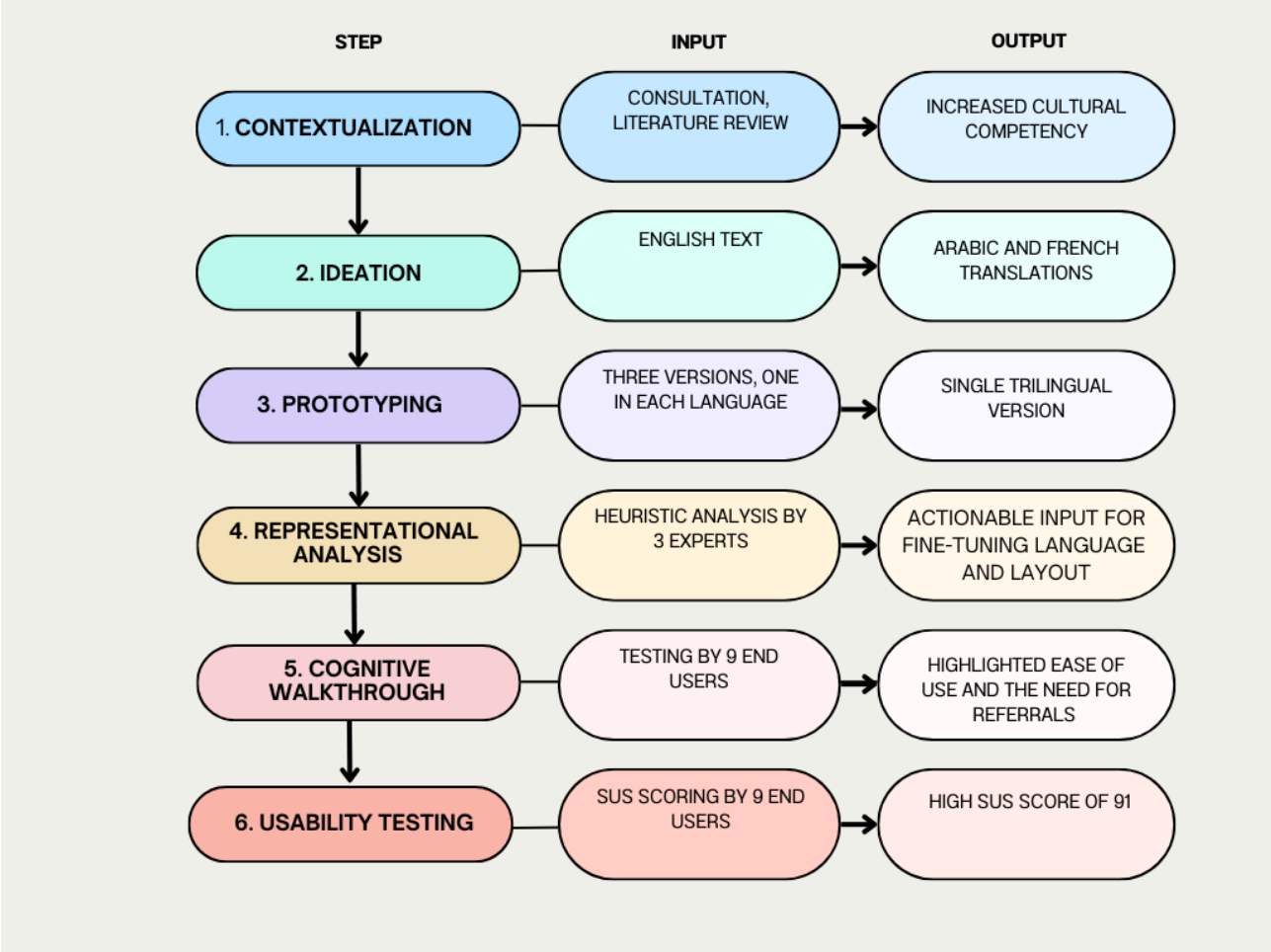

Our experience reflects the literature [20,21] regarding the usefulness of multiple methods, despite challenges. Using multiple UCD methods at each stage of the process gave us richer information than we would have had with only one method at each stage, as others have described [18,19]. The time invested was significantly more than it would have taken with one method only, but the benefits are great because the information gathered will direct our next steps and reduce the chances of needing more changes later. The value of using multiple methods in UCD was clearly demonstrated in the process of designing and evaluating a prototype data collection to submit information about incidents of VASWs. Using multiple methods in the initial steps of contextualization and ideation led to multiple revisions in these early stages. Using multiple methods in prototyping and evaluating the prototype afforded the opportunity to collect informative input from people in different roles, including end users and informatics professionals. Results from each method aligned such that the representational analysis, consistency inspection, and heuristic analysis reinforced ways to improve the prototype, reinforcing the input from each source.

Partnering with Nawara enabled the adaptation and the discovery of the need for increased data literacy about digital personal security, including secure applications and browser history. This is in line with other literature that sex workers have adopted digital technology to address occupational safety and health [35] and shared information with each other [36] about benefits as well as new problems related to privacy [35,37]. We are unaware of other studies aside from our own implementing cognitive walkthrough with sex worker end users [16].

The iterative process of UCD contributed to the favorable evaluations of the trilingual version. UCD with multiple methods allows us to incorporate local sensitivity into the specific language and vocabulary used in this trilingual prototype, reflecting both success [18,23,24] and difficulties, including timing stakeholder involvement and managing heterogeneity among stakeholders [20,21] reported by others.

The heuristic analysis primarily identified problems with language, and end user comments offered confirmation of ways to make ReportVASW more helpful to end users; these solutions will be applied in the next steps. Additionally, UCD was useful in evaluating the ReportVASW prototype interface because adaptation is necessary to bring something initially designed for use elsewhere to the MENA context; input from end users will improve this adaptation. Evaluations of the interface using heuristic analysis and end user scenario-based testing will inform the revisions to the prototype.

The heuristic analysis offered actionable recommendations; most of the information offered was concrete and included suggestions that can be easily incorporated, for example, ways to formulate specific questions. Heuristic analysis further afforded interesting discussion related to the varied backgrounds of the analysts. For example, the most technically skilled of the

XSL•FO
RenderX

analysts disagreed that some things merited high scores, because language issues are easy to address. However, other reviewers countered that the ease of addressing these language issues does not diminish the importance of making these changes for end users who read Arabic. Arabic reads in the opposite direction from English and French, and the directional navigation arrows point the wrong way in Arabic, such that clicking the arrow to advance for an Arabic reader actually goes to the previous screen. This may not be possible to address without a change in Qualtrics software to enable the addition of labels to the directional arrows in a variety of orthographies.

End user's interested and substantive responses indicate a need for ReportVASW, provided that referral lists prove helpful and that the information collected is acted upon in advocacy. One complication is that few services for survivors of violence exist in Morocco, and therefore those available are overstretched.

Sharing the information collected about people who commit VASWs with people who may be targeted was identified as one possible benefit, with one end user saying, "If we had this five years ago, we could have prevented a lot of violence."

Users said that ReportVASW is easy to use for those who are educated and literate, but that sex workers who are not able to read will have difficulty. It may be possible for others who know about their victimization to help them by reading the questions and entering the information for them, in line with sex workers' peer-to-peer education promoting occupational health and safety [35-37]. In addition to issues of literacy, internet use is greater in wealthy countries than in low- and middle-income countries [38], although greater smartphone use everywhere may diminish this disparity; both text and voice messages have proven effective in reaching sex workers in low- and middle-income countries [39-41].

Limitations

Double the ideal range of 4 to 5 end users [22] evaluated ReportVASW using cognitive walkthrough and SUS, leading to repeated input. We recognize the possibility of implicit bias in the implementation of the SUS scale, as the facilitator read the questions and asked for SUS scale responses orally, rather than end users writing in their responses.

Conclusions and Next Steps

Follow-up is essential to the findings and implications of the project. We have received actionable recommendations through the cognitive walkthrough and the heuristic analysis that indicate clear, urgent next steps. The agreement between the task analysis of the cognitive walkthrough and the input from the cognitive walkthroughs and the heuristic analysis included many recommendations addressing aesthetics, usability, data collection, and other input about ways to improve uptake and also increase end users' confidence that ReportVASW is benign and not used for surveillance. Immediate next steps based on this input include:

- Adding all the input in changes to screening and data collection offered by end users, including adding branching questions suggested during cognitive walkthroughs.
- Making the urgent changes identified in the heuristic analysis, including fine-tuning language.
- Developing ways to share information, including a "bad date report" and a resource guide to services.

The most important next step will be to link users to services that could be helpful in the aftermath of violence, including the long-term aftermath, involving long-term effects of violence such as chronic disease [1] and posttraumatic stress disorder [10]. Sex workers indicated that sharing links to services would meet this need for survivors who are literate. There are few services for sex workers, and many existing health service providers stigmatize sex workers, presenting an obstacle to access, and for this reason, end users must be involved in the decisions about services included in order to identify service providers that do not stigmatize or discriminate against sex workers.

These changes and next steps will be undertaken before piloting the data collection tool in Morocco. The pilot will collect data and provide an opportunity to ask more general users to rate ReportVASW using the SUS. Building more evidence will contribute to understanding the reasons for sex workers to report VASWs, so that the data collected can be used in advocacy.

We now have a product that we feel confident in testing. End users confirmed the need for an application like ReportVASW and that developing the data collection tool is worth pursuing, and informatics personnel reinforced the feasibility and offered insight to improve its design and utility. The use of multiple methods to evaluate the prototype contributed to a greater understanding than any single method alone.

Acknowledgments

We would like to thank the end users who agreed to undertake the usability testing, and Sakina Ctehr for her help recruiting people to test the tool. We would like to thank Imane Jaakouk for undertaking the heuristic analysis. Usability testing was supported by the Association for Progressive Communication's Small Grants Program. We thank Tahani Abbas for her assistance with the Arabic translation. We would like to thank Rafaela Villacres and Sadaf Ahmed for undertaking the heuristic analysis of the original English version.

Data Availability

Data generated using heuristic analysis during this study are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Scenarios used for the cognitive walkthrough.
[DOCX File, 14 KB - formative_v9i1e65210_app1.docx]

References

1. Decker M, Rouhani S, Park JN, et al. Incidence and predictors of violence from clients, intimate partners and police in a prospective US-based cohort of women in sex work. Occup Environ Med 2021 Mar;78(3):160-166. [doi: 10.1136/oemed-2020-106487]
2. Deering KN, Amin A, Shoveller J, et al. A systematic review of the correlates of violence against sex workers. Am J Public Health 2014 May;104(5):e42-e54. [doi: 10.2105/AJPH.2014.301909] [Medline: 24625169]
3. Population size estimation of key populations august 2013. National AIDS program of Sierra Leone. URL: https://www.nas.gov.sl/publication/140-population-size-estimation-of-key-populations-august-2013 [accessed 2025-07-05]
4. Ditmore MH. “Caught Between the Tiger and the Crocodile”: Cambodian sex workers’ experiences of structural and physical violence. Stud Gend Sex 2014 Jan 2;15(1):22-31. [doi: 10.1080/15240657.2014.877726]
5. Nawara Women’s Network for the Middle East and North Africa. How sex work laws are implemented on the ground and their impact on sex workers: Morocco case study. Global Network of Sex Work Projects. URL: https://www.nswp.org/country/morocco [accessed 2025-07-05]
6. Stenersen MR, Thomas K, McKee S. Police harassment and violence against transgender & gender diverse sex workers in the United States. J Homosex 2024 Feb 23;71(3):828-840. [doi: 10.1080/00918369.2022.2132578] [Medline: 36228168]
7. Fehrenbacher AE, Park JN, Footer KHA, Silberzahn BE, Allen ST, Sherman SG. Exposure to police and client violence among incarcerated female sex workers in Baltimore City, Maryland. Am J Public Health 2020 Jan;110(S1):S152-S159. [doi: 10.2105/AJPH.2019.305451] [Medline: 31967867]
8. BC-wide “bad date” reporting system aims to improve sex worker safety. CBC News. URL: https://www.cbc.ca/news/canada/british-columbia/b-c-bad-date-database-sex-worker-safety-1.5897871 [accessed 2025-07-05]
9. Ditmore M, Allman D. Who Is Helsinki?”: sex workers recommend better communication in clinical trials. Health Educ Rev 2011;26(3):466-475. [doi: 10.1093/her/cyq087]
10. Park JN, Decker MR, Bass JK, et al. Cumulative violence and PTSD symptom severity among urban street-based female sex workers. J Interpers Violence 2021 Nov;36(21-22):10383-10404. [doi: 10.1177/0886260519884694] [Medline: 31679445]
11. Goldenberg SM, Morgan Thomas R, Forbes A, Baral S, editors. Sex Work, Health, and Human Rights: Global Inequities, Challenges, and Opportunities for Action: Springer International Publishing; 2021. [doi: 10.1007/978-3-030-64171-9]
12. Carr C, King LM, Maizel J, et al. Strategies and interventions used to prevent violence against sex workers in the United States: a scoping review using the social-ecological model. Trauma Violence Abuse 2024 Jul;25(3):2436-2451. [doi: 10.1177/15248380231214786]
13. Ditmore M. Aye myanmar association’s community-based intervention to combat violence against sex workers, 2017-2020. : United Nations Trust Fund to End Violence Against Women, Asia Pacific Network of Sex Workers, and Aye Myanmar Association; 2020 URL: https://apnsw.info/wp-content/uploads/2022/09/Evaluation-report.pdf [accessed 2025-08-01]
14. Brody C, Chhoun P, Tuot S, et al. A mobile intervention to link young female entertainment workers in Cambodia to health and gender-based violence services: randomized controlled trial. J Med Internet Res 2022 Jan 4;24(1):e27696. [doi: 10.2196/27696] [Medline: 34982716]
15. SANAC sex worker program. Apple store. 2016. URL: https://apps.apple.com/us/app/sanac-sex-worker-program/id1173413049 [accessed 2025-07-05]
16. Ditmore MH, Florez-Arango JF. User-centered design for designing and evaluating a prototype of a data collection tool to submit information about incidents of violence against sex workers: multiple methods approach. JMIR Hum Factors 2024 Oct 9;11:e53557 [FREE Full text] [doi: 10.2196/53557] [Medline: 39382561]
17. Final strategic plan. : Nawara Women’s Network for the Middle East and North Africa; 2023.
18. Lyon AR, Munson SA, Renn BN, et al. Use of human-centered design to improve implementation of evidence-based psychotherapies in low-resource communities: protocol for studies applying a framework to assess usability. JMIR Res Protoc 2019 Oct 9;8(10):e14990. [doi: 10.2196/14990] [Medline: 31599736]
19. Graham AK, Munson SA, Reddy M, et al. Integrating user-centered design and behavioral science to design a mobile intervention for obesity and binge eating: mixed methods analysis. JMIR Form Res 2021 May 10;5(5):e23809. [doi: 10.2196/23809] [Medline: 33970114]
20. Noergaard B, Sandvei M, Rottmann N, et al. Development of a web-based health care intervention for patients with heart disease: lessons learned from a participatory design study. JMIR Res Protoc 2017 May 17;6(5):e75. [doi: 10.2196/resprot.7084] [Medline: 28526674]

XSL•FO
RenderX

21. Cornet VP, Toscos T, Bolchini D, et al. Untold stories in user-centered design of mobile health: practical challenges and strategies learned from the design and evaluation of an app for older adults with heart failure. JMIR Mhealth Uhealth 2020 Jul 21;8(7):e17703. [doi: 10.2196/17703] [Medline: 32706745]
22. Nielsen J. Estimating the number of subjects needed for a thinking aloud test. Int J Hum Comput Stud 1994 Sep;41(3):385-397. [doi: 10.1006/ijhc.1994.1065]
23. Schnall R, Rojas M, Bakken S, et al. A user-centered model for designing consumer mobile health (mHealth) applications (apps). J Biomed Inform 2016 Apr;60:243-251. [doi: 10.1016/j.jbi.2016.02.002] [Medline: 26903153]
24. De Vito Dabbs A, Myers BA, Mc Curry KR, et al. User-centered design and interactive health technologies for patients. Comput Inform Nurs 2009;27(3):175-183. [doi: 10.1097/NCN.0b013e31819f7c7c] [Medline: 19411947]
25. 10 usability heuristics for user interface design. Nielsen Norman Group. URL: https://www.nngroup.com/articles/ten-usability-heuristics/ [accessed 2025-07-05]
26. Uribe-Ocampo S, Torres EA, Luna IF, Florez-Arango JF, Smith JW. SIM-CIG: a serious game to practice and improve clinical guidelines adoption based on computer-interpretable guidelines. Stud Health Technol Inform 2019 [FREE Full text] [doi: 10.3233/SHTI190753]
27. Zhang J, Walji MF. TURF: toward a unified framework of EHR usability. J Biomed Inform 2011 Dec;44(6):1056-1067. [doi: 10.1016/j.jbi.2011.08.005] [Medline: 21867774]
28. Thukral J, Ditmore MH. Revolving door. sex workers project. : Sex Workers Project of the Urban Justice Center; 2003.
29. New Mexico (2009): Qualtrics. Qualtrics. 2020. URL: https://www.qualtrics.com [accessed 2025-07-05]
30. Ditmore MH, Maternick A, Zapert K. The road north: the role of gender, poverty and violence in trafficking from Mexico to the US. : Urban Justice Center Sex Workers Project; 2012.
31. Brooke J. Usability. Digital.gov. URL: https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html [accessed 2023-03-30]
32. Gao M, Kortum P, Oswald FL. Multi-language toolkit for the system usability scale. Int J Hum Comput Interact 2020 Dec 13;36(20):1883-1901. [doi: 10.1080/10447318.2020.1801173]
33. Ditmore M. Making international mechanisms work for MENA communities. : Global Fund MENA Platform, ITPC MENA, Nawara Women's Network, UNDP; 2019 URL: https://nawara.org/downloads/2019/report-gf-platform-november-2019-en.pdf [accessed 2025-08-01]
34. Al Rasheed A. Action plan on gender-based violence in sudan. : UNDP; 2022.
35. Bernier T, Shah A, Ross LE, Logie CH, Seto E. The use of information and communication technologies by sex workers to manage occupational health and safety: scoping review. J Med Internet Res 2021 Jun 24;23(6):e26085. [doi: 10.2196/26085] [Medline: 34185001]
36. Smart service provider's guide to ICT and sex work. Global Network of Sex Work Projects. 2017. URL: https://www.nswp.org/resource/nswp-smart-guides/smart-service-providers-guide-ict-and-sex-work [accessed 2024-09-20]
37. Panchanadeswaran S, Unnithan AM, Chacko S, Brazda M, Kuruppu S. What's technology got to do with it? exploring the impact of mobile phones on female sex workers' lives and livelihood in India. Gend Technol Dev 2017 May 4;21(1-2):152-167. [doi: 10.1080/09718524.2017.1385318]
38. Kraidy M. The internet as a mass communication medium. J Mass Commun 2011;2 [FREE Full text]
39. Brody C, Dhaliwal S, Tuot S, Johnson M, Pal K, Yi S. Are text messages a feasible and acceptable way to reach female entertainment workers in Cambodia with health messages? a cross-sectional phone survey. JMIR Mhealth Uhealth 2016 May 20;4(2):e52. [doi: 10.2196/mhealth.5297] [Medline: 27207374]
40. Thomas B, Closson EF, Biello K, et al. Development and open pilot trial of an HIV-prevention intervention integrating mobile-phone technology for male sex workers in Chennai, India. Arch Sex Behav 2017 May;46(4):1035-1046. [doi: 10.1007/s10508-015-0665-3] [Medline: 26714684]
41. Nyatsanza F, McSorley J, Murphy S, Brook G. "It's all in the message": the utility of personalised short message service (SMS) texts to remind patients at higher risk of STIs and HIV to reattend for testing-a repeat before and after study. Sex Transm Infect 2016 Aug;92(5):393-395. [doi: 10.1136/sextrans-2015-052216] [Medline: 26670912]

Abbreviations

GBV: gender-based violence
MENA: Middle East and North Africa
mHealth: mobile health
SUS: System Usability Scale
UCD: user-centered design
VASW: violence against sex workers

XSL•FO
RenderX

Edited by A Mavragani; submitted 08.08.24; peer-reviewed by J Koken, T Bernier; revised version received 19.03.25; accepted 21.03.25; published 22.08.25.

Please cite as:
Ditmore M, El Karouaoui A, Florez-Arango JF
Designing and Evaluating a Prototype of a Trilingual Data Collection Tool for the Middle East and North Africa (MENA) Region to Collect Data About Violence Against Sex Workers: Multiple Methods Approach in User-Centered Design
JMIR Form Res 2025;9:e65210
URL: https://formative.jmir.org/2025/1/e65210
doi:10.2196/65210

© Melissa Ditmore, Amal El Karouaoui, Jose Fernando Florez-Arango. Originally published in JMIR Formative Research (https://formative.jmir.org), 22.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Relationship of Hair Cortisol Concentration With Perceived and Somatic Stress Indices: Cross-Sectional Pilot Study

Sharon H Bergquist[1], MD; Danyang Wang[1], MPH; Brad Pearce[2], PhD; Alicia K Smith[3], PhD; Allison Hankus[2], MPH; David L Roberts[1], MD; Miranda A Moore[1,4], PhD

[1]Department of Medicine, School of Medicine, Emory University, 1365 Clifton Rd, Atlanta, GA, United States

[2]Rollins School of Public Health, Emory University, Atlanta, GA, United States

[3]Department of Gynecology and Obstetrics, School of Medicine, Emory University, Atlanta, GA, United States

[4]Department of Family and Preventive Medicine, School of Medicine, Emory University, Atlanta, GA, United States

Corresponding Author:
Sharon H Bergquist, MD
Department of Medicine, School of Medicine, Emory University, 1365 Clifton Rd, Atlanta, GA, United States

Abstract

Background: Hair cortisol is an emerging biomarker of chronic stress. However, the psychological and physiological aspects of chronic stress that are reflected in hair cortisol concentration (HCC) have not been fully determined. Since physiological responses to stress do not always align with how stress is perceived, we conducted this study to evaluate whether HCC correlates with neuroendocrine stress indicators or stress perceptions.

Objective: This study aimed to evaluate whether subjective (Perceived Stress Scale and Connor-Davidson Resilience Scale) and objective (plasma cortisol/dehydroepiandrosterone-sulfate [DHEA-S] and cortisol/high-sensitivity C-reactive protein) determinants of stress and resilience correlate with HCC.

Methods: In this cross-sectional pilot validity study, scatter plots and Spearman correlation coefficients were used to measure the direction and magnitude of the relationship between stress and resilience measures among 51 predominantly male participants. In a subset (n=24), we performed a step-wise regression modeling approach to isolate the association of perceived and somatic stress on hair cortisol.

Results: Bivariate correlations showed a weak inverse association of HCC with Perceived Stress Scale (Spearman correlation $\rho=-0.14$, $P=.52$) and a stronger positive association with somatic neuroendocrine stress indices cortisol/DHEA-S ($\rho=0.24$, $P=.25$) and cortisol/high-sensitivity C-reactive protein ($\rho=0.21$, $P=.35$). In linear regression models, HCC showed the strongest association with cortisol/DHEA-S ($r^2=0.10$, $P=.13$, 1.01^{β} 1.01, 95% CI 0.99 - 1.01). This relationship remained when age, gender, hair washing frequency, hair dye or bleach use, diabetes mellitus, obesity, cardiovascular disease, anxiety, medication use, and endocrine disorders were considered.

Conclusions: Our results do not indicate a statistically significant association (at the $P<.05$ threshold) between HCC and stress perception or somatic measures of neuroendocrine response.

(JMIR Form Res 2025;9:e63811) doi:10.2196/63811

KEYWORDS

hair cortisol concentration; perceived stress; somatic stress; resilience; physiological stress

Introduction

Stress and its related comorbid diseases are responsible for a large proportion of health care expenditure [1,2] and disability worldwide [3]. Psychiatric disorders, including stress-related disorders, are predicted to be second only to ischemic heart disease as the leading cause of disability by 2020 [4]. Stress can be defined as the self-perception that an individual does not have the resources to cope or respond to a threat [5]. The duration of stress exposure is a key determinant of the disease risk attributed to stress. Acute stress is an adaptive response to a challenge and is generally not harmful to health [6]. However, when stress is severe, or experienced continuously or repeatedly, it can lead to psychological and physical illness [5,7]. While the detrimental effects of chronic stress on health are well-recognized, measuring stress burden remains challenging [8]. Biomarkers of chronic stress reactivity are needed for early risk determination and targeted risk management in clinical practice [9].

Current instruments used for assessing stress burden include subjective and objective measures, each of which has limitations. Subjective self-assessments rely on a conscious perception of stress, which can be influenced by genetic, developmental,

XSL•FO
RenderX

cultural, social, and resilience factors. Hence the perception of the quality and intensity of a stressor is not consistently correlated with physiologic stress reactivity [7,8]. The commonly used objective assessment of stress is via measurement of the glucocorticoid cortisol, the downstream effector released by the neuroendocrine hypothalamus-pituitary-adrenal (HPA) axis. Unfortunately, capturing long-term cortisol secretion is methodologically challenging. Traditional measures of salivary and plasma cortisol provide insight into cortisol level at a single point in time [9], and are subject to diurnal fluctuation [10]. Cortisol sampling over 1 day can be accomplished with a 24-hour urine collection. In addition to being logistically burdensome, this short time interval also fails to capture long-term cortisol output.

Since cortisol is incorporated into growing hair, hair is emerging as a novel matrix for measuring retrospective cortisol secretion over months [9,11,12]. Each 1 cm hair segment, beginning from the proximal end, approximates a month's cortisol production. While accumulating evidence supports the validity of hair cortisol concentration (HCC) as an index of long-term systemic cortisol secretion and its reliability across repeated assessments [10,11], 2 central gaps remain in its clinical applicability [9-11]. The first is determining the covariates that can influence HCC. Research has identified age, sex, hair washing frequency, hair treatment, and oral contraceptive use as relevant covariates [12]. Health conditions and other medications can also influence HCC [10]. The second is understanding the stress-related determinants of HCC [12]. Thus far, the correlation between validated measures of perceived stress and HCC has been inconsistent [9,12,13]. Psychological resilience has been shown to moderate the indirect association of perceived stress with illness severity via HCC [14] but little research is available on the relationship between HCC and physiologic stress indices.

We hypothesize that the strength of the correlation of HCC with somatic measures of stress will be greater than with psychologic measures of stress. We additionally speculate that the relationship of HCC with somatic stress constructs will be stronger among constructs that take into account neuroendocrine resilience factors and neuroinflammatory responses. Current evidence indicates that during the stress response, dehydroepiandrosterone (DHEA) and its sulfate ester, dehydroepiandrosterone-sulfate (DHEA-S), collectively referred to as DHEA(S), are cosynthesized with cortisol and released from the adrenal glands [15]. Several lines of evidence suggest DHEA has anabolic and neuroprotective effects that antagonize cortisol [16,17], and the ratio of cortisol to DHEA(S) has been considered an index of "net steroid activity" [18]. Cortisol/DHEA(S) may be a better predictor of stress outcomes than either hormone alone [18,19] and a higher ratio of cortisol/DHEA(S) has been shown to predict a higher risk of disease [18,20].

Another factor potentially moderating the association between stress reactivity and HCC is the chronicity of stress, which has been proposed to result in glucocorticoid receptor resistance [21] and consequent failure to down-regulate the inflammatory response [21]. The ratio of cortisol to high-sensitivity C-reactive protein (hs-CRP) has been examined as an alternate index that captures the integrity of homeostatic regulation between the HPA axis and inflammatory processes [22].

Investigating levels of HCC, cortisol, DHEA-S, and hs-CRP simultaneously may provide important clinical insights; although this has rarely been done in clinical studies [23]. The first aim of this study is to examine the interrelation between HCC and subjective measures of stress (Perceived Stress Scale [PSS]) and resilience (Connor Davidson Resilience Scale [CD-RISC]), and different objective measures of neuroendocrine stress reactivity and inflammatory balance (cortisol/DHEA-S, cortisol/hs-CRP). The secondary aim is to assess the health-related and other covariates that can influence HCC.

Methods

Recruitment

We conducted a prospective, cross-sectional cohort, pilot study of 51 participants from patients seen at the executive health program at Emory University in Atlanta, Georgia between May 23, 2017 and October 19, 2017. The executive health program provides comprehensive health examinations primarily to senior-level corporate executives. During the nursing intake process, the 499 patients seen in the clinic in the time period were invited to participate, and written informed consent was obtained and the nature of the procedures was fully explained. Participants completed 2 survey assessments (totaling approximately 10 min) and provided a 2 mL tube of blood in addition to providing clinical histories and undergoing tests obtained during routine care. Furthermore, 24 participants also agreed to provide a pencil-wide hair sample used for hair cortisol analysis. The patients did not receive their study results and there were no follow-ups to the initial study data collection.

Ethical Considerations

Participation was voluntary and patients were notified that participation would not impact their clinical care. No payment was offered for participation. The protocol for this study was approved by the Emory University Institutional Review Board (IRB 00092476). This study was conducted in accordance with the latest version of the Declaration of Helsinki. All data were deidentified for publication. During the nursing intake process, patients were invited to participate, and written informed consent (Multimedia Appendix 1) was obtained, and the nature of the procedures was fully explained.

Psychological Stress Measures

The PSS 10-item designed by Cohen et al [24] was used for assessing perceived stress over the past month. This psychometrically validated 10-item self-reported questionnaire is the most widely used psychological instrument for measuring the degree to which an individual appraises situations in one's life as stressful. Each item is forward or reverse scored from 0 to 4 based on the experienced frequency of each item. Scores on the PSS can range from 0 to 40 with higher scores indicating higher perceived stress. The US population probability sample validated mean PSS value was 12.1 (SD 5.1) for males overall and 12.6 (SD 6.1) for males aged 45 - 54 years [25,26]. The CD-RISC 25-item has robust psychometric properties for assessing resilience, viewed as a measure of stress-coping ability

XSL•FO
RenderX

[27]. It is comprised of 25 items, each rated on a 5-point scale (0 - 4), with higher scores reflecting greater resilience. The total possible scores range from 0 to 100. Scores were categorized by quintile based on results of the general US population, with the lowest quintile of resilience Q1 (0 - 73), Q2 (74-82), Q3 (83 - 90), and Q4 (91-100) [27]. Furthermore, 2 additional questions, rated on a 1 - 10 Likert scale, were also obtained: (1) "How would you rate your stress level?" and (2) "How would you rate how you respond to stress?"

Cortisol, DHEA(S), and Hs-CRP

Blood samples were obtained early morning (6:35 AM to 9:36 AM) in all participants and timing of blood drawn was recorded to reduce cortisol variability from diurnal cortisol rhythms. Blood was also collected in the fasting state to avoid the potential of meal effect on cortisol. For plasma cortisol and DHEA(S), a 2 mL serum separator tube was processed. Plasma cortisol and DHEA are reported in nanograms per milliliter (ng/mL). DHEA(S) is reported in micrograms per milliliter (µg/mL). Plasma cortisol was determined by enzyme-linked immunosorbent assay (ELISA) per the manufacturer protocol (IBL America IB79343). The cortisol assay dynamic range was 10 - 800 ng/mL with a sensitivity of 3.79 ng/mL. Plasma DHEA and DHEA(S) were determined by discrete ELISAs per the manufacturer protocols (IBL IB79332; IBL IB79342). The DHEA assay dynamic range was 0.3 - 30 ng/mL with a sensitivity of 0.07 ng/mL. The DHEA(S) assay dynamic range was 0.03 - 10 µg/mL with a sensitivity of 0.002 µg/mL. Hs-CRP was processed as part of the routine physical examination laboratory assessment by Emory Medical Laboratory, the fully accredited and licensed clinical laboratory operating within Emory Healthcare.

Hair Collection and Analysis

Hair samples were collected from a subset of participants (n=24), limited by willingness to participate and ability to contribute sufficient hair volume in this predominantly male cohort. Those who provided hair samples were given a questionnaire that solicited the frequency of hair washing; use of scalp treatments; and use of hair dyes, bleaching products, permanent straighteners, or curling products in the previous 3 months. A pencil-wide distribution (approximately 50 hairs) was cut with scissors as close as possible to the posterior vertex scalp to yield the lowest intraindividual coefficient of variation for cortisol concentration [28]. The hair sample was placed in a collection kit, with the root end indicated. Hair length was measured from the root end and isolated. The samples were weighed, washed in isopropanol, dried, and transferred to metal bead lysing matrix tubes (MP 6925 - 100). The transferred sample weight was recorded. The lysing tubes were run for cycles of 40s in a grinding machine (BIO101 Savant FastPrep FP120 [Savant Instruments]) until each sample reached uniform powder consistency. Upon grinding completion, cortisol was extracted in methanol overnight. The samples were centrifuged to pellet the powdered hair and the supernatant was isolated. The extracted cortisol supernatant was placed in a vacuum evaporator (Savant DNA Speed Vac DNA120 [Savant Instruments]) set at RT dry for 100 min to completely remove methanol from the sample. The cortisol extract was reconstituted in ELISA assay diluent (Salimetrics PN8005). Cortisol was measured via immunoassay (Salimetrics 1 - 3002), in a single batch, using internal positive controls as standards, to circumvent potential interassay variability among immunoassays. The Salimetrics cortisol assay dynamic range was 0.3 - 3 µg/dL with a sensitivity of 0.007 µg/dL.

Demographic and clinical characteristics, including the presence of mental and physical conditions, were obtained from the medical record.

Statistical Analysis

Descriptive statistics analyses were performed to obtain mean (SD) or number (percentage) for sociodemographic characteristics, stress variables, and medical conditions and use of medications that may alter cortisol production. As many of our variables are clinically limited to a specific range, the following variables contained outliers and were truncated: plasma DHEA (2 observations truncated at 15.0 ng/mL) and plasma DHEA(S) (2 truncated at 5 µg/mL). In addition, normality tests were performed to determine if each variable was well-modeled by normal distributions. Based on this analysis, plasma DHEA and plasma DHEA(S) were log-transformed for analysis. We used scatter plots and Spearman correlation coefficients (ρ) to measure the direction and magnitude of the relationship between stress and resilience measures.

To isolate the impact of perceived and somatic stress on HCC from the impact of other factors impacting HCC, we used a step-wise regression modeling approach. As a first step, we used simple linear regression modeling of each perceived and somatic neuroendocrine stress construct regressed on HCC to determine which variables had the strongest independent association. Next, we used multivariate regression modeling to calculate fully adjusted associates controlling for factors impacting HCC. Based on previous studies, the covariates included were age, gender, frequency of hair washing, use of hair dyes or bleach, topical and systemic medication (dichotomous indicator of any use), prediabetes or diabetes mellitus, obesity, subclinical or clinical cardiovascular disease, anxiety, and endocrine disorders [10,12]. Race was not included in our model as there was limited variability in the sample.

Other mental health conditions (depression, post-traumatic stress disorder, and bipolar illness) were also queried but not included in the regression model due to an insufficient number of observations with known data among participants who provided hair samples. Exponentiation, based on the variable form, of the β-coefficient with 95% CIs and t values were calculated. We performed sensitivity analyses for the variables in our models to determine the minimum effect size that could yield a statistically significant result with a significance level of 0.05 and power of 0.80 for our sample size using Cohen f^2 measures. All study variables were available for all study participants; thus, we had no missing data to exclude. All data analyses were conducted using StataCorp LLC StataSE 16 software [29].

XSL•FO
RenderX

Results

Demographic and Stress Characteristics

The average age of the full sample of participants in our study was 53.7 (SD 8.8 y; range 37 - 77 y, Table 1). The majority were white (48/51, 94%) males (46/51, 92%). Our sample PSS values were slightly lower than the US national average at 11.7 (SD 5.4; range 3.0 - 28.0). The mean CD-RISC score in our sample was 82.6 (SD 12.4; range 46.0 - 100.0), which was higher than the mean score of 80.4 (SD 12.8) in the US general population [28]. Mean levels of cortisol, DHEA, and DHEA(S) were 225.3 (SD 67.6) ng/mL, 4.9 (SD 3.2) ng/mL, and 2.2 (SD 1.2) µg/mL, respectively.

Demographic, stress, and clinical characteristics were not significantly different in the 24 participants in whom sufficient hair volume was available for analysis than in the full sample of participants (Table 1). The average HCC was 32.32 (SD 45.50; range 1.29 - 202.41) ng/g hair. Among the 24 participants, the average age was 54.8 (SD 10.1; range 39 - 77) years. Furthermore, 83% (20/24) were males and 95% (23/24) were White. The average PSS value was 11.9 (SD 6.0; range 3.0 - 28.0). The mean CD-RISC score was 85.0 (SD 14.0; range 46.0 - 100.0). Mean levels of cortisol, DHEA, and DHEA(S) were 227.2 (SD 64.8) ng/mL, 4.6 (SD 3.1) ng/mL, and 2.2 (SD 1.0) µg/mL, respectively (Table 1).

XSL•FO
RenderX

Table . Demographic, stress, and clinical characteristics (this prospective, cross-sectional cohort, pilot study includes 51 patients seen in the executive health clinical program at Emory Healthcare in Atlanta, Georgia between May 23, 2017 and October 19, 2017).

Characteristics	Full sample (n=51)		Regression sample (n=24)		Nonregression sample (n=27)		*P* value[h]
	Statistical value	Range	Statistical value	Range	Statistical value	Range	
Age (years), mean (SD)	53.7 (8.8)	37-77	54.8 (10.1)	39-77	52.8 (7.6)	37-66	.43
Sex, n (%)							
Female	5 (10)	—[i]	4 (17)	—	1 (4)	—	.18
Male	46 (90)	—	20 (83)	—	26 (96)	—	—
Race, n (%)							
White	48 (94)	—	23 (96)	—	25 (93)	—	≥.99
Other	3 (6)	—	1 (4)	—	2 (7)	—	—
Self-rated stress (Likert scale; n=50), mean (SD)	6.2 (2.1)	2-10	6.3 (2.2)	2-10	6.1 (1.9)	3-10	.71
Self-rated response to stress (Likert scale; n=50), mean (SD)	2.7 (1.5)	1-8	2.6 (1.7)	1-8	2.9 (1.3)	1-6	.47
Perceived Stress Scale-10 (total score), mean (SD)	11.7 (5.4)	3-28	11.9 (6.0)	3-28	11.6 (5.0)	2-23	.86
Connor-Davidson Resilience Scale-25 (total score), mean (SD)	82.6 (12.4)	46-100	85.0 (14.0)	46-100	80.4 (10.5)	55-98	.19
Q1[j] (0 - 73), n (%)	11 (22)	—	4 (17)	—	7 (26)	—	.02
Q2[k] (74-82), n (%)	12 (24)	—	3 (13)	—	9 (33)	—	—
Q3[l] (83-90), n (%)	13 (26)	—	5 (21)	—	8 (30)	—	—
Q4[m] (91-100), n (%)	15 (29)	—	12 (50)	—	3 (11)	—	—
Hair cortisol (ng/g; n=24), mean (SD)	32.3 (45.5)	1.3-203.4	32.3 (45.5)	1.3-202.4	—	—	—
Hair bleached or dyed, n (%)	6 (25; n=24)	—	5 (22; n=23)	—	—	—	—
Hair washing frequency (times/week), mean (SD)	6.3 (1.7; n=24)	1-7	6.2 (1.7; n=23)	1-7	—	—	—
Plasma cortisol (ng/ml), mean (SD)	225.3 (67.6)	102.0-366.6	227.2 (64.8)	114.1-330.2	223.7 (71.1)	102.0-366.6	.86
Plasma DHEA[n] (ng/ml), mean (SD)	4.9 (3.2)	0.9-15.0[o]	4.6 (3.1)	0.9-15.0[o]	5.2 (3.3)	1.4-15.0[o]	.55

XSL•FO
RenderX

Characteristics	Full sample (n=51)		Regression sample (n=24)		Nonregression sample (n=27)		*P* value[h]
	Statistical value	Range	Statistical value	Range	Statistical value	Range	
Plasma DHEA(S[p]) (µg/ml), mean (SD)	2.2 (1.2)	0.0-5.0[o]	2.2 (1.0)	0.1-3.8	2.2 (1.3)	0.0-5.0[o]	.80
Medical condition							
Obesity (BMI≥30 kg/m^2), n (%)	13 (26)	—	7 (30)	—	6 (22)	—	.57
Prediabetes or diabetes mellitus, n (%)	12 (24)	—	10 (42)	—	2 (7)	—	.004
Subclinical or clinical cardiovascular disease, n (%)	16 (31)	—	9 (38)	—	7 (26)	—	.37
Endocrine disorders, n (%)	5 (10)	—	2 (8)	—	3 (11)	—	≥.99
Generalized anxiety disorder, n (%)	5 (10)	—	3 (13)	—	2 (7)	—	.66
Sleep (h/night), mean (SD)	6.7 (1.0; n=19)	5-8	6.7 (1.0; n=8)	5-8	6.7 (1.0; n=11)	5-8	.93
Medications, n (%)							
Topical scalp medications							
Steroids	0 (0)	—	0 (0)	—	0 (0)	—	—
Nonsteroids	1 (2)	—	1 (4)	—	0 (0)	—	—
Oral medications							
OCP[q] or HRT[r]	1 (2)	—	0 (0)	—	1 (4)	—	—
SSRI[s] or SNRI[t]	1 (2)	—	1 (4)	—	0 (0)	—	—
Testosterone or other hormones	1 (2)	—	0 (0)	—	1 (4)	—	—

[h]*P* values for differences between the regression sample and nonregression sample, using *t* test, chi-square test, or Fisher exact test.

[i]Not applicable.

[j]Q1: quintile 1.

[k]Q2: quintile 2.

[l]Q3: quintile 3.

[m]Q4: quintile 4.

[n]DHEA: dehydroepiandrosterone.

[o]range after truncation.

[p]DHEA(S): collective term for dehydroepiandrosterone and dehydroepiandrosterone-sulfate.

[q]OCP: oral contraceptive pills.

[r]HRT: hormone replacement therapy.

[s]SSRI: selective serotonin reuptake inhibitor.

[t]SNRI: serotonin and norepinephrine reuptake inhibitor.

XSL•FO
RenderX

Hair Cortisol and Its Relationship With Perceived and Physiologic Stress Reactivity Measures

There was a significant inverse association between PSS and CD-RISC scores (ρ=−0.33, *P*<.05, Table 2). However, PSS and the ratio of PSS/CD-RISC showed weak inverse associations with HCC and did not reach statistical significance (ρ=−0.14 and −0.10, *P*=.52 and .63, respectively). The association was stronger between HCC and the neuroendocrine stress indices of cortisol/DHEA(S) (ρ=0.24, *P*=.25) and cortisol/hs-CRP (ρ=0.21, *P*=.35), but the strength of the positive association was weak to moderate and did not reach significance (Table 2 and Figure 1).

Table . Bivariate Spearman correlations between subjective and objective stress-related determinants (this prospective, cross-sectional cohort, pilot study includes 24 patients seen in the executive health clinical program at Emory Healthcare in Atlanta, Georgia between May 23, 2017 and October 19, 2017).

Stress-related determinants	1	2	3	4	5	6	7
PSS[a]-10 score	—[b]	—	—	—	—	—	—
CD-RISC-25[c] score	−0.33[d]	—	—	—	—	—	—
PSS-10 or CD-RISC-25	0.96[e]	−0.54[e]	—	—	—	—	—
HCC[fg]	−0.14	−0.05	−0.10	—	—	—	—
Plasma cortisol	−0.08	0.07	−0.11	0.07	—	—	—
Cortisol/DHEA[hi]	−0.06	−0.05	−0.05	0.09	0.04	—	—
Cortisol/DHEA(S)[ij]	−0.04	−0.02	−0.03	0.24	0.33[d]	0.55[e]	—
Cortisol/hs-CRP[k]	−0.12	−0.09	−0.13	0.21	0.30[d]	−0.03	0.14

[a]PSS: Perceived Stress Scale.

[b]Not applicable.

[c]CD-RISC: Connor Davidson Resilience Scale.

[d]Significance level *P*<.05.

[e]Significance level *P*<.001.

[f]HCC: hair cortisol concentration.

[g]n=24.

[h]DHEA: dehydroepiandrosterone.

[i]Variable truncated.

[j]DHEA(S): collective term for dehydroepiandrosterone and dehydroepiandrosterone-sulfate.

[k]hs-CRP: high-sensitivity C-reactive protein.

XSL•FO
RenderX

Figure 1. Correlation between log hair cortisol concentration and predictors. Log transformed hair cortisol concentration does not show a linear relationship with (A) Perceived Stress Scale-10 score but does show a weak to moderate linear correlation with (B) log transformed cortisol/DHEA(S) ratio. DHEA(S): collective term for dehydroepiandrosterone and dehydroepiandrosterone-sulfate; HCC: hair cortisol concentration; PSS: Perceived Stress Scale.

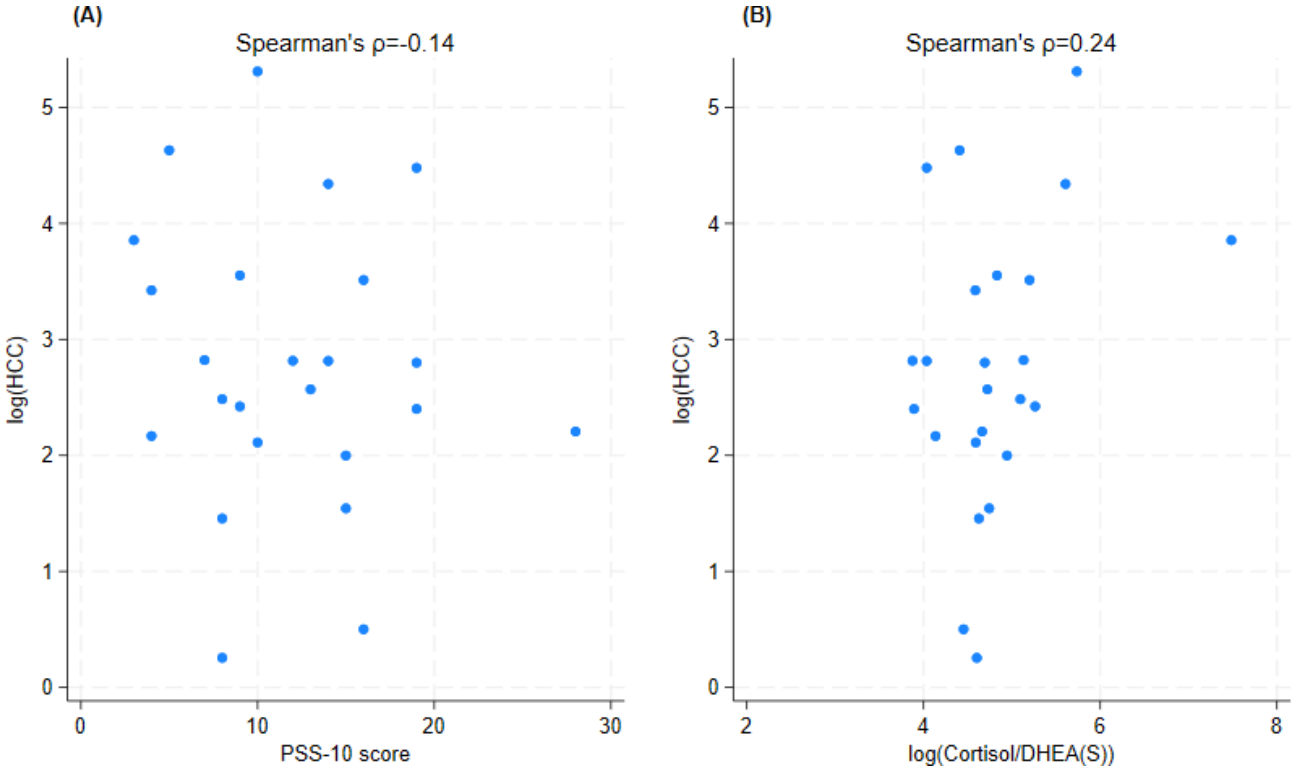

This prospective, cross-sectional cohort, pilot study includes 24 patients seen in the executive health clinical program at Emory Healthcare in Atlanta, Georgia between May 23, 2017 and October 19, 2017.

In simple linear regression models using hair cortisol as the outcome variable and PSS, cortisol, cortisol/DHEA, cortisol/DHEA(S), and cortisol/hs-CRP as the respective predictor variables, HCC showed the strongest association with cortisol/DHEA(S) (r^2=0.10, $F_{1,22}$=2.51, P=.13, 1.01^{β} 1.01, 95% CI 0.998 - 1.012, Table 3). A moderate positive relationship remained when age, gender, frequency of hair washing, use of hair dyes or bleach, prediabetes or diabetes mellitus, obesity, subclinical or clinical cardiovascular disease, anxiety, endocrine disorders, and medications were used as control variables (Table 3). In the fully adjusted model, each 1% increase in cortisol/DHEA(S) was associated with a 1% increase in HCC (1.01^{β} 1.00, 95% CI 0.996 - 1.011, t_2=1.00, P=.34). In addition, the control variables with the strongest association to hair cortisol were history of an endocrine disease, subclinical or clinical cardiovascular disease, male, shampoo frequency, and age.

In sensitivity analyses, the minimum effect size for yielding a statistically significant result with our sample size was Cohen f^2 of 0.358 for simple linear regressions. Our observed Cohen f^2 ranged from 0.001 to 0.114. For multiple linear regressions, the minimum Cohen f^2 measuring the local effect size of cortisol/DHEA(S) that can be detected at 80% power is 0.412. The observed Cohen f^2 for cortisol/DHEA(S) was 0.091, indicating a small effect.

XSL•FO
RenderX

Table . Simple and multiple linear regression analyses (this prospective, cross-sectional cohort, pilot study includes 24 patients seen in the executive health clinical program at Emory Healthcare in Atlanta, Georgia between May 23, 2017 and October 19, 2017).

Regression determinants	Log (HCC)[a]				
	β coefficient	e^{β} or 1.01^{β} (95% CI of e^{β} or 1.01^{β})	Regression *t* statistics[b]	R^2	Cohen f^2
Simple linear regressions					
PSS-10[c] score	−0.025	0.98 (0.89-1.07)	−0.58	0.02	0.02
CD-RISC-25[d] score	−0.003	0.997 (0.96-1.034)	−0.16	0.001	0.001
Serum cortisol	0.001	1.00 (0.99-1.01)	0.30	0.004	0.004
log (Cortisol/DHEA)[ef]	0.26	1.00 (0.99-1.01)	0.48	0.01	0.01
log (Cortisol/DHEA(S[g]))	0.52	1.01 (0.998-1.01)	1.58	0.10	0.11
log (Cortisol/hs-CRP)[h]	0.33	1.00 (0.998-1.01)	1.42	0.09	0.10
Multiple linear regression					
log (Cortisol/DHEA(S))	0.33	1.00 (0.996-1.01)	1.00	—[i]	0.09
Hair washing frequency	0.05	1.05 (0.71-1.55)	0.28	—	0.01
Hair bleached or dyed	−0.28	0.80 (0.13-4.76)	−0.28	—	0.01
Age	0.01	1.01 (0.94-1.08)	0.33	—	0.01
Male	0.20	1.22 (0.08-18.09)	0.17	—	0.002
Prediabetes or diabetes	−0.89	0.41 (0.12-1.45)	−1.56	—	0.22
Obesity	−1.33	0.27 (0.07-1.05)	−2.12	—	0.41
Subclinical or clinical CVD[j]	0.18	1.197 (0.25-5.63)	0.26	—	0.01
Anxiety	−1.26	0.28 (0.02-3.83)	−1.06	—	0.10
Endocrine disorders	1.97	7.19 (0.51-101.38)	1.64	—	0.25
On medications	−1.52	0.22 (0.03-1.55)	−1.71	—	0.27

[a]HCC: hair cortisol concentration.

[b]The multiple linear regression has a single df for the entire model, that is, df=13.

[c]PSS-10: Perceived Stress Scale-10.

[d]CD-RISC-25: Connor Davidson Resilience Scale-25.

[e]DHEA: dehydroepiandrosterone.

[f]Variable truncated.

[g]DHEA(S): collective term for dehydroepiandrosterone and dehydroepiandrosterone-sulfate.

[h]hs-CRP: high-sensitivity C-reactive protein.

[i]Not applicable.

[j]CVD: cardiovascular disease.

XSL•FO
RenderX

Discussion

Principal Results

The primary aim of this study was to examine the relationship between HCC and subjective stress perception and neuroendocrine indicators derived from blood samples. We did not find a statistically significant association between HCC with somatic stress indices or with perceived stress at the $P<.05$ level. We did find several weak to moderate nonsignificant correlations, with a stronger relationship between HCC and other somatic indices of stress rather than with perceived stress. Interestingly, HCC showed the strongest positive association with cortisol/DHEA(S). While we hypothesized that somatic stress indicators would show a stronger association with HCC, our results did not support a significant difference. Of interest, we found weak evidence that neuroendocrine indices that include the modulating effect of resilience factors may complement HCC as a clinical tool for measuring chronic stress burden.

Although HCC is emerging as a useful biomarker of chronic stress, the elements of cortisol production and stress response regulation reflected in HCC remain undetermined. Previous investigations of stress-related determinants of HCC have focused primarily on subjective self-evaluations of stress rather than on somatic neuroendocrine parameters of stress reactivity. To the best of our knowledge, Qiao et al [23] have conducted the only study evaluating HCC and DHEA/cortisol simultaneously, with all measurements obtained from the hair matrix. Their results showed higher HCC levels and lower DHEA to cortisol ratios among a high-stress group. The positive association we found between HCC and cortisol/DHEA(S) is consistent with the findings of Qiao et al [23] and lends insight into the neuroendocrine stress variables that may be predictors of HCC. Interestingly, we found a stronger correlation of HCC with cortisol/DHEA(S) rather than cortisol/DHEA; potentially due to the longer half-life and lower clearance of DHEA-S compared with DHEA. Although the variance in HCC attributed to cortisol/DHEA(S) was of small magnitude, the strength of the association may have been greater if the time period and collections (number of days, collections per day) of cortisol and DHEA(S) were extended to more than a single point in time. Previous reports have shown that salivary cortisol has a higher correlation to HCC when the time period of collection is 30 days [30] rather than 1 or 2 sampling days [12].

We also found a weak negative correlation between PSS and the Cortisol/hs-CRP. Chronic stress and depression are associated with reduced integrity of immune responses. Somewhat paradoxically, these mental conditions may be associated with immune suppression or excessive inflammation [31,32]. Cortisol and inflammatory mediators, such as CRP, have a reciprocal relationship such that cortisol is the primary neuroendocrine hormone responsible for keeping inflammation in check, and numerous inflammatory mediators drive the HPA axis leading to an increased level of cortisol. Thus, physical and mental health are somewhat dependent on a balance between cortisol and inflammatory processes, which are necessary for combating infectious challenges. Some researchers have argued that the ratio of cortisol to CRP is a better indicator of physical and mental health than either of these biomarkers alone [32,33].

In terms of clinical applicability, we found a positive association of HCC with obesity, prediabetes and diabetes, cardiovascular disease, and endocrine conditions, and an inverse association with anxiety, which is consistent with previous reports [10,12,34,35]. A recent literature review supports an association between HCC and post-traumatic stress disorder, depressive disorders, and ongoing social and family stress [36]. While the mechanism by which cortisol enters the hair shaft is not fully understood [9], the overlap between the somatic and mental health conditions associated with HCC and the ones associated with plasma cortisol support that the mechanism is primarily via blood.

Limitations

This study has several additional limitations. The primary limitation in our study was the small sample size, which may have limited the ability of our correlations to reach statistical significance. Second, the cross-sectional design and single point-in-time data collection restrict the implications of our findings. In addition, the predominance of males and recruitment among business executives limit generalizability. The largely male cohort also reduced the feasibility of hair sample collection in the clinical setting. Although we used well-validated surveys of subjective stress and resilience, other self-rated instruments could have been used to assess different stressor categories, such as response to global stress as well as stressful life events. Due to the variability of hair length among participants, different lengths of hair were obtained, representing different durations of hair cortisol accumulation in our weighed samples. Finally, the data collection is not recent.

Comparison With Previous Work

Although we interpret our results with caution due to the lack of statistical significance, the weak inverse association found between PSS and HCC suggests a potential methodologic advantage of objective stress measures for evaluating stress burden in clinical practice and that retrospective stress perception may not be a sufficient vulnerability marker. This finding is consistent with previous studies that have shown an inconclusive, often contradictory association between perceived stress and HCC [10]. In a review of HCC with subjective stress measures, less than half of the studies reported a significant association [37]. Potential explanations include relatively small sample sizes, heterogeneity of the study populations, diversity of stress questionnaires, and variability in hair lengths. In addition, questionnaire-based assessments may be influenced by awareness about affective states, social desirability, and retrospection bias [38]. Retrospective self-perception of stress may also differ from stress reactivity elicited by contextual factors such as anticipation and social evaluative threat.

Our regression results corroborate the robustness of HCC against potential demographic and hair treatment confounders [10,12]. In a meta-analysis of human studies using HCC, Stalder et al [12] identified a significant positive association between HCC and age, male sex, and less frequent hair washing. Hair treatments were negatively associated with HCC. However, the

XSL•FO
RenderX

association with each of these variables was weak [12]. We similarly found a weak association of HCC with age, male sex, hair washing frequency, and use of hair treatments. We were unable to explore the association with race due to the small number of non-White participants in our study sample. In addition to demographic and hair treatment variables, medications may influence plasma cortisol. Similar to our results, they have not been found to be a significant inclusion factor in HCC [10,12].

Future Directions

Future, larger studies may benefit from the addition of biometric markers in guiding the assessment of stress-related health risk and management strategies. In addition, since cortisol is only one output of the complex HPA axis, the addition of an index of net glucocorticoid may have utility as an adjunct biomarker for chronic stress. More studies are needed to determine the types of psychological stressors and elements of cortisol production and regulation reflected in HCC and potentially identify a complement of neuroendocrine somatic stress biomarkers that are more apt to yield insight into the mechanism of stress-related disease vulnerability than HCC alone.

Conclusions

While this study was a preliminary exploration into the elements of stress perception and reactivity measured by hair cortisol, further studies are needed to determine the subjective and blood measured correlates measured by hair cortisol and assess its utility in a clinical setting.

Acknowledgments

We would like to thank the staff at Emory Executive Health for sample collections used in this study, Chad Robichaux (MPH), for his assistance with data management and extraction, and Dominique Munroe for formatting assistance.

Funding for this study was provided by the John and Mary Brock Discovery Fund.

Data Availability

The datasets generated during and/or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

MAM reports receiving funding for the US Department of Health and Human Services Health Resources and Services Administration; US Agency for Healthcare Research and Quality; Ardmore Institute of Health; Georgia State; and the Alzheimer's Association. SHB, DW, BP, AKS, AH, and DLR declare no conflicts of interest.

Multimedia Appendix 1
Informed consent.
[DOC File, 101 KB - formative_v9i1e63811_app1.doc]

References

1. Rice DP, Miller LS. Health economics and cost implications of anxiety and other mental disorders in the United States. Br J Psychiatry 1998 Jul;173(S34):4-9. [doi: 10.1192/S0007125000293458]
2. Trautmann S, Rehm J, Wittchen HU. The economic costs of mental disorders: Do our societies react appropriately to the burden of mental disorders? EMBO Rep 2016 Sep;17(9):1245-1249. [doi: 10.15252/embr.201642951] [Medline: 27491723]
3. Kalia M. Assessing the economic impact of stress--the modern day hidden epidemic. Metab Clin Exp 2002 Jun;51(6 Suppl 1):49-53. [doi: 10.1053/meta.2002.33193] [Medline: 12040542]
4. Murray CJL, Lopez AD. The Global Burden of Disease: A Comprehensive Assessment of Mortality and Disability from Diseases, Injuries, and Risk Factors in 1990 and Projected to 2020 ; Summary: Harvard School of Public Health; 1996.
5. Cohen S, Janicki-Deverts D, Miller GE. Psychological stress and disease. JAMA 2007 Oct 10;298(14):1685-1687. [doi: 10.1001/jama.298.14.1685] [Medline: 17925521]
6. Sapolsky RM, Romero LM, Munck AU. How do glucocorticoids influence stress responses? Integrating permissive, suppressive, stimulatory, and preparative actions. Endocr Rev 2000 Feb;21(1):55-89. [doi: 10.1210/edrv.21.1.0389] [Medline: 10696570]
7. Schneiderman N, Ironson G, Siegel SD. Stress and health: psychological, behavioral, and biological determinants. Annu Rev Clin Psychol 2005;1:607-628. [doi: 10.1146/annurev.clinpsy.1.102803.144141] [Medline: 17716101]
8. Epel ES, Crosswell AD, Mayer SE, et al. More than a feeling: a unified view of stress measurement for population science. Front Neuroendocrinol 2018 Apr;49:146-169. [doi: 10.1016/j.yfrne.2018.03.001] [Medline: 29551356]
9. Russell E, Koren G, Rieder M, Van Uum S. Hair cortisol as a biological marker of chronic stress: current status, future directions and unanswered questions. Psychoneuroendocrinology 2012 May;37(5):589-601. [doi: 10.1016/j.psyneuen.2011.09.009] [Medline: 21974976]

XSL•FO
RenderX

10. Wester VL, van Rossum EFC. Clinical applications of cortisol measurements in hair. Eur J Endocrinol 2015 Oct;173(4):M1-10. [doi: 10.1530/EJE-15-0313] [Medline: 25924811]
11. Stalder T, Kirschbaum C. Analysis of cortisol in hair--state of the art and future directions. Brain Behav Immun 2012 Oct;26(7):1019-1029. [doi: 10.1016/j.bbi.2012.02.002] [Medline: 22366690]
12. Stalder T, Steudte-Schmiedgen S, Alexander N, et al. Stress-related and basic determinants of hair cortisol in humans: a meta-analysis. Psychoneuroendocrinology 2017 Mar;77:261-274. [doi: 10.1016/j.psyneuen.2016.12.017] [Medline: 28135674]
13. Prado-Gascó V, de la Barrera U, Sancho-Castillo S, de la Rubia-Ortí JE, Montoya-Castilla I. Perceived stress and reference ranges of hair cortisol in healthy adolescents. PLoS One 2019;14(4):e0214856. [doi: 10.1371/journal.pone.0214856] [Medline: 30947316]
14. Lehrer HM, Steinhardt MA, Dubois SK, Laudenslager ML. Perceived stress, psychological resilience, hair cortisol concentration, and metabolic syndrome severity: a moderated mediation model. Psychoneuroendocrinology 2020 Mar;113:104510. [doi: 10.1016/j.psyneuen.2019.104510] [Medline: 31911349]
15. Charney DS. Psychobiological mechanisms of resilience and vulnerability: implications for successful adaptation to extreme stress. AJP 2004 Feb 1;161(2):195-216. [doi: 10.1176/appi.ajp.161.2.195]
16. Taylor MK. Dehydroepiandrosterone and dehydroepiandrosterone sulfate: anabolic, neuroprotective, and neuroexcitatory properties in military men. Mil Med 2013 Jan;178(1):100-106. [doi: 10.7205/milmed-d-12-00296] [Medline: 23356127]
17. Maninger N, Wolkowitz OM, Reus VI, Epel ES, Mellon SH. Neurobiological and neuropsychiatric effects of dehydroepiandrosterone (DHEA) and DHEA sulfate (DHEAS). Front Neuroendocrinol 2009 Jan;30(1):65-91. [doi: 10.1016/j.yfrne.2008.11.002] [Medline: 19063914]
18. Jin RO, Mason S, Mellon SH, et al. Cortisol/DHEA ratio and hippocampal volume: a pilot study in major depression and healthy controls. Psychoneuroendocrinology 2016 Oct;72:139-146. [doi: 10.1016/j.psyneuen.2016.06.017] [Medline: 27428086]
19. Morgan CA 3rd, Southwick S, Hazlett G, et al. Relationships among plasma dehydroepiandrosterone sulfate and cortisol levels, symptoms of dissociation, and objective performance in humans exposed to acute stress. Arch Gen Psychiatry 2004 Aug;61(8):819-825. [doi: 10.1001/archpsyc.61.8.819] [Medline: 15289280]
20. Hechter O, Grossman A, Chatterton RT. Relationship of dehydroepiandrosterone and cortisol in disease. Med Hypotheses 1997 Jul;49(1):85-91. [doi: 10.1016/s0306-9877(97)90258-9] [Medline: 9247914]
21. Cohen S, Janicki-Deverts D, Doyle WJ, et al. Chronic stress, glucocorticoid receptor resistance, inflammation, and disease risk. Proc Natl Acad Sci U S A 2012 Apr 17;109(16):5995-5999. [doi: 10.1073/pnas.1118355109] [Medline: 22474371]
22. Suarez EC, Sundy JS, Erkanli A. Depressogenic vulnerability and gender-specific patterns of neuro-immune dysregulation: what the ratio of cortisol to C-reactive protein can tell us about loss of normal regulatory control. Brain Behav Immun 2015 Feb;44:137-147. [doi: 10.1016/j.bbi.2014.09.008] [Medline: 25241020]
23. Qiao S, Li X, Zilioli S, et al. Hair measurements of cortisol, DHEA, and DHEA to cortisol ratio as biomarkers of chronic stress among people living with HIV in China: known-group validation. PLoS ONE 2017;12(1):e0169827. [doi: 10.1371/journal.pone.0169827] [Medline: 28095431]
24. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. J Health Soc Behav 1983 Dec;24(4):385-396. [Medline: 6668417]
25. Cohen S. Perceived stress in a probability sample of the united states. In: The Social Psychology of Health, in The Claremont Symposium on Applied Social Psychology: Sage Publications, Inc; 1988:31-67 URL: https://www.cmu.edu/dietrich/psychology/stress-immunity-disease-lab/scales/pdf/cohen,-s.--williamson,-g.-1988.pdf [accessed 2025-06-09]
26. Cohen S, Janicki‐deverts D. Who's stressed? Distributions of psychological stress in the United States in probability samples from 1983, 2006, and 2009 1. J Applied Social Pyschol 2012 Jun;42(6):1320-1334 [FREE Full text] [doi: 10.1111/j.1559-1816.2012.00900.x]
27. Connor KM, Davidson JRT. Development of a new resilience scale: the Connor-Davidson Resilience Scale (CD-RISC). Depress Anxiety 2003;18(2):76-82. [doi: 10.1002/da.10113] [Medline: 12964174]
28. Sauvé B, Koren G, Walsh G, Tokmakejian S, Van Uum SHM. Measurement of cortisol in human hair as a biomarker of systemic exposure. Clin Invest Med 2007;30(5):E183-E191. [doi: 10.25011/cim.v30i5.2894] [Medline: 17892760]
29. StataCorp, stata statistical software: release 16. : StataCorp LP, College Station, TX; 2019.
30. Short SJ, Stalder T, Marceau K, et al. Correspondence between hair cortisol concentrations and 30-day integrated daily salivary and weekly urinary cortisol measures. Psychoneuroendocrinology 2016 Sep;71:12-18. [doi: 10.1016/j.psyneuen.2016.05.007] [Medline: 27235635]
31. Blume J, Douglas SD, Evans DL. Immune suppression and immune activation in depression. Brain Behav Immun 2011 Feb;25(2):221-229. [doi: 10.1016/j.bbi.2010.10.008] [Medline: 20955778]
32. Sharpley CF, Bitsika V, McMillan ME, Jesulola E, Agnew LL. Does the cortisol: CRP ratio inform the measurement of individual burden of illness for depression in community samples? Journal of Affective Disorders Reports 2021 Jan;3:100058. [doi: 10.1016/j.jadr.2020.100058]

XSL•FO
RenderX

33. Bouras M, Roquilly A, Mahé PJ, et al. Cortisol total/CRP ratio for the prediction of hospital-acquired pneumonia and initiation of corticosteroid therapy in traumatic brain-injured patients. Crit Care 2019 Dec 5;23(1):394. [doi: 10.1186/s13054-019-2680-6] [Medline: 31805967]
34. Manenschijn L, Schaap L, van Schoor NM, et al. High long-term cortisol levels, measured in scalp hair, are associated with a history of cardiovascular disease. J Clin Endocrinol Metab 2013 May;98(5):2078-2083. [doi: 10.1210/jc.2012-3663] [Medline: 23596141]
35. Feller S, Vigl M, Bergmann MM, Boeing H, Kirschbaum C, Stalder T. Predictors of hair cortisol concentrations in older adults. Psychoneuroendocrinology 2014 Jan;39:132-140. [doi: 10.1016/j.psyneuen.2013.10.007] [Medline: 24275012]
36. Botschek T, Hußlein V, Peters EMJ, Brosig B. Hair cortisol as outcome parameter for psychological and neuropsychiatric interventions-a literature review. Front Psychiatry 2023;14:1227153. [doi: 10.3389/fpsyt.2023.1227153] [Medline: 37881597]
37. Staufenbiel SM, Penninx BWJH, Spijker AT, Elzinga BM, van Rossum EFC. Hair cortisol, stress exposure, and mental health in humans: a systematic review. Psychoneuroendocrinology 2013 Aug;38(8):1220-1235. [doi: 10.1016/j.psyneuen.2012.11.015] [Medline: 23253896]
38. Mauss IB, Levenson RW, McCarter L, Wilhelm FH, Gross JJ. The tie that binds? Coherence among emotion experience, behavior, and physiology. Emotion 2005 Jun;5(2):175-190. [doi: 10.1037/1528-3542.5.2.175] [Medline: 15982083]

Abbreviations:

CD-RISC: Connor Davidson Resilience Scale
DHEA: dehydroepiandrosterone
DHEA(S): collective term for dehydroepiandrosterone and dehydroepiandrosterone-sulfate
DHEA-S: dehydroepiandrosterone-sulfate
ELISA: enzyme-linked immunosorbent assay
HCC: hair cortisol concentration
HPA: hypothalamus-pituitary-adrenal
hs-CRP: high-sensitivity C-reactive protein
PSS: Perceived Stress Scale

Edited by A Mavragani; submitted 30.06.24; peer-reviewed by R Pickler; revised version received 13.03.25; accepted 16.03.25; published 11.06.25.

Please cite as:
Bergquist SH, Wang D, Pearce B, Smith AK, Hankus A, Roberts DL, Moore MA
Relationship of Hair Cortisol Concentration With Perceived and Somatic Stress Indices: Cross-Sectional Pilot Study
JMIR Form Res 2025;9:e63811
URL: https://formative.jmir.org/2025/1/e63811
doi:10.2196/63811

© Sharon H Bergquist, Danyang Wang, Brad Pearce, Alicia K Smith, Allison Hankus, David L Roberts, Miranda A Moore. Originally published in JMIR Formative Research (https://formative.jmir.org), 11.6.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Perceptions and Experiences of Developing Prediabetes Awareness and Participating in the National Diabetes Prevention Program Among Hispanic US Adults: Qualitative Study

Kevin L Joiner[1], PhD; Mackenzie Adams[2], MPH; Edith Beltrán Minehan[3], PhD; Matthew J O'Brien[4], MD, MSc

[1]School of Nursing, University of Michigan, 400 N Ingalls St, Ann Arbor, MI, United States

[2]School of Public Health, University of Michigan, Ann Arbor, MI, United States

[3]School of Academic Advancement, Madison Area Technical College, Madison, WI, United States

[4]Institute of Public Health and Medicine, Feinberg School of Medicine, Northwestern University, Evanston, IL, United States

Corresponding Author:
Kevin L Joiner, PhD
School of Nursing, University of Michigan, 400 N Ingalls St, Ann Arbor, MI, United States

Abstract

Background: Prediabetes conveys an increased risk for subsequently developing type 2 diabetes (T2D). The National Diabetes Prevention Program (DPP) is a widely available intensive behavioral intervention that decreases the risk of developing T2D in adults with prediabetes. Data are needed to inform approaches to increase prediabetes awareness and National DPP participation. Few studies have explored perceptions and experiences of prediabetes diagnosis and National DPP participation, and none have focused on Hispanic adults and participation in the National DPP as implemented by a community-based organization.

Objective: This study aims to explore perceptions and experiences of developing prediabetes awareness and participating in the National DPP among Hispanic US adults.

Methods: The sample was recruited from participants in the National DPP as implemented in Spanish by a community-based organization in the upper Midwest. Semistructured interviews were conducted by telephone in April and May 2021. A qualitative descriptive approach was used. Data from the interviews were reviewed, coded, and integrated into themes to reflect the narratives elicited in the interviews.

Results: A total of 16 interviews were conducted. The mean age of the participants was 46 (SD 6, range 34 - 55) years. Most (n=15) identified as female. The majority (n=15) reported having been born in Mexico. More than two-thirds (n=11) had a level of educational attainment of high school completion or less. Nearly half (n=7) reported not having health insurance. Qualitative description resulted in the emergence of four main themes: (1) processing the news of having prediabetes, (2) deciding on treatment for T2D primary prevention, (3) valuing language and cultural congruence in the National DPP, and (4) appreciating action-oriented knowledge gained during National DPP participation. Participants described the emotional impact of becoming aware of having prediabetes. National DPP lifestyle coaches' outreach and recruitment efforts on a local radio program and a Facebook Live (Meta Platforms) broadcast helped raise awareness of prediabetes and influence attitudes toward participation in the National DPP. Values and cultural beliefs appeared to contribute to perceptions and experiences of participating in the National DPP. Participants were inclined to share information about the National DPP with others in their community.

Conclusions: This study presents some of the first evidence exploring perceptions and experiences of developing prediabetes awareness and participating in the National DPP among Hispanic US adults. The findings can inform approaches to increase prediabetes awareness and National DPP participation among Hispanic US adults.

(JMIR Form Res 2025;9:e66964) doi:10.2196/66964

KEYWORDS

prediabetes; Hispanic; prevention; diabetes; diabetes prevention; awareness; older adult; elder; aging; diabetes prevention program; diabetic; type 2 diabetes; type 1 diabetes; endocrinology; diabetes mellitus; community-based; qualitative study; interview; United States

Introduction

Prediabetes is a state of elevated blood glucose that conveys an increased risk for subsequently developing type 2 diabetes (T2D) [1]. In the United States, approximately 97.6 million adults have prediabetes (38% of the entire population) [2]. Among adults with prediabetes, an estimated 29% will develop T2D within 3 years, and 70% will develop T2D within their lifetime without an effective strategy for primary prevention [3,4]. Lifestyle

XSL•FO
RenderX

modification is an effective primary prevention strategy, as demonstrated in a landmark clinical trial of adults with prediabetes and overweight or obesity. Participants randomized to receive the Diabetes Prevention Program (DPP) intensive lifestyle intervention experienced a 58% lower risk of developing T2D over 3 years than the placebo group [5].

Based on this promising evidence, the Centers for Disease Control and Prevention is leading an effort to offer the National DPP, which is derived from the evidence-based intensive lifestyle intervention [6]. The National DPP is implemented in-person and online across the country by community-based organizations, for-profit businesses, health care systems, and insurers [6,7]. The out-of-pocket cost of participating in the National DPP is covered through Medicare, commercial insurance, or employer wellness programs [6]. There are also organizations that enable participation without any out-of-pocket cost through support from philanthropic funds and grants [8].

Despite the high risk that prediabetes portends for subsequent development of T2D and the large-scale effort to provide access to an effective T2D primary prevention strategy, less than 1% of adults with prediabetes have participated in the National DPP [8]. Compared with White non-Hispanic adults with prediabetes, adults with prediabetes of underrepresented racial and ethnic groups are less likely to participate in the National DPP [7,8]. Hispanic adults have the lowest uptake rate into the National DPP among major racial and ethnic groups. Despite making up 15.4% of the US adult population with prediabetes, Hispanic adults only constitute 9% of individuals who have participated in the National DPP [2,8].

Racial and ethnic disparities in National DPP participation are likely multifactorial, stemming from patient-, provider-, and system-level factors. National survey data suggest that only about 20% of adults who meet the glycemic criteria for prediabetes are aware that they have the condition [2]. Even when adults are aware of having prediabetes, evidence suggests that many may not receive counseling from a health care provider about their subsequent risk for developing T2D and the preventive benefits participation in an intensive lifestyle intervention might offer [9]. Factors hindering effective provider-patient discussions about prediabetes and the National DPP include limited educational materials, time constraints during clinic visits, provider perceptions of potential financial constraints to participation, and provider perceptions of patient low self-efficacy for lifestyle change [10-14]. An analysis of 2016 and 2017 National Health Interview Survey data found that less than 5% of adults with diagnosed prediabetes received a referral to the National DPP [15].

Qualitative data from Hispanic adults with prediabetes who have participated in the National DPP are needed to understand their perceptions and experiences about becoming aware of having prediabetes, as well as making an informed choice about initiating and continuing participation in the National DPP. Previous research in the context of T2D prevention in Hispanic adults has explored perceptions regarding the risk of developing diabetes and preferences for prediabetes treatment options [16,17]. In addition, qualitative studies have provided insight into perceptions and experiences of participation in the National DPP as implemented by an integrated health care system among adults with diverse ethnic backgrounds [18,19]. Research in the setting of the National DPP implemented by a community-based organization can augment these earlier qualitative findings.

The objective of this study was to explore perceptions and experiences of prediabetes awareness and National DPP participation among Hispanic adults. The findings from this study will inform future interventions to increase prediabetes awareness and National DPP participation in a high-risk population while helping promote racial and ethnic equity in the prevention of T2D.

Methods

Overview

A qualitative descriptive approach was used [20]. This approach is intended to systematically summarize individual participants' specific and unique perspectives and experiences and report them in detailed descriptions that are closely linked to the words of the participants. Semistructured interviews were used to collect in-depth perspectives and experiences from the participants, allowing their responses to be considered within context [21]. Specifically, follow-up questions were included to elicit more detail about how perspectives and experiences might be influenced by language and cultural factors [22]. The research team was led by the first author, who is a nurse practitioner with a PhD in nursing and formal training and experience in qualitative methods. Additional team members included a research assistant with a PhD in Spanish and a research assistant with a master's degree in public health, both of whom had experience but no formal training in qualitative methods.

The sample was recruited from among current and completed participants of the National DPP, which was implemented in Spanish by a community-based organization, the National Kidney Foundation of Michigan, in Kent County, an urban county in West Michigan (April and May 2021). According to the US 2020 Census, the population of Kent County, Michigan, was approximately 650,000, with 11.7% of residents identifying as having Hispanic or Latino ethnicity [23].

The eligibility criteria were being 18 years or older, identifying as having Hispanic or Latino ethnicity, having a history of prediabetes, and participating in the National DPP. Participating in the National DPP was defined as attending at least one session. There were no exclusion criteria. One of the lifestyle coaches who facilitated the National DPP in Spanish, distributed study flyers in English and Spanish through email to National DPP participants. Those interested in learning more about the study completed an online screener in which they provided their contact information. Individuals who supplied their contact information in the screener were reached by telephone by the first author, who speaks fluent Spanish and identifies as being non-Hispanic White. The first author was unknown to the participants. Those who agreed to participate were subsequently enrolled in the study.

After obtaining verbal informed consent, the first author offered to conduct the semistructured interview in English or Spanish,

XSL•FO
RenderX

depending on the participant's preference. All of the participants chose to complete their interview in Spanish. Textbox 1 presents the semistructured interview guide in English. Sociodemographic and clinical data collected included age, gender identity, country of birth, English language proficiency, language spoken at home, educational attainment, insurance status, and annual household income.

Textbox 1. Interview guide.

Interview questions:

- When did you first learn you had prediabetes?
- Before your diagnosis, what did you know about prediabetes?
- What do you know about the consequences of prediabetes?
- What was your reaction when you were diagnosed?
- Tell me about what it was like when you were diagnosed. How did it make you feel?
- How did you learn about the National Diabetes Prevention Program (DPP)? Did you know someone who had joined?
- Which organization presented the National DPP that you participated in?
- How did your diagnosis influence your decision to participate in the National DPP?
- What made you want to join the National DPP? What gave you motivation?
- If you were to try to convince someone to join the program, what would you tell them?
- Did you encounter obstacles or barriers to participating? Did you have to make any sacrifices to be able to attend? Did you need to find childcare, miss work, or find transportation? How did you overcome these barriers?
- Was there any aspect of the program that made it easier for you to join?
- Which aspects of the National DPP did you like the most or find the most helpful?
- Which aspects of the National DPP did you not like or not find helpful?
- What do you think you learned or gained from the program?

A research assistant, a native Spanish speaker of Mexican heritage who identifies as having Hispanic ethnicity, transcribed the audio recordings of the interviews and translated them from Spanish into English. The research assistant, well-acquainted with the Spanish-language dialects used by the participants, took special care in the translation. The translated transcripts were reviewed by the first author and updated for accuracy when needed. Independently, the research assistant and the other research assistant, who identifies as having non-Hispanic White ethnicity and race, read all of the transcripts and created a list of possible codes with a description for each code. Based on the list of the codes and their descriptions, a codebook was developed by the 2 research assistants and the first author through consensus. The research assistants then independently applied the codes to text segments using ATLAS.ti (Scientific Software Development) computer-aided qualitative data analysis software [24]. The rationale for using computer-aided qualitative data analysis software was to assist in applying the codes and organizing and managing the textual data. As new concepts were identified inductively, additional codes were added to the codebook. After all the transcripts were coded, discrepancies in coding were identified and discussed by the research assistants and the first author. Any differences in coding were adjudicated by the first author and resolved through consensus. Using coding reports, patterns were identified, and a consensus was reached on the themes to be reported. During the analysis, data reduction tables were used to facilitate the iteration of concepts and the development of themes. In addition, audit trails were used to document analytical processes. Thematic saturation was achieved when additional evidence no longer changed the overall description of the themes. Standards for Reporting Qualitative Research endorsed by the EQUATOR (Enhancing the Quality and Transparency of Health Research) Network were followed [25].

Ethical Considerations

The study was approved by the institutional review Board of the University of Michigan (HUM00192113). Informed verbal consent was obtained from all participants before the interviews were conducted. All transcripts were deidentified during transcription, before analysis. Unique identifiers were used to anonymize quotes from individual participants. Each study participant received a US $20 Amazon gift card as compensation for their time spent participating.

Results

Overview

A total of 16 participant interviews were conducted. Table 1 presents the characteristics of the participants. The mean age was 46 (SD 6, range 34 - 55). Most (n=15) identified as female. The majority (n=15) reported being born in Mexico. More than two-thirds (n=11) had a level of educational attainment of high school completion or less. Nearly half (n=7) reported not having health insurance.

Qualitative description resulted in the emergence of four main themes: (1) processing the news of having prediabetes, (2) deciding on treatment for T2D primary prevention, (3) valuing

XSL•FO
RenderX

language and cultural congruence in the National DPP, and (4) appreciating action-oriented knowledge gained during National DPP participation.

Table . Participant demographics (N=16).

Characteristic	Values
Age (years), mean (SD; range)	46 (6; 34-55)
Gender, n	
Male	1
Female	15
Country of birth, n	
Mexico	15
Puerto Rico	1
English proficiency, n	
I don't speak English	3
Not well	11
Well	1
Very well	1
Language spoken at home, n	
Spanish	13
Spanish and English	3
Educational attainment, n	
Elementary school (grades 1 - 8)	2
High school (grades 9 - 12)	9
College or graduate school	5
Insurance status, n	
Private insurance	3
Public insurance	4
No insurance	7
Annual household income (US $), mean (range)	35,500 (17,000-96,000)

[a]Not applicable.

Theme 1: Processing the News of Having Prediabetes

Participants shared a range of unique perceptions and experiences upon becoming aware that they had prediabetes, often through specific events. Commonly, individuals were informed by their health care provider based on the results of a glycated hemoglobin (HbA_{1c}) screening test performed during a routine physical exam. Other participants reported developing awareness of having prediabetes based on a point-of-care HbA_{1c} test performed at a health fair. In one instance, a participant discovered her prediabetes following a pregnancy complicated by gestational diabetes:

> *My prediabetes started… well, 9 years ago. I was pregnant. My son is now 9 years old. Back then, they told me I had gestational diabetes. And after that, I went back to the doctor after giving birth and all of that. Then they told me that I was in the prediabetes [range]. And, I still have prediabetes now.* [Participant 16]

Participants who learned about their prediabetes during a routine physical exam or a visit to a health fair described the emotions they experienced while processing the news of having prediabetes. One woman described how she was caught off guard by the diagnosis of prediabetes during a routine physical examination:

> *I reacted with, well, surprise and also, a little bit of sadness and anger. Because one thinks they're okay until they get checked. All the time, every year, I get my physical, and before, it never turned out that way… that I had prediabetes.* [Participant 11]

For some participants, their diagnosis was a chance to begin understanding prediabetes and making lifestyle changes, including attempting to "eat healthier" and "exercise more." The words of one woman conveyed the sentiment of the participants:

XSL•FO
RenderX

> *I felt distressed. Sad. But this distress and this sadness, I focused it in a positive direction.* [Participant 6]

Several participants explained that the news of having prediabetes was followed by concern about their future quality of life if they were to develop T2D. Nearly all the participants had at least one close family member with T2D, which informed their views of what it would be like to live with diabetes. For example, one woman explained:

> *Yes, I knew [about T2D] because my mother-in-law died from it. And a lot of family on my husband's side [have T2D]. And I knew that they had to inject insulin and check their blood and things like that. I knew a little.* [Participant 6]

In one case, a participant reported that her health care provider outlined a potential future medication scenario for her if she went on to develop T2D:

> *He told me that I was prediabetic and that if I kept going, the next step would be diabetes, and that I would have to take the medicines that are necessary, and possibly, after that, I would have to take insulin. And that is when I said "No… I don't want that." I had to do something with myself so I would not get to the point of having diabetes.* [Participant 2]

Theme 2: Deciding on Treatment for T2D Primary Prevention

Many participants reported being concerned about the risk of subsequently developing T2D and considering possible paths to T2D primary prevention. The only options presented by the health care provider were typically losing weight through eating healthier meals and being more physically active or taking medication. When asked if the individual who had delivered their prediabetes diagnosis had offered the National DPP or any other support for achieving modest weight loss, a few participants reported that they were referred to a nutritionist. However, most participants reported not being offered any support for healthy lifestyle changes and weight loss. To many, at the time of their prediabetes diagnosis, it seemed that the only option was to try lifestyle changes on their own. One woman described her experience choosing between making lifestyle changes and taking medication:

> *I first went to my appointment, and they drew blood. Then, later, I came back for the results. It was then that the doctor told me what was going on and that I should try to get better, lose weight, and eat healthier. Because the doctor knows I am very reluctant to take medicines. I said: "No, I will do it on my own, lose weight, and do it.* [Participant 2]

One woman explained:

> *It was not my doctor [from who I heard about the National DPP]. My doctor would only say: "You need this [to eat healthier and lose weight]. You need that [to take a medication]." She never told me that there is a diabetes prevention program.* [Participant 1]

Participants expressed that the prospect of having to eat healthier and lose weight on their own, without support, seemed overwhelming. One woman shared her frustration:

> *I want to lose weight, but I cannot. Meanwhile, for example, during Holy Week, my clothes did not fit me, and I thought: What can I do? Sometimes, I do not have money to buy new clothes, and it's very unflattering.* [Participant 7]

Another woman stated:

> *You see that we want to lose weight, and we try to diet. But that only lasts, what, 15 days to a month.* [Participant 8]

The majority of participants learned about the National DPP through public health campaigns after being diagnosed with prediabetes. For example, most participants reported finding out about the National DPP through a local Spanish-language program on the radio or a local broadcast on Facebook Live in which a media personality spoke with a local lifestyle coach. What was immediately appealing to many study participants was how the lifestyle coach prepared them for initiating and participating in the National DPP. The lifestyle coach discussed information about the National DPP that was essential so that they could make an informed choice about participation, including that it would teach them about topics such as "how to eat healthy food" and "how to count calories." The lifestyle coach also discussed how the National DPP was a free, group-based program. One woman recalled her experience finding out about the National DPP:

> *I heard about it [the National DPP] on the radio. It was on the program La Poderosa. She [the local lifestyle coach] was saying that peoples' bodies are becoming larger because we have bad habits. And … what caught my attention was that it [T2D] is caused by the decisions that people make … and that people need counseling, some motivation. And I said: "Oh, that sounds interesting!" … I wrote down the number, called, and registered myself.* [Participant 6]

Another woman described her experience finding out about the program:

> *I responded to the video that she [the lifestyle coach] made on El Informador [Facebook Live program]. She said that she was going to teach us how to take care of ourselves better so that we wouldn't develop it [T2D]. And we would be changing our lifestyle and eating healthier. And that she would support us.* [Participant 8]

In addition, a couple of participants who became aware that they had prediabetes at a health fair reported that the National DPP had been offered to them as an opportunity for T2D primary prevention.

Theme 3: Valuing Language and Cultural Congruence in the National DPP

Participants reported that an important factor in their engagement with the National DPP was the linguistic and cultural tailoring for Hispanic adults. The fact that the program

XSL•FO
RenderX

was in Spanish was described as essential. In addition to the language, participants expressed their appreciation that the program was culturally salient in other ways. One participant shared that the National DPP offered her access to information that she felt that she was missing from visits with her health care provider. She expressed that this was especially important since many Hispanic women like herself may not visit a health care provider to manage their health. She explained:

> *To tell you the truth, as a Hispanic, you take things differently. One does not go to the doctor just because. You go to the doctor when you are sick and when you feel bad. You do not go to get a physical or anything. So, I learned a lot in it [the National DPP] about what you should look for in the labels. What you must do and all that, I learned it there.* [Participant 11]

When considering how being Hispanic might present unique challenges or opportunities to engage with the National DPP, one woman identified her country of origin as impacting her relationship with food choices. She explained:

> *I'm Mexican, and we are used to eating… no not eating, ingesting, too much. We're accustomed to beans, rice, meat… steak, chicken, hamburgers, or whatever it may be, …. and always tortillas. We ingest a lot of carbohydrates. … And one eats and eats, and one doesn't pay attention to how much one eats. What I learned from the women [the lifestyle coaches] has been of great importance. I've learned a lot. Thank God.* [Participant 12]

Some participants perceived that being an immigrant created challenges related to gaining weight and eating behaviors. One woman explained:

> *I should not gain any more weight because it is very easy here [in the US] to gain weight. That was the only thing I thought. That I won't gain weight.* [Participant 14]

Another woman explained:

> *Because the problems of Hispanic people are different from the problems of people who were born here. Having left your country, the pain of leaving your parents, your culture, all of that, is painful, and that also causes depression and not eating appropriately.* [Participant 6]

Theme 4: Appreciating Action-Oriented Knowledge Gained During National DPP Participation

Many participants appreciated the action-oriented knowledge they gained during National DPP participation, which promoted their attempts at healthy lifestyle change. One woman explained:

> *From the beginning, when we started with the teacher, she taught us that we need to be detectives of nutrients and food labels. I had never paid attention to that before. I had tried to look for things with less sugar and stuff, but never stopped to look at ingredient by ingredient and the amount of fat. This is what we also studied in class. I tell you, for me, that was something new… She also taught us about exercise, and she showed us what we could do at home.* [Participant 7]

Another woman stated:

> *I've seen nutritionists before, and I was the type of person who wanted to be told, 'Ok, in the morning, you will eat this; at noon, you will eat that; and at night, you will eat this.' But no one tells you that … She [the lifestyle coach] is teaching us how to read food labels to find out about calories, sugars, and fats.* [Participant 8]

A participant recalled the early challenges of learning about prediabetes and reported that understanding biological mechanisms offered in the National DPP enhanced his motivation to make healthy lifestyle changes. He noted:

> *For me, at first, it was very complicated. But then after that, when relating how one eats to how sugars rise, and knowing that the pancreas regulates sugars, it's then when one starts to understand why, before eating and after eating, one's sugar levels are different.* [Participant 11]

Participants discussed how they perceived that they were able to carry over the focus on the empowerment of the National DPP to other family members.

> *To tell you the truth, it [the National DPP] did help me because I have … small children. … I do not want what happened to me to happen to them. I mean, it [the National DPP] helps you. It helps you. It helps make your family better. They [the lifestyle coaches] help you to take care of your family.* [Participant 3]

Some of the participants seemed to feel confident sharing the knowledge that they gained from the National DPP with others in the community. One woman recounted:

> *In my experience, I have gotten to know more people and told them, when they say "prediabetes," you have to take it seriously…and try to change your way of living and lifestyle.* [Participant 3]

Another woman explained:

> *There should be more, like, more programs. Like more publicity. Give us more information.… They should give programs like this [the National DPP] more promotion. Also, tell people more so they know about it. Because the truth is, one doesn't know about them. Look after this, the truth is that I sent a lot of people there. I told them: "Go. Go. It's very important."* [Participant 1]

Discussion

Principal Findings

This qualitative study presents some of the first evidence exploring perceptions and experiences of developing prediabetes awareness and participating in the National DPP among Hispanic US adults. The findings emphasize the emotional impact of becoming aware of having prediabetes and highlight the significance of dynamic outreach and recruitment by

XSL•FO
RenderX

National DPP lifestyle coaches. Outreach by National DPP lifestyle coaches, who leveraged local community forums, including a local radio program and a Facebook Live broadcast, was pivotal in raising awareness of prediabetes and influencing attitudes toward participation in the National DPP. Values and cultural beliefs among US Hispanic communities appeared to contribute to perceptions and experiences of participating in the National DPP. Some participants expressed a strong inclination to share information about the National DPP with others in their community.

Participants vividly remembered the moment they found out that they had prediabetes and the emotions evoked by the diagnosis. Emotions such as surprise, distress, and concern about their future quality of life were commonly reported. This resonates with findings in other communities where similar emotional reactions to developing prediabetes awareness have been documented [26]. The familiarity with T2D and the way T2D had disrupted the lives of close family members and friends was also consistently and clearly communicated, which complements previous research on culturally relevant issues among Hispanic adults in the United States [27]. When reflecting on their views before joining the National DPP and how participating had affected them, the participants highlighted a previous lack of knowledge regarding prediabetes and a previous feeling of lack of control over their trajectory toward developing T2D. For many, the diagnosis of prediabetes served as a catalyst for participating in the National DPP and taking proactive measures such as healthier eating and increased physical activity. These findings highlight the importance of providing clear, easily understood, complete information to enhance individuals' optimism regarding the risk that prediabetes portends for developing T2D and the effectiveness of the National DPP in mitigating this risk.

Guidelines recommend primary care providers play an essential role in T2D primary prevention through prediabetes screening, counseling, and support for initiating National DPP participation [28,29]. However, the findings suggest that primary care providers often fail to effectively counsel Hispanic adults with prediabetes about their risk for developing T2D and the benefits of the National DPP [9,15]. This gap could be attributed to limited awareness among primary care providers about the National DPP and the community-based organizations implementing it [13,14]. Participants echoed this concern, indicating that many did not learn about the National DPP from their health care providers but through other channels.

The role of trusted community figures and effective communication through local media emerged as a key factor in participants' decisions to join the National DPP. Information disseminated through appearances by lifestyle coaches on a local Spanish-language radio program, and a Facebook Live broadcast provided essential insights into the National DPP, making it accessible and appealing. This aligns with previous research that highlights the positive impact of social media on health promotion and National DPP recruitment, particularly within Hispanic communities where the use of social media is prevalent [30,31].

Participants valued the National DPP implemented by the community-based organization for its linguistic and cultural congruence, noting that it resonated with their Hispanic heritage. Cultural factors like traditional diets and the challenges faced by immigrants in adopting healthier lifestyles were common themes. The National DPP's emphasis on action-oriented knowledge, teaching participants about reading food labels, and understanding the biological mechanisms of prediabetes, empowered them to make informed health choices. This sense of control that they gained extended beyond their own personal health, influencing family members and potentially enhancing health consciousness in the community.

Participants' appreciation of the National DPP extended to its broader impacts on family and community. Promotion through family and friends was also highlighted as a valuable channel for reaching more people, including men who might be hesitant about participating or facilitating their spouse's participation. Participants suggested that more supportive efforts and more comprehensive promotion could help inform others in their community about prediabetes and combat fatalism, which was viewed as a cultural belief that might have an impact on how treatments such as the National DPP for primary prevention of T2D are viewed [32].

Offering the National DPP at no out-of-pocket cost helped to alleviate socioeconomic barriers, making the program accessible to those with fewer resources [33]. Lifestyle coaches marketing directly to adults with prediabetes could help overcome potential financial concerns that health care providers may have regarding patient National DPP participation [34]. This approach holds promise for reaching underserved communities and warrants further exploration of coordinated interventions between health care systems and community outreach.

Limitations

This study has several limitations. The Hispanic adults with prediabetes who took part in the study had participated in the National DPP as implemented by a community-based organization, which was in Spanish and culturally tailored for the Hispanic community. Therefore, this study does not explore perceptions and experiences of Hispanic adults with prediabetes when the National DPP is not accessible in the community in a linguistically and culturally tailored manner. The community-based organization was supported by philanthropic and grant funding, which allowed Hispanic adults with prediabetes to participate in the National DPP without any out-of-pocket cost. This study included an overrepresentation of women and individuals who identify as Mexican American compared with the overall Hispanic adult population in the United States [35]. The overwhelming majority of all National DPP participants (75%) nationwide are women [8]. Notwithstanding these limitations, the qualitative description offers clinicians, researchers, and policymakers, insight into the perceptions and experiences of an at-risk population currently underserved by the National DPP, upon which future studies may expand. Further research in this area should consider sampling from a larger number of National DPP sites in geographically dispersed areas. Expanded representation of adults of diverse racial, ethnic, and cultural backgrounds, as

XSL•FO
RenderX

well as male adults, who likely have unique needs and challenges related to being reached and enrolled in the National DPP, are also opportunities for future research.

Future Directions

A community-based intervention for Hispanic adults that strengthens prediabetes awareness and provides counseling about participating in the National DPP may increase uptake into the National DPP. The American Diabetes Association and the Centers for Disease Control and Prevention advocate for patient-centered approaches in primary care to strengthen awareness of prediabetes and support decisions about National DPP participation [6,29]. The potential for local communication channels, including radio programs and social media broadcasts, to amplify health care services' National DPP recruitment and enrollment efforts at scale is also important [36]. In total, 66% of Hispanic adults use Facebook [31,37]. Having an interview with a National DPP lifestyle coach on a radio program or a broadcast on Facebook Live leverages community forums and requires much less investment of resources than health care services calling many people to ask them to join the National DPP [38]. Interventions that foster joint efforts through multiple communication channels have the potential to make the path from prediabetes diagnosis to National DPP participation smoother and more effortless and reach underrepresented racial and ethnic groups [30].

Conclusions

Efforts to increase National DPP participation among Hispanic adults with prediabetes should be tailored to these populations' unique social, cultural, and linguistic needs. This study is helpful for primary care providers, community organizations, and researchers aiming to boost prediabetes awareness and National DPP participation. Participants in this study described their perceptions and experiences of being diagnosed with prediabetes and what they gained from participating in the National DPP, which provides insights for future efforts to advance T2D primary prevention. Overall, participants reported finding the National DPP valuable and well-suited to their needs. First, the enrollment information should be a multipronged effort, including coming through respected, vibrant communication forums that serve the needs of the community. Second, a National DPP lifestyle coach who is a community member would be trusted easily; however, if a lifestyle coach would like to be helpful to a Hispanic community, there must be an awareness of the cultural, social, and linguistic issues pertaining to this community. Finally, it is important to note that the empowerment that the National DPP brings to individuals and families has its core in the community. Empowerment comes from managing the diagnosis of prediabetes with lifestyle changes that affect and impact family and loved ones. Hispanic adults with prediabetes who participate in the National DPP don't just decrease their own risk for developing T2D; they also become passionate and engaged leaders who have the potential to improve others' lives in the community.

Acknowledgments

This study would not have been possible without the support of the National Kidney Foundation of Michigan. The authors acknowledge the use of ChatGPT for editing and refining language in manuscript preparation and take full responsibility for the content and interpretations.

Data Availability

The dataset generated and analyzed during this qualitative research study is not publicly available due to concerns about participant privacy and the sensitive nature of the information collected.

Conflicts of Interest

None declared.

Checklist 1
Standards for Reporting Qualitative Research (SRQR) checklist.
[DOCX File, 22 KB - formative_v9i1e66964_app1.docx]

References

1. Tabák AG, Herder C, Rathmann W, Brunner EJ, Kivimäki M. Prediabetes: A high-risk state for diabetes development. Lancet 2012 Jun 16;379(9833):2279-2290. [doi: 10.1016/S0140-6736(12)60283-9] [Medline: 22683128]
2. National Diabetes Statistics Report 2024. Centers for Disease Control and Prevention. 2024. URL: https://www.cdc.gov/diabetes/php/data-research/index.html [accessed 2024-09-27]
3. Nathan DM, Davidson MB, DeFronzo RA, et al. Impaired fasting glucose and impaired glucose tolerance: implications for care. Diabetes Care 2007 Mar;30(3):753-759. [doi: 10.2337/dc07-9920] [Medline: 17327355]
4. Gerstein HC, Santaguida P, Raina P, et al. Annual incidence and relative risk of diabetes in people with various categories of dysglycemia: A systematic overview and meta-analysis of prospective studies. Diabetes Res Clin Pract 2007 Dec;78(3):305-312. [doi: 10.1016/j.diabres.2007.05.004] [Medline: 17601626]

XSL•FO
RenderX

5. Knowler WC, Barrett-Connor E, Fowler SE, et al. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. N Engl J Med 2002 Feb 7;346(6):393-403. [doi: 10.1056/NEJMoa012512] [Medline: 11832527]
6. National Diabetes Prevention Program. Centers for Disease Control and Prevention. 2024. URL: https://www.cdc.gov/diabetes-prevention/index.html [accessed 2024-07-08]
7. Cannon MJ, Ng BP, Lloyd K, Reynolds J, Ely EK. Delivering the National Diabetes Prevention Program: Assessment of enrollment in in-person and virtual organizations. J Diabetes Res 2022;2022:2942918. [doi: 10.1155/2022/2942918] [Medline: 35118160]
8. Gruss SM, Nhim K, Gregg E, Bell M, Luman E, Albright A. Public health approaches to Type 2 diabetes prevention: the US National Diabetes Prevention Program and beyond. Curr Diab Rep 2019 Aug 5;19(9):78. [doi: 10.1007/s11892-019-1200-z] [Medline: 31385061]
9. Cunningham A, Li E, Silverio A, Mills G, Han J, Waters A. Patient and provider prediabetes knowledge, attitudes, and behavior in a large urban family medicine practice. Ann Fam Med 2022 Apr 1;20(20 Suppl 1):2748. [doi: 10.1370/afm.20.s1.2748] [Medline: 36857128]
10. Madrigal L, Haardörfer R, Kegler MC, et al. A structural equation model of CFIR inner and outer setting constructs, organization characteristics, and national DPP enrollment. Implement Sci Commun 2023 Nov 17;4(1):142. [doi: 10.1186/s43058-023-00522-3] [Medline: 37978574]
11. Papalii M, Ely E. 333-OR: What motivates people to enroll in the National Diabetes Prevention Program Lifestyle Change Program. Diabetes 2023 Jun 20;72(Supplement_1). [doi: 10.2337/db23-333-OR]
12. AuYoung M, Moin T, Richardson CR, Damschroder LJ. The Diabetes Prevention Program for Underserved Populations: A brief review of strategies in the real world. Diabetes Spectr 2019 Nov;32(4):312-317. [doi: 10.2337/ds19-0007] [Medline: 31798288]
13. Tseng E, Greer RC, O'Rourke P, et al. National Survey of Primary Care Physicians' Knowledge, Practices, and Perceptions of Prediabetes. J Gen Intern Med 2019 Nov;34(11):2475-2481. [doi: 10.1007/s11606-019-05245-7] [Medline: 31502095]
14. Tseng E, Greer RC, O'Rourke P, et al. Survey of primary care providers' knowledge of screening for, diagnosing and managing prediabetes. J Gen Intern Med 2017 Nov;32(11):1172-1178. [doi: 10.1007/s11606-017-4103-1] [Medline: 28730532]
15. Ali MK, McKeever Bullard K, Imperatore G, et al. Reach and use of Diabetes Prevention Services in the United States, 2016-2017. JAMA Netw Open 2019 May 3;2(5):e193160. [doi: 10.1001/jamanetworkopen.2019.3160]
16. Joiner KL, Sternberg RM, Kennedy CM, Fukuoka Y, Chen JL, Janson SL. Perception of risk for developing diabetes among foreign-born Spanish-speaking US Latinos. Diabetes Educ 2016 Aug;42(4):418-428. [doi: 10.1177/0145721716646204] [Medline: 27150605]
17. O'Brien MJ, Moran MR, Tang JW, et al. Patient perceptions about prediabetes and preferences for diabetes prevention. Diabetes Educ 2016 Dec;42(6):667-677. [doi: 10.1177/0145721716666678] [Medline: 27621093]
18. Ritchie ND, Phimphasone-Brady P, Sauder KA, Amura CR. Perceived barriers and potential solutions to engagement in the National Diabetes Prevention Program. ADCES in Practice 2021 Jan;9(1):16-20. [doi: 10.1177/2633559X20966275]
19. Harrison CR, Phimphasone-Brady P, DiOrio B, et al. Barriers and facilitators of National Diabetes Prevention Program engagement among women of childbearing age: a qualitative study. Diabetes Educ 2020 Jun;46(3):279-288. [doi: 10.1177/0145721720920252] [Medline: 32597384]
20. Sandelowski M. Whatever happened to qualitative description? Res Nurs Health 2000 Aug;23(4):334-340. [doi: 10.1002/1098-240X(200008)23:4<334::AID-NUR9>3.0.CO;2-G]
21. DeJonckheere M, Vaughn LM. Semistructured interviewing in primary care research: A balance of relationship and rigour. Fam Med Com Health 2019 Mar;7(2):e000057. [doi: 10.1136/fmch-2018-000057]
22. Hispanic Americans' trust in and engagement with science. Pew Research Center. 2022. URL: https://www.pewresearch.org/science/2022/06/14/hispanic-americans-trust-in-and-engagement-with-science/ [accessed 2024-09-27]
23. U.S. Census Bureau. Quick Facts: Kent County, Michigan. 2024. URL: https://www.census.gov/quickfacts/fact/table/kentcountymichigan/PST04022424 [accessed 2025-02-03]
24. Friese S. Qualitative Data Analysis with ATLAS.ti, 3rd edition: SAGE Publications Ltd; 2019.
25. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014 Sep;89(9):1245-1251. [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]
26. Skoglund G, Nilsson BB, Olsen CF, Bergland A, Hilde G. Facilitators and barriers for lifestyle change in people with prediabetes: A meta-synthesis of qualitative studies. BMC Public Health 2022 Mar 21;22(1):553. [doi: 10.1186/s12889-022-12885-8] [Medline: 35313859]
27. Caban A, Walker EA. A systematic review of research on culturally relevant issues for Hispanics with diabetes. Diabetes Educ 2006;32(4):584-595. [doi: 10.1177/0145721706290435] [Medline: 16873596]
28. Pronk NP, Remington PL, Community Preventive Services Task Force. Combined diet and physical activity promotion programs for prevention of diabetes: Community preventive services task force recommendation statement. Ann Intern Med 2015 Sep 15;163(6):465-468. [doi: 10.7326/M15-1029] [Medline: 26168073]
29. ElSayed NA, Aleppo G, Bannuru RR, et al. 3. Prevention or Delay of Diabetes and Associated Comorbidities: Standards of Care in Diabetes—2024. Diabetes Care 2024 Jan 1;47(Supplement_1):S43-S51. [doi: 10.2337/dc24-S003]

XSL•FO
RenderX

30. Jalkanen K, Järvenpää R, Tilles-Tirkkonen T, et al. Comparison of communication channels for large-scale type 2 diabetes risk screening and intervention recruitment: empirical study. JMIR Diabetes 2021 Sep 9;6(3):e21356. [doi: 10.2196/21356] [Medline: 34499036]
31. Americans' social media use. Pew Research Center. 2024. URL: https://www.pewresearch.org/internet/2024/01/31/americans-social-media-use/ [accessed 2024-09-27]
32. Moreira T, Hernandez DC, Scott CW, Murillo R, Vaughan EM, Johnston CA. Susto, Coraje, y Fatalismo: Cultural-bound beliefs and the treatment of diabetes among socioeconomically disadvantaged Hispanics. Am J Lifestyle Med 2018 Jan;12(1):30-33. [doi: 10.1177/1559827617736506] [Medline: 29541008]
33. Improving access to diabetes prevention. Center for Disease Control and Prevention. 2024. URL: https://www.cdc.gov/diabetes/health-equity/improving-access-type-2-diabetes-prevention-program.html [accessed 2024-08-05]
34. Thomas TW, Golin CE, Kinlaw AC, et al. Did the 2015 USPSTF abnormal blood glucose recommendations change clinician attitudes or behaviors? A mixed-method assessment. J Gen Intern Med 2022 Jan;37(1):15-22. [doi: 10.1007/s11606-021-06749-x] [Medline: 33826060]
35. Hispanic/Latino health. U.S. Department of Health and Human Services Office of Minority Health. 2024. URL: https://minorityhealth.hhs.gov/hispaniclatino-health [accessed 2024-07-17]
36. Mora J, Romo R, Dempsey S, et al. Engaging the community served: A U.S. Cancer Center'S Facebook live cancer awareness campaign for Spanish-speaking Latinos during COVID-19. Cancer Causes Control 2023 Dec;34(12):1037-1042. [doi: 10.1007/s10552-023-01755-2] [Medline: 37490139]
37. Social media use in 2021. Pew Research Center. 2021. URL: https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/ [accessed 2024-09-27]
38. Philis-Tsimikas A, Fortmann A, Lleva-Ocana L, Walker C, Gallo LC. Peer-led diabetes education programs in high-risk Mexican Americans improve glycemic control compared with standard approaches: A Project Dulce promotora randomized trial. Diabetes Care 2011 Sep;34(9):1926-1931. [doi: 10.2337/dc10-2081] [Medline: 21775748]

Abbreviations

DPP: Diabetes Prevention Program
EQUATOR: Enhancing the Quality and Transparency of Health Research
HbA_{1c}: glycated hemoglobin
T2D: type 2 diabetes

Edited by A Mavragani; submitted 27.09.24; peer-reviewed by B Gibson, TB Loeb; revised version received 26.12.24; accepted 27.12.24; published 11.02.25.

Please cite as:
Joiner KL, Adams M, Minehan EB, O'Brien MJ
Perceptions and Experiences of Developing Prediabetes Awareness and Participating in the National Diabetes Prevention Program Among Hispanic US Adults: Qualitative Study
JMIR Form Res 2025;9:e66964
URL: https://formative.jmir.org/2025/1/e66964
doi:10.2196/66964

© Kevin L Joiner, Mackenzie Adams, Edith Beltrán Minehan, Matthew J O'Brien. Originally published in JMIR Formative Research (https://formative.jmir.org), 11.2.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Disease Awareness in Patients With Type 2 Diabetes: Analysis of Baseline Data From the SMART-Finder Observational Study

Christian Mueller[1], PhD; Thomas Neusser[1], PhD; Inga Thate-Waschke[1], PhD; Julia Nowicki[2], BSc; Tomasz Plominski[2], MSc; Regine Griesinger[3]; Stefanie Kessner[3]; Stephan Martin[4], MD

[1]Pharmaceuticals Medizin, Pharmaceuticals, BAYER Vital GmbH, K56, Leverkusen, Germany

[2]smartpatient GmbH, Munich, Germany

[3]Emmes Biopharma Germany GmbH, Munich, Germany

[4]Verbund Katholischer Kliniken Duesseldorf (VKKD) und Westdeutsches Diabetes- und Gesundheitszentrum (WDGZ), Duesseldorf, Germany

Corresponding Author:
Christian Mueller, PhD
Pharmaceuticals Medizin, Pharmaceuticals, BAYER Vital GmbH, K56, Leverkusen, Germany

Abstract

Background: Chronic kidney disease (CKD) is a common comorbidity of type 2 diabetes mellitus (T2DM). Data on the determination of CKD-related biomarkers among patients with T2DM in a real-life setting within Germany are limited.

Objective: We aimed to determine the prevalence of CKD and risk factors, availability of urine albumin-to-creatinine ratio (UACR) and estimated glomerular filtration rate (eGFR) values, treatment satisfaction, and quality of life among patients with T2DM in Germany.

Methods: SMART-Finder is a retrospective and prospective, observational, digital, patient-centered cohort study being performed as part of the routine use of an adherence-supporting app. This baseline analysis' observation period was from August to November 2023. Patients with T2DM in Germany who actively used the MyTherapy app; allowed push notifications; and documented use of diabetes medications, renin-angiotensin system inhibitors, finerenone, and/or blood glucose test strips were eligible for inclusion. Study materials (background information, electronic consent form, and laboratory and electronic questionnaires) were provided to eligible patients via app push notifications. Participants completed an electronic case report form that included questions on their blood pressure; their most recent UACR, eGFR, and glycated hemoglobin (HbA_{1c}) values in the past 12 months; the EQ-5D-5L; and the Diabetes Treatment Satisfaction Questionnaire. The primary outcome was the proportion of patients with a UACR of ≥30 mg/g.

Results: Of 9527 invited eligible patients, 101 completed the electronic case report form (male: n=61; female: n=40; age: mean 54.2, SD 11.4 y). Of these, 1 female patient and 5 male patients reported their UACR values; 3 (all male) had a UACR of ≥30 mg/g. The remaining 95 patients reported that their health care professionals had not provided UACR measurements. Only 9 (8.9%) patients were aware of their latest eGFR values (3 patients: 15 - 44 mL/min/1.73 m^2; 6 patients: 45 - 89 mL/min/1.73 m^2), 90 provided HbA_{1c} values (80 patients: ≥6.0%), 46 had a systolic blood pressure of ≥130 mm Hg, and 83 reported former or current nephrotoxic medication intake. The mean EQ-5D-5L index score was 0.7 (SD 0.3; range –0.1 to 1.0; 50 patients). The mean Diabetes Treatment Satisfaction Questionnaire score was 28.8 (SD 6.8; range 9.0-36.0; 49 patients).

Conclusions: Patients with T2DM who were using an adherence-supporting app in Germany lacked awareness of CKD-related biomarkers but had high knowledge of self-manageable biomarkers (eg, blood pressure, serum fasting glucose, and HbA_{1c} values). Our results suggest that treating physicians either do not test for UACRs and eGFRs or do not inform patients about the results. Nonadherence to diagnostic testing guidelines and a lack of physician-patient communication put patients at risk. Another reason for this health literacy imbalance may be the focus on HbA_{1c} instead of kidney comorbidity in patient education material. Future goals for diabetes management must include guideline-compliant testing of CKD-related biomarkers and open physician-patient communication.

International Registered Report Identifier (IRRID): RR2-10.2196/44996

(JMIR Form Res 2025;9:e60246) doi:10.2196/60246

KEYWORDS

app-based documentation; chronic kidney disease; CKD; disease awareness; MyTherapy; type 2 diabetes; type 2 diabetes mellitus; urine albumin-to-creatinine ratio screening; patient; observational study; Germany; quality of life; treatment; therapy; physician-patient communication

XSL•FO
RenderX

Introduction

Chronic kidney disease (CKD) is a common comorbidity of type 2 diabetes mellitus (T2DM), which is a leading cause of death and disability worldwide [1-8]. Patients with CKD show persistently decreased estimated glomerular filtration rates (eGFRs) or persistently elevated urine albumin excretion [9-11], which may progress to end-stage renal disease [12]. A high-normal urine albumin-to-creatinine ratio (UACR) is associated with a significantly increased risk of all-cause mortality [13].

Effective therapeutic approaches that can influence the development of CKD in patients with T2DM are available [14-16]. The current consensus guideline of the American Diabetes Association (ADA) and the Kidney Disease: Improving Global Outcomes (KDIGO) Work Group recommends renin-angiotensin system inhibitors, sodium-glucose cotransporter 2 inhibitors (SGLT2is), statins, and metformin as first-line drug therapies for patients with T2DM and CKD [17]. Glucagon-like peptide-1 receptor agonists are recommended if SGLT2is and metformin do not achieve sufficient glycemic control. For patients with persistent albuminuria despite standard treatment with glucose-lowering and antihypertensive medications, a nonsteroidal mineralocorticoid receptor antagonist is recommended in addition to baseline therapy [17].

The early identification of patients with diabetes who are at increased risk for renal disease is highly recommended [18,19]. Therefore, the eGFR (based on serum creatinine) and albuminuria (based on the UACR) should be assessed once per year in patients with T2DM [20,21]. As the UACR demonstrates a high degree of within-individual variability among individuals, regularly collecting samples for UACR testing may improve capacity for monitoring changes over time in clinical and research settings [22]. However, data on the determination of UACRs in a real-life setting and the prevalence of CKD among patients with T2DM in Germany are limited [23-26].

The SMART-Finder study was designed to determine the prevalence of CKD and risk factors, availability of UACR and eGFR values, satisfaction with treatment, and quality of life among patients with T2DM using an adherence-supporting app in Germany [27]. In this analysis of baseline data from the SMART-Finder study, the primary objective was to assess the proportion of patients with T2DM who had an elevated UACR (≥30 mg/g; albumin-to-creatinine ratio [ACR] stages A2 and A3). As secondary objectives, the proportion of patients without UACR test data was determined, blood pressure and nephrotoxic comedication status were assessed, and quality of life was analyzed. Disease awareness and treatment satisfaction were assessed as further secondary objectives.

Methods

Study Design

As previously described [27], the recruitment of patients and documentation of data for this retrospective and prospective, observational, digital, patient-centered cohort study were carried out without the assistance of a health care professional (HCP), as part of the routine use of the MyTherapy app (smartpatient GmbH) by patients with T2DM in Germany. The app helps patients manage their treatment-related tasks and access information on other health topics, including comorbidities, by using their mobile device. The app provides the user with a daily "To do" list of tasks, including those regarding medication intake, and asks the user to confirm whether each task has been completed [27]. Patients who documented the following medications or medical devices within the app were considered eligible for participation during initial selection: metformin, acarbose, dipeptidyl-peptidase-4 inhibitor, glucagon-like peptide-1 receptor agonist, renin-angiotensin system inhibitor, SGLT2i, basal insulin, nonsteroidal mineralocorticoid receptor antagonist, sulfonylurea, and blood glucose test strips.

To ensure that patients were properly informed, all eligible patients who actively used the app and allowed push notifications during the app's regular operation received content via push notifications during an awareness phase, which lasted from August 2 to 15, 2023. This content comprised 3 videos about the scope and goals of the SMART-Finder study, the risk of CKD in patients with T2DM, and the impairment of kidney function in patients with CKD; a study description page; a document for collecting laboratory values; and an electronic consent form for explaining, in detail, which data were used for this study and obtaining consent for the use of pseudonymized data. After completion of the consent form, an electronic case report form (eCRF), the EQ-5D-5L, and the Diabetes Treatment Satisfaction Questionnaire (DTSQ) were provided in the "Discover" section of the app (patients were notified about this via push notifications as well) on August 18, 21, and 24, 2023, respectively. The eCRF included questions on patients' residential region, age, gender (self-reported as 1 of 3 options: male, female, or diverse), weight, height, smoking status, last systolic and diastolic blood pressures, serum fasting glucose level, and most recent available laboratory values within the last 12 months (UACR, eGFR, and glycated hemoglobin [HbA_{1c}]). If patients did not know these laboratory values, they were encouraged to request them from their treating physician.

The observation period for the baseline analysis was from August 18, 2023, until the cutoff date (November 20, 2023). Patients who signed the informed consent form and completed at least the eCRF were eligible for baseline analysis.

Ethical Considerations

Ethics approval for this study was granted by the Ethics Committee of the North Rhine Medical Association (Ärztekammer Nordrhein) on October 13, 2022 (approval number: 2022263). All patients were offered the informed consent form; of the 103 who signed the form, 2 withdrew consent. All patient data were handled in a pseudonymized manner, and the privacy of patients was ensured. Upon completion of the baseline eCRF (regardless of EQ-5D-5L or DTSQ completion), patients received a €10 (US $10.80) voucher for Shop Apotheke as compensation for their time spent on data documentation.

XSL•FO
RenderX

Statistical Analysis

Data were analyzed descriptively with SAS (release 9.4 or higher; SAS Institute Inc). UACRs were categorized based on KDIGO ACR stages [6], with the addition of a fourth level indicating nephrotic-range albuminuria [28,29] (stage A1: UACR of <30 mg/g; stage A2: UACR of 30 - 300 mg/g; stage A3: UACR of 301 - 3000 mg/g; stage A4: UACR of >3000 mg/g). Serum fasting glucose levels were categorized based on the ADA diagnostic criteria for diabetes and prediabetes [30], and eGFRs were reported by KDIGO category [6]. The EQ-5D-5L index score was calculated by using a value set for the German general population [31].

Results

Study Population

During the awareness campaign, push notifications for study- and indication-related content were sent out to a cohort of patients with T2DM (Figure 1), whereof 8799 to 9527 patients received these push notifications. Of these 8799 to 9527 patients, 103 provided consent, and 2 withdrew consent. A total of 101 patients completed the eCRF and were therefore included in the baseline analysis.

Figure 1. Patient recruitment and eligibility in the SMART-Finder observational study. Push notifications were sent out to all eligible patients with T2DM (N=9527). The number of patients receiving each push notification could be influenced by factors such as internet connectivity or patients having their devices switched off at the time of the notification. Therefore, 8921 (93.6%), 8799 (92.4%), 8999 (94.5%), 8716 (91.5%), and 9034 (94.8%) of the 9527 eligible patients received push notifications for the article on the SMART-Finder study, the laboratory values handout, the video on kidney function in CKD, the video on the SMART-Finder study, and the video on risk of CKD, respectively. CKD: chronic kidney disease; DTSQ: Diabetes Treatment Satisfaction Questionnaire; eCRF: electronic case report form; T2DM: type 2 diabetes mellitus.

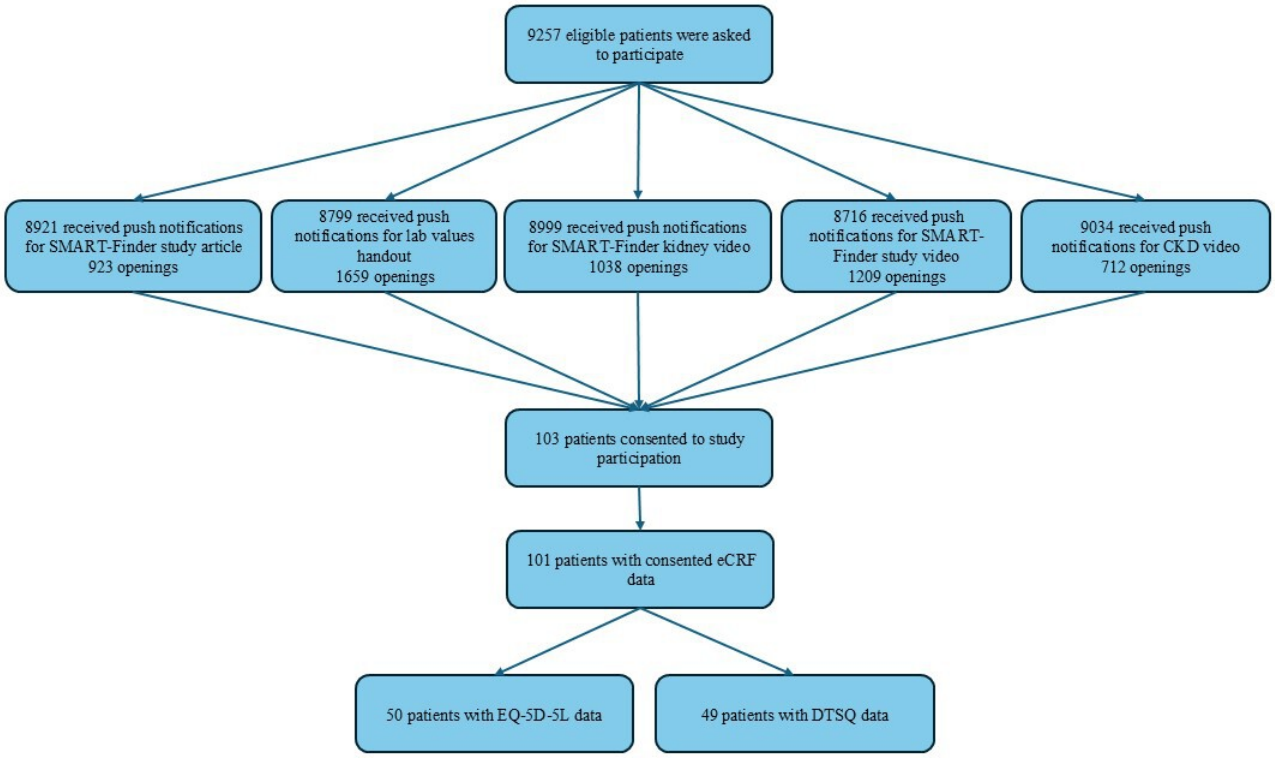

Of the 101 participating patients with T2DM, 61 were male, and 40 were female. The study population had a mean age of 54.2 (SD 11.4; range 22 - 76) years, a mean weight of 103.9 (SD 21.7; range 45 - 149) kg, and a mean BMI of 34 (SD 7.0; range 16.9 - 54.7) kg/m^2. Further, 26 patients were smokers, 43 were former smokers, and 32 had never smoked.

Laboratory Values and Blood Pressure

Of the 101 patients in the analyzed cohort, only 1 female patient and 5 male patients reported their UACR values (Table 1); 3 of the patients (all male) had a UACR of ≥30 mg/g. The remaining 95 (94.1%) patients reported that their HCPs had not provided UACR measurements. For eGFR values, a similar finding was observed—only 9 (8.9%) of the 101 patients were aware of their latest values.

XSL•FO
RenderX

Table . Classification of UACRs[a] in female and male patients.

	Total patients (N=101), n (%[b])	Patients stratified by UACR, n (%[b])				
		Missing data (n=95)	<30 mg/g (n=3)	30 - 300 mg/g (n=2)	301 - 3000 mg/g (n=1)	>3000 mg/g (n=0)
Female	40 (39.6)	39 (41)	1 (33)	0 (0)	0 (0)	0 (0)
Male	61 (60.4)	56 (59)	2 (67)	2 (100)	1 (100)	0 (0)

[a]UACR: urinary albumin-to-creatinine ratio.

[b]Percentages were calculated by using the n values in the column headings as the denominator.

In contrast to laboratory values provided by HCPs, self-determined test results were more commonly reported. In total, 65 (64.4%) patients provided their serum fasting glucose levels, and 90 (89.1%) patients provided their latest HbA_{1c} values (Table 2).

Table . eGFR[a], HbA_{1c}[b,c], and serum fasting glucose[d] at baseline.

	Patients (N=101), n (%)
eGFR at baseline, mL/min/1.73m^2	
Missing	92 (91.1)
15 - 29	1 (1)
30 - 44	2 (2)
45 - 59	2 (2)
60 - 89	4 (4)
HbA_{1c} at baseline, %	
Missing	11 (10.9)
<6.0	9 (8.9)
6.0 - 6.5	29 (28.7)
6.6 - 7.0	16 (15.8)
7.1 - 7.5	16 (15.8)
7.6 - 8.0	6 (5.9)
8.1 - 8.5	3 (3)
8.6 - 9.0	8 (7.9)
>9.0	3 (3)
Serum fasting glucose summary at baseline, mg/dL	
Missing	36 (35.6)
<100	7 (6.9)
100 - 126	38 (37.6)
>126	20 (19.8)

[a]eGFR: estimated glomerular filtration rate.

[b]HbA_{1c}: glycated hemoglobin.

[c]To convert HbA_{1c} to mmol/mol, multiply by 10.93 and subtract 23.5 mmol/mol.

[d]To convert serum fasting glucose to mmol/L, divide by 18.018.

Systolic and diastolic blood pressures are presented in Table 3. In total, 31 (30.7%) patients had a high-normal systolic blood pressure, 15 (14.9%) patients showed systolic hypertension, and 13 (12.9%) patients showed diastolic hypertension according to the 2023 European Society of Hypertension guidelines for the management of arterial hypertension (assuming office-based blood pressure measurements) [32].

XSL•FO
RenderX

Table . Systolic and diastolic blood pressures at baseline.

	Patients (N=101), n (%)
Systolic blood pressure at baseline, mm Hg	
≤119	18 (17.8)
120 - 129	37 (36.6)
130 - 139	31 (30.7)
140 - 159	13 (12.9)
160 - 179	2 (2)
Diastolic blood pressure at baseline, mm Hg	
55 - 59	4 (4)
60 - 64	4 (4)
65 - 69	7 (6.9)
70 - 74	14 (13.9)
75 - 79	11 (10.9)
80 - 84	36 (35.6)
85 - 89	12 (11.9)
90 - 99	9 (8.9)
100 - 109	4 (4)

Medication Usage

Use of SGLT2is, the nonsteroidal mineralocorticoid receptor antagonist finerenone, hypertension medication (defined by anatomic therapeutic chemical codes), and nephrotoxic medication (as listed by Patel and Sapra [33]) was reported by 44, 0, 80, and 83 patients, respectively. All patients confirmed their T2DM diagnosis at baseline. Further, 81 reported hypertension (77 with treated hypertension and 4 with untreated hypertension) and 12 reported CKD as comorbidities. Reported comedication data for all patients are presented in Multimedia Appendix 1.

Quality of Life and Treatment Satisfaction

A total of 50 patients completed the EQ-5D-5L. The mean EQ-visual analog scale score was 64.4 (SD 24.4; range 3.0 - 99.0; on a scale of 0 - 100, where 0 is the worst and 100 is the best health state possible), and the mean EQ-5D-5L index score was 0.7 (SD 0.3; range –0.1 to 1.0). The DTSQ was completed by 49 patients, with a mean DTSQ score of 28.8 (SD 6.8; range 9.0 - 36.0) at baseline.

Discussion

Principal Results and Comparison With Prior Work

With the SMART-Finder study, we are investigating the self-reporting of UACR, eGFR, and HbA_{1c} values in a cohort of patients with T2DM using an adherence-supporting app in Germany. This baseline analysis provides insights into the implementation of treatment guidelines and HCP-patient interactions.

We observed a low participation rate in our cohort of patients, who were taking responsibility for their health status by regularly tracking adherence. Of the 9527 patients who were contacted, only 712 (7.5%) to 1659 (17.4%) patients engaged with at least one aspect of the awareness campaign by clicking on a video or the written information content in the app. The final participation rate—8.4% (101 of 1209 patients who saw the SMART-Finder study video)—is comparable to those observed in other awareness campaigns; for example, 10.2% (804/7865) and 10.1% (802/7920) of invited participants completed a survey on a public health campaign for increasing kidney health awareness in a Canadian province [34].

In contrast to activities that can be self-managed by patients (blood pressure measurement and testing of HbA_{1c} and serum fasting glucose levels), about which our cohort appears to be well informed, our results suggest a dramatic lack of awareness of renal function status in this population at risk of CKD. This is of particular concern, since almost half of the cohort (46/101, 45.5%) had a systolic blood pressure of ≥130 mm Hg, and 82.2% (83/101) reported former or current intake of nephrotoxic medication. Further, more than 90% of the patients were unaware of their eGFR (92/101, 91.1%) and UACR (95/101, 94.1%), but 89.1% (90/101) of the patients knew their HbA_{1c} level, suggesting that treating physicians either do not test for UACRs and eGFRs or do not inform patients about the results. Both a lack of adherence to guidelines for diagnostic testing and a lack of HCP-patient communication put patients at risk. Another reason for this imbalance in health literacy may be the focus on HbA_{1c}, rather than on kidney comorbidity, in patient education material.

Limitations

The presented baseline analysis has some limitations. There is a selection bias, as only users of the adherence-supporting app, that is, those who seemed to have experience with using

XSL•FO
RenderX

technology, were able to participate. Additionally, the unconventional method of data collection, with patients requesting data from HCPs, may have introduced reporting bias from the HCPs. Further, due to the low participation rate and the lack of UACR and eGFR values, our findings on CKD risk in the observed cohort of patients may not be generalizable to all patients with T2DM. However, the missing data show a lack of awareness in a high-risk population. Even though the annual eGFR and albuminuria testing rates in Germany are 96.5% and 77.2%, respectively, our results suggest that among German patients with diabetes, knowledge regarding their kidney function and disease is still insufficient [35].

Conclusions

Baseline data from the SMART-Finder study, which were derived from patients with T2DM who were using an adherence-supporting app in Germany, showed a dramatic lack of awareness of kidney-related biomarkers (UACR and eGFR) but revealed high knowledge of HbA_{1c} values. As only informed patients are empowered patients, future goals for diabetes management must include guideline-compliant testing of CKD-related biomarkers and open communication between HCPs and patients.

Acknowledgments

This study was sponsored by Bayer Vital GmbH, Leverkusen, Germany. It was performed in collaboration with smartpatient GmbH, Munich, Germany (the provider of the MyTherapy app)—a subsidiary of Shop Apotheke, one of the largest internet-based pharmacies in Germany—and EMMES Biopharma Germany, who assisted with the statistical analysis. The funding organization was the sponsor of this study. Bayer provided support for designing and conducting the study; managing data collection and interpretation; and managing manuscript preparation, review, and approval. The authors would like to thank Claire Mulligan, PhD (Beacon Medical Communications Ltd, Brighton, United Kingdom) for editing the final draft of this manuscript (funded by Bayer Vital GmbH, Leverkusen, Germany).

Data Availability

Data sharing can be requested from the sponsor of this study, and it will be managed according to internal standard operating procedures and the General Data Protection Regulation of the European Union.

Authors' Contributions

CM and SK had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. CM and TN contributed to the concept and design of the study. All authors contributed to the acquisition, analysis, or interpretation of data. CM drafted the manuscript. ITW, SM, TN, JN, RG, and SK critically reviewed the manuscript for important intellectual content. SK and RG contributed to statistical analysis. CM and TN obtained funding. TP, JN, and TN provided administrative, technical, or material support. SM supervised the study.

Conflicts of Interest

CM, TN, and ITW are employees of Bayer Vital GmbH. JN and TP are employees of smartpatient GmbH. RG and SK are employees of EMMES Biopharma Germany. SM received consultancy fees from Bayer Vital GmbH.

Multimedia Appendix 1
Reported concomitant medications.
[DOCX File, 46 KB - formative_v9i1e60246_app1.docx]

References

1. Song MK, Davies NM, Roufogalis BD, Huang THW. Management of cardiorenal metabolic syndrome in diabetes mellitus: a phytotherapeutic perspective. J Diabetes Res 2014;2014:313718. [doi: 10.1155/2014/313718] [Medline: 24818164]
2. Galicia-Garcia U, Benito-Vicente A, Jebari S, et al. Pathophysiology of type 2 diabetes mellitus. Int J Mol Sci 2020 Aug 30;21(17):6275. [doi: 10.3390/ijms21176275] [Medline: 32872570]
3. Roden M, Shulman GI. The integrative biology of type 2 diabetes. Nature 2019 Dec;576(7785):51-60. [doi: 10.1038/s41586-019-1797-8] [Medline: 31802013]
4. Stumvoll M, Goldstein BJ, van Haeften TW. Type 2 diabetes: principles of pathogenesis and therapy. Lancet 2005 Apr;365(9467):1333-1346. [doi: 10.1016/S0140-6736(05)61032-X] [Medline: 15823385]
5. Ronco C, Haapio M, House AA, Anavekar N, Bellomo R. Cardiorenal syndrome. J Am Coll Cardiol 2008 Nov 4;52(19):1527-1539. [doi: 10.1016/j.jacc.2008.07.051] [Medline: 19007588]
6. Kidney Disease: Improving Global Outcomes (KDIGO) Diabetes Work Group. KDIGO 2020 clinical practice guideline for diabetes management in chronic kidney disease. Kidney Int 2020 Oct;98(4S):S1-S115. [doi: 10.1016/j.kint.2020.06.019] [Medline: 32998798]

XSL•FO
RenderX

7. Perkovic V, Agarwal R, Fioretto P, et al. Management of patients with diabetes and CKD: conclusions from a "Kidney Disease: Improving Global Outcomes" (KDIGO) controversies conference. Kidney Int 2016 Dec;90(6):1175-1183. [doi: 10.1016/j.kint.2016.09.010] [Medline: 27884312]
8. GBD 2021 Diabetes Collaborators. Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: a systematic analysis for the Global Burden of Disease Study 2021. Lancet 2023 Jul 15;402(10397):203-234. [doi: 10.1016/S0140-6736(23)01301-6] [Medline: 37356446]
9. Wilmer WA, Rovin BH, Hebert CJ, Rao SV, Kumor K, Hebert LA. Management of glomerular proteinuria: a commentary. J Am Soc Nephrol 2003 Dec;14(12):3217-3232. [doi: 10.1097/01.asn.0000100145.27188.33] [Medline: 14638920]
10. Chen TK, Knicely DH, Grams ME. Chronic kidney disease diagnosis and management: a review. JAMA 2019 Oct 1;322(13):1294-1304. [doi: 10.1001/jama.2019.14745] [Medline: 31573641]
11. Chronic Kidney Disease Prognosis Consortium, Matsushita K, van der Velde M, et al. Association of estimated glomerular filtration rate and albuminuria with all-cause and cardiovascular mortality in general population cohorts: a collaborative meta-analysis. Lancet 2010 Jun 12;375(9731):2073-2081. [doi: 10.1016/S0140-6736(10)60674-5] [Medline: 20483451]
12. Hashmi MF, Benjamin O, Lappin SL. End-Stage Renal Disease: StatPearls Publishing; 2023. [Medline: 29763036]
13. Mahemuti N, Zou J, Liu C, Xiao Z, Liang F, Yang X. Urinary albumin-to-creatinine ratio in normal range, cardiovascular health, and all-cause mortality. JAMA Netw Open 2023 Dec 1;6(12):e2348333. [doi: 10.1001/jamanetworkopen.2023.48333] [Medline: 38113044]
14. Garcia Sanchez JJ, Thompson J, Scott DA, et al. Treatments for chronic kidney disease: a systematic literature review of randomized controlled trials. Adv Ther 2022 Jan;39(1):193-220. [doi: 10.1007/s12325-021-02006-z] [Medline: 34881414]
15. Bao W, Zhang M, Li N, Yao Z, Sun L. Efficacy and safety of finerenone in chronic kidney disease associated with type 2 diabetes: a systematic review and meta-analysis of randomized clinical trials. Eur J Clin Pharmacol 2022 Dec;78(12):1877-1887. [doi: 10.1007/s00228-022-03408-w] [Medline: 36273065]
16. Shiau CH, Tsau LY, Kao CC, et al. Efficacy and safety of sodium-glucose cotransporter-2 inhibitors in patients with chronic kidney disease: a systematic review and meta-analysis. Int Urol Nephrol 2024 Apr;56(4):1359-1381. [doi: 10.1007/s11255-023-03789-6] [Medline: 37752340]
17. de Boer IH, Khunti K, Sadusky T, et al. Diabetes management in chronic kidney disease: a consensus report by the American Diabetes Association (ADA) and Kidney Disease: Improving Global Outcomes (KDIGO). Kidney Int 2022 Nov;102(5):974-989. [doi: 10.1016/j.kint.2022.08.012] [Medline: 36202661]
18. Kovesdy CP. Epidemiology of chronic kidney disease: an update 2022. Kidney Int Suppl (2011) 2022 Apr;12(1):7-11. [doi: 10.1016/j.kisu.2021.11.003] [Medline: 35529086]
19. Gheith O, Farouk N, Nampoory N, Halim MA, Al-Otaibi T. Diabetic kidney disease: world wide difference of prevalence and risk factors. J Nephropharmacol 2015 Oct 9;5(1):49-56. [Medline: 28197499]
20. American Diabetes Association. 11. Microvascular complications and foot care: standards of medical care in diabetes-2019. Diabetes Care 2019 Jan;42(Suppl 1):S124-S138. [doi: 10.2337/dc19-S011] [Medline: 30559237]
21. Cosentino F, Grant PJ, Aboyans V, et al. 2019 ESC guidelines on diabetes, pre-diabetes, and cardiovascular diseases developed in collaboration with the EASD. Eur Heart J 2020 Jan 7;41(2):255-323. [doi: 10.1093/eurheartj/ehz486] [Medline: 31497854]
22. Rasaratnam N, Salim A, Blackberry I, et al. Urine albumin-creatinine ratio variability in people with type 2 diabetes: clinical and research implications. Am J Kidney Dis 2024 Jul;84(1):8-17.e1. [doi: 10.1053/j.ajkd.2023.12.018] [Medline: 38551531]
23. Hagnäs M, Sundqvist H, Jokelainen J, et al. The prevalence of chronic kidney disease and screening of renal function in type 2 diabetic patients in Finnish primary healthcare. Prim Care Diabetes 2020 Dec;14(6):639-644. [doi: 10.1016/j.pcd.2020.05.005] [Medline: 32527663]
24. Manski-Nankervis JAE, Thuraisingam S, Lau P, et al. Screening and diagnosis of chronic kidney disease in people with type 2 diabetes attending Australian general practice. Aust J Prim Health 2018 Jul;24(3):280-286. [doi: 10.1071/PY17156] [Medline: 29807557]
25. Chu L, Bhogal SK, Lin P, et al. AWAREness of diagnosis and treatment of chronic kidney disease in adults with type 2 diabetes (AWARE-CKD in T2D). Can J Diabetes 2022 Jul;46(5):464-472. [doi: 10.1016/j.jcjd.2022.01.008] [Medline: 35739044]
26. Chu CD, McCulloch CE, Banerjee T, et al. CKD awareness among US adults by future risk of kidney failure. Am J Kidney Dis 2020 Aug;76(2):174-183. [doi: 10.1053/j.ajkd.2020.01.007] [Medline: 32305206]
27. Mueller C, Schürks M, Neußer T, et al. Identification of patients with elevated urine albumin-to-creatinine ratio levels in a type 2 diabetes mellitus cohort based on data submitted by patients via a smartphone app (SMART-Finder): protocol for an observational study. JMIR Res Protoc 2023 Apr 5;12:e44996. [doi: 10.2196/44996] [Medline: 37018036]
28. Kodner C. Diagnosis and management of nephrotic syndrome in adults. Am Fam Physician 2016 Mar 15;93(6):479-485. [Medline: 26977832]
29. Kuritzky L, Toto R, Van Buren P. Identification and management of albuminuria in the primary care setting. J Clin Hypertens (Greenwich) 2011 Jun;13(6):438-449. [doi: 10.1111/j.1751-7176.2010.00424.x] [Medline: 21649844]
30. American Diabetes Association. 2. Classification and diagnosis of diabetes: standards of medical care in diabetes-2021. Diabetes Care 2021 Jan;44(Suppl 1):S15-S33. [doi: 10.2337/dc21-S002] [Medline: 33298413]

XSL•FO
RenderX

31. Ludwig K, Graf von der Schulenburg JM, Greiner W. German value set for the EQ-5D-5L. Pharmacoeconomics 2018 Jun;36(6):663-674. [doi: 10.1007/s40273-018-0615-8] [Medline: 29460066]
32. Mancia G, Kreutz R, Brunström M, et al. 2023 ESH guidelines for the management of arterial hypertension The Task Force for the management of arterial hypertension of the European Society of Hypertension: endorsed by the International Society of Hypertension (ISH) and the European Renal Association (ERA). J Hypertens 2023 Dec 1;41(12):1874-2071. [doi: 10.1097/HJH.0000000000003480] [Medline: 37345492]
33. Patel JB, Sapra A. Nephrotoxic Medications: StatPearls Publishing; 2023. [Medline: 31985937]
34. Ryz K, Tangri N, Verrelli M, et al. A before and after cross-sectional analysis of a public health campaign to increase kidney health awareness in a Canadian province. BMC Res Notes 2015 Nov 20;8:695. [doi: 10.1186/s13104-015-1662-2] [Medline: 26590133]
35. Kellerer M, Wanner C. Kidney diseases in type 2 diabetes mellitus : overview and implementation of guidelines, position papers and practical recommendations for diagnostics and monitoring. Inn Med (Heidelb) 2023 Dec;64(12):1211-1217. [doi: 10.1007/s00108-023-01610-1] [Medline: 37955643]

Abbreviations

ACR: albumin-to-creatinine ratio
ADA: American Diabetes Association
CKD: chronic kidney disease
DTSQ: Diabetes Treatment Satisfaction Questionnaire
eCRF: electronic case report form
eGFR: estimated glomerular filtration rate
HbA_{1c}: glycated hemoglobin
HCP: health care professional
KDIGO: Kidney Disease: Improving Global Outcomes
SGLT2i: sodium-glucose cotransporter 2 inhibitor
T2DM: type 2 diabetes mellitus
UACR: urine albumin-to-creatinine ratio

Edited by A Mavragani; submitted 07.05.24; peer-reviewed by JC Galle, JH Shubrook; revised version received 08.01.25; accepted 09.01.25; published 18.02.25.

Please cite as:
Mueller C, Neusser T, Thate-Waschke I, Nowicki J, Plominski T, Griesinger R, Kessner S, Martin S
Disease Awareness in Patients With Type 2 Diabetes: Analysis of Baseline Data From the SMART-Finder Observational Study
JMIR Form Res 2025;9:e60246
URL: https://formative.jmir.org/2025/1/e60246
doi:10.2196/60246

© Christian Mueller, Thomas Neusser, Inga Thate-Waschke, Julia Nowicki, Tomasz Plominski, Regine Griesinger, Stefanie Kessner, Stephan Martin. Originally published in JMIR Formative Research (https://formative.jmir.org), 18.2.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

A Very Low–Carbohydrate Program in Adults With Metabolic Dysfunction–Associated Steatotic Liver Disease and Phospholipase Domain–Containing Protein 3 Risk Genotype: Pre-Post Intervention Study

Laura R Saslow[1], PhD; Jamie Krinock[1], BS; Alison O'Brien[1], MPH; Kaitlyn Raymond[1], BS; Hovig Bayandorian; Judith T Moskowitz[2], MPH, PhD; Jennifer Daubenmier[3], PhD; Antonino Oliveri[4], PhD; Deanna J Marriott[5], PhD; Dina H Griauzde[6,7], MD, MSc; Elizabeth K Speliotes[4,8], MPH, MD, PhD

[1]Department of Health Behavior and Biological Sciences, School of Nursing, University of Michigan, Ann Arbor, MI, United States

[2]Department of Medical Social Sciences, Feinberg School of Medicine, Northwestern University, Chicago, IL, United States

[3]Institute of Holistic Health Studies, San Francisco State University, San Francisco, CA, United States

[4]Division of Gastroenterology, Department of Internal Medicine, University of Michigan Medical School, Medical Sciences Building II, Room 4741, Ann Arbor, MI, United States

[5]Department of Systems, Populations and Leadership, School of Nursing, University of Michigan, Ann Arbor, MI, United States

[6]VA Ann Arbor Healthcare System, Ann Arbor, MI, United States

[7]Department of Internal Medicine, University of Michigan Medical School, Ann Arbor, MI, United States

[8]Gilbert S Omenn Department of Computational Medicine and Bioinformatics, University of Michigan Medical School, Ann Arbor, MI, United States

Corresponding Author:
Elizabeth K Speliotes, MPH, MD, PhD
Division of Gastroenterology, Department of Internal Medicine, University of Michigan Medical School, Medical Sciences Building II, Room 4741, Ann Arbor, MI, United States

Abstract

Background: Insulin resistance and the G allele of rs738409 interact to create a greater risk of metabolic dysfunction–associated steatotic liver disease.

Objective: This study aims to confirm that one promising way to reduce insulin resistance is by following a very low–carbohydrate (VLC) dietary pattern.

Methods: Adults with rs738409-GG or -CG with liver steatosis and elevated liver function tests, were taught an ad libitum VLC diet, positive affect and mindful eating skills, goal setting, and self-monitoring and given feedback and coaching for 4 months. We measured liver steatosis, anthropometric, serum metabolic diet adherence, and quality of life measures.

Results: In this small pilot trial, of the 11 participants enrolled, 9 (82%) participants completed outcomes. All 11 participants viewed at least 1 session of the intervention, and 8 (73%) participants viewed at least half of the sessions. Among the 9 participants who provided 4-month self-report information, intervention satisfaction was high (mean 6.22, 95% CI 5.58-6.85), with 5 (56%) participants rating the intervention the top score, and 4 (44%) participants reporting they did not plan to stop following the VLC diet. Across participants with a 4-month hepatic liver fat percent measurement, the percent change in liver fat was −33.17% (95% CI −86.48 to 20.14), and in only the participants who were adherent to the eating pattern, the percent change in liver fat was −53.12% (95% CI −71.25 to −34.99). Amongst participants with a 4-month hepatic liver fat percent measurement, 6 out of 8 (75%) participants were considered responders, with a relative decline in liver fat ≥30%, and of the 9 participants with a 4-month body weight, 9 (100%) participants lost ≥5% of their body weight. There were no serious adverse events.

Conclusions: Results suggest the feasibility, acceptability, and preliminary efficacy of the VLC intervention in adults with higher genetic risk for metabolic dysfunction–associated steatotic liver disease, although there is a need for further studies given the small sample size and the high risk of substantial biases in this small pilot study.

(JMIR Form Res 2025;9:e60051) doi:10.2196/60051

KEYWORDS

metabolic dysfunction–associated steatotic liver disease; ketogenic diet; low carbohydrate; adult; genotype; insulin; insulin resistance; metabolic dysfunction; dietary pattern; type 2 diabetes; T2DM; single-arm pilot trial; liver function test; genome; non-alcoholic fatty liver disease

XSL•FO
RenderX

Introduction

Metabolic dysfunction–associated steatotic liver disease (MASLD) is an increasingly prevalent disease afflicting about 25% of US adults and more than 50% of people with type 2 diabetes [1]. MASLD is caused by the deposition of excess fat in the liver, not due to alcohol, and can lead to advanced liver disease in the form of inflammation (metabolic dysfunction–associated steatohepatitis), fibrosis/cirrhosis (scarring), and hepatocellular carcinoma (liver cancer) [2]. Advanced MASLD is present in more than 90% of severely obese individuals [3]. It is associated with a shorter lifespan [4] and is expected to become the leading indication for liver transplantation in the United States [5].

Several dietary approaches have been suggested to treat MASLD, but no single approach has been found to be superior. For example, time-restricted eating and low-calorie, low-fructose, low-carbohydrate, very low–carbohydrate (VLC), and Mediterranean diets have all been studied in the context of MASLD [6-8].

We and others have shown that MASLD is heritable and the strongest common risk factor is having the I148M variant for the patatin-like phospholipase domain–containing protein 3 (PNPLA3) gene [9]. We have previously shown that individuals who carry this allele have a multiplicative risk of developing MASLD if they also have insulin resistance. The largest gene-environment meta-analysis study to date included approximately 15,000 individuals of diverse ancestry and demonstrated that rs738409-GG individuals had a 57% increased risk of MASLD if their insulin levels were in the highest versus lowest quartile of insulin, whereas rs738409-CC individuals had only a 32% increased risk, suggesting the presence of gene-environment interaction. Despite being at high risk of liver disease, individuals with rs738409-GG or -CG and insulin resistance are not being specifically targeted by programs aiming to reduce insulin resistance.

One promising way to reduce insulin resistance is through the use of a VLC eating pattern, also known as a ketogenic or “keto” eating pattern, which is a very reduced carbohydrate, moderate protein, and higher fat eating pattern. A VLC eating pattern has been found to be effective for reducing insulin resistance, body weight, inflammation, intrahepatic lipid content, and de novo fatty acid synthesis and fatty liver deposits, with some preliminary research showing the benefit of a VLC for MASLD [10-14]. On the other hand, high-carbohydrate, low-fat diets are related to greater liver inflammation or damage [15,16].

The National Institutes of Health’s obesity-related behavioral intervention trials model for behavioral intervention development [17] encourages preliminary testing of an intervention in a subpopulation before conducting a full-scale clinical trial. In this research, we built on previous research findings demonstrating that a VLC eating pattern can be taught using a web-based program that includes positive affect skills, which aim to increase the frequency with participants experience positive emotions, to improve intervention adherence and satisfaction [18,19]; mindful eating, to help participants cope with emotional and hedonic-driven eating, drivers of weight gain [20-22]; and other strategies including goal setting and self-monitoring; personalized feedback and support from a coach; social support; reminders; and booster messages [23].

We hypothesized that a VLC eating pattern might be able to achieve MASLD improvement or even reversal in adults with steatosis or mild fibrosis, especially in the higher-risk subpopulations of rs738409-GG or -CG individuals. As a first step, in this small pilot trial of adults with MASLD, we examined intervention feasibility, intervention acceptability, and physical and patient-reported outcomes of rs738409-GG or -CG individuals.

Methods

Study Design

This was a single-arm pilot trial of an online, VLC dietary intervention.

Ethical Considerations

The research was approved by the University of Michigan Institutional Review Board (HUM00154361) and registered at ClinicalTrials.gov (NCT05010070). The study followed the ethical standards of the responsible committee on human experimentation (institutional and national) and abided by the Helsinki Declaration of 1975, as revised in 2000. All participants provided informed consent and could withdraw from the study at any point. No writing assistance was used for this manuscript. Recruitment began on August 31, 2021, and the outcomes of the trial were finalized by October 3, 2022.

Participant Recruitment and Screening

Potential participants were contacted based on their rs738409 genotype in the Michigan Genomics Initiative (MGI), a research initiative that collects and genotypes blood samples and allows researchers to link this genetic information to patients. As part of the initiative, participants had already provided informed consent for broad research purposes [24]. Participants interested in our pilot study were referred to a webpage that described the trial and linked to an online screening survey (Qualtrics). The screening survey included questions used to assess eligibility. If individuals passed this, they were sent a video describing the goals, procedure, and pros and cons of participating in the trial. If they continued to express interest in the trial, they consented, baseline measures were collected, and then they were enrolled in the trial. Inclusion criteria included a baseline magnetic resonance imaging with liver steatosis but not cirrhosis, having been identified based on information from the MGI database, and having elevated liver function tests. Exclusion criteria included non-MASLD causes of elevated liver function tests; use of exogenous insulin; planned or history of weight loss surgery; active substance use or untreated mental health condition that could pose a safety risk; advanced medical conditions including as current chemotherapy, heart failure, or kidney failure; type 1 diabetes; Cushing syndrome; adherence to a vegan or dietary vegetarian; pregnant or planning to get pregnant in the next 6 months; currently enrolled in a weight loss program or other investigative study that might conflict with this research; taking medications known to cause weight gain or loss; or recent decompensation/hospitalization.

XSL•FO
RenderX

Intervention

The online intervention included 16 weekly emails with a video and handouts, as well as text messages and email-based coaching. At baseline, participants received a body weight scale (Bodytrace) and urine ketone strips. Participants were also sent VLC cookbooks by mail 3 times throughout the intervention [25-27]. We taught participants to follow a VLC eating pattern, consisting of 20 - 35 net (nonfiber) grams of carbohydrates per day. At week 6, we encouraged participants to become physically active for at least 150 minutes of moderate-intensity physical activity per week as well as to sleep enough to feel well rested. The intervention also taught skills related to feeling more positive affect and eating mindfully [28]. Throughout, we asked participants to track their dietary intake and to step on the body weight scale at least once a week. The program's email-based coach provided feedback and support and we sent near daily, unique text messages that provided reminders of the intervention content.

Data Collection

At baseline, participants were asked to self-weigh using the scale, complete a fasting blood draw and magnetic resonance imaging at Michigan Medicine, complete an online survey (Qualtrics); and complete a 2-day dietary recall with a trained dietitian by phone. At 4 months, at the end of the intervention, these were repeated. Prior to each week's didactic session, we asked participants questions about their dietary adherence and experiences. Procedures associated with the research were either covered by the study or the participants' insurance.

Outcome Measures

Intervention Feasibility

Trial and Intervention Retention

We defined trial retention as the percentage of participants who completed the postintervention outcomes divided by the total number of trial participants enrolled. We assessed intervention retention based on intervention participation: (1) viewing at least 1 session, and (2) being active in the program, defined as viewing at least 8 out of 16 (50%) of the sessions.

Dietary Adherence

At baseline and 4 months, participants took part in two 24-hour dietary recalls. We considered participants adherent to the eating pattern if their net carbohydrates at 4 months were below 60 grams of carbohydrates per day and their overall calories did not increase by more than 400 kcal per day (as this was ad libitum, in which participants are asked to eat when hungry and stop when full, which typically leads to lower hunger).

Intervention Acceptability

General Intervention Satisfaction

We asked participants, "How would you rate your overall satisfaction with the program?" Response options ranged from 1="not at all satisfied" to 7="very satisfied." To assess potential acceptability, we also asked participants to answer, "How long can you see yourself following your assigned diet?"

Intervention Skills Satisfaction

Participants rated their satisfaction with the positive affect skills and, separately, the mindfulness skills. Response options ranged from 1="don't include them, they were not helpful" to 7="you must include them, they were very helpful."

Qualitative Feedback

At 4 months, using open-ended questions, we asked participants about their experiences with the intervention and recommendations for improvement.

Patient-Centered Outcomes

Change in Liver Percent Fat and Fibrosis

We used magnetic resonance imaging to assess the percent fat of the left and right liver lobes, which we then combined. We also used magnetic resonance imaging to quantify the fibrosis in the liver. We examined what percent of participants lost at least 30% of their liver fat, a clinically significant threshold [29]. We a priori registered a reduction in liver fat as our primary outcome of the trial. We therefore examined this outcome for all participants and for adherent participants.

Body Weight Changes

Participants self-weighed throughout the trial and this weight was sent automatically to the trial's staff through the scale's connection to a cellular network. Mean change in weight and BMI (using baseline self-reported height) was calculated at 4 months compared with baseline. Mean percent weight loss was defined as (weight at 4 months – baseline weight) / (baseline weight) *100. We examined what percent of participants lost at least 5% of their body weight, a clinically significant threshold [30].

Blood Test Changes

We assessed liver function with aspartate aminotransferase and alanine aminotransferase, glucose control with glycated hemoglobin (HbA_{1c}) and fasting glucose, fasting insulin, Homeostatic Model Assessment for Insulin Resistance (HOMA-IR), and lipids (triglycerides, high-density lipoprotein or HDL, and low-density lipoprotein or LDL), and inflammation (with C-reactive protein). We calculated the mean change at 4 months compared with baseline.

Self-Rated Change in Health

We asked participants to rate how their health changed over the intervention by answering the questions, "How much do you think your health has changed as a result of participating in this program?" (Response options ranged from 1="very much worse" to 7="very much better").

Changes in Chronic Liver Disease Questionnaire (CLDQ)

We measured change in MASLD-related quality of life with the chronic liver disease questionnaire (CLDQ), which assesses 6 dimensions, rated from 1 to 7, with a higher number reflecting worse symptoms: abdominal symptoms (bloating, discomfort, and pain), activity (trouble with eating, physical ability), emotional (poor mood, sleep, and ability to concentrate), fatigue,

XSL•FO
RenderX

systemic symptoms (body pain, dry mouth, itching, and muscle cramps), and worry about MASLD [31].

Changes in Psychological Factors

We assessed psychological changes from baseline to 4 months for positive affect with the Scale of Positive and Negative Experience [32], mindful eating with the reliance on hunger and satiety cues subscale of the Intuitive Eating Scale-2 [33], and stress-based eating with the Palatable Eating Motives Scale coping subscale [34].

Statistical Analysis

For statistical significance, we followed the guidance for the use of frequentist inferential statistics in public health [35]. We completed an intention-to-treat analysis with all available data and analyses that were limited to only adherent individuals, with no imputation for missing data. Means and 95% CIs were calculated for continuous variables.

Results

Sample Characteristics

Of the 97 individuals who expressed interest in study participation and were screened for study eligibility, 55 did not meet inclusion criteria (including, as the top reasons for ineligibility: 13 were taking insulin, 11 had prior weight loss surgery or were taking weight loss medications, 8 had normal liver function tests, 6 had cirrhosis of the liver, 3 had heart failure or another related heart condition), 12 were eligible based on the prescreening survey but did not reply to follow-up requests, and 5 were eligible but declined to participate. Eleven participants enrolled in the trial.

At baseline, participants were 38 - 77 years old (mean 55.54 years, SD 11.85 years), 8 out of 11 (72.7%) participants were male, most were White and not Hispanic (10/11, 90.9%), with 1 being White and Hispanic. Of the 11 participants, 7 (63.6%) participants were rs738409-GG and 4 (36.4%) participants were rs738409-CG. Other baseline values are in Table 1.

XSL•FO
RenderX

Table . Baseline for all participants (n=11).

Outcomes	Values
Age (years), mean (SD)	55.54 (11.85)
Sex (male), n (%)	8 (73)
Race (White and not Hispanic), n (%)	10 (91)
rs738409-GG (vs rs738409-CG), n (%)	7 (63.6)
Weight (kg), mean (95% CI)	96.19 (83.24-109.14)
BMI (kg/m^2), mean (95% CI)	33.79 (29.60-37.98)
Liver lobe fat (%), mean (95% CI)	23.10 (14.25-37.98)
Liver lobe fibrosis, mean (95% CI)	2.46 (2.03-2.89)
AST[a] (U/L), mean (95% CI)	49.91 (24.85-74.97)
ALT[b] (U/L), mean (95% CI)	78.36 (36.90-119.83)
HbA_{1c}[c] (mmol/mol), mean (95% CI)	40.22 (35.86-44.58)
Glucose (mg/dL), mean (95% CI)	101.55 (91.44-111.65)
Insulin (mIU/mL), mean (95% CI)	21.15 (14.01-28.28)
HOMA-IR[d], mean (95% CI)	5.44 (3.51-7.37)
Triglycerides (mg/dL) mean (95% CI)	127.82 (87.70-167.94)
HDL[e] (mg/dL), mean (95% CI)	48.00 (35.56-60.44)
LDL[f] (mg/dL), mean (95% CI)	84.82 (63.13-106.51)
C-reactive protein (mg/dL), mean (95% CI)	0.79 (0.24-1.34)
CLDQ[g], mean (95% CI)	
Overall	2.62 (1.93-3.31)
Abdominal symptoms	2.49 (1.65-3.32)
Activity	2.33 (1.55-3.12)
Emotional function	2.60 (1.80-3.39)
Fatigue	2.98 (2.12-3.84)
Systemic symptoms	2.43 (1.67-3.19)
Worry about MASLD[h]	2.75 (1.78-3.73)
Positive affect, mean (95% CI)	24.18 (20.99-27.37)
Mindful eating, mean (95% CI)	2.88 (2.48-3.28)
Stress-based eating, mean (95% CI)	1.70 (1.24-2.17)
Energy, kilocalories, mean (95% CI)	1901.92 (1547.20-2256.64)
Carbohydrates (g), mean (95% CI)	192.40 (139.66-245.15)
Fiber (g), mean (95% CI)	17.23 (12.18-22.28)
Net (Non-fiber) Carbohydrates (g), mean (95% CI)	175.17 (123.97-226.37)
Total fat (g), mean (95% CI)	85.99 (67.55-104.43)
Monounsaturated fat (g), mean (95% CI)	30.38 (23.13-37.63)
Polyunsaturated fat (g), mean (95% CI)	17.26 (11.37-23.15)
Saturated Fat (g), mean (95% CI)	30.88 (23.74-38.02)
Protein (g), mean (95% CI)	80.31 (67.57-94.06)

[a]AST: aspartate aminotransferase.
[b]ALT: alanine aminotransferase.

XSL•FO
RenderX

[c]HbA_{1c}: glycated hemoglobin.

[d]HOMA-IR: Homeostatic Model Assessment for Insulin Resistance.

[e]HDL: high-density lipoprotein.

[f]LDL: low-density lipoprotein.

[g]CLDQ: chronic liver disease questionnaire.

[h]MASLD: metabolic dysfunction–associated steatotic liver disease.

Intervention Feasibility

Trial and Intervention Retention

Of the 11 participants who were enrolled, 9 (82%) participants completed study outcomes. All 11 participants viewed at least 1 session of the intervention, and 8 (73%) participants viewed at least half of the sessions.

Dietary Adherence

Nine participants completed the pre- and post-intervention dietary recalls. Changes in nutrient intake are shown in Table 2. In Table 3, we show results for just the 6 participants who were adherent to the eating pattern.

XSL•FO
RenderX

Table . Change in outcomes from baseline to post for all participants with outcome data. For all outcomes n=9, except for liver lobe fat % (n=8), liver lobe fibrosis (n=6), and C-reactive protein (n=4).

Outcomes	Baseline (mean, 95% CI)	Post (mean, 95% CI)	Percent change (mean, 95% CI)	Change (mean, 95% CI)
Weight (kg)	96.19 (83.24 to 109.14)	85.50 (74.63 to 96.37)	−10.88 (−13.06 to −8.69)	−10.69 (−13.37 to −8.00)
BMI (kg/m^2)	32.93 (28.03 to 37.83)	29.26 (25.29 to 33.22)	−10.85 (−13.01 to −8.69)	−3.68 (−4.74 to −2.62)
Liver lobe fat (%)	19.13 (12.22 to 26.04)	12.06 (5.64 to 18.48)	−33.17 (−86.48 to 20.14)	−7.06 (−15.50 to 1.37)
Liver lobe fibrosis	2.41 (2.04 to 2.78)	2.09 (1.59 to 2.59)	−13.22 (−31.39 to 4.95)	−0.33 (−0.82 to 0.17)
AST[a] (U/L)	50.22 (18.48 to 81.97)	41.33 (2.88 to 79.78)	−24.30 (−40.89 to −7.70)	−8.89 (−19.52 to 1.74)
ALT[b] (U/L)	79 (26.35 to 131.65)	54.44 (−4.97 to 113.86)	−35.16 (−56.14 to −14.17)	−24.56 (−49.79 to 0.68)
HbA_{1c}[c] (mmol/mol)	40.22 (35.86 to 44.58)	36.33 (32.83 to 39.83)	−9.40 (−13.50 to −5.30)	−3.89 (−5.75 to -2.03)
Glucose (mg/dL)	105.56 (96.32 to 114.79)	95.56 (84.86 to 106.25)	−9.55 (−15.79 to −3.30)	−10.00 (−16.33 to -3.67)
Insulin (mIU/mL)	22.36 (14.01 to 30.7)	14.49 (9.04 to 19.94)	−30.71 (−55.17 to −6.25)	−7.87 (−15.44 to −0.29)
HOMA-IR[d]	5.88 (3.69 to 8.07)	3.54 (2.11 to 4.97)	−36.75 (−60.54 to −12.96)	−2.34 (−4.30 to −0.38)
Triglycerides (mg/dL)	123.89 (86.9 to 160.87)	107.89 (64.33 to 151.45)	−14.32 (−25.91 to −2.74)	−16.00 (−28.74 to −3.26)
HDL[e] (mg/dL)	50.67 (35.63 to 65.7)	47.78 (37.57 to 57.99)	−2.50 (−14.87 to 9.88)	−2.89 (−9.62 to 3.84)
LDL[f] (mg/dL)	80.11 (55.84 to 104.38)	83.44 (65.01 to 101.88)	9.33 (−9.12 to 27.78)	3.33 (−9.75 to 16.42)
C-reactive protein (mg/dL)	1.55 (0.59 to 2.52)	1.18 (0.36 to 2.00)	−23.58 (−59.94 to 12.77)	−0.38 (−1.07 to 0.32)
CLDQ[g]				
Overall	2.33 (1.66 to 3.01)	1.68 (1.47 to 1.89)	−20.47 (−41.16 to 0.22)	−0.65 (−1.40 to 0.09)
Abdominal symptoms	2.67 (1.55 to 3.78)	1.37 (1.07 to 1.67)	−33.15 (65.25 to −1.05)	−1.30 (−2.48 to −0.11)
Activity	2.33 (1.49 to 3.17)	1.81 (1.16 to 2.47)	−9.88 (−45.53 to 25.77)	−0.52 (−1.58 to 0.55)
Emotional function	2.08 (1.58 to 2.58)	1.54 (1.21 to 1.87)	−22.55 (−39.55 to −5.55)	−0.54 (−1.11 to 0.03)
Fatigue	2.72 (1.62 to 3.83)	1.71 (1.13 to 2.29)	−27.05 (−49.50 to −4.59)	−1.01 (−2.02 to −0.01)
Systemic symptoms	2.31 (1.39 to 3.24)	1.82 (1.37 to 2.27)	−11.94 (−36.28 to 12.40)	−0.49 (−1.17 to 0.19)
Worry about MASLD[h]	2.18 (1.43 to 2.92)	1.82 (1.17 to 2.47)	−8.20 (−32.58 to 16.18)	−0.36 (−1.11 to 0.40)
Positive affect	25.56 (23.45 to 27.66)	26.78 (24.58 to 28.98)	5.01 (−1.11 to 11.13)	1.22 (−0.26 to 2.70)
Mindful eating	2.93 (2.44 to 3.41)	3.52 (3.24 to 3.79)	26.77 (−2.03 to 55.57)	0.59 (0.16 to 1.03)
Stress-based eating	1.69 (1.1 to 2.29)	1.31 (0.88 to 1.73)	−18.55 (−39.01 to 1.90)	−0.39 (−0.81 to 0.03)
Energy (kilocalories)	1752.25 (1401 to 2103.51)	1585.48 (1231.09 to 1939.88)	−6.97 (−29.49 to 15.55)	−166.77 (−551.21 to 217.67)
Carbohydrates (g)	171.37 (117.21 to 225.53)	88.53 (29.72 to 147.35)	−51.07 (−68.51 to −33.63)	−82.84 (−122.62 to −43.05)
Fiber (g)	17.35 (10.91 to 23.8)	18.88 (12.64 to 25.12)	22.28 (−28.19 to 72.76)	1.53 (−6.41 to 9.46)
Net (Non-fiber) Carbohydrates (g)	154.02 (102.54 to 205.5)	69.65 (13.66 to 125.65)	−58.62 (−76.36 to −40.89)	−84.36 (−123.63 to −45.09)
Total fat (g)	77.77 (60.71 to 94.83)	100.01 (71.62 to 128.4)	31.45 (−7.46 to 70.37)	22.24 (−4.99 to 49.47)
Monounsaturated fat (g)	28.2 (20.46 to 35.93)	34.43 (25.12 to 43.73)	25.24 (−4.70 to 55.18)	6.23 (−2.36 to 14.82)
Polyunsaturated fat (g)	13.95 (10.19 to 17.72)	18.47 (12.15 to 24.78)	41.43 (−22.53 to 105.39)	4.51 (−1.94 to 10.96)
Saturated fat (g)	28.53 (20.59 to 36.48)	38.48 (25.74 to 51.22)	38.58 (−5.55 to 82.71)	9.95 (−2.82 to 22.72)
Protein (g)	79.81 (63.11 to 96.51)	87.04 (72.31 to 101.76)	13.98 (−13.11 to 41.07)	7.22 (−13.07 to 27.51)

[a]AST: aspartate aminotransferase.

[b]ALT: alanine aminotransferase.

[c] HbA_{1c}: glycated hemoglobin.

[d]HOMA-IR: Homeostatic Model Assessment for Insulin Resistance.

XSL•FO
RenderX

[e]HDL: high-density lipoprotein.
[f]LDL: low-density lipoprotein.
[g]CLDQ: chronic liver disease questionnaire.
[h]MASLD: metabolic dysfunction–associated steatotic liver disease.

XSL•FO
RenderX

Table . Change in outcomes from baseline to post for participants who were adherent to the eating pattern. For all outcomes n=6, except for C-reactive protein (n=3).

Outcomes	Baseline, mean (95% CI)	Post, mean (95% CI)	Percent change, mean (95% CI)	Change, mean (95% CI)
Weight (kg)	96.83 (83.93 to 109.74)	85.15 (74.36 to 95.94)	−11.99 (−14.33 to −9.65)	−11.69 (−14.59 to −8.78)
BMI (kg/m^2)	34.05 (26.76 to 41.34)	29.92 (23.83 to 36.01)	−11.94 (−14.21 to −9.66)	−4.13 (−5.47 to −2.8)
Liver lobe fat (%)	21.67 (10.77 to 32.57)	10.92 (3.05 to 18.79)	−53.12 (−71.25 to −34.99)	−10.75 (−16.38 to −5.12)
Liver lobe fibrosis	2.35 (1.76 to 2.94)	1.88 (1.49 to 2.28)	−17.73 (−37.41 to 1.94)	−0.47 (−0.98 to 0.05)
AST[a] (U/L)	41.67 (31.01 to 52.32)	26.33 (20.79 to 31.87)	−33.94 (−52.93 to −14.95)	−15.33 (−25.72 to −4.95)
ALT[b] (U/L)	71 (35.6 to 106.4)	32.33 (22.52 to 42.15)	−47.49 (−70.67 to -24.31)	−38.67 (−70.38 to −6.95)
HbA_{1c}[c] (mmol/mol)	38.33 (34.48 to 42.18)	35.33 (31.6 to 39.07)	−7.73 (−13.15 to −2.32)	−3.00 (−5.10 to −0.90)
Glucose (mg/dL)	104.33 (90.18 to 118.49)	96.83 (80.54 to 113.12)	−7.36 (−15.76 to 1.05)	−7.50 (−15.65 to 0.65)
Insulin (mIU/mL)	25.3 (13.27 to 37.33)	15.08 (6.04 to 24.12)	−40.52 (−72.62 to −8.42)	−10.22 (−21.53 to 1.10)
HOMA-IR[d]	6.63 (3.43 to 9.82)	3.77 (1.45 to 6.09)	−44.02 (−77.02 to −11.03)	−2.86 (−5.79 to 0.08)
Triglycerides (mg/dL)	126.67 (64.35 to 188.98)	111.5 (38.45 to 184.55)	−13.59 (−32.80 to 5.61)	−15.17 (−36.73 to 6.40)
HDL[e] (mg/dL)	50.83 (25.58 to 76.08)	47.33 (30.03 to 64.64)	−2.60 (−23.00 to 17.80)	−3.50 (−14.46 to 7.46)
LDL[f] (mg/dL)	77 (50.1 to 103.9)	86.5 (63.52 to 109.48)	16.92 (−8.88 to 42.72)	9.50 (−5.70 to 24.70)
C-reactive protein (mg/dL)	1.83 (0.71 to 2.95)	1.43 (-0.57 to 3.44)	−17.16 (−74.65 to 40.33)	−0.40 (−1.71 to 0.91)
CLDQ[g]				
Overall	2 (1.62 to 2.38)	1.7 (1.38 to 2.02)	−12.21 (−38.44 to 14.02)	−0.30 (−0.85 to 0.24)
Abdominal symptoms	2.44 (1.07 to 3.82)	1.28 (0.87 to 1.69)	−31.67 (−81.09 to 17.76)	−1.17 (−2.61 to 0.28)
Activity	1.89 (1.17 to 2.61)	1.67 (0.73 to 2.61)	−2.92 (−56.30 to 50.46)	−0.22 (−1.41 to 0.96)
Emotional function	1.88 (1.6 to 2.15)	1.65 (1.17 to 2.12)	−12.96 (−30.42 to 4.50)	−0.23 (−0.54 to 0.09)
Fatigue	2.12 (1.02 to 3.21)	1.57 (0.8 to 2.33)	−18.50 (−48.68 to 11.67)	−0.55 (−1.41 to 0.31)
Systemic symptoms	2 (1.36 to 2.64)	1.83 (1.26 to 2.4)	−3.96 (−39.18 to 31.27)	−0.17 (−0.75 to 0.42)
Worry about liver disease	1.93 (0.99 to 2.88)	2.03 (1.02 to 3.05)	5.42 (−7.76 to 18.59)	0.10 (−0.22 to 0.42)
Positive affect	25.83 (22.42 to 29.24)	27.33 (24.24 to 30.42)	6.28 (−3.59 to 16.15)	1.50 (−0.87 to 3.87)
Mindful eating	3.03 (2.62 to 3.43)	3.56 (3.1 to 4.01)	18.12 (5.48 to 30.75)	0.53 (0.15 to 0.90)
Stress-based eating	1.71 (0.75 to 2.67)	1.29 (0.66 to 1.92)	−20.79 (−39.39 to −2.20)	−0.42 (−0.85 to 0.01)
Energy (kilocalories)	1836.48 (1298.33 to 2374.62)	1385.79 (933.4 to 1838.17)	−24.28 (−38.40 to −10.16)	−450.69 (−777.83 to −123.54)
Carbohydrates (g)	169.25 (111.67 to 226.82)	60.25 (39.73 to 80.76)	−64.14 (−72.48 to −55.79)	−109.00 (−150.88 to -67.13)
Fiber (g)	19.11 (9.35 to 28.86)	16.86 (7.49 to 26.22)	−5.84 (−61.83 to 50.15)	−2.25 (−13.11 to 8.61)
Net (Non-fiber) Carbohydrates (g)	150.14 (95.45 to 204.83)	43.39 (28.69 to 58.08)	−70.04 (−79.77 to −60.31)	−106.75 (−153.97 to −59.54)
Total fat (g)	84.12 (58.22 to 110.01)	90.09 (49.99 to 130.19)	5.53 (−30.18 to 41.24)	5.97 (−23.95 to 35.90)
Monounsaturated fat (g)	30.1 (17.68 to 42.52)	32.04 (17.99 to 46.1)	7.89 (−29.08 to 44.86)	1.94 (−9.69 to 13.57)
Polyunsaturated fat (g)	15.64 (10.15 to 21.14)	14.96 (6.91 to 23)	−9.79 (−35.91 to 16.34)	−0.69 (−3.89 to 2.52)
Saturated fat (g)	30.58 (17.87 to 43.3)	35.17 (18 to 52.34)	17.23 (−32.59 to 67,06)	4.58 (−12.01 to 21.18)
Protein (g)	86.15 (60.68 to 111.63)	80.55 (61.78 to 99.33)	−4.24 (−23.72 to 15.24)	−5.60 (−25.7 to 14.51)

[a]AST: aspartate aminotransferase.

[b]ALT: alanine aminotransferase.

[c]HbA_{1c}: glycated hemoglobin.

[d]HOMA-IR: Homeostatic Model Assessment for Insulin Resistance.

XSL•FO
RenderX

[e]HDL: high-density lipoprotein.

[f]LDL: low-density lipoprotein.

[g]CLDQ: chronic liver disease questionnaire.

Intervention Acceptability

General Intervention Satisfaction

Of the 9 participants who provided 4-month self-report information, intervention satisfaction was high (mean 6.22, 95% CI 5.58-6.85), with 5 (56%) participants rating the intervention the top score. Only 1 (11%) participant reported that they would stop the assigned eating pattern as soon as the study was over, with 4 (44%) participants reporting that they intended to continue it for at least another few months, and 4 (44%) participants stating that they did not plan to ever stop following it.

Intervention's Skills Satisfaction

The 9 participants who provided 4-month self-report information rated their satisfaction with the positive affect skills (mean 4.78, 95% CI 03.53-6.03), with 1 (11%) participant rating them the top score. Participants rated their satisfaction with the mindfulness skills (mean 4.44, 95% CI 3.09-5.79), with 1 (11%) participant rating them the top score.

Qualitative Feedback

Participants noted barriers that reduced their adherence to and motivation for the intervention, including lack of time, carbohydrate cravings, fatigue, vacations, and social situations. They mentioned factors that supported their adherence to and motivation for the intervention, including accountability by being in the intervention, the program's supportive coach, feeling that the "diet was manageable," weight loss success, and an overall desire to improve their own health. Opinions were mixed about the positive affect and mindful eating skills: one person said that they did not need them and another thought that the "skill building around mindfulness is essential."

Participants noted positive health changes such as weight loss, increased energy, and a changed attitude about food and the "need/desire to eat food that is not good for me." They noted that their family, friends, and physicians were impressed with their weight loss, health changes, and ability to stick with the program. Some comments included, "My physician noticed I lost weight and said I looked much better. She also encouraged me to continue this after the program is over." "My primary care provider reported he was happy for the change and that my blood tests show I am now healthier than ever."

Secondary Patient-Centered Outcomes

Adverse Events

No serious adverse events were reported, but adverse events, possibly unrelated, included diarrhea (later discovered to be a Clostridioides difficile infection), a flare-up of dermatitis, and a broken foot.

Change in Liver Percent Fat, Body Weight, and Blood Tests

In Figure 1 we show the changes in liver lobe fat percent across participants. In Table 2 we show the changes with values for all available participants. For the participants with this outcome, the percent change in liver fat was −33.17% (95% CI −86.48 to 20.14). In Table 3 we show these changes for only the participants who were adherent to the eating pattern. For the adherent participants, the percent change in liver fat was −53.12% (95% CI −71.25 to −34.99).

XSL•FO
RenderX

Figure 1. Change in liver lobe fat percent from baseline to post across participants.

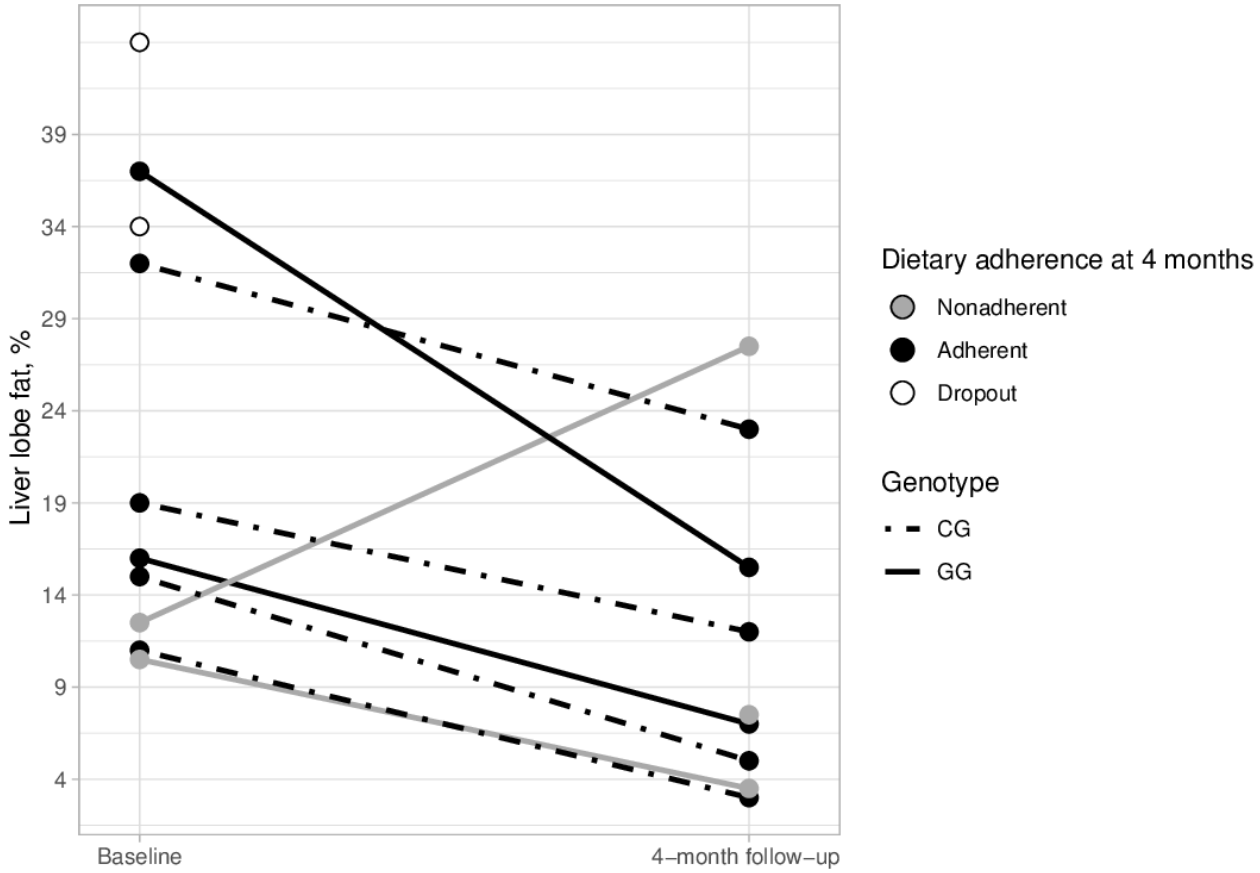

Achievement of ≥30% Liver Fat Loss

Amongst participants with a 4-month hepatic liver fat percent measurement, 6 out of 8 (75%) participants were considered responders, with a relative decline in liver fat ≥30%, with one person not counted losing 28%.

Achievement of ≥5% Body Weight Loss

Amongst participants with a 4-month body weight, 9 out of 9 (100%) of participants lost at least 5% of their body weight.

Self-Rated Change in Health

Two participants reported their health as having gotten a little better, 6 reported that their health was much better, and 1 reported that their health was very much better.

Changes in Chronic Liver Disease Questionnaire and Psychological Factors

Across all participants (Table 2) and for participants who were adherent (Table 3), changes were in the expected, salutary direction.

A summary of the study design and results is shown in Figure 2.

XSL•FO
RenderX

Figure 2. Graphical table of contents. PNPLA3: phospholipase domain–containing protein 3; VLC: very low–carbohydrate.

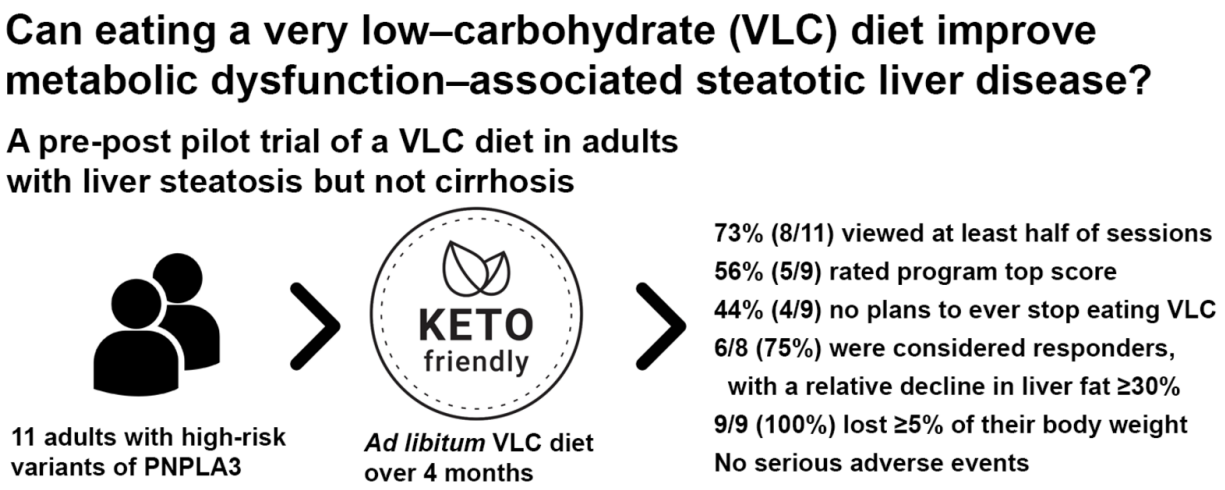

Discussion

Principal Findings

Results suggest the feasibility, acceptability, and preliminary efficacy of the VLC intervention in adults with higher genetic risk for MASLD, although there is a need for further studies given the small sample size and the high risk of substantial biases in this small pilot study. One concern of the VLC diet is its tolerability. Here we show that the diet was well tolerated with 5 out of 9 participants rating it at the highest satisfaction score. Only 1 out of 9 said they would stop the assigned dietary pattern as soon as the study was over. In contrast, 4 out of 9 said they would continue to follow the diet indefinitely. This is consistent with data where close to 60% of individuals with relapsing multiple sclerosis continued on a VLC or low-carb diet following a trial of the diet, particularly if they had lost a lot of weight and had decreased fatigue [36]. The mindfulness training was less helpful to participants with only 1 out of 9 rating it with the top score. They noted challenges with adhering to the diet as time, carbohydrate cravings, fatigue, vacations, and social situations which need to be addressed as part of a larger program to change eating patterns.

Several previous trials have examined the impact of a VLC eating pattern for adults with MASLD or metabolic dysfunction–associated steatohepatitis, and they found similar results. For example, in a 1-month trial of a VLC diet with 27 participants with MASLD, left hepatic lobe volume decreased by 20% [37]. In a 3-month trial of a Spanish and Mediterranean–adapted VLC diet in 12 men, liver steatosis improved in 93% of participants, and complete fatty liver regression was observed in 21% of participants [12]. In a 6-month trial of a VLC diet with 5 participants, 4 (80%) showed improvements in liver histology, steatosis, inflammatory grade, and fibrosis [10]. In a 2-month trial of a very low-calorie, VLC diet with 20 participants with obesity, the prevalence of liver steatosis (proton density fat fraction>5.4%) decreased from 70% to 30%. Likewise, 2 other nonrandomized, pre-post trials of hypocaloric VLC diets, a 6.5-month [38] and an 8-week trial, both showed improvements in liver steatosis [39].

Five previous trials, with a total of 179 people, have randomized with MASLD to a Mediterranean diet and measured liver-based fat with magnetic resonance imaging. These varied in length of time, with 1 being 1.5 months long, 2 being 3 months long, and 3 being 6 months long. Overall, they have found an average of 12% - 60% relative reduction in hepatic fat percentage [40-44]. Our results and the results of previous VLC trials suggest that a VLC, ketogenic diet may be at least as effective as a Mediterranean diet, although future randomized trials are needed to assess this. Our study is also in line with a previous study where carriers of PNPLA3 G risk allele were able to lose substantial liver fat on a low carbohydrate diet [45].

XSL•FO
RenderX

Limitations

One of our major limitations is the fact that our sample size was small. Our recruitment was limited by the fact that we reached out to adults who had already been genotyped by the MGI. Instead, if we had genotyped interested participants, it may have been easier to recruit more participants. We also required that some of the study-related tests, which could be justified as medical expenses as part of regular health care, including the baseline magnetic resonance imaging and most of the blood tests, be covered by the participants' health insurance. However, not all potential participants were willing to add these charges to their insurance, because it would still incur costs to the participants. In addition, the trial was conducted during the COVID-19 pandemic, which may have influenced who was willing or able to participate or their ability to adhere to dietary instructions.

Moreover, this trial's generalizability is limited, as most participants were White, non-Hispanic men. Future recruitment efforts should focus on recruiting a more diverse population. This trial lacked a control group to which participants could have been randomized to, and thus statements cannot be made about causality.

Another important limitation is duration. As with other trials in this area, participants were tracked for less than a year, which does not allow for the understanding of the longer-term impacts of this nutritional approach nor whether participants may maintain dietary adherence longer term [10,12,37-44]. A growing body of literature including our paper showed that both insulin resistance and the PNPLA3 risk allele can drive the progression of liver disease [46,47]. A limitation of our study is that whether PNPLA3-driven liver fat accumulation causes liver insulin resistance cannot be determined here. Finally, the lack of a PNPLA3 CC control group limits our ability to determine whether the outcomes would be even more improved in the G versus C carriers.

Conclusions and Future Directions

Thus, results supported the feasibility, acceptability, and preliminary efficacy of the VLC intervention in adults thought to have a higher genetic risk for MASLD. Overall, there is a need for longer-term studies to assess the sustainability of benefits and potential long-term risks associated with this way of eating. Moreover, future studies should aim to include a more diverse population, explore the biological mechanisms by which the VLC diet affects MASLD, consider the role of the PNPLA3 genotype, and integrate omics technologies to identify modulated biomarkers and pathways, thereby enhancing the generalizability and depth of the findings.

Acknowledgments

We thank our dedicated participants, Lara Khadr for her help with the figure, Samuel Handelman for his help with the initial grant, and the Michigan Genomics Initiative. For this research, LRS and EKS were supported by R03 DK124741 and P30 DK089503, and Core Services was used, supported by grant DK089503 to the University of Michigan. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authors' Contributions

LRS and EKS were involved in the conceptualization, project administration, and methodology of the study. AO and JK were involved in the project administration of the study. KR was the dietary coach. LRS completed the analyses with help from DJM. All authors reviewed and assisted with writing revisions. The authors read and approved the final manuscript.

Conflicts of Interest

LRS' partner, HB, is an inventor of the software used in this study, which purchased a software services agreement for its use. The other authors have no conflicts of interest to declare.

References

1. Amiri Dash Atan N, Koushki M, Motedayen M, et al. Type 2 diabetes mellitus and non-alcoholic fatty liver disease: a systematic review and meta-analysis. Gastroenterol Hepatol Bed Bench 2017;10(Suppl1):S1-S7. [Medline: 29511464]
2. Lazo M, Hernaez R, Eberhardt MS, et al. Prevalence of nonalcoholic fatty liver disease in the United States: the third national health and nutrition examination survey, 1988-1994. Am J Epidemiol 2013 Jul 1;178(1):38-45. [doi: 10.1093/aje/kws448] [Medline: 23703888]
3. Crespo J, Fernández-Gil P, Hernández-Guerra M, et al. Are there predictive factors of severe liver fibrosis in morbidly obese patients with non-alcoholic steatohepatitis? Obes Surg 2001 Jun 1;11(3):254-257. [doi: 10.1381/096089201321336548]
4. Portillo P, Yavuz S, Bril F, Cusi K. Role of insulin resistance and diabetes in the pathogenesis and treatment of nonalcoholic fatty liver disease. Curr Hepatology Rep 2014 Jun;13(2):159-170. [doi: 10.1007/s11901-014-0229-3]
5. Charlton MR, Burns JM, Pedersen RA, Watt KD, Heimbach JK, Dierkhising RA. Frequency and outcomes of liver transplantation for nonalcoholic steatohepatitis in the United States. Gastroenterology 2011 Oct;141(4):1249-1253. [doi: 10.1053/j.gastro.2011.06.061] [Medline: 21726509]

XSL•FO
RenderX

6. Katsagoni CN, Papachristou E, Sidossis A, Sidossis L. Effects of dietary and lifestyle interventions on liver, clinical and metabolic parameters in children and adolescents with non-alcoholic fatty liver disease: a systematic review. Nutrients 2020 Sep 19;12(9):2864. [doi: 10.3390/nu12092864] [Medline: 32961669]
7. Parra-Vargas M, Rodriguez-Echevarria R, Jimenez-Chillaron JC. Nutritional approaches for the management of nonalcoholic fatty liver disease: an evidence-based review. Nutrients 2020 Dec 17;12(12):3860. [doi: 10.3390/nu12123860] [Medline: 33348700]
8. Saeed N, Nadeau B, Shannon C, Tincopa M. Evaluation of dietary approaches for the treatment of non-alcoholic fatty liver disease: a systematic review. Nutrients 2019 Dec 16;11(12):3064. [doi: 10.3390/nu11123064] [Medline: 31888132]
9. Speliotes EK, Yerges-Armstrong LM, Wu J, et al. Genome-wide association analysis identifies variants associated with nonalcoholic fatty liver disease that have distinct effects on metabolic traits. PLoS Genet 2011 Mar;7(3):e1001324. [doi: 10.1371/journal.pgen.1001324] [Medline: 21423719]
10. Tendler D, Lin S, Yancy WS Jr, et al. The effect of a low-carbohydrate, ketogenic diet on nonalcoholic fatty liver disease: a pilot study. Dig Dis Sci 2007 Feb;52(2):589-593. [doi: 10.1007/s10620-006-9433-5] [Medline: 17219068]
11. Mardinoglu A, Wu H, Bjornson E, et al. An integrated understanding of the rapid metabolic benefits of a carbohydrate-restricted diet on hepatic steatosis in humans. Cell Metab 2018 Mar 6;27(3):559-571. [doi: 10.1016/j.cmet.2018.01.005] [Medline: 29456073]
12. Pérez-Guisado J, Muñoz-Serrano A. The effect of the Spanish ketogenic mediterranean diet on nonalcoholic fatty liver disease: a pilot study. J Med Food 2011;14(7-8):677-680. [doi: 10.1089/jmf.2011.0075] [Medline: 21688989]
13. Sevastianova K, Santos A, Kotronen A, et al. Effect of short-term carbohydrate overfeeding and long-term weight loss on liver fat in overweight humans. Am J Clin Nutr 2012 Oct;96(4):727-734. [doi: 10.3945/ajcn.112.038695] [Medline: 22952180]
14. Hudgins LC, Hellerstein M, Seidman C, Neese R, Diakun J, Hirsch J. Human fatty acid synthesis is stimulated by a eucaloric low fat, high carbohydrate diet. J Clin Invest 1996 May 1;97(9):2081-2091. [doi: 10.1172/JCI118645] [Medline: 8621798]
15. Solga S, Alkhuraishe AR, Clark JM, et al. Dietary composition and nonalcoholic fatty liver disease. Dig Dis Sci 2004 Oct;49(10):1578-1583. [doi: 10.1023/b:ddas.0000043367.69470.b7] [Medline: 15573908]
16. Kang H, Greenson JK, Omo JT, et al. Metabolic syndrome is associated with greater histologic severity, higher carbohydrate, and lower fat diet in patients with NAFLD. Am J Gastroenterol 2006 Oct;101(10):2247-2253. [doi: 10.1111/j.1572-0241.2006.00719.x] [Medline: 17032189]
17. Czajkowski SM, Powell LH, Adler N, et al. From ideas to efficacy: the ORBIT model for developing behavioral treatments for chronic diseases. Health Psychol 2015 Oct;34(10):971-982. [doi: 10.1037/hea0000161] [Medline: 25642841]
18. Carr A, Cullen K, Keeney C, et al. Effectiveness of positive psychology interventions: a systematic review and meta-analysis. J Posit Psychol 2021 Nov 2;16(6):749-769. [doi: 10.1080/17439760.2020.1818807]
19. Fredrickson BL, Joiner T. Positive emotions trigger upward spirals toward emotional well-being. Psychol Sci 2002 Mar;13(2):172-175. [doi: 10.1111/1467-9280.00431] [Medline: 11934003]
20. Mason AE, Epel ES, Aschbacher K, et al. Reduced reward-driven eating accounts for the impact of a mindfulness-based diet and exercise intervention on weight loss: data from the SHINE randomized controlled trial. Appetite 2016 May 1;100:86-93. [doi: 10.1016/j.appet.2016.02.009] [Medline: 26867697]
21. Koenders PG, van Strien T. Emotional eating, rather than lifestyle behavior, drives weight gain in a prospective study in 1562 employees. J Occup Environ Med 2011 Nov;53(11):1287-1293. [doi: 10.1097/JOM.0b013e31823078a2] [Medline: 22027541]
22. Lazarevich I, Irigoyen Camacho ME, Velázquez-Alva MDC, Zepeda Zepeda M. Relationship among obesity, depression, and emotional eating in young adults. Appetite 2016 Dec 1;107:639-644. [doi: 10.1016/j.appet.2016.09.011] [Medline: 27620648]
23. Willmott TJ, Pang B, Rundle-Thiele S, Badejo A. Weight management in young adults: systematic review of electronic health intervention components and outcomes. J Med Internet Res 2019 Feb 6;21(2):e10265. [doi: 10.2196/10265] [Medline: 30724736]
24. Zawistowski M, Fritsche LG, Pandit A, et al. The Michigan genomics initiative: a biobank linking genotypes and electronic clinical records in Michigan medicine patients. Cell Genom 2023 Feb 8;3(2):100257. [doi: 10.1016/j.xgen.2023.100257] [Medline: 36819667]
25. MarcAurele L, Volo L. Low Carb Yum 5-Ingredient Keto: 120+ Easy Recipes: Houghton Mifflin Harcourt; 2020:222.
26. Ketchum C. The Everyday Ketogenic Kitchen: 150+ Inspirational Low-Carb, High-Fat Recipes to Maximize Your Health: Victory Belt Publishing Inc; 2017:384.
27. Krampf M. The Wholesome Yum Easy Keto Cookbook: 100 Simple Low Carb Recipes: 10 Ingredients or Less: Harmony Books; 2019:256.
28. Saslow LR, Moskowitz JT, Mason AE, et al. Intervention enhancement strategies among adults with type 2 diabetes in a very low-carbohydrate web-based program: evaluating the impact with a randomized trial. JMIR Diabetes 2020 Sep 9;5(3):e15835. [doi: 10.2196/15835] [Medline: 32902391]

XSL•FO
RenderX

29. Stine JG, Munaganuru N, Barnard A, et al. Change in MRI-PDFF and histologic response in patients with nonalcoholic steatohepatitis: a systematic review and meta-analysis. Clin Gastroenterol Hepatol 2021 Nov;19(11):2274-2283. [doi: 10.1016/j.cgh.2020.08.061] [Medline: 32882428]
30. Blackburn G. Effect of degree of weight loss on health benefits. Obes Res 1995 Sep;3 Suppl 2:211s-216s. [doi: 10.1002/j.1550-8528.1995.tb00466.x] [Medline: 8581779]
31. Younossi ZM, Guyatt G, Kiwi M, Boparai N, King D. Development of a disease specific questionnaire to measure health related quality of life in patients with chronic liver disease. Gut 1999 Aug;45(2):295-300. [doi: 10.1136/gut.45.2.295] [Medline: 10403745]
32. Diener E, Wirtz D, Biswas-Diener R, et al. New measures of well-being. In: Assess Well-Being Springer: Springer; 2009:247-266. [doi: 10.1007/978-90-481-2354-4_12]
33. Tylka TL, Kroon Van Diest AM. The intuitive eating scale-2: item refinement and psychometric evaluation with college women and men. J Couns Psychol 2013 Jan;60(1):137-153. [doi: 10.1037/a0030893] [Medline: 23356469]
34. Burgess EE, Turan B, Lokken KL, Morse A, Boggiano MM. Profiling motives behind hedonic eating. preliminary validation of the palatable eating motives scale. Appetite 2014 Jan;72:66-72. [doi: 10.1016/j.appet.2013.09.016] [Medline: 24076018]
35. Rovetta A, Mansournia MA, Vitale A. For a proper use of frequentist inferential statistics in public health. Glob Epidemiol 2024 Dec;8:100151. [doi: 10.1016/j.gloepi.2024.100151] [Medline: 39021384]
36. Merkel M, Velez-Carrasco W, Hudgins LC, Breslow JL. Compared with saturated fatty acids, dietary monounsaturated fatty acids and carbohydrates increase atherosclerosis and VLDL cholesterol levels in LDL receptor-deficient, but not apolipoprotein E-deficient, mice. Proc Natl Acad Sci U S A 2001 Nov 6;98(23):13294-13299. [doi: 10.1073/pnas.231490498] [Medline: 11606787]
37. Schiavo L, Pilone V, Rossetti G, Barbarisi A, Cesaretti M, Iannelli A. A 4-week preoperative ketogenic micronutrient-enriched diet is effective in reducing body weight, left hepatic lobe volume, and micronutrient deficiencies in patients undergoing bariatric surgery: a prospective pilot study. Obes Surg 2018 Aug;28(8):2215-2224. [doi: 10.1007/s11695-018-3145-8] [Medline: 29502279]
38. Belopolsky Y, Khan MQ, Sonnenberg A, Davidson DJ, Fimmel CJ. Ketogenic, hypocaloric diet improves nonalcoholic steatohepatitis. J Transl Int Med 2020 Mar;8(1):26-31. [doi: 10.2478/jtim-2020-0005] [Medline: 32435609]
39. Rinaldi R, De Nucci S, Castellana F, et al. The effects of eight weeks' very low-calorie ketogenic diet (VLCKD) on liver health in subjects affected by overweight and obesity. Nutrients 2023 Feb 6;15(4):825. [doi: 10.3390/nu15040825] [Medline: 36839183]
40. Ryan MC, Itsiopoulos C, Thodis T, et al. The Mediterranean diet improves hepatic steatosis and insulin sensitivity in individuals with non-alcoholic fatty liver disease. J Hepatol 2013 Jul;59(1):138-143. [doi: 10.1016/j.jhep.2013.02.012] [Medline: 23485520]
41. Properzi C, O'Sullivan TA, Sherriff JL, et al. Ad libitum mediterranean and low-fat diets both significantly reduce hepatic steatosis: a randomized controlled trial. Hepatology 2018 Nov;68(5):1741-1754. [doi: 10.1002/hep.30076] [Medline: 29729189]
42. Marin-Alejandre BA, Abete I, Cantero I, et al. The metabolic and hepatic impact of two personalized dietary strategies in subjects with obesity and nonalcoholic fatty liver disease: the fatty liver in obesity (FLiO) randomized controlled trial. Nutrients 2019 Oct 22;11(10):2543. [doi: 10.3390/nu11102543] [Medline: 31652512]
43. Abbate M, Mascaró CM, Montemayor S, et al. Energy expenditure improved risk factors associated with renal function loss in NAFLD and MetS patients. Nutrients 2021 Feb 15;13(2):629. [doi: 10.3390/nu13020629] [Medline: 33672073]
44. George E, Reddy A, Nicoll A, et al. Impact of a Mediterranean diet on hepatic and metabolic outcomes in non-alcoholic fatty liver disease: the MEDINA randomized controlled trial. J Gastroenterol Hepatol 2021;36(3):160-161. [doi: 10.1111/liv.15264]
45. Sevastianova K, Kotronen A, Gastaldelli A, et al. Genetic variation in PNPLA3 (adiponutrin) confers sensitivity to weight loss-induced decrease in liver fat in humans. Am J Clin Nutr 2011 Jul;94(1):104-111. [doi: 10.3945/ajcn.111.012369] [Medline: 21525193]
46. Barata L, Feitosa MF, Bielak LF, et al. Insulin resistance exacerbates genetic predisposition to nonalcoholic fatty liver disease in individuals without diabetes. Hepatol Commun 2019 Jul;3(7):894-907. [doi: 10.1002/hep4.1353] [Medline: 31334442]
47. Luukkonen PK, Qadri S, Ahlholm N, et al. Distinct contributions of metabolic dysfunction and genetic risk factors in the pathogenesis of non-alcoholic fatty liver disease. J Hepatol 2022 Mar;76(3):526-535. [doi: 10.1016/j.jhep.2021.10.013] [Medline: 34710482]

Abbreviations

CLDQ: chronic liver disease questionnaire
MASLD: metabolic dysfunction–associated steatotic liver disease
MGI: Michigan Genomics Initiative
PNPLA3: phospholipase domain–containing protein 3

XSL•FO
RenderX

VLC: very low–carbohydrate

Edited by A Rovetta; submitted 30.04.24; peer-reviewed by D Kaminska, GP Caviglia; revised version received 24.09.24; accepted 10.10.24; published 10.01.25.

Please cite as:
Saslow LR, Krinock J, O'Brien A, Raymond K, Bayandorian H, Moskowitz JT, Daubenmier J, Oliveri A, Marriott DJ, Griauzde DH, Speliotes EK
A Very Low–Carbohydrate Program in Adults With Metabolic Dysfunction–Associated Steatotic Liver Disease and Phospholipase Domain–Containing Protein 3 Risk Genotype: Pre-Post Intervention Study
JMIR Form Res 2025;9:e60051
URL: https://formative.jmir.org/2025/1/e60051
doi:10.2196/60051

© Laura R Saslow, Jamie Krinock, Alison O'Brien, Kaitlyn Raymond, Hovig Bayandorian, Judith T Moskowitz, Jennifer Daubenmier, Antonio Oliveri, Deanna J Marriott, Dina H Griauzde, Elizabeth K Speliotes. Originally published in JMIR Formative Research (https://formative.jmir.org), 10.1.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Continuous Glucose Monitoring Among Patients With Type 1 Diabetes in Rwanda (CAPT1D) Phase I: Prospective Observational Feasibility Study

Jason Baker[1], MD; Giacomo Cappon[2], PhD; Jean Claude Habineza[3,4], MD; Corey H Basch[1], MPH, EdD; Steven Mey[1], BS; Diana L Malkin-Washeim[1], PhD; Christian Schuetz[1], MD, PhD; Niyonsenga Simon Pierre[5], MSc; Etienne Uwingabire[3], BSN; Alvera Mukamazimpaka[3], BSc; Paul Mbonyi[3], MPH; Sandhya Narayanan[1], EdD

[1]Diabetes Empowerment International, New York, NY, United States

[2]Department of Information Engineering, University of Padova, Padova, Italy

[3]Rwanda Diabetes Association, Kigali, Rwanda

[4]School of Health Sciences, University of Rwanda, Kigali, Rwanda

[5]Rwanda Biomedical Center, Kigali, Rwanda

Corresponding Author:
Sandhya Narayanan, EdD
Diabetes Empowerment International, New York, NY, United States

Abstract

Background: The development of minimally invasive continuous glucose monitoring systems (CGMs) has transformed diabetes management. CGMs have shown clinical significance by improving time in the euglycemic range, decreasing rates of hypoglycemia, and improving hemoglobin A_{1c} (HbA_{1c}). In Rwanda, CGMs are currently not routinely used, and no clinical studies of CGM use were identified in the literature.

Objective: This study aims to determine the impact and feasibility of real-time CGM use among people living with type 1 diabetes (T1D) in Rwanda through assessment of sensor use, time in range, rates of hypoglycemia and hyperglycemia, HbA_{1c}, and rates of diabetes-related hospitalizations over time.

Methods: The Continuous Glucose Monitoring Among Patients with Type 1 Diabetes in Rwanda (CAPT1D) study is a single-arm, prospective observational study conducted at the Rwandan Diabetes Association clinic in Kigali, Rwanda, aiming to assess the impact and feasibility of CGM use in Rwanda. A cohort of 50 participants diagnosed with T1D were enrolled. Participants were at least 21 years old, undergoing multiple daily insulin therapy, and not currently pregnant. Phase I of the study was conducted over 12 months, using the Dexcom G6 CGM. Phase II and Phase III extended CGM use for an additional 6 months respectively, using the next-generation Dexcom G7 CGM. Here, we report the quantitative results of the Phase I study.

Results: Participants used the sensor for >80% of the time throughout the study period. A significant increase in time in range was observed within 3 months and sustained over 12 months. HbA_{1c} decreased significantly in 3 months and stayed lower throughout the 12-month period. Mean HbA_{1c} levels decreased by 2.8% at 6 months ($P<.001$) and 3.2% at 12 months ($P<.001$). A total of 12 diabetes-related hospitalizations were reported during the study period. No cases of diabetic ketoacidosis or episodes of severe hypoglycemia occurred.

Conclusions: Significant and meaningful improvements in key glycemic indices indicate the potential feasibility and impact of a CGM among people living with T1D in Rwanda. Future studies could be designed to include pre- and postintervention analysis to determine the effectiveness in terms of complications and costs.

(JMIR Form Res 2025;9:e64585) doi:10.2196/64585

KEYWORDS

CGM; feasibility; type 1 diabetes; self-management; complications; Rwanda; time in range; hypoglycemia; continuous glucose monitoring; single-arm; feasibility study; diabetes; diabetes mellitus; euglycemic; HbA1c; observational study; quantitative; adult; health; public health; health informatics

XSL•FO
RenderX

Introduction

The International Diabetes Federation reports that there are 537 million adults aged 20‑79 years living with diabetes worldwide, with estimates that this number will increase to 700 million by 2045 [1]. There are more than 19 million adults with diabetes living in the African region [1], and in the International Diabetes Federation's latest report, Rwanda had a prevalence of diabetes registered at 2.7% (168,900/6,358,500) [2]. For those people with diabetes classified as having type 1 diabetes (T1D), the only treatment is through the administration of insulin exogenously. In addition, people with type 2 diabetes may need intensive insulin therapy with disease progression or poor management with noninsulin therapies. In both cases, glucose monitoring is necessary [3] to prevent or treat acute (hypoglycemia, hyperglycemia, and diabetic ketoacidosis [DKA]) or chronic (microvacular and macrovascular) complications, and more frequent monitoring has been shown to have better diabetes outcomes [4].

Traditional methods of glucose monitoring among people with diabetes have been through finger-stick blood sugar checks, referred to as self-monitoring blood glucose (SMBG), involving pricking one's finger with a small needle at different times during the day. The blood drawn is then transferred to a test strip which is read by a blood glucose monitor to display the glucose value at that point in time. There may be several obstacles related to SMBG—an invasive method that can be accompanied by discomfort and inconvenience [5,6] and relies on glucose meters that vary in accuracy [3]. Furthermore, SMBG provides very limited insight into glycemic trends, as only a few data points per day are available for decision-making [6].

In the absence of regular information regarding glucose levels, the primary way to understand how well diabetes is managed is through a blood test known as the hemoglobin A_{1c} (HbA_{1c}) test, which measures the percentage of glycated hemoglobin in red blood cells. This is an average over a 3‑ to 4-month period and provides a retrospective evaluation of therapeutic and self-management efficacy [7]. As such, it does not provide meaningful information regarding frequent or sustained high or low blood glucose values. Therefore, it is not the best tool to assist people with diabetes with treatment interventions for daily fluctuations in blood glucose [8].

The development of minimally invasive continuous glucose monitoring systems (CGMs) has transformed diabetes management [9]. CGMs provide continuous readings throughout the day and provide meaningful insight to people with diabetes regarding the time in range (TIR), as well as alerts when the glucose values veer from the range into higher (time above range [TAR]) or lower (time below range [TBR]) levels. These alerts then allow for immediate treatment action, thereby reducing the time spent in the times outside of the euglycemic range [8].

CGMs have shown clinical significance by improving time in the euglycemic range, decreasing rates of hypoglycemia and HbA_{1c} levels [10], and lowering rates of emergency department visits and hospitalizations for hypoglycemia [11]. Improved outcomes have also been shown when a CGM is introduced in the first year after diagnosis with the improvements sustained over time, for as long as 7 years [12].

In Rwanda, CGMs are currently not routinely used, and no clinical studies of CGM use were identified in the literature. The study objectives were to determine the impact and feasibility of real-time CGM use among people living with T1D in Rwanda through assessment of sensor use; TIR; HbA_{1c}; and rates of hypoglycemia, diabetes-related hospitalizations (DRHs), and DKA.

Methods

Study Design

The Continuous Glucose Monitoring Among Patients with Type 1 Diabetes in Rwanda (CAPT1D) study is a single-arm, prospective observational study conducted at the Rwandan Diabetes Association (RDA) clinic in Kigali, Rwanda, aiming to assess the impact and feasibility of CGM use in Rwanda.

The study protocol involved a run-in phase with blinded CGM (Dexcom G6, Dexcom Inc), during which participants received instructions in August 2022. Subsequently, continuous glucose monitoring was unblinded, and participants underwent a 12-month period during which they visited the RDA clinic every 20 days to upload CGM data onto the Dexcom Clarity (Dexcom Inc) platform. Participants were queried for and self-reported episodes of DKA (date, symptoms, treatment), instances of hypoglycemia, and DRHs at each study visit.

Study Sample and Recruitment

A convenience sample of established patients of the RDA was selected using a 2-stage process. First, the RDA study staff created a list of patients who met the inclusion criteria, as follows: older than 21 years, not pregnant, and treated with multiple daily doses of insulin. Second, this list was polled to identify those participants who were willing and able to commute to the RDA clinic as frequently as needed for orientation, to change sensors, and to upload data. Individuals were informed prior to enrollment that they would not be paid to participate in the study, but that transportation costs to and from the clinic will be reimbursed. The first 50 participants who agreed to participate were then enrolled in the study. The sample size was determined by the capacity of the staff and the availability of equipment. Phase I of the study was conducted over 12 months, using the Dexcom G6 CGM. Phase II and Phase III extended CGM use for an additional 6 months respectively, using the next-generation Dexcom G7 CGM. The number of participants was reduced to 47 when 3 participants withdrew from the study at month 6. Thereafter, the study sample was further reduced to 42 as a result of 5 participants joining another study where they received additional education possibly biasing the results of CAPT1D. To handle the fact that several patients were removed from month 6 on, statistical analysis was adjusted with the last observation carried forward method, that is, the last available measurement for a patient is carried forward to later time points. Here we report the quantitative results of the Phase I study.

XSL•FO
RenderX

Ethical Considerations

Ethical approval was obtained from the University of Rwanda School of Medicine Ethical Committee, and the study was registered with the institutional review board under number #271/CMHSIRB/2022. Participants were provided with all details regarding the study's objectives, procedures, potential risks, and benefits before requesting consent. At this time, any questions were addressed, and privacy and confidentiality policies were explained. A consent form outlining the study's scope, potential risks, the voluntary nature of participation, and the right to withdraw at any time was provided and signed by each participant. Additionally, consent to use photographs and other recorded material in published material was obtained. To this end, all data collected for this study have been fully anonymized and deidentified to ensure the privacy and confidentiality of all participants, and to prevent linkage of the data to individual participants. Participants were compensated varied amounts for their transportation to and from the clinic, provided light snacks and beverages during clinic visits, and received the CGMs for no cost. There is no identification of individual participants/users in any aspect of this manuscript.

Statistical Methods

CGM data were preprocessed and analyzed using AGATA [13], an open-source software tool that computes glucose control metrics in accordance with the definitions of the International Consensus on CGM-derived metrics for clinical trials [14]. Initially, CGM data were preprocessed by aligning each data point on a uniform 5-minute time grid and linearly interpolating data gaps up to 30 minutes.

Subsequently, AGATA was used to extract a set of well-established metrics [14] quantifying glucose control, including 5 metrics for glucose variability (mean, SD, coefficient of variation, glucose management indicator, and glucose risk index); 6 metrics describing the percentage of time spent in specific glucose ranges (level 1 and level 2 TBR [TBR_{L1} and TBR_{L2}], defined as <70 mg/dL and <55 mg/dL respectively; level 1 and level 2 TAR [TAR_{L1} and TAR_{L2}], defined as >180 mg/dL and >250 mg/dL, respectively; TIR, defined as 70-180 mg/dL; and time in tight range [TITR], defined as 70-140 mg/dL); and a metric quantifying CGM use, namely the percentage of time the CGM was active in the considered period ($\%CGM_{USE}$).

To align CGM data to HbA_{1c} samples collected every 3 months, each metric has been calculated by aggregating CGM data over 3-month periods and compared with the baseline computed during the initial run-in session at month 0.

Hereafter, for the sake of brevity, we will label M0 the baseline run-in session, while we will use M3, M6, M9, and M12 to represent the metric values computed using data from M0 to month 3, month 3 to month 6, and so forth.

Results were presented as the mean (SD) if the metric distribution followed a Gaussian distribution (determined by the Lillefors test at a significance level of 5%), or as the median (IQR) if not. Changes were evaluated by comparing M0 versus M6, M6 versus M12, and M0 versus M12 using a 2-tailed *t* test with a significance level of 5% for Gaussian-distributed metrics. Alternatively, the Wilcoxon rank sum test with a significance level of 5% was used.

Results

The results are illustrated in Table 1. Participants used the sensor for >80% of the time throughout the study period. TIR was defined as 70 - 180 mg/dL. The absolute change from baseline to 6 months for TIR was an increase of 9.9% (P<.001), and from baseline to 12 months, the absolute change was an increase of 14.7% (P<.001). Thus, the significant improvement in TIR at month 6 was maintained (indeed increased) at 12 months. TITR was defined as 70 - 140 mg/dL. The absolute change from baseline to 6 months for TITR was an increase of 7.1%, and that from baseline to 12 months was an increase of 10.4% (P<.001) (Figure 1). Time in hypoglycemia or TBR was defined as <70 mg/dL. From baseline to 6 months, this increased by 0.7%, which was not statistically significant (P=.31), and from baseline to 12 months, there was an increase of 2.5% (P<.001). The TAR is defined as >180 mg/dL. The absolute change from baseline to 6 months for TAR was a decrease of 9.7% (P<.01), and that from baseline to 12 months was a decrease of 15.3% (P<.001). Mean HbA_{1c} levels decreased by 2.8% at 6 months (P<.001) and 3.2% at 12 months (P<.001) (Figure 2). A total of 12 DRHs were reported during the study period. More detailed analysis of these hospitalizations is underway. No cases of DKA or episodes of severe hypoglycemia occurred.

XSL•FO
RenderX

Table . Glycemic indicators at baseline and at 3, 6, 9, and 12 months from the Continuous Glucose Monitoring Among Patients with Type 1 Diabetes in Rwanda study, a single-arm, prospective observational study conducted at the Rwandan Diabetes Association clinic in Kigali, Rwanda (2022 - 2023).

Metrics	Baseline (N=46)		M3 (N=47)		M6 (N=47)		M9 (N=42)		M12 (N=42)		*P* value (baseline vs M12)	*P* value (baseline vs M6)	*P* value (M6 vs M12)
	Mean (SD)	Median (IQR)	Mean (SD)	Median (IQR)	Mean (SD)	Median (IQR)	Mean (SD)	Median (IQR)	Mean (SD)	Median (IQR)			
Percentage of sensor data	__[a]	99.8941 (98.7353 - 100)	—	84.4136 (77.8819 - 91.8981)	—	83.8927 (73.1684 - 93.2301)	—	86.5239 (81.3619 - 95.2238)	—	88.0305 (73.5031 - 95.7562)	<.001	<.001	.49
Time in range (%)	30.9998 (15.0695)	—	42.4303 (14.683)	—	40.8986 (12.8338)	—	44.0248 (12.4312)	—	45.6591 (12.0551)	—	<.001	<.001	.08
Time in tight range (%)	18.098 (9.5095)	—	25.8216 (10.056)	—	25.2415 (8.7807)	—	27.1875 (8.7808)	—	28.4829 (8.6731)	—	<.001	<.001	.08
Time below range (%)	—	2.369 (0.1768 - 6.4643)	—	2.146 (1.2232 - 6.1304)	—	3.0203 (1.054 - 6.4066)	—	2.9182 (1.2722 - 7.7068)	—	4.8789 (1.602 - 8.9319)	<.001	.31	.17
Time above range (%)	64.1901 (17.409)	—	53.5695 (16.3889)	—	54.5215 (14.4949)	—	50.9759 (13.8959)	—	48.8406 (13.6259)	—	<.001	<.001	.06
Mean glucose (mg/dL)	238.9856 (53.6842)	—	—	193.5024 (181.2758 - 232.7176)	209.5659 (40.2107)	—	200.1775 (34.7531)	—	194.077 (32.4736)	—	<.001	<.001	.05
Glucose management indicator	9.0265 (1.2841)	—	—	7.9386 (7.6461 - 8.8766)	8.3228 (0.96184)		8.0982 (0.83129)	—	7.9523 (0.77677)	—	<.001	<.001	.05
Glucose risk index	—	99.5992 (76.2825 - 100)	75.5983 (19.3713)	—	78.7051 (17.9064)	—	—	81.6141 (57.678 - 90.9717)	73.6757 (19.2475)	—	<.001	<.001	.21
Coefficient of variation	41.107 (10.4761)	—	42.7715 (7.7309)	—	43.8174 (7.1405)	—	44.2991 (6.5068)	—	44.6247 (6.6217)	—	.07	.15	.58
SD (mg/dL)	94.2823 (19.0105)	—	87.6943 (17.0699)	—	90.4366 (15.6548)	—	88.2446 (17.5455)	—	86.2282 (17.1804)	—	<.001	.29	.23
Hypoglycemic events per week (n)	—	4.4336 (0.70961 - 9.1829)	—	4.6668 (1.9445 - 7.2531)	—	4.978 (2.3529 - 7.7781)	—	4.9002 (2.6445 - 9.1781)	6.4021 (4.3053)	—	.14	.50	.38
Hyperglycemic events per week (n)	16.2707 (4.8316)	—	16.5623 (4.4013)	—	16.5226 (4.8881)	—	17.6488 (4.882)	—	16.7118 (5.7817)	—	.70	.80	.87

[a]Not applicable.

XSL•FO
RenderX

Figure 1. Box plot distributions of time in tight range were obtained at M0 (gray), M3 (red), M6 (light blue), M9 (green), and M12 (violet) from the Continuous Glucose Monitoring Among Patients with Type 1 Diabetes in Rwanda study, a single-arm, prospective observational study conducted at the Rwandan Diabetes Association clinic in Kigali, Rwanda (2022 - 2023). Median values are highlighted and connected by a black solid line.

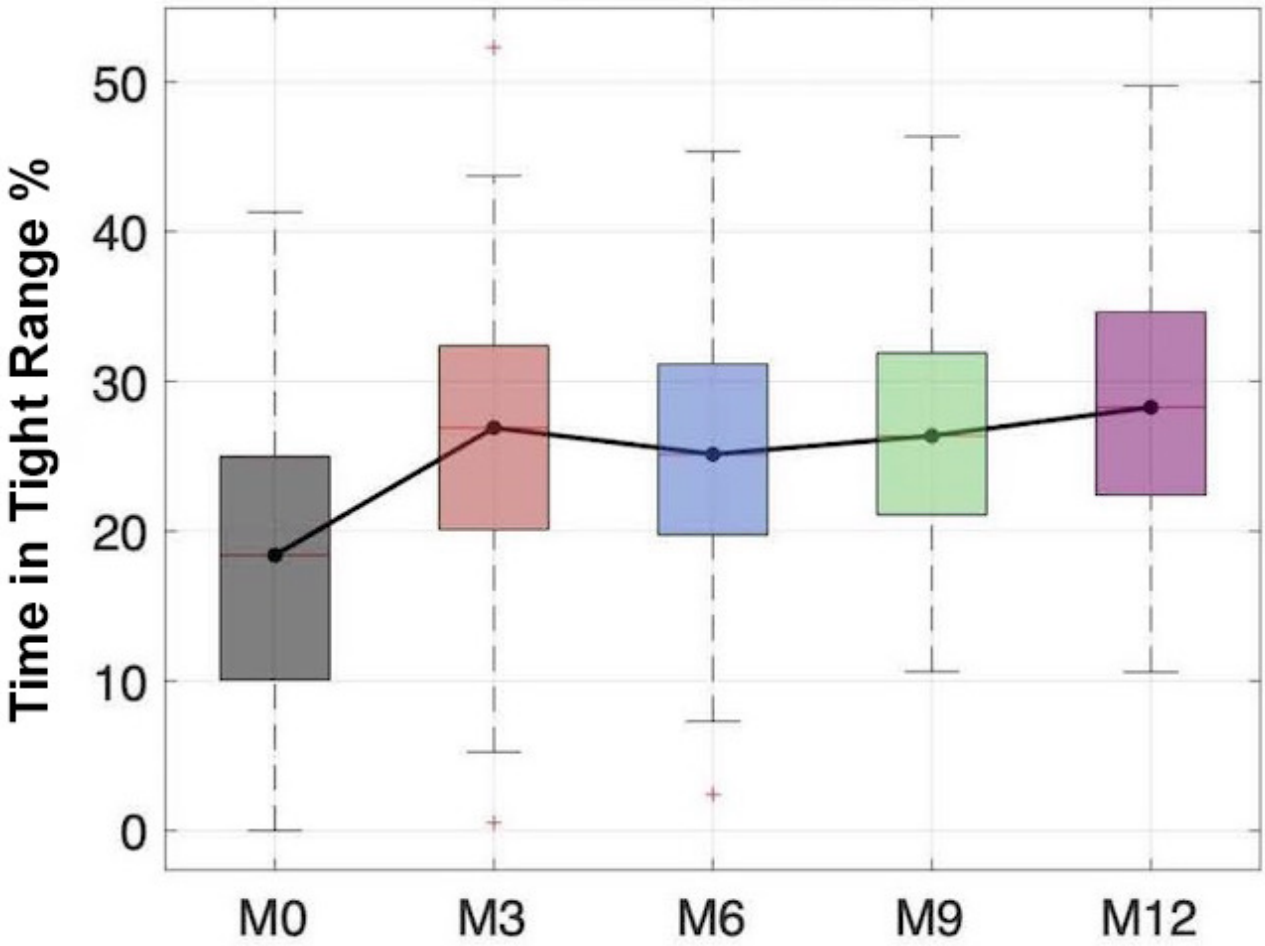

XSL•FO
RenderX

Figure 2. Box plot distributions of HbA_{1c} were obtained at M0 (gray), M3 (red), M6 (light blue), M9 (green), and M12 (violet) from the Continuous Glucose Monitoring Among Patients with Type 1 Diabetes in Rwanda study, a single-arm, prospective observational study conducted at the Rwandan Diabetes Association clinic in Kigali, Rwanda (2022 - 2023). Median values are highlighted and connected by a black solid line. HbA_{1c}: hemoglobin A_{1c}.

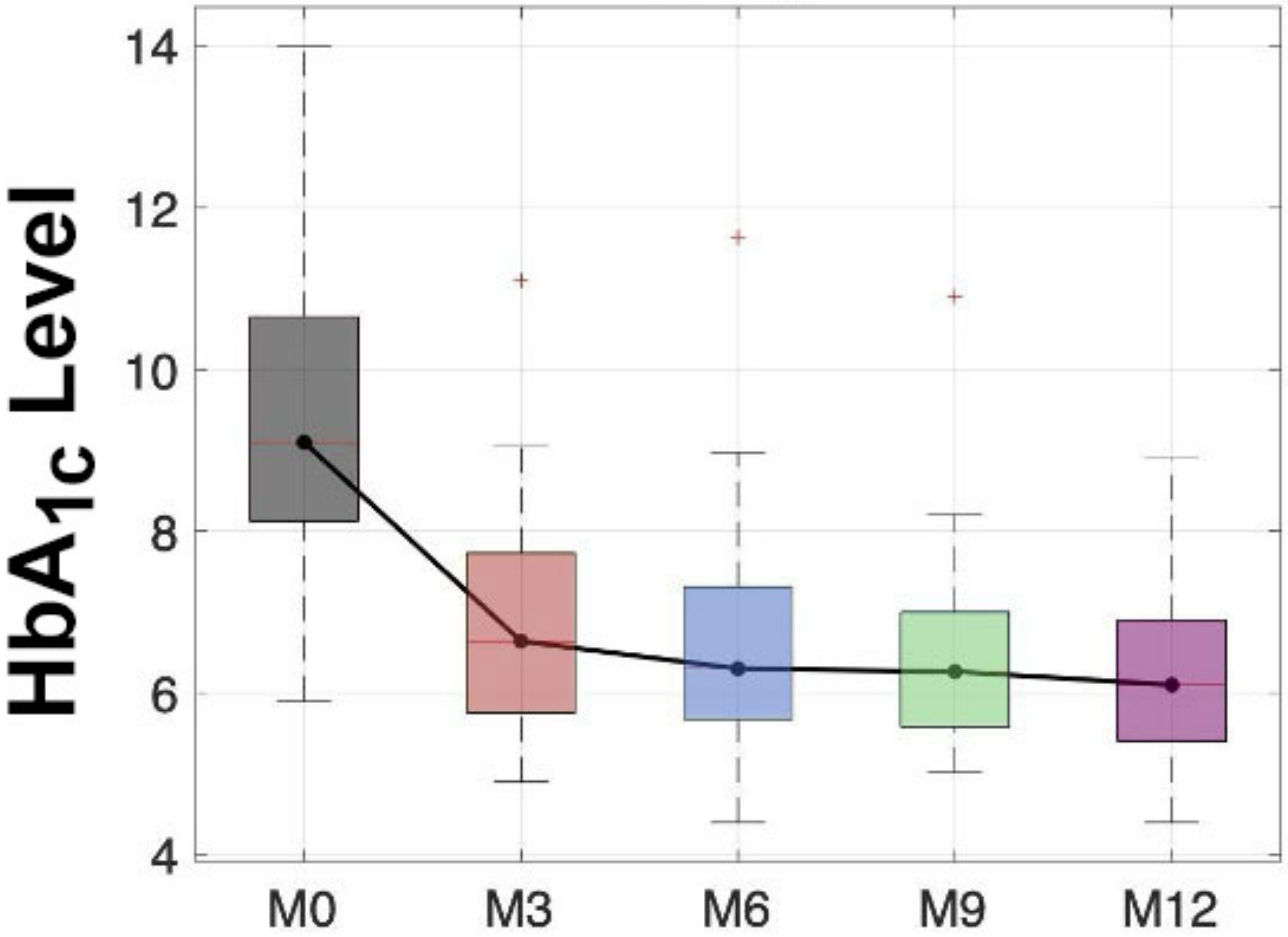

Discussion

The findings from Phase I of the CGM use among individuals with T1D in Rwanda provide insight into the impact and feasibility of real-time CGM use. As noted previously, the key objectives of this study included evaluating sensor use; TIR; HbA_{1c} levels; and rates of hypoglycemia, DRH, and DKA. Study participants demonstrated high levels of sensor use (80% of the study period). As noted in the results, TIR improved with increased levels at 6 and 12 months. HbA_{1c} levels also improved with marked decreases at 6 and 12 months. Of note, there was a mild, nonstatistically significant increase in TBR, which is indicative of hypoglycemia, but there was also a statistically significant reduction in TAR at 6 and 12 months. There was no report of severe hypoglycemia or DKA events, and only 12 DRHs were reported during the study. In sum, these data indicate a low level of adverse effects.

The low attrition rate, the high percentage of sensor use, and the willingness of participants to return to the clinic every 20 days for data uploads indicated the feasibility of continuous glucose monitoring in Rwanda among patients with T1D. The findings from this study augment those from other studies conducted in African countries showing that continuous glucose monitoring is a feasible alternative to SMBG [15,16].

In a study aiming to assess the feasibility of CGM use in Malawi, one of the challenges identified was the inability of participants to change their own sensors [16]. However, in this study, participants were given sufficient training so that they could change their sensors on their own. Additionally, participants visited the clinic frequently per study protocol, providing ample opportunity for study staff to help provide guidance on sensor removal and insertion.

A systematic review of 22 randomized controlled trials comparing continuous glucose monitoring and SMBG found a significant reduction in HbA_{1c} with the CGM group [17]. Although this study was not interventional, the observed results showed a similar significant reduction in HbA_{1c} with the use of a CGM.

Further, the findings from this study are aligned with other studies showing a reduction in HbA_{1c} with the use of a CGM [10,18] and consistent with prior studies showing that the use of a CGM increased TIR, a desirable outcome for people with diabetes[18].

XSL•FO
RenderX

There were several limitations of this study that limited validity and generalizability. The single-arm observational design lacks a control group, which is consistent with a feasibility study but nonetheless limits between-group comparisons, the ability to isolate the effect of CGM use, the ability to conduct complete comparison analyses, and the ability to generalize findings.

While we were able to glean insight from this convenience sample, the sample was small at the outset, and the further reduction due to attrition impacts statistical power and effect size. Our use of the last observation carried forward method for missing data could potentially lead to bias.

As with any study that uses self-reported information, bias can be introduced and accuracy can be compromised. Notwithstanding these limitations, the length of this study, conducted over a 1-year period, provided a meaningful amount of data points for analysis. These preliminary results can serve as a foundation for future studies that can include control groups and interventions. This work is particularly relevant given the dearth of studies to date related to CGM use in people with T1D in Rwanda.

The results suggest that implementing CGM technology in T1D patients in Rwanda is a feasible and promising method for improving glycemic outcomes among people with T1D. Within 3 months, results indicate improvements in glycemic management, with the improvements sustained over the 1-year study period, as evidenced by better TIR, TITR, and TAR and lower HbA_{1c} levels. These outcomes indicate potential for better diabetes self-management, thereby impacting rates of diabetes-related complications and their associated personal and social costs. Results from this study can be used to conduct more rigorous interventional studies to confirm the beneficial impact of a CGM on health outcomes for people with diabetes.

Data Availability

Data are available upon reasonable request to the corresponding author.

Conflicts of Interest

SM and SN are employees of Dexcom and own company shares. JB is a speaker for Dexcom and a consultant for Sanofi (as a speaker) and Insulet (on the advisory board).

References

1. IDF Diabetes Atlas. International Diabetes Federation. URL: https://diabetesatlas.org/data/en/ [accessed 2024-07-13]
2. IDF Members: Rwanda. International Diabetes Federation. URL: https://idf.org/our-network/regions-and-members/africa/members/rwanda/ [accessed 2024-07-13]
3. American Diabetes Association Professional Practice Committee. 7. Diabetes technology: standards of care in diabetes-2024. Diabetes Care 2024 Jan 1;47(Suppl 1):S126-S144. [doi: 10.2337/dc24-S007] [Medline: 38078575]
4. Miller KM, Beck RW, Bergenstal RM, et al. Evidence of a strong association between frequency of self-monitoring of blood glucose and hemoglobin A1c levels in T1D exchange clinic registry participants. Diabetes Care 2013 Jul;36(7):2009-2014. [doi: 10.2337/dc12-1770] [Medline: 23378621]
5. Patton SR. Adherence to glycemic monitoring in diabetes. J Diabetes Sci Technol 2015 May;9(3):668-675. [doi: 10.1177/1932296814567709] [Medline: 25591853]
6. Didyuk O, Econom N, Guardia A, Livingston K, Klueh U. Continuous glucose monitoring devices: past, present, and future focus on the history and evolution of technological innovation. J Diabetes Sci Technol 2021 May;15(3):676-683. [doi: 10.1177/1932296819899394] [Medline: 31931614]
7. Cappon G, Vettoretti M, Sparacino G, Facchinetti A. Continuous glucose monitoring sensors for diabetes management: a review of technologies and applications. Diabetes Metab J 2019 Aug;43(4):383-397. [doi: 10.4093/dmj.2019.0121] [Medline: 31441246]
8. Dovc K, Battelino T. Time in range centered diabetes care. Clin Pediatr Endocrinol 2021;30(1):1-10. [doi: 10.1297/cpe.30.1] [Medline: 33446946]
9. Rodbard D. Continuous glucose monitoring: a review of recent studies demonstrating improved glycemic outcomes. Diabetes Technol Ther 2017 Jun;19(S3):S25-S37. [doi: 10.1089/dia.2017.0035] [Medline: 28585879]
10. Nemlekar PM, Hannah KL, Green CR, Norman GJ. Association between adherence, A1C improvement, and type of continuous glucose monitoring system in people with type 1 diabetes or type 2 diabetes treated with intensive insulin therapy. Diabetes Ther 2024 Mar;15(3):639-648. [doi: 10.1007/s13300-023-01529-8] [Medline: 38289464]
11. Karter AJ, Parker MM, Moffet HH, Gilliam LK, Dlott R. Association of real-time continuous glucose monitoring with glycemic control and acute metabolic events among patients with insulin-treated diabetes. JAMA 2021 Jun 8;325(22):2273-2284. [doi: 10.1001/jama.2021.6530] [Medline: 34077502]
12. Champakanath A, Akturk HK, Alonso GT, Snell-Bergeon JK, Shah VN. Continuous glucose monitoring initiation within first year of type 1 diabetes diagnosis is associated with improved glycemic outcomes: 7-year follow-up study. Diabetes Care 2022 Mar 1;45(3):750-753. [doi: 10.2337/dc21-2004] [Medline: 35018417]
13. Cappon G, Sparacino G, Facchinetti A. AGATA: a toolbox for automated glucose data analysis. J Diabetes Sci Technol 2024 Sep;18(5):1109-1121. [doi: 10.1177/19322968221147570] [Medline: 36602030]

XSL•FO
RenderX

14. Battelino T, Alexander CM, Amiel SA, et al. Continuous glucose monitoring and metrics for clinical trials: an international consensus statement. Lancet Diabetes Endocrinol 2023;11(1):42-57. [doi: 10.1016/S2213-8587(22)00319-9]
15. McClure Yauch L, Velazquez E, Piloya-Were T, Wainaina Mungai L, Omar A, Moran A. Continuous glucose monitoring assessment of metabolic control in east African children and young adults with type 1 diabetes: a pilot and feasibility study. Endocrinol Diabetes Metab 2020 Jul;3(3):e00135. [doi: 10.1002/edm2.135] [Medline: 32704558]
16. Gomber A, Valeta F, Coates MM, et al. Feasibility of continuous glucose monitoring in patients with type 1 diabetes at two district hospitals in Neno, Malawi: a randomised controlled trial. BMJ Open 2024 May 6;14(5):e075554. [doi: 10.1136/bmjopen-2023-075554] [Medline: 38719319]
17. Teo E, Hassan N, Tam W, Koh S. Effectiveness of continuous glucose monitoring in maintaining glycaemic control among people with type 1 diabetes mellitus: a systematic review of randomised controlled trials and meta-analysis. Diabetologia 2022 Apr;65(4):604-619. [doi: 10.1007/s00125-021-05648-4] [Medline: 35141761]
18. Beck RW, Riddlesworth T, Ruedy K, et al. Effect of continuous glucose monitoring on glycemic control in adults with type 1 diabetes using insulin injections: the DIAMOND randomized clinical trial. JAMA 2017 Jan 24;317(4):371-378. [doi: 10.1001/jama.2016.19975] [Medline: 28118453]

Abbreviations

CAPT1D: Continuous Glucose Monitoring Among Patients with Type 1 Diabetes in Rwanda
CGM: continuous glucose monitoring system
DKA: diabetic ketoacidosis
DRH: diabetes-related hospitalization
HbA_{1c}: hemoglobin A_{1c}
RDA: Rwandan Diabetes Association
SMBG: self-monitoring blood glucose
T1D: type 1 diabetes
TAR: time above range
TBR: time below range
TIR: time in range
TITR: time in tight range

Edited by A Mavragani; submitted 21.07.24; peer-reviewed by I Milani, JA Aloi; revised version received 21.11.24; accepted 25.11.24; published 21.01.25.

Please cite as:
Baker J, Cappon G, Habineza JC, Basch CH, Mey S, Malkin-Washeim DL, Schuetz C, Simon Pierre N, Uwingabire E, Mukamazimpaka A, Mbonyi P, Narayanan S
Continuous Glucose Monitoring Among Patients With Type 1 Diabetes in Rwanda (CAPT1D) Phase I: Prospective Observational Feasibility Study
JMIR Form Res 2025;9:e64585
URL: https://formative.jmir.org/2025/1/e64585
doi:10.2196/64585

© Jason Baker, Giacomo Cappon, Jean Claude Habineza, Corey H Basch, Steven Mey, Diana L Malkin-Washeim, Christian Schuetz, Niyonsenga Simon Pierre, Etienne Uwingabire, Alvera Mukamazimpaka, Paul Mbonyi, Sandhya Narayanan. Originally published in JMIR Formative Research (https://formative.jmir.org), 21.1.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Developing Assessments for Key Stakeholders in Pediatric Congenital Heart Disease: Qualitative Pilot Study to Inform Designing of a Medical Education Toy

Neda Barbazi[1], MArch; Ji Youn Shin[1], PhD; Gurumurthy Hiremath[2], MD; Carlye Anne Lauff[1], PhD

[1]Department of Design Innovation, College of Design, University of Minnesota, Twin Cities, Minneapolis, MN, United States

[2]Division of Cardiology, Department of Pediatrics, Medical School, University of Minnesota, Twin Cities, Minneapolis, MN, United States

Corresponding Author:
Carlye Anne Lauff, PhD
Department of Design Innovation, College of Design
University of Minnesota, Twin Cities
13 Rapson Hall, 89 Church Street SE
Minneapolis, MN, 55455-0213
United States
Phone: 1 724 889 5511
Fax: 1 612 625 7525
Email: carlye@umn.edu

Abstract

Background: Congenital heart disease (CHD) is a birth defect of the heart that requires long-term care and often leads to additional health complications. Effective educational strategies are essential for improving health literacy and care outcomes. Despite affecting around 40,000 children annually in the United States, there is a gap in understanding children's health literacy, parental educational burdens, and the efficiency of health care providers in delivering education.

Objective: This qualitative pilot study aims to develop tailored assessment tools to evaluate educational needs and burdens among children with CHD, their parents, and health care providers. These assessments will inform the design of medical education toys to enhance health management and outcomes for pediatric patients with CHD and key stakeholders.

Methods: Through stakeholder feedback from pediatric patients with CHD, parents, and health care providers, we developed three tailored assessments in two phases: (1) iterative development of the assessment tools and (2) pilot testing. In the first phase, we defined key concepts, conducted a literature review, and created initial drafts of the assessments. During the pilot-testing phase, 12 participants were recruited at the M Health Fairview Pediatric Specialty Clinic for Cardiology—Explorer in Minneapolis, Minnesota, United States. We gathered feedback using qualitative methods, including cognitive interviews such as think-aloud techniques, verbal probing, and observations of nonverbal cues. The data were analyzed to identify the strengths and weaknesses of each assessment item and areas for improvement.

Results: The 12 participants included children with CHD (n=5), parents (n=4), and health care providers (n=3). The results showed the feasibility and effectiveness of the tailored assessments. Participants showed high levels of engagement and found the assessment items relevant to their education needs. Iterative revisions based on participant feedback improved the assessments' clarity, relevance, and engagement for all stakeholders, including children with CHD.

Conclusions: This pilot study emphasizes the importance of iterative assessment development, focusing on multistakeholder engagement. The insights gained from the development process will guide the creation of tailored assessments and inform the development of child-led educational interventions for pediatric populations with CHD.

(JMIR Form Res 2025;9:e63818) doi:10.2196/63818

KEYWORDS

assessment; congenital heart disease; children health literacy; health education; caregiving; patient-centered care; design; qualitative pilot; children health; educational interventions

XSL•FO
RenderX

Introduction

Background

Congenital heart disease (CHD) is a heart abnormality present at birth, requiring intensive medical care and often leading to life-threatening complications [1]. It affects around 40,000 newborns annually in the United States and 1% globally, posing lifelong challenges for children, families, and health care systems [1-4]. Despite significant advancements in diagnosis and treatment, CHD remains a chronic condition that demands ongoing care and specialized educational resources [5,6]. Tailored education is critical for addressing the needs of children with CHD, providing accessible educational support for parents, and equipping health care providers with efficient educational strategies. However, no tools currently exist to understand children's health literacy, assess parental educational burdens, or measure health care providers' efficiency in delivering education [7,8].

Children with CHD face unique barriers in understanding their condition due to developmental challenges, such as difficulty grasping abstract medical concepts or relating them to their lived experiences. These limitations can hinder their ability to adhere to treatment plans, actively participate in care, or respond effectively during emergencies and transitions to adult care [4,6,9-11]. Despite these challenges, pediatric patients are often excluded from health literacy efforts, as most resources are designed for parents or caregivers. This exclusion highlights the critical need for age-appropriate tools that empower children to engage with their care actively, ultimately improving adherence and long-term health outcomes [7,12-16].

Parents, meanwhile, bear a significant educational burden. They must interpret medical jargon, simplify it for their child, and act as intermediaries with health care providers, all while managing the emotional strain of caregiving and the cognitive load of understanding complex medical information. This burden can increase stress, reduce caregiving effectiveness, and impact family well-being. Low parental health literacy further compounds these challenges, as it is linked to medical errors and poorer health outcomes for children [17-24]. Furthermore, health care providers face the challenge of balancing clear communication with time and resource constraints, often struggling to deliver education tailored to the needs of both children and parents [11,25-27]. Addressing these gaps is essential for improving communication, care coordination, and health outcomes for children with CHD and their families [7,28].

Study Objectives

As part of a multiphase research project, this pilot study aims to close these gaps by developing tailored assessment tools for key stakeholders in pediatric CHD education. This research comprises two phases: (1) iterative development of the assessment tools and (2) pilot testing to create, refine, and validate 3 CHD-specific assessments based on stakeholder feedback in a real-world setting. We gathered feedback using cognitive interviews and observations to assess each assessment item [29-34]. We iteratively developed face-to-face and computer-based assessments based on this feedback from all stakeholders, including children. The assessments measure changes in children's knowledge, parental educational burdens, and health care provider efficiency before and post interventions. The results show that developing and refining these assessments are essential before introducing our medical education toy. The objectives of this study are 3-fold:

1. To assess changes in knowledge of children with CHD before and after educational interventions to empower pediatric engagement in health care innovation.
2. To measure the educational burden experienced by parents before and after interventions, focusing on their needs and challenges in navigating CHD-related educational responsibilities.
3. To assess the efficiency and effectiveness of health care providers in communicating essential information to children with CHD and their families to improve care coordination and patient outcomes.

Prior Work

Assessments Developed for Pediatric Patients With CHD, Parents, and Health Care Providers

Assessments are essential for evaluating the health status and practices of pediatric patients, their parents, and health care providers across various domains [35-37]. These domains include health-related quality of life, emotional well-being, physical or psychological health, social support, and behavioral problems. For pediatric patients aged 8-18 years, assessments often involve parent proxies completing tools for younger children [38-43]. Commonly used tools in this context include the Pediatric Quality of Life Inventory (PedsQL), Pediatric Cardiac Quality of Life Inventory (PCQLI) [44-47], Child Health Questionnaire (CHQ), and Child Behavior Checklist (CBCL). However, there is a notable gap in health literacy measures for children younger than 9 years [48]. While some tools, such as Food Label Literacy for Applied Nutrition Knowledge (FLLANK) questionnaire [49], exist, none are specifically tailored to disease populations such as those with CHD.

Parental well-being and caregiving burden assessments are also critical for understanding and supporting effective caregiving. However, the specific burden related to education remains underexplored. Existing tools, such as the Parenting Stress Index/Parental Stress Scale (PSI/PSS) and the Family Impact Module (FIM) of the PedsQL, focus on parental stress, caregiving difficulty, and family functioning [50-53]. Similarly, tools such as the Caregiver Health Self-Assessment, which aims to improve the caregiver-provider dyadic relationship (commonly for older adult patients), do not address the educational needs of caregivers of patients with CHD [54-56].

For health care providers, efficiency assessments—measuring the ability to minimize wasted time and maximize outcomes—remain an ongoing challenge despite progress in evaluating physician and hospital care effectiveness [57,58]. While tools such as Key Performance Indicators (KPIs), Lean Six Sigma (LSS), and Performance Metrics are widely used to evaluate provider and hospital performance, their primary focus is on patient care outcomes (eg, patient satisfaction). These methods often overlook the educational challenges faced by

XSL•FO
RenderX

health care providers themselves [59-63]. Addressing these gaps can enhance our understanding of the needs and well-being of CHD stakeholders, including children, enabling the development of more targeted and effective educational interventions. More details about these assessments are available in Multimedia Appendix 1.

Empowering Pediatric Engagement in Health Care Interventions

Engaging children effectively in health care interventions and assessments is challenging due to their unique developmental hurdles [64-67]. Younger children face difficulties participating because of limited attention spans and cognitive abilities, requiring reliance on parental feedback as a proxy for designing and testing interventions [68,69]. To address these challenges, studies suggest using interviews, focus groups, and activity-based methods [70-74].

Prior studies in design, human-computer interaction, and health care domains indicate that adolescents aged 13-17 years engage well in interviews and focus groups. In contrast, younger children, particularly those aged 4-12 years, benefit more from creative techniques such as activity-based methods [72-78]. Furthermore, studies highlight the importance of social factors, such as ongoing support, in alleviating children's potential stress during research activities [79]. Empowering children through age-appropriate strategies, coupled with family facilitation, enhances their sense of ownership in health care decision-making. These approaches not only improve engagement but also strengthen the research process by addressing children's developmental needs and promoting their active participation [79-83].

Methods

We developed and refined 3 CHD-specific assessments to evaluate children's knowledge, parental educational burdens, and health care provider efficiency in delivering education. This process included defining key concepts, conducting a literature review, and designing initial assessment tools. These assessments were subsequently pilot-tested with stakeholders at the M Health Fairview Pediatric Specialty Clinic for Cardiology—Explorer in Minneapolis, Minnesota, United States, to ensure validity and relevance.

Assessment Development

CHD Health Literacy Children Assessment

We developed the CHD Health Literacy Children Assessment (CHD-HLCA) to evaluate CHD health literacy in children aged 4-10 years through pre- and posteducational intervention. This tool draws inspiration from the FLLANK questionnaire [84] and incorporates storytelling techniques to enhance engagement [85]. The assessment consists of 10 questions featuring simple black and white icons without color or intricate details to reduce visual distractions and avoid potential psychological and physiological influences of color. Instead, children use colored markers to answer, fostering their engagement and enthusiasm during the assessment process [86,87].

This self-report assessment is administered with the assistance of the research team or parents, particularly for younger children. The assessment covers various knowledge dimensions, including understanding CHD, preparing for doctor visits, and practicing self-care [49]. Questions offer 2 comparison options to minimize complexity, using simple icons and illustrations to reduce distractions and prevent cognitive overload. A "can't tell" option is also included to ensure that children feel comfortable expressing uncertainty. Furthermore, a Likert scale question with Smiley Face Likerts assesses their general knowledge about their heart [88], a common method for children's surveys. The assessment is conducted face-to-face with parental involvement, fostering a supportive environment where children can freely express themselves through drawing, crafting, or pointing to icons. It takes approximately 5-10 minutes per child and evaluates 4 key domains: Conceptual Knowledge, Comprehension, Appraisal, and Application/function specific to CHD (Figure 1 [84]).

XSL•FO
RenderX

Figure 1. Final children's assessment, inspired by the Food Label Literacy for Applied Nutrition Knowledge (FLLANK) questionnaire [84].

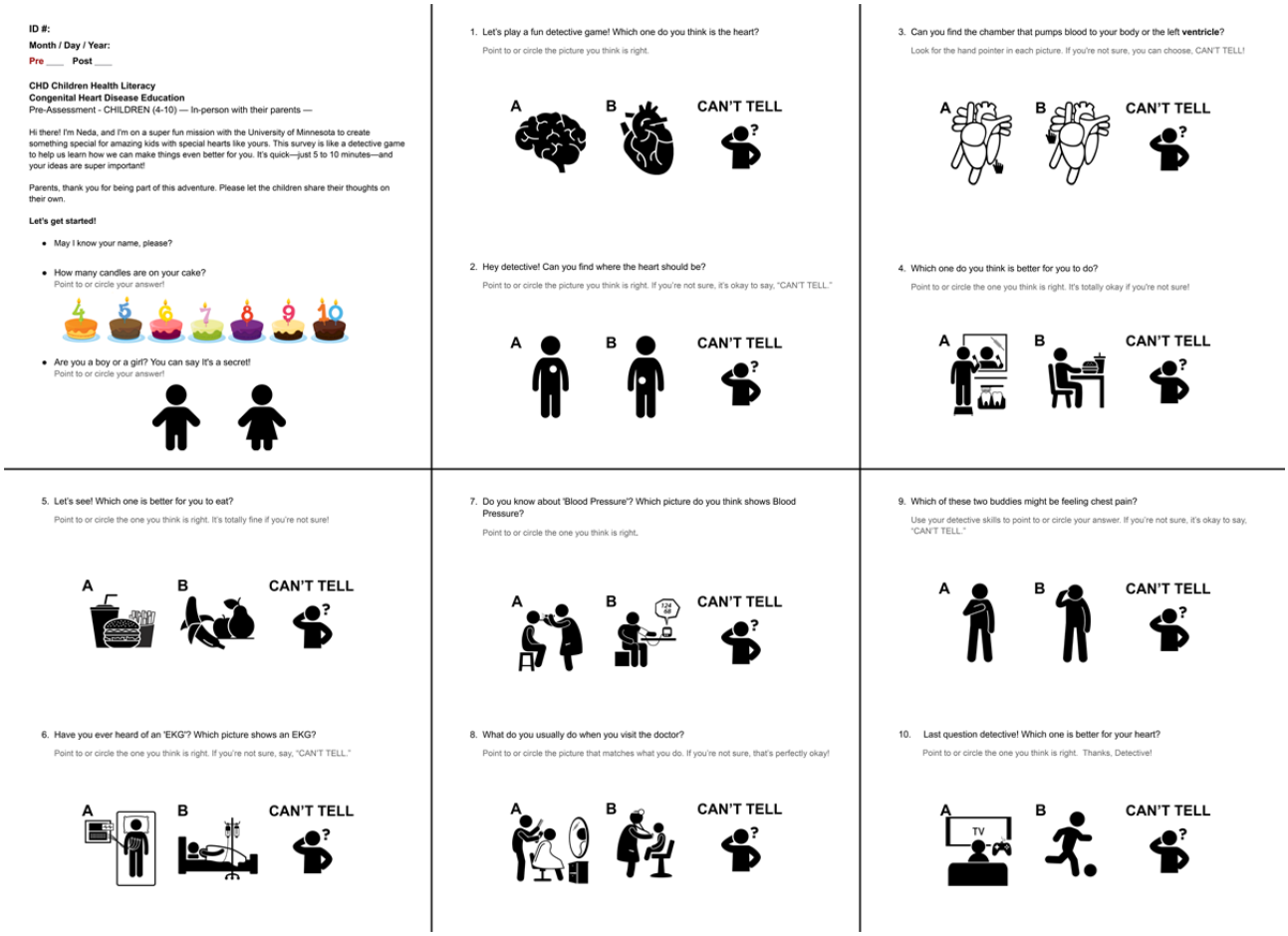

CHD Parental Educational Burden Assessment

The CHD Parental Educational Burden Assessment (CHD-PEBA) is a 30-question assessment designed to evaluate parental educational responsibilities. It uses Likert scales, multiple-choice questions, matrix format, and open-ended responses to gather both quantitative and qualitative data on parental challenges. Inspired by Neuro-QoL and the Caregiver Health Self-Assessment Questionnaire [54,89,90], this self-report assessment examines parental understanding, coping mechanisms, and support needs in educating children about CHD. The assessment evaluates parents' knowledge of CHD, their perception of their children's understanding of the condition, and their confidence and stress levels in providing educational support. It also explores preferred information sources, information-seeking behaviors, and the time and effort parents dedicate to educating their children. Demographic data are collected to contextualize responses. Administered on the web, the survey takes approximately 5-10 minutes to complete. It serves as both a pre-and postintervention tool, measuring changes in parental educational burden over time.

CHD Healthcare Provider Educational Efficiency Assessment

The self-reported CHD Healthcare Provider Educational Efficiency Assessment (CHD-HEEA) health care providers' assessment is a 20-question self-report assessment to evaluate health care providers' educational efficiency in educating children with CHD and their families. It assesses the practices of pediatric cardiologists, fellows, nurses, and child life specialists through various question formats, including Likert scales, matrix questions, and multiple-choice and open-ended responses. Inspired by tools such as the Physician Task Checklist and Lean Six Sigma [60,63,91], the assessment examines educational practices, challenges, and strategies. Key areas of evaluation include the time and effort spent on educating children with CHD and families, perceptions of current educational methods, use of preappointment materials and self-education resources, encouragement of questions from caregivers and children, and strategies to streamline education while maintaining information accessibility. It also collects some demographic data to provide context for the responses. Administered on the web, the survey takes approximately 5-10 minutes to complete. It serves as a pre- and postassessment tool to evaluate changes in health care provider practices after educational interventions to optimize care coordination and patient education. Additional details about the 3 assessments are provided in Multimedia Appendix 2.

Pilot Testing of Assessment Tools

During the pilot phase, we tested 3 different assessments with CHD stakeholders at the M Health Fairview Pediatric Specialty Clinic for Cardiology—Explorer.

Study Population

The study defined specific criteria for each stakeholder group. Children with CHD were required to be between 4 and 10 years

XSL•FO
RenderX

of age, diagnosed by a health care provider, and fluent in English. Parents or guardians had to speak English and be confirmed as caregivers by a health care provider. Health care providers must be actively involved in CHD care and fluent in English.

Participants in this pilot study included children with CHD (n=5), their parents (n=4), and health care providers (n=3). Among the children were four 7-year-olds (3 males and 1 female) and one 5-year-old female (n=5). The parental and caregiver group comprised 5 mothers aged 25-44 years, including 2 White/Caucasian and 3 African American/Black participants, all with bachelor's degrees and reporting middle to high-level incomes (n=5). However, one of the parents did not complete the assessment, resulting in a final parental group size (n=4). The health care provider group included 2 pediatric cardiology fellows and 1 physician assistant student, aged 25-44 years, with 1-5 years of experience in pediatric cardiology (n=3).

Study Recruitment and Informed Consent

The medical team facilitated communication with potential children with CHD and their families through purposive sampling. Parents signed parental permission forms for their children and children provided assent. We administered the children's assessment while their parents were present but ensured that the parents did not guide their answers. After providing their own informed consent, parents received a link to complete the computer-based parental assessment. They could complete it at the clinic using the researcher's laptop or later at home. Parents shared their experiences and provided feedback on the assessment, including its clarity and length, either during or after completion. Health care providers also participated by completing their assessments and providing feedback on their experiences. All participants received a gift card for their time. We transcribed participants' input or feedback and anonymized results and transcripts to protect privacy.

Data Collection

We used qualitative methods, including cognitive interviews, observational notes, and interactive techniques, to gather feedback from children, parents, and health care providers [29,30,32,34]. These methods captured participants' impressions, behaviors, and engagement, providing rich data for iterative refinements of the assessments.

Logistics and Setup

To accommodate the clinic setting and time constraints, we conducted parent and child assessments simultaneously. Parents began slightly later than their children to ensure that the children were comfortable and understood the process before starting. This approach minimized parental influence on children's responses while saving time. After completing their assessments, children engaged in painting activities, allowing parents to complete their assessments without distractions. For health care providers, assessments were scheduled flexibly, either on the web or in person, to accommodate their busy schedules.

Cognitive and Observational Methods

Cognitive interviews and verbal probing revealed how participants interpreted and responded to survey questions. As a qualitative method widely used in survey design, cognitive interviews helped identify ambiguities, cognitive challenges, and difficulties in understanding. Participants verbalized their thoughts concurrently (using the think-aloud technique) or retrospectively (recalling and explaining their thought processes after completing the assessment) [29-34]. Observational techniques captured nonverbal cues, such as hesitation, frustration, or excitement, along with body language and engagement levels. These observations were particularly valuable for younger children, who often struggled with abstract concepts or articulating their thoughts [76]. Combining these methods provided detailed feedback that informed refinements to our assessments.

Engaging Children

We used developmentally appropriate and interactive methods to engage children effectively. A researcher read questions aloud to younger children, with parents assisting if the child felt uncomfortable interacting with the research team. To make the activity engaging, we incorporated storytelling and asked children to role-play as detectives solving questions. Prompts such as "Hey, detective! Let's figure out which, what, or where!" encouraged participation. Children used markers and craft materials, such as multicolor pom-poms, to point to their answers, making the process interactive and enjoyable. For older children who could read independently, we encouraged self-guided responses and asked clarifying questions to explore their thought processes. The think-aloud technique and verbal probing provided deeper insights into their reasoning [70-74]. To ensure that children understood the questions rather than guessing, we rephrased questions after they provided answers. For example, if a child pointed to an icon for chest pain, we asked, "Do you know where your chest is?" and used physical prompts to confirm their understanding. This approach validated their answers and often prompted children to share personal experiences, enriching the feedback [66,67,76,83].

Refining the Assessments

We iteratively refined the assessments, tailoring questions to each stakeholder group by focusing on their thoughts, understanding, relevance, completeness, survey length, and overall experience.

For the children's assessment, we prioritized meeting the developmental needs of children aged 4-10 years based on criteria provided by health care providers. These criteria outlined essential knowledge for this age group while addressing the additional challenges faced by younger or newly diagnosed children. Testing revealed that younger children often struggled with multistep questions, so we simplified these into single, clear actions. Language and visuals were adjusted for clarity while retaining enough detail to engage older children. Older children provided feedback on question relevance and reflected on how they might have responded when younger, offering insights that shaped refinements.

Each testing round informed adjustments to ensure that the questions were clear, precise, and engaging for all stakeholders in CHD pediatric care. These assessments now support pre- and postassessment stages to validate interventions, including our

XSL•FO
RenderX

ongoing study of medical educational toys for CHD pediatric populations. Table 1 outlines the development and testing process, with sample prompts provided in Multimedia Appendix 3.

Table 1. Iterative development and pilot testing of congenital heart disease (CHD) assessments.

Step	Description	Stakeholders involved	Outcomes
1. Define key concepts	Define health literacy, parental educational burden, and health care provider efficiency	Research team, pediatric cardiologist	Established key concepts for the assessment framework
2. Conduct literature review	Review existing best practices and assessments	Research team	Insights for developing new assessments
3. Develop initial assessments	Draft assessments for children, parents, and health care providers	Research team, pediatric cardiologist	Initial assessments ready for review
4. Pilot testing—first iteration	Test initial assessments with children with CHD, parents, and health care providers	Children with CHD, parents, and health care providers	Feedback on relevance, clarity, and appropriateness
5. Analyze feedback	Analyze feedback from pilot testing	Research team, pediatric cardiologist	Identify strengths, weaknesses, and improvements
6. Refine assessments	Modify assessments based on feedback	Research team	Improved assessments
7. Pilot testing—further iterations	Conduct additional testing and refinement	Children with CHD, parents, and health care providers	Continuous improvement and validation
8. Final analysis	Analyze all data and feedback to finalize assessments	Research team, pediatric cardiologist	Validated assessments for implementation

Data Analysis

Our research team held weekly meetings to analyze assessment data and stakeholder feedback. Using qualitative methods, including thematic analysis, we identified common themes from feedback and observational data [92,93]. The first author coded the data to highlight relevant themes, focusing on the assessment's reliability, engagement, and responsiveness to meet all stakeholders' needs. During each round of pilot testing, we prioritized feedback using four criteria: (1) its potential impact on the assessment's effectiveness, (2) alignment with intended outcomes, (3) feasibility of incorporating changes, and (4) stakeholder preferences. Feedback was categorized by feasibility, clarity, relevance to educational needs, and participant engagement, then ranked by the frequency and significance of reported issues. We implemented revisions iteratively, prioritizing aspects that most improved comprehension and ease of response. This systematic process enhanced the assessments' accuracy and relevance across stakeholder groups. During weekly meetings, the team reviewed how each piece of feedback aligned with the established criteria and worked collaboratively to resolve discrepancies. This iterative process led to the development of a set of codes such as "strengths," "weaknesses," and "areas of improvement." These ongoing discussions and thematic analyses refined the assessments and informed adjustments for subsequent pilot tests [94].

Ethical Considerations

The study received ethical approval from the University of Minnesota institutional review board (STUDY00020670). The medical team facilitated communication with potential participants—children with CHD and their families—through purposive sampling. Informed consent was obtained from all participants: parents also signed parental permission forms for their children, and children provided assent. Participants were informed about the study's purpose, procedures, and their right to withdraw at any time. Data were anonymized and deidentified during transcription and analysis, with all personal information securely stored in adherence to institutional guidelines. Participants received a gift card as compensation for their time, ensuring fairness and transparency.

Results

We identified consistent themes across three assessments: (1) assessment engagement, (2) relevance and structure of assessment, and (3) opportunities of assessments. These themes guided iterative revisions before each new pilot test. In this section, we summarize the findings from (1) children's assessments, (2) parents' assessments, and (3) health care providers' assessments. Some feedback from key stakeholders on 3 survey experiences can be found in Multimedia Appendix 2.

Assessments Engagement

Both children (C) and parents (P) actively engaged in the children's assessment process. Parents were surprised by their children's interest and enjoyment, with one parent expressing, "I wasn't sure she'd even talk. Wow!" It appeared that the interactive elements, such as choosing pens or markers and storytelling, not only boosted curiosity and made them concentrate on doing the assessment but also eased the anxiety and fear of children. For instance, C3 quickly transitioned from nervousness about another medical procedure to excitement upon seeing colorful markers, exclaiming, "Yay, I can paint here in the doctor's office!" Initially, she hid under the bed due to fear when we entered the examination room. However, she relaxed and became comfortable after seeing the markers and being invited to answer the questions using different colors. Even children like C2, initially focused on their iPad, became

XSL•FO
RenderX

engaged, asking, “Can I use all colors? I can answer anything like this!”

Parents appreciated the playful and instructive design of the assessments, with one commenting, “It’s like teaching by itself through playtime. She’s telling me all you've asked her.” After finishing their assessments, we provided paper to children like C4, whose mother was also being assessed. This engagement prompted them to draw, becoming so engrossed to continue that they requested us to stay there even after the doctor’s arrival (Figure 2). Initially, parental influence impacted C1’s responses, but gentle interventions and prioritizing children’s uninfluenced responses over accuracy fostered independent responses in other pilot tests. Through face-to-face interaction using cognitive interviews akin to semistructured interviews [95,96], we collected less biased data from children rather than their parents, maintained attention, and reduced parental influence.

Figure 2. Children’s interactions with children’s assessments and parental responses to their assessments at the Explorer clinic.

Similarly, in the clinic, parents provided feedback after their children’s assessments through a think-aloud and verbal probing method [29,32] while answering the computer-based parents’ assessment using Typeform (Typeform SL). During the assessment, P4 noted, “You know what to ask.” P3 mentioned that they would change only some of the wording. However, despite this feedback, P3 described the survey as “clear, easy, feeling good; finally, somebody asks!” This feedback highlighted the ease of use, clear directions, and meeting the participants’ needs, facilitating effective data collection. The health care providers’ assessment also received positive feedback for its engaging interface, with H1 noting, “Much better than usual surveys,” and H3 stating, “...really interactive with photos and video, and buttons. It’s easy....” Such feedback underscores the effectiveness of engaging participants and facilitating data collection.

Relevance and Structure of Assessment

The effectiveness of engaging all stakeholders through assessments depends on their relevance and well-structured design. Maintaining high relevance across all assessments, we noticed an interesting trend involving the Smiley Face Likert question during the children’s assessment. Despite our efforts to adjust the question wording, children consistently chose the happiest face in the initial version of the assessment. They emphasized their positive feelings associated with smiley faces by saying, “I chose it because I like it!” We expected this issue but still tested the smiley face Likert question due to its common use in children’s surveys. To improve engagement and reliability, we found that comparison questions, structured like a right or wrong game format, were more effective. Children showed high concentration levels and often asked, “Is this right?” They felt like participating in a game, increasing their involvement in the assessment process. Incorporating feedback from our medical team and parents highlighting the assessment’s educational value, we replaced the Likert question with a prompt offering a choice between physical activity and screen time (Figure 1 [84]), addressing another relevant habit for children with CHD. We also adjusted the question order to observe response variations, albeit with limited reliability due to its implementation with only participants C2 and C3 immediately after the initial version.

P2, P3, P4, and P5 actively engaged with the questions in the parents’ assessment, finding them clear, easy, and relevant. However, P2 noted assumptions in specific questions regarding the prior receipt of educational material during doctor visits. They proposed a preliminary question to confirm whether educational material was used before assessing its appropriateness (Figure 3). They asked, “How do you know if we got any educational material before asking how good it was? It’d make more sense to check if we received any and then ask about it.” Interestingly, when explicitly asked about this question, P3, a follow-up patient, did not express the same concern. Another participant suggested improving the context of the question about comfort levels during clinic visits (follow-ups or surgeries), which we addressed in the later version. We condensed the questionnaire to 25 items in the latest parent assessments by adding a matrix question.

Figure 3. Congenital Heart Disease Parental Educational Burden Assessment (CHD-PEBA) logic model question.

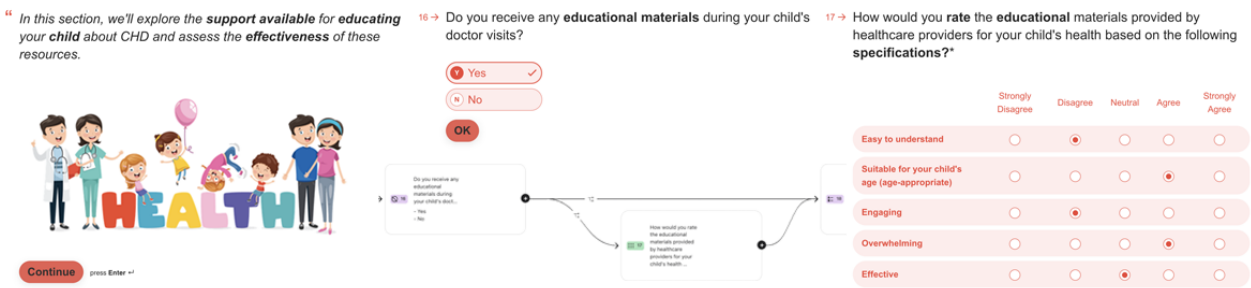

XSL•FO
RenderX

In response to initial feedback on the health care providers' assessment, H1 proposed categorizing questions differently for new and follow-up patients. All participants provided positive feedback regarding the matrix questions (Figure 4). H2 preferred to keep the existing matrix questions but suggested reordering them for better clarity and efficiency. They emphasized that combining related aspects into single questions helps respondents compare options and find the correct answer more easily: "Having one question for each of these aspects would make the questionnaire longer and harder to compare when answering; it just seems more practical this way." Furthermore, H2 recommended adding a specific demographic—vulnerable parents—to measure parental CHD information instead of using a general category. These suggestions were integrated into version 03, reducing the length to 15 items.

Figure 4. Congenital Heart Disease Healthcare Provider Educational Efficiency Assessment (CHD-HEEA) matrix questions.

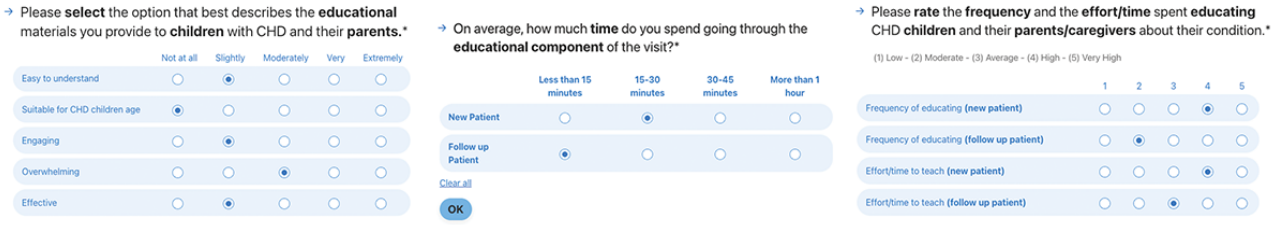

Opportunities of Assessments

We used patients' wait time between their initial screenings and the meeting with their physician to complete the assessments. This integration not only streamlined care coordination and reduced wait times but also engaged families and alleviated worries, as noted by P4, "It keeps us good busy!" This suggests that our design intervention (an educational toy) can be implemented during patients' wait time as well. The assessments also doubled as educational tools for children, engaging them in learning about their condition and easing anxiety. Parents found solace in sharing burdens, while health care providers gained insights into their practices and sought potential solutions. These assessments exceeded their initial purposes, offering education, support, and reflection opportunities for child-led approaches in pediatric care.

Discussion

Principal Results

During face-to-face interviews, we identified areas for refinement in assessment items, such as clarifying the context of health care experiences for children's comfort-level rating or specifying the type of visit for health care providers. Observational notes highlighted instances where assumptions in specific questions, such as the prior receipt of educational material during doctor visits, caused confusion among interviewers. To alleviate this issue, we proposed preliminary questions. Despite these challenges, participants generally interpreted the majority of survey items as intended, with occasional adjustments to wording choices for better comprehension. Furthermore, we changed the order of questions based on feedback and observations to improve the structure of the assessment.

The pilot testing evaluated the feasibility and effectiveness of the tailored assessments. Iterative revisions further improved their clarity and appropriateness. We ensured the validity of the assessments through a collaborative and iterative approach, incorporating perspectives from all stakeholders. As a result, the assessments offer a comprehensive understanding of educational needs and burdens, providing valuable insights to guide the development of targeted educational interventions.

Comparison With Prior Work

The tailored assessments developed in this pilot study address a gap in the existing literature by focusing on the educational needs of pediatric patients with CHD, their parents, and health care providers [7,8,27]. While existing assessment tools mainly measure health-related quality of life and emotional well-being, they offer limited insight into health literacy and the unique educational challenges associated with CHD [38-56]. Moreover, these assessments are primarily developed quantitatively and lack qualitative insights into educational needs and burdens. We used qualitative methods to explore how respondents interpret, understand, and respond to specific survey items. This approach offered a more comprehensive understanding of the questions that assessments should address [29-32,93-97].

Through collaboration with stakeholders, including children, we developed and refined assessments to evaluate children's knowledge, parental educational burdens, and health care provider efficiency in pediatric CHD care. Including children in the development process was crucial due to the lack of tailored educational materials for children with CHD and the complexities involved in assessing this population [64-67,98]. While parents were present during the children's assessments to provide comfort, their involvement was carefully managed to ensure that it did not interfere with the children's meaningful participation. This approach contrasts with prior methods, where parents often act as proxies, potentially biasing the results [68,69]. Feedback from all stakeholders was collected post assessment to gain insights into their experiences and further improve the assessments.

We conducted pilot tests for two reasons: (1) to refine the assessments before involving more stakeholders, specifically children, and (2) to ensure the assessments were well adapted to meet all stakeholders' needs before the actual study [99,100]. This pilot study revealed critical gaps in existing tools and advanced the methodology for developing educational assessments. Emphasizing collaborative, iterative, and direct engagement with children, parents, and health care providers

XSL•FO
RenderX

ultimately leads to the design of tailored assessments. These assessments are crucial for informing the development of effective, child-led educational interventions for pediatric populations with CHD, demonstrating feasibility before broader implementation.

The assessments developed in this study have potential applications in routine clinical practice, offering health care providers a structured tool to assess and enhance CHD-related health literacy among pediatric patients and their families. By integrating these assessments into preappointment resources or waiting room activities, health care providers can identify educational gaps early and address them proactively. Furthermore, the assessments could complement existing educational interventions, providing feedback on the effectiveness of child- and family-centered resources to improve health literacy. This integration would not only inform the design of more effective child- and family-centered resources but also support continuous improvement of educational interventions. This approach aligns with broader goals in pediatric health care to support lifelong health management through early literacy interventions.

Limitations

Although the pilot testing provided valuable insights, several limitations are acknowledged. First, despite efforts to minimize parental influence, their presence may have still impacted some children's responses. Future research should explore strategies to reduce this influence further and ensure more independent responses from children. Second, the sample distribution was not spread evenly across age groups, with only 1 child aged 5 years and 4 children aged 7 years. This uneven distribution may affect the representativeness of the findings. Future studies should aim for a more balanced age distribution. Third, conducting assessments for both children and parents at the clinic during their visit might have influenced their responses, as they were exposed to medical procedures. Furthermore, the researcher's presence during the assessment of parents and children could have influenced their feedback. Finally, while general accessibility was considered, specific adaptations for educational disabilities were not within the scope of this pilot study.

Conclusions

This pilot study aims to improve educational interventions and care coordination through tailored assessments designed by stakeholder feedback, including affected children. The findings emphasize the importance of interactive and qualitative methods that foster multistakeholder engagement and ensure question relevance. These assessments can potentially enhance child-led interventions and improve outcomes for patients with CHD from childhood through to adult care. However, the iterative development process highlighted several challenges, such as managing parental presence to encourage independent responses from children and the logistical complexities of recruiting diverse participants within clinical settings. It is also important to ensure that assessments for younger children are read aloud without influencing their responses. Integrating these assessments into real-world clinical workflows requires adaptability to avoid disrupting patient care. Addressing these practical challenges will be essential to scaling and sustaining the use of these assessments in diverse health care contexts.

Acknowledgments

The authors gratefully acknowledge the funding provided by the Kusske Design Initiative (KDI) and Dr Gwenyth Fischer, MD. Their sincere thanks extend to the M Health Fairview Explorer Clinic and all pilot participants. Finally, they would like to thank contributions from former students Grace Rubas, Jessica Jenkins, Jonathan Jakubas, Levi Skelton, and Andy Thai.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Existing assessment tools.
[DOCX File , 21 KB - formative_v9i1e63818_app1.docx]

Multimedia Appendix 2
Three congenital heart disease–specific assessments.
[DOCX File , 3612 KB - formative_v9i1e63818_app2.docx]

Multimedia Appendix 3
Sample prompts and feedback from data collection.
[DOCX File , 23 KB - formative_v9i1e63818_app3.docx]

References

1. Centers for Disease Control and Prevention. Data and statistics on congenital heart defects. 2023. URL: https://www.cdc.gov/heart-defects/data/index.html [accessed 2023-09-22]

XSL•FO
RenderX

2. GBD 2017 Congenital Heart Disease Collaborators. Global, regional, and national burden of congenital heart disease, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet Child Adolesc Health 2020;4(3):185-200 [FREE Full text] [doi: 10.1016/S2352-4642(19)30402-X] [Medline: 31978374]
3. Liu Y, Chen S, Zühlke L, Black GC, Choy M, Li N, et al. Global birth prevalence of congenital heart defects 1970-2017: updated systematic review and meta-analysis of 260 studies. Int J Epidemiol 2019;48(2):455-463 [FREE Full text] [doi: 10.1093/ije/dyz009] [Medline: 30783674]
4. Chong LSH, Fitzgerald DA, Craig JC, Manera KE, Hanson CS, Celermajer D, et al. Children's experiences of congenital heart disease: a systematic review of qualitative studies. Eur J Pediatr 2018;177(3):319-336. [doi: 10.1007/s00431-017-3081-y] [Medline: 29327140]
5. Lee A, Bailey B, Cullen-Dean G, Aiello S, Morin J, Oechslin E. Transition of care in congenital heart disease: ensuring the proper handoff. Curr Cardiol Rep 2017;19(6):55. [doi: 10.1007/s11886-017-0859-5] [Medline: 28477319]
6. Moons P, De Volder E, Budts W, De Geest S, Elen J, Waeytens K, et al. What do adult patients with congenital heart disease know about their disease, treatment, and prevention of complications? A call for structured patient education. Heart 2001;86(1):74-80 [FREE Full text] [doi: 10.1136/heart.86.1.74] [Medline: 11410567]
7. Barbazi N, Shin JY, Hiremath G, Lauff CA. Exploring Health Educational Interventions for Children with Congenital Heart Disease: A Scoping Review. JMIR Pediatr Parent 2024 Jan 24;8:e64814 [FREE Full text] [doi: http://dx.doi.org/10.2196/64814]
8. Wray J, Maynard L. The needs of families of children with heart disease. J Dev Behav Pediatr 2006;27(1):11-17. [doi: 10.1097/00004703-200602000-00002] [Medline: 16511363]
9. Veldtman GR, Matley SL, Kendall L, Quirk J, Gibbs JL, Parsons JM, et al. Illness understanding in children and adolescents with heart disease. Heart 2000;84(4):395-397 [FREE Full text] [doi: 10.1136/heart.84.4.395] [Medline: 10995408]
10. Goossens E, Fieuws S, Van Deyk K, Luyckx K, Gewillig M, Budts W, et al. Effectiveness of structured education on knowledge and health behaviors in patients with congenital heart disease. J Pediatr 2015;166(6):1370-1376.e1. [doi: 10.1016/j.jpeds.2015.02.041] [Medline: 25841537]
11. Burns J, Higgins C, Ganigara M, Kalivas B, Basken A. Health literacy in CHD. Cardiol Young 2022;32(9):1369-1372. [doi: 10.1017/S1047951122001020] [Medline: 35611812]
12. van der Heijden Z, de Gooijer F, Camps G, Lucassen D, Feskens E, Lasschuijt M, et al. User requirements in developing a novel dietary assessment tool for children: mixed methods study. JMIR Form Res 2024;8:e47850 [FREE Full text] [doi: 10.2196/47850] [Medline: 38300689]
13. Francis-Oliviero F, Loubières C, Grové C, Marinucci A, Shankland R, Salamon R, et al. Improving children's mental health literacy through the cocreation of an intervention and scale validation: protocol for the CHILD-Mental Health literacy research study. JMIR Res Protoc 2023;12:e51096 [FREE Full text] [doi: 10.2196/51096] [Medline: 37796588]
14. Gupta A, Cafazzo JA, IJzerman MJ, Swart JF, Vastert S, Wulffraat NM, et al. Genomic health literacy interventions in pediatrics: scoping review. J Med Internet Res 2021;23(12):e26684 [FREE Full text] [doi: 10.2196/26684] [Medline: 34951592]
15. Shin JY, Holtz B. Identifying opportunities and challenges: how children use technologies for managing diabetes. 2020 Presented at: IDC '20: Interaction Design and Children; 2020 June 24; London, United Kingdom p. 495-507 URL: https://dl.acm.org/doi/10.1145/3392063.3394444
16. Shin JY, Holtz BE. Towards better transitions for children with diabetes: user experiences on a mobile health app. 2019 Presented at: Proceedings of the 18th ACM International Conference on Interaction Design and Children; 2019 June 15; Boise, ID. [doi: 10.1145/3311927.3325319]
17. DeHoff BA, Staten LK, Rodgers RC, Denne SC. The role of online social support in supporting and educating parents of young children with special health care needs in the United States: a scoping review. J Med Internet Res 2016;18(12):e333 [FREE Full text] [doi: 10.2196/jmir.6722] [Medline: 28007689]
18. Shin JY, Kedroske J, Vue R, Sankaran R, Chaar D, Churay T, et al. Design considerations for family-centered health management: preliminary findings with pediatric BMT patients. 2018 Presented at: Proceedings of the 17th ACM Conference on Interaction Design and Children; 2018 June 22; Trondheim, Norway. [doi: 10.1145/3202185.3210781]
19. Chaar D, Shin JY, Mazzoli A, Vue R, Kedroske J, Chappell G, et al. A mobile health app (Roadmap 2.0) for patients undergoing hematopoietic stem cell transplant: qualitative study on family caregivers' perspectives and design considerations. JMIR Mhealth Uhealth 2019;7(10):e15775 [FREE Full text] [doi: 10.2196/15775] [Medline: 31651402]
20. Rodts ME, Unaka NI, Statile CJ, Madsen NL. Health literacy and caregiver understanding in the CHD population. Cardiol Young 2020;30(10):1439-1444. [doi: 10.1017/S1047951120002243] [Medline: 32746956]
21. Tennant R, Allana S, Mercer K, Burns CM. Exploring the experiences of family caregivers of children with special health care needs to inform the design of digital health systems: formative qualitative study. JMIR Form Res 2022;6(1):e28895 [FREE Full text] [doi: 10.2196/28895] [Medline: 34989692]
22. Shin JY, Okammor N, Hendee K, Pawlikowski A, Jenq G, Bozaan D. Development of the socioeconomic screening, active engagement, follow-up, education, discharge readiness, and consistency (SAFEDC) model for improving transitions of care: participatory design. JMIR Form Res 2022;6(4):e31277 [FREE Full text] [doi: 10.2196/31277] [Medline: 35412461]

XSL•FO
RenderX

23. Hölgyesi Á, Luczay A, Tóth-Heyn P, Muzslay E, Világos E, Szabó AJ, et al. The impact of parental electronic health literacy on disease management and outcomes in pediatric type 1 diabetes mellitus: cross-sectional clinical study. JMIR Pediatr Parent 2024;7:e54807 [FREE Full text] [doi: 10.2196/54807] [Medline: 38506893]
24. Sakamoto M, Ishikawa H, Suzuki A. Evaluation of parents' use of a child health care information app and their health literacy: cross-sectional study. JMIR Pediatr Parent 2024;7:e48478 [FREE Full text] [doi: 10.2196/48478] [Medline: 38623727]
25. Castor C, Lindkvist RM, Hallström IK, Holmberg R. Health care professionals' experiences and views of eHealth in pediatric care: qualitative interview study applying a theoretical framework for implementation. JMIR Pediatr Parent 2023;6:e47663 [FREE Full text] [doi: 10.2196/47663] [Medline: 37851500]
26. Lindström NB, Pozo RR. Perspectives of nurses and doulas on the use of information and communication technology in intercultural pediatric care: qualitative pilot study. JMIR Pediatr Parent 2020;3(1):e16545 [FREE Full text] [doi: 10.2196/16545] [Medline: 32181748]
27. Arya B, Glickstein JS, Levasseur SM, Williams IA. Parents of children with congenital heart disease prefer more information than cardiologists provide. Congenit Heart Dis 2013;8(1):78-85 [FREE Full text] [doi: 10.1111/j.1747-0803.2012.00706.x] [Medline: 22891764]
28. Van Orne J. Care coordination for children with medical complexity and caregiver empowerment in the process: a literature review. J Spec Pediatr Nurs 2022;27(3):e12387. [doi: 10.1111/jspn.12387] [Medline: 35671389]
29. Willis GB, Artino AR. What do qur respondents think we're Asking? using cognitive interviewing to improve medical education surveys. J Grad Med Educ 2013;5(3):353-356 [FREE Full text] [doi: 10.4300/JGME-D-13-00154.1] [Medline: 24404294]
30. Balza JS, Cusatis RN, McDonnell SM, Basir MA, Flynn KE. Effective questionnaire design: how to use cognitive interviews to refine questionnaire items. J Neonatal Perinatal Med 2022;15(2):345-349 [FREE Full text] [doi: 10.3233/NPM-210848] [Medline: 35034911]
31. Kedroske J, Koblick S, Chaar D, Mazzoli A, O'Brien M, Yahng L, et al. Development of a national caregiver health survey for hematopoietic stem cell transplant: qualitative study of cognitive interviews and verbal probing. JMIR Form Res 2020;4(1):e17077 [FREE Full text] [doi: 10.2196/17077] [Medline: 32012037]
32. Eccles DW, Arsal G. The think aloud method: what is it and how do I use it? Qual Res Sport Exerc Health 2017;9(4):514-531. [doi: 10.1080/2159676x.2017.1331501]
33. Presser S, Couper MP, Lessler JT, Martin E, Martin J, Rothgeb JM, et al. Methods for testing and evaluating survey questions. Public Opin Q 2004;68(1):109-130. [doi: 10.1093/poq/nfh008]
34. Collins D. Pretesting survey instruments: an overview of cognitive methods. Qual Life Res 2003;12(3):229-238. [doi: 10.1023/a:1023254226592] [Medline: 12769135]
35. Mingjie Z, Li W, Jianxin Z. Developing a computer system for health assessment. 2009 Presented at: Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Huma; 2009 November 26; New York, NY p. 303-306. [doi: 10.1145/1655925.1655981]
36. Morris M, Intille S. HCI challenges in health assessment. 2005 Presented at: CHI '05 Extended Abstracts on Human Factors in Computing Systems; 2005 April 7; New York, NY. [doi: 10.1145/1056808.1057122]
37. Ricci M, Cimini A, Gli?ovi? K, Jiménez RJ, Medina-García R, Steinböck M, et al. Quality of life assessment methodology in TeNDER project. 2022 Presented at: Proceedings of the 15th International Conference on Pervasive Technologies Related to Assistive Environments; 2022 June 29; New York, NY. [doi: 10.1145/3529190.3534743]
38. Ayers TS, Sandler IN, West SG, Roosa MW. Children?s Coping Strategies Checklist. 1991. URL: https://psycnet.apa.org/doiLanding?doi=10.1037%2Ft42024-000 [accessed 2024-12-13]
39. ScienceDirect Topics. Child behavior checklist—an overview. URL: https://www-sciencedirect-com.ezp2.lib.umn.edu/topics/psychology/child-behavior-checklist [accessed 2024-03-12]
40. Haley SM, Coster WI, Kao YC, Dumas HM, Fragala-Pinkham MA, Kramer JM, et al. Lessons from use of the pediatric evaluation of disability inventory: where do we go from here? Pediatr Phys Ther 2010;22(1):69-75 [FREE Full text] [doi: 10.1097/PEP.0b013e3181cbfbf6] [Medline: 20142708]
41. Horner-Johnson W, Krahn G, Andresen E, Hall T, Rehabilitation ResearchTraining Center Expert Panel on Health Status Measurement. Developing summary scores of health-related quality of life for a population-based survey. Public Health Rep 2009;124(1):103-110 [FREE Full text] [doi: 10.1177/003335490912400113] [Medline: 19413032]
42. Hullmann SE, Ryan JL, Ramsey RR, Chaney JM, Mullins LL. Measures of general pediatric quality of life: Child Health Questionnaire (CHQ), DISABKIDS Chronic Generic Measure (DCGM), KINDL-R, Pediatric Quality of Life Inventory (PedsQL) 4.0 Generic Core Scales, and Quality of My Life Questionnaire (QoML). Arthritis Care Res (Hoboken) 2011;63 Suppl 11:S420-S430 [FREE Full text] [doi: 10.1002/acr.20637] [Medline: 22588762]
43. Child Outcomes Research Consortium. Strengths and Difficulties Questionnaire (SDQ). URL: https://www.corc.uk.net/outcome-experience-measures/strengths-and-difficulties-questionnaire-sdq/ [accessed 2024-12-20]
44. Marino BS, Tomlinson RS, Wernovsky G, Drotar D, Newburger JW, Mahony L, Pediatric Cardiac Quality of Life Inventory Testing Study Consortium. Validation of the pediatric cardiac quality of life inventory. Pediatrics 2010;126(3):498-508 [FREE Full text] [doi: 10.1542/peds.2009-2973] [Medline: 20805147]

XSL•FO
RenderX

45. Pearson. Pediatric Evaluation of Disability Inventory. URL: https://www.pearsonassessments.com/store/usassessments/en/Store/Professional-Assessments/Developmental-Early-Childhood/Pediatric-Evaluation-of-Disability-Inventory/p/100000505.html [accessed 2024-12-20]
46. Cleveland Clinic Children's. Pediatric Cardiac Quality of Life Inventory (PCQLI). URL: https://my.clevelandclinic.org/pediatrics/medical-professionals/pediatric-cardiac-quality-life-inventory [accessed 2024-12-20]
47. PedsQL TM (Pediatric Quality of Life Inventory TM). URL: https://www.pedsql.org/about_pedsql.html [accessed 2024-12-20]
48. Health Literacy Tool Shed. URL: https://healthliteracy.tuftsmedicine.org/ [accessed 2024-12-20]
49. Reynolds JS, Treu JA, Njike V, Walker J, Smith E, Katz CS, et al. The validation of a food label literacy questionnaire for elementary school children. J Nutr Educ Behav 2012;44(3):262-266. [doi: 10.1016/j.jneb.2011.09.006] [Medline: 22406011]
50. Child Outcomes Research Consortium. Parental Stress Scale (PSS). URL: https://www.corc.uk.net/outcome-experience-measures/parental-stress-scale-pss/ [accessed 2024-12-20]
51. American Psychological Association. Parenting Stress Index. URL: https://www.apa.org/pi/about/publications/caregivers/practice-settings/assessment/tools/parenting-stress [accessed 2024-12-20]
52. Ríos M, Zekri S, Alonso-Esteban Y, Navarro-Pardo E. Parental stress assessment with the Parenting Stress Index (PSI): a systematic review of its psychometric properties. Children (Basel) 2022;9(11):1649 [FREE Full text] [doi: 10.3390/children9111649] [Medline: 36360377]
53. Varni JW, Sherman SA, Burwinkle TM, Dickinson PE, Dixon P. The PedsQL family impact module: preliminary reliability and validity. Health Qual Life Outcomes 2004;2:55 [FREE Full text] [doi: 10.1186/1477-7525-2-55] [Medline: 15450120]
54. HealthinAging. org. Caregiver Self Assessment Questionnaire. URL: https://www.healthinaging.org/tools-and-tips/caregiver-self-assessment-questionnaire [accessed 2024-12-20]
55. Epstein-Lubow G, Gaudiano BA, Hinckley M, Salloway S, Miller IW. Evidence for the validity of the American Medical Association's caregiver self-assessment questionnaire as a screening measure for depression. J Am Geriatr Soc 2010;58(2):387-388. [doi: 10.1111/j.1532-5415.2009.02701.x] [Medline: 20370867]
56. Hanmer J, Jensen RE, Rothrock N, HealthMeasures Team. A reporting checklist for HealthMeasures' patient-reported outcomes: ASCQ-Me, Neuro-QoL, NIH Toolbox, and PROMIS. J Patient Rep Outcomes 2020;4(1):21 [FREE Full text] [doi: 10.1186/s41687-020-0176-4] [Medline: 32215788]
57. de Brantes F. Measuring Provider Efficiency Version 1. 2004. URL: https://www.commonwealthfund.org/sites/default/files/documents/___media_files_publications_other_2004_dec_measuring_provider_efficiency__version_1_0__a_collaborative_multi_stakeholder_effort_measurproviderefficiency1_12312004_pdf.pdf [accessed 2004-12-03]
58. Palmer S, Torgerson DJ. Economic notes: definitions of efficiency. BMJ 1999;318(7191):1136 [FREE Full text] [doi: 10.1136/bmj.318.7191.1136] [Medline: 10213735]
59. Bailie R, Bailie J, Larkins S, Broughton E. Editorial: continuous quality improvement (CQI)—advancing understanding of design, application, impact, and evaluation of CQI approaches. Frontiers in Public Health. 2017. URL: https://www.frontiersin.org/journals/public-health/articles/10.3389/fpubh.2017.00306 [accessed 2024-02-24]
60. Henrique DB, Godinho Filho M. A systematic literature review of empirical research in Lean and Six Sigma in healthcare. Total Qual Manag Bus Excell 2018;31(3-4):429-449. [doi: 10.1080/14783363.2018.1429259]
61. Hussey PS, de Vries H, Romley J, Wang MC, Chen SS, Shekelle PG, et al. A systematic review of health care efficiency measures. Health Serv Res 2009;44(3):784-805 [FREE Full text] [doi: 10.1111/j.1475-6773.2008.00942.x] [Medline: 19187184]
62. O'Neill SM, Hempel S, Lim YW, Danz MS, Foy R, Suttorp MJ, et al. Identifying continuous quality improvement publications: what makes an improvement intervention 'CQI'? BMJ Qual Saf 2011;20(12):1011-1019 [FREE Full text] [doi: 10.1136/bmjqs.2010.050880] [Medline: 21727199]
63. Rathi R, Vakharia A, Shadab M. Lean Six Sigma in the healthcare sector: a systematic literature review. Mater Today Proc 2022;50:773-781 [FREE Full text] [doi: 10.1016/j.matpr.2021.05.534] [Medline: 35155129]
64. Shepard L, Kagan SL, Wurtz E. Principles and recommendations for early childhood assessments. National Education Goals Panel. 1998. URL: https://eric.ed.gov/?id=ED416033 [accessed 2024-03-16]
65. Committee on the Science of Children Birth to Age 8: Deepening and Broadening the Foundation for Success, Board on Children, Youth, and Families, Institute of Medicine. In: National Research Council. Transforming the Workforce for Children Birth Through Age 8: A Unifying Foundation. Washington, DC: National Academies Press; 2015.
66. Banker A, Lauff C. Usability testing with children: history of best practices, comparison of methods & gaps in literature. URL: https://dl.designresearchsociety.org/drs-conference-papers/drs2022/researchpapers/225 [accessed 2022-06-25]
67. Vang PC, Lauff CA. Reflections on data collection during toy prototype development in a design studio course. 2024 Presented at: Proceedings of the 23rd Annual ACM Interaction Design and Children Conference; 2024 June 20; New York, NY. [doi: 10.1145/3628516.3659422]
68. Koutná V, Blatný M, Jelínek M. Concordance of child self-reported and parent proxy-reported posttraumatic growth in childhood cancer survivors. Cancers (Basel) 2021;13(16):4230 [FREE Full text] [doi: 10.3390/cancers13164230] [Medline: 34439384]

XSL•FO
RenderX

69. Smriti D, Kao TSA, Rathod R, Shin JY, Peng W, Williams J, et al. Motivational interviewing conversational agent for parents as proxies for their children in healthy eating: development and user testing. JMIR Hum Factors 2022;9(4):e38908 [FREE Full text] [doi: 10.2196/38908] [Medline: 36206036]
70. Angell C, Alexander J, Hunt JA. 'Draw, write and tell': a literature review and methodological development on the 'draw and write' research method. J Early Child Res 2014;13(1):17-28. [doi: 10.1177/1476718x14538592]
71. McWhirter J. The draw and write technique as a versatile tool for researching children's understanding of health and well-being. Int J Health Promot Educ 2014;52(5):250-259. [doi: 10.1080/14635240.2014.912123]
72. Teela L, Verhagen LE, van Oers HA, Kramer EEW, Daams JG, Gruppen MP, et al. Pediatric patient engagement in clinical care, research and intervention development: a scoping review. J Patient Rep Outcomes 2023;7(1):32 [FREE Full text] [doi: 10.1186/s41687-023-00566-y] [Medline: 36988738]
73. Wang Q, Hay M, Clarke D, Menahem S. Adolescents' drawings of their cardiac abnormality. Cardiol Young 2011;21(5):556-561. [doi: 10.1017/S1047951111000461] [Medline: 21554829]
74. Noonan RJ, Boddy LM, Fairclough SJ, Knowles ZR. Write, draw, show, and tell: a child-centred dual methodology to explore perceptions of out-of-school physical activity. BMC Public Health 2016;16:326 [FREE Full text] [doi: 10.1186/s12889-016-3005-1] [Medline: 27080384]
75. Flynn R, Walton S, Scott SD. Engaging children and families in pediatric health research: a scoping review. Res Involv Engagem 2019;5:32 [FREE Full text] [doi: 10.1186/s40900-019-0168-9] [Medline: 31700676]
76. Druin A. The role of children in the design of new technology. Behav Inf Technol 2002;21(1):1-25. [doi: 10.1080/01449290110108659]
77. Sanders EBN, Brandt E, Binder T. A framework for organizing the toolstechniques of participatory design. 2010 Presented at: Proceedings of the 11th Biennial Participatory Design Conference; 2010 November 29; New York, NY p. 195-198. [doi: 10.1145/1900441.1900476]
78. Read JC, MacFarlane S. Using the fun toolkit and other survey methods to gather opinions in child computer interaction. 2006 Presented at: Proceedings of the 2006 Conference on Interaction Design and Children; 2006 June 9; New York, NY p. A. [doi: 10.1145/1139073.1139096]
79. Bradbury-Jones C, Taylor J. Engaging with children as co-researchers: challenges,counter-challenges and solutions. Int J Soc Res Methodol 2013;18(2):161-173. [doi: 10.1080/13645579.2013.864589]
80. Abdelghani R, Law E, Desvaux C, Oudeyer PY, Sauzéon H. Interactive environments for training children's curiosity through the practice of metacognitive skills: a pilot study. : Association for Computing Machinery; 2023 Presented at: Proceedings of the 22nd Annual ACM Interaction Design and Children Conference; 2023 June 23; New York, NY p. 495-501. [doi: 10.1145/3585088.3593880]
81. Boberg M, Karapanos E, Holopainen J, Lucero A. PLEXQ: Towards a Playful Experiences Questionnaire. New York, NY: Association for Computing Machinery; 2015 Presented at: Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play; 2015 October 7; London. [doi: 10.1145/2793107.2793124]
82. Kellett M, Forrest R, Dent N, Ward S. 'Just teach us the skills please, we'll do the rest': empowering ten - year - olds as active researchers. Children Soc 2006;18(5):329-343. [doi: 10.1002/chi.807]
83. Druin A. Cooperative inquiry: developing new technologies for children with children. Proc SIGCHI Conf Hum Factors Comput Syst New York, NY. New York, NY: Association for Computing Machinery; 1999 Presented at: Proceedings of the SIGCHI conference on Human Factors in Computing Systems; 1999 May 20; Pittsburgh, PA. [doi: 10.1145/302979.303166]
84. Yale-Griffin Prevention Research Center. Food Label Literacy for Applied Nutrition Knowledge (FLLANK). URL: https://yalegriffinprc.griffinhealth.org/products-resources/prc-products/food-label-literacy [accessed 2023-12-20]
85. Stals S, Voysey I, Baillie L. Let's make this fun!: activities to motivate children teens to complete questionnaires. New York, NY: Association for Computing Machinery; 2024 Presented at: ACMIEEE International Conference on Human-Robot Interaction; 2024 March 15; Boulder, CO. [doi: 10.1145/3610978.3640675]
86. Gaines KS, Curry ZD. The inclusive classroom: the effects of color on learning and behavior. J Fam Consum Sci Educ 2011;29(1):46-57 [FREE Full text]
87. Barbazi N, Wang CX. Hum factors Aging Spec Needs. Perceiving through colors: visual supports for children with autism 2023;88:104-112. [doi: 10.54941/ahfe1003667]
88. Hall L, Hume C, Tazzyman S. Five degrees of happinessffective smiley face likert scales for evaluating with children. New York, NY: Association for Computing Machinery; 2016 Presented at: Proceedings of the 15th International Conference on Interaction Design and Children; 2016 June 24; Manchester, UK p. 311-321. [doi: 10.1145/2930674.2930719]
89. National Cancer Institute at the National Institutes of Health. Health measures. URL: https://healthcaredelivery.cancer.gov/healthmeasures/ [accessed 2024-12-20]
90. HealthMeasures. Neuro-QoL. 2024. URL: https://www.healthmeasures.net/explore-measurement-systems/neuro-qol [accessed 2024-02-17]
91. Stein DM, Wrenn JO, Stetson PD, Bakken S. What "to-do" with physician task lists: clinical task model development and electronic health record design implications. AMIA Annu Symp Proc 2009;2009:624-628 [FREE Full text] [Medline: 20351930]

XSL•FO
RenderX

92. Braun V, Clarke V. Thematic analysis. In: APA Handbook of Research Methods in Psychology Vol 2. Research Designs: Quantitative, Qualitative, Neuropsychological, and Biological. Washington, DC: American Psychological Association; 2012:57-71.
93. Creswell JW, Creswell JD. Research Design: Qualitative, Quantitative, and Mixed Methods Approaches. Washington DC: SAGE Publications; 2017.
94. Martin B, Hanington B. Universal Methods of DesignWays to Research Complex Problems, Develop Innovative Ideas, and Design Effective Solutions. Gloucester MA: Rockport Publishers; 2012.
95. Creswell JW, Poth CN. Qualitative Inquiry and Research Design: Choosing Among Five Approaches. Washington, DC: SAGE Publications; 2016.
96. DeJonckheere M, Vaughn LM. Semistructured interviewing in primary care research: a balance of relationship and rigour. Fam Med Community Health 2019;7(2):e000057 [FREE Full text] [doi: 10.1136/fmch-2018-000057] [Medline: 32148704]
97. de Souza RH, Dorneles CF. Searchingranking questionnaires: an approach to calculate similarity between questionnaires. New York, NY: Association for Computing Machinery; 2019 Presented at: Proceedings of the ACM Symposium on Document Engineering 2019; 2019 September 26; Berlin, Germany p. 1-9. [doi: 10.1145/3342558.3345390]
98. Coyne I. Children's participation in consultations and decision-making at health service level: a review of the literature. Int J Nurs Stud 2008;45(11):1682-1689. [doi: 10.1016/j.ijnurstu.2008.05.002] [Medline: 18706560]
99. Hassan ZA, Schattner P, Mazza D. Doing a pilot study: why is it essential? Malays Fam Physician 2006;1(2-3):70-73 [FREE Full text] [Medline: 27570591]
100. Schroder C, Medves J, Paterson M, Byrnes V, Chapman C, O'Riordan A, et al. Development and pilot testing of the collaborative practice assessment tool. J Interprof Care 2011;25(3):189-195. [doi: 10.3109/13561820.2010.532620] [Medline: 21182434]

Abbreviations

CBCL: Child Behavior Checklist
CHD: congenital heart disease
CHD-HEEA: Congenital Heart Disease Healthcare Provider Educational Efficiency Assessment
CHD-HLCA: Congenital Heart Disease Health Literacy Children Assessment
CHD-PEBA: Congenital Heart Disease Parental Educational Burden Assessment
CHQ: Child Health Questionnaire
FIM: Family Impact Module
FLLANK: Food Label Literacy for Applied Nutrition Knowledge
KPI: Key Performance Indicator
LSS: Lean Six Sigma
PCQLI: Pediatric Cardiac Quality of Life Inventory
PedsQL: Pediatric Quality of Life Inventory
PSI/PSS: Parenting Stress Index/Parental Stress Scale

Edited by A Mavragani; submitted 30.06.24; peer-reviewed by F Yang, LB Carey; comments to author 13.11.24; revised version received 02.12.24; accepted 02.12.24; published 27.01.25.

Please cite as:
Barbazi N, Shin JY, Hiremath G, Lauff CA
Developing Assessments for Key Stakeholders in Pediatric Congenital Heart Disease: Qualitative Pilot Study to Inform Designing of a Medical Education Toy
JMIR Form Res 2025;9:e63818
URL: https://formative.jmir.org/2025/1/e63818
doi:10.2196/63818
PMID:

©Neda Barbazi, Ji Youn Shin, Gurumurthy Hiremath, Carlye Anne Lauff. Originally published in JMIR Formative Research (https://formative.jmir.org), 27.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Effect of the Reassured Self-Compassion–Based School Program on Anxiety, Video Game Addiction, and Body Image Among Rural Female Adolescents: Retrospective Study

Areeg Zuair[1], PhD

Community Health Nursing Department, College of Nursing, Taibah University, Medina, Saudi Arabia

Corresponding Author:
Areeg Zuair, PhD
Community Health Nursing Department
College of Nursing
Taibah University
Prince Naif Ibn Abdulaziz Rd
Medina, 42351
Saudi Arabia
Phone: 966 594800400
Email: Azuair@taibahu.edu.sa

Abstract

Background: The COVID-19 pandemic has amplified mental health challenges among adolescents, particularly in rural areas with limited access to services. In response, the Saudi government launched mental health campaigns and mandated schools to implement mental health programs. However, the effectiveness of these programs remains largely unreported.

Objective: This study aims to determine the prevalence of anxiety disorder symptoms, video game addiction, and body image dissatisfaction, as well as to examine the effect of a school-based program, The Reassured Self, grounded in compassion-focused therapy, on anxiety symptoms, video game addiction, and body image dissatisfaction among rural adolescent females in Saudi Arabia.

Methods: A retrospective secondary analysis of pre-post intervention data was used. All female middle school students (N=133; age: mean 13.7, SD 1.01 years) in grades 1-3 from a rural setting were included, with no exclusion criteria. Participants were recruited as part of a school-mandated mental health program. Data were collected at baseline (2 weeks before the intervention) and 2-3 weeks post intervention during school hours in a quiet classroom setting using teacher-supervised, printed surveys. Survey completion was voluntary, and students exhibiting high distress based on post data analysis were referred to the school health counselor for support. The intervention consisted of 3 sessions (30-35 minutes each) delivered over 2 weeks. Validated Arabic versions of the Spence Children's Anxiety Scale, Game Addiction Scale, and Body Image Discrepancy Assessment were used to measure outcomes 2 weeks post intervention (April to June 2024). Data were analyzed using linear mixed-effects regression models. Data were analyzed using linear mixed effects regression models.

Results: Among the students, 25.8% (32/124) were identified as having high anxiety levels, with 19.3% (24/124) falling into the clinical concern or very high clinical concern categories. Additionally, 26% (32/123) were classified as problem gamers, while 14% (18/123) were categorized as addicted gamers. A smaller proportion, 5.1% (6/117), expressed a strong desire for thinness. Compared to pre-intervention levels, students exhibited significant reductions in anxiety symptoms (mean difference –4.11, 95% CI –6.98 to –1.24; *P*=.005) and video game addiction (mean difference –1.62, 95% CI –2.83 to –0.42; *P*=.009) following the program. However, changes in body image dissatisfaction were minimal and not statistically significant (mean difference 0.067, 95% CI –0.046 to 0.18; *P*=.057). High satisfaction scores, with a mean of 28.49 out of 35 (SD 6.31), indicated strong acceptability and cultural alignment of the intervention. High satisfaction scores indicated strong acceptability and cultural alignment with the intervention.

Conclusions: The results highlight the potential for compassion-focused school programs to address mental health disparities in underserved adolescent populations. Future research should explore the long-term effects and scalability of such culturally adapted interventions.

(JMIR Form Res 2025;9:e68840) doi:10.2196/68840

XSL•FO
RenderX

KEYWORDS

adolescents; rural; compassion-focused therapy; mental health; Saudi Arabia; school

Introduction

The COVID-19 pandemic has profoundly impacted adolescent mental health worldwide, notably increasing anxiety disorder symptoms in this population [1-3]. With societies implementing lockdowns, enforcing social distancing, and facing unprecedented uncertainties, adolescents were exposed to substantial stressors that affected their psychological well-being [1-3]. A meta-analysis across 29 samples and 80,879 youth, conducted post the pandemic, highlighted a global increase in anxiety, with clinically significant symptoms nearly doubling, from 11.6% prepandemic to 20.5% post the pandemic [1]. This surge underscores the pandemic's enduring psychological effects on adolescents' mental health. Adolescence, a crucial developmental phase, often marks the onset of mental health disorders, with 50% emerging by age 14 and 75% by the age of 24 years [4]. Without timely intervention, these conditions can have long-lasting impacts on well-being and life trajectories.

Female adolescents have been disproportionately affected in the post–COVID-19 era, facing higher anxiety prevalence and greater symptom severity, necessitating targeted interventions that address their unique mental health needs [1,5]. Modifiable risk factors for anxiety disorders include game addiction and body image dissatisfaction, both of which significantly affect female adolescents [3,6]. Studies report an increase in gaming addiction during and after the pandemic, with prevalence estimates ranging between 2.3% and 29.4% [6]. Research shows a bidirectional relationship where excessive gaming escalates anxiety, and anxious adolescents use gaming as a coping mechanism [7]. Although many studies focus on male adolescents, emerging evidence indicates that female adolescents are similarly at risk for gaming addiction and associated mental health concerns [8,9]. This highlights the need for gender-specific research and interventions. Reports also indicate an increase in disordered eating and body image issues since the COVID-19 outbreak, partly due to increased exposure to appearance-focused social media and heightened internalization of social pressures [3,5].

In Saudi Arabia, the mental health landscape among adolescents is concerning, with 40.1% experiencing at least one lifetime disorder, 26.84% affected by anxiety, and 7.06% affected by eating disorders [10]. Female adolescents, particularly those in rural areas, are at higher risk for anxiety and body dissatisfaction, with a significantly higher anxiety prevalence than males [11]. For instance, 14.1% (424/3008) of Saudi females are at risk for generalized anxiety disorder, compared to 11.4% (342/3007) of males [11].

Rural female adolescents face additional barriers. Limited access to mental health services and cultural stigma exacerbate untreated mental health disorders, with only 0.5% of at-risk individuals in Saudi receiving treatment [11]. This calls for culturally appropriate, accessible interventions tailored to the specific needs of rural female adolescents in Saudi Arabia. Low-intensity mental health interventions that target anxiety, game addiction, and body image issues are critical to address these unmet needs.

Currently, a gap exists in the Saudi literature regarding school-based mental health and game addiction interventions for adolescents. Despite research on gaming addiction prevalence and related mental health issues, there is limited research on intervention programs, particularly those tailored to Saudi adolescents and female-specific needs [12]. Additionally, comprehensive programs that address both gaming addiction and associated mental health issues like anxiety are scarce, with most research focusing on treatment rather than prevention [8,12]. Also, our search failed to identify any mental health school program among Saudi adolescents. Additionally, there is a notable gap in research on the use of self-compassion interventions in populations outside of Western, educated, industrialized, rich, and democratic countries [13]. In Saudi Arabia, specifically, no reports were found on community- or school-based programs for female adolescents using any mental health intervention methods [10,12-16]. This study aims to fill this gap by providing valuable data on an underserved population: rural female adolescents, who often lack access to mental health services. Designed within the Saudi cultural context, this school-based intervention addresses interconnected issues of anxiety, game addiction, and body image dissatisfaction. The program emphasizes self-reassurance skills, empowering adolescents to manage their mental health independently—an approach particularly relevant in a cultural setting where mental health stigma may discourage individuals from seeking external help [15,17]. By fostering resilience, this intervention addresses critical mental health needs within this underserved population, contributing to the limited research in the area and offering insights that could inform future health policies and practices in Saudi Arabia.

To address these gaps, this study aimed to identify the prevalence of anxiety disorder symptoms, game addiction, and body image dissatisfaction among rural female adolescents; evaluate the efficacy of a compassion-focused therapy (CFT) school-based mental health intervention, The Reassured Self, in reducing anxiety disorder symptoms, game addiction, and body image dissatisfaction among rural female adolescents; and measure the acceptability of The Reassured Self intervention among rural female adolescents.

Methods

Design and Participants

This study used a retrospective design using existing data collected through a predesigned program implemented in a public school setting for quality improvement purposes (see Multimedia Appendix 1 for the checklist used to report the results of this retrospective database study, as recommended by the International Society for Pharmacoeconomics and Outcomes Research Task Force [18]). A pre-post design with no control group was used to evaluate the efficacy of a school-based mental health program, The Reassured Self. The

XSL•FO
RenderX

intervention was conducted with middle school female students in an underserved rural area in the Medina region, Saudi Arabia. Saudi Arabian schools are required to offer health programs to students; however, reports on the efficacy and acceptability of these programs are limited. This study documents a partnership between a non-profit organization and a school to develop a mental health program and evaluates its efficacy and acceptability for future development.

Most rural areas in Saudi Arabia have a single girls' school. This project was implemented for all female middle school students (N=133) in grades 1-3, with no exclusion criteria, as the program was mandatory for all students (see Multimedia Appendix 2 for details on data completeness for intervention participants).

Intervention Program: The Reassured Self

The Reassured Self is a culturally adapted, low-intensity intervention program designed for Saudi Arabian students. Based on the principles of self-compassion, the program integrates cultural and religious values with evidence-based practices to promote mental and emotional resilience. It addresses anxiety, game addiction, and body image dissatisfaction through tailored strategies. The program was delivered to middle school female students in a classroom environment, fostering peer support and shared learning experiences. Each class received three 30- to 35-minute sessions within 2 weeks. To ensure consistency, a trained nurse with experience on delivering the program to other schools delivered the same program to all classes. The intervention content focused on 5 core components, as outlined in Table 1.

Table 1. Detailed overview of the topics and descriptions covered in The Reassured Self intervention program[a].

	Topic	Description
1	Foundational concepts of self-compassion	• Self-blaming vs Self-reassurance: Introduces the difference between self-blame and self-compassion, guiding students to develop a supportive inner dialogue. • Religious integration: Emphasizes Islamic principles written in the Quran and Hadith about compassion, assurance, self-blaming, and gratitude. Morning and evening supplications (Athkar) were incorporated for self-affirmation. • Examples: “O reassured soul, return to your Lord well-pleased and pleasing to Him” (Surah Al-Fajr, 89:27-28) to emphasize fostering inner peace and self-assurance. • “And I swear by the self-blaming soul” (Surah Al-Qiyamah, 75:2) to highlight self-reflection without destructive self-blame. Gratitude practices (Shukr): Students participated in gratitude journaling exercises, drawing on the verse: “If you are grateful, I will certainly give you more” (Surah Ibrahim, 14:7).
2	Building resilience	• Challenge as growth: Teaches students to view challenges as opportunities for growth and resilience by using Quranic verses about patience, effort, and striving (Sabr and Sa’y). For example: “And that there is not for man except that [good] for which he strives” (Surah An-Najm, 53:39).
3	Addressing game addiction	• Understanding dopamine and addiction: Explains how game addiction provides a temporary dopamine boost, leading to a false sense of achievement. • Healthier coping strategies: Encourages real-life connections, exercises, and activities as healthier sources of satisfaction.
4	Reframing bullying	• Understanding bullying behavior: Frames bullying as often stemming from the bully’s own self-blame, helping students empathize and avoid internalizing harmful behaviors. • Responding with resilience: Guides students to respond to negativity with self-respect and compassion, fostering an empathetic school environment.
5	Practical applications and exercises	• Self-compassion exercises: Includes activities like gratitude practices, physical self-care routines, and stress-management tools such as using stress-relief items. • Exercise as self-care: Promotes physical exercise as a form of self-respect and care, enhancing both mental and physical well-being. • Social skills for conflict resolution: Involves role-playing scenarios to provide practical guidance on maintaining self-respect in interactions with others, including handling criticism or bullying.

[a]The program integrates principles of self-compassion and Islamic values to foster emotional resilience, reduce anxiety, address game addiction, and promote healthy coping strategies. Each topic focuses on specific aspects of self-reassurance, resilience building, and practical exercises to enhance mental well-being in a culturally sensitive manner.

Measurements and Data Collection

Data were collected at 2 time points: baseline (2 weeks before the intervention) and postintervention (2-3 weeks after the intervention). Assessments were conducted during school hours in a quiet classroom setting. The teacher distributed the printed surveys (paper-and-pencil format) to students, who were given 45 minutes to complete the questionnaires. The teacher remained available in the classroom to answer any questions. After the follow-up data collection, students exhibiting high levels of psychological distress, identified through their questionnaire responses, were referred to the school health counselor at that time for additional support.

Anxiety Symptoms

To assess clinically significant levels of anxiety symptoms, the Spence Children's Anxiety Scale (SCAS)–Arabic version was administered [19]. The SCAS is a 38-item questionnaire rated

XSL•FO
RenderX

on a 4-point scale (0-3; never-always), with total scores reflecting the sum of responses [20]. This scale encompasses symptoms of the most prevalent *DSM-5* (*Diagnostic and Statistical Manual of Mental Disorders* [Fifth Edition]) anxiety disorders and is widely used in both clinical practice and research [21]. The SCAS includes 6 subscales that measure different dimensions of anxiety: separation anxiety, social phobia, obsessive-compulsive symptoms, panic or agoraphobia, physical injury fears, and generalized anxiety. Total anxiety was calculated by summing the scores from the individual subscales. In this study, the SCAS demonstrated excellent internal consistency (α=0.94). *t*-Scores and cutoffs for the SCAS are available on the SCAS website and are categorized by age and gender [22]. For girls aged 12-15 years, a total score cutoff of ≥60 (84th percentile) suggests clinically significant anxiety symptoms.

Video Game Addiction

A newly modified Arabic translation of the 7-item version of the Game Addiction Scale (GAS) was used to measure video game addiction [23]. Responses were given on a 5-point scale (1="never" to 5="very often"), with a total score range of 7-35. The original Arabic version of the GAS served as the basis for this translation [24]. In this study, the GAS-Arabic demonstrated good internal consistency (α=0.81).

Video Game Addiction Status

To classify participants based on their level of video game engagement, responses to the core 4-item criteria for addiction (withdrawal, relapse, conflict, and problems) were evaluated according to methods recommended in recent literature [25]. Participants were divided into 3 groups: addicted gamers, problem gamers, and engaged gamers. Respondents who scored 3 or higher on all 4 items were classified as Addicted Gamers. Those who scored 3 or higher on 2 or 3 of the 4 items were categorized as problem gamers, indicating some signs of problematic gaming behavior. Respondents who did not meet the criteria for either addicted or problem gamers were classified as engaged gamers, reflecting low engagement without signs of problematic behavior.

Body Image Dissatisfaction

Body image dissatisfaction was assessed using a series of silhouettes representing 4 different body sizes [26]. Students selected the silhouette that best represented their ideal body size and the one that most resembled their current body size. The difference between these selections (ideal vs current) indicated the level of body image discrepancy, with scores ranging from –3 to +3. A negative score suggested a desire for thinness (preference for a smaller body size), while a positive score indicated a desire for fullness (preference for a larger body size) [27]. Participants were categorized as follows: high drive for thinness (scores less than –1.5), low drive for thinness (scores between –1.5 and 0), and no drive for thinness (scores of 0 or higher).

Program Reaction

Participant feedback was collected using the Kirkpatrick model's level 1 evaluation 2-3 weeks post the program. The level 1 evaluation, also known as Reaction, assesses participants' immediate responses to a training or intervention [28]. It captures how participants felt about the experience, evaluating satisfaction, perceived ease and clarity, engagement (perceived effectiveness), and perceived relevance of the content [28]. A 7-item Likert scale was used, with responses ranging from 1 (totally disagree) to 5 (totally agree), resulting in a total score range of 7-35.

Statistical Analysis

Overview

Descriptive statistics (means, SDs, and frequencies) were calculated for all variables. A series of linear mixed models were used to examine the effects of the intervention on 3 dependent variables—anxiety symptoms, video game addiction, and body image discrepancy—measured at 2 time points (before and post the intervention). The models included time as a fixed effect and student ID as a random effect to account for within-subject variability across repeated measures. Restricted maximum likelihood estimation was used to fit the models. Grade levels were tested as fixed effects, but neither demonstrated significant associations with the outcomes or interactions with time. Consequently, these variables were excluded from the model. Pairwise comparisons of the estimated marginal means were conducted using least significant difference adjustment. The assumption of homoscedasticity (constant variance) appeared mostly satisfied, and the residuals were approximately normally distributed. Linearity between time and each dependent variable was reasonable, as no strong nonlinear patterns were observed. Also, correlation analyses were conducted to identify potential confounding variables, including age, baseline anxiety, body image discrepancy, and gaming addiction.

Participants with valid data on anxiety symptoms, video game addiction, or body image discrepancy at either baseline or follow-up were included in the analysis. Linear mixed models use all available data, regardless of missing outcome data, under the assumption that data are missing at random. Data were found to be missing completely at random, as indicated by Little's MCAR test (χ^2_{135}=110.5, *P*=.94). No cases were missing more than 75% of data, which is the recommended threshold for analyzing the SCAS.

Power Analysis

A post hoc power analysis was conducted using G*Power (version 3.1.9.7) to evaluate the achieved power for detecting changes in anxiety levels, game addiction, and body image discrepancy across pre- and postintervention time points. The paired samples *t* test was selected as the appropriate statistical test for the pre-post design, focusing on differences between dependent means. Effect sizes (dz) were calculated as 0.25, 0.24, and 0.076 based on the observed mean differences and SDs of the difference scores for anxiety, game addiction, and body image discrepancy, respectively.

The analysis was performed with an alpha level of 0.05 and total sample sizes of 133, 132, and 130 participants for anxiety, game addiction, and body image discrepancy, respectively. The results indicated that the study had an 89% and 86% probability of detecting a true effect for anxiety and game addiction,

XSL•FO
RenderX

respectively. However, the post hoc power for body image discrepancy was approximately 13.8%, indicating that the study had very low power to detect such a small effect. This suggests that the study may not have been adequately powered to identify significant changes in body image discrepancy.

Ethical Considerations

Given the retrospective nature of the study, all students participated in the program as it was mandated by the school health unit as part of the standard curriculum. Written consent was obtained from students specifically for their participation in completing the surveys associated with the program. Information about the surveys, including their purpose and voluntary nature, was communicated to the students in an age-appropriate manner. This included verbal explanations by teachers and written summaries that were easy to understand. Students were assured that completing the surveys was voluntary and would not affect their grades or school standing. To encourage participation, all students received a small gift (a relaxation squeeze toy) after completing the program satisfaction survey. Parental consent was not required, as the program itself was a required school activity, and the decision to waive parental consent for the surveys followed ethical guidelines for school-based health programs. This approach was approved by the University of Taibah Nursing College Ethical Review Committee (IRB No. TUCN-REC 11 2024). To ensure the privacy and confidentiality of participants, all data from school project records were securely stored and deidentified prior to analysis.

Results

Students' Characteristics

Table 2 presents the descriptive statistics for the study variables, including age, anxiety symptoms, video game addiction, and body image discrepancy. A total of 133 students participated in the program. However, only 93% (124/133) completed the preanxiety symptoms survey, 92% (123/133) completed the video game addiction, and 88% (117/133) completed the body image discrepancy assessment (see Multimedia Appendix 2 for details on data completeness for intervention participants). The mean age of participants was 13.72 (SD 1.01) years. Total anxiety symptoms had a mean score of 47.60 (SD 22.02), indicating a moderate level of anxiety across the sample. Among the anxiety subscales, panic or agoraphobia symptoms were the highest, with a mean of 9.38 (SD 6.76), followed by obsessive-compulsive symptoms (mean 8.79, SD 4.54) and generalized anxiety symptoms (mean 8.66, SD 4.55). Physical injury fears (mean 7.64, SD 3.75) and separation anxiety (mean 6.31, SD 3.88) were also prominent. Social phobia symptoms had a mean of 6.82 (SD 4.12), indicating a relatively lower prevalence compared to other anxiety types.

Table 2. Descriptive statistics for age, anxiety symptoms, video game addiction, and body image discrepancy among rural adolescent female students[a].

Descriptive statistics	Mean (SD) values
Age (years; n=133)	13.72 (1.01)
Total anxiety symptoms (n=124)	47.60 (22.02)
Generalized anxiety symptoms	8.66 (4.56)
Physical injury fears symptoms	7.64 (3.75)
Separation anxiety symptoms	6.31 (3.88)
Panic agoraphobia symptoms	9.38 (6.76)
Social phobia symptoms	6.82 (4.12)
Obsessive-compulsive symptoms	8.79 (4.54)
Video game addiction (n=123)	15.57 (6.67)
Body image discrepancy (n=117)	–0.197 (0.87)

[a]Descriptive statistics summarizing the mean and SD for age, anxiety symptoms, video game addiction, and body image discrepancy among rural adolescent female students who participated in The Reassured Self intervention program.

The mean score for video game addiction was 15.57 (SD 6.67), indicating a moderate level of engagement or potential addiction to gaming. The body image discrepancy score had a mean of –0.20 (SD 0.87), suggesting a slight drive for thinness among the sample, with negative values indicating a preference for a thinner body size compared to current perception. Correlations between baseline variables showed a significant, moderate positive relationship between anxiety and gaming addiction (r=0.457, P<.001). Age and body image discrepancy demonstrated weak, nonsignificant correlations with other variables.

Characterizing Mental Health Status

The distribution of anxiety symptoms among students reveals that a substantial portion experience elevated or clinically concerning levels across various anxiety types (Table 3). Specifically, 25.8% (32/124) of students had a high anxiety level, and 19.3% (24/124) fall within the clinical concern or very high clinical concern category, indicating that a significant subset may require attention. Generalized anxiety is particularly prevalent, with 33.1% (41/124) of students classified in the clinical concern or very high clinical concern categories. In addition, fears related to physical injury show one of the most severe patterns, with 33.9% (42/124) of students in the very high clinical concern category, highlighting this as a critical

XSL•FO
RenderX

area for mental health support. Separation anxiety also shows a high prevalence, with 33.9% (42/124) of students in the clinical or very high clinical concern categories. Symptoms of panic and agoraphobia are notable as well, with 28.2% (35/124) of students in the very high clinical concern category, indicating a significant impact. Obsessive-compulsive symptoms present another area of concern, with 21% (26/124) of students in the very high clinical concern category. In contrast, social phobia has a lower prevalence of severe concern, with only 2.4% (3/124) of students in the very high clinical concern category.

Table 3. Frequency and percentage distribution of anxiety levels, video game status, and body image dissatisfaction among rural adolescent female students.

Anxiety levels (N=124)[a]	Normal range	Elevated anxiety	Clinical concern	Very high clinical concern
Anxiety symptoms, n (%)	92 (74.2)	8 (6.5)	5 (4)	19 (15.3)
Generalized anxiety, n (%)	61 (49.2)	22 (17.7)	25 (20.2)	16 (12.9)
Physical injury fears, n (%)	38 (30.6)	20 (16.1)	24 (19.4)	42 (33.9)
Separation anxiety, n (%)	55 (44.4)	27 (21.8)	16 (12.9)	26 (21)
Panic and agoraphobia, n (%)	54 (43.5)	18 (14.5)	17 (13.7)	35 (28.2)
Social phobia, n (%)	93 (75)	12 (9.7)	16 (12.9)	3 (2.4)
Obsessive-compulsive, n (%)	38 (30.6)	30 (24.2)	30 (24.2)	26 (21)
Video game status (N=123)[a], n (%)	Engaged gamers: 73 (59)	Problem gamers: 32 (26)	Addicted gamers: 18 (14)	Addicted gamers: 18 (14)
Body image dissatisfaction (N=117)[a], n (%)	No drive for thinness: 77 (65.8)	Low drive for thinness: 34 (29.1)	High drive for thinness: 6 (5.1)	High drive for thinness: 6 (5.1)

[a]Total number of completed surveys.

Overall, the findings suggest that while some anxiety symptoms present moderate concern, specific areas—particularly physical injury fears, separation anxiety, and obsessive-compulsive symptoms—are more prevalent at clinically concerning levels, underscoring the need for focused mental health support in these domains.

Regarding body image dissatisfaction, the results indicate a slight drive for thinness among most of the students. Approximately 29.1% (34/117) of students exhibit a moderate inclination toward weight reduction, while a smaller proportion, 5.1% (6/117), express a strong desire for thinness.

Characterizing Video Game Status

The data on video game status among students reveals varying levels of gaming involvement, with a significant portion showing signs of problematic or addictive behavior (Table 3). Overall, 59% (73/123) of the students are categorized as engaged gamers, indicating regular gaming without notable adverse effects. However, 26% (32/123) are identified as problem gamers, suggesting a level of gaming that may interfere with daily responsibilities or personal well-being. Additionally, 14% (18/123) of students fall into the addicted gamers category, indicating a more severe level of dependency that likely impacts various aspects of their lives.

Effectiveness of The Reassured Self Program

The Reassured Self intervention demonstrated a positive effect in reducing anxiety levels among students, as shown by a significant decrease in anxiety scores from preintervention to follow-up (see Table 4). The effect size, with a mean difference of –4.11, was small but statistically significant (P=.005). Additionally, the intervention had a beneficial impact on reducing video game addiction, with an estimated mean reduction of –1.62 in game addiction scores, which was also statistically significant (P=.009). In contrast, although the intervention led to a slight decrease in the drive for thinness, this change was not statistically significant (P=.057), suggesting a minimal effect on body image dissatisfaction. Overall, these findings indicate that the intervention was effective in reducing both anxiety and game addiction among students, while its impact on body image discrepancy remains inconclusive.

Table 4. Changes in anxiety symptoms, video game addiction, and body image dissatisfaction from baseline to follow-up[a].

Outcome	Baseline, mean (SE) [95% CI]	Follow-up, mean (SE) [95% CI]	Within group differences change from follow-up–baseline	
			Mean difference [95% CI]	*P* value
Anxiety symptoms (n=133)	47.137 (1.93) [43.03 to 50.95]	43.32 (2.13) [38.81 to 47.25]	–4.11 [–6.98 to –1.24]	.005
Video game addiction (n=132)	15.45 (0.59) [14.28 to 16.63]	13.83 (0.61) [12.63 to 15.03]	–1.62 [–2.83 to –0.42]	.009
Body image dissatisfaction (n=130)	–0.19 (0.08) [–0.35 to 0.038]	–0.125 (0.07) [–0.27 to 0.02]	0.067 [–0.046 to 0.18]	.057

[a]This table presents the results of linear mixed-effects models, which accounted for repeated measures within individuals, to evaluate changes in anxiety symptoms, video game addiction, and body image dissatisfaction from baseline to follow-up among participants of The Reassured Self intervention program. Adjustments for multiple comparisons were made using the least significant difference method.

XSL•FO
RenderX

Program Reaction

The program received positive feedback across multiple dimensions, as reflected in the high mean scores for various aspects of participant satisfaction (Table 5). The overall satisfaction sum score ranged from 7 to 35, with a mean of 28.49 (SD 6.31), indicating a generally favorable response. The overall results suggest that participants responded positively to the program, finding it engaging, clear, and psychologically beneficial, with a strong likelihood to recommend it to others.

Table 5. Descriptive statistics for reaction measures of The Reassured Self program (N=123)[a].

Measure	Mean (SD) values
Overall satisfaction score	28.49 (6.31)
Perceived ease and clarity	4.41 (0.88)
Enjoyment and satisfaction	4.16 (1.14)
Psychological benefits gained	4.14 (1.11)
Program relevance and value	3.9 (1.29)
Teaching effectiveness	3.97 (1.13)
Engagement and attention retention	4.15 (1.09)
Likelihood of recommending the program	4.2 (1.20)

[a]This table presents the mean scores and SDs for various aspects of participant satisfaction with The Reassured Self program. The consistently high scores across all measures reflect positive feedback and strong acceptability of the intervention.

Discussion

Principal Findings

This study conducted a retrospective analysis of data collected from female adolescents who participated in a CFT-based school program, The Reassured Self, in a rural setting. The program, collaboratively implemented by a nonprofit organization and an academic institution, aimed to reduce anxiety, video game addiction, and body image dissatisfaction among rural Saudi female adolescents, addressing critical mental health challenges within this population. The findings revealed a high prevalence of anxiety and video game addiction, consistent with national averages. The intervention effectively reduced anxiety and video game addiction symptoms but showed no improvement in body image dissatisfaction, highlighting its potential as a health education tool.

Anxiety Reduction

The overall high anxiety symptom prevalence of 25.8% (32/124) and the high severe anxiety symptoms of 19.3% (24/124) in our sample are comparable to global postpandemic figures (20.5%) [1] and align closely with the recent national prevalence of lifetime anxiety disorders among adolescents (26.84%) [10]. Our results underscore the urgent need for interventions targeting adolescent anxiety in underserved areas. Recent literature argues that school well-being programs should consider children's and adolescents' anxiety levels as a key metric for measuring program impact [29]. The Reassured Self program demonstrated a significant reduction in anxiety symptoms, suggesting that a culturally adapted CFT approach can effectively address anxiety. This aligns with recent meta-analyses and systematic reviews, which consistently show that compassion-based interventions reduce anxiety and psychological distress among adolescents [30-32].

This study's findings also align with recent evidence on compassion-based school interventions. For instance, Maratos et al [29] evaluated a Compassionate Mind Training (CMT-Pupils) intervention among Year 7 students aged 11-12 and found significant improvements in anxiety levels compared to the control group. Similar to our intervention, CMT-Pupils emphasized emotional regulation and prosocial behaviors through compassion-based practices, including mindfulness and self-compassion. Their study highlighted the importance of promoting kindness, emotional awareness, and classroom inclusivity, which aligns with the core mechanisms of The Reassured Self program.

Further supporting the importance of self-compassion in adolescent mental health, Hammad et al [16] conducted a cross-sectional study in Saudi Arabia that explored the relationship between perceived stress and self-compassion among adolescents during the COVID-19 pandemic. Their findings revealed a significant negative correlation ($r=-0.460$, $P<.001$) between perceived stress and self-compassion levels, suggesting that higher self-compassion is associated with lower levels of stress. These results underscore the role of self-compassion as a protective factor for adolescent mental health, reinforcing the theoretical foundation of The Reassured Self program, which aims to reduce anxiety by fostering self-compassion and emotional regulation. CFT is rooted in evolutionary psychology and attachment theory, incorporating mindfulness through activities designed to foster self-compassion toward oneself and others, particularly in relation to self-criticism, self-blame, and shame [17,31,33]. Recommended techniques and exercises to develop self-compassion include self-talk and understanding the evolved nature of the human mind, along with the origins of self-criticism and self-blame [17]. Therefore, the core focus of The Reassured Self program is to introduce students to the concepts of self-blaming and self-reassurance, fostering a supportive inner dialogue and emotional regulation.

However, while The Reassured Self program demonstrated reductions in anxiety, our study did not directly measure

XSL•FO
RenderX

emotional regulation or self-compassion processes—key mechanisms observed in Petrocchi et al [33] and other self-compassion studies. Tools such as the Forms of Self-Criticizing/Attacking and Self-Reassuring Scale, developed by Gilbert et al [34] (the founder of CFT), provide validated measures for assessing self-criticism and self-reassurance. Incorporating these tools in future studies would enable a more comprehensive evaluation of how The Reassured Self program works. Specifically, measuring changes in emotional regulation, self-reassurance, and self-criticism could elucidate the mechanisms driving the reduction in anxiety and gaming addiction. As literature suggests, a decrease in self-criticism and self-blame is strongly associated with a reduction in anxiety [17,31,33]. Including these intermediate outcomes will help validate the program's effectiveness and provide deeper insights into its psychological processes. Overall, our findings suggest that culturally relevant, compassion-based programs can provide a viable solution for reducing anxiety in rural, underserved contexts. By addressing both the reduction of anxiety symptoms and fostering emotional regulation, such interventions hold promise as scalable and effective mental health solutions for adolescents. Also, our approach is supported by recent findings that advocate for brief, targeted interventions to effectively engage adolescents, particularly in an era where social media usage has impacted attention spans [35]. A recent meta-analysis of brief school mental health interventions found that programs lasting fewer than 4 sessions or 250 minutes resulted in significant, albeit small, improvements in students' mental health when assessed 1 month, 6 months, and 1 year post intervention [35].

Reduction in Video Game Addiction

The high baseline prevalence of video game addiction (14.6%, 18/123) among rural female adolescents in our study underscores the need for targeted mental health interventions in this area. This prevalence is close to global averages, which show approximately 12% of adolescents are at risk for problematic gaming, highlighting the specific challenges faced by this population [36]. However, the prevalence in our study is lower than that in the recent national report for the Madinah region (28%), and the overall national prevalence of 20.45% (897/4431) [37].

The program’s focus on self-compassion and emotional resilience appears to have positively impacted gaming behaviors, as evidenced by a significant decrease in video game addiction scores. Furthermore, the significant correlation between anxiety and gaming addiction underscores their interconnectedness, aligning with existing research that highlights gaming as a coping mechanism for anxiety [7].

This finding supports research suggesting that interventions emphasizing emotional regulation and coping skills such as exercise can reduce problematic gaming behaviors [38]. For example, a recent study investigating the relationship between internet gaming disorder, sleep quality, self-compassion, physical activity, and psychological distress among young adults aged 18-24 years found that internet gaming disorder was positively associated with interpersonal stress and depressive symptoms. Additionally, self-compassion and exercise behavior were found to mediate the relationship between depressive symptoms and sleep quality [39].

By providing alternative coping mechanisms, The Reassured Self may help reduce adolescents' reliance on gaming to manage emotional distress, addressing a critical mental health concern in this demographic. However, as discussed earlier, the lack of measurement of changes in emotional regulation, or adapting exercise behaviors, which could be assessed using validated tools, limits the conclusions that can be drawn about the program's overall efficacy.

Body Image Dissatisfaction

The notable prevalence of body image dissatisfaction, with 5.1% (6/117) of students showing a strong drive for thinness, reveals another area for potential intervention. Although the program’s impact on body image dissatisfaction was minimal and nonsignificant, the slight reduction in drive for thinness suggests that longer or more targeted interventions may more effectively address body image issues. Body image is deeply influenced by social and cultural factors, and this program provides a foundation for future interventions to build upon by incorporating more specific strategies to address body image concerns in rural contexts. While CFT has shown success in improving body image in other studies [30], the short duration of our program may have limited its impact on this complex issue. Additionally, this study found an average of 1 size discrepancy among the students, aligning with other research indicating that rural students may have low body image discrepancy [40,41]. Future iterations of the program may need to incorporate more targeted components to address body image issues specific to the cultural context.

Program Reaction

The integration of Islamic values and culturally sensitive practices, such as gratitude and self-compassion, likely enhanced the intervention's acceptability and effectiveness. Student feedback indicated high levels of satisfaction regarding the program's ease and clarity, perceived effectiveness, and content relevance. The concept of self-compassion, rooted in Islamic culture and religious practice, fosters self-reassurance and reduces self-blame [16]. Cultural and contextual considerations played a crucial role in the program’s design and implementation, which aligned with current recommendations for culturally adapting mental health interventions through self-compassion [17]. In addition, the program's success in a rural setting addresses a critical gap in mental health services for underserved populations and supports the need for global mental health interventions [42,43].

Limitations and Recommendations

This study’s findings have important implications for designing school-based mental health programs in rural or underserved areas. To ensure the effective delivery and sustainability of such interventions, several strategies should be prioritized: (1) training mental health professionals, educators, and community facilitators is critical for delivering the intervention effectively. (2) Partnerships with local organizations and schools can enhance resource availability and facilitate recruitment in resource-limited settings. (3) Leveraging community

XSL•FO
RenderX

partnerships: the collaboration demonstrated in this study—between a nonprofit organization, an academic institution, and a local school—provides a viable model for delivering mental health interventions in underserved areas. This approach aligns with broader recommendations to leverage community partnerships to enhance mental health support in schools [43,44]. (4) Advocacy efforts should focus on investments in culturally adapted, school-based mental health interventions to ensure they can reach underserved populations and address mental health disparities in context-specific and meaningful ways. (5) Future interventions should explore digital education methods, such as webinars, web-based workshops, and gaming-based mental health curricula, to enhance accessibility and engagement, particularly among rural students [45,46]. (6) Future research should investigate the intervention's long-term effects, including its impact on academic performance (eg, grades, school engagement) and social functioning (eg, relationships with peers, family members, and teachers). Understanding whether mental health improvements translate into better academic and social outcomes would provide a more holistic evaluation of the program's effectiveness. (7) Further exploration of specific cultural factors, such as family dynamics, social stigma, and rural-urban differences, is necessary to understand their contributions to the program's effectiveness. Adapting the intervention for broader cultural and socioeconomic contexts is critical for expanding its applicability.

However, several limitations should be considered when interpreting these results. The retrospective design and lack of a control group limit causal inferences. Additionally, the short follow-up period (2-3 weeks) may not capture long-term effects or potential relapses [17]. The specific demographic focus on rural Saudi female adolescents may also limit generalizability to other populations. To address these limitations and build on this study, future research should (1) incorporate randomized controlled trial designs with active control groups to establish causality; (2) extend follow-up periods to evaluate the sustainability of intervention benefits over time; (3) compare The Reassured Self program with other evidence-based approaches, such as cognitive-behavioral therapy and mindfulness-based stress reduction, to assess relative effectiveness and adaptability; (4) examine the specific components of the program and their relative contributions to outcomes to refine and optimize the intervention; and (5) investigate the interplay between cultural factors and program outcomes to ensure effective adaptations for different cultural and socioeconomic contexts.

Conclusions

Adolescents in rural Saudi Arabia face significant mental health challenges, with high rates of anxiety, video game addiction, and, to a lesser extent, body image dissatisfaction. This study's findings underscore the potential of a culturally adapted CFT program, The Reassured Self, to address these issues. Collaboratively implemented by a non-profit organization and an academic institution, the program successfully reduced anxiety symptoms and problematic gaming behaviors among rural female adolescents, providing a promising model for school-based mental health interventions in underserved areas.

The integration of Islamic values and culturally sensitive practices, such as gratitude and self-compassion, likely enhanced the program's acceptability and relevance, as reflected in the high levels of student satisfaction with the intervention's content and delivery. The Reassured Self demonstrates the feasibility and effectiveness of a culturally sensitive, compassion-focused intervention for adolescent mental health in rural Saudi Arabia. The promising findings of The Reassured Self program highlight its potential for widespread implementation in schools and communities, particularly in underserved areas. By addressing the unique challenges faced by this underserved population, the program contributes valuable insights for global mental health initiatives aimed at reducing disparities in mental health access and support. Future efforts to refine and scale the intervention can inform broader applications in diverse cultural and socioeconomic contexts.

Acknowledgments

The author would like to express sincere gratitude to the Ministry of Education School Health Department in Medina and the participating school for their invaluable support.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Checklist for Reporting Results of Retrospective Database Studies from the ISPOR Task Force.
[PDF File (Adobe PDF File), 54 KB - formative_v9i1e68840_app1.pdf]

Multimedia Appendix 2
Data Completeness for Intervention Participants.
[PDF File (Adobe PDF File), 36 KB - formative_v9i1e68840_app2.pdf]

XSL•FO
RenderX

References

1. Racine N, McArthur BA, Cooke JE, Eirich R, Zhu J, Madigan S. Global prevalence of depressive and anxiety symptoms in children and adolescents during COVID-19: a meta-analysis. JAMA Pediatr 2021;175(11):1142-1150 [FREE Full text] [doi: 10.1001/jamapediatrics.2021.2482] [Medline: 34369987]
2. Triplett NS, Woodard GS, Johnson C, Nguyen JK, AlRasheed R, Song F, et al. Stakeholder engagement to inform evidence-based treatment implementation for children's mental health: a scoping review. Implement Sci Commun 2022;3(1):82. [doi: 10.1186/s43058-022-00327-w] [Medline: 35906675]
3. Rodrigue C, Rodgers RF, Carbonneau N, Bégin C, Dion J. COVID-19-related distress, body image, and eating behaviors: a cross-sectional explanatory model. BMC Psychol 2024;12(1):117 [FREE Full text] [doi: 10.1186/s40359-024-01613-z] [Medline: 38431696]
4. Kessler RC, Berglund P, Demler O, Jin R, Merikangas KR, Walters EE. Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the national comorbidity survey replication. Arch Gen Psychiatry 2005;62(6):593-602. [doi: 10.1001/archpsyc.62.6.593] [Medline: 15939837]
5. Firoozjah MH, Shahrbanian S, Homayouni A, Hower H. Comparison of eating disorders symptoms and body image between individual and team sport adolescent athletes during the COVID-19 pandemic. J Eat Disord 2022;10(1):119 [FREE Full text] [doi: 10.1186/s40337-022-00644-4] [Medline: 35962428]
6. Han TS, Cho H, Sung D, Park MH. A systematic review of the impact of COVID-19 on the game addiction of children and adolescents. Front Psychiatry 2022;13:976601 [FREE Full text] [doi: 10.3389/fpsyt.2022.976601] [Medline: 36061296]
7. Kaya A, Türk N, Batmaz H, Griffiths MD. Online gaming addiction and basic psychological needs among adolescents: the mediating roles of meaning in life and responsibility. Int J Ment Health Addict 2023:1-25 [FREE Full text] [doi: 10.1007/s11469-022-00994-9] [Medline: 36643385]
8. Bumozah HS, Al-Quwaidhi AJ, Al-Ghadeeb R. Prevalence and risk factors of internet gaming disorder among female secondary school students in Al-Ahsa, Kingdom of Saudi Arabia. Cureus 2023;15(6):e40375 [FREE Full text] [doi: 10.7759/cureus.40375] [Medline: 37456410]
9. Alghamdi MH, Alghamdi MM. Prevalence of internet gaming disorder among intermediate and high school students in Albaha, Saudi Arabia: a cross-sectional study. Cureus 2023;15(4):e37115 [FREE Full text] [doi: 10.7759/cureus.37115] [Medline: 37153233]
10. Altwaijri Y, Kazdin AE, Al-Subaie A, Al-Habeeb A, Hyder S, Bilal L, et al. Lifetime prevalence and treatment of mental disorders in Saudi youth and adolescents. Sci Rep 2023;13(1):6186 [FREE Full text] [doi: 10.1038/s41598-023-33005-5] [Medline: 37061556]
11. Alhabeeb AA, Al-Duraihem RA, Alasmary S, Alkhamaali Z, Althumiri NA, BinDhim NF. National screening for anxiety and depression in Saudi Arabia 2022. Front Public Health 2023;11:1213851 [FREE Full text] [doi: 10.3389/fpubh.2023.1213851] [Medline: 37441650]
12. Basurrah AA, Di Blasi Z, Lambert L, Murphy M, Warren MA, Setti A, et al. The effects of positive psychology interventions in Arab countries: a systematic review. Appl Psychol Health Well Being 2023;15(2):803-821. [doi: 10.1111/aphw.12391] [Medline: 35856920]
13. Kotera Y, Kirkman A, Beaumont J, Komorowska MA, Such E, Kaneda Y, et al. Self-compassion during COVID-19 in non-WEIRD countries: a narrative review. Healthcare (Basel) 2023;11(14):2016 [FREE Full text] [doi: 10.3390/healthcare11142016] [Medline: 37510457]
14. Mariee AA, Abd Elhamid EA, Al-Dhmashi ASF, Alshahrani FF, El-Said HDA. Integrating of mental health interventions in school health in Saudi Arabia: policy brief. Int J Nov Res Healthc Nurs 2024;11(1):303-306. [doi: 10.5281/zenodo.10868098]
15. Chatterjee K, Srivastava K, Prakash J, Dangi A. Positive mental health for all serving the under-served. Ind Psychiatry J 2023;32(2):234-239 [FREE Full text] [doi: 10.4103/ipj.ipj_132_22] [Medline: 38161446]
16. Hammad SS, Alzhrani MD, Almulla HA. Adolescents' perceived stress of COVID-19 and self-compassion in Saudi Arabia: a cross-sectional study. Int J Nurs Sci 2023;10(2):215-220 [FREE Full text] [doi: 10.1016/j.ijnss.2023.03.008] [Medline: 37095851]
17. Neff KD. Self-compassion: theory, method, research, and intervention. Annu Rev Psychol 2023;74(1):193-218 [FREE Full text] [doi: 10.1146/annurev-psych-032420-031047] [Medline: 35961039]
18. Motheral B, Brooks J, Clark MA, Crown WH, Davey P, Hutchins D, et al. A checklist for retrospective database studies--report of the ISPOR task force on retrospective databases. Value Health 2003;6(2):90-97 [FREE Full text] [doi: 10.1046/j.1524-4733.2003.00242.x] [Medline: 12641858]
19. Ali N, Snell T. Spence children's anxiety scale ? Arabic. 1994. URL: https://www.scaswebsite.com/arabic/ [accessed 2025-02-05]
20. Spence SH. Structure of anxiety symptoms among children: a confirmatory factor-analytic study. J Abnorm Psychol 1997;106(2):280-297. [doi: 10.1037/0021-843x.106.2.280]
21. Galán-Luque T, Serrano-Ortiz M, Orgilés M. Factor structure and psychometric properties of the spence children's anxiety scale: a 25-Year systematic review. Child Psychiatry Hum Dev 2023. [doi: 10.1007/s10578-023-01566-1] [Medline: 37490107]

XSL•FO
RenderX

22. Spence SH. Using the SCAS-Child T-Scores. URL: https://www.scaswebsite.com/portfolio/scas-child-t-scores/ [accessed 2025-02-05]
23. Lemmens JS, Valkenburg PM, Gentile DA. The internet gaming disorder scale. Psychol Assess 2015;27(2):567-582. [doi: 10.1037/pas0000062] [Medline: 25558970]
24. Asaad T, Morsy KE, Hasan H, El Din HM, El Meguid MA. Game addiction scale for adolescents: Arabic version validation. Addict Disord Their Treat 2019;18(6):223-228. [doi: 10.1097/adt.0000000000000167]
25. Brunborg GS, Hanss D, Mentzoni RA, Pallesen S. Core and peripheral criteria of video game addiction in the game addiction scale for adolescents. Cyberpsychol Behav Soc Netw 2015;18(5):280-285 [FREE Full text] [doi: 10.1089/cyber.2014.0509] [Medline: 25826043]
26. Simeon DT, Rattan RD, Panchoo K, Kungeesingh KV, Ali AC, Abdool PS. Body image of adolescents in a multi-ethnic caribbean population. Eur J Clin Nutr 2003;57(1):157-162. [doi: 10.1038/sj.ejcn.1601515] [Medline: 12548311]
27. Latiff AA, Muhamad J, Rahman RA. Body image dissatisfaction and its determinants among young primary-school adolescents. J Taibah Univ Med Sci 2018;13(1):34-41 [FREE Full text] [doi: 10.1016/j.jtumed.2017.07.003] [Medline: 31435300]
28. Kirkpatrick D, Kirkpatrick J. Evaluating Training Programs: The Four Levels. San Francisco: Berrett-Koehler Publishers; 2006.
29. Maratos FA, Wood W, Cahill R, Tronco Hernández YAT, Matos M, Gilbert P. A mixed-methods study of compassionate mind training for pupils (CMT-Pupils) as a school-based wellbeing intervention. Mindfulness 2024;15(2):459-478. [doi: 10.1007/s12671-024-02303-y]
30. Turk F, Waller G. Is self-compassion relevant to the pathology and treatment of eating and body image concerns? A systematic review and meta-analysis. Clin Psychol Rev 2020;79:101856. [doi: 10.1016/j.cpr.2020.101856] [Medline: 32438284]
31. Vidal J, Soldevilla JM. Effect of compassion-focused therapy on self-criticism and self-soothing: a meta-analysis. Br J Clin Psychol 2023;62(1):70-81 [FREE Full text] [doi: 10.1111/bjc.12394] [Medline: 36172899]
32. Tali E, Potharst ES, de Bruin EI, Utens EMWJ. Self-compassion and anxiety in adolescents with and without anxiety disorder. Children (Basel) 2023;10(7):1181 [FREE Full text] [doi: 10.3390/children10071181] [Medline: 37508678]
33. Petrocchi N, Ottaviani C, Cheli S, Matos M, Baldi B, Basran JK, et al. The impact of compassion-focused therapy on positive and negative mental health outcomes: results of a series of meta-analyses. Clin Psychol Sci Pract 2024;31(2):230-247. [doi: 10.1037/cps0000193]
34. Gilbert P, Clarke M, Hempel S, Miles J, Irons C. Criticizing and reassuring oneself: an exploration of forms, styles and reasons in female students. Br J Clin Psychol 2004;43(Pt 1):31-50. [doi: 10.1348/014466504772812959] [Medline: 15005905]
35. Cohen KA, Ito S, Ahuvia IL, Yang Y, Zhang Y, Renshaw TL, et al. Brief school-based interventions targeting student mental health or well-being: a systematic review and meta-analysis. Clin Child Fam Psychol Rev 2024;27(3):732-806. [doi: 10.1007/s10567-024-00487-2] [Medline: 38884838]
36. Boniel-Nissim M, Tynjälä J, Gobiņa I, Furstova J, van den Eijnden RJJM, Marino C, et al. Adolescent use of social media and associations with sleep patterns across 18 European and North American countries. Sleep Health 2023;9(3):314-321 [FREE Full text] [doi: 10.1016/j.sleh.2023.01.005] [Medline: 36804326]
37. Alfaleh A, Alzaher A, Alkattan A, Alabdulkareem K, Ibrahim MH. Prevalence of video gaming disorder in Saudi Arabia: a school-based national study. J Egypt Public Health Assoc 2024;99(1):20 [FREE Full text] [doi: 10.1186/s42506-024-00165-9] [Medline: 39164516]
38. King DL, Delfabbro PH, Perales JC, Deleuze J, Király O, Krossbakken E, et al. Maladaptive player-game relationships in problematic gaming and gaming disorder: a systematic review. Clin Psychol Rev 2019;73:101777. [doi: 10.1016/j.cpr.2019.101777] [Medline: 31707185]
39. Wong MYC, Yuan GF, Liu C, Lam SKK, Fung HW. The relationship between internet gaming disorder, sleeping quality, self-compassion, physical activity participation and psychological distress: a path analysis. Glob Ment Health (Camb) 2024;11:e67. [doi: 10.1017/gmh.2024.36] [Medline: 39220110]
40. Błaszczyk-Bębenek E, Jagielski P, Nowak R, Ostachowska- Gąsior A. Dissatisfaction and discrepancy in body-image perceptions among adolescents aged 10–17 years from rural areas of southern poland – a cross-sectional study. Ann Agric Environ Med 2024. [doi: 10.26444/aaem/193461]
41. Paria B, Dwivedi M, Roy SK. A cross-sectional study on body image perception and self-esteem among adolescent girls in urban and rural Areas of Kolkata, West Bengal, India. J Clin Diagn Res 2023:LC11-LC14. [doi: 10.7860/jcdr/2023/62667.18697]
42. Barnett ML, Gonzalez A, Miranda J, Chavira DA, Lau AS. Mobilizing community health workers to address mental health disparities for underserved populations: a systematic review. Adm Policy Ment Health 2018;45(2):195-211 [FREE Full text] [doi: 10.1007/s10488-017-0815-0] [Medline: 28730278]
43. Mongelli F, Georgakopoulos P, Pato MT. Challenges and opportunities to meet the mental health needs of underserved and disenfranchised populations in the United States. Focus (Am Psychiatr Publ) 2020;18(1):16-24 [FREE Full text] [doi: 10.1176/appi.focus.20190028] [Medline: 32047393]

XSL•FO
RenderX

44. Castillo EG, Ijadi-Maghsoodi R, Shadravan S, Moore E, Mensah MO, Docherty M, et al. Community interventions to promote mental health and social equity. Curr Psychiatry Rep 2019;21(5):35 [FREE Full text] [doi: 10.1007/s11920-019-1017-0] [Medline: 30927093]
45. van Lotringen C, Lusi B, Westerhof GJ, Ludden GDS, Kip H, Kelders SM, et al. The role of compassionate technology in blended and digital mental health interventions: systematic scoping review. JMIR Ment Health 2023;10:e42403 [FREE Full text] [doi: 10.2196/42403] [Medline: 37027207]
46. Jenson C, Wolff SF, Milkovich LM. Effects of a preventive mental health curriculum embedded into a scholarly gaming course on adolescent self-esteem: prospective matched pairs experiment. JMIR Serious Games 2023;11:e48401 [FREE Full text] [doi: 10.2196/48401] [Medline: 38059568]

Abbreviations

CFT: compassion-focused therapy
CMT-Pupils: Compassionate Mind Training
DSM-5: Diagnostic and Statistical Manual of Mental Disorders (Fifth Edition)
GAS: Game Addiction Scale
SCAS: Spence Children's Anxiety Scale

Edited by A Mavragani; submitted 15.11.24; peer-reviewed by R Ghammem, A Alenezi; comments to author 12.12.24; revised version received 06.01.25; accepted 28.01.25; published 19.02.25.

Please cite as:
Zuair A
Effect of the Reassured Self-Compassion–Based School Program on Anxiety, Video Game Addiction, and Body Image Among Rural Female Adolescents: Retrospective Study
JMIR Form Res 2025;9:e68840
URL: https://formative.jmir.org/2025/1/e68840
doi:10.2196/68840
PMID:

©Areeg Zuair. Originally published in JMIR Formative Research (https://formative.jmir.org), 19.02.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Machine Learning–Based Prediction of Delirium and Risk Factor Identification in Intensive Care Unit Patients With Burns: Retrospective Observational Study

Ryo Esumi[1*], MD, PhD; Hiroki Funao[2*], MSc; Eiji Kawamoto[1], MD, PhD; Ryota Sakamoto[3], PhD; Asami Ito-Masui[1], MD, PhD; Fumito Okuno[4], MD; Toru Shinkai[4], MD, PhD; Atsuya Hane[4], MD; Kaoru Ikejiri[4], MD; Yuichi Akama[1], MD, PhD; Arong Gaowa[1], PhD; Eun Jeong Park[1], PhD; Ryo Momosaki[5], MD, PhD; Ryuji Kaku[6], MD, PhD; Motomu Shimaoka[1], MD, PhD

[1]Department of Molecular Pathobiology and Cell Adhesion Biology, Mie University Graduate School of Medicine, Mie University, Tsu, Japan
[2]Department of Practical Nursing, Mie University Graduate School of Medicine, Tsu city, Japan
[3]Department of Medical Informatics, Mie University Hospital, Tsu city, Japan
[4]Department of Emergency and Critical Care Center, Mie University Hospital, Tsu city, Japan
[5]Department of Rehabilitation Medicine, Mie University Graduate School of Medicine, Tsu city, Japan
[6]Department of Anesthesiology, Mie University Hospital, Tsu city, Japan
[*]these authors contributed equally

Corresponding Author:
Eiji Kawamoto, MD, PhD
Department of Molecular Pathobiology and Cell Adhesion Biology, Mie University Graduate School of Medicine
Mie University
Edobashi 2-174
Tsu, 5140001
Japan
Phone: 81 0592321111
Email: a_2.uk@mac.com

Abstract

Background: The incidence of delirium in patients with burns receiving treatment in the intensive care unit (ICU) is high, reaching up to 77%, and has been associated with increased mortality rates. Therefore, early identification of patients at high risk of delirium onset is essential for improving treatment strategies.

Objective: This study aimed to create a machine learning model for predicting delirium in patients with burns during their ICU stay using patient data from the first day of ICU admission and identify predictive factors for ICU delirium in patients with burns.

Methods: This study focused on 82 patients with burns aged ≥18 years who were admitted to the ICU at Mie University Hospital for ≥24 hours between January 2015 and June 2023. In total, 70 variables were measured in patients upon ICU admission and used as explanatory variables in the ICU delirium prediction model. Delirium was assessed using the Intensive Care Delirium Screening Checklist every 8 hours after ICU admission. A total of 10 different machine learning methods were used to predict ICU delirium. Multiple receiver operating characteristic curves were plotted for various machine learning models, and the area under the curve (AUC) for each was compared. In addition, the top 15 risk factors contributing to delirium onset were identified using Shapley additive explanations analysis.

Results: Among the 10 machine learning models tested, logistic regression (mean AUC 0.906, SD 0.073), support vector machine (mean AUC 0.897, SD 0.056), k-nearest neighbor (mean AUC 0.894, SD 0.060), neural network (mean AUC 0.857, SD 0.058), random forest (mean AUC 0.850, SD 0.074), adaptive boosting (mean AUC 0.832, SD 0.094), gradient boosting machine (mean AUC 0.821, SD 0.074), and naïve Bayes (mean AUC 0.827, SD 0.095) demonstrated the highest accuracy in predicting ICU delirium. Specifically, 24-hour urine output (from ICU admission to 24 hours), oxygen saturation, burn area, total bilirubin level, and intubation upon ICU admission were identified as the major risk factors for delirium onset. In addition, variables, such as the proportion of white blood cell fractions, including monocytes; methemoglobin concentration; and respiratory rate, were identified as important risk factors for ICU delirium.

XSL•FO
RenderX

Conclusions: This study demonstrated the ability of machine learning models trained using vital signs and blood data upon ICU admission to predict delirium in patients with burns during their ICU stay.

(JMIR Form Res 2025;9:e65190) doi:10.2196/65190

KEYWORDS

burns; delirium; intensive care unit; machine learning; prediction model; artificial intelligence; AI

Introduction

Background

Delirium is a significant complication in patients in the intensive care unit (ICU) and is recognized as an urgent medical need requiring treatment and prevention. Delirium is defined as acute brain dysfunction associated with underlying conditions characterized by fluctuating bouts of impaired consciousness, attention, and cognition. This condition is frequently observed in patients admitted to the ICU, with delirium occurring in 10% to 50% of patients [1]. Delirium is an independent predictor of poor outcomes, and there is currently no established specific treatment, making early diagnosis and prevention critically important [2]. In particular, the incidence of delirium in patients with burns can reach 77% [3], with reports indicating that 30% of patients who develop delirium respond effectively to prevention and treatment [4]. These facts underscore the importance of identifying patients at high risk of delirium and implementing preventive measures.

In recent years, research on ICU delirium prediction using artificial intelligence technology has advanced, with a particular focus on the application of various machine learning algorithms. These algorithms, such as random forest (RF), support vector machine (SVM), and gradient boosting, can be used to develop predictive models for ICU delirium [5]. While these algorithms themselves do not identify relevant features, they can be combined with feature importance analysis techniques such as Shapley additive explanations (SHAP) to determine which variables contribute most significantly to the predictions. This approach allows for the analysis of large volumes of patient data with speed and accuracy beyond human capacity. However, research on the prediction of ICU delirium in patients with burns remains underdeveloped, and very few studies have been conducted in this field.

Objectives

This study aimed to demonstrate whether it is possible to predict ICU delirium in patients with burns using machine learning. Specifically, we hypothesized that a machine learning model using clinical data such as vital signs and blood test results could predict delirium in ICU patients. The null hypothesis was that these models would not be superior to random chance in predicting ICU delirium.

While ICU delirium prediction potential for multiple machine learning models, as well as the model with the highest accuracy, was assessed, the second objective of this study was to identify risk factors for ICU delirium in patients with burns and contribute to the development of more effective prevention and treatment strategies.

This ICU delirium prediction approach using machine learning has the potential to support the early detection of ICU delirium in patients with burns and, ultimately, improve patient outcomes.

Methods

Patient Demographics and Data Collection

This was a retrospective observational study focused on predicting delirium in patients with burns admitted to the ICU. This study included 82 patients with burns aged ≥18 years who were admitted to the Mie University Hospital ICU for ≥24 hours between January 2015 and June 2023. The sample size of 82 patients was determined by including all patients with burns who were admitted to the ICU within the study period.

Patients were retrospectively included based on the inclusion criteria, which ensured the comprehensive capture of all eligible cases during the study period. This approach minimized selection bias and allowed for a representative sample of the population of patients with burns in our ICU.

Physiological, biochemical, and clinical data collected from these patients upon ICU admission were used to extract 70 explanatory variables. This study aimed to develop a model using these 70 variables to predict delirium onset during ICU admission, assess its accuracy, and identify the risk factors contributing to each model.

The Definition, Diagnostic Criteria, and Standard Assessment of ICU Delirium

ICU delirium, also known as ICU psychosis, is an acute, fluctuating change in consciousness and cognition that occurs frequently in patients who are critically ill. ICU delirium is characterized by disturbances in attention, awareness, and cognitive function. These disturbances are often temporary and reversible but can lead to prolonged ICU stays, increased morbidity and mortality, and long-term cognitive impairments if not properly managed.

The standard assessment of ICU delirium involves the use of validated diagnostic tools to ensure accurate detection and timely intervention. In total, 2 widely recognized tools are the Confusion Assessment Method for the Intensive Care Unit (CAM-ICU) and the Intensive Care Delirium Screening Checklist (ICDSC).

The CAM-ICU is a structured diagnostic tool specifically designed for use in the ICU setting. It is based on the Confusion Assessment Method and has been modified for the critical care environment. The CAM-ICU assesses four key features: (1) acute onset of mental status changes or a fluctuating course, (2) inattention, (3) disorganized thinking, and (4) altered level of consciousness. A positive CAM-ICU diagnosis requires the

XSL•FO
RenderX

presence of both features 1 and 2 and either feature 3 or 4. This tool is favored for its ease of use and quick administration, making it suitable for frequent assessments.

The ICDSC is another tool used to screen for delirium in ICU patients, consisting of an 8-item checklist that assesses various cognitive and behavioral symptoms associated with delirium. The items are (1) altered level of consciousness, (2) inattention, (3) disorientation, (4) hallucinations or delusions, (5) psychomotor agitation or retardation, (6) inappropriate speech or mood, (7) sleep and wake cycle disturbances, and (8) symptom fluctuation. Each item is scored based on its presence within the previous 24 hours, and a total score of ≥4 indicates the presence of delirium. The ICDSC provides a comprehensive overview of the patient's condition over a longer period than the CAM-ICU.

In this study, delirium was assessed using the internationally recognized ICDSC every 8 hours after ICU admission. Delirium was diagnosed when the ICDSC score was ≥4 points. We opted for a binary classification approach because this method simplifies the model's output to either *delirium present* or *delirium absent*, facilitating smoother decision-making in clinical settings. Another reason for this choice was the diversity in patient conditions in real-world clinical environments, where a few specific atypical cases can skew the predictions of regression models. Binary classification is less susceptible to the influence of such outliers, enabling the development of a more robust model. This method allowed for accurate determination of the presence and severity of delirium.

Development of Machine Learning Models

In total, 10 different algorithms were used to develop the machine learning models: logistic regression (LR), RF, SVM, neural network, k-nearest neighbor (KNN), decision tree, naïve Bayes, adaptive boosting (AdaBoost), gradient-boosting machine (GBM), and linear discriminant analysis (LDA). These models were selected based on the area under the receiver operating characteristic curve (AUC) to compare the accuracy of delirium prediction.

Ethical Considerations

This study was approved by the ethics committee of Mie University (H2020-164), ensuring compliance with ethical guidelines for clinical research. This study underwent ethical review and was approved as a clinical research study. For informed consent, we applied the opt-out method, allowing participants to refuse the inclusion of their data in this study. If participants chose to opt out, their data were excluded from the analysis. To protect privacy and confidentiality, all study data were fully anonymized. Data extracted from electronic medical records were deidentified, ensuring that individual patients could not be identified. No compensation was provided to participants. As the study used an opt-out approach, participants had the right to refuse the inclusion of their data without any consequences.

Feature Importance Analysis Using SHAP

In this study, we used SHAP to interpret the feature importance of our machine learning models. SHAP provides a unified measure of each variable's contribution to the prediction outcome. By applying SHAP, we identified the top 15 predictors of ICU delirium across all models, which allowed for a detailed understanding of the factors most associated with delirium onset. The SHA*P* values helped clarify the influence of variables such as 24-hour urine output, total bilirubin (T-bil) levels, and respiratory rate on the predictive performance of the models. SHAP analysis also enabled clinicians to visualize and interpret the risk factors, enhancing the practical applicability of the model's predictions.

Data Preprocessing for Machine Learning

Handling Missing Values

In this study, missing values were not handled. This approach ensured that all features had complete data, thus preserving the dataset's integrity without excluding any records.

Data Splitting

The dataset was loaded from a CSV file using the *Pandas* library. It was then split into explanatory variables and the target variable, where the target variable was delirium. The data were divided into 2 subsets: 80% was allocated for training the model, and 20% was reserved for internal validation. To evaluate the model's performance, the dataset was further split into folds using stratified k-fold cross-validation, which ensures that the proportion of classes remains consistent across folds.

Data Standardization

The explanatory variables in both the training and test sets were standardized using *StandardScaler* from the *sklearn* library to ensure that each feature had a mean of 0 and an SD of 1. This preprocessing step is crucial for models sensitive to feature scaling.

Model Selection and Initialization

Several machine learning models were initialized for this study, including SVM, neural network, KNN, decision tree, naïve Bayes, AdaBoost, GBM, LDA, LR, and RF. Each model was imported from the *sklearn* or *SciPy* library and initialized with appropriate default parameters.

Hyperparameter Tuning

For specific models, such as SVM, neural network, KNN, decision tree, and RF, hyperparameter tuning was performed using grid search with cross-validation. The *GridSearchCV* class from the *sklearn* library was used to search for the best combination of hyperparameters, optimizing the model's performance based on the AUC metric.

Model Training and Evaluation

Each model was trained using stratified k-fold cross-validation to ensure robust performance evaluation. For models with hyperparameter grids defined, grid search was applied to find the best hyperparameter set. The best-performing model from the grid search was then used to fit the data in each fold. The model's performance was evaluated using the AUC score on the test subset within each fold.

XSL•FO
RenderX

Statistical Analysis

Continuous variables are reported as medians with IQRs and were compared using the Mann-Whitney *U* test or the Kruskal-Wallis test, depending on the number of groups. Categorical variables are expressed as counts and percentages, with comparisons made using the Fisher exact test or the chi-square test, as appropriate. All statistical analyses were conducted using the SPSS software (version 21; IBM Corp), and statistical significance was defined as $P<.05$.

To evaluate differences in model performance, pairwise 2-tailed *t* tests were applied to the AUC scores using the *ttest_ind* function from the *scipy.stats* module. In addition, a 1-way ANOVA was conducted using the *f_oneway* function to identify statistically significant differences across models.

Model performance results were visualized through bar charts displaying the mean AUC scores along with their SDs for each model. These visualizations were created using the *Matplotlib* library. A heat map was also generated to illustrate pairwise *P* values from the *t* tests, offering a clear representation of the statistical significance of the differences between models. The ANOVA *P* value was similarly calculated and displayed.

Results

Comparative Analysis of ICU Delirium in Patients With Burns

Compared to patients without delirium (Table 1), patients with burns with ICU delirium were older (age: median 77.0, IQR 69.5-84.5 years vs median 60.5, IQR 37.5-73.0 years; $P<.001$), were more likely to have airway burns (12/32, 38% vs 8/50, 16%; $P=.03$), and experienced longer ICU stays (median 2.5, IQR 1.8-11.2 days vs 2.0, IQR 1.0-3.0 days; $P=.009$). The mortality rate was significantly higher in the delirium group (7/32, 22% vs 0%; $P=.001$), and these patients more frequently required intubation (20/32, 62% vs 4/50, 8%; $P<.001$). The delirium group also had greater burn areas (median 16%, IQR 9.75%-34% vs 8%, IQR 4.6%-17%; $P=.007$) and burn indexes (median 9.5, IQR 4.7-25.1 vs median 4.0, IQR 1.0-10.0; $P=.002$).

XSL•FO
RenderX

Table 1. Characteristics of patients with burns with intensive care unit (ICU) delirium—comparison of explanatory factors between patients with and without deliriuma.

	Burns with delirium (n=32)	Burns without delirium (n=50)	*P* value
Age (y), median (IQR)	77.0 (69.5-84.5)	60.5 (37.5-73.0)	*<.001* [b]
Sex (male), n (%)	20 (63)	32 (64)	.90
Height (cm), median (IQR)	160.5 (150.0-170.0)	162.2 (158.0-168.0)	.42
Weight (kg), median (IQR)	54.4 (44.4-63.3)	60.2 (50.5-67.1)	.07
BMI (kg/m^2), median (IQR)	21.2 (19.0-23.4)	23.1 (19.7-24.9)	.06
Airway burns, n (%)	12 (38)	8 (16)	*.03* [b]
Length of ICU stay (days), median (IQR)	2.5 (1.8-11.2)	2.0 (1.0-3.0)	*.009* [b]
Mortality rate, n (%)	7 (22)	0 (0)	*.001* [b]
Intubation, n (%)	20 (62)	4 (8)	*<.001* [b]
Burn area (%), median (IQR)	16 (9.75-34)	8 (4.6-17)	*.007* [b]
Burn index, median (IQR)	9.5 (4.7-25.1)	4.0 (1.0-10.0)	*.002* [b]
WBC[c] (10^3 per µL), median (IQR)	11.9 (9.0-18.5)	8.8 (6.5-11.3)	*.006* [b]
RBC[d] (10^6 per µL), median (IQR)	4.3 (3.7-5.3)	4.4 (4.0-4.9)	.51
Hemoglobin (g/dL), median (IQR)	13.5 (11.9-15.9)	14.6 (14.0-15.9)	*.04* [b]
Hematocrits (%), median (IQR)	40 (35-44)	43 (41-46.7)	*.009* [b]
MCV[e] (fL), median (IQR)	92.0 (89.9-93.8)	92.3 (87.7-95.6)	.95
MCH[f] (pg), median (IQR)	30.8 (30.0-32.4)	31.4 (29.7-32.9)	.71
MCHC[g] (%), median (IQR)	33.8 (33.2-34.6)	33.9 (33-34.6)	.77
Platelet count (10^3 per µL), median (IQR)	205.5 (165.0-314.8)	249.5 (202.2-302.0)	.50
Neutrophils (%), median (IQR)	77 (73.9-83.7)	74.4 (64.6-80.9)	*.04* [b]
Lymphocytes (%), median (IQR)	15.8 (8.4-17.9)	16.8 (12.6-23.6)	*.02* [b]
Monocytes (%), median (IQR)	6.6 (5.2-7.7)	6.6 (5.7-7.2)	.73
Eosinophils (%), median (IQR)	0.6 (0.4-1.4)	1.2 (0.6-1.9)	*.04* [b]
Basophils (%), median (IQR)	0.4 (0.2-0.5)	0.4 (0.2-0.5)	.84
Neutrophil count (10^3 per µL), median (IQR)	8985 (6265-10,830)	6840.0 (4449.2-9029.0)	*.03* [b]
Lymphocyte count (10^3 per µL), median (IQR)	1550 (1130-1943)	1550.0 (1178.2-2145.0)	.55
Monocyte count (10^3 per µL), median (IQR)	767 (478-993)	575.0 (447.5-767.0)	*.04* [b]
Eosinophil count (10^3 per µL), median (IQR)	70.0 (37.5-111.0)	90.0 (40.0-167.7)	.22
Basophil count (10^3 per µL), median (IQR)	38.5 (30.0-62.5)	37.0 (20.0-50.0)	.13
APTT[h] (seconds), median (IQR)	30.8 (25.9-36.8)	27.4 (24.9-29.9)	*.02* [b]
PT[i] (seconds), median (IQR)	12.2 (11.5-13.6)	11.3 (10.8-12.1)	*.001* [b]
PT (%)[j], median (IQR)	94.8 (77-104.1)	105.4 (97-116.9)	*.001* [b]
PT-INR[k], median (IQR)	1.0 (0.9-1.1)	0.9 (0.9-1.0)	*.002* [b]
Fibrinogen (mg/dL), median (IQR)	299.0 (251.2-374.2)	272.5 (225.2-327.7)	.20
D-dimer (µg/mL), median (IQR)	4.0 (1.5-8.5)	1.0 (0.2-8.2)	*.02* [b]
pH, median (IQR)	7.3 (7.3-7.4)	7.4 (7.3-7.4)	*.04* [b]

XSL•FO
RenderX

	Burns with delirium (n=32)	Burns without delirium (n=50)	*P* value
PCO_2[l] (mm Hg), median (IQR)	40.0 (31.7-45.0)	36.5 (33.2-39.7)	.28
PO_2 (mm Hg), median (IQR)	152.5 (93.5-322.7)	123.0 (83.2-166.0)	*.03* [b]
SO_2[m] (%), median (IQR)	99.1 (98.3-99.6)	97.6 (94.9-99)	*.01* [b]
HCO_3^-[n] (mmol/L), median (IQR)	23.1 (18.9-25.6)	23.1 (20.5-24.4)	.76
Anion gap (mmol/L), median (IQR)	10.4 (4.4-17.9)	13.1 (10.4-17.2)	.24
O_2-Hb[o] (%), median (IQR)	96.0 (92.5-97.3)	95.9 (90.8-97.1)	.59
CO-Hb[p] (%), median (IQR)	1.6 (1.1-3.4)	1.7 (0.7-3.4)	.75
MetHb[q] (%), median (IQR)	0.9 (0.5-1.1)	0.4 (0.3-0.9)	*.004* [b]
Ionized calcium (mmol/L), median (IQR)	1.1 (1.0-1.1)	1.1 (1.0-1.1)	.60
Lactate (mmol/L), median (IQR)	3.0 (1.5-4.9)	2.3 (1.8-3.3)	.45
TP[r] (g/dL), median (IQR)	6.4 (5.9-6.7)	7.0 (6.2-7.5)	*.006* [b]
Albumin (g/dL), median (IQR)	3.6 (3.1-3.9)	4.1 (3.2-4.4)	*.004* [b]
BUN[s] (mg/dL), median (IQR)	16.9 (12.8-22.7)	13.4 (11.0-19.0)	.01
Creatinine (mg/dL), median (IQR)	0.8 (0.6-1.0)	0.7 (0.5-0.8)	*.03* [b]
eGFR[t] (mL per minute per 1.73 m^2), median (IQR)	67.2 (47.8-87.2)	83.7 (59.9-101.0)	*.02* [b]
Sodium (mmol/L), median (IQR)	139.0 (138.0-141.0)	140.0 (138.0-141.0)	>.99
Potassium (mmol/L), median (IQR)	4.0 (3.7-4.5)	3.9 (3.7-4.4)	.58
CL[u] (mmol/L), median (IQR)	105.0 (102.0-107.0)	104.0 (103.0-106.7)	.57
Calcium (mg/dL), median (IQR)	8.5 (8.2-8.9)	8.9 (8.4-9.2)	*.009* [b]
AST[v] (U/L), median (IQR)	39.5 (22.7-60.5)	30.5 (24.2-38.0)	.30
ALT[w] (U/L), median (IQR)	23.0 (14.0-43.5)	20.5 (14.5-29.0)	.46
LDH[x] (U/L), median (IQR)	294.5 (216.7-630.7)	255.5 (197.7-303.0)	.05
ALP[y] (U/L), median (IQR)	215.5 (154.2-263.5)	177.0 (91.5-221.7)	.05
T-bil[z] (mg/dL), median (IQR)	0.9 (0.7-1.3)	0.6 (0.4-0.8)	*<.001* [b]
Glucose (mg/dL), median (IQR)	147.0 (120.5-179.2)	136.0 (113.2-180.5)	.64
CPK[aa] (U/L), median (IQR)	249.5 (96.5-645.7)	129.0 (80.0-243.0)	.06
AMY[ab] (U/L), median (IQR)	85.5 (57.7-134.0)	80.5 (61.0-103.5)	.46
CRP[ac] (mg/dL), median (IQR)	0.4 (0.0-5.5)	0.1 (0.0-1.85)	*.03* [b]
Daily urinary output (mL), median (IQR)	412.5 (215.0-908.2)	1493.0 (985.0-1977.0)	*<.001* [b]
Respiratory rate (breaths per minute), median (IQR)	19.5 (17.7-21.0)	16.0 (14.0-20.0)	*.01* [b]
sBP[ad] (mmHg), median (IQR)	146.5 (115.5-160.2)	141.0 (124.2-157.0)	.84
dBP[ae] (mmHg), median (IQR)	81.0 (64.7-88.5)	74.0 (65.5-86.5)	.99
HR[af] (beats per minute), median (IQR)	97.0 (81.0-113.5)	88.5 (81.2-99.0)	.12

XSL•FO
RenderX

	Burns with delirium (n=32)	Burns without delirium (n=50)	*P* value
BT[ag] (°C), median (IQR)	37.0 (36.3-37.5)	37.0 (36.8-37.4)	.23

[a]This table presents the demographic, clinical, and laboratory characteristics of adult patients with burns who were admitted to the ICU at Mie University Hospital between January 2015 and June 2023. This study compared patients who developed ICU delirium with those who did not, highlighting differences in vital signs, burn severity, inflammatory markers, coagulation parameters, organ function indicators, and other physiological variables. The findings aimed to identify key risk factors for ICU delirium and provide insights into the pathophysiology and early prediction of delirium in patients with burns who are critically ill. Statistical comparisons were conducted using median and IQR values, with a *P* value of <.05 considered statistically significant.

[b]Statistically significant at *P*<.05.

[c]WBC: white blood cell count.

[d]RBC: red blood cell count.

[e]MCV: mean corpuscular volume.

[f]MCH: mean corpuscular hemoglobin.

[g]MCHC: mean corpuscular hemoglobin concentration.

[h]APTT: activated partial thromboplastin time.

[i]PT: prothrombin time.

[j]PT percentage.

[k]PT-INR: prothrombin time international normalized ratio.

[l]PCO_2: partial pressure of carbon dioxide.

[m]SO_2: saturation of oxygen.

[n]HCO_3–: bicarbonate.

[o]O_2-Hb: oxygenated hemoglobin.

[p]CO-Hb: carboxyhemoglobin.

[q]MetHb: methemoglobin.

[r]TP: total protein.

[s]BUN: blood urea nitrogen.

[t]eGFR: estimated glomerular filtration rate.

[u]CL: chloride.

[v]AST: aspartate aminotransferase.

[w]ALT: alanine aminotransferase.

[x]LDH: lactate dehydrogenase.

[y]ALP: alkaline phosphatase.

[z]T-bil: total bilirubin.

[aa]CPK: creatine phosphokinase.

[ab]AMY: amylase.

[ac]CRP: C-reactive protein.

[ad]sBP: systolic blood pressure.

[ae]dBP: diastolic blood pressure.

[af]HR: heart rate.

[ag]BT: body temperature.

Laboratory parameters further demonstrated significant differences. The delirium group showed higher white blood cell counts (median 11.9, IQR 9.0-18.5 per µL vs median 8.8, IQR 6.5-11.3 per µL; *P*=.006), lower hemoglobin (median 13.5, IQR 11.9-15.9 g/dL vs median 14.6, IQR 14.0-15.9 g/dL; *P*=.04), and lower hematocrit levels (median 40%, IQR 35%-44% vs median 43%, IQR 41%-46.7%; *P*=.009). Neutrophil percentages were elevated (median 77%, IQR 73.9%-83.7% vs median 74.4%, IQR 64.6%-80.9%; *P*=.04), whereas lymphocyte (median 15.8%, IQR 8.4%-17.9% vs median 16.8%, IQR 12.6%-23.6%; *P*=.02) and eosinophil (median 0.6%, IQR 0.4%-1.4% vs median 1.2%, IQR 0.6%-1.9%; *P*=.04) percentages were lower in the delirium group. Absolute neutrophil counts were higher (median 8985, IQR 6265-10,830 cells per µL vs median 6840, IQR 4449.2-9029.0 cells per µL; *P*=.03), as were monocyte counts (median 767, IQR 478-993 cells per µL vs median 575.0, IQR 447.5-767.0 cells per µL; *P*=.04).

Coagulation parameters revealed that activated partial thromboplastin time was prolonged in the delirium group (median 30.8, IQR 25.9-36.8 seconds vs median 27.4, IQR 24.9-29.9 seconds; *P*=.02). Prothrombin time (PT) and PT percentage also differed significantly (median 12.2, IQR 11.5-13.6 seconds vs median 11.3, IQR 10.8-12.1 seconds with *P*=.001 and median 94.8%, IQR 77%-104.1% vs median 105.4%, IQR 97%-116.9% with *P*=.001, respectively), as did the PT international normalized ratio (median 1.0, IQR 0.9-1.1 vs median 0.9, IQR 0.9-1.0; *P*=.002). D-dimer levels were

XSL•FO
RenderX

elevated (median 4.0, IQR 1.5-8.5 mg/L vs median 1.0, IQR 0.2-8.2 mg/L; *P*=.02).

Acid-base and oxygenation parameters showed a lower pH (median 7.3, IQR 7.3-7.4 vs median 7.4, IQR 7.3-7.4; *P*=.04) and higher PO_2 (median 152.5, IQR 93.5-322.7 mmHg vs median 123.0, IQR 83.2-166.0 mmHg; *P*=.03) in the delirium group. Oxygen saturation was also significantly higher (median 99.1%, IQR 98.3%-99.6% vs median 97.6%, IQR 94.9%-99%; *P*=.01), as was methemoglobin (MetHb; median 0.9%, IQR 0.5%-1.1% vs median 0.4%, IQR 0.3%-0.9%; *P*=.004).

Biochemical markers indicated lower total protein (median 6.4, IQR 5.9-6.7 g/dL vs median 7.0, IQR 6.2-7.5 g/dL; *P*=.006) and albumin (median 3.6, IQR 3.1-3.9 g/dL vs median 4.1, IQR 3.2-4.4 g/dL; *P*=.004) levels in the delirium group. Creatinine levels were higher (median 0.8, IQR 0.6-1.0 mg/dL vs median 0.7, IQR 0.5-0.8 mg/dL; *P*=.03), and estimated glomerular filtration rate was lower (median 67.2, IQR 47.8-87.2 mL per minute per 1.73 m^2 vs median 83.7, IQR 59.9-101.0 mL per minute per 1.73 m^2; *P*=.02). Serum calcium was lower (median 8.5, IQR 8.2-8.9 mg/dL vs median 8.9, IQR 8.4-9.2 mg/dL; *P*=.009), and T-bil levels were elevated (median 0.9, IQR 0.7-1.3 mg/dL vs median 0.6, IQR 0.4-0.8 mg/dL; *P*<.001). C-reactive protein levels were significantly higher in the delirium group (median 0.4, IQR 0.0-5.5 mg/dL vs median 0.1, IQR 0.0-1.85 mg/dL; *P*=.03).

Finally, daily urinary output was markedly reduced in the delirium group (median 412.5, IQR 215.0-908.2 mL vs median 1493.0, IQR 985.0-1977.0 mL; *P*<.001), and respiratory rates were higher (median 19.5, IQR 17.7-21.0 breaths per minute vs median 16.0, IQR 14.0-20.0 breaths per minute; *P*=.01).

Visual Examination of the Data Using Violin Plots

Next, the distribution of the data between the 2 groups (those with and without ICU delirium) was visually examined using violin plots. As shown in Figure 1, violin plots visually represent the density and range of the data, allowing for the identification of data dispersion and bimodality, that is, the shape of the distribution. Specifically, it became evident that urine output decreased proportionally, serving as a risk factor for ICU delirium.

XSL•FO
RenderX

Figure 1. Comparative analysis of explanatory factors for intensive care unit (ICU) delirium in patients with burns using violin plots. This figure illustrates the distribution of key clinical and laboratory variables among 2 groups of ICU patients with burns: those who developed delirium (with delirium) and those who did not (without delirium). Violin plots visualize the distribution and density of each variable, highlighting differences in factors such as age, burn area, total bilirubin (T-bil), methemoglobin (MetHb) levels, daily urinary output, and respiratory rate. These comparisons emphasize the significance of various explanatory factors in predicting ICU delirium, aiding in the identification of potential risk indicators for clinical decision-making. Alb: albumin; APTT: activated partial thromboplastin time; BUN: blood urea nitrogen; Ca: calcium; CPK: creatine phosphokinase; CRP: C-reactive protein; eGFR: estimated glomerular filtration rate; Hb: hemoglobin; Hct: hematocrit level; PT-%: prothrombin time percentage; PT: prothrombin time; PT-INR: PT international normalized ratio; sO2: oxygen saturation; TP: total protein; WBC: white blood cell count.

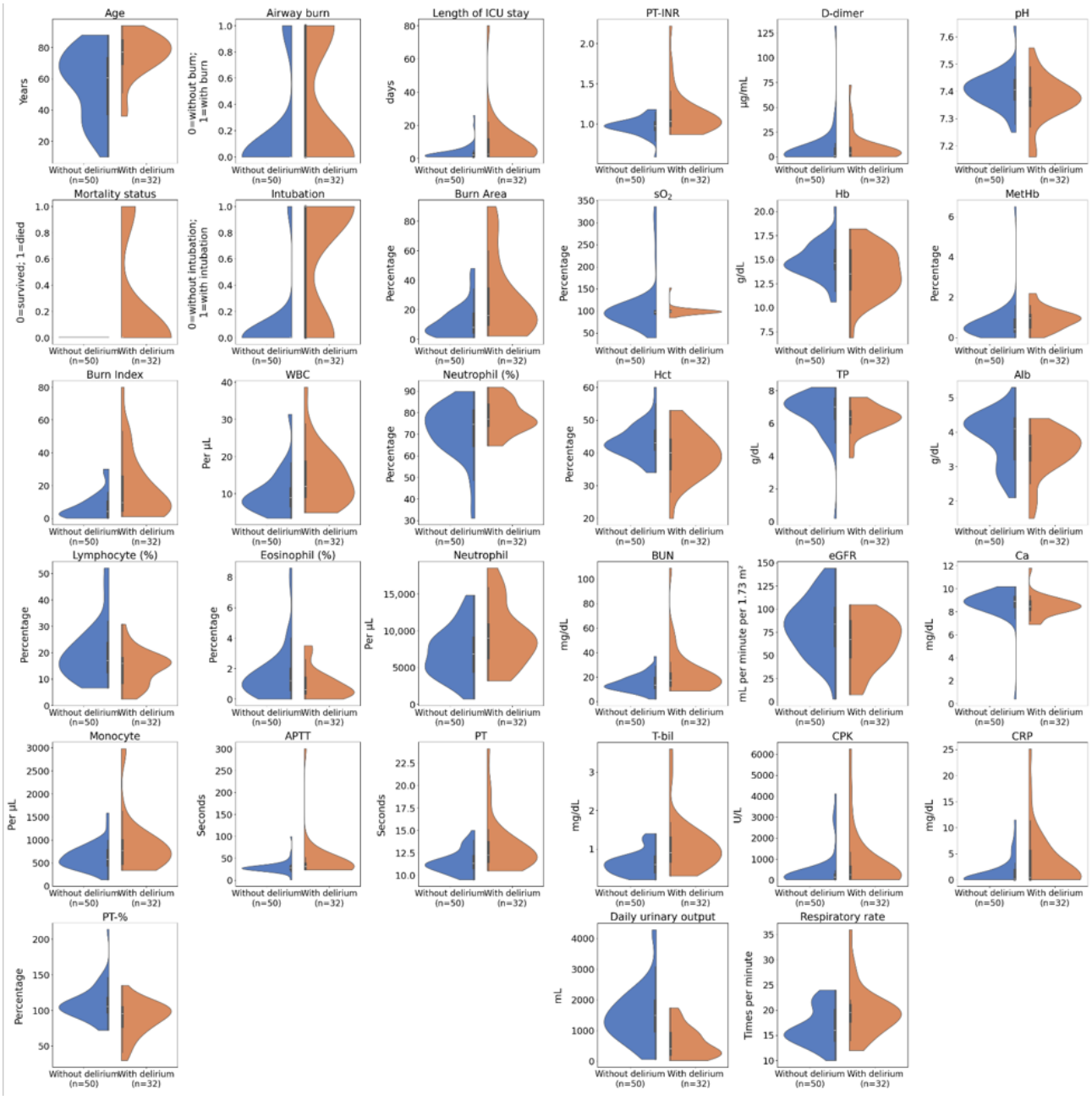

Machine Learning Model Evaluation for Delirium Prediction

In this study, we evaluated 10 different machine learning models for delirium prediction using 2 performance metrics: the AUC and Matthews correlation coefficient (MCC). The results are summarized in Table 2 and Figure 2.

XSL•FO
RenderX

Table 2. Performance comparison of machine learning models across various metrics[a].

Model	AUC[b], mean (SD)	MCC[c], mean (SD)	True positives	True negatives	False positives	False negatives	Accuracy	Precision	Recall	F_1-score
Support vector machine	0.897 (0.056)	0.157 (0.202)	0.6	10.0	0.0	5.8	0.647	0.400	0.095	0.150
Neural network	0.857 (0.058)	0.522 (0.238)	4.2	8.4	1.6	2.2	0.757	0.750	0.652	0.664
K-nearest neighbor	0.894 (0.060)	0.529 (0.311)	3.4	9.6	0.4	3.0	0.793	0.733	0.524	0.605
Decision tree	0.729 (0.033)	0.417 (0.059)	4.2	7.6	2.4	2.2	0.720	0.655	0.624	0.634
Naïve Bayes	0.827 (0.095)	0.411 (0.351)	3.0	9.2	0.8	3.4	0.744	0.683	0.462	0.533
AdaBoost[d]	0.832 (0.094)	0.493 (0.126)	4.0	8.4	1.6	2.4	0.756	0.753	0.614	0.651
Gradient boosting machine	0.821 (0.074)	0.486 (0.026)	3.8	8.6	1.4	2.6	0.768	0.796	0.586	0.651
Linear discriminant analysis	0.660 (0.114)	0.337 (0.163)	3.8	7.4	2.6	2.6	0.684	0.601	0.590	0.583
Logistic regression	*0.906* [e] *(0.073)*	*0.625 (0.162)*	4.8	8.6	1.4	1.6	*0.818*	0.797	*0.743*	*0.755*
Random forest	0.850 (0.074)	0.505 (0.217)	3.4	9.2	0.8	3.0	0.757	*0.811*	0.529	0.617

[a]This table presents the predictive performance of 10 machine learning models trained on clinical and laboratory data from patients with burns admitted to the intensive care unit (ICU) at Mie University Hospital. Model performance was evaluated using the mean area under the curve, the mean Matthews correlation coefficient, accuracy, precision, recall, and F_1-score, providing a comprehensive assessment of each model's ability to predict ICU delirium onset. The table also reports the true positive, true negative, false positive, and false negative counts, offering insights into each model's sensitivity and specificity. Logistic regression demonstrated the highest predictive performance (area under the curve=0.906), whereas decision tree and linear discriminant analysis showed relatively lower predictive power. These findings highlight the potential of machine learning in early risk stratification for ICU delirium, emphasizing the importance of selecting an optimal predictive model for clinical application.

[b]AUC: area under the curve.

[c]MCC: Matthews correlation coefficient.

[d]AdaBoost: adaptive boosting.

[e]Values in italics indicate the best performance for each metric. Logistic regression exhibited the highest AUC, MCC, accuracy, recall, and F_1 -score, demonstrating superior overall performance.

XSL•FO
RenderX

Figure 2. Receiver operating characteristic (ROC) curves for machine learning models predicting intensive care unit (ICU) delirium. This figure shows ROC curves for 10 machine learning models, illustrating sensitivity and 1-specificity from 5-fold stratified cross-validation. The legend presents mean area under the curve (AUC) values with SDs, indicating predictive accuracy. Logistic regression (LR; mean AUC 0.906, SD 0.073) achieved the highest performance, whereas decision tree and linear discriminant analysis (LDA) performed poorly. The diagonal line (AUC 0.5) represents random guessing. Models with curves near the top left corner demonstrate superior predictive ability for ICU delirium in patients with burns. AdaBoost: adaptive boosting; GBM: gradient-boosting machine; KNN: k-nearest neighbor; SVM: support vector machine.

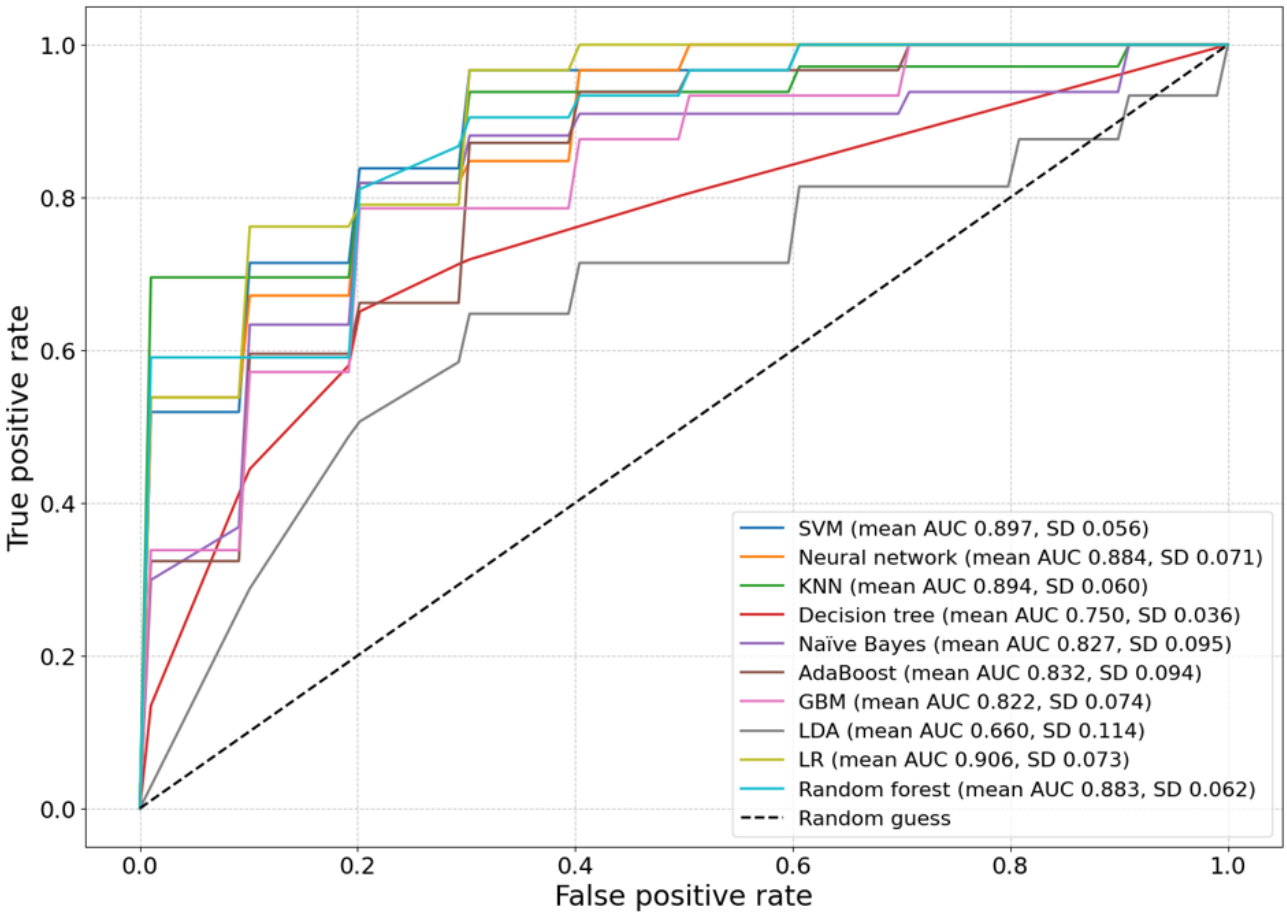

The LR model showed the best overall performance, achieving the highest mean AUC of 0.906 (SD 0.073) and the highest mean MCC of 0.625 (SD 0.162). This suggests that LR provides a good balance between discriminative power and classification accuracy for delirium prediction. The SVM and KNN models also showed strong discriminative ability, with mean AUC values of 0.897 (SD 0.056) and 0.894 (SD 0.060), respectively. However, the SVM model's relatively low MCC score of 0.157 (SD 0.202) indicates that it may struggle with precise classification, possibly due to class imbalance or suboptimal threshold selection. Neural network and RF demonstrated comparable performance, with mean AUC values of 0.857 (SD 0.058) and 0.850 (SD 0.074), respectively. Their MCC scores were also similar, suggesting consistent performance across both metrics. The decision tree model and LDA showed the lowest performance among the evaluated models, with mean AUC values of 0.729 (SD 0.033) and 0.660 (SD 0.114), respectively. This suggests that these models may not capture the complexity of the delirium prediction task as effectively as the other algorithms. It is noteworthy that, while some models (eg, SVM) achieved high AUC scores, their corresponding MCC scores were relatively low. This discrepancy highlights the importance of using multiple evaluation metrics to gain a comprehensive understanding of model performance, especially in potentially imbalanced classification tasks such as delirium prediction. In summary, LR emerged as the most promising model for delirium prediction in this comparative analysis. However, the strong performance of several other models, such as KNN and neural network, suggests that ensemble methods or model stacking could potentially yield further improvements in predictive accuracy (Table 2).

Performance Metrics of Machine Learning Models in ICU Delirium Prediction

On the basis of the provided classification performance metrics, LR showed the highest overall performance in terms of accuracy (0.818), precision (0.797), recall (0.743), and F_1-score (0.755), indicating a robust balance between sensitivity and specificity. Although KNN showed slightly higher accuracy (0.793) than most models, its recall was lower (0.524), which may limit its utility in detecting positive cases.

Among ensemble methods, GBM and AdaBoost performed well, with GBM achieving a relatively high precision (0.796), reflecting its ability to minimize false positives. RF also yielded strong precision (0.811) but showed lower recall (0.529), suggesting a higher risk of missing positive cases.

XSL•FO
RenderX

Neural network models achieved a competitive balance across all metrics, notably F_1-score (0.664), which highlights their consistency in handling both sensitivity and specificity. However, SVM and LDA demonstrated weaker recall, suggesting that these models are less suited for tasks requiring high sensitivity.

We then compared the mean AUCs across the models to evaluate their overall discriminative ability. When compared with decision tree, SVM (*P*=.002), neural network (*P*=.005), KNN (*P*=.003), LR (*P*=.003), and RF (*P*=.005) showed a significantly better performance. Naïve Bayes (*P*=.04), AdaBoost (*P*=.03), and GBM (*P*=.03) also demonstrated a significantly better performance than that of decision tree. LDA showed a comparable performance to that of decision tree (*P*=.99; Figure 3).

XSL•FO
RenderX

Figure 3. (A) Mean area under the curve (AUC) scores and (B) statistical significance across machine learning models. This figure compares the predictive performance of 10 machine learning models for intensive care unit (ICU) delirium in patients with burns. The bar graph shows mean AUC scores with SDs from 5-fold stratified cross-validation. Among the models, logistic regression (LR; AUC=0.906) performed best, whereas decision tree and linear discriminant analysis (LDA) had lower AUCs. The heat map displays pairwise *P* values from 2-sample 2-tailed t tests, with darker cells indicating significant differences (*P*<.05). This figure highlights model performance and statistical differences, aiding in selecting effective classifiers for ICU delirium prediction. AdaBoost: adaptive boosting; GBM: gradient-boosting machine; KNN: k-nearest neighbor; SVM: support vector machine.

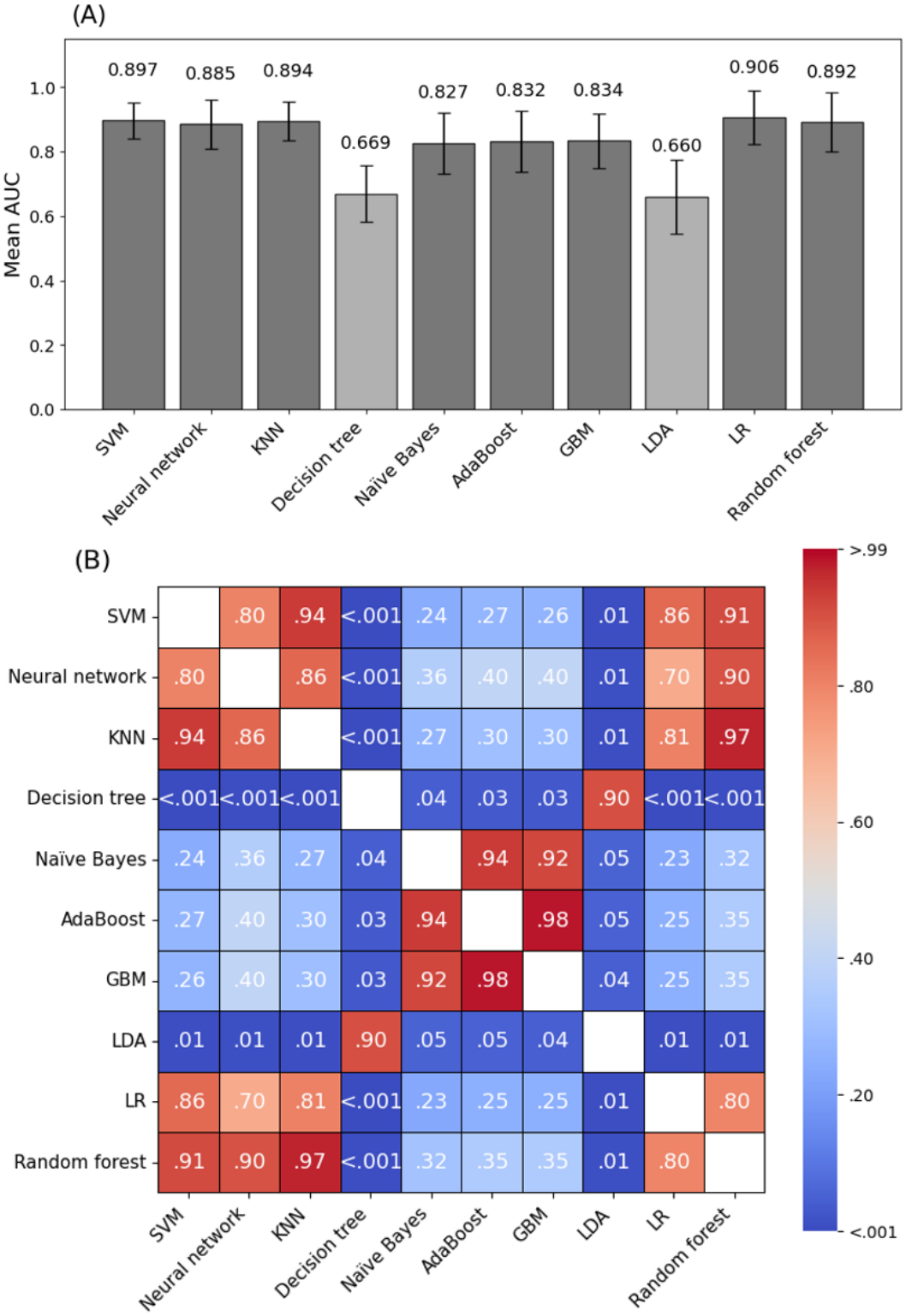

Similarly, when compared with LDA, SVM (*P*=.006), neural network (*P*=.01), KNN (*P*=.007), LR (*P*=.007), and RF (*P*=.009) exhibited a significantly better performance. GBM (*P*=.04) and AdaBoost (*P*=.047) also showed a significantly better performance than that of LDA, whereas naïve Bayes showed a marginally better performance (*P*=.05). Decision tree demonstrated a comparable performance to that of LDA (*P*=.99).

In summary, LR was the most balanced model across all metrics, making it a preferred choice for clinical applications where both

XSL•FO
RenderX

precision and recall are critical. Ensemble methods such as GBM and AdaBoost offer strong alternatives, particularly for optimizing precision (Table 2).

Identification of High-Risk Factors for ICU Delirium Using SHAP Analysis

We used SHAP analysis to identify the top 15 high-risk factors for ICU delirium in each validated machine learning model (Multimedia Appendix 1). This comprehensive analysis revealed distinct sets of risk factors across different models, providing valuable insights into the complex nature of ICU delirium. Key findings included the following:

- LR (best overall performance; AUC=0.906; MCC=0.625): identified daily urine output, eosinophil count, age, and fibrinogen levels as key risk factors.
- Neural network (AUC=0.857; MCC=0.522): highlighted lactate dehydrogenase levels, daily urine output, inhalation injury, neutrophil count, and platelet count as key risk factors.
- SVM (AUC=0.897; MCC=0.157): uniquely identified burn area and length of ICU stay as significant factors.
- KNN (AUC=0.894): emphasized neutrophil and monocyte percentages as key risk factors.
- Decision tree: despite its lower performance, it identified endotracheal intubation and D-dimer levels as risk factors.
- Other models: consistently highlighted hematological parameters, daily urine output, and age across multiple models. Novel factors such as MetHb levels (LDA) and anion gap (RF) were also identified.

Discussion

Machine Learning–Based Prediction of ICU Delirium in Patients With Burns: LR Performance and SHAP Analysis for Risk Factor Identification

Our study demonstrated that LR effectively predicted ICU delirium in patients with burns using clinical data, including vital signs and blood biomarkers. Among the machine learning models, LR achieved the highest predictive accuracy, confirming that ICU delirium risk can be assessed through computational modeling (Figures 2 and 3).

Furthermore, SHAP analysis identified key risk factors associated with ICU delirium, highlighting T-bil, MetHb, daily urine output, and leucocyte fractions as novel predictors alongside established factors such as burn area and tracheal intubation (Multimedia Appendix 1). These findings provide crucial insights for developing targeted prevention and treatment strategies, emphasizing the importance of respiratory, renal, hepatic, and inflammatory function management in patients with burns who are critically ill.

The Role of T-Bil Levels in Predicting Delirium in ICU Patients With Burns: Connections With Cholestasis and Inflammation

T-bil levels were identified as a risk factor for the development of delirium in ICU patients with burns. However, the direct relationship between burns and cholestasis remains unclear. Cholestasis often occurs after burns, and patients with burns who have increased bilirubin levels without a corresponding increase in alkaline phosphatase and gamma-glutamyl transferase levels face a higher risk of mortality. Furthermore, intrahepatic cholestasis is observed in half of patients with severe burns [6]. Cholestasis is also associated with hypoxic hepatitis [7]. In addition, hypovolemic shock observed in severe burns may be involved in the elevation of T-bil levels [8]. Increases in interleukin (IL)-6 and tumor necrosis factor-α, which are observed in the early stages of severe burns [9], have been reportedly associated with hyperbilirubinemia [10] and organ dysfunction [11]. Inflammatory cytokines such as IL-6 and tumor necrosis factor-α may reduce the expression of bile transporters on the canalicular membrane of hepatocytes, leading to an increase in T-bil levels [12,13]. These research findings support the validity of our study, which identified an increase in T-bil levels as an important risk factor for predicting ICU delirium in patients with burns.

MetHb Level as a Novel Indicator of ICU Delirium in Patients With Burns: Insights From Sepsis and Hemolysis Research

MetHb level was identified as an explanatory factor. No previous studies have clearly demonstrated the involvement of MetHb level as a risk factor for ICU delirium in patients with burns. However, several studies have shown the involvement of MetHb in delirium in patients with sepsis, which, similarly to burns, can cause a cytokine storm. As is well known, patients with sepsis have a high incidence of delirium, and in these patients, nitric oxide is released into the bloodstream due to ischemia-reperfusion stimulation. Nitric oxide is converted into MetHb and nitrates; as a result, the concentration of MetHb in the blood is a useful marker for the onset of sepsis or septic shock [14]. However, the molecular mechanisms through which MetHb causes delirium in patients with sepsis remain largely unknown [15]. In patients with severe conditions, such as trauma or infection, intracellular hemolysis may occur, leading to anemia. Anemia, which develops relatively early in severe conditions, is thought to result from damaged red blood cells processed by the reticuloendothelial system. Acute hemolysis leads to an increase in free hemoglobin in the blood. Subsequently, free hemoglobin and heme are released into the circulatory system, and the wound interstitium is rapidly converted into MetHb by oxidants. The increase in MetHb levels is more pronounced in the ischemia-reperfusion areas, where activated macrophages and neutrophils accumulate [16]. Therefore, MetHb produced by ischemia-reperfusion injury and hemolysis, as observed in severe conditions such as sepsis, may affect leucocyte cell adhesion, phagocytic ability, and metabolic activation and may be involved in ICU delirium. Previous research has shown a stronger correlation among the total amount of hemoglobin [17,18], red blood cell count [19], and delirium. Thus, hemoglobin and MetHb levels in the blood are important factors influencing delirium.

The Impact of Decreased Urine Output on Delirium Risk in ICU Patients With Burns: Insights From SHAP Analysis

Using SHAP analysis, we identified that a decreased daily urine output within 24 hours of ICU admission is a risk factor for

XSL•FO
RenderX

ICU delirium in patients with burns. Interestingly, daily urine output was identified as an important risk factor in 50% (5/10) of the machine learning models evaluated (Multimedia Appendix 1). The finding that a decrease in 24-hour urine volume is a risk factor for delirium in patients with burns seems reasonable considering that acute kidney injury can potentially increase the risk of delirium 10-fold in patients who are critically ill [20]. For example, urine output is an important indicator of renal function and hydration status. In patients with burns, a decrease in urine output may indicate insufficient renal perfusion or dehydration. Both factors contribute to the development of delirium [21]. Furthermore, patients with burns often require multiple medications, such as sedatives and analgesics. Decreased urine output can affect the metabolism and excretion of these drugs, potentially leading to the accumulation of psychoactive substances that can induce or exacerbate delirium [22]. In addition, a decrease in urine output can lead to electrolyte imbalance, which is known to cause neurological dysfunction and delirium. Electrolyte abnormalities such as hyponatremia or hypernatremia can occur in patients with burns because of fluid shifts and inadequate fluid replacement [23].

Leukocyte Biomarkers as Indicators of Delirium in ICU Patients With Burns: The Role of Inflammatory Response

Our study identified the number or proportion of neutrophils and monocytes in the leucocyte fraction as risk factors for delirium development (Multimedia Appendix 1). This suggests that exposure of leucocytes to a cytokine storm due to excessive stress from burns contributes to delirium in patients with burns. Numerous studies have used leucocyte biomarkers to diagnose delirium in the past [24]. Inflammatory biomarkers and brain-specific metabolic biomarkers have been extensively studied in delirium, and inflammatory cytokines and activation markers of astrocytes and glial cells (IL-6, IL-8, IL-10, tumor necrosis factor-α, C-reactive protein, and S100β protein levels) positively correlate with longer duration of delirium, severity of delirium, and higher in-hospital mortality [25]. In addition, elevated levels of IL-8 and S100β protein have been associated with increased mortality in patients with delirium [26]. In a mouse model with delirium, the infiltration of bone marrow–derived monocytes into the blood-brain barrier [27] and activation of microglia were observed [28]. Although the number and proportion of neutrophils and monocytes were identified as risk factors for ICU delirium, these leucocytes may be involved in the production of inflammatory cytokines and contribute to the onset of ICU delirium.

Respiratory Rate as a Predictor of ICU Delirium in Patients With Burns: New Insights and Implications

In our study, an increase in the respiratory rate was a risk factor for ICU delirium. Patients with ICU delirium had a higher median respiratory rate of 19.5 (IQR 17.7-21.0) breaths per minute compared to 16.0 (IQR 14.0-20.0) breaths per minute in patients without delirium (P=.01; Table 1). To our knowledge, no studies have clearly established a link between delirium and respiratory rate. Delirium is generally recognized as a common complication in patients with respiratory failure in the ICU. The incidence of delirium in the ICU ranges from 10% to 78%, with most cases occurring in patients receiving mechanical ventilation. This suggests a significant overlap between respiratory complications and the occurrence of delirium; however, a direct correlation between an increase in respiratory rate and delirium has not been explicitly stated [29]. Inhalation injuries occur in approximately one-third of burn hospital admissions and contribute to a high mortality rate (50%) in patients with burns. Therefore, an increase in the respiratory rate may be associated with carbon monoxide poisoning and chemical tracheobronchitis due to the inhalation of toxic combustion products and generally correlates with a higher mortality rate. Unfortunately, many patients with burns receive high-concentration oxygen therapy from emergency teams before being transported to the ICU or emergency room. Therefore, carboxyhemoglobin levels are often adjusted to lower levels, and PO_2 is frequently high during treatment, which is why carboxyhemoglobin and PO_2 were not identified as risk factors in our model. Therefore, it might be appropriate to consider respiratory rate as a potential risk factor for ICU delirium in future studies [30,31].

Advancing ICU Delirium Research: The Prediction of Delirium in ICU Patients Model and the Need for Machine Learning Approaches in Patients With Burns

In the field of delirium research, the Prediction of Delirium in ICU Patients (PRE-DELIRIC) model is considered a seminal study [32]. In the ICU, the PRE-DELIRIC model uses 10 identified risk factors (age, Acute Physiology and Chronic Health Evaluation II score, admission group, coma, infection, metabolic acidosis, use of sedatives and morphine, blood urea nitrogen, and emergency admission) and predicts delirium with an AUC of 0.87 (95% CI 0.85-0.89) within 24 hours of ICU admission [32]. Furthermore, the model by Lanzhou University [33] heavily relies on patients' detailed past medical histories, making data collection challenging in busy clinical settings such as emergency rooms and ICUs, where obtaining comprehensive patient histories, diagnoses, and treatments can be difficult [34]. Therefore, it is crucial to establish machine learning models that can accurately predict conditions with multifactorial risk factors, such as ICU delirium, using data that are easily obtainable during emergency department visits, such as vital signs and blood data. Despite this need, a delirium prediction model for ICU patients with burns using machine learning has not yet been developed. Predictive models for diseases such as ICU delirium, which involve numerous risk factors, stand to benefit significantly from machine learning's capability to perform multifactorial analyses, surpassing traditional biostatistical methods. Our proposed machine learning model can more effectively evaluate complex interactions among multivariate data, which is essential for accurately predicting conditions with multifactorial risk factors such as ICU delirium. Therefore, although our study had a small number of cases, it is considered valuable for the development of a machine learning–based ICU delirium prediction model and the identification of risk factors.

Comprehensive Model Evaluation for ICU Delirium

When evaluating the performance of different models, it is important to consider not only the AUC but also other metrics,

XSL•FO
RenderX

such as accuracy, precision, recall, and F_1-score. In a specialized medical environment such as the ICU, some metrics may become more important than others when dealing with specific diseases. For example, precision is important if avoiding false detections of delirium is crucial. Conversely, if it is vital to avoid missing cases of delirium, recall should be emphasized. In our study, we considered all these metrics comprehensively and selected the model that best suited the objectives of the research and clinical demands.

Among the 10 models we adjusted, the LR model was found to be the most balanced and high performing. Notably, it demonstrated the highest values across key metrics, including AUC, MCC, accuracy, and F_1-score, suggesting that it provides the most reliable predictions from various perspectives.

Strengths

The explanatory variables in our study were based on blood data collected immediately upon the arrival of patients with burns at the emergency outpatient clinic. Therefore, there was a time lag between the collection of these data and the collection of data at the time of delirium diagnosis using the ICDSC. However, we believe that our model, which accurately predicts the development of delirium during an ICU stay based on blood data and vital signs at the time of patient arrival, can be easily interpreted by clinicians and has high general applicability. This allows clinicians to predict the incidence of ICU delirium, which significantly affects the prognosis of severe burns from early in the patient's hospitalization, enabling the initiation of early interventions for patients at high risk of ICU delirium.

Limitations

Demerit of Binary Classification for Delirium

A binary classification model was used for predicting delirium, which simplified the model's output to either *delirium present* or *delirium absent*. This approach facilitated decision-making in clinical settings by providing clear, binary outcomes. Conversely, using a regression model would require predicting specific ICDSC scores and interpreting these scores to assess patient status. Thus, binary classification often offers greater practicality in busy clinical environments due to its straightforward interpretation. However, regression analysis may be more appropriate for predicting continuous outcomes, such as the ICDSC score. This method offers detailed information by producing continuous values, allowing for a nuanced understanding of the severity of ICU delirium, ranging from mild to severe. For instance, the difference in delirium between an ICDSC score of 3 and 4 might be minimal, whereas the difference between scores of 1 and 7 indicates substantially different symptoms. Therefore, future research should investigate the benefits of developing regression models that predict ICDSC scores.

Model Selection Criteria for Small Datasets

This study examined delirium in patients with burns who were critically ill admitted to the ICU and did not include patients with missing data, resulting in a small number of cases. In analyses with small datasets, such as in our 82 cases, it is generally recommended to avoid overfitting and choose a simpler model. In this study, we were cautious about using complex models (such as deep neural networks or RF with many trees) that are prone to overfitting with limited data. Therefore, simple models are generally recommended for studies with few cases. Typically, models with regularization effects (such as LR) are effective in preventing overfitting. In our study, LR demonstrated a high accuracy (0.818), indicating that overfitting was well controlled during model creation. Linear SVMs have also been proposed to prevent overfitting. In addition, LDA has linear boundaries, is computationally fast, and can sometimes provide relatively stable results even with a small amount of data.

Limitations and Future Directions for External Validation

The performance of our model was discussed using data from patients in the ICU in a single hospital; however, an external validation was not conducted. Therefore, our study is limited by its single-institution setting. Future research could strengthen the reliability of our delirium prediction model for patients with burns in the ICU by conducting external validation using datasets from other hospitals.

Enhancing Predictive Power: Inclusion of Diverse Variables

Our machine learning model incorporated 70 explanatory variables, including patients' vital signs and blood biomarkers. However, unlike the PRE-DELIRIC model, which includes medication history and environmental factors, our study did not consider these variables. Expanding the range of input variables in future research may further enhance the predictive accuracy of delirium risk assessment.

Conclusions

Our study underscores the clinical utility of machine learning in predicting ICU delirium in patients with burns, demonstrating that LR provided the highest predictive accuracy among the tested models. Using SHAP analysis, we identified both well-established and novel risk factors, such as T-bil, MetHb, urine output, and leucocyte fractions, offering new insights into the complex pathophysiology of delirium. These findings suggest that early identification of patients at high risk using readily available clinical data upon ICU admission could facilitate proactive intervention strategies, potentially reducing morbidity and improving patient outcomes.

Beyond its immediate clinical applications, this study highlights the need for integrating machine learning into real-world ICU decision-making systems. Traditional delirium prediction models such as PRE-DELIRIC rely on a limited set of predefined variables, whereas machine learning models can dynamically incorporate diverse clinical parameters, enabling real-time risk stratification. This adaptability is particularly relevant in burn care, where patients exhibit highly variable and rapidly evolving physiological changes.

Moving forward, the integration of multi-institutional external validation is crucial to ensure the generalizability and robustness of our predictive model across diverse ICU settings. In addition, further research should explore the mechanistic pathways linking

XSL•FO
RenderX

identified risk factors with delirium onset, which could pave the way for personalized prevention and treatment strategies. Ultimately, this study provides a foundation for the next generation of delirium risk prediction models, emphasizing the potential of artificial intelligence–driven clinical decision support to enhance patient care in critical care medicine.

Acknowledgments

This study was supported by the Japan Society for the Promotion of Science (grants JP 22K06910, 22K09117, 22K09138, 19KK0224, 18K08916, and 21K09015). The authors want to thank their colleagues in the intensive care unit at Mie University Hospital. They also appreciate Dr Sakamoto, a PhD candidate in engineering, and his colleagues for their help in creating the machine learning model and Shapley additive explanations analysis. In addition, the authors benefited from the cooperation of the Dr Shimaoka Laboratory for Molecular Biology in biostatistics and molecular biological descriptions. They would like to express their gratitude to Dr Sakamoto and his team, who are experts in machine learning, for their invaluable contributions to improving the accuracy of the authors' machine learning models. In addition, they extend their thanks to the members of Dr Shimaoka Laboratory for Molecular Biology for their insightful comments on the manuscript, which significantly enhanced its quality.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to the need to protect participant confidentiality but are available from the corresponding author on reasonable request. In addition, the Python code used for the machine learning analyses, including modifications and optimizations conducted using generative artificial intelligence, is provided in Multimedia Appendices 2-14. Further details regarding the code can be obtained by contacting the corresponding author.

Authors' Contributions

RE and EK were primarily responsible for manuscript preparation and drafting of the original text. MS oversaw the overall direction and planning of the research. The other coauthors, RS, AIM, FO, TS, AH, KI, YA, HF, AG, EJP, RM, and RK, provided supervision and were instrumental in shaping the research, analysis, and manuscript refinement.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Comparison of feature Shapley additive explanations (SHAP) values across multiple machine learning models. The 15 most influential features in each model's predictions are displayed. A higher SHA*P* value indicates a stronger impact of the feature on the model's predictions.
[DOCX File , 23 KB - formative_v9i1e65190_app1.docx]

Multimedia Appendix 2
Python code used in this study for generating the violin plots comparing explanatory factors for intensive care unit delirium in patients with burns.
[DOCX File , 17 KB - formative_v9i1e65190_app2.docx]

Multimedia Appendix 3
Python code used in this study for generating the receiver operating characteristic curves illustrating machine learning model performance in predicting intensive care unit delirium.
[DOCX File , 15 KB - formative_v9i1e65190_app3.docx]

Multimedia Appendix 4
Python code used in this study for generating the bar graph and heat map comparing machine learning model performance in predicting intensive care unit delirium.
[DOCX File , 16 KB - formative_v9i1e65190_app4.docx]

Multimedia Appendix 5
Adaptive boosting code for Shapley additive explanations analysis.
[DOCX File , 14 KB - formative_v9i1e65190_app5.docx]

XSL•FO
RenderX

Multimedia Appendix 6
Decision tree code for Shapley additive explanations analysis.
[DOCX File , 15 KB - formative_v9i1e65190_app6.docx]

Multimedia Appendix 7
K-nearest neighbor code for Shapley additive explanations analysis.
[DOCX File , 15 KB - formative_v9i1e65190_app7.docx]

Multimedia Appendix 8
Linear discriminant analysis code for Shapley additive explanations analysis.
[DOCX File , 18 KB - formative_v9i1e65190_app8.docx]

Multimedia Appendix 9
Light gradient-boosting machine code for Shapley additive explanations analysis.
[DOCX File , 15 KB - formative_v9i1e65190_app9.docx]

Multimedia Appendix 10
Logistic regression code for Shapley additive explanations analysis.
[DOCX File , 17 KB - formative_v9i1e65190_app10.docx]

Multimedia Appendix 11
Naïve Bayes code for Shapley additive explanations analysis.
[DOCX File , 14 KB - formative_v9i1e65190_app11.docx]

Multimedia Appendix 12
Neural network code for Shapley additive explanations analysis.
[DOCX File , 17 KB - formative_v9i1e65190_app12.docx]

Multimedia Appendix 13
Random forest code for Shapley additive explanations analysis.
[DOCX File , 15 KB - formative_v9i1e65190_app13.docx]

Multimedia Appendix 14
Support vector machine code for Shapley additive explanations analysis.
[DOCX File , 16 KB - formative_v9i1e65190_app14.docx]

References

1. Watt J, Tricco AC, Talbot-Hamon C, Pham B, Rios P, Grudniewicz A, et al. Identifying older adults at risk of delirium following elective surgery: a systematic review and meta-analysis. J Gen Intern Med 2018 Apr 26;33(4):500-509 [FREE Full text] [doi: 10.1007/s11606-017-4204-x] [Medline: 29374358]
2. Devlin JW, Skrobik Y, Gélinas C, Needham DM, Slooter AJ, Pandharipande PP, et al. Clinical practice guidelines for the prevention and management of pain, agitation/sedation, delirium, immobility, and sleep disruption in adult patients in the ICU. Crit Care Med 2018 Sep;46(9):e825-e873 [FREE Full text] [doi: 10.1097/CCM.0000000000003299] [Medline: 30113379]
3. Agarwal V, O Neill PJ, Cotton BA, Pun BT, Haney S, Thompson J, et al. Prevalence and risk factors for development of delirium in burn intensive care unit patients. J Burn Care Res 2010;31(5):706-715. [doi: 10.1097/bcr.0b013e3181eebee9]
4. Guo Z, Liu J, Li J, Wang X, Guo H, Ma P, et al. Postoperative delirium in severely burned patients undergoing early escharotomy. J Burn Care Res 2017;38(1):e370-e376. [doi: 10.1097/bcr.0000000000000397]
5. Hur S, Ko RE, Yoo J, Ha J, Cha WC, Chung CR. A machine learning-based algorithm for the prediction of intensive care unit delirium (PRIDE): retrospective study. JMIR Med Inform 2021 Jul 26;9(7):e23401 [FREE Full text] [doi: 10.2196/23401] [Medline: 34309567]
6. de Tymowski C, Dépret F, Soussi S, Nabila M, Vauchel T, Chaussard M, et al. Contributing factors and outcomes of burn-associated cholestasis. J Hepatol 2019 Sep;71(3):563-572. [doi: 10.1016/j.jhep.2019.05.009] [Medline: 31152758]
7. Jäger B, Drolz A, Michl B, Schellongowski P, Bojic A, Nikfardjam M, et al. Jaundice increases the rate of complications and one-year mortality in patients with hypoxic hepatitis. Hepatology 2012 Dec 24;56(6):2297-2304. [doi: 10.1002/hep.25896] [Medline: 22706920]

XSL•FO
RenderX

8. Soussi S, Dépret F, Benyamina M, Legrand M. Early hemodynamic management of critically ill burn patients. Anesthesiology 2018 Sep;129(3):583-589. [doi: 10.1097/ALN.0000000000002314] [Medline: 29958240]
9. Csontos C, Foldi V, Pálinkas L, Bogar L, Röth E, Weber G, et al. Time course of pro- and anti-inflammatory cytokine levels in patients with burns--prognostic value of interleukin-10. Burns 2010 Jun;36(4):483-494. [doi: 10.1016/j.burns.2009.10.009] [Medline: 20045261]
10. Jones A, Selby PJ, Viner C, Hobbs S, Gore ME, McElwain TJ. Tumour necrosis factor, cholestatic jaundice, and chronic liver disease. Gut 1990 Aug 01;31(8):938-939 [FREE Full text] [doi: 10.1136/gut.31.8.938] [Medline: 2387521]
11. Sprague AH, Khalil RA. Inflammatory cytokines in vascular dysfunction and vascular disease. Biochem Pharmacol 2009 Sep 15;78(6):539-552 [FREE Full text] [doi: 10.1016/j.bcp.2009.04.029] [Medline: 19413999]
12. Whiting JF, Green RM, Rosenbluth AB, Gollan JL. Tumor necrosis factor-alpha decreases hepatocyte bile salt uptake and mediates endotoxin-induced cholestasis. Hepatology 1995 Oct;22(4 Pt 1):1273-1278. [doi: 10.1016/0270-9139(95)90639-8] [Medline: 7557881]
13. Green RM, Whiting JF, Rosenbluth AB, Beier D, Gollan JL. Interleukin-6 inhibits hepatocyte taurocholate uptake and sodium-potassium-adenosinetriphosphatase activity. Am J Physiol Gastrointest Liver Physiol 1994 Dec 01;267(6):G1094-G1100. [doi: 10.1152/ajpgi.1994.267.6.g1094]
14. Ohashi K, Yukioka H, Hayashi M, Asada A. Elevated methemoglobin in patients with sepsis. Acta Anaesthesiol Scand 1998 Jul 31;42(6):713-716. [doi: 10.1111/j.1399-6576.1998.tb05306.x] [Medline: 9689279]
15. Atterton B, Paulino MC, Povoa P, Martin-Loeches I. Sepsis associated delirium. Medicina (Kaunas) 2020 May 18;56(5):240 [FREE Full text] [doi: 10.3390/medicina56050240] [Medline: 32443606]
16. Schaer DJ, Buehler PW, Alayash AI, Belcher JD, Vercellotti GM. Hemolysis and free hemoglobin revisited: exploring hemoglobin and hemin scavengers as a novel class of therapeutic proteins. Blood 2013 Feb 21;121(8):1276-1284 [FREE Full text] [doi: 10.1182/blood-2012-11-451229] [Medline: 23264591]
17. Kijima E, Kayama T, Saito M, Kurosaka D, Ikeda R, Hayashi H, et al. Pre-operative hemoglobin level and use of sedative-hypnotics are independent risk factors for post-operative delirium following total knee arthroplasty. BMC Musculoskelet Disord 2020 May 02;21(1):279 [FREE Full text] [doi: 10.1186/s12891-020-03206-4] [Medline: 32359366]
18. Liu YM, Huang H, Gao J, Zhou J, Chu HC. Hemoglobin concentration and post-operative delirium in elderly patients undergoing femoral neck fracture surgery. Front Med (Lausanne) 2021 Jan 5;8:780196 [FREE Full text] [doi: 10.3389/fmed.2021.780196] [Medline: 35071265]
19. Prasko J, Horacek R, Krnacova B, Latalova K. Delirium as a complication of the surgical intensive care. Neuropsychiatr Dis Treat 2016 Sep;Volume 12:2425-2434. [doi: 10.2147/ndt.s115800]
20. Zipser CM, Deuel J, Ernst J, Schubert M, Weller M, von Känel R, et al. Predisposing and precipitating factors for delirium in neurology: a prospective cohort study of 1487 patients. J Neurol 2019 Dec 13;266(12):3065-3075. [doi: 10.1007/s00415-019-09533-4] [Medline: 31520105]
21. Pang H, Kumar S, Ely EW, Gezalian MM, Lahiri S. Acute kidney injury-associated delirium: a review of clinical and pathophysiological mechanisms. Crit Care 2022 Aug 27;26(1):258 [FREE Full text] [doi: 10.1186/s13054-022-04131-9] [Medline: 36030220]
22. Wan R, McKenzie CA, Taylor D, Camporota L, Ostermann M. Acute kidney injury as a risk factor of hyperactive delirium: a case control study. J Crit Care 2020 Feb;55:194-197. [doi: 10.1016/j.jcrc.2019.10.013] [Medline: 31756583]
23. Arieff AI, Massry SG. Calcium metabolism of brain in acute renal failure. Effects of uremia, hemodialysis, and parathyroid hormone. J Clin Invest 1974 Feb 1;53(2):387-392. [doi: 10.1172/jci107571]
24. Toft K, Tontsch J, Abdelhamid S, Steiner L, Siegemund M, Hollinger A. Serum biomarkers of delirium in the elderly: a narrative review. Ann Intensive Care 2019 Jul 01;9(1):76 [FREE Full text] [doi: 10.1186/s13613-019-0548-1] [Medline: 31263968]
25. Khan BA, Perkins AJ, Prasad NK, Shekhar A, Campbell NL, Gao S, et al. Biomarkers of delirium duration and delirium severity in the ICU. Crit Care Med 2020 Mar;48(3):353-361 [FREE Full text] [doi: 10.1097/CCM.0000000000004139] [Medline: 31770149]
26. Khan BA, Perkins AJ, Gao S, Hui SL, Campbell NL, Farber MO, et al. The confusion assessment method for the ICU-7 delirium severity scale: a novel delirium severity instrument for use in the ICU. Crit Care Med 2017 May;45(5):851-857 [FREE Full text] [doi: 10.1097/CCM.0000000000002368] [Medline: 28263192]
27. Terrando N, Eriksson LI, Ryu JK, Yang T, Monaco C, Feldmann M, et al. Resolving postoperative neuroinflammation and cognitive decline. Ann Neurol 2011 Dec 21;70(6):986-995 [FREE Full text] [doi: 10.1002/ana.22664] [Medline: 22190370]
28. Feng X, Valdearcos M, Uchida Y, Lutrin D, Maze M, Koliwad SK. Microglia mediate postoperative hippocampal inflammation and cognitive decline in mice. JCI Insight 2017 Apr 06;2(7):e91229 [FREE Full text] [doi: 10.1172/jci.insight.91229] [Medline: 28405620]
29. Fu X, Wang L, Wang G, Liu X, Wang X, Ma S, et al. Delirium in elderly patients with COPD combined with respiratory failure undergoing mechanical ventilation: a prospective cohort study. BMC Pulm Med 2022 Jul 09;22(1):266 [FREE Full text] [doi: 10.1186/s12890-022-02052-5] [Medline: 35810306]
30. Won YH, Cho YS, Joo SY, Seo CH. Respiratory characteristics in patients with major burn injury and smoke inhalation. J Burn Care Res 2022 Jan 05;43(1):70-76. [doi: 10.1093/jbcr/irab085] [Medline: 34142710]

XSL•FO
RenderX

31. Żwierełło W, Piorun K, Skórka-Majewicz M, Maruszewska A, Antoniewski J, Gutowska I. Burns: classification, pathophysiology, and treatment: a review. Int J Mol Sci 2023 Feb 13;24(4):3749 [FREE Full text] [doi: 10.3390/ijms24043749] [Medline: 36835171]
32. van den Boogaard M, Pickkers P, Slooter AJ, Kuiper MA, Spronk PE, van der Voort PH, et al. Development and validation of PRE-DELIRIC (PREdiction of DELIRium in ICu patients) delirium prediction model for intensive care patients: observational multicentre study. BMJ 2012 Feb 09;344(feb09 3):e420 [FREE Full text] [doi: 10.1136/bmj.e420] [Medline: 22323509]
33. Chen Y, Du H, Wei BH, Chang XN, Dong CM. Development and validation of risk-stratification delirium prediction model for critically ill patients: a prospective, observational, single-center study. Medicine (Baltimore) 2017 Jul 17;96(29):e7543-e7590 [FREE Full text] [doi: 10.1097/MD.0000000000007543] [Medline: 28723773]
34. Green C, Bonavia W, Toh C, Tiruvoipati R. Prediction of ICU delirium: validation of current delirium predictive models in routine clinical practice. Crit Care Med 2019 Mar;47(3):428-435. [doi: 10.1097/CCM.0000000000003577] [Medline: 30507844]

Abbreviations

AdaBoost: adaptive boosting
AUC: area under the receiver operating characteristic curve
CAM-ICU: Confusion Assessment Method for the Intensive Care Unit
GBM: gradient-boosting machine
ICDSC: Intensive Care Delirium Screening Checklist
ICU: intensive care unit
IL: interleukin
LDA: linear discriminant analysis
LR: logistic regression
MCC: Matthews correlation coefficient
MetHb: methemoglobin
PRE-DELIRIC: Prediction of Delirium in Intensive Care Unit Patients
PT: prothrombin time
RF: random forest
SHAP: Shapley additive explanations
SVM: support vector machine
T-bil: total bilirubin

Edited by A Mavragani; submitted 08.08.24; peer-reviewed by J Nalepa; comments to author 19.09.24; revised version received 06.01.25; accepted 03.02.25; published 05.03.25.

Please cite as:
Esumi R, Funao H, Kawamoto E, Sakamoto R, Ito-Masui A, Okuno F, Shinkai T, Hane A, Ikejiri K, Akama Y, Gaowa A, Park EJ, Momosaki R, Kaku R, Shimaoka M
Machine Learning–Based Prediction of Delirium and Risk Factor Identification in Intensive Care Unit Patients With Burns: Retrospective Observational Study
JMIR Form Res 2025;9:e65190
URL: https://formative.jmir.org/2025/1/e65190
doi:10.2196/65190
PMID:39895101

©Ryo Esumi, Hiroki Funao, Eiji Kawamoto, Ryota Sakamoto, Asami Ito-Masui, Fumito Okuno, Toru Shinkai, Atsuya Hane, Kaoru Ikejiri, Yuichi Akama, Arong Gaowa, Eun Jeong Park, Ryo Momosaki, Ryuji Kaku, Motomu Shimaoka. Originally published in JMIR Formative Research (https://formative.jmir.org), 05.03.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Race and Ethnicity in Facebook Images and Text: Thematic Analysis

Shaniece Criss[1], ScD; Sarah M Gonzales[1], BS; Heran Mane[2], BS; Katrina Makres[2], MPH; Dalmondeh D Nayreau[1], BA; Vaishnavi Bharadwaj[1]; Hannah G Kim[1]; Thu T Nguyen[2], ScD

[1]Department of Health Sciences, Furman University, Greenville, SC, United States

[2]School of Public Health, Epidemiology and Biostatistics, University of Maryland, College Park, MD, United States

Corresponding Author:
Shaniece Criss, ScD
Department of Health Sciences
Furman University
3300 Poinsett Hwy
Greenville, SC, 29613
United States
Phone: 1 8642943606
Email: shaniece.criss@furman.edu

Abstract

Background: Social media platforms, such as Facebook, provide a dynamic public space where users of various racial and ethnic backgrounds share content related to identity, politics, and other social issues. These platforms allow racially minoritized groups to both challenge racial silencing and express cultural pride. At the same time, they expose users to racism and stereotypes that can negatively affect their mental and physical health through psychosocial stress. Given the rise of multimodal communication, it is essential to study both images and text to fully understand how race and ethnicity are discussed in digital spaces.

Objective: This exploratory, descriptive study aimed to investigate how people discuss race and ethnicity on Facebook and specifically examine themes related to cultural pride, solidarity, racism, antiracism, and politics using qualitative content analysis of race- and ethnicity-related Facebook posts with images and text. These themes reflect how individuals construct identity, engage with other social identities, and navigate sociopolitical discourse in digital spaces.

Methods: We conducted a qualitative content analysis using a hybrid inductive-deductive approach. A total of 500 multimodal Facebook posts were randomly sampled using CrowdTangle, with 100 posts from each year between 2019 and 2023. Each post included both image and text and contained at least 1 race- or ethnicity-related keyword. Posts were uploaded to GitHub for storage and to Label Studio for coding. An iteratively developed codebook guided the analysis, focusing on representations of race and ethnicity, the continuum of race-related discourse, and topical content. All posts were double coded until an 80% interrater agreement was reached. The remaining discrepancies were resolved through coder consensus to ensure reliability and consistency. Themes were solidified through thematic analysis.

Results: Across 500 Facebook posts from 2019 to 2023, nearly one-third lacked clear racial specificity, with 19.8% (99/500) unrelated to race and 11.2% (56/500) mentioning no specific racial or ethnic group. Among the identified groups, Hispanic, multiracial, and immigrant communities were the most frequently referenced. Common themes included US politics, cultural pride, racism and stereotypes, and antiracism. Political content was the most crosscutting theme, while cultural pride and racism-related discourse varied by group. Antiracism posts reflected the national response to racial justice movements. These findings highlight the nuanced and evolving nature of race-related discourse on social media.

Conclusions: It can be complicated to interpret image-based posts because of the subtle ways in which an image may reference race and ethnicity but does not explicitly mention it, or when there is a contradiction in the ideas portrayed in the image versus the text. Decoding this process on Facebook can help researchers boost the positive impacts and reduce the harmful effects of racism on social media.

(JMIR Form Res 2025;9:e62713) doi:10.2196/62713

KEYWORDS

Facebook; race; ethnicity; racism; antiracism; cultural pride; solidarity; qualitative research

XSL•FO
RenderX

Introduction

Race and Social Media

A total of 68% of adults in the United States reported that they have used Facebook at some point, with the racial and ethnic breakdown of users being 69% White, 67% Asian, 66% Hispanic, and 64% Black [1]. Facebook activity varies by race, with racially minoritized groups practicing more content creation and engagement related to race and ethnicity compared to White users and engaging more in the explicit self-presentation of racial and ethnic identities as salient [2-5]. This pattern of online content engagement creates resistance to the racial silencing of marginalized groups by dominant color-blind ideologies of broader society because racially minoritized groups can actively lead or participate in race- and ethnicity-related discourse [4,5].

However, while social media can provide opportunities to resist racial silencing and promote the voices of marginalized groups, it can also be used to propagate harmful and racist ideologies. Racism and the resulting stress for marginalized groups operate on intrapersonal, interpersonal, and organizational levels, and interpersonal racism can be disseminated through and observed on social media [6]. Vicarious racism, or awareness of racism against one's friends, family, or racial group, creates stress and negatively impacts both physical and mental health outcomes [6,7]. Vicarious racism can also be mediated by social media. Racism has profound effects on the health outcomes of marginalized groups and is not only thought to correlate with stress and poorer health outcomes but is also thought to be a causal factor.

Engaging with race- and ethnicity-related content on social network sites is influenced by the anonymity or nonanonymity of the site, with Facebook predominantly being the latter [8]. The anonymity of social network sites allows users to more openly perpetuate hostility and negative impulses, whereas users of nonanonymous sites present more socially desirable traits of themselves [5,8]. Facebook's nonanonymous status and the representation of racially minoritized groups on the platform provide support for the spiral of silence theory, which suggests that when faced with increasing social pressures, individuals with unpopular perspectives withhold their opinions (eg, racist ideals) due to fear of social isolation [8]. However, recent research suggests the spiral of silence theory is weakened by social media because nonanonymous Facebook users can engage in subtle or covert racism instead of overt racism in order to minimize social isolation [6,8]. It is also important to note that social media users are exposed to a multitude of content, especially within online echo chambers, which may disrupt user perception of what is considered popular or unpopular. User disorientation further weakens the spiral of silence theory and the formation of extreme views regarding race online may influence offline perceptions of racism. The racial formation theory holds that racial categories can be created, destroyed, and transformed by society; hence, social media posts that reinforce racist or stereotypical ideologies can spread a negative racial concept among social media users [9,10]. Both racial formation theory and the subversion of the spiral of silence theory guided the research team to further investigate racism as a major content theme.

Social media users are encouraged to share unpopular opinions when they perceive that their opinions are correct, believe that the topic of engagement is of heightened importance, or hold a particular personal concern, regardless of social pressures [8]. As a result, while racist discourse may still be prominent on social media, there remains a high potential to counter hateful content through collective calibration; that is, norm setting and policing harmful content [6,8]. Finally, the nature of social media networks, as characterized by racial homophily and a higher percentage of weak or strong ties, also has implications for engagement with ethno-racial content [2]. Racially minoritized groups report having weaker ties, usually for building social capital, which can lead to less intense race engagement. White individuals, by contrast, report having stronger ties, which means ethno-racial content creation or engagement happens within a reference group with similar values, perspectives, and a sense of solidarity [2,11]. There is a gap in the literature regarding the potential use of social media platforms to form stronger ties within minoritized groups over time, such as the cultivation of a sense of solidarity, pride, or other common values.

Images in Social Media Analysis

Although images and image-based content on social media platforms circulate routinely, their contribution to social media communication can extend beyond entertainment. Expression of personal opinion, community engagement, and support for public causes are increasingly seen in image-based content, to the extent that videos and images are as common as written posts on social media [12,13]. The uses of social media, heightened by the use of images, contribute to the creation of collective group identity as well as group solidarity [14,15]. This makes social media platforms, such as Facebook, X (formerly known as Twitter), and Instagram, powerful settings for content creation and the dissemination of messages on diverse subjects and groups of people.

Image-based content on Facebook, such as memes, may contain implicit and explicit messages of racism or hate, frequently influenced by stereotypes of marginalized groups. However, memes often convey negative sentiments toward marginalized groups in a mitigated or ironic manner, often incorporating sarcasm [13,16]. Humor framing helps mask their hateful intent and avoid detection by content moderators, which aids their widespread distribution [17,18]. Unfortunately, the difficulty in detecting racism in image-based social media posts can perpetuate stereotypes and intolerance. For instance, Windisch and Simi [15] found that sharing humorous images with undertones of racism facilitated a sense of solidarity among White supremacists in a more entertaining manner than hate speeches or publications. The use of irony and humor in memes can allow social media users to express hateful sentiments toward marginalized groups without risking major repercussions. Therefore, image-based content can reveal common racial sentiments and public attitudes about stereotypes, which are multifaceted [19]. In addition, the viewer's interpretation of

XSL•FO
RenderX

posts is subjective and based on their personal identities, making users' intentions difficult to identify.

Gaps and Study Aim

There is a need to clearly describe qualitative content themes in Facebook posts regarding race and ethnicity [20,21]. Conventional studies and content moderation have often relied on text-based content. However, the advent of social media platforms rich with multimodal content has underscored the need to look beyond text and incorporate other modalities, such as images, to capture a more comprehensive and nuanced understanding of trends and sentiments expressed on these platforms. Using qualitative content analysis of race- and ethnicity-related Facebook posts with images and text, this exploratory, descriptive study seeks to investigate how people discuss race and ethnicity on Facebook and specifically examine themes related to cultural pride, solidarity, racism and antiracism, and politics. These themes reflect how individuals construct identity, engage with other social identities, and navigate sociopolitical discourse in digital spaces.

Methods

Sample

The relevant data include Facebook posts with images and text. For the remainder of the paper, the data will be referred to as posts. The study sample consisted of posts with terms related to race and ethnicity posted from 2019 to 2023. The data were collected using CrowdTangle, a public insight tool by Meta [22,23]. CrowdTangle tracks publicly available social media posts made by pages and groups covering 2 million public Instagram accounts, about 20 thousand of the most active subreddits, and over 7 million pages and groups [22,23]. For this study, a sample of 100 randomly selected Facebook posts per year that contain 1 or more race-related key terms was analyzed. Beyond setting the date range, the collection was further refined by restricting the data to posts in English, from Facebook pages that were primarily managed in the United States, and where at least 1 keyword from a race- and ethnicity-related keyword list (Multimedia Appendix 1) appeared in the post text or corresponding image text. To generate the keyword list, the research team used an online database of racial slurs [24], as well as a keyword list from a study that evaluated race-related online discourse [25] and categorized it by racial and ethnic groups. For this study, posts were randomly selected, specifically 100 Facebook posts per year that had images and that contained 1 or more race-related key terms.

Data Management

For each year, a file was created containing the image URLs and their corresponding texts through GitHub. Subsequently, the data were uploaded to Label Studio Enterprise, where a project was created for each year, and the files containing the images, URLs, and texts were uploaded. The Label Studio platform offers a customizable and user-friendly interface that enables annotators to code efficiently and conduct interrater reliability through percent agreement [26].

Codebook Development

The initial draft of the codebook was informed by the literature reviewed in the introduction, which highlights key frameworks and findings related to race and racism in digital and social media contexts. Drawing from this foundational literature, we identified initial constructs and categories for analysis, including racism, antiracism, solidarity, and various racial and ethnic group representations. The research team used a hybrid inductive-deductive approach by applying concepts from the literature while also refining categories through close reading and iterative coding of posts from our sample. The research team labeled 50 posts together and made revisions based on the piloted posts. Next, 2 researchers coded 100 posts, and the team met to make revisions to the codebook. The final codebook consisted of 3 major coding dimensions, as presented in Textbox 1.

XSL•FO
RenderX

Textbox 1. Coding dimensions.

Race and ethnicity referenced

- Arab, Middle Eastern, Muslim, Asian, Black, Hispanic, indigenous, Immigrants, Jewish, White, discussed Black and White, multiple races, or no race specified.
- Each post was categorized based on which racial or ethnic groups were explicitly or implicitly referenced in either the image or the accompanying text. These categories were selected to capture the diverse ways in which race and ethnicity appeared across the dataset, whether through direct identification, symbolic representation, or discursive framing. Posts that mentioned more than one racial group were coded as "discussed Black and White" or "multiple races, multiethnic, or mixed," depending on the content. Posts that did not refer to any specific group but still addressed racial issues were coded as "no specific race mentioned." This approach allowed for both specificity in capturing group-related discourse and flexibility in acknowledging the complex, intersectional nature of many posts.

Continuum of race-related discourse

- Racist, antiracist, solidarity, and not racist.
- Posts were evaluated by the continuum of race-related discussions with classifications of racist posts, antiracism, solidarity, and posts that were not racist. *Racist* posts included aspects of discrimination, stereotyping, criticism of critical race theory, and slavery. The *antiracist* posts featured condemning racial discrimination, antihate, pointing out hypocrisy, and acknowledging White privilege. The *solidarity* response option included posts with aspects of cultural pride, representation, or unifying events. *Not racist* posts had no obvious connection to racism.

Topic of the post

- For example, cultural pride, immigration, religion, US politics, military, food, daily life, racism, and antiracism.
- There were a multitude of topics. *Cultural pride* consisted of positive cultural sentiments, including promoting civil rights and attending events. The *immigration* code was assigned to posts related to residents of other countries entering the United States, with documentation and without documentation. *Military* posts related to the US military, veterans, or military history. *Religion* referred to posts about beliefs, practices, and communities associated with religion, spirituality, and prayers. *US politics* included posts about national politics and politicians (eg, Kamala Harris, Ketanji Jackson-Brown, and Donald Trump), as well as local government and policies. Posts were coded as *global politics* if they mentioned foreign relations or politics in other countries. Posts about *food* included advertisements and admiration of food. *Racism* referred to posts that discussed discrimination, stereotypes, and offenses related to race and ethnicity. *Antiracism* referred to posts that focused on denouncing racist ideologies and fighting against inequitable treatment.

Data Coding and Analysis

Extensive multisession discussions about the posts were held. All posts were double coded within Label Studio [26], and the team met to ensure that at least 80% interrater reliability was reached by calculating the percent agreement. After the initial coding process was complete, 2 coders met to reach a 100% consensus within the entire dataset. Once the complete dataset had the same coding designation, a thematic analysis was conducted [27]. Each researcher individually identified themes in the data, and as a group, solidified the final themes in the data, using the various perspectives of the different racial backgrounds and life experiences of the study team.

Ethical Considerations

This study was determined not to be human participant research by the University of Maryland, College Park institutional review board (2072551-1). In addition, the social media posts were anonymized, upholding user privacy.

Results

Table 1 provides count data on the frequencies of posts mentioning various racial and ethnic categories in 2019, 2020, 2021, 2022, and 2023 in the sample. Table 1 shows the total percentage of the 500 posts categorized by racial and ethnic groups over the 5-year period. Notably, the 2 most frequent classifications were posts coded as "not relevant" to race or ethnicity, which made up 19.8% (99/500), and posts that referenced no specific racial or ethnic group, accounting for 11.2% (56/500). These 2 categories together comprised nearly one-third of the dataset, highlighting the prevalence of generalized or ambiguous racial discourse in the sample. Among posts with identifiable racial or ethnic references, the most frequently mentioned groups were Hispanic (n=54, 10.8%), multiple races, multiethnic, or mixed (n=53, 10.6%), and immigrants (n=48, 9.6%). Posts referencing Black individuals or communities accounted for 9% (n=45) of the posts, followed by Asian (n=38, 7.6%) and Arab, Middle Eastern, and Muslim groups (n=32, 6.4%). Less commonly represented groups included White (n=26, 5.2%), Jewish (n=20, 4%), and Indigenous or Native American (n=10, 2%). Posts that discussed both Black and White racial groups simultaneously made up 3.8% (n=19) of the data. Race-related language was present in a significant number of public Facebook posts. Explicit identification of racial or ethnic groups was inconsistent, with many posts either lacking specific group references or being tangentially related to race.

XSL•FO
RenderX

Table 1. Count and percentage sample of Facebook posts referencing different race and ethnicity categories by year (2019-2023).

Race and ethnicity category	2019 (n=100), n	2020 (n=100), n	2021 (n=100), n	2022 (n=100), n	2023 (n=100), n	Over the 5-year period (N=500), n (%)
Arab, Middle Eastern, and Muslim	2	5	17	5	3	32 (6.4)
Asian	7	8	5	7	11	38 (7.6)
Black	6	11	9	9	10	45 (9)
Hispanic	7	15	11	14	7	54 (10.8)
Immigrants	16	8	10	6	8	48 (9.6)
Indigenous or Native American	1	2	2	1	4	10 (2)
Jewish	7	2	1	3	7	20 (4)
White	4	3	6	5	8	26 (5.2)
Discussed Black and White	5	4	5	4	1	19 (3.8)
Discussed multiple races (multiethnic or mixed)	16	16	8	6	7	53 (10.6)
No specific race mentioned	13	11	11	11	10	56 (11.2)
Not relevant	16	15	15	29	24	99 (19.8)

Table 2 presents common content themes across racial and ethnic categories, reflecting the underlying narratives, sentiments, and issues most frequently associated with each group across the 2019 to 2023 sample. These themes emerged from the topic codes in the codebook and were shaped through an interpretive analysis of the narrative patterns within each racial and ethnic group. US politics was the most prevalent crosscutting theme, appearing in posts associated with 7 (64%) of the 11 racial and ethnic categories: Arab, Middle Eastern, Muslim, Black, Hispanic, immigrants, White, discussed Black and White, and no specific race mentioned. Political content typically referenced elections, immigration policy, public figures, and social justice debates, often with explicit or implied racial framing. Cultural pride emerged as a prominent theme in 4 categories: Asian, Black, Hispanic, and Indigenous or Native American. These posts frequently celebrated cultural heritage, community events, and milestones in representation, such as historic appointments or honors. Posts highlighting racism and stereotypes were especially prominent in content referencing Black, Hispanic, and multiracial identities, often in the context of public discourse on systemic racism, xenophobia, and media portrayals. These posts spiked during moments of national attention to police violence, border policies, and racialized political rhetoric. The theme of antiracism appeared distinctly in posts referencing Black individuals, discussions involving both Black and White groups, and posts where no specific race was mentioned. These posts included calls to action, personal reflections, and educational content, with a notable peak in frequency during the summer of 2020, coinciding with widespread mobilizations following the killing of George Floyd. These patterns display how racial and ethnic discourse on Facebook is shaped by both identity-specific issues and broader societal events that cut across groups.

XSL•FO
RenderX

Table 2. Common content themes by race and ethnicity category in the sample (2019 to 2023)[a].

Race and ethnicity category	Over the 5-year period (N=500), n (%)	Common content themes (2019-2023)
Arab, Middle Eastern, and Muslim	32 (6.4)	US military deaths in Afghanistan, religion, and hijabs
Asian	38 (7.6)	Cultural pride, food, COVID-19, and stereotypes
Black	45 (9)	Racism, antiracism, US politics, cultural pride, Black history, Kamala Harris, and Ketanji Brown Jackson
Hispanic	54 (10.8)	Food, US politics, United States-Mexico border, immigration, stereotypes, Donald Trump, and cultural pride
Immigrants	48 (9.6)	US politics, politics for and against immigration, refugee and immigrant crises, and anti-immigrant sentiments
Indigenous or Native American	10 (2)	Cultural pride, Native American proverbs and blessings, and events and celebrations
Jewish	20 (4)	Solidarity—Christians and Jews and support for Jewish people
White	26 (5.2)	US politics, association with the United States. Republican party, antiracism, daily life, humor, and sarcasm—“redneck”
Discussed Black and White	19 (3.8)	US politics, antiracism, and addressing historical and current racism—global and the United States
Discussed multiple races	53 (10.6)	Conflict between races (eg, historical racism, double standards, “human race,” or colorblindness)
No specific race mentioned	56 (11.2)	Racism, antiracism, US politics, and political commentary

[a]Sample sizes for each race and ethnicity category are based on the total number of relevant Facebook posts coded from 2019 to 2023 (Table 1). Not relevant posts accounted for 19.8% (99/500) of the total posts.

To further illustrate how these themes were expressed in the data, The following sections provide representative examples of posts aligned with the 4 major themes: cultural pride, racism and stereotypes, antiracism, and US politics.

Textbox 2 highlights posts expressing cultural pride, often through celebration of heritage, historical milestones, or representation in national media.

Textbox 3 displays posts reflecting racism and stereotypes, including explicit or coded language that reinforces bias or discrimination, often masked in humor.

By contrast, Textbox 4 contains examples of posts incorporating antiracism, as evidenced by their condemnation of racist ideologies and views on inequitable treatment.

Finally, Textbox 5 contains examples of posts referring to various issues related to US politics, including commentary on elections, immigration policy, racial justice, and political figures.

Textbox 2. Examples of cultural pride memes in Facebook posts (2019-2023).

- This victory is special for Hideki.

- Congratulations to the Honorable Ketanji Brown Jackson who was just confirmed to the Supreme Court of the United States in a 53-47 Senate vote! Once she’s sworn in, Justice Jackson will become the first Black woman on the highest court in the nation in its 232 year history. There have been 115 Justices in this country. 108 white men, 5 white women, and 2 Black men. I’m thrilled to witness this moment. #BlackGirlMagic #SupremeCourtJustice #Herstory #KetanjiBrownJackson #BlackFactsxARYE

- Cuban American artist Celia Cruz will be the first Latina singer featured as part of the American Women Quarters Program, the US Mint announced. Read more: https://cnn.it/3YwPaDX:=:https://www.cnn.com/style/article/celia-cruz-us-quarter/index.html

- Next weekend, Cherokee Nation will join the Eastern Band of Cherokee Indians and the United Keetoowah Band of Cherokee Indians to share the story of the Cherokee people. The tribes host Cherokee Days at the Smithsonian’s National Museum of the American Indian in Washington, DC. For more information, visit CherokeeDays.com.

XSL•FO
RenderX

Textbox 3. Examples of racism and stereotypes memes in Facebook posts (2019-2023).

- Life hacks
- Southern Breeze
- Poll: Many Americans Know Illegals Create Risk of Mass Casualty Internal Attack on US Ranging from 9/11 Style To A More Systematic Nationwide Uprising
- Tik tok is wild...you just open the app and wham someone is lecturing you on a random non issue they made up in their head while combing a double ended dildo through their hair edit: the comment section is a dumpster fire and thats specificly because tiktok posts are curated to be beacons for anger dumping. thats the whole app meta concept.

Textbox 4. Examples of antiracism memes in Facebook posts (2019-2023).

- The two failings America has that are at the core of our destruction are racism and misogyny. If we have a movement that exposes how wrong that is and how oneness can save US as a society our world would be a much better place. Start by rejecting those bigotries by VOTING out anyone who espouses those beliefs. 2020
- Grow [antiracism: Fixed to Growth Mindset Fixed-Comfort: “I don’t know where to start or what to say.” Growth-Courage: “First I will listen/read/watch. I will speak against injustice.” Fixed-Comfort: “I don’t want to get it wrong or get called out.” Growth-Courage: “I will make mistakes, no doubt about it. I will be grateful for the lesson.” Fixed-Comfort: “It won’t make a difference what I do. Nothing is going to change.” Growth-Courage: “Things happen when I take risks and become part of something bigger.” Fixed-Comfort: “I don’t get involved in politics. I don’t have time.” Growth-Courage: “This is a human rights issue. This matters. I will make time.”]
- Your belief that you haven’t personally benefited from white privilege doesn’t mean white privilege doesn't exist. (What people mostly mean when they make statements like “white privilege never helped me!” Is that they weren’t economically well off enough to benefit from class privilege, too.;
- None of us are free, until all of us are free.

XSL•FO
RenderX

Textbox 5. Examples of US politics memes in Facebook posts (2019-2023).

- Truth

- [Chris]

- Shameful and disgusting.

- #BREAKING — Four Proud Boys members found guilty of seditious conspiracy in Jan. 6 trial — https://nbcnews.to/42rUoCu:=:https://www.nbcnews.com/politics/justice-department/jury-reaches-verdict-proud-boys-seditious-conspiracy-trial-rcna81129 This is #BreakingNews worthy.

Discussion

Principal Findings

The rapid advancement of technology has profoundly reshaped social media platforms and the features they offer, creating a dynamic and constantly evolving landscape. Communication on these platforms is increasingly multimodal, incorporating various models such as text, images, and videos, which not only enhances the richness of content but also increases its interpretive complexity. This evolution necessitates a more intricate and critical understanding of how to engage with online content beyond traditional text-based evaluations, enabling a more nuanced interpretation of cultural and sentiment trends.

In this study, we qualitatively explored Facebook posts related to race and ethnicity, highlighting how images provide a cultural context and subtle cues that text alone cannot convey. This study advances previous literature by explicitly integrating multimodal analysis, which reveals complex, layered meanings in racial discourse often overlooked in text-only studies. For example, the nuanced use of humor and sarcasm in images demonstrates how stereotypes can be both masked and reinforced in ways that complicate interpretation. Furthermore, this work highlights a sustained expression of cultural pride across multiple racial and ethnic groups outside of crisis moments. This finding broadens the understanding of how identity affirmation operates continuously on social media, rather than solely during episodic activism. Finally, by mapping thematic content across a 5-year span and multiple groups, the study captures the interplay between identity-specific concerns and wider societal events shaping discourse on Facebook instead of examining one point in time.

Chronology of Posts and Germane Current Events

The research team described qualitative content themes in Facebook posts regarding race and ethnicity, providing an overview of the landscape of image-based content creation and engagement. The qualitative content themes that were highly represented throughout the sample and across multiple race and ethnicity categories included US politics, cultural pride, racism, and antiracism. Approximately 19.8% (99/500) of the study sample was not relevant as they did not contain explicit or allusions to any race- and ethnicity-related information based on the research team's ability to analyze the post. These posts may have appeared in the sample due to words having different meanings based on the context of the posts, such as sanctuary in terms of immigration versus animal sanctuaries.

The years 2019 to 2023 were an important time period for race-related discussions. Some of the most consequential moments include national discussions of immigration (eg, Trump's administration immigration policies), the killing of George Floyd and a rise in the Black Lives Matter movement, increased racism against Asians being blamed for the COVID-19 pandemic, and sharp racial health disparities in the pandemic outcomes, among other cultural events that occurred [28-32]. The increase in mentions of Arab, Middle Eastern, and Muslim-related terms in 2021 may have been associated with the 20th anniversary of the September 11, 2001, terrorist attacks (commonly 9/11) in the United States [33].

Cultural Pride

Social media can influence public perceptions on a variety of issues, and the tone of posts can reinforce or challenge common perspectives [34,35]. The findings of the study indicate that posts with a predominantly positive or neutral tone often included aspects of cultural pride (Textbox 2). Posts, such as the "Hideki" and the "Cuban American artist Celia Cruz" in Textbox 2, portray Hideki (of Japanese ancestry) and Cuban Americans as resourceful and talented. Studies have found that the use of a positive tone in posts elicits a positive response from social media users, who are more receptive to the message and likely to respond in a positive manner as a result [36]. Expression of cultural pride through image-based content can not only allow ethnic groups to denounce color-blind ideologies through racial silencing but can also foster a greater sense of self-esteem and competence among racial minoritized groups [5].

Previous research on social media found a similar theme regarding cultural pride. One study examining ethnic media

XSL•FO
RenderX

found that ethnic social media use is often associated with increased ethnic pride and performance, unlike mainstream media use, which is associated with low self-esteem for ethnic users [37]. Other studies have found that individuals belonging to racially minoritized groups resist racial silencing through posts associated with cultural pride [5,38]. While these studies analyze user profiles, a study examining themes on #BlackLivesMatter tweets (regardless of user identity) found that while many posts expressed negative sentiments, posts with references to pride, optimism, and positive emotions were also prominent [39]. These results are similar to the study's findings that positivity and cultural pride constituted a substantial portion of Facebook image-based content that contained terms related to race and ethnicity.

Racism and Stereotypes

Often, in the absence of images, the text alone may appear neutral or benign. However, the addition of an image can dramatically alter the perspective of the text, bringing to light the true message and intention of a post. A common theme from the study found that Facebook posts perpetuate the stereotyping of racially minoritized groups. Results indicate that humor masking is often used to conceal hateful intentions and stereotyping, as also seen in other studies [18-20]. For instance, in Textbox 3, the phrase "life hacks" alone lacks interpretability without additional context. However, the inclusion of an image showing cheat sheets glued to fingernails, along with the image text "cheating level: Asian," reveals a deeper meaning. The post subtly conceals and reinforces the typical "smart Asian" stereotype with humor, thus aiding the widespread distribution of such posts. Prior studies have also found, consistent to the findings in this study, that posts containing aspects of stereotyping or racism also tend to have neutral tones (as seen in the "poll: many Americans know..." post in Textbox 3), so that their opinionated content can be classified as "informational" or even as a positive stereotype [40]. Positive stereotypes, as in the "cheating level: Asian" post, attribute the success of racially minoritized groups to their racial and ethnic identity rather than their abilities and strengths [40].

Masked stereotyping can create false and negative perceptions of racially minoritized groups, as suggested by the racial formation theory [9,10]. According to cultivation theory, long-term exposure can distort users' views of what topics relate to different races and their perception of others' stances in relation to those topics, possibly working synergistically with social media's undermining of the spiral of silence theory. In addition to the masking of racist or stereotypical posts, social media makes it easy to spread hateful messages through online communication (rather than in-person conversation) and the presence of like-minded online acquaintances, which creates a clustering of supporters of racist and stereotypical ideologies [41,42]. This both facilitates and normalizes racial microaggression on social media, which can create incorrect and hate-filled concepts of marginalized groups, as postulated by the racial formation theory [43].

Furthermore, this normalization perpetuates stress for targeted groups [3] due to the difference between how members of racially minoritized groups perceive their own identity and how that identity is perceived and expressed by others. In turn, the compounding negative effects on mental health could adversely affect physical health and social, economic, and political outcomes for marginalized groups, thus serving as a social determinant of health [6,7]. Therefore, it is necessary for social media users to be familiar with the different ways stereotypical and racist ideologies can manifest themselves through posts, especially if those ideologies and hateful intentions are masked in image-based content.

Antiracism

Many antiracist posts on social media include a call to action, as can be seen in Textbox 4 (ie, the "fixed to growth mindset" and "none of us are free..." images) [44-46]. The call to action usually encourages the reader to decentralize authority and challenge implicit racism embedded in the status quo. Disseminating antiracist content via social media can provide a great opportunity to promote antiracism and influence real-life behavior on both interpersonal and intrapersonal levels. Interpersonally, people moved by social media posts are encouraged to oppose racism in everyday life [45]. Antiracism engagement on social media has positive mental health effects, such as decreased depression and anxiety, in Black, White, and Hispanic people [47]. Many antiracist posts use sarcasm and humor to convey their messages. So, while sarcasm is often used to mask racist ideology and vent negative emotions, it can also be used to promote antiracist ideals and have positive effects on mental health [13,48].

Social Media and Political Views

Social media plays an important role in shaping a nation's collective political will, or the will of a subgroup within a country [34]. For instance, disinformation often contributes to anti-immigrant racism. Neutral and informative content, such as the "breaking news" image in Textbox 5, cannot replace traditional media as a source of political knowledge. In its current state, social media reporting is inadequate to dispel disinformation [49]. Furthermore, social media frequently normalizes expressing anti-immigrant opinions not held by the majority, weakening the spiral of silence theory [8]. The normalization of anti-immigrant fervor also strengthens the cultivation theory, as users' views of how others respond to issues, such as immigration, are distorted by long-term media exposure.

Once a consensus is reached through various means, many collective political wills make up the political landscape. Politicians gauge what views are popular and monitor their relations with the public using social media [50,51]. The posts captioned "[Chris]" and "shameful and disgusting" in Textbox 5 are examples of social media posts that represent user views of politicians and political parties. However, the political landscape and its constituent collective wills are ephemeral. Social media may initially draw attention to current events that encourage racism or antiracism, but interests can be transitory [52,53].

Strengths and Limitations

This study offers several strengths that contribute to its overall rigor and relevance. First, the inclusion of both images and text,

XSL•FO
RenderX

rather than relying on text alone, provided a more holistic view of how race and ethnicity were represented on Facebook. Second, all posts were double coded with an interrater agreement target of at least 80%, followed by consensus coding to achieve 100% alignment, ensuring a high level of reliability and consistency throughout the analysis. Finally, the 5-year time span of the dataset enabled the study to capture trends across key sociopolitical moments.

The study also had several limitations. The sample consisted of 100 posts per year, which allowed for manageable and detailed analysis, but a larger sample size may have revealed a wider range of perspectives and discourse patterns. With 500 posts, the sample size limited generalizability, particularly for specific race and ethnicity categories. For example, the "Indigenous or Native American" category included 10 (2%) posts, so readers should be cautioned against drawing strong conclusions from small subgroups. Facebook is known to overrepresent racially minoritized groups relative to their proportion of the US population. This representation may have influenced the content trends observed, but a larger sample might reveal different or additional thematic patterns.

To enhance representativeness, posts were randomly selected from a broad pool of publicly available Facebook content through CrowdTangle. It should be noted that the sample did not capture discussions from private groups and interpersonal conversations on Facebook. These other forums may have provided a more holistic perspective of the content. Another limitation is that Facebook may have removed very negative or harmful content before data collection. Even though these types of posts might have significant reach before deletion, they would not appear in the CrowdTangle dataset. Therefore, it is possible that the sample could have skewed more positive or affirming than the actual discourse.

Furthermore, the study relied on a keyword-based sampling strategy, which may have excluded posts with more implicit or nuanced references to race. To mitigate this, the research team developed the keyword list using both academic literature and a public database, and refined it during pilot testing. Even with an extensive keyword list, it is still possible that some posts addressing race and ethnicity did not include any of the specified keywords and were excluded. To strengthen the development of the keyword strategy and interpretation of data, the research team included members who identify as Asian, Black, and Hispanic.

Conclusions

The results from the study highlight 2 key findings. Image-based posts and content on Facebook highlight positive ideas about racial and ethnic groups, such as cultural pride and antiracism; negative ideas, such as racism; and biased posts about US politics. This study demonstrated that it is more complicated to interpret image-based posts because of the subtle ways in which an image may reference race and ethnicity but does not explicitly mention it, or when there is a contradiction in the ideas portrayed in the image versus the text. That is, an image and its accompanying text can align in ideation, or a negative image could be accompanied by a positive text that opposes it or vice versa. Given that the findings of the study describe content themes regarding race and ethnicity, further research could examine themes in posts as related to other sociodemographic factors, such as gender identity, sexual orientation, age, socioeconomic status, and intersecting identities.

Future research could also expand the scope of this study by conducting a quantitative content analysis of image-based posts (including the prevalence of certain content themes) and by conducting a more thorough analysis of posts' tone and sentiment. Subsequent studies may also find it beneficial to examine how social media users post about a specific race or ethnicity to better understand users' attitudes or intentions toward a particular marginalized group. Finally, previous studies have found that identical posts are received differently in mainstream social media and social media subcultures or social media groups centered around common cultures [54]. However, how users post about sociodemographic factors in mainstream social media, in contrast to subgroups, is an unexplored terrain that future studies could explore. It would be valuable for future studies to expand the dataset using artificial intelligence tools to classify a larger volume of posts from platforms such as CrowdTangle. In conclusion, increasing our ability to decode posts related to race and ethnicity on Facebook can help inform research and intervention to boost the positive impacts and reduce the harmful effects of racism on social media.

Acknowledgments

Research reported in this publication was supported by the National Institute on Minority Health and Health Disparities (R00MD012615 [TTN] and R01MD015716 [TTN]). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Data Availability

The datasets generated or analyzed during this study are available in the data repositories [22,23].

Conflicts of Interest

None declared.

Multimedia Appendix 1

XSL•FO
RenderX

Race terms used in keyword filtering.
[DOCX File , 25 KB - formative_v9i1e62713_app1.docx]

References

1. Schaeffer K. 5 facts about how Americans use Facebook, two decades after its launch. Pew Research Center. URL: https://www.pewresearch.org/short-reads/2024/02/02/5-facts-about-how-americans-use-facebook-two-decades-after-its-launch/?gad_source=1&gad_campaignid=22285027172 [accessed 2025-05-29]
2. Mesch GS. Race, ethnicity and the strength of Facebook ties. J Youth Stud 2017 Oct 28;21(5):575-589. [doi: 10.1080/13676261.2017.1396303]
3. Brinkman N, Jacobi L. The projection of racial identity on social network. J Intercult Commun 2020 Mar 30;20(1):52-64. [doi: 10.36923/jicc.v20i1.297]
4. Bindra VG, DeCuir-Gunby JT. Race in cyberspace: college students' moral identity and engagement with race-related issues on social media. Urban Rev 2020 Feb 27;52(3):541-561. [doi: 10.1007/S11256-020-00560-4]
5. Grasmuck S, Martin J, Zhao S. Ethno-racial identity displays on Facebook. J Comput Mediat Commun 2009;15(1):158-188. [doi: 10.1111/j.1083-6101.2009.01498.x]
6. Criss S, Michaels EK, Solomon K, Allen AM, Nguyen TT. Twitter fingers and echo chambers: exploring expressions and experiences of online racism using Twitter. J Racial Ethn Health Disparities 2021 Oct 15;8(5):1322-1331 [FREE Full text] [doi: 10.1007/s40615-020-00894-5] [Medline: 33063284]
7. Criss S, Kim M, De La Cruz MM, Thai N, Nguyen QC, Hswen Y, et al. Vigilance and protection: how Asian and Pacific Islander, Black, Latina, and Middle Eastern women cope with racism. J Racial Ethn Health Disparities 2024 Apr 14;11(2):773-782 [FREE Full text] [doi: 10.1007/s40615-023-01560-2] [Medline: 36917397]
8. Chaudhry I, Gruzd A. Expressing and challenging racist discourse on Facebook: how social media weaken the "spiral of silence" theory. Policy Internet 2019 Jan 07;12(1):88-108. [doi: 10.1002/poi3.197]
9. Daniels J. Race and racism in internet studies: a review and critique. New Media Soc 2012 Dec 10;15(5):695-719 [FREE Full text] [doi: 10.1177/1461444812462849]
10. McKay DL, Vinyeta K, Norgaard KM. Theorizing race and settler colonialism within U.S. sociology. Sociol Compass 2020 Aug 04;14(9):33. [doi: 10.1111/soc4.12821]
11. Hopkins AE. Effects of the "spiral of silence" in digital media. Inq J 2015;7(9):23. [doi: 10.4324/9781003395805-19]
12. Kligler-Vilenchik N, Thorson K. Good citizenship as a frame contest: Kony2012, memes, and critiques of the networked citizen. New Media Soc 2016 Jul 09;18(9):1993-2011. [doi: 10.1177/1461444815575311]
13. Nissenbaum A, Shifman L. Meme templates as expressive repertoires in a globalizing world: a cross-linguistic study. J Comput Mediat Commun 2018;23(5):294-310. [doi: 10.1093/jcmc/zmy016]
14. Milan S. From social movements to cloud protesting: the evolution of collective identity. Inf Commun Soc 2015 May 21;18(8):887-900. [doi: 10.1080/1369118x.2015.1043135]
15. Windisch S, Simi P. More than a joke: white supremacist humor as a daily form of resistance. Deviant Behav 2022 Mar 04;44(3):381-397. [doi: 10.1080/01639625.2022.2048216]
16. Sharma S, Alam F, Akhtar MS, Dimitrov D, Da San Martino G, Firooz H, et al. Detecting and understanding harmful memes: a survey. arXiv Preprint posted online on May 9, 2022 [FREE Full text] [doi: 10.24963/ijcai.2022/781]
17. Sălcudean M. Visual humor through internet memes (ii) from harmless humour to the discriminatory potential of (anti)memes: case study: the transgender bathroom debate. transilvania. 2020 Dec 15. URL: https://revistatransilvania.ro/wp-content/uploads/2021/01/Transilvania-11-12.2020.12-Sa%CC%86lcudean.pdf [accessed 2025-05-29]
18. Pérez R. The Souls of White Jokes: How Racist Humor Fuels White Supremacy. Stanford, CA: Stanford University Press; 2022.
19. Petsko CD, Rosette AS, Bodenhausen GV. Through the looking glass: a lens-based account of intersectional stereotyping. J Pers Soc Psychol 2022 Oct;123(4):763-787. [doi: 10.1037/pspi0000382] [Medline: 35025602]
20. Al-Ghadir AI, Azmi AM, Hussain A. A novel approach to stance detection in social media tweets by fusing ranked lists and sentiments. Inf Fusion 2021 Mar;67:29-40. [doi: 10.1016/j.inffus.2020.10.003]
21. Kawintiranon K, Singh L. Knowledge enhanced masked language model for stance detection. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologiesy. 2021 Presented at: NAACL '21; June 6-11, 2021; Virtual Event p. 4725-4735 URL: https://aclanthology.org/2021.naacl-main.376.pdf
22. Shiffman N, Silverman B. CrowdTangle opens public application for academics. Meta Research. URL: https://research.facebook.com/blog/2020/7/crowdtangle-opens-public-application-for-academics/ [accessed 2025-03-04]
23. Transparency center. CrowdTangle. URL: https://transparency.meta.com/researchtools/other-datasets/crowdtangle [accessed 2025-05-08]
24. Home page. The Racial Slur Database. URL: http://www.rsdb.org/ [accessed 2023-04-28]
25. Bartlett J, Krasodomski-Jones A. counter-speech examining content that challenges extremism online. Demos. 2015. URL: https://demos.co.uk/wp-content/uploads/2015/10/Counter-speech.pdf [accessed 2023-04-28]
26. Label studio. HumanSignal. URL: https://labelstud.io [accessed 2023-09-04]

XSL•FO
RenderX

27. Braun V, Clarke V. Thematic analysis. In: Cooper H, Camic PM, Long DL, Panter AT, Rindskopf D, Sher KJ, editors. APA Handbook of Research Methods in Psychology. Volume 2. New York, NY: American Psychological Association; 2012:57-71.
28. Pierce S. Immigration-related policy changes in the first two years of the Trump administration. Migration Policy Institute. URL: https://www.migrationpolicy.org/sites/default/files/publications/ImmigrationChangesTrumpAdministration-FinalWEB.pdf [accessed 2024-05-18]
29. Pineo R. Immigration crisis: the United States under President Donald J. Trump. J Dev Soc 2020 Jan 30;36(1):7-40. [doi: 10.1177/0169796x19896905]
30. Callister AH, Galbraith Q, Galbraith S. Immigration, deportation, and discrimination: Hispanic political opinion since the election of Donald Trump. Hisp J Behav Sci 2019 Apr 25;41(2):166-184. [doi: 10.1177/0739986319840717]
31. Parker K. Amid protests, majorities across racial and ethnic groups express support for the black lives matter movement. Pew Research Center. URL: https://www.pewresearch.org/social-trends/2020/06/12/amid-protests-majorities-across-racial-and-ethnic-groups-express-support-for-the-black-lives-matter-movement/ [accessed 2024-05-17]
32. Wood D. As pandemic deaths add up, racial disparities persist -in some cases worsen. NPR. URL: https://www.npr.org/sections/health-shots/2020/09/23/914427907/as-pandemic-deaths-ad [accessed 2025-05-29]
33. Mohamed B. Muslims are a growing presence in U.S., but still face negative views from the public. Pew Research Center. URL: https://www.pewresearch.org/short-reads/2021/09/01/muslims-are-a-growing-presence-in-u-s-but-still-face-negative-views-from-the-public/ [accessed 2024-05-20]
34. Crivellaro C, Comber R, Bowers J, Wright P, Olivier P. A pool of dreams: Facebook, politics and the emergence of a social movement. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2014 Presented at: CHI '14; April 26-May 1, 2014; Toronto, ON p. 3573-3582 URL: https://dl.acm.org/doi/10.1145/2556288.2557100 [doi: 10.1145/2556288.2557100]
35. Perloff RM. Social media effects on young women's body image concerns: theoretical perspectives and an agenda for research. Sex Roles 2014 May 29;71(11-12):363-377. [doi: 10.1007/S11199-014-0384-6]
36. Amirinasab M, Mehrazeen A, Masih AA. The tone of market participants' opinions via social media and capital market reaction. Iran J Account Audit Fin 2022;6(4):45-60 [FREE Full text]
37. Ramasubramanian S, Doshi MJ, Saleem M. Mainstream versus ethnic media: how they shape ethnic pride and self-esteem among ethnic minority audiences. Int J Commun 2017;11:52 [FREE Full text] [doi: 10.31237/osf.io/dxfe5]
38. Lee C, Chau D. Language as pride, love, and hate: archiving emotions through multilingual Instagram hashtags. Discourse Context Media 2018 Apr;22:21-29. [doi: 10.1016/j.dcm.2017.06.002]
39. Field A, Park CY, Theophilo A, Watson-Daniels J, Tsvetkov Y. An analysis of emotions and the prominence of positivity in #BlackLivesMatter tweets. Proc Natl Acad Sci U S A 2022 Aug 30;119(35):e2205767119 [FREE Full text] [doi: 10.1073/pnas.2205767119] [Medline: 35998217]
40. Gunturi US, Kumar A, Ding X, Rho EH. Linguistically differentiating acts and recalls of racial microaggressions on social media. Proc ACM Hum Comput Interact 2024 Apr 26;8(CSCW1):1-36. [doi: 10.1145/3637366]
41. Bliuc AM, Faulkner N, Jakubowicz A, McGarty C. Online networks of racial hate: a systematic review of 10 years of research on cyber-racism. Comput Human Behav 2018 Oct;87:75-86. [doi: 10.1016/j.chb.2018.05.026]
42. Keum BT, Miller MJ. Racism on the internet: conceptualization and recommendations for research. Psychol Violence 2018 Nov;8(6):782-791. [doi: 10.1037/vio0000201]
43. Williams A, Oliver C, Aumer K, Meyers C. Racial microaggressions and perceptions of internet memes. Comput Human Behav 2016 Oct;63:424-432. [doi: 10.1016/j.chb.2016.05.067]
44. Ferguson A. Redefining antiracism: learning from activists to sharpen academic language. Sociol Compass 2022 Dec 10;17(1):92. [doi: 10.1111/soc4.13057]
45. Lentin A, Humphry J. Antiracism apps: framing understandings and approaches to antiracism education and intervention. Inf Commun Soc 2016 Oct 06;20(10):1539-1553. [doi: 10.1080/1369118x.2016.1240824]
46. Deepak AC, Biggs MJ. Intimate technology: a tool for teaching anti-racism in social work education. J Ethn Cult Diversit Soc Work 2011 Jan;20(1):39-56. [doi: 10.1080/15313204.2011.545944]
47. Brannon TN. Racism hurts, can antiracism heal?: positive mental health correlates of antiracist engagement. PNAS Nexus 2023 Oct;2(10):pgad309 [FREE Full text] [doi: 10.1093/pnasnexus/pgad309] [Medline: 37799326]
48. Petray TL, Collin R. Your privilege is trending: confronting whiteness on social media. Soc Media Soc 2017 May 10;3(2):205630511770678. [doi: 10.1177/2056305117706783]
49. Shehata A, Strömbäck J. Learning political news from social media: network media logic and current affairs news learning in a high-choice media environment. Commun Res 2018 Jan 17;48(1):125-147. [doi: 10.1177/0093650217749354]
50. McGregor SC. "Taking the temperature of the room": how political campaigns use social media to understand and represent public opinion. Public Opin Q 2020;84(S1):236-256. [doi: 10.1093/poq/nfaa012]
51. Gilardi F, Gessler T, Kubli M, Müller S. Social media and political agenda setting. Polit Commun 2021 May 01;39(1):39-60. [doi: 10.1080/10584609.2021.1910390]

XSL•FO
RenderX

52. Dixon PJ, Dundes L. Exceptional injustice: Facebook as a reflection of race- and gender-based narratives following the death of George Floyd. Social Sciences 2020 Dec 15;9(12):231. [doi: 10.3390/socsci9120231]
53. Roberts J. Empathy cultivation through (Pro)social media: a counter to compassion fatigue. Journal Media 2021 Dec 10;2(4):819-829. [doi: 10.3390/journalmedia2040047]
54. Ems L, Gonzales AL. Subculture-centered public health communication: a social media strategy. New Media Soc 2016 Jul 10;18(8):1750-1767. [doi: 10.1177/1461444815570294]

Edited by A Mavragani; submitted 30.05.24; peer-reviewed by C Nguyen-Truong, M Bailey, X Ding; comments to author 20.05.25; revised version received 17.07.25; accepted 11.08.25; published 04.09.25.

Please cite as:
Criss S, Gonzales SM, Mane H, Makres K, Nayreau DD, Bharadwaj V, Kim HG, Nguyen TT
Race and Ethnicity in Facebook Images and Text: Thematic Analysis
JMIR Form Res 2025;9:e62713
URL: https://formative.jmir.org/2025/1/e62713
doi:10.2196/62713
PMID:40906523

©Shaniece Criss, Sarah M Gonzales, Heran Mane, Katrina Makres, Dalmondeh D Nayreau, Vaishnavi Bharadwaj, Hannah G Kim, Thu T Nguyen. Originally published in JMIR Formative Research (https://formative.jmir.org), 04.09.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Sexual Health Self-Management Intervention (Psychosexual Educational Partners Program) for Couples With a History of Breast and Gynecological Cancer: Mixed Methods Feasibility Study

Noël M Arring[1], RN, DNP, PhD; Jennifer Barsky Reese[2], PhD; Carolyn Lafferty[1], PhD; Debra L Barton[1], RN, PhD; Jeanne Carter[3], PhD

[1]College of Nursing, University of Tennessee at Knoxville, Knoxville, TN, United States

[2]Cancer Prevention and Control Program, Fox Chase Cancer Center, Philadelphia, PA, United States

[3]Gynecology Service, Department of Surgery and Department of Psychiatry & Behavioral Sciences, Memorial Sloan Kettering Cancer Center, New York, NY, United States

Corresponding Author:
Noël M Arring, RN, DNP, PhD
College of Nursing
University of Tennessee at Knoxville
1412 Circle Dr.
Knoxville, TN, 37996
United States
Phone: 1 865 974 1988
Email: narring@utk.edu

Abstract

Background: Women with breast or gynecological cancer and their intimate partners often face sexual problems in their relationships. Accessing care for sexual health problems is challenging for several reasons (eg, limited trained health care providers and privacy concerns), making self-management approaches highly promising.

Objective: This study assessed the feasibility of the Psychosexual Educational Partners Program (PEPP), a 6-week sexual health self-management intervention for women treated for breast or gynecological cancer and their intimate partners.

Methods: A mixed methods single-arm, repeated measures design was used for this study. An attrition rate of ≤25% was considered feasible. Intervention experiences were assessed via interviews, and preliminary effects on the Dyadic Sexual Communication Scale, relationship quality measured by the Revised Dyadic Adjustment Scale, and sexual health measured by PROMIS Sexual Function and Satisfaction version 2.0 were explored quantitatively.

Results: In total, 7 (77%) of the 9 couples completed the study through week 6 and provided both pre- and poststudy data, resulting in an attrition rate of 22% (2/9), which met the feasibility benchmark for attrition of 25% or less. The following two themes emerged: (1) PEPP *helped us start difficult conversations* and impacted *emotional and physical intimacy*. The intervention adherence was 85%. Dyadic Sexual Communication Scale scores improved with a mean change score of 6.64 (SD 9.65) and a Cohen *d* of 0.69. Revised Dyadic Adjustment Scale scores declined slightly, with a mean change score of –0.93 (SD 3.41) and a Cohen *d* of 0.27. PROMIS Sexual Function and Satisfaction version 2.0 scores showed small improvements for women on desire, with a mean change score of 2.36 (SD 6.24) and a Cohen *d* of 0.38. Similarly, for women, the satisfaction mean change score was 2.20 (SD 8.22) and a Cohen *d* of 0.27. For intimate partners, a small effect was found for desire, but in this instance, desire decreased with a mean change score of –1.57 (SD 6.09) and a Cohen *d* of 0.26.

Conclusions: The findings support PEPP as a feasible intervention for improving sexual communication. If proven effective in a randomized controlled trial, it has the potential to address the critical gap in supportive care among female cancer survivors.

Trial Registration: ClinicalTrials.gov NCT05070299; https://clinicaltrials.gov/study/NCT05070299

(JMIR Form Res 2025;9:e75743) doi:10.2196/75743

XSL•FO
RenderX

KEYWORDS

breast cancer; cancer; couples-based; dyad; gynecological cancer; oncology; partners; self-management; sexual communication; sexual health

Introduction

Background

As many as 90% of women with a history of gynecological cancer and 70% of women with a history of breast cancer report issues with sexual health [1]. These concerns encompass both negative physical symptoms and a woman's cognitive and emotional sexual connection [1-3]. Changes in intimacy include sexual communication, sexual interest and satisfaction [2], and one's appearance or body and self-image [3].

In addition, partners of cancer survivors often experience distress related to cancer-related sexual problems, as well as negative effects on their own sexual functioning [4-6]. It has long been recognized that the quality of a partnered relationship is a significant predictor of a woman's sexual health after breast cancer [2,7]. To understand the individual and interdependent experiences of female cancer survivors and their partners, it is critical to examine how such couples navigate their experiences within the context of their relationships [8,9]. According to the theory of dyadic illness management [10], patients and their partners navigate the illness together. For couples facing life after cancer, the challenges can be both physical and psychological [10,11] in nature. Recent systematic reviews highlight the efficacy of a couple-based approach for improving relationship dynamics in the presence of adequate dyadic coping after cancer [12-14]. A couple-based approach is also consistent with the 2017 American Society of Clinical Oncology sexual health guidelines, which call for health care providers to include the partner if the patient wishes [15].

Access to couple-based psychosexual interventions is often restricted due to the limited number of health care providers equipped to deliver these interventions. Therefore, self-management approaches hold promise for reaching a greater number of patients while enhancing the cost-effectiveness, flexibility, and accessibility of the intervention. In addition, self-management strategies for sexual health among cancer survivors are promising [16-18]. On the basis of our previous research and clinical data from an established female sexual medicine program [15], our interdisciplinary team developed the Psychosexual Educational Partners Program (PEPP), a self-managed intervention.

Objectives

The primary aim of this study was to assess the feasibility of the PEPP, a 6-week sexual health self-management intervention for couples and women treated for breast or gynecological cancer who report cancer-related sexual changes and their intimate partners. We selected an attrition rate of 25% or less as our *a priori* feasibility benchmark. In addition, the study evaluated participants' intervention experiences using a qualitative investigation and explored the impact of the intervention on sexual communication, relationship quality, and sexual health.

Methods

Study Design

This feasibility study of the PEPP, a sexual health self-management intervention for couples, included a mixed methods single-arm, repeated measures design.

Ethical Considerations

This study was approved by the University of Michigan Institutional Review Board (HUM00202063) and registered on ClinicalTrials.gov (NCT05070299). All participants provided written informed consent between March and July 2022. All study data were deidentified before analysis. Compensation for study participation was up to US $50 per person or US $100 per couple.

Setting and Recruitment Procedures

This study was conducted at the University of Michigan, a National Cancer Institute–designated comprehensive cancer center located in Ann Arbor, Michigan. Recruitment involved physician referrals, the University of Michigan Health research portal (a clinical trial volunteer database), cancer registry mailings, and community outreach (eg, flyers and presentations). Once a potential participant reached out to the study team, their eligibility was assessed using a checklist derived from the inclusion and exclusion criteria described in the following sections. Women who met the eligibility requirements were sent information about the study's purpose, procedures, and expectations, along with their partners. If both expressed interest in participating, they were sent a consent form and scheduled for a consent visit. Consent visits were conducted via a secure Health Insurance Portability and Accountability Act–compliant video platform or by phone. Consent was obtained using an institutional review board–approved electronic system during this session. Once consent was obtained, the participants were sent the PEPP workbook and scheduled for their baseline visit.

Participants

The study participants were enrolled as couples, consisting of women who had been treated for breast or gynecological cancer and their intimate partners. Eligible participants were females with a history (any stage) of breast or gynecological cancer aged ≥18 years who had completed primary treatment (chemotherapy, radiation, or surgery) ≥3 months and ≤5 years before registration, reported that they had experienced a negative change in communication or intimacy since their diagnosis or treatment, and were able to read and write in English. Participants must also have had a stable partner, defined as anyone with whom the woman has had an intimate relationship for at least 3 months before her cancer diagnosis. Concurrent adjuvant endocrine therapy or HER2-targeted therapy was allowed. Exclusion criteria included individuals on an antidepressant for less than 30 days before registration, or with an unstable dose (expected change in dose or treatment); a history of sexual abuse; a psychiatric disorder such as major

XSL•FO
RenderX

depressive disorder, bipolar disorder, obsessive compulsive disorder, or schizophrenia; or enrollment in another study addressing sexual health.

Intervention

PEPP, a sexual health self-management intervention for couples, was developed by a multidisciplinary team and informed by clinical data from an established female sexual medicine program [15]. PEPP included a self-administered workbook with 3 chapters normalizing intimacy after cancer, reconnecting, and expanding communication and intimacy (Table 1) and appendices written at a reading level of grade 6.5. Phone call check-ins were completed at the end of each module with at least one member to assess the couples' adherence to the intervention components, identify adverse events, and address questions or concerns. Couples were instructed to complete one chapter every 2 weeks but were allotted additional time if needed.

Table 1. Workbook chapters included in the Psychosexual Educational Partners Program.

Title	Content	Goals
Chapter 1: Normalizing intimacy after cancer	• Physical and emotional sequelae of cancer treatment • Identifying challenges and barriers • Positive thinking and unhelpful thoughts • Mindfulness exercise, including progressive muscle relaxation	• Improved understanding of sexual health challenges created by cancer and cancer treatment • Improved awareness of challenges (physical, emotional, and communication) experienced by partner • Improved mindfulness
Chapter 2: Reconnecting	• Answering "What is Mindfulness?" • Sensate focus 101 • Goal-setting activities	• Increased awareness about ways to approach physical intimacy • Practical skills for goal setting as a couple
Chapter 3: Expanding communication and intimacy	• Communication 101 • Mindfulness, including guided imagery of leaves on a stream activity	• Increased knowledge of verbal and nonverbal communication

The PEPP workbook was written as a motivating, nonjudgmental guide providing information to support couples as they work toward rebuilding intimacy after cancer and creating healthy behavioral changes together. To enhance couples' motivation, 4 key components of motivational interviewing were integrated into the workbook: partnership, acceptance, compassion, and evocation [19]. First, the workbook begins with an overview of sexual health after cancer and then acknowledges a variety of physical and emotional challenges that couples often face. The goal is to help couples engage with the material according to their level of readiness and comfort (partnership). Couples are routinely encouraged to modify their activities to suit their preferences and needs (acceptance). The couples' experiences, interests, and needs are prioritized when they are asked questions or presented with activities (compassion). Finally, through carefully crafted *PEPP talks* and activities, the workbook provides a vehicle for couples to think about what is important to them and articulate their own steps toward behavior change (evocation).

Data Collection

Following the informed consent process, each participant completed a demographic survey and the Dyadic Sexual Communication Scale (DSC), Revised Dyadic Adjustment Scale (RDAS), and PROMIS Sexual Function and Satisfaction version 2.0 (PROMIS SexFS). Data were collected using the REDCap (Research Electronic Data Capture; Vanderbilt University) at baseline and week 6. Participant experience interviews were conducted using the Health Insurance Portability and Accountability Act–compliant Zoom (Zoom Video Communications) within 14 days of workbook completion.

Study Measures

Attrition was defined as any participant who did not provide data for both pre- and postquantitative measures. Participants were considered to have completed the study if they provided quantitative and qualitative data in week 6. Our *a priori* benchmark was that an attrition rate of 25% or less would be considered feasible [20].

To evaluate the participants' experiences with the intervention, each couple participated in an interview within 14 days at the end of week 6. The interview questions were based on the experience of the research team. Questions prompted responses ranging from simple feedback on the length of the intervention to more complex input about its impact. Questions were developed to assess participants' experiences with the workbook, such as perceived usefulness, time commitment, and relevance of the content.

The DSC is a 13-item instrument that assesses sexual communication between partners. The revised scale uses a 5-point response option where 3 is neutral [21]. Items were reverse scored so that high scores reflected more positive communication related to their sexual relationship. Possible scores range from 13 to 65, with higher scores representing a perception of better sexual communication with their partners. The original scale discriminated between couples with and without sexual problems, with a Cronbach α of 0.87 [22].

The RDAS is a 14-item scale with 3 themes (satisfaction, cohesion, and consensus) that measures an individual's perception of the quality of a relationship with an intimate partner using a 6-point Likert scale [23]. The scores range from 0 to 69. Higher scores indicate more positive dyadic adjustment.

XSL•FO
RenderX

Scores ≤47 indicate relationship distress. This scale was found to have high reliability, with a Cronbach α of 0.90 [23].

PROMIS SexFS V2 measures sexual activities, symptoms, functioning, and evaluation of sexual experiences [24]. General screener items ask about sexual activity and reasons for not engaging in sexual activity in the past 30 days. Female participants answered the 14-item version, and male participants answered the 10-item version. On both instruments, the desire and satisfaction subscales each included 2 items, and raw scores for each subscale ranged from 2 to 10. Raw scores for the desire and satisfaction domains were converted to T-scores using PROMIS T-score look-up tables, with a mean of 50 (SD 10) in the referent population. Higher scores indicate more of the domains being assessed. In the validation study, Cronbach α scores ranged from 0.85 to 0.98 [25]. This measure has been used in other studies on dyadic sexual health interventions with similar reliability and validity [26].

Data Analysis

All demographic and outcome variables were summarized using descriptive statistics (mean, SD, percentage, and frequency). To evaluate the intervention's impact on sexual communication, relationship quality, and sexual health, effect sizes were calculated with Cohen *d* from baseline to week 6. Means, variance, 95% CIs, and effect sizes were calculated to better estimate what sample size is needed for a larger trial with at least 80% power to detect at least a medium effect size. PROMIS domain scores can only be calculated using lookup tables when participants respond to all items in the domain. Therefore, missing data on the PROMIS scale resulted in participants being dropped from the analyses. SPSS statistical software (version 27; IBM Corp) was used for the analyses.

Participant interviews were conducted by one study team member (CL) to evaluate the intervention and the next steps in its delivery. The interviewer was known to the participants through recruitment and biweekly phone check-ins. Qualitative data were deidentified and summarized using interview questions. In total, 12 items prompted simple feedback that was quantified, and descriptive statistics were summarized. Three open-ended items could be described as "what" questions [27], such as "What, if any, impact has the workbook made on your ability to communicate with one another about difficult topics?" To respond to these exploratory items, a qualitative description approach was adopted, and content analysis was conducted to examine the 3 open-ended questions [27]. Two study team members (NMA and CL) independently coded the data with a focus on straight description, where coding remained consistent with the language used by participants [28]. The codes were reviewed by the coding team, and discrepancies were resolved through consensus.

Results

Sample Demographics

In total, 9 couples, including 1 gynecological and 8 breast cancer survivors, were recruited for 18 participants (Figure 1). Participant ages ranged from 30 to 75 years; 17 (94%) identified as White, 1 (6%) as Hispanic; 15 (83%) reported a college degree or higher; 14 (78%) reported an income of US $100,000 or more; and 8 (89%) couples were married.

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) diagram of the Psychosexual Educational Partners Program. LTFU: loss to follow-up.

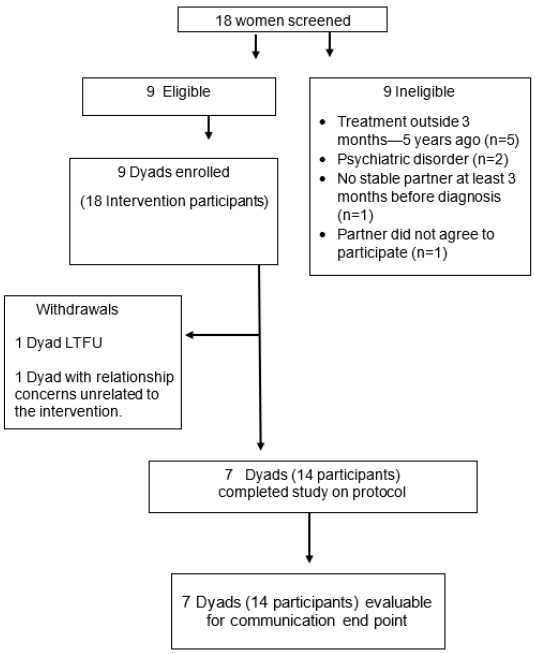

Feasibility

In total, 7 (77%) of the 9 couples, for 14 participants, completed the study through week 6 and provided both pre- and poststudy data, resulting in an attrition rate of 22% (2/9) of couples, which met the feasibility benchmark for attrition of 25% or less. Of the 2 (22%) couples who did not complete the study, 1 (11%) couple withdrew in week 4 to seek couples therapy for concerns unrelated to the intervention, the other was lost to follow-up, and one couple included the study's only gynecological cancer survivor.

XSL•FO
RenderX

The 7 (77%) couples who completed the study read the content in the 3 chapters. There was an 85% completion rate for the 17 activities included in the workbook. In addition, 2 (29%) of the 7 couples completed less than 50% of the activities. Moreover, 5 (71%) of the 7 couples completed all 17 activities.

Participant Experience: Qualitative Description

Codes based on participant language and supporting quotations are provided in Table 2, and a full qualitative analysis is presented in Multimedia Appendix 1. The code "helped us start difficult conversations" emerged in response to 2 of the 3 questions. Overall, half (7/14, 50%) of the participants reported that the 6-week time frame was just right. In total, 8 (57%) of the 14 participants stated that it was a helpful guide for discussions.

Table 2. Qualitative description of the Psychosexual Educational Partners Program: open-ended questions, themes, and supporting quotes (N=14).

Questions	Answers
Question 1: How (or in what way) did the workbook match your understanding of the sexual health challenges created by cancer and cancer treatments?	• It was normalizing (n=2, 14%) • "I'm not alone in this. This is a common thing. it's not like my fault or our fault that...we have these challenges, but...this is like a normal issue." [Survivor] • "It helped me to understand that what we were experiencing was not just normal, but actually pretty common...I wish I'd known this before. It would have just been helpful to know that lots of women, lots of couples go through this." [Survivor] • It gave us a new way to think about sexual health (n=3,21%) • "I didn't think about...you know, that bullet point had been affected by...cancer and treatment and things like that. So there were some things that I was like. 'Oh, yeah that that makes sense that I hadn't thought about before.'" [Survivor] • "[M]aking sure intimacy is defined not just as a goal to sexual intercourse, but a goal to physical [and] emotional connection." [Partner] • Helped us start difficult conversations (n=3, 21%) • "Where it was tremendously useful was in getting us to...discuss things that I didn't want to discuss. It's more me, [my partner] will talk about anything, certainly to do with sex and I have been very reticent and this was really forcing me to. We talked about things." [Survivor]
Question 2: What, if any, impact has the workbook made on your ability to communicate with one another about difficult topics?	• Helped us start difficult conversations (n=2, 14%) • "Before there was kind of [an] elephant in the room, and this...forced us to talk about it." [Survivor] • Opportunities for practice (n=6, 43%) • "I think...the frequency of the conversations just sort of lends itself to just a closer connection and communication when we're doing it on a regular basis."[Survivor] • Communication tools (n=6, 43%) • "I statements have been really helpful...So it's been helpful to just like take a second to pause and reframe what I want to say to him." [Survivor] • "I think that we have spent a lot more time communicating and kind of maybe pausing and reflecting on each other's experience, and just like thinking about it from both perspectives."[Partner]
Question 3: What, if any, impact has the workbook made in the level of intimacy of your relationship?	• Emotional and physical intimacy (n=5, 36%) • "The emotional intimacy piece...has like really facilitated some of the physical intimacy things as well, and so they're kind of hand in hand." [Partner] • "[H]aving those discussions in and of itself is sort of an intimate experience." [Survivor] • Helpful (n=4, 29%) • "I think we're headed in the right direction, and that was super helpful as well." [Survivor] • "It was helpful. I think it was helpful." [Partner]

Sexual Health Outcomes

The DSC improved over the 6 weeks of the study (Table 3). Possible scores for these items ranged from 1 to 5, where higher scores indicated greater agreement. The overall mean score at baseline was 45.00 (SD 11.06), and at 6 weeks it was 51.64 (SD 11.71), with a mean change score of 6.64 (SD 9.65) and a Cohen *d* of 0.69.

XSL•FO
RenderX

Table 3. Sexual health outcomes of the Psychosexual Educational Partners Program at baseline and week 6.

	Values, mean (SD)	Values, mean change (SD)	Cohen *d*
Dyadic Sexual Communication Scale (n=14)			
Baseline	45.00 (11.06)	__[a]	—
6 wk	51.64 (11.71)	6.64 (9.65)	0.69
Revised Dyadic Adjustment Scale (n=14)			
Baseline	51.14 (6.32)	—	—
6 wk	50.21 (6.78)	−0.93 (3.41)	0.27
PROMIS Sexual Function and Satisfaction version 2.0			
Women			
Desire (n=7)			
Baseline	42.31 (9.87)	—	—
6 wk	44.67 (8.62)	2.36 (6.24)	0.38
Satisfaction (n=6)			
Baseline	41.60 (3.30)	—	—
6 wk	43.80 (7.75)	2.20 (8.22)	0.27
Partners			
Desire (n=7)			
Baseline	55.63 (8.61)	—	—
6 wk	54.06 (9.71)	−1.57 (6.09)	0.26
Satisfaction (n=6)			
Baseline	44.63 (6.90)	—	—
6 wk	45.60 (4.95)	0.97 (6.92)	0.14

[a]Not applicable.

The RDAS scores declined slightly, with a mean score at baseline of 51.14 (SD 6.32), and at 6 weeks it was 50.21 (SD 6.78), with a mean change score of –0.93 (SD 3.41) and a Cohen *d* of 0.27. The 6-week score was highly correlated with the baseline score of 0.867 ($P<.001$).

The PROMIS SexFS scores showed small improvements for women for desire and satisfaction. The mean score for desire at baseline for the women was 42.31 (SD 9.87), and at 6 weeks it was 44.67 (SD 8.62), with a mean change score of 2.36 (SD 6.24) and a Cohen *d* of 0.38. The mean score for satisfaction at baseline for the women was 41.60 (SD 3.30), and at 6 weeks it was 43.80 (SD 7.75), with a mean change score of 2.20 (SD 8.22) and a Cohen *d* of 0.27. For partners, a small effect was found for desire, but in this instance, desire decreased from 55.63 (SD 8.61) at baseline to 54.06 (SD 9.71) at 6 weeks, reflecting a mean change score of –1.57 (SD 6.09) and a Cohen *d* of 0.26.

Discussion

Principal Findings

Overall, there was a 77% retention rate and 85% intervention adherence, which supports the feasibility of PEPP, a novel sexual health self-management intervention for couples surviving breast and gynecological cancer. Preliminary efficacy analyses indicate that the intervention yielded a medium effect on sexual communication, as measured by the DSC scale (d=0.69). In addition, among the women, small effect sizes were observed on the PROMIS SexFS subscales for desire (d=0.38) and satisfaction (d=0.27). These findings support the feasibility and potential efficacy of PEPP in enhancing sexual health outcomes among breast cancer survivors and their intimate partners.

The intervention was well received by the participants, with a retention rate of 77% (7/9), which surpassed the benchmark. These findings are consistent with those found in a recent systematic review of 19 trials of couple-based intervention studies by Li et al [13] and contribute to the literature by demonstrating the positive reception of a sexual health self-management intervention, whereas most previous interventions in this vein involved a trained facilitator [13]. Using a self-management approach could also be appealing to couples because it offers privacy when seeking help for sexual problems, a sensitive issue that some couples may feel uncomfortable discussing with others outside the relationship [29]. Some self-management interventions for sexual health among cancer survivors have shown promise but have been limited by high attrition [16-18]. This study used a low-touch (every 2-week phone check-in) approach, which we believe influenced our retention rate. Other studies addressing sexual

XSL•FO
RenderX

or psychosocial concerns among cancer survivors who used a similar approach reported high retention rates [30-32].

Findings from the qualitative interviews identified specific ways in which couples perceived the self-management intervention as helpful in promoting sexual communication and intimacy. On the basis of participant feedback, it is recommended that PEPP be updated by shifting some of the activities of chapters 2-3 and updating some of the imagery used. The quantitative results are promising and suggest that workbook components designed to improve couples' coping skills in overcoming communication barriers may have been effective, although the findings are preliminary. Taken together, these findings support the need for further research into the effectiveness of the PEPP intervention for improving sexual health communication among couples learning to rebuild intimacy after cancer.

Surprisingly, despite some promising increases in sexual communication and small improvements in desire among women and satisfaction among both groups on the PROMIS SexFS, we found that scores on the RDAS declined, with a small effect found for this change. Couples were in the nondistressed range on the RDAS at baseline and at 6 weeks. This small decline, as well as the small decrease in desire reported by partners on the PROMIS SexFS measure, may be a result of discussing sexual issues, which can be challenging for many couples. Because the couples included in the study were in nondistressed relationships, it could be that relationship quality may not change over the course of a 6-week study. It would be interesting to examine whether baseline relationship quality could serve as a moderator of the effects on sexual outcomes in future studies.

Clinical Implications

Sexual health issues are often an unmet need for many women with a history of cancer. This was a small single-arm feasibility study of PEPP, a sexual health self-management intervention for couples. The PEPP intervention appears to be a promising and potentially easily disseminated self-management intervention; however, additional testing is needed to assess its efficacy. If found to be effective, it could serve as a tangible tool that health care providers could provide to patients and their partners to help them discuss and manage sensitive issues in their sexual relationships. While research is needed to assess its efficacy, it is vitally important for clinicians to initiate conversations about sexual health to guide patients to currently available resources.

Research Implications

While the phone call check-ins at the end of each module provided a low-touch method for staying in contact with participants, recent research exploring self-management interventions supports the use of text or email reminders [33-36] or weekly assessments [37], which could improve the scalability of a self-managed intervention such as PEPP. Future research should continue to evaluate the most effective way to continue to engage participants in self-managed interventions and identify predictors for which couples are best served by self-managed couples' interventions.

A major advantage of self-management interventions is the participation of couples who would not otherwise join due to discomfort in discussing sexual issues with a person outside the relationship or have time constraints limiting their ability to travel to seek help. However, the couples who participated in this study were likely to be highly intrinsically motivated; therefore, the data may not be generalizable to all couples with intimacy concerns following breast or gynecological cancer. Future research should assess potential moderators, including sexual distress, dyadic adjustment, and relationship quality.

Study Limitations

As this was a feasibility study with a small sample size, there was limited ability to generalize these findings or determine the efficacy of the intervention, as there was no control condition or comparison group. Furthermore, the sample was largely White, well-educated, and of high socioeconomic strata. In addition, the final sample included only breast cancer survivors, which could signify that this may not be feasible for gynecological cancer survivors. Another potential limitation of this study is that the women were only eligible if their partners agreed to participate with them. Therefore, women with partners who are less supportive may be underrepresented.

Conclusions

The findings support the potential benefit of the PEPP intervention in helping couples work to re-establish intimacy after breast and gynecological cancer and cancer treatment. Research is needed in a larger, more diverse population to evaluate the efficacy of the PEPP intervention compared to a control condition, to better understand its impact on sexual communication, and to test strategies that can support the wider dissemination of such an intervention.

Acknowledgments

This study was supported by the Breast Cancer Research Foundation. The funder was not involved in the study design, data collection, analysis, interpretation, or writing of the manuscript. JC was supported in part by the National Institutes of Health/National Cancer Institute Cancer Center Support Grant P30CA008748.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to the small sample size but are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Authors' Contributions

JC, DLB, and NMA contributed to the conceptualization of this study. The methodology was developed by NMA, DLB, JBR, and JC. Funding acquisition was overseen by DLB, resources and supervision were provided by NMA and DLB, and project administration was provided by NMA. Investigation and formal analysis were conducted by DLB, NMA, and CL, and data curation was performed by NMA and CL. All authors contributed to the validation, visualization, interpretation, writing, and reviewing of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Descriptive summaries of participant experiences.
[DOCX File , 23 KB - formative_v9i1e75743_app1.docx]

References

1. Robison K, Kulkarni A, Dizon DS. Sexual health in women affected by gynecologic or breast cancer. Obstet Gynecol 2024 Apr 01;143(4):499-514. [doi: 10.1097/AOG.0000000000005506] [Medline: 38207333]
2. Ganz PA, Desmond KA, Belin TR, Meyerowitz BE, Rowland JH. Predictors of sexual health in women after a breast cancer diagnosis. J Clin Oncol 1999 Aug;17(8):2371-2380. [doi: 10.1200/JCO.1999.17.8.2371] [Medline: 10561299]
3. Berterö C, Chamberlain Wilmoth M. Breast cancer diagnosis and its treatment affecting the self: a meta-synthesis. Cancer Nurs 2007;30(3):194-202; quiz 203. [doi: 10.1097/01.NCC.0000270707.80037.4c] [Medline: 17510582]
4. Maleki M, Mardani A, Ghafourifard M, Vaismoradi M. Changes and challenges in sexual life experienced by the husbands of women with breast cancer: a qualitative study. BMC Womens Health 2022 Aug 02;22(1):326 [FREE Full text] [doi: 10.1186/s12905-022-01906-8] [Medline: 35918700]
5. Reese JB, Zimmaro LA, McIlhenny S, Sorice K, Porter LS, Zaleta AK, et al. Coping with changes to sex and intimacy after a diagnosis of metastatic breast cancer: results from a qualitative investigation with patients and partners. Front Psychol 2022;13:864893 [FREE Full text] [doi: 10.3389/fpsyg.2022.864893] [Medline: 35465532]
6. Cui P, Ai J, Chen X, Cheng C, Shi J, Li S, et al. Dyadic effects of perceived burden and psychological distress on quality of life among Chinese advanced cancer patients and their caregivers. Sci Rep 2025 Apr 09;15(1):12180 [FREE Full text] [doi: 10.1038/s41598-025-96861-3] [Medline: 40204850]
7. Vegunta S, Kuhle CL, Vencill JA, Lucas PH, Mussallem DM. Sexual health after a breast cancer diagnosis: addressing a forgotten aspect of survivorship. J Clin Med 2022 Nov 14;11(22):6723 [FREE Full text] [doi: 10.3390/jcm11226723] [Medline: 36431200]
8. Arring N, Barton DL, Reese JB. Clinical practice strategies to address sexual health in female cancer survivors. J Clin Oncol 2023 Nov 01;41(31):4927-4936. [doi: 10.1200/JCO.23.00523] [Medline: 37535889]
9. Zhou Y, Che CC, Chong MC, Zhao H, Hou Y, Li J. Understanding the patient-spouse communication experience during chemotherapy for gastric cancer based on the couple's adaptation to cancer intimacy model: a directed content analysis. Support Care Cancer 2025 Jun 14;33(7):578. [doi: 10.1007/s00520-025-09516-4] [Medline: 40515864]
10. Lyons KS, Lee CS. The theory of dyadic illness management. J Fam Nurs 2018 Feb;24(1):8-28 [FREE Full text] [doi: 10.1177/1074840717745669] [Medline: 29353528]
11. Di Mattei VE, Perego G, Taranto P, Mazzetti M, Marotta E, Candiani M, et al. The long-term effects of cancer treatment on sexuality and couple relationships. Fam Process 2021 Sep;60(3):853-865. [doi: 10.1111/famp.12604] [Medline: 33030242]
12. Valente M, Chirico I, Ottoboni G, Chattat R. Relationship dynamics among couples dealing with breast cancer: a systematic review. Int J Environ Res Public Health 2021 Jul 07;18(14):7288 [FREE Full text] [doi: 10.3390/ijerph18147288] [Medline: 34299738]
13. Li M, Zhang L, Li N, Luo X, Chen J, Zhang Y, et al. Effects of couple-based dyadic interventions on breast cancer patients and their intimate partners: a systematic review and meta-analysis. J Adv Nurs 2023 Sep;79(9):3192-3213. [doi: 10.1111/jan.15639] [Medline: 36918983]
14. Traa MJ, De Vries J, Bodenmann G, Den Oudsten BL. Dyadic coping and relationship functioning in couples coping with cancer: a systematic review. Br J Health Psychol 2015 Feb;20(1):85-114. [doi: 10.1111/bjhp.12094] [Medline: 24628822]
15. Carter J, Lacchetti C, Andersen BL, Barton DL, Bolte S, Damast S, et al. Interventions to address sexual problems in people with cancer: American Society of Clinical Oncology clinical practice guideline adaptation of Cancer Care Ontario guideline. J Clin Oncol 2018 Feb 10;36(5):492-511. [doi: 10.1200/JCO.2017.75.8995] [Medline: 29227723]
16. Schover LR, Strollo S, Stein K, Fallon E, Smith T. Effectiveness trial of an online self-help intervention for sexual problems after cancer. J Sex Marital Ther 2020;46(6):576-588. [doi: 10.1080/0092623X.2020.1762813] [Medline: 32400321]
17. Brotto LA, Dunkley CR, Breckon E, Carter J, Brown C, Daniluk J, et al. Integrating quantitative and qualitative methods to evaluate an online psychoeducational program for sexual difficulties in colorectal and gynecologic cancer survivors. J Sex Marital Ther 2017 Oct 03;43(7):645-662. [doi: 10.1080/0092623X.2016.1230805] [Medline: 27592509]

XSL•FO
RenderX

18. Sella T, Snow C, Freeman H, Poorvu PD, Rosenberg SM, Partridge AH. Young, empowered and strong: a web-based education and supportive care intervention for young women with breast cancer across the care continuum. JCO Clin Cancer Inform 2021 Aug;5:933-943. [doi: 10.1200/CCI.21.00067] [Medline: 34473546]
19. Miller WR, Rollnick S. Motivational Interviewing: Helping People Change Third Edition. New York, NY: Guilford Press; 2013.
20. Reese JB, Bober SL, Sorice KA, Handorf E, Chu CS, Middleton D, et al. Starting the conversation: randomized pilot trial of an intervention to promote effective clinical communication about sexual health for gynecologic cancer survivors. J Cancer Surviv 2024 Jun;18(3):800-809. [doi: 10.1007/s11764-022-01327-4] [Medline: 36604391]
21. Zacchilli TL. The relationship between conflict and communication, sex, relationship satisfaction, and other relational variables in dating relationships. Texas Tech University. 2007. URL: https://ttu-ir.tdl.org/items/dc9faeab-246c-4a14-8676-dc342c144800 [accessed 2025-09-25]
22. Catania J. Dyadic sexual communication scale. In: Milhausen RR, Sakaluk JK, Fisher TD, Davis CM, Yarber WL, editors. Handbook of Sexuality-Related Measures. New York, NY: Routledge; 2019.
23. Busby DM, Christensen C, Crane DR, Larson JH. A revision of the Dyadic Adjustment Scale for use with distressed and nondistressed couples: construct hierarchy and multidimensional scales. J Marital Fam Ther 1995;21(3):289-308 [FREE Full text] [doi: 10.1111/j.1752-0606.1995.tb00163.x] [Medline: 618698290]
24. PROMIS sexual function and satisfaction manual. HealthMeasures. 2018 Jun 3. URL: https://www.healthmeasures.net/images/PROMIS/manuals/PROMIS_Sexual_Function_and_Satisfaction_Measures_User_Manual_v1.0_and_v2.0.pdf [accessed 2025-09-25]
25. Weinfurt KP, Lin L, Bruner DW, Cyranowski JM, Dombeck CB, Hahn EA, et al. Development and initial validation of the PROMIS(®) sexual function and satisfaction measures version 2.0. J Sex Med 2015 Sep;12(9):1961-1974. [doi: 10.1111/jsm.12966] [Medline: 26346418]
26. Reese JB, Smith KC, Handorf E, Sorice K, Bober SL, Bantug ET, et al. A randomized pilot trial of a couple-based intervention addressing sexual concerns for breast cancer survivors. J Psychosoc Oncol 2019;37(2):242-263 [FREE Full text] [doi: 10.1080/07347332.2018.1510869] [Medline: 30580675]
27. Sandelowski M. Whatever happened to qualitative description? Res Nurs Health 2000 Aug;23(4):334-340. [doi: 10.1002/1098-240x(200008)23:4<334::aid-nur9>3.0.co;2-g] [Medline: 10940958]
28. Neergaard MA, Olesen F, Andersen RS, Sondergaard J. Qualitative description - the poor cousin of health research? BMC Med Res Methodol 2009 Jul 16;9:52 [FREE Full text] [doi: 10.1186/1471-2288-9-52] [Medline: 19607668]
29. Hay CM, Donovan HS, Hartnett EG, Carter J, Roberge MC, Campbell GB, et al. Sexual health as part of gynecologic cancer care: what do patients want? Int J Gynecol Cancer 2018 Nov;28(9):1737-1742 [FREE Full text] [doi: 10.1097/IGC.0000000000001376] [Medline: 30358703]
30. Körner A, Roberts N, Steele RJ, Brosseau DC, Rosberger Z. A randomized controlled trial assessing the efficacy of a self-administered psycho-educational intervention for patients with cancer. Patient Educ Couns 2019 Apr;102(4):735-741. [doi: 10.1016/j.pec.2018.12.002] [Medline: 30545649]
31. Fergus K, Ahmad S, Gardner S, Ianakieva I, McLeod D, Stephen J, et al. Couplelinks online intervention for young couples facing breast cancer: a randomised controlled trial. Psychooncology 2022 Mar;31(3):512-520. [doi: 10.1002/pon.5836] [Medline: 34669239]
32. Perz J, Ussher JM, Australian Cancer and Sexuality Study Team. A randomized trial of a minimal intervention for sexual concerns after cancer: a comparison of self-help and professionally delivered modalities. BMC Cancer 2015 Sep 09;15:629. [doi: 10.1186/s12885-015-1638-6] [Medline: 26353787]
33. Fry JP, Neff RA. Periodic prompts and reminders in health promotion and health behavior interventions: systematic review. J Med Internet Res 2009 May 14;11(2):e16. [doi: 10.2196/jmir.1138] [Medline: 19632970]
34. De Leon E, Fuentes LW, Cohen JE. Characterizing periodic messaging interventions across health behaviors and media: systematic review. J Med Internet Res 2014 Mar 25;16(3):e93 [FREE Full text] [doi: 10.2196/jmir.2837] [Medline: 24667840]
35. Crafoord MT, Fjell M, Sundberg K, Nilsson M, Langius-Eklöf A. Engagement in an interactive app for symptom self-management during treatment in patients with breast or prostate cancer: mixed methods study. J Med Internet Res 2020 Aug 10;22(8):e17058 [FREE Full text] [doi: 10.2196/17058] [Medline: 32663140]
36. Rimmer B, Brown MC, Sotire T, Beyer F, Bolnykh I, Balla M, et al. Characteristics and components of self-management interventions for improving quality of life in cancer survivors: a systematic review. Cancers (Basel) 2023 Dec 19;16(1):14. [doi: 10.3390/cancers16010014] [Medline: 38201442]
37. Murphy KM, Sauer C, Yang D, Hass N, Novakovic K, Helfand B, et al. The development of iManage-PC, an online symptom monitoring and self-management tool for men with clinically localized prostate cancer. Cancer Nurs 2022;45(1):E309-E319. [doi: 10.1097/NCC.0000000000000948] [Medline: 33867430]

Abbreviations

DSC: Dyadic Sexual Communication Scale
PEPP: Psychosexual Educational Partners Program

XSL•FO
RenderX

PROMIS SexFS: PROMIS Sexual Function and Satisfaction version 2.0
RDAS: Revised Dyadic Adjustment Scale
REDCap: Research Electronic Data Capture

Edited by A Mavragani; submitted 09.04.25; peer-reviewed by K Abbott-Anderson; comments to author 22.05.25; revised version received 02.07.25; accepted 16.09.25; published 21.10.25.

Please cite as:
Arring NM, Reese JB, Lafferty C, Barton DL, Carter J
A Sexual Health Self-Management Intervention (Psychosexual Educational Partners Program) for Couples With a History of Breast and Gynecological Cancer: Mixed Methods Feasibility Study
JMIR Form Res 2025;9:e75743
URL: https://formative.jmir.org/2025/1/e75743
doi:10.2196/75743
PMID:

©Noël M Arring, Jennifer Barsky Reese, Carolyn Lafferty, Debra L Barton, Jeanne Carter. Originally published in JMIR Formative Research (https://formative.jmir.org), 21.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Developing a Community of Practice to Provide Care Coordination and Address Health-Related Social Needs for Veterans Receiving Care in Community-Based Settings: Program Development and Survey Study

Carolyn Turvey[1,2,3], PhD; Natalie Suiter[1,3], BS; Rhonda Fellows[4], RN, MSN; Shawna Domeyer[5], MSW, JD; Lindsey Fuhrmeister[1,2,3], BA; Amanda Heeren[1,2,3], MSW, LISW; Kimberly McCoy[1,3], MS; M Bryant Howren[1,3,6], MPH, PhD

[1]Veterans Rural Health Resource Center-Iowa City, Veterans Health Administration Office of Rural Health, Department of Veterans Affairs, Iowa City, IA, United States

[2]Carver College of Medicine, Department of Psychiatry, The University of Iowa, Iowa City, IA, United States

[3]Center for Access & Delivery Research and Evaluation, VA Iowa City Health Care System, Iowa City, IA, United States

[4]Van Buren County Hospital, Keosauqua, IA, United States

[5]Crescent Community Health Center, Dubuque, IA, United States

[6]Carver College of Medicine, Department of Internal Medicine, University of Iowa, Iowa City, United States

Corresponding Author:
Carolyn Turvey, PhD
Veterans Rural Health Resource Center-Iowa City
Veterans Health Administration Office of Rural Health
Department of Veterans Affairs
601 Highway 6
Iowa City, IA, 52246
United States
Phone: 1 319 338 0581 ext 637714
Email: carolyn.turvey@va.gov

Abstract

Background: Approximately half of US veterans receive care outside of US Department of Veterans Affairs (VA) Veterans Health Administration facilities—a proportion expected to rise due to the Promise to Address Comprehensive Toxics Act and expanded use of VA-purchased community care.

Objective: This paper describes the structure and impact of the Veterans Care Coordination in Community Settings (VetCoor) program. VetCoor was implemented in 2 non-VA community health centers, and we explored setting, staffing, and treatment targets to enhance veteran care and inform broader dissemination.

Methods: VetCoor embedded a coordinator within community-based, non-VA health care settings to improve veteran identification and address unmet medical needs. Coordinators also connected veterans with VA and local resources addressing health-related social needs. VetCoor included training on veteran needs and military culture. It also held a monthly community of practice call where coordinators shared best practices and met with facility representatives to learn about VA services.

Results: From May 2021 to September 2023, a total of 220 veterans participated, engaging in 773 sessions. Of these 220 veterans, 73 (33.2%) received VA enrollment assistance; 54 (24.5%) were referred for medical care; and 82 (37.3%) received care coordination, including medication reconciliation assistance. They also received assistance with transportation (46/220, 20.9%) nutrition and food access (42/220, 19.1%), housing and repair (42/220, 19.1%), and utility payment support (31/220, 14.1%). Common barriers to veterans seeking care were perceptions that enrolling in the VA took resources from veterans more in need and confusion regarding discharge papers required for enrollment.

Conclusions: VetCoor supported rural veterans' health care and health-related social needs using dedicated coordinators. This model addresses resource gaps, fosters VA-community collaboration, and aligns with the VA's expanding benefits under the Promise to Address Comprehensive Toxics Act.

(JMIR Form Res 2025;9:e80654) doi:10.2196/80654

XSL•FO
RenderX

KEYWORDS

veterans; rural health care; health-related social needs; care coordination; US Department of Veterans Affairs; Veterans Health Administration; community care

Introduction

Approximately half of all veterans receive health care outside of US Department of Veterans Affairs (VA) Veterans Health Administration facilities [1-3]. Some veterans do not enroll in the VA, whereas others enroll but seek access at both VA and community health care settings due to dual insurance coverage—referred to herein as "dual-use veterans." Veterans enrolled in the VA may also receive community-based health care reimbursed by the VA, also known as purchased care. The VA allows veterans to seek care locally but reimbursed by the VA if they live far enough away from a VA facility or they have an extended wait for needed care. Recent programs such as the Maintaining Internal Systems and Strengthening Integrated Outside Networks Act [4] and the original Veterans Access, Choice, and Accountability Act of 2014 [5] have greatly expanded VA purchased care and, thereby, the total number of dual-use veterans.

Rural veterans often qualify for VA purchased care in the community due to the eligibility criterion based on distance from a VA facility. Nonetheless, they often find themselves in health care physician and nursing shortage areas, which is exacerbated by ongoing rural facility closures [6,7]. In light of the rural health care access crisis, the federal government provides enhanced reimbursement for federally qualified health centers (FQHCs), critical access hospitals (CAHs), or rural health clinics (RHCs). Therefore, rural veterans often combine care from the VA with local FQHCs, CAHs, or RHCs.

This paper describes the development and implementation of Veterans Care Coordination in Community Settings (VetCoor), a program designed to address cross-system coordination between the VA and rural health care facilities. VetCoor builds on work addressing the essential components of cross-system coordination [8-12], as well as previous pilot studies testing models to improve VA and rural community physician coordination [11,13-16]. The aim of the program is to develop and test novel relationships between the VA and community providers and engage veterans in the full range of their potential VA benefits.

Cross-system coordination is currently being explored within the VA as well as through models such as the relational coordination theory by Gittell et al [8] or the emphasis on relational, informational, and management continuity of care by Haggerty et al [9,17]. In this context, Miller et al [18] explored cross-system care coordination by interviewing veterans and community providers about their wishes for optimal care coordination. Veterans valued highly the dedication of VA staff yet also wanted enhanced access to local non-VA providers. Veterans reported confusion about overall responsibility for care coordination between the 2 health care systems. Community providers, when asked what would reduce such confusion, expressed their wish to have contact information for a single VA staff person highly knowledgeable in care coordination policies and procedures. They also desired contact information for the specific providers treating shared patients. Garvin et al [19], in a systematic review of dual-use care in rural settings, identified 2 policy and administrative characteristics contributing to successful interorganizational care coordination with the VA: alignment and understanding between partners and a formal contractual agreement.

Hempel et al [10] convened care coordination leadership within the VA and used a modified Delphi panel to identify research, practice, and policy recommendations for cross-system care coordination. This process yielded a recommendation for developing measures of shared care. They recommended enhanced communication technologies, determining best points of contact between health care facilities, and cross-system education about the unique cultures and patient care approaches within the 2 (or more) systems.

VetCoor aims to address many of these identified aspects of cross-system care coordination and builds on prior pilot work exploring feasible models for VA collaboration with community providers. Howren et al [14,15] and Lee et al [16] both addressed the care coordination needs of rural veterans specifically. Howren et al [14,15] improved identification of veterans seeking rural care facilities through an enhanced screening process at patient check-in, thereby improving their ability to identify shared care. The programs by both Howren et al [14,15] and Lee et al [16] developed key points of contact between the VA and community providers and provided targeted training for rural providers about veteran-specific concerns, such as military posttraumatic stress disorder or the specific needs of female veterans.

VetCoor replicates and expands upon these models, with the aim of identifying and testing new screening, staffing, and intervention parameters to increase impact while enhancing feasibility. Improving feasibility will allow for later scaling up to enroll many community providers and further enhance the reach of VA services in rural settings. Both the original pilot by Howren et al [14,15] and the VetCoor implementation detailed in this paper received funding from the Veterans Health Administration's Office of Rural Health to enhance access to and quality of care (New Office of Rural Health Management and Analysis Database PROJ-04143).

Methods

VetCoor Implementation Overview

VetCoor starts with the identification of rural non-VA community providers who likely serve many veterans yet are not located close to VA medical centers. Once potential collaborative sites are identified, VetCoor leads meet with non-VA community facility leadership to discuss the benefits of VetCoor, including funding for a 0.5 full-time equivalent (FTE) care coordinator, mentorship regarding identifying patients who are veterans, and collaboration with both VA and

XSL•FO
RenderX

community services to best address the needs of these veterans. If agreeable, VetCoor representatives initiate the contracting process and work with the community facility to identify potential VetCoor coordinators. Coordinator onboarding includes training about developing workflows to identify veterans in need of services, care coordination strategies for working with the VA and other community providers, veteran and military culture, and addressing health-related social needs. Coordinators are also trained on reporting requirements to convey how many veterans they are serving and what services are provided. Ongoing mentorship is provided for up to 3 years for facilities engaging in VetCoor. As part of this mentorship, coordinators participate in monthly community of practice virtual meetings where presentations by local VA staff describe VA services and coordinators share their best practices between themselves. Descriptions of each of these steps in developing VetCoor at community sites are provided in the following sections.

Selection of Setting

Identification of potential non-VA RHCs made use of the VA geographic information systems [20,21]. Geospatial mapping allowed VetCoor leads to identify geographic regions where many veterans reside yet there are no VA facilities within 30 miles or more. Once these areas were identified, the VA could map where FQHCs, CAHs, and RHCs are located. Potential community partners were vetted to ensure that they were providers endorsed by the VA's Care in the Community program. Endorsed providers have already been vetted by the VA for quality standards and, therefore, are eligible for reimbursement through the VA's Care in the Community program. Once potential community partners were identified, VetCoor leads met with the VA's local and regional strategic planners to ensure that the VA did not already have future plans to add any facility or community-based outpatient clinic in the target region.

VetCoor leads met with community facility leadership to describe the program and established the contract for the 0.5 FTE. VetCoor leads also conveyed the demands made on the coordinator during this period, including onboarding training, reporting requirements, and meeting attendance and cadence. If agreed, VetCoor initiated the contract processes with the community facility and informed the local VA of this new relationship.

This manuscript describes the implementation of VetCoor in 2 unique community-based health care settings. The first was a CAH serving a rural community with a population of 1000. This rural clinic is located 42 miles from the nearest VA community-based outpatient clinic and 86 miles from the nearest VA hospital. The second was an FQHC serving a town of 60,000 and located in the same town as a VA community-based outpatient clinic. This FQHC was located 86 miles from the nearest VA hospital. Implementation started in May 2021, and data were collected between May 2021 and September 2023.

Identifying VetCoor Coordinators

The original pilot program by Howren et al [14,15] used nurses in a patient navigator model to assist with VA health care access and care coordination. In this replication, we explored expanding this role to include other health care delivery professionals. One of the sites participating in this project designated a nurse as the VetCoor project lead, and the other, due to local nursing staff shortages, designated a nonlicensed care coordinator. This nonlicensed care coordinator was in a career transition from law to social work but had no formal social work training at the time of the pilot. This enabled us to evaluate the feasibility and impact of engaging the designated care coordinator at the site and not insisting on nursing credentials. If successful, this would greatly improve the scalability of the VetCoor program in low-resource settings and align with other care coordination programs that use nonnursing staff such as community health workers or staff with Bachelor's-level training [22,23].

Onboarding Coordinators

During project orientation, coordinators met with the project leads (CT, NS, AH, and LF) to discuss implementing the veteran screen at check-in, progress in enrolling veterans, initial veteran contacts and assessment of unmet medical needs, and processes for referral for VA Care in the Community reimbursement. They were also referred to online resources educating them about VA culture, such as the VHA TRAIN Learning Collaborative program [24], or online resources provided by specific VA program offices. For example, the VA mental health services line has many online resources regarding the identification and management of posttraumatic stress and depression [25]. VetCoor coordinators were also provided with online resources describing the VA's Care in the Community program, clarifying complex authorization processes for their health care facilities and how to determine when veterans may be eligible for VA Care in the Community.

Community of Practice

Once a month, coordinators and VA project staff attended a community of practice meeting. These were virtual meetings as attendees were located across broad geographic settings. Each meeting addressed different subjects relevant to VetCoor, such as VA enrollment, female veterans' health, or VA Care in the Community. A list of meetings and topics is shown in Table 1. The meetings began with brief presentations, and then coordinators were free to share best practices and address recent challenges in meeting veterans' needs. Presenters were often from VA facilities serving the general region of the community partners to continue to build these relationships. VetCoor aimed to build strong VA facility and community health care facility ties so that when the VetCoor mentoring ended after 3 years, community providers would be fully aware of VA resources and how to collaborate with their local VA. For example, the local facility enrollment and benefits lead presented the VA enrollment policy in broad brushstrokes but then provided his contact information for any follow-up questions or if coordinators had specific enrollment questions for a veteran with whom they were working. Often, facility public relations staff or Care in the Community staff attended these meetings.

XSL•FO
RenderX

Table 1. Topics and timeline of community of practice presentations.

Time point	Topic	Presenter or discussant
Month 1	Introductions and presentation of the basic VetCoor[a] structure	VetCoor lead and coordinators
Month 2	Relevant VHA[b] TRAIN content and other online resources	VetCoor lead
Month 3	Ways of approaching veterans for VetCoor	VetCoor lead and experienced coordinators
Month 4	Understanding processes for VA[c] enrollment	VA benefits officer from local facility
Month 5	Navigating Care in the Community referrals	Facility Care in the Community representative and experienced coordinator
Month 6	Understanding the VA formulary and formulary change requests	Facility pharmacist and experienced coordinator
Month 7	Veteran transportation benefits	Facility transportation officer
Month 8	Homeless veteran programs and resources	Facility homeless program representative
Month 9	VA women's health	Facility women's health program lead
Month 10	Coordinating care between the VA and the community	All current coordinators and representative patient-aligned care team nurse case manager
Month 11	Telehealth resources within the VA, including VA Video Connect, My HealtheVet, mobile apps, and patient-generated health data	Office of Rural Health telehealth expert
Month 12	Yearly wrap-up: identifying strengths and areas for improvement of the current program and community of practice model	VetCoor lead and all coordinators

[a]VetCoor: Veterans Care Coordination in Community Settings.

[b]VHA: Veterans Health Administration.

[c]VA: US Department of Veterans Affairs.

The community of practice also helped build relationships between coordinators. Discussions started during the community of practice call were often continued between coordinators after these calls without the participation of a VA representative. This allowed for candid sharing of experiences working with the VA and best practices in care coordination.

Veteran Identification and VetCoor Participation

A multifaceted approach identified veterans in the participating community-based health care settings. First, we implemented a mandatory screening question in the community partner's electronic health record. This health IT innovation was implemented at patient check-in for all patients' medical visits and aided in determining which patients were veterans. The electronic mandatory screening question was as follows: "Have you served in the United States military or armed forces? This includes Air Force, Army, Coast Guard, Marines, Navy, National Guard, or Reserves." Howren et al [14,15] developed this process to replace previous veteran screens that simply asked, "Are you a Veteran?" The latter question did not explicitly state the full range of service (eg, service in the National Guard and Reserve or service without deployment) that potentially renders one eligible for the range of VA benefits. Moreover, some veterans may not self-identify as such for personal reasons, including but not limited to if they were not in combat, do not believe that their service "counted," or strongly believe that they do not meet income criteria for VA benefits.

VetCoor coordinators also facilitated identification of veterans by building facility and community awareness of the program. They attended routine staff meetings at their facilities, obtaining referrals from providers who identified veterans in need in the course of their regular clinical practice. Coordinators also developed relationships with their local veterans service officers and veteran service organizations such as Veterans of Foreign Wars.

Once identified, veterans were invited to meet with the VetCoor coordinator to complete an interview during which veterans were assessed for current VA enrollment status and unmet medical or health-related social needs [26,27]. When needs related to VA benefits and enrollment or other medical or health-related social needs were identified, the coordinator made and assisted with appropriate referrals. They often provided care coordination across VA and community health care providers. If health-related social needs were identified, such as food insecurity, vocational training, housing, utilities, and transportation, both VA and community resources were provided. There was no limit to the number of encounters between a VetCoor coordinator and a veteran in addressing these needs.

VetCoor coordinators uniformly recommended that all unenrolled veterans who agreed to participate in VetCoor try to enroll for VA benefits. They did so as the criteria for benefits have changed greatly with recent legislation and it is very hard to predict who is eligible and what services they may be eligible for. For example, under the Promise to Address Comprehensive

XSL•FO
RenderX

Toxics Act, a veteran who is suicidal may be reimbursed for inpatient mental health treatment even if they are not enrolled in the VA at the time of admission. Coordinators addressed unmet medical needs for veterans enrolled in the VA by combining both community and VA resources and arranging for VA-reimbursed care when indicated. Among veteran participants for whom patient health care coordination with the VA was indicated, coordinators first obtained veteran release of information permissions according to their respective facilities' health information exchange policy and practices.

If, in spite of the recommendation to enroll, veterans did not wish to apply for VA benefits, this was respected. VetCoor coordinators identified state and locally based health care and health-related social services. This included state and county-based veteran benefit offices; Medicaid if indicated; and local health, housing, and food insecurity resources. In this way, VetCoor was able to address unmet needs in veterans whom the VA was not able to assist, per veteran preference.

VetCoor Coordinator Reporting

To facilitate the characterization of VetCoor participants and services received, a project-specific Qualtrics survey (Qualtrics International Inc) [28] was developed. Coordinators did not have experience with research data collection and, at 0.5 FTE, often held multiple roles that did not allow for extensive reporting requirements. The Qualtrics survey was developed with input from the coordinators to ensure that it was not too burdensome yet captured key components of their meetings with veterans. Using a secure link to the VA, the coordinators reported encounter-specific, nonidentifiable demographic data and a description of the services provided to the participating veteran. The survey comprised radio buttons, drop-down menus, and open-text boxes.

Veteran Participant Characteristics

To protect veteran anonymity in reporting data to the VA, veterans were identified using unique ID numbers only, and VA program staff did not have the ability to cross-reference the ID to a veteran's name or any VA medical records. Initial contact and return encounter dates were indicated by fiscal quarter (3-month period within a year) rather than exact encounter dates. Age ranges were reported in 10- to 20-year increments, not as the actual age or birth date. Sex, race, and ethnicity were collected.

Program and Service Use Characteristics

Coordinators provided a narrative account of unique and return encounters with veterans via an open-text field. Narratives could include but were not limited to location of the encounter (ie, via phone or in person), participants in the encounter (ie, veteran, spouse, or physician), presenting concern (ie, need for medical or social service or VA benefit eligibility), actions taken (ie, care coordinated or referral made), and unmet needs that would require further follow-up. Narratives could be of any length.

Analysis of Coordinator Narratives

Veteran participation data were analyzed using descriptive statistics and inductive and deductive coding of the narratives. Given that data were collected over the course of 17 months, analysis took place in phases. In the first analysis, 3 VA project team members (CT, LF, and AH) used inductive reasoning and together reviewed 30 narratives to determine activities that took place in the VetCoor encounters. These activities were added to a codebook, and consensus was reached across all 30 narratives. The analysts then divided the remaining narratives and individually coded the activities, adding codes to the codebook if new activities were identified. After individual coding was completed, the analysts met to discuss unclear narratives and reach a consensus on codes. The resulting code list was thematically analyzed, and the final codes (activities) were organized into 2 categories: health care–related needs addressed and health-related social needs addressed. For the remainder of the project, narratives were coded using deductive reasoning based on the established codebook by project team members LF and AH.

Ethical Considerations

This program was determined to constitute quality improvement and deemed not human subject research through review by the University of Iowa Human Subjects Review Board and formal written participant consent was not required. This manuscript adheres to the Standards for Quality Improvement Reporting Excellence guidelines for qualitative reporting. Data will be destroyed at the close of the study except for what is required to be kept in accordance with the National Archives and Records Administration published in the Veterans Health Administration's Record Control Schedule 10-1.

Results

Veteran and VetCoor Use Characteristics

From May 2021 to September 2023, a total of 220 veterans were identified via an embedded electronic health record screen, referral from providers within participating sites, and veteran self-referral. Veterans participating in this program engaged in a total of 773 sessions with VetCoor coordinators (Table 2). A session was considered to be any visit between the VetCoor coordinator and a veteran addressing VA benefits, health-related issues, or health-related social needs. Sessions were conducted in person or via phone or video technologies. The demographics of the participants reflect the general veteran population in rural Midwest regions, with most being male, White or non–Hispanic/Latino, and aged ≥55 years. In total, 45% (99/220) of the participants resided in rural or highly rural areas.

XSL•FO
RenderX

Table 2. Veterans Care Coordination in Community Settings number of unique veterans and number of sessions by age, gender, race, ethnicity, and rurality.

Characteristic	Unique veterans (n=220), n (%)	Sessions (n=773), n (%)	Ratio of sessions per unique veteran
Age (years)			
17-35	16 (7.3)	47 (6.1)	2.94
36-55	43 (19.5)	167 (21.6)	3.88
56-65	80 (36.4)	352 (45.5)	4.40
66-75	43 (19.5)	124 (16.0)	2.88
≥76	38 (17.3)	83 (10.7)	2.18
Sex			
Female	31 (14.1)	132 (17.1)	4.26
Male	187 (85)	638 (82.5)	3.41
Did not endorse male or female	2 (0.9)	3 (0.4)	2.00
Race			
American Indian or Alaska Native	1 (0.5)	3 (0.4)	3.00
Asian	1 (0.5)	2 (0.3)	2.00
Black	7 (3.2)	23 (3.0)	3.29
Multiracial	2 (0.9)	15 (1.9)	7.50
White	207 (94.1)	728 (94.2)	3.52
Preferred not to answer	2 (0.9)	2 (0.3)	1.00
Ethnicity			
Hispanic or Latino	1 (0.5)	10 (1.3)	10.00
Non–Hispanic or Latino	217 (98.6)	761 (98.4)	3.51
Preferred not to answer	2 (0.9)	2 (0.3)	1.00
Rurality			
Rural or highly rural	99 (45.0)	307 (39.7)	3.10
Urban	118 (53.6)	462 (59.8)	3.92
Unknown	3 (1.4)	4 (0.5)	1.33

VA Enrollment

As shown in Table 3, half (110/220, 50%) of the participants were not enrolled in the VA when they joined the program. Of those 110 participants, 73 (66.4%) were open to discussing enrollment options. Some opted out of enrollment due to a strong belief in their ineligibility (5/110, 4.5%), whereas others declined based on personal preference (3/110, 2.7%).

XSL•FO
RenderX

Table 3. Health care–related and care coordination tasks addressed.

Task addressed	Unique veterans (n=220), n (%)	Sessions per task (n=773), n (%)
VA[a] enrollment		
Not enrolled in the VA	110 (50)	341 (44.1)
Follow-up for those not enrolled N (% of unenrolled unique veterans and sessions)		
Agreed to discuss enrollment	73 (66.4)	126 (36.9)
Initiated enrollment	27 (24.5)	30 (8.8)
Declined; veteran believed they were ineligible	5 (4.5)	7 (2)
Declined; veteran did not wish to enroll in the VA	3 (2.7)	3 (0.1)
Medical referral N (% of total sample [N=220] and % of total sessions [N=773])		
Referred for medical care to address unmet needs	54 (24.5)	109 (14.1)
Settings the veteran was referred to		
Referred for Care in the Community purchased by the VA	41 (18.6)	69 (8.9)
Referred for Care in the Community that was not reimbursed by the VA but paid for through other means	22 (10.0)	29 (3.8)
Referred to care at local VA facility	7 (3.2)	12 (1.6)
Care coordination N (% of total sample [N=220] and % of total sessions [N=773])	82 (37.3)	158 (20.4)
Between VA and community facility	21 (9.5)	27 (3.5)
Coordination of services within the community facility	59 (26.8)	94 (12.2)
Coordination with other non-VA community providers	19 (8.6)	33 (4.3)
Coordination with home health	9 (4.1)	14 (1.8)
Medication reconciliation	20 (9.1)	33 (4.3)

[a]VA: US Department of Veterans Affairs.

Medical Care Referrals

Among medical care referrals, 18.6% (41/220) were directed to the Care in the Community program. A total of 10% (22/220) were referred to non-VA community providers with costs covered by the veteran's other insurance or out-of-pocket payments. In total, 3.2% (7/220) of the veterans were referred back to the VA for care.

Care Coordination

Coordination efforts were primarily with providers located in the same community-based facilities as our coordinators (59/220, 26.8%). A total of 9.5% (21/220) of the participants received care coordination that facilitated a combination of community and VA care.

Health-Related Social Needs

Many veterans indicated health-related social needs (Table 4). The most common needs included transportation (46/220, 20.9%), food insecurity (42/220, 19.1%), housing repairs (42/220, 19.1%), and utility assistance (31/220, 14.1%). In addition to facilitating enrollment in VA benefits, coordinators assisted in applications for Medicaid (20/220, 9.1%) and social security (16/220, 7.3%).

XSL•FO
RenderX

Table 4. Health-related social needs addressed.

Need addressed	Unique veterans (n=220), n (%)	Sessions (n=773), n (%)	Number of sessions per health-related social need, mean (SD)
Transportation	46 (20.9)	95 (12.3)	2.1 (1.6)
Food insecurity	42 (19.1)	99 (12.8)	2.4 (1.9)
Housing repair and expenses	42 (19.1)	55 (7.1)	1.3 (0.8)
Utility payment assistance	31 (14.1)	79 (10.2)	2.5 (3.9)
Applying for Medicaid	20 (9.1)	25 (3.2)	1.2 (0.4)
Housing or homelessness assistance	19 (8.6)	56 (7.2)	2.9 (3.6)
Applying for social security	16 (7.3)	46 (6.0)	2.9 (3.7)
Access to legal assistance	11 (5.0)	18 (2.3)	1.6 (1.0)
Reducing medical debt	9 (4.1)	13 (1.7)	1.4 (0.7)
Assistance with job seeking	8 (3.6)	12 (1.6)	1.5 (0.9)
Assistance with family and dependent needs	2 (0.9)	3 (0.4)	1.5 (0.7)
Education benefits	1 (0.5)	1 (0.1)	1.0 (—[a])

[a]Not applicable.

Recurring Issues in Care Coordination and Assisting Veterans

Care in the Community Reimbursement Denials

Early meetings between VetCoor VA program staff and community health care partners included representatives from the health care partners' clinical leadership as well as administrative and billing representatives. In conducting this program, key areas of confusion arose regarding VA reimbursement for community-based care. For example, one common area of confusion was VA reimbursement for postacute care in the community. Community partners often assumed that VA authorization for an acute admission in the emergency department or an inpatient visit extended to any prescribed postacute care, such as follow-up primary care visits and diagnostic or rehabilitation services. This is not accurate as separate authorization of postdischarge care is required. This type of information was disseminated in our regular coordinator meetings to prevent costly mistakes in which often the veteran is left responsible for an unexpected medical cost.

Veteran Reluctance to Enroll in the VA

In light of the continually evolving legislation regarding veteran eligibility for benefits, our program strongly encouraged all interested unenrolled veterans to apply for VA benefits. One of the biggest barriers for veterans was the belief that others were more deserving of VA benefits. Coordinators explained to veterans that their enrolling in the VA did not take this opportunity away from other veterans and, in fact, it often led to greater support for the local VA. Discharge status was reviewed with veterans for accuracy, and if it was not accurate, VetCoor coordinators assisted in the process of getting the discharge status reviewed and possibly changed. Coordinators also emphasized the potential VA benefits across the life span. They reminded veterans who were currently healthy that they may at some point need the extensive wraparound services for older veterans available through the VA's geriatrics and extended care services. These benefits are of great assistance to veterans' loved ones as well as veterans themselves. Such arguments appealed to veterans' value of service and putting others before themselves.

Administrative Barriers to Veteran Enrollment

Both the local veterans service officers and the VetCoor coordinator assisted with completing the 1010EZ form—the VA's application for benefits—and putting together the required documentation. For example, many veterans identified in our program needed ongoing assistance related to their DD214 forms, which summarize a military service member's duty service and related information, including discharge status. Some veterans had lost their DD214 and were not aware of how to obtain a replacement form, whereas others believed that the discharge status on the DD214 was incorrect. VetCoor coordinators assisted the veterans in addressing these issues, including applying for a discharge status review when indicated.

Veterans Choosing Not to Enroll in the VA

VetCoor assisted all veterans regardless of VA enrollment or eligibility status. For those who preferred non-VA care resources, coordinators referred them to programs from both public agencies and nonprofit organizations. In addition to government-sponsored programs such as Medicaid, many states have veteran relief funds such as the Veterans' Trust fund in Iowa or the Military Family Relief Fund in Indiana. Moreover, cities, counties, and states often have programs, either publicly or privately funded, to assist with utilities, rent, housing or car repair, and access to fulfill basic needs. Coordinators turned to these resources while conveying that VA enrollment would be available at a later time if the veteran changed their mind.

XSL•FO
RenderX

Discussion

Principal Findings

VetCoor successfully replicates and expands upon the program by Howren et al [14,15] in addressing key components of cross-system care coordination. First, community providers needed to better assess whether a patient was a veteran. The enhanced screen embedded in the check-in process accomplished this. Community outreach, local presentations, and physician referrals also identified veterans who would then present for care coordination services at their local health care facility. Through VetCoor, coordinators were trained on VA as well as military culture. This included extensive training on the VA's Care in the Community program and reimbursement processes. Personal relationships were forged between the community providers and key contacts at their local VA facility through ongoing care coordination combined with the community of practice meetings.

Monthly community of practice meetings educated coordinators on the VA's housing, transportation, suicide prevention, women's health, and Care in the Community programs. VetCoor coordinators accessed these services and also tapped into state and county resources, building up a safety net for both enrolled and unenrolled veterans. No less importantly, the community of practice allowed coordinators to share frustrations as well as best practices. For example, these meetings helped identify areas of confusion, such as postacute care authorizations or navigating the VA formulary.

A significant addition to VetCoor relative to the programs by Howren et al [14,15] and Lee et al [16] was the focus on veteran health-related social needs. The VA is currently expanding eligibility criteria and benefits addressing housing, transportation, and vocational needs. Aligning with community providers who primarily treat underserved patients enabled us to identify both enrolled and not enrolled veterans with needs related to food security, transportation, and housing. VetCoor then worked with veterans and both VA and community systems to identify resources to improve the quality of life for these veterans.

Finally, community providers engaging in VetCoor often focused on VA purchased care as a key role of VetCoor coordinators. However, this recent iteration emphasized how VetCoor's goal is to connect veterans with a broad array of both community and VA benefits and integrate the convenience of local care with the documented high-quality care available at VA medical centers [29-32]. For example, the recommendation for all unenrolled veterans identified in our program to apply for VA benefits was based on current and future legislation that aims to broadly expand veteran eligibility for services. Under the Comprehensive Prevention, Access to Care, and Treatment Act, beginning in January 2023, veterans in an acute suicidal crisis can receive up to 30 days of inpatient care and 90 days of outpatient care regardless of VA enrollment status. Similarly, the VA's housing benefits through the Department of Housing and Urban Development–VA Supportive Housing have significantly expanded eligibility to include veterans who served just 1 day in the military. Combining both community and VA resources is necessary in light of dwindling wraparound services in rural areas, yet such collaboration runs counter to more competitive models seen in health care systems vying to remain solvent.

Limitations

While in development, this program faces significant limitations. Data collection remains limited due to the time demands for our coordinators practicing in busy community settings that do not necessarily have a tradition of research data collection. The data presented are based on coordinator self-report and global coding of narratives, which significantly limits the methodological rigor of this study. As this was the initial pilot phase of the study, we limited burdensome research data collection to increase the likelihood of community partner participation. Therefore, full evaluation of the program remains descriptive at this time. Currently, we are addressing this shortcoming in new sites embedded within a larger health care system where research resources are more readily available. We are planning data extraction from the medical record combined with repeated standard assessment data collected from program participants.

Future Directions

VetCoor is expanding to 5 new sites in the Upper Midwest using both nurse and non–clinically trained care coordinators. Coordinators from previous sites who have now completed the initial 3-year contract remain engaged and available to new coordinators for training and to participate in the community of practice meetings. As we expand to new sites, we are exploring more sophisticated data collection methods that do not increase coordinator workload, such as data use agreements regarding medical record data and brief administration of standardized measures of satisfaction, health, and well-being at standard time points. We are also interviewing stakeholders directly about their experience. In this context, we are developing evaluation metrics based on the reach, effectiveness, adoption, implementation, and maintenance framework [32]. Preliminary analyses will explore whether participating veterans improve on brief measures of overall self-reported health and whether the quality of care received improves according to standard measures of quality. Veteran experience and barriers to and facilitators of broader expansion will also be collected using qualitative methods.

Conclusions

Recent legislation (Promise to Address Comprehensive Toxics and Comprehensive Prevention, Access to Care, and Treatment Acts) extends VA benefits to Vietnam War, Gulf War, and post-9/11 veterans and those exposed to military toxins. This increases the likelihood that more veterans will turn to the VA for care—whether within VA facilities or through care purchased by the VA from community providers. The VA and community providers are facing enormous challenges meeting the needs of their communities in the context of physician and nursing shortages. VetCoor promotes enhanced collaboration and communication, with a joint focus on building a network to support veterans with extensive needs both within and outside VA purview. Future program iterations will explore integrated

XSL•FO
RenderX

continuing medical education, cross-system triage, and wraparound services.

Acknowledgments

ChatGPT (OpenAI) was used to assist with grammar and language editing in early drafts. The final manuscript was reviewed and revised by all authors. This work was funded by the US Department of Veterans Affairs (VA) Office of Rural Health Veterans Rural Health Resource Center–Iowa City "Developing a Community of Practice between VA and Rural Community Health Centers to Enhance Coordinated Care for Rural Veterans" project (New Office of Rural Health Management and Analysis Database PROJ-04143). The views expressed in this paper are those of the authors and do not necessarily reflect the position or policy of the VA or the US government.

Data Availability

The datasets generated or analyzed during this study are not publicly available as their collection and use are governed by the terms of funding provided by the US Department of Veterans Affairs, which imposes restrictions on data access and distribution. Questions about specific aspects of the data could be submitted to the primary author (CT).

Authors' Contributions

Conceptualization: CT (lead), MBH (supporting)
Data curation: CT (lead), NS (supporting), KM (supporting), LF (supporting), AH (supporting), MB (supporting)
Formal analysis: CT (lead), KM (equal)
Funding acquisition: CT
Investigation: CT (lead), NS (supporting), RF (supporting), SD (supporting), LF (supporting), AH (supporting), MBH (supporting)
Methodology: CT (lead), KM (equal)
Project administration: CT (lead), NS (equal), RF (supporting), SD (supporting), LF (supporting), AH (supporting)
Resources: CT
Supervision: CT
Validation: CT (lead), NS (supporting), RF (supporting), SD (supporting), LF (supporting), AH (supporting), MBH (supporting)
Visualization: CT (lead), LF (supporting), KM (supporting)
Writing—original draft: CT (lead), NS (supporting), RF (supporting), SD (supporting), LF (supporting), AH (supporting), MBH (supporting), KM (supporting)
Writing—review and editing: CT (lead), NS (supporting), LF (supporting), AH (supporting)

Conflicts of Interest

None declared.

References

1. 2023 survey of veteran enrollees' health and use of health care. US Department of Veterans Affairs. URL: https://www.va.gov/VHASTRATEGY/SOE2023/2023_Survey_of_Veteran_Enrollees_Health_and_Use_of_Health_Care_Main_Results_Report.pdf [accessed 2025-05-29]
2. Shen Y, Hendricks A, Wang F, Gardner J, Kazis LE. The impact of private insurance coverage on veterans' use of VA care: insurance and selection effects. Health Serv Res 2008 Feb 07;43(1 Pt 1):267-286 [FREE Full text] [doi: 10.1111/j.1475-6773.2007.00743.x] [Medline: 18211529]
3. Yoon J, Vanneman ME, Dally SK, Trivedi AN, Phibbs CS. Veterans' reliance on VA care by type of service and distance to VA for nonelderly VA-medicaid dual enrollees. Med Care 2019 Mar;57(3):225-229 [FREE Full text] [doi: 10.1097/MLR.0000000000001066] [Medline: 30676354]
4. The VA mission act of 2018 (VA maintaining systems and strengthening integrated outside networks act). House Committee on Veterans' Affairs. 2018. URL: https://veterans.house.gov/uploadedfiles/va_mission_act_summary.pdf [accessed 2025-05-29]
5. Veterans access, choice, and accountability act of 2014. US Congress. 2014. URL: https://www.congress.gov/bill/113th-congress/house-bill/3230 [accessed 2025-05-29]
6. Maganty A, Byrnes ME, Hamm M, Wasilko R, Sabik LM, Davies BJ, et al. Barriers to rural health care from the provider perspective. Rural Remote Health 2023 May;23(2):7769 [FREE Full text] [doi: 10.22605/RRH7769] [Medline: 37196993]
7. Malayala SV, Vasireddy D, Atluri P, Alur RS. Primary care shortage in medically underserved and health provider shortage areas: lessons from Delaware, USA. J Prim Care Community Health 2021 Feb 10;12:2150132721994018 [FREE Full text] [doi: 10.1177/2150132721994018] [Medline: 33567941]

XSL•FO
RenderX

8. Gittell JH, Godfrey M, Thistlethwaite J. Interprofessional collaborative practice and relational coordination: improving healthcare through relationships. J Interprof Care 2013 May 19;27(3):210-213. [doi: 10.3109/13561820.2012.730564] [Medline: 23082769]
9. Haggerty JL, Roberge D, Freeman GK, Beaulieu C. Experienced continuity of care when patients see multiple clinicians: a qualitative metasummary. Ann Fam Med 2013;11(3):262-271 [FREE Full text] [doi: 10.1370/afm.1499] [Medline: 23690327]
10. Hempel S, Ganz D, Saluja S, Bolshakova M, Kim T, Turvey C, et al. Care coordination across healthcare systems: development of a research agenda, implications for practice, and recommendations for policy based on a modified Delphi panel. BMJ Open 2023 May 17;13(5):e060232 [FREE Full text] [doi: 10.1136/bmjopen-2021-060232] [Medline: 37197809]
11. Hynes DM, Govier DJ, Niederhausen M, Tuepker A, Laliberte AZ, McCready H, et al. Understanding care coordination for veterans with complex care needs: protocol of a multiple-methods study to build evidence for an effectiveness and implementation study. Front Health Serv 2023 Aug 15;3:1211577 [FREE Full text] [doi: 10.3389/frhs.2023.1211577] [Medline: 37654810]
12. Mattocks KM, Cunningham K, Elwy AR, Finley EP, Greenstone C, Mengeling MA, et al. Recommendations for the evaluation of cross-system care coordination from the VA state-of-the-art working group on VA/non-VA care. J Gen Intern Med 2019 May 16;34(Suppl 1):18-23 [FREE Full text] [doi: 10.1007/s11606-019-04972-1] [Medline: 31098968]
13. Abrams TE, Alexander B, Flores A, Howren MB. Veterans utilizing a federally qualified health center: a clinical snapshot. Mil Med Res 2022 Apr 13;9(1):18 [FREE Full text] [doi: 10.1186/s40779-022-00379-y] [Medline: 35418105]
14. Howren MB, Kazmerzak D, Kemp RW, Boesen TJ, Capra G, Abrams TE. Identification of military veterans upon implementation of a standardized screening process in a federally qualified health center. J Community Health 2020 Jun 16;45(3):465-468. [doi: 10.1007/s10900-019-00761-3] [Medline: 31620908]
15. Howren MB, Kazmerzak D, Pruin S, Barbaris W, Abrams TE. Behavioral health screening and care coordination for rural veterans in a federally qualified health center. J Behav Health Serv Res 2022 Jan 25;49(1):50-60 [FREE Full text] [doi: 10.1007/s11414-021-09758-0] [Medline: 34036516]
16. Lee PW, Lee RE, Markle P, Shirley EA, Welch P. Dissemination of a care collaboration project. Fed Pract 2015 Oct;32(10):38-42 [FREE Full text] [Medline: 30766027]
17. Haggerty JL, Pineault R, Beaulieu M, Brunelle Y, Gauthier J, Goulet F, et al. Practice features associated with patient-reported accessibility, continuity, and coordination of primary health care. Ann Fam Med 2008 Mar 01;6(2):116-123 [FREE Full text] [doi: 10.1370/afm.802] [Medline: 18332403]
18. Miller CJ, Shin M, Pugatch M, Kim B. Veteran perspectives on care coordination between veterans affairs and community providers: a qualitative analysis. J Rural Health 2021 Mar 21;37(2):437-446 [FREE Full text] [doi: 10.1111/jrh.12526] [Medline: 33085119]
19. Garvin LA, Pugatch M, Gurewich D, Pendergast JN, Miller CJ. Interorganizational care coordination of rural veterans by veterans affairs and community care programs: a systematic review. Med Care 2021 Jun 01;59(Suppl 3):S259 [FREE Full text] [doi: 10.1097/MLR.0000000000001542] [Medline: 33976075]
20. Cowper D, Yu W, Kuebeler M, Kubal JD, Manheim LM, Ripley BA. Using GIS in government: an overview of the VHA's Healthcare Atlas, FY-2000. J Med Syst 2004 Jun;28(3):257-269. [doi: 10.1023/b:joms.0000032843.52406.2f] [Medline: 15446616]
21. Cowper-Ripley DC, Jia H, Wang X, Freytes IM, Hale-Gallardo J, Castaneda G, et al. Trends in VA telerehabilitation patients and encounters over time and by rurality. Fed Pract 2019 Mar;36(3):122-128 [FREE Full text] [Medline: 30983852]
22. Kessler M, Thumé E, Facchini LA, Moro LC, Tomasi E. Calidad de la visita domiciliaria del agente comunitario de salud en la atención básica y factores asociados. Rev Latino Am Enfermagem 2024;32:e4398. [doi: 10.1590/1518-8345.7203.4397]
23. Segal L, Donato R, Richardson J, Peacock S. Strengths and limitations of competitive versus non-competitive models of integrated capitated fundholding. J Health Serv Res Policy 2002 Jul 01;7 Suppl 1(1_suppl):S56-S64. [doi: 10.1258/135581902320176485] [Medline: 12175436]
24. VHA train. Institute for Learning, Education and Development (ILEAD). URL: https://www.va.gov/VHALEADERSHIPDEVELOPMENT/Resources/VHA_Train.asp [accessed 2025-05-29]
25. Mental health. US Department of Veterans Affairs. URL: https://www.mentalhealth.va.gov/ [accessed 2025-05-29]
26. Russell LE, Cohen AJ, Chrzas S, Halladay CW, Kennedy MA, Mitchell K, et al. J Gen Intern Med 2023 Oct 10;38(13):2906-2913 [FREE Full text] [doi: 10.1007/s11606-023-08181-9] [Medline: 37165261]
27. Weir RC, Proser M, Jester M, Li V, Hood-Ronick CM, Gurewich D. Collecting social determinants of health data in the clinical setting: findings from national PRAPARE implementation. J Health Care Poor Underserved 2020;31(2):1018-1035. [doi: 10.1353/hpu.2020.0075] [Medline: 33410822]
28. Understand every customer. Qualtrics. URL: https://www.qualtrics.com/ [accessed 2025-05-29]
29. Anhang Price R, Sloss EM, Cefalu M, Farmer CM, Hussey PS. Comparing quality of care in veterans affairs and non-veterans affairs settings. J Gen Intern Med 2018 Oct 25;33(10):1631-1638 [FREE Full text] [doi: 10.1007/s11606-018-4433-7] [Medline: 29696561]

XSL•FO
RenderX

30. Ohl ME, Richardson Miell K, Beck BF, Mecham B, Bailey G, Mengeling M, et al. Mortality among US veterans admitted to community vs veterans health administration hospitals for COVID-19. JAMA Netw Open 2023 May 01;6(5):e2315902 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.15902] [Medline: 37252740]
31. Vanneman ME, Wagner TH, Shwartz M, Meterko M, Francis J, Greenstone CL, et al. Veterans' experiences with outpatient care: comparing the veterans affairs system with community-based care. Health Aff (Millwood) 2020 Aug 01;39(8):1368-1376 [FREE Full text] [doi: 10.1377/hlthaff.2019.01375] [Medline: 32744943]
32. King DK, Shoup JA, Raebel MA, Anderson CB, Wagner NM, Ritzwoller DP, et al. Planning for implementation success using RE-AIM and CFIR frameworks: a qualitative study. Front Public Health 2020 Mar 3;8:59 [FREE Full text] [doi: 10.3389/fpubh.2020.00059] [Medline: 32195217]

Abbreviations

CAH: critical access hospital
FQHC: federally qualified health center
FTE: full-time equivalent
RHC: rural health clinic
VA: Veterans Health Administration
VetCoor: Veterans Care Coordination in Community Settings

Edited by J Sarvestan; submitted 16.07.25; peer-reviewed by D Boykin, M Rigby; comments to author 19.08.25; revised version received 16.10.25; accepted 17.10.25; published 31.10.25.

Please cite as:
Turvey C, Suiter N, Fellows R, Domeyer S, Fuhrmeister L, Heeren A, McCoy K, Howren MB
Developing a Community of Practice to Provide Care Coordination and Address Health-Related Social Needs for Veterans Receiving Care in Community-Based Settings: Program Development and Survey Study
JMIR Form Res 2025;9:e80654
URL: https://formative.jmir.org/2025/1/e80654
doi:10.2196/80654
PMID:

©Carolyn Turvey, Natalie Suiter, Rhonda Fellows, Shawna Domeyer, Lindsey Fuhrmeister, Amanda Heeren, Kimberly McCoy, M Bryant Howren. Originally published in JMIR Formative Research (https://formative.jmir.org), 31.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Impact of the Practice of Playing Video Games on Technical Skills Development in Preclinical Dental Education: Preliminary Cross-Sectional Observational Study

Roselyne Clouet[1*], DDS; Thomas Remaud[1*], DDS; Camille Boeffard[1], DDS; Samuel Serisier[2], DDS, PhD; Alexis Gaudin[2*], DDS, PhD; Tony Prud'homme[2*], DDS, PhD

[1]Service d'Odontologie (UIC11), Centre hospitalier universitaire de Nantes, Nantes Université, Nantes, France

[2]Regenerative Medicine and Skeleton (UMR 1229), Centre hospitalier universitaire de Nantes (INSERM), Nantes Université, Nantes, France

[*]these authors contributed equally

Corresponding Author:
Tony Prud'homme, DDS, PhD
Regenerative Medicine and Skeleton (UMR 1229)
Centre hospitalier universitaire de Nantes (INSERM)
Nantes Université
1 Place Alexis Ricordeau
Nantes, 44000
France
Phone: 33 677489648
Email: tony.prudhomme@univ-nantes.fr

Abstract

Background: Video games are becoming increasingly accessible and occupy a prominent position among students' leisure activities. Recent studies have demonstrated that engagement with video games can facilitate the development of specific abilities, such as visuospatial skills and hand-eye coordination. Thus, it seemed relevant to investigate whether the practice of playing video games could enhance the technical capabilities of novice dental students.

Objective: The objective of this study was to ascertain whether dental students who identified themselves as video game players exhibited differences in fundamental technical skills in comparison to students who identified as nonplayers. This preliminary study aimed to validate the methodology and measurement tools for a subsequent prospective study.

Methods: Second-year dental students who were novices in the field of preclinical dental practice were divided into 2 groups: one designated nonplayers and the other players. The visual, motor, and cognitive coordination of the students was assessed using 3 different tests. The initial assessment focused on evaluating spatial ability, while the subsequent assessments addressed arm-hand coordination and the velocity of execution. The study data were collected in September 2021.

Results: This study included 92 second-year dental students (mean age 19.49, SD 0.8 years). Thirty-seven students were placed in the player group (40%), while 55 were placed in the nonplayer group (60%). The findings showed no statistically significant differences between the two groups when the 3 distinct tests were administered. The nonplayer group made fewer mistakes on the test evaluating spatial ability (P=.051) and achieved a higher score on the modified Precision Manual Dexterity Test, which evaluates arm-hand coordination, compared to the player group (P=.17), but without statistical significance. The nonplayer group took more time to perform the pulpotomy during the third test, which evaluated execution speed, compared to the player group, again without significance (P=.87).

Conclusions: This study did not demonstrate significant differences between player and nonplayer dental students who participated in the study with regard to their fundamental technical abilities in a preclinical training environment. Nevertheless, it validated the feasibility of a methodology for a future longitudinal study to concentrate on the evolution of acquiring technical skills during preclinical training in these two populations. Consequently, further investigation is warranted to determine the potential impact of video games on the acquisition of surgical skills, including in dentistry.

(JMIR Form Res 2025;9:e80082) doi:10.2196/80082

XSL•FO
RenderX

KEYWORDS

video games; dental education; surgical education; simulation; preclinical training

Introduction

Since the inception of the video game industry in the 1970s, the significance of these games in the lives of users and players has increased exponentially. Video games have become a major component of popular culture, with their use becoming pervasive from a young age and subsequently becoming an integral part of people's lives. It is evident that video games have the capacity to exert a beneficial influence on society [1]. In 2022, players constituted almost 39% of the global population, and in countries with a high or very high Human Development Index, up to 71% of the population plays at least occasionally [2]. Consequently, players can be categorized according to their distinct gaming habits.

Despite the remarkable capacity of humans to acquire novel skills and adapt their behavior in response to experience, performance enhancements are frequently constrained by the parameters of the training environment [3]. Nevertheless, research conducted over the past decade has demonstrated that training involving action video games has the potential to yield learning that transfers beyond the immediate training context [3]. A significant number of psychological and educational studies have been conducted on the psychomotor effects that may be observed in individuals engaging in this practice, particularly among those rely on visual, motor, and cognitive coordination in their gaming activities.

A number of earlier studies reported that participation in video gaming was associated with enhancements in psychomotor functioning and other competencies. Indeed, players have been shown to outperform nonplayers in terms of the precision of their arm-hand movements and the efficiency with which they complete tasks [4,5]. Players have also been shown to track faster-moving objects, detect changes in visual short-term memory more efficiently, switch tasks more rapidly, and carry out mental rotation tasks more effectively than nonplayers [6]. Furthermore, players appear to exhibit increased perceptual reaction speeds without compromising performance accuracy [7].

However, it should not be forgotten that video games may also exert negative effects. The problems associated with them are mainly related to excessive use, including the potential for addiction, aggression resulting from violent games, and problems linked to a more sedentary lifestyle (obesity, musculoskeletal disorders, eye strain) [8,9].

On the other hand, students are increasingly fond of learning through games. Recent studies highlight the growing role of serious games as effective educational tools in dental education, enhancing engagement and learning outcomes. Studies have shown that gamified applications, such as ESKILLD for laser dentistry, SimOL for oral lesion diagnosis, and virtual environments like the Gamified Virtual Dental Clinic, significantly improve students' satisfaction and knowledge acquisition. Although learning outcomes are often comparable to traditional teaching, serious games consistently promote higher motivation and positive learning experiences among dental students [10-13]. It is therefore relevant for dental educators to embrace these approaches in order to adapt to a student population whose abilities may be influenced by video games, but who are also receptive to and more interested in teaching methods that use serious games.

It is therefore pertinent to consider whether the extensive use of video games in medical practice exerts any influence on clinical competencies [14]. A number of studies have been conducted in order to observe the possible interactions between surgical skills and video gaming practice, with a particular focus on laparoscopic surgery. The results of these studies demonstrated a high degree of correlation with those from other fields, as video game skill correlated with higher laparoscopic surgical skills [15]. Indeed, prior experience with video games has been demonstrated to enhance baseline performance in laparoscopic simulator skills [14,16]. Nevertheless, it is evident that there exists a methodological heterogeneity amongst the various trials that have examined the impact of video games on surgical skills among medical students [17]. The conclusions are varied and occasionally contradictory. Consequently, the literature does not permit conclusions to be drawn on the consequences of a particular video game practice on the learning of technical skills in surgery. There is a paucity of evidence to support the hypothesis that the use of video games enhances surgical simulation performance [18]. Nevertheless, it appears that medical students hold a favorable opinion of the use of video games in a medical teaching context [19].

To date, there have been no studies examining video game use patterns among dental students and correlating them with preclinical technical performance using standardized testing. Preclinical training constitutes an indispensable element of dental education, with the objective of cultivating students' manual dexterity and fundamental technical skills, which are essential for the successful execution of dental procedures. Indeed, it is imperative that they are able to comprehend the diverse anatomical configurations of teeth as depicted in medical imagery. Furthermore, concomitant development of arm-hand and visuospatial coordination is critical, particularly in the context of dental care using indirect vision [20].

In view of the growing literature pertaining to the contribution of video games in the medical field, we hypothesized that playing such games could influence preclinical learning in dentistry. In this pilot study, we included a wide variety of game genres. Rather than restricting inclusion to traditional action games (first-person shooter or role-playing games), we also included sports games, strategy games, racing games, platform games, and even smartphone games. The objective of this work was to ascertain whether the practice of playing video games could enhance the technical capabilities of dental novice students, thereby directly impacting their competencies during the preclinical training phase. Furthermore, given the absence of a preexisting standardized methodology, it was necessary to adapt evaluation tools for use in the dental field. A significant

XSL•FO
RenderX

part of this study was therefore also dedicated to applying specifically designed or adapted psychomotor tests to assess skills that are directly relevant to dental education.

Methods

Study Design

This was an observational, descriptive, single-center, controlled, cross-sectional, and single-blinded preliminary study.

Ethical Considerations

The research was approved by the Ethics, Deontology and Scientific Integrity Committee of Nantes University (IORG0011023). The regulatory qualification for noninterventional research was submitted to and validated by the committee (reference number 26062023-1). The recruitment of participants was voluntary, with written consent obtained. Data were anonymized. No compensation was provided to participants.

Inclusion and Exclusion Criteria

Included students were in their second year of study in dentistry and were the legal age of majority. This ensured that participants were novices in preclinical training, minimizing the effect of skill improvement that appears with time. At the Faculty of Dental Surgery of Nantes, no hands-on practical training occurs in the first year. Therefore, we aimed to assess the students' baseline technical performance.

The exclusion criteria were as follows: students who were repeating their second year of dental studies, students who were younger than 18 years, and students with medical conditions that could impair their ability to complete a practical training course under the same conditions as the other participants.

Sample Design

The 2 investigators made an oral presentation on the study's organization and objectives.

After the inclusion criteria were applied, the participants were invited to complete a questionnaire to determine their allocation to either the player or nonplayer group. Participants who indicated that they engaged in play "several times a day," "daily," "regularly" and "occasionally" were categorized as "players." In contrast, those who played "rarely" or "never" were designated as nonplayers. Two extreme subcategories were also defined for participants, including those who indicated that they engaged in play "several times a day," who were categorized as extreme players, and those who "never" played, who were designated as extreme nonplayers. Participants were coded anonymously to ensure that the investigator remained blinded.

Outcomes

The primary outcome of this study was the influence of video games on the preclinical abilities of novice dental students. Comparisons between players and nonplayers were conducted for three skill domains that are particularly relevant in preclinical training: (1) 3D spatial representation, (2) accuracy of technical gestures, and (3) speed of execution. The secondary outcome was to compare these same domains between the extreme players and extreme nonplayers.

Procedure and Outcome Measurements

Following the recruitment stage, participants completed a preliminary questionnaire to verify their eligibility and to allocate them to the appropriate groups. Subsequently, all participants were evaluated according to 3 distinct tests, corresponding to the 3 abilities mentioned above (Figure 1).

Figure 1. Flowchart of the study procedure.

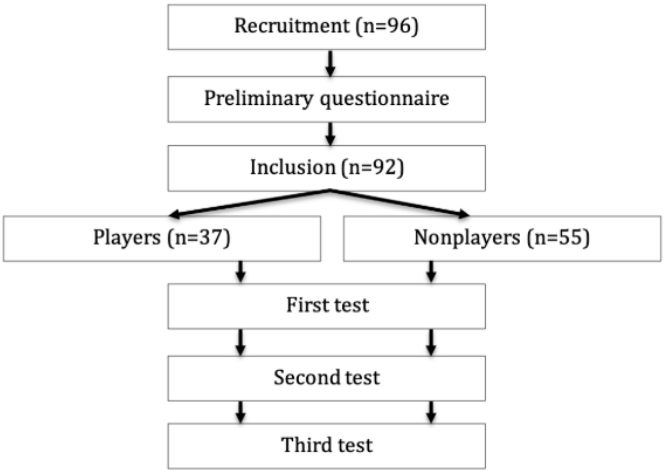

The first evaluation comprised a task in which participants were required to match 3D and 2D images of the same tooth. This task served to assess the participants' capacity to recognize an object from diverse perspectives and to evaluate their visuospatial abilities. Participants were asked to associate videos of axial and sagittal sections of a micro–computed tomography scan and 3D scanned models with the corresponding standardized anatomical board representing the 5 faces of the tooth (occlusal, mesial, distal, vestibular, and palatal/lingual; Figure 2). For each tooth, the participants received a point if they failed to correctly associate the projected 3D video representation with the correct 2D representation. The task consisted of accurately identifying 9 teeth, resulting in 9 assessments per participant. As one point was assigned for each error, higher scores indicated poorer performance.

XSL•FO
RenderX

Figure 2. Presentation of the documents used in the first test. Top: participants associate videos of axial and sagittal sections of micro–computed tomography scans and 3D scanned models. Bottom: standardized anatomical sheets representing the 5 faces of the tooth.

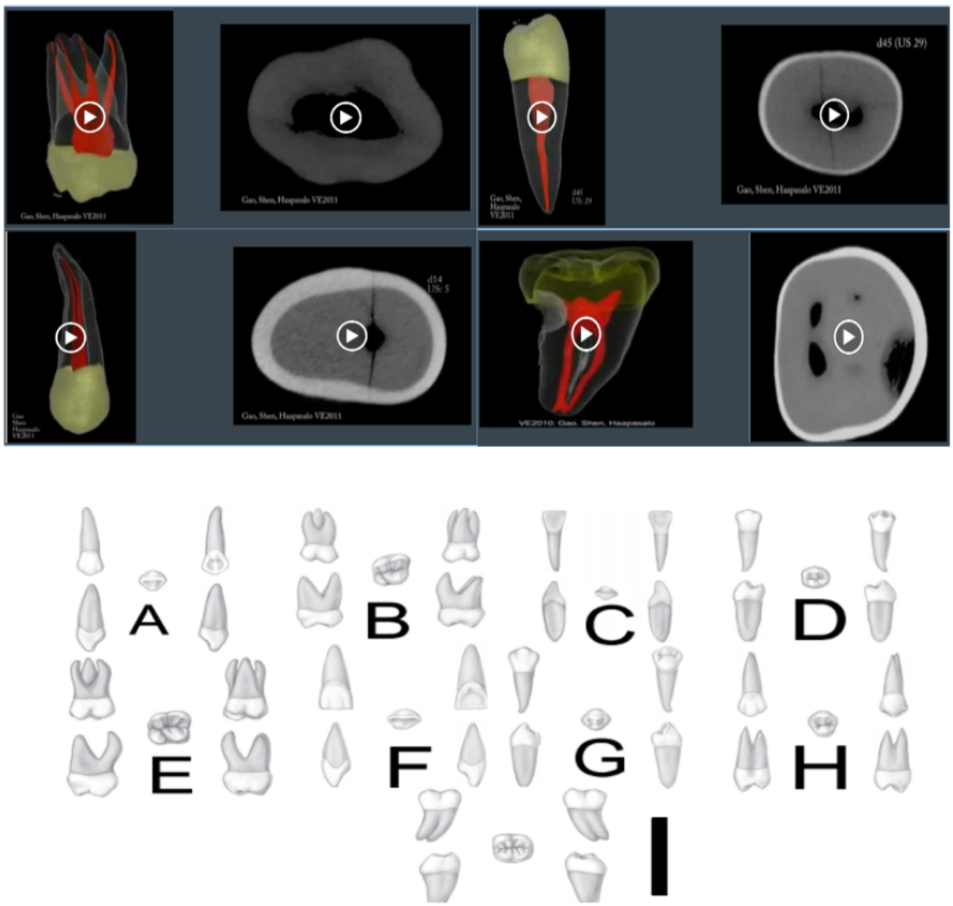

The second test assessed gesture precision and hand-eye coordination. We used a modified version of the Precision Manual Dexterity Test (PMD Test) as described by Bowers et al [21]. The test involved performing the most precise possible penetrations of 80 targets printed on a sheet of paper stretched over a soft support, using a handheld 10/100 steel endodontic file. Participants used their right hand to perform the task on the 4 rectangles located on the right side of the sheet (40 targets) and their left hand for the 4 rectangles on the left side (40 targets) (Figure 3). The test sheet was divided into 8 target zones, each containing 10 randomly arranged targets with an enlarged diameter of 1 mm, based on the original PMD Test, which was performed under optical magnification (×8 and ×2.5). Scoring was as follows: 3 points were awarded when the penetration point was entirely within the target, 2 points when more than 50% but less than 100% of the point was within the target, 1 point when less than 50% was within the target, and zero when the point was completely outside the target. Each sheet was visually compared to a prepunched reference sheet under backlighting for scoring. The maximum total score for each participant was 240 points.

For the third test, the participants were required to undertake a complete pulpotomy procedure on a temporary mandibular molar resin model. The objective of this evaluation was to ascertain the efficiency with which a technical act was executed. In order to simulate clinical conditions and ensure maximum realism, a resin model of a right temporary second mandibular molar with a pulp chamber (Frasaco GmbH) was used (Figure 4). All participants received standardized written and visual instructions. The investigator demonstrated the procedure before the participants to clarify the success criterion. The success criterion was defined as the complete absence of any overhang above the chamber. Participants activated a stopwatch at the start of the procedure and stopped it upon reaching the chamber and notified the investigators, after which they were asked to remove any remaining overhangs. They stopped the stopwatch whenever they believed the success criterion had been met and requested an assessment from the investigators. The investigators then visually inspected the chamber opening to determine whether it met the quality requirement. If not, participants were instructed to continue; if acceptable, the procedure was deemed complete. The time (in seconds) required by the participants to correctly perform this pulpotomy by gaining access to the pulp chamber without producing an overhang was measured.

XSL•FO
RenderX

Figure 3. Illustration of the second test (modified Precision Manual Dexterity Test).

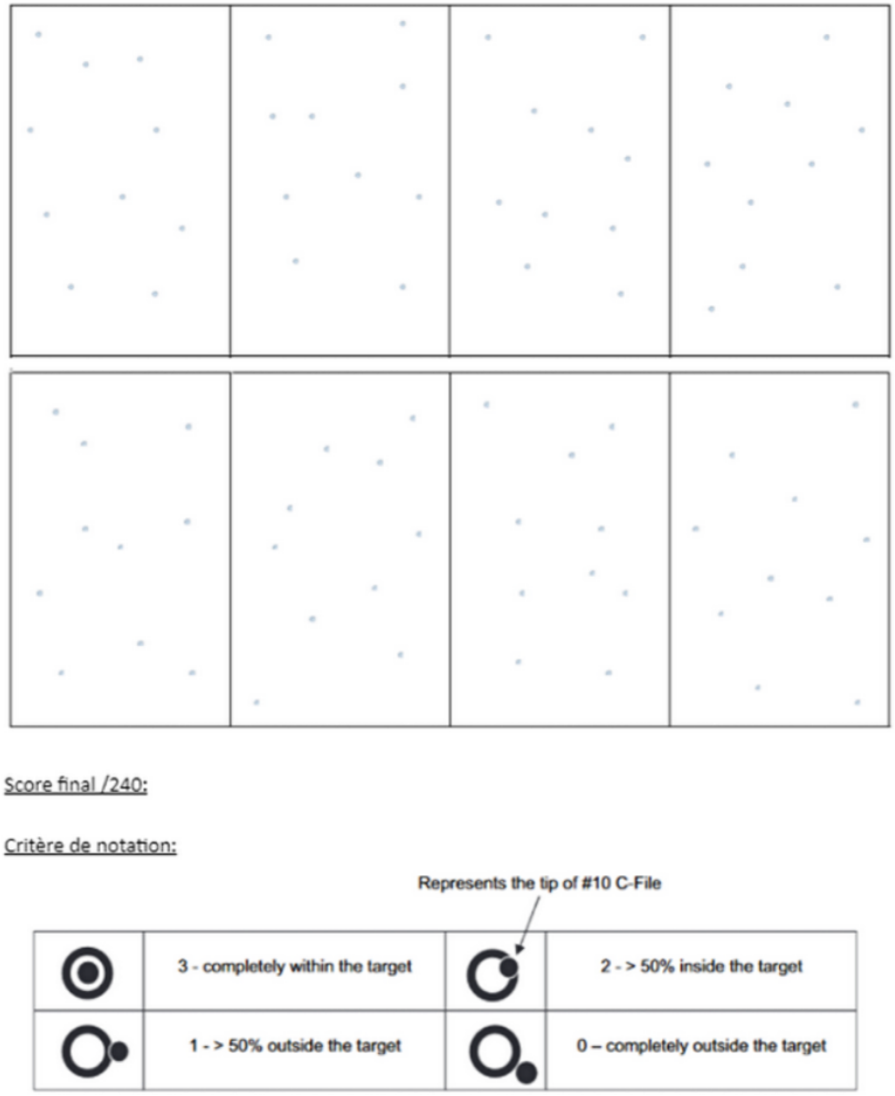

Figure 4. Illustration of the third test.

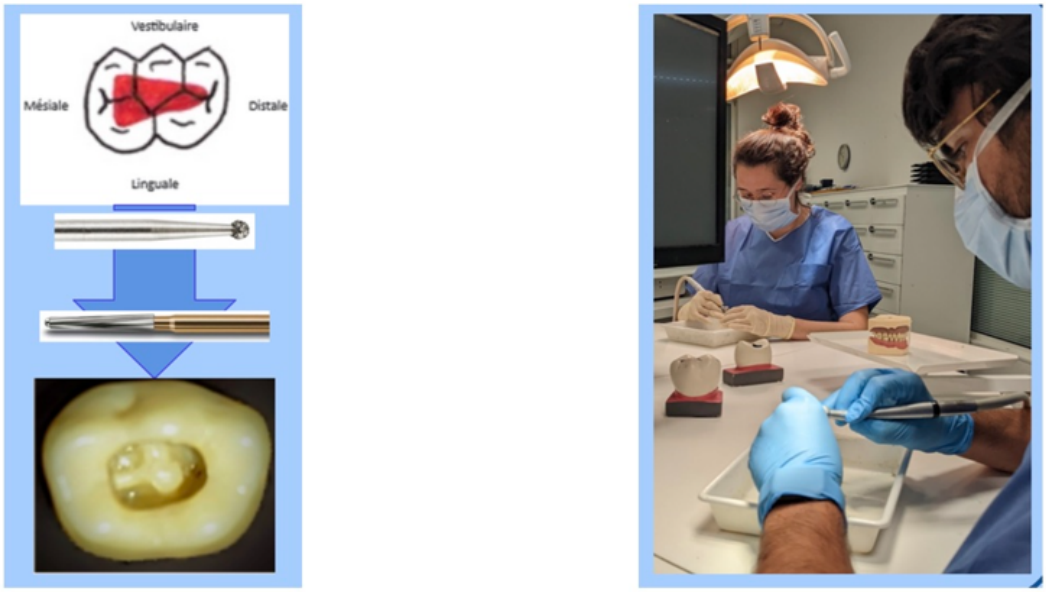

XSL•FO
RenderX

The comprehensibility and ease of use of the 3 tests were assessed by the 3 evaluators present during the sessions.

Statistical Methods

Quantitative variables were described in terms of means and SDs. The collected data were then subjected to comparative analysis between the player and nonplayer groups. The parametric statistical test used in this study was the 2-tailed Student *t* test for independent samples. A further analysis was conducted between the 2 extreme groups: extreme nonplayers (subgroup "never") and extreme players (subgroups "several times a day" and "daily"), using nonparametric tests, such as the Mann-Whitney *U* test for independent samples. The statistical significance threshold was set at $P<.05$.

Results

Participant Flow

Ninety-two second year dental students were included in this study (mean age 19.49, SD 0.8 years; median age 20, IQR 19-21 years). Most of the students were women (n=66, 72%). Group allocation was based on responses to the question "How often do you think you play video games?" Thirty-seven participants were placed in the player group (40%), while 55 were placed in the nonplayer group (60%).

Analysis of Outcome Measurements

There was no statistical difference between the player and nonplayer groups on the first test, assessing visuospatial ability ($P=.051$), and the modified PMD Test ($P=.17$), even if the nonplayer group tended to make slightly fewer mistakes on the first test and obtained a higher score on the second one. There was still no statistical difference on the third test ($P=.87$), but conversely, the nonplayer group tended to require slightly more time to perform the pulpotomy compared with the player group (Table 1).

Table 1. Comparison of scores on the 3 tests between the player and nonplayer groups.

	Nonplayers (n=55), mean (SD)	Players (n=37), mean (SD)	*P* value
Anatomy test, number of mistakes (out of 9)	2.75 (2.12)	3.67 (2.02)	.051
Modified Precision Manual Dexterity Test, score (out of 240)	146.65 (23.4)	139.36 (20.66)	.17
Pulpotomy test, time (in seconds)	311.85 (137.3)	306.92 (148.34)	.87

In addition, a subgroup analysis between the extremes of each group showed that the extreme nonplayer group appeared to make significantly fewer mistakes on the recognition test than the extreme player group ($P=.048$). However, there was no statistical difference between the extreme nonplayer group on the modified PMD Test ($P=.22$) or in the time required to perform the pulpotomy ($P>.99$) compared to the extreme player group (Table 2).

Table 2. Comparison of scores on the 3 tests between the extreme player and extreme nonplayer groups.

	Extreme nonplayers (n=34), mean (SD)	Extreme players (n=9), mean (SD)	*P* value
Anatomy test, number of mistakes (out of 9)	2.5 (2.09)	4 (1.87)	.048
Modified Precision Manual Dexterity Test, score (out of 240)	147.09 (24.34)	138.67(16.23)	.22
Pulpotomy test, time (in seconds)	296.7 (123.32)	321.56 (190.38)	>.99

Discussion

Principal Findings

The potential of video games as a tool to enhance manual dexterity and cognitive skills has been a subject of considerable interest. This study attempted to explore whether this potential may extend to the field of dentistry.

The first test was conceived with the objective of appraising the visuospatial aptitudes of the participants. In contrast to our findings, Mallow et al [22] demonstrated that participants who were novices in bronchoscopy performed better on a mental rotation test of a 3D figure than nonplayers prior to any training. To measure this, the authors used a psychometric test (the Purdue Spatial Visualization Test), which revealed a significant discrepancy between the 2 groups with respect to their baseline performance. In contrast, Kennedy et al [23] did not demonstrate a substantial discrepancy in mental capacity for 3D representations between players and nonplayers.

The second test assessed subjects' psychomotor skills with a modified PMD Test. Kennedy et al [23] also showed that students who played video games at least 7 hours per week demonstrated significantly better psychomotor skills than students who did not play video games regularly, suggesting that video gaming experience may facilitate surgical learning [23]. According to Rosser et al [15], students who were players made fewer mistakes than nonplayers. In addition, if they played more than 3 hours per week, their performance increased. Sammut et al [14] observed 3 surgical tasks performed by young physicians and evaluated them. The results were correlated with the physicians' previous experience with video games; the researchers found that the physicians who had played video games had a better accuracy rate.

XSL•FO
RenderX

The third test aimed to evaluate the speed of performing a technical procedure. Contrary to our results, Dye et al [7] suggested that action video games could significantly reduce reaction time without compromising accuracy. Moreover, according to Rosser et al [15], player participants executed a task more quickly than nonplayers. Shane et al [5] also demonstrated that students with video game experience had a shorter learning curve than those without, especially when learning 2 concurrent tasks. Without any medical knowledge, and with only a few instructions on how to perform the task, adolescent players were faster than experienced surgeons on simulators, emphasizing the importance of the ease with which these tools can be used [24]. Thus, despite a lack of medical knowledge, handling the training tools was easier for the adolescents than for the experienced surgeons.

A lot of effort has been devoted to quantify the impact of game-based training on cognitive abilities, and a significant proportion of research has been dedicated to exploring the formative role of video gaming and its influence on the learning process [22,25]. As hypothesized, video games may serve as a complementary training tool in the acquisition of surgical skills; this is supported by their historical use in simulation-based training [17]. Moreover, the transfer effect increases as visual similarity increases [26]. Thus, performance in spatially intense and visually challenging video games could be a predictor of surgical simulation outcomes. Past studies have also highlighted the importance of video gaming typology in participants' previous experience [27,28]. Players differ according to their habits, such as the type of games they play or the frequency with which they play them. Indeed, as has been demonstrated in previous studies, players of action games appear to demonstrate superior visuomotor control and faster information processing speeds than players of nonaction games [27,29]. Indeed, surgical performance criteria were enhanced most in players of action games. It is posited that action games help reduce the processing time of information requiring a rapid response, without altering the motor response [30]. However, most studies do not clearly distinguish between gaming experience and gaming skill, which remain poorly defined concepts [14].

Harrington et al [31] investigated the relationship between visuospatial performance and nonsurgical experiences during the learning process. Participants were assessed during the skill acquisition phase and the maintenance phase. Players were found to perform better than other participants during the maintenance phase. Middleton et al [32] suggested that for educational purposes, practicing video games could improve bimanual dexterity, help train the surgeon's nondominant hand, and accelerate the acquisition of basic surgical skills. Psychologically, this enhanced performance could relate to the "flow" concept described by Csikszentmihalyi, an immersive mental state of deep concentration and satisfaction [33]. This contributes to the rationale for active and immersive learning strategies. Nevertheless, contradictory results also exist [18,34], preventing definitive conclusions. In several studies, performance improvements in players appeared only after repeated attempts, suggesting a potentially accelerated learning curve rather than superior baseline skills.

Limitations

Several methodological limitations of this study should be considered. There were unequal sample sizes between groups, with the player group being smaller. This may have reduced the statistical power and precision of the results. As this was the first study on this topic in the field of dental education, we drew comparisons with research on medical students at more advanced levels, limiting generalizability. Discipline-specific testing tools were used, which may not be valid in the context of video game–related research, hindering comparisons with the existing literature. Furthermore, no restrictions were placed on game types or platforms—given that different genres develop different skills, this may have obscured specific effects of action-heavy games. Self-reported gaming frequency may also lack accuracy, and categorical ranges (eg, "occasionally") were not precisely standardized.

Improvements should be considered for future studies, such as a longitudinal design, which would allow evaluation of learning progression over time [31,35]. This preliminary work has facilitated the validation of 2 of the 3 measurement tools for such a longitudinal study, which could aim to observe the learning curve for technical skills among dental students in relation to their video game–playing habits. Indeed, the simulation of technical skills using a pulpotomy test and the modified PMD Test appear to be relevant measurement tools. However, the visuospatial aptitude test does not appear to be an appropriate instrument for the population under study. As this was a pilot study, we tried to create a tool that would evaluate the concept of visuospatial assessment but be specific to the field of dentistry and easy for students to use and understand. However, the lack of scope (the total score was only 9) and the difficulty students had in understanding it during assessment sessions, as well as its lack of validation, lead us to question its future use. Vandenberg and Kuse [36] introduced an alternative method, the mental rotation test, which aims to assess these skills prior to the start of studies. Collecting participants' feedback on task difficulty could identify barriers and refine test calibration. Nevertheless, this preliminary study allowed identification of key methodological issues ahead of the now-published protocol for a prospective longitudinal investigation [37].

Conclusion

The results of this preliminary study did not demonstrate that regular video game playing improved basic preclinical performance in novice dental students. Nevertheless, this work enabled the refinement and validation of assessment instruments for further research. Future longitudinal studies are necessary to determine whether dental students who engage in gaming acquire technical skills more efficiently during preclinical training than students who do not.

XSL•FO
RenderX

Acknowledgments

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article. No part of this manuscript was written using generative artificial intelligence.

Data Availability

The data sets generated during and/or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

TR, RC, and TP contributed to the design and methodology of the study. AG was responsible for statistical analysis. CB and SS contributed to the drafting of the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Mark G. The educational benefits of videogames. Educ Health 2002;20(3):47-51 [FREE Full text]
2. Games Market Reports and Forecasts. NewZoo. URL: https://newzoo.com/products/reports/global-games-market-report [accessed 2025-11-26]
3. Green CS, Bavelier D. Learning, attentional control, and action video games. Curr Biol 2012 Mar 20;22(6):R197-R206 [FREE Full text] [doi: 10.1016/j.cub.2012.02.012] [Medline: 22440805]
4. Borecki L, Tolstych K, Pokorski M. Computer games and fine motor skills. In: Pokorski M, editor. Respiratory Regulation - Clinical Advances. Dordrecht: Springer; 2013.
5. Shane MD, Pettitt BJ, Morgenthal CB, Smith CD. Should surgical novices trade their retractors for joysticks? Videogame experience decreases the time needed to acquire surgical skills. Surg Endosc 2008 May;22(5):1294-1297. [doi: 10.1007/s00464-007-9614-0] [Medline: 17972136]
6. Boot WR, Kramer AF, Simons DJ, Fabiani M, Gratton G. The effects of video game playing on attention, memory, and executive control. Acta Psychol (Amst) 2008 Nov;129(3):387-398. [doi: 10.1016/j.actpsy.2008.09.005] [Medline: 18929349]
7. Dye MW, Green CS, Bavelier D. Increasing speed of processing with action video games. Curr Dir Psychol Sci 2009;18(6):321-326 [FREE Full text] [doi: 10.1111/j.1467-8721.2009.01660.x] [Medline: 20485453]
8. Kühn S, Romanowski A, Schilling C, Lorenz R, Mörsen C, Seiferth N, et al. The neural basis of video gaming. Transl Psychiatry 2011 Nov 15;1(11):e53 [FREE Full text] [doi: 10.1038/tp.2011.53] [Medline: 22833208]
9. Anderson CA, Shibuya A, Ihori N, Swing EL, Bushman BJ, Sakamoto A, et al. Violent video game effects on aggression, empathy, and prosocial behavior in eastern and western countries: a meta-analytic review. Psychol Bull 2010 Mar;136(2):151-173. [doi: 10.1037/a0018251] [Medline: 20192553]
10. Khorasanchi M, Hoseinzadeh M, Khadem Rezaeian M, Kazemian A, Moradi A, Sarabadani J. The development of a serious game for laser applications in dentistry and the evaluation of dental students' satisfaction. BMC Med Educ 2024 May 28;24(1):583 [FREE Full text] [doi: 10.1186/s12909-024-05563-3] [Medline: 38807167]
11. Fernández-Gómez F, Cosin-Villanueva M, Almiñana-Pastor P, López-Roldán A. A comparative analysis of game-based learning and conventional learning in dental education. J Dent Educ 2025 Mar;89(3):414-420. [doi: 10.1002/jdd.13747] [Medline: 39436274]
12. Peanchitlertkajorn S, Reynolds PA, Chaisawas M, Krungthong Y, Boonkaew B, Sipiyaruk K. The impact of an online gamified virtual tour on cognitive enhancement in dental practice management. Sci Rep 2024 Nov 06;14(1):26975 [FREE Full text] [doi: 10.1038/s41598-024-75128-3] [Medline: 39505891]
13. Buajeeb W, Reynolds PA, Boontub H, Tangmanpuwadol Y, Sipiyaruk K. Comparison of the effectiveness of a serious game and pre-recorded lecture in diagnosis and treatment planning of oral lesions for dental students. Sci Rep 2024 Dec 27;14(1):30641 [FREE Full text] [doi: 10.1038/s41598-024-83433-0] [Medline: 39730901]
14. Sammut M, Sammut M, Andrejevic P. The benefits of being a video gamer in laparoscopic surgery. Int J Surg 2017 Sep;45:42-46 [FREE Full text] [doi: 10.1016/j.ijsu.2017.07.072] [Medline: 28733118]
15. Rosser JC, Lynch PJ, Cuddihy L, Gentile DA, Klonsky J, Merrell R. The impact of video games on training surgeons in the 21st century. Arch Surg 2007 Feb;142(2):181-6; discusssion 186. [doi: 10.1001/archsurg.142.2.181] [Medline: 17309970]
16. Datta R, Chon S, Dratsch T, Timmermann F, Müller L, Plum PS, et al. Are gamers better laparoscopic surgeons? Impact of gaming skills on laparoscopic performance in "Generation Y" students. PLoS One 2020;15(8):e0232341 [FREE Full text] [doi: 10.1371/journal.pone.0232341] [Medline: 32845892]
17. Gupta A, Lawendy B, Goldenberg MG, Grober E, Lee JY, Perlis N. Can video games enhance surgical skills acquisition for medical students? A systematic review. Surgery 2021 Apr;169(4):821-829. [doi: 10.1016/j.surg.2020.11.034] [Medline: 33419578]

XSL•FO
RenderX

18. Glassman D, Yiasemidou M, Ishii H, Somani BK, Ahmed K, Biyani CS. Effect of playing video games on laparoscopic skills performance: a systematic review. J Endourol 2016 Feb;30(2):146-152. [doi: 10.1089/end.2015.0425] [Medline: 26414043]
19. Kron FW, Gjerde CL, Sen A, Fetters MD. Medical student attitudes toward video games and related new media technologies in medical education. BMC Med Educ 2010 Jun 24;10:50 [FREE Full text] [doi: 10.1186/1472-6920-10-50] [Medline: 20576125]
20. Hegarty M, Keehner M, Khooshabeh P, Montello DR. How spatial abilities enhance, and are enhanced by, dental education. Learn Individ Differ 2009 Jan;19(1):61-70. [doi: 10.1016/j.lindif.2008.04.006]
21. Bowers DJ, Glickman GN, Solomon ES, He J. Magnification's effect on endodontic fine motor skills. J Endod 2010 Jul;36(7):1135-1138. [doi: 10.1016/j.joen.2010.03.003] [Medline: 20630285]
22. Mallow C, Shafiq M, Thiboutot J, Yu DH, Batra H, Lunz D, et al. Impact of video game cross-training on learning bronchoscopy. A pilot randomized controlled trial. ATS Scholar 2020 Jun;1(2):134-144. [doi: 10.34197/ats-scholar.2019-0015oc]
23. Kennedy A, Boyle E, Traynor O, Walsh T, Hill A. Video gaming enhances psychomotor skills but not visuospatial and perceptual abilities in surgical trainees. J Surg Educ 2011;68(5):414-420. [doi: 10.1016/j.jsurg.2011.03.009] [Medline: 21821223]
24. Fanning J, Fenton B, Johnson C, Johnson J, Rehman S. Comparison of teenaged video gamers vs PGY-I residents in obstetrics and gynecology on a laparoscopic simulator. J Minim Invasive Gynecol 2011;18(2):169-172. [doi: 10.1016/j.jmig.2010.11.002] [Medline: 21215709]
25. Khimani K, Koshevarova V, Mathew AA, Gupta AK, Schmitz-Brown M, Gupta PK. Video game warm-up may suggest improvement in ophthalmic surgical simulator performance in surgically naïve students. Clin Ophthalmol 2022;16:2561-2568 [FREE Full text] [doi: 10.2147/OPTH.S369697] [Medline: 35978609]
26. Schlickum MK, Hedman L, Enochsson L, Kjellin A, Felländer-Tsai L. Systematic video game training in surgical novices improves performance in virtual reality endoscopic surgical simulators: a prospective randomized study. World J Surg 2009 Nov;33(11):2360-2367. [doi: 10.1007/s00268-009-0151-y] [Medline: 19649553]
27. Li L, Chen R, Chen J. Playing action video games improves visuomotor control. Psychol Sci 2016 Aug;27(8):1092-1108. [doi: 10.1177/0956797616650300] [Medline: 27485132]
28. de Araujo TB, Silveira FR, Souza DLS, Strey YTM, Flores CD, Webster RS. Impact of video game genre on surgical skills development: a feasibility study. J Surg Res 2016 Mar;201(1):235-243. [doi: 10.1016/j.jss.2015.07.035] [Medline: 26850208]
29. Green CS, Bavelier D. Effect of action video games on the spatial distribution of visuospatial attention. J Exp Psychol Hum Percept Perform 2006 Dec;32(6):1465-1478 [FREE Full text] [doi: 10.1037/0096-1523.32.6.1465] [Medline: 17154785]
30. van Ravenzwaaij D, Boekel W, Forstmann BU, Ratcliff R, Wagenmakers E. Action video games do not improve the speed of information processing in simple perceptual tasks. J Exp Psychol Gen 2014 Oct;143(5):1794-1805 [FREE Full text] [doi: 10.1037/a0036923] [Medline: 24933517]
31. Harrington CM, Dicker P, Traynor O, Kavanagh DO. Visuospatial abilities and fine motor experiences influence acquisition and maintenance of fundamentals of laparoscopic surgery (FLS) task performance. Surg Endosc 2018 Nov;32(11):4639-4648. [doi: 10.1007/s00464-018-6220-2] [Medline: 29777356]
32. Middleton KK, Hamilton T, Tsai P, Middleton DB, Falcone JL, Hamad G. Improved nondominant hand performance on a laparoscopic virtual reality simulator after playing the Nintendo Wii. Surg Endosc 2013 Nov;27(11):4224-4231. [doi: 10.1007/s00464-013-3027-z] [Medline: 23760943]
33. Peifer C, Wolters G, Harmat L, Heutte J, Tan J, Freire T, et al. A scoping review of flow research. Front Psychol 2022;13:815665 [FREE Full text] [doi: 10.3389/fpsyg.2022.815665] [Medline: 35465560]
34. Harper JD, Kaiser S, Ebrahimi K, Lamberton GR, Hadley HR, Ruckle HC, et al. Prior video game exposure does not enhance robotic surgical performance. J Endourol 2007 Oct;21(10):1207-1210. [doi: 10.1089/end.2007.9905] [Medline: 17949327]
35. Giannotti D, Patrizi G, Di Rocco G, Vestri AR, Semproni CP, Fiengo L, et al. Play to become a surgeon: impact of Nintendo Wii training on laparoscopic skills. PLoS One 2013;8(2):e57372 [FREE Full text] [doi: 10.1371/journal.pone.0057372] [Medline: 23460845]
36. Vandenberg SG, Kuse AR. Mental rotations, a group test of three-dimensional spatial visualization. Percept Mot Skills 1978 Oct;47(2):599-604. [doi: 10.2466/pms.1978.47.2.599] [Medline: 724398]
37. Clouet R, Gaudin A, Tourtelier J, Gogendeau C, Serisier S, Prud'homme T. Analyzing the influence of video game and music engagement on technical skills acquisition in dental students in preclinical phase: protocol for a prospective, controlled, longitudinal study. JMIR Res Protoc 2024 Sep 13;13:e55738 [FREE Full text] [doi: 10.2196/55738] [Medline: 39269750]

Abbreviations

PMD Test: Precision Manual Dexterity Test

XSL•FO
RenderX

Edited by A Schwartz, M Balcarras; submitted 03.07.25; peer-reviewed by M Marty, A Gazis, H Zou; comments to author 17.10.25; revised version received 20.11.25; accepted 20.11.25; published 03.12.25.

Please cite as:
Clouet R, Remaud T, Boeffard C, Serisier S, Gaudin A, Prud'homme T
Impact of the Practice of Playing Video Games on Technical Skills Development in Preclinical Dental Education: Preliminary Cross-Sectional Observational Study
JMIR Form Res 2025;9:e80082
URL: https://formative.jmir.org/2025/1/e80082
doi:10.2196/80082
PMID:

©Roselyne Clouet, Thomas Remaud, Camille Boeffard, Samuel Serisier, Alexis Gaudin, Tony Prud'homme. Originally published in JMIR Formative Research (https://formative.jmir.org), 03.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Changes in the Neighborhood Built Environment and Chronic Health Conditions in Washington, DC, in 2014-2019: Longitudinal Analysis

Quynh C Nguyen[1], PhD; Riki Doumbia[2*]; Thu T Nguyen[3*], SCD; Xiaohe Yue[3*], MS; Heran Mane[3*], BS; Junaid Merchant[3*], PhD; Tolga Tasdizen[4*], PhD; Mitra Alirezaei[4*], PhD; Pankaj Dipankar[1*], PhD; Dapeng Li[5*], PhD; Penchala Sai Priya Mullaputi[3*], MS; Amrutha Alibilli[3*]; Yulin Hswen[3,6,7*], SCD; Xin He[3*], PhD

[1]National Institute of Nursing Research, National Institutes of Health, Bethesda, MD, United States

[2]Urban Studies Program, Brown University, Providence, RI, United States

[3]Department of Epidemiology and Biostatistics, University of Maryland School of Public Health, College Park, MD, United States

[4]Department of Electrical and Computer Engineering, Scientific Computing and Imaging Institute, University of Utah, Salt Lake City, UT, United States

[5]Department of Geography and the Environment, University of Alabama, Tuscaloosa, AL, United States

[6]Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, CA, United States

[7]College of Computer, Mathematical, and Natural Sciences (CMNS), Artificial Intelligence Interdisciplinary Institute (AIM), University of Maryland, College Park, MD, United States

[*]these authors contributed equally

Corresponding Author:
Quynh C Nguyen, PhD
National Institute of Nursing Research
National Institutes of Health
31 Center Drive, Room 5B03
Bethesda, MD, 20892
United States
Phone: 1 503 309 1098
Email: quynh.nguyen3@nih.gov

Abstract

Background: Google Street View (GSV) images offer a unique and scalable alternative to in-person audits for examining neighborhood built environment characteristics. Additionally, most prior neighborhood studies have relied on cross-sectional designs.

Objective: This study aimed to use GSV images and computer vision to examine longitudinal changes in the built environment, demographic shifts, and health outcomes in Washington, DC, from 2014 to 2019.

Methods: In total, 434,115 GSV images were systematically sampled at 100 m intervals along primary and secondary road segments. Convolutional neural networks, a type of deep learning algorithm, were used to extract built environment features from images. Census tract summaries of the neighborhood built environment were created. Multilevel mixed-effects linear models with random intercepts for years and census tracts were used to assess associations between built environment changes and health outcomes, adjusting for covariates, including median age, percentage male, percentage Hispanic, percentage African American, percentage college educated, percentage owner-occupied housing, and median household income.

Results: Washington, DC, experienced a shift toward higher-density housing, with non-single-family homes rising from 66% to 72% of the housing stock. Single-lane roads increased from 37% to 42%, suggesting a shift toward more sustainable and compact urban forms. Gentrification trends were reflected in a rise in college-educated residents (16%-41%), a US $17,490 increase in the median household income, and a US $159,600 increase in property values. Longitudinal analyses revealed that increased construction activity was associated with lower rates of obesity, diabetes, high cholesterol, and cancer, while growth in non-single-family housing was correlated with reductions in the prevalence of obesity and diabetes. However, neighborhoods with higher proportions of African American residents experienced reduced construction activity.

XSL•FO
RenderX

Conclusions: Washington, DC, has experienced significant urban transformation, marked by substantial changes in neighborhood built environments and demographic shifts. Urban development is associated with reduced prevalence of chronic conditions. These findings highlight the complex interplay between urban development, demographic changes, and health, underscoring the need for future research to explore the broader impacts of neighborhood built environment changes on community composition and health outcomes. GSV imagery, along with advances in computer vision, can aid in the acceleration of neighborhood studies.

(JMIR Form Res 2025;9:e74195) doi:10.2196/74195

KEYWORDS

urban development; Google Street View; temporal analysis; health outcomes; computer vision; artificial intelligence

Introduction

Background

Neighborhood characteristics have been identified as critical determinants of various health behaviors and outcomes [1,2]. Previous studies have examined neighborhood walkability, streetscape design, residential density, and land use mix [3-5]. A previous Google Street View (GSV) study analyzing 31 million images found that neighborhoods with more non-single-family homes, an indicator of mixed land use, are associated with lower rates of diabetes and obesity [6]. Another North American study found that a greater land use mix is associated with reduced obesity prevalence [7]. Walkability has been associated with higher physical activity levels, reduced obesity and diabetes risk, and a lower waist-to-hip ratio [8-11]. A study conducted by Carson et al [12] involving 690 participants from 380 census block groups in Baltimore, Seattle, and Washington, DC, metropolitan areas found that increasing neighborhood walkability is associated with increased youth total physical activity [12]. Other studies, both domestic and international, have identified links between built environment characteristics and health outcomes in youth [13-15].

Sidewalks play a critical role in pedestrian safety by providing a designated space separate from vehicular traffic [16]. The risk of pedestrian-vehicle collisions is more than twice as high in areas lacking sidewalks compared to those with them [17]. Sidewalks are particularly important in neighborhoods where children play outside with limited adult supervision [18]. A national investigation [19] by Johns Hopkins University analyzed the effects of lane width on traffic safety across various urban settings and found that lane widths of 10 feet are most appropriate in urban areas, as they balance safety and mobility. Wider lanes (over 11 feet) may lead to unintended speeding and occupy valuable space that could be allocated to other modes of transportation, such as wider sidewalks or bike lanes. These findings suggest that narrower lanes can contribute to traffic calming and improve pedestrian safety [17-19]. The ABQ Streets Project evaluated alternative street designs to retrofit typical residential streets in Albuquerque [20]. The study [20] assessed existing neighborhood design features and proposed modifications, including the addition of sidewalks and the implementation of single-lane roads; the findings suggested that such designs can improve pedestrian safety and accessibility. Collectively, these studies highlight the importance of infrastructural elements to promote safety, accessibility, and livability in urban environments.

Although much of the existing literature has been cross-sectional, offering only a snapshot in time and limiting our ability to assess the dynamic nature of neighborhoods, longitudinal research is essential for identifying how neighborhood changes over time impact health. One longitudinal cohort study (2013-2018) [21] in low-income, predominantly African American Pittsburgh neighborhoods found that enhancements in urban design, such as public art, landscaping, and signage, are linked to reduced wakefulness after sleep onset, highlighting the potential impact of environmental aesthetics on sleep quality. Another longitudinal study [22] found that people living in neighborhoods with high walkability, reduced urban sprawl, and increased recreational access have a lower mean BMI; the meta-analysis of various longitudinal studies further demonstrated that changes in built characteristics, such as enhanced walkability and access to amenities in urban redevelopment, lead to substantial improvements in cardiometabolic health [22].

Although these studies offer valuable insights, many rely on resource-intensive in-person audits or traditional geographic information system (GIS) data, which may lack the spatial granularity or visual detail needed to capture microlevel environmental change. As such, there remains a critical gap in longitudinal neighborhood health research that leverages scalable, low-cost, and reproducible data sources, such as GSV imagery processed with computer vision, to examine built environment changes. Additionally, previous studies have largely relied on cross-sectional surveys, government property records, and satellite-derived geospatial datasets (eg, National Land Cover Database) [23-25]. Although these approaches certainly provide valuable public health surveillance insights, they have critical limitations: self-reported surveys may be subject to bias, official records may lack real-time information, and satellite images may not have street-level granularity. In contrast, GSV offers large-scale, real-time, street-level data with widespread coverage. GSV images reveal reliable physical features that are consistent with topological field assessments [26]. Furthermore, GSV built environment characteristics have been linked with chronic health outcomes by county and zip code levels [27,28].

For researchers interested in neighborhood built environment characteristics and longitudinal changes, GSV images provide a unique data source to examine these characteristics without the need for time-consuming and expensive in-person audits over multiple years. We used GSV images from 2014 to 2019 to quantify longitudinal changes in neighborhood built environments over a 6-year study period in Washington, DC.

XSL•FO
RenderX

We selected Washington, DC, due to its rapid urban development and evolving neighborhood characteristics during the study period. Between 2000 and 2013, Washington, DC, was rated as the most intensely gentrifying city in the United States, with approximately 20,000 African Americans displaced from their neighborhoods [29]. Data from 2012 to 2017 ranked the city of Washington, DC, in 13th place, due to gentrification increasing in other US cities [30].

Study Aims

This study aimed to investigate changes in the built environment in Washington, DC, derived from GSV images. Furthermore, the study examined how changes in built environment characteristics are associated with health outcomes, including the chronic disease burden (eg, obesity, diabetes) and mental health. The study also explored links between urban development and the neighborhood sociodemographic composition, including shifts in racial/ethnic minorities and the median household income.

Study Hypotheses

We hypothesized the following:

- Urban development increases, as indicated by built environment characteristics derived from GSV images.
- Urban development is associated with a decreased chronic disease burden (obesity, diabetes) and mental distress.
- Changes in the sociodemographic neighborhood composition are associated with changes in the built environment.

Methods

Google Street View Data Collection

We collected historical GSV images from Washington, DC, spanning 2014-2019. In total, 434,115 GSV images were collected using the GSV Image application programming interface (API). The GSV API is a tool from Google Maps that allows users to access and download 360° panoramic images of streets worldwide. These images are captured by GSV cars, equipped with advanced cameras, making detailed visual data available. Static images were retrieved by sending API calls with parameters such as year, latitude/longitude coordinates, and camera angle [31]. Images were sampled every 100 m along all primary and secondary roads, with four images captured at each point (facing north, south, east, and west) to create a full view of the built environment.

For longitudinal GSV data collection, we used the Python package *Streetview* [32] to retrieve panorama IDs: unique identifiers for GSV panoramic images, including the year captured. Using these panorama IDs with the GSV Static API, we were able to collect corresponding GSV images for each geocoordinate or its nearest available point. We used the same sampling coordinates for each year of GSV data collection (ie, images were downloaded to capture the same location for each year between 2014 and 2019). However, although we used the same coordinates to sample images each year, image availability could differ depending on whether an updated image was available for any given year of data collection. On average, GSV data were available for 158 census tracts (88% of all census tracts) in Washington, DC, across the years of the study period.

Built Environment Indicators

We selected neighborhood characteristics that were shown in prior research to influence health outcomes [33] and that could be robustly identified using computer vision models. Additionally, we incorporated less commonly studied indicators, such as stop signs and road construction, to expand the existing neighborhood literature. In total, we examined the following six built environment indicators to represent (1) urban development (non-single-family homes, single-lane roads, construction, two or more cars) and (2) walkability (sidewalks, streetlights, stop signs). In this study, stop signs were used as a measure of intersection density, increased pedestrian safety, and area walkability, all of which are associated with better health outcomes [34]. Road construction was used as a proxy for ongoing urbanization and changes to the built environment. Since residents of sprawling areas tend to live in single-family homes, non-single-family homes were used as an indicator of nearby dense residential and commercial buildings and mixed land use, which have been connected to better health outcomes [6,35]. Single-lane roads were used to indicate residential streets or areas with lower levels of urban development. The presence of two or more cars was used to represent higher traffic density. Sidewalks and streetlights served as walkability indicators, which have been linked to improved health outcomes [36-39].

Image Data and Quality Control Analyses

Next is a summary of our image analysis; for additional details, please refer to prior publications [40,41]. To create training and test datasets for the computer vision models, 18,000 GSV images from the national data collection were manually annotated. Each image was labeled with a binary “yes/no” value for the presence of six built environment indicators: (1) non-single family homes (single family detached house vs other building type), (2) single-lane roads (one-lane road for motor vehicles vs two or more lanes for traffic demarcated by lines), (3) two or more cars (two or more motor vehicles; some vehicles may be in the background), (4) sidewalks (presence of a sidewalk on at least one side of the road), (5) streetlights (presence of a lamp fixture mounted on a pole spaced along a street or road); and (6) road construction (presence of road construction indicated by orange signs, traffic control devices such as barriers or cones, construction equipment and workers). A value of “yes” indicated the presence of the indicator (eg, sidewalk), while “no” indicated its absence. The principal investigator and three graduate research assistants conducted the data annotation. Interrater agreement was above 85% for all neighborhood indicators. Annotation proceeded in small batches (eg, 100 images per batch). To ensure labeling consistency, the annotators first labeled the same batch, and interrater agreement was calculated for each indicator. Discrepancies were resolved through group discussion and consensus. If 85% interrater reliability was not reached, the labelers were assigned a new batch of images to annotate. Training was iterative until 85% interrater reliability was achieved for each neighborhood indicator, at which point the annotators were asked to label 3000-5000 images

XSL•FO
RenderX

independently. These data were then divided into training (80%) and test (20%) datasets.

To generate census tract–level summaries for each GSV-derived built environment characteristic, we used the latitude and longitudinal coordinates associated with each image to assign them to a census tract. Next, for each census tract, we calculated the percentage of the total number of images that contained a given built environment indicator (eg, number of images with a sidewalk/total number of images) × 100 = percentage with sidewalks. We created tertiles and classified each census tract based on its percentage and year of data collection, with the lowest tertile serving as the reference group.

A deep convolutional neural network architecture, Visual Geometry Group (VGG)-19 and Residual Network (ResNet)-18, in TensorFlow was trained with sigmoid cross-entropy with logits as the loss function. These computer vision models were run on 434,115 GSV images from Washington, DC, to produce inferences for the six built environment characteristics. Quality control statistics were as follows: non-single-family homes (accuracy 82%), single-lane roads (accuracy 88%), road construction (accuracy 96%, F_1-score 62%), two or more cars (accuracy 88%, F_1-score 79%), sidewalks (accuracy 84%, F_1-score 81%), streetlights (accuracy 88%, F_1-score 60%), and stop signs (accuracy 90%, F_1-score 89%). Non-single-family homes and single-lane roads were derived from a previous study that did not report F_1-scores [40]. For each year of data, GSV-derived built environment indicators were categorized into tertiles, with the lowest tertile serving as the reference group. This was done to examine potential dose-response associations and to allow for potential changes in the directionality of the associations between tertiles in the built environment characteristics and health outcomes. Using tertiles also allowed comparison with prior GSV-based studies [27,40,42,43]. Sensitivity analyses used linear (continuous) forms of built environment characteristics.

Census Tract Health Outcomes and Sociodemographic Characteristics

Census tract–level health outcomes data for 2014-2019 were obtained from the Centers for Disease Control and Prevention's (CDC) PLACES (Population Level Analysis and Community Estimates) program [44]. PLACES is a collaboration between the CDC, the Robert Wood Johnson Foundation, and the CDC Foundation to provide model-based estimates of health measures for census tracts. Additionally, census tract sociodemographic characteristics were obtained from the 1-year estimates of the American Community Survey (ACS) for the period 2014-2019. ACS census tract–level 1-year estimates were also obtained for median property values and median gross rent costs from the database on "Selected Housing Characteristics." Prior to 2022, the ACS did not adjust home values for inflation, so we did not adjust the values for inflation to preserve the accuracy of the original 1-year estimates.

Statistical Analysis

We calculated descriptive statistics to examine yearly changes in built environment characteristics, sociodemographics, and health outcomes from 2014 to 2019. Longitudinal analyses were conducted using multilevel mixed-effects linear models for each health outcome, with tertiles for six built environment characteristics as the independent variables of interest. These models included random intercepts for census tracts and random slopes for years (centered on 2016). The six built environment indicators analyzed were single-lane roads, the presence of two or more cars, streetlights, non-single-family homes, sidewalks, and road construction. Models controlled for fixed effects, including census tract, median age, percentage male, percentage Hispanic, percentage African American, percentage college educated, median household income, and percentage owner-occupied housing as covariates. All covariates were standardized to have a mean of 0 (SD 1). We also ran separate multilevel mixed-effects linear models to assess whether changes in neighborhood sociodemographics were associated with changes in built environment characteristics using the same random effect's structure and covariate adjustments.

Ethical Considerations

The study was conducted in accordance with the Declaration of Helsinki and was approved by the Institutional Review Board (IRB) of the University of Maryland (protocol number 1074955-8; IRB approval date June 9, 2022).

Results

Temporal Trends in the Built Environment in Washington, DC (2014-2019)

Built environment characteristics changed over the study period, reflecting shifts in the urban landscape of Washington, DC (Figure 1 and Table 1). The proportion of non-single-family homes (a proxy of mixed land use) increased from 70.94% in 2014 to 72.48% in 2019 (Figure 2). The prevalence of single-lane roads rose from 39.53% in 2014 to 42.16% in 2019. Other built environment characteristics (sidewalks, streetlights, two or more cars) increased from 2014 to 2016 and then declined from 2017 to 2019, resulting in lower prevalence rates in 2019 compared to 2014. For instance, the prevalence of sidewalks increased from 2014 (64.12%) to 2016 (67.8%) and then decreased from 2017 to 2019, reaching a prevalence lower than the baseline of 60.83% in 2019. Sensitivity analyses using another GSV-derived walkability indicator (presence of crosswalks) also revealed a similar trend: increasing prevalence between 2014 and 2016, followed by a decrease from 2016 to 2019 (Table S1 in Multimedia Appendix 1). Similarly, streetlight presence increased from 34.85% in 2014 to 39.75% in 2016 but then decreased to 32.55% in 2019. The proportion of two or more cars in GSV images (a proxy of traffic density) increased from 54.19% in 2014 to 56.36% in 2016 but then decreased to 48.65% in 2019 (Figure 2 and Table 1). There was a slight decrease in the prevalence of stop signs, from 2.58% in 2014 to 2.09% in 2019, and a slight decrease in road construction, from 4.54% in 2014 to 3.22% in 2019.

XSL•FO
RenderX

Figure 1. Temporal change in built environment characteristics across census tracts in Washington, DC, 2014-2019.

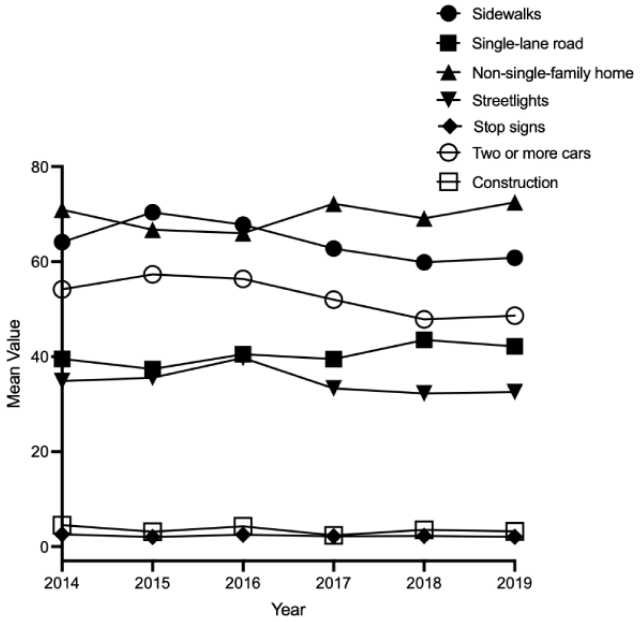

XSL•FO
RenderX

Table 1. Descriptive statistics of neighborhood built characteristics, census tract characteristics, and health outcomes in Washington, DC[a].

Characteristics	Years					
	2014	2015	2016	2017	2018	2019
Built environment characteristics						
Single-lane roads (%), mean (SD)	39.53 (21.22)	37.39 (27.74)	40.53 (23.65)	39.50 (22.72)	43.54 (17.21)	42.16 (15.19)
Stop signs (%), mean (SD)	2.58 (4.82)	2.03 (5.85)	2.51 (5.19)	2.21 (4.07)	2.26 (3.30)	2.09 (1.58)
Two or more cars (%), mean (SD)	54.19 (23.54)	57.35 (28.31)	56.36 (24.49)	52.00 (27.15)	47.85 (18.23)	48.65 (15.81)
Streetlights (%), mean (SD)	34.85 (21.10)	35.57 (27.04)	39.75 (24.76)	33.33 (22.29)	32.26 (15.56)	32.55 (13.90)
Non-single-family homes (%), mean (SD)	70.94 (24.35)	66.74 (31.14)	66.01 (27.19)	72.19 (26.78)	69.12 (22.93)	72.48 (19.86)
Sidewalks (%), mean (SD)	64.12 (24.16)	70.39 (28.79)	67.80 (23.66)	62.77 (27.89)	59.87 (18.76)	60.83 (15.50)
Road construction (%), mean (SD)	4.54 (10.90)	3.16 (10.21)	4.28 (11.09)	2.39 (5.35)	3.55 (5.53)	3.22 (3.94)
Census tracts, n	162	143	152	155	166	169
Census tract characteristics						
Population size (n), mean (SD)	3405 (1370)	3418 (1382)	3428 (1399)	3457 (1382)	3463 (1386)	3468 (1390)
Median age (years), mean (SD)	35.24 (6.79)	35.02 (6.34)	34.93 (6.08)	34.84 (6.00)	34.82 (5.76)	34.79 (5.67)
Percentage male (%), mean (SD)	47.25 (6.10)	47.23 (6.22)	47.47 (5.89)	47.59 (5.88)	47.51 (5.77)	47.46 (5.66)
Percentage African American (%), mean (SD)	55.06 (35.14)	53.87 (34.97)	52.77 (34.68)	51.85 (34.59)	51.11 (34.40)	50.15 (34.19)
Percentage Hispanic (%), mean (SD)	8.76 (8.64)	9.17 (8.61)	9.59 (8.42)	9.89 (8.33)	10.08 (8.44)	10.28 (8.22)
Median household income (US $), mean (SD)	74,026 (41,041)	75,466 (40,972)	78,360 (43,295)	82,936 (45,084)	86,504 (44,706)	91,516 (46,229)
Percentage college educated (%), mean (SD)	15.82 (9.34)	15.92 (9.14)	16.17 (9.27)	16.78 (9.15)	40.05 (24.43)	40.96 (24.29)
Percentage owner-occupied housing (%), mean (SD)	38.35 (20.54)	38.74 (21.12)	38.25 (21.39)	38.65 (20.96)	38.99 (20.85)	39.04 (21.16)
Census tracts, n	178	178	178	178	178	178
Adult health outcomes						
Obesity (%), mean (SD)	27.40 (7.74)	27.78 (8.01)	28.01 (7.86)	28.34 (8.14)	30.00 (7.95)	26.95 (7.85)
Diabetes (%), mean (SD)	10.18 (4.49)	10.37 (4.54)	9.70 (4.21)	10.15 (4.44)	10.70 (4.63)	9.20 (4.04)
High blood pressure (%), mean (SD)	31.53 (10.18)	31.53 (10.18)	31.28 (9.98)	31.65 (9.66)	32.14 (9.79)	32.14 (9.79)
High cholesterol (%), mean (SD)	32.02 (5.39)	32.02 (5.39)	28.93 (4.91)	29.34 (4.93)	27.73 (4.77)	27.73 (4.77)
Cancer (%), mean (SD)	5.18 (1.48)	4.90 (1.41)	4.99 (1.43)	5.71 (1.62)	5.62 (1.62)	5.10 (1.56)
Poor mental health days (%), mean (SD)	12.17 (3.82)	11.73 (3.55)	12.11 (3.79)	14.04 (4.16)	13.78 (3.91)	15.38 (3.25)
Census tracts, n	154	154	154	174	174	174

[a]Data sources: (1) Google Street View (GSV) built environment characteristics, (2) American Community Survey (ACS) census tract sociodemographic characteristics, and (3) Centers for Disease Control (CDC) Prevention Population Level Analysis and Community Estimates (PLACES) health outcomes.

Figure 2 displays the geographic distribution of GSV-derived built environment characteristics across census tracts in Washington, DC, and across time. Stop signs appeared more evenly distributed across Washington, DC, in 2014 than in 2019, while the converse was true for road construction, which seemed more dispersed in 2019 than in 2014. Non-single-family homes gained even more prominence over time, except in the northwestern parts of Washington, DC. The presence of two or

XSL•FO
RenderX

more cars on city streets decreased over time, especially in the southern areas of Washington, DC.

Figure 2. Geographic distribution of built environment characteristics across census tracts in Washington, DC, in 2014 and 2019.

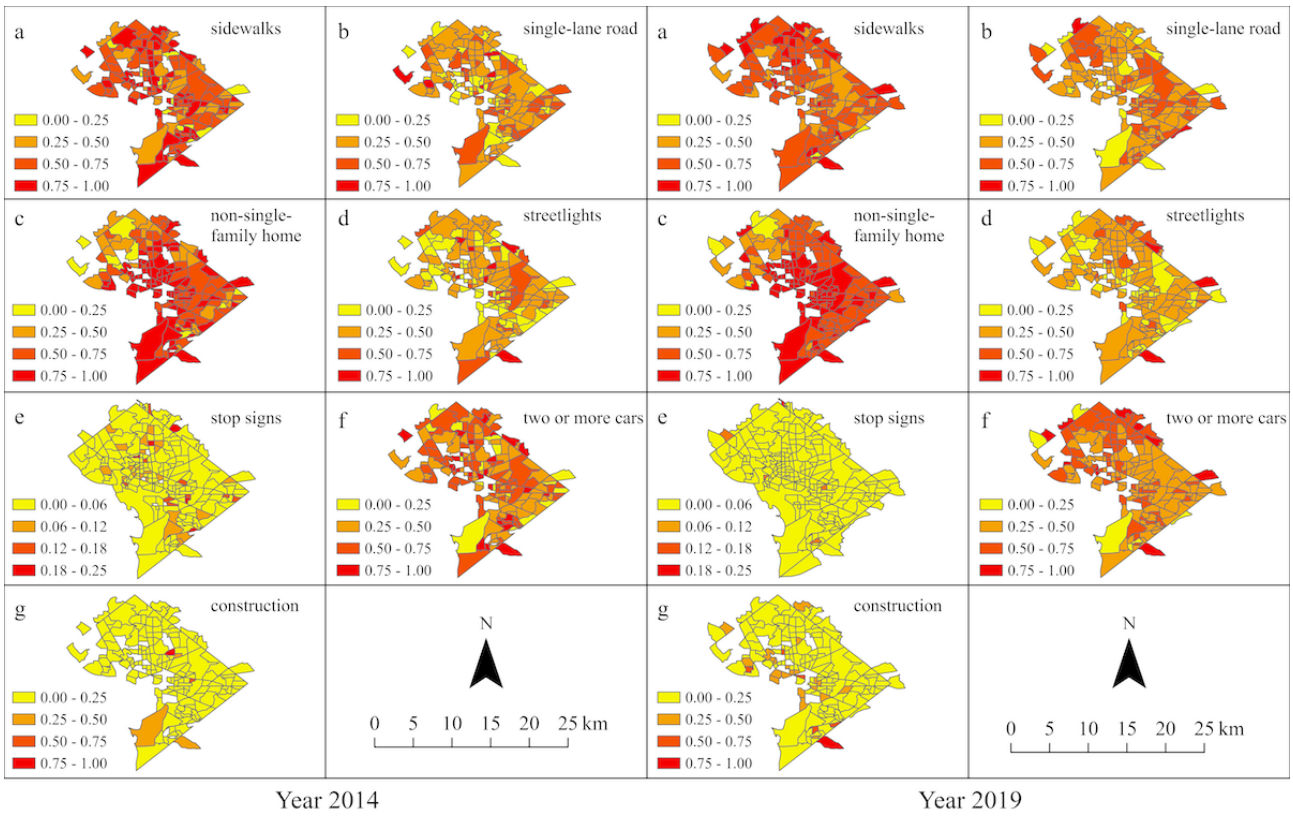

Changes in Property Values, Housing Costs, and Sociodemographics in Washington, DC

Over the study period, median property values in Washington, DC, rose significantly, increasing from US $486,900 in 2014 to US $646,500 in 2019 (a difference of US $159,600; Figure 3A). Similarly, median monthly rent costs increased by almost 20%, rising from US $1360 in 2014 to US $1603 in 2019 (Figure 3B).

XSL•FO
RenderX

Figure 3. Housing prices in Washington, DC, 2014-2019: (A) median property values and (B) median gross rent costs. Data source: American Community Survey.

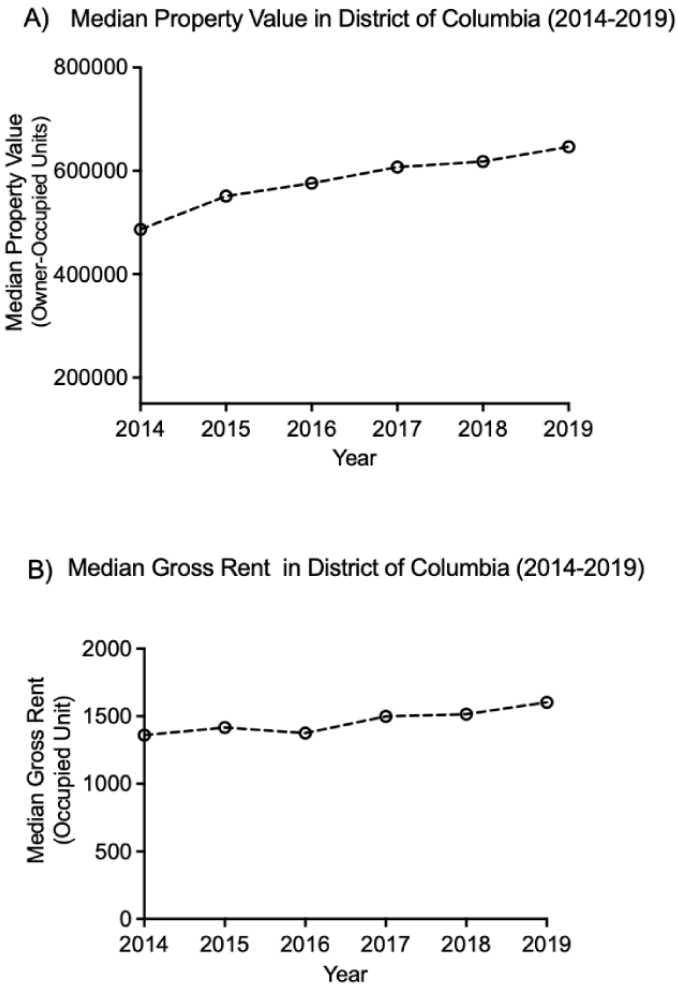

Dramatic shifts occurred in socioeconomic characteristics, with the percentage of college-educated residents increasing from 15.83% to 40.96% and the median household income increasing by US $17,490 over these 6 years (Table 1). The proportion of Washington, DC, residents who are African American declined from 55.06% in 2014 to 50.15% in 2019, while the Hispanic population increased from 8.76% to 10.28% during the same period (Table 1). The median age, owner-occupied housing rates, and the percentage of male residents remained relatively stable (Table 1).

Changes in Health Outcomes in Washington, DC

High cholesterol rates steadily declined from 32.02% in 2014 to 27.73% in 2019 (Table 1). Cancer prevalence was highest in 2017 at 5.71% and then declined to 5.1% in 2019. High blood pressure was relatively stable, with a prevalence of 31.53% in 2014 and 32.14% in 2019. Obesity rates showed a consistent upward trend until 2018, before dropping by ~3% in 2019. A similar trend was seen for diabetes, which saw a high of 10.7% in 2018, followed by a decline in 2019 (9.2%). Poor mental health days increased steadily throughout the study period from 12.17% in 2014 to 15.38% in 2019.

Longitudinal Analyses

Census tracts with the highest proportion of non-single-family homes (ie, mixed land use) experienced a 0.39% decrease in obesity and a 0.2% decrease in diabetes rates compared to those with the lowest proportion of non-single-family homes (Table 2). Increased road construction activity was associated with a 0.91% decrease in obesity rates and a 0.41% decrease in diabetes rates, comparing the second tertile to the first. Road construction activity in the highest tertile was associated with a 0.17% decrease in high cholesterol prevalence compared to the lowest tertile. Contrary to initial hypotheses, sidewalks had mixed associations (Tables 2 and 3). An increase in stop signs was linked with a reduction in high cholesterol (second tertile only) and marginally with a decrease in obesity but was associated with an increase in poor mental health days (Tables 2 and 3). Higher proportions of African American residents were associated with an increased prevalence of chronic conditions and poor mental health days (Table 3), with an SD increase in the African American population associated with a 6.81% rise in obesity (Table 2). A higher proportion of owner-occupied housing was associated with reduced rates of chronic diseases and poor mental health days (Table 3). The median household income was associated with reduced prevalence of obesity, diabetes, and high cholesterol (Tables 2 and 3).

XSL•FO
RenderX

Table 2. Built environment predictors of adult health outcomes (obesity, diabetes, high blood pressure)[a].

Predictors	Obesity	Diabetes	High blood pressure
	Change % (95% CI)[b]	Change % (95% CI)[b]	Change % (95% CI)[b]
Year (centered 2016)	0.16 (0.08 to 0.24)[c]	–0.05 (–0.10 to 0.00)[c]	0.11 (0.05 to 0.17)[c]
Single-lane roads			
Third tertile (highest)	0.26 (–0.05 to 0.57)	0.09 (–0.08 to 0.27)	0.05 (–0.08 to 0.18)
Second tertile	0.11 (–0.15 to 0.37)	0.07 (–0.08 to 0.21)	0.03 (–0.07 to 0.14)
Stop signs			
Third tertile (highest)	–0.20 (–0.44 to 0.04)[d]	–0.03 (–0.17 to 0.10)	0.07 (–0.03 to 0.17)
Second tertile	–0.31 (–0.66 to 0.04)[d]	–0.02 (–0.22 to 0.18)	0.07 (–0.07 to 0.22)
Two or more cars			
Third tertile (highest)	–0.22 (–0.57 to 0.14)	–0.15 (–0.35 to 0.05)[d]	–0.18 (–0.33 to –0.04)[c]
Second tertile	–0.19 (–0.48 to 0.10)	–0.15 (–0.32 to 0.01)	–0.07 (–0.19 to 0.05)
Streetlights			
Third tertile (highest)	0.09 (–0.21 to 0.38)	0.06 (–0.10 to 0.23)	0.00 (–0.12 to 0.13)
Second tertile	–0.04 (–0.31 to 0.22)	0.01 (–0.13 to 0.16)	0.09 (–0.02 to 0.19)
Non-single-family homes			
Third tertile (highest)	–0.39 (–0.74 to –0.03)[c]	–0.20 (–0.39 to 0.00)	–0.11 (–0.26 to 0.03)
Second tertile	–0.15 (–0.44 to 0.14)	–0.13 (–0.30 to 0.03)	–0.04 (–0.16 to 0.08)
Sidewalks			
Third tertile (highest)	0.20 (–0.16 to 0.56)	0.12 (–0.08 to 0.32)	0.06 (–0.08 to 0.21)
Second tertile	0.34 (0.06 to 0.61)[c]	0.12 (–0.03 to 0.28)	0.02 (–0.09 to 0.14)
Road construction			
Third tertile (highest)	–0.01 (–0.24 to 0.23)	0.00 (–0.13 to 0.13)	0.06 (–0.03 to 0.16)
Second tertile	–0.91 (–1.34 to –0.48)[c]	–0.41 (–0.66 to –0.16)[c]	0.19 (–0.01 to 0.38)[d]
Census tract sociodemographics			
Percentage male	0.00 (–0.20 to 0.21)	–0.12 (–0.24 to 0.01)	–0.10 (–0.23 to 0.02)
Median age	–0.19 (–0.46 to 0.08)	0.35 (0.18 to 0.52)[c]	0.09 (–0.08 to 0.26)
Percentage African American	6.81 (6.34 to 7.27)[c]	3.74 (3.44 to 4.04)[c]	3.85 (3.27 to 4.43)[c]
Percentage Hispanic	0.68 (0.42 to 0.95)[c]	0.41 (0.24 to 0.57)[c]	0.61 (0.42 to 0.79)[c]
Percentage college educated	0.51 (0.31 to 0.71)[c]	0.23 (0.12 to 0.34)[c]	0.31 (0.21 to 0.40)[c]
Median household income	–0.75 (–1.17 to –0.32)[c]	–0.49 (–0.75 to –0.24)[c]	–0.72 (–0.97 to –0.47)[c]
Percentage owner-occupied housing	–0.80 (–1.17 to –0.42)[c]	–0.11 (–0.35 to 0.13)	–0.31 (–0.64 to 0.02)[d]

[a]Data source for health outcomes: Centers for Disease Control and Prevention (CDC) Population Level Analysis and Community Estimates (PLACES).

[b]Multilevel mixed-effects linear models were run for each outcome separately with random intercepts for census tract levels (153 tracts) and random slopes for years (centered 2016). Models controlled for census tract median age, percentage male, percentage Hispanic, percentage African American, percentage with a college degree, median household income, and percentage owner-occupied housing. Built environment characteristics were categorized into tertiles, with the lowest tertile serving as the reference group. Census tract sociodemographic characteristics were standardized with a mean of 0 (SD 1).

[c]$P<.05$.

[d]$P<.10$.

XSL•FO
RenderX

Table 3. Built environment predictors of adult health outcomes (high cholesterol, cancer, mental health days)[a].

Predictors	High cholesterol	Cancer	Mental health days
	Change % (95% CI)[b]	Change % (95% CI)[b]	Change % (95% CI)[b]
Year (centered 2016)	–1.16 (–1.23 to –1.09)[c]	0.07 (0.05 to 0.09)[c]	0.72 (0.65 to 0.79)[c]
Single-lane roads			
Third tertile (highest)	0.11 (–0.11 to 0.33)	–0.03 (–0.10 to 0.05)	0.08 (–0.15 to 0.32)
Second tertile	0.03 (–0.15 to 0.21)	–0.02 (–0.08 to 0.05)	0.13 (–0.06 to 0.33)
Stop signs			
Third tertile (highest)	–0.11 (–0.28 to 0.05)	0.04 (–0.02 to 0.10)	0.30 (0.12 to 0.48)[c]
Second tertile	–0.25 (–0.51 to 0.00)[c]	–0.03 (–0.12 to 0.05)	0.25 (–0.02 to 0.51)[d]
Two or more cars			
Third tertile (highest)	–0.04 (–0.29 to 0.21)	0.07 (–0.02 to 0.15)	0.13 (–0.14 to 0.40)
Second tertile	–0.06 (–0.27 to 0.14)	0.03 (–0.04 to 0.10)	0.05 (–0.17 to 0.27)
Streetlights			
Third tertile (highest)	0.04 (–0.17 to 0.24)	–0.03 (–0.10 to 0.04)	0.05 (–0.18 to 0.27)
Second tertile	0.14 (0.04 to 0.33)	–0.02 (–0.08 to 0.04)	0.18 (–0.02 to 0.38)[d]
Non-single-family homes			
Third tertile (highest)	0.13 (–0.12 to 0.38)	–0.01 (–0.09 to 0.08)	0.18 (–0.09 to 0.45)
Second tertile	0.03 (–0.18 to 0.23)	–0.01 (–0.08 to 0.06)	0.14 (–0.08 to 0.35)
Sidewalks			
Third tertile (highest)	0.26 (0.01 to 0.52)[c]	–0.08 (–0.16 to 0.01)[d]	–0.21 (–0.48 to 0.06)
Second tertile	–0.07 (–0.26 to 0.12)	–0.04 (–0.11 to 0.03)	–0.21 (–0.41 to 0.00)[c]
Road construction			
Third tertile (highest)	–0.17 (–0.33 to –0.01)[c]	–0.04 (–0.09 to 0.02)	0.01 (–0.17 to 0.18)
Second tertile	0.21 (–0.10 to 0.53)	–0.10 (–0.20 to 0.00)[c]	0.17 (–0.17 to 0.51)
Census tract sociodemographics			
Percentage male	–0.21 (–0.39 to –0.03)[c]	–0.01 (–0.06 to 0.04)	0.10 (–0.06 to 0.27)
Median age	0.42 (0.18 to 0.67)[c]	0.08 (0.01 to 0.15)[c]	–0.57 (–0.79 to –0.36)[c]
Percentage African American	3.01 (2.46 to 3.57)[c]	0.07 (–0.11 to 0.25)	2.45 (2.08 to 2.82)[c]
Percentage Hispanic	0.67 (0.42 to 0.93)[c]	–0.02 (–0.10 to 0.05)	0.16 (–0.05 to 0.36)
Percentage college educated	0.61 (0.46 to 0.76)[c]	–0.05 (–0.10 to 0.00)[c]	0.05 (–0.11 to 0.20)
Median household income	0.23 (–0.13 to 0.59)	0.01 (–0.10 to 0.11)	–0.14 (–0.49 to 0.20)
Percentage owner-occupied housing	0.07 (–0.34 to 0.48)	0.29 (0.16 to 0.42)[c]	–1.13 (–1.42 to –0.83)[c]

[a]Data source for health outcomes: Centers for Disease Control and Prevention Population Level Analysis and Community Estimates (CDC PLACES).

[b]Multilevel mixed-effects linear models were run for each outcome separately with random intercepts for census tract levels (153 tracts) and random slopes for years (centered 2016). Models controlled for census tract median age, percentage male, percentage Hispanic, percentage African American, percentage with a college degree, median household income, and percentage owner-occupied housing. Built environment characteristics were categorized into tertiles, with the lowest tertile serving as the reference group. Census tract sociodemographic characteristics were standardized with a mean of 0 (SD 1).

[c]$P<.05$.

[d]$P<.10$.

In sensitivity analyses, we implemented regression analyses using continuous (linear forms) variables of the built environment characteristics rather than tertiles (Table S2 in Multimedia Appendix 1). Results from the linear models were

XSL•FO
RenderX

generally consistent with the tertile-based analyses. However, some associations did not reach statistical significance in regression models with linear forms of the built environment characteristics. For instance, although the second tertile of sidewalks was positively associated with obesity and negatively linked with poor mental health days compared to the first tertile, these associations were not statistically significant when sidewalks were modeled as a continuous variable. Additionally, some associations that were statistically significant in the third tertile (compared to the first tertile) were not statistically significant in the linear form of the variable (ie, non-single-family homes and diabetes; construction and high cholesterol). This divergence in patterns detected could stem from tertiles allowing for nonlinear associations and the examination of dose-response associations.

Table 4 shows the results of analyses examining the relationship between changes in neighborhood sociodemographics and the built environment. A 1 SD increase in the proportion of African American residents was linked with a 1.41% decrease in road construction activity and a marginally significant reduction in non-single-family homes, a proxy for mixed land use. A 1 SD increase in the median age was linked to a 4.94% decrease in non-single-family homes and a 0.86% decrease in road construction activity, suggesting lower urban development in census tracts with older populations. In contrast, a 1 SD increase in owner-occupied housing was associated with a 4.1% reduction in non-single-family homes, suggesting that neighborhoods that had higher owner occupancy were less likely to be mixed land use (Table 4). A heatmap of correlations with 2019 data are presented in Figure S1 in Multimedia Appendix 1. Directionality of associations was similar. For instance, the median age was negatively correlated with non-single-family homes (r=–0.51; Figure S1 in Multimedia Appendix 1).

Table 4. Sociodemographic predictors of built environment changes[a].

Predictors	Percentage non-single-family homes (%), change (95% CI)	Percentage road construction (%), change (95% CI)	Percentage sidewalks (%), change (95% CI)
Year (centered 2016)	0.33 (–0.56 to 1.22)	–0.56 (–0.95 to –0.17)[b]	–1.36 (–2.39 to –0.32)[b]
Census tract sociodemographics			
Percentage male (%)	3.39 (1.35 to 5.43)[b]	0.40 (–0.31 to 1.11)	–0.72 (–2.54 to 1.11)
Median age (%)	–4.94 (–7.44 to –2.43)[b]	–0.86 (–1.66 to –0.06)[b]	–1.14 (–3.17 to 0.89)
Percentage African American (%)	–3.95 (–8.05 to 0.15)[c]	–1.41 (–2.81 to –0.02)[b]	–2.39 (–5.96 to 1.19)
Percentage Hispanic (%)	–0.70 (–3.03 to 1.63)	–0.05 (–0.78 to 0.68)	2.06 (0.20 to 3.92)[b]
Percentage college educated (%)	2.18 (–0.16 to 4.52)[c]	0.83 (–0.17 to 1.83)	–0.44 (–3.05 to 2.17)
Median household income (%)	–3.19 (–7.51 to 1.13)	–1.37 (–2.87 to 0.12)[c]	–2.98 (–6.79 to 0.84)
Percentage owner-occupied housing (%)	–4.10 (–7.38 to –0.83)[b]	0.34 (–0.70 to 1.38)	2.59 (–0.07 to 5.24)[c]

[a]Multilevel mixed-effects linear models were run for each built environment characteristic separately with random intercepts for census tract levels (153 tracts) and random slopes for years (centered 2016). Models controlled for the census tract median age, percentage male, percentage Hispanic, percentage African American, percentage college educated, and percentage owner-occupied housing.

[b]$P<.05$.

[c]$P<.10$.

Discussion

Changes in the Neighborhood Environment

During the period from 2014 to 2019, Washington, DC, experienced significant urban transformation, marked by substantial changes in neighborhood built environments and demographic shifts. The urban landscape shifted toward higher-density housing, with non-single-family homes rising from 66.01% to 72.48% of the housing stock. The observed growth in non-single-family housing aligns with large-scale mixed-use developments, such as CityCenterDC, the Wharf, and the Anacostia Riverwalk Trail. In addition, the prevalence of single-lane roads increased from 37.39% to 42.16%, indicating a shift toward more sustainable and compact urban forms, which promote slower speeds, reduce car dependency, and increase walkable neighborhoods. Furthermore, other indicators of walkability, such as sidewalks and streetlights, increased from 2014 to 2016. This particularly aligns with early efforts to help improve walkability through multiple initiatives, such as Vision Zero DC (2014) and the Pedestrian Master Plan. However, later declines in walkability indicators contrast with the goals that were laid out according to these plans (Vision Zero DC, Pedestrian Master Plan).

Moreover, during this time period, Washington, DC, underwent rapid gentrification. The median household income increased by US $17,000, reflecting the influx of wealthier residents, and the percentage of the population with a college education increased from 15.82% to 40.96%. Meanwhile, the percentage of African American residents decreased from 55.06% to 50.15%. The decline in the proportion of the African American population, alongside rising income and education levels, is an indicator of gentrification, but it also raises some concerns about racial displacement and equity in the context of urban development [45]. Hwang et al [46] found that gentrification is

XSL•FO
RenderX

not strongly associated with increased displacement of lower-income residents; however, the authors found different location outcomes by race, suggesting that racial stratification in the housing market has a dominant effect regardless of gentrification [46].

Built Environment and Health

The longitudinal analysis performed in this study showed associations with changes in the built environment and various health outcomes in Washington, DC, that provide insights into the possible mechanisms through which urban development can influence population health. One of the important findings of this study is that an increase in non-single-family homes is linked with lower prevalence of obesity and diabetes. This is consistent with prior research analyzing 1.4 million GSV images integrated with electronic health records in Utah, which showed that residents in neighborhoods with more non-single-family homes had a 10%-27% lower prevalence of diabetes, uncontrolled diabetes, hypertension, and obesity [27]. In addition to physical health benefits, prior research has also demonstrated associations between greater land use mix and mental health; a 2021 systematic review [47] found that greater land use mix is associated with decreased psychological distress. Non-single-family homes, typically located in high-population-density areas and mixed-use urban areas, are close to shops and workplaces and promote the use of public transportation; this encourages walking and cycling as part of daily routines. As a result, active transportation in these neighborhoods leads to greater walkability and physical activity than in car-dependent suburbs dominated by single-family homes. In addition, the residents of high-density areas have access to parks, community centers, and recreational facilities that promote outdoor physical activity [48].

Our results align with prior research suggesting that built environment improvements can yield measurable health benefits. For instance, Nguyen et al [42] used GSV imagery to assess neighborhood characteristics and reported that areas classified as more rural or underdeveloped exhibit higher rates of obesity, diabetes, poor self-rated health, and premature mortality, underscoring the link between urban development and population health. Building on this, the health benefits observed in our study may be indicative of well-implemented urbanization programs that enhance access to amenities and services through new construction projects. We found that an increase in road construction is associated with decreased prevalence of obesity, diabetes, high cholesterol, and cancer. These developments can facilitate healthy behavioral changes by preserving or increasing opportunities for physical activity, particularly in areas affected by road closures [49]. For instance, a prior study [50] has shown that when roads are closed to vehicular traffic, residents often utilize the space for recreation.

Furthermore, we identified that the presence of stop signs is associated with a decrease in chronic conditions. Although these associations do not imply causality, they may reflect well-developed urban environments that support healthier lifestyles. This interpretation is partially supported by previous cross-sectional studies that have linked dense areas and traffic-calming measures, such as traffic lights and stop signs, with reduced vehicle speeds and traffic volumes, enhanced safety, and improved quality of life [51,52]. Previous research also suggests that neighborhoods with safer traffic environments and a variety of destinations are more likely to promote physical activity, which can contribute to better health outcomes [53].

In this study, sidewalks were positively associated with obesity but negatively linked with poor mental health days, but these findings were limited to the second tertile versus the first tertile and were not significant when modeled as a continuous variable. Previous studies also indicate a highly complex relationship between walkability indicators and various chronic conditions. For example, Keralis et al [6] analyzed 31 million GSV images and highlighted an increased prevalence of obesity, diabetes, and inactivity associated with the second tertile of crosswalk scores in contrast with a decreased prevalence associated with the third tertile. However, when modeling crosswalks as a continuous linear variable, there was a significant negative correlation between crosswalks and chronic conditions. These mixed findings illustrate the complexity of correlating walkability indicators with health outcomes. On the one hand, sidewalks are associated with urbanicity, where there is a higher prevalence of chronic conditions [54]; on the other hand, they are associated with increased physical activity, which can reduce the risk of negative health outcomes [55]. These mixed findings underscore that sidewalks are proxies for multiple contextual factors, including urban density and opportunities for physical activity. Therefore, the varying associations of sidewalks and health outcomes warrant further investigation and increased efforts to adjust for potential confounding.

Finally, our analyses revealed that neighborhoods with higher proportions of African American residents are associated with reduced road construction and marginally fewer non-single-family homes. This is consistent with gentrification patterns reported for Washington, DC. This is only partially addressed by equity frameworks in transportation plans, such as MoveDC (Washington Post Gentrification Report, MoveDC Equity Framework). Nicoletti et al [56] also reported that urban communities with larger shares of racial minorities, lower income, and less educational attainment face persistent disparities in spatial access to infrastructure. These findings underscore the need to address structural inequities in infrastructure investment and urban planning.

Structural factors have profoundly shaped the housing market in Washington, DC, producing persistent racial disparities in homeownership, property values, and housing access. Beginning in the 1930s, the Homeowners' Loan Corporation (HOLC) established "Residential Security Maps." Neighborhoods were assigned a letter grade from A (safest investment) to D (hazardous), with different colors indicating their level of risk for lenders. African American neighborhoods were consistently outlined in red ("hazardous"). This practice of redlining denied residents access to home loans, business capital, and construction funds, thereby entrenching long-term disinvestment [57-60]. In the mid-20th century, "urban renewal" programs further displaced thousands of African American residents [61]. Entire neighborhoods were demolished to make way for commercial development and highways to the suburbs. The city offered little or no compensation or relocation assistance.

XSL•FO
RenderX

Following these demolitions, the city concentrated public housing projects in the eastern part of the district, particularly east of the Anacostia River, deepening patterns of racial and economic segregation [62]. These historical policies and practices shaped modern-day neighborhoods in Washington, DC. Today, Washington, DC, remains one of the most racially segregated cities in the United States, with distinct east-west divisions [63]. These housing and planning practices not only entrenched segregation but also restricted access to health-promoting resources, such as quality housing, neighborhood amenities, education, and health care [64,65].

Strengths of the Study

This study is one of the few to use a longitudinal study design to examine the potential impact of changing built environments, neighborhood sociodemographics, and health outcomes. In addition, we uniquely leveraged innovative data sources and computer vision methods to enable the analysis of millions of street view scenes in Washington, DC. The use of publicly available GSV images offers a useful case study for conducting cost-effective, large neighborhood studies with data not typically included in public health studies. We collected and analyzed four million GSV images to assess changes in neighborhood built environments over 6 years. Neighborhood studies have traditionally relied on costly and time-intensive neighborhood audits or neighborhood surveys [66,67]. Alternatively, other research teams have also set up specially designed camera equipment on vehicles to enable photo data collection during drives through neighborhoods [68]. However, the cost of renting or purchasing that equipment can be prohibitively expensive, in addition to the expertise and staff time needed for operating the equipment. Moreover, primary data collection is often feasible for studies involving certain neighborhoods within a city and is less feasible for comprehensively capturing images across multiple cities or states or conducting national neighborhood studies.

Limitations of the Study

Nonetheless, our study is subject to some notable limitations. Focusing on a single city may limit statistical power and reduce external validity, as findings may not fully extend to other urban contexts with different social, economic, or planning structures. Future multicity analyses will be important for testing the robustness of these detected patterns and enhancing their policy relevance. Washington, DC, was chosen as a case example, given its well-known gentrification and urban development happening in recent years in the city and to investigate links between built environment changes, sociodemographic changes, and health outcomes using a longitudinal design.

In this study, health outcomes were analyzed separately. In future studies, if the relationships between multiple outcome variables are of interest, multivariate multilevel mixed-effects models can be applied by including additional random effects for multiple correlated outcomes. Additionally, although this study used built environment indicators theoretically and empirically supported by the literature as being important to health (eg, indicators of walkability, such as sidewalks), future studies may incorporate additional features of the neighborhood environment, such as environmental hazards, recreational resources, and indicators of dilapidation.

Moreover, our image data collection was limited by the coverage of GSV for Washington, DC, for our study period. Although the same sampling coordinates were used to collect GSV images for each year, the availability of yearly image updates differed across areas in Washington, DC. GSV coverage data ranged from 143 to 169 of the 178 census tracts in Washington, DC, from 2014 to 2019. Health outcome estimates from CDC PLACES were available for approximately 154 tracts in 2014-2016 and up to 174 tracts in subsequent years.

Partial GSV coverage likely reflects uneven update patterns across neighborhoods, wherein areas with greater commercial activity or higher traffic volumes are prioritized for imagery updates, whereas predominantly residential or lower-income areas may experience reduced update frequency or lower-resolution coverage. This limitation has been noted in prior studies using GSV imagery for neighborhood health research [69]. Similarly, the incomplete availability of health estimates in earlier years likely reflects methodological constraints, including the evolving nature of small-area estimation models and the exclusion of tracts with insufficient population sizes or high uncertainty. As the modeling approach matured and incorporated more robust data sources, coverage improved in later years [70]. These gaps in coverage could have contributed to bias in the analyses. Variation in image update availability could lead to noncorrespondence between changes in the prevalence of a feature in imagery and the true change of that feature in the real world. Nonetheless, the overall data coverage was still high, ranging from 88%-92% of census tracts in Washington, DC.

In addition, CDC PLACES uses a multilevel regression and poststratification (MRP) method to generate estimates for places with smaller populations, such as census tracts and ZIP Code Tabulation Areas. These model-based datasets provide a more granular and comprehensive view of community health than traditional surveillance data, enabling detailed analysis of chronic diseases, preventive services use, health risk behaviors, disabilities, and health-related social needs at the local level [44,71]. However, reliance on modeled health estimates (rather than direct health data) could have introduced error or bias in the analysis.

To derive sociodemographic information, we used the 1-year estimates from the ACS data instead of the 5-year estimates. Although 1-year estimates offer more precise year-by-year demographic snapshots, they come with trade-offs in accuracy due to smaller sample sizes. The 1-year estimate provides estimates for geographic areas with populations of 65,000 or more, and the smaller sample size could result in less precision. Future studies could consider using 5-year estimates as a supplement for greater precision. Despite these limitations, the analytic sample ultimately encompassed approximately 88% of census tracts in Washington, DC, representing substantial spatial coverage and supporting the generalizability of study findings across DC.

XSL•FO
RenderX

Conclusion

This study used GSV images and computer vision to examine changes in neighborhood built environments, demographics, and health outcomes in Washington, DC, from 2014 to 2019. By analyzing over 430,000 images, we assessed urban development and walkability indicators at the census tract level. We observed notable shifts in the demographic composition of residents and concurrent changes in neighborhood built environments and health outcomes. These patterns underscore the critical connections between urban policy and public health. Key findings showed that increased construction activity and non-single-family housing are associated with reductions in the prevalence of obesity and diabetes, highlighting the potential health benefits of urban development. In contrast, walkability indicators revealed mixed associations: sidewalks were unexpectedly linked to higher obesity and cholesterol rates but to lower rates of poor mental health. Neighborhoods with larger African American populations saw less infrastructure investment, highlighting ongoing racial disparities in urban development. By integrating longitudinal imagery with public health data, this study demonstrates the feasibility of using scalable, low-cost digital tools to monitor neighborhood change and its health impacts. These findings contribute to urban health research and underscore the need for more equitable infrastructure planning. As issues such as gentrification, zoning, residential location in proximity to environmental hazards, and walkability remain at the forefront of urban discourse, it is imperative that future studies investigate the connections between these indicators and health outcomes.

Acknowledgments

The research reported in this publication was supported by the National Library of Medicine (award number R01LM012849 to QCN) and the National Institute on Minority Health and Health Disparities (award numbers R01MD015716 to TTN and R01MD016037 to QCN). Additionally, this research was supported (in part) by the Intramural Research Program of the National Institutes of Health (NIH; grant number ZIA NR000043-01 to QCN). The contributions of the NIH author(s) are considered Works of the United States Government. The findings and conclusions presented in this paper are those of the authors and do not necessarily reflect the views of the NIH or the US Department of Health and Human Services. The funders played no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data Availability

Population Level Analysis and Community Estimates 2021 and the American Community Survey are publicly available. Google Street View neighborhood-level data can be accessed online [72].

Authors' Contributions

Conceptualization, funding acquisition, and project administration were managed by QCN; methodology, QCN, TT, MA, and DL; software and visualization, DL; validation, QCN, XY, and HM; formal analysis, QCN and MA.; data curation, XY, HM, and RD; writing—original draft preparation, QCN, TT, and RD; and writing—review and editing, XY, HM, JM, MA, PD, DL, TTN, PPM, AA, YH, and XH. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Time trends in walkability indicators in census tracts across Washington DC, 2014-2019; correlations between built environment characteristics and census tract characteristics, 2019; built environment predictors of adult health outcomes.
[DOCX File , 83 KB - formative_v9i1e74195_app1.docx]

References

1. Phan L, Yu W, Keralis JM, Mukhija K, Dwivedi P, Brunisholz KD, et al. Google Street View derived built environment indicators and associations with state-level obesity, physical activity, and chronic disease mortality in the United States. Int J Environ Res Public Health 2020 May 22;17(10):3659 [FREE Full text] [doi: 10.3390/ijerph17103659] [Medline: 32456114]
2. Ulijaszek S. Physical activity and the human body in the (increasingly smart) built environment. Obes Rev 2018 Dec 03;19 Suppl 1(S1):84-93 [FREE Full text] [doi: 10.1111/obr.12790] [Medline: 30511510]
3. Cerin E, Cain KL, Conway TL, Van Dyck D, Hinckson E, Schipperijn J, et al. Neighborhood environments and objectively measured physical activity in 11 countries. Med Sci Sports Exerc 2014 Dec;46(12):2253-2264 [FREE Full text] [doi: 10.1249/MSS.0000000000000367] [Medline: 24781892]
4. Kerr J, Emond JA, Badland H, Reis R, Sarmiento O, Carlson J, et al. Perceived Neighborhood Environmental Attributes Associated with Walking and Cycling for Transport among Adult Residents of 17 Cities in 12 Countries: The IPEN Study. Environ Health Perspect 2016 Mar;124(3):290-298 [FREE Full text] [doi: 10.1289/ehp.1409466] [Medline: 26186801]

XSL•FO
RenderX

5. Sallis JF, Cain KL, Conway TL, Gavand KA, Millstein RA, Geremia CM, et al. Is Your Neighborhood Designed to Support Physical Activity? A Brief Streetscape Audit Tool. Prev Chronic Dis 2015 Sep 03;12:E141 [FREE Full text] [doi: 10.5888/pcd12.150098] [Medline: 26334713]
6. Keralis JM, Javanmardi M, Khanna S, Dwivedi P, Huang D, Tasdizen T, et al. Health and the built environment in United States cities: measuring associations using Google Street View-derived indicators of the built environment. BMC Public Health 2020 Feb 12;20(1):215 [FREE Full text] [doi: 10.1186/s12889-020-8300-1] [Medline: 32050938]
7. Mackenbach JD, Rutter H, Compernolle S, Glonti K, Oppert J, Charreire H, et al. Obesogenic environments: a systematic review of the association between the physical environment and adult weight status, the SPOTLIGHT project. BMC Public Health 2014 Mar 06;14:233 [FREE Full text] [doi: 10.1186/1471-2458-14-233] [Medline: 24602291]
8. Wang Z, Shi Q, Yan X, Tang J, Zhang J. There was a similar U-shaped nonlinear association between waist-to-height ratio and the risk of new-onset hypertension: findings from the CHNS. Front Nutr 2023;10:1304521 [FREE Full text] [doi: 10.3389/fnut.2023.1304521] [Medline: 38156282]
9. Frank LD, Adhikari B, White KR, Dummer T, Sandhu J, Demlow E, et al. Chronic disease and where you live: built and natural environment relationships with physical activity, obesity, and diabetes. Environ Int 2022 Jan;158:106959 [FREE Full text] [doi: 10.1016/j.envint.2021.106959] [Medline: 34768046]
10. Creatore MI, Glazier RH, Moineddin R, Fazli GS, Johns A, Gozdyra P, et al. Association of neighborhood walkability with change in overweight, obesity, and diabetes. JAMA 2016 May 24;315(20):2211-2220 https://pubmed.ncbi.nlm.nih.gov/27218630/. [doi: 10.1001/jama.2016.5898] [Medline: 27218630]
11. Nichani V, Turley L, Vena JE, McCormack GR. Associations between the neighbourhood characteristics and body mass index, waist circumference, and waist-to-hip ratio: findings from Alberta's Tomorrow Project. Health Place 2020 Jul;64:102357 [FREE Full text] [doi: 10.1016/j.healthplace.2020.102357] [Medline: 32479266]
12. Carson JR, Conway TL, Perez LG, Frank LD, Saelens BE, Cain KL, et al. Neighborhood walkability, neighborhood social health, and self-selection among U.S. adults. Health Place 2023 Jul;82:103036 [FREE Full text] [doi: 10.1016/j.healthplace.2023.103036] [Medline: 37244155]
13. Carlson JA, Sallis JF, Kerr J, Conway TL, Cain K, Frank LD, et al. Built environment characteristics and parent active transportation are associated with active travel to school in youth age 12-15. Br J Sports Med 2014 Dec 21;48(22):1634-1639 [FREE Full text] [doi: 10.1136/bjsports-2013-093101] [Medline: 24659503]
14. Kyttä AM, Broberg AK, Kahila MH. Urban environment and children's active lifestyle: SoftGIS revealing children's behavioral patterns and meaningful places. Am J Health Promot 2012 May 01;26(5):e137-e148 [FREE Full text] [doi: 10.4278/ajhp.100914-quan-310] [Medline: 22548432]
15. Nordbø ECA, Raanaas RK, Nordh H, Aamodt G. Neighborhood green spaces, facilities and population density as predictors of activity participation among 8-year-olds: a cross-sectional GIS study based on the Norwegian Mother and Child Cohort Study. BMC Public Health 2019 Oct 30;19(1):1426 [FREE Full text] [doi: 10.1186/s12889-019-7795-9] [Medline: 31666049]
16. McMahon PJ, Zegeer CV, Duncan C, Knoblauch RL, Stewart JR, Khattak AJ. An analysis of factors contributing to "walking along roadway" crashes research study and guidelines for sidewalks and walkways. Federal Highway Administration, US Department of Transportation. 2002 Feb 1. URL: https://rosap.ntl.bts.gov/view/dot/50265 [accessed 2025-11-21]
17. Knoblauch RL, Tustin BH, Smith SA, Steven A, Pietrucha MT. Investigation of exposure based pedestrian accident areas: crosswalks, sidewalks, local streets and major arterials. Federal Highway Administration, US Department of Transportation. 1988 Sep 1. URL: https://rosap.ntl.bts.gov/view/dot/68336 [accessed 2025-11-21]
18. Roberts I. Adult accompaniment and the risk of pedestrian injury on the school-home journey. Inj Prev 1995 Dec 01;1(4):242-244 [FREE Full text] [doi: 10.1136/ip.1.4.242] [Medline: 9346039]
19. Karim DM. Narrower lanes, safer streets. 2015 Presented at: Canadian Institute of Transportation Engineers (CITE) 2015 Annual Conference; November 11-13, 2015; Regina, Saskatchewan URL: https://www.researchgate.net/publication/277590178_Narrower_Lanes_Safer_Streets
20. Rowangould G, Ferenchak N. ABQ Streets Project: creating alternative residential street designs. National Academies of Sciences. 2020. URL: https://trid.trb.org/view/1769925 [accessed 2025-11-21]
21. Kim B, Troxel WM, Dubowitz T, Hunter GP, Ghosh-Dastidar B, Chaix B, et al. Neighborhood built environment and sleep health: a longitudinal study in low-income and predominantly African-American neighborhoods. Am J Epidemiol 2023 May 05;192(5):736-747 [FREE Full text] [doi: 10.1093/aje/kwad016] [Medline: 36691683]
22. Chandrabose M, Rachele JN, Gunn L, Kavanagh A, Owen N, Turrell G, et al. Built environment and cardio-metabolic health: systematic review and meta-analysis of longitudinal studies. Obes Rev 2019 Jan 25;20(1):41-54 [FREE Full text] [doi: 10.1111/obr.12759] [Medline: 30253075]
23. Grazuleviciene R, Andrusaityte S, Dėdelė A, Grazulevicius T, Valius L, Rapalavicius A, et al. Urban environment and health: a cross-sectional study of the influence of environmental quality and physical activity on blood pressure. Int J Environ Res Public Health 2021 Jun 06;18(11):6126 [FREE Full text] [doi: 10.3390/ijerph18116126] [Medline: 34204097]
24. Becker DA, Browning MH, Kuo M, Van Den Eeden SK. Is green land cover associated with less health care spending? Promising findings from county-level Medicare spending in the continental United States. Urban For Urban Green 2019 May;41:39-47 [FREE Full text] [doi: 10.1016/j.ufug.2019.02.012]

XSL•FO
RenderX

25. Gerken M, Leahu-Aluas O, Jolly J. Built environment indicators and health - interactive map tool. GitHub. 2024 Jun 6. URL: https://mattgerken.github.io/dc-built-environment-map-tool/#Methodology [accessed 2025-11-21]
26. Mooney S, Bader MDM, Lovasi GS, Neckerman KM, Teitler JO, Rundle AG. Validity of an ecometric neighborhood physical disorder measure constructed by virtual street audit. Am J Epidemiol 2014 Sep 15;180(6):626-635 [FREE Full text] [doi: 10.1093/aje/kwu180] [Medline: 25122584]
27. Nguyen QC, Belnap T, Dwivedi P, Deligani AHN, Kumar A, Li D, et al. Google Street View images as predictors of patient health outcomes, 2017–2019. Big Data Cogn Comput 2022 Mar;6(1):15 [FREE Full text] [doi: 10.3390/bdcc6010015] [Medline: 36046271]
28. Plascak JJ, Schootman M, Rundle AG, Xing C, Llanos AAM, Stroup AM, et al. Spatial predictive properties of built environment characteristics assessed by drop-and-spin virtual neighborhood auditing. Int J Health Geogr 2020 May 29;19(1):21 [FREE Full text] [doi: 10.1186/s12942-020-00213-5] [Medline: 32471502]
29. Richardson J, Mitchell B, Franco J. Shifting neighborhoods: gentrification and cultural displacement in American cities. National Community Reinvestment Coalition (NCRC). 2019. URL: https://www.researchgate.net/publication/340278957_Shifting_Neighborhoods_Gentrification_and_cultural_displacement_in_American_cities [accessed 2025-11-21]
30. Richardson J, Mitchell B, Edlebi J. Gentrification and disinvestment 2020. National Community Reinvestment Coalition (NCRC). 2020. URL: https://ncrc.org/gentrification20/ [accessed 2025-11-21]
31. Platform G. Street View Static API overview. Google Maps Platform. URL: https://developers.google.com/maps/documentation/streetview/overview [accessed 2025-11-21]
32. robolyst/streetview. GitHub. URL: https://github.com/robolyst/streetview [accessed 2025-11-24]
33. Arcaya MC, Tucker-Seeley RD, Kim R, Schnake-Mahl A, So M, Subramanian S. Research on neighborhood effects on health in the United States: a systematic review of study characteristics. Soc Sci Med 2016 Nov;168:16-29 [FREE Full text] [doi: 10.1016/j.socscimed.2016.08.047] [Medline: 27637089]
34. Paglione L, Donato MA, Cofone L, Sabato M, Appolloni L, D'Alessandro D. The healthy city reimagined: walkability, active mobility, and the challenges of measurement and evaluation. Urban Sci 2024 Sep 29;8(4):157 [FREE Full text] [doi: 10.3390/urbansci8040157]
35. Resnik DB. Urban sprawl, smart growth, and deliberative democracy. Am J Public Health 2010 Oct;100(10):1852-1856 [FREE Full text] [doi: 10.2105/ajph.2009.182501]
36. Ewing R, Pendall R, Chen D. Measuring sprawl and its impact. Smart Growth America. 2002. URL: https://www.smartgrowthamerica.us/documents/MeasuringSprawl.PDF [accessed 2025-11-21]
37. Galster G. On the nature of neighbourhood. Urban Stud 2001 Nov 01;38(12):2111-2124 [FREE Full text] [doi: 10.1080/00420980120087072]
38. Downs A. Still Stuck in Traffic: Coping with Peak-Hour Traffic Congestion. Washington, DC: Brookings Institution Press; 2000.
39. Frank LD, Schmid TL, Sallis JF, Chapman J, Saelens BE. Linking objectively measured physical activity with objectively measured urban form: findings from SMARTRAQ. Am J Prev Med 2005 Feb;28(2 Suppl 2):117-125 [FREE Full text] [doi: 10.1016/j.amepre.2004.11.001] [Medline: 15694519]
40. Nguyen QC, Sajjadi M, McCullough M, Pham M, Nguyen TT, Yu W, et al. Neighbourhood looking glass: 360º automated characterisation of the built environment for neighbourhood effects research. J Epidemiol Community Health 2018 Mar 15;72(3):260-266 [FREE Full text] [doi: 10.1136/jech-2017-209456] [Medline: 29335255]
41. Yue X, Antonietti A, Alirezaei M, Tasdizen T, Li D, Nguyen L, et al. Using convolutional neural networks to derive neighborhood built environments from Google Street View images and examine their associations with health outcomes. Int J Environ Res Public Health 2022 Sep 24;19(19):12095 [FREE Full text] [doi: 10.3390/ijerph191912095] [Medline: 36231394]
42. Nguyen QC, Khanna S, Dwivedi P, Huang D, Huang Y, Tasdizen T, et al. Using Google Street View to examine associations between built environment characteristics and U.S. health outcomes. Prev Med Rep 2019 Jun;14:100859 [FREE Full text] [doi: 10.1016/j.pmedr.2019.100859] [Medline: 31061781]
43. Nguyen QC, Keralis JM, Dwivedi P, Ng AE, Javanmardi M, Khanna S, et al. Leveraging 31 million Google Street View images to characterize built environments and examine county health outcomes. Public Health Rep 2021 Nov 19;136(2):201-211 [FREE Full text] [doi: 10.1177/0033354920968799] [Medline: 33211991]
44. PLACES: local data for better health. Centers for Disease Control and Prevention. 2022. URL: https://www.cdc.gov/places/index.html [accessed 2025-11-21]
45. Brink-Johnson A, Lubin J. Center for Urban and Racial Equity. 2020. URL: https://urbanandracialequity.org/wp-content/uploads/2020/08/Structural-Racism-in-Washington-DC-1.pdf [accessed 2025-11-21]
46. Hwang J, Zhang IH. The reign of racialized residential sorting: gentrification and residential mobility in the twenty-first century. City Community 2024 Oct 13;24(2):83-110 [FREE Full text] [doi: 10.1177/15356841241276390]
47. Park Y, McMorris BJ, Pruinelli L, Song Y, Kaas MJ, Wyman JF. Use of geographic information systems to explore associations between neighborhood attributes and mental health outcomes in adults: a systematic review. Int J Environ Res Public Health 2021 Aug 14;18(16):8597 [FREE Full text] [doi: 10.3390/ijerph18168597] [Medline: 34444345]

XSL•FO
RenderX

48. Physical activity and the built environment. Physiopedia. 2022. URL: https://www.physio-pedia.com/Physical_Activity_and_the_Built_Environment [accessed 2025-08-25]
49. Pedestrian safety and work zone standards covered and open walkways. District Department of Transportation. 2007. URL: https://ddot.dc.gov/sites/default/files/dc/sites/ddot/publication/attachments/pedestrian_safety_and_work_zone_standards_covered_and_open_walkways_july_2010.pdf? [accessed 2025-11-21]
50. Kingham S, Curl A, Banwell K. Streets for transport and health: the opportunity of a temporary road closure for neighbourhood connection, activity and wellbeing. J Transport Health 2020 Sep;18:100872 [FREE Full text] [doi: 10.1016/j.jth.2020.100872]
51. Ewing R, Dumbaugh E. The built environment and traffic safety: a review of empirical evidence. J Plann Lit 2009 May 21;23(4):347-367 [FREE Full text] [doi: 10.1177/0885412209335553]
52. Brown V, Moodie M, Carter R. Evidence for associations between traffic calming and safety and active transport or obesity: a scoping review. J Transport Health 2017 Dec;7:23-37 [FREE Full text] [doi: 10.1016/j.jth.2017.02.011]
53. Smith M, Mavoa S, Ikeda E, Hasanzadeh K, Zhao J, Rinne TE, et al. Associations between children's physical activity and neighborhood environments using GIS: a secondary analysis from a systematic scoping review. Int J Environ Res Public Health 2022 Jan 18;19(3):1033 [FREE Full text] [doi: 10.3390/ijerph19031033] [Medline: 35162057]
54. McCormack GR, Shiell A, Giles-Corti B, Begg S, Veerman J, Geelhoed E, et al. The association between sidewalk length and walking for different purposes in established neighborhoods. Int J Behav Nutr Phys Act 2012 Aug 01;9(1):92 [FREE Full text] [doi: 10.1186/1479-5868-9-92] [Medline: 22853008]
55. Wei J, Wu Y, Zheng J, Nie P, Jia P, Wang Y. Neighborhood sidewalk access and childhood obesity. Obes Rev 2021 Feb 07;22 Suppl 1(Suppl 1):e13057 [FREE Full text] [doi: 10.1111/obr.13057] [Medline: 32638457]
56. Nicoletti L, Sirenko M, Verma T. Disadvantaged communities have lower access to urban infrastructure. Environ Plann B: Urban Anal City Sci 2022 Oct 04;50(3):831-849 [FREE Full text] [doi: 10.1177/23998083221131044]
57. Aaronson D, Hartley D, Mazumder B. The effects of the 1930s HOLC "redlining" maps. Am Econ J: Econ Policy 2021 Nov 01;13(4):355-392 [FREE Full text] [doi: 10.1257/pol.20190414]
58. Mitchell B, Franco J. HOLC "redlining" maps: the persistent structure of segregation and economic inequality. National Community Reinvestment Coalition (NCRC). 2018. URL: https://ncrc.org/holc/ [accessed 2025-11-21]
59. An B, Orlando AW, Rodnyansky S. The physical legacy of racism: how redlining cemented the modern built environment. SSRN Journal 2019 Dec 30:1-39 [FREE Full text] [doi: 10.2139/ssrn.3500612]
60. Wilson B. Urban heat management and the legacy of redlining. J Am Plann Assoc 2020 May 22;86(4):443-457 [FREE Full text] [doi: 10.1080/01944363.2020.1759127]
61. Asch CM, Musgrove GD. Chocolate City: A History of Race and Democracy in the Nation's Capital. Chapel Hill, NC: University of North Carolina Press; 2017.
62. Gwam P, Cohen M. Combating the legacy of segregation in the nation's capital. Urban Institute. URL: https://www.urban.org/racial-equity-analytics-lab/structural-racism-explainer-collection/combating-legacy-segregation-nations-capital [accessed 2025-11-21]
63. Menendian S, Gailes A, Gambhir S. The roots of structural racism: twenty-first century racial residential segregation in the United States. Othering and Belonging Institute. 2021 Jun 21. URL: https://belonging.berkeley.edu/roots-structural-racism [accessed 2025-11-21]
64. Williams DR, Collins C. Racial residential segregation: a fundamental cause of racial disparities in health. Public Health Rep 2001 Sep;116(5):404-416 [FREE Full text] [doi: 10.1016/s0033-3549(04)50068-7]
65. Bailey ZD, Krieger N, Agénor M, Graves J, Linos N, Bassett MT. Structural racism and health inequities in the USA: evidence and interventions. Lancet 2017 Apr;389(10077):1453-1463 [FREE Full text] [doi: 10.1016/s0140-6736(17)30569-x] [Medline: 28402827]
66. Diez Roux AV. Investigating neighborhood and area effects on health. Am J Public Health 2001 Nov;91(11):1783-1789 [FREE Full text] [doi: 10.2105/ajph.91.11.1783] [Medline: 11684601]
67. Sampson RJ, Morenoff JD, Gannon-Rowley T. Assessing "neighborhood effects": social processes and new directions in research. Annu Rev Sociol 2002 Aug;28(1):443-478 [FREE Full text] [doi: 10.1146/annurev.soc.28.110601.141114]
68. Bruno R, Nurchis M. Robust and efficient data collection schemes for vehicular multimedia sensor networks. 2013 Presented at: IEEE 14th International Symposium on "A World of Wireless, Mobile and Multimedia Networks" (WoWMoM); 2013; Madrid, Spain URL: https://ieeexplore.ieee.org/document/6583399 [doi: 10.1109/wowmom.2013.6583399]
69. Rzotkiewicz A, Pearson AL, Dougherty BV, Shortridge A, Wilson N. Systematic review of the use of Google Street View in health research: major themes, strengths, weaknesses and possibilities for future research. Health Place 2018 Jul;52:240-246 [FREE Full text] [doi: 10.1016/j.healthplace.2018.07.001] [Medline: 30015181]
70. 500 cities project all measures 2014. Centers for Disease Control and Prevention. 2021. URL: https://www.arcgis.com/home/item.html?id=ea8b721cf9034814bce067ddefd21ecc [accessed 2025-11-21]
71. Zhang X, Holt JB, Yun S, Lu H, Greenlund KJ, Croft JB. Validation of multilevel regression and poststratification methodology for small area estimation of health indicators from the behavioral risk factor surveillance system. Am J Epidemiol 2015 Jul 15;182(2):127-137 [FREE Full text] [doi: 10.1093/aje/kwv002] [Medline: 25957312]

XSL•FO
RenderX

72. Google Street View maps: census tract level. ArcGIS. URL: https://experience.arcgis.com/experience/0013c4d9acb54327a198c44692511e68/page/Census-Tract-Level [accessed 2025-11-24]

Abbreviations

ACS: American Community Survey
API: application programming interface
CDC: Centers for Disease Control and Prevention
GSV: Google Street View
PLACES: Population Level Analysis and Community Estimates

Edited by J Sarvestan; submitted 19.03.25; peer-reviewed by F Pennoni, S Mooney, M Lyu; comments to author 18.06.25; revised version received 06.08.25; accepted 26.09.25; published 10.12.25.

Please cite as:
Nguyen QC, Doumbia R, Nguyen TT, Yue X, Mane H, Merchant J, Tasdizen T, Alirezaei M, Dipankar P, Li D, Mullaputi PSP, Alibilli A, Hswen Y, He X
Changes in the Neighborhood Built Environment and Chronic Health Conditions in Washington, DC, in 2014-2019: Longitudinal Analysis
JMIR Form Res 2025;9:e74195
URL: https://formative.jmir.org/2025/1/e74195
doi:10.2196/74195
PMID:

©Quynh C Nguyen, Riki Doumbia, Thu T Nguyen, Xiaohe Yue, Heran Mane, Junaid Merchant, Tolga Tasdizen, Mitra Alirezaei, Pankaj Dipankar, Dapeng Li, Penchala Sai Priya Mullaputi, Amrutha Alibilli, Yulin Hswen, Xin He. Originally published in JMIR Formative Research (https://formative.jmir.org), 10.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Mapping the Availability of Rehabilitation Providers Using Public Licensure and Population Data for a Geographic Information System–Based Approach to Workforce Planning: Cross-Sectional Feasibility Study

Madeline Ratoza[1], PT, DPT, PhD; Rupal M Patel[2], PT, PhD; Julia Chevan[3], PT, DPT, MPH, PhD; Wayne Brewer[2], PT, MPH, PhD; Katy Mitchell[4], PT, PhD

[1]University of St. Augustine for Health Sciences, Austin, TX, United States
[2]Texas Woman's University, Houston, TX, United States
[3]Springfield College, Springfield, MA, United States
[4]Texas Woman's University, Denton, TX, United States

Corresponding Author:
Madeline Ratoza, PT, DPT, PhD
University of St. Augustine for Health Sciences
5401 La Crosse Ave
Austin, TX
United States
Phone: 1 7372023279
Email: madeline.ratoza@gmail.com

Abstract

Background: Access to rehabilitation services is a critical yet under-studied dimension of health equity. Among the 6 domains of access, health care provider availability, defined as the presence of sufficient health care providers to meet population needs, is particularly underexplored in rehabilitation professions such as physical and occupational therapy. Current data reporting often lacks the geographic granularity required for effective workforce planning.

Objective: The purpose of this study was to demonstrate the feasibility of mapping rehabilitation provider availability at the census tract level using geographic information systems and integrating public licensure and population data to inform equitable workforce planning.

Methods: A descriptive, cross-sectional study was conducted using publicly available state licensure data for physical and occupational therapists and demographic data from the American Community Survey. Residential addresses of rehabilitation providers were geocoded and matched to 2020 census tracts. Population-to-provider ratios were calculated and mapped using choropleth and bivariate mapping techniques. Population-to-provider ratios were calculated per tract and summarized overall and by rurality using 2020 Rural-Urban Commuting Area (RUCA) codes (urban: RUCA of 1-3; rural: RUCA of ≥4). The spatial dependence of ratios was tested using a spatial autocorrelation statistic, the global Moran *I*, in ArcGIS Pro using edge contiguity neighbors and row standardization.

Results: Across 6896 tracts, ratios ranged from 4.5 to 11,147 persons per provider (median 1131, IQR 537-2501). By rurality, urban tracts (n=5734, 83.1%) had a median ratio of 1141 (IQR 2054), and rural tracts (n=1162, 16.9%) had a median ratio of 1093 (IQR 1690), indicating a broadly similar central tendency with somewhat greater variability in urban areas. The population-to-provider ratio exhibited significant positive spatial autocorrelation (global Moran I=0.305; Z=40.28; P<.001), consistent with clustered pockets of high and low availability rather than random dispersion.

Conclusions: A replicable geographic information system protocol can integrate licensure and demographic data to produce interpretable population-to-provider metrics and spatial diagnostics at the census-tract level. In Texas, rehabilitation workforce availability is spatially clustered and not explained solely by an urban-rural divide, underscoring the value of small-area mapping for equitable workforce planning and policy decisions.

(JMIR Form Res 2025;9:e85025) doi:10.2196/85025

XSL•FO
RenderX

KEYWORDS

geographic information systems; rehabilitation workforce; spatial analysis; workforce planning; health service access

Introduction

Access to health care is a key social determinant of health that directly influences outcomes across populations [1-6]. Within rehabilitation professions such as physical therapy (PT) and occupational therapy (OT), access is shaped by a combination of system-level and individual-level factors [7,8]. The concept of access has been widely studied and is frequently organized around the theory of access by Penchansky and Thomas [7], which outlines 6 interrelated dimensions: availability, accessibility, accommodation, affordability, acceptability, and awareness [8]. Among these, availability, or the extent to which a sufficient rehabilitation workforce is geographically and operationally present to meet population needs, is an underexamined dimension. Because rehabilitation care typically involves multiple visits across an episode of care, rehabilitation provider availability is essential to ensuring consistent, effective treatment [9].

The distribution of rehabilitation providers has direct implications for health outcomes. Limited availability can contribute to preventable decline in function, increased pain, and prolonged dependence on higher levels of care. The World Health Organization's Rehabilitation 2030 initiative highlights this concern, identifying a profound unmet global need for rehabilitation [8]. The 2019 Global Burden of Diseases, Injuries, and Risk Factors study estimated that 2.41 billion individuals have conditions that would benefit from rehabilitation, which contributed to a 69% increase in years lived with disability since 1990 [10,11]. In North and South America, 1 in 3 people could benefit from rehabilitation, and in the United States, approximately 5 in 11 individuals have conditions requiring these services [10,11]. This growing need highlights the importance of equitable workforce distribution as a prerequisite for access to care.

Despite the centrality of provider availability to access, rehabilitation disciplines such as PT and OT are not included in federal shortage designation frameworks such as health professional shortage areas or medically underserved areas and populations [12-15]. These designations, which are used for primary care, dentistry, mental health, and maternity care, rely on population-to-provider ratios and community indicators of need, such as poverty or infant mortality, and are spatially designated by the Health Resources and Services Administration as geographic information system (GIS)–compatible shapefiles [14-16]. However, no comparable structure exists for the rehabilitation workforce, limiting visibility into how rehabilitation provider availability aligns with population need for rehabilitation services.

Existing workforce data sources, such as state licensure registries, are often reported only at the county level and rarely integrate demographic or health indicators [17-19]. These limitations obscure local variation and hinder efforts to identify areas of unmet need. Recent workforce projections in PT have highlighted the lack of precise practice location and individual-level demographic data as a limitation in modeling physical therapist supply and distribution [20]. GISs provide a powerful tool to evaluate health care access by combining location-based data with population characteristics and social indicators [21-24].

Recent GIS-based workforce studies across diverse health professions demonstrate the value of spatial methods for identifying disparities in provider availability. National analyses in the United States have revealed substantial geographic variation in accessibility to health care providers and dental clinics, including the emergence of "clinic deserts" in low-capacity regions [25,26]. Internationally, GIS-based assessments in Taiwan, Brazil, China, and Ethiopia have documented maldistribution of rehabilitation and primary care resources, highlighting clusters of poor accessibility and mismatches between service capacity and population need [27-31]. Together, these studies illustrate that workforce inequities often reflect fine-grained spatial patterns rather than simple urban-rural divides, underscoring the importance of small-area analysis. However, few studies have applied these methods to the rehabilitation workforce in the United States or used licensure data to generate tract-level availability estimates, leaving a notable gap that this study addressed.

Prior studies have used GISs to assess rehabilitation access in urban and rural contexts, revealing disparities in proximity to care, travel time, and service availability in relation to population health needs [32-36]. However, these studies have typically focused on either small metropolitan areas or broad, aggregated regions and have not offered standardized, reproducible methods for data collection, integration, and analysis at a smaller geographic scale. There is a growing need for research that documents transparent, scalable approaches to workforce mapping in rehabilitation. At present, no census tract–level GIS mapping exists for the rehabilitation workforce, leaving a gap in understanding how rehabilitation provider availability aligns with population need. Census tracts, which are smaller statistical subdivisions averaging 4000 residents, offer a valuable unit of analysis for this type of inquiry [37,38]. This feasibility study addressed that gap by developing and testing a transparent, replicable analytic approach for rehabilitation workforce mapping.

The objective of this study was to describe the feasibility of a GIS-based methodology to assess the availability of rehabilitation providers at the census tract level. This methodology systematically integrates state licensure data with open access demographic datasets to geocode provider locations and links them to area-level indicators of population need. The resulting spatial framework enables calculation of population-to-provider ratios, visualization of workforce distribution, and identification of small-area disparities in provider availability. By outlining each step, from data acquisition and cleaning to spatial analysis and mapping, this feasibility study is designed to support future research,

XSL•FO
RenderX

workforce planning, and equity-focused policy development in rehabilitation.

Methods

Study Design

This was a descriptive, cross-sectional observational study designed to evaluate the availability of rehabilitation providers across census tracts using GISs. The primary objective was to estimate and compare population-to-provider ratios at the census tract level and examine how these ratios relate to the sociodemographic characteristics of each tract. This study used a methodology that combined state licensure data (obtained via paid public record requests) with open access demographic datasets. These data sources were integrated using spatial mapping techniques to facilitate small-area workforce analysis. Texas was chosen as the exemplar state because of its demographic diversity; pronounced urban-rural differences; and availability of comprehensive, high-quality licensure data. All processed datasets were stored locally in both CSV and geodatabase (.gdb) formats in ArcGIS Pro (Esri) to support spatial analysis.

When conducting GIS-based analyses of workforce availability, researchers should consider the geographic and demographic context of the region under study. Factors such as geographic land area, population density, and urban-rural distribution shape both the delivery and interpretation of health services and provider access.

Aggregated geographic units such as counties or zip codes, while commonly used, can mask meaningful local variation in provider availability and introduce misclassification bias. Zip codes, for example, often group together heterogeneous populations by combining areas of affluence with areas of poverty, which can distort analyses of access and need [39]. In contrast, census tracts, which typically include populations of 1200 to 8000 people and are designed to be socioeconomically homogeneous, provide a more precise basis for equitable spatial analysis [38].

The use of small-area units such as census tracts requires detailed data on provider location and population characteristics. This study's methodology is designed to support census tract–level analysis when such data are available. Ultimately, GIS-based access assessments must balance granularity and feasibility to generate meaningful, context-sensitive insights.

Ethical Considerations

This study methodology used secondary, deidentified data from publicly available sources, including licensure databases and census-based demographic datasets. No identifiable confidential information was disclosed, and no direct interaction with human subjects took place. Before implementation, this study was approved by the Texas Woman's University Institutional Review Board as exempt under title 45 of the Code of Federal Regulations part 46.104(d) as the research involved publicly available or nonidentifiable data. Institutional review board approval was obtained before conducting data analysis. All data were deidentified before data cleaning and analysis. The institutional review approval is available in Multimedia Appendix 1.

Data Sources

This study integrated information from two primary data sources: (1) state-level licensure data for rehabilitation professionals and (2) population and demographic data from the US Census Bureau's American Community Survey (ACS). For this study, we used Texas as the example state in the analysis.

Licensure data were obtained from the Executive Council of Physical Therapy and Occupational Therapy Examiners through publicly available open record requests in Texas in 2022 [40]. This dataset included rehabilitation provider information, including residential mailing address (including street, city, state, and zip code); license type; issue and expiration dates; and basic demographic characteristics such as gender and ethnicity, where available. Personally identifiable information (eg, name and license number) were removed before analysis. Rehabilitation provider addresses were used for spatial analysis, and all data were stored and processed in a deidentified format. Although the licensure dataset included an optional work or practice address field, this information was often incomplete or inaccurate as many licensees either repeated their residential address or left the field blank. Therefore, residential mailing addresses were used as a consistent and complete proxy for spatial analysis.

Demographic and population data were sourced from the US Census Bureau's 2020 ACS [41], a publicly available dataset providing census tract–level estimates for population size and characteristics such as race, ethnicity, age, disability status, insurance coverage, and income. Population estimates were obtained from the 2020 ACS dataset to match the licensure data time frame and provide reliable census tract–level denominators for ratio calculations. These data were downloaded as geodatabase files for integration with GIS software (eg, ArcGIS Pro).

These 2 data sources were used to estimate population-to-provider ratios and assess workforce availability at the census tract level.

Data Aggregation and Cleaning

After obtaining licensure and demographic data, a multistep cleaning and aggregation process was conducted to ensure accurate spatial representation of the rehabilitation workforce.

Licensure Data Cleaning and Inclusion Criteria

Licensure datasets typically include all actively licensed rehabilitation professionals within a state or jurisdiction, including individuals practicing via interstate licensure compacts and those residing out of state. To ensure an accurate geographic representation of rehabilitation providers residing and working within the region of interest (Texas) the dataset was filtered based on the following criteria: (1) inclusion of individuals with an in-state residential address; (2) exclusion of licensees reporting only out-of-state addresses or inactive or retired licensure status; (3) retention of individuals with in-state residential addresses but missing workplace address data, with

XSL•FO
RenderX

acknowledgment of potential uncertainty regarding employment status; and (4) manual review and correction of ambiguous or mislabeled geographic entries, such as addresses geocoded incorrectly or entries containing missing street-level information.

Figure 1 illustrates the selection and cleaning process, including inclusion and exclusion steps for licensure data. Filtering and cleaning procedures were conducted using R (version 4.3.1 or higher; R Foundation for Statistical Computing) within RStudio (Posit PBC). The code used for data cleaning is documented and available in Multimedia Appendices 2 and 3.

Figure 1. Licensure data cleaning process.

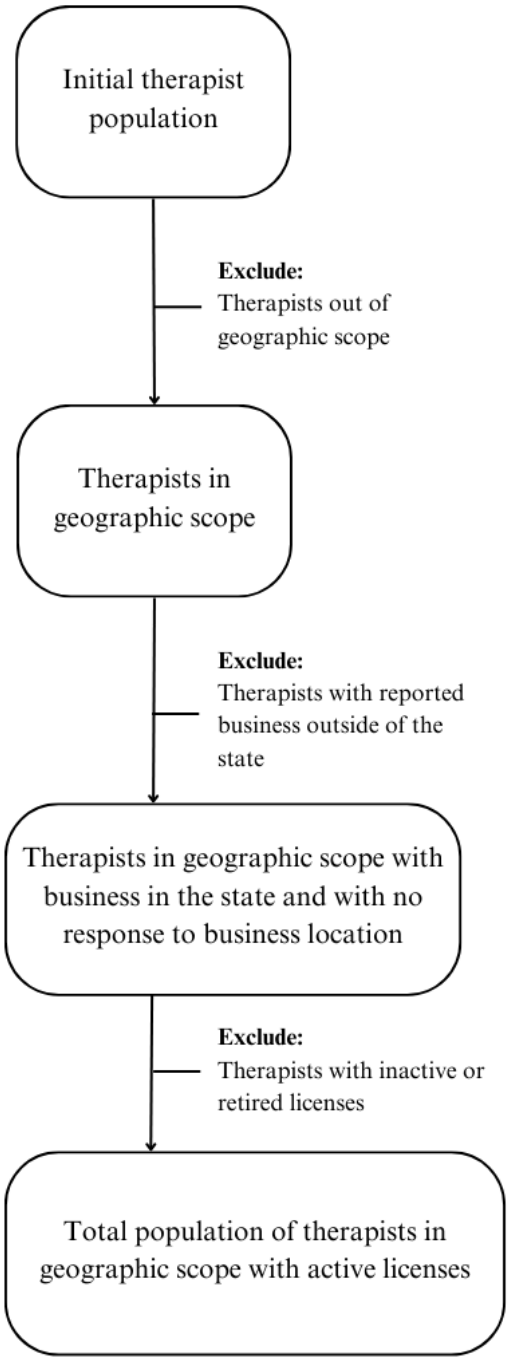

XSL•FO
RenderX

Software

All geospatial procedures were conducted using ArcGIS Pro (version 3.3), a GIS platform that allows for spatial data visualization, management, and analysis. Data cleaning and preprocessing were completed in RStudio before import into ArcGIS.

Demographic and Census Data Cleaning

Census tract–level data from the ACS were filtered to match the relevant geographic units used for licensure data aggregation. Relevant demographic variables (eg, population size, poverty rate, disability prevalence, and insurance coverage) were cleaned and formatted to allow for merging with provider datasets using census tract–level geographic identifiers.

Geospatial Analysis Methods

Overview

In ArcGIS, for the purpose of visualization and analysis, the first step is to create a feature layer, which is a collection of similar geographic features that are grouped together [42]. Thus, feature layers were created for individual PT and OT datasets. These were combined into 1 feature layer with both PT and OT data that were used for analysis. Next, an ArcGIS geoprocessing tool called a spatial join that joins attributes from one feature to another based on the spatial relationship was added for data management. This spatial join was conducted for the combined PT and OT feature layer to match each rehabilitation provider spatially to the census tract in which their residential address was located. The join count resulting from this spatial join represents the number of providers residing in each of the 6896 census tracts from the 2020 census. This dataset was then merged with 2020 census tracts that included over 70 ACS demographic variables [43]. The full code for geocoding and spatial joins is included in Multimedia Appendix 4.

Calculating the Population-to-Provider Ratio

To estimate the population-to-provider ratio, a basic calculation of demand over supply was used. The population in the census tract, or the measure of the demand for services (d_i), was divided by the provider join count, or the number of physical and occupational therapists residing in the census tract (p_i):

Additional adjustments were made to address zero-population and zero-provider tracts. Census tracts with zero population were excluded from ratio calculations as no residents indicate no potential demand for services. Because the population-to-provider ratio includes provider count in the denominator, tracts with zero providers would otherwise yield undefined values. To retain these tracts, which represent communities with residents but no local providers, a continuity correction was applied by substituting a denominator value of 1 for tracts with residents but zero providers. This approach preserved meaningful variation across all populated tracts while preventing division-by-zero and maintaining comparability across the dataset.

Although this method slightly attenuates extreme ratio values, it allows for visualization and analysis of areas where service absence is a critical feature of access inequity. This correction was selected over data exclusion or imputation to preserve geographic completeness and interpretability of spatial clustering. All adjustments were implemented in ArcGIS using a Python code block (Python Software Foundation), and the code is provided in Multimedia Appendix 4.

Next, using the adjusted provider count, the population-to-provider ratio was calculated by creating a new field and calculating that field by dividing the total population per census tract by the adjusted join count.

Descriptive Mapping

The population-to-provider ratios across census tracts were investigated using descriptive statistics and choropleth and bivariate maps. A choropleth map is a thematic map in which areas are styled to represent the variation in a single variable. A bivariate map displays 2 variables using variations in colors or symbols, where each variable is assigned a different graded color scheme in a 3 × 3 table that is included in the top right-hand corner of the bivariate map [44,45]. A graduated color choropleth map was created to visually analyze the distribution of the population-to-provider ratios across the state of Texas by quintiles. Population-to-provider ratios were calculated and categorized into quintiles to facilitate relative comparison across areas following the approach used by Brown et al [46], who used quintile-based classifications to highlight spatial variation in the absence of established adequacy benchmarks. Comparing the highest and lowest quintiles facilitates the examination of factors associated with substantial difference in provider availability across areas. Bivariate choropleth maps were then created to visually describe the relationship between population-to-provider ratios and the percentage of disability in each census tract as defined by the ACS. Bivariate choropleth maps allow for the visualization of the specific areas that may be most in need of providers based on specific population demographics.

Statistical Analyses

Descriptive statistics were used to describe the variability in population-to-provider ratio across all census tracts within the state of Texas. To describe differences in provider availability by rurality, census tracts were classified using 2020 Rural-Urban Commuting Area (RUCA) codes [47]. Tracts with a primary RUCA code of <4 were categorized as urban, and those with codes of ≥4 were categorized as rural. Each tract's population-to-provider ratio was summarized by rurality using measures of central tendency and dispersion (median, IQR, minimum, and maximum). All analyses were conducted in R (version 4.3) using the *dplyr* and *ggplot2* packages. The R code for the RUCA analysis is included in Multimedia Appendix 5.

To evaluate whether the spatial distribution of access to rehabilitation providers exhibited clustering across the study area, a global Moran *I* statistic was computed for the population-to-provider ratio at the census tract level using ArcGIS Pro (version 3.3). Because Texas census tracts are polygonal units that vary substantially in geographic size, with

XSL•FO
RenderX

very large tracts in rural regions and very small tracts in dense urban areas, a contiguity-based conceptualization of spatial relationships (edges only) was selected instead of a fixed distance threshold. This approach defines neighbors as tracts sharing a common boundary segment, ensuring that all tracts, regardless of their physical size, are evaluated based on direct adjacency rather than distance alone. Row standardization was applied to account for variability in the number of neighbors across tracts. The analysis tested whether the distribution of the population-to-provider ratio was spatially random, clustered, or dispersed using 999 random permutations to assess significance. *Z* scores and *P* values were used to evaluate departures from spatial randomness.

Measures of Feasibility

The following measures of feasibility were assessed throughout the study: ability to obtain licensure data and the percentage of rehabilitation providers included after cleaning, the time and resources needed to clean and geocode the data, the percentage of provider addresses that were successfully geocoded, the ability to merge licensure data with ACS data, the percentage of census tracts that were matched, the feasibility of the geospatial analysis methods, and the relevance of the results to workforce planning and their replication.

Results

Overview

Across 6896 tracts, ratios ranged from 4.5 to 11,147 persons per rehabilitation provider (median 1131, IQR 537-2501). Tracts with the lowest ratios clustered in the Texas Medical Center (eg, census tract 313102), whereas the highest ratios were observed in Lubbock (002100) and east of Dallas (census tract 012000). The distribution of population-to-provider ratios across the state of Texas is shown in Figure 2.

Figure 2. Distribution of population-to-provider ratios for rehabilitation providers.

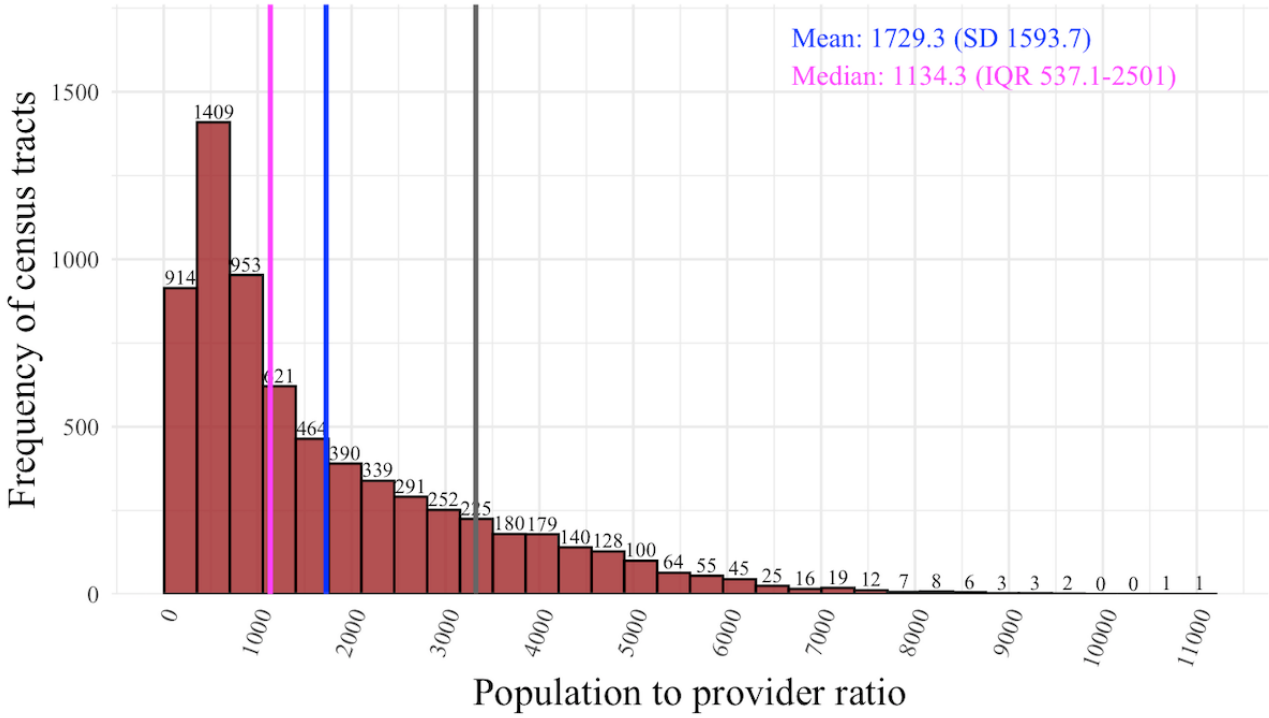

When stratified by rurality, urban tracts (5734/6896, 83.1%) had a median ratio of 1141 (IQR 2054), whereas rural tracts (1162/6896, 16.9%) had a median ratio of 1093 (IQR 1690). These values indicate broadly similar distributions of provider availability between urban and rural areas, with slightly greater variability in urban tracts. Box plots (Figure 3) show the distribution of census tract–level population-to-provider ratios for rehabilitation providers classified by RUCA code. Ratios are displayed on a logarithmic scale to account for skewness in the data. Median ratios and IQRs are similar across rural and urban tracts, suggesting comparable variability in provider availability across both settings.

The global Moran *I* statistic for the population-to-provider ratio across census tracts was 0.305 (Z=40.28; P<.001), indicating significant positive spatial autocorrelation and less than a 1% chance that this clustered pattern was a result of random chance. The ArcGIS Moran *I* results are included in Multimedia Appendix 6.

Graduated and bivariate choropleth maps were used to visualize provider availability and its intersection with social and demographic indicators. A graduated color choropleth map (Figure 4) displays population-to-provider ratios across all census tracts in Texas, with lighter shades indicating higher provider availability (lower population-to-provider ratio) and darker shades indicating lower provider availability.

XSL•FO
RenderX

Figure 3. Distribution of population-to-provider ratios fro rehabilitation providers by rurality. RUCA: Rural-Urban Commuting Area.

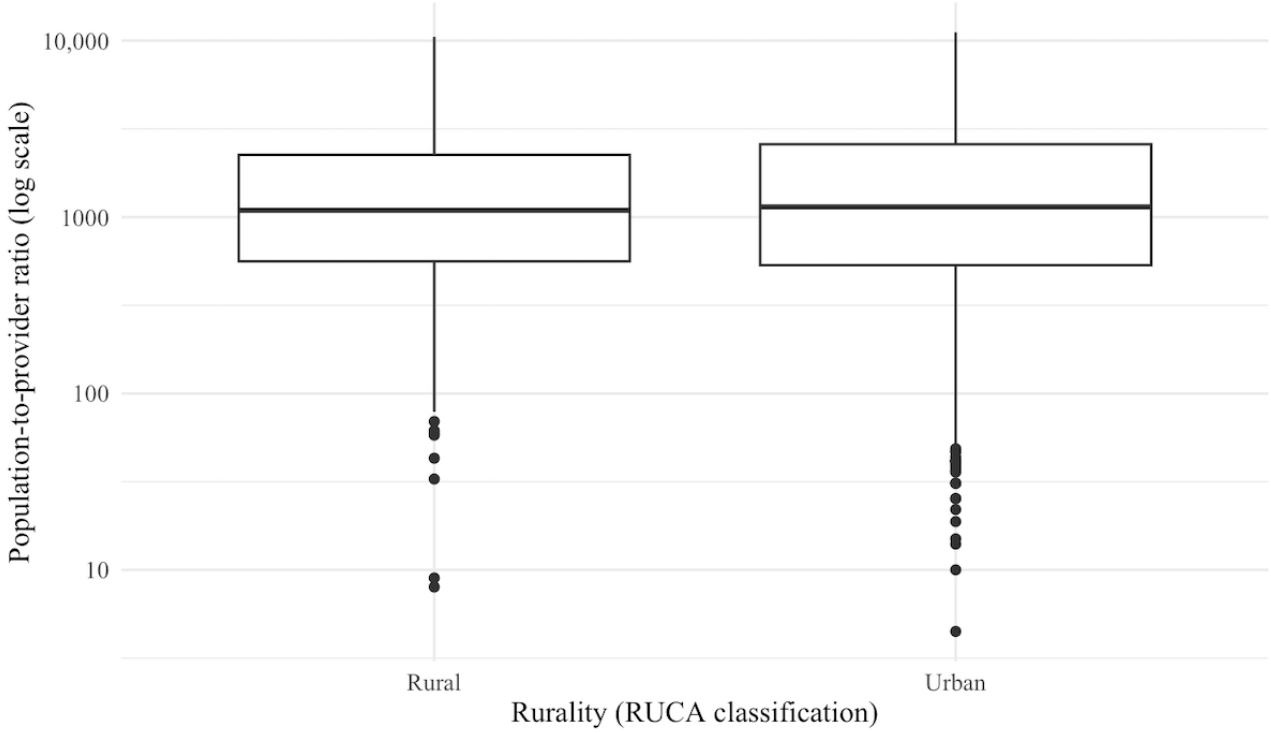

Figure 4. Spatial distribution of population-to-provider ratios for rehabilitation providers across Texas census tracts (combined occupational therapy and physical therapy providers).

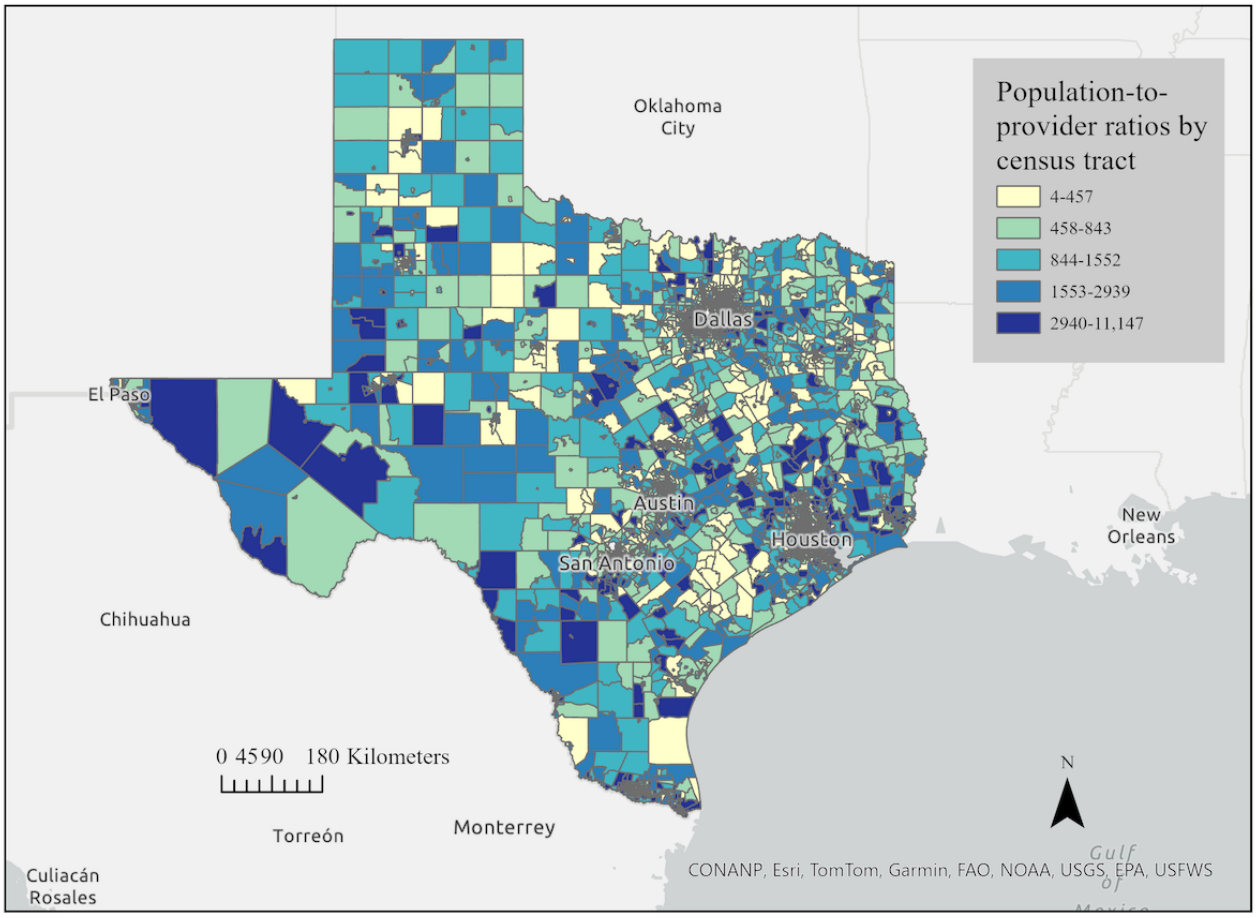

To further illustrate equity-related patterns, bivariate choropleth maps overlay these ratios with variables such as percentage of disability, percentage of minority populations, and percentage of the population below the poverty level. For example,

XSL•FO
RenderX

disability data were drawn from 6 ACS questions addressing hearing, vision, cognitive ability, ambulatory ability, self-care, and independent living difficulty [48]; respondents answering "yes" to any item were classified as having a disability.

As shown in Figure 5, the overlay of population-to-provider ratios with disability prevalence reveals regional variability across the state, although not a strict urban-rural divide. While areas of west and south Texas include clusters of high population-to-provider ratios, similar patterns also appear in parts of urban and periurban counties. Conversely, some rural tracts demonstrate moderate ratios despite small provider counts, reflecting the influence of low population density on the calculation. Regions in dark blue reflect regions where there is a high population-to-provider ratio coinciding with high prevalence of disability, suggestive of census tracts where there is a mismatch of population need and provider availability.

Figure 5. Bivariate choropleth map by quintile of the relationship between population-to-provider ratio (for rehabilitation providers) and percentage of disability within each census tract in Texas.

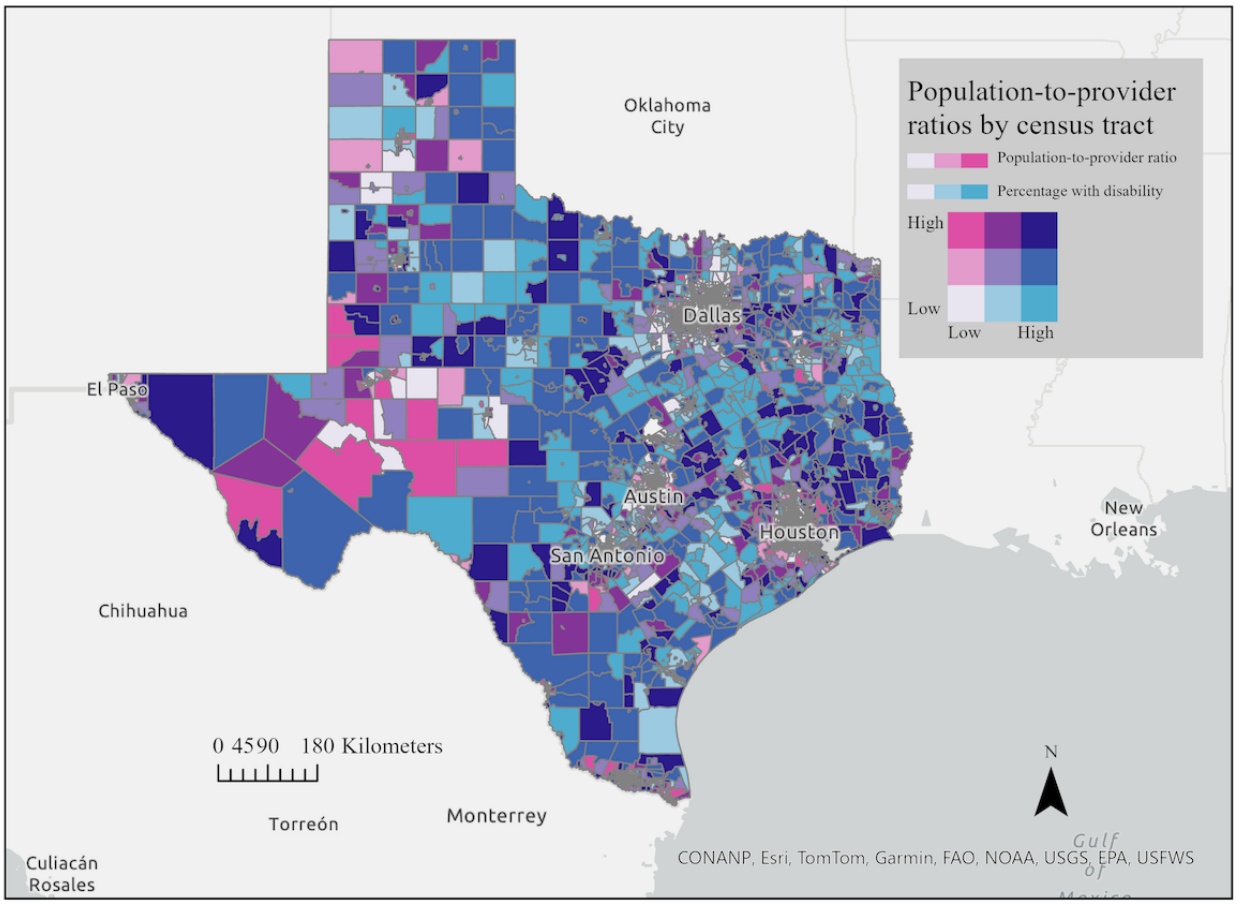

Feasibility

Feasibility was assessed throughout each phase of the methods, including data acquisition, data processing, descriptive mapping, and statistical analysis. Table 1 illustrates the measures of feasibility assessed in this study.

XSL•FO
RenderX

Table 1. Feasibility results.

Domain	Indicator	Results
Data acquisition	Ability to obtain licensure data; percentage of licensed rehabilitation providers included after cleaning	Licensure data were purchased and accessible in 2022; 28,613 PTs[a] (88% of initial entries) and 16,501 OTs[b] (87% of initial entries) were retained after cleaning.
Data processing	Time and resources needed to clean and geocode data	Data cleaning required manual review of address inconsistencies and development of flowcharts and R code. Processing took approximately 4 weeks of graduate assistant time with regular oversight by the investigators.
Geocoding success	Percentage of rehabilitation provider addresses successfully geocoded	Nearly all addresses were successfully geocoded; only 9 PT and 5 OT addresses (<0.05% of total entries) were excluded because they could not be verified.
Data integration	Ability to merge licensure data with ACS[c] data; percentage of tracts matched	Rehabilitation provider data (28,613 PTs and 16,501 OTs) were successfully matched to 6896 Texas census tracts and merged with ≥70 ACS variables, with no unmatched tracts.
Analysis of feasibility	Ability to generate population-to-provider ratios, maps, and bivariate overlays	Ratios were calculated across all tracts; maps and overlays were successfully created using ArcGIS. Analyses were replicable using shared R code and ArcGIS workflows.
Acceptability and utility	Relevance for workforce planning; potential for replication	Methods allowed for visualization of geographic disparities in provider availability. The workflows (R scripts, flow diagrams, and verification steps) are reproducible and adaptable to other states and health professions, although feasibility will depend on state-specific licensure access rules.

[a]PT: physical therapist.

[b]OT: occupational therapist.

[c]ACS: American Community Survey.

Discussion

Principal Findings

Rehabilitation provider availability showed substantial geographic variation, with tract-level population-to-provider ratios ranging from 4.5 to 11,147 (median 1131, IQR 2010) and similar central tendencies across rural and urban areas, yet the strong positive spatial autocorrelation (Moran I=0.305; P<.001) indicated clear clustering of high- and low-availability regions.

These patterns align with findings from other GIS-based workforce studies across a range of health professions, which similarly report uneven spatial distribution of health care providers and localized pockets of limited access. National analyses in the United States have identified substantial variation in spatial accessibility to health care providers and dental clinics [25,26], whereas international assessments in Taiwan, Brazil, China, and Ethiopia have documented comparable clustering of high- and low-access areas across rehabilitation and community care services [27-31]. Although these studies differ in setting and methodological approach, they consistently demonstrate that workforce maldistribution emerges at small geographic scales and often reflects underlying sociodemographic gradients. This study extends this body of work to the US rehabilitation workforce by applying comparable geospatial methods to tract-level licensure data, situating our findings within a broader pattern of spatially concentrated workforce inequities.

Statewide patterns in this study demonstrated striking variation in population-to-provider ratios across census tracts, revealing local disparities that county-level summaries fail to capture. Previous statewide reports from the Texas Health Professions Resource Center have suggested that availability challenges are largely a rural issue [17,18]. In 2022, the average population-to-provider ratio for physical therapists was 1593:1 in metropolitan counties and 2870:1 in nonmetropolitan counties, whereas for occupational therapists, it was 2778:1 in metropolitan areas and 5611:1 in nonmetropolitan areas. These data painted a picture of a binary divide between urban and rural access.

When we examined rehabilitation provider availability at the census tract level rather than the more common county designation, our findings diverged from prevailing assumptions. Although state-level reports show markedly higher population-to-provider ratios in nonmetropolitan counties than in metropolitan ones, our tract-level analysis found similar median values for urban and rural tracts, albeit with greater variability in urban areas. This pattern mirrors findings for other health professions: for example, a spatial analysis of primary care access in Philadelphia, Pennsylvania, revealed that geographic access disparities often overlay racial and socioeconomic patterns rather than simply urban versus rural gradients [46]. By shifting to smaller geographic units, our study revealed that within-urban and within-rural heterogeneity matters because access gaps may persist in urban neighborhoods that are often assumed to be well served and, conversely, some rural tracts may fare better than county-level data suggest.

The Moran I statistic confirms that rehabilitation provider availability is spatially clustered rather than randomly distributed. In other words, tracts with low provider availability tend to be near one another, forming regional patterns of limited access. A comparable spatial analysis of rehabilitation human resources in China demonstrated this clustering effect across provinces, underscoring that workforce maldistribution may follow geographic spillovers rather than isolated pockets [49]. However, in our Texas analysis, these clusters did not

XSL•FO
RenderX

correspond neatly to urban or rural classifications, indicating that factors beyond simple rurality may drive spatial patterns of provider scarcity. Although further research should investigate these determinants in depth, the observed clustering suggests that improving access will likely require coordinated, regionally targeted interventions rather than isolated, tract-by-tract fixes.

Although this analysis was based on provider residential addresses, the presence of strong spatial clustering indicates that these data still meaningfully represent patterns of service availability. It is unlikely that large numbers of providers are commuting long distances into the clusters of low access identified in this study. Instead, these regions likely represent genuine shortages in the local rehabilitation workforce. Recognizing and addressing these gaps is essential not only from an equity standpoint but also from a broader population health perspective. Limited access to rehabilitation care can lead to preventable disability, loss of function, and reduced participation in work and community life [50,51].

The GIS-based maps generated in this study extend beyond visualization: they function as actionable tools for workforce assessment and planning. Similar spatial approaches have been applied in other areas of public health to identify regions where need and provider availability intersect, informing policy and resource allocation decisions. For example, bivariate choropleth maps have been used in cancer control to guide the placement of screening resources in areas with low service capacity and high disease burden [52] and in studies of lung cancer screening to identify US counties with both high smoking prevalence and low provider density [53]. Such mapping frameworks demonstrate the potential of GIS-based methods to translate complex population and workforce data into actionable insights. Applying this approach to rehabilitation introduces an equivalent capacity to identify where population need and provider scarcity converge. By pinpointing census tracts with high disability prevalence and low provider availability, these maps enable decision-makers to prioritize areas for resource allocation, community partnerships, and program expansion. Beyond workforce planning, they also provide a framework for longitudinal monitoring, allowing policymakers to observe how workforce distribution shifts in response to educational pipelines, policy incentives, or population change over time. In this way, the maps produced in this study serve not only as evidence of current inequities but also as instruments for promoting future workforce equity in rehabilitation.

Strengths and Limitations

The approach presented in this paper is replicable; the step-by-step procedures, along with the inclusion of code in the multimedia appendices, are intentional so that the methods can be applied in a variety of geographic contexts or across different health professions. This transparency supports reproducibility and scalability in future research. Additionally, the novel method emphasizes high geographic resolution by conducting analyses at the census tract level. This finer scale allows for a more precise identification of geographic disparities than approaches that rely on broader units such as counties or zip codes [38,39], offering a more detailed understanding of provider availability.

There are several limitations to this study. First, it focused exclusively on availability, one domain within the theory of access by Penchansky and Thomas [7], and did not account for realized access (the actual entry of a patient into the health care system) [54] or outcomes. The use of residential addresses as a proxy for provider practice location introduced potential misclassification, and the population-to-provider ratio did not capture factors such as transportation infrastructure, health insurance coverage, or service capacity.

Additionally, the observational nature of the analysis precludes causal inferences about relationships between availability and health disparities. This study also prioritized comparison of high- and low-availability areas using quintiles, which may not account for regional variations in land mass, population density, or local health systems.

Finally, publicly available licensure and census data are subject to reporting errors and definitional inconsistencies. These data limitations should be acknowledged in future research and addressed, where possible, through improved workforce data infrastructure.

Future Applications

The methods used in this study are adaptable and scalable for use in other professions (eg, speech-language pathology and behavioral health) and other geographic regions. Future studies can build on this framework by incorporating clinic location data, travel time analyses, and patient-level outcomes to provide a more complete picture of access to care. Integration of GISs into national workforce planning tools could support more equitable distribution of health services and guide data-driven decisions at local, state, and federal levels.

Beyond immediate replication, this work lays the foundation for a broader research agenda in rehabilitation workforce equity. The GIS-based approach demonstrated in this paper establishes a platform for developing rehabilitation-specific shortage area indexes, evaluating spatial alignment between workforce supply and population need, and examining equity in workforce diversity. These next steps extend beyond feasibility to inform national policy and education initiatives aimed at improving access for underserved communities. The ability to acquire, clean, and link licensure and census data demonstrates both the practicality and potential of applying this method to larger, longitudinal studies that track changes in provider availability and access over time. In this way, the results directly inform workforce planning by providing an evidence-based foundation for policies and programs aimed at achieving a more equitable distribution of rehabilitation providers.

Conclusions

This study demonstrates the feasibility and value of integrating licensure and demographic data to assess rehabilitation provider availability at the census tract level, which is a novel approach in this field. Quantitatively, tract-level population-to-provider ratios ranged from 4.5 to 11,147 (median 1131, IQR 2010) and demonstrated significant spatial clustering (Moran I=0.305; P<.001). The results reveal that provider availability in Texas is spatially clustered and influenced by factors beyond

XSL•FO
RenderX

urban-rural status, highlighting the need for targeted, regionally informed interventions.

By establishing a replicable, GIS-based framework, this work advances both workforce equity research and geospatial health methods, offering a foundation for small-area analyses that move beyond traditional county-level approaches. These findings underscore the potential of spatial mapping to identify underserved communities and guide data-driven workforce planning and policy development that promote equitable access to rehabilitation care.

Acknowledgments

This work was completed in partial fulfillment of the requirements for the doctor of philosophy degree at Texas Woman's University. The author gratefully acknowledges Hiro Chang and Andrea Koh for their invaluable assistance with data cleaning and verification. Additional support for this project was provided by the Center on Health Services Training and Research funded by the Foundation for Physical Therapy Research.

Funding

This project was supported by the Center on Health Services Training and Research Pilot Study Grant. Funding was allocated for data acquisition, software licensing, and research assistant time. Additional resources may be required for broader implementation or extension of the methodology across other health care provider types or states. No commercial or private interests were involved in the design or intended dissemination of this study.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Institutional review board approval letter from Texas Woman's University for this research protocol.
[PDF File (Adobe PDF File), 151 KB - formative_v9i1e85025_app1.pdf]

Multimedia Appendix 2
R code for cleaning physical therapist licensure data, creating subsets for spatial and statistical analysis.
[PDF File (Adobe PDF File), 36 KB - formative_v9i1e85025_app2.pdf]

Multimedia Appendix 3
R code for cleaning occupational therapist licensure data, creating subsets for spatial and statistical analysis.
[PDF File (Adobe PDF File), 32 KB - formative_v9i1e85025_app3.pdf]

Multimedia Appendix 4
Python code block for ArcGIS reclassifications.
[PDF File (Adobe PDF File), 18 KB - formative_v9i1e85025_app4.pdf]

Multimedia Appendix 5
R code for Rural-Urban Commuting Area classification analysis.
[PDF File (Adobe PDF File), 62 KB - formative_v9i1e85025_app5.pdf]

Multimedia Appendix 6
Moran I ArcGIS results output.
[PDF File (Adobe PDF File), 144 KB - formative_v9i1e85025_app6.pdf]

References

1. Braaten AD, Hanebuth C, McPherson H, Smallwood D, Kaplan S, Basirico D, et al. Social determinants of health are associated with physical therapy use: a systematic review. Br J Sports Med 2021 Nov 03;55(22):1293-1300. [doi: 10.1136/bjsports-2020-103475] [Medline: 34083223]

XSL•FO
RenderX

2. Magnusson DM, Rethorn DZ, Bradford EH, Maxwell J, Ingman MS, Davenport TE, et al. Population health, prevention, health promotion, and wellness competencies in physical therapist professional education: results of a modified Delphi study. Phys Ther 2020 Aug 31;100(9):1645-1658 [FREE Full text] [doi: 10.1093/ptj/pzaa056] [Medline: 32280980]
3. Rethorn ZD, Cook C, Reneker JC. Social determinants of health: if you aren't measuring them, you aren't seeing the big picture. J Orthop Sports Phys Ther 2019 Dec;49(12):872-874. [doi: 10.2519/jospt.2019.0613] [Medline: 31789121]
4. Rethorn ZD, Rethorn TJ, Cook CE, Sharpe JA, Hastings SN, Allen KD. Association of burden and prevalence of arthritis with disparities in social risk factors, findings from 17 US states. Prev Chronic Dis 2022 Feb 17;19:E08 [FREE Full text] [doi: 10.5888/pcd19.210277] [Medline: 35175917]
5. Douthit N, Kiv S, Dwolatzky T, Biswas S. Exposing some important barriers to health care access in the rural USA. Public Health 2015 Jun;129(6):611-620. [doi: 10.1016/j.puhe.2015.04.001] [Medline: 26025176]
6. McGibbon E, Etowa J, McPherson C. Health-care access as a social determinant of health. Can Nurse 2008 Sep;104(7):22-27. [Medline: 18856224]
7. Penchansky R, Thomas JW. The concept of access: definition and relationship to consumer satisfaction. Med Care 1981 Feb;19(2):127-140. [doi: 10.1097/00005650-198102000-00001] [Medline: 7206846]
8. Saurman E. Improving access: modifying Penchansky and Thomas's theory of access. J Health Serv Res Policy 2016 Jan 15;21(1):36-39. [doi: 10.1177/1355819615600001] [Medline: 26377728]
9. Kava CM, Fishleder S, Steinman L, Petrescu-Prahova M, Schrodt L, Harris JR. Provider-patient communication and involvement in physical activity programs among patients receiving physical therapy services: a qualitative study. J Appl Gerontol 2020 Sep 02;39(9):1000-1007 [FREE Full text] [doi: 10.1177/0733464819847402] [Medline: 31043115]
10. Cieza A, Causey K, Kamenov K, Hanson SW, Chatterji S, Vos T. Global estimates of the need for rehabilitation based on the Global Burden of Disease study 2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet 2021 Dec 19;396(10267):2006-2017 [FREE Full text] [doi: 10.1016/S0140-6736(20)32340-0] [Medline: 33275908]
11. WHO rehabilitation need estimator. Institute for Health Metrics and Evaluation. URL: https://vizhub.healthdata.org/rehabilitation/ [accessed 2022-10-16]
12. MUA find. Health Resources & Services Administration. URL: https://data.hrsa.gov/tools/shortage-area/mua-find [accessed 2023-10-10]
13. Ricketts TC, Goldsmith LJ, Holmes GM, Randolph RM, Lee R, Taylor DH, et al. Designating places and populations as medically underserved: a proposal for a new approach. J Health Care Poor Underserved 2007 Aug;18(3):567-589. [doi: 10.1353/hpu.2007.0065] [Medline: 17675714]
14. HPSA find. Health Resources & Services Administration. URL: https://data.hrsa.gov/tools/shortage-area/hpsa-find [accessed 2025-01-20]
15. Shortage designation scoring criteria. Health Resources & Services Administration. URL: https://bhw.hrsa.gov/workforce-shortage-areas/shortage-designation/scoring [accessed 2025-03-18]
16. Criteria for determining priorities among health professional shortage areas. Federal Register, Health Resources & Services Administration. URL: https://www.federalregister.gov/d/03-13478 [accessed 2025-05-29]
17. Physical therapists 2022 - factsheet A. Texas Department of State Health Services, Center for Health Statistics, Health Professions Resource Center. URL: https://www.dshs.texas.gov/sites/default/files/chs/hprc/publications/2022/Physical%20Therapists%202022%20-%20Factsheet%20A.pdf [accessed 2025-03-18]
18. Occupational therapists 2022 - fact sheet A. Texas Department of State Health Services, Center for Health Statistics, Health Professions Resource Center. URL: https://www.dshs.texas.gov/sites/default/files/chs/hprc/publications/2022/Occupational%20Therapist%202022%20-%20Fact%20Sheet%20A.pdf [accessed 2025-03-18]
19. Physical therapy workforce reports. Minnesota Department of Health, Office of Rural Health and Primary Care. URL: https://www.health.state.mn.us/data/workforce/pt/index.html [accessed 2025-03-18]
20. Zarek P, Ruttinger C, Armstrong D, Chakrabarti R, Hess DR, Manal TJ, et al. Current and projected future supply and demand for physical therapists from 2022 to 2037: a new approach using microsimulation. Phys Ther 2025 Mar 03;105(3):pzaf014. [doi: 10.1093/ptj/pzaf014] [Medline: 40037340]
21. Shaw N, McGuire S. Understanding the use of geographical information systems (GIS) in health informatics research: a review. J Innov Health Inform 2017 Jun 23;24(2):940 [FREE Full text] [doi: 10.14236/jhi.v24i2.940] [Medline: 28749317]
22. Dulin MF, Ludden TM, Tapp H, Blackwell J, de Hernandez BU, Smith HA, et al. Using Geographic Information Systems (GIS) to understand a community's primary care needs. J Am Board Fam Med 2010 Jan 05;23(1):13-21 [FREE Full text] [doi: 10.3122/jabfm.2010.01.090135] [Medline: 20051538]
23. Edward J, Biddle DJ. Using geographic information systems (GIS) to examine barriers to healthcare access for Hispanic and Latino immigrants in the U.S. South. J Racial Ethn Health Disparities 2017 Apr 29;4(2):297-307. [doi: 10.1007/s40615-016-0229-9] [Medline: 27129855]
24. Hutson SN. Understanding social, legal, economic, and spatial barriers to healthcare access in El Paso County, Texas colonias: an examination of structural violence using mixed methods [dissertation]. University of Colorado. 2017. URL: https://www.proquest.com/openview/a73e6eac773503479552dc387094e713/1?pq-origsite=gscholar&cbl=18750 [accessed 2025-12-11]

XSL•FO
RenderX

25. Naylor KB, Tootoo J, Yakusheva O, Shipman SA, Bynum JP, Davis MA. Geographic variation in spatial accessibility of U.S. healthcare providers. PLoS One 2019 Apr 9;14(4):e0215016 [FREE Full text] [doi: 10.1371/journal.pone.0215016] [Medline: 30964933]
26. Rahman MS, Blossom JC, Kawachi I, Tipirneni R, Elani HW. Dental clinic deserts in the US: spatial accessibility analysis. JAMA Netw Open 2024 Dec 02;7(12):e2451625 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.51625] [Medline: 39714842]
27. Wu HC, Tseng MM, Lin CC. Assessment on distributional fairness of physical rehabilitation resource allocation: geographic accessibility analysis integrating google rating mechanism. Int J Environ Res Public Health 2020 Oct 18;17(20):7576 [FREE Full text] [doi: 10.3390/ijerph17207576] [Medline: 33081016]
28. Tseng MH, Wu HC. Accessibility assessment of community care resources using maximum-equity optimization of supply capacity allocation. Int J Environ Res Public Health 2021 Jan 28;18(3):1153 [FREE Full text] [doi: 10.3390/ijerph18031153] [Medline: 33525529]
29. Pereira AG, Ferreira AF, Merey LS, Carli AD, Gomes AM, Santos ML. Assistance gaps in physical rehabilitation: spatial analysis of physical therapy services and users' households in a Brazilian capital. Rev Bras Epidemiol 2022 May 23;25:e220011 [FREE Full text] [doi: 10.1590/1980-549720220011.2] [Medline: 35613249]
30. Qiu N, Zhang T, Cheng J. Examining the impact of spatial accessibility to rehabilitation facilities on the degree of disability: a heterogeneity perspective. SSM Popul Health 2023 Aug 09;23:101489 [FREE Full text] [doi: 10.1016/j.ssmph.2023.101489] [Medline: 37588767]
31. Hierink F, Oladeji O, Robins A, Muñiz MF, Ayalew Y, Ray N. A geospatial analysis of accessibility and availability to implement the primary healthcare roadmap in Ethiopia. Commun Med (Lond) 2023 Oct 07;3(1):140 [FREE Full text] [doi: 10.1038/s43856-023-00372-z] [Medline: 37805668]
32. Chan DV, Mann A, Gopal S. Applying environmental context to rehabilitation research using geographic information systems and global positioning systems geospatial technologies. Rehabil Res Policy Educ 2021 Mar 26;35(1):33-50 [FREE Full text] [doi: 10.1891/re-19-39] [Medline: 34306839]
33. Rozvar C, Judd D, Magnusson D. Disparities in geographic proximity to outpatient physical therapist services in a large metropolitan area. Health Policy 2019;3(19):5-12 [FREE Full text]
34. Gao F, Foster M, Liu Y. Disability concentration and access to rehabilitation services: a pilot spatial assessment applying geographic information system analysis. Disabil Rehabil 2019 Oct 04;41(20):2468-2476. [doi: 10.1080/09638288.2018.1468931] [Medline: 29726287]
35. Lakhani A, Parekh S, Gudes O, Grimbeek P, Harre P, Stocker J, et al. Disability support services in Queensland, Australia: identifying service gaps through spatial analysis. Appl Geogr 2019 Sep;110:102045. [doi: 10.1016/j.apgeog.2019.102045]
36. Pak SS, Ratoza M, Cheuy V. Examining rehabilitation access disparities: an integrated analysis of electronic health record data and population characteristics through bivariate choropleth mapping. BMC Health Serv Res 2024 Feb 07;24(1):170 [FREE Full text] [doi: 10.1186/s12913-024-10649-1] [Medline: 38321457]
37. Geography: education resources. United States Census Bureau. URL: https://www2.census.gov/geo/pdfs/education/ [accessed 2025-05-29]
38. Glossary. United States Census Bureau. URL: https://www.census.gov/programs-surveys/geography/about/glossary.html#par_textimage_13 [accessed 2023-08-03]
39. Pearson J, Jacobson C, Ugochukwu N, Asare E, Kan K, Pace N, et al. Geospatial analysis of patients' social determinants of health for health systems science and disparity research. Int Anesthesiol Clin 2023 Jan 01;61(1):49-62 [FREE Full text] [doi: 10.1097/AIA.0000000000000389] [Medline: 36480649]
40. Open records request. Executive Council of Physical Therapy and Occupational Therapy Examiners. URL: https://ptot.texas.gov/open-records-request/ [accessed 2023-01-23]
41. American Community Survey (ACS). United States Census Bureau. URL: https://www.census.gov/programs-surveys/acs [accessed 2024-11-24]
42. Feature layers. ArcGIS Enterprise. URL: https://enterprise.arcgis.com/en/portal/latest/use/feature-layers.htm [accessed 2024-07-09]
43. TxDOT open data portal. ArcGIS Hub. URL: https://gis-txdot.opendata.arcgis.com/search?collection=Dataset [accessed 2024-11-07]
44. Choropleth map definition. Esri. URL: https://support.esri.com/en-us/gis-dictionary/choropleth-map [accessed 2024-09-12]
45. Bivariate map definition. Esri. URL: https://support.esri.com/en-us/gis-dictionary/bivariate-map [accessed 2024-09-12]
46. Brown EJ, Polsky D, Barbu CM, Seymour JW, Grande D. Racial disparities in geographic access to primary care in Philadelphia. Health Aff (Millwood) 2016 Aug 01;35(8):1374-1381. [doi: 10.1377/hlthaff.2015.1612] [Medline: 27503960]
47. Dobis EA, Sanders A. Rural-urban commuting area codes. Economic Research Service. URL: https://www.ers.usda.gov/data-products/rural-urban-commuting-area-codes [accessed 2025-11-10]
48. How disability data are collected from the American community survey. United States Census Bureau. URL: https://www.census.gov/topics/health/disability/guidance/data-collection-acs.html [accessed 2024-09-12]

XSL•FO
RenderX

49. Chen C, Chen T, Zhao N, Dong S. Regional maldistribution of human resources of rehabilitation institutions in China Mainland based on spatial analysis. Front Public Health 2022 Nov 8;10:1028235 [FREE Full text] [doi: 10.3389/fpubh.2022.1028235] [Medline: 36424956]
50. Rehabilitation. World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/rehabilitation [accessed 2023-02-13]
51. Neill R, Shawar YR, Ashraf L, Das P, Champagne SN, Kautsar H, et al. Prioritizing rehabilitation in low- and middle-income country national health systems: a qualitative thematic synthesis and development of a policy framework. Int J Equity Health 2023 May 17;22(1):91 [FREE Full text] [doi: 10.1186/s12939-023-01896-5] [Medline: 37198596]
52. Biesecker C, Zahnd WE, Brandt HM, Adams SA, Eberth JM. A bivariate mapping tutorial for cancer control resource allocation decisions and interventions. Prev Chronic Dis 2020 Jan 02;17:E01 [FREE Full text] [doi: 10.5888/pcd17.190254] [Medline: 31895673]
53. Liu B, Sze J, Li L, Ornstein KA, Taioli E. Bivariate spatial pattern between smoking prevalence and lung cancer screening in US counties. Int J Environ Res Public Health 2020 May 13;17(10):3383 [FREE Full text] [doi: 10.3390/ijerph17103383] [Medline: 32413964]
54. Zahnd WE, Del Vecchio N, Askelson N, Eberth JM, Vanderpool RC, Overholser L, et al. Definition and categorization of rural and assessment of realized access to care. Health Serv Res 2022 Jun 07;57(3):693-702 [FREE Full text] [doi: 10.1111/1475-6773.13951] [Medline: 35146771]

Abbreviations

ACS: American Community Survey
GIS: geographic information system
OT: occupational therapy
PT: physical therapy
RUCA: Rural-Urban Commuting Area

Edited by A Schwartz, M Balcarras; submitted 29.09.25; peer-reviewed by MI Kamali, AC Teodoro; comments to author 24.10.25; accepted 25.11.25; published 23.12.25.

Please cite as:
Ratoza M, Patel RM, Chevan J, Brewer W, Mitchell K
Mapping the Availability of Rehabilitation Providers Using Public Licensure and Population Data for a Geographic Information System–Based Approach to Workforce Planning: Cross-Sectional Feasibility Study
JMIR Form Res 2025;9:e85025
URL: https://formative.jmir.org/2025/1/e85025
doi:10.2196/85025
PMID:

©Madeline Ratoza, Rupal M Patel, Julia Chevan, Wayne Brewer, Katy Mitchell. Originally published in JMIR Formative Research (https://formative.jmir.org), 23.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Estimating the Frequency of Inpatient Adverse Events Using a 2-Step Retrospective Chart Review: Initial Observational Cohort Study

Syed Sabih Ul Hassan[1], MBBS; Fozia Asif[1], MBBS, MBA, MSc; Farwa Ayub[1], MBBS; Ghazal Haque[1], MBBS; Asad Latif[1], MD, MPH

Aga Khan University, Karachi, Sindh, Pakistan

Corresponding Author:
Asad Latif, MD, MPH
Aga Khan University
Stadium Road, P. O. Box 3500
Karachi, Sindh
Pakistan
Phone: 92 21 3493 0051
Email: asad.latif@aku.edu

Abstract

Background: Lower- and middle-income countries bear a disproportionate share of the global burden of adverse events in health care. Despite this, patient safety research is predominantly conducted in high-income countries with well-developed health care systems, resulting in evidence and methodologies that have limited applicability in resource-constrained settings.

Objective: This pilot study primarily aimed to identify the most suitable methodology for a full-scale study to detect inpatient adverse events at a tertiary care hospital in a lower- to middle-income country. Second, we aimed to use our experience with this study to further adapt the selected methodology to our setting.

Methods: This external pilot study used a 2-step retrospective chart review methodology. Two separate screening tools, tools 1A and 1B, were compared to identify which performed better in our setting. We reviewed the medical records of patients who were discharged between January 1, 2019, and December 31, 2019, from a tertiary care hospital in a lower- to middle-income country in South Asia. The main outcome of interest was the rate of adverse events among hospitalized patients reported as the total number of adverse events experienced per 100 admissions.

Results: A total of 100 medical records were screened using tool 1A, with the mean patient age being 39.2 (SD 27.7) years and the mean length of stay being 3.3 (SD 2.8) days. Only 1 adverse event was identified using tool 1A, resulting in an adverse event rate of 1 event per 100 hospital admissions. Tool 1B was also used to screen a total of 100 medical records. The mean patient age was found to be 39.8 (SD 28.4) years, with the mean length of stay being 3.5 (SD 3.4) days. A total of 30 adverse events were identified across 22 patient files, with 18 (60%) considered preventable, resulting in an adverse event rate of 30 events per 100 hospital admissions.

Conclusions: This study demonstrates that tool 1B, adapted from the Global Trigger Tool for Measuring Adverse Events, represents an appropriate and sensitive methodology to identify adverse events among hospitalized patients in a lower- to middle-income country. Furthermore, the findings and experiences of this study were used to improve the design and procedures of our research methodology before implementation in a full-scale study.

International Registered Report Identifier (IRRID): RR2-10.1136/bmjopen-2023-076971

(JMIR Form Res 2025;9:e82512) doi:10.2196/82512

KEYWORDS

patient safety; patient harm; adverse events; inpatients; retrospective studies; low- and middle-income country

XSL•FO
RenderX

Introduction

Unsafe patient care during health care delivery is one of the 10 leading causes of death and disability worldwide [1]. Worldwide, the frequency of adverse events in hospitalized patients has been estimated to range between 2.9% and 21.9%, with approximately half of these events considered to be preventable [2]. The most frequent patient safety events encountered in hospitalized patients include medication errors; health care–associated infections; unsafe surgical procedures, injections, and blood transfusions; venous thromboembolism; and diagnostic and radiation errors [3]. In the United States alone, medical errors result in health care costs of approximately US $20 billion and contribute to more than 250,000 deaths annually [4,5]. It is estimated that the global economic cost of unsafe care exceeds US $1 trillion annually [6]. However, the burden of unsafe care disproportionately affects low- and middle-income countries (LMICs), where it is responsible for 134 million adverse events and 2.6 million deaths annually [7]. In LMICs, poor-quality care now represents a greater barrier to reducing mortality than limited access to care, with approximately 60% of deaths associated with treatable conditions now resulting from deficiencies in the quality of care delivered [8].

Despite this long-standing disparity in burden, the vast majority of patient safety research has originated from high-income countries (HICs), underscoring a broader global trend in which scientific knowledge production is dominated by a small number of high-income nations [9,10]. However, in recent years, global awareness of the need to enhance patient safety standards has resulted in a modest increase in patient safety research in LMICs [11,12]. It is now widely recognized that local researchers are better able to establish health research priorities that are relevant to their populations; design contextually appropriate interventions; and effectively bridge the gap among evidence, policy, and implementation [13,14]. Therefore, a crucial first step in addressing unsafe care in LMICs is to generate locally derived evidence by identifying adverse events within health care settings, thereby determining the true burden and nature of the problem.

Worldwide, a range of methodologies and tools have been developed to detect health care–associated harm [15]. However, keeping resource constraints in LMICs in mind, it is important to use an adverse event detection methodology that is both accurate and feasible. Historically, most studies concerned with adverse events have used various medical record review methodologies that have greater sensitivity than other methods, including voluntary reporting [16,17]. The 2 most popular tools, derived from the Harvard Medical Practice Study and the Global Trigger Tool for Measuring Adverse Events, have been repeatedly validated and compared in studies conducted in HICs [2]. However, it is expected that, for a number of reasons, including rudimentary medical record documentation systems and a lack of available expertise, the results of those studies are not applicable to LMICs [7]. Furthermore, to the best of our knowledge, there are no real-world data comparing the 2 aforementioned tools exclusively from LMICs. Therefore, the primary aim of this external pilot study was to assess which of the 2 methodologies is most suitable for detecting adverse events in a tertiary health care setting in an LMIC. A secondary aim was to draw on the experience gained during the pilot to refine the selected methodology, including improving the data collection tools and workflow processes, in preparation for a larger study that will measure the incidence of adverse events in hospitalized patients.

Methods

Study Design

This pilot study used an observational cohort design where the medical records of patients discharged between January 1, 2019, and December 31, 2019, were randomized and retrospectively reviewed to identify adverse events experienced by inpatients. We piloted 2 separate trigger lists to allow us to select the most appropriate methodology for our larger study. Our published protocol provides more details regarding our larger study [18].

Study Setting

The study site is a 710-bed private, tertiary care teaching hospital that is located in an LMIC in South Asia and provides care to over 50,000 inpatients and 900,000 outpatients annually from all over the country. The hospital is organized into 11 clinical service lines: (1) heart, lung, and vascular; (2) kidney and bladder; (3) mind and brain; (4) cancer care; (5) internal medicine; (6) eye and ear, nose, and throat; (7) gastrointestinal and surgery; (8) anesthesiology, operating room, and central sterile supply department; (9) children's hospital; (10) women's health care; and (11) musculoskeletal and sports medicine.

Sampling Technique

A sample size of 100 patients was used to detect methodological considerations to inform the design of the main study [19]. To ensure that the sample reflected the patient population at the study site, we first determined how many records to draw from each clinical service. This was done by allocating the sample in proportion to the number of admissions that each service received during the study period. For example, a total of 53,307 patients were admitted to the study site in 2019. Of these 53,307 patients, 10,543 (19.8%) were admitted to the internal medicine service. To calculate the number of internal medicine patients included in the total sample, we divided the number of patients admitted to internal medicine (ie, 10,543) by the total number of patients admitted to the study site (ie, 53,307) and multiplied the result by our total sample size (ie, 100). Therefore, 20 patients admitted to the internal medicine service were included in the total sample.

Next, for each service, a list of all admitted patients was prepared and sorted in order of admission date. To select a representative sample from each clinical service, we used systematic sampling. For every service, we calculated a sampling interval by dividing the total number of admissions in that service by the number of records needed from that service (this value is referred to as K). Next, we selected every Kth patient from the ordered lists until the required number of records was reached. For example, to calculate the sampling interval (K) for internal medicine, we divided the total number of admissions to internal medicine (ie, 10,543) by the number

XSL•FO
RenderX

of records needed from internal medicine for the total sample (ie, 20). Therefore, we selected every 527th patient admitted to internal medicine in 2019 to reach a total of 20 patients.

A total of 100 files were reviewed for tool 1A, and these patients are referred to as cohort A. However, 4% (4/100) of the patient files from cohort A were unavailable when tool 1B was applied, so they were replaced with 4 files from the same clinical services associated with those patients; these 100 patients are referred to as cohort B. In total, 104 patient files were reviewed.

Definitions

We define an adverse event as "unintended physical injury resulting from or contributed to by medical care that requires additional monitoring, treatment, or hospitalization, or that results in death" [20].

Data Collection

Overview

Our data collection approach used a 2-stage review process that is recommended by the World Health Organization for collecting data on adverse events using retrospective record reviews [21]. During the first stage, we assessed the medical records against a list of triggers present in the screening tools. The triggers used are specific occurrences, such as clinical outcomes, laboratory tests, and interventions, that are closely associated with adverse events and, therefore, aid reviewers in identifying iatrogenic harm. For this study, we used 2 separate screening tools: tool 1A and tool 1B. Each positive trigger was subsequently checked for its association with an adverse event. If an adverse event was confirmed, we performed a detailed review of the identified event using tool 2, the adverse event assessment tool. The tools were converted from paper-based instruments into electronic templates on the browser-based software REDCap (Research Electronic Data Capture; Vanderbilt University) so that they could be accessed on laptops, computers, and tablets [22].

Data were initially collected by a single reviewer (SH), a medical doctor by training, who had trained on 10 test cases before starting data collection. Each week, the collected data and identified potential adverse events were reviewed by 4 additional medical doctors, including the senior authors (F Asif and AL), who hold postgraduate qualifications and have extensive experience in patient safety and quality improvement. Consensus among all reviewers was required before any incident was designated as an adverse event to minimize bias. Furthermore, this iterative review process promoted consistency and progressive standardization across the study period.

Screening Tools (Tools 1A and 1B)

The screening tools are primarily made up of questions regarding patient characteristics and their corresponding trigger lists. The first trigger list, contained in tool 1A, was adapted from the work by Letaief et al [23]. The trigger list used in tool 1B was adapted from the Global Trigger Tool for Measuring Adverse Events developed by the Institute for Healthcare Improvement [24].

Adverse Event Assessment Tool (Tool 2)

Tool 2 was used to delineate the characteristics of the adverse events identified, including a confidence score for preventability. Tool 2 was also adapted from the work by Letaief et al [23]. A single patient file may contain evidence of multiple adverse events, and every adverse event had a corresponding tool 2 applied.

Data Sources

The hospital uses physical patient files as well as electronic databases for information such as medication orders and laboratory test results to maintain detailed patient records. We used the physical files and all available electronic databases to conduct our study.

Data Analysis

The main outcome measure was the rate of adverse events in the inpatient setting. The rate was reported as the number of adverse events per 100 hospital admissions. Furthermore, we reported the frequency of the different types of adverse events identified. Events with preventability scores above 3 (out of a maximum of 6) were classified as preventable. Data analysis was carried out using Stata (version 14; StataCorp).

Ethical Considerations

This study was approved by the Ethics Review Committee at Aga Khan University (2023-6324-24566), covering secondary analysis of existing data without need for additional consent. We generated unique identifiers against which all findings were documented, and every effort was made to maintain the anonymity of the patients whose files were reviewed. Data were stored on encrypted, institution-approved cloud services that restricted access to only the investigating team.

Patient and Public Involvement

Patients or the public were not involved in the design, conduct, reporting, or dissemination plans of this research.

Reporting Guideline

We used the Strengthening the Reporting of Observational Studies in Epidemiology reporting guidelines to draft and edit this manuscript. The checklist has been included in Multimedia Appendix 1 [25].

Results

The mean ages of patients in cohort A and cohort B were 39.2 (SD 27.7) and 39.8 (SD 28.4) years, respectively. Male individuals constituted 45% (45/100) of cohort A and 46% (46/100) of cohort B. The mean length of hospital stay for cohort A was 3.3 (SD 2.8) days compared to 3.5 (SD 3.4) days for cohort B. In cohort A, 36% (36/100) of the patients were admitted electively, whereas 39% (39/100) were admitted acutely. Conversely, in cohort B, 33% (33/100) of the patients were admitted electively, and 38% (38/100) were admitted acutely. Patient characteristics are summarized in Table 1.

XSL•FO
RenderX

Table 1. Comparison of demographics and tool performance (N=100).

	Tool 1A	Tool 1B
Patient characteristics		
Age (y), mean (SD)	39.2 (27.7)	39.8 (28.4)
Sex, n (%)		
Female	55 (55)	54 (54)
Male	45 (45)	46 (46)
Hospital stay (d), mean (SD)	3.3 (2.8)	3.5 (3.4)
Admission status, n (%)		
Elective	36 (36)	33 (33)
Acute	39 (39)	38 (38)
Direct[a]	21 (21)	22 (22)
Did not know	4 (4)	7 (7)
Tool performance		
Time taken per file (min), median (IQR)	19 (13-27)	29 (19-57)
Total number of positive triggers	114	260
Number of patients with positive triggers	75	87
Total number of adverse events	1	30
Number of patients with adverse events	1	22

[a]The patient presented to an outpatient clinic and was directly admitted after the consultation.

A median time of 19 (IQR 13-27) minutes was required to review each file using tool 1A. A total of 114 positive triggers were found, resulting in the identification of 1 adverse event, which translates to an adverse event rate of 1 event per 100 hospital admissions. For tool 1B, a median time of 29 (IQR 19-57) minutes was required per file, and a total of 260 positive triggers were identified along with 30 adverse events, with 22% (22/100) of the patients experiencing at least one adverse event. Of these 30 events, 18 (60%) were considered preventable. The performance of both tools is compared in Table 1.

The most commonly identified trigger using tool 1A was >1 admission within 30 days, which was positive in a total of 27% (27/100) of the patients. Difference between diagnosis at admission and diagnosis at discharge was positive in 25% (25/100) of the patients, and National Early Warning Score 2 of ≥3 or Pediatric Early Warning Score of ≥2 on the day of discharge was positive in 21% (21/100) of the patients, representing the second and third most commonly identified triggers, respectively. The most frequently identified triggers in tool 1A are listed in Table 2.

XSL•FO
RenderX

Table 2. Most common positive triggers identified in cohort A using tool 1A (N=100).a

Trigger	Patients, n (%)
>1 admission within 30 days	27 (27.0)
Difference between diagnosis at admission and diagnosis at discharge	25 (25.0)
NEWS2[b] of ≥3 or PEWS[c] of ≥2 on the day of discharge (n=89)	21 (23.6)
>1 visit to the operating room within the last 6 months	13 (13.0)
Cardiac or respiratory arrest; low Apgar score	9 (9.0)
Injury or complications related to the termination of pregnancy or labor and delivery, including neonatal complications (n=32)	3 (9.4)
LAMA[d]	3 (3.0)
Unplanned transfer from general care to intensive care or higher dependency (n=84)	3 (3.6)
Development of neurological deficit not present on admission	2 (2.0)
Patient care delayed or lesser treatment given because the patient was unable to pay	2 (2.0)

[a]Some triggers were not applicable to all patients (eg, triggers for surgical care were not applicable to patients who did not undergo surgery).

[b]NEWS2: National Early Warning Score 2.

[c]PEWS: Pediatric Early Warning Score.

[d]LAMA: leave against medical advice.

Antiemetic administration, which occurred with 60% (60/100) of the patients, was the most frequently identified trigger using tool 1B, followed by transfusion of blood products, which was found in 19% (19/100) of the patients. The next most common trigger was time in the emergency department greater than 6 hours, which was found to be positive for a total of 18% (18/100) of the patients. The most commonly identified triggers using tool 1B are listed in Table 3.

Table 3. Most common positive triggers identified in cohort B using tool 1B (N=100).a

Trigger	Patients, n (%)
Antiemetic administration	60 (60.0)
Transfusion of blood products	19 (19.0)
Time in ED[b] of >6 hours (n=39)	18 (46.2)
Oxytocic agent administration (n=16)	13 (81.3)
Administration of glucagon or glucose of ≥10% (n=20)	13 (65.0)
Vitamin K administration	13 (13.0)
Readmission within 30 days	12 (12.0)
Admission to intensive care postoperatively	11 (11.0)
Code, arrest, or rapid response call generated	11 (11.0)
Decrease of >25% in hemoglobin or hematocrits	9 (9.0)

[a]Some triggers were not applicable to all patients (eg, triggers for surgical care were not applicable to patients who did not undergo surgery).

[b]ED: emergency department.

Of a total of 30 events identified, surgical- or procedure-related complications were the most frequently identified type of adverse event, making up 40% (12/30) of all events identified. In this category, the most common event was surgery- or procedure-related hemorrhage, which was identified in 6% (6/100) of the patients and comprised 20% (6/30) of all events. Failure of diagnosis or treatment was the second most frequent category, making up 26.7% (8/30) of all events identified, followed by adverse medication events, which comprised 23.3% (7/30) of all events identified. The least common type of adverse event was health care–associated infections, only being identified in 1% (1/100) of the patients. Both tools 1A and 1B identified the health care–associated infection. The adverse events identified are summarized in Table 4.

XSL•FO
RenderX

Table 4. Categorization of all adverse events identified (N=30).

Category	Events, n (%)
Surgery- or procedure-related complications	12 (40.0)
Hemorrhage	6 (20.0)
CNS[a] complications	3 (10.0)
Iatrogenic injury	2 (6.7)
Postoperative atelectasis	1 (3.3)
Failure of diagnosis or treatment	8 (26.7)
Delay in diagnosis or treatment	5 (16.7)
Retinopathy of prematurity	2 (6.7)
Ventilator-associated injury	2 (6.7)
Failed extubation	1 (3.3)
Adverse medication event	7 (23.3)
Bleeding events	4 (13.3)
Drug-induced parkinsonism	1 (3.3)
Transfusion reaction	1 (3.3)
Oversedation	1 (3.3)
Health care–associated infection	1 (3.3)[b]

[a]CNS: central nervous system.

[b]This adverse event was detected by both tool 1A and tool 1B.

Discussion

This pilot study aimed to identify an appropriate methodology that could be used to accurately estimate the rate at which adverse events occur in hospitalized patients in LMICs. Using tool 1A, we obtained an adverse event rate of 1 event per 100 hospital admissions. However, using tool 1B, we obtained an adverse event rate of 30 events per 100 admissions, with a total of 22% (22/100) of the patients experiencing at least one adverse event. The most common type of adverse events identified were surgical- or procedure-related complications, with surgery- or procedure-related hemorrhage being the most frequently identified adverse event.

There is a lack of authoritative work regarding the incidence of adverse events in hospitalized patients in LMICs. However, the National Academies of Sciences, Engineering, and Medicine have estimated that 134 million adverse events occur across 531 million hospital admissions annually in LMICs, resulting in an adverse event rate of approximately 25 events per 100 hospital admissions [7]. However, the first methodology used, tool 1A, obtained an adverse event rate of 1 event per 100 hospital admissions. This is also in contrast to the work by Letaief et al [23], from which tool 1A was adapted. Their study found that, of 620 patient records reviewed, 62 (10%) contained evidence of at least one adverse event [23]. This suggests that tool 1A might be an inappropriate choice for estimating the rate of adverse events in our setting. We believe that this could be because the trigger list was originally developed and validated in HICs and, therefore, has limited applicability in our context. Differences in documentation practices, health care processes, and clinical workflows between HICs and LMICs may reduce the sensitivity and relevance of certain triggers. For example, variations in record-keeping systems, diagnostic resources available, and patterns of care delivery may lead to underdetection or misclassification of adverse events when using tools designed for high-resource settings [26].

The second methodology used, tool 1B, identified a total of 30 adverse events across 100 hospital admissions, which aligns more closely with regional estimates and, therefore, may be more suitable for use in our setting [7]. Moreover, Bates et al [27] conducted a similar study across 11 Boston hospitals using a trigger list that was also adapted from the Global Trigger Tool for Measuring Adverse Events. In their weighted sample of 2809 admissions, they identified at least one adverse event in 23.6% of the admissions, which is in line with our finding of 22% (22/100). This demonstrates that the tool may be applicable across diverse health care settings and is potentially suitable for future benchmarking efforts. It is worth noting that Bates et al [27] identified adverse medication events as the most common type of adverse event, accounting for 39% of all events detected. In contrast, surgical- or procedure-related complications were the most frequently identified type of adverse event in our study, comprising 40% (12/30) of all events recorded. We believe that this might reflect the difference between the quality of surgical care provided in LMICs compared with HICs [28]. Lack of access to timely surgical intervention, low surgical case volumes, and a lack of specialists are often cited as the main contributors to unsafe surgical care in LMICs [28].

An important consideration with tool 1B is that it generates a substantial number of positive triggers, many of which do not ultimately correspond to true adverse events yet still necessitate

XSL•FO
RenderX

a detailed review of medical records. However, this broad approach may be preferable within nascent patient safety systems, especially in LMICs. By casting a wider net, tool 1B reduces the risk of missing relevant events and allows organizations to assess and prioritize identified events based on relevance and impact. Additionally, analyzing and addressing these triggers through a system-based approach may promote wider organizational improvements, yielding safety benefits and fostering resilience that may not be possible by relying solely on reactive approaches to safety issues [29].

We would like to acknowledge the salient limitations of this methodology. Retrospective chart reviews are limited by the fact that they only use data that were previously collected during the delivery of routine care and, therefore, cannot capture the full range of harm experienced by patients. For example, any adverse events that manifest after discharge but do not result in further patient encounters will be missed. Moreover, this issue is further exacerbated by poor documentation culture and practices in LMICs, which result in missing information, difficulty in interpreting information (such as due to the use of nonapproved abbreviations), and questionable authenticity of the information recorded. However, despite these limitations, retrospective chart reviews remain the most appropriate choice for our setting as they are relatively inexpensive; use existing data; and allow us to capture events that manifest after short latency periods, such as postoperative complications. Finally, we would like to acknowledge that tools 1A and 1B were not applied to the same set of medical records. However, as no adverse events were detected in the records involved, any potential impact on the findings is expected to be minimal.

In conclusion, this pilot study enabled us to identify and select well-defined, validated tools for use in our larger study, which will allow for reliable measurement of the incidence of adverse events among hospitalized patients. The methodological and operational lessons learned will directly inform our larger study and future capacity-building efforts in data quality, training, and surveillance systems [30]. Furthermore, by building local capacity for adverse event detection, we contribute to broader efforts to strengthen health systems by identifying priority areas for quality improvement, guiding resource allocation, and informing the design of context-appropriate interventions to reduce preventable harm. At the policy level, evidence generated through such locally led initiatives can support the development of national patient safety strategies, establish frameworks for data-driven decision-making, and integrate patient safety indicators into routine health system performance monitoring.

Acknowledgments

All authors declared that they had insufficient or no funding to support open access publication of this manuscript, including from affiliated organizations or institutions, funding agencies, or other organizations. JMIR Publications provided APF support for the publication of this article.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this paper.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

All persons listed as authors met the International Committee of Medical Journal Editors authorship criteria. All authors made significant contributions and revisions to the manuscript and provided complete assent for publication. AL and F Asif were responsible for the conceptualization of the study and for supervising all activities. SUHS, GH, and F Ayub made major contributions toward designing the methodology, conducting the study, and analyzing the results. Each author is responsible for the content and has read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
STROBE checklist.
[PDF File (Adobe PDF File), 274 KB - formative_v9i1e82512_app1.pdf]

References

1. Global patient safety action plan 2021-2030: towards eliminating avoidable harm in health care. World Health Organization. 2021. URL: https://apps.who.int/iris/handle/10665/343477 [accessed 2023-04-28]

XSL•FO
RenderX

2. Schwendimann R, Blatter C, Dhaini S, Simon M, Ausserhofer D. The occurrence, types, consequences and preventability of in-hospital adverse events - a scoping review. BMC Health Serv Res 2018 Jul 04;18(1):521 [FREE Full text] [doi: 10.1186/s12913-018-3335-z] [Medline: 29973258]
3. Dhingra-Kumar N, Brusaferro S, Arnoldo L. Patient safety in the world. In: Donaldson L, Ricciardi W, Sheridan S, Tartaglia R, editors. Textbook of Patient Safety and Clinical Risk Management. Cham, Switzerland: Springer; Dec 15, 2020:93-98.
4. Rodziewicz TL, Houseman B, Vaqar S, Hipskind JE. Medical error reduction and prevention. In: StatPearls. Treasure Island, FL: StatPearls Publishing; 2024.
5. Makary MA, Daniel M. Medical error-the third leading cause of death in the US. BMJ 2016 May 03;353:i2139. [doi: 10.1136/bmj.i2139] [Medline: 27143499]
6. Flott K, Fontana G, Darzi A. The global state of patient safety. Imperial College London. 2019. URL: https://www.imperial.ac.uk/media/imperial-college/institute-of-global-health-innovation/GlobalStateofPS_DIGITAL_16Sep19[2].pdf [accessed 2025-10-24]
7. National Academies of Sciences, Engineering, and Medicine, Health and Medicine Division, Board on Health Care Services, Board on Global Health, Committee on Improving the Quality of Health Care Globally. Crossing the Global Quality Chasm: Improving Health Care Worldwide. Washington, DC: National Academies Press; Aug 28, 2018.
8. Kruk ME, Gage AD, Arsenault C, Jordan K, Leslie HH, Roder-DeWan S, et al. High-quality health systems in the Sustainable Development Goals era: time for a revolution. Lancet Glob Health 2018 Nov;6(11):e1196-e1252 [FREE Full text] [doi: 10.1016/S2214-109X(18)30386-3] [Medline: 30196093]
9. Woods WA, Watson M, Ranaweera S, Tajuria G, Sumathipala A. Under-representation of low and middle income countries (LMIC) in the research literature: ethical issues arising from a survey of five leading medical journals: have the trends changed? Glob Public Health 2023 Jan;18(1):2229890 [FREE Full text] [doi: 10.1080/17441692.2023.2229890] [Medline: 37401751]
10. Schreiber M, Klingelhöfer D, Groneberg DA, Brüggmann D. Patient safety: the landscape of the global research output and gender distribution. BMJ Open 2016 Feb 12;6(2):e008322 [FREE Full text] [doi: 10.1136/bmjopen-2015-008322] [Medline: 26873042]
11. Elmontsri M, Banarsee R, Majeed A. Improving patient safety in developing countries - moving towards an integrated approach. JRSM Open 2018 Nov 12;9(11):2054270418786112 [FREE Full text] [doi: 10.1177/2054270418786112] [Medline: 30455962]
12. Al-Worafi Y. Patient Safety in Developing Countries: Education, Research, Case Studies. Boca Raton, FL: CRC Press; 2023.
13. Beran D, Byass P, Gbakima A, Kahn K, Sankoh O, Tollman S, et al. Research capacity building-obligations for global health partners. Lancet Glob Health 2017 Jun;5(6):e567-e568 [FREE Full text] [doi: 10.1016/S2214-109X(17)30180-8] [Medline: 28495256]
14. Malekzadeh A, Michels K, Wolfman C, Anand N, Sturke R. Strengthening research capacity in LMICs to address the global NCD burden. Glob Health Action 2020 Dec 31;13(1):1846904 [FREE Full text] [doi: 10.1080/16549716.2020.1846904] [Medline: 33373280]
15. Murff HJ, Patel VL, Hripcsak G, Bates DW. Detecting adverse events for patient safety research: a review of current methodologies. J Biomed Inform 2003;36(1-2):131-143 [FREE Full text] [doi: 10.1016/j.jbi.2003.08.003] [Medline: 14552854]
16. Classen DC, Resar R, Griffin F, Federico F, Frankel T, Kimmel N, et al. 'Global trigger tool' shows that adverse events in hospitals may be ten times greater than previously measured. Health Aff (Millwood) 2011 Apr;30(4):581-589. [doi: 10.1377/hlthaff.2011.0190] [Medline: 21471476]
17. Gandhi TK, Seger DL, Bates DW. Identifying drug safety issues: from research to practice. Int J Qual Health Care 2000 Feb;12(1):69-76. [doi: 10.1093/intqhc/12.1.69] [Medline: 10733086]
18. Ul Hassan SS, Asif F, Haque G, Ayub F, Munir T, Khan F, et al. Estimating the frequency of inpatient adverse events using a two-step retrospective chart review: a study protocol. BMJ Open 2023 Dec 06;13(12):e076971 [FREE Full text] [doi: 10.1136/bmjopen-2023-076971] [Medline: 38056938]
19. Kunselman AR. A brief overview of pilot studies and their sample size justification. Fertil Steril 2024 Jun;121(6):899-901 [FREE Full text] [doi: 10.1016/j.fertnstert.2024.01.040] [Medline: 38331310]
20. Adverse events, near misses, and errors. Agency for Healthcare Research and Quality, US Department of Health and Human Services, Rockville, MD. 2019. URL: https://psnet.ahrq.gov/primer/adverse-events-near-misses-and-errors [accessed 2025-12-18]
21. Assessing and tackling patient harm: a methodological guide for data-poor hospitals. World Health Organization. 2010 Sep 08. URL: https://www.who.int/publications/i/item/9789241500388 [accessed 2023-05-03]
22. Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O'Neal L, REDCap Consortium. The REDCap consortium: building an international community of software platform partners. J Biomed Inform 2019 Jul;95:103208 [FREE Full text] [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]
23. Letaief M, El Mhamdi S, El-Asady R, Siddiqi S, Abdullatif A. Adverse events in a Tunisian hospital: results of a retrospective cohort study. Int J Qual Health Care 2010 Oct;22(5):380-385. [doi: 10.1093/intqhc/mzq040] [Medline: 20685729]

XSL•FO
RenderX

24. Griffin FA, Resar RK. IHI global trigger tool for measuring adverse events (second edition). Institute for Healthcare Improvement. 2009. URL: https://www.ihi.org/library/white-papers/ihi-global-trigger-tool-measuring-adverse-events [accessed 2023-04-20]
25. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Ann Intern Med 2007 Oct 16;147(8):573-577 [FREE Full text] [doi: 10.7326/0003-4819-147-8-200710160-00010] [Medline: 17938396]
26. Reidpath DD, Allotey P. The problem of 'trickle-down science' from the Global North to the Global South. BMJ Glob Health 2019 Jul 24;4(4):e001719 [FREE Full text] [doi: 10.1136/bmjgh-2019-001719] [Medline: 31406597]
27. Bates DW, Levine DM, Salmasian H, Syrowatka A, Shahian DM, Lipsitz S, et al. The safety of inpatient health care. N Engl J Med 2023 Jan 12;388(2):142-153. [doi: 10.1056/NEJMsa2206117] [Medline: 36630622]
28. Meara JG, Leather AJ, Hagander L, Alkire BC, Alonso N, Ameh EA, et al. Global Surgery 2030: evidence and solutions for achieving health, welfare, and economic development. Lancet 2015 Aug 08;386(9993):569-624. [doi: 10.1016/S0140-6736(15)60160-X] [Medline: 25924834]
29. Mossburg SE, Weaver SJ, Pillari M, Daugherty Biddison E. Manifestations of high-reliability principles on hospital units with varying safety profiles: a qualitative analysis. J Nurs Care Qual 2019;34(3):230-235 [FREE Full text] [doi: 10.1097/NCQ.0000000000000368] [Medline: 30480611]
30. Hassan ZA, Schattner P, Mazza D. Doing a pilot study: why is it essential? Malays Fam Physician 2006 Aug 31;1(2-3):70-73 [FREE Full text] [Medline: 27570591]

Abbreviations

HIC: high-income country
LMIC: low- or middle-income country
REDCap: Research Electronic Data Capture
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by A Mavragani; submitted 16.08.25; peer-reviewed by M Al Zoubi, MW Perroud Jr; comments to author 08.10.25; revised version received 06.12.25; accepted 12.12.25; published 29.12.25.

Please cite as:
Ul Hassan SS, Asif F, Ayub F, Haque G, Latif A
Estimating the Frequency of Inpatient Adverse Events Using a 2-Step Retrospective Chart Review: Initial Observational Cohort Study
JMIR Form Res 2025;9:e82512
URL: https://formative.jmir.org/2025/1/e82512
doi:10.2196/82512
PMID:

©Syed Sabih Ul Hassan, Fozia Asif, Farwa Ayub, Ghazal Haque, Asad Latif. Originally published in JMIR Formative Research (https://formative.jmir.org), 29.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Misinformation About Medical Cannabis in YouTube Videos: Systematic Review

Shivank Khare, BSc (Hons); Simon Erridge, BSc (Hons). MBBS; Swathikan Chidambaram, BSc (Hons), MBBS, MRCS; Mikael Hans Sodergren, MBChB, DIC, PhD

Department of Surgery and Cancer, Academic Surgical Unit, Imperial College London, 10th Floor QEQM, St Mary's Hospital, South Wharf Road, London, United Kingdom

Corresponding Author:
Mikael Hans Sodergren, MBChB, DIC, PhD
Department of Surgery and Cancer, Academic Surgical Unit, Imperial College London, 10th Floor QEQM, St Mary's Hospital, South Wharf Road, London, United Kingdom

Abstract

Background: YouTube has become a major source of health information, with 2.5 billion monthly users. Despite efforts taken to promote reliable sources, misinformation remains prevalent, particularly regarding medical cannabis.

Objective: This study aims to evaluate the quality and reliability of medical cannabis information on YouTube and to examine the relationship between video popularity and content quality.

Methods: A systematic review of YouTube videos on medical cannabis was conducted. Search terms were selected based on Google Trends, and 800 videos were retrieved on July 8, 2024. After applying exclusion criteria, 516 videos were analyzed. Videos were categorized by content creators: (1) nonmedical educational channels, (2) medical education channels, and (3) independent users. Two independent reviewers (SK and SE) assessed content quality using the DISCERN grade and the Health on the Net (HON) code. Statistical analysis included one-way ANOVA and Pearson correlation coefficient.

Results: Of the 516 videos analyzed, 48.5% (n=251) were from the United States, and 17.2% (n=89) from the United Kingdom. Only 12.2% (n=63) were produced by medical education channels, while 84.3% (n=435) were by independent users. The total views reached 119 million, with nonmedical educational channels having the highest median views with 274,957 (IQR 2161-546,887) and medical education channels having the lowest median views at 5721 (IQR 2263 - 20,792.50). The mean DISCERN and HON code scores for all videos were 34.63 (SD 9.49) and 3.93 (SD 1.20), respectively. Nonmedical educational creators had the highest DISCERN score (mean 47.78, SD 10.40) and independent users had the lowest score (mean 33.5, SD 8.50; $P<.001$). Similarly, nonmedical educational creators had the highest HON code score (mean 5.33, SD 1.22), while independent users had the lowest (mean 3.78, SD 1.10; $P=.007$). Weak positive correlations were found between video views and DISCERN scores ($r=0.34$, $P<.001$) and likes and DISCERN scores ($r=0.30$, $P<.001$).

Conclusions: YouTube is a key source of information on medical cannabis, but the credibility of videos varies widely. Independent users attract the highest viewers but have reduced reliability according to the DISCERN and HON scores. Educational channels, despite increased reliability received the least engagement. The weak correlation between views and content quality emphasizes the need for content moderation to ensure that the most reliable and accurate information on health issues is widely disseminated. Future research should identify strategies to promote verified sources of information and limit misinformation.

(JMIR Form Res 2025;9:e76723) doi:10.2196/76723

KEYWORDS

cannabis; consumer health information; health literacy; health education; health promotion

Introduction

The therapeutic use of cannabis has undergone a dramatic transformation in public perception and legal status over the past decade. As such, cannabis is now one of the most consumed drugs. According to the 2014 National Survey on Drug Use and Health, cannabis was the most used illicit substance in the United States, with 13.2% of individuals aged ≥12 years reporting cannabis use in the past year and 8.4% in the past month [1]. Similar trends are also seen in Europe, Asia, and Australasia [2-4]. This shift reflects both evolving legislative frameworks and growing public acceptance of medical cannabis [5]. However, the rapid normalization of medical cannabis has outpaced the generation of high-quality evidence on its efficacy and safety. Current randomized controlled trials are highly heterogenous, limited by small sample sizes and use short follow-up assessments [6]. Consequently, there is a paucity of clinical guidance and education available to inform clinicians and their patients about medical cannabis. Within this context,

XSL•FO
RenderX

YouTube—as the world's second-most popular social media network [7]—has become a key platform for distributing health-related information and shaping patient understanding of this class of medications.

The democratization of health information through social media platforms presents unique challenges for medical professionals and public health authorities. Unlike peer-reviewed literature or guidance from clinical associations, YouTube's algorithms prioritize engagement metrics over accuracy, potentially amplifying sensationalist or commercially motivated content. A study of COVID-19 misinformation on YouTube highlighted that misleading information was least prevalent in professional or government videos; however, these had lower views and engagement [8]. In response to misinformation propagated during the COVID-19 pandemic, YouTube has implemented community guidelines to help mitigate medical misinformation. These guidelines classify misinformation into three categories: disease prevention, medical treatments, and denial of specific health conditions or their consequences in accordance with the World Health Organization (WHO) [9]. Despite these regulatory efforts, misinformation remains prevalent on YouTube, indicating an ongoing challenge in producing accurate and reliable medical information.

Medical cannabis has emerged as a topic of considerable interest on YouTube, fueled by growing research into its potential therapeutic applications for various conditions, including pain, anxiety, and other chronic illnesses [10]. Misinformation about medical cannabis can be detrimental to public perception. Studies indicate that news articles and blogs often exhibit a strong positive bias toward the efficacy of cannabinoids in pain management [11]. This may foster unrealistic treatment expectations based on online information. In addition, analysis from the International Cannabis Policy Study highlights a reduced perception of some potential risks associated with cannabis [11]. This raises concerns about potential harm to individuals who may struggle to distinguish between evidence-based data and unverified claims, widening the gap between perceptions and the ground truth.

A prior analysis of YouTube videos highlighted that, out of 66 videos on cannabis for chronic pain, there was a wide range of video quality and reliability. The videos of the highest quality were most likely to be produced by physicians. However, due to the limited sample size and focus on chronic pain, this analysis limits insights into engagement with misleading or harmful videos. Consequently, there is an ongoing paucity of information on the quality and reliability of information on cannabis and health on YouTube and other social media platforms [12]. This review therefore provides a timely assessment of the quality of medical cannabis information available on YouTube. The primary aim is to assess the objective quality of the videos with the highest reach on the platform. In addition, the study aims to investigate the relationship between video popularity and the accuracy and quality of the information presented.

Methods

A systematic review was performed of the content published on YouTube (Alphabet Inc) on medical cannabis. The conduct and reporting of this study were performed in line with Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) [13].

Ethical Considerations

This study involved a retrospective review of publicly accessible YouTube videos and did not involve human participants or identifiable private information. As such, formal ethical approval was not required. This was confirmed by the Imperial College London ethics panel. All data used in the study were publicly available and did not contain any personally identifiable information. Videos were analyzed anonymously, and no attempts were made to contact content creators or link content to specific individuals.

Search Strategy and Data Extraction

To ensure relevancy of the search strategy for medical cannabis, key terms were derived from an assessment of worldwide Google Trends (Google LLC) data from the 12 months preceding July 8, 2024, to ensure the highest coverage of cannabis-based medicinal products. These included 'medical cannabis,' 'medical marijuana,' 'medical weed' and 'prescription cannabis.' These search terms were then entered into the United Kingdom version of YouTube (Alphabet Inc) on July 8, 2024, using incognito mode to prevent third-party cookies and browsing history from affecting the results.

The top 200 videos of each search term were noted using Microsoft Excel spreadsheet with the key attributes: title, number of views, country of origin, duration, age, channel name, number of likes, and number of comments. Exclusion criteria included videos shorter than one minute, older than five years (based on upload date), duplicate videos, not in English, or unrelated to cannabis-based medicinal products, such as those not addressing their biochemistry, effects or social topics.

Data extraction was performed by two independent reviewers (SK and SE), with disagreements planned to be resolved by a third author (SC). Data extraction included video views, likes, channel, and country of origin. The videos were subsequently categorized into the following three categories based on the channel types and purpose: educational channels organized by nonmedical professionals (eg, TED or TEDx Talks); education channels organized by medical professionals (eg, Mayo Clinic or Curaleaf Clinic); and independent nonmedical users and groups (eg, independent content creators).

Outcome Measures

The primary outcome was the quality and reliability of the video content. This was determined using the DISCERN Grade and Health on the Net (HON) code. These tools have been used previously to assess health care information based on social media influence [14]. The DISCERN Grade consists of 16 questions, with each question scored on a scale of 1 to 5 points, therefore making a total of 80 points. It is divided into three sections, with the first section evaluating the reliability of the

XSL•FO
RenderX

paper, the second assessing the quality of information on treatment choices and the final section examining the overall quality of the publication [15]. All sections of the DISCERN grade are included in the sum score. The HON code is used to ensure health and medical websites maintain honesty and serve as reliable sources of information backed up by evidence. There are eight key principles, and each video is scored on a scale from zero to eight, where zero indicates noncompliance, and eight represents full compliance [16]. The videos were assessed independently by SK and SE. Any discrepancies were planned to be resolved through discussion with a senior author (SC). Secondary outcomes included assessment of video reach and engagement through video views and likes.

Statistical Analysis

The YouTube engagement metrics were summarized using median (IQR), while DISCERN grade scores, and HON code scores were summarized using mean (SD). Statistical analysis was performed using PRISM (version 10; GraphPad software). One-way ANOVA was conducted to compare DISCERN and HON scores between different content creators assuming unequal variances.

Pairwise comparisons were conducted using Tukey test to identify specific differences between groups. The Tukey method was applied for multiple comparisons to control for Type I error and was conducted when a statistically significant finding was present on ANOVA. Pearson correlation coefficient was used to determine the association between the DISCERN Grade and the number of views. Statistical significance was defined as P value <.05.

Results

The YouTube platform was searched on July 8, 2024, with the four search phrases yielding 800 videos in total. Two hundred and fifty-seven videos were duplicates and were subsequently removed. A further 27 videos did not meet the inclusion criteria. This led to 516 videos being included in the final analysis (Multimedia Appendix 1). The video search and data extraction can be depicted in the illustration in Figure 1.

Figure 1. Flow chart showing the video selection process for information pertaining to medical cannabis including the exclusion criteria and video duplicates.

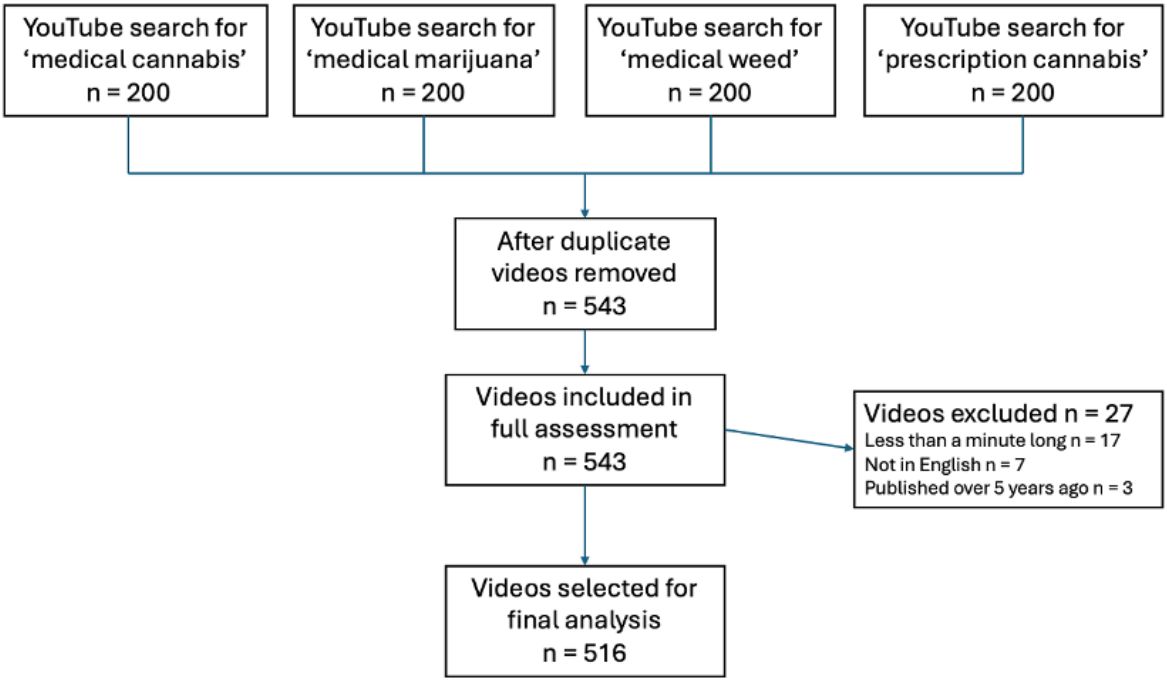

Most videos originated from the United States of America (n=251, 48.65%), followed by Europe (n=94, 18.22%). Seventeen (3.29%) videos had more than a million views. Only 12.2% (n=63) of videos were produced by medical professionals, with most videos produced by independent organizations (n=435; 84.3%). The important baseline characteristics are mentioned below in Table 1, including the country of origin, channel type, and engagement metrics.

XSL•FO
RenderX

Table . Baseline characteristics of YouTube videos (N=516).

Characteristic	YouTube videos (N=516)
Country of Origin, n (%)	
Australia	33 (6.4)
Canada	16 (3.1)
Denmark	2 (0.4)
Germany	3 (0.6)
India	1 (0.2)
Israel	1 (0.2)
Korea	1 (0.2)
Malaysia	1 (0.2)
New Zealand	9 (1.7)
Norway	1 (0.2)
Philippines	4 (0.8)
Qatar	1 (0.2)
Singapore	1 (0.2)
South Africa	5 (1.0)
Sri Lanka	1 (0.2)
United Kingdom	89 (17.2)
United States of America	251 (48.5)
Unknown	96 (18.6)
Channel types, n (%)	
Educational (Nonmedical)	18 (3.5)
Educational (Medical)	63 (12.2)
Independent users	435 (84.3)
Engagement metrics, median (IQR)	
Views	9,605 (4,347-39,029)
Views per day	12 (4-45)
Likes	169 (67-590)

Table 2 presents the engagement metrics and quality assessment indicators for YouTube videos addressing medical cannabis misinformation, with a combined total of 119,155,949 views since upload. Among the three video categories, independent users contributed most of the total views (91,660,510), followed by educational (nonmedical) creators with 23,026,930 views, and educational (medical) sources with 4,468,509 views.

XSL•FO
RenderX

Table . Engagement metrics, adherence to the HON code of conduct, and DISCERN scores for YouTube videos on medical cannabis misinformation..

Parameter	Educational (Nonmedical) (n=18)	Educational (Medical) (n=63)	Independent users (n=435)	*P* value[a]
Engagement metrics				
Total views, n	23,026,930	4,468,509	91,650,510	__[b]
Views per video, median (IQR)	274,957 (2,161-546,887)	5,721 (2,263–20,793)	10,447 (4,625-38,557)	.001
Likes, median (IQR)	6,063 (28–9,060)	133.5 (49-338)	174 (69-594)	<.001
Quality assessment indicators, mean (SD)				
HON[c] adherence total score	5.22 (1.22)	4.59 (1.43)	3.78 (1.10)	.007
DISCERN total score	47.78 (10.40)	38.59 (12.73)	33.50 (8.50)	<.001

[a]*P* values are calculated using one-way ANOVA.

[b]Not applicable.

[c]HON: Health on the Net foundation.

When examining views per video, educational (nonmedical) creators had the highest median view number with 274,957 (IQR 2161-546,887) views, substantially outperforming both independent users with 10,447 (IQR 4625-38,557) views and educational (medical) creators with 5721 views (IQR 2263-20,793) views. Post hoc pairwise comparisons revealed that the differences between independent and nonmedical sources (*P*<.001), as well as between medical and nonmedical sources (*P*<.001), were statistically significant. There was no statistically significant difference between independent and medical sources (*P*=.67).

A similar trend was observed for likes, where educational (nonmedical) creators garnered the most engagement with 6063 (IQR 28–9060) likes, followed by independent users with 174 (IQR 69-594) likes, and lastly, educational (medical) creators with 134 (49-338) likes. Tukey multiple comparisons test confirmed that independent sources significantly differed from nonmedical sources (*P*<.001), and medical sources also significantly differed from nonmedical sources (*P*<.001). However, there was no difference between independent and medical creators (*P*=.83).

In terms of quality assessment, educational (nonmedical) videos demonstrated the highest standards, with HON adherence scores averaging 5.22 (SD 1.22) and DISCERN scores at 47.78 (SD 10.40). In comparison, independent users produced content with the lowest quality scores (HON: mean 3.78, SD 1.10; DISCERN: mean 33.50, SD 8.50), while educational (medical) sources occupied a middle ground (HON: mean 4.59, SD 1.43; DISCERN: mean 38.59, SD 12.73). Post hoc testing showed significant differences in DISCERN scores across all groups, including independent versus medical (*P*<.001), independent versus nonmedical (*P*<.001), and medical versus nonmedical sources (*P*<.001). Regarding HON adherence, educational (nonmedical) sources outperformed both independent users (*P*<.001) and medical creators (*P*<.001), though the difference between medical and independent sources was not statistically significant (*P*=.07).

The mean DISCERN and HON code scores across all videos were 34.63, SD 9.49 and 3.93, SD 1.20, respectively. Pearson correlation coefficient analysis revealed a low positive correlation (*r*=0.34; *P*<.001) between YouTube video views and DISCERN scores and a low positive correlation of (*r*=0.30; *P*<.001) between likes and DISCERN scores (Figure 2).

Figure 2. The relationship between engagement metrics and DISCERN grade score. (A) A scatter plot with a positive trend line indicating a weak correlation (0.34) between an increased number of views and DISCERN grade scores. (B) Scatter plot with a positive trend line with a weak correlation (0.30) between increased likes and DISCERN grade scores. Correlations were assessed using Pearson correlation coefficient.

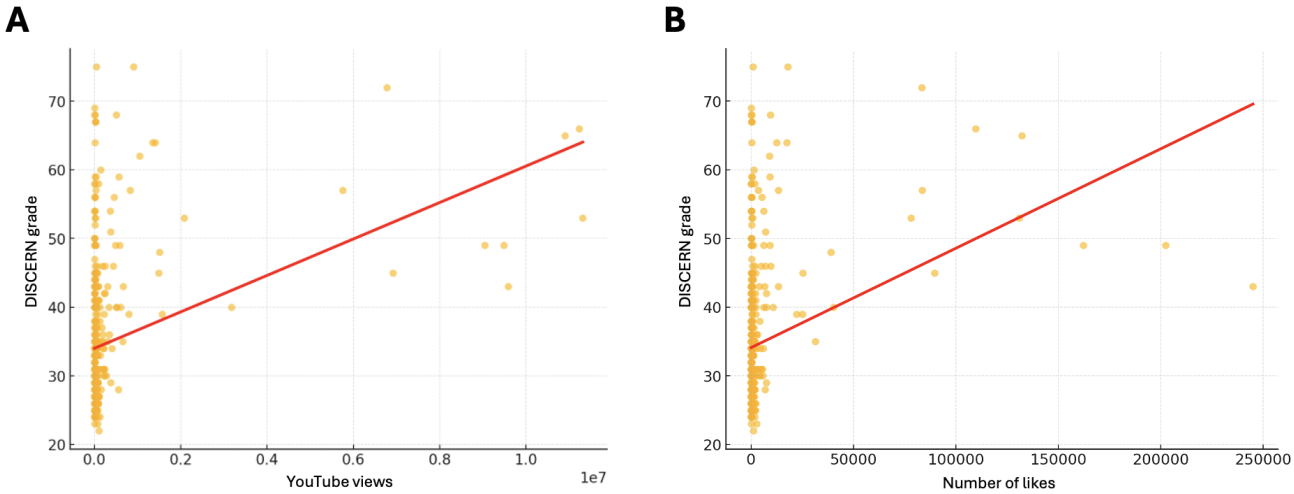

XSL•FO
RenderX

Discussion

Principal Findings

This study aimed to evaluate the quality of information presented in YouTube videos about medical cannabis and assess whether popularity correlates with content reliability. The findings reveal a notable disconnect between viewer engagement and video quality, with independent users gathering the highest views but offering the least reliable content, as indicated by lower DISCERN and HON scores. These results emphasize the persistent challenge of misinformation on digital platforms, even in an era of increasing regulation.

The overall mean DISCERN score across all videos was 34.63, which, on the 80-point scale, suggests generally poor reliability. In comparison, other studies examining health topics such as varicoceles and testicular torsion exhibited similar concerns [17,18]. The analysis by Hong et al [18] expressed frustration as from the nonduplicated videos, an overwhelming majority were expressed to be educational but were rated as poor on the DISCERN score. A comparable pattern was observed in the present analysis with only 12.2% of the videos originating from medical or expert sources, with the overall mean DISCERN score. This suggests that misinformation about medical cannabis is not an isolated issue but part of a broader trend of unreliable health communication online.

A noticeable finding in this study was videos from medical education sources were less popular and had lower quality and reliability scores in comparison to nonmedical educational sources. This is a challenge also seen in other areas of health content on YouTube [19]. One reason could be medical sources often use technical language and longer video formats suited for academic purposes but less engaging to the public. In contrast, nonmedical creators may adopt language and entertaining visuals, which enhances the relatability and shareability for lay audiences. Moreover, while the quality and reliability of medical videos was higher than independent creators, it was lower than that of the nonmedical educational content. Many of the nonmedical sources were from news organizations; therefore, the introduction of clear editorial policies for medical channels may help further enhance their quality. While independent creators received lower DISCERN and HON code scores, they had higher engagement. Evidence demonstrates that subjective language and content that falls under entertainment categories are more likely to receive a higher number of views and likes [20]. In addition, posting outside of business hours is more likely to receive higher engagement. Independent creators are less bound by professional standards of objectivity and are also more likely to focus on changing posting frequency and timings to maximize video engagement. YouTube has introduced initiatives to try and help improve the reach of content produced by accredited health care professionals. It is also important that this content continues to be as engaging and meets viewers at their level for effective dissemination of health information on this particular social media platform [21].

Furthermore, a study on COVID-19 vaccine videos found that approximately 11% of the most-viewed videos contradicted the reference standards from the WHO or the Centers for Disease Control and Prevention. These videos had significantly lower modified DISCERN (mDISCERN) scores compared to factual ones, yet they garnered far more likes and engagement [22]. Similarly, this analysis identified several of the most viewed videos were of cannabis content published by nonexpert independent users with low reliability.

The weak positive correlation between engagement metrics (views and likes) and DISCERN scores (r=0.34 and r=0.30, respectively) further confirms that popular videos are not necessarily trustworthy. Moreover, visual analysis of correlation plots highlights that the r value is likely impacted by several outlier videos. This is consistent with research by Stimpson and Ortega [23], which found that a large proportion of social media users (82%) perceive false or misleading health information to be prevalent on these platforms, with 67% admitting they are unable to assess the accuracy of the information they encounter [23]. This increases the risk of misinformed health decisions by patients and the general public, but may also increase anxiety among viewers [24]. Videos that over illustrate the therapeutic benefits of cannabis while downplaying risks, such as adverse effects or drug interactions, may lead individuals to experiment with unproven treatments without medical guidance. For example, some cancer patients believe cannabis can treat their cancer directly, despite a lack of scientific evidence, potentially leading to worse outcomes if they prioritize unproven cannabis use over chemotherapy or immunotherapy [25].

YouTube enforces strict policies on cannabis-related content, prohibiting videos that display or promote the consumption of cannabis and typically permitting educational, documentary, or news-focused discussions if they do not encourage use or provide instructions for consumption [26]. While YouTube's approach to misinformation prohibits health claims that contradict recognized authorities and removes content with serious risk of harm, the platform does not systematically check creators for medical credentials or require source acknowledgment, resulting in only partial alignment with the more stringent transparency and authority requirements of the HON code.

To combat the spread of misinformation, the WHO is working with the Council of Medical Specialty Societies and the National Academy of Medicine on Phase II of a project on identifying credible sources of health information on social media [27]. The project aims to provide social media platforms with standardized criteria to distinguish reliable health information, addressing challenges like inconsistent vetting processes and the proliferation of unverified claims. However, systemic issues like algorithmic bias and regulatory inertia continue to remain unresolved.

Several limitations of this study should be acknowledged. First, the video selection was restricted to English-language content, which limits the generalisability of the findings to non-English speaking populations and other geographic regions. Additionally, while the DISCERN tool provides a structured framework for assessing information quality, it allows for subjective interpretation, particularly when assigning partial scores between 2 and 4. This introduces some variability in how

XSL•FO
RenderX

each criterion is evaluated. The study also did not consider channel-level data or account characteristics, which may have implications for both the reach and perceived credibility of the videos hosted on that channel. Lastly, this study focused solely on YouTube and did not examine other social media platforms such as Instagram, TikTok, or Facebook, which are also social media sites where individuals may be exposed to unverified health information.

Conclusion

In conclusion, the quality of information about medical cannabis among the most viewed videos on YouTube remains poor. This mirrors similar patterns observed during the COVID-19 pandemic and in studies of other health topics. This study contributes to a growing body of evidence suggesting that user engagement does not strongly correlate with informational quality, posing ongoing challenges for public health communication. In the future, it is important YouTube and other social media sites implement policy changes to mitigate the spread of misinformation in all fields, particularly health, and to regulate the content by health care professionals more frequently. Medical professionals should also be encouraged to engage more actively on social platforms and implement strategies to make their content engaging as well as informative to compete with low-quality content.

Acknowledgments

We used the generative AI tool ChatGPT by OpenAI to assist with grammar and spelling checks during the preparation of this manuscript [28]. No generative artificial intelligence tools were used for content generation, data analysis, or interpretation of results. No external financial support or grants were received from any public, commercial, or not for-profit entities for the research, authorship, or publication of this article.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files. The data that support the findings of this study are also available in the public domain on YouTube. These data were derived from publicly accessible videos retrieved using predefined search terms. A full list of video URLs and the associated extracted metadata (eg, views, likes, upload date, channel category) is provided in the supplementary file submitted with this manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
516 YouTube videos included in the final analysis.
[XLSX File, 78 KB - formative_v9i1e76723_app1.xlsx]

Checklist 1
PRISMA checklist.
[PDF File, 109 KB - formative_v9i1e76723_app2.pdf]

References

1. Azofeifa A, Mattson ME, Schauer G, McAfee T, Grant A, Lyerla R. National estimates of marijuana use and related indicators - National Survey on Drug Use and Health, United States, 2002-2014. MMWR Surveill Summ 2016 Sep 2;65(11):1-28. [doi: 10.15585/mmwr.ss6511a1] [Medline: 27584586]
2. Mills L, Arnold JC, Suraev A, et al. Medical cannabis use in Australia seven years after legalisation: findings from the online Cannabis as Medicine Survey 2022–2023 (CAMS-22). Harm Reduct J ;21(1). [doi: 10.1186/s12954-024-00992-1]
3. Manthey J, Freeman TP, Kilian C, López-Pelayo H, Rehm J. Public health monitoring of cannabis use in Europe: prevalence of use, cannabis potency, and treatment rates. Lancet Reg Health Eur 2021 Nov;10:100227. [doi: 10.1016/j.lanepe.2021.100227] [Medline: 34806072]
4. Hasin D, Walsh C. Trends over time in adult cannabis use: a review of recent findings. Curr Opin Psychol 2021 Apr;38:80-85. [doi: 10.1016/j.copsyc.2021.03.005] [Medline: 33873044]
5. Mills L, Arnold JC, Suraev A, et al. Medical cannabis use in Australia seven years after legalisation: findings from the online Cannabis as Medicine Survey 2022-2023 (CAMS-22). Harm Reduct J 2024 May 28;21(1):104. [doi: 10.1186/s12954-024-00992-1] [Medline: 38807133]
6. Banerjee R, Erridge S, Salazar O, et al. Real world evidence in medical cannabis research. Ther Innov Regul Sci 2022 Jan;56(1):8-14. [doi: 10.1007/s43441-021-00346-0] [Medline: 34748204]
7. 23 essential YouTube statistics you need to know in 2025. Social Shepherd. URL: https://thesocialshepherd.com/blog/youtube-statistics [accessed 2025-02-15]

XSL•FO
RenderX

8. Li HOY, Bailey A, Huynh D, Chan J. YouTube as a source of information on COVID-19: a pandemic of misinformation? BMJ Glob Health 2020 May;5(5):e002604. [doi: 10.1136/bmjgh-2020-002604] [Medline: 32409327]
9. A long term vision for YouTube's medical misinformation policies. YouTube Official Blog. URL: https://blog.youtube/inside-youtube/a-long-term-vision-for-medical-misinformation-policies/ [accessed 2025-02-15]
10. Therapeutic effects of cannabis and cannabinoids. In: The Health Effects of Cannabis and Cannabinoids: The Current State of Evidence and Recommendations for Research: National Academies Press (US); 2017. URL: https://www.ncbi.nlm.nih.gov/books/NBK425767/ [accessed 2025-09-23]
11. Gedin F, Blomé S, Pontén M, et al. Placebo response and media attention in randomized clinical trials assessing cannabis-based therapies for pain: a systematic review and meta-analysis. JAMA Netw Open 2022 Nov 1;5(11):e2243848. [doi: 10.1001/jamanetworkopen.2022.43848] [Medline: 36441553]
12. Etumuse B, Greer M, Onyemachi J, et al. Medical misinformation and quality of public video content on cannabis for chronic pain management: a cross-sectional analysis of the YouTube platform. J Pain Res 2024;17:3577-3586. [doi: 10.2147/JPR.S479200] [Medline: 39526076]
13. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
14. Noll D, Green L, Asokan G, et al. Is YouTubeTM a good source of information for patients to understand laparoscopic fundoplication? Ann R Coll Surg Engl 2023 Apr;105(4):365-371. [doi: 10.1308/rcsann.2022.0039] [Medline: 36374324]
15. DISCERN. Oxford Population Health. URL: https://www.ndph.ox.ac.uk/research/research-groups/applied-health-research-unit-ahru/discern/discern [accessed 2025-04-07]
16. Health On the Net Foundation Code of Conduct for medical and health web sites. URL: https://www.mfd.lv/HONcode_en.htm [accessed 2025-04-07]
17. Bai G, Pan X, Zhao T, Chen X, Liu G, Fu W. Quality assessment of YouTube videos as an information source for testicular torsion. Front Public Health 2022;10:905609. [doi: 10.3389/fpubh.2022.905609] [Medline: 35664123]
18. Hong HS, Lang JJ, Damodaran S, Sindhwani P. Assessing information on YouTube™ as a quality source for the treatment of varicoceles. Indian J Urol 2021;37(4):339-344. [doi: 10.4103/iju.iju_201_21] [Medline: 34759526]
19. Chidambaram S, Maheswaran Y, Chan C, et al. Misinformation about the human gut microbiome in YouTube videos: cross-sectional study. JMIR Form Res 2022 May 16;6(5):e37546. [doi: 10.2196/37546] [Medline: 35576578]
20. Munaro AC, Hübner Barcelos R, Francisco Maffezzolli EC, Santos Rodrigues JP, Cabrera Paraiso E. To engage or not engage? The features of video content on YouTube affecting digital consumer engagement. J of Consumer Behaviour 2021 Sep;20(5):1336-1352. [doi: 10.1002/cb.1939]
21. Apply to be a source in YouTube health features. YouTube Help. URL: https://support.google.com/youtube/answer/12796915?hl=en-GB [accessed 2025-08-16]
22. Li HOY, Pastukhova E, Brandts-Longtin O, Tan MG, Kirchhof MG. YouTube as a source of misinformation on COVID-19 vaccination: a systematic analysis. BMJ Glob Health 2022 Mar;7(3):e008334. [doi: 10.1136/bmjgh-2021-008334] [Medline: 35264318]
23. Stimpson JP, Ortega AN. Social media users' perceptions about health mis- and disinformation on social media. Health Aff Sch 2023 Oct;1(4):qxad050. [doi: 10.1093/haschl/qxad050] [Medline: 38107206]
24. Pan W, Liu D, Fang J. An examination of factors contributing to the acceptance of online health misinformation. Front Psychol 2021;12:630268. [doi: 10.3389/fpsyg.2021.630268] [Medline: 33732192]
25. As more people with cancer use medical cannabis, oncologists face questions they struggle to answer. National Cancer Institute. URL: https://www.cancer.gov/news-events/cancer-currents-blog/2024/medical-cannabis-oncologists-patients-communication [accessed 2025-04-07]
26. Illegal or regulated goods or services policies. YouTube Help. URL: https://support.google.com/youtube/answer/9229611?hl=en-GB [accessed 2025-08-16]
27. Phase II - global principles for identifying credible sources of health information on social media. World Health Organization. URL: https://www.who.int/teams/digital-health-and-innovation/digital-channels/phase-ii--global-principles-for-identifying-credible-sources-of-health-information-on-social-media [accessed 2025-04-07]
28. ChatGPT. Open AI. URL: https://chatgpt.com/ [accessed 2025-09-23]

Abbreviations

HON: Health on the Net
PRISMA-ScR: Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews
WHO: World Health Organization

XSL•FO
RenderX

Edited by A Mavragani; submitted 29.04.25; peer-reviewed by S Gunther, Y Cui; revised version received 08.09.25; accepted 09.09.25; published 06.10.25.

Please cite as:
Khare S, Erridge S, Chidambaram S, Sodergren MH
Misinformation About Medical Cannabis in YouTube Videos: Systematic Review
JMIR Form Res 2025;9:e76723
URL: https://formative.jmir.org/2025/1/e76723
doi:10.2196/76723

©Shivank Khare, Simon Erridge, Swathikan Chidambaram, Mikael Hans Sodergren. Originally published in JMIR Formative Research (https://formative.jmir.org), 6.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Perceived Risk and Fashion on the Intention to Adopt Wireless Earbuds in the United States Using a Partial Least Squares-Structural Equation Modeling Approach: Empirical Study

Edward M Lee[1*], DBA; Chandrasekar Subramaniam[2*], PhD; Sungjune Park[2*], PhD

[1]Bank of America, Charlotte, NC, United States

[2]Department of Business Information Systems and Operations Management, University of North Carolina Charlotte, 9201 University City Blvd, Charlotte, NC, United States

[*]all authors contributed equally

Corresponding Author:
Chandrasekar Subramaniam, PhD
Department of Business Information Systems and Operations Management, University of North Carolina Charlotte, 9201 University City Blvd, Charlotte, NC, United States

Abstract

Background: The number of studies on the use of smart wearables has increased dramatically in recent years. However, aspects including personal safety and fashion perspectives of wearable devices have not yet been adequately addressed in the literature. There have been debates regarding the potential health risks and fashionability of using wearable devices. Regardless of the actual impact of such devices, these aspects may influence users' perceptions toward the purchase and use of wearable technology.

Objective: This paper addresses the following research question: How do perceptions of risk and fashion affect the user's intention to purchase and use wireless earbuds?

Methods: A survey was administered to assess perceptions on health and privacy risks, fashionability, and wearable comfort of wireless earbuds, alongside questions on behavioral intention regarding their purchase and use. All questions were adapted from prior research and measured using a 7-point Likert scale. The final sample of 205 responses was analyzed using the partial least squares method with Smart-PLS software.

Results: Perceived health risk (P=.015), perceived fashionability (P<.001), and wearable comfort (P=.007) had a significant impact on a consumer's intention to purchase wireless earbuds. Privacy risk did not have a significant impact on intention to purchase. Intention to purchase had a significant impact on intention to use (P<.001).

Conclusions: As new types of emerging technology are introduced to the market, technology acceptance models should evolve to better understand consumers' perceptions toward these new technologies, from both academic and practical points of view.

(JMIR Form Res 2025;9:e56887) doi:10.2196/56887

KEYWORDS

wireless earbuds; health risk; wearable comfort; fashionability; intention to buy

Introduction

Background

Wearable technology is no longer unfamiliar to consumers. According to the International Data Corporation, the worldwide market for wearable devices grew by approximately 20% in 2021, with around 530 million units sold in the third quarter of 2021. Of these units, 339 million were earwear. According to Niknejad et al [1], the publication of studies relating to smart wearables has also increased dramatically in recent years. Less than 50 studies on wearable devices were published in 2013, rising to 404 in 2022 [2]. These studies covered topics ranging from technical issues, user behavior, design, security, and privacy to the issue of social acceptability. However, personal safety perspectives of wearable devices have not been adequately addressed in the literature to date.

Debates continue regarding the health risks of wearable devices. For instance, some researchers have raised concerns about wireless earbuds due to their high levels of radio-frequency radiation, while others have contested this risk [3]. Other researchers have suggested that powerful batteries close to the human body for extended periods could cause leukemia, but this has also been contested [4]. Regardless of the actual health risks associated with wearable devices, these debates can influence and shape users' perceptions toward wearable technologies [5]. It is important to understand how perceived health risks impact consumers' perceptions of wearable devices.

XSL•FO
RenderX

A unique feature of wearable devices is that they are not only considered technical devices but also as fashion items. Fashion-related factors affect consumers in their evaluation of wearable technologies [6]. Wireless earbuds are available in different colors, and smartwatches have options such as changeable straps for consumer design preference. The roles of perceived risk and fashion in a consumer's choice to purchase and use wearable devices have not been studied in the literature.

This paper addresses the following research question: *Do perceptions of technology, risk, and fashionability affect the user's intention to purchase wireless earbuds?* This study will fill the gap in the literature of a systematic understanding of the impact of perceived risk and fashionability on wearable device purchase and use. The results of our study will also help the development of safer wearable devices. The remainder of the paper is organized as follows. The next section reviews the relevant literature and presents the research gaps. The section "Research Model and Hypotheses" presents our research model and hypotheses, while "Method" describes our research method and results. The research findings are then discussed, and we conclude by presenting the study's limitations and future research opportunities.

Literature Review

The underlying theory for our study is the Unified Theory of the Adoption and Use of Technology 2 (UTAUT 2). The UTAUT 2 theory is complemented by the literature on the perceived risks and fashionability of wearable devices to build our model. This section synthesizes the current literature in the above 3 research streams.

Wearable devices or technologies are mobile technologies that are integrated into clothing and accessories, incorporating wireless connectivity for access, interaction, and the exchange of information anytime and anywhere [7]. Wearable technology is used in various industries. For example, end consumers wear wireless earbuds to listen to music or answer phone calls, while the health care industry uses wearable trackers for patient and disease management. This study will focus on the consumers' perspectives and their behavioral intention to purchase and use wireless earbuds, a type of wearable device.

The underlying theory, UTAUT 2, was developed to understand technology acceptance in a consumer use context [8]. UTAUT 2 extends the original UTAUT model and incorporates three additional constructs: hedonic motivation, price value, and habit. Hedonic motivation is "the fun or pleasure derived from using a technology" and plays a critical role in determining technology acceptance and use [9]. Price value is "the consumers' cognitive tradeoff between the perceived benefits of the applications and the monetary cost for using them" [10]. Habit is "the extent to which people tend to perform behaviors automatically because of learning" [11].

Literature on wearable device use has covered themes including user behavior and aspects of the wearables themselves, such as technological features, design, security and privacy issues, and social acceptability [1]. For example, Grym et al [12] studied the feasibility of smart wristband wearable use among first-time pregnant women and found that the device usage was similar during the second and third trimesters but decreased during the postpartum period. The authors identified various reasons for the decreased usage, including problems with charging the devices, a perception of the devices as uncomfortable, and fear of scratching their babies with the devices.

Previous research on user behavior has examined users' willingness relating to different aspects, such as their intention to purchase, adopt, and use technologies, alongside their actual use and user experiences [1]. For example, Li et al [13] found that among adults older than 60 years, perceived usefulness, compatibility, facilitating conditions, and self-reported health status positively affected their intention to use smart wearable technologies. Nunes and Arruda Filho [14] examined users' Google Glass adoption and identified three categories of users: socially satisfied, socially constrained, and early adopters. The early adopters were subcategorized as enthusiasts and visionaries. Dehghani and Kim [15] proposed three key factors, such as screen size, uniqueness, and design, that affected participants' current and potential future purchase intention of smartwatches. From these studies, it is clear that to study the purchase intention of wireless earbuds, we must complement the foundation of UTAUT 2 with additional variables relevant to the use context of wearables.

The health risks associated with wearable devices are highly debated in the literature. One researcher raised concerns about the placement of wireless earbuds in the ear canal, which exposes tissues in the head to relatively high levels of radio-frequency radiation, eventually increasing the chance of cancer [3]. However, another researcher, who has studied the effects of wireless radiation on human health, wrote that there is no health risk from wearable devices and that the arguments about adverse health effects of wearables have no credibility [3]. Researchers have also raised concerns about having powerful batteries close to the human body for an extended period, as being too close to power lines over long times may cause leukemia [4]. In 2015, a group of 250 scientists filed a petition to the World Health Organization (WHO) and United Nations (UN) regarding the health risks of electromagnetic field radio waves generated by wireless devices [3]. However, Heid points out that the WHO and other public health organizations have not found "any clear evidence for health hazards at exposure levels below international limits" [16].

Regardless of the actual health risks of wearable devices, these debates on adverse health effects could influence and shape users' perceptions of the technology [5]. Attitudes are formed by an individual's belief regarding the consequences of using a particular technology, whether those consequences are perceived or real [5]. The information available about a particular technology affects the individual's perceptions and impacts their intention to use that technology [5]. Thus, perceptions regarding health risks can affect consumer behavior, regardless of the actual level of health risk [17].

Due to the COVID-19 pandemic, consumers may be more sensitive to the health risks than before, regardless of the actual health risks of products or technology. For instance, a rumor spread on social media that the pandemic was caused by 5th generation wireless network technology [18]; this was refuted

XSL•FO
RenderX

by the United States Federal Emergency Management Agency [18]. As consumers become more sensitive to health risks, the perception of these risks must be studied in the wearable technology. This is an aspect lacking in the literature on wearable technology [1].

The Bluetooth technology, used to power almost all wireless earbuds, expands the attack surface and is subject to security and privacy risks [19]. The privacy attack types include passive eavesdropping and active eavesdropping. When wireless earbuds are used for activities beyond listening to music, such as controlling devices using voice commands, they are vulnerable to compromising the listener's security and privacy. Researchers have shown that perceived privacy risks have a negative impact on the value perception of wearable devices and on the intention to purchase and use these devices [20].

Perceived fashionability, the perception of the design component of a product, has been known to affect consumer purchase and use of a product [21]. Wearable accessories can strongly impact the user's physical appearance [6]. Therefore, how consumers view the design of wearable technology devices can impact their intention of using the device. When Apple announced their first wireless headphones called AirPods, people disliked the aesthetic and shared jokes on social media, comparing their design to a toothbrush head [22]. Meanwhile, wearable comfort refers to the "consumers' overall subjective assessment of the physical feeling from wearing" a device [6]. Comfortable wearable devices increase usage enjoyment. This is related to users' expectations of functionalities for certain apparel, such as raincoats [6].

Despite the jokes circulating on the Internet regarding the design of Apple's AirPods, competitors such as smartphone manufacturers have launched their own wireless earbuds. Wireless earbuds are now more sophisticated, with additional features to control other devices, and are likely to increase in popularity among consumers. Therefore, studying consumers' fashion perceptions of wireless earbuds would provide additional insights into their purchase intention.

Research Model and Hypotheses

In this section, we present our research model on consumers' intention to purchase and use wireless earbuds, incorporating antecendents from UTAUT 2, perceived risk and perceived fashion. The conceptual model is shown in Figure 1. Below we present our hypotheses.

Figure 1. Conceptual model.

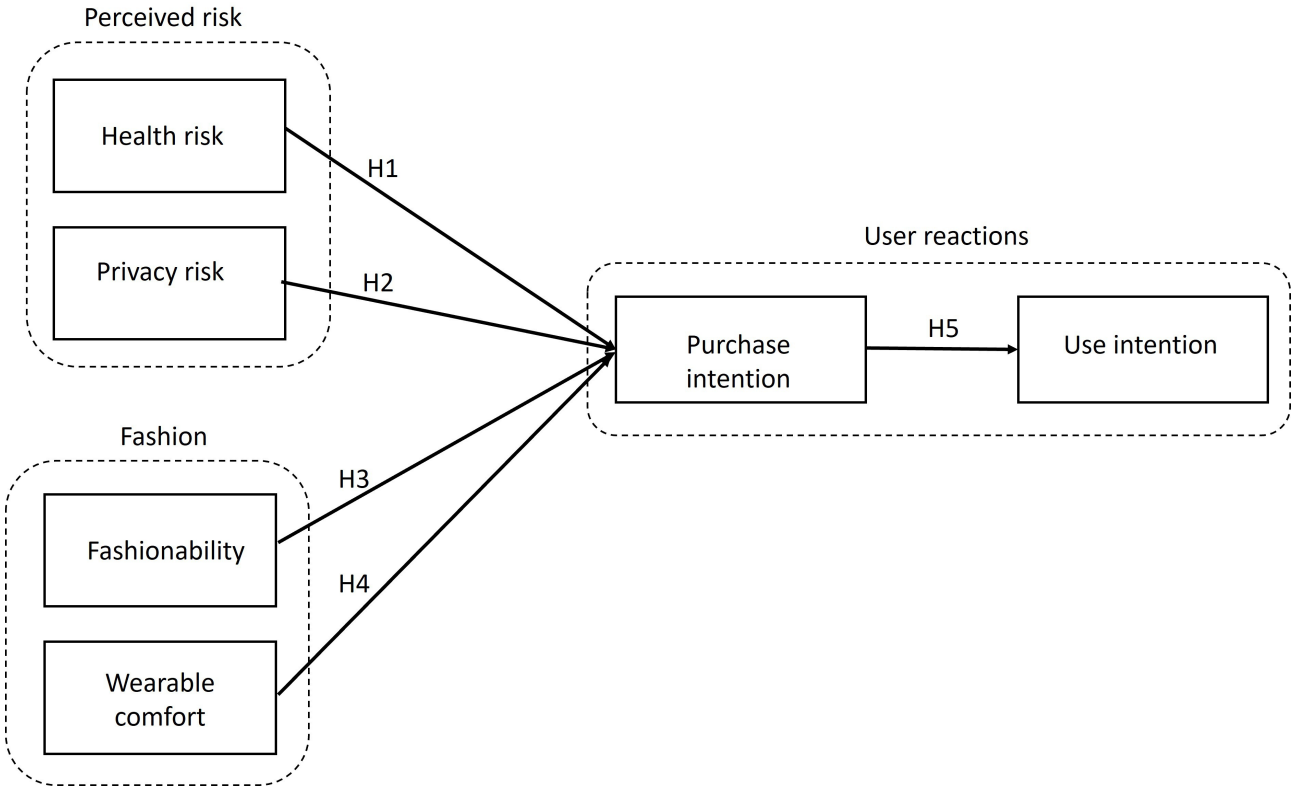

Despite ongoing debates on the health risks of wearable devices, there is no consensus regarding the actual health effects of using wireless earbuds. While some researchers have expressed concerns about the extended use of wireless technology electronics and rechargeable batteries, others claim a lack of evidence to support these health risks, positing wearable devices as safe for use [3,4,16]. Regardless, such debates on health risks potentially influence consumer perceptions of wearable technology. Consumers are more likely to purchase wireless earbuds if they believe them to be safe. In contrast, consumers are less likely to purchase wireless earbuds if they think long-term use could be harmful to their health due to exposure to wireless radiation. Hence, our first hypothesis is as follows:

H1: Perceived Health Risk Has a Negative Influence on the Behavioral Intention to Purchase Wireless Earbuds

Perceived privacy risk is another factor that could influence consumers' attitudes towards wearable technology [23].

XSL•FO
RenderX

Although wireless earbuds generally only play music, they are connected to a main device, such as a smartphone, through Bluetooth wireless technology. This means third parties might be able to gain access to a smartphone by hacking connected wireless earbuds [19,24]. Wireless earbuds may also allow eavesdropping on sensitive conversations of a consumer and compromise their privacy [19]. Regardless of the actual possibility of such hacking, a consumer's sensitivity regarding privacy protection could affect how they view wearable devices. If a consumer is concerned with personal information being collected by companies or being exposed to hackers, they may prefer to use wired headsets rather than wireless earbuds. Therefore, we hypothesize that:

H2: Perceived Privacy Risk Has a Negative Influence on the Behavioral Intention to Purchase Wireless Earbuds

Wearable technologies are not only evaluated based on their functional quality but also on their perceived fashionability and wearable comfort. Herz and Rauschnabel [6] identified three broad categories of fashion-related factors that explain why people chose particular apparel: perceived fashionability, wearable comfort, and functional quality.

Perceived fashionability is the perception of the design component of a product and has been known to impact consumer behavior [21]. Consumers may choose wearable technology devices based on how they fit with their outfits, and the ownership of wearable accessories can strongly impact one's physical appearance [6]. Therefore, the design and color of wearable devices can impact a consumer's intention to use them. For example, the design of Apple's AirPods was initially criticized on social media, and this could have negatively affected their sales. The study hypothesizes that:

H3: Perceived Fashionability Has a Positive Influence on the Behavioral Intention to Purchase Wireless Earbuds

Wearable comfort is the overall subjective assessment of the physical feeling from wearing a device [6]. Since wireless earbuds are typically worn for extended periods, comfort is a critical factor in their evaluation. Consumers may be less likely to use wearable technology devices that are uncomfortable or hinder their daily activities. For example, a consumer would not wear wireless earbuds if they made their ears uncomfortable during jogging or exercising. Therefore, the study hypothesizes that:

H4: Wearable Comfort Has a Positive Influence on the Behavioral Intention to Purchase Wireless Earbuds

According to Ajzen [22] , intentions are presumed to be an indicator of to what extent people are willing to perform a certain behavior. Consumers who intend to adopt wireless earbuds will consider purchasing them. Consumers who already own wireless earbuds may consider repurchasing more for different purposes. For example, a consumer may consider ordering a second pair of wireless earbuds so that one set can be used at work and the other can be used while exercising. There could also be cases of consumers who already own wireless earbuds who consider buying next-generation devices as an upgrade. Following the intention to purchase, consumers will consider how frequently they would use wireless earbuds. Therefore, it is hypothesized that behavioral intention to purchase drives use intentions.

H5: Behavioral Intention to Purchase Wireless Earbuds Has a Significant Influence on Behavioral Intention to Use Wireless Earbuds

In summary, the main model includes perceived health risk and privacy risk to understand the impact of perceived risk on consumers' attitudes toward wireless earbuds. In addition, fashionability and wearable comfort from Herz and Rauschnabel [6] were added to incorporate consumers' perceptions from an aesthetic view. Figure 1 is the conceptual model of this study. The definitions of the model constructs are provided in Table 1.

Table . Definition of perception constructs.

Construct	Study	Definition
Health risk	Jacoby and Kaplan [25]	"The risk to the buyer's or other's safety in using products."
Privacy risk	Featherman and Pavlou [26]	"Potential loss of control over personal information, such as when information about you is used without your knowledge or permission."
Fashionability	Homburg et al [21]	"The perception of the design component."
Wearable comfort	Herz and Rauschnabel [6]	"Consumers' overall subjective assessment of the physical feeling from wearing."

In addition to the main variables for the constructs, three control variables were included in this study: age, experience, and awareness of health risks. Age was controlled as there is a higher chance of younger consumers having better knowledge of wearable technology, impacting their reactions toward the technology. Consumers' reactions toward wireless earbuds would most likely change according to whether they have tried or own earbuds; thus, experience was controlled for. Reactions toward wireless earbuds could be different between consumers who are aware or not aware of potential health risks; thus, awareness of wireless earbuds' potential health risks was also controlled.

XSL•FO
RenderX

Methods

Instrument Development

Each scale used was adapted from previous research. The survey items are included in the Multimedia Appendix 1. The scales for perceived health risk, perceived privacy risk, perceived fashionability, wearable comfort, and behavioral intention to purchase are adapted from Herz and Rauschnabel [6]. Each survey item was measured using a 7-point Likert scale ranging from 1 (“strongly disagree”) to 7 (“strongly agree”). Age was measured in years and experience was coded using a 0 or 1 dummy variable, for which 0 represents “does not own wireless earbuds” and 1 represents “owns wireless earbuds”. Awareness of health risks was coded using a 0 or 1 dummy variable, for which 0 represents “not aware” and 1 represents “aware” of the potential health risks of wireless earbuds. For additional insights, optional questions included the duration of ownership and feedback relating to the purchase and use of wireless earbuds. Duration of ownership was measured in months. Except for duration of ownership and feedback, participants were required to answer every question to prevent missing values.

Data Collection

The survey included a short description of wireless earbuds and their main functionalities for participants who are not familiar with the devices. Based on the response to the control question regarding experience (whether or not the participant owns wireless earbuds), participants were assigned different question sets; these were mostly identical but used slightly different terms applicable to the participant’s situation. The quality of the survey questions was reviewed and improved, based on responses to a pilot study conducted with 23 participants. For example, in the initial survey, there was confusion about the meaning of “re-purchasing wireless earbuds” for those who had already owned a pair; thus, a brief description was added (“additional pair for different occasions or upgrading to next model”).

The finalized survey was distributed to panelists through Qualtrics, an American experience management company. To gather a diverse and informative sample, we used a purposive sampling strategy. Specifically, 2 distinct groups were targeted, consumers who currently own wireless earbuds and consumers who do not. This deliberate selection allowed the collection of data from both segments of the market to gain a comprehensive understanding of consumer intentions. Participants were all located in the United States and were 18 years old or above. A total of 304 participants accessed the online survey, and 276 of them completed it; however, 71 completed surveys were removed due to incomplete answers. Thus, a total of 205 completed survey datasets were analyzed. The sample size of 205 meets the suggested minimum threshold of “ten times the largest number of structural paths directed at a particular latent construct in the structural model” [24].

Measurement and Structural Analysis

Partial least squares (PLS) analysis was used to test the model, as PLS is capable of testing the effects of several interaction terms [27]. Using Smart-PLS software (version 3.3.3), the measurement model and structural model were examined. The measurement model, also called the outer model, describes the relationships between the constructs and indicator variables [28]. The structural model, also called the inner model, represents the constructs and describes the relationships between them [28].

The reliability and validity of the measurement model were evaluated by assessing internal consistency reliability, indicator reliability, convergent validity, and discriminant validity . Cronbach alpha, a traditional criterion for internal consistency, provides an estimate of reliability based on the intercorrelations of the observed indicator variables. For reliability to be acceptable, Cronbach alpha for internal consistency reliability should be higher than 0.7. Indicator reliability is determined by the size of the outer loading. For indicator reliability, the standardized indicator loadings should be 0.7 or higher.

Convergent validity is the degree to which a measure correlates positively with alternative measures of the same construct. The average variance extracted (AVE), a common measure to establish convergent validity on the construct level, is defined as “the grand mean value of the squared loadings of the indicators associated with the construct.” For convergent validity, AVE should be 0.50 or higher, meaning the construct explains more than half of the variance of its indicators on average. Discriminant validity is the degree to which a construct is truly distinct from other constructs by empirical standards. There are two measures of discriminant validity: cross-loadings and the Fornell–Larcker criterion. Cross-loadings describe the correlation between an indicator and other constructs in the model; generally, an indicator’s outer loading on its associated construct should be greater than its correlation with any other construct. Meanwhile, the Fornell–Larcker criterion compares the square root of each construct’s AVE value with its correlations with other constructs. The square root of each construct’s AVE should be greater than its highest correlation with any other construct in the model.

For evaluating the structural model, the coefficient of determination (R^2 value) and the level and significance of the path coefficients are the primary criteria [24]. R^2 is a “measure of the model’s predictive power and is calculated as the squared correlation between a specific endogenous construct’s actual and predicted values”. For endogenous latent variables, R^2 values of 0.75, 0.50, or 0.25 can be respectively described as substantial, moderate, or weak. The level and significance of the path coefficients are assessed through bootstrapping. In this process, subsamples are randomly drawn from the original sample with replacement [28]. This helps determine whether each relationship in the model is statistically significant and whether they support the hypotheses [24]. For this study, 300 subsamples were set for bootstrapping.

In examining the measurement and structural models, survey items with low loadings were removed in order to improve the constructs’ reliability and validity. Removed survey items are indicated in the Multimedia Appendix 1. The PLS algorithm in Smart-PLS provides the reports of construct reliability and validity, discriminant validity, collinearity statistics (variance inflation factor), and R^2. Bootstrapping provides the specific

XSL•FO
RenderX

indirect effects report for mediation effects and the path coefficients report for path significance.

Ethical Considerations

This study was approved by the Office of Research Protections and Integrity, University of North Carolina at Charlotte (review number 21-0223). The study was approved with the statement "Exemption Category: 2. Survey, interview, public observation".

Results

Demographic Statistics

The survey included questions about age, ownership of wireless earbuds, and awareness of health risks. The youngest respondent was 18 years old while the oldest was 89 years. The average age of the respondents was 43 (SD 14.29) years, with the majority falling in the age range of 30-49 years. Among those who owned wireless earbuds, ownership duration ranged from 1 month to 8 years and 4 months, with an average of a year and 4 months. Out of the 205 respondents, 63.4% (130/205) owned wireless earbuds, and 56.6% (116/205) were aware of the potential health risks associated with them.

Measurement and Structural Model Results

The construct reliability and validity reports provided each construct's Cronbach alpha and AVE. Cronbach α for every construct exceeded 0.80; thus, all constructs can be said to be reliable. Every construct exhibited an AVE higher than 0.50, suggesting convergent validity. Table 2 presents Cronbach α and AVE values for each construct. Discriminant validity is also presented in the form of the Fornell-Larcker criterion in Table 2. The diagonal cells show the square root of AVE of each construct. No constructs have exceptionally high correlations. This means that the constructs were truly distinct from each other by empirical standards. The correlation between perceived fashionability and wearable comfort is higher than between other variables. However, according to the collinearity statistics report, every variance inflation factor value was below 5.; Thus, we conclude that the constructs are acceptable for further analysis.

Table . Discriminant validity using the Fornell–Larcker criterion.

Model Construct	Purchase intention	Use intention	Fashionability	Health risk	Privacy risk	Wearable comfort
Purchase intention	0.901					
Use intention	0.840	0.924				
Fashionability	0.668	0.658	0.920			
Health risk	–0.422	–0.442	–0.319	0.748		
Privacy risk	–0.137	–0.175	–0.110	0.534	0.774	
Wearable comfort	0.642	0.626	0.676	–0.406	–0.152	0.879
Cronbach α	0.884	0.914	0.820	0.806	0.845	0.853
Average Variance Extracted (AVE)	0.812	0.853	0.847	0.560	0.599	0.773
Composite reliability	0.928	0.946	0.917	0.834	0.852	0.911

The PLS result shows that R^2 and adjusted R^2 values for purchase intention were 0.560 and 0.544, respectively, and R^2 and adjusted R^2 of behavioral intention to use were 0.757 and 0.747, respectively. Therefore, the model's predictive power is substantial. The specific indirect effects report provides the significance of the mediator. As shown in Table 3, this was statistically significant (P<.05), suggesting that behavioral intention to purchase mediates the relationship of perceived health risk, perceived fashionability, and wearable comfort with behavioral intention to use. However, the mediation of the relationship between perceived privacy risk and behavioral intention to use by behavioral intention to purchase was not found to be statistically significant. The path coefficients report provides the significance of each path in the model. As shown in Table 3, the P values for the indirect effects of perceived fashionability, perceived health risk, and wearable comfort on behavioral intention to use, mediated through behavioral intention to purchase, were all less than .05. The relationship between behavioral intention to purchase and behavioral intention to use was also statistically significant (Table 4). Hypotheses H1, H3, H4, and H5 are therefore supported by the data. However, the P value for the indirect effect of perceived privacy risk was greater than .05. Therefore, H2 is not supported. The overall results are illustrated in Figure 2.

XSL•FO
RenderX

Table . Specific indirect effects.

Path	Original sample	Sample mean (SD)	*P* value
Privacy Risk → Purchase Intention → Use Intention	0.035	0.023 (0.042)	.40
Health Risk → Purchase Intention → Use Intention	−0.114	−0.108 (0.048)	.02
Fashionability → Purchase Intention → Use Intention	0.230	0.223 (0.061)	<.001
Wearable Comfort → Purchase Intention → Use Intention	0.167	0.167 (0.067)	.01

Table . Path coefficients.

Paths	Original sample	Sample mean (SD)	*P* value
Purchase Intention → Use Intention	0.624	0.621 (0.076)	<.001
Fashionability → Purchase Intention	0.368	0.359 (0.084)	<.001
Fashionability → Use Intention	0.124	0.122 (0.061)	.04
Health Risk → Purchase Intention	–0.183	–0.174 (0.075)	.02
Health Risk → Use Intention	–0.049	–0.053 (0.047)	.30
Privacy Risk → Purchase Intention	0.057	0.040 (0.069)	.41
Privacy Risk → Use Intention	–0.022	–0.020 (0.052)	.68
Wearable Comfort → Purchase Intention	0.268	0.269 (0.098)	.007
Wearable Comfort → Use Intention	0.059	0.061 (0.054)	.278

XSL•FO
RenderX

Figure 2. Final model results – path coefficients and statistical significance. *P<.05, **P<.01, ***P<.001

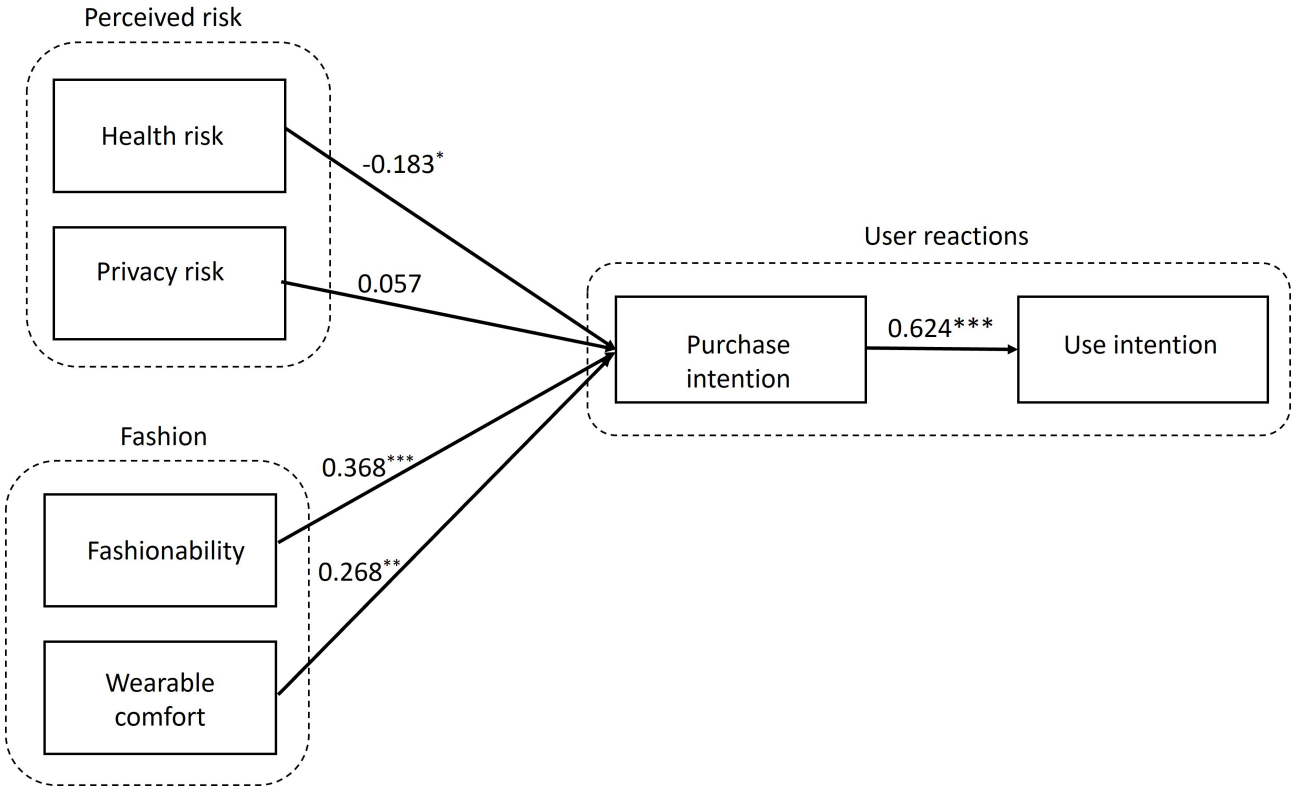

Table 4 presents the direct path coefficients, further highlighting the relationships among variables. Perceived fashionability had a direct effect on use intention, indicating a complementary partial mediation, whereas the direct effects of perceived health risk and wearable comfort on use intention were not statistically significant, indicating a full mediation.

Discussion

Principal Findings

Based on the results of this study, perceived health risk, perceived fashionability, and wearable comfort have a significant impact on consumers' decisions in terms of purchasing wireless earbuds. The significance of perceived health risk fills a gap in the literature on wearable technology, which has not been previously adequately studied. Perceived health risks had a significant negative impact on the intention to purchase wireless earbuds; however, they did not have a significant impact on the intention to use the devices. This could mean that even before considering using wireless earbuds, consumers may not purchase them at all due to the potential health risks. However, once consumers had purchased wireless earbuds, the perceived health risks did not seem to affect their subsequent usage. Poor sales impact business; thus, it is important for developers to make sure that their technology and products are fully tested and safe to use before introducing them to the market. In addition, producers should help consumers understand that their products are safe to use. Perceived privacy risk did not have a significant impact on either the intention to purchase or use wireless earbuds. This could mean that consumers do not consider potential privacy risks when they purchase and use wireless earbuds, or they may not be aware of the risk. In the qualitative responses to the survey, none of the respondents mentioned privacy risks among their concerns. In future research, adding a survey question asking if the respondent is aware of any privacy risks from using wearable devices would provide more insight.

The significance of perceived fashionability and wearable comfort reaffirms that consumers view wearable devices as not only technical devices but also as fashion items. Perceived fashionability has a significant impact on the intention to both purchase and use wireless earbuds. This means consumers may not purchase and use a wearable device if they think the product is ugly. Developers should make sure to produce wearable devices that are not only highly technical but also fashionable. Furthermore, wearable devices must not only be fashionable but also comfortable to wear.

Comparing the path coefficients and effect size between the three significant variables (perceived health risk, perceived fashionability, and wearable comfort) provides insight into how impactful each variable is. Wearable comfort had the highest path coefficient (0.344), with perceived fashionability second (0.285) and perceived health risk the lowest (−0.202). The effect size, also known as f-squared, indicates the relevance of variables in explaining selected endogenous constructs [28]. A higher f-squared value means larger effects of the specific variable on endogenous constructs. As with the comparison of path coefficients, wearable comfort exhibited the highest effect size (0.052), with perceived fashionability second (0.041) and perceived health risk the lowest (0.039). The results of the path coefficients and effect sizes imply that wearable comfort has the most impact on the consumer's intention to purchase wireless earbuds. In other words, consumers consider the fashion of wireless earbuds more than the wireless earbuds' potential health risks.

XSL•FO
RenderX

Conclusion

Certain limitations of this study provide opportunities for future research. First, the sample group for the survey was recruited by Qualtrics, which hired respondents through their panel program. Respondents were paid to complete the survey. Studying a larger and truly randomized sample group may help generalize the results. In addition, with a larger sample group, it would be possible to compare results between those who already own and those who do not own a wearable device. The results could vary depending on the type of wearable device, the type of consumer, and how or why the device was purchased. The impact of perceptions could be different when intending to purchase or use wireless earbuds compared to other devices like smartwatches. Adding technology type to the model in future research with a larger sample group could provide a more generalizable result. Different types of consumers will have different perceptions and needs. For example, some survey respondents did not see the need to purchase and use wireless earbuds due to hearing conditions (hearing issues or hearing aids). Patients who need to track their health conditions may have different motivations for using wearable devices. Motivation for purchasing a wearable device would also be different if the buyer and the actual user were different people. For example, perceptions toward wearable devices could be different between parents who are purchasing the device for their children and the children themselves as the actual users of the device. Consumers with different education levels, salaries, or jobs may also have different perceptions toward wearable devices.

The availability of alternatives could be another factor that affects the acceptance of wearable devices. For example, wired headsets are an alternate option to wireless earbuds for listening to music. However, no alternative provides the same functionality that smart glasses offer. A number of respondents also mentioned that they are opposed to using wireless earbuds because they think they could be a distraction or block the user's surroundings. These concerns could be considered new types of perceived risk. Conducting qualitative as well as quantitative research in the future could provide more insights and ideas for wearable technologies. As technology advances, new types of emerging technology will be introduced to the market. Technology acceptance models and studies must also evolve to understand consumers' perceptions toward these new technologies from both academic and practical points of view.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Survey Items.
[DOC File, 33 KB - formative_v9i1e56887_app1.doc]

References

1. Niknejad N, Ismail WB, Mardani A, et al. A comprehensive overview of smart wearables: the state of the art literature, recent advances, and future challenges. Eng Appl Artif Intell 2020 Apr;90:103529. [doi: 10.1016/j.engappai.2020.103529]
2. Zhang N, Peng Y, Guo Q. Visual analysis of research trends and hotspots in wearable electronic devices in the medical field: a bibliometric study. Digit Health 2024;10:20552076241305233. [doi: 10.1177/20552076241305233] [Medline: 39679002]
3. Waugh R. Wireless headphones like apple airpods 'could pose cancer risk', scientists warn. Yahoo News. 2019 Mar 11. URL: https://news.yahoo.com/wireless-headphones-like-apple-airpods-pose-cancer-risk-scientists-warn-144751987.html [accessed 2025-03-27]
4. Bilton N. The health concerns in wearable tech. The New York Times. 2015 Mar 19. URL: https://www.nytimes.com/2015/03/19/style/could-wearable-computers-be-as-harmful-as-cigarettes.html [accessed 2025-03-27]
5. Ajzen I, Fishbein M. Understanding Attitudes and Predicting Social Behavior: Prentice-Hall; 1980.
6. Herz M, Rauschnabel PA. Understanding the diffusion of virtual reality glasses: the role of media, fashion and technology. Technol Forecast Soc Change 2019 Jan;138:228-242. [doi: 10.1016/j.techfore.2018.09.008]
7. Bower M, Sturman D. What are the educational affordances of wearable technologies? Comput Educ 2015 Oct;88:343-353. [doi: 10.1016/j.compedu.2015.07.013]
8. Venkatesh V, Thong JY, Xu X. Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology. MIS Q 2012;36(1):157-178. [doi: 10.2307/41410412]
9. Brown SA, Venkatesh V. Model of adoption of technology in households: a baseline model test and extension incorporating household life cycle. MIS Q 2005;29(3):399-426. [doi: 10.2307/25148690]
10. Dodds WB, Monroe KB, Grewal D. Effects of price, brand, and store information on buyers' product evaluations. J Mark Res 1991 Aug;28(3):307-319. [doi: 10.1177/002224379102800305]
11. Limayem M, Hirt SG, Cheung CMK. How habit limits the predictive power of intention: the case of information systems continuance. MIS Q 2007;31(4):705-737. [doi: 10.2307/25148817]

XSL•FO
RenderX

12. Grym K, Niela-Vilén H, Ekholm E, et al. Feasibility of smart wristbands for continuous monitoring during pregnancy and one month after birth. BMC Pregnancy Childbirth 2019 Jan 17;19(1):34. [doi: 10.1186/s12884-019-2187-9] [Medline: 30654747]
13. Li J, Ma Q, Chan AH, et al. Health monitoring through wearable technologies for older adults: smart wearables acceptance model. Appl Ergon 2019 Feb;75:162-169. [doi: 10.1016/j.apergo.2018.10.006] [Medline: 30509522]
14. Nunes GS, Arruda Filho EJM. Consumer behavior regarding wearable technologies: Google Glass. Innov Manag Rev 2018 Aug 30;15(3):230-246. [doi: 10.1108/INMR-06-2018-0034]
15. Dehghani M, Kim KJ. The effects of design, size, and uniqueness of smartwatches: perspectives from current versus potential users. Behav Inform Technol 2019 Nov 2;38(11):1143-1153. [doi: 10.1080/0144929X.2019.1571111]
16. Heid M. Are airpods and other bluetooth headphones safe? Elemental. 2019 Mar 7. URL: https://elemental.medium.com/are-airpods-and-other-bluetooth-headphones-safe-214a0449e13a [accessed 2025-03-27]
17. Cocosila M, Turel O, Archer N, et al. Perceived health risks of 3G cell phones: do users care? Commun ACM 2007 Jun;50(6):89-92. [doi: 10.1145/1247001.1247026]
18. Duffy C. Why conspiracy theorists think 5G is bad for your health and why experts say not to worry. CNN Business. 2020 Jun 14. URL: https://www.cnn.com/2020/06/14/tech/5g-health-conspiracy-debunked/index.html [accessed 2025-03-27]
19. Barua A, Al Alamin MA, Hossain M, et al. Security and privacy threats for bluetooth low energy in IoT and wearable devices: a comprehensive survey. IEEE Open J Commun Soc 2022;3:251-281. [doi: 10.1109/OJCOMS.2022.3149732]
20. Mathavan B, Vafaei-Zadeh A, Hanifah H, et al. Understanding the purchase intention of fitness wearables: using value-based adoption model. Asia-Pac J Bus Adm 2024 Jan 2;16(1):101-126. [doi: 10.1108/APJBA-04-2022-0166]
21. Homburg C, Schwemmle M, Kuehnl C. New product design: concept, measurement, and consequences. J Mark 2015 May;79(3):41-56. [doi: 10.1509/jm.14.0199]
22. Ajzen I. The theory of planned behavior. Organ Behav Hum Decis Process 1991 Dec;50(2):179-211. [doi: 10.1016/0749-5978(91)90020-T]
23. Luo X, Li H, Zhang J, et al. Examining multi-dimensional trust and multi-faceted risk in initial acceptance of emerging technologies: an empirical study of mobile banking services. Decis Support Syst 2010 May;49(2):222-234. [doi: 10.1016/j.dss.2010.02.008]
24. Hair JF, Ringle CM, Sarstedt M. PLS-SEM: Indeed a Silver Bullet. J Mark Theory Pract 2011 Apr;19(2):139-152. [doi: 10.2753/MTP1069-6679190202]
25. Jacoby J, Kaplan LB. The components of perceived risk. ACR special volumes 1972;10:382-393.
26. Featherman MS, Pavlou PA. Predicting e-services adoption: a perceived risk facets perspective. Int J Hum Comput Stud 2003 Oct;59(4):451-474. [doi: 10.1016/S1071-5819(03)00111-3]
27. Chin WW, Marcolin BL, Newsted PR. A partial least squares latent variable modeling approach for measuring interaction effects: results from a Monte Carlo simulation study and an Electronic-Mail Emotion/Adoption study. Inf Syst J 2003 Jun;14(2):189-217. [doi: 10.1287/isre.14.2.189.16018]
28. Hair JF Jr, Hult GTM, Ringle C, et al. A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM): Sage Publications; 2016.

Abbreviations

AVE: average variance extracted
PLS: partial least squares
UN: United Nations
UTAUT 2: Unified Theory of the Adoption and Use of Technology 2
WHO: World Health Organization

Edited by K Bond; submitted 29.01.24; peer-reviewed by HS Lom, T Miron-Shatz; revised version received 31.03.25; accepted 25.04.25; published 19.06.25.

Please cite as:
Lee EM, Subramaniam C, Park S
Perceived Risk and Fashion on the Intention to Adopt Wireless Earbuds in the United States Using a Partial Least Squares-Structural Equation Modeling Approach: Empirical Study
JMIR Form Res 2025;9:e56887
URL: https://formative.jmir.org/2025/1/e56887
doi:10.2196/56887

XSL•FO
RenderX

© Edward M Lee, Chandrasekar Subramaniam, Sungjune Park. Originally published in JMIR Formative Research (https://formative.jmir.org), 19.6.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Internet Access and Use Among Dementia Carers and the People They Support in Australia: Cross-Sectional Survey

Emilie Cameron[1,2], BSc, PhD; Elise Mansfield[1,2], BA (Psych), PhD; Ama Ampofo[1,2], BSc, MSc, PhD; Andrea Coda[2,3], BSc, PhD; Allison Boyes[1,2], BA (Psych), MPH, PhD

[1]Health Behaviour Research Collaborative, School of Medicine and Public Health, College of Health, Medicine and Wellbeing, University of Newcastle, Callaghan, NSW, Australia

[2]Equity in Health and Wellbeing Research Program, Hunter Medical Research Institute, New Lambton Heights, Australia

[3]School of Health Sciences, College of Health, Medicine and Wellbeing, University of Newcastle, Ourimbah, Australia

Corresponding Author:
Emilie Cameron, BSc, PhD
Health Behaviour Research Collaborative, School of Medicine and Public Health, College of Health, Medicine and Wellbeing, University of Newcastle, Callaghan, NSW, Australia

Abstract

Background: Dementia is a major public health priority due to its increasing prevalence and the considerable burden it places on individuals living with the condition and their carers. Internet-based tools can help carers and the people they support to manage daily tasks, access resources and support, track health data, and connect with health care professionals. However, the benefits of these tools will only be realized if the intended users have access to the internet and are confident in its use.

Objective: This study aimed to examine the patterns of internet access and use among dementia carers in Australia and the people they support.

Methods: A cross-sectional survey of carers providing informal support to a person diagnosed with dementia and living in the community was conducted. Carers were recruited through geriatric clinics, aged care providers, and community organizations between July 2018 and June 2020. Consenting carers self-completed a multitopic survey that included study-specific questions about their access to the internet, confidence using the internet, and whether the person with dementia they support was able to use the internet. Multivariate regression analysis identified sociodemographic factors associated with carers' internet access.

Results: A total of 172 carers (consent rate 47%) with an average age of 71.8 (SD 10.91) years completed the survey. Most carers (126/155, 81%) had access to the internet; however, 31% (47/153) reported they were not at all confident in using it. The likelihood of carers having access to the internet decreased with carers' age (OR [odds ratio] 0.87, 95% CI 0.80 - 0.96; P=.003). Male carers were 4 times more likely to have internet access than female carers (OR 4.32, 95% CI 1.00 - 18.6; P=.05). Similarly, carers with private health insurance (an indicator of individual socioeconomic status) were 8 times more likely to have internet access than those without private health insurance (OR 8.45, 95% CI 2.74 - 26; P<.001). Only 17% (29/166) of carers perceived that the person with dementia they support was able to use the internet independently or with assistance.

Conclusions: Despite high rates of internet access among carers, significant age, gender, and socioeconomic disparities were found, and a notable proportion lacked confidence in its use. Internet use among people with dementia was limited. The digital disparities identified in this study suggest that targeted training to build digital literacy to effectively use internet resources and co-design new technologies is needed. The findings further highlight that in this digital era, alternative methods to deliver dementia care and resources are essential to accommodate those who are unable to access or are less comfortable using the internet.

(JMIR Form Res 2025;9:e68333) doi:10.2196/68333

KEYWORDS

Alzheimer disease; caregiver; support person; technology; online tools

Introduction

Dementia is a major public health concern worldwide, contributing significantly to disability, dependency, and reduced quality of life among older adults. The term dementia refers to a group of symptoms that affect memory and other cognitive functions, mood, behavior, and the ability to perform activities of daily living. Alzheimer disease is the most common form of dementia [1]. As of 2022, an estimated 56 million people worldwide were living with dementia, a figure projected to rise to 78 million by 2030 due to population aging, with global care costs set to exceed US $2.8 trillion [2]. People living with dementia often have complex physical, psychosocial, and practical needs. The progressive nature of dementia can lead to

XSL•FO
RenderX

loss of independence, increased risk of hospitalization, and social isolation [3,4]. Informal caregivers, often family members, play a crucial role in supporting the person with dementia. However, caregivers frequently report difficulties in accessing clear information, appropriate support services, and respite care. Many experience high levels of stress, anxiety, depression, and social isolation, along with uncertainties about how to manage the behavioral and medical challenges associated with dementia [5-9].

Recent years have seen a seismic shift toward digital health technology as part of the delivery of health care and care services for older adults, including services provided to people with dementia and their carers [10]. These include telehealth consultations, online access to medical records and information, wearable health tracking devices, and web-based peer support platforms. For both people with dementia and their carers, such tools offer promising opportunities to improve care, reduce burden, and promote autonomy [11,12]. Digital solutions confer significant potential advantages over traditional face-to-face modes of support: they allow access to integrated care and support in a timely manner, which may help to de-escalate behavioral symptoms and prevent crisis situations, allow for greater accessibility of support for those living in rural or remote areas, can be accessed at a time and place that suits the user, can be tailored to individual needs and preferences, are easily modified to take into account visual or hearing impairments (eg, option to present information as video or text, read-aloud features), and are likely to be more affordable at a time of increased financial strain [12-15]. In recognition of the multitude of potential benefits of digital technologies, international action plans such as the United Nations Decade of Healthy Ageing: Plan of Action (2021 - 2030) advocate for the use of digital technologies to expand access to quality care for older adults [16].

Evidence from intervention studies supports the effectiveness of internet-based programs in improving outcomes for carers, including reduced depressive symptoms, lower perceived stress, increased self-efficacy, and improved confidence in decision-making [17-19]. Similarly, for people with dementia, digital tools have shown potential to enhance well-being and cognitive functioning [20,21]. However, many technologies shown to be effective in controlled research settings have yet to be widely adopted in practice. Reasons provided by older adults for not using the internet include not having a need or interest in it, a lack of confidence, and not having access [22]. Factors such as age, gender, education, income, remoteness, health status including mental health, and digital literacy influence technology use among older adults [23-26]. Reluctance to engage with digital solutions can stem from mistrust, unfamiliarity, or perceived complexity [11,14,27]. Consequently, many carers and people with dementia may be missing out on the benefits of online interventions.

To date, few studies have examined population-level internet use among people with dementia, with existing research largely focusing on specific interventions (eg, cognitive training apps) or usage in early-stage dementia [20,21,27-31]. Studies from various countries suggest rising internet usage among carers [32-34], for example, more than 90% reported internet use for caregiving purposes in Portugal in 2022, but these findings are often based on online convenience samples, which may not reflect broader populations [34]. The extent of carer's confidence and training needs remains underexplored. This study aimed to address these gaps by exploring how confident carers of people with dementia are in using the internet, the factors associated with having internet access, and whether the person they support is able to use the internet.

Methods

Design

A cross-sectional survey of carers of people with dementia was conducted.

Participants

Eligible carers were aged 18 years and older, self-identified as able to read and write English, and were the primary source of informal practical and emotional support to a person living in the community who had received a confirmed diagnosis of dementia from a medical professional.

Recruitment and Data Collection

Participants were recruited as part of a larger study examining dementia carer needs [5]. Potentially eligible carers were identified through multiple settings in 2 states of Australia (New South Wales and Victoria) including 10 private geriatrician clinics, 3 community-based respite centers for people with dementia, 6 carer support groups, 34 community-based social groups for older people, and an aged care provider. Staff members within these settings reviewed records and provided potential participants with a study information pack including a cover letter, information statement, hard copy of the survey, and reply-paid envelope for the return of the survey directly to the research team. Alternatively, the study was advertised within the setting, and interested carers completed a study eligibility form, which was returned to the research team to mail out the study information pack. The survey could be completed by pen and paper or online if preferred. A reminder including a second copy of the survey was sent after 4 weeks to nonresponders. Data were collected between July 2018 and June 2020.

Measures

Participants self-reported their sex, age, postcode, highest level of education, private health insurance status, concession card status, presence of health conditions, and relationship to the person with dementia and completed a measure of depression severity (Patient Health Questionnaire-9 [PHQ-9]) [35]. Participants were asked whether they have access to the internet (Yes or No), how confident they are in using the internet (Very confident or Somewhat confident or not at all confident), and whether the person they support would be able to use the internet if they wanted to (Yes on their own or Yes with help or No).

Analysis

Analyses were conducted in R version 4.3.1 (2023-06-16 ucrt; R Foundation for Statistical Computing). Descriptive statistics are reported as counts and proportions for categorical variables and mean and SD for continuous variables. Postcode was used

XSL•FO
RenderX

to calculate an Accessibility/Remoteness Index of Australia Plus (ARIA+) score indicating whether people live in a major city or regionally [36]. Postcode was also used to estimate the degree of socioeconomic disadvantage using the Socio-Economic Index for Areas (SEIFA) Index of Advantage and Disadvantage (2016) [37].

Factors related to internet access were first modeled individually and then all the factors were included in a multivariable logistic regression model. These factors included age, depression score (PHQ-9), sex, ARIA+ score, SEIFA decile, highest level of education, relationship to person with dementia, living arrangements, private health insurance status, and the presence of health conditions.

Ethical Considerations

The study was approved by the Human Research Ethics Committees of Hunter New England Health (17/05/17/4.07) and the University of Newcastle (H-2018 - 0308). Participation in the study was voluntary. Completion and return of the survey was taken as implied consent to participate. All personal information was stored separately from responses collected in the survey and only deidentified data were analyzed. Participants were not compensated for their participation in the study.

Results

Sociodemographic Characteristics

A total of 172 carers completed the survey (consent rate 47%). Participant characteristics are shown in Table 1. Most participants were female (124/172, 74%) lived in a major city (118/140, 84%) and had an average age of 71.8 (SD 10.91) years. The mean PHQ-9 score was 5.50 (SD 5.37), which is classified as mild depression.

XSL•FO
RenderX

Table . Sociodemographic and internet characteristics of study participants (N=172).

Characteristics	Values[a]
Age (years), mean (SD)	71.81 (10.91)
PHQ-9[b] score, mean (SD)	5.5 (5.37)
Sex, n (%)	
Female	124 (74)
Male	44 (26)
ARIA+ score[c], n (%)	
Major city	140 (84)
Regional	27 (16)
SEIFA[d], n (%)	
Most disadvantaged (decile≤4)	102 (61)
Least disadvantaged (decile>4)	65 (39)
Education, n (%)	
High school or below	67 (40)
Vocational training or University	99 (60)
Relationship to person with dementia, n (%)	
Partner	140 (83)
Parent	22 (13)
Other	7 (4.1)
Living arrangement, n (%)	
Live separately to person they support	12 (7.1)
Live with person they support	157 (93)
Private health insurance, n (%)	
No	42 (25)
Yes	126 (75)
Health conditions, n (%)	
None or 1 health condition	93 (56)
>1 health conditions	73 (44)
Internet access, n (%)	
No	29 (19)
Yes	126 (81)
Confidence using internet, n (%)	
Very confident	52 (34)
Somewhat confident	54 (35)
Not at all confident	47 (31)
Person with dementia able to use internet, n (%)	
Yes, on their own	13 (7.8)
Yes, with help	16 (9.6)
No	137 (83)

[a]Totals do not all add to N=172 due to missing responses.

[b]PHQ-9: Patient Health Questionnaire-9.

[c]ARIA+: Accessibility/Remoteness Index of Australia Plus.

[d]SEIFA: Socio-Economic Index for Areas.

XSL•FO
RenderX

Internet Access and Use

Most carers (126/155, 81%) indicated they had access to the internet. Overall, 34% (52/153) of carers reported being very confident, and 35% (52/153) reported being somewhat confident using the internet. The majority of people with dementia (114/166, 83%) were reported as not able to use the internet. A further 9.6% (16/166) of people with dementia were reported as being able to use the internet with help, and only 7.8% (13/166) were reported as able to use the internet independently.

Factors Associated with Carers' Internet Access

As shown in Table 2, after adjusting for other factors, age, sex, and private health insurance status were significantly associated with internet access among carers of people with dementia. For every 1-year increase in age, there was a 13% reduction in the odds of a carer having internet access (odds ratio [OR] 0.87, 95% CI 0.80 - 0.96; P=.003). Male carers were 4 times more likely to have internet access than females (OR 4.32, 95% CI 1.00 - 18.6; P=.05). Similarly, carers with private health insurance were 8 times more likely to have internet access compared to those who did not (OR 8.45; 95% CI 2.74 - 26; P<.001).

XSL•FO
RenderX

Table . Sociodemographic factors associated with having internet access among carers of people with dementia (n=155).

Characteristics	Crude		Adjusted	
	OR[a] (95% CI[b])	*P* value[c]	OR (95% CI)	*P* value
Age	0.94 (0.89 - 0.99)	.01	0.87 (0.80 - 0.96)	.003
PHQ-9 score[d]	1.11 (1.01 - 1.22)	.05	1.09 (0.96 - 1.24)	.20
Sex				
Female	Reference	—[e]	Reference	—
Male	2.27 (0.80 - 8.12)	.16	4.32 (1.00 - 18.6)	.05
ARIA+ score[f]				
Major city	Reference	—	Reference	—
Regional	0.76 (0.25 - 2.87)	.66	0.81 (0.18 - 3.69)	.80
SEIFA[g]				
Most disadvantaged	Reference	—	Reference	—
Least disadvantaged	0.76 (0.33 - 1.78)	.52	1.21 (0.41 - 3.61)	.70
Education				
High school or below	Reference	—	Reference	—
Vocational training or University	1.82 (0.80 - 4.14)	.15	1.42 (0.50 - 4.04)	.50
Relationship to person with dementia				
Partner	Reference	—	Reference	—
Parent	5.35 (0.69 - 41.6)	.11	4.31 (0.16 - 119)	.40
Other	1.07 (0.11 - 9.97)	.95	0.46 (0.02 - 11.2)	.60
Living arrangement				
Live separately	Reference	—	Reference	—
Live with person they support	0.52 (0.03 - 3.02)	.55	32.7 (0.58 - 1,845)	.09
Private health insurance				
No	Reference	—	Reference	—
Yes	5.06 (2.15 - 12.2)	<.001	8.45 (2.74 - 26.0)	<.001
Carer's health conditions				
None or 1 health condition	Reference	—	Reference	—
>1 health conditions	0.72 (0.31 - 1.65	.43	1.17 (0.40 - 3.43)	.80

[a]OR: odds ratio

[b]CI: Confidence Interval

[c]*P* values <.05 are significant.

[d]PHQ-9: PHQ-9: Patient Health Questionnaire-9.

[e]Not applicable.

[f]ARIA+: Accessibility/Remoteness Index of Australia Plus.

[g]SEIFA: Socio-Economic Index for Areas.

XSL•FO
RenderX

Discussion

Principal Findings

We found that most (126/172, 81%) carers in our study had access to the internet. The proportion was higher than expected given the average age of the carers in our study was 71 (SD 10.91) years. Data from the Australian Bureau of Statistics suggest that in 2018, 62% of people older than the age of 65 years were internet users and this increased to 76% in 2022 [38]. Other studies reporting on internet use among dementia carers have tended to have a younger cohort and often use internet-based recruitment methods [32-34]. Compared to national data, our sample contained more carers who were the partner of the person with dementia, were older, and were residing in major cities [5]. Our findings suggest that it is feasible to leverage the increasing rates of internet access among older people, including dementia carers, to provide dementia care services, such as e-learning modules, virtual support groups, telehealth services, and online care plans.

Consistent with other studies illustrating the digital divide, our cross-sectional survey found that younger carers were more likely to have internet access than older carers [22,24,32,33]. Men were also more likely to have internet access than women. As a majority of carers of people with dementia are women [39], there is a need to further explore reasons for reduced uptake and develop strategies to overcome barriers to uptake among this group. Those with private health insurance were also more likely to have internet access, compared to those without. Having private health insurance is an indicator of individual socioeconomic status [40]. This finding likely reflects the costs involved in accessing digital technologies including initial financial outlay for equipment and ongoing charges for internet connection [11]. These findings suggest that dementia care resources and services must continue to be accessible through a variety of channels, particularly for carers from vulnerable population groups who often face social and economic barriers to digital inclusion.

Part of eHealth literacy theory posits that in order to effectively use internet technology, carers not only need access to the internet but also the self-efficacy to use it [41]. Accordingly, a systematic review showed that interventions designed to increase self-efficacy in the use of eHealth among older adults were associated with increases in their ability to manage health concerns [42]. We found that only one-third of dementia carers feel very confident in using the internet. There is an opportunity to help carers feel more confident in using the internet through successful eHealth literacy training programs [23,24] such as “Be Connected”—an Australian Government initiative to improve digital literacy for older adults [43]. Promoting the use of such programs may assist in improving uptake of and engagement with a range of digital health platforms. In addition, ensuring carers are provided adequate training in the use of new technologies as they are introduced will help them to feel comfortable, less frustrated, and more likely to use them [11]. Future research should consider the factors associated with confidence in using the internet and ways to build digital confidence so that the full potential of internet-based support systems can be realized.

While internet access among carers in our study was high, our findings revealed that only 17% of people with dementia were able to use the internet either independently or with assistance. This proportion is consistent with data from the Australian Bureau of Statistics which identified 12% of people with dementia as internet users in 2018 and 27% in 2022 [38]. While internet use is low among people with dementia, some individuals do engage online, benefiting from social connection, sharing stories, and accessing additional support [27,29,30]. Despite this, research has primarily focused on the development and testing of interventions aimed at carers, with little attention to the benefits and potential for people with dementia themselves or ways carers and people with dementia could use the technology together [25]. A systematic review of 90 studies found that people with dementia often required support from carers to navigate internet technologies [31]. There is a risk that increased use of internet-based solutions will increase the demands placed on carers and reduce the input of the person with dementia into their own care [31,44]. These findings highlight the need to carefully consider who will be using digital technologies and whether their needs are being met [28,30,45]. Co-designing solutions with people with dementia should be a priority to ensure digital interventions are relevant, practical, and aligned with their abilities, preferences, and online behaviors. Their involvement in the design process can lead to more effective and acceptable support tools.

Limitations and Future Directions

Our sample included carers who were linked with support systems, such as respite services and geriatricians, and may therefore not be representative of the population caring for people with dementia. The sample was older than the general population of carers in Australia and included few people outside major cities and none from remote locations. It also did not include those who could not read and write English. These groups may face further barriers to the use of internet-based innovations. The sample size was small due to the difficulties of recruiting people from vulnerable populations [46]. Although the majority of data were collected prior to the commencement of the COVID-19 pandemic, rates of internet access among carers were comparable with rates of internet access for older Australians from 2022 [38]. While internet access and user confidence are important factors in ensuring uptake and engagement with digital health programs, other potential barriers to internet use among these populations, such as level of eHealth literacy, fears about scams and security of personal data, and affordability of internet access could be explored in future studies.

Conclusions

Internet use among carers was found to be high, suggesting that digital health interventions are increasingly feasible for this population. However, disparities in access and confidence remain, particularly among older, female, and socioeconomically disadvantaged carers. Internet use among people with dementia, however, was limited. The findings reinforce the importance of providing dementia care through diverse modalities. To

XSL•FO
RenderX

maximize the benefits of digital health technologies, future efforts should prioritize equity in access, build digital literacy through targeted training, and ensure that online support systems are co-designed to ensure that they meet the needs of both carers and people with dementia.

Acknowledgments

We would like to acknowledge the carers who took part in this research as well as the geriatricians, aged care providers, respite facilities, and community organizations who assisted with recruitment. We received data management and statistical assistance from Sandra Dowley, Vera Tse, and Lucy Leigh. This study was funded by an NHMRC Boosting Dementia Research Grant (APP1137807). Infrastructure support was provided by the HMRI Equity in Health and Wellbeing Research Program.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

References

1. Global status report on the public health response to dementia. : World Health Organization; 2021 URL: https://www.who.int/publications/i/item/9789240033245 [accessed 2025-07-28]
2. A blueprint for dementia research. World Health Organization. 2022. URL: https://www.who.int/publications/i/item/9789240058248 [accessed 2025-07-28]
3. Mansfield E, Cameron E, Carey M, et al. Prevalence and type of unmet needs experienced by people living with dementia. J Alzheimers Dis 2022;87(2):833-842. [doi: 10.3233/JAD-215183] [Medline: 35404274]
4. Monin JK, Jorgensen TD, MacNeil Vroomen JL. Self-reports and caregivers' proxy reports of unmet needs of persons with dementia: implications for both partners' health-related quality of life. Am J Geriatr Psychiatry 2020 Mar;28(3):363-367. [doi: 10.1016/j.jagp.2019.10.006] [Medline: 31708379]
5. Mansfield E, Cameron EC, Boyes AW, et al. Prevalence and type of unmet needs experienced by carers of people living with dementia. Aging Ment Health 2023 May 4;27(5):904-910. [doi: 10.1080/13607863.2022.2053833]
6. Hargreaves S, Sbaffi L, Ford N. Information seeking amongst informal caregivers of people with dementia: a qualitative study. JD 2023 Mar 6;79(2):281-300. [doi: 10.1108/JD-03-2022-0066]
7. Macleod A, Tatangelo G, McCabe M, You E. "There isn't an easy way of finding the help that's available." Barriers and facilitators of service use among dementia family caregivers: a qualitative study. Int Psychogeriatr 2017 May;29(5):765-776. [doi: 10.1017/S1041610216002532] [Medline: 28351450]
8. Brodaty H, Donkin M. Family caregivers of people with dementia. Dialogues Clin Neurosci 2009;11(2):217-228. [doi: 10.31887/DCNS.2009.11.2/hbrodaty] [Medline: 19585957]
9. Lindeza P, Rodrigues M, Costa J, Guerreiro M, Rosa MM. Impact of dementia on informal care: a systematic review of family caregivers' perceptions. BMJ Support Palliat Care 2020 Oct 14;14(e1):e38-e49. [doi: 10.1136/bmjspcare-2020-002242] [Medline: 33055092]
10. Abdulazeem H, Borges do Nascimento IJ, Weerasekara I, et al. Use of digital health technologies for dementia care: bibliometric analysis and report. JMIR Ment Health 2025 Feb 10;12:e64445. [doi: 10.2196/64445] [Medline: 39928936]
11. Watson J, Hanna K, Talbot C, et al. A systematic review of digital access to post-diagnostic health and social care services for dementia. Int J Geriatr Psychiatry 2024 Jul;39(7):e6119. [doi: 10.1002/gps.6119] [Medline: 39004777]
12. Gross N, Byers V, Geiger S. Digital health's impact on integrated care, carer empowerment and patient-centeredness for persons living with dementia. Health Policy Technol 2021 Sep;10(3):100551. [doi: 10.1016/j.hlpt.2021.100551]
13. Ezeamii VC, Okobi OE, Wambai-Sani H, et al. Revolutionizing healthcare: how telemedicine is improving patient outcomes and expanding access to care. Cureus 2024 Jul;16(7):e63881. [doi: 10.7759/cureus.63881] [Medline: 39099901]
14. Teles S, Ferreira A, Paúl C. Attitudes and preferences of digitally skilled dementia caregivers towards online psychoeducation: a cross-sectional study. Behav Inf Technol 2023 Mar 12;42(4):345-359. [doi: 10.1080/0144929X.2021.2021285]
15. Angelopoulou E, Papachristou N, Bougea A, et al. How telemedicine can improve the quality of care for patients with alzheimer's disease and related dementias? A narrative review. Medicina (Kaunas) 2022 Nov 22;58(12):1705. [doi: 10.3390/medicina58121705] [Medline: 36556907]
16. UN Decade of Healthy Ageing: Plan of action (2021-2030). : United Nations; 2021 URL: https://cdn.who.int/media/docs/default-source/decade-of-healthy-ageing/decade-proposal-final-apr2020-en.pdf [accessed 2025-07-28]
17. Hopwood J, Walker N, McDonagh L, et al. Internet-based interventions aimed at supporting family caregivers of people with dementia: systematic review. J Med Internet Res 2018 Jun 12;20(6):e216. [doi: 10.2196/jmir.9548] [Medline: 29895512]

XSL•FO
RenderX

18. Leng M, Zhao Y, Xiao H, Li C, Wang Z. Internet-based supportive interventions for family caregivers of people with dementia: systematic review and meta-analysis. J Med Internet Res 2020 Sep 9;22(9):e19468. [doi: 10.2196/19468] [Medline: 32902388]
19. Ottaviani AC, Monteiro DQ, Oliveira D, et al. Usability and acceptability of internet-based interventions for family carers of people living with dementia: systematic review. Aging Ment Health 2022 Oct 3;26(10):1922-1932. [doi: 10.1080/13607863.2021.1975095]
20. Park H, Ha J. Effect of digital technology interventions for cognitive function improvement in mild cognitive impairment and dementia: a systematic review and meta‐analysis. Res Nurs Health 2024 Aug;47(4):409-422. [doi: 10.1002/nur.22383] [Medline: 38567389]
21. Ditton A, Alodan H, Fox C, Evans S, Cross J. Exploring the effectiveness and experiences of people living with dementia interacting with digital interventions: a mixed methods systematic review. Dementia (London) 2025 Apr;24(3):506-551. [doi: 10.1177/14713012241302371] [Medline: 39604136]
22. Use of information technology by people with disability, older people and primary carers. Australian Bureau of Statistics. 2020. URL: https://www.abs.gov.au/articles/use-information-technology-people-disability-older-people-and-primary-carers [accessed 2025-08-18]
23. Zhao YC, Zhao M, Song S. Online health information seeking behaviors among older adults: systematic scoping review. J Med Internet Res 2022 Feb 16;24(2):e34790. [doi: 10.2196/34790] [Medline: 35171099]
24. Estacio EV, Whittle R, Protheroe J. The digital divide: examining socio-demographic factors associated with health literacy, access and use of internet to seek health information. J Health Psychol 2019 Oct;24(12):1668-1675. [doi: 10.1177/1359105317695429] [Medline: 28810415]
25. Burns P, Jones SC, Iverson D, Caputi P. Riding the wave or paddling in the shallows? Understanding older Australians' use of the internet. Health Promot J Austr 2012 Aug;23(2):145-148. [doi: 10.1071/he12145] [Medline: 23088477]
26. Cotten SR, Ford G, Ford S, Hale TM. Internet use and depression among retired older adults in the United States: a longitudinal analysis. J Gerontol B Psychol Sci Soc Sci 2014 Sep;69(5):763-771. [doi: 10.1093/geronb/gbu018] [Medline: 24671896]
27. White C, Dunn R, Wolverson E, Zafeiridi E. The views of carers, people living with dementia and healthcare practitioners about the value of online information and peer support. Int J Care Caring 2020;4(4):549-571. [doi: 10.1332/239788220X15934528490504]
28. Talbot CV, Briggs P. The use of digital technologies by people with mild-to-moderate dementia during the COVID-19 pandemic: a positive technology perspective. Dementia (London) 2022 May;21(4):1363-1380. [doi: 10.1177/14713012221079477] [Medline: 35333111]
29. Talbot CV, O'Dwyer ST, Clare L, Heaton J, Anderson J. How people with dementia use twitter: a qualitative analysis. Comput Human Behav 2020 Jan;102:112-119. [doi: 10.1016/j.chb.2019.08.005]
30. Lee AR, McDermott O, Orrell M. Understanding barriers and facilitators to online and app activities for people living with dementia and their supporters. J Geriatr Psychiatry Neurol 2023 Sep;36(5):366-375. [doi: 10.1177/08919887221149139] [Medline: 36597870]
31. Yi JS, Pittman CA, Price CL, Nieman CL, Oh ES. Telemedicine and dementia care: a systematic review of barriers and facilitators. J Am Med Dir Assoc 2021 Jul;22(7):1396-1402. [doi: 10.1016/j.jamda.2021.03.015] [Medline: 33887231]
32. Kim H. Understanding internet use among dementia caregivers: results of secondary data analysis using the US caregiver survey data. Interact J Med Res 2015 Feb 23;4(1):e1. [doi: 10.2196/ijmr.3127] [Medline: 25707033]
33. Efthymiou A, Papastavrou E, Middleton N, Markatou A, Sakka P. How caregivers of people with dementia search for dementia-specific information on the internet: survey study. JMIR Aging 2020 May 19;3(1):e15480. [doi: 10.2196/15480] [Medline: 32427105]
34. Teles S, Paúl C, Costa-Santos C, Ferreira A. Use of dementia and caregiving-related internet resources by informal caregivers: a cross-sectional study. Front Med (Lausanne) 2022;9:978635. [doi: 10.3389/fmed.2022.978635] [Medline: 36186787]
35. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613. [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
36. Australian Statistical Geography Standard (ASGS): volume 2 - Indigenous structure, July 2016 (ABS cat no 1270.0.55.002). Australian Bureau of Statistics. 2016. URL: https://www.abs.gov.au/ausstats/abs@.nsf/mf/1270.0.55.002 [accessed 2025-07-28]
37. Census of population and housing: Socio-Economic Indexes for Areas (SEIFA), Australia, 2016 (ABS cat no 2033.0.55.001). Australian Bureau of Statistics. 2016. URL: https://www.abs.gov.au/ausstats/abs@.nsf/mf/2033.0.55.001 [accessed 2025-07-28]
38. Microdata and tablebuilder: Disability, Ageing and Carers, Australia. Australian Bureau of Statistics. 2024. URL: https://www.abs.gov.au/statistics/microdata-tablebuilder/available-microdata-tablebuilder/disability-ageing-and-carers-australia [accessed 2025-08-18]
39. Dementia in Australia. Australian Institute of Health and Welfare. 2024. URL: https://www.aihw.gov.au/reports/dementia/dementia-in-aus [accessed 2025-07-24]

XSL•FO
RenderX

40. Banks E, Jorm L, Lujic S, Rogers K. Health, ageing and private health insurance: baseline results from the 45 and up study cohort. Aust New Zealand Health Policy 2009 Jul 13;6(1):16. [doi: 10.1186/1743-8462-6-16] [Medline: 19594895]
41. Norman CD, Skinner HA. eHealth literacy: essential skills for consumer health in a networked world. J Med Internet Res 2006 Jun 16;8(2):e9. [doi: 10.2196/jmir.8.2.e9] [Medline: 16867972]
42. Pourrazavi S, Kouzekanani K, Bazargan-Hejazi S, et al. Theory-based eHealth literacy interventions in older adults: a systematic review. Arch Public Health 2020;78:72. [doi: 10.1186/s13690-020-00455-6] [Medline: 32793345]
43. McCosker A, Tucker J, Critchley C, Hiruy K, Walshe J, Suchowerska R, et al. Improving the digital inclusion of older Australians: the social impact of be connected. : Social Innovation Research Institute and the Centre for Social Impact, Swinburne University; 2020 URL: https://www.dss.gov.au/system/files/resources/improving-digital-inclusion-older-australians-social-impact-be-connected-16-june-2020.pdf [accessed 2025-07-28]
44. Tuijt R, Rait G, Frost R, Wilcock J, Manthorpe J, Walters K. Remote primary care consultations for people living with dementia during the COVID-19 pandemic: experiences of people living with dementia and their carers. Br J Gen Pract 2021 Aug;71(709):e574-e582. [doi: 10.3399/BJGP.2020.1094] [Medline: 33630749]
45. Neal D, van den Berg F, Planting C, et al. Can use of digital technologies by people with dementia improve self-management and social participation? A systematic review of effect studies. J Clin Med 2021 Feb 5;10(4):604. [doi: 10.3390/jcm10040604] [Medline: 33562749]
46. Beattie E, Fielding E, O'Reilly M, Brooks D, MacAndrew M, McCrow J. Recruitment of individuals with dementia and their carers for social research: lessons learned from nine studies. Res Gerontol Nurs 2018 May 1;11(3):119-128. [doi: 10.3928/19404921-20180308-01] [Medline: 29634849]

Abbreviations

ARIA+: Accessibility/Remoteness Index of Australia Plus
CI: Confidence Interval
OR: odds ratio
PHQ-9: Patient Health Questionnaire-9
SD: Standard deviation
SEIFA: Socio-Economic Index for Areas

Edited by A Mavragani; submitted 03.11.24; peer-reviewed by AS Gilbert, W Tang; revised version received 26.06.25; accepted 27.06.25; published 22.08.25.

Please cite as:
Cameron E, Mansfield E, Ampofo A, Coda A, Boyes A
Internet Access and Use Among Dementia Carers and the People They Support in Australia: Cross-Sectional Survey
JMIR Form Res 2025;9:e68333
URL: https://formative.jmir.org/2025/1/e68333
doi:10.2196/68333

© Emilie Cameron, Elise Mansfield, Ama Ampofo, Andrea Coda, Allison Boyes. Originally published in JMIR Formative Research (https://formative.jmir.org), 22.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Adolescents' and Young Adults' Perceptions of a Pop-Up Aimed at Combating the Spread of E-Cigarette Misinformation on Social Media: Cross-Sectional Survey Study

Jessica Maturo[1,2], BA; Shivani Mathur Gaiha[1,3], BA, MSc, PhD

[1]Division of Adolescent/Young Adult Medicine, Department of Pediatrics, Boston Children's Hospital, 1 Autumn St, 5th Fl, Boston, MA, United States

[2]Boston University School of Public Health, Boston, MA, United States

[3]Department of Pediatrics, Harvard Medical School, Boston, MA, United States

Corresponding Author:
Shivani Mathur Gaiha, BA, MSc, PhD
Division of Adolescent/Young Adult Medicine, Department of Pediatrics, Boston Children's Hospital, 1 Autumn St, 5th Fl, Boston, MA, United States

Abstract

Background: Social media is an important source of e-cigarette–related information for adolescents and young adults. However, misinformation is being shared across platforms, which may encourage e-cigarette use.

Objective: This study aims to examine adolescent and young adult perceptions of a novel pop-up on social media that provides links to vaping–related health information from credible sources (eg, CDC).

Methods: Between December 2023 and March 2024, participants aged 13‑24 years (N=5326) completed a web-based cross-sectional survey. Participants were asked to select from 4 positive (eg, useful to share, helpful to check health-related information) and 4 negative or neutral sentiments (eg, usually ignore such notifications, would not trust it) to reflect their perceptions about a mock pop-up that provided a link to e-cigarette–related information on social media.

Results: More participants endorsed positive sentiments than negative or neutral sentiments in relation to the pop-up. Specifically, 1078 (20.8%) endorsed two or more positive sentiments, and 690 (13.3%) endorsed two or more negative or neutral sentiments when searching for "vaping" on social media; similarly 913 (17.6%) participants endorsed two or more positive sentiments and 690 (13.3%) endorsed two or more negative or neutral sentiments when viewing e-cigarette–related posts in their feed (all $P<.001$). Among those who were searching for e-cigarette–related information, participants aged 13-18 years were more likely to endorse at least two positive sentiments compared to those aged 19‑24 years (ie, 646, 22.0% vs 423, 19.2%, respectively), those who had never used e-cigarettes compared with those who had ever used them (ie, 674, 23.6% vs 404, 17.3%, respectively), and those who last used e-cigarettes more than 30 days ago compared with those who had used them in the past 30 days (ie, 187, 19.8% vs 217, 15.6%, respectively). Similarly, among participants who viewed e-cigarette–related posts in their feed, those who had never used e-cigarettes were more likely to endorse two or more positive perceptions compared to those who had ever used e-cigarettes (ie, 563, 19.7% vs 350, 15.0%), and those who had last used e-cigarettes more than 30 days ago (ie, 159, 16.9% vs 191, 13.7%) were more likely to endorse two or more positive perceptions compared to those who had used them in the past 30 days (all $Ps<.001$). Participants who had never used e-cigarettes were also less likely to trust pop-ups, compared to those who had ever used them, whether while searching for e-cigarette-related information on social media or while seeing e-cigarette-related posts in their feed (approximately, 19%vs 24%). There were no significant differences in the endorsement of negative or neutral sentiments. However, participants aged 13‑18 years were more likely to ignore such notifications while viewing e-cigarette–related posts in their social media feed compared to those aged 19‑24 years (ie, 850, 28.9% vs 563 25%); further, participants who had never used e-cigarettes were less trusting of e-cigarette–related information while searching for vaping or in their feed.

Conclusions: Positive perception of a social media pop-up indicates its potential to prevent e-cigarette–related misinformation. Further development of a pop-up requires strategies to better engage and inform adolescents and young adults, specifically younger individuals, as they may be more likely to disregard pop-ups, and older individuals and those who used e-cigarettes in the past 30 days, as they are potentially more dismissive of such information.

(JMIR Form Res 2025;9:e73193) doi:10.2196/73193

KEYWORDS

electronic nicotine delivery systems; vaping; adolescents and young adults; social media; misinformation; prevention; intervention; acceptability; tobacco control

XSL•FO
RenderX

Introduction

E-cigarette–related misinformation is widely shared on social media, which can lower perceived health risks and undermine adolescents' and young adults' quit attempts [1]. As extensive social media use is associated with e-cigarette use among adolescents and young adults [2-4], and seeking information about e-cigarettes predicts future use [5], it is crucial that the information being accessed is accurate and up to date.

Currently, social media platforms combat e-cigarette–related misinformation by restricting branded content by influencers, requiring nicotine health warnings on advertisements [6] and removing reported content [7]. During the COVID-19 pandemic, Instagram implemented graphic pop-ups that directed users to websites of health authorities and other evidence-based resources, which helped connect people with trustworthy information [8]. To date, only one study assessed young adults' perceptions of links to e-cigarette cessation resources embedded within fictitious provaping influencer posts. Participants who viewed the link had lower odds of susceptibility to e-cigarette use, and higher odds of believing that the product would not cause positive experiences and would cause negative experiences, with no impact on intention to use or desire to quit using e-cigarettes [9]. This study expands upon research to inform social media users about e-cigarette use, by examining adolescent and young adult perceptions of the utility and acceptability of a pop-up that appears both while searching for e-cigarettes on social media and below e-cigarette–related posts, providing a link to information about the harms of tobacco and vaping from their social media feeds.

Methods

Participant Recruitment

Between December 2023 and March 2024, the study team administered a national, cross-sectional survey to an online panel of participants aged 13-24 years through Qualtrics. . Recruitment initially aimed for a 3:2 ratio of participants aged 13 - 17 and 18 - 24-years; however, this restriction was lifted on February 9, 2024, due to low response rates among younger participants. Sampling quotas were used to align sex and race/ethnicity distributions with the latest US Census data. Of 27,525 individuals who accessed the survey, 12,919 were ineligible, 8,064 were excluded for providing incomplete responses, and 1216 were excluded for failing quality checks (eg, completing the survey in under half the median time).

Sample

The participants (N=5326) had a mean age of 18.4 (SD 3.2) years; 3013 (56.6%) identified as female, and 3207 (60.2%) were non-Hispanic White. Approximately one-quarter (1418; 26.6%) of participants had used an e-cigarette in the past 30 days, 970 (18.2%) had used an e-cigarette but not in the past 30 days, and 2938 (55.2%) had never used an e-cigarette.

Ethical Considerations

The study was approved by Stanford University's School of Medicine IRB (Protocol Number IRB-54761). We collected e-consent from participants 18 years and above and assent from participants under 18 years for this anonymous survey (with a waiver of documentation of parental consent). Each participant could complete the survey only once and received up to US $5 or a panel-specific incentive.

Measures

E-cigarette use status was measured by asking participants whether they had ever used a nicotine e-cigarette, even one or two puffs. Response options included "yes, in the past 30 days," "yes, but not in the past 30 days," and "no."

Participants were shown two images of a fictitious pop-up tagged as "trusted resources" with a link to "vaping–related information" from the Centers for Disease Control and Prevention (CDC), World Health Organization (WHO), and United Nations Children's Fund (UNICEF) (Figure 1). They were then asked what they would think about seeing this pop-up appear when: (1) searching for vaping on social media ("Imagine you were searching for vaping on social media and a link came up like in the image shown below. What would you think?") and (2) looking at social media posts in their feed ("Imagine you were looking at social media posts and a link came up like in the image shown below. What would you think?"). Eight response options exhibiting 4 positive and 4 negative or neutral perceptions were provided; participants could select all that applied (see list of all perceptions in tabulated results). Additionally, participants had the option to select 'other' and write-in their own response. We also asked participants where they would prefer a link or trusted resource to appear.

XSL•FO
RenderX

Figure 1. Two images of fictitious social media pop-ups appearing when searching for information about "vaping" and when viewing posts in one's social media feed.

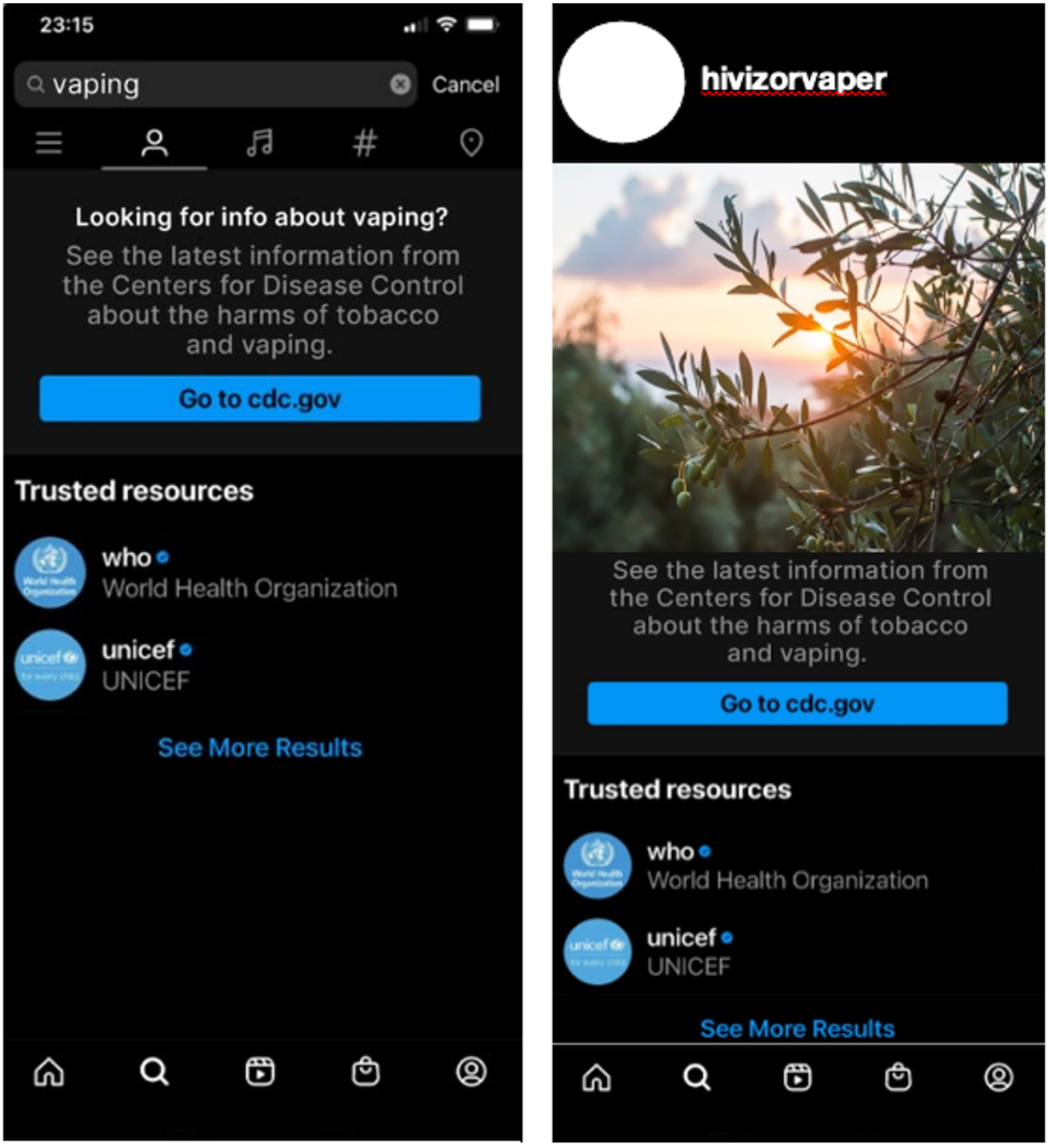

Analysis

Descriptive analyses examined the proportion of participants endorsing individual perceptions as well as two or more positive and negative or neutral sentiments. Chi-square analyses were used to examine differences in the proportion of participants endorsing individual perceptions as well as two or more positive and negative or neutral perceptions across age groups (13 - 18 y vs 19 - 24 y) and by e-cigarette use status (those who have never used vs those who have used; those who have used in the past 30 d vs those who used but not in the past 30 d). A significance level of $P<.05$ was used to determine if the proportions differed statistically between groups (ie, one group was more or less likely to endorse a particular perception than the other).

Results

Perceptions of a Social Media Pop-Up

Among participants who viewed a pop-up intended to appear while searching for information on e-cigarettes (n=5186), positive sentiments endorsed were: the pop-up was good for checking health-related information (24.9%), might help prevent misinformation (24.2%), provided useful information about e-cigarettes that is a click away (23.2%), and would help to share information with peers (15.1%) (Table 1). Negative or

XSL•FO
RenderX

neutral sentiments endorsed included: usually ignoring such notifications (27.3%), not trusting the pop-up (22.2%), perceiving that it would not help with the real problem that lies with sellers online or on social media (16.6%), and believed that it was for people who did not really know anything about vaping (15.7%). When asked about searching for information, 20.8% endorsed two or more positive sentiments and 13.3% endorsed two or more negative or neutral sentiments ($P<.001$).

XSL•FO
RenderX

Table . Participants endorsing positive and negative perceptions towards a pop-up, stratified by age and e-cigarette use status.

Endorsing perceptions towards the pop-up	Participants (N=5326), n (%)	Age (N=5326), n (%)		E-Cigarette use status (N=5326), n (%)		E-cigarette users (n=2388), n (%)	
		13 - 18 years (n =3037)	19 - 24 years (n=2289)	Never used (n=2938)	Ever used (n=2388)	Used >30 days ago (n=970)	Past 30-day use (n=1418)
While searching for "vaping" on social media (n=5186)							
Two or more positive perceptions	1078 (20.8)	646 (22.0)[a]	432 (19.2)	674 (23.6)[a]	404 (17.3)	187 (19.8)[a]	217 (15.6)
Two or more negative or neutral perceptions	690 (13.3)	391 (13.3)	299 (13.3)	375 (13.2)	315 (13.5)	112 (11.9)	203 (14.6)
Positive perceptions							
Good to check health-related information about vaping	1293 (24.9)	755 (25.7)	538 (24.0)	784 (27.5)[a]	509 (21.8)	203 (21.5)	306 (22.0)
May help prevent misinformation	1254 (24.2)	734 (25.0)	520 (23.2)	741 (26.0)[a]	513 (22.0)	236 (25.0)[a]	277 (19.9)
Useful information about vaping that is a click away	1203 (23.2)	719 (24.5)[a]	484 (21.6)	697 (24.5)[a]	506 (21.7)	217 (23.0)	289 (20.8)
Will help to share information with friends and others my age	785 (15.1)	448 (15.2)	337 (15.0)	435 (15.3)	350 (15.0)	148 (15.7)	202 (14.5)
Negative/neutral perceptions							
I usually ignore these notifications	1417 (27.3)	828 (28.2)	589 (26.2)	729 (25.6)[a]	688 (29.5)	259 (27.4)	429 (30.8)
I don't think I will trust it	1153 (22.2)	657 (22.3)	496 (22.1)	700 (24.4)[a]	453 (19.4)	198 (21.0)	255 (18.3)
It won't help with the real problem which is people who sell online or on social media	861 (16.6)	492 (16.7)	369 (16.4)	462 (16.2)	399 (17.1)	165 (17.5)	234 (16.8)
I think that's for people who really don't know anything about vaping	812 (15.7)	420 (14.3)[a]	392 (17.5)	420 (14.7)[a]	392 (16.8)	149 (15.78)	243 (17.5)
While looking at social media posts in their feed (n=5189)							
Two or more positive	913 (17.6)	533 (18.1)	380 (16.9)	563 (19.7)[a]	350 (15.0)	159 (16.9)[a]	191 (13.7)
Two or more negative/neutral	690 (13.3)	395 (13.4)	295 (13.1)	370 (13.0)	320 (13.7)	117 (12.5)	203 (14.6)
Positive perceptions							

XSL•FO
RenderX

Endorsing perceptions towards the pop-up	Participants (N=5326), n (%)	Age (N=5326), n (%)		E-Cigarette use status (N=5326), n (%)		E-cigarette users (n=2388), n (%)	
		13 - 18 years (n =3037)	19 - 24 years (n=2289)	Never used (n=2938)	Ever used (n=2388)	Used >30 days ago (n=970)	Past 30-day use (n=1418)
Good to check health-related information about vaping	1134 (21.9)	650 (22.1)	484 (21.5)	663 (23.2)[a]	471 (20.2)	192 (20.5)	279 (20.0)
May help prevent misinformation	1105 (21.3)	661 (22.5)	444 (19.8)	654 (22.9)[a]	451 (19.3)	194 (20.7)	257 (18.4)
Useful information about vaping that is a click away	1101 (21.2)	612 (20.8)	489 (21.8)	617 (21.6)	484 (20.7)	185 (19.7)	299 (21.4)
Will help to share information with friends and others my age	763 (14.7)	428 (14.6)	335 (14.9)	433 (15.2)	330 (14.1)	130 (13.8)	200 (14.3)
Negative/neutral perceptions							
I usually ignore these notifications	1413 (27.2)	850 (28.9)[a]	563 (25.0)	767 (26.9)	646 (27.7)	251 (26.7)	395 (28.3)
I don't think I will trust it	1144 (22.1)	669 (22.8)	475 (21.1)	695 (24.3)[a]	449 (19.2)	202 (21.5)[a]	247 (17.7)
It won't help with the real problem which is people who sell online or on social media	862 (16.6)	472 (16.1)	390 (17.4)	456 (16.0)	406 (17.4)	159 (16.9)	247 (17.7)
I think that's for people who really don't know anything about vaping	873 (16.8)	462 (15.7)[a]	411 (18.3)	428 (15.0)[a]	445 (19.1)	163 (17.4)	282 (20.2)

[a]*P* values significant at *P*=.05 level.

[b]χ^2 analyses exclude missing data.

[c]The maximum percentage missing across all groups was 2.63%.

[d]χ^2 analyses assessed differences in the proportion of participants endorsing two or more positive versus two or more negative/neutral perceptions. Participants were significantly more likely to endorse two or more positive perceptions than negative/neutral perceptions when searching for vaping information and looking at posts in the feed (both *P*<.0001).

While viewing e-cigarette–related posts, a pop-up intended to appear in their feed (n=5189), positive sentiments endorsed by participants included that the pop-up was good to check health-related information (21.9%), might help prevent misinformation (21.3%), provided useful information about e-cigarettes that was a click away (21.2%), and would help to share information with peers (14.7%). Negative or neutral sentiments endorsed were that they usually ignore such notifications (27.2%); would not trust it (22.1%); it would not help with the real problem, which is sellers online or on social media (16.6%); and that it is for people who did not really know anything about vaping (16.8%). While viewing e-cigarette-related posts, 17.6% endorsed two or more positive sentiments and 13.3% endorsed two or more negative or neutral sentiments (*P*<.001).

Differences in Perceptions Based on Age Group and E-Cigarette Use Status

While searching for "vaping" on social media, 13 - 18-year-old participants were more likely to endorse two or more positive perceptions compared to 19 - 24-year old participants (22.0% vs 19.2%, respectively). Furthermore, 13 - 18-years old participants were more likely to think the pop-up offered useful information just a click away (24.5% vs 21.6%) and less likely to believe that the pop-up was for people who did not know anything about vaping (14.3% vs 17.5%) compared to

XSL•FO
RenderX

19 - 24-year-old participants. Participants who had never used e-cigarettes compared to those who had ever used (23.6% vs 17.3%) and those with e-cigarette use>30 days ago compared to those with past-30-day use (19.8% vs 15.6%) were more likely to endorse two or more positive perceptions while searching for e-cigarette information. Participants who had never used e-cigarettes were significantly more likely to find pop-ups good for checking health-related information about e-cigarettes (27.5% vs 21.8%), helpful in preventing misinformation (26.0% vs 22.0%), and useful in providing information about e-cigarettes that was a click away (24.5% vs 21.7%) compared to those who ever used. Moreover, those who had never used e-cigarettes were more likely to say they would not trust the pop-up (24.4% vs 19.4%) and less likely to say they usually ignored notifications on social media (25.6% vs 29.5%) compared to those who had ever used them. Participants who used e-cigarettes >30 days ago were more likely to say that the pop-up may help prevent misinformation compared to those with past-30-day use (25.0% vs 19.9%, respectively).

While viewing social media posts in their feed, 13 - 18-year-old participants were more likely to say they usually ignored such notifications (28.9% vs 25.0%) and were less likely to think that the pop-up was for people who did not know anything about vaping (15.7% vs 18.3%) compared to 19 - 24-year-olds. Participants who had never used e-cigarettes compared to those who had used (19.7% vs 15.0%), and those with use >30 days ago compared to those with past-30-day use (16.9% vs 13.7%) were more likely to endorse two or more positive perceptions. Those who had never used e-cigarettes were more likely to say the pop-up was good to check health-related information about vaping (23.2% vs 20.2%) and may prevent misinformation (19.3% vs 22.9%) compared to those who have ever used. In addition, participants who never used e-cigarettes compared to those who had ever used (24.3% vs 19.2%) and those who had used>30 days ago compared to those with past-30-day use (21.5% vs 17.7%) were more likely to say they would not trust the pop-up.

Location Preference for Link or Trusted Resource

When asked where they would prefer a link or a trusted resource to appear (n=5,136), 25.2% participants indicated below the post but above comments, 20.7% at the top of the post with the post slightly greyed out, 19.6% as the next post, reel, or TikTok, 18.8% as a pop-up, and 15.8% below the post and comments.

Discussion

Overall, participants viewed the pop-up more positively than negatively or neutrally. The pop-up was viewed more favorably by those who had never used e-cigarettes than those who had used e-cigarettes, by those who used >30 days ago than current users, and by younger participants (under 18 y) compared to those older than them. One explanation is that adolescents and young adults who use e-cigarettes may be less likely to worry about associated health risks and therefore may not perceive the pop-up as beneficial or necessary [10].

Our finding that those who had never used e-cigarettes were less likely to trust the pop-up compared to those who had used, aligns with research demonstrating that noncigarette smokers were less likely to trust health information from online sources compared to current or former cigarette smokers [11]. Future research should assess if adolescents and young adults who have never used e-cigarettes are more likely to seek out information about vaping than those who had used them, which could potentially explain why they held specific positive perceptions.

Furthermore, older participants were less likely to have two or more positive perceptions compared to younger individuals. One explanation is that older adolescents may have more experience using social media, and as a result they may be less likely to click on pop-ups due to the association between pop-up windows and unwanted and inappropriate content, malware, spam advertisements, and high distractibility [1213]. Another potential reason why older participants were less likely to have positive perceptions of the pop-up is that because older adolescents and young adults (now aged 18 - 24 y) were in middle or high school during the peak of the youth e-cigarette epidemic in 2019, which may have likely resulted in increased exposure to e-cigarette-use behavior and normalized use or anti-e-cigarette and related messaging, and that they may have more entrenched beliefs about e-cigarettes. Thus, they may be less responsive to a pop-up providing additional health information about e-cigarettes in the present day.

One limitation of this formative study is that because the pop-up was presented hypothetically, it is unclear how participants would perceive the pop-up in real-time on social media. In addition, although roughly five percent of the participants did not answer the question or wrote-in their own response, it is possible that the list of response options did not exhaustively capture all possible perceptions.

Nonetheless, our study on adolescent and young adult perceptions can inform the potential development of a social media pop-up to prevent misinformation and e-cigarette use. Additional research is needed on how to reach out to adolescents and young adults on social media and strategies to make pop-ups or other notifications more favorable among adolescents and young adults, especially older individuals and those with past 30-day use who appear less receptive to information in such notifications. Further, it may be beneficial to co-design social media outreach with youth to ensure that e-cigarette–related information and notifications are not disregarded by younger individuals on social media. Moreover, examining factors that facilitate or inhibit trust in e-cigarette–related information on social media can provide information and recommendations for social media policy initiatives.

XSL•FO
RenderX

Acknowledgments

The research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health under Award Number R00CA267477. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the Food and Drug Administration. Ms. Maturo was supported by the Boston University School of Public Health Generation Health Award.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to potential reidentification risks, but are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

References

1. Silver N, Kierstead E, Tran B, Sparrock L, Vallone D, Schillo B. Belief and recall of nicotine as therapeutic for COVID-19 may undermine e-cigarette quitting behavior. Health Educ Behav 2022 Dec;49(6):929-933. [doi: 10.1177/10901981221109127] [Medline: 35848331]
2. Richardson MK, Agbonlahor O, Hart JL, Mattingly DT. Extensive social media use and frequency of current e-cigarette use among US youth. Drug and Alcohol Dependence Reports 2024 Sep;12:100266. [doi: 10.1016/j.dadr.2024.100266]
3. Pokhrel P, Fagan P, Herzog TA, et al. Social media e-cigarette exposure and e-cigarette expectancies and use among young adults. Addict Behav 2018 Mar;78:51-58. [doi: 10.1016/j.addbeh.2017.10.017] [Medline: 29127784]
4. Lee J, Krishnan-Sarin S, Kong G. Social media use and subsequent e-cigarette susceptibility, initiation, and continued use among US adolescents. Prev Chronic Dis 2023 Sep 7;20:E78. [doi: 10.5888/pcd20.220415] [Medline: 37676857]
5. Yang Q, Liu J, Lochbuehler K, Hornik R. Does seeking e-cigarette information lead to vaping? Evidence from a national longitudinal survey of youth and young adults. Health Commun 2019 Feb 23;34(3):298-305. [doi: 10.1080/10410236.2017.1407229]
6. Industry influencer: how tobacco content is infiltrating social media. Truth Initiative. 2023 Jul 20. URL: https://truthinitiative.org/research-resources/tobacco-pop-culture/industry-influencer-how-tobacco-content-infiltrating-social [accessed 2024-11-07]
7. Terms and policies: community guidelines. Instagram. URL: https://help.instagram.com/477434105621119/?helpref=uf_share [accessed 2024-11-07]
8. Rosen G. An update on our work to keep people informed and limit misinformation about COVID-19. Meta. 2020 Apr 16. URL: https://about.fb.com/news/2020/04/covid-19-misinfo-update/ [accessed 2024-11-07]
9. Vogel EA, Unger JB, Vassey J, Barrington-Trimis JL. Effects of a nicotine warning label and vaping cessation resources on young adults' perceptions of pro-vaping Instagram influencer posts. Addict Behav 2024 Feb;149:107888. [doi: 10.1016/j.addbeh.2023.107888] [Medline: 37857044]
10. Rohde JA, Noar SM, Horvitz C, Lazard AJ, Cornacchione Ross J, Sutfin EL. The role of knowledge and risk beliefs in adolescent e-cigarette use: a pilot study. Int J Environ Res Public Health 2018 Apr 23;15(4):830. [doi: 10.3390/ijerph15040830] [Medline: 29690606]
11. Rutten LJF, Blake K, Hesse BW, Ackerson LK. Isolated and skeptical: social engagement and trust in information sources among smokers. J Cancer Educ 2011 Sep;26(3):465-473. [doi: 10.1007/s13187-011-0201-7] [Medline: 21340632]
12. Potential risks of content, features, and functions: the science of how social media affects youth.: American Psychological Association; 2024 Apr. URL: https://www.apa.org/topics/social-media-internet/youth-social-media-2024 [accessed 2025-01-24]
13. Bozzola E, Spina G, Agostiniani R, et al. The use of social media in children and adolescents: scoping review on the potential risks. Int J Environ Res Public Health 2022 Aug 12;19(16):9960. [doi: 10.3390/ijerph19169960] [Medline: 36011593]

Abbreviations

CDC: Centers for Disease Control and Prevention

XSL•FO
RenderX

Edited by A Mavragani; submitted 27.02.25; peer-reviewed by EA Vogel, SP Uppu; revised version received 09.06.25; accepted 11.06.25; published 19.08.25.

Please cite as:
Maturo J, Gaiha SM
Adolescents' and Young Adults' Perceptions of a Pop-Up Aimed at Combating the Spread of E-Cigarette Misinformation on Social Media: Cross-Sectional Survey Study
JMIR Form Res 2025;9:e73193
URL: https://formative.jmir.org/2025/1/e73193
doi:10.2196/73193

© Jessica Maturo, Shivani Mathur Gaiha. Originally published in JMIR Formative Research (https://formative.jmir.org), 19.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Fast Healthcare Interoperability Resources (FHIR)–Based Interoperability Design in Indonesia: Content Analysis of Developer Hub's Social Networking Service

Lukman Heryawan[1], PhD; Yukiko Mori[2], MD, PhD; Goshiro Yamamoto[3], PhD; Naoto Kume[2], PhD; Lutfan Lazuardi[4], PhD; Anis Fuad[4], DEA; Tomohiro Kuroda[2], Prof Dr

[1]Department of Computer Science and Electronics, Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada, Sekip utara BLS 21, Sleman, Indonesia

[2]Division of Medical Information Technology and Administration Planning, Kyoto University Hospital, Kyoto University, Kyoto, Japan

[3]Preemptive Medicine and Lifestyle Related Disease Research Center, Kyoto University Hospital, Kyoto, Japan

[4]Health Policy and Management, Faculty of Medicine Public Health and Nursing, Universitas Gadjah Mada, Sleman, Indonesia

Corresponding Author:
Lukman Heryawan, PhD
Department of Computer Science and Electronics, Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada, Sekip utara BLS 21, Sleman, Indonesia

Abstract

Background: Interoperability in health care is a critical aspect for the exchange of health information. The Fast Healthcare Interoperability Resources (FHIR) framework has become widely adopted to provide interoperable data exchange in the health care industry. The COVID-19 pandemic has demonstrated the significance of interoperable data in tracking patients who have contracted the virus and keeping track of the vaccinated population. Indonesia is one of the many countries that have implemented interoperable data systems to track patients with COVID-19, and it has aspirations to expand the system to other use cases, particularly in the primary health care setting. The primary health care providers in Indonesia include *Puskesmas* (community health centers) and private clinics.

Objective: To promote interoperable health data exchange in the primary health care sector, the Indonesian government has launched the Satusehat project. The goal of the Satusehat platform is to make health data in Indonesia interoperable and exchangeable between health care organizations, particularly *Puskesmas* and private clinics.

Methods: For a successful implementation of the Satusehat platform in *Puskesmas* and private clinics, it is crucial to understand the challenges that may arise. This study analyzed the pain points of the Satusehat platform based on a content analysis of the Satusehat Social Networking Service Telegram group messages. The study revealed the pain points and suggested existing approaches to address them, which can be used as a proposed design of interoperability for *Puskesmas* and private clinics, making it easier for these organizations to adopt the Satusehat platform.

Results: The pain points identified in this study include issues with the FHIR server, problems with FHIR profile selection, and the mapping of electronic medical record data into standardized data, such as mapping into the Systematized Nomenclature of Medicine Clinical Terminology. The results show that the value of the mapping issue is 37, profile issue 9, and server issue 61. Among the 3 categories, server issues had the highest population, followed by mapping issues and then profile issues. To address these issues, the study proposed practical approaches, including a federated architecture for the FHIR server instead of a centralized architecture, an FHIR writer and FHIR viewer system inspired by the Standardized Structured Medical Record Information eXchange system in Japan, and an FHIR conversion framework that integrates with our FHIR writer and FHIR viewer system.

Conclusions: These proposed solutions can help resolve the pain points identified in the study and help the advancement of the Satusehat platform implementation in *Puskesmas* and private clinics in Indonesia. We believed that the proposed solutions have potential to be adopted to other countries with similar issues when conducting nationwide project in health care interoperability design.

(JMIR Form Res 2025;9:e51270) doi:10.2196/51270

KEYWORDS

interoperability; Fast Healthcare Interoperability Resources (FHIR); COVID-19; Satusehat; Indonesia; interoperability design; social networking; health information; content analysis; private clinic

XSL•FO
RenderX

Introduction

Background

Health information exchange (HIE) involves the transfer of electronic health data between different health information systems, including the electronic medical record (EMR) system, the electronic health record (EHR) system, and the personal health record (PHR) system. Notably, when the data are controlled by a health facility, an EMR system transitions into an EHR system. Conversely, if the data are managed by the patient or consumer themselves, it transforms into a PHR system [1]. To facilitate this exchange, HIE requires a data exchange or data interoperability framework that provides a standardized format for transferring health data [2]. One of the organizations that provides such interoperability frameworks is Health Level Seven (HL7) [3]. HL7 is a nonprofit organization that develops interoperability frameworks, such as HL7 FHIR (Fast Healthcare Interoperability Resources). It is a newer standard that provides a modern and flexible approach to exchanging health information. It uses a Representational State Transfer web services approach and supports both structured and unstructured data [4]. There have been several studies and reports on the implementation of EHR systems using the FHIR standard in different countries. One study, focused on the implementation of an FHIR-based EHR in Japan, highlighted some of the challenges faced during the implementation process, including lack of standardization and compatibility between different systems, difficulty of data migration, and the need for improved communication and cooperation between health care providers [5]. Another study provided the significance of standardization, data quality, and the need for technical support. The study also offered recommendations for improving the implementation and adoption of FHIR-based EHR systems [6,7].

The Ministry of Health (MoH) manages the EHR system in Indonesia, while the EMR system is independently managed by medical providers, such as hospitals, *Puskesmas* (community health centers), and private clinics. The PHR system, on the other hand, is provided by the MoH to the citizens of Indonesia [8]. In 2022, the expansion of EMR implementation to encompass all types of health care facilities was enacted through MoH decree number 24/2022. This regulation is a part of the digital health transformation blueprint released by the MoH [9].

A significant driver in the adoption of EMR, EHR, and PHR in Indonesia has been the COVID-19 pandemic. The PeduliLindungi app, developed as a tool for tracking, tracing, and monitoring during the pandemic, has become a crucial part of this initiative. This app has been used by more than 100 million users, making it interoperable with hundreds of laboratories and 17 telemedicine service providers. Most recently, through the MoH of the Republic of Indonesia decree number HK.01.07/MENKES/1559/2022, health data management is now under the MoH. Data from the PeduliLindungi app have been migrated to the Satusehat platform, as shown in Figure 1, which will integrate the PHR system and EMR system into a single nationwide EHR system [10]. The Satusehat Mobile platform enables the public to track COVID-19 vaccinations, immunization records, and medication information from EMRs of various health care facilities connected to this platform, using the FHIR HL7 standard. Furthermore, the EMR systems will be able to exchange and share data with ease through the Satusehat platform.

Legal basis for the implementation of the Satusehat system in Indonesia has been further strengthened with the enactment of the Health Law number 17/2023. This legislation specifically references the National Health Information System, which is interpreted as a system based on the Satusehat platform. The forthcoming derivative laws are expected to lay out the structural and operational details, ensuring a coherent and efficient implementation of these systems across the country.

The Satusehat platform represents the cornerstone of Indonesia's nationwide EHR system [11]. This platform aims to facilitate the exchange and sharing of health data, ultimately leading to better patient care and outcomes. In Indonesia, the Satusehat platform places a strong emphasis on primary health care. This is caused by the referral health care system in place, where most patients' health data begin with their first visit to a primary health care facility such as a *Puskesmas* or private clinic. These facilities are typically the first point of contact for patients seeking medical attention. If the primary health care facilities are unable to provide the necessary treatment or care, the patient will be referred to a higher level of care, such as a hospital. This referral system is depicted in Figure 2, which shows the progression of patients through different levels of care within the Indonesian health care system [12].

XSL•FO
RenderX

Figure 1. Satusehat platform's architecture. PeduliLindungi app as a PHR is integrated to FHIR server with view only. On the other hand, hospital information system with data from EMR could view or write with other health care applications. The FHIR server which is handled by the Ministry of Health acts as a bridge between health care digital instances. EMR: electronic health record; FHIR: Fast Healthcare Interoperability Resources; MoH: Ministry of Health; PHR: personal health record.

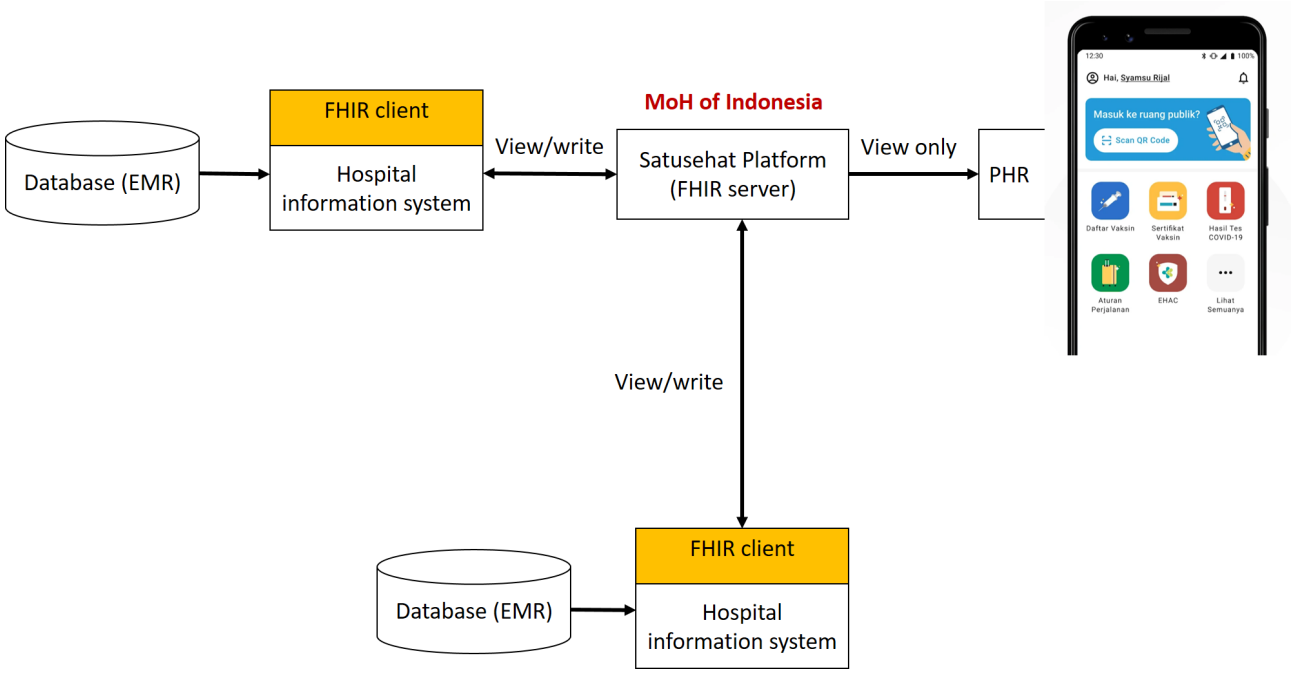

Figure 2. Indonesia's health care tiered referral system. Primary care facilities such as *Puskesmas* or private clinics as the base and entry point of health care service. Patient progress or referred to higher level, including secondary of tertiary facilities, if needed. This system is implemented with HL7 FHIR between tiers, thorough IT system, and regulation. HL7 FHIR: Health Level Seven Fast Healthcare Interoperability Resources.

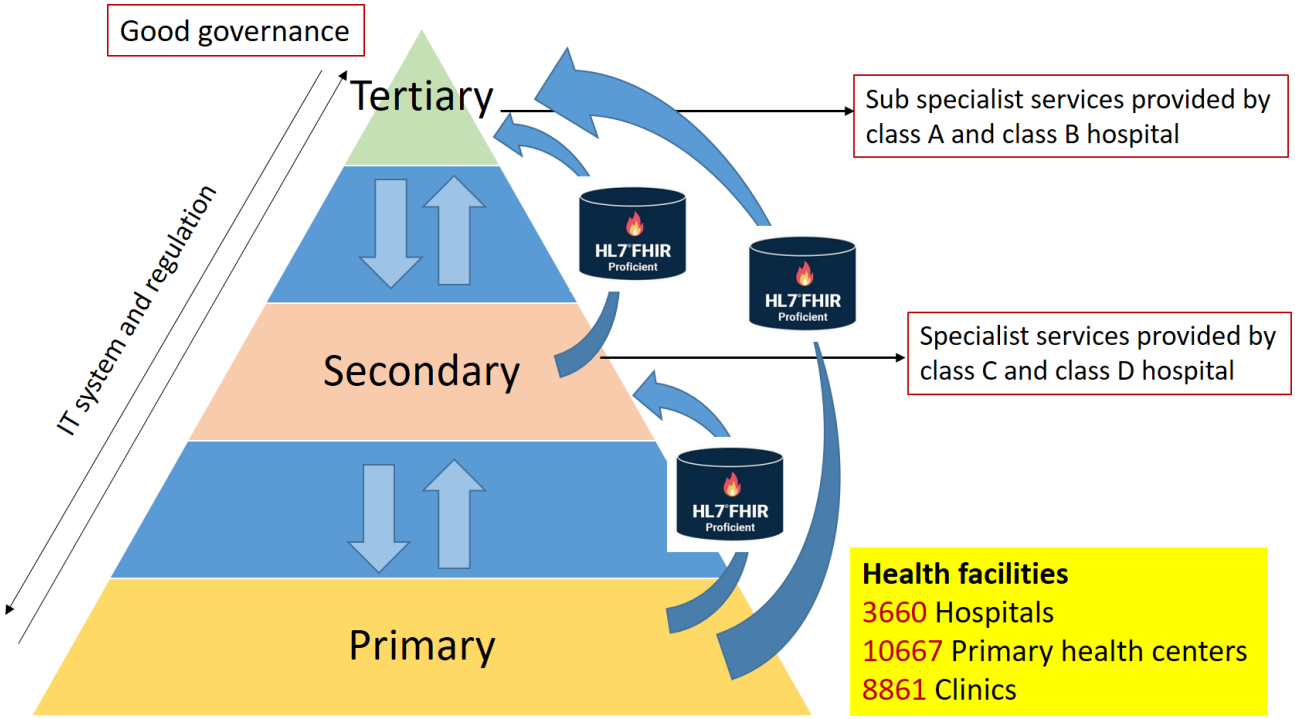

The Satusehat platform consists of 4 primary components: the FHIR server, the terminology server, the master data server, and the developer hub as shown in Figure 3. The FHIR server is responsible for providing an HL7 FHIR application programming interface (API) for health data exchange. The terminology server serves as a database, providing standardized data references for diagnoses using ICD-10 (*International Statistical Classification of Diseases, Tenth Revision*), observations using Logical Observation Identifiers Names and Codes (LOINC), and other clinical terms using Systematized Nomenclature of Medicine Clinical Terminology (SNOMED-CT). The master data server is a database that contains the nationwide identification numbers of patients or citizen ID, practitioner data, health facilities data, pharmacies data, financing data, and health service data. The developer hub provides a sandbox environment for developers to test the FHIR

XSL•FO
RenderX

API offered by the MoH. Developers can experiment and test data exchange scenarios, such as immunization, using the provided FHIR API. The developer hub also includes community tools, such as a video-sharing platform, a website, and a Telegram group, to provide support and assistance to developers while they test the Satusehat FHIR API [13].

Figure 3. Satusehat platform components that consist of FHIR server, terminology server, master data server, and developer hub. FHIR is responsible for health data exchange, terminology server provides diagnoses terminology standardization, master data stores patient and health care instances data (practitioners and infrastructures), and developer hub is for support center for developers when using FHIR application programming interface. HL7 FHIR: Health Level Seven Fast Healthcare Interoperability Resources.

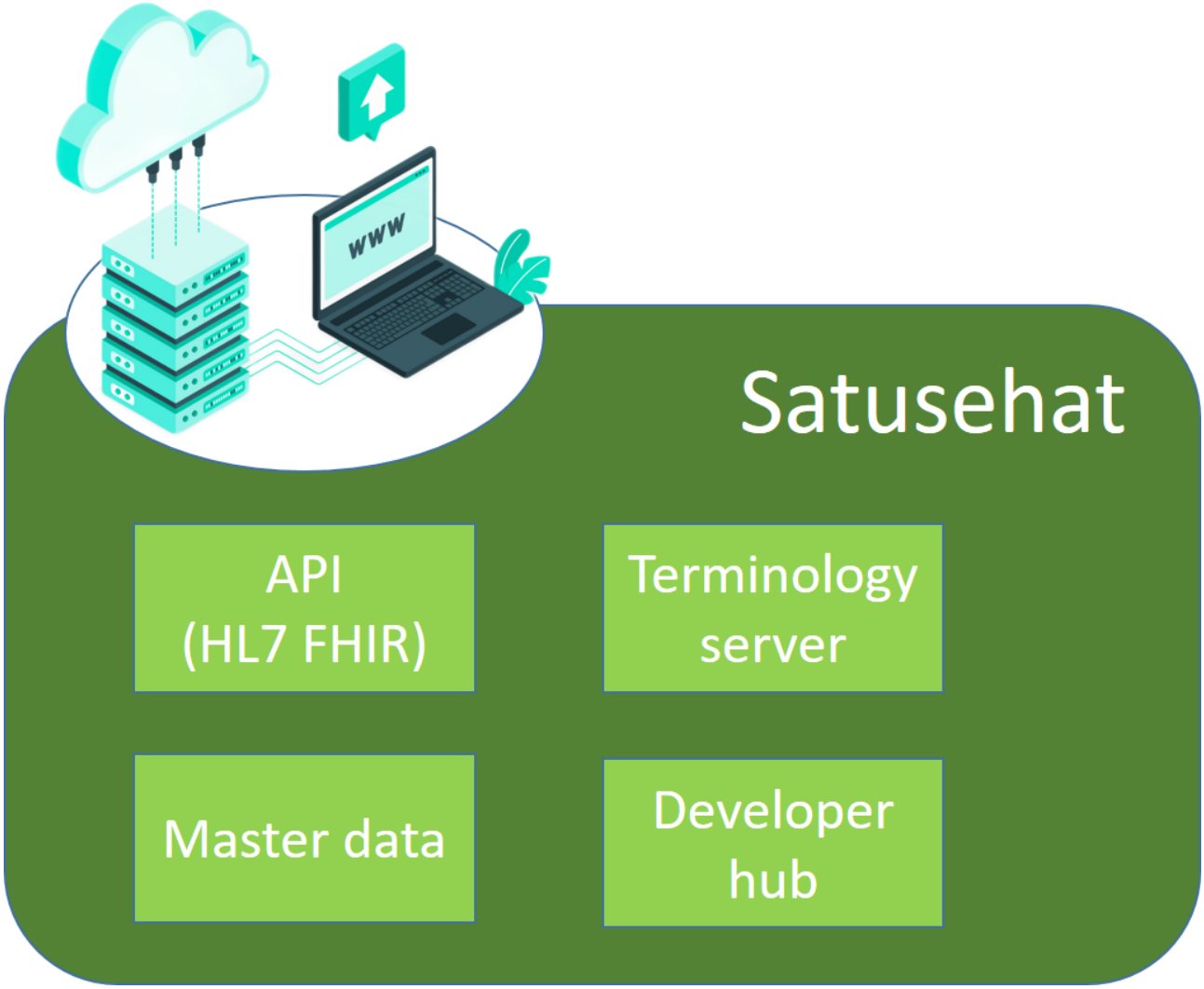

To ensure a successful implementation of the Satusehat platform in *Puskesmas* and private clinics, it is crucial to identify potential pain points in integrating EMR into the platform from a developer's perspective. Understanding the underlying architecture and structure of the FHIR standard is necessary for effective integration with EHR systems [14]. Another reason for pain points is the lack of resources, as implementing FHIR standards can be resource-intensive, requiring time and expertise. Developers may face constraints in terms of budget and man power, which can make it difficult to effectively implement the standard [15]. In addition, the lack of support from other stakeholders, including health care providers and technical experts, can also pose challenges for developers in implementing FHIR standards [16]. The Indonesian MoH is currently in the trial phase of the Satusehat platform. The Digital Transformation Office (DTO) of the MoH conducts workshops and training sessions for EMR system developers and vendors [17]. DTO also provides a website that provides step-by-step tutorials on how to connect and exchange data from EMR systems to the Satusehat platform using the FHIR framework, including the standardized terminologies used for mapping health data to standardized terminologies, such as LOINC for observations, *ICD-10* for diagnoses, and SNOMED-CT for other health terminologies. To provide interactive support, DTO also provides a Telegram group that developers can join to ask for support when integrating their EMR systems into the Satusehat platform. Based on the activities in the Telegram group, which mostly involve questions about how to integrate EMR systems into the Satusehat platform, we conclude that there are some pain points when integrating EMR systems into the Satusehat platform. This conclusion came from analysis through the Telegram group including a question and answer session with the DTO representative.

XSL•FO
RenderX

Objectives

The purpose of this study is to identify the challenges and difficulties encountered by developers that are present in situ during the integration process of EMR systems into the Satusehat platform. To accomplish this, we will analyze the communication content in the Telegram group of the Satusehat developer hub, which serves as an interactive community platform for developers. By exploring the pain points communicated by the developers in this forum, we aim to gain a deeper understanding of the barriers and obstacles encountered during the integration process. Once the pain points have been revealed, the study proposes practical approaches that can be used to ensure a successful integration into the Satusehat platform. These approaches are then combined to form an interoperability design of an FHIR-based system for EMR in *Puskesmas* and private clinics. The study is organized into 4 sections. The first section introduces the Satusehat platform, related work, and how pain points will be revealed. The second section describes the content analysis methods used in the study. The third section presents the results, and the fourth section discusses the practical approaches that will be proposed as an interoperability design to ensure a smoother integration into the Satusehat platform.

Methods

Study Design

To uncover the pain points that are addressed in the developer hub, we conducted a content analysis of the Telegram group's contents. We decided to conduct the analysis to the specific telegram group because the group is officially managed by the DTO. It is the most active group that contains developer discussions about EMR integration to the Satusehat platform. Content analysis is a research method used to make inferences by interpreting and coding written, spoken, or visual material [18]. The analysis was conducted at a Telegram group that consists of 1067 active members to this day. This study focused on analyzing conversations from December 20, 2022, the date the group was created, until January 29, 2023. The end date was chosen because the amount of extracted data is sufficient to support the experiment, which is content analysis. Extending the time span of chat extraction could potentially enhance the volume of data collected and strengthen the robustness of our conclusions, as the amount of data sampled in an experiment highly depends on the subject of the study [19]. In this instance, our focus was on Telegram as a social network service. A total of 219 chats were chosen during this period. However, we selected only 107 chats related to our research purpose. These chats primarily involved developers discussing issues encountered while integrating their EMR systems with the Satusehat platform using FHIR standards. In our research, we focused on issue-related chats, such as "http server error 500," "what FHIR profile should be used for chief complaints," and "SNOMED-CT or LOINC for mapping blood glucose." We excluded nonissue chats that contained simple responses such as "Ok" or "thank you." To collect the chats, we used the Telegram menu to export them into HTML files, which we later transferred to an Excel file. The collected chats were later manually labeled by 2 of the researchers (LH and ADW), who are trained to categorize the data according to agreed and specific rules for these issue-related chats. The labeler was chosen with taking the research design, the amount of data being analyzed, and the complexity of the coded classification process of the study into account. In addition, the coding process is ensured to be consistent and reliable with involving 2 researchers including the author and his colleagues (LH and ADW) with a background in information technology and understanding of the Satusehat platform. This established intercoder reliability in order to achieve valid interpretation of the data [20]. After conducting our content analysis and categorizing the issues, we developed a proposed approach for each issue category as shown in Figure 4. These proposed approaches are feasible incorporated into the existing Satusehat platform architecture, potentially resulting in an interoperability design that is less burdensome for developers. By less burdensome, we mean that the design can be easily implemented by novice developers, which are commonly found in private clinics and *Puskesmas* (community health centers). This design can be referred to as the FHIR-based Satusehat platform interoperability design for private clinics and *Puskesmas*. By reducing the burden of the interoperability design, the integration of EMR systems into the Satusehat platform in Indonesia will be smoother. It is important to note that there are other factors that can influence the smooth implementation beyond the technical scope this study. For the sake of the study's conciseness, we will focus only on the technical issues from a developer's perspective.

XSL•FO
RenderX

Figure 4. The approach method of the study. Given the developer hub, as the support and assistance forum for developers to test FHIR, the Telegram group chat content is extracted. Then, content analysis is conducted that resulted in categorization of issues and proposed solutions for each issues. API: application programming interface; HL7 FHIR: Health Level Seven Fast Healthcare Interoperability Resources.

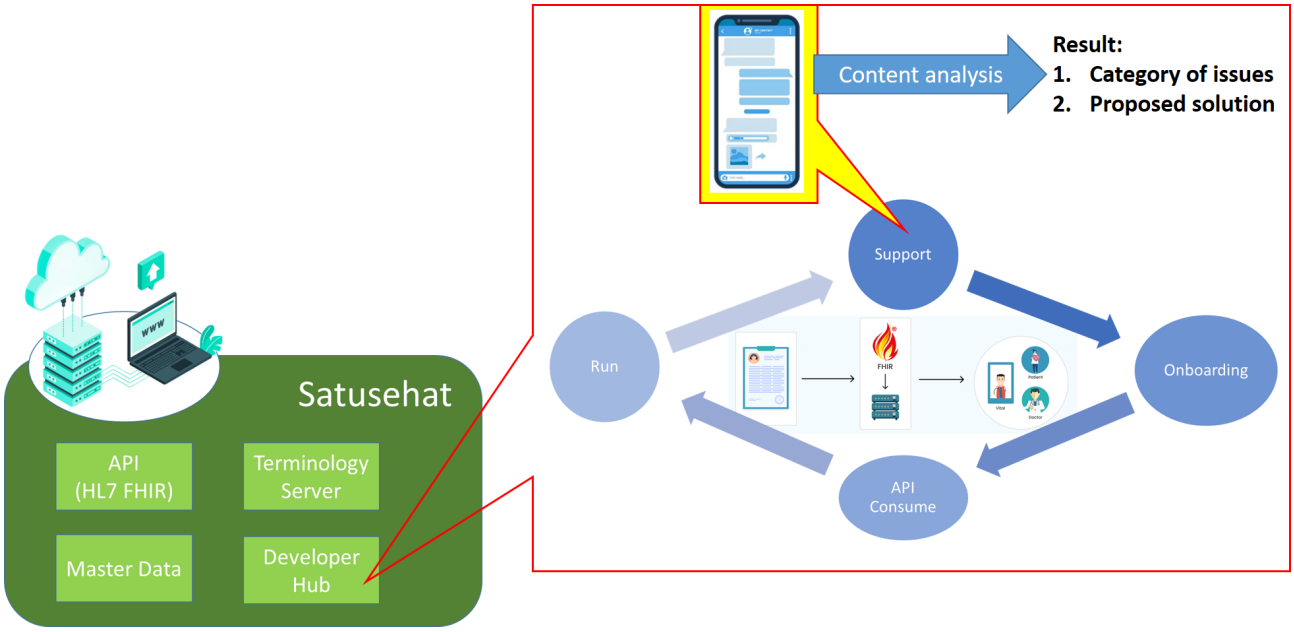

Ethical Considerations

Public Telegram group of the Satusehat developer hub contents was used for content analysis in this research. There were no any extra compensation given to conduct the analysis since the content itself is public. We acknowledge the ethical concerns that follows while analyzing data that may raise privacy issues. An original study relating to framework that can be used by researchers to consider ethical issues in web-based research also included how privacy policy and content that were analyzed on specific web-based platform should be reviewed. The collected data have to be accessible to public, focused to the content, and not the individual who posted the content [21]. In addition, taking into account the public nature of the group and also Telegram's privacy policy relating to public chats, we believe that group members implicitly consented to having their data accessible publicly [22]. Following the logics provided in the framework and Telegram's privacy policy, the exemption of additional ethics review is feasible for this study. This study was reviewed by the Medical and Health Research Ethics Committee Faculty of Medicine, Public Health and Nursing, Universitas Gadjah Mada-Dr. Sardjito General Hospital, Indonesia (KE/FK/0204/EC/2023).

The focus of this study is to analyze the content of the public developer hub and not on the individual who posted in the developer hub. Therefore, anonymization is applied to the data that are collected and presented in this study to ensure group participant's data privacy and confidentiality. According to other literatures, it is stated that waiver of consent could be requested when no identifying information is presented. Also, permission of further use of the data is permitted if the data are freely available on public forum [23]. Considering aforementioned points, acquiring consent was not necessary for this research. However, it is important to get consent in general for research using data from people. In our case, it was not possible to get informed consent due to the scope of our research.

Results

Overview

The results of our study are divided into 2 sections for clarity and organization. The first section outlines the pain points and issues with the Satusehat platform, which were identified through content analysis of the Satusehat Social Networking Service Telegram group messages. These pain points include issues with the FHIR server, difficulties in selecting the appropriate FHIR profile, and challenges in mapping EMR data into standardized data formats such as SNOMED-CT terminology [24]. Implementing and maintaining an FHIR server may be challenging due to the interoperability, security concerns, and scalability. Selecting the appropriate FHIR profile gives difficulties in aligning with specific use cases, balancing flexibility, and considering long-term maintenance. Mapping EMR data into standardized data formats such as SNOMED-CT could be challenging due to data complexity, granularity, updates, and achieving semantic interoperability. The second section of our study presents proposed designs that aim to address the pain points identified in the first section. We believe that the proposed designs will help overcome the challenges that may arise during the implementation of the Satusehat platform in primary health care settings, such as *Puskesmas* and private clinics. These proposed designs include a federated architecture for the FHIR server to ensure reliable and distributed server load management, an FHIR writer and viewer system inspired by the successful Standardized Structured Medical Record Information eXchange (SS-MIX) system in Japan [25], and an FHIR conversion framework that integrates with our FHIR writer and FHIR viewer system.

XSL•FO
RenderX

Pain Points

Based on content analysis of the Satusehat Social Networking Service Telegram group messages, we identified 3 main types of issues faced by developers: server issues, profile issues, and mapping issues as illustrated in Multimedia Appendix 1. Among the 3 categories, server issues had the highest population with 61 chats, followed by mapping issues with 37 identified chats and then profile issues with 9 chats. Our data collection method allowed us to focus on the issues and concerns that developers faced when integrating their EMR systems with Satusehat platform using FHIR standards.

Proposed Design

The Satusehat platform is built on a centralized architecture, which is illustrated in Figure 5. To ensure authentication, the platform uses an OAuth server [14]. Once a user is authenticated, they can access the FHIR resources available on the platform through the FHIR API. However, based on the developers' perspective, connecting to the Satusehat platform can be challenging. We have identified this issue as a server issue. Sometimes, the Satusehat platform server fails to respond to the user's request, resulting in errors such as HTTP Error 401—Unauthorized, HTTP status 403—Forbidden, HTTP status 404—Not Found, and HTTP status 500—Internal Server Error [26].

To overcome this issue, we propose a federated server architecture instead of a centralized server architecture [27], as shown in Figure 6. In this approach, the FHIR resources are stored in local or regional servers that are subordinate to the central server. Periodically, the central server inquiries about the FHIR resources and copies them to the central server. This approach ensures high reliability for both the central and local or regional servers, resulting in a more robust exchange of FHIR resources. By implementing the federated server architecture, the exchange of FHIR resources will be more reliable, making it easier for authenticated users to post, get, put, and patch the FHIR resources from the Satusehat platform.

The issue of mapping is a significant challenge faced by developers working on the Satusehat platform. Health care facilities in Indonesia use various local terminologies, which are often in Indonesian or local tribal languages, such as Javanese. In contrast, standardized terminologies such as *ICD-10* for diagnosis and *ICD-9-CM* (*International Classification of Diseases, Ninth Revision, Clinical Modification*) for medical procedures are in non-Indonesian languages, such as English [28]. This presents a significant challenge in mapping the various local terminologies with their many variations into a single standardized terminology in English. We believe that this issue is not only unique to Indonesia but also affects other countries whose national language is not English, such as Japan, Taiwan, and Thailand [29]. To overcome this challenge, we propose an approach that is similar to the one used in Japan. The Japanese approach uses the FHIR conversion framework [30]. In this approach, the conversion is done only on the terminology code and not on the terminology text. The written terminology still uses the local language, such as Japanese, but the code follows the standardized code. This approach reduces the complexity of the mapping task significantly. The FHIR conversion framework is illustrated in Figure 7. By implementing this approach, we can ensure that the Satusehat platform can handle the mapping of various local terminologies into standardized terminologies in a more efficient and effective way.

In addition, automating the code conversion process is much simpler than the text conversion process because the variations in code are relatively small. This approach not only reduces the workload for developers but also ensures the accuracy and consistency of the mapping process. The last issue concerns the challenge of selecting and applying the correct FHIR profiles to EMR data. While there is an implementation guide available to guide users in selecting the appropriate profiles for specific use cases, such as outpatient medical resumes and immunization records, manually selecting and applying the profiles can still be difficult. To address this issue, we propose a solution that involves feeding EMR data into a system that automatically selects the correct FHIR profile and builds the content of that profile based on the input data. Similar approaches have been implemented in countries that have standardized their health data using HL7 standards, such as Japan's use of HL7 V 2.5 to standardize various health data into a storage system called the SS-MIX [28]. The SS-MIX system has been successful in facilitating HIE in Japan. Since both SS-MIX and FHIR are from HL7, we believe that it would be feasible to adopt the SS-MIX system and modify it to use HL7 FHIR structures instead of HL7 V 2.5 structures. This is a practical solution as there are already open structures and guidelines available to implement SS-MIX. We simply need to modify the structure to use HL7 FHIR data.

XSL•FO
RenderX

Figure 5. Satusehat current architecture. The platform was built on centralized architecture and database [27]. Authenticated institutions or individuals are able to access the Fast Healthcare Interoperability Resources in the central database. Writer identifies this as the root of server issue stated in developer hub.

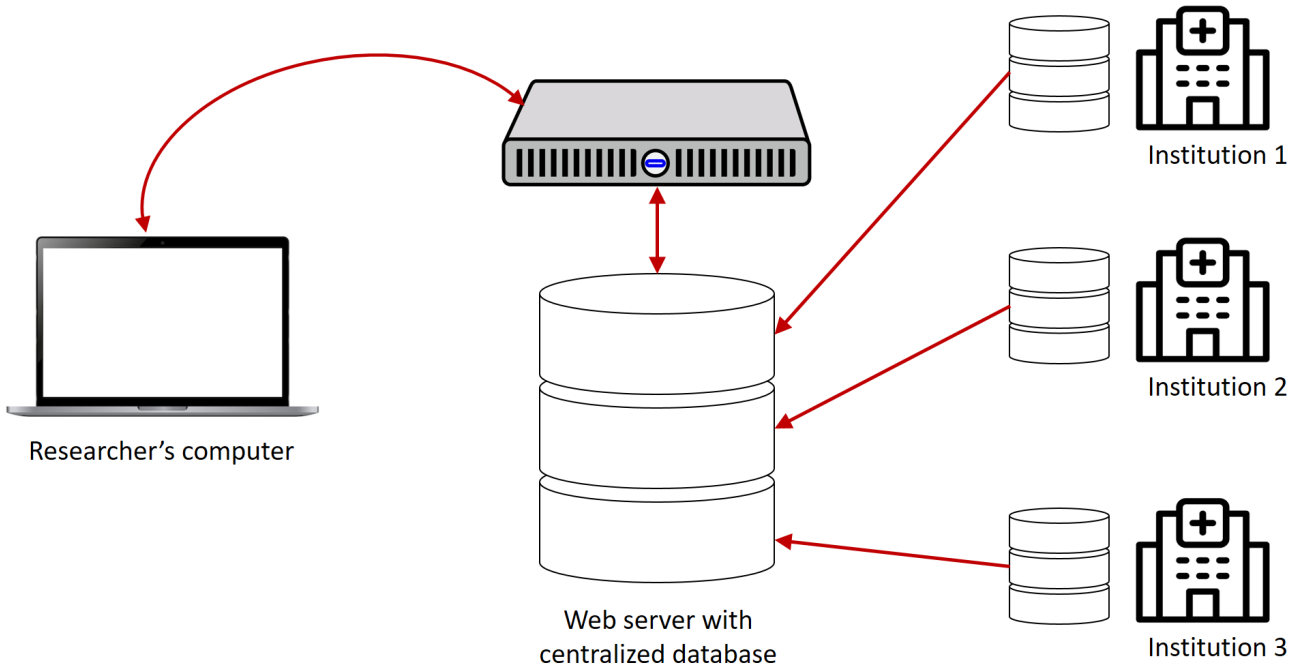

Figure 6. Our proposed solution: Satusehat with federated architecture [27]. The central server periodically fetch and add Fast Healthcare Interoperability Resources that are stored in local servers with the aim to enhance reliability and robustness of the platform.

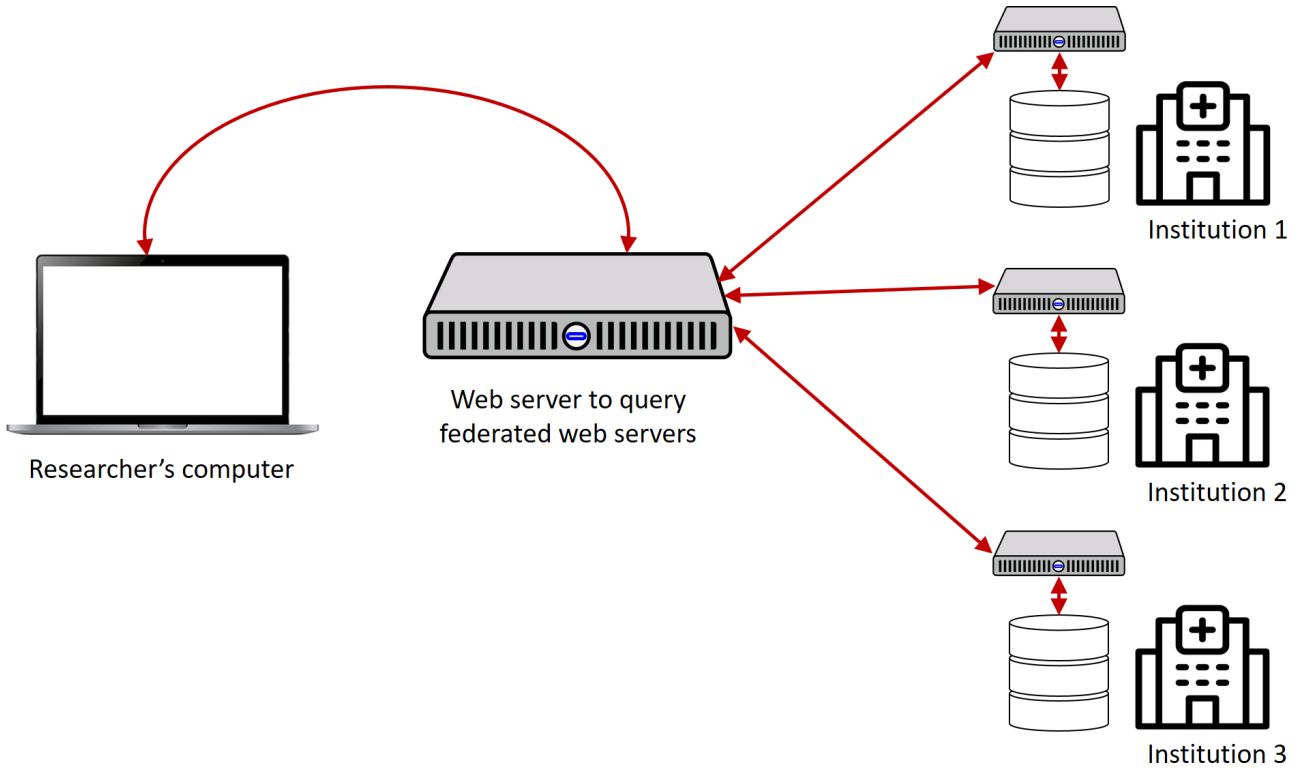

XSL•FO
RenderX

Figure 7. FHIR conversion framework to address mapping issues [14]. Japanese terminology for medication data, in the name attribute (2mg.CERCINE TABLETS) is preserved in hospital EMR to ensure readability. The corresponding standardized code (such as *ICD-10* [*International Statistical Classification of Diseases, Tenth Revision*]) is mapped into the FHIR server with terminology mapping. EMR: electronic medical record; FHIR: FHIR: Fast Healthcare Interoperability Resources

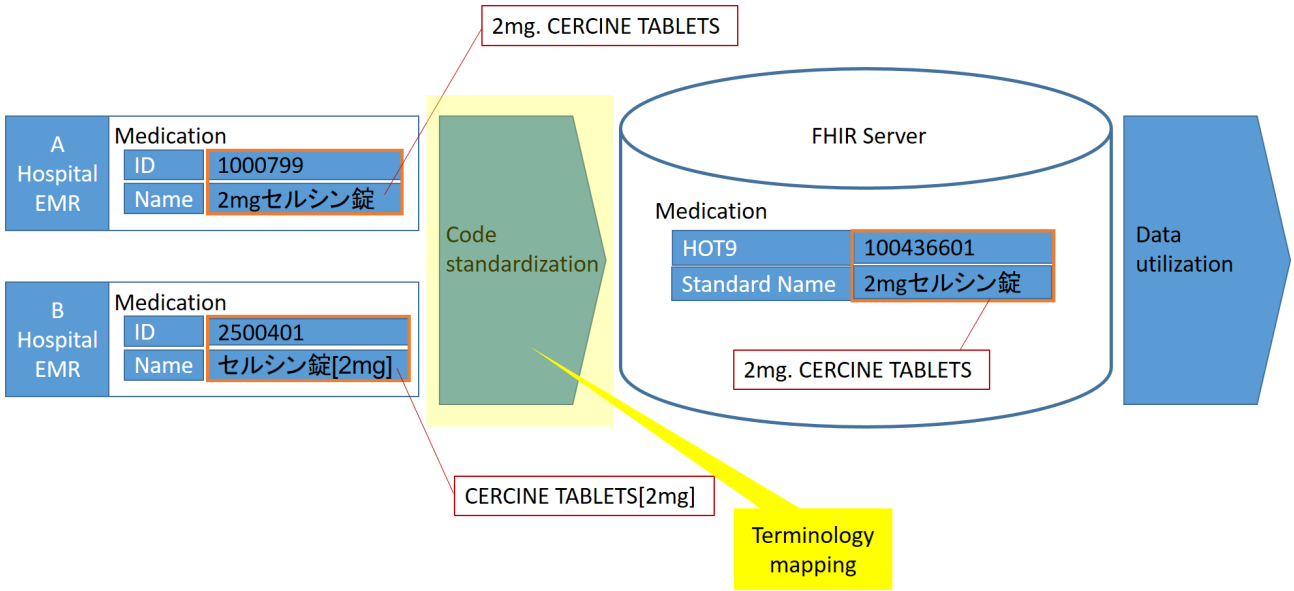

Discussion

Principal Findings

Server issue, mapping issue, and profile issue are identified as the most frequent problems that developers faced when integrating EMR and Satusehat with FHIR protocol. The proposed design for an interoperable FHIR-based Satusehat platform tailored to the needs of *Puskesmas* (community health centers) and private clinics is depicted in Figure 8, which illustrates the combined approaches that have led to a more reliable and user-friendly platform. The design integrates several key features, including a federated architecture that distributes the data, a data mapping system that ensures consistency across diverse health care terminologies, and an automated selection and implementation of appropriate FHIR profiles for EMR data.

Figure 8. Our proposed design for FHIR-based Satusehat platform. It consists of federated architecture between central server and hospital information system in local or regional server as FHIR client, local terminology system mapping, and automated FHIR profile selection. Fast Healthcare Interoperability Resources; IHS: Indonesia Health Services; PHR: personal health record.

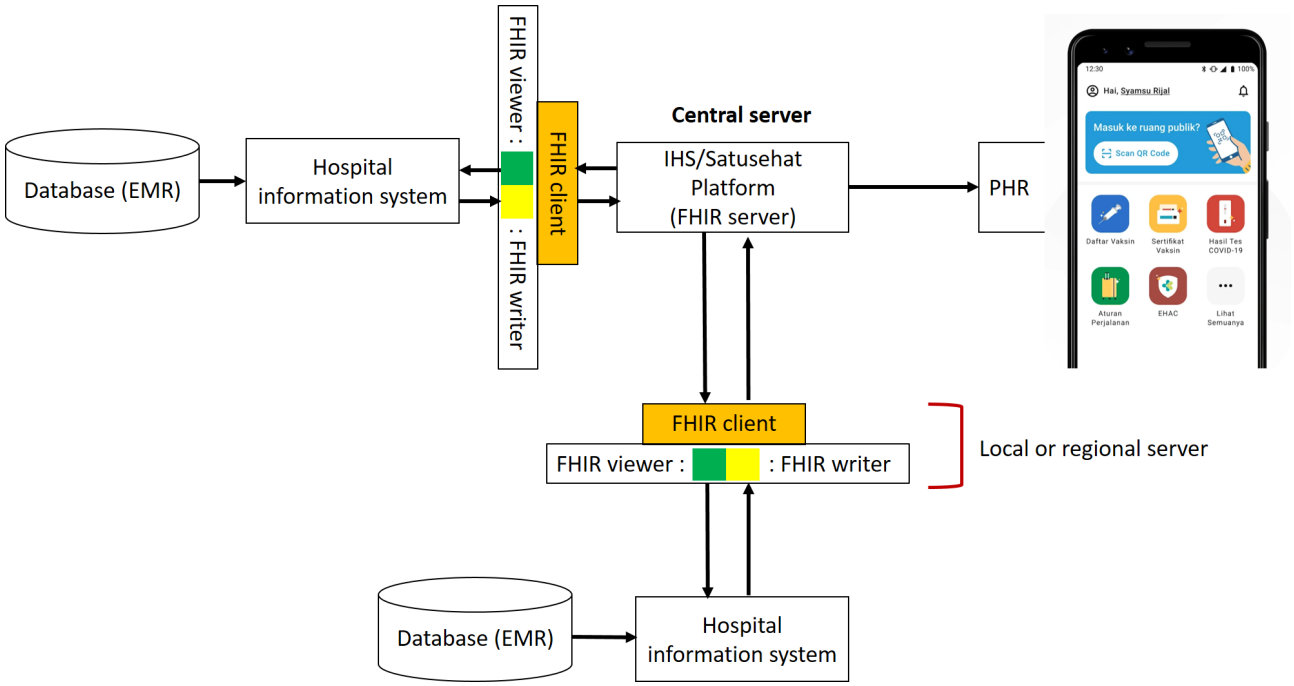

To address the mapping and profile issues related to data, we propose an FHIR writer and viewer system that combines the FHIR conversion framework and the SS-MIX system. This system can be deployed on both local and regional servers to ensure accessibility for health care facilities such as *Puskesmas*, hospitals, and private clinics. To use this system, hospital

XSL•FO
RenderX

information systems or EMR systems feed data into the FHIR writer subsystem. The FHIR writer then converts the local code into standardized code and selects the appropriate profile to be inserted with the standardized data. The FHIR profile, filled with standardized data, is then stored or prepared in the local or regional server and is ready to be queried by a central server. If health care facilities want to view patient data from their own facilities, their EMR system can make a request to the local or regional server to view the patient data. If the data are not available in the local or regional server, the server continues the request to the central server to retrieve the data using FHIR API. In this way, the local or regional server acts as an FHIR client, and the central server acts as an FHIR server in terms of the client-server connection [29]. Our proposed FHIR writer and viewer system provides a standardized approach to data mapping and FHIR profile selection, ensuring consistency and ease of use for health care facilities. It also allows for efficient querying of patient data across multiple facilities, increasing the interoperability and accessibility of health data. The issue of server architecture is a critical consideration in developing an efficient and reliable health information system. In particular, the ability to handle large volumes of data and ensure seamless data exchange across different health care providers is crucial. The approach to addressing this challenge lies in the use of federated architecture. Federated architecture offers a scalable and flexible solution to server issues by distributing the load across multiple servers [30]. This ensures that the server load is manageable even during peak periods, such as when there is an automatic processing of EMR data. By spreading the load, federated architecture also enhances the reliability of the system, making it more robust and resilient. Moreover, federated architecture is an ideal solution for countries such as Indonesia, which has a vast archipelago with numerous islands separated by sea. With federated architecture, health facilities in remote areas can be connected to a centralized health information system, enabling the efficient exchange of health data and improving patient care. To fully leverage the benefits of federated architecture, an automated FHIR writer or viewer system can be implemented. This system enables the automatic conversion of local codes to standardized codes, the selection of appropriate FHIR profiles, and the storage and retrieval of patient data from local and regional servers. The system can be easily deployed and customized to suit the specific needs of *Puskesmas*, private clinics, and hospitals. By using this approach, health care providers can efficiently access patient data, resulting in improved health care outcomes for patients. Although our proposed design is based on a small amount of chats in the Satusehat Telegram group, we recognize that this group is the most interactive channel for developers seeking support in integrating their EMR systems into the Satusehat platform. Therefore, despite the limited data, we believe that most of the issues related to integrating EMR systems into Satusehat using the FHIR framework have been recorded in the Satusehat Telegram group. By analyzing and addressing these issues, we were able to develop a design that effectively addresses the common challenges faced by developers. We also recognize that ongoing monitoring and analysis of the Satusehat Telegram group will be critical for identifying and addressing new issues that arise as the platform continues to evolve. Our proposed design for the FHIR-based Satusehat platform can be a valuable solution for other countries facing similar challenges in implementing HIE systems. While we drew inspiration from the design of Japan's HIE system, which shares similar characteristics with Indonesia, such as a non-English language and multiple islands, our approach is adaptable and scalable to meet the needs of other countries. By collaborating and sharing knowledge, countries facing similar issues can learn from each other to enhance the implementation of FHIR-based platforms such as Satusehat. This approach could foster the development of a regional-level HIE system in ASEAN countries or even beyond, incorporating other Asian countries such as Japan, Taiwan, and India, which have already implemented HIE systems [31]. By working together, we are able to create a more efficient and effective HIE system that benefits all participating countries.

Limitations

The content analysis in the study was limited into specific time frame (December 20, 2022, to January 29, 2023) and a small size sample of 107 chats. It was originated from the developer hub's Telegram group that was managed by the DTO of Indonesia MoH. The telegram also addresses only the technical pain points from developers' perspective to integrate their EMR to Satusehat platform. Also, the proposed approaches based on the identified pain points have not been implemented to the existing Satusehat platform.

Future Works

After gaining an understanding of the challenges faced by the Satusehat platform and proposing an interoperability design for the platform, our next step is to develop an FHIR writer or viewer system that is tailored to the Satusehat platform. To accomplish this, we plan to collaborate with an EMR system developer for the purpose of collecting EMR data. These data will be used to train and test the FHIR writer or viewer system, to determine whether the EMR data can be standardized based on the FHIR-based Satusehat platform, and to determine whether it can be viewed back by the same EMR system. We are also interested in collaborating with researchers from Japan, whose approach to building the SS-MIX system has inspired our proposed design. By working together, we aim to create an FHIR-compliant version of SS-MIX that can not only be used in the Satusehat platform in Indonesia but also has the potential to be implemented in Japan. With this joint collaboration, the FHIR-based SS-MIX can facilitate cross-country border HIE by ensuring interoperability and seamless exchange of health data between the 2 countries.

Conclusions

For ensuring successful implementation of the Satusehat platform, this paper proposes content analysis as a way to identify pain points that are faced by developers during development. Analyzing Satu Sehat developers' discussions on social networking system, we have identified 3 top pain points about Satusehat platform: server issues, data mapping, and difficulties managing user profiles. We have showed how using federated server architecture, an FHIR writer and an FHIR viewer, and an FHIR conversion framework could address the

XSL•FO
RenderX

mentioned issues. The proposed architecture for solving pain points gives several suggestions as a blueprint on how to reduce developer workload when committing integration, increasing the robustness of the resources exchange, and ensure accuracy and consistency on mapping and also profiling process. All of this will further help the advancement of Satusehat platform implementation in *Puskesmas* (community health centers) and private clinics in Indonesia.

Acknowledgments

This research was conducted without the use of ChatGPT or any similar generative artificial intelligence model for ideation, method execution, manuscript writing, and referencing process. This work is the result of collaborative works and within the authors. The information and sources that are included in this work have been carefully checked for accuracy. This research was made possible by the collaboration with Medical informatics Lab in Kyoto University, Japan, and Software and Data Engineering Lab in Universitas Gadjah Mada, Indonesia. This work was supported by the Department of Computer Science and Electronics, Universitas Gadjah Mada, under the Publication Funding Year 2025.

Data Availability

The datasets generated during and/or analyzed during this study are not publicly available due to the Satusehat privacy policy but are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Most prominent pain points perceived by developers when testing Fast Healthcare Interoperability Resources from the Telegram group (developer hub).
[PNG File, 29 KB - formative_v9i1e51270_app1.png]

References

1. Sonoda T. Evolution of electronic medical record solutions. Fujitsu Sci Tech J 2011 Jan 1;47:19-27.
2. Ishigure Y. Trends, standardization, and interoperability of healthcare information. NTT Tech Rev 2011;9(4):57-62. [doi: 10.53829/ntr201104gls]
3. Joyia GJ, Akram MU, Akbar CN, Maqsood MF. Evolution of health level-7: a survey. Presented at: Proceedings of the 2018 International Conference on Software Engineering and Information Management; Jan 4, 2018; Casablanca, Morocco p. 118-123. [doi: 10.1145/3178461.3178480]
4. Mandel JC, Kreda DA, Mandl KD, Kohane IS, Ramoni RB. SMART on FHIR: a standards-based, interoperable apps platform for electronic health records. J Am Med Inform Assoc 2016 Sep;23(5):899-908. [doi: 10.1093/jamia/ocv189] [Medline: 26911829]
5. Xiao D, Song C, Nakamura N, Nakayama M. Development of an application concerning fast healthcare interoperability resources based on standardized structured medical information exchange version 2 data. Comput Methods Programs Biomed 2021 Sep;208:106232. [doi: 10.1016/j.cmpb.2021.106232] [Medline: 34174764]
6. Faulkenberry JG, Luberti A, Craig S. Electronic health records, mobile health, and the challenge of improving global health. Curr Probl Pediatr Adolesc Health Care 2022 Jan;52(1):101111. [doi: 10.1016/j.cppeds.2021.101111] [Medline: 34969611]
7. Dullabh P, Heaney-Huls K, Hovey L, et al. The technology landscape of patient-centered clinical decision support—where are we and what is needed? Stud Health Technol Inform 2022 Jun 6;290:350-353. [doi: 10.3233/SHTI220094] [Medline: 35673033]
8. Harahap NC, Handayani PW, Hidayanto AN. The challenges in integrated personal health record adoption in Indonesia: a qualitative analysis of regulatory perspectives. Presented at: 2021 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS); Oct 2021; Jakarta, Indonesia p. 169-174. [doi: 10.1109/ICIMCIS53775.2021.9699353]
9. Ministry of Health of the Republic of Indonesia. Blueprint for digital health transformation strategy 2024. 2021. URL: https://dto.kemkes.go.id/ENG-Blueprint-for-Digital-Health-Transformation-Strategy-Indonesia%202024.pdf [accessed 2024-02-25]
10. Chong HY, Allotey PA, Chaiyakunapruk N. Current landscape of personalized medicine adoption and implementation in Southeast Asia. BMC Med Genomics 2018 Oct 26;11(1):94. [doi: 10.1186/s12920-018-0420-4] [Medline: 30367635]
11. Saputra Y, Ashila MN, Muliarini P. Readiness and acceptance of electronic medical records among health professionals in indonesia. Presented at: Proceedings of International Conference of Graduate School on Sustainability; Nov 4-5, 2022; Malang, Indonesia p. 81-92.

XSL•FO
RenderX

12. Handayani PW, Saladdin IR, Pinem AA, Azzahro F, Hidayanto AN, Ayuningtyas D. Health referral system user acceptance model in Indonesia. Heliyon 2018 Dec;4(12):e01048. [doi: 10.1016/j.heliyon.2018.e01048] [Medline: 30603690]
13. Postman. URL: https://www.postman.com/satusehat/satusehat-public/overview [accessed 2023-02-19]
14. Lee YL, Lee HA, Hsu CY, Kung HH, Chiu HW. Implement an international interoperable PHR by FHIR—a Taiwan innovative application. Sustainability 2020;13(1):198. [doi: 10.3390/su13010198]
15. Harbi A. Health care expert's readiness to implement National Unified Medical Records (NUMR) system in the United Arab Emirates: a qualitative study. IJCAI 2021;45(5). [doi: 10.31449/inf.v45i5.3358]
16. Bender D, Sartipi K. HL7 FHIR: an agile and restful approach to healthcare information exchange. Presented at: Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems; Jun 2013; Porto, Portugal p. 326-331.
17. DTO kementerian kesehatan—home page. dto.kemkes.go.id. URL: https://dto.kemkes.go.id/ [accessed 2023-02-19]
18. Renz SM, Carrington JM, Badger TA. Two strategies for qualitative content analysis: an intramethod approach to triangulation. Qual Health Res 2018 Apr;28(5):824-831. [doi: 10.1177/1049732317753586] [Medline: 29424274]
19. Kim H, Jang SM, Kim SH, Wan A. Evaluating sampling methods for content analysis of Twitter data. Soc Media + Soc 2018 Apr;4(2):205630511877283. [doi: 10.1177/2056305118772836]
20. O'Connor C, Joffe H. Intercoder reliability in qualitative research: debates and practical guidelines. Int J Qual Methods 2020 Jan 1;19:160940691989922. [doi: 10.1177/1609406919899220]
21. Cilliers L, Viljoen K. A framework of ethical issues to consider when conducting internet-based research. SA J Inform Manage 2021;23(1):a1215. [doi: 10.4102/sajim.v23i1.1215]
22. Telegram. Telegram privacy policy. URL: https://telegram.org/privacy/id?setln=en [accessed 2024-04-23]
23. Tripathy JP. Secondary data analysis: ethical issues and challenges. Iran J Public Health 2013 Dec;42(12):1478-1479. [Medline: 26060652]
24. Döring M. Evaluating an approach for mapping FHIR profiles to research protocols. Doctoral dissertation.: Hochschulbibliothek der Technischen Hochschule Köln; 2018.
25. Japan Association for Medical Informatics. SS-MIX2 standardized storage' explanation of the structure and guidelines for implementation ver. 1.2. 2014. URL: https://www.jami.jp/jamistd/docs/SSMIX2/descript-implemglonSS-MIX2_V1.2.pdf [accessed 2023-02-19]
26. TIBCO. TIBCO product documentation. docs.tibco.com. URL: https://docs.tibco.com/pub/bwpluginhl7/8.1.0/doc/html/GUID-20E40FAF-6F0B-41D6-8257-3387607178DA.html [accessed 2023-02-19]
27. Vesteghem C, Brøndum RF, Sønderkær M, et al. Implementing the FAIR Data Principles in precision oncology: review of supporting initiatives. Brief Bioinform 2020 May 21;21(3):936-945. [doi: 10.1093/bib/bbz044] [Medline: 31263868]
28. Quan H, Li B, Saunders LD, et al. Assessing validity of ICD-9-CM and ICD-10 administrative data in recording clinical conditions in a unique dually coded database. Health Serv Res 2008 Aug;43(4):1424-1441. [doi: 10.1111/j.1475-6773.2007.00822.x] [Medline: 18756617]
29. Aljunid SM, Srithamrongsawat S, Chen W, et al. Health-care data collecting, sharing, and using in Thailand, China mainland, South Korea, Taiwan, Japan, and Malaysia. Value Health 2012;15(1 Suppl):S132-S138. [doi: 10.1016/j.jval.2011.11.004] [Medline: 22265060]
30. Hisashi Osanai—effort to internationalize a FHIR implementation guide in Japan | devdays. www.youtube.com. 2022. URL: https://www.youtube.com/watch?v=HwAxcdKfbxY [accessed 2023-02-19]
31. Dornan L, Pinyopornpanish K, Jiraporncharoen W, Hashmi A, Dejkriengkraikul N, Angkurawaranon C. Utilisation of electronic health records for public health in Asia: a review of success factors and potential challenges. Biomed Res Int 2019;2019:7341841. [doi: 10.1155/2019/7341841] [Medline: 31360723]

Abbreviations

API: application programming interface
DTO: Digital Transformation Office
EHR: electronic health record
EMR: electronic medical record
FHIR: Fast Healthcare Interoperability Resources
HIE: health information exchange
HL7: Health Level Seven
***ICD-10*:** *International Statistical Classification of Diseases, Tenth Revision*
***ICD-9-CM*:** *International Classification of Diseases, Ninth Revision, Clinical Modification*
LOINC: Logical Observation Identifiers Names and Codes
MoH: Ministry of Health
PHR: personal health record
SNOMED-CT: Systematized Nomenclature of Medicine Clinical Terminology
SS-MIX: Standardized Structured Medical Record Information eXchange

XSL•FO
RenderX

Edited by A Mavragani; submitted 27.07.23; peer-reviewed by B Senst, G Grieve, J Frimpong; revised version received 20.05.24; accepted 26.05.24; published 21.04.25.

Please cite as:
Heryawan L, Mori Y, Yamamoto G, Kume N, Lazuardi L, Fuad A, Kuroda T
Fast Healthcare Interoperability Resources (FHIR)–Based Interoperability Design in Indonesia: Content Analysis of Developer Hub's Social Networking Service
JMIR Form Res 2025;9:e51270
URL: https://formative.jmir.org/2025/1/e51270
doi:10.2196/51270

© Lukman Heryawan, Yukiko Mori, Goshiro Yamamoto, Naoto Kume, Lutfan Lazuardi, Anis Fuad, Tomohiro Kuroda. Originally published in JMIR Formative Research (https://formative.jmir.org), 21.4.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Crisis Communication About the Maui Wildfires on TikTok: Content Analysis of Engagement With Maui Wildfire–Related Posts Over 1 Year

Jim P Stimpson[1,2], PhD; Aditi Srivastava[3]; Ketan Tamirisa[4]; Joseph Keawe aimoku Kaholokula[5], PhD; Alexander N Ortega[6], PhD

[1]Department of Health Economics, Systems, and Policy, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd, Dallas, TX, United States

[2]Peter O'Donnell Jr. School of Public Health, University of Texas Southwestern Medical Center, Dallas, TX, United States

[3]Frisco High School, Frisco, TX, United States

[4]Washington University in St. Louis, St Louis, MO, United States

[5]Department of Native Hawaiian Health, John A. Burns School of Medicine, University of Hawai i at Mānoa, Honolulu, HI, United States

[6]Thompson School of Social Work & Public Health, University of Hawai i at Mānoa, Honolulu, HI, United States

Corresponding Author:
Jim P Stimpson, PhD
Department of Health Economics, Systems, and Policy, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd, Dallas, TX, United States

Abstract

Background: The August 2023 wildfire in the town of Lāhainā on the island of Maui in Hawai i caused catastrophic damage, affecting thousands of residents, and killing 102 people. Social media platforms, particularly TikTok, have become essential tools for crisis communication during disasters, providing real-time crisis updates, mobilizing relief efforts, and addressing misinformation. Understanding how disaster-related content is disseminated and engaged with on these platforms can inform strategies for improving emergency communication and community resilience.

Objective: Guided by Social-Mediated Crisis Communication theory, this study examined TikTok posts related to the Maui wildfires to assess content themes, public engagement, and the effectiveness of social media in disseminating disaster-related information.

Methods: TikTok posts related to the Maui wildfires were collected from August 8, 2023, to August 9, 2024. Using TikTok's search functionality, we identified and reviewed public posts that contained relevant hashtags. Posts were categorized into 3 periods: during the disaster (August 8 to August 31, 2023), the immediate aftermath (September 1 to December 31, 2023), and the long-term recovery (January 1 to August 9, 2024). Two researchers independently coded the posts into thematic categories, achieving an interrater reliability of 87%. Engagement metrics (likes and shares) were analyzed to assess public interaction with different themes. Multivariable linear regression models were used to examine the associations between log-transformed likes and shares and independent variables, including time intervals, video length, the inclusion of music or effects, content themes, and hashtags.

Results: A total of 275 TikTok posts were included in the analysis. Most posts (132/275, 48%) occurred in the immediate aftermath, while 76 (27.6%) were posted during the long-term recovery phase, and 24.4% (n=67) were posted during the event. Posts during the event garnered the highest average number of likes (mean 75,092, SD 252,759) and shares (mean 10,928, SD 55,308). Posts focused on "Impact & Damage" accounted for the highest engagement, representing 36.8% (4,090,574/11,104,031) of total likes and 61.2% (724,848/1,184,049) of total shares. "Tourism Impact" (2,172,991/11,104,031, 19.6% of likes; 81,372/1,184,049, 6.9% of shares) and "Relief Efforts" (509,855/11,104,031, 4.6% of likes; 52,587/1,184,049, 4.4% of shares) were also prominent themes. Regression analyses revealed that videos with "Misinformation & Fake News" themes had the highest engagement per post, with a 4.55 coefficient for log-shares (95% CI 2.44-6.65), while videos about "Tourism Impact" and "Relief Efforts" also showed strong engagement (coefficients for log-likes: 2.55 and 1.76, respectively).

Conclusions: TikTok is an influential tool for disaster communication, amplifying both critical disaster updates and misinformation, highlighting the need for strategic content moderation and evidence-based messaging to enhance the platform's role in crisis response. Public health officials, emergency responders, and policy makers can leverage TikTok's engagement patterns to optimize communication strategies, improve real-time risk messaging, and support long-term community resilience.

(JMIR Form Res 2025;9:e67515) doi:10.2196/67515

XSL•FO
RenderX

KEYWORDS

social media; public health; disasters; Hawaii; media; post; communication; disaster; disaster communication; wildfire; information; dissemination; engagement; content analysis; content; metrics; misinformation; community; support

Introduction

Background

The wildfire that began on August 8, 2023, in the town of Lāhainā on Maui island in Hawai i brought widespread destruction, disrupted the lives of residents, destroyed homes, killed 102 people, and altered the environment [1-4]. Wildfires are increasingly becoming a serious public health concern due to their devastating impacts on the environment, communities, and health infrastructure [5-7]. The growing frequency and intensity of wildfires have been, in part, attributed to climate change, which creates hotter and drier conditions conducive to fire outbreaks [8-10]. Understanding public perceptions and engagement with these disasters is crucial for enhancing response, recovery, and resiliency [11].

In the digital age, social media platforms have emerged as critical tools for crisis communication, particularly during natural disasters and public health emergencies, providing situational updates, relief information, emotional support, and avenues for public mobilization [12-19]. TikTok has emerged as a key platform for public engagement during crises, offering a unique short-form video format that encourages personal storytelling and emotional connection [20-26]. TikTok's algorithm promotes high-engagement posts, which amplifies diverse perspectives. Previous studies, such as Sharma and De Choudhury's [20] exploration of TikTok in public health, underscore its value in disseminating information and fostering community resilience. However, limited research has focused on its role during natural disasters, especially compared with platforms like Facebook or Instagram [13,14]. The rise of mis and disinformation on social media also poses challenges, which potentially complicates disaster response and relief efforts [27-31].

Despite TikTok's increasing influence, few studies have focused on its role in communicating and shaping public discourse during natural disasters. One study analyzed TikTok videos from official accounts featured in the COVID-19 information hub, focusing on how credible information was disseminated to combat misinformation during the pandemic, highlighting the platform's role in delivering accessible, authoritative public health messaging [32]. Another examined trending TikTok content during the early days of the US COVID-19 pandemic, emphasizing how the platform facilitated social connection and coping through creative and relatable content during the "shelter in place" period [33]. A third study evaluated TikTok videos tagged with #ClimateChange to assess the platform's use in raising awareness and driving discussions on environmental issues [34]. Collectively, these studies underscore TikTok's growing importance as a tool for disseminating information, fostering public engagement, and influencing perceptions on global challenges but more information is needed about its use during natural disasters.

This study applies Social-Mediated Crisis Communication (SMCC) theory to understand how TikTok was used to disseminate and engage with information about the August 2023 Maui wildfires. SMCC theory explains how people seek, share, and respond to crisis information in digital environments [35]. SMCC theory posits that during a crisis, individuals rely on traditional media, social media influencers, and peer networks to make sense of unfolding events. The theory also emphasizes the role of organizational crisis communicators, but in the case of decentralized platforms like TikTok, crisis-related information is largely driven by user-generated content rather than official sources. This shift highlights the need to analyze which themes and message characteristics drive engagement on social media during disasters [35,36].

Applying SMCC theory, this study examined how content themes, video features (eg, video length, music, and effects), and hashtags influence engagement metrics, including likes and shares. Specifically, we assessed whether certain types of crisis messages align with SMCC theory's key information-sharing mechanisms. Instructing information refers to messages that provide practical guidance or factual updates, such as evacuation notices and safety alerts. Adjusting information includes messages that help users emotionally process the crisis, including personal narratives and community responses. Bolstering information consists of messages that reinforce community solidarity, fundraising efforts, and expressions of resilience. Prior research indicates that adjusting and bolstering messages often garner higher engagement on social media platforms, as users seek emotional support and opportunities to contribute to recovery efforts [36-38]. For instance, a study analyzing social media interactions during crises found that content providing emotional support and community solidarity received substantial user engagement [37]. Conversely, while instructing information—such as safety alerts and factual updates—is crucial for public safety, it tends to receive comparatively lower engagement, especially on entertainment-focused platforms like TikTok [23-26]. Research on TikTok's user engagement patterns suggests that content aligning with entertainment and emotional resonance is more likely to engage users than purely informational posts [38].

By applying SMCC theory, this study seeks to identify which crisis communication strategies gain traction on TikTok, offering insights into how social media platforms shape public understanding and response during natural disasters. The findings have implications for emergency responders, public health officials, and community organizations aiming to optimize social media engagement in future crises. Natural disasters, like the Maui wildfire, are typically seen as external and uncontrollable, which places greater emphasis on the need for empathetic and supportive communication rather than blame attribution [35-38]. Understanding how social media users adapt these strategies in real time contributes to advancing best practices for disaster communication [12-15].

XSL•FO
RenderX

Research Questions

Guided by SMCC theory, this study examined public engagement with TikTok content related to the Maui wildfire to understand its role in digital health communication during disasters. Specifically, we address three key questions: (1) how do TikTok posts related to the Maui wildfire align with SMCC's instructing, adjusting, and bolstering information-sharing mechanisms and what are the implications for public health messaging during crises; (2) which content themes generate the highest engagement among TikTok users throughout different phases of the disaster and how do these themes contribute to public health awareness, misinformation spread, and community resilience; and (3) how does engagement evolve over time from the immediate disaster response to the long-term recovery period and what does this reveal about the role of social media for public health communication after the immediate danger of crisis is past? By examining these research questions, this study contributes to the growing body of literature on digital health communication in disaster contexts, focusing on TikTok as an influential platform shaping disaster-related communication strategies. Understanding public engagement patterns across distinct temporal phases provides actionable insights for public health agencies, emergency responders, and policy makers to strengthen community resilience, improve evidence-based health messaging, and optimize the use of social media platforms for disaster response and recovery.

Methods

Data Collection

We collected TikTok posts about the Maui wildfire from August 8, 2023, to August 9, 2024, covering both the immediate aftermath of the wildfire and the 1-year anniversary. This timeframe was chosen to align with the phases emphasized in SMCC theory, which highlights the evolving nature of crisis communication on social media, emphasizing both the acute crisis period and the extended postcrisis discourse that shapes public understanding of response and recovery [35]. Posts were identified using a combination of disaster-related hashtags, including #MauiStrong and #MauiWildfires, and were sampled until thematic saturation was achieved [39].

To ensure diverse and representative content, posts were selected at random from the top search results for each hashtag or hashtag combination, prioritizing those with high visibility. For each hashtag, all available posts were thoroughly reviewed and included if they were relevant to the wildfire. Duplicate posts were removed to prevent redundancy and posts unrelated to the wildfire were excluded. After reaching thematic saturation, the final data set consisted of 275 unique posts, which was sufficient to capture the breadth of content shared about the wildfire [39].

Measures

The outcome variables for this study were likes and shares, which represent the primary engagement metrics for TikTok posts. Likes captured the number of times users expressed approval of a post, and shares indicated how often a post was redistributed by users to others on the platform. To address outliers in these engagement metrics, we applied a log transformation to both likes and shares. This transformation helped normalize the distribution of the variables and reduced the influence of extreme values, ensuring a more robust analysis. The log-transformed values were used as dependent variables in the regression models to evaluate factors influencing engagement.

Posts were coded with a variety of themes, and each post could be coded with multiple themes. Two researchers independently coded the posts to ensure reliability and minimize bias. Discrepancies in coding were resolved through discussion and consensus. The level of agreement between the 2 researchers was measured using the κ reliability statistic, which resulted in a high agreement rate of 87%. This high level of interrater reliability confirms the consistency and accuracy of the thematic classification [40].

We identified several major themes in the TikTok posts related to the Maui wildfire, each reflecting different aspects of the event and its aftermath. One prominent theme was "Impact & Damage", with posts documenting the destruction caused by the wildfires. These posts often included powerful visuals, such as videos and images of affected areas, illustrating the extent of the devastation. The "Personal Narratives & Interviews" theme featured individuals sharing their experiences of living through the wildfire. These posts were often deeply emotional, with appeals for continued attention to the needs of those affected. Personal stories ranged from immediate reactions during the wildfire to reflections on the ongoing recovery process, amplifying the voices of those directly impacted by the disaster. The theme of "Community, Solidarity, & Tribute" was also widely represented. These posts conveyed messages of support and solidarity from people around the world. Many stories highlighted local heroes, community efforts, and the resilience of those affected by the wildfire. A common subtheme within these posts was the presentation of "before the fire" scenes, which offered a sentimental and solemn perspective on what was lost and underscored the emotional impact of the disaster.

Another key theme was "Relief Efforts", which encompassed content aimed at promoting fundraisers, donations, and other forms of assistance. Posts in this category highlighted the work of both local and national organizations providing aid to affected communities. They also documented the progress (or lack thereof) in recovery efforts, which served as calls to action for further support from the public. These posts were integral in mobilizing resources and raising awareness about ongoing needs in the aftermath of the wildfire. Posts related to "Government & Policy Responses" critiqued or praised the governmental handling of the wildfire. Discussions centered on the effectiveness of emergency preparedness and response efforts, with users offering their perspectives on the adequacy of relief efforts, resource allocation, and communication from officials during and after the crisis. "Tourism Impact" was another important theme.

As Maui is a major tourist destination and most tourists to the island visit Lāhainā, many posts discussed how the wildfires affected the tourism industry. Both travelers and locals shared their thoughts on the changing landscape and economy,

XSL•FO
RenderX

reflecting on how the wildfires disrupted the flow of visitors and altered the island's economic prospects. Posts categorized under "Environmental & Climate Commentary" explored the role of climate change in exacerbating wildfires. Users discussed how changing weather patterns and prolonged droughts have contributed to the frequency and intensity of such natural disasters. Many of these posts also advocated for policy changes aimed at mitigating future risks, calling for stronger environmental protections and climate action to prevent similar events from occurring.

Finally, posts under the "Informational" theme provided detailed overviews of the wildfire. These posts explained the causes, effects, and other key details of the disaster, often serving as educational content for users looking to understand the broader context and implications of the wildfire. These posts helped disseminate factual and descriptive information about the event. In contrast, the theme of "Misinformation & Fake News" was also present. Posts categorized as "Misinformation" were identified based on 2 criteria: posts that discussed and attempted to debunk misinformation about the fires by explaining why the claims were inaccurate and posts that directly promoted misinformation, with the accuracy of their claims verified through fact-checking conducted by our research team. Any post meeting either of these criteria was classified as "Misinformation."

We included several other measures to assess how structural and temporal characteristics of posts influenced engagement levels (Likes and Shares). Video length, representing the duration of each TikTok post, was measured in seconds and included in the analyses as a continuous variable. A squared term for video length was also included to account for potential nonlinearity in its relationship with engagement metrics. Additionally, the presence of music or effects was included as a binary variable, capturing whether a post used TikTok's music or special effects features. Event timing was categorized into 3 periods: during the disaster (August 8 to August 31, 2023), the immediate aftermath (September 1 to December 31, 2023), and the long-term recovery (January 1 to August 9, 2024). Posts were coded based on their publication date relative to these time frames. This segmentation aligns with SMCC's theoretical framework, enabling insights into how content types and engagement patterns shift over time [35].

We categorized individual hashtags into thematic groups to analyze engagement patterns related to the Maui wildfire. Each hashtag group was designed to capture distinct aspects of public discourse and response. Posts were assigned to groups based on predefined hashtags, with a binary variable indicating whether a post contained at least 1 relevant hashtag. The "Support" group included hashtags such as #MauiStrong, #PrayForMaui, and #MauiSupport, capturing expressions of solidarity and general support for those affected. The "Relief" group encompassed hashtags like #DonateForMaui, #Fundraiser, #MauiRelief, #DisasterRelief, and #RedCross, reflecting calls for donations, fundraising efforts, and relief activities. The "WildfireContent" group included hashtags such as #MauiWildfires, #Wildfire, #WildfireAwareness, and #FirePrevention, focusing on wildfire-specific content, including awareness and prevention efforts. The "SafetyAlerts" group comprised hashtags like #StaySafe and #EmergencyAlert, emphasizing safety information and emergency updates. The "MauiLocations" group included hashtags referencing specific locations such as #Lahaina, #Kaanapali, #WestMaui, and #Hawaii, highlighting geographically specific content. The "AdvocacyAndUpdates" group encompassed hashtags such as #EnvironmentalJustice, #NewsUpdate, and #BreakingNews, reflecting advocacy-related posts and updates on unfolding events. Each group was assigned a binary variable, where a post was flagged as belonging to a group if it contained at least one of the associated hashtags. Logical functions in spreadsheet software were used to implement this classification efficiently, which ensured consistent assignment of posts to their respective groups.

Data Analysis

Engagement metrics were analyzed to assess immediate reactions and the reach of content to viewers [41,42]. The analysis strategy began with descriptive statistics to provide an overview of the distribution of engagement metrics (likes and shares) across key categories, such as time intervals and hashtag groups. We summarized engagement metrics by time interval, showing the number of posts, average likes and shares, and their distribution across the event duration, immediate aftermath, and long-term recovery phases. Crosstabulations of hashtags and content themes were conducted to examine how engagement metrics varied across different combinations of hashtag groups and content themes. These tables found in Multimedia Appendix 1 highlighted the total and relative contributions of each category to overall engagement.

The multivariable analyses included linear regression models predicting the log-transformed likes and shares. Independent variables included time intervals, video length (and its squared term), the presence of music or effects, content themes, and hashtag groups. This approach assessed the relative contribution of these variables to engagement metrics while adjusting for potential confounding factors. To test the robustness of the results, we replicated the linear regression models using robust regression, which is less sensitive to outliers and provides an alternative estimation strategy. The results of the robustness check in Multimedia Appendix 2 were largely consistent with the primary analyses, confirming the stability of key findings. This study adheres to the STROBE (Strengthening the Reporting of Observational studies in Epidemiology) guidelines for reporting observational studies.

Ethical Considerations

The University of Hawai i's Office of Research Compliance determined that the study does not meet the definition of human participants research and therefore does not require institutional review board approval or oversight.

Results

The analysis shown in Table 1 examined social media engagement metrics across the event time intervals. A total of 275 observations were included in the analysis. Most posts (132/275, 48%) occurred in the immediate aftermath, while 76 (27.6%) were posted during the long-term recovery phase, and

XSL•FO
RenderX

24.4% (n=67) were posted during the event. Posts during the event garnered the highest average number of likes, with a mean of 75,092 (SD 252,759). Engagement declined in subsequent periods, with posts in the "Immediate Aftermath" receiving an average of 40,641 likes, and posts in the "Long-Term Recovery" averaging 8685 likes. Across all intervals, the average number of likes was 40,203. Similarly, posts during the event achieved the highest average number of shares, with a mean of 10,928 (SD 55,308). This figure dropped to 3111 in the "Immediate Aftermath" and further declined to 493 in the "Long-Term Recovery" period. The overall average number of shares across all intervals was 4292.

Table . Engagement metrics (likes and shares) by time interval. Content analysis of 275 TikTok posts from August 8, 2023, to August 9, 2024.

	Value, n (%)	Likes, mean (SD)	Shares, mean (SD)
During event (August 8 to August 31, 2023)	67 (24.4)	75,092 (252,759)	10,928 (55,308)
Immediate aftermath (September 1 to December 31, 2023)	132 (48.0)	40,641 (184,994)	3111 (10,365)
Long-term recovery (January 1 to August 9, 2024)	76 (27.6)	8685 (27,244)	493 (1,207)
Total	275 (100.0)	124,418	14,532

Table 2 presents the results from the linear regression analysis of log-transformed likes. Video length was positively associated with likes, with the linear term (coefficient=0.01, *P*=.05) indicating a slight increase in engagement for longer videos. However, the quadratic term (coefficient=0.00, *P*=.03) suggests diminishing returns for extremely long videos. Event timing showed that engagement was significantly lower during the immediate aftermath of the disaster (coefficient=−0.96, *P*=.01) and during the long-term recovery phase (coefficient=−1.47, *P*<.001) compared with the disaster period. Videos featuring music or effects had a marginally lower engagement, though this relationship was not statistically significant (coefficient=−0.56, *P*=.09). Regarding content themes, several categories significantly increased engagement. Videos focused on "Impact & Damage" (coefficient=1.55, *P*=.01), "Informational" content (coefficient=1.32, *P*=.04), "Misinformation & Fake News" (coefficient=3.24, *P*<.001), "Personal Narratives & Interviews" (coefficient=1.66, *P*=.02), "Relief Efforts" (coefficient=1.62, *P*=.02), and "Tourism Impact" (coefficient=2.53, *P*<.001) all received higher levels of likes compared with the reference group. In contrast, themes such as "Environmental & Climate Commentary" and "Government & Policy Response" did not significantly impact engagement. For hashtags, the "Support" and "WildfireContent" groups were associated with significantly higher engagement, with coefficients of 1.90 (*P*<.001) and 1.87 (*P*<.001), respectively, compared with the reference group. However, the "MauiLocations" and "Relief" hashtags did not exhibit significant effects.

XSL•FO
RenderX

Table . Multivariable linear regression of log-transformed likes. Content analysis of 275 TikTok posts from August 8, 2023, to August 9, 2024.

		Coefficient	*P* value	95% CI
Video length		0.01	.05	0.00 to 0.02
Video length square		0.00	.03	0.00 to 0.00
Event timing				
	During disaster	—[a]	—	—
	Immediate aftermath	−0.96	.01	−1.71 to −0.21
	Long-term recovery	−1.47	<.001	−2.31 to −0.63
Music and effects		−0.56	.09	−1.21 to 0.09
Content theme				
	Community, Solidarity, & Tribute	—	—	—
	Environmental & Climate Commentary	0.04	.97	−2.19 to 2.27
	Government & Policy Response	0.95	.23	−0.61 to 2.51
	Impact & Damage	1.55	.01	0.33 to 2.77
	Informational	1.32	.04	0.05 to 2.58
	Misinformation & Fake News	3.24	<.001	1.22 to 5.26
	Personal Narratives & Interviews	1.66	.02	0.32 to 3.00
	Relief Efforts	1.62	.02	0.23 to 3.00
	Tourism Impact	2.53	<.001	1.02 to 4.05
Hashtag group				
	AdvocacyAndUpdates	—	—	—
	MauiLocations	0.00	>0.99	−0.89 to 0.90
	Relief	−0.93	.12	−2.09 0.24
	Support	1.90	<.001	0.97 to 2.84
	WildfireContent	1.87	<.001	0.95 to 2.78
Constant		6.02	<.001	4.55 to 7.49

[a]Not applicable (reference category).

Table 3 presents the linear regression analysis for log-transformed shares. Video length was not significantly associated with shares. The Immediate Aftermath period was negatively associated with shares (coefficient=−0.82, *P*=.04), and long-term recovery showed an even stronger negative effect (coefficient=−1.51, *P*<.001), both indicating fewer shares during these periods compared with the baseline (during event). Content themes also had significant effects on shares. Specifically, “Impact & Damage,” “Informational,” and “Misinformation & Fake News” content themes had large positive associations with shares (coefficients of 2.46, 1.86, and 4.55, respectively, all with *P* values <.01), suggesting that content related to these themes is shared more frequently. “Personal Narratives & Interviews,” “Relief Efforts,” and “Tourism Impact” also showed significant positive associations with shares, further emphasizing the popularity of content related to these themes. Finally, certain hashtag groups influenced shares. “Relief” had a negative impact on shares (coefficient=−1.68, *P*=.01), while “Support” and “WildfireContent” both had positive effects (coefficients of 2.28 and 1.64, respectively, both with *P* values <.01), indicating that these types of hashtags contribute to higher share rates relative to the reference group.

XSL•FO
RenderX

Table . Multivariable linear regression of log-transformed shares. Content analysis of 275 TikTok posts from August 8, 2023, to August 9, 2024.

		Coefficient	*P* value	95% CI
Video length		0.01	.08	0.00 to 0.02
Video length square		0.00	.08	0.00 to 0.00
Event timing				
	During disaster	—[a]	—	—
	Immediate aftermath	−0.82	.04	−1.60 to −0.04
	Long-term recovery	−1.51	<.001	−2.39 to −0.64
Music and effects		−0.43	.21	−1.11 to 0.25
Content theme				
	Community, Solidarity, & Tribute	—	—	—
	Environmental & Climate Commentary	0.44	.71	−1.89 to 2.77
	Government & Policy Response	1.62	.05	0.00 to 3.25
	Impact & Damage	2.46	<.001	1.19 to 3.73
	Informational	1.86	.01	0.54 to 3.18
	Misinformation & Fake News	4.55	<.001	2.44 to 6.65
	Personal Narratives & Interviews	2.11	<.001	0.72 to 3.51
	Relief Efforts	2.38	<.001	0.94 to 3.82
	Tourism Impact	2.96	<.001	1.37 to 4.54
Hashtag group				
	AdvocacyAndUpdates	—	—	—
	MauiLocations	0.19	.70	−0.75 to 1.12
	Relief	−1.68	.01	−2.89 to −0.46
	Support	2.28	<.001	1.30 to 3.25
	WildfireContent	1.64	<.001	0.68 to 2.60
Constant		2.70	<.001	1.17 to 4.23

[a]Not applicable (reference category).

Discussion

Principal Findings

This study provides novel insights into how TikTok serves as a platform for public engagement during and after a natural disaster, using the 2023 Maui wildfire as a case study. Guided by SMCC theory, we examined how content themes, hashtags, and video features influence user engagement metrics, such as likes and shares. Our findings contribute to the growing body of literature on social media's role in crisis communication while offering unique contributions that expand upon prior studies [43-49].

This study reveals that TikTok serves as a dynamic platform for real-time information dissemination and public engagement during and after natural disasters. Posts emphasizing themes such as "Impact & Damage" and "Personal Narratives & Interviews" garnered the highest engagement metrics, illustrating the platform's capacity to emotionally and visually document the crisis. Content highlighting tourism and economic impacts also performed strongly, reflecting the broader public's interest in understanding the disaster's ripple effects. However, posts addressing critical but less emotional themes, such as "Government & Policy Response" and "Relief Efforts," attracted relatively lower engagement, highlighting a potential gap in how these topics resonate with TikTok audiences.

Engagement trends varied across the disaster timeline. While a prior study on the Maui wildfire focused on the immediate aftermath of the disaster, we analyzed engagement over a 1-year period, encompassing both crisis and postcrisis phases [38]. This extended time frame enabled us to explore the evolution of public engagement and highlighted key shifts in user priorities over time. Posts shared during the crisis phase generated the highest engagement, underscoring the public's urgent demand

XSL•FO
RenderX

for real-time updates. Engagement declined in the immediate aftermath and long-term recovery periods, reflecting waning public attention. These patterns emphasize the importance of strategically timing communication to sustain public interest and engagement throughout the recovery phase. The observed temporal trends also underscore the need for innovative strategies to re-engage audiences as the crisis progresses into recovery [35].

Engagement was predominantly driven by themes that conveyed the emotional impact and visual magnitude of the disaster, reinforcing the role of emotionally resonant content in shaping public discourse [37,38]. This pattern aligns with SMCC theory, which emphasizes the importance of adjusting information in crisis communication—content that helps audiences process the event and fosters emotional connection [35]. Moreover, informational posts about safety alerts and evacuation notices garnered relatively low engagement despite their critical importance for public safety. In contrast, adjusting and bolstering information, such as personal narratives and fundraising efforts, received significantly higher engagement, reflecting the public's preference for emotionally resonant and actionable content. These findings suggest that while instructing information is vital during the early crisis phase, sustained engagement may rely on content that fosters emotional connection and community solidarity [35,36].

Our results underscore the importance of visually compelling and emotionally resonant content in driving engagement during disasters [50-52]. This aligns with a recent study that found audiovisual vividness was a key predictor of engagement on TikTok during the same wildfire [38]. Both studies highlight the role of survivor-led narratives in capturing public attention, suggesting that user-generated content may resonate more strongly than traditional news media posts. However, while prior work used a digital storytelling framework to analyze the role of narrative persuasion and visual storytelling, our study extends the literature by applying SMCC theory to examine how instructing, adjusting, and bolstering messages align with engagement patterns [35-38]. This distinction situates our findings within the broader literature on crisis communication strategies.

Furthermore, our study highlights the persistent challenge of misinformation during disasters [21,27-29,53]. Posts categorized under misinformation themes, while limited in number, garnered substantial engagement, a finding that aligns with a recent study that found that posts leveraging emotional storytelling and high audiovisual vividness were particularly effective at drawing audience engagement [38]. This finding underscores the broader concern in public health regarding the virality of misinformation on social media, where false claims about disaster response, relief efforts, or health risks can undermine public trust in official sources [27-31]. Proactive strategies are needed to combat misinformation and amplify credible sources during crisis events. Integrating insights from both SMCC theory and the digital storytelling literature, our findings suggest that pairing information with compelling visual narratives may enhance its reach and impact, particularly on entertainment-driven platforms like TikTok.

Finally, our study contributes to the ongoing discourse on the role of video features in driving engagement [25,26,50,51]. The platform's algorithm-driven reach and short-form video format provide unique opportunities to disseminate impactful messages that resonate with diverse audiences [34,38,54]. Consistent with prior research, we find that shorter videos with clear and emotionally engaging messages are more likely to capture public attention [55]. These findings align with broader trends in TikTok engagement metrics and highlight the need for tailored communication strategies that align with platform-specific user behaviors. Tailored strategies to amplify these themes could enhance the platform's utility in disaster response [20].

Limitations

While 2 researchers independently coded the posts to ensure reliability and minimize bias, the interpretation of themes still involves subjective judgment. Even with our high interrater reliability, some residual bias may persist. The 1-year timeframe of the study provides a snapshot but might miss evolving trends over a longer duration. Additionally, the focus on TikTok, with its distinct user demographics and content styles, limits the generalizability of the findings to other social media platforms such as X (formerly Twitter), Facebook, or Instagram. The unique characteristics of TikTok's user base might not reflect broader social media trends. Furthermore, language and cultural barriers could have resulted in the underrepresentation of posts in languages other than English or those deeply rooted in specific cultural contexts, limiting the diversity of perspectives analyzed.

We acknowledge that TikTok's algorithmic influence on post visibility is a significant factor, as the platform curates content based on user behavior, preferences, and engagement, potentially creating biases in the posts available for analysis. As a result, our data set may disproportionately reflect content with high algorithmic visibility rather than a fully representative sample of all relevant posts. Additionally, regional and language biases in content discovery may affect the results, as TikTok's algorithms and search functionalities prioritize content tailored to specific geographic regions and linguistic preferences. This could result in the underrepresentation of videos from certain regions or in languages less prominent on the platform. The impact of paid promotion or boosted content is another limitation. Videos that received paid promotions may have artificially higher visibility and engagement, skewing the data toward commercially sponsored content. Our study did not distinguish between organically popular posts and those boosted through paid promotion, which could influence the analysis of engagement trends.

Implications for Future Research and Practice

The findings have important implications for disaster communication and public health response strategies. Understanding how engagement evolves over time and informs messaging that sustains public interest beyond the acute crisis phase. Disaster communicators can leverage TikTok's ability to amplify emotionally resonant content while addressing misinformation through targeted campaigns. Future research should explore cross-platform comparisons to identify complementary roles of different social media platforms.

XSL•FO
RenderX

Additionally, research on the effectiveness of public health interventions in countering misinformation and improving digital health literacy is needed to develop best practices for social media–based disaster communication.

Conclusion

In conclusion, this study builds upon prior work by integrating theoretical insights from SMCC theory with empirical analyses of TikTok engagement during a major natural disaster. TikTok is an influential tool for real-time communication, emotional support, and advocacy during natural disasters. By examining themes related to the Maui wildfire, this study demonstrated how users respond to and interact with crisis-related content in real time. While TikTok's strengths lie in its ability to visually and emotionally connect with audiences, addressing challenges such as misinformation and sustaining engagement throughout the recovery phase is essential. Strengthening digital crisis communication strategies, particularly in sustaining engagement beyond the immediate response phase, is critical for improving disaster preparedness and recovery efforts. Future research should further investigate best practices for leveraging social media in disaster preparedness, response, and community resilience.

Acknowledgments

JPS, JKK, and ANO were supported by the National Institute on Minority Health and Health Disparities (NIMHD) at the National Institutes of Health (NIH) under award number R61MD019939. The content is solely the responsibility of the authors and does not necessarily represent the official views of NIMHD. The funders had no role in the study design, data analysis, decision to publish, or preparation of the manuscript.

Data Availability

The data underlying the results presented in the study are publicly available from TikTok.

Authors' Contributions

All authors contributed to the study's conception and design. Data collection and analysis were performed by AS and KT. JPS wrote the first draft of the manuscript. All authors contributed to subsequent drafts of the manuscript. All authors read and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Distribution of likes and shares.
[DOCX File, 34 KB - formative_v9i1e67515_app1.docx]

Multimedia Appendix 2
Multivariable robust regression of log-transformed likes and shares.
[DOCX File, 24 KB - formative_v9i1e67515_app2.docx]

References

1. Farrant DN, Trauernicht C, Kagawa-Viviani A, Giambelluca TW, D'Antonio CM. On the anniversary of the Maui fires, a call for indigenous land care to mitigate future disasters. Proc Natl Acad Sci USA 2024 Sep 24;121(39):e2401045121. [doi: 10.1073/pnas.2401045121]
2. Welch L, Ramones M. 102 angels looking down on us: Maui marks 1 year since wildfire disaster. Hawaii News Now. URL: https://www.hawaiinewsnow.com/2024/07/09/1-year-anniversary-maui-wildfires-approaches-events-take-place-honor-victims [accessed 2024-07-09]
3. Suresh A, Tu L, Stanford FC. The 2023 Maui wildfires and the mental health effects of climate-induced relocation. Am J Public Health 2024 Nov;114(11):1184-1187. [doi: 10.2105/AJPH.2024.307815] [Medline: 39236278]
4. Holliday R, Krishnamurti LS, Jordan SE, Sia MA, Brenner LA, Monteith LL. The health and social impacts of the Maui wildfires: post-disaster care from a sociocultural lens. Hawaii J Health Soc Welf 2024 Mar;83(3):85-87. [doi: 10.62547/KXYZ1785] [Medline: 38456158]
5. Balch JK, Bradley BA, Abatzoglou JT, Nagy RC, Fusco EJ, Mahood AL. Human-started wildfires expand the fire niche across the United States. Proc Natl Acad Sci U S A 2017 Mar 14;114(11):2946-2951. [doi: 10.1073/pnas.1617394114] [Medline: 28242690]
6. Wen J, Heft-Neal S, Baylis P, Boomhower J, Burke M. Quantifying fire-specific smoke exposure and health impacts. Proc Natl Acad Sci U S A 2023 Dec 19;120(51):e2309325120. [doi: 10.1073/pnas.2309325120] [Medline: 38085772]

XSL•FO
RenderX

7. Gould CF, Heft-Neal S, Johnson M, Aguilera J, Burke M, Nadeau K. Health effects of wildfire smoke exposure. Annu Rev Med 2024 Jan 29;75:277-292. [doi: 10.1146/annurev-med-052422-020909] [Medline: 37738508]
8. Bowman D, Kolden CA, Abatzoglou JT, et al. Wildfires: climate change, deforestation, and the aridification of wildlands. Science 2020;370(6519):eabb5365. [doi: 10.1126/science.abb5365]
9. Xu R, Yu P, Abramson MJ, et al. Wildfires, global climate change, and human health. N Engl J Med 2020 Nov 26;383(22):2173-2181. [doi: 10.1056/NEJMsr2028985] [Medline: 33034960]
10. Lambrou N, Kolden C, Loukaitou-Sideris A, Anjum E, Acey C. Social drivers of vulnerability to wildfire disasters: a review of the literature. Landsc Urban Plan 2023 Sep;237:104797. [doi: 10.1016/j.landurbplan.2023.104797]
11. Hertelendy AJ, Howard C, Sorensen C, et al. Seasons of smoke and fire: preparing health systems for improved performance before, during, and after wildfires. Lancet Planet Health 2024 Aug;8(8):e588-e602. [doi: 10.1016/S2542-5196(24)00144-X] [Medline: 39122327]
12. Alexander DE. Social media in disaster risk reduction and crisis management. Sci Eng Ethics 2014 Sep;20(3):717-733. [doi: 10.1007/s11948-013-9502-z] [Medline: 24306994]
13. Houston JB, Hawthorne J, Perreault MF, et al. Social media and disasters: a functional framework for social media use in disaster planning, response, and research. Disasters 2015 Jan;39(1):1-22. [doi: 10.1111/disa.12092] [Medline: 25243593]
14. Lachlan KA, Spence PR, Lin X, Najarian K, Del Greco M. Social media and crisis management: CERC, search strategies, and Twitter content. Comput Human Behav 2016 Jan;54:647-652. [doi: 10.1016/j.chb.2015.05.027]
15. Saroj A, Pal S. Use of social media in crisis management: a survey. Int J Disaster Risk Reduct 2020 Sep;48:101584. [doi: 10.1016/j.ijdrr.2020.101584]
16. Clark N, Boersma K, Bonati S, et al. Exploring the impacts of social media and crowdsourcing on disaster resilience. Open Res Eur 2021;1:60. [doi: 10.12688/openreseurope.13721.3] [Medline: 38380383]
17. Ogie RI, James S, Moore A, Dilworth T, Amirghasemi M, Whittaker J. Social media use in disaster recovery: a systematic literature review. Int J Disaster Risk Reduct 2022 Feb;70:102783. [doi: 10.1016/j.ijdrr.2022.102783]
18. Lam NSN, Meyer M, Reams M, et al. Improving social media use for disaster resilience: challenges and strategies. Int J Disaster Risk Reduct 2023 Oct 2;16(1):3023-3044. [doi: 10.1080/17538947.2023.2239768]
19. Seneviratne K, Nadeeshani M, Senaratne S, Perera S. Use of social media in disaster management: challenges and strategies. Sustainability 2024 Jan;16(11):4824. [doi: 10.3390/su16114824]
20. Sharma R, De Choudhury M. TikTok and the evolving public health landscape. Nat Med 2021;27(1):8-10. [doi: 10.1038/s41591-020-01170-1]
21. Kirkpatrick CE, Lawrie LL. TikTok as a source of health information and misinformation for young women in the United States: survey study. JMIR Infodemiology 2024 May 21;4:e54663. [doi: 10.2196/54663] [Medline: 38772020]
22. Comp G, Dyer S, Gottlieb M. Is TikTok the next social media frontier for medicine? AEM Educ Train 2021 Jul;5(3). [doi: 10.1002/aet2.10532] [Medline: 34095694]
23. Kong W, Song S, Zhao YC, Zhu Q, Sha L. TikTok as a health information source: assessment of the quality of information in diabetes-related videos. J Med Internet Res 2021 Sep 1;23(9):e30409. [doi: 10.2196/30409] [Medline: 34468327]
24. Samuel L, Kuijpers K, Bleakley A. TherapyTok for depression and anxiety: a quantitative content analysis of high engagement TikTok videos. J Adolesc Health 2024 Jun;74(6):1184-1190. [doi: 10.1016/j.jadohealth.2024.02.002] [Medline: 38493396]
25. Lau N, Srinakarin K, Aalfs H, Zhao X, Palermo TM. TikTok and teen mental health: an analysis of user-generated content and engagement. J Pediatr Psychol 2025 Jan 1;50(1):63-75. [doi: 10.1093/jpepsy/jsae039] [Medline: 38950415]
26. Xiong M, Young AT, Fernandez A, Jones JS, Kirby AM. Understanding breast cancer surgery through Tiktok: an analysis of video content and viewer engagement. Ann Surg Oncol 2024 Jul;31(7):4182-4184. [doi: 10.1245/s10434-024-15268-y] [Medline: 38592623]
27. Vosoughi S, Roy D, Aral S. The spread of true and false news online. Science 2018 Mar 9;359(6380):1146-1151. [doi: 10.1126/science.aap9559] [Medline: 29590045]
28. Lewandowsky S, Ecker UKH, Cook J. Beyond misinformation: understanding and coping with the “post-truth” era. J Appl Res Mem Cogn 2017;6(4):353-369. [doi: 10.1016/j.jarmac.2017.07.008]
29. Starbird K, Maddock J, Orand M, Achterman P, Mason RM. Rumors, false flags, and digital vigilantes: misinformation on twitter after the 2013 boston marathon bombing. Presented at: iConference 2014; Mar 4-7, 2014; Berlin. [doi: 10.9776/14308]
30. George JO, Elayan S, Sykora M, et al. The role of social media in building pandemic resilience in an urban community: a qualitative case study. Int J Environ Res Public Health 2023 Sep 4;20(17):6707. [doi: 10.3390/ijerph20176707] [Medline: 37681847]
31. Xie L, Pinto J, Zhong B. Building community resilience on social media to help recover from the COVID-19 pandemic. Comput Human Behav 2022 Sep;134:107294. [doi: 10.1016/j.chb.2022.107294] [Medline: 35431426]
32. Li Y, Guan M, Hammond P, Berrey LE. Communicating COVID-19 information on TikTok: a content analysis of TikTok videos from official accounts featured in the COVID-19 information hub. Health Educ Res 2021 Jul 12;36(3):261-271. [doi: 10.1093/her/cyab010] [Medline: 33667311]
33. Unni Z, Weinstein E. Shelter in place, connect online: trending TikTok content during the early days of the U.S. COVID-19 pandemic. J Adolesc Health 2021 May;68(5):863-868. [doi: 10.1016/j.jadohealth.2021.02.012] [Medline: 33824071]

XSL•FO
RenderX

34. Basch CH, Yalamanchili B, Fera J. #Climate change on TikTok: a content analysis of videos. J Community Health 2022 Feb;47(1):163-167. [doi: 10.1007/s10900-021-01031-x] [Medline: 34545460]
35. Jin Y, Austin L, Liu BF. Social‐Mediated crisis communication research: how information generation, consumption, and transmission influence communication processes and outcomes. In: Coombs WT, Holladay SJ, editors. The Handbook of Crisis Communication, 1st edition: Wiley; 2022:151-167. [doi: 10.1002/9781119678953]
36. Lauran N, Kunneman F, Van de Wijngaert L. Connecting social media data and crisis communication theory: a case study on the chicken and the egg. J Risk Res 2020 Oct 2;23(10):1259-1277. [doi: 10.1080/13669877.2019.1628097]
37. Tian Y, Yang J. Deny or bolster? A comparative study of crisis communication strategies between Trump and Cuomo in COVID-19. Public Relat Rev 2022 Jun;48(2):102182. [doi: 10.1016/j.pubrev.2022.102182] [Medline: 35368968]
38. Liu W, Zhao X, Zhan M(, Hernandez S. Streaming disasters on TikTok: examining social mediated crisis communication, public engagement, and emotional responses during the 2023 Maui wildfire. Public Relat Rev 2024 Dec;50(5):102512. [doi: 10.1016/j.pubrev.2024.102512]
39. Saunders B, Sim J, Kingstone T, et al. Saturation in qualitative research: exploring its conceptualization and operationalization. Qual Quant 2018;52(4):1893-1907. [doi: 10.1007/s11135-017-0574-8] [Medline: 29937585]
40. Gisev N, Bell JS, Chen TF. Interrater agreement and interrater reliability: key concepts, approaches, and applications. Res Social Adm Pharm 2013;9(3):330-338. [doi: 10.1016/j.sapharm.2012.04.004] [Medline: 22695215]
41. McCay-Peet L, Quan-Haase A. A model of social media engagement: user profiles, gratifications, and experiences. In: The SAGE Handbook of Social Media Research Methods: SAGE; 2017:199-214.
42. Neiger BL, Thackeray R, Van Wagenen SA, et al. Use of social media in health promotion: purposes, key performance indicators, and evaluation metrics. Health Promot Pract 2012 Mar;13(2):159-164. [doi: 10.1177/1524839911433467] [Medline: 22382491]
43. Muniz-Rodriguez K, Ofori SK, Bayliss LC, et al. Social media use in emergency response to natural disasters: a systematic review with a public health perspective. Disaster Med Public Health Prep 2020 Feb;14(1):139-149. [doi: 10.1017/dmp.2020.3] [Medline: 32148219]
44. Mitcham D, Taylor M, Harris C. Utilizing social media for information dispersal during local disasters: the communication hub framework for local emergency management. Int J Environ Res Public Health 2021 Oct 14;18(20):10784. [doi: 10.3390/ijerph182010784] [Medline: 34682529]
45. Simon T, Adini B, El-Hadid M, Goldberg A, Aharonson-Daniel L. The race to save lives: demonstrating the use of social media for search and rescue operations. PLoS Curr 2014 Nov 6;6:ecurrents.dis.806848c38f18c6b7b0037fae3cd4edc5. [doi: 10.1371/currents.dis.806848c38f18c6b7b0037fae3cd4edc5] [Medline: 25685618]
46. Gómez-Casillas A, Gómez Márquez V. The effect of social network sites usage in climate change awareness in Latin America. Popul Environ 2023 May 1;45(2):7. [doi: 10.1007/s11111-023-00417-4] [Medline: 37152891]
47. Meng Y, Chung D, Zhang A. The effect of social media environmental information exposure on the intention to participate in pro-environmental behavior. PLoS ONE 2023;18(11):e0294577. [doi: 10.1371/journal.pone.0294577] [Medline: 37972040]
48. Maran DA, Begotti T. Media exposure to climate change, anxiety, and efficacy beliefs in a sample of Italian university students. Int J Environ Res Public Health 2021 Sep 4;18(17):9358. [doi: 10.3390/ijerph18179358] [Medline: 34501946]
49. Shan S, Zhao F. Social media-based urban disaster recovery and resilience analysis of the Henan deluge. Nat Hazards (Dordr) 2023 May 23:1-29. [doi: 10.1007/s11069-023-06010-0] [Medline: 37360801]
50. Chen Q, Min C, Zhang W, Ma X, Evans R. Factors driving citizen engagement with government TikTok accounts during the COVID-19 pandemic: model development and analysis. J Med Internet Res 2021 Feb 4;23(2):e21463. [doi: 10.2196/21463] [Medline: 33481756]
51. Cheng Z, Li Y. Like, comment, and share on TikTok: exploring the effect of sentiment and second-person view on the user engagement with TikTok news videos. Soc Sci Comput Rev 2024 Feb;42(1):201-223. [doi: 10.1177/08944393231178603]
52. Basch CH, Hillyer GC, Yalamanchili B, Morris A. How TikTok is being used to help individuals cope with breast cancer: cross-sectional content analysis. JMIR Cancer 2022 Dec 6;8(4):e42245. [doi: 10.2196/42245] [Medline: 36472899]
53. Basch CH, Basch CE, Hillyer GC, Meleo-Erwin ZC. Social media, public health, and community mitigation of COVID-19: challenges, risks, and benefits. J Med Internet Res 2022 Apr 12;24(4):e36804. [doi: 10.2196/36804] [Medline: 35380539]
54. Chavda J. Who U.S. adults follow on TikTok. Pew Research Center. URL: https://www.pewresearch.org/data-labs/2024/10/08/who-u-s-adults-follow-on-tiktok/ [accessed 2025-02-19]
55. Dai X, Wang J. Effect of online video infotainment on audience attention. Humanit Soc Sci Commun 2023;10(1):421. [doi: 10.1057/s41599-023-01921-6]

Abbreviations

SMCC: Social-Mediated Crisis Communication
STROBE: Strengthening the Reporting of Observational studies in Epidemiology

XSL•FO
RenderX

Edited by A Mavragani; submitted 14.10.24; peer-reviewed by KP Hocevar, R Peng; revised version received 29.01.25; accepted 30.01.25; published 04.03.25.

Please cite as:
Stimpson JP, Srivastava A, Tamirisa K, Kaholokula JK, Ortega AN
Crisis Communication About the Maui Wildfires on TikTok: Content Analysis of Engagement With Maui Wildfire–Related Posts Over 1 Year
JMIR Form Res 2025;9:e67515
URL: https://formative.jmir.org/2025/1/e67515
doi:10.2196/67515

© Jim P Stimpson, Aditi Srivastava, Ketan Tamirisa, Joseph Keawe aimoku Kaholokula, Alexander N Ortega. Originally published in JMIR Formative Research (https://formative.jmir.org), 4.3.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Informal Caregivers Connecting on the Web: Content Analysis of Posts on Discussion Forums

Michelle L Foster[1], MPH; Chinenye Egwuonwu[1,2], MPH, MD; Erin Vernon[3], MA, PhD; Mohammad Alarifi[1,4], MS, PhD; M Courtney Hughes[1], MS, PhD

[1]School of Health Studies, Northern Illinois University, DeKalb, IL, United States

[2]Department of Internal Medicine, Piedmont Athens Regional, Athens, GA, United States

[3]Department of Economics, Seattle University, Seattle, WA, United States

[4]Radiological Sciences Department, College of Applied Medical Sciences, King Saud University, Riyadh, Saudi Arabia

Corresponding Author:
M Courtney Hughes, MS, PhD
School of Health Studies, Northern Illinois University, DeKalb, IL, United States

Abstract

Background: About 53 million adults in the United States offer informal care to family and friends with disease or disability. Such care has an estimated economic value of US $600 million. Most informal caregivers are not paid nor trained in caregiving, with many experiencing higher-than-average levels of stress and depression and lower levels of physical health. Some informal caregivers participate in web-based forums related to their caregiving role.

Objective: This study aimed to explore how informal caregivers use easy-to-access caregiving web-based forums, including the types of information they share and seek from others. It also aimed to gain insights into the informal caregiver experience from the content these informal caregivers posted.

Methods: The study population consisted of participants who posted on 5 web-based forums for informal caregivers between February and April 2024. Researchers extracted the first 6 responses to the first 20 questions and comments to appear posted by the informal caregivers in each of the 5 forums, removing any individually identifying information. We used a codebook thematic analysis approach to examine the data with Dedoose (SocioCultural Research Consultants). Researchers independently read all posts and coded the data. The author group discussed the codes, reiteratively refined them, and identified themes within the data.

Results: The data consisted of 100 initial posts and 600 responses. Over half of the initial posts included specific questions, with the remaining initial posts sharing experiences or reflections. Posts ranged in length from a sentence to more than 500 words. Domains identified included handling interpersonal challenges, navigating complicated systems, gathering tactical coping strategies, managing emotions, and connecting with others in similar situations. Negative interpersonal interactions were mentioned 123 times, with 77 posts describing challenging situations with extended family. Posters inquired about accessing resources, with health care and health insurance included 51 times, while legal and financial concerns were addressed 124 times. Caregiving challenges were mentioned hundreds of times, including discussion of hygiene (n=18), nutrition (n=21), and desire for a caregiving break (n=47). Posters expressed emotion in their comments 180 times, which included 32 mentions of guilt and 26 mentions of positive emotion. The importance of web-based group support was mentioned 301 times.

Conclusions: Informal caregivers play an essential role in society. Many experience multifaceted challenges related to their caregiving role, and some turn to the internet for community. Accessing web-based discussion forums is a low-barrier method for informal caregivers to connect with others who may be experiencing similar emotions and challenges. Gaining a greater understanding of the ways informal caregivers seek advice and offer support to one another provides insight into the challenges they face. The domains identified on these forums may be helpful, as clinicians provide information to care recipients and their informal caregivers along their health journeys.

(JMIR Form Res 2025;9:e64757) doi:10.2196/64757

KEYWORDS

informal caregivers; family caregivers; discussion forum; caregiver support; support group; social support; caregiver navigation; content analysis; adults; United States; informal care; codebook; thematic analysis; web-based discussion; web-based forums; clinicians; medical care; peer-to-peer support; web-based communities; caregivers

XSL•FO
RenderX

Introduction

About 53 million informal caregivers assist individuals in need of medical care and activities of daily living in the United States [1]. The unpaid value of this care was about US $600 billion, up from a 2017 estimate of about US $470 billion value of unpaid informal care. Informal caregivers are generally unpaid and untrained, and historically, these caregivers performed household tasks and personal care activities in contrast to formal caregivers who were part of the health system [2]. Research shows that more than 58% of informal caregivers are responsible for medical or nursing tasks for their care recipients [3].

Informal caregiving is associated with both benefits and burdens for caregivers. Many informal caregivers experience higher levels of stress and depression and have lower levels of physical health than their counterparts [4-6]. Care recipients with poor levels of physical and mental health were associated with worse caregiving experiences [5,6]. Furthermore, informal caregivers with older age or mental health issues have been shown to have worse caregiving experiences [5,7].

Most informal caregivers have internet access and use it to obtain health information [8]. There is significant research regarding the use of web-based tools among informal caregivers [9-18]. Some research used web-based surveys to ask about informal caregiver needs [10,11,17], while others evaluated the use or content of various purpose-built social support service interventions [13-15,18]. The qualitative research in this space has focused on specific topics related to caregiving, such as well-being during COVID-19 [19], advance care planning [20], and developing the identity of an informal caregiver [21]. Many other qualitative analyses have focused on particular diagnoses, such as care recipients with Alzheimer disease or dementia [22-25], posttraumatic stress disorder [26], stroke [27,28], and cancer [29-31]. Qualitative research about general topics addressed in informal caregiver forums focused only on 1 discussion platform (Reddit) and was evaluated with the use of artificial intelligence [32].

The previous studies investigating informal caregivers' use of web-based forums have found that informal caregivers use web-based support groups and forums to find emotional support [13-15,19,20,22-28,30], request information [14,20,23,26-28,30], discuss symptoms and conditions [23,24,27,28], interact with other caregivers [13,15,19-21,23,26,30], and talk about the challenges of caregiving [15,19,20,22,23,25,26]]. Additionally, all existing studies but one [27] evaluated data from 1 specific web-based forum rather than more than 1 forum. The study herein aims to help fill the gap in research that examines how informal caregivers use caregiving web-based forums that are not restricted to just 1 forum. This is also the first study to examine the forums that appear first when a caregiver searches for chronic illness caregiver support forums on Google. This approach can shed light on how caregivers interact with one another across multiple caregiving web-based forums.

This study explores how informal caregivers used a variety of popular, easy-to-access web-based forums. Specifically, we sought to discover the types of information individuals posting to web-based informal caregiver forums shared and sought from others. Insights into the informal caregiver experience from exploring multiple discussion forums can inform health care providers and other community groups so they may better design interventions and policies responsive to caregiver needs.

Methods

Study Population

The study population consists of participants who posted on publicly accessible web-based forums for informal caregivers between February and April 2024. Descriptions of the forums (AgingCare [33], Caregiver Action Network [34], Facebook Dementia Caregivers Support Group [35], Facebook Empowering Caregivers Support Group [36], and Mayo Clinic Caregivers Forum [37]) are presented in Table 1. We selected these forums because they appeared as the top 5 forums in a web-based Google search using the phrase "caregiver support forum and chronic illness."

XSL•FO
RenderX

Table . Description of web-based caregiver forums from each forum's website.

Forum name	Description	Details
AgingCare [33]	"Connects families who are caring for aging parents, spouses, or other elderly loved ones with the information and support they need to make informed caregiving decisions."	• Caregivers seek mutual support for medical challenges that occur with aging such as dementia • Caregivers offer valuable suggestions based on firsthand experience managing these conditions
Caregiver Action Network[a] [34]	"The nation's leading family caregiver organization working to improve the quality of life for the more than 90 million Americans who care for loved ones with chronic conditions, disabilities, or disease, as well as those supporting the living needs of older adults."	• Caregivers seek support for general, day-to-day issues in caregiving • Caregivers share personal experiences while caring for their loved ones with disabilities or disease
Facebook Dementia Caregivers Support Group [35]	"A group for caregivers taking care of Dementia and Alzheimer's afflicted loved ones. Designed to stimulate conversation and ensure all members a safe haven in posting personal feelings."	• Caregivers seek support for general challenges with caregiving • Caregivers seek support for caring for loved ones with neurocognitive behavioral changes and degenerative diseases such as delirium and dementia
Facebook Empowering Caregivers Support Group [36]	"A public group created so that there is a place for family members to turn to when in need of mental, emotional, and informational support when it comes to taking care of aging family members."	• Caregivers seek support for the day-to-day challenges of aging family members • Caregivers seek support on specific information such as health system challenges and legal issues
Mayo Clinic Connect [37]	"An online community, connecting patients and family caregivers with each other. You do not have to be a Mayo Clinic patient or caregiver to join the conversations."	• Many forum members are caregivers of patients with chronic medical conditions, such as cancer, who receive treatment at the Mayo Clinic, though this is not required • Caregivers typically seek support for medical challenges, such as medication side effects

[a]Caregiver Action Network's caregiver forum was on its website at the time of the research. It has since moved to Facebook.

Data Analysis

Two researchers (MA and CE) extracted the first 6 responses from the first 20 questions and comments to appear posted by the caregivers in each of the 5 forums. We chose this method based on the passive or unobtrusive method, which involves techniques that allow researchers to collect data without actively engaging with participants or influencing the conversation [38]. We did not extract questions and responses that did not contain comments relating to caregiver support (spam posts) because they were irrelevant to the research topic. To protect the identity of all posters on the web-based forum, we removed any names that appeared in the comments.

We used a subjective approach to gain insights into the caregiver forum participants' perceptions of their experiences. Specifically, we adopted an approach based on grounded theory using an iterative inductive process for our qualitative thematic analysis. As such, the number of comments and posts collected were adjusted iteratively until 2 authors (MLF and CE) separately believed that thematic saturation within forums' comments and posts was reached [39]. Similarly, the authors began with 2 forums and then added the next highest publicly accessible forum listed on Google until they achieved thematic saturation across forums. Using a codebook thematic analysis approach to examine the data [40] with Dedoose (SocioCultural Research Consultants), 2 researchers (MLF and CE) read through all the posts and coded the data. The entire author group discussed the codes, refined them reiteratively, and identified themes within the data. The author group also resolved any coding disagreements and resulting frequencies of each theme through discussions and collective agreements. They also examined the data to ensure that the chosen themes accurately reflected the data. Additionally, our author group comprised researchers with varying caregiving experiences. However, none of the researchers have intense caregiving experience that would deeply influence the research.

Ethical Considerations

The institutional review boards at Northern Illinois University and Seattle University determined the study to be exempt from institutional review board review in accordance with federal regulation criteria. The data used in this study were collected from forums that are easy to access and generally open to the public. All data were deidentified before analysis, including removing usernames and other naming data to maintain anonymity and confidentiality. There was no contact between the researchers and the posters.

XSL•FO
RenderX

Results

Overview

The data consisted of 100 initial posts and 600 responses to the initial posts or conversations stemming from the initial posts. Over half of the initial posts included specific questions, while the others shared reflections or experiences. The length of the initial posts and responses varied from a few with just a short sentence to others that were a couple of paragraphs. A few stretched beyond 500 words—a full page of single-spaced text. Response lengths were similar, with ranges from a few words to more than 500.

Posters (our term for individuals posting anything, including initial posts and responses) to the forums discussed care recipients with various diseases and cited dementia and cancer the most. We found some variation in the care recipient's disease based on the forum used. The Mayo Clinic Caregivers Forum included more clinical posts, with posters sharing specific disease states. In the Facebook Dementia Caregivers Support Group, some posters asked about what medications had worked for other care recipients and for strategies to obtain particular treatments, with almost all having to do with dementia. In the other forums, the care recipient's diseases often were not apparent, with many posters omitting such details.

Initial posts often included significant details describing the poster's life with their care recipient. Some posters responded with advice or feedback on the situation, while others shared their own experiences with their care recipients. Below, we present 5 key domains that resulted from the themes.

Domains

Handling Interpersonal Challenges

In these forums, posters shared their varying daily interactions, ranging from toileting and feeding to medication management and household chores. These interactions sometimes led to interpersonal challenges with the person receiving care or others involved in the household, with such negative interactions recorded 123 times in these posts. One poster noted his care recipient wife was upset with how he did things around the house.

> *She is very exacting on how she wants things done. I am hard of hearing and she gets upset when she has to repeat herself. If I guess at what she wants and get it wrong, oh well, you get the picture. I have to control myself better and be there for her. I have to swallow my pride and do what is right for her. It is not easy.* [Mayo Clinic Connect]

In total, 10 posts described challenges related to changing relationship roles. Husbands and wives recounted working to feed their spouses and care for them throughout the day. The posters shared their feelings, which often included frustration and sadness about these changes.

> *I lost an equal parent, as well as my romantic and "relational" partner, partly. It's a loss that is fully understood only by other spousal caregivers. It hasn't gotten any easier after 6 years.* [Caregiver Action Network]

Some posters also described negative situations with extended family members, with such interactions reported 77 times. Posters shared that some siblings were only concerned about financial arrangements after death, while other siblings were described as expressing negative attitudes and overall dissatisfaction with care.

> *She is very difficult to deal with. She questions everything and does not trust me to care for our parents in a manner that is in THEIR best interest. Without details suffice to say my husband and I gave up our dreams to take care of my parents but she berates me, tries to put my other siblings against me, says I lie, and accuses me of taking things.* [Facebook Empowering Caregivers Support Group]

Other posters described a lack of family support for the care recipient and feeling isolated as the main caregiver. This was highlighted 33 times in these posts, with some posters, noting that they did not get breaks or have family visitors who could relieve the ever-present demands of caregiving. Other posters wrote that this constant responsibility negatively impacted their relationships with their own spouses and children.

> *My husband and kids are not getting the best version of me. I can't stand having sex now because I spend so long caring for my parents the idea of sharing intimacy with my husband is actually repulsive to me. I don't enjoy anything that I'm doing.* [AgingCare]

Navigating Complicated Systems

Posters asked and responded to questions about accessing resources related to health care, living arrangements, and financial and legal systems. They mentioned health care and health insurance systems in 51 posts. Other posters responded by sharing their experiences with the health care system as a guide for others and sometimes suggesting various social services. Information on accessing home health aides, hospice care, and various medicines was requested 47 times in posts. Ten discussed necessary steps for care recipients to qualify for nursing home access through Medicaid, such as spending down funds or putting assets into a trust.

> *I can't wait until my mom is approved for Medicaid. She's in the nursing home. We [are] just waiting on that. I spent all her money. Most of it went to the nursing home. Also when I sent information to Medicaid I sent copies of all checks to show where her money went.* [Facebook Dementia Caregivers Support Group]

> *I have been her sole caregiver for about 5 years and I know caring for her now is beyond my capabilities. What do people do when they can no longer keep their parent in the home? Money is a huge issue, her insurance doesn't want to cover anything. I don't know what to do at this point. She needs more than I can give her.* [Facebook Empowering Caregivers Support Group]

XSL•FO
RenderX

Posters expressed concern about the quality of care hospice or nursing home workers provided to their care recipients 23 times and asked for feedback on how to advocate. They described their situation and asked for advice on what to do when the nursing care was not working out as expected.

> *The nurses have asked that when I come to visit I help them get her to the shower ... I can't be there all the time to make sure she gets up, takes a shower, just participates with life? The CNAs try to help, the nurses have told me that they can't force her because they could get into trouble.* [AgingCare]

Posters responded with messages stating how they understand these caregivers' concerns. They offered ideas on how they had dealt with similar situations and shared details about similar struggles they faced in working with care homes or hospice providers.

Posters described legal and financial concerns related to caregiving 124 times and asked others to weigh in on their options. Posters sometimes asked how to work within the legal system to help their loved ones. Some individuals receiving care wanted the informal caregiver to inherit a greater portion of assets upon their death, and posters asked about how to make the care recipient's wishes clear. Respondents offered suggestions to these legal and financial concerns, often urging posters to consult with professionals.

> *If your mom is of sound mind and wants to make this change, that's her decision. With the no-contest order and the doctor's assessment, legal challenges from your siblings might not hold up. Considering your exhaustion, it might be worth seeking legal advice to prepare, but ultimately, your well-being is the priority.* [Facebook Empowering Caregivers Support Group]

Gathering Tactical Coping Strategies

Posters described hundreds of challenges related to their caregiving roles. They asked questions about how to address these concerns, and respondents offered ideas or shared how they had responded in similar circumstances.

One common theme was difficulties related to hygiene among those receiving care, highlighted in 18 posts. Posters described challenges related to body odors and getting care recipients to shower or bathe. One poster asked for help with their care recipient who was touching foods in their kitchen pantry, which caused the poster concern about contamination of foods due to dirty hands. Fellow posters pitched in with various ideas.

> *Find a lock for the pantry but leave snacks out for him on the counter that are just for him. My Mom goes prowling for snacks at night so we leave some out for her.* [Facebook Dementia Caregivers Support Group]

> *I keep extra bags of chips because I can't eat out of the ones already opened. I have to constantly remind my [loved one] to wash her hands so I know if I'm not here they are probably not getting washed and she takes her dentures in and out of her mouth which totally grosses me out.* [Facebook Dementia Caregivers Support Group]

Concern about a loved one eating too much sugar or other nutrition concerns were mentioned in 21 posts. Respondents said that they experienced the same problems and highlighted that desire for sweets is common among people with dementia. Other posters responded with a variety of recipes and food ideas.

Posters expressed a desire for a break from caregiving duties 47 times. Respondents offered helpful tips, such as enlisting community or church volunteers. Enrollment in palliative or hospice care was mentioned in 55 posts, and posters shared their experiences engaging in those services.

> *It wasn't until she was in home hospice that they were ever able conquer her incredibly high anxiety. We certainly realized no course of care is ever perfect and always knew palliative care was her only choice given her situation. We never regretted the care she received.* [Mayo Clinic Connect]

Managing Emotions

Posters expressed emotions in their comments 180 times. Some described themselves as being "stressed," "overwhelmed," "weary," or "experiencing burnout." They talked about their emotions, which often included anger, depression, sadness, and resentfulness. Posters mentioned feelings of guilt 32 times, and some described guilt for having negative feelings about caregiving and their loved ones.

> *I feel so guilty for the anger and resentment I feel. I always thought in our 50's we would be traveling and enjoying ourselves and there is none of that. Whenever I think he is finally improving enough that we can do something enjoyable, he comes up with a new problem or complaint, which leads to more Dr. appointments, tests and procedures ... I feel sorry for myself and wonder what horrible things I did to deserve the life I have now.* [Caregiver Action Network]

Posters said they cried frequently 7 times. They described facing challenging decisions about their loved one's care and the emotional toll of such choices.

> *My heart is breaking because I'm not sure if I'm ready to let him go but at the same time I know I can't keep living like this, I'm feeling guilty cuz it feels like I'm giving up on him ... This is the hardest decision I've ever had to deal with but I know it's the best decision, but then why do I feel so guilty?* [Mayo Clinic Connect]

Isolation was another common feeling described in 33 posts. Some lived in rural areas where there were few community groups or other social outlets. Others described a lack of connection to friends and other family members and difficulty engaging in employment.

> *My friendships are starting to be strained, because they don't understand why "I'm hardly ever available" and quite frankly I'm starting to get bitter how they can just stay out as long as they want*

XSL•FO
RenderX

> *without worry. Romantic relationships are nonexistent, as most girls immediately become disinterested when they find out about my home situation. And in general I've found it harder to relate with other people, as my only social contact is in class or work.* [Caregiver Action Network]

Posters shared negative emotions 123 times and reported anger about the way their care recipient's medical condition had impacted their lives, calling the situation unfair. Posters also described helping their care recipient deal with negative emotions 40 times, including anger about their diagnosis. Some posters detailed navigating emotional outbursts from the care recipient that were emotionally hurtful to the poster.

> *The last day that my Dad was home before going to a hospice facility, he kicked me when I was trying to clean him after a BM. He also punched at me and said "I'm going to beat you up!" Which wasn't his normal personality- he was the sweetest, gentlest Dad usually. I know he didn't feel good and he was embarrassed to need that much help, but I'm human and it still hurt my feelings.* [Facebook Empowering Caregivers Support Group]

A few posters expressed negative emotions toward themselves—saying they were "horrible" or a "monster" when describing difficult decisions involved in caregiving. This was often expressed along with frustration. One poster detailed how her husband had been put on hospice, then taken off, and lived more than 4 more years facing many health challenges. "WHY IS HE YET ALIVE[?]" she asked. Another described her internal wishes.

> *I pray every day my mother has a heart attack or a massive stroke.* [AgingCare]

While the majority of posts with emotion had a negative tone, 26 were positive. For example, some posters expressed fulfillment in their caregiving duties. Others said they were proud of the care they had been able to offer their loved ones, and some noted personal growth during the caregiving experience.

> *Over the years, I can see a huge change in myself. For one, I am more patient now. However, the goal line of ENOUGH patience keeps moving further away! I believe that my husband is providing me an opportunity to become a better person.* [Mayo Clinic Connect]

Connecting With Others in Similar Situations

Many posters expressed a desire to connect with others in similar situations, with web-based group support mentioned 301 times. As specific posters communicated more with one another, their comments took on a more familiar tone. A few even offered each other direct assistance.

> *If you are a man of faith, use it. If not, you may want to consider it and there are those who can help you with that. I'll help you with that if I can. There is LOTS of help out there and there are LOTS of people that understand and are ready to be your friend and your lifeline.* [Caregiver Action Network]

> *Know that there are people that care. Like [name redacted] said, he will help, I will too, we are both reaching out to you, you are not alone.* [Caregiver Action Network]

Spiritual conversations appeared in posts 36 times. In some instances, the posters shared a positive spiritual statement as an initial post. In other instances, respondents promised to pray for the initial poster regarding the challenging situation or their care recipient's medical condition.

> *As a caregiver myself to my husband, I've never felt more useless in my life. Feels like all you can do is watch and wait. There are days I'm afraid to go to sleep. I do pray for healing over your husband, and peace that surpasses all understanding for you. Bless you both.* [Mayo Clinic Connect]

Posters mentioned feeling lonely and shared their challenges with that across 20 posts. They brought up hobbies they no longer engaged in and not having time for friends or socialization. While respondents said that they had similar difficulty making time for themselves, they often prescribed self-care to one another.

> *Coping with the isolation of being a 24/7 caregiver is very challenging. It's crucial to seek support from friends, family, or here with caregiving friends who understand your journey. Schedule regular breaks, explore local resources, delegate tasks, and set boundaries, which are essential for your well-being.* [Facebook Empowering Caregivers Support Group]

Posters also commiserated about the many tasks involved in caregiving. They described their day-to-day activities and the challenges associated with them. Respondents said they do the same things and appreciated knowing others were facing similar circumstances.

> *I feel like I am forever walking uphill. Frankly, one of the most helpful things I'm finding on here is that I am not alone. I don't really need anyone to tell me what to do or fix this ... just to validate my feelings.* [Caregiver Action Network]

Discussion

Principal Findings

Our qualitative analysis of posts across 5 web-based caregiver forums showed that informal caregivers participate in these communities to share and receive moral and practical support. These forums functioned as a space for caregivers to disclose their daily struggles with those in similar circumstances. This was true across forums that offered more anonymity (AgingCare, Caregiver Action Network, and Mayo Clinic Caregivers Forum) but was also seen in the 2 Facebook groups that had the potential to be tied to caregivers' personal lives.

Our research identified 5 key domains explaining why people posted on these caregiver forums. Within these themes were consistent comments about being heard and supported,

XSL•FO
RenderX

highlighting the positive benefits that popular publicly accessible web-based caregiver forums can provide. Caregivers described putting significant time and effort into their caregiving, and their posts and stories showed how the role was a critical part of their lives. While the 5 domains we identified were similar to those found in prior research on web-based caregiver support that focused on particular topics related to caregiving, one website, or on caregivers of people with a specific diagnosis [17-30], this analysis demonstrated the consistency of these themes across several caregiver forums.

Many caregivers in this analysis experienced significant interpersonal tensions, including resentment and guilt when relating to the care recipient or other family members. Caregiving is often characterized by intense, intimate relationships with unexpected role changes and needs [41]. Research shows that spousal caregiving, in particular, can be more intense, with the informal caregiver experiencing feelings of depression and guilt [42,43]. Research also shows that caregivers caring for individuals younger than them may find more emotional, physical, and financial difficulties than their counterparts [44]. It is possible that caregivers anticipate caring for someone younger than them for a longer time, making it more difficult than caring for someone older. Similarly, our research indicates many caregivers are experiencing challenges with their relationships. These caregivers are reaching out on the web to vent and share their experiences with others, and this research shows that there are many places on the web where caregivers may find community among fellow caregivers likely to understand the relational intricacies and difficult situations they experience.

Informal caregivers used web-based forums to aid in navigating the maze of legal, financial, and health care systems. Previous research shows that informal caregivers experience significant difficulty identifying resources in the fragmented and complex health system and find a lack of consistency in access to community and home care [45]. In addition, there is insufficient community and health infrastructure for sharing resources with the caregiver community [45]. Caregivers use forums for aid related to logistical issues of caregiving in areas including legal, financial, and health care topics [23,32]. In our research, the problem-solving required to navigate these systems often stumped many caregivers and left them feeling frustrated and isolated. While respondents were frequently able to provide helpful steps to guide the original poster toward the desired outcome, our findings support the need for additional informal caregiver support toward navigating the necessary solutions within health and community systems.

In our study, informal caregivers expressed persistent concerns about managing practical daily matters. The range of such daily demands on caregivers is vast, with tasks such as feeding and bathing, paying bills and shopping, caring for property and financial matters, and offering encouragement and companionship. Informal caregivers who help with a greater number of activities of daily living or independent activities of daily living experience more caregiver burden [46]. Additionally, caregiving often includes more medical-focused tasks; yet, practical daily tasks were the dominant type mentioned by posters in our study. Research indicates that support with medical-focused tasks from effective interventions and health care providers can help informal caregivers improve their caregiving experience [47,48]. Web-based caregiver forums are likely a better source for advice for practical daily tasks than medical-focused tasks, given that providing advice on the latter involves having medical expertise.

Every web-based forum we reviewed included numerous caregivers who expressed gratitude and encouragement related to connecting with others walking a similar path in life. Whether it was offering a step-by-step suggestion on how to address a problem, thanking each other for sharing, or simply saying, “Hang in there,” it was clear that the posters in our study appreciated the connection the forum provided. Similarly, a previous study analyzing web-based caregiver forums during COVID-19 identified a common theme of “celebrating” the virtual way of life [19]. In that study, posters highlighted the sense of community in the forum and how engaging with the site had helped them with caregiving during the pandemic.

Our analysis highlighted web-based forums as places where caregivers shared thoughts and sensitive caregiving details that could be considered socially inappropriate—things they might not say in an interview or to someone they know—such as comments on family disagreements, personal hygiene, and the caregiver asking “Why is he yet alive” about her spouse. This finding is consistent with another study that found participants in web-based caregiver groups felt safe sharing thoughts that were not socially acceptable [49]. Additionally, research shows that reciprocity, altruism, and empathy have a greater influence on knowledge sharing on web-based forums for nonhealth professionals than health professionals [50]. These motivations help explain the candid statements made by posters in our study and suggest that the informal caregivers want to help one another.

Many of the posts on the caregiver web-based forums in our analysis were emotionally charged, and posters seemed to use these communities as a way to vent negative emotions. Spousal caregivers, in particular, have been found to share more negative emotion and anxiety than other informal caregivers [22]. Furthermore, previous research found that informal caregivers who share negative aspects of caregiving on the web may ask others to confirm that their feelings are shared [32]. One reason that sharing frustrations on the web may help informal caregivers is that the negative emotions can be posted as soon as they experience the frustration [51].

Our study comes after the COVID-19 pandemic, a challenging time for informal caregivers, who often experienced increased burden from illness for themselves and their care recipient, had to navigate pandemic-related policies like medical office closings, and experienced social isolation [52,53]. Web-based discussion forums provided a medium for connecting with others and gaining information during the pandemic. Older adults, in particular, became more comfortable with using the internet and other technologies during the pandemic, with digital connectivity playing an increased role in their lives even after the pandemic [54]. With many older adults caring for spouses, siblings, and friends, access to web-based caregiver forums is

XSL•FO
RenderX

a tool that aging adults, for one, can continue to give and receive support.

Limitations

Our study had some limitations. Due to the anonymity of the web-based forum posters, information on their descriptive characteristics could not be obtained. Additionally, the web-based caregiver forums we examined were based in the United States. While the internet is available globally, it is likely that the posters were primarily based in the United States and experiencing caregiving in that country. This may limit the generalizability of the study findings on a global scale. For example, many health system challenges highlighted in the study are specific to those in the United States. In addition, this research captures caregivers' posts at one point in time and may not be generalizable in the future, as the caregiver landscape changes over time. Finally, it is possible that some posters were not informal caregivers, even though the content seemed to be posted by informal caregivers.

Conclusions

This paper contributes meaningfully to the current state of the art by providing a rich understanding of informal caregiving through an innovative lens, that of web-based forums. From this work, researchers and practitioners can be better informed about the needs of informal caregivers beyond what such caregivers may mention in interviews or medical appointments. By illuminating the lived experiences shared in web-based forums, the study advances knowledge about caregiving.

Informal caregivers play an essential role in society. Each year, informal caregivers in the United States provide as much as 36 billion hours of care to adults with limitations in daily activities, and such caregiving is linked to significant emotional, physical, and financial burdens to informal caregivers [55]. Since most informal caregivers have internet access [8], accessing web-based discussion forums is a low-barrier method for them to reach a community of people who may be experiencing similar emotions and challenges. With the number of American people aged 65 years and older projected to jump from 53 million [1] to nearly 95 million by 2060, the need for informal caregiving is projected to grow substantially [56]. Our study provides insights into how informal caregivers use web-based caregiver forums and can inform health care providers and community groups who provide information, education, and support to patients and their informal caregivers. In summary, web-based caregiver forums enable informal caregivers to use modern technology to help meet their support needs and, in the process, help to show them that they are not alone in facing caregiving challenges.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

References

1. AARP and National Alliance for Caregiving. Caregiving in the United States 2020. AARP. 2020. URL: https://www.aarp.org/pri/topics/ltss/family-caregiving/caregiving-in-the-united-states/ [accessed 2025-01-09]
2. Reinhard SC, Young HM, Levine C, Kelly K, Choula RB, Accius J. Home alone revisited: family caregivers providing complex care. AARP. 2019. URL: https://www.aarp.org/pri/topics/ltss/family-caregiving/home-alone-family-caregivers-providing-complex-chronic-care/ [accessed 2024-12-30]
3. Keeton VF, Trask J, Whitney R, Bell JF. Overburdened and underprepared: medical/nursing task performance among informal caregivers in the United States. J Gerontol Nurs 2020 Sep 1;46(9):25-35. [doi: 10.3928/00989134-20200811-05] [Medline: 32845344]
4. Coughlin SS, Datta B. Predictors of unhealthy physical and mental days among informal cancer caregivers: results from the 2019 Behavioral Risk Factor Surveillance System survey. Support Care Cancer 2022 Mar;30(3):2163-2171. [doi: 10.1007/s00520-021-06635-6]
5. Pristavec T. The burden and benefits of caregiving: a latent class analysis. Gerontologist 2019 Nov 16;59(6):1078-1091. [doi: 10.1093/geront/gny022] [Medline: 29659788]
6. Pinquart M, Sörensen S. Correlates of physical health of informal caregivers: a meta-analysis. J Gerontol B Psychol Sci Soc Sci 2007 Mar;62(2):126-137. [doi: 10.1093/geronb/62.2.p126] [Medline: 17379673]
7. Petry SE, Lara L, Boucher NA. Older caregivers: who they are and how to support them. J Aging Soc Policy 2024 Jul 3;36(4):589-602. [doi: 10.1080/08959420.2022.2051683] [Medline: 35290168]
8. Shaffer KM, Chow PI, Cohn WF, Ingersoll KS, Ritterband LM. Informal caregivers' use of internet-based health resources: an analysis of the health information national trends survey. JMIR Aging 2018 Dec 18;1(2):e11051. [doi: 10.2196/11051] [Medline: 31518244]
9. Sitges-Maciá E, Bonete-López B, Sánchez-Cabaco A, Oltra-Cucarella J. Effects of e-Health training and social support interventions for informal caregivers of people with dementia—a narrative review. Int J Environ Res Public Health 2021 Jul 21;18(15):7728. [doi: 10.3390/ijerph18157728]

XSL•FO
RenderX

10. Denham AMJ, Wynne O, Baker AL, et al. An online survey of informal caregivers' unmet needs and associated factors. PLoS ONE 2020;15(12):e0243502. [doi: 10.1371/journal.pone.0243502] [Medline: 33301483]
11. Kim H, Lee S, Cheon J, Hong S, Chang M. A comparative study to identify factors of caregiver burden between baby boomers and post baby boomers: a secondary analysis of a US online caregiver survey. BMC Public Health 2018 May 2;18(1):579. [doi: 10.1186/s12889-018-5488-4] [Medline: 29720119]
12. Anderson JG, Hundt E, Dean M, Keim-Malpass J, Lopez RP. "The church of online support": examining the use of blogs among family caregivers of persons with dementia. J Fam Nurs 2017 Feb;23(1):34-54. [doi: 10.1177/1074840716681289] [Medline: 27920340]
13. Blusi M, Kristiansen L, Jong M. Exploring the influence of internet-based caregiver support on experiences of isolation for older spouse caregivers in rural areas: a qualitative interview study. Int J Older People Nurs 2015 Sep;10(3):211-220. [doi: 10.1111/opn.12074] [Medline: 25425070]
14. Bateman DR, Brady E, Wilkerson D, Yi EH, Karanam Y, Callahan CM. Comparing crowdsourcing and friendsourcing: a social media-based feasibility study to support Alzheimer disease caregivers. JMIR Res Protoc 2017 Apr 10;6(4):e56. [doi: 10.2196/resprot.6904] [Medline: 28396304]
15. Diefenbeck CA, Klemm PR, Hayes ER. "Anonymous meltdown": content themes emerging in a nonfacilitated, peer-only, unstructured, asynchronous online support group for family caregivers. Comput Inform Nurs 2017 Dec;35(12):630-638. [doi: 10.1097/CIN.0000000000000376] [Medline: 28742535]
16. McKechnie V, Barker C, Stott J. The effectiveness of an internet support forum for carers of people with dementia: a pre-post cohort study. J Med Internet Res 2014 Feb 28;16(2):e68. [doi: 10.2196/jmir.3166] [Medline: 24583789]
17. Tanis M, Das E, Fortgens-Sillmann M. Finding care for the caregiver? Active participation in online health forums attenuates the negative effect of caregiver strain on wellbeing. Communications 2011;36(1). [doi: 10.1515/comm.2011.003]
18. Horrell B, Stephens C, Breheny M. Online research with informal caregivers: opportunities and challenges. Qual Res Psychol 2015 Jul 3;12(3):258-271. [doi: 10.1080/14780887.2015.1040318]
19. Lindroos A, Douglas-Smith N. Exploring informal caregivers' well-being during COVID-19 through online discussion forums. Scand J Caring Sci 2024 Mar;38(1):104-113. [doi: 10.1111/scs.13199] [Medline: 37522268]
20. Menke M, Wagner AJM, Kinnebrock S. Communicative care in online forums: how burdened informal caregivers seek mediated social support. Int J Commun 2020;14:21 [FREE Full text]
21. Andréasson F, Andreasson J, Hanson E. Developing a carer identity and negotiating everyday life through social networking sites: an explorative study on identity constructions in an online Swedish carer community. Ageing Soc 2018 Nov;38(11):2304-2324. [doi: 10.1017/S0144686X17000551]
22. Donnellan WJ, Warren JG. Emotional word use in informal carers of people living with dementia: linguistic analysis of online discussion forums. JMIR Aging 2022 Jun 17;5(2):e32603. [doi: 10.2196/32603] [Medline: 35713942]
23. Du Y, Dennis B, Liu J, et al. A conceptual model to improve care for individuals with Alzheimer's disease and related dementias and their caregivers: qualitative findings in an online caregiver forum. J Alzheimers Dis 2021;81(4):1673-1684. [doi: 10.3233/JAD-210167] [Medline: 33967054]
24. Scharett E, Madathil KC, Lopes S, et al. An investigation of the information sought by caregivers of Alzheimer's patients on online peer support groups. Cyberpsychol Behav Soc Netw 2017 Oct;20(10):640-657. [doi: 10.1089/cyber.2017.0274]
25. Kruk B. 'I can't bear the thought that he might not recognise me': personal narratives as a site of identity work in the online Alzheimer's support group. Commun Med 2015;12(2-3):273-286. [doi: 10.1558/cam.18453] [Medline: 29048867]
26. Ferrell EL, Russin SE, Hardy RM. Informal caregiving experiences in posttraumatic stress disorder: a content analysis of an online community. J Community Psychol 2019 May;47(4):757-771. [doi: 10.1002/jcop.22151] [Medline: 30592051]
27. Lobo EH, Johnson T, Frølich A, et al. Utilization of social media communities for caregiver information support in stroke recovery: an analysis of content and interactions. PLoS One 2022;17(1):e0262919. [doi: 10.1371/journal.pone.0262919] [Medline: 35081150]
28. De Simoni A, Shanks A, Balasooriya-Smeekens C, Mant J. Stroke survivors and their families receive information and support on an individual basis from an online forum: descriptive analysis of a population of 2348 patients and qualitative study of a sample of participants. BMJ Open 2016 Apr 6;6(4):e010501. [doi: 10.1136/bmjopen-2015-010501] [Medline: 27053271]
29. Washington KT, Oliver DP, Benson JJ, et al. Factors influencing engagement in an online support group for family caregivers of individuals with advanced cancer. J Psychosoc Oncol 2020;38(3):235-250. [doi: 10.1080/07347332.2019.1680592] [Medline: 31690247]
30. Benson JJ, Oliver DP, Washington KT, et al. Online social support groups for informal caregivers of hospice patients with cancer. Eur J Oncol Nurs 2020 Feb;44:101698. [doi: 10.1016/j.ejon.2019.101698] [Medline: 31816508]
31. Ihrig A, Renner T, Muck T, et al. Online support groups offer low-threshold backing for family and friends of patients with prostate cancer. Eur J Cancer Care (Engl) 2019 Mar;28(2):e12982. [doi: 10.1111/ecc.12982] [Medline: 30569592]
32. Shoults CC, Rutherford MW, Kemp AS, et al. Analysis of caregiver burden expressed in social media discussions. Int J Environ Res Public Health 2023 Jan 20;20(3):1933. [doi: 10.3390/ijerph20031933] [Medline: 36767297]
33. Aging Care Caregiver Forum. AgingCare. URL: https://www.agingcare.com/caregiver-forum [accessed 2025-1-9]

XSL•FO
RenderX

34. Caregiver Action Network Community. Facebook. URL: https://www.facebook.com/groups/201714486219468 [accessed 2024-07-27]
35. Dementia Caregivers Support Group. Facebook. URL: https://www.facebook.com/groups/672984902717938/ [accessed 2024-07-27]
36. Empowering Caregivers. Facebook. URL: https://www.facebook.com/groups/empoweringcaregivers/ [accessed 2024-07-27]
37. Caregivers: discussions. Mayo Clinic Connect. URL: https://connect.mayoclinic.org/group/caregivers/ [accessed 2024-07-27]
38. Shaw EK. The use of online discussion forums and communities for health research. Fam Pract 2020 Sep 5;37(4):574-577. [doi: 10.1093/fampra/cmaa008] [Medline: 32050014]
39. Lim WM. What is qualitative research? An overview and guidelines. Aust Mark J 2024;Jul 25:14413582241264619. [doi: 10.1177/14413582241264619]
40. Clarke V, Braun V. Using thematic analysis in counselling and psychotherapy research: a critical reflection. Couns Psychother Res 2018 Jun;18(2):107-110. [doi: 10.1002/capr.12165]
41. Schulz R, Eden J, Committee on Family Caregiving for Older Adults, Board on Health Care Services, Health and Medicine Division, National Academies of Sciences, Engineering, and Medicine. Families Caring for an Aging America: National Academies Press (US); 2016. [doi: 10.17226/23606]
42. Baumbach A, Hughes MC, Liu Y. Challenges and coping strategies in transitioning from caregiving to widowhood: a systematic review. Res Aging 2024;46(9-10):535-547. [doi: 10.1177/01640275241254396] [Medline: 38742924]
43. Hansen T, Slagsvold B. The psychological effects of providing personal care to a partner: a multidimensional perspective. Health Psychol Res 2013 Apr 18;1(2):e25. [doi: 10.4081/hpr.2013.e25] [Medline: 26973910]
44. Huo M, Kim K. Lasting impact of relationships on caregiving difficulties, burden, and rewards. J Soc Pers Relat 2023 Jan;40(1):288-310. [doi: 10.1177/02654075221118095]
45. Kim B, Wister A, O'dea E, Mitchell BA, Li L, Kadowaki L. Roles and experiences of informal caregivers of older adults in community and healthcare system navigation: a scoping review. BMJ Open 2023 Dec 9;13(12):e077641. [doi: 10.1136/bmjopen-2023-077641] [Medline: 38070939]
46. Riffin C, Van Ness PH, Wolff JL, Fried T. Multifactorial examination of caregiver burden in a national sample of family and unpaid caregivers. J Am Geriatr Soc 2019 Feb;67(2):277-283. [doi: 10.1111/jgs.15664] [Medline: 30452088]
47. Sugihara Y, Sugisawa H. Influence of medical care tasks on subjective burden and gain among older adults' family caregivers: structural equation modeling for testing the role of formal and informal support. BMC Geriatr 2023 Oct 6;23(1):628. [doi: 10.1186/s12877-023-04348-5] [Medline: 37803249]
48. Hughes MC, Afrin S, Hamlish T. Effectiveness of skill-building interventions for informal caregivers of adults with cancer: a systematic review. J Cancer Educ 2023 Apr;38(2):390-397. [doi: 10.1007/s13187-022-02236-x] [Medline: 36372814]
49. Daynes-Kearney R, Gallagher S. Online support groups for family caregivers: scoping review. J Med Internet Res 2023 Dec 13;25:e46858. [doi: 10.2196/46858] [Medline: 38090796]
50. Zhang X, Liu S, Deng Z, Chen X. Knowledge sharing motivations in online health communities: a comparative study of health professionals and normal users. Comput Human Behav 2017 Oct;75:797-810. [doi: 10.1016/j.chb.2017.06.028]
51. White MH, Dorman SM. Online support for caregivers. Analysis of an internet Alzheimer mailgroup. Comput Nurs 2000;18(4):168-176. [Medline: 10939186]
52. Hughes MC, Liu Y, Baumbach A. Impact of COVID-19 on the health and well-being of informal caregivers of people with dementia: a rapid systematic review. Gerontol Geriatr Med 2021;7:23337214211020164. [doi: 10.1177/23337214211020164] [Medline: 34104687]
53. Liu Y, Hughes MC, Wang H. Financial train, health behaviors, and psychological well-being of family caregivers of older adults during the COVID-19 pandemic. PEC Innov 2024 Dec;4:100290. [doi: 10.1016/j.pecinn.2024.100290] [Medline: 38799257]
54. Martin W, Collett G, Bell C, Prescott A. Ageing, the digital and everyday life during and since the Covid-19 pandemic. Front Psychol 2023;14:1168340. [doi: 10.3389/fpsyg.2023.1168340] [Medline: 37829072]
55. Reinhard SC, Caldera S, Houser A, Choula R. Valuing the invaluable: 2023 update. ARCH National Respite Network and Resource Center.: AARP; 2023. URL: https://archrespite.org/wp-content/uploads/2023/03/valuing-the-invaluable-2023-update.doi_.10.26419-2Fppi.00082.006.pdf [accessed 2025-01-09]
56. Fact sheet: how the build back better framework will support the sandwich generation. The White House. 2021. URL: https://www.whitehouse.gov/briefing-room/statements-releases/2021/09/21/fact-sheet-how-the-build-back-better-framework-will-support-the-sandwich-generation/ [accessed 2024-12-27]

XSL•FO
RenderX

Edited by A Mavragani; submitted 28.07.24; peer-reviewed by B Davis, DB Loo, J Johnson; revised version received 06.12.24; accepted 08.12.24; published 17.01.25.

Please cite as:
Foster ML, Egwuonwu C, Vernon E, Alarifi M, Hughes MC
Informal Caregivers Connecting on the Web: Content Analysis of Posts on Discussion Forums
JMIR Form Res 2025;9:e64757
URL: https://formative.jmir.org/2025/1/e64757
doi:10.2196/64757

© Michelle L Foster, Chinenye Egwuonwu, Erin Vernon, Mohammad Alarifi, M Courtney Hughes. Originally published in JMIR Formative Research (https://formative.jmir.org), 17.1.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Planned Behavior in the United Kingdom and Ireland Online Medicine Purchasing Context: Mixed Methods Survey Study

Bernard D Naughton[1,2], MPharm, PhD

[1]School of Pharmacy and Pharmaceutical Sciences, Trinity College Dublin, University of Dublin, Dublin, Ireland

[2]Centre for Pharmaceutical Medicine Research, School of Cancer and Pharmaceutical Science, Kings College London, London, United Kingdom

Corresponding Author:
Bernard D Naughton, MPharm, PhD
School of Pharmacy and Pharmaceutical Sciences
Trinity College Dublin, University of Dublin
College Green, Dublin 2
Dublin, D02 PN40
Ireland
Phone: 353 18962824
Email: bernard.naughton@tcd.ie

Abstract

Background: Online medicine purchasing is a growing health care opportunity. However, there is a scarcity of available evidence through a behavioral lens, which addresses why consumers buy medicines online. Governments try to influence online medicine purchasing behavior using health campaigns. However, there are little data regarding specific online medicine purchasing behaviors to support these campaigns.

Objective: The theory of planned behavior explains that perceived behavioral control (PBC), attitudes, and norms contribute to intentions, leading to behaviors. This study challenges these assumptions, by testing them in an online medicine purchasing context. We asked: What is the role of attitudes, norms, and PBC in an online medicine purchasing context.

Methods: An anonymous online snowball convenience sample survey, including open and closed questions concerning online medicine purchasing, was implemented. The data were thematically analyzed until data saturation. The emerging themes were applied to each individual response, as part of a case-by-case narrative analysis.

Results: Of the 190 consumers from the United Kingdom and Ireland who consented to participate in the study, 46 participants had purchased medicines online, 9 of which were illegal sales. Of the 113 participants who demonstrated an intention to purchase, 42 (37.2%) completed a purchase. There were many cases in which participants demonstrated an intention to buy medicines online, but this intention did not translate to a purchasing behavior (71/190, 37.4%). Reasons for consumers progressing from intention to behavior are suggested to be impacted by PBC and attitudes. Qualitative data identified access to medicine as a factor encouraging online medicine purchasing behaviors and a facilitator of behavior transition. Despite understanding the importance of why some medicines required a prescription, which is described as an example of legal and health norms, and despite suspicion and concern categorized as negative attitudes in this paper, some participants were still buying products illegally online. Risk reduction strategies were performed by 17 participants (17/190, 9%). These strategies facilitated a transition from intention to behavior.

Conclusions: The study results indicate that a consumer's intention to buy does not automatically translate to a purchasing behavior online; instead, a transition phase exists. Second, consumers followed different pathways to purchase and used risk reduction practices while transitioning from an intention to a behavior. Finally, owing to the covert nature of online medicine purchasing, norms do not appear to be as influential as PBC and attitudes in an online medicine purchasing setting. Understanding how a consumer transitions from an intention to a behavior could be useful for researchers, health care professionals, and policymakers involved in public health campaigns. We encourage future research to focus on different consumer behavior pathways or ideal types, rather than taking a blanket approach to public health campaigns.

(JMIR Form Res 2025;9:e55391) doi:10.2196/55391

KEYWORDS

planned behavior; consumer behavior; perceived behavioral control; attitudes; online purchasing; medicine

XSL•FO
RenderX

Introduction

Background

Authenticity is an important consideration when buying any product online, especially medicines, as inappropriate access and poor-quality medicines can have catastrophic implications for human health [1], with some sales causing death [2,3] including several coroner report cases [4]. Furthermore, online medicine purchasing has increased in recent times due to increased convenience, access, and an improved consumer experience [5,6]. From 1999 to 2003 in the United States, there was an increase in online sales of prescription drugs from $160 million to $3.2 billion [7], and the act of purchasing medicine online is likely to increase in line with internet access. In the United Kingdom in 2020, 96% of households had internet access, up from 2017, when 90% of UK households had internet access, and 2006, when only 57% of households had internet access [8]. These increases in internet access have also influenced the rate at which goods and services are bought online, with 87% of adults having bought goods or services online in 2020, up from 53% in 2008 [8].

Despite the dangers and increasing rates of online medicine purchases, there has been a scarcity of research to examine online medicine purchasing behavior. This paper examined online medicine purchasing through a behavioral model. The theory of planned behavior literature describes that past behavior, attitudes, perceived behavioral control (PBC), and norms affect a consumer's intention and that this intention leads to a behavior. We challenge this assumption in the online medicine purchasing context and propose that there are unexplored contributing factors that affect a consumer's likelihood of moving from an intention to a behavior. We did not aim to consider specific process variables such as the role of internal and external drivers on the translation from intention to behavior. Instead, we argued the presence of a transition phase that exists between an intention and a behavior and the core factors that specifically influence this transition. Although we see the merit in more granular research, it was not within the scope of this paper to pursue this.

Theory of Planned Behavior

Overview

This article examined human behavior in an online medicine purchasing context. Through a theory of planned behavior lens, we asked: What does the relationship between intentions and behaviors look like when buying medicine online? The theory of planned behavior is a social cognitive model in psychology that concerns intentional behavior; it was constructed by Ajzen in 1985 [9] and 1991 [10] and was originally known as the theory of reasoned action [11-15]. The theory identifies 3 determinants of intentions: (1) the attitude toward the behavior, which can be either positive or negative; (2) subjective norms, which are a person's perceptions of the social expectations to adopt a behavior; and (3) PBC, which is a person's perception of how easy or hard it might be to perform a behavior [16]. Since the book published by Fishbein and Ajzen in 2010 [13], this theory has also been known as the reasoned action approach. Although this theory was constructed over 50 years ago, it is still used in contemporary times, with Hagger et al [17] applying the reasoned action approach to health behavior in 2018 and contributing "past behavior" to the framework [18].

Theory of Planned Behavior: Strengths and Limitations

The theory of planned behavior is the most influential and highly cited model for the research and prediction of human behavior and is therefore a suitable theory to apply in this study [19]. However, like most behavioral theories, it has its limitations and criticisms in the literature when used in qualitative research [19]. One of the main limitations, which we addressed in this study, is the intention-behavior gap, as several studies consider intentions to be a poor indicator of behaviors [18,20] even over short time frames. The theory also does not take into consideration emotions or irrationality. The absence of emotion is important, especially when dealing with the purchase of health products as one's health or the health of a loved one can have significant emotional impacts [21]. The lack of a focus on irrational behavior is an area of the model on which this study touched [19].

Online Medicine Purchasing Context

The at-distance medicine sales context is different to in-person sales for a variety of reasons. In an at-distance sales environment, attitudes, which are affected by suspicion or distrust, are augmented due to diminished seller-consumer accountability and a lack of physical interaction between a consumer and a health care professional, a lack of trust between the consumer and seller, and a fear of online fraud, often due to reports of rogue online traders [22-25]. Furthermore, norms are silenced at the point of sale, as the purchase usually occurs in isolation and is removed from direct and indirect social influence due to consumers commonly purchasing and administering the medicine covertly. Finally, PBC is augmented in an online context, due to the ability to buy products that may be prohibited or restricted, resulting in some patients buying illegal medicines online that they would not otherwise have access to. This context causes an augmentation of PCB when compared to in-person purchasing. We wished to understand if the transition from intention to behavior is as straightforward as explained in the literature or if there is a further unexplored nuance when applied to the online medicine purchasing context [26].

The constructs of planned behavior are relevant to the purchase of counterfeit luxury goods [27-29]. There are also many studies in the medical sciences literature that investigate issues concerning online pharmacies [5,22,30-38]. Based upon our literature review, there is a scarcity of published research concerning online medicine purchasing behavior research examined through a theoretical lens. This lack of research may be due to difficulties accessing patients who are willing to admit to buying medicines online for fear their activity may be illegal.

As part of a preprint report of a qualitative study conducted in the United Kingdom in 2023, Almomani and colleagues [39] interviewed 20 participants who had bought prescription-only medicines (POM) from the internet after 2019. They indeed acknowledged the reluctancy of consumers to disclose such activity because of its illegal connotations. The same research

XSL•FO
RenderX

group also explored consumers' beliefs by reviewing news media articles [40]. During analysis, they applied the theory of planned behavior model and observed behavioral beliefs (risks and benefits of purchasing), social beliefs (social influencing factors [eg, health care professionals and influencers]), and control beliefs (facilitators of purchasing [eg, accessibility, medicine shortages]) as influencing the likelihood of people buying POMs online. Hall and Antonopoulos [41] also conducted some behavioral research in this context. They used a virtual ethnography approach that provided qualitative evidence of the link between falsified medicine and online sales through the advancement of the field of demand dynamics and provided qualitative data that support their propositions and the findings of this paper.

This work is insightful but could be built upon through a more in-depth use of theoretical lenses. The rise of internet purchasing, the effect of poor-quality medicines on health [2], and the risks associated with buying potent medicines on the web [34,37] make this research and context a challenging area of unique and increasing importance in online consumer behavior research, where qualitative data are difficult to attain. The aim of this study was to gain a broad understanding of online medicine purchasing behaviors through the lens of planned behavior. In doing so, we questioned the role of attitudes, norms, and PBC in an online medicine purchasing context and how they affect the intention-behavior gap.

Methods

Study Design

This study used a mixed methods survey study design that posed open and closed questions to 190 members of the public. Due to the illegal nature of some online medicine purchases and the challenges of obtaining data from this population, an online convenience snowball sample survey was implemented. This approach was chosen to encourage anonymous participation. This included broad open-ended and closed-ended questions regarding the purchase of medicines online using Google Forms (Multimedia Appendix 1). The theory was developed subsequent to data collection [42]. This mixed methods survey was designed according to the CHERRIES (Checklist for Reporting Results of Internet E-Surveys) checklist [43].

Data Analysis

The study involved a qualitative relational analysis that used thematic evaluation and inductive reasoning to identify themes to help better understand the relationship between intentions and behaviors in the online medicine purchasing context. Quantitative analysis used percentages to depict descriptive quantitative data and was generated from closed answer questions. This quantitative data supported the qualitative inductive reasoning, which proposed a novel transition phase between intentions and behaviors and the factors affecting this transition. Based on the emerging themes, a case-by-case narrative analysis was performed to further understand how these themes affected individual consumer purchasing journeys. This involved analyzing each case to understand the role that each theme played in their transition from intention to behavior. Percentages were rounded to 1 decimal place.

Sampling Methodology

The recruitment approach was intended for scale and breadth using an established methodology [44,45] involving a convenience snowballing sampling method. Convenience sampling has its advantages: It is a useful tool for exploratory research in areas where there is little or no existing literature or where resources are limited. Convenience sampling can also generate large amounts of data in a short space of time and can identify hard-to-reach populations, such as willing participants who have purchased medicines online [46]. The survey contained closed questions to record the cohort demographics and open questions concerning the purchase of online medicines to facilitate qualitative data analysis. The data were collected in spring 2018 prior to the COVID-19 pandemic, which is a study limitation and should be considered within the context of that era. The inclusion criteria in this study were English-speaking internet users as per the methodology in [45].

Study Recruitment

This was an open survey disseminated via social media. Therefore, this study was specifically representative of social media users and may not be representative of the wider public. The survey took approximately 10 minutes to complete, and the data were stored in an online database. No personally identifiable information was gathered. An invite to this survey was posted internally by researchers' academic affiliations and the lead researcher's professional LinkedIn and X (previously known as Twitter) and personal Facebook accounts. Reposting was encouraged to promote a snowballing effect. The survey contained a review step that allowed participants to go back to amend their responses. Cookies and IP checks were not used in this study. Only completed surveys were analyzed. The survey was posted and reposted through social media outlets for 4 weeks by the lead researcher, his contacts, and their connections. Payments or incentives were not provided to participants.

Study Pilot and Consent

An anonymous, online, mixed methods survey, with a randomized response recording, was developed by the lead researcher and piloted with members of the research team and professional colleagues of the lead researcher to improve the survey questions. This included a review to ensure the online survey operated as intended, the questions were clear, and the survey was free of typos or grammatical errors (Multimedia Appendix 1). As part of the consenting process, the participants read study information and ticked a box consenting to being involved in the research. (Multimedia Appendix 1).

Product and Sale Type

A free-text option in the survey allowed participants to identify the medicine or medicine type that they purchased online. These medicines were separated into the 3 most common UK legal classes: POM, pharmacy-only (P) medicines, and general sales list medicines. Some medicine classes differ based on the strength, quantity, formulation, and country in which they are purchased. Therefore, the medicines mentioned in this study were listed according to the most frequent class for each product in the United Kingdom. The medicines type, whether clinical questions were asked at the time of purchase, and whether a

XSL•FO
RenderX

prescription was required were recorded to understand if the medicine was purchased legally or illegally. For the purposes of this study, an illegal sale was classified as one in which consumers were not asked for a prescription or asked medical questions when buying products that were considered as POM within the purchasing country.

A P medicine is one that does not require a prescription but can only be sold in a pharmacy. A POM medicine is one that requires a prescription and can only be supplied in a pharmacy. In cases where medicine class was uncertain (such as when the drug was listed as "Hair loss," which could be a P medicine (minoxidil) or POM medicine (finasteride), the drug was classified as P.

Ethical Considerations and Approval

This research was classified according to the National Institute for Health and Care Research guidelines. It was reviewed and received ethical approval from Keele University (reference number ERP2369). Participants provided informed consent to participate. Study data were anonymous. Participants did not receive compensation for participation.

Results

This study involved a survey that generated qualitative and quantitative results that informed the contributions of this paper. The key themes were subjective norms, attitudes, and PBC.

Demographics

This study attracted 227 respondents, of which 190 were living in the United Kingdom or Ireland and consented to the study (Table 1). The largest group (78/190, 41.1%) of participants was aged between 30 years and 39 years. There were more female respondents than male respondents (Table 1). The majority (135/190, 71.1%) of respondents lived in the United Kingdom, and 28.9% (55/190) lived in Ireland (Table 1). The study also included respondents with a variety of employment statuses and education levels. The majority (170/190, 89.5%) of respondents were employed full time or part time. Full-time students (13/190, 6.8%) and the unemployed (7/190, 3.7%) were also represented (Table 1). About one-third of participants (55/190, 28.9%) had health insurance, 64.7% (123/190) of participants lived in a country that provided free health care, and 6.3% (12/190) of participants did not have health insurance and did not live in a country that provided free health care.

XSL•FO
RenderX

Table 1. Demographic characteristics of the study participants (N=190).

Characteristics	Results, n (%)
Gender	
Male	67 (35.3)
Female	123 (64.7)
Age (years)	
19-29	64 (33.7)
30-39	78 (41.1)
40-49	29 (15.3)
50-59	14 (7.4)
60-69	4 (2.1)
70-79	1 (0.5)
Country	
United Kingdom	135 (71.1)
Ireland	55 (28.9)
Employment status	
Full time	144 (75.8)
Part time	26 (13.7)
Full-time student	13 (6.8)
Unemployed	7 (3.7)
Education level	
Primary school	1 (0.5)
Secondary school	24 (12.6)
Undergraduate	85 (44.7)
Postgraduate	66 (34.7)
PhD	14 (7.4)
Health insurance status	
Free national health care	123 (64.7)
Health insurance	55 (28.9)
Pay for health care when required	12 (6.3)

Medicines Purchased

Of the 190 study participants, 46 had purchased medicines online. These medicines included a wide variety such as antibiotics, antimalarials, controlled drugs such as diazepam, and cosmetic medicines such as minoxidil for hair loss and Orlistat for weight loss (Multimedia Appendix 2). As all study participants were from the United Kingdom and Ireland, these medicines were categorized according to UK medicine classifications.

Health and Legal Norms

The participants in this study were a largely well-educated population from the United Kingdom and Ireland (Table 1) who understood why prescriptions were needed. They could identify doctors, pharmacists, dentists, opticians, and nurses or a mixture of all 5 as the main professionals legally permitted to supply a prescription or drugs. The majority (189/190, 99.5%) of respondents knew that all medicines did not require a prescription and were aware of reasons why some medicines required a prescription and others did not. Respondents, which are numbered (#), gave many legitimate reasons why prescriptions were needed, for example:

> *Monitor drug usage,...ensure correct medications are given.* [#159]
>
> *To legally obtain drugs.* [#152]
>
> *To prevent risk to patients.* [#141]
>
> *To allow you to legally obtain medicines.* [#79]

Although the norm within the cohort was an understanding that prescriptions were required for the safe supply of medicine, of the 46 respondents who bought medicines online, 18 of those respondents (39%) purchased POMs (according to the UK classification). Of those 18 participants, 17 either had health insurance or free health care provided by the state. Of the 18

XSL•FO
RenderX

participants who bought POM online, 78% (17/18) were not asked for a prescription when purchasing these POM online, and 50% (9/18) were not asked for a prescription or asked medical questions when buying POM, making these 9 sales likely to be illegal. In summary, an interesting contradiction existed. Despite an understanding of the safety reasons for why prescriptions were needed, some participants still completed an illegal purchase of a medicine online without a prescription.

Attitudes: Cost, Convenience, and Access

Participants were asked whether they would consider buying medicines online, and 59% (112/190) said that they would. In another question, participants were asked "On a scale of 1 to 10 (with 1 being unlikely and 10 being highly likely), how likely are you to buy a medicine online?" The total cohort gave an average mean response of 4.2, and those who had previously bought medicines online described a high likelihood, a mean average of 7.6, of buying medicines in the future.

There were several reasons in this study why consumers purchased their medicines online. Participants selected from a predetermined list of reasons for positive or negative attitudes that motivated consumers to buy medicine online, or they could free-type their response. The 2 reasons that contributed to positive attitudes toward online medicine purchasing behavior were convenience (18/46, 39%) and cost (22/46, 48%). A free-text space in this question resulted in 17 other responses. These responses largely focused on medicine access issues or due to their general practitioner (GP) not being willing to supply the medicine:

> *The specific medication that is most useful for my condition is not available in my country.* [#162]
>
> *Dr will only prescribe two weeks of treatment, despite chronic condition.* [#66]

The point relating to a GP's unwillingness to supply (diazepam) was mentioned in this study a second time, when the same participant stated:

> *I was desperate as local GP will only prescribe very short-term treatment in low doses.* [#66]

Furthermore, despite buying and taking medicines bought online, 26% (12/46) of participants had negative attitudes toward online medicine purchasing. They explained that they were concerned that the medicine they purchased could be fake, yet these participants progressed to purchase.

PBC: Risk Reduction Measures

The participants demonstrated that they had the autonomy and capacity to mitigate the risks associated with online medicine purchasing. In Table 2, we see a variety of practices used by our study participants to do so.

Table 2. Risk reduction practices used by this study population to support the online purchase of medicines.

Risk reduction practice	Example quotes (participant number)
No checks	• "None, just had bought from it (online site) many times before but not medicines" (#34) • "The product is manufactured from the same supplier that a prescription could have given me. They are cheaper, and a prescription is not required in this case. So, the safety of the website is not a factor to me." (#137) • "I assumed it was safe." (#59) • "None, because the product works (this is bad I know)" (#24) • "Word of mouth reviews" (#4)
Recommendations and reviews	• "By looking for recommendations on online forums" (#74) • "Asked friends where they purchased from" (#3)
Using a well-known or reputable website or company	• "It was the website of a pharmacy I used to shop when I lived there." (#140) • "Bought through an online pharmacy that is linked to a high street pharmacy. They asked a variety of health-related questions." (#94) • "Bought from Amazon, reliable source" (#183)
Other risk-mitigating methods	• "Google search or company registration number check" (#158) • "I conducted extensive research to ensure that the website was genuine, authentic, regulated and safe." (#162) • "Research the company I was buying from. Check their address on Google earth, recognise the name of the company, ensure site was safe (Https) and not a redirect" (#144) • "Checked their registration from link to gov (government) website" (#37) • "Register with regulator GMP (Good Manufacturing Practice) approved" (#175)

Practices to support a move from an intention to a purchasing behavior are categorized in this study as a form of PBC. From the cohort of 46 participants who purchased medicines online, there were a variety of methods used by respondents to support the purchasing behavior. These risk reduction practices are categorized as themes, represented as quotes, and summarized in Table 2. These practices include reading reviews and using a well-known or trusted website among other methods.

Case-by-Case Narrative Analysis

As part of our case-by-case narrative analysis, which focused on the transition between intentions and behavior, we identified behavioral pathways or, as we call them, behavioral ideal types. In Table 3, we see that all participants in our study could be categorized into 6 patterns of behaviors or ideal types, which all involved an intention, an attitude (positive or negative), the presence or absence of risk reduction practices, and the presence

XSL•FO
RenderX

or absence of the completed behavior. Although norms play an important role in a consumer's intention to buy medicine online, they do not appear, in our data, to influence the transition from an intention to a purchasing behavior. This is possibly due to a lack of social influence at the point of purchase, as online medicine purchasing does not require social interaction with a pharmacist or other customers.

Table 3. The 6 ideal types identified in this study that describe the transition between intentions and behavior, mediated by attitudes and perceived behavioral control based on behavioral pathways revealed as part of the case-by-case narrative analysis (N=190).

Consumer ideal type	Intention	Attitude	Perceived behavioral control	Behavior	Participants, n (%)
1	Present	Unassessed	Unassessed	Absent	71 (37.4)
2	Present	Absence of a negative attitude	No risk reduction practices	Present	20 (10.5)
3	Present	Negative attitude	No risk reduction practices	Present	5 (2.6)
4	Present	Negative attitude	Risk reduction practices	Present	4 (2.1)
5	Present	Absence of a negative attitude	Risk reduction practices	Present	13 (6.8)
6	Absent	Unassessed	Unassessed	Absent	77 (40.5)

The case-by-case narrative analysis involved a review of each complete individual survey response to identify purchasing pathways from intentions to a purchasing behavior. On some consumers' pathways from intention to behavior, they encountered a negative attitude postintention, and in some ideal types, they used practices to reduce risk. Consumers commonly had either no intention to purchase and did not (77/190, 40.5%; ideal type 6) or had intentions to purchase and did not (71/190, 37.4%; ideal type 1). These individuals were not asked questions regarding attitudes or PBC as they had no experience with online medicine purchasing. There was also evidence that participants with intentions to purchase had no negative attitudes toward the purchasing behavior; therefore, they did not need to implement a risk reduction practice and completed the purchase (20/190, 10.5%; ideal type 2). There were also participants (ideal type 3), who had an intention to purchase had a negative attitude toward the behavior, and, without reducing the risk, completed the purchase (5/190, 2.6%). There were also 4 participants (4/190, 2.1%) who had an intention to purchase, experienced a negative attitude, and used a risk reduction practice before completing the purchasing behavior (ideal type 4). Ideal type 5 (13/190, 6.8%) included individuals who had an intention to purchase, appeared to demonstrate no negative attitudes, implemented a risk reduction practice in any case, and completed the purchase. Ideal type 3 is an example of irrational behavior, as these study participants had an intention to purchase, then had negative attitudes toward the behavior, and, without taking any steps to manage that attitude, they still purchased. When asked "What measures, if any, did or do you take, to make sure the website that you are buying from is or was safe?", one participant replied "None because the product works (this is bad I know)' which typifies irrational behavior" (participant #24).

Discussion

Themes

This study aimed to gain a broad understanding of online medicine purchasing behaviors through the theory of planned behavior lens with a focus on how intentions lead to behaviors and the intention-behavior gap. The dominant themes running through this paper concerned (1) health and legal norms; (2) attitudes toward online purchasing (positive: convenience and cost; negative: concern or suspicion); and (3) PBC, which presented as practices to minimize risk. The study also proposes that a transition stage (Figure 1) exists between intentions and behavior, which appears, in our data, to be influenced by attitudes and risk reduction practices but less so by norms.

Figure 1. An adapted theory of planned behavior model with red lines representing this study's proposed contributions, showing attitudes and perceived behavioral control as mediators of the transition from intention to behavior.

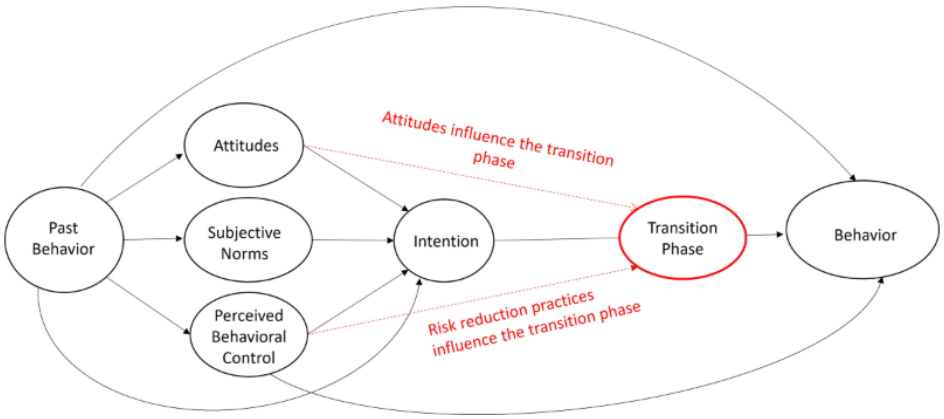

XSL•FO
RenderX

Theoretical Contribution

This paper identified qualitative and quantitative patterns within our data that support the use of the theory of planned behavior in this context [9,10] and described the influence of attitude, PBC, and norms on intentions and behavior. However, within this context, we suggest that the relationship between intention and behavior contains more nuance than is covered by Ajzen [9,10] and others [13,17]. We therefore adapted the model by Hagger et al [17] (Figure 1) to include our propositions, which are highlighted in red. Other studies support our proposal that intentions do not automatically translate to behaviors [26]. However, this paper reasons that a transition stage exists, where positive attitudes toward purchasing, such as convenience and cost, promote a transition from intention to behavior. Negative attitudes, which can emerge after the initial intention due to concern or suspicion, can inhibit a transition from intention to behavior [47], whereas PBC in the form of using risk reduction practices, such as reading reviews to mitigate negative attitudes, plays an important role in an actor's successful psychological transition from an intention to a purchasing behavior. These propositions are supported in part by the work of Sniehotta [26] and aim to help address the intention to behavior gap in the literature (Figure 1).

This proposed transition phase is interesting and suggests that intentions can change before completing a purchase. If this transition stage exists, then there is an opportunity to influence consumers between the intention and behavior stages, which is an opportunity for public health campaigners or health care professionals to encourage or discourage practices that influence the transition phase.

Norms: Health and Legal

There are several subjective norms that may affect online medicine purchasing behaviors. Norms differ in an online context because consumers have access to products that they would not normally have access to, due to availability issues or laws that prevent their purchase without a legal and valid prescription. In our study, the 2 most prominent norms that emerged included legal and health norms. There was no evidence from this study context that social norms negatively affected a participant's transition from intention to behavior. This may be due to the standard medical treatment setting, where consumers are influenced by health care professionals to conform to norms during diagnosis and medicine supply, or by fellow consumers within stores. However, this is not the case online. Remote purchasing of medicine online is a step removed from the influences in a traditional purchasing environment, whereby the patient is without pressure to conform to social norms as they do not see the clinician or pharmacist face to face and are free to request whatever medicine they wish, in whatever quantity they desire. The effects of being refused a request online are unlikely to be as impactful as being refused in person by a health care professional who may know you and your family or have full access to your personal details and records. Therefore, it is understandable that the consumers in this study that transitioned from an intention to a behavior did not raise any negative normative influences throughout the transition.

If the medicine received from an online pharmacy is legitimate, the risks relate to addiction, inadequate medical supervision, and poor condition management. If the medicine is poor quality or falsified, there are additional risks such as side effects, overdose, and death. Buying medicine online is not a problem if the website is legitimate. Although several consumers answered yes to "At any stage were you ever concerned that the product(s) you were buying may have been fake, falsified, substandard or of poor quality?", 7 of these (#3, #4, #24, #66, #74, #168, #182) still progressed to purchasing a product. From this evidence, we propose that the usual societal norms and legal medicine access norms in the online context do not appear to inhibit the transition from intention to behavior.

Attitudes: Cost, Convenience, and Access

We present, in this paper, that attitudes toward purchasing behavior were influenced by cost, convenience, and access, which is also described in other studies [22].

Hall and Antonopoulos [41] described cases of patients purchasing medicines online due to inconvenience or because they were struggling to see a GP, when it was possible to see the GP:

> *After trying to get a doctor's appointment for several weeks with no luck, I gave up and decided to go online to buy the drug I wanted.*

The idea that consumers were taking control of the situation themselves due to cost and convenience was also covered in another study by Almomani et al [22]:

> *There is no need to make an appointment with your GP. Well, it takes a couple of weeks to get an appointment in our surgery.* [2QL]

Medicine access was a qualitative theme that commonly emerged from the survey data, which also correlates with a study by Cicero and Ellis [32] and is argued to impact attitudes. The 2 reasons for buying tramadol online in the empirical study by Cicero and Ellis were (1) difficulty finding a doctor to prescribe the opioid medicine and (2) doctors who would not prescribe enough. Both points are access-related issues that affect PBC, appear to be augmented in the context of online addictive drug purchasing, and increase intentions to purchase. In the case of addictive medicines, access appears to be a key driver of intentions. Another example from this study was related to diazepam, and in that case, the participant who bought the diazepam online also explained that they were motivated to do so because their doctor refused to supply enough of the medicine.

In this case, the participant went on to buy this medicine online from a website that did not require an online consultation or a prescription. This is a case of a patient bypassing the GP to obtain medicines, which demonstrates autonomy. This is a common case that is also reflected in quotes gathered by Hall and Antonopoulos in their 2016 study [41].

The issue of limited access affects the participant's attitude toward the behavior and may be a driver to regain PBC that encourages online purchasing behavior which appears to be common in this study and the wider literature.

XSL•FO
RenderX

PBC and Risk Reduction Practices

We propose that PBC affects the psychological transition from intention to behavior. Consumers can assert their own control in a purchasing context to mitigate against risks that develop postintention and prebehavior, such as a consumer noticing something suspicious about an online pharmacy. Their PBC allows them to check the website URL against a list of approved websites or check other customer reviews to mitigate their negative attitudes toward that purchase [48-51].

Hall and Antonopoulos [41] identified the importance of finding a trusted illegal source.

In order to find a "Reputable" supplier, a consumer commonly relied on the use of risk reduction practices before purchase, which is a form of PBC. These practices included their own research, the company's reputation, and online reviews (Table 2). We reason that postintention risk reduction practices influence consumer transitions from intention to behavior. This is supported by evidence from Filieri et al [48] and Jin et al [49], who discussed the role of product reviews in online decision-making. When consumers encounter convincing information to mitigate identified risks, such as online reviews, they tend to complete the transition from initial intention to behavior [48]. Of course, there will be exceptions to this transition scenario, whereby consumers' intentional and logical behaviors are overtaken by impulsiveness behaviors [52], which is covered by consumer ideal type 3 in Table 3.

Behavioral Pathways: Public Health Policy Implications

A consumer's transition from intention to behavior, depicted as consumer ideal types in Table 3, acts as a lens through which to see consumer purchasing behavior patterns. This study has its limitations; however, these ideal types could be useful tools to help understand how different consumer groups transition from an intention to a behavior. This may assist with targeting individual psychological processes associated with norms, negative attitudes, PBC, and the facilitation of behavioral change through targeted public health campaigns. The current public health campaign policies concerning the purchase of falsified medicine online but do not appear to consider the different attitudes and PBC practices that affect individuals. Instead, public health campaigns either send an educational message about the risks of buying medicines online or try to incite fear into customers to deter them from the online purchase of medicine. Viewing consumers through different ideal types will provide public health campaign marketers with a new structured strategy to help them target specific behavior types. For example, consumers who demonstrate a negative attitude toward the online purchasing behavior but tend to continue from intention to behavior when adopting a risk reduction practice are unlikely to find a public health campaign that incites fear or suspicion as an effective deterrent. In those cases, these individuals would be best served by public health campaigns that describe safe and effective risk reduction practices such as explaining ways to verify the legitimacy of the website or the medicine they are purchasing.

Retail Pharmacy Marketing Implications

For legitimate online pharmacies, it may be useful to create marketing approaches that develop upon the principles described in this paper. These approaches may include the reassurance of consumers with negative attitudes and sign-posting reliable risk reduction practices to help reassure their customers. Reassurance may include regular blogs or posts concerning the risks of buying from illegal pharmacies; risk reduction practices to support patients, which may include searching for privacy or security statements [25]; or methods to verify online pharmacies or medicines they receive from pharmacies. These practices could be monitored using consumer metrics to understand which strategies were most effective at converting intentions to behaviors [53]. This argument may also hold true for marketers in other industries, as they too could run campaigns to raise awareness of illegal websites and ways to verify the legitimacy of an online retailer or the products they sell.

Limitations

This study is based on a relatively small sample of participants from the United Kingdom and Ireland and could benefit from being repeated as part of a large, multicenter, triangulated, mixed methods study that includes interviews, focus groups, and observations across different regions. In addition, the data were collected in 2018; therefore, the findings should be considered in light of that era. This study was also based on social media users, who may not represent a suitable cross-section of society. However, considering the sensitive nature of the subject, it can be difficult to recruit participants to such a study. This study is a first step to identifying the presence of certain factors facilitating or inhibiting a consumer's transition from an intention to a behavior in an online medicine purchasing context.

In conclusion, intentions do not always directly translate to a behavior in an online medicine purchasing context, which is an important consideration for researchers in this area. Instead, we suggest that the transition is mediated by PBC and attitudes. In many cases, those buying medicines online were concerned that their medicine may be a fake but still proceeded to purchase, which is concerning from a public health perspective. In cases where a complete transition from intention to behavior did occur, it was often in the absence of a negative attitude or the presence of a negative attitude coupled with a practice to reduce risk. The arguments in this article support the presence of a transition phase when buying medicine online. It is hoped that this paper's contribution will help public health campaigners and online pharmacy marketers to successfully target specific behavior types and support the use of consumer risk mitigation tools.

Data Availability

Primary study data are available on request.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
The questionnaire applied in this study.
[DOCX File , 19 KB - formative_v9i1e55391_app1.docx]

Multimedia Appendix 2
Medicines purchased by respondents in this study.
[DOCX File , 15 KB - formative_v9i1e55391_app2.docx]

References

1. Cinelli MD, LeBoeuf RA. Keeping it real: How perceived brand authenticity affects product perceptions. J Consum Psychol 2019 Jul 05;30(1):40-59. [doi: 10.1002/jcpy.1123]
2. Philby C. Death of medical student Sarah Houston after taking banned slimming drug Dinitrophenol highlights dangers of buying pills online, warns Government. Independent. 2013 Apr 23. URL: http://www.independent.co.uk/news/uk/home-news/death-of-medical-student-sarah-houston-after-taking-banned-slimming-drug-dinitrophenol-highlights-8584597.html [accessed 2025-02-02]
3. Forster K. Young mother dies after 'taking dodgy diet pills sold on Facebook'. Independent. 2017 Jul 05. URL: http://www.independent.co.uk/news/health/diet-pills-facebook-dead-young-mother-24-years-old-lucero-garza-mexico-monterrey-avitia-cobrax-a7824206.html [accessed 2025-02-02]
4. Aronson JK, Ferner RE, Richards GC. Deaths attributed to the use of medications purchased online. BMJ Evid Based Med 2022 Feb 21;27(1):60-64. [doi: 10.1136/bmjebm-2021-111759] [Medline: 34933931]
5. Fittler A, Ambrus T, Serefko A, Smejkalová L, Kijewska A, Szopa A, et al. Attitudes and behaviors regarding online pharmacies in the aftermath of COVID-19 pandemic: At the tipping point towards the new normal. Front Pharmacol 2022 Dec 23;13:1070473 [FREE Full text] [doi: 10.3389/fphar.2022.1070473] [Medline: 36642991]
6. Pantano E, Priporas C. The effect of mobile retailing on consumers' purchasing experiences: A dynamic perspective. Computers in Human Behavior 2016 Aug;61:548-555. [doi: 10.1016/j.chb.2016.03.071]
7. Jena AB, Goldman DP. Growing Internet use may help explain the rise in prescription drug abuse in the United States. Health Aff (Millwood) 2011 Jun;30(6):1192-1199 [FREE Full text] [doi: 10.1377/hlthaff.2011.0155] [Medline: 21565838]
8. Internet access – households and individuals, Great Britain: 2017. Office for National Statistics. URL: https://www.ons.gov.uk/peoplepopulationandcommunity/householdcharacteristics/homeinternetandsocialmediausage/bulletins/internetaccesshouseholdsandindividuals/2017 [accessed 2025-02-02]
9. Ajzen I. From Intentions to Actions: A Theory of Planned Behavior. In: Kuhl J, Beckmann J, editors. Action Control. SSSP Springer Series in Social Psychology. Berlin, Germany: Springer; 1985:11-39.
10. Ajzen I. The theory of planned behavior. Organizational Behavior and Human Decision Processes 1991 Dec;50(2):179-211 [FREE Full text] [doi: 10.1016/0749-5978(91)90020-T]
11. Ajzen I, Fishbein M. Understanding Attitudes and Predicting Social Behavior. Upper Saddle River, NJ: Prentice Hall; 1980.
12. Bamberg S, Ajzen I, Schmidt P. Choice of travel mode in the theory of planned behavior: the roles of past behavior, habit, and reasoned action. Basic and Applied Social Psychology 2010 Jun 07;25(3):175-187. [doi: 10.1207/s15324834basp2503_01]
13. Fishbein M, Ajzen I. Predicting and Changing Behavior: The Reasoned Action Approach. New York, NY: Psychology Press; 2010.
14. Gold GJ. Review of predicting and changing behavior: the reasoned action approach. The Journal of Social Psychology 2011 Apr 21;151(3):382-385. [doi: 10.1080/00224545.2011.563209]
15. Sheppard BH, Hartwick J, Warshaw PR. The theory of reasoned action: a meta-analysis of past research with recommendations for modifications and future research. Journal of Consumer Research 1988 Dec;15(3):325-343. [doi: 10.1086/209170]
16. Ajzen I, Madden TJ. Prediction of goal-directed behavior: attitudes, intentions, and perceived behavioral control. Journal of Experimental Social Psychology 1986;22(5):453-474. [doi: 10.1016/0022-1031(86)90045-4]
17. Hagger MS, Polet J, Lintunen T. The reasoned action approach applied to health behavior: role of past behavior and tests of some key moderators using meta-analytic structural equation modeling. Soc Sci Med 2018 Sep;213:85-94. [doi: 10.1016/j.socscimed.2018.07.038] [Medline: 30064092]
18. McEachan R, Taylor N, Harrison R, Lawton R, Gardner P, Conner M. Meta-analysis of the reasoned action approach (RAA) to understanding health behaviors. Ann Behav Med 2016 Aug 11;50(4):592-612 [FREE Full text] [doi: 10.1007/s12160-016-9798-4] [Medline: 27169555]
19. Ajzen I. The theory of planned behaviour: reactions and reflections. Psychol Health 2011 Sep;26(9):1113-1127. [doi: 10.1080/08870446.2011.613995] [Medline: 21929476]

XSL•FO
RenderX

20. Kor K, Mullan BA. Sleep hygiene behaviours: an application of the theory of planned behaviour and the investigation of perceived autonomy support, past behaviour and response inhibition. Psychol Health 2011 Sep;26(9):1208-1224. [doi: 10.1080/08870446.2010.551210] [Medline: 21678170]
21. Conner M, Armitage CJ. Extending the theory of planned behavior: a review and avenues for further research. J Applied Social Pyschol 2006 Jul 31;28(15):1429-1464. [doi: 10.1111/j.1559-1816.1998.tb01685.x]
22. Almomani H, Raza A, Patel N, Donyai P. Reasons that lead people to buy prescription medicines on the internet: a systematic review. Front Pharmacol 2023 Aug 31;14:1239507 [FREE Full text] [doi: 10.3389/fphar.2023.1239507] [Medline: 37719862]
23. Benedicktus RL, Brady MK, Darke PR, Voorhees CM. Conveying trustworthiness to online consumers: reactions to consensus, physical store presence, brand familiarity, and generalized suspicion. Journal of Retailing 2010 Dec;86(4):322-335. [doi: 10.1016/j.jretai.2010.04.002]
24. Hoffman DL, Novak TP, Peralta M. Building consumer trust online. Commun. ACM 1999 Apr;42(4):80-85. [doi: 10.1145/299157.299175]
25. Pan Y, Zinkhan GM. Exploring the impact of online privacy disclosures on consumer trust. Journal of Retailing 2006 Jan;82(4):331-338. [doi: 10.1016/j.jretai.2006.08.006]
26. Sniehotta FF. Towards a theory of intentional behaviour change: plans, planning, and self‐regulation. British J Health Psychol 2010 Dec 24;14(2):261-273. [doi: 10.1348/135910708x389042]
27. Gistri G, Romani S, Pace S, Gabrielli V, Grappi S. Consumption practices of counterfeit luxury goods in the Italian context. J Brand Manag 2009 Feb 6;16(5-6):364-374. [doi: 10.1057/bm.2008.44]
28. Phau I, Teah M, Lee A. Targeting buyers of counterfeits of luxury brands: a study on attitudes of Singaporean consumers. J Target Meas Anal Mark 2009 Mar 27;17(1):3-15. [doi: 10.1057/jt.2008.25]
29. Wilcox K, Kim HM, Sen S. Why do consumers buy counterfeit luxury brands? Journal of Marketing Research 2009 Apr 01;46(2):247-259. [doi: 10.1509/jmkr.46.2.247]
30. Bate R, Jin G, Mathur A. In whom we trust: the role of certification agencies in online drug markets. The B. E. Journal of Economic Analysis & Policy 2013;14(1):111-150. [doi: 10.1515/bejeap-2013-0085]
31. Cho S, Park HJ, Lee JH, Do J, Heo S, Jo JH, et al. Determination of anabolic-androgenic steroid adulterants in counterfeit drugs by UHPLC-MS/MS. J Pharm Biomed Anal 2015 Jul;111:138-146. [doi: 10.1016/j.jpba.2015.03.018] [Medline: 25880245]
32. Cicero TJ, Ellis MS. Health outcomes in patients using no-prescription online pharmacies to purchase prescription drugs. J Med Internet Res 2012 Dec 06;14(6):e174 [FREE Full text] [doi: 10.2196/jmir.2236] [Medline: 23220405]
33. Fittler A, Bősze G, Botz L. Evaluating aspects of online medication safety in long-term follow-up of 136 Internet pharmacies: illegal rogue online pharmacies flourish and are long-lived. J Med Internet Res 2013 Sep 10;15(9):e199 [FREE Full text] [doi: 10.2196/jmir.2606] [Medline: 24021777]
34. Gaudiano MC, Borioni A, Antoniella E, Valvo L. Counterfeit Adderall containing aceclofenac from internet pharmacies. J Forensic Sci 2016 Jul 06;61(4):1126-1130. [doi: 10.1111/1556-4029.13095] [Medline: 27364297]
35. Liang BA, Mackey TK, Lovett KM. Illegal "no prescription" internet access to narrow therapeutic index drugs. Clin Ther 2013 May;35(5):694-700. [doi: 10.1016/j.clinthera.2013.03.019] [Medline: 23597722]
36. Orizio G, Merla A, Schulz PJ, Gelatti U. Quality of online pharmacies and websites selling prescription drugs: a systematic review. J Med Internet Res 2011 Sep 30;13(3):e74 [FREE Full text] [doi: 10.2196/jmir.1795] [Medline: 21965220]
37. Rodomonte AL, Gaudiano MC, Antoniella E, Lucente D, Crusco V, Bartolomei M, et al. Counterfeit drugs detection by measurement of tablets and secondary packaging colour. J Pharm Biomed Anal 2010 Oct 10;53(2):215-220. [doi: 10.1016/j.jpba.2010.03.044] [Medline: 20547026]
38. Wang T, Hoag SW, Eng ML, Polli J, Pandit NS. Quality of antiretroviral and opportunistic infection medications dispensed from developing countries and Internet pharmacies. J Clin Pharm Ther 2015 Feb;40(1):68-75 [FREE Full text] [doi: 10.1111/jcpt.12226] [Medline: 25381836]
39. Almomani H, Patel N, Donyai P. Reasons that lead people to end up buying fake medicines on the internet: qualitative interview study. JMIR Form Res 2023 Feb 16;7:e42887 [FREE Full text] [doi: 10.2196/42887] [Medline: 36795460]
40. Almomani H, Patel N, Donyai P. News media coverage of the problem of purchasing fake prescription medicines on the internet: thematic analysis. JMIR Form Res 2023 Mar 21;7:e45147 [FREE Full text] [doi: 10.2196/45147] [Medline: 36943354]
41. Hall A, Antonopoulos GA. Fake Meds Online: The Internet and the Transnational Market in Illicit Pharmaceuticals. London, England: Palgrave Macmillan; 2016.
42. Corbin J, Strauss A. Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory. Thousand Oaks, CA: SAGE Publications Inc; 2008.
43. Eysenbach G. Improving the quality of web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34 [FREE Full text] [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
44. Middleton A, Morley KI, Bragin E, Firth HV, Hurles ME, Wright CF, et al. Attitudes of nearly 7000 health professionals, genomic researchers and publics toward the return of incidental results from sequencing research. Eur J Hum Genet 2016 Jan 29;24(1):21-29 [FREE Full text] [doi: 10.1038/ejhg.2015.58] [Medline: 25920556]

XSL•FO
RenderX

45. Middleton A, Bragin E, Parker M, Study D. Finding people who will tell you their thoughts on genomics-recruitment strategies for social sciences research. J Community Genet 2014 Oct 18;5(4):291-302 [FREE Full text] [doi: 10.1007/s12687-014-0184-2] [Medline: 24535681]
46. Baltar F, Brunet I. Social research 2.0: virtual snowball sampling method using Facebook. Internet Research 2012 Jan 27;22(1):57-74. [doi: 10.1108/10662241211199960]
47. Broniarczyk SM, Griffin JG. Decision difficulty in the age of consumer empowerment. J Consum Psychol 2014 May 24;24(4):608-625. [doi: 10.1016/j.jcps.2014.05.003]
48. Filieri R, Hofacker CF, Alguezaui S. What makes information in online consumer reviews diagnostic over time? The role of review relevancy, factuality, currency, source credibility and ranking score. Computers in Human Behavior 2018 Mar;80:122-131. [doi: 10.1016/j.chb.2017.10.039]
49. Jin L, Hu B, He Y. The recent versus the out-dated: an experimental examination of the time-variant effects of online consumer reviews. Journal of Retailing 2014 Dec;90(4):552-566. [doi: 10.1016/j.jretai.2014.05.002]
50. Khare A, Labrecque LI, Asare AK. The assimilative and contrastive effects of word-of-mouth volume: an experimental examination of online consumer ratings. Journal of Retailing 2011 Mar;87(1):111-126. [doi: 10.1016/j.jretai.2011.01.005]
51. Senecal S, Nantel J. The influence of online product recommendations on consumers' online choices. Journal of Retailing 2004 Jan;80(2):159-169. [doi: 10.1016/j.jretai.2004.04.001]
52. Baumeister RF. Yielding to temptation: self‐control failure, impulsive purchasing, and consumer behavior. Journal of Consumer Research 2002 Mar;28(4):670-676. [doi: 10.1086/338209]
53. Zenetti G, Klapper D. Advertising effects under consumer heterogeneity – the moderating role of brand experience, advertising recall and attitude. Journal of Retailing 2016 Sep;92(3):352-372. [doi: 10.1016/j.jretai.2016.02.004]

Abbreviations

CHERRIES: Checklist for Reporting Results of Internet E-Surveys
GP: general practitioner
P: pharmacy-only
PBC: perceived behavioral control
POM: prescription-only medicines

Edited by A Mavragani; submitted 11.12.23; peer-reviewed by A Fittler, D Dinh; comments to author 02.05.24; revised version received 23.05.24; accepted 05.01.25; published 21.02.25.

Please cite as:
Naughton BD
Planned Behavior in the United Kingdom and Ireland Online Medicine Purchasing Context: Mixed Methods Survey Study
JMIR Form Res 2025;9:e55391
URL: https://formative.jmir.org/2025/1/e55391
doi:10.2196/55391
PMID:

©Bernard D Naughton. Originally published in JMIR Formative Research (https://formative.jmir.org), 21.02.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Improving Suicidal Ideation Detection in Social Media Posts: Topic Modeling and Synthetic Data Augmentation Approach

Hamideh Ghanadian[1], BSc, MSc, PhD; Isar Nejadgholi[2], BSc, MSc, PhD; Hussein Al Osman[1], BSc, MSc, PhD

[1]School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, ON, Canada

[2]National Research Council Canada, Ottawa, ON, Canada

Corresponding Author:
Hamideh Ghanadian, BSc, MSc, PhD
School of Electrical Engineering and Computer Science
University of Ottawa
75 Laurier Ave E
Ottawa, ON, K1N 6N5
Canada
Phone: 1 6134000702
Email: hghanadian36@gmail.com

Abstract

Background: In an era dominated by social media conversations, it is pivotal to comprehend how suicide, a critical public health issue, is discussed online. Discussions around suicide often highlight a range of topics, such as mental health challenges, relationship conflicts, and financial distress. However, certain sensitive issues, like those affecting marginalized communities, may be underrepresented in these discussions. This underrepresentation is a critical issue to investigate because it is mainly associated with underserved demographics (eg, racial and sexual minorities), and models trained on such data will underperform on such topics.

Objective: The objective of this study was to bridge the gap between established psychology literature on suicidal ideation and social media data by analyzing the topics discussed online. Additionally, by generating synthetic data, we aimed to ensure that datasets used for training classifiers have high coverage of critical risk factors to address and adequately represent underrepresented or misrepresented topics. This approach enhances both the quality and diversity of the data used for detecting suicidal ideation in social media conversations.

Methods: We first performed unsupervised topic modeling to analyze suicide-related data from social media and identify the most frequently discussed topics within the dataset. Next, we conducted a scoping review of established psychology literature to identify core risk factors associated with suicide. Using these identified risk factors, we then performed guided topic modeling on the social media dataset to evaluate the presence and coverage of these factors. After identifying topic biases and gaps in the dataset, we explored the use of generative large language models to create topic-diverse synthetic data for augmentation. Finally, the synthetic dataset was evaluated for readability, complexity, topic diversity, and utility in training machine learning classifiers compared to real-world datasets.

Results: Our study found that several critical suicide-related topics, particularly those concerning marginalized communities and racism, were significantly underrepresented in the real-world social media data. The introduction of synthetic data, generated using GPT-3.5 Turbo, and the augmented dataset improved topic diversity. The synthetic dataset showed levels of readability and complexity comparable to those of real data. Furthermore, the incorporation of the augmented dataset in fine-tuning classifiers enhanced their ability to detect suicidal ideation, with the F_1-score improving from 0.87 to 0.91 on the University of Maryland Reddit Suicidality Dataset test subset and from 0.70 to 0.90 on the synthetic test subset, demonstrating its utility in improving model accuracy for suicidal narrative detection.

Conclusions: Our results demonstrate that synthetic datasets can be useful to obtain an enriched understanding of online suicide discussions as well as build more accurate machine learning models for suicidal narrative detection on social media.

(JMIR Form Res 2025;9:e63272) doi:10.2196/63272

KEYWORDS

suicidal ideation detection; large language models; synthetic data; topic modeling; social media

XSL•FO
RenderX

Introduction

Suicide is a pressing global public health issue with far-reaching consequences for individuals, families, and communities. The World Health Organization [1] has highlighted the alarming increase in suicide rates in recent years, emphasizing the urgent need for effective prevention strategies [2]. Traditionally, identifying individuals at risk of suicide has relied on clinical assessments and crisis hotlines. However, the widespread adoption of social media has opened new avenues for early detection and intervention [3].

Natural language processing (NLP) techniques have emerged as a promising tool for suicide detection, leveraging machine learning to analyze textual data shared on social media platforms [4-6]. Social media datasets have become invaluable in the study of suicidal ideation. As more people express their thoughts and emotions online, these datasets offer a unique glimpse into individuals' digital lives, providing rich data for research and mental health support [7]. Researchers use these datasets to analyze text, identifying patterns and linguistic cues associated with suicidal ideation [8-10].

However, while social media provides a wealth of information, it is not without its limitations. For instance, social media platforms tend to attract a younger demographic, which may skew the data and make them less representative of older populations. Additionally, the open nature of social media may encourage users to share certain aspects of their experiences while discouraging the disclosure of more personal or sensitive factors [11]. These biases can have significant implications for machine learning models trained on social media data. If the data are not representative of all demographics or fail to capture less frequently discussed factors, the models may struggle to generalize effectively. For example, a model trained on data that overrepresent certain themes or demographics may perform poorly when encountering posts from underrepresented groups or those addressing less common topics. This could lead to false negatives, where the model fails to identify at-risk individuals, or biased outputs, where the model perpetuates stereotypes or stigmatization present in the training data.

To address these limitations, it is crucial to complement social media datasets with information from other reliable sources. Within psychology, extensive research has been conducted to understand the underlying psychological, social, and environmental factors contributing to suicidal thoughts and behaviors. Several investigations have explored how psychological factors, including depression, anxiety, a sense of hopelessness, and feelings of low self-worth, influence the emergence of suicidal thoughts [12-14]. These studies have investigated the strong association between suicidal thoughts and conditions like depression [15,16], bipolar disorder [17], borderline personality disorder [18], and substance abuse [19,20]. The studies used a combination of quantitative and qualitative methods to explore the factors contributing to suicide [21-24].

Our objective in this study is to identify gaps between the risk factors represented in social media data and those established in psychology literature. We hypothesize that social media data do not fully capture empirically established factors related to suicidal ideation, particularly those that may be underrepresented due to demographic or disclosure biases. To examine this, we conducted a comprehensive analysis of suicide-related social media datasets using topic modeling techniques to identify themes and topics [25]. We then compared the extracted themes with factors outlined in psychology literature to locate any significant gaps. When such gaps were identified, we explored the potential of augmenting real-world data with synthetic data generated by advanced language models like GPT-3.5 Turbo to improve the representation of underreported factors. The use of synthetic data has gained traction in recent years, particularly with the advent of generative models that can produce high-quality, realistic data [26,27]. This study presents a novel approach for identifying and mitigating topic gaps in suicidal ideation detection models, aiming to improve their generalizability and fairness.

Methods

Study Methodology

In this subsection, we elaborate on our study methodology. Figure 1 illustrates how the study methodology was developed. Our approach included the following steps:

- Step 1 (unsupervised topic modeling): We performed unsupervised topic modeling to discover the topics and risk factors mostly discussed on social media.
- Step 2 (domain knowledge extraction): We extracted relevant social risk factors from psychology literature to ground the content assessment of datasets in these factors.
- Step 3 (guided topic modeling): We compared topics extracted from social media and the literature. We shifted toward supervised guided topic modeling on social media using discovered topics from the literature to identify underrepresented and missing risk factors.
- Step 4 (synthetic and augmented dataset): We extended and annotated the existing synthetic datasets by incorporating additional risk factors that were missing from existing datasets. Moreover, we created an augmented dataset by combining synthetic data with real social media data to enhance its generalizability and topic coverage.
- Step 5 (dataset evaluation): To evaluate the utility of synthetic data, we analyzed both real and synthetic datasets by assessing the complexity, readability, and diversity of the text within each dataset. Additionally, we trained state-of-the-art classifiers using real-world, synthetic, and augmented datasets. Subsequently, the performance of these classifiers was evaluated by testing them on both real-world and synthetic test sets.

XSL•FO
RenderX

Figure 1. Overview of the multistep methodology used in this study to assess topic gaps in suicidal ideation detection from Reddit social media posts.

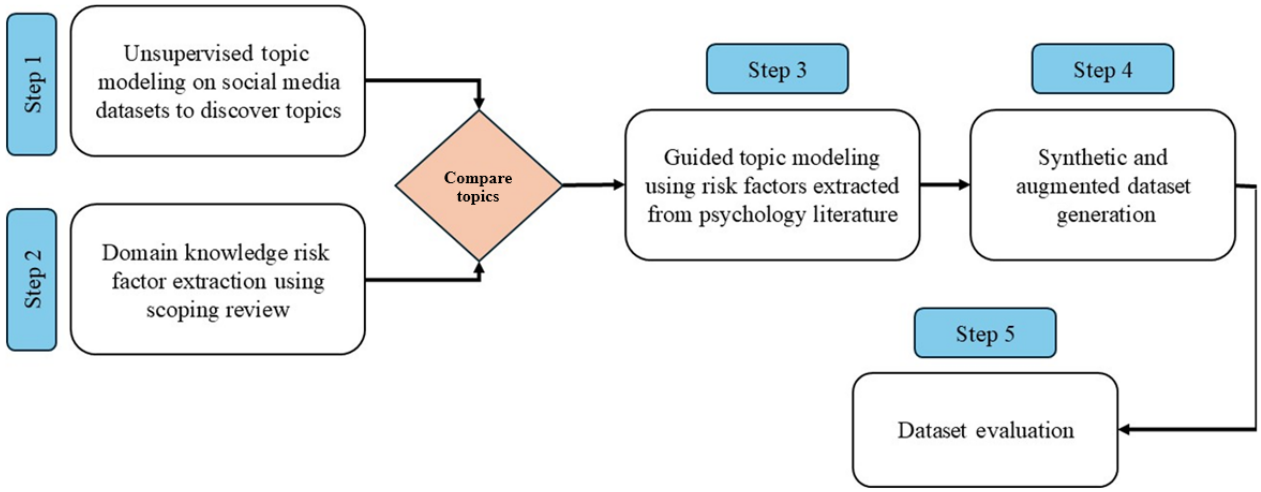

Unsupervised BERTopic

In this subsection, we present the methodology for conducting topic modeling using BERTopic [28], a state-of-the-art deep learning approach for discovering topics in text data. Our objective was to assess the quality, diversity, and coverage of datasets collected from social media in relation to suicidal ideation.

BERTopic leverages transformer-based language models as embedding models, combined with clustering methods for topic extraction [29]. BERTopic integrates transformer-based techniques with term frequency–inverse document frequency (TF-IDF) to create compact, interpretable clusters, preserving the most relevant terms in topic descriptions. This approach leverages deep learning and is mostly used with the sentence transformers embedding model, which supports document embedding extraction in more than 50 languages [30]. BERTopic's topic modeling procedure involves 3 main stages: document embeddings, document clustering, and topic representation.

Text Embedding

Text embedding refers to the vector representation of text within a multidimensional space where textual contents conveying similar meanings exhibit similar embeddings. In this project, we used the "*SentenceTransformers*" Python framework for state-of-the-art sentence and text embeddings [31]. Although there are many models for text embeddings, we used the sentence transformers "*all-MiniLM-L6-v2*" model as it has been shown to be one of the best-performing models within the BERTopic framework [32].

Topic Extraction

In this step, similar documents or sentences are grouped based on their content. It is a method of organizing large amounts of textual information into meaningful categories or clusters, providing a high-level overview of the information contained within. Before clustering the embeddings, a dimensionality reduction is implemented, as embeddings are often high in dimensionality. In this work, we used the UMAP (Uniform Manifold Approximation and Projection) algorithm to reduce the dimensionality of the embeddings because it can capture both the local and global structures of high-dimensional data in lower-dimensional space [33]. It has also been proven that for short text clustering, UMAP demonstrates superior results [34,35]. The hyperparameter space involved in UMAP is manually inspected, and based on the performance of the model and presented topics, the best parameters are selected. The number of neighbors, the number of components, and the minimum distance of each component were selected as 15, 5, and 0, respectively.

Following the dimensionality reduction of our input embeddings, we needed to cluster them into groups of similar embeddings to extract our topics. The method used in this paper is HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise), introduced by Campello et al [36]. This method is based on the density clustering method that finds clusters of different shapes and identifies outliers where possible. Similar to UMAP, the parameters of this model are manually inspected, and the proper parameters are chosen. There is no automated method for determining values for HDBSCAN. The parameters should be set manually based on domain knowledge and understanding of the dataset. Hence, we selected 10 and 5 as the minimum cluster size and sample number in each cluster, respectively.

Keyword Extraction for Topics

In this step, class-based term frequency–inverse document frequency (C-TF-IDF) scores are used to identify a set of keywords that represent the topic for a better interpretation of the topic's content. TF-IDF is a technique used to extract features from text documents, which is achieved by combining 2 components: term frequency and inverse document frequency. Term frequency represents the simple word count within a document, treating each word count as a feature, and it is calculated by dividing the number of times the term occurs in the document by the total number of terms in the document. Inverse document frequency gauges the informativeness of specific words by measuring their frequency within a document relative to their frequency across all other documents. C-TF-IDF is similar to TF-IDF but is adopted for multiple classes by joining all documents per class. Thus, each class is converted

XSL•FO
RenderX

to a single document instead of a set of documents. After calculating the C-TF-IDF scores for all words in each topic, words with the highest scores are chosen as keywords associated with that topic.

Social and Psychological Knowledge Extraction

We conducted a scoping review to offer a current and thorough synthesis of psychology studies, aiming to identify topics, themes, and risk factors related to the sensitive and complex issue of suicidal ideation. Scoping reviews serve as a method to assess the scope of literature on a particular subject, providing insights into the available research and offering a broad overview of its focus [37]. Additionally, they highlight the types of evidence that guide practice in the field and examine how the research has been conducted [38]. Our protocol was developed using the scoping review methodological framework proposed by Arksey and O'Malley [39] and further refined by Peters et al [40].

Research Question

The central research question guiding this review was as follows: "What are the most frequently reported risk factors associated with suicidal ideation in the psychology and mental health literature?" By answering this question, the review seeks to provide a clearer understanding of the key factors that contribute to the emergence of suicidal thoughts, ultimately informing future research and interventions aimed at preventing suicide and supporting individuals at risk.

Search Strategy

We scoured prominent academic databases, such as PubMed, PsycArticles, ScienceDirect, and Google Scholar, employing a systematic approach. Our search strategy involved using identical keywords across all databases: "suicidal ideation" and "suicide risk factors."

Eligibility Criteria

To be eligible for inclusion, studies were required to be peer-reviewed review or systematic review publications that investigated risk factors associated with suicidal ideation. Eligible studies had to be published between January 2014 and August 2024, address all forms of suicidal behavior including ideation and attempts, and focus on the general population.

Studies were excluded if they focused on specific groups that could introduce additional factors into the analysis, were published in a language other than English, or were not peer-reviewed publications, such as book chapters.

Study Selection

All identified publications were initially screened for relevance based on abstract and title. Subsequently, the full text of selected publications was assessed for eligibility. Furthermore, the reference lists of eligible papers were used to identify additional studies.

This exploration allowed us to identify a wide array of relevant review papers, forming the foundation of our research. Subsequently, we reported the most common topics among all the selected research articles. Based on our analysis of the literature, several social and psychological factors were consistently reported in relation to suicidal ideation in psychology. These topics have not been listed in a specific order of importance but represent the consistently reported themes in the literature reviewed.

Mental Health Disorders and Personality Traits

"Depression" is a well-documented and significant risk factor for suicide. The persistent feelings of sadness and emotional pain that characterize depression can lead individuals to contemplate or attempt suicide as a means of escape from their suffering. It is vital to recognize the signs of depression, offer support, and connect individuals to mental health professionals and resources for effective treatment and intervention [41-43].

"Anxiety" disorders are commonly associated with suicidal ideation. The chronic emotional distress and physical symptoms associated with severe anxiety can contribute to the development of suicidal thoughts [44].

"Posttraumatic stress disorder" (PTSD) is strongly associated with an increased risk of suicidal ideation. Individuals who have experienced traumatic events may struggle with the emotional aftermath, including intrusive memories, hyperarousal, and avoidance of reminders, which can lead to constant emotional distress and a sense of being overwhelmed. These constant feelings of traumatic memories can contribute to thoughts of suicide [45,46].

"Bipolar disorder," formerly called manic depression, is a mental health condition that causes extreme mood swings. These include emotional highs (also known as mania) and lows (also known as depression) [47]. Suicide attempts and completed suicide are significantly more common in patients with bipolar disorder when compared with the general population [47,48].

"Schizophrenia" is a severe mental disorder that disrupts both cognitive and social functioning, often resulting in the onset of additional health conditions [49]. Schizophrenia is strongly linked to an increased risk of suicidal ideation. The distress caused by its symptoms, feelings of isolation, and perceived loss of control over one's mind can contribute to a deep sense of hopelessness. Additionally, the stigma and social withdrawal often associated with schizophrenia may exacerbate these feelings, further increasing the risk of suicide [50,51].

"Borderline personality disorder" is a prevalent mental health condition linked to elevated suicide rates, significant functional impairment, frequent co-occurrence with other mental disorders, and extensive treatment needs. Long-term outcome studies of patients with borderline personality disorder have documented a high rate of suicide completion [52,53].

The expression and experience of "anger" have been reported as influential factors in suicidal ideation. Unresolved anger and intense emotional turmoil can drive individuals toward suicidal thoughts and actions. Anger, when left unmanaged, can escalate distress and lead to impulsive decisions with dire consequences. Recognizing and managing anger are crucial facets of suicide detection [41,54].

"Perfectionism," marked by excessively high standards and self-criticism, has been identified as a psychological factor related to suicidal ideation. The relentless pursuit of unattainable

XSL•FO
RenderX

standards can lead to feelings of inadequacy and despair, increasing the risk of suicidal ideation. Recognizing the need for balance and self-compassion is pivotal in addressing this risk factor [41].

Feelings of "hopelessness," characterized by a pervasive sense of despair and an inability to envision a better future, can be a potent predictor of suicidal behavior. Those burdened by overwhelming hopelessness may see suicide as the only means of escape from their emotional suffering [41,42,55].

Substance Abuse

Alcohol and drug misuse is a significant risk factor for suicidal ideation and behavior [56]. The connection between substance abuse and suicide is complex, as substances like drugs and alcohol can impair judgment and exacerbate underlying emotional distress. Individuals who struggle with addiction may turn to substances as a means of coping with psychological pain, and when combined with impaired decision-making, this can increase the likelihood of suicidal thoughts and actions [57-59].

Sociodemographic Status

Prolonged "unemployment" can erode self-esteem, create financial difficulties, and contribute to feelings of hopelessness, increasing the risk of suicide. Government support and job assistance programs can help mitigate this risk [55]. "Financial hardships" can trigger intense stress, making individuals vulnerable to suicide. Economic support and resources are instrumental in addressing this risk factor [44,60].

"Education pressure," including exams and expectations, can lead to emotional turmoil and an increased risk of suicidal ideation, particularly among students. Educational institutions must provide resources for coping with academic stress [61,62].

"Sexual minority" individuals, such as those who identify as lesbian, gay, bisexual, transgender, queer (LGBTQ+), often face unique stressors related to their sexual orientation or gender identity [63]. Discrimination, prejudice, and stigma can lead to feelings of isolation, rejection, and psychological distress. Research consistently shows that sexual minority individuals are at a higher risk for suicidal ideation and attempts compared to their heterosexual counterparts [64-67].

Abuse

"Bullying," including physical, verbal, or cyber bullying, has consistently emerged as a significant topic related to suicidal ideation. The experience of bullying can lead to social isolation, low self-esteem, and emotional distress, contributing to the development of suicidal thoughts [44,68-70].

"Sexual abuse" is a recognized risk factor for suicide, which includes any sexual activity that occurs without consent, also referred to as sexual assault or sexual violence [71]. Dissociation is a common response to sexual abuse, and higher levels of dissociation have been associated with self-harm, suicidal thoughts, and suicide attempts [72,73].

Family-Related Issues

"Family-related stressors," such as conflict, dysfunctional dynamics, and poor communication, can significantly impact an individual's emotional well-being and contribute to suicidal thoughts. Strengthening family relationships and providing support to those affected are essential in mitigating this risk factor [43,55].

Difficulties in "relationships," including conflicts, breakups, and marital dissatisfaction, have been reported as significant topics in relation to suicidal ideation. Relationship problems can contribute to emotional distress and feelings of hopelessness, leading to thoughts of suicide [55].

The death of family members or friends has been reported as a risk factor associated with suicidal ideation. Grief, feelings of loneliness, and a sense of being unable to cope with the loss can increase the risk of suicidal thoughts [74].

Racism

Studies have consistently highlighted the significant impact of racial discrimination on suicidal ideation. Experiencing racism and racial prejudice can increase the risk of suicidal thoughts [75-77].

Immigration

The process of immigration, with its cultural adjustments, isolation, and uncertainty, can intensify stress and emotional distress, increasing the risk of suicide among immigrants. Providing support and resources tailored to the immigrant experience is essential [78,79].

Dementia

Dementia, particularly in its advanced stages, can lead to significant cognitive and emotional challenges. Individuals with dementia may experience confusion, memory loss, and personality changes, which can be distressing for both them and their caregivers. The experience of losing one's cognitive abilities and identity can contribute to feelings of hopelessness and despair, leading to thoughts of suicide [80-82].

Chronic Physical Problems

Living with chronic physical health conditions can be emotionally taxing, and individuals facing such challenges may be more susceptible to suicidal ideation. Chronic pain, disability, and limitations in physical functioning can erode one's quality of life and lead to a sense of hopelessness [83]. Coping with the constant demands of managing a chronic condition can also contribute to emotional distress [84,85].

Guided BERTopic

In this research, we began by applying unsupervised BERTopic to the social media datasets to identify key topics discussed within them. Next, we conducted a scoping review to examine established suicidal risk factors. We then compared the topics from the social media analysis with risk factors extracted from psychology literature. During this process, we observed that the effectiveness of topic modeling techniques in capturing domain-specific terms may be limited. To address this, we used guided topic modeling to ensure that certain specialized terms are detected and appropriately represented in the topics.

Guided topic modeling (supervised topic modeling) is an extension of traditional topic modeling that incorporates external information or guidance to influence the topic discovery process.

XSL•FO
RenderX

In this project, we employed guided BERTopic [86], which introduces external information in the form of seed words to guide the algorithm in discovering topics that align with the specified guidance. This guidance helps to improve the relevance and coherence of the identified topics. Here, we used discovered suicidal risk factors from psychology literature in the Social and Psychological Knowledge Extraction section as categories to guide the topic modeling process. Each category comes with sets of seed words that are suggested using "Meriam-Webster Dictionary" [87] as the most related keywords for each category. This deliberate approach allowed us to align the resulting topics with known psychological factors that contribute to suicide risk, creating a more focused and domain-informed representation.

Guided BERTopic involves 2 primary steps. First, the BERTopic algorithm generates embeddings for each seeded topic by concatenating them and passing them through the document embedder. As shown in Figure 2, these embeddings are compared to the existing document embeddings using cosine similarity and are then assigned a label. If a document is most similar to a seeded topic, it receives that topic's label; otherwise, if it is most similar to the average document embedding, it will be categorized as an outlier (–1 label). UMAP then applies these labels in a semisupervised manner, guiding the dimensionality reduction process to emphasize the distinctions between seeded topics and potentially identify outliers, thereby steering the topic creation more effectively toward the seeded topics. Second, all words in the seed topic list are assigned to a multiplier with a value greater than 1. These multipliers are applied to increase the inverse document frequency values of the words across all topics by boosting the likelihood of a seeded topic word appearing in a topic. After having generated our topics using C-TF-IDF, we employed the *KeyBERTInspired* extraction technique that leverages BERT embeddings and simple cosine similarity to find keywords and key phrases that are most similar to a document. Then, we selected keywords with the highest C-TF-IDF score as representative seed topics for the guided topic modeling.

Figure 2. Architecture of guided BERTopic using seed terms derived from psychology literature. Illustration of how seed words based on established suicide risk factors from the literature were incorporated into the guided BERTopic algorithm to steer topic discovery in Reddit social media posts for suicidal ideation detection. IDF: inverse document frequency.

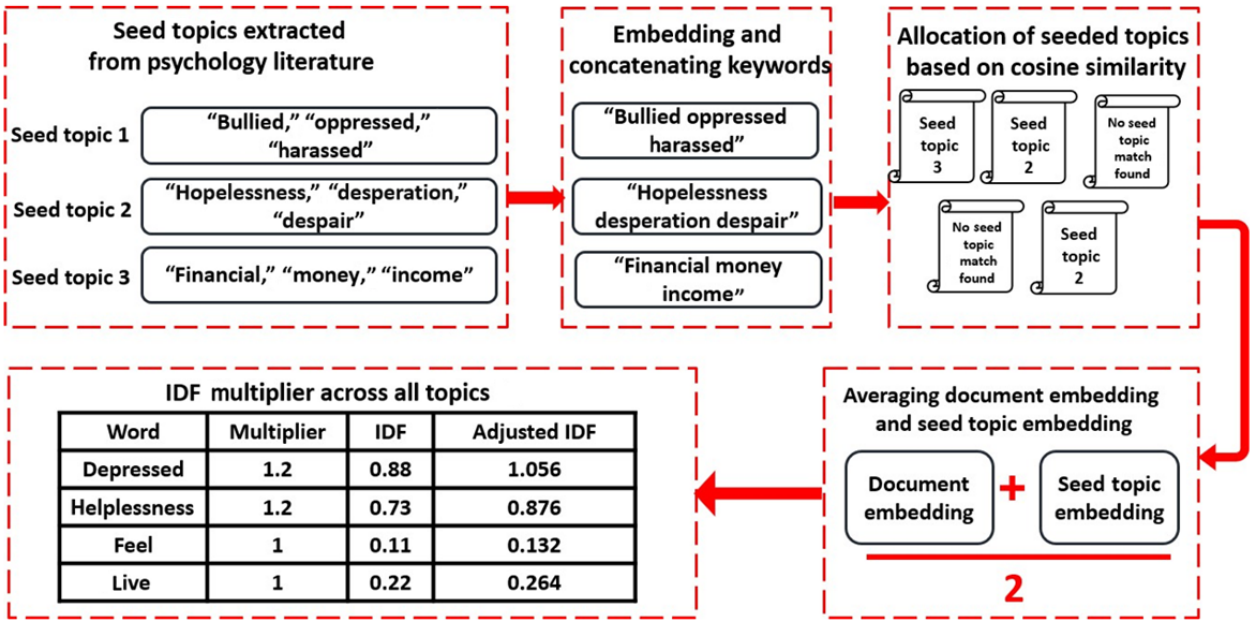

Word	Multiplier	IDF	Adjusted IDF
Depressed	1.2	0.88	1.056
Helplessness	1.2	0.73	0.876
Feel	1	0.11	0.132
Live	1	0.22	0.264

Synthetic Data Generation

Synthetic data generation provides a practical solution for addressing data availability limitations by creating artificially generated data that closely resemble real-world data. Ghanadian et al [26] employed 3 generative large language models, namely GPT-3.5 Turbo, Flan-T5, and Llama2, to generate a synthetic dataset related to suicide. Their findings indicated that GPT-3.5 Turbo outperformed other generative models in creating synthetic data, that is, a classifier trained on a dataset generated by GPT-3.5 Turbo achieved higher *F*-scores when tested on the real-world and synthetic test sets. In our research, we used the dataset presented in the study by Ghanadian et al [26], along with domain knowledge–based risk factors we extracted, to generate a new dataset using GPT-3.5 Turbo. Moreover, we augmented 30% of the real dataset with a new synthetic dataset to report the practicality of the generated data for suicidal ideation detection.

Dataset Analysis

For textual datasets, assessing the diversity of semantics and the sentence structure of textual content is crucial, even more so when dealing with synthetic datasets. To measure these qualities, we used 3 sets of metrics: complexity, readability, and entropy. Complexity refers to the intricacy and sophistication of the language used in a text, encompassing factors such as sentence structure, vocabulary richness, and syntactic intricacies. Readability pertains to the ease with which a text can be comprehended by its intended audience, considering elements such as sentence length, word difficulty, and overall coherence. Understanding complexity and readability in synthetic datasets aids in ensuring that the generated text

XSL•FO
RenderX

aligns with linguistic patterns observed in real-world data. Moreover, these parameters facilitate an assessment of the synthetic dataset, specifically regarding the incorporation of suitable language complexities. This evaluation allows us to examine if synthetic data replicate language patterns akin to those found in genuine, human-generated content. Furthermore, entropy is a metric that measures the unpredictability of a text. It assesses the information content or disorder present in the dataset. In simpler terms, entropy in text datasets gauges how diverse or varied the words or characters are within the dataset. High entropy indicates a higher degree of unpredictability, suggesting a wider range of different words or characters used in the text. Conversely, low entropy implies a more predictable or ordered text with fewer variations in the words or characters used.

To measure the readability of text, we used the Flesch reading ease test [88], which quantifies readability based on sentence length and the number of syllables per word. As a benchmark, a high score indicates that the text is easily understood by an average 11-year-old, while a low score indicates that the text is best understood by university graduates. Table 1 presents the definition criteria for the Flesch reading ease test score.

Table 1. Flesch reading ease test score definition criteria.

Score	Difficulty
90-100	Very easy
80-89	Easy
70-79	Fairly easy
60-69	Standard
50-59	Fairly difficult
30-49	Difficult
0-29	Very confusing

Additionally, we used the type-token ratio [89] to assess the complexity of a text using lexical diversity measures. The basic idea behind that measure is that if the text is more complex, the author uses a more varied vocabulary, so there is a larger number of unique words [90].

Another selected metric was Shannon entropy [91], which is calculated based on the frequency of occurrence of different characters, words, or other linguistic units within the text. In word-based analysis, higher entropy suggests a wider range of vocabulary, showing greater linguistic diversity [92]. Calculation of Shannon entropy involves summing the probabilities of each word occurrence in the text, weighted by the logarithm of the inverse of these probabilities. The Shannon entropy H for a set of words with probabilities $p_1, p_2, ..., p_n$ is calculated as follows:

In this equation, H represents the calculated entropy value for the given set of words, and p_i represents the probability of the ith word occurring in the text. The sum extends over all unique words. By computing Shannon entropy in text analysis, one can gain insights into the richness, diversity, and complexity of the language used within the text dataset.

Datasets

In this section, we review the specifications of the datasets assessed in this work.

University of Maryland Reddit Suicidality Dataset

We used the University of Maryland Reddit Suicidality Dataset (UMD) [93,94] collected from Reddit. Reddit is an online website and forum for anonymous discussion on a wide variety of topics. The UMD is a collection of Reddit posts and comments created by individuals who expressed suicidal thoughts or behaviors. The dataset contains over 100,000 posts and comments collected from various subreddits, including those related to mental health and suicide prevention, such as *Depression* [95] and *SucideWatch* [96] subreddits. The data were collected over a period of several years and include the content of posts and comments, as well as the location and timing of the posts. This dataset contains annotations at the user level, using a 4-point scale to indicate the severity of suicide risk: (1) *no risk*, (2) *low risk*, (3) *moderate risk*, and (4) *high risk*. The goal of this annotation is to assess the risk level of individuals through an examination of their online activities. This task necessitates minimal data, with users generally contributing only a limited number of posts on SuicideWatch. Among the 993 labeled users, 496 made at least one post on the SuicideWatch subreddit. The remaining 497 users served as control subjects. Since the provided labels were user-level labels, we aggregated all the posts of each user into a single data point, through the concatenation of all the posts made by a particular user. Using a binary classification approach similar to that by Ghanadian et al [26], we applied binarization to the UMD. In accordance with the 4 class definitions, the categories labeled as "*no risk*" and "*low risk*" were categorized as nonsuicidal, while those labeled as "*moderate risk*" and "*high risk*" were considered suicidal. A total of 490 anonymous posts with binary labels were employed in this paper.

The UMD has been repeatedly used by researchers to develop and test NLP algorithms and machine learning models in order to identify and analyze patterns in online communication related to suicide risk [97]. Ji et al [98] proposed a method for improving text representation by incorporating sentiment scores based on lexicon analysis and latent topics. Additionally, they introduced the use of relation networks for the detection of suicidal ideation and mental disorders, leveraging relevant risk indicators. In another paper, Ji et al [99] used 2 pretrained

XSL•FO
RenderX

masked language models, MentalBERT and MentalRoBERTa, specifically designed to support machine learning in the mental health care research field. The authors assessed these domain-specific models along with various pretrained language models on multiple mental disorder detection benchmarks. The results showed that using language representations pretrained in the mental health domain can enhance the performance of mental health detection tasks, highlighting the potential benefits of these models for the mental health care research community.

Knowledge Aware Assessment Dataset

Gaur et al [100] developed an annotated gold standard dataset of 500 Reddit users out of 2181 potentially suicidal users, using their content from mental health–related subreddits within the time frame of 2005 to 2016. The dataset consists of 5 different categories of suicidality, including suicidal ideation, suicidal behavior, actual attempt, suicide indicator, and supportive.

Suicidal ideation refers to thoughts of suicide, which may involve concerns related to suicide risk factors, such as job loss or the end of a significant relationship. Suicidal behavior is defined as actions that carry a higher risk, such as self-harm (either current or historical), active planning to commit suicide, or a history of institutionalization for mental health reasons. An actual attempt encompasses any deliberate action that could potentially lead to intentional death. This includes, but is not limited to, instances where an individual sought help, reconsidered their decision, or publicly expressed thoughts of suicide. The suicide indicator category serves as a classification method to distinguish individuals who use at-risk language from those who are actively experiencing general or acute symptoms. Often, users converse in supportive conversations and share their personal histories while using language from the clinical lexicon. The supportive category pertains to individuals engaging in discussions without expressing any history of being at risk, either in the past or at present.

2021 Reddit Dataset

In addition to the existing datasets, we collected a new dataset of suicidal social media posts from the Reddit platform using the Reddit application programming interface [101]. Specifically, we focused on the “SuicideWatch” subreddit to gather posts related to suicide and analyzed the topics discussed within this subreddit.

The initial data collection was conducted on September 11, 2021, with the goal of gathering 2500 posts published between May 1, 2021, and September 1, 2021. Subsequently, we conducted extensive text preprocessing, which included removing links, eliminating duplicates, handling special characters, removing stop words, filtering out irrelevant and noninformative posts, performing lemmatization, and conducting spellchecking. After these preprocessing steps, our dataset consisted of 2052 unlabeled posts from the “SuicideWatch” subreddit, which we used for the purpose of topic modeling in this study.

Synthetic Dataset

Ghanadian et al [26] generated a synthetic dataset to enhance the performance of state-of-the-art suicide detection approaches by augmenting real-world datasets. They harnessed the capabilities of 3 generative large language models, namely, GPT-3.5 Turbo, Flan-T5, and Llama 2, to generate synthetic data for the detection of suicidal ideation.

In this paper, we generated a dataset comprising 748 suicide-related posts. As discussed in the Social and Psychological Knowledge Extraction section, our updated suicidal risk factors and topics, derived from an exhaustive literature search, have led to an extension of the synthetic dataset initially created by Ghanadian et al [26]. We found 5 more risk factors associated with suicidal ideation in the literature. We prompted GPT-3.5 Turbo to generate both suicidal and nonsuicidal instances to add to the existing dataset. This dataset is binary, categorized into suicidal and nonsuicidal classes, with annotations independently provided by 2 expert human annotators. A notable 90% of the labels, initially generated by GPT-3.5, were agreed upon by the human annotators. However, for the remaining 10% of the data, the labels were altered based on the decision of the annotators. In cases where both annotators agreed on a label, that label was retained. Conversely, when disagreements arose, the annotators engaged in discussions to ultimately reach a consensus on the appropriate label.

Augmented Dataset

Data augmentation involves enriching a dataset by introducing variations to its existing instances or generating entirely new instances. This process is designed to enhance the diversity and quality of the dataset, which, in turn, can lead to improved model performance and generalization. Ghanadian et al [26] reported that when synthetic data were augmented with 30% of the UMD training set, the fine-tuned model outperformed the model trained with the full UMD. This ratio represents the minimum amount of augmentation required for the ALBERT model to surpass the performance results of the model trained solely on the UMD. Hence, we augmented 30% of the synthetic dataset with the UMD to achieve a balance between diversity and quality in the training dataset.

The augmented dataset was used to fine-tune the pretrained state-of-the-art models and then was evaluated on 2 separate testing sets. The first testing set was 20% of the UMD before any training. The second testing set was generated and annotated by Ghanadian et al [26]. They used several synthetic datasets to create the testing set for their evaluation.

Ethical Considerations

This study involved the analysis of suicidal posts from social media, which is a sensitive and ethically complex task. We obtained ethics approval for the secondary use of data from the Research Ethics Board at the University of Ottawa (approval number: H-02-23-8967). This approval confirms that our research complies with the ethical guidelines for working with human-derived data and protects the identity of individuals whose posts are analyzed. The data used in this study are from the UMD, which was accessed with authorization from its creators. This dataset was approved by the University of Maryland’s Institutional Review Board. The dataset consists of publicly available social media posts collected from Reddit, where users are anonymous by default. To further protect user

XSL•FO
RenderX

privacy, Reddit usernames in the dataset have been replaced with numeric identifiers [102]. Since this study involves the secondary analysis of publicly available and anonymized social media data, informed consent from individual users was not required.

Results

Overview

We present the results according to this project's main research questions. First, we present the discovered topics in 3 real datasets collected from social media, including the UMD, Knowledge Aware Assessment dataset, and 2021 Reddit dataset uncovered by unsupervised BERTopic. Second, we present the results of a scoping review and the reference table for each risk factor. Figure 3 presents the details of our scoping review process. Third, we present the results of topic discovery using guided BERTopic and a comparison between unsupervised and guided BERTopic. Finally, we provide an insightful breakdown, demonstrating the distribution of each topic within our synthetic dataset and the augmented dataset.

Figure 3. Scoping review workflow for extracting suicide risk factors from psychology literature.

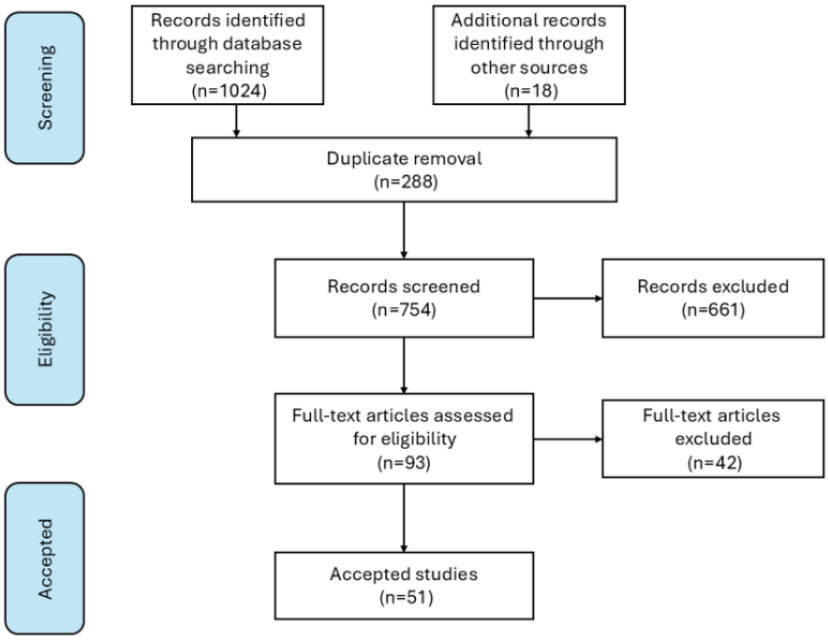

To delve deeper into our investigation, we report the performance of the fine-tuned model on these synthetic and augmented datasets, offering a comprehensive comparison between the topic-diverse synthetic dataset and real dataset.

Unsupervised Topic Modeling

In this section, we assess the topic diversity in all datasets using basic BERTopic. Table 2 reports the most occurring topics and keywords in the selected datasets using basic BERTopic. Some topics are common across all 3 datasets, while others are not discussed. For example, the topic of *education pressure* is discussed in the UMD and 2021 SuicideWatch datasets but is not found in the Knowledge Aware Assessment dataset. The output of topic modeling consisted of clusters of keywords, each associated with a probability of their relevance to the cluster. The title of each topic was chosen based on the highest probability keywords within the cluster.

XSL•FO
RenderX

Table 2. Distribution of suicide-related psychological and social risk factors identified through unsupervised BERTopic modeling in Reddit datasets.

Topics and risk factors extracted from psychology	Dataset		
	UMD[a] (SW[b] subreddit) (n=490), n	Knowledge Aware Assessment dataset (n=500), n	2021 SW dataset (n=2050), n
Mental health disorders			
Depression	98	78	756
Anxiety	45	73	244
PTSD[c]	0	0	4
Sociodemographic status			
Unemployment	18	5	0
Financial crisis	0	21	0
Education pressure	15	0	28
Abuse			
Being bullied	5	0	6
Family domain			
Family issues	50	14	34
Relationship problems	17	21	30
Personality and psychological traits			
Hopelessness	9	0	8
Racism	0	0	17
Substance abuse	4	0	0
Chronic physical problems	0	1	3

[a]UMD: University of Maryland Reddit Suicidality Dataset.

[b]SW: SuicideWatch.

[c]PTSD: posttraumatic stress disorder.

Scoping Review Knowledge Extraction

In our scoping review of psychology literature, we initially identified 1042 articles related to suicide. After applying the inclusion and exclusion criteria outlined in Figure 3, a total of 51 studies were retained for analysis. Table 3 provides a list of references we investigated to extract underlying risk factors of suicidal ideation. This table organizes all the studies reporting the identified risk factors. Some of these risk factors fall under broader categories, allowing for a more structured understanding. In the Social and Psychological Knowledge Extraction section, we have provided a description of each extracted risk factor and how it plays a role in suicidal ideation.

XSL•FO
RenderX

Table 3. List of risk factors for suicidal ideation extracted through a scoping review of 51 psychology and mental health studies.

Category	Studies
Mental health disorders	
Depression	[12,41-43,103-111]
Anxiety	[44,106-108,110,112-114]
Bipolar	[12,17,103,109-111]
Schizophrenia	[12,103,105,107,109,110,113]
Borderline	[12,106,110,115,116]
PTSD[a]	[45,46,107]
Sociodemographic status	
Unemployment	[12,55,107-109]
Education pressure	[61,62,109,111]
Financial crisis	[44,60,107,108,111,114]
Sexual minority stigma	[63-67,107,111]
Abuse	
Being bullied	[44,68-70]
Sexual abuse	[107,109,111,117,118]
Family domain	
Death of loved ones	[74,108,109]
Family conflicts	[43,55,110]
Relationship problems	[55,109,110]
Personality and psychological traits	
Hopelessness	[12,41,42,55,108,119]
Anger	[41,54,107,108]
Perfectionism	[41,107,109]
Chronic physical pain	[83-85]
Dementia	[80-82]
Racism	[75-77]
Immigration	[12,78,79,109,111]
Substance abuse	[56-59,110,111,114]

[a]PTSD: posttraumatic stress disorder.

Guided Topic Modeling

In unsupervised topic modeling, we investigated the underlying topics within each social media dataset. A comparison between the extracted topics in the Unsupervised Topic Modeling and Scoping Review Knowledge Extraction sections revealed that many topics reported in psychology literature were not discovered in social media using unsupervised BERTopic. The absence of these topics was evident only when compared to the risk factors in psychology literature, and without this comparison, these important gaps in the discussion might have been overlooked.

To address this, we employed guided topic modeling, which focuses on specific risk factors of interest. This approach ensures that important but less frequently mentioned topics are identified and included in the analysis. Hence, the topics from psychology were used as seed topics for the discovery of the topics in these datasets. Table 4 presents the list of seed words used in guided topic modeling to extract topics from the UMD. We conducted seed topic extractions separately for each of the 3 social media datasets. It is important to note that this list of words does not need to be exhaustive and only serves as a hint about the topic for the topic modeling algorithm.

XSL•FO
RenderX

Table 4. List of manually curated seed words representing 23 risk factor categories used to guide topic modeling of suicide-related Reddit posts, grounded in terms extracted from psychology studies and verified using the Merriam-Webster Dictionary.

Category	Seed words
Depression	Depressed, sadness, and mentally ill
Anxiety	Anxious, phobia, and stress
Schizophrenia	Delusion
Bipolar	Manic and temperamental
Borderline	Petulant and impulsive
PTSD[a]	War, memory, and accident
Sexual abuse	Assault and rape
Being bullied	Bullying, oppress, and alone
Family issues	Family, parents, mom, and dad
Relationship problems	Wife, husband, partner, girlfriend, and boyfriend
Death of loved ones	Loss, grief, and mourn
Anger	Outrage and annoyed
Perfectionism	Perfection and expectations
Hopelessness	Despair
Unemployment	Job, work, and poverty
Financial crisis	Money and income
Education pressure	College, school, and overwhelmed
Sexual minority stigma	LGBTQ+[b] and identity
Racism	Discrimination, justice, bias, and hate
Substance abuse	Alcohol, drug, and opioid
Chronic physical pain	Constant, hurt, and escape
Immigration	Moving, loneliness, and culture
Dementia	Alzheimer

[a]PTSD: posttraumatic stress disorder.

[b]LGBTQ+: lesbian, gay, bisexual, transgender, queer.

Real Dataset Topic Verification

Here, we aim to understand the distribution of topics in real data using guided topic modeling. Three real datasets collected from social media were investigated for suicide-related risk factors and topics. Table 5 presents the suicide topics in each real dataset along with the number of posts for each topic. Since we were interested in the distribution of these topics in relation to suicidal thoughts in labeled datasets (UMD and Knowledge Aware suicidality), we only considered the suicidal classes for topic modeling and evaluation. The results in Table 5 show that while social media provides a vast and dynamic platform for individuals to express their thoughts and experiences, it may not always comprehensively reflect the nuanced and scientifically established suicide-related topics discussed in academic psychology literature. The conversational nature of social media often includes a wide range of personal narratives, opinions, and language that may or may not align with the structured and research-driven topics found in psychology literature.

XSL•FO
RenderX

Table 5. Distribution of suicidal risk factors extracted using guided BERTopic in real datasets.

Topics and risk factors extracted from psychology	Dataset		
	UMD[a] (SW[b] subreddit) (n=490), n	Knowledge Aware Assessment dataset (n=500), n	2021 SW dataset (n=2050), n
Mental health disorders			
Depression	115	79	806
Anxiety	52	80	297
Bipolar	5	2	2
Schizophrenia	0	0	0
Borderline	0	2	0
PTSD[c]	5	0	4
Sociodemographic status			
Unemployment	30	25	0
Financial crisis	0	30	0
Education pressure	30	0	33
Sexual minority stigma	0	0	0
Abuse			
Being bullied	8	0	13
Sexual abuse	2	0	7
Family domain			
Death of loved ones	0	0	0
Family issues	55	14	55
Relationship problems	25	26	38
Personality and psychological traits			
Anger	0	0	0
Perfectionism	0	0	0
Hopelessness	9	0	17
Racism	0	0	36
Substance abuse	7	0	0
Immigration	0	0	0
Chronic physical problems	0	2	3
Dementia	0	0	0

[a]UMD: University of Maryland Reddit Suicidality Dataset.

[b]SW: SuicideWatch.

[c]PTSD: posttraumatic stress disorder.

Although social media can offer valuable insights into real-world expressions of mental health concerns, researchers need to carefully interpret and validate social media data to ensure their reliability and relevance to the broader body of psychology literature on suicide. Underrepresentation of certain topics in social media data, specifically stigmatized topics, such as conversations around sexual minorities and racism, can lead to models that underperform in cases where users break the stigma and talk about these issues.

Topic Verification on Synthetic and Augmented Datasets

We employed guided topic modeling to verify if this method reveals the suicide topics and risk factors in the synthetic and augmented datasets. Note that these topics were used to generate the synthetic data. However, the number of topics discovered by the guided topic modeling did not sum up to the total number of posts, which was 908 for the synthetic dataset and 1055 for the augmented dataset. This discrepancy arose because some posts were associated with multiple topics and represented more than one risk factor. Table 6 displays the suicide topics within each synthetic and augmented dataset, along with the respective number of posts for each topic. We observed that topic modeling

XSL•FO
RenderX

showed a relatively equal distribution of topics in the synthetic dataset. Table 6 illustrates how effectively guided BERTopic extracted the topics within the documents. Furthermore, these results demonstrate the significance of the generation of topic-diverse synthetic and augmented datasets for various research and applications within the domain of suicide-related studies.

Table 6. Topic coverage in synthetic and augmented suicidal ideation datasets using guided BERTopic.

Topics and risk factors extracted from psychology	Dataset	
	Synthetic dataset (n=908), n	Augmented dataset (n=1055), n
Mental health disorders		
Depression	62	173
Anxiety	46	68
Bipolar	19	20
Schizophrenia	23	23
Borderline	23	23
PTSD[a]	22	21
Sociodemographic status		
Unemployment	20	35
Financial crisis	21	39
Education pressure	21	20
Sexual minority stigma	41	42
Abuse		
Being bullied	39	39
Sexual abuse	18	20
Family domain		
Death of loved ones	40	41
Family issues	19	20
Relationship problems	40	43
Personality and psychological traits		
Anger	39	38
Perfectionism	48	48
Hopelessness	36	44
Racism	39	39
Immigration	20	18
Substance abuse	40	46
Chronic physical problems	56	61
Dementia	39	48

[a]PTSD: posttraumatic stress disorder.

Dataset Analysis and Evaluation

In this subsection, we first analyze the quality of each dataset based on 3 indicators introduced in the Dataset Analysis section. The comparison between synthetic and nonsynthetic datasets (UMD, Knowledge Aware, and 2021 Reddit) presented in Table 7 reveals distinct patterns in terms of complexity, readability, and Shannon entropy. The synthetic dataset and augmented dataset exhibited relatively high complexity values of 72.77 and 70.27, respectively, which were comparable to that of the 2021 Reddit dataset (72.87) but surpassed those of the UMD and Knowledge Aware dataset, indicating that synthetic data are linguistically or structurally more complex. In terms of content diversity, measured by Shannon entropy, all datasets showed similar means, but the synthetic dataset was significantly more uniform, with very low deviation (σ=0.05), compared to the nonsynthetic datasets. Based on the results presented in Table 7, synthetic datasets tend to be more complex and are more controlled and consistent, whereas nonsynthetic datasets provide a broader range of readability and content diversity.

XSL•FO
RenderX

Table 7. Linguistic characteristics of real, synthetic, and augmented datasets for suicidal ideation detection.

Dataset	Complexity (type-token ratio)		Readability (Flesch score)		Shannon entropy	
	μ	σ	μ	σ	μ	σ
UMD[a]	61.18	16.15	82.49	10.87	4.31	0.17
Knowledge Aware dataset	49.04	19.53	82.51	15.27	4.17	0.75
2021 Reddit dataset	72.87	20.13	75.76	48.02	4.27	0.30
Synthetic dataset	72.69	18.40	75.80	10.34	4.27	0.05
Augmented dataset	70.26	17.18	77.38	10.80	4.28	0.05

[a]UMD: University of Maryland Reddit Suicidality Dataset.

Moreover, Table 8 indicates that the readability of the synthetic and augmented datasets was more challenging than that of the real datasets. Specifically, 36% of the synthetic dataset fell into the "easy" and "very easy" categories, while 67% of the UMD fell into these categories. In contrast, 33% of the UMD and 64% of the synthetic dataset fell into the "fairly easy," "standard," "fairly difficult," and "difficult" categories. Furthermore, we conducted a statistical *t* test with a significance level of .05 on the readability, complexity, and entropy metrics, comparing the synthetic dataset with each of the real datasets. Our findings indicated a significant difference in Shannon entropy between the synthetic and real datasets. However, no significant differences were observed in terms of readability and complexity between the real and synthetic datasets.

Table 8. Distribution of Flesch reading ease scores across all datasets.

Flesch reading ease interval	UMD[a], n	Knowledge Aware dataset, n	2021 Reddit dataset, n	Synthetic dataset, n	Augmented dataset, n
Very confusing	2	0	10	0	0
Difficult	1	3	30	0	5
Fairly difficult	5	5	50	50	50
Standard	30	25	200	130	140
Fairly easy	100	140	400	270	300
Easy	200	240	600	170	270
Very easy	100	50	350	80	120

[a]UMD: University of Maryland Reddit Suicidality Dataset.

Additionally, we created a binary classifier for detecting suicidal ideation using different datasets. For that, we fine-tuned ALBERT [120], a transformer-based model optimized for performance and speed, using the UMD and our synthetic and augmented datasets, to evaluate the effect of topics and risk factor inclusion on the performance of a pretrained model. To fine-tune these models, we used the Huggingface library [121]. The Huggingface library is an open-source library and data science platform that provides tools to build, train, and deploy machine learning models. We compared our classification results with a baseline ALBERT model fine-tuned on the UMD by Ghanadian et al [10]. We used the Trainer [122] class from Huggingface transformers for feature-complete training in PyTorch. The hyperparameters were selected based on the default values commonly used in similar studies. The final hyperparameters used in our experiments were as follows: learning rate, $2e^{-5}$; batch size, 4; dropout rate, 0.1; and maximum sequence length, 512.

Table 9 presents the performance results of fine-tuning the ALBERT model using the synthetic and augmented training datasets. The model fine-tuned on the augmented dataset outperformed the other 2 models, achieving F_1-scores of 0.91 and 0.90 on the UMD and synthetic testing subsets, respectively. Achieving a high F_1-score is important when evaluating classifiers, especially with imbalanced datasets. A high F_1-score indicates that the model is both precise and effective at identifying at-risk cases. This is critical in sensitive domains like mental health, where failing to detect or incorrectly identifying individuals at risk can have serious real-world consequences.

XSL•FO
RenderX

Table 9. Performance metrics of suicidal ideation classifiers trained on real, synthetic, and augmented data.

Test set metrics	Nonsynthetic	Synthetic	
	UMD[a] dataset	ChatGPT dataset	Augmented dataset
UMD test subset			
Accuracy	0.87	0.72	0.88
F_1-score	0.87	0.81	0.91
Synthetic test subset			
Accuracy	0.67	0.87	0.90
F_1-score	0.70	0.86	0.90

[a]UMD: University of Maryland Reddit Suicidality Dataset.

Discussion

This study investigated whether social media data adequately capture empirically established risk factors associated with suicidal ideation and whether synthetic data augmentation can improve the representation of underrepresented risk factors in suicide detection models. Our findings reveal that real-world social media datasets lack coverage of several critical risk factors, particularly those related to marginalized communities, financial crises, racism, and chronic health conditions. By incorporating insights from psychology literature and leveraging guided topic modeling, we identified these gaps and generated synthetic data to enhance topic diversity. Our results demonstrate that augmenting real-world datasets with topic-diverse synthetic data improves the performance of machine learning models in detecting suicidal ideation, as evidenced by an increase in the F_1-score from 0.87 to 0.91 on the UMD test subset and from 0.70 to 0.90 on the synthetic test subset. These findings support our hypothesis that social media data alone do not fully represent the nuances of suicide risk factors and that synthetic data augmentation can lead to more accurate and fair suicidal ideation detection models.

Traditional data collection methods, specifically self-reports, are subject to various kinds of biases, including social desirability, recall, and self-perception biases. While the use of naturalistic data collected from social media mitigates some of the self-report biases and incorporates contextual information, it, unfortunately, introduces other types of biases, mainly selection bias, platform-induced behavior, and public image curation bias. Many demographics are underrepresented on social media, and even if present, they might shy away from expressing their real self. On the other hand, traditional methods are more controlled, and as we have shown in our literature review, larger numbers of topics and demographics have been covered under those studies. For that reason, we compared the social media datasets to risk factors collected in traditional studies to identify underrepresented risk factors.

This study assessed real-world datasets from social media for their coverage of suicidal ideation topics and risk factors. Additionally, it reviewed psychology literature to identify expert-validated risk factors, suicidal language, and terminology. Integrating this knowledge into NLP models can enhance their contextual awareness in identifying and understanding suicidal ideation in social media posts. Leveraging insights from psychology also supports the development of ethically responsible artificial intelligence systems for addressing sensitive mental health issues.

Unsupervised topic modeling and a scoping review of psychology literature provided 2 sets of suicide topics and risk factors. Table 2 summarizes 13 suicide topics identified across 3 social media datasets, while Table 3 lists 23 risk factors extracted from the literature review. These findings offer valuable insights into the context of suicidal ideation. However, the comparison reveals a significant limitation in social media datasets, which is the absence of critical topics essential for addressing suicidal ideation. This gap could create a barrier to the development of robust NLP models for analyzing suicide-related content.

To address the limitations of unsupervised topic modeling, given the risk factors discovered in our scoping review, we performed guided topic modeling on 3 social media datasets (Table 5) for providing a more comprehensive overview of the suicide topics and risk factors present in the data. The results indicate that certain topics missed during unsupervised topic modeling were identified through guided topic modeling. This issue arises when less frequently discussed topics are overshadowed by more dominant ones, causing them to be overlooked by unsupervised methods. In contrast, guided topic modeling can specifically target and extract these less prominent yet critical topics.

Furthermore, we observed that there was a discrepancy in the distribution of topic representations across the datasets. As an example, let us consider the UMD in Table 5, where the topic “anxiety” appears 10.4 times more than the topic “PTSD.” Such disparities in topic representations could lead to an inherent bias within the models. Biases can occur when certain topics are overrepresented or underrepresented. These biases will potentially impact the effectiveness and fairness of NLP models, as they may disproportionately emphasize or neglect certain aspects of suicidal ideation. Therefore, addressing these data limitations and imbalances is crucial to ensure the development of NLP models that provide accurate and equitable insights into this critical issue.

Moreover, we observed that suicidal narratives related to dementia, sexual minority stigmas, immigration, death of loved ones, perfectionism, and anger were never discussed in these datasets, and topics of being bullied, PTSD, substance abuse,

XSL•FO
RenderX

and chronic physical problems were rarely mentioned. Financial crises and racism, which are 2 common and important risk factors of suicide, were only discussed in 1 of the datasets, with a relatively low number of examples. The most represented topics in these datasets were depression and anxiety. This observation is not surprising since many other risk factors lead to mental pressure. In social media conversations, users mostly talk about depression and anxiety, which are surface-level symptoms of what they are experiencing and not the root causes that led to these feelings.

In this study, we harnessed the innovative approach of synthetic dataset creation as a means to enhance the fairness and accuracy of NLP models in the context of detecting suicidal ideation. Using the extracted risk factors from psychology, we constructed a synthetic dataset that comprehensively represents the entire spectrum of risk factors associated with suicidal ideation. To leverage the strengths of both real and synthetic data, we augmented 30% of the UMD by incorporating our synthetic dataset. Table 6 reports the distribution of psychological topics within our synthetic and augmented datasets. We applied guided topic modeling on the synthetic and augmented datasets primarily as a sanity check to show that guided topic modeling performs as expected and, therefore, to analyze the content of datasets according to expected topics. In other words, our goal was to verify whether all relevant topics and risk factors were included in the dataset and could be identified by a cluster- and transformer-based model.

Understanding complexity and readability in synthetic datasets helps in ensuring that the generated text aligns with linguistic patterns observed in real-world data. Moreover, these parameters facilitate an assessment of the synthetic dataset, specifically regarding the incorporation of suitable language complexities. This evaluation allows us to examine whether synthetic data replicates language patterns akin to those found in genuine, human-generated content.

Shannon entropy serves as a quantitative measure of diversity, reflecting the range of vocabulary in a dataset. Our study's finding of a lower Shannon entropy in synthetic datasets despite higher complexity prompts a nuanced discussion. This discrepancy indicates that, despite intricate language patterns, synthetic datasets might lack the diverse lexical richness found in real datasets. Thus, the relationship between diversity and Shannon entropy suggests that achieving linguistic complexity does not guarantee a broad vocabulary range. Hence, augmenting a real dataset with synthetic data can leverage the advantages of both datasets, incorporating the linguistic complexity characteristic of synthetic datasets and the broad range of diversity inherent in real datasets. Moreover, the relationship between diversity and complexity in datasets reveals a fascinating interplay. While complexity often indicates intricate language structures, the presence of a diverse range of expressions and ideas enhances the overall diversity of the dataset. However, as seen in Table 7, a higher complexity in synthetic datasets does not necessarily translate to a higher diversity, as reflected by a lower Shannon entropy. This suggests that complexity might be influenced more by the intricacy of language patterns than by a broad lexical spectrum.

In order to guarantee the quality of the augmented and synthetic datasets, we fine-tuned the ALBERT model, as a binary classifier, using the aforementioned datasets. Following this fine-tuning, we proceeded to rigorously evaluate the model's performance on 2 distinct testing subsets. As depicted in Table 9, the results underscore an interesting trend. The model fine-tuned on the UMD exhibited a higher performance when evaluated on an in-domain testing subset but showed a lower performance when tested on an out-of-domain testing subset. A similar pattern emerged when the model was fine-tuned on the ChatGPT synthetic dataset. However, an intriguing twist arose with the augmented dataset. This dataset notably enhanced the model's performance across both the in-domain and out-of-domain testing subsets. This observation implies that the synergistic inclusion of both real-world and synthetically generated data can yield a more diverse NLP model. This research insight illuminates the potential advantages of a hybrid approach, emphasizing the importance of leveraging the complementary strengths of both real and synthetic datasets to improve the overall performance and adaptability of NLP models in the critical domain of suicidal ideation detection.

Our study has some limitations that should be acknowledged. First, high-quality annotated datasets collected from social media are scarce, so this work exclusively used the Reddit dataset. While valuable information on this topic is shared on other social media platforms, such as X and Facebook, to the best of our knowledge, there is no publicly available large annotated dataset that includes such data. Creating new datasets involves both data collection and a careful annotation process performed by multiple experts, making it an expensive undertaking. Second, due to the limitation of data availability and the fact that the classifier was trained solely on Reddit data, it may not perform as effectively on datasets from other social media platforms. Future work should incorporate data from platforms like X to provide broader insights and improve generalizability.

This study highlights the value of integrating domain knowledge from psychology with computational techniques to improve suicidal ideation detection in social media. By examining 3 real-world datasets, we observed that several well-established risk factors, such as chronic illness, racism, immigration stress, and identity-based stigma, were either absent or rarely discussed. To address these topic gaps, we applied guided topic modeling informed by psychology literature and constructed a synthetic dataset with broader thematic representation. Our evaluations showed that this synthetic data, when combined with real data, enhanced topic diversity and preserved linguistic quality. The classifier fine-tuned on the augmented dataset significantly outperformed models trained solely on real or synthetic data, improving the F_1-score to 0.91 on the UMD and 0.90 on the synthetic test set. These results suggest that synthetic augmentation not only fills in content gaps but also leads to more effective and inclusive detection models.

The findings of this study underscore the importance of integrating psychological insights into NLP models to improve the detection of suicidal ideation on social media. By addressing gaps in topic coverage and leveraging synthetic data augmentation, this work contributes to the development of more

XSL•FO
RenderX

robust and ethical artificial intelligence tools for mental health applications. Future research should continue refining these approaches by expanding dataset sources and enhancing model interpretability to ensure that technology is both effective and fair in handling sensitive mental health issues.

Acknowledgments

This research was conducted with the support of the University of Ottawa and the Natural Sciences and Engineering Research Council of Canada through the Discovery Grant program. We sincerely thank the anonymous reviewers for their insightful feedback, which helped improve the quality of this work. We also extend our gratitude to the creators of the University of Maryland Reddit Suicidality Dataset for providing access to the data, as well as to the researchers whose contributions to suicide risk assessment have informed and enriched our study.

Data Availability

The generated synthetic dataset and the code used in this study are publicly available on our GitHub repository [123]. The other datasets used in this research, including the University of Maryland Reddit Suicidality Dataset (UMD), are not publicly available and require permission from the creators of the dataset for access. More information on obtaining the UMD can be found on the UMD website [102]. Researchers interested in accessing the UMD should contact the dataset creators directly for permission and further instructions. By making these resources accessible, we aim to promote transparency, reproducibility, and further research on suicidal ideation detection.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[PDF File (Adobe PDF File), 77 KB - formative_v9i1e63272_app1.pdf]

References

1. Suicide. World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/suicide [accessed 2025-05-19]
2. O'Connor RC, Nock MK. The psychology of suicidal behaviour. The Lancet Psychiatry 2014 Jun;1(1):73-85. [doi: 10.1016/s2215-0366(14)70222-6]
3. Ghosal S, Jain A. Depression and suicide risk detection on social media using fastText embedding and XGBoost classifier. Procedia Computer Sci 2023;218:1631-1639. [doi: 10.1016/j.procs.2023.01.141]
4. Sawhney R, Joshi H, Gandhi S, Shah RR. Towards ordinal suicide ideation detection on social media. In: Proceedings of the 14th ACM International Conference on Web Search and Data Mining. 2021 Presented at: WSDM '21; March 8-12, 2021; Virtual Event, Israel. [doi: 10.1145/3437963.3441805]
5. Fernandes AC, Dutta R, Velupillai S, Sanyal J, Stewart R, Chandran D. Identifying suicide ideation and suicidal attempts in a psychiatric clinical research database using natural language processing. Sci Rep 2018 May 09;8(1):7426. [doi: 10.1038/s41598-018-25773-2] [Medline: 29743531]
6. Bejan CA, Ripperger M, Wilimitis D, Ahmed R, Kang J, Robinson K, et al. Improving ascertainment of suicidal ideation and suicide attempt with natural language processing. Sci Rep 2022 Sep 07;12(1):15146. [doi: 10.1038/s41598-022-19358-3] [Medline: 36071081]
7. Buddhitha P, Inkpen D. Multi-task learning to detect suicide ideation and mental disorders among social media users. Front Res Metr Anal 2023 Apr 17;8:1152535 [FREE Full text] [doi: 10.3389/frma.2023.1152535] [Medline: 37138946]
8. Boonyarat P, Liew DJ, Chang Y. Leveraging enhanced BERT models for detecting suicidal ideation in Thai social media content amidst COVID-19. Information Processing & Management 2024 Jul;61(4):103706. [doi: 10.1016/j.ipm.2024.103706]
9. Singh LG, Mao J, Mutalik R, Middleton SE. Extracting and Summarizing Evidence of Suicidal Ideation in Social Media Contents Using Large Language Models. In: Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology. 2024 Presented at: CLPsych 2024; March 2024; St. Julians, Malta.
10. Ghanadian H, Nejadgholi I, Al Osman H. ChatGPT for suicide risk assessment on social media: quantitative evaluation of model performance, potentials and limitations. arXiv. Preprint posted online on June 15, 2023 [FREE Full text] [doi: 10.18653/v1/2023.wassa-1.16]
11. Skaik R, Inkpen D. Suicide ideation estimators within Canadian provinces using machine learning tools on social media text. J Adv Inf Technol 2021;12(4):357-362. [doi: 10.12720/jait.12.4.357-362]
12. Franklin JC, Ribeiro JD, Fox KR, Bentley KH, Kleiman EM, Huang X, et al. Risk factors for suicidal thoughts and behaviors: a meta-analysis of 50 years of research. Psychol Bull 2017 Feb;143(2):187-232. [doi: 10.1037/bul0000084] [Medline: 27841450]

XSL•FO
RenderX

13. Bentley KH, Franklin JC, Ribeiro JD, Kleiman EM, Fox KR, Nock MK. Anxiety and its disorders as risk factors for suicidal thoughts and behaviors: a meta-analytic review. Clin Psychol Rev 2016 Feb;43:30-46 [FREE Full text] [doi: 10.1016/j.cpr.2015.11.008] [Medline: 26688478]
14. Bayliss LT, Christensen S, Lamont-Mills A, du Plessis C. Suicide capability within the ideation-to-action framework: a systematic scoping review. PLoS One 2022;17(10):e0276070 [FREE Full text] [doi: 10.1371/journal.pone.0276070] [Medline: 36301944]
15. Orsolini L, Latini R, Pompili M, Serafini G, Volpe U, Vellante F, et al. Understanding the complex of suicide in depression: from research to clinics. Psychiatry Investig 2020 Mar;17(3):207-221 [FREE Full text] [doi: 10.30773/pi.2019.0171] [Medline: 32209966]
16. Kalin NH. Insights into suicide and depression. Am J Psychiatry 2020 Oct 01;177(10):877-880. [doi: 10.1176/appi.ajp.2020.20081207] [Medline: 32998546]
17. Costa LDS, Alencar AP, Nascimento Neto PJ, dos Santos M, da Silva CGL, Pinheiro SDFL, et al. Risk factors for suicide in bipolar disorder: a systematic review. J Affect Disord 2015 Jan 01;170:237-254. [doi: 10.1016/j.jad.2014.09.003] [Medline: 25261630]
18. Paris J. Suicidality in borderline personality disorder. Medicina (Kaunas) 2019 May 28;55(6):223 [FREE Full text] [doi: 10.3390/medicina55060223] [Medline: 31142033]
19. Lee KH, Jun JS, Kim YJ, Roh S, Moon SS, Bukonda N, et al. Mental health, substance abuse, and suicide among homeless adults. J Evid Inf Soc Work 2017 Jul 05;14(4):229-242. [doi: 10.1080/23761407.2017.1316221] [Medline: 28678621]
20. Homan S, Gabi M, Klee N, Bachmann S, Moser A, Duri' M, et al. Linguistic features of suicidal thoughts and behaviors: a systematic review. Clin Psychol Rev 2022 Jul;95:102161 [FREE Full text] [doi: 10.1016/j.cpr.2022.102161] [Medline: 35636131]
21. Arensman E, Larkin C, McCarthy J, Leitao S, Corcoran P, Williamson E, et al. Psychosocial, psychiatric and work-related risk factors associated with suicide in Ireland: optimised methodological approach of a case-control psychological autopsy study. BMC Psychiatry 2019 Sep 06;19(1):275 [FREE Full text] [doi: 10.1186/s12888-019-2249-6] [Medline: 31492119]
22. Brockie T, Kahn-John M, Mata Lopez L, Bell E, Brockie T, Brockie T, et al. A mixed-methods study protocol on factors contributing to suicide clusters among Native American youth in a northern plains reservation. Front Public Health 2023 Jan 8;11:1281109 [FREE Full text] [doi: 10.3389/fpubh.2023.1281109] [Medline: 38259800]
23. Grattidge L, Hoang H, Mond J, Lees D, Visentin D, Auckland S. Exploring community-based suicide prevention in the context of rural Australia: a qualitative study. Int J Environ Res Public Health 2023 Feb 01;20(3):2644 [FREE Full text] [doi: 10.3390/ijerph20032644] [Medline: 36768008]
24. Liu L, Pollock NJ, Contreras G, Tonmyr L, Thompson W. Pandemic-related impacts and suicidal ideation among adults in Canada: a population-based cross-sectional study. Health Promot Chronic Dis Prev Can 2023 Mar 15;43(3):105-118 [FREE Full text] [doi: 10.24095/hpcdp.43.3.01] [Medline: 36630582]
25. Kumar M, Dredze M, Coppersmith G, De Choudhury M. Detecting Changes in Suicide Content Manifested in Social Media Following Celebrity Suicides. In: HT '15: Proceedings of the 26th ACM Conference on Hypertext & Social Media. 2015 Presented at: 26th ACM Conference on Hypertext & Social Media; September 1-4, 2015; Guzelyurt, Northern Cyprus. [doi: 10.1145/2700171.2791026]
26. Ghanadian H, Nejadgholi I, Osman HA. Socially aware synthetic data generation for suicidal ideation detection using large language models. IEEE Access 2024;12:14350-14363. [doi: 10.1109/access.2024.3358206]
27. Li Z, Zhu H, Lu Z, Yin M. Synthetic data generation with large language models for text classification: potential and limitations. arXiv. Preprint posted online on October 11, 2023 [FREE Full text] [doi: 10.18653/v1/2023.emnlp-main.647]
28. BERTopic. GitHub. URL: https://maartengr.github.io/BERTopic/index.html [accessed 2025-05-19]
29. Maarten G. BERTopic: neural topic modeling with a class-based TF-IDF procedure. arXiv. Preprint posted online on March 11, 2022 [FREE Full text]
30. Egger R, Yu J. A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify Twitter posts. Front Sociol 2022 May 6;7:886498 [FREE Full text] [doi: 10.3389/fsoc.2022.886498] [Medline: 35602001]
31. Reimers N, Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019 Presented at: 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing; November 2019; Hong Kong, China. [doi: 10.18653/v1/d19-1410]
32. Pretrained Models. SBERT. URL: https://www.sbert.net/docs/sentence_transformer/pretrained_models.html [accessed 2025-05-19]
33. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv. Preprint posted online on February 9, 2018 [FREE Full text]
34. Asyaky MS, Mandala R. Improving the Performance of HDBSCAN on Short Text Clustering by Using Word Embedding and UMAP. 2021 Presented at: 8th International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA); September 29-30, 2021; Bandung, Indonesia.

XSL•FO
RenderX

35. Oskolkov N. Dimensionality reduction. In: Egger R, editor. Applied Data Science in Tourism. Tourism on the Verge. Cham: Springer; 2022:151-167.
36. Campello RJGB, Moulavi D, Sander J. Density-based clustering based on hierarchical density estimates. In: Pei J, Tseng VS, Cao L, Motoda H, Xu G, editors. Advances in Knowledge Discovery and Data Mining. PAKDD 2013. Lecture Notes in Computer Science, vol 7819. Berlin, Heidelberg: Springer; 2013:160-172.
37. Munn Z, Peters MDJ, Stern C, Tufanaru C, McArthur A, Aromataris E. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. BMC Med Res Methodol 2018 Nov 19;18(1):143 [FREE Full text] [doi: 10.1186/s12874-018-0611-x] [Medline: 30453902]
38. Armstrong R, Hall BJ, Doyle J, Waters E. Cochrane Update. 'Scoping the scope' of a cochrane review. J Public Health (Oxf) 2011 Mar 23;33(1):147-150. [doi: 10.1093/pubmed/fdr015] [Medline: 21345890]
39. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodology 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
40. Peters MDJ, Godfrey CM, Khalil H, McInerney P, Parker D, Soares CB. Guidance for conducting systematic scoping reviews. Int J Evid Based Healthc 2015 Sep;13(3):141-146. [doi: 10.1097/XEB.0000000000000050] [Medline: 26134548]
41. Boergers J, Spirito A, Donaldson D. Reasons for adolescent suicide attempts: associations with psychological functioning. J Am Acad Child Adolesc Psychiatry 1998 Dec;37(12):1287-1293. [doi: 10.1097/00004583-199812000-00012] [Medline: 9847501]
42. Klonsky ED, May AM, Saffer BY. Suicide, suicide attempts, and suicidal ideation. Annu Rev Clin Psychol 2016 Mar 28;12(1):307-330. [doi: 10.1146/annurev-clinpsy-021815-093204] [Medline: 26772209]
43. Vilhjalmsson R, Kristjansdottir G, Sveinbjarnardottir E. Factors associated with suicide ideation in adults. Soc Psychiatry Psychiatr Epidemiol 1998 Mar 25;33(3):97-103. [doi: 10.1007/s001270050028] [Medline: 9540383]
44. Lázaro-Pérez C, Munuera Gómez P, Martínez-López J, Gómez-Galán J. Predictive factors of suicidal ideation in Spanish university students: a health, preventive, social, and cultural approach. J Clin Med 2023 Feb 02;12(3):1207. [doi: 10.3390/jcm12031207] [Medline: 36769853]
45. Florez IA, Allbaugh LJ, Harris CE, Schwartz AC, Kaslow NJ. Suicidal ideation and hopelessness in PTSD: spiritual well-being mediates outcomes over time. Anxiety Stress Coping 2018 Jan 28;31(1):46-58. [doi: 10.1080/10615806.2017.1369260] [Medline: 28846030]
46. Noor N, Pao C, Dragomir-Davis M, Tran J, Arbona C. PTSD symptoms and suicidal ideation in US female firefighters. Occup Med (Lond) 2019 Dec 31;69(8-9):577-585. [doi: 10.1093/occmed/kqz057] [Medline: 31064010]
47. Goldstein TR, Ha W, Axelson DA, Goldstein BI, Liao F, Gill MK, et al. Predictors of prospectively examined suicide attempts among youth with bipolar disorder. Arch Gen Psychiatry 2012 Nov 01;69(11):1113-1122 [FREE Full text] [doi: 10.1001/archgenpsychiatry.2012.650] [Medline: 22752079]
48. Goodwin FK, Jamison KR. Manic-Depressive Illness: Bipolar Disorders and Recurrent Depression. Oxford, England: Oxford University Press; 2007.
49. Schultz SH, North SW, Shields CG. Schizophrenia: a review. Am Fam Physician 2007 Jun 15;75(12):1821-1829 [FREE Full text] [Medline: 17619525]
50. Hawton K, Sutton L, Haw C, Sinclair J, Deeks JJ. Schizophrenia and suicide: systematic review of risk factors. Br J Psychiatry 2005 Jul 02;187(1):9-20. [doi: 10.1192/bjp.187.1.9] [Medline: 15994566]
51. Sher L, Kahn RS. Suicide in schizophrenia: an educational overview. Medicina (Kaunas) 2019 Jul 10;55(7):361 [FREE Full text] [doi: 10.3390/medicina55070361] [Medline: 31295938]
52. Paris J. Chronic suicidality among patients with borderline personality disorder. Psychiatr Serv 2002 Jun;53(6):738-742. [doi: 10.1176/appi.ps.53.6.738] [Medline: 12045312]
53. Brodsky BS, Malone KM, Ellis SP, Dulit RA, Mann JJ. Characteristics of borderline personality disorder associated with suicidal behavior. Am J Psychiatry 1997 Dec;154(12):1715-1719. [doi: 10.1176/ajp.154.12.1715] [Medline: 9396951]
54. Jang J, Park J, Oh K, Lee K, Kim MS, Yoon M, et al. Predictors of suicidal ideation in a community sample: roles of anger, self-esteem, and depression. Psychiatry Res 2014 Apr 30;216(1):74-81. [doi: 10.1016/j.psychres.2013.12.054] [Medline: 24507544]
55. Lee J, Lee M, Liao S, Chang C, Sung S, Chiang H, et al. Prevalence of suicidal ideation and associated risk factors in the general population. Journal of the Formosan Medical Association 2010 Feb;109(2):138-147. [doi: 10.1016/s0929-6646(10)60034-4]
56. Landberg J. Alcohol and suicide in eastern Europe. Drug Alcohol Rev 2008 Jul 29;27(4):361-373. [doi: 10.1080/09595230802093778] [Medline: 18584385]
57. Mino A, Bousquet A, Broers B. Substance abuse and drug-related death, suicidal ideation, and suicide: a review. Crisis 1999 Jan;20(1):28-35. [doi: 10.1027//0227-5910.20.1.28] [Medline: 10365504]
58. Rizk MM, Galfalvy H, Miller JM, Milak M, Parsey R, Grunebaum M, et al. Characteristics of depressed suicide attempters with remitted substance use disorders. J Psychiatr Res 2021 May;137:572-578 [FREE Full text] [doi: 10.1016/j.jpsychires.2020.10.041] [Medline: 33158553]

XSL•FO
RenderX

59. Andersson HW, Mosti MP, Nordfjærn T. Suicidal ideation among inpatients with substance use disorders: Prevalence, correlates and gender differences. Psychiatry Res 2022 Nov;317:114848 [FREE Full text] [doi: 10.1016/j.psychres.2022.114848] [Medline: 36116184]
60. Bryan CJ, Bryan AO. Financial strain, suicidal thoughts, and suicidal behavior among US military personnel in the National Guard. Crisis 2019 Nov;40(6):437-445. [doi: 10.1027/0227-5910/a000592] [Medline: 31030554]
61. Farabaugh A, Bitran S, Nyer M, Holt DJ, Pedrelli P, Shyu I, et al. Depression and suicidal ideation in college students. Psychopathology 2012 May 22;45(4):228-234. [doi: 10.1159/000331598] [Medline: 22627683]
62. Okechukwu FO, Ogba KTU, Nwufo JI, Ogba MO, Onyekachi BN, Nwanosike CI, et al. Academic stress and suicidal ideation: moderating roles of coping style and resilience. BMC Psychiatry 2022 Aug 12;22(1):546. [doi: 10.1186/s12888-022-04063-2] [Medline: 35962365]
63. Hatchel T, Polanin JR, Espelage DL. Suicidal thoughts and behaviors among LGBTQ youth: meta-analyses and a systematic review. Archives of Suicide Research 2019 Oct 10;25(1):1-37. [doi: 10.1080/13811118.2019.1663329]
64. Kaufman EA, Meddaoui B, Seymour NE, Victor SE. The roles of minority stress and thwarted belongingness in suicidal ideation among cisgender and transgender/nonbinary LGBTQ+ individuals. Arch Suicide Res 2023 Oct 10;27(4):1296-1311. [doi: 10.1080/13811118.2022.2127385] [Medline: 36214309]
65. Sutter M, Perrin PB. Discrimination, mental health, and suicidal ideation among LGBTQ people of color. J Couns Psychol 2016 Jan;63(1):98-105. [doi: 10.1037/cou0000126] [Medline: 26751158]
66. Rhoades H, Rusow JA, Bond D, Lanteigne A, Fulginiti A, Goldbach JT. Homelessness, mental health and suicidality among LGBTQ youth accessing crisis services. Child Psychiatry Hum Dev 2018 Aug 10;49(4):643-651. [doi: 10.1007/s10578-018-0780-1] [Medline: 29322361]
67. de Lange J, Baams L, van Bergen DD, Bos HM, Bosker RJ. Minority stress and suicidal ideation and suicide attempts among LGBT adolescents and young adults: a meta-analysis. LGBT Health 2022 Jun 01;9(4):222-237. [doi: 10.1089/lgbt.2021.0106] [Medline: 35319281]
68. Luo Z, Wang J, Zhou Y, Mao Q, Lang B, Xu S. Workplace bullying and suicidal ideation and behaviour: a systematic review and meta-analysis. Public Health 2023 Sep;222:166-174. [doi: 10.1016/j.puhe.2023.07.007] [Medline: 37544128]
69. Husky MM, Bitfoi A, Carta MG, Goelitz D, Koç C, Lesinskiene S, et al. Bullying involvement and suicidal ideation in elementary school children across Europe. J Affect Disord 2022 Feb 15;299:281-286. [doi: 10.1016/j.jad.2021.12.023] [Medline: 34915081]
70. Zaborskis A, Ilionsky G, Tesler R, Heinz A. The association between cyberbullying, school bullying, and suicidality among adolescents. Crisis 2019 Mar;40(2):100-114. [doi: 10.1027/0227-5910/a000536] [Medline: 30109963]
71. Finkelhor D. Sexual abuse: a sociological perspective. Child Abuse Negl 1982 Jan;6(1):95-102. [doi: 10.1016/0145-2134(82)90055-2] [Medline: 6892290]
72. Brokke SS, Bertelsen TB, Landrø NI, Haaland VO. The effect of sexual abuse and dissociation on suicide attempt. BMC Psychiatry 2022 Jan 10;22(1):29 [FREE Full text] [doi: 10.1186/s12888-021-03662-9] [Medline: 35012509]
73. Lindert J, von Ehrenstein OS, Grashow R, Gal G, Braehler E, Weisskopf MG. Sexual and physical abuse in childhood is associated with depression and anxiety over the life course: systematic review and meta-analysis. Int J Public Health 2014 Apr 12;59(2):359-372. [doi: 10.1007/s00038-013-0519-5] [Medline: 24122075]
74. Peteet JR, Maytal G, Rokni H. Unimaginable loss: contingent suicidal ideation in family members of oncology patients. Psychosomatics 2010 Mar 23;51(2):166-170. [doi: 10.1176/appi.psy.51.2.166] [Medline: 20332292]
75. Keum BT, Wong MJ, Salim-Eissa R. Gendered racial microaggressions, internalized racism, and suicidal ideation among emerging adult Asian American women. Int J Soc Psychiatry 2023 Mar;69(2):342-350 [FREE Full text] [doi: 10.1177/00207640221089536] [Medline: 35411802]
76. Robinson WL, Whipple CR, Jason LA, Flack CE. African American adolescent suicidal ideation and behavior: the role of racism and prevention. J Community Psychol 2021 Jul;49(5):1282-1295 [FREE Full text] [doi: 10.1002/jcop.22543] [Medline: 33675671]
77. Keum BT, Wong MJ. COVID-19 anti-Asian racism, perceived burdensomeness, thwarted belongingness, and suicidal ideation among Asian American Emerging Adults. Int Rev Psychiatry 2023;35(3-4):302-309 [FREE Full text] [doi: 10.1080/09540261.2023.2182186] [Medline: 37267028]
78. Ratkowska KA, De Leo D. Suicide in immigrants: an overview. Open J Med Psychol 2013;02(03):124-133. [doi: 10.4236/ojmp.2013.23019]
79. Hovey JD. Acculturative stress, depression, and suicidal ideation in Mexican immigrants. Cultur Divers Ethnic Minor Psychol 2000 May;6(2):134-151. [doi: 10.1037/1099-9809.6.2.134] [Medline: 10910528]
80. Naismith H, Howard R, Stewart R, Pitman A, Mueller C. Suicidal ideation in dementia: associations with neuropsychiatric symptoms and subtype diagnosis. International Psychogeriatrics 2022 Apr;34(4):399-406. [doi: 10.1017/s1041610222000126]
81. Cui R, Maxfield M, Fiske A. Dementia-related anxiety and coping styles associated with suicidal ideation. Aging Ment Health 2020 Nov 04;24(11):1912-1915. [doi: 10.1080/13607863.2019.1636205] [Medline: 31271051]
82. Armstrong MJ, Moore K, Jacobson CE, Bedenfield N, Patel B, Sullivan JL. Frequency of suicidal ideation and associated clinical features in Lewy body dementia. Parkinsonism Relat Disord 2021 Sep;90:33-37. [doi: 10.1016/j.parkreldis.2021.07.029] [Medline: 34352608]

XSL•FO
RenderX

83. Fässberg MM, Cheung G, Canetto SS, Erlangsen A, Lapierre S, Lindner R, et al. A systematic review of physical illness, functional disability, and suicidal behaviour among older adults. Aging Ment Health 2016 Sep 18;20(2):166-194 [FREE Full text] [doi: 10.1080/13607863.2015.1083945] [Medline: 26381843]
84. Vasiliadis H, D'Aiuto C, Lamoureux-Lamarche C, Pitrou I, Gontijo Guerra S, Berbiche D. Pain, functional disability and mental disorders as potential mediators of the association between chronic physical conditions and suicidal ideation in community living older adults. Aging Ment Health 2022 Apr 23;26(4):791-802. [doi: 10.1080/13607863.2021.1913478] [Medline: 33890523]
85. Racine M. Chronic pain and suicide risk: A comprehensive review. Prog Neuropsychopharmacol Biol Psychiatry 2018 Dec 20;87(Pt B):269-280. [doi: 10.1016/j.pnpbp.2017.08.020] [Medline: 28847525]
86. Guided Topic Modeling. GitHub. URL: https://maartengr.github.io/BERTopic/getting_started/guided/guided.html [accessed 2025-05-19]
87. Merriam-Webster. URL: https://www.merriam-webster.com/ [accessed 2025-05-19]
88. Flesch RF. Estimating the comprehension difficulty of magazine articles. J General Psychol 1943 Jan;28(1):63-80. [doi: 10.1080/00221309.1943.10545195]
89. Chotlos JW. IV. A statistical and comparative analysis of individual written language samples. Psychological Monographs 1944;56(2):75-111. [doi: 10.1037/h0093511]
90. Richards B. Type/Token Ratios: what do they really tell us? J Child Lang 1987 Jun 17;14(2):201-209. [doi: 10.1017/s0305000900012885] [Medline: 3611238]
91. Shannon CE. A mathematical theory of communication. Bell Syst Technical J 1948;27(3):379-423 [FREE Full text] [doi: 10.1002/j.1538-7305.1948.tb01338.x]
92. Snow EL, Allen LK, Jacovina ME, Crossley SA, Perret CA, McNamara DS. Keys to detecting writing flexibility over time: entropy and natural language processing. Learning Analytics 2016 Feb 18;2(3):40-54. [doi: 10.18608/jla.2015.23.4]
93. Zirikly A, Resnik P, Uzuner O, Hollingshead K. CLPsych 2019 Shared Task: Predicting the Degree of Suicide Risk in Reddit Posts. In: Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology. 2019 Presented at: Sixth Workshop on Computational Linguistics and Clinical Psychology; June 2019; Minneapolis, MN.
94. Shing HC, Nair S, Zirikly A, Friedenberg M, Daumé H, Resnik P. Expert, Crowdsourced, and Machine Assessment of Suicide Risk via Online Postings. In: Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic. 2018 Presented at: Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic; June 2018; New Orleans, LA. [doi: 10.18653/v1/w18-0603]
95. Depression subreddit. Reddit. URL: https://www.reddit.com/r/depression/ [accessed 2025-05-19]
96. SuicideWatch subreddit. Reddit. URL: https://www.reddit.com/r/SuicideWatch/ [accessed 2025-05-19]
97. Walsh CG, Ribeiro JD, Franklin JC. Predicting risk of suicide attempts over time through machine learning. Clinical Psychological Science 2017 Apr 11;5(3):457-469. [doi: 10.1177/2167702617691560]
98. Ji S, Li X, Huang Z, Cambria E. Suicidal ideation and mental disorder detection with attentive relation networks. Neural Comput & Applic 2021 Jun 24;34(13):10309-10319. [doi: 10.1007/s00521-021-06208-y]
99. Ji S, Zhang T, Ansari L, Fu J, Tiwari P, Cambria E. MentalBERT: publicly available pretrained language models for mental healthcare. arXiv. Preprint posted online on October 29, 2021 [FREE Full text]
100. Gaur M, Alambo A, Sain JP, Kursuncu U, Thirunarayan K, Kavuluru R, et al. Knowledge-aware assessment of severity of suicide risk for early intervention. 2019 Presented at: WWW '19: The World Wide Web Conference; May 13-17, 2019; San Francisco, CA. [doi: 10.1145/3308558.3313698]
101. API documentation. Reddit. URL: https://www.reddit.com/dev/api/ [accessed 2025-05-19]
102. The University of Maryland Reddit Suicidality Dataset. GitHub. URL: https://psresnik.github.io/umd_reddit_suicidality_dataset.html [accessed 2025-05-19]
103. Giupponi G, Innamorati M, Baldessarini RJ, De Leo D, de Giovannelli F, Pycha R, et al. Factors associated with suicide: Case-control study in South Tyrol. Compr Psychiatry 2018 Jan;80:150-154 [FREE Full text] [doi: 10.1016/j.comppsych.2017.09.010] [Medline: 29091781]
104. Lasserre AM, Marti-Soler H, Strippoli MF, Vaucher J, Glaus J, Vandeleur CL, et al. Clinical and course characteristics of depression and all-cause mortality: a prospective population-based study. J Affect Disord 2016 Jan 01;189:17-24. [doi: 10.1016/j.jad.2015.09.010] [Medline: 26402343]
105. Moitra M, Santomauro D, Degenhardt L, Collins PY, Whiteford H, Vos T, et al. Estimating the risk of suicide associated with mental disorders: a systematic review and meta-regression analysis. J Psychiatr Res 2021 May;137:242-249 [FREE Full text] [doi: 10.1016/j.jpsychires.2021.02.053] [Medline: 33714076]
106. Harris EC, Barraclough BM. Suicide as an outcome for mental disorders. A meta-analysis. Br J Psychiatry 1997 Mar;170:205-228. [doi: 10.1192/bjp.170.3.205] [Medline: 9229027]
107. Langford VM. Risk factors for suicide in men. Nurs Clin North Am 2023 Dec;58(4):513-524. [doi: 10.1016/j.cnur.2023.06.010] [Medline: 37832996]
108. Favril L, Yu R, Geddes JR, Fazel S. Individual-level risk factors for suicide mortality in the general population: an umbrella review. Lancet Public Health 2023 Nov;8(11):e868-e877. [doi: 10.1016/s2468-2667(23)00207-4]

XSL•FO
RenderX

109. Pemau A, Marin-Martin C, Diaz-Marsa M, de la Torre-Luque A, Ayad-Ahmed W, Gonzalez-Pinto A, et al. Risk factors for suicide reattempt: a systematic review and meta-analysis. Psychol Med 2024 Apr 16;54(9):1897-1904. [doi: 10.1017/s0033291724000904]
110. Grover C, Huber J, Brewer M, Basu A, Large M. Meta-analysis of clinical risk factors for suicide among people presenting to emergency departments and general hospitals with suicidal thoughts and behaviours. Acta Psychiatr Scand 2023 Dec 30;148(6):491-524. [doi: 10.1111/acps.13620] [Medline: 37904016]
111. Jha S, Chan G, Orji R. Identification of risk factors for suicide and insights for developing suicide prevention technologies: a systematic review and meta-analysis. Hum Behav Emerging Technologies 2023 Mar 13;2023:1-18. [doi: 10.1155/2023/3923097]
112. Anderberg J, Bogren M, Mattisson C, Brådvik L. Long-term suicide risk in anxiety-The Lundby Study 1947-2011. Arch Suicide Res 2016 Jul 02;20(3):463-475. [doi: 10.1080/13811118.2015.1057663] [Medline: 26954705]
113. Athey A, Overholser J, Bagge C, Dieter L, Vallender E, Stockmeier CA. Risk-taking behaviors and stressors differentially predict suicidal preparation, non-fatal suicide attempts, and suicide deaths. Psychiatry Res 2018 Dec;270:160-167 [FREE Full text] [doi: 10.1016/j.psychres.2018.09.032] [Medline: 30253320]
114. Page A, Morrell S, Hobbs C, Carter G, Dudley M, Duflou J, et al. Suicide in young adults: psychiatric and socio-economic factors from a case-control study. BMC Psychiatry 2014 Mar 06;14:68 [FREE Full text] [doi: 10.1186/1471-244X-14-68] [Medline: 24597482]
115. Lieb K, Zanarini MC, Schmahl C, Linehan MM, Bohus M. Borderline personality disorder. Lancet 2004;364(9432):453-461. [doi: 10.1016/S0140-6736(04)16770-6] [Medline: 15288745]
116. Zanarini M. Prediction of the 10-year course of borderline personality disorder. Am J Psychiatry 2006 May 01;163(5):827. [doi: 10.1176/appi.ajp.163.5.827]
117. Braquehais MD, Oquendo MA, Baca-García E, Sher L. Is impulsivity a link between childhood abuse and suicide? Compr Psychiatry 2010 Mar;51(2):121-129. [doi: 10.1016/j.comppsych.2009.05.003] [Medline: 20152291]
118. Devries KM, Mak JYT, Child JC, Falder G, Bacchus LJ, Astbury J, et al. Childhood sexual abuse and suicidal behavior: a meta-analysis. Pediatrics 2014 May;133(5):e1331-e1344. [doi: 10.1542/peds.2013-2166] [Medline: 24733879]
119. Wolfe KL, Nakonezny PA, Owen VJ, Rial KV, Moorehead AP, Kennard BD, et al. Hopelessness as a predictor of suicide ideation in depressed male and female adolescent youth. Suicide Life Threat Behav 2019 Feb 21;49(1):253-263 [FREE Full text] [doi: 10.1111/sltb.12428] [Medline: 29267993]
120. ALBERT Base v2. Hugging Face. URL: https://huggingface.co/albert/albert-base-v2 [accessed 2025-05-19]
121. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. HuggingFace's transformers: state-of-the-art natural language processing. arXiv. Preprint posted online on October 9, 2019 [FREE Full text] [doi: 10.18653/v1/2020.emnlp-demos.6]
122. Trainer. Hugging Face. URL: https://huggingface.co/docs/transformers/main_classes/trainer [accessed 2025-05-19]
123. Addressing Topic Gaps to Improve Suicide Ideation Detection in Social Media Posts A Topic Modeling. GitHub. URL: https://github.com/Hamideh-ghanadian/Addressing-Topic-Gaps-to-Improve-Suicide-Ideation-Detection-in-Social-Media-Posts-A-Topic-Modeling- [accessed 2025-05-20]

Abbreviations

C-TF-IDF: class-based term frequency–inverse document frequency
HDBSCAN: Hierarchical Density-Based Spatial Clustering of Applications with Noise
NLP: natural language processing
PTSD: posttraumatic stress disorder
TF-IDF: term frequency–inverse document frequency
UMAP: Uniform Manifold Approximation and Projection
UMD: University of Maryland Reddit Suicidality Dataset

Edited by A Mavragani; submitted 14.06.24; peer-reviewed by C Entwistle, C Haag, B Gong; comments to author 15.07.24; revised version received 10.10.24; accepted 08.05.25; published 11.06.25.

Please cite as:
Ghanadian H, Nejadgholi I, Al Osman H
Improving Suicidal Ideation Detection in Social Media Posts: Topic Modeling and Synthetic Data Augmentation Approach
JMIR Form Res 2025;9:e63272
URL: https://formative.jmir.org/2025/1/e63272
doi:10.2196/63272
PMID:

XSL•FO
RenderX

©Hamideh Ghanadian, Isar Nejadgholi, Hussein Al Osman. Originally published in JMIR Formative Research (https://formative.jmir.org), 11.06.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Analysis of the Political Viewpoint of Policy Statements From Professional Medical Organizations Using ChatGPT With GPT-4: Cross-Sectional Study

Ben Knudsen[1], MD; Amr Madkour[2], MD; Preetam Cholli[3], MD; Alyson Haslam[4], PhD; Vinay Prasad[4], MD, MPH

[1]The George Washington University School of Medicine and Health Sciences, Washington, DC, United States

[2]Department of Obstetrics and Gynecology, The George Washington University, Washington, DC, United States

[3]Department of Internal Medicine, Emory University, Atlanta, GA, United States

[4]Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, CA, United States

Corresponding Author:
Vinay Prasad, MD, MPH
Department of Epidemiology and Biostatistics
University of California San Francisco
UCSF Mission Bay Campus | Mission Hall: Global Health & Clinical Sciences Building
550 16th St, 2nd Fl
San Francisco, CA, 94158
United States
Phone: 1 (971) 219 1888
Email: Vinayak.prasad@ucsf.edu

Abstract

Background: Professional medical organizations publish policy statements that are used to impact legislation or address societal issues. Many organizations are nonpartisan, yet it is uncertain whether their policy statements balance liberal and conservative values.

Objective: This study aims to evaluate the political viewpoint of policy statements from 6 influential medical organizations, including the American Academy of Pediatrics, American College of Surgeons, American Psychiatric Association, American College of Obstetricians and Gynecologists, American College of Physicians, and American Academy of Family Physicians.

Methods: Between December 2023 and February 2024, policy statements from the 6 organizations were identified and evaluated using ChatGPT with GPT-4 to reduce bias. Each statement was pasted into a new ChatGPT session following the phrase “Does this text align with a liberal or conservative viewpoint?” Two authors reviewed each response and categorized the statement as liberal, probably liberal, neutral, probably conservative, or conservative.

Results: One-third of policy statements (529/1592, 33.2%) were found to be aligned with a political ideology. Among these 529 statements, 516 (97.5%) were liberal or probably liberal and 13 (2.5%) were conservative or probably conservative. For each organization, among policy statements with a political leaning, the percentage of liberal or probably liberal statements was as follows: 100% (69/69) for the American Academy of Pediatrics, 100% (24/24) for the American College of Obstetricians and Gynecologists, 100% (12/12) for the American College of Surgeons, 99% (72/73) for the American Psychiatric Association, 97% (174/180) for the American Academy of Family Physicians, and 96% (165/171) for the American College of Physicians.

Conclusions: One in 3 policy statements from these 6 professional organizations align with a partisan political viewpoint. Among these, positions are 40 times more likely to be liberal or probably liberal than conservative or probably conservative. Whether or not organizations are politically neutral and seek viewpoint diversity warrants further exploration.

(JMIR Form Res 2025;9:e66204) doi:10.2196/66204

KEYWORDS

medical organizations; health policy; policy statements; public health; ChatGPT; AI; chatbot; political viewpoint; political alignment; political ideology

XSL•FO
RenderX

Introduction

Background

Professional medical organizations have guided the medical community for well over a century and continue to evolve in size, function, and influence [1]. Currently, in addition to creating clinical guidelines, holding conferences [2], and publishing research [3], organizations develop policy and position statements that are used to advocate for legislative changes [4]. Political advocacy often occurs in the form of lobbying [5] and, to a lesser extent, contributions to political candidates or parties [6].

In 2023, among lobbyists for health professionals, the American Medical Association (AMA) was the top donor at US $21,215,000, followed by the American Academy of Family Physicians (AAFP) with US $3,152,917 in contributions [5]; thus, some medical organizations have leverage to impact legislation and engage in political activities.

Although some overlap between politics and medicine is unavoidable, there have been growing concerns, most notably during the COVID-19 pandemic [7-9], about clinical and public health recommendations becoming associated with a political party. Global polls suggest that 2 in 3 people worry about the politicization of medical science [10]. This has resulted in widespread effects, ranging from growing distrust among the public toward federal and other public health institutions, as well as scientific experts, to increased emotional burden and stress among practicing physicians during patient encounters [11].

Several major medical organizations, such as the AMA [12], American Academy of Pediatrics (AAP) [13,14], American College of Surgeons (ACS) [15], American College of Physicians (ACP) [16], American College of Obstetricians and Gynecologists (ACOG) [17], and others, label themselves as nonpartisan entities while simultaneously publishing policy statements that center on political issues. It is uncertain whether these policy statements are nonpartisan or bipartisan, given that some medical organizations have recently (and historically) [18] taken stances on important political issues [19], including climate change [20-24] and universal healthcare coverage [25-27]. This may indicate a preference for policies favored by one political party.

Objectives

Given the influence and involvement of medical organizations in politics, their endorsement of political issues, and the negative effects of the politicization of medicine, we sought to characterize the political viewpoint (liberal vs conservative) of all policy statements from the following 6 prominent medical organizations using ChatGPT with GPT-4 (OpenAI): AAP, ACS, ACOG, ACP, AAFP, and American Psychiatric Association (APA). Our analysis seeks to provide the first descriptive analysis of the political alignment of policy preferences of medical societies. To reduce potential bias and more objectively ascertain the political leaning of each statement, a third-party method—ChatGPT—was used to code the political valence of all statements.

Methods

Overview

We used ChatGPT with GPT-4 to analyze the political alignment of policy and position statements (referred to as policy statements) from 6 medical organizations, including AAP, ACS, APA, ACOG, ACP, and AAFP. ChatGPT is a chatbot designed to understand input presented by users and provide a humanlike text response.

We limited our analysis to flagship organizations for medical specialties that are considered core rotations in medical school (pediatrics, surgery, psychiatry, obstetrics and gynecology, internal medicine, and family medicine). This cross-sectional study was performed in accordance with the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) statement.

Policy Statement Inclusion Criteria

We identified all policy statements listed on the websites of the 6 organizations from December 2023 to February 2024. The following bulleted list describes the digital location of the policy statements that were retrieved for each medical organization:

- For the AAP, policy statements available under the Policy Statements section [28] through December 2023 were included. Documents under the following headings were not included: Clinical Reports, Clinical Practice Guidelines, AAP News, Statements of Endorsement, Technical Reports, or AAP Policy Books.
- For the ACS, statements available under the Statements heading [29] through December 2023 were included.
- For the APA, position statements under the Policy Finder section [30] through December 2023 were included.
- For the ACOG, reports under the headings Statements of Policy and Position Statements [31] through December 2023 were included.
- For the ACP, policy statements located in the Policy Compendium [32] through December 2023 were included. The Policy Compendium is a collection of “policy statements of a national scope adopted by the ACP Board of Regents” and listed as an “up-to-date summary of ACP’s policy positions.” Other policy statements or papers from the ACP were not considered.
- For the AAFP, policies, position statements, and position papers located in the full policy listing [33] through February 2024 were included.

Policy statements that were duplicates, old versions, or unable to be viewed or evaluated were excluded. For the AAP, definitions, prevention guidelines, AAP retired or reaffirmed, and vaccine schedule updates were excluded. The name, date (most recent), and status of each policy statement were obtained. Many policy statements were a revision of a prior statement. In these cases, the most recent version was used. Older versions were not readily available.

Political Viewpoint Determination

The whole text of each policy statement (excluding references and acknowledgments) was pasted into a new ChatGPT session

XSL•FO
RenderX

(with prior queries having been deleted) following the phrase "Does this text align with a liberal or conservative viewpoint?" The full ChatGPT response was recorded. Safeguards were used to prevent ChatGPT from being influenced by prior queries. Specifically, the first ChatGPT output was always used; the option to regenerate a response was never selected; previous ChatGPT sessions were always deleted and never present when performing the analysis for each subsequent policy statement; and after each policy statement evaluation, the ChatGPT session was deleted, and a new session was initiated.

ChatGPT was used to evaluate political alignment to reduce bias that may be present in human reviewers and for its ability to quickly analyze long and detailed documents. In addition, ChatGPT has performed highly on various professional and academic examinations [34] and is growing in popularity in medical research [35]. It is reported that ChatGPT may have a left-leaning political bias [36-38]. However, newer versions of ChatGPT [38,39], specifically GPT 3.5-turbo and GPT 4, have shown less political bias than earlier models. Despite ChatGPT's observed political bias, our methodology did not prompt ChatGPT to provide its own opinion; rather, we queried ChatGPT to match ideas in a text to well-established political values, which likely bypassed ChatGPT's inherent political preferences. Motoki et al [36] showed that ChatGPT was capable of correctly identifying liberal and conservative values, which is the foundation of our study.

ChatGPT responses did not always definitively label a statement as either liberal or conservative. Thus, each ChatGPT response was analyzed independently and separately by 2 authors (BK and AM), and the policy statement was determined to be one of the following: liberal, probably liberal, neutral, probably conservative, or conservative. If both authors agreed, the political classification was assigned to the policy statement. If the 2 authors disagreed, a third author (AH) independently analyzed and assigned a political label to the statement. Then, the third author's political classification was compared to that of the first 2 authors (BK and AM). The final classification was determined to be the political category that received the most votes. For example, if BK labeled a statement as liberal and AM labeled a statement as probably liberal, a third author, AH, evaluated the statement. If AH found the statement to be liberal, then the liberal category had received 2 votes, and the final classification for that policy statement was liberal.

Full transcripts of the ChatGPT results are available in Multimedia Appendix 1. The following bullet points provide further explanation regarding the criteria used to categorize the ChatGPT outputs:

- Liberal or conservative—the ChatGPT output clearly defined the political alignment of the policy statement. ChatGPT described principles in the policy statement that were associated with either liberal or conservative views. The explanation was nonambiguous. The explanation did not indicate that any part of the policy statement aligned with the opposing political perspective.
- Probably liberal or probably conservative—the ChatGPT output was nuanced and did not obviously indicate that the policy statement was strictly liberal or conservative. The ChatGPT explanation indicated that some, but not all, ideas in the policy statement were associated with a political ideology. ChatGPT recognized that some ideas in the policy statement aligned with both liberal and conservative views while also suggesting that 1 ideology was more apparent and discussed with more importance.
- Neutral—ChatGPT clearly indicated that the policy statement was neutral, and its contents were not political in nature. The ChatGPT output described that the policy statement contained both liberal and conservative ideologies but did not report that 1 set of ideas was more prominent.

As a quality assurance check for ChatGPT outputs, random policy statements were read by the authors after the ChatGPT query to ensure the ChatGPT responses were representative of the policy statement.

Textbox 1 contains examples of parts of the ChatGPT output that were key to categorizing the statement into a political category. Policy statements were not evaluated based on their original content but were assessed based on their associated ChatGPT output.

XSL•FO
RenderX

Textbox 1. Examples of key elements of ChatGPT output that were used for classifying the political alignment of policy statements.

Liberal

- "The text you provided appears to align more with a liberal viewpoint. This conclusion is based on several key aspects of the text..."
- "Key aspects that suggest a liberal alignment include..."
- "The text provided primarily aligns with a liberal viewpoint..."
- "The text you provided expresses a viewpoint that is generally aligned with liberal or progressive ideologies. This alignment is evident in several key aspects..."
- "The text you provided presents a viewpoint that is more commonly associated with liberal ideologies, although it's important to note that such categorizations can vary and are not always clear-cut. Here are some key reasons why this text aligns more closely with a liberal viewpoint..."
- "Conservative viewpoints, in contrast, might emphasize..."

Probably liberal

- "Overall, while the text does not explicitly align with a strictly liberal or conservative viewpoint, its emphasis on community involvement, addressing social determinants of health, and advocating for systemic changes in public health policy might resonate more with liberal perspectives..."
- "In summary, while the text itself is focused on educational criteria and does not explicitly endorse a political ideology, some of the values it promotes—such as a focus on underserved communities, inclusivity, and a comprehensive approach to healthcare—might resonate more with ideologies typically associated with progressive or liberal viewpoints..."
- "Thus, while the text aligns with principles often associated with liberal viewpoints, the call for transparency in healthcare transcends traditional political boundaries and can be supported by individuals with a wide range of political beliefs..."
- "Therefore, while some elements of the text may align more closely with liberal viewpoints, particularly in terms of regulatory and public health perspectives, the overall concern for reducing the impact of tobacco use is a bipartisan issue that transcends traditional liberal-conservative divides..."

Neutral

- "The text you provided doesn't align distinctly with either a liberal or conservative viewpoint..."
- "The content is neutral and objective, presenting facts and guidelines without any discernible bias towards a liberal or conservative viewpoint..."
- "However, the text you've provided is centered around the structure and philosophy of care rather than explicit political arguments or policy prescriptions..."
- "The text you provided primarily aligns with a viewpoint that emphasizes patient autonomy, informed consent, and ethical considerations in healthcare, which are principles that can be appreciated across the political spectrum..."
- "In summary, the text discusses healthcare and legal reform in a way that doesn't neatly fit into a liberal or conservative box. It seems to prioritize practical solutions and systemic improvement over ideological positions..."
- "In essence, while the text does not neatly fit into a single political ideology, elements of it can be appreciated or critiqued from both liberal and conservative perspectives..."

Probably conservative

- "The text primarily aligns with principles that could be appreciated by both sides for different reasons but leans more towards a conservative viewpoint..."
- "The ACP's stance against mandatory Medicare assignment by states might not align neatly with traditional liberal views, as it advocates for less state control over Medicare policies..."
- "In essence, the ACP's position might be seen as aligning more with a conservative viewpoint on regulatory issues, advocating for professional discretion over government mandates..."

Conservative

- "The text you've provided appears to align with a more conservative viewpoint..."
- "The text reflects a conservative viewpoint towards healthcare professional roles and scope of practice..."
- "This stance aligns more closely with a perspective that values structured oversight, professional standards, and integration within traditional healthcare frameworks, which can be seen as conservative in the context of healthcare policy..."
- "The text you provided represents a viewpoint that is generally aligned with conservative principles..."

Here is an example of how an ambiguous case was evaluated. The following is the ChatGPT output for an ACP policy statement titled "Proper Use of Accepted Drugs":

The text expresses a viewpoint advocating for physician autonomy and clinical judgment in determining the proper usage of approved drugs,

XSL•FO
RenderX

rather than strict legislative or administrative controls. This perspective aligns more with principles typically associated with a liberal viewpoint, which often emphasizes individual professional judgment and less restrictive government intervention in professional practices. However, some aspects of this stance could be embraced by conservatives who support limited government regulation in professional affairs. Ultimately, the alignment might vary based on the broader context of healthcare policies and the specific beliefs of individuals within each ideological group.

This statement was classified as probably liberal. There is a blurry and unclear delineation between liberal and conservative values. The output suggests that both viewpoints may align with parts of the policy statement depending on specific beliefs. However, given the emphasis on liberal values, such as individual professional judgment and less restrictive government intervention in professional practices, it was determined that overall, the policy statement was probably liberal.

Data Analysis

Among all medical organizations and for each individual organization, the percentage of statements categorized as liberal, probably liberal, neutral, probably conservative, or conservative was calculated.

Descriptive statistics are provided. Graphs were created using Excel (Microsoft Corporation). The κ coefficient was found using R (version 4.2.2; R Foundation for Statistical Computing) to determine the degree of agreement between the political categories assigned to policy statements by the 2 reviewers.

Secondary Analysis: ChatGPT Reliability

Identical queries and the words used to prompt ChatGPT have been shown to influence its output; thus, a secondary analysis was performed from December 2024 to January 2025 using the following scenarios to evaluate ChatGPT's reliability. In total, 150 random policy statements were reanalyzed using ChatGPT or Grok (xAI) with either the same prompt or an alternative prompt. This is approximately 10% of all policy statements, and this number was chosen for feasibility purposes. Grok is a chatbot with capabilities similar to ChatGPT but differs in various ways that include the developer, philosophy, and integration. In addition to its popularity and accessibility, Grok was chosen due to the possibility that it has a conservative bias, as its founder, Elon Musk, has recently been associated with the Republican party. However, this is speculative, and studies have indicated that Grok has less political bias than ChatGPT. Nonetheless, Grok is a chatbot that may exhibit a conservative leaning and thus was selected as a comparator to ChatGPT. The following list describes 3 scenarios used in the secondary analysis to evaluate ChatGPT reliability:

1. Policy statements were requeried in ChatGPT with the original prompt ("Does this text align with a liberal or conservative viewpoint?").
2. Policy statements were requeried in ChatGPT with an alternative prompt ("Please describe if you think the following statement leans toward a politically liberal or conservative stance."). This entire scenario was repeated twice with the same policy statements to gather more data.
3. Policy statements were requeried in Grok with the original prompt.

All safeguards to protect against ChatGPT bias as described in the Political Viewpoint Determination section were used. In addition, memory (this is a feature within ChatGPT settings that allows the program to learn from user responses) was cleared and disabled to prevent any prior queries within ChatGPT from influencing the output for the policy statement evaluations. A political category (liberal, probably liberal, neutral, probably conservative, or conservative) was then assigned to the policy statement by BK (classification from the original analysis was hidden). The political label was then compared to the original analysis.

The percentage of policy statements with the same political label was calculated. For statements with discordant results, the statement from the secondary analysis was characterized as either more liberal, less liberal, more conservative, or less conservative compared to the original analysis by BK. For example, if a statement was labeled neutral in the original analysis and liberal or probably liberal in the secondary analysis, it was placed in the more liberal category. Alternatively, if a statement was labeled as liberal in the original analysis and probably liberal or neutral in the secondary analysis, it was placed in the less liberal category.

Ethical Considerations

In accordance with 45 Code of Federal Regulations §46.102 (f), this study was not submitted for institutional review board approval because it involved publicly available data and did not involve individual patient data. This study did not involve individual patients, and therefore, informed consent was not required.

Results

Primary Analysis

Collectively, there were 1672 statements across the 6 high-profile medical organizations evaluated (AAP, ACS, APA, ACOG, ACP, and AAFP). After duplicates, old versions, statements that were unable to be viewed, and others (from AAP) were excluded, 95% (1592/1672) of the policy statements were included in the final analysis, which were either revised or originally published between 1986 and 2023.

For AAP, 91.6% (294/321) of the policy statements published or revised between 2010 and 2023 were included. Of the 294 policy statements, there were 225 (76.5%) neutral and 69 (23.5%) liberal or probably liberal policy statements (Table 1). The κ coefficient between reviewers was 0.841 ($P<.001$). Of liberal or probably liberal policy statements, the most common topics were related to health care (11/69, 16%), social justice (10/69, 15%), and public health (chemical exposure, immunization, nutrition, smoking, etc; 10/69, 15%; Table 2).

XSL•FO
RenderX

Table 1. Policy statements from each medical organization stratified by political categorization (liberal, probably liberal, neutral, probably conservative, or conservative).

Medical organization and all combined	Liberal, n (%)	Probably liberal, n (%)	Neutral, n (%)	Probably conservative, n (%)	Conservative, n (%)
American Academy of Pediatrics (n=294)	53 (18)	16 (5.4)	225 (76.5)	0 (0)	0 (0)
American College of Surgeons (n=106)	7 (6.6)	5 (4.7)	94 (88.7)	0 (0)	0 (0)
American Psychiatric Association (n=222)	55 (24.8)	17 (7.7)	149 (67.1)	0 (0)	1 (0.5)
American College of Obstetricians and Gynecologists (n=37)	22 (59)	2 (5)	13 (35)	0 (0)	0 (0)
American College of Physicians (n=524)	142 (27.1)	23 (4.4)	353 (67.4)	2 (0.4)	4 (0.8)
American Academy of Family Physicians (n=409)	143 (35)	31 (7.6)	229 (55)	2 (0.5)	4 (1)
All organizations combined (N=1592)	422 (26.5)	94 (5.9)	1063 (66.8)	4 (0.3)	9 (0.6)

Table 2. Themes of liberal and probably liberal policy statements. Policy statements were categorized by political theme based on their content. The number of liberal or probably liberal policy statements from each medical organization is presented, categorized by political topic. A few policy statements did not fall under these categories, and thus, they were excluded from the figure.

Political theme	AAP[a]	ACS[b]	APA[c]	ACOG[d]	ACP[e]	AAFP[f]
Health care (access, financing, efficiency, reform, etc)	11	2	7	5	38	27
Social justice (racial equity, inclusion, diversity, and LGBTQ[g] equity)	10	2	16	2	13	45
Public health	10	1	6	2	25	20
Reproductive rights	4	1	4	7	4	10
Education and medical training	5	__[h]	—	—	6	12
Criminal justice and crime	—	—	12	2	3	5
Immigration	2	—	5	1	5	5
Substance use	3	2	6	—	2	2
Pharmaceuticals	—	—	—	—	12	1
Firearms	2	1	4	1	2	2
Mental health	1	—	4	—	1	4
Social determinants of health	3	—	1	—	4	2
Clinical care	3	—	1	—	1	4
Climate and environment	2	—	1	1	2	2
COVID-19	—	—	1	1	5	1
Domestic violence	—	1	—	—	1	4
Employee rights	1	1	1	1	—	2
Poverty, welfare, and homelessness	2	—	1	—	—	2
Alternative treatments	—	—	—	—	1	2
Foreign policy and national security	—	—	—	—	3	1

[a]AAP: American Academy of Pediatrics.
[b]ACS: American College of Surgeons.
[c]APA: American Psychiatric Association.
[d]ACOG: American College of Obstetricians and Gynecologists.
[e]ACP: American College of Physicians.
[f]AAFP: American Academy of Family Physicians.
[g]LGBTQ: lesbian, gay, bisexual, transgender, queer.
[h]Not applicable.

XSL•FO
RenderX

For the ACS, 93.8% (106/113) of the policy statements published or revised between 1986 and 2023 were included. Of the 106 policy statements, there were 94 (88.7%) neutral and 12 (11.3%) liberal or probably liberal policy statements (Table 1). The κ coefficient between reviewers was 0.716 (*P*<.001). The topics of some of the liberal or probably liberal policy statements were health care (2/12, 17%), social justice (2/12, 17%), substance use (2/12, 17%), public health (1/12, 8%), reproductive rights (1/12, 8%), firearms (1/12, 8%), and employee rights (1/12, 8%; Table 2).

For the APA, 92.9% (222/239) of the policy statements published or revised between 1993 and 2023 were included. Of the 222 policy statements, there were 149 (67.1%) neutral, 72 (32.4%) liberal or probably liberal, and 1 (0.5%) conservative policy statement (Table 1). The κ coefficient between reviewers was 0.861 (*P*<.001). The conservative policy statement was about opposing cannabis as a treatment for posttraumatic stress disorder. The most common topics among liberal or probably liberal policy statements were social justice (16/72, 22%), criminal justice and crime (12/72, 17%), and health care (7/72, 10%; Table 2).

For the ACOG, 100% (37/37) of the policy statements published or revised between 2016 and 2023 were included. Of the 37 policy statements, there were 13 (35%) neutral and 24 (65%) liberal or probably liberal policy statements (Table 1). The κ coefficient between reviewers was 0.89 (*P*<.001). Among liberal or probably liberal statements, the most common topics were reproductive rights (7/24, 29%) and health care (5/24, 21%; Table 2).

For the ACP, 95.1% (524/551) of the policy statements published or revised between 2008 and 2023 were included. Of the 524 policy statements, there were 353 (67.4%) neutral, 165 (31.5%) liberal or probably liberal, and 6 (1.1%) conservative or probably conservative policy statements (Table 1). The κ coefficient between reviewers was 0.847 (*P*<.001). The most common political issues discussed among liberal or probably liberal policy statements were health care (38/165, 23%), public health (25/165, 15.1%), social justice (13/165, 7.9%), and pharmaceuticals (12/165, 7.3%; Table 2).

For the AAFP, 99.5% (409/411) of the policy statements published or revised between 2008 and 2023 were included. Of the 409 policy statements, there were 229 (55.0%) neutral, 174 (42.5%) liberal or probably liberal, and 6 (1.5%) conservative or probably conservative policy statements (Table 1). The κ coefficient between reviewers was 0.855 (*P*<.001). The most common political issues discussed among liberal or probably liberal policy statements were social justice (45/174, 25.9%), health care (27/174, 15.5%), and public health (20/174, 11.5%; Table 2).

Overall, of the 1592 policy statements, 1063 (66.8%) were neutral, 422 (26.5%) were liberal, and 94 (5.9%) were probably liberal. Of the 1592 policy statements, there were only 9 (0.6%) and 4 (0.3%) conservative and probably conservative policy statements, respectively. Examples of liberal and conservative policy statements with a selection of their ChatGPT output are shown in Multimedia Appendix 2.

Secondary Analysis

In total, 150 random policy statements were reanalyzed using an identical prompt, an alternative prompt, or Grok and compared to the original analysis (Table 3). Some of the policy statements (50/150, 33.3%) were analyzed twice using the alternative prompt (scenario 3); thus, the analysis included 200 total data points. Overall, the percentage agreement between the 2 analyses ranged from 54% to 66%. Among the policy statements coded differently, the majority (72/80, 90%) were categorized as more liberal in the secondary analysis. There were 0 policy statements that moved from either liberal or probably liberal to conservative or probably conservative or vice versa. Among statements within the "more liberal" group, most policy statements (58/72, 81%) changed from neutral (original analysis) to liberal (secondary analysis). There were 0 neutral statements in the original analysis that changed to conservative or probably conservative. There were 0 statements that were less conservative in the secondary analysis.

Table 3. Results from the secondary analysis. Scenarios describe the prompt and chatbot used to reevaluate policy statements. The percentage of policy statements coded with either the same, more liberal, less liberal, or more conservative political label is described. Footnotes explain the prompts and the types of political category change within the groups: more liberal, less liberal, and more conservative.

Scenario	Classification of policy statements relative to the original analysis, n/N (%)			
	Same	More liberal[a]	Less liberal[b]	More conservative[c]
Original prompt[d]—original prompt (ChatGPT)	33/50 (66)	14/17 (82)	3/17 (18)	0/17 (0)
Original prompt—new prompt[e] (ChatGPT)	30/50 (60)	19/20 (95)	0/20 (0)	1/20 (5)
Original prompt—new prompt (×2; ChatGPT)	27/50 (54)	20/23 (87)	2/23 (9)	1/23 (4)
Original prompt—original prompt (Grok)	30/50 (60)	19/20 (95)	1/20 (5)	0/20 (0)

[a]More liberal: neutral to liberal; neutral to probably liberal; probably liberal to liberal.

[b]Less liberal: liberal to neutral; liberal to probably liberal; probably liberal to neutral.

[c]More conservative: probably conservative to conservative.

[d]Original prompt: Does this text align with a liberal or conservative viewpoint?

[e]New prompt: Please describe if you think the following statement leans toward a politically liberal or conservative stance.

Among 71 policy statements originally coded as liberal or conservative (the most distinct categories), 67 (94%) were coded

XSL•FO
RenderX

the same in the secondary analysis. In contrast, among 119 statements originally coded as neutral, 51 (42.9%) were coded the same in the secondary analysis.

Discussion

Principal Findings

We found that 33.2% (529/1592) of policy statements from 6 highly regarded medical organizations were nonneutral and correlated with a set of political values. Of these statements, the majority (516/529, 97.5%) were aligned with a liberal or probably liberal viewpoint, whereas only 2.5% (13/529) aligned with a conservative or probably conservative viewpoint. This equates to 40 times more liberal than conservative policy statements.

For each organization, among policy statements with a political leaning, the percentage of liberal or probably liberal statements was 100% (69/69) for the AAP, 100% (24/24) for the ACOG, 100% (12/12) for the ACS, 99% (72/73) for the APA, 97% (174/180) for the AAFP, 96% (165/171) for the ACP (Figure 1). The number of conservative or probably conservative policy statements, which totaled 13, were from only 3 of the 6 organizations: the APA, ACP, and AAFP.

Figure 1. Policy statements that are liberal, probably liberal, conservative, or probably conservative for each medical organization among only policy statements found to have a political alignment. Dark blue depicts liberal, light blue depicts probably liberal, dark red depicts conservative, and light red depicts probably conservative. Blue and red were chosen due to the association with liberal and conservative political ideology, respectively, as represented in modern US politics. AAFP: American Academy of Family Physicians; AAP: American Academy of Pediatrics; ACOG: American College of Obstetricians and Gynecologists; ACP: American College of Physicians; ACS: American College of Surgeons; APA: American Psychiatric Association.

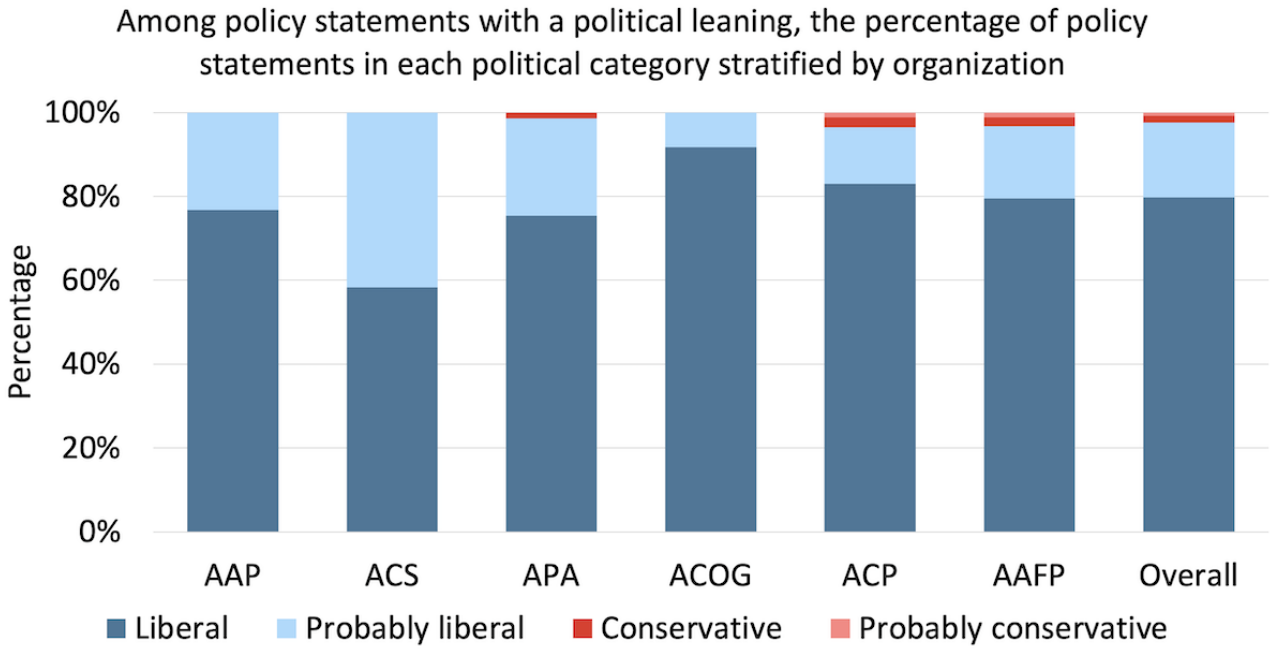

Our results may be unexpected, given that physicians from both major political parties are represented within each medical specialty [40,41]. Our results differ and represent an imbalance in favor of liberal values. For example, data published in 2016 in the *New York Times* [41] reported that about 47% of obstetrician and gynecologist physicians are registered as Republican; however, 100% (24/24) of the nonneutral policy statements published by ACOG, the largest organization for this specialty, lean liberal. Similarly, the same report showed that 52% of family medicine physicians are registered as Republican; however, the AAFP published 29 times more liberal or probably liberal than conservative or probably conservative policy statements. The same trend is true for the other organizations included in this study.

Across organizations, there were different proportions of liberal policy statements. For example, the ACOG had the highest (22/37, 59% liberal), while the ACS had the lowest (7/106, 6.6% liberal). There are a few explanations that may account for this difference. First, particular fields may attract individuals with a common political ideology. For example, psychiatry may recruit individuals with a liberal political preference due to the focus on mental health. Second, the nature of the profession may impact and alter the political ideology of an individual. People training and practicing obstetrics and gynecology may become proponents of medical abortion after witnessing the devastating complications of unsafe abortions. Finally, members of the medical organizations or individuals that write policy statements may have a common political ideology that does not capture alternative viewpoints in the field. This reflects limited viewpoint diversity, which some have identified as a problem within the field of public health [42]. Future research could seek to evaluate if causal relationships exist between individuals' political orientation and policy positions. This could be evaluated within and between medical organizations. Moreover, it may be valuable to understand how political preferences of individuals within specialties change over time and measure how this affects favorability toward policies. As noted earlier, demographics, such as age, gender, background, socioeconomic status, and professional experiences, among others, may influence policy preferences, and additional research exploring these relationships would be an important contribution.

XSL•FO
RenderX

Medical organizations have an extensive history of involvement in politics [43]; however, it is unclear how their politicization has changed over time. A cross-sectional analysis in 2024 found that 64% (35/55) of medical organizations had statements related to gender-affirming care, and 97% of the statements were supportive [44]. Furthermore, a report in 2022 found that 45% (50/111) of the US medical organizations had at least 1 type of content related to climate change on their website [45]. Finally, in our analysis, we found that 67% (4/6) of the organizations had policy statements related to COVID-19 that aligned with a liberal ideology. Although outside the scope of this analysis, it appears that medical organizations may be becoming more politicized, especially because of the COVID-19 pandemic. Research expanding on this work could aim to evaluate the political alignment of policy statements over time to explore how viewpoints have changed.

There are potential upsides and downsides to having medical societies issue policy statements. As a potential benefit, these can catalyze important policy discussions that may lead to improved advocacy efforts and outcomes. For example, pediatricians may rightfully believe that they have a duty to advocate against firearm shootings, given the negative effects on children and society. Psychiatrists may feel compelled by a responsibility to advocate for an improved criminal justice system and legal procedures, given the vulnerability of those with psychiatric disorders and their overrepresentation within the criminal justice system.

For potential downsides, it is unknown whether these statements achieve tangible change, and the effect on public trust, particularly among conservatives, may be impacted. Surveys show a major decline in public confidence in medical experts and organizations over the past 5 decades [46], and specifically in recent years, this decline has been particularly pronounced among Republicans [47,48].

Some may argue that our findings are explained merely by the fact that conservative viewpoints do not seek to maximize or promote public health, and ergo, it is natural that organizations' policy preferences are liberal. However, we contend that this point of view itself lacks empirical support and may represent the underlying bias present in organizational views. Fundamentally, conservative and liberal policy positions are typically disagreements about the solutions to societal problems, rather than the enumeration of the problems themselves. Indeed, one striking example of this is the Affordable Care Act, which addresses the problems of uninsured Americans with a marketplace solution and was originally offered by the conservative Heritage Foundation. Similar to the Affordable Care Act, there may be other potentially valuable conservative proposals, but excluding these views from professional organizations may result in missed opportunities to advance medicine and public health. Ultimately, our paper merely documents the political leaning of policy statements from high-profile medical societies and does not seek to adjudicate a reason for their alignment or their purpose.

The secondary analysis showed that ChatGPT has variable reliability depending on the content of the policy statement. For example, policy statements coded as liberal or conservative in the original analysis had 94% agreement in the secondary analysis when using the original prompt, an alternative prompt, or a different chatbot, Grok (Table 3). Liberal and conservative values represent opposite ends of the spectrum and embody principles that have minimal overlap. Thus, we found that ChatGPT had minimal variability when reanalyzing policy statements with clear political ideology. In contrast, neutral policy statements had 43% agreement in political coding. Ultimately, this is unsurprising given that policy statements without a uniform ideology often contained numerous political ideas that could be interpreted through a liberal or conservative lens based on the reader. ChatGPT frequently responded with this disclaimer.

Notably, among policy statements with discordant results, 90% (72/80) were coded as more liberal compared to the original analysis. There were 0 neutral statements in the original analysis that were coded as conservative or probably conservative in the secondary analysis. Thus, it is likely that the net vector of bias in these political statements lies in the liberal direction, which aligns with our main findings. Although these findings strengthen the study results, we do not have a concrete answer for why statements in the secondary analysis almost always shifted in the liberal direction relative to the original classification. Admittedly, we are not experts in artificial intelligence and are unfamiliar with the internal coding and processes within ChatGPT or Grok. Thus, it is hard to speculate about the reason for the results. It is possible that this represents an inherent bias that exists within various chatbots. Although the same version of ChatGPT was used in the secondary analysis, it is possible that there were minor changes to the code or internal algorithm that influenced our results. Considering these results, it is possible that our original analysis may have undercounted the number of liberal leaning statements and captured only the lower bound of the total number of politically biased statements.

Strengths and Limitations

Our study has at least 3 strengths and 5 limitations. To our knowledge, our work is the first to comprehensively evaluate most policy statements from these 6 professional organizations for political alignment. As opposed to selectively picking policy statements to analyze, we sampled most or all (in some cases) policy statements that the organizations published on their website. Second, the use of ChatGPT allowed us to reduce bias that may have been present in human reviewers. Finally, at least 2 independent authors reviewed each output from ChatGPT to determine the final political alignment, and their agreement was highly correlated, indicating high reliability in our classification.

For limitations, we evaluated policy statements from only 6 medical organizations; thus, our results may not be generalizable to other professional groups. Second, we were unable to analyze the trend in the political alignment of policy statements over time. Many policy statements were revisions (or updates) of established policy statements, and the original text was unavailable. Third, we used a single prompt for all ChatGPT sessions and used the first ChatGPT response for our analysis. ChatGPT responses can be variable depending on the text used in each session and can change with repeated, identical queries

XSL•FO
RenderX

[36]. Thus, our results are limited to a single prompt and do not account for variability in ChatGPT's output; however, using a single prompt and only the first ChatGPT output, we maintained a uniform process for all policy statement evaluations. Fourth, most policy statements were long documents that discussed various ideas and covered many topics. It is uncertain if ChatGPT considered all points in each policy statement equally or selectively focused on a few. If the latter is true, then our results may not represent the entire scope of each policy statement. However, there is no reason to suspect that ChatGPT was selective in its analysis. Moreover, some policy statements were read in full by the authors after the ChatGPT query, and the ChatGPT outputs were representative of the policy statements. Finally, some reports have indicated that ChatGPT has a left-leaning bias, which may have influenced our results. However, when ChatGPT's output was compared to an alternative chatbot, Grok, of all the statements categorized as liberal by ChatGPT, Grok agreed 95% of the time. Although it is possible that Grok is also politically biased, it is unlikely that separate chatbots with different developers would produce almost identical results. Moreover, the secondary analysis showed that ChatGPT and Grok interpreted many more policy statements to be aligned with a liberal ideology than in the main analysis. Thus, we are more confident that political bias in our study was not overestimated.

Conclusions

One-third of the policy statements from 6 medical organizations champion a political view. When this occurs, the organization's statement is 40 times more likely to espouse a liberal rather than conservative viewpoint. Policy may inherently be political in nature, but our study shows a strong preference for liberal viewpoints in medical organizations. Notably, many of these organizations explicitly state that they are nonpartisan. It remains to be explored whether these positions primarily reflect alignment with evidence-based policies or instead indicate a lack of viewpoint diversity within the organizations.

Disclaimer

This paper was written when VP was a University of California, San Francisco employee and before service at the US Food and Drug Administration (FDA). The views do not reflect those of the US FDA or the US government. This work was supported by ORMJMP Foundation.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files. The ChatGPT output for all policy statements analyzed in this study is available in Multimedia Appendix 1.

Authors' Contributions

BK and VP conceptualized the study design. BK obtained and abstracted data. BK and AM reviewed and confirmed data. BK performed data analysis. BK and AH performed statistical analysis. BK wrote the first draft of the manuscript. PC contributed to the discussion. BK, AM, PC, AH, and VP reviewed subsequent and final versions of the manuscript.

Conflicts of Interest

VP receives research funding from Arnold Ventures through a grant made to the University of California, San Francisco and royalties for books and writing from Johns Hopkins Press, MedPage, and the Free Press. VP declares consultancy roles with UnitedHealthcare and OptumRX; hosts the podcasts Plenary Session, VPZD, and Sensible Medicine; writes the newsletters Sensible Medicine, the Drug Development Letter, and VP's Observations and Thoughts; and runs the YouTube channel Vinay Prasad MD MPH, which collectively earns revenue on the platforms Patreon, YouTube, and Substack. The other authors report no conflicts of interest.

Multimedia Appendix 1
Full ChatGPT output for all policy statements analyzed.
[PDF File (Adobe PDF File), 4047 KB - formative_v9i1e66204_app1.pdf]

Multimedia Appendix 2
Examples of part of the ChatGPT output for liberal and conservative policy statements.
[DOCX File , 27 KB - formative_v9i1e66204_app2.docx]

References

1. AMA history. American Medical Association. URL: https://www.ama-assn.org/about/ama-history/ama-history [accessed 2024-03-15]
2. Antonelli M, Coopersmith CM. Professional medical societies: do we have any conflict of interest with industry? Intensive Care Med 2018 Oct 13;44(10):1762-1764. [doi: 10.1007/s00134-018-5304-8] [Medline: 30006897]

XSL•FO
RenderX

3. Dalsing MC. Industry working with physicians through professional medical associations. J Vasc Surg 2011 Sep;54(3 Suppl):41S-46S [FREE Full text] [doi: 10.1016/j.jvs.2011.04.068] [Medline: 21872115]
4. Marcotte L, Moriates C, Milstein A. Professional organizations' role in supporting physicians to improve value in health care. JAMA 2014 Jul 16;312(3):231-232. [doi: 10.1001/jama.2014.6762] [Medline: 24901503]
5. Health professionals lobbying profile. OpenSecrets. URL: https://www.opensecrets.org/federal-lobbying/industries/summary?cycle=2023&id=H01 [accessed 2024-03-20]
6. Health professionals summary. OpenSecrets. URL: https://www.opensecrets.org/industries//indus?ind=H01 [accessed 2024-03-20]
7. Stroebe W, vanDellen MR, Abakoumkin G, Lemay EP, Schiavone WM, Agostini M, et al. Politicization of COVID-19 health-protective behaviors in the United States: longitudinal and cross-national evidence. PLoS One 2021;16(10):e0256740 [FREE Full text] [doi: 10.1371/journal.pone.0256740] [Medline: 34669724]
8. Bolsen T, Palm R. Politicization and COVID-19 vaccine resistance in the U.S. Prog Mol Biol Transl Sci 2022;188(1):81-100 [FREE Full text] [doi: 10.1016/bs.pmbts.2021.10.002] [Medline: 35168748]
9. Prasad V, Haslam A. COVID-19 vaccines: history of the pandemic's great scientific success and flawed policy implementation. Monash Bioeth Rev 2024 Jun 09;42(1):28-54 [FREE Full text] [doi: 10.1007/s40592-024-00189-z] [Medline: 38459404]
10. Trust barometer trust and health. Edelman. URL: https://www.edelman.com/sites/g/files/aatuss191/files/2024-04/2024%20Edelman%20Trust%20Barometer%20Trust%20and%20Health%20FINAL.pdf [accessed 2024-06-23]
11. Sinsky CA, Trockel M, Carlasare LE, West CP, Wang H, Tutty M, et al. Politicization of medical care, burnout, and professionally conflicting emotions among physicians during COVID-19. Mayo Clin Proc 2023 Nov;98(11):1613-1628 [FREE Full text] [doi: 10.1016/j.mayocp.2023.07.007] [Medline: 37923520]
12. Southwick R. AMA chief: stop the harassment and threats against doctors. MedicalEconomics. URL: https://www.medicaleconomics.com/view/ama-chief-stop-the-harassment-and-threats-against-doctors [accessed 2024-03-12]
13. American Academy of Pediatrics (AAP). InfluenceWatch. URL: https://www.influencewatch.org/non-profit/american-academy-of-pediatrics/#:~:text=political%20advocacy%20campaigns.-,Political%20Activity,health%20issues%2C%20and%20gun%20control [accessed 2024-03-10]
14. Stein F. AAP statement on protecting immigrant children. American Academy of Pediatrics. URL: https://www.aap.org/en/news-room/news-releases/aap/2017/aap-statement-on-protecting-immigrant-children/ [accessed 2024-03-10]
15. ACSPA-SurgeonsPAC. American College of Surgeons. URL: https://www.facs.org/advocacy/get-involved/surgeonspac/ [accessed 2024-03-10]
16. Doherty RB. Why ACP engages in politics, without being partisan. I.M. Matters from ACP. URL: https://immattersacp.org/archives/2021/07/why-acp-engages-in-politics-without-being-partisan.htm [accessed 2024-03-15]
17. American College of Obstetricians and Gynecologists. Elections. ACOG. URL: https://www.acog.org/advocacy/get-involved/elections#:~:text=ACOG%20is%20a%20nonpartisan%20organization,possible%20in%20preparation%20for%20elections
18. Collins MC. Is universal coverage the American way? Fam Pract Manag 2000;7(7):33-38 [FREE Full text] [Medline: 11010609]
19. Lynch JF. Medical associations as social partners in support of democracy. Ann Am Thorac Soc 2021 Jul;18(7):1106-1108 [FREE Full text] [doi: 10.1513/AnnalsATS.202101-030IP] [Medline: 33448902]
20. Global climate change and children's health: AAP recommendations. American Academy of Pediatrics. URL: https://tinyurl.com/ms5krwrd [accessed 2024-03-15]
21. Position statement on mental health and climate change. American Psychiatric Association. URL: https://www.psychiatry.org/getattachment/0ce71f37-61a6-44d0-8fcd-c752b7e935fd/Position-Mental-Health-Climate-Change.pdf
22. Crowley R, Mathew S, Hilden D, Mathew S, Hilden D, Beachy M, et al. Environmental health: a position paper from the American College of Physicians. Ann Intern Med 2022 Nov;175(11):1591-1593. [doi: 10.7326/m22-1864]
23. Parker CL. Slowing global warming: benefits for patients and the planet. American Family Physician. URL: https://www.aafp.org/pubs/afp/issues/2011/0801/p271.html [accessed 2024-03-20]
24. Addressing climate change. American College of Obstetricians and Gynecologists. URL: https://www.acog.org/clinical-information/policy-and-position-statements/position-statements/2021/addressing-climate-change [accessed 2024-03-20]
25. Statement on healthcare reform. American College of Surgeons. 2017 May 01. URL: https://www.facs.org/about-acs/statements/health-care-reform/ [accessed 2024-03-19]
26. Position statement on universal health care in the United States. American Psychiatric Association. URL: https://www.psychiatry.org/getattachment/00b3d511-064e-4675-908a-bcceb0a27cae/Position-Universal-Health-Care.pdf [accessed 2024-03-25]
27. Crowley R, Daniel H, Cooney TG, Engel LS. Envisioning a better U.S. health care system for all: coverage and cost of care. Ann Intern Med 2020 Jan 21;172(2_Supplement):S7. [doi: 10.7326/m19-2415]
28. Home page. American Academy of Pediatrics. URL: https://publications.aap.org/pediatrics/pages/policy?autologincheck=redirected [accessed 2025-05-29]
29. ACS quality and safety conference. American College of Surgeons. URL: https://www.facs.org/ [accessed 2025-05-29]
30. Home page. American Psychiatric Association. URL: https://www.psychiatry.org/ [accessed 2025-05-29]

XSL•FO
RenderX

31. Get on-demand access to ACSM content. American College of Obstetricians and Gynecologists. URL: https://www.acog.org/ [accessed 2025-05-29]
32. Where we stand. American College of Physicians. URL: https://www.acponline.org/advocacy/where-we-stand [accessed 2025-05-29]
33. Policies. American Academy of Family Physicians. URL: https://www.aafp.org/about/policies.html
34. GPT-4. OpenAI. URL: https://openai.com/index/gpt-4-research [accessed 2024-06-23]
35. Gande S, Gould M, Ganti L. Bibliometric analysis of ChatGPT in medicine. Int J Emerg Med 2024 Apr 04;17(1):50. [doi: 10.1186/s12245-024-00624-2] [Medline: 38575866]
36. Motoki F, Pinho Neto V, Rodrigues V. More human than human: measuring ChatGPT political bias. Public Choice 2023 Aug 17;198(1-2):3-23 [FREE Full text] [doi: 10.1007/S11127-023-01097-2]
37. Choudhary T. Political bias in large language models: a comparative analysis of ChatGPT-4, Perplexity, Google Gemini, and Claude. IEEE Access 2025;13:11341-11379. [doi: 10.1109/access.2024.3523764]
38. Fujimoto S, Takemoto K. Revisiting the political biases of ChatGPT. Front Artif Intell 2023 Oct 20;6:1232003 [FREE Full text] [doi: 10.3389/frai.2023.1232003] [Medline: 37928447]
39. Webber E, Rutinowski J, Jost N, Pauly M. Is GPT-4 less politically biased than GPT-3.5? A renewed investigation of ChatGPT's political biases. arXiv. preprint posted online October 28, 2024 [FREE Full text]
40. Bonica A, Rosenthal H, Rothman DJ. The political polarization of physicians in the United States: an analysis of campaign contributions to federal elections, 1991 through 2012. JAMA Intern Med 2014 Aug 01;174(8):1308-1317. [doi: 10.1001/jamainternmed.2014.2105] [Medline: 24887456]
41. Sanger-katz M. Your surgeon is probably a Republican, your psychiatrist probably a Democrat. The New York Times. URL: https://www.nytimes.com/2016/10/07/upshot/your-surgeon-is-probably-a-republican-your-psychiatrist-probably-a-democrat.html [accessed 2024-06-23]
42. VanderWeele TJ. Viewpoint diversity in public health. Front Public Health 2023 Aug 31;11:1263767 [FREE Full text] [doi: 10.3389/fpubh.2023.1263767] [Medline: 37719726]
43. Wilkerson JD, Carrell D. Money, politics, and medicine: the American Medical PAC's strategy of giving in U.S. house races. J Health Polit Policy Law 1999 Apr;24(2):335-355. [doi: 10.1215/03616878-24-2-335] [Medline: 10321360]
44. LaValley MN, Eble DJ, Kuzon Jr WM, Peters BR, Skokan AJ, Ettinger RE, et al. Making a statement: positions of professional medical organizations towards gender-affirming care. Ann Surg 2024 May 10:2024. [doi: 10.1097/SLA.0000000000006342] [Medline: 38726665]
45. Bush T, Jensen WA, Katsumoto TR. U.S. medical organizations and climate change advocacy: a review of public facing websites. BMC Public Health 2022 Oct 21;22(1):1950 [FREE Full text] [doi: 10.1186/s12889-022-14339-7] [Medline: 36271371]
46. Blendon RJ, Benson JM. Trust in medicine, the health system and public health. Daedalus 2022;151(4):67-82. [doi: 10.1162/daed_a_01944]
47. Kennedy B, Tyson A, Funk C. Americans' trust in scientists, other groups declines. Pew Research Center. URL: https://www.pewresearch.org/science/2022/02/15/americans-trust-in-scientists-other-groups-declines/ [accessed 2024-06-23]
48. Borenstein S, Fingerhut H. Americans' trust in science now deeply polarized, poll shows. AP News. URL: https://apnews.com/article/coronavirus-pandemic-science-health-covid-19-pandemic-4e99139d995581319dffab4107627a5e [accessed 2024-06-23]

Abbreviations

AAFP: American Academy of Family Physicians
AAP: American Academy of Pediatrics
ACOG: American College of Obstetricians and Gynecologists
ACP: American College of Physicians
ACS: American College of Surgeons
APA: American Psychiatric Association
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

XSL•FO
RenderX

Edited by A Mavragani; submitted 06.09.24; peer-reviewed by T Akindahunsi, S Fitzek, S Mukerjee, Y Sarikhani; comments to author 24.11.24; revised version received 16.02.25; accepted 30.04.25; published 13.06.25.

Please cite as:
Knudsen B, Madkour A, Cholli P, Haslam A, Prasad V
Analysis of the Political Viewpoint of Policy Statements From Professional Medical Organizations Using ChatGPT With GPT-4: Cross-Sectional Study
JMIR Form Res 2025;9:e66204
URL: https://formative.jmir.org/2025/1/e66204
doi:10.2196/66204
PMID:

©Ben Knudsen, Amr Madkour, Preetam Cholli, Alyson Haslam, Vinay Prasad. Originally published in JMIR Formative Research (https://formative.jmir.org), 13.06.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Use of Technology by the Rural-Dwelling Caregivers of People Living With Dementia to Support Caregiving: Qualitative Interview Study

Anna Jolliff[1], MS, LMHC; Sarah Boucher[2]; Jordan R Hill[2], PhD; Kristen Allen-Watts[3], MPH, PhD; Miriam J Rodriguez[2], PhD; Christian Elliott[4], BSc; Matthew Zuraw[4], MBA; Nicole E Werner[1], PhD

[1]Center for Research and Innovation in Systems Safety, Department of Anesthesiology, Vanderbilt University Medical Center, Nashville, TN, United States

[2]Health & Wellness Design, Indiana University School of Public Health - Bloomington, Bloomington, IN, United States

[3]The Division of General Internal Medicine and Population Science, School of Medicine, University of Birmingham, Birmingham, AL, United States

[4]CareVirtue Technologies, Inc, San Diego, CA, United States

Corresponding Author:
Anna Jolliff, MS, LMHC
Center for Research and Innovation in Systems Safety
Department of Anesthesiology
Vanderbilt University Medical Center
2525 West End Avenue
Suite 1200
Nashville, TN, 37203
United States
Phone: 1 615 936 598
Email: anna.jolliff@vumc.org

Abstract

Background: Family caregivers of people living with dementia who live in rural areas face challenges distinctly different from those experienced by their urban counterparts. Technology is a promising but underused method for delivering interventions to rural caregivers.

Objective: This study aims to describe the technologies used by rural caregivers of people living with dementia, how these caregivers use technology to support caregiving, and the barriers and facilitators they encounter in using technology.

Methods: We conducted online, semistructured interviews with rural caregivers of people living with dementia. The objective of the primary study was to understand how caregivers access support, including the types of support used, strategies used to find support, and any unmet support needs. This study was a secondary analysis focused exclusively on caregivers' technology use. Summative, content, and thematic analyses were used to understand patterns of technology use and the associated facilitators.

Results: A total of 19 caregivers were included in this study, of whom 14 (74%) were female. The average age was 66.4 (SD 10.1) years. The 5 most frequently endorsed technologies were phones for calling (n=19, 100%) and texting (n=14, 74%), websites (n=17, 89%), television and movies (n=15, 79%), and emails (n=13, 68%). Technology provided the following types of support: informational (n=19, 100%), emotional (n=13, 68%), instrumental (n=12, 63%), entertainment (n=10, 53%), safety (n=6, 32%), and caregiver personal health (n=4, 21%). Thematic analysis yielded 4 characteristics of caregivers that facilitated technology use, including access to technology, technological savviness, preference for technology, and having a technology broker. In addition, analyses yielded 4 characteristics of technology that facilitated technology use, including appeal, efficiency, ease of use, and trustworthiness.

Conclusions: All rural caregivers in this study used technology to access caregiving support, which suggests that this population is prepared for remote interventions. Our findings can be used to help determine optimal delivery and content of these interventions and instruct interventionists on caregiver and technology characteristics that may facilitate uptake. Rural caregivers may be receptive to interventions delivered through texts, websites, and videos, particularly those that focus on meeting informational and instrumental needs.

(JMIR Form Res 2025;9:e77231) doi:10.2196/77231

XSL•FO
RenderX

KEYWORDS

family caregivers; informal caregivers; rural populations; rural health; digital health; information technology; qualitative research

Introduction

Background

In the United States, approximately 12 million Americans care for someone living with dementia, totaling 19 billion hours of care annually [1]. While this care is integral to the well-being of people living with dementia, it has consequences for family caregivers. Family caregivers face numerous social, emotional, physical, and financial challenges [1]. In rural areas, family caregivers contend with challenges different from those of their urban counterparts, such as difficulty finding transportation to medical appointments, lower health literacy, fewer available caregiving resources, and less awareness of existing resources [2]. It is imperative to develop interventions uniquely suited for and accessible by this population of rural caregivers.

One promising method of delivering interventions to rural caregivers is through technology. Recent studies on the use of information and communication technologies by older adult caregivers living in rural areas have demonstrated numerous benefits of technology use, including improved access to services, enhanced social support, and a reduced sense of isolation [3,4]. However, technology-based interventions for rural caregivers of people living with dementia are scarce. One recent review of 11 studies on health interventions for rural caregivers of people living with dementia found that less than half used technology and, of those that did use it, only 2 interventions leveraged technology that was more advanced than a phone call (one used online education and another video recordings) [5]. Similarly, a recent systematic review of 30 technology-based interventions for rural family caregivers of people living with dementia found that most interventions were delivered via the phone [6]. The extent to which other forms of technology can be leveraged to support caregiving by rural caregivers of people living with dementia is yet unknown.

Understanding current patterns of technology use is a necessary first step toward designing useful digital interventions. To date, there is little research describing the technology use patterns and preferences of rural caregivers of people living with dementia. It has been established that older adults' use of technology is on the rise, with ownership of smartphones, tablets, smart televisions, home assistants, and wearables increasing among people aged between 60 and 69 years [7]. Furthermore, research suggests that older adults in rural areas can use technology to obtain support, including emotional, informational, and instrumental support [8-10]. However, technology use in rural areas may be threatened by the lack of access to high-speed internet and cost [7]. In addition, caregivers of people living with dementia may face barriers to technology use above and beyond those of other rural adults, including more demands on their time and less support in using technology [11].

In sum, there is reason to think that technology-based interventions could meet the many challenges faced by rural caregivers of people living with dementia. However, research is needed to understand viable technology-based methods of delivery as well as barriers to and facilitators of using technology to access support within this population.

This Study

This study asked the following research questions:

1. What technologies do rural caregivers describe using to support caregiving?
2. For what supportive purposes do rural caregivers describe using technology?
3. What are the barriers and facilitators described by rural caregivers when using technology for support?

Methods

Study Design

We conducted remote, semistructured interviews with 19 rural caregivers of people living with dementia from November 2022 to January 2023. The objective of the primary study was to understand how caregivers access support, including the types of support used, strategies used to find support, and any unmet support needs. The development of the interview guide was informed by this study's objective. An example question was, "Can you tell me what services and resources you know about in your community?" The interviewer would then probe to understand the process of finding out about these resources, including who helped the caregiver and any technologies used. Questions were initially drafted by the research team and then refined based on input from the study's strategic advisory board. The strategic advisory board comprised 7 professionals who worked with caregivers of people living with Alzheimer disease or related dementias who lived in rural areas. Members were selected to ensure diverse representation in terms of geographic region, resource provision, roles, and expertise. This study was a secondary analysis that used qualitative methods to describe caregivers' use of technology to access support.

Ethical Considerations

Ethics approval for this study was granted by the Indiana University Bloomington Institutional Review Board. The board considered this study to involve minimal risk to human participants and thus granted expedited review (16214). A waiver of signed consent was approved. Participants received a study information sheet via email that outlined the risks and benefits of participation; then, they provided verbal consent to participate during a screening call. The study information sheet noted that the researchers might use participants' interview data for secondary analyses without obtaining additional consent. Participants were encouraged to avoid the use of names or identifying information in interviews. When identifying information was shared, it was redacted from the transcript such that the final dataset was deidentified. Participants who completed the interview received a US $50 electronic gift card, which was emailed to them within 1 week of completing the interview.

XSL•FO
RenderX

Setting and Participants

We used purposive and network sampling to recruit rural caregivers of people living with dementia in collaboration with community organizations serving rural areas, such as area agencies on aging, as well as in partnership with our strategic advisory board. Representatives from agencies and strategic advisors were invited to share the research opportunity electronically and in person. Recruitment contexts included digital newsletters and emails as well as organizations' electronic and in-person support group meetings and education sessions. The inclusion criteria included being an unpaid caregiver of someone living with dementia, identifying their geographic area as rural, being aged at least 18 years, and having access to a phone to conduct the interview. Interested participants emailed or called the research team. A research team member scheduled the enrollment visit phone call and emailed the participant a copy of the study information sheet. During the enrollment call, participants were briefed on the study and invited to ask questions. They then provided verbal informed consent and scheduled the interview. All interviews took place over the phone or Zoom (Zoom Communications, Inc), according to the caregiver's preference. Participants had no relationship with researchers before the study commencement. No participants dropped out after enrollment.

Study Procedure

Interviews lasted for 60 to 90 minutes. Each interview was attended by a lead facilitator (AJ or NEW, both of whom identify as women) and a notetaker (JRH or MZ, who identify as a woman and a man, respectively). At the beginning of the interview, the interviewer and notetaker briefly shared with the participant their own professional and personal interest in the project. Participants completed an electronic survey in which they provided demographic details. All interviews were audio recorded and transcribed for analysis.

Analysis

All analyses took place in NVivo (version 14; Lumivero). To answer research question 1, "What technologies do rural caregivers describe using to support caregiving?" we completed a summative analysis in which we documented each type of technology that caregivers used [12]. Transcripts were first reviewed line by line, and each passage describing technology use was extracted. These passages were then coded for the types of technology used. For example, if a caregiver described attending a support group over Zoom, it was coded that they used Zoom, and if they asked for help in a group message, it was coded that they used the text messaging app on their phone. Two coders (AJ and SB) coded a 20% subsample of transcripts to ensure consistency in coding. The remaining 14 transcripts were coded by either one of the coders.

To answer research question 2, "For what supportive purposes do rural caregivers describe using technology?" we performed a content analysis that combined deductive and inductive approaches [12]. As with the first research question, all passages in which technology use was described were extracted and coded. The initial codebook included 3 common types of support described in the caregiving literature: emotional, instrumental, and informational [13,14]. Two coders identified all passages in which caregivers described using technology for support and then coded these segments for the type of support sought. For example, if a participant described going online to find information about local resources for caregivers, this was coded as "informational support," whereas if they texted a friend to request respite care, this was coded as "instrumental support." Again, 2 coders (AJ and SB) coded a 20% subsample of transcripts to ensure consistency in coding. Coders made a note of any questions or challenges encountered during coding and discussed these in weekly coding meetings. Questions that could not be resolved between the two coders were brought to the whole research team for consensus-based discussion. Types of support that did not fit into the existing coding scheme were assigned a new code. The final codebook can be found in the Results section.

To answer research question 3, we used reflexive thematic analysis guided by the following research question: "What are the barriers and facilitators experienced by rural dementia caregivers when using technology for support?" We used this approach because we did not expect to identify a single "correct" interpretation of the data but rather an interpretation that was the product of coders' perspectives [15]. Two coders (AJ and SB) separately read each transcript line by line and identified passages in which barriers to and facilitators of technology use seemed to have been described. Coders individually assigned a name to the preliminary barrier or facilitator. At weekly intervals, coders met to discuss the codes that they had created based on the passages they had identified. A codebook was created and iterated upon to reflect the coders' evolving understanding of the data. After 50% of the transcripts had been coded, 3 additional members of the research team (NEW, MJR, and KA-W) coded a transcript each using the codebook in progress and discussed the extent to which they felt the codebook reflected their interpretation of the data. The goal of this discussion was not consensus but rather crystallization, that is, a multifaceted perspective on the data based on the reflections of multiple coders [16].

Data interpretation was shaped by the analysis team members' perspectives. The analysis team included researchers with doctoral degrees in health education and promotion (KA-W), clinical neuropsychology (MJR), and human factors and applied cognition (NEW); a research coordinator with a master's degree in counseling psychology (AJ); and an undergraduate student studying community health (SB). Data analysts' training inevitably shaped our impressions of the data [15]. This is particularly true for research questions 2 and 3, for which data analysis required more interpretation. Our recognition and categorization of support seeking and technology use as well as our understanding of barriers to and facilitators of use were interpretations based on our professional and, to some extent, personal experiences. Interpretation was further required because this study was a secondary analysis; in the interviews, we did not explicitly ask all the research questions that we later set out to answer through secondary analysis.

XSL•FO
RenderX

Results

Demographic Information

Of the 19 caregivers, 14 (74%) were female. Caregivers' average age was 66.4 (SD 10.1) years. Within the sample, 11 (58%) caregivers cared for a partner, 6 (32%) cared for a parent or parent-in-law, 1 (5%) cared for a child, and 1 (5%) cared for a friend. Table 1 presents complete participant demographic information.

Table 1. Demographic information for a sample of rural caregivers of people living with dementia (N=19).

Characteristic	Values
Age (y), mean (SD)	66.4 (10.1)
Gender[a], n (%)	
Man	5 (26)
Woman	14 (74)
Race and ethnicity[a], n (%)	
American Indian or Alaska Native	1 (5)
Non-Hispanic White	18 (95)
State, n (%)	
South Dakota	13 (68)
Wisconsin	3 (16)
California	2 (11)
Washington	1 (5)
Education, n (%)	
Graduate or professional degree	6 (32)
Bachelor's or associate's degree	10 (53)
Some college or no degree	2 (11)
High school diploma or GED[b]	1 (5)
Employment, n (%)	
Working full time	4 (21)
Working part time	1 (5)
Retired	9 (47)
Other	5 (26)
Income (US $), n (%)	
<25,000	3 (16)
25,000-49,999	2 (11)
50,000-74,999	8 (42)
75,000-99,999	1 (5)
100,000-149,999	3 (16)
Prefer not to say	2 (11)

[a]Only 2 response options were selected by participants.

[b]GED: general educational development.

Technologies Used to Support Caregiving

The 5 most frequently mentioned technologies by the 19 caregivers were phones for calling (n=19, 100%) and texting (n=14, 74%), websites (n=17, 89%), television and movies (n=15, 79%), and emails (n=13, 68%). Table 2 presents a comprehensive list of the technologies mentioned by 1 or more caregivers.

XSL•FO
RenderX

Table 2. Technologies used by rural caregivers of people living with dementia to support caregiving (N=19).

Type of technology	Caregivers, n (%)
Phone	
Call app	19 (100)
Text app (individual or group)	14 (74)
Other app	3 (16)
Website	17 (89)
Television and movies	15 (79)
Email	13 (68)
Videoconferencing software	
Zoom or webinar	11 (58)
Telehealth	4 (21)
View clicks (video calling)	1 (5)
Social media (including YouTube)	6 (32)
Audio (music and audiobook)	4 (21)
Surveillance system	
Camera	4 (21)
Door alarm	2 (11)
Motion-detecting lights	1 (5)
Sensor on the walker	1 (5)
Speaker	1 (5)
Call button	1 (5)
Smart watch	3 (16)
Videogames	3 (16)
Tablet	2 (11)
Personal assistant	1 (5)
Computer application	1 (5)
Digital pill dispenser	1 (5)

Purposes of Technology Use

In order of frequency, technology use was described for the following purposes by the 19 caregivers: informational support (n=19, 100%), emotional support (n=13, 68%), instrumental support (n=12, 63%), entertainment (n=10, 53%), safety (n=6, 32%), and caregiver personal health (n=4, 21%). The purposes of entertainment, safety, and caregiver personal health were not part of our initial coding framework but were arrived at inductively. Table 3 presents examples of each type of support and the percentage of caregivers who described using technology for that type of support.

XSL•FO
RenderX

Table 3. Types of technology-based support obtained by rural caregivers of people living with dementia (N=19).

Type of support	Definition	Examples from transcripts	Caregivers, n (%)
Informational support	Technology to facilitate knowledge sharing	Researching aging or dementia online, participating in live webinars about dementia care, receiving email newsletters from area agencies on aging, and requesting information in online support groups	19 (100)
Emotional support	Technology to facilitate the expression of empathy or caring	Commiserating in support groups, calling church members with prayer requests, expressions of caring from Facebook friends, and texting other caregivers to ask how they are doing	13 (68)
Instrumental support	Technology to facilitate the provision of tangible aid	Texting family or friends to request respite care, going online to find an attorney, calling a friend to help when the care recipient has had a fall, calling government agencies to request services, and watching instructional YouTube videos	12 (63)
Entertainment	Technology to amuse or divert the care recipient	Recording care recipient's favorite television shows, allowing friends and family to video call care recipient, and care recipient playing video games alone or with others	10 (53)
Safety	Technology for the protection of the care recipient	Sensor to alert caregiver when care recipient gets out of bed at night, smartwatch to enable care recipient to easily call recipient, cameras to monitor care recipient when away from home, door alarms to know if care recipient is leaving, and checking on care recipient's safety over text	6 (32)
Caregiver personal health	Technology to facilitate the physical or emotional well-being of the caregiver	Listening to a podcast or audiobook while driving the care recipient, caregiver monitoring their own sleep with a smartwatch, and taking an online weight loss class	4 (21)

Barriers to and Facilitators of Technology Use

Through reflexive thematic analysis, we perceived 8 factors that could serve as a facilitator of or a barrier to technology use for rural caregivers. We divided these themes into 2 imperfect categories: characteristics of caregivers and characteristics of technology. We perceived the following characteristics of caregivers: access to technology, technological savviness, preference for technology, and access to a technology broker. Furthermore, we perceived the following characteristics of technology: appeal, efficiency, ease of use, and trustworthiness. These categories are imperfect because, for example, trustworthy technology could also reflect a trusting caregiver, while a caregiver's ability to access technology could reflect an accessible technology. However, these categories were valuable in that they facilitated the organization of ideas. Multimedia Appendix 1 presents definitions of each thematic barrier or facilitator and example quotations from participants.

Discussion

Principal Findings

In this study, we sought to explore the technology use of rural caregivers of people living with dementia, including the technologies they described using, the supportive purposes for which they used technology, and any factors that seemed to serve as a barrier to or facilitator of use. Our findings suggest that rural caregivers' technology use goes well beyond making phone calls and watching television to include texting, visiting websites, emailing, videoconferencing, social media use, and more. Every caregiver we interviewed used technology for informational support, and more than half of them used technology for emotional support (13/19, 68%), instrumental support (12/19, 63%), and entertainment (10/19, 53%). Thematic analyses yielded 8 factors that, when present, seemed to facilitate technology use. Taken together, these findings can be used to help determine the optimal delivery and content of technology-based interventions for the population of rural caregivers of people living with dementia.

In combination with existing research, our findings suggest that the current array of technology-based caregiver interventions may underestimate this population's level of technological engagement and ingenuity [17]. Currently, most technology-based interventions for caregivers, including rural caregivers, use phone calls as the technological medium [18]. In this study, while it was true that all 19 rural caregivers used phone calls, many also used websites (n=17, 89%), group or individual texting (n=14, 74%), emails (n=13, 68%), and videoconferencing (n=11, 58%). There are benefits to these technologies as mediums for intervention delivery. Unlike phone call–based interventions, web- and email-based interventions can be self-paced and contain more information than can be delivered via phone [18]. Web-based interventions can also be made more accessible than phone-based interventions by adhering to web content accessibility guidelines, which offer instruction on presenting content in multiple, customizable formats [19]. In text messaging–based interventions, contact is often more frequent than in phone call–based interventions; therefore, participants may feel more continuously supported [20]. Furthermore, the text message "thread" automatically creates a record of all resources shared with the caregiver [20]. Finally, interventions delivered via videoconferencing have the power to facilitate deeper interpersonal connections among caregivers who would otherwise be geographically isolated [21,22]. We therefore advocate for the development and testing of new technology-based interventions for rural caregivers of people living with dementia. We further advocate for using technologies that are already familiar to many caregivers (eg, websites, text messaging, and online support groups) to disseminate information about more advanced technologies with unique affordances, such as caregiver-focused communities

XSL•FO
RenderX

on TikTok [23] and virtual reality experiences [24] that educate dementia caregivers.

Previous research has reviewed the content of interventions for caregivers and has found that the content predominantly focuses on psychosocial outcomes, such as burden and self-efficacy [5]. However, there is a lack of interventions for rural caregivers that focus on providing the instrumental aid that is much needed in rural areas, such as addressing the effects of poverty and increasing access to health care services (eg, medical specialties and home nursing) [5]. In this study, 63% (12/19) of the caregivers used technology, such as web, text, email, and videoconferencing, to find instrumental assistance. Furthermore, all caregivers in this study used technology to find information. However, interventions for caregivers of people living with dementia that focus on education and information sharing are outnumbered by those focusing on psychosocial outcomes [25]. Our results support the development of technology-based interventions that provide instrumental and informational aid to rural caregivers of people living with dementia, with objectives such as finding long-term care; accessing local, state, and federal caregiver programs; and obtaining professional assistance with financial, legal, and health care planning that may be scant in rural areas [26]. It is also possible that smartphone apps and websites that are already commercially or freely available could meet caregivers' instrumental and informational needs. However, the sheer number of resources may overwhelm busy caregivers [27]. Our findings suggest that caregivers would benefit from clinical, community, and technological brokers (eg, medical professionals, older adult service providers, and user experience designers) who can screen existing resources on the metrics that mattered to caregivers in this study, such as trustworthiness and the ease of use.

Many of the facilitators of and barriers to technology use yielded by our analyses overlap with existing research on technology use among older adults. With both rural caregivers and the larger population of older adults, facilitative characteristics include having access to technology (including being able to afford it), technological savviness or knowledge, interest in technology, and the presence of "brokers" (family and friends) who can guide technology use [7,17,28,29]. For both groups, characteristics that act as barriers to technology use include the technology being untrustworthy and difficult to use [7]. This suggests that the existing efforts to improve technology design for older adults may generalize to rural caregivers [30]. However, to reach rural older adults, efforts to help older adults get online must be improved [31]. In our sample, all participants had access to home internet; however, national estimates of rural Americans' access to broadband internet at home range from 72% to 86% [32,33]. Interest in technology-based caregiving interventions is likely to increase as individuals from Generation X and Generation Y become caregivers, given that their youngest members (aged 45 and 29 y, respectively, as of 2025) are either digital natives themselves or report high engagement with technology [34].

Limitations

This study had some limitations. Although our sample was representative of dementia caregivers in that it was predominantly composed of women [35], it also disproportionately consisted of White individuals. Therefore, our results lack the critical perspective of caregivers in rural areas from other racial and ethnic groups. Second, in this study, we adopted an inclusive definition of rurality, considering caregivers to be rural if they self-identified as such. However, there are many distinct definitions of rurality; for example, in the United States, the US Census Bureau, the Centers for Disease Control and Prevention, and the United States Department of Agriculture all define rurality differently [36]. While we do not believe our operationalization of rurality to be a weakness of this study, it is important to keep it in mind while interpreting results [36]. Third, participating in a phone- or web-based interview was a condition of participation. This biased our sample toward caregivers who were more technologically connected. Future research should expand the in-person recruitment contexts to include more spaces in which rural caregivers may be found, such as religious organizations, medical offices, and grocery stores, and data collection contexts to include homes, libraries, and community centers. Finally, because this was a secondary analysis, our data collection procedure was not designed to understand exactly which technologies caregivers used or the breadth of their technology use. Therefore, participants may have had broader experience with technology than is reflected in our findings. Future quantitative research should adopt a more comprehensive measure of technology use and distribute it to larger samples of caregivers to further elucidate rural caregiver technology use.

Conclusions

Our findings underscore the need for interventions that leverage technologies, such as websites, text messaging, and videoconferencing, to support rural caregivers. Caregivers in our study already used these technologies to obtain many types of support, suggesting that technology-enabled interventions targeting common areas of need for rural caregivers would be well received by this population. To further understand the technology use patterns and preferences of rural caregivers, future research must prioritize in-person or otherwise offline strategies for recruiting and engaging these caregivers. The resulting interventions may be as varied and nuanced in their deployment of technology and support as rural caregivers are themselves.

Acknowledgments

The authors are grateful to the caregivers who completed interviews and the strategic advisory board members who informed their approach to this project.

XSL•FO
RenderX

Data Availability

The datasets generated or analyzed during this study are not publicly available due to the sensitivity of the data but are available from the corresponding author on reasonable request.

Authors' Contributions

AJ and NEW contributed to conceptualization. AJ, SB, MJR, KA-W, and NEW contributed to formal analysis. NEW, MZ, and CE contributed to funding acquisition. AJ, JRH, and MZ contributed to investigation. NEW contributed to methodology and supervision. AJ contributed to writing the original draft. AJ, JRH, SB, MJR, KA-W, and NEW contributed to reviewing and editing the original draft.

Conflicts of Interest

CE and MZ are co-founders of CareVirtue. All other authors declare no conflicts of interest.

Multimedia Appendix 1
Barriers to and facilitators of technology use among rural caregivers of people living with dementia.
[DOCX File , 19 KB - formative_v9i1e77231_app1.docx]

References

1. Alzheimer's Association. 2025 Alzheimer's disease facts and figures. Alzheimers Dement 2025 Apr 29;21(4):e70235. [doi: 10.1002/alz.70235]
2. Shuffler J, Lee K, Fields N, Graaf G, Cassidy J. Challenges experienced by rural informal caregivers of older adults in the United States: a scoping review. J Evid Based Soc Work (2019) 2023 Jul 04;20(4):520-535. [doi: 10.1080/26408066.2023.2183102] [Medline: 37330681]
3. Blusi M, Asplund K, Jong M. Older family carers in rural areas: experiences from using caregiver support services based on Information and Communication Technology (ICT). Eur J Ageing 2013 Sep 29;10(3):191-199 [FREE Full text] [doi: 10.1007/s10433-013-0260-1] [Medline: 28804294]
4. Lyu M, Zhao Q, Yang Y, Hao X, Qin Y, Li K. Benefits of and barriers to telehealth for the informal caregivers of elderly individuals in rural areas: a scoping review. Aust J Rural Health 2022 Aug 23;30(4):442-457. [doi: 10.1111/ajr.12869] [Medline: 35460580]
5. Braley T, Eisenhauer C, Rowland SA. Interventions for rural caregiver health: an integrative review. Res Gerontol Nurs 2022 Mar 30;15(2):101-108. [doi: 10.3928/19404921-20220204-01] [Medline: 35148209]
6. Ruggiano N, Brown EL, Li J, Scaccianoce M. Rural dementia caregivers and technology: what is the evidence? Res Gerontol Nurs 2018 Jul 01;11(4):216-224. [doi: 10.3928/19404921-20180628-04] [Medline: 30036405]
7. Kakulla B. Older adults embrace tech for entertainment and day-to-day living: 2022 tech trends and the 50-plus. AARP Research. 2021 Dec 06. URL: https://www.aarp.org/pri/topics/technology/internet-media-devices/2022-technology-trends-older-americans/ [accessed 2025-05-09]
8. Mehrotra A, Huskamp HA, Souza J, Uscher-Pines L, Rose S, Landon BE, et al. Rapid growth in mental health telemedicine use among rural medicare beneficiaries, wide variation across states. Health Aff (Millwood) 2017 May 01;36(5):909-917. [doi: 10.1377/hlthaff.2016.1461] [Medline: 28461359]
9. Azizi Z, Broadwin C, Islam S, Schenk J, Din N, Hernandez MF, et al. Digital health interventions for heart failure management in underserved rural areas of the United States: a systematic review of randomized trials. J Am Heart Assoc 2024 Jan 16;13(2). [doi: 10.1161/jaha.123.030956]
10. Batsis JA, Petersen CL, Clark MM, Cook SB, Kotz D, Gooding TL, et al. Feasibility and acceptability of a technology-based, rural weight management intervention in older adults with obesity. BMC Geriatr 2021 Jan 12;21:44 [FREE Full text] [doi: 10.1186/s12877-020-01978-x]
11. Atoyebi O, Eng JJ, Routhier F, Bird ML, Mortenson WB. A systematic review of systematic reviews of needs of family caregivers of older adults with dementia. Eur J Ageing 2022 Jan 13;19(3):381-396 [FREE Full text] [doi: 10.1007/s10433-021-00680-0] [Medline: 36052180]
12. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005 Nov 30;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
13. MaloneBeach EE, Zarit SH. Dimensions of social support and social conflict as predictors of caregiver depression. Int Psychogeriatr 1995 Mar 30;7(1):25-38 [FREE Full text] [doi: 10.1017/s1041610295001827] [Medline: 7579018]
14. Evans M, Donelle L, Hume-Loveland L. Social support and online postpartum depression discussion groups: a content analysis. Patient Educ Couns 2012 Jun 30;87(3):405-410. [doi: 10.1016/j.pec.2011.09.011] [Medline: 22019021]
15. Braun V, Clarke V. Toward good practice in thematic analysis: avoiding common problems and be(com)ing a knowing researcher. Int J Transgend Health 2022 Oct 25;24(1):1-6 [FREE Full text] [doi: 10.1080/26895269.2022.2129597] [Medline: 36713144]

XSL•FO
RenderX

16. Varpio L, Ajjawi R, Monrouxe LV, O'Brien BC, Rees CE. Shedding the cobra effect: problematising thematic emergence, triangulation, saturation and member checking. Med Educ 2017 Jan 30;51(1):40-50. [doi: 10.1111/medu.13124] [Medline: 27981658]
17. Block L, Gilmore-Bykovskyi A, Jolliff A, Mullen S, Werner NE. Exploring dementia family caregivers' everyday use and appraisal of technological supports. Geriatr Nurs 2020 Nov 30;41(6):909-915 [FREE Full text] [doi: 10.1016/j.gerinurse.2020.06.019] [Medline: 32665095]
18. Graven LJ, Glueckauf RL, Regal RA, Merbitz NK, Lustria ML, James BA. Telehealth interventions for family caregivers of persons with chronic health conditions: a systematic review of randomized controlled trials. Int J Telemed Appl 2021 May 21;2021:3518050 [FREE Full text] [doi: 10.1155/2021/3518050] [Medline: 34093704]
19. Web content accessibility guidelines (WCAG) 2.0. World Wide Web Consortium. 2021 May 21. URL: http://www.w3.org/TR/WCAG20/ [accessed 2025-08-21]
20. Perales-Puchalt J, Ramírez-Mantilla M, Fracachán-Cabrera M, Vidoni ED, Ellerbeck EF, Ramírez AS, et al. A text message intervention to support latino dementia family caregivers (CuidaTEXT): feasibility study. Clin Gerontol 2024 Jan 30;47(1):50-65. [doi: 10.1080/07317115.2022.2137449] [Medline: 36268684]
21. Holtz BE, Mitchell KM, Strand D, Hirko K. Perceptions of telehealth-based cancer support groups at a rural community oncology program. J Cancer Educ 2024 Aug 30;39(4):418-425. [doi: 10.1007/s13187-024-02428-7] [Medline: 38539005]
22. Banbury A, Nancarrow S, Dart J, Gray L, Parkinson L. Telehealth interventions delivering home-based support group videoconferencing: systematic review. J Med Internet Res 2018 Feb 02;20(2):e25 [FREE Full text] [doi: 10.2196/jmir.8090] [Medline: 29396387]
23. Wells K. On #dementia TikTok, family caregivers find support and bring the disease to light. npr. 2022 Oct 11. URL: https://www.npr.org/sections/health-shots/2022/10/11/1102436118/on-dementia-tiktok-family-caregivers-find-support-and-bring-the-disease-to-light [accessed 2025-08-06]
24. AFA's dementia virtual reality experience. Alzheimer's Foundation of America. URL: https://alzfdn.org/afas-dementia-experience/ [accessed 2025-08-06]
25. Huggins M, Pesut B, Puurveen G. Interventions for caregivers of older adults with dementia living in the community: a rapid review of reviews. Can J Aging 2023 Feb 17;42(3):425-433. [doi: 10.1017/s0714980823000016]
26. Soong A, Au ST, Kyaw BM, Theng YL, Tudor Car L. Information needs and information seeking behaviour of people with dementia and their non-professional caregivers: a scoping review. BMC Geriatr 2020 Feb 14;20(1):61 [FREE Full text] [doi: 10.1186/s12877-020-1454-y] [Medline: 32059648]
27. Ali S, Alizai H, Hagos DJ, Rubio SR, Calabia D, Serrano Jimenez P, et al. mHealth apps for dementia, Alzheimer disease, and other neurocognitive disorders: systematic search and environmental scan. JMIR Mhealth Uhealth 2024 Jul 03;12:e50186 [FREE Full text] [doi: 10.2196/50186] [Medline: 38959029]
28. Haase KR, Cosco T, Kervin L, Riadi I, O'Connell ME. Older adults' experiences with using technology for socialization during the COVID-19 pandemic: cross-sectional survey study. JMIR Aging 2021 Apr 23;4(2):e28010 [FREE Full text] [doi: 10.2196/28010] [Medline: 33739929]
29. Grewal KS, Gowda-Sookochoff R, Peacock S, Cammer A, McWilliams LA, Spiteri RJ, et al. Perspectives on technology use in the context of caregiving for persons with dementia: qualitative interview study. JMIR Form Res 2024 Dec 13;8:e63041 [FREE Full text] [doi: 10.2196/63041] [Medline: 39671589]
30. DeAngelis T. Optimizing tech for older adults. Monit Psychol 2021 Jul 01;52(5):54 [FREE Full text]
31. Aging connected: exposing the hidden connectivity crisis for older adults. Older Adults Technology Services. URL: https://oats.org/wp-content/uploads/2021/01/Aging-Connected-Exposing-the-Hidden-Connectivity-Crisis-for-Older-Adults.pdf [accessed 2025-04-28]
32. Vogels EA. Some digital divides persist between rural, urban and suburban America. Pew Research Center. 2021 Aug 09. URL: https://pewrsr.ch/3k5aU6J [accessed 2025-04-28]
33. Pender J, Goldstein J, Mahoney-Nair D, Charankevich H. Three USDA rural broadband programs: areas and populations served. U.S. Department of Agriculture. 2023 Oct 30. URL: https://search.nal.usda.gov/discovery/fulldisplay?context=L&vid=01NAL_INST:MAIN&docid=alma9916410631607426 [accessed 2025-08-20]
34. Gen X Crossroads: navigating digital innovation and traditional values. Media Culture. 2023 Nov 30. URL: https://www.mediaculture.com/insights/audience-insights-report-gen-x [accessed 2025-04-28]
35. Women and Alzheimer's. Alzheimer's Association. URL: https://www.alz.org/alzheimers-dementia/what-is-alzheimers/women-and-alzheimer-s [accessed 2025-04-28]
36. Long JC, Delamater PL, Holmes GM. Which definition of rurality should I use?: the relative performance of 8 federal rural definitions in identifying rural-urban disparities. Med Care 2021 Oct 01;59(Suppl 5):S413-S419 [FREE Full text] [doi: 10.1097/MLR.0000000000001612] [Medline: 34524237]

XSL•FO
RenderX

Edited by A Mavragani; submitted 04.06.25; peer-reviewed by C Jensen, C Turvey; comments to author 17.07.25; revised version received 06.08.25; accepted 11.08.25; published 24.09.25.

Please cite as:
Jolliff A, Boucher S, Hill JR, Allen-Watts K, Rodriguez MJ, Elliott C, Zuraw M, Werner NE
The Use of Technology by the Rural-Dwelling Caregivers of People Living With Dementia to Support Caregiving: Qualitative Interview Study
JMIR Form Res 2025;9:e77231
URL: https://formative.jmir.org/2025/1/e77231
doi:10.2196/77231
PMID:

©Anna Jolliff, Sarah Boucher, Jordan R Hill, Kristen Allen-Watts, Miriam J Rodriguez, Christian Elliott, Matthew Zuraw, Nicole E Werner. Originally published in JMIR Formative Research (https://formative.jmir.org), 24.09.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Medicine Beyond Machines: Viewpoint on the Art of Thinking in the Age of AI

Isabella B B Ferreira[1,2*], MD; Rodrigo C Menezes[2*], MD, PhD; Luis Cláudio L Correia[2*], MD, PhD; Bruno B Andrade[1,2*], MD, PhD

[1]Departamento de Pós-Graduação, Escola Bahiana de Medicina e Saúde Pública, Salvador, Brazil

[2]Laboratório de Pesquisa Clínica e Translacional, Instituto Gonçalo Moniz, Fundação Oswaldo Cruz, Rua Waldemar Falcao, 121, Candeal, Salvador, Brazil

[*]all authors contributed equally

Corresponding Author:
Bruno B Andrade, MD, PhD
Departamento de Pós-Graduação, Escola Bahiana de Medicina e Saúde Pública, Salvador, Brazil

Abstract

The widespread adoption of large language models is increasingly shaping clinical decision-making by altering how physicians engage with data and reasoning. While these tools enhance diagnostic capacity, streamline workflows, and support learning, their misuse may diminish critical, contextual, and humanized thinking, reducing physicians to passive validators of algorithmic outputs. This paper explores the evolution of medical cognition and proposes strategies for integrating artificial intelligence in ways that preserve cognitive autonomy, such as structuring information, reducing bias, and strengthening metacognition. We argue that artificial intelligence should serve as a "cognitive stethoscope," a tool that refines reasoning without compromising its essence.

(JMIR Form Res 2025;9:e76669) doi:10.2196/76669

KEYWORDS

artificial intelligence; AI; critical thinking; medicine; medical education

Introduction

Before the advent of antibiotics, anesthetics, or hard technologies, medicine was built on the pillars of care and curiosity. Over time, this innate curiosity became intertwined with innovation, driving continuous improvement. Technology has since expanded human capabilities far beyond biological limits, reconfiguring the way in which we diagnose, treat, and even think.

Now, we stand at the threshold of another technological leap: the incorporation of artificial intelligence (AI) into clinical practice, particularly through large language models (LLMs) [1]. These tools go beyond providing information; they analyze patterns, process data, and suggest diagnoses and therapeutic approaches, fundamentally reshaping the relationship between physicians and technology. In this viewpoint, we argue that, unlike earlier innovations, which extended physicians' senses without altering how they think, LLMs act directly on cognition itself. Used passively, they risk narrowing physicians' judgment; used deliberately, they may serve as a "cognitive stethoscope," refining clinical reasoning without supplanting it.

AI encompasses a broad field of systems that mimic human cognitive functions. Among its most influential branches is machine learning (ML), where algorithms learn from data to identify patterns and make predictions without explicit programming. A modern subset of ML, deep learning (DL), uses neural networks with multiple layers to detect complex, abstract representations, especially useful in tasks such as image or speech recognition. LLMs, in turn, are a specialized form of DL model trained on vast textual datasets, enabling applications such as clinical summarization, triage, and decision support [2].

The clinical potential of these tools is undeniable. However, their growing presence raises a fundamental question: to what extent does AI enhance medical cognition, and to what extent might it undermine it? Early studies suggest that, when integrated as a complement rather than a replacement, AI can improve diagnostic accuracy and efficiency without eroding independent reasoning [3]. Still, its impact depends less on raw capability than on how it is integrated into practice, an unsettled and increasingly urgent challenge.

However, medicine is more than a probabilistic exercise. Clinical decision-making is a complex, multidimensional art form that relies on judgment, inference, and human perception, aspects that algorithms have yet to replicate. In this paper, we explore the evolution of medical cognition and what distinguishes human thought from AI's role in clinical practice, focusing on AI's influence on diagnostic reasoning and therapeutic decision-making, while also considering its implications for education and workflow. Our goal is to orient

XSL•FO
RenderX

physicians and medical students toward a deliberate, critical engagement with AI.

Scientific Revolution and the Transformation of Medical Thinking

Before modern science, medicine was upheld by experimentation, beliefs, and individual observations. Diseases were attributed to miasmas or humoral imbalances, and treatments rested more on assumption than evidence [4]. Then, science began to unveil the invisible. In the 19th century, John Snow mapped London's cholera outbreaks, tracing the source to a contaminated water pump. His work demonstrated that patterns hidden in chaos could guide medical reasoning, laying the foundation for modern epidemiology [5].

With the discovery of pathogenic microorganisms, medicine shifted from speculation to specificity [6]. Diseases such as tuberculosis were no longer seen as a "tragic fate" but as conditions caused by identifiable agents such as *Mycobacterium tuberculosis* [7]. This marked a new era in which tools such as the stethoscope, laboratory assays, and imaging technologies expanded clinical perception and enhanced our senses, allowing clinicians to hear internal sounds, decode biochemical markers, and visualize organs and tissues.

As medicine evolved, new technologies and interventions were evaluated through rigorous scientific methods. Intuition gave way to evidence, to randomized clinical trials and systematic reviews, transforming care from trial and error into a practice grounded in probabilities and measurable outcomes. This shift produced landmark successes; for example, antiretroviral therapy dramatically reduced mortality among patients with AIDS. Initially facing a 5-year mortality rate of 80% to 90%, individuals living with HIV now approach the life expectancy of the general population [8].

At the same time, medicine has become more complex. Where once it was clear when to administer insulin, drain an abscess, or amputate a gangrenous limb, today's clinicians must navigate subtler trade-offs involving marginal benefits, long-term risks, and alignment with patients' preferences. Physicians must now integrate clinical history and physical examination with imaging, laboratory values, and probabilistic models, a cognitive load demanding both critical judgment and scientific literacy.

The pace of change has also accelerated. Whereas new knowledge once took an average of 17 years to reach textbooks and guidelines in the early 2000s, discoveries are now rapidly adopted into practice [9]. This acceleration creates a new challenge: managing the exponential growth of medical knowledge that now unfolds faster than the human capacity for absorption.

The Anatomy of Medical Thinking: From Gut Feelings to Organic Neural Networks

Unlike a purely logical or algorithmic process, clinical reasoning is dynamic, shaped by perception and inference. It develops in layers, each influenced by internal factors such as knowledge, bias, and emotion and external factors such as context, time, and access to resources. Even before a patient speaks, the clinical gaze captures unstructured information, such as how the patient enters the room or their posture when sitting. This first impression exists at the intersection of intuition and technique, where every detail, no matter how subtle, becomes a potential piece in the diagnostic puzzle.

Then, history taking unfolds as a guided dialogue. Each question refines emerging hypotheses, with the physician adjusting in real time to what is said and what is left unsaid [10,11]. This process goes beyond cataloging symptoms; it involves understanding how the patient perceives their condition within the narrative of their life. Because human behavior is complex and motivations are not binary, accounts may be exaggerated; minimized; or reshaped by fear, shame, or secondary gain. Therefore, the clinician must interpret language, silences, and inconsistencies critically, recognizing that each may alter the trajectory of reasoning.

After anamnesis and physical examination, the delicate process of synthesis begins. Physicians must balance 2 extremes: reasoning too broadly opens up endless possibilities, leading to irrelevant hypotheses, unnecessary tests, and unwarranted interventions, whereas narrowing too quickly risks dismissing critical diagnoses. Recognizing uncertainty is essential as biology is seldom exact. The medical mind operates within layers of probability and judgment, constantly reassessing hypotheses as new information emerges, considering both the plausibility of each possibility and the potential impact of diagnostic errors.

Refining this process requires metacognition: the capacity to monitor one's own thinking. Although fundamental, it is often neglected in clinical practice [12]. Biases are intrinsic to human reasoning. Anchoring traps thought in the first plausible diagnosis, availability exaggerates the weight of recently observed or common diseases, and confirmation bias narrows attention to evidence that supports an initial hunch [13]. Therefore, structured cognitive models and deliberate reflection are essential safeguards, helping physicians recognize and resist these tendencies.

Clinical decision-making is also shaped by forces beyond the individual. Time constraints, fatigue, and cognitive overload can favor speed over accuracy, impairing analytical depth, whereas clinical settings impose distinct demands: rapid stabilization in the emergency room or slower, narrative-rich deliberation in the clinic. In every case, physicians must balance evidence with the patient's context, preferences, and values, revealing that clinical reasoning is never a purely technical exercise but a human negotiation [14].

Thus, the greatest challenge to medical reasoning is not AI itself but the stagnation of cognitive training. Physicians are taught protocols and guidelines, which prescribe "what to do," but receive little formal training in "how to think" in the face of uncertainty [13]. As a result, medicine continues to rely on heuristics and experience even as modern complexity demands structured frameworks for risk interpretation, trade-off management, and adaptive reasoning.

XSL•FO
RenderX

The Impact of AI on Medical Thinking: Cognitive Expansion or Reduction?

AI has been part of medical practice for decades, from specialized diagnostic systems to ML algorithms that analyze vast clinical, genomic, and epidemiological datasets to uncover hidden patterns [15,16]. In radiology, DL techniques such as convolutional neural networks improve diagnostic accuracy by detecting subtle anomalies that may elude even experienced clinicians [17]. Meanwhile, ML-based models strengthen real-time disease surveillance by processing structured data streams such as case notifications and hospital admissions [18].

However, a turning point emerged with the widespread adoption of LLMs. Unlike earlier predictive algorithms that operated within rigid, preprogrammed parameters and required structural inputs, LLMs interact through natural language, making them accessible for daily clinical tasks such as case analysis and documentation [19]. This ease of use shifts AI from background support to direct influence on decision-making, a change that demands careful scrutiny [20].

Despite their enhanced capacities, LLMs do not "think" in the way humans do. In fact, LLMs recognize patterns, calculate probabilities, and predict the most statistically probable outcome based on previous data. While this can resemble clinical reasoning, it lacks causal inference; they do not ask "why" something happens [21]. Clinical reasoning, in contrast, is adaptive and prospective, integrating mechanisms, trade-offs, and patient narratives to generate and prioritize hypotheses.

This distinction is fundamental in medicine as a model cannot reassess its own assumptions or correct embedded biases. When training datasets contain distortions due to historical disparities, underrepresentation, or flawed methodologies, LLMs amplify these inaccuracies, presenting biased outputs as statistical truth. For instance, a widely used US algorithm underestimated the health needs of Black patients by equating health care cost with health status, thus favoring White patients for additional care [22]. Similarly, generative pretrained transformers have been shown to generate racially and gender-biased recommendations when presented with identical clinical cases, further deepening global health inequities [23].

Clinical reasoning is not just retrospective but adaptive, especially in resource-limited or culturally specific settings. Physicians adjust decisions as evidence and constraints evolve, whereas LLMs remain tied to preexisting data [24]. Consequently, AI trained on generalized datasets may miss local realities such as limited access to medications, diagnostic tools, and specialized care or cultural barriers. For instance, a model might recommend a colonoscopy for iron deficiency anemia in a rural area, overlooking prohibitive costs or reluctance toward invasive procedures. A clinician, aware of these factors, may instead pursue noninvasive tests or negotiate feasible next steps with the patient.

Another limitation is AI's inability to assimilate unrecorded or deliberately adjusted elements as clinical reasoning relies on subtle cues that often go unregistered in formal records [15]. Compounding this, LLMs often present responses with an unwarranted level of confidence, creating an illusion of precision, a phenomenon called "hallucination" [19]. In contrast, experienced physicians recognize that uncertainty is intrinsic to clinical reasoning, continuously refining hypotheses as new information emerges.

Recognizing these pitfalls, leading regulators have begun to shape "good practices" of AI use in health care. The European Union's AI Act (2024/1689) designates LLM-based clinical support tools as high-risk devices, mandating data governance, public model documentation, and continuous postmarket surveillance to catch adaptive drift [25]. In parallel, the US Food and Drug Administration's guidance principles require developers of AI-enabled medical software to specify in advance how and within what limits the model may evolve after approval and demonstrate that such updates will not erode safety or performance [26]. Both frameworks insist on (1) robust quality management to ensure the representativeness and traceability of the training data; (2) meaningful explainability so that clinicians and patients can understand a model's logic and limitations; and (3) empowering human oversight, signaling a shift toward emerging requirements for the safe and ethical deployment of clinical AI.

The Cognitive Stethoscope: How to Hear Across AI's Patterns

The integration of LLMs into medical practice is increasingly likely. However, their impact will depend less on technical capacity than on how physicians choose to engage with them. The danger is not simply that AI can make mistakes but that physicians may stop thinking critically, failing to question, refine, or override their suggestions when necessary. Nonetheless, when properly used, AI can strengthen medical reasoning, serving as a strategic tool to organize thought processes, reduce biases, and validate hypotheses while preserving human judgment (Table 1).

Consider a scenario of a patient with vague abdominal pain. A physician, facing the constraints we have already discussed, may lean toward a common diagnosis such as gastritis and might miss rare but serious conditions. An LLM, drawing from a broader context, could flag mesenteric ischemia as a possibility. While the model cannot determine causality, its input may prompt the physician to revisit the case; order a critical test; and, ultimately, save the patient from a missed diagnosis.

XSL•FO
RenderX

Table . Pitfalls and successes in physicians' uses of artificial intelligence (AI) for clinical reasoning.

Aspect	Pitfalls	Mitigation	Successes
Initial assessment	Neglect of clinical cognitive skills such as perception and inference due to premature reliance on AI	Emphasize clinical observation training and delay AI consultation	Enhances data collection by structuring initial patient information efficiently
Hypothesis formation	Overreliance on AI-generated differential diagnoses, potentially overlooking emergent situations or specific sociocultural contexts	Cross-check AI output with situational context and patient narrative	Expands differential diagnosis lists, reducing premature closure and anchoring bias
Decision-making	Passive acceptance of AI recommendations without critical evaluation	Use structured decision frameworks that require justification of AI suggestions	Provides up-to-date evidence and guideline-based recommendations to support decisions
Error handling	Overconfidence in AI outputs due to training data biases or lack of uncertainty calibration	Mandate local calibration studies and introduce uncertainty estimates in outputs	Identifies inconsistencies, flags potential errors, and reduces biases
Patient engagement	Risk of depersonalization by automating interactions with patients or overuse during consultation	Define "AI-free" phases in consultation to focus on human connection	Automates administrative tasks, allowing for more time for meaningful patient interactions
Longitudinal patient management	Fragmented decision-making due to AI operating on isolated encounters	Integrate AI tools with electronic health records to enable continuity-aware insights	Reveals longitudinal patterns that might be missed in episodic human reasoning
Clinical teaching and medical training	Reduced development of diagnostic intuition in trainees due to AI overuse	Balance AI use with mandatory non-AI problem-solving sessions in medical education	Functions as an adaptive learning tool, offering case-based scenarios, interactive feedback, and personalized study recommendations
Ethical decision-making	Inadequate integration of patient values or ethical nuances into AI-driven outputs	Incorporate shared decision-making tools and human review to contextualize ethical trade-offs	Summarizes ethical frameworks, guidelines, and case precedents to aid in ethical deliberation
AI in emergency and high-stakes situations	Delayed or misguided responses in emergencies due to insufficient data for novel cases	Establish protocols where human judgment overrides AI in high-stakes or novel settings	Enhances rapid triage and alerts in time-sensitive, high-risk clinical scenarios

In practice, AI should contribute to clinical reasoning in distinct yet complementary ways. It can rapidly analyze vast datasets to highlight potential diagnoses, suggest additional differential diagnoses, and flag atypical findings that might be overlooked. However, these benefits depend on integration. A recent study showed that an LLM outperformed physicians on a structured diagnostic task when used independently, but embedding it in real-time workflows did not improve diagnostic performance compared to conventional resources, likely due to suboptimal prompting, lack of training, or poor alignment with clinicians' existing habits [20].

Therefore, physicians remain central. They must deliberately decide which AI-generated insights warrant further pursuit, weighing severity, likelihood and the patient's context, narrative, and environment. Structured decision-making frameworks can support this process by prompting clinicians to actively justify or question AI suggestions. While algorithms can model probabilities and outcomes, only physicians can interpret them against ethical, contextual, and practical considerations.

While these cautions are essential, they should not obscure the settings in which AI may extend cognitive reach. In low-resource environments, where subspecialty expertise or advanced diagnostics are scarce, AI can broaden differential diagnoses and guide pragmatic next steps. In medical education, it can serve as a scaffold for reasoning, prompting students to articulate hypotheses, test alternatives, and reflect on biases. Beyond these domains, AI can also reinforce safety checks by flagging drug interactions, contraindications, or unintended side effects that warrant monitoring, such as postural hypotension after antihypertensive therapy, prompting proactive follow-up [27].

In ethically sensitive decisions, pairing AI recommendations with shared decision-making tools and explicit human oversight may ensure that care remains aligned with the values and expectations of each patient. Predefined override protocols can be established for situations in which AI may struggle, such as novel presentations or emergencies, reinforcing that human judgment is the final authority in clinical care.

AI can also relieve administrative burdens. AI-powered tools such as ambient documentation systems have shown potential to reduce documentation time for certain subgroups of clinicians, such as those in family medicine [28]. Although these systems have not yielded broad efficiency gains across all users in previous studies, they suggest that AI can meaningfully improve workflow in selected contexts, particularly when adopted with consistent engagement. When linked to longitudinal patient data, AI can further support continuity and coherence, helping decisions extend beyond episodic encounters. This frees time and cognitive bandwidth, allowing physicians to refocus on

XSL•FO
RenderX

what technology cannot replace: attentive listening, understanding, and the therapeutic relationship.

Far from replacing the human element, AI can be an ally in restoring the focus on what has always been essential in medicine: the connection between physician and patient. The real opportunity lies in integrating it strategically as a tool for critical reflection. Doubt must remain central to the physician-AI interaction as the physician of the future will not be the one who merely consults AI for answers but the one who knows when to trust it; when to challenge it; and, most importantly, when to think beyond it.

Conclusions

No previous innovation has had the potential to reshape medical cognition as profoundly as LLMs. To use AI as a cognitive stethoscope is to hear patterns more clearly while never mistaking the instrument for the act of listening itself. The path forward demands deliberate strategies. Clinicians must cultivate metacognitive habits to question and refine AI outputs; educators should embed AI into training as a scaffold for reasoning, prompting learners to articulate hypotheses, test alternatives, and confront biases; and policymakers and developers must ensure transparency, safety, and equity in AI design, with oversight mechanisms to guard against adaptive drift and systemic bias. Future inquiry should examine how AI influences clinical reasoning in real-world practice, evaluate its role in low-resource environments where expertise is scarce, and create regulatory frameworks that adapt alongside evolving models. Only then will it be clear whether AI strengthens cognition or quietly reshapes it in unintended ways.

Authors' Contributions

All authors wrote, made edits to, reviewed, and approved the final version of the manuscript for submission.

Conflicts of Interest

None declared.

References

1. Clusmann J, Kolbinger FR, Muti HS, et al. The future landscape of large language models in medicine. Commun Med (Lond) 2023 Oct 10;3(1):141. [doi: 10.1038/s43856-023-00370-1] [Medline: 37816837]
2. FDA digital health and artificial intelligence glossary – educational resource. U.S. Food & Drug Administration. 2024. URL: https://www.fda.gov/science-research/artificial-intelligence-and-medical-products/fda-digital-health-and-artificial-intelligence-glossary-educational-resource [accessed 2025-05-02]
3. McDuff D, Schaekermann M, Tu T, et al. Towards accurate differential diagnosis with large language models. Nature 2025 Jun;642(8067):451-457. [doi: 10.1038/s41586-025-08869-4] [Medline: 40205049]
4. Hope B. The greatest benefit to mankind: a medical history of humanity from antiquity to the present. BMJ 1998 Feb 28;316(7132):713. [doi: 10.1136/bmj.316.7132.713a]
5. Tulchinsky TH. John Snow, cholera, the Broad Street pump; waterborne diseases then and now. Case Stud Public Health 2018:77-99. [doi: 10.1016/B978-0-12-804571-8.00017-2]
6. Sakai T, Morimoto Y. The history of infectious diseases and medicine. Pathogens 2022 Oct 4;11(10):1147. [doi: 10.3390/pathogens11101147] [Medline: 36297204]
7. Barberis I, Bragazzi NL, Galluzzo L, Martini M. The history of tuberculosis: from the first historical records to the isolation of Koch's bacillus. J Prev Med Hyg 2017 Mar;58(1):E9-E12. [doi: 10.15167/2421-4248/jpmh2017.58.1.728] [Medline: 28515626]
8. Trickey A, Sabin CA, Burkholder G, et al. Life expectancy after 2015 of adults with HIV on long-term antiretroviral therapy in Europe and North America: a collaborative analysis of cohort studies. Lancet HIV 2023 May;10(5):e295-e307. [doi: 10.1016/S2352-3018(23)00028-0] [Medline: 36958365]
9. Morris ZS, Wooding S, Grant J. The answer is 17 years, what is the question: understanding time lags in translational research. J R Soc Med 2011 Dec;104(12):510-520. [doi: 10.1258/jrsm.2011.110180] [Medline: 22179294]
10. Duffy FD. Dialogue: the core clinical skill. Ann Intern Med 1998 Jan 15;128(2):139-141. [doi: 10.7326/0003-4819-128-2-199801150-00011] [Medline: 9441575]
11. Helman CG. Disease versus illness in general practice. J R Coll Gen Pract 1981 Sep;31(230):548-552. [Medline: 7328537]
12. Marcum JA. An integrated model of clinical reasoning: dual‐process theory of cognition and metacognition. J Eval Clin Pract 2012 Oct;18(5):954-961. [doi: 10.1111/j.1365-2753.2012.01900.x] [Medline: 22994991]
13. Corrao S, Argano C. Rethinking clinical decision-making to improve clinical reasoning. Front Med (Lausanne) 2022 Sep 8;9:900543. [doi: 10.3389/fmed.2022.900543] [Medline: 36160131]
14. Stewart M, Brown JB, Weston W, McWhinney IR, McWilliam CL, Freeman T, editors. Patient-Centered Medicine: Transforming the Clinical Method, 3rd edition: CRC Press; 2013. [doi: 10.1201/b20740-1]
15. Giordano C, Brennan M, Mohamed B, Rashidi P, Modave F, Tighe P. Accessing artificial intelligence for clinical decision-making. Front Digit Health 2021 Jun 25;3:645232. [doi: 10.3389/fdgth.2021.645232] [Medline: 34713115]

XSL•FO
RenderX

16. Eloranta S, Boman M. Predictive models for clinical decision making: deep dives in practical machine learning. J Intern Med 2022 Aug;292(2):278-295. [doi: 10.1111/joim.13483] [Medline: 35426190]
17. Yamashita R, Nishio M, Do RK, Togashi K. Convolutional neural networks: an overview and application in radiology. Insights Imaging 2018 Aug;9(4):611-629. [doi: 10.1007/s13244-018-0639-9] [Medline: 29934920]
18. Wiens J, Shenoy ES. Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology. Clin Infect Dis 2018 Jan 6;66(1):149-153. [doi: 10.1093/cid/cix731] [Medline: 29020316]
19. Thirunavukarasu AJ, Ting DS, Elangovan K, Gutierrez L, Tan TF, Ting DS. Large language models in medicine. Nat Med 2023 Aug;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
20. Goh E, Gallo R, Hom J, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. JAMA Netw Open 2024 Oct 1;7(10):e2440969. [doi: 10.1001/jamanetworkopen.2024.40969] [Medline: 39466245]
21. Stead WW. Clinical implications and challenges of artificial intelligence and deep learning. JAMA 2018 Sep 18;320(11):1107-1108. [doi: 10.1001/jama.2018.11029] [Medline: 30178025]
22. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019 Oct 25;366(6464):447-453. [doi: 10.1126/science.aax2342] [Medline: 31649194]
23. Zack T, Lehman E, Suzgun M, et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. Lancet Digit Health 2024 Jan;6(1):e12-e22. [doi: 10.1016/S2589-7500(23)00225-X] [Medline: 38123252]
24. Gao Y, Myers S, Chen S, et al. Uncertainty estimation in diagnosis generation from large language models: next-word probability is not pre-test probability. JAMIA Open 2025 Jan 10;8(1):ooae154. [doi: 10.1093/jamiaopen/ooae154] [Medline: 39802674]
25. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) no 300/2008, (EU) no 167/2013, (EU) no 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (text with EEA relevance). European Union. 2024. URL: http://data.europa.eu/eli/reg/2024/1689/oj/eng [accessed 2024-06-04]
26. Software as a Medical Device (SaMD). U.S. Food & Drug Administration. 2018 Apr 12. URL: https://www.fda.gov/medical-devices/digital-health-center-excellence/software-medical-device-samd [accessed 2025-06-04]
27. Choudhury A, Asan O. Role of artificial intelligence in patient safety outcomes: systematic literature review. JMIR Med Inform 2020 Jul 24;8(7):e18599. [doi: 10.2196/18599] [Medline: 32706688]
28. Liu TL, Hetherington TC, Dharod A, et al. Does AI-powered clinical documentation enhance clinician efficiency? A longitudinal study. NEJM AI 2024 Nov 22;1(12). [doi: 10.1056/AIoa2400659]

Abbreviations

AI: artificial intelligence
DL: deep learning
LLM: large language model
ML: machine learning

Edited by J Sarvestan; submitted 28.04.25; peer-reviewed by A Uchenna, E Oluwagbade, Z Yang; revised version received 18.08.25; accepted 29.08.25; published 09.10.25.

Please cite as:
Ferreira IBB, Menezes RC, Correia LCL, Andrade BB
Medicine Beyond Machines: Viewpoint on the Art of Thinking in the Age of AI
JMIR Form Res 2025;9:e76669
URL: https://formative.jmir.org/2025/1/e76669
doi:10.2196/76669

© Isabella B B Ferreira, Rodrigo C Menezes, Luis Cláudio L Correia, Bruno B Andrade. Originally published in JMIR Formative Research (https://formative.jmir.org), 9.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

mHealth as a Key Component of a New Model of Primary Care for Older Adults

Jean Woo[1,2], MD; Ruby Yu[1,2], PhD; Maggie Wong[2], MPH; Ken Cheung[2], MSc; Nicole Fung[2], PhD

[1]Department of Medicine and Therapeutics, Faculty of Medicine, Chinese University of Hong Kong, 9/F, Lui Che Woo Clinical Sciences Building, Prince of Wales Hospital, Shatin, China (Hong Kong)

[2]Jockey Club Institute of Ageing, Chinese University of Hong Kong, Shatin, China (Hong Kong)

Corresponding Author:
Jean Woo, MD
Department of Medicine and Therapeutics, Faculty of Medicine, Chinese University of Hong Kong, 9/F, Lui Che Woo Clinical Sciences Building, Prince of Wales Hospital, Shatin, China (Hong Kong)

Abstract

With population aging, an increase in total life expectancy at birth (TLE) should ideally be accompanied by an equal increase in health span (HS), or by a trend in increasing HS/TLE ratio. Hong Kong has one of the longest life expectancies in the world; however, there is a trend of declining HS/TLE ratio, such that the absolute number of people with dependencies is increasing. To address this challenge, the World Health Organization proposed the model of integrated care for older people (ICOPE) that combines both health and social elements in community care, using the measurement of intrinsic capacity (IC) as a metric for monitoring the performance in different countries. The use of technology is essential in achieving a wide coverage of the population in assessing IC, followed by an individually tailored plan of action. This model can be adapted to different health and social care systems in different countries. Hong Kong has an extensive network of community centers, where the basic assessment may be based, followed by further assessments and personalized activities, and referral to medical professionals may only be needed in the presence of disease. Conversely, the medical sector may refer patients to the community for activities designed to optimize the various domains of IC. Such a model of care has the potential to address manpower shortage and mitigate inequalities in healthy aging, as well as enable the monitoring of physiological systems in community-dwelling adults using digital biomarkers as a metric of IC.

(JMIR Form Res 2025;9:e82262) doi:10.2196/82262

KEYWORDS

intrinsic capacity; life expectancy; health span; integrated care of older persons; healthy aging; primary care

Demographic Shift Resulting in Increasing Dependency

As in many developed economies, Hong Kong is experiencing a rapid demographic shift, with its aging population growing at an unprecedented rate [1]. As of 2021, approximately 20% of Hong Kong's population was aged 65 years or above, and this figure is projected to rise to over 35% by 2046 [2]. This trend poses significant challenges to the health care system, elder care services, and informal caregivers, making it increasingly difficult to meet the complex health and social needs of older adults. As a result, an increase in health span has not matched the pace of increase in total life expectancy, resulting in an increase in absolute numbers of people with dependency needs [3].

The Urgency of the Problem

In spite of the existence of an excellent health system that is low-cost and free to those in poverty, a system that should theoretically mitigate health inequalities in the context of extreme wealth inequality, this dependency burden has resulted in health inequality due to unequal access [4]. Resources—such as clinical screenings, rehabilitation, and caregiver support—remain limited with many barriers to access. Many older adults, particularly those with mobility constraints or financial difficulties, struggle to access timely and appropriate care. The burden also extends to family caregivers, who often lack adequate support and knowledge, leading to increased stress and burnout [5]. At the same time, training programs in skills to manage age-related health conditions of personnel in elder care are not fit for purpose, and the number of such staff is not sufficient. Health and social care systems are also fragmented, such that early signs of functional decline may go unnoticed, leading to preventable complications and hospitalizations.

Need for Integration of Health Social Care Using a Stepped-Care Model, Shifting

XSL•FO
RenderX

From Hospital to Community, From Sickness to Prevention

To address these challenges, a structured and multitiered approach is essential. The stepped-care model provides a framework where different levels of intervention are applied based on the severity of a person's needs. This aligns with the Integrated Care for Older People (ICOPE) model, developed by the World Health Organization (WHO), which emphasizes early detection, prevention, and community-based support to promote healthy aging and independent living [6]. The ICOPE model focuses on monitoring key domains such as mobility, cognition, psychological well-being, nutrition, vision, and hearing to identify risks and implement timely interventions [7], in contrast to conventional healthcare that prioritizes diagnosing and managing individual diseases in isolation. Monitoring and managing declines in the domains of intrinsic capacity (IC), such as hearing, vision, memory, and mobility, can greatly affect quality of life and overall health [8]. By prioritizing prevention and proactive care, this innovative framework encourages a shift from reactive disease treatment to a proactive, function-based approach, empowering older adults to take charge of their health and well-being [9]. Preventive strategies for geriatric syndromes can significantly impact adverse health outcomes and dependency, as these syndromes increase hospital and outpatient utilizations [4].

Building upon IC as a monitoring tool for healthy aging, the ICOPE model supports the optimization of IC and functional ability of older adults, addressing their social and health needs in a holistic and person-centered approach [10]. The care pathway consists of four steps for delivering integrated care: basic assessment, in-depth assessment, development of a personalized care plan, implementation, and monitoring.

Technology Has a Key Role in the New Model of Primary Care in Raising Health Literacy

Technology has an important role in this care model, in enabling population-wide coverage. The concept of mHealth as a crucial tool for achieving healthy aging has been promoted by the WHO, particularly through its mAgeing program, as part of its global strategy to support older adults in maintaining their IC and functional ability [11]. Increasing adoption of technology and smartphones among older adults presents an opportunity to enhance health care accessibility through digital solutions. For example, smartphone adoption among Hong Kong's older adults rose significantly from 68% in 2020 to 90% in 2022 [12]. Digital tools can bridge the gap in service accessibility, providing self-management support, health education, and real-time resources that empower both older adults and caregivers.

Currently, various barriers exist for the successful implementation of the ICOPE care pathway in primary care. These include the lack of active engagement of stakeholders from community health and social care sectors and ageism, regarding age-related changes and declines in intrinsic capacity as an inevitable accompaniment of aging, focus on chronic disease and not taking into account geriatric syndromes, and fragmentation of care delivery systems into medical, and social compartments, hospitals and community settings, and absence of funding mechanism [13].

The Hong Kong Research Initiative in Using Technology as a Key Component of Providing a New Model of Integrated Health and Social Care in the Primary Care Setting

In order to address the unmet needs of Hong Kong's aging population and to promote equitable healthy aging, a project was initiated to raise health literacy relating to prevention, assessment, and management of IC declines and geriatric syndromes, among older adults, informal and formal caregivers, using information technology to create a mobile application covering assessment, and available neighborhood services (Multimedia Appendix 1). It is linked to an extensive series of topics focusing on commonly encountered problems, the underlying cause, and evidence-based management strategies. The design and development of the mobile application included repeated feedback from potential users: older adults themselves, informal and formal carers, through user testing sessions and focus group discussions. Specific age-friendly features were incorporated to facilitate adoption among older adults (Table 1).

The iHealth Screen mobile app integrates three key modules to support the ICOPE care pathway: (1) a health screening module covering IC and other 11 geriatric tests such as sarcopenia, cognition, and nutrition; (2) a self-care education hub offering credible content developed by geriatrician and nurses; and (3) a GPS-enabled community resource repository that connects users to 28 categories of medical and social care services across Hong Kong. These services include community health services, rehabilitation support, assistive tool rentals, barrier-free transportation, and carer support services, among others.

XSL•FO
RenderX

Table . Age-friendly design of iHealth Screen.

Key aspects	Age-friendly features
Font size	Enable standard and large font size options to suit for individual needs
Buttons and icons	Utilize large buttons and easily recognizable icons—such as a muscle icon for sarcopenia test and a brain icon for cognitive function tests— to facilitate older adults' comprehension
Color contrast	Use of high-contrast color, such as dark black text on a light yellow-orange background for clear visibility
Navigation	A chatbot-style interface guides users through health tests and community resource searches using predefined options. This design resembles familiar messaging and daily-use apps, eliminating the need for manual text input
Voice instruction	Voice-over instructions with text and images for particular health tests requiring additional steps, such as sarcopenia and cognitive tests, to support users who may benefit from auditory explanations
In-app onboarding	Offers welcome guidance and interactive walkthroughs to assist first-time users in familiarizing with the app's features and navigation

The health screening module enables users to complete assessments and receive instant results along with tailored educational recommendations. It incorporates standardized questionnaires and an embedded algorithm that processes predefined user responses to generate results and guidance. For instance, intrinsic capacity is assessed using a validated local questionnaire [14] that measures five domains: cognitive capacity, locomotor capacity, vitality, psychological capacity, and sensory capacity. The results are presented as a composite score ranging from 0 (no impairment) to 5 (impairment in all domains), and are categorized into three levels: 0 (robust), 1 - 2 (mild impairment), and 3 - 5 (severe impairment). Impaired domains are clearly highlighted to prompt user attention and guide follow-up actions. Users can also review past results and access an integrated health report to monitor changes over time and support early detection and intervention.

The mobile app is freely accessible to the public on both Android and iOS platforms. To ensure ease of use, individual users are not required to create personal accounts and can access all features directly. Upon first use, users are invited to complete a brief, optional demographic survey, including age group, gender, education level, and self-identified role. Each user is assigned a unique, anonymized ID for backend tracking of app usage, without collecting names or sensitive personal data. All data is securely stored on the university's web hosting infrastructure, ensuring security measures to prevent unauthorized access or data breaches. A comprehensive privacy policy is published within the app to inform users of their rights and ensure compliance with ethical and legal standards.

In addition to individual use, the app offers an institutional interface designed for potential community partners to integrate ICOPE care into routine operations. This module allows organizations to create accounts for their members, monitor health conditions, and plan targeted healthy aging initiatives. For institutional users, informed consent is obtained prior to collecting any personal information beyond the standard anonymous data. This ensures that all data collection is conducted transparently and ethically, with explicit user permission.

In particular, the usefulness of the mobile application was explored in various community centers, which provide various social support services as well as group activities. These are run by nongovernment organizations (NGOs) covering most areas of Hong Kong, with medical, nursing, and allied health input to varying extents. To adapt to local needs, where many foreign domestic helpers are employed to help with the care of older adults, translations into Tagalog, Indonesian, and Putonghua have been incorporated into the educational resources. An indicator of the perceived usefulness of this tool is provided by the following user statistics between November 2024 to March 2025: 7000 downloads of the app from app distribution platforms; 3300 users completed IC and related tests; and 2000 community resource searches.

1820 users completed 4626 IC, frailty, sarcopenia, and nutrition; and 2177 community resources searches by 1198 users, with the most frequently accessed being nutrition consultation, carer support service, followed by rehabilitation, assistive tool rental, and fall prevention. Qualitative feedback from older adults and caregivers highlighted that the app provides an accessible and consistent self-care tool for monitoring health and accessing resources, challenging the assumption that older individuals cannot use mHealth tools independently.

Despite positive engagement indicators, sustained use of the app remains a challenge. Limited awareness of health management often reduces motivation for regular app usage, as the need for health assessments is not always perceived as necessary. The value of standardized tools for monitoring intrinsic capacity and geriatric health is also not widely understood, with a preference for practical health information and demonstrations over completing assessments. Community outreach has been identified as an effective strategy to improve mHealth adoption, emphasizing the need for educational initiatives that not only focus on the technical aspects of app usage but also highlight the value of healthy aging and proactive health monitoring.

The creation of this product provides a tool that may be used as a key component of a new model of integrated primary care for older people, building on existing health and social care

XSL•FO
RenderX

infrastructure. Its use promotes empowerment for older adults and their carers, self-management, raising health literacy related to aging matters, participation in health-promoting group activities, linkage to existing medical and social services, other community resources, as well as reduces carer stress. The importance of raising health literacy cannot be emphasized enough, as one of the most significant issues in enabling health inclusivity for as wide a population as possible. It has been shown that a modest increase in health literacy may deliver significant health gains that can be translated into increasing a country's GDP as well as reducing health care costs [15]. This is supported by a recent literature review showing that health literacy combined with planned physical activity can increase self-efficacy, cognitive function, quality of life, and reduce frailty scores, falls risk, and depression scores [16]. However, providers need to ensure that technology does not create new barriers for older people to access information and services. In the development of this digital tool, great care has been placed in the design precisely to minimize this risk.

Creation of this digital tool enables the effective implementation of the ICOPE model for healthy aging in Hong Kong, following the principles of the model. The ICOPE model articulates the principles: countries have different economic developments, health and social care systems, so that the actual model may vary. In the case of Hong Kong, the tool may form an important link between the medical systems (hospitals, outpatient clinics) with the large network of social centers (eg, NGOs), based on the principles of social prescribing (Multimedia Appendix 2). This tool could be used as a first step to determine needs, followed by actions that may largely take place in NGO centers in terms of educational programs, more detailed functional assessments, and regular participation in group activities such as frailty prevention programs. The advantage of the app goes beyond the provision of just another tool, in its potential in countering ageism by raising health literacy relating to management of geriatric syndromes and the aging process itself, and also facilitates stakeholder engagement by offering a complementary pathway to continuing community care. The establishment of this new model of care has the potential to address the lack of manpower, in addition to mitigating inequalities in healthy aging. It may also enable the monitoring of physiological systems in community-dwelling adults using digital biomarkers, a trend that is transforming health care and personal health management [17].

Acknowledgments

The authors declare that no part of the manuscript used generative AI.

Funding

This work received funding from the Hong Kong Jockey Club Charities Trust.

Authors' Contributions

JW: Conceptualization, writing original draft, review & editing.
RY, MW, KC and NF: writing – review & editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
iHealth Screen.
[PNG File, 1346 KB - formative_v9i1e82262_app1.png]

Multimedia Appendix 2
A new model of care.
[PNG File, 484 KB - formative_v9i1e82262_app2.png]

References

1. Olshansky SJ, Willcox BJ, Demetrius L, Beltrán-Sánchez H. Implausibility of radical life extension in humans in the twenty-first century. Nat Aging 2024 Nov;4(11):1635-1642. [doi: 10.1038/s43587-024-00702-3] [Medline: 39375565]
2. Hong Kong population projections 2022-2046. Hong Kong Census and Statistic Department. 2023. URL: https://www.censtatd.gov.hk/en/data/stat_report/product/B1120015/att/B1120015092023XXXXB01.pdf [accessed 2025-12-14]
3. Yu R, Leung J, Lum CM, et al. A comparison of health expectancies over 10 years: implications for elderly service needs in Hong Kong. Int J Public Health 2019 Jun;64(5):731-742. [doi: 10.1007/s00038-019-01240-1] [Medline: 31032531]
4. Woo J. Policy Implications of Population Ageing in Hong Kong in the UN Decade of Healthy Aging. Am J Med Public Health 2023;4(5):1056 [FREE Full text]

XSL•FO
RenderX

5. Chan CY, Cheung G, Martinez-Ruiz A, et al. Caregiving burnout of community-dwelling people with dementia in Hong Kong and New Zealand: a cross-sectional study. BMC Geriatr 2021 Apr 20;21(1):261. [doi: 10.1186/s12877-021-02153-6] [Medline: 33879099]
6. Guidance on person-centred assessment and pathways in primary care. World Health Organization. 2019. URL: https://iris.who.int/bitstream/handle/10665/326843/WHO-FWC-ALC-19.1-eng.pdf?sequence=17 [accessed 2025-08-10]
7. Cesari M, Araujo de Carvalho I, Amuthavalli Thiyagarajan J, et al. Evidence for the domains supporting the construct of intrinsic capacity. J Gerontol A Biol Sci Med Sci 2018 Nov 10;73(12):1653-1660. [doi: 10.1093/gerona/gly011] [Medline: 29408961]
8. Zhang S, Tange C, Huang ST, et al. Multi-trajectories of intrinsic capacity and their effect on higher-level functional capacity, life satisfaction, and self-esteem in community-dwelling older adults: the NILS-LSA. J Nutr Health Aging 2025 Jan;29(1):100432. [doi: 10.1016/j.jnha.2024.100432] [Medline: 39615397]
9. Woo J. Take Charge of Your Own Ageing: The Chinese University of Hong Kong Press; 2024.
10. Integrated Care for Older People (ICOPE): Guidance for Person-Centred Assessment and Pathways in Primary Care, 2nd edition: World Health Organization; 2024. URL: https://iris.who.int/bitstream/handle/10665/380175/9789240103726-eng.pdf?sequence=1 [accessed 2025-08-10]
11. A Handbook on How to Implement mAgeing: World Health Organization; 2018. URL: https://iris.who.int/bitstream/handle/10665/274576/9789241514125-eng.pdf?sequence=1 [accessed 2025-08-10]
12. Thematic household survey report no.77 information technology usage and penetration. Hong Kong Census and Statistics Department. 2023. URL: https://www.censtatd.gov.hk/en/data/stat_report/product/B1130201/att/B11302772023XXXXB0100.pdf [accessed 2025-08-10]
13. Yu R, Leung G, Lai D, et al. Assessing the readiness for implementing the World Health Organization's ICOPE approach in Hong Kong: perspectives from social care and policy stakeholders. J Frailty Aging 2023;12(2):126-133. [doi: 10.14283/jfa.2023.3] [Medline: 36946709]
14. Yu R, Leung G, Leung J, et al. Prevalence and distribution of intrinsic capacity and its associations with health outcomes in older people: The Jockey Club Community eHealth Care Project in Hong Kong. J Frailty Aging 2022;11(3):302-308. [doi: 10.14283/jfa.2022.19] [Medline: 35799437]
15. Tucker P. Health drives wealth: the economic impact of health inclusivity. Economist Impact. URL: https://impact.economist.com/projects/health-inclusivity-index/documents/health_inclusivity_index_phase3_report.pdf [accessed 2025-12-15]
16. Tandilangi TP, Rekawati E, Eriska W. Improving functional status of older adults through health literacy media: a systematic review. bik 2024;17(2):218-226. [doi: 10.23917/bik.v17i2.5063]
17. Lu JK, Wang W, Mahadzir MDA, et al. Digital biomarkers of ageing for monitoring physiological systems in community-dwelling adults. Lancet Healthy Longev 2025 Jun;6(6):100725. [doi: 10.1016/j.lanhl.2025.100725] [Medline: 40517785]

Abbreviations

HS: health span
IC: intrinsic capacity
ICOPE: integrated care for older people
NGO: nongovernment organizations
TLE: total life expectancy at birth
WHO: World Health Organization

Edited by A Schwartz, M Balcarras; submitted 12.08.25; peer-reviewed by A Sudaryanto, S Eshun; revised version received 12.11.25; accepted 24.11.25; published 19.12.25.

Please cite as:
Woo J, Yu R, Wong M, Cheung K, Fung N
mHealth as a Key Component of a New Model of Primary Care for Older Adults
JMIR Form Res 2025;9:e82262
URL: https://formative.jmir.org/2025/1/e82262
doi:10.2196/82262

© Jean Woo, Ruby Yu, Maggie Wong, Ken Cheung, Nicole Fung. Originally published in JMIR Formative Research (https://formative.jmir.org), 19.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction

XSL•FO
RenderX

in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Hype vs Reality in the Integration of Artificial Intelligence in Clinical Workflows

Alaa Abd-Alrazaq[1*], PhD; Barry Solaiman[2], PhD; Yosra Magdi Mekki[3], MD; Dena Al-Thani[4], PhD; Faisal Farooq[5], PhD; Metab Alkubeyyer[6], MD; Mohamed Ziyad Abubacker[7], MD; Rawan AlSaad[1], PhD; Sarah Aziz[1], MSc; Ahmed Serag[8], PhD; Rajat Thomas[9], PhD; Javaid Sheikh[1], MD; Arfan Ahmed[1*], PhD

[1]AI Center for Precision Health, Weill Cornell Medical College in Qatar, Education City, Doha, Qatar
[2]College of Law, Hamad bin Khalifa University, Doha, Qatar
[3]Department of Basic Medical Sciences, College of Medicine, QU Health, Qatar University, Doha, Qatar
[4]College of Science and Engineering, Hamad bin Khalifa University, Doha, Qatar
[5]Qatar Computational Research Institute (QCRI), Hamad Bin Khalifa University, Doha, Qatar
[6]Department of Radiology and Medical Imaging, King Khalid University Hospital, King Saud University, Riyadh, Saudi Arabia
[7]Department of Radiology, Al Ahli Hospital, Doha, Qatar
[8]AI Innovation Lab, Weill Cornell Medical College, Doha, Qatar
[9]Weill Cornell Medical College, Doha, Qatar
[*]these authors contributed equally

Corresponding Author:
Arfan Ahmed, PhD
AI Center for Precision Health, Weill Cornell Medical College in Qatar, Education City, Doha, Qatar

Abstract

Artificial intelligence (AI) has the capacity to transform health care by improving clinical decision-making, optimizing workflows, and enhancing patient outcomes. However, this potential remains limited by a complex set of technological, human, and ethical barriers that constrain its safe and equitable implementation. This paper argues for a holistic, systems-based approach to AI integration that addresses these challenges as interconnected rather than isolated. It identifies key technological barriers, including limited explainability, algorithmic bias, integration and interoperability issues, lack of generalizability, and difficulties in validation. Human factors such as resistance to change, insufficient stakeholder engagement, and education and resource constraints further impede adoption, whereas ethical and legal challenges related to liability, privacy, informed consent, and inequity compound these obstacles. Addressing these issues requires transparent model design, diverse datasets, participatory development, and adaptive governance. Recommendations emerging from this synthesis are as follows: (1) establish standardized international regulatory and governance frameworks; (2) promote multidisciplinary co-design involving clinicians, developers, and patients; (3) invest in clinician education, AI literacy, and continuous training; (4) ensure equitable resource allocation through dedicated funding and public-private partnerships; (5) prioritize multimodal, explainable, and ethically aligned AI development; and (6) focus on long-term evaluation of AI in real-world settings to ensure adaptive, transparent, and inclusive deployment. Adopting these measures can align innovation with accountability, enabling health care systems to harness AI's transformative potential responsibly and sustainably to advance patient care and health equity.

(JMIR Form Res 2025;9:e70921) doi:10.2196/70921

KEYWORDS

artificial intelligence; AI; clinical workflow; challenges; solutions; technology; human factors; ethics; regulation; health care

Introduction

Artificial intelligence (AI) is reshaping clinical health care, yet real-world uptake remains limited. The gap between hype and reality stems from fragmented integration that treats technology, people, and ethics as separate tracks. This viewpoint argues that responsible adoption requires a holistic integration framework that advances technological robustness, human alignment, and ethical and regulatory accountability together. Progress in any 1 domain without the others undermines impact; therefore, coordinated advancement across all 3 is needed to translate promise into clinical value. AI tools already influence high-stakes clinical decisions in imaging, triage, treatment selection, and remote monitoring. Their rapid diffusion, sometimes ahead of regulation and robust evidence, underscores the need for a principled, holistic integration framework. Prior reviews catalog challenges but rarely explain how the domains interact or how to coordinate solutions across stakeholders [1-4].

XSL•FO
RenderX

This paper contributes by leading the conversation from what the challenges are to how they can be collectively resolved.

Figure 1 summarizes three interlocking domains of AI integration: technological, human, and ethical and legal. Each domain includes principal challenges, for example, explainability and bias; interoperability and generalizability; workforce engagement and training; and liability, privacy, consent, equity, and conflicts of interest. Together, these dependencies determine whether AI becomes a trusted clinical partner or remains a research accessory. To preserve concision, detailed technical explanations, case examples, and jurisdictional analyses are provided in Appendix A in Multimedia Appendix 1. The translation gap reflects misalignment between rapid technical progress and slower social, ethical, and institutional adaptation. Four principles guide the analysis: validation parity (AI systems should meet evidence standards comparable to drugs and devices), human and AI complementarity (tools should augment rather than replace clinician judgment), transparency and accountability (models and processes must be traceable, with clear responsibility across developers, institutions, and regulators), and equitable infrastructure (fair access to data, talent, and computing so that adoption does not widen disparities). These principles frame each domain discussed next and anchor the targeted solutions.

Figure 1. Challenges of integrating artificial intelligence (AI) in clinical workflows.

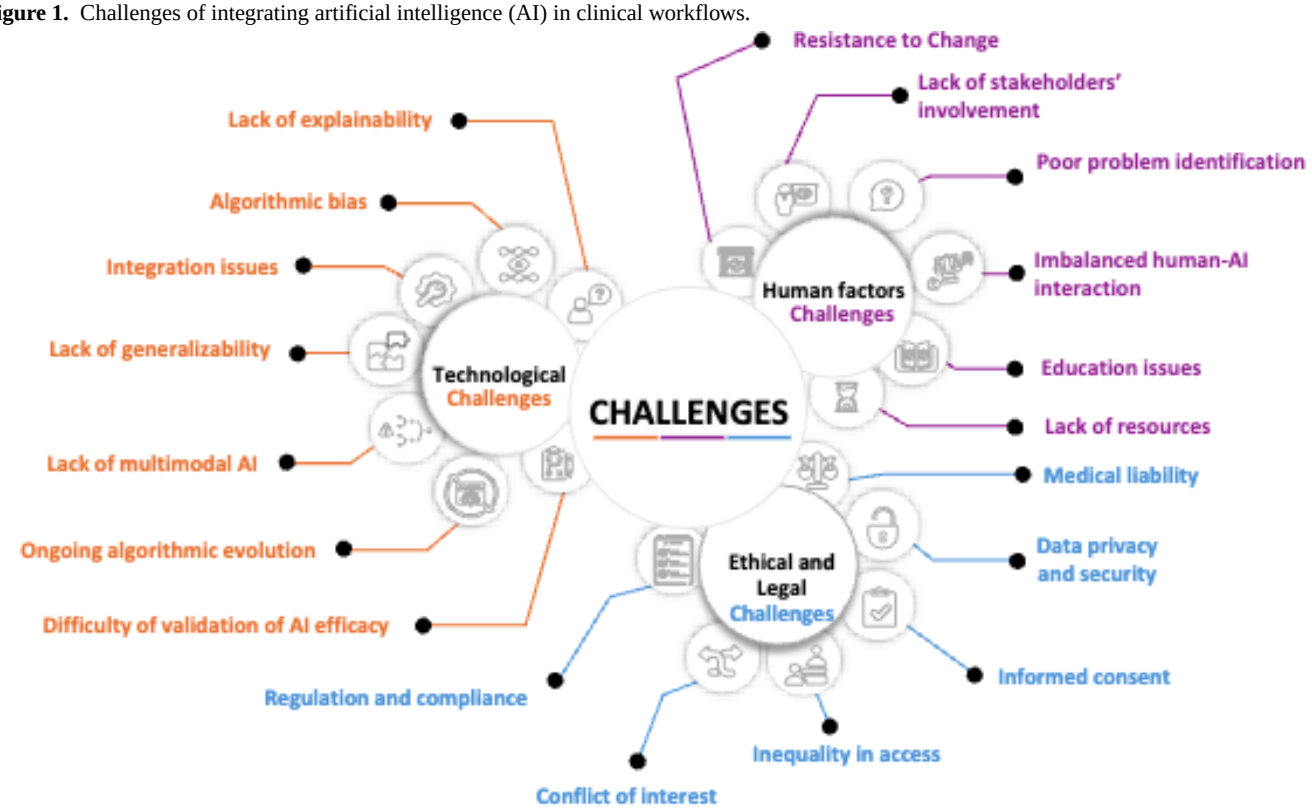

We write for clinicians, policymakers, ethicists, and technologists and move beyond describing obstacles to explain why partial fixes fail and how coordinated approaches can succeed. The paper aims to clarify how technical design links to human trust and ethical legitimacy; to translate this into strategies that turn prototypes into regulated clinical tools; and to offer concise, actionable recommendations for policy, practice, and research. We address 3 domains aligned with Figure 1 (ie, technological foundations, human and organizational dimensions, and ethical and regulatory context), synthesize cross-cutting implications, and conclude with concrete recommendations. By reframing AI integration as a holistic sociotechnical problem rather than a purely technological one, we seek to bridge the gap between innovation and implementation and support accountable, equitable adoption. A summary of the challenges, proposed solutions, and policy implications discussed in this article is outlined in Table 1. We include a glossary of technical terms in Multimedia Appendix 2.

XSL•FO
RenderX

Table . Summary of challenges, proposed solutions, and policy implications.

Challenge	Proposed solutions (authors' position)	Policy or practice implications
Technological		
Lack of explainability	Mandate regulatory-grade validation and audits and embed XAI[a] principles for transparency	Require FDA[b] or EMA[c]-style approval processes, develop explainability standards for health care AI[d], fund XAI research, and require transparency reports
Algorithmic bias	Diverse and representative training data, debiasing techniques, and continuous auditing	Fund diverse dataset creation, institutionalize fairness audits, and enforce transparency in dataset composition
Lack of generalizability	Require multi-institutional validation, use of foundational models, and continuous monitoring	Make external validation mandatory and incentivize data sharing across institutions
Integration issues	Adopt interoperability standards (FHIR[e], OMOP[f], HL7[g], DICOM[h]), modular API[i]-based design, and NLP[j] for unstructured data	Promote universal interoperability standards and incentivize legacy system upgrades
Lack of multimodal AI	Invest in multimodal ML[k] and HPC[l] or cloud infrastructure and integrate structured and unstructured	Allocate funding for R&D[m] and infrastructure and set guidelines for multimodal validation
Ongoing evolution	Continuous monitoring for drift, clinician-adjustable thresholds, embedded confidence scores, and fail-safe mechanisms	Mandate postdeployment auditing and regulate update protocols for "locked" versus adaptive algorithms
Difficulty of validation	Standardized validation frameworks, prospective RCTs[n] for AI, and data repositories	Develop AI-specific RCT guidelines and fund longitudinal data repositories for validation
Human		
Resistance to change	Comprehensive training, education, and communication; frame AI as collaborative tool; and user-centered design	Develop national AI curricula for health care, incentivize adoption through training subsidies, and develop certification standards for AI usability
Lack of stakeholders' involvement	Multidisciplinary development teams, "shadow clinician" immersion, and iterative feedback loops	Require co-design processes in procurement and implementation
Poor problem identification	Evidence-based prioritization and align AI initiatives with clinical and operational goals	Establish hospital AI governance committees with diverse stakeholder representation
Imbalanced human-AI interaction	Position AI as decision support, establish clear usage protocols, embed fail-safes, and training on limitations	Develop national guidelines on human oversight, set standards for AI confidence thresholds, and integrate AI interaction into clinical training
Education and training issues	Update medical curricula; tailored, role-specific training; and lifelong learning	Accredit AI in health courses and provide subsidies for continuing AI education
Lack of resources	Strategic funding and partnerships, train existing staff, phased implementation, and plan infrastructure before deployment	Allocate dedicated budgets for AI workforce and infrastructure, ensure equity in resource allocation, and incentivize public-private partnerships
Ethical or legal		
Liability concerns	Define legal responsibility across developers, providers, and institutions and regulatory approval for developers	Develop national and international liability frameworks for AI and clarify insurer responsibilities for AI-related claims
Data privacy and security	Strengthen consent models, data anonymization, and blockchain and advanced security	Strengthen HIPAA[o] or GDPR[p]-style protections and enforce auditable consent processes
Informed consent issues	Mandate disclosure of AI use, dynamic consent models, and simplify explanations	Legally require patient notification of AI involvement and standardize consent forms for AI
Inequality in access	Ensure equity-driven design and subsidized deployment in underresourced settings; develop cost-effective, open-source AI; infrastructure investment; and diverse training data	Fund equitable AI deployment, mandate accessibility assessments for approved AI tools, policy frameworks to prioritize underserved populations

XSL•FO
RenderX

Challenge	Proposed solutions (authors' position)	Policy or practice implications
Conflict of interest	Mandatory transparency and disclosure, independent oversight bodies, and ethical training	Enforce strict conflict of interest declarations for procurement and research and require third-party evaluation
Regulation and compliance	Adaptive regulatory models and international regulatory harmonization	Create agile regulatory bodies and promote WHO[q], FDA, or EMA coordination

[a]XAI: explainable artificial intelligence.
[b]FDA: Food and Drug Administration.
[c]EMA: European Medicines Agency.
[d]AI: artificial intelligence.
[e]FHIR: Fast Healthcare Interoperability Resources.
[f]OMOP: Observational Medical Outcomes Partnership.
[g]HL7: Health Level 7.
[h]DICOM: Digital Imaging and Communications in Medicine.
[i]API: application programming interface.
[j]NLP: natural language processing.
[k]ML: machine learning.
[l]HPC: high performance computing.
[m]R&D: research and design.
[n]RCT: randomized controlled trial.
[o]HIPAA: Health Insurance Portability and Accountability Act.
[p]GDPR: General Data Protection Regulation.
[q]WHO: World Health Organization.

Technological Challenges

Lack of Explainability

The lack of explainability in AI systems poses a major barrier to their integration into clinical workflows. Deep learning (a type of machine learning that uses artificial neural networks to learn from data, similar to the way we learn) models often operate as "black boxes," producing predictions without transparent reasoning. This opacity limits clinicians' ability to validate system outputs, undermining trust and accountability. When clinicians cannot trace the basis for a decision, potential errors or biases remain hidden, impeding clinical validation and adoption.

Lack of explainability is often misunderstood. Explainability does not require every model to be interpretable at the bedside; rather, it demands a traceable and auditable development process that documents how the model produces outcomes and how its reliability is verified. Given identical inputs, the system should produce reproducible outputs within a known error margin, and this process must be validated through rigorous testing.

Regulatory bodies such as the Food and Drug Administration (FDA) should develop validation and approval processes for AI systems analogous to those for drugs and medical devices, ensuring reliability and safety. At the same time, explainable AI techniques play a complementary role by offering interpretable insights—highlighting patient features that most influenced a prediction or generating visual explanations to support clinical review. The optimal pathway, therefore, combines rigorous validation to guarantee reproducibility with targeted transparency to build user trust. This balanced black box functionality at the point of care coupled with full traceability behind the scenes is essential for sustainable clinical integration.

Algorithmic Bias

AI systems can produce biased outcomes that reflect and reinforce inequities in the underlying data. The sources of algorithmic bias are multifaceted and include historical inequities present in training datasets, socioeconomic disparities, and unequal access to health care services. Using unrepresentative data can lead to skewed model outputs that disadvantage certain populations. In health care, such bias may cause unequal care delivery and poorer outcomes for underrepresented groups. For example, a risk prediction tool used in US hospitals systematically underreferred Black patients to specialized care programs compared with White patients with equivalent medical needs. The algorithm used health care costs as a proxy for illness severity, assuming lower spending meant better health. In reality, lower expenditures reflected systemic barriers to care rather than improved outcomes, resulting in underallocation of resources to minority patients and amplifying inequities.

Reducing algorithmic bias requires more than technical adjustments. Models should be trained on datasets that are diverse and representative of the populations they serve. When such data are unavailable, methods such as resampling or data augmentation can be used to improve representativeness. However, these approaches must be applied cautiously, as they may introduce new biases if synthetic data reproduce flawed assumptions. Developers can also apply importance weighting to assign higher training value to underrepresented samples, improving model fairness.

XSL•FO
RenderX

Another critical step is eliminating proxy variables that encode protected attributes, such as race or socioeconomic status. Techniques such as blinding or data transformation can help obscure these attributes. Bias mitigation must also extend beyond the development phase: continuous auditing and multidisciplinary oversight involving ethicists, clinicians, and data scientists are necessary to identify emergent biases and sustain fairness.

A frequently cited parallel exists in non-AI health care systems, such as the revision of the estimated glomerular filtration rate formula, which until 2021 included race-based adjustments that were later removed to promote equity. AI systems, unlike static formulas, can adapt much faster once bias is detected, but only if the feedback mechanisms are properly implemented [5].

Lack of Generalization

AI models often perform well in controlled research environments but struggle to replicate that performance in real-world clinical settings. This discrepancy arises because development and deployment environments differ in data quality, patient populations, and workflow practices. Real-world data are frequently noisy, incomplete, and inconsistent, leading to reduced reliability when AI systems are transferred between institutions. A model that functions effectively in 1 hospital may fail elsewhere due to variation in clinical protocols and documentation standards. The problem is compounded by the dynamic nature of health care, where evolving patient demographics and treatment practices can quickly outdate static models.

A notable example is IBM Watson for Oncology, one of the most widely publicized failures of AI in health care [6]. Trained primarily on synthetic data from a single cancer center, the system failed to generalize to diverse patient populations and local treatment practices. Its recommendations were inconsistent with clinical standards, leading to its withdrawal from the market. The Watson case underscores that training on narrow or synthetic datasets undermines generalizability and erodes trust in AI systems. To improve generalizability, models should be trained and validated on datasets from multiple institutions and populations, using both cross-validation and external validation techniques. Broader data sharing across health care systems remains essential, although challenging in practice. Continuous monitoring of deployed AI systems should become standard practice to detect performance drift early. Data quality, encompassing completeness, consistency, timeliness, and validity, is as critical as dataset size and should be prioritized equally. Clinician feedback during validation is indispensable, as practitioners provide contextual insights absent from retrospective data. Developing adaptive models capable of learning from new data and evolving clinical environments can further enhance reliability. Techniques such as site-specific fine-tuning, local threshold calibration, and transfer learning enable customization without sacrificing overall model robustness [7].

Integration Issues

Health care environments often depend on diverse legacy systems and electronic health record (EHR) platforms, which rarely communicate seamlessly because each system uses distinct data formats, protocols, and standards. For AI systems to operate effectively, they must interact with these heterogeneous infrastructures. Achieving such interoperability is complex and resource-intensive owing to inconsistent protocols and data formats across platforms. The problem is further complicated by the nature of health care data itself, which includes structured fields, free-text clinical notes, imaging, and sensor data. Integrating these heterogeneous sources into unified, machine-readable formats remains a major technical and operational challenge.

Another barrier is ensuring real-time data exchange. Clinical workflows rely on up-to-date information, but legacy systems often cannot sustain the speed or reliability required for AI-assisted decision-making [8]. Compatibility gaps between AI applications and existing clinical software may require extensive customization and ongoing maintenance, making implementation resource-intensive and sometimes impractical.

Addressing these integration barriers requires coordinated technical and organizational strategies. Developers should assess existing technological infrastructure before deployment to ensure compatibility and minimize operational disruption. Standardized data models and exchange frameworks, such as the Observational Medical Outcomes Partnership and openEHR, provide structured approaches for harmonizing EHR data [9]. These frameworks enhance interoperability by creating common data structures that facilitate consistent data sharing and algorithm training across sites. Likewise, adopting interoperability standards such as Fast Healthcare Interoperability Resources enables structured data exchange across clinical systems, promoting real-time integration of AI tools into workflows. This not only improves compatibility and communication across platforms but also enhances the timeliness and reliability of clinical decision support.

Lack of Multimodal AI

Most AI systems in health care still rely on a single data type, such as radiological images for disease detection. While unimodal systems have produced valuable insights, they often fail to capture the multidimensional nature of patient health. Modern health care generates diverse data sources, including clinical records, imaging, genomic sequences, wearable sensor data, and patient-reported outcomes, which together provide a more complete understanding of health and disease [10]. When AI models analyze only 1 modality, they risk overlooking critical context, such as comorbidities, lifestyle factors, or environmental influences, leading to incomplete or less accurate predictions. For example, chest computed tomography–based algorithms often struggle to differentiate overlapping pulmonary pathologies that could be better distinguished by integrating imaging with laboratory and clinical data.

Developing multimodal AI systems presents technical and operational challenges. Integrating heterogeneous data requires advanced architectures capable of aligning different data structures and ensuring consistent scaling and normalization. Combining structured and unstructured data also demands significant computational resources and data harmonization.

XSL•FO
RenderX

These requirements highlight the need for investments in infrastructure and algorithmic innovation.

Improving multimodal AI capacity depends on sustained research investment in machine learning methods that can effectively process heterogeneous datasets. Standardization of data formats and the establishment of interoperable frameworks are essential to link imaging, text, and sensor-derived data. Expanding high-performance computing infrastructure, including cloud-based and distributed systems, will also be necessary to handle the scale of multimodal data. Ultimately, the integration of structured and unstructured clinical data with social, behavioral, and wearable inputs offers a path toward a holistic view of patient care. This approach promises improved diagnostic precision and personalized interventions, representing a key step in realizing the full potential of AI in health care.

Ongoing Algorithmic Evolution

AI algorithms are designed to optimize their performance by continually evolving and learning new rules and techniques. While this adaptability offers potential advantages, it introduces serious risks when applied to health care. One concern is that evolving AI models may optimize for short-term metrics that do not align with patient well-being or clinical goals. For example, a predictive model might learn superficial correlations, such as associating positive outcomes with specific hospital equipment rather than genuine clinical indicators. Such behavior illustrates the danger of algorithms identifying statistical patterns that lack causal or clinical relevance.

The autonomous nature of adaptive AI also raises safety concerns. As these models explore new strategies, they may inadvertently adopt unvalidated or unsafe practices without adequate oversight. Another challenge involves data drift, the gradual change in input data distributions over time, which can degrade model performance even if the system was initially trained on large, heterogeneous datasets [11,12]. Changes in imaging protocols, scanner technology, or patient demographics can all contribute to declining accuracy, highlighting the need for continuous vigilance.

Mitigating these risks requires ongoing model monitoring rather than 1-time validation. Continuous surveillance should detect deviations from expected behavior and trigger corrective actions promptly. Structured feedback loops between clinicians and developers are essential to ensure that observed workflow disruptions or unexpected model behaviors inform retraining and improvement. Confidence assessment mechanisms should also be embedded directly into AI systems, preventing models from generating outputs when uncertainty exceeds safe thresholds. Fail-safe measures must be implemented to prevent unreliable predictions from influencing clinical decisions. Real-time monitoring for data drift is equally important; drift detection should enable dynamic recalibration of model thresholds to maintain performance under changing conditions.

Difficulty of Validation of AI Efficacy

Validating the efficacy of AI systems in health care remains a major challenge. Robust validation demands real-world testing within clinical workflows while ensuring minimal disruption, a process that requires time, resources, and collaboration from health care professionals who may remain skeptical of emerging technologies. Establishing rigorous evidence through randomized controlled trials, long considered the gold standard for medical evaluation, is especially difficult for AI. Unlike static interventions, AI models evolve through continuous learning, complicating the definition of stable evaluation criteria. Additionally, ensuring data diversity and representativeness poses logistical and ethical barriers, as collaboration across multiple institutions and jurisdictions is often constrained by privacy regulations and data governance requirements. Retrospective datasets, which remain the predominant source for AI development, can introduce hidden biases, limiting the generalizability of findings to real-world clinical contexts.

To advance trustworthy AI, validation must be systematic and standardized. Data-sharing frameworks and cross-institutional collaborations should underpin the creation of centralized, anonymized repositories for longitudinal and prospective validation. Establishing regulatory and technical guidance for conducting AI-specific RCTs is crucial for ensuring that models undergo testing under real-world conditions. Co-developed validation protocols, jointly designed by developers, clinicians, and regulators, can help ensure clinical relevance and reliability. Transparency and reproducibility must also be enforced through open reporting of algorithms, datasets, and validation outcomes, enabling peer scrutiny and evidence-based refinement. Ultimately, regulatory endorsement by independent bodies such as the FDA and the World Health Organization should represent the benchmark for AI systems before clinical deployment, ensuring global credibility and patient safety.

Human Factors

Resistance to Change

Health care professionals' resistance to AI can significantly impede its integration into clinical workflows. This resistance stems from several interrelated causes. Clinicians may hesitate to trust AI systems because of skepticism about their accuracy, reliability, and capacity to make sound clinical judgments. Such skepticism is often linked to a limited understanding of how AI systems function and concerns regarding transparency and explainability. Fear of accountability for adverse outcomes when relying on AI recommendations also contributes to reluctance. In addition, some professionals worry that AI may reduce their autonomy, challenge their expertise, or even replace their roles. The introduction of AI tools often disrupts established routines and requires adaptation to new processes, which can be perceived as burdensome and time-consuming. Moreover, clinicians express concern that AI could depersonalize care, diminishing empathy and the human connection central to medicine. Finally, institutional incentives and recognition programs can encourage adoption by aligning AI use with professional development and organizational goals (see sections Lack of Explainability, Education and Training Issues, Liability Concerns, Data Privacy and Security, Imbalanced Human-AI Interaction, Integration Issues).

Lack of Stakeholders' Involvement

The successful development and deployment of AI systems in health care depend on the early and sustained participation of

XSL•FO
RenderX

diverse stakeholders, including clinicians, patients, administrators, data scientists, and IT professionals. When these groups are insufficiently engaged, integration efforts often encounter resistance, inefficiencies, and misaligned goals. Clinicians who are not included in development may distrust AI systems or find them poorly adapted to real-world workflows, resulting in limited adoption. Patient involvement is equally important, as transparency regarding data privacy, security, and ethical use helps build confidence and preserve autonomy. Administrators and technical teams ensure compliance with legal and infrastructure requirements, whereas their absence may lead to implementation failures or regulatory nonconformity. A lack of coordinated engagement can also create disconnects between developers who prioritize technical optimization and health care professionals who value usability and clinical outcomes.

Addressing these challenges requires formalized, multidisciplinary collaboration throughout design, testing, and deployment. Development teams should include health care providers, patients, ethicists, data scientists, and engineers to foster trust, ownership, and shared accountability. Immersive approaches, such as assigning "shadow clinicians," in which developers or observers follow clinicians during daily practice, can offer valuable insights into workflow realities and user behaviors. These experiences inform the creation of intuitive interfaces that align closely with clinical routines and reduce cognitive load. Continuous feedback loops between clinicians, administrators, and AI developers should also be institutionalized to sustain improvement. Collaboration models such as the Google Health-Mayo Clinic partnership demonstrate that co-design and iterative engagement can significantly enhance usability, accuracy, and clinician acceptance [13] (see section Lack of Explainability).

Poor Problem Identification and Prioritization

Health care organizations often face challenges in identifying and prioritizing where AI can deliver the most value. The sector's complexity, with diverse patient needs and rapidly changing priorities, makes selecting suitable problems difficult. Successful integration requires alignment among clinicians, administrators, and IT professionals, each of whom may define value differently. Clinicians tend to emphasize diagnostic accuracy and clinical outcomes, administrators focus on cost efficiency and resource optimization, and IT teams prioritize feasibility and infrastructure compatibility. Misalignment among these perspectives can delay decision-making and impede adoption.

Resource limitations add further complexity. Many organizations face budget constraints, workforce shortages, and limited digital maturity, which hinder deployment. Prioritization must therefore balance clinical needs with technological readiness and available capacity. Poorly chosen initiatives risk diverting resources from higher impact opportunities, whereas flexible governance structures are essential to adapt priorities as clinical and operational contexts evolve. The COVID-19 pandemic exemplified this need for adaptability, as priorities shifted rapidly from long-term analytics to real-time applications such as outbreak tracking and hospital resource allocation [14]. The experience of IBM Watson for Drug Discovery illustrates how poor problem definition can undermine even advanced technologies. The project targeted a highly complex domain, drug discovery, where success rates are inherently low. Its broad objective produced results that were either redundant or not actionable, leading to limited adoption and eventual discontinuation [15].

Effective prioritization begins with multidisciplinary teams that combine clinical, technical, and administrative expertise. Evidence-based frameworks should guide the selection of use cases that address measurable clinical or operational needs. Projects that integrate seamlessly into workflows and demonstrate early, tangible impact are most likely to succeed and maintain stakeholder support. Continuous evaluation and real-world feedback enable organizations to refine focus areas and sustain alignment between AI initiatives and evolving health care demands.

Imbalanced Human-AI Interaction

Determining the appropriate level of reliance on AI recommendations remains a key challenge in clinical practice. Overreliance can cause clinicians to accept outputs uncritically, neglecting clinical cues and diminishing independent judgment. This complacency risks diagnostic errors and weakens clinical reasoning skills over time. A well-documented example of this challenge is the Epic Sepsis Model developed by Epic Systems to identify patients at risk of sepsis. Many US hospitals deployed it, but several studies revealed that the tool frequently generated false alarms and missed a substantial number of true sepsis cases [16,17]. Clinicians who trusted it blindly risked patient harm, whereas those who ignored its alerts overlooked potential benefits. The episode underscored the importance of calibrated trust and clear human oversight in clinical AI use. Health care institutions must foster balanced human-AI interaction through 3 key strategies. First, AI systems should function strictly as decision support tools rather than autonomous decision-makers, ensuring that ultimate responsibility remains with clinicians. Second, comprehensive training programs should emphasize both the capabilities and limitations of AI, helping clinicians critically interpret system outputs and maintain accountability. Third, protocols and governance frameworks must delineate when and how to use AI, reserving complex or ethically sensitive decisions for human judgment. Finally, explainability remains vital: transparent models with interpretable outputs, such as visual saliency maps or confidence indicators, build user trust and encourage appropriate reliance. These measures ensure that AI augments rather than replaces clinical expertise, sustaining both safety and human-centered care (see sections Education and Training Issues, Regulation and Compliance Hurdles, and Lack of Explainability).

Education and Training Issues

The successful adoption of AI technologies in health care depends heavily on clinicians' ability to understand, interpret, and apply these tools effectively. However, many professionals lack sufficient training to do so. The absence of structured AI curricula in most medical education programs has created a knowledge gap that limits understanding of AI methodologies, capabilities, and limitations. This gap fosters hesitation and

XSL•FO
RenderX

mistrust toward AI, as clinicians often feel uncertain about technologies they do not fully comprehend. Continuous professional development is also limited, as the rapid pace of AI innovation makes it difficult for institutions to maintain up-to-date training programs. Moreover, education rarely addresses the differing needs of health care roles. Physicians, nurses, administrators, and IT staff engage with AI in distinct ways, yet most training remains generic and insufficiently tailored to these contexts.

Effective solutions begin with reforming medical and health sciences education to include core AI competencies. Undergraduate and postgraduate curricula should incorporate foundational modules on AI methodology, data ethics, and computational reasoning, emphasizing practical applications and critical evaluation of AI outputs. Health care institutions should complement academic reform with ongoing, role-specific professional training that balances theoretical understanding with hands-on experience. Successful examples include dedicated AI workshops at Harvard Medical School, the Cambridge Center for AI in Medicine, and Weill Cornell Medicine–Qatar, which integrate interdisciplinary instruction and practical simulation. These initiatives demonstrate the value of combining technical and ethical education through cross-faculty collaboration. Cultivating a culture of continuous learning is equally essential: clinicians must be equipped to adapt to evolving technologies, reflecting the broader shift toward lifelong digital competence across the health workforce.

Lack of Resources

Resource limitations, manpower, time, and finances pose significant barriers to integrating AI into clinical practice. Implementing AI systems often requires additional staff to manage data entry, monitoring, and tool maintenance. Most health care institutions already face staff shortages and heavy workloads, leaving little capacity for new responsibilities. The expert input required to train and validate AI algorithms is both expensive and time-consuming, as data labeling and annotation typically rely on highly skilled clinicians. These tasks divert specialists from clinical duties and increase operational costs. Maintaining AI systems further compounds the burden because updates, monitoring, and recalibration require sustained financial investment and technical expertise.

A well-documented example is the Google Health AI for diabetic retinopathy detection, which achieved strong research results but underperformed in real-world use due to resource constraints. In rural Thai clinics, limited access to high-quality imaging equipment and reliable internet connectivity reduced accuracy and slowed analysis. These limitations highlight that even well-designed systems depend on adequate infrastructure, training, and funding for successful deployment.

Overcoming resource constraints requires long-term strategic planning. Institutions should adopt phased implementation to distribute costs and workload, whereas policymakers establish funding mechanisms that support infrastructure, data curation, and workforce development. Public-private partnerships can accelerate progress by pooling resources and expertise. Finally, equitable resource allocation is essential to ensure that underresourced health care settings benefit from AI innovations, preventing the deepening of global health disparities.

Ethical and Legal Considerations

Liability Concerns

The integration of AI into health care has complicated traditional concepts of medical liability. When patient harm occurs following an AI recommendation, responsibility may be unclear—should it rest with the physician, the hospital, the developer, or the regulator who approved the system? Most jurisdictions still apply conventional tort law, assigning liability primarily to clinicians, even when decisions rely on algorithms that they neither designed nor fully understand. This creates uncertainty and discourages adoption.

Legal scholars have argued that hospitals should share responsibility under vicarious liability because physicians often operate under institutional authority. This approach simplifies adjudication and reflects the shared enterprise of AI-assisted care. However, current regimes rarely provide patients with a direct path to seek redress from developers (see section Regulation and Compliance Hurdles).

A pragmatic interim model is to assign joint liability to clinicians and hospitals, with insurers or institutional authorities covering claims. This balances accountability, promotes transparency, and provides legal certainty for health care professionals. Over time, national and international frameworks should evolve toward shared responsibility models that clearly delineate obligations across clinicians, institutions, and AI developers. Regulators should ensure that liability rules remain adaptive, reflecting the dynamic learning nature of AI systems and the distributed accountability inherent in their deployment.

Data Privacy and Security

AI systems depend on vast amounts of sensitive patient data, raising major privacy and cybersecurity concerns. Large-scale data aggregation across EHRs, imaging repositories, and wearable devices increases exposure risks, making health care AI a prime target for breaches and unauthorized access. Research has shown that adversarial attacks can extract identifiable training data from AI systems, revealing patient information and compromising confidentiality. Even anonymized datasets are vulnerable, as sophisticated algorithms can reidentify individuals through pattern matching and cross-referencing. (Issues of data ownership and control further complicate the landscape, as patients, providers, and technology developers have competing interests in how medical data are used, shared, and monetized).

The 2016 DeepMind–Royal Free London NHS collaboration illustrates these challenges. Data from 1.6 million patients were shared without sufficient transparency or explicit consent, leading to significant public backlash when Google assumed direct control of the project [18]. This case underscored the risks of inadequate oversight in public-private partnerships and the urgent need for stronger governance.

Existing regulatory frameworks, such as Health Insurance Portability and Accountability Act in the United States and

XSL•FO
RenderX

General Data Protection Regulation in Europe, remain insufficient to address AI-specific risks. Their compliance models were not designed for dynamic, data-intensive AI systems, and many countries have adopted them merely for market harmonization. We argue that next-generation frameworks must move beyond compliance checklists to embed proactive accountability and patient-centered control. This includes integrating medical confidentiality laws into AI data protection regimes, as seen in Qatar's health legislation, which should serve as a model for linking data protection to ethical practice [19].

Emerging technologies such as blockchain can enhance traceability and ensure auditability of AI data use, improving both security and trust. However, without harmonized global data localization standards, cross-border AI operations will remain fragmented. A coordinated international approach, combining technical safeguards, legal harmonization, and ethical oversight, is essential for protecting privacy while enabling innovation in global health AI.

Informed Consent Issues

Informed consent is a cornerstone of ethical medical practice, ensuring patients understand and agree to proposed interventions. However, integrating AI into health care complicates this principle. In several jurisdictions, clinicians are not legally required to inform patients when AI tools are used in diagnosis or treatment. This omission undermines transparency and can erode trust. The opacity of "black box" models makes it difficult for clinicians to explain how AI arrives at specific recommendations or decisions, further challenging patient comprehension. The dynamic nature of AI compounds the issue, as continuously learning systems evolve over time, potentially invalidating earlier explanations and raising questions about whether consent remains truly informed. There is also uncertainty about whether patients must reconsent when their data are reused to validate or refine AI systems (see sections Lack of Explainability and Ongoing Algorithmic Evolution).

To safeguard autonomy, informed consent laws should be amended to establish a universal baseline requiring that patients be notified whenever AI is used in their care. Patients must receive clear, comprehensible explanations about how AI systems operate, what data they use, and potential risks, including algorithmic bias and privacy concerns. Information should be provided through written materials, visual aids, or digital formats tailored to different literacy levels. Health care providers must be trained to communicate these concepts effectively, supported by developers who enhance explainability to facilitate such dialogue (see section Education and Training Issues).

Dynamic consent models offer a viable path forward by allowing patients to update preferences over time and choose specific permissions for data use, including training, validation, and development. Patients should also retain the right to refuse AI involvement in their care altogether. Regulators must enforce standardized consent guidelines to ensure patients are informed consistently, regardless of jurisdiction. In institutional settings where explicit consent is impractical, such as hospitals using AI-based tracking, implied consent may apply, but clear notification should remain mandatory. Establishing hospital ethics committees focused on AI oversight would help ensure ethical compliance and accountability.

Inequality in Access to Health Care

AI has the potential to widen existing disparities in health care access. Underresourced clinics and hospitals, especially in rural or low-income areas, often lack the infrastructure required for AI adoption, including reliable internet connectivity, high-performance computing, and trained personnel. This digital divide limits their ability to implement AI systems, creating inequities in access to advanced diagnostics and personalized care. Moreover, many AI models are trained on datasets that underrepresent minority populations, leading to biased performance and diagnostic inaccuracies for these groups. Such biases can reinforce structural inequities, erode trust, and result in poorer outcomes for marginalized communities. The high cost of developing and maintaining AI systems further exacerbates disparities, allowing well-funded institutions to deploy advanced tools while resource-limited facilities fall behind.

Ensuring equitable access requires deliberate and globally coordinated action. AI developers must prioritize designing cost-effective, scalable, and open-source models that can operate on basic hardware and limited bandwidth. Training datasets must be diverse and representative to minimize bias and enhance performance across populations. Governments and nonprofit organizations should allocate targeted funding to strengthen digital infrastructure and workforce capacity in low-resource settings, whereas public-private partnerships can facilitate technology transfer and cost-sharing mechanisms. Policymakers should also adopt proactive regulatory measures that mandate equitable distribution of AI resources, ensuring that underserved regions are prioritized in national digital health strategies. Through these efforts, AI can evolve from a technology that risks deepening inequity into a tool that actively promotes fairness, inclusivity, and universal access to high-quality care.

Conflicts of Interest

Conflicts of interest in AI arise when the personal, financial, or institutional interests of health care providers, developers, or policymakers compromise patient-centered decision-making. Financial relationships between clinicians and AI companies can bias procurement and use, as physicians involved in AI development may favor tools in which they hold investments or royalties. Partnerships between developers and health care institutions can also blur boundaries between clinical integrity and commercial gain, particularly when companies sponsor trials or offer incentives that influence results or adoption decisions. The collapse of Babylon Health, a UK startup once valued at $4.2 billion, illustrates the dangers of unchecked conflicts. Despite early concerns from regulators and researchers, Babylon secured NHS contracts and political endorsements through lobbying and financial contributions [20]. Its rapid expansion, driven more by political and financial support than by proven clinical efficacy, ended in bankruptcy in 2023, eroding public trust and wasting substantial investment. The case highlights how opaque financial and political

XSL•FO
RenderX

relationships can distort AI adoption, undermining evidence-based medicine and public confidence.

Mitigating such risks requires strict transparency and governance measures. Mandatory disclosure of all financial ties between clinicians, institutions, and AI developers should be a regulatory requirement. Hospitals should establish internal ethics committees to oversee AI procurement and usage decisions, ensuring these are guided by patient safety and clinical benefit rather than commercial interest. Independent auditing of AI trials and deployment outcomes should be routine to identify potential biases. Finally, separating the roles of developers and clinical decision-makers, through independent evaluation bodies, would safeguard neutrality and promote integrity in AI adoption.

Regulation and Compliance Hurdles

The absence of unified, AI-specific regulation remains one of the most significant barriers to safe and effective implementation of AI in health care. Existing laws, while robust in data protection, do not adequately address the dynamic and adaptive nature of AI, creating uncertainty around liability, validation, and accountability. Regulatory fragmentation compounds these challenges. For example, US frameworks such as HIPAA and the FDA approval processes differ markedly from the EU's GDPR and the AI Act, resulting in conflicting expectations and inconsistent oversight for developers and health care providers [20]. In countries lacking formal approval mechanisms, startups may find themselves in regulatory limbo, unable to bring products to market or compete with foreign firms holding recognized certifications. At the same time, opaque "black box" models and the absence of standardized evaluation metrics make it difficult for regulators to ensure safety and reliability across diverse health care contexts.

We argue that an adaptive and collaborative regulatory framework is urgently required. This should prioritize (1) international harmonization, promoting coordination between bodies such as the World Health Organization, the FDA, the European Medicines Agency, and regional regulators to develop common AI validation and approval standards; (2) agile regulation, establishing rapid-update mechanisms that can evolve alongside technological innovation while preserving patient safety; and (3) transparent oversight, mandating algorithmic explainability, performance auditing, and public reporting of outcomes. Additionally, regulators must embed context-aware evaluation, ensuring that assessments reflect real-world clinical variation rather than controlled laboratory conditions. By embracing adaptive, risk-based regulation, policymakers can safeguard innovation while maintaining accountability and equity in global health care AI deployment.

Conclusions

The integration of AI into health care holds transformative potential but remains hindered by complex technological, human, and ethical challenges. Despite promising advances, real-world adoption has been slow, with successful implementations still the exception rather than the norm. This paper provides a holistic framework categorizing the major challenges, technological, human, and ethical or legal, and presents evidence-informed, actionable strategies to address them.

From a technological perspective, progress depends on explainable, unbiased, and interoperable systems that can generalize across populations and settings. Human challenges, including resistance to change, insufficient stakeholder engagement, poor problem definition, and limited training, must be addressed through education, collaboration, and resource investment. Ethical and legal barriers, including liability, privacy, consent, inequality, conflicts of interest, and fragmented regulation, require adaptive governance and robust safeguards to protect patient rights while enabling innovation.

In practice, we prioritize actions in three tiers: first, governance and safety; second, workflow and people; and third, equity and scale. This translates to institutional oversight and validation parity with postmarket monitoring stakeholder co-design and role-specific education for fit-for-workflow and calibrated human-AI interaction, and equity through interoperability and infrastructure in low-resource settings, diverse datasets, and cross-jurisdictional coordination.

The path forward demands sustained, interdisciplinary collaboration among clinicians, developers, regulators, and policymakers. AI must be viewed not as a technological end point but as a continually evolving tool within a sociotechnical ecosystem. By combining transparent design, equitable access, and adaptive regulation, health care systems can move beyond hype to responsible implementation. Realizing AI's full potential will depend not only on technical breakthroughs but also on collective commitment to ethics, accountability, and human-centered care.

Acknowledgments

The authors declare the use of generative AI in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: Proofreading and editing. The GAI tool used was ChatGPT-4. Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes. Declaration submitted by all authors. All intellectual content, study design, data collection, analysis, and final interpretations are the sole responsibility of the authors.

Funding

The authors received no specific funding for this work.

XSL•FO
RenderX

Data Availability

Data sharing is not applicable to this article as no data sets were generated or analyzed during this study.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed technical explanations, case examples, and jurisdictional analyses.
[DOCX File, 33 KB - formative_v9i1e70921_app1.docx]

Multimedia Appendix 2
Glossary of technical terms.
[DOCX File, 17 KB - formative_v9i1e70921_app2.docx]

References

1. Younis HA, Eisa TAE, Nasser M, et al. A systematic review and meta-analysis of artificial intelligence tools in medicine and healthcare: applications, considerations, limitations, motivation and challenges. Diagnostics (Basel) 2024 Jan 4;14(1):109. [doi: 10.3390/diagnostics14010109] [Medline: 38201418]
2. Romagnoli A, Ferrara F, Langella R, Zovi A. Healthcare systems and artificial intelligence: focus on challenges and the international regulatory framework. Pharm Res 2024 Apr;41(4):721-730. [doi: 10.1007/s11095-024-03685-3] [Medline: 38443632]
3. Singh RP, Hom GL, Abramoff MD, Campbell JP, Chiang MF, AAO Task Force on Artificial Intelligence. Current challenges and barriers to real-world artificial intelligence adoption for the healthcare system, provider, and the patient. Transl Vis Sci Technol 2020 Aug;9(2):45. [doi: 10.1167/tvst.9.2.45] [Medline: 32879755]
4. Guan J. Artificial intelligence in healthcare and medicine: promises, ethical challenges and governance. Chin Med Sci J 2019 Jun 30;34(2):76-83. [doi: 10.24920/003611] [Medline: 31315747]
5. Delgado C, Baweja M, Crews DC, et al. A unifying approach for GFR estimation: recommendations of the NKF-ASN task force on reassessing the inclusion of race in diagnosing kidney disease. Am J Kidney Dis 2022 Feb;79(2):268-288. [doi: 10.1053/j.ajkd.2021.08.003] [Medline: 34563581]
6. Gerke S, Minssen T, Cohen G. Ethical and legal challenges of artificial intelligence-driven healthcare. Artif Intell Healthc 2020:295-336. [doi: 10.1016/B978-0-12-818438-7.00012-5]
7. Yu S, Farooq F, van Esbroeck A, Fung G, Anand V, Krishnapuram B. Predicting readmission risk with institution-specific prediction models. Artif Intell Med 2015 Oct;65(2):89-96. [doi: 10.1016/j.artmed.2015.08.005] [Medline: 26363683]
8. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. Nat Med 2019 Jan;25(1):30-36. [doi: 10.1038/s41591-018-0307-0] [Medline: 30617336]
9. Bönisch C, Kesztyüs D, Kesztyüs T. Harvesting metadata in clinical care: a crosswalk between FHIR, OMOP, CDISC and openEHR metadata. Sci Data 2022 Oct 28;9(1):659. [doi: 10.1038/s41597-022-01792-7] [Medline: 36307424]
10. Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ. Multimodal biomedical AI. Nat Med 2022 Sep;28(9):1773-1784. [doi: 10.1038/s41591-022-01981-2] [Medline: 36109635]
11. Davis SE, Greevy RA, Lasko TA, Walsh CG, Matheny ME. Detection of calibration drift in clinical prediction models to inform model updating. J Biomed Inform 2020 Dec;112:103611. [doi: 10.1016/j.jbi.2020.103611] [Medline: 33157313]
12. Sahiner B, Chen W, Samala RK, Petrick N. Data drift in medical machine learning: implications and potential remedies. Br J Radiol 2023 Oct;96(1150):20220878. [doi: 10.1259/bjr.20220878] [Medline: 36971405]
13. Mayo clinic selects google as strategic partner for health care innovation, cloud computing. : Mayo Clinic; 2024 URL: https://newsnetwork.mayoclinic.org/discussion/mayo-clinic-selects-google-as-strategic-partner-for-health-care-innovation-cloud-computing/ [accessed 2025-12-05]
14. Pesapane F, Codari M, Sardanelli F. Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine. Eur Radiol Exp 2018 Oct 24;2(1):35. [doi: 10.1186/s41747-018-0061-6] [Medline: 30353365]
15. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
16. Kamran F, Tjandra D, Heiler A, et al. Evaluation of sepsis prediction models before onset of treatment. NEJM AI 2024 Feb 22;1(3):AIoa2300032. [doi: 10.1056/AIoa2300032]
17. Wong A, Otles E, Donnelly JP, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. JAMA Intern Med 2021 Aug 1;181(8):1065-1070. [doi: 10.1001/jamainternmed.2021.2626] [Medline: 34152373]

XSL•FO
RenderX

18. Murdoch B. Privacy and artificial intelligence: challenges for protecting health information in a new era. BMC Med Ethics 2021 Sep 15;22(1):122. [doi: 10.1186/s12910-021-00687-3] [Medline: 34525993]
19. Solaiman B, Vansweevelt T, Glover-Thomas N. Chapter 7: data protection, privacy, and confidentiality in qatar's health system. In: Privacy and Medical Confidentiality in Healthcare: A Comparative Analysis: Edward Elgar Publishing; 2023:114-139. [doi: 10.4337/9781035309436.00012]
20. Babylon Health: the failed AI wonder app that "dazzled" politicians. The Week. 2023. URL: https://theweek.com/health/babylon-health-the-failed-ai-wonder-app-that-dazzled-politicians [accessed 2025-08-27]

Abbreviations

AI: artificial intelligence
EHR: electronic health record

Edited by J Sarvestan; submitted 06.01.25; peer-reviewed by BS Ibrahim, ER Chaudhuri, H Saeidnia, M Alsahli; revised version received 29.10.25; accepted 30.10.25; published 12.12.25.

Please cite as:
Abd-Alrazaq A, Solaiman B, Mekki YM, Al-Thani D, Farooq F, Alkubeyyer M, Abubacker MZ, AlSaad R, Aziz S, Serag A, Thomas R, Sheikh J, Ahmed A
Hype vs Reality in the Integration of Artificial Intelligence in Clinical Workflows
JMIR Form Res 2025;9:e70921
URL: https://formative.jmir.org/2025/1/e70921
doi:10.2196/70921

© Alaa Abd-Alrazaq, Barry Solaiman, Yosra Magdi Mekki, Dena Al-Thani, Faisal Farooq, Metab Alkubeyyer, Mohamed Ziyad Abubacker, Rawan AlSaad, Sarah Aziz, Ahmed Serag, Rajat Thomas, Javaid Sheikh, Arfan Ahmed. Originally published in JMIR Formative Research (https://formative.jmir.org), 12.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Leveraging Artificial Intelligence for Clinical Study Matching: Key Threads for Interweaving Data Science and Implementation Science

Andrew James Goodwin[1], MD, MSCR; Sara Ann Armstrong[2], MD; David Ptak[3]; Kenneth Catchpole[4], PhD; Jihad S Obeid[5], MD; Paul M Heider[5], PhD

[1]Division of Pulmonary, Critical Care, Allergy and Sleep Medicine, Medical University of South Carolina, Suite 816 CSB, MSC 630, 96 Jonathan Lucas St., Charleston, SC, United States

[2]Department of Medicine, Medical University of South Carolina, Charleston, SC, United States

[3]Biomedical Informatics Center, Medical University of South Carolina, Charleston, SC, United States

[4]Department of Anesthesia and Perioperative Medicine, Medical University of South Carolina, Charleston, SC, United States

[5]Department of Public Health Sciences, Medical University of South Carolina, Charleston, SC, United States

Corresponding Author:
Andrew James Goodwin, MD, MSCR
Division of Pulmonary, Critical Care, Allergy and Sleep Medicine, Medical University of South Carolina, Suite 816 CSB, MSC 630, 96 Jonathan Lucas St., Charleston, SC, United States

Abstract

Artificial intelligence holds the potential to enhance the efficiency of clinical research. Yet, like all innovations, its impact is dependent upon target user uptake and adoption. As efforts to leverage artificial intelligence for clinical trial screening become more widespread, it is imperative that implementation science principles be incorporated in both the design and roll-out of user-facing tools. We present and discuss implementation themes considered to be highly relevant by target users of artificial intelligence–enabled clinical trial screening platforms. The identified themes range from design features that optimize usability to collaboration with tool designers to improve transparency and trust. These themes were generally mapped to domains of existing implementation science frameworks such as the Consolidated Framework for Implementation Research. Designers should consider incorporating an implementation science framework early in the development process to not only ensure a user-centered design but to inform how tools are integrated into existing clinical research workflows.

(JMIR Form Res 2025;9:e71831) doi:10.2196/71831

KEYWORDS

artificial intelligence; machine learning; clinical trials; implementation science; clinical research

Introduction

Artificial intelligence (AI) presents a transformative opportunity to optimize health care through machine learning and advanced data analysis to assist clinicians in diagnosis, prevention, and treatment planning. In theory, well-designed AI should allow us to focus on processes that require higher level decision-making while cognitively offloading us from simpler and time-consuming work [1]. These same principles can be applied to clinical research in which AI can enhance efficiency through automation. As AI development has accelerated at an exponential pace, it has led to the opportunity for it to support all manner of complex research workflows, such as disease recognition [2], outcome prediction [3,4], treatment discovery [5], and clinical trial eligibility matching [6-11]. Despite these advances, the optimal strategies to achieve uptake and drive utilization of AI-enabled tools by clinical researchers remain underexplored.

Combining Data Science and Implementation Science

As we collectively move toward an era of AI-enabled clinical research, we should be mindful of potential pitfalls of using AI and pursue strategies to mitigate them whenever possible. The incorporation of implementation science principles into the design, building, and dissemination of AI-enabled clinical research tools will be critical to these efforts. The Consolidated Framework for Implementation Research (CFIR) offers a useful schema to help identify key targetable factors associated with the adoption of new tools and processes [12]. When applied to the uptake of an AI-enabled trial matching tool, CFIR highlights several potentially relevant targets across each domain (Figure 1). Targets such as leadership engagement, knowledge and beliefs about a tool, the adaptability of a tool, and the planning around implementation all illustrate the importance of a preliminary intake process during which design and study teams

XSL•FO
RenderX

can align on knowledge, design needs, and workflow integration. Such alignment is necessary to ensure that the correct data are correctly displayed using a correct prioritization scheme. Furthermore, this process could improve readiness to implement by increasing transparency and reducing mistrust in an unknown algorithm.

Figure 1. Identified targets for the implementation of an artificial intelligence (AI)-enabled clinical trial matching tool mapped to the domains of the consolidated framework for implementation research model.

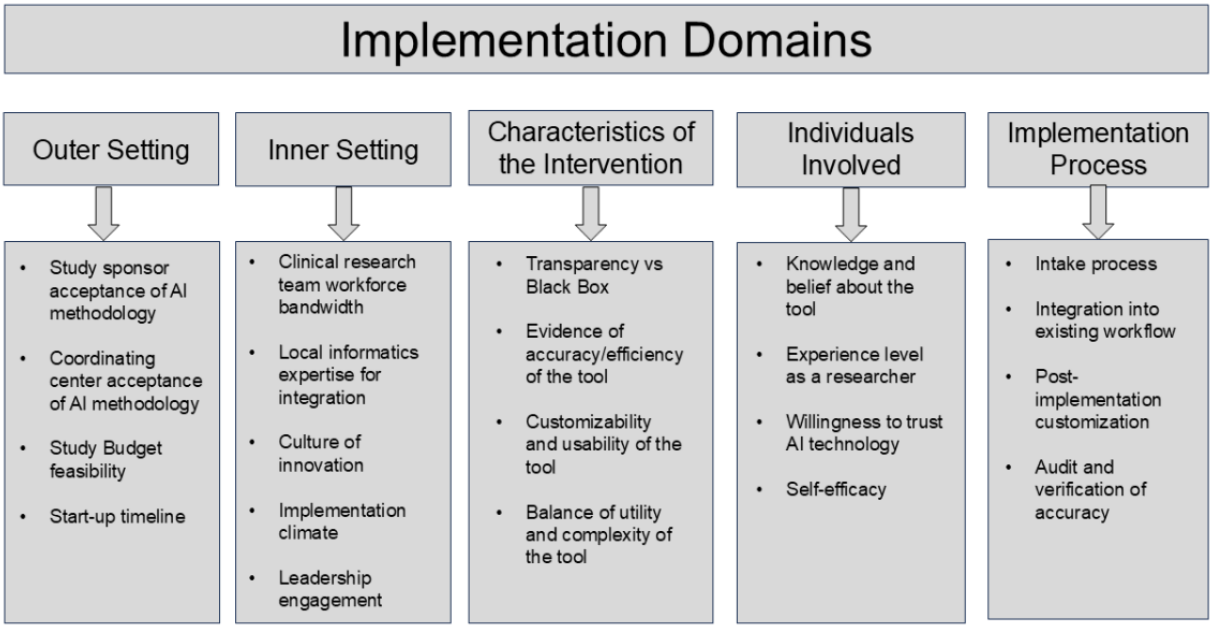

As part of a larger effort to examine how AI-enabled clinical trial optimization tools can be developed and successfully implemented by research teams, we performed a pilot project aimed to design an AI-enabled prototype clinical trial patient matching tool. The matching tool was constructed of concepts derived from structured and unstructured (eg, clinical notes) data that were individually mapped to the inclusion and exclusion criteria of four test trials. Unstructured data were analyzed by natural language processing informed by clinical annotators while structured data were mapped to concepts by a clinical trialist and informaticists. A consensus approach was used in cases where more than one data point could be mapped to a concept (eg, an ICD-10 code for “fever” versus documented hyperthermia). The screening tool was integrated into a data pipeline with interoperability with our institution’s research data warehouse and was engineered to return potential trial candidates to a prototype user interface. A complete description of the developmental methods and the model’s performance will be reported elsewhere.

During the development process, our team convened focus groups with clinical research coordinators to understand their workflows, identify needs, and solicit their perspectives on the use of AI in clinical research (Multimedia Appendix 1). Eight research coordinators with a wide range in both duration and breadth of role experience participated in one session each during which they were provided a demonstration of the screening tool and asked structured questions. Several key themes emerged from these discussions, which should be considered when creating or adapting an AI-enabled clinical research management tool. These included the importance of integrating existing workflows into product design, the need for customizability, and a desire for transparency to achieve trust in the tool. We expand on these critical features below.

Integration With Existing Workflows

Clinical research teams frequently employ strategies to optimize screening efficiency. One common approach identified during our focus groups is the use of “gatekeeper” criteria. These criteria tend to be frequent causes of exclusion from eligibility, and thus, are prioritized in the screening process. Coordinators revealed that this prioritization allows research teams to expend less time excluding large volumes of subjects while allowing a smaller pool of potential subjects through the “gate” where a deeper examination of eligibility is required. Research personnel also target their searches to include a potentially eligible population. Provided examples of targeted searching included screening a specific provider’s clinic panel if they frequently manage patients with the disease of interest, screening an inpatient unit where patients with a particular condition are commonly cohorted, or screening inpatient lists, such as an inpatient infection prevention list, for a study targeting a communicable infectious disease. These targeted approaches can provide concentrated pools of potential subjects while minimizing the time required for searching. Gatekeeper strategies may also be considered in the prioritization scheme of how AI tools display potential subjects to clinical research end users (ie, prioritizing or ranking the presentation of subjects who have met key gatekeeper criteria to focus time-constrained users). Incorporating these and other commonly employed

XSL•FO
RenderX

screening approaches into the design of an AI-supported matching tool could potentially provide several benefits including (1) the more rapid identification of subjects facilitating enrollment in trials with narrow eligibility windows, (2) improved efficiency allowing for trial performance in resource-constrained settings such as rural hospitals, and (3) the expansion of the potential subject pool for trials with rarer eligibility criteria.

Customizability

The unique nature of each study's eligibility criteria, procedures, and data capture necessitate that AI-enabled support tools be customizable. A "one size fits all" approach will not only lack precision across trials but will also fail to meet the needs and expectations of a study team, and thus, limit uptake. In our discussions, clinical research coordinators identified several potential customizable features that should be considered when building an AI-enabled support tool. Often, these corresponded to unique screening bottlenecks that a study may present to a team. For example, for studies with the inclusion criteria that are rare or that require significant time to identify, a patient matching tool that prioritizes the recall of these criteria could provide great benefit. Alternatively, for studies with commonly found inclusion criteria, a tool that prioritizes precision across multiple exclusion criteria may be preferred to narrow the pool of potential candidates. Coordinators also highlighted the importance of recognizing whether mutable criteria exist, which may exclude a potential subject from eligibility at one screening time point but have the potential to change over time. If present, a patient matching tool could be customized with a "watch list" feature where subjects excluded by mutable criteria can be monitored longitudinally for changes in study candidacy.

An additional customizable feature that may enhance the value of a matching tool is the ability to easily screen across multiple studies. Research teams frequently screen for multiple trials in a disease space simultaneously with team members varying in their screening approaches (ie, screen all potential patients for a specific trial first versus screen an individual patient for all potential trials first). A tool that is adaptable enough to accommodate for such variation in workflow may lead to better uptake. Finally, research coordinators expressed an interest in being able to modify the tool after implementation. Gatekeeper criteria are not always evident until screening begins for a trial, and these criteria can differ across sites. Thus, the ability to alter which criteria the screening tool is prioritizing based on lessons learned can improve its utility.

Transparency

Another consistent theme identified by clinical research coordinators was whether they could develop sufficient trust in an AI-enabled screening tool to effectively use it [13]. Specific concerns included both the potential for inappropriate subjects to be recommended by a screening tool or for a tool to fail to identify appropriate candidates. These concerns are amplified by the "black box problem" of AI in which a lack of transparency diminishes trust in an algorithm and impedes its utilization [14,15]. However, research coordinators did articulate that trust in an AI-enabled screening tool could be enhanced through (1) coordination with the design team to improve algorithm transparency, (2) line of sight to where an eligibility criteria was identified by the algorithm, and (3) positive experiences with the algorithm during which its accuracy is demonstrated.

Conclusions

Generalizability should be a cornerstone of the tool design and should be considered when designing how a tool can integrate on the front end with an electronic medical record and on the back end with an electronic source database. Research coordinators at our site identified several key themes central to the successful design of an AI-enabled tool, which mapped to implementation science domains. However, other research teams in different environments and with different workflows may identify a distinct set of design priorities. This highlights the potential importance of applying a rigorous implementation science framework to the tool design process.

Many clinical trials within a disease space contain similar or identical inclusion and exclusion criteria such that screening models may be adaptable between different studies. A system designed to optimize reusability would reduce development time and offer greater confidence in the validity of a model for the research team. The pace of innovation in the field of AI should also be considered in the design of clinical research tools to avoid rapid obsolescence. We favor the incorporation of modular design to facilitate upgrades without necessitating a complete rebuild. Furthermore, we advocate for the routine use of common data models as well as the inclusion of application programming interfaces at all natural joints in a tool. These future-proofing efforts will enhance the longevity of AI-based programs and their value to clinical research teams.

In summary, we currently stand at a key juncture on our path toward leveraging artificial intelligence in the performance of clinical research. As we take these initial steps, we should be open to the myriad lessons that implementation science offers and mindful of the specific needs of clinical research teams. By seeking these insights, we can realize the potential of AI to enhance the accuracy and efficiency of future studies. However, ignorance of these insights will delay the uptake of AI-enabled tools indefinitely and should be done at our peril.

Authors' Contributions

Conceptualization: AJG (lead), PMH (equal)
Data Curation: AJG (lead), PMH (equal), KC (equal), DP (supporting)
Project Administration: AJG (lead), PMH (equal), DP (supporting)

XSL•FO
RenderX

Writing – original draft: AJG (lead), SAA (supporting)
Writing – review and editing: AJG (lead), PMH (equal), SAA (supporting), DP (supporting), KC (supporting), JSO (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Focus group guide.
[DOCX File, 15 KB - formative_v9i1e71831_app1.docx]

References

1. Grinschgl S, Neubauer AC. Supporting cognition with modern technology: distributed cognition today and in an AI-enhanced future. Front Artif Intell 2022;5:908261. [doi: 10.3389/frai.2022.908261] [Medline: 35910191]
2. Han Z, Wei B, Zheng Y, Yin Y, Li K, Li S. Breast cancer multi-classification from histopathological images with structured deep learning model. Sci Rep 2017;7(1):4172. [doi: 10.1038/s41598-017-04075-z]
3. Wang S, Chen A, Yang L, et al. Comprehensive analysis of lung cancer pathology images to discover tumor shape and boundary features that predict survival outcome. Sci Rep 2018;8(1):10393. [doi: 10.1038/s41598-018-27707-4]
4. Dumbrava EI, Meric-Bernstam F. Personalized cancer therapy-leveraging a knowledge base for clinical decision-making. Cold Spring Harb Mol Case Stud 2018 Apr;4(2):a001578. [doi: 10.1101/mcs.a001578] [Medline: 29212833]
5. Paul D, Sanap G, Shenoy S, Kalyane D, Kalia K, Tekade RK. Artificial intelligence in drug discovery and development. Drug Discov Today 2021 Jan;26(1):80-93. [doi: 10.1016/j.drudis.2020.10.010] [Medline: 33099022]
6. Askin S, Burkhalter D, Calado G, El Dakrouni S. Artificial intelligence applied to clinical trials: opportunities and challenges. Health Technol (Berl) 2023;13(2):203-213. [doi: 10.1007/s12553-023-00738-2] [Medline: 36923325]
7. Beck JT, Rammage M, Jackson GP, et al. Artificial intelligence tool for optimizing eligibility screening for clinical trials in a large community cancer center. JCO Clin Cancer Inform 2020 Jan;4:50-59. [doi: 10.1200/CCI.19.00079] [Medline: 31977254]
8. Alexander M, Solomon B, Ball DL, et al. Evaluation of an artificial intelligence clinical trial matching system in Australian lung cancer patients. JAMIA Open 2020 Jul;3(2):209-215. [doi: 10.1093/jamiaopen/ooaa002] [Medline: 32734161]
9. Meystre SM, Heider PM, Cates A, et al. Piloting an automated clinical trial eligibility surveillance and provider alert system based on artificial intelligence and standard data models. BMC Med Res Methodol 2023 Apr 11;23(1):88. [doi: 10.1186/s12874-023-01916-6] [Medline: 37041475]
10. Meystre SM, Heider PM, Kim Y, Aruch DB, Britten CD. Automatic trial eligibility surveillance based on unstructured clinical data. Int J Med Inform 2019 Sep;129:13-19. [doi: 10.1016/j.ijmedinf.2019.05.018] [Medline: 31445247]
11. Stubbs A, Uzuner Ö. New approaches to cohort selection. J Am Med Inform Assoc 2019 Nov 1;26(11):1161-1162. [doi: 10.1093/jamia/ocz174] [Medline: 31613362]
12. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. Implementation Sci 2009 Dec;4(1):50. [doi: 10.1186/1748-5908-4-50]
13. Jackson S, Panteli N. Trust or mistrust in algorithmic grading? An embedded agency perspective. Int J Inf Manage 2023 Apr;69:102555. [doi: 10.1016/j.ijinfomgt.2022.102555]
14. Carabantes M. Black-box artificial intelligence: an epistemological and critical analysis. AI & Soc 2020 Jun;35(2):309-317. [doi: 10.1007/s00146-019-00888-w]
15. Zednik C. Solving the black box problem: a normative framework for explainable artificial intelligence. Philos Technol 2021 Jun;34(2):265-288. [doi: 10.1007/s13347-019-00382-7]

Abbreviations

AI: artificial intelligence
CFIR: Consolidated Framework for Implementation Research

XSL•FO
RenderX

Edited by A Mavragani, J Sarvestan; submitted 28.01.25; peer-reviewed by CA Brunette, S Naicker; revised version received 22.09.25; accepted 23.09.25; published 30.10.25.

Please cite as:
Goodwin AJ, Armstrong SA, Ptak D, Catchpole K, Obeid JS, Heider PM
Leveraging Artificial Intelligence for Clinical Study Matching: Key Threads for Interweaving Data Science and Implementation Science
JMIR Form Res 2025;9:e71831
URL: https://formative.jmir.org/2025/1/e71831
doi:10.2196/71831

© Andrew James Goodwin, Sara Ann Armstrong, David Ptak, Kenneth Catchpole, Jihad S Obeid, Paul M Heider. Originally published in JMIR Formative Research (https://formative.jmir.org), 30.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Evaluating the Role of Artificial Intelligence in Breast Self-Examination

Shirish Nayyar, MBBS

General Internal Medicine, King's College Hospital NHS Foundation Trust, Denmark Hill, London, United Kingdom

Corresponding Author:
Shirish Nayyar, MBBS
General Internal Medicine, King's College Hospital NHS Foundation Trust, Denmark Hill, London, United Kingdom

Abstract

This viewpoint explores the role of conversational artificial intelligence (AI) in educating the public on breast self-examination, using an interaction with DeepSeek AI as an example, where 6 AI-generated responses to commonly asked questions were compared with guidelines published by organizations such as the World Health Organization. Although the AI provided clear, accessible, and evidence-aligned responses consistent with professional guidance, limitations included oversimplification and absence of multimedia resources. These findings suggest that AI can support public health education but should be complemented by physician oversight and evidence-based resources for responsible use.

(JMIR Form Res 2025;9:e82550) doi:10.2196/82550

KEYWORDS

breast; cancer; self assessment; AI; DeepSeek; artificial intelligence; breast cancer

With the growing accessibility of artificial intelligence (AI), the general population is increasingly engaging in conversations with AI chatbots, especially to understand topics related to their body and well-being, hence making it important for the medical community to evaluate and guide the dissemination of related content. This sparked my curiosity to explore one such interaction with DeepSeek AI to generate a general health advice session revolving around breast self-examination.

Breast cancer is a prevalent global health issue, ranking as the second most common cancer in women and the leading cause of cancer-related mortality in women [1]. A breast self-exam, while not a substitute for medical screening, helps one become familiar with how one's breasts look and feel, making it easier to detect unusual changes, potentially leading to seeking medical aid, early detection, and a better prognosis.

DeepSeek AI is an artificial research company whose aim is to provide cost-efficient but high-performance language models. It offers models at a lower cost than OpenAI's GPT-4 and is being increasingly used for medical research and development, hence my decision to use this model in a manner aligned with my purpose.

In my recent conversation with DeepSeek AI, I asked the bot 6 commonly asked questions about breast self-examination; the summary of the responses to the same are presented in Table 1.

XSL•FO
RenderX

Table . Commonly asked questions about breast self-examination and a summary of DeepSeek AI's responses.

Question	Response
Is self-examination for breast cancer important?	The AI[a] advised that self-examination plays a key role in early detection but should not replace clinical screenings such as mammograms or professional examinations. It helps individuals become aware of their breast health and detect changes like lumps, dimpling, or discharge between clinical exams. However, it has limitations, as it cannot identify all cancers—some of which may be detected earlier through imaging—and may occasionally cause false alarms due to normal hormonal variations.
How often should I screen myself for breast cancer?	The AI advised that I should perform a self-breast examination once a month to check for any changes, ideally 3 - 5 days after my period or on a fixed date if postmenopausal. It explained that doing this monthly helps me track what is normal for my breasts and detect any unusual changes early, such as lumps, dimpling, or nipple discharge. If I find a new hard lump, skin or nipple changes, or persistent swelling or pain, I should see a doctor, as many lumps can be benign. It also reminded me that self-exams are meant to complement, not replace, clinical screenings. I was recommended to follow standard screening guidelines—mammograms typically beginning at age 40 - 50 (earlier if high-risk) and clinical breast exams every 1 - 3 years in my 20s-30s, then annually after 40. For individuals at higher risk, such as those with a family history of breast cancer or BRCA gene mutations, earlier or more frequent imaging (like magnetic resonance imaging or ultrasound) may be recommended.
How do I screen myself for breast cancer?	The AI outlined that performing a breast self-exam correctly is essential for detecting early changes. It provided a structured, step-by-step guide comprising 3 components: visual check (observing in the mirror for changes in size, dimpling, or nipple appearance), manual check while lying down (using 3 finger pads with varying pressure in circular, vertical, or wedge motions covering the entire breast and armpit), and a shower check (repeating circular motions as soap aids movement). It distinguished between normal findings—soft, slightly lumpy tissue—and warning signs such as hard lumps, skin changes, or bloody discharge. The AI recommended not panicking if a lump is found, as most are benign, but to consult a doctor if the lump is new, hard, or persists beyond 1 menstrual cycle. It concluded by emphasizing that breast self-exams should be done monthly, ideally after the menstrual period, and should complement mammography (recommended from ages 40 - 50). For those at high risk, additional imaging such as magnetic resonance imaging or ultrasound may be considered.
Is there more than one way to screen myself for breast cancer?	DeepSeek AI outlined 3 main techniques for performing a breast self-exam, emphasizing that different methods may suit different individuals. The circular (spiral) method involves making small circular motions from the outer edge of the breast toward the nipple, ensuring coverage of the upper chest and underarm for systematic tissue assessment. The vertical strip method entails moving the fingers up and down in vertical lines across the breast, from the collarbone to the bra line and from the side to the underarm and is noted by the American Cancer Society (ACS) as the most thorough approach. The wedge (pie-slice) method involves dividing the breast into "hour-like" sections extending from the nipple outward, allowing a focused examination of each area. For better detection, DeepSeek AI advised performing the exam both while standing and lying down, using soap in the shower for smoother motion, and applying 3 levels of pressure—light, medium, and firm—to assess tissues at varying depths. The optimal time for conducting a breast self-exam is 3 - 5 days after menstruation ends. Individuals are encouraged to compare both breasts for symmetry, document any monthly changes, and begin mammograms at the age of 40, or earlier if they are at high risk.

XSL•FO
RenderX

Question	Response
At what age do I start to screen myself for breast cancer?	Breast self-awareness, which involves noticing how one's breasts normally look and feel, can begin in the teenage years or 20s, while formal monthly self-examinations are typically recommended from age 20. During the teens to 20s, individuals are encouraged to develop breast awareness, with monthly exams being optional unless high-risk. In the 20s to 30s, monthly self-exams may be considered, along with clinical exams every 1 to 3 years, and any changes should be reported promptly. From age 40 onward, self-exams should continue, with annual mammograms added between ages 40 and 50, and high-risk individuals may require additional imaging such as magnetic resonance imaging or ultrasound. Special circumstances include starting screening 10 years earlier than the age at which a family member was diagnosed if high-risk, consulting a doctor during pregnancy or breastfeeding for breast changes, and initiating earlier screening for those with genetic risks such as BRCA mutations or prior radiation exposure. Although self-examinations support early detection, they do not replace clinical evaluations, and screening schedules should be tailored based on individual risk factors.
What findings during self examination should make me see a doctor?	During breast self-exams, certain findings should prompt medical evaluation. Urgent warning signs include a new hard, immovable lump; skin changes such as dimpling, orange-peel texture, or unhealed sores; nipple changes like sudden inversion, bloody discharge, or crusting; unexplained persistent pain in one spot; and sudden asymmetric swelling or distortion. Normal variations include cyclical tenderness or lumpiness before the period, symmetrical texture changes, and small movable lumps that vary with the menstrual cycle. It is advisable to monitor changes through 1 to 2 menstrual cycles, document the size, location, and persistence, and seek prompt evaluation for rapid changes, combined symptoms such as a lump with discharge, or noncyclical symptoms. Most findings are benign, such as cysts or fibroadenomas, but follow-up with clinical examinations or mammograms is essential. High-risk patients may require additional imaging.

[a]AI: artificial intelligence.

This dialogue highlighted both the potential and the limitations of AI in patient education.

Overall, I felt that the responses to the questions asked were quite comprehensive and detailed, using language that the general population can understand and implement. The responses used words and techniques used by various credible agencies promoting breast self-examination, such as Breastcancer.org and the World Health Organization [2]. The responses also gave an easy-to-understand explanation about the technique of the examination, outlining various alternatives, in line with what is advised by medical professionals and institutions [3]. Additionally, every response stated the fact that breast self-examination was not a substitute for medical examination or investigations, but a complementary practice [4].

Points for further improvement that were clear to me included the provision of links to videos published by credible sites and organizations showing this practice, as 65% of the general population are considered to be visual learners [5]. There is also a risk of oversimplification of the process, especially in cases where the anatomy of breast tissue differs among individuals (such as those with dense breast tissue). There is also not enough emphasis on the fact that breast self-examination is not a substitute for medical examination and screening, which could be reinforced by mentioning certain facts, such as that self-examination, in some studies, is only linked to early detection, but not a reduction in mortality [6].

The utility of AI in medicine also has been raising some ethical and regulatory concerns; for instance, models trained on biased data may reproduce or amplify racial, gender, socioeconomic, or geographic disparities [7]. In addition, chatbots often collect conversational and usage data, bringing about concerns regarding privacy. Methods to mitigate these could be limitation of data collection (collect only what is necessary for the educational task) [8] alongside testing on real-world pilots and external datasets, with integration of these in the bot's database [9]. Another easy and reliable mitigation strategy would be the use of transparent documentation (data sheets) describing the training data, as well as obtaining clear consent for educational data reuse.

In conclusion, while AI-based conversations can aid in spreading awareness about and reinforcing health behaviors like breast self-screening, they must be complemented by physician oversight, public education campaigns, and accessibility to diagnostic services. Another way to evaluate this could be via the upcoming "evaluator bots," some of which have been shown to have the same efficacy as human evaluators [10]. Certain facts and percentages must be used to support AI-based responses and stress some important aspects rather than just mentioning them.

I hope this perspective contributes to the ongoing discourse on digital health tools and their responsible integration into medicine.

XSL•FO
RenderX

Acknowledgments

The generative AI tool DeepSeek AI was used to generate a conversation between a user and the AI regarding breast self-examination, as detailed in the study. The conversation outputs were subsequently reviewed and refined by the author.

Funding

No funding was provided for this work.

Data Availability

All data generated or analyzed during this study are included in this published article.

Conflicts of Interest

None declared.

References

1. Breast cancer. World Health Organization. 2025. URL: https://www.who.int/news-room/fact-sheets/detail/breast-cancer#:~:text=In%202022%2C%20there%20were%202.3,use%20and%20postmenopausal%20hormone%20therapy [accessed 2025-08-14]
2. American Cancer Society recommendations for the early detection of breast cancer. American Cancer Society. URL: https://www.cancer.org/cancer/types/breast-cancer/screening-tests-and-early-detection/american-cancer-society-recommendations-for-the-early-detection-of-breast-cancer.html [accessed 2025-11-18]
3. Henderson JA, Duffee D. Breast Examination Techniques: StatPearls - NCBI Bookshelf; 2023. URL: https://www.ncbi.nlm.nih.gov/books/NBK459179/ [accessed 2023-01-16]
4. Kösters JP, Gøtzsche PC. Regular self-examination or clinical examination for early detection of breast cancer. Cochrane Libr 2003;2010(1). [doi: 10.1002/14651858.CD003373]
5. Jawed S, Amin HU, Malik AS, Faye I. Classification of visual and non-visual learners using electroencephalographic alpha and gamma activities. Front Behav Neurosci 2019;13:86. [doi: 10.3389/fnbeh.2019.00086] [Medline: 31133829]
6. National Cancer Institute (US). Breast cancer screening (PDQ). In: PDQ Cancer Information Summaries: NCBI Bookshelf. URL: https://www.ncbi.nlm.nih.gov/books/NBK65906/ [accessed 2025-04-10]
7. Farhud DD, Zokaei S. Ethical issues of artificial intelligence in medicine and healthcare. Iran J Public Health 2021 Nov;50(11):i-v. [doi: 10.18502/ijph.v50i11.7600] [Medline: 35223619]
8. Konnoth F. AI and Data Protection Law in Health: Edward Elgar Publishing eBooks; 2024:111-129. [doi: 10.4337/9781802205657.ch07]
9. Chen F, Wang L, Hong J, Jiang J, Zhou L. Unmasking bias in artificial intelligence: a systematic review of bias detection and mitigation strategies in electronic health record-based models. J Am Med Inform Assoc 2024 Apr 19;31(5):1172-1183. [doi: 10.1093/jamia/ocae060] [Medline: 38520723]
10. Choo S, Yoo S, Endo K, Truong B, Son MH. Advancing clinical chatbot validation using AI-powered evaluation with a new 3-bot evaluation system: instrument validation study. JMIR Nurs 2025 Feb 27;8:e63058. [doi: 10.2196/63058] [Medline: 40014000]

Abbreviations

AI: artificial intelligence

Edited by A Bucher; submitted 17.08.25; peer-reviewed by A Gupta, R Mamidipally; revised version received 05.11.25; accepted 07.11.25; published 27.11.25.

Please cite as:
Nayyar S
Evaluating the Role of Artificial Intelligence in Breast Self-Examination
JMIR Form Res 2025;9:e82550
URL: https://formative.jmir.org/2025/1/e82550
doi:10.2196/82550

XSL•FO
RenderX

© Shirish Nayyar. Originally published in JMIR Formative Research (https://formative.jmir.org), 27.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Embedding Passive Monitoring Into Global Health and Longitudinal Patient Care

Angelika Papanicolaou[1]; Laura Gaetano[1], PhD; Olivia Yu[2], PhD; Graham Jones[3], PhD

[1]Neuroscience Development, Novartis AG, Basel, Switzerland

[2]Clinical Innovation, Novartis Pharmaceuticals, San Diego, CA, United States

[3]Tufts University Medical Center, 800 Washington Street, Boston, MA, United States

Corresponding Author:
Graham Jones, PhD
Tufts University Medical Center, 800 Washington Street, Boston, MA, United States

Abstract

A multitude of digital health tools have been developed to monitor, record, and predict health-related events in healthy subjects and patients. In clinical settings, although promising advances have resulted in near-term benefits, their use in longer-term studies is often limited due to the level of friction and burden imposed on the subject, often requiring active engagement by the patient with digital devices and/or its interfaces. Herein, we outline how smart ring technologies could form the anchor point for passive patient monitoring systems by offering a near-ideal compromise between device form factor and data capturing capacity. By using wireless technologies, such devices could form integral components of a hub-and-spoke health monitoring system, feeding data to cloud-based patient electronic health records and allowing push–pull actions through bidirectional communication. Such capabilities could have immediate utility in the longitudinal monitoring of patients diagnosed with slow progressing disease such as cardiovascular and neurodegenerative conditions. Moreover, if integrated through provisioned federated wireless networks, the technology could become components of global health care. To be successful, such a grand challenge would naturally require multiple technological, financial, and data privacy obstacles to be overcome. In support of this vision, we outline practical considerations for the development of such systems for specific applications and potential next steps for implementation.

(JMIR Form Res 2025;9:e81580) doi:10.2196/81580

KEYWORDS

remote monitoring; digital health technologies; global health; patient data; health data; smart ring

Introduction

Digital health technologies have developed rapidly over the past decade, fueled by dramatic improvements in the capability of miniaturized sensors coupled with ever increasing network and computing power [1]. This has resulted in a wide range of options, from commodity-scale consumer-worn products to elaborate clinical grade devices able to track and monitor multiple health-related signals [2]. Although a vast majority of these devices are deployed as performance trackers by healthy individuals motivated to monitor and assess specific personal targets (eg, step counting, heart rate, or sleep cycles), there is increasing use in clinical trials to enable study coordinators to monitor patient data through a combination of passive signals and patient-reported outcomes [3]. For short-term studies, this can be an extremely valuable approach, allowing the gathering of data in real-world environments and the potential for continual monitoring to augment episodic evaluations by physicians. For longer-term studies, however, patient adherence and persistence using the devices often wanes over time, rendering the data unreliable, and this is exacerbated in studies that utilize multiple devices [4]. For these reasons, there is growing interest in the use of fully passive monitoring devices, where following setup, little or no engagement is necessary [5]. For tightly controlled clinical studies, where costs are borne by sponsors, this has opened up new vistas for the development of highly sophisticated systems that provide high-precision clinical-grade data longitudinally [6]. Though attractive, for population-level studies and particularly those in developing nations, economic considerations likely preclude such options. In such instances, a compromise among device cost, simplicity, and features (ie, durability, precision, reporting frequency, and patient affinity) needs to be considered, and recent developments in smart ring technology render them well suited to address a variety of uses [7]. Representing a near-minimal form-factor for a personal wearable, contemporary smart rings now offer a wide range of health-sensing functions ranging from heart rate, actigraphy, sleep, and temperature derived from algorithms that are continually refined and updated [8-10]. Although smart rings are far from clinical grade standards, they can provide useful information; moreover, owing to form factors that blend well (thus not signaling to onlookers that the person may be wearing a medical device) and good battery life, which requires only periodic charging, they offer a viable and patient friendly approach to long-term health monitoring. Data have the potential

XSL•FO
RenderX

to be accessed from the cloud, via interfaces to smart phones, or through connection to a device reader, giving considerable flexibility to the level of patient engagement and burden. Extrapolating further, one could envision the deployment of such devices as fully passive autonomous monitors, providing cloud-based periodic or real-time updates on clinical trial metrics, or in the case of long-term monitoring of slow progressing diseases (eg, cardiometabolic or neurodegenerative), forming diarized components of the patient's electronic health record [11-13]. Herein, we outline how such systems could become embedded in personal health tracking, potential considerations for adoption, and mechanisms to incentivize deployment in various components of the health care ecosystem.

Discussion

Overview

The increased miniaturization of health-based electronic sensors potentially allows myriad monitoring instruments to be embedded in a single device. Among recent developments, consumer smart rings now host a number of sophisticated analytical tools, reflecting a compromise between cost and functionality (Figure 1). These typically include photoplethysmography/pulse shape measurement, temperature, and motion/inertial measurement unit sensors. There are also numerous prototype sensors in development that have shown promise for custom analyte detection, which might be incorporated into ring architectures including galvanic monitors [14], nano-tattoo–based circuits [15], and custom analytic cells [16]. For powering these devices, there is the potential to extend duty cycles through the use of solar cell technology and a wide range of emerging options for wireless data transmission from such devices. Although the commercial focus lies in the development of these products for the direct-to-consumer market, an attractive proposition could be to deploy a standardized version for use at scale in global health. This would require that data and device standards be identified, and a minimum set of sensors and functionality be established for patient monitoring. This could, for example, reflect current aggregated health care needs of national priority (eg, cardiovascular monitoring) or sensors fine tuned for regional needs including epidemic, pandemic, or episodic flares of specific diseases where tracking is important.

Figure 1. Hypothetical components of a ring-based passive monitor. DTC: direct to cloud HRV=heart rate variability; IMU: inertial measurement unit; PPG: photoplethysmography.

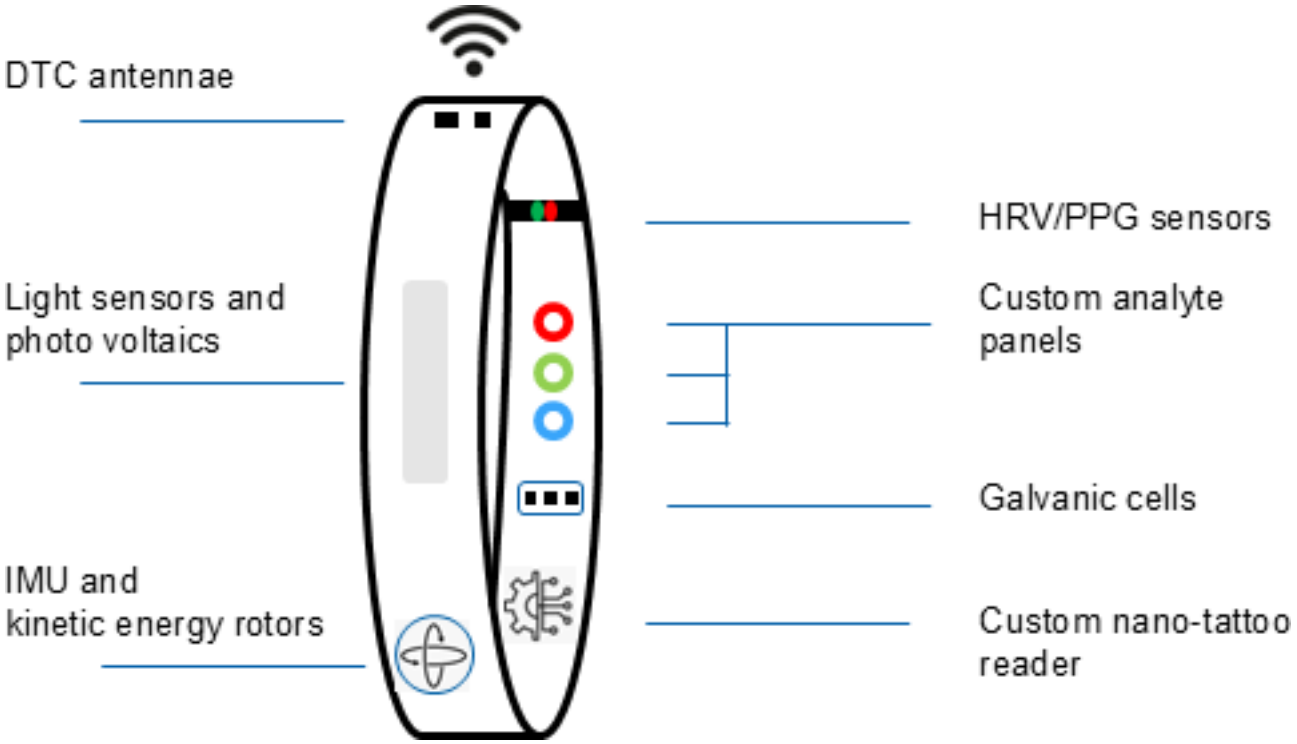

In terms of connectivity, various options exist for wireless-based communication between the device and a data hub/host. The most straightforward of these is via a Bluetooth link to a nearby smartphone, with the option to bounce data directly to the cloud (Figure 2). In regions where 4G and 5G networks are available, this could utilize direct-to-cloud connectivity through 802.11 Wi-Fi protocols on local area networks (2.4 GHz band). However, on the global scale where network infrastructure is limited, this might capitalize on low-power area networks (LoRaWAN), which have been proven successful in a number of nations, allowing data to be transmitted in defined packets at intervals [17]. Yet another option could be through connection to the new low-orbit satellite–based networks being developed such as Star Link or the Inmarsat system [18]. In theory, a multitude of options could be made available, with the possibility of a standardized smart ring capable of communicating with a data hub anywhere in the world at any time. Data collected or stored would be ideally linked to databases formatted for electronic health records to allow passive, real-time longitudinal monitoring of people affiliated with country-specific health systems, or patients as they show progress as engaged users of the health system. Examples of

XSL•FO
RenderX

such monitoring could include the monitoring of vital signs, providing emergency alerts to health providers (and tied to global system for mobile communication (GSM) location–based tracking for emergency responders), and monitoring of mobility (eg, during a pandemic crisis). Clearly, any approach of this nature would require compliance with health data privacy architectures including provisions regulated by bodies such as HIPAA (Health Insurance Portability and Accountability Act), GDPR (General Data Protection Regulation), and the World Health Organization but the benefits could justify such a challenge [19].

Figure 2. Wireless networking options through ring-based passive monitors. BLE: bluetooth low energy.

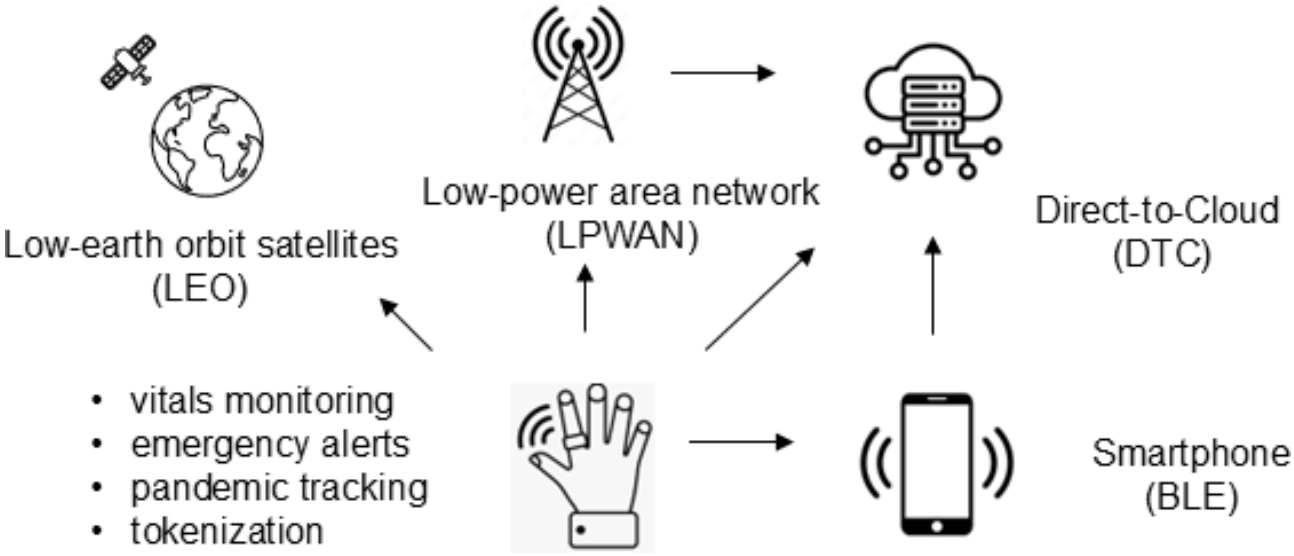

Incentivizing and Monetizing

The ultimate goal of developing a standardized, commoditized, affordable smart ring product capable of wireless communication across the globe presents a Herculean challenge. However, if introduced according to the above tenets, this could realistically have a marked beneficial impact on global health. As such, it may be attractive to one or more of the emerging cadre of billionaire philanthropists who have demonstrated their capacity to address societal health needs. One could envision the initial development costs being ultimately offset by sponsor nation states, who stand to achieve savings in health care costs while they elevate the standard of care of their citizens [20]. Another potential for revenue funding and also incentivizing subjects to embrace smart ring technology could be the tokenization of personal data for financial gain. Leaving aside the ethical considerations, it is conceivable that trial sponsors or epidemiologists would have interest in the monitoring of data from a large segment of the population for a fixed period of time. This is the basis of compensation for participants in traditional clinical trials, but the scale of the smart ring network globally could offer a quantum leap in opportunities. Such opportunities would need to be carefully structured to adhere to ethical guidelines, but the tactic of sale of personal data is already established in the consumer world, and this could offer a compensatory mechanism that some participants could presumably embrace [21]. Of course, in projecting associated costs globally, it will be important to assess the total economic burden of ownership, which will include device, supply chain, lifecycle updates, and the impact of e-waste. Although future-leaning, the reality of such a device eventually being realized is potentially high. Sensor, wireless, and power technologies are all advancing rapidly, and the key human variable for ring adherence (ie, finger diameter) is relatively narrow (typically 14 - 23 mm).

Technical Hurdles to Be Addressed

In order for the envisioned device to be realized, a number of technical challenges will need to be overcome with regard to device engineering.

Network Connectivity

Currently, smart rings transmit data packets via Bluetooth connectivity to smart devices, but the potential for other modalities exist. As described in Figure 2, these include direct-to-cloud, LoRaWAN, and potentially satellite-based systems. A key consideration relates to monetizing these connections, as data continuity is a central component of longitudinal monitoring. While many people enjoy free access to Wi-Fi networks in residences and commercial settings, it may become necessary to provision sponsored services in some situations. For example, patients agreeing to share vital health data to a research organization or sponsor in return for dedicated cellular or Wi-Fi access.

Minimum Viable Sensor Units

A key consideration would be the agreement on a minimum set of measures for patient tracking, and the sensors required to enable this. For example, this could be pulse rate, temperature, and actigraphy in a standard battery of assessments. Data could be directly uploaded or stored locally on a miniature hard drive, and then uploaded when network connectivity (low-power wide-area network or bluetooth) was optimal. Given the cost benefits of having a unified model of the device, one solution to facilitate this yet allow additional options to be utilized by other users could be design of a base unit with latent options built in that are only activated on demand. Such product line

XSL•FO
RenderX

engineering is often used in the automotive industry, where wiring circuitry is built in on the assembly line independent of final features and options ultimately selected by the customer [22]. In this case, the base and optional features on the ring menu could be selected by state health program needs. Other considerations relate to accuracy and reliability of the sensors themselves, particularly if medical decision-making is intended. Clinical-grade outputs should be a long-term objective for such devices, ideally independent of environmental challenges, such as temperature and humidity fluxes, shock sensitivity, and power surges.

Power Supply and Consumption

Advances in the efficiency of rechargeable battery systems continues apace as does miniaturization and improvement in charging cycle intervals [23]. Ideally, the ring charging interval would be >1 month, with the potential either for a cradle-type device or though wireless chargers that are becoming more widely available. Supplementing these systems could be surface-mounted solar-cells that harvest energy for emergency use on demand, for storage on mini-capacitors, or potentially recharging the main battery stack [24]. Other energy-harvesting technologies could be incorporated, including kinetic energy from hand motion, which might be possible through inertial measurement unit sensors [25]. In parallel, refined approaches to circuit design and engineering could serve to reduce power consumption on the device, for example, resting/standby modes that capture and transmit data at defined intervals.

Developing Long-term Product Affinity

A key consideration to encourage long-term use in participants will be to arrive at form factors that are functional, esthetically pleasing, and culturally adaptable. Obvious considerations include ring size, material, colors, and the potential for personalization as witnessed in the smartphone market. In theory, and assuming the monetary value was low, the trading-in of rings, for example, at different age points, when BMI changes, or owing to personal preference, could be possible in a mass market, but would require (1) the transaction-style upload of personal health data on demand, and (2) an adequate refurbishment program including the sterilization of the ring. Learning from the smartphone industry would be insightful as a number of repurposing and recycling strategies have been reported [26]. Another means to increase product affinity would be to incorporate dual purpose functionality, for example, adding a security layer that confirmed the person's identity and thus allowed user access or entry provisions, including entry to buildings [27]. Similarly, the potential for GSM/GPS location tracking functionality could be appealing for family members, school age participants, or care givers.

Regulatory Frameworks

A major consideration for any device used for health tracking will be regulatory compliance. Well-defined pathways have been outlined by health authorities including the United States Food and Drug Administration and The European Medicines Agency on medical devices and the software used to support them. The degree of classification relates to the intended use of data derived, with a naturally high bar for devices used to inform medical decision-making (Figure 3) [28]. Additionally, efforts are underway to utilize the potential of digital devices in support of home care and monitoring, and it can be expected that regulatory policies will evolve as these technologies advance and become more pervasive [29]. Likewise, the sovereignty and privacy of data generated from such devices are subject to governance and regulations, which are often country-specific and require end users to adhere to strict guidelines on capture, storage, use, and accessibility. The deployment of monitoring devices will need to be mindful of current and future regulations, particularly if longitudinal use is desired [30]. Finally, the interoperability frameworks for wireless devices represents another evolving consideration for intended use, with myriad protocols such as FHIR (Fast Healthcare Interoperability Resources), SNOMED CT (Systematized Nomenclature of Medicine – Clinical Terms), and IEEE (The Institute of Electrical and Electronics Engineers) 11073 governing deployment including cross-border applications [31].

Figure 3. Challenges and opportunities for scaled deployment of digital passive health monitors.

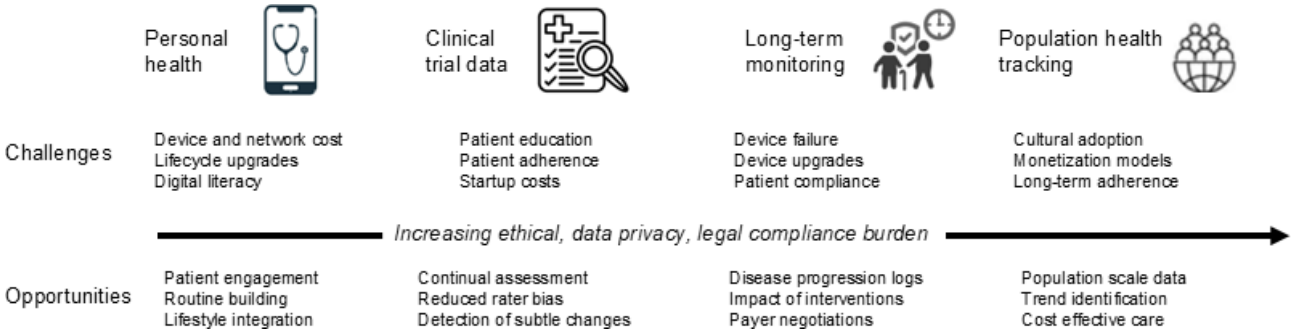

The Pathway to Development

Although a multitude of obstacles need to be addressed, there is evident potential for the deployment of smart ring technologies in long-term health monitoring. In terms of the pathways for development, each of the potential use cases envisioned presents unique challenges and opportunities and the ethical, legal, and data privacy burdens for each are expected to reflect scale and intended use (Figure 3). Nonetheless, given striking advances in health technology, computational and communication networks, and health literacy, it seems inevitable that such frameworks will evolve naturally. In order to propel such opportunities forward, we advocate enhanced dialog between the technology community, health advocacy groups, global health organizations, and philanthropic leaders. We encourage readers of this journal to jump start and advance this

XSL•FO
RenderX

dialog as members of the citizen science community, as the impact is potentially significant. Given the evident potential for such a device to play a transformative role in health care reform, we believe this call-to-arms for the device engineering community is timely and warranted and look forward to active dialog to progress this vision.

Conclusions

Smart ring technology has advanced dramatically over the past decade, and when coupled with new wireless connectivity, offers the potential for use in patient monitoring at individual- through to population-level scales. We posit that by arriving at common standards for design, function, and manufacture, a singular device could be identified for global deployment in health care at a mass scale. To enable such a vision will require active and sustained dialog with myriad stakeholders, seed funding provision from philanthropists, and endorsement from nation member states. Mechanisms for monetization can be envisioned to sustain such an initiative and integrate it into standard health care, and the near- and long-term benefits of this technology could be substantial. The momentum for this may also come through political agendas. A recent vision articulated by the US Health and Human Services Secretary that all Americans wear a digital health device within the next four years may help fuel interest in the approach [32], and the medical community is already providing input on the implementation [33].

Acknowledgments

References cited were selected using search engines (SciFinder, Google Scholar) on the basis of scientific relevance, with priority afforded to most recent exemplars. Generative AI was not used in any portion of the manuscript.

Conflicts of Interest

LG and GJ are employees of Novartis Pharmaceuticals and may hold common stock in the company. Views represented are those of the authors.

References

1. Yeung AWK, Torkamani A, Butte AJ, et al. The promise of digital healthcare technologies. Front Public Health 2023;11:1196596. [doi: 10.3389/fpubh.2023.1196596] [Medline: 37822534]
2. Mittermaier M, Venkatesh KP, Kvedar JC. Digital health technology in clinical trials. NPJ Digit Med 2023 May 18;6(1):88. [doi: 10.1038/s41746-023-00841-8] [Medline: 37202443]
3. Collado-Borrell R, Escudero-Vilaplana V, Narrillos-Moraza Á, Villanueva-Bueno C, Herranz-Alonso A, Sanjurjo-Sáez M. Patient-reported outcomes and mobile applications. A review of their impact on patients' health outcomes. Farm Hosp 2022 Apr 27;46(3):173-181. [Medline: 36183211]
4. Sauter C, Diemar J, Terebaite V. Usability of digital health devices in clinical trials. Expert Rev Pharmacoecon Outcomes Res 2024 Dec;24(10):1097-1099. [doi: 10.1080/14737167.2024.2384545] [Medline: 39076077]
5. Popp Z, Low S, Igwe A, et al. Shifting from active to passive monitoring of Alzheimer disease: the state of the research. J Am Heart Assoc 2024 Jan 16;13(2):e031247. [doi: 10.1161/JAHA.123.031247] [Medline: 38226518]
6. Kabelac Z, Tarolli CG, Snyder C, et al. Passive monitoring at home: a pilot study in Parkinson disease. Digit Biomark 2019;3(1):22-30. [doi: 10.1159/000498922] [Medline: 32095766]
7. Fiore M, Bianconi A, Sicari G, et al. The use of smart rings in health monitoring—a meta-analysis. Appl Sci (Basel) 2024;14(23):10778. [doi: 10.3390/app142310778]
8. Kim J, Chang SA, Park SW. First-in-human study for evaluating the accuracy of smart ring based cuffless blood pressure measurement. J Korean Med Sci 2024 Jan 15;39(2):e18. [doi: 10.3346/jkms.2024.39.e18]
9. Krishna HS, Starwin A, Starvin A. Smart rings in health monitoring: technological advances and clinical applications of finger-worn biometric devices. IOSR Journal of Dental and Medical Sciences 2025;24(9):26-28. [doi: 10.9790/0853-2409012628]
10. Wang Z, Yu R, Wang X, et al. Computing with smart rings: a systematic literature review. arXiv. Preprint posted online on Feb 4, 2025. [doi: 10.48550/arXiv.2502.02459]
11. Moon DU, Lee Y, Lütt A, Lee S, Lee E. Adjunctive smart ring monitoring during digital cognitive behavioral therapy for insomnia. Sci Rep 2025 Oct 30;15(1):37934. [doi: 10.1038/s41598-025-24312-0] [Medline: 41168443]
12. Azim Zadegan M, Aromaa E, Montonen T, Sahlström R, Eriksson P, Leinonen V. Smart rings in remote monitoring at home: older patients' perceptions and expectations. Health Expect 2025 Oct;28(5):e70407. [doi: 10.1111/hex.70407] [Medline: 40888147]
13. Linh VTN, Han S, Koh E, Kim S, Jung HS, Koo J. Advances in wearable electronics for monitoring human organs: bridging external and internal health assessments. Biomaterials 2025 Mar;314:122865. [doi: 10.1016/j.biomaterials.2024.122865] [Medline: 39357153]
14. Volpes G, Valenti S, Genova G, et al. Wearable ring-shaped biomedical device for physiological monitoring through finger-based acquisition of electrocardiographic, photoplethysmographic, and galvanic skin response signals: design and preliminary measurements. Biosensors (Basel) 2024 Apr 20;14(4):205. [doi: 10.3390/bios14040205] [Medline: 38667198]

XSL•FO
RenderX

15. Pirrera A, Giansanti D. Smart tattoo sensors 2.0: a ten-year progress report through a narrative review. Bioengineering (Basel) 2024 Apr 13;11(4):376. [doi: 10.3390/bioengineering11040376] [Medline: 38671797]
16. Gao F, Liu C, Zhang L, et al. Wearable and flexible electrochemical sensors for sweat analysis: a review. Microsyst Nanoeng 2023;9:1. [doi: 10.1038/s41378-022-00443-6] [Medline: 36597511]
17. Jones GB, Bryant A, Wright J. Democratizing global health care through scalable emergent (beyond the mobile) wireless technologies. JMIR Biomed Eng 2022 Feb 11;7(1):e31079. [doi: 10.2196/31079] [Medline: 38875689]
18. Shaengchart Y, Kraiwanit T. Starlink satellite project impact on the Internet provider service in emerging economies. Research in Globalization 2023 Jun;6:100132. [doi: 10.1016/j.resglo.2023.100132]
19. Xiang D, Cai W. Privacy protection and secondary use of health data: strategies and methods. Biomed Res Int 2021;2021:6967166. [doi: 10.1155/2021/6967166] [Medline: 34660798]
20. Winders WT, Garbern SC, Bills CB, et al. The effects of mobile health on emergency care in low- and middle-income countries: a systematic review and narrative synthesis. J Glob Health 2021 Apr 3;11:04023. [doi: 10.7189/jogh.11.04023] [Medline: 33828846]
21. Maher MA, Khan IA. From sharing to selling: challenges and opportunities of establishing a digital health data marketplace using blockchain technologies. Blockchain Healthc Today 2022;5. [doi: 10.30953/bhty.v5.184] [Medline: 36779023]
22. Wozniak L, Clements P. How automotive engineering is taking product line engineering to the extreme. 2015 Jul 20 Presented at: SPLC ’15; Nashville Tennessee p. 327-336. [doi: 10.1145/2791060.2791071]
23. Sarmah S, Kakati BK, Deka D. Recent advancement in rechargeable battery technologies. WIREs Energy & Environment 2023 Mar;12(2):e461. [doi: 10.1002/wene.461]
24. Kim KA, Bagci FS, Dorsey KL. Design considerations for photovoltaic energy harvesting in wearable devices. Sci Rep 2022 Oct 28;12(1):18143. [doi: 10.1038/s41598-022-22232-x] [Medline: 36307451]
25. Mahboubi H, Sayrafian K, Aghdam AG. Adaptive maximization of harvested kinetic energy for small wearable medical sensors. 2023 Presented at: ICC 2023 - IEEE International Conference on Communications; Rome, Italy p. 4441-4445. [doi: 10.1109/ICC45041.2023.10279411]
26. Maldonado-Romo J, Méndez JI, Rojas M, Montesinos L, Ponce P, Molina A. Circular economy strategy based on digital technologies and the S4 framework: a case study on smartphones. International Journal of Sustainable Engineering 2024 Dec 31;17(1):295-305. [doi: 10.1080/19397038.2024.2336008]
27. Behera SK. Chipless RFID sensors for wearable applications: a review. IEEE Sensors J 2022 Jan 15;22(2):1105-1120. [doi: 10.1109/JSEN.2021.3126487]
28. Shi L, Xuan D, Jakovljevic M. A review on the evolving environment of medical device real-world evidence regulation on market access in the USA. Cost Eff Resour Alloc 2024 Oct 25;22(1):75. [doi: 10.1186/s12962-024-00582-9] [Medline: 39456032]
29. Imagine home as a health care hub through the idea lab. Idea Lab. URL: https://www.fda.gov/medical-devices/home-health-care-hub/idea-lab [accessed 2025-11-12]
30. Barbaria S, Jemai A, Ceylan H, Muntean RI, Dergaa I, Boussi Rahmouni H. Advancing compliance with HIPAA and GDPR in healthcare: a blockchain-based strategy for secure data exchange in clinical research involving private health information. Healthcare (Basel) 2025 Oct 15;13(20):2594. [doi: 10.3390/healthcare13202594] [Medline: 41154272]
31. Martins D, Lewerenz S, Carmo A, Martins H. Interoperability of telemonitoring data in digital health solutions: a scoping review. Front Digit Health 2025;7:1502260. [doi: 10.3389/fdgth.2025.1502260] [Medline: 40330872]
32. RFK Jr wants a wearable on your wrist. Portico. URL: www.politico.com/newsletters/future-pulse/2025/06/24/rfk-jr-wants-a-wearable-on-your-wrist-00419190 [accessed 2025-11-12]
33. Will RFK Jr.’s wearables push work? Becker’s Health IT. URL: www.beckershospitalreview.com/healthcare-information-technology/digital-health/will-rfk-jr-s-wearables-push-work/ [accessed 2025-11-12]

Abbreviations

FHIR: Fast Healthcare Interoperability Resources
GDPR: General Data Protection Regulation
GSM: global system for mobile communication
HIPAA: Health Insurance Portability and Accountability Act
LoRaWAN: low-power area network
SNOMED CT: Systematized Nomenclature of Medicine – Clinical Terms

XSL•FO
RenderX

Edited by A Schwartz, M Balcarras; submitted 30.07.25; peer-reviewed by D Rajendra Prasad, S Spethmann; revised version received 11.12.25; accepted 11.12.25; published 30.12.25.

Please cite as:
Papanicolaou A, Gaetano L, Yu O, Jones G
Embedding Passive Monitoring Into Global Health and Longitudinal Patient Care
JMIR Form Res 2025;9:e81580
URL: https://formative.jmir.org/2025/1/e81580
doi:10.2196/81580

© Angelika Papanicolaou, Laura Gaetano, Olivia Yu, Graham Jones. Originally published in JMIR Formative Research (https://formative.jmir.org), 30.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Applications of Self-Driving Vehicles in an Aging Population

Sara Shu[1,2], DO; Benjamin K P Woo[2,3,4], MD

[1]Department of Community Internal Medicine, Geriatrics and Palliative, Mayo Clinic, 200 First Street SW, Rochester, MN, United States
[2]University of California, Los Angeles, Los Angeles, CA, United States
[3]Asian American Studies Center, University of California, Los Angeles, Los Angeles, CA, United States
[4]Chinese American Health Promotion Laboratory, University of California, Los Angeles, Los Angeles, CA, United States

Corresponding Author:
Sara Shu, DO
Department of Community Internal Medicine, Geriatrics and Palliative, Mayo Clinic, 200 First Street SW, Rochester, MN, United States

Abstract

The proportion of older adult drivers is increasing and represents a growing population that must contemplate reducing driving and eventually stopping driving. The advent of self-driving vehicles opens vast possibilities with practical and far-reaching applications for our aging population. Advancing technologies in transportation may help to overcome transportation barriers for less mobile individuals, transcend social and geographical isolation, and improve resource and medical access. Herein, we propose various applications and benefits that self-driving vehicles have in maintaining independence and autonomy specifically for our aging population to preserve aging.

(JMIR Form Res 2025;9:e66180) doi:10.2196/66180

KEYWORDS

self-driving; driverless; driver; autonomous vehicles; car; transportation; mobility; travel; vehicle; driving; artificial intelligence; gerontology; geriatric; older; elderly; aging; healthy aging; older adult; autonomy; independence; aging in place; health equity

Introduction

In the United States, driving remains a symbol of independence and is an important means to maintain a connection to and identity within the community. For many community-dwelling older adults, private vehicles remain the primary mode of transportation. In the United States, there are over 55 million licensed automobile drivers over 65 years of age, comprising over 25% of the total driving population [1]. Motor vehicle collisions are among the most common causes of nonfatal and fatal injuries in the United States in older adults [2,3]. Older adults have higher mortality compared to a younger population with similar motor vehicle collision-related injuries [4]. The association between increased dangers of driving and collisions in the older adult population is well-known. With increasing age comes impairments in vision and visuospatial and reaction times, all of which make it more dangerous and difficult to navigate safely [5]. Primary care providers and associated interdisciplinary teams are in a unique position to assess the fitness of older adults to drive safely and provide alternative transportation recommendations [6,7]. Developments in technology may also provide tools to assist with the assessment of driving safety. For example, speech data analysis during interactions with voice assistants has been suggested to accurately predict future accident experiences in older drivers [8].

Environmental factors such as congestion of roads and the distance to the locations of goods and services can influence driving reduction and cessation in older adults [9]. Even medical events requiring hospitalization have been shown to correlate with immediate or eventual driving reduction or cessation, which ultimately compromises the autonomy of individuals [10]. Moreover, driving cessation has been associated with declines in general health and mental, physical, social, and cognitive function, and has even been associated with increased risk of admission to long-term care facilities and mortality [11,12]. In this paper, we aim to provide an updated review of current and proposed applications of self-driving vehicles, particularly through the integration and application of mobility as a service to assist with prolonging the autonomy of aging persons, as well as review the limitations and future directions that have yet to be explored.

Towards Age-Friendly Communities

Over the past several decades, there has been a push to develop communities and city infrastructure to be more supportive of an aging population in areas such as housing, transportation, and access to health care and social services. There is great heterogeneity in the driving challenges and concerns of older adult drivers in different geographical settings: urban, suburban, and rural [13]. Differences in community infrastructure have shown varying densities in the incidence of fatal crashes for older drivers. For example, the majority of fatal crashes occurred

XSL•FO
RenderX

on rural roads, while suburban and urban roads conferred a different likelihood of crash location and mechanisms, such as multivehicle crashes, as well as those within intersections versus multilane roads [14]. The movement towards age-friendly communities has driven the development of the World Health Organization Global Network for Age-Friendly Cities and Communities to provide a framework of considerations when developing age-friendly environments [15]. These include the development of technologies, services, and systems to transcend physical and social barriers that can promote the inclusion and continued integration and engagement of older adults in the community. As there are increasing needs in the rapidly growing older adult population, there will be an increased need to develop innovative strategies to meet them.

Self-driving vehicles hold promise in maintaining autonomy and maintaining independence to allow living in and remaining at home for as long as possible. When the option of driving autonomy is exhausted, one begins to recruit assistance from social networks comprising relatives, neighbors, and friends. However, as social networks organically decrease among an aging population, socially isolated individuals may not have the luxury to call on such assistance. The reliance on public transportation as a primary mode of transportation is not always feasible nor uniformly available across the United States. Extreme temperatures and climates during the winter and summer months may furthermore be a barrier for older adults to use public transportation. Volunteer transportation programs, senior services, or service operations that charge a small fee are available but are met with similar limitations in access depending on the local infrastructure and individual socioeconomic status. Therefore, finding a reliable and affordable means of transportation remains a great challenge in our aging society.

Applications of Self-Driving Vehicles

Self-driving (also known as autonomous) vehicles present a promising alternative and emerging transportation option for our older adult population [16,17]. Semiautonomous vehicles already have applications for use as cognitive-assistive devices for community-dwelling older adults owing to features such as onboard navigation, parking assistance, lane departure, and collision avoidance technologies [18]. As fully autonomous vehicle technology continues to develop, the applications are far reaching. By freeing the need to engage another community-dwelling adult or one in the social network, the availability of transportation options will inadvertently increase. In metropolitan areas where self-driving vehicles have been launched as mobility services, many individuals are already using these mobility services for commuting to and from work. The use of this technology in the older population can help to close the divide in populations otherwise socially isolated and enable them to once again be independent through their community and be able to engage in the community-based or group activities they once enjoyed. It would significantly relieve the caregiver's burden of having to provide transportation to and from medical appointments, the grocery store, barber shop, or other social functions and leisure activities. Similar mobility service efforts are seen in the European Union and Asia, where working towards cooperative, connected, and automated mobility is being optimized to embolden digital connectivity between vehicles, the transportation infrastructure, and other road users. These so-called cooperative intelligent transport systems are the foundation for the widespread use and applications of these technologies to provide services for otherwise homebound community-dwelling older adults with limitations that would preclude them from independent and safe utilization of their own motor vehicles.

As the market is already saturated with home food and medical supply delivery services, self-driving vehicles can eventually be equipped to undertake these tasks, thereby relieving the current workforce to engage in other areas of need. As much of the home delivery services rely on assistance from community-dwelling adults, these individuals would now have increased caregiving capacities for their aging community and loved ones, to serve in capacities where technology is unable to provide assistance. Some of the country's largest ride-share companies are already pioneering the driverless vehicle space by partnering with home-based care agencies [19]. By increasing the capacity and affordability of these services, the use of these technologies can help to provide solutions to building more equitable age-friendly, disability-friendly communities.

Similar advantages of self-driving vehicles can be extended to mobility-isolated persons or those with disabilities. In order to optimize accessibility to these populations, car designs should be optimized to be age-friendly and wheelchair accessible, such as being equipped with a ramp or capacity for the vehicle to lower itself to the curbside. Vehicles such as Toyota Sienna's "Autono-MaaS" (Autonomous-Mobility as a Service) vehicles are already incorporating these features in partnership with transportation companies to provide increased transportation options for those with disabilities. Companies such as May Mobility and Via have launched new and free self-driving shuttle services called "Accessibili-D" for the local inhabitants of Detroit, Michigan aged 65 years and older or those with disabilities [20]. These on-demand shuttle services have deployed similar services in Texas, Arizona, and throughout Michigan and have expanded into retirement communities. The "goMARTI" (Minnesota's Autonomous Rural Transit Initiative) in Grand Rapids, Minnesota, was the first pilot to provide free, on-demand rides with the American Disabilities Act (ADA)-compliant vehicles to improve mobility to those otherwise mobility-isolated in rural areas and during inclement weather [21].

With the continued development of artificial intelligence (AI) technology, medical capabilities could eventually be installed in autonomous vehicles to be able to detect and respond to medical emergencies. They could be equipped with emergency medical supplies in addition to sensors and technology that could detect the need to redirect to the nearest medical center, connect with a remote telemedicine provider, or provide basic equipment such as oxygen or epinephrine pens when indicated. Vehicles could even be engineered to detect mood fluctuations including depressed, anxious, or agitated states, and be equipped with robotic animal interventions for temporal therapy or connect individuals to telemedicine assistance or crisis response teams. These so-called autonomous mobile clinics equipped

XSL•FO
RenderX

with AI software, medical diagnostic detection, equipment, and telemedicine capabilities could reshape and facilitate prompt medical care and can even help to bring medical care to once-difficult-to-reach, more rural populations [22,23]. Epidemiologically conscious uses of self-driving vehicles have already been piloted in the days of the COVID-19 pandemic, such as contactless goods transportation and transporting patients in isolation to prevent exposure to other high-infection risks. Nevertheless, such widespread integration and the adoption of clinical applications of AI have yet to be subject to thorough clinical validation and approval by the appropriate regulatory authorities [24].

Challenges and Limitations

The limitations of self-driving vehicles are many, and as far as our imaginations can carry us considering the tremendous potential these technologies have to offer, so too do foreseen issues and concerns arise. Hypothetical issues of the unknown have guided various expert opinions. Driving principles in regard to road user safety are expected to be upheld; however, there are countless scenarios by which these self-driving vehicles must navigate tensions between obeying the traffic code and avoiding human harm. This highlights the importance of the translation of these frameworks into the engineering of such self-driving vehicles to guarantee consistency and safety in behavior that is driven by automated technology [25]. Moreover, in addition to predicting the unpredictability of road interactions, passenger safety is needed as well. For example, in caring for an older adult with dementia, one with limited mobility, or one with mental health concerns, it is important to have the assurance of safety for the otherwise vulnerable adult, such as precluding drop-off into an unfavorable environment. For example, a person with compromised cognitive abilities may not recognize that it may be unsafe to be outside for extended periods of time in a heatwave, or a person with impulsivity may not comprehend the dangers of going to a casino. Such ethical dilemmas tread the fine line between autonomy and paternalism, especially when such technology works to expand the independence and autonomy of an increasingly diverse, vulnerable population. Whether autonomous vehicles should have the ability to override the decisions of their passengers under the pretense of responsibility for safety over vulnerable populations or risk the repercussions for invading the right to autonomy has yet to be explored.

The implications of AI software on passenger privacy in autonomous vehicles will always be a concern. As the software of these autonomous vehicles could potentially be targets of cyberattacks, the kind of protections that would be placed on ensuring the safety of its passengers is important. Moreover, the use of these technologies would result in varying perceptions and attitudes from different potential users, all of which are also important. Various small-scale studies have captured such perceptions in the older adult population, including in persons with dementia [26-31]. It is important, therefore, to address targeted user populations' concerns about the cost, functionality, performance, and trustworthiness of the automated vehicles to suit their needs. Social determinants of health can also affect the accessibility of these technologies and have the potential to further marginalize underserved communities and populations; therefore, affordability needs to be considered. By partnering with local, state, and national departments of transportation and agencies on aging, solutions are needed for the promotion of equity, accessibility, and inclusion of all the mobility needs of older adults as well as individuals with disabilities and caregivers, regardless of the socioeconomic status.

Conclusion

In summary, the advent of self-driving, or autonomous, vehicles can have beneficial effects on how individuals are able to age while ensuring autonomy. We are presently at a unique crossroads of personalizing this developing technology to facilitate and optimize independence for the older adult population and maintain autonomy for as long as possible. Engaging the proper stakeholders may enable the rapid development of these technologies to promote a more age-friendly society.

Conflicts of Interest

None declared.

References

1. Federal Highway Administration, Department of Transportation (US). Distribution of licensed drivers 2022. : FHWA URL: https://www.fhwa.dot.gov/policyinformation/statistics/2022/ [accessed 2024-06-07]
2. Platts-Mills TF, Hunold KM, Esserman DA, Sloane PD, McLean SA. Motor vehicle collision-related emergency department visits by older adults in the United States. Acad Emerg Med 2012 Jul;19(7):821-827. [doi: 10.1111/j.1553-2712.2012.01383.x] [Medline: 22724382]
3. Dellinger AM, Stevens JA. The injury problem among older adults: mortality, morbidity and costs. J Safety Res 2006;37(5):519-522. [doi: 10.1016/j.jsr.2006.10.001] [Medline: 17140602]
4. Perdue PW, Watts DD, Kaufmann CR, Trask AL. Differences in mortality between elderly and younger adult trauma patients: geriatric status increases risk of delayed death. J Trauma 1998 Oct;45(4):805-810. [doi: 10.1097/00005373-199810000-00034] [Medline: 9783625]
5. Anstey KJ, Windsor TD, Luszcz MA, Andrews GR. Predicting driving cessation over 5 years in older adults: psychological well-being and cognitive competence are stronger predictors than physical health. J Am Geriatr Soc 2006 Jan;54(1):121-126. [doi: 10.1111/j.1532-5415.2005.00471.x] [Medline: 16420208]

XSL•FO
RenderX

6. Pomidor A, editor. Clinician's Guide to Assessing and Counseling Older Drivers, 4th edition: The American Geriatrics Society; 2019. URL: https://geriatricscareonline.org/toc/clinicians-guide-to-assessing-and-counseling-older-drivers-4th-edition/B047/ [accessed 2024-04-08]
7. Hill LJN, Pignolo RJ, Tung EE. Assessing and counseling the older driver: a concise review for the generalist clinician. Mayo Clin Proc 2019 Aug;94(8):1582-1588. [doi: 10.1016/j.mayocp.2019.03.023] [Medline: 31378232]
8. Yamada Y, Shinkawa K, Kobayashi M, et al. Using speech data from interactions with a voice assistant to predict the risk of future accidents for older drivers: prospective cohort study. J Med Internet Res 2021 Apr 8;23(4):e27667. [doi: 10.2196/27667] [Medline: 33830066]
9. Vivoda JM, Heeringa SG, Schulz AJ, Grengs J, Connell CM. The influence of the transportation environment on driving reduction and cessation. Gerontologist 2017 Oct 1;57(5):824-832. [doi: 10.1093/geront/gnw088] [Medline: 27342439]
10. Gaulton TG, Neuman MD, Brown RT, Betz ME. Association of hospitalization with driving reduction and cessation in older adults. J Am Geriatr Soc 2021 Aug;69(8):2231-2239. [doi: 10.1111/jgs.17178] [Medline: 33864381]
11. Chihuri S, Mielenz TJ, DiMaggio CJ, et al. Driving cessation and health outcomes in older adults. J Am Geriatr Soc 2016 Feb;64(2):332-341. [doi: 10.1111/jgs.13931] [Medline: 26780879]
12. Qin W, Xiang X, Taylor H. Driving cessation and social isolation in older adults. J Aging Health 2020 Oct;32(9):962-971. [doi: 10.1177/0898264319870400] [Medline: 31431103]
13. Payyanadan RP, Lee JD, Grepo LC. Challenges for older drivers in urban, suburban, and rural settings. Geriatrics (Basel) 2018 Mar 22;3(2):14. [doi: 10.3390/geriatrics3020014] [Medline: 31011058]
14. Zwerling C, Peek-Asa C, Whitten PS, Choi SW, Sprince NL, Jones MP. Fatal motor vehicle crashes in rural and urban areas: decomposing rates into contributing factors. Inj Prev 2005 Feb;11(1):24-28. [doi: 10.1136/ip.2004.005959] [Medline: 15691985]
15. The global network for age-friendly cities and communities: looking back over the last decade,looking forward to the next. : World Health Organization; 2018 URL: https://www.who.int/publications/i/item/WHO-FWC-ALC-18.4 [accessed 2025-04-08]
16. Abdi S, de Witte L, Hawley M. Emerging technologies with potential care and support applications for older people: review of gray literature. JMIR Aging 2020 Aug 11;3(2):e17286. [doi: 10.2196/17286] [Medline: 32780020]
17. Shu S, Woo BKP. Use of technology and social media in dementia care: current and future directions. World J Psychiatry 2021 Apr 19;11(4):109-123. [doi: 10.5498/wjp.v11.i4.109] [Medline: 33889536]
18. Knoefel F, Wallace B, Goubran R, Sabra I, Marshall S. Semi-autonomous vehicles as a cognitive assistive device for older adults. Geriatrics (Basel) 2019 Nov 16;4(4):63. [doi: 10.3390/geriatrics4040063] [Medline: 31744041]
19. Donlan A, Home Health Care News. Driverless cars finding new lane in home-based care. 2020 Mar 1. URL: https://homehealthcarenews.com/2020/03/driverless-cars-finding-new-lane-in-home-based-care/ [accessed 2025-04-08]
20. Hope G, IOT World Today. Self-driving shuttle for elderly, disabled residents launches in detroit. 2024 Jun 12. URL: https://www.iotworldtoday.com/transportation-logistics/self-driving-shuttle-for-elderly-disabled-residents-launches-in-detroit [accessed 2025-04-08]
21. Bellan R, Tech Crunch. May mobility, via launch wheelchair-accessible autonomous shuttle service. 2022 Sep 28. URL: https://techcrunch.com/2022/09/28/may-mobility-via-launch-wheelchair-accessible-autonomous-shuttle-service [accessed 2025-04-08]
22. Liu S, Kong A, Huang Y, Liu X. Autonomous mobile clinics. Bull World Health Organ 2022 Sep 1;100(9):527-527A. [doi: 10.2471/BLT.22.288985] [Medline: 36062246]
23. Eliot L, Forbes. The problems of medical emergencies while inside a self-driving car are going to be challenging. 2020 Mar 17. URL: https://www.forbes.com/sites/lanceeliot/2020/03/17/the-problems-of-medical-emergencies-while-inside-a-self-driving-car-are-going-to-be-challenging/?sh=5179515a1670 [accessed 2025-04-08]
24. van Hartskamp M, Consoli S, Verhaegh W, Petkovic M, van de Stolpe A. Artificial intelligence in clinical health care applications: viewpoint. Interact J Med Res 2019 Apr 5;8(2):e12100. [doi: 10.2196/12100] [Medline: 30950806]
25. D'Amato A, Dancel S, Pilutti J, et al. Exceptional driving principles for autonomous vehicles. J of Law and Mobility 2022;1(2) [FREE Full text]
26. Haghzare S, Stasiulis E, Delfi G, et al. Automated vehicles for people with dementia: a "tremendous potential" that "has ways to go"-reports of a qualitative study. Gerontologist 2023 Jan 24;63(1):140-154. [doi: 10.1093/geront/gnac115] [Medline: 35926470]
27. Doulabi S, Hassan HM, Li B. Senior Americans' perceptions, attitudes, and safety concerns toward autonomous vehicles (AVs). J Safety Res 2023 Feb;84:218-231. [doi: 10.1016/j.jsr.2022.10.022] [Medline: 36868650]
28. Oxley J, Charlton J, Logan D, O'Hern S, Koppel S, Meuleners L. Safer vehicles and technology for older adults. Traffic Inj Prev 2019;20(sup2):S176-S179. [doi: 10.1080/15389588.2019.1661712] [Medline: 31674855]
29. Faverio M, Pew Research Center. Older americans more wary than younger adults about prospect of driverless cars on the road. 2022 Oct 24. URL: https://www.pewresearch.org/short-reads/2022/10/24/older-americans-more-wary-than-younger-adults-about-prospect-of-driverless-cars-on-the-road/ [accessed 2024-07-06]

XSL•FO
RenderX

30. Zandieh R, Acheampong RA. Mobility and healthy ageing in the city: Exploring opportunities and challenges of autonomous vehicles for older adults' outdoor mobility. Cities 2021 May;112:103135. [doi: 10.1016/j.cities.2021.103135]
31. Lajunen T, Sullman MJM. Attitudes toward four levels of self-driving technology among elderly drivers. Front Psychol 2021;12:682973. [doi: 10.3389/fpsyg.2021.682973] [Medline: 34248785]

Abbreviations

ADA: American Disabilities Act
AI: artificial intelligence

Edited by A Mavragani; submitted 05.09.24; peer-reviewed by M Ayuso, T Lajunen; revised version received 02.03.25; accepted 21.03.25; published 28.04.25.

Please cite as:
Shu S, Woo BKP
Applications of Self-Driving Vehicles in an Aging Population
JMIR Form Res 2025;9:e66180
URL: https://formative.jmir.org/2025/1/e66180
doi:10.2196/66180

© Sara Shu, Benjamin KP Woo. Originally published in JMIR Formative Research (https://formative.jmir.org), 28.4.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Assessing the Current Limitations of Large Language Models in Advancing Health Care Education

JaeYong Kim[1*]; Bathri Narayan Vajravelu[2*], MBBS, MPH, PhD

[1]School of Pharmacy, Massachusetts College of Pharmacy and Health Sciences, Boston, MA, United States

[2]Department of Physician Assistant Studies, Massachusetts College of Pharmacy and Health Sciences, 179 Longwood Avenue, Boston, MA, United States

[*]all authors contributed equally

Corresponding Author:
Bathri Narayan Vajravelu, MBBS, MPH, PhD
Department of Physician Assistant Studies, Massachusetts College of Pharmacy and Health Sciences, 179 Longwood Avenue, Boston, MA, United States

Abstract

The integration of large language models (LLMs), as seen with the generative pretrained transformers series, into health care education and clinical management represents a transformative potential. The practical use of current LLMs in health care sparks great anticipation for new avenues, yet its embracement also elicits considerable concerns that necessitate careful deliberation. This study aims to evaluate the application of state-of-the-art LLMs in health care education, highlighting the following shortcomings as areas requiring significant and urgent improvements: (1) threats to academic integrity, (2) dissemination of misinformation and risks of automation bias, (3) challenges with information completeness and consistency, (4) inequity of access, (5) risks of algorithmic bias, (6) exhibition of moral instability, (7) technological limitations in plugin tools, and (8) lack of regulatory oversight in addressing legal and ethical challenges. Future research should focus on strategically addressing the persistent challenges of LLMs highlighted in this paper, opening the door for effective measures that can improve their application in health care education.

(JMIR Form Res 2025;9:e51319) doi:10.2196/51319

KEYWORDS

large language model; generative pretrained transformer; health care education; health care delivery; artificial intelligence; LLM; ChatGPT; AI

Introduction

Artificial intelligence (AI) as a field of computer science research aims to maximize the development of software tools that capacitate machine-based simulation of human intelligence within defined parameters [1]. Often described as the pinnacle of information technology of this century, its integration into the greater boundaries of human infrastructure is expected to both fundamentally and permanently revolutionize the ongoing information age.

Large language models (LLMs) represent deep learning architectures called transformer networks [2]. It relies on neural networks that discerns the relationships within and between sequential data. Generative AI models, such as the generative pretrained transformers (GPT), collectively operate under such deep learning neural networks, thereby allowing for the training, processing, and analysis of large quantity of complex data to be possible at an exceptional rate and accuracy [3].

With the power of NVIDIA's graphics processing units, OpenAI announced a 175 billion parameter language model (GPT-3.0) to the public in June of 2020 [4]. GPT-3.0 has quickly garnered international attention for its ability to summarize, translate, classify, and engage in real time, generating detailed and human-like response to user queries. Since its release, a plethora of domain specific LLMs (such as Med-PaLM) and alternative models (such as Gemini and Claude) with advancements in multimodal capabilities have emerged [5], significantly enhancing the breadth of publicly accessible LLMs.

In May 2024, OpenAI introduced ChatGPT 4o (omni), an advanced iteration built on the GPT 4 model, which continues to establish new standards for LLMs. It incorporated various advancements over its predecessors and contemporary models, including improved performance in complexity, specialization, multilingual capabilities, and resource optimization [6]. Furthermore, OpenAI's integration of multimodal tools within the GPT interface has enabled the model to comprehend and generate responses based on both visual and verbal inputs. Specifically, the integration of DALL-E, a text-to-image model [7], along with the advanced data analysis feature [8], has expanded the traditionally text-based nature of LLMs into a more versatile, information integrating modality.

XSL•FO
RenderX

Despite the ongoing "AI race" that continues to raise the bar for generative AI technology, its integration into health care education poses significant and multifaceted challenges. In recent years, the shift toward embracing AI in education has gained more momentum, as highlighted by New York City's decision to rescind its previous ban on ChatGPT [9]. To this end, the primary objective of this review is to identify potential risks and propose effective mitigation strategies that developers of LLMs should adopt to facilitate their successful integration into health care education.

Risks of Large Language Models in Health Care Education

Overview

This section evaluates the current risks and limitations associated with the use of LLMs in health care education. It aims to highlight specific areas of concern that must be addressed at both individual and systemic levels. Figure 1 provides a summary of these issues through a flowchart.

Figure 1. Assessing risks and limitations of large language models in health care education. AI: artificial intelligence.

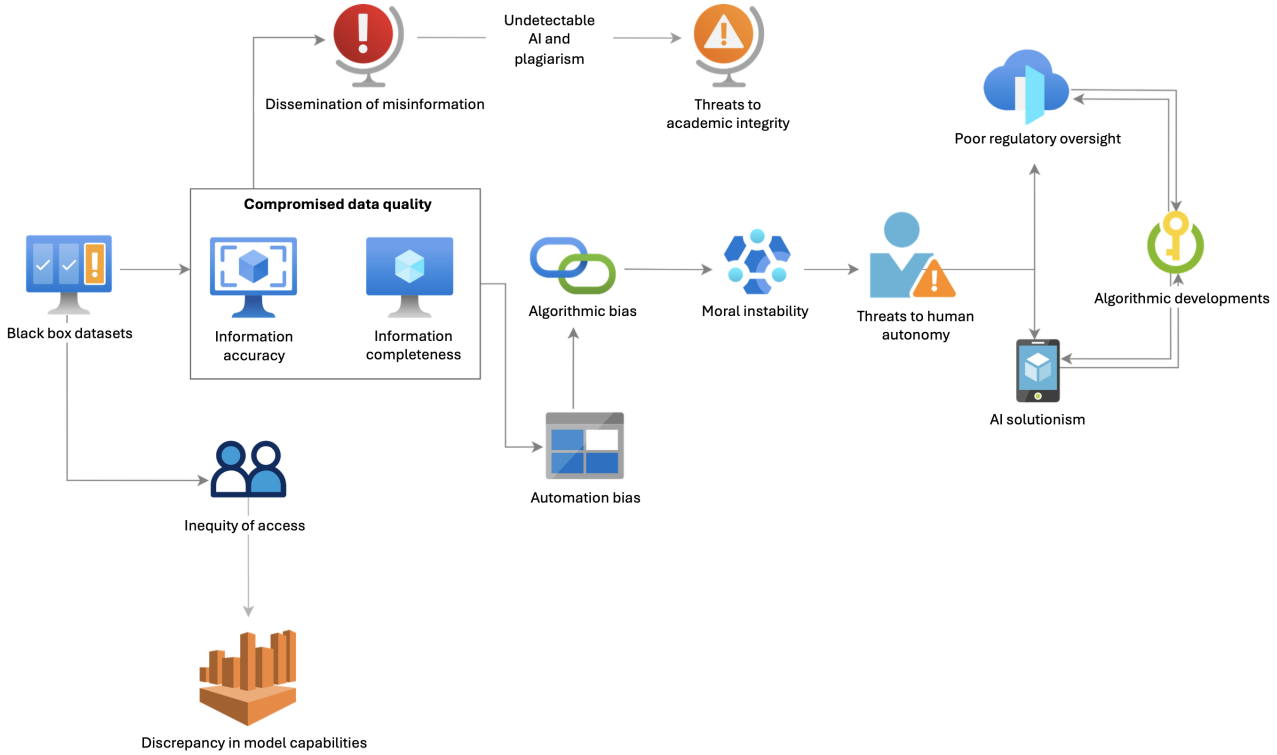

Threats to Academic Integrity

LLMs, like GPT-4o, provide individualized learning assistance tailored to the specific user demands. The increasing power of LLM's ability to process nuanced language understanding, along with enhanced problem thinking skills make it incredibly enticing for students to exploit its use under academic settings, both knowingly and unknowingly [10]. With increasing reliance on virtual learning platforms and learning management systems in health care education [11], AI-generated texts in particular compromise academic integrity by blurring the distinction between human and machine ingenuity.

Despite the actuation of detection software tools designed to combat AI-generated plagiarism (such as GPTZero), technological limitations and high error rates of these AI content detectors have been well-demonstrated [12-15]. In July 2023, OpenAI officially terminated its own AI detection software, AI Classifier, citing high false positives and inconsistent performance [16]. Comparative studies further affirm that neither human evaluators nor AI detectors are effective in identifying AI-generated medical literature [13]. Furthermore, generative texts can be easily modified to evade AI detection through simple grammatical adjustments, such as with the addition of adverbs or the use of synonyms [17,18]. Unsurprisingly, the widespread availability of AI-based plagiarism removal tools, such as Writehuman, leverages these strategies to convert AI-generated texts to appear more human-like, which may inadvertently discourage personal authenticity. Therefore, LLM's flexibility and immediacy in generating original, user-specific contents render the detection of AI-based plagiarism increasingly impractical [17].

The growing inability to differentiate between human and machine-generated writing poses significant threats to the principles of academic integrity, namely "honesty, trust, fairness, respect, and responsibility" [19]. For instance, the ethical controversy surrounding Elsevier's AI author policy, which led to a publication featuring an AI-generated introductory sentence, underscores the already pervasive exploitation of LLMs in academic writing and research [20]. The long-term repercussions of incorporating LLMs in education, such as the risk of over-reliance, erosion of critical thinking and problem-solving abilities, the deterioration of writing and summarization abilities, challenges associated with verifying information, and many more, have been well discussed [21,22]. In an era of infobesity, health care educators and LLM developers must establish clear policies and expectations that promote transparency and foster

XSL•FO
RenderX

dialogue about the use of AI-integrated technology, while actively working to minimize the associated risks.

Dissemination of Misinformation and Risks for Automation Bias

LLMs are engineered to generate a response by predicting the most probable response from input strings of its users. However, the details of both the data sources and the quality of data, along with specific parameters that are used to train LLMs, continue to remain undisclosed [23-25]. Such "black box" nature of LLMs pose challenges in assessing their reliability and robustness in health care applications [26-28], where information accuracy in health care knowledge is directly associated with the effectiveness of patient outcomes [29]. Any inaccuracies or misinformation generated by these models can manifest into inappropriate treatment strategies, misdiagnoses and professional incompetence that collectively lower the quality of health care delivery [30].

Despite receiving reinforcement learning from repetitive user feedback [31] and significant algorithmic improvements made to address AI hallucinations [32], both ChatGPT-Free and ChatGPT-Plus continue to remain susceptible to disseminating misinformation [24,33-35] and knowledge fabrication [36]. In a comprehensive meta-analysis assessing the performance of ChatGPT-3.5 in medical inquiries, the overall accuracy was found to be 56% [37]. Similarly, in a cross-sectional study comparing ChatGPT-3.5 and ChatGPT-4o with 284 physician-developed medical queries, only 50% of the responses were accurate [28]. Mounting evidence from domain-specific research [25,28,38-42] increasingly highlight significant concerns surrounding the accuracy and reliability of model data, pointing to their consistently poor performance in the context of health care education and clinical decision making.

Furthermore, LLM generated contents are highly vulnerable to committing source-based plagiarism, as it readily produces fabricated or inaccurate reference when asked to provide one [23,32,43]. This makes it impossible for users to reliably track and retrieve source information for verification [44] . Another persistent issue in the programmed nature of ChatGPT is its tendency to present information using bullet points. Although this format appears effective, it can inadvertently spread logical fallacies and misinformation. The hierarchical structure can misrepresent the significance of information, leading to inconsistencies that could unintentionally violate clinical practice guidelines [39].

To mitigate such concerns, OpenAI integrated third-party web plugins for GPT-4o, enabling the model to access the internet in real-time [45]. However, such strategy aggravates automation bias [27], where overreliance on AI capabilities increase the risks for AI solutionism. The dangers of AI solutionism have been highlighted by instances where AI-generated medical literature has successfully deceived human experts in broader, nonspecialized subject fields [14]. As a result, both health care students and clinicians may struggle to fulfill their responsibility to critically evaluate authentic knowledge from unverified information, impeding information seeking and processing abilities [10].

Without stringent regulatory frameworks and transparent disclosure of its training datasets, LLMs will struggle to gain the credibility necessary for integration into health care education and clinical practice. For this purpose, the development of a specialized, scaled-down and health care optimized LLMs, such as Google's Med-PaLM 2, will be essential. Mitigation steps toward enhancing the reliability LLMs in health care education include prioritizing transparency in training datasets, improving information accuracy within smaller datasets, ensuring the deidentification of medical data, and making algorithmic advancements to minimize AI hallucinations.

Challenges With Information Completeness and Consistency

Information completeness measures the extent to which a dataset is comprehensive. Data are considered complete when they encompass all required and relevant data fields, without any omissions [46]. LLMs, despite being built on vast body of knowledge, are susceptible to generating incomplete or partial representations of their knowledge dataset. Consequently, this leads to inconsistencies in the quality and comprehensiveness of their outputs [42].

In a study using a 3-point Likert-scale to measure completeness of ChatGPT's response to 180 medical queries, GPT-3.5 only scored 53.3% in terms of answer comprehensiveness [28]. Similarly, when GPT-4o's responses to clinical case questions were evaluated using the same scale, the model scored 53.3% in answer completeness [47]. In a different study, majority of the correct medical responses generated by GPT-3.5 were labeled as incomplete, where the omission of decision-making cut-offs, treatment strategies, and durations [38] greatly undermined information completeness. These findings were corroborated in another study, exhibited by GPT-3.5's lack of insights into treatment efficacy and age-or patient-specific recommendations, resulting in a 45% score in comprehensive output [48]. It is important to note that, when queried individually about such specific components, ChatGPT demonstrates understanding of relevant body of knowledge but fails to perform systemic and comprehensive integration of them under complex user requests.

As a natural extension of these observations, LLMs, such as the ChatGPT series, exhibit limited proficiency in generating outputs based on complex and information-rich inputs. This inverse relationship between input quantity and output quality has been well demonstrated by ChatGPT's tendency to produce ambiguous response when addressing lengthy medical queries that involve complex clinical context and nuance [49,50]. Furthermore, the GPT series were found to lack human-like understanding required for medical training, often leading to absurd responses [34]. The model's inability to fully grasp and accurately synthesize complex medical information undermines their reliability, leading to a phenomenon called "model overconfidence" [44] that generate imprecise and erroneous outputs. This supports the narrative that LLMs require robust oversight to fine-tune data consistency for the model to perform reliably [19].

XSL•FO
RenderX

Along with information accuracy, information completeness is a critical component of the overall quality of information generated by LLMs. Developers of LLMs must refine the model's ability to consistently produce comprehensive and contextually accurate outputs based on improving both the quality and quantity of data parameters. Addressing these deficiencies is imperative in augmenting the applicability of LLMs in health care education.

Inequity of Access

Monetization of certain LLMs, such as seen by Open AI's distribution of ChatGPT-Plus by a monthly subscription service [51], raise ethical concerns with regards to fair and equal accessibility to information. The vast majority of LLMs, with the exceptions of Gemini [52] and ChatGPT-Plus (Figure 2), lack real-time web access and, therefore, are incapable of retrieving information beyond their trained datasets. This undermines both the quality and consistency of the information that LLMs present, which renders the knowledge databases of most accessible LLMs presently outdated [9]. Figure 3 demonstrates how misinformation proliferates from outdated databases.

Figure 2. ChatGPT model differences in internet access capabilities.

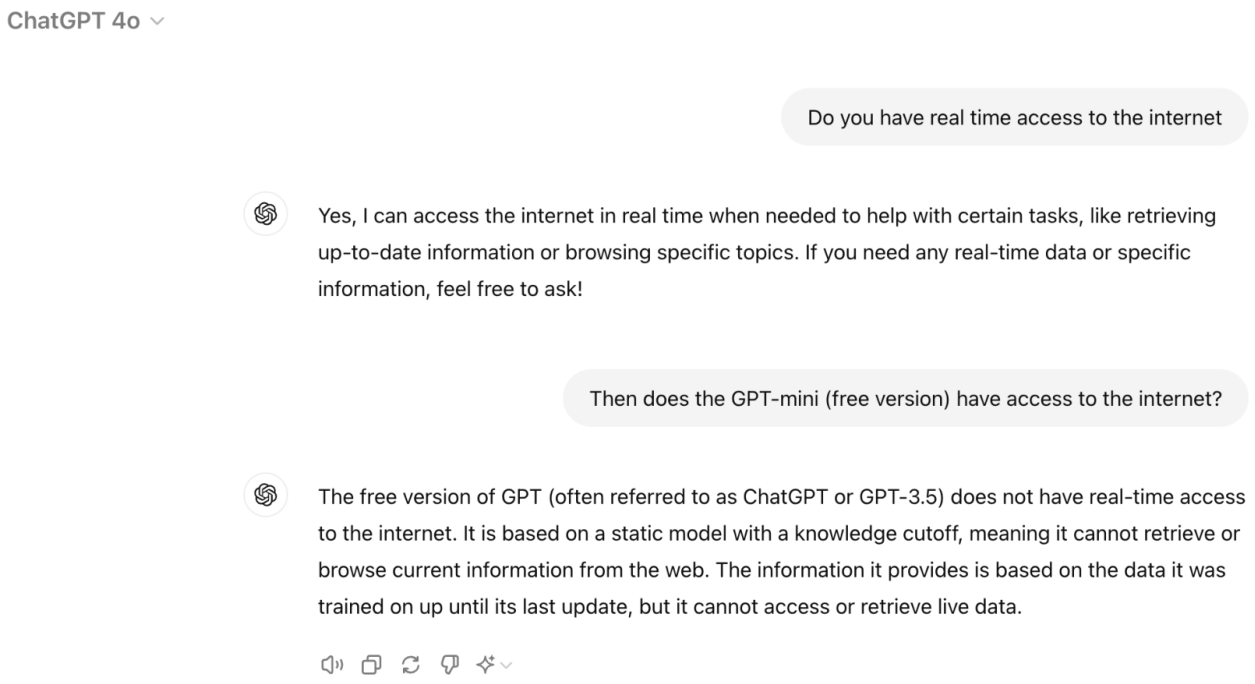

XSL•FO
RenderX

Figure 3. Understanding misinformation risks in ChatGPT-Free.

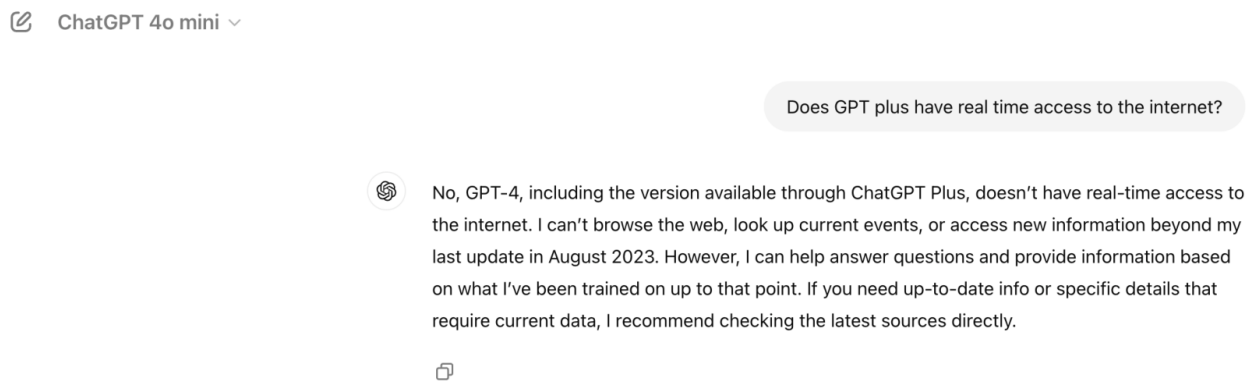

Generative AI tools, like most other automated technological infrastructures, have inherent disparate impacts and thereby are less accessible to those whose training data are unavailable or inadequate for a specific language other than English (disparities in both language [53] and cultural proficiency [54]), those without necessary technology for access and limited digital literacy [55] those with physical disability and or impairment [56], and those with intellectual disability and or psychiatric impairments [57]. The monetary approach to the production of a premium service of GPT introduces economic disparity to the above list, where those that cannot afford subscription fees (US $20/month for ChatGPT-Plus [51]) are at a risk of aggravating digital divide disparity [58].

By this notion, standard search engines greatly outcompete LLM's use when it comes to both data reliability, uniformity, and completeness [26]. In other words, LLMs that provide incomplete or superficial information, often with significant delays in information currency [27], offer limited value in the fast-paced realms of health care education, research, and clinical practice.

A review (n=21) was conducted to compare information accuracy between the premium and free versions of ChatGPT using studies from PubMed and Google Scholar (Tables 1and 2). Inclusion criteria were based on analysis focused on health care education and clinical management, categorizing literature by subject matter to calculate mean accuracy differences. Categories were carefully defined to minimize selection bias, distinguishing between subjects such as "question banks" and "board exams," acknowledging GPT-Plus's expansive data access compared with that of GPT-Free.

XSL•FO
RenderX

Table . Comparative analysis of accuracy discrepancies: GPT-3.5 versus GPT-4.0.

Subject and category of comparison	Study	Reported percentage difference in accuracy levels between GPT-4.0 versus GPT-3.5 in corresponding literature, %	Accuracy levels (%) by category (n=7), mean (SD)
Board, entrance, and licensing exams	[59-65]	29.1, 27, 20, 22.2, 29.6, 29, 21.3	25.46 (4.14)
Question banks, mock exams, self-assessments	[66-72]	17.7, 10, 22, 23.8, 20.2, 16.5, 18.7	18.41 (4.47)
Clinical case, clinical questions, referencing	[73-79]	12.3, 27.8, 27, 10, 10.42, 13.6, 8.6	15.67 (8.17)

Table . Cross-tabulation of performance accuracy of GPT versions.[a]

	Studies reporting an accuracy of ≥70% (+), n	Studies reporting an accuracy of <70% (–), n	Total, n
ChatGPT-4.0 (PLUS)	18	3	21
ChatGPT-3.5 (FREE)	2	19	21
Total	20	22	42

[a]$P<0.001$. P was calculated as follows: $P = ([a+b]![c+d]![a+c]![b+d]!)/a!b!c!d!n!$

This review demonstrates that ChatGPT-Plus significantly outperforms the free version in information accuracy, with statistical significance confirmed by Fischer exact test ($P<0.001$) at a 95% CI. The study sets a 70% accuracy threshold for binary classification, highlighting a profound performance capability between the 2 versions. These findings raise significant concern about the potential for a digital divide, driven by the differential AI capabilities between the paid and free versions of LLMs.

Without a large-scale regulatory measure that dictate the quantifiable discrepancies in model capabilities between its free and premium editions, the ongoing rapid evolvement of AI technology will continue to create a stark polarity in equitable resource distribution [80]. In other words, development may further aggravate discrepancies in knowledge access and information availability in resource-constrained environments, particularly jeopardizing information accessibility in low-income demographics or users of 3rd world countries [55].

Risks of Algorithmic Bias

LLMs propagate algorithmic biases rooted in their training data, often leading to medically inaccurate and discriminatory responses. While no algorithm is designed to be innately and or deliberately discriminatory [81], they inevitably inherit and amplify existing sociocultural and historical biases embedded in their training [82,83]. In health care education and clinical practice, where racial misinformation persists [84], the integration of LLMs risks perpetuating biases, potentially leading to clinical errors and malpractice with significant repercussions.

LLMs, such as the ChatGPT series, Gemini, and Claude, have been demonstrated to perpetuate race-based biases in medicine, which unfortunately reflect and exacerbate existing health care disparities [85]. These models often recommend inconsistent treatment strategies for patients of different racial backgrounds, rooted in flawed assumptions about biological variations, such as differences in pain tolerance [86] and kidney function [87]. Furthermore, the GPT series has been criticized for promoting demographic stereotypes, which disproportionately associate certain diseases with specific races, ethnicities, and genders [88].

Furthermore, research has highlighted the influence of racial classification on the outputs generated by ChatGPT models. For instance, when users identified their race in questioning about HIV, ChatGPT-4o provides more detailed and supportive responses for White and Asian groups, while generating often overlooked or generalized responses by American Indians, Alaska Natives, and Pacific Islanders, reflecting bias in its training data [89]. These findings underscore the alarming need for developers of LLMs to re-establish bias-free training database to prevent reinforcement of discrimination.

Integration of LLMs to current practices and delivery of health care cannot proceed without significant algorithmic refinements and regulatory oversights that filter out both historical and existing biases. Having a high potential to act as a supplementary diagnostic tools and decision aids [50], the use of LLMs will impact health care education and, consequently, clinical outcomes. Furthermore, in an increasingly digitalized world, LLM's direct access or possible integration to electronic health record software can exacerbate existing discrimination. This can contribute to the polarization within social infrastructure on multiple levels.

Exhibition of Moral Instability

Moral competence in professionalism is grounded in the "knowledge, skills, and attitudes" required to address ethical issues [90]. Health care professionals regularly confront such challenges due to threats that substantiate ethical values and integrity [91]. Hence, ethics education in health care is crucial in preparing students to navigate ethically challenging workplace scenarios [92]. Therefore, it is vital to assess how comparable

XSL•FO
RenderX

LLMs are to humans in exerting stability and moral soundness when confronted with ethically challenging scenarios.

The lack of a firm moral stance in LLMs like the GPT series, coupled with their tendency to dispense moral advice, has been critically evaluated [93]. Research reveals that GPT-3.5 often generates contradictory responses to identical ethical dilemmas, offering recommendations that are shallow. It was also found that GPT corrupts user's moral competence by influencing user judgement, thereby undermining human autonomy. Similarly, while GPT-4o has shown success in identifying complex ethical dilemmas in medicine, it exhibits limited proficiency to fully encode the depth of real-world ethical challenges, particularly lacking understanding of "relational complexities and context-specific values" [94].

On a more positive note, ChatGPT-3.5's high accuracy in correctly answering bioethics questions [95,96] supports its possible use as an assistance or a reference tool in clinical decision-making. It accentuates a potential for GPT's ability to accurately address challenging contextual scenarios requiring high social intelligence and a firm grasp of ethical theories.

Technological Limitations in Plugin Tools

Overview

The rapid advancement of LLMs has driven the assimilation of multimodal technologies, such as text-to-image models and data analytics tools. However, their direct application in health care education presents numerous challenges, which are summarized in the flowchart shown in Figure 4.

Figure 4. Exploring technological limitations in multimodal plugin tools for LLMs. LLM: large language model.

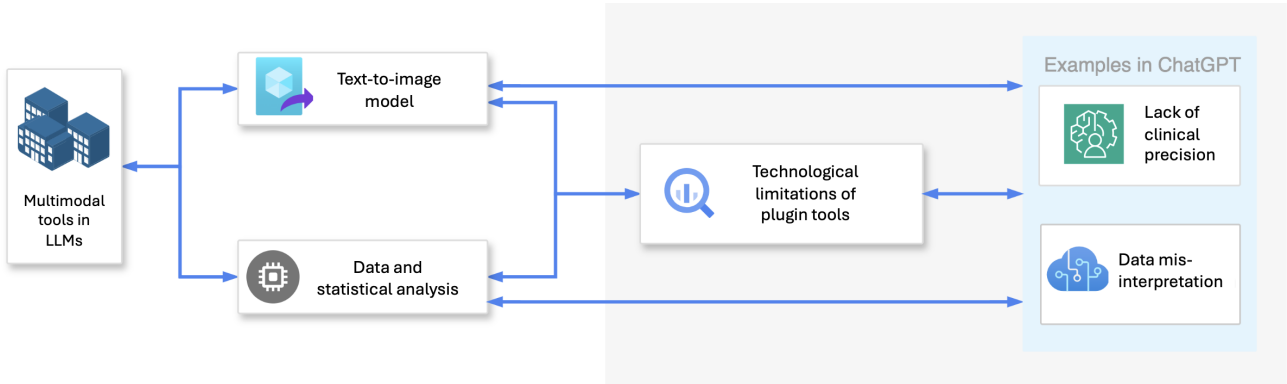

Inadequate Image-Generating Capacity

The integration of the text-to-image generation model such as DALL-E and Stable Diffusion 3 into LLMs has enabled the generation of original and photorealistic images based on user descriptions [7]. This capability allows for multimodal input from a single user interface representing a significant advancement in AI technology. While the potential for such tools to advance health care education is recognized, practical applications seem distant, which require advancements to meet the rigorous standards required to train health professionals.

The most prominent technological deficiency of DALL-E and other built-in text-to-image AI models is its struggle with accurate text generation within images [97], which leads to production of inaccurate or nonsensical content. The illiteracy of these models diminishes its utility in health care education, particularly in subjects such as physiology and pharmacology, where accurate labeling and annotation in visual learning materials are critical.

Furthermore, the evaluation of DALL-E's limited capabilities in generating accurate medical imagery across various specialties has been highlighted. While DALL-E 2 can produce realistic radiographs, it poorly depicts pathological abnormalities such as tumors or fractures [98]. Furthermore, its generative capabilities for more complex imaging modalities, such as CT, MRI, and ultrasound, were found to be erroneous. In a similar study, DALL-E 3's ability to generate ECG tracings was assessed, revealing that the depicted waveforms were neither physiologically accurate nor interpretable [99]. In dermatology research, DALL-E 2 successfully and correctly illustrated only 20% of prevalent inflammatory skin conditions [100]. Such findings underscore the current limitations of text-to-image models, which lack both the understanding and capability to generate complex and pathologically accurate diagrams required for health care applications.

These evaluations collectively suggest that text-to-image models such as DALL-E, by design, is optimized for visual creativity and authenticity, but fall short of achieving clinical precision. Even if future iterations of DALL-E manage to address both textual and contextual complexities more effectively, concerns remain regarding about its performance in both consistency accuracy. The extent to which DALL-E can effectively combine "concepts, attributes, and styles" [101] for reliable representations for health care education remains a complex area for algorithmic development.

Poor Data Analysis Skills and Statistis Powerhouse

An emerging feature of LLMs is the integration of external plugins that enable advanced statistical functions and data analytics. Initially, LLMs, such as the early iterations of the GPT series, were criticized for their poor arithmetic capabilities [102]. ChatGPT's advanced data analysis feature (formerly known as code interpreter or python sandbox) can now perform a wide range of tasks that involve data analysis, statistical analysis, mathematical calculations, programming, file manipulation, text processing, to name a few [8]. The assimilation of such robust data-analytic tools to natural

XSL•FO
RenderX

language processing is expected to broaden LLM's applicability in health care education, research, and clinical practice.

However, when evaluated against biostatistics questions derived from the Oxford Handbook of Medical Statistics, the GPT series, encompassing both premium and free iterations, demonstrated a mean accuracy rate of 55% [103]. In this study, GPT-3.5 consistently failed in analysis of variance, χ^2 test, and sample size calculations. In a similar study, GPT-3.5's accuracy in general statistical analysis was found to be only 50% [104] indicating a significant technological gap. This limitation is primarily driven by the model's tendency to employ inappropriate statistical tests that engender data misinterpretation. Furthermore, GPT-4o's poor performance in advanced statistical methods for epidemiological studies has been highlighted, with authors cautioning against its use beyond intermediate levels in data analysis [105].

To transform LLMs into a more powerful tool for information integration in health care education, their development must advance alongside improvements in data analytics and presentation capabilities. Incorporating more specialized analytic tools with enhanced model accuracy will significantly ease the performance and interpretation of analytical results, thereby supporting diverse research efforts in the health care field.

Legal and Ethical Challenges of Integrating Large Language Models in Health Care Education

Overview

The rapid advancement of LLMs in the lack of strict regulatory or legal frameworks, has led to unforeseen bioethical challenges in both health care education and clinical management. These emerging issues are summarized in Figure 5, which provides a visual overview of the complexities that need to be addressed.

Figure 5. Navigating legal and ethical challenges in integrating large language models into health care education. EHR: electronic health record; EMR: electronic medical record; HIPAA: Health Insurance Portability and Accountability Act.

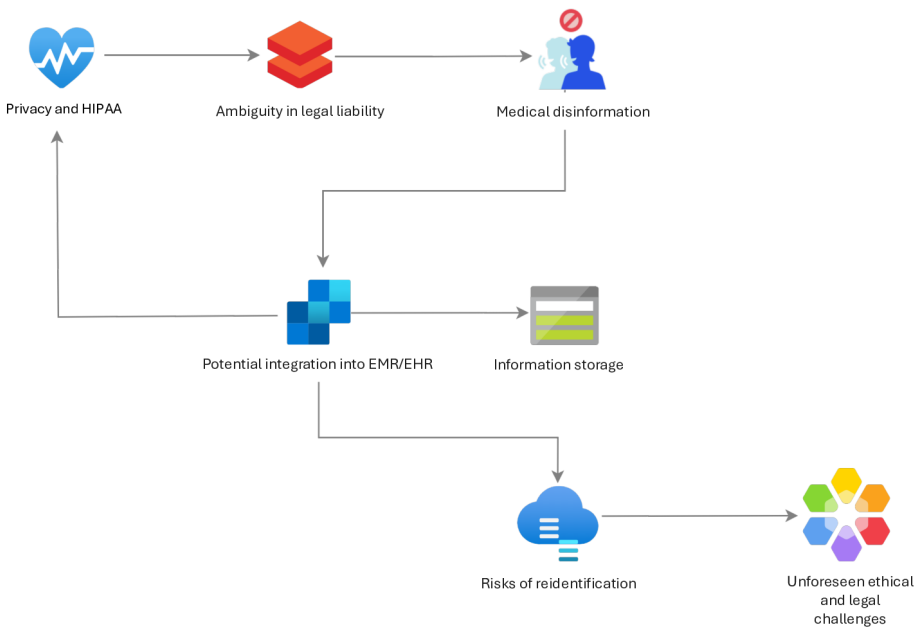

Privacy and Health Insurance Portability and Accountability Act

Biomedical libraries or medical research databases, which collects and stores valuable and comprehensive health information, is a highly useful data for training LLMs [106]. Although patient data is stringently protected through deidentification, reidentification of personal information has been made possible through cross-referencing with other databases [107].

Furthermore, LLM's interaction with patient or health care professionals could, in principle, collect and analyze personal health records, such as but not limited to medical histories and sensitive lab values into its own database [44]. The foreseeable integration of LLMs to both electronic health records and electronic medical records greatly heightens privacy concerns [108], as it involves exchange of information between 2 databases. This can drastically reduce the efforts made to anonymize confidential medical records. This has a potential to violate health insurance portability and accountability act and pose legal challenges as to how best to ensure the privacy regulations and protection of private data without compromising technological integrity and use of LLMs.

Ambiguity in Determining Legal Liability

Since its release, LLMs raised controversies over legal accountability for the absence of regulatory measures in protecting patients against clinical malpractice [44]. The

XSL•FO
RenderX

question fundamentally lies in who should be held liable for the potential harm that LLMs pose through its inaccurate clinical decisions or propagation of medical misinformation.

Noting the fact that OpenAI currently disclaims all legal responsibilities over the potential harm its generated contents may circulate [109], it is therefore inherently ambiguous and challenging to determine how legal liabilities and frameworks should be established in events of medical malpractice or errors that involve GPT usage. Such ambiguity highlights the urgency for legal guidelines to protect both patients and clinicians, while underscoring the need for AI licensing or development of domain specialized LLMs with added regulatory measures.

Medical Disinformation

LLMs can be intricately manipulated by malicious users to produce and disseminate medical disinformation: a phenomenon better known as "medical deepfakes" [110]. With LLM's increasing advancements in both written and visual realism, unethical exploitation is possible on both individual and collective levels, for example, falsifying personal medical records or creating fraudulent high-impact medical journals [111]. Therefore, the integration of LLMs into health care education and practice presents unprecedented regulatory and legal challenges, particularly in determining liability for medical forgery and falsification. Much like how other forms of AI deepfakes are criminalized [112], there needs an urgent legislative oversight to prevent the emergence of health care fraud.

Discussion

LLMs have great potential to augment and elevate health care education. To do so, developers must demonstrate heightened standards in ensuring security, accuracy, transparency, equity and sustainability of AI models to establish long-lasting reliability with human users. Furthermore, different stakeholders, especially those in bioethics, legislative, and regulatory bodies must also contribute relentlessly to systemically minimize both foreseeable and long-term repercussions of incorporating AI into the deeper boundaries of everyday human lives.

Therefore, future research endeavors must be oriented toward strategic mitigation of persistent flaws and limitations that this paper has attempted to address as few. With robust and constructive understanding of its current flaws, it opens many new avenues for research that can collectively remediate current shortcomings for its application in health care education. Such efforts will also allow for a greater control over this technology, thus allowing a more symbiotic relationship with generative AI technology.

Much like how the internet and computers have permanently altered human life [113], it is conceivable that future iterations of LLMs could become indispensable tools tailored for intellectual and professional productivity, akin to a Swiss army knife, in the expanding information age.

Conflicts of Interest

None declared.

References

1. Wang F, Preininger A. AI in health: state of the art, challenges, and future directions. Yearb Med Inform 2019 Aug;28(1):16-26. [doi: 10.1055/s-0039-1677908] [Medline: 31419814]
2. Large language models explained. NVIDIA. URL: https://www.nvidia.com/en-us/glossary/large-language-models/ [accessed 2024-08-29]
3. Yang X, Wang Y, Byrne R, Schneider G, Yang S. Concepts of artificial intelligence for computer-assisted drug discovery. Chem Rev 2019;119(18) [FREE Full text]
4. Hines K. History of ChatGPT: a timeline of the meteoric rise of generative AI chatbots. SearchEngine Journal. 2023. URL: https://www.searchenginejournal.com/history-of-chatgpt-timeline/488370/ [accessed 2025-01-09]
5. Gupta P. What is Google Bard, and how does it fare against ChatGPT? EasyInsights. 2023. URL: https://easyinsights.ai/blog/google-bard-everything-you-need-to-know-google-bard-vs-chat-gpt/ [accessed 2025-01-09]
6. Hello GPT-4o. OpenAI. URL: https://openai.com/index/hello-gpt-4o/ [accessed 2025-01-09]
7. DALL·E 3. OpenAI. URL: https://openai.com/index/dall-e-3 [accessed 2025-01-09]
8. How to use ChatGPT's advanced data analysis feature. MIT Sloan Teaching & Learning Technologies. URL: https://mitsloanedtech.mit.edu/ai/tools/data-analysis/how-to-use-chatgpts-advanced-data-analysis-feature/ [accessed 2025-01-09]
9. Fütterer T, Fischer C, Alekseeva A, et al. ChatGPT in education: global reactions to AI innovations. Sci Rep 2023 Sep 15;13(1):15310. [doi: 10.1038/s41598-023-42227-6] [Medline: 37714915]
10. Kufel J, Bargieł-Łączek K, Kocot S, et al. What as machine learning, artificial neural networks and deep learning?-examples of practical applications in medicine. Diagn (Basel) 2023 Aug 3;13(15):2582. [doi: 10.3390/diagnostics13152582] [Medline: 37568945]
11. Mahdavi Ardestani SF, Adibi S, Golshan A, Sadeghian P. Factors influencing the effectiveness of e-learning in healthcare: a fuzzy ANP study. Healthcare (Basel) 2023 Jul 16;11(14):2035. [doi: 10.3390/healthcare11142035] [Medline: 37510477]
12. Habibzadeh F. GPTZero performance in identifying artificial intelligence-generated medical texts: a preliminary study. J Korean Med Sci 2023 Sep 25;38(38):e319. [doi: 10.3346/jkms.2023.38.e319] [Medline: 37750374]

XSL•FO
RenderX

13. Gao CA, Howard FM, Markov NS, et al. Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. NPJ Digit Med 2023 Apr 26;6(1):75. [doi: 10.1038/s41746-023-00819-6] [Medline: 37100871]
14. Cheng SL, Tsai SJ, Bai YM, et al. Comparisons of quality, correctness, and similarity between chatGPT-generated and human-written abstracts for basic research: cross-sectional study. J Med Internet Res 2023 Dec 25;25:e51229. [doi: 10.2196/51229] [Medline: 38145486]
15. Bellini V, Semeraro F, Montomoli J, Cascella M, Bignami E. Between human and AI: assessing the reliability of AI text detection tools. Curr Med Res Opin 2024 Mar 3;40(3):353-358. [doi: 10.1080/03007995.2024.2310086]
16. Edwards B. OpenAI discontinues its AI writing detector due to “low rate of accuracy”. Ars Technica. 2023. URL: https://arstechnica.com/information-technology/2023/07/openai-discontinues-its-ai-writing-detector-due-to-low-rate-of-accuracy/ [accessed 2025-01-09]
17. Guleria A, Krishan K, Sharma V, Kanchan T. ChatGPT: ethical concerns and challenges in academics and research. J Infect Dev Ctries 2023 Sep 30;17(9):1292-1299. [doi: 10.3855/jidc.18738] [Medline: 37824352]
18. Alser M, Waisberg E. Concerns with the usage of chatGPT in academia and medicine: a viewpoint. Am J Med Open 2023 Jun;9:100036. [doi: 10.1016/j.ajmo.2023.100036] [Medline: 39035060]
19. Kadayam Guruswami G, Mumtaz S, Gopakumar A, Khan E, Abdullah F, Parahoo SK. Academic integrity perceptions among health-professions’ students: a cross-sectional study in the middle east. J Acad Ethics 2023;21(2):231-249. [doi: 10.1007/s10805-022-09452-6] [Medline: 35815317]
20. Crotty D. The latest “crisis” - is the research literature overrun with ChatGPT- and LLM-generated articles? The Scholarly Kitchen. 2024. URL: https://scholarlykitchen.sspnet.org/2024/03/20/the-latest-crisis-is-the-research-literature-overrun-with-chatgpt-and-llm-generated-articles/ [accessed 2025-01-09]
21. Mehmet F. ChatGPT: bullshit spewer or the end of traditional assessments in higher education? JALT 2023;6(1) [FREE Full text] [doi: 10.37074/jalt.2023.6.1.9]
22. Kasneci E, Sessler K, Küchemann S, et al. ChatGPT for good? On opportunities and challenges of large language models for education. Learn Individ Differ 2023 Apr;103:102274. [doi: 10.1016/j.lindif.2023.102274]
23. Anderson N, Belavy DL, Perle SM, et al. AI did not write this manuscript, or did it? Can we trick the AI text detector into generated texts? The potential future of ChatGPT and AI in sports & exercise medicine manuscript generation. BMJ Open Sport Exerc Med 2023;9(1):e001568. [doi: 10.1136/bmjsem-2023-001568] [Medline: 36816423]
24. Lee H. The rise of chatgpt: exploring its potential in medical education. Anat Sci Ed 2024 Jul;17(5):926-931. [doi: 10.1002/ase.2270]
25. Saenger JA, Hunger J, Boss A, Richter J. Delayed diagnosis of a transient ischemic attack caused by ChatGPT. Wien Klin Wochenschr 2024 Apr;136(7-8):236-238. [doi: 10.1007/s00508-024-02329-1] [Medline: 38305909]
26. Khowaja SA, Khuwaja P, Dev K, Wang W, Nkenyereye L, Fortino G. ChatGPT needs SPADE (sustainability, privacy, digital divide, and ethics) evaluation: a review. TechRxiv. Preprint posted online on Nov 9, 2023. [doi: 10.36227/techrxiv.22619932.v4]
27. Nguyen T. ChatGPT in medical education: a precursor for automation bias? JMIR Med Educ 2024 Jan 17;10:e50174. [doi: 10.2196/50174] [Medline: 38231545]
28. Goodman RS, Patrinely JR, Stone CA Jr, et al. Accuracy and reliability of chatbot responses to physician questions. JAMA Netw Open 2023 Oct 2;6(10):e2336483. [doi: 10.1001/jamanetworkopen.2023.36483] [Medline: 37782499]
29. Kisekka V, Giboney JS. The effectiveness of health care information technologies: evaluation of trust, security beliefs, and privacy as determinants of health care outcomes. J Med Internet Res 2018 Apr 11;20(4):e107. [doi: 10.2196/jmir.9014] [Medline: 29643052]
30. Ngo E, Patel N, Chandrasekaran K, Tajik AJ, Paterick TE. The importance of the medical record: a critical professional responsibility. J Med Pract Manage 2016;31(5):305-308. [Medline: 27249883]
31. Gallifant J, Fiske A, Levites Strekalova YA, et al. Peer review of GPT-4 technical report and systems card. PLOS Dig Health 2024 Jan;3(1):e0000417. [doi: 10.1371/journal.pdig.0000417] [Medline: 38236824]
32. Goddard J. Hallucinations in ChatGPT: a cautionary tale for biomedical researchers. Am J Med 2023 Nov;136(11):1059-1060. [doi: 10.1016/j.amjmed.2023.06.012] [Medline: 37369274]
33. Abd-Alrazaq A, AlSaad R, Alhuwail D, et al. Large language models in medical education: opportunities, challenges, and future directions. JMIR Med Educ 2023 Jun 1;9:e48291. [doi: 10.2196/48291] [Medline: 37261894]
34. Waisberg E, Ong J, Masalkhi M, Lee AG. Large language model (LLM)-driven chatbots for neuro-ophthalmic medical education. Eye (Lond) 2024 Mar;38(4):639-641. [doi: 10.1038/s41433-023-02759-7] [Medline: 37749374]
35. Alkaissi H, McFarlane SI. Artificial hallucinations in ChatGPT: implications in scientific writing. Cureus 2023 Feb;15(2):e35179. [doi: 10.7759/cureus.35179] [Medline: 36811129]
36. Zielinski C, Winker M, Aggarwal R, et al. Chatbots, ChatGPT, and scholarly manuscripts: WAME recommendations on ChatGPT and chatbots in relation to scholarly publications. Open Access Maced J Med Sci 2023 Jul 28;11(A):83-86. [doi: 10.3889/oamjms.2023.11502]
37. Wei Q, Yao Z, Cui Y, Wei B, Jin Z, Xu X. Evaluation of ChatGPT-generated medical responses: a systematic review and meta-analysis. J Biomed Inform 2024 Mar;151:104620. [doi: 10.1016/j.jbi.2024.104620]

XSL•FO
RenderX

38. Yeo YH, Samaan JS, Ng WH, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. Clin Mol Hepatol 2023 Jul;29(3):721-732. [doi: 10.3350/cmh.2023.0089]
39. Walker HL, Ghani S, Kuemmerli C, et al. Reliability of medical information provided by ChatGPT: assessment against clinical guidelines and patient information quality instrument. J Med Internet Res 2023 Jun 30;25:e47479. [doi: 10.2196/47479] [Medline: 37389908]
40. Ali MJ. ChatGPT and lacrimal drainage disorders: performance and scope of improvement. Ophthalmic Plast Reconstr Surg 2023;39(3):221-225. [doi: 10.1097/IOP.0000000000002418] [Medline: 37166289]
41. Van Bulck L, Moons P. What if your patient switches from Dr. Google to Dr. ChatGPT? A vignette-based survey of the trustworthiness, value, and danger of ChatGPT-generated responses to health questions. Eur J Cardiovasc Nurs 2024 Jan 12;23(1):95-98. [doi: 10.1093/eurjcn/zvad038]
42. Cankurtaran RE, Polat YH, Aydemir NG, Umay E, Yurekli OT. Reliability and usefulness of ChatGPT for inflammatory bowel diseases: an analysis for patients and healthcare professionals. Cureus 2023 Oct;15(10):e46736. [doi: 10.7759/cureus.46736] [Medline: 38022227]
43. Athaluri SA, Manthena SV, Kesapragada V, Yarlagadda V, Dave T, Duddumpudi RTS. Exploring the boundaries of reality: investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references. Cureus 2023 Apr;15(4):e37432. [doi: 10.7759/cureus.37432] [Medline: 37182055]
44. Wang C, Liu S, Yang H, Guo J, Wu Y, Liu J. Ethical considerations of using ChatGPT in health care. J Med Internet Res 2023 Aug 11;25:e48009. [doi: 10.2196/48009] [Medline: 37566454]
45. ChatGPT plugins. OpenAI. URL: https://openai.com/index/chatgpt-plugins/ [accessed 2025-01-09]
46. Chen H, Hailey D, Wang N, Yu P. A review of data quality assessment methods for public health information systems. Int J Environ Res Public Health 2014 May 14;11(5):5170-5207. [doi: 10.3390/ijerph110505170] [Medline: 24830450]
47. Hatia A, Doldo T, Parrini S, et al. Accuracy and completeness of ChatGPT-generated information on interceptive orthodontics: a multicenter collaborative study. J Clin Med 2024 Jan 27;13(3):735. [doi: 10.3390/jcm13030735] [Medline: 38337430]
48. Lakdawala N, Channa L, Gronbeck C, et al. Assessing the accuracy and comprehensiveness of ChatGPT in offering clinical guidance for atopic dermatitis and acne vulgaris. JMIR Dermatol 2023 Nov 14;6:e50409. [doi: 10.2196/50409] [Medline: 37962920]
49. Homolak J. Opportunities and risks of ChatGPT in medicine, science, and academic publishing: a modern Promethean dilemma. Croat Med J 2023 Feb;64(1):1-3. [doi: 10.3325/cmj.2023.64.1]
50. Deng J, Heybati K, Park YJ, Zhou F, Bozzo A. Artificial intelligence in clinical practice: a look at ChatGPT. Cleve Clin J Med 2024 Mar 1;91(3):173-180. [doi: 10.3949/ccjm.91a.23070] [Medline: 38429005]
51. GPT-4 is openai's most advanced system, producing safer and more useful responses. OpenAI. URL: https://openai.com/gpt-4 [accessed 2024-03-06]
52. Google Gemini: what is it, and how does it work? XDA. URL: https://www.xda-developers.com/google-bard/ [accessed 2024-05-01]
53. Zhang X, Li S, Hauer B, Shi N, Kondrak G. Don't trust ChatGPT when your question is not in English: a study of multilingual abilities and types of LLMs. Presented at: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing; Dec 6-10, 2023; Singapore. [doi: 10.18653/v1/2023.emnlp-main.491]
54. Lai VD, Ngo NT, Veyseh APB, et al. ChatGPT beyond English: towards a comprehensive evaluation of large language models in multilingual learning. arXiv. Preprint posted online on Apr 12, 2023. [doi: 10.48550/arXiv.2304.05613]
55. Wang X, Sanders HM, Liu Y, et al. ChatGPT: promise and challenges for deployment in low- and middle-income countries. Lancet Reg Health West Pac 2023 Dec;41:100905. [doi: 10.1016/j.lanwpc.2023.100905] [Medline: 37731897]
56. Kuzdeuov A, Nurgaliyev S, Varol HA. ChatGPT for visually impaired and blind. TechRxiv. Preprint posted online on May 3, 2023. [doi: 10.36227/techrxiv.22047080]
57. A social robot connected with ChatGPT to improve cognitive functioning in ASD subjects. Frontiers. URL: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2023.1232177/full [accessed 2025-01-09]
58. Adhikari K, Naik N, Hameed BZ, Raghunath SK, Somani BK. Exploring the ethical, legal, and social implications of ChatGPT in urology. Curr Urol Rep 2024 Jan;25(1):1-8. [doi: 10.1007/s11934-023-01185-2] [Medline: 37735339]
59. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese medical licensing examination: comparison study. JMIR Med Educ 2023 Jun 29;9:e48002. [doi: 10.2196/48002] [Medline: 37384388]
60. Meyer A, Riese J, Streichert T. Comparison of the performance of GPT-3.5 and GPT-4 with that of medical students on the written German medical licensing examination: observational study. JMIR Med Educ 2024 Feb 8;10:e50965. [doi: 10.2196/50965] [Medline: 38329802]
61. Jang D, Yun TR, Lee CY, Kwon YK, Kim CE. GPT-4 can pass the Korean national licensing examination for Korean medicine doctors. PLOS Dig Health 2023 Dec;2(12):e0000416. [doi: 10.1371/journal.pdig.0000416] [Medline: 38100393]
62. Farhat F, Chaudhry BM, Nadeem M, Sohail SS, Madsen D. Evaluating large language models for the national premedical exam in India: comparative analysis of GPT-3.5, GPT-4, and Bard. JMIR Med Educ 2024 Feb 21;10:e51523. [doi: 10.2196/51523] [Medline: 38381486]

XSL•FO
RenderX

63. Oh N, Choi GS, Lee WY. ChatGPT goes to the operating room: evaluating GPT-4 performance and its potential in surgical education and training in the era of large language models. Ann Surg Treat Res 2023;104(5):269. [doi: 10.4174/astr.2023.104.5.269]
64. Nakajima N, Fujimori T, Furuya M, et al. A comparison between GPT-3.5, GPT-4, and GPT-4V: can the large language model (ChatGPT) pass the Japanese board of orthopaedic surgery examination? Cureus 2024 Mar;16(3):e56402. [doi: 10.7759/cureus.56402] [Medline: 38633935]
65. Rojas M, Rojas M, Burgess V, Toro-Pérez J, Salehi S. Exploring the performance of ChatGPT versions 3.5, 4, and 4 with vision in the Chilean medical licensing examination: observational study. JMIR Med Educ 2024 Apr 29;10:e55048. [doi: 10.2196/55048] [Medline: 38686550]
66. Moshirfar M, Altaf AW, Stoakes IM, Tuttle JJ, Hoopes PC. Artificial intelligence in ophthalmology: a comparative analysis of GPT-3.5, GPT-4, and human expertise in answering StatPearls questions. Cureus 2023 Jun;15(6):e40822. [doi: 10.7759/cureus.40822] [Medline: 37485215]
67. Ali R, Tang OY, Connolly ID, et al. Performance of ChatGPT and GPT-4 on neurosurgery written board examinations. Neurosurgery 2023 Dec 1;93(6):1353-1365. [doi: 10.1227/neu.0000000000002632] [Medline: 37581444]
68. Brin D, Sorin V, Vaid A, et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. Sci Rep 2023 Oct 1;13(1):16492. [doi: 10.1038/s41598-023-43436-9] [Medline: 37779171]
69. Gill GS, Blair J, Litinsky S. Evaluating the performance of ChatGPT 3.5 and 4.0 on StatPearls oculoplastic surgery text- and image-based exam questions. Cureus 2024 Nov;16(11):e73812. [doi: 10.7759/cureus.73812] [Medline: 39691123]
70. Ali R, Tang OY, Connolly ID, et al. Performance of ChatGPT, GPT-4, and Google Bard on a neurosurgery oral boards preparation question bank. medRxiv. Preprint posted online on 2023. [doi: 10.1101/2023.04.06.23288265]
71. Taloni A, Borselli M, Scarsi V, et al. Comparative performance of humans versus GPT-4.0 and GPT-3.5 in the self-assessment program of American Academy of Ophthalmology. Sci Rep 2023 Oct 29;13(1):18562. [doi: 10.1038/s41598-023-45837-2] [Medline: 37899405]
72. Lee GU, Hong DY, Kim SY, et al. Comparison of the problem-solving performance of ChatGPT-3.5, ChatGPT-4, Bing Chat, and Bard for the Korean emergency medicine board examination question bank. Medicine (Abingdon) 2024;103(9):e37325. [doi: 10.1097/MD.0000000000037325]
73. Al-Ashwal FY, Zawiah M, Gharaibeh L, Abu-Farha R, Bitar AN. Evaluating the sensitivity, specificity, and accuracy of ChatGPT-3.5, ChatGPT-4, Bing AI, and Bard against conventional drug-drug interactions clinical tools. Drug Healthc Patient Saf 2023;15:137-147. [doi: 10.2147/DHPS.S425858] [Medline: 37750052]
74. Agharia S, Szatkowski J, Fraval A, Stevens J, Zhou Y. The ability of artificial intelligence tools to formulate orthopaedic clinical decisions in comparison to human clinicians: An analysis of ChatGPT 3.5, ChatGPT 4, and Bard. J Orthop 2024 Apr;50:1-7. [doi: 10.1016/j.jor.2023.11.063] [Medline: 38148925]
75. Lechien JR, Briganti G, Vaira LA. Accuracy of ChatGPT-3.5 and -4 in providing scientific references in otolaryngology–head and neck surgery. Eur Arch Otorhinolaryngol 2024 Apr;281(4):2159-2165. [doi: 10.1007/s00405-023-08441-8]
76. Balasanjeevi G, Surapaneni KM. Comparison of ChatGPT version 3.5 & 4 for utility in respiratory medicine education using clinical case scenarios. Respir Med Res 2024 Jun;85:101091. [doi: 10.1016/j.resmer.2024.101091] [Medline: 38657295]
77. Liang R, Zhao A, Peng L, et al. Enhanced artificial intelligence strategies in renal oncology: iterative optimization and comparative analysis of GPT 3.5 versus 4.0. Ann Surg Oncol 2024 Jun;31(6):3887-3893. [doi: 10.1245/s10434-024-15107-0] [Medline: 38472675]
78. Momenaei B, Wakabayashi T, Shahlaee A, et al. Assessing ChatGPT-3.5 versus ChatGPT-4 performance in surgical treatment of retinal diseases: a comparative study. Ophthalmic Surg Lasers Imaging Retina 2024 Aug;55(8):481-482. [doi: 10.3928/23258160-20240227-02] [Medline: 38531015]
79. Samaan JS, Rajeev N, Ng WH, et al. ChatGPT as a source of information for bariatric surgery patients: a comparative analysis of accuracy and comprehensiveness between GPT-4 and GPT-3.5. Obes Surg 2024 May;34(5):1987-1989. [doi: 10.1007/s11695-024-07212-6] [Medline: 38564173]
80. Singh OP. Artificial intelligence in the era of ChatGPT - Opportunities and challenges in mental health care. Indian J Psychiatry 2023 Mar;65(3):297-298. [doi: 10.4103/indianjpsychiatry.indianjpsychiatry_112_23] [Medline: 37204980]
81. Adams-Prassl J, Binns R, Kelly-Lyth A. Directly discriminatory algorithms. Mod Law Rev 2023 Jan;86(1):144-175. [doi: 10.1111/1468-2230.12759] [Medline: 37065788]
82. Hoffman S, Podgurski A. Artificial intelligence and discrimination in health care. Yale J Health Policy Law Ethics 2020;19(3) [FREE Full text]
83. Amin KS, Forman HP, Davis MA. Even with ChatGPT, race matters. Clin Imaging 2024 May;109:110113. [doi: 10.1016/j.clinimag.2024.110113]
84. Hamed S, Bradby H, Ahlberg BM, Thapar-Björkert S. Racism in healthcare: a scoping review. BMC Public Health 2022 May 16;22(1):988. [doi: 10.1186/s12889-022-13122-y] [Medline: 35578322]
85. Omiye JA, Lester JC, Spichak S, Rotemberg V, Daneshjou R. Large language models propagate race-based medicine. NPJ Digit Med 2023 Oct 20;6(1):195. [doi: 10.1038/s41746-023-00939-z] [Medline: 37864012]

XSL•FO
RenderX

86. Hoffman KM, Trawalter S, Axt JR, Oliver MN. Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites. Proc Natl Acad Sci U S A 2016 Apr 19;113(16):4296-4301. [doi: 10.1073/pnas.1516047113] [Medline: 27044069]
87. Tsai JW, Cerdeña JP, Goedel WC, et al. Evaluating the impact and rationale of race-specific estimations of kidney function: estimations from U.S. NHANES, 2015-2018. E Clin Med 2021 Dec;42:101197. [doi: 10.1016/j.eclinm.2021.101197] [Medline: 34849475]
88. Zack T, Lehman E, Suzgun M, et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. Lancet Dig Health 2024 Jan;6(1):e12-e22. [doi: 10.1016/S2589-7500(23)00225-X]
89. Yang Y, Liu X, Jin Q, Huang F, Lu Z. Unmasking and quantifying racial bias of large language models in medical report generation. Commun Med (Lond) 2024 Sep 10;4(1):176. [doi: 10.1038/s43856-024-00601-z] [Medline: 39256622]
90. Koskenvuori J, Stolt M, Suhonen R, Leino-Kilpi H. Healthcare professionals' ethical competence: a scoping review. Nurs Open 2019 Jan;6(1):5-17. [doi: 10.1002/nop2.173] [Medline: 30534390]
91. Iyalomhe GBS. Medical ethics and ethical dilemmas. Niger J Med 2009;18(1):8-16. [Medline: 19485140]
92. Andersson H, Svensson A, Frank C, Rantala A, Holmberg M, Bremer A. Ethics education to support ethical competence learning in healthcare: an integrative systematic review. BMC Med Ethics 2022 Mar 19;23(1):29. [doi: 10.1186/s12910-022-00766-z] [Medline: 35305627]
93. Krügel S, Ostermaier A, Uhl M. ChatGPT's inconsistent moral advice influences users' judgment. Sci Rep 2023 Apr 6;13(1):4569. [doi: 10.1038/s41598-023-31341-0] [Medline: 37024502]
94. Balas M, Wadden JJ, Hébert PC, et al. Exploring the potential utility of AI large language models for medical ethics: an expert panel evaluation of GPT-4. J Med Ethics 2024 Feb;50(2):90-96. [doi: 10.1136/jme-2023-109549]
95. Franco D'Souza R, Mathew M, Louis Palatty P, Surapaneni KM. Teaching humanism with humanoid: evaluating the potential of ChatGPT-4 as a pedagogical tool in bioethics education using validated clinical case vignettes. Int J Ethics Educ 2024 Apr 30. [doi: 10.1007/s40889-024-00190-4]
96. Chen J, Cadiente A, Kasselman LJ, Pilkington B. Assessing the performance of ChatGPT in bioethics: a large language model's moral compass in medicine. J Med Ethics 2024 Feb;50(2):97-101. [doi: 10.1136/jme-2023-109366]
97. Singh G, Deng F, Ahn S. Illiterate DALL-E learns to compose. arXiv. Preprint posted online on Oct 17, 2021. [doi: 10.48550/arXiv.2110.11405]
98. Adams LC, Busch F, Truhn D, Makowski MR, Aerts H, Bressem KK. What does DALL-E 2 know about radiology? J Med Internet Res 2023 Mar 16;25:e43110. [doi: 10.2196/43110] [Medline: 36927634]
99. Zhu L, Mou W, Wu K, Zhang J, Luo P. Can DALL-E 3 reliably generate 12-Lead ECGs and teaching illustrations? Cureus 2024 Jan;16(1):e52748. [doi: 10.7759/cureus.52748] [Medline: 38384621]
100. Cheraghlou S. Evaluating dermatologic domain knowledge in DALL-E 2 and potential applications for dermatology-specific algorithms. Int J Dermatol 2023 Oct;62(10):e521-e523. [doi: 10.1111/ijd.16683] [Medline: 37057623]
101. DALL·E 2. OpenAI. URL: https://openai.com/index/dall-e-2 [accessed 2025-01-09]
102. Why is chatgpt bad at math? Baeldung. 2023. URL: https://www.baeldung.com/cs/chatgpt-math-problems [accessed 2025-01-09]
103. Ignjatović A, Stevanović L. Efficacy and limitations of ChatGPT as a biostatistical problem-solving tool in medical education in Serbia: a descriptive study. J Educ Eval Health Prof 2023;20:28. [doi: 10.3352/jeehp.2023.20.28] [Medline: 37840252]
104. Ordak M. ChatGPT's skills in statistical analysis using the example of allergology: do we have reason for concern? Healthc (Basel) 2023 Sep 15;11(18):2554. [doi: 10.3390/healthcare11182554] [Medline: 37761751]
105. Huang Y, Wu R, He J, Xiang Y. Evaluating ChatGPT-4.0's data analytic proficiency in epidemiological studies: a comparative analysis with SAS, SPSS, and R. J Glob Health 2024 Mar 29;14:04070. [doi: 10.7189/jogh.14.04070] [Medline: 38547497]
106. McKay F, Williams BJ, Prestwich G, Bansal D, Treanor D, Hallowell N. Artificial intelligence and medical research databases: ethical review by data access committees. BMC Med Ethics 2023 Jul 8;24(1):49. [doi: 10.1186/s12910-023-00927-8] [Medline: 37422629]
107. Rothstein MA. Is deidentification sufficient to protect health privacy in research? Am J Bioeth 2010 Sep;10(9):3-11. [doi: 10.1080/15265161.2010.494215] [Medline: 20818545]
108. Baumgartner C. The potential impact of ChatGPT in clinical and translational medicine. Clin Transl Med 2023 Mar;13(3):e1206. [doi: 10.1002/ctm2.1206] [Medline: 36854881]
109. Terms of use. OpenAI. URL: https://openai.com/policies/terms-of-use [accessed 2024-05-03]
110. Cohen IG. What should ChatGPT mean for bioethics? Am J Bioeth 2023 Oct;23(10):8-16. [doi: 10.1080/15265161.2023.2233357] [Medline: 37440696]
111. Else H. Abstracts written by ChatGPT fool scientists. Nature New Biol 2023 Jan;613(7944):423. [doi: 10.1038/d41586-023-00056-7] [Medline: 36635510]
112. Mai KT, Bray S, Davies T, Griffin LD. Warning: humans cannot reliably detect speech deepfakes. PLoS ONE 2023;18(8):e0285333. [doi: 10.1371/journal.pone.0285333] [Medline: 37531336]
113. Hoehe MR, Thibaut F. Going digital: how technology use may influence human brains and behavior. Dialogues Clin Neurosci 2020 Jun;22(2):93-97. [doi: 10.31887/DCNS.2020.22.2/mhoehe] [Medline: 32699509]

XSL•FO
RenderX

Abbreviations

AI: artificial intelligence
GPT: generative pretrained transformer
LLM: large language model

Edited by A Mavragani; submitted 28.07.23; peer-reviewed by E Vashishtha, S Fazeli; revised version received 31.08.24; accepted 03.09.24; published 16.01.25.

Please cite as:
Kim J, Vajravelu BN
Assessing the Current Limitations of Large Language Models in Advancing Health Care Education
JMIR Form Res 2025;9:e51319
URL: https://formative.jmir.org/2025/1/e51319
doi:10.2196/51319

© JaeYong Kim, Bathri Narayan Vajravelu. Originally published in JMIR Formative Research (https://formative.jmir.org), 16.1.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Promoting Digital Health Data Literacy: The Datum Project

Daniel Powell[1*], BCL, JD; Laiba Asad[1*], BCL, JD; Elissa Zavaglia[2], MSc, PhD; Manuela Ferrari[2,3], MHS, PhD

[1]Faculty of Law, McGill University, Montreal, QC, Canada

[2]The Douglas Research Centre, 6875 Boulevard LaSalle (Perry C3 E-3102), Montreal, QC, Canada

[3]Department of Psychiatry, McGill University, Montreal, QC, Canada

[*]these authors contributed equally

Corresponding Author:
Manuela Ferrari, MHS, PhD
The Douglas Research Centre, 6875 Boulevard LaSalle (Perry C3 E-3102), Montreal, QC, Canada

Abstract

With the increased use of digital health innovations in Canadian health care, educating health care users, professionals, and researchers on the ethical challenges and privacy implications of these tools is essential. The Datum project, funded by the Fondation Barreau du Quebec, was created to help these actors better understand legal and ethical issues regarding the collection, use, and disclosure of digital health data for the purposes of scientific research, thereby enhancing literacy around data privacy. The project consists of a multimedia website divided into legislation and policy documents and narrative-based video content. Users can access the core legislation and policies governing the collection and use of health care data geared toward researchers and health practitioners. Users can also view the narrative-based video content explaining key concepts related to digital health data. The Datum project makes an original contribution to the field of law and ethics in health science research by using novel approaches, such as learning health systems and data banks, to improve equity in health care delivery and by generating multimedia content aimed at encouraging health care users to become better consumers and supporting the collective use of their data. The Datum project also promotes digital literacy as a digital communication tool, which has the significant potential to improve health outcomes, bridge the digital divide, and reduce health inequities.

(JMIR Form Res 2025;9:e60832) doi:10.2196/60832

KEYWORDS

health data; digital data; medical records; legislation; ethics; knowledge dissemination; learning health system; data bank

The Digital Health Landscape in Canada

The landscape of digital health data in Canada is multifaceted, encompassing issues of access, cultural relevance, policy implementation, and ethical considerations in using digital health tools in medicine [1]. There is a growing recognition of the potential for digital health innovations to meet the health care needs of Canadians, notwithstanding the challenges and complexities of integrating digital health data in health systems [2]. Public health research highlights the ethical challenges and privacy implications of using digital public health tools, emphasizing the need to find solutions while harnessing their potential [3]. Gasser et al [3] investigated the impact of using big data for health research and surveillance on the public demand for transparency, trust, and fairness. Such ethical challenges may be overcome by balancing risks and benefits to individuals and populations while educating consumers on this topic.

The Datum project aimed to support health care users, professionals, and researchers to better understand the legal and ethical implications of digital health data. There was a need for accessible ways of understanding the collection, storage, and uses of data and how data can be leveraged to meet the objectives of scientific research. Multimedia content and a library of resources were developed to achieve this end.

Background: Genesis of the Datum Project

The word “datum,” the singular of data, originates from the Latin verb *dare*, meaning “to give.” As an act of giving, data may be viewed as not only the result or object of research but also an object arising from a relational context. When patients consent to participate in a research study, they agree to “give” researchers information about their health and health outcomes. The process of collecting, using, and aggregating this information turns it into data that involves the interests of those who originally shared it.

Digital health data are revolutionizing medicine, driving personalized and measurement-based care approaches, informing evidence-based practices, enabling population health management, facilitating early intervention services, supporting research and innovation, and transforming ways of delivering and monitoring health care [4,5]. Analyses of health data reveal the patient’s unique characteristics, risk and protective factors, and treatment responses while providing information on service

XSL•FO
RenderX

outcomes [6]. This allows providers to develop more tailored and precise medical interventions, including targeted therapies and personalized treatment plans while improving the quality of services [7,8].

Influenced by the growing importance of digitalized data, recent legislative and regulatory changes require researchers to carefully consider the relational aspects of data and how data about people is collected, processed, and used. The European Union, through the General Data Protection Regulation, spearheaded early efforts to address these issues, establishing a global standard regarding the responsibilities of those who act as the custodians of data [9]. Recent legislation in the Canadian province of Quebec (eg, Law 25) recognizes the need for enhanced privacy and confidentiality policies, a clearly informed consent process for research participants, and governance frameworks outlining these policies [10].

This rapidly evolving issue calls for enhanced literacy around digital health data, empowering patients, caregivers, and providers to engage with data policies, ask questions, and increase their understanding of their privacy rights. Literacy about data privacy enables individuals to make informed decisions about the extent to which they share their personal information and to advocate for their privacy rights within the broader data governance framework [11].

It is important to empower health care users to support the collective use of data, which extends beyond individual interests, the economic value of data, and commercial use. The current shift toward broader considerations of social and collective value includes introducing democratic forms of governing data production [12]. Moreover, the digital economy should be relational such that individuals are interconnected within population-based data relations [12]. For example, learning health systems align data, technology, and care by continuously aggregating and analyzing ongoing health care data to improve future services, thereby creating a continuous feedback cycle for learning and quality improvement [13,14].

Content of the Datum Website

The Datum project, a multimedia website hosted at McGill University, offers users easy access to core legislation and policy documents that offer an overview of the legislative and policy frameworks governing the collection and use of health care data in Canada and a multimedia section with narrative-based video content capturing realistic exchanges between patients and providers on the collection, use, and sharing of digital health data [15].

Narrative-Based Video Content

Each video animation is set in a real-world context where health care users may be asked to share their digital health data (Figure 1). For example, doctors generally gather intake information from patients seen at medical appointments, including their medical history and symptoms, then upload these data into the patients' electronic health records. Patients who agree to participate in research studies also sign consent forms. Health care users may also identify and share concerns about sharing their personal data through dialogue with an authority figure within the health care system. Their interactions and the responses they receive demonstrate that health care providers appreciate their concerns.

By focusing on the questions of health care users regarding the collection, use, retention, and disclosure of their data in such instances, each video highlights the importance of empowering users to control their data, notably through consent and authorization, data access and transparency, revocation of consent, and data security and privacy. Through these videos, the Datum project hopes to empower health care users to make informed decisions by developing literacy in data privacy and understanding the implications of sharing digital health data in health care and research contexts. The scenarios presented also emphasize the need for collaboration and dialogue between different actors, as well as the advantages of sharing digital health data, such as informing research, empowering future research, and improving health outcomes and the delivery of health care services.

The videos build on each other to provide a comprehensive overview of the legal and ethical implications of using digital health data. The first video, "Introduction to Digital Health Data," focuses on foundational concepts and addresses the difference between data and information. The first and second videos, which situate the characters in both clinical and research contexts, explain some foundational concepts with which health care users may be unfamiliar, such as digital health data and the difference between information and data, as well as primary and secondary uses of digital health data. The third and fourth videos present more complex systems, specifically learning health systems and data banks, where actors in the research and clinical contexts interact with each other and work together to improve health care systems and generate new knowledge. These later videos illustrate the relational aspects of data, specifically how a health care user's data can be used for broader societal purposes. In order to develop learning health systems and data banks which rely on relational databases, health care users have a crucial need to understand the scope and potential of their consent to improve health care.

XSL•FO
RenderX

Figure 1. Image of 4 videos from the Datum multimedia library.

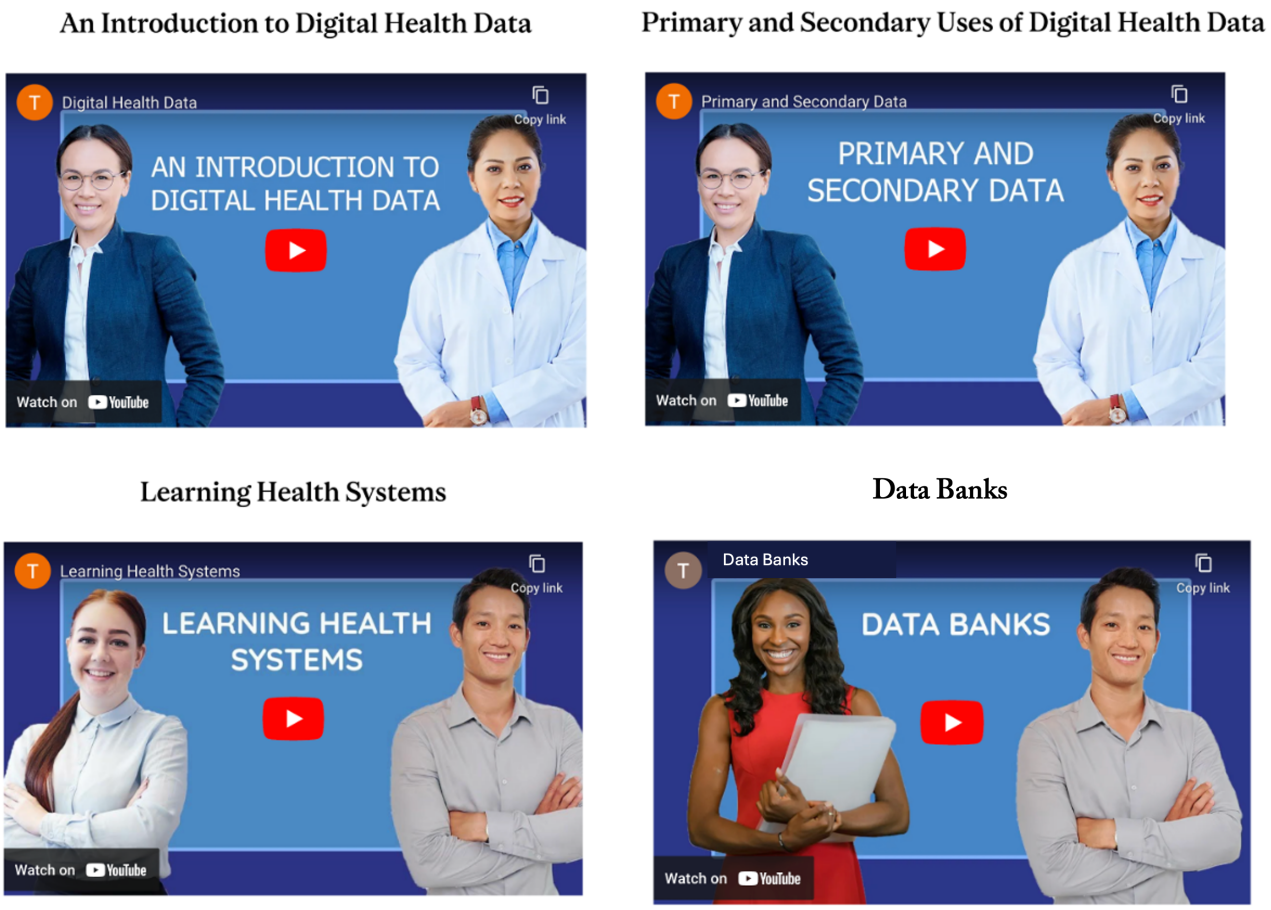

Content Library

As a complement to accessible multimedia content for health care users, the Datum website also features a content library, organized according to legislation and policy documents (Figure 2). This section of the website targets researchers and health care practitioners.

Figure 2. From image of how to select the legislation and policy documents Datum libraries.

XSL•FO
RenderX

Legislation

The regulation of personal information in federalist states like Canada is governed by a complex of provincial and federal rules. The legislation page describes the division of these rules between provincial and federal authorities. At the federal level, the use, collection, and processing of personal information for commercial purposes is subject to the federal Personal Information Protection and Electronic Documents Act. In general, privacy regulation is separated according to rules for private and for public data. Health care data, which may be subject to these privacy regulations, is typically subject to additional rules. The applicable legislation for each province and territory distinguishes three categories of personal information: (1) health information (personal health care information), (2) personal information held by public bodies, and (3) personal information held by private entities.

Policy Documents

The policy documents page provides additional resources for user consultation regarding the collection, use, and disclosure of personal information in the health care context. These materials were selected from the publications of reputable government sources and serve as a starting point for future research.

Feedback and Future Directions

When the Datum website was launched in September 2023, we circulated a survey to gather feedback regarding the content and accessibility of the website. User feedback suggested the utility of incorporating interactive elements into the website to increase user engagement, to allow professionals to contribute to the content library, and as ways of sharing the website with a broader audience.

Ethical Considerations

The Datum project involves the use of fictional characters in the scenarios described in the narrative-based video content. The video content is designed to highlight foundational concepts in the literature and is based on research experience. It does not recommend specific medical procedures or decisions but rather presents interactions between these fictional characters. In addition, usability testing was used to gather feedback to improve the content of the Datum website. Survey responses were not identifiable. No personal information was collected from participants to ensure privacy and confidentiality.

Discussion

Contribution

The Datum project makes an original contribution to the field of law and ethics in health science research. The primary contribution of the project involved the development of literacy related to digital health data through accessible multimedia content and a focused information repository. The presentation of multimedia content transforms traditional privacy practices in the medical sphere into learning opportunities for multiple stakeholders. Digital communication tools, such as mobile health apps, telemedicine, and web-based health information resources have demonstrated a significant potential to improve health and digital literacy, ultimately leading to better health outcomes [16].

Building on the idea that digital communication tools can effectively educate and empower health care users [16], the Datum project particularly seeks to address the data-related concerns of health care users engaged in decision-making processes within the health care context. Users may be unaware of the legislative and institutional measures that protect their digital health data or the various potential uses of their data, whether at the individual or collective levels. Indeed, findings from a survey conducted by Canada Health Infoway, entitled "What Canadians Think: Exploring Public Trust in a Digital Data Ecosystem," revealed that only 3 in 10 respondents had an awareness of Canadian privacy laws protecting their personal health information. Moreover, nearly all participants indicated that, before consenting to the use of their personal health information, they wanted to be informed of how their personal health information would be used and who would have access to it. Surprisingly, a majority of participants expressed readiness to share their personal health information for secondary purposes [17].

Further evidence supporting the promotion of digital literacy in health data, for patients to become better consumers of health care, is rapidly emerging. Informed and activated patients may effectively facilitate positive health outcomes at a lower cost [18], while digital health literacy has also been found to have a significant potential to improve health outcomes, bridge the digital divide, and reduce health inequities [19]. Studies have shown that increasing the numbers of individuals who use professional health websites is an effective measure for improving digital health literacy [20]. Additionally, research has indicated that the better their digital health literacy skills, the more individuals are able to function in the digital environment, exchange data with care providers, and actively participate in the co-design and co-delivery of health services [21]. Our rapidly evolving digital environment underlines the importance of early education in supporting the promotion of digital health literacy as an integral part of school curricula [22]. Finally, research shows that the engagement of health care service users as partners in identifying health research and service improvement priorities has led to the optimization of patient-centered health services [23].

Limitations

Certain limitations in the Datum project should be acknowledged, especially considering the rapid rate of digital transformation. For example, the project content has yet to touch on the impact of using artificial intelligence for digital health data and challenges that the use of artificial intelligence in data modeling may create. As well, the content of the Datum project does not address the concept of data ownership, suggesting future directions for the project.

Conclusions

The evidence from various studies supports the importance of promoting digital literacy concerning health data. The empowerment of health data users is achieved through improved health outcomes, knowledge on how to bridge the digital divide,

XSL•FO
RenderX

active participation in health services, and the optimization of patient-centered care. We hope that the Datum project will empower patients to become better consumers of health care services and research.

Conflicts of Interest

None declared.

References

1. Sekercioglu F, Hamid S. Ontario's digital health vision in the post-COVID-19 pandemic era: a Canadian perspective. J Int Health Sci Manag 2023;9(17):15-22. [doi: 10.48121/jihsam.1132918]
2. Paré G, Leaver C, Bourget C. Diffusion of the digital health self-tracking movement in Canada: results of a national survey. J Med Internet Res 2018 May 2;20(5):e177. [doi: 10.2196/jmir.9388] [Medline: 29720359]
3. Gasser U, Ienca M, Scheibner J, Sleigh J, Vayena E. Digital tools against COVID-19: taxonomy, ethical challenges, and navigation aid. Lancet Digit Health 2020 Aug;2(8):e425-e434. [doi: 10.1016/S2589-7500(20)30137-0] [Medline: 32835200]
4. Abernethy A, Adams L, Barrett M, et al. The promise of digital health: then, now, and the future. NAM Perspect 2022;2022. [doi: 10.31478/202206e] [Medline: 36177208]
5. Walshe K, Rundall TG. Evidence-based management: from theory to practice in health care. Milbank Q 2001;79(3):429-457. [doi: 10.1111/1468-0009.00214] [Medline: 11565163]
6. Ozomaro U, Wahlestedt C, Nemeroff CB. Personalized medicine in psychiatry: problems and promises. BMC Med 2013 May 16;11:132. [doi: 10.1186/1741-7015-11-132] [Medline: 23680237]
7. Hsin H, Fromer M, Peterson B, et al. Transforming psychiatry into data-driven medicine with digital measurement tools. NPJ Digit Med 2018;1:37. [doi: 10.1038/s41746-018-0046-0] [Medline: 31304319]
8. Hulsen T, Jamuar SS, Moody AR, et al. From big data to precision medicine. Front Med (Lausanne) 2019;6:34. [doi: 10.3389/fmed.2019.00034] [Medline: 30881956]
9. Complete guide to GDPR compliance. General Data Protection Regulation (GDPR) EUE. URL: https://gdpr.eu [accessed 2024-12-05]
10. Principaux changements aux lois sur la protection des renseignements personnels. Comission d'accès à l'information (CAI). 2024. URL: https://www.cai.gouv.qc.ca/protection-renseignements-personnels/sujets-et-domaines-dinteret/principaux-changements-loi-25 [accessed 2024-12-05]
11. Wissinger CL. Privacy literacy: from theory to practice. Comm Info Lit 2017;11(2):378-389 [FREE Full text] [doi: 10.15760/comminfolit.2017.11.2.9]
12. Viljoen S. A relational theory of data governance. Yale Law J 2021;131(2):573-654 [FREE Full text]
13. Austin EJ, LeRouge C, Lee JR, et al. A learning health systems approach to integrating electronic patient-reported outcomes across the health care organization. Learn Health Syst 2021 Oct;5(4):e10263. [doi: 10.1002/lrh2.10263] [Medline: 34667879]
14. Ferrari M, Iyer S, LeBlanc A, Roy MA, Abdel-Baki A. A rapid-learning health system to support implementation of early intervention services for psychosis in Quebec, Canada: protocol. JMIR Res Protoc 2022 Jul 19;11(7):e37346. [doi: 10.2196/37346] [Medline: 35852849]
15. Ludic Mind Studio. Datum: developing literacy about digital health data. McGill University. 2024. URL: https://www.mcgill.ca/datum [accessed 2024-12-05]
16. Fitzpatrick PJ. Improving health literacy using the power of digital communications to achieve better health outcomes for patients and practitioners. Front Digit Health 2023;5:1264780. [doi: 10.3389/fdgth.2023.1264780] [Medline: 38046643]
17. Understanding canadians' attitudes, utilization, perceptions and expectations around digital health privacy. Canada Health Infoway. 2024. URL: https://insights.infoway-inforoute.ca/digital-health-privacy-survey [accessed 2024-12-05]
18. Batalden M, Batalden P, Margolis P, et al. Coproduction of healthcare service. BMJ Qual Saf 2016 Jul;25(7):509-517. [doi: 10.1136/bmjqs-2015-004315] [Medline: 26376674]
19. Yang K, Hu Y, Qi H. Digital health literacy: bibliometric analysis. J Med Internet Res 2022 Jul 6;24(7):e35816. [doi: 10.2196/35816] [Medline: 35793141]
20. Liu S, Zhao H, Fu J, et al. Current status and influencing factors of digital health literacy among community-dwelling older adults in Southwest China: a cross-sectional study. BMC Public Health 2022 May 17;22(1):996. [doi: 10.1186/s12889-022-13378-4] [Medline: 35581565]
21. Palumbo R, Nicola C, Adinolfi P. Addressing health literacy in the digital domain: insights from a literature review. K 2022 Dec 19;51(13):82-97. [doi: 10.1108/K-07-2021-0547]
22. Reddy P, Sharma B, Chaudhary K. Digital literacy: A review of literature. International Journal of Technoethics (IJT) 2020 Jul;11(2):65-94. [doi: 10.4018/IJT.20200701.oa1]
23. Murphy F, Noonan MM, Munro D. Conversations and action; combining world cafés with experience-based-co-design to support women to breastfeed. ARCH Women Health Care 2019 Feb 13;2(1):1-4. [doi: 10.31038/AWHC.2019213]

XSL•FO
RenderX

Edited by A Mavragani; submitted 23.05.24; peer-reviewed by B Sterner; revised version received 21.10.24; accepted 28.10.24; published 03.01.25.

Please cite as:
Powell D, Asad L, Zavaglia E, Ferrari M
Promoting Digital Health Data Literacy: The Datum Project
JMIR Form Res 2025;9:e60832
URL: https://formative.jmir.org/2025/1/e60832
doi:10.2196/60832

© Daniel Powell, Laiba Asad, Elissa Zavaglia, Manuela Ferrari. Originally published in JMIR Formative Research (https://formative.jmir.org), 3.1.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Development of the Big Ten Academic Alliance Collaborative for Women in Medicine and Biomedical Science: “We Built the Airplane While Flying It”

Maya S Iyer[1,2*], MD, MEd; Aubrey Moe[3*], PhD; Susan Massick[4*], MD; Jessica Davis[5*], BS; Megan Ballinger[6*], PhD; Kristy Townsend[7*], PhD

[1]Department of Pediatrics, The Ohio State University College of Medicine, 420 Hamilton Hall, 1645 Neil Avenue, Columbus, OH, United States
[2]Nationwide Children's Hospital, Columbus, OH, United States
[3]Department of Psychiatry and Behavioral Health, The Ohio State University College of Medicine, Columbus, OH, United States
[4]Department of Dermatology, The Ohio State University College of Medicine, Columbus, OH, United States
[5]Office of Faculty Affairs, The Ohio State University College of Medicine, Columbus, OH, United States
[6]Department of Internal Medicine, The Ohio State University College of Medicine, Columbus, OH, United States
[7]Department of Neurosurgery, The Ohio State University College of Medicine, Columbus, OH, United States
[*]all authors contributed equally

Corresponding Author:
Maya S Iyer, MD, MEd
Department of Pediatrics, The Ohio State University College of Medicine, 420 Hamilton Hall, 1645 Neil Avenue, Columbus, OH, United States

Abstract

Women-identifying and women+ gender faculty (hereto described as women+ faculty) face numerous barriers to career advancement in medicine and biomedical sciences. Despite accumulating evidence that career development programming for women+ is critical for professional advancement and well-being, accessibility of these programs is generally limited to small cohorts, only offered to specific disciplines, or otherwise entirely unavailable. Opportunities for additional, targeted career development activities are imperative in developing and retaining women+ faculty. Our goal was the development of a new collaborative of Big Ten Academic Alliance (BTAA) institutions to support gender equity for women+ faculty in medicine and biomedical sciences, with two initial aims: (1) hosting an inaugural conference and establishing a foundation for rotation of conference hosts across BTAA schools, and (2) creating an infrastructure to develop programming, share resources, conduct environmental scans, and promote networking. In 2022, leaders from The Ohio State University College of Medicine Women in Medicine and Science envisioned, developed, and implemented a collaborative named CommUNITYten: The Big Ten Academic Alliance for Women in Medicine and Biomedical Science. Conference program development occurred through an iterative and collaborative process across external and internal task forces alongside industry partners. We developed a fiscal model to guide registration fees, budget tracking, and solicitation of conference funding from academic and industry sponsors. Attendees completed postconference surveys assessing speaker or workshop effectiveness and suggestions for future events. Finally, we developed an environmental scan survey to assess gender equity needs and existing programming across BTAA institutions. In June 2024, The Ohio State University hosted the inaugural CommUNITYten conference in Columbus, Ohio, featuring 5 keynote presentations, 9 breakout sessions, and networking opportunities across one and a half days of curated programming. Nearly 180 people attended, with representation from 9 BTAA institutions, 6 industry companies, staff, and trainees. Postconference surveys showed 50% (n=27) of respondents were likely to attend another in-person conference and suggested future conference topics. The environmental scan survey launched in October 2024. We successfully established the CommUNITYten collaborative and hosted the inaugural conference. Establishing key stakeholders from each BTAA institution, obtaining sponsorship, and detailed conference planning and partnerships were critical in ensuring realization of this collaborative. The conference brought together leaders, faculty, staff, trainees, and industry partners from across the country and met the initial goal of networking, sharing resources, and building community for women+ faculty. These efforts lay a robust foundation for the BTAA CommUNITYten collaborative to foster ongoing collaboration, innovation, and progress in the years to come. Given the importance of steady improvements, this viewpoint may further guide the efforts of other individuals, groups, and leadership supporting women+ as they consider approaches and strategies advocating for gender equity at the national level.

(JMIR Form Res 2025;9:e65561) doi:10.2196/65561

XSL•FO
RenderX

KEYWORDS

collaborative; gender equity; women in medicine; women in science; biomedical science; women; women+; gender; medicine; university; faculty; accessibility; career; equity; networking; opportunity; retaining; programming; Big Ten Academic Alliance; BTAA; academic alliance

Introduction

Women-identifying and women+ gender faculty (women+ is inclusive of different identities pertaining to gender, hereto described as women+ faculty) in medicine and biomedical sciences face numerous barriers to career advancement. These barriers include fewer invitations to keynote speaking engagements, reduced academic promotion rates, decreased leadership opportunities, and disparities in research funding and other research metrics [1-4]. For women with intersectional identities, these barriers multiply [5]. There are few prominent medicine-based national career development programs in the United States targeted to women, such as the Association of American Medical Colleges (AAMC) Early-Career and Mid-Career programs, the Hedwig van Ameringen Executive Leadership in Academic Medicine (ELAM) program, the University of Michigan's Rudi Ansbacher Advancing Women in Academic Medicine Leadership Scholars Program, and Harvard's Career Advancement and Leadership Skills for Women, with metrics demonstrating postparticipation career success and advancement [6-8]. Women who participate in these career development programs are more likely than men and nonparticipant women to be promoted to associate professor, and as likely as men and more likely than nonparticipant women to be promoted to full professor within 10 years [9]. Career development program participation also improves retention of women in the academic workforce and leadership opportunities, while expanding participants' networks and increasing visibility, both internally and externally at their institutions [10].

However, career development programs such as those from the AAMC and ELAM can be cost prohibitive and are often limited to one to two participants per institution [10]. Additionally, these programs do not accommodate faculty in medicine-adjacent fields such as biomedical engineering, veterinary medicine, and nursing. Attendance at shorter duration gender-specific conferences for women broadly in the medical and biomedical fields may provide similar benefits through delivery of content, networking, and professional support that are needed for career advancement [11]. Moreover, bridging the divide among health-related professions and bringing together women+ faculty across these fields may create a more effective cohort of professionals who can work synergistically to advance gender equity across academic institutions.

In 2022, leaders (all authors) from The Ohio State University College of Medicine's (OSUCOM's) Women in Medicine and Science (WIMS) created the Big Ten Academic Alliance (BTAA) collaborative for women in medicine and biomedical sciences with the goal of developing a network of women+ faculty in the BTAA to collaborate, share resources, provide support, and connect in a community that could be stronger together in combating gender barriers in these disciplines. Herein, we detail how we, as the leaders of this collaborative, developed a group of women+ individuals and respective institutions across the BTAA; how we crafted the inaugural conference; and how we are actively conducting an environmental scan survey and planning continuity of programming through conference host handoff, sharing of materials, and webinars.

Collaborative Development

Background and Overview

As a member of the Big Ten Conference, The Ohio State University (OSU) felt as if it was the natural place to start planning a collaborative to bring together peer institutions across the BTAA, an academic outcropping of this sports network. The BTAA was founded in 1958 [12]. Other BTAA conferences already exist in disciplines such as neuroscience; lipids; music education; women+ in technology; and Women in Science, Technology, Engineering, and Mathematics. The BTAA integrates across an existing network of deans, with central coordination and support in Rosemont, Illinois. The BTAA currently includes 14 universities. We chose to cast a broad net for the collaborative development and included institutions from all 19 prior, current, and future universities in the BTAA (Table 1). This BTAA collaborative of women in medicine and biomedical sciences would enable similar-sized universities to share effective programming and resources and work together to address academic gender inequity.

XSL•FO
RenderX

Table . The BTAA[a] institutions.

	Member of BTAA	Representatives on ETF (n)[b,c]
Indiana University	✓	4
Michigan State University	✓	3
Northwestern University	✓	3
Pennsylvania State University	✓	3
Purdue University	✓	2
Rutgers University	✓	1
The Ohio State University	✓	4
University of California, Los Angeles	✓[d]	1
University of Maryland	✓	2
University of Nebraska-Lincoln	✓	3
University of Chicago	✓[c]	1
University of Illinois	✓	2
University of Iowa	✓	1
University of Michigan	✓	2
University of Minnesota	✓	2
University of Oregon	✓[d]	0
University of Southern California	✓[d]	1
University of Washington	✓[d]	1
University of Wisconsin-Madison	✓	1

[a]BTAA: Big Ten Academic Alliance.

[b]ETF: external task force.

[c]Historical member.

[d]New BTAA member.

As we generated the vision and mission for this initiative, we benefited by learning from prior speakers hosted by OSUCOM WIMS, as well as from internal advisors. Key suggestions from our advisors were to ensure early discussions with planning partners, include broad representation of organizing bodies, gather data that could be published as part of the initiative, and listen to key stakeholders. We differentiated ourselves from existing groups, such as the BTAA Women in Science, Technology, Engineering, and Mathematics and the AAMC Group on Women in Medicine and Science (GWIMS) by focusing on support of gender equity for faculty across medicine and biomedical sciences, with additional shared resources and interactions among trainees, staff, and relevant industry members.

We also envisioned that once the CommUNITYten collaborative was formed, this group could work together on initiatives that would benefit the BTAA's future growth toward gender equity. Given the persistent national trend of women being promoted to full professor less frequently than men [1,13], we hope that this collaborative will provide future opportunities for networking, obtaining external talk and letters of evaluation for promotion, while also cultivating opportunities for networking, mentorship, collaboration, and sponsorship. Finally, we prioritized inclusivity and cultivating a partnership of interested industry companies, allies, staff, and leaders. Women faculty cannot solve the problems of gender inequity alone; we need the involvement and support of the entire community.

To solicit membership to this collaborative's leadership group, OSUCOM WIMS reached out to faculty representatives of the AAMC GWIMS to obtain contacts at the 19 BTAA institutions. Not all BTAA schools are part of GWIMS, so we also searched publicly available web pages dedicated to supporting women faculty at these institutions to find individuals who may be interested in planning the inaugural conference. OSUCOM WIMS also hosted a lunch roundtable at the 2022 AAMC Lead, Serve, Inspire annual meeting to gather additional participants [14]. In addition, we also relied on personal networks and connections to identify faculty members involved in initiatives that would align with the conference objectives. We were ultimately able to obtain at least one name and current contact information from 18 of the 19 BTAA institutions (Table 1). We included these individuals in all future email correspondence and meeting invitations. We also engaged with four industry representatives to bolster the networking component of this collaborative. These individuals together formed our external task force (ETF). We also created an internal task force (ITF)

XSL•FO
RenderX

of faculty from OSU for development and planning for the inaugural conference. Figure 1 displays the CommUNITYten organizational chart.

Figure 1. Reporting structure of BTAA CommUNITYten collaborative. BTAA: Big Ten Academic Alliance; ETF: external task force; ITF: internal task force.

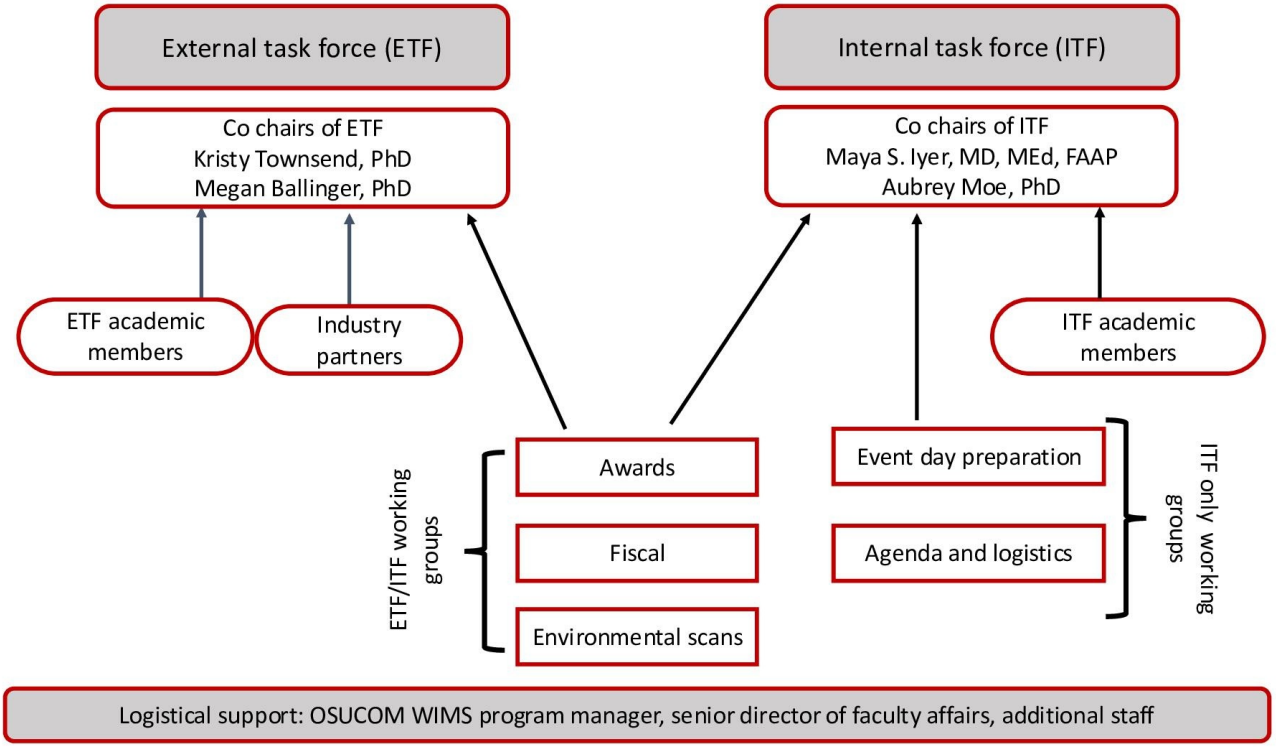

About ETF

The ETF began meeting by Zoom (Zoom Communications, Qumu Corporation) in June 2023. The ETF was led by two coauthors (MB and KT). During these calls, we refined the collaborative's mission, solicited speaker votes for the initial conference, and discussed and voted on logistics such as registration fees, the cadence of ETF meetings, solicitation of sponsorships, and communication tactics.

Industry Partners

In many cases, industry companies in pharma, biotech, or medicine are ahead of academics in their diversity, inclusion, and equity efforts. As we spoke with our four initial industry planning partners (Abbott, Bayer, Entrada, and Novartis), we learned that their efforts to educate on neurodiversity and to support women employee groups represented opportunities for academia to learn from industry. Furthermore, since academic-industry partnerships are key to successful academic medicine and biomedical research, and since these relationships can often be informal and dominated by men, we wanted to include industry attendees to bolster the networking goals of the conference. Industry could also benefit from learning about resources and best practices for gender equity in academic settings, and also benefit from networking opportunities for recruitment of potential employees. As part of our partnership with industry, we planned to have one panel discussion at the inaugural conference to learn more about industry equity initiatives, career opportunities, and translational research across academic-industry partnerships.

About ITF

An ITF was specifically established to plan for the inaugural CommUNITYten conference and to execute key deliverables and tasks related to event logistics. This task force was led by two coauthors (MSI and AM). Given that OSUCOM WIMS led the establishment of the CommUNITYten collaborative, the ITF members were solicited from existing WIMS members, but invites were also extended to working groups from our academic sponsors and ETF members. The ITF met monthly and was further divided into working groups. This approach permitted work delegation to specific individuals and allowed for identification of specific leaders for each working group, providing leadership opportunities. A description of the five working groups is below. The ITF also worked closely with the WIMS' program coordinator (coauthor JD), OSUCOM staff, and marketing teams.

ITF Working Groups

Agenda and Logistics

This working group focused on aspects of agenda planning, speaker communications, and coordination of timing for multiple tracks of speakers, seminars, and events during the day-and-a-half conference. Engagement and ongoing coordination with other working groups (fiscal and event day preparation) were critical to maintaining alignment of the agenda with CommUNITYten's goals and budgetary resources.

XSL•FO
RenderX

Event Day Preparation

This working group oversaw event logistics and worked directly under the WIMS program coordinator (coauthor JD). Specific items managed by this group included coordination of speaker transportation, audio or visual access and support, creation of program materials (including printed products and conference swag) and providing conference day assistance to attendees.

ITF and ETF Joint Working Groups

Awards

This group was led by two OSUCOM WIMS faculty members and had two additional BTAA members. Given the underrecognition of women recipients among societal and specialty-based awards [15-17], this working group intentionally developed three awards for the inaugural conference to recognize trainees and faculty at all academic ranks: (1) catalyst award, (2) rising star, and (3) luminary. The catalyst award is granted to a trainee (postdoctoral fellow, clinical fellow, or resident) who has demonstrated excellence, drive, and dedication in promoting women+ individuals in medicine and biomedical science or research. The rising star award recognizes early- or midcareer faculty (within 10 y of a full-time faculty position or rank of assistant professor) who have similar advanced women in their fields. The luminary award recognizes senior faculty members who have substantially contributed to the promotion and advancement of women+ individuals within or external to the nominee's institution. Through an iterative process, this group developed these award descriptions, announcements, and scoring rubrics. These awards were presented at the inaugural conference.

Fiscal

This working group developed the fiscal model to support the CommUNITYten conference and was coled by OSUCOM's senior director of faculty affairs and coauthor KT in conjunction with the entire ETF. Starting with an initial US $20,000 investment from OSUCOM WIMS, the fiscal working group developed a registration model to generate needed revenue (with a fee schedule based on input and votes from the ETF), and solicited academic and industry sponsorships. Fundraising for the inaugural CommUNITYten conference involved a 3-pronged approach: (1) internal solicitation of host institution's colleges and centers or institutes; (2) external solicitation of industry, nonprofit, or community organizations; and (3) academic sponsorship from other BTAA schools. Together, the fundraising from registrations and sponsorships would form the available budget for conference. Since we were fundraising as we were building the conference program, there was an iterative process whereby later fiduciary decisions (catering choices, for example) were dictated closer to the conference date and based on projections of confirmed incoming funds. By June 2024, we had raised approximately US $130,000 for the CommUNITYten collaborative.

To solicit sponsors, we generated a sponsor package that included an overview of the intent of the meeting and what would be gained by sponsoring across several options of sponsorship tiers (Multimedia Appendix 1). An Excel (Microsoft Corp) file was maintained to track sponsor calls, verbal commitments, and follow-up correspondence, or the "income" for the meeting. This file was compared to an Excel outline of expected expenditures (which would become a "budget," once figures were finalized). The goal was to increase income above the numbers in the projected budget, in time for the deadline to finalize space and catering contracts. The last component of this effort was generating a dedicated account at OSU to hold funds separate from the college operating budget. This would allow for easy transfer of funds to the next host institution. We prioritized transparency in our efforts, with detailed recordkeeping and fiscal responsibility that was shared with the ETF. As an example of fiscal responsibility, our initial keynote choice had a prohibitively high speaker fee, and we now hope to have this speaker for a webinar in the future using remaining carryover funds now that the inaugural conference is over.

Environmental Scan Surveys

This working group was led by two coauthors (MB and SM) in conjunction with the entire ETF. The group met monthly for 6 months to develop an institutional review board–approved research study protocol to ascertain current trends and faculty experiences across the BTAA institutions related to gender equity in medicine and biomedical science. The purpose for conducting a multi-institutional survey was to ask questions specifically on workplace climate, career promotion, leadership development, discrimination or bias, retention, mentorship, sponsorship, scholarship, pay equity, care responsibilities, and lactation or reproductive care services. This survey was piloted among non-BTAA faculty members to ensure understanding. The survey was launched in October 2024 and sent to conference participants or registrants, members of the ETF and other BTAA faculty contacts. Responses will be analyzed, and aggregated data will be shared among the participating BTAA institutions. Survey findings may identify opportunities for improvement, strategic interventions, and targeted programming that can benefit faculty across disciplines and shape programming initiatives for upcoming BTAA CommUNITYten conferences, future CommUNITYten collaborative work, and potentially a partnership with the BTAA and AAMC GWIMS going forward.

CommUNITYten 2024 Inaugural Conference

The inaugural "CommUNITYten: Big Ten Academic Alliance Conference for Women in Medicine and Biomedical Science" was held on Friday June 7 and Saturday June 8 at the Ohio Union in Columbus, Ohio. Day 1 had 171 attendees and day 2 had 78 attendees, out of 220 total registrants. The final agenda for the conference is listed in Multimedia Appendix 2. The conference included 5 keynote presentations, 9 breakout sessions, coffee breaks with interactive activities (mindfulness, conversations surrounding people of difference, and a session with our Buckeye Paws service dogs), and a networking event on the first evening of the conference. Breakout sessions included presentations on topics relevant to the collaborative mission (eg, gender equity, developing a professional narrative, and effective science communication) as well as panel events on publishing, allyship from men, and using social media as a

XSL•FO
RenderX

tool to establish professional presence. Finally, we awarded 1 catalyst, 1 luminary, and 2 rising star awards.

We conducted postconference evaluations (N=54 respondents). Among the respondents, on a 5-point Likert scale, 50% (n=27) reported that they would be likely to attend a future in-person BTAA CommUNITYten conference. In addition, 50% (n=27) stated that they would be likely to attend a webinar series offered by BTAA CommUNITYten. Nearly 60% (n=31) of respondents heard about the conference from a direct invitation. Open-ended comments showed that future topics should include negotiation; setting boundaries; navigating career changes; lesbian, gay, bisexual, transgender, or queer issues; creating networks; and clinical and translational research. Areas for improvement included shortening the conference to a one-day event, going beyond cis-normative topics, and including more time for networking.

Challenges and Lessons Learned

The initial challenge in creating the CommUNITYten collaborative was "building the plane while flying it." Our fiscal model was dependent on funds generated from registration costs, sponsorships, and the seed money from OSUCOM WIMS—all of which were unknowns when we initially embarked on this process. As a result, we had to establish our incoming funds flow simultaneously as we solicited speakers and panelists. We could not build a website until we had a set agenda. In addition, communicating with our ETF was challenging at the tail end of the academic year due to conflicts with scheduled meetings; loss of contacts (ie, individuals leaving their institutions due to retirement and promotions); and other academic obligations such as commencements, other conferences, and personal vacations.

We were fortunate to have the OSUCOM WIMS program coordinator (coauthor JD) serve as a project manager across all task forces and working groups. However, we recommend hosts of similar events bring on a full-time equivalent to support a host institution in planning a conference of this scale. Textbox 1 provides information on all the aspects of the conference that our program coordinator oversaw, including tasks carried out with our coleads, working group members, and support staff. Table 2 presents conference attendees' institutional affiliations.

XSL•FO
RenderX

Textbox 1. Conference planning and execution logistics. Logistical items listed were largely carried out by The Ohio State University College of Medicine (OSUCOM) Women in Medicine and Science's program coordinator for the inaugural CommUNITYten conference, with some items carried out by coleads, OSUCOM support staff, or working group members.

Conference planning

- Create an online file system (Microsoft Teams) to share information across internal and external partners.
- Schedule internal task force (ITF), external task force (ETF), marketing, and coleads meetings; send out reminder emails, meeting agendas, meeting notes, and postmeeting tasks.
- Research and secure space and catering orders.
- Preview facility or venue and work with venue event planner to confirm number of rooms, room layouts, technology needs, day of logistics, and build a timeline to meet venue deadlines.
- Collaborate with coleads and ETF and ITF on go/no go decisions such as event date or times, registration platform and wording, continuing medical education, fiscal model, Big Ten Academic Alliance (BTAA) logo usage, etc.
- Advertise through numerous channels, such as emails, campus-wide university newsletter, OSUCOM newsletters, and Twitter. Externally, the coleads advertised the conference through the Executive Leadership in Academic Medicine EDGE newsletter.
- Work with venue IT team for audio or video needs.
- Decide on and secure decorations (photo backdrop, centerpieces, or stage décor).
- Secure photographer.

Fiscal

- Schedule meetings for Women in Medicine and Science director and associate director to meet with The Ohio State University deans, center directors, and other leaders to request sponsorship.
- Meet with coleads and fiscal working group leads to review finances.
- Secure sponsor information: logo, marketing contact, fiscal contact, and contact for dissemination of information post conference.

Speakers or agenda

- Collate and track list of potential speakers, as determined by ETF, across price ranges and topics.
- Follow up on emails from coleads to potential speakers asking and securing availability, interest, and fees.
- Develop speaker intake forms (name, credentials, talk title and objectives, contact info, IT needs, travel needs, and payment forms).
- Track speaker information from intake through postevent payment process.
- Work with speakers on contracts and send to legal for approval.
- Set up logistics meetings with panelist groups and their moderators to understand their needs and desired speaker introductions.
- Set up logistics meetings with speakers and session presenters on talk time, topic, format, and logistics.
- Book speaker travel (lodging, flights, and transportation to the venue) and create individual speaker itineraries.
- Finalize agenda.
- Create day of speaker itineraries including travel, meetings, green room locations, and arrange shuttles or transport to and from the hotel to venue.
- Collect items for speaker gift bags to bring to the conference.
- Process speaker payments and honorariums, including receipts from speakers and create expense reports for travel reimbursement.

Marketing

- Create a comprehensive communication plan: Work with host institution marketing teams to create marketing language, create save the date PowerPoint (Microsoft Corp) slides, flyers, and email templates.
- Once approval is received from BTAA to use logo; work with BTAA representative to secure swag delivery before conference.
- Create registration site (Cvent) that includes deadlines, refund policy, option for single day registration versus full conference only, what information you want to gather from participants, what registration confirmation emails should look like, ability for registered participants to use an app or have a separate postregistration website, what sponsor logos will look like, what reports you will to need to monitor registration.
- Get input from coleads and event day logistics working group on aesthetic of Cvent site.
- Work with marketing team regularly to create and update QR codes, URL short links, content for social media posts, websites, and develop and pull daily reports from Cvent or conference registration website.
- Collaborate with marketing and event day preparation working group to install communications on electronic boards across institution.
- Add communications to electronic boards across institution.

XSL•FO
RenderX

- Create name tags.

Working groups (awards, environmental scan, event day preparation, fiscal, and speaker or agenda and logistics)

- Set up monthly group meetings, create or gather agendas, send reminder emails, and meeting notes and tasks.
- Develop a survey tool to capture award nominations and deadlines.
- Design and order plaques for the award recipients.
- Determine when and how to recognize awardees in event schedule with speaker or agenda working group.
- Market call for nomination emails to registered attendees and internally to participating colleagues.

Day-of-conference logistics

- Create run of show with day of assignments (décor, registration desk, speaker movement, monitor catering, audio-visual, space issues, and postevent clean-up).
- Inventory and gather supplies to take to venue (banners, signage, BTAA swag, speaker bags, BTAA notebooks, and pens).
- Create postevent surveys for feedback with coleads.
- Create rolling slide decks and master slide decks for each day, collaborate with audio-visual team at venue to ensure slide decks run smoothly.

Postconference

- Create and send thank you emails to speakers with coleads.
- Distribute postevent surveys for feedback for those that may have left early.
- Process venue or service-provider payments.
- Set up debrief meetings for ITF and ETF.
- Collate event survey data to send to coleads for analysis.

Table . Conference attendees' institutional affiliations.

Group or institution	Friday June 7, 2024 (n)	Saturday June 8, 2024 (n)
Industry	6	0
Indiana University	3	4
Northwestern University	6	2
Pennsylvania State University	1	1
Rutgers University	2	2
The Ohio State University	134	56
University of Illinois	2	0
University of Maryland	3	3
University of Michigan	6	4
Univeristy of Nebraska	3	3
Non–Big Ten Academic Alliance institution	5	3
Total	171	78

After hosting the conference, we learned several lessons to improve future conference gatherings. First, future conferences in this series could benefit from selecting annual conference titles and themes to align content around specific goals. Many speakers expressed a desire for this information to tailor their talks to align with the conference's intentions and goals beyond the broad gender equity focus. In addition, we learned the importance of engaging the institutional marketing team from the outset. Their support is crucial for successful outreach, within the institution and to external networks via LinkedIn. Before launching the registration platform, having a complete program agenda and speaker bios helps convert website visitors to registrants when they can see the full conference content. We also learned that shortening the conference to 1 day and avoiding programming on a weekend could boost attendance. Finally, if possible, we found that a 2-year timeline is needed to secure sponsorships and plan logistics associated with the in-person meeting.

XSL•FO
RenderX

Future Directions

One of the pillars in the development of this conference was to establish a network of like-minded women+ interested in bolstering gender equity in medicine and biomedical science. In order to provide equitable leadership opportunities and accessibility of the in-person conference geographically, the ETF decided that the host institution would rotate every other year for in-person conferences. In the interim nonconference years, a virtual webinar series will continue the momentum and the networks established by this collaborative, with allocated budget voted on by the ETF and ITF from our rollover funds. We anticipate that each BTAA institution will be responsible for sponsoring a speaker for the interim year webinars. Additionally, we hope that results from the environmental scan, once completed, will be presented in upcoming meetings or publications and used as a needs assessment for future CommUNITYten conferences in determining conference themes, workshops, sessions, and speakers.

In summary, the BTAA CommUNITYten collaborative successfully launched in 2024. Establishing key stakeholders from each BTAA institution, obtaining sponsorship, and detailed conference planning and partnerships were critical in ensuring realization of this group. The inaugural conference brought together leaders, faculty, staff, trainees, and industry partners from across the country and met the initial goal of networking and developing interinstitutional community. Our future initiatives of conducting environmental scans and sharing resources are in progress, as well as development of a national webinar series. These efforts lay a robust foundation for the BTAA CommUNITYten collaborative to foster ongoing collaboration, innovation, and progress in the years to come.

Acknowledgments

We would like to thank Dean Carol Bradford, MD, MS, and Vice Dean Tatiana Oberyszyn, PhD, for their unwavering support of this initiative. We would also like to thank Kathy Tober, PhD, for her logistical support of the conference, for coleading the Fiscal Task Force, and for overseeing all financial aspects of the initial event. In addition, we would like to thank Alex Hildebrandt, MA, Kim Knight, BA, and the members of the College of Medicine Marketing Team for their excellent support of this event. The coauthors wish to thank all members of the internal task force, external task force, and working groups that made the inaugural conference successful. Finally, we acknowledge the key contributions of our inaugural sponsors. Generative artificial intelligence was not used in any portion of writing this paper.

Authors' Contributions

MSI did the conceptualization, data curation, formal analysis, funding acquisition, methodology, supervision, writing of the original draft, and reviewing and editing of the writing. AM handled writing of the original draft, and reviewing and editing of the writing. SM handled the reviewing and editing of the writing. JD worked on the project administration, writing of the original draft, and reviewing and editing of the writing. MB wrote the original draft and reviewed and edited the writing. KT assisted in conceptualization, data curation, formal analysis, funding acquisition, methodology, supervision, writing of the original draft, and reviewing and editing of the writing. MSI and AM served as co-first authors and MB and KT served as co-senior authors for this manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Sponsorship tiers.
[DOCX File, 21 KB - formative_v9i1e65561_app1.docx]

Multimedia Appendix 2
Conference agenda.
[PDF File, 7174 KB - formative_v9i1e65561_app2.pdf]

References

1. Richter KP, Clark L, Wick JA, et al. Women physicians and promotion in academic medicine. N Engl J Med 2020 Nov 26;383(22):2148-2157. [doi: 10.1056/NEJMsa1916935] [Medline: 33252871]
2. Carr PL, Raj A, Kaplan SE, Terrin N, Breeze JL, Freund KM. Gender differences in academic medicine: retention, rank, and leadership comparisons from the National Faculty Survey. Acad Med 2018 Nov;93(11):1694-1699. [doi: 10.1097/ACM.0000000000002146] [Medline: 29384751]
3. Amrein K, Langmann A, Fahrleitner-Pammer A, Pieber TR, Zollner-Schwetz I. Women underrepresented on editorial boards of 60 major medical journals. Gend Med 2011 Dec;8(6):378-387. [doi: 10.1016/j.genm.2011.10.007] [Medline: 22153882]

XSL•FO
RenderX

4. Marcelin JR, Bares SH, Fadul N. Improved infectious diseases physician compensation but continued disparities for women and underrepresented minorities. Open Forum Infect Dis 2019 Feb;6(2):ofz042. [doi: 10.1093/ofid/ofz042] [Medline: 30815507]
5. Verduzco-Gutierrez M, Wescott S, Amador J, Hayes AA, Owen M, Chatterjee A. Lasting solutions for advancement of women of color. Acad Med 2022 Nov 1;97(11):1587-1591. [doi: 10.1097/ACM.0000000000004785] [Medline: 35731581]
6. Helitzer DL, Newbill SL, Morahan PS, et al. Perceptions of skill development of participants in three national career development programs for women faculty in academic medicine. Acad Med 2014 Jun;89(6):896-903. [doi: 10.1097/ACM.0000000000000251] [Medline: 24871241]
7. The Hedwig van Ameringen Executive Leadership in Academic Medicine. Drexel University College of Medicine. URL: https://drexel.edu/medicine/academics/womens-health-and-leadership/elam/ [accessed 2023-12-11]
8. Woitowich NC, Graff SL, Swaroop M, Jain S. Gender-specific conferences and symposia: a putative support structure for female physicians. J Womens Health (Larchmt) 2020 Aug;29(8):1136-1141. [doi: 10.1089/jwh.2019.7859] [Medline: 31697577]
9. Chang S, Guindani M, Morahan P, Magrane D, Newbill S, Helitzer D. Increasing promotion of women faculty in academic medicine: impact of national career development programs. J Womens Health (Larchmt) 2020 Jun;29(6):837-846. [doi: 10.1089/jwh.2019.8044] [Medline: 32466701]
10. Chang S, Morahan PS, Magrane D, et al. Retaining faculty in academic medicine: the impact of career development programs for women. J Womens Health (Larchmt) 2016 Jul;25(7):687-696. [doi: 10.1089/jwh.2015.5608] [Medline: 27058451]
11. Salem V, McDonagh J, Avis E, Eng PC, Smith S, Murphy KG. Scientific medical conferences can be easily modified to improve female inclusion: a prospective study. Lancet Diabetes Endocrinol 2021 Sep;9(9):556-559. [doi: 10.1016/S2213-8587(21)00177-7] [Medline: 34332698]
12. History of the Big Ten Alliance. Big Ten Academic Alliance. URL: https://btaa.org/about/history [accessed 2024-05-15]
13. 2018-2019: the state of women in academic medicine: exploring pathways to equity. AAMC. 2018. URL: https://www.aamc.org/data-reports/faculty-institutions/report/state-women-academic-medicine [accessed 2024-05-15]
14. Learn serve lead: the AAMC annual meeting. AAMC. URL: https://www.aamc.org/career-development/events/learn-serve-lead [accessed 2024-12-07]
15. Jacobs JW, Bibb LA, Allen ES, et al. Women and non-white people among Lasker Award recipients from 1946 to 2022: cross sectional study. BMJ 2023 May 17;381:e074968. [doi: 10.1136/bmj-2023-074968] [Medline: 37197781]
16. Neikirk K, Vue Z, Vue N, et al. Disparities in funding for Nobel Prize awards in medicine and physiology across nationalities, races, and gender. J Cell Physiol 2024 Jul;239(7):e31157. [doi: 10.1002/jcp.31157] [Medline: 38225913]
17. Silver JK, Slocum CS, Bank AM, et al. Where are the women? the underrepresentation of women physicians among recognition award recipients from medical specialty societies. PM R 2017 Aug;9(8):804-815. [doi: 10.1016/j.pmrj.2017.06.001] [Medline: 28606837]

Abbreviations

AAMC: Association of American Medical Colleges
BTAA: Big Ten Academic Alliance
ELAM: Executive Leadership in Academic Medicine
ETF: external task force
GWIMS: Group on Women in Medicine and Science
ITF: internal task force
OSU: The Ohio State University
OSUCOM: The Ohio State University College of Medicine
WIMS: Women in Medicine and Science

Edited by A Mavragani; submitted 25.08.24; peer-reviewed by D Helitzer, GS Booth; revised version received 14.11.24; accepted 14.11.24; published 23.01.25.

Please cite as:
Iyer MS, Moe A, Massick S, Davis J, Ballinger M, Townsend K
Development of the Big Ten Academic Alliance Collaborative for Women in Medicine and Biomedical Science: "We Built the Airplane While Flying It"
JMIR Form Res 2025;9:e65561
URL: https://formative.jmir.org/2025/1/e65561
doi:10.2196/65561

XSL•FO
RenderX

© Maya S Iyer, Aubrey Moe, Susan Massick, Jessica Davis, Megan Ballinger, Kristy Townsend. Originally published in JMIR Formative Research (https://formative.jmir.org), 23.1.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Applying Critical Discourse Analysis to Cross-Cultural Mental Health Recovery Research

Yasuhiro Kotera[1,2], PhD; Riddhi Daryanani[3], MSc; Oliver Skipper[3], MSc; Jonathan Simpson[3], MSc; Simran Takhi[1], MSc; Merly McPhilbin[1], MSc; Benjamin-Rose Ingall[1], MSc; Mariam Namasaba[3], PhD; Jessica Jepps[3], MSc; Vanessa Kellermann[3], MSc; Divya Bhandari[4], MSc; Yasutaka Ojio[5], PhD; Amy Ronaldson[3], PhD; Estefania Guerrero[6], MSc; Tesnime Jebara[3], PhD; Claire Henderson[3], PhD; Mike Slade[1,7], PhD; Sara Vilar-Lluch[8], PhD

[1]Institute of Mental Health, School of Health Sciences, University of Nottingham, Nottingham, United Kingdom
[2]Center for Infectious Disease Education and Research, Osaka University, Suita, Japan
[3]Health Service and Population Research Department, King's College London, London, United Kingdom
[4]Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA, United States
[5]Department of Community Mental Health & Law, National Institute of Mental Health, Tokyo, Japan
[6]Department of Social Psychology and Quantitative Psychology, University of Barcelona, Barcelona, Spain
[7]Faculty of Nursing and Health Sciences, Nord University, Namsos, Norway
[8]School of English, University of Nottingham, Nottingham, United Kingdom

Corresponding Author:
Yasuhiro Kotera, PhD
Institute of Mental Health
School of Health Sciences
University of Nottingham
Triumph Road
Nottingham, NG7 2TU
United Kingdom
Phone: 44 01158231294
Email: yasuhiro.kotera@nottingham.ac.uk

Abstract

The purpose of this paper is to demonstrate how critical discourse analysis (CDA) frameworks can be used in cross-cultural mental health recovery research. CDA is a qualitative approach that critically appraises how language contributes to producing and reinforcing social inequalities. CDA regards linguistic productions as reflecting, consciously or unconsciously, the narrators' understandings of, or attitudes about, phenomena. Mental health recovery research aims to identify and address power differentials, making CDA a potentially relevant approach. However, CDA frameworks have not been widely applied to mental health recovery research. We adapted established CDA frameworks to our cross-cultural mental health recovery study. The adapted methodology comprises (1) selecting discourses that indicate positive changes and (2) considering sociocultural practices informed by relevant cultural characteristics identified in our previous research, without placing value judgments. Our adapted framework can support cross-cultural mental health recovery research that uses CDA.

(JMIR Form Res 2025;9:e64087) doi:10.2196/64087

KEYWORDS

critical discourse analysis; cross-cultural mental health recovery research; linguistic analysis; social inequality; mental health; recovery research; language; social inequalities; qualitative analytical approach; linguistic expressions; discourse; analysis; framework; inequalities; CDA

Critical Discourse Analysis and Mental Health

Critical discourse analysis (CDA) is a qualitative analytical approach that critically appraises how language contributes to the production and reproduction of social inequalities through the examination of authentic uses of language [1,2]. CDA considers that linguistic expressions reflect the speakers' and writers' conscious or unconscious perceptions or opinions towards phenomena [1,2]. CDA used in mental health research

XSL•FO
RenderX

is grounded in the belief that people's mental illness experiences are represented in the language they use [3]. This belief underscores the value of analyzing various forms of textual expression, such as first-person narratives and other linguistic representations. CDA can offer profound insights into the personal accounts of mental illness experiences and the societal perceptions surrounding specific mental illnesses.

The application of CDA to mental health has become an important area of research. Informed by corpus linguistics approaches (software-assisted examination of large collections of digitized texts), scholars have delved into media depictions of mental health conditions, as evidenced by Karaminis et al's [4] examination of autism portrayal in the British press. Likewise, Balfour [5] explored schizophrenia representations in the British press, while Price [6] scrutinized the perpetuation of mental health stereotypes resulting from media portrayals. The critical approach of CDA turns linguistic analysis into an engaged social enterprise; in these studies, it reveals engrained prejudices towards mental health conditions reproduced in mainstream media and how stereotypical representations have evolved over the years.

Moreover, researchers have turned their attention to the rich landscape of digital informal interactions among service users and nonprofessional caregivers. Digital forums and support groups have become significant spaces where individuals share their experiences, exchange information, and provide emotional support related to illnesses. For instance, based on corpus methods, Hunt and Brookes [7] shed light on discussions surrounding anorexia, depression, and diabulimia, while Jones et al [8] appraise the experiences of individuals with psychosis and bipolar disorder. Corpus-informed CDA is actively used to identify patterns of language use in large text-data collections, which deepens understanding of how language shapes our perceptions of mental health and illness.

First-person narratives of mental health experiences are another emerging field. This includes investigations into autobiographical accounts, such as stylistic studies on the experiences of voice hearing narrated by individuals with schizophrenia [9,10]. These stylistic studies have provided refined typologies of voices in auditory hallucinations in schizophrenia, bringing new insights into the phenomenology of the condition. Additionally, CDA studies have evaluated naturally occurring narratives, such as the examination of men's accounts of depression [11]. CDA research on naturally occurring narratives highlights the importance of lived-experience insights in gaining a more nuanced understanding of living with a mental health condition, both from individual and social perspectives. CDA studies, and discourse studies more generally, have also examined the construction of the self in interaction—eg, Meehan and MacLachlan [12] on schizophrenia, and Kłosińska and Ziółkowska [13] on dementia. Studies on self-identity construction among individuals with dementia, for example, have provided evidence that a sense of self is maintained in the later stages of dementia. This is achieved through the reliance on previous social identities and lexical repertoires to give meaning to the present [13]. A tenet of CDA is that discourse is ideological. In studying narratives of illness and recovery, this may involve examining how dominant discourses around identity (eg, gender, ethnicity) can influence the experience of illness, potentially affecting treatment and recovery (eg, normative ideologies of masculinity may underpin men's experience of depression [11]).

CDA has been employed extensively within mental health research. Nevertheless, the specific application of CDA within one area of mental health research, namely "mental health recovery research," remains largely unexplored to date [14]. In this paper, we will present one example of using established CDA frameworks in our cross-cultural mental health recovery study [15].

Mental Health Recovery

In mental health research, mental health recovery has been receiving attention [16]. Mental health recovery is "a deeply personal, unique process of changing one's attitudes, values, feelings, goals, skills, and/or roles" and "a way of living a satisfying, hopeful, and contributing life even within the limitations caused by illness" [17]. Approaches oriented towards mental health recovery, known as "recovery-oriented approaches," have been implemented in mental health services globally, and their positive impacts have been reported, such as empowerment, hope, autonomy, and reduced stigma in service users [18,19], as well as positive attitudes and understanding towards service users in mental health staff [20]. Recovery-oriented approaches are also helpful in understanding the mental health experiences of people with mental disorders such as bipolar disorders [21,22]. Owing to these benefits, mental health recovery has been embedded in many national policies [23-29].

Both CDA and mental health recovery value highlighting and addressing inequalities. CDA's fundamental principle is to identify inequalities in linguistic and social practices [30]. Similarly, mental health recovery aims to address power differentials in mental health practice and places the person living with a mental health experience at the center [31]. Mental health recovery is deeply rooted in social justice principles, empowering marginalized voices, reducing stigma and discrimination, and promoting equality [32]. This overlap between CDA and mental health recovery research suggests their compatibility, and yet the application of CDA to this field remains relatively unexplored. The current status of psychiatric discourse is diagnostic-driven rather than person-centered, marginalizing service users' lived experience rather than empowering them, biologically deficit-focused rather than strength-based, and regarding service users as passive patients rather than as individuals with unique needs, goals, and human rights [33]. Research on mental health recovery, empowered by CDA, has the potential to drive cultural changes in the current psychiatric discourse.

Our Study

Recovery colleges (RCs) are one of the recovery-oriented approaches that are actively used globally [34] and are regarded as a mental health innovation [35]. RCs are learning-based

XSL•FO
RenderX

support systems offering information, social support, and skill development for people with mental health symptoms, carers, and staff [36]. RCs are open to anyone interested in mental health and recovery, such as service users, carers, and professionals. People can directly self-enroll in courses offered at an RC. RCs are often free to use (eg, those offered by NHS) or may charge a modest fee (eg, £3 [equivalent to US$ 3.73 as of January 28, 2025] in some RCs in Japan). The duration and length of RC courses vary (eg, a one-off 2-hour session or 5 weekly 3-hour mindfulness sessions). RCs are advertised through various channels such as mental health services, digital platforms (eg, social media), community outreach (eg, flyers), and word of mouth [15,34]. Since the first RC was developed in England in 2009, there have been 221 RCs established in 28 countries within 15 years [34].

Our recent RC study was informed by CDA frameworks to be able to highlight textual emphases in the RC promotional texts in Japan and England and compared them [15]. We explored how RCs are promoted to the public in Japan and England and compared textual and thematic differences from cross-cultural perspectives. The textual emphases found in each country were in line with the cultural characteristics associated with the RC operational model [37], namely collectivism versus individualism and long-term orientation versus short-term orientation [38]. The texts in RCs in Japan emphasized collectivism (eg, “learn together”) and long-term orientation (eg, highlighting the presence of difficulties now), whereas those in RCs in England emphasized Individualism (eg, “self-management”) and short-term orientation (eg, focusing on skill acquisition) [15].

Specific recommendations were made for the RC operational model, which comprises 12 components [37]. The RC operational model was aligned with individualism and short-term orientation [38]. Our findings indicate some of the components may need to change in order to include under-recognized cultural characteristics such as collectivism and long-term orientation. People in Japan see the RC promotional texts that emphasize collectivism and long-term orientation and attend an RC expecting to engage with activities associated with collectivism and long-term orientation. RCs in Japan need to adapt their operations to meet those expectations, which work disadvantageously to achieve some of the components. For example, one of the components—component 3—evaluates whether an RC actively enquires about the student’s individual needs. This activity is more accepted in individualism than in collectivism. In collectivism, actively enquiring about individual needs can be regarded as rude because expressing individual needs may violate group harmony [39]. Therefore, we proposed that component 3 should include collectivism to ensure RCs oriented to this cultural characteristic are included.

Critical Discourse Analysis in Recovery Research

Our cross-cultural recovery research was informed by established CDA frameworks [1,2,40]. CDA involves textual analysis [1]. In our study, we examined whether the promotional texts of 61 RCs in England and those of all the 13 RCs in Japan reflected any cultural characteristics in their descriptions of recovery and support provided. The analysis was supported by corpus linguistics tools, which involve a computer-assisted examination of digitized texts (eg, [41,42]). In the analysis, we (1) considered the contexts of production and reception [38] and (2) interpreted the results in relation to sociocultural practices (Table 1). We adapted 2 stages of traditional CDA to better fit cross-cultural mental health recovery research: “selection of the discourse” and “consideration of the sociocultural practice” (see Table 1 for the original versions and adaptations).

Since our study considered RCs, our “selection of the discourse” focused on texts highlighting positive changes rather than inequalities (ie, RCs promotional material), as informed by Bartlett’s and Martin’s positive approach [43,44]. Traditionally, CDA studies predominantly focus on highlighting inequalities and discriminatory practices [46]. This approach is not fully applicable to recovery research that aims not only to highlight individuals experiencing difficulty and societal injustice encompassing social, political, and rights aspects around mental health and recovery [47] but also to underscore the positive impacts on people living with mental health symptoms [19]. Many people living with mental health symptoms, and those who work with them, are well aware of inequalities [48]. In addition to recognizing inequalities, they need changes that address inequalities [35]. Therefore, we have employed a positive approach and considered textual productions aimed at promoting inclusive mental health recovery practices (ie, RCs).

The second adaptation was made when considering the relevance of textual findings for sociocultural practice. The “consideration of sociocultural practice” was informed by cross-cultural findings. In our previous study [38], we used Hofstede’s cross-cultural indexes [45] to identify cultural characteristics that were associated with the RC operational model. These cultural characteristics informed how we interpreted the results of the textual analysis, which revealed that different themes were emphasized in RCs in Japan and England. Drawing our interpretation on empirical cross-cultural research also allowed us to reduce potential value judgments. Cross-cultural theories aim to *understand* different cultures, thus seeking to explain cultural differences without placing value judgments on them [49]. In highlighting inequalities reproduced in texts, many CDA studies have critically evaluated those aspects of advantaged groups that perpetuate discrimination. This explicit critical stance has led to criticisms of being biased for formulating value judgments according to a predefined agenda [50,51]. These evaluative assessments, however, can conflict with the aim of cross-cultural theories to understand different cultures. Therefore, our interpretation of textual differences was informed by empirical cross-cultural research [38] and a cross-cultural theory [45] to move away from value judgments.

The implications of adopting these 2 adaptations—“selection of the discourse” and “consideration of the sociocultural practice”—resulted in tangible recommendations for positive changes. Specific changes were proposed in the discussion section of our study paper, such as the suggestion for component 3 to include collectivism as noted above [15].

XSL•FO
RenderX

Table 1. Adapted version of the analytical framework for critical discourse analysis [1,2,40].

Stage of analysis	Description	How it is addressed in our work
Preparation (1): select the discourse	Select a discourse according to your research interests. Originally focused on addressing social injustices and inequalities, (critical) discourse analysis is also applied to discourses that promote positive social change [43,44] **Original version** Select a discourse related to injustice or inequality in society.	Selection of recovery colleges (RCs) promotional texts, a relatively new mental health support system that promotes individual empowerment and recovery through learning in community.
Preparation (2): data gathering	Select data sources, consider any ethical implications involved in data gathering, and prepare the data for analysis.	Identification of RCs in England and Japan, and retrieval of relevant descriptions from their information websites, notably focusing on those presenting RCs and recovery to the public. Translation of Japan RCs descriptions into English. Preparation of 2 .TXT files[a] (Japan RCs and England RCs datasets) to be used with the software Sketch Engine.
Text analysis (micro-level)	According to research interests: Identify the major underlying themes and subthemes. Examine linguistic choices used to represent social actors or events. Examine the stance taken by the author or speaker. Examine whether the text includes references to other texts (intertextuality).	Corpus linguistics-based analysis supported with the software Sketch Engine to retrieve keywords (single and multi-words) and identification of key themes for each dataset. Examination of concordances (keywords in context) guided by the RQs (ie, construal of RCs and recovery in England and Japan RCs promotional texts). Focus on both the portrayal of the RCs and the role attribution to service users.
Discourse practice (meso-level)	Examine the contexts of production and reception of the text. Consider the goal of the text, who has produced it, and the putative audience.	The context of production (RCs as mental health intervention) and the values of producers (RC managerial staff) and putative audience (RC students) have been studied [38]. For the purpose of this study (ie, construal of RC and recovery), the linguistic analysis has not considered characteristics of the textual register (promotional texts of medical services).
Sociocultural practice (macro-level)	Consider relevant sociocultural or historical factors (ie, context of production) that have conditioned the text. Consider whether the text reflects any sociocultural values. **Original version** Examine social relations that control the production of the text; in addition, examine the reciprocal relations (how the texts affect social practices and structures). How do social practices inform the arguments in the text? How does the text in turn influence social practices?	Examination of RCs as a new mental health support system and its main underpinning philosophies. Interpretation of the main themes emerging from the linguistic analysis based on Hofstede's cultural dimensions theory [45] and Kotera et al's [38] study on the impact of culture on the RC operational model.

[a]TXT files: text files.

Conclusions

As both CDA and mental health recovery research share the common goal of identifying and addressing inequalities, our adaptation of established CDA frameworks to cross-cultural mental health recovery research represents a crucial step forward. This paper offered valuable insights to researchers seeking to explore recovery in cross-cultural contexts using CDA. We hope our approach will foster a more evidence-based understanding of recovery and reduce inequalities.

Acknowledgments

We would like to thank Nigel Henderson who helped facilitate the completion of RC surveys in Scotland. We thank the RECOLLECT Lived Experience Advisory Panel (LEAP) who provided input into the design of the survey and interpretation of results. MS acknowledges the support of NIHR Nottingham Biomedical Research Centre.

Conflicts of Interest

None declared.

References

1. Fairclough N. Critical Discourse Analysis. 2nd ed. London, England: Routledge; 2010.

XSL•FO
RenderX

2. Fairclough N. Language and Power. 3rd ed. London, England: Routledge; 2014.
3. Demjén Z, Atkins S, Semino E. Researching Language and Health. London: Routledge; 2023.
4. Karaminis T, Gabrielatos C, Maden-Weinberger U, Beattie G. Portrayals of autism in the British press: a corpus-based study. Autism 2023;27(4):1092-1114 [FREE Full text] [doi: 10.1177/13623613221131752] [Medline: 36317371]
5. Balfour J. Representing Schizophrenia in the Media. London, England: Routledge; 2023.
6. Price H. The Language of Mental Illness. Cambridge, England: Cambridge University Press; 2022.
7. Hunt D, Brookes G. Corpus, Discourse and Mental Health. London, England: Bloomsbury Academic; 2020.
8. Jones S, Atanasova D, Dodd S, Flowers S, Rosala-Hallas A, Robinson H, et al. Use of an online forum for relatives of people with psychosis and bipolar disorder: mixed methods study. JMIR Ment Health 2022;9(10):e35837 [FREE Full text] [doi: 10.2196/35837] [Medline: 36264621]
9. Demjén Z, Semino E. Stylistics: Mind style in an autobiographical account of schizophrenia. In: Brookes G, Hunt D, editors. Analysing Health Communication: Discourse Approaches. London: Palgrave; 2021.
10. Demjén Z, Semino E. Henry's voices: the representation of auditory verbal hallucinations in an autobiographical narrative. Med Humanit 2015;41(1):57-62. [doi: 10.1136/medhum-2014-010617] [Medline: 25505160]
11. Galasinski D. Men's Discourses of Depression. Basingstoke, England: Palgrave Macmillan; 2008.
12. Meehan T, MacLachlan M. Self construction in schizophrenia: a discourse analysis. Psychol Psychother 2010;81(2):131-142. [doi: 10.1348/147608307x256777]
13. Kłosińska U, Ziółkowska J. Discourses of work life in narrative identities of people with advanced dementia. Int J Geriatr Psychiatry 2023;38(5):e5921. [doi: 10.1002/gps.5921] [Medline: 37173832]
14. Mancini MA, Rogers R. Narratives of recovery from serious psychiatric disabilities: a critical discourse analysis. Crit Approach Discourse Anal Disciplines 2015;1(2):35-40. [doi: 10.4135/9781849208765.n8]
15. Kotera Y, Miyamoto Y, Vilar-Lluch S, Aizawa I, Reilly O, Miwa A, et al. Cross-cultural comparison of recovery college implementation between Japan and England: corpus-based discourse analysis. Int J Ment Health Addict 2024 [FREE Full text] [doi: 10.1007/s11469-024-01356-3]
16. Slade M, Amering M, Farkas M, Hamilton B, O'Hagan M, Panther G, et al. Uses and abuses of recovery: implementing recovery-oriented practices in mental health systems. World Psychiatry 2014;13(1):12-20 [FREE Full text] [doi: 10.1002/wps.20084] [Medline: 24497237]
17. Anthony WA. Recovery from mental illness: the guiding vision of the mental health service system in the 1990s. Psychosocial Rehabil J 1993;16(4):11-23. [doi: 10.1037/h0095655]
18. Ellison ML, Belanger LK, Niles BL, Evans LC, Bauer MS. Explication and definition of mental health recovery: a systematic review. Adm Policy Ment Health 2018;45(1):91-102. [doi: 10.1007/s10488-016-0767-9] [Medline: 27709376]
19. Dell NA, Long C, Mancini MA. Models of mental health recovery: an overview of systematic reviews and qualitative meta-syntheses. Psychiatr Rehabil J 2021;44(3):238-253. [doi: 10.1037/prj0000444] [Medline: 33734781]
20. McPherson P, Lloyd-Evans B, Dalton-Locke C, Killaspy H. A systematic review of the characteristics and efficacy of recovery training for mental health staff: implications for supported accommodation services. Front Psychiatry 2021;12:624081 [FREE Full text] [doi: 10.3389/fpsyt.2021.624081] [Medline: 34054593]
21. Jagfeld G, Lobban F, Humphreys C, Rayson P, Jones SH. How people with a bipolar disorder diagnosis talk about personal recovery in peer online support forums: corpus framework analysis using the POETIC framework. JMIR Med Inform 2023;11:e46544 [FREE Full text] [doi: 10.2196/46544] [Medline: 37962520]
22. Jagfeld G, Lobban F, Marshall P, Jones SH. Personal recovery in bipolar disorder: systematic review and 'best fit' framework synthesis of qualitative evidence - a POETIC adaptation of CHIME. J Affect Disord 2021;292:375-385 [FREE Full text] [doi: 10.1016/j.jad.2021.05.051] [Medline: 34139411]
23. Department of Health and Ageing. Fourth National Mental Health Plan: An Agenda for Collaborative Government Action in Mental Health 2009. Canberra: Commonwealth of Australia; 2009.
24. Mental Health Commission of Canada. Changing Directions, Changing Lives: The Mental Health Strategy for Canada. Calgary: Mental Health Commission of Canada; 2012.
25. Department of Health Social Services and Public Safety of Northern Ireland. In: Revised Service Framework for Mental Health and Wellbeing 2018-2021. Belfast: Department of Health Social Services and Public Safety; 2018.
26. New Freedom Commission on Mental Health. Achieving the promise: transforming mental health care in America report. U.S. Department of Health and Human Services. 2003. URL: https://dmh.mo.gov/media/pdf/achieving-promise-transforming-mental-health-care-america-report [accessed 2025-02-06]
27. No health without mental health: a cross-government mental health outcomes strategy for people of all ages. HM Government. 2011. URL: https://assets.publishing.service.gov.uk/media/5a7c348ae5274a25a914129d/dh_124058.pdf [accessed 2025-02-06]
28. Slade M, Adams N, O'Hagan M. Recovery: past progress and future challenges. Int Rev Psychiatry 2012;24(1):1-4 [FREE Full text] [doi: 10.3109/09540261.2011.644847] [Medline: 22385420]
29. National Mental Health Development Plan No. 1 (2018-2037). Strategy and Planning Division, Department of Mental Health. Department of Mental Health-Thailand. Nonthaburi: Department of Mental Health-Thailand; 2019. URL: https://dmh-elibrary.org/items/show/450 [accessed 2025-01-23]

XSL•FO
RenderX

30. Blommaert J, Bulcaen C. Critical discourse analysis. Annu Rev Anthropol 2000;29(1):447-466 [FREE Full text] [doi: 10.1146/annurev.anthro.29.1.447]
31. Slade M, Longden E. Empirical evidence about recovery and mental health. BMC Psychiatry 2015;15(1):285 [FREE Full text] [doi: 10.1186/s12888-015-0678-4] [Medline: 26573691]
32. Slade M. Personal Recovery and Mental Illness: A Guide for Mental Health Professionals. Cambridge: Cambridge University Press; 2009.
33. Crowe M. Psychiatry and/or recovery: a critical analysis. Int J Ment Health Nurs 2022;31(6):1542-1551 [FREE Full text] [doi: 10.1111/inm.13072] [Medline: 36379908]
34. Hayes D, Hunter-Brown H, Camacho E, McPhilbin M, Elliott RA, Ronaldson A, et al. Organisational and student characteristics, fidelity, funding models, and unit costs of recovery colleges in 28 countries: a cross-sectional survey. Lancet Psychiatry 2023;10(10):768-779. [doi: 10.1016/s2215-0366(23)00229-8]
35. Whitley R, Shepherd G, Slade M. Recovery colleges as a mental health innovation. World Psychiatry 2019;18(2):141-142 [FREE Full text] [doi: 10.1002/wps.20620] [Medline: 31059628]
36. Kotera Y, Ronaldson A, Hayes D, Hunter-Brown H, McPhilbin M, Dunnett D, et al. Cross-cultural insights from two global mental health studies: self-enhancement and ingroup biases. Int J Ment Health Addiction 2024. [doi: 10.1007/s11469-024-01307-y]
37. Toney R, Knight J, Hamill K, Taylor A, Henderson C, Crowther A, et al. Development and evaluation of a recovery college fidelity measure. Can J Psychiatry 2019;64(6):405-414 [FREE Full text] [doi: 10.1177/0706743718815893] [Medline: 30595039]
38. Kotera Y, Ronaldson A, Hayes D, Hunter-Brown H, McPhilbin M, Dunnett D, et al. 28-country global study on associations between cultural characteristics and recovery college fidelity. Npj Ment Health Res 2024;3(1):46. [doi: 10.1038/s44184-024-00092-9] [Medline: 39379618]
39. Kotera Y, Ozaki A, Miyatake H, Tsunetoshi C, Nishikawa Y, Kosaka M, et al. Qualitative investigation into the mental health of healthcare workers in Japan during the COVID-19 pandemic. Int J Environ Res Public Health 2022;19(1):568 [FREE Full text] [doi: 10.3390/ijerph19010568] [Medline: 35010828]
40. Mullet DR. A general critical discourse analysis framework for educational research. J Adv Acad 2018;29(2):116-142. [doi: 10.1177/1932202x18758260]
41. Hunston S. Corpora in Applied Linguistics. Cambridge: Cambridge University Press; 2002.
42. Flowerdew L. Corpus-based discourse analysis. The Routledge Handbook of Discourse Analysis. London: Routledge; 2023.
43. Bartlett T. Positive discourse analysis 1. In: Flowerdew J, Richardson JE, editors. The Routledge Handbook of Critical Discourse Studies. London: Routledge; 2017.
44. Martin JR. Positive discourse analysis: solidarity and change. Revista Canaria Estudios Ingleses 2004;49:179-200. [doi: 10.25145/j.recaesin]
45. Hofstede G, Minkov M. Values Survey Module 2013 Manual. Geert Hofstede BV. URL: https://geerthofstede.com/wp-content/uploads/2016/07/Manual-VSM-2013.pdf [accessed 2025-01-23]
46. Liu K, Guo F. A review on critical discourse analysis. Theory Pract Lang Stud 2016;6(5):1076. [doi: 10.17507/tpls.0605.23]
47. Llewellyn-Beardsley J, Rennick-Egglestone S, Callard F, Crawford P, Farkas M, Hui A, et al. Characteristics of mental health recovery narratives: systematic review and narrative synthesis. PLoS One 2019;14(3):e0214678 [FREE Full text] [doi: 10.1371/journal.pone.0214678] [Medline: 30921432]
48. Monk-Cunliffe J. How can we overcome health inequalities in psychiatry? BJPsych Bull 2023;47(4):228-230. [doi: 10.1192/bjb.2023.49] [Medline: 37401250]
49. Anjum G, Aziz M. Advancing equity in cross-cultural psychology: embracing diverse epistemologies and fostering collaborative practices. Front Psychol 2024;15:1368663 [FREE Full text] [doi: 10.3389/fpsyg.2024.1368663] [Medline: 38638521]
50. Widdowson HG. Discourse analysis: a critical view. Lang Literature 1995;4(3):157-172. [doi: 10.1177/096394709500400301]
51. Breeze R. Critical discourse analysis and its critics. Pragmatics 2011;21(4):493-525 [FREE Full text] [doi: 10.1075/prag.21.4.01bre]

Abbreviations

CDA: critical discourse analysis
RC: recovery college
TXT file: text file

XSL•FO
RenderX

Edited by A Mavragani; submitted 08.07.24; peer-reviewed by C Entwistle; comments to author 20.09.24; revised version received 04.10.24; accepted 11.10.24; published 21.02.25.

Please cite as:
Kotera Y, Daryanani R, Skipper O, Simpson J, Takhi S, McPhilbin M, Ingall BR, Namasaba M, Jepps J, Kellermann V, Bhandari D, Ojio Y, Ronaldson A, Guerrero E, Jebara T, Henderson C, Slade M, Vilar-Lluch S
Applying Critical Discourse Analysis to Cross-Cultural Mental Health Recovery Research
JMIR Form Res 2025;9:e64087
URL: https://formative.jmir.org/2025/1/e64087
doi:10.2196/64087
PMID:

©Yasuhiro Kotera, Riddhi Daryanani, Oliver Skipper, Jonathan Simpson, Simran Takhi, Merly McPhilbin, Benjamin-Rose Ingall, Mariam Namasaba, Jessica Jepps, Vanessa Kellermann, Divya Bhandari, Yasutaka Ojio, Amy Ronaldson, Estefania Guerrero, Tesnime Jebara, Claire Henderson, Mike Slade, Sara Vilar-Lluch. Originally published in JMIR Formative Research (https://formative.jmir.org), 21.02.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Are Dating App Algorithms Making Men Lonely and Does This Present a Public Health Concern?

Eric Balki[1], BSc, MSc, MA, PhD

Faculty of Health and Medicine, Department of Health, Lancaster University, Lancaster, United Kingdom

Corresponding Author:
Eric Balki, BSc, MSc, MA, PhD
Faculty of Health and Medicine
Department of Health
Lancaster University
Innovation One, Sir John Fisher Dr, Bailrigg
Lancaster, LA1 4AT
United Kingdom
Phone: 44 1524 593763
Email: e.balkhi@lancaster.ac.uk

Abstract

During the pandemic, dating apps emerged as essential platforms connecting users amid social isolation, experiencing rapid growth in engagement and profile creation. This paper examines the evolution of these apps, highlighting their shift from facilitating offline encounters to promoting match accumulation for revenue. In particular, the study investigates gender disparities, addictive behaviors, and algorithmic match throttling that disproportionately impact men's psychological well-being. Drawing on evidence linking dating app use to increased depression and anxiety, the analysis calls for regulatory intervention to eliminate pay-for-advantage models and ensure fair, healthy user experiences, thereby mitigating adverse public health outcomes.

(JMIR Form Res 2025;9:e70594) doi:10.2196/70594

KEYWORDS

dating apps; mental health; men; algorithm; anxiety; depression; loneliness

This paper deals with issues revolving around heterosexual dating, which currently forms the largest marketplace in dating apps [1]. During the pandemic with frequent lockdowns, dating apps provided an opportunity for people to connect [2]. In a short span of time, there was a surge in dating profiles on these apps. Just like Peloton, the home exercise machine company that found a way of appealing to users to use stationary bikes and treadmills, dating apps attracted a vast range of young adults looking for a connection or relationship, at a time when isolation and loneliness was high [3].

At the start of the pandemic, the way the apps were working was quite simple. When two users accepted each other's profile, they had a match. Browsing through and evaluating other user's profiles was the central activity, eventually leading to an offline encounter where possible [4]. Dating apps, however, slowly started to not facilitate offline encounters but rather feed into the user's desire of accumulating matches [5].

Even before the pandemic, more than half of the users on dating apps reported not going on in-person dates, and this percentage was especially higher for women [6], with them looking for validation, gratification, and a boost in self-worth instead [7]. It is very likely this has worsened even further, with a large majority of users not going on in-person, offline dates [8].

Even those who are now in relationships find it hard to quit these apps—keeping them for gratification purposes [9], especially in women who accumulate matches to satisfy the need for belongingness or as a replacement for relational or sexual intimacy, getting value from social acceptance and approval [10].

In the last two years, the problem has exacerbated significantly. Women report being inundated with so many matches that it is hard from them to make a decision on choosing someone to engage with properly [11]. At the same time, men are on the opposite end of the spectrum, where they get very few responses and must purchase expensive, paid features and subscriptions, which allow them to get a limited amount of priority over nonpaying members [12].

Men continue to form the largest group of paid subscribers to dating apps [12]. Dating apps are focused on maximizing profit activities, especially as many of them are now owned by publicly listed companies [13]. Losing a subscriber means losing revenue, so it is conjectured that dating apps are motivated to keep their

XSL•FO
RenderX

paying customers as long as they possibly can, as a successful offline encounter mediated by the dating app would lead to the loss of that subscriber [12]. Unlike other social media platforms, there is no upside in successful encounters from a revenue-generating perspective. Men are also more likely to quit a dating app if they have found someone in person [12].

Dating apps are like casinos in a way, in that they have to strategize where the reward needs to be—just enough to keep users coming back for more, but the reward cannot be so high that users walk away and not return, which can increase addictive behavior [14]. They have gamified the process of meeting a partner, invoking "gambler" tendencies, thus leading to behaviors that are addictive or even compulsive [15]. Now, sophisticated artificial intelligence (AI) tools are being deployed by dating apps to maximize subscriber revenue, which is potentially causing excessive harm to the psychological profiles of a large number of men and hindering the development of healthy relationships, which could become a large public health concern [16]. Thus, there is both algorithmic and natural throttling of the mutual acceptance matching process on these apps, which is driving an increasing frustration in men as well as impacting their female users [16].

This is the fallacy in the approach by dating app companies, as the pool of paid subscribers will eventually, and inevitably, fall as the frustration leads users to find alternative ways of finding dates.

It has become extremely important for researchers to study the psychological traits and well-being of users of dating apps [17], especially with the changes and advances in algorithms and use of AI. As things stand, there are very few studies that have looked at dating app use; the algorithms being deployed by such companies; match throttling; and the associated psychological outcomes, especially in men.

There is a need for studies that move away from simply associating correlations between dating app use and psychological outcomes, including mere use associated with decreased well-being. Authors need to distinguish which specific activities, features, or experiences may be leading to undesired effects. While it may not always be possible to understand how the backend of the apps works or what algorithms may be at play, real-time testing with users could reveal basic issues revolving around authentic matching [18]. The design of such studies will have to be carefully considered to ensure the cross-sectional design does not end up ruling out self-selection, that is, attracting lonely individuals and those in fear of being single [19].

The majority of the population are reliant on dating apps to meet prospective partner, and such apps are having an impact on their mental health; these factors may very well make this into a public health concern. The Centers for Disease Control and Prevention's [20] Household Pulse Survey estimates that depression and anxiety are now affecting 34.2% of the population, making them the two most common mental health disorders in the United States, and to some extent, the increased numbers have been associated with dating app use [21].

Some studies have linked dating apps directly with higher depression and anxiety, especially the frequency and length of use [17]. Symptoms from these disorders can further impact social skills, making the formation of healthy relationships even more difficult, thus leading to a downward spiral in the populace [18].

This is an urgent problem that needs regulatory action from the Western governments, where dating app algorithms should provide fair conditions to all users. Match throttling and pay-for-advantage models should be disallowed as they have a disproportionately negative impact on the health of young men.

If such action is not taken, then over time we will have a population of young, productive men who grow tired of using dating apps to find a partner and come out feeling lonely, depressed, and anxious, thus impacting in other areas of their life and their ability to be productive. Therefore, rather than helping people find partners, dating apps may be failing the very users they rely upon for their income.

Acknowledgments

This work was not funded.

Authors' Contributions

The author listed has made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Conflicts of Interest

None declared.

References

1. Bandinelli C, Gandini A. Dating apps: the uncertainty of marketised love. Cultural Sociology 2022 Jan 10;16(3):423-441. [doi: 10.1177/17499755211051559]
2. Gibson AF. Exploring the impact of COVID-19 on mobile dating: critical avenues for research. Soc Personal Psychol Compass 2021 Nov;15(11):e12643 [FREE Full text] [doi: 10.1111/spc3.12643] [Medline: 34899975]
3. Weldon M. The Time We Have: Essays on Pandemic Living. Evanston, IL: Northwestern University Press; 2024.

XSL•FO
RenderX

4. Cummings JJ, Mays K. Trait motivational reactivity as a predictor of online dating app behavior. Computers in Human Behavior 2021 Aug;121:106775. [doi: 10.1016/j.chb.2021.106775]
5. Bandinelli C, Bandinelli A. What does the app want? a psychoanalytic interpretation of dating apps' libidinal economy. Psychoanal Cult Soc 2021 Apr 19;26(2):181-198. [doi: 10.1057/s41282-021-00217-5]
6. Timmermans E, Courtois C. From swiping to casual sex and/or committed relationships: exploring the experiences of Tinder users. The Information Society 2018 Mar 08;34(2):59-70. [doi: 10.1080/01972243.2017.1414093]
7. Wimark T, Hedlund D. From individuals to emotional drones: technology-driven change in the collective conditioning of intimacy. New Media & Society 2021 Nov 18;26(1):236-252. [doi: 10.1177/14614448211057108]
8. Toma CL. Online dating and psychological wellbeing: a social compensation perspective. Curr Opin Psychol 2022 Aug;46:101331. [doi: 10.1016/j.copsyc.2022.101331] [Medline: 35349878]
9. Sharabi LL, Timmermans E. Why settle when there are plenty of fish in the sea? Rusbult's investment model applied to online dating. New Media & Society 2020 Jul 10;23(10):2926-2946. [doi: 10.1177/1461444820937660]
10. Cross SE, Madson L. Models of the self: self-construals and gender. Psychol Bull 1997 Jul;122(1):5-37. [doi: 10.1037/0033-2909.122.1.5] [Medline: 9204777]
11. Thomas MF, Binder A, Stevic A, Matthes J. 99+ Matches but a spark ain't one: adverse psychological effects of excessive swiping on dating apps. Telematics and Informatics 2023 Mar;78:101949. [doi: 10.1016/j.tele.2023.101949]
12. O'Gorman K, Wilson MJ, Seidler Z. Swiping into the void: tears, fears, and motivations of young men on dating apps. In: Seilder Z, editor. Masculinities and Mental Health in Young Men. Cham, Switzerland: Palgrave Macmillan, Springer Nature; 2024:229-262.
13. Nickenig A. The market of love: dating economies from early modern matchmaking to Tinder. In: Schomacher E, Söffner J, editors. The Culture of Money: Implications for Contemporary Economics. London, United Kingdom: Routledge; 2024.
14. Mackinnon L. Love, games and gamification: gambling and gaming as techniques of modern romantic love. Theory, Culture & Society 2022 Mar 17;39(6):121-137. [doi: 10.1177/02632764221078258]
15. Isisag A. Mobile dating apps and the intensive marketization of dating: gamification as a marketizing apparatus. In: Bagchi R, Block L, Lee L, editors. Advances in Consumer Research, Volume 47. Duluth, MN: Association for Consumer Research; 2019:135-141.
16. Fletcher R, Tzani C, Ioannou M. The dark side of artificial intelligence – risks arising in dating applications. Assessment and Development Matters 2024 Mar 01;16(1):17-23. [doi: 10.53841/bpsadm.2024.16.1.17]
17. Holtzhausen N, Fitzgerald K, Thakur I, Ashley J, Rolfe M, Pit SW. Swipe-based dating applications use and its association with mental health outcomes: a cross-sectional study. BMC Psychol 2020 Mar 04;8(1):22 [FREE Full text] [doi: 10.1186/s40359-020-0373-1] [Medline: 32127048]
18. Braithwaite S, Holt-Lunstad J. Romantic relationships and mental health. Curr Opin Psychol 2017 Feb;13:120-125. [doi: 10.1016/j.copsyc.2016.04.001] [Medline: 28813281]
19. Lauckner C, Truszczynski N, Lambert D, Kottamasu V, Meherally S, Schipani-McLaughlin AM, et al. "Catfishing," cyberbullying, and coercion: an exploration of the risks associated with dating app use among rural sexual minority males. Journal of Gay & Lesbian Mental Health 2019 Mar 29;23(3):289-306. [doi: 10.1080/19359705.2019.1587729]
20. Chin K, Edelstein R, Vernon PA. Attached to dating apps: attachment orientations and preferences for dating apps. Mobile Media & Communication 2018 Jun 18;7(1):41-59. [doi: 10.1177/2050157918770696]
21. National Center for Health Statistics. Anxiety and depression. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/nchs/covid19/pulse/mental-health.htm [accessed 2024-12-25]

Abbreviations

AI: artificial intelligence

Edited by A Mavragani; submitted 26.12.24; peer-reviewed by V Sideropoulos; comments to author 06.02.25; revised version received 25.02.25; accepted 11.03.25; published 07.04.25.

Please cite as:
Balki E
Are Dating App Algorithms Making Men Lonely and Does This Present a Public Health Concern?
JMIR Form Res 2025;9:e70594
URL: https://formative.jmir.org/2025/1/e70594
doi:10.2196/70594
PMID:

XSL•FO
RenderX

©Eric Balki. Originally published in JMIR Formative Research (https://formative.jmir.org), 07.04.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Addressing the Ethical, Legal, and Social Issues of Healthtech in Education: Insights From Japan

Motofumi Sumiya[1,2,3], PhD; Tomoko Nishimura[1,2,4], PhD; Kyoko Aizaki[1,5], PhD; Ikue Hirata[1,2,6], PhD; Nobuaki Tsukui[1,2,4], MA; Yuko Osuka[1,4], PhD; Manabu Wakuta[1,4], PhD; Atsushi Senju[1,2], PhD

[1]Research Center for Child Mental Development, Hamamatsu University School of Medicine, Shizuoka, Japan

[2]United Graduate School of Child Development, Osaka University, Kanazawa University, Hamamatsu University School of Medicine, Chiba University and University of Fukui, Osaka, Japan

[3]Department of Clinical Psychology, Hanazono University, Kyoto, Japan

[4]Institute of Child Developmental Science Research, Shizuoka, Japan

[5]Faculty of Global Studies, Reitaku University, Chiba, Japan

[6]Faculty of Education and Integrated Arts and Sciences, Waseda University, Tokyo, Japan

Corresponding Author:
Atsushi Senju, PhD
Research Center for Child Mental Development
Hamamatsu University School of Medicine
1-20-1 Handayama, Chuo-ku, Hamamatsu city
Shizuoka, 431-3192
Japan
Phone: 81 53 435 2088
Fax: 81 53 435 2291
Email: senju@hama-med.ac.jp

Abstract

The increasing application of health technology (healthtech) in educational settings, particularly for monitoring students' mental health, has garnered significant attention. These technologies, which range from wearable devices to digital mental health screenings, offer new opportunities for enhancing student well-being and strengthening support systems. Numerous studies have explored the ethical, legal, and social issues (ELSIs) of healthtech in the field of psychiatry, highlighting its potential benefits while also acknowledging the inherent complexities and risks that demand careful consideration. However, the ELSIs related to the use of healthtech in educational settings remain largely overlooked and insufficiently addressed. This study provides an overview of items that should be considered by researchers, teachers, and education boards or committees to promote healthtech in the educational context. By adapting existing ELSI frameworks from educational technology and digital health, this study systematically reviews ethical concerns surrounding healthtech in schools. Expert consultations were conducted through a project consisting of members with expertise related to healthtech, including developers, a teacher, a school counselor, and university researchers, leading to the identification of 52 ELSI concerns categorized into 8 domains: consent, rights and privacy, algorithms, information management, evaluation, use, role of public institutions, and relationships with private companies. Using Japan as a case study, we examine regulatory and cultural factors affecting healthtech adoption in schools. The findings reveal critical challenges, such as ensuring informed consent for minors, protecting student privacy, preventing biased algorithmic decision-making, and maintaining transparency in data management. In addition, institutional factors, including the role of public education policies and private-sector involvement, shape the ethical landscape of healthtech implementation. This study highlights the need for multistakeholder collaboration to establish guidelines that balance innovation with ethical responsibility. The study underscores the need for a multifaceted approach to mitigate risks such as data misuse, inequitable access, and algorithmic bias, ensuring the ethical and effective use of healthtech in education. The fundamental ELSI framework for healthtech, including privacy, consent, and algorithms, can be applied to educational systems worldwide, while aspects related to public education policies should be considered in accordance with the specific context of each country and culture. Incorporating healthtech into the educational system helps address the barriers associated with traditional approaches, including limited resources, cost constraints, and logistical challenges. Researchers from universities and healthtech companies, along with educators and other stakeholders, should ensure that healthtech projects consider diverse ELSI concerns at every stage before and during implementation.

(JMIR Form Res 2025;9:e72781) doi:10.2196/72781

XSL•FO
RenderX

KEYWORDS

health technology; ethical, legal, and social issues; education; school; well-being

Introduction

Overview

The integration of health technology (healthtech) into school education has introduced new opportunities for monitoring students' mental health and enhancing their well-being. Innovations such as wearable devices that track physical activity (see Au et al [1] and Zhang et al [2] for a systematic review) and daily self-reported mental health screening, provide the potential to revolutionize the educational landscape [3,4]. However, these advancements also raise ethical, legal, and social issues (ELSIs) that require careful consideration. For example, it is imperative to address concerns regarding data privacy, equitable access to healthtech services, and the potential misuse and mishandling of sensitive health information when introducing these technologies in public educational settings. Addressing these concerns ensures that healthtech functions as a supportive tool for students, educators, and parents, rather than a source of harm or inequity. Although many studies have considered ELSIs for healthtech in the field of psychiatry, no study has specifically examined them in the educational context. Accordingly, this study provides an overview of items that should be considered by researchers, teachers, and education boards or committees to promote healthtech in the educational context.

Importance of Healthtech in Education

There is an increasing emphasis on the role of mental health screening in school education in bridging the gap between the significant mental health needs of students and the limited access to appropriate care. According to Soneson et al [5], identifying mental health challenges in school-aged children early can significantly enhance access to timely interventions and address the widespread issue of underidentification, where only a small percentage of affected children receive adequate support. Similarly, Weist et al [6] emphasized that systematic screening in schools not only enhances outreach to at-risk youth but also integrates mental health initiatives into the educational environment, fostering both mental well-being and academic progress. Effective implementation ensures that such programs align with public health approaches, facilitating early detection and prevention while reducing barriers to learning and combating the stigma surrounding mental health issues.

With the rapid development of digital technology in recent years and its increasing use during the COVID-19 pandemic, mental health screening in schools has become progressively digitalized. The incorporation of healthtech into school-based mental health screening presents a promising solution for addressing the barriers associated with traditional approaches, including limited resources, cost constraints, and logistical challenges. For example, Nishimura et al [3] highlighted the use of tablet devices and health-monitoring apps to examine and analyze students' daily health and emotional well-being through self-report measures in primary and junior high schools in Japan. They emphasized that the digitalization of data enabled the efficient transition from traditional paper-based data to an online system, facilitating the analysis of both long-term health trends and short-term variations. They also observed that using simple algorithms, such as rolling *z* scores calculated from the 7-day rolling average and SD, to alert students about their health status enabled the rapid identification of data fluctuations. This, in turn, facilitated the early prediction of mental health issues and the timely provision of interventions. These efforts were supported by Japan's Global and Innovation Gateway for All School Program, which equips students with tablet devices. The introduction of digital technology into school education also has the potential to reduce the burden on teachers, enhance the efficiency and sustainability of observations, and ensure equitable and individually optimized learning environments.

With regard to leveraging healthtech for school-based mental health screening, Porter et al [4] examined the effectiveness of the Digital Health Contact (DHC), an online self-report tool used in schools in the East Midlands of England. Using a structured questionnaire, this tool enables students to provide data on their health and well-being, including indicators of mental health concerns. They introduced a "red flag" system that automatically flags certain answers or keywords when they are included in free-text responses to highlight potential issues using an automated algorithm, directing at-risk students for further evaluation by public health nurses. Porter et al [4] also determined that the DHC enhanced the ability to identify unmet mental health needs among adolescents, particularly in schools with a higher number of students from diverse linguistic backgrounds, where traditional methods often fell short. In addition, while the study noted an increase in the overall number of referrals, it emphasized that the DHC complemented traditional referral pathways, enabling a more structured and proactive approach to providing mental health support. Therefore, although these two studies [3,4] have different objectives and detection algorithms, both further highlight the transformative potential of healthtech in schools, with Nishimura et al [3] focusing on identifying quantitative variation in data and Porter et al [4] emphasizing mechanisms for immediate identification of specific student problems and providing direct support.

However, several issues must be considered when using innovative and novel tools in the context of institutional barriers and traditional rules. Woodrow et al [7,8] highlighted both the promise and challenges of implementing healthtech, such as the DHC in school settings, by conducting qualitative studies with semistructured interviews. From the students' perspective, the DHC fosters honest communication regarding health and well-being by ensuring privacy and reducing the stigma associated with seeking support [8]. Similarly, stakeholders such as commissioners, service providers, and health care staff emphasized the tool's effectiveness in identifying unmet health needs and its ability to operate efficiently in resource-limited environments [7]. Importantly, both studies highlight critical ELSI concerns that must be addressed to ensure their responsible use. Concerns regarding data privacy, informed consent,

XSL•FO
RenderX

equitable access, and cultural adaptability have been identified as key challenges [7,8]. These findings reinforce the need to introduce robust ELSI frameworks to navigate healthtech in educational settings, ensuring that they promote inclusivity, trust, and ethical practices while maximizing potential benefits.

Discussion of ELSIs in Healthtech

Overview

To the best of our knowledge, ELSIs surrounding the use of healthtech in educational contexts have not been systematically evaluated in the literature, although they have been widely explored in the context of psychiatry or mental health informatics [9-14]. These studies have not only highlighted the potential benefits of healthtech but also flagged the inherent complexities and risks that require careful navigation. For example, Shen et al [12] proposed an ethics checklist for deep phenotyping research in psychiatry, combining extensive data collection (such as 24/7 data on location, movement, and emails, along with brain scans, genetics, genomics, neuropsychological assessments, and clinical interviews) with advanced techniques for analyzing these data. The checklist addresses key domains such as informed consent, equity, and privacy, urging researchers to explicitly document their decisions and consult diverse stakeholders. This procedural rigor is crucial in ensuring that healthtech is designed to uphold participants' rights while addressing systemic inequities stemming from unequal access to digital health solutions. Therefore, by considering a comprehensive ELSI framework based on these perspectives and prioritizing transparency, inclusivity, and accountability at every stage of technology development and deployment, healthtech can advance mental health care and serve as a model for ethically responsible innovations.

Seiferth et al [15] provide a comprehensive set of international guidelines for the development, implementation, and evaluation of healthtech. They emphasize a user-centered and participatory design approach, ensuring that digital solutions address the specific needs of their target populations while enhancing accessibility and usability. The authors identified key considerations including data security, privacy, and transparency, highlighting the importance of building trust among users through clear communication and adherence to ethical standards. They argue that safeguarding user data is critical for maintaining trust, particularly given the sensitive nature of mental health information. The guidelines recommend using advanced security measures such as end-to-end encryption, pseudonymization, and regular security audits to protect against unauthorized access and data breaches. Privacy considerations extend to the ethical collection, storage, and use of personal data. They advocate transparent data practices, including clear communication with users regarding how their data will be managed, who will have access to them, and for what purposes. This transparency must be embedded in the informed consent process to enable users to make informed decisions about their participation in healthtech interventions. In addition, the authors highlight the need for real-time user notifications and accessible options for participants to withdraw their consent or modify their data-sharing preferences. These practices ensure compliance with data protection regulations, such as the General Data Protection Regulation, while fostering a culture of accountability and respect for user autonomy. By addressing these considerations, Seiferth et al [15] provide guidelines aimed at creating a trustworthy environment that encourages the adoption and sustainable use of healthtech for mental health.

First Step Toward Discussing ELSIs With Regard to the Use of Healthtech in the Educational Context

Although healthtech has penetrated the field of education globally, ELSIs surrounding the use of these technologies have received limited attention. Researchers from universities and healthtech companies, educators, and stakeholders ensure that healthtech projects consider diverse ELSI concerns at every stage before and during implementation. When using healthtech in the educational context, diverse factors must be considered compared to its use in other settings. For example, both students and their parents must be considered when obtaining consent for healthtech use. In addition, because school systems, regulations, and laws related to schools vary by country, it is necessary to consider realistic use within the existing framework of the country where the school is located. For example, in the Japanese education system, with a few exceptions, teaching curricula and assessments are required to adhere to the Courses of Study established by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) [16]. Accordingly, in Japan, it can be challenging for local governments or individual schools to flexibly modify educational content or instructional hours or introduce new technologies and systems. These issues are specific to educational settings and need to be considered, in addition to ELSI concerns related to healthtech for mental health or psychiatry. However, as noted in previous paragraphs, healthtech is increasingly being integrated into the field of education worldwide, yet ELSI considerations regarding the use of these technologies have been largely overlooked.

This study examines the issues to be considered when implementing healthtech in Japanese educational settings. Given the need to thoroughly consider the characteristics of Japanese educational settings, we reference a project that examined the ELSIs of implementing educational technology in Japan [17]. Kano et al [17] discuss issues related to the use of digital technology in education, particularly within the context of Japanese-style public education. Their study, conducted through a project involving experts in ELSIs—such as ethicists, legal and constitutional scholars, social psychologists, and digital technology developers—explores 101 ELSI concerns. Our project is supported by the Japan Science and Technology Agency–Mirai Program under the "Society Optimized for Diversity" mission area. This study was conducted by a team comprising healthtech developers, a teacher, a school counselor, and university researchers, with the goal of creating a comprehensive ELSI framework for healthtech in school education. The study design was approved by the ethics committees of Hamamatsu University School of Medicine. To achieve this, we first analyzed the 101 ELSI concerns identified in educational technology research [17]. Each concern was systematically reviewed and assessed for its relevance in healthtech. We conducted a focus group discussion following a structured procedure: individuals with relatively homogeneous

XSL•FO
RenderX

attributes, knowledge, and characteristics relevant to the research topic were selected. A moderator facilitated the discussion based on a prepared guideline, encouraging interactions among participants and eliciting their emotions, attitudes, and thoughts on the designated themes. On the basis of participants' collective input, items from EdTech101 that were not relevant to healthtech were eliminated, while those with strong relevance were grouped and refined. Keywords were assigned to each item, forming the basis for organizing the concerns into 8 thematic categories. This process was grounded in thematic analysis [18,19]. Therefore, we identified 52 key issues, categorized into 8 topics related to healthtech (Table 1; a detailed version is provided in Multimedia Appendix 1, along with the Japanese-language original version). The topics range across (1) consent, (2) rights and privacy, (3) algorithms, (4) information acquisition and management, (5) evaluation, (6) use, (7) public institutions, and (8) private companies (Figure 1), with each containing 4 to 11 issues. In the following section, we discuss the characteristics of each topic and the ELSIs for each.

Figure 1. Phased considerations and institutional roles in health technology implementation.

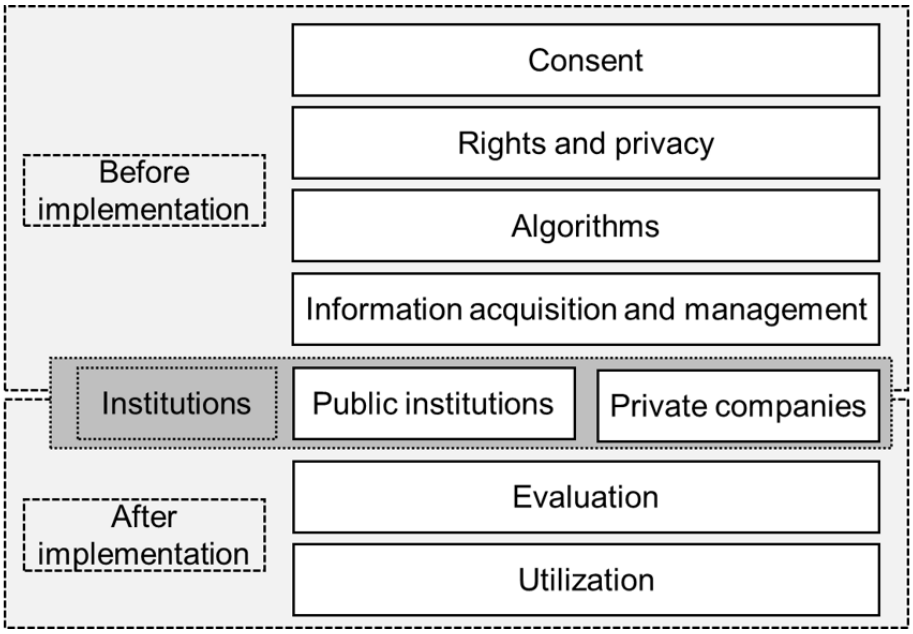

XSL•FO
RenderX

Table 1. Ethical, legal, and social issues on health technology (healthtech) in the educational context.

Category and issues (when data collection and mental health assessments are conducted as a school activity)	Key laws[a]
Consent	
Is guardian informed consent[b] required? If so, is the informed assent[c] of students also obtained?	1
Does the information used in informed consent include risk assessment[d]?	1
Even in the case where a guardian has not consented to data acquisition from a healthtech service, are students notified of the details of guardian consent/non-consent?	1
Should an "opt-out[e]" format be applied for consent?	1
Do students view the facial features and attitudes of teachers and urge their guardians to give their consent, or do they express support?	1
Rights and privacy	
Is there recognition of a right not to be evaluated by healthtech?	1
To whom, to what extent, and how are the evaluation and reasons disclosed.	1
Is the so-called right to be forgotten[f] (e.g., the deletion of information collected by healthtech) recognized?	1 and 2
Do inferences regarding the internal mind (attitudes, emotions, etc.) of students constitute a failure to protect the freedom of thought and conscience of the concerned persons?	1
Is this not considered an infringement of the teacher's right to "educational freedom[g]"?	3
Does using expression or voice recognition to record the details of student information entail making sensitive information available to others, with a risk of violating privacy rights?	2
Algorithms	
Is third-party oversight (auditing, inspection) guaranteed for algorithms[h]?	1 and 4
Do mechanisms exist for corrections where, due to inaccurate profiling, mistakes have occurred in the evaluation process?	1
Can the system consider the need to ensure that students who deviate from the norm are not assessed unfavorably?	1
Is mental health expertise or third-party certification system required to develop assessment items and algorithms of healthtech? If so, what are the criteria for determining it?	5
Information acquisition and management	
Should sensitive personal information[i] about students with developmental disabilities and mental health history be acquired and used to screen and sort schools?	1
Is the information management system regarding information accumulation and management as well as its linking with other information sufficient?	2
Who, where, and how long should the information collected by healthtech be stored?	__[j]
How should requests for the cessation of use[k], etc., of collected data be treated?	6
Will not schools and teachers emulate the healthtech methods to voluntarily collect excessive personal data?	6
Evaluation	
When an evaluation includes parameters other than just healthtech assessment, how can it be guaranteed that factors indubitably related to prejudice/discrimination do not influence the evaluation?	1
Does making health/mental condition "visible" promote hierarchization of students by grades?	3
Is there no danger of so-called overmeasurement from quantifying things that are difficult to quantify?	3
Is there a risk that visualizing mental health assessments provides a biased view of students? If visualization (with charts, etc.) focuses on mental health assessments rather than school life, is it not possible that more importance will be given to students' abilities, leading to an over-bias on the importance of a person's background?	2
If such a face recognition system encourages a majority of students to adopt expression methods and emotional expressions that tend to result in good score, does this not hinder the diversity of self-expression?	7
How should one consider and treat cases where sensitive individual information is used to infer, via profiling, that the concerned individual requires special care or consideration?	6
What are the grounds for making all problems identical for everyone? Should items and assessment methods not reflect regional or age differences?	—

XSL•FO
RenderX

Category and issues (when data collection and mental health assessments are conducted as a school activity)	Key laws[a]
The use of technologies is expected to increase equality in evaluation of interests, motivation, and attitudes compared with the subjective evaluation of human teachers. However, will objective quantification (scoring) of such things as developmental disorders and/or mental illness result in such conditions being treated in a fixed, inflexible way?	—
With the accumulation of recordings of a student's past problematic behavior, even if that student shows growth and development (maturation) over time, will that student not be "stuck with" that reputation, etc., making appropriate evaluations in the future unlikely?	—
Can mental health history and abilities be linked to personality evaluation?	—
When the "optimal" value for "individual optimization[l]" recommended by a system is not proven with evidence, is there not a risk that the "individual optimization" as designated by the system in its first introduction and use becomes a type of de facto standard[m], locking in those values as "standards?"	—
Use	
Are there assurances that the results of evaluations performed to promote the development of a student are not used in selection screening?	1
It is acceptable if healthtech is used by a teacher as a support tool, but if its influence surpasses that of human judgment, would that lead to "inappropriate control" by HealthTech?	3
Is healthtech truly used for educational purposes, and not just nominally so?	8
Will classifications (for determining members of specific school classes, etc.) based on differences in assessment by HealthTech ultimately not lead to discrimination?	9
If students' posture and attitude are judged based on recognition accuracy (i.e., some faces tend to be recognized by face recognition as actively claiming and others as expressionless), does this constitute discrimination?	9
Should educational results from healthtech be recognized as the educational performance of teachers?	10
Is it appropriate to use data obtained through facial recognition[n], such as a teacher's level of concentration and facial expressions, as criteria for teacher evaluation?	10
How should the usage of anonymously processed information[o] be treated?	6
Is it necessary to add healthtech functions that prioritize student guidance elements, such as those that help improve postures or those that boost students' self-worth by encouraging them to modify their behavior?	—
Will the teacher's duties not be taken over in those areas where healthtech can serve as a substitute or alternate performer?	—
Public institutions	
If there is evidence of educational benefits, is the national government obliged to introduce an educational environment using healthtech?	11
In the case of digital terminals (tablets, LAN extension inside the student's home etc.), will the cost of purchasing related educational materials be privately or publicly funded?	11
If, for example, a system is established where subsidies and other payment standards vary depending on whether or not a certain school has introduced healthtech promoted by the Ministry of Education, Culture, Sports, Science and Technology, could this lead to the penetration of the government's message to children via public education?	3
Can assistance monies be paid to an educational institution operated by an healthtech enterprise (i.e., an enterprise not designated under article 1 of the School Education Act[p,q])?	8
Is it appropriate for the national government to set uniform evaluation standards?	7
Does the mass introduction of healthtech services entail expenditures of public monies for private businesses?	4
Does using healthtech to teach mental health count as health education in school curricula?	4
Private companies	
Is it suitable for healthtech provided by private enterprises to be incorporated into educational contents and/or methods?	3
If healthtech is introduced without sufficient knowledge of ICT[r], schools may be unable to fully understand its content and ongoing changes, with the result that they merely follow what the tech provider instructs.	8
Will the value systems and expressions of private companies be linked with the evaluation standards for children/students that are used in public education?	2 and 7

XSL•FO
RenderX

Category and issues (when data collection and mental health assessments are conducted as a school activity)	Key laws[a]
If a stockholding company, not an incorporated educational institution, possesses multifaceted data of students, is there a risk that the data may be used beyond education-related purposes?	12

[a]Key laws (relevant laws and regulations); 1: Constitution, Article 26, Paragraph 1: “All people shall have the right to receive an equal education correspondent to their ability,” 2: Constitution, Article 13: “Personal rights” and “Right to privacy,” 3: Basic Act on Education, Article 16: “Education must not be subject to improper controls,” 4: Characteristics of Japanese-style public education not necessarily grounded in law, 5: Article 3 of the Educational Personnel Certification Act: 'Educational personnel must be those who have been granted a certificate in accordance with the provisions of this Act, 6: Act on the Protection of Personal Information, 7: Basic Act on Education, Article 1: “Education must be provided with the aim of fully developing the individual character,” 8: Constitution, Article 89: “Expenditures of public money, and limits on its usage (appropriation),” 9: Constitution, Article 14: “Equality under the law” and “Prohibition of discrimination,” 10: Every Article and item (paragraph) of the Local Public Service Act stipulated by this Act,” 11: Constitution, Article 26, Paragraph 1: “Right to receive education “Paragraph 2: “Compulsory education shall be free,” 12: School Education Act, Article 2: “Schools shall be established only by the national government, local governments, and school corporations specified by Article 3 of the Private Schools Act.”

[b]Informed consent: providing consent after having been fully informed and satisfied about the purpose, risks, benefits, and alternatives of the data collection or intervention.

[c]Informed assent: In addition to parental consent (informed consent) for minors and others, the child’s consent should be obtained after he or she understands and accepts the purpose, risks, benefits, and alternatives to data collection or intervention.

[d]Risk assessment: Assessing potential risks and harms that may arise from implementing HealthTech in advance, and considering measures to mitigate those risks.

[e]Opt-out: A process in which individuals, in a situation where they are automatically included in a specific service or data processing, request to be excluded based on personal decision.

[f]Right to be forgotten: The right of an individual to request the deletion of their personal information stored by an organization or service provider when they ask for it to be erased.

[g]Freedom of educational policy: Schools and teachers have the right to freely decide on their educational policy. A famous court case is the Asahikawa School Achievement Test. The “Asahikawa Gakute Judgment” in the 1950s is a leading case involving the freedom of education of school teachers (here, a teacher who opposed the National Achievement Test [or “Gakute”] was charged with obstructing the execution of public affairs).

[h]Algorithm: A set of specific methods, processes, or rules established to address a particular issue or conduct analysis.

[i]Sensitive personal information: Personal information that requires particularly careful handling, such as health status, disability, nationality, criminal history, or beliefs.

[j]Not applicable.

[k]Request for suspension of use: The right of an individual to request the suspension or deletion of personal information when its use is illegal or inappropriate.

[l]Individual optimization: Providing optimal measures and resources based on an individual's needs and characteristics.

[m]De facto standards: Technologies or norms recognized as the default through widespread adoption and proven track record, without undergoing formal standardization.

[n]Face recognition: Analyzing facial features and using it to identify the individual.

[o]Anonymously Processed Information: Data generated by removing or replacing necessary identifying elements from personal information to ensure that specific individuals cannot be identified.

[p]Schools falling under School Education Act 1: Educational institutions (called “schools”) established under article 1 of the “School Education Act.” Specifically, this refers to kindergarten, elementary, junior high, compulsory education, high, secondary education, and special needs education schools, universities, and college of technology (KOSEN).

[q]Schools that do not fall under School Education Act 1: Educational institutions not positioned under article 1 of the “School Education Act.” Specifically, they are international, driving, and preparatory schools.

[r]ICT (information and communication technology): This term collectively refers to technologies and tools related to information processing and communication, such as computers, the Internet, smartphones, and cloud services.

Consent

The consent obtained by the implementer of healthtech from the participants is an important topic that anyone involved in healthtech needs to address. Shen et al [12] emphasized the importance of obtaining consent after providing participants with sufficient information. Furthermore, when implementing healthtech in educational settings, it is essential to consider who should provide consent for participation and how it should be obtained. First, obtaining informed consent from parents respects their responsibility toward minors and strengthens their trust in the implementation of healthtech. Through this process, parents gain an opportunity to understand the mechanism, purpose, and risks surrounding healthtech, which can provide them with a sense of security. In addition, seeking informed consent from students allows their opinions and autonomy to be respected while fostering an understanding of the importance of decision-making. However, parents’ lack of understanding or low technical literacy may hinder a suitable consent process. For instance, if consent forms contain excessive technical jargon, parents may struggle to comprehend the content and potentially reject participation. Similarly, when obtaining students’ assent, their ability to fully understand the mechanisms and risks of technology depends on their age and developmental stage. Providing appropriate explanations and support is indispensable. If these efforts are insufficient, there is a risk that consent and assent in the school setting will become merely procedural,

XSL•FO
RenderX

failing to achieve their intended purposes (Table 1 provides relevant concerns).

Another critical consideration is managing situations where parents withhold consent for healthtech-based data collection. Should students still have access to healthtech, and if so, should they be informed about their parents' consent decisions? In public education, allowing students to access healthtech even when parental consent has been withheld may be regarded as a student-centered approach that prioritizes their best interests. For instance, in school health management programs providing psychological support or screenings, it may be essential to ensure that students do not miss necessary assistance due to a lack of parental consent. Furthermore, it is crucial to avoid scenarios where specific students are excluded from participation, as this could exacerbate feelings of isolation or stigma associated with healthtech use. Conversely, disregarding parental nonconsent and permitting students to use healthtech could undermine familial trust. Whether or not to inform students about their parents' decision requires careful deliberation. Informing students could trigger conflicts or resistance toward their parents, while withholding such information could invite criticism for the lack of transparency or ethical oversight in educational settings.

In light of these issues, opt-out systems could be considered an alternative approach to consent, especially in public schools, where initiatives often involve a large number of students. An opt-out approach assumes consent unless parents or students explicitly indicate their objections, which offers several advantages. For example, it can eliminate the need to collect explicit consent from students and their parents, significantly reducing the time and effort required. By including every student unless they actively opt out, the approach increases overall participation and enhances the comprehensiveness of the collected data [20,21]. This facilitates an accurate understanding of the school's overall situation, which can inform both individual support and broader policy improvements. Furthermore, because participation is considered the default, opt-out systems can help reduce the social stigma often associated with seeking psychological support, thereby improving access to mental health and welfare services.

In Japan, to address these challenges, public schools are required to operate in accordance with established educational regulations, including the School Education Act and the Act on the Protection of Personal Information. When implementing healthtech, schools must ensure proper management based on the guidelines set by the Board of Education and the MEXT. Under the School Health and Safety Act, schools are responsible for ensuring the well-being of students, which includes providing these health-related interventions. For screening related to physical health, it is clearly stated that an explanation to parents and students and parental consent is required if additional medical screening items are added or if privacy-related content is included. However, there is no clear provision for obtaining consent in the implementation of healthtech. This is an ongoing issue for consideration, including opt-out methods.

Rights and Privacy

Rights and privacy are critical for effective healthtech implementation. In the field of psychiatry, Wies et al [11] suggested that concerns regarding privacy breaches may erode user trust in healthtech, leading to reluctance in sharing data, ultimately diminishing treatment effectiveness. Similarly, in the context of healthtech use in educational settings, it is essential to prioritize the rights and privacy of students. Self-determination, privacy protection, and data management rights are critical elements to safeguard the healthy development and safety of students (Table 1).

First, when implementing healthtech, it is crucial to examine whether students should be granted the right to refuse to undergo mental health assessment. Granting students the right to opt out of evaluations is ethically significant in terms of respecting their self-determination and protecting their privacy. This is particularly important when screening for mental health issues or psychological stress, as students may not feel comfortable disclosing their feelings and conditions. Despite these concerns, forcing evaluations can result in psychological distress. However, broadly recognizing the right to refuse evaluation carries the risk of neglecting issues that require intervention, leaving problems unresolved. Therefore, when introducing healthtech to students, it is necessary to establish clear procedures and alternatives for cases where a student refuses evaluation, ensuring that their decision does not result in disadvantages or inequity.

Next, if healthtech includes the capability to detect sensitive information through nonverbal cues, such as facial expressions or voice, careful consideration is required to avoid potential privacy infringements. For instance, even if students do not intend to disclose their emotions or internal struggles, the system may inadvertently collect and analyze such information covertly. To mitigate this risk, the scope and purpose of data collection must be clearly defined, and comprehensive explanations must be provided to both students and their parents. Furthermore, it is imperative to provide opt-out options during the consent process, allowing participants to decline sharing sensitive information. In cases where interventions or actions are based on information identified by the system, transparency is crucial, as is the implementation of mechanisms that enable both students and parents to review and understand the collected data and subsequent actions.

Finally, it is essential to consider whether the "right to be forgotten"—allowing for the deletion of data collected by healthtech—should be recognized. Data collected in educational contexts, particularly concerning mental health or psychological traits, may include sensitive information. Considering the potential risk such information may pose in the future, a system enabling students or their parents to request data deletion is necessary. However, exercising the right to be forgotten can disrupt the consistency of school health programs or learning support initiatives. Consequently, establishing a framework that minimizes these disruptions when responding to deletion requests is essential. For instance, anonymizing the requested data for deletion and restricting their use in statistical analyses without personal identifiers can be a viable approach. Moreover,

XSL•FO
RenderX

it is crucial to clearly outline policies and procedures for data deletion to ensure students, and their parents are informed in advance about the process, its implications, and how deletion requests will be managed.

Algorithms

The guidelines for developing healthtech in the field of psychiatry emphasize the importance of a user-centered design that actively involves end users, health professionals, and other stakeholders in the preliminary stages of development [15]. This also applies to the field of education, where the design and implementation of algorithms must involve diverse stakeholders and be carefully considered from multiple perspectives. First, ensuring the auditability of algorithms by third parties is critical for maintaining transparency and trust in the system. Healthtech tools implemented in educational settings have a direct impact on how students' health and learning outcomes are evaluated. If algorithms lack transparency, there is a risk of being unfair or biased. For instance, if an algorithm operates as a "black box," stakeholders may not understand its decision-making criteria, making it challenging to address unjust evaluations. To mitigate this, mechanisms allowing third-party verification of the algorithm design and decision-making processes should be established. In addition, explaining the auditing process in an accessible manner to schools and families can help build trust among parents and students.

Next, it is important to establish a mechanism to evaluate the accuracy of the assessments derived from the algorithms, as well as a system to correct them if they are found to be inaccurate. There is always a possibility of errors owing to the quality of the input data or system bugs. Therefore, there should be processes that allow students, parents, and teachers to challenge the evaluation results and ensure prompt corrections. For example, when inaccurate results are detected, a system that enables experts to intervene appropriately can protect students' rights while enhancing the system's reliability.

Finally, it is important to ensure that algorithms based on standard datasets do not evaluate students with unique characteristics. For instance, students who require special support or come from diverse cultural and social backgrounds may face the risk of unfair treatment based on algorithm-based evaluations. To address this, it is imperative to design algorithms incorporating fairness into their framework. Organizing datasets that account for diverse backgrounds and implementing measures to mitigate harmful biases are critical. Furthermore, combining quantitative algorithmic results with qualitative assessments by teachers and experts can lead to comprehensive and equitable evaluations. By addressing these concerns, healthtech systems can better support students' diverse needs while ensuring fairness, transparency, and trustworthiness in educational settings (Table 1).

Information Acquisition and Management

It is essential to carefully consider the ELSI concerns related to the collection and management of information through healthtech [11,12,15]. In an educational setting, the acquisition of sensitive personal data on developmental disorders or mental health conditions requires thorough deliberation. While such information is important for enhancing algorithms' predictive capabilities related to health conditions, considering its highly sensitive nature, even with student or parental consent, the purposes and methods of its use must be clearly defined. Specifically, if such information is used for class organization or decision-making in educational policies, it is crucial to eliminate any possibility of unfair discrimination or inequitable treatment. To this end, it is necessary to obtain explicit consent from individuals or their parents and ensure that the information collected is used exclusively for educational support. Moreover, a strict privacy protection framework should be established to ensure that sensitive information is not disclosed to other students or teachers.

When institutions beyond schools and boards of education, such as university researchers or private companies, are involved in the design and operation of healthtech, data handling becomes increasingly complicated, necessitating stricter management systems. First, the retention period for this information must be clearly defined. If it is excessively lengthy, the risk of data being used in unforeseen ways increases. For example, data related to a student's health condition or psychological characteristics could potentially be shared with third parties to influence their future academic or career opportunities. To prevent such risks, the information retention period should be limited to the minimum necessary to achieve its purpose, and the data should be promptly deleted once the period expires. In addition, it is crucial to clearly define who is responsible for data management and assign responsibilities accordingly. For instance, when university researchers use data, clear agreements or contracts should specify the research purpose and scope of use to prevent its inappropriate use. However, when private companies manage healthtech, strict legal regulations and contracts are required to ensure that data are not exploited for commercial purposes that go beyond the intended scope of educational or health-related benefits. For instance, using student data to enhance algorithms that improve child protection and welfare may be considered legitimate use. However, repurposing the data for targeted advertising, user profiling for commercial products, or selling aggregated datasets to third parties should be strictly prohibited to prevent the misuse of sensitive student information.

Furthermore, access to the information collected should be restricted to a minimum number of relevant personnel. For example, only those directly supporting students, such as teachers and counselors, should be allowed to access the data. External individuals, including researchers, corporate staff, and members of the board of education, should be granted access only to anonymized data. The risk of identifying individual students can be minimized by anonymizing and encrypting the data.

When information collected through healthtech is linked to other data (eg, academic test scores or family background information), it is necessary to verify that the process and management system for linking these data are adequate. Although the linked data can potentially contribute to comprehensive support for individual students, improper handling may lead to privacy violations or reinforce bias. For example, there is a risk that a particular student may be unfairly labeled as "problematic." To avoid such risks, it is essential that

XSL•FO
RenderX

the data-linking process is conducted fairly and transparently, with clear definitions of the purposes for which the linked data will be used. By addressing these considerations, healthtech can be used to promote mental health while safeguarding students' privacy and rights.

Evaluation

When healthtech evaluates students' mental health, it is possible that their responses may converge toward answers perceived as "favorable" or "likely to receive a good evaluation." For example, when questions about stress or anxiety are included, students may conceal their true feelings and choose responses that they believe are more acceptable to evaluators, driven by the desire to present themselves in a better light [22]. Such convergence can undermine the recognition of individual differences and diverse values, instead encouraging self-expression that conforms to standardized criteria. In addition, this may result in genuine mental health issues being overlooked. To address this, it is crucial to include open-ended questions in addition to multiple-choice formats that provide children with the opportunity to express their feelings and circumstances in their own words.

Moreover, quantifying elements such as mental health and psychological traits—domains that are inherently multifaceted and context-dependent—could lead to the problem of "overquantification." These areas are not easily reduced to simple numerical values, and quantification attempts may inadvertently reinforce misunderstandings or biases. For instance, the richness and diversity of students' emotions and behaviors may be overlooked when they are reduced to scores or indices. In addition, when evaluation results are presented as numerical values, educators and parents may overrely on these figures, leading to rigid educational policies or support strategies. To prevent this, it is imperative to avoid excessive reliance on quantified results. Instead, establishing a system that combines these results with interpretations based on observations of daily behaviors and attitudes by teachers and parents can facilitate comprehensive evaluations.

Finally, concerns have been raised that the "individual optimization" recommended by certain systems—meaning the provision of tailored resources and interventions based on a student's unique needs and characteristics—may become a "de facto standard," that is, a widely adopted default practice not formally validated, thereby solidifying specific values as universal norms. For example, a risk-alert algorithm designed for one system may be adopted elsewhere without verifying its suitability in the new system. In such cases, the system's recommendations can be widely adopted in educational settings, potentially narrowing the diverse possibilities for each student and hindering flexible educational approaches. To mitigate this risk, it is essential to design algorithm evaluation criteria that incorporate diverse perspectives and establish mechanisms for regularly reviewing optimization recommendations. In addition, it is important to ensure that teachers and parents retain the ability to make flexible judgments regarding the system's recommendations.

Use

The evaluation results provided by healthtech should be collected primarily to support students' development and well-being. However, using these results as selection materials for admission or academic progress can help identify students who may require early individual support and prepare to provide appropriate assistance after enrollment. However, when the evaluation results are used as these selection criteria, students and their parents may become overly conscious of the evaluation, potentially distorting its original purpose of promoting growth and providing support. In addition, using evaluation results in the selection process may lead to the labeling of certain scores or traits, unfairly categorizing students as "unfit" or "problematic." Furthermore, data related to mental health or developmental conditions are highly sensitive, and their use in selection processes can increase the risk of violating the privacy of students. Therefore, using the evaluation results from healthtech for selection purposes requires a thorough and careful discussion.

In screenings or interventions conducted through healthtech, students who report issues or difficulties are more likely to receive risk assessment and support, whereas those who fail to report such issues may be overlooked. For example, students with introverted personalities or difficulty expressing themselves may struggle to articulate their challenges, leading the system to incorrectly determine that no problems exist. Conversely, students who frequently report issues may attract excessive attention, potentially leading to disproportionate allocation of limited support resources. The determination of support should not rely solely on the frequency or severity of the reported issues. It is also necessary to consider how to support students who struggle with self-expression or face latent challenges by supplementing self-report assessments with observational data from teachers and parents, daily behavior records, and consideration of factors such as students' age, sex, and other demographics.

The use of data collected through healthtech should also be discussed in terms of its application as anonymized data. Although the use of anonymized data has significant educational, academic, and social potential, schools may not actively promote their use, potentially amid concerns about the risk of reidentification of individuals and the perception among parents and students that anonymized data constitutes "surveillance" or "data exploitation." To enable flexible use of anonymized information, it is essential to clearly define its purpose, establish robust data management systems, and ensure careful consideration and transparency. With these measures in place, anonymized information can be used to allow all stakeholders to share their benefits.

Nature as Public Institutions

Topics such as privacy and consent, as discussed earlier, extend beyond ELSI concerns in educational settings and are also relevant in fields such as mental health and psychiatry. However, the nature of public institutions should also be considered, especially in light of the increasing implementation of healthtech in schools. For example, careful consideration of multiple perspectives regarding the financial burden of digital devices

XSL•FO
RenderX

is necessary. With digital transformation advancing rapidly in the education sector, there is a need to develop a sustainable support model that goes beyond merely providing hardware. When allocating public funds, it is crucial to consider not only the initial investment but also long-term maintenance, update, and support costs. Comprehensive support schemes can include initiatives by local governments and educational committees, cost-sharing models in collaboration with corporations, and the establishment of social support funds. Concrete support measures are particularly important in ensuring that students from economically vulnerable households are not excluded. For example, strategies such as device-leasing systems, subsidies for low-income households, and shared-use systems developed by local communities and educational institutions should be actively explored.

In Japan, schools established based on the School Education Act are primarily eligible for financial support. Schools that fall under article 1 of this law, commonly referred to as "Article One Schools," include elementary, junior high, high, and special support schools, which operate in accordance with the standards set by the MEXT. These standards encompass aspects such as curriculum, facilities, and teacher placement. Only schools that met these criteria were considered eligible for public financial support and subsidies. However, many alternative schools that do not meet these criteria—and therefore do not fall under article 1 of the School Education Act—are not eligible for public financial support. Consequently, providing healthtech-related subsidies to these schools presents complex challenges in terms of educational diversity and fairness. Incorporating multiple perspectives is crucial when considering subsidies for healthtech. First, there is a need to clarify the purpose and effects of these subsidies. A rigorous evaluation is necessary to determine the extent to which these measures contribute to specific goals, such as student health. Establishing guidelines, including ensuring transparency in the use of funds, periodic verification of effectiveness, and mandating performance reports, is essential.

Furthermore, careful consideration should be given to ensuring alignment with long-term educational policies. Establishing national evaluation and recommendation standards for mental health requires particular caution, considering that students' mental health is closely related to individual dignity, privacy, and diversity during the developmental stages. Diverse perspectives must be considered when establishing national standards. First, objective evaluation indicators should be developed based on scientific evidence. Simultaneously, a flexible framework is essential for respecting individual diversity and preventing uniform assessments. Therefore, a comprehensive approach that considers the age, developmental stage, cultural background, and individual characteristics is critical. Moreover, mental health standards must not promote stigma or prejudice but foster understanding and support. Multilayered considerations are necessary, including interdisciplinary evaluations by experts, reflections on the perspectives of those directly affected and their families, ongoing research and standard updates, and strict information management to safeguard privacy.

Relationships With Private Companies

As an issue related to public institutions, it is essential to carefully examine their relationships with private companies. Kucirkova [23] observed that educational technologies typically tend to overlook the principles of learning sciences while including distracting elements or manipulative designs that could reduce learning effectiveness and have an indirect negative impact on mental health. This issue appears to be a potential concern for healthtech as well. There are several other issues to be considered. For instance, if educational institutions do not fully understand the nature of healthtech or the changes made to it, they risk becoming overly dependent on service providers and compromising their autonomy. Private companies operate through multiple mechanisms. First, while determining technical specifications and functionalities, there is a risk that companies' commercial logic may take precedence over the needs of educational settings. Furthermore, through ongoing updates and maintenance contracts, schools may become increasingly dependent on specific corporate ecosystems in the long term. In addition, during data collection and analysis, companies may monopolize critical insights regarding educational processes, thereby gaining substantial influence over de facto educational policy formation.

Another concern is that private companies' values and evaluation criteria may infiltrate public education, raising serious concerns regarding the neutrality and public nature of education. Through healthtech, the commercial and ideological values of companies may inadvertently permeate students' minds. Evaluation criteria in education should be based on pedagogical perspectives and holistic development of children and should not be distorted by the profit motives or specific values of private enterprises. Therefore, the process of introducing healthtech must involve the participation of educational scholars, ethics experts, and parental representatives to establish multilayered checks and balances to ensure that technology does not deviate from the essential purposes of education.

Moreover, the student data collected through healthtech includes highly sensitive and personal information, and the misuse of such data by private companies for purposes other than those originally intended has the potential to seriously violate students' human rights. Therefore, strict legal regulations, transparent consent processes, and rigorous data management and protection mechanisms are essential. Specifically, this includes clearly limiting data use purposes, prohibiting third-party sharing, conducting regular audits, implementing comprehensive consent processes involving both students and their parents, and ensuring the right to data deletion.

Conclusions

So far, we have emphasized that using healthtech in Japanese school education requires addressing ELSIs from eight key perspectives: (1) consent, (2) rights and privacy, (3) algorithms, (4) data collection and management, (5) evaluation, (6) use, (7) nature of public institutions, and (8) relationships with private companies. The applicability of informed consent and opt-out mechanisms has been discussed from the perspective of consent, focusing on the challenge of respecting the intentions of both

XSL•FO
RenderX

parents and children. In terms of rights and privacy, guaranteeing the right to refuse evaluation and the right to data deletion has been emphasized, requiring systems to ensure these rights. Third-party auditing, error correction mechanisms, and fair and unbiased evaluation criteria are essential algorithms. For information collection and management, appropriate handling of sensitive data, limitations on storage periods, and prevention of unauthorized data use are critical challenges. In evaluation, considerations include the risk of evaluations suppressing diversity, the dangers of "overquantification," and the potential fixation of individual optimization as a de facto standard. Use focuses on the appropriateness of using evaluation results for selection purposes, disparities in support between children who express difficulties and those who do not, and cautious consideration of anonymized data use. Furthermore, the public nature of schools necessitates multilayered oversight mechanisms to maintain neutrality and public accountability in education while integrating technology. Finally, relationships with private companies underscore the importance of ensuring transparency in preventing the influence of commercial values on education.

This project incorporated 52 issues spanning 8 topics based on Ed Tech 101 [17] (Table 1). Figure 1 provides a visual summary of the topics to be considered before and after implementation, as well as the appropriate institutions for these discussions. It is worth noting that many issues span more than 1 topic. For example, the issue "Will not schools and teachers emulate the methods of healthtech to voluntarily collect excessive personal information?" relates not only to information collection but also to consent, further rights, and privacy. This finding underscores the need to examine each issue from multiple perspectives. The table (Table 1; Multimedia Appendix 1) serves as a reference for diverse stakeholders in healthtech, including school teachers, boards of education, university researchers, and developers in private companies. To facilitate discussion, we have included explanations for technical terms such as "informed consent" and "assent," as well as less-intuitive concepts such as "risk assessment." Furthermore, as many of the issues are related to laws such as the Constitution, the Basic Act on Education, and the Act on the Protection of Personal Information, we have also provided references to relevant legislation. One relevant case is the Asahikawa Standardized Testing Case, where the court held that teachers' freedom of education, while constitutionally protected, can be reasonably limited by public interest and institutional frameworks. This suggests that healthtech should be implemented within educational governance structures, not solely at individual discretion. Another relevant case concerned the disclosure of bullying survey responses, in which access was limited due to the risk of student reidentification based on handwriting. This highlights the need to address reidentification risks in healthtech design to protect student privacy, even when data are anonymized.

The introduction of healthtech brings new possibilities to educational settings [3,4]; however, it must be implemented in a fair and ethical manner. As a domestic example of a policy model for healthtech governance in education, we referred to the Education Data Utilization Roadmap [24] issued by Japan's Digital Agency. This document outlines core principles, such as clarity of purpose, interoperability, and informed consent, all of which are highly relevant to the ELSI framework for implementing healthtech in schools. Notably, it emphasizes that data use should be designed not for administrative efficiency or vendor convenience but based on the best interests of children. This national-level policy serves as a concrete reference for aligning healthtech governance with ethical and educational priorities. Stakeholders involved in healthtech are encouraged to refer to this table, engage in thorough discussions tailored to their specific contexts, and base their practices on these deliberations.

To address the risks identified in this study, a phased approach is recommended. In the short term, concrete guidelines for obtaining consent should be developed, and efforts should be made to enhance the transparency of algorithmic decision-making. Pilot implementation of third-party audits could be introduced in selected educational settings. In the medium term, mechanisms for auditing and validating healthtech tools should be gradually established at the level of education boards or national authorities. Systems should also be put in place to collect feedback from users—including students, parents, and teachers—and to allow for correction requests. In the long term, policies promoting fairness and bias reduction should be integrated into educational guidelines. In addition, the sharing of international data privacy standards and best practices should be encouraged to support responsible and effective use of healthtech.

This study primarily focuses on Japan's education system, which is characterized by a relatively centralized structure governed by national standards and policies. As such, the generalizability of the identified ELSI concerns and recommendations may vary in countries with decentralized or locally governed education systems. Further research is needed to explore how these considerations might differ in diverse educational contexts. However, the insights gained here provide valuable implications for discussions on the implementation of healthtech in educational settings worldwide. The ELSIs of healthtech—such as privacy, consent, and algorithmic transparency—are fundamental principles that can be applied to educational systems globally. However, its implementation must be adapted to each country's legal framework and cultural contexts, particularly considering differences in public education policies and operational guidelines. For example, in the European Union, the General Data Protection Regulation establishes strict standards for data protection, including specific provisions for data processing in educational contexts. In the United States, the Family Educational Rights and Privacy Act governs the access to and privacy of student education records. These frameworks provide useful reference points for understanding how ELSI principles are operationalized in diverse legal systems and highlight the importance of adapting healthtech practices to meet international privacy and consent standards. Future research should further explore the applicability and challenges of healthtech across different educational systems through international comparisons, contributing to the development of more universal and practical guidelines.

XSL•FO
RenderX

Acknowledgments

This work was supported by a JSPS KAKENHI grant (24K05786) to TN, the Japan Agency for Medical Research and Development (AMED, Japan) grant (23gn0110071h0001) to TN and MW, and a JST-Mirai Program grant (23-221037845).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Ethical, legal, and social issues on health technology in the educational context (English/Japanese).
[PDF File (Adobe PDF File), 1076 KB - formative_v9i1e72781_app1.pdf]

References

1. Au WW, Recchia F, Fong DY, Wong SH, Chan DK, Capio CM, et al. Effect of wearable activity trackers on physical activity in children and adolescents: a systematic review and meta-analysis. Lancet Digit Health 2024 Sep;6(9):e625-e639 [FREE Full text] [doi: 10.1016/S2589-7500(24)00139-0] [Medline: 39112110]
2. Zhang W, Xiong K, Zhu C, Evans R, Zhou L, Podrini C. Promoting child and adolescent health through wearable technology: a systematic review. Digit Health 2024;10:20552076241260507 [FREE Full text] [doi: 10.1177/20552076241260507] [Medline: 38868368]
3. Nishimura T, Wakuta M, Osuka Y, Tsukui N, Hirata I, Takahashi M, et al. Early detection of students' mental health issues from a traditional daily health observation scheme in Japanese schools and its digitalization. Front Public Health 2024;12:1430011 [FREE Full text] [doi: 10.3389/fpubh.2024.1430011] [Medline: 39314787]
4. Porter A, d'Apice K, Albers P, Woodrow N, Fairbrother H, Breheny K, et al. The impact of the newly developed school-based 'digital health contact'-evaluating a health and wellbeing screening tool for adolescents in England. PLoS One 2024;19(1):e0297016 [FREE Full text] [doi: 10.1371/journal.pone.0297016] [Medline: 38215072]
5. Soneson E, Howarth E, Ford T, Humphrey A, Jones PB, Thompson Coon J, et al. Feasibility of school-based identification of children and adolescents experiencing, or at-risk of developing, mental health difficulties: a systematic review. Prev Sci 2020 Jul;21(5):581-603 [FREE Full text] [doi: 10.1007/s11121-020-01095-6] [Medline: 32062764]
6. Weist MD, Rubin M, Moore E, Adelsheim S, Wrobel G. Mental health screening in schools. J Sch Health 2007 Mar;77(2):53-58. [doi: 10.1111/j.1746-1561.2007.00167.x] [Medline: 17222155]
7. Woodrow N, Fairbrother H, Breheny K, d'Apice K, Albers PN, Mills C, et al. Exploring the potential of a school-based online health and wellbeing screening tool: professional stakeholders' perspectives and experiences. BMC Public Health 2022 Mar 16;22(1):324 [FREE Full text] [doi: 10.1186/s12889-022-12748-2] [Medline: 35168580]
8. Woodrow N, Fairbrother H, D'Apice K, Breheny K, Albers P, Mills C, et al. Exploring the potential of a school-based online health and wellbeing screening tool: young people's perspectives. Int J Environ Res Public Health 2022 Mar 29;19(7):4062 [FREE Full text] [doi: 10.3390/ijerph19074062] [Medline: 35409747]
9. Bartlett Ellis R, Wright J, Miller LS, Jake-Schoffman D, Hekler EB, Goldstein CM, et al. Lessons learned: beta-testing the digital health checklist for researchers prompts a call to action by behavioral scientists. J Med Internet Res 2021 Dec 22;23(12):e25414 [FREE Full text] [doi: 10.2196/25414] [Medline: 34941548]
10. Doraiswamy PM, London E, Varnum P, Harvey B, Saxena S, Tottman S, et al. Empowering 8 billion minds: enabling better mental health for all via the ethical adoption of technologies. NAM Perspect 2019;2019:31478 [FREE Full text] [doi: 10.31478/201910b] [Medline: 34532674]
11. Wies B, Landers C, Ienca M. Digital mental health for young people: a scoping review of ethical promises and challenges. Front Digit Health 2021;3:697072 [FREE Full text] [doi: 10.3389/fdgth.2021.697072] [Medline: 34713173]
12. Shen FX, Silverman BC, Monette P, Kimble S, Rauch SL, Baker JT. An ethics checklist for digital health research in psychiatry: viewpoint. J Med Internet Res 2022 Mar 09;24(2):e31146 [FREE Full text] [doi: 10.2196/31146] [Medline: 35138261]
13. Subbian V, Galvin HK, Petersen C, Solomonides A. Ethical, legal, and social issues (ELSI) in mental health informatics. In: Tenenbaum JD, Ranallo PA, editors. Mental Health Informatics: Enabling a Learning Mental Healthcare System. Cham, Switzerland: Springer; 2021:479-503.
14. Unsworth H, Dillon B, Collinson L, Powell H, Salmon M, Oladapo T, et al. The NICE evidence standards framework for digital health and care technologies - developing and maintaining an innovative evidence framework with global impact. Digit Health 2021;7:20552076211018617 [FREE Full text] [doi: 10.1177/20552076211018617] [Medline: 34249371]
15. Seiferth C, Vogel L, Aas B, Brandhorst I, Carlbring P, Conzelmann A, et al. How to e-mental health: a guideline for researchers and practitioners using digital technology in the context of mental health. Nat Mental Health 2023 Aug 07;1(8):542-554. [doi: 10.1038/S44220-023-00085-1]

XSL•FO
RenderX

16. Improvement of academic abilities (courses of study). Ministry of Education, Culture, Sports, Science, and Technology. URL: https://www.mext.go.jp/en/policy/education/elsec/title02/detail02/1373859.htm [accessed 2025-05-29]
17. Kano K. Peta meaning: expectations for educational data utilization EdTech and ELSI (ethical, legal, and social issues). Inf Process 2024;64(7):63 [FREE Full text] [doi: 10.20729/00226351]
18. Oyake K, Sue K, Sumiya M, Tanaka S. Physical therapists use different motivational strategies for stroke rehabilitation tailored to an individual's condition: a qualitative study. Phys Ther 2023 Jun 05;103(6):pzad034 [FREE Full text] [doi: 10.1093/ptj/pzad034] [Medline: 37017336]
19. Sumiya M, Igarashi K, Miyahara M. Emotions surrounding friendships of adolescents with autism spectrum disorder in Japan: a qualitative interview study. PLoS One 2018;13(2):e0191538 [FREE Full text] [doi: 10.1371/journal.pone.0191538] [Medline: 29408894]
20. Chartier M, Stoep AV, McCauley E, Herting JR, Tracy M, Lymp J. Passive versus active parental permission: implications for the ability of school-based depression screening to reach youth at risk. J Sch Health 2008 Mar;78(3):157-164 [FREE Full text] [doi: 10.1111/j.1746-1561.2007.00278.x] [Medline: 18307611]
21. Villarreal V, Castro-Villarreal F, Peterson LS, Bear M, Cortes DM, Escobedo T. Meta-analysis of proportions of students screened and identified in mental health multiple-gate screening research. Sch Psychol Rev 2022 Aug 23;52(2):130-143. [doi: 10.1080/2372966x.2022.2106155]
22. Van de Mortel TF. Faking it: social desirability response bias in self-report research. Aust J Adv Nurs 2008;25(4):40-48 [FREE Full text]
23. Kucirkova N. Debate: response to "should academics collaborate with digital companies to improve young people's mental health". Child Adolesc Ment Health 2023 May;28(2):336-337. [doi: 10.1111/camh.12648] [Medline: 36948604]
24. Roadmap on the utilization of data in education. Digital Agency. URL: https://cio.go.jp/sites/default/files/uploads/documents/digital/20220307_en_education_outline_01.pdf [accessed 2025-05-29]

Abbreviations

DHC: Digital Health Contact
ELSIs: ethical, legal, and social issues
healthtech: health technology
MEXT: Ministry of Education, Culture, Sports, Science and Technology

Edited by A Mavragani; submitted 18.02.25; peer-reviewed by C Chikwe, M Adnan; comments to author 01.05.25; revised version received 08.06.25; accepted 10.06.25; published 18.07.25.

Please cite as:
Sumiya M, Nishimura T, Aizaki K, Hirata I, Tsukui N, Osuka Y, Wakuta M, Senju A
Addressing the Ethical, Legal, and Social Issues of Healthtech in Education: Insights From Japan
JMIR Form Res 2025;9:e72781
URL: https://formative.jmir.org/2025/1/e72781
doi:10.2196/72781
PMID:

©Motofumi Sumiya, Tomoko Nishimura, Kyoko Aizaki, Ikue Hirata, Nobuaki Tsukui, Yuko Osuka, Manabu Wakuta, Atsushi Senju. Originally published in JMIR Formative Research (https://formative.jmir.org), 18.07.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Exploring the Role of Telemedicine in Duchenne Muscular Dystrophy: Benefits and Challenges

Eliza Wasilewska[1], MD, PhD; Andrzej Wasilewski[2], MR; Alessandro Onofri[3], MD; Jan Wasilewski[4], MSc; Dominika Sabiniewicz-Ziajka[5], MD, PhD; Agnieszka Sobierajska-Rek[6], MSc, PhD; Jarosław Meyer-Szary[6], MD, PhD; Karolina Śledzińska[7], MD, PhD; Jolanta Wierzba[7], MD, PhD; Marek Niedoszytko[1], MD, PhD; Sylwia Małgorzewicz[8], MD, PhD

[1]Department of Allergology and Pulmonology, University Clinical Center Gdańsk (Accredited Duchenne Center), Medical University of Gdańsk, Gdansk, Poland

[2]Student Scientific Association of Medical Chemistry and Immunochemistry, Wroclaw Medical University, Wrocław, Poland

[3]Pediatric Pulmonology & Cystic Fibrosis Unit, Respiratory Intermediate Care Unit, Sleep and Long-Term Ventilation Unit, Bambino Gesù Children's Hospital, Scientific Institute for Research, Rome, Italy

[4]Rochester Institute of Technology, Rochester, NY, United States

[5]Department of Rehabilitation, Medical University of Gdańsk, Gdańsk, Poland

[6]Department of Pediatric Cardiology, University Clinical Center (Accredited Duchenne Center), Medical University of Gdańsk, Gdańsk, Poland

[7]Department of Paediatrics, Haematology and Oncology, University Clinical Center (Accredited Duchenne Center), Medical University of Gdansk, Gdańsk, Poland

[8]Department of Clinical Nutrition, University Clinical Center Gdańsk (Accredited Duchenne Center), Medical University of Gdańsk, Gdańsk, Poland

Corresponding Author:
Sylwia Małgorzewicz, MD, PhD
Department of Clinical Nutrition
University Clinical Center Gdańsk (Accredited Duchenne Center)
Medical University of Gdańsk
Sklodowskiej-Curie 3a
Gdańsk, 80-210
Poland
Phone: 1 +48583492723
Email: sylwiam@gumed.edu.pl

Abstract

Duchenne muscular dystrophy (DMD) is the most frequent, progressive disease caused by a genetic defect that leads to the production of a nonfunctional form of dystrophin, thereby causing premature death. Ways to improve, adapt, and facilitate the care of people with DMD are still being explored. This viewpoint, developed by an accredited Duchenne center, aims to present current telemedicine options specifically tailored for patients with DMD and to discuss the advantages and limitations of these approaches across various health care domains. As one of the first centers in Poland to implement such an approach, the accredited Duchenne center provides targeted home-based care by using digital platforms and telemedicine tools. Additionally, we explore the potential of telemedicine to support different types of remote communication, including provider-to-provider, between patient/caregiver and provider, and between patient/caregiver and patient/caregiver interactions. This model has the potential to significantly enhance access to specialized care and improve the continuity and quality of life for those living with DMD.

(JMIR Form Res 2025;9:e77698) doi:10.2196/77698

KEYWORDS

neuromuscular disorders; Duchenne muscular dystrophy; telemedicine; eHealth; digital medicine

Introduction

A patient with Duchenne muscular dystrophy (DMD) has unique specific needs since the disease is not only chronic but also progressive and leads to premature death in the second or third decade of life. The essence of DMD is progressive muscle weakness caused by a genetic defect that leads to the production of a nonfunctional form of dystrophin. As a result, muscle cells are irreversibly damaged [1,2].

XSL•FO
RenderX

To date, there is no effective causal treatment for patients with DMD. Gene therapies such as exon skipping, stop codon read-through, or adeno-associated virus gene therapy are promising but limited by many factors, for example, the type of mutation [3]. Therefore, the basic treatment still remains—apart from corticosteroid therapy, multidisciplinary specialist care. Due to the interplay of muscles and organs, an interdisciplinary care team (according to the guidelines) should include a neurologist, pulmonologist, physiotherapist, cardiologist, dietitian, psychologist, endocrinologist, and other health workers [4,5]. The frequency of the follow-up visits depends on the patient's clinical condition (at least once a year), but the increase in the survival rate has resulted in an increase in the severity of the disease and the need for more frequent contact with health care professionals [1,2,4]. This may be inconvenient or impossible, particularly for people with reduced mobility and/or who live a long way from the research center. Ways to improve, adapt, and facilitate the care of people with DMD are still being explored.

One way may be the use of eHealth, which is a rapidly growing field and can support many health care processes. As of 2012, only 31% of the hospitals and 15% of the outpatient clinics were using telemedicine services in the European Union [6,7]. Currently, digital health interventions are available for most people in the United States, Europe, Canada, and Australia [8,9]. Notably, the pandemic was a catalyst for the development of new forms of virtual medical care, and it significantly accelerated the introduction of telemedicine in patients with DMD.

This up-to-date viewpoint based on the experience of our accredited center and collaboration with other centers for children with DMD aims to provide current telemedicine options dedicated to patients with DMD and discusses the pros and cons in different health care areas. We also explore the possible use of telemedicine services at a distance for communication conducted between 2 or more providers (provider-to-provider telemedicine), between patients/caregivers and providers (client-to-provider telemedicine), and between 2 or more patients/caregivers (client-to-client telemedicine).

In this paper, we have focused only on telemedicine due to its rapid development and increasing widespread use, even though it is only one of the components of the broadly understood eHealth according to the World Health Organization (WHO) [10].

Types of Telemedicine

Client-To-Provider Telemedicine

The rapid increase in the use of telemedicine during the COVID-19 pandemic was evident, and many countries implemented individualized DMD telecare approaches [11-16]. For example, in Italy, remote clinical evaluation and administration of functional scales were lacking for neuromuscular diseases before the pandemic [17], but these increased rapidly during the pandemic. Similarly, in France, teleconsultation guidelines underwent rapid changes, allowing various forms of teleconsultation, including telephone, whereas videoconferencing was the primary method used before the pandemic [18]. In Norway, at the same time, 14% of the medical professionals used telemedicine to contact a patient with a rare disease and monitor its development during the pandemic [19].

Many well-organized telemedicine programs have been related to other neuromuscular diseases. In 2014, in Massachusetts, TelePALS (Telemedicine for People with Amyotrophic Lateral Sclerosis) was performed with follow-up structured care via video televisits [20]. In another study in 2019, a trained nurse traveled to patients' homes and performed and recorded video assessments for later review by members of the multidisciplinary team [21]. Yet another project in 2018 and 2020 included a mobile app for weekly surveys about symptoms, complications, noninvasive ventilation use, and gastrostomy use to communicate with clinic staff [22,23]. These projects showed that telemedicine is feasible and acceptable to both health care professionals and patients with amyotrophic lateral sclerosis. A telemedicine program in 2018 for another neuromuscular disease, facio-scapulo-humeral muscular dystrophy, has also shown improvement in the quality of life and the possibility of reducing hospitalizations [24].

Provider-To-Provider Telemedicine

Telemedicine has opened up opportunities for remote connections between health care professionals. The e-community of health care providers has become a common tool not only for scientific purposes but also for everyday clinical practice involving patients with DMD [25]. In Italy, during the COVID-19 pandemic, efforts were made to create a network for professionals and less-experienced doctors to enhance their knowledge of DMD [26]. This network also facilitated education through webinars and collaboration among distant researchers, projects, and clinical trials. Collaborative efforts between different centers to create large databases have helped in developing new medicines and policies for rare diseases. As of May 14, 2023, over 300 clinical trials on DMD have been registered on the clinical trial platforms (clinicaltrials.gov and clinicaltrialsregister.eu).

Another example of collaborative cooperation is the European Reference Networks established in 2017. These networks connect medical specialists across different disciplines through a dedicated information technology platform and telemedicine tools, effectively forming a virtual advisory board. EURO-NMD (European Reference Network for the thematic grouping of rare neuromuscular diseases) is a European Reference Network dedicated to rare neuromuscular diseases, aiming to facilitate diagnostic processes, standardize therapeutic approaches, and improve the efficiency and accessibility of health care systems [27] (Figure 1).

XSL•FO
RenderX

Figure 1. EURO-NMD (European Reference Network for the thematic grouping of rare Neuromuscular Diseases). ERN: European Reference Network; NMD: neuromuscular disease.

Client-To-Client Telemedicine

Nearly 2 decades ago, the first teleproject was proposed in England to provide computer equipment and connectivity for families with DMD to reduce social isolation and to create a window into the world [28]. However, not all families were willing to participate in the program at that time. Since then, significant progress has been made in the use of electronic devices, equipment, social media, and applications. Organizations like the Parent Project of Muscular Dystrophy have developed their websites [29]. Technological barriers are diminishing, as mobile computing equipment becomes smaller and more affordable and wireless connectivity becomes more reliable and widespread. Currently, smartphones offer many possibilities to improve communication, access, and participation, and give a unique opportunity to directly deliver functionality to people with disabilities and their caregivers.

Applications of Telemedicine in Different Areas of DMD Health Care

Neurology

Neurological examinations for DMD require personal contact between doctors and patients, making virtual examinations during teleconsultations challenging [30]. Available data comparing virtual examinations to traditional physical examinations are limited, and standardized protocols for televisits are lacking in the literature [31-34].

To address these challenges, researchers have developed their televisit systems and protocols. Structured surveys such as the Brook and Vignos scales or Egen Scale Classification used for clinical evaluation and functionality can be useful for collecting standardized data [35-38]. Additionally, other scales commonly used in clinical trials (the 6-Minute Walk Test, Quantitative Muscle Testing, the DMD Functional Ability Self-Assessment Tool, or the Performance of the Upper Limb Scale) can be utilized in telemedicine for assessing muscle function and progression of disease [39-41]. For children or adult patients with advanced-stage DMD, the presence of a telepresenter during the virtual examination can be helpful [17,28].

Although telemedicine became prevalent during the pandemic, hybrid programs combining telemedicine with in-person care emerged after the pandemic. Nowadays, it is recommended that the initial diagnostic visit for patients with DMD takes place in-person at the center, with possible follow-up teleconsultations [42,43].

Pulmonology

Respiratory function is crucial for the quality and length of life of patients with DMD; therefore, annual evaluation of respiratory function is recommended for patients with gait preservation and every 6 months for those without or even more frequent if clinical symptoms dictate [44]. This includes collecting information on the type, number, and duration of respiratory infections, presence of cough reflex, daytime hypoventilation symptoms and sleep quality, as well as pulmonary function tests, including spirometry, peak cough flow, respiratory muscle strength (sniff nasal inspiratory pressure, maximal static inspiratory and expiratory pressures), and sleep study (polysomnography) with overnight capnography. Some of these procedures can be supported by telemedicine, for example, during virtual visits, health care professionals can provide information on respiratory care and collect data on chest infections and symptoms of hypoventilation or ineffective sleep (preferably using validated questionnaires, eg, Sleep-Disordered Breathing Questionnaire In Neuromuscular Patients [SiNQ-5]) [45].

In a case series involving 3 patients with neuromuscular disorders—including one with DMD—video consultations, supported by baseline physiological monitoring, facilitated early identification of respiratory complications and helped avoid hospitalization [46]. Further, telemonitoring of pulmonary function by using portable devices at home (e-spirometry with assessment of forced vital capacity), in addition to traditional spirometry in reference centers, can be employed in patients with DMD [47,48].

Web-based care services could potentially be beneficial for adolescent and young adults in the later stages of the disease who require ventilatory support, aiding in decision-making and providing timely interventions for respiratory issues. A study conducted during the pandemic found that weekly video consultations (averaged 23 minutes) and ventilator telemonitoring successfully addressed most problems related to general clinical status and ventilation in a group of 21 children (including 5 with neuromuscular diseases), who relied on long-term ventilatory support (noninvasive and invasive) and living at home [49]. That study also identified challenges with telemedicine, particularly the absence of sufficient technology at home, such as low internet signal or lack of internet connectivity, which affected 17% of the participants [49]. Studies presenting comprehensive virtual respiratory care dedicated exclusively to patients with DMD are generally lacking. However, data from pediatric patients with other

XSL•FO
RenderX

conditions requiring long-term respiratory support at home indicate that although clinical complications are common and respiratory decompensation is the most common cause, telemedicine can facilitate early diagnosis and treatment of life-threatening events [50]. In addition, it has been shown to reduce the frequency of hospitalizations and improve caregivers' sense of security and confidence in home care among ventilated patients [51].

Cardiology

DMD can lead to significant cardiac complications, including cardiomyopathy, arrhythmias, and heart failure, which are all potentially lethal [52,53]. Early diagnosis and treatment of cardiomyopathy leads to favorable ventricular remodeling [54]. Because adverse myocardial changes can occur at a younger age than previously thought and due to the importance of diagnosis before overt cardiac dysfunction, yearly cardiac screening is recommended [54]. Cardiac diagnostics with electrocardiography, echocardiography, Holter monitoring, and magnetic resonance imaging require in-person visits. However, telemedicine can provide access to specialized care for patients who may not have local access to cardiologists [55]. Telemedicine enables remote monitoring of cardiac function and arrhythmias by using wearable sensors and implantable devices, allowing for timely interventions to prevent or treat cardiac complications [56].

Nutrition

Proper nutrition is vital for patients with DMD. Eating disorders at the early stage of the disease due to steroid therapy and immobilization cause concern for excessive body weight. At later stages of the disease, there is a risk of malnutrition due to the weakening of the muscles of the gastrointestinal tract [4]. Teleconsultations can involve assessing the nutritional status, monitoring diet and fluid intake, and providing nutritional counseling. Web-based surveys such as the Food Frequency Questionnaire-6 item questionnaire [57] are useful in collecting dietary data. Software applications can assist patients in monitoring their dietary goals and measurements [58]. These interventions can address both excessive body weight in the early stages of the disease and the risk of malnutrition in later stages.

Psychology

Psychological support is needed at every stage of chronic disease. Studies [59] have shown that psychological and peer support via telehealth programs such as video sessions are useful and helpful for people with DMD and their caregivers. Telehealth programs can include health promotion causing improvement in the health-related quality of life. One study demonstrated the benefits of virtually delivered health coaching sessions (the first session was face-to-face, followed by 4 virtual sessions) related to physical activity and diet for males with DMD [60].

Telerehabilitation

According to the World Physiotherapy and the International Network of Physical Therapy Regulatory Authorities, the purpose of digital physical therapy practice is to facilitate the effective delivery of physical therapy services by improving access to care and information and managing health care resources [61]. Telerehabilitation consists of consulting, diagnostics, and therapy conducted with the use of interactive digital technologies in real-time connection (synchronous) or through prerecorded videos (asynchronous). Caregiver involvement is crucial for individuals with DMD to ensure the proper execution of exercises during telerehabilitation [62]. Web-based video instructions are more acceptable than live workshops [63]. Another study showed that 8-week telerehabilitation sessions enhanced muscle strength in ambulatory patients with DMD and that regular web-based meetings with physiotherapists were more effective than prerecorded or streaming video exercises [64]. Telerehabilitation can be supported by telemonitoring in patients with DMD (eg, cardiovascular parameters with electrocardiogram [65], overall physical activity level, step activity with accelerometers [66], effects of respiratory rehabilitation with home spirometers [47]). Virtual reality, including video games, has proven to be a valuable assistive technology for home-based rehabilitation [47].

Terminal Phase

Some studies [67-69] have shown that telemedicine can be used at every stage of chronic disease, including patients in the terminal phase of the disease, both in adults and children. This is important because patients with DMD can present hospice phase and have a premature death at a young age. Hospice programs utilizing telemedicine have mostly focused on children with oncological diseases, but the benefits extend to patients with neuromuscular diseases even in the form of simple teleconsultations [67-73]. Devices with touch screens, large letters, and easy-to-understand pictograms have proven useful in telemedicine for patients in the terminal phase [74]. Despite this, patients with end-stage DMD (requiring always multidimensional palliative support) remain underrepresented in telehealth research. Formative studies are needed to develop and evaluate integrated telemedicine approaches that address the specific needs of these patients.

Home health and nursing services may complement telemedicine in end-stage DMD by enabling in-home assessments, medication administration, and laboratory sample collection. This type of care can serve as an effective bridge between virtual supervision and hands-on clinical support, ensuring continuity for patients with complex needs. Evidence from other chronic conditions shows that nurse-led telehomecare can reduce hospitalizations and improve the quality of life [75,76]. However, this hybrid approach remains underexplored in neuromuscular disorders. Future studies should evaluate its feasibility by using indicators such as hospital admissions, clinical stability, and caregiver satisfaction.

Summary

Telemedicine provides numerous benefits for individuals with DMD, their caregivers, and health care professionals. It facilitates seamless communication among medical professionals from different centers, resulting in faster information exchange and the sharing of valuable experiences. By creating communities on social media platforms, individuals with DMD and their caregivers can find mutual support and gain a broader

XSL•FO
RenderX

perspective of the world. Telemedicine has found applications in many areas of DMD health care, including neurology, pulmonology, cardiology, dietetics, psychology, and even terminally ill patients. Telerehabilitation with virtual reality can be especially attractive to patients with DMD.

A limitation of telemedicine is the absence of standardized televisit protocols and a comprehensive virtual physical examination, including neurological assessments. As a result, the initial diagnostic visit necessitates in-person presence. Telerehabilitation is also not free of limitations. For example, it requires the caregiver's assistance, and for optimal outcomes, real-life physiotherapy cannot be fully replaced by web-based treatment. Technical barriers and privacy concerns can also limit the use of telemedicine. Although there is no substitute for personal contact with a doctor, it seems that patients with DMD can benefit greatly from digital health. Furthermore, patients with DMD and their family caregivers are relatively young and therefore can more easily adapt to the daily use of technology.

Clinical Consequences

Telemedicine cannot replace face-to-face visits—especially the first diagnostic visit must take place at the hospital. Telemedicine can support DMD health care when hospital visits are not possible. Telemedicine can be helpful as a window to the world for patients and their families and a quick way to exchange information and consultations between medical centers.

Authors' Contributions

Conceptualization: EW
Writing – original draft preparation: EW, AW, JW, DS, JMS
Writing – reviewing: AO, MN, SM
Editing: EW, AW
Selection and preparation of literature: ASR, KS, JW, MN
All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

None declared.

References

1. Birnkrant DJ, Bushby K, Bann CM, Apkon SD, Blackwell A, Brumbaugh D, et al. Diagnosis and management of Duchenne muscular dystrophy, part 1: diagnosis, and neuromuscular, rehabilitation, endocrine, and gastrointestinal and nutritional management. The Lancet Neurology 2018 Mar;17(3):251-267. [doi: 10.1016/s1474-4422(18)30024-3]
2. Birnkrant DJ, Bushby K, Bann CM, Alman BA, Apkon SD, Blackwell A, et al. Diagnosis and management of Duchenne muscular dystrophy, part 2: respiratory, cardiac, bone health, and orthopaedic management. The Lancet Neurology 2018 Apr;17(4):347-361. [doi: 10.1016/s1474-4422(18)30025-5]
3. Elangkovan N, Dickson G. Gene therapy for Duchenne muscular dystrophy. JND 2021 Nov 30;8(s2):S303-S316. [doi: 10.3233/jnd-210678]
4. Birnkrant DJ, Bushby K, Bann CM, Apkon SD, Blackwell A, Colvin MK, DMD Care Considerations Working Group. Diagnosis and management of Duchenne muscular dystrophy, part 3: primary care, emergency management, psychosocial care, and transitions of care across the lifespan. Lancet Neurol 2018 May;17(5):445-455 [FREE Full text] [doi: 10.1016/S1474-4422(18)30026-7] [Medline: 29398641]
5. Wasilewska E, Małgorzewicz S, Sobierajska-Rek A, Jabłońska-Brudło J, Górska L, Śledzińska K, et al. Transition from childhood to adulthood in patients with Duchenne muscular dystrophy. Medicina (Kaunas) 2020 Aug 24;56(9):426 [FREE Full text] [doi: 10.3390/medicina56090426] [Medline: 32846887]
6. Oh J, Park Y, Jo EC, Kim S. Current status and progress of telemedicine in Korea and other countries. Healthc Inform Res 2015 Oct;21(4):239-243 [FREE Full text] [doi: 10.4258/hir.2015.21.4.239] [Medline: 26618029]
7. Scott Kruse C, Karem P, Shifflett K, Vegi L, Ravi K, Brooks M. Evaluating barriers to adopting telemedicine worldwide: a systematic review. J Telemed Telecare 2016 Oct 16;24(1):4-12. [doi: 10.1177/1357633x16674087]
8. Craig J, Patterson V. Introduction to the practice of telemedicine. J Telemed Telecare 2005 Jan 01;11(1):3-9. [doi: 10.1258/1357633053430494]
9. Singhal A, Cowie M. What is e-Health? e-Journal of Cardiology Practice Vol 18. URL: https://www.escardio.org/Journals/E-Journal-of-Cardiology-Practice/Volume-18/what-is-e-health [accessed 2024-10-28]
10. Global observatory for eHealth. World Health Organization. URL: https://www.who.int/observatories/global-observatory-for-ehealth#_blank [accessed 2024-10-29]
11. Mauri E, Abati E, Musumeci O, Rodolico C, D'Angelo MG, Mirabella M, Italian Association of Myology. Estimating the impact of COVID-19 pandemic on services provided by Italian Neuromuscular Centers: an Italian Association of Myology survey of the acute phase. Acta Myol 2020 Jun;39(2):57-66. [doi: 10.36185/2532-1900-008] [Medline: 32904925]

XSL•FO
RenderX

12. Bamaga AK, Alghamdi F, Alshaikh N, Altwaijri W, Bashiri FA, Hundallah K, et al. Consensus statement on the management of Duchenne muscular dystrophy in Saudi Arabia during the coronavirus disease 2019 pandemic. Front Pediatr 2021;9:629549 [FREE Full text] [doi: 10.3389/fped.2021.629549] [Medline: 33681102]
13. Levine H, Prais D, Aharoni S, Nevo Y, Katz J, Rahmani E, et al. COVID-19 in advanced Duchenne/Becker muscular dystrophy patients. Neuromuscul Disord 2021 Jul;31(7):607-611 [FREE Full text] [doi: 10.1016/j.nmd.2021.03.011] [Medline: 34053847]
14. Veerapandiyan A, Wagner KR, Apkon S, McDonald CM, Mathews KD, Parsons JA, et al. The care of patients with Duchenne, Becker, and other muscular dystrophies in the COVID-19 pandemic. Muscle Nerve 2020 Jul;62(1):41-45 [FREE Full text] [doi: 10.1002/mus.26902] [Medline: 32329920]
15. Costamagna G, Abati E, Bresolin N, Comi GP, Corti S. Management of patients with neuromuscular disorders at the time of the SARS-CoV-2 pandemic. J Neurol 2021 May;268(5):1580-1591 [FREE Full text] [doi: 10.1007/s00415-020-10149-2] [Medline: 32804279]
16. Spina E, Trojsi F, Tozza S, Iovino A, Iodice R, Passaniti C, Digital Technologies, WebSocial Media Study Group of the Italian Society of Neurology (SIN). How to manage with telemedicine people with neuromuscular diseases? Neurol Sci 2021 Sep;42(9):3553-3559 [FREE Full text] [doi: 10.1007/s10072-021-05396-8] [Medline: 34173087]
17. Giannotta M, Petrelli C, Pini A. Telemedicine applied to neuromuscular disorders: focus on the COVID-19 pandemic era. Acta Myol 2022;41(1):6. [doi: 10.36185/2532-1900-066]
18. Solé G, Salort-Campana E, Pereon Y, Stojkovic T, Wahbi K, Cintas P, FILNEMUS COVID-19 study group. Guidance for the care of neuromuscular patients during the COVID-19 pandemic outbreak from the French Rare Health Care for Neuromuscular Diseases Network. Rev Neurol (Paris) 2020 Jun;176(6):507-515 [FREE Full text] [doi: 10.1016/j.neurol.2020.04.004] [Medline: 32354651]
19. Hagen K, Hummelvoll G. Practical videoconference training: experience from a Norwegian resource centre for rare disorders. Commun Med 2015;12(2-3):199-209. [doi: 10.1558/cam.16690] [Medline: 29048147]
20. Van De Rijn M, Paganoni S, Levine-Weinberg M, Campbell K, Swartz Ellrodt A, Estrada J, et al. Experience with telemedicine in a multi-disciplinary ALS clinic. Amyotroph Lateral Scler Frontotemporal Degener 2018 Feb;19(1-2):143-148. [doi: 10.1080/21678421.2017.1392577] [Medline: 29250986]
21. Pulley MT, Brittain R, Hodges W, Frazier C, Miller L, Matyjasik-Liggett M, et al. Multidisciplinary amyotrophic lateral sclerosis telemedicine care: the store and forward method. Muscle Nerve 2019 Jan;59(1):34-39. [doi: 10.1002/mus.26170] [Medline: 29802746]
22. Hobson EV, Baird WO, Partridge R, Cooper CL, Mawson S, Quinn A, et al. The TiM system: developing a novel telehealth service to improve access to specialist care in motor neurone disease using user-centered design. Amyotroph Lateral Scler Frontotemporal Degener 2018 Aug;19(5-6):351-361 [FREE Full text] [doi: 10.1080/21678421.2018.1440408] [Medline: 29451026]
23. Helleman J, Van Eenennaam R, Kruitwagen ET, Kruithof WJ, Slappendel MJ, Van Den Berg LH, et al. Telehealth as part of specialized ALS care: feasibility and user experiences with "ALS home-monitoring and coaching". Amyotroph Lateral Scler Frontotemporal Degener 2020 May;21(3-4):183-192 [FREE Full text] [doi: 10.1080/21678421.2020.1718712] [Medline: 32003245]
24. Portaro S, Calabrò RS, Bramanti P, Silvestri G, Torrisi M, Conti-Nibali V, et al. Telemedicine for facio-scapulo-humeral muscular dystrophy: a multidisciplinary approach to improve quality of life and reduce hospitalization rate? Disabil Health J 2018 Apr;11(2):306-309. [doi: 10.1016/j.dhjo.2017.09.003] [Medline: 28967584]
25. How telehealth can help patients with muscular dystrophy. Rare Disease Advisor. URL: https://www.rarediseaseadvisor.com/conferences/mda-2022/how-telehealth-can-help-patients-with-muscular-dystrophy; [accessed 2024-10-29]
26. Astrea G, Marinella G, Agosto C, Gagliardi D, Grandis M, Giuliano M, et al. How to define and enhance diagnostic and assistance pathways in neuromuscular diseases during the COVID-19 pandemic: the concept of network. Acta Myol 2021 Dec;40(4):172-176 [FREE Full text] [doi: 10.36185/2532-1900-060] [Medline: 35047757]
27. EURO-NMD. European Reference Network. URL: https://ern-euro-nmd.eu/educational-resources/ [accessed 2025-01-10]
28. Soutter J, Hamilton N, Russell P, Russell C, Bushby K, Sloper P, et al. The Golden Freeway: a preliminary evaluation of a pilot study advancing information technology as a social intervention for boys with Duchenne muscular dystrophy and their families. Health Soc Care Community 2004 Jan;12(1):25-33. [doi: 10.1111/j.1365-2524.2004.00465.x]
29. Parent Project Muscular Dystrophy. URL: https://www.parentprojectmd.org/ [accessed 2024-10-29]
30. Wechsler LR. Advantages and limitations of teleneurology. JAMA Neurol 2015 Mar;72(3):349-354. [doi: 10.1001/jamaneurol.2014.3844] [Medline: 25580942]
31. Contesse MG, Hodges J, Staunton H, Lowes LP, Elmankabadi B, Elder SJ, et al. Home-based video assessment of ease of movement for patients with Duchenne: interviews with physical therapists to select movement tasks. Physiother Res Int 2023 Jan 30;28(3):e1993. [doi: 10.1002/pri.1993] [Medline: 36718112]
32. Contesse MG, Lowes LP, White MK, Dalle Pazze L, McSherry C, Alfano LN, et al. Development of Duchenne Video Assessment scorecards to evaluate ease of movement among those with Duchenne muscular dystrophy. PLoS One 2022;17(4):e0266845 [FREE Full text] [doi: 10.1371/journal.pone.0266845] [Medline: 35417501]

XSL•FO
RenderX

33. Contesse MG, Sapp ATL, Apkon SD, Lowes LP, Dalle Pazze L, Leffler MG. Reliability and construct validity of the Duchenne Video Assessment. Muscle Nerve 2021 Aug;64(2):180-189 [FREE Full text] [doi: 10.1002/mus.27335] [Medline: 34050939]
34. White MK, Leffler M, Rychlec K, Jones C, McSherry C, Walker L, et al. Adapting traditional content validation methods to fit purpose: an example with a novel video assessment and training materials in Duchenne muscular dystrophy (DMD). Qual Life Res 2019 Nov;28(11):2979-2988 [FREE Full text] [doi: 10.1007/s11136-019-02245-2] [Medline: 31302840]
35. Brooke MH, Griggs RC, Mendell JR, Fenichel GM, Shumate JB, Pellegrino RJ. Clinical trial in Duchenne dystrophy. I. The design of the protocol. Muscle Nerve 1981;4(3):186-197. [doi: 10.1002/mus.880040304] [Medline: 7017401]
36. Vignos PJ. Management of progressive muscular dystrophy in childhood. JAMA 1963 Apr 13;184(2):89-96. [doi: 10.1001/jama.1963.03700150043007] [Medline: 13997180]
37. Miliani Capelini C, da Silva TD, Tonks J, Watson S, Boscolo Alvarez MP, Del Ciello de Menezes L, et al. Improvements in motor tasks through the use of smartphone technology for individuals with Duchenne muscular dystrophy. NDT 2017 Aug;13:2209-2217. [doi: 10.2147/ndt.s125466]
38. Ansary AM, Martinez JN, Scott JD. The virtual physical exam in the 21st century. J Telemed Telecare 2019 Nov 06;27(6):382-392. [doi: 10.1177/1357633x19878330]
39. Bushby K, Connor E. Clinical outcome measures for trials in Duchenne muscular dystrophy: report from International Working Group meetings. Clin Investig (Lond) 2011 Sep;1(9):1217-1235 [FREE Full text] [doi: 10.4155/cli.11.113] [Medline: 22639722]
40. Herrington WG, Goldsack JC, Landray MJ. Increasing the use of mobile technology-derived endpoints in clinical trials. Clin Trials 2018 Jun;15(3):313-315 [FREE Full text] [doi: 10.1177/1740774518755393] [Medline: 29400066]
41. Fujii T, Takeshita E, Iwata Y, Yajima H, Nozaki F, Mori M, et al. Cumulative jerk as an outcome measure in nonambulatory Duchenne muscular dystrophy. Brain Dev 2019 Oct;41(9):796-802. [doi: 10.1016/j.braindev.2019.06.002] [Medline: 31213334]
42. Saporta MA, Granit V, Lewis R, Benatar M. Yes, we can: neuromuscular examination by telemedicine. Muscle Nerve 2020 Dec;62(6):E83-E85. [doi: 10.1002/mus.27056] [Medline: 32875561]
43. Grogan J, Simmons Z. Telemedicine for the care of neuromuscular disorders. Curr Treat Options Neurol 2020 May 11;22(6):1 [FREE Full text] [doi: 10.1007/s11940-020-00625-5]
44. Childs A, Turner C, Astin R, Bianchi S, Bourke J, Cunningham V, et al. Development of respiratory care guidelines for Duchenne muscular dystrophy in the UK: key recommendations for clinical practice. Thorax 2024 Apr 15;79(5):476-485 [FREE Full text] [doi: 10.1136/thorax-2023-220811] [Medline: 38123347]
45. Steier J, Jolley C, Seymour J, Teschler H, Luo Y, Polkey M, et al. Screening for sleep-disordered breathing in neuromuscular disease using a questionnaire for symptoms associated with diaphragm paralysis. Eur Respir J 2011 Feb;37(2):400-405 [FREE Full text] [doi: 10.1183/09031936.00036210] [Medline: 20595146]
46. Zamarrón C, Morete E, González F. Telemedicine system for the care of patients with neuromuscular disease and chronic respiratory failure. Arch Med Sci 2014 Oct 27;10(5):1047-1051 [FREE Full text] [doi: 10.5114/aoms.2014.46223] [Medline: 25395959]
47. Wasilewska E, Sobierajska-Rek A, Małgorzewicz S, Soliński M, Jassem E. Benefits of telemonitoring of pulmonary function-3-month follow-up of home electronic spirometry in patients with Duchenne muscular dystrophy. J Clin Med 2022 Feb 06;11(3):856 [FREE Full text] [doi: 10.3390/jcm11030856] [Medline: 35160307]
48. Wasilewska E, Sobierajska-Rek A, Małgorzewicz S, Soliński M, Szalewska D, Jassem E. Is it possible to have home e-monitoring of pulmonary function in our patients with Duchenne muscular dystrophy in the COVID-19 pandemic?-a one center pilot study. Int J Environ Res Public Health 2021 Aug 26;18(17):8967 [FREE Full text] [doi: 10.3390/ijerph18178967] [Medline: 34501557]
49. Onofri A, Pavone M, De Santis S, Verrillo E, Caggiano S, Ullmann N, et al. Telemedicine in children with medical complexity on home ventilation during the COVID-19 pandemic. Pediatr Pulmonol 2021 Jun;56(6):1395-1400 [FREE Full text] [doi: 10.1002/ppul.25289] [Medline: 33524228]
50. Muñoz-Bonet JI, López-Prats JL, Flor-Macián EM, Cantavella T, Domínguez A, Vidal Y, et al. Medical complications in a telemedicine home care programme for paediatric ventilated patients. J Telemed Telecare 2020;26(7-8):462-473. [doi: 10.1177/1357633X19843761] [Medline: 31023136]
51. Muñoz-Bonet JI, López-Prats JL, Flor-Macián EM, Cantavella T, Bonet L, Domínguez A, et al. Usefulness of telemedicine for home ventilator-dependent children. J Telemed Telecare 2018 Dec 11;26(4):207-215 [FREE Full text] [doi: 10.1177/1357633x18811751]
52. Birnkrant DJ, Ararat E, Mhanna MJ. Cardiac phenotype determines survival in Duchenne muscular dystrophy. Pediatr Pulmonol 2015 Jun 10;51(1):70-76. [doi: 10.1002/ppul.23215]
53. Paramsothy P, Wang Y, Cai B, Conway KM, Johnson NE, Pandya S, et al. Selected clinical and demographic factors and all-cause mortality among individuals with Duchenne muscular dystrophy in the Muscular Dystrophy Surveillance, Tracking, and Research Network. Neuromuscul Disord 2022 Jun;32(6):468-476 [FREE Full text] [doi: 10.1016/j.nmd.2022.04.008] [Medline: 35597713]

XSL•FO
RenderX

54. Spurney C, Shimizu R, Morgenroth LP, Kolski H, Gordish-Dressman H, Clemens PR, CINRG Investigators. Cooperative International Neuromuscular Research Group Duchenne Natural History Study demonstrates insufficient diagnosis and treatment of cardiomyopathy in Duchenne muscular dystrophy. Muscle Nerve 2014 Aug;50(2):250-256 [FREE Full text] [doi: 10.1002/mus.24163] [Medline: 24395289]
55. Grant B, Wallace JG, Hobson RA, Craig BG, Mulholland HC, Casey FA. Telemedicine applications for the regional paediatric cardiology service in Northern Ireland. J Telemed Telecare 2002 Jan 01;8(2_suppl):31-33. [doi: 10.1177/1357633x020080s214]
56. Imberti JF, Tosetti A, Mei DA, Maisano A, Boriani G. Remote monitoring and telemedicine in heart failure: implementation and benefits. Curr Cardiol Rep 2021 May 07;23(6):55 [FREE Full text] [doi: 10.1007/s11886-021-01487-2] [Medline: 33959819]
57. Freedman LS, Commins JM, Moler JE, Willett W, Tinker LF, Subar AF, et al. Pooled results from 5 validation studies of dietary self-report instruments using recovery biomarkers for potassium and sodium intake. Am J Epidemiol 2015 Apr 01;181(7):473-487 [FREE Full text] [doi: 10.1093/aje/kwu325] [Medline: 25787264]
58. Wills AM, Garry J, Hubbard J, Mezoian T, Breen CT, Ortiz-Miller C, et al. Nutritional counseling with or without mobile health technology: a randomized open-label standard-of-care-controlled trial in ALS. BMC Neurol 2019 May 29;19(1):104 [FREE Full text] [doi: 10.1186/s12883-019-1330-6] [Medline: 31142272]
59. Sadasivan A, Warrier M, Vengalil S, Thomas P, Nalini A. Telehealth services for an adolescent with Duchenne muscular dystrophy (DMD) during the COVID-19 pandemic. Australian Social Work 2024 May 06;78(1):108-114. [doi: 10.1080/0312407X.2024.2340469]
60. McPherson AC, McAdam L, Keenan S, Schwellnus H, Biddiss E, DeFinney A, et al. A feasibility study using solution-focused coaching for health promotion in children and young people with Duchenne muscular dystrophy. Dev Neurorehabil 2018 Feb;21(2):121-130. [doi: 10.1080/17518423.2017.1289271] [Medline: 28272972]
61. Lee A, Finnin K, Holdsworth L, Milette D, Peterson C. Report of the WCPT/INPTRA digital physical therapy practice task force. World Confederation for Physical Therapy and International Network of Physiotherapy Regulatory Authorities. URL: https://world.physio/sites/default/files/2020-06/WCPT-INPTRA-Digital-Physical-Therapy-Practice-Task-force-March2020.pdf [accessed 2025-01-10]
62. Sobierajska-Rek A, Mański, Jabłońska-Brudło J, Śledzińska K, Wasilewska E, Szalewska D. Respiratory telerehabilitation of boys and young men with Duchenne muscular dystrophy in the COVID-19 pandemic. Int J Environ Res Public Health 2021 Jun 08;18(12):6179 [FREE Full text] [doi: 10.3390/ijerph18126179] [Medline: 34200985]
63. Sobierajska-Rek A, Mański, Jabłońska-Brudło J, Śledzińska K, Ucińska A, Wierzba J. Establishing a telerehabilitation program for patients with Duchenne muscular dystrophy in the COVID-19 pandemic. Wien Klin Wochenschr 2021 Apr;133(7-8):344-350 [FREE Full text] [doi: 10.1007/s00508-020-01786-8] [Medline: 33346889]
64. Kenis-Coskun O, Imamoglu S, Karamancioglu B, Kurt K, Ozturk G, Karadag-Saygi E. Comparison of telerehabilitation versus home-based video exercise in patients with Duchenne muscular dystrophy: a single-blind randomized study. Acta Neurol Belg 2022 Oct;122(5):1269-1280 [FREE Full text] [doi: 10.1007/s13760-022-01975-4] [Medline: 35616780]
65. Peretti A, Amenta F, Tayebati SK, Nittari G, Mahdi SS. Telerehabilitation: review of the state-of-the-art and areas of application. JMIR Rehabil Assist Technol 2017 Jul 21;4(2):e7 [FREE Full text] [doi: 10.2196/rehab.7511] [Medline: 28733271]
66. Arteaga D, Donnelly T, Crum K, Markham L, Killian M, Burnette WB, et al. Assessing physical activity using accelerometers in youth with Duchenne muscular dystrophy. JND 2020 Jun 02;7(3):331-342. [doi: 10.3233/jnd-200478]
67. Bradford N, Herbert A, Walker R, Pedersen L, Hallahan A, Irving H, et al. Home telemedicine for paediatric palliative care. Stud Health Technol Inform 2010;161:10-19. [Medline: 21191154]
68. Monterosso L, Kristjanson LJ, Aoun S, Phillips MB. Supportive and palliative care needs of families of children with life-threatening illnesses in Western Australia: evidence to guide the development of a palliative care service. Palliat Med 2007 Dec;21(8):689-696. [doi: 10.1177/0269216307083032] [Medline: 18073255]
69. Bradford NK, Armfield NR, Young J, Smith AC. Paediatric palliative care by video consultation at home: a cost minimisation analysis. BMC Health Serv Res 2014 Jul 28;14(1):328 [FREE Full text] [doi: 10.1186/1472-6963-14-328] [Medline: 25069399]
70. Holmen H, Riiser K, Winger A. Home-based pediatric palliative care and electronic health: systematic mixed methods review. J Med Internet Res 2020 Feb 28;22(2):e16248 [FREE Full text] [doi: 10.2196/16248] [Medline: 32130127]
71. Bradford NK, Young J, Armfield NR, Herbert A, Smith AC. Home telehealth and paediatric palliative care: clinician perceptions of what is stopping us? BMC Palliat Care 2014 Jun 16;13(1):29. [doi: 10.1186/1472-684x-13-29]
72. Bird M, Li L, Ouellette C, Hopkins K, McGillion MH, Carter N. Use of synchronous digital health technologies for the care of children with special health care needs and their families: scoping review. JMIR Pediatr Parent 2019 Nov 21;2(2):e15106 [FREE Full text] [doi: 10.2196/15106] [Medline: 31750840]
73. Weaver MS, Neumann ML, Navaneethan H, Robinson JE, Hinds PS. Human touch via touchscreen: rural nurses' experiential perspectives on telehealth use in pediatric hospice care. J Pain Symptom Manage 2020 Nov;60(5):1027-1033 [FREE Full text] [doi: 10.1016/j.jpainsymman.2020.06.003] [Medline: 32525081]

XSL•FO
RenderX

74. Duursma F, Schers HJ, Vissers KC, Hasselaar J. Study protocol: optimization of complex palliative care at home via telemedicine. A cluster randomized controlled trial. BMC Palliat Care 2011 Aug 9;10(1):13. [doi: 10.1186/1472-684x-10-13]
75. Liang HY, Hann Lin L, Yu Chang C, Mei Wu F, Yu S. Effectiveness of a nurse-led tele-homecare program for patients with multiple chronic illnesses and a high risk for readmission: a randomized controlled trial. J Nurs Scholarsh 2021 Mar;53(2):161-170. [doi: 10.1111/jnu.12622] [Medline: 33507626]
76. Bulto LN. The role of nurse‐led telehealth interventions in bridging healthcare gaps and expanding access. Nursing Open 2024 Jan 11;11(1):e2092-e2092. [doi: 10.1002/nop2.2092]

Abbreviations

DMD: Duchenne muscular dystrophy
EURO-NMD: European Reference Network for the thematic grouping of rare Neuromuscular Diseases
SiNQ-5: Sleep-Disordered Breathing Questionnaire In Neuromuscular Patients
TelePALS: Telemedicine for People with Amyotrophic Lateral Sclerosis
WHO: World Health Organization

Edited by A Mavragani; submitted 18.05.25; peer-reviewed by BF Birnbaum, Y Ishikawa; comments to author 07.07.25; revised version received 14.07.25; accepted 16.07.25; published 19.09.25.

Please cite as:
Wasilewska E, Wasilewski A, Onofri A, Wasilewski J, Sabiniewicz-Ziajka D, Sobierajska-Rek A, Meyer-Szary J, Śledzińska K, Wierzba J, Niedoszytko M, Małgorzewicz S
Exploring the Role of Telemedicine in Duchenne Muscular Dystrophy: Benefits and Challenges
JMIR Form Res 2025;9:e77698
URL: https://formative.jmir.org/2025/1/e77698
doi:10.2196/77698
PMID:

©Eliza Wasilewska, Andrzej Wasilewski, Alessandro Onofri, Jan Wasilewski, Dominika Sabiniewicz-Ziajka, Agnieszka Sobierajska-Rek, Jarosław Meyer-Szary, Karolina Śledzińska, Jolanta Wierzba, Marek Niedoszytko, Sylwia Małgorzewicz. Originally published in JMIR Formative Research (https://formative.jmir.org), 19.09.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Surveillance of Twitter Data on COVID-19 Symptoms During the Omicron Variant Period: A Sentiment Analysis

Kaiyue Zhang[1,2*], BS, MPH; Zhaojin Guo[3*], BS, MPH; Yujie Ai[1,3,4,5], BEng, MSc; An-Ran Li[1], BS, MPH; Anlin Li[3,6], BS, MSc; Ziyu Liu[3], MSc; Yittie Yi Ting Tse[1,5,7], MSc, MPH; Xinyu Zhou[1], BS, MPH; Taoran Liu[3], PhD; Chuxi Xiong[8], BSc; Jian Huang[9,10,11], PhD; Wai-kit Ming[3,12], PhD

[1]School of Public Health, Li Ka Shing Faculty of Medicine, University of Hong Kong, Hong Kong, China (Hong Kong)

[10]Bioinformatics Institute (BII), ASTAR, Singapore, Singapore

[11]Human Potential Translational Research Programme, National University of Singapore, Singapore, Singapore

[12]Institute of Global Governance and Innovation for a Shared Future, City University of Hong Kong, Hong Kong, China (Hong Kong)

[2]Department of Paediatrics and Adolescent Medicine, School of Clinical Medicine, University of Hong Kong, Hong Kong, China (Hong Kong)

[3]Department of Infectious Diseases and Public Health, Jockey Club College of Veterinary Medicine and Life Sciences, City University of Hong Kong, To Yuen Building, 31 To Yuen Street, Hong Kong, China (Hong Kong)

[4]Department of Family Medicine and Primary Care, Li Ka Shing Faculty of Medicine, School of Clinical Medicine, University of Hong Kong, Hong Kong, China (Hong Kong)

[5]Department of Pharmacology and Pharmacy, Li Ka Shing Faculty of Medicine, University of Hong Kong, Hong Kong, China (Hong Kong)

[6]Department of Medical Record, Peking University Third Hospital, Beijing, China

[7]Department of Health, The Government of the Hong Kong Special Administrative Region, Hong Kong, China (Hong Kong)

[8]Department of Science, University of Waterloo, Waterloo, ON, Canada

[9]Institute for Human Development and Potential (IHDP), ASTAR, Singapore, Singapore

[*]these authors contributed equally

Corresponding Author:
Wai-kit Ming, PhD
Department of Infectious Diseases and Public Health, Jockey Club College of Veterinary Medicine and Life Sciences, City University of Hong Kong, To Yuen Building, 31 To Yuen Street, Hong Kong, China (Hong Kong)

Abstract

Background: The global outbreak of COVID-19 has significantly impacted health care systems and has necessitated timely access to information for effective decision-making by health care authorities. Conventional methods for collecting patient data and analyzing virus mutations are resource-intensive. In the current era of rapid internet development, information on COVID-19 infections could be collected by a novel approach that leverages social media, particularly Twitter (subsequently rebranded X).

Objective: The aim of this study was to analyze the trending patterns of tweets containing information about various COVID-19 symptoms, explore their synchronization and correlation with conventional monitoring data, and provide insights into the evolution of the virus. We categorized tweet sentiments to understand the predictive power of negative emotions of different symptoms in anticipating the emergence of new Omicron subvariants and offering real-time assistance to affected individuals.

Methods: Relevant user tweets from 2022 containing information about COVID-19 symptoms were extracted from Twitter. Our fine-tuned RoBERTa model for sentiment analysis, achieving 99.7% accuracy for sentiment analysis, was used to categorize tweets as negative, positive, or neutral. Joinpoint regression analysis was used to examine the trends in weekly negative tweets related to COVID-19 symptoms, aligning these trends with the transition periods of SARS-CoV-2 Omicron subvariants from 2022. Real-time Twitter users with negative sentiments were geographically plotted. A total of 105,934 tweets related to fever, 120,257 to cough, 55,790 to headache, 101,220 to sore throat, 3410 to vomiting, and 5913 to diarrhea were collected.

Results: The most prominent topics of discussion were fever, sore throat, and headache. The weekly average daily tweets exhibited different fluctuation patterns in different stages of subvariants. Specifically, fever-related negative tweets were more sensitive to Omicron subvariant evolution, while discussions of other symptoms declined and stabilized following the emergence of the BA.2 variant. Negative discussions about fever rose to nearly 40% at the beginning of 2022 and showed 2 distinct peaks during the absolute dominance of BA.2 and BA.5, respectively. Headache and throat-related negative sentiment exhibited the highest levels among the analyzed symptoms. Tweets containing geographic information accounted for 1.5% (1351/391,508) of all collected data, with negative sentiment users making up 0.35% (5873/391,508) of all related tweets.

Conclusions: This study underscores the potential of using social media, particularly tweet trends, for real-time analysis of COVID-19 infections and has demonstrated correlations with major symptoms. The degree of negative emotions expressed in

XSL•FO
RenderX

tweets is valuable in predicting the emergence of new Omicron subvariants of COVID-19 and facilitating the provision of timely assistance to affected individuals.

(JMIR Form Res 2025;9:e66237) doi:10.2196/66237

KEYWORDS

COVID-19; social media surveillance; Twitter; natural language processing; real-time analysis; sentiment analysis; virus mutations; geographic tracking

Introduction

Symptom surveillance is crucial for public health and monitoring the transmission dynamics of infectious diseases [1]. Surveillance involves indicator-based and event-based surveillance [2], of which the former uses established protocols to collect structured data based on correlations between cases and hospital diagnostic data. Currently, the identification and variation of virus symptoms are studied in large populations in hospitals. However, delays in testing and reporting among institutions impact the timeliness of decision-making, hinder the prompt mobilization of resources and public education after determining the symptoms of mutated viruses, and potentially lead to losses and public panic [3].

Event-based surveillance can be performed using nonconventional data sources, such as social media, news, and public health networks, and excels in the timely identification of emerging public health issues. There is a trend of growing reliance on social media and search engines for pandemic-related data collection. Health systems are integrating conventional and advanced techniques to obtain more immediate and comprehensive data, which enhances the formulation of health policy.

Since the H1N1 influenza epidemic in 2009, social media has started to play a pivotal role in surveillance. For instance, HealthMap combines formal sources such as the World Health Organization (WHO) and the Centers for Disease Control and Prevention (CDC) with informal data sources such as Twitter to reduce reporting delays [4]. A study on the 2010 cholera outbreak in Haiti assessed the correlation between unofficial data sources (eg, HealthMap and Twitter) and government-reported cases. The effective reproduction number was also estimated using these nonofficial sources [5]. Twitter has also been applied to monitor dengue fever in analyses focusing on various aspects of tweets to identify potential cases [6], especially sociodemographic characteristics [7].

Social media surveillance has continued to progress for other infectious diseases, including Ebola [8], Middle East respiratory syndrome [9], and Zika [10]. Retrospective prediction models based on Google and Twitter data have demonstrated potential in the early detection of suspected Zika cases [10]. In the era of the COVID-19 pandemic, research has shifted focus toward understanding risk factors for the disease [11], its symptoms and mental health impacts [12], and using sentiment analysis to combat the spread of misinformation [13]. Mackey et al [14] used machine learning to identify COVID-19 symptoms and related experiences from Twitter, highlighting testing challenges and underreporting issues. A recurrent neural network has also been used to classify sentiment and analyze manifestations [15].

Wu et al [16] demonstrated the dynamic changes of COVID-19 symptoms over time, from the early typical respiratory symptoms to the later musculoskeletal and nervous system symptoms. We propose a Twitter-based symptom surveillance method for COVID-19 infections focusing on 6 typical symptoms. We chose Twitter (subsequently rebranded as X) as a data source to collect public information on people with COVID-19 from 2022. A previous study collected a large number of tweets related to the Omicron variant and classified them based on sentiment, language, source, and type of tweet, among other aspects [17]. We also analyzed the sentiment of their tweets to determine the severity of the impact for each individual. Our model can detect the emotions of users with COVID-19 on a geographic map in real time. This method monitors the dynamic epidemiological situation, especially the evolution of symptoms of Omicron subvariants, and explores correlations with public sentiment. The findings could assist in making timely governmental decisions for strengthening pandemic prevention and control.

Methods

Overview

This study used quantitative and qualitative methods for collecting and analyzing tweets related to COVID-19 to synchronize, forecast, and track the evolution of the different COVID-19 symptoms. The focus was on symptoms, the emotions of those who have been infected, and geographical information. We identified tweets related to COVID-19 symptoms as our key search criterion and collected tweets containing any reference to the occurrence of the disease and its symptoms from January 1, 2022, to December 31, 2022. The data were prepared in 3 steps: (1) data selection, (2) data collection, and (3) data cleaning.

Data Selection

We collected tweets related to COVID-19 symptoms using the following key search terms: “fever,” “cough,” “headache,” “throat,” and “vomit,” which are related to common symptoms of COVID-19 according to the CDC in October 2022 [18]. We collected data if any of these terms and the term “COVID” appeared in a published tweet. We chose daily Twitter data from 2022 because of its higher maturity in test technologies and monitoring systems, which supported the meaningful comparison between Twitter data and clinical data.

XSL•FO
RenderX

Data Collection

During the data collection phase, daily tweets in English were gathered from Twitter using Python scripts. The advanced search function on Twitter was used to extract comprehensive datasets, followed by a data-validation process to maintain collection accuracy. Additional data crawling was conducted to reinforce the integrity of the dataset. Official numbers of weekly confirmed cases were reported by the WHO [19]. We used 3 continuous modules to analyze COVID-19 symptoms for the real-time surveillance of people infected with COVID-19: temporal analysis, sentiment analysis, and user-feature analysis.

Sentiment Analysis

After data collection, we preprocessed the data of each tweet using Python (Python Software Foundation) and R (The R Foundation), which filtered out tweets with significant amounts of garbled characters before inputting them into the sentiment analysis model. The sentiments expressed in short tweets are generally more direct, which makes it easier to determine the sentiment category. To accurately determine their sentiment, both training and inference were conducted on all tweets using the latest natural language processing (NLP) model, the RoBERTa base model [20], in the Hugging Face repository.

We identified the tweets of users who were infected with COVID-19, predicted the probability of different types of emotions in each tweet (positive, neutral, and negative), and selected the category with the highest probability as the sentiment result. We used Omicron subvariants data from the weekly epidemiology reports of the WHO to create a geographical map of Omicron subvariants' proportions over 1 year. The WHO data are obtained through geographical coverage and align with our worldwide Twitter data. The accuracy of sentiment analysis for the 3 categories was 99.7%.

We performed a joinpoint regression to analyze changes in counts of weekly negative tweets. The subvariant transition periods are set as the week during which a new Omicron subvariant replaces the previous subvariant, plus 1 week before and after. We aim to identify whether the significant joinpoints captured by the joinpoint regression fall within these subvariant transition periods.

Geographical Analysis

In the completed map, each red dot represents a tweet related to COVID-19 symptoms. Geographic information about Twitter users is obtained by using the Twitter API to see what geographic information exists, or by analyzing metadata or tweets to infer geographic information. The map provides a quick display of the daily count of negative tweets and the tally of tweets concerning COVID-19 in a specific region while summarizing the negative, positive, and neutral sentiments. A line graph shows the fluctuations in the quantity of COVID-19–related tweets regionally over time.

Ethical Considerations

This study did not involve human participants or interventions, and thus no informed consent was required. The analysis relied solely on publicly available Twitter data, ensuring the anonymity of individuals. Any personal information collected was handled in accordance with ethical guidelines, and no identifiable user data are disclosed in this paper.

Results

Real-Time Analysis of Tweets Related to COVID-19 Symptoms From 2022

The collected data encompassed 6 symptoms mentioned in tweets, of which 105,934 tweets were related to fever, 120,257 were related to cough, 55,790 were related to headache, 101,220 were related to sore throat, 3410 were related to vomiting, and 5913 were related to diarrhea (Figure 1).

XSL•FO
RenderX

Figure 1. A flowchart of screening for COVID-19 symptom–related tweets. NLP: natural language processing; WHO: World Health Organization.

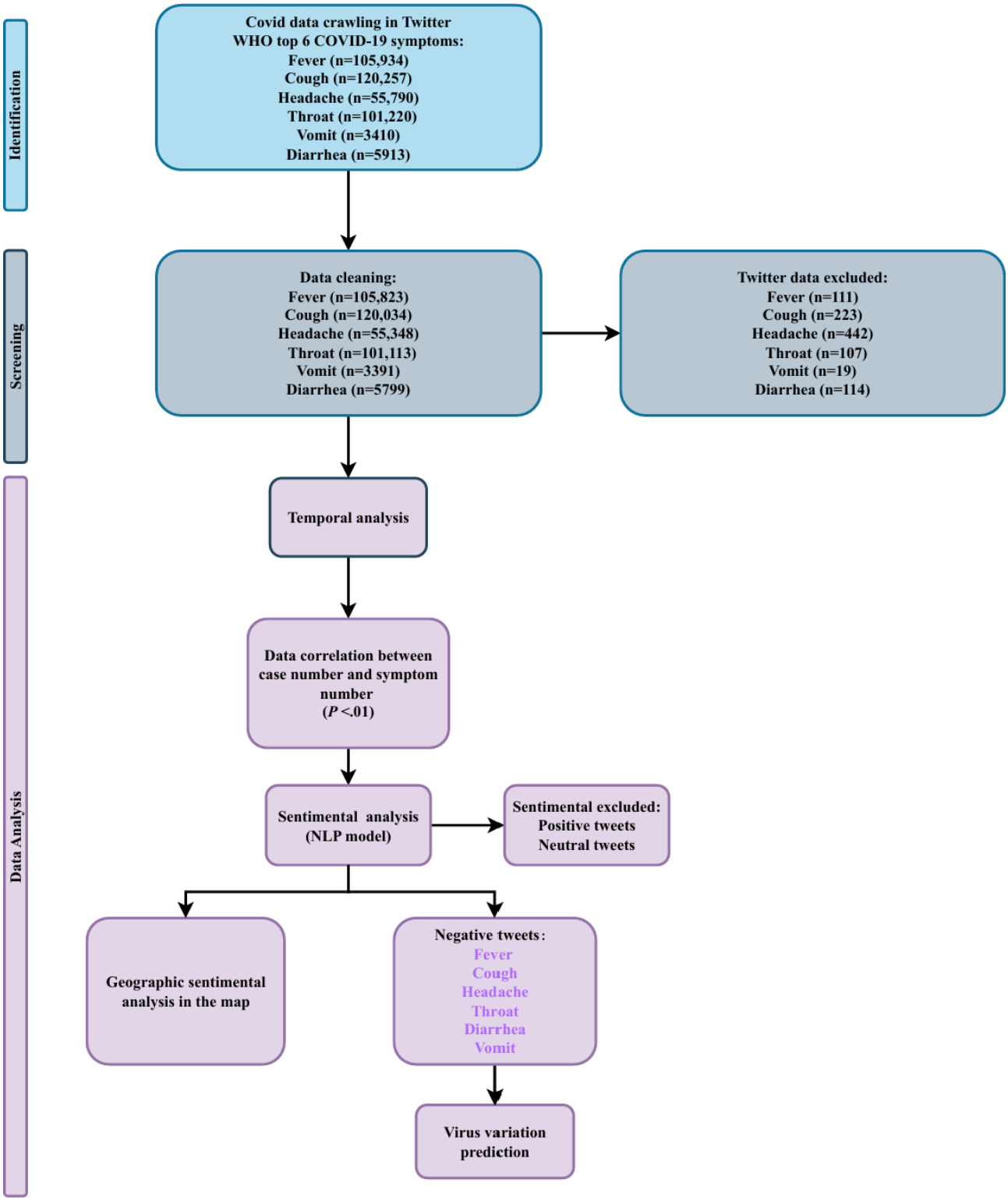

Figure 2 shows the trend of the weekly ratio of symptom-related tweets and new cases of COVID-19 worldwide. We divided the year 2022 into 4 periods based on the prevalence of different variants. There were 4 peaks in the number of new cases throughout the year, which occurred in January, March, July, and December. Furthermore, the fluctuations in ratios of weekly symptom tweets demonstrated a certain level of synchronous pattern with the evolution of Omicron variants and waves of new cases.

XSL•FO
RenderX

Figure 2. Weekly average severity of daily tweets about symptoms and confirmed cases.

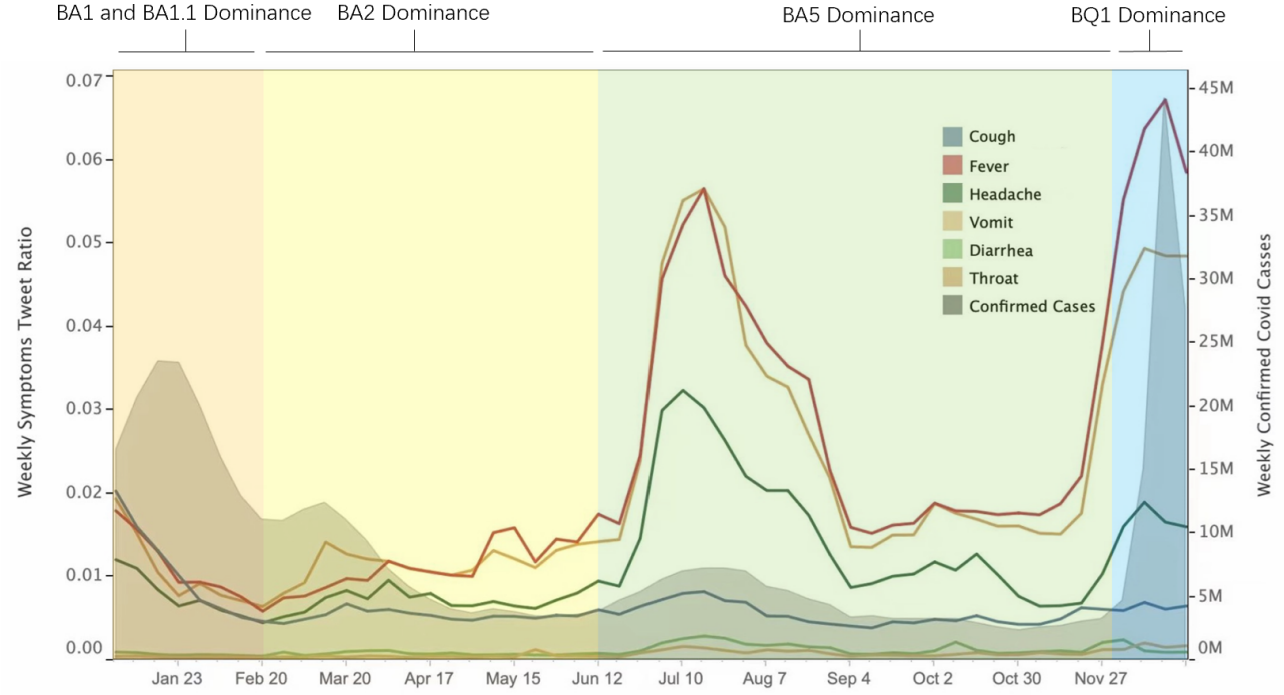

According to the gathered symptom surveillance data, discussions surrounding fever, sore throat, and headache were the most prominent, which was consistent with real-world observations. Fever was common from the beginning of the pandemic, whereas sore throat and headache became more prevalent during the Omicron period [21,22] and attracted increased attention from multiple public health authorities [23].

In Figure 2, the periods of analysis are as follows: (1) period 1 (January 1-February 20): dominance of BA.1 and BA.1.1, (2) period 2 (February 20-June 12): BA.2 dominance, (3) period 3 (June 12-December 1): BA.5 dominance, and (4) period 4 (December 1-12): BQ1 and XBB dominance. The weekly average of daily symptom severity is the calculated number of tweets for each symptom as a proportion of tweets for all symptoms. Gray shaded area: weekly confirmed cases from the WHO. To match the 2 values of the vertical axis (weekly average daily), we performed vertical axis matching to display the image.

The confirmed cases (Figure 2, gray shaded area) increased only in the first 2 weeks during period 1 and then continued to decrease because this period was in the later stages of the dominance of the BA.1 and BA.1.1 subvariants. Restrictions on travel and activities and vaccination measures were implemented as early as November-December 2021. Correspondingly, the tweets related to all symptoms demonstrated a decreasing trend with little difference among them.

During period 2, the BA.2 subvariant emerged as the dominant strain and superseded BA.1 and BA.1.1. Although the increased transmissibility of BA.2 [24] contributed to a second peak in confirmed cases, the risk of reinfection with BA.2 was lower [24], and the severity did not escalate compared to earlier periods. Consequently, there were some slight fluctuations in Twitter discussions. At the same time, tweets related to sore throat, fever, and headache began to dominate during this period.

By mid-2022, in period 3, the symptom surveillance data showed a peak that corresponds to a surge in new cases in the real world, particularly in terms of fever, throat discomfort, and headache. Tweets related to other symptoms also exhibited relatively elevated peaks compared to their baseline levels. During this period, the BA.5 subvariant swiftly supplanted BA.2 as the predominant strain. Notably, BA.5 was characterized by enhanced immunity and heightened transmissibility [25-28].

Some research indicates that individuals infected with the BA.5 subvariant are more prone to experiencing several common symptoms that we have selected [29], which reinforces the robustness of our analysis. Heightened public sensitivity on social media outpaced the actual occurrence of the disease and may have been influenced by sore throats becoming a primary symptom and growing awareness of the Omicron subvariants. The dynamic interplay between subvariant characteristics and public discourse is evident.

In period 4, the BQ1 subvariant dominated, which was accompanied by the emergence of XBB. Research indicates that BQ.1 and XBB have higher effective reproduction numbers [30] and transmissibility compared to BA.2 and BA.5 [31]. A peak of newly reported cases in December 2022 occurred in the week of December 5-11 and reached its highest point in the week of December 19-25. The Twitter data promptly revealed this conspicuous peak, and the top 3 most discussed symptoms were more sensitive.

Sentiment Analysis of Tweets Related to COVID-19 Symptoms From 2022

Among the 8 subvariants, BA.1, BA.1.1, BA.2, BA.5, and BQ.1 notably dominated at various points during 2022. At the beginning of the year, BA.1 had already started to be gradually

XSL•FO
RenderX

replaced by BA.1.1, but BA.2 quickly emerged and maintained a proportion greater than 50% for an extended period of 16 weeks. Subsequently, BA.5 began to replace BA.2 as the dominant subvariant, sustaining a proportion greater than 50% for an even longer duration of 20 weeks. By the end of the year, BQ.1 surpassed BA.5, while XBB also emerged.

Negative discussions about fever rose to nearly 40% at the beginning of the year and subsequently exhibited 2 distinct peaks during the periods when BA.2 and BA.5 were absolutely dominant, respectively. The prompt public response to the symptom of fever during the early stages of subvariant evolution may be attributed to its role as the first symptom to manifest after infection with COVID-19 [32].

One characteristic of the Omicron subvariants is that the majority of symptoms closely resemble those of the cold, including headaches and sore throat [33]. This is illustrated in Figure 3, where these 2 symptoms exhibit the highest levels of negative sentiment.

Figure 3. Specific symptom sentiment analysis with proportions of Omicron subvariants. The weekly negative rate is the calculated number of negative tweets for each symptom as a proportion of all tweets for the specific symptom. The evolution of the proportions of Omicron subvariants was determined based on the data on Omicron subvariant proportions from World Health Organization weekly reports.

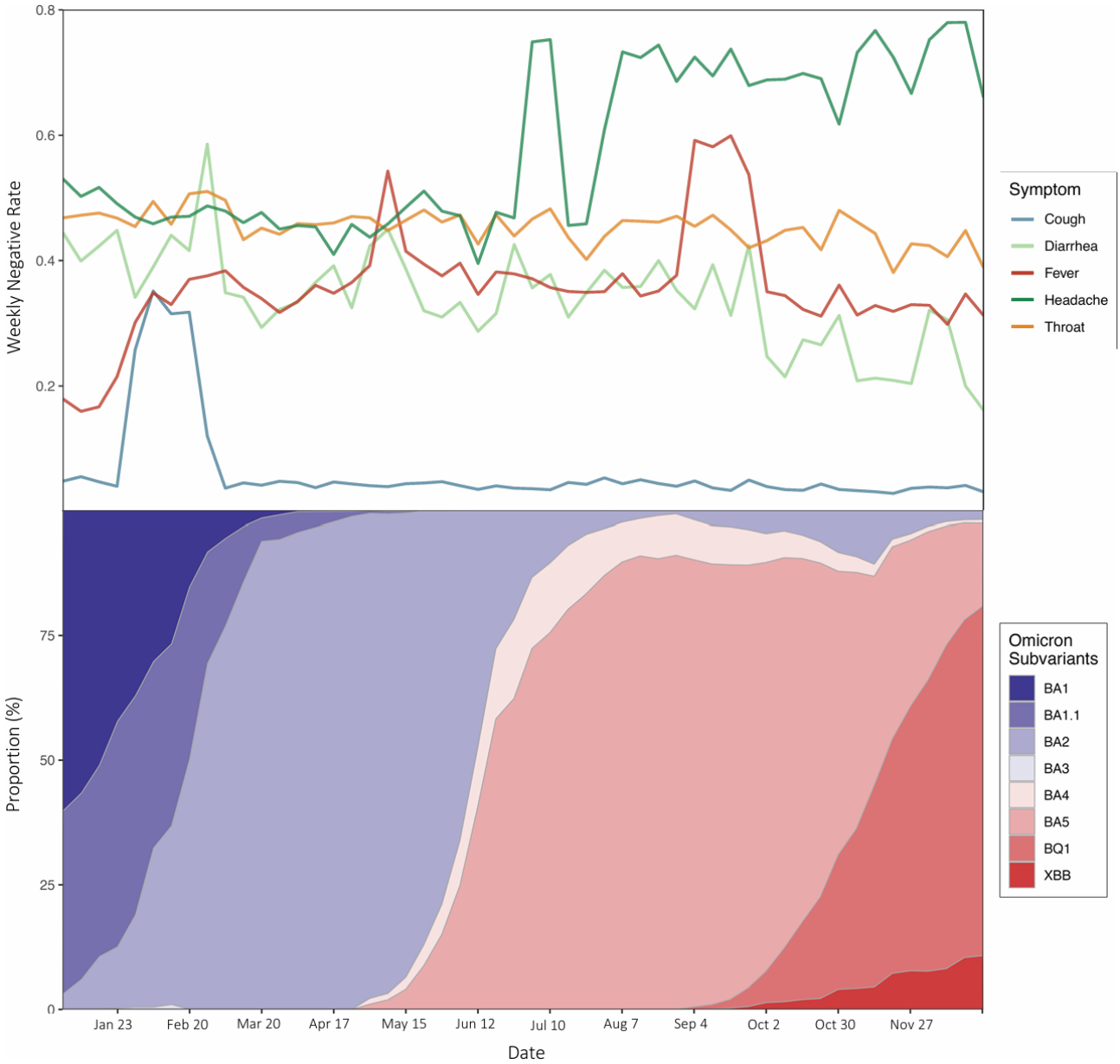

Joinpoint Regression of the Number of Negative Tweets of COVID-19 Symptoms From 2022

Cough exhibited a transient rise in negative sentiment during only the initial stages of BA.2, while negative attitudes in this regard remained consistently low and stable during other periods (Figure 4). Given its persistent presence as a symptom of COVID-19 throughout the pandemic, the public might have developed a relatively comprehensive understanding of both the symptom itself and its associated management strategies. This symptom typically manifests in the middle stages of the infection [32], and its prevalence had already begun to decline with the Omicron variant [33].

XSL•FO
RenderX

Figure 4. Joinpoint regression analysis of weekly negative tweets: alignment with subvariant transition periods. Significant joinpoints, which indicate a significant change in the trend of negative tweet counts, are highlighted with the specific week in black squares. The background's shaded regions represent the transition periods of major SARS-CoV-2 Omicron subvariants throughout 2022.

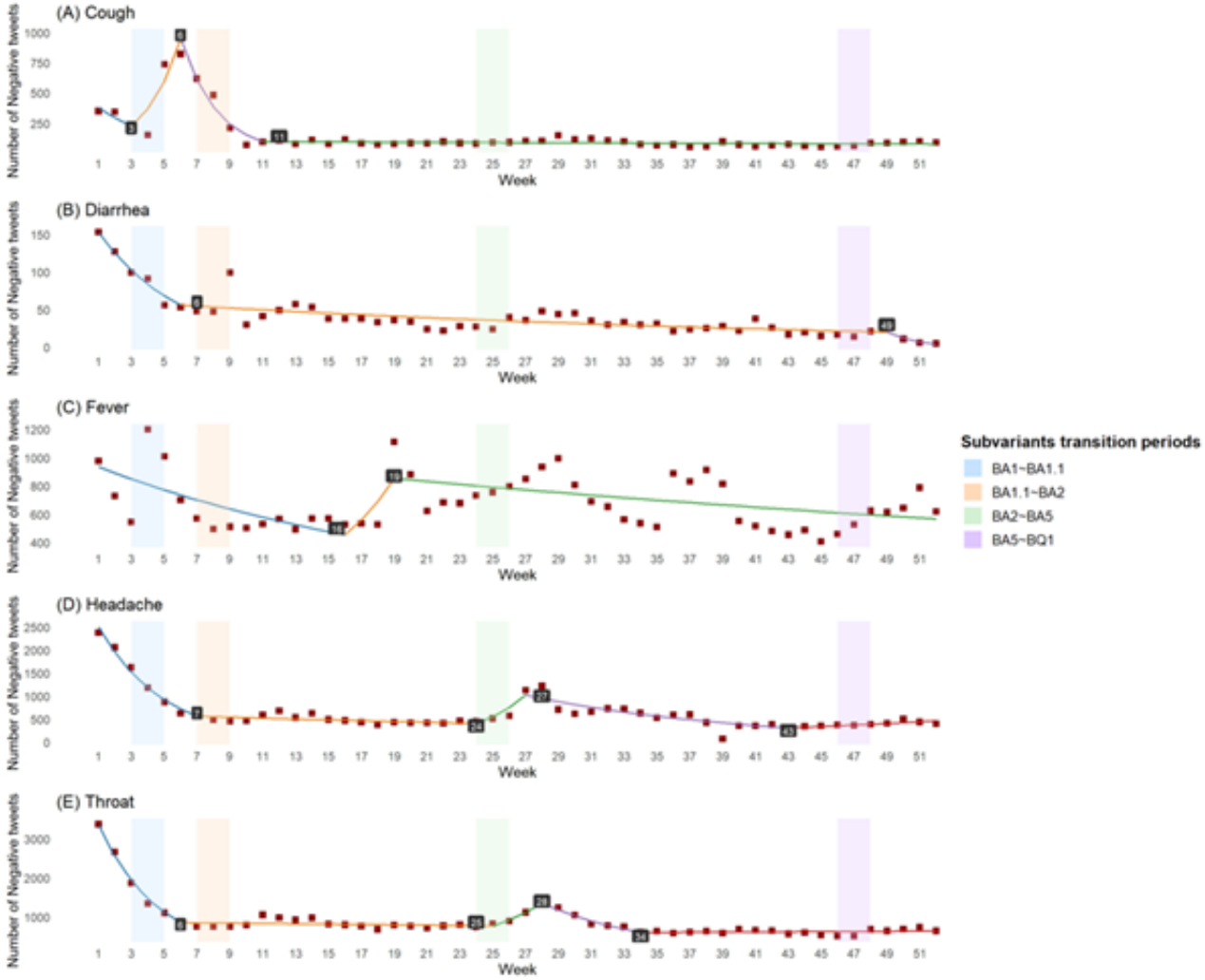

The joinpoint analysis revealed similar trends for diarrhea, headache, and throat-related discussions (Figure 4). These discussions declined as the proportion of BA.1 decreased and stabilized before BA.2 became dominant. However, only headache and throat-related negative discussions exhibited minimal fluctuations during the later phase of BA.2-BA.5.

Fever was able to capture more of the subvariants' evolution, with negative tweets exhibiting a peak in sentiment during the later stages of each subvariant transition (Figure 4). Model estimates of the joinpoint regression can be seen in Multimedia Appendix 1.

An Interactive Map of Opinion Analysis and Sentiment Visualization of COVID-19 Tweets

Social media platforms have become an essential channel for people to express their views and emotions about the epidemic. We developed an interactive map demo using the United States as an example to collect Twitter data, perform sentiment analysis through NLP, and link the results to corresponding geographic locations (Figure 5A). Of all the data we collected, tweets containing geographic information accounted for 1.5% (1351/391,508), while negative sentiment users accounted for 0.35% (5873/391,508) of all related tweets. This demo link is available at Tableau Public [34]. This system provides a visual representation that provides an intuitive understanding of attitudes and emotions toward COVID-19 across different regions in the United States. Users can click on a specific location on the map to obtain detailed information about emotions from comments about COVID-19 (Figure 5B).

XSL•FO
RenderX

Figure 5. (A) An interactive map of opinion analysis and sentiment visualization of COVID-19 tweets showing a sample of users with negative tweets. (B) Users can click on a specific location on the map to obtain detailed information about emotions from comments about COVID-19.

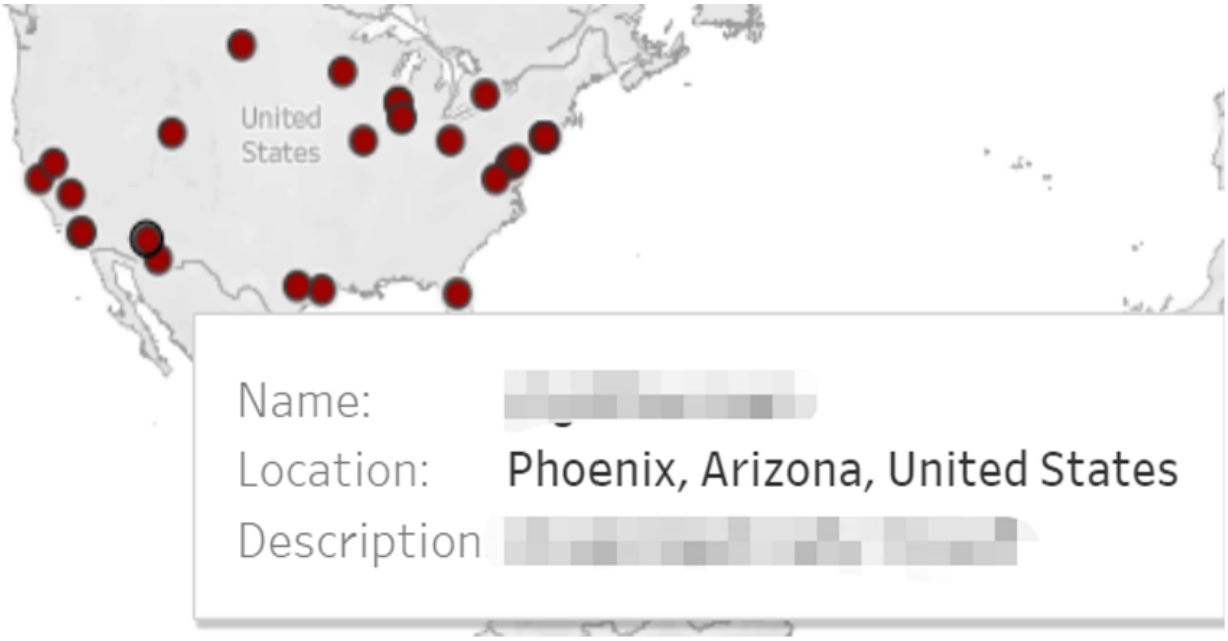

(A)

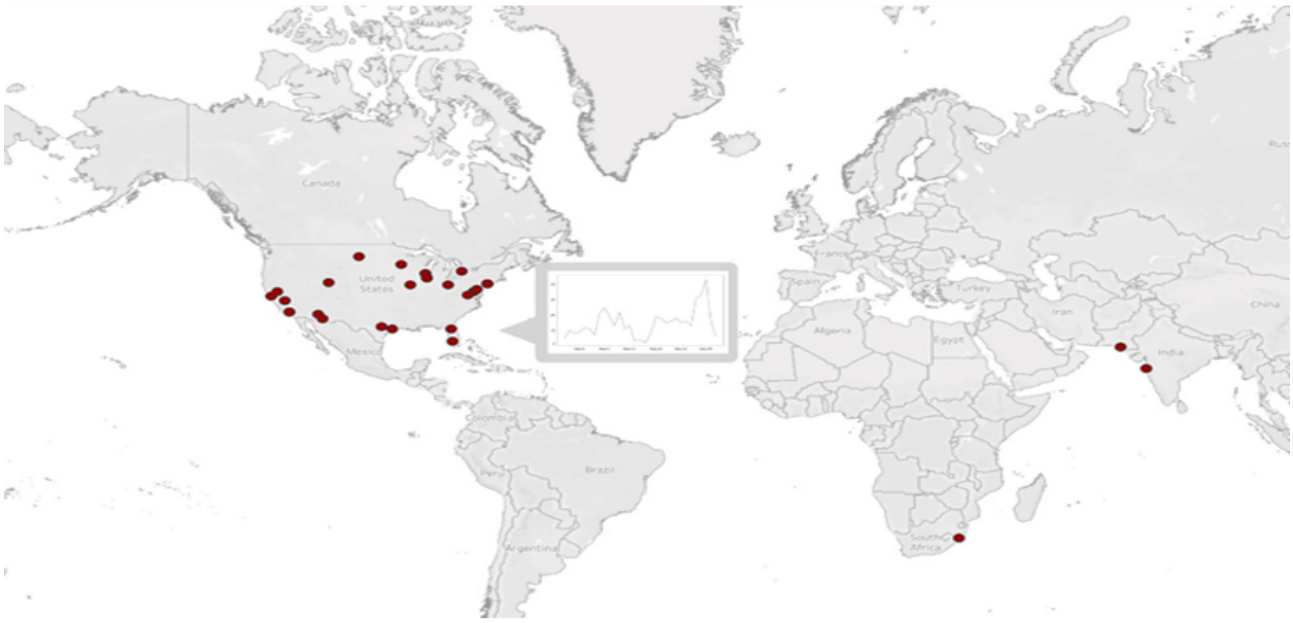

(B)

The interactive map could deepen the understanding of people's opinions and emotions and provide valuable national information for decision-making, opinion management, and social intervention. In decision-making and public opinion management, analyzing the distribution of different opinions and emotional tendencies in the United States could help decision-makers better understand attitudes and opinions about COVID-19 in different places. This could help to formulate more effective policies to meet the needs of the public and improve the acceptance of policies. In terms of social intervention, interactive maps of the United States could help the government better understand the distribution of views. As a result, state governments could implement more targeted interventions, such as providing psychological support, disseminating information, and mobilizing communities to address mental health and social issues related to COVID-19.

Discussion

Principal Findings

Our findings corroborate clinical observations that Omicron subvariants shifted symptom profiles toward cold-like

XSL•FO
RenderX

manifestations (eg, sore throat and headache), consistent with reports by Menni et al [21] and Schulze et al [22]. The heightened negative sentiment around headache and sore throat (Figure 3) likely reflects their unexpected prevalence during Omicron, contrasting with earlier variants. Notably, fever-related negativity proved uniquely sensitive to subvariant evolution, exhibiting dual peaks during BA.2 and BA.5 dominance. This suggests public social media reactivity may serve as a leading indicator for viral evolution—a novel extension of prior infodemiology studies [eg, 16,17]. However, the sustained high negative sentiment toward symptoms such as headache (Figure 3) requires nuanced interpretation. While headache prevalence increased during Omicron acute infection [21], its persistent discussion throughout 2022 may also reflect growing public awareness of Long COVID, where headache is a predominant symptom affecting most cases [35,36]. This "twin-demic" context, where acute infection waves coexist with Long COVID burden, introduces complexity in attributing symptom trends solely to variant dynamics.

The real-time geographic dashboard (Figure 5) operationalizes a key advantage of social media surveillance: overcoming delays in conventional reporting. For instance, our system detected the December 2022 case surge 1‑2 weeks before WHO confirmation [19], enabling proactive resource targeting. This addresses critical gaps identified in traditional systems, such as lags in resource allocation during BA.5's rapid spread [31]. Furthermore, sentiment-driven resource allocation (eg, directing mental health support to regions with rising negative tweets) offers a scalable complement to clinical surveys. Our method also expands on Mackey et al's [14] symptom detection framework by integrating sentiment as a proxy for symptom severity and subvariant impact, though future models should distinguish acute versus chronic symptom manifestations through keyword filtering (eg, "long-hauler" and "post-COVID") or duration-based classification.

Our study has 4 key limitations. First, the exclusive reliance on 2022 English-language tweets introduces demographic and linguistic biases, as Twitter users skew toward younger, urban, technologically adept populations. This limits generalizability to non-English speakers, older adults, and rural communities. Future work must integrate multilingual platforms (eg, Weibo and Facebook) and extend temporal coverage to strengthen model robustness. Second, clinical generalizability is constrained by the underrepresentation of pediatric populations. While children exhibit distinct Omicron symptom severity [10], our adult-dominated dataset cannot capture this nuance. Subpopulation-specific analyses require complementary clinical data.

Third, geographic sparsity remains a challenge: only 1.5% (1351/391,508) of tweets contained usable location data, and inferred geodata may introduce biases. Resource allocation based solely on these sparse signals risks misprioritization. Cross-validation with electronic health records and refined geolocation inference algorithms is essential. Fourth, global representativity is compromised by overlapping biases in Twitter data (overrepresenting Anglophone nations) and WHO variant surveillance (skewed toward high-income countries). The observed correlations between sentiment and subvariants may reflect region-specific epidemiological patterns rather than global dynamics. Future research must prioritize intentional sampling from underrepresented regions and multiplatform data fusion to validate cross-cultural applicability. In addition, our analysis did not explicitly account for Long COVID–related symptom reporting. Given that 2022 marked a peak in Long COVID cases (WHO COVID-19 Weekly Epidemiological Update, July 2022), symptom discussions, particularly for neurological manifestations like headache, likely conflated acute infection and chronic sequelae. This conflation may partially explain the sustained high negative sentiment for certain symptoms beyond variant transition periods (eg, headache negativity remained >30% during BA.5 decline in Figure 3).

Beyond COVID-19, this study establishes a scalable framework for infodemiology-driven pandemic response. The integration of real-time social media analytics with conventional surveillance creates a synergistic system capable of (1) accelerating outbreak detection (eg, leveraging symptom sentiment as an early-warning signal), (2) optimizing precision public health interventions (eg, dynamic resource allocation based on geotagged emotional hotspots), and (3) enabling rapid evaluation of policy impacts (eg, monitoring sentiment shifts postvaccine rollout). As novel pathogens emerge, such systems could mitigate the "infodemic" paradox, where data abundance coexists with decision-making delays, by converting public discourse into actionable intelligence. Future iterations will expand linguistic parameters, incorporate clinical correlates, and develop predictive algorithms to establish a global public health intelligence platform, ultimately bridging the gap between digital epidemiology and frontline response. The framework's adaptability is particularly valuable for diseases with long-term sequelae (eg, Long COVID), where real-time tracking of chronic symptoms through social discourse could accelerate therapeutic development and patient support.

Conclusions

In summary, our method allows rapid surveillance of changes in the geographical COVID-19 symptoms. This was accomplished using sentiment analysis of tweets from people who have been infected with COVID-19 to detect the geographical severity rate. At the individual level, it can detect the real-time severity of COVID-19 and could make it possible to provide medical and mental assistance. It also provides a new solution for coping with the urgent emergence of a new pandemic.

Acknowledgments

The authors thank the individuals who participated in data collection and validation in this paper. We used artificial intelligence tools only for language polishing and grammar checking.

XSL•FO
RenderX

Data Availability

The datasets and code generated or analyzed during the study are available from the corresponding author (WM) on reasonable request. The data will be shared for noncommercial purposes after being fully anonymized.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Model estimates of joinpoint regression.
[PDF File, 130 KB - formative_v9i1e66237_app1.pdf]

References

1. Chan EH, Brewer TF, Madoff LC, et al. Global capacity for emerging infectious disease detection. Proc Natl Acad Sci USA 2010 Dec 14;107(50):21701-21706. [doi: 10.1073/pnas.1006219107] [Medline: 21115835]
2. Velasco E, Agheneza T, Denecke K, Kirchner G, Eckmanns T. Social media and internet-based data in global systems for public health surveillance: a systematic review. Milbank Q 2014 Mar;92(1):7-33. [doi: 10.1111/1468-0009.12038] [Medline: 24597553]
3. Strengthening surveillance of and response to foodborne diseases: a practical manual. Stage 2: strengthening indicator-based surveillance. World Health Organization. 2017. URL: https://www.who.int/publications/i/item/9789241513258 [accessed 2025-08-30]
4. Brownstein JS, Freifeld CC, Chan EH, et al. Information technology and global surveillance of cases of 2009 H1N1 influenza. N Engl J Med 2010 May 6;362(18):1731-1735. [doi: 10.1056/NEJMsr1002707] [Medline: 20445186]
5. Chunara R, Andrews JR, Brownstein JS. Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak. Am J Trop Med Hyg 2012 Jan;86(1):39-45. [doi: 10.4269/ajtmh.2012.11-0597] [Medline: 22232449]
6. Gomide J, Veloso A, Meira W, et al. Dengue surveillance based on a computational model of spatio-temporal locality of Twitter. Presented at: WebSci '11; Jun 15-17, 2011; Koblenz, Germany URL: https://dl.acm.org/doi/proceedings/10.1145/2527031 [accessed 2025-07-23] [doi: 10.1145/2527031.2527049]
7. Nsoesie EO, Flor L, Hawkins J, et al. Social media as a sentinel for disease surveillance: what does sociodemographic status have to do with it? PLoS Curr 2016 Dec 7;8:ecurrents.outbreaks.cc09a42586e16dc7dd62813b7ee5d6b6. [doi: 10.1371/currents.outbreaks.cc09a42586e16dc7dd62813b7ee5d6b6] [Medline: 28123858]
8. Odlum M, Yoon S. What can we learn about the Ebola outbreak from tweets? Am J Infect Control 2015 Jun;43(6):563-571. [doi: 10.1016/j.ajic.2015.02.023] [Medline: 26042846]
9. Shin SY, Seo DW, An J, et al. High correlation of Middle East respiratory syndrome spread with Google search and Twitter trends in Korea. Sci Rep 2016 Sep 6;6(1):32920. [doi: 10.1038/srep32920] [Medline: 27595921]
10. McGough SF, Brownstein JS, Hawkins JB, Santillana M. Forecasting Zika incidence in the 2016 Latin America outbreak combining traditional disease surveillance with search, social media, and news report data. PLoS Negl Trop Dis 2017 Jan;11(1):e0005295. [doi: 10.1371/journal.pntd.0005295] [Medline: 28085877]
11. Jeon J, Baruah G, Sarabadani S, Palanica A. Identification of risk factors and symptoms of COVID-19: analysis of biomedical literature and social media data. J Med Internet Res 2020 Oct 2;22(10):e20509. [doi: 10.2196/20509] [Medline: 32936770]
12. Oyebode O, Ndulue C, Adib A, et al. Health, psychosocial, and social issues emanating from the COVID-19 pandemic based on social media comments: text mining and thematic analysis approach. JMIR Med Inform 2021 Apr 6;9(4):e22734. [doi: 10.2196/22734] [Medline: 33684052]
13. Medford RJ, Saleh SN, Sumarsono A, Perl TM, Lehmann CU. An "infodemic": leveraging high-volume Twitter data to understand early public sentiment for the coronavirus disease 2019 outbreak. Open Forum Infect Dis 2020 Jul;7(7):ofaa258. [doi: 10.1093/ofid/ofaa258] [Medline: 33117854]
14. Mackey T, Purushothaman V, Li J, et al. Machine learning to detect self-reporting of symptoms, testing access, and recovery associated with COVID-19 on Twitter: retrospective big data infoveillance study. JMIR Public Health Surveill 2020 Jun 8;6(2):e19509. [doi: 10.2196/19509] [Medline: 32490846]
15. Nemes L, Kiss A. Social media sentiment analysis based on COVID-19. J Inform Telecommun 2021 Jan 2;5(1):1-15. [doi: 10.1080/24751839.2020.1790793]
16. Wu J, Wang L, Hua Y, et al. Trend and co-occurrence network of COVID-19 symptoms from large-scale social media data: infoveillance study. J Med Internet Res 2023 Mar 14;25:e45419. [doi: 10.2196/45419] [Medline: 36812402]
17. Thakur N, Han CY. An exploratory study of tweets about the SARS-CoV-2 Omicron Variant: insights from sentiment analysis, language interpretation, source tracking, type classification, and embedded URL detection. COVID 2022;2(8):1026-1049. [doi: 10.3390/covid2080076]

XSL•FO
RenderX

18. Crawley AW, Divi N, Smolinski MS. Using timeliness metrics to track progress and identify gaps in disease surveillance. Health Secur 2021;19(3):309-317. [doi: 10.1089/hs.2020.0139] [Medline: 33891487]
19. Symptoms of COVID-19. Centers for Disease Control and Prevention. 2022 Oct 26. URL: https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html [accessed 2025-07-23]
20. RoBERTa base model. Hugging Face. URL: https://huggingface.co/FacebookAI/roberta-base [accessed 2025-07-23]
21. Menni C, Valdes AM, Polidori L, et al. Symptom prevalence, duration, and risk of hospital admission in individuals infected with SARS-CoV-2 during periods of omicron and delta variant dominance: a prospective observational study from the ZOE COVID study. Lancet 2022 Apr 23;399(10335):1618-1624. [doi: 10.1016/S0140-6736(22)00327-0] [Medline: 35397851]
22. Schulze H, Bayer W. Changes in symptoms experienced by SARS-CoV-2-infected individuals – from the first wave to the Omicron Variant. Front Virol 2022;2. [doi: 10.3389/fviro.2022.880707]
23. Iacobucci G. Covid-19: UK adds sore throat, headache, fatigue, and six other symptoms to official list. BMJ 2022 Apr 4;377:o892. [doi: 10.1136/bmj.o892] [Medline: 35379691]
24. Weekly epidemiological update on COVID-19. : World Health Organization; 2022 Jun 22 URL: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports [accessed 2025-07-23]
25. Tuekprakhon A, Nutalai R, Dijokaite-Guraliuc A, et al. Antibody escape of SARS-CoV-2 Omicron BA.4 and BA.5 from vaccine and BA.1 serum. Cell 2022 Jul 7;185(14):2422-2433. [doi: 10.1016/j.cell.2022.06.005] [Medline: 35772405]
26. Hachmann NP, Miller J, Collier ARY, et al. Neutralization escape by SARS-CoV-2 Omicron subvariants BA.2.12.1, BA.4, and BA.5. N Engl J Med 2022 Jul 7;387(1):86-88. [doi: 10.1056/NEJMc2206576] [Medline: 35731894]
27. Shrestha LB, Foster C, Rawlinson W, Tedla N, Bull RA. Evolution of the SARS-CoV-2 omicron variants BA.1 to BA.5: implications for immune escape and transmission. Rev Med Virol 2022 Sep;32(5):e2381. [doi: 10.1002/rmv.2381] [Medline: 35856385]
28. Kang SW, Park H, Kim JY, et al. Comparison of the clinical and virological characteristics of SARS-CoV-2 Omicron BA.1/BA.2 and omicron BA.5 variants: a prospective cohort study. J Infect 2023 May;86(5):e148-e151. [doi: 10.1016/j.jinf.2023.01.015] [Medline: 36669564]
29. Goller KV, Moritz J, Ziemann J, et al. Differences in clinical presentations of Omicron infections with the Lineages BA.2 and BA.5 in Mecklenburg-Western Pomerania, Germany, between April and July 2022. Viruses 2022 Sep 13;14(9):2033. [doi: 10.3390/v14092033] [Medline: 36146837]
30. Tamura T, Ito J, Uriu K, et al. Virological characteristics of the SARS-CoV-2 XBB variant derived from recombination of two Omicron subvariants. Nat Commun 2023 May 16;14(1):2800. [doi: 10.1038/s41467-023-38435-3] [Medline: 37193706]
31. Kimura I, Yamasoba D, Tamura T, et al. Virological characteristics of the SARS-CoV-2 Omicron BA.2 subvariants, including BA.4 and BA.5. Cell 2022 Oct 13;185(21):3992-4007. [doi: 10.1016/j.cell.2022.09.018] [Medline: 36198317]
32. Larsen JR, Martin MR, Martin JD, Kuhn P, Hicks JB. Modeling the onset of symptoms of COVID-19. Front Public Health 2020;8:473. [doi: 10.3389/fpubh.2020.00473] [Medline: 32903584]
33. Londoners with cold symptoms more likely to have Covid, says expert. The Guardian. 2021. URL: https://www.theguardian.com/world/2021/dec/15/londoners-with-cold-symptoms-more-likely-to-have-covid-expert-omicron [accessed 2025-08-30]
34. Location&description. Tableau Public. URL: https://public.tableau.com/app/profile/sia7698/viz/Locationdescription/Sheet8 [accessed 2025-09-02]
35. Tana C, Bentivegna E, Cho SJ, et al. Long COVID headache. J Headache Pain 2022 Aug 1;23(1):93. [doi: 10.1186/s10194-022-01450-8] [Medline: 35915417]
36. Gómez-Dabó L, Melgarejo-Martínez L, Caronna E, Pozo-Rosich P. Headache in COVID-19 and Long COVID: to know facts for clinical practice. Curr Neurol Neurosci Rep 2023 Oct;23(10):551-560. [doi: 10.1007/s11910-023-01296-w] [Medline: 37665495]

Abbreviations

CDC: Centers for Disease Control and Prevention
NLP: natural language processing
WHO: World Health Organization

XSL•FO
RenderX

Edited by A Mavragani; submitted 07.09.24; peer-reviewed by A Jarynowski; revised version received 17.06.25; accepted 27.06.25; published 23.09.25.

Please cite as:
Zhang K, Guo Z, Ai Y, Li AR, Li A, Liu Z, Tse YYT, Zhou X, Liu T, Xiong C, Huang J, Ming WK
Surveillance of Twitter Data on COVID-19 Symptoms During the Omicron Variant Period: A Sentiment Analysis
JMIR Form Res 2025;9:e66237
URL: https://formative.jmir.org/2025/1/e66237
doi:10.2196/66237

© Kaiyue Zhang, Zhaojin Guo, Yujie Ai, An-Ran Li, Anlin Li, Ziyu Liu, Yittie Yi Ting Tse, Xinyu Zhou, Taoran Liu, Chuxi Xiong, Jian Huang, Wai-kit Ming. Originally published in JMIR Formative Research (https://formative.jmir.org), 23.9.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Monitoring People With COVID-19 at Home With the COVIDFree@Home Program: Feasibility Cohort Study

Andrea S Gershon[1,2,3,4], MSc, MD; Alex Mariakakis[5], PhD; Eyal de Lara[5], PhD; Joseph Munn[2], PhD; Maryann Calligan[6], MA; Daniyal Liaqat[5,7], PhD; Salaar Liaqat[5], PhD; Junlin Chen[5], MSc; Teresa To[3,4,8], PhD; Philip W Lam[2], MSc, MD; Andrew Simor[2], MD; Adrienne K Chan[1,2,3], MD, MPH; Nisha Andany[2], MD, MPH; Sameer Masood[1], MD; Nick Daneman[2], MSc, MD; Tiffany Chan[2,9], MD; Christopher Graham[9], MD; Vikram Comondore[10,11,12], MD; Andre de Moulliac[13], MBA; Alice Y Tu[6], MA; Robert Wu[1,6], MSc, MD

[1]Department of Medicine, Temerty Faculty of Medicine, University of Toronto, Toronto, ON, Canada
[10]Department of Medicine, William Osler Health System, Brampton, ON, Canada
[11]Department of Medicine, Toronto Metropolitan University, Toronto, ON, Canada
[12]Division of Respirology, McMaster University, Hamilton, ON, Canada
[13]IQVIA (United States), Durham, NC, United States
[2]Department of Medicine, Sunnybrook Health Sciences Centre, V1 06, 2075 Bayview Avenue, Toronto, ON, Canada
[3]Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada
[4]ICES, Toronto, ON, Canada
[5]Department of Computer Science, University of Toronto, Toronto, ON, Canada
[6]Medicine, University Health Network, Toronto, ON, Canada
[7]Vector Institute for Artificial Intelligence, Toronto, ON, Canada
[8]Child Health Evaluative Sciences, The Hospital for Sick Children, Toronto, ON, Canada
[9]Department of Medicine, Trillium Health Partners, Toronto, ON, Canada

Corresponding Author:
Andrea S Gershon, MSc, MD
Department of Medicine, Temerty Faculty of Medicine, University of Toronto, Toronto, ON, Canada

Abstract

Background: During the COVID-19 pandemic, many with acute infection isolated at home, with a small but significant number requiring hospitalization. At the time, since the pathogen was fairly unknown, clinicians were uncertain about which patients would rapidly deteriorate and need hospitalization. We developed the COVIDFree@Home smartphone app and clinician dashboard to monitor and support people managing at home with acute COVID-19 infection. It was uncertain whether such an app would be used by patients and whether it would support patient care. This knowledge would inform telemedicine and digital health tools being used to deliver care to patients remotely at that time.

Objective: This study aimed to determine the feasibility of using a smartphone app and clinician dashboard for remote clinical monitoring of people with COVID-19 at home.

Methods: A feasibility study set at 3 hospital sites (University Health Network, Sunnybrook Health Sciences Centre, and Trillium Health Partners) between 2020 and 2022 was conducted. Participants newly diagnosed with COVID-19 were asked to enter data into a smartphone app called COVIDFree@Home twice daily for 10 days while isolating at home. Their data, including symptoms, temperature, and oxygen saturation, were monitored on a clinician-facing dashboard. The primary outcome of feasibility was the number of patients who used the app. We also examined patient satisfaction through a survey questionnaire.

Results: A total of 431 patients were recruited, out of which 229 (56.5%) were females and the average age was 38.9 (SD 12.8) years. There were 376 (87.2%) participants who used the app to report symptoms or oxygen saturation at least once. Among these participants, 373 (99.2%) reported symptoms and 363 (96.5%) reported oxygen saturation. Participants reported symptoms an average of 1.7 (SD 1.1) times per day for a median of 5 (IQR 3‑8) days. Oxygen saturation levels were reported 1.5 times per day for a median of 6 (IQR 4‑9) days. There were 19 hospitalizations (4.4%) among study participants. Most patients felt comfortable using the app, felt reassured their data was being monitored and did not have privacy concerns.

Conclusions: Patients with acute COVID-19 infection engaged with a remote home monitoring platform, however, not at the recommended frequency or duration. Remote patient monitoring of acute respiratory infection appears viable and can offer patients reassurance. It has the potential to reduce strain on the health care system during future pandemics, but further evidence is required to demonstrate improved health outcomes.

XSL•FO
RenderX

Trial Registration: Clinicaltrials.gov NCT04453774; https://clinicaltrials.gov/study/NCT04453774

(JMIR Form Res 2025;9:e69140) doi:10.2196/69140

KEYWORDS

COVID-19; remote monitoring; mobile application; oxygen saturation; pandemic

Introduction

The COVID-19 pandemic had an unprecedented global impact on human health and health care systems. The influx of patients overburdened hospitals and led to disruptions in health care, demanding quick adaptations in how care was delivered [1,2]. Particularly in the initial phase of the pandemic, people with COVID-19 isolated at home, with a small but significant number deteriorating and requiring hospitalization. With home isolation becoming common and health care resources being limited, people turned to remote care solutions to support patients outside of hospital settings [3]. Since the pathogen was fairly unknown, clinicians were uncertain about which patients would rapidly deteriorate and need hospitalization. Many clinicians were concerned about a phenomenon referred to as “silent hypoxia,” wherein patients had low oxygen saturation levels without experiencing shortness of breath. This led to groups recommending that patients be followed at home with oximeters [4-6].

In response to this, a transdisciplinary team of researchers, computer scientists, and clinicians (infectious disease, respirology, and general internal medicine) developed the COVIDFree@Home remote monitoring program to support and study people recovering from COVID-19 for the duration of their illness. This program consisted of a patient-facing smartphone app for data collection and a clinician-facing dashboard that enabled health care providers to remotely monitor a roster of patients. The uptake of such a program by patients, as well as their satisfaction with these technologies in the context of the pandemic, was unknown.

During the pandemic, telemedicine and digital health tools subsequently emerged as means to deliver care to patients [7], but few were evaluated. A systematic review of remote monitoring programs for COVID-19 found 23 descriptive reports of remote monitoring systems that varied significantly in the patient groups they studied, the technology they used, and the outcome monitored [8]. Of the 23 reports, only 9 assessed patient satisfaction, and only 12 assessed patient adherence. None evaluated effectiveness. Another review examined 241 studies of remote monitoring of patients with COVID-19 and found a lack of evidence of effectiveness [9]. Of these, only 73 evaluated patient experience. This review did not assess adherence. Lack of evaluation of previous remote monitoring programs represents a lost opportunity to learn from past experiences to improve monitoring in the future.

We conducted a prospective study to document the feasibility of using the COVIDFree@Home program for patients with acute respiratory infection, specifically those newly diagnosed with COVID-19 infection isolating at home. Our secondary objective was to determine patient satisfaction with the monitoring. This information can be used to inform future remote clinical monitoring solutions.

Methods

Study Design

We conducted a multisite, prospective clinical study to evaluate the feasibility of using a smartphone app for remote clinical monitoring of patients with COVID-19 recovering at home. Patients were recruited from 3 Ontario health networks: University Health Network, Sunnybrook Health Sciences Centre, and Trillium Health Partners from October 12, 2020, to June 23, 2022 (ClinicalTrials.gov NCT04453774).

Participants

Patients were recruited from COVID-19 ambulatory testing sites at the 3 hospital networks. In the course of regular clinical care, those who received a positive SARS-CoV-2 polymerase chain reaction test result were contacted by phone by clinicians for a routine online synchronous medical assessment. Patients were invited to participate in our study during these online assessments.

For those with acute COVID-19 who were interested in participating, a member of the research team contacted them by phone to go through the inclusion and exclusion criteria and informed consent process. Patients were invited to participate if they were 18 years of age or older, English-speaking, within 6 days of testing positive for COVID-19, and experiencing symptoms of COVID-19. Those who were receiving palliative care or had significant comorbidities that would confound symptoms were excluded but still received regular care. Patients were followed for the duration of their isolation period (10 days after symptom onset) or until their symptoms resolved, whichever was earlier.

Study Intervention: COVIDFree@Home Program

The COVIDFree@Home program consisted of a patient-facing smartphone app and a clinician-facing dashboard. Patients were asked to download the smartphone app to their personal devices; if they did not own one, smartphones were available to be loaned by the study team, but this did not occur in our study. Patients were also provided with a thermometer and digital pulse oximeter.

After taking readings with the devices, they were instructed to input the resulting oxygen saturation level, heart rate, and temperature data into the app while also noting the symptoms they were experiencing. Patients were asked to do this twice daily for routine assessments and input additional records whenever they experienced a notable change in health (Figure 1). Patients were instructed that their information was reviewed daily, that they should not depend on it for urgent medical

XSL•FO
RenderX

attention, and that they should use their usual processes (eg, calling their family physician or going to the emergency department) as appropriate.

Figure 1. Screenshots from the patient-facing COVIDFree@Home mobile app, which allows patients to report their symptoms, temperature, oxygen saturation, and heart rate.

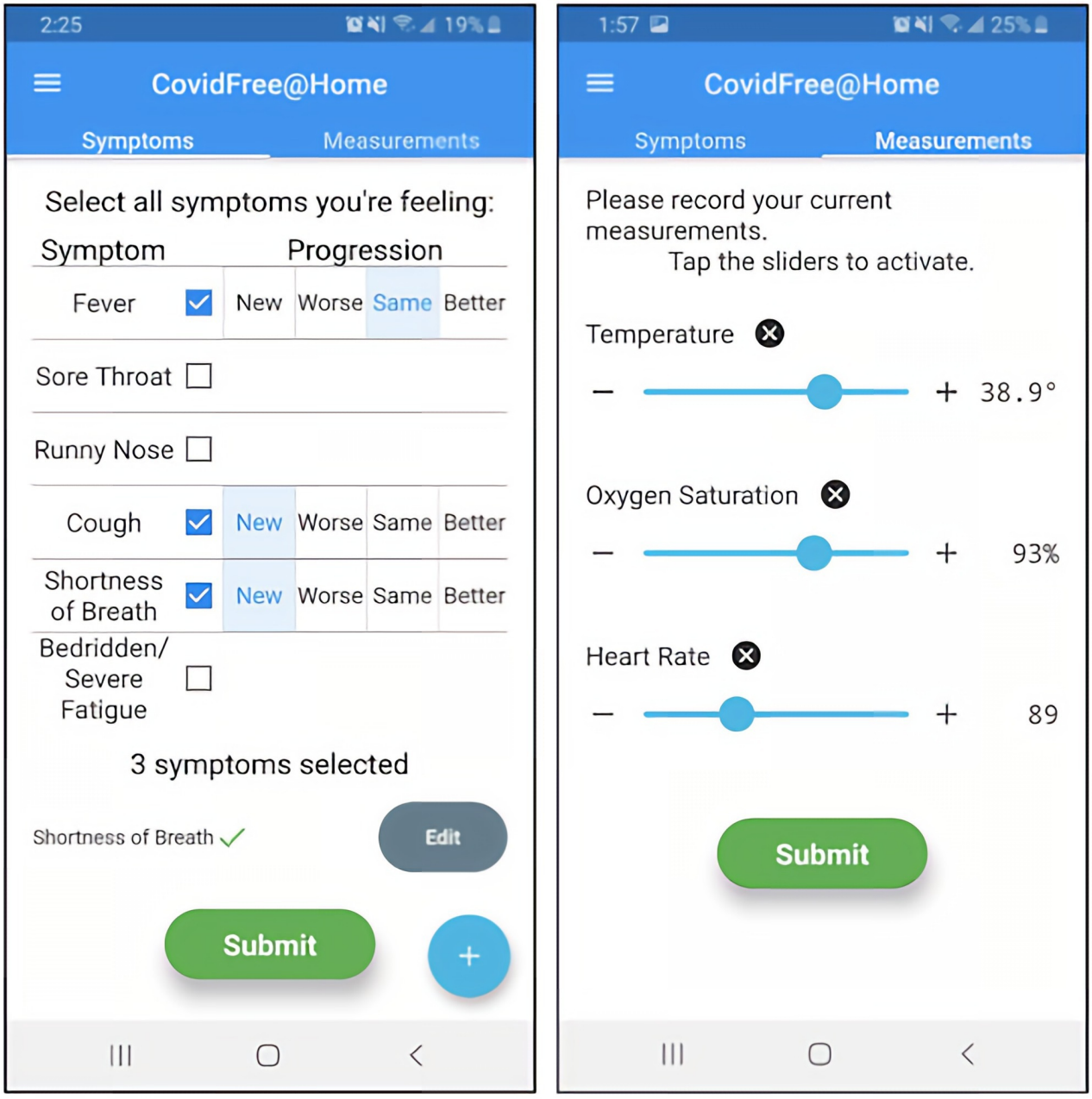

Individual patient information was displayed on a clinician-facing dashboard accessed via a secure website by clinicians (Figure 2). Out-of-range values were visually highlighted to enable rapid identification of patients who were potentially deteriorating. Email alerts were also sent to clinicians automatically for out-of-range values as notifications. In addition, physicians reviewed their assigned patients' data and then signed off on it at least once per day. This sign-off could be seen by patients on their app, thus providing reassurance that their data was being monitored. To further facilitate interaction between patients and clinicians, there was a messaging feature between the app (patients) and the dashboard (clinicians). A physician monitored patients at each site. Escalation procedures were at the discretion of the monitoring physician but usually included calling the participant with a potential request that they go to the emergency department, asking them to keep monitoring, or providing reassurance that their health status is normal.

XSL•FO
RenderX

Figure 2. Screenshot of a page from the clinician-facing COVIDFree@Home dashboard. This screen displays age, oxygen saturation, temperature, heart rate, and notes on symptoms for a roster of patients being monitored.

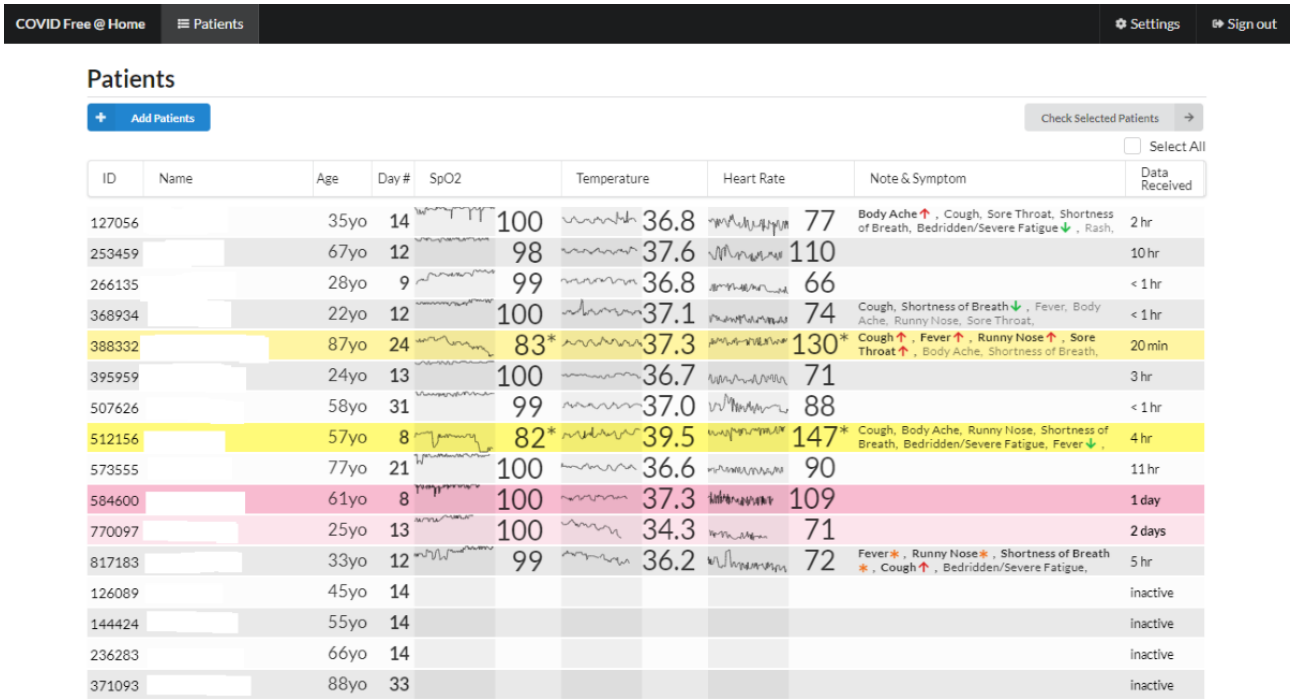

Study Outcomes

As this was a feasibility study, the primary outcome was adherence to using the app at least once to report symptoms or oxygen saturation. Reporting of oxygen saturation was also selected as the main outcome as it was the most relevant vital indicator for our targeted population. Secondary outcomes included the number of days the app was used and the average entries per day. Patients were asked to report if they were hospitalized through the app. We also asked patients to report on their experience using the smartphone app in an end-of-study questionnaire. In this questionnaire, they were asked to rate their agreement with a series of statements about using the COVIDFree@Home app on a 5-point Likert scale ranging from 1=strongly disagree to 5=strongly agree. Participants were also given the opportunity to provide free-form feedback in a text field.

Patient Characteristics

Demographics and comorbidities were collected via questionnaires during study enrollment. Ethnicity was collected to understand the differences among patients with COVID-19 who might experience marginalization.

Data Analysis

We used descriptive statistics to explore the characteristics of our patient population, usage of the app, and other secondary outcomes including patient experience. Quantitative data from the questionnaire were summarized according to median and IQR using Python (Python Software Foundation). Missing data and those lost to follow-up were reported. Free-form text data was reviewed using thematic coding, and representative quotes were selected for each theme.

Ethical Considerations

This study received Research Ethics Board approval through Clinical Trials Ontario (study ID 3185) and the associated study sites. Written consent was obtained from all patients prior to participation. Privacy and confidentiality protection were in place with study data deidentified. There was no compensation for participation.

Results

Patient Population

We consented 431 patients to participate in the study and provided them with the COVIDFree@Home app. Of these, 229 (56.5%) were reported as female, and the average age was 38.9 (SD 12.8) years (Table 1). A total of 185 of the 431 (42.9%) patients reported at least 1 comorbidity, with the most common being obesity (14.9%), asthma (8.6%), hypertension (7.1%), and diabetes (4%). Of the 98 participants who reported their ethnicity (22.7%), 30 (30.6%) identified as white and 39 (39.8%) as Asian.

XSL•FO
RenderX

Table . Baseline characteristics of the COVIDFree@Home project participants recruited at 3 sites in Ontario, Canada, between October 12, 2020, and June 23, 2022.

Characteristics	Using the app (n=376)	Not using the app (n=55)	Total (n=431)
Sex, n (%)			
Female	194 (51.6)	35 (63.6)	229 (56.5)
Missing or not reported[a]	—[b]	—	26
Age (years), mean (SD)	38.6 (12.5)	40.8 (14.5)	38.9 (12.8)
Missing or not reported	—	—	18
Self-reported comorbidities, n (%)			
Obesity	53 (15.3)	6 (12.2)	59 (14.9)
Asthma	31 (8.9)	3 (6.1)	34 (8.6)
Hypertension	26 (7.5)	2 (4.1)	28 (7.1)
Diabetes	14 (4)	2 (4.1)	16 (4)
Mental illness	14 (4)	4 (8.2)	12 (3)
Missing or not reported	—	—	35
Self-reported ethnicity, n (%)			
Asian	—	—	39 (39.8)
White	—	—	30 (30.6)
Middle Eastern	—	—	17 (17.3)
Black	—	—	6 (6.1)
Latin American	—	—	3 (3.1)
Other	—	—	3 (3.1)
Missing or not reported	—	—	333
COVID-19 variant by date[c], n (%)			
Alpha or Beta	209 (55.6)	34 (61.8)	243 (56.4)
Delta	55 (14.6)	9 (16.4)	64 (14.8)
Omicron	112 (29.8)	12 (21.8)	124 (28.8)
MRC[d] Dyspnea scale at baseline, n (%)			
No dyspnea	279 (86.9)	41 (87.2)	320 (87)
Slight dyspnea	34 (10.6)	2 (4.3)	36 (9.8)
Moderate dyspnea	8 (2.5)	3 (6.4)	11 (3)
Severe dyspnea	0 (0)	1 (2.1)	1 (0.3)
Missing or not reported	—	—	63

[a]Not reported due to large number of missing data.

[b]Not available.

[c]Determined by most prevalent variant at date of enrollment.

[d]MRC: Medical Research Council.

App Use

Overall, there were a total of 4121 entries for symptoms (average per patient 11, SD 12.9), 4128 for oximetry and heart rate (average per patient 11.4, SD 12.5), and 3598 for temperature (average per patient 10.3, SD 12.3). The average time that passed before participants began using the app after receiving a positive polymerase chain reaction test result was 2.1 (SD 3) days, including the time participants were invited and enrolled into the study. There were 376 participants (87.2%) who used the app at least once (Figure 3; Table 2). Among these participants, 373 (99.2%) reported symptoms, and 363 (96.5%) reported oxygen saturation. On average, participants entered symptoms 1.7 (SD 1.1) times and measured oxygen saturation 1.5 times (SD 1) per day. Most participants (346/376, 92%) reported symptoms and oxygen saturation for at least 6 days after enrollment.

XSL•FO
RenderX

Figure 3. Summary of participant engagement with various parts of the COVIDFree@Home app.

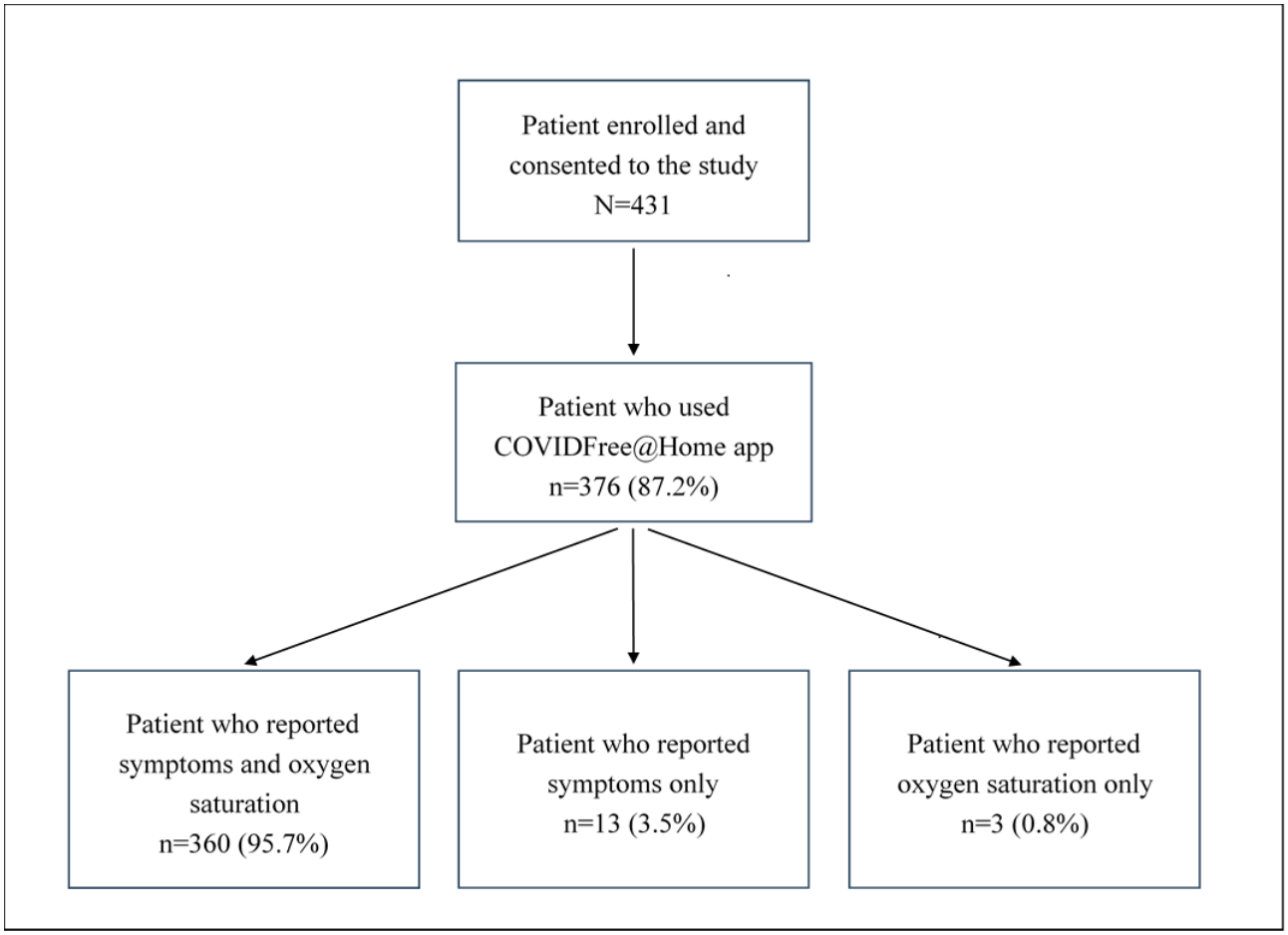

Table . App usage of patients who used the app at least once (n=376).

Usage **measure**	**Values** (n=376)
Participants who used the app to report symptoms, n (%)	373 (99.2)
Daily measures taken with the app, mean (SD)	1.7 (1.1)
Total days of app use, median (IQR)	5 (3-8)
Days from positive test to enrollment, mean (SD)	2.1 (3)
Days from enrollment to first entry, mean (SD)	5.5 (2.4)
Participants who used the pulse oximeter, n (%)	363 (96.5)
Daily measures taken using the pulse oximeter, mean (SD)	1.5 (1)
Total days of oximeter use, median (IQR)	6 (4-9)

Of the 363 participants who recorded oxygen saturation, 45 (12.4%) experienced at least 1 oxygen saturation level below 92%, and 13 (3.6%) below 88%. Among all study participants, 19 of 431 (4.4%) were subsequently hospitalized, and no deaths were reported.

Patient Experience

Of the 376 participants who used the COVIDFree@Home app, 143 (38%) completed the patient experience questionnaire at the end of the study (Table 3). On a 5-point Likert scale ranging from “1: strongly disagree” to “5: strongly agree,” participants were comfortable using the app (Q2: median 4, IQR 3 - 5), did not need assistance (Q4: median 1, IQR 1 - 1), and felt reassured their data was being monitored (Q7: median 5, IQR 4 - 5). Most did not have privacy concerns about reporting their health information through the app (Q12: median 1, IQR 1 - 2). Eighty-one respondents (56.6%) agreed that the app helped manage their COVID-19 infection, and 84 (58.7%) agreed that it improved communication with their health care providers. Participant free-form feedback was generally positive, with many noting the reassurance that monitoring provided (Textbox 1). Suggestions by some participants included being able to view one’s own data, particularly with trend lines to show historical changes.

XSL•FO
RenderX

Table . Questionnaire that participants completed to report their experience with the COVIDFree@Home app. Each question was answered along a 5-point Likert scale (1=strongly disagree, 2=somewhat disagree, 3=neither disagree nor agree, 4=somewhat agree, 5=strongly agree).

Statements	Median **response** (**IQR**)
I have used a virtual app for my medical care before using the COVID-Free@Home app	1 (1-3)
I was comfortable using electronic apps for my health prior to participating in this study	4 (3-5)
After participating in this study, I am comfortable using electronic apps for my health	5 (4-5)
I needed a lot of help to use the app (from friends, family, technical support)	1 (1-1)
I had technical difficulties due to technical problems with the app	1 (1-2)
I had difficulties due to internet connectivity	1 (1-1)
I felt reassured knowing my health data was being monitored	5 (4-5)
Entering data twice a day is a burden	2 (1-3)
The app helped me manage my COVID infection	4 (3-5)
The app made me more aware of my symptoms	4 (4-5)
The app improved communication with my health care provider	4 (3-5)
I had privacy concerns that my health information was being uploaded	1 (1-2)
I would like to be able to see the information that I entered in the app	4 (3-5)

Textbox 1. Selected feedback from participants on the COVIDFree@Home app.

- "It gave me an unbelievable amount of piece (sic) of mind to see that someone was watching my entries. My biggest fear was not knowing if I should seek further medical attention."
- "This app was very important to my family. My wife had much worse symptoms than I did and the feedback we received gave us peace of mind that we were on the road to recovery."
- "I was very impressed with the response from the medical team when my symptoms became a concern. Using this app kept me out of the hospital and provided me with oxygen and puffers to help with my recovery."
- "It would have been powerful (and helpful to me) to see a trend analysis to show me how my data was tracking over time. Not sure why you wouldn't empower patients to see their own data and how it changes. Brings much better awareness to the situation."

Discussion

Principal Findings

We conducted a multisite, prospective clinical study to determine the feasibility of using a smartphone app to remotely monitor patients with respiratory disease isolating at home during a pandemic. We found that 87% of participants reported symptoms and oxygen saturation using the app at least once and continued to use it for a median of 5 days. In terms of user experience, the feedback received from participants was positive overall. In assessing patient satisfaction with app-based monitoring, we found that most respondents were comfortable using the platform, found it helpful in managing their COVID-19 condition, and reported minimal privacy concerns. Taken together, these findings highlight the feasibility of using a smartphone app to support patients in managing their health and offering emotional comfort through remote monitoring during a pandemic.

The burden of COVID-19 infection has been dramatically reduced due to less aggressive variants, vaccine development, and new therapeutics for COVID-19. Unlike early in the pandemic, COVID-19 infection is now more predictable, and hospitalization is less likely. However, 6 million people are still hospitalized and between 400,000 to 700,000 people die from COVID-19, influenza, and respiratory syncytial virus per year globally [10,11]. Most of these are frail, immunosuppressed, or both. We need remote monitoring that can help these people and reduce the burden on health care systems. Our findings can be used to create and improve upon monitoring systems that patients will use and that are effective.

Our findings that most participants used an app to report health information and were satisfied with the experience are consistent with the literature. In a systematic review of 23 studies using remote patient monitoring in people with acute COVID-19, adherence was generally high in the 12 that evaluated it [8]. Similarly, patient acceptance was reported in the 9 studies that evaluated it. In a large review of 164 telemonitoring papers, 73 mentioned patient experience and found it positive [9]. Of note, this review found no evidence of effectiveness in improving outcomes of reduced ED visits or hospitalizations. While a systematic review of reviews of mobile apps for COVID-19 found many reviews of many apps, few focused on remote

XSL•FO
RenderX

monitoring of infection [12]. A recent review of remote monitoring of COVID-19 found that 4 of 13 total studies used a smartphone self-monitoring app [13-17]. Of these, 1 reported satisfaction which was predominantly positive [14], and 1 provided information on usage [15]. Our study adds value by providing a prospective analysis of the feasibility, usage, and patient experience of remote monitoring using a smartphone app—a format of remote monitoring for which studies of feasibility have been limited [18].

The strengths of our study were its ability to prospectively evaluate a real-world clinically active remote monitoring program. It also has limitations that warrant emphasis. First, we were not able to recruit all people who were seen with COVID-19 infection at our sites, and thus, our findings may not be generalizable to the general population. This was due to a combination of participants not being able or interested in participating. We found this to be worse during times when COVID-19 surged, and it was made worse because participants had to be enrolled in a limited window of time right after their infection was diagnosed. A second limitation was that—due to the nature of how our study was set up to accommodate the pandemic—we were unable to determine which patients did not engage with COVIDFree@Home. We also found low questionnaire response rates to some questions, namely participants' ethnicity, limited our ability to collect complete demographic data. This left us unable to be certain who found our remote monitoring program inaccessible or unacceptable. For example, while older people are more likely to experience severe illness, we hypothesize that they may have been less likely to participate because they were less comfortable with technology due to diminished digital health literacy. A third limitation is that we were unable to determine the effectiveness of the intervention due to there being too few adverse clinical outcomes and a lack of control arm. Future studies should assess feasibility and effectiveness of remote monitoring in diverse socioeconomic groups including older populations. Fourth, we did not evaluate health care provider experience alongside patient experience. To scale such a system, it would be important to engage health care workers who would do the monitoring.

Conclusions

Our multisite clinical study indicated that our remote patient monitoring intervention, the COVIDFree@Home app and clinician dashboard, was feasible and well-received according to quantitative usage metrics and qualitative feedback. Future research should use remote monitoring data to develop digital biomarkers that predict patient deterioration so that interventions can be implemented proactively to prevent hospitalizations [19].

Acknowledgments

The COVIDFree@Home project was supported by the Canadian Institutes of Health Research (CIHR) VR4-172746, Coronavirus Variants Rapid Response Network (CoVaRR-Net) grant (#175622), and the PSI Foundation.

Data Availability

The datasets generated or analyzed during this study are not available for sharing due to permissions not being obtained for sharing in the informed consent.

Conflicts of Interest

ASG, ED, DL, SL, JC, and RW were involved in the development of the innovation evaluated in this study. The other authors declare no conflicts of interest.

References

1. Okafor NM, Thompson I, Venkat V, et al. Evaluating the feasibility, adoption, cost-effectiveness, and sustainability of telemedicine interventions in managing COVID-19 within low-and-middle-income countries (LMICs): a systematic review. PLOS Digit Health 2025 Apr;4(4):e0000771. [doi: 10.1371/journal.pdig.0000771] [Medline: 40198616]
2. Garfan S, Alamoodi AH, Zaidan BB, et al. Telehealth utilization during the Covid-19 pandemic: a systematic review. Comput Biol Med 2021 Nov;138:104878. [doi: 10.1016/j.compbiomed.2021.104878] [Medline: 34592585]
3. Chaiyachati KH, Shea JA, Ward M, et al. Patient and clinician perspectives of a remote monitoring program for COVID-19 and lessons for future programs. BMC Health Serv Res 2023 Jun 27;23(1):698. [doi: 10.1186/s12913-023-09684-1] [Medline: 37370059]
4. Teo J. Early detection of silent hypoxia in Covid-19 pneumonia using smartphone pulse oximetry. J Med Syst 2020 Jun 19;44(8):134. [doi: 10.1007/s10916-020-01587-6] [Medline: 32562006]
5. Lam PW, Sehgal P, Andany N, et al. A virtual care program for outpatients diagnosed with COVID-19: a feasibility study. CMAJ Open 2020;8(2):E407-E413. [doi: 10.9778/cmajo.20200069] [Medline: 32447283]
6. Levitan RM. Pulse oximetry as a biomarker for early identification and hospitalization of COVID-19 pneumonia. In: Kline JA, editor. Acad Emerg Med 2020 Aug;27(8):785-786. [doi: 10.1111/acem.14052] [Medline: 32779867]
7. Shaver J. The state of telehealth before and after the COVID-19 pandemic. Prim Care 2022 Dec;49(4):517-530. [doi: 10.1016/j.pop.2022.04.002] [Medline: 36357058]

XSL•FO
RenderX

8. Joyce D, De Brún A, Symmons SM, Fox R, McAuliffe E. Remote patient monitoring for COVID-19 patients: comparisons and framework for reporting. BMC Health Serv Res 2023 Aug 3;23(1):826. [doi: 10.1186/s12913-023-09526-0] [Medline: 37537615]
9. Cornelis J, Christiaens W, de Meester C, Mistiaen P. Remote patient monitoring at home in patients with COVID-19: narrative review. JMIR Nurs 2024 Nov 19;7:e44580. [doi: 10.2196/44580] [Medline: 39287362]
10. Mazur NI, Terstappen J, Baral R, et al. Respiratory syncytial virus prevention within reach: the vaccine and monoclonal antibody landscape. Lancet Infect Dis 2023 Jan;23(1):e2-e21. [doi: 10.1016/S1473-3099(22)00291-2] [Medline: 35952703]
11. Influenza (seasonal). World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/influenza-(seasonal) [accessed 2024-06-12]
12. Holl F, Schobel J, Swoboda WJ. Mobile apps for COVID-19: a systematic review of reviews. Healthcare (Basel) 2024 Jan 8;12(2):139. [doi: 10.3390/healthcare12020139] [Medline: 38255029]
13. Lara B, Kottler J, Olsen A, Best A, Conkright J, Larimer K. Home monitoring programs for patients testing positive for SARS-CoV-2: an integrative literature review. Appl Clin Inform 2022 Jan;13(1):203-217. [doi: 10.1055/s-0042-1742370] [Medline: 35172373]
14. Annis T, Pleasants S, Hultman G, et al. Rapid implementation of a COVID-19 remote patient monitoring program. J Am Med Inform Assoc 2020 Aug 1;27(8):1326-1330. [doi: 10.1093/jamia/ocaa097] [Medline: 32392280]
15. Gordon WJ, Henderson D, DeSharone A, et al. Remote patient monitoring program for hospital discharged COVID-19 patients. Appl Clin Inform 2020 Oct;11(5):792-801. [doi: 10.1055/s-0040-1721039] [Medline: 33241547]
16. Medina M, Babiuch C, Card M, et al. Home monitoring for COVID-19. Cleve Clin J Med 2020 Jun 11. [doi: 10.3949/ccjm.87a.ccc028] [Medline: 32409432]
17. Yordanov Y, Dechartres A, Lescure X, et al. Covidom, a telesurveillance solution for home monitoring patients with COVID-19. J Med Internet Res 2020 Oct 22;22(10):e20748. [doi: 10.2196/20748] [Medline: 33006938]
18. Pandit JA, Radin JM, Quer G, Topol EJ. Smartphone apps in the COVID-19 pandemic. Nat Biotechnol 2022 Jul;40(7):1013-1022. [doi: 10.1038/s41587-022-01350-x] [Medline: 35726090]
19. Sim I. Mobile devices and health. N Engl J Med 2019 Sep 5;381(10):956-968. [doi: 10.1056/NEJMra1806949] [Medline: 31483966]

Abbreviations

Edited by A Mavragani; submitted 13.02.25; peer-reviewed by E Ross, R Nourse; revised version received 06.08.25; accepted 11.08.25; published 08.10.25.

Please cite as:
Gershon AS, Mariakakis A, de Lara E, Munn J, Calligan M, Liaqat D, Liaqat S, Chen J, To T, Lam PW, Simor A, Chan AK, Andany N, Masood S, Daneman N, Chan T, Graham C, Comondore V, de Moulliac A, Tu AY, Wu R
Monitoring People With COVID-19 at Home With the COVIDFree@Home Program: Feasibility Cohort Study
JMIR Form Res 2025;9:e69140
URL: https://formative.jmir.org/2025/1/e69140
doi:10.2196/69140

© Andrea S Gershon, Alex Mariakakis, Eyal de Lara, Joseph Munn, Maryann Calligan, Daniyal Liaqat, Salaar Liaqat, Junlin Chen, Teresa To, Philip W Lam, Andrew Simor, Adrienne K Chan, Nisha Andany, Sameer Masood, Nick Daneman, Tiffany Chan, Christopher Graham, Vikram Comondore, Andre de Moulliac, Alice Y Tu, Robert Wu. Originally published in JMIR Formative Research (https://formative.jmir.org), 8.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Trends and Shifts in Swedish Telemedicine Consultations During the Pre–COVID-19, COVID-19, and Post–COVID-19 Periods: Retrospective Observational Study

Adaora Uloma Asomugha*, BTECH, MSc, DMed; Annamaria Pakai*, PhD

Doctoral School of Health Sciences, University of Pécs, Vörösmarty M u 4, Pécs, Hungary

*all authors contributed equally

Corresponding Author:
Adaora Uloma Asomugha, BTECH, MSc, DMed
Doctoral School of Health Sciences, University of Pécs, Vörösmarty M u 4, Pécs, Hungary

Abstract

Background: In recent times, the telemedicine landscape has changed dramatically; it serves as a bridge, connecting health care providers and patients, especially during challenges such as the recent COVID-19 pandemic.

Objective: This study seeks to explore the Swedish telemedicine landscape in terms of primary patient symptoms for teleconsultation and the patterns of telemedicine use in the periods before COVID-19, during COVID-19, and after COVID-19, including the primary care use dynamics with respect to the teleconsultations done.

Methods: Secondary data was used in this observational retrospective study. The study population consisted of Swedish residents who had online telemedicine consultations. Telemedicine consultations were divided by text and video delivery; the period of analysis ranged from November 2018 to June 2023. The statistical methods used for the data analysis were descriptive analysis, 2-way cross tabulation, and a generalized linear model.

Results: During the pandemic, the number of teleconsultations concerning general, unspecified symptoms increased in comparison to the other analyzed symptoms, signaling a change in care-seeking behavior under epidemiological pressure. General health-related issues were the most pronounced symptom across all periods: 186.9 of 1000 consultations before COVID-19, 1264.6 of 1000 consultations during COVID-19, and 319.2 of 1000 consultations after COVID-19. There was no significant main effect of COVID-19 period on the number of telemedicine consultation meetings (F_2=1.653; P=.38). The interaction effect between delivery type and period was statistically significant (F_2=14.723; P<.001).

Conclusions: The findings are in favor of the COVID-19 pandemic having had a considerable effect on telemedicine use. Telemedicine could subsequently be used more often for general health consultations and acute conditions. Video consultations were more prominent because of the importance of bidirectional communication. The study suggests that there was a transformation of patterns of demand for health care; there is a necessity for health care systems to respond to these changes.

(JMIR Form Res 2025;9:e60294) doi:10.2196/60294

KEYWORDS

telemedicine; COVID-19; health care consultation; Sweden; digital health; patient behavior; video consultation; telehealth trends

Introduction

Despite it's recent popularity, the idea of telemedicine has it's roots in the early twentieth century [1,2]. Early on, telemedicine referred to the use of technologies like radio to enable patient access to medical care in remote locations. However, with the advancement of technology, telemedicine has evolved and has been optimized to use sophisticated platforms that incorporate advanced computing and communication technologies [3,4]. Today, the world is in dire need of telemedicine, which can bridge the gap between patients and health care providers and has become a realistic alternative to the traditional medical consultation [5]. Studies have shown that telemedicine is effective in providing timely interventions at reasonable cost; for instance, telemedicine substantially improved health outcomes and reduced hospitalization rates among patients with chronic illness [6-8].

Telemedicine has recently gained even more prominence and is being used more and more. For instance, the World Health Organization [9] estimated that telemedicine has been integrated into more than 60% of its member countries' health care systems, proving the acceptance and use of the technology. Similarly, a Precedence Research Market Report [10] projected that the telemedicine industry will grow at a compound annual growth rate of 17.16% from 2022 to 2030 and forecast a value of US $225 billion. Telemedicine adoption differs by area; North

XSL•FO
RenderX

America has led the way, thanks in part to it's technological infrastructure and legislative frameworks, but emerging countries are seeing an increase in telemedicine programs. In particular, telemedicine has played a key role in the fight against diseases within endemic regions and in the delivery of important health care interventions in some parts of Africa and Asia [11].

In Sweden, telemedicine has grown in response to progressive legislation, state-of-the-art technology, and a continued commitment to patient-centered treatment [12]. With people older than 65 years expected to comprise 20% of the population of Sweden by 2050, there is a looming lack of health care resources [13]. Sweden therefore needs to deploy telemedicine solutions quickly to meet the needs of an aging population. Telemedicine for remote monitoring and chronic disease treatment is well placed to become a key approach to dealing with this demographic shift.

The COVID-19 pandemic presented health systems all over the globe, including in Sweden, with unprecedented challenges. Prior to the pandemic, telemedicine was sufficiently flexible to address a wide range of symptoms and care requirements [14]. The pandemic period unquestionably indicated that this flexibility not only offered a solid basis for treatment but also highlighted the ability of telemedicine to adapt and scale in the face of global issues [15]. Telemedicine consultations increased dramatically in the first half of the pandemic, with some health care providers reporting an increase of more than 300% compared with the same period the year before [16]. Despite this, the postpandemic world showed that telemedicine, once looked upon as the future, is now very much a present reality, leading the way in setting the course of patient care, beyond what was conceived of in the past [17]. This new epoch provides us a beacon of hope and a new field of research as technology, medical expertise, and patient centricity converge into a medical model.

The aims of this study were to, first, elaborate on what patient symptoms prompted their use of telemedicine consultations (using a respectful approach) and explore the differences, if any, between the pre–COVID-19, COVID-19, and post–COVID-19 pandemic periods. Second, we aimed to assess differences in patterns of telemedicine consultations in the pre–COVID-19, COVID-19, and post–COVID-19 pandemic periods. Third, we aimed to analyze changes in the primary care use of telemedicine and the pattern of changes in this use in the pre–COVID-19, COVID-19, and post–COVID-19 pandemic periods. The COVID-19 pandemic was an unprecedented occurrence that necessitated telemedicine consultations and led to the widespread use of telemedicine. This study provides a critical analysis of trends and shifts in Swedish use of telemedicine consultations over time.

Methods

Study Design

An observational, retrospective design was used for this study. In a situation where experimentation is not controlled, researchers can observe trends or correlations without affecting the system [18]. This study involved telemedicine consultation formats and the symptoms noted during consultations in the pre–COVID-19 period of November 2018 to November 2019, the COVID-19 period of December 2019 to July 2022, and the post–COVID-19 period of August 2022 to June 2023.

Data and Population

The population for this study was all residents of Sweden in the periods under study. Data was collected from Kry Primary Care Limited, a leading health start-up and telemedicine provider that offers insightful data for research purposes and promotes accessible health care across several countries, including Sweden, the United Kingdom, France, Germany, Norway, and Spain.

Commonly, researchers use a primary methodology for obtaining aggregated and anonymized data from Kry Primary Care's large pool of telemedicine consultations, which conform to strict ethical guidelines and privacy regulations to ensure patient privacy. The dataset is a detailed record of all telemedicine consultations during the period of interest for this research. The data were recorded and collated by a designated research assistant in the organization and accessed after necessary approvals had been granted by the organization at different stages of verification and approval.

Kry Primary Care's datasets are a rich resource for telemedicine and include patient demographics, clinical outcomes, and use patterns. Kry Primary Care operates through a mobile app that patients can use to remotely consult physicians, as well as psychologists and other health care specialists, removing the need for conventional physical appointments.

The data from Kry Primary Care comprise the records of text-based consultations and video meetings, as well as the symptoms diagnosed by the health care providers after the consultations. Symptoms of skin rash, poisonous animal bites, and acne were categorized as skin disorders; sore throat, cold and flu, cough, fever, diarrhea, sinusitis, and eye infections were categorized as COVID-19–like symptoms; pollen and other allergies were categorized as allergy symptoms; chlamydia, urinary tract infections, and contraceptive prescriptions were categorized as reproductive health–related issues; and general health, revisits, and prescription renewals were categorized as general health-related issues. Other conditions included sleep disorders and threadworm infections.

Statistical Plan

Descriptive analyses of text and video consultations performed during the periods before, during, and after COVID-19 were prepared. Two-way cross-tabulation was used to summarize the symptoms and compare them before, during, and after COVID-19. These methods were used to provide an efficient summary of the trends across time periods and consultation types.

A 2-way, between-subjects, generalized linear model was used to establish significant interaction effects and main effects between the periods and meeting consultation styles. A generalized linear model offered necessary flexibility to this study given the complex nature of health care, which involves nonnormal distributions. The Tukey's HSD test was used post

XSL•FO
RenderX

hoc to determine the levels of each factor that were significantly different from each other, controlling for type I errors across multiple comparisons. SPSS (version 25; IBM Corp) was used for the data analysis.

Ethical Considerations

This research was exempted from human subjects research ethics review due to the nature of the dataset, which contains anonymized secondary data collected by Kry Primary Care. The company uses deidentified data from patients who consent to the company's terms of service and privacy policy for research purposes. Consequently, additional informed consent was not required; likewise, no compensation was provided, as the secondary data did not involve any personal interaction and did not include access to patient-identifiable information. This secondary analysis did not need a review from an institutional review board, given that the data were collected in Sweden and the study was conducted in Sweden. According to Swedish Ethical Review Authority [19], "If you have designed a study so that no sensitive personal data or personal data about violations of the law will be received and you do not intend to process such data, no ethical review is needed." Consequently, the authors obtained permission only from Kry Primary Care to use the data for this research.

Results

Patients' Symptoms That Prompted Telemedicine Consultations

Table 1 shows that the most pronounced symptom category across all periods was general health-related issues: 186.9 of 1000 consultations before COVID-19, 1264.6 of 1000 consultations during COVID-19, and 319.2 of 1000 consultations after COVID-19; more general health-related issues were reported during COVID-19 than after COVID-19 and before COVID-19. Other symptoms that prompted telemedicine consultations included skin disorders, COVID-19–like symptoms, allergies, reproductive health–related issues, sleep disorders, and threadworm infections. There were more COVID-19–like symptoms, such as sore throat, cold and flu, cough, fever, diarrhea, sinusitis, and eye infections, during COVID-19 (314.1/1000 consultations) and after COVID-19 (110.4/1000 consultations) and fewer occurrences before COVID-19 (141.4/1000 consultations).

Table . Patient symptoms per 1000 consultations before COVID-19, during COVID-19, and after COVID-19.

	Before COVID-19, consultations per 1000	During COVID-19, consultations per 1000	After COVID-19, consultations per 1000
Skin disorders	129.4	409.0	95.7
COVID-19–like symptoms	141.4	378.9	161.0
Allergies	21.6	0.6	14.9
Reproductive health–related issues	41.9	314.1	110.4
General health-related issues	186.9	1,264.6	319.2
Sleep disorders	9.7	—[a]	—
Threadworms	15.4	4.5	—

[a]Not applicable.

Differences in the Pattern of Telemedicine Consultations

The data showed that there were 2 patterns of telemedicine consultations in Sweden: text meetings and video meetings. There were more video meetings (mean 66,415.38, SD 23,737.20) than text meetings (mean 7247.73, SD 4747.52) across the 3 periods (before COVID-19, during COVID-19, and after COVID-19). There were significantly more text meetings during COVID-19 (mean 8428.31, SD 5564.29) than after COVID-19 (7965.27, SD 4747.52) or before COVID-19 (mean 3734.54, SD 2053.39). Likewise, more video meetings were recorded during COVID-19 (mean 76,977.19, SD 21,475.66) than after COVID-19 (mean 71,459.73, SD 5162.53) or before COVID-19 (38,364.15, SD 15,832.76), as presented in Table 2.

XSL•FO
RenderX

Table . Descriptive statistics for patterns of telemedicine consultations among patients in Sweden before, during, and after COVID-19.

Meeting types and periods		Meetings, mean (SD)
Text meetings		
	Before COVID-19	3,734.54 (2,053.387)
	During COVID-19	8,428.31 (5,564.290)
	After COVID-19	7,965.27 (1,361.069)
	Mean of Totals	7,247.73 (4,747.516)
Video meetings		
	Before COVID-19	38,364.15 (15,832.756)
	During COVID-19	76,077.19 (21,475.659)
	After COVID-19	71,459.73 (5,162.527)
	Mean of Totals	66,415.38 (23,767.203)
Totals		
	Before COVID-19	21,049.35 (20,836.101)
	During COVID-19	42,252.75 (37,475.717)
	After COVID-19	39,712.50 (32,702.515)
	Mean of Totals	36,831.55 (34,265.881)

A 2-way, between-subject, generalized linear model was used to assess the main and interaction effects of period (before, during, and after COVID-19) and meeting style. As presented in Table 3, there was no statistically significant main effect of period on the number of telemedicine consultation meetings (F_2=1.653; P=.38); in essence, the number of meetings in any period was not statistically different. The main effect of meeting type (text or video) on the number of consultations was statistically significant (F_1=28.682; P=.03) with a large effect size (η=0.934) at a 95% CI, indicating that video meetings were significantly more frequently used for telemedicine consultations than text meetings. The interaction effect between meeting type and period was statistically significant (F_2=14.723; P<.001) with a small effect size (η=0.22); for the pre–COVID-19 period, there were more video meetings than text meetings; likewise, during and after COVID-19, there were more video consultations than text meetings.

Pairwise comparisons of the interaction effect further proved that there was a significant increase in video meetings during the COVID-19 period compared to the pre–COVID-19 period. Similarly, there was a significant increase in video meetings during the post–COVID-19 period compared to the pre–COVID-19 period, while there was no difference in consultations between the COVID-19 and post–COVID-19 periods. Meanwhile, period did not have any significant effect on text meetings.

Table . ANOVA table of treatment means for telemedicine consultations among patients in Sweden before, during, and after COVID-19.

Source	Mean square	*F* test (*df*)	*P* value
Intercept	1.066×10^{11}	26.86 (1)	.04
Meeting type	6.902×10^{10}	28.682 (1)	.03
Period	4.269×10^{10}	1.653 (2)	.38
Meeting type × period	2.583×10^{10}	14.724 (2)	<.001

Changes in the Primary Care Use of Telemedicine

Primary care use before, during, and after COVID-19 involved either text meetings or video meetings. During all 3 periods, the number of video meetings was significantly higher than text meetings. The mean difference was the lowest before COVID-19 (mean 34,629.62; P<.001) and highest during COVID-19 (mean 67,648.88; P<.001); it then decreased after COVID-19 (mean 63,494.46, P<.001), as shown in Table 4.

XSL•FO
RenderX

Table . Mean difference in number of primary care uses of telemedicine consultation among patients in Sweden before, during, and after COVID-19.

	Mean (SE) difference between video and text meetings	*P* value
Before COVID-19	34,629.615 (5,195.856)[a]	<.001
During COVID-19	67,648.875 (3,311.721)[a]	<.001
After COVID-19	63,494.455 (5,648.491)[a]	<.001

[a]Significant at 95% level of confidence.

Discussion

Overview

This study aimed to investigate symptoms that prompted Swedish patients to seek telemedicine consultations and explore the differences, if any, before, during, and after the COVID-19 pandemic. The study further assessed differences in the pattern of telemedicine consultations before, during, and after the COVID-19 pandemic and analyzed changes in the primary care use of telemedicine and patterns of change in use in these periods.

Our findings provide evidence that, during the COVID-19 pandemic, there was a pronounced increase in telemedicine consultations for general health-related problems compared to the pre–COVID-19 and post–COVID-19 periods. As telemedicine was the only available alternative to access health care in the context of the public health measures taken during COVID-19, it is possible to correlate this marked elevation in teleconsultations for general health-related issues with other recorded symptoms. This is in contrast to the substantial reduction during the post–COVID-19 period, which mirrored the winding down of strategies for adapting and stabilizing health care services. It contradicts the findings of Kambris et al [20] that the execution of lockdown guidelines, along with a fear of visiting health clinics, resulted in the relocation of health care-seeking behavior to telemedicine. Given the increased risks of exposure associated with accessing health care in person during the COVID-19 pandemic, it seems that people reached out for a virtual consultation for general concerns about their health in order to limit their possible exposure to the virus.

As expected, the number of appointments for symptoms possibly caused by COVID-19 (sore throat, cold, flu, and fever), for example, increased during the pandemic when compared to the pre–COVID-19 period, given the level of concern and awareness at the time. However, these consultations maintained a consistent presence after COVID-19 due to the enduring role of telemedicine in addressing acute, potentially infectious conditions. This finding resonates with the existing literature underscoring the surge in telemedicine use during the pandemic due to an overlap between the symptoms of COVID-19 and those of other respiratory illnesses, such as the common cold or flu [21]. While individuals were unable to distinguish COVID-19 from other respiratory infections without the help of a medical professional, they were more amenable to communicate with health care professionals through digital platforms to ensure appropriate diagnosis and guidance because these platforms provided easy access to health care and were unhindered by physical boundaries. It was thus a key finding that postpandemic use of telemedicine persisted for general health and acute conditions, possibly indicating a permanent shift in health care–seeking behavior.

Our results further showed a clear preference for video meetings over text meetings across all study periods. This shows all the more the preference for meetings between health care providers and seekers that are close to home, intuitive, and clear; video meetings can improve the precision and quality of health care delivery. The results from this study reflect earlier research showing that face-to-face communication is more effective in building rapport, developing trust, and delivering nonverbal cues [22], all of which are important in the context of health care [23]. In addition, the findings corroborate [24] that technology advancement, improved internet connection, and the acceptance of telemedicine as a competitive choice to in-person consultations have improved the use of video meetings in the health care industry. Health care providers and searchers looking to avoid traveling physical distances have made the choice of video sessions because the human connection still needs to be preserved, even in the virtual world.

Furthermore, there was an increase in text and video consultations during the COVID-19 pandemic; this is consistent with the global requirement for remote health care solutions in the face of pandemic-related limits [25]. The number of text meetings increased dramatically throughout the COVID-19 period, as compared to the pre–COVID-19 period, and further increased into the post–COVID-19 period. This trend was mirrored in video meetings, which peaked during COVID-19 and underscored the essential role that visually engaging platforms will play in maintaining health care continuity in future circumstances similar to the COVID-19 pandemic. The rise of cell phone–optimized text-based communication in the health care sector is shown by the large difference in number of text consultations between the pre–COVID-19 and post–COVID-19 periods. As suggested by the findings of Hylock and Zeng [26] and Woodward et al [27], text consultation is effective and convenient for follow-up therapy and nonemergency medical matters; patients can consult with their health care provider asynchronously without an in-person meeting and can reschedule conveniently. Importantly, the ongoing use of text- and video-based meetings after COVID-19 demonstrates that the pandemic has affected the ways in which we seek health care and that we need to expand this research to better contextualize it.

This study further found that there was no statistically significant difference in the number of telemedicine consultations. This

XSL•FO
RenderX

result differs from what we expected, because previous studies have shown how effective telemedicine can be in times of medical emergency [28-30]. The lack of significant differences in the number of telemedicine consultations might be due to the digital divide; for example, demographic groups that have limited access to technology and internet services might have found it harder to take part in telemedicine consultations, resulting in bias in our data. In addition, the consistent number of consultations across the different periods may have been due to patient preferences and comfort levels with telemedicine. Some suggested that they preferred in-person visits because they liked the opportunity to speak directly to their health care provider or practitioner or because these visits gave them greater confidence in the accuracy of diagnoses, as they were made based on in-person physical examinations rather than video only. The disparity between patients' preferences and the potential benefits of telemedicine emphasize that incorporating patient views is critical to deploying new medical technology. There seems to have been no appreciable shift in the quantity of consultations when comparing different COVID-19 stages.

Technology is a part of today's digital age, especially in health care. By removing geographical challenges and providing flexible scheduling, video meetings make health care accessible to people with mobility difficulties or living in remote places [31]. Across all time periods in this study, some patients preferred video meetings with their physicians more than any other modality of consultation, suggesting an innate bias toward contact modalities that offer interactiveness and visual richness versus those that deemphasize the human-connection aspect of health care interactions. Several factors may have contributed to the popularity of online preferences. The innate propensity toward more interactive consultation modalities may be a desire to improve health care providers' and patients' ability to communicate and interact with one another. Moreover, robust and efficient health care service delivery is possible with video meetings, as they allow real-time engagement and visible facial expressions and body language clues [32].

Limitations

The results of this study should be interpreted with caution because of the presence of a few limitations. First, the data came only from telemedicine consultations in Sweden, and may not be generalizable to other locations because of confounding factors such as differences in health care systems and patient preferences outside of Sweden. As a result, the findings should not be generalized without proper qualification; this opens an opportunity for further studies to replicate the study with a wider geographical scope.

Second, the study data were collected retrospectively from telemedicine consultations, which may have resulted in errors or omissions in the results due to variability in the accuracy and completeness of symptoms and diagnoses recorded through consultations. In future research, this limitation can be mitigated, for example by comparing the recorded symptoms and diagnoses in the telemedicine consultations with other sources of medical data, such as electronic health records, patient self-reports, or follow-up consultations, while juxtaposing the records against patient feedback.

Third, this study was limited in its ability to explain the causes of the observed variances and similarities in telemedicine consultations across time. Further research could investigate the socioeconomic factors that may have influenced patients to use telemedicine during the COVID-19 outbreak.

Fourth, some text meetings were converted to video meetings as a function of necessity in specific cases, but the dataset did not specify which meetings were converted from text to video consultations. Hence, there is a likelihood that this could have affected the results.

Conclusions

This study examined symptoms for telemedicine consultations in Sweden and their trends before, during, and after the COVID-19 pandemic. During COVID-19, general health-related teleconsultations increased, perhaps because of public health concerns and a desire to minimize exposure. During the pandemic, there was a surge in COVID-19 consultations which demostrates that telemedicine can help manage symptoms such as sore throat and fevers. The results show the preference of the sample for video meetings over text meetings, which emphasizes the need to prioritize the visual and interactive aspects of communication in health care delivery. Socioeconomic considerations may have influenced patient participation in telemedicine during the COVID-19 pandemic, based on these findings. More research on public perceptions and attitudes affecting video preferences in telemedicine may help explain post–COVID-19 health care–seeking habits. Finally, further studies could examine telemedicine adoption barriers, such as the digital divide and access to technology and internet services, in an effort to increase telehealth service use for all demographics.

Data Availability

The datasets analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

AUA wrote the Introduction and analyzed the dataset. AP was instrumental in writing the rest of the manuscript. All authors read and approved the final manuscript. This research is self-funded.

Conflicts of Interest

None declared.

XSL•FO
RenderX

References

1. Dyb K, Halford S. Placing globalizing technologies: telemedicine and the making of difference. Sociology 2009 Apr;43(2):232-249. [doi: 10.1177/0038038508101163]
2. El-Miedany Y. Telehealth and telemedicine: how the digital era is changing standard health care. Smart Homecare Technol Telehealth 2017;4:43-51. [doi: 10.2147/SHTT.S116009]
3. Carlo FD, Sociali A, Picutti E, et al. Telepsychiatry and other cutting edge technologies in Covid-19 pandemic: bridging the distance in mental health assistance. Authorea 2020 May 21. [doi: 10.22541/au.159007666.69069059]
4. Manju R, Harinee P, Gangolli SS, Bhuvana N. Evolution of computational intelligence in modern medicine for health care informatics. In: Translating Healthcare Through Intelligent Computational Methods 2023:395-411. [doi: 10.1007/978-3-031-27700-9_24]
5. Orozco-Beltran D, Sánchez-Molla M, Sanchez JJ, Mira JJ, ValCrònic Research Group. Telemedicine in primary care for patients with chronic conditions: the ValCrònic quasi-experimental study. J Med Internet Res 2017 Dec 15;19(12):e400. [doi: 10.2196/jmir.7677] [Medline: 29246881]
6. Şahin E, Yavuz Veizi BG, Naharci MI. Telemedicine interventions for older adults: a systematic review. J Telemed Telecare 2024 Feb;30(2):305-319. [doi: 10.1177/1357633X211058340] [Medline: 34825609]
7. Kvedar J, Coye MJ, Everett W. Connected health: a review of technologies and strategies to improve patient care with telemedicine and telehealth. Health Aff (Millwood) 2014 Feb;33(2):194-199. [doi: 10.1377/hlthaff.2013.0992] [Medline: 24493760]
8. Bohingamu Mudiyanselage S, Stevens J, Watts JJ, et al. Personalised telehealth intervention for chronic disease management: a pilot randomised controlled trial. J Telemed Telecare 2019 Jul;25(6):343-352. [doi: 10.1177/1357633X18775850] [Medline: 29793387]
9. The World Health Report 2010. World Health Organization (WHO). URL: https://www.who.int/publications/i/item/9789241564021 [accessed 2025-03-27]
10. Telemedicine market size to reach USD 225 billion by 2030. Precedence Research. URL: https://www.precedenceresearch.com/telemedicine-market [accessed 2023-05-10]
11. Alhassan GN, Öztürk İ, Adedoyi˙n FF, Bekun FV. Telehealth as a panacea amidst global pandemic (COVID-19) in Africa. Düzce Tıp Fakültesi Dergisi 2021;23(Special Issue):43-47. [doi: 10.18678/dtfd.898373]
12. Heshmati A, Tsionas M, Rashidghalam M. An assessment of the Swedish health system's efficiency during the COVID-19 pandemic. Int J Healthc Manag 2023 Jul 3;16(3):336-352. [doi: 10.1080/20479700.2022.2102184]
13. Sweden, almost a country for old men. Istituto per gli Studi di Politica Internazionale. URL: https://www.ispionline.it/en/publication/sweden-almost-a-country-for-old-men-106929 [accessed 2023-03-20]
14. Omboni S, Padwal RS, Alessa T, et al. The worldwide impact of telemedicine during COVID-19: current evidence and recommendations for the future. Connect Health 2022 Jan 4;1:7-35. [doi: 10.20517/ch.2021.03] [Medline: 35233563]
15. Patrício Silva AL, Prata JC, Walker TR, et al. Increased plastic pollution due to COVID-19 pandemic: challenges and recommendations. Chem Eng J 2021 Feb 1;405:126683. [doi: 10.1016/j.cej.2020.126683] [Medline: 32834764]
16. Sclafani AP, Shomorony A, Stewart MG, Grunstein E, Overdevest JB. Telemedicine lessons learned during the COVID-19 pandemic: the augmented outpatient otolaryngology teleconsultation. Am J Otolaryngol 2021;42(4):102960. [doi: 10.1016/j.amjoto.2021.102960] [Medline: 33610921]
17. Lindeman DA, Kim KK, Gladstone C, Apesoa-Varano EC. Technology and caregiving: emerging interventions and directions for research. Gerontologist 2020 Feb 14;60(Suppl 1):S41-S49. [doi: 10.1093/geront/gnz178] [Medline: 32057082]
18. Thyer BA. Quasi-experimental group designs. In: Quasi-Experimental Research Designs: Oxford Academic; 2012:77-106. [doi: 10.1093/acprof:oso/9780195387384.003.0003]
19. Conditions for conducting research on personal data with or without ethical review. Swedish Ethical Review Authority. 2024. URL: https://etikprovningsmyndigheten.se/en/faq/jag-har-planerat-att-forska-om-personuppgifter-i-en-studie-som-inte-ska-vara-ett-forskningsprojekt-kan-jag-behandla-personuppgifter-utan-etikprovning/ [accessed 2025-03-27]
20. Kambris MEK, AlBlooshi S, Elamin A, Ajja R. Healthcare services utilization and seeking behaviour during COVID- 19 pandemic in the UAE – a cross-sectional study. Research Square. 2023. URL: https://www.researchsquare.com/article/rs-2710887/v1 [accessed 2025-03-27]
21. Agricola E, Beneduce A, Esposito A, et al. Heart and lung multimodality imaging in COVID-19. JACC Cardiovasc Imaging 2020 Aug;13(8):1792-1808. [doi: 10.1016/j.jcmg.2020.05.017] [Medline: 32762885]
22. Sethuraman K. Doctor-patient communication: an overview. In: Communication Skills in Clinical Practice (Doctor-Patient Communication): Jaypee Digital; 2001. [doi: 10.5005/jp/books/11607_2]
23. Pinedo-Torres I, Garcia-Villasante EJ, Gutierrez-Ortiz CC, et al. The doctor-patient relationship and barriers in non-verbal communication during teleconsultation in the era of COVID-19: a scoping review. F1000Res 2023;12:676. [doi: 10.12688/f1000research.129970.1]
24. Haleem A, Javaid M, Singh RP, Suman R. Telemedicine for healthcare: capabilities, features, barriers, and applications. Sens Int 2021;2:100117. [doi: 10.1016/j.sintl.2021.100117] [Medline: 34806053]

XSL•FO
RenderX

25. Wasserman D, Carli V. Evidence-based suicide prevention strategies during and after the COVID-19 pandemic. In: Wasserman D, editor. Oxford Textbook of Suicidology and Suicide Prevention: Oxford Academic; 2021:795-800. [doi: 10.1093/med/9780198834441.003.0096]
26. Hylock RH, Zeng X. A blockchain framework for patient-centered health records and exchange. J Med Internet Res 2019;21(8):e13592. [doi: 10.2196/13592] [Medline: 31471959]
27. Woodward M, De Pennington N, Grandidge C, McCulloch P, Morgan L. Development and evaluation of an electronic hospital referral system: a human factors approach. Ergonomics 2020 Jun;63(6):710-723. [doi: 10.1080/00140139.2020.1748232] [Medline: 32220218]
28. Ward MM, Jaana M, Natafgi N. Systematic review of telemedicine applications in emergency rooms. Int J Med Inform 2015 Sep;84(9):601-616. [doi: 10.1016/j.ijmedinf.2015.05.009] [Medline: 26072326]
29. Napi NM, Zaidan AA, Zaidan BB, Albahri OS, Alsalem MA, Albahri AS. Medical emergency triage and patient prioritisation in a telemedicine environment: a systematic review. Health Technol 2019 Nov;9(5):679-700. [doi: 10.1007/s12553-019-00357-w]
30. Zhu X, Merchant KAS, Mohr NM, Wittrock AJ, Bell AL, Ward MM. Real-time learning through telemedicine enhances professional training in rural emergency departments. Telemed J E Health 2021 Apr;27(4):441-447. [doi: 10.1089/tmj.2020.0042] [Medline: 32552479]
31. Kiianytsia VV. The role of telemedicine during the COVID-19 pandemic. In: New Trends and Unresolved Issues of Preventive and Clinical Medicine: Baltija Publishing; 2020. [doi: 10.30525/978-9934-588-81-5-2.46]
32. Almathami HKY, Win KT, Vlahu-Gjorgievska E. Barriers and facilitators that influence telemedicine-based, real-time, online consultation at patients' homes: systematic literature review. J Med Internet Res 2020 Feb 20;22(2):e16407. [doi: 10.2196/16407] [Medline: 32130131]

Abbreviations

Edited by A Mavragani; submitted 10.05.24; peer-reviewed by A Shah, U Sinha; revised version received 08.11.24; accepted 10.11.24; published 16.05.25.

Please cite as:
Asomugha AU, Pakai A
Trends and Shifts in Swedish Telemedicine Consultations During the Pre–COVID-19, COVID-19, and Post–COVID-19 Periods: Retrospective Observational Study
JMIR Form Res 2025;9:e60294
URL: https://formative.jmir.org/2025/1/e60294
doi:10.2196/60294

© Adaora Uloma Asomugha, Annamaria Pakai. Originally published in JMIR Formative Research (https://formative.jmir.org), 16.5.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Analysis of Social Media Perceptions During the COVID-19 Pandemic in the United Kingdom: Social Listening Study (2019-2022)

Marzieh Araghi[1], PhD; Arron Sahota[1], MPharm (Hons); Maciej Czachorowski[1], PhD; Kevin Naicker[1], PhD; Natalie Bohm[1], PhD; Katie Phillipps[2], BA (Hons); James Gaddum[2], BSc; Erica Jane Cook[3], PhD

[1]Pfizer Ltd, Walton Oaks, Dorking Rd, Surrey, United Kingdom

[2]Ipsos, 3 Thomas More Square, London, United Kingdom

[3]School of Psychology, University of Bedfordshire, Luton, United Kingdom

Corresponding Author:
Marzieh Araghi, PhD
Pfizer Ltd, Walton Oaks, Dorking Rd, Surrey, United Kingdom

Abstract

Background: Social media listening can be leveraged to obtain authentic perceptions about events, their impact, guidelines, and policies. There has been to date no research that has examined the experiences of patients with COVID-19 from diagnosis to treatment using social media listening in the United Kingdom.

Objective: This study aimed to assess public perceptions, insights, and sentiments throughout the patient journey from diagnosis to treatment during the COVID-19 pandemic.

Methods: A comprehensive search query was designed to retrieve social media data that referred to COVID-19 and treatment. The search was conducted using the social media monitoring tool, Synthesio (Ipsos). Data were retrospectively collected for the period covering September 2019 to September 2022 from Twitter (subsequently rebranded X), Facebook, Instagram, and YouTube as well as 126 public forums (including Health Unlocked, Mums Net, The Student Room, and Patient Forums UK). Available data in the United Kingdom expressed in the English language were collected and filtered, generating a final dataset consisting of 31,319 posts from an overall initial dataset of 706,634 posts. Complimentary Google trend analyses of search terms mentioning COVID-19 treatments were also performed.

Results: Social media posts related to COVID-19 symptoms accounted for 6% of overall posts, compared to 35% of posts related to testing, 25% of posts related to diagnosis, and 32% of posts related to treatment. Overall, the trend observed from social media posts relating to COVID-19 treatment extracted in Synthesio was largely congruent with the trend of COVID-19 searches on Google, indicating a potential relationship between public discourse and social media and internet search behavior.

Conclusions: The findings from this study have the potential to inform decision-making regarding public health interventions, communication strategies, and health care policies in the United Kingdom during future public health emergencies.

(JMIR Form Res 2025;9:e63997) doi:10.2196/63997

KEYWORDS

COVID-19; social media; perceptions; sentiments; United Kingdom; patient experiences; diagnosis; treatment; SARS-COV-2; coronavirus; respiratory; infectious; pulmonary; pandemic; Synthesio; data; Google; trends; health information; public health; search query; Ipsos; named-entity recognition

Introduction

The World Health Organization declared COVID-19 a global pandemic on March 11, 2020, with more than 118,000 confirmed cases in 114 countries and 4291 deaths [1]. In the United Kingdom, the first confirmed case of COVID-19 was reported on January 29, 2020 [2] with the first death recorded on March 5, 2020 [3]. To mitigate the spread of COVID-19, a national lockdown was implemented in the United Kingdom on March 23, 2020 [4,5].

Temporary closure of health care facilities because of the national lockdown introduced new challenges for patients with COVID-19 such as accessing health services and public health information. Dissemination of public health information and guidelines relating to preventive measures and vaccination by the UK government was challenged by the complexity of pandemic risk communication and the need to reach across the diverse UK population [6,7]. In such a scenario, social media platforms became a useful and vital substitute for sharing public health information, evidenced by the rapid increase in the

XSL•FO
RenderX

number of internet users during this period in the United Kingdom. The social media platforms including Facebook, Instagram, Twitter (subsequently rebranded X), WhatsApp as well as videoconferencing surged in popularity [8]. At the peak of the pandemic in 2020, an estimated 79% of "eligible users" (those aged 13 years and older) were active on social media, and the number of users reached 3.8 billion worldwide [9]. Thus, social media has played a positive and indispensable role in providing health information to the public [10].

Although the importance of social media for disseminating vital public health information cannot be overestimated, a concurrent rise in health misinformation was also observed. Social media posts relating to the origin of the virus, its pathogenesis, and transmissibility saturated social media platforms worldwide in a phenomenon termed as "infodemic" [11]. Processing social media activities may offer insights for monitoring public experience about different stages of the COVID-19 pandemic and how the health guidelines are perceived despite the flood of digital misinformation.

In the United Kingdom, previous social media listening studies have explored topics including COVID-19 vaccine hesitancy [12-15], the experience with health care [16], and the impact of COVID-19 on mental health [17]. However, to date, there has been no research examining the experiences of patients with COVID-19 from diagnosis to treatment. This research aims to address this gap, to provide useful information on how social media can be most effectively harnessed by health authorities to disseminate important public health messages, particularly to the UK's diverse population. The study described here was undertaken to collect and collate social media posts during the COVID-19 pandemic in the United Kingdom with the objective to assess public perceptions, insights, and sentiments throughout the patient pathway.

Methods

Search Strategy and Data Collection

To investigate discussions surrounding COVID-19 treatment on social media, we collected data from various social media platforms using a comprehensive search query (Table S1 in Multimedia Appendix 1). Data collection spanned from September 2019 to September 2022 and focused on capturing data from the United Kingdom, specifically English language content. Below, we outline our data sources and search strategy.

- Social media platforms and web-based forums: we used the proprietary social media monitoring tool, Synthesio [18] (Ipsos), to extract data meeting our search criteria from Twitter/X, Facebook, Instagram, YouTube, and public forums. Synthesio collects data from over 800 million data sources, but this search was restricted to just those located in the United Kingdom or social media users from the United Kingdom, resulting in inclusion of data from 126 public forums (such as Health Unlocked, Mums Net, The Student Room, and Patient Forums UK).
- Google Trends analysis: To complement the social media and forum data, we conducted Google Trends analysis, which allowed us to observe trends in public interest regarding COVID-19 treatments over time, again using search terms mentioning COVID-19 treatment.

This approach yielded a dataset of 709,634 posts.

Data Cleaning and Processing

Following the search, the generated dataset of 709,634 posts covering September 2019 to September 2021 was refined via a multistep process.

The steps are listed in Textbox 1.

Textbox 1. Summary of data cleaning and processing procedures.

1. Irrelevant content removal: the first step involved filtering out irrelevant posts to minimize noise and bias in the dataset, reducing the dataset to 325,770 posts. This included:
   - Automated filtering: We used Synthesio's built-in "noise filters" to automatically remove posts categorized as job offers, vulgar content, scams, fraud, financial news, advertisements, content and games, and autoposting.
   - Manual removal: We further refined the dataset by manually removing posts identified as spam, advertisements, and news blog articles during a review of the dataset.
2. Duplicate removal: to avoid overrepresentation of specific topics or sentiments, we identified and removed duplicate posts or Retweets. Non-COVID-19 and treatment-related posts were also excluded, reducing the dataset to 107,796 posts.
3. Relevance labeling and machine learning: to enhance data quality, we implemented a machine learning–based approach:
   - Manual tagging: A random sample of 600 posts was manually classified as "relevant" or "irrelevant" to COVID-19 treatment discussions.
   - Model training: Using this manually tagged dataset, we trained a Convolutional Neural Network (CNN), a deep learning algorithm, to automatically categorize posts. The model was then applied to the remaining dataset. Model evaluation metrics (Table S2 in Multimedia Appendix 1) show an overall F1-score of 0.787. While the recall for relevant posts was low, potentially resulting in some relevant posts being filtered out, we prioritized accurately identifying and removing irrelevant content due to the dataset's size.

Text Processing and Normalization

To prepare the data for analysis, we performed several text processing steps using a Python pipeline [19]:

1. Sentence segmentation: dividing posts into individual sentences for a granular analysis.
2. Duplicate removal: identifying and removing duplicate sentences.

XSL•FO
RenderX

3. Tokenization: breaking sentences into words or noun phrases for analysis.
4. Lemmatization: converting words to their base forms to standardize text.
5. Additional cleaning: removing personally identifiable information, unifying text case, and eliminating stop words (eg, “at,” “on”).

Selection of Themes and Conversations

For data categorization, we used Named Entity Recognition (NER) to classify unstructured text into predefined categories. We then identified key terms (~5000 in total) based on frequency (eg, “COVID” or “patient”). The list was manually reviewed and structured into a 3-tier hierarchy:

- Overarching group → Net → Codes and Terms
  - Example: Coronavirus → Variants → Delta, Omicron

Themes included symptoms, infection, testing, diagnosis, and treatment. Synonyms and related terms were grouped accordingly. Through this process, 89.7% of posts were categorized.

Data Analysis

With all posts in the dataset categorized, we analyzed the data using 2 approaches listed in Textbox 2:

Textbox 2. Summary of data analysis findings: overall trends and patient journey stages.

- **Overall trends**
  - Frequency tables were generated to highlight dominant discussion themes.
- **Patient journey stages**
  - Posts were classified into symptoms, testing, diagnosis, and treatment categories using the Synthesio platform. Researchers manually reviewed and tagged posts to identify thematic trends [20].
  - Sentiment analysis [21] was conducted using a convolutional neural network–based model within Synthesio. Posts were classified as positive, negative, or neutral, with the model continuously refined via user feedback.

Validity and Reliability

To cross-validate social media trends, we examined Google search data using AnswerThePublic [21], Google Trends [22], and Ahrefs [23]. These platforms provided insights into search volume and query types related to COVID-19 treatment, allowing us to compare social media discussions with public search behavior.

Ethical Considerations

This study did not require human participant research review, as it did not meet the criteria for such oversight. All data were sourced from publicly accessible social media platforms. However, due to the inclusion of health-related content concerning COVID-19, additional privacy safeguards were implemented. Personally identifiable information was protected by using obfuscated identifiers; social media post IDs and usernames were redacted during data collection.

Results

Overview

Following the search, the generated dataset of 709,634 posts covering September 2019 to September 2021 was refined via a multistep process (Figure 1). This comprehensive cleaning process resulted in a final dataset of 31,319 posts.

XSL•FO
RenderX

Figure 1. Flowchart of data process and analysis.

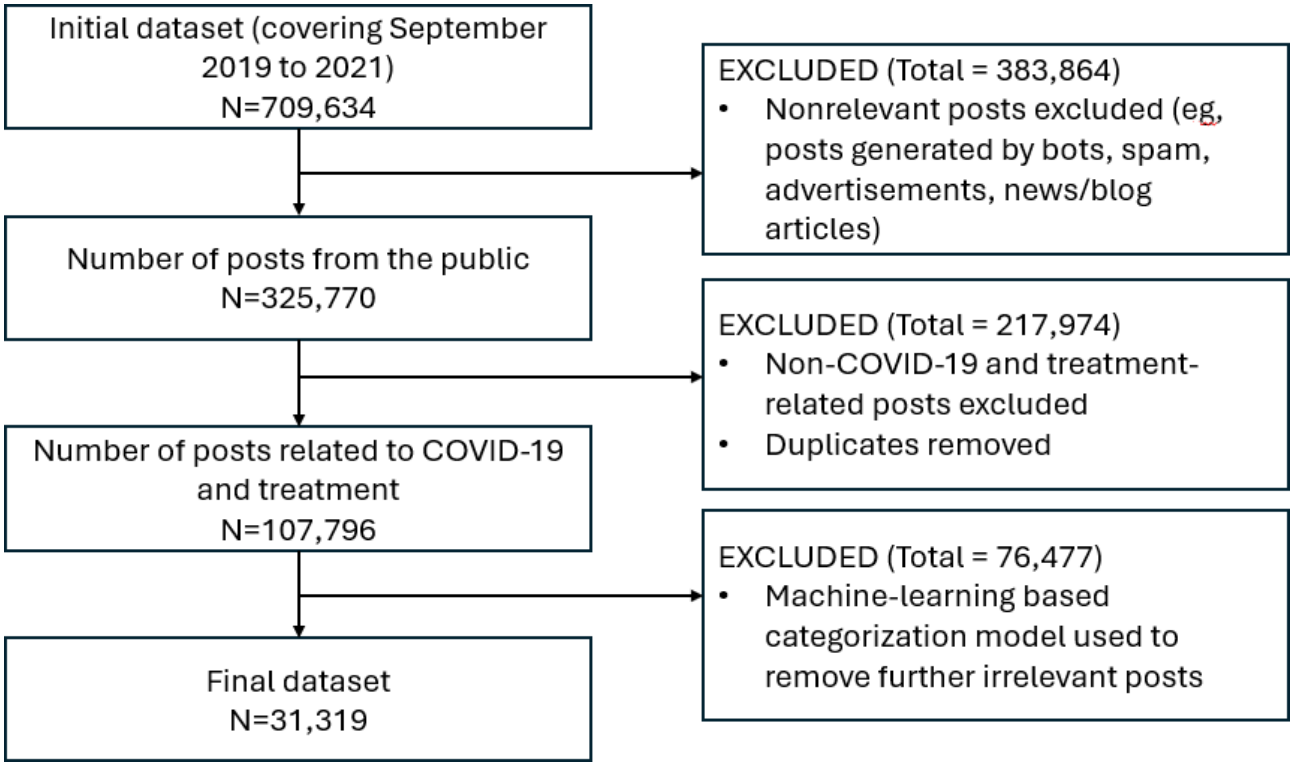

Overall Trends on COVID-19 Social Media Activity

The final dataset is primarily composed of data from Twitter (subsequently rebranded X), with a minimal contribution from forums and other social media platforms such as Facebook, Instagram or YouTube.

Trends in social media discussions closely aligned with Google search trends. Social media posts emerged in December 2019, with the highest peak (107,796 posts) in March 2020, coinciding with the first UK lockdown [4]. Discussion volume decreased in late 2020 but spiked in response to COVID-19 waves, new variants, and treatment announcements (Figure 2). Synthesio data trends closely mirrored Google search patterns (Figure 2).

Notably, the Omicron variant was more prominent in social media discussions (3824 posts, 12.0%) compared to Delta (894 posts, 2.9%) (Figure S1 in Multimedia Appendix 1).

Figure 2. Total volume of posts around COVID-19 (Synthesio) and total volume of searches on COVID-19 (Google search). PM: Prime Minister.

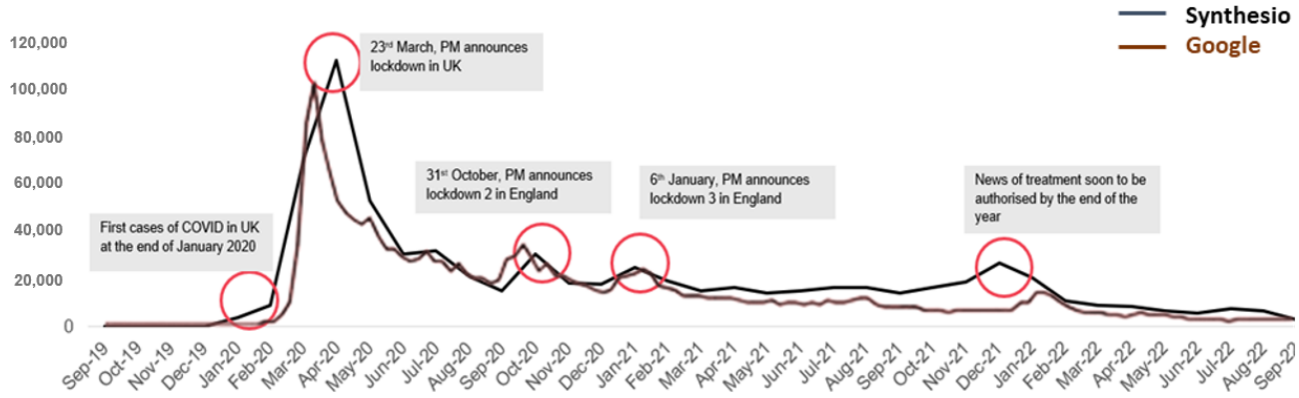

The Patient Journeys

Thematic analysis was performed on the cleaned dataset (n=31,319) to further examine the search trends and volume on predetermined stages of the treatment pathway, that is, symptoms → testing → diagnosis → treatment.

Symptoms

Discussions about COVID-19 symptoms were prevalent, with 49% of posts mentioning general symptoms, with 17% specifically referring to "COVID-19 symptoms." Frequently mentioned symptoms included cough, fever, fatigue, pain, anxiety, breathing difficulties, and respiratory issues [24].

Testing

Conversations about COVID-19 testing accounted for 35% of discussions, with a primary focus on test accessibility, long wait times, and delays in receiving polymerase chain reaction (PCR) test results. Sentiment toward testing shifted over time. Initially, discussions were largely neutral, but sentiment became positive in July 2021 as PCR and rapid tests became widely available. However, a peak in negative sentiment occurred in March 2022,

XSL•FO
RenderX

coinciding with the discontinuation of free rapid testing in the United Kingdom (Figure 3).

Figure 3. Sentiment trend of COVID-19 testing.

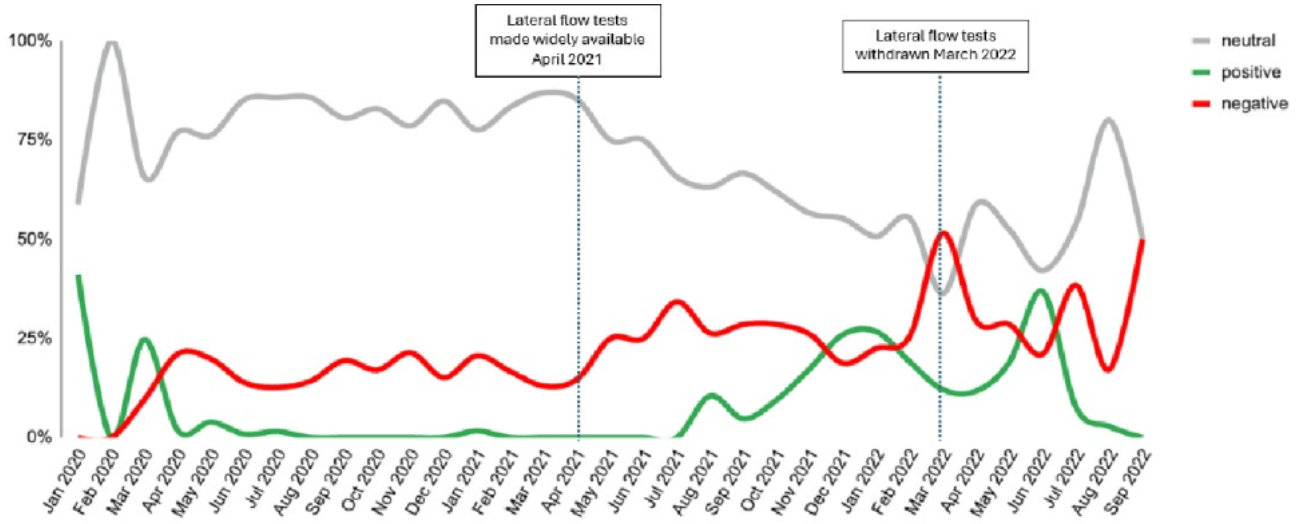

Diagnosis

The topic of COVID-19 diagnosis appeared in 25% of posts, with discussions primarily centered on test results and subsequent treatment options. Conversations about diagnosis peaked during major pandemic waves in October 2020, January 2021, and December 2021. Over time, sentiment became more positive, increasing from 2% to 19% by late 2021. However, negative sentiment remained prominent, ranging from 22% to 36% (Figure 4).

Figure 4. Sentiment trend of COVID-19 diagnosis.

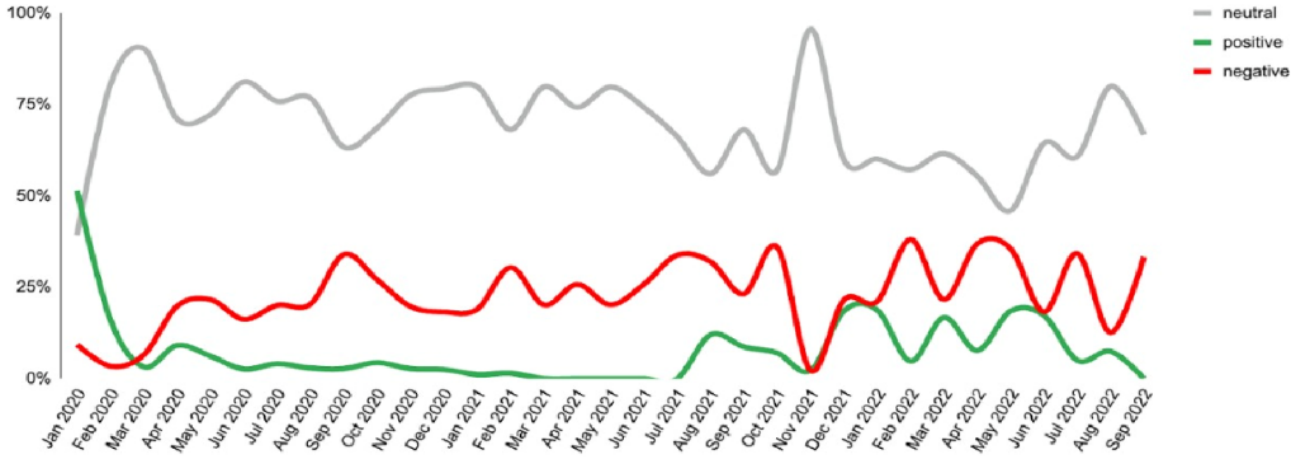

Treatment

Discussions about the treatment made up 32% of the dataset, often referring to medications in general terms such as “antiviral” (20% of total posts on treatment) or “medications” (39% of total posts on treatment). Misinformation was evident, as terms like “antimalarial” and “antibiotics” were frequently mentioned in the context of COVID-19 treatment (15% and 4% of total posts on treatment, respectively).

The analysis of top-searched terms and questions related to COVID-19 treatment on Google showed that eligibility for new COVID-19 treatment ranked highest, reflecting public interest in access to emerging therapies (Table S3 in Multimedia Appendix 1).

The volume of treatment-related discussions peaked in December 2021, following announcements from the UK National Health Service regarding treatment availability. Sentiment analysis indicated that discussions were mostly negative (9% - 12%) until July 2021. From August 2021 onward, positive and negative sentiment balanced out, despite the availability of antiviral treatments from April 2022 [25] (Figure 5).

XSL•FO
RenderX

Figure 5. Sentiment trend of COVID-19 treatment.

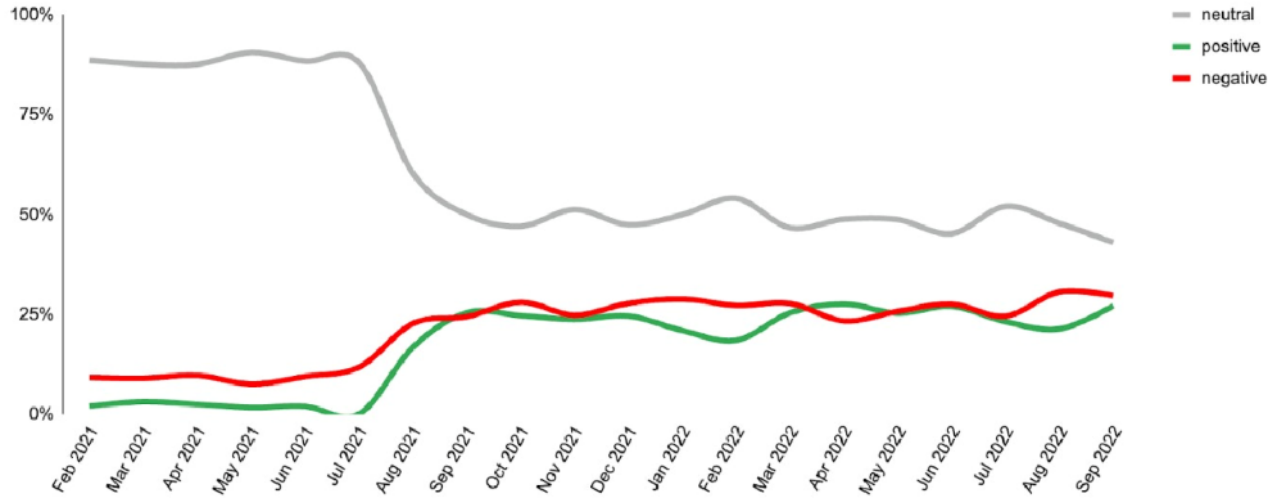

Discussion

Principal Findings

Our study examined public perceptions of COVID-19 symptoms, testing, diagnosis, and treatment through social media between September 2019 and September 2022. Public interest fluctuated, with peaks occurring during lockdowns, variant outbreaks, and treatment availability. The analysis of public perception of COVID-19 over time showed that the public attention fluctuated, with peaks occurring when lockdowns were introduced and when new variants emerged. The COVID-19 pandemic created an unparalleled sense of uncertainty and widespread health anxiety across the globe [26]. Despite or perhaps due to the vast amount of available information, uncertainty remains regarding the characteristics of the virus, as well as the emergence of new waves and variants. Studies show that information-seeking and knowledge-sharing practices increase rapidly in response to emergencies: since infectious diseases such as HIV or AIDS, acute respiratory syndrome, and Asian bird flu have emerged, the need for information by the public continues to increase [27-30]. Since social media is more frequently used than other media sources to acquire information on COVID-19, the government should have guidelines in place to ensure that reliable and accurate digital health information reaches diverse populations and those with lower levels of education.

Our findings show that COVID-19 symptoms represented 6% of social media expressions. People shared and sought validation for their symptoms, possibly reflecting anxiety, uncertainty, or a need for peer assurance. Posts related to symptoms ranged from less severe, like a cold, to potentially more severe, such as high fever, difficulty in breathing, and fatigue. Searches related to symptoms were closely linked to severity and age. In another study, a survey of over 2000 adults highlighted the awareness of main symptoms but not of others including skin rash, muscle and body aches, diarrhea, headache, nausea, and vomiting. Younger age groups in particular were least aware of COVID symptoms [31]. This lack of awareness may have led to increased anxiety and fear among the public, and to some extent a positive change in people's health behaviors toward preventive measures [32].

Discussions about testing were common. The sentiment toward testing was neutral through the pandemic but turned positive from July 2021 and then became negative in the United Kingdom in March 2022. Diagnostic tests played a crucial role in COVID-19 management worldwide [33]. The public encountered challenges in accessing tests for diagnosis of COVID-19 infection particularly in the earlier days of the pandemic. Public frustration with testing delays was further aggravated as the pandemic progressed and test sites became overwhelmed. Major challenges encountered by health authorities included addressing the ever-increasing demand for testing and real-time PCR limitations (including being resource-intensive and time-consuming), which may have been the main source of public frustration with the test. Despite anxiety with regards to testing, social media activity eased during mid-2021, and public sentiment to COVID-19 testing also started turning positive. This aligned with the second phase in the evolution of the COVID-19 testing strategy which occurred several months into the pandemic, making diagnosis more affordable, easier, and faster than molecular tests [34].

The increase in negative sentiment in March 2022 could have been due to stopping free rapid COVID-19 tests in the United Kingdom. Ending free access to COVID-19 rapid tests could have contributed to health inequality, disproportionately affecting low-income and marginalized communities who may struggle to afford testing. This has the possibility of creating barriers to early detection and isolation, increasing the risk of transmission, particularly in high-risk settings [35]. Ensuring affordable, widespread testing is crucial for early intervention and outbreak control. Investing in community-based diagnostics, digital health solutions, and stronger public health data integration can improve disease monitoring and promote health equity, supporting better preparedness for future crises.

The diagnostic-related posts were less frequent and mostly involved sharing test results, which may reflect the normalization of COVID-19 infection over time. As COVID-19 became an ongoing part of daily life, public concern around

XSL•FO
RenderX

diagnosis likely diminished, especially after the removal of widespread testing programs. With fewer policy changes and declining case severity, individuals who may have viewed COVID-19 as less urgent led to reduced engagement in discussions about diagnostic tools. This shift suggests that as disease becomes endemic, public interest in testing decreases, which has implications for maintaining awareness and ensuring continued access to diagnostic resources for effective disease monitoring.

Our analyses showed a large volume of queries about COVID-19 treatment, including its availability, effectiveness, and eligibility criteria. Initially, the public had a negative attitude toward possible side effects of antiviral treatments for COVID-19, and "rebound" was found to be the main concern for patients. Patients who tested positive for COVID-19 were also anxious regarding their eligibility to receive the available treatment. This concern was highlighted in the discussions due to the limited availability of the antiviral drugs, particularly for patients with mild-to-moderate symptoms. Difficulty accessing treatment and lack of access to information about treatment were also documented elsewhere as being particularly challenging for vulnerable communities in the United Kingdom throughout the pandemic, which may have influenced these sentiments [36].

Comparison With Prior Work

Our findings align with previous studies showing that social media trends mirror Google search behavior [37]. Similar studies have confirmed that digital platforms effectively capture real-time public interest in health topics [38-41]. However, social media also contributed to misinformation, highlighting the importance of public health communication strategies [42]. Of note, the UK government and WHO had collaborative campaigns and created a series of social media infographics and messages to explain the safety profile of COVID-19 vaccines [43]. In addition, the British Broadcasting Company launched a successful project called "Stop the Spread" during May and June 2020; the purpose of which was to increase the public's awareness of the magnitude of COVID-19 misinformation and motivate them to limit the spread and associated consequences.

Limitations

One key limitation of this study is the demographic representation of social media users. Twitter data do not fully reflect the general population, and user characteristics, such as age, gender, and location could not be inferred. Although some geographic analysis was conducted, the findings did not show significant differences and were therefore not included in the study. Another limitation is the observational nature of the study. The comparison between social media trends and Google searches is descriptive rather than a statistical correlation analysis. To ensure clarity, we explicitly state this in the study to avoid any causal implications. Finally, there are constraints related to the data source. The use of Synthesio, a proprietary tool, limits the reproducibility of the study. In addition, while efforts were made to filter noise and spam, some irrelevant content may persist due to the inherent limitations of natural language processing techniques.

Future Work

The nature of internet-based conversation about COVID-19 is rapidly changing, providing insights regarding the public's thoughts and needs. Governments and public health authorities can leverage these insights use these insights to enhance crisis communication, combat misinformation, and promote vaccine uptake. By monitoring social media discussions, policy makers can identify emerging concerns and misconceptions, allowing them to tailor public health messaging accordingly. For example, during COVID-19 peaks, public uncertainty about symptoms, testing, and treatments could have been addressed through targeted social media campaigns to provide accurate information and alleviate confusion.

In addition, our findings highlight the role of social media in both spreading reliable information and amplifying misinformation. Health authorities should proactively track trending misinformation, implement fact-checking initiatives, and collaborate with platforms to promote credible sources. The success of initiatives like the WHO's "Stop the Spread" campaign underscores the potential impact of coordinated efforts to counteract health misinformation and enhance public understanding.

Social media insights can also help inform vaccine communication strategies by identifying public concerns and hesitations regarding treatments and vaccinations. By analyzing discussions and trends, policy makers can develop targeted campaigns that address specific fears, correct misunderstandings, and promote informed decision-making. Integrating social media insights with traditional public health surveillance systems can further strengthen early detection of emerging health threats. By leveraging these real-time insights, governments can proactively engage with the public, reduce uncertainty, and build trust during health crises.

Future research could also build upon and deepen the types of analysis we conducted for this study. For example, a more comprehensive sentiment analysis could be conducted delving into a deeper range of emotions beyond the positive, neutral, and negative categorization used in this research [44,45], or further exploring how the different proportions of post related to each journey theme changed over time in line with previous studies [45,46]. We also note that while we did not analyze social media platform content individually against Google activity, future similar types of research could further investigate differences by social media platforms, Google activity, or both to clarify which platforms bear the most similarities in trends, particularly regarding conversations around health and health emergencies.

Conclusions

Throughout our social listening analysis on several social media platforms, we had a promising overview of patients with COVID-19 journey from diagnosis to treatment. The study highlights the significant role of social media in shaping public perceptions of COVID-19 and underscores its potential as a tool for public health interventions. By leveraging social media analytics, policy makers can gain valuable real-time insights into public sentiment, allowing for more effective crisis

XSL•FO
RenderX

communication, misinformation management, and targeted health campaigns. Future public health strategies should incorporate social listening tools to identify trends, address misinformation, and foster trust through clear and responsive communication.

Acknowledgments

This study was sponsored by Pfizer. Editorial and medical writing support was provided by Rebecca West at Ipsos and was funded by Pfizer. This research was conceived by MA, AS, MC, KN, NB, and EJC. Project management, data collection, and analysis were led by KP and JG. MA led drafting of the manuscript.

Data Availability

The datasets generated during and/or analyzed during this study are not publicly available due to being conducted within Ipsos' proprietary social listening platform but are available from the corresponding author on reasonable request.

Authors' Contributions

This research was conceived by MA, AS, MC, KN, NB, and EJC. Project management, data collection, and analysis were led by KP and JG. MA led drafting of the manuscript.

Conflicts of Interest

MA, AS, MC, and NB are employees of Pfizer and hold Pfizer stock. KN was an employee of Pfizer at the time of writing this manuscript and hold Pfizer stock. JG and KP are employees of Ipsos and have no conflicts of interest to disclose. EJC is an employee of University of Bedfordshire and has no conflict of interest to disclose.

Multimedia Appendix 1
Supplementary Materials.
[DOCX File, 70 KB - formative_v9i1e63997_app1.docx]

References

1. WHO director-general's opening remarks at the media briefing on COVID-19. World Health Organization. 2020. URL: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020 [accessed 2025-06-17]
2. Wright O. Coronavirus: how the UK dealt with its first Covid case. BBC News. 2021. URL: https://www.bbc.co.uk/news/uk-england-55622386 [accessed 2025-06-17]
3. Mahase E. Covid-19: UK records first death, as world's cases exceed 100 000. BMJ 2020 Mar 6;368:m943. [doi: 10.1136/bmj.m943] [Medline: 32144096]
4. Johnson B. Prime minister's statement on coronavirus (COVID-19). Prime Minister's Office. 2020. URL: https://www.gov.uk/government/speeches/pm-address-to-the-nation-on-coronavirus-23-march-2020 [accessed 2025-06-17]
5. Onyeaka H, Anumudu CK, Al-Sharify ZT, Egele-Godswill E, Mbaegbu P. COVID-19 pandemic: a review of the global lockdown and its far-reaching effects. Sci Prog 2021;104(2):368504211019854. [doi: 10.1177/00368504211019854] [Medline: 34061685]
6. Berg SH, O'Hara JK, Shortt MT, et al. Health authorities' health risk communication with the public during pandemics: a rapid scoping review. BMC Public Health 2021 Jul 15;21(1):1401. [doi: 10.1186/s12889-021-11468-3] [Medline: 34266403]
7. Abba-Aji M, Stuckler D, Galea S, McKee M. Ethnic/racial minorities' and migrants' access to COVID-19 vaccines: a systematic review of barriers and facilitators. J Migr Health 2022;5:100086. [doi: 10.1016/j.jmh.2022.100086] [Medline: 35194589]
8. Abbas J, Wang D, Su Z, Ziapour A. The role of social media in the advent of COVID-19 pandemic: crisis management, mental health challenges and implications. Risk Manag Healthc Policy 2021;14:1917-1932. [doi: 10.2147/RMHP.S284313] [Medline: 34012304]
9. Kemp S. Digital 2020: global digital yearbook. DATAREPORTAL. 2020. URL: https://datareportal.com/reports/digital-2020-global-digital-yearbook [accessed 2025-06-17]
10. Fang J, Wang X, Wen Z, Zhou J. Fear of missing out and problematic social media use as mediators between emotional support from social media and phubbing behavior. Addict Behav 2020 Aug;107:106430. [doi: 10.1016/j.addbeh.2020.106430]
11. Zarocostas J. How to fight an infodemic. Lancet 2020 Feb 29;395(10225):676. [doi: 10.1016/S0140-6736(20)30461-X] [Medline: 32113495]
12. Cheng T, Han B, Liu Y. Exploring public sentiment and vaccination uptake of COVID-19 vaccines in England: a spatiotemporal and sociodemographic analysis of Twitter data. Front Public Health 2023;11:1193750. [doi: 10.3389/fpubh.2023.1193750] [Medline: 37663835]

XSL•FO
RenderX

13. Lanyi K, Green R, Craig D, Marshall C. COVID-19 vaccine hesitancy: analysing Twitter to identify barriers to vaccination in a low uptake region of the UK. Front Digit Health 2021;3:804855. [doi: 10.3389/fdgth.2021.804855] [Medline: 35141699]
14. Fazel S, Zhang L, Javid B, Brikell I, Chang Z. Harnessing Twitter data to survey public attention and attitudes towards COVID-19 vaccines in the UK. Sci Rep 2021 Dec 14;11(1):23402. [doi: 10.1038/s41598-021-02710-4] [Medline: 34907201]
15. Jones LF, Bonfield S, Farrell J, Weston D. Understanding the public's attitudes toward COVID-19 vaccines in Nottinghamshire, United Kingdom: qualitative social media analysis. J Med Internet Res 2023 Mar 29;25(25):e38404. [doi: 10.2196/38404] [Medline: 36812390]
16. Ainley E, Witwicki C, Tallett A, Graham C. Using Twitter comments to understand people's experiences of UK health care during the COVID-19 pandemic: thematic and sentiment analysis. J Med Internet Res 2021 Oct 25;23(10):e31101. [doi: 10.2196/31101] [Medline: 34469327]
17. Marshall C, Lanyi K, Green R, Wilkins GC, Pearson F, Craig D. Using natural language processing to explore mental health insights from UK tweets during the COVID-19 pandemic: infodemiology study. JMIR Infodemiology 2022;2(1):e32449. [doi: 10.2196/32449] [Medline: 36406146]
18. See what you are missing. Synthesio. URL: https://www.synthesio.com/ [accessed 2025-06-17]
19. Python language reference, version 3.9. Python Software Foundation. URL: https://www.python.org/ [accessed 2025-06-29]
20. Saunders CH, Sierpe A, von Plessen C, et al. Practical thematic analysis: a guide for multidisciplinary health services research teams engaging in qualitative analysis. BMJ 2023 Jun 8;381:e074256. [doi: 10.1136/bmj-2022-074256] [Medline: 37290778]
21. AnswerThePublic: that free visual keyword research & content ideas tool. AnswerThePublic. URL: https://answerthepublic.com/ [accessed 2025-06-29]
22. Google trends. Google. URL: https://trends.google.com/trends/ [accessed 2025-06-29]
23. Ahrefs - SEO tools & resources to grow your search traffic. Ahrefs. URL: https://ahrefs.com/ [accessed 2025-06-29]
24. Coronavirus (COVID-19) infection survey, UK statistical bulletins. Office for National Statistics, UK. URL: https://tinyurl.com/38pf3hnr [accessed 2025-06-17]
25. Highest risk COVID-19 patients receive 'brilliant' new antivirals at home. National Health Service. URL: https://www.england.nhs.uk/2022/04/highest-risk-covid-19-patients-receive-brilliant-new-antivirals-at-home/ [accessed 2025-06-17]
26. Autenrieth LK, Asselmann E, Melzig CA, Benke C. Fear of COVID-19 predicts increases in anxiety, depressive symptoms, health anxiety, psychosocial distress, and loneliness: findings from a prospective two-year follow-up study. J Psychiatr Res 2024 Sep;177:162-168. [doi: 10.1016/j.jpsychires.2024.07.018] [Medline: 39024740]
27. Groseclose SL, Buckeridge DL. Public health surveillance systems: recent advances in their use and evaluation. Annu Rev Public Health 2017 Mar 20;38:57-79. [doi: 10.1146/annurev-publhealth-031816-044348] [Medline: 27992726]
28. LaPelle NR, Luckmann R, Simpson EH, Martin ER. Identifying strategies to improve access to credible and relevant information for public health professionals: a qualitative study. BMC Public Health 2006 Apr 5;6:89. [doi: 10.1186/1471-2458-6-89] [Medline: 16597331]
29. Nölke L, Mensing M, Krämer A, Hornberg C. Sociodemographic and health-(care-)related characteristics of online health information seekers: a cross-sectional German study. BMC Public Health 2015 Jan 29;15:31. [doi: 10.1186/s12889-015-1423-0] [Medline: 25631456]
30. Wagner AC, Hart TA, McShane KE, Margolese S, Girard TA. Health care provider attitudes and beliefs about people living with HIV: initial validation of the Health Care Provider HIV/AIDS Stigma Scale (HPASS). AIDS Behav 2014 Dec;18(12):2397-2408. [doi: 10.1007/s10461-014-0834-8] [Medline: 24965675]
31. Jones DL, Rhymes JM, Green E, et al. Poor air passenger knowledge of COVID-19 symptoms and behaviour undermines strategies aimed at preventing the import of SARS-CoV-2 into the UK. Sci Rep 2023 Mar 1;13(1):3494. [doi: 10.1038/s41598-023-30654-4] [Medline: 36859503]
32. Montazeri A, Mohammadi S, M Hesari P, et al. Exposure to the COVID-19 news on social media and consequent psychological distress and potential behavioral change. Sci Rep 2023 Sep 14;13(1):15224. [doi: 10.1038/s41598-023-42459-6] [Medline: 37710006]
33. Rai P, Kumar BK, Deekshit VK, Karunasagar I, Karunasagar I. Detection technologies and recent developments in the diagnosis of COVID-19 infection. Appl Microbiol Biotechnol 2021 Jan;105(2):441-455. [doi: 10.1007/s00253-020-11061-5] [Medline: 33394144]
34. Mathew BJ, Vyas AK, Khare P, et al. Laboratory diagnosis of COVID-19: current status and challenges. Iran J Microbiol 2021 Feb;13(1):1-7. [doi: 10.18502/ijm.v13i1.5485] [Medline: 33889356]
35. Batista C, Hotez P, Amor YB, et al. The silent and dangerous inequity around access to COVID-19 testing: a call to action. EClinicalMedicine 2022 Jan;43:101230. [doi: 10.1016/j.eclinm.2021.101230] [Medline: 34927038]
36. Experiences of accessing healthcare services during Covid-19. NHS. URL: https://notts.icb.nhs.uk/wp-content/uploads/2021/05/Integrated-Recovery-Engagement-Report-v1.pdf [accessed 2025-06-17]
37. Shin SY, Seo DW, An J, et al. High correlation of Middle East respiratory syndrome spread with Google search and Twitter trends in Korea. Sci Rep 2016 Sep 6;6:32920. [doi: 10.1038/srep32920] [Medline: 27595921]
38. Liu PL. COVID-19 information seeking on digital media and preventive behaviors: the mediation role of worry. Cyberpsychol Behav Soc Netw 2020 Oct;23(10):677-682 [FREE Full text] [doi: 10.1089/cyber.2020.0250] [Medline: 32498549]

XSL•FO
RenderX

39. Parikh PA, Shah BV, Phatak AG, et al. COVID-19 pandemic: knowledge and perceptions of the public and healthcare professionals. Cureus 2020 May 15;12(5):e8144. [doi: 10.7759/cureus.8144] [Medline: 32550063]
40. Rouvinen H, Turunen H, Lindfors P, et al. Online health information-seeking behaviour and mental well-being among Finnish higher education students during COVID-19. Health Promot Int 2023 Dec 1;38(6):daad143. [doi: 10.1093/heapro/daad143] [Medline: 37935171]
41. Neely S, Eldredge C, Sanders R. Health information seeking behaviors on social media during the COVID-19 pandemic among American social networking site users: survey study. J Med Internet Res 2021 Jun 11;23(6):e29802. [doi: 10.2196/29802] [Medline: 34043526]
42. Joseph AM, Fernandez V, Kritzman S, et al. COVID-19 misinformation on social media: a scoping review. Cureus 2022 Apr;14(4):e24601. [doi: 10.7759/cureus.24601] [Medline: 35664409]
43. Fighting misinformation in the time of COVID-19, one click at a time. World Health Organization. 2021. URL: https://www.who.int/news-room/feature-stories/detail/fighting-misinformation-in-the-time-of-covid-19-one-click-at-a-time [accessed 2025-06-17]
44. Medford RJ, Saleh SN, Sumarsono A, Perl TM, Lehmann CU. An “Infodemic”: leveraging high-volume Twitter data to understand early public sentiment for the coronavirus disease 2019 outbreak. Open Forum Infect Dis 2020 Jul;7(7):ofaa258. [doi: 10.1093/ofid/ofaa258] [Medline: 33117854]
45. Lyu JC, Han EL, Luli GK. COVID-19 vaccine-related discussion on Twitter: topic modeling and sentiment analysis. J Med Internet Res 2021 Jun 29;23(6):e24435. [doi: 10.2196/24435] [Medline: 34115608]
46. Kobayashi R, Takedomi Y, Nakayama Y, et al. Evolution of public opinion on COVID-19 vaccination in Japan: large-scale Twitter data analysis. J Med Internet Res 2022 Dec 22;24(12):e41928. [doi: 10.2196/41928] [Medline: 36343186]

Abbreviations

CNN: convolutional neural network
NER: named entity recognition
PCR: polymerase chain reaction

Edited by A Mavragani; submitted 05.07.24; peer-reviewed by N Thakur, R Kobayashi; revised version received 12.05.25; accepted 14.05.25; published 30.07.25.

Please cite as:
Araghi M, Sahota A, Czachorowski M, Naicker K, Bohm N, Phillipps K, Gaddum J, Cook EJ
Analysis of Social Media Perceptions During the COVID-19 Pandemic in the United Kingdom: Social Listening Study (2019-2022)
JMIR Form Res 2025;9:e63997
URL: https://formative.jmir.org/2025/1/e63997
doi:10.2196/63997

© Marzieh Araghi, Arron Sahota, Maciej Czachorowski, Kevin Naicker, Natalie Bohm, Katie Phillipps, James Gaddum, Erica Jane Cook. Originally published in JMIR Formative Research (https://formative.jmir.org), 30.7.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Economic Deterioration and Social Factors Affecting Mental Health During the COVID-19 Pandemic in Japan: Web-Based Cross-Sectional Survey

Kentaro Hori[1*]; Yosuke Yamada[2*]; Hideyuki Namba[3*]; Misaka Kimura[4*]; Hiroyuki Fujita[5*]; Heiwa Date[1*]

[1]Faculty of Data Science, Shiga University, 1-1-1 Banba, Hikone City, Japan

[2]Department of Physical Activity Research, National Institutes of Biomedical Innovation, Health and Nutrition, Settsu City, Japan

[3]Sports and Health Education Division, Center for Education in Liveral Arts and Sciences, Osaka University, Toyonakashi City, Japan

[4]Institute for Active Health, Kyoto University of Advanced Science, Kyoto City, Japan

[5]Faculty of Bioenvironmental Science, Kyoto University of Advanced Science, Kyoto City, Japan

[*]all authors contributed equally

Corresponding Author:
Kentaro Hori
Faculty of Data Science, Shiga University, 1-1-1 Banba, Hikone City, Japan

Abstract

Background: The socioeconomic impact of the COVID-19 pandemic has severely affected individuals' mental health. However, the factors that mitigate or exacerbate the mental health effects of economic deterioration remain underexplored.

Objective: This paper analyzes survey data from the second wave of the COVID-19 pandemic in Japan, a period during which women workers were reported to be economically and psychologically vulnerable. The analysis examined factors that mitigate or amplify the impact of COVID-19-induced economic deterioration on mental health, testing 3 hypotheses based on the conservation of resources theory and the stress buffering model: the negative impact of economic deterioration on mental health is greater for individuals with less social support compared to those with more social support (hypothesis 1); the negative impact of economic deterioration on mental health is greater for individuals experiencing more negative interactions compared to those experiencing fewer (hypothesis 2); and the buffering effect of social support is stronger in women than in men, with women receiving less social support experiencing greater mental health impacts from economic deterioration (hypothesis 3).

Methods: A web-based survey was conducted by an internet research company in Japan from June to July 2020. A balanced sample of 250 men and 250 women was recruited from each of the following age groups: 20-29, 30-39, 40-49, 50-59, 60-69, and 70-79 years. The analysis focused on working men and women aged 20 - 50 years (n=1238). Psychological distress was measured using the K6 scale. Economic deterioration was defined as a decrease in income compared to the prepandemic levels, and scales for social support and negative interactions were included. Logistic regression analysis was performed, using K6≥9 as the dependent variable, with interaction terms for each hypothesis sequentially incorporated.

Results: In the best-fitting model determined by the Bayesian Information Criterion, a significant association was observed between the interaction of COVID-19-induced economic deterioration and social support with K6 scores (odds ratio [OR] 0.90, 95% CI 0.81 - 0.99). However, in other models, the interaction between economic deterioration and negative interactions (OR 1.01, 95% CI 0.90 - 1.13) as well as the 3-way interaction involving economic deterioration, social support, and gender (OR 1.13, 95%CI 0.92 - 1.39) were not significant. The average marginal effect of economic deterioration was statistically significant for social support scores ranging from 4 to 10. The average marginal effect was 0.11 when social support was 4 (95% CI 0.03 - 1.20; *P*=.009) and 0.028 when social support was 10 (95% CI 0.00 - 0.06; *P*=.047).

Conclusions: The adverse impact of economic deterioration on mental health was more pronounced among individuals with lower levels of social support. These findings support hypothesis 1.

(JMIR Form Res 2025;9:e65204) doi:10.2196/65204

KEYWORDS

COVID-19; mental health; economic deterioration; social support; Japan

XSL•FO
RenderX

Introduction

The spread of COVID-19 forced people to change their lifestyles, raising significant concerns about mental health deterioration. Those who faced unemployment or had a history of chronic or mental illnesses were particularly susceptible to worsening mental health [1-5]. The socioeconomic downturn caused by the pandemic contributed to an increase in suicides, with unemployment, economic insecurity, and poverty having profound effects [2,3,5]. In Japan, where the suicide rate was already high before the pandemic, the number of suicides had been decreasing but began to rise again after the outbreak [6]. Analyses of suicide and depression risk factors in Japan identified economic deterioration from socioeconomic stagnation as a key contributor [7,8].

During the pandemic, individuals were exposed to negative events such as economic decline, while restrictions on interpersonal relationships reduced available social support, adversely affecting mental health [1,9-12]. For example, mothers raising children faced increased stress due to limited contact with grandparents, a crucial source of support [2]. Conversely, individuals with stronger social support were better able to manage loneliness under these restrictions [13,14], alleviate anxiety about the pandemic among pregnant women [15], and mitigate the impact of work-related stress on mental health [16].

Negative interactions, such as domestic violence during stay-at-home orders [17] and deteriorating relationships with family or partners [18], also emerged, highlighting new challenges in close relationships. Previous studies have revealed that such interactions negatively impacted mental health [19,20].

As outlined, previous studies have demonstrated that social support and negative interactions directly influenced mental health during the pandemic. However, the role of factors mitigating or exacerbating the negative impact of economic deterioration on mental health remains insufficiently explored, particularly in the context of limited interpersonal interactions. Social support and negative interactions, as discussed below, may serve as such factors.

Incorporating the conservation of resources hypothesis and the stress buffering model may provide new insights into social support during periods of widespread infection. Hobfoll's conservation of resources hypothesis defines resources as elements valued by individuals in society, including material resources (eg, money and housing), social resources (eg, social support and status), and psychological resources (eg, personal achievement and autonomy). The hypothesis posits that people feel stress when they lose or face the potential loss of valued resources, as these are essential for daily life and goal achievement [21-23]. For example, loss of income or employment (material resources) threatens stability and induces anxiety and stress. A study in Israel during the pandemic using the conservation of resources hypothesis found that individuals who experienced economic deterioration and loss of resources such as income and socioeconomic status showed declines in mental health [24].

In addition, the conservation of resources theory suggests that individuals facing resource loss may use other resources, such as social support, to cope and reduce stress. Social support, in particular, is a vital resource individuals rely on in emergencies to prevent further resource loss [21]. Those with stronger social support are better equipped to manage stress through assistance and information from others [22].

The stress buffering model provides a practical and statistical explanation for the benefits of social support. According to this model, social support not only has a direct positive effect on mental health but also serves as an indirect protective factor, alleviating stress and psychological burden during stressful events [25-29]. The model posits that social support helps resolve stressful situations and reduces the perception of stress through positive interpersonal relationships. Studies testing this model typically use interaction terms between stressful events and social support as explanatory variables [25,28,29].

Based on these theories, it can be anticipated that social support may mitigate the negative impact of economic deterioration on mental health during the COVID-19 pandemic. To verify this, it is necessary to analyze the interaction term between economic deterioration and social support. However, previous studies have focused primarily on the main effects of economic deterioration and social support, without considering their interaction [7,17,30].

Conversely, negative interactions tend to have the opposite effect of social support, with experiences such as criticism or excessive demands from others worsening mental health [26,31]. According to the conservation of resources theory, economic deterioration, which signifies the depletion of critical resources such as income, induces stress. Under such circumstances, negative interactions are likely to amplify the risk of mental health deterioration. However, studies on negative interactions during the pandemic have not analyzed the interaction between economic deterioration and negative interactions, leaving their synergistic effects unexamined [19,20].

This paper addresses these gaps by analyzing these interaction terms and testing the following hypotheses.

First, hypothesis 1: the negative impact of economic deterioration due to the COVID-19 pandemic on mental health is stronger for individuals with less social support compared to those with more social support.

Second, hypothesis 2: the negative impact of economic deterioration due to the COVID-19 pandemic on mental health is stronger for individuals with more negative interactions compared to those with fewer negative interactions.

Furthermore, this study examined hypotheses related to the high suicide rate among women during the pandemic in Japan. The first wave of the pandemic occurred in April 2020, followed by a larger second wave beginning in July 2020. During the second wave, women's mental health deteriorated significantly, with the female suicide rate increasing approximately 5 times that of men, marking a distinct shift [8,32,33]. This deterioration was partly attributed to the substantial economic damage in industries predominantly employing women, such as service, retail, and tourism [8,33]. Historically, the suicide rate among

XSL•FO
RenderX

Japanese women has not exceeded that of men, but during this period, women in the workforce may have been more likely than men to face economic and psychological vulnerability.

Drawing on the conservation of resources theory, female workers likely perceived income loss as a significant threat during the second wave. Among these women, those with strong social support may have been better equipped to cope with the stress of income loss, benefiting from its buffering effects. Therefore, this study used data from a survey conducted in Japan between June and July 2020, which covers the second wave. It analyzed a 3-way interaction by adding a gender variable to the interaction term of hypothesis 1 to test the third hypothesis: the buffering effect of social support is more pronounced in women than in men. Specifically, among women, the negative impact of economic deterioration due to the COVID-19 pandemic on mental health is stronger for those with less social support than for those with more social support.

This paper examined the 3 hypotheses above, focusing on the factors that mitigate or amplify the negative impact of economic deterioration caused by COVID-19 on mental health, and to clarify how social support and negative interactions function as such factors. This understanding is critical for developing targeted interventions and policies to support mental health during future crises.

Methods

Data Source

This study used primary data from a web-based survey conducted among residents of Japan. The survey was administered in a questionnaire format through Cross Marketing Inc, a Japanese internet research company. Cross Marketing Inc and its partner companies maintain an active panel of over 5 million people, who have registered their sociodemographic information in the company's database and responded to at least one survey within the past year [34].

Data collection was conducted from June to July 2020 using Cross Marketing Inc's web-based response system. The company provided respondents with a survey URL via email requesting their participation. First, individuals who agreed to participate underwent a screening survey regarding their demographic attributes, such as gender and age (n=19,301). Next, stratified sampling was performed, evenly allocating 250 men and 250 women from each age group (20s, 30s, 40s, 50s, 60s, and 70s), resulting in a total sample size of 3000 participants. This nonprobabilistic sampling was used to minimize data shortages for specific age and gender groups; however, it has limitations in ensuring representativeness. Respondents were selected in order of earliest response, and the survey continued until the target number of participants was reached, excluding inconsistent or invalid responses. The sample size was determined primarily based on feasibility and financial constraints.

In this study, economic deterioration was considered a key explanatory variable. In Japan, many individuals in their 20s to 50s work full-time and earn income, making them more susceptible to economic issues. Therefore, samples from individuals in their 60s and 70s were excluded, leaving a dataset of 2000 respondents.

Economic deterioration was analyzed based on changes in personal income, which required focusing on income earners. Respondents who had been unemployed before the pandemic and remained unemployed at the time of the survey (including full-time homemakers) were excluded. Those who were employed before the pandemic but unemployed during the survey period were also excluded due to their small number (n=17). These exclusions reduced the sample to 1513 respondents.

Following listwise deletion for missing values, the final dataset included 1238 respondents, with most exclusions resulting from non-responses to the household income question (n=254). In total, 18.2% (275/1513; 275 excluded) of the sample was excluded in this process.

Measurements

The K6 scale was used to assess mental health, serving as the dependent variable. This scale measures psychological distress over the past 30 days with 6 items rated on a 5-point scale, producing total scores from 0 to 24, with higher scores indicating greater distress [35]. Furukawa et al [36] validated the reliability of the K6 scale in the Japanese population. Following their methodology, a cutoff score of ≥9 was used to screen for mood and anxiety disorders.

The primary independent variables for hypothesis testing were economic deterioration due to COVID-19, social support, and negative interactions. Economic deterioration was measured by asking, "How has your income (total pre-tax income from work and non-work sources) changed as of April compared to before the COVID-19 pandemic? Please select the option that best describes your situation. If your salary is yet to be confirmed (eg, not yet deposited), please answer based on your expectations." The response options were: 1=increased (expected to increase), 2=slightly increased (expected to slightly increase), 3=almost unchanged, 4=slightly decreased (expected to slightly decrease), and 5=Decreased (expected to decrease). The response distribution was: 1 (n=21), 2 (n=37), 3 (n=735), 4 (n=197), and 5 (n=248). Categories 1 and 2 were combined with Category 3 and recorded as 0, while Categories 4 and 5 were combined and recorded as 1, resulting in binary variables. The correlation between economic deterioration and household income (a continuous variable) was weak (r=−0.15).

Social support was assessed using a 4-item, 4-point scale, with total scores ranging from 4 to 16, with higher scores indicating greater social support. The items were: "How much do your friends really care about you?" "How much do they understand your feelings?" "How much can you rely on them for help if you have a serious problem?" and "How much can you open up to them if you need to talk about your worries?" Response options were: 1=not at all, 2=a little, 3=some, and 4=a lot [37,38]. The Cronbach α was 0.90.

Negative interactions were measured using a 4-item, 4-point scale, with total scores ranging from 4 to 16, where higher scores indicated more negative interactions. The items were: "How often do your friends make too many demands on you?" "How

XSL•FO
RenderX

often do they criticize you?" "How often do they let you down when you are counting on them?" and "How often do they get on your nerves?" Response options were: 1=never, 2=rarely, 3=sometimes, and 4=often [37,38]. The Cronbach α was 0.84.

The control variables used in this study were sex (1=female and 0=male), age, marital status (1=married, 0=single, divorced, or widowed), education level (with 3 categories: university or graduate school; junior college, technical college, or specialized training school; and high school or middle school), last year's household income (log-transformed for analysis), employment conditions (regular employment, nonstandard employment, and self-employment), presence of chronic illness (1=no and 0=yes), and residential area (1=urban and 0=rural).

Ethical Considerations

This study was approved by the ethics committees of the Kyoto University of Advanced Science (KUAS 20‐3) and the National Institutes of Biomedical Innovation, Health, and Nutrition (NIBIOHN 202).

Informed consent was obtained on the research company Cross Marketing Inc's website before participants completed the survey. Participants were presented with an information sheet on the web-based survey landing page, and only those who confirmed that they had read the information sheet and agreed to participate were included in the study.

Before transferring the data to researchers, the research company removed participants' names, addresses, contact information, and any other details that could potentially be used to identify individuals.

Subjects registered with the research company could receive point rewards for participating in the survey (Cross Marketing Inc does not publicly disclose the number of points respondents receive for completing surveys).

Statistical Analyses

A logistic regression model was used to evaluate the relationship between economic deterioration, social support, negative interactions, and mental health. Control variables included sex, age, marital status, education level, last year's household income, employment conditions, presence of chronic illness, and residential area. In addition, interaction terms between economic deterioration and social support, as well as between economic deterioration and negative interactions, were added sequentially. A final model with a 3-way interaction term among economic deterioration, social support, and the female dummy variable was also examined. Statistical analyses were conducted using R (version 4.2.2; R Foundation for Statistical Computing). Listwise deletion was applied to data with missing values, yielding a final sample size of 1238 individuals.

Results

Table 1 presents the descriptive statistics, while Table 2 shows the results of the logistic regression analysis. In Model 1 of Table 2, the main effects of economic deterioration due to COVID-19, social support, and negative interactions were evaluated. The odds ratio for economic deterioration due to COVID-19 was 1.49 (95% CI 1.12‐1.98), for social support it was 0.96 (95% CI 0.91‐1.00), and for negative interactions it was 1.22 (95% CI 1.15‐1.29).

XSL•FO
RenderX

Table . Descriptive statistics.

Characteristics	Female (n=505, 40.8%)	Male (n=733, 59.2%)	Total (N=1238)
Age (years), n (%)			
20-39	131 (25.9)	154 (21)	285 (23)
30-39	125 (24.8)	192 (26.2)	317 (25.6)
40-49	129 (25.5)	193 (26.3)	322 (26)
50-59	120 (23.8)	194 (26.5)	314 (25.4)
Marital status, n (%)			
Married	273 (54.1)	384 (52.4)	657 (53.1)
Single, divorced, or widowed	232 (45.9)	349 (47.6)	581 (46.9)
Educational background, n (%)			
University or graduate school	221 (43.8)	470 (64.1)	691 (55.8)
Junior college, technical college, or specialized training school	156 (30.9)	109 (14.9)	265 (21.4)
High school or middle school	128 (25.3)	154 (21)	282 (22.8)
Household income (¥), n (%)[a]			
Less than 2 million	43 (8.5)	50 (6.8)	93 (7.5)
2 to 3.99 million	146 (28.9)	149 (20.3)	295 (23.8)
4 to 6.99 million	142 (28.1)	265 (36.2)	407 (32.9)
7 to 9.99 million	120 (23.8)	152 (20.7)	272 (22)
Over 10 million	54 (10.7)	117 (16)	171 (13.8)
Employment conditions, n (%)			
Regular employment	252 (49.9)	566 (77.2)	818 (66.1)
Nonstandard employment	231 (45.7)	109 (14.9)	340 (27.5)
Self-employment	22 (4.4)	58 (7.9)	80 (6.5)
Chronic disease, n (%)			
Yes	107 (21.2)	191 (26.1)	940 (75.9)
No	398 (78.8)	542 (73.9)	298 (24.1)
Residence area, n (%)			
Urban	306 (60.6)	443 (60.4)	749 (60.5)
Rural	199 (39.4)	290 (39.6)	489 (39.5)
Economic deterioration due to COVID-19, n (%)			
Worsened	198 (39.2)	247 (33.7)	445 (35.9)
Not worsened	307 (60.8)	486 (66.3)	793 (64.1)
Social support			
Mean (SD)	10.0 (3.0)	9.1 (2.7)	9.5 (2.9)
Median (range)	10 (4 - 16)	8 (4 - 16)	9 (4 - 16)
Negative interaction			
Mean (SD)	8.7 (2.6)	8.4 (2.3)	8.5 (2.8)
Median (range)	8.0 (4 - 16)	8.0 (4 - 16)	9.0 (4 - 16)
K6 scale, n (%)			
K6 ≥9	145 (28.7)	176 (24.0)	321 (25.9)
K6 <9	360 (71.3)	557 (76)	917 (74.1)

XSL•FO
RenderX

[a]The exchange rate for Japanese yen (¥) to US dollars (US $) during the data collection period (June to July 2020): US $1=¥110.

Table . Results of binomial logistic regression (K6 ≥9 as the dependent variable).[a]

	Model 1[b]	
	OR[c] (95% CI)	*P* value
Female (vs male)	1.13 (0.84-1.53)	.42
Age 20-29 (vs 50-59) years	3.90 (2.45- 6.28)	<.001
Age 30-39 (vs 50-59) years	2.96 (1.91- 4.63)	<.001
Age 40-49 (vs 50-59) years	2.12 (1.39-3.24)	<.001
Married (vs single, divorced, or widowed)	1.29 (0.94-1.78)	.12
University or graduate school (vs high school or middle school)	0.79 (0.55-1.13)	.19
Junior college, technical college, or specialized training school (vs high school or middle school)	0.86 (0.57-1.29)	.47
Household income	0.65 (0.52-0.81)	<.001
Nonstandard employment (vs regular employment)	1.04 (0.74-1.47)	.82
Self-employment (vs regular employment)	0.92 (0.50-1.64)	.79
No chronic disease (vs yes)	0.50 (0.36-0.69)	<.001
Urban (vs rural)	0.81 (0.61-1.08)	.15
Economic deterioration due to COVID-19 (vs not worsened)	1.49 (1.12-1.98)	.006
Social support	0.96 (0.91-1.00)	.08
Negative interaction	1.22 (1.15-1.29)	<.001

[a]The Bayesian Information Criterion (BIC) was 1378.15. After listwise deletion of missing data, 1238 respondents were included.

[b]The variance inflation factors (VIFs) ranged from 1.077 to 2.386 in Model 1.

[c]OR: adjusted odds ratio.

Table 3 and Table 4 present the results of the models with interaction terms. Model 2 includes the interaction term between economic deterioration due to COVID-19 and social support, Model 3 includes the interaction term between economic deterioration due to COVID-19 and negative interactions, Model 4 includes both interaction terms, and Model 5 includes the 3-way interaction term among economic deterioration due to COVID-19, social support, and the female dummy variable. The interaction between economic deterioration due to COVID-19 and social support was significant in Model 2 (OR 0.90, 95% CI 0.81 - 0.99) and Model 4 (OR 0.89, 95% CI 0.81 - 0.99). In Model 5, the 3-way interaction term was not significant. However, the interaction between economic deterioration due to COVID-19 and social support remained significant (OR 0.84, 95% CI 0.73 - 0.96), confirming its effect. A supplementary analysis, not shown in Tables 3 and 4, was conducted by adding the 3-way interaction term among economic deterioration due to COVID-19, negative interactions, and the female dummy variable to Model 5, which proved nonsignificant (OR 0.83, 95%CI 0.66 - 1.05; *P*=.12).

XSL•FO
RenderX

Table . Results of binomial logistic regression including interaction terms (K6 ≥9 as the dependent variable) for Model 2 and Model 3.[a]

Variables	Model 2[b]		Model 3[b]	
	OR[c] (95% CI)	*P* value	OR (95% CI)	*P* value
Female (vs male)	1.13 (0.83-1.52)	.44	1.13 (0.84-1.53)	.42
Age 20-29 (vs 50-59) years	3.95 (2.48-6.38)	<.001	3.90 (2.45-6.28)	<.001
Age 30-39 (vs 50-59) years	2.97 (1.92-4.66)	<.001	2.96 (1.91-4.63)	<.001
Age 40-49 (vs 50-59) years	2.11 (1.39-3.24)	.001	2.11 (1.39-3.24)	<.001
Married (vs single, divorced, or widowed)	1.29 (0.94-1.78)	.12	1.29 (0.94-1.78)	.12
University or graduate school (vs high school or middle school)	0.77 (0.54-1.11)	.17	0.79 (0.55-1.13)	.19
Junior college, technical college, or specialized training school (vs high school or middle school)	0.86 (0.57-1.29)	.46	0.86 (0.57-1.29)	.46
Household income	0.66 (0.52-0.82)	<.001	0.65 (0.52-0.81)	<.001
Nonstandard employment (vs regular employment)	1.04 (0.74-1.47)	.82	1.04 (0.74-1.47)	.82
Self-employment (vs regular employment)	0.95 (0.51-1.70)	.88	0.92 (0.50-1.64)	.79
No chronic disease (vs yes)	0.50 (0.36-0.68)	<.001	0.50 (0.36-0.69)	<.001
Urban (vs rural)	0.81 (0.60-1.08)	.14	0.81 (0.61-1.08)	.15
Economic deterioration due to COVID-19 (vs not worsened)	1.47 (1.11-1.96)	.008	0.96 (0.91-1.00)	.08
Social support	1.00 (0.94-1.06)	.94	1.48 (1.11-1.98)	.008
Negative interaction	1.22 (1.15-1.29)	.008	1.21 (1.13-1.30)	<.001
Economic deterioration due to COVID-19×social support	0.90 (0.81-0.99)	.03	—[d]	—
Economic deterioration due to COVID-19×negative interaction	—	—	1.01 (0.90-1.13)	.84

[a]The Bayesian Information Criterion (BIC) was 1380.62 for Model 2 and 1385.22 for Model 3. After listwise deletion of missing data, 1238 respondents were included for both models.

[b]The variance inflation factors (VIFs) ranged from 1.045 to 2.395 in Model 2, and 1.077 to 2.386 in Model 3.

[c]OR: adjusted odds ratio.

[d]Not applicable.

XSL•FO
RenderX

Table . Results of binomial logistic regression including interaction terms (K6 ≥9 as the dependent variable) for Model 4 and Model 5.[a]

Variables	Model 4[b]		Model 5[b]	
	OR[c] (95% CI)	*P* value	OR (95% CI)	*P* value
Female (vs male)	1.12 (0.83-1.52)	.45	1.02 (0.70-1.49)	.90
Age 20-29 (vs 50-59) years	3.95 (2.48-6.39)	<.001	4.02 (2.51-6.52)	<.001
Age 30-39 (vs 50-59) years	2.97 (1.92-4.66)	<.001	3.07 (1.97-4.84)	<.001
Age 40-59 (vs 50-59) years	2.11 (1.39-3.24)	.001	2.11 (1.39-3.25)	<.001
Married (vs single, divorced, or widowed)	1.29 (0.94-1.78)	.12	1.28 (0.93-1.78)	.13
University or graduate school (vs high school or middle school)	0.77 (0.54-1.11)	.16	0.75 (0.53-1.09)	.13
Junior college, technical college, or specialized training school (vs high school or middle school)	0.86 (0.57-1.28)	.45	0.85 (0.57-1.28)	.45
Household income	0.66 (0.52-0.82)	<.001	0.66 (0.52-0.83)	<.001
Nonstandard employment (vs regular employment)	1.04 (0.73-1.47)	.83	1.04 (0.73-1.47)	.83
Self-employment (vs regular employment)	0.96 (0.51-1.71)	.88	0.94 (0.50-1.69)	.84
No chronic disease (vs yes)	0.50 (0.36-0.69)	<.001	0.50 (0.36-0.70)	<.001
Urban (vs rural)	0.81 (0.60-1.08)	.15	0.81 (0.60-1.08)	.15
Economic deterioration due to COVID-19 (vs not worsened)	1.46 (1.09-1.95)	.01	1.29 (0.88-1.90)	.19
Social support	1.00 (0.94-1.06)	.95	1.11 (1.01-1.21)	.03
Negative interaction	1.21 (1.12-1.30)	<.001	1.21 (1.13-1.31)	<.001
Economic deterioration due to COVID-19×social support	0.89 (0.81-0.99)	.03	0.84 (0.73-0.96)	.01
Economic deterioration due to COVID-19×negative interaction	1.03 (0.92-1.15)	.67	1.02 (0.91-1.14)	.79
Economic deterioration due to COVID-19×female	—[d]	—	1.27 (0.72-2.25)	.41
Social support×female	—	—	0.81 (0.71-0.92)	.001

XSL•FO
RenderX

Variables	Model 4[b]		Model 5[b]	
	OR[c] (95% CI)	*P* value	OR (95% CI)	*P* value
Economic deterioration due to COVID-19×social support×female	—	—	1.13 (0.92-1.39)	.25

[a]The Bayesian Information Criterion (BIC) was 1387.55 for Model 4 and 1396.19 for Model 5. After listwise deletion of missing data, 1238 respondents were included for both models.

[b]The variance inflation factors (VIFs) ranged from 1.080 to 2.332 in Model 4, and 1.081 to 3.163 in Model 5.

[c]Odds ratio: adjusted odds ratio.

[d]Not applicable.

Finally, since the 2-way interaction term between economic deterioration due to COVID-19 and social support was significant, the average marginal effect of economic deterioration was calculated to elucidate this interaction. Model 2 demonstrated the best Bayesian Information Criterion, and thus, the average marginal effect was derived from this model.

The average marginal effects of economic deterioration due to COVID-19 are shown in Figure 1. Based on the 95% CIs, the average marginal effects were significant when the number of social support items ranged from 4 to 10. The average marginal effect was 0.11 when social support was 4 (95% CI 0.03 - 1.20; *P*=.009) and 0.028 when social support was 10 (95% CI 0.00 - 0.06; *P*=.047). This finding suggests that the negative impact of economic deterioration on mental health is amplified when social support is low.

Figure 1. Average marginal effects of economic deterioration due to COVID-19 with 95% CI.

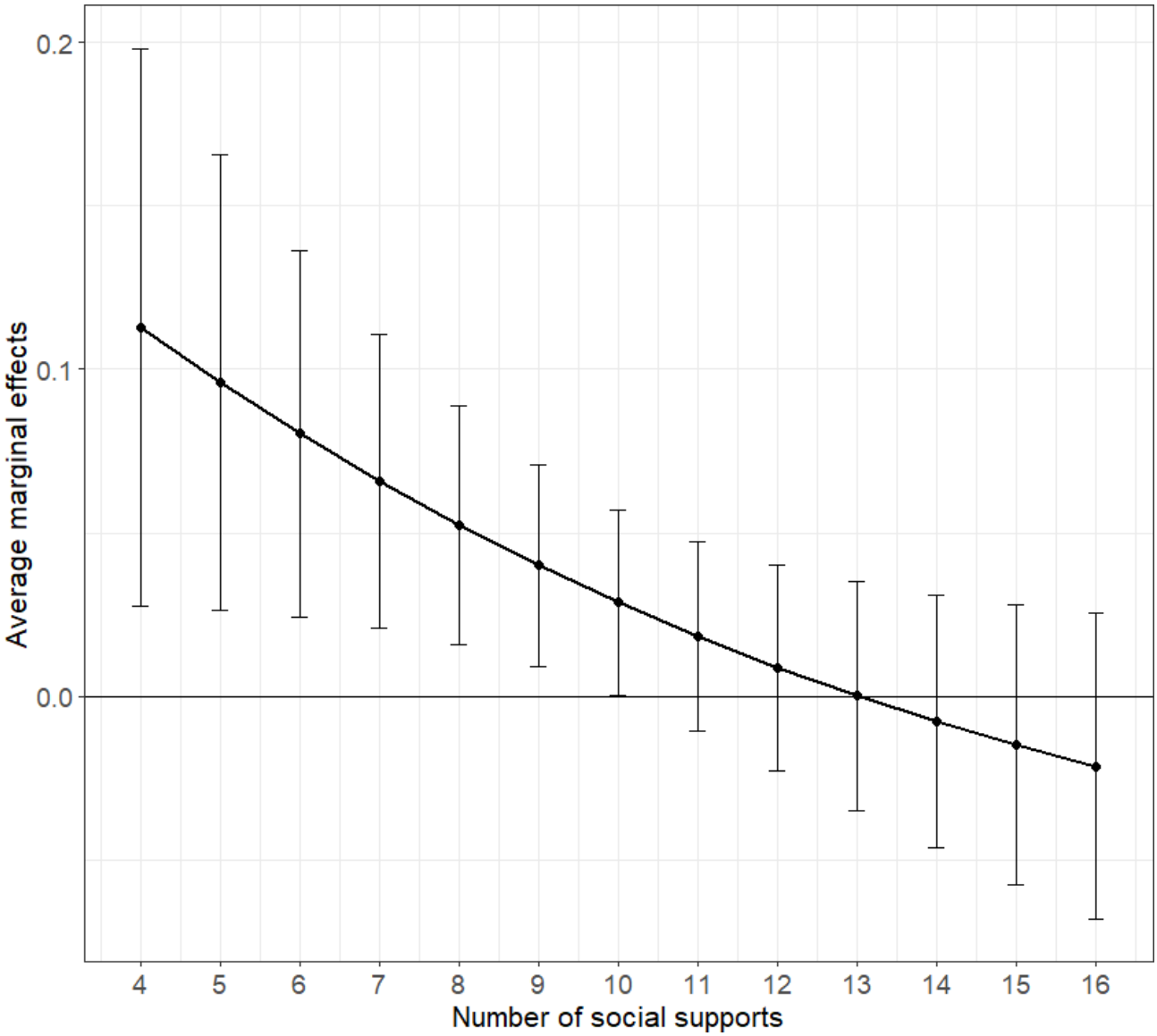

XSL•FO
RenderX

Discussion

Principal Findings

The primary finding of this study is that the negative impact of economic deterioration due to COVID-19 on mental health is amplified when social support is low, supporting hypothesis 1.

Previous studies have established that economic deterioration due to COVID-19 and social support affect mental health [2,7,14-17,30], but have examined the main effects of each variable. These analyses do not identify protective factors that mitigate the negative impact of economic deterioration. By contrast, this study suggests that social support functions as a protective factor. While social support has been widely recognized for its protective effects across various contexts [25-29], this study further confirms its role in mitigating the impact of economic deterioration during the pandemic.

The main effect of negative interactions was statistically significant in all models. Unlike previous studies [19,20], this study also analyzed the interaction between economic deterioration and negative interactions. However, this interaction term was not statistically significant, indicating that negative interactions do not exacerbate the negative impact of economic deterioration. Thus, hypothesis 2 was rejected. Nevertheless, the main effect of negative interactions remained consistently significant, aligning with previous studies [19,20] on the impact of negative interactions during the pandemic. Even before the pandemic, the influence of negative interactions was considered more significant than social support [31]. This suggests that the impact of negative interactions on mental health is consistently strong and should be recognized as a potent stressor, irrespective of the pandemic context.

In addition, building on previous studies that reported the economic and psychological vulnerability of female workers in Japan during the second wave of the COVID-19 pandemic [8,32,33], this study focused on women during this period. Specifically, drawing on the conservation of resources theory [21-23], hypothesis 3 was proposed to examine whether women with greater social support were better able to cope with this stressful event and benefit from its buffering effects. However, the 3-way interaction term required to test this hypothesis was not statistically significant, leading to its rejection.

By contrast, the persistent significance of the interaction between economic deterioration and social support suggests that the buffering effect of social support was effective regardless of gender. This result likely reflects the broader reality that, during this period, both men and women faced economic difficulties and social pressures, emphasizing the importance of social support for the mental health of both genders. In Japan, traditional gender roles, which position men as breadwinners and women as homemakers or caregivers, remain deeply ingrained [39]. These norms have long linked economic difficulties to suicide among working men in Japan, even before the outbreak [40,41]. During the pandemic, these issues continued to contribute to suicide rates [18], indicating that men faced significant pressures from both economic burdens and societal expectations tied to gender roles. Consequently, while women's mental health deteriorated significantly during the second wave [8,32,33], men also faced the risk of losing critical resources, such as income. Therefore, social support likely served as an essential buffer for maintaining the mental health of both men and women.

Limitations

This study has several limitations. First, its cross-sectional design restricted the ability to establish causal relationships. Individuals with preexisting poor mental health may struggle with work and interpersonal relationships, increasing their susceptibility to economic difficulties. As such, a bidirectional relationship, rather than a unidirectional causal link, should be considered.

Second, although the sample was drawn based on age and gender group allocations, it was limited to individuals willing to participate in a web-based survey, which may have introduced sampling bias. It is likely that individuals with severely poor health, less motivation to participate in web-based surveys, or heavy workloads and limited free time were underrepresented.

Third, this analysis excluded individuals who were employed before the pandemic but unemployed at the time of the survey, due to the small number of such individuals (n=17), which made meaningful analysis difficult. Therefore, people who lost their jobs due to the pandemic were not considered. Possible reasons for this exclusion include a lack of time or resources among the unemployed to participate or limited access to the internet. This study focused on individuals employed both before the pandemic and at the time of the survey, allowing it to demonstrate the severity of the mental health challenges faced by workers experiencing economic hardship.

Conclusion

No previous studies have specifically addressed the factors that mitigate or amplify the negative impact of COVID-19-induced economic deterioration on mental health or explored gender differences in these effects. This paper analyzed data from a Japanese survey conducted during the second wave of the pandemic, a time when women's mental health was believed to be in decline. The results revealed that the impact of economic deterioration was more pronounced when social support was limited, regardless of gender, and that negative interactions consistently showed significant main effects. These results suggest the ongoing need for targeted support for individuals with limited social support during societal crises such as a pandemic, irrespective of gender.

Data Availability

The data sets generated and analyzed during this study are not publicly available due to the terms of participant consent and data use agreements.

XSL•FO
RenderX

Conflicts of Interest

None declared.

References

1. Brooks SK, Webster RK, Smith LE, et al. The psychological impact of quarantine and how to reduce it: rapid review of the evidence. Lancet 2020 Mar 14;395(10227):912-920. [doi: 10.1016/S0140-6736(20)30460-8] [Medline: 32112714]
2. Fegert JM, Vitiello B, Plener PL, Clemens V. Challenges and burden of the coronavirus 2019 (COVID-19) pandemic for child and adolescent mental health: a narrative review to highlight clinical and research needs in the acute phase and the long return to normality. Child Adolesc Psychiatry Ment Health 2020;14(1):20. [doi: 10.1186/s13034-020-00329-3] [Medline: 32419840]
3. Moreno C, Wykes T, Galderisi S, et al. How mental health care should change as a consequence of the COVID-19 pandemic. Lancet Psychiatry 2020 Sep;7(9):813-824. [doi: 10.1016/S2215-0366(20)30307-2] [Medline: 32682460]
4. Vindegaard N, Benros ME. COVID-19 pandemic and mental health consequences: systematic review of the current evidence. Brain Behav Immun 2020 Oct;89:531-542. [doi: 10.1016/j.bbi.2020.05.048] [Medline: 32485289]
5. Pathirathna ML, Nandasena H, Atapattu A, Weerasekara I. Impact of the COVID-19 pandemic on suicidal attempts and death rates: a systematic review. BMC Psychiatry 2022 Jul 28;22(1):506. [doi: 10.1186/s12888-022-04158-w] [Medline: 35902951]
6. Oka M. Why did the suicide rate for Japanese women rise so rapidly in 2020: understanding regional and gender differences in the rise of suicide rates and a study of the causes. Occup Ment Health 2023;31(1):36-41. [doi: 10.57339/jjomh.31.1_36]
7. Fukase Y, Ichikura K, Murase H, Tagaya H. Depression, risk factors, and coping strategies in the context of social dislocations resulting from the second wave of COVID-19 in Japan. BMC Psychiatry 2021 Jan 12;21(1):33. [doi: 10.1186/s12888-021-03047-y] [Medline: 33435930]
8. Tanaka T, Okamoto S. Increase in suicide following an initial decline during the COVID-19 pandemic in Japan. Nat Hum Behav 2021 Feb;5(2):229-238. [doi: 10.1038/s41562-020-01042-z] [Medline: 33452498]
9. Saltzman LY, Hansel TC, Bordnick PS. Loneliness, isolation, and social support factors in post-COVID-19 mental health. Psychol Trauma 2020 Aug;12(S1):S55-S57. [doi: 10.1037/tra0000703] [Medline: 32551762]
10. Blake H, Bermingham F, Johnson G, Tabner A. Mitigating the psychological impact of COVID-19 on healthcare workers: a digital learning package. Int J Environ Res Public Health 2020 Apr 26;17(9):2997. [doi: 10.3390/ijerph17092997] [Medline: 32357424]
11. Chao RC. Managing perceived stress among college students: the roles of social support and dysfunctional coping. J of College Counseling 2012 Apr;15(1):5-21. [doi: 10.1002/j.2161-1882.2012.00002.x]
12. Segrin C, McNelis M, Swiatkowski P. Social skills, social support, and psychological distress: a test of the social skills deficit vulnerability model. Hum Commun Res 2016 Jan;42(1):122-137. [doi: 10.1111/hcre.12070]
13. Bu F, Steptoe A, Fancourt D. Loneliness during lockdown: trajectories and predictors during the COVID-19 pandemic in 35,712 adults in the UK. medRxiv. Preprint posted online on Nov 11, 2020. [doi: 10.1101/2020.05.29.20116657]
14. Christ CC, Gray JM. Factors contributing to adolescents' COVID-19-related loneliness, distress, and worries. Curr Psychol 2022 Feb 13;43(9):1-12. [doi: 10.1007/s12144-022-02752-5] [Medline: 35194361]
15. Lebel C, MacKinnon A, Bagshawe M, Tomfohr-Madsen L, Giesbrecht G. Elevated depression and anxiety symptoms among pregnant individuals during the COVID-19 pandemic. J Affect Disord 2020 Dec 1;277:5-13. [doi: 10.1016/j.jad.2020.07.126] [Medline: 32777604]
16. Shi LSB, Xu RH, Xia Y, Chen DX, Wang D. The impact of COVID-19-related work stress on the mental health of primary healthcare workers: the mediating effects of social support and resilience. Front Psychol 2021;12:800183. [doi: 10.3389/fpsyg.2021.800183] [Medline: 35126252]
17. Every-Palmer S, Jenkins M, Gendall P, et al. Psychological distress, anxiety, family violence, suicidality, and wellbeing in New Zealand during the COVID-19 lockdown: a cross-sectional study. PLOS ONE 2020;15(11):e0241658. [doi: 10.1371/journal.pone.0241658] [Medline: 33147259]
18. Koda M, Harada N, Eguchi A, Nomura S, Ishida Y. Reasons for suicide during the COVID-19 pandemic in Japan. JAMA Netw Open 2022 Jan 4;5(1):e2145870. [doi: 10.1001/jamanetworkopen.2021.45870] [Medline: 35099547]
19. Hoffman J, Liddell BJ, Keegan D, et al. The impact of COVID-19 stressors on refugee mental health and well-being in the context of sustained displacement. Am J Orthopsychiatry 2023;93(2):144-155. [doi: 10.1037/ort0000664] [Medline: 36634011]
20. Snetselaar RS, Liber JM, Geurts SM, Koning IM. Examination of risk exposure models during COVID-19 in relation to youth life satisfaction and internalizing symptoms. Sci Rep 2022 Sep 28;12(1):16252. [doi: 10.1038/s41598-022-20661-2] [Medline: 36171349]
21. Hobfoll SE. Conservation of resources. A new attempt at conceptualizing stress. Am Psychol 1989 Mar;44(3):513-524. [doi: 10.1037//0003-066x.44.3.513] [Medline: 2648906]
22. Hobfoll SE. Social and psychological resources and adaptation. Rev Gen Psychol 2002;6(4):307-324. [doi: 10.1037//1089-2680.6.4.307]

XSL•FO
RenderX

23. Halbesleben JRB, Neveu JP, Paustian-Underdahl SC, Westman M. Getting to the "COR": understanding the role of resources in conservation of resources theory. J Manag 2014;40(5):1334-1364. [doi: 10.1177/0149206314527130]
24. Egozi Farkash H, Lahad M, Hobfoll SE, Leykin D, Aharonson-Daniel L. Conservation of resources, psychological distress, and resilience during the COVID-19 pandemic. Int J Public Health 2022;67:1604567. [doi: 10.3389/ijph.2022.1604567] [Medline: 36119444]
25. Cohen S, Wills TA. Stress, social support, and the buffering hypothesis. Psychol Bull 1985 Sep;98(2):310-357. [doi: 10.1037/0033-2909.98.2.310] [Medline: 3901065]
26. Cohen S, Underwood LG, Gottlieb BH, editors. Social Support Measurement and Intervention: A Guide for Health and Social Scientists: Oxford University Press; 2000. [doi: 10.1093/med:psych/9780195126709.001.0001]
27. Thoits PA. Mechanisms linking social ties and support to physical and mental health. J Health Soc Behav 2011 Jun;52(2):145-161. [doi: 10.1177/0022146510395592] [Medline: 21673143]
28. Hsieh CM, Tsai BK. Effects of social support on the stress-health relationship: gender comparison among military personnel. Int J Environ Res Public Health 2019 Apr 12;16(8):1317. [doi: 10.3390/ijerph16081317] [Medline: 31013723]
29. Green ZA, Faizi F, Jalal R, Zadran Z. Emotional support received moderates academic stress and mental well-being in a sample of Afghan university students amid COVID-19. Int J Soc Psychiatry 2022 Dec;68(8):1748-1755. [doi: 10.1177/00207640211057729] [Medline: 34903066]
30. Huang C, Feng Q, Zhang B, et al. Income and social support related with mental health during COVID-19 outbreak in China. Medicine (Baltimore) 2022 Mar 11;101(10):e29022. [doi: 10.1097/MD.0000000000029022] [Medline: 35451402]
31. Lincoln KD. Social support, negative social interactions, and psychological well-being. Soc Serv Rev 2000 Jun 1;74(2):231-252. [doi: 10.1086/514478] [Medline: 26594064]
32. Fujihara S, Tabuchi T. The impact of COVID-19 on the psychological distress of youths in Japan: A latent growth curve analysis. J Affect Disord 2022 May 15;305:19-27. [doi: 10.1016/j.jad.2022.02.055] [Medline: 35218863]
33. Ueda M, Nordström R, Matsubayashi T. Suicide and mental health during the COVID-19 pandemic in Japan. J Public Health (Oxf) 2022 Aug 25;44(3):541-548. [doi: 10.1093/pubmed/fdab113] [Medline: 33855451]
34. Internet research (internet-based research, web surveys). Cross Marketing Inc. 2025. URL: https://www.cross-m.co.jp/service/marketing-research/net-research [accessed 2025-02-20]
35. Kessler RC, Andrews G, Colpe LJ, et al. Short screening scales to monitor population prevalences and trends in non-specific psychological distress. Psychol Med 2002 Aug;32(6):959-976. [doi: 10.1017/s0033291702006074] [Medline: 12214795]
36. Furukawa TA, Kawakami N, Saitoh M, et al. The performance of the Japanese version of the K6 and K10 in the World Mental Health Survey Japan. Int J Methods Psychiatr Res 2008;17(3):152-158. [doi: 10.1002/mpr.257] [Medline: 18763695]
37. Schuster TL, Kessler RC, Aseltine RH. Supportive interactions, negative interactions, and depressed mood. Am J Community Psychol 1990 Jun;18(3):423-438. [doi: 10.1007/BF00938116] [Medline: 2264558]
38. Ryff CD, Kitayama S, Karasawa M, Markus HR, Kawakami N, Coe C. Survey of midlife development in Japan (MIDJA), April-September 2008 (ICPSR 30822). : Inter-University Consortium for Political and Social Research; 2022. [doi: 10.3886/ICPSR30822.v3]
39. Belarmino M, Roberts MR. Japanese gender role expectations and attitudes: a qualitative analysis of gender inequity. J Int Women's Stud 2019;20(7):272-288 [FREE Full text]
40. Koo J, Cox WM. An economic interpretation of suicide cycles in Japan. Contemp Econ Policy 2008 Jan;26(1):162-174. [doi: 10.1111/j.1465-7287.2007.00042.x]
41. Suicides in 2019. National Police Agency. 2020. URL: https://www.npa.go.jp/safetylife/seianki/jisatsu/R02/R01_jisatuno_joukyou.pdf [accessed 2024-11-30]

Abbreviations

KUAS: Kyoto University of Advanced Science
NIBIOHN: National Institutes of Biomedical Innovation, Health, and Nutrition
OR: odds ratio

Edited by A Mavragani; submitted 08.08.24; peer-reviewed by N Sugaya, S Fujihara; revised version received 26.03.25; accepted 10.04.25; published 10.06.25.

Please cite as:
Hori K, Yamada Y, Namba H, Kimura M, Fujita H, Date H
Economic Deterioration and Social Factors Affecting Mental Health During the COVID-19 Pandemic in Japan: Web-Based Cross-Sectional Survey
JMIR Form Res 2025;9:e65204
URL: https://formative.jmir.org/2025/1/e65204
doi:10.2196/65204

XSL•FO
RenderX

© Kentaro Hori, Yosuke Yamada, Hideyuki Namba, Misaka Kimura, Hiroyuki Fujita, Heiwa Date. Originally published in JMIR Formative Research (https://formative.jmir.org), 10.6.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Explainable Machine Learning Framework for Dynamic Monitoring of Disease Prognostic Risk: Retrospective Cohort Study

Tetsuo Ishikawa[1,2,3,4,5,6*], PhD; Masahiro Shinoda[4*], MD, PhD; Megumi Oya[1,3,4,6*], MD, PhD; Koichi Ashizaki[1,4], BA; Shinichiro Ota[4], MD, PhD; Kenichi Kamachi[4], MD; Kazuhiro Sakurada[1,2], PhD; Eiryo Kawakami[1,3,4,6], MD, PhD; Masaharu Shinkai[4], MD, PhD

[1]Predictive Medicine Special Project, RIKEN Center for Integrative Medical Sciences, RIKEN, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Japan
[2]Department of Extended Intelligence for Medicine, The Ishii-Ishibashi Laboratory, Keio University School of Medicine, Tokyo, Japan
[3]Department of Artificial Intelligence Medicine, Graduate School of Medicine, Chiba University, Chiba, Japan
[4]Department of Respiratory Medicine, Tokyo Shinagawa Hospital, Tokyo, Japan
[5]Collective Intelligence Research Laboratory, Graduate School of Arts and Sciences, The University of Tokyo, Tokyo, Japan
[6]Division of Applied Mathematical Science, RIKEN Center for Interdisciplinary Theoretical and Mathematical Sciences, RIKEN, Yokohama, Japan
[*]these authors contributed equally

Corresponding Author:
Eiryo Kawakami, MD, PhD
Predictive Medicine Special Project, RIKEN Center for Integrative Medical Sciences, RIKEN, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Japan

Abstract

Background: Patients' clinical status often evolves rapidly after an initial diagnosis, with each patient exhibiting a distinct disease trajectory. As a result, static risk scores fall short in supporting timely interventions—an issue highlighted by COVID-19, where deaths have stemmed from heterogeneous pathways such as pneumonia, multiorgan failure, or exacerbation of preexisting conditions.

Objective: This study aims to propose a dynamic prognostic risk assessment framework based on longitudinal data collected during hospitalization, using COVID-19 as an example. Our aim was to develop and validate an interpretable framework that (1) screens prognosis at admission and (2) dynamically updates mortality risk throughout hospitalization, thereby providing clinicians with early, explainable warnings while minimizing additional cognitive load.

Methods: In this retrospective study, we extracted electronic medical records of 382 COVID-19 cases treated at Tokyo Shinagawa Hospital between January 27 and September 30, 2020. At admission, gradient boosting decision trees (Light Gradient Boosting Machine) were used to predict the maximum clinical deterioration, including death, based on data available at initial diagnosis. Model performance was evaluated using the area under the receiver operating characteristic curve (AUC). For in-hospital monitoring, random survival forests (RSF) were trained on a longitudinal dataset that combined static demographic characteristics with serially measured vital signs and laboratory results. The model dynamically assessed daily mortality risk by calculating a 7-day cumulative hazard function, with risk scores recalculated each day during hospitalization. RSF accuracy was evaluated in an independent one-third test set using the concordance index (C-index), an integrated Brier score (1-50 days), and mean time-dependent AUC. SurvSHAP(t), an extension of Shapley Additive Explanations, was applied to provide time-dependent explanations of each variable's contribution to the prediction.

Results: The prediction at initial diagnosis showed good agreement with the actual severity outcomes (AUC of 0.717 for predicting hospitalization/severity ≥2; 0.878 for severity ≥3; 0.951 for severity ≥4; 0.952 for severity ≥5; and 0.970 for death/severity=6), although some cases exhibited discrepancies between the predicted and actual prognoses. The dynamic mortality risk assessment during hospitalization using the RSF achieved a test-set C-index of 0.941, an integrated Brier score of 0.315, and a mean time-dependent AUC of 0.936. This dynamic assessment was able to distinguish between dead and surviving patients as early as 1 - 2 weeks before the outcome. Early in hospitalization, C-reactive protein was an important risk factor for mortality; during the middle period, peripheral oxygen saturation (SpO_2) gained importance; and immediately before death, platelets and β-D-glucan were the primary risk factors.

Conclusions: Integrating static admission triage with daily, explainable RSF predictions enables early identification of patients with COVID-19 at high risk of deterioration. By surfacing phase-specific, actionable predictors, the framework supports timely interventions and more efficient resource allocation. Prospective, multicenter studies are warranted to validate its generalizability and clinical impact.

XSL•FO
RenderX

(JMIR Form Res 2025;9:e65585) doi:10.2196/65585

KEYWORDS

COVID-19; prognostic modeling; electronic health records; time-varying biomarkers; clinical decision support; interpretable AI; random survival forests

Introduction

During hospitalization, various adverse events and complications arise from the disease and its treatment, sometimes leading to death [1]. Therefore, health care professionals take great care not to overlook signs of deterioration based on blood tests and vital-sign measurements [2,3]. The recent surge in hospital admissions driven by COVID-19 and population aging has intensified this burden [4,5]. Predicting patient outcomes and issuing timely alerts can reduce the burden and enable the advanced preparation of scarce medical resources. To frame the scope of this study, we ask a single overarching question: How can clinicians anticipate deterioration early enough to intervene, while minimizing additional cognitive load?

The task is complicated by the heterogeneity of risk factors and signs of deterioration, which makes uniform risk assessment challenging, even for patients with the same disease [6]. Taking COVID-19 as an example, causes of death range from pneumonia to multiorgan failure and exacerbation of preexisting diseases [7]. Detecting such diverse paths to deterioration requires a personalized monitoring strategy that tracks multiple prognostic biomarkers over time. This clinical heterogeneity defines the modeling gap that data-driven methods must fill.

Recent reports indicate that machine learning models using electronic health records can accurately predict individual deterioration and mortality, thereby aiding health care professionals in clinical decision-making [8,9]. However, most existing prognostic models are trained on static snapshots, struggle to generalize across heterogeneous populations, or fail to provide interpretable insights. While deep learning models applied to longitudinal intensive care unit data can assess risks dynamically [10,11], their “black-box” nature restricts bedside adoption [12]. Accordingly, the next generation of prognostic tools must be both dynamic and interpretable to earn clinical trust.

Real-world deployment adds further hurdles: missing data, irregular sampling, inconsistent documentation, and limited computation resources in busy wards. Methods must therefore be accurate, explainable, and robust to data imperfections, while remaining fast enough for routine use. Taken together, limited interpretability, data-quality issues, and computational constraints form a 3-fold barrier that this work seeks to overcome.

To address these limitations, we leverage random survival forests (RSFs) to capture time-varying trajectories and pair them with SurvSHAP(t), an extension of Shapley Additive Explanations (SHAP) that provides context-specific, time-dependent rationale [13]. In contrast to models that assume linear relationships (eg, Cox proportional hazards), RSF flexibly models nonlinear, time-dependent interactions; SurvSHAP(t) embeds explanation directly alongside prediction, closing the trust gap.

Building on these insights, we first confirm the predictability of prognosis at initial diagnosis with a machine learning model that anticipates the highest COVID-19 severity from baseline data. We then propose a dynamic prognostic-risk framework that continuously updates survival estimates using longitudinal information (Figure 1) and visualizes the drivers of individual risk. This 2-stage design—static triage at admission followed by dynamic risk updates—constitutes the narrative arc of the paper.

We therefore present an interpretable, dynamic prognostic framework designed to enhance clinicians’ understanding of patient status, enable precise treatment allocation, and optimize the use of medical resources.

XSL•FO
RenderX

Figure 1. Conceptual diagram illustrating the diversity of pathological changes. In this study, we first assess the ability to predict a patient's maximum COVID-19 severity using data from the initial diagnosis. We then introduce a dynamic prognostic risk assessment framework that continuously monitors hospitalized patients based on their longitudinal data. BP: blood pressure; CRP: C-reactive protein; ECMO: extracorporeal membrane oxygenation; GBDT: gradient boosting decision trees; Plt: platelet.

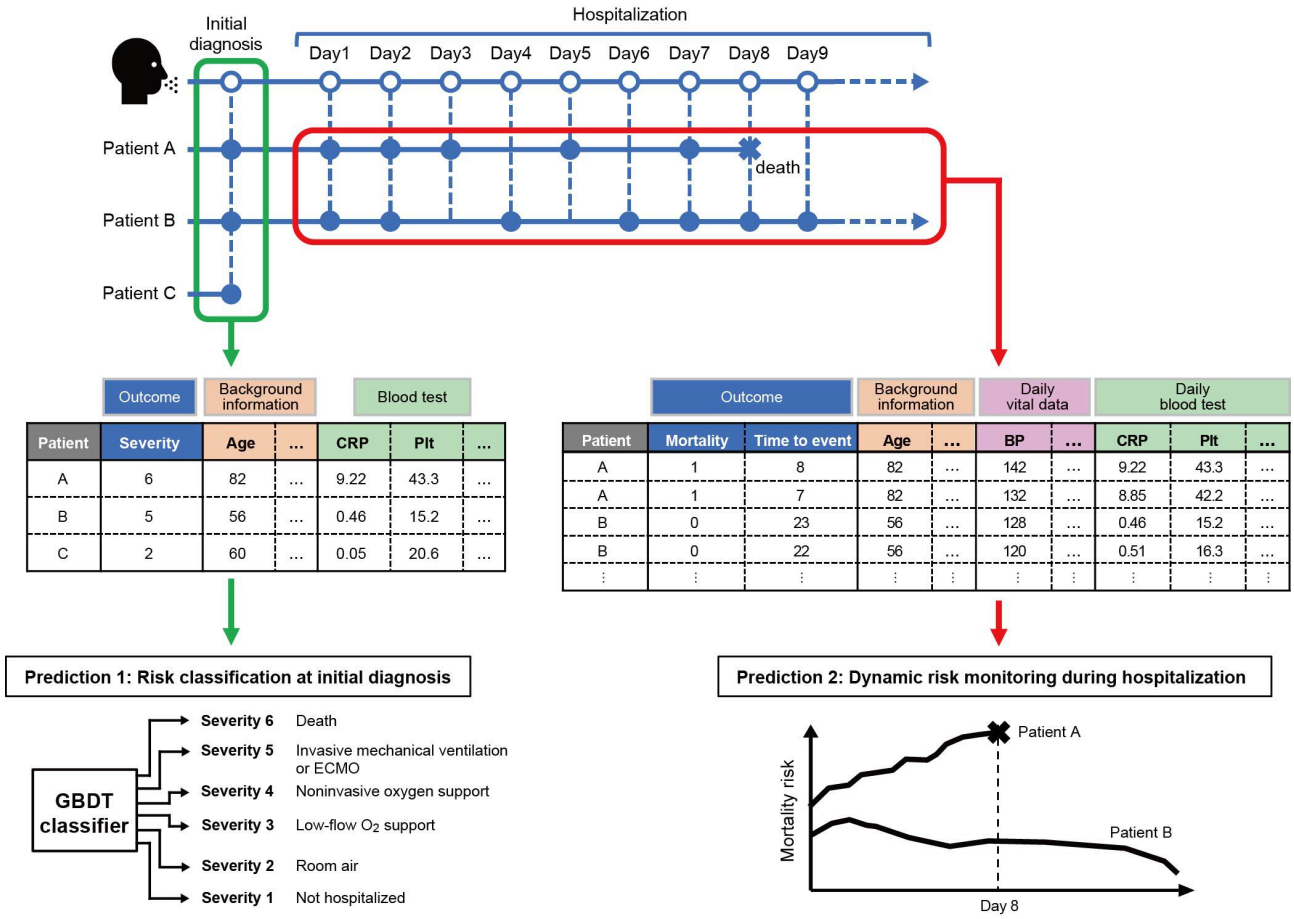

Methods

Study Design, Setting, and Population

This retrospective cohort study screened 382 patients with COVID-19 treated between January 2020 and September 2020 at the Department of Respiratory Medicine, Tokyo Shinagawa Hospital. Patients aged under 18 years and those with unknown severity changes during hospitalization or clinical outcomes were excluded from the analysis. This included cases in which the prognosis was unclear because of hospital transfer, observation censoring, or insufficient documentation of the time point at which severity changed. Inpatients with a discrepancy of more than 4 days between the date of first diagnosis and the date of admission were excluded from the prognostic screening but were included in the dynamic risk assessment. The included cases were randomly divided into training and test datasets at a 2:1 ratio using a stratified sampling method to preserve the distribution of severity outcomes.

For risk classification at initial diagnosis, patients who experienced severity changes were included in the dataset. Overall, 201 outpatients and 161 inpatients were included. We extracted data from the electronic medical records, including 84 variables, such as symptoms, background information, and blood and urine biomarkers, for prognostic screening at initial diagnosis (Multimedia Appendix 1).

For the dynamic prognostic risk assessment framework, the dataset included 182 inpatients with longitudinal data collected during hospitalization. From the electronic medical records, 72 variables, including blood and urine biomarkers, vital signs, and background information during hospitalization, were extracted for dynamic risk assessment (Multimedia Appendix 2). The median length of stay for inpatients was 11 days.

Ethical Considerations

This study was approved by the local institutional review board of RIKEN and Tokyo Shinagawa Hospital (approval number 20-A-06). It was conducted in accordance with the 1964 Declaration of Helsinki and its later amendments, or comparable ethical standards. The requirement for informed consent was waived because of the retrospective nature of the study. All data used in this study were deidentified before analysis. Specifically, any directly identifiable personal information (eg, names, addresses, or medical record numbers) was removed. The deidentification process adhered to relevant institutional and national data protection guidelines. As the data were not anonymized in the strictest sense, but were instead deidentified with safeguards in place, individual participants cannot be readily identified by the investigators.

Severity Classification

We adopted the oxygen-support status scores proposed by Grein et al [14] to classify patient severity. The score uses an ordinal

XSL•FO
RenderX

scale from 1 to 6 based on the level of oxygenation: (1) discharged or not hospitalized; (2) room air; (3) low-flow oxygen support; (4) noninvasive intervention, including nasal high-flow oxygen therapy or noninvasive positive pressure ventilation; (5) invasive mechanical ventilation or extracorporeal membrane oxygenation (ECMO); and (6) death. In addition to the original score, where 1 indicated discharge status, outpatients were assigned a score of 1.

Machine Learning for Prognostic Classification at Initial Diagnosis

Gradient boosting decision trees (GBDT) [15,16] were used to predict the highest severity from data at initial diagnosis. GBDT is related to random forest and uses boosting to enhance predictive performance. Among various GBDT implementations, the Python package (Python Foundation) Light Gradient Boosting Machine (LightGBM; version 4.5.0) was used because of its superior computing speed. We included all variables with less than 15% missing values, given that COVID-19 severity and mortality are influenced by a wide range of factors. Although 265 biomarkers were recorded at the first visit, most were not measured in nonhospitalized patients. To ensure that the selected biomarkers would be applicable across a broad clinical spectrum, we included only laboratory test variables that were routinely measured in both hospitalized and nonhospitalized (outpatient) patients at the time of initial diagnosis. This approach was intended to support early-stage risk prediction and avoid biasing the model toward inpatient-specific features. Biomarkers available in less than 50% of the patient population were excluded to reduce the impact of missing data on model performance, resulting in 53 biomarkers collected from 4 days before to 1 day after the initial diagnosis. Missing values were imputed using the missForest package (version 1.5) in R (R Foundation) [17].

As a preprocessing step, variables with absolute Spearman correlation coefficients of 0.85 or higher in the training data were excluded, retaining only those with the strongest correlation with severity. While this univariate filtering approach does not account for potential multivariate interactions among features, it was adopted to reduce feature redundancy and multicollinearity, which could otherwise impair the stability and efficiency of model training.

LightGBM hyperparameters were optimized using Bayesian optimization with repeated cross-validation via Optuna version 4.2.1 [18]. We applied the best model to the test dataset to obtain predicted probabilities and plotted the corresponding receiver operating characteristic curve. The area under the receiver operating characteristic curve was calculated to evaluate the predictive performance of the initial diagnosis model for each severity classification. To evaluate the importance of variables in the prediction, we used SHAP values [13] for the test data. Importantly, when models were developed for different severity subgroups, hyperparameter optimization and SHAP analysis were performed separately for each group to account for subgroup-specific characteristics.

To validate the advantage of the selected LightGBM model, we initially implemented and evaluated 5 machine learning algorithms for predicting hospital admission (ie, threshold ≥2): LightGBM, random forest, support vector machine, Elastic Net, and decision tree. Among these, LightGBM and random forest demonstrated the highest predictive performance for binary classification (Figure S1 in Multimedia Appendix 3 and Table S3 in Multimedia Appendix 4). However, random forest required substantially longer training and inference times. Considering the need for computational efficiency and scalability in clinical applications, we selected LightGBM for subsequent analyses involving higher severity cut-offs.

Dynamic Risk Assessment of Mortality During Hospitalization

We used survival models, including Cox proportional hazards models and RSFs [19], to evaluate patient mortality risk during hospitalization. The Cox proportional hazards model analyzes time-to-event data and associated predictors under the proportional hazards assumption. To accommodate a large number of predictors, regularized versions of the Cox model were used, including Cox-LASSO (least absolute shrinkage and selection operator) (L1), Cox-Ridge (L2), and Cox ElasticNet (a combination of L1 and L2 penalties).

To emulate real-time risk assessment, we implemented a dynamic updating mechanism in which the model recalculates cumulative hazard scores on each day of hospitalization. For each patient, the RSF model is applied iteratively at time points t, t+1, t+2, and so on, using only data observed up to the corresponding time. This time-restricted evaluation prevents data leakage from future observations and allows us to assess how well the model performs when used prospectively in a clinical setting. The RSF estimates the cumulative hazard function (CHF) by aggregating the outputs of multiple survival trees constructed using the random forest algorithm. Like random forest, RSF is robust to outliers and enables accurate risk assessment for event occurrence. To dynamically predict mortality risk, 72 variables were used in the analysis, including background factors (eg, age and BMI), as well as biomarkers and vital signs measured over time during hospitalization. We used variables that were measured in most patients (>180 patients for biomarkers and >150 patients for background and vital information). Missing values were imputed using the last observation carried forward method; if no prior values were available, continuous variables were filled with the median, and categorical variables with the mode. The Python (Python Foundation) package scikit-survival, which includes implementations of Cox and RSF models, was used for the analysis. To evaluate the dynamic risk of mortality, we used the 7-day CHF (ie, the estimated probability of death within 7 days) calculated using the RSF as an index of mortality risk. The 7-day CHF was calculated for each patient on each day of hospitalization.

The performance of the survival models was evaluated using 3 metrics: concordance index (C-index), integrated Brier score, and mean cumulative/dynamic AUC (mean AUC).

In clinical practice, it is important to identify the factors that increase the risk of mortality. The variable importance of the RSF models was calculated for the test data using SurvSHAP(t) [20], which provides a time-dependent explanation of the survival function predicted using SHAP. SurvSHAP(t) allowed

XSL•FO
RenderX

us to determine how each variable affects the survival function at each time point. We obtained the mean aggregated SurvSHAP(t) across all days of hospitalization, which showed how each variable contributed to the prediction for every patient at every time point.

Variable Selection for Dynamic Risk Assessment

To balance clinical applicability with the computational cost of SurvSHAP(t), we performed variable selection during the RSF training process. First, we trained an RSF model using all 72 variables on the training data and calculated permutation importance scores using the implementation in the scikit-learn Python package. The RSF model was then retrained using the top 10 variables with the highest permutation importance.

Additional technical details of the machine learning implementation not covered in the main text are presented in Multimedia Appendix 5.

Results

Patient Demographics

The median age of the 382 patients with COVID-19 was 39 years; 233 were male and 149 were female (Table 1). Of the 51 inpatients who required oxygen, 30 required low-flow oxygen, 3 needed high-flow oxygen, 8 were treated with invasive ventilation/ECMO, and 10 died.

Table . Baseline characteristics of patients with COVID-19 by severity outcome.

	Severity outcome of COVID-19					
	1: Not admitted	2: Room air	3: Oxygen administration	4: Noninvasive ventilation	5: ECMO[a]/invasive ventilation	6: Death
Number of cases, n	201	130	30	3	8	10
Length of stay (days), median (range)	N/A[b]	10 (2-52)	20 (11-36)	47 (31-48)	39 (25-73)	23 (12-37)
Age (years), median (range)	34 (18-93)	39 (19-98)	59 (30-93)	70 (68-85)	58 (49-75)	79 (50-94)
Sex, n (%)						
Female	75 (37)	60 (46)	9 (30)	2 (67)	2 (25)	1 (10)
Male	126 (63)	70 (54)	21 (70)	1 (33)	6 (75)	9 (90)

[a]ECMO: extracorporeal membrane oxygenation.

[b]N/A: not applicable.

Prognostic Predictability of COVID-19 at Initial Diagnosis

To evaluate the predictability of COVID-19 prognosis at initial diagnosis, we constructed prognostic models based on the following information available at that time: symptoms, background information, and biomarkers. Using the severity scale defined by Grein et al [14], we assigned a score to each patient at initial presentation and, if hospitalized, at subsequent time points. The highest score observed across these time points was used as the outcome to be predicted. The AUCs represent the performance of the model in predicting whether patients would reach each severity threshold during hospitalization: 0.717 for severity ≥2, 0.878 for ≥3, 0.951 for ≥4, 0.952 for ≥5, and 0.970 for 6 (ie, death), as shown in Figure 2A. When we examined the concordance between the predicted probability and the actual severity outcome, we found that most patients with an actual severity of ≥3 had a high predicted probability of severity of ≥2. By contrast, patients with an actual severity of 1 or 2 showed a wide distribution of predicted probabilities—from low to high—indicating no clear distinction between the 2 (Figure 2B, top panel). For the other predictions, the predicted probabilities and actual severity showed good concordance; however, some cases remained where the severity risk was not accurately determined (eg, patients who ultimately had mild disease but were predicted to have a high risk of severe illness; Figure 2B).

XSL•FO
RenderX

Figure 2. Screening for severity prognosis at the initial diagnosis. (A) Receiver operating characteristic curves and the AUC values for prediction models of severity outcomes based on the oxygen-support status score. The 95% CIs for the AUC are shown in parentheses. (B) Box and jitter plots showing the prediction probabilities that test cases exceed the severity threshold (indicated by the dashed line). The center line of the box indicates the median; box limits represent the upper and lower quartiles; whiskers extend to 1.5× the IQR. Each dot represents an individual observation. (C) Variable importance in each predictive model evaluated using SHAP values. Bar plots show the mean absolute SHAP value for each variable across individual patients in the test dataset. (D) Heatmap of symptoms observed at initial diagnosis. Columns represent individual patients, ordered by severity level and sorted to minimize missing values and symptom absence. (E) Box and jitter plots showing the distribution of key continuous variables across severity groups. Plot elements are defined as in panel B. Alb: albumin; ALP: alkaline phosphatase; AMY: amylase; APTT: activated partial thromboplastin time; AUC: area under the receiver operating characteristic curve; Baso-%: basophil %; CK: creatine kinase; Cl: chloride; Cre: creatinine; CRP: C-reactive protein; eGFR-Cre: estimated glomerular filtration rate-creatinine; Eosino-%: eosinophil %; GLU: blood glucose; LDH: lactate dehydrogenase; Lympho-Count: lymphocyte count; MCH: mean corpuscular hemoglobin; MCHC: mean corpuscular hemoglobin concentration; MCV: mean corpuscular volume; Mono-Count: monocyte count; MPV: mean platelet volume; N/A: not available; PDW: platelet distribution width; Plt: platelet; PT: prothrombin time; Neutro-%: neutrophil %; RDW: red cell distribution width; SHAP: Shapley Additive Explanations; T-Bil: total bilirubin; TP: total protein; UA: uric acid; UN: urea nitrogen; WBC: white blood cell; γ-GTP: gamma-glutamyl transpeptidase.

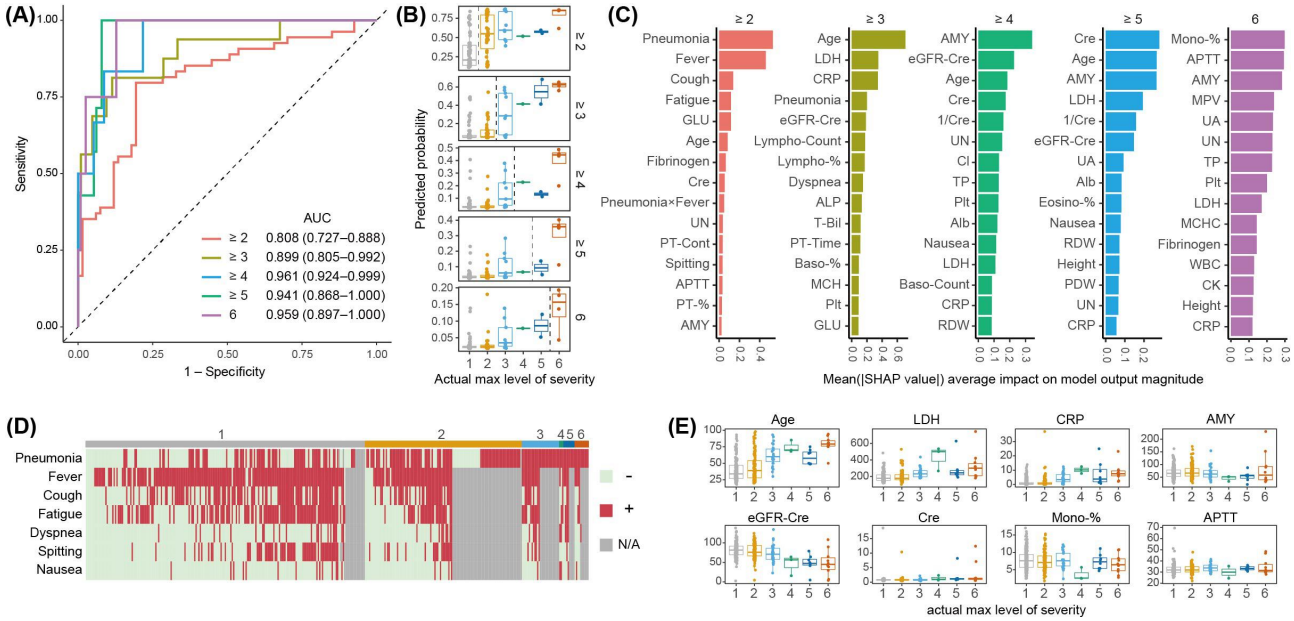

Subsequently, we evaluated the variable importance of each prediction. The presence of pneumonia was identified as the most important predictor for whether severity would be ≥2 (Figure 2C, leftmost panel). Most hospitalized patients with a severity of ≥2 showed pneumonia at initial diagnosis, and all patients with a severity of ≥3 had pneumonia (Figure 2D). Conversely, as the target severity level of prediction increased, the importance of symptoms decreased. Instead, biomarkers such as the lymphocyte count, prothrombin time (PT), C-reactive protein (CRP), creatinine, and amylase levels emerged as the top predictors (Figure 2C). BMI was also an important factor in predicting whether the severity was ≥4 or ≥5. Age remained an important factor across all severity levels. Among the important predictors, we examined the distribution of age and biomarkers according to actual severity prognosis. Age and BMI were higher in patients with a severity of ≥3 than in others, while biomarkers such as PT, CRP, creatinine, red blood cell volume distribution width, and blood glucose values were higher in patients with a severity of ≥4 (Figure 2E). Lymphocyte counts, estimated glomerular filtration rate (eGFR)-creatinine, and albumin values were lower in more severe cases than in less severe cases, and platelet counts were particularly low in mortality cases (Figure 2E). Amylase levels showed a peculiar distribution: they were low in patients with severity levels 4, 5, and 6, but extremely high in some patients with a severity level of 6 (Figure 2E).

Dynamic Mortality Risk Assessment Based on Longitudinal Data During Hospitalization

Outcome screening based on information from the initial diagnosis was accurate but incomplete. This suggests that a patient's prognosis is not fully determined at the time of diagnosis and may change depending on the disease course and treatment after admission. When we investigated changes in severity status after admission, we found that oxygen administration and noninvasive ventilation were typically initiated within 5 days after admission, whereas invasive ventilation was often introduced more than 5 days after admission; death generally occurred more than 20 days after admission (Figure 3A). Additionally, many deaths occurred without invasive ventilation (Figure 3A, top panel). This is likely because most mortality cases involved older adults who were not eligible for invasive ventilation or ECMO, even when their condition deteriorated. These observations suggest that, during the weeks between admission and death or discharge, patients undergo changes in their condition that cannot be fully assessed based on oxygen-support status.

We then used survival models, including RSF [19], to evaluate mortality risk during hospitalization. Initially, we attempted to develop a Cox proportional hazards model using all 72 variables, but the model could not be created due to multicollinearity. Therefore, we developed Cox proportional hazards models with L1 (LASSO), L2 (Ridge), and combined L1+L2 (Elastic Net)

XSL•FO
RenderX

penalties. The Cox model achieved the highest performance, as evaluated by the C-index, when the Elastic Net penalty was applied. Its performance was comparable to that of the RSF model using all 72 variables (Table 2). When we retrained an RSF model using the top 10 important variables selected during the RSF learning process, as described in the "Methods" section, the performance improved substantially. We also constructed a Cox model based on the same top 10 variables selected by RSF. Although this model exhibited markedly better performance compared with the regularized Cox model, it did not match the predictive accuracy of the RSF model using the same variables. These findings highlight the critical importance of variable selection in enhancing model performance, as well as the value of accounting for interactions and nonlinear effects—factors that the Cox model does not incorporate.

Figure 3. Dynamic mortality risk assessment during hospitalization. (A) Changes in severity status during hospitalization for each severity outcome group. The plot shows how the number of patients receiving procedures such as ventilation or oxygen administration changes over time after admission. (B) Daily mortality risk trajectories for each patient for each day of hospitalization in the test dataset. Mortality risk is shown as the 7-day CHF (ie, the risk of death within 7 days), calculated using the RSF. Filled circles in mortality cases indicate the date of last observation. (C) Daily mortality risk trajectories for each patient for each day of hospitalization in the training dataset. (D) Important variables contributing to the RSF-based mortality prediction. SurvSHAP(t) provides a time-dependent explanation of how each variable influences each patient's CHF. The left panel shows the mean aggregated SurvSHAP(t) for each case, that is, the mean value of aggregated SurvSHAP(t) across all days of hospitalization. Each circle represents the contribution of a variable to the prediction for each patient in the test dataset. The right panel shows the mean absolute value of SurvSHAP(t) across all patients in the test dataset. (E) Mean aggregated SurvSHAP(t) for each case (left), and mean aggregated |SurvSHAP(t)| across all cases (right) in the training dataset. AMY: amylase; Ca: calcium; CHF: cumulative hazard function; CRP: C-reactive protein; Plt: platelets; PT: prothrombin time; RSF: random survival forests; SpO_2: peripheral oxygen saturation

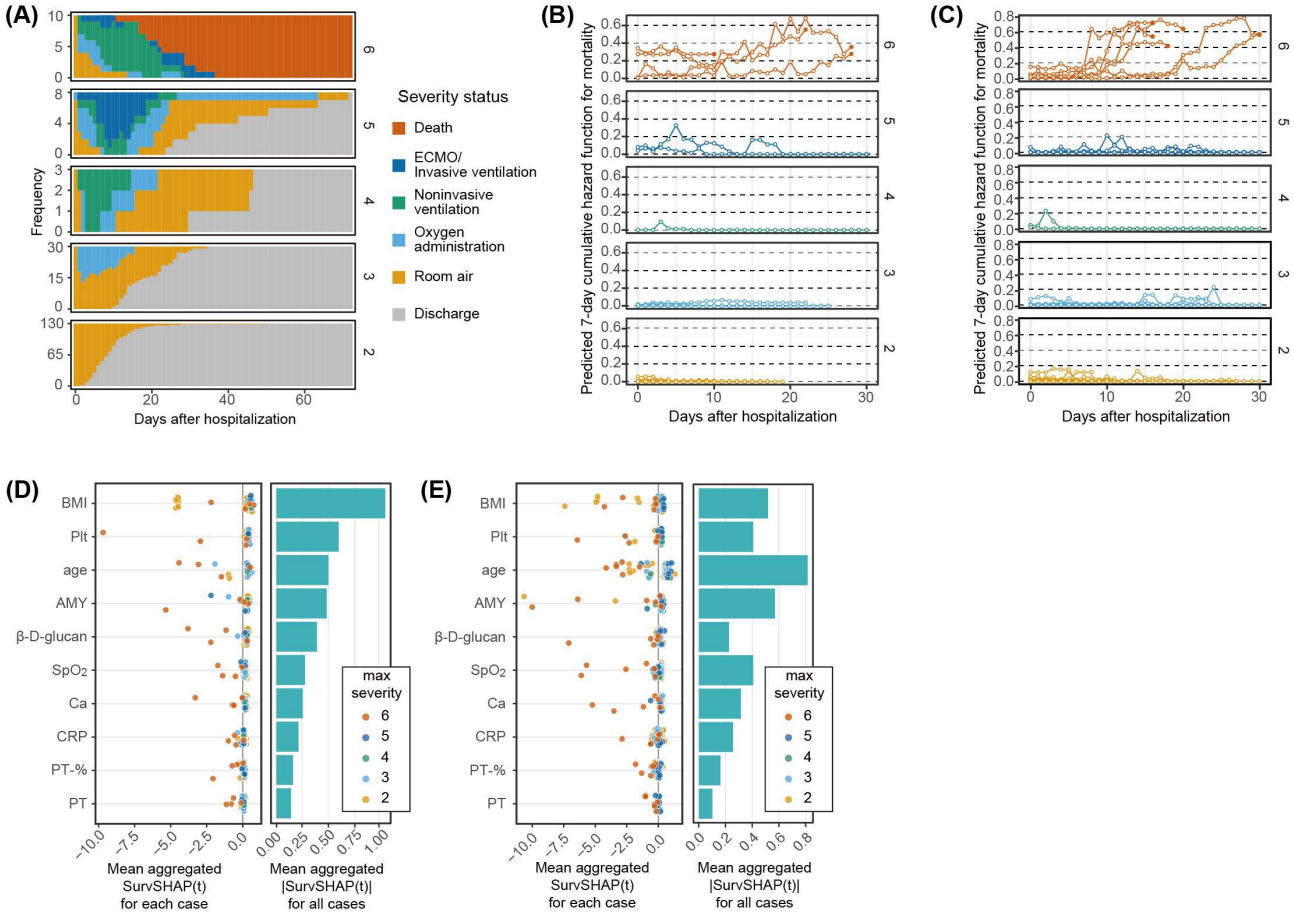

For all 4 mortality cases in the test dataset, an increase in mortality risk was observed 1 - 2 weeks before the outcome, and the CHF (ie, the estimated probability of death within 7 days) reached approximately 0.3 or higher at the time of death (Figure 3B). Conversely, in patients recovering from invasive or noninvasive ventilation, the CHF also increased around 1 week after admission, similar to the mortality cases, but then decreased and rarely exceeded 0.2. In mild cases that progressed with oxygen administration or room air, there was little increase in CHF. Similar changes in CHF were observed in the training dataset, which included 6 deaths, suggesting consistency of the predictive model between the training and test datasets (Figure 3C).

The average contribution of the variables to the RSF prediction for each patient was evaluated using the mean aggregated SurvSHAP(t). In addition to background factors such as age and BMI, blood test variables measured multiple times during hospitalization, such as platelets, amylase, and β-D-glucan, had the highest importance in mortality prediction (Figure 3D). Age, amylase level, and platelet count were identified as important predictors of mortality at initial diagnosis. Most of the important variables showed high contributions only in mortality cases;

XSL•FO
RenderX

however, some, such as age and BMI, showed a high mean aggregated SurvSHAP(t) even in milder cases, suggesting that they are nonspecific factors. Conversely, important blood test variables such as β-D-glucan, platelets, and calcium (Ca) contributed less in some mortality cases, suggesting heterogeneity of deterioration. In the training data, these predictors also showed high contributions, especially in mortality cases, although there was some shuffling in their rankings (Figure 3E).

Table . Performance metrics of mortality risk prediction models.

Model	Concordance index	Integrated Brier score	Cumulative area under the receiver operating characteristic curve
Cox-LASSO[a]	0.715	1.14	0.560
Cox-Ridge	0.595	1.53	0.481
Cox ElasticNet	0.845	1.31	0.731
Random survival forests with all 72 variables	0.843	0.664	0.753
Random survival forests with the top 10 variables	*0.941* [b]	0.315	*0.936*
Cox with the top 10 variables	0.919	*0.272*	0.882

[a]LASSO: least absolute shrinkage and selection operator.

[b]Italicized values indicate the model with the best performance for each indicator.

We chose the 7-day time frame because it is clinically meaningful in the context of inpatient care, where monitoring and interventions typically follow a weekly cycle. It provides a practical horizon for anticipating deterioration and making timely care decisions. In addition, the 7-day window offered a suitable balance between the number of outcome events and prediction stability during model development. We also evaluated the data over different periods (eg, 1- and 14-day windows). Although the absolute values of the CHFs differed among the 1-, 7-, and 14-day prediction windows, the temporal patterns within each severity group were approximately consistent (see Figure S2 in Multimedia Appendix 3).

Rationale for Estimated Mortality Risk in Severe Cases

For the mortality risk assessment of patients with COVID-19, a machine learning model based on blood markers was proposed in Nature Machine Intelligence (NMI) 2020 [21]. This model was built using samples obtained immediately before death or discharge, and lactate dehydrogenase, lymphocytes, and CRP were identified as key features. Mortality risk assessment using a decision tree using these 3 features was proposed in the study to facilitate clinical application (Figure 4A). After approximately 10 days of hospitalization, the RSF and NMI models showed equivalent performance as measured by accuracy and F_1-score. However, immediately following hospitalization, the RSF model outperformed the NMI model, indicating its superiority in early prognostic prediction (Figure 4B and C).

To examine the factors contributing to mortality risk over time, we calculated SurvSHAP(t) daily for the severe cases in the test dataset. Patients and time points showed different combinations of factors associated with changes in CHF (Figure 4D). For example, in young patients #2 and #6, peripheral oxygen saturation (SpO_2) was associated with an increase in CHF immediately after admission. However, the contribution of SpO_2 decreased following the initiation of ECMO or invasive ventilation. By contrast, in patients #3 and #4, who were both older and not eligible for ECMO or invasive ventilation, the contribution of SpO_2 was higher than 10 days after hospitalization. β-D-glucan showed a greater contribution to mortality risk 1 - 2 weeks before death than at any other time in patients #3, #4, and #6. Platelets had a high contribution in patients #4 and #6, and BMI also contributed to the mortality risk in these patients. Of note, in patient #4, BMI was measured for the first time on day 7, leading to a rapid increase in its contribution on that day. Calcium and blood amylase levels were elevated immediately before death, suggesting electrolyte abnormalities and multiorgan failure. CRP, a major predictor of mortality at initial diagnosis, contributed to risk assessment immediately after hospitalization.

XSL•FO
RenderX

Figure 4. Validation of the RSF model for mortality risk assessment. (A) Mortality risk assessment for COVID-19 based on 3 key predictors identified in a previous study: LDH, lymphocytes, and CRP. (B, C) Comparison of the performance of the RSF model developed in this study with the previously proposed NMI model. Accuracy (B) and F_1-score (C) for distinguishing survival and death cases are shown for each day after hospitalization, based on 61 COVID-19 cases in the test dataset. (D) Changes in the RSF-predicted mortality risk during hospitalization, along with daily SurvSHAP(t) of important predictors for severely ill patients in the test dataset. Six representative severe cases are shown, including 2 ECMO survivors and 4 deaths. The daily SurvSHAP(t) values for each variable correspond to changes in RSF mortality risk. The oxygen-support status is displayed between the CHF plot and the aggregated SurvSHAP(t) plot for each variable, as in Figure 3A: room air is shown in yellow, oxygen administration in light blue, NIV in green, and IMV/ECMO in blue. The light blue and black bars at the top of the graph indicate predictions by the previous NMI model, with light blue for predicted survival and black for predicted death. AMY: amylase; Ca: calcium; CHF: cumulative hazard function; CRP: C-reactive protein; ECMO: extracorporeal membrane oxygenation; IMV: invasive mechanical ventilation; LDH: lactate dehydrogenase; LFNC: low-flow nasal cannula; NIV: noninvasive ventilation; NMI: Nature Machine Intelligence; Plt: platelet; PT: prothrombin time; RA: room air; RSF: random survival forests; SpO_2: peripheral oxygen saturation.

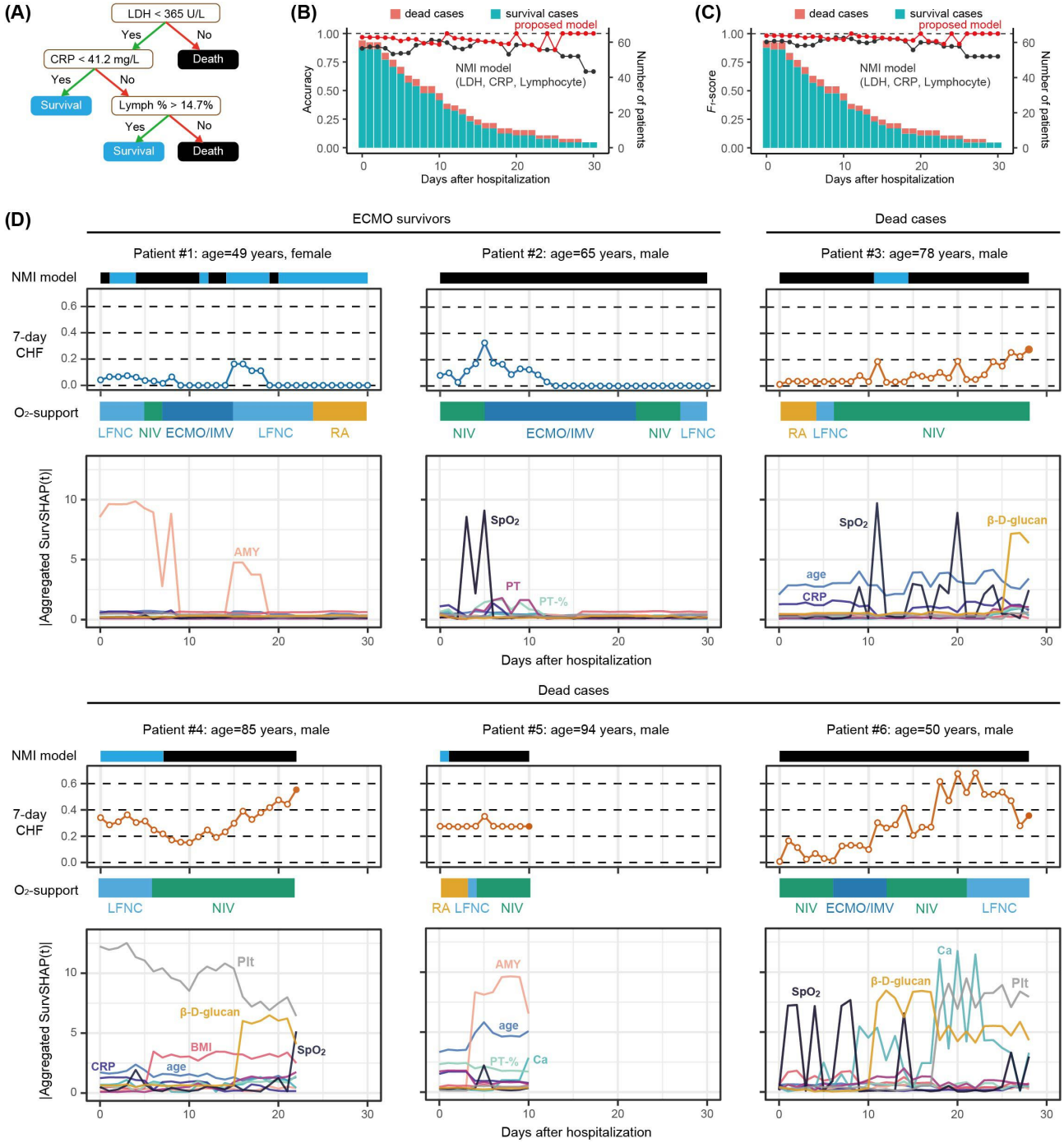

XSL•FO
RenderX

Discussion

Principal Findings

In this study, we developed a framework to assess the prognostic risks of individual patients with COVID-19. Although we were able to predict patient prognosis with reasonable accuracy based on symptoms and blood test results at initial diagnosis, the predicted risk and actual prognosis deviated for some patients, suggesting that prognostic risk may change during hospitalization. Our proposed RSF model demonstrated high performance in predicting in-hospital mortality as a binary outcome, with an F_1-score exceeding 0.95 based on daily predictions made immediately after hospitalization (threshold: CHF>0.2). Visualization of predictors using SurvSHAP(t) made the model explainable for each patient, with consistent predictors observed in both the training and test datasets. Such a dynamic prognostic framework can provide health care providers with real-time guidance to support clinical decision-making during hospitalization.

In the prognostic model at initial diagnosis, we did not predefine a single severity cut-off to dichotomize the outcome. Instead, we explored multiple binary classification tasks based on different thresholds (eg, severity≥1, ≥2,..., ≥5) derived from the original 7-point ordinal outcome scale (0 - 6). This approach was chosen to comprehensively assess how predictive performance varies across different levels of clinical severity and to allow for flexible interpretation depending on the clinical context of interest. The partial predictability of disease progression at the time of initial diagnosis can be explained by key variables. CRP levels were important for identifying moderate and severe cases that may require oxygen administration. There was a clear difference in CRP levels between severity levels 3 and 4, but little difference between severity levels ≥4 (Figure 2E). The same was observed for eGFR-creatinine and lymphocyte values, with clear distinctions between severity levels 3 and 4, but minimal differences at higher severities. This suggests that inflammation, reflected by CRP and lymphocytes, and renal dysfunction, reflected by eGFR-creatinine, serve as entry points for severe disease; however, further progression in severity and mortality is mainly influenced by other factors. These findings are consistent with previous studies reporting that elevated CRP and decreased lymphocyte count or eGFR are associated with early COVID-19 severity [22,23]. Our results, however, further suggest that the progression to higher severity and mortality is predominantly driven by additional factors beyond these early indicators. Interestingly, creatinine was not ranked among the top 15 features only in the severity prediction model with the threshold ≥3, while it appeared consistently for other thresholds (Figure 2C). This phenomenon may be attributed to the high degree of information overlap between creatinine and eGFR-creatinine, the latter being a composite measure that incorporates both age and sex and may have provided a more robust signal under this specific threshold. Platelet and amylase levels are the most likely factors associated with the progression of disease severity and mortality, given their importance in prognostic models during hospitalization. Platelet counts were characteristically lower in several mortality cases than in other severe cases, and amylase levels were markedly elevated in some fatal cases (Figure 2E). Fatal cases characterized by these factors may represent patients in whom signs of poor prognosis are already present at initial diagnosis. However, it should be noted that some deaths did not show such signs at initial diagnosis, making them difficult to distinguish from severe or even mild cases.

Although the initial diagnosis model demonstrated fair discriminatory ability overall (AUC 0.717 for predicting severity ≥2), its performance remains below the generally accepted threshold for strong clinical utility (eg, ≥0.8), particularly when applied to low-risk patients. While this AUC is clearly better than random chance (AUC 0.5), it may be insufficient for dependable clinical decision-making when used in isolation. The relatively lower performance in low-risk groups suggests that predicting hospital admission based solely on initial laboratory values may be inherently limited, and that more nuanced clinical judgment—incorporating symptoms, physical findings, and contextual factors—may be necessary in such cases.

Several RSF predictors that varied early in mortality cases were identified. Platelet and PT values were among the top 10 early predictors in the RSF model for COVID-19 mortality. Patients with COVID-19 are prone to arterial and venous thrombosis, and in severe cases, disseminated intravascular coagulation (DIC) can complicate their condition and lead to death [24]. DIC was observed in only 0.6% of COVID-19 survivors, but it occurred frequently in 71.4% of mortality cases [22]. Our results support the idea that coagulation abnormalities, which may lead to DIC, are early predictors of COVID-19 mortality. Amylase was also identified as an important predictor of dynamic mortality risk. It is mainly secreted by the pancreas and salivary glands. Previous studies have reported elevated blood amylase levels in patients with severe COVID-19 [23]. Blood amylase levels are regulated by the balance between amylase production and clearance; elevated levels suggest damage to the producing tissues or impaired kidney function related to clearance. Such tissue damage may be a key event in the death of patients with COVID-19. Changes in blood cell composition were included among the RSF predictors, but these changes were not specific to patients who died (Figure 4). Platelets form platelet-neutrophil complexes when activated, which trigger neutrophil release via neutrophil extracellular traps [25,26]. Neutrophilia is not an independent risk factor for mortality but may act as a predisposing factor for DIC, ultimately leading to death. β-D-glucan is not only an indicator of fungal infection but also a marker of sepsis-related gastrointestinal leakage [27]. Even in the absence of fungemia or bacteremia, blood β-D-glucan levels during COVID-19 infection can induce neutrophil extracellular traps and may be associated with hypercytokinemia and severe inflammation [28]. By integrating these dynamically changing risk factors using RSF, we were able to accurately assess how close each patient was to death at each time point.

Previous studies have used the approach of building models based on data immediately before outcomes such as death or hospital discharge, and then applying these models to earlier time points to predict prognosis [21]. Although this approach can identify factors directly associated with mortality and achieves good predictive accuracy near the outcome, it has not

XSL•FO
RenderX

performed well for early prognostic prediction. Even among patients with COVID-19, at least 10 factors are associated with death, and their importance shifts over time: CRP and SpO_2 are more influential in the earlier phases, whereas β-D-glucan, platelets, and calcium become more important just before death. In patients with COVID-19, viral proliferation is considered the primary pathogenic mechanism during the first few days after onset, whereas the inflammatory response and coagulation abnormalities due to host immunity become the dominant factors after the first week [29]. This shift necessitates early administration of antiviral drugs and neutralizing antibodies during the initial phase, and the use of steroids and anticoagulation therapy after the first week [30,31]. For effective clinical decision-making based on the phase of hospitalization, it is essential to present mortality risk with distinct, phase-specific rationales.

Our prognostic framework can be applied to various aspects of disease prevention and treatment. Early prognostic prediction immediately after hospitalization, which screens for potentially severe cases with high sensitivity, can be used to estimate the number of hospital beds and the amount of medical equipment required. Additionally, owing to the recent shortage of hospital beds, the number of patients receiving treatment at home is increasing, and cases of sudden deterioration or death during home care have become a cause for concern. By using the RSF model to identify early signs of mortality risk, it may be possible to reduce deaths during home treatment and ensure safer management outside the hospital. Coagulation factors, which are key predictors in the RSF mortality model, represent potent targets for early intervention. Heparin administration has been reported to improve the prognosis of severe DIC in patients with COVID-19 [32]. The RSF model may help determine the appropriate timing for initiating anticoagulation therapy and evaluate treatment effects over time.

Although our study utilized a retrospective dataset, we designed the model to simulate real-time clinical application by dynamically updating the CHF at each time point during hospitalization (eg, day t, t+1, t+2). At each time point, only clinical data available up to that day were used, thereby emulating real-world scenarios in which patient information accumulates sequentially. The RSF model recalculates risk scores daily, enabling the timely identification of clinical deterioration and supporting dynamic decision-making. This time-updating mechanism reflects the evolving clinical status of patients and is essential for informing appropriate interventions and care adjustments during hospitalization.

While our framework was developed using data from patients with COVID-19, the underlying approach—time-varying survival modeling with interpretable outputs—is applicable to other acute or chronic diseases characterized by dynamic clinical trajectories. However, we acknowledge that this generalizability has not yet been empirically validated. Future work should explore its application to other conditions, such as sepsis, pneumonia, or heart failure, using disease-specific datasets to evaluate the framework's adaptability and robustness. To facilitate clinical integration, the model outputs could be embedded into electronic health record systems as real-time dashboards that display individualized risk trajectories and highlight contributing variables. For example, dynamic risk scores and SurvSHAP(t)-based feature summaries could be visualized alongside vital signs or laboratory trends to support triage decisions, early warnings, or escalation of therapy. Although we did not conduct a formal usability study, preliminary feedback from a small group of physicians who viewed the SurvSHAP(t) visualizations suggested that the temporal feature contributions enhanced their understanding of the rationale behind changes in risk over time. Future work should include structured surveys or qualitative interviews to assess how explainability tools influence clinician trust, understanding, and adoption in real-world settings.

In the interpretation of SurvSHAP(t) values, negative values indicate that a variable contributes to an increase in the predicted cumulative hazard, corresponding to a higher mortality risk. Conversely, positive values suggest a protective effect by lowering the predicted risk. SHAP-based feature importance, measured using both the mean absolute SHAP values (|SHAP|) and the mean aggregated absolute SurvSHAP(t) values, which average patient-level contributions over time, reflects the expected magnitude of each variable's influence on the model output across all observations. This provides insight into how sensitive the model's predictions are to each variable in aggregate and can be considered a model-based global sensitivity analysis method. In other words, a variable with a larger mean (aggregated) absolute SHAP value is expected to induce greater changes in the model's output when perturbed, indicating higher sensitivity to that feature.

As an additional aspect of sensitivity analysis, we examined the impact of different missing data scenarios. Preliminary results showed that, although prediction performance occasionally improved slightly without imputation, SHAP analysis revealed that the model was leveraging whether certain symptom variables were measured, essentially using the pattern of missingness as a predictive case. This indicates that the missingness mechanism was likely not at random and may have been biasedly associated with patient outcomes. To prevent overfitting to such data collection artifacts and ensure that the model relied on clinically meaningful features rather than exploiting missingness as a proxy, we opted to train all final models on imputed data.

We recognize that comorbidities may serve as confounders, influencing both predictor variables and patient outcomes. Although comorbidities were included as input features when available, we did not explicitly adjust for them using causal inference frameworks. Future extensions of this work should incorporate structured comorbidity indices (eg, Charlson or Elixhauser indices) and consider techniques such as inverse probability weighting or propensity score adjustment to reduce bias and enhance causal interpretability.

We designed the model with real-time applicability in mind. To assess its suitability for clinical implementation, we quantified the computation time (Table S3 in Multimedia Appendix 6). In our implementation, the learning and inference times for RSF were approximately 1.89 and 0.25 seconds, respectively. Cox regression required approximately 0.10

XSL•FO
RenderX

seconds for learning and 0.01 seconds for inference, making RSF roughly 20 times slower in both processes. However, the computational times for RSF remain sufficiently short to enable inference in real-world clinical settings and do not present a barrier to real-time clinical application. On the other hand, calculating the contribution of variables to the prediction results using SurvSHAP(t) took approximately 20,000 seconds (ie, over 5 hours) for the test dataset of 828 records. Even when converted to computation time per record, the processing remains substantial and may pose a barrier to its use in actual clinical practice. To support future clinical applications, a more efficient and faster implementation of SurvSHAP(t) is desirable.

To evaluate potential biases in dynamic mortality predictions across different demographic groups, we conducted subgroup analyses based on sex and age categories, as summarized in Table 3. Specifically, we assessed performance in the following subgroups: males only, a younger age group (<65 years), and an older age group (≥75 years). The results indicated a slight decrease in predictive performance among males compared with the overall cohort. By contrast, the younger age group showed a marginal performance improvement, likely reflecting the lower incidence of mortality in this subgroup. Conversely, the older age group exhibited a slight reduction in predictive performance, possibly due to the higher mortality rate in this population. Nonetheless, the performance differences across these subgroups were relatively minor compared with the overall analysis, suggesting that the model does not display substantial bias with respect to sex or age. It should also be noted that our dataset consisted exclusively of Japanese patients, limiting our ability to evaluate predictive fairness across different ethnic groups. We acknowledge this limitation and recognize the need for further validation in more diverse populations.

Table . Performance of mortality prediction in subgroup analyses.

Subgroup	Concordance index
All	0.941
Males only	0.919
Younger age group (<65 years)	0.957
Older age group (≥75 years)	0.864

This study had several limitations. First, the study cohort was relatively small, and data were obtained from a single hospital. As machine learning model performance and robustness are highly influenced by data size and diversity, our models may be prone to overfitting to this specific cohort. As a result, the identified predictive patterns could lack stability and generalizability when applied to other populations without retraining. Furthermore, the substantial reliance on data imputation to address missing values, particularly for certain biomarkers, introduces an additional layer of potential model instability and overfitting risk, as the models may inadvertently learn patterns from the imputation process itself rather than from true underlying biological signals. Consequently, the model will require updates and rigorous revalidation as the study expands to multiple centers and includes a larger sample size, ideally with efforts to minimize the extent of missing data. Second, we included only patients with COVID-19 diagnosed in 2020. Therefore, it is necessary to evaluate the generalizability of these models, including their applicability to different SARS-CoV-2 variants and the effects of vaccination. Notably, the data used in this study were collected during a period when vaccination was rarely administered, and can thus be considered valuable in representing disease trajectories unaffected by vaccination. Third, the study treated daily records from the same patient as independent data points for training, which presents a technical challenge for RSF. Incorporating the risk estimated on the previous day may enable more accurate predictions.

To advance the clinical utility and generalizability of the model, we plan to conduct a multicenter study involving hospitals with diverse patient populations and care protocols. This will enable retraining and external validation of the model across different clinical settings. Additionally, we aim to prospectively evaluate the model in a clinical pilot study, focusing on real-time integration, clinician usability, and assessment of treatment impact through interventional trials.

Conclusions

We developed and validated an interpretable 2-model framework to enhance prognostic support for patients with COVID-19. The framework integrates a LightGBM model for predicting disease severity at admission with an RSF model that provides daily mortality risk estimates throughout hospitalization. These complementary models are designed to address distinct clinical needs at different stages of patient care.

Central to our approach is interpretability through SurvSHAP(t), which identified actionable, phase-specific predictors, such as elevated CRP levels at admission, declining SpO_2 during progression, and late-stage indicators such as platelet loss accompanied by elevated β-D-glucan. These time-dependent explanations empower clinicians to anticipate and respond to evolving patient conditions with greater precision.

By combining early disease stratification with continuous, explainable mortality prediction, our framework facilitates timely interventions and strategic resource allocation. This demonstrates significant potential to improve clinical decision-making in dynamic care environments and may serve as a foundation for managing future infectious disease outbreaks with heterogeneous trajectories, pending prospective multi-institutional validation.

XSL•FO
RenderX

Acknowledgments

We thank Elsevier Language Editing Services for reviewing the manuscript to eliminate grammatical and spelling errors and to ensure adherence to correct scientific English. This study was supported by grants from the Japan Society for the Promotion of Science (JSPS KAKENHI: grants JP20K21837, JP21K02356, and JP25K21344 to TI); the Japan Science and Technology Agency (JST) Moonshot R&D Program (grant JPMJMS2025 to EK); the JST CREST Program (grant JPMJCR20H4 to EK); and the Japan Agency for Medical Research and Development (AMED: grants JP21wm0325007, JP20fk0108412, JP20fk0108413, and JP21gm5010003 to EK). The funders had no role in the study design, data collection, data analysis and interpretation, or writing of the manuscript. The authors confirm that no generative artificial intelligence tools were used in the original drafting of the manuscript. During the revision phase only, artificial intelligence–based language polishing assistance was obtained from ChatGPT (model o3, OpenAI, 2025), Claude (Sonnet 4, Anthropic, 2025), and Gemini (2.5 Pro, Google, 2025). All suggestions provided by these tools were independently reviewed, fact-checked, and manually edited by the authors before inclusion. The authors remain fully responsible for the accuracy and integrity of the final text. This study was presented as a poster at the 6th Workshop on Virus Dynamics, held in July 2023.

Data Availability

The datasets used or analyzed during this study are available from the corresponding author (EK) upon reasonable request. The underlying code and the training/test datasets are not publicly available but may be made available to qualified researchers upon reasonable request from the corresponding author.

Authors' Contributions

TI, M Shinoda, KS, EK, and M Shinkai designed the study. M Shinoda, KA, SO, KK, and M Shinkai collected the data. TI, MO, and EK conducted the data analysis. All authors contributed to data interpretation, manuscript writing, and critical review, and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Descriptive statistics of 84 variables—including background information, symptoms, and blood and urine test data—of patients with COVID-19 for prognostic screening at initial diagnosis.
[XLSX File, 18 KB - formative_v9i1e65585_app1.xlsx]

Multimedia Appendix 2
Descriptive statistics of 72 variables—including background information, vital signs, and blood and urine test data—of patients with COVID-19 for dynamic mortality risk assessment during hospitalization.
[XLSX File, 15 KB - formative_v9i1e65585_app2.xlsx]

Multimedia Appendix 3
Receiver operating characteristic curves, confusion matrices, and absolute values of the cumulative hazard functions.
[PDF File, 541 KB - formative_v9i1e65585_app3.pdf]

Multimedia Appendix 4
Performance of hospitalization prediction models, including Light Gradient Boosting Machine, random forest, support vector machine, Elastic Net, and decision tree.
[PDF File, 120 KB - formative_v9i1e65585_app4.pdf]

Multimedia Appendix 5
Detailed methods covering missing data imputation, hyperparameter optimization, ensemble averaging, evaluation metrics, model selection and interpretability rationale, and clinical justification of selected predictors.
[PDF File, 125 KB - formative_v9i1e65585_app5.pdf]

Multimedia Appendix 6
Comparison of computation time for training, prediction, and Shapley Additive Explanations–based interpretation in survival analysis.
[PDF File, 333 KB - formative_v9i1e65585_app6.pdf]

XSL•FO
RenderX

References

1. Forster AJ, Murff HJ, Peterson JF, Gandhi TK, Bates DW. The incidence and severity of adverse events affecting patients after discharge from the hospital. Ann Intern Med 2003 Feb 4;138(3):161-167. [doi: 10.7326/0003-4819-138-3-200302040-00007] [Medline: 12558354]
2. Odell M, Victor C, Oliver D. Nurses' role in detecting deterioration in ward patients: systematic literature review. J Adv Nurs 2009 Oct;65(10):1992-2006. [doi: 10.1111/j.1365-2648.2009.05109.x] [Medline: 20568317]
3. Mok WQ, Wang W, Liaw SY. Vital signs monitoring to detect patient deterioration: an integrative literature review. Int J Nurs Pract 2015 May;21 Suppl 2:91-98. [doi: 10.1111/ijn.12329] [Medline: 26125576]
4. Shang Y, Pan C, Yang X, et al. Management of critically ill patients with COVID-19 in ICU: statement from front-line intensive care experts in Wuhan, China. Ann Intensive Care 2020 Jun 6;10(1):73. [doi: 10.1186/s13613-020-00689-1] [Medline: 32506258]
5. Adhikari NKJ, Fowler RA, Bhagwanjee S, Rubenfeld GD. Critical care and the global burden of critical illness in adults. Lancet 2010 Oct 16;376(9749):1339-1346. [doi: 10.1016/S0140-6736(10)60446-1] [Medline: 20934212]
6. Perotte A, Ranganath R, Hirsch JS, Blei D, Elhadad N. Risk prediction for chronic kidney disease progression using heterogeneous electronic health record data and time series analysis. J Am Med Inform Assoc 2015 Jul;22(4):872-880. [doi: 10.1093/jamia/ocv024] [Medline: 25896647]
7. Renu K, Prasanna PL, Valsala Gopalakrishnan A. Coronaviruses pathogenesis, comorbidities and multi-organ damage - a review. Life Sci 2020 Aug 15;255:117839. [doi: 10.1016/j.lfs.2020.117839] [Medline: 32450165]
8. Nielsen AB, Thorsen-Meyer HC, Belling K, et al. Survival prediction in intensive-care units based on aggregation of long-term disease history and acute physiology: a retrospective study of the Danish National Patient Registry and electronic patient records. Lancet Digit Health 2019 Jun;1(2):e78-e89. [doi: 10.1016/S2589-7500(19)30024-X] [Medline: 33323232]
9. Johnson AEW, Ghassemi MM, Nemati S, Niehaus KE, Clifton DA, Clifford GD. Machine learning and decision support in critical care. Proc IEEE Inst Electr Electron Eng 2016 Feb;104(2):444-466. [doi: 10.1109/JPROC.2015.2501978] [Medline: 27765959]
10. Tomašev N, Glorot X, Rae JW, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. Nature New Biol 2019 Aug;572(7767):116-119. [doi: 10.1038/s41586-019-1390-1] [Medline: 31367026]
11. Thorsen-Meyer HC, Nielsen AB, Nielsen AP, et al. Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records. Lancet Digit Health 2020 Apr;2(4):e179-e191. [doi: 10.1016/S2589-7500(20)30018-2] [Medline: 33328078]
12. Quinn TP, Jacobs S, Senadeera M, Le V, Coghlan S. The three ghosts of medical AI: can the black-box present deliver? Artif Intell Med 2022 Feb;124:102158. [doi: 10.1016/j.artmed.2021.102158] [Medline: 34511267]
13. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Presented at: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems; Dec 4-9, 2017; Long Beach, CA p. 4768-4777.
14. Grein J, Ohmagari N, Shin D, et al. Compassionate use of remdesivir for patients with severe COVID-19. N Engl J Med 2020 Jun 11;382(24):2327-2336. [doi: 10.1056/NEJMoa2007016] [Medline: 32275812]
15. Ke G, Meng Q, Finley T, et al. LightGBM: a highly efficient gradient boosting decision tree. In: Guyon I, editor. Advances in Neural Information Processing Systems: Curran Associates, Inc; 2017, Vol. 30.
16. Friedman JH. Greedy function approximation: A gradient boosting machine. Ann Statist 2001;29(5):1189-1232. [doi: 10.1214/aos/1013203451]
17. Stekhoven DJ, Bühlmann P. MissForest--non-parametric missing value imputation for mixed-type data. Bioinformatics 2012 Jan 1;28(1):112-118. [doi: 10.1093/bioinformatics/btr597] [Medline: 22039212]
18. Akiba T, Sano S, Yanase T, et al. Optuna: a next-generation hyperparameter optimization framework. Presented at: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; Aug 4-8, 2019; Anchorage, AK p. 2623-2631. [doi: 10.1145/3292500.333070]
19. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. Ann Appl Stat 2008 Sep;2(3):841-860. [doi: 10.1214/08-AOAS169]
20. Krzyziński M, Spytek M, Baniecki H, Biecek P. SurvSHAP(t): time-dependent explanations of machine learning survival models. Knowl Based Syst 2023 Feb;262:110234. [doi: 10.1016/j.knosys.2022.110234]
21. Yan L, Zhang HT, Goncalves J, et al. An interpretable mortality prediction model for COVID-19 patients. Nat Mach Intell 2020;2(5):283-288. [doi: 10.1038/s42256-020-0180-7] [Medline: 32285025]
22. Tang N, Li D, Wang X, Sun Z. Abnormal coagulation parameters are associated with poor prognosis in patients with novel coronavirus pneumonia. J Thromb Haemost 2020 Apr;18(4):844-847. [doi: 10.1111/jth.14768] [Medline: 32073213]
23. Liu F, Long X, Zhang B, Zhang W, Chen X, Zhang Z. ACE2 expression in pancreas may cause pancreatic damage after SARS-CoV-2 infection. Clin Gastroenterol Hepatol 2020 Aug;18(9):2128-2130. [doi: 10.1016/j.cgh.2020.04.040]
24. Fraissé M, Logre E, Pajot O, Mentec H, Plantefève G, Contou D. Thrombotic and hemorrhagic events in critically ill COVID-19 patients: a French monocenter retrospective study. Crit Care 2020 Jun 2;24(1):275. [doi: 10.1186/s13054-020-03025-y] [Medline: 32487122]
25. Guo L, Rondina MT. The era of thromboinflammation: platelets are dynamic sensors and effector cells during infectious diseases. Front Immunol 2019;10:2204. [doi: 10.3389/fimmu.2019.02204] [Medline: 31572400]

XSL•FO
RenderX

26. Kimball AS, Obi AT, Diaz JA, Henke PK. The emerging role of NETs in venous thrombosis and immunothrombosis. Front Immunol 2016;7:236. [doi: 10.3389/fimmu.2016.00236] [Medline: 27446071]
27. Leelahavanichkul A, Worasilchai N, Wannalerdsakun S, et al. Gastrointestinal leakage detected by serum (1→3)-β-D-glucan in mouse models and a pilot study in patients with sepsis. Shock 2016 Nov;46(5):506-518. [doi: 10.1097/SHK.0000000000000645] [Medline: 27172153]
28. Saithong S, Worasilchai N, Saisorn W, et al. Neutrophil extracellular traps in severe SARS-CoV-2 infection: a possible impact of LPS and (1→3)-β-D-glucan in blood from gut translocation. Cells 2022 Mar 24;11(7):1103. [doi: 10.3390/cells11071103] [Medline: 35406667]
29. Siddiqi HK, Mehra MR. COVID-19 illness in native and immunosuppressed states: a clinical-therapeutic staging proposal. J Heart Lung Transplant 2020 May;39(5):405-407. [doi: 10.1016/j.healun.2020.03.012] [Medline: 32362390]
30. Gandhi RT, Lynch JB, Del Rio C. Mild or Moderate COVID-19. N Engl J Med 2020 Oct 29;383(18):1757-1766. [doi: 10.1056/NEJMcp2009249] [Medline: 32329974]
31. Murakami N, Hayden R, Hills T, et al. Therapeutic advances in COVID-19. Nat Rev Nephrol 2023;19(1):38-52. [doi: 10.1038/s41581-022-00642-4]
32. Tang N, Bai H, Chen X, Gong J, Li D, Sun Z. Anticoagulant treatment is associated with decreased mortality in severe coronavirus disease 2019 patients with coagulopathy. J Thromb Haemost 2020 May;18(5):1094-1099. [doi: 10.1111/jth.14817] [Medline: 32220112]

Abbreviations

AUC: area under the receiver operating characteristic curve
C-index: concordance index
CHF: cumulative hazard function
CRP: C-reactive protein
DIC: disseminated intravascular coagulation
ECMO: extracorporeal membrane oxygenation
eGFR: estimated glomerular filtration rate
GBDT: gradient boosting decision trees
LASSO: least absolute shrinkage and selection operator
LightGBM: Light Gradient Boosting Machine
NMI: Nature Machine Intelligence
PT: prothrombin time
RSF: random survival forests
SHAP: Shapley Additive Explanation
SpO_2: peripheral oxygen saturation

Edited by J Sarvestan; submitted 20.08.24; peer-reviewed by A Mete, PH Kasani, S Venkatesh; revised version received 26.05.25; accepted 28.05.25; published 07.08.25.

Please cite as:
Ishikawa T, Shinoda M, Oya M, Ashizaki K, Ota S, Kamachi K, Sakurada K, Kawakami E, Shinkai M
Explainable Machine Learning Framework for Dynamic Monitoring of Disease Prognostic Risk: Retrospective Cohort Study
JMIR Form Res 2025;9:e65585
URL: https://formative.jmir.org/2025/1/e65585
doi:10.2196/65585

© Tetsuo Ishikawa, Masahiro Shinoda, Megumi Oya, Koichi Ashizaki, Shinichiro Ota, Kenichi Kamachi, Kazuhiro Sakurada, Eiryo Kawakami, Masaharu Shinkai. Originally published in JMIR Formative Research (https://formative.jmir.org), 7.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Quantifying the Regional Disproportionality of COVID-19 Spread: Modeling Study

Kenji Sasaki[1], PhD; Yoichi Ikeda[1], PhD; Takashi Nakano[1,2], PhD

[1]Center for Infectious Disease Education and Research, Osaka University, Co-creation BLDG. D88-1, 2-1 Yamadaoka, Suita, Osaka, Japan
[2]Research Center for Nuclear Physics, Osaka University, Ibaraki, Osaka, Japan

Corresponding Author:
Kenji Sasaki, PhD
Center for Infectious Disease Education and Research, Osaka University, Co-creation BLDG. D88-1, 2-1 Yamadaoka, Suita, Osaka, Japan

Abstract

Background: The COVID-19 pandemic has caused serious health, economic, and social consequences worldwide. Understanding how infectious diseases spread can help mitigate these impacts. The Theil index, a measure of inequality rooted in information theory, is useful for identifying geographic disproportionality in COVID-19 incidence across regions.

Objective: This study focused on capturing the degrees of regional disproportionality in incidence rates of infectious diseases over time. Using the Theil index, we aim to assess regional disproportionality in the spread of COVID-19 and detect epicenters where the number of infected individuals was disproportionately concentrated.

Methods: To quantify the degree of disproportionality in the incidence rates, we applied the Theil index to the publicly available data of daily confirmed COVID-19 cases in the United States over a 1100-day period. This index measures relative disproportionality by comparing daily regional case distributions with population proportions, thereby identifying regions where infections are disproportionately concentrated.

Results: Our analysis revealed a dynamic pattern of regional disproportionality in the confirmed cases by monitoring variations in regional contributions to the Theil index as the pandemic progressed. Over time, the index reflected a transition from localized outbreaks to widespread transmission, with high values corresponding to concentrated cases in some regions. We also found that the peaks in the Theil index often preceded surges in confirmed cases, suggesting its potential utility as an early warning signal.

Conclusions: This study demonstrated that the Theil index is one of the effective indices for quantifying regional disproportionality in COVID-19 incidence rates. Although the Theil index alone cannot fully capture all aspects of pandemic dynamics, it serves as a valuable tool when used alongside other indicators such as infection and hospitalization rates. This approach allows policy makers to monitor regional disproportionality efficiently, offering insights for early intervention and targeted resource allocation.

(JMIR Form Res 2025;9:e59230) doi:10.2196/59230

KEYWORDS

infectious disease; COVID-19; epidemiology; public health; SARS-CoV-2; pandemic; inequality measure; information theory; Kullback-Leibler divergence

Introduction

The COVID-19 pandemic has caused serious health problems and has had major economic and social consequences worldwide. It has highlighted the need to understand regional disparities in infection rates to strengthen public health responses since infection dynamics are influenced by factors such as population density, socioeconomic conditions, and health care infrastructure [1,2]. Numerous indicators and models have been proposed to address the problem, and mechanisms for the spread of the infection and intervention measures to control the pandemic have been studied [3-9].

Several recent studies have investigated regional differences in COVID-19 prevalence [10-13]. Differences in the prevalence rates between regions highlight the need to understand regional inequalities in pandemic response strategies. Effectively addressing these disparities requires accurate quantification and understanding of regional disproportionalities in daily confirmed COVID-19 cases.

In the field of economics, various indicators have been developed to measure resource and income inequality, including an index proposed by Theil, which incorporated information theory [14]. Manz and Mansmann [15] have demonstrated the importance of using inequality indices for monitoring changes in geographic inequality; for instance, the Theil index was used to track geographic disproportionality over time during the COVID-19 pandemic, providing important insights for public health policy.

XSL•FO
RenderX

The aim of this paper is to quantify the interregional disproportionality in the number of confirmed cases using the Theil index, which mathematically corresponds to the Kullback-Leibler (KL) divergence in information theory [16]. The Theil index is an effective method of measuring the degree of disproportionality and objectively assessing biases in the interregional distribution of infected individuals.

Methods

Overview

We analyzed the time trends of daily COVID-19–confirmed cases in the United States over 1100 days since the first reported case on January 21, 2020 [17]. Data are taken from the COVID-19 data repository at the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University [18]. US state population data were obtained from the US Census Bureau website [19]. Population changes due to migration, births, and deaths were not considered in the analysis.

The Disproportionality Measure: Theil Index

The Theil index is commonly applied in various fields including economics, sociology, and information theory. It quantifies the relative differences between various components of a dataset. In the context of regional analysis of the confirmed cases, the Theil index can be employed to evaluate the distribution of infected individuals across different regions. In this study, we use the Theil index to identify regions with disproportionate numbers of confirmed cases relative to their population size.

The discrete form of the Theil index is expressed as:

$$T=\sum_{i=1}^{N} t_i=\sum_{i=1}^{N} p_i \ln \frac{p_i}{q_i}$$

where N is the total number of regions being considered and ln is the natural logarithm. The Theil index T is composed of a sum of t_i which is a partial contribution from region i. The discrete probability distribution, p_i, in region i is defined as the ratio of daily confirmed cases region i to the total confirmed cases across all regions for that day. Similarly, the population ratio, q_i, in region i is the ratio of the population in region i to the total population across all regions.

The Theil index, which is mathematically related to the KL divergence, is a nonsymmetric metric that measures the relative entropy or informational difference between two distributions. It is sensitive to the interregional variations in the distribution of the confirmed cases, with its maximum value attained when the confirmed cases are concentrated in areas with the smallest population proportion. Consequently, the index tends to exhibit higher values when a small number of regions account for a large share of the confirmed cases, and conversely, lower values when the confirmed cases are more evenly distributed across regions. Notably, it remains nonnegative and reaches a minimum value of 0 only when the two distributions are identical. Therefore, applying the Theil index to the time-series data of the confirmed cases, and monitoring changes in the index over time, we quantified the degree of spread of COVID-19 cases and assessed whether the confirmed cases were disproportionately concentrated relative to the regional population sizes over time.

Ethical Considerations

This study used publicly available, deidentified COVID-19 data from CSSE at Johns Hopkins University [18], and therefore, additional ethics approval and informed consent were not required. The aggregated data ensured privacy and confidentiality, and no direct human participants were involved; thus, no compensation was provided. No identifiable information appears in any images or materials.

Results

To address fluctuations in the Theil index caused by data aggregation inconsistencies during holidays across different regions, the 7-day average of confirmed COVID-19 cases was used instead of raw data.

Figure 1 illustrates a 2-axis graph showing the time trends of the Theil index (left axis) and the number of confirmed cases in a logarithmic scale (right axis, logarithmic scale). The horizontal axis represents the number of days elapsed (denoted by d in the text) since the date of the first reported case in the United States.

In Figure 1, there are eight notable surges of the confirmed cases, occurring at approximately d=80 (first), 180 (second), 350 (third), 450 (fourth), 580 (fifth), 720 (sixth), 900 (seventh), and 1080 (eighth), respectively. The presence of multiple peaks in the Theil index indicates that infected individuals were concentrated in specific regions during the period, and the degree of this concentration can be assessed by examining the numerical values. However, it is important to note that changes in the Theil index simply indicate the degree of regional disproportionality in the confirmed cases rather than absolute increases or decreases in the number of infected individuals. Therefore, this indicator is most effective when interpreted in conjunction with actual trends in the number of confirmed cases.

Before the first peak, the number of confirmed cases was quite low, and the Theil index fluctuated erratically. As d increased near the first peak, the Theil index gradually decreased, reaching a local minimum around d=120. This suggested that the initially localized epidemic began to spread throughout the US during the early stages of the global pandemic. Similar trends were observed during subsequent surges, such as slight a increase in the Theil index before the peak, followed by a decrease. This could be seen as a precursor to a surge in the number of infected individuals. This finding aligns with previous research by Ikeda, Sasaki, and Nakano [7].

The following examples provide interesting insights; when the Theil index value was high and the number of confirmed cases was low (d=60, 550, etc), it indicated that the infectious disease was localized and beginning to spread to various regions. Conversely, when the index was low and the number of confirmed cases was high (d=750, etc), it indicated that there was no obvious epicenter of the infectious disease, with the number of confirmed cases increasing relatively and evenly across different regions.

The contributions to the Theil index from each region (t_i), calculated from the number of cases on each date, were arranged

XSL•FO
RenderX

in chronological order and visualized using a heatmap, as shown in Figure 2. Regions with a high proportion of confirmed cases are represented in red, while regions with a low proportion are colored blue. Notably, there are long intervals between the deep red patches in some regions such as California, Florida, and New York. Particularly, the periods of intense infection represented by these deep red patches were not repeated at short intervals. This phenomenon is of great importance in infectious disease management. Once a major epidemic in an area has subsided, the interval between subsequent outbreaks provides an opportunity to rebuild the health care infrastructure and implement preventive measures before the occurrence of the next epidemic.

Based on the observations from Figure 2, the epicenter of infectious diseases as indicated by the red patches alternates between New York, California, and Florida. This insight is crucial for understanding the underlying mechanisms of the spread of infectious diseases in the future. Furthermore, after d=750, both the red and blue colors fade over time, indicating the absence of a single epicenter, and a widespread outbreak of COVID-19. This pattern suggests the ineffectiveness of countermeasures against the spread of infectious diseases under these circumstances.

Figure 3A shows the contributions to the Theil index by region at d=60. The horizontal axis in the figure shows the state code (as listed in Multimedia Appendix 1). There is a significant contribution to the Theil index from New York State compared to the other regions. Figure 3B shows that at this point confirmed cases were highly localized in these regions.

There were relatively large negative contributions to the Theil index from California, Florida, and Texas, which were regions with high population ratios. It is interesting to note that there was little risk of infection in these regions at that point; however, the number of infected individuals rapidly increased following the concentration of confirmed cases in New York.

Figure 4 shows the contributions to the Theil index from each region at d=550 and 750. At d=550 shown in Figure 4A, the Theil index reaches a peak, and the trend of confirmed cases is increasing. This suggests that a new epidemic is emerging, mainly in Florida and Louisiana. However, their contributions are significantly smaller compared to New York at d=60, as seen in Figure 3. This indicates that regional disproportionality is much less pronounced than in the early stage of the COVID-19 pandemic. It is also interesting to look at data on d=750 as shown in Figure 4B, when confirmed cases in the United States are at their maximum. Although several regions show large contributions to the Theil index, the epicenter of COVID-19 is no longer obvious, suggesting that COVID-19 cases are uniformly distributed across the country.

Figure 1. Time trends of the Theil index on the left axis and the 7-day average number of the confirmed cases on the right axis on a logarithmic scale are shown in the red and blue curves, respectively. The horizontal axis is the number of days elapsed since January 21, 2020.

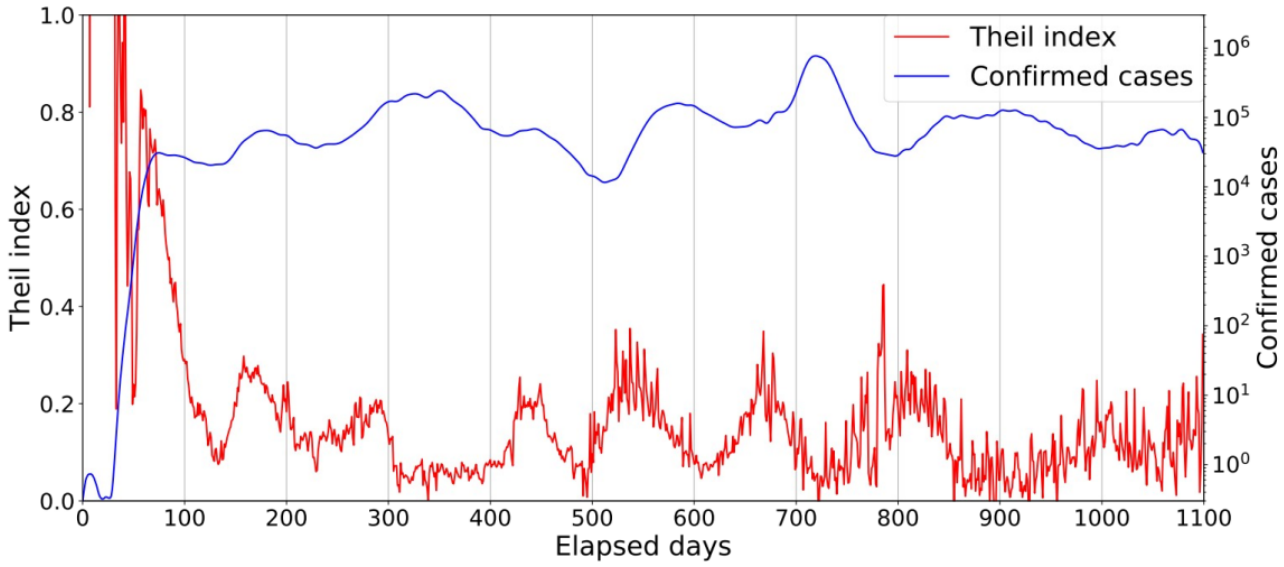

XSL•FO
RenderX

Figure 2. Partial contributions to the Theil index from each region, t_i, are displayed in a heatmap over time. The horizontal axis represents the number of days elapsed since January 21, 2020. The vertical axis shows the names of states in the United States. The positive (high concentration of incidences) and negative (low concentration) contributions to the Theil index correspond to deep red and blue colors, respectively.

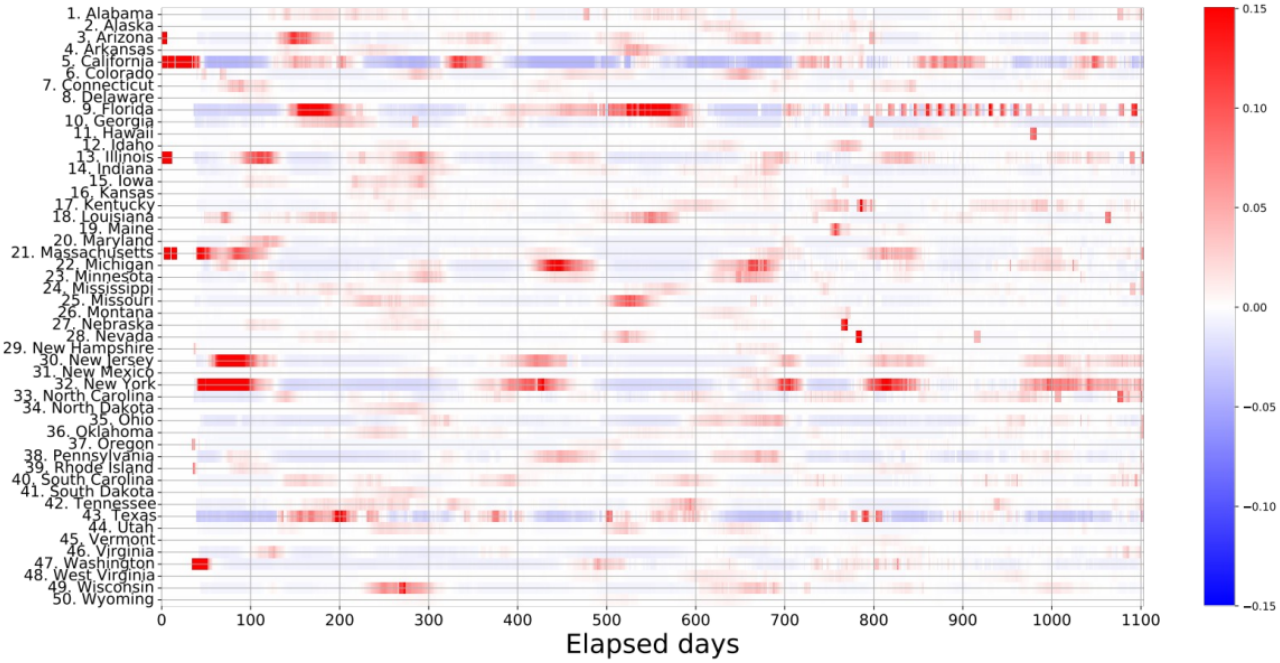

Figure 3. Partial contributions to the Theil index from each region, t_i, at d=60. The horizontal axis shows the state code given in Multimedia Appendix 1. (A) Comparison of the distribution of the confirmed cases and population. The vertical axis shows the ratio of a part to the whole region for populations and for the confirmed cases. (B) Contributions to the Theil index from each region. The vertical axis shows the strength of the contribution to the Theil index. The significantly high value of partial contribution to the Theil index is highlighted in orange.

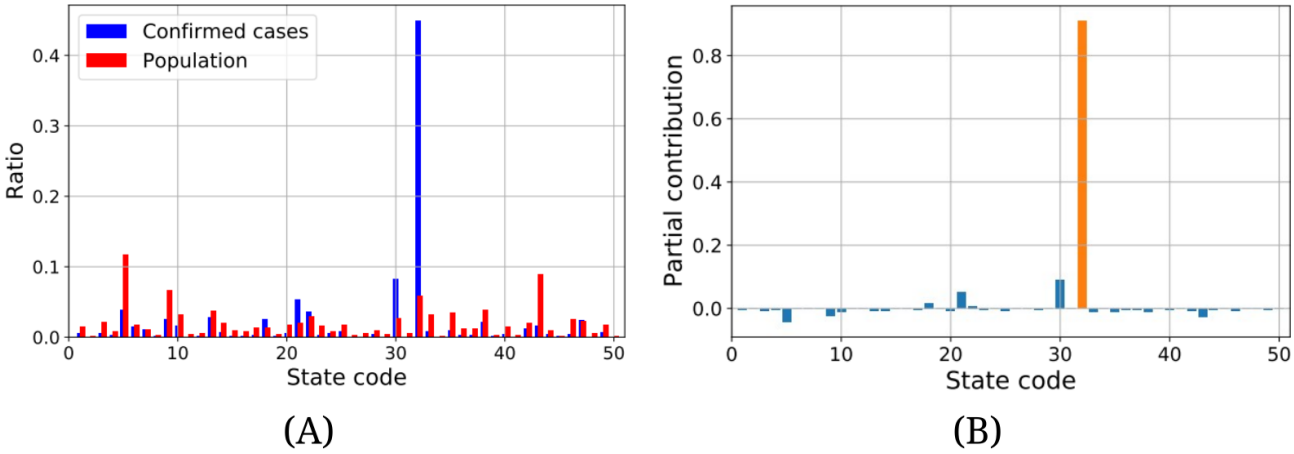

XSL•FO
RenderX

Figure 4. Partial contributions of the Theil index from each region, t_i at a specific date. The vertical axis shows the strength of contribution to the Theil index. The horizontal axis shows the state code given in Multimedia Appendix 1. (A) Contributions of the Theil index at d=550. (B) Contributions of the Theil index at d=750. The significantly high values of partial contribution to the Theil index are highlighted in orange.

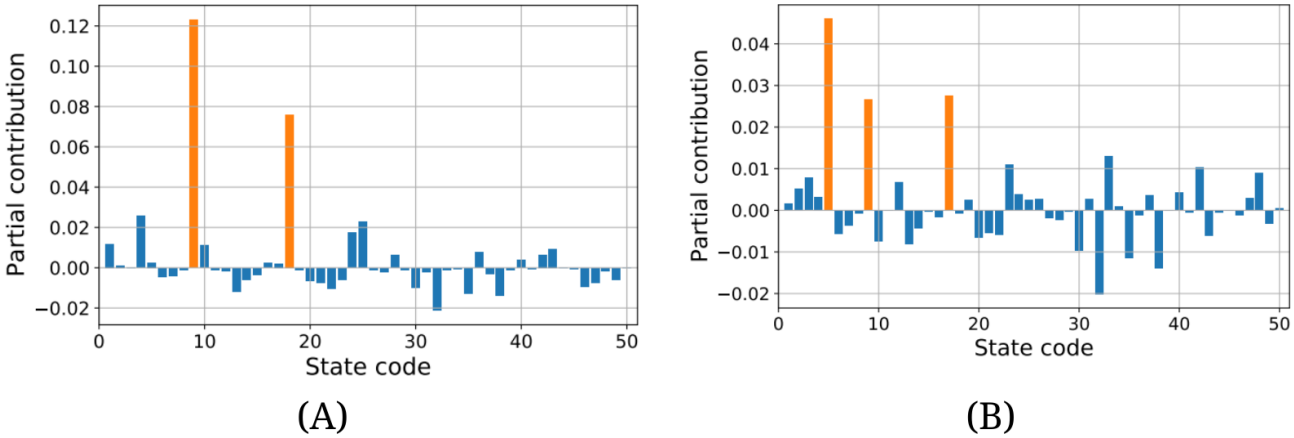

Discussion

This study demonstrates the utility of the Theil index for quantifying regional disproportionalities in the distribution of COVID-19 cases. It offers an intuitive and efficient approach to identifying hotspots and monitoring the spread of infection. However, certain limitations may affect result interpretation.

The accuracy of the analysis depends on data quality; factors such as underreporting, delays in case confirmation, and regional differences in testing capacity may introduce biases into case counts. These issues could potentially impact the calculated Theil index and the assessment of regional disproportionalities.

Additionally, this study focuses primarily on confirmed cases rather than new infections, limiting its capacity to predict future spread. Therefore, the Theil index alone may not be sufficient for determining the timing and location of public health interventions, such as isolation measures. To support comprehensive policy-making, it should be used alongside other indicators, such as infection rates, hospitalization rates, and health care capacity.

Conventional spatiotemporal analysis methods [20,21] are widely used in epidemiology and public health to track infectious disease spread and visualize infection clusters over time in specific regions. These established tools effectively detect geographical clusters, identify areas with unusually high incidence, and reveal disease hotspots within defined spatial ranges.

In contrast, our method offers two distinct advantages. First, an increase in the Theil index acts as a precursor to a surge in the number of infected individuals. Second, it quantifies regional disproportionalities in incidence rates at any given time. Unlike conventional methods that emphasize physical distance and spatial proximity, our approach treats regions as discrete categories to calculate incidence rate disproportionalities. Although simple, this approach provides an intuitive way to identify epicenters at a lower computational cost compared to spatiotemporal scanning, enabling us to detect early surges in confirmed cases and pinpoint regions with concentrated infections.

For instance, the concentration of COVID-19 cases in New York at d=60 as shown in Figure 3A and B, cannot be overlooked when considering infection control. The lockdown was implemented in New York City [22] and coincided with a period when the contribution to the Theil index was concentrated in New York State. Although it is challenging to assess the direct impact of lockdown using the Theil index alone, the timing appears appropriate based on the pattern of concentration of confirmed cases.

Integrating our method with additional data sources, such as mobility patterns and health care capacity, will enhance pandemic response strategies, particularly for early intervention and efficient resource allocation.

In conclusion, this study demonstrates the application of the Theil index in quantifying regional disproportionalities in confirmed cases and monitoring their evolution over time. By analyzing confirmed case data in the United States, we have identified patterns of disproportionalities, specified epicenters, and characterized localized outbreaks.

Continued monitoring and analysis of regional differences in COVID-19 transmission remain essential, especially considering emerging variants and evolving public health responses. Our findings highlight the importance of understanding the regional dynamics of infected individuals for effective pandemic response interventions.

Incorporating the findings of this study will help policy makers refine strategies and address the diverse needs of different regions, ultimately increasing the effectiveness of pandemic response efforts and mitigating the impact of future health crises.

Lastly, the decomposability of the Theil index makes it possible to quantify and compare disproportionality in groups with specific characteristics, such as age, vaccination coverage, and health care accessibility. Identifying these disproportionalities will provide important insights for future pandemic responses.

XSL•FO
RenderX

Acknowledgments

We thank the members of the Division of Scientific Information and Public Policy (SiPP) at the Center for Infectious Disease Education and Research (CiDER) Osaka University for useful discussions. This research was supported by The Nippon Foundation—Osaka University Project for Infectious Disease Prevention.

Data Availability

Data were derived from public resources.

Conflicts of Interest

None declared.

Multimedia Appendix 1
US states codes.
[PNG File, 58 KB - formative_v9i1e59230_app1.png]

References

1. Lopez L, Hart LH, Katz MH. Racial and ethnic health disparities related to COVID-19. JAMA 2021 Feb 23;325(8):719-720. [doi: 10.1001/jama.2020.26443] [Medline: 33480972]
2. Adhikari S, Pantaleo NP, Feldman JM, Ogedegbe O, Thorpe L, Troxel AB. Assessment of community-level disparities in coronavirus disease 2019 (COVID-19) infections and deaths in large US metropolitan areas. JAMA Netw Open 2020 Jul 1;3(7):e2016938. [doi: 10.1001/jamanetworkopen.2020.16938] [Medline: 32721027]
3. Perkins TA, España G. Optimal control of the COVID-19 pandemic with non-pharmaceutical interventions. Bull Math Biol 2020 Sep 4;82(9):118. [doi: 10.1007/s11538-020-00795-y] [Medline: 32888118]
4. Nakano T, Ikeda Y. Novel indicator to ascertain the status and trend of COVID-19 spread: modeling study. J Med Internet Res 2020 Nov 30;22(11):e20144. [doi: 10.2196/20144] [Medline: 33180742]
5. Buchy P, Buisson Y, Cintra O, et al. COVID-19 pandemic: lessons learned from more than a century of pandemics and current vaccine development for pandemic control. Int J Infect Dis 2021 Nov;112:300-317. [doi: 10.1016/j.ijid.2021.09.045] [Medline: 34563707]
6. Stenseth NC, Dharmarajan G, Li R, Shi ZL, Yang R, Gao GF. Lessons learnt from the COVID-19 pandemic. Front Public Health 2021;9:694705. [doi: 10.3389/fpubh.2021.694705] [Medline: 34409008]
7. Ikeda Y, Sasaki K, Nakano T. A new compartment model of COVID-19 transmission: the broken-link model. Int J Environ Res Public Health 2022 Jun 3;19(11):6864. [doi: 10.3390/ijerph19116864] [Medline: 35682447]
8. Pandemic preparedness. Nature. 2022 Oct 26. URL: https://www.nature.com/collections/jaacfgeief [accessed 2024-03-14]
9. Sasaki K, Ikeda Y, Nakano T. The effects of behavioral restrictions on the spread of COVID-19. Rep 2022;5(4):37. [doi: 10.3390/reports5040037]
10. Jinjarak Y, Ahmed R, Nair-Desai S, Xin W, Aizenman J. Accounting for global COVID-19 diffusion patterns, January-April 2020. Econ Disaster Clim Chang 2020;4(3):515-559. [doi: 10.1007/s41885-020-00071-2] [Medline: 32901228]
11. Mollalo A, Vahedi B, Rivera KM. GIS-based spatial modeling of COVID-19 incidence rate in the continental United States. Sci Total Environ 2020 Aug 1;728:138884. [doi: 10.1016/j.scitotenv.2020.138884] [Medline: 32335404]
12. Villanustre F, Chala A, Dev R, et al. Modeling and tracking Covid-19 cases using big data analytics on HPCC system platform. J Big Data 2021;8(1):33. [doi: 10.1186/s40537-021-00423-z] [Medline: 33614394]
13. Yue H, Hu T. Geographical detector-based spatial modeling of the COVID-19 mortality rate in the continental United States. Int J Environ Res Public Health 2021 Jun 25;18(13):6832. [doi: 10.3390/ijerph18136832] [Medline: 34202168]
14. Theil H. Economics and Information Theory: North-Holland Publishing Company: Amsterdam, North-Holland; 1967.
15. Manz KM, Mansmann U. Inequality indices to monitor geographic differences in incidence, mortality and fatality rates over time during the COVID-19 pandemic. PLoS ONE 2021;16(5):e0251366. [doi: 10.1371/journal.pone.0251366] [Medline: 33984055]
16. Kullback S, Leibler RA. On information and sufficiency. Ann Math Statist 1951 Mar;22(1):79-86. [doi: 10.1214/aoms/1177729694]
17. Holshue ML, DeBolt C, Lindquist S, et al. First case of 2019 novel coronavirus in the United States. N Engl J Med 2020 Mar 5;382(10):929-936. [doi: 10.1056/NEJMoa2001191] [Medline: 32004427]
18. COVID-19 data repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. Github. URL: https://github.com/CSSEGISandData/COVID-19 [accessed 2024-03-14]
19. National population totals and components of change: 2020-2023. US Census Bureau. URL: https://www.census.gov/data/datasets/time-series/demo/popest/2020s-national-total.html [accessed 2024-03-14]
20. Kulldorff M, Heffernan R, Hartman J, Assunção R, Mostashari F. A space-time permutation scan statistic for disease outbreak detection. PLoS Med 2005 Mar;2(3):e59. [doi: 10.1371/journal.pmed.0020059] [Medline: 15719066]

XSL•FO
RenderX

21. Desjardins MR, Hohl A, Delmelle EM. Rapid surveillance of COVID-19 in the United States using a prospective space-time scan statistic: detecting and evaluating emerging clusters. Appl Geogr 2020 May;118:102202. [doi: 10.1016/j.apgeog.2020.102202] [Medline: 32287518]
22. Huang Y, Li R. The lockdown, mobility, and spatial health disparities in COVID-19 pandemic: a case study of New York City. Cities 2022 Mar;122:103549. [doi: 10.1016/j.cities.2021.103549] [Medline: 35125596]

Abbreviations

CSSE: Center for Systems Science and Engineering
KL: Kullback-Leibler

Edited by A Mavragani; submitted 06.04.24; peer-reviewed by A Inostrosa-Psijas, Z Guo; revised version received 04.11.24; accepted 06.11.24; published 03.01.25.

Please cite as:
Sasaki K, Ikeda Y, Nakano T
Quantifying the Regional Disproportionality of COVID-19 Spread: Modeling Study
JMIR Form Res 2025;9:e59230
URL: https://formative.jmir.org/2025/1/e59230
doi:10.2196/59230

© Kenji Sasaki, Yoichi Ikeda, Takashi Nakano. Originally published in JMIR Formative Research (https://formative.jmir.org), 3.1.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Increasing COVID-19 Testing and Vaccination Uptake in the Take Care Texas Community-Based Randomized Trial: Adaptive Geospatial Analysis

Kehe Zhang[1,2], MS; Jocelyn V Hunyadi[1,2], MPH; Marcia C de Oliveira Otto[3], PhD; Miryoung Lee[4], PhD; Zitong Zhang[1], MS, PhD; Ryan Ramphul[3], PhD; Jose-Miguel Yamal[1], PhD; Ashraf Yaseen[1], PhD; Alanna C Morrison[3], PhD; Shreela Sharma[3,5], PhD; Mohammad Hossein Rahbar[3,6,7], PhD; Xu Zhang[7], PhD; Stephen Linder[8], PhD; Dritana Marko[9], MSc, MD; Rachel White Roy[10], MPH, PhD; Deborah Banerjee[11], PhD; Esmeralda Guajardo[12], MA; Michelle Crum[13], PhD; Belinda Reininger[4], DrPH; Maria E Fernandez[14], PhD; Cici Bauer[1,2], MS, PhD

[1]Department of Biostatistics and Data Science, School of Public Health, The University of Texas Health Science Center at Houston, 1200 Pressler St., RAS-E819, Houston, TX, United States

[10]Harris County Public Health, Houston, TX, United States

[11]Houston Health Department, Houston, TX, United States

[12]Cameron County Public Health, San Benito, TX, United States

[13]Department of Preventive Medicine and Population Health, School of Medicine, The University of Texas at Tyler, Tyler, TX, United States

[14]Department of Health Promotion and Behavior Sciences, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX, United States

[2]Center for Spatial-Temporal Modeling for Applications in Population Sciences, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX, United States

[3]Department of Epidemiology, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX, United States

[4]Department of Epidemiology, School of Public Health, The University of Texas Health Science Center at Houston, Brownsville, TX, United States

[5]Center for Health Equity, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX, United States

[6]Biostatistics/Epidemiology/Research Design Component, Center for Clinical and Translational Sciences, The University of Texas Health Science Center at Houston, Houston, TX, United States

[7]Department of Internal Medicine, Division of Clinical and Translational Sciences, McGovern Medical School, The University of Texas Health Science Center at Houston, Houston, TX, United States

[8]Department of Management, Policy and Community Health, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX, United States

[9]Department of Quantitative and Qualitative Health Sciences, University of Texas School of Public Health San Antonio, San Antonio, TX, United States

Corresponding Author:
Cici Bauer, MS, PhD
Department of Biostatistics and Data Science, School of Public Health, The University of Texas Health Science Center at Houston, 1200 Pressler St., RAS-E819, Houston, TX, United States

Abstract

Background: Geospatial data science can be a powerful tool to aid the design, reach, efficiency, and impact of community-based intervention trials. The project titled *Take Care Texas* aims to develop and test an adaptive, multilevel, community-based intervention to increase COVID-19 testing and vaccination uptake among vulnerable populations in 3 Texas regions: Harris County, Cameron County, and Northeast Texas.

Objective: We aimed to develop a novel procedure for adaptive selections of census block groups (CBGs) to include in the community-based randomized trial for the Take Care Texas project.

Methods: CBG selection was conducted across 3 Texas regions over a 17-month period (May 2021 to October 2022). We developed persistent and recent COVID-19 burden metrics, using real-time SARS-CoV-2 monitoring data to capture dynamic infection patterns. To identify vulnerable populations, we also developed a CBG-level community disparity index, using 12 contextual social determinants of health (SDOH) measures from US census data. In each adaptive round, we determined the priority CBGs based on their COVID-19 burden and disparity index, ensuring geographic separation to minimize intervention "spillover." Community input and feedback from local partners and health workers further refined the selection. The selected CBGs were then randomized into 2 intervention arms—multilevel intervention and just-in-time adaptive intervention—and 1 control arm, using covariate adaptive randomization, at a 1:1:1 ratio. We developed interactive data dashboards, which included

XSL•FO
RenderX

maps displaying the locations of selected CBGs and community-level information, to inform the selection process and guide intervention delivery. Selection and randomization occurred across 10 adaptive rounds.

Results: A total of 120 CBGs were selected and followed the stepped planning and interventions, with 60 in Harris County, 30 in Cameron County, and 30 in Northeast Texas counties. COVID-19 burden presented substantial temporal changes and local variations across CBGs. COVID-19 burden and community disparity exhibited some common geographical patterns but also displayed distinct variations, particularly at different time points throughout this study. This underscores the importance of incorporating both real-time monitoring data and contextual SDOH in the selection process.

Conclusions: The novel procedure integrated real-time monitoring data and geospatial data science to enhance the design and adaptive delivery of a community-based randomized trial. Adaptive selection effectively prioritized the most in-need communities and allowed for a rigorous evaluation of community-based interventions in a multilevel trial. This methodology has broad applicability and can be adapted to other public health intervention and prevention programs, providing a powerful tool for improving population health and addressing health disparities.

(JMIR Form Res 2025;9:e62802) doi:10.2196/62802

KEYWORDS

COVID-19 testing; COVID-19 vaccination; study design; community-based interventions; geospatial analysis; public health; social determinants of health; data dashboard

Introduction

The COVID-19 pandemic created a public health crisis, prompting discussions on how to best mitigate its far-reaching effects in a timely and efficient manner. From the start of the pandemic (March 2020) to May 2021 (the beginning of this study), there were over 30 million cumulative COVID-19 cases [1], over 2 million hospitalizations, and over half a million deaths in the United States alone [2]. During natural and health-related disasters (eg, pandemics and epidemics), vulnerable populations have historically endured a disproportionate level of negative health outcomes [3-7]. Extensive research has revealed similar impacts of the COVID-19 pandemic on vulnerable populations, particularly older adults, racial and ethnic minority groups, individuals who are immunocompromised, those with chronic conditions, and individuals with lower socioeconomic status and education attainment [8-17]. For example, a higher risk of severe COVID-19 symptoms, hospitalization, and mortality is associated with the presence of underlying medical conditions and older age [8,9]. Further, multiple studies have assessed population vulnerability through the Centers for Disease Control and Prevention (CDC) Social Vulnerability Index (SVI). For example, COVID-19 risk increases with the overall SVI score and the scores for specific SVI domains, notably those for the minority status and language domains [10,11,16]. Specifically, Hispanic and Black individuals have experienced higher rates of infection, hospitalization, and mortality when compared to non-Hispanic White individuals [12-14,18]. Racial and ethnic minority groups tend to be more susceptible to COVID-19, partially due to economic and social factors such as limited financial resources; insecure housing; higher housing density; reliance on public transportation; and additional inequities in nutrition, jobs, environments, and access to health care and health-related resources [13,14,19,20]. These findings underscore the importance and complex interplay of social determinants of health (SDOH) with respect to COVID-19 outcomes.

Health disparities research has consistently shown that geography—where individuals live, work, and play—matters, and it can greatly influence mortality, morbidity, life expectancy, and various aspects of health [21]. Geographic disparities in COVID-19 health outcomes have also been well documented. Previous studies have demonstrated county-level disparities in SARS-CoV-2 testing, infection, and fatality rates [22-24]. Research conducted by Das et al [25] reported geographic disparities in COVID-19 incidence across various zip code tabulation areas (ZCTAs) in the greater St. Louis area, Missouri. Das et al [25] also reported that the strength of the associations between the area proportion of the population working in agriculture and COVID-19 incidence varied by different ZCTAs. In a separate study, Bauer et al [26] developed a novel approach to identifying testing disparities by census block groups (CBGs), revealing significant geographical disparities in the context of smaller areas. Further, geographical disparities may not be constant and instead shift over time. For example, from 2019 to 2020, temporal trends in confirmed COVID-19 incidence and mortality differed between the most and least vulnerable counties in the United States [27]. Another study reported significant changes in the spatial-temporal patterns of the health care accessibility, test positivity rate of confirmed COVID-19 cases, and case fatality ratio in Texas [28].

Effective and real-time identification of vulnerable, at-risk populations that are characterized by diverse characteristics or reside within specific geographical regions is essential to address and eliminate disparities in COVID-19 health outcomes. In the context of health care, where resources are often limited and must be efficiently and equitably allocated, questions regarding site or population selection become paramount. Achieving this goal requires (1) the identification of metrics to accurately measure the multifaceted nature of disparities with available datasets and (2) a robust method to summarize all metrics and provide a meaningful output with practical use. The latter has been achieved through various approaches, such as assigning weights to each metric, which are often provided by domain experts, and constructing an index from the weighted sum of all metrics.

XSL•FO
RenderX

This paper presents an innovative protocol for adaptive geospatial selections in a community-based intervention program—the Take Care Texas project—funded by the National Institutes of Health (NIH) Rapid Access to Diagnostics for Underserved Populations (RADx-UP) program. Our protocol leveraged a data-driven approach for the rapid adaptation and deployment of multilevel, just-in-time adaptive interventions to improve COVID-19 testing and vaccination uptake in Texas. In the *Methods* section, we focus on 2 key areas. First, we outline the criteria and metrics for identifying the vulnerable populations via SDOH and geographical measures. Second, we detail the methodology for selecting the local communities based on CBGs, which involved using a phased, multistage approach. Designed for real-time prioritization, our protocol addresses the immediate needs of communities most affected by the changing landscape of the COVID-19 crisis.

Methods

Study Region and Populations

Our Take Care Texas project includes 3 geographically distinct sites within Texas, each with unique demographic characteristics (Multimedia Appendix 1). Southeast Texas, represented by the Houston area within Harris County, has an ethnically diverse population (N=4.73 million), covering both urban and suburban communities. This region serves as a significant residence for various racial and ethnic minority groups and is recognized for having one of the largest metropolitan economies in the United States. The South Texas site, represented by Cameron County (population: N=423,029) in the Rio Grande Valley, is primarily Hispanic, with a considerable proportion of the population (22.6%) living below the federal poverty line, exceeding the national average (11.5%) [29]. The Northeast Texas (NETX) region covers an extensive area, including the Smith, Gregg, Henderson, Van Zandt, Anderson, Wood, and Rains counties. NETX is characterized by a mix of small cities and rural communities, with a lower population density when compared to other sites. Anderson County was excluded from this study due to insufficient SARS-CoV-2 infection data coverage at the time of analysis. The demographic distribution across these regions reflects a wide socioeconomic and cultural spectrum, offering a comprehensive understanding of the diverse population within Texas. There are 2144 CBGs in Harris County, 222 CBGs in Cameron County, and 349 CBGs from the NETX region, according to the 2010 census data.

Overview of Priority CBG Selection Process

The process for identifying priority CBGs (PBGs) in this study is presented in Figure 1. The process began with the identification of geographic areas experiencing significant COVID-19 burden, using SARS-CoV-2 surveillance data and population estimates at the ZCTA level. We then identified CBGs within these areas that exhibited high socioeconomic disparities, using the community disparity index we constructed to achieve the desired spatial granularity. Detailed descriptions for determining disease burden and community disparities are provided in the *Persistent and Recent COVID-19 Burdens* and *CBG-Level Community Disparity Index* sections.

In response to the dynamic nature of the COVID-19 pandemic, our identification and selection process was conducted monthly with a near–real-time data feed from the local public health department for timely and effective intervention planning. Table 1 illustrates the stepped planning and intervention calendar. Between May 2021 and October 2022, a total of 10 rounds of CBG selection were conducted for each of the three study sites. In each round, among the CBGs within the identified high–disease burden ZCTAs, we selected the most vulnerable CBGs based on the community disparity index (ie, CBGs in the top 25% for this metric). We then excluded previously randomized CBGs, their neighbors (ie, CBGs that share a common geographic boundary), and any CBGs with small populations (total population<500) to form the PBG list, which served as the sampling universe. From the PBGs, we sampled geographically nonadjacent sets of CBGs (3 CBGs as a set), ensuring suitable geographic separation to mitigate potential intervention "spillover," since the study goal was to compare 2 distinct intervention strategies to a control arm. We sought feedback from the community health workers and staff in the three regions, to determine if the potential CBGs presented safety, geography, and access concerns, by showing maps of the areas and then discussing the areas and documenting concerns. When there was lack of clarity about an area, we reached out via phone calls and emails to community experts in our advisory groups (health department leaders, city leaders, and local nonprofits serving the areas) to obtain their feedback as well. CBGs that posed challenges in reaching community members (such as CBGs with primarily gated communities or business districts), which would prevent implementing the intervention protocol (eg, door-to-door visits), were excluded from the sampling frame. This collaborative approach ensured that our selection process was informed by both data-driven analysis and community and implementer insights. The final selected CBGs within each set were then assigned to 1 of 2 intervention groups (multilevel intervention and just-in-time adaptive intervention) or a control group through baseline covariate adaptive randomization, balancing disparity index and population size, as well as vaccination coverage when such data were available. These CBGs then underwent a 1-month planning phase, followed by a 2-month intervention period. In each round, this study selected and randomized 6 CBGs (2 sets) for Harris County due to its larger geographic area and population, 3 CBGs for Cameron County, and 3 CBGs for NETX counties.

XSL•FO
RenderX

Figure 1. Workflow of the adaptive geospatial selection of CBGs for intervention in the Take Care Texas study. The selection of CBGs for intervention was conducted each round. CBG: census block group; NETX: Northeast Texas; PBG: priority census block group; ZCTA: zip code tabulation area.

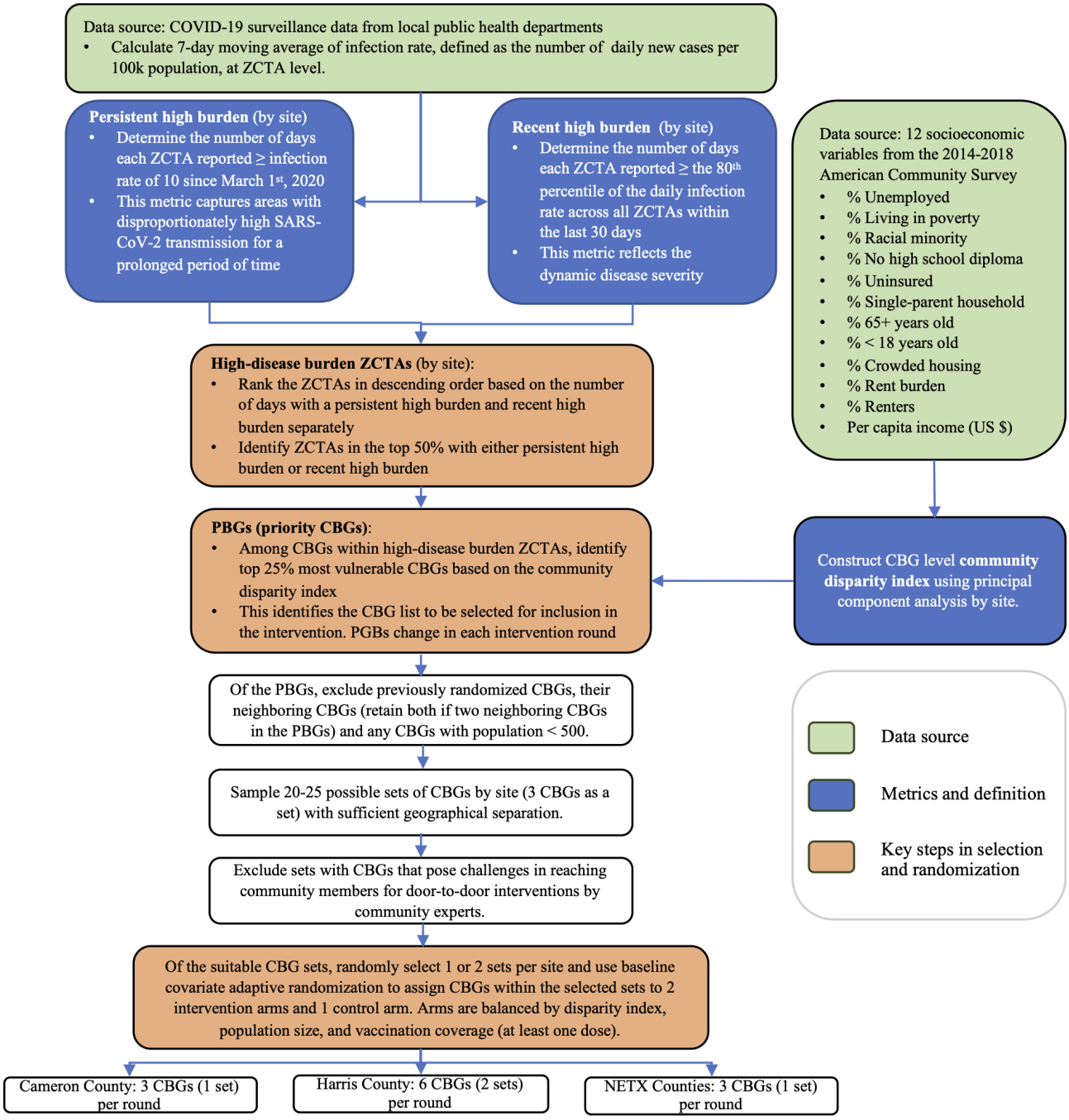

XSL•FO
RenderX

Table . The stepped planning and intervention calendar by month. Each round, selected census block groups were randomized into 1 control arm and 2 intervention arms—multilevel intervention (MILI) and just-in-time adaptive intervention (JITAI).

	2021								2022					
	May	June	July	August	September	October	November	December	January	February	March	April	May	June
Round 1														
Control	✓[a]	✓[b]	✓[b]											
MILI	✓[a]	✓[b]	✓[b]											
JITAI	✓[a]	✓[b]	✓[b]											
Round 2														
Control		✓[a]	✓[b]	✓[b]										
MILI		✓[a]	✓[b]	✓[b]										
JITAI		✓[a]	✓[b]	✓[b]										
Round 3														
Control			✓[a]	✓[b]	✓[b]									
MILI			✓[a]	✓[b]	✓[b]									
JITAI			✓[a]	✓[b]	✓[b]									
Round 4														
Control				✓[a]	✓[b]	✓[b]								
MILI				✓[a]	✓[b]	✓[b]								
JITAI				✓[a]	✓[b]	✓[b]								
Round 5														
Control					✓[a]	✓[b]	✓[b]							
MILI					✓[a]	✓[b]	✓[b]							
JITAI					✓[a]	✓[b]	✓[b]							
Round 6														
Control						✓[a]	✓[b]	✓[b]						
MILI						✓[a]	✓[b]	✓[b]						
JITAI						✓[a]	✓[b]	✓[b]						
Round 7														
Control									✓[a]	✓[b]	✓[b]			
MILI									✓[a]	✓[b]	✓[b]			
JITAI									✓[a]	✓[b]	✓[b]			
Round 8														
Control										✓[a]	✓[b]	✓[b]		
MILI										✓[a]	✓[b]	✓[b]		
JITAI										✓[a]	✓[b]	✓[b]		

XSL•FO
RenderX

	2021								2022					
	May	June	July	August	September	October	November	December	January	February	March	April	May	June
Round 9														
Control											✓[a]	✓[b]	✓[b]	
MILI											✓[a]	✓[b]	✓[b]	
JITAI											✓[a]	✓[b]	✓[b]	
Round 10														
Control												✓[a]	✓[b]	✓[b]
MILI												✓[a]	✓[b]	✓[b]
JITAI												✓[a]	✓[b]	✓[b]

[a]Planning months.

[b]Intervention months.

Persistent and Recent COVID-19 Burdens

The persistent and recent high COVID-19 burdens were quantified by using SARS-CoV-2 surveillance data obtained from local public health departments. We computed a 7-day moving average for infection rate, which was defined as the daily new cases per 100,000 population. For each ZCTA, the persistent disease burden was evaluated based on the number of days on which ≥10 new cases per 100,000 population were reported, as per the common COVID-19 risk level threshold [30], starting from March 1, 2020, up to the respective analysis date for each selection round. To evaluate the recent disease burden, we first calculated the 80th percentile of daily new cases per 100,000 population across all ZCTAs within each study site, using data from the 30 days preceding each analysis date. We then counted the number of days on which the infection rate for each ZCTA exceeded this 80th percentile threshold. In each selection round and for each site, we ranked the ZCTAs, in descending order, based on their number of days with (1) a persistent high burden and (2) a recent high burden. The ranking was performed for each metric separately, and the ZCTAs in the top 50% for either metric were identified as high–disease burden ZCTAs.

CBG-Level Community Disparity Index

We developed the CBG-level community disparity index due to the lack of spatial granularity in existing indices, such as the SVI, which did not meet our study's needs. Using 12 SDOH measures from the 2014 - 2018 American Community Survey 5-year estimates [31], our community disparity index covered the socioeconomic, demographic, and housing dimensions of SDOH (Multimedia Appendix 2). It integrated economic stability factors (employment and poverty rates, per capita income, education level, and health insurance coverage), demographic characteristics (percentages of individuals younger than 18 years, individuals older than 65 years, individuals in single-parent households, and racial minority individuals), and housing factors (percentage of renter-occupied units, rent burden, and crowded housing).

We used principal component analysis (PCA) to construct the community disparity index, similar to the approach used by Kolak et al [32]. PCA is a data-driven approach that synthesizes multiple pieces of information into one combined measure, while simultaneously handling high correlation and minimizing information loss [32-35]. We kept the principal components with eigenvalues exceeding 1, following the Kaiser criteria. The scores of these selected principal components were weighted according to the variance they accounted for and then summed, with higher scores representing a greater community disparity. The resulting indices were then ranked and scaled from 0 to 1 within each study region, providing a relative ranking of CBGs. All data were processed and analyzed with R statistical software (v4.1.2; R Foundation for Statistical Computing) [36].

Interactive Data Dashboards

We developed ArcGIS (Environmental Systems Research Institute Inc) and R Shiny (Posit PBC) data dashboards that provided key geospatial information and relevant COVID-19 data for this study. Feedback was received from the dashboard's end users, including community health workers. The dashboards included maps that showed community disparity indices, PBGs, intervention-prioritized CBGs, COVID-19 testing sites, road maps, and SARS-CoV-2 infection and testing rates at the CBG and ZCTA levels. These tools enabled community health workers to identify the boundaries of the priority areas and access the relevant COVID-19 data for effective on-ground intervention messaging and referral to needed services. Both dashboards were designed to be interactive to facilitate strategic planning efforts and were password protected for internal team use only. The dashboard was especially useful during the planning phase and facilitated the creation of a community profile document, which community health workers used to share real-time data and deliver messaging tailored to the communities.

Ethical Considerations

The research protocol was approved by the UTHealth Committee for the Protection of Human Subjects

XSL•FO
RenderX

(HSC-SPH-20 - 1372). Data user agreements were established between UTHealth and the public health departments of Harris County, Cameron County, and NETX counties. This study did not involve direct contact with participants, as data were aggregated at the CBG level, so no individual-level informed consent was required. Data were deidentified, and all analyses followed data privacy guidelines.

Results

Table 2 presents summary statistics of the population density and 12 SDOH variables used to construct the community disparity index for CBGs in Harris County, Cameron County, and NETX counties, in comparison to the whole state of Texas. Harris County, with the highest population density (2600 per km^2) and per capita income (US $33,500), also has the highest percentage of renters (44.9%). Cameron County, on the other hand, has a higher percentage of crowded housing (11.2%), individuals without high school diplomas (35.1%), uninsured individuals (28.0%), people living in poverty (30.9%), and racial minority individuals (90.0%). NETX has a higher percentage of people aged 65 years and older (18.2%) but a much lower proportion of racial minority individuals compared to Texas overall (35% vs 57.7%).

Between March 2021 and March 2023, we identified 5 distinct COVID-19 waves, with 3 occurring within the intervention period of our study, specifically around August 2021, December 2021, and July 2022 (Multimedia Appendix 3). The overall infection rates varied across the three study regions, highlighting the importance of site-specific burden assessment. The temporal trends of SARS-CoV-2 infection rates differed among the sites when the pandemic started but became more similar during the intervention period, that is, from May 2021 to October 2022. Harris County consistently reported higher infection rates than those reported by the other two regions throughout this study.

Figure 2 displays maps of the CBG-level community disparity index, highlighting substantial variations across and within the study regions. The leading four PCA components captured a significant portion of the variance in the 12 SDOH measures. For instance, in Harris County, these components explained 70.96% of the variance, with the first principal component alone accounting for 43.45%, and were primarily influenced by socioeconomic factors (percentage of individuals with no high school diploma, percentage of individuals living under poverty, per capita income, percentage of uninsured individuals, and percentage of racial minority individuals). Consequently, we derived the standardized index scores by using the first four PCA components.

The maps of the COVID-19 burden and community disparity index shared similar geographic patterns, with high COVID-19 burdens often found in ZCTAs with elevated community disparity (Multimedia Appendix 4). However, distinctive patterns also existed. For instance, during the first selection round for Harris County in May 2021, several ZCTAs located in the western and eastern parts of the county experienced high COVID-19 burden, despite their low community disparity scores. This highlights the complex interplay between the pandemic impact and SDOH, as COVID-19 can affect diverse communities, including those traditionally considered less socially vulnerable. Therefore, it is crucial to leverage both SDOH data and COVID-19 surveillance data to dynamically identify and prioritize areas for interventions.

Over the 17-month period from May 2021 to October 2022, we selected and randomized 120 CBGs for intervention across 10 rounds, with 60 CBGs in Harris County, 30 in Cameron County, and 30 in NETX counties (Table 1). The population size of the randomized CBGs ranged from 615 to 11,321. The combination of the community disparity index and real-time COVID-19 surveillance data allowed for the identification and selection of areas—those with the highest disease burden among the most vulnerable populations—for the adaptive, multilevel, just-in-time intervention. This study also observed substantial changes in the PBG list across different rounds, with more CBGs identified in the later rounds of selection, corresponding to increased infection rates during the pandemic, as shown in Figure 3.

Multimedia Appendix 5 illustrates the R Shiny and ArcGIS dashboards that we developed to provide a dynamic visualization of the data from this study. These interactive dashboards, along with the community profiles we developed, were used by the intervention team to tailor their strategies. By leveraging these tools, the team was able to adapt culturally relevant messages and materials, enhance door-to-door education by community health workers in selected CBGs, and better inform social marketing campaigns to promote COVID-19 testing and vaccination efforts.

XSL•FO
RenderX

Table . Summary statistics of the American Community Survey variables (2014-2018 5-year estimates) [31] used in community disparity index creation at the census block group (CBG) level for 3 regions in Texas and Texas overall.

	Texas regions			Texas overall
	Cameron County (CBGs: n=222)	Harris County (CBGs: n=2144)	NETX[a] counties (CBGs: n=349)	
Population (n)				27,900,000
Mean (CV[b] [%])	1900 (83.7)	2150 (83.5)	1540 (48.6)	—[c]
Median (IQR)	1550 (967-2350)	1770 (1210-2470)	1370 (1060-1890)	—
Population density (per km^2)				43.1
Mean (SD [%])	1420 (86.5)	2600 (97.5)	546 (136.5)	—
Median (minimum, maximum)	1180 (345, 2130)	2020 (1180, 3090)	209 (41.4, 870)	—
Crowded housing (%)				4.79
Mean (CV [%])	11.2 (74.0)	6.28 (123.8)	3.45 (141.4)	—
Median (IQR)	9.84 (4.82-16.3)	3.68 (0-9.46)	1.82 (0-5.24)	—
Individuals without a high school diploma (%)				16.8
Mean (CV [%])	35.1 (45.1)	21.5 (84.1)	16.7 (74.3)	—
Median (IQR)	36.0 (23.2-47.8)	17.0 (5.65-34.5)	14.1 (7.93-21.5)	—
Renters (%)				38.1
Mean (CV [%])	36.0 (56.4)	44.9 (65.1)	32.2 (64.9)	—
Median (IQR)	34.0 (20.9-46.1)	40.2 (20.5-66.5)	27.4 (17.2-43.7)	—
Rent burden (%)				44.4
Mean (CV [%])	46.8 (45.9)	43.7 (52.2)	39.7 (61.6)	—
Median (IQR)	46.4 (33.3-60.9)	45.5 (29.0-58.2)	39.1 (22.2-56.3)	—
Individuals younger than 18 years (%)				26.2
Mean (CV [%])	29.1 (29.6)	25.3 (36.2)	23.7 (36.4)	—
Median (IQR)	30.1 (23.6-35.4)	25.8 (19.5-31.5)	23.0 (18.2-29.0)	—
Individuals older than 65 years (%)				12.0
Mean (CV [%])	15.5 (69.0)	11.3 (66.7)	18.2 (53.7)	—
Median (IQR)	13.2 (9.67-18.4)	9.99 (6.28-14.5)	17.2 (11.0-23.5)	—
Individuals in single-parent households (%)				14.7
Mean (CV [%])	18.3 (62.5)	17.2 (83.3)	14.0 (92.8)	—
Median (IQR)	17.1 (9.94-25.8)	14.4 (6.23-25.3)	10.8 (4.72-20.9)	—
Unemployed individuals (%)				5.41
Mean (CV [%])	7.05 (91.1)	6.45 (91.4)	5.65 (103.5)	—
Median (IQR)	5.85 (2.60-9.76)	5.16 (2.54-8.89)	4.39 (1.47-7.87)	—
Per capita income (US $)				30,100
Mean (CV [%])	16,400 (45.3)	33,500 (78.2)	25,800 (41.8)	—
Median (IQR)	14,500 (10,700-20,200)	24,200 (16,700-40,700)	25,400 (17,600-30,700)	—
Uninsured individuals (%)				17.1
Mean (CV [%])	28.0 (38.4)	20.9 (67.3)	17.6 (54.8)	—
Median (IQR)	28.4 (20.2-35.3)	19.5 (9.25-30.9)	16.0 (10.6-23.4)	—
Individuals living in poverty (%)				14.3
Mean (CV [%])	30.9 (46.9)	16.2 (81.2)	16.6 (71.9)	—

XSL•FO
RenderX

	Texas regions			Texas overall
	Cameron County (CBGs: n=222)	Harris County (CBGs: n=2144)	NETX[a] counties (CBGs: n=349)	
Median (IQR)	31.1 (19.1-42.1)	13.2 (5.63-23.9)	14.5 (7.93-22.4)	—
Racial minority individuals (%)				57.7
Mean (CV [%])	90.0 (15.6)	69.3 (40.0)	35.0 (77.2)	—
Median (IQR)	94.2 (87.3-98.4)	79.7 (46.5-94.0)	26.1 (14.4-50.9)	—

[a]NETX: Northeast Texas.

[b]CV: coefficient of variation.

[c]Not applicable.

Figure 2. Census block group–level community disparity index developed for the community-based intervention by study region. The community disparity index was constructed by using principal component analysis to combine 12 social determinants of health measures from the American Community Survey 2014-2018 5-year estimates [31]. The index was scaled to range from 0 to 1 within each study area, with higher index values indicating greater community disparity.

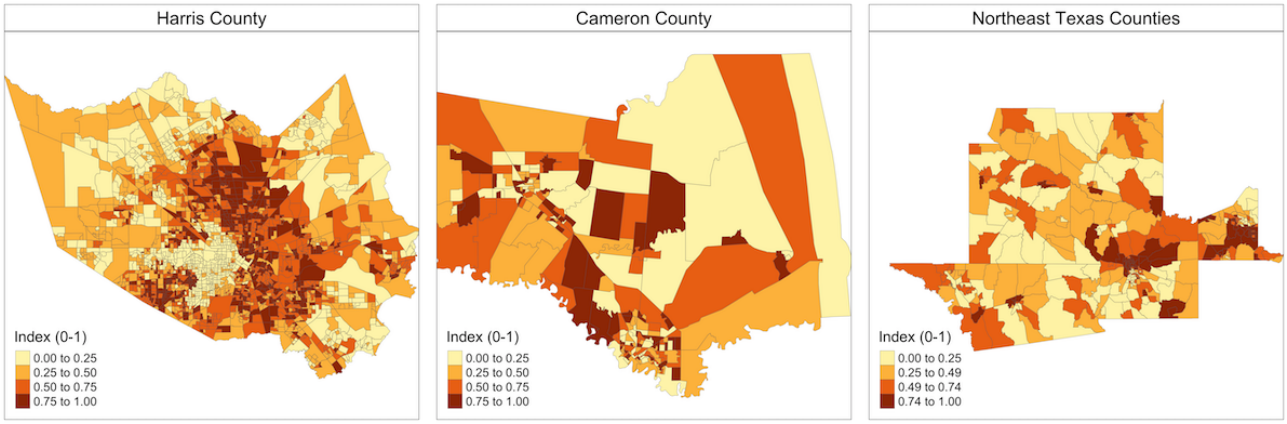

XSL•FO
RenderX

Figure 3. Priority CBGs for the community-based intervention by study region and selection rounds: highlighted CBGs were identified as candidates for inclusion in R1, R2, R4, and R7. CBG: census block group; NETX: Northeast Texas; R: round.

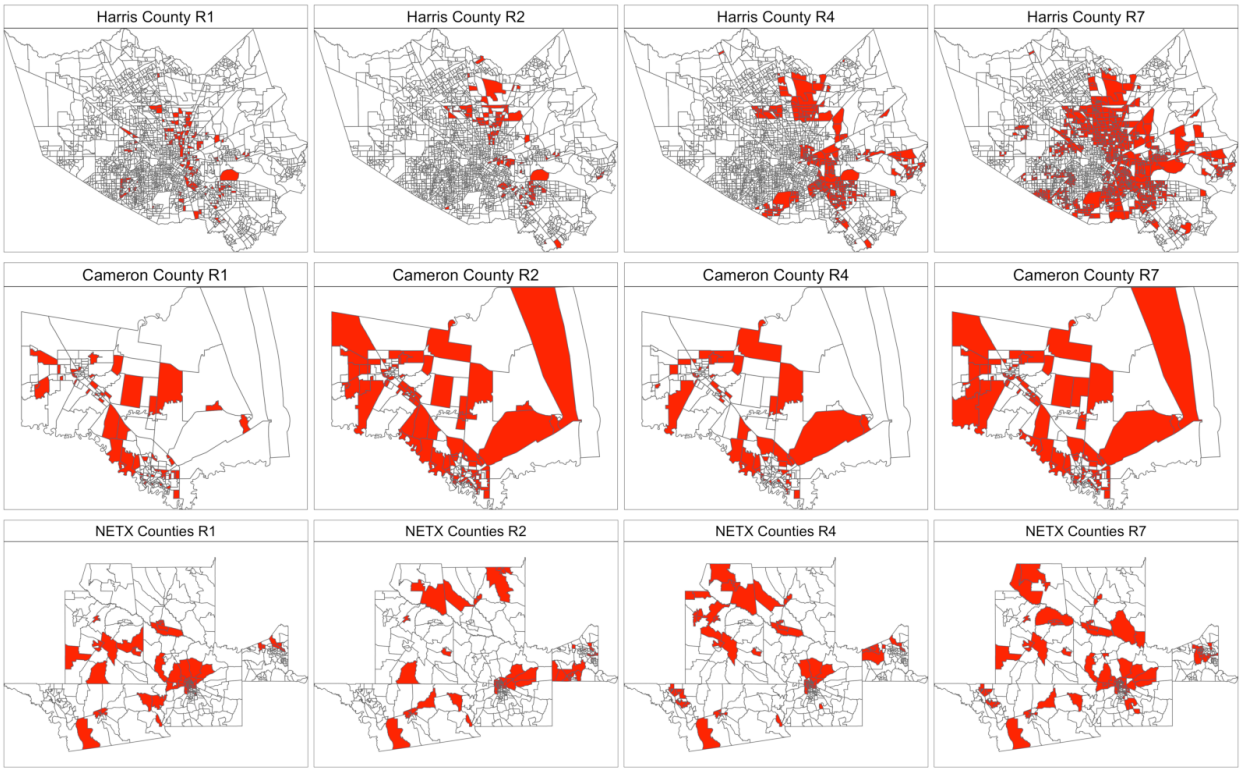

Discussion

Principal Findings

In this study, we used geospatial data science methods to develop a protocol for the adaptive selection of 120 CBGs in 3 regions of Texas, which occurred across 10 rounds over a 17-month period (May 2021 to October 2022). This protocol, which was developed for the Take Care Texas project, integrated real-time dynamic data on SARS-CoV-2 infection rates and contextual SDOH to identify and prioritize the CBGs most in need of tailored activities and resources to improve COVID-19 testing and vaccination. The adaptive selection of CBGs prioritized the most vulnerable communities and allowed for the rigorous evaluation of community-based interventions in a multilevel trial.

Strengths and Limitations

Our protocol demonstrated several unique strengths in integrating public health surveillance data and geospatial data science. First, our study offered spatial granularity at the level of CBGs, which has rarely been seen in other studies [11,16,37]. Such granularity could potentially result in more tailored interventions that are best suited to the local communities for which they were developed. Second, we incorporated both the reported SARS-CoV-2 infection rates and the community SDOH, reflecting the socioeconomic and housing conditions that contributed to health disparities. This allowed us to prioritize communities experiencing not only high disease burden but also high social vulnerability that could potentially impact community resilience and the ability to adapt to challenges presented by the pandemic. Third, the dynamic nature of our protocol, particularly the initial selection of high–disease burden ZCTAs, allowed for real-time adaptation of the selection strategies in response to the changing pandemic, thereby enabling us to prioritize the most in-need and most vulnerable communities effectively. Finally, the protocol was strengthened by the input from and collaboration among community members and frontline community health workers, which were critical in designing and implementing interventions. The final CBG selection was made by community partners considering on-the-ground intervention feasibility. In the context of this study and site selection, the input from our community partners and community health workers supplemented the data-driven approaches [38]. Although multicriteria decision-making methods have been previously used for site selection problems [39-42], including the integration of objective criteria with experts' qualitative input to determine site suitability, some multicriteria decision-making techniques have been critiqued for arbitrary and irrational rankings of alternatives [43] and questionable mathematical validity [44]. By contrast, previous

XSL•FO
RenderX

work has shown that PCA is a robust approach for integrating various vulnerability and health indicators and providing reliable unequal weighting [33,34] to help guide the selection of local communities for health care interventions. Future studies should continue to investigate best practices for engaging communities in trial design and implementation.

Our protocol also presented several limitations. First, the COVID-19 pandemic highlighted significant challenges in data reporting due to the absence of established surveillance data infrastructures for COVID-19 and other rapidly spreading communicable diseases. The state reporting requirements for data elements and definitions changed over time, complicating the consistent use of data across study regions. In our study, we directly collaborated with individual county public health departments and processed and harmonized data across different study regions, which were time-consuming processes. Second, the time lag commonly seen in surveillance databases posed challenges in accurately identifying current hot spots, and incomplete demographic data hindered the assessment of health equity throughout the COVID-19 pandemic by delaying the analysis of COVID-19 health outcomes by subgroups and across more granular geographic scales [45,46]. The Congressional Research Service's 2020 report highlighted the need to modernize public health reporting systems to improve the timeliness and accuracy of public health data reporting [45]. However, differences in data standards, technical capacity, and regional policy priorities, particularly in the early stages of the COVID-19 pandemic, limited effective data sharing across study regions. Third, our protocol required accurate and complete address information for geographical granularity in identifying the high–disease burden priority areas, and when address information was missing, we may not have been able to include the SARS-CoV-2 case records in determining the PBGs. Finally, finding geographically separated CBGs in later rounds of selection could be difficult in small counties with fewer CBGs, such as Cameron County. Traditionally, administrative boundaries are used to define spatial units, yet alternative approaches may better capture actual community layouts. For example, Tuson et al [47] developed the Spatial Targeting Algorithm to generate flexible priority regions not restricted by administrative boundaries. However, while flexible boundaries may better capture neighborhoods, they may not provide SDOH information, which is typically accessible only at the administrative level. Additional work is needed to develop more flexible spatial units, with or without administrative boundaries as a constraint, for population-based research.

Study Implications

Despite these challenges, our study provided valuable protocols for data processing and analytical support to an intervention team working with local public health departments and organizations. It demonstrated a useful approach to ensuring that the most vulnerable communities participated in a multilevel intervention trial for increasing COVID-19 testing and vaccination uptake. Our proposed protocol could be adapted to other population-based randomized trials for which rapid assessment and allocation of finite resources are imperative. For example, geospatial approaches have been used to assess cancer screening rates among a vulnerable population in Ontario, Canada [48] and determine priority clinics to allocate funding for improving up-to-date colorectal cancer screening rates [49]. The integration of PCA and the real-time nature of our protocol can also be adapted to other population-based trials. When tackling multifaceted and complex problems, our protocol demonstrates that PCA is effective at robustly incorporating a wide range of health indicators. Further, the real-time nature of our protocol allows for ongoing, continuous assessment and adjustment in allocation of resources, provided that accurate and timely data are available. This could be applied to study recruitment protocols or other intervention-based trials.

Conclusion

The selection of optimal sites for health care and public health interventions, especially under limited resources, is a crucial challenge. Our protocol contributes to existing research by integrating disease surveillance and SDOH data, the robust PCA method, and community engagement to identify locations where COVID-19 testing interventions for vulnerable populations are most needed. The proposed data-driven approach could be adapted to other disease control and prevention programs to improve population health and reduce disease burden in areas with disparities.

Acknowledgments

This work was supported by the National Institutes of Health (NIH) Rapid Access to Diagnostics for Underserved Populations (RADx-UP) initiative (NIH 3UL1TR003167-02S1, UL1TR003167-02S6, and U01TR004355). This project is partially supported by UTHealth Houston Center for Clinical and Translational Sciences/Clinical and Translational Awards. We acknowledge the important contributions of the community health workers and staff of Take Care Texas in providing expertise, valuable feedback, and guidance during the revision. These staff include Ibette Cavazos, Lisa Mitchell-Bennett, Maribel Sifuentes, Carol Killingley, Carlton Allen, and LuBeth Perez. We also thank all the officials of the counties for providing their support to the Take Care Texas project.

Authors' Contributions

MCOO, ML, ACM, SS, MC, BR, MEF, and CB conceptualized the study design. KZ, MCOO, ML, ZZ, RR, JMY, AY, SL, DM, RWR, DB, EG, MC, and CB collected and processed the data. KZ, JVH, MCOO, ZZ, and CB analyzed and visualized the data. KZ, JVH, and CB wrote the initial draft of the manuscript. KZ, JVH, MCOO, ML, JMY, ACM, SS, MHR, XZ, DM, MC, BR, MEF, and CB edited and revised the manuscript. ML, MC, BR, MEF, and CB supervised the study and acquired funding.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Map of Texas counties with 3 study regions highlighted. Harris County (yellow) has the most ethnically diverse population in Texas. Cameron County (blue), located on the southern border of Texas, is primarily Hispanic. A large proportion of the population in Northeast Texas (NETX; orange) is White.
[PNG File, 494 KB - formative_v9i1e62802_app1.png]

Multimedia Appendix 2
Metrics for identifying vulnerable populations, definitions, data sources, and available spatial scale.
[DOCX File, 24 KB - formative_v9i1e62802_app2.docx]

Multimedia Appendix 3
Temporal trends of weekly COVID-19 cases per 100,000 population in Harris County, Cameron County, and Northeast Texas (NETX) counties, with vertical lines indicating the start and end of the planning and intervention periods for the Take Care Texas project. Historical Texas COVID-19 data were obtained from the Texas Department of State Health Services.
[PNG File, 350 KB - formative_v9i1e62802_app3.png]

Multimedia Appendix 4
Maps of the census block group–level community disparity index developed in this protocol, overlaid with high–disease burden zip code tabulation areas (shown in red geographical boundaries) in the first round of selection (May 2021) for the three study regions.
[PNG File, 1603 KB - formative_v9i1e62802_app4.png]

Multimedia Appendix 5
Screenshots of the (A) R shiny dashboard and (B) ArcGIS dashboard developed for the Take Care Texas study. These data dashboards were designed to be interactive to best facilitate strategic planning efforts and were password protected for internal team use only.
[PDF File, 233 KB - formative_v9i1e62802_app5.pdf]

References

1. WHO COVID-19 dashboard. World Health Organization. URL: https://covid19.who.int [accessed 2023-11-29]
2. COVID data tracker. Centers for Disease Control and Prevention. URL: https://covid.cdc.gov/covid-data-tracker [accessed 2023-11-29]
3. Paramanathan P, Abbas M, Huda SA, Huda S, Mortazavi M, Taravati P. Comparing racial health disparities in pandemics a decade apart: H1N1 and COVID-19. Future Healthc J 2021 Nov;8(3):e722-e728. [doi: 10.7861/fhj.2021.0030] [Medline: 34888475]
4. Quinn SC, Kumar S, Freimuth VS, Musa D, Casteneda-Angarita N, Kidwell K. Racial disparities in exposure, susceptibility, and access to health care in the US H1N1 influenza pandemic. Am J Public Health 2011 Feb;101(2):285-293. [doi: 10.2105/AJPH.2009.188029] [Medline: 21164098]
5. D'Adamo A, Schnake-Mahl A, Mullachery PH, Lazo M, Diez Roux AV, Bilal U. Health disparities in past influenza pandemics: a scoping review of the literature. SSM Popul Health 2023 Mar;21:101314. [doi: 10.1016/j.ssmph.2022.101314] [Medline: 36514788]
6. Howell J, Elliott JR. Damages done: the longitudinal impacts of natural hazards on wealth inequality in the United States. Soc Probl 2019 Aug 14;66(3):448-467. [doi: 10.1093/socpro/spy016]
7. Climate change and social vulnerability in the United States: a focus on six impacts. United States Environmental Protection Agency. 2021. URL: www.epa.gov/cira/social-vulnerability-report [accessed 2023-11-28]
8. Sanyaolu A, Okorie C, Marinkovic A, et al. Comorbidity and its impact on patients with COVID-19. SN Compr Clin Med 2020;2(8):1069-1076. [doi: 10.1007/s42399-020-00363-4] [Medline: 32838147]
9. Souch JM, Cossman JS. A commentary on rural-urban disparities in COVID-19 testing rates per 100,000 and risk factors. J Rural Health 2021 Jan;37(1):188-190. [doi: 10.1111/jrh.12450] [Medline: 32282964]
10. Karaye IM, Horney JA. The impact of social vulnerability on COVID-19 in the U.S.: an analysis of spatially varying relationships. Am J Prev Med 2020 Sep;59(3):317-325. [doi: 10.1016/j.amepre.2020.06.006] [Medline: 32703701]
11. Khazanchi R, Beiter ER, Gondi S, Beckman AL, Bilinski A, Ganguli I. County-level association of social vulnerability with COVID-19 cases and deaths in the USA. J Gen Intern Med 2020 Sep;35(9):2784-2787. [doi: 10.1007/s11606-020-05882-3] [Medline: 32578018]

XSL•FO
RenderX

12. Millett GA, Jones AT, Benkeser D, et al. Assessing differential impacts of COVID-19 on black communities. Ann Epidemiol 2020 Jul;47:37-44. [doi: 10.1016/j.annepidem.2020.05.003] [Medline: 32419766]
13. Aleligne YK, Appiah D, Ebong IA. Racial disparities in coronavirus disease 2019 (COVID-19) outcomes. Curr Opin Cardiol 2021 May 1;36(3):360-366. [doi: 10.1097/HCO.0000000000000847] [Medline: 33657019]
14. Mackey K, Ayers CK, Kondo KK, et al. Racial and ethnic disparities in COVID-19-related infections, hospitalizations, and deaths : a systematic review. Ann Intern Med 2021 Mar;174(3):362-373. [doi: 10.7326/M20-6306] [Medline: 33253040]
15. Upshaw TL, Brown C, Smith R, Perri M, Ziegler C, Pinto AD. Social determinants of COVID-19 incidence and outcomes: a rapid review. PLoS One 2021 Mar 31;16(3):e0248336. [doi: 10.1371/journal.pone.0248336] [Medline: 33788848]
16. Biggs EN, Maloney PM, Rung AL, Peters ES, Robinson WT. The relationship between social vulnerability and COVID-19 incidence among Louisiana census tracts. Front Public Health 2021 Jan 20;8:617976. [doi: 10.3389/fpubh.2020.617976] [Medline: 33553098]
17. Bauer C, Zhang K, Lee M, et al. Census tract patterns and contextual social determinants of health associated with COVID-19 in a Hispanic population from South Texas: a spatiotemporal perspective. JMIR Public Health Surveill 2021 Aug 5;7(8):e29205. [doi: 10.2196/29205] [Medline: 34081608]
18. Tai DBG, Shah A, Doubeni CA, Sia IG, Wieland ML. The disproportionate impact of COVID-19 on racial and ethnic minorities in the United States. Clin Infect Dis 2021 Feb 16;72(4):703-706. [doi: 10.1093/cid/ciaa815] [Medline: 32562416]
19. Caraballo C, Ndumele CD, Roy B, et al. Trends in racial and ethnic disparities in barriers to timely medical care among adults in the US, 1999 to 2018. JAMA Health Forum 2022 Oct 7;3(10):e223856. [doi: 10.1001/jamahealthforum.2022.3856] [Medline: 36306118]
20. Radley DC, Baumgartner JC, Collins SR, Zephyrin LC, Schneider EC. Achieving racial and ethnic equity in U.S. health care: a scorecard of state performance. The Commonwealth Fund. 2021 Nov 18. URL: https://www.commonwealthfund.org/publications/scorecard/2021/nov/achieving-racial-ethnic-equity-us-health-care-state-performance [accessed 2023-11-29]
21. National Academies of Sciences, Engineering, and Medicine, Health and Medicine Division, Board on Population Health and Public Health Practice, Committee on Community-Based Solutions to Promote Health Equity in the United States. The state of health disparities in the United States. In: Baciu A, Negussie Y, Geller A, Weinstein JN, editors. Communities in Action: Pathways to Health Equity: National Academies Press; 2017. URL: https://www.ncbi.nlm.nih.gov/books/NBK425844 [accessed 2023-12-07] [doi: 10.17226/24624] [Medline: 28418632]
22. Deb Nath N, Khan MM, Schmidt M, Njau G, Odoi A. Geographic disparities and temporal changes of COVID-19 incidence risks in North Dakota, United States. BMC Public Health 2023 Apr 20;23(1):720. [doi: 10.1186/s12889-023-15571-5] [Medline: 37081453]
23. Khan MM, Odoi A, Odoi EW. Geographic disparities in COVID-19 testing and outcomes in Florida. BMC Public Health 2023 Jan 11;23(1):79. [doi: 10.1186/s12889-022-14450-9] [Medline: 36631768]
24. Jackson SL, Derakhshan S, Blackwood L, et al. Spatial disparities of COVID-19 cases and fatalities in United States counties. Int J Environ Res Public Health 2021 Aug 4;18(16):8259. [doi: 10.3390/ijerph18168259] [Medline: 34444007]
25. Das P, Igoe M, Lenhart S, et al. Geographic disparities and determinants of COVID-19 incidence risk in the greater St. Louis Area, Missouri (United States). PLoS One 2022 Sep 28;17(9):e0274899. [doi: 10.1371/journal.pone.0274899] [Medline: 36170339]
26. Bauer C, Li X, Zhang K, et al. A novel Bayesian spatial-temporal approach to quantify SARS-CoV-2 testing disparities for small area estimation. Am J Public Health 2023 Jan;113(1):40-48. [doi: 10.2105/AJPH.2022.307127] [Medline: 36516388]
27. Neelon B, Mutiso F, Mueller NT, Pearce JL, Benjamin-Neelon SE. Spatial and temporal trends in social vulnerability and COVID-19 incidence and death rates in the United States. PLoS One 2021 Mar 24;16(3):e0248702. [doi: 10.1371/journal.pone.0248702] [Medline: 33760849]
28. Park J, Michels A, Lyu F, Han SY, Wang S. Daily changes in spatial accessibility to ICU beds and their relationship with the case-fatality ratio of COVID-19 in the state of Texas, USA. Appl Geogr 2023 May;154:102929. [doi: 10.1016/j.apgeog.2023.102929] [Medline: 36960405]
29. U.S. Census Bureau QuickFacts: Cameron County, Texas. United States Census Bureau. 2023. URL: https://www.census.gov/quickfacts/cameroncountytexas [accessed 2023-06-09]
30. Harvard Global Health Institute, Edmond J. Safra Center for Ethics. Key metrics for COVID suppression: a framework for policy makers and the public. Global Epidemics. 2020 Jul 1. URL: https://globalepidemics.org/wp-content/uploads/2020/09/key_metrics_and_indicators_v5-1.pdf [accessed 2025-01-28]
31. 2014-2018 ACS 5-year estimates. United States Census Bureau. URL: https://www.census.gov/programs-surveys/acs/technical-documentation/table-and-geography-changes/2018/5-year.html [accessed 2023-06-09]
32. Kolak M, Bhatt J, Park YH, Padrón NA, Molefe A. Quantification of neighborhood-level social determinants of health in the continental United States. JAMA Netw Open 2020 Jan 3;3(1):e1919928. [doi: 10.1001/jamanetworkopen.2019.19928] [Medline: 31995211]
33. Niu Y, Li Z, Gao Y, et al. A systematic review of the development and validation of the heat vulnerability index: major factors, methods, and spatial units. Curr Clim Change Rep 2021;7(3):87-97. [doi: 10.1007/s40641-021-00173-3] [Medline: 34745843]

XSL•FO
RenderX

34. Pathak S, Wang H, Seals K, Adusumilli NC, Holston D. Self-assessed health status and obesity vulnerability in rural Louisiana: a cross-sectional analysis. PLoS One 2023 Jun 16;18(6):e0287181. [doi: 10.1371/journal.pone.0287181] [Medline: 37327219]
35. Lieberman-Cribbin W, Tuminello S, Flores RM, Taioli E. Disparities in COVID-19 testing and positivity in New York City. Am J Prev Med 2020 Sep;59(3):326-332. [doi: 10.1016/j.amepre.2020.06.005] [Medline: 32703702]
36. R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing. 2022. URL: https://www.R-project.org/ [accessed 2025-01-28]
37. Desjardins MR, Hohl A, Delmelle EM. Rapid surveillance of COVID-19 in the United States using a prospective space-time scan statistic: detecting and evaluating emerging clusters. Appl Geogr 2020 May;118:102202. [doi: 10.1016/j.apgeog.2020.102202] [Medline: 32287518]
38. Cromley E, Kleinman LC, Ramos MA, et al. A community-engaged approach to select geographic areas for interventions to reduce health disparities. Prog Community Health Partnersh 2011;5(3):299-305. [doi: 10.1353/cpr.2011.0033] [Medline: 22080778]
39. Mardani A, Hooker RE, Ozkul S, et al. Application of decision making and fuzzy sets theory to evaluate the healthcare and medical problems: a review of three decades of research with recent developments. Expert Syst Appl 2019 Dec;137:202-231. [doi: 10.1016/j.eswa.2019.07.002]
40. Boyacı A, Şişman A. Pandemic hospital site selection: a GIS-based MCDM approach employing Pythagorean fuzzy sets. Environ Sci Pollut Res Int 2022 Jan;29(2):1985-1997. [doi: 10.1007/s11356-021-15703-7] [Medline: 34357491]
41. Tripathi AK, Agrawal S, Gupta RD. Comparison of GIS-based AHP and fuzzy AHP methods for hospital site selection: a case study for Prayagraj City, India. GeoJournal 2022;87(5):3507-3528. [doi: 10.1007/s10708-021-10445-y] [Medline: 34075269]
42. Sotoudeh-Anvari A. The applications of MCDM methods in COVID-19 pandemic: a state of the art review. Appl Soft Comput 2022 Sep;126:109238. [doi: 10.1016/j.asoc.2022.109238] [Medline: 35795407]
43. Asadabadi MR, Chang E, Saberi M. Are MCDM methods useful? A critical review of Analytic Hierarchy Process (AHP) and Analytic Network Process (ANP). Cogent Engineering 2019 May 31;6(1). [doi: 10.1080/23311916.2019.1623153]
44. Zhü K. Fuzzy analytic hierarchy process: fallacy of the popular methods. Eur J Oper Res 2014 Jul;236(1):209-217. [doi: 10.1016/j.ejor.2013.10.034]
45. Sekar K, Napili A. Tracking COVID-19: U.S. public health surveillance and data. : Congressional Research Service; 2020 R46588 URL: https://crsreports.congress.gov/product/pdf/R/R46588 [accessed 2025-01-28]
46. Krieger N, Gonsalves G, Bassett MT, Hanage W, Krumholz HM. The fierce urgency of now: closing glaring gaps in US surveillance data on COVID-19. Health Affairs. 2020 Apr 14. URL: https://www.healthaffairs.org/do/10.1377/forefront.20200414.238084/full/ [accessed 2024-02-02]
47. Tuson M, Yap M, Kok MR, et al. Improving the efficiency of geographic target regions for healthcare interventions. Int J Environ Res Public Health 2022 Nov 9;19(22):14721. [doi: 10.3390/ijerph192214721] [Medline: 36429437]
48. Lofters AK, Gozdyra P, Lobb R. Using geographic methods to inform cancer screening interventions for South Asians in Ontario, Canada. BMC Public Health 2013 Apr 26;13:395. [doi: 10.1186/1471-2458-13-395] [Medline: 23622426]
49. Zhan FB, Liu Y, Yang M, et al. Using GIS to identify priority sites for colorectal cancer screening programs in Texas health centers. Prev Chronic Dis 2023 Mar 2;20:E10 [FREE Full text] [doi: 10.5888/pcd20.220205] [Medline: 36862605]

Abbreviations

CBG: census block group
CDC: Centers for Disease Control and Prevention
NETX: Northeast Texas
NIH: National Institutes of Health
PBG: priority census block group
PCA: principal component analysis
RADx-UP: Rapid Access to Diagnostics for Underserved Populations
SDOH: social determinants of health
SVI: Social Vulnerability Index
ZCTA: zip code tabulation area

XSL•FO
RenderX

Edited by A Mavragani; submitted 31.05.24; peer-reviewed by JA Yaun, L Zhu; revised version received 24.10.24; accepted 01.11.24; published 11.02.25.

Please cite as:
Zhang K, Hunyadi JV, de Oliveira Otto MC, Lee M, Zhang Z, Ramphul R, Yamal JM, Yaseen A, Morrison AC, Sharma S, Rahbar MH, Zhang X, Linder S, Marko D, Roy RW, Banerjee D, Guajardo E, Crum M, Reininger B, Fernandez ME, Bauer C
Increasing COVID-19 Testing and Vaccination Uptake in the Take Care Texas Community-Based Randomized Trial: Adaptive Geospatial Analysis
JMIR Form Res 2025;9:e62802
URL: https://formative.jmir.org/2025/1/e62802
doi:10.2196/62802

© Kehe Zhang, Jocelyn V Hunyadi, Marcia C de Oliveira Otto, Miryoung Lee, Zitong Zhang, Ryan Ramphul, Jose-Miguel Yamal, Ashraf Yaseen, Alanna C Morrison, Shreela Sharma, Mohammad Hossein Rahbar, Xu Zhang, Stephen Linder, Dritana Marko, Rachel White Roy, Deborah Banerjee, Esmeralda Guajardo, Michelle Crum, Belinda Reininger, Maria E Fernandez, Cici Bauer. Originally published in JMIR Formative Research (https://formative.jmir.org), 11.2.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Investigating Reddit Data on Type 2 Diabetes Management During the COVID-19 Pandemic Using Latent Dirichlet Allocation Topic Modeling and Valence Aware Dictionary for Sentiment Reasoning Analysis: Content Analysis

Meghan Nagpal[1,2], MSc, PEng; Niloofar Jalali[3], PhD; Diana Sherifali[4], PhD, RN, CDE; Plinio Morita[1,2,3], PhD, PEng; Joseph A Cafazzo[1,2,5,6], PhD, PEng

[1]Institute of Health Policy, Management, & Evaluation, Dalla Lana School of Public Health, University of Toronto, Health Sciences Building, 4th Floor, Toronto, ON, Canada

[2]Centre for Digital Therapeutics, Techna Institute, University Health Network, Toronto, ON, Canada

[3]School of Public Health & Health Systems, University of Waterloo, Kitchener, ON, Canada

[4]School of Nursing, McMaster University, Hamilton, ON, Canada

[5]Department of Computer Science, University of Toronto, Toronto, ON, Canada

[6]Institute of Biomedical Engineering, Faculty of Applied Science & Engineering, University of Toronto, Toronto, ON, Canada

Corresponding Author:
Meghan Nagpal, MSc, PEng
Institute of Health Policy, Management, & Evaluation, Dalla Lana School of Public Health, University of Toronto, Health Sciences Building, 4th Floor, Toronto, ON, Canada

Abstract

Background: Type 2 diabetes (T2D) is a chronic disease that can be partially managed through healthy behaviors. However, the COVID-19 pandemic impacted how people managed T2D due to work and school closures and social isolation. Moreover, individuals with T2D were at increased risk of complications from COVID-19 and experienced worsened mental health due to stress and anxiety.

Objective: This study aims to synthesize emerging themes related to the health behaviors of people living with T2D, and how they were affected during the early stages of the COVID-19 pandemic by examining Reddit forums dedicated to people living with T2D.

Methods: Data from Reddit forums related to T2D, from January 2018 to early March 2021, were downloaded using the Pushshift API; support vector machines were used to classify whether a post was made in the context of the pandemic. Latent Dirichlet allocation topic modelling was performed to identify topics of discussion across the entire dataset and a subsequent iteration was performed to identify topics specific to the COVID-19 pandemic. Sentiment analysis using the VADER (Valence Aware Dictionary for Sentiment Reasoning) algorithm was performed to assess attitudes towards the pandemic.

Results: From all posts, the identified topics of discussion were classified into the following themes: managing lifestyle (sentiment score 0.25, 95% CI 0.25-0.26), managing blood glucose (sentiment score 0.19, 95% CI 0.18-0.19), obtaining diabetes care (sentiment score 0.19, 95% CI 0.18-0.20), and coping and receiving support (sentiment score 0.34, 95% CI 0.33-0.35). Among the COVID-19–specific posts, the topics of discussion were coping with poor mental health (sentiment score 0.04, 95% CI −0.01 to0.11), accessing doctor and medications and controlling blood glucose (sentiment score 0.14, 95% CI 0.09-0.20), changing food habits during the pandemic (sentiment score 0.25, 95% CI 0.20-0.31), impact of stress on blood glucose levels (sentiment score 0.03, 95% CI −0.03 to 0.08), changing status of employment and insurance (sentiment score 0.17, 95% CI 0.13-0.22), and risk of COVID-19 complications (sentiment score 0.09, 95% CI 0.03-0.14). Overall, posts classified as COVID-19–related (0.12, 95% CI 0.01-0.15) were associated with a lower sentiment score than those classified as nonCOVID (0.25, 95% CI 0.24-0.25). This study was limited due to the lack of a method for assessing the demographics of users and verifying whether users had T2D.

Conclusions: Themes identified from Reddit data suggested that the COVID-19 pandemic significantly influenced how people with T2D managed their disease, particularly in terms of accessing care and dealing with the complications of the virus. Overall, the early stages of the pandemic negatively impacted the attitudes of people living with T2D. This study demonstrates that social media data can be a qualitative data source for understanding patient perspectives.

(JMIR Form Res 2025;9:e51154) doi:10.2196/51154

XSL•FO
RenderX

KEYWORDS

diabetes; diabetes mellitus; DM; COVID-19; pandemics; social media; health behavior; health knowledge; attitudes; practice; self-management; patient-generated health data; perspective; T2DM

Introduction

Background

Diabetes is a serious metabolic condition in which the body experiences elevated blood glucose levels, which can result in serious complications such as cardiovascular disease, kidney disease, stroke, eye disease, foot ulcers, nerve damage, and amputation. The World Health Organization [1] states that high blood glucose levels are the third leading cause of premature mortality. Type 2 diabetes (T2D) is characterized by insulin resistance or insufficient production of insulin. Research suggests that the risks of complications in people living with T2D can be mitigated through self-management, including glucose levels monitoring, optimized nutrition, regular physical activity, and taking prescribed medications [2,3].

Managing T2D During the COVID-19 Pandemic

With the emergence of the COVID-19 pandemic in March 2020, nearly the entire global population was affected by social distancing measures that included business closures, remote schooling and work, prohibition of large crowds, limited socialization, and virtual outpatient diabetes care delivery. Beyond the fear and anxiety that manifested [4], individuals with T2D were at an increased risk of complications from COVID-19, as the virus thrives in an environment of high blood glucose [5]. T2D was considered among the most prevalent chronic conditions where patients were at serious risk of hospitalization or death if they contracted the virus [6]. This was primarily due to increased inflammation from obesity and insulin resistance, and prevalence of other comorbidities such as hypertension, cardiovascular disease, dyslipidemia, and older age [7]. Because T2D is mostly managed by lifestyle behaviors, COVID-19 restriction measures reduced physical activity and increased other unhealthy behaviors among those living with T2D and other chronic conditions [4,8]. Moreover, there was a concern about limited access to health care for people with T2D, as telemedicine replaced many in-person visits [9]. Overall, people with T2D exhibited worsening glucose levels during lockdown periods [10].

Rationale and Study Objectives

Given that the pandemic impacted the lifestyle behaviors of individuals with T2D, it is reasonable to assume that there were changes in how they managed their condition, which contributed to added stress and anxiety. Many people with T2D use social medial forums to discuss how they manage their condition through sharing information on diet, symptoms, research findings, and recipes, while seeking peer support [11]. Social media also serves as a public data source to gauge sentiment and topics of discussion during the initial lockdown period. The objectives of this study were to synthesize emerging themes from peer discussion on Reddit, a social media forum with dedicated communities for individuals coping with T2D. Using data harnessed through the Pushshift application programming interface (API), this study aimed to understand how their attitudes and diabetes management evolved during the early stages of the pandemic. Reddit was chosen over other social networking platforms due to the segregation of special interest communities, such as those focused on diabetes, and the availability of the Pushshift API as an open-source tool for data scraping.

Methods

Overview

Support vector machines were used to classify Reddit posts related to the COVID-19 pandemic. Latent Dirichlet allocation (LDA) topic modeling [12] and sentiment analysis using the VADER (Valence Aware Dictionary for Sentiment Reasoning) algorithm [13] were performed to achieve the study objectives. Algorithms were executed through Python scripts in the Jupyter Notebook platform version 6 (Linux Foundation), with final extractions outputted as CSV files and further analyzed through Microsoft Excel. No predefined protocol was used for this experiment, given the novelty of social media analysis at the time of the study.

Data Source and Collection

For this study, 3 communities on Reddit were examined: r/type2diabetes, r/diabetes_t2, and r/diabetes [14-16]. From the r/diabetes community, only posts tagged with the “flair” and “type 2 diabetes” were included. Raw data of 100,887 posts between January 1, 2018, to March 5, 2021, from these 3 communities were collected from Reddit in March 2021 via the Pushshift application programming interface (API) using Python scripts in Jupyter Notebook [17]. Data was exported as CSV files and examined in Microsoft Excel. Posts with less than 5 words, posts with the label “removed” and those containing special characters were removed from the dataset using Microsoft Excel filtering tools to avoid bias in the algorithm.

Classification of Posts

Manual examination of the dataset found that the first post about the COVID-19 pandemic was written on February 28, 2020, and coincidentally, the last post about the pandemic in the dataset was made on February 28, 2021. In total, 48,988 posts were published during this period. Within the dataset, posts were manually searched for terms related to the COVID-19 pandemic . These terms included “COVID”, “coronavirus”, “pandemic”, “social distancing”, “lockdown”, “quarantine”, “toilet paper”, “unemploy” and “unemployed”, “work and working from home”, “telehealth”, “vaccine”, “sanitizer”, and “mask.”

Posts that contained these terms in the text body were manually evaluated for context and labeled as COVID or nonCOVID. In total, 9803 posts were manually classified, of which 2065 were labeled as COVID and 7738 as nonCOVID; these posts were subsequently classified using support vector machines. An additional 818 posts made in the context of the COVID-19 pandemic were identified, bringing the total number of

XSL•FO
RenderX

pandemic-specific posts to 2883. The remaining unclassified posts from the identified pandemic period were labeled as nonCOVID.

Data Analysis

The LDA topic modeling algorithm [18] was applied in two iterations to identify topics of discussion by obtaining clustering words belonging to a single topic. A total of 85,266 posts analyzed, with 2682 being specific to the COVID-19 pandemic. The first iteration of topic modeling was performed on the entire processed dataset of 85,266 posts made between January 1, 2018, and March 5, 2021. The second iteration was performed on a subset of this dataset comprising the 2682 posts specifically related to the COVID-19 pandemic.

Sentiment analysis was performed to understand the subjective emotions or sentiment associated with each post, using a normalized compound score between −1 and 1. The compound score thresholds for classifying text into 3 sentiment categories were as per the VADER algorithm [19]: positive (≥0.05), neutral (between −0.05 and 0.05), and negative (≤−0.05). Further statistical analysis was performed with R software version 4 (R Foundation for Statistical Computing), to calculate mean sentiment scores and 95% CI.

Ethical Considerations

According to the University of Toronto's research ethics guidelines, ethical approval was not sought for this study as per exemptions in Section 1, as Reddit was assumed to be a public data source with no reasonable expectation of privacy [20]. Moreover, there was no direct interaction between the researchers and participants; therefore, ethical approval was not necessary, as per Section 2 of the guidelines [20]. The data used for qualitative analysis were scraped directly from Reddit. As users do not need authentication to view Reddit forums, it was assumed that the users who made these posts were aware that they would be displayed publicly. As this is a publicly available dataset, participants were not compensated.

Data for this study was collected through the Pushshift API. Over 100 published research studies have already used Reddit data extracted through this API [21]. Proferes et al [22] conducted a systematic analysis of 727 manuscripts that used Reddit as a data source and found that fewer than 15% mentioned undergoing any form of ethical review.

While the usernames of the post authors' were obtained, they were assumed to be pseudonyms and not their actual names. However, we acknowledge that some users may have integrated their real names into their usernames; in this study, the usernames were removed to deidentify the data. Reddit does not provide identifiable characteristics of individual users, such as their names, genders, or geographical locations. However, it is possible that some users may include identifiable information within their posts. In this study, such identifiable information was not extracted or analyzed. Since the posts could not be traced to any individual, no compensation was provided.

Results

Topics of Discussion

The final processed dataset consisted of 85,266 posts made between January 1, 2018, and March 3, 2021. The study methodology is shown in Figure 1. Data were collected from Reddit diabetes communities through the Pushshift API and classified as COVID or nonCOVID. LDA topic modelling and sentiment analysis was conducted on processed text. The topics were manually labeled and categorized into broader themes as shown in Table 1.

Figure 1. Study methodology conducted in early 2021. LDA: latent Dirichlet allocation; SVM: support vector machine.

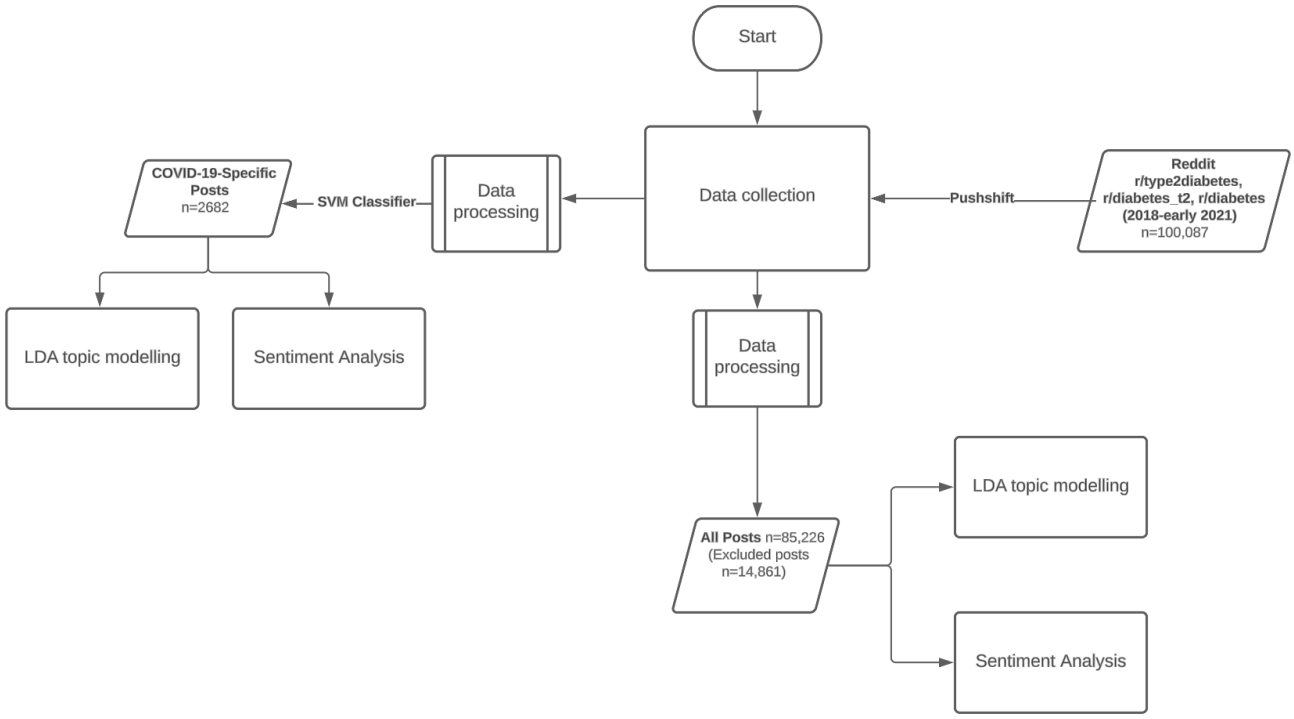

XSL•FO
RenderX

Table . Words extracted from topic modeling of processed data of all Reddit posts in diabetes communities between January 2018–March 2021.

Themes and topic names	Words
Theme 1: Managing blood glucose	
Using blood glucose meters	meter, insurance, reading, check, strip, libre, sensor, finger, cover, cgm, free, pay, app, stick, test, track, buy, freestyle, monitor, difference
Controlling blood glucose	glucose, blood, high, glucose, level, normal, low, range, control, number, spike, check, average, mgdl, raise, point, fast, affect, measure, test
Controlling HbA_{1C} levels	ac, month, weight, year, lose, start, med, diagnose, metformin, week, diet, stop, lb, change, medication, drop, loss, exercise, pound, control
Rationale for glucose variability	insulin, type, body, resistance, gt, disease, people, medication, risk, cure, increase, reverse, cell, case, pancreas, term, treatment, long, produce, diabetic
Theme 2: Managing lifestyle	
Exercising for diabetes care	work, time, exercise, walk, bit, long, pretty, day, start, minute, run, ill, thing, week, stress, home, stay, big, lot, hit
Timing meals and snacks	fast, hour, meal, bg, metformin, morning, time, number, week, mg, night, eat, effect, day, start, low, dinner, breakfast, sleep, reading
Eating low carbohydrate or ketogenic diet	diet, low, keto, carb, fat, body, protein, calorie, lot, liver, carbs, healthy, process, cholesterol, reduce, energy, high, carbohydrate, intake, muscle
Controlling carbohydrates and alcohol	eat, carbs, food, drink, water, cut, meal, carb, avoid, lot, spike, soda, fiber, drinking, limit, small, alcohol, gram, sweet, ate
Options for food	bread, rice, add, cheese, egg, potato, meat, cream, fruit, veggie, chicken, taste, pasta, recipe, snack, butter, nut, coffee, salad, milk
Theme 3: Coping and receiving support	
Attitudes towards disease	change, life, great, lot, thing, work, dr, make, learn, hope, luck, live, easy, control, love, healthy, manage, hard, lifestyle, sound
Caregiving for family and history	feel, time, family, year, felt, thought, happen, care, make, deal, mom, friend, turn, kind, super, shit, die, guess, hard, suck
Theme 4: Obtaining diabetes care	
Interacting with care team	doctor, test, result, advice, talk, doc, give, diagnose, endo, read, appointment, wait, endocrinologist, follow, diagnosis, time, thought, adjust, mention, idea
Educating for self-management	people, post, health, question, medical, understand, research, information, study, read, answer, agree, link, support, comment, patient, diabetic, base, opinion, group
Managing complications/comorbidities	issue, problem, symptom, pain, foot, effect, bad, eye, year, damage, heart, experience, kidney, stomach, hand, stop, happen, shot, neuropathy, vision

A subset of 2682 posts written between February 28, 2020, and February 28, 2021, was classified as COVID-specific. The topics were manually labeled as shown in Table 2.

XSL•FO
RenderX

Table . Words extracted from topic modeling of processed data from all COVID–related Reddit posts in diabetes communities from 28 February 2020 to 28 February 2021.

Topic Name	Words
Coping with poor mental health	feel, day, week, work, weight, lose, walk, bit, symptom, bad, lot, gym, night, happen, end, great, covid, felt, ill, diagnosis
Accessing doctor and medications and controlling blood glucose	doctor, test, ac, year, insulin, month, metformin, diagnose, low, glucose, diet, week, med, exercise, change, level, stop, reading, check, medication
Changing food habits during the pandemic	eat, food, carbs, lot, meal, low, carb, diet, hard, hour, thing, water, keto, fast, cut, easy, glucose, add, stuff, rice
Impact of stress on blood glucose levels	blood, glucose, high, time, stress, number, long, exercise, bg, body, problem, morning, normal, control, sleep, make, level, effect, change, kind
Changing status of employment and insurance	work, home, hospital, today, time, year, talk, give, advice, wait, insurance, visit, guess, state, order, meter, live, strip, situation, friend
Risk of COVID complications	covid, type, health, care, risk, sick, control, issue, virus, disease, diabetic, question, pandemic, mask, hand, case, infection, wear, patient, home

Sentiment Analysis

Using the VADER algorithm, the compound sentiment score of each post was determined. The sentiment across all posts in the dataset was positive with a mean compound sentiment score of 0.23 (95% CI 0.23-0.24).

Comparison of Sentiment of COVID Versus NonCOVID Posts

The sentiment was compared between posts that were classified as COVID and nonCOVID. The mean compound scores of COVID-related posts remained within the positive threshold (sentiment score 0.12, 95% CI 0.01-0.15) but was lower than that of nonCOVID-related posts (sentiment score 0.25, 95% CI 0.24-0.25).

Sentiment by Theme

Coping and receiving support had the highest mean compound score (sentiment score 0.34, 95% CI 0.33-0.35), followed by managing lifestyle (sentiment score 0.25, 95% CI 0.25-0.26), obtaining diabetes care (sentiment score 0.19, 95% CI 0.18 - 0.20), and managing blood glucose (0.19, 95% CI 0.18 - 0.19) (Figure 2). All mean compound scores fall in the threshold to be classified as positive, with coping and receiving support demonstrating the strongest positive intensity score.

Figure 2. Mean compound sentiment score of all posts, sorted by theme and classified as covid-related, noncovid-related, or combined.

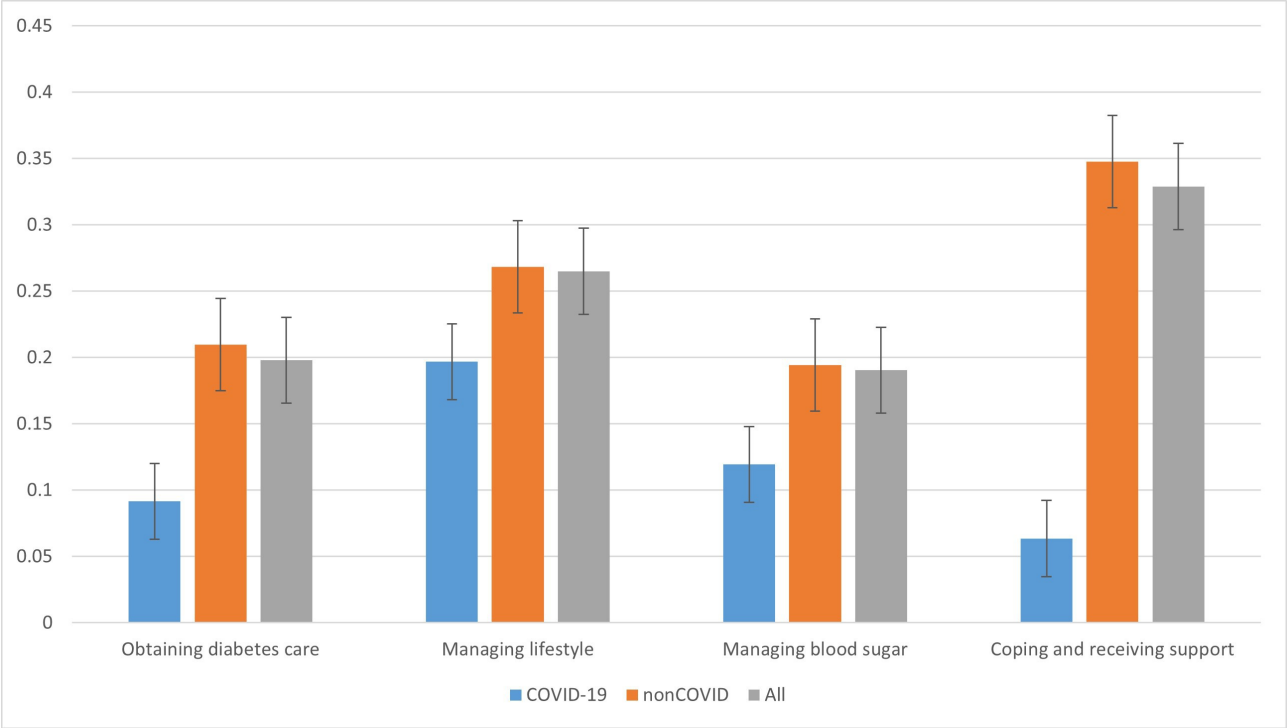

XSL•FO
RenderX

Sentiment by Theme Amongst COVID-Classified and NonCOVID Posts

Mean compound scores were further analyzed among posts classified as COVID-related or as nonCOVID-related. Coping and receiving support had the highest mean compound score (sentiment score 0.35, 95% CI 0.33 - 0.36), followed by managing lifestyle (sentiment score 0.27, 95% CI 0.26 - 0.28), obtaining diabetes care (sentiment score 0.21, 95% CI 0.20 - 0.22), and managing blood glucose (0.19, 95% CI 0.18 - 0.20). However, for COVID-related posts, coping and receiving support had the lowest mean compound score (sentiment score 0.06, 95% CI −0.02 to 0.13), which was slightly above the threshold for positive classification. Ranked from lowest to highest, obtaining diabetes care had the second-lowest mean compound score (sentiment score 0.09, 95% CI 0.05 - 0.13), followed by managing blood glucose (sentiment score 0.12, 95% CI 0.08 - 0.16), and managing lifestyle (sentiment score 0.20, 95% CI 0.15 - 0.24).

Sentiment by COVID-Specific Topics

Sentiment was further analyzed across the topics generated for the COVID-related posts (Table 3).

Table . Mean compound sentiment score of COVID-related posts sorted by topic.

Topic	Compound sentiment score, mean (95% CI)
Accessing doctor and medications and controlling blood sugar	0.14 (0.09 - 0.20)
Changing status of employment and insurance	0.17 (0.13 - 0.22)
Changing food habits during the pandemic	0.25 (0.20 - 0.31)
Coping with poor mental health	0.04 (−0.01 to 0.11)
Risk of COVID-19 complications	0.09 (0.03 - 0.14)
Impact of stress on blood glucose levels	0.03 (–0.03 to 0.08)

Impact of stress on blood glucose levels and coping with poor mental health had mean compound scores that were within the threshold of being classified as neutral (sentiment score 0.03, 95% CI −0.03 to 0.08 and 0.04, 95% CI −0.01 to 0.11, respectively). These were followed by risk 0f COVID complications (sentiment score 0.09, 95% CI 0.03 - 0.14), accessing doctor and medications and controlling blood glucose (sentiment score 0.14, 95% CI 0.09 - 0.20), and changing status of employment and insurance (sentiment score 0.17, 95% CI 0.13 - 0.22), ranked from lowest to highest. Finally, changing food habits had the highest mean compound score (sentiment score 0.25, 95% CI 0.20 - 0.31).

Discussion

Our study confirmed that T2D subreddits were used for peer support to discuss themes related to diabetes management in depth. Oyebode and Orji [23] and Griffis et al [24] found that the themes for discussion of diabetes management on online platforms focused on nutrition, lifestyle, symptoms, treatments, research, risk factors, and supplements or remedies. Additionally, we found that topics of discussion also included glycemic control and support-seeking behaviors. The overall emerging themes identified in this study include health behaviors such as glycemic control, exercise, and nutrition, along with attitudes towards diabetes, interactions with family and caregivers, management of comorbidities, and educational resources. Topics of discussion were specific to the COVID-19 pandemic and were generally associated with lower sentiment, with specific discussions focusing on glycemic and lifestyle management under unique circumstances, including changing food habits, increased stress, deteriorated mental health, difficulties in accessing health care, and fear of complications due to COVID-19. Additional themes of discussion included access to diabetes care during the pandemic and the impact on employment and insurance during this period.

Topics Discussed on Reddit Communities

Topics discussed on Reddit communities for people managing T2D included glycemic control, exercise, and nutrition, indicating that users were discussing health behaviors. Additionally, users discussed their attitudes toward their disease, interactions with their families and caregivers, management of comorbidities, and shared educational resources. This indicated that online forums provide opportunities for holistic discussions that consider the person as a whole, including the social and psychological factors impacting the management of diabetes. The possibilities for such discussions are endless, which may not be captured in structured clinical settings.

COVID-Specific Discussions on Reddit Forums

Topics extracted from COVID-specific posts were about accessing medical care, the impact of stress on glycemic control, changing food habits, poor mental health, and fear of complications due to COVID. This was in line with a literature review of conventional qualitative studies that assessed the impact of the pandemic on people with T2D [8]. Additionally, we found that the changing status of employment and insurance status was also a topic of discussion among the COVID-specific posts. These data supports our finding that discussions among people with T2D on Reddit offer insights from a holistic perspective, considering various aspects of a person's life related to their disease.

Sentiment Analysis

For all posts in the dataset, the mean compound sentiment score as per the VADER algorithm was 0.23 (95% CI, 0.23 - 0.24). As this score is above the threshold of 0.05, we conclude that, on an average, posts in the Reddit communities for people with diabetes were positive.

A significant difference in sentiment was observed between the COVID-related and nonCOVID posts. Overall, the mean

XSL•FO
RenderX

compound score of posts labeled as "nonCOVID" was 0.24 (95% CI 0.24 - 0.25), while posts labeled as COVID had a mean compound score of 0.12 (95% CI 0.01 - 0.15), which was significantly lower. While we cannot conclude that the COVID-labeled posts were overall negative in tone, their sentiment intensity was lower.

Among the nonCOVID posts, coping and receiving support was associated with the highest mean compound score (sentiment score 0.35, 95% CI 0.33 - 0.36). However, for COVID-related posts, coping and receiving support had the lowest mean compound score (sentiment score 0.06, 95% CI −0.01 to 0.13), slightly above the threshold to be classified as positive. This indicated that the pandemic negatively impacted people with T2D.

While specifically analyzing the topics generated only among the COVID-related posts, changing food habits (sentiment score 0.25, 95% CI 0.20 - 0.31) had the highest mean compound score, indicating that even though eating patterns were impacted during the pandemic, there was peer support and camaraderie in discussions about food choices. Conversely, the impact of stress on blood glucose levels and coping with poor mental health had the lowest mean compound scores (sentiment score 0.03, 95% CI −0.03 to 0.08 and 0.05, 95% CI −0.01 to 0.11, respectively). These low sentiment scores suggest that the pandemic resulted in anxiety that affected glycemic control and resulted in poorer mental health outcomes.

Impact of the Pandemic on Managing T2D

Our study identified key themes for managing T2D during the pandemic from Reddit communities and the associated sentiment toward these themes.

Access to Diabetes Care During the Pandemic

During the COVID-19 pandemic, accessing care was perceived as a significant barrier to managing T2D, as more clinical visits were through web-based platforms. Khader, Jabeen, and Namoju [25] found that in a study of 1582 participants in India, the frequency of clinical visits was reduced in 87.28% of participants, and 87.81% of participants did not have access to health services. Al-Sofiani et al [26] found that among 568 participants in the Arab Gulf region, a lack of communication with the health care providers was associated with higher odds of depression and anxiety among people with diabetes. These findings were further confirmed through our sentiment analysis. Among all posts, the mean compound sentiment score for posts related to the theme of obtaining diabetes care was 0.19 (95% CI 0.18 - 0.20); however, among the COVID-specific posts, this score significantly dropped to 0.09 (95% CI 0.05 - 0.13).

Further, topic modeling among COVID-specific posts also identified access to doctor and controlling blood glucose as a topic for discussion. While barriers to accessing health care providers existed, another barrier including the fear of acquiring a COVID-19 infection and avoiding hospitalization in potentially dangerous situations was identified. Further research should examine the impact of reduced in-person health care visits among people with T2D.

Impact on Employment During the Pandemic

Efforts to curb the spread of the virus resulted in employers requiring employees to work from home [27]. As this was a disruption to daily life, Reddit discussions suggested that this disruption impacted the health behaviors of individuals managing T2D.

In addition to work-from-home measures, the pandemic also resulted in the loss of employment due to the shutting down of businesses to curb the spread of the virus. Unemployment not only affected income and standard of living but also decreased their sense of purpose [28], potentially impacting health behaviors due to increased stress. Furthermore, many unemployed people were also impacted by changes to their health insurance caused by job loss [29,30]. From Reddit discussions, it could be inferred that loss of insurance created barriers to obtaining medications, blood glucose meters, and test strips, ultimately affecting glycemic control for people with T2D.

Glycemic Control During the Pandemic

Managing blood glucose was a major theme for discussion across our dataset. Glycemic control is a major component of self-management of T2D and was expected to be a major theme of discussion among peers. Our analysis found that specifically during the pandemic, users of the Reddit communities for managing T2D reported that the pandemic-related stress affected their blood glucose levels. Laboratory studies have demonstrated that psychological stressors are linked to hyperglycemia [31-33].

Overall, managing blood glucose was the theme with the lowest mean sentiment score (0.19, 95% CI 0.18 - 0.19), which was further reduced among the COVID-related posts (0.12, 95% CI 0.08 - 0.16). This correlation suggested that attitudes toward managing blood glucose are associated with lower sentiment compared to other themes and that the pandemic further increased concerns about glycemic management. Specifically, the impact of stress on blood glucose levels had the lowest sentiment score (0.03, 95% CI −0.03 to 0.08), suggesting that users had increased anxiety about glycemic control.

Additionally, the topic associated with accessing a doctor also included terms related to controlling blood glucose. This suggested that reduced access to a health care provider may have affected glycemic control in people with T2D through limited A1c monitoring or medication access during the pandemic. Changes to employment were associated with reduced access to health insurance, which further resulted in reduced access to medications, blood glucose meters, and test strips, affecting glycemic control.

Lifestyle Management During the Pandemic

In our dataset, lifestyle management was the most frequent theme of discussion, with a positive mean sentiment score of 0.25 (95% CI 0.25 - 0.26). Among the COVID-specific posts, this sentiment score dropped to 0.20 (95% CI 0.15 - 0.24) but was overall positive. Changing food habits during the pandemic was associated with the highest mean sentiment score (0.25, 95% CI 0.20 - 0.31), indicating that increased unhealthy food

XSL•FO
RenderX

consumption may have been a coping mechanism associated with positive emotions through the stressful time.

Interestingly, the terms “gym” and “walk” were included among the COVID-specific posts under the topic coping with poor mental health. This suggests that reduced physical activity was attributed to gym closures in the initial months of the pandemic and people relied on walking as a means of physical activity. Furthermore, physical activity is also another means of coping with stressors and many people were unable to be physically active while staying at home. Further investigation by clinicians is needed to help people with T2D cope during changes such as pandemic-related lockdowns, so that they could continue to maintain a healthy lifestyle during stressful times.

Mental Health During the Pandemic

The mental health impact of this pandemic were expected to be long-term due to the extreme measures necessary to prevent the spread of the virus and the resulting economic implications [34]; people with T2D were no exceptions to this. In our study, we found that while coping and receiving support was associated with the highest mean sentiment overall in our dataset (sentiment score 0.33, 95% CI 0.33 - 0.35), among the COVID-19–related posts, it was associated with the lowest mean sentiment (sentiment score 0.06, 95% CI −0.01 to 0.13). This finding suggests that while Reddit forums for managing T2D have been helpful overall for people with the disease, the pandemic negatively impacted how people coped and their attitudes toward their disease. Specifically, among the COVID-19–related posts, discussions about coping with poor mental health and risk of COVID complications were associated with lower mean sentiment scores (0.05, 95% CI −0.01 to 0.11 and 0.09, 95% CI 0.03-0.14, respectively). Although our study demonstrated that people with T2D used peer support to cope with the stressors of the pandemic, it also highlights the negative impact caused by these psychological stressors. Moreover, the implications of potentially being exposed to the virus posed an additional source of anxiety, given the risk of serious COVID-19 complications among people with T2D. As discussed in this paper, the implications of poor mental health and anxiety regarding COVID-19 exposure affected glycemic control and lifestyle management, both of which are necessary to prevent complications from diabetes. Furthermore, the anxiety related to COVID-19 exposure served as a barrier to seeking care at urgent care centres and hospitals when necessary.

Summary of Findings

This study builds upon the work of Nagpal et al [11], which suggests that public social media forums can serve as a form of patient-generated health data for individuals with chronic conditions such as T2D. Social media forums such as Reddit provide a platform for individuals with chronic diseases to discuss their condition, particularly during times of collective chaos, such as the global COVID-19 pandemic; the insights generated from these discussions provide a valuable source of qualitative knowledge. Consistent with the findings from conventional qualitative studies, this study found that health behaviors of people with T2D were impacted during the pandemic, along with declines in mental health [8], as reflected in the sentiment scores. In particular, the pandemic raised concerns about COVID-19 exposure and access to health care. However, our analysis also indicated that employment and insurance were impacted during the pandemic, affecting how people managed their diabetes care. This finding suggests that social media data can be a valuable source of qualitative data to capture health behaviors during particular periods, such as the global COVID-19 pandemic. As the world continues to evolve due to geopolitics, wars, diseases, and climate change, social media can be a source of qualitative data to gauge attitudes toward chronic disease management.

Limitations of the Study

While topic modeling and sentiment analysis were performed on the dataset, researchers did not manually examine each post for context and cannot comment on the quality of discussions within the forum. Additionally, we cannot confirm that all users who contributed posts were people living with T2D. This study only examined 3 Reddit communities mentioned earlier, and did not include other subreddits related to COVID-19 or mental health.

Furthermore, there is no information about the demographics of users specific to the diabetes forums of Reddit, and we assumed that their demographics were similar to those of general Reddit users. Under this assumption, there may be a sampling bias in our findings, as Reddit users are predominantly male and within the 18 - 49 age demographic [35,36]. By country, the United States has the largest number of Reddit users, with other users mostly residing in high-income, English-speaking countries [37]. However, we cannot determine the specific demographics of the users in the Reddit diabetes communities and have assumed that the stated demographics represent the entire Reddit platform.

Given that COVID-19 implications such as number of cases, lockdown policies, and public attitudes toward the pandemic varied by geographical areas, even within the same country, it was beyond the scope of this study to correlate our findings with these factors, as we did not have geographical information of the users.

Conclusion

Themes of Discussion

Using LDA topic modeling from 85,266 posts made between January 2020 and early March 2021 on Reddit forums for people managing T2D, we found 14 topics of discussion that could be assigned into four themes: managing blood glucose, managing lifestyle, coping and receiving support, and obtaining diabetes care. These findings are in line with recommended health behaviors for T2D management that include glycemic control, healthy nutrition, increasing physical activity, reducing sedentary time, and taking prescribed medications. Additionally, discussions encompassed attitudes toward the disease, interactions with family members and caregivers, managing comorbidities, and the sharing of educational resources. These findings validate that Reddit forums for diabetes management not only serve as a support network for people with T2D, but also facilitate discussions that are holistic in nature and extend beyond treating the symptoms of the disease. Among COVID-19–specific posts, topics of discussion included coping

XSL•FO
RenderX

with poor mental health, accessing doctor and medications and controlling blood glucose, changing food habits during the pandemic, impact of stress on blood glucose levels, changing status of employment and insurance, and risk of COVID-19 complications

Impact on Attitudes Toward T2D Management

The posts classified as COVID-19–related had significantly lower mean sentiment scores than those classified as nonCOVID-19–related, suggesting that the pandemic negatively affected attitudes toward T2D management. Among posts classified as COVID-19–related, topics related to mental health and the impact of stress on glycemic control were associated with lower sentiment scores, whereas posts relating to changing food habits were associated with higher sentiment scores.

Significance of Findings

Our study demonstrates that social media analysis is a simpler and effective method for understanding the behaviors and attitudes of people managing chronic diseases such as diabetes, compared to other qualitative inquiry methods. Topics of discussion and sentiment can potentially be generated in real time and over a long period, directly capturing the perspectives of the patient; this is unlike clinical visits, where data is generated from the perspective of the clinician and only presents a snapshot of the patient's condition. The benefit of social media posts is that they offer infinite possibilities for discussion and provide a more holistic perspective, potentially capturing insights that would have otherwise been overlooked in structured surveys or interviews.

This study has implications on both macro- and microlevels. On a macro scale, extracting data from social media forums can provide an epidemiological overview of behavioral trends of people with T2D, potentially guiding public health policy and formulating treatment guidelines, particularly during public health emergencies such as the COVID-19 pandemic.

Acknowledgments

This study was conducted as part of MSN's master's thesis project and funding was provided by the Centre for Digital Therapeutics. No generative AI was used for this paper.

Data Availability

The datasets generated during or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: MN, NJ, PM, DS, JAC
Data curation: MN, NJ
Formal analysis: MN, NJ, PM
Funding acquisition: JAC
Investigation: MN
Project administration: JAC
Supervision: JAC
Validation: DS
Writing – original draft: MN
Writing – review & editing: MN,
MN received guidance on the methodology from NJ and PPM. DS serves as the clinical advisor for this study. All listed authors approved the final version of this paper.

Conflicts of Interest

None declared.

References

1. Global Health Risks: Mortality and Burden of Disease Attributable to Selected Major Risks: World Health Organization; 2009. URL: https://www.who.int/publications/i/item/9789241563871 [accessed 2025-02-09]
2. Sherifali D, Viscardi V, Bai JW, Ali RMU. Evaluating the effect of a diabetes health coach in individuals with type 2 diabetes. Can J Diabetes 2016 Feb;40(1):84-94. [doi: 10.1016/j.jcjd.2015.10.006] [Medline: 26827684]
3. Diabetes Canada Clinical Practice Guidelines Expert Committee. Diabetes Canada 2018 Clinical practice guidelines for the prevention and management of diabetes in Canada. Can J Diabetes 2018;42(Suppl1):S1-325.
4. Mental Health and COVID-19. World Health Organization. 2021 Aug 23. URL: https://www.who.int/teams/mental-health-and-substance-use/covid-19 [accessed 2025-02-09]
5. COVID-19 and diabetes. International Diabetes Federation. 2021 Aug 23. URL: https://www.idf.org/aboutdiabetes/what-is-diabetes/covid-19-and-diabetes/1-covid-19-and-diabetes.html [accessed 2025-02-09]

XSL•FO
RenderX

6. Corrao S, Pinelli K, Vacca M, Raspanti M, Argano C. Type 2 diabetes mellitus and COVID-19: a narrative review. Front Endocrinol (Lausanne) 2021;12:609470. [doi: 10.3389/fendo.2021.609470] [Medline: 33868163]
7. Rajpal A, Rahimi L, Ismail-Beigi F. Factors leading to high morbidity and mortality of COVID-19 in patients with type 2 diabetes. J Diabetes 2020 Dec;12(12):895-908. [doi: 10.1111/1753-0407.13085] [Medline: 32671936]
8. Nagpal MS, Jalali N, Sherifali D, Morita PP, Cafazzo JA. Managing type 2 diabetes during the COVID-19 pandemic: scoping review and qualitative study using systematic literature review and reddit. Interact J Med Res 2024 Aug 8;13:e49073. [doi: 10.2196/49073] [Medline: 39116432]
9. Amsah N, Md Isa Z, Ahmad N, Abdul Manaf MR. Impact of COVID-19 pandemic on healthcare utilization among patients with type 2 diabetes mellitus: a systematic review. Int J Environ Res Public Health 2023 Mar 4;20(5):4577. [doi: 10.3390/ijerph20054577] [Medline: 36901588]
10. Eberle C, Stichling S. Impact of COVID-19 lockdown on glycemic control in patients with type 1 and type 2 diabetes mellitus: a systematic review. Diabetol Metab Syndr 2021 Sep 7;13(1):95. [doi: 10.1186/s13098-021-00705-9] [Medline: 34493317]
11. Nagpal MS, Barbaric A, Sherifali D, Morita PP, Cafazzo JA. Patient-Generated data analytics of health behaviors of people living with type 2 diabetes: scoping review. JMIR Diabetes 2021 Dec 20;6(4):e29027. [doi: 10.2196/29027] [Medline: 34783668]
12. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. J Mach Learn Res 2003 Mar 1;3:993-1022. [doi: 10.1162/jmlr.2003.3.4-5.993]
13. vaderSentiment 3.3.2. PyPI. URL: https://pypi.org/project/vaderSentiment [accessed 2021-08-24]
14. R/type2diabetes. Reddit. URL: https://www.reddit.com/r/type2diabetes/ [accessed 2021-08-24]
15. R/diabetes_t2. Reddit. URL: https://www.reddit.com/r/diabetes_t2/ [accessed 2021-08-24]
16. R/diabetes. Reddit. URL: https://www.reddit.com/r/diabetes/ [accessed 2021-08-24]
17. Pushshift reddit API documentation. GitHub. URL: https://github.com/pushshift/api [accessed 2024-08-24]
18. Blei DM. Probabilistic topic models. Commun ACM 2012 Apr 1;55(4):77-84. [doi: 10.1145/2133806.2133826]
19. Bonta V, Kumaresh N, Janardhan N. A comprehensive study on lexicon based approaches for sentiment analysis. AJCST 2019;8(S2):1-6. [doi: 10.51983/ajcst-2019.8.S2.2037]
20. Activities exempt from human ethics review. University of Toronto - Division of the Vice-President, Research & Innovation. URL: https://research.utoronto.ca/ethics-human-research/activities-exempt-human-ethics-review [accessed 2024-05-30]
21. Baumgartner J, Zannettou S, Keegan B, Squire M, Blackburn J. The pushshift reddit dataset. proceedings of the international AAAI conference on web and social media, 14(1), 830-839. 2020 Presented at: Proceedings of the International AAAI Conference on Web and Social Media p. 830-839. [doi: 10.1609/icwsm.v14i1.7347]
22. Proferes N, Jones N, Gilbert S, Fiesler C, Zimmer M. Studying reddit: a systematic overview of disciplines, approaches, methods, and ethics. Social Media + Society 2021 Apr;7(2):205630512110190. [doi: 10.1177/20563051211019004]
23. Oyebode O, Orji R. Detecting factors responsible for diabetes prevalence in Nigeria using social media and machine learning. Presented at: 2019 15th International Conference on Network and Service Management (CNSM); 2019; Halifax, NS, Canada p. 1-4. [doi: 10.23919/CNSM46954.2019.9012679]
24. Griffis H, Asch DA, Andrew Schwartz H, Ungar L, Buttenheim AM, Barg FK, et al. Using social media to track geographic variability in language about diabetes: infodemiology analysis. JMIR Diabetes 2020;5(1):e14431. [doi: 10.2196/14431]
25. Khader MA, Jabeen T, Namoju R. A cross sectional study reveals severe disruption in glycemic control in people with diabetes during and after lockdown in India. Diabetes Metab Syndr 2020;14(6):1579-1584. [doi: 10.1016/j.dsx.2020.08.011] [Medline: 32858476]
26. Al-Sofiani ME, Albunyan S, Alguwaihes AM, Kalyani RR, Golden SH, Alfadda A. Determinants of mental health outcomes among people with and without diabetes during the COVID-19 outbreak in the Arab Gulf Region. J Diabetes 2021 Apr;13(4):339-352. [doi: 10.1111/1753-0407.13149] [Medline: 33351239]
27. Savić D. COVID-19 and work from home: digital transformation of the workforce. Grey Journal (TGJ) 2020 Jun 1;16(2):101-104 [FREE Full text]
28. Unemployment and mental health. Institute for Work & Health. 2009 Aug 24. URL: https://www.iwh.on.ca/sites/iwh/files/iwh/reports/iwh_issue_briefing_mental_health_2009.pdf [accessed 2021-08-24]
29. Gangopadhyaya A, Garrett AB. Unemployment, health insurance, and the COVID-19 recession. SSRN Journal 2024 Apr 1. [doi: 10.2139/ssrn.3568489]
30. Garrett AB, Gangopadhyaya A. How the COVID-19 recession could affect health insurance coverage. SSRN Journal 2020 May 4. [doi: 10.2139/ssrn.3598558]
31. Goetsch VL, VanDorsten B, Pbert LA, Ullrich IH, Yeater RA. Acute effects of laboratory stress on blood glucose in noninsulin-dependent diabetes. Psychosom Med 1993;55(6):492-496. [doi: 10.1097/00006842-199311000-00004] [Medline: 8310109]
32. Faulenbach M, Uthoff H, Schwegler K, Spinas GA, Schmid C, Wiesli P. Effect of psychological stress on glucose control in patients with type 2 diabetes. Diabet Med 2012 Jan;29(1):128-131. [doi: 10.1111/j.1464-5491.2011.03431.x] [Medline: 21883440]

XSL•FO
RenderX

33. Bruce DG, Chisholm DJ, Storlien LH, Kraegen EW, Smythe GA. The effects of sympathetic nervous system activation and psychological stress on glucose metabolism and blood pressure in subjects with type 2 (non-insulin-dependent) diabetes mellitus. Diabetologia 1992 Sep;35(9):835-843. [doi: 10.1007/BF00399929] [Medline: 1397778]
34. Iqbal SZ, Li B, Onigu-Otito E, Naqvi MF, Shah AA. The Long-Term Mental Health Effects of COVID-19. Psychiatr Ann 2020 Dec;50(12):522-525. [doi: 10.3928/00485713-20201103-01]
35. Dean B. Reddit usage and growth statistics: how many people use reddit in 2021. Backlinko. URL: https://backlinko.com/reddit-users [accessed 2021-08-24]
36. Agrawal A. The user demographics of reddit: the official app. Medium. 2021 Aug 24. URL: https://medium.com/@sm_app_intel/the-user-demographics-of-reddit-the-official-app-7e2e18b1e0e1 [accessed 2025-02-09]
37. Reddit: traffic by country. Statista. URL: https://www.statista.com/statistics/325144/reddit-global-active-user-distribution/ [accessed 2021-08-24]

Abbreviations

API: Application Programming Interface
LDA: Latent Dirichlet Allocation
T2D: Type 2 Diabetes
VADER: Valence Aware Dictionary for Sentiment Reasoning

Edited by A Mavragani; submitted 22.07.23; peer-reviewed by C Porter, P Khorasani; revised version received 31.12.24; accepted 02.01.25; published 21.02.25.

Please cite as:
Nagpal M, Jalali N, Sherifali D, Morita P, Cafazzo JA
Investigating Reddit Data on Type 2 Diabetes Management During the COVID-19 Pandemic Using Latent Dirichlet Allocation Topic Modeling and Valence Aware Dictionary for Sentiment Reasoning Analysis: Content Analysis
JMIR Form Res 2025;9:e51154
URL: https://formative.jmir.org/2025/1/e51154
doi:10.2196/51154

© Meghan Nagpal, Niloofar Jalali, Diana Sherifali, Plinio Morita, Joseph A Cafazzo. Originally published in JMIR Formative Research (https://formative.jmir.org), 21.2.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

COVID-19 Public Health Communication on X (Formerly Twitter): Cross-Sectional Study of Message Type, Sentiment, and Source

Sana Parveen[1], PhD; Agustin Garcia Pereira[2], MSc; Nathaly Garzon-Orjuela[1], PhD; Patricia McHugh[3], PhD; Aswathi Surendran[4], PhD; Heike Vornhagen[2], PhD; Akke Vellinga[1], PhD

[1]School of Public Health, Physiotherapy and Sports Science, University College Dublin, Belfield, Dublin 4, Ireland, Dublin, Ireland

[2]Data Science Institute, University of Galway, Galway, Ireland

[3]J.E. Cairnes School of Business & Economics, University of Galway, Galway, IrelandIreland

[4]School of Medicine, Ollscoil na Gaillimhe – University of Galway, Galway, Ireland

Corresponding Author:
Sana Parveen, PhD
School of Public Health, Physiotherapy and Sports Science, University College Dublin, Belfield, Dublin 4, Ireland, Dublin, Ireland

Abstract

Background: Social media can be used to quickly disseminate focused public health messages, increasing message reach and interaction with the public. Social media can also be an indicator of people's emotions and concerns. Social media data text mining can be used for disease forecasting and understanding public awareness of health-related concerns. Limited studies explore the impact of type, sentiment and source of tweets on engagement. Thus, it is crucial to research how the general public reacts to various kinds of messages from different sources.

Objective: The objective of this paper was to determine the association between message type, user (source) and sentiment of tweets and public engagement during the COVID-19 pandemic.

Methods: For this study, 867,485 tweets were extracted from January 1, 2020 to March 31, 2022 from Ireland and the United Kingdom. A 4-step analytical process was undertaken, encompassing sentiment analysis, bio-classification (user), message classification and statistical analysis. A combination of manual content analysis with abductive coding and machine learning models were used to categorize sentiment, user category and message type for every tweet. A zero-inflated negative binomial model was applied to explore the most engaging content mix.

Results: Our analysis resulted in 12 user categories, 6 message categories, and 3 sentiment classes. Personal stories and positive messages have the most engagement, even though not for every user group; known persons and influencers have the most engagement with humorous tweets. Health professionals receive more engagement with advocacy, personal stories/statements and humor-based tweets. Health institutes observe higher engagement with advocacy, personal stories/statements, and tweets with a positive sentiment. Personal stories/statements are not the most often tweeted category (22%) but have the highest engagement (27%). Messages centered on shock/disgust/fear-based (32%) have a 21% engagement. The frequency of informative/educational communications is high (33%) and their engagement is 16%. Advocacy message (8%) receive 9% engagement. Humor and opportunistic messages have engagements of 4% and 0.5% and low frequenciesof 5% and 1%, respectively. This study suggests the optimum mix of message type and sentiment that each user category should use to get more engagement.

Conclusions: This study provides comprehensive insight into Twitter (rebranded as X in 2023) users' responses toward various message type and sources. Our study shows that audience engages with personal stories and positive messages the most. Our findings provide valuable guidance for social media-based public health campaigns in developing messages for maximum engagement.

(JMIR Form Res 2025;9:e59687) doi:10.2196/59687

KEYWORDS

public health communication; surveillance; COVID-19; SARS-CoV-2; coronavirus; respiratory; infectious; pulmonary; pandemic; public health messaging; healthcare information; social media; tweets; text mining; data mining; social marketing; infoveillance; intervention planning

XSL•FO
RenderX

Introduction

Public health communication is the scientific research, strategic transmission, and critical evaluation of health information to promote public health [1]. Public health communication initiatives can result in change by increasing awareness, boosting knowledge and forming attitudes when initiatives are well-planned, meticulously carried out, and sustained over time [2].

Social media can be used to quickly disseminate focused public health messages, increasing message reach and interaction with the general public [3,4]. Identifying and understanding information needs, false information, hate speech and discrimination, adherence to precautions, and where concerns lay, aids in the customization of public health strategy and eventually, the development of more informed interventions [5].

Social media can be a valuable resource for learning about people's emotions, concerns and exchanging information. This was shown for instance, when Ebola broke out in Nigeria and public health institutions assisted in containing the Ebola outbreak by tracking social media interactions and disseminating accurate information about the illness [6]. Social media allows public health institutions to track outbreaks in real time and Twitter (rebranded as X in 2023) has been frequently used as a communication tool [7]. The features and status of disease outbreaks can be predicted and explained using information from social media sites and user-generated information has supported the development of early response methods [8].

Social media data text mining can be used for disease forecasting and understanding public awareness of health-related concerns [8]. However, it is still unclear how different social media messages are shared and interpreted or whether different sources (individuals or institutions) communicate efficiently.

During the COVID-19 pandemic [9], social media successfully informed and increased public awareness about this new phenomenon [10]. However, there were considerable differences in the preferred social media platforms, message formats and source sender types [11].

An important concern during the COVID-19 pandemic was the spread of misinformation on social media [12]. Research shows that promoting more messages from reputable, authoritative sources on social media is one of the best ways to prevent misinformation [13].

Examining the content of social media messages provides valuable and timely insights regarding public awareness levels and their needs [14]. While there have been many studies analyzing the tweets on various health issues, there has been limited studies to explore the impact of type, sentiment and source of posts on engagement in public health communication. This study used tweets sent during the pandemic to explore 3 research questions:

- Which sources of information are effective in public health communication?
- What are the most effective types of messages in public health communication?
- Which message type should different sources use to improve engagement?

Methods

Data Collection

Data were extracted from Twitter using a Python script to communicate with Twitter Rest API (application programming interface) using the "Search" endpoint. The "query" parameter was used to filter the results based on the "has:geo" tag, the "place_country:UK"/ "place_country:IE" tag (for Ireland and the United Kingdom), and "lang:en" tag (for English). The "start_time" and "end_time" parameters were used to filter the posts from January 1, 2020 to March 31, 2022.

A basic search was conducted with phrases such as "coronavirus," "SARS-CoV-2," and "COVID-19" on Twitter. Using an iterative method, keywords were added and removed in order to find the most suitable for the data search. The final list of 10 keywords for data extraction included "COVID-19," "pandemic," "SARS-CoV-2," "coronavirus," "SARS-CoV-2 virus," "social distancing," "self-isolation," "self-quarantine," "quarantine," and "new variant." The total number of tweets extracted totaled 867,485.

Data Analysis

Due to the extensive volume of data, a systematic approach was adopted using a 4-step analytical process including sentiment analysis, user-classification, message classification and statistical analysis (Table 1). A combination of manual coding and machine learning (ML) models were used for sentiment, user and message classification. This allowed for multiple rounds of coding to increase the robustness of our results. This methodology is also used by Kummervold et al in their study which shows that by utilizing machine learning models, they could almost exactly match the accuracy of a single human coder when it came to tweet classification. Their research indicates that this automated method, which is dependable and accurate, may be able to guide potentially useful and essential interventions while also freeing up important time and resources for carrying out similar analyses [15].

XSL•FO
RenderX

Table . Phase wise description of the data analysis process.

Phase	Tasks	Methods	Outcome
Data collection	• Identification of keywords related to COVID-19	• Identification of 15 keywords. • After discussion among research team, reduced to 10 keywords for data extraction.	• Final keywords
	• Data extraction	• An academic researcher access was applied for with Twitter. • 867,485 tweets extracted.	
Sentiment analysis	• Allocate sentiment to each tweet	• Random selection of 260 tweets for sentiment allocation (positive, negative, and neutral). • A total of 3 coders manually assigned sentiments to 1 set of 260 random tweets. • Using majority voting, out of 260 tweets, coders agreed on 247 tweets (2 or 3 coders voted the same). • Kappa statistic calculated between all coders and the models. • RoBERTa-base model (highest agreement) applied to the whole dataset.	• Manual coded tweets • Kappa statistic results • Tweets sentiment assigned—full dataset
User classification	• Identify user categories	• User categories identified based on literature on social media supported public health interventions. • Definition of 5 user categories: health institute, health professional, influencer, researcher, and public.	• User categories
	• Coding user profiles	• Abductive coding by 4 Manual coders on 1000 randomly selected tweets to identify new user categories. • After 3 rounds of manual coding, most discrepancies or disagreements resolved through discussion. • Using majority voting, out of 250 tweets, coders agreed on 165 tweets (at least 3 coders voted the same). • A total of 3 ML models selected to assign user categories to same set of 250 tweets used in Round 3 of manual coding. • Bag of words were defined for each user category for user categorization by the ML models. • Accuracy score and F_1-score calculated between ML models and manual coders. • SetFit model (highest agreement) applied to the whole dataset.	• User classification coding—Round 1, 2, and 3 • Bag of words • Accuracy and F_1-score • Tweets users assigned—full dataset

XSL•FO
RenderX

Phase	Tasks	Methods	Outcome
Message classification	• Identify message categories	• Message categories defined based on review of public health communication literature. • Definition of 5 message categories: humor, shock/disgust, informative/educational, opportunistic, and personal stories.	• Message categories
	• Coding tweets	• Abductive coding by 5 manual coders on randomly selected tweets in 3 rounds: 100 tweets in round 1, 100 tweets in round 2, and 250 tweets in round 3. • After 2 rounds of manual coding, discrepancies or disagreements resolved through discussion. • A total of 6 final message categories after addition and reduction of categories over 3 rounds. • Using majority voting, out of 250 tweets, coders agreed on 203 tweets (3 or more coders voted the same). • GPT-3 and Setfit model to assign message categories to set of 250 tweets from the final round using 6 message categories. • Accuracy score and F_1-score calculated between models and manual coder. • SetFit model (highest agreement) applied to the whole dataset.	• Manual coding—Round 1, 2, and 3 • Tweets message assigned—Tweet users assigned
Statistical analysis	• Engagement calculation	• Exclusion of zero-follower users (n=537). • Calculation of engagement (sum of likes, replies, retweets and quoted tweet count divided by the respective user's followers)	
	• Zero-inflation model	• Exclusion of public user group (outside of objective of research) due to 87% zero-engagement. • Application of zero-inflated Poisson and zero-inflated negative binomial model due to remainder of 26% zero-engagement. • Selection of zero-inflated negative binomial model (best fit) with informative/educational message type and neutral sentiment as comparator.	• Final model results

Our cross-sectional observational study aims to provide guidance for social media-based public health campaigns in developing messages for maximum engagement. The study adhered to the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) checklist for cross-sectional studies (Checklist 1). Public health interventions are mostly run by governmental agencies or health institutes but can also involve collaborations with researchers, influencers, artists, etc.

XSL•FO
RenderX

Therefore, it is important to study how the public engages with different types of messages from various sources. To ensure a clear distinction and understanding of public engagement, tweets by the general public were removed from our final analysis to focus on engagement received by tweets from all other sources.

Sentiment Analysis

A subset of 260 tweets was randomly extracted for manual annotation by three coders with sentiment labels - positive, negative or neutral (Multimedia Appendix 1). Subsequently, we used 3 ML sentiment analysis models to detect the sentiment for the same tweet subset, which were: distilbert-base-uncased-finetuned-sst-2-english [16], finite automata/bertweet-base-sentiment-analysis [17], and RoBERTa-base for Sentiment Analysis [18].

The kappa statistic was calculated between the 3 coders and between each coder and model. The model with the highest agreement with manual coders based on accuracy score [19] and F_1-score [20] was applied to the full dataset.

User Classification

A directed content analysis with abductive coding was used to explore which sources of information are effective in public health communication. User categorization was based on the profile description and categorized according to an adaptation of the user categories from Cole-Lewis [21]: health institute, health professional, influencer, researcher and public.

User coding was performed in 3 cycles. The profile description of 1000 randomly selected tweets were classified by 4 coders with an overlap of 10% for double coding, that is, coders independently assigned codes to the same set of tweets. The 1000 profile descriptions were categorized into 6 distinct categories with a directed content analysis approach [22] with an additional category “others.” This “others” category was further defined through exploring the profile descriptions, resulting in a total of 12 categories.

The second round aimed to ensure consistency and accuracy between the coders. In total, 100 randomly selected profile descriptions of tweets were allocated to 12 user categories by each coder (Table 2) and agreements, disagreements and potential additions to the categories were discussed. A “bag of words” was compiled for each user category, providing descriptive insights into the characteristics of the users (Table 2 and Multimedia Appendix 2).

XSL•FO
RenderX

Table . User category and bag of words used for classification.

Index	User category	Profile keywords–"bag of words" (Version 6)
1	Health institute	ICGP[a]; WHO[b]; NHS[c]; public health; ECDC[d]; HSE[e]; HPSC[f]; clinic; hospital
2	Health professional	health worker; GP[g]; consultant; nurse; MD[h]; specialist; physician; clinician; surgeon
3	University/Researcher	university; researcher; PhD; student; scientist; academic professor; academia; principal lecturer
4	Influencer	influencer; blogger; vlogger; coach; YouTube
5	Teacher	teacher; teach; school; education; children
6	Politician	politics; government; governor; cabinet; council; minister; councilor; ambassador; MP;[i] Secretary of state; Fine Gael; TD[j]; Mayor
7	Sports	football; rugby; run; swim; tennis; exercise; sports club; pickleball
8	Journalist	journalist; report; news; reporter; columnist; reviewer; media; correspondent; editor
9	Charity	charity; church; ngo; foundation; donations
10	Public	community; union; group; nature; lover; adventure; travel; live; life; world; freedom; farm; pet; cat; dog; walks
11	Known personality	views my own, "True" in verified status
12	Artist	artist; actor; actress; music; writer; singer; photography; movie; sing; play; cinema, orchestra

[a]ICGP: Irish College of General Practitioners.
[b]WHO: World Health Organization.
[c]NHS: National Health Service.
[d]ECDC: European Centre for Disease Prevention and Control.
[e]HSE: Health Service Executive
[f]HPSC: Health Protection Surveillance Centre.
[g]GP: General Practitioner.
[h]MD: Doctor of Medicine.
[i]MP: Member of Parliament.
[j]TD: Teachta Dála.

In the third round, 250 randomly selected profile descriptions were coded by 4 coders and compared with 3 different ML models using the same bag of words: Lbl2TransformerVec model (unsupervised) [23], SGRank model [24], and SetFit model [23].

The ML models classified many influencers as "Public," which was rectified by allocating influencer to any user with more than 3000 followers. The 3 models' performance was evaluated using the profile descriptions of 250 tweets (third round of coding) from 4 coders. Majority voting was used to generate a final coding for the 250 profile descriptions and filtering out profiles that had less than 3 coders agreement. A final dataset of 165 profile descriptions was left after this filtering. The 3 models performance was compared with the manual allocation and the final classification was based on the majority allocation.

Message Classification

Similar to the user classification, a directed content analysis with abductive coding was applied. A review of literature resulted in the identification of 5 message categories from a study by Gough et al [25], which are humor, shock/disgust, educational/informative, personal stories, and opportunistic. The dataset was divided into 2 subsets—public tweets and nonpublic tweets—to determine which type of messages the public engages with most.

A total of 3 manual coding cycles were applied, starting the allocation of 100 random nonpublic tweets to the five message categories and an "other" category. In addition, 2 additional categories emerged, namely, fear and advocacy.

In Round 2, a total of 5 coders allocated 100 tweets to 7 message categories and discussed agreements and disagreements leading to a refinement of the message categories. Fear-based messaging was combined with shock/disgust, personal stories were combined with personal statements and a "not enough information" category was added. In the final round, 250 tweets were categorised into 7 message categories, which are humor, shock/disgust/fear-based, educational/informative, personal

XSL•FO
RenderX

stories/statements, opportunistic, advocacy, and not enough info, and compared to 2 ML models (GPT-3 [OpenAI] and SetFit; Multimedia Appendix 3).

Statistical Analysis

For each tweet, engagement was calculated as the sum of likes, replies, retweets, and quoted tweet count divided by the respective user's follower count [26,27].

For 87% of the tweets, the engagement was zero due to the lack of followers or likes, replies, or quotes (mainly public users). The final dataset excluded this user category resulting in the reduction to 26% zeros. Zero-inflated Poisson and zero-inflated negative binomial models were applied.

Ethical Considerations

Only publicly available data were extracted for this study. All personal identifiable information was deidentified during the data cleaning process.

Results

The total number of tweets extracted was 867,485, which reduced to 802,042 after deleting duplicates. Majority of the tweets (729,619) came from the United Kingdom, 72,350 from Ireland, and 73 were without a location.

The follower count for the dataset showed a wide range, from a maximum of 14,065,098 followers to 0 followers, with an average of 4296 followers. Accounts (users) without followers were investigated to identify inactive accounts or bots and 537 users were removed from the dataset.

Public user category tweets (430,760) were removed from the final dataset as they were not a user category of interest. The final dataset included 370,745 tweets.

Sentiment Analysis

Out of 260 tweets, coders agreed on 247 tweets (this is when 2-3 coders voted the same), which were compared with the accuracy score and F_1-score (weighted) of the 3 ML models. The RoBERTa-base (Model 3) performed the best and was used to assign sentiment to all tweets (Table 3).

Sentiment was negative for 138,379 tweets (37.3%), positive for 84,939 (22.9%), and neutral for 147,427 tweets (39.8%). Positive sentiment tweets had the highest engagement (29.3%), followed by negative (26.7%) and neutral (20.4%; Table 4).

Table . Accuracy and F_1-score for sentiment models with machine learning models.

Model	Accuracy	F_1-score (weighted)
1–Distilbert-base-uncased-finetuned-sst-2-english	0.68	0.61
2–Finite automata/bertweet-base-sentiment-analysis	0.58	0.61
3–RoBERTa-base	0.74	0.76

Table . Frequency and engagement for each sentiment category.

Sentiment	Engagement, %	Frequency, %
Positive	29.3	22.9
Negative	26.7	37.3
Neutral	20.4	39.8

User Classification

Model 3 (assisted) achieved an accuracy score of 0.73 when 4 to 5 coders were in agreement, and achieved an accuracy score of 0.77 when there was at least one match with one coder—192/250 tweets (Table 5).

Health professionals (8%) significantly outnumbered health institutes (1%), while the number of influencers (32%) was more than twice that of journalists (15%) and politicians (13%). Artists (6%) outnumbered individuals associated with sports (2%), and teachers (3%) were one-third in comparison to university/researchers (14%), see Table 6.

Table 7 shows the frequency and percentage of engagement for each user category and provides valuable insights into their impact and effectiveness in engaging audiences. Health professionals have the highest level of engagement (15%), followed by university/researchers (13%) despite a lower frequency (8%). Even though influencers have high frequency (32%) it does not translate into high engagement with influencers having a lower percentage of engagement at 12%.

Journalists and politicians account for 15% and 13%, respectively, with 8% engagement each. Teachers, known personalities. and charities received engagement of 6%, 4%, and 3%, respectively, at lower frequencies. Sports and health Institutes have the lowest engagement each at 1%.

XSL•FO
RenderX

Table . Accuracy and F_1-score for user classification with machine learning models.

Model	Accuracy	F_1-score (weighted)
1-Lbl2TransformerVec	0.26	0.21
2-SGRank	0.43	0.43
3-SetFit	0.69	0.67
3-SetFit (assisted)	0.73	0.73

Table . User categories and their frequency.

User categories	Frequency, n	Frequency, %
Influencer	119,711	32.3
Journalist	54,384	14.7
University/Researcher	50,275	13.6
Politician	47,694	12.9
Health professional	28,963	7.8
Artist	21,730	5.9
Known personality	15,095	4.1
Teacher	12,019	3.2
Charity	11,788	3.2
Sports	5659	1.5
Health institute	3427	0.9

Table . Frequency and engagement for each user category excluding public.

User category	Engagement, %	Frequency, %
Health institute	0.7	0.9
Sports	1.4	1.5
Charity	3.2	3.2
Teacher	5.5	3.2
Known personality	4.1	4.1
Artist	6.1	5.9
Health professional	15.2	7.8
Politician	7.8	12.9
University/Researcher	12.7	13.6
Journalist	8.2	14.7
Influencer	11.5	32.3

Message Classification

The 2 models' performance was evaluated using 250 tweets from 5 coders. Majority voting was used to generate a final coding for the 250 tweets, and filtering out tweets that had less than 3 coders agreement. In addition, 11 tweets were removed because they were identified by the coders as "Not enough information" to code. A final dataset of 203 tweets was left after this filtering. SetFit outperformed GPT-3 and this model achieved 80% accuracy when considering a match with at least 1 coder (192/239 tweets; Table 8).

Personal stories/statements have the highest engagement at 27%, but are not the most frequent tweeted category (22%; Table 9). Shock/disgust/fear-based messages (32%) have 21% engagement. Informative/educational messages have high frequencies (33%) and have 16% engagement. Advocacy messages (8%) have 9% engagement and humor and opportunistic messages have engagements of 4% and 0.5%, and low frequencies 5% and 1%, respectively (Table 10).

XSL•FO
RenderX

Table . Accuracy and F_1-score for user classification with machine learning (ML) models.

	4 and 5 coders agreement		3, 4, or 5 coders agreement	
	Accuracy	F_1-score (weighted)	Accuracy	F_1-score (weighted)
GPT-3	0.60	0.60	0.51	0.49
SetFit	0.74	0.74	0.64	0.62

Table . Frequency and engagement for each message type.

Message type	Engagement, %	Frequency, %
Opportunistic	0.5	0.7
Humor	3.7	4.7
Advocacy	9.1	8.5
Personal stories/statements	26.7	21.8
Shock/disgust/fear-based	21	31.9
Informative/educational	15.5	32.5

Table . Message types frequency and engagement percentage.

Message category	Example tweet	Frequency, n	Engagement, %
Informative/ educational	Coronavirus Daily Update: As at 06 Mar 2022, in the Isle of Man there have been 23328 confirmed cases. #coronavirus #iom #coronaupdate	120,317	15.5
Shock/disgust/ fear-based	@Mysturji @AntacsB @Keir_Starmer General strike? We’re already basically on one. Take to the streets? And die of a pandemic?	118,238	21
Personal stories/ statements	Celebrating my end of Self-isolation period with some improvised gluten free macaroni and, er, fussili and cheese. #norecipe #foodie #satisfied @ Dublin, Ireland	80,881	26.7
Advocacy	‚ÄòEach person who has died in this pandemic is a loved person, a life gone too soon and a family torn apart.' Hold a public inquiry into the Government’s handling of the Covid-19 pandemic #Covid19 - Sign the Petition!	31,340	9.1
Humor	The government says we need to exercise social distancing, stay indoors as much as possible and behave like other people might be carrying a disease like this is all new, but I‚Äôve been doing it since about 2002 #introvert #hermit	17,302	3.7
Opportunistic	In light of #pandemic financial challenges for families, I considered the 10% property tax base increase on Waterford #households trying to recover from the pandemic morally wrong in 2020. Yesterday I voted against #LPT 10% increase & my party again!	2667	0.5

XSL•FO
RenderX

Statistical Analysis

A zero-inflated negative binomial model was applied with informative/educational tweets and neutral sentiment as reference categories. Health professionals received more engagement with advocacy, personal stories/statements and humor. Health Institutes observe higher engagement with advocacy, personal stories/statements and tweets with a positive sentiment (Multimedia Appendix 4).

Journalists and teachers observe higher engagement with advocacy, personal stories/statements and humor while artists have more engagement with shock/disgust/fear-based messages and positive sentiment. Charity organizations, universities, and researchers have more engagement with advocacy and personal stories/statements but not with shock/disgust/fear tweets.

Known personalities have most engagement with advocacy and humor-based, which is opposite for sports entities. Sports entities have more engagement with personal stories/statements and positive sentiment. Politicians and influencers have more engagement with advocacy, personal stories/statements and humor, while they should avoid opportunistic tweets. Table 11 shows the message type and sentiment each user category should tweet and avoid to get maximum engagement.

Table . Findings of zero-inflated model.

User category	Message type to post	Message type to avoid
Health Institute	Advocacy, personal stories/statements, and positive sentiment	Opportunistic, shock/disgust/fear-based, and negative sentiment
Artist	Shock/disgust/fear-based and positive sentiment	Humor and opportunistic
Charity	Advocacy and personal stories/statements	Humor and negative sentiment
Journalist	Advocacy, personal stories/statements and humor	Shock/disgust/fear-based and negative sentiment
Known personality	Advocacy and humor	Shock/ disgust/fear-based
Teacher	Advocacy, personal stories/statements and humor	Shock/disgust/fear-based and negative sentiment
Health professional	Advocacy and personal stories/statements	Opportunistic and shock/disgust/fear-based
Influencer	Advocacy, personal stories/statements and Humor	Opportunistic and negative sentiment
Sports	Personal stories/statements and positive sentiment	Shock/disgust/fear-based
Politician	Advocacy, personal stories/statements, and humor	Opportunistic and negative sentiment
University/Researcher	Advocacy and personal stories/statements	Shock/disgust/fear-based and negative sentiment

Discussion

Principal Findings

This analysis of type, sentiment, and source of COVID-19–related tweets showed what type of tweets have the most engagement for each user group. Overall, personal stories and positive messages have most engagement, even though not for every user group; known persons and influencers have most engagement with humorous tweets. Journalists and teachers were found to have flexibility in their posting strategies, with advocacy, personal stories/statements, and humor being effective for them as well. Artists garnered the most engagement with shock/disgust/fear-based and positive sentiment posts, while charity organizations and universities/researchers had advocacy and personal stories/statements as more engaging message types. Known personalities and sports entities also displayed varying engagement for message types and sentiments, with the former having advocacy and humor-based posts more engaging, while the latter should focus on personal stories/statements and positive sentiment. Politicians and influencers, on the other hand, exhibited higher engagement with advocacy, personal stories/statements, and shock/disgust/fear-based messages, while avoiding opportunistic content.

Positive sentiments in public health messages typically evoke feelings of hope, encouragement, and trust among users, leading to increased sharing behavior [28]. Users are more inclined to engage with and share positive messages that resonate with them emotionally, as they perceive such content as uplifting and supportive [28]. This explains the high engagement received by positive sentiment tweets despite having lower frequency, as shown in Table 4 in this study. Our initial manual categorization ensured a reliable baseline for sentiment analysis in this study. The frequency of negative sentiment tweets and their engagement reflects the widespread anxiety, fear, and uncertainty during the pandemic [29]. The high engagement of tweets with negative sentiments further emphasizes the need to consider emotional content in information dissemination on social media platforms such as Twitter. In this context, future studies should explore the specific impact of tweets by analyzing the responses they generate, including retweets with quotes and replies. This approach would provide additional insights and allow us to evaluate whether tweets achieve their intended impact.

The findings of this study revealed distinct patterns of public engagement across different user categories and message types. Trusted sources are important in shaping public behavior and engagement with health information during crises [30]. Overall,

XSL•FO
RenderX

in the case of users, health professionals received high engagement during the pandemic whereas health institutes received the lowest engagement, maybe reflecting their different use of messages. This highlights the importance of using specific message types for each user category to achieve engagement, as recommended by our study. In addition, health experts, such as general practitioners, communicate health information with greater credibility and persuasiveness than nonexperts or institutes [31].

Internet-based health communities and social media platforms influence public health behaviors and engagement levels [32]. The definition and expansion of sources (user categories) in this study provides a framework to develop specific messaging by user group. Our findings reiterate the importance of understanding audience and tailoring engagement strategies based on category-specific behaviors for optimal engagement [33-35].

Comparison With Previous Work

Most studies have focused on the content of the tweets to understand public reactions or sentiment [36,37], trends during the COVID-19 pandemic [38,39], or vaccinations [40-42].

In one study, Twitter data were explored using machine learning to examine how public opinions and discussions changed throughout the COVID-19 epidemic [36]. Trends in social media discussions during the pandemic were explored using sentiment analysis and topic modeling, which produced useful information about the public discussion around the pandemic [38]. These studies focused mainly on the content of tweets and their engagement but did not explore their association with sources, message type, or sentiments.

The classification of tweets based on types enriched the analysis by capturing the multifaceted nature of communication during the pandemic and provided an understanding of how different message categories influence public engagement and sharing behavior. This study also added to the existing literature by introducing 2 new message categories—advocacy and personal statements—while modifying the existing categories to enable application and provide guidance in framing future public health communications.

Public health literature on message types is very limited. A scoping review on the health risk communication with the public during a pandemic found a lack of studies on the modes of communication [43]. One study discussed the framing of effective COVID-19 messages to connect individuals to authoritative content, emphasising the importance of positive and gain-framed messages [44]. Similarly, personal experiences increased the salience of public health messaging, particularly in promoting sanitation and hygiene practices [45]. Public health messaging during the lockdown in New Zealand showed the importance of consistent messaging principles such as transparency, timeliness, empathy, and clarity [46]. None of these studies used defined message categories and the definition and recommendation of message types for different user categories is a major contribution of our study. This will help content creators, particularly health intervention planners, in choosing the right mix of message and sentiment type to increase their engagement.

A similar study of social media messages explored account type and message structure, taking elements such as hashtags, hyperlinks, mentions, and any images or videos into account but only counting retweets as engagement [8]. They found that tweets with hashtags, videos, and pictures were retweeted more often, while tweets with links had fewer retweets. Furthermore, tweets with sentiment were more frequently retweeted than tweets with neutral sentiments. In our study, the user profile and engagement were explored through engagement metrics such as likes, retweets, reply count, and quote count. We found health institutes to be the least engaged user category, while Xie et al [8] found national health authorities received more engagement when compared with provincial accounts. However, their analysis was limited to the study of the official (national and provincial) public health agencies' Sina Weibo posts only (a China-only microblogging platform).

This study's novel approach lies in the examination of engagement for a 2-year period across different user categories for different message types and sentiment, providing insights into public's response toward different messages and sources. The volume of data and methodology used allows us to provide insights that were not addressed in the existing body of work. This differs from the current studies with similar research objectives, such as a comparative study between Poland and Jordan [47] on social media's role focused on the disparities in platform choices and message efficacy. Another study categorized COVID-19-related tweets into themes to understand public sentiment. The study identified 5 themes for message categories, which were general information, health information, expressions, humor and others, but used a small dataset [48]. Furthermore, a study examined Canadian public health tweets, revealing that tweets promoting action garnered more engagement than purely informational ones. However, retweets were used as the measure for engagement, unlike our study, where we took into consideration a more comprehensive method for calculating engagement [49]. This is another strength of our study, where we compared engagement across message types and user categories instead of solely depending on likes for comparison.

Our findings suggest a correlation between message type, sentiment, source credibility, and engagement. Our study shows that audience engages with personal stories and positive messages the most. Also, with varied users, different types of messages yield engagement. Our study provides guidance for social media–based public health campaigns for developing messages for maximum engagement.

Limitations

We included just 3 factors (sentiment, user type, and message type) for the analysis, which was leading to variance in our analysis. Other factors not recorded or captured may also influence engagement, for instance, time or day of posts, hashtags, images, etc. For our study, we also excluded tweets from the public user category, which was almost 50% of the dataset as it was not required to address this study's research questions.

XSL•FO
RenderX

Conclusion

Our study provides a framework to develop social media messages according to sentiment and message type for different users. Health professionals and institutes and other users can build on the results to improve effective communication through social media channels.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request

Conflicts of Interest

None declared.

Multimedia Appendix 1
Sentiment analysis.
[XLSX File, 44 KB - formative_v9i1e59687_app1.xlsx]

Multimedia Appendix 2
User classification.
[XLSX File, 89 KB - formative_v9i1e59687_app2.xlsx]

Multimedia Appendix 3
Message classification.
[XLSX File, 80 KB - formative_v9i1e59687_app3.xlsx]

Multimedia Appendix 4
Results of zero-inflated model.
[XLSX File, 11 KB - formative_v9i1e59687_app4.xlsx]

Checklist 1
STROBE checklist for cross-sectional studies.
[DOC File, 94 KB - formative_v9i1e59687_app5.doc]

References

1. Bernhardt JM. Communication at the core of effective public health. Am J Public Health 2004 Dec;94(12):2051-2053. [doi: 10.2105/ajph.94.12.2051] [Medline: 15569948]
2. Hornik RC, editor. Public Health Communication: Evidence for Behavior Change: Routledge; 2002. [doi: 10.4324/9781410603029]
3. Plackett R, Kaushal A, Kassianos AP, et al. Use of social media to promote cancer screening and early diagnosis: scoping review. J Med Internet Res 2020 Nov 9;22(11):e21582. [doi: 10.2196/21582] [Medline: 33164907]
4. Khan Y, Tracey S, O'Sullivan T, Gournis E, Johnson I. Retiring the flip phones: exploring social media use for managing public health incidents. Disaster Med Public Health Prep 2019 Dec;13(5-6):859-867. [doi: 10.1017/dmp.2018.147] [Medline: 31218988]
5. Jang H, Rempel E, Roth D, Carenini G, Janjua NZ. Tracking COVID-19 discourse on twitter in North America: infodemiology study using topic modeling and aspect-based sentiment analysis. J Med Internet Res 2021 Feb 10;23(2):e25431. [doi: 10.2196/25431] [Medline: 33497352]
6. Carter M. How twitter may have helped Nigeria contain ebola. BMJ 2014 Nov 19;349:g6946. [doi: 10.1136/bmj.g6946] [Medline: 25410185]
7. Bartlett C, Wurtz R. Twitter and public health. J Public Health Manag Pract 2015;21(4):375-383. [doi: 10.1097/PHH.0000000000000041] [Medline: 24356087]
8. Xie J, Liu L. Identifying features of source and message that influence the retweeting of health information on social media during the COVID-19 pandemic. BMC Public Health 2022 Dec;22(1):805. [doi: 10.1186/s12889-022-13213-w]
9. Ciotti M, Ciccozzi M, Terrinoni A, Jiang WC, Wang CB, Bernardini S. The COVID-19 pandemic. Crit Rev Clin Lab Sci 2020 Sep;57(6):365-388. [doi: 10.1080/10408363.2020.1783198] [Medline: 32645276]
10. Al-Dmour H, Masa'deh R, Salman A, Abuhashesh M, Al-Dmour R. Influence of social media platforms on public health protection against the COVID-19 pandemic via the mediating effects of public health awareness and behavioral changes: integrated model. J Med Internet Res 2020 Aug 19;22(8):e19996. [doi: 10.2196/19996] [Medline: 32750004]

XSL•FO
RenderX

11. Gan CCR, Feng S, Feng H, et al. #WuhanDiary and #WuhanLockdown: gendered posting patterns and behaviours on Weibo during the COVID-19 pandemic. BMJ Glob Health 2022 Apr;7(4):e008149. [doi: 10.1136/bmjgh-2021-008149] [Medline: 35414567]
12. Kouzy R, Abi Jaoude J, Kraitem A, et al. Coronavirus goes viral: quantifying the COVID-19 misinformation epidemic on twitter. Cureus 2020 Mar 13;12(3):e7255. [doi: 10.7759/cureus.7255] [Medline: 32292669]
13. Llewellyn S. Covid-19: how to be careful with trust and expertise on social media. BMJ 2020 Mar 25;368:m1160. [doi: 10.1136/bmj.m1160] [Medline: 32213480]
14. Lenoir P, Moulahi B, Azé J, Bringay S, Mercier G, Carbonnel F. Raising awareness about cervical cancer using twitter: content analysis of the 2015 #SmearForSmear campaign. J Med Internet Res 2017 Oct 16;19(10):e344. [doi: 10.2196/jmir.8421] [Medline: 29038096]
15. Kummervold PE, Martin S, Dada S, et al. Categorizing vaccine confidence with a transformer-based machine learning model: analysis of nuances of vaccine sentiment in twitter discourse. JMIR Med Inform 2021 Oct 8;9(10):e29584. [doi: 10.2196/29584] [Medline: 34623312]
16. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv. Preprint posted online on Oct 2, 2019. [doi: 10.48550/arXiv.1910.01108]
17. Pérez JM, Rajngewerc M, Giudici JC, et al. Pysentimiento: a python toolkit for opinion mining and social NLP tasks. arXiv. Preprint posted online on Jun 17, 2021. [doi: 10.48550/arXiv.2106.09462]
18. Loureiro D, Barbieri F, Neves L, Espinosa Anke L, Camacho-collados J. TimeLMs: diachronic language models from twitter. Presented at: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics; May 22-27, 2022; Dublin, Ireland URL: https://aclanthology.org/2022.acl-demo [doi: 10.18653/v1/2022.acl-demo.25]
19. Huilgol P. Accuracy vs F1-score. Medium. 2019. URL: https://medium.com/analytics-vidhya/accuracy-vs-f1-score-6258237beca2 [accessed 2025-03-07]
20. Kundu R. F1 score in machine learning: intro & calculation. V7labs. 2022. URL: https://www.v7labs.com/blog/f1-score-guide [accessed 2025-03-07]
21. Cole-Lewis H, Pugatch J, Sanders A, et al. Social listening: a content analysis of e-cigarette discussions on twitter. J Med Internet Res 2015 Oct 27;17(10):e243. [doi: 10.2196/jmir.4969] [Medline: 26508089]
22. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005 Nov;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
23. Schopf T, Braun D, Matthes F. Semantic label representations with lbl2vec: a similarity-based approach for unsupervised text classification. In: Web Information Systems and Technologies: Springer, Cham; 2023:59-73. [doi: 10.1007/978-3-031-24197-0_4]
24. Eldallal A, Barbu E. BibRank: automatic keyphrase extraction platform using metadata. Information 2023;14(10):549. [doi: 10.3390/info14100549]
25. Gough A, Hunter RF, Ajao O, et al. Tweet for behavior change: using social media for the dissemination of public health messages. JMIR Public Health Surveill 2017 Mar 23;3(1):e14. [doi: 10.2196/publichealth.6313] [Medline: 28336503]
26. Semiz G, Berger PD. Determining the factors that drive twitter engagement-rates. ABR 2017;5(2) [FREE Full text] [doi: 10.14738/abr.52.2700]
27. Katie Sehl KM. Engagement rate calculator. IndiKit. 2024. URL: https://www.indikit.net/document/371-engagement-rate-calculator [accessed 2025-03-07]
28. Voorveld HAM, van Noort G, Muntinga DG, Bronner F. Engagement with social media and social media advertising: the differentiating role of platform type. J Advert 2018 Jan 2;47(1):38-54. [doi: 10.1080/00913367.2017.1405754]
29. Melton CA, White BM, Davis RL, Bednarczyk RA, Shaban-Nejad A. Fine-tuned sentiment analysis of COVID-19 vaccine-related social media data: comparative study. J Med Internet Res 2022 Oct 17;24(10):e40408. [doi: 10.2196/40408] [Medline: 36174192]
30. Latkin CA, Dayton L, Miller JR, et al. Behavioral and attitudinal correlates of trusted sources of COVID-19 vaccine information in the US. Behav Sci (Basel) 2021 Apr 20;11(4):56. [doi: 10.3390/bs11040056] [Medline: 33924118]
31. Jucks R, Thon FM. Better to have many opinions than one from an expert? Social validation by one trustworthy source versus the masses in online health forums. Comput Human Behav 2017 May;70:375-381. [doi: 10.1016/j.chb.2017.01.019]
32. Chen X. Online health communities influence people's health behaviors in the context of COVID-19. PLOS ONE 2023;18(4):e0282368. [doi: 10.1371/journal.pone.0282368] [Medline: 37053287]
33. Lustria MLA, Noar SM, Cortese J, Van Stee SK, Glueckauf RL, Lee J. A meta-analysis of web-delivered tailored health behavior change interventions. J Health Commun 2013;18(9):1039-1069. [doi: 10.1080/10810730.2013.768727] [Medline: 23750972]
34. Maibach EW. In: Parrott RL, editor. Designing Health Messages: Approaches from Communication Theory and Public Health Practice: SAGE Publication; 1995. [doi: 10.4135/9781452233451]
35. Lutkenhaus RO, Jansz J, Bouman MP. Tailoring in the digital era: Stimulating dialogues on health topics in collaboration with social media influencers. Digit Health 2019;5:2055207618821521. [doi: 10.1177/2055207618821521] [Medline: 30729023]

XSL•FO
RenderX

36. Xue J, Chen J, Hu R, et al. Twitter discussions and emotions about the COVID-19 pandemic: machine learning approach. J Med Internet Res 2020 Nov 25;22(11):e20550. [doi: 10.2196/20550] [Medline: 33119535]
37. Xue J, Chen J, Chen C, Zheng C, Li S, Zhu T. Public discourse and sentiment during the COVID 19 pandemic: Using Latent Dirichlet Allocation for topic modeling on Twitter. PLoS ONE 2020;15(9):e0239441. [doi: 10.1371/journal.pone.0239441] [Medline: 32976519]
38. Boon-Itt S, Skunkan Y. Public perception of the COVID-19 pandemic on Twitter: sentiment analysis and topic modeling study. JMIR Public Health Surveill 2020 Nov 11;6(4):e21978. [doi: 10.2196/21978] [Medline: 33108310]
39. Worrall AP, Kelly C, O'Neill A, et al. Online search trends influencing anticoagulation in patients with COVID-19: observational study. JMIR Form Res 2021 Aug 31;5(8):e21817. [doi: 10.2196/21817] [Medline: 34292865]
40. Liu S, Liu J. Understanding behavioral intentions toward COVID-19 vaccines: theory-based content analysis of tweets. J Med Internet Res 2021 May 12;23(5):e28118. [doi: 10.2196/28118] [Medline: 33939625]
41. Hussain A, Tahir A, Hussain Z, et al. Artificial intelligence-enabled analysis of public attitudes on Facebook and Twitter toward COVID-19 vaccines in the United Kingdom and the United States: observational study. J Med Internet Res 2021 Apr 5;23(4):e26627. [doi: 10.2196/26627] [Medline: 33724919]
42. Lyu JC, Han EL, Luli GK. COVID-19 vaccine-related discussion on twitter: topic modeling and sentiment analysis. J Med Internet Res 2021 Jun 29;23(6):e24435. [doi: 10.2196/24435] [Medline: 34115608]
43. Berg SH, O'Hara JK, Shortt MT, et al. Health authorities' health risk communication with the public during pandemics: a rapid scoping review. BMC Public Health 2021 Jul 15;21(1):1401. [doi: 10.1186/s12889-021-11468-3] [Medline: 34266403]
44. Pattison AB, Reinfelde M, Chang H, et al. Finding the facts in an infodemic: framing effective COVID-19 messages to connect people to authoritative content. BMJ Glob Health 2022 Feb;7(2):e007582. [doi: 10.1136/bmjgh-2021-007582] [Medline: 35131808]
45. Pakhtigian EL, Downs-Tepper H, Anson A, Pattanayak SK. COVID-19, public health messaging, and sanitation and hygiene practices in rural India. J Water Sanit Hyg Dev 2022 Nov 1;12(11):828-837. [doi: 10.2166/washdev.2022.134]
46. Officer TN, McKinlay E, Imlach F, Kennedy J, Churchward M, McBride-Henry K. Experiences of New Zealand public health messaging while in lockdown. Aust N Z J Public Health 2022 Dec;46(6):735-737. [doi: 10.1111/1753-6405.13297] [Medline: 36190208]
47. Abuhashesh MY, Al-Dmour H, Masa'deh R, et al. The role of social media in raising public health awareness during the pandemic COVID-19: an international comparative study. Informatics (MDPI) 2021;8(4):80. [doi: 10.3390/informatics8040080]
48. Karmegam D, Mapillairaju B. What people share about the COVID-19 outbreak on Twitter? An exploratory analysis. BMJ Health Care Inform 2020 Nov;27(3):e100133. [doi: 10.1136/bmjhci-2020-100133] [Medline: 33214193]
49. Slavik CE, Buttle C, Sturrock SL, Darlington JC, Yiannakoulias N. Examining tweet content and engagement of Canadian public health agencies and decision makers during COVID-19: mixed methods analysis. J Med Internet Res 2021 Mar 11;23(3):e24883. [doi: 10.2196/24883] [Medline: 33651705]

Abbreviations

API: application programming interface
ML: machine learning
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by A Mavragani; submitted 19.04.24; peer-reviewed by E Willis, R Chandrasekaran; revised version received 18.12.24; accepted 18.12.24; published 19.03.25.

Please cite as:
Parveen S, Pereira AG, Garzon-Orjuela N, McHugh P, Surendran A, Vornhagen H, Vellinga A
COVID-19 Public Health Communication on X (Formerly Twitter): Cross-Sectional Study of Message Type, Sentiment, and Source
JMIR Form Res 2025;9:e59687
URL: https://formative.jmir.org/2025/1/e59687
doi:10.2196/59687

© Sana Parveen, Agustin Garcia Pereira, Nathaly Garzon-Orjuela, Patricia McHugh, Aswathi Surendran, Heike Vornhagen, Akke Vellinga. Originally published in JMIR Formative Research (https://formative.jmir.org), 19.3.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Food Access in New York City During the COVID-19 Pandemic: Social Media Monitoring Study

Leah Butz[1,2], BA; Charles Platkin[1,2], MPH, JD, PhD; Jonathan Chin[1,2], MA, MFA; Juan Pablo Chavez Salas[2], MPhil, RDN; Ellie Serres[2], BSc; May May Leung[1,2], RDN, PhD

[1]Hunter College NYC Food Policy Center, Hunter College, New York, NY, United States

[2]Hunter College, New York, NY, United States

Corresponding Author:
Leah Butz, BA
Hunter College NYC Food Policy Center
Hunter College
2180 Third Avenue Ste 508
New York, NY, 10035
United States
Phone: 1 212 396 7744
Email: leahbutz@gmail.com

Abstract

Background: The COVID-19 pandemic exacerbated issues of poverty and food insecurity in New York City, and many residents experienced difficulty accessing available resources to help them get food on the table. Social media presents an opportunity to observe and understand the barriers people face in accessing affordable, nutritious, and culturally appropriate foods.

Objective: This study aims to explore the food access discourse during the COVID-19 pandemic on Twitter (subsequently rebranded as X) in New York City by analyzing publicly available tweets posted from March 1, 2020, to March 31, 2021.

Methods: Tweets posted by individuals in New York City during the first 13 months of the COVID-19 pandemic were collected using the observation platform Talkwalker. We categorized a list of multiple keywords into related groups (search strings). Data were cleaned to keep only tweets relevant to food insecurity and food access in New York City and remove duplicate tweets. The software Botometer was used to remove accounts considered to be bots. Topic modeling was used to group these tweets into relevant themes, which were analyzed. The top viral tweets (ie, tweets that received the highest number of retweets in New York City) from this period were further analyzed.

Results: We identified 6 major themes (with subthemes) that emerged from the analysis (in order of popularity): community efforts, public assistance programs, grocery shopping and food workers, school foods, millions go hungry, and food justice. Interesting terms that emerged from the data were also identified. Overall, quantities of tweets increased in correlation with current events, such as the closure of New York City public schools; the expansion of the Supplemental Nutrition Assistance Program and unemployment benefits; the proliferation of mutual aid groups in the spring of 2020; and the May Day Instacart, Amazon, and Target strike in 2020.

Conclusions: Findings revealed that in the earliest months of the COVID-19 pandemic, Twitter users in New York City quickly responded to the wave of need by sharing information and resources about food access in their communities. Some users turned to Twitter to either solicit or offer help finding food. Furthermore, the platform lent itself to many conversations about the policies enacted on a federal, state, and city level to help feed New Yorkers in need. Future research on this topic should include an analysis of social media posting on platforms such as Facebook, as well as languages other than English. Results from this type of research can provide information to community leaders and elected officials to better address future crises.

(JMIR Form Res 2025;9:e49520) doi:10.2196/49520

KEYWORDS

Twitter; food access; food insecurity; New York City; topic modeling; social media; artificial intelligence

XSL•FO
RenderX

Introduction

Background

The COVID-19 pandemic was one of the causes of a global food emergency that only intensified an already fragile food environment for many of our communities' most susceptible members [1]. Food banks and soup kitchens saw unprecedented demand, and enrollment in federal assistance programs such as the Supplemental Nutrition Assistance Program (SNAP) and the Supplemental Nutrition Program for Women, Infants, and Children increased [1]. Food supply also decreased dramatically, due in part to new difficulties in transportation and storage [2], all the while emergency food relief providers were faced with increasing operational costs [3]. These struggles disproportionately impacted low-income communities, who were more likely to lose income and reduce spending on food [1,4]. Malnutrition (including both undernutrition and overnutrition) can increase an individual's risk of chronic diet-related diseases, which in turn can increase the risk of COVID-19 and corresponding adverse health outcomes [5].

New York City specifically saw a major food crisis during the COVID-19 pandemic [6]. Food insecurity and lack of access to affordable, healthy foods persist and exert a disproportionate burden on low-income neighborhoods and racial and ethnic minority communities in New York City. The onset of the COVID-19 pandemic only exacerbated these issues as New York City officials reported that "2 million or more" residents faced food insecurity in May 2020 [7]. This was especially acute among Black people, Indigenous people, and other racial and ethnic minorities [8]. While New York City worked quickly to feed hungry residents, existing infrastructure was not equipped to meet this overwhelming need, and many residents still struggled for basic needs [9,10]. Many food-insecure households in New York City further experienced instability and insecurity in their jobs and health care at the same time.

It can be difficult to qualitatively analyze the experiences and opinions of individuals struggling with food access issues contemporary to the experiences themselves. However, social media (Twitter [subsequently rebranded as X] specifically) presents an opportunity to gather large amounts of real-time data from its users, as it can be used as a means of instant communication and information sharing. This allows researchers to obtain large amounts of data and insight into the opinions held by individuals during a discrete period. According to a survey of Twitter users conducted by Pew Research Center in 2018, individuals of all races and genders use the platform; however, users tend to be "younger, more educated and more likely to be Democrats than general public" [11]. Furthermore, Pew noted that 80% of all tweets originate from only 10% of all users, as the median Twitter user posts only about twice per month [11]. Previous research has shown that Twitter is frequently used to share and disseminate information to followers, while acknowledging the potential for the platform to be used to solicit information and assistance [12,13].

Numerous studies have attempted to track the posts made on Twitter that occurred during the COVID-19 pandemic, and some researchers have determined that social media could be a useful tool to observe and analyze the knowledge and opinions held by people during a public health emergency [14-16]. However, there is minimal social media–related research specific to food-related topics, presenting an opportunity to understand the barriers people face accessing affordable, nutritious, culturally appropriate foods and point those in need toward resources available to them [17,18]. Social media monitoring could be a useful asset to specifically understand the food environments and needs of individuals and families during such a time and potentially connect these people with resources available in their communities in real time.

Objective

This study aims to explore the food access discourse during the COVID-19 pandemic on Twitter in New York City by analyzing publicly available tweets posted from March 1, 2020, to March 31, 2021.

Methods

Data Collection

Tweet collection began with determining keywords to search for, which was informed by the literature [14,19,20]. Expert knowledge informed initial keywords to search using the social media observation platform Talkwalker; this platform generates word clouds and lists sample posts containing keywords. These outputs were reviewed to search for and identify additional keywords using a snowball method, adding or removing keywords to include in our search depending on how relevant the related outputs were. Key phrases were searched in both English and Spanish. (Talkwalker has a proprietary artificial intelligence model used to determine a post's language.) Slang keywords were added where applicable. For example, the term *freedge*, a portmanteau of *free* and *fridge*, was added to our list of search terms when use of this term was discovered in conversations about community fridges. Some terms were removed when it was determined that they would return too many irrelevant tweets. For example, in searching for tweets that talked about SNAP, called "SNAP" (uppercase), the term "snap" (lowercase) was removed after reviewing a sample of 100 tweets containing "snap" (lowercase) and determining none of them referred to SNAP.

These preliminary searches eventually resulted in lists of multiple keywords that were categorized into related subgroups, as informed by Kim et al [19]. These related subgroups are called *search strings* because similar key phrases were strung together with Boolean operators when conducting the search and collecting data using Talkwalker. A list of the search strings and the Boolean operators used to conduct these searches can be found in Multimedia Appendix 1. When searching using social media, there must be a balance between 2 criteria: "retrieval precision (how much of retrieved data are relevant) and retrieval recall (how much of the relevant data are retrieved)," as noted by Kim et al [19]. For the purposes of initial searches, retrieval recall was prioritized. Therefore, more tweets (and consequently more irrelevant tweets) were collected, and irrelevant tweets (also called *noise*) were removed during the data cleaning process. Tweets collected were posted between March 1, 2020, and March 31, 2021.

XSL•FO
RenderX

Data Cleaning

Before analysis could occur, the data had to be cleaned to ensure all tweets were relevant to the study purpose. Tweets were deemed relevant to the study if they related to food insecurity and food access in New York City. Talkwalker has a multitiered process to determine the geolocation of tweets. First, Talkwalker will check their own internal database of users and authors and their locations. If the user is not listed there, they try to obtain location data from Twitter application programming interface if provided in the form of geographical coordinates. If these 2 steps yield no results, then it will check the user's biographical location data. Finally, if still no location data can be found, then it will use other context clues (such as the username or the language) and place the tweet in the most likely country of origin's capital (Talkwalker Product Support Representative, email, August 2024).

At least 50% of the time spent on this study was used to clean the data. Spam, porn, and irrelevant tweets were removed in a rigorous data cleaning process. Talkwalker has a feature that detects and categorizes spam tweets based on the type of spam the tweet is; any tweets categorized as "Social Media Account Promotions," "Promotions," "Job Offers," and "Real Estate" were removed. Talkwalker has another feature that rates the "porn level" of any given tweet; tweets with a "porn level" higher than 0 were removed after reviewing a random sample of tweets with porn levels higher than 0 and finding none of the tweets to be relevant. Furthermore, posts with a porn level higher than 0 made up a negligible amount of less than 2%.

The data cleaning process included filtering out tweets that contained certain key phrases and hashtags not associated with food access and food insecurity. For example, in searching for tweets that talked about the SNAP, many tweets were retrieved about Snapchat (a social media company whose stock exchange ticker symbol is $SNAP), so all tweets that contained exactly the key phrases *$SNAP* and *Snapchat* were removed. The list of key phrases to remove was determined through a systematic review of a random sample of 100 tweets and adding irrelevant key phrases to a filter that removed them. A random sample of 100 tweets was continually generated and reviewed until at least 90% of the tweets in the sample were considered relevant to the purposes of the study.

During the data cleaning process, the search strings "Healthy Foods (En) & (Sp)" and "Fast Food" from the analysis were removed after random samples of the retrieved tweets were determined to be irrelevant to the purpose of the study. Furthermore, tweets retrieved by the "Community Fridges" and "Community Gardens" search strings were grouped together for analysis into one category, called "community efforts," because these datasets comprised tweets with similar content. Finally, 3 search strings were "reverse filtered": tweets that did not contain specific key phrases were removed. "Food Justice" was reverse filtered to only include tweets that contained at least one of the following phrases: *food*, *nutrition*, *meal*, *meals*, *nutritional*, *diet*, or *dietary*. "School Food" was reverse filtered to only include tweets that contained at least one of the following phrases: *food*, *nutrition*, *meal*, *meals*, *nutritional*, *diet*, *dietary*, *lunch*, *lunches*, *breakfast*, *breakfasts*, *eat*, *feed*, *feeds*, or *eats*. Finally, "Food Insecurity" was reverse filtered to only include tweets that contained the phrase *food insecurity*. Figure 1 depicts the number of tweets retrieved by the initial search, the number of tweets remaining after data cleaning, and the number of unique tweets that were analyzed.

Figure 1. Flowchart of dataset cleaning process for collected tweets dated March 1, 2020, through March 31, 2021, in New York City.

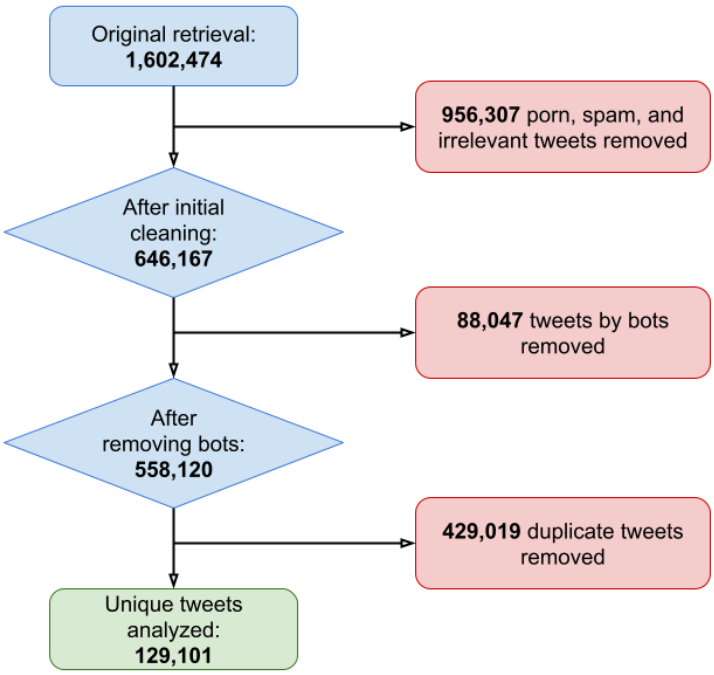

Furthermore, tweets generated by bots ("partially or fully automated accounts on social media platforms," according to Martini et al [21]) were removed using an application programming interface called Botometer. Botometer is a tool that "checks the activity of a Twitter account and gives it a score based on the extent to which it matches accounts that use

XSL•FO
RenderX

automation" [22]. In other words, Botometer provides a score to a Twitter account based on how likely it is that the account is a bot; higher scores indicate a higher likelihood that the account is automated and not run by a human. Botometer was chosen to help determine which accounts were bots or not because it is the predominant tool used in the literature [11,23,24].

Botometer was used to score every Twitter account attached to the retrieved tweets (>600,000 accounts). Because Botometer does not decide whether a given account is a bot, instead only providing a score as to the likelihood that the account is a bot, a cutoff score to sort bots from nonbots had to be determined. Botometer provides 2 scores to each account checked: *raw* and *complete automation probability*. The raw score is the score that Botometer has assigned in terms of likelihood that an account is a bot. Higher raw scores mean a higher likelihood that an account is a bot, and lower raw scores mean a lower likelihood that an account is a bot. However, Botometer notes that raw scores "don't provide enough information by themselves to classify an account." Therefore, the complete automation probability score can be used to ascertain the "probability...that an account with this score or greater is controlled by software, i.e., is a bot." For the purposes of this report, score refers to the raw score provided by Botometer. Various methods were considered to determine the Botometer cutoff score. A 2018 paper from the Pew Research Center analyzing bots used a Botometer cutoff score of 0.43 [11]. In addition, 3 research assistants were given a random sample of 150 accounts to decide "bot or not" based on attributes such as profile appearance, quantity of tweets and retweets, frequency of posting, and interactions with other users [25]. Interrater reliability was measured using Light's implementation of the κ coefficient [26]. Furthermore, Amazon Mechanical Turk (MTurk) was used to crowdsource bot distinction for a random sample of 1366 accounts. However, due to various limitations to using MTurk (such as the inability to control for variation in graders by language proficiency or establish a shared definition for what qualifies as a bot) [27-29], ultimately the MTurk decisions were discarded and the research assistants' determinations were used to test the Botometer thresholds. After exploration of both the research assistants' cutoff score and the Pew Research Center paper cutoff score, and taking into consideration the performance metrics of specificity, sensitivity, accuracy, precision, and recall, a cutoff score of 0.43 was used to eliminate bot accounts, which removed approximately 11% of accounts.

Estimations of how many accounts on Twitter are bots were considered to justify the cutoff point. Each year, Twitter files a report with the US Securities and Exchange Commission noting how many active accounts on the website it considers to be automated, alleging about 5% of accounts are bots [30]. However, many researchers believe a more accurate estimate is 9% to 20% of all accounts [25,31]. When research assistants manually annotated accounts, they determined that approximately 11% of accounts were bots, roughly aligning with the literature.

After removing irrelevant tweets, tweets by bots, and duplicate tweets (refer to the Topic Modeling Analysis section for reasoning regarding the removal of duplicate tweets), 8.06% of our initial retrieval remained. As depicted in Figure 1, the search strings initially retrieved 1,602,474 tweets. Of these tweets, 956,307 tweets classified as porn, spam, or irrelevant to the study were removed; 88,047 tweets authored by bots were removed; and 429,019 duplicate tweets were removed. This left 129,101 unique, relevant tweets in the dataset for analysis.

Before topic modeling could occur, the tweets had to be further cleaned to optimize the results. Contractions (such as *I'm*, *it's*, and *we're*) were expanded (such as *I am*, *it is*, and *we are*). Punctuation and "stop words"—very common words in a particular language with little semantic value, such as *a*, *of*, and *to* [32]—were removed. The full list of stop words removed came from Natural Language Toolkit, the leading open-source natural language processing platform [33]. Emojis, hashtags, and tagged usernames (preceded by the @ symbol) were not removed.

Topic Modeling Analysis

Topic modeling is a type of machine learning technique commonly used to evaluate data coming from a variety of text sources or documents to determine the words and features that the text sources or documents share [34,35]. A topic model scans a set of text sources or documents and groups them into *clusters* (sometimes called *topics*) based on shared words and phrases (called *keywords*).

A topic modeling technique called Gibbs Sampling Dirichlet Multinomial Mixture (GSDMM) was used because it is optimal for short texts and has been used in the literature [15,16,36-38]. Specifically, a GSDMM topic model publicly available on the software development sharing platform GitHub was used [39].

In addition, after much discussion and consideration, duplicate tweets and retweets from the topic modeling analysis were removed from the analysis. This resulted in a dataset wherein every tweet is distinct. This decision was made after some preliminary practice topic modeling was run and resulted in clusters entirely made up of a single tweet that had been retweeted several times. To better see the connections between the various *distinct* tweets, these retweets were removed. This left behind only one copy of each tweet in the dataset.

GSDMM topic modeling was then run on the dataset with the following parameters: α and (parameters that affect the likeliness of tweets to sort into different clusters) set at .1, as recommended by Yin and Wang [37]. The number of iterations was set to 40, as Yin and Wang [37] noted that GSDMM does not require a high number of iterations for the model to reach convergence (ie, for the model to sort tweets into clusters with a high level of confidence). Each iteration is slightly different and more refined than the previous; iteration 40 is much more specific than iteration 1, but not significantly better than iteration 39.

Running GSDMM topic modeling resulted in a model unique to the dataset. This model revealed 50 clusters and their associated keywords. However, this model did not indicate the cluster of each individual tweet in the dataset, so then the model was used again on each tweet individually, producing a list of probabilities of that tweet belonging in each cluster. The tweet

XSL•FO
RenderX

was then assigned to the cluster with the highest probability. The code used in the implementation of data cleaning and topic modeling can be found in a designated GitHub repository [40].

A qualitative content analysis of the clusters was then conducted. Informed by Ridhwan and Hargreaves [15], clusters that contained similar tweets and conversations were grouped into themes. This resulted in the emergence of 6 major themes. Each cluster was assigned to a theme. Only the top 30 most populated clusters were analyzed. Clusters determined to not be relevant to the study's purpose were not assigned to a theme. For example, there was a cluster composed entirely of tweets quoting the Langston Hughes poem "Harlem," which contains the line "Does it stink like rotten meat?" [41]. This cluster was eliminated from analysis. Furthermore, none of the clusters analyzed had any Spanish keywords, and the tweets analyzed were almost exclusively in English. After sorting the relevant clusters into their respective themes, the tweets were reviewed again to pull out significant subthemes within each major theme.

Viral Tweets Analysis

While removing retweets was essential in ensuring the topic modeling clusters were not subsumed by viral tweets, excluding retweets from analysis altogether seemed like it would not lead to an accurate reflection of the conversations about food security and food access that took place on the platform. Thus, the content of and conversations instigated by certain viral tweets in New York City were analyzed. To do this, the top 10 most duplicated tweets (ie, the tweets with the most retweets occurring in New York City) were pulled from the dataset before retweets were removed. The replies and quote tweets posted on each viral tweet were then analyzed. Similar to the analysis of the clusters in topic modeling, in reviewing the replies and quote tweets, certain themes that characterized the conversations spurred by these tweets emerged.

Interesting Terminology

During the review of the dataset, certain noteworthy terms and phrases were recognized. This included slang terms and words used in a nontraditional sense. These terms were pulled and investigated further for their use and potential implications on future social media research.

Ethical Considerations

This study involved observations of large swaths of publicly available data, generally not considered to be "human subjects research." Approval from an institutional research board was not needed. The proposed research did not require CUNY HRPP (City University of New York Human Research Protection Program) or Institutional Review Board review as we were not engaged in research involving human subjects, in accordance with CUNY guidelines [42]. All data were anonymized before analysis and publication. No compensation was provided to the individuals who wrote the tweets that were analyzed.

Results

Topic Modeling Analysis

Multimedia Appendix 2 shows the 6 themes and related subthemes that emerged from GSDMM modeling. Further information for each theme is also included, such as common words or roots and example tweets.

Figure 2 shows the spread of tweets over the 13-month study period and the relative quantities of tweets in each of the 6 themes. Across all themes, there was a large spike in tweets in March and April 2020, which began to taper off during May 2020. The quantity of tweets stabilized from June through October 2020, experiencing relatively smaller spikes in November and December 2020, corresponding with the increased number of COVID-19 infections during this period.

Figure 2. Distribution of tweets per theme by week from March 1, 2020, to March 31, 2021, shown by the relative number of tweets posted that fall within each theme.

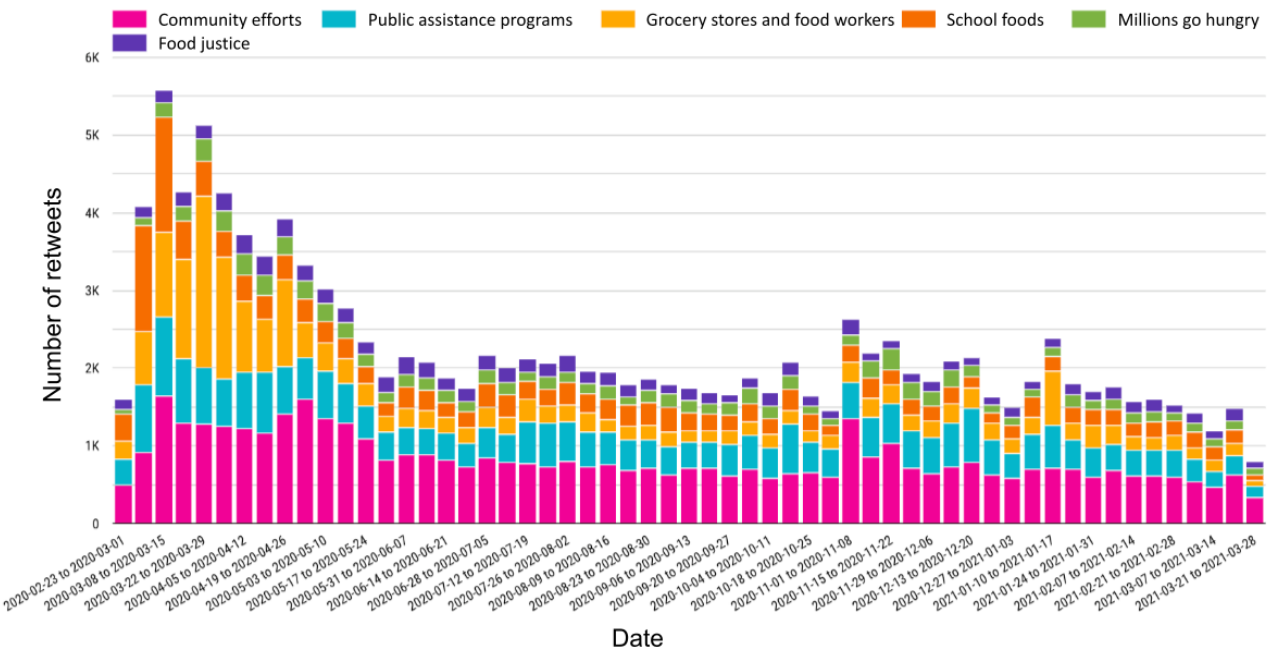

XSL•FO
RenderX

Overall, the theme containing the most tweets was "community efforts," as noted in Multimedia Appendix 2. This theme contained tweets related to community-based organizations providing assistance and relief to New Yorkers in need. As shown in Figure 2, the quantity of tweets related to this theme increased quickly in March 2020 and then began to taper off toward the end of May 2020. During this period, community efforts (such as community fridges, mutual aid organizations, and other nonprofits) quickly scaled up to meet the needs of New Yorkers. Another spike in tweets related to this theme occurred in the middle of November 2020, ahead of the holiday season and amid increased COVID-19 infection rates.

The second most populated theme was "public assistance programs," which included any tweets related to programs such as SNAP and pandemic electronic benefit transfer (P-EBT). Figure 2 shows that these conversations increased quickly in the early months of the pandemic, when SNAP benefits were temporarily increased for all recipients after passage by the US Congress of the Coronavirus Aid, Relief, and Economic Security Act in April 2020 [43], but tweets tapered off by June 2020. Minor spikes can be seen in July 2020, when the P-EBT was expanded to all public school students in New York City [44], and in December 2020, during the holiday season. Twitter users also worried about electronic benefit transfer (EBT) cards not having sufficient funds to purchase food, stores being unable to process EBT transactions, and the purchasing decisions made by people receiving SNAP benefits.

The third most populated theme was "grocery stores and food workers." Tweets belonging to this theme pertained to shopping habits during the pandemic and protections for people working in the food industry during the pandemic. Some users posted about the benefit of web-based grocery delivery services (such as Instacart and Amazon Fresh) accepting SNAP [45,46]. This also included a very small number of tweets regarding the safety of products at the grocery store and a misinformation theory that certain packaged frozen foods were contaminated with COVID-19. This theme saw the highest number of tweets in April and May 2020, as there was a nationwide strike by grocery store and other food workers on May 1, 2020. Furthermore, this theme saw a smaller spike in January 2021, when Teamsters at the Hunts Point Produce Market went on strike [47].

The fourth most populated theme was "school foods." This theme contains tweets about how students were going to access meals without in-person schooling and discussions about P-EBT or grab-and-go meals. This theme saw a large spike in March 2020, when schools were closed in New York City, and smaller spikes in July 2020 and September 2020, other moments of uncertainty about students' futures.

The fifth theme with the most tweets was titled "millions go hungry." This theme is characterized by tweets containing information about the pandemic's effects on food insecurity but on a broader scale (ie, not specific to New York City). Users commonly tweeted about the "millions" of households that will experience food insecurity because of the COVID-19 pandemic, drawing attention to the inadequacy of existing infrastructures (such as food pantries, benefit programs, and charities) to meet the expanded demand. This theme was fairly consistent throughout the period analyzed.

The sixth and final theme that emerged from the data was "food justice." Tweets in this theme related to the relationship between social justice or progressive ideology and food. This theme saw spikes from May through August 2020, in the wake of nationwide Black Lives Matter protests spurred by the murder of George Floyd. This theme also saw a smaller spike in November 2020, during the days leading up to the November 3, 2020, federal election.

Analysis of Viral Tweets

Multimedia Appendix 3 shows the top 10 viral tweets pulled from the dataset along with the number of likes and retweets they each received, as well as a brief summary of the types of conversations found in the replies to these tweets. Top viral tweets received 6946 to 68,784 retweets and 24,794 to 298,839 likes. Most tweets came from users with follower counts in the thousands to millions, but 1 viral tweet about a major worker strike in the food industry was posted by a user with only 956 followers at the time.

Figure 3 shows the top 10 viral tweets over time, mapped to the timelines when they went viral. The content of these tweets was analyzed relative to contemporary news and current events.

Figure 3. Viral tweets over time by number of retweets, with added context about current events occurring at similar times to the tweets. NY: New York; NYC: New York City.

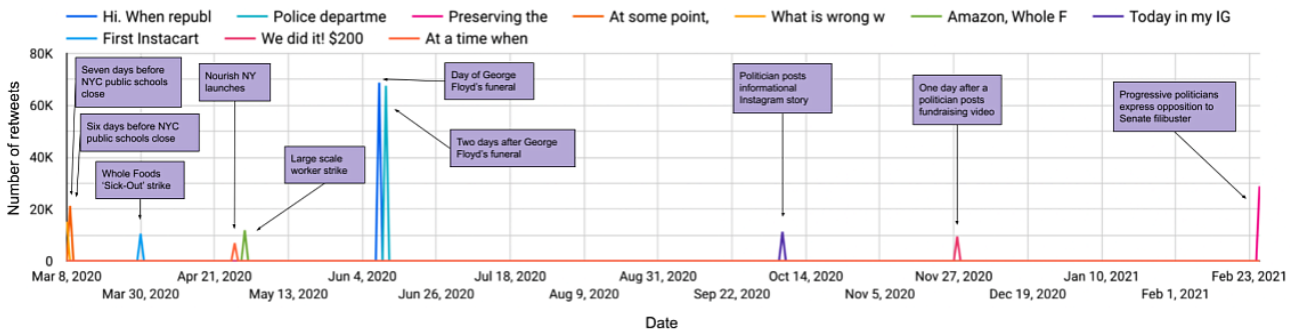

The first 2 tweets to go viral in the dataset occurred in early March 2020, a few days before Mayor Bill de Blasio announced the closure of New York City public schools as a precaution against COVID-19 [48]. Conversations instigated by these tweets primarily related to students from low-income families and concerns that they might not be able to have reliable meals without in-person schooling.

XSL•FO
RenderX

Two more top viral tweets related to large-scale food worker strikes that occurred during the early months of the pandemic. On March 31, 2020, workers at Whole Foods staged a nationwide "sick-out" strike (a strike wherein employees call out from work sick) to demand COVID-19 precautions [49]. A little over a month later, on May 1, 2020, workers for Amazon, Instacart, Whole Foods, and others staged a strike to protest the lack of protections and benefits to frontline workers [50].

Many viral tweets were by or about politicians and related to food access and food policy. Tweets of this kind instigated conversations about individuals' political leanings and support. Furthermore, these tweets tended to garner more disagreement within the replies and quote tweets. One tweet by the same politician referred to a push by Democrats in February 2021 to eliminate the Senate filibuster [51]. Another referred to the passing of a state initiative to reroute surplus agricultural output to food banks in the state, called Nourish New York.

Another tweet by a politician related to the Black Panther Breakfast Program [52], which led to discussions about free and reduced-price school meals in the 21st century and the relationship between food and racial justice. Two viral tweets occurred during the early June 2020 Black Lives Matter protests, especially in the days surrounding George Floyd's funeral. These tweets prompted conversations about police brutality, one of which directly compared police budgets to school budgets and the affordability of school food.

Interesting Terminology

In reviewing the tweets, certain interesting terms emerged from the data that could help guide social media studies and interventions in the future.

- "School lunch debt" is a phenomenon wherein students who are unable to pay for their daily school lunches and accrue a balance they owe to the school. Some schools would cease to provide lunch to students whose "lunch debts" were too high; example tweet: "Please explain f—school lunch debt. Why are 7 year olds in debt over chicken nuggets."
- A "struggle meal" refers to a meal that someone would eat during times when money is tight; example tweet: "Facts, my struggle meal was one slice of bread with butter. Or if I was lucky chopped hot dogs w. Ketchup. And now that's basically my snack."
- People in need of food assistance sometimes turned to Twitter to ask other users for help, usually through gifts of cash. These types of posts varied quite a bit in the language used, but almost all of them included a link to a CashApp, PayPal, or Venmo account; example tweet: "The only thing that could possibly be good about this Xmas would be if you see my entering 1000x for cashapp blessings and helping me so I don't lose my home. [username] Unemployed, no money, can't pay rent, get food, will be homeless. Been in survival mode too long. Breaking."
- With the proliferation of community fridges during the pandemic came new terminology related to them, including the word "freedge"—a portmanteau of "free" and "fridge"; example tweet: "In my neighborhood in NYC there is a Freedge—a free refrigerator where we can all put food for others to take as they wish."
- In the context of government assistance, critics of the support used the term "freeloading" often to point toward people who were taking advantage of the benefits for a long time or without an effort to get off them; example tweet: "I know so many women that have been freeloading off government assistance and bragging about their benefits for years."

Discussion

Principal Findings

This study analyzed tweets during the first 13 months of the COVID-19 pandemic related to food security and food access in New York City. The themes that emerged from the data include topics related to food justice, food policy, food workers, and food relief efforts. Overall, quantities of tweets increased in correlation with current events, such as the closure of New York City public schools [53]; the expansion of SNAP and unemployment benefits [43,54]; the proliferation of mutual aid groups in April and June of 2020 [55]; and the May Day Instacart, Amazon, and Target strike in 2020 [56]. The study highlighted some of the common opinions held, emotions felt, and phrases used by New Yorkers from March 2020 to March 2021 around these issues.

In the earliest months of the pandemic (March, April, and May 2020), many tweets were posted about food access and community responses to the COVID-19 pandemic. This included changes in shopping habits and increased awareness of food resources such as SNAP benefits, food pantries, and community fridges. Due to stay-at-home orders and worries about catching COVID-19 from shoppers at the grocery store, web-based grocery shopping became increasingly popular during this time [57,58]. However, with this increased demand for web-based shopping came delayed grocery deliveries and errors in grocery orders, about which Twitter users frequently complained. Other users posted about the opening of new grocery stores in their communities and the expansion of grocery delivery areas.

During the early months of the pandemic, many individuals saw food shortages [59,60], but some users noted that Spanish, Mexican, and Chinese grocery stores had many products in stock while more mainstream grocery store chains had empty shelves. Larger food distributors, including large-scale meat processing plants and fruit and vegetable production facilities [61,62], were especially susceptible to pandemic-caused closures and service interruptions [63,64]. It is possible that some of these stores rely on more direct supply chains that were not as heavily impacted by pandemic closures, thus allowing the store to continue purchasing inventory directly from farmers and producers without experiencing as many interruptions as larger grocery store chains [65,66]. Furthermore, another possible reason that smaller ethnic grocery stores were more likely to have products on the shelves than larger supermarkets is due to the stockpiling that many individuals participated in during the early months of the pandemic [67]. Someone interested in stockpiling large amounts of product is less likely to shop at a

XSL•FO
RenderX

small ethnic market, rather they would go to a large, multidepartment store that carries a larger variety of products.

Government officials also became involved in Twitter conversations, sharing information about new policies and resources available to New Yorkers. For example, in the early months of the pandemic, dairy farms were forced to dump old milk as the lockdown restricted the market for dairy products [68]. In response to this, the Governor announced a new policy to combat food waste and food shortages simultaneously, called Nourish New York [69]. The Nourish New York initiative "reroutes surplus agricultural products to populations who need them most through NY's network of food banks," according to the New York State Department of Agriculture and Markets website [69]. Farmers and other food producers in New York State share information about surplus food products and are connected to food banks that purchase these items to distribute to people in need. The initiative was highly successful and has since become permanent [70]. The supply chain disruptions and food shortages caused by the pandemic, especially in large-scale plants and facilities [63,64], showed a need to strengthen regional food chains to create a more resilient system [63].

Conversations were often politicized: some users talked about fears that Republican politicians sought to cut SNAP budgets or make enrolling in the program more difficult. SNAP benefits are already often insufficient to meet a household's full food needs [71-73]. In fact, toward the end of each month, food pantries have reported increased need from clients, many of whose SNAP benefits have likely run out [74]. This phenomenon was also seen in the data: some Twitter users in need posted about running out of money on their EBT cards and needing assistance toward the end of the month, asking others who saw the posts to help them out with gifts of cash. The need for extra money to purchase food has persisted beyond the first year of the pandemic; inflation [75] and the end of expanded SNAP benefits in 2023 [76] have shown continual barriers for people needing assistance acquiring food. There are clear implications on potential policy of these obstacles to meeting basic needs faced by many of the country's most susceptible populations.

There were also numerous debates about what kinds of foods SNAP beneficiaries should be allowed to purchase with their benefits; in November 2020, a celebrity posted a tweet about using EBT cards to purchase only healthy items, drawing criticism from users who wrote that no one should be allowed to police what others eat. Currently, SNAP recipients can purchase any food items for the household, such as fruits, vegetables, meats, dairy, grains, and "accessory foods" [77]—this includes cooking supplements such as oil, baking powder, and sugar, as well as snacks, desserts, and packaged beverages [78]. However, hot food items are generally not SNAP eligible. In 2021, the Restaurant Meals Program, a state option to allow certain populations to purchase hot food with SNAP benefits, was passed [79]; as of publication, the program is not available in all parts of New York State and is in the process of rolling out in New York City. Municipalities not opting into this program can leave many unhoused, disabled, and older adult New Yorkers who are unable to prepare or cook food with another barrier to accessing nutritious, warm meals.

Negative perception of the choices made by SNAP beneficiaries regarding what foods they choose to purchase using benefits is ubiquitous, resulting in a stigmatization of the program that often dissuades eligible people from enrolling [80,81]. It also feeds into the detrimental misconception that people receiving benefits are "freeloading," one of the terms found in the data. However, these arguments that SNAP beneficiaries should not be allowed to purchase "junk food" or other *unhealthy* items fail to recognize the food environment of many people enrolled in the program. Many underresourced communities with higher rates of poverty have fewer full-service grocery stores and more bodegas than affluent communities, so individuals in these communities are less likely to shop at supermarkets with larger inventories of "healthy foods" [82-85]. Furthermore, *healthy foods* are often perceived as more expensive than *unhealthy foods*, another barrier to healthy eating [86-89]. Consider the term "struggle meal" as found in the data: many of these struggle meals were described as quite calorically dense and comprised processed or prepackaged foods. Lack of access to affordable healthy foods contributes to the higher rates of diet-related diseases, such as diabetes and obesity, seen in underresourced communities.

School foods and getting meals to children was a top concern for Twitter users, especially when public schools closed in March 2020, during the summer vacation of 2020, and when schools were supposed to reopen in September 2020. Many users lamented how for some students, their free school breakfast and lunch was the only guaranteed meal they had each day. The P-EBT program was significant in helping families with children to acquire food [90,91]. Because New York City has universal free school meals, all families with children enrolled in public school received preloaded P-EBT cards, regardless of need. Automatic enrollment in the P-EBT program proved to be an effective way to ensure that all children in need received the benefit and positive health impacts associated [92]. Furthermore, some well-off families expressed interest in donating their surplus P-EBT funds to others in need [93,94]. P-EBT is also a more cost-effective measure for school districts than grab-and-go meals, another program meant to replace the meals lost by pandemic school closures [95]. However, many schools faced administrative obstacles in implementing the P-EBT program, leaving many families without benefits for as much as half the school year [96].

Some users also mentioned a phenomenon called "school lunch debt," wherein students are denied breakfast or lunch at school because they owe money for past lunches [97,98]. Some Twitter users, including 1 viral tweet with tens of thousands of likes and retweets, posted about the Black Panther Party's Breakfast Program of the 1960s and 70s [99] as a reference point for free school meals. Meals for schoolchildren is a particularly salient topic because of the various harmful effects of food insecurity on children's health and development [100-102]. Research has shown that food insecurity during childhood is associated with health, emotional, behavioral, and academic problems that might persist into adulthood [100-102]. Children from food-insecure households are also more likely to be suspended or expelled from school [103]; increased rates of suspension and expulsion, in turn, are associated with the school-to-prison pipeline

XSL•FO
RenderX

[104,105]. Future research should be explored into the relationship between food insecurity in a community and the school-to-prison pipeline.

Viral tweet analysis offers an interesting look into what posts and ideas gained the most traction and public discourse. While tweets from ordinary users without large follower counts provide an important glimpse into the opinions and perspectives of residents during this time, viral tweets are, by their nature, seen by far more users and have the potential to contribute more heavily to the public understanding of the topics they discuss [106]. Therefore, users with large followings are better able to expose other Twitter users to important and underdiscussed topics including urban food insecurity, labor activism, and nutrition assistance programs.

There is often a gap between individuals in need of nutrition assistance and organizations, programs, and services that exist to provide this necessary assistance. Often, individuals most in need of help are unable to access the resources available within their communities, whether due to limited hours or availability, lack of awareness, or stigma [107-109]. Therefore, social media provides a unique platform to quickly connect individuals and families in need with support from their own neighborhoods, as well as normalize and destigmatize the use of these services. By combining location data with an understanding of keywords and terminology used by people in need, service providers have the potential to intervene and offer their services more quickly and efficiently to their target populations. Furthermore, by understanding what makes certain posts go viral, hunger activists and food relief service providers might better be able to spread awareness of their efforts within affected communities. In addition, given the range of the authors of the viral tweets—politicians, academics, labor union activists, nutritionists and personal trainers, political analysts, and fashion journalists—leveraging collaborations with influencers across various sectors could further enhance the reach, delivery, and impact of critical services.

Limitations

While this study successfully analyzed tweets about food insecurity and food access in New York City during the COVID-19 pandemic, there were certain limitations. First, in cleaning the tweets of bots, it was learned that bot detection is not an exact science. Some tweets might have been flagged as bots even if the users were real, and some tweets by bots might have evaded detection and were kept in the dataset. Second, Talkwalker, the platform used to pull tweets, was not good at reading images, despite the initial study conceptualization to include an analysis of images.

Previous social media monitoring studies have demonstrated the prevalence of misinformation on platforms [110-112]. Therefore, it is possible that some tweets analyzed contained incorrect information. To better spread reliable information during a public health emergency such as the COVID-19 pandemic, social media is an invaluable tool for policy makers to understand the way public health information is shared. Future social media monitoring studies should consider and analyze any misinformation found within the dataset.

Finally, for the analysis of viral tweets, the number of likes and retweets each tweet accumulated was looked at. However, these numbers reflect these metrics from when the data were collected in Winter 2022. Therefore, it is possible that by the end of the time analyzed, March 31, 2021, these figures were lower and that more likes and retweets were added in the 7 months before the data were collected. However, it is unlikely that this altered the figures significantly. This is because as time goes on, newer tweets push these viral tweets down on the home page, and users are therefore less likely to see them [113].

This study only looked at social media posts on Twitter, but conversations on other platforms, such as Facebook and Instagram, could be beneficial to monitor in a similar way. Analyzing images posted on social media in addition to text would also be beneficial. The quantity of tweets in Spanish was so low that most Spanish tweets were ignored by topic modeling software. Future research should separate Spanish and English tweets before topic modeling to avoid this issue. Furthermore, it is possible that Spanish-speaking New York City residents do not use Twitter at the same rates as English-speaking residents, or they may use Twitter in English and other platforms in Spanish; therefore, future research should hone in on social media platforms used more prominently by Spanish speakers.

Since this study was originally conducted, Twitter was purchased and rebranded as a new social media platform called X. This raises questions about the reproducibility of the study now that Twitter, as it was once known, no longer exists. Talkwalker representatives stated that all data and accounts from Twitter were not deleted or otherwise removed in the change to X (Talkwalker Product Support Representative, email, August 2024). Any tweets no longer in existence on X are a result of the user's decision to remove them. Therefore, any researcher who would like to retrieve the same data used in this study would need only the same Talkwalker access and Boolean search strings used here. However, the General Data Protection Regulation (GDPR), implemented in the European Union in 2018, requires companies to delete any stored digital data if a user requests so [114,115]. Therefore, if a user deletes a post or account from X, Talkwalker is also required to delete any data they have associated with that account or post. The GDPR only affects companies operating in the European Union, and because Talkwalker is a global company, this policy does apply to them. Talkwalker also chooses to abide by these standards for users and posts located outside of the European Union (Talkwalker Product Support Representative, email, August 2024). While it is unknown exactly why Talkwalker chooses to do so, there are several reasons a company may do this. One potential explanation is that the GDPR can be seen as the highest standard for digital rights, and by adhering to it in countries with different standards, they are not only meeting the requirements of the host country but also providing protections above and beyond that.

It is unclear what the full impact of the change from Twitter to X will be on future social media listening research, including studies such as this one. Since the rebrand to X, many users migrated away from the platform and onto other social media websites, including sites such as Bluesky, Threads, and Mastodon [116]. Some analysts estimated that Twitter lost as

XSL•FO
RenderX

many as 1 million users in the months after the company's ownership change that will eventually lead to the rebranding to X [117]. However, many of these alternative sites were far less successful than Twitter, and some users eventually migrated back to X. Many users have concluded that the transition from Twitter to X had little significant impact on the experience of using the social media platform [116,118].

In sum, if any accounts or posts were deleted since data for this study were initially retrieved, they would be unavailable to future researchers. This is not a new phenomenon. For many studies involving computational communication science, as more time passes from the data collection stage of a study, less and less data are available to future researchers as a result of account and post deletions, changed privacy settings, and user suspensions [119].

Conclusions

Findings reveal that in the earliest months of the COVID-19 pandemic, Twitter users in New York City quickly responded to the wave of need by sharing information and resources about food access in their communities. Some users turned to Twitter to either solicit or offer help finding food. Furthermore, the platform lent itself to many conversations about the policies enacted on a federal, state, and city level to help feed New Yorkers in need. Future research on this topic should include an analysis of social media posting on platforms such as Facebook, TikTok, and Instagram, as well as in languages other than English. Social media data can be used to explore and better understand the food security and food access challenges and concerns of local communities and how such challenges and concerns may evolve over the course of the pandemic. Community leaders and elected officials could use such data to inform more responsive and tailored initiatives and services to better address such crises.

Acknowledgments

Funding for this study was provided by the New York City Council and a Professional Staff Congress—City University of New York Enhanced grant.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Boolean search strings.
[DOCX File , 16 KB - formative_v9i1e49520_app1.docx]

Multimedia Appendix 2
Themes and related subthemes from topic modeling analysis of collected tweets, presented with the common words or roots, the percentage of tweets from the dataset that fall into each theme, and at least 1 anonymized example tweet per subtheme (N=129,101).
[DOCX File , 21 KB - formative_v9i1e49520_app2.docx]

Multimedia Appendix 3
Top 10 viral tweets, in chronological order of original post, presented with each tweet's total number of likes and retweets at the time of data collection and a brief qualitative summary of the content posted in replies and quote tweets.
[DOCX File , 33 KB - formative_v9i1e49520_app3.docx]

References

1. Huizar MI, Arena R, Laddu DR. The global food syndemic: the impact of food insecurity, malnutrition and obesity on the healthspan amid the COVID-19 pandemic. Prog Cardiovasc Dis 2021;64:105-107 [FREE Full text] [doi: 10.1016/j.pcad.2020.07.002] [Medline: 32653438]
2. Azuma AM, Gilliland S, Vallianatos M, Gottlieb R. Food access, availability, and affordability in 3 Los Angeles communities, Project CAFE, 2004-2006. Prev Chronic Dis 2010 Mar;7(2):A27 [FREE Full text] [Medline: 20158956]
3. Feeding America network faces soaring demand, plummeting supply due to COVID-19 crisis. Feeding America. URL: https://www.feedingamerica.org/about-us/press-room/soaring-demand-plummeting-supply [accessed 2024-08-11]
4. Karpman M, Zuckerman S, Gonzalez D, Kenney G. The COVID-19 pandemic is straining families' abilities to afford basic needs. Urban Institute. URL: https://www.urban.org/sites/default/files/publication/102124/the-covid-19-pandemic-is-straining-families-abilities-to-afford-basic-needs_4.pdf [accessed 2024-04-29]
5. Naja F, Hamadeh R. Nutrition amid the COVID-19 pandemic: a multi-level framework for action. Eur J Clin Nutr 2020 Aug;74(8):1117-1121 [FREE Full text] [doi: 10.1038/s41430-020-0634-3] [Medline: 32313188]
6. Cohen N. Food crisis as a tool for social change: lessons from New York City's COVID-19 response. Urban Gov 2022 Jun;2(1):173-177. [doi: 10.1016/j.ugj.2022.03.001]

XSL•FO
RenderX

7. Mann B. In New York City, 2 million residents face food insecurity, officials say. National Public Radio (NPR). URL: https://www.npr.org/sections/coronavirus-live-updates/2020/05/21/860312565/in-new-york-city-2-million-residents-face-food-insecurity-officials-say [accessed 2023-05-02]
8. Mazzeo J, Al Abdeen Qusair Z, Gadhoke P, Freiberg T, Brenton BP, Sedlacek A, et al. A tale of two cities during the COVID-19 pandemic: evaluating food insecurity in Chicago and New York City. J Racial Ethn Health Disparities 2023 Aug;10(4):1703-1720 [FREE Full text] [doi: 10.1007/s40615-022-01355-x] [Medline: 35831703]
9. The Hunter College NYC Food Policy Center, The CUNY Urban Food Policy Institute. NY Food 20/20: vision, research, and recommendations during COVID-19 and beyond. CUNY Academic Works. URL: https://academicworks.cuny.edu/sph_pubs/273/ [accessed 2024-04-29]
10. Gadhoke P, Brenton BP, Sadeghzadeh S, Huang SY, Lane L, Mathieu D, et al. The intersectionality of food insecurity, job insecurity, healthcare insecurity, and mental health challenges in NYC households during COVID-19. Ecol Food Nutr 2024;63(2):112-134. [doi: 10.1080/03670244.2024.2308669] [Medline: 38421255]
11. Wojcik S, Hughes A. Sizing up Twitter users. Pew Research Center. URL: https://www.pewresearch.org/internet/wp-content/uploads/sites/9/2019/04/twitter_opinions_4_18_final_clean.pdf [accessed 2019-04-24]
12. Mosleh M, Pennycook G, Arechar AA, Rand DG. Cognitive reflection correlates with behavior on Twitter. Nat Commun 2021 Feb 10;12(1):921 [FREE Full text] [doi: 10.1038/s41467-020-20043-0] [Medline: 33568667]
13. Mohammadi E, Thelwall M, Kwasny M, Holmes KL. Academic information on Twitter: a user survey. PLoS One 2018;13(5):e0197265 [FREE Full text] [doi: 10.1371/journal.pone.0197265] [Medline: 29771947]
14. Abd-Alrazaq A, Alhuwail D, Househ M, Hamdi M, Shah Z. Top concerns of tweeters during the COVID-19 pandemic: infoveillance study. J Med Internet Res 2020 Apr 21;22(4):e19016 [FREE Full text] [doi: 10.2196/19016] [Medline: 32287039]
15. Ridhwan KM, Hargreaves CA. Leveraging Twitter data to understand public sentiment for the COVID‐19 outbreak in Singapore. Int J Inf Manag Data Insights 2021 Nov;1(2):100021 [FREE Full text] [doi: 10.1016/j.jjimei.2021.100021]
16. Boon-Itt S, Skunkan Y. Public perception of the COVID-19 pandemic on Twitter: sentiment analysis and topic modeling study. JMIR Public Health Surveill 2020 Nov 11;6(4):e21978 [FREE Full text] [doi: 10.2196/21978] [Medline: 33108310]
17. de Choudhury M, Sharma S, Kiciman E. Characterizing dietary choices, nutrition, and language in food deserts via social media. In: Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing. 2016 Presented at: CSCW '16; February 27-March 2, 2016; San Francisco, CA p. 1157-1170 URL: https://dl.acm.org/doi/10.1145/2818048.2819956 [doi: 10.1145/2818048.2819956]
18. Steils N, Obaidalahe Z. “Social food”: food literacy co-construction and distortion on social media. Food Policy 2020 Aug;95:101932. [doi: 10.1016/j.foodpol.2020.101932]
19. Kim Y, Huang J, Emery S. Garbage in, garbage out: data collection, quality assessment and reporting standards for social media data use in health research, infodemiology and digital disease detection. J Med Internet Res 2016 Feb 26;18(2):e41 [FREE Full text] [doi: 10.2196/jmir.4738] [Medline: 26920122]
20. Vydra S, Kantorowicz J. Tracing policy-relevant information in social media: the case of twitter before and during the COVID-19 crisis. Stat Politics Policy 2022;12(1):87-127. [doi: 10.1515/spp-2020-0013]
21. Martini F, Samula P, Keller TR, Klinger U. Bot, or not? Comparing three methods for detecting social bots in five political discourses. Big Data Soc 2021 Aug 23;8(2):22. [doi: 10.1177/20539517211033566]
22. Botometer Pro API. RapidAPI. URL: https://rapidapi.com/OSoMe/api/botometer-pro/details [accessed 2022-07-22]
23. Xu W, Sasahara K. Characterizing the roles of bots on Twitter during the COVID-19 infodemic. J Comput Soc Sci 2022 Aug 30;5(1):591-609 [FREE Full text] [doi: 10.1007/s42001-021-00139-3] [Medline: 34485752]
24. Gruzd A, Mai P. Going viral: how a single tweet spawned a COVID-19 conspiracy theory on Twitter. Big Data Soc 2020 Jul 20;7(2):25. [doi: 10.1177/2053951720938405]
25. Varol O, Ferrara E, Davis C, Menczer F, Flammini A. Online human-bot interactions: detection, estimation, and characterization. In: Proceedings of the 11th International AAAI Conference on Web and Social Media. 2017 Presented at: AAAI/WSM '17; May 15-18, 2017; Montréal, QC p. 280-289 URL: https://cdn.aaai.org/ojs/14871/14871-28-18390-1-2-20201228.pdf [doi: 10.1609/icwsm.v11i1.14871]
26. kappam.light: light's Kappa for m raters. RDocumentation. URL: https://www.rdocumentation.org/packages/irr/versions/0.84.1/topics/kappam.light [accessed 2023-05-02]
27. Stritch JM, Pedersen MJ, Taggart G. The opportunities and limitations of using Mechanical Turk (MTURK) in public administration and management scholarship. Int Public Manag J 2017 Jan 19;20(3):489-511. [doi: 10.1080/10967494.2016.1276493]
28. Chambers S, Anthony-McMann P. A primer for conducting survey research using MTurk: tips for the field. Int J Adult Vocat Educ Train 2016;7(2):54-73. [doi: 10.4018/IJAVET.2016040105]
29. Dennis SA, Goodson BM, Pearson C. Online worker fraud and evolving threats to the integrity of MTurk data: a discussion of virtual private servers and the limitations of IP-based screening procedures. Behav Res Account 2019:63. [doi: 10.2308/bria-18-044]

XSL•FO
RenderX

30. Quarterly report pursuant to section 13 or 15(D) of the securities exchange Act of 1934. United States Securities and Exchange Commission. URL: https://www.sec.gov/ix?doc=/Archives/edgar/data/1418091/000141809122000075/twtr-20220331.htm [accessed 2022-08-03]
31. O'Leary L. How bad is Twitter's bot problem? The Slate Group. URL: https://slate.com/technology/2022/06/how-bad-are-bots-on-twitter.html [accessed 2023-05-03]
32. Nothman J, Qin H, Yurchak R. Stop word lists in free open-source software packages. In: Proceedings of the 2018 Workshop for NLP Open Source Software. 2018 Presented at: NLP-OSS '18; July 20, 2018; Melbourne, Australia p. 7-12 URL: https://aclanthology.org/W18-2502.pdf [doi: 10.18653/v1/w18-2502]
33. seblier/NLTK's list of english stopwords. GitHub. URL: https://gist.github.com/sebleier/554280 [accessed 2022-07-27]
34. Vayansky I, Kumar SA. A review of topic modeling methods. Inf Sys 2020 Dec;94:101582. [doi: 10.1016/j.is.2020.101582]
35. Liu L, Tang L, Dong W, Yao S, Zhou W. An overview of topic modeling and its current applications in bioinformatics. Springerplus 2016;5(1):1608 [FREE Full text] [doi: 10.1186/s40064-016-3252-8] [Medline: 27652181]
36. Bowers B. A unique approach to short text clustering (algorithmic theory). Medium. URL: https://towardsdatascience.com/a-unique-approach-to-short-text-clustering-part-1-algorithmic-theory-4d4fad0882e1 [accessed 2020-01-24]
37. Yin J, Wang J. A dirichlet multinomial mixture model-based approach for short text clustering. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. 2014 Presented at: KDD '14; August 24-27, 2014; New York, NY p. 233-242 URL: https://dl.acm.org/doi/10.1145/2623330.2623715 [doi: 10.1145/2623330.2623715]
38. Mazarura J, de Waal A. A comparison of the performance of latent Dirichlet allocation and the Dirichlet multinomial mixture model on short text. In: Proceedings of the 2016 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference. 2016 Presented at: PRASA-RobMech '16; November 30-December 2, 2016; Stellenbosch, South Africa p. 1-6 URL: https://ieeexplore.ieee.org/document/7813155 [doi: 10.1109/robomech.2016.7813155]
39. rwalk / gsdmm: GSDMM: short text clustering. GitHub. URL: https://github.com/rwalk/gsdmm [accessed 2022-07-22]
40. Chin J. jonathan-chin / food-chatter-code. GitHub. URL: https://github.com/jonathan-chin/food-chatter-code [accessed 2023-05-31]
41. Harlem. Poetry Foundation. URL: https://www.poetryfoundation.org/poems/46548/harlem [accessed 2024-04-29]
42. HRPP-104 CUNY HRPP Policy: Exempt Review of Human Subjects Research. City University of New York. 2024 Dec. URL: https://cuny.policystat.com/policy/17202761/latest/#autoid-69q6p
43. Moss K, Wexler A, Dawson L, Long M, Kates J, Cubanski J, et al. The coronavirus aid, relief, and economic security act: summary of key health provisions. KFF. URL: https://www.kff.org/coronavirus-covid-19/issue-brief/the-coronavirus-aid-relief-and-economic-security-act-summary-of-key-health-provisions/ [accessed 2023-05-03]
44. Katz M. Every public school student is getting a $420 food stamp card, regardless of income. What should higher earners do? Gothamist. URL: https://gothamist.com/food/every-public-school-student-getting-420-food-stamp-card-regardless-income-what-should-higher-earners-do [accessed 2023-05-03]
45. SNAP benefits expanded for online grocery store orders. Fox 5. URL: https://www.fox5dc.com/news/snap-benefits-expanded-for-online-grocery-store-orders [accessed 2023-05-03]
46. Jones JW. Online supplemental nutrition assistance program (SNAP) purchasing grew substantially in 2020. Economic Research Service. 2020. URL: https://www.ers.usda.gov/amber-waves/2021/july/online-supplemental-nutrition-assistance-program-snap-purchasing-grew-substantially-in-2020 [accessed 2024-04-29]
47. Solis M. 1,400 union workers at Bronx supermarket strike over hourly wages. ABC Inc. URL: https://abc7ny.com/hunts-point-supermarket-produce-market-teamsters-union-bronx/9881692/ [accessed 2023-05-03]
48. New York City to close all school buildings and transition to remote learning. City of New York. URL: https://www.nyc.gov/office-of-the-mayor/news/151-20/new-york-city-close-all-school-buildings-transition-remote-learning [accessed 2023-05-03]
49. Gurley LK. Whole foods employees are staging a nationwide 'sick-out'. VICE. URL: https://www.vice.com/en/article/5dmeka/whole-foods-employees-are-staging-a-nationwide-sick-out [accessed 2023-05-03]
50. Saxena J. Workers from whole foods, Instacart, Amazon, and more plan may day strike. Eater. URL: https://www.eater.com/2020/4/29/21241320/workers-from-whole-foods-instacart-amazon-target-organize-may-day-strike [accessed 2023-05-03]
51. Levine M, Barrón-López L, Ferris S. Progressives push to squash Senate filibuster after minimum wage defeat. POLITICO. URL: https://www.politico.com/news/2021/02/26/progressives-minimum-wage-defeat-471772 [accessed 2023-05-03]
52. Blakemore E. How the black panthers' breakfast program both inspired and threatened the government. The History Channel. URL: https://www.history.com/news/free-school-breakfast-black-panther-party [accessed 2023-05-03]
53. Shapiro E. New York City public schools to close to slow spread of coronavirus. The New York Times. URL: https://www.nytimes.com/2020/03/15/nyregion/nyc-schools-closed.html [accessed 2023-05-03]
54. Luhbi T. Stimulus bill offers $600 a week to the unemployed for 4 months. CNN. URL: https://www.cnn.com/2020/03/25/politics/senate-stimulus-unemployment-benefits-coronavirus/index.html [accessed 2023-05-03]
55. Laterman K. Hunger is on the rise. Food donors are getting creative. The New York Times. URL: https://www.nytimes.com/2020/03/13/nyregion/snap-benefits-food-insecurity-nyc.html [accessed 2024-04-29]

XSL•FO
RenderX

56. Neuman S. Essential workers plan may day strikes; others demand end to COVID-19 lockdowns. NPR. URL: https://www.npr.org/sections/coronavirus-live-updates/2020/05/01/848931228/essential-workers-plan-may-day-strikes-others-demand-end-to-covid-19-lockdowns [accessed 2023-05-03]
57. Tyrväinen O, Karjaluoto H. Online grocery shopping before and during the COVID-19 pandemic: a meta-analytical review. Telemat Inform 2022 Jul;71:101839 [FREE Full text] [doi: 10.1016/j.tele.2022.101839] [Medline: 35607591]
58. Grashuis J, Skevas T, Segovia MS. Grocery shopping preferences during the COVID-19 pandemic. Sustainability 2020 Jul 02;12(13):5369. [doi: 10.3390/su12135369]
59. Kumar V. Adjusting to the new normal: challenges of the food sector in the wake of COVID-19. J Supply Chain Manag Log Manag 2020 Dec 1;3(2):163. [doi: 10.69554/ptfh2059]
60. de Paulo Farias D, Dos Santos Gomes MG. COVID-19 outbreak: what should be done to avoid food shortages? Trends Food Sci Technol 2020 Aug;102:291-292 [FREE Full text] [doi: 10.1016/j.tifs.2020.06.007] [Medline: 32834501]
61. Anderson JD, Mitchell JL, Maples JG. Invited review: lessons from the COVID-19 pandemic for food supply chains. Applied Animal Science 2021 Dec;37(6):738-747 [FREE Full text] [doi: 10.15232/aas.2021-02223] [Medline: 36812219]
62. Luckstead J, Nayga RM, Snell HA. Labor issues in the food supply chain amid the COVID-19 pandemic. Appl Econ Perspect Policy 2021 Mar;43(1):382-400 [FREE Full text] [doi: 10.1002/aepp.13090] [Medline: 33173571]
63. Hobbs JE. Food supply chains during the COVID‐19 pandemic. Can J Agric Econ 2020 May 03;68(2):171-176. [doi: 10.1111/cjag.12237]
64. Rizou M, Galanakis IM, Aldawoud TM, Galanakis CM. Safety of foods, food supply chain and environment within the COVID-19 pandemic. Trends Food Sci Technol 2020 Aug;102:293-299 [FREE Full text] [doi: 10.1016/j.tifs.2020.06.008] [Medline: 32834502]
65. Hartke K. How Chinatown businesses are using farm-to-market ingenuity to survive. The Washington Post. URL: https://www.washingtonpost.com/food/2023/05/09/chinatown-market-bargains/ [accessed 2023-05-16]
66. Imbruce V. From Farm to Canal Street: Chinatown's Alternative Food Network in the Global Marketplace. New York, NY: Cornell University Press; 2015.
67. Food supply chains and COVID-19: impacts and policy lessons. Organisation for Economic Co-operation and Development. URL: https://www.oecd.org/coronavirus/policy-responses/food-supply-chains-and-covid-19-impacts-and-policy-lessons-71b57aea/ [accessed 2023-05-03]
68. Pietzold J. NY partnering with state dairy producers to give excess milk to food banks, those in need. CNY Central. URL: https://cnycentral.com/news/local/ny-partnering-with-state-dairy-producers-to-process-distribute-excess-milk-to-food-banks [accessed 2024-04-29]
69. Nourish New York. New York State. URL: https://agriculture.ny.gov/NourishNY [accessed 2023-03-15]
70. Governor Hochul signs legislation codifying the Nourish New York program into law. New York State. URL: https://www.governor.ny.gov/news/governor-hochul-signs-legislation-codifying-nourish-new-york-program-law [accessed 2023-05-03]
71. Christensen G, Bronchetti ET. Local food prices and the purchasing power of SNAP benefits. Food Policy 2020 Aug;95:101937. [doi: 10.1016/j.foodpol.2020.101937]
72. Bronchetti E, Christensen G, Hansen B. Variation in food prices and SNAP adequacy for purchasing the thrifty food plan. University of Kentucky. URL: https://uknowledge.uky.edu/ukcpr_papers/115 [accessed 2024-04-29]
73. Ziliak JP. Modernizing SNAP benefits. The Hamilton Project. URL: https://gattonweb.uky.edu/Faculty/Ziliak/final_ziliak_modernizing_snap_benefits.pdf [accessed 2023-05-03]
74. Byrne AT, Just DR. The other half: an examination of monthly food pantry cycles in the context of SNAP benefits. Appl Econ Perspect Policy 2021 Jan 28;43(2):716-731. [doi: 10.1002/aepp.13150]
75. Consumer price index. US Bureau of Labor Statistics. URL: https://www.bls.gov/cpi/ [accessed 2023-05-03]
76. Frequently asked questions for the emergency allotment of SNAP benefits. Office of Temporary and Disability Assistance. URL: https://otda.ny.gov/SNAP-COVID-19/Frequently-Asked-Questions.asp [accessed 2023-05-03]
77. What can SNAP buy? United States Food and Nutrition Service. URL: https://www.fns.usda.gov/snap/eligible-food-items [accessed 2023-05-03]
78. Accessory foods list. United States Department of Agriculture. URL: https://fns-prod.azureedge.us/sites/default/files/resource-files/Accessory_Foods_List.pdf [accessed 2023-05-03]
79. Restaurant meals program. United States Department of Agriculture. URL: https://www.fns.usda.gov/snap/retailer/restaurant-meals-program [accessed 2023-05-03]
80. Gaines-Turner T, Simmons JC, Chilton M. Recommendations from SNAP participants to improve wages and end stigma. Am J Public Health 2019 Dec;109(12):1664-1667. [doi: 10.2105/AJPH.2019.305362] [Medline: 31622134]
81. Leung MM, Tomczyk D, Eilender E, Li C, Aguilera Y, Butz L, et al. Capturing the state of food access in NYC during COVID-19 recovery: insights on preparing for the next crisis. NY Food 2025. Tisch Center for Food Education & Policy. 2025. URL: https://www.nycfoodpolicy.org/wp-content/uploads/2022/10/Food-Access-Covid19-Recovery_v3c_pages.pdf [accessed 2024-04-29]
82. Gustafson AA, Allen JE, Schoenberg NE, Swanson M. The relationship between neighborhood food environment and food store choice on purchasing habits among SNAP and lower income households, USDA FoodAPS data. University of Kentucky. URL: https://uknowledge.uky.edu/ukcpr_papers/110 [accessed 2023-05-03]

XSL•FO
RenderX

83. Gustafson A. Shopping pattern and food purchase differences among Supplemental Nutrition Assistance Program (SNAP) households and non-supplemental nutrition assistance program households in the United States. Prev Med Rep 2017 Sep;7:152-157 [FREE Full text] [doi: 10.1016/j.pmedr.2017.06.005] [Medline: 28660124]
84. Elbel B, Tamura K, McDermott ZT, Duncan DT, Athens JK, Wu E, et al. Disparities in food access around homes and schools for New York City children. PLoS One 2019;14(6):e0217341 [FREE Full text] [doi: 10.1371/journal.pone.0217341] [Medline: 31188866]
85. Ruff RR, Akhund A, Adjoian T. Small convenience stores and the local food environment: an analysis of resident shopping behavior using multilevel modeling. Am J Health Promot 2016;30(3):172-180. [doi: 10.4278/ajhp.140326-QUAN-121] [Medline: 25806566]
86. Haws KL, Reczek RW, Sample KL. Healthy diets make empty wallets: the healthy = expensive intuition. J Consum Res 2017 Jan 12:ucw078. [doi: 10.1093/jcr/ucw078]
87. Carlson A, Frazao E. Are healthy foods really more expensive? It depends on how you measure the price. U.S. Department of Agriculture, Economic Research Service. URL: https://www.ers.usda.gov/publications/pub-details?pubid=44679#:~:text=For%20all%20metrics%20except%20the,little%20to%20meeting%20dietary%20recommendations) [accessed 2023-05-03]
88. Daniel C. Is healthy eating too expensive?: how low-income parents evaluate the cost of food. Soc Sci Med 2020 Mar;248:112823 [FREE Full text] [doi: 10.1016/j.socscimed.2020.112823] [Medline: 32062569]
89. Howse E, Hankey C, Allman-Farinelli M, Bauman A, Freeman B. 'Buying salad is a lot more expensive than going to McDonalds': young adults' views about what influences their food choices. Nutrients 2018 Jul 30;10(8):25 [FREE Full text] [doi: 10.3390/nu10080996] [Medline: 30061513]
90. Cadenhead JW, McCarthy JE, Nguyen TT, Rodriguez M, Koch PA. Qualitative study of participation facilitators and barriers for emergency school meals and pandemic electronic benefits (P-EBT) in an urban setting during COVID-19. Nutrients 2022 Aug 16;14(16):25 [FREE Full text] [doi: 10.3390/nu14163358] [Medline: 36014864]
91. Fraser KT, Pereira J, Poppendieck J, Tavarez E, Berg J, Freudenberg N. Pandemic EBT in New York State: lessons from the 2019-2020 academic year and recommendations for 2020-2021 and beyond. CUNY Urban Food Policy Institute and Hunger Free America. URL: https://cunyurbanfoodpolicy.org/wp-content/uploads/2022/04/PEBT_Policy_Brief_online_01192021.pdf [accessed 2023-05-03]
92. Acciai F, Ohri-Vachaspati P, Yedidia MJ. Schools' participation in the community eligibility provision affects students' receipt of emergency benefits during the COVID-19 pandemic. Nutrients 2022 Nov 21;14(22):25 [FREE Full text] [doi: 10.3390/nu14224919] [Medline: 36432605]
93. Bellafante G. Why giving food stamps to the rich is not a terrible idea. The New York Times. URL: https://www.nytimes.com/2020/09/18/nyregion/coronavirus-food-stamps-nyc-schools.html [accessed 2023-05-03]
94. DesOrmeau T. Some high-income families are getting EBT cards during the pandemic. Here's why. MLive. URL: https://www.mlive.com/public-interest/2020/05/some-high-income-families-are-getting-ebt-cards-during-the-pandemic-heres-why.html [accessed 2024-04-29]
95. Kenney EL, Walkinshaw LP, Shen Y, Fleischhacker SE, Jones-Smith J, Bleich SN, et al. Costs, reach, and benefits of COVID-19 pandemic electronic benefit transfer and grab-and-go school meals for ensuring youths' access to food during school closures. JAMA Netw Open 2022 Aug 01;5(8):e2229514 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.29514] [Medline: 36044212]
96. Waxman E, Gupta P, Pratt E, Lyons M, Green C. Documenting pandemic EBT for the 2020-21 school year: state perspectives on implementation challenges and lessons for the future. Urban Institute. 2020. URL: https://eric.ed.gov/?id=ED616283 [accessed 2024-04-29]
97. Karnaze A. You Are Where You Eat: discrimination in the National School Lunch Program. NW U Law Rev 2018;113(3):629-666 [FREE Full text]
98. Atkinson A. Philando castile, state violence, and school lunch debt: a meditation. NYU Law Rev (Forthcoming) 2021 [FREE Full text]
99. Lateef H, Androff D. "Children can't learn on an empty stomach": the black panther party's free breakfast program. J Sociol Soc Welf 2017 Jan 01;44(4):25. [doi: 10.15453/0191-5096.3883]
100. Thomas MM, Miller DP, Morrissey TW. Food insecurity and child health. Pediatrics 2019 Oct;144(4):3514. [doi: 10.1542/peds.2019-0397] [Medline: 31501236]
101. Gallegos D, Eivers A, Sondergeld P, Pattinson C. Food insecurity and child development: a state-of-the-art review. Int J Environ Res Public Health 2021 Aug 26;18(17):e31 [FREE Full text] [doi: 10.3390/ijerph18178990] [Medline: 34501578]
102. Shankar P, Chung R, Frank DA. Association of food insecurity with children's behavioral, emotional, and academic outcomes: a systematic review. J Dev Behav Pediatr 2017;38(2):135-150. [doi: 10.1097/DBP.0000000000000383] [Medline: 28134627]
103. Jackson DB, Testa A. Household food insecurity and preschool suspension/expulsion in the United States. Prev Med 2020 Dec;141:106283. [doi: 10.1016/j.ypmed.2020.106283] [Medline: 33035546]
104. Cuellar AE, Markowitz S. School suspension and the school-to-prison pipeline. Int Rev Law Econ 2015 Aug;43:98-106. [doi: 10.1016/j.irle.2015.06.001]

XSL•FO
RenderX

105. Mallett CA. The school-to-prison pipeline: a critical review of the punitive paradigm shift. Child Adolesc Soc Work J 2015 Apr 28;33(1):15-24 [FREE Full text] [doi: 10.1007/S10560-015-0397-1]
106. Jenders M, Kasneci G, Naumann F. Analyzing and predicting viral tweets. In: Proceedings of the 22nd International Conference on World Wide Web. 2013 Presented at: WWW '13; May 13-17, 2013; Rio de Janeiro, Brazil p. 657-664 URL: https://dl.acm.org/doi/10.1145/2487788.2488017 [doi: 10.1145/2487788.2488017]
107. Carroll G, Keene D, Santilli A, Johannes J, Ickovics J, O'Connor Duffany K. Availability, accessibility, utilization: in-depth interviews with food insecure residents and emergency food providers in New Haven, CT. J Hunger Environ Nutr 2018 Dec 10;14(1-2):240-251. [doi: 10.1080/19320248.2018.1555072]
108. Ginsburg ZA, Bryan AD, Rubinstein EB, Frankel HJ, Maroko AR, Schechter CB, et al. Unreliable and difficult-to-access food for those in need: a qualitative and quantitative study of urban food pantries. J Community Health 2019 Feb;44(1):16-31 [FREE Full text] [doi: 10.1007/s10900-018-0549-2] [Medline: 30019196]
109. Byrne AT, Just DR, Barrett CB. But it came from a food pantry: product stigma and quality perceptions of food pantry offerings. Agric Econ 2023 Jan 08;54(2):327-344. [doi: 10.1111/agec.12755]
110. Tsao SF, Chen H, Tisseverasinghe T, Yang Y, Li L, Butt ZA. What social media told us in the time of COVID-19: a scoping review. Lancet Digit Health 2021 Mar;3(3):e175-e194 [FREE Full text] [doi: 10.1016/S2589-7500(20)30315-0] [Medline: 33518503]
111. Goel A, Gupta L. Social media in the times of COVID-19. J Clin Rheumatol 2020 Sep;26(6):220-223 [FREE Full text] [doi: 10.1097/RHU.0000000000001508] [Medline: 32852927]
112. Cinelli M, Quattrociocchi W, Galeazzi A, Valensise CM, Brugnoli E, Schmidt AL, et al. The COVID-19 social media infodemic. Sci Rep 2020 Oct 06;10(1):16598 [FREE Full text] [doi: 10.1038/s41598-020-73510-5] [Medline: 33024152]
113. Hodas NO, Lerman K. The simple rules of social contagion. Sci Rep 2014 Mar 11;4:4343 [FREE Full text] [doi: 10.1038/srep04343] [Medline: 24614301]
114. Hoofnagle CJ, van der Sloot B, Borgesius FZ. The European Union general data protection regulation: what it is and what it means. Inf Commun Technol Law 2019 Feb 10;28(1):65-98. [doi: 10.1080/13600834.2019.1573501]
115. Regulation (EU) 2016/679 (general data protection regulation). Intersoft Consulting. URL: https://gdpr-info.eu/ [accessed 2024-08-22]
116. Jeong U, Nirmal A, Jha K, Xu Tang S, Bernard H, Liu H. User migration across multiple social media platforms. arXiv Preprint posted online on September 22, 2023 [FREE Full text] [doi: 10.1137/1.9781611978032.51]
117. Stoke-Walker C. Twitter may have lost more than a million users since Elon Musk took over. MIT Technology Review. URL: https://www.technologyreview.com/2022/11/03/1062752/twitter-may-have-lost-more-than-a-million-users-since-elon-musk-took-over/ [accessed 2024-09-21]
118. Schukow CP, Punjabi LS, Abdul-Karim FW. #PathX: #PathTwitter's transformation and a discussion on different social media platforms used by pathologists in 2024. Adv Anat Pathol 2024 Jul 01;31(4):275-277. [doi: 10.1097/PAP.0000000000000424] [Medline: 38047394]
119. Knöpfle P, Schatto-Eckrodt T. The challenges of replicating volatile platform-data studies: replicating Schatto-Eckrodt et al. (2020). Media Commun 2024 Apr 15;12:362. [doi: 10.17645/mac.7789]

Abbreviations

EBT: electronic benefit transfer
GDPR: General Data Protection Regulation
GSDMM: Gibbs Sampling Dirichlet Multinomial Mixture
MTurk: Mechanical Turk
P-EBT: pandemic electronic benefit transfer
SNAP: Supplemental Nutrition Assistance Program

Edited by A Mavragani; submitted 01.06.23; peer-reviewed by L Mcintyre, Y Wang; comments to author 16.07.24; revised version received 10.10.24; accepted 17.11.24; published 09.05.25.

Please cite as:
Butz L, Platkin C, Chin J, Chavez Salas JP, Serres E, Leung MM
Food Access in New York City During the COVID-19 Pandemic: Social Media Monitoring Study
JMIR Form Res 2025;9:e49520
URL: https://formative.jmir.org/2025/1/e49520
doi:10.2196/49520
PMID:

XSL•FO
RenderX

©Leah Butz, Charles Platkin, Jonathan Chin, Juan Pablo Chavez Salas, Ellie Serres, May May Leung. Originally published in JMIR Formative Research (https://formative.jmir.org), 09.05.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Comparing Media and Law Enforcement Reports on Anti-Asian Hate Incidents During the COVID-19 Pandemic: Data Visualization Approach

Young Ji Yoon[1], PhD; Su Hyun Shin[2], PhD; Dongwook Kim[3], MA, MSW; Hee Yun Lee[4], PhD

[1]Department of Social Work, Colorado State University Pueblo, Pueblo, CO, United States
[2]Department of Family and Consumer Studies, The University of Utah, Salt Lake City, UT, United States
[3]School of Social Work, Arizona State University, Phoenix, AZ, United States
[4]School of Social Work, University of Georgia, Athens, GA, United States

Corresponding Author:
Young Ji Yoon, PhD
Department of Social Work
Colorado State University Pueblo
2200 Bonforte Boulevard
Pueblo, CO, 81001
United States
Phone: 1 7195492847
Email: youngji.yoon@csupueblo.edu

Abstract

Background: During the COVID-19 pandemic, anti-Asian hate incidents (AAHIs) increased conspicuously. Literature reports discrepancies in how crimes are reported differently in media and law enforcement data, emphasizing potential biases and inconsistencies in AAHI reporting. Understanding the discrepancies in AAHI reporting between the two sources is crucial for improving documentation procedures and addressing systemic issues in reporting mechanisms.

Objective: This study aimed to (1) present the monthly trends in AAHI counts reported by media and law enforcement data from 2020 to 2021, (2) investigate variations in AAHI counts across states and counties for each year, (3) examine discrepancies in AAHI reporting between the two sources at state and county levels during the same period, and (4) delineate differences in the types and geographic distribution of incidents as represented by the two sources.

Methods: This study used two data sources for AAHIs, media data (n=1288) from The Asian American Foundation and law enforcement data (n=1086) from the Federal Bureau of Investigation, for the 2020-2021 period. Descriptive analyses were conducted to evaluate monthly trends, state and county-level variations, and differences in incident types and locations. Ratios of reported incidents between the two sources were calculated to assess discrepancies. Temporal trends were contextualized within key sociopolitical events to offer insights into reporting dynamics.

Results: First, both media and law enforcement data presented a sharp increase in reported AAHIs following the first confirmed COVID-19 case in the United States, peaking around March 2020, coinciding with controversial political rhetoric. A second peak occurred from March to April 2021, immediately following the pandemic's peak, and was followed by a decline as the situation improved. Second, in the two data sources, the state-level analysis indicated that California, Texas, New York, and Washington consistently reported the highest AAHI counts. In 2021, there were notable increases in reported incidents in states such as Wisconsin and Illinois. County-level data revealed persistent high counts in California, particularly in Los Angeles County. Ratios of AAHI counts between the two data sources presented significant discrepancies, with higher ratios in California and New York. Finally, the analysis of incident types revealed that media data reported a higher proportion of harassment (477/1288, 37%), while the law enforcement data reported more property-related incidents (239/1086, 22%). Regarding location types, media data frequently reported incidents in public areas (515/1288, 40%) and businesses (361/1288, 28%), whereas law enforcement data reported more incidents occurring in residential settings (201/1086, 18.5%).

Conclusions: This study highlighted significant trends and disparities in AAHI reporting between media and law enforcement data, underscoring the need for a nuanced understanding of how these incidents were reported. Practice and policy implications suggested fostering community engagement to support Asian communities while enhancing the accuracy and consistency of hate crime reporting.

XSL•FO
RenderX

(JMIR Form Res 2025;9:e70881) doi:10.2196/70881

KEYWORDS

anti-Asian hate incidents; hate crimes; Asian American and Pacific Islanders; COVID-19; United States

Introduction

Historical Background of Anti-Asian Hate Incidents

During the public health emergency of the COVID-19 pandemic, the United States has witnessed a significant rise in racially motivated hate incidents involving harassment and physical assault targeting Asian American and Pacific Islanders (AAPIs) [1]. These incidents reflect deeper societal issues of animosity, othering, misinformation, and negative stereotypes toward these minority groups [2-4].

In the late 19th century, Asian immigrants came to the United States for low-cost labor in the mining, agriculture, and garment industries, enduring harsh conditions and being unfairly blamed for disease outbreaks such as tuberculosis and cholera [5-7]. During this period, "yellow peril" fears emerged, centering on the perceived threat to European culture resulting from the expansion of Asian influence, particularly on the West Coast [5]. This led to the enactment of the 1882 Chinese Exclusion Act and the 1924 Immigration Act, which restricted immigration from Asian countries [7,8]. With the onset of World War II and the post-Pearl Harbor attack, Japanese Americans, including citizens and people younger than 18 years of age, were unfairly forced into concentration camps due to suspicions about their allegiance to the Japanese Empire [9]. This historical context underscores the persistent and systemic nature of anti-Asian sentiment in the United States, which has been exacerbated during times of crisis, such as the COVID-19 pandemic.

With the advancements in promoting racial justice and equality in recent decades, AAPIs have become the fastest-growing racial and ethnic group in the United States [10]. However, a considerable number of AAPIs continue to contend with a feeling of not achieving full acceptance in American society, frequently being subject to the perception of being perennial outsiders [11,12]. Another stereotype, the "model minority," has been applied to AAPIs who have been more successful, partly due to higher educational attainment and income compared to other racial or ethnic minority groups in the United States [13,14].

According to the Othering Theory [15], a racially and socially dominant group tends to distinguish or relegate a nondominant group by designating them as an outgroup, aberrant from the prevailing culture and system [16,17]. When this bias merges with the recognition that the outgroup is implicated in catastrophic historical instances (eg, economic recession and acts of terrorism), the negative sentiments and outward manifestations directed toward the outgroup can be magnified [18]. For example, following the 9/11 attacks, hate crimes witnessed an upsurge, particularly targeting Arab Americans perceived to be affiliated with the group involved in the terrorist attack based on their ethnicity, culture, and religion [19]. Similarly, the COVID-19 pandemic may have provoked the dissemination of blame toward AAPIs, which in turn has contributed to a surge in anti-Asian hate incidents (AAHIs) in the United States.

A Surge of AAHIs During the COVID-19 Pandemic

Amid the COVID-19 pandemic, the perception of AAPIs as perennial outsiders has been further exacerbated. Between January and March 2020, terms such as "Wuhan virus" or "Chinese virus" were used in news media and by government officials, potentially stigmatizing AAPIs [20-22], particularly those of Chinese or Asian appearance, by associating them with virus transmission. The news media in the United States shed light on the sanitation issues in Wuhan's wet markets, drawing attention to the consumption of wild animals as a potential cause of the coronavirus [23,24]. While the World Health Organization (WHO) recommended in February 2020 that media organizations use the terms not specifying a certain country or region to indicate the virus, hate crimes targeting AAPIs swelled [25].

In 2019, prior to the pandemic, the Federal Bureau of Investigation (FBI) hate crime data documented 7314 hate crimes, and 55.8% of single-bias incidents stemmed from race, ethnicity, or ancestry [26]. After the outbreak of the pandemic, hate crimes presented about a 13% increase, with 8263 and 8327 incidents reported in 2020 and 2021, respectively, compared to 2019. During the same period, more than 62% of single-bias hate crime offenses were due to race, ethnicity, or ancestry during the same period [25,27]. A survey conducted during the pandemic reported that about 34% of the Asian respondents in the United States experienced bias (32.3% of noncriminal and 19.5% of criminal bias) [2]. Verbal insults were the most common form of noncriminal bias (23.7%), while threats of physical violence were the most common form of criminal bias (13.1%) [2].

The years 2020 and 2021 were marked by the outbreak of COVID-19, public health measures, politicians' remarks, vaccine rollouts, and prevalence of virus variants [28,29]. Han et al [1] reported on changes in anti-Asian hate crimes by comparing incidents between 2019 and 2020. However, their study covered only a segment of the COVID-19 period, and a research gap persists in studying AAHIs over shorter intervals across an extended period during the pandemic. Given this, it is imperative to investigate the temporal patterning of AAHIs following the outbreak of a pandemic, as this may enable researchers to discern connections between the data and seismic historical events. Such an investigation will reveal how specific pandemic-related events may have influenced the frequency of AAHIs, offering critical insights into the factors that may exacerbate or mitigate these incidents.

Discrepancies in AAHIs Reported Between Media and Law Enforcement Data

The literature underscored the need for researchers to explore AAHIs, drawing on diverse data sources. According to a study

XSL•FO
RenderX

examining police reporting among hate crime victims within the initial 2 months of the pandemic, less than 29% of Asian victims were likely to report their bias experiences to the police, compared to 39%, 41%, and 42% of Hispanic, Black, and White victims, respectively [30]. This suggests that a substantial number of AAHI occurrences do not enter police criminal procedures, indicating the need to explore AAHIs using various data sources to gain insights into multiple phases of these incidents.

For example, Curiel et al [31] compared the frequency of different types of crime reported on Twitter (currently known as "X") with crime data in several Latin American countries. On social media, about 42.2% and 28.3% of posts were about violent crimes and murder, while only 9.8% and 2.4% were about property and sexual crimes, respectively [31]. Conversely, in official crime data, the majority (65.9%) were about property crimes, while 2.5%, 1.5%, and 0.1% were about sexual crimes, violent crimes, and murder, respectively [31]. The disparity in reported crimes between social media and crime data suggested that media coverage of crimes conveyed fear, caution, or perceptions about crimes that went beyond the incidents themselves.

Recently, several studies have reported empirical evidence of AAHIs during COVID-19 using diverse sources. Lantz and Wenger [2] compared AAHI cases reported by Asians and non-Asians through a web-based survey for the first 2 months of COVID-19. Han et al [1] presented AAHIs reported to police in four major US cities with large Asian populations (New York, San Francisco, Seattle, and Washington, DC) in 2019 and 2020. Prior research on AAHIs throughout the pandemic has covered limited regions or relied on data from surveys [1,2].

Despite these studies, there is a notable lack of effort to examine both mass media and law enforcement data pertaining to the AAHIs. Law enforcement data collects and compiles hate incidents that have been investigated and documented by law enforcement agencies, providing an official record of reported incidents [25]. In contrast, news media data may include a broader range of incidents not reported to law enforcement [32-34]. For example, Chadee and Chadee [35] discovered a disparity in crime reporting between police data and newspapers, using a comparative analysis of the reporting ratios from both sources. Their findings revealed that crimes targeting individuals were overreported in newspapers compared to official police reports, while crimes involving property were underreported [35]. Discrepancies between media and law enforcement data can highlight potential biases in reporting or law enforcement practices [33]. Additionally, comparing types and locations of AAHIs between media and law enforcement data can provide insights into potential discrepancies in reporting, the visibility of incidents in different contexts, and the factors that may influence the occurrence and documentation of these hate incidents in various settings.

This Study

Based on this groundwork, the purpose of this study is to investigate the disparities in reporting AAHI counts between media and law enforcement data at both state and county levels during the COVID-19 pandemic. By examining these discrepancies, the study aims to elucidate potential biases in reporting practices. Our research is crucial for initiating discussions among scholars and policy makers to address this societal issue by providing compelling evidence. To achieve the study's aim, we will first illustrate the monthly trends of AAHI counts reported in media and law enforcement data between 2020 and 2021. We will then show the state and county variations in AAHI counts reported each year by both media and law enforcement data, as well as the discrepancies in AAHI count reporting between the two sources at the state and county levels in both years. Additionally, we will investigate the differences in the types and locations of incidents reported in the media and national hate crime data.

Information about safety and exposure to hate crimes is a crucial social determinant of health, shaping behavioral and psychological factors that influence population well-being. While this study does not directly assess these health outcomes, disparities in data sources provide insights that may inform public health discourse and future research on the intersection of crime reporting and community health.

Methods

Data and Variables

This study used incident-level data compiled by The Asian American Foundation (TAAF) and the FBI. TAAF collected AAHI data reported in various news media between 2020 and 2021. The data included the media source, date, type, description, location of each incident, and the victim's gender. TAAF set up and scanned Google News Alerts and Feedly weekly or biweekly to identify AAHI-related news reported in media and blog websites (eg, NextShark) using keywords related to AAPIs and hate incident attributes. Staff members further manually read headlines of articles published in the media and selected those that met the definition of an AAHI case. Although the dataset includes news from AAPI-focused outlets such as AsAmNews and Asian Dawn, only articles published in English were included in the final dataset.

The FBI provided a comprehensive set of incident-level hate crime data, including details on bias types and motivations, criminal acts and types, race or ethnicity, age, and gender of offenders, victims, and arrestees, as well as incident types and locations. The FBI's data collection process involves two steps. First, a responding officer determines whether an incident is suspected to be bias-motivated. If so, the case is then forwarded to a second-level judgment officer or unit to review, and then reported to the FBI Uniform Crime Reporting program [36]. Although Congress passed the Hate Crime Statistics Act on April 23, 1990, mandating the Attorney General to collect hate crime data from law enforcement at all levels, submitting this data to the FBI Uniform Crime Reporting program remains "voluntary" [36]. Consequently, bias-motivated offenses data might not include all the potential AAHIs if some law enforcement agencies do not participate in the program.

To provide an overview of AAHIs during the COVID-19 period, this study used two variables: (1) AAHI counts and (2) the ratio of AAHIs reported in media data to those in law enforcement

XSL•FO
RenderX

data. First, AAHI counts reported in the media were coded as one if a case exhibited bias-motivated acts committed, in whole or in part, because of actual or perceived race or ethnicity, where either: (1) the perpetrator explicitly indicated anti-Asian animus during the act, or (2) contextual evidence strongly suggested that the victim was targeted due to their AAPI identity, as perceived by the AAPI community. Property-related crimes targeting institutions or areas (eg, Chinatown) associated with the AAPI community were also included. Based on these selection criteria, 1288 AAHI cases were collected from media reports. Similarly, in the FBI data, AAHI counts were coded as one based on bias motivation codes assigned to anti-Asian hate intent, resulting in 1086 cases during the same period.

To assess the discrepancy in AAHI counts reported between the two data sources, the study calculated the ratio of AAHI counts reported in the TAAF data to those reported in the FBI data. To avoid having zero values in the denominator, 0.1 was added to both the numerator and the denominator. Thus, the ratio is 1 if both the TAAF and FBI data report 0 AAHI cases. Due to this adjustment, the absolute value of the ratio is not meaningful by itself. However, observing whether the ratio increases or decreases over time or whether it is greater or less than 1 provides valuable insights.

Ethical Considerations

The dataset used in this study did not include any personal identifiers and, therefore, posed minimal risk to participants. Accordingly, the study was determined to be exempt from the Institutional Review Board review at the researcher's institution (23-11-7119). The researchers acquired permission to use data collected by TAAF.

Analysis

For data analyses, this study used STATA (version 18.0; Stata Corp LLC). To illustrate the time trends of AAHI counts from 2020 to 2021, we first aggregated AAHI counts by month and data source. Then, we plot the AAHI counts against each month for each data source. Key events (eg, the first confirmed COVID-19 case in the United States) were also included in the plot to help contextualize and understand the observed patterns.

To illustrate state-level variations in AAHI phenomena, we aggregated AAHI counts by state, year, and data source and then drew heat maps. Additionally, we depicted the ratio of AAHI counts reported between the two sources using heat maps, based on the mean values by state and year. Similarly, county-level analyses were conducted by aggregating AAHI counts by county, year, and data source, followed by creating heat maps. The ratio of AAHI counts reported between the two data sources was also illustrated on heat maps, based on the mean values by county and year.

We then illustrated a bar chart showing the percentage of incident types and incident locations by data source. Due to differences in the terminology used by media sources and FBI data to describe incident types and locations, we standardized the language across both datasets and grouped specific categories into broader ones, as detailed in Tables S1 and S2 in Multimedia Appendix 1. When both the FBI and news media sources reported multiple offense types within a single incident, we coded the first offense type listed as the primary offense, using it to categorize the incident. For example, if an incident was recorded as "Aggravated Assault; Intimidation," it was categorized under Assault. To ensure consistency and reliability, four researchers collaborated on the coding process and cross-checked each categorization. To examine whether incident types and locations differ between data sources, Pearson chi-square statistics were also calculated to test for statistical differences.

Results

Time Trends of AAHI Counts

Figure 1 presents the monthly counts of AAHIs reported in both news media and law enforcement data from January 2020 to December 2021. The data points for news media are represented by blue circles connected by solid lines, while law enforcement data are represented by red diamonds connected by dashed lines. Following the first confirmed COVID-19 case in the United States on January 20, 2020, there was a sharp increase in the number of reported incidents in both data sources. The number of incidents peaks around March 2020, coinciding with a politician's anti-Asian hashtags, immediately following the WHO's declaration of a global pandemic, which sparked controversy. This initial spike reflects a surge in hate incidents as the pandemic began to take hold and public discourse around its origins intensified.

XSL•FO
RenderX

Figure 1. Monthly trends in AAHI counts in the United States, 2020-2021. AAHI: anti-Asian hate incident.

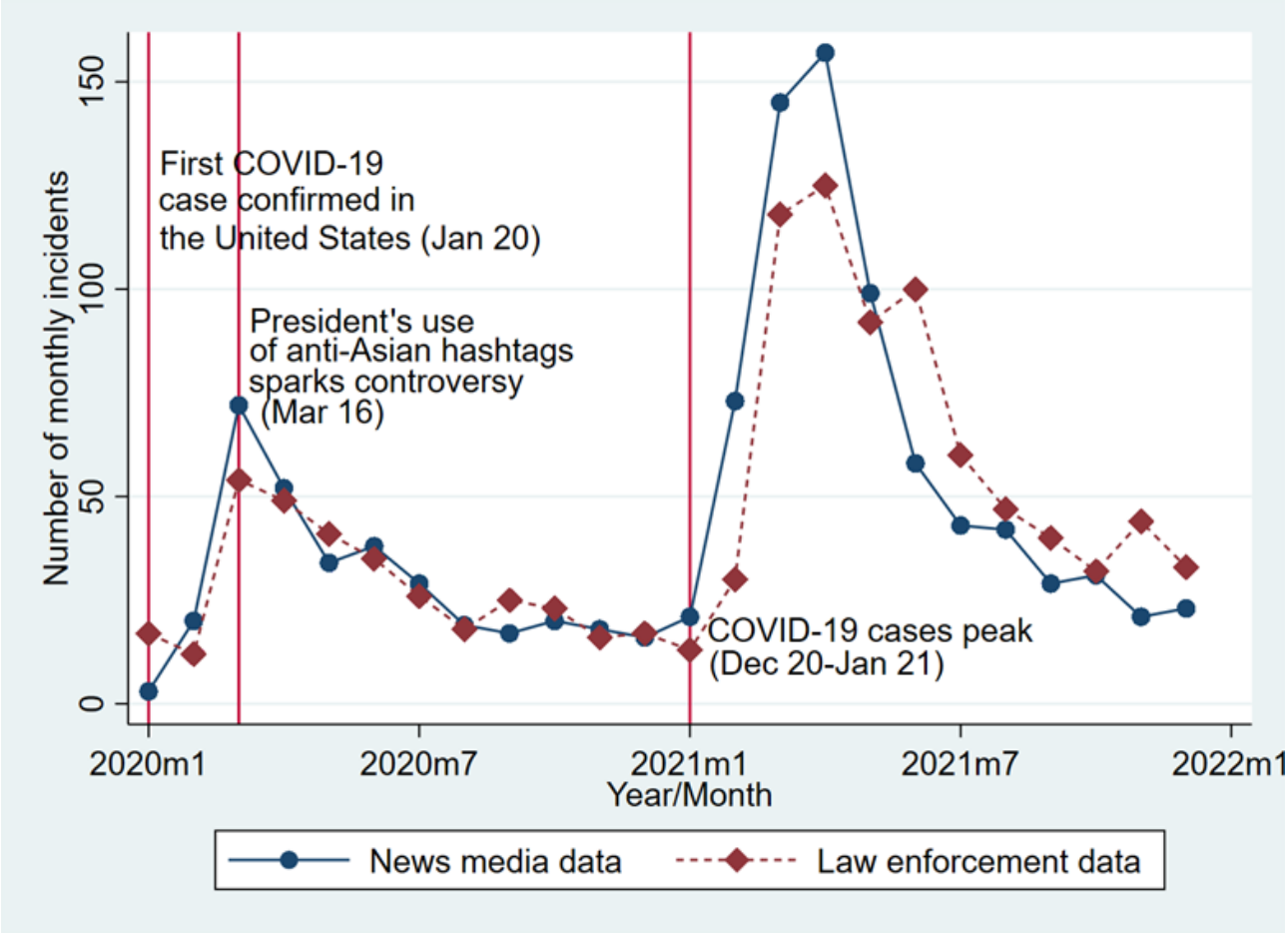

After the initial spike, there is a gradual decline in the number of reported incidents from April 2020 to October 2020. During this period, both data sources show a similar trend. This decrease suggests a temporary reduction in the intensity or visibility of such incidents following the initial outbreak.

A second peak in reported incidents is observed around March and April 2021, followed by the peak of COVID-19 cases in the United States during December 2020 and January 2021. This spike is more pronounced in the news media data compared to the law enforcement data. The AAHIs' peak, followed by the pandemic's peak, suggests that increased stress and fear during this period may have contributed to a resurgence of hate incidents. Following the second peak, both data sources show a noticeable decline in reported incidents throughout 2021. The trends between the two data sources remain relatively aligned, though news media reports generally show higher incident numbers than law enforcement data. This overall decline suggests a reduction in hate incidents as the pandemic situation improved and public attention shifted.

State Variations of AAHI Counts

The analysis of AAHI counts reported in news media for the years 2020 and 2021 reveals several key findings (Figure 2). In 2020, California, Texas, New York, and Washington reported the highest numbers of AAHI, indicated by the darkest green shades on the heat map. The central United States exhibited lower reported counts, with many states in lighter green shades. The Northeast and West Coast regions, along with Texas, showed significant concentrations of reported incidents. In 2021, these high counts persisted in California, Texas, New York, and Washington, maintaining their dark green shading. Additionally, increased reported incidents were observed in states such as Wisconsin and Illinois, which displayed darker shades compared to the previous year. The central United States remained predominantly in lighter shades, though some states experienced increases in reported incidents.

XSL•FO
RenderX

Figure 2. AAHI counts by states, news media data, 2020 and 2021. AAHI: anti-Asian hate incident.

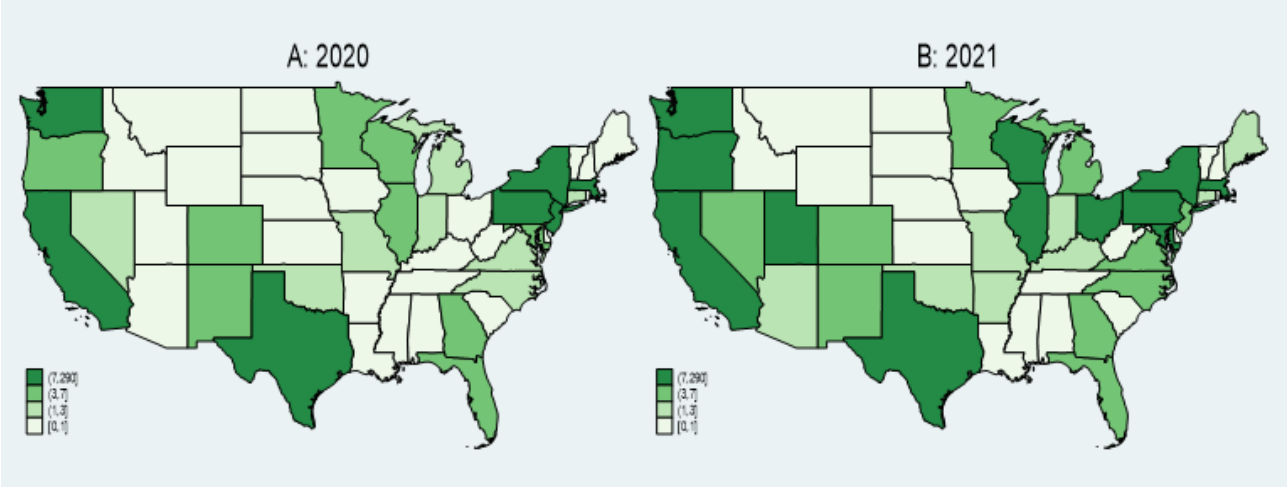

Figure 3 illustrates AAHI counts reported in FBI data across various states for the years 2020 (A) and 2021 (B), respectively. States such as California, Texas, New York, and Washington reported the highest numbers of AAHIs in both 2020 and 2021, as indicated by the darkest blue shades on the maps. This indicates that these states consistently faced significant levels of hate incidents against the AAPIs. In 2021, there was an increase in reported incidents in several states compared to 2020. Notably, states like Tennessee and West Virginia displayed darker shades in 2021, suggesting a rise in reported incidents. The central United States, which predominantly showed lighter shades in 2020, indicating lower counts, saw some states shift to darker shades in 2021. This suggests increased reporting or occurrence of hate incidents in these regions over the year. Parts of the Southeastern United States also exhibited increased reported incidents from 2020 to 2021, with states such as North Carolina becoming more prominent in the data. Along with the consistently high counts in California, Texas, New York, and Washington, the Northeast and parts of the West Coast continued to show significant concentrations of reported incidents in both years.

Figure 3. AAHI counts by state, FBI data, 2020 and 2021. AAHI: anti-Asian hate incident; FBI: Federal Bureau of Investigation.

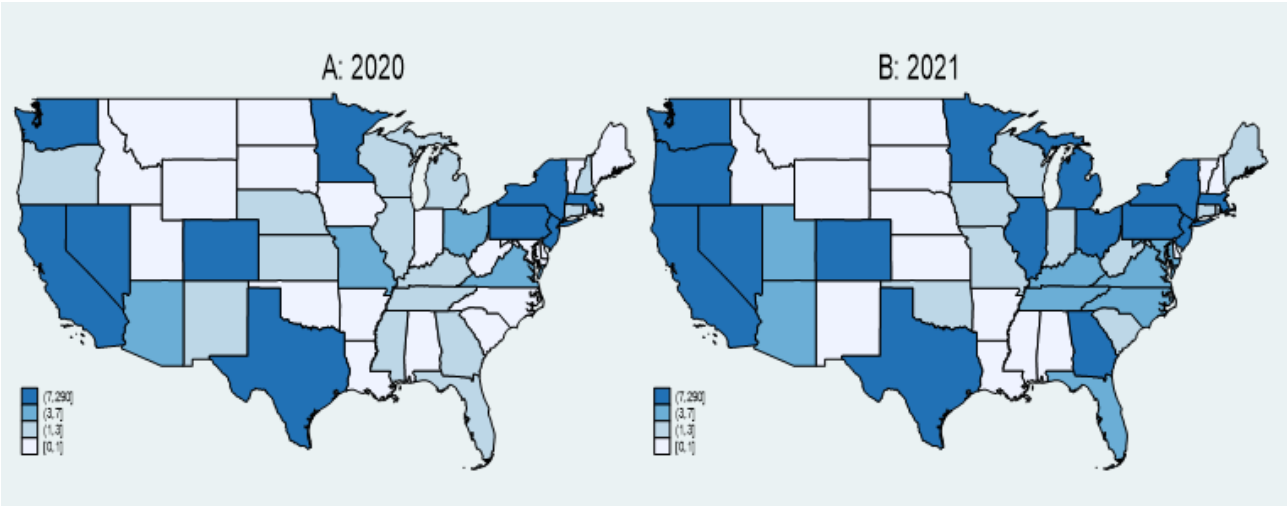

Figure 4 displays the mean ratio of AAHI counts reported between news media and FBI data for 2020 (A) and 2021 (B). The ratio is calculated by dividing monthly county-level AAHI counts reported in news media by those reported in FBI data, with 0.1 added to both the numerator and denominator. We then averaged the ratio by state and by incident year. These heat maps reveal notable trends and disparities between news media and FBI reporting across different states.

XSL•FO
RenderX

Figure 4. Mean ratio of AAHI counts reported by news media to FBI data by state, 2020 and 2021. AAHI: anti-Asian hate incident; FBI: Federal Bureau of Investigation.

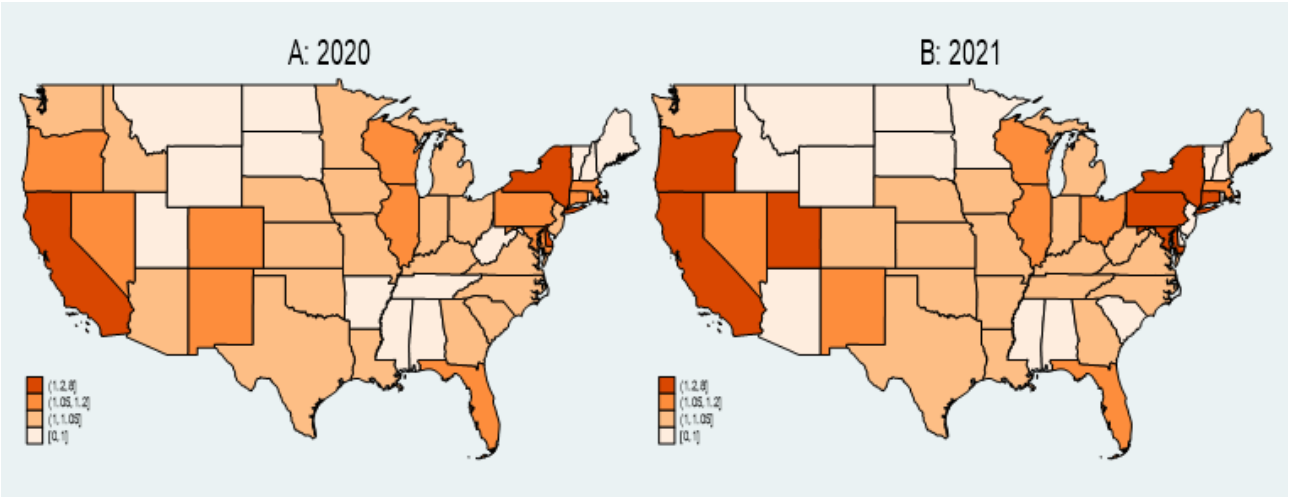

In 2020, states such as California, New York, and Maryland exhibited the highest ratios, indicated by the darkest orange shades on the map. This suggests that these states had a significantly higher number of AAHI incidents reported in the news media than the counts reported by the FBI. Conversely, the central United States and parts of the Southeast generally showed lower ratios, represented by lighter orange and beige shades. This indicates a closer alignment between news media and FBI-reported counts or lower overall counts in these regions.

In 2021, the trend of high ratios persisted in states such as California and New York, which continued to display the darkest orange shades. This indicates ongoing significant discrepancies between news media and FBI-reported counts in these states. Additionally, some states, such as Utah, showed an increase in their ratios compared to 2020, indicated by darker shades in the 2021 heat map. This suggests increased news media reporting relative to FBI data in these states. Overall, the central United States and several Southeastern states maintained lower ratios, with lighter shades similar to the previous year, reflecting more consistency between media and FBI-reported counts or lower levels of reporting overall.

County Variations of AAHI Counts

Figures 5 and 6 provide heat maps illustrating AAHI counts reported in news media across counties for the years 2020 and 2021, respectively. In 2020, the highest counts of AAHI incidents were concentrated in a few key areas, most notably in California, particularly in counties such as Los Angeles. Other notable areas with relatively high counts included some counties in Washington state, New York, and a few scattered regions across the country. However, most counties across the central and southeastern United States reported very few or no incidents, as indicated by the lighter shades or absence of shading on the map.

XSL•FO
RenderX

Figure 5. AAHI counts by county from news media data, 2020. AAHI: anti-Asian hate incident.

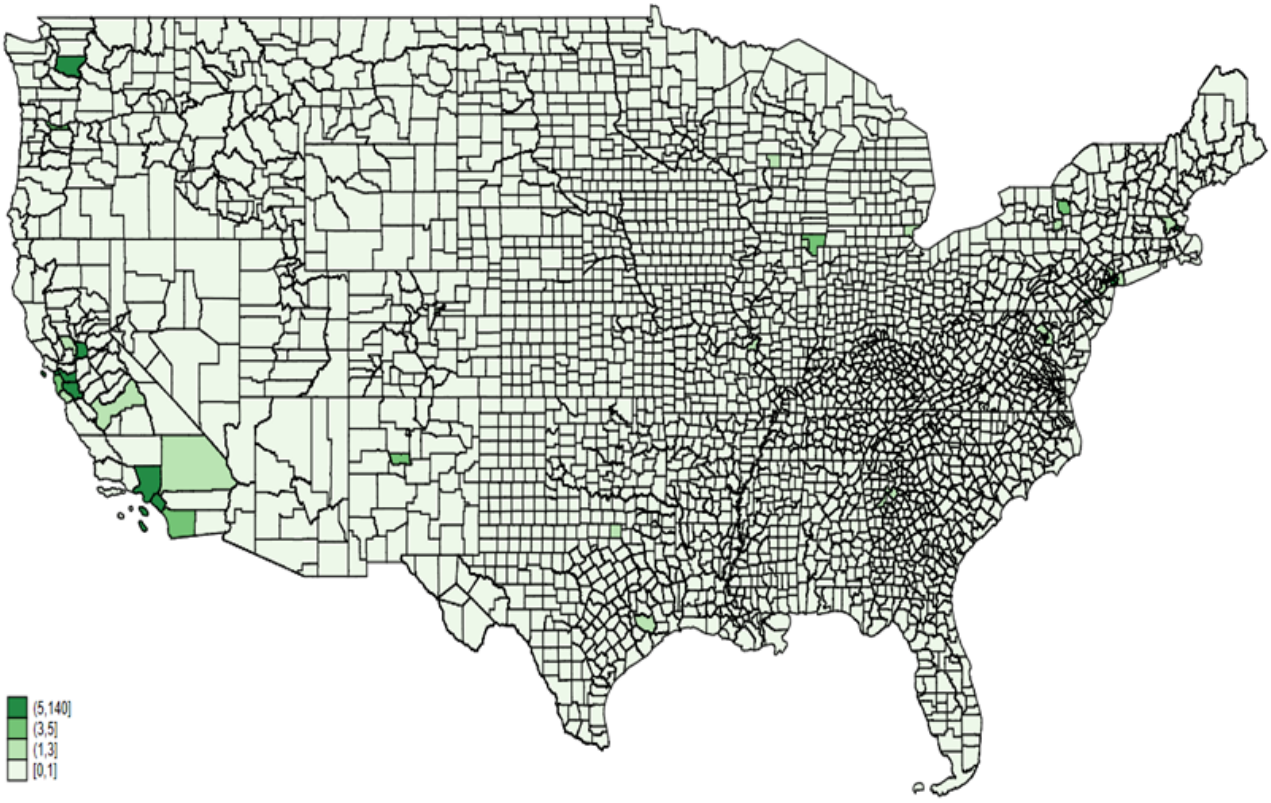

Figure 6. AAHI counts by county from news media data, 2021. AAHI: anti-Asian hate incident.

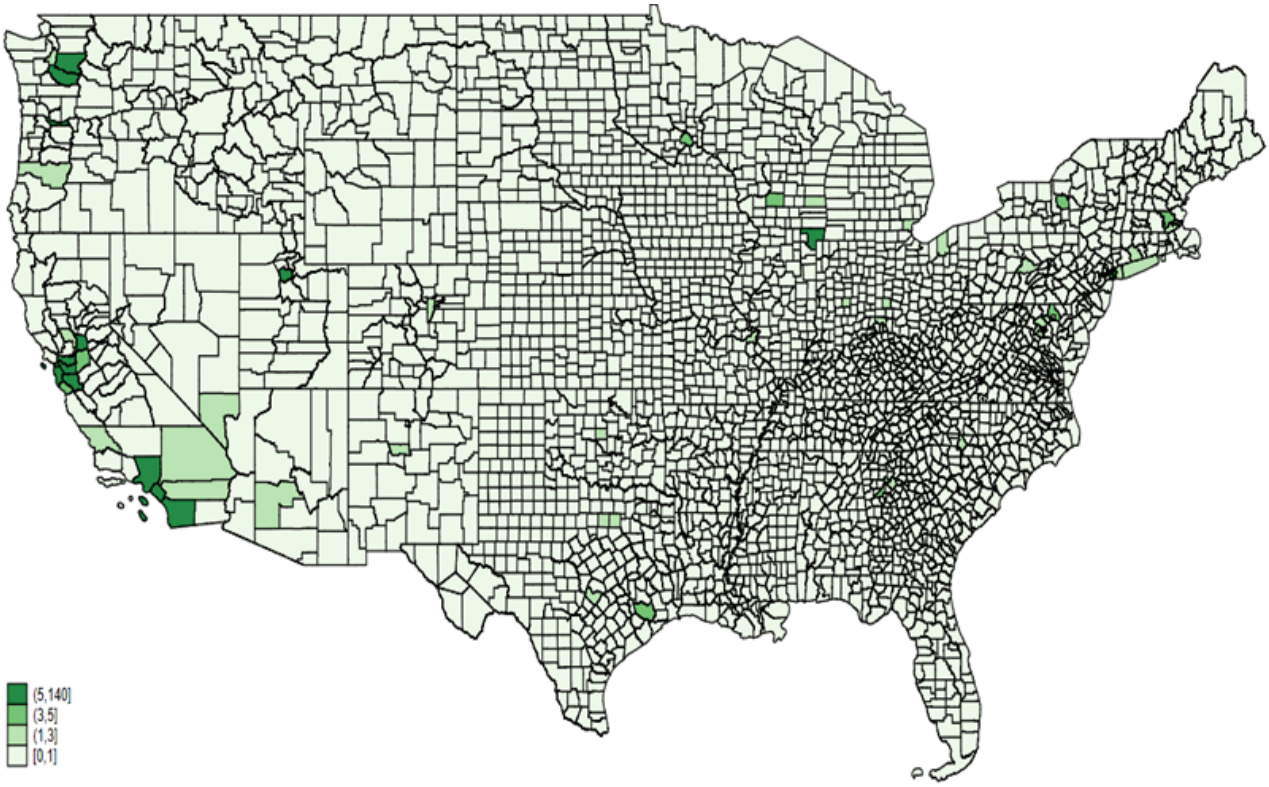

In 2021, the pattern of high counts in California persisted, with Los Angeles County continuing to show significant numbers of incidents. There was a slight increase in the number of counties reporting incidents across various states, suggesting a broader geographical spread of reported incidents. Washington state and New York again showed notable counts, along with increased reporting in some other states such as Colorado and Illinois. Despite this increase, many counties in the central and southeastern United States still reported lower counts than the coastal and more populated regions.

XSL•FO
RenderX

Figures 7 and 8 illustrate AAHI counts recorded by law enforcement across counties for the years 2020 and 2021, respectively. In 2020, the highest counts of AAHI incidents reported in law enforcement data were concentrated in a few key areas, most notably in California, particularly in counties such as Los Angeles and San Francisco. Other areas with notable counts included counties in Washington state and a few scattered regions in the Northeast, such as New York and Pennsylvania. However, the vast majority of counties across the central and southeastern United States reported very few or no incidents, as indicated by the lighter shades or absence of shading on the map. This suggests that incidents were either less frequent or reported in these regions during 2020.

Figure 7. AAHI counts by county from FBI data, 2020. AAHI: anti-Asian hate incident; FBI: Federal Bureau of Investigation.

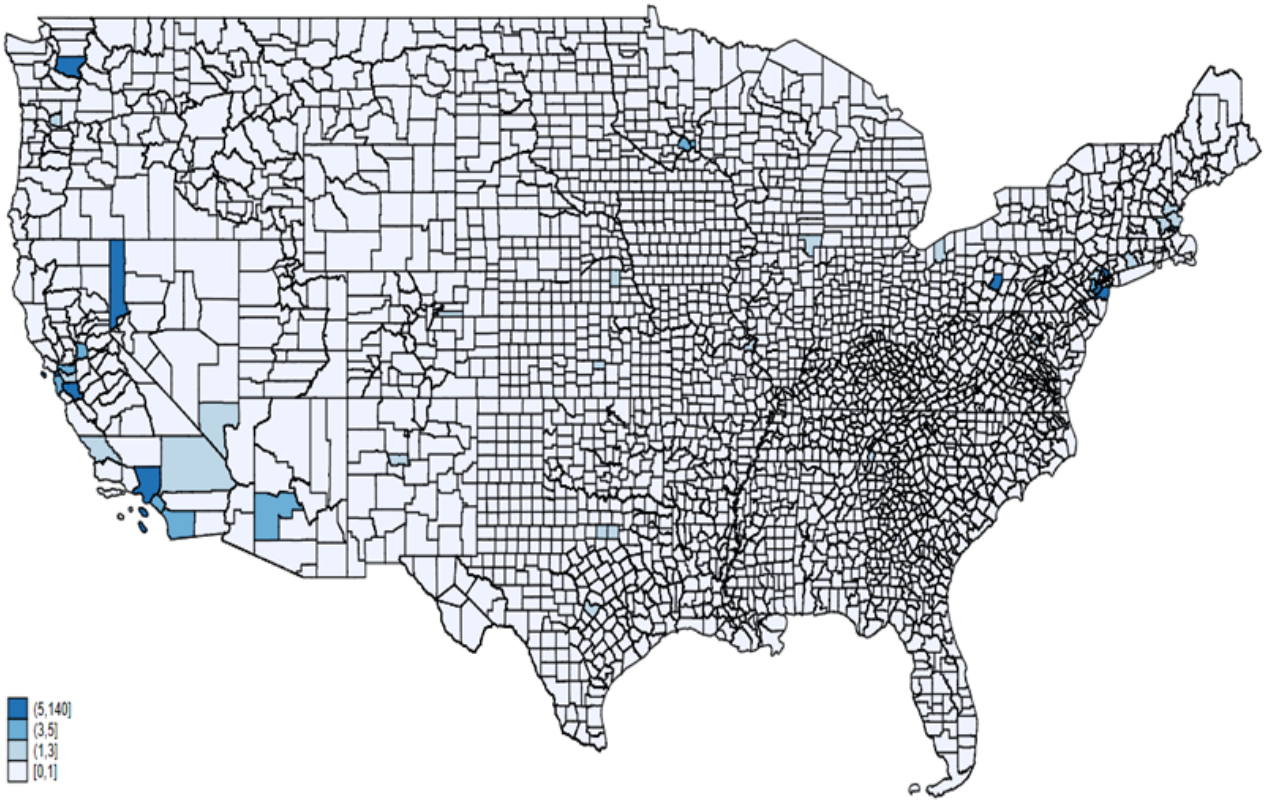

XSL•FO
RenderX

Figure 8. AAHI counts by county from FBI data, 2021. AAHI: anti-Asian hate incident; FBI: Federal Bureau of Investigation.

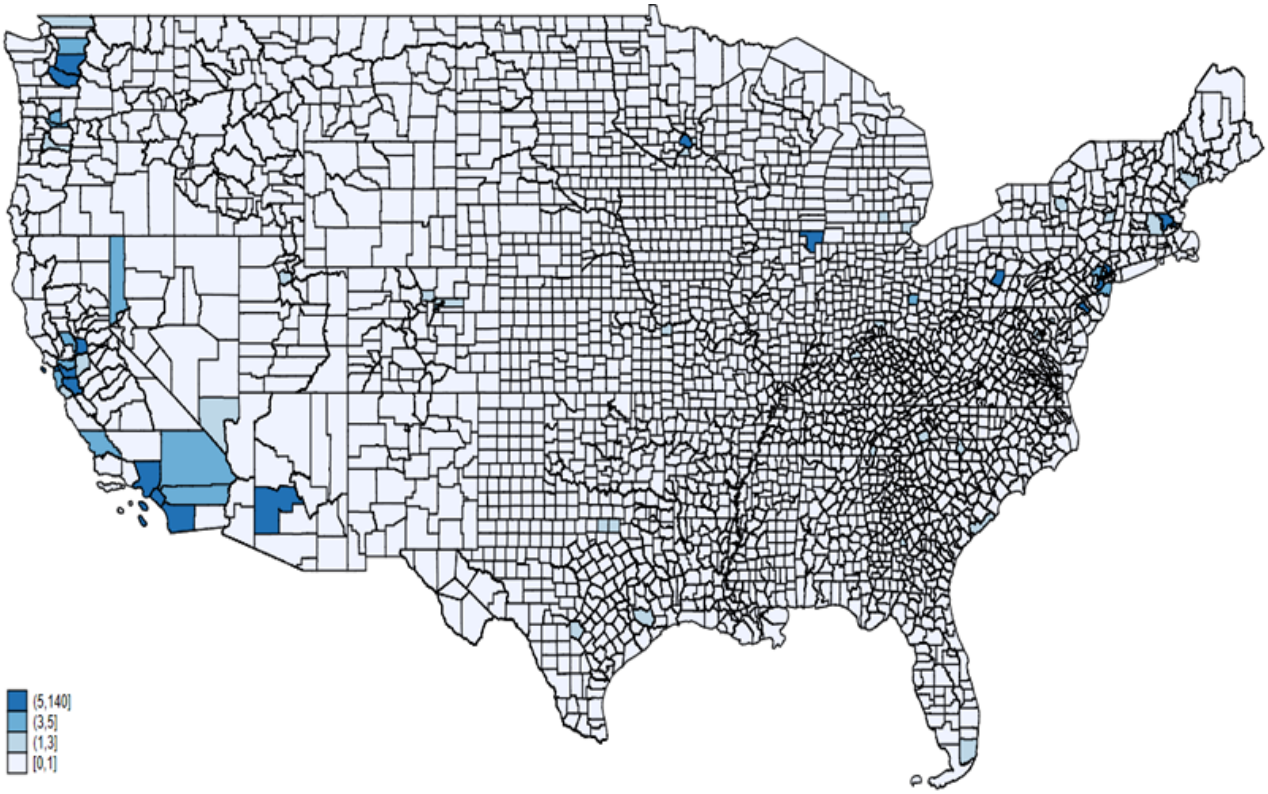

In 2021, the pattern of high counts in California persisted, with counties like Los Angeles and San Francisco continuing to show significant numbers of incidents. A slight increase in the number of counties reporting incidents across various states suggests a broader geographical spread of reported incidents. Washington state and New York again showed notable counts and increased reporting in some other states, such as Illinois and Minnesota. Despite this increase, many counties in the central and southeastern United States still reported lower counts compared to the coastal and more populous regions, indicating a continued disparity in the geographical distribution of reported incidents.

Figures 9 and 10 show the mean ratio of AAHI counts reported in two data sources by county for 2020 and 2021, respectively. In 2020, the highest ratios, represented by the darkest orange shades, are concentrated in a few areas, particularly in California, including counties such as Los Angeles and San Francisco. This indicates a significantly higher number of incidents reported in the news media compared to FBI data. Other areas with notably high ratios include counties in Washington state, Arizona, and a few scattered regions in the Northeast, such as parts of Pennsylvania and New York. Most counties across the central and southeastern United States show lower ratios, indicated by lighter shades, suggesting more consistency between the two data sources or lower overall reporting levels.

XSL•FO
RenderX

Figure 9. Mean ratio of AAHI counts reported by news media to FBI data by county, 2020. AAHI: anti-Asian hate incident; FBI: Federal Bureau of Investigation.

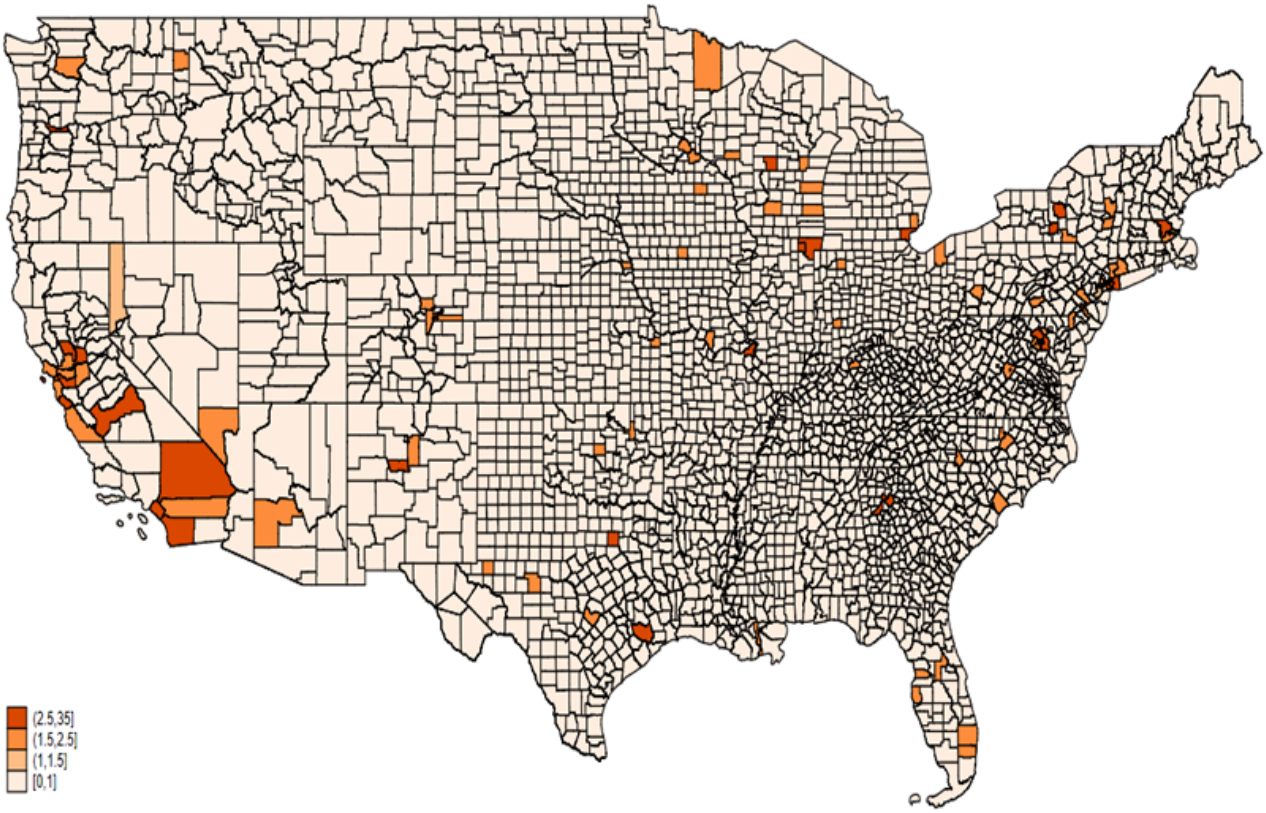

Figure 10. Mean ratio of AAHI counts reported by news media to FBI data by county, 2021. AAHI: anti-Asian hate incident; FBI: Federal Bureau of Investigation.

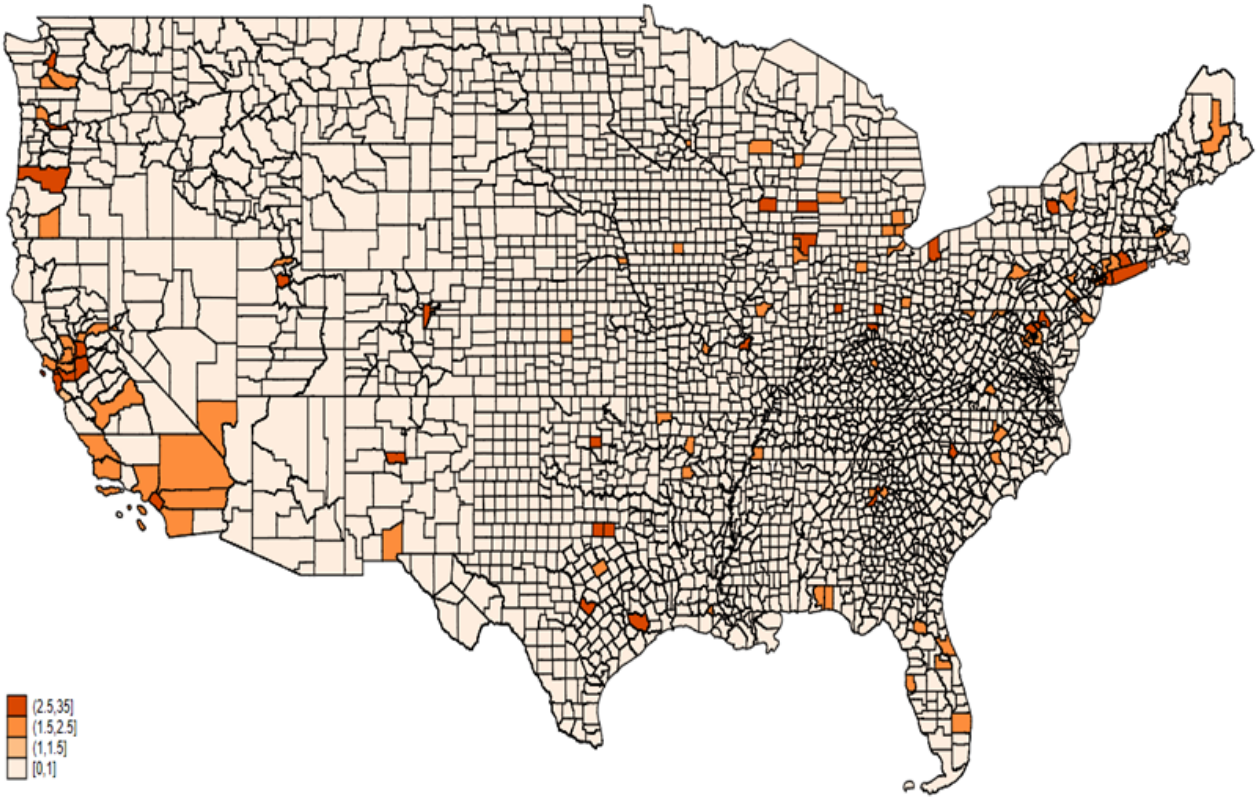

In 2021, the pattern of high ratios persisted in California, with counties like Los Angeles and San Francisco continuing to show significant discrepancies between the news media and FBI data. There was an increase in the number of counties with higher ratios across various states, such as Colorado, Illinois, and parts of the Southeast, indicating a growing disparity in reporting

XSL•FO
RenderX

between the two sources in these regions. The central United States still displayed mostly lower ratios, with some increases in counties within Texas and Arizona, suggesting regional variations in the consistency of reporting.

Incident Types by Data Sources

Table 1 compares the distribution of AAHI types reported in news media versus those reported by the FBI, highlighting differences in the proportions of various incident types between the two sources. In terms of assault incidents, the news media reported that 45% (580/1288) of incidents were assaults, while the FBI reported a slightly higher proportion at 47% (510/1086). This indicates that both sources identify assault as the most common type of incident, with nearly identical proportions. For harassment incidents, there is a noticeable difference between the two sources. The news media reported that 37% (477/1288) of incidents were classified as harassment, whereas the FBI reported a lower proportion of 30% (326/1086). This suggests that the news media reports a higher proportion of harassment incidents than the FBI. For property-related incidents, the FBI reported a significantly higher proportion at 22% (239/1086), compared to just 17% (219/1288) reported by the news media. This indicates that the FBI data includes many more property-related incidents than the news media reports. When we conducted Pearson chi-squares analysis, the value was 32.9072, which is significant at $P<.001$. This suggests that the types of incidents reported statistically differ between the two data sources.

Table 1. Percentage of incident types by data sources.

	News media (n=1288), n (%)	Law enforcement (n=1086), n (%)
Assault	580 (45)	510 (47)
Harassment	477 (37)	326 (30)
Property-related	219 (17)	239 (22)
Shunning	13 (1)	0 (0)
Other	0 (0)	11 (1)

Incident Location Types by Data Sources

Table 2 compares the distribution of AAHIs by location type as reported in news media versus those reported by the FBI, highlighting differences in the proportions of incidents occurring in various settings between the two sources. In the news media reports, the most common location for AAHI incidents was public areas, accounting for 40% (515/1288) of incidents. This is significantly higher than the proportion reported by the FBI, where 33% (358/1086) of incidents occurred in public areas. This discrepancy suggests that news media reports might focus more on incidents occurring in public settings than FBI reports. Businesses were the second most common location for AAHI incidents in news media, with 28% (361/1288) of incidents reported in this setting. In contrast, the FBI reported a slightly lower proportion of 22% (239/1086) for incidents occurring in businesses. Both sources highlight businesses as a major location for AAHI incidents, though the proportions differ. A notable discrepancy is observed in transportation settings, where 13% (167/1288) of incidents were reported in the news media compared to just 2% (22/1086) in FBI data. In contrast, incidents in home or housing settings accounted for 9% (116/1288) of news media reports but 18.5% (201/1086) of FBI reports. Schools were another notable location, but both sources had smaller proportions. The news media reported 3% (39/1288) of incidents occurring in schools, while the FBI reported a slightly lower proportion of 1% (11/1086). Similarly, the news media reported 3% (39/1288) of incidents occurring in colleges or universities, compared to 1% (11/1086) in the FBI data. This indicates that incidents in educational settings are less frequently reported compared to public areas and businesses. Additionally, health care settings and hospitals or clinics were reported more frequently by the FBI (22/1086, 2%) compared to the news media (8/1288, 0.6%). The Pearson chi-square statistic is 214.1602, significant at $P<.001$, implying that AAHI incident location types differ significantly between the two data sources.

Table 2. Percentage of incident locations by data sources.

Locations	News media (n=1288), n (%)	Law enforcement (n=1086), n (%)
Public area	515 (40)	358 (33)
Business	361 (28)	239 (22)
Transportation	167 (13)	22 (2)
Home or housing	116 (9)	201 (18.5)
College or university	39 (3)	11(1)
School	39 (3)	11 (1)
Place of worship	21 (1.6)	7 (0.6)
Health care setting	8 (0.6)	22 (2)
Online	5 (0.4)	11 (1)
Other	10 (0.8)	195 (18)

XSL•FO
RenderX

Discussion

Principal Findings

Our result demonstrates significant fluctuations in the number of reported AAHIs over the course of the COVID-19 pandemic. Key triggers such as the initial outbreak, controversial political statements, and peaks in COVID-19 cases appear to correlate with spikes in reported incidents. Both news media and law enforcement data exhibit similar overall trends. After May 2021, the crossover point in this sample, AAHIs reported in law enforcement data were modestly higher than those in media reports. This shift may be attributed to the Centers for Disease Control and Prevention's announcement that fully vaccinated individuals in the United States no longer needed to wear masks, which potentially led to increased social interactions. Consequently, media coverage of AAHIs might have shifted focus or decreased due to changing public priorities.

The relationship between spikes in reported incidents and key events such as the initial outbreak of COVID-19, controversial political statements, and peaks in COVID-19 cases underscores the influence of societal and political factors in driving hate incidents. This evidence is congruent with previous literature indicating that when prejudice against a minority group is compounded by the recognition of the outgroup's association with significant historical events, negative sentiments and expressions directed toward the outgroup can be intensified [18]. Additionally, our findings suggest that public messaging and political rhetoric may significantly influence the occurrence and reporting of AAHIs. Previous literature indicated that politicians' narratives to tighten the border and curb immigration had increased violence toward specific racial, ethnic, or spiritual communities [37]. Similarly, anti-Muslim tweets from a prominent political figure have been shown to increase xenophobic tweets by his followers, the number of cable news mentions of Muslims, and hate crimes targeting this population [38]. Content analysis has demonstrated a significant increase in tweets blaming Chinese and other Asians for the virus, reinforcing the Yellow Peril and perpetual foreigner stereotypes, following the use of phrases such as "Chinese virus" or "Kung flu" in a prominent politician's remarks [22]. While our analysis is descriptive, it provides valuable context that sounds the alarm for politicians and government officials to recognize the impact of their rhetoric, which can shape public opinion and incite further hostility and violence, such as hate incidents.

In both media and FBI data, states such as California, Texas, New York, and Washington consistently reported the highest numbers of incidents, highlighting lingering challenges in these regions. County-level analyses indicate that both media and FBI data consistently reported the highest numbers of AAHI incidents in Los Angeles and San Francisco counties for both 2020 and 2021. This indicates a significant overlap between media and law enforcement reporting in these areas. The high number of AAHI incidents can be attributed to the significant Asian populations in these regions. In 2019, California had the largest Asian population in the nation, with approximately 6.7 million people, followed by New York (1.9 million), Texas (1.6 million), New Jersey (958,000), and Washington (852,000) [39]. Together, these five states were home to 55% of the Asian population in the United States. Community organizations can leverage these findings to advocate for more resources and support services for AAPI communities in these high-incident areas [40]. This includes mental health services, legal assistance, and community safety programs. The consistent identification of these high-incident areas enables targeted interventions by policy makers, community leaders, and law enforcement, allowing for a more effective allocation of resources to prevent and address AAHI.

Both media and FBI data showed an increase in the geographical spread of reported incidents in 2021 compared to 2020. However, there are discrepancies in the overall number of incidents reported. Media data tend to show a wider distribution of lower-count incidents across more counties, suggesting more extensive coverage or awareness. In contrast, FBI data show fewer counties with reported incidents. Additionally, both datasets consistently show lower counts in the central and southeastern United States, with fewer reported incidents in these regions. These findings highlight the need for further research to explore the underlying reasons for discrepancies in reporting between media and law enforcement. It is essential to examine whether lower counts in certain regions reflect actual prevalence or simply lower reporting practices. Understanding these factors can help develop plans to bridge the gap and ensure comprehensive and full documentation of AAHI.

The authors uncovered notable differences in how incident types are reported by news media versus the FBI. While both sources agree on the high prevalence of assault incidents, the news media reports a substantially higher proportion of harassment. In contrast, the FBI data includes a much higher proportion of property-related incidents. Our results align with previous findings showing that law enforcement data have a higher proportion of property-related crime reporting compared to social or news media data [31,35]. News media may prioritize and highlight incidents that are more sensational or resonate more with public interest through editorial decisions, including those involving harassment. In contrast, property-related incidents, such as vandalism or property damage, are more concrete to document and classify under hate crime statutes. Because the losses are more visible and tangible, individuals are more likely to report these incidents to law enforcement. Additionally, Asian Americans tend to underreport their hate incident experiences to police because they may believe that the damage is minor, even when it is not, or they are unsure if they have been affected [30]. As a result, minor incidents, categorized as harassment, are likely underreported to law enforcement due to these two major reasons.

Our results reveal significant differences in the reported locations of AAHI incidents between news media and FBI data. While both sources identify public areas and businesses as common locations, the news media emphasizes public area incidents more heavily, whereas the FBI data highlights incidents in residential settings and others. News media may prioritize incidents that occur in public areas because these incidents are more likely to capture public attention and provoke strong emotional reactions. Public area incidents often involve multiple witnesses and can be more triggering, making them

XSL•FO
RenderX

more newsworthy. In contrast, individuals may be more likely to report incidents that occur in residential settings to law enforcement due to the personal and private nature of these incidents. Property damage, home invasions, or harassment at home are often perceived as serious threats to personal safety, leading to formal police reports.

In sum, our descriptive analyses reveal a discrepancy in AAHI reporting between the two sources, with the FBI data showing fewer incidents compared to media data. A growing body of research emphasizes the obstacles that hinder the reporting and processing of hate crimes [32,34,41]. A study identified reasons for not reporting violent bias experiences to police, including perceptions that reports were not important to police, that they were personal matters, that police were inefficient, that police were biased, that it was not clear it was a crime, and that individuals were advised not to report [42]. Another study described that the underreporting of crimes was related to fear of consequences, a lack of trust in law enforcement agencies, and previous police discrimination experiences [32,34]. Tessler et al [41] posited that recent Asian immigrants may not fully grasp and navigate the legal system or know how to report crimes, especially those charged as hate crimes. Studies also suggest that Asian Americans might engage in social withdrawal as a defensive strategy driven by fear, leading to a self-imposed limitation on their movements and a reluctance to report instances of hate crime experiences [2,30]. Additionally, this underreporting may be attributed to the varying definitions and legal frameworks for hate crimes across different states. A study reported that states with more comprehensive definitions of hate crimes report a greater incidence of hate incidents [43]. This underscores the need for more comprehensive and consistent hate crime reporting and classification to better understand and address hate-fueled attacks.

Practice and Policy Implications

Our findings underscore the impact of societal and political influences on AAHIs and their reporting in both media and law enforcement data, leading to the following practice and policy implications. First, practitioners and policy makers should prioritize building stronger relationships with AAPI communities to encourage the reporting of hate crimes. This can be achieved by fostering trust through community outreach programs, providing organized and detailed information on how to report incidents, and ensuring support systems are in place for those affected. Addressing barriers to reporting, such as fear of retaliation or distrust in law enforcement, is essential for increasing reporting rates. Additionally, given the observed discrepancies between media and law enforcement reports, it is crucial to enhance mechanisms for reporting and tracking AAHIs. This includes developing standardized criteria for reporting and classifying hate crimes across jurisdictions to improve consistency and accuracy. Implementing a consistent and organized reporting system, along with educating law enforcement personnel about this system, could help reconcile discrepancies between media and law enforcement data and ensure comprehensive documentation of incidents.

From a public health standpoint, our findings contribute to building awareness of AAHIs, which can inform and guide public health professionals in addressing their mental and physical health consequences. Racially motivated hate not only harms those affected but also disrupts the social fabric, affecting their families, friends, and broader community. Understanding how hate crimes and incidents are identified and framed is a crucial step toward building health professionals' capacity to recognize trauma associated with hate crimes and provide culturally competent care. Additionally, public health campaigns can leverage this awareness to foster solidarity, encourage community-based reporting, and promote a sustained response to mitigate the impact of hate crimes. At the policy level, it is vital to integrate anti-hate crime initiatives into broader public health strategies. Policy makers should advocate for laws that provide adequate protection for AAPI individuals and communities, ensuring resources for mental health care and community healing.

Limitations

This study has several limitations that should be acknowledged. First, the reliance on two primary data sources, media reports compiled by TAAF and law enforcement data provided by the FBI, may introduce inherent biases. Media reports could overrepresent sensational or newsworthy incidents, while law enforcement data might underrepresent incidents due to underreporting by individuals affected or inconsistencies in hate crime classification across jurisdictions. The use of open-source media data without standardized criteria for article selection may result in observation bias, potentially leading to an incomplete representation of relevant incidents. Moreover, online hate incidents, which are increasingly significant, fall outside the scope of this study. Second, the study's descriptive analysis does not account for potential confounding variables, such as pandemic-related countermeasures (eg, lockdowns and vaccine rollout) that could influence the observed trends and discrepancies in AAHI reporting. Variations in state laws, differences in local media coverage, and the varying capacity of law enforcement agencies to investigate and report hate crimes were not controlled for, which could contribute to discrepancies between the data sources. Third, the study focuses solely on reported incidents, missing the experiences of AAPIs who may have faced discrimination or hate incidents but did not report them. This limitation suggests that the actual prevalence of AAHIs could be higher than what is documented. Additionally, excluding social media data may have resulted in an incomplete picture of the AAHI landscape during the pandemic. Furthermore, although the dataset includes English-language content from AAPI-focused outlets such as AsAmNews and Asian Dawn, non-English articles were excluded. This may lead to underrepresentation of incidents covered only in other languages. Finally, the generalizability of the findings is limited by the focus on the COVID-19 pandemic period. While the study provides valuable insights into AAHI trends during this specific timeframe, the findings may not fully apply to other periods or the broader context of anti-Asian sentiment outside of a pandemic setting. Future research should address these limitations by incorporating more data sources, controlling for additional variables, and extending the analysis to other periods and social contexts.

XSL•FO
RenderX

Conclusions

This study provides critical insights into the reporting discrepancies and trends of AAHIs during the COVID-19 pandemic, highlighting the complex interplay between societal, political, and media influences on the documentation of hate crimes. Our findings reveal significant fluctuations in reported AAHIs corresponding with key events, such as the onset of the pandemic, peaks in COVID-19 cases, and politically charged rhetoric, underscoring the impact of public discourse on the prevalence and visibility of hate crimes. The observed discrepancies between media reports and law enforcement data, particularly in incident types and locations, suggest that the current methods of documenting and reporting these incidents are insufficient and may fail to capture the full scope of the issue.

Moreover, the geographical analysis indicates persistent high counts of AAHIs in states and counties with significant AAPI populations, emphasizing the need for targeted interventions in these areas. However, the disparities between media and law enforcement data across different regions and the underreporting of incidents in certain states and counties underscore the necessity for a more standardized and comprehensive approach to hate crime documentation. Such an approach would improve the accuracy and reliability of data, enabling policy makers and community leaders to develop more effective strategies to combat hate crimes.

Addressing the discrepancies in AAHI reporting is paramount for advancing our understanding of hate crimes and fostering a safer environment for AAPI communities. Future research should focus on enhancing the consistency of hate crime data collection across various sources, while also exploring the underlying reasons for underreporting and regional disparities. By implementing standardized reporting systems and fostering stronger community-law enforcement relationships, we can work toward a more accurate representation of hate crimes, ultimately contributing to the development of informed and effective policies aimed at eradicating anti-Asian hate and promoting social justice.

Acknowledgments

We extend our gratitude to The Asian American Foundation (TAAF) for providing the datasets and invaluable insights that shaped this Data Story. We also express our special thanks to DataFace for their critical contributions to the development of the Data Story. The researchers conducted final editing and grammar checks with the assistance of ChatGPT-4 [44]. This study was supported by the Robert Wood Johnson Foundation's Data Visualization of Structural Racism and Place Fund. The perspectives articulated in this manuscript are those of the authors and do not necessarily represent those of the Foundation.

Data Availability

The datasets used in this study are publicly available in the FBI UCR repository [45] and the TAAF repository [46].

Conflicts of Interest

None declared.

Multimedia Appendix 1
The categorization of incident types and locations by data sources.
[DOCX File , 20 KB - formative_v9i1e70881_app1.docx]

References

1. Han S, Riddell JR, Piquero AR. Anti-Asian American hate crimes spike during the early stages of the COVID-19 pandemic. J Interpers Violence 2023;38(3-4):3513-3533 [FREE Full text] [doi: 10.1177/08862605221107056] [Medline: 35657278]
2. Lantz B, Wenger MR. Anti-Asian xenophobia, hate crime victimization, and fear of victimization during the COVID-19 pandemic. J Interpers Violence 2023;38(1-2):NP1088-NP1116 [FREE Full text] [doi: 10.1177/08862605221086651] [Medline: 35544766]
3. Roberto KJ, Johnson AF, Rauhaus BM. Stigmatization and prejudice during the COVID-19 pandemic. Adm Theory Prax 2020;42(3):364-378. [doi: 10.1080/10841806.2020.1782128]
4. Vidgen B, Botelho A, Broniatowski D, Guest E, Hall M, Margetts H. Detecting East Asian prejudice on social media. ArXiv. Preprint posted online on May 8, 2020 2020. [doi: 10.48550/arXiv.2005.03909]
5. Lee E. The "Yellow Peril" and Asian exclusion in the Americas. Pac Hist Rev 2007;76(4):537-562. [doi: 10.1525/phr.2007.76.4.537]
6. Lui MT. The Chinatown Trunk Mystery: Murder, Miscegenation, and Other Dangerous Encounters in Turn-of-the-Century. New York City: Princeton University Press; 2005.
7. Chinese immigration and the Chinese Exclusion Acts. Office of the Historian, Foreign Service Institute. URL: https://history.state.gov/milestones/1866-1898/chinese-immigration [accessed 2025-08-18]
8. Stanciu C. Native acts, immigrant acts: citizenship, naturalization, and the performance of civic identity during the progressive era. J Gilded Age Prog Era 2020;20(2):252-276. [doi: 10.1017/s1537781420000080]

XSL•FO
RenderX

9. Luther CA. Reflections of cultural identities in conflict. Journalism Hist 2019;29(2):69-81. [doi: 10.1080/00947679.2003.12062623]
10. Key facts about Asian Americans, a diverse and growing population. Pew Research Center. 2021. URL: https://www.pewresearch.org/short-reads/2021/04/29/key-facts-about-asian-americans/ [accessed 2021-04-29]
11. Cheryan S, Monin B. "Where are you really from?": Asian Americans and identity denial. J Pers Soc Psychol 2005;89(5):717-730. [doi: 10.1037/0022-3514.89.5.717] [Medline: 16351364]
12. Li Y, Nicholson HL. When "model minorities" become "yellow peril"—othering and the racialization of Asian Americans in the COVID-19 pandemic. Sociol Compass 2021;15(2):e12849 [FREE Full text] [doi: 10.1111/soc4.12849] [Medline: 33786062]
13. Lee A. Model minority. In: Weiss G, Murphy AV, Salamon G, editors. 50 Concepts for a Critical Phenomenology. Evanston: Northwestern University Press; 2019:231-236.
14. The rise of Asian Americans. Pew Research Center. 2012. URL: https://web.archive.org/web/20160305005915/http://www.pewsocialtrends.org/files/2013/01/SDT_Rise_of_Asian_Americans.pdf [accessed 2025-08-18]
15. Weis L. Identity formation and the processes of "Othering": unraveling sexual threads. J Educ Found 1995;9(1):17 [FREE Full text]
16. Gover AR, Harper SB, Langton L. Anti-Asian hate crime during the COVID-19 pandemic: exploring the reproduction of inequality. Am J Crim Justice 2020;45(4):647-667 [FREE Full text] [doi: 10.1007/s12103-020-09545-1] [Medline: 32837171]
17. Jefferson T. What is racism? Othering, prejudice and hate-motivated violence. Int J Crime Justice Soc Democr 2015;4(4):120-135. [doi: 10.5204/ijcjsd.v4i4.261]
18. Byers BD, Jones JA. The impact of the terrorist attacks of 9/11 on anti-islamic hate crime. J Ethn Crim Justice 2007;5(1):43-56. [doi: 10.1300/j222v05n01_03]
19. Disha I, Cavendish JC, King RD. Historical events and spaces of hate: hate crimes against Arabs and Muslims in post-9/11 America. Soc Probl 2011;58(1):21-46. [doi: 10.1525/sp.2011.58.1.21]
20. John T, Guy J. A visual guide to the Wuhan coronavirus. CNN. 2020. URL: https://www.cnn.com/2020/01/22/world/wuhan-coronavirus-visual-guide-intl/index.html [accessed 2025-08-18]
21. Palmer E. 'Chinese coronavirus' trends as Kevin McCarthy accused of being racist over COVID-19 remarks. Newsweek. 2020. URL: https://www.newsweek.com/chinese-coronavirus-kevin-mccarthy-tucker-carlson-covid-twitter-1491383 [accessed 2025-08-18]
22. Zheng J, Zompetti JP. ‘I’m not a virus’: Asian hate in Donald Trump’s rhetoric. Asian J Commun 2023;33(5):470-503. [doi: 10.1080/01292986.2023.2246509]
23. Gomera M. How to prevent outbreaks of zoonotic diseases like COVID-19. Aljazeera. URL: https://www.aljazeera.com/indepth/opinion/prevent-outbreaks-zoonotic-diseases-covid-19-200404130822685.html [accessed 2020-04-26]
24. Mackenzie JS, Smith D. COVID-19: a novel zoonotic disease caused by a coronavirus from China: what we know and what we don't. Microbiol Aust 2020:MA20013. [doi: 10.1071/MA20013] [Medline: 32226946]
25. Hate crime statistics. Federal Bureau of Investigation. 2021. URL: https://www.justice.gov/crs/highlights/2021-hate-crime-statistics [accessed 2025-08-18]
26. Hate crime statistics, 2019. Federal Bureau of Investigation. 2021. URL: https://www.fbi.gov/news/press-releases/fbi-releases-2019-hate-crime-statistics [accessed 2025-08-18]
27. Hate crime statistics, 2020. Federal Bureau of Investigation. 2022. URL: https://www.justice.gov/crs/highlights/2020-hate-crimes-statistics [accessed 2025-08-18]
28. Federal emergency authorities. MACPAC. 2021. URL: https://tinyurl.com/2a7d8a6b [accessed 2021-03-18]
29. McNamara LA, Wiegand RE, Burke RM, Sharma AJ, Sheppard M, Adjemian J, et al. Estimating the early impact of the US COVID-19 vaccination programme on COVID-19 cases, emergency department visits, hospital admissions, and deaths among adults aged 65 years and older: an ecological analysis of national surveillance data. Lancet 2022;399(10320):152-160. [doi: 10.1016/s0140-6736(21)02226-1]
30. Lantz B, Wenger MR. Are Asian victims less likely to report hate crime victimization to the police? Implications for research and policy in the wake of the COVID-19 pandemic. Crime Delinq 2021;68(8):1292-1319. [doi: 10.1177/00111287211041521]
31. Curiel RP, Cresci S, Muntean CI, Bishop SR. Crime and its fear in social media. Palgrave Commun 2020;6(1). [doi: 10.1057/s41599-020-0430-7]
32. Erentzen C, Schuller R. Exploring the dark figure of hate: experiences with police bias and the under-reporting of hate crime. Can J Criminol Crim Justice 2020;62(2):64-97. [doi: 10.3138/cjccj.2019-0035]
33. Feldman JM, Gruskin S, Coull BA, Krieger N. Correction: quantifying underreporting of law-enforcement-related deaths in United States vital statistics and news-media-based data sources: a capture-recapture analysis. PLoS Med 2017;14(10):e1002449 [FREE Full text] [doi: 10.1371/journal.pmed.1002449] [Medline: 29073133]
34. Vergani M, Navarro C. Hate crime reporting: the relationship between types of barriers and perceived severity. Eur J Crim Policy Res 2021;29(1):111-126. [doi: 10.1007/s10610-021-09488-1]
35. Chadee D, Chadee M. Media and fear of crime: an integrative model. In: Weiss G, Murphy AV, Salamon G, editors. Psychology of Fear, Crime and the Media. Milton Park: Psychology Press; 2015:58-78.

XSL•FO
RenderX

36. Global Law Enforcement Support Section (GLESS), Crime and Law Enforcement Statistics Unit (CLESU). Hate crime data collection guidelines and training manual. Criminal Justice Information Services (CJIS) Division Uniform Crime Reporting (UCR) Program. 2022. URL: https://www.justice.gov/hatecrimes/resource/hate-crime-data-collection-guidelines-and-training-manual-version-30 [accessed 2022-03-01]
37. Hodwitz O, Massingale K. Rhetoric and hate crimes: examining the public response to the Trump narrative. Behav Sci Terror Polit Aggress 2021;15(3):303-320. [doi: 10.1080/19434472.2021.1936121]
38. Müller K, Schwarz C. From hashtag to hate crime: twitter and antiminority sentiment. Am Econ J Appl Econ 2023;15(3):270-312. [doi: 10.1257/app.20210211]
39. Budiman A, Ruiz NG. Key facts about Asian Americans, a diverse and growing population. Pew Research Center. Washington, D.C; 2021. URL: https://www.pewresearch.org/short-reads/2021/04/29/key-facts-about-asian-americans/ [accessed 2025-08-18]
40. Francisco C, Cruz M, Phuong A, Jeung R, Yoo G. Healing in community and responding with leadership: addressing the pandemic and anti-Asian hate through community service learning. AAPI Nexus Policy Pract Community 2022;19(1-2):A [FREE Full text]
41. Tessler H, Choi M, Kao G. The anxiety of being Asian American: hate crimes and negative biases during the COVID-19 pandemic. Am J Crim Justice 2020;45(4):636-646 [FREE Full text] [doi: 10.1007/s12103-020-09541-5] [Medline: 32837158]
42. Pezzella FS, Fetzer MD, Keller T. The dark figure of hate crime underreporting. Am Behav Sci 2019:000276421882384. [doi: 10.1177/0002764218823844]
43. Stacey M. The effect of law on hate crime reporting: the case of racial and ethnic violence. Am J Crim Just 2015;40(4):876-900. [doi: 10.1007/s12103-015-9289-3]
44. OpenAI. ChatGPT-4. Large language model. 2023. URL: https://openai.com/chatgpt
45. Federal Bureau of Investigation. The Hate Crime Master File, 2020 and 2021. FBI Crime Data Explorer. URL: https://cde.ucr.cjis.gov/LATEST/webapp/#/pages/explorer/crime/hate-crime
46. The Asian American Foundation. Media Reporting. GitHub. URL: https://github.com/taaforg/media-reporting

Abbreviations

AAHI: anti-Asian hate incident
AAPI: Asian American and Pacific Islander
FBI: Federal Bureau of Investigation
TAAF: The Asian American Foundation
WHO: World Health Organization

Edited by A Mavragani; submitted 04.01.25; peer-reviewed by L Fang, T Sugawara; comments to author 05.03.25; revised version received 28.04.25; accepted 02.06.25; published 17.09.25.

Please cite as:
Yoon YJ, Shin SH, Kim D, Lee HY
Comparing Media and Law Enforcement Reports on Anti-Asian Hate Incidents During the COVID-19 Pandemic: Data Visualization Approach
JMIR Form Res 2025;9:e70881
URL: https://formative.jmir.org/2025/1/e70881
doi:10.2196/70881
PMID:

©Young Ji Yoon, Su Hyun Shin, Dongwook Kim, Hee Yun Lee. Originally published in JMIR Formative Research (https://formative.jmir.org), 17.09.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Dental Undergraduate Students' Perceptions of Blended Learning in the COVID-19 and Post–COVID-19 Years: Survey Study

Xin Li[1], DSC; Li Peng[1], MSc; Heng Tu[1], MM; Lijia Liu[1], MM; Xianglong Han[1], DDS; Ling Ye[1], DDS; Jing Yang[1], DDS

State Key Laboratory of Oral Diseases, National Clinical Research Center for Oral Diseases, West China Hospital of Stomatology, Sichuan University, Chengdu, Sichuan, China

Corresponding Author:
Jing Yang, DDS
State Key Laboratory of Oral Diseases, National Clinical Research Center for Oral Diseases, West China Hospital of Stomatology, Sichuan University
No. 14 Renminnan Rd.
Chengdu, Sichuan, 610041
China
Phone: 86 18084868648
Email: yangj7@scu.edu.cn

Abstract

Background: The outbreak of the COVID-19 pandemic created significant challenges but also a unique opportunity, accelerating the evolution of higher education, including dental education. This encouraged dental education to adopt more flexible modes like blended learning.

Objective: This study aimed to explore senior undergraduate dental students' views on blended learning during and after the COVID-19 pandemic and to identify modifiable factors influencing their engagement.

Methods: A survey was conducted among final-year undergraduate students at a top-ranking dental school in mainland China during the fall semesters of 2020-2021 and 2023-2024. The survey assessed satisfaction with blended learning, preferences for engagement, strengths compared to purely online or offline teaching, and factors influencing engagement during and after the pandemic.

Results: Response rates were 75% (85/114) in 2020 and 73% (47/64) in 2023. Blended learning was used in 53% (26/49) of evaluated courses. High satisfaction was reported by 82% (93/114) in 2020 and 59% (38/64) in 2023, with significant differences between high- and low-satisfaction groups (P<.001). Satisfaction with specific course types and learning activities was analyzed. Factors associated with higher satisfaction were evaluated using Pearson correlation. Students acknowledged the strengths of blended learning over online- or offline-only formats. In total, 70% (80/114) in 2020 and 61% (39/64) in 2023 expressed a desire to participate in blended dental education. Factors decreasing engagement included unstable technical support (68/114, 60% in 2020 vs 26/64, 41% in 2023), poor online-offline integration (58/114, 51% vs 34/64, 53%), lack of motivation (51/114, 45% vs 24/64, 38%), and insufficient teacher-student interaction (44/114, 39% vs 20/64, 31%). Factors increasing engagement included high-quality learning materials (76/114, 67% vs 43/64, 67%) and improved technical environments (62/114, 54% vs 35/64, 55%).

Conclusions: Final-year dental students were generally satisfied with blended learning and recognized its strengths compared to purely online or offline formats, both during and after the pandemic. More efforts are required to enhance students' potential engagement in blended learning for the future.

(JMIR Form Res 2025;9:e63453) doi:10.2196/63453

KEYWORDS

blended learning; undergraduate dental education; student satisfaction; dental student; Chinese university; COVID-19

Introduction

Undergraduate dental education aims to provide dental students with basic dental sciences and high-quality clinical practice to guarantee their competency [1,2]. Learning modes in dental education in China have continuously developed over the past decades. Generally, undergraduate dental courses are concentrated in the last 2 years, especially the final year, which mainly focuses on intensive and comprehensive training to

XSL•FO
RenderX

develop students' competence in clinical practice [2,3]. Traditional undergraduate dental education relies on textbook-based learning, focusing on classic textbooks in a specialized field for knowledge acquisition, but some of this information may be outdated, and students may develop fixed ways of thinking [2]. In face-to-face learning, students receive lectures from teachers in the classroom and usually acquire knowledge passively [4]. To alter the teacher-centered mode, problem-based learning and case-based learning questionnaires have been introduced, which support a learner-centered mode to foster active learning and clinical decision-making [5-7]. In recent years, advances in internet infrastructure have broadened the availability of internet services and prompted the application of online learning in modern dental education, exhibiting comparable or even better effectiveness compared to traditional face-to-face learning in dental education [8,9].

Blended learning, a learning mode that has grown increasingly popular in higher education since the 1990s [10-13], is generally defined as the integration of face-to-face learning with online learning [14]. This learning mode is designed to fully harness the merits of online learning and offline learning, enhance the effectiveness and delivery of educational systems, and facilitate learning activation and support [12,15]. While heterogeneity among different surveys exists, influencing factors of 3 dimensions, namely, learning dimensions, teaching dimensions, and curriculum dimensions, have been identified as accounting for student satisfaction with blended learning experiences in higher education in China [16]. Hitherto, only a few panel studies have focused on students' opinions about the implementation of blended learning in dental education compared with face-to-face learning or online learning modes. For example, in early preclinical dental education for prosthetic dentistry, students highly evaluated blended learning with high satisfaction for learning content and outcomes, and appreciated blended learning for the achievement of learning objectives with the aid of e-learning tools [17]. In third-year undergraduate dental education for orthodontics, blended learning combining the use of plaster models with digital 3D models received positive feedback from students, and the learning performance outcomes were highly evaluated [18]. In final-year undergraduate dental education for oral radiology, dental students appreciated the benefits of blended learning in effectiveness, motivation, and engagement compared to the face-to-face only mode [19]. These findings are encouraging regarding the value of blended learning in undergraduate dental curricula.

The outbreak of COVID-19, an infectious disease caused by SARS-CoV-2, started in China in December 2019, was declared by the World Health Organization as a public health emergency of international concern in January 2020, and was assessed as a pandemic in March 2020 [20,21]. The immediate impacts of the COVID-19 pandemic on medical education are largely attributed to the modification of the educational calendar, online shifts of preclinical theoretical curricula, limitations on clinical rotations, and reliance on e-learning technology for clinical practical training [7]. By contrast, students were faced with unexpected challenges related to the COVID-19 pandemic [22,23]. Studying the learning mode differences between the COVID-19 pandemic and the post–COVID-19 years might provide valuable information for medical education. The competency of health professions students is also important [24], although it was not included in this study.

Indeed, interpreting the effects of the COVID-19 pandemic and postpandemic period on the learning outcomes and competency acquisition of students could provide clues for future dental education [25,26]. Existing studies on dental students' self-evaluation of blended learning during the COVID-19 pandemic have reported that most students were very satisfied with the blended learning mode for improving their learning effectiveness, learning efficiency, and personal learning pace [27]. However, disadvantages of blended learning, such as limited social contact, have also been disclosed [27]. The blended mode used in examinations during the COVID-19 pandemic has also been acknowledged by students for assessment feasibility, test anxiety alleviation, and academic performance improvement [28]. Under the context of the Healthy China Strategy proposed by the National Congress of the Communist Party of China, progress in the construction and application of blended learning in medical education has substantially contributed to the theoretical and practical development of this learning mode [29]. Although the self-reported learning experiences of students might be subjective, they serve as important feedback on the learning mode [1,30]. It is worth noting that for the COVID-19 and post–COVID-19 periods, differences in learning environments might alter students' needs for the efficiency, availability, and quality of knowledge delivery, which would provide information for the reform of the learning mode under an unexpected public health emergency [22].

However, students' perceptions of the blended learning mode for undergraduate dental education in both the COVID-19 and post–COVID-19 periods have not been revealed. This study aimed to explore students' satisfaction and suggestions regarding the blended learning mode in final-year undergraduate dental education during both the COVID-19 and post–COVID-19 periods. Using a questionnaire survey, we gathered students' satisfaction with blended learning among final-year undergraduate dental students in the West China School of Stomatology of Sichuan University, assessed students' self-reported desire to engage in blended learning activities, and identified modifiable factors that might influence their potential engagement. Our study reported the learning modes of dental education in the top-ranking dental university during the COVID-19 and post–COVID-19 period.

Methods

Study Settings and Participant Criteria

This survey study was undertaken at the West China School of Stomatology of Sichuan University. This school is recognized as a pioneer of modern dental education in China and has consistently ranked first in stomatology evaluations in mainland China [31]. The dental science program at the West China School of Stomatology is the top-ranked discipline among all dental colleges and universities, according to the Ministry of Education of mainland China. The undergraduate dental science

XSL•FO
RenderX

(5-year) program is intended to nurture students with basic medicine and clinical medicine knowledge, train students with basic theories and practical skills of stomatology, and qualify students with solid expertise on the diagnosis, treatment, and prevention of common and frequently occurring oral diseases. The course structures and learning modes in this school are considered to represent the highest level of undergraduate dental education in China.

The survey recruited final-year undergraduate dental students who majored in dental science at the West China School of Stomatology of Sichuan University during the end of the fall semester of academic years 2020-2021 and 2023-2024, as representative time points for COVID-19 (December 2019-May 2023) and post–COVID-19 (May 2023), respectively. In total, 248 and 226 final-year undergraduate dental students enrolled in the fall semesters of the academic year 2020 to 2021 and the academic year 2023 to 2024 were invited to participate in the survey. The inclusion criteria were students who (1) were enrolled in the undergraduate stomatology (5 years) program at the West China School of Stomatology of Sichuan University; (2) had completed all previous comprehensive theoretical and practical training courses in the past 4 years; and (3) had smartphone, tablet, and laptop devices with internet access during the data collection process. The exclusion criteria were students who (1) were transfers and (2) were on long-term leave or had dropped out during the study periods (COVID-19 period: December 2019 to May 2023; post–COVID-19 period: May 2023 onward).

Students' Exposure to Blended Learning

Participants accessed their blended learning experiences during their final-year study in the fall semester of the academic year 2020 to 2021 and the academic year 2023 to 2024. Courses in the undergraduate stomatology (5 years) program, which is one of China's double first-class disciplines and is jointly run with the West China Hospital of Stomatology of Sichuan University, were evaluated in this study. This program aims to cultivate talents who have basic medical theoretical and clinical knowledge; master oral medicine theories and clinical operation skills; are engaged in the national medical practitioner qualification certification; and have good professionalism in oral clinical diagnosis, treatment, and prevention. The curricula that were evaluated in this survey consisted of 49 courses divided into 4 types, namely, general courses (10/49, 20%), specialized foundation courses (22/49, 45%), specialized courses (14/49, 29%), and practical courses (3/49, 6%).

Survey Design and Pilot Test

This study used a self-administered, anonymous, electronic survey, including a survey invitation letter, participant informed consent, a questionnaire of 20 questions, and a free-response item. The first section of the questionnaire collected information on students' majors and year of study, their self-perceived knowledge levels for blended learning, and the frequency with which they reported using blended learning in their studies. The second section evaluated the types of curricula using blended learning and the detailed implementation of the learning processes. In the third section, the items measured students' satisfaction with blended learning and factors influencing their satisfaction [32,33]. The following section assessed students' perceptions of the strengths of blended learning compared with online or offline learning and evaluated their desire to engage in blended learning. The final section asked about factors that might increase or decrease students' potential engagement [9]. The free-response item was included to gather rich insights into students' thinking about blended learning. Items assessing students' self-perceived knowledge levels of blended learning were rated using a 4-point Likert scale (ranging from not knowledgeable to very knowledgeable) [32]. Items assessing students' satisfaction with blended learning were rated using a 5-point Likert scale (ranging from very dissatisfied to very satisfied) [34]. Items assessing students' desire to participate in blended learning were rated using desire as a predictor of engagement. Items assessing factors influencing students' desire to engage in blended learning were evaluated using multiple-choice question options.

The survey was developed on a web-based survey platform, WJX.cn, in mainland China, based on a skip-logic design. The items were developed by a panel of 12 experienced educators, student representatives, and academic affairs experts and were piloted on 10 subjects for 2 rounds without iterative revision. The online survey was conducted following the CHERRIES (Checklist for Reporting Results of Internet E-Surveys) guidelines.

Data Collection

The survey recruitment was conducted during 2 separate 2-week recruitment periods (January 18 to January 31, 2021, and December 27, 2023, to January 10, 2024). The survey link was included in an invitation letter and was distributed to students via QQ, which is an online platform well accepted by undergraduate students for its instant messaging and online organizing functions [35]. The response settings were rigidly specified so that a respondent could submit only 1 response, and up to 3 reminders were sent to nonresponders during the recruitment period.

Statistical Analysis

All statistical analyses were performed using SPSS software (version 22.0). The response rates included only fully completed surveys from respondents. Descriptive statistics for each outcome variable were reported as mean, mean difference (MD), and standard deviation (SD) for continuous quantitative variables and frequencies (or percentages) for categorical variables. We conducted the Shapiro-Wilk normality test and found that the student satisfaction data approximately followed a normal distribution across outcome variables (overall blended learning, individual curricula, and specific learning activities). Then, the 1-tailed t test was adopted as a parametric test to examine differences between groups by comparing the overall -satisfaction group, the high satisfaction group (HSG, defined as satisfaction responses as very satisfied and satisfied), and the low satisfaction group (LSG, defined as satisfaction responses as very dissatisfied, dissatisfied, and neutral) with a 95% CI. Pearson correlation analysis was used to examine factors associated with higher satisfaction from respondents. The chi-square test, a common nonparametric test for data not assumed to be normally distributed, was used to compare the

XSL•FO
RenderX

strengths of blended learning with offline or online learning only between different groups.

Ethical Considerations

The survey was approved by the Academic Affairs Division, West China School of Stomatology, Sichuan University Institutional Review Board (WCHSIRB-D-2024-481). The electronic survey invitation letter included a description of the study's objectives and procedures, informed participants that participation was voluntary and anonymous, and clarified that the survey results would be used for research purposes only. Informed consent was obtained from participants upon survey submission. No compensation was given for participating in this study.

Results

Among the 474 final year undergraduate students recruited in 2020 and 2023, 38% (178/474) completed the survey, with response rates of 46% (114/248) in 2020 and 28% (64/226) in 2023. Of these respondents, 75% (85/114) in 2020 and 73% (47/64) in 2023 described themselves as very knowledgeable or knowledgeable about blended learning. Considering that one response might choose multiple course types, we also analyzed the knowledge levels of responses in 2020 and 2023 (Table 1).

Table 1. Characteristics and knowledge levels of the survey in 2020 and 2023.

Curricula	Overall sample[a], n	Respondents, n (%)			
		Very knowledgeable	Knowledgeable	Somewhat knowledgeable	Not knowledgeable
2020					
General courses	38	3 (8)	22 (58)	13 (34)	0
Specialized foundation courses	53	8 (15)	33 (62)	12 (23)	0
Specialized courses	94	13 (14)	63 (67)	17 (18)	1 (1)
Practical courses	26	5 (19)	19 (73)	1 (4)	1 (4)
2023					
General courses	42 (66)	9 (21)	23 (55)	10 (24)	0
Specialized foundation courses	38 (59)	5 (13)	23 (61)	10 (26)	0
Specialized courses	50 (78)	6 (12)	28 (56)	16 (32)	0
Practical courses	32 (50)	5 (16)	19 (59)	8 (25)	0

[a]The overall sample sizes of specific courses were counted as the sum of all respondents who chose this course.

As for the implementation of blended learning, students reported shifts in the online and offline distribution patterns for specific learning procedures (Multimedia Appendix 1). In general, preclass activities, student sign-in, and in-class quizzes showed less online integration in 2023 compared to 2020. By contrast, teacher-student interactions, student-student interactions, and teachers' feedback did not follow this pattern, as these procedures showed a decrease in online engagement and an increase in offline engagement in 2023 compared with 2020, which differed from most other teaching procedures. Moreover, practical teaching and final examinations shifted from online to offline and showed almost the same trend in both 2020 and 2023.

Table 2 summarizes students' satisfaction with blended learning in 2020 and 2023. Overall, students' satisfaction decreased from 2020 to 2023, as 82% (93/114) versus 59% (38/64) of the students were in the HSG in 2020 versus 2023, and statistically significant differences were found between the HSG and the LSG in both years (MD 1.634 and 95% CI 1.502-1.767 in 2020 and MD 2.116 and 95% CI 2.054-2.177 in 2023). Specifically, compared to 2020, students in 2023 were most satisfied with specialized foundation courses (40/53, 75% in the HSG) and moderately satisfied with general courses (26/38, 68% in the HSG) and specialized courses (63/94, 67% in the HSG), whereas they felt most dissatisfied with practical courses (9/26, 35% in the HSG), and statistically significant differences were found for the mean scores of all curricula (P<.001).

XSL•FO
RenderX

Table 2. Students' satisfaction with blended learning for different curriculum types and specific learning activities in 2020 (n=114) and 2023 (n=64).

Terms	Overall (2020 vs 2023)				
	Overall, mean (SD)	HSG[a], mean (SD)	LSG[b], mean (SD)	Mean difference (95% CI)	*P* value
Overall curricula	3.79 (0.926) vs 2.36 (1.155)	4.29 (0.457) vs 3.98 (0.513)	2.66 (0.684) vs 1.87 (0.776)	1.634 (1.502-1.767) vs 2.116 (2.054-2.177)	<.001 vs <.001
General courses	3.68 (0.933) vs 2.55 (1.951)	4.19 (0.402) vs 4.27 (0.450)	2.58 (0.793) vs 1.03 (1.425)	1.609 (1.268-1.950) vs 3.238 (3.567-2.909)	<.001 vs <.001
Specialized foundation courses	3.87 (0.900) vs 4.31 (1.720)	4.28 (0.452) vs 4.23 (0.429)	2.62 (0.768) vs 1.12 (1.452)	1.660 (1.341-1.979) vs 3.108 (3.370-2.846)	<.001 vs <.001
Specialized courses	3.75 (0.950) vs 2.92 (1.721)	4.29 (0.455) vs 4.25 (0.440)	2.65 (0.709) vs 1.59 (1.478)	1.641 (1.495-1.786) vs 2.656 (3.030-2.282)	<.001 vs <.001
Practical courses	3.50 (1.105) vs 1.80 (1.904)	4.18 (0.393) vs 4.16 (0.375)	2.22 (0.833) vs 0.80 (1.307)	1.954 (1.516-2.392) vs 3.358 (3.570-3.145)	<.001 vs <.001
Preclass activities	3.79 (0.814) vs 3.75 (0.667)	4.30 (0.462) vs 4.17 (0.377)	2.88 (0.400) vs 2.96 (0.213)	1.423 (1.405-1.442) vs 1.212 (1.189-1.235)	<.001 vs <.001
Teaching activities					
Class content	3.92 (0.754) vs 3.77 (0.636)	4.31 (0.463) vs 4.14 (0.347)	2.94 (0.354) vs 2.95 (0.224)	1.367 (1.342-1.393) vs 1.186 (1.186-1.187)	<.001 vs <.001
Teaching strategies	3.82 (0.815) vs 3.77 (0.611)	4.31 (0.464) vs 4.14 (0.351)	2.87 (0.409) vs 3.00 (0.000)	1.435 (1.409-1.461) vs 1.140 (1.032-1.247)	<.001 vs <.001
Examination design	3.82 (0.748) vs 3.60 (0.750)	4.31 (0.464) vs 4.17 (0.378)	2.98 (0.154) vs 2.86 (0.356)	1.330 (1.269-1.391) vs 1.310 (1.320-1.300)	<.001 vs <.001
Postclass feedback	3.83 (0.740) vs 3.69 (0.639)	4.26 (0.441) vs 4.13 (0.335)	2.95 (0.329) vs 2.96 (0.204)	1.314 (1.304-1.324) vs 1.167 (1.146-1.187)	<.001 vs <.001
Student engagement	3.71 (0.849) vs 3.77 (0.661)	4.29 (0.459) vs 4.13 (0.344)	2.85 (0.470) vs 2.90 (0.315)	1.446 (1.418-1.475) vs 1.238 (1.287-1.190)	<.001 vs <.001
Teacher-student interactions	3.75 (0.818) vs 3.64 (0.721)	4.31 (0.465) vs 4.16 (0.370)	2.91 (0.412) vs 2.89 (0.326)	1.396 (1.385-1.405) vs 1.273 (1.283-1.263)	<.001 vs <.001
Integration between online and offline learning	3.78 (0.750) vs 3.70 (0.728)	4.24 (0.432) vs 4.17 (0.381)	2.93 (0.350) vs 2.87 (0.344)	1.318 (1.306-1.330) vs 1.301 (1.329-1.273)	<.001 vs <.001

[a]HSG: high satisfaction group.

[b]LSG: low satisfaction group.

Regarding students' attitudes toward individual learning activities in blended learning (Table 2), statistically significant differences were found in the overall mean satisfaction scores for all learning activity items (P<.001). In addition, statistically significant differences (P<.001) were found between the HSG and the LSG in the mean scores for all learning activity items in both 2020 and 2023. Students expressed the highest satisfaction for class content, with an overall mean satisfaction score of 3.92 (SD 0.754) in 2020 versus 3.77 (SD 0.636) in 2023. In addition, students were more satisfied with student engagement during blended learning in 2023 (mean 3.71, SD 0.849) compared to 2020 (mean 3.77, SD 0.661). By contrast, students expressed lower satisfaction with learning activities such as preclass activities, teaching strategies, examination design, postclass feedback, teacher-student interactions, and integrations between online and offline learning in 2023 compared to 2020.

Factors correlated with students' satisfaction with blended learning were analyzed using Pearson correlation analysis (Multimedia Appendix 2). Among different online learning materials, MOOC (r=0.19; P=.04) was significantly correlated with students' satisfaction in 2020, while live lectures (r=0.28; P=.03) and other materials (r=0.29; P=.02) were significantly correlated with students' satisfaction in 2023. Regarding offline learning materials, designated learning materials (r=0.20; P=.03) were significantly associated with higher satisfaction with blended learning in 2020. Other materials, such as clinical cases, literature materials, students' self-chosen materials, and clinical teaching models, were not related to student satisfaction in 2020. The 2023 results were similar to those of 2020, with most offline learning materials showing no correlation with students' satisfaction, except for other materials (r=0.29; P=.02). In addition, among all the teaching procedures, teachers' feedback was significantly correlated with higher student satisfaction (r=0.32; P=.01) in 2023. Other face-to-face instructions, including collaborative learning, student-student interactions, teacher-student interactions, in-class quizzes, and so on, had no significant correlation with students' satisfaction in 2020 and 2023. Moreover, there were also significant associations between students' satisfaction and active learning strategies, such as turn and talk (r=0.39; P=.001), pause procedures during lecture (r=0.35; P=.004), and other strategies (r=0.29; P=.02) in 2023,

XSL•FO
RenderX

while only simulation exercises were significantly correlated with higher student satisfaction in 2020 (r=0.19; P=.04).

As shown in Table 3, students perceived the strengths of blended learning compared to online learning or offline learning only. Compared to traditional offline learning in 2020 versus 2023, students valued blended learning for its ability to enhance learning motivation (P<.001), communication skills (P=.008), practical skills (P<.001), course interest (P=.04), time and space flexibility (P<.001), and new media literacy skills (P=.002). When compared to online learning in 2020 versus 2023, students appreciated blended learning for improving learning motivation (P=.04) and new media literacy skills (P=.02).

Table 3. Students' self-perceived strengths of blended learning compared to offline learning or online learning only in 2020 (N=114) versus 2023 (N=64).

Terms	Blended vs offline (2020 vs 2023)		Blended vs online (2020 vs 2023)		*P* value (2020 vs 2023)
	n (%)	95% CI	n (%)	95% CI	
Knowledge acquisition	63 (55) vs 30 (47)	46.1-64.1 vs 34.3-59.4	61 (54) vs 37 (58)	44.4-62.4 vs 45.4-70.3	.40 vs .22
Learning motivation	84 (74) vs 45 (70)	64.9-80.9 vs 58.8-81.8	54 (47) vs 29 (45)	38.4-56.5 vs 32.8-57.9	<.001 vs .04
Communication skills	44 (39) vs 34 (53)	30.2-47.8 vs 40.6-65.7	62 (54) vs 27 (42)	45.3-63.2 vs 29.8-54.6	.008 vs .22
Collaboration skills	53 (47) vs 35 (55)	37.6-55.6 vs 42.2-67.2	43 (38) vs 29 (45)	29.4-46.9 vs 32.8-57.9	.09 vs .29
Practical skills	17 (15) vs 27 (42)	9.5-22.6 vs 29.8-54.6	54 (47) vs 34 (53)	9.5-22.6 vs 40.6-65.7	<.001 vs .22
Interest in courses	25 (22) vs 19 (30)	15.3-30.4 vs 18.2-41.2	37 (33) vs 20 (31)	25.6-41.5 vs 19.6-42.9	.04 vs .85
Time and space flexibility	82 (72) vs 35 (55)	63.1-79.4 vs 42.2-67.2	48 (42) vs 24 (38)	33.5-51.3 vs 25.3-49.7	<.001 vs .05
New media literacy skills	54 (47) vs 32 (50)	38.4-56.5 vs 37.4-62.6	33 (29) vs 19 (30)	21.4-37.9 vs 18.2-41.2	.002 vs .02
Expanded horizons and mindset	30 (26) vs 20 (31)	19.1-35.1 vs 19.6-42.9	23 (20) vs 14 (22)	13.8-28.5 vs 11.5-32.3	.14 vs .23
None of the above	11 (10) vs 8 (13)	5.5-16.5 vs 4.2-20.8	8 (7) vs 5 (8)	3.6-13.2 vs 1.1-14.6	.24 vs .38

Most students expressed a preference for incorporating blended learning in final-year undergraduate dental education, with 70% (80/114) of the students in 2020 and 61% (39/64) in 2023 indicating a desire to engage in blended learning. Multimedia Appendix 3 summarizes that the leading factors decreasing students' potential engagement in blended learning were the same in both 2020 and 2023, namely, unstable technical support (68/114, 60% in 2020 vs 26/64, 41% in 2023), poor integration between online and offline learning (58/114, 51% in 2020 vs 34/64, 53% in 2023), a lack of learning motivation (51/114, 45% in 2020 vs 24/64, 38% in 2023), and a lack of teacher-student interaction (44/114, 39% in 2020 vs 20/64, 31% in 2023). Compared to 2020, there was a decline in unstable technical support, a lack of learning motivation, a lack of teacher-student interaction, and a lack of teaching materials, which decreased students' desire to engage in blended learning. By contrast, the influences of improper performance assessment, insufficient teaching ability, and limited teaching supervision and evaluation on decreasing students' engagement were upregulated in 2023.

On the other hand, the top factors increasing students' potential engagement in blended learning were the same in both 2020 and 2023, namely, providing high-quality learning materials (76/114, 67% in 2020 vs 43/64, 67% in 2023) and the improvement of the technical environment (62/114, 54% in 2020 vs 35/64, 55% in 2023). Compared to 2020, the influence of the use of teacher incentives (59/114, 52% in 2020 vs 22/64, 34% in 2023) on increasing students' engagement decreased in 2023. By contrast, the impacts of emphasizing evaluation in the learning process (51/114, 45% in 2020 vs 33/64, 52% in 2023) and establishing effective classroom schedules (48/114, 42% in 2020 vs 31/64, 48% in 2023) on enhancing students' engagement increased in 2023.

For the free-response items, students suggested ways to enhance engagement in blended learning, including involving students in curriculum design and learning material development, more emphasis on problem solving during face-to-face instruction, as well as offering incentives during the learning process, etc.

Discussion

Principal Findings

This study demonstrates that final-year undergraduate dental students had a generally satisfied attitude toward blended learning in the COVID-19 and post–COVID-19 years. Students acknowledged the strengths of blended learning in multiple aspects compared to online or offline learning only. Meanwhile, students desired to engage in blended learning and reported factors that might influence their potential engagement. These findings reveal students' self-reported satisfaction and

XSL•FO
RenderX

perceptions regarding blended learning for undergraduate dental education and provide practical clues for enhancing student engagement in blended learning in the circumstance of the COVID-19 pandemic and beyond.

At its peak, the COVID-19 pandemic significantly impacted higher education by necessitating rapid shifts to online or blended learning in health professional training [7,9,36]. Participants generally acknowledged that the diverse methodologies in online-based learning during the COVID-19 pandemic advanced the modernization of health professional education [37]. In addition, comparable or even greater effectiveness for knowledge acquisition than nonblended learning has been observed in health professionals, although heterogeneity should be noted [38]. Similar positive effects of blended learning have been reported in health professions students for learning effectiveness [39], graduate medical students for knowledge gain [40], and undergraduate medicine and nursing students for knowledge outcomes and clinical skills training [7,41]. Serving as one of the most prevalent teaching approaches for medical education during the pandemic, blended learning encountered challenges such as students' limited acceptance of technology-assisted learning, difficulties in student engagement during the learning process, and issues related to cognition and retention [42]. Moreover, the sudden transition of medical education to online-only or blended learning modes during the pandemic might interrupt academic interactions and trigger unstructured learning patterns among health professional students and arouse their need for retaining certain course components, such as practical training in traditional face-to-face learning [43].

In our survey study, more students showed a generally satisfied attitude toward blended learning in 2020 compared to 2023 (Table 2). Other studies have shown that students' age, computer experience, and technology acceptance might affect their satisfaction with online learning [44,45], while failure to incorporate these aspects in our survey might bias the results. In fact, final-year dental students complete most of their theoretical and practical studies during their junior year or earlier, making them familiar with online learning environments and technology for blended learning, which helps minimize bias in our study. Notably, students expressed the lowest degree of satisfaction toward practical courses in both 2020 and 2023 (Table 2), consistent with the finding that the number of students who desired to engage in practical courses involving blended learning was the least among all curriculum types in both years (data not shown). However, in another study assessing student perceptions of blended learning in practical training in orthodontic education, students expressed general acceptance and a preference for engagement in the blended learning mode compared with both traditional solid models and e-models [18]. Specifically, the blended learning mode that combines electrical 3D simulation models with traditional plaster models cultivates critical thinking and haptic sensitivity to facilitate clinical practice [18]. Another retrospective study concluded that the integration of a flipped classroom approach into blended learning in a dental anatomy course could effectively improve students' scores in practical assessments, although not in theoretical assessments [46]. These studies support the adoption of blended learning in practical training for undergraduate dental education. Moreover, the learning of experiential forms of knowledge is also important for dental skills [47,48]. By contrast, in the COVID-19 pandemic context, the emphasis on patients' knowledge of dental emergencies, disease transmission, and needed preparations underscored the need for dental students to enhance their professional preparedness and knowledge for clinical practice [38]. This might fuel senior dental students' aspiration to acquire clinical skills, particularly during the post–COVID-19 period. Future instruction on the rationality of course design, the availability of reliable online resources, and the delivery of in-class activities is necessary for the implementation of blended learning in dental courses with practical components, improving learner acceptance and performance outcomes.

In our study, students showed the lowest satisfaction for teacher-student interactions in both 2020 and 2023 (Table 2), and teacher-student interactions were identified as a key factor influencing student engagement in blended learning (Multimedia Appendix 3). According to the social constructivist theory [49], interactions between a student and their teachers and peers during blended learning could stimulate effective learning and thus improve student learning [50]. Blended learning can also integrate the advantages of synchronous teacher-student interactions with asynchronous time and space flexibility to foster the engagement of students and teachers in the learning process [51,52]. Indeed, teacher-student interactions contribute to improved student engagement and satisfaction with blended learning compared to online learning and face-to-face learning [50]. During the COVID-19 pandemic, students reported insufficient interactions with teachers for preclinical practice in dental education, and such reduced teacher-student interactions partially impacted students' overall satisfaction with online learning-involved processes [9,50,53]. Barriers such as challenges in selecting helpful peers, limited and delayed teacher feedback and support, as well as difficulty in discerning relevant information, particularly in asynchronous online environments, might lead to students' dissatisfaction with teacher-student interaction experiences in blended learning.

Students greatly appreciated the strengths of blended learning over online or offline learning only (Table 3). In the past decade, online learning resources and platforms have substantially facilitated the successful implementation of blended learning in higher education [54]. During the COVID-19 pandemic, more high-quality online courses were freely developed, opened, and shared among different schools to enhance public accessibility for educators and students [55]. In recent years, it is encouraging that more high-quality online learning platforms with communication components have been available for web-based courses in dental education, such as Rain Classroom [56] and PMPHMOOC [57]. These platforms could be accessed by students and teachers on campus or remotely via mobile devices, such as laptops, tablets, and phones [9]. Moreover, social software tools, such as Tencent QQ chat groups [9] and WeChat apps [58], also demonstrate an educational function by promoting multidirectional interactions, including teacher-student interactions, even during the COVID-19 pandemic [11]. Although these platforms are mostly limited to

XSL•FO
RenderX

China and might not be applicable to other countries, the fast-growing evolution of these online learning platforms and social tools is progressing toward meeting the needs of dental students and educators in the COVID-19 and post–COVID-19 periods.

Students expressed a desire to engage in blended learning in both 2020 and 2023 and reported several factors that might influence their potential engagement. More than half of the students indicated the importance of tech for blended learning in both years (Multimedia Appendix 3). As mentioned earlier, much effort is needed to facilitate the implementation of higher education due to the COVID-19 pandemic [59]. However, the availability of newly developed educational resources is still limited. Moreover, their applications and adaptations to a wider range require more time and practice [35,60]. Previous studies have shown a strong relationship between the learning environment and student learning behaviors, self-perceived satisfaction, and academic performance [61-64]. As a well-established element highly susceptible to the environment, self-regulated learning imposes higher standards for students to adapt to the inevitably varied COVID-19 pandemic and postpandemic world than before [19,36]. Efforts have been made to cultivate self-regulated skills in students for a web-based environment, such as the provision of tutorials before starting a course or the supply of real-time skill guidance within a course. Our findings also suggest that with the incorporation of blended learning, a significant number of students viewed the impact of the development of the learning environment positively when faced with the unprecedented COVID-19 and post–COVID-19 periods.

Limitations

There are several limitations of this study. First, the survey was performed in a single dental school and might not reflect the blended learning experiences of other schools. Second, our survey did not record students' health status during the COVID-19 and post-COVID-19 periods.

Conclusions

This survey study demonstrates that final-year undergraduate dental students are generally satisfied with blended learning in both the COVID-19 and post–COVID-19 years. Although students prefer to engage in blended learning compared with online or offline learning only, their desire to participate might vary depending on multiple factors. Dental schools, policy makers, and educators could increase students' engagement in blended learning by incorporating more teacher-student interactions, establishing standardized assessment mechanisms, providing dedicated learning materials and platforms, and promoting self-regulated learning among dental students. The COVID-19 pandemic has raised an alarm for dental education to reform into a more flexible and adaptable mode, and this survey provides clues for developing a new face of blended learning–involved dental education in the context of the pandemic and beyond.

Acknowledgments

This study was supported by the Research and Development Program of the West China Hospital of Stomatology, Sichuan University (; October 2020). The authors thank Le Hong for assistance with the design of the survey items and data collection.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Authors' Contributions

XL, LY, and JY conceived the study and collaboratively developed the research protocol. XL, LL, and JY supported the comprehensive search and extraction of studies. XL, LP, HT, and LL conducted the data analysis. XL, LP, and LL drafted the manuscript. XL, LP, XH, and JY contributed to the interpretation of the data, revised the manuscript, and approved the final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Characteristics of online vs offline distribution patterns of specific teaching procedures in blended learning in 2020 vs 2023.
[DOCX File , 17 KB - formative_v9i1e63453_app1.docx]

Multimedia Appendix 2
Characteristics of online vs offline distribution patterns of specific teaching procedures in blended learning in 2020 vs 2023.
[DOCX File , 20 KB - formative_v9i1e63453_app2.docx]

Multimedia Appendix 3
Factors influencing final-year undergraduate dental students' potential engagement in blended learning during and after COVID-19 in 2020 and 2023.

XSL•FO
RenderX

[PNG File , 195 KB - formative_v9i1e63453_app3.png]

References

1. Chen Y, Deng J, Li B, Yang Y, He Z, Ye L, et al. Curriculum setting and students' feedback of pre-clinical training in different dental schools in China-a national-wide survey. Eur J Dent Educ 2022 Feb 09;26(1):28-35. [doi: 10.1111/eje.12669] [Medline: 33511722]
2. Wang YH, Zhao Q, Tan Z. Current differences in dental education between Chinese and Western models. Eur J Dent Educ 2017 Nov 24;21(4):e43-e49. [doi: 10.1111/eje.12216] [Medline: 27339198]
3. Manogue M, McLoughlin J, Christersson C, Delap E, Lindh C, Schoonheim-Klein M, et al. Curriculum structure, content, learning and assessment in European undergraduate dental education - update 2010. Eur J Dent Educ 2011 Aug;15(3):133-141 [FREE Full text] [doi: 10.1111/j.1600-0579.2011.00699.x] [Medline: 21762317]
4. Gatt G, Attard NJ. Multimodal teaching methods for students in dentistry: a replacement for traditional teaching or a valuable addition? A three-year prospective cohort study. BMC Med Educ 2023 Jun 02;23(1):401 [FREE Full text] [doi: 10.1186/s12909-023-04377-z] [Medline: 37268949]
5. Du GF, Li CZ, Shang SH, Xu XY, Chen HZ, Zhou G. Practising case-based learning in oral medicine for dental students in China. Eur J Dent Educ 2013 Nov 01;17(4):225-228. [doi: 10.1111/eje.12042] [Medline: 24127763]
6. Huang B, Zheng L, Li C, Li L, Yu H. Effectiveness of problem-based learning in Chinese dental education: a meta-analysis. J Dent Educ 2013 Mar;77(3):377-383. [Medline: 23486905]
7. Li L, Liu X, Chen Z, Wang L, Lian X, Zou H. The application of a case-based social media-assisted teaching method in cariology education: comparative study. J Med Internet Res 2021 Aug 13;23(8):e29372 [FREE Full text] [doi: 10.2196/29372] [Medline: 34397390]
8. Santos GN, Leite AF, Figueiredo PT, Pimentel NM, Flores-Mir C, de Melo NS, et al. Effectiveness of e-learning in oral radiology education: a systematic review. J Dent Educ 2016 Sep;80(9):1126-1139. [Medline: 27587580]
9. Wang K, Zhang L, Ye L. A nationwide survey of online teaching strategies in dental education in China. J Dent Educ 2021 Feb 20;85(2):128-134 [FREE Full text] [doi: 10.1002/jdd.12413] [Medline: 32954532]
10. Mangione S, Nieman LZ, Greenspon LW, Margulies H. A comparison of computer-assisted instruction and small-group teaching of cardiac auscultation to medical students. Med Educ 1991 Sep;25(5):389-395. [doi: 10.1111/j.1365-2923.1991.tb00086.x] [Medline: 1758315]
11. Rouse DP. The effectiveness of computer-assisted instruction in teaching nursing students about congenital heart disease. Comput Nurs 2000;18(6):282-287. [Medline: 11105402]
12. Tomej K, Liburd J, Blichfeldt BS, Hjalager AM. Blended and (not so) splendid teaching and learning: higher education insights from university teachers during the Covid-19 pandemic. Int J Educ Res Open 2022;3:100144 [FREE Full text] [doi: 10.1016/j.ijedro.2022.100144] [Medline: 36161268]
13. Vallée A, Blacher J, Cariou A, Sorbets E. Blended learning compared to traditional learning in medical education: systematic review and meta-analysis. J Med Internet Res 2020 Aug 10;22(8):e16504 [FREE Full text] [doi: 10.2196/16504] [Medline: 32773378]
14. Bonnel WB, Starling CK, Wambach KA, Tarnow K. Blended roles: preparing the advanced practice nurse educator/clinician with a web-based nurse educator certificate program. J Prof Nurs 2003 Nov;19(6):347-353 [FREE Full text] [doi: 10.1016/s8755-7223(03)00130-3] [Medline: 14689391]
15. Chen M, Ye L, Weng Y. Blended teaching of medical ethics during COVID-19: practice and reflection. BMC Med Educ 2022 May 11;22(1):361 [FREE Full text] [doi: 10.1186/s12909-022-03431-6] [Medline: 35545784]
16. Cheng X, Mo W, Duan Y. Factors contributing to learning satisfaction with blended learning teaching mode among higher education students in China. Front Psychol 2023;14:1193675 [FREE Full text] [doi: 10.3389/fpsyg.2023.1193675] [Medline: 37457098]
17. Reissmann DR, Sierwald I, Berger F, Heydecke G. A model of blended learning in a preclinical course in prosthetic dentistry. J Dent Educ 2015 Feb;79(2):157-165. [Medline: 25640620]
18. Ho AC, Liao C, Lu J, Shan Z, Gu M, Bridges SM, et al. 3-Dimensional simulations and student learning in orthodontic education. Eur J Dent Educ 2022 Aug 03;26(3):435-445 [FREE Full text] [doi: 10.1111/eje.12718] [Medline: 34528742]
19. Kavadella A, Tsiklakis K, Vougiouklakis G, Lionarakis A. Evaluation of a blended learning course for teaching oral radiology to undergraduate dental students. Eur J Dent Educ 2012 Feb 17;16(1):e88-e95. [doi: 10.1111/j.1600-0579.2011.00680.x] [Medline: 22251359]
20. WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020. World Health Organization. 2020 Mar 11. URL: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020 [accessed 2025-10-12]
21. Coronavirus disease (COVID-19) pandemic. World Health Organization. URL: https://www.who.int/europe/emergencies/situations/covid-19 [accessed 2025-10-12]
22. Chilton JK, Hanks S, Watson HR. A blended future? A cross-sectional study demonstrating the impacts of the COVID-19 pandemic on student experiences of well-being, teaching and learning. Eur J Dent Educ 2024 Feb 28;28(1):170-183 [FREE Full text] [doi: 10.1111/eje.12934] [Medline: 37379447]

XSL•FO
RenderX

23. Choi B, Jegatheeswaran L, Minocha A, Alhilani M, Nakhoul M, Mutengesa E. The impact of the COVID-19 pandemic on final year medical students in the United Kingdom: a national survey. BMC Med Educ 2020 Jun 29;20(1):206 [FREE Full text] [doi: 10.1186/s12909-020-02117-1] [Medline: 32600460]
24. Long N, Wolpaw DR, Boothe D, Caldwell C, Dillon P, Gottshall L, et al. Contributions of health professions students to health system needs during the COVID-19 pandemic: potential strategies and process for U.S. medical schools. Acad Med 2020 Nov;95(11):1679-1686 [FREE Full text] [doi: 10.1097/ACM.0000000000003611] [Medline: 32701558]
25. Pellerone M. Self-perceived instructional competence, self-efficacy and burnout during the Covid-19 pandemic: a study of a group of Italian school teachers. Eur J Investig Health Psychol Educ 2021 Jun 01;11(2):496-512 [FREE Full text] [doi: 10.3390/ejihpe11020035] [Medline: 34708818]
26. Garzon Diaz K, Villamil Duarte A, Velasco Forero S, Herrera Jacobo S. Post-pandemic face-to-face learning environments for undergraduate students: a scoping review protocol. PLoS One 2024;19(11):e0309932 [FREE Full text] [doi: 10.1371/journal.pone.0309932] [Medline: 39535990]
27. Nijakowski K, Lehmann A, Zdrojewski J, Nowak M, Surdacka A. The effectiveness of the blended learning in conservative dentistry with endodontics on the basis of the survey among 4th-year students during the COVID-19 pandemic. Int J Environ Res Public Health 2021 Apr 25;18(9):4555 [FREE Full text] [doi: 10.3390/ijerph18094555] [Medline: 33923047]
28. Hong S, Go B, Rho J, An S, Lim C, Seo DG, et al. Effects of a blended design of closed-book and open-book examinations on dental students' anxiety and performance. BMC Med Educ 2023 Jan 13;23(1):25 [FREE Full text] [doi: 10.1186/s12909-023-04014-9] [Medline: 36635682]
29. Ma J, Jiang X, Wang J, Liang Z, Sun Z, Qian H, et al. The construction and application of a blended teaching model under the strategic background of healthy China. Biochem Mol Biol Educ 2022 Jan;50(1):114-119 [FREE Full text] [doi: 10.1002/bmb.21591] [Medline: 34854213]
30. Davey J, Bryant ST, Dummer PM. The confidence of undergraduate dental students when performing root canal treatment and their perception of the quality of endodontic education. Eur J Dent Educ 2015 Nov;19(4):229-234. [doi: 10.1111/eje.12130] [Medline: 25490882]
31. Global ranking of academic subjects. ShanghaiRanking. URL: https://www.shanghairanking.com/rankings/gras/2021 [accessed 2021-06-29]
32. Bucklin BA, Asdigian NL, Hawkins JL, Klein U. Making it stick: use of active learning strategies in continuing medical education. BMC Med Educ 2021 Jan 11;21(1):44 [FREE Full text] [doi: 10.1186/s12909-020-02447-0] [Medline: 33430843]
33. AlRuthia Y, Alhawas S, Alodaibi F, Almutairi L, Algasem R, Alrabiah HK, et al. The use of active learning strategies in healthcare colleges in the Middle East. BMC Med Educ 2019 May 14;19(1):143 [FREE Full text] [doi: 10.1186/s12909-019-1580-4] [Medline: 31088430]
34. Ren L, Ren J, Liu C, He M, Qiu X. Policy goals of contract arrangements in primary care in jeopardy: a cross-sectional consumer satisfaction survey of community residents in Hangzhou, China. Front Public Health 2022;10:800612 [FREE Full text] [doi: 10.3389/fpubh.2022.800612] [Medline: 35586005]
35. Zhao Y, Watterston J. The changes we need: education post COVID-19. J Educ Chang 2021 Feb 18;22(1):3-12 [FREE Full text] [doi: 10.1007/s10833-021-09417-3] [Medline: 38624465]
36. Lee M, An SY, Ihm J. Dental students' satisfaction with web-based learning during the initial phase of the COVID-19 pandemic: mixed methods study. J Med Internet Res 2024 Mar 08;26:e50278 [FREE Full text] [doi: 10.2196/50278] [Medline: 38457808]
37. Iyer PH. Corona virus disease (COVID-19): lessons learned impact on the education of health professionals. Adv Exp Med Biol 2024;1458:233-246. [doi: 10.1007/978-3-031-61943-4_15] [Medline: 39102200]
38. Liu Q, Peng W, Zhang F, Hu R, Li Y, Yan W. The effectiveness of blended learning in health professions: systematic review and meta-analysis. J Med Internet Res 2016 Jan 04;18(1):e2 [FREE Full text] [doi: 10.2196/jmir.4807] [Medline: 26729058]
39. Cook DA, Levinson AJ, Garside S, Dupras DM, Erwin PJ, Montori VM. Internet-based learning in the health professions: a meta-analysis. JAMA 2008 Sep 10;300(10):1181-1196. [doi: 10.1001/jama.300.10.1181] [Medline: 18780847]
40. Jwayyed S, Stiffler KA, Wilber ST, Southern A, Weigand J, Bare R, et al. Technology-assisted education in graduate medical education: a review of the literature. Int J Emerg Med 2011 Aug 08;4(1):51 [FREE Full text] [doi: 10.1186/1865-1380-4-51] [Medline: 21824405]
41. Enoch LC, Abraham RM, Singaram VS. Factors that enhance and hinder the retention and transfer of online pre-clinical skills training to facilitate blended learning. Adv Med Educ Pract 2023;14:919-936 [FREE Full text] [doi: 10.2147/AMEP.S398376] [Medline: 37645657]
42. Kaup S, Jain R, Shivalli S, Pandey S, Kaup S. Sustaining academics during COVID-19 pandemic: the role of online teaching-learning. Indian J Ophthalmol 2020 Jun;68(6):1220-1221 [FREE Full text] [doi: 10.4103/ijo.IJO_1241_20] [Medline: 32461490]
43. O'Brien L, Tighe J, Doroud N, Barradell S, Dowling L, Pranata A, et al. "Burnout felt inevitable": experiences of university staff in educating the nursing and allied health workforce during the first COVID-19 waves. Front Public Health 2023 Jan 30;11:1082325 [FREE Full text] [doi: 10.3389/fpubh.2023.1082325] [Medline: 36794063]

XSL•FO
RenderX

44. McCutcheon K, Lohan M, Traynor M, Martin D. A systematic review evaluating the impact of online or blended learning vs. face-to-face learning of clinical skills in undergraduate nurse education. J Adv Nurs 2015 Feb 19;71(2):255-270. [doi: 10.1111/jan.12509] [Medline: 25134985]
45. Moule P, Ward R, Lockyer L. Nursing and healthcare students' experiences and use of e-learning in higher education. J Adv Nurs 2010 Dec;66(12):2785-2795. [doi: 10.1111/j.1365-2648.2010.05453.x] [Medline: 20946565]
46. Bakr MM, Massey WL, Massa HM. Flipping a dental anatomy course: a retrospective study over four years. Educ Res Int 2016;2016:1-9. [doi: 10.1155/2016/7097398]
47. Whipp JL, Ferguson DJ, Wells LM, Iacopino AM. Rethinking knowledge and pedagogy in dental education. J Dent Educ 2000 Dec;64(12):860-866. [Medline: 11197948]
48. Vanka A, Vanka S, Wali O. Flipped classroom in dental education: a scoping review. Eur J Dent Educ 2020 May 18;24(2):213-226. [doi: 10.1111/eje.12487] [Medline: 31808231]
49. Westerlaken M, Christiaans-Dingelhoff I, Filius RM, de Vries B, de Bruijne M, van Dam M. Blended learning for postgraduates; an interactive experience. BMC Med Educ 2019 Jul 30;19(1):289 [FREE Full text] [doi: 10.1186/s12909-019-1717-5] [Medline: 31362735]
50. Ullah R, Siddiqui F, Adnan S, Afzal AS, Sohail Zafar M. Assessment of blended learning for teaching dental anatomy to dentistry students. J Dent Educ 2021 Jul;85(7):1301-1308 [FREE Full text] [doi: 10.1002/jdd.12606] [Medline: 33794027]
51. Blond N, Chaux AG, Hascoët E, Lesclous P, Cloitre A. Blended learning compared to traditional learning for the acquisition of competencies in oral surgery by dental students: a randomized controlled trial. Eur J Dent Educ 2024 Nov 31;28(4):943-954. [doi: 10.1111/eje.13030] [Medline: 39083448]
52. Sellnow-Richmond D, Strawser MG, Sellnow DD. Student perceptions of teaching effectiveness and learning achievement: a comparative examination of online and hybrid course delivery format. Commun Teach 2019 Oct 20;34(3):248-263. [doi: 10.1080/17404622.2019.1673456]
53. Di Giacomo P, Di Paolo C. COVID-19 and dental distance-based education: students' perceptions in an Italian university. BMC Med Educ 2021 Aug 02;21(1):414 [FREE Full text] [doi: 10.1186/s12909-021-02840-3] [Medline: 34340662]
54. Schlenz MA, Schmidt A, Wöstmann B, Krämer N, Schulz-Weidner N. Students' and lecturers' perspective on the implementation of online learning in dental education due to SARS-CoV-2 (COVID-19): a cross-sectional study. BMC Med Educ 2020 Oct 09;20(1):354 [FREE Full text] [doi: 10.1186/s12909-020-02266-3] [Medline: 33036592]
55. Quinn B, Field J, Gorter R, Akota I, Manzanares MC, Paganelli C, et al. COVID-19: the immediate response of European academic dental institutions and future implications for dental education. Eur J Dent Educ 2020 Nov 13;24(4):811-814 [FREE Full text] [doi: 10.1111/eje.12542] [Medline: 32394605]
56. Rain classroom web version. Rain Classroom. URL: https://www.yuketang.cn/web [accessed 2025-10-12]
57. People's health MOOC homepage. People's Health MOOC. URL: http://www.pmphmooc.com/#/home [accessed 2025-10-12]
58. Luan H, Wang M, Sokol RL, Wu S, Victor BG, Perron BE. A scoping review of WeChat to facilitate professional healthcare education in Mainland China. Med Educ Online 2020 Dec 23;25(1):1782594 [FREE Full text] [doi: 10.1080/10872981.2020.1782594] [Medline: 32573367]
59. Wang C, Cheng Z, Yue XG, McAleer M. Risk management of COVID-19 by universities in China. J Risk Financial Manag 2020 Feb 19;13(2):36. [doi: 10.3390/jrfm13020036]
60. Hattar S, AlHadidi A, Sawair FA, Alraheam IA, El-Ma'aita A, Wahab FK. Impact of COVID-19 pandemic on dental education: online experience and practice expectations among dental students at the University of Jordan. BMC Med Educ 2021 Mar 08;21(1):151 [FREE Full text] [doi: 10.1186/s12909-021-02584-0] [Medline: 33685451]
61. Divaris K, Barlow PJ, Chendea SA, Cheong WS, Dounis A, Dragan IF, et al. The academic environment: the students' perspective. Eur J Dent Educ 2008 Feb 15;12(s1):120-130. [doi: 10.1111/j.1600-0579.2007.00494.x]
62. DiMattio MJ, Hudacek SS. Educating generation Z: psychosocial dimensions of the clinical learning environment that predict student satisfaction. Nurse Educ Pract 2020 Nov;49:102901. [doi: 10.1016/j.nepr.2020.102901] [Medline: 33086139]
63. Han F, Ellis RA. The relations between self-reported perceptions of learning environment, observational learning strategies, and academic outcome. J Comput High Educ 2023 Aug 11;35(1):111-125 [FREE Full text] [doi: 10.1007/s12528-022-09333-2] [Medline: 35971519]
64. Wang H, Li D, Gu C, Wei W, Chen J. Research on high school students' behavior in art course within a virtual learning environment based on SVVR. Front Psychol 2023 Jul 14;14:1218959 [FREE Full text] [doi: 10.3389/fpsyg.2023.1218959] [Medline: 37519396]

Abbreviations

CHERRIES: Checklist for Reporting Results of Internet E-Surveys
HSG: high satisfaction group
LSG: low satisfaction group
MD: mean difference

XSL•FO
RenderX

Edited by J Sarvestan; submitted 20.06.24; peer-reviewed by GM Santaella, A Alsaegh; comments to author 24.09.24; revised version received 19.11.24; accepted 28.08.25; published 28.11.25.

Please cite as:
Li X, Peng L, Tu H, Liu L, Han X, Ye L, Yang J
Dental Undergraduate Students' Perceptions of Blended Learning in the COVID-19 and Post–COVID-19 Years: Survey Study
JMIR Form Res 2025;9:e63453
URL: https://formative.jmir.org/2025/1/e63453
doi:10.2196/63453
PMID:

©Xin Li, Li Peng, Heng Tu, Lijia Liu, Xianglong Han, Ling Ye, Jing Yang. Originally published in JMIR Formative Research (https://formative.jmir.org), 28.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Correction: Breath-Focused Mindfulness and Compassion Training in Parent-Child Dyads: Pilot Intervention Study

Satish Jaiswal[1], PhD; Jason Nan[1], PhD; Seth Dizon[1], BS; Jessica O Young[1], BS; Suzanna R Purpura[1], MS; James K Manchanda[1], BS; Dhakshin Ramanathan[2], MD, PhD; Dennis J Kuo[1], MS, MD; Jyoti Mishra[1], MBA, PhD

[1]Department of Psychiatry, University of California, San Diego, 9500 Gilman Drive, Guava Building, Room 130, La Jolla, CA, United States
[2]VA San Diego Medical Center, San Diego, CA, United States

Corresponding Author:
Satish Jaiswal, PhD
Department of Psychiatry, University of California, San Diego, 9500 Gilman Drive, Guava Building, Room 130, La Jolla, CA, United States

Related Article:

Correction of: https://formative.jmir.org/2025/1/e69607

(JMIR Form Res 2025;9:e81122) doi:10.2196/81122

In "Breath-Focused Mindfulness and Compassion Training in Parent-Child Dyads: Pilot Intervention Study" (JMIR Form Res 2025;9:e69607), the authors noted one error.

In Table 1 of the originally published paper, the age ranges of parents and children were swapped. Table 1 with the age range corrected for "Parents (n=24)" and "Children (n=24)" is displayed below:

Table . Summary of demographics and baseline mental health for parent-child dyad study participants.[a]

Demographics and baseline mental health	Parents (n=24)	Children (n=24)
Age (years)		
Mean (SD)	44.5 (6.5)	9.5 (3.27)
Range	28-54	5-15

[a]Parental stress was measured using the 7 stress items on the 21-item Depression Anxiety Stress Scale (DASS-21), anxiety was measured on the 7-item General Anxiety Disorder (GAD-7) scale, and depression was measured on the 9-item Patient Health Questionnaire scale (PHQ-9).

The correction will appear in the online version of the paper on the JMIR Publications website, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Submitted 22.07.25; this is a non–peer-reviewed article;accepted 23.07.25; published 07.08.25.

Please cite as:
Jaiswal S, Nan J, Dizon S, Young JO, Purpura SR, Manchanda JK, Ramanathan D, Kuo DJ, Mishra J
Correction: Breath-Focused Mindfulness and Compassion Training in Parent-Child Dyads: Pilot Intervention Study
JMIR Form Res 2025;9:e81122
URL: https://formative.jmir.org/2025/1/e81122
doi:10.2196/81122

© Satish Jaiswal, Jason Nan, Seth Dizon, Jessica O Young, Suzanna R Purpura, James K Manchanda, Dhakshin Ramanathan, Dennis J Kuo, Jyoti Mishra. Originally published in JMIR Formative Research (https://formative.jmir.org), 7.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: Latino Parents' Reactions to and Engagement With a Facebook Group–Based COVID-19 Vaccine Promotion Intervention: Mixed Methods Pilot Study

Anna I González-Salinas[1], BSc, MBA, MM; Elizabeth L Andrade[1], BA, MPH, DrPH; Lorien C Abroms[1], BSc, MA, DSC; Kaitlyn Gómez[2], BA; Carla Favetto[1]; Valeria M Gómez[1], BS; Karen K Collins[1], PhD

[1]George Washington University, Washington, DC, United States

[2]California State University, Fullerton, CA, United States

Corresponding Author:
Anna I González-Salinas, BSc, MBA, MM
George Washington University
950 New Hampshire Avenue
Washington, DC, 20052
United States
Phone: 1 (202) 994 7400
Email: agonzalez985@gwmail.gwu.edu

Related Article:

Correction of: https://formative.jmir.org/2024/1/e51331

(JMIR Form Res 2025;9:e76107) doi:10.2196/76107

In "Latino Parents' Reactions to and Engagement With a Facebook Group–Based COVID-19 Vaccine Promotion Intervention: Mixed Methods Pilot Study" (JMIR Form Res 2024;8:e51331) the authors made one update.

The Acknowledgements section has been changed from:

> *This research was funded by the National Institutes of Health Community Engagement Alliance Against COVID-19 Disparities (NIH CEAL; grant 1OT2HL156812-01; multiple Principal Investigators: Mark Edberg and ELA), and the George Washington University Institute for Data, Democracy, and Politics (internal grant; multiple Principal Investigators: LCA and ELA).*

To:

> *This research was, in part, funded by the National Institutes of Health (NIH) Agreement OT2HL158287. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the NIH. This research was also funded, in part, by The George Washington University Institute for Data, Democracy, and Politics (multiple Principal Investigators LCA and ELA).*

The correction will appear in the online version of the paper on the JMIR Publications website, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Submitted 19.04.25; this is a non–peer-reviewed article;accepted 22.04.25; published 28.04.25.

Please cite as:
González-Salinas AI, Andrade EL, Abroms LC, Gómez K, Favetto C, Gómez VM, Collins KK
Correction: Latino Parents' Reactions to and Engagement With a Facebook Group–Based COVID-19 Vaccine Promotion Intervention: Mixed Methods Pilot Study
JMIR Form Res 2025;9:e76107
URL: https://formative.jmir.org/2025/1/e76107
doi:10.2196/76107
PMID:

XSL•FO
RenderX

©Anna I González-Salinas, Elizabeth L Andrade, Lorien C Abroms, Kaitlyn Gómez, Carla Favetto, Valeria M Gómez, Karen K Collins. Originally published in JMIR Formative Research (https://formative.jmir.org), 28.04.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: Advancing Mental Health and Psychological Support for Health Care Workers Using Digital Technologies and Platforms

Related Article:

Correction of: https://formative.jmir.org/2021/6/e22075/

(JMIR Form Res 2025;9:e84386) doi:10.2196/84386

Following the publication of "Advancing Mental Health and Psychological Support for Health Care Workers Using Digital Technologies and Platforms" [1], concerns were raised regarding corresponding author Jiancheng Ye's rate of self-citation within the article's reference list. Further investigation into this matter identified that numerous self-citations were added to the manuscript following acceptance of the article without adequate disclosure.

The author was contacted for a response regarding this matter. The author indicated that the added references were pertinent to the article and its findings, and provided documentation justifying the inclusion of the identified self-citations.

Following review of this response, the JMIR Publications Editorial Office determined that several of the added references were not relevant to the article or were used to support a generic statement where another reference could have been used instead of self-citation. Therefore, the JMIR Publications Editorial Office is issuing this corrigendum to remove the following references:

- Reference 1: Ye J. Health information system's responses to COVID-19 pandemic in China: a national cross-sectional study. Appl Clin Inform 2021 Mar 19;12(2):399-406 [doi: 10.1055/s-0041-1728770] [Medline: 34010976]
- Reference 9: Ye J. Pediatric mental and behavioral health in the period of quarantine and social distancing with COVID-19. JMIR Pediatr Parent 2020 Jul 28;3(2):e19867 [doi: 10.2196/19867] [Medline: 32634105
- Reference 15: Ye J, Wang Z, Hai J. Social networking service, patient-generated health data, and population health informatics: patterns and implications for using digital technologies to support mental health. medRxiv. Preprint posted online June 18, 2021 URL: https://www.medrxiv.org/content/10.1101/2021.06.11.21258777v1
- Reference 19: Feinglass J, Wang JA, Ye J, Tessier R, Kim H. Hospital care for opioid use in Illinois, 2016-2019. J Behav Health Serv Res 2021 Jan 27:1-13 [doi: 10.1007/s11414-020-09748-8] [Medline: 33502670]
- Reference 32: Ye J, Ma Q. The effects and patterns among mobile health, social determinants, and physical activity: a nationally representative cross-sectional study. AMIA Jt Summits Transl Sci Proc 2021
- Reference 37: Ye J, Sanchez-Pinto LN. Three data-driven phenotypes of multiple organ dysfunction syndrome preserved from early childhood to middle adulthood. AMIA Annu Symp Proc 2020;2020:1345-1353 [Medline: 33936511]

The JMIR Publications Editorial Office regrets that these issues were not identified prior to publication.

The correction will appear in the online version of the paper on the JMIR Publications website together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Reference

1. Ye J. Advancing mental health and psychological support for health care workers using digital technologies and platforms. JMIR Form Res 2021 Jun 30;5(6):e22075 [FREE Full text] [doi: 10.2196/22075] [Medline: 34106874]

Submitted 18.09.25; this is a non–peer-reviewed article;accepted 18.09.25; published 08.10.25.

Please cite as:
Correction: Advancing Mental Health and Psychological Support for Health Care Workers Using Digital Technologies and Platforms
JMIR Form Res 2025;9:e84386
URL: https://formative.jmir.org/2025/1/e84386
doi:10.2196/84386
PMID:

XSL•FO
RenderX

©JMIR Publications Editorial Office. Originally published in JMIR Formative Research (https://formative.jmir.org), 08.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: A Neuroaffirmative, Self-Determination Theory–Based Psychosocial Intervention for Adults With Attention-Deficit/Hyperactivity Disorder: Randomized Feasibility Study

Rebecca Elizabeth Champ[1], BA, MA, PhD; Rita Wengorovius Meneses[2], LicPs, PhD; Marios Adamou[1*], MD, MSc, MA, LLM, PhD; Warren Gillibrand[1*], RN, MSc, PhD; Sally Arrey[1*], RLDN, MSc, PhD; Barry Tolchard[3*], RLDN, MSc, PhD

[1]University of Huddersfield, Huddersfield, United Kingdom
[2]Marketing Aplicado Lda (Portugal), Lisbon, Portugal
[3]University of Teeside, North Yorkshire, Middlesbrough, United Kingdom
[*]these authors contributed equally

Corresponding Author:
Rebecca Elizabeth Champ, BA, MA, PhD
University of Huddersfield
Queensgate
West Yorkshire
Huddersfield, HD1 3DH
United Kingdom
Phone: 44 1484 422288
Email: rebecca@rebeccachamp.co.uk

Related Article:

Correction of: https://formative.jmir.org/2025/1/e69943

(JMIR Form Res 2025;9:e87679) doi:10.2196/87679

In "A Neuroaffirmative, Self-Determination Theory–Based Psychosocial Intervention for Adults With Attention-Deficit/Hyperactivity Disorder: Randomized Feasibility Study" the authors noted one error.

In the originally published manuscript, an incorrect image appeared for Figure 1. The originally published image is available in Multimedia Appendix 1.

In the corrected manuscript, Figure 1 appears as follows:

Figure 1. Overview of ADAPT Framework pilot study assessment time and frequency of intervention sessions.

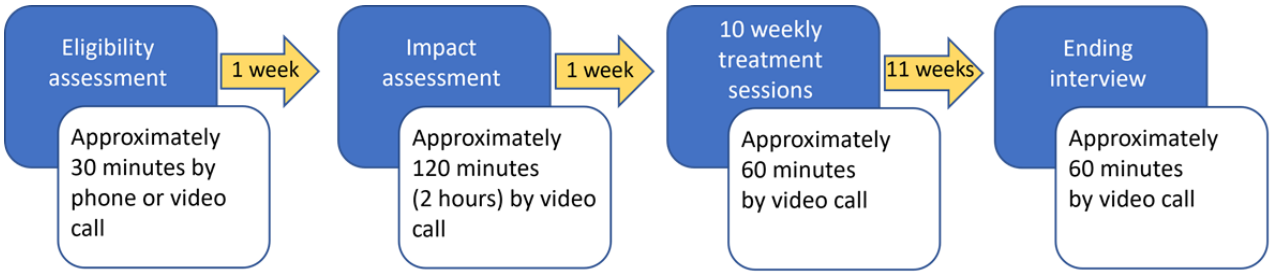

The correction will appear in the online version of the paper on the JMIR Publications website on December 4th 2025, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Multimedia Appendix 1
Originally published figure.

XSL•FO
RenderX

[PNG File , 161 KB - formative_v9i1e87679_app1.png]

Submitted 12.11.25; this is a non–peer-reviewed article;accepted 12.11.25; published 03.12.25.

Please cite as:
Champ RE, Meneses RW, Adamou M, Gillibrand W, Arrey S, Tolchard B
Correction: A Neuroaffirmative, Self-Determination Theory–Based Psychosocial Intervention for Adults With Attention-Deficit/Hyperactivity Disorder: Randomized Feasibility Study
JMIR Form Res 2025;9:e87679
URL: https://formative.jmir.org/2025/1/e87679
doi:10.2196/87679
PMID:

©Rebecca Elizabeth Champ, Rita Wengorovius Meneses, Marios Adamou, Warren Gillibrand, Sally Arrey, Barry Tolchard. Originally published in JMIR Formative Research (https://formative.jmir.org), 03.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Correction: Effectiveness, Usability, and Satisfaction of a Self-Administered Digital Intervention for Reducing Depression, Anxiety, and Stress in a University Community in the Andean Region of Peru: Randomized Controlled Trial

Rosario Yslado-Méndez[1*], PhD; Stefan Escobar-Agreda[2*], MD; David Villarreal-Zegarra[3,4], MSc; Wilfredo Manuel Trejo Flores[5], MSc; Junior Duberli Sánchez-Broncano[1], PhD; Ana Lucia Vilela-Estrada[3,4], MSc; Jovanna Hasel Olivares Córdova[1], MD, MSc; C Mahony Reategui-Rivera[6], MD; Claudia Alvarez-Yslado[7], BEd; Leonardo Rojas-Mezarina[2], MD, MSc

[1]Facultad de Ciencias Médicas, Universidad Nacional Santiago Antúnez de Mayolo, Av. Centenario 200, Huaraz, Peru
[2]Unidad de Telesalud, Facultad de Medicina, Universidad Nacional Mayor de San Marcos, Lima, Peru
[3]Digital Health Research Center, Instituto Peruano de Orientación Psicológica, Lima, Peru
[4]Universidad Científica del Sur, Lima, Peru
[5]Facultad de Ciencias, Universidad Nacional Santiago Antúnez de Mayolo, Huaraz, Peru
[6]Escuela de Posgrado, Universidad San Ignacio de Loyola, Lima, Peru
[7]Facultad de Ciencias Sociales, Educación, Comunicación, Universidad Nacional Santiago Antúnez de Mayolo, Huaraz, Peru
[*]these authors contributed equally

Corresponding Author:
Rosario Yslado-Méndez, PhD
Facultad de Ciencias Médicas, Universidad Nacional Santiago Antúnez de Mayolo, Av. Centenario 200, Huaraz, Peru

Related Article:

Correction of: https://formative.jmir.org/2025/1/e71465

Abstract

(JMIR Form Res 2025;9:e87717) doi:10.2196/87717

In "Effectiveness, Usability, and Satisfaction of a Self-Administered Digital Intervention for Reducing Depression, Anxiety, and Stress in a University Community in the Andean Region of Peru: Randomized Controlled Trial" [1], the authors made several adjustments.

The corresponding author's phone number was changed to the following:

51 042640020

The affiliation for authors RY-M, JDS-B, and JHOC has been changed from the following:

Facultad de Ciencias Sociales, Educación, Comunicación, Universidad Nacional Santiago Antúnez de Mayolo, Huaraz, Peru

Their affiliation now reads as:

Facultad de Ciencias Médicas, Universidad Nacional Santiago Antúnez de Mayolo, Huaraz, Peru

This has resulted in the change of affiliation 1. As author CA-Y is still affiliated with the original affiliation, it remains in the affiliation list as affiliation 7.

The affiliation for WMTF has been changed from:

Facultad de Ciencias Sociales, Educación, Comunicación, Universidad Nacional Santiago Antúnez de Mayolo, Huaraz, Peru

To read:

Facultad de Ciencias, Universidad Nacional Santiago Antúnez de Mayolo, Huaraz, Peru

Additionally, affiliation 4 has been added to the affiliations for ALVE. All other affiliations have been renumbered according to these amendments.

The correction will appear in the online version of the paper on the JMIR Publications website, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

XSL•FO
RenderX

Reference

1. Yslado-Méndez R, Escobar-Agreda S, Villarreal-Zegarra D, et al. Effectiveness, usability, and satisfaction of a self-administered digital intervention for reducing depression, anxiety, and stress in a university community in the Andean Region of Peru: randomized controlled trial. JMIR Form Res 2025 Oct 15;9:e71465. [doi: 10.2196/71465] [Medline: 41092084]

Submitted 13.11.25; this is a non–peer-reviewed article;accepted 13.11.25; published 22.12.25.

Please cite as:
Yslado-Méndez R, Escobar-Agreda S, Villarreal-Zegarra D, Flores WMT, Sánchez-Broncano JD, Vilela-Estrada AL, Córdova JHO, Reategui-Rivera CM, Alvarez-Yslado C, Rojas-Mezarina L
Correction: Effectiveness, Usability, and Satisfaction of a Self-Administered Digital Intervention for Reducing Depression, Anxiety, and Stress in a University Community in the Andean Region of Peru: Randomized Controlled Trial
JMIR Form Res 2025;9:e87717
URL: https://formative.jmir.org/2025/1/e87717
doi:10.2196/87717

© Rosario Yslado-Méndez, Stefan Escobar-Agreda, David Villarreal-Zegarra, Wilfredo Manuel Trejo Flores, Junior Duberli Sánchez-Broncano, Ana Lucia Vilela-Estrada, Jovanna Hasel Olivares Córdova, C Mahony Reategui-Rivera, Claudia Alvarez-Yslado, Leonardo Rojas-Mezarina. Originally published in JMIR Formative Research (https://formative.jmir.org), 22.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Correction: Online Health Information–Seeking Behaviors Among the Chongqing Population: Cross-Sectional Questionnaire Study

Honghui Rong[1*], MSc; Lu Lu[1*], MSc; Miao He[2], MSc; Tian Guo[2], MSc; Xian Li[2], MSc; Qingliu Tao[2], MSc; Yixin Li[2], MSc; Chuanfen Zheng[1], MSc; Ling Zhang[1], MSc; Fengju Li[1], MSc; Dali Yi[1], MSc; Enyu Lei[1], BS; Ting Luo[1], MSc; Qinghua Yang[2], MSc; Ji-an Chen[1], PhD

[1]Department of Health Education, School of Military Preventive Medicine, Army Medical University (Third Military Medical University), Gaotanyan Street No 30, Shapingba District, Chongqing, China

[2]Chongqing Health Education Institute, Chongqing, China

[*]these authors contributed equally

Corresponding Author:
Ji-an Chen, PhD
Department of Health Education, School of Military Preventive Medicine, Army Medical University (Third Military Medical University), Gaotanyan Street No 30, Shapingba District, Chongqing, China

Related Article:

Correction of: https://formative.jmir.org/2025/1/e56028

(JMIR Form Res 2025;9:e77500) doi:10.2196/77500

In "Online Health Information–Seeking Behaviors Among the Chongqing Population: Cross-Sectional Questionnaire Study" (JMIR Form Res 2025;9:e56028) the authors made one correction.

The affiliation of author Lu Lu, Chuanfen Zheng, Fengju Li, and Ting Luo, has been changed from affiliation 2:

Chongqing Health Education Institute

to affiliation 1:

Department of Health Education, School of Military Preventive Medicine, Army Medical University (Third Military Medical University), Chongqing, China

The correction will appear in the online version of the paper on the JMIR Publications website, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Submitted 15.05.25; this is a non–peer-reviewed article;accepted 15.05.25; published 30.05.25.

Please cite as:
Rong H, Lu L, He M, Guo T, Li X, Tao Q, Li Y, Zheng C, Zhang L, Li F, Yi D, Lei E, Luo T, Yang Q, Chen JA
Correction: Online Health Information–Seeking Behaviors Among the Chongqing Population: Cross-Sectional Questionnaire Study
JMIR Form Res 2025;9:e77500
URL: https://formative.jmir.org/2025/1/e77500
doi:10.2196/77500

© Honghui Rong, Lu Lu, Miao He, Tian Guo, Xian Li, Qingliu Tao, Yixin Li, Chuanfen Zheng, Ling Zhang, Fengju Li, Dali Yi, Enyu Lei, Ting Luo, Qinghua Yang, Ji-an Chen. Originally published in JMIR Formative Research (https://formative.jmir.org), 30.5.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Correction: Evaluating Bard Gemini Pro and GPT-4 Vision Against Student Performance in Medical Visual Question Answering: Comparative Case Study

Jonas Roos[1], MD; Ron Martin[2], MD; Robert Kaczmarczyk[3], MD

[1]Department of Orthopedics and Trauma Surgery, University Hospital of Bonn, Venusberg-Campus 1, 53127, Bonn, Germany

[2]Department of Plastic and Hand Surgery, Burn Center, BG Clinic Bergmannstrost, Halle (Saale), Germany

[3]Department of Dermatology and Allergy, Technical University of Munich, Munich, Germany

Corresponding Author:
Jonas Roos, MD
Department of Orthopedics and Trauma Surgery, University Hospital of Bonn, Venusberg-Campus 1, 53127, Bonn, Germany

Related Article:

Correction of: https://formative.jmir.org/2024/1/e57592

(JMIR Form Res 2025;9:e71664) doi:10.2196/71664

In "Evaluating Bard Gemini Pro and GPT-4 Vision Against Student Performance in Medical Visual Question Answering: Comparative Case Study" (JMIR Form Res 2024;8:e57592) the authors noted two errors.

Multimedia Appendices 1 and 2 contained identical files. Therefore, the Multimedia Appendix 1 file has been changed to the correct supplementary inference code.

Additionally, the "query" column of Multimedia Appendix 2 has been removed.

The corrections will appear in the online version of the paper on the JMIR Publications website on February 11, 2025, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Multimedia Appendix 1
Inference code.
[ZIP File, 2 KB - formative_v9i1e71664_app1.zip]

Multimedia Appendix 2
Responses from the models.
[ZIP File, 492 KB - formative_v9i1e71664_app2.zip]

Submitted 23.01.25; this is a non–peer-reviewed article;accepted 28.01.25; published 11.02.25.

Please cite as:
Roos J, Martin R, Kaczmarczyk R
Correction: Evaluating Bard Gemini Pro and GPT-4 Vision Against Student Performance in Medical Visual Question Answering: Comparative Case Study
JMIR Form Res 2025;9:e71664
URL: https://formative.jmir.org/2025/1/e71664
doi:10.2196/71664

© Jonas Roos, Ron Martin, Robert Kaczmarczyk. Originally published in JMIR Formative Research (https://formative.jmir.org), 11.2.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: Acceptability and Initial Adoption of the Heart Observation App for Infants With Congenital Heart Disease: Qualitative Study

Elin Hjorth Johansen[1,2], RN, MSCN; Elin Børøsund[3,4], RN, PhD; Ingeborg Martinsen Østen[1], RN, MSc; Henrik Holmstrøm[2,5], MD, PhD; Anne Moen[6], RN, PhD

[1]Neonatal Intensive Care Unit, Division of Children and Adolescent Medicine, Oslo University Hospital, Oslo, Norway
[2]Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo, Norway
[3]Department of Digital Health Research, Division of Medicine, Oslo University Hospital, Oslo, Norway
[4]Department of Nursing and Health Sciences, Faculty of Health and Social Sciences, University of South-Eastern Norway,, Drammen, Norway
[5]Department of Cardiology, Division of Children and Adolescent Medicine, Oslo University Hospital, Oslo, Norway
[6]Department of Public Health Science, Institute of Health and Society, Faculty of Medicine, University of Oslo, Oslo, Norway

Corresponding Author:
Elin Hjorth Johansen, RN, MSCN
Neonatal Intensive Care Unit
Division of Children and Adolescent Medicine
Oslo University Hospital
E2 4th Floor
Sognsvannsveien 20
Oslo, 0372
Norway
Phone: 47 98676884
Email: ehjorth@ous-hf.no

Related Article:

Correction of: https://formative.jmir.org/2023/1/e45920

(JMIR Form Res 2025;9:e70280) doi:10.2196/70280

In "Acceptability and Initial Adoption of the Heart Observation App for Infants with Congenital Heart Disease: Qualitative Study" (JMIR Form Res 2023; 7:e45920) the authors noted that author Elin Hjorth Johansen's affiliations need to be adjusted.

The following affiliation was added for author Elin Hjorth Johansen:

[2]Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo, Norway

The correction will appear in the online version of the paper on the JMIR Publications website on January 10, 2025, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Submitted 18.12.24; this is a non–peer-reviewed article;accepted 19.12.24; published 10.01.25.

Please cite as:
Hjorth Johansen E, Børøsund E, Martinsen Østen I, Holmstrøm H, Moen A
Correction: Acceptability and Initial Adoption of the Heart Observation App for Infants With Congenital Heart Disease: Qualitative Study
JMIR Form Res 2025;9:e70280
URL: https://formative.jmir.org/2025/1/e70280
doi:10.2196/70280
PMID:

XSL•FO
RenderX

©Elin Hjorth Johansen, Elin Børøsund, Ingeborg Martinsen Østen, Henrik Holmstrøm, Anne Moen. Originally published in JMIR Formative Research (https://formative.jmir.org), 10.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: Assessing and Improving Data Integrity in Web-Based Surveys: Comparison of Fraud Detection Systems in a COVID-19 Study

Stephen Bonett[1], RN, PhD; Willey Lin[1], MB; Patrina Sexton Topper[1], BSN, MS, PhD; James Wolfe[1], MS; Jesse Golinkoff[1], MPH; Aayushi Deshpande[2], MPhil; Antonia Villarruel[1], RN, PhD; José Bauermeister[1], MPH, PhD

[1]School of Nursing, University of Pennsylvania, Philadelphia, PA, United States

[2]Department of Psychology, Ashoka University, Sonepat, India

Corresponding Author:
Stephen Bonett, RN, PhD
School of Nursing
University of Pennsylvania
418 Curie Boulevard
Philadelphia, PA, 19104
United States
Phone: 1 2155734299
Email: stepdo@nursing.upenn.edu

Related Article:

Correction of: https://formative.jmir.org/2024/1/e47091/

(JMIR Form Res 2025;9:e76462) doi:10.2196/76462

In "Assessing and improving data integrity in web-based surveys: Comparison of fraud detection systems in a COVID-19 study" (JMIR Form Res 2024;8:e47091) the authors made one change.

The Acknowledgments section has been updated from:

> *This work was supported by funding from the National Institutes of Health Agreement OT2HL16156 as part of the Community Engagement Alliance Against COVID-19 Disparities, PhillyCEAL.*

to:

> *This research was, in part, funded by the National Institutes of Health (NIH) Agreement OT2HL158287. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the NIH.*

The correction will appear in the online version of the paper on the JMIR Publications website, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Submitted 23.04.25; this is a non–peer-reviewed article;accepted 25.04.25; published 30.04.25.

Please cite as:
Bonett S, Lin W, Sexton Topper P, Wolfe J, Golinkoff J, Deshpande A, Villarruel A, Bauermeister J
Correction: Assessing and Improving Data Integrity in Web-Based Surveys: Comparison of Fraud Detection Systems in a COVID-19 Study
JMIR Form Res 2025;9:e76462
URL: https://formative.jmir.org/2025/1/e76462
doi:10.2196/76462
PMID:

XSL•FO
RenderX

©Stephen Bonett, Willey Lin, Patrina Sexton Topper, James Wolfe, Jesse Golinkoff, Aayushi Deshpande, Antonia Villarruel, José Bauermeister. Originally published in JMIR Formative Research (https://formative.jmir.org), 30.04.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: Identifying Themes for an Initial Beta Version of a Mobile Health App for Latino and Native Hawaiian/Pacific Islander Communities: Co-Design and Community-Based Participatory Research in a Code to Community Study

Christina K Holub[1*], MPH, PhD; Amy L Barrera[1*], MPH; Rosalva Romero Gonzalez[2*], BA; Diane Hoang[1*], MPH; Luna Prieto[2*], BA; Samuelu Fesili[1*], BA; Tiana Smith[1*], BA; Harleen Kaur[1*], MPH; Cassandra Surban[1*], MPH; Michael Markidis[*], BS; Tana Lepule[3*], BA; Konane Martinez[2*], PhD

[1]Department of Public Health, California State University, San Marcos, San Marcos, CA, United States
[2]National Latino Research Center, California State University, San Marcos, San Marcos, CA, United States
[3]Pacific Islander Collective San Diego, San Diego, CA, United States
[*]all authors contributed equally

Corresponding Author:
Christina K Holub, MPH, PhD
Department of Public Health
California State University, San Marcos
333 S Twin Oaks Valley Road
San Marcos, CA, 92096
United States
Phone: 1 760 750 4295
Email: cholub@csusm.edu

Related Article:

Correction of: https://formative.jmir.org/2025/1/e76178/

(JMIR Form Res 2025;9:e79194) doi:10.2196/79194

In "Identifying Themes for an Initial Beta Version of a Mobile Health App for Latino and Native Hawaiian/Pacific Islander Communities: Co-Design and Community-Based Participatory Research in a Code to Community Study" (JMIR Form Res 2025;9:e76178), the authors note one omitted row in Table 1.

Table 1 has been revised to include the theme "generational approaches", its definition, and participant quotes as the last row.

Below is the added row with its definition and participant quotes:

Table 1. Themes, definitions, and participant quote examples (added row).

Theme	Definition; participant describes...	Example
Generational Approaches		
Intergenerational and multigenerational	Approaches for the different age groups or generations	• "And I think I'd agree with that really having the younger generation, you know…They're the ones that are being exposed to…nutrition classes in school… and really encouraging students to go home and share this with your parents. Go home and share this with…mom and dad or grandma and grandpa." [Latino participant] • "…If you had exclusively for that age group meeting. I don't think putting them all together in a forum would be the ideal." [NHPI participant]

The correction will appear in the online version of the paper on the JMIR Publications website, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

XSL•FO
RenderX

Submitted 16.06.25; this is a non–peer-reviewed article;accepted 23.06.25; published 02.07.25.

Please cite as:
Holub CK, Barrera AL, Romero Gonzalez R, Hoang D, Prieto L, Fesili S, Smith T, Kaur H, Surban C, Markidis M, Lepule T, Martinez K
Correction: Identifying Themes for an Initial Beta Version of a Mobile Health App for Latino and Native Hawaiian/Pacific Islander Communities: Co-Design and Community-Based Participatory Research in a Code to Community Study
JMIR Form Res 2025;9:e79194
URL: https://formative.jmir.org/2025/1/e79194
doi:10.2196/79194
PMID:

©Christina K Holub, Amy L Barrera, Rosalva Romero Gonzalez, Diane Hoang, Luna Prieto, Samuelu Fesili, Tiana Smith, Harleen Kaur, Cassandra Surban, Michael Markidis, Tana Lepule, Konane Martinez. Originally published in JMIR Formative Research (https://formative.jmir.org), 02.07.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: “Smartwatch-Based Ecological Momentary Assessment for High-Temporal-Density, Longitudinal Measurement of Alcohol Use (AlcoWatch): Feasibility Evaluation”

Chris Stone[1,2], BSc, RSci; Sally Adams[3], PhD; Robyn E Wootton[2,4,5,6], PhD; Andy Skinner[1], PhD

[1]Integrative Cancer Epidemiology Programme, Bristol Medical School, University of Bristol, Bristol, United Kingdom
[2]School of Psychological Science, University of Bristol, Bristol, United Kingdom
[3]School of Psychology, University of Birmingham, Birmingham, United Kingdom
[4]Nic Waals Institute, Lovisenberg Diaconal Hospital, Oslo, Norway
[5]Medical Research Council Integrative Epidemiology Unit, Bristol Medical School, University of Bristol, Bristol, United Kingdom
[6]PsychGen Centre of Genetic Epidemiology and Mental Health, Norwegian Institute of Public Health, Oslo, Norway

Corresponding Author:
Chris Stone, BSc, RSci
School of Psychological Science
University of Bristol
12a Priory Road
Bristol, BS8 1TU
United Kingdom
Phone: 44 07983 317748
Fax: 44 0117 928 8588
Email: chris.stone@bristol.ac.uk

Related Article:

Correction of: https://formative.jmir.org/2025/1/e63184

(JMIR Form Res 2025;9:e82194) doi:10.2196/82194

In “Smartwatch-Based Ecological Momentary Assessment for High-Temporal-Density, Longitudinal Measurement of Alcohol Use (AlcoWatch): Feasibility Evaluation” (JMIR Form Res 2025;9:e63184) the authors have updated the email address of the corresponding author.

Corresponding Author Chris Stone’s email address has been updated to: chris.stone@bristol.ac.uk

The correction will appear in the online version of the paper on the JMIR Publications website, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Submitted 11.08.25; this is a non–peer-reviewed article;accepted 20.08.25; published 01.10.25.

Please cite as:
Stone C, Adams S, Wootton RE, Skinner A
Correction: “Smartwatch-Based Ecological Momentary Assessment for High-Temporal-Density, Longitudinal Measurement of Alcohol Use (AlcoWatch): Feasibility Evaluation”
JMIR Form Res 2025;9:e82194
URL: https://formative.jmir.org/2025/1/e82194
doi:10.2196/82194
PMID:41032354

XSL•FO
RenderX

©Chris Stone, Sally Adams, Robyn E Wootton, Andy Skinner. Originally published in JMIR Formative Research (https://formative.jmir.org), 01.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: Overcoming Language Barriers in Paramedic Care With an App Designed to Improve Communication With Foreign-Language Patients: Nonrandomized Controlled Pilot Study

Frank Müller[1], Dr med; Dominik Schröder[1], MSc; Eva Maria Noack[1], Dr sc

Department of General Practice, University Medical Center Göttingen, Göttingen, Germany

Corresponding Author:
Eva Maria Noack, Dr sc
Department of General Practice
University Medical Center Göttingen
Humboldtallee 38
Göttingen, 37073
Germany
Phone: 49 55139 68193
Email: evamaria.noack@med.uni-goettingen.de

Related Article:

Correction of: https://formative.jmir.org/2023/1/e43255

(JMIR Form Res 2025;9:e83353) doi:10.2196/83353

In "Overcoming Language Barriers in Paramedic Care With an App Designed to Improve Communication With Foreign-Language Patients: Nonrandomized Controlled Pilot Study" [1] the authors made the following correction.

The first two sentences in the Results section of the Abstract were originally published as:

> *A total of 22 LGP patients were recruited into the interventional group and 122 into control group 1. The control group 2 included 27,212 German-speaking patients.*

They have now been corrected to:

> *A total of 22 LGP patients were recruited into the interventional group and 112 into control group 1. The control group 2 included 23,045 German-speaking patients.*

The correction will appear in the online version of the paper on the JMIR Publications website, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Reference

1. Müller F, Schröder D, Noack E. Overcoming Language Barriers in Paramedic Care With an App Designed to Improve Communication With Foreign-Language Patients: Nonrandomized Controlled Pilot Study. JMIR Form Res 2023 Mar 23;7(1):e43255 [FREE Full text] [doi: 10.2196/43255] [Medline: 36951895]

Submitted 01.09.25; this is a non–peer-reviewed article;accepted 03.09.25; published 21.10.25.

Please cite as:
Müller F, Schröder D, Noack EM
Correction: Overcoming Language Barriers in Paramedic Care With an App Designed to Improve Communication With Foreign-Language Patients: Nonrandomized Controlled Pilot Study
JMIR Form Res 2025;9:e83353
URL: https://formative.jmir.org/2025/1/e83353
doi:10.2196/83353
PMID:

XSL•FO
RenderX

©Frank Müller, Dominik Schröder, Eva Maria Noack. Originally published in JMIR Formative Research (https://formative.jmir.org), 21.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Communication Strategies for Tick-Borne Disease Prevention Among US Cattle Producers: Quasi-Experimental Study on Message Framing and Source Credibility

Cheng-Xian Yang[1], PhD; Lauri M Baker[2], PhD; Jessica Czipulis[3], MSci

[1]Graduate Institute of Bio-Industry Management, College of Agriculture and Natural Resources, National Chung Hsin University, Taichung City, Taiwan

[2]Department of Agricultural Education and Communication, Center for Public Issues Education in Agriculture and Natural Resources, University of Florida, 1408 Sabal Palm Drive, Level 2, P.O. Box 110126, Gainesville, FL, United States

[3]Department of Agricultural Education and Communication, College of Agricultural and Life Sciences, University of Florida, Gainesville, FL, United States

Corresponding Author:
Lauri M Baker, PhD
Department of Agricultural Education and Communication, Center for Public Issues Education in Agriculture and Natural Resources, University of Florida, 1408 Sabal Palm Drive, Level 2, P.O. Box 110126, Gainesville, FL, United States

Abstract

Background: *Haemaphysalis longicornis* (Asian longhorned tick) presents a growing threat to cattle health in the United States, causing anemia, weight loss, and even death. Despite the expanding distribution of this tick and the need for prevention, there is limited research on how to effectively communicate tick-borne disease risks to agricultural communities, particularly cattle producers. Social media represents a potentially impactful channel for risk communication; however, its utility and optimal message design for this population remain insufficiently understood.

Objective: This study aimed to evaluate the effectiveness of various digital communication strategies in influencing US cattle producers' engagement, risk perceptions, and self-reported preventive intentions regarding tick-borne diseases.

Methods: A quasi-experimental study with a 2×3 factorial design was conducted with 116 cattle producers in the United States. Participants were randomly assigned to view simulated Facebook posts that varied by message source (extension agent vs cattle producer) and message framing (prevention-oriented, fear-based, or neutral). Participants reported their likelihood of engaging with the post (using an emoji reaction, commenting, or sharing), their perceptions (fear and attitudes), and their intention to adopt preventive behaviors. Data were analyzed using descriptive statistics, ANOVA, and Spearman rank-order correlations.

Results: Across all groups, participants showed a clear preference for low-effort engagement, most frequently using emoji reactions rather than commenting or sharing. Descriptive trends suggested that posts from cattle producers, especially those framed as prevention oriented, elicited higher mean engagement. Prevention-oriented messages were also associated with greater concern about tick risks and more favorable attitudes toward prevention. As reflected in the qualitative feedback, fear-based posts were often viewed as exaggerated or less credible. Social media engagement showed moderate correlations with attitudes (ρ=0.52-0.64) and preventive intentions, with attitudes more strongly associated with behavioral intent than fear.

Conclusions: This study provides preliminary evidence that prevention-oriented digital messages, especially when shared by credible sources, may foster more positive attitudes and greater willingness to undertake tick prevention among cattle producers, compared to fear-based or neutral content. The findings highlight the need for balanced, context-rich messaging tailored to the preferences and realities of agricultural audiences. By integrating quantitative and qualitative results, this study informs the design of more effective social media–based interventions for tick-borne disease prevention in agricultural settings.

(JMIR Form Res 2025;9:e77239) doi:10.2196/77239

KEYWORDS

cattle producers; Haemaphysalis longicornis; message framing; quasi-experiment; risk communication; tick prevention

Introduction

Background

Haemaphysalis longicornis, also known as the Asian longhorned tick, is originally from East Asia but has established invasive populations in regions such as Australia, New Zealand, and several Pacific islands. This is an invasive species that was first detected in the United States in 2017 and has since been reported in 19 states, particularly in the eastern and southeastern regions [1-5]. The rapid proliferation of *H longicornis* across the United

XSL•FO
RenderX

States can be attributed to its broad host range and its ability to reproduce rapidly through a parthenogenetic life cycle, allowing females to reproduce without the need for males [5-9].

H longicornis is considered a significant pest, particularly for livestock, where it poses substantial challenges to agricultural productivity and animal health. It is particularly dangerous to the livestock industry because it is a known vector of *Theileria orientalis* Ikeda, a hemoparasite that causes bovine theileriosis, in the United States [2,6,8,10-13]. Heavy infestations of *H longicornis* on cattle can have direct, detrimental impacts on cattle health, including significant blood loss, anemia, reduced production, weight loss, and even death [2,4,5,7,8]. Due to its expanding range and ability to carry pathogens, this tick species is considered a growing concern for the agricultural industry and public health in the United States [1,14].

Effectively managing the risks posed by *H longicornis* to the livestock industry requires comprehensive tick control strategies tailored for cattle producers. Recommended practices include mechanical approaches such as tick surveillance, manual removal, and vegetation management to limit tick habitats, as well as chemical measures like the application of on-animal acaricides or tick repellents [10,12,15,16]. Maintaining a closed herd is also advised to prevent the introduction and spread of ticks among livestock [10]. However, for these preventive strategies to be adopted effectively, raising awareness of tick-borne diseases and the risks they pose is essential. Educational interventions targeted at livestock producers can improve knowledge and awareness and promote the consistent use of proven prevention and control methods [17-19].

Supporting this need, prior research has shown that individuals with higher knowledge levels and greater awareness and concern about tick-borne diseases are more likely to engage in preventive behaviors [20-23]. Nevertheless, a widespread lack of confidence in the effectiveness of tick prevention products such as acaricides and repellents may discourage their use [23]. It has been reported that cattle producers often apply acaricides only occasionally and in response to visible tick presence rather than as part of a consistent, preventive routine [24]. Previous studies recommend that educational efforts address disease transmission routes, tick biology, and validated control strategies to support more systematic management practices. These findings highlight the importance of communication in shaping not only knowledge but also behavior.

Building on this insight, this study approached the issue from a risk communication perspective to better understand how information about tick-borne disease prevention in the context of *H longicornis* is perceived and acted upon by cattle producers. The goal was to inform more effective communication strategies that enhance awareness, improve risk perception, and ultimately encourage the adoption of preventive behaviors among this high-risk population.

Literature Review and Theoretical Framework

Effective disease communication strategies can enhance public understanding, influence decision-making, and promote behaviors that support disease prevention [25-27]. To ensure maximum impact, communicators should customize their strategies to fit the specific behaviors and needs of target populations, including those at risk for tick-borne diseases, and account for varying contexts [28-30]. For example, digital technologies and online platforms have rapidly expanded the ways in which risk information can be delivered and received.

The rise of social media has revolutionized global communication, especially in fields like agriculture, where these platforms serve as powerful tools for disseminating technology and information [31]. Social media, with its potential for real-time updates and interactive engagement, is suited for risk communication, including that for tick-borne diseases [32-34]. These features allow communicators to effectively engage with the public, fostering a better understanding and management of tick risks [19,35]. Gupta et al [17] further indicated that social media appears to be an effective communication channel for spreading knowledge about tick-borne diseases. Platforms such as Facebook (Meta Platforms, Inc), X (formerly Twitter; X Corp), and YouTube (Google Inc) have become essential channels for sharing scientific information, making them indispensable for communicating risks [36,37].

The interactive nature of social media promotes risk perception and informed decision-making during disease outbreaks. Previous studies have pointed out that social media content about vector-borne diseases is positively related to users' awareness and knowledge of the risk and preventive actions [38-40]. Although previous research suggests that cattle producers in the United States generally have a lower preference for and lower trust in social media [41], it also highlights that information shared through social media by credible sources such as peer cattle producers and beef industry organizations is often considered reliable [42]. Therefore, while social media may not be the preferred communication tool for this demographic, its importance in reaching broader audiences and conveying critical disease prevention messages cannot be underestimated.

In the context of US cattle producers, extension agents and peer cattle producers represent two of the most influential yet contrasting sources in terms of credibility and practical relevance. Extension agents have institutional expertise and are scientific authority figures. They serve as formal representatives of land-grant universities or government agricultural agencies. Their credibility stems from professional training, research-based knowledge, and access to current scientific information. Many producers view them as reliable and unbiased sources of practical advice. Effective communication by extension agents, especially when combined with interpersonal skills and an understanding of producer needs, can enhance the adoption of innovations and best practices [43-45]. In contrast, peer cattle producers are trusted for their practical experience and shared challenges. Producers see peers as credible because they have faced similar decisions in the same environment. Peer learning and observation strongly influence behavioral change, as agricultural producers trust advice from those living comparable realities [46,47]. Unlike extension agents, peers shape behavior through social connections and proven local success, offering a grassroots influence that complements expert-driven approaches.

XSL•FO
RenderX

Additionally, designing communication content that resonates with the target audience enhances the effectiveness of risk communication. Framing theory is often used to explain how the presentation of information through communication channels influences how audiences interpret and respond to a message. A frame highlights certain aspects of an issue while downplaying others, thus shaping the audience's perception of the topic [48]. In the context of disease communication, gain-framed messages focus on prevention and present outcomes as clear and low-risk, conveying the idea that the disease is manageable. On the other hand, loss-framed messages, which emphasize the potential health risks, tend to communicate greater uncertainty and evoke stronger negative emotions such as fear [19,49-51]. Gain-framed appeals, which highlight the advantages of compliance, have statistically higher engagement rates and are more persuasive than loss-framed appeals, but the difference in effectiveness is quite small [52,53]. While this finding is supported by several studies [49,54], other research has demonstrated that framing messages around the severity of risk can also be effective in encouraging preventive actions [50,55]. Li et al [56] found that social media posts combining both positive and negative content generate more comments and reactions than those with purely negative content.

Purpose

This exploratory study aimed to identify effective social media communication strategies for cattle producers in the United States and to explore the factors related to their risk-preventive intentions regarding the impact of *H longicornis* on the livestock industry. Adoption of personal preventive actions varies depending on individuals' perceptions of tick-borne disease risks and their demographic characteristics. Therefore, targeted health promotion communication is essential for high-risk groups exposed to ticks. This area remains underresearched, and communicators should focus on ensuring that these at-risk populations adopt necessary prevention behaviors [15-17,19,30,57]. The following research questions were developed to align with the research objectives.

1. Which message framing is observed to engage cattle producers more when presenting the risks of *H longicornis*?
2. How do cattle producers perceive different types of message framing regarding fear, attitudes, and consideration of preventive behavioral intentions regarding the risks of *H longicornis*?
3. What are the associations among social engagement, perceptions, and preventive behavioral intentions to identify the key variables in cattle producers' responses?

Methods

Population and Sample

This quantitative study surveyed cattle producers in the United States using an online survey distributed via Qualtrics. Given the difficulty in reaching the target population, both online and offline methods were used to gather responses through convenience sampling. At the annual convention of the National Cattlemen's Beef Association (NCBA) in February 2024 in Orlando, Florida, we collected data directly from cattle producers. They were asked to complete the online questionnaire using iPads (Apple) and QR codes, resulting in 68 responses. Additionally, from May to June 2024, we collected 48 more responses through Facebook ads and advertisements in cattle magazines. A total of 116 valid responses were obtained in this study. Nonprobability sampling was adopted as a suitable method for making population estimates in this context because it allows targeting specific, hard-to-reach populations while maintaining a broad outreach, even though it does not involve random selection [58].

Research Design and Instrument

At the beginning of the questionnaire, we asked participants to identify the primary class of commodity they grow or produce. If "cattle" was not selected as one of the primary commodities, the respondent was excluded from the study to ensure that the collected sample aligned with our target population. Variables such as location and annual sales for the agricultural operation were also collected to gather comprehensive background information on the participants. Additionally, a pretest question, which required a response on a 5-point Likert scale, was asked to understand the respondent's support for tick risk education.

The study used a quasi-experimental approach. To ensure complete and transparent reporting, this study adhered to the Checklist for Reporting Results of Internet E-Surveys (CHERRIES) [59], appropriate for quasi-experimental studies and web-based interventions. Specifically, we followed the recommendations for reporting internet-based surveys to ensure clear description of the participant recruitment and allocation procedures and data collection methods. These guidelines supported consistent and detailed reporting of our study methods and helped ensure that the study is interpretable and reproducible by others.

A 2×3 factorial design was implemented to examine the influence of message source (extension agent versus cattle producer) and message framing (prevention-oriented, fear-based, or neutral) on producer perceptions and behavioral intentions related to tick prevention. This design created 6 experimental conditions, with each participant randomly assigned to view 1 message from each source, totaling 2 messages per participant. Random assignment was used to minimize selection bias and balance potential confounders across groups.

To guage cattle producers' perceptions of social media posts, this study developed a survey instrument focused on creating realistic simulated Facebook posts. Zeoob, an online educational tool that can generate simulated social media posts [60], was chosen for this purpose. Six posts were created, all related to *H longicornis*, and attributed to 2 personas—an extension agent and a cattle producer. Each persona shared 3 posts with different content. The content framing varied from a prevention-oriented message, which was gain framed (M1: "It is crucial to spray your pastures and animals to prevent ticks from infecting the herd."), to a fear-based message, which was loss framed (M2: "Watch out! Cow-killing ticks are spreading across the U.S."), and included a middle-ground approach (M3: "You don't want to learn the hard way. Entire herds across the U.S. have been lost already.").

XSL•FO
RenderX

Simulated posts were presented via the online survey platform. We excluded responses that were completed too quickly to ensure that respondents had sufficient exposure time to view each post. Participants could not revisit previous messages. Participants could view only the posts assigned to their experimental condition and had no opportunity to view posts from other subgroups.

The survey questions assessed participants' likelihood of social engagement after viewing the social media post. Participants were asked how they would react, including choosing an emoji, commenting, or sharing the post. The likelihood of engagement was measured on a 5-point Likert scale (1=unlikely, 2=somewhat likely, 3=likely, 4=very likely, and 5=extremely likely). If participants indicated that they were at least somewhat likely to react by liking the post, they were further asked to select which emoji they might use. These emojis, modeled after those on Facebook, included 7 options such as like, love, and sad.

We also assessed their perceptions of the post, including their fear of *H longicornis* and the diseases it spreads as well as their attitudes supporting proactive preventive intentions and raising awareness about tick-related risks. To assess the participants' preventive behavioral intentions in response to tick-borne diseases, we focused on 2 key measures: removing ticks through surveillance and spraying cattle with on-animal acaricides or tick repellents [10]. Each measure was assessed with 2 questions, all answered on a 5-point Likert scale (1=strongly disagree to 5=strongly agree).

Moreover, to enrich our findings, we included open-ended questions to capture the participants' thoughts and reflections on the simulated social media posts viewed in their assigned experimental condition. Each respondent answered up to 28 questions, including 6 about their baseline characteristics as well as 2 sets of quasi-experimental questions (11 questions each). Figure 1 depicts the research process and the simulated social media posts.

Figure 1. Research process and illustration of the quasi-experiment.

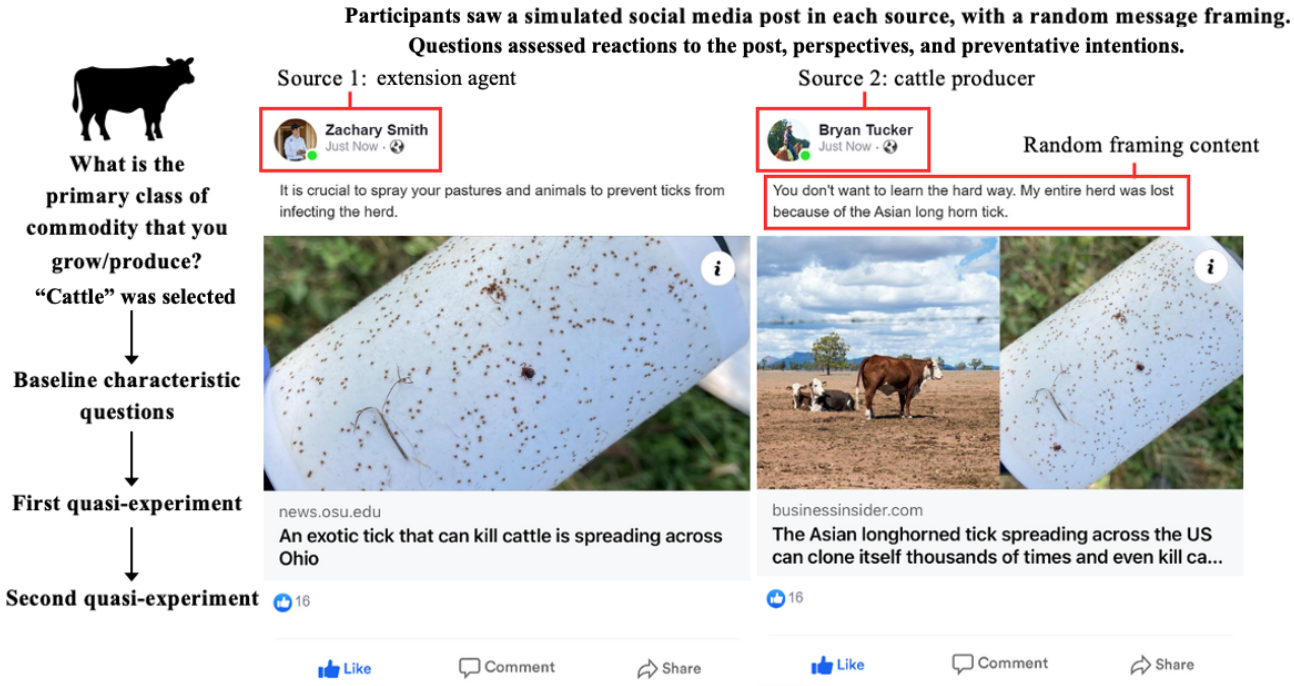

Data Analysis

SPSS software (version 29.0; IBM Corp) was used to analyze the quantitative data. The random assignment method was expected to distribute potential confounders evenly. We conducted ANOVA tests to compare baseline demographic variables and pretest attitudes across the randomized groups, using group assignment as the independent variable. All comparisons were nonsignificant, indicating that the groups did not differ in these baseline measures. To answer our first and second research questions, descriptive statistics were first used to explain how participants responded to different messages. ANOVA tests with post hoc comparisons were completed using the Scheffe correction. For the third research question, Spearman rank-order correlations were used to examine the associations between their social engagement on social media, perceptions of the post, and reported preventive intentions, given the ordinal nature of the Likert-type items.

Qualitative data were analyzed using a thematic analysis approach. First, all responses were read in full to ensure familiarization with the content. Initial codes that captured distinct concepts or sentiments were identified. Then, through an iterative process, similar codes were grouped into broader themes that illustrated cattle producers' perspectives on message tone, credibility, perceived relevance, and barriers or motivators to engagement. Representative quotes were selected to illustrate each major theme. This procedure ensured a rigorous and transparent analysis of qualitative inputs, complementing the quantitative survey results.

XSL•FO
RenderX

Ethical Considerations

This study was reviewed and approved by the Institutional Review Board at the University of Florida (IRB#202201562). Written informed consent was obtained electronically from each participant prior to their participation. Participants were briefed about the study objective, procedures, voluntary nature of participation, and the right to decline or withdraw at any time without penalty. All data collected were anonymized before analysis. Participants were given a tick removal kit for their participation in the study. The value of each kit was under US $2.

Results

Descriptive Statistics of Agricultural Operation Characteristics

Due to convenience sampling, most of the respondents came from southeastern states in the United States, such as Florida (30/116, 25.9%), Tennessee (20/116, 17.2%), and Kentucky (14/116, 12.1%). The agricultural operations of most cattle producers (98/116, 84.5%) were located in rural areas. Approximately half of the respondents (55/116, 47.4%) had annual sales of less than US $50,000, followed by those with sales between US $50,001 and US $150,000 (24/116, 20.7%). These figures reflect the national data from the 2022 Census of Agriculture, which reported an average annual agricultural income of US $79,790 [61]. Regarding their role in agriculture, 89 participants (76.7%) identified as owner or operator. Detailed characteristics of the sample are shown in Table 1.

Table . Characteristics of the sample (N=116).

Characteristic	Participants, n (%)
State of origin	
Florida	30 (25.9)
Tennessee	20 (17.2)
Kentucky	14 (12.1)
Texas	9 (7.8)
Virginia	9 (7.8)
Georgia	8 (6.9)
Other	26 (22.4)
Agricultural operation location	
Rural area	98 (84.5)
Urban or suburban area outside the city limits	14 (12.1)
Urban or suburban area inside the city limits	4 (3.4)
Agricultural operation annual sales	
Less than US $50,000 in gross annual sales	55 (47.4)
Between US $50,001 and US $150,000 in gross annual sales	24 (20.7)
Between US $150,001 and US $249,999 in gross annual sales	12 (10.3)
Between US $250,000 and US $500,000 in gross annual sales	13 (11.2)
More than US $500,000 in gross annual sales	12 (10.3)
Role in agriculture	
Owner/operator	89 (76.7)
Hired labor	9 (7.8)
Consumer	4 (3.4)
Other	14 (12.1)

Social Engagement on Social Media

To answer the first research question, in the quasi-experiment, we assessed cattle producers' self-reported engagement with different simulated social media posts, specifically through choosing an emoji, commenting, and sharing the post. While the ANOVA did not reveal statistically significant differences across the 6 combinations of message source and message type, we observed several noteworthy descriptive patterns in the engagement data.

Overall, the results indicated that cattle producers are most likely to engage with posts through emoji reactions compared

XSL•FO
RenderX

to commenting and sharing. Among the various emojis available on Facebook, the "Like" emoji remained the most frequently used in all experimental conditions, even compared to newer emojis. Commenting was generally less likely across all combinations of message source and type. The mean scores for commenting were consistently lower than those for emoji reactions and sharing, suggesting that while cattle producers may react to or share posts, they are less likely to engage in discussions or leave comments on these posts.

Further examination showed some variation in the mean engagement scores across the combinations of message source and type and highlighted some trends. For instance, the highest average likelihood score for emoji engagement (mean 2.89, SD 1.30) was observed for the cattle producer × M1 (cattle producer, prevention-oriented message) group, while the lowest (mean 2.38, SD 1.33) was observed for the extension agent × M2 (extension agent, fear-based message) group. Similar descriptive differences were noted for the sharing and commenting behaviors. However, it is important to note that these observed patterns did not reach statistical significance in this analysis and thus should be interpreted with caution as preliminary trends rather than as definitive group differences. Table 2 presents the descriptive results for social engagement across all experimental conditions.

Table . Likelihood of social engagement with simulated Facebook posts among participants in 6 experimental conditions.

Type of engagement	Likelihood score[a], mean (SD)					
	EA[b] × M1[c] (n=40)	EA × M2[d] (n=34)	EA × M3[e] (n=40)	CP[f] × M1 (n=36)	CP × M2 (n=32)	CP × M3 (n=38)
Emoji	2.60 (1.37)	2.38 (1.33)	2.81 (1.37)	2.89 (1.30)	2.42 (1.17)	2.76 (1.46)
Comment	2.05 (1.36)	1.62 (1.13)	1.86 (1.24)	2.00 (1.29)	1.72 (1.05)	2.03 (1.40)
Share	2.33 (1.39)	2.35 (1.43)	2.30 (1.52)	2.42 (1.42)	2.03 (1.18)	2.63 (1.53)

[a]Likelihood was rated on a 5-point Likert scale.

[b]EA: extension agent.

[c]M1: prevention-oriented message.

[d]M2: neutral message.

[e]M3: fear-based message.

[f]CP: cattle producer.

Perceptions of Simulated Social Media Posts and Preventive Intentions

To answer the second research question, we assessed several important dimensions, including participants' fear and concern about tick-related risks, attitudes supporting preventive behaviors and local awareness, and their self-reported intentions to take preventive actions following exposure to the social media posts. Across these outcomes, while ANOVA results did not indicate significant differences between groups, we identified some descriptive patterns that may offer preliminary insights for future studies.

Regarding the respondent's perception of the post, it was observed that the source of the message alone did not make a difference. However, when combined with the content of the message, distinct trends began to emerge. Specifically, when either an extension agent or a cattle producer used a prevention-oriented message to communicate the risks of *H longicornis* to the cattle, participants expressed higher levels of concern about tick risks. They also demonstrated more positive attitudes toward managing the risks associated with ticks. Conversely, when a fear-based message was employed as the communication strategy, it led to lower perceptions of post credibility or effectiveness.

Regarding responses related to preventive intentions, this study observed distinct patterns across different messaging sources and types. An analysis of responses to messages from an extension agent showed that the neutral message proved effective in prompting preventive intentions. For example, the extension agent × M2 condition had a relatively high mean score for examining gear and cattle for ticks (mean 3.94, SD 0.75) and using tick repellents (mean 3.88, SD 1.02). However, when cattle producers were the source of the message, the data revealed somewhat mixed outcomes. The prevention-oriented message had a relatively high mean score for encouraging cattle producers to examine gear and cattle for ticks as a precaution (mean 3.94, SD 0.89), but this communication strategy was less effective in promoting the use of tick repellents to avoid tick bites (mean 3.62, SD 1.23). Table 3 shows the perceptions of the posts and preventive behavioral intentions among participants in various experimental conditions.

XSL•FO
RenderX

Table . Perceptions of simulated Facebook posts and preventive intentions among participants in 6 experimental conditions.

Statement	Rating[a], mean (SD)					
	EA[b] × M1[c] (n=39)	EA × M2[d] (n=33)	EA × M3[e] (n=40)	CP[f] × M1 (n=34)	CP × M2 (n=33)	CP × M3 (n=34)
Fear						
Worries about cattle health	3.77 (0.90)	3.52 (1.09)	3.60 (0.84)	3.85 (0.93)	3.73 (0.67)	3.50 (1.02)
Fear of tick risks	3.28 (0.92)	3.09 (0.95)	3.22 (0.89)	3.53 (0.93)	3.39 (0.75)	3.32 (1.07)
Attitude						
Attitude toward educating the community about tick risks	4.08 (0.74)	3.88 (1.17)	4.07 (0.94)	4.09 (0.90)	4.00 (0.71)	3.91 (0.90)
Attitude toward regular checks for ticks in the surroundings	4.08 (0.87)	4.15 (0.83)	3.68 (1.07)	3.97 (0.87)	3.94 (0.75)	3.79 (0.85)
Preventive intentions						
Regular examination of gear and cattle for ticks	3.77 (0.87)	3.94 (0.75)	3.50 (1.13)	3.94 (0.89)	3.70 (0.81)	3.82 (0.97)
Use of tick repellents to avoid tick bites	3.67 (0.90)	3.88 (1.02)	3.67 (1.07)	3.62 (1.23)	3.82 (0.81)	3.97 (0.83)

[a]Statements were rated on a 5-point Likert scale.

[b]EA: extension agent.

[c]M1: prevention-oriented message.

[d]M2: neutral message.

[e]M3: fear-based message.

[f]CP: cattle producer.

Open-Ended Responses From Cattle Producers

To interpret and supplement the quantitative results, we further asked participants for their opinions on the simulated social media posts to gain deeper insights into their thoughts on the experimental conditions. Some responses were brief affirmations, such as “Good job” or “Great info.” These responses suggest that certain producers viewed the posts as useful and well presented, even if the messages did not always prompt further engagement or detailed feedback. Conversely, some participants offered minimal or neutral input, with responses such as “No” or “N/A.” Such replies indicate that not all respondents felt strongly about the content or were motivated to provide additional commentary.

After excluding brief or ambiguous responses, thematic analysis was conducted to identify distinct concepts emerging from the qualitative data. This approach yielded 3 dominant themes. The most frequently occurring theme was criticism of sensationalism and scare tactics. Respondents expressed concern over the use of alarming or exaggerated language in the social media posts. For instance, comments such as “Seems like a fear tactic instead of informational,” “Over exaggerated,” and “Sounds like a scare tactic by PETA [People for the Ethical Treatment of Animals].” All comments about the content being exaggerated were about posts with a fear-based or neutral message, with slightly more comments directed at posts with fear-based messages. This underscores the perception that fear-based messaging may undermine credibility, potentially resulting in skepticism or disengagement among the target audience. While some participants acknowledged that such messages were effective in capturing attention, several emphasized that communication should prioritize education over sensationalism, as reflected in the following comment: “The post should be more educational rather than shocking.”

Another closely related but different theme was social media skepticism and fatigue. Under this theme, participants expressed doubt not only about the content of the posts but also about the reliability and intentions behind social media communication in general. Comments such as “[It was] too much of a scare tactic to get someone to click the link” and “Less clickbait, obviously untrue” reflect broader concerns about the prevalence of clickbait in social media environments. Moreover, comments such as “I would look for verification on something other than Facebook before I reacted” suggest a distrust of information on social media. The primary distinction between this theme and the previous one lies in the focus: while criticism of sensationalism centers on the content and style of the specific posts, social media skepticism and fatigue reflect a deeper mistrust toward the platform itself and weariness with manipulative online communication strategies.

XSL•FO
RenderX

Additionally, a prominent theme was the desire for more detailed and contextual information. Several participants explicitly stated the need for "more information" or greater context, as exemplified by comments such as "Could use more information in general," "I would like it to elaborate more," and "I think more context should be included!" These responses may suggest that while social media posts can raise initial awareness, producers are looking for more substantive and actionable content that offers practical guidance on tick risks and prevention strategies. The call for additional information likely stems from the perception that the information in the simulated posts lacked sufficient detail or depth to fully inform decision-making or address producers' informational needs. Additionally, some respondents highlighted the importance of factual accuracy, expressing a preference for information backed by evidence, through responses such as "Would react to see back up verification that it was real." All comments under the theme of desire for more detailed and contextual information were directed at posts with fear-based or neutral messages, with the latter receiving more comments.

Associations Between Social Engagement, Perceptions, and Preventive Intentions

The analysis of social engagement, which included using emoji reactions, commenting, and sharing, revealed associations with risk perceptions, attitudes, and reported preventive intentions. Emoji reactions were highly correlated with both sharing (ρ=0.83, P<.001) and commenting (ρ=0.80, P<.001), suggesting that participants who engaged with posts through emojis were also more inclined to engage further by sharing or commenting. Furthermore, social engagement exhibited meaningful associations with perceptions and preventive actions. Participants' interactions with social media posts showed low to moderate correlations with their perceptions and subsequent behavioral intentions. For instance, those who were more willing to share the post tended to have moderate concerns about the health risks ticks pose to their cattle (ρ=0.32, P <.001) and were also more likely to engage in preventive intentions such as regularly examining their gear and cattle for ticks (ρ=0.25, P<.001) and using tick repellents (ρ=0.29, P<.001).

The results also indicated that fears and attitudes toward managing the risks associated with ticks had a positive association with preventive intentions. The association between fear and preventive intentions was low to moderate. For instance, producers' worries about cattle health because of ticks were moderately correlated with regular examination of gear and cattle for ticks (ρ=0.34, P<.001) and the use of tick repellents to avoid tick bites (ρ=0.33, P<.001). On the other hand, attitudes formed after viewing the posts were closely linked to the likelihood of engaging in preventive intentions against tick-borne risks, showing substantial to very strong associations (ρ=0.52 - 0.64). All correlations reported in this study were positive. While these correlations point to meaningful links between risk perception, engagement, and behavioral intentions, it is important to emphasize that given the observational design, these associations do not imply prediction or causation. Table 4 shows the associations between social engagement, perceptions, and preventive intentions among the experimental conditions.

XSL•FO
RenderX

Table . Associations[a] between various aspects of social engagement, perceptions, and preventive intentions.

	Emoji use	Comment	Share	Worries about cattle health	Fear of tick risks	Attitude toward educating the community about tick risks	Attitude toward regular checks for ticks in the surroundings	Regular examination of gear and cattle for ticks	Use of tick repellents to avoid tick bites
Emoji use, ρ	1.00	0.80[b]	0.83[b]	0.27[b]	0.35[b]	0.13	0.08	0.15[c]	0.16[d]
Comment, ρ	0.80[b]	1.00	0.79[b]	0.22[b]	0.30[b]	0.10	0.10	0.20[b]	0.21[b]
Share, ρ	0.83[b]	0.79[b]	1.00	0.32[b]	0.37[b]	0.18[b]	0.13	0.25[b]	0.29[b]
Worries about cattle health, ρ	0.27[b]	0.22[b]	0.32[b]	1.00	0.57[b]	0.46[b]	0.38[b]	0.34[b]	0.33[b]
Fear of tick risks, ρ	0.35[b]	0.30[b]	0.37[b]	0.57[b]	1.00	0.32[b]	0.32[b]	0.20[b]	0.27[b]
Attitude toward educating the community about tick risks, ρ	0.13	0.10	0.18[b]	0.46[b]	0.32[b]	1.00	0.54[b]	0.53[b]	0.53[b]
Attitude toward regular checks for ticks in the surroundings, ρ	0.08	0.10	0.13	0.38[b]	0.32[b]	0.54[b]	1.00	0.64[b]	0.52[b]
Regular examination of gear and cattle for ticks, ρ	0.15[c]	0.20[b]	0.25[b]	0.34[b]	0.20[b]	0.53[b]	0.64[b]	1.00	0.58[b]
Use of tick repellents to avoid tick bites, ρ	0.16[d]	0.21[b]	0.29[b]	0.33[b]	0.27[b]	0.53[b]	0.52[b]	0.58[b]	1.00

[a]Strength of the association is indicated by the value of the correlation coefficient [62]: 0.01-0.09, negligible; 0.10 - 0.29, low; 0.30 - 0.49, moderate; 0.50 - 0.69, substantial; >0.70, very strong.

[b]$P<.001$.

[c]$P=.03$.

[d]$P=.02$.

Discussion

This study provides a clearer understanding of the effectiveness of digital risk communication strategies among cattle producers, specifically regarding social media engagement, risk perception, and preventive intentions in response to messages about *H longicornis*. While some observed differences did not reach statistical significance in our quasi-experimental study, several noteworthy descriptive trends and participant insights emerged from the qualitative data, providing valuable direction for both future research and practical science communication. These findings fill a gap in research on the social media health education interventions on people's perceptions and preventive intentions [16].

First, various message frames under different conditions were analyzed to better inform communication strategies targeting cattle producers in the United States regarding the health risks posed by *H longicornis* to their livestock. Results showed that cattle producers were more likely to react to posts using emojis than to share or comment, suggesting a preference for low-effort engagement when processing tick-related information. This is consistent with previous digital communication research, indicating that while such responses can enhance message visibility [63], they may not fully reflect deeper cognitive engagement. The "Like" emerged as the most frequently used reaction across all experimental conditions. This aligns with prior research suggesting that clicking the "Like" button is often a cognitively automatic response [63,64]. Users may favor it because it requires minimal mental effort and serves as a default expression of acknowledgment or general approval without necessitating deeper emotional processing or critical engagement.

Contrary to the findings of O'Keefe and Jensen [53], we did not observe statistically significant differences in social media

XSL•FO
RenderX

engagement rates between prevention-oriented and fear-based messages, nor did gain-framed messages lead to higher engagement [56]. This result aligns with the findings of Yang et al [19], who observed no clear distinction in social media engagement when tick risks were communicated using either gain or loss framing. Nevertheless, both gain- and loss-framed messages showed a trend of higher levels of social engagement than neutral messages communicating the risks associated with *H longicornis*, regardless of whether the message source was an extension agent or a fellow cattle producer. This contrasts with the findings of Li et al [56], which reported that posts with a middle-ground approach containing both positive and negative content generated higher social engagement. We infer that this descriptive difference may be due to the special nature of the case. Agricultural risks, which often directly affect livestock and livelihoods, may prompt cattle producers to respond more strongly to content that emphasizes these risks than to content that presents risks mildly. This finding highlights that emotionally charged and prevention-oriented content may attract more attention than neutral content, although these differences were not statistically significant in this sample. Despite this finding, qualitative data revealed that some participants have a distrust of the social media platform itself and experienced information fatigue, which in turn affected their responses to the messages.

The results also revealed descriptive patterns suggesting associations between different communication strategies and cattle producers' perceptions of *H longicornis* risks. Although the message source alone did not appear to substantially affect perceptions, the combination of source and content framing showed some preliminary trends. Specifically, prevention-oriented (gain-framed) messages, whether delivered by an extension agent or fellow cattle producer, were associated with higher levels of concern about tick risks and more positive attitudes toward community education and routine livestock checks. This observation is generally consistent with previous research on prospect theory [49,51,54], which indicates that gain-framed information tends to be more persuasive when promoting actionable outcomes. In this study, the prevention-oriented messages highlighted specific actions that could mitigate risks, potentially supporting the positive attitudinal responses.

Conversely, fear-based (loss-framed) messaging tended to elicit lower perceptions of post credibility or effectiveness. Our qualitative findings further enrich the understanding of communication effectiveness. Some cattle producers reported being less inclined to engage with or trust posts they viewed as sensationalized and did not prefer excessive use of clickbait. Despite acknowledging that emotionally charged language draws attention, several participants cautioned that such approaches can undermine trust or provoke skepticism, echoing broader findings in health communication that fear-based messaging, while memorable, may backfire among certain segments of the population [29,40,52,65,66]. These insights provide actionable guidance that effective digital outreach should balance the need to capture attention with the delivery of accurate, context-rich information that empowers rather than alarms the audience [31,33].

With respect to reported preventive intentions, differences among message types were subtle and did not reach statistical significance. Descriptively, neutral messaging from an extension agent appeared to be relatively more effective in encouraging intentions to undertake both mechanical and chemical tick prevention measures. In contrast, fear-based messaging was associated with a trend of lower engagement in preventive intentions. Notably, prevention-oriented messages from cattle producers were more commonly associated with precautionary behavioral intentions, such as checking gear and cattle for ticks. However, fewer participants reported using more intensive measures, such as tick repellents, following these messages. These trends suggest that while producers may rely on peer recommendations for certain preventive practices [46,47], additional reinforcement or information may be needed to encourage the use of more comprehensive preventive methods such as chemical treatments.

For the last research question, the associations among social engagement, perceptions, and preventive intentions were examined. First, it was found that participants who engaged with posts showed low to moderate correlations with their perceptions, including both their fear of tick-borne diseases and attitudes toward managing tick-related risks. This suggests that as cattle producers develop stronger perceptions of the issue, they are more inclined to interact with related posts, which in turn may amplify their resonance with the content. Furthermore, the study revealed that higher social engagement was also associated with increased implementation of both mechanical and chemical tick prevention measures. This implies the effectiveness of social media not only in engaging the public [32,34] but also in disseminating information about tick-borne diseases [17,19,35]. As this is a cross-sectional study, we cannot infer a direct causal relationship between social media interaction and cattle producers' perceptions and preventive intentions. The findings of this study present a potential avenue to be explored in greater depth in future research.

Additionally, the study revealed moderate associations between participants' fear of ticks harming cattle and their intentions to regularly inspect the cattle and gear or apply tick repellents. This is consistent with Olechnowicz et al [57], who highlight the motivating influence of emotional discomfort caused by ticks, as well as with other research showing that increased awareness and concern about tick-borne diseases can promote preventive practices [20-22]. These findings suggest that tick prevention messaging may benefit from incorporating emotional appeals so long as messages remain accurate and clear. On the basis of framing theory [19,51] and our results, we recommend pairing risk-based messages with practical, actionable strategies that empower producers, rather than relying solely on fear. When combined with evidence-based control recommendations [23,24], such an approach is more likely to motivate cattle producers to adopt systematic and sustained tick prevention practices.

This exploratory study has certain limitations. The sample—primarily drawn from the southeastern United States—was modest in size, and baseline participant characteristics were not fully assessed, which may limit the generalizability of the findings. However, the combination of

XSL•FO
RenderX

quantitative and qualitative results yields practical recommendations for those seeking to improve digital health risk communication in agricultural sectors. We recommend using larger, more diverse samples and assessing more baseline characteristics in future research, to validate these findings across different regions and agricultural contexts and to explore the potential differences in risk perception and preventive behaviors across diverse agricultural populations.

In the future, communicators should consider the challenge of providing sufficient and practical information about tick-borne diseases within the constraints of brief social media posts. While concise, visually engaging messages are necessary, they must also convey credibility and substance in order to build trust and motivate action. There is no single framing or strategy that works for all target audiences, which makes it an ongoing challenge to effectively communicate the risks of ticks and tick-borne diseases within a limited number of characters in a post while also encouraging the audience to seek more information or take preventive actions [28]. These challenges are important areas for further exploration in future research.

Funding

This work was supported by Cooperative Agreement Number U01CK000662, funded by the Centers for Disease Control and Prevention.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: CXY, LMB
Data curation: CXY, LMB
Formal analysis: CXY
Funding acquisition: LMB
Methodology: CXY
Project administration: CXY
Resources: LMB
Supervision: LMB
Visualization: CXY, JC
Writing – original draft: CXY
Writing – review & editing: LMB, JC

Conflicts of Interest

None declared.

References

1. Where ticks live. Centers for Disease Control and Prevention. 2025. URL: https://www.cdc.gov/ticks/about/where-ticks-live.html [accessed 2025-08-06]
2. Egizi A, Bulaga-Seraphin L, Alt E, et al. First glimpse into the origin and spread of the Asian longhorned tick, Haemaphysalis longicornis, in the United States. Zoonoses Public Health 2020 Sep;67(6):637-650. [doi: 10.1111/zph.12743] [Medline: 32638553]
3. Rainey T, Occi JL, Robbins RG, Egizi A. Discovery of Haemaphysalis longicornis (Ixodida: Ixodidae) parasitizing a sheep in New Jersey, United States. J Med Entomol 2018 May 4;55(3):757-759. [doi: 10.1093/jme/tjy006] [Medline: 29471482]
4. The longhorned tick: what you need to know story map. Animal and Plant Health Inspection Service - USDA. 2024. URL: https://www.aphis.usda.gov/livestock-poultry-disease/cattle/ticks/asian-longhorned/asian-longhorned-tick-what-you-need-know [accessed 2024-09-07]
5. Monitoring Haemaphysalis longicornis, the Asian longhorned tick, populations in the United States. : United States Department of Agriculture; 2024 URL: https://www.aphis.usda.gov/sites/default/files/monitoring-h-longicornis-plan-final-january-2024.pdf [accessed 2024-09-07]
6. Beard CB, Occi J, Bonilla DL, et al. Multistate infestation with the exotic disease-vector tick Haemaphysalis longicornis - United States, August 2017-September 2018. MMWR Morb Mortal Wkly Rep 2018 Nov 30;67(47):1310-1313. [doi: 10.15585/mmwr.mm6747a3] [Medline: 30496158]
7. Heath ACG. Biology, ecology and distribution of the tick, Haemaphysalis longicornis Neumann (Acari: Ixodidae) in New Zealand. N Z Vet J 2016 Jan;64(1):10-20. [doi: 10.1080/00480169.2015.1035769] [Medline: 25849758]
8. Schappach BL, Krell RK, Hornbostel VL, Connally NP. Exotic Haemaphysalis longicornis (Acari: Ixodidae) in the United States: biology, ecology, and strategies for management. J Integr Pest Manag 2020 Oct 30;11(1):21. [doi: 10.1093/jipm/pmaa019]

XSL•FO
RenderX

9. Tufts DM, Goodman LB, Benedict MC, Davis AD, VanAcker MC, Diuk-Wasser M. Association of the invasive Haemaphysalis longicornis tick with vertebrate hosts, other native tick vectors, and tick-borne pathogens in New York City, USA. Int J Parasitol 2021 Feb;51(2-3):149-157. [doi: 10.1016/j.ijpara.2020.08.008] [Medline: 33130214]
10. Butler RA, Trout Fryxell RT. Management of Haemaphysalis longicornis (Acari: Ixodidae) on a cow-calf farm in East Tennessee, USA. J Med Entomol 2023 Nov 14;60(6):1374-1379. [doi: 10.1093/jme/tjad121] [Medline: 37738311]
11. Dinkel KD, Herndon DR, Noh SM, et al. A U.S. isolate of Theileria orientalis, Ikeda genotype, is transmitted to cattle by the invasive Asian longhorned tick, Haemaphysalis longicornis. Parasit Vectors 2021 Mar 16;14(1):157. [doi: 10.1186/s13071-021-04659-9] [Medline: 33726815]
12. Eleftheriou A, Beckett J, Bai N, Pesapane R. An established population of Asian longhorned ticks (Acari: Ixodidae) in Ohio, USA. J Med Entomol 2023 Sep 12;60(5):1126-1130. [doi: 10.1093/jme/tjad104] [Medline: 37551419]
13. Thompson AT, White S, Shaw D, et al. Theileria orientalis Ikeda in host-seeking Haemaphysalis longicornis in Virginia, U.S.A. Ticks Tick Borne Dis 2020 Sep;11(5):101450. [doi: 10.1016/j.ttbdis.2020.101450] [Medline: 32723633]
14. Beard CB, Visser SN, Petersen LR. The need for a national strategy to address vector-borne disease threats in the United States. J Med Entomol 2019 Sep 3;56(5):1199-1203. [doi: 10.1093/jme/tjz074] [Medline: 31505668]
15. de la Fuente J, Estrada-Peña A, Rafael M, et al. Perception of ticks and tick-borne diseases worldwide. Pathogens 2023 Oct 19;12(10):1258. [doi: 10.3390/pathogens12101258] [Medline: 37887774]
16. Eisen L. Personal protection measures to prevent tick bites in the United States: knowledge gaps, challenges, and opportunities. Ticks Tick Borne Dis 2022 Jul;13(4):101944. [doi: 10.1016/j.ttbdis.2022.101944] [Medline: 35364518]
17. Gupta S, Eggers P, Arana A, et al. Knowledge and preventive behaviors towards tick-borne diseases in Delaware. Ticks Tick Borne Dis 2018 Mar;9(3):615-622. [doi: 10.1016/j.ttbdis.2018.01.006] [Medline: 29428491]
18. Herdiana H, Sari JFK, Whittaker M. Intersectoral collaboration for the prevention and control of vector borne diseases to support the implementation of a global strategy: a systematic review. PLoS One 2018;13(10):e0204659. [doi: 10.1371/journal.pone.0204659] [Medline: 30303996]
19. Yang CX, Baker LM, McLeod-Morin A. Tweet tweet tick: a quantitative content analysis of risk communication about ticks on Twitter. Front Commun 2024;9:1339356. [doi: 10.3389/fcomm.2024.1339356]
20. Cuadera MKQ, Mader EM, Safi AG, Harrington LC. Knowledge, attitudes, and practices for tick bite prevention and tick control among residents of Long Island, New York, USA. Ticks Tick Borne Dis 2023 May;14(3):102124. [doi: 10.1016/j.ttbdis.2023.102124] [Medline: 36764054]
21. Hansen MF, Sørensen PK, Sørensen AE, Krogfelt KA. Can protection motivation theory predict protective behavior against ticks? BMC Public Health 2023 Jun 22;23(1):1214. [doi: 10.1186/s12889-023-16125-5] [Medline: 37349761]
22. Vasić A, Bjekić J, Veinović G, et al. Knowledge, attitudes, and practices on tick-borne encephalitis virus and tick-borne diseases within professionally tick-exposed persons, health care workers, and general population in Serbia: a questionnaire-based study. Int J Environ Res Public Health 2022 Jan 13;19(2):867. [doi: 10.3390/ijerph19020867] [Medline: 35055686]
23. Zöldi V, Turunen T, Lyytikäinen O, Sane J. Knowledge, attitudes, and practices regarding ticks and tick-borne diseases, Finland. Ticks Tick Borne Dis 2017 Oct;8(6):872-877. [doi: 10.1016/j.ttbdis.2017.07.004] [Medline: 28778675]
24. Namgyal J, Tenzin T, Checkley S, et al. A knowledge, attitudes, and practices study on ticks and tick-borne diseases in cattle among farmers in a selected area of eastern Bhutan. PLoS One 2021;16(2):e0247302. [doi: 10.1371/journal.pone.0247302] [Medline: 33617588]
25. Cairns G, de Andrade M, MacDonald L. Reputation, relationships, risk communication, and the role of trust in the prevention and control of communicable disease: a review. J Health Commun 2013;18(12):1550-1565. [doi: 10.1080/10810730.2013.840696] [Medline: 24298887]
26. Ophir Y, Jamieson KH. The effects of media narratives about failures and discoveries in science on beliefs about and support for science. Public Underst Sci 2021 Nov;30(8):1008-1023. [doi: 10.1177/09636625211012630] [Medline: 34000907]
27. Sellnow TL, Seeger MW. Theorizing Crisis Communication, 2nd edition: Wiley-Blackwell; 2020.
28. Brady RM, Lemieux CJ, Doherty ST. Ticks and lyme disease in natural areas: a segmentation analysis of visitor perceptions of risk and preferred communication strategies. J Outdoor Recreat Tour 2024 Sep;47:100794. [doi: 10.1016/j.jort.2024.100794]
29. Koinig I, Kohler S. On the relationship between skepticism towards and reactance to health messages: the special case of online communication on tick-borne encephalitis. Front Commun 2021;6:737800. [doi: 10.3389/fcomm.2021.737800]
30. Yang CX, Baker LM, McLeod-Morin A. Trending ticks: using Google Trends data to understand tickborne disease prevention. Front Public Health 2024;12:1410713. [doi: 10.3389/fpubh.2024.1410713] [Medline: 38939559]
31. Bhattacharjee S, Raj S. Social media: shaping the future of agricultural extension and advisory services. : GFRAS; 2016 URL: https://www.researchgate.net/publication/330397283_Social_media_Shaping_the_future_of_agricultural_extension_and_advisory_services [accessed 2025-11-24]
32. Avidar R, Ariel Y, Malka V, Levy EC. Smartphones, publics, and OPR: do publics want to engage? Public Relat Rev 2015 Jun;41(2):214-221. [doi: 10.1016/j.pubrev.2014.11.019]
33. Malecki KMC, Keating JA, Safdar N. Crisis communication and public perception of COVID-19 risk in the era of social media. Clin Infect Dis 2021 Feb 16;72(4):697-702. [doi: 10.1093/cid/ciaa758] [Medline: 32544242]

XSL•FO
RenderX

34. Taylor M, Kent ML. Dialogic engagement: clarifying foundational concepts. J Public Relat Res 2014 Oct 20;26(5):384-398. [doi: 10.1080/1062726X.2014.956106]
35. Yang CX, Baker LM, McLeod-Morin A. Ticks on X: uncovering sources and engagement rate. Presented at: Proceedings of the International Crisis and Risk Communication Conference; Mar 11-13, 2024; Orlando, FL p. 108-111. [doi: 10.69931/JYCW3062]
36. Hwong YL, Oliver C, Van Kranendonk M, Sammut C, Seroussi Y. What makes you tick? The psychology of social media engagement in space science communication. Comput Human Behav 2017 Mar;68:480-492. [doi: 10.1016/j.chb.2016.11.068]
37. Pollett S, Rivers C. Social media and the new world of scientific communication during the COVID-19 pandemic. Clin Infect Dis 2020 Nov 19;71(16):2184-2186. [doi: 10.1093/cid/ciaa553] [Medline: 32396623]
38. Chan MPS, Winneg K, Hawkins L, Farhadloo M, Jamieson KH, Albarracín D. Legacy and social media respectively influence risk perceptions and protective behaviors during emerging health threats: a multi-wave analysis of communications on Zika virus cases. Soc Sci Med 2018 Sep;212:50-59. [doi: 10.1016/j.socscimed.2018.07.007] [Medline: 30005224]
39. Gamboa J, Lamb MM, de la Cruz P, Bull S, Olson D. Using social media to increase preventative behaviors against arboviral diseases: a pilot study among teens in the Dominican Republic. Mhealth 2019;5:30. [doi: 10.21037/mhealth.2019.07.03] [Medline: 31559275]
40. Zhang J, Wu HC, Chen L, Su Y. Effect of social media use on food safety risk perception through risk characteristics: exploring a moderated mediation model among people with different levels of science literacy. Front Psychol 2022;13:963863. [doi: 10.3389/fpsyg.2022.963863] [Medline: 36248473]
41. Tweeten JF, Paulsen TH. Perceptions of communication tools as defined by Iowa cattle producers. NACTA J 2020;65:176-184 [FREE Full text]
42. Gillespie JL. U.S. beef producers' current use and perceptions of social media as a communications tool [Master's thesis]. : Oklahoma State University; 2011 URL: https://openresearch.okstate.edu/server/api/core/bitstreams/32fc7eca-04b6-4daf-87eb-64a1c834a0e1/content [accessed 2025-11-21]
43. Begho T, Glenk K, Anik AR, Eory V. A systematic review of factors that influence farmers' adoption of sustainable crop farming practices: lessons for sustainable nitrogen management in South Asia. J Sustain Agric Environ 2022 Jun;1(2):149-160. [doi: 10.1002/sae2.12016]
44. Ofuoku AU. Influence of extension agents' and farmers' communications factors on the effectiveness poultry technology messages. Trop Agric Res Ext 2012;15(1):14-23. [doi: 10.4038/tare.v15i1.5238]
45. Seli L, Akbar M, Arianto A. Supporting and inhibiting factors of agricultural extension interpersonal communication competence in Enrekang Regency. KnE eng 2024:586-595. [doi: 10.18502/keg.v6i1.15440]
46. BenYishay A, Mobarak AM. Social learning and incentives for experimentation and communication. Rev Econ Stud 2019 May 1;86(3):976-1009. [doi: 10.1093/restud/rdy039]
47. Lasdun V, Harou AP, Magomba C, Guereña D. Peer learning and technology adoption in a digital farmer-to-farmer network. J Dev Econ 2025 Sep;176:103496. [doi: 10.1016/j.jdeveco.2025.103496]
48. Entman RM. Framing: toward clarification of a fractured paradigm. J Commun 1993 Dec 1;43(4):51-58. [doi: 10.1111/j.1460-2466.1993.tb01304.x]
49. Hameleers M. Prospect theory in times of a pandemic: the effects of gain versus loss framing on risky choices and emotional responses during the 2020 coronavirus outbreak – evidence from the US and the Netherlands. Mass Commun Soc 2021 Jul 4;24(4):479-499. [doi: 10.1080/15205436.2020.1870144]
50. Sleigh J, Amann J, Schneider M, Vayena E. Qualitative analysis of visual risk communication on Twitter during the COVID-19 pandemic. BMC Public Health 2021 Apr 28;21(1):810. [doi: 10.1186/s12889-021-10851-4] [Medline: 33906626]
51. Tversky A, Kahneman D. Advances in prospect theory: cumulative representation of uncertainty. J Risk Uncertain 1992 Oct;5(4):297-323. [doi: 10.1007/BF00122574]
52. O'Keefe DJ, Jensen JD. The relative persuasiveness of gain-framed and loss-framed messages for encouraging disease prevention behaviors: a meta-analytic review. J Health Commun 2007;12(7):623-644. [doi: 10.1080/10810730701615198] [Medline: 17934940]
53. O'Keefe DJ, Jensen JD. Do loss-framed persuasive messages engender greater message processing than do gain-framed messages? A meta-analytic review. Commun Stud 2008 Feb 21;59(1):51-67. [doi: 10.1080/10510970701849388]
54. Gantiva C, Jiménez-Leal W, Urriago-Rayo J. Framing messages to deal with the COVID-19 crisis: the role of loss/gain frames and content. Front Psychol 2021;12:568212. [doi: 10.3389/fpsyg.2021.568212] [Medline: 33584464]
55. Cho H, Boster FJ. Effects of gain versus loss frame antidrug ads on adolescents. J Commun 2008 Sep;58(3):428-446. [doi: 10.1111/j.1460-2466.2008.00393.x]
56. Li S, Coduto KD, Song C. Comments vs. one-click reactions: seeking and perceiving social support on social network sites. J Broadcast Electron Media 2020 Dec 1;64(5):777-793. [doi: 10.1080/08838151.2020.1848181]
57. Olechnowicz C, Leahy J, Gardner A, Sponarski CC. Perceived vulnerability for Lyme disease questionnaire: a social science tool for understanding tick-borne disease attitudes. Ticks Tick Borne Dis 2023 Mar;14(2):102120. [doi: 10.1016/j.ttbdis.2023.102120] [Medline: 36696753]
58. Baker R, Brick JM, Bates NA, et al. Summary report of the AAPOR task force on non-probability sampling. J Surv Stat Methodol 2013 Nov 1;1(2):90-143. [doi: 10.1093/jssam/smt008]

XSL•FO
RenderX

59. Eysenbach G. Improving the quality of web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34. [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
60. Fake tweet/X generator, Instagram posts generator and many more. Zeoob. URL: https://zeoob.com [accessed 2024-09-07]
61. 2022 Census of Agriculture. National Agricultural Statistics Service United States Department of Agriculture. 2024. URL: https://www.nass.usda.gov/Publications/AgCensus/2022 [accessed 2025-03-24]
62. Davis JA. Elementary Survey Analysis: Prentice-Hall; 1971.
63. Sumner EM, Ruge-Jones L, Alcorn D. A functional approach to the Facebook Like button: an exploration of meaning, interpersonal functionality, and potential alternative response buttons. New Media & Society 2017 Mar 20;20(4):1451-1469. [doi: 10.1177/1461444817697917]
64. Alhabash S, Almutairi N, Lou C, Kim W. Pathways to virality: psychophysiological responses preceding likes,shares, comments, and status updates on Facebook. Media Psychol 2019 Mar 4;22(2):196-216. [doi: 10.1080/15213269.2017.1416296]
65. Overbey KN, Jaykus LA, Chapman BJ. A systematic review of the use of social media for food safety risk communication. J Food Prot 2017 Sep;80(9):1537-1549. [doi: 10.4315/0362-028X.JFP-16-345] [Medline: 28805456]
66. Zeballos Rivas DR, Lopez Jaldin ML, Nina Canaviri B, Portugal Escalante LF, Alanes Fernández AMC, Aguilar Ticona JP. Social media exposure, risk perception, preventive behaviors and attitudes during the COVID-19 epidemic in La Paz, Bolivia: a cross sectional study. PLoS One 2021;16(1):e0245859. [doi: 10.1371/journal.pone.0245859] [Medline: 33481945]

Abbreviations

CHERRIES: Checklist for Reporting Results of Internet E-Surveys
NCBA: National Cattlemen's Beef Association
PETA: People for the Ethical Treatment of Animals

Edited by A Mavragani; submitted 09.05.25; peer-reviewed by EM Mader, HM Wang; revised version received 18.09.25; accepted 24.09.25; published 08.12.25.

Please cite as:
Yang CX, Baker LM, Czipulis J
Communication Strategies for Tick-Borne Disease Prevention Among US Cattle Producers: Quasi-Experimental Study on Message Framing and Source Credibility
JMIR Form Res 2025;9:e77239
URL: https://formative.jmir.org/2025/1/e77239
doi:10.2196/77239

© Cheng-Xian Yang, Lauri M Baker, Jessica Czipulis. Originally published in JMIR Formative Research (https://formative.jmir.org), 8.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Firearm Violence and Health in Policymaker Discourse: Mixed Methods Social Media Analysis

Vivek A Ashok[1,2,3,4], MSHP, MD; William J K Vervilles[5], MPH, MD; Katherine S Kellom[2,6], BA; Anyun Chatterjee[2,6], MPH; Isabella Ntigbu[7,8], MD; Okechi Boms[5], MD; Andrea Szabo[5], MD; Joel A Fein[3,7,9], MPH, MD; Therese S Richmond[10,11,12], RN, PhD; Jonathan Purtle[13], MPH, MSc, DrPH; Matthew D Kearney[10,14], MPH, PhD; Zachary F Meisel[10,12,15], MSHP, MPH, MD

[1]Division of General and Community Pediatrics, Cincinnati Children's Hospital Medical Center, 3333 Burnet Ave, Cincinnati, OH, United States
[10]Leonard Davis Institute of Health Economics, University of Pennsylvania, Philadelphia, PA, United States
[11]Department of Biobehavioral Health Sciences, School of Nursing, University of Pennsylvania, Philadelphia, PA, United States
[12]Penn Injury Science Center, University of Pennsylvania, Philadelphia, PA, United States
[13]Department of Public Health Policy and Management, School of Global Public Health, New York University, New York, NY, United States
[14]Department of Family Medicine and Community Health, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States
[15]Department of Emergency Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States
[2]PolicyLab, Children's Hospital of Philadelphia, Philadelphia, PA, United States
[3]The Center for Violence Prevention, Children's Hospital of Philadelphia, Philadelphia, PA, United States
[4]Department of Pediatrics, School of Medicine, University of Cincinnati, Cincinnati, OH, United States
[5]Combined Internal Medicine-Pediatrics Residency Program, Department of Medicine, University of Pennsylvania, Philadelphia, PA, United States
[6]Qualitative Research Core, Children's Hospital of Philadelphia, Philadelphia, PA, United States
[7]Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States
[8]Emergency Medicine Residency Program, School of Medicine, Brown University, Providence, RI, United States
[9]Division of Emergency Medicine, Department of Pediatrics, Children's Hospital of Philadelphia, Philadelphia, PA, United States

Corresponding Author:
Vivek A Ashok, MSHP, MD
Division of General and Community Pediatrics, Cincinnati Children's Hospital Medical Center, 3333 Burnet Ave, Cincinnati, OH, United States

Abstract

Background: Since 2019, firearm violence has remained the leading cause of death for US children and adolescents aged 1‐19 years. This crisis has spurred action from policymakers, health professionals, and advocates. However, political polarization has contributed to divergent views on the causes and appropriate responses to firearm violence. Communication by elected officials, especially on social media, plays a critical role in shaping public opinion and policy agendas. Understanding how state policymakers discuss firearm violence, including the use of causal blame, calls to action, and health-related narratives, can inform more effective public health strategies.

Objective: This study aimed to examine how Pennsylvania state legislators discuss firearms and firearm violence on social media and assess the extent to which their messaging aligns with public health perspectives.

Methods: We conducted a 2-phase mixed methods analysis of X (formerly known as Twitter; X Corp) posts by Pennsylvania state legislators from May 27, 2017, to July 26, 2022. Posts were grouped into 3 time periods surrounding the Tree of Life Synagogue mass shooting in Pittsburgh. Using a Boolean search strategy, we identified 4573 posts related to firearms and firearm violence. After removing reposts and non-English content, we randomly sampled 1491 (32.6%) original posts authored by 152 unique legislators. Posts were coded using a structured codebook based on the Multiple Streams Framework to capture rhetorical framing, causal blame, and policy content. Interrater reliability was high (Holsti coefficient >0.8). We used chi-square tests and multivariable logistic regression to assess associations between rhetorical elements and policy mentions, adjusting for time period.

Results: Mass shootings were the most frequently referenced category of firearm violence, peaking after the Tree of Life shooting (22/43, 51% vs 91/118, 77.1% vs 140/220, 63.6%; *P*=.004), while firearm suicide was rarely discussed. Posts using advocacy frames were nearly 5 times more likely to mention policy (adjusted odds ratio [aOR] 4.67, 95% CI 3.55‐6.16), whereas those referencing mass shootings (aOR 0.54, 95% CI 0.37‐0.77) or emotional appeals (aOR 0.53, 95% CI 0.40‐0.69) were significantly less likely to do so. Most posts used general advocacy (aOR 2.97, 95% CI 2.13-4.13) and vague blame (aOR 8.26, 95% CI 6.02‐11.35), resulting in nonspecific policy suggestions. Posts that attributed blame to firearm access were strongly associated

XSL•FO
RenderX

with specific policy proposals (aOR 6.37, 95% CI 4.29-9.47) and inversely associated with general policy mentions (aOR 0.26, 95% CI 0.17 - 0.42). Only 9.4% (133/1422) of posts used health frames; when present, they more often referenced physical consequences (58/133, 43.6% vs 216/1358, 15.9%; *P*<.001).

Conclusions: Pennsylvania legislators primarily focused on mass shootings and relied on emotional or symbolic language without proposing specific policies. Health frames were rare and typically focused on consequences rather than prevention. Findings highlight an opportunity to support policymakers with health-informed messaging strategies to promote actionable firearm violence prevention policies, particularly those addressing prevention.

(JMIR Form Res 2025;9:e80397) doi:10.2196/80397

KEYWORDS

firearms; violence; social media; communication; health policy; prevention

Introduction

Firearm violence and its devastating physical, psychological, and economic consequences in the United States have galvanized policymakers, health and public health professionals, researchers, law enforcement, and the judicial system to enact policies aimed at stopping this epidemic [1-5]. However, despite growing attention, deeply entrenched political polarization surrounding firearms has led to divergent views on the causes and consequences of firearm violence and the policies proposed to address it [6,7]. These divisions persist amid a steady rise in firearm violence, now the leading cause of death among children and adolescents aged 1 - 19 years [4]. Moreover, Non-Hispanic Black or African American adolescents and young adults experience significantly higher rates of interpersonal firearm violence, while American Indian and Alaskan Native youth experience disproportionately high rates of firearm suicide, reflecting broader societal inequities [3,8-11]. This burden is often overlooked in policy discourse and media coverage, especially on social media, where selective framing can skew public perception [12]. Nonetheless, social media remains a powerful force in influencing policymaking through effective health and public health communication [13].

Elected officials play a central role in shaping the policy environment and represent a critical upstream political determinant of health [13,14]. Increasingly, legislators and their staff have turned to social media as a tool for issue engagement and political action [15-19]. Understanding how policymakers perceive, define, and communicate about firearm violence is essential to informing effective and equitable policy solutions. Social media offers a unique window into policymaker perspectives and directly impacts health policy agenda setting that is increasingly used in policy research [13,20-26]. It allows unfiltered access to policymakers' unscripted viewpoints and positions without the framing and gatekeeping of traditional media [13]. Specifically, social media can reveal what causes are emphasized, what solutions are proposed, how certain incidents elicit reactions, and how specific populations are represented. Through this medium, policymakers can elevate public understanding of firearm violence; however, they can also perpetuate misinformation and misleading narratives that undermine evidence-based prevention strategies [27,28].

Despite the growing role of social media in shaping public discourse and policy agendas, no prior study has systematically examined how policymakers characterize firearm violence online. Most legislative action around firearm regulation and violence prevention occurs at the state level [29]. Understanding how the content and framing of state-level policymaker messages have evolved over time, how rhetorical strategies, such as causal blame and calls to action, correspond with policy discussions, and the degree to which firearm violence is conceptualized through and aligns with a health or public health lens is necessary to address this epidemic. Addressing these gaps is critical for many reasons. The ways in which firearm violence is framed—whether as a health or public health issue, a criminal justice concern, or a partisan talking point—influence the development and uptake of policies [30-38]. Furthermore, biased or inaccurate representations can advance policies that are misguided, ineffective, or inequitable by further entrenching health inequities and promoting stigma and bias of certain groups [27,28]. To better understand how recent sociopolitical discourse may align or misalign with public health and health care realities, we analyzed the language used by Pennsylvania state legislators in their social media posts from 2017 through 2022.

Methods

Setting, Data Collection, and Sampling

We conducted a 2-phased mixed methods analysis of X (formerly known as Twitter) posts related to firearms and firearm violence by Pennsylvania state legislators. This involved a systematic qualitative coding process using an iteratively developed codebook, followed by quantitative content analysis to identify trends in message framing, associations, and policy engagement. In order to capture the breadth of our discussion, firearm violence was broadly defined to include firearm suicide, any interpersonal firearm injury, and unintentional firearm injury.

We focused our analysis geographically on the Commonwealth of Pennsylvania for its heterogeneity in political party affiliation, urban and rural representation, firearm-related cultural norms, and diverse demographic makeup. We focused on state legislators because state-level policymaking has the most direct influence on firearm-related legislation, funding, and implementation, and because state legislative discourse reflects the primary locus of policy debate in this domain. We collected posts from May 27, 2017, to July 26, 2022, selecting a time frame that allowed us to examine the potential influence of major local and global events. We analyzed three distinct

XSL•FO
RenderX

15-month periods: (1) prior to the Tree of Life Synagogue shooting in Pittsburgh (from May 27, 2017 to October 26, 2018), (2) following that shooting through the onset of the COVID-19 pandemic (from October 27, 2018 to February 26, 2020), and (3) during the pandemic (from February 27, 2020 to July 26, 2022). Based on prior literature, using a local mass shooting as a fulcrum for time-varying analysis allows us to capture periods when firearm violence garners different levels of attention [39]. Furthermore, the COVID-19 pandemic was marked by an overall rise in firearm violence and firearm ownership and offered a distinct contextual period independent of the mass shooting event [40,41].

Our unit of analysis was individual X posts. Data were sourced from Quorum (Quorum Analytics), a public affairs software program that tracks policymaker social media engagements; data are available from 2007 to present [42]. As is standard practice, the research team paid a fee to access the Quorum database. Quorum adheres to public data usage guidelines set by X. The database provides metadata on the source of each post, the policymaker's political party affiliation, the policymaker's represented district, and the date posted.

Using Quorum's advanced search features, we limited our query to Pennsylvania state legislators and the study period of interest. Pennsylvania state legislators created a total of 401,477 posts. Drawing on existing literature and in consultation with academic librarians, we created a Boolean string to capture posts relevant to discourse around firearms and firearm violence. For example, we included any post including the terms "gun(s)," "firearm(s)," "shooting(s)," AND "violence," "safety," etc (full list of search terms is provided in Figure 1) [43]. This search yielded 4573 eligible X posts, from which we excluded reposts and non-English posts. From the remaining pool, we drew a final simple random sample of 1491 (32.6%) posts authored by 152 unique users for in-depth coding and quantitative analysis. The final sample size was determined post hoc, once thematic saturation was achieved, defined as the point at which no new themes or codes emerged across successive batches of coded posts. This number also provided adequate representation across the study period to permit descriptive comparisons and multivariable logistic regression analyses. Each post was manually reviewed during coding to ensure relevance to the study topic.

Figure 1. Identification of X posts by Pennsylvania state legislators describing firearms and firearm violence, 2017-2022.

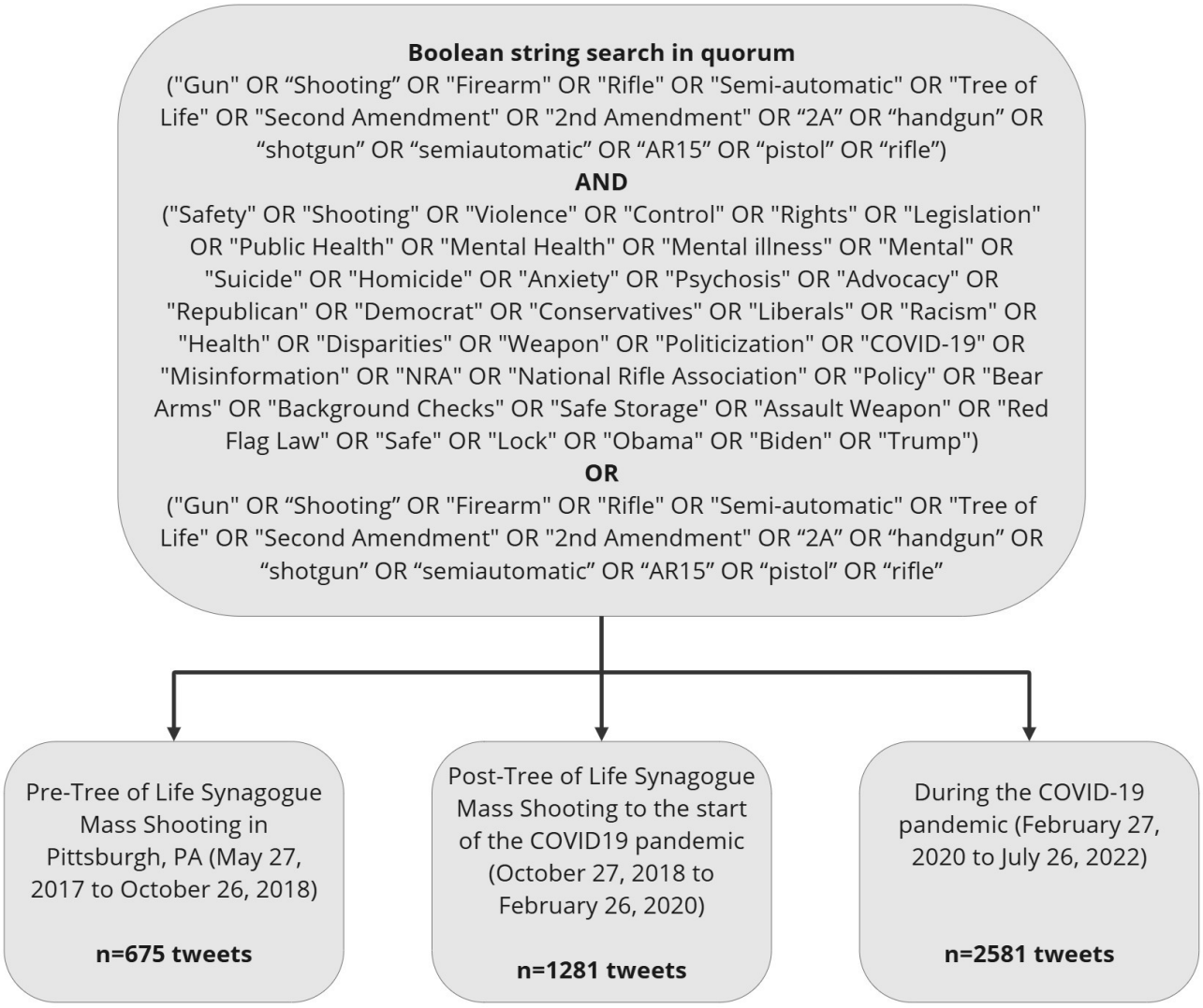

XSL•FO
RenderX

Qualitative Analysis

Codebook Development and Theoretical Framework

Five researchers (VAA, KSK, IN, MDK, and WJKV), with a range of expertise in firearm injury prevention research, health services research, social media, qualitative methods, health and medicine, and public health, collaboratively developed a 56-item codebook (Multimedia Appendix 1). Our coding framework was grounded in the Kingdon [44] Multiple Streams Framework (MSF) for Policy Agenda Setting (Figure 2 [45-47]). The MSF model has been used extensively in social and political sciences to evaluate policy agenda setting and offers an innovative way to conceptualize the pathway to health policy creation [47,48].

Figure 2. Kingdon's multiple streams framework adapted for health policy discourse on firearm violence among Pennsylvania state legislators (2017-2022).

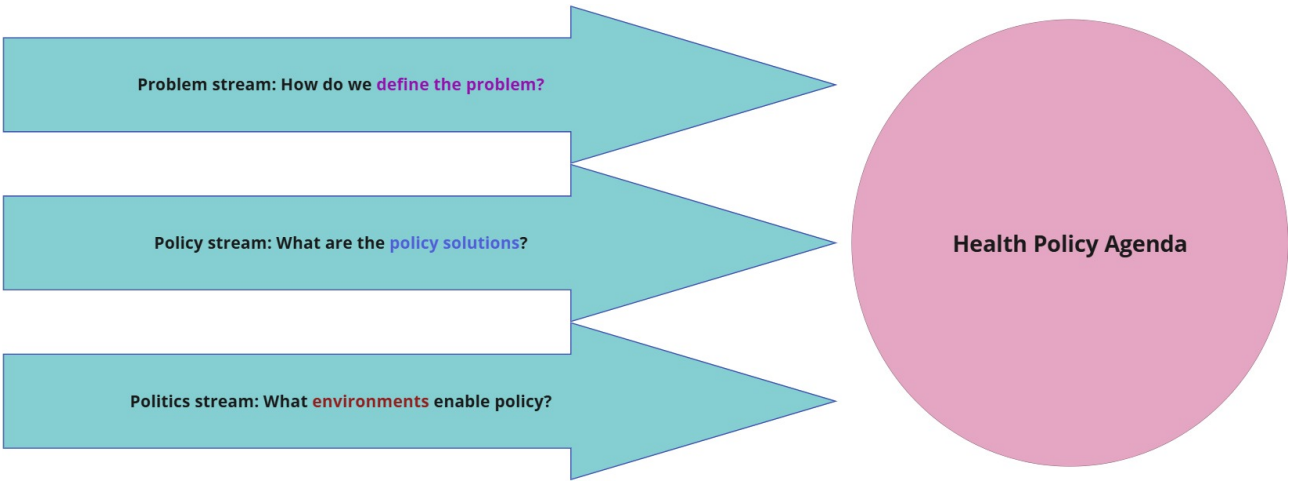

The MSF model describes 3 "streams" that influence policy agenda setting: the problem stream, policy stream, and politics stream. The problem stream refers to how a problem is defined and socially constructed. The policy stream relates to the development and advocacy of specific solutions. The politics stream refers to the sociopolitical forces that shape policy adoption. Our analysis focused on the problem and policy streams. We excluded the politics stream from our coding schema to maintain analytic focus on the content of the posts themselves and to minimize coder bias that could arise from subjective interpretations of external political context. However, recognizing that posts are not created in a vacuum and often reflect broader geopolitical events, we used the politics stream post hoc to contextualize our findings within relevant sociopolitical moments. Within the problem stream, we used the Policy Frames Codebook, created by Boydstun et al [49], to identify salient frames used to describe firearms and firearm violence (Table 1).

XSL•FO
RenderX

Table . Frames, definitions, and example X posts by Pennsylvania state legislators (2017-2022).

Frame	Definition	Example X post
Morality	Any perspective—or policy objective or action (including proposed action)—that is compelled by moral doctrine or interpretation, logic, duty, honor, honesty or integrity, righteousness or any other sense of ethics or social responsibility. A calculated way for someone to think about right and wrong rather than emotion. Often, the phrase "common sense" is used to connote logic.	"Thoughts and prayers do not protect our children. I will not participate in these actions when we can pass common sense gun laws TODAY."
Advocacy or Endorsement	A way of presenting information that promotes a particular viewpoint, policy, or action. It is designed to persuade the audience to support a specific cause, agenda, or position. This type of frame often includes subjective elements, such as endorsements, recognition of efforts, calls to action, and expressions of support or opposition. This frame can be general or specific advocacy. For example, it can include advocating for general gun violence prevention (viewpoint) or protecting the rights of firearm owners through legislation (policy). It can also include advocating or promoting one's own actions toward advancing a specific goal. A vote or sponsorship of a bill is an endorsement.	"We cannot wait until the next Columbine, or Sandy Hook, or Parkland, or Uvalde. We need to act now. Lives depend on us. Help send a message to the Majority leaders and urge them to advance gun-safety legislation"
Emotional Appeal or Moral Outrage	Uses words and phrases to evoke sympathy, sorrow, and a sense of tragedy. Uses phrases to express injustice and moral outrage, which can motivate audiences to act against an issue. Using facts or statistics to highlight the sheer scope of an issue and evoke a sense of moral imperative.	"Today is a day of mourning and reflection as we mark one year since the mass shooting at the Tree of Life building in Pittsburgh. Eleven were killed and six others injured in the deadliest attack on the Jewish community in the United States. Remember. Repair. Together."
Cultural Identity	Relates to social norms, trends, values, and customs. For example, it can include hunting and sporting events and shooting season in Pennsylvania. It may also refer to tradition and firearms as heritage items.	"Rifle season starts Saturday and it's important to remember to be safe while hunting or hiking in PA's outdoors. Have fun and good luck to all the hunters out there!"
Equality or Equity	Equality or inequality with which laws, punishment, rewards, and resources are applied or distributed among individuals or groups. Also, the balance between the rights or interests of one individual or group compared to another individual or group.	"The signs and voices chanted about black lives and gun violence, about justice reform, about the systemic underfunding of schools in black and brown neighborhoods. about healthcare, housing, transportation, minimum wage. So much undone. (Yet to be done.)"
Criminology and Policing	Framing gun violence within the context of crime and law enforcement. Specifically, if the problem or solutions are grounded in crime prevention or law enforcement interventions (eg, longer sentences for perpetrators or criminal penalties for gun-related offenses); portrayals of gun violence that are intertwined with the criminal justice system (eg, highlighting gang violence and armed robberies); or involvement of law enforcement as primary responders (as opposed to health care workers or community members).	"Within Philadelphia, effective and non-lethal policing is also essential to healing communities that have been historically overpoliced and over incarcerated. It will take all stakeholders to end the scourge of gun violence! #EnoughIsEnough"

XSL•FO
RenderX

Frame	Definition	Example X post
Constitutionality and Jurisprudence	The constraints imposed on or freedoms granted to individuals, government, and corporations via the Constitution, Bill of Rights, and other amendments, or judicial interpretation. This deals specifically with the authority of government to regulate and the authority of individuals or corporations to act independently of government. For example, civil libertarianism, citing encroachment on constitutional rights and individual liberties, most commonly the Second Amendment. Local laws or policies wrapped in legal or lawmaker language. Citations of legal cases are also included in this frame.	"I was proud to stand up for the Second Amendment rights of law-abiding Pennsylvanians today. The bills we approved will allow lawful gun owners to carry without a permit and maintain uniformity of state laws in every municipality in the state."
Security, Defense, or Protection	Safety, threats to security, and protection of one's person, family, in-group, nation, etc. Generally, it is an action or a call to action that can be taken to protect the welfare of a person, group, nation, sometimes from a not yet-manifested threat. In this frame, the coder asks if the post is using rhetoric related to security, defense, or protection to persuade you to take a stance or delineates specific steps to provide that security, defense, or protection. For example, this frame could include lauding the use of a firearm or gun for self-protection or protecting the community, or protecting the community from gun violence itself.	"I've introduced legislation to ensure that we are doing everything we can to protect the residents of the commonwealth. I will continue to be an advocate for legislative initiatives that respect the Second Amendment while promoting gun safety."
Public Opinion	References to general social attitudes, polling, and demographic information, as well as implied or actual consequences of diverging from or "getting ahead of" public opinion or polls. Often uses the phrase "most people support."	"People demand that we act & pass common sense gun violence prevention bills like ERPO."
Health and Public Health	Is the public health approach to curb gun violence, as described by the Centers for Disease Control and Prevention (CDC), mentioned (must meet one criterion): (1) Does the post cite statistics or the importance of research, generating data, risk factors, etc? (2) Does the post mention using evidence-based practices or science? (3) Does the post mention evaluating the implementation of the methods used to prevent firearm violence?	Example "Did you know? Studies suggest that waiting period law may reduce firearms suicide rates by 7% to 11%. That's around 2500 lives a year."
Partisan	Defines the problems or solutions of gun violence through a partisan lens by touting one political party over another or castigating a specific political party.	"This is absolutely ridiculous. We are talking about thousands of lives lost due to gun violence. We will not be silenced. We will keep fighting for what's right! Senate Republicans are now trying to silence [deidentified]. This is a typical move the GOP makes when they are clearly losing the argument."
Informational	The entire post should convey information, which could include statistics, public health recommendations, and information about an event, campaign, or vigil. While neutral on the surface, it may not be felt as neutral since authors tend to use jarring statistics to drive home a point. There should be no partisan language in a post that is information-framed.	"SB565 reduces the age a person can carry a firearm from 21 to 18 and does not require a license to openly carry firearms in cities like Philadelphia."
Religiosity or Spirituality	Relates to tying the right to bear a firearm or firearm violence to religiosity or spirituality. For example, "God given right" or "may God bless this state"	"Practically every gun tragedy is a preventable tragedy, if you don't believe the myth that our Creator intended us to have the God-like ability to destroy human life"

The codebook was developed iteratively through group discussion and consensus. Guided by the MSF, the research team deductively coded posts using these predefined frames and inductively refined the codebook through group discussion

XSL•FO
RenderX

and consensus. New or modified frames (eg, health and public health, criminology and policing, and advocacy or endorsement) were added as themes emerged during pilot coding. We coded a subset of 6.7% (100/1491) of the total analytic sample to refine and validate our codebook. We reviewed the data to ensure that key concepts within the problem and policy streams were sufficiently represented in the sample and well captured by our codebook. Throughout this process, discrepancies in interpretation were resolved with team-based consensus and subsequent revisions.

Coding Process

A final sample of 1491 posts was coded using a structured instrument developed in REDCap (Research Electronic Data Capture; Vanderbilt University). Seven coders (VAA, IN, WJKV, OB, AS, KSK, and AC) participated in this process, with VA serving as the reference coder. To assess interrater reliability, 10.6% (159/1491) posts of the final sample were coded by at least 2 team members. We calculated the Holsti coefficient for agreement, achieving average coefficients greater than 0.8 across domains. Consistent with qualitative coding conventions for larger teams, 1 coder (VA) served as the dominant coder, and all other coders were trained to achieve a Holsti coefficient >0.8 relative to VA. Each coding pair (VA and a secondary coder) independently coded a subset of 10 - 20 posts, reviewed discrepancies, and iterated until consensus and the reliability threshold were achieved. Double-coded subsets were retrospectively amended to reflect consensus codes. To prevent coder drift, all coders participated in one to two double-coding sessions for every 250 - 500 posts coded.

To minimize potential coder bias, all coders were blinded to the identity and political affiliation of the poster. In addition, embedded hyperlinks were not coded, as these sometimes revealed the identity or political affiliation of the author.

Measures and Statistical Analysis

Coded variables were applied at the post level, classified as binary or categorical, and entered into REDCap. Within the problem stream, we measured whether posts referenced general descriptors of violence, including the type of incident, location, and characteristics of violence survivor or perpetrators if a specific interpersonal firearm violence incident was discussed. These characteristics are conceptualized as part of the problem definition. Mentions of violence survivor or perpetrator identity were categorized by factors, if known or shared, such as age, race and ethnicity, gender identity or sexual orientation, mental health status, and involvement with the criminal justice system. Posts expressing symbolic responses, such as “thoughts and prayers,” solidarity, or remembrance, were also coded. We identified calls to action, including general appeals for change, specific policy recommendations, or encouragement to vote. Thirteen rhetorical frames were coded, encompassing moral arguments, advocacy language, emotional appeals, criminology and policing narratives, public health framing, cultural identity, partisanship, and equity. To capture all post elements, codes were nonmutually exclusive and could be used in combination with each other.

Within the policy stream, we coded whether posts assigned causal blame (ie, attribution of responsibility for a given action to specific individuals, groups, or systems) to any of 17 distinct sources, such as mental health, access to firearms, weak laws, racism, political parties, the firearm lobby, substance use, or social media. We defined causal blame as statements assigning responsibility for firearm violence to specific actors or systems (eg, weak laws and the firearm lobby), whereas rhetorical frames referred to the broader narrative lens through which the issue was discussed (eg, advocacy, emotion, and partisanship). These are conceptually and analytically distinct. We also coded references to the consequences of firearm violence, including mental and physical health impacts, economic costs, the cycle of violence, secondary trauma, and harm to community well-being. Finally, we assessed whether posts endorsed or proposed specific policies, such as (1) restrictive or permissive firearm ownership laws, (2) support or funding for community violence intervention programs or law enforcement, (3) initiatives targeting the firearm industry, and (4) policies addressing mental health or substance use.

We merged our coded dataset with the original Quorum dataset using unique identifiers to retain metadata on post date and author. We calculated the proportion of X posts that referenced each code and used chi-square tests to assess whether these proportions varied significantly across the three 15-month time frames. We applied Bonferroni correction for multiple comparisons with a significance level of P=0.02 when appropriate. All statistical analysis was performed using STATA 18.0 (StataCorp LLC) software.

We conducted multivariable logistic regression, adjusting for time period, to examine the associations between policy mentions and rhetorical elements, such as calls to action, types of shootings, causal blame attributions, and thematic frames. While rhetorical frames, causal blame attributions, and types of firearm violence may co-occur in individual posts, they were coded as distinct constructs and modeled as independent predictors in our regression analyses. Our primary model examined predictors of any policy-related content being mentioned, using variables selected based on descriptive prevalence and refined through significance testing and model fit. Then, we ran a series of additional models to assess whether specific categories of violence, causal blame attributions, or frames were individually associated with policy-related content.

To further explore the relationship between rhetorical framing and policy specificity, we stratified our policy outcome variable to distinguish between general (nonspecific) and specific policy proposals and re-estimated our model to examine predictors of each category. Predictors included categories of firearm violence mentioned, types of call to action, framing devices, and causal blame attributions. Finally, we conducted a subgroup analysis comparing X posts that did and did not use health frames, using chi-square tests to identify which topics were more likely to be mentioned when firearm violence was conceptualized as a health issue. Adjusted odds ratios (aORs) with 95% CIs were reported, using robust SEs.

XSL•FO
RenderX

Ethical Considerations

The dataset contained only publicly available information on X from the offices of elected officials, and therefore did not meet the criteria for human subjects research. All posts, data, and user IDs were deidentified for this study.

Results

Descriptive Results

We analyzed 1491 unique X posts by Pennsylvania legislators related to firearms and firearm violence. Of these, 210 (14%) were posted 15 months prior to the Tree of Life Synagogue shooting, 415 (27.8%) in the 15 months following the shooting through the onset of the COVID-19 pandemic, and 866 (58%) during the first 15 months of the pandemic. We present results using the MSF, starting with the problem stream (Table 2) and the policy stream (Table 3).

XSL•FO
RenderX

Table . The problem stream–unadjusted comparisons of specific topic mentions on X (formerly known as Twitter) by Pennsylvania state legislators about firearms and firearm violence during 3 study time periods (2017-2022).

Topic	Time period 1 (n=210), n (%)	Time period 2 (n=415), n (%)	Time period 3 (n=866), n (%)	Totals (n=1491), n (%)	*P* values
Descriptions of firearm violence					
Mentions of specific type of firearm violence	43 (20.5)	118 (28.4)	220 (25.4)	381 (25.6)	.10
Mass shooting[a,b]	22 (51.2)	91 (77.1)	140 (63.6)	253 (66.4)	.004
Suicide or intentional self-inflicted harm	3 (7)	11 (9.3)	18 (8.2)	32 (8.4)	.88
Unintentional or accidental injury	1 (2.3)	2 (1.7)	8 (3.6)	11 (2.9)	.58
Interpersonal or community violence[b]	11 (25.6)	13 (11)	55 (25)	79 (20.7)	.007
Other type of shooting	6 (14)	4 (3.4)	22 (10)	32 (8.4)	.04
Mentions of specific locations of shooting	48 (22.9)	107 (25.8)	204 (23.6)	358 (24.1)	.62
School or school grounds	11 (22.9)	12 (11.2)	39 (19.1)	62 (17.3)	.12
Home	1 (2.1)	0 (0)	3 (1.5)	4 (1.1)	.40
Place of worship[a,b,c]	1 (2.1)	52 (48.6)	38 (18.6)	91 (25.4)	<.001
Neighborhood or community	3 (6.2)	5 (4.7)	14 (6.9)	22 (6.1)	.75
Place of business	3 (6.2)	5 (4.7)	17 (8.3)	25 (7)	.47
City or county[b]	28 (58.3)	47 (43.9)	120 (58.8)	195 (54.5)	.04
Other location	2 (4.2)	0 (0)	4 (2)	6 (1.7)	.15
Mentions of a vigil, remembrance event, or campaign (such as #WearOrange)	14 (7.6)	47 (12.6)	77 (10.1)	138 (9.3)	.17
Mentions of relationship of shooter to violence survivor	5 (2.4)	6 (1.4)	34 (3.9)	45 (3)	.04
Intimate partner[b,c]	4 (80)	3 (50)	1 (2.9)	8 (17.8)	<.001
Parent or guardian	0 (0)	0 (0)	0 (0)	0 (0)	__[d]
Family member	0 (0)	0 (0)	1 (2.9)	1 (2.2)	.85
Law enforcement	1 (20)	1 (16.7)	19 (55.9)	21 (46.7)	.09
Other authority figure	0 (0)	0 (0)	1 (2.9)	1 (2.2)	.85
Unknown to violence survivor	0 (0)	1 (16.7)	5 (14.7)	6 (13.3)	.64
Other relationship	0 (0)	1 (16.7)	9 (26.5)	10 (22.2)	.40
Expressions of symbolic responses	34 (16.2)	91 (21.9)	169 (19.5)	294 (19.7)	.23
“Thoughts and Prayers”[a]	17 (50)	22 (24.2)	51 (30)	90 (30.6)	.02

XSL•FO
RenderX

Topic	Time period 1 (n=210), n (%)	Time period 2 (n=415), n (%)	Time period 3 (n=866), n (%)	Totals (n=1491), n (%)	*P* values
Solidarity[b]	11 (32.4)	26 (28.6)	76 (45)	113 (38.4)	.03
Remembrance[c]	7 (20.6)	58 (63.7)	68 (40.2)	133 (45.2)	<.001
Mentions of calls to action	61 (29)	103 (24.8)	293 (33.8)	457 (30.7)	.004
General call to action	46 (75.4)	80 (77.7)	205 (70)	331 (72.4)	.28
Specific call to action	31 (51.7)	43 (41.7)	130 (44.5)	204 (44.8)	.46
Calls to vote[c]	9 (29)	5 (12.2)	14 (10.9)	28 (13.9)	.03
Issue framing	199 (94.8)	390 (94)	833 (96.2)	1422 (95.4)	.35
Morality	10 (5)	26 (6.7)	50 (6)	86 (6)	.73
Advocacy or endorsement	135 (67.8)	283 (72.6)	633 (76)	1051 (73.9)	.049
Emotional appeal or moral outrage[c]	76 (38.2)	171 (43.8)	412 (49.5)	659 (46.3)	.008
Cultural identity	3 (1.5)	13 (3.3)	27 (3.2)	43 (3)	.40
Equality or equity	4 (2)	12 (3.1)	26 (3.1)	42 (3)	.70
Criminology and policing	8 (4)	16 (4.1)	46 (5.5)	70 (4.9)	.46
Constitutionality and jurisprudence	14 (7)	32 (8.2)	75 (9)	121 (8.5)	.65
Security, defense, or protection	24 (12.1)	41 (10.5)	81 (9.7)	146 (10.3)	.61
Public opinion	7 (3.5)	13 (3.3)	38 (4.6)	58 (4.1)	.55
Health and public health[b]	20 (10.1)	50 (12.8)	63 (7.6)	133 (9.4)	.01
Partisan[b,c]	17 (8.5)	28 (7.2)	181 (21.7)	226 (15.9)	<.001
Informational[c]	35 (17.6)	50 (12.8)	84 (10.1)	169 (11.9)	.01
Religiosity or spirituality	4 (2)	10 (2.6)	14 (1.7)	28 (2)	.55

[a]Post hoc pairwise comparisons significant for Time 1 versus Time 2.

[b]Post hoc pairwise comparisons significant for Time 2 versus Time 3.

[c]Post hoc pairwise comparisons significant for Time 1 versus Time 3.

[d]Not available.

XSL•FO
RenderX

Table . The policy stream-unadjusted comparisons of specific topic mentions on X (formerly known as Twitter) by Pennsylvania state legislators about firearms and firearm violence during 3 study time periods (2017-2022).

Topic	Time period 1 (n=210), n (%)	Time period 2 (n=415), n (%)	Time period 3 (n=866), n (%)	Totals (n=1491), n (%)	*P* values
Causal blame or causal attributions	100 (47.6)	180 (43.4)	487 (56.2)	767 (51.4)	<.001
Mental health or illness	2 (2)	7 (3.9)	22 (4.5)	31 (4)	.50
Weak firearm laws	40 (40)	79 (43.9)	199 (40.9)	318 (41.5)	.74
Increased ownership, access, or availability of firearms	16 (16)	27 (15)	106 (21.8)	149 (19.4)	.10
Decreased possession of firearms	0 (0)	0 (0)	5 (1)	5 (0.7)	.24
Sociopolitical and structural determinants of health	0 (0)	2 (1.1)	12 (2.5)	14 (1.8)	.20
Stereotyped, Biased, or Racist beliefs of a group of persons[a,b]	0 (0)	10 (5.6)	34 (7)	44 (5.7)	.02
Guns in the hands of wrong people	9 (9)	13 (7.2)	19 (3.9)	41 (5.3)	.05
Gang Violence or Organized crime	0 (0)	0 (0)	1 (0.2)	1 (0.1)	.75
Social media	0 (0)	0 (0)	1 (0.2)	1 (0.1)	.75
Individuals or groups of policymakers (eg, Philadelphia Attorney General)	15 (15)	20 (11.1)	83 (17)	118 (15.4)	.17
Causal blame attributed to the violence survivor	0 (0)	0 (0)	1 (0.2)	1 (0.1)	.75
Substance use	0 (0)	0 (0)	0 (0)	0 (0)	__[c]
Firearm industry and commercial practices	0 (0)	2 (1.1)	8 (1.6)	10 (1.3)	.41
US firearm culture	2 (2)	2 (1.1)	6 (1.2)	10 (1.3)	.80
The firearm lobby[b]	13 (13)	9 (5)	14 (2.9)	36 (4.7)	<.001
A specific political party or affiliation[d,b]	8 (8)	7 (3.9)	131 (26.9)	146 (19)	<.001
Hate or hate crimes[a]	0 (0)	13 (7.2)	26 (5.3)	39 (5.1)	.03
Other causal blame category	27 (27)	40 (22.2)	86 (17.7)	153 (19.9)	.07
Consequences	44 (21)	129 (31.1)	298 (34.4)	471 (31.6)	<.001
Economic or Financial	0 (0)	2 (1.6)	5 (1.7)	7 (1.5)	.69
Physical harm or disability	25 (56.8)	78 (60.5)	171 (57.4)	274 (58.2)	.82
Mental health or illness	1 (2.3)	4 (3.1)	12 (4)	17 (3.6)	.79
Furthering the cycle of violence	10 (22.7)	23 (17.8)	73 (24.5)	106 (22.5)	.32
Secondary trauma to cosurvivors of violence	12 (27.3)	32 (24.8)	79 (26.5)	123 (26.1)	.92

XSL•FO
RenderX

Topic	Time period 1 (n=210), n (%)	Time period 2 (n=415), n (%)	Time period 3 (n=866), n (%)	Totals (n=1491), n (%)	*P* values
General individual or community well-being[a]	10 (22.7)	11 (8.5)	50 (16.8)	71 (15.1)	.03
Other consequence	0 (0)	3 (2.3)	7 (2.3)	10 (2.1)	.59
Policies mentioned	102 (48.6)	196 (47.2)	494 (57)	792 (53.1)	.002
Generic of nonspecific policies (reduction programs, initiatives, or reforms)	68 (66.7)	120 (61.2)	288 (58.3)	476 (60.1)	.27
Restrictive firearm policies	21 (20.6)	55 (28.1)	115 (23.3)	191 (24.1)	.28
Permissive firearm policies[d]	7 (6.9)	6 (3.1)	55 (11.1)	68 (8.6)	.002
Funding for CVIP/HVIP programs[d,b,e]	2 (2)	5 (2.6)	45 (9.1)	52 (6.6)	<.001
Funding for firearm research	1 (1)	0 (0)	4 (0.8)	5 (0.6)	.43
Address mental health and illness	1 (1)	4 (2)	12 (2.4)	17 (2.1)	.65
Address substance use disorder	0 (0)	0 (0)	2 (0.4)	2 (0.3)	.55
Funding or support for law enforcement[a]	3 (2.9)	0 (0)	8 (1.6)	11 (1.4)	.09
Hold firearm industry accountable	1 (1)	1 (0.5)	3 (0.6)	5 (0.6)	.88
Hold the gun lobby accountable	2 (2)	3 (1.5)	1 (0.2)	6 (0.8)	.06
Address structural determinants of violence	0 (0)	1 (0.5)	7 (1.4)	8 (1)	.31
Other policy mentioned	5 (4.9)	16 (8.2)	28 (5.7)	49 (6.2)	.40

[a]Post hoc pairwise comparisons significant for Time 1 versus Time 2.

[b]Post hoc pairwise comparisons significant for Time 1 versus Time 3.

[c]Not applicable.

[d]Post-hoc pairwise comparisons significant for Time 2 versus Time 3.

[e]CVIP: Community-Based Violence Intervention Program/HVIP: Hospital-Based Violence Intervention Program.

Among 381 posts that mentioned a specific category of firearm violence, mass shootings were the most common (253, 66.4%), with statistically significant variation across the 3 time periods (22/43, 51% vs 91/118, 77.1% vs 140/220, 63.6%; *P*=.004). Interpersonal or community violence was the next most frequently mentioned category (79, 20.7%), with a significant increase after the onset of the pandemic. Suicide and unintentional injury were rarely discussed.

When specific locations were mentioned (358 posts), legislators most often named cities or counties (195/358, 54.5%) and places of worship (91/358, 25.4%). Mentions of places of worship peaked following the Tree of Life Synagogue shooting. Additionally, a significant portion of posts (138/1491, 9.3%) described vigils or remembrance events, aligning with the symbolic frame of “Remembrance,” which was the most commonly expressed symbolic response (133/294, 45.2% posts) and increased significantly over time. “Solidarity” and “Thoughts and Prayers” were also common expressions (113/294, 38.4% and 90/294, 30.6%) without significant changes over time. Pennsylvania legislators also mentioned “Calls to Action” in 457/1491 (30.7%) posts of the time, with the majority representing general calls to action (331/457, 72.4%) rather than specific ones (204/457, 14.2%).

The most common frames included advocacy or endorsement (1051/1422, 73.9% posts), emotional appeal or moral outrage (659/1422, 46.3%), partisan framing (226/1422, 15.9%), and

XSL•FO
RenderX

informational framing (169/1422, 11.9%). Of these, partisan frames increased significantly by 14.5% ($P<.001$) during the COVID-19 pandemic, while emotional and informational frames differed significantly between the first and third time periods. Health or public health frames were rare, appearing in less than 1 in 10 (133/1422, 9.4%) posts. Pairwise comparisons between frames indicated that posts using advocacy or emotional frames were significantly less likely to co-occur with informational frames ($r=-0.24$; $P<.001$ and $r=-0.18$; $P<.001$, respectively). Religion and cultural frames were positively correlated ($r=0.15$; $P<.001$).

Among the 767/1491 (51.4%) posts that attributed causal blame, the most common attributions were to weak firearm laws—a nonspecific category where authors criticized laws as generally insufficient without referencing a particular statute or policy (318/767, 41.5% posts)—and to increased firearm ownership, access, or availability (149/767, 19.4% posts). Legislators also frequently blamed individual policymakers (118/767, 15.4%) or political parties (146/767, 19%), with blame attribution toward political parties significantly increasing by 23% during the COVID-19 pandemic period. A majority of posts (792/1491, 53.1%) included a mention of at least one nonspecific or specific policy. Nonspecific policy mentions (eg, "gun reform," "initiatives," or "common-sense laws") were cited more often than specific policy proposals (476/792, 60.1%). Among posts describing policies, nearly a quarter (191/792, 24.1%) referenced policy proposals that restrict access to firearms, such as background checks or extreme risk protection orders.

Analytic Results: Primary Model

Adjusting for time period, our analysis of the relationship between any policy mention and rhetorical predictors derived from our descriptive content analysis revealed several features that were independently and significantly associated with policy-related content (Figure 3). Certain types of rhetorical framing and causal blame attributions were associated with policy mentions. For instance, posts that used advocacy frames had an almost 5 times greater odds of including any policy mention (aOR 4.67, 95% CI 3.55 - 6.16) in addition to those that used partisan rhetorical framing (aOR 2.89, 95% CI 1.91 - 4.38). Additionally, posts casting causal blame on weak firearm laws (aOR 7.88, 95% CI 5.14 - 12.09) were significantly more likely to include any policy mention. In contrast, posts referencing mass shootings (aOR 0.54, 95% CI 0.37 - 0.77) or using emotional frames (aOR 0.53, 95% CI 0.40 - 0.69) were less likely to mention policy. The odds of any policy mention remained stable after holding time constant throughout all 3 time periods. Neither of the latter time periods was significantly associated with policy mentions, suggesting associations between any policy mention and rhetorical elements were consistent across time.

Figure 3. Adjusted odds ratios for predictors of any policy mention in X posts by Pennsylvania state legislators (2017-2022).

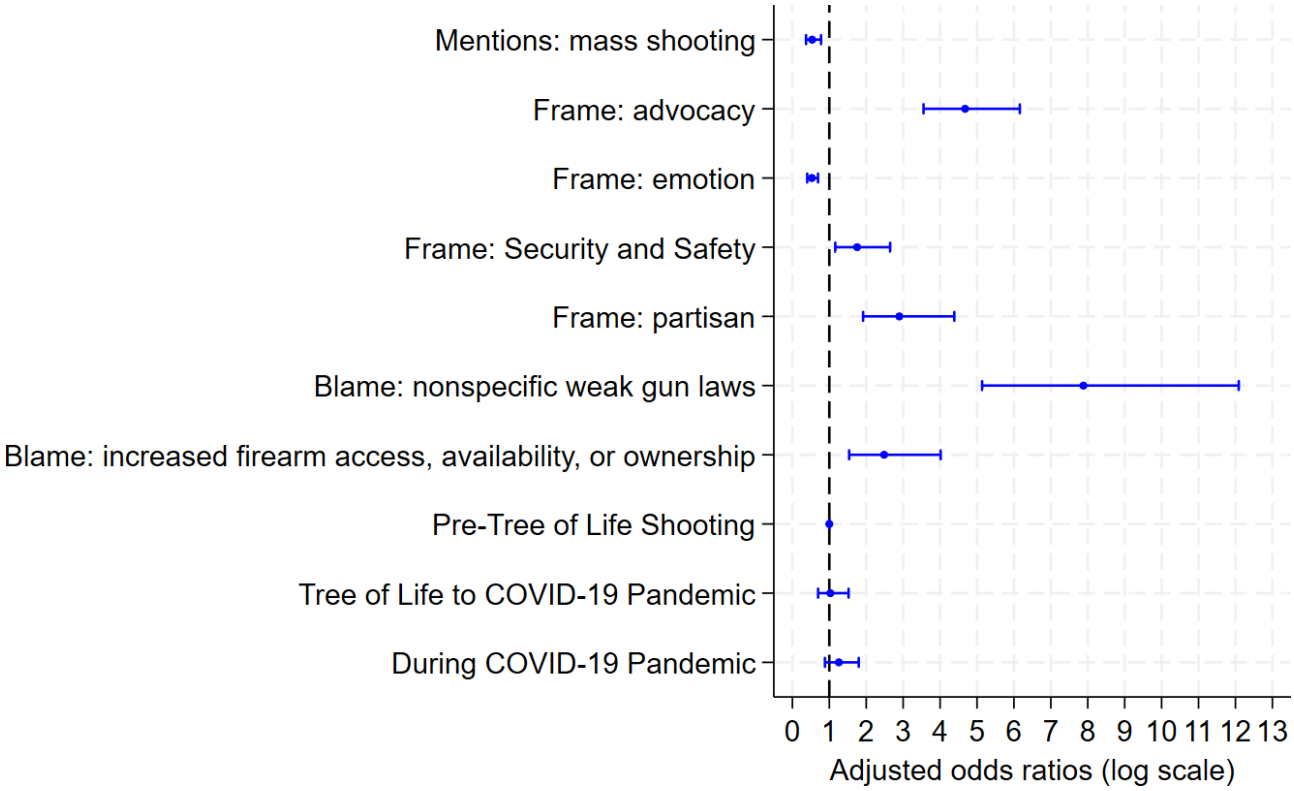

Analytic Results: Stratified Policy Model

When stratifying by policy specificity, we observed notable differences (Figure 4). Posts using advocacy frames were significantly associated with both nonspecific (aOR 2.97, 95% CI 2.13 - 4.13) and specific policy mentions (aOR 2.77, 95% CI 1.88 - 4.06). Causal blame directed toward weak firearm laws independently predicted nonspecific policy mentions (aOR 8.26, 95% CI 6.02 - 11.35) but was inversely associated with specific policy mentions (aOR 0.50, 95% CI 0.35 - 0.72). In contrast, posts specifying blame toward increased access, availability, or ownership of firearms were more likely to reference specific policies (aOR 6.37, 95% CI 4.29 - 9.47) and were negatively associated with nonspecific policy mentions

XSL•FO
RenderX

(aOR 0.26, 95% CI 0.17 - 0.42). General calls to action were also associated with increased odds of nonspecific policy mentions (aOR 1.71, 95% CI 1.28 - 2.27) and decreased odds of specific policy mentions (aOR 0.50, 95% CI 0.35 - 0.72). Finally, posts referencing mass shootings were the least likely to mention specific policies (aOR 0.18, 95% CI 0.10 - 0.35).

Figure 4. Adjusted odds ratios for predictors of generic versus specific policy mentions in X posts by Pennsylvania state legislators (2017-2022).

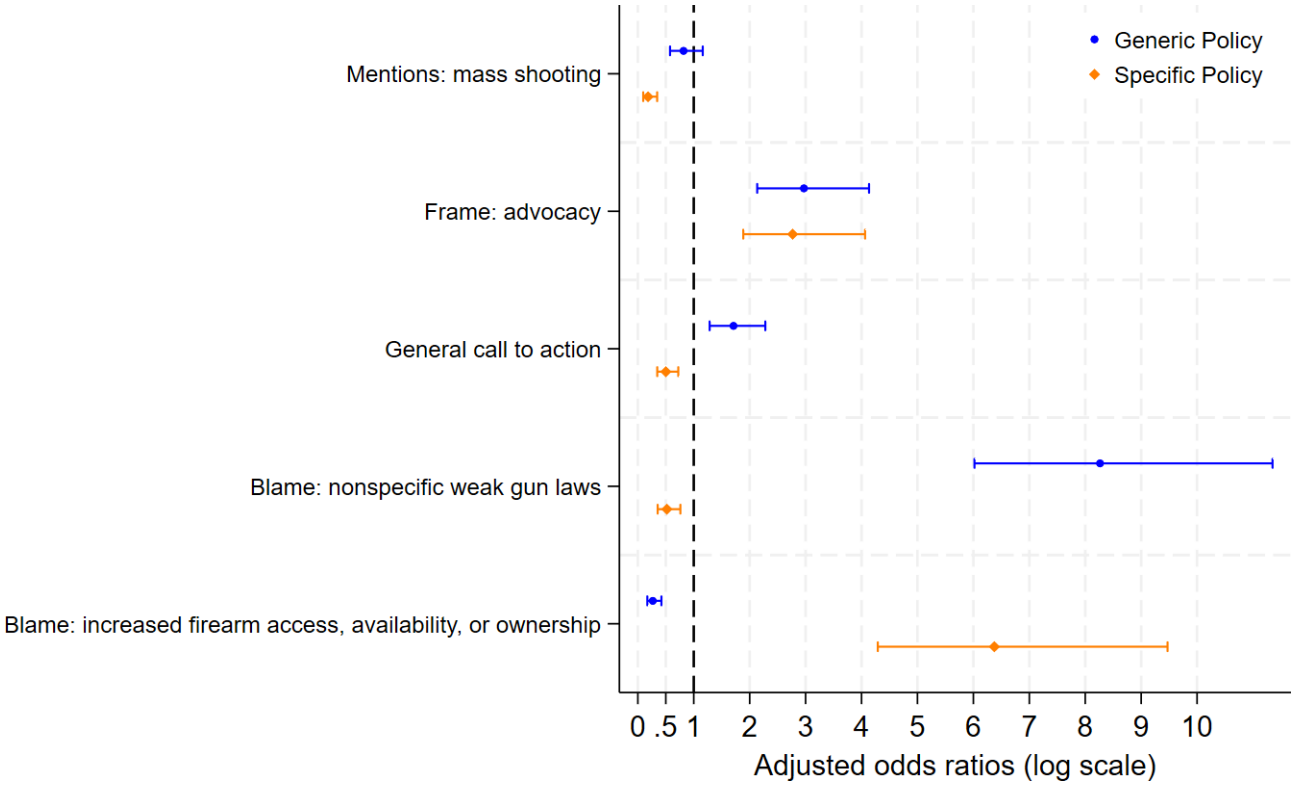

Individual Causal Blame, Shooting Type, and Frame Models

Three additional models, adjusted for time period, identified independent relationships between any policy mention and type of shooting, causal blame attributions, and rhetorical framing.

In the model assessing the relationship between shooting type, posts referencing mass shootings had a significantly lower odds of any policy mention (aOR 0.35, 95% CI 0.26 - 0.47), while posts about firearm suicide had an almost 3 times greater odds of including a policy mention (aOR 2.81, 95% CI 1.18 - 6.70). Other categories of firearm violence were not significantly associated with policy mentions. Posts made during the COVID-19 period had higher odds of including a policy mention compared to those posted prior to the Tree of Life shooting (aOR 1.51, 95% CI 1.11 - 2.05; *P*=.009).

In the causal blame model, posts that attributed blame to increased firearm availability, accessibility, and ownership (aOR 3.26, 95% CI 1.98 - 5.37) or to a specific political party (2.71, 95% CI 1.66 - 4.43) were significantly associated with any policy mention. Although only 4% (31/767) of posts attributed firearm violence to mental health, these posts had significantly higher odds of any policy mention (aOR 4.60, 95% CI 1.55 - 13.61). Other causal blame attributions were not associated with mentions of policy. The time period was not significantly associated with the likelihood of policy mention.

Finally, in the rhetorical frame model, posts using advocacy or endorsement (aOR 6.19, 95% CI 4.69 - 8.16), public opinion (aOR 2.41, 95% CI 1.25 - 4.67), security (aOR 2.56, 95% CI 1.68 - 3.90), constitutionality and jurisprudence (aOR 2.40, 95% CI 1.51 - 3.82), and partisan frames (aOR 3.62, 95% CI 2.45 - 5.37) had a greater odds of including any policy mention. In contrast, emotion (aOR 0.58, 95% CI 0.46 - 0.75) and firearm culture frames (aOR 0.28, 95% CI 0.13 - 0.63) were associated with lower odds of any policy mention. The time period was not a significant predictor of policy mention.

Subgroup Analysis: Health Frame Comparison

Of the 1491 posts analyzed, 133 (9.4%) used health or public health frames. A greater percentage of posts that used health frames, compared to those that did not, mentioned suicide (120/133, 9% vs 20/1358, 1.5%; *P*<.001) and mental health consequences (9/133, 6.8% vs 8/1358, 0.6%; *P*<.001) (Example posts provided in Table 1). Physical consequences, such as mortality or disability, were referenced in 43.6% (58/133) of health-framed posts, compared to just 15.9% (216/1358) of posts without health framing (*P*<.001). Notably, causal blame directed toward a political party was less common in health-framed posts (5/133, 3.8%) than in those without health framing (141/1358, 10.4%; *P*<.001). Comparisons are illustrated in Figure 5.

XSL•FO
RenderX

Figure 5. Descriptive unadjusted comparisons of significant characteristics by health frame usage in X posts by Pennsylvania state legislators (2017-2022).

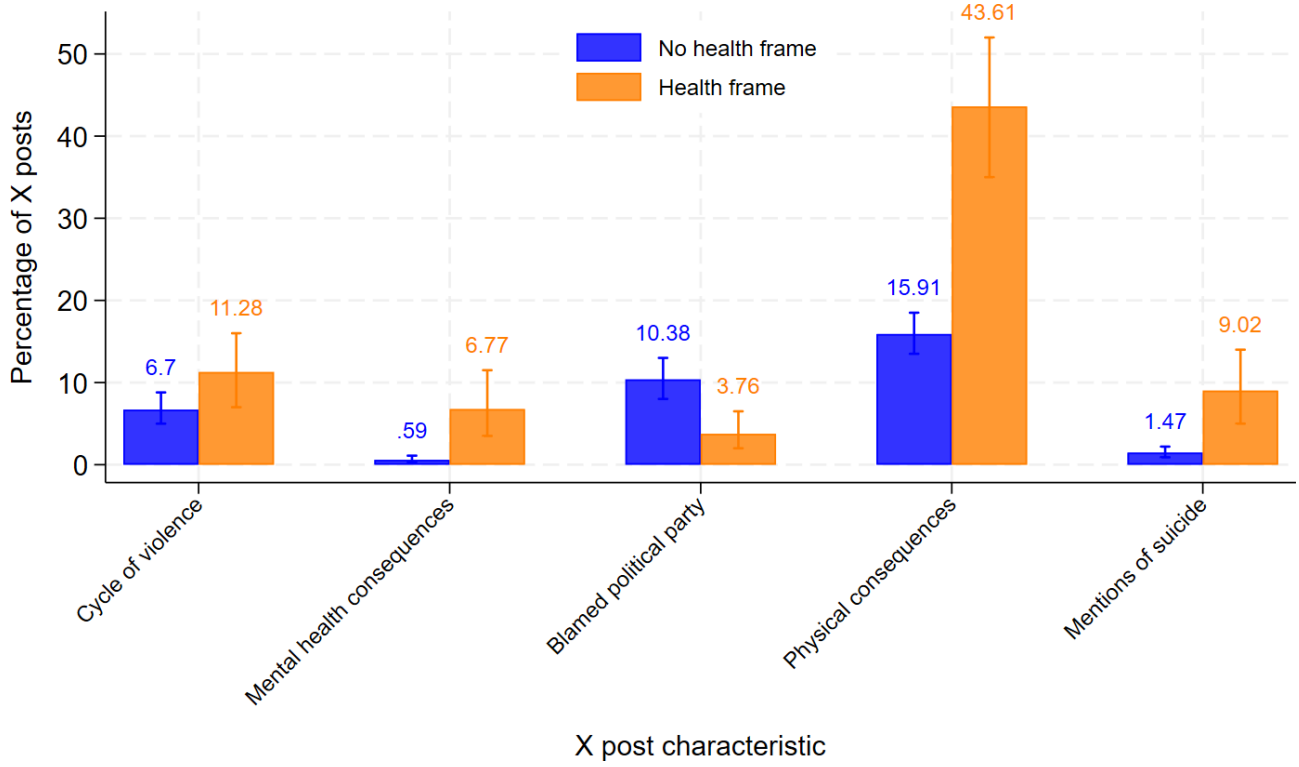

Discussion

Principal Findings

This mixed methods analysis of X posts from Pennsylvania state legislators revealed key patterns in how firearm violence is framed and discussed in the political sphere. Legislators most frequently focused on mass shootings, often pairing them with symbolic expressions of remembrance and emotional appeals, and rarely connected these tragedies to specific policy proposals. Emotional frames, particularly those evoking moral outrage, were common, yet typically lacked actionable next steps. Health frames were rarely used. Of note, partisan rhetorical framing increased over the study period. Concurrently, causal blame was increasingly directed toward political parties and individual policymakers. Most posts relied on general advocacy rhetoric and nonspecific causal attributions that, in turn, yielded vague or generic policy suggestions. These patterns provide insight into the prevailing rhetorical strategies used by policymakers and underscore the opportunities for more evidence-informed communication, especially approaches grounded in public health, to more meaningfully address firearm violence.

Despite only accounting for 1% - 2% of firearm injuries, mass shootings dominate public discourse, overshadowing more prevalent forms of firearm violence, such as suicide [50]. Between 2019 and 2023 in Pennsylvania, 56% of firearm deaths were by suicide, 40% by homicide, and 1% by unintentional injury [51]. This is particularly salient for rural communities, where firearm suicide rates are high and health infrastructure is often limited [52]. While suicides account for the majority of firearm fatalities, most nonfatal firearm injuries result from interpersonal or community violence [53]. This may represent a gap in legislators' knowledge or lack of acknowledgment of violence epidemiology and can undermine support for interventions targeted at the leading drivers of firearm injury [54-56].

Policymakers may also struggle with how to engage in ethical, nontrauma-inducing communication about suicide, contributing to a void in public rhetoric. This presents an opportunity for public health experts, communication specialists, and policymakers to cocreate trauma-informed messaging and reframe policy narratives around firearm suicide prevention, such as those provided by the American Foundation for Suicide Prevention, which is an approach that could foster broader support for evidence-informed injury prevention strategies [57]. In parallel, these same groups can collaborate to highlight the underlying structural drivers of community firearm violence, rather than focusing on event-specific details that risk retraumatization, to advance upstream, evidence-informed approaches to reduce both fatal and nonfatal firearm injuries [58].

Much of the current rhetoric by Pennsylvania legislators falls within a generalized advocacy frame characterized by emotional appeals in the aftermath of mass shootings. While these emotions are certainly valid and highlight the devastating toll of firearms, they are rarely accompanied by concrete calls to action or specific policy solutions. Instead, policymakers often rely on vague language, blaming "weak laws" or issuing nonspecific calls for change, without tying these claims to clear legislative proposals. This pattern, particularly when paired with highly emotional framing, may diminish the utility of such messages for constituents, colleagues, and coalition-building efforts. Policymakers may choose to remain nonspecific to appeal to a broad base of constituents with diverse political and ideological views in order to secure their electoral seat. For example, in a study of members of parliament (MPs) in the United Kingdom House of Commons, speeches by electorally

XSL•FO
RenderX

vulnerable and junior MPs contained higher levels of emotive language compared to MPs with electoral security [59]. These emotions were mainly anger, disgust, fear, and sadness, which mirrors rhetoric used to characterize mass shootings. Recognizing that policymakers, like the broader public, process tragedy through emotion underscores the need to respond with empathy rather than condescension. A valuable next step is to explore how emotional responses can be intentionally channeled into public health communication that drives specific policy outcomes.

Policymakers may view social media as an ill-suited platform for nuanced policy discussion, even though prior research suggests it can influence policy agendas. Our data indicate that policymakers often use social media to emphasize partisan talking points rather than share information grounded in public health or scientific evidence. This is not surprising. As social media use has grown, platform algorithms have amplified divisive content, reinforcing users' preexisting beliefs and deepening political polarization, including among policymakers themselves [60]. These dynamics unfold alongside growing mistrust in both political and public health institutions. In this context, and amid pressure to appeal to voter bases critical for nomination and electoral success, policymakers may feel compelled to conform to dominant partisan and polarized rhetoric, limiting their ability or willingness to communicate in ways that promote shared understanding or evidence-informed policy.

Despite broader political dynamics, current messaging techniques present an opportunity for refinement. For instance, when legislators directly attributed firearm violence to factors, such as firearm availability, accessibility, or ownership, they were more likely to propose concrete policy responses. This suggests that greater specificity in causal attribution can support more actionable policy discourse. A strengths-based approach, which builds on policymakers' existing communication skills while encouraging improved clarity, accuracy, and alignment with public values, to message crafting could help policymakers pivot from general advocacy toward more effective communication strategies rooted in specificity, accuracy, and resonance with public values [61-64].

Although health-related frames were only used in 9.4% of legislators' posts, they tended to appear in posts discussing the mental and physical consequences of firearm violence and were more likely to reference empirical evidence rather than partisan rhetoric. However, noticeably, health frames were used reactively to discuss the consequences of firearm violence rather than focus on firearm violence prevention, which is central to a public health approach. Health and public health frames have demonstrated success in advancing policy agendas, particularly when rooted in the authority and credibility of health care professionals [65]. Additionally, centering health garners favor from individuals across ideological spectrums for firearm injury prevention policies, such as extreme risk protection orders [66]. A study by Ojo et al [65] compared content, affect, and authorship of posts using the #ThisIsOurLane hashtag with #GunViolence over 1 year. Posts authored by health care providers and containing the #ThisIsOurLane hashtag were more likely to be framed through a health or public health-specific lens, connote positive emotions, and to contain more action-oriented content. This highlights the potential of health frames to amplify evidence-informed, preventive, and solutions-oriented messaging in the firearm policy space.

In addition to incorporating health frames, legislators might benefit from communication strategies that emphasize shared values, common goals, and inclusive identities, all of which can reduce polarization and foster greater message acceptance [60]. For example, framing firearm safety as a way to protect "Pennsylvanians," families, children, and community well-being may appeal across ideological divides. Self-affirmation and trust-building techniques, such as acknowledging shared concerns and engaging constituents with empathy, may also help bridge divides. Similarly, emphasizing social norms (eg, "most parents support secure firearm storage") and humanizing the communicator (eg, "as a parent and former law enforcement officer, I worry about...") can increase message relatability and trust [60]. These strategies, shown to build common ground and reduce defensiveness, may help legislators communicate more effectively about contentious topics while still advancing specific, evidence-informed policy solutions. Further, resources exist to guide journalists in ethical, accurate, and equitable reporting on mass shootings and community firearm violence [58,67]. Although these guidelines are designed for journalists, their principles, when coupled with media literacy training, could inform the development of evidence-based communication guidelines for policymakers, who also play a key role in shaping public discourse and preventing harm.

Public health experts, clinicians, and clinician-researchers have a longstanding tradition of patient-centered communication and are well-positioned to collaborate with policymakers by translating complex evidence into accessible messages that minimize harm. Social media represents a natural extension of health professionals' communication skills; however, few health professionals have training in public-facing media literacy or large-scale message framing. By integrating the expertise of health care professionals, the skills of communication specialists, and the platforms of policymakers, there is a clear opportunity to advance health-oriented framing and reshape firearm policy narratives to emphasize prevention as a form of health promotion [68,69]. For example, in the Agree to Agree campaign, health professionals partnered with the Ad Council, a communication firm, to develop public-facing messaging that helps caregivers discuss secure firearm storage with other caregivers [70]. Similarly, Da El Siguiente Paso (Take the Next Step), developed in collaboration with the Ad Council, offers a culturally relevant secure storage campaign for Hispanic firearm owners [71]. This bidirectional relationship also offers an opportunity to humanize both policymakers and health professionals, fostering trust and reducing stereotypes on both sides [60].

Building on this, future studies should broaden the scope of analysis by tracking engagement metrics from constituents and the general public, incorporating more states, and examining longer time frames. Additionally, similar to chart-stimulated recall used in health services research, involving policymakers in qualitative analysis of their own posts may foster reflective discourse that identifies barriers and facilitators to incorporating health framing and evidence into public communications. There

XSL•FO
RenderX

is also considerable potential to leverage machine learning and large language models to support this work at scale. This descriptive analysis lays the groundwork for cocreating more accurate, ethical, and equitable portrayals of firearm violence on social media. Future research can further explore how health framing and survivor testimony influence message effectiveness and policy uptake, particularly the relationship between health-specific language and specific evidence-informed policy mentions.

Our findings highlight the potential importance of concerted collaboration between policymakers, advocates, researchers, and community partners to amplify survivor stories, honor lives lost, and promote specific, actionable steps for reform. Messages that combine health framing, evidence-informed policy, and emotional resonance may be especially effective in shifting public opinion and mobilizing support for firearm injury prevention. Enhancing the precision and relevance of firearm-related messaging, grounded in health and public health principles, can help bridge the gap between tragedy and legislative action. Providing policymakers and their staff with communication templates rooted in data and health equity may ultimately strengthen efforts to advance firearm violence prevention policies and promote individual and community well-being.

Limitations

Our study has several limitations. First, our sample was limited to X posts and did not include other traditional or social media platforms that many policymakers also use to communicate. We also did not track user engagement metrics, such as reposts or likes, which could help assess the reach and resonance of specific messages. We opted to analyze the frames and solutions proposed in order to offer greater insight into policymaker perspectives rather than examining engagement metrics. Because we conducted a qualitative content analysis, incorporating multiple sources and studying engagement patterns was beyond the scope of this project. Second, the advocacy and endorsement frame was broadly coded to include general advocacy toward a cause or solution, specific policy advocacy, and self-advocacy. This broad categorization may have contributed to its strong correlation with policy mentions—both general and specific. The advocacy frame warrants more granular analysis and will be explored in future work. Third, assessing individual policymaker attributes or district characteristics was outside the scope of this study. However, future research could examine how district demographics and political context may influence rhetorical strategies. Fourth, regression results should be interpreted with caution due to the relatively small sample size. This limitation informed our decision to rely primarily on descriptive prevalence rather than model-driven inference. These associations should be interpreted as correlational rather than causal, due to the cross-sectional nature of the data. Fourth, our analysis did not account for external influences on social media rhetoric, such as concurrent events in the news cycle or geopolitical developments, which may have shaped the timing and content of posts. Fifth, although the study was not formally powered for inferential hypothesis testing, the sample was considered sufficiently large to ensure stability of model estimates and meaningful comparison between timepoints. Finally, our use of a priori frame definitions may have limited the capture of emergent frames; however, this approach ensured consistency and replicability across a large dataset and was balanced by iterative refinement during codebook development.

Conclusion

By analyzing the language state legislators use when discussing firearm violence on social media, this study offered an empirical foundation to understand how political discourse aligns, or misaligns, with the health and public health realities. In describing firearm violence, Pennsylvania state legislators primarily focused on mass shootings, often relying on general advocacy rhetoric and vague causal attributions, which led to nonspecific policy proposals. This pattern persisted across all 3 time periods included in our analysis, even after adjusting for time using indicator variables in our regression models. These findings highlight opportunities for policymakers to leverage their strengths in communication, such as conveying emotion and personalizing complex issues, while becoming better informed about the epidemiology of firearm violence and evidence-based solutions. Adopting health-informed communication strategies can help advance specific policy recommendations to prevent all forms of firearm violence, especially firearm suicide.

Acknowledgments

Support for this project through data acquisition was provided by the Center for Health Economics of Treatment Interventions for Substance Use Disorder, HCV (Hepatitis C Virus), and HIV (CHERISH), a National Institute on Drug Abuse research center (P30 DA040500) and in partnership with the Research-to-Policy Collaboration, affiliated with The Pennsylvania State University's Edna Bennett Pierce Prevention Research Center.

Funding

VAA's work on this project was supported by the Ruth L Kirchstein-National Service Research Award through the Health Resources and Services Administration (grant T32-HP46118-02), the Children's Hospital of Philadelphia 2024-2025 PolicyLab and Clinical Futures Pilot Grant, and the University of Pennsylvania Perelman School of Medicine's National Clinician Scholars Program. The funders had no role in the preparation, review, or approval of the manuscript and decision to submit the manuscript for publication.

XSL•FO
RenderX

Data Availability

The datasets generated or analyzed during this study are available in a public GitHub repository [72].

Authors' Contributions

Conceptualization: VAA, JAF, TSR, JP, ZFM
Data curation: VAA, WJKV, KSK, IN, MDK
Formal analysis: VAA, WJKV, KSK, AC, IN, OB, AS, MDK
Funding acquisition: VAA
Methodology: VAA, JAF, TSR, ZFM, WJKV, KSK, AC, IN, MDK
Project administration: VAA
Software: VAA
Visualization: VAA
Writing–original draft: VAA
Writing–review & editing: VAA, WJKV, KSK, AC, IN, OB, AS, JAF, TSR, JP, MDK, ZFM

Conflicts of Interest

None declared.

Multimedia Appendix 1
Codebook used for qualitative content analysis.
[DOCX File, 28 KB - formative_v9i1e80397_app1.docx]

References

1. Roberts BK, Nofi CP, Cornell E, Kapoor S, Harrison L, Sathya C. Trends and disparities in firearm deaths among children. Pediatrics 2023 Sep 1;152(3):e2023061296. [doi: 10.1542/peds.2023-061296] [Medline: 37599647]
2. Kegler SR, Simon TR, Zwald ML, et al. Vital signs: changes in firearm homicide and suicide rates — United States, 2019–2020. MMWR Morb Mortal Wkly Rep 2022 May 13;71(19):656-663. [doi: 10.15585/mmwr.mm7119e1] [Medline: 35550497]
3. Lee LK, Fleegler EW, Goyal MK, et al. Firearm-related injuries and deaths in children and youth: injury prevention and harm reduction. Pediatrics 2022 Oct 8;150(6):e2022060070. [doi: 10.1542/peds.2022-060070] [Medline: 36207776]
4. Goldstick JE, Cunningham RM, Carter PM. Current causes of death in children and adolescents in the United States. N Engl J Med 2022 May 19;386(20):1955-1956. [doi: 10.1056/NEJMc2201761] [Medline: 35443104]
5. Miller GF, Barnett SBL, Florence CS, McDavid Harrison K, Dahlberg LL, Mercy JA. Costs of fatal and nonfatal firearm injuries in the U.S., 2019 and 2020. Am J Prev Med 2024 Feb;66(2):195-204. [doi: 10.1016/j.amepre.2023.09.026] [Medline: 38010238]
6. Oliphant JB. Bipartisan support for some gun proposals, stark partisan divisions on many others. Pew Research Center. 2017 Jun. URL: http://pewrsr.ch/2rY7Tgk [accessed 2025-11-05]
7. Pearson‐Merkowitz S, Dyck JJ. Crime and partisanship: how party ID muddles reality, perception, and policy attitudes on crime and guns. Soc Sci Q 2017 Jun;98(2):443-454. [doi: 10.1111/ssqu.12417]
8. Impact of gun violence on historically marginalized communities. Everytown for Gun Safety. 2024. URL: https://www.everytown.org/issues/impact-of-gun-violence-on-historically-marginalized-communities/ [accessed 2025-11-05]
9. Truman J, Morgan R. Violent victimization by sexual orientation and gender identity, 2017–2020. Bureau of Justice Statistics. 2022 Jun. URL: https://bjs.ojp.gov/document/vvsogi1720.pdf [accessed 2025-11-05]
10. Web-Based Injury Statistics Query and Reporting System (WISQARS). Centers for Disease Control and Prevention. URL: https://wisqars.cdc.gov/ [accessed 2024-07-08]
11. Kaczkowski W, Kegler SR, Chen MS, Zwald ML, Stone DM, Sumner SA. Notes from the Field : Firearm Suicide Rates, by Race and Ethnicity — United States, 2019–2022. MMWR Morb Mortal Wkly Rep 2023 Dec 1;72(48):1307-1308. [doi: 10.15585/mmwr.mm7248a3]
12. Zagorski N. Gun violence is an underrecognized social determinant of health. PN 2022 Dec 1;57(12). [doi: 10.1176/appi.pn.2022.12.12.43]
13. Gilardi F, Gessler T, Kubli M, Müller S. Social media and political agenda setting. Polit Commun 2022 Jan 2;39(1):39-60. [doi: 10.1080/10584609.2021.1910390]
14. Dawes D. The Political Determinants of Health: Johns Hopkins University Press; 2020. [doi: 10.56021/9781421437903]
15. Yeung D. Social media as a catalyst for policy action and social change for health and well-being: viewpoint. J Med Internet Res 2018 Mar 19;20(3):e94. [doi: 10.2196/jmir.8508] [Medline: 29555624]
16. Hemphill L, Russell A, Schöpke‐Gonzalez AM. What drives U.S. congressional members' policy attention on Twitter? Policy Internet 2021 Jun;13(2):233-256 [FREE Full text] [doi: 10.1002/poi3.245]

XSL•FO
RenderX

17. Blum R, Cormack L, Shoub K. Conditional Congressional communication: how elite speech varies across medium. Political Sci Res Method 2023 Apr;11(2):394-401. [doi: 10.1017/psrm.2022.28]
18. Payson J, Casas A, Nagler J, Bonneau R, Tucker JA. Using social media data to reveal patterns of policy engagement in state legislatures. State Politics Policy Q 2022 Dec;22(4):371-395. [doi: 10.1017/spq.2022.1]
19. Cook JM. Twitter adoption and activity in U.S. legislatures: a 50-state study. Am Behav Sci 2017 Jun;61(7):724-740. [doi: 10.1177/0002764217717564]
20. Peeters J, Van Aelst P, Praet S. Party ownership or individual specialization? A comparison of politicians' individual issue attention across three different agendas. Party Politics 2021 Jul;27(4):692-703. [doi: 10.1177/1354068819881639]
21. Purtle J, Soltero M, Crane ME, McSorley AMM, Knapp M, Drapeau CW. State legislator social media posts about the 988 suicide and crisis lifeline. JAMA Netw Open 2023 Oct 2;6(10):e2339845. [doi: 10.1001/jamanetworkopen.2023.39845] [Medline: 37883089]
22. Stokes DC, Purtle J, Meisel ZF, Agarwal AK. State legislators' divergent social media response to the opioid epidemic from 2014 to 2019: longitudinal topic modeling analysis. J Gen Intern Med 2021 Nov;36(11):3373-3382. [doi: 10.1007/s11606-021-06678-9] [Medline: 33782896]
23. Engel-Rebitzer E, Stokes DC, Meisel ZF, Purtle J, Doyle R, Buttenheim AM. Partisan differences in legislators' discussion of vaccination on twitter during the COVID-19 era: natural language processing analysis. JMIR Infodemiology 2022;2(1):e32372. [doi: 10.2196/32372] [Medline: 35229075]
24. Abrams MP, Pelullo AP, Meisel ZF, Merchant RM, Purtle J, Agarwal AK. State and federal legislators' responses on social media to the mental health and burnout of health care workers throughout the COVID-19 pandemic: natural language processing and sentiment analysis. JMIR Infodemiology 2023;3:e38676. [doi: 10.2196/38676] [Medline: 37013000]
25. Tai YC, Buma R, Desmarais BA. Official yet questionable: examining misinformation in U.S. state legislators' tweets. J Inf Technol Politics 2024 Oct;21(4):597-609. [doi: 10.1080/19331681.2023.2257682]
26. Peterson DAM, Tavanapong W, Qi L, Sukul A, Khaleel M. The public‐facing policy agenda of state legislatures: the communication of public policy via twitter. Policy Stud J 2023 Aug;51(3):551-571 [FREE Full text] [doi: 10.1111/psj.12485]
27. Khullar D. Social media and medical misinformation: confronting new variants of an old problem. JAMA 2022 Oct 11;328(14):1393-1394. [doi: 10.1001/jama.2022.17191] [Medline: 36149664]
28. Harvey A. Combatting health misinformation and disinformation: building an evidence base. Health Aff Forefront 2021. [doi: 10.1377/forefront.20211118.932213]
29. Degli Esposti M, Goldstick J, Gravel J, et al. How have firearm laws changed in states with unexpected decreases or increases in firearm homicide, 1990–2019? SSM Popul Health 2023 Jun;22:101364. [doi: 10.1016/j.ssmph.2023.101364]
30. Rozel JS, Mulvey EP. The link between mental illness and firearm violence: implications for social policy and clinical practice. Annu Rev Clin Psychol 2017 May 8;13:445-469. [doi: 10.1146/annurev-clinpsy-021815-093459] [Medline: 28375722]
31. Gun violence: prediction, prevention, and policy. : American Psychological Association; 2013 Dec URL: https://www.apa.org/pubs/reports/gun-violence-prevention [accessed 2025-11-05]
32. Swanson JW, McGinty EE, Fazel S, Mays VM. Mental illness and reduction of gun violence and suicide: bringing epidemiologic research to policy. Ann Epidemiol 2015 May;25(5):366-376. [doi: 10.1016/j.annepidem.2014.03.004] [Medline: 24861430]
33. Beltzer ML, Moulder RG, Baker C, Comer K, Teachman BA. Effects of Mass Shootings on Mental Illness Stigma in the United States. Pers Soc Psychol Bull 2023 Aug;49(8):1231-1247. [doi: 10.1177/01461672221097180] [Medline: 35658698]
34. McGinty EE, Kennedy-Hendricks A, Choksy S, Barry CL. Trends in news media coverage of mental illness in the United States: 1995–2014. Health Aff (Millwood) 2016 Jun 1;35(6):1121-1129. [doi: 10.1377/hlthaff.2016.0011] [Medline: 27269031]
35. Wilson LC, Ballman AD, Buczek TJ. News content about mass shootings and attitudes toward mental illness. Journal Mass Commun Q 2016 Sep;93(3):644-658. [doi: 10.1177/1077699015610064]
36. Brucato G, Appelbaum PS, Hesson H, et al. Psychotic symptoms in mass shootings v . mass murders not involving firearms: findings from the Columbia mass murder database. Psychol Med 2022 Nov;52(15):3422-3430. [doi: 10.1017/S0033291721000076]
37. Parham-Payne W. The role of the media in the disparate response to gun violence in America. J Black Stud 2014 Nov;45(8):752-768. [doi: 10.1177/0021934714555185]
38. Armstrong M, Carlson J. Speaking of trauma: the race talk, the gun violence talk, and the racialization of gun trauma. Palgrave Commun 2019 Oct 1;5(1):112. [doi: 10.1057/s41599-019-0320-z]
39. Reny TT, Newman BJ, Holbein JB, Hassell HJG. Public mass shootings cause large surges in Americans' engagement with gun policy. PNAS Nexus 2023 Dec;2(12):pgad407. [doi: 10.1093/pnasnexus/pgad407] [Medline: 38111824]
40. Donnelly M, Kuza C, Sargent B, et al. Firearm violence surrounding the COVID-19 pandemic: a reopening phenomenon. J Surg Res 2023 May;285:168-175. [doi: 10.1016/j.jss.2022.12.017] [Medline: 36680877]
41. Miller M, Zhang W, Azrael D. Firearm purchasing during the COVID-19 pandemic: results from the 2021 National Firearms Survey. Ann Intern Med 2022 Feb;175(2):219-225. [doi: 10.7326/M21-3423] [Medline: 34928699]
42. Quorum. 2023. URL: https://www.quorum.us/ [accessed 2024-09-30]

XSL•FO
RenderX

43. Demszky D, Garg N, Voigt R, et al. Analyzing Polarization in Social Media: Method and Application to Tweets on 21 Mass Shootings. Presented at: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers); Jun 3-5, 2019; Minneapolis, Minnesota p. 2970-3005 URL: http://aclweb.org/anthology/N19-1 [doi: 10.18653/v1/N19-1304]
44. Kingdon J. Alternatives, and Public Policies, 2nd edition: Pearson Education Limited; 2014.
45. Hoefer R. The multiple streams framework: Understanding and applying the problems, policies, and politics approach. J of Pol Practice & Research 2022 Mar;3(1):1-5. [doi: 10.1007/s42972-022-00049-2]
46. Mhazo AT, Maponga CC. Agenda setting for essential medicines policy in sub-Saharan Africa: a retrospective policy analysis using Kingdon's multiple streams model. Health Res Policy Syst 2021 May 3;19(1):72. [doi: 10.1186/s12961-021-00724-y] [Medline: 33941199]
47. Cairney P, Jones MD. Kingdon's multiple streams approach: what is the empirical impact of this universal theory? Policy Stud J 2016 Feb;44(1):37-58 [FREE Full text] [doi: 10.1111/psj.12111]
48. Jones MD, Peterson HL, Pierce JJ, et al. A river runs through it: a multiple streams meta-review. Policy Stud J 2016 Feb;44(1):13-36 [FREE Full text] [doi: 10.1111/psj.12115]
49. Boydstun AE, Card D, Gross J, Resnick P, Smith NA. Tracking the development of media frames within and across policy issues. : Carnegie Mellon University; 2014. [doi: 10.1184/R1/6473780.V1]
50. Metzl JM, MacLeish KT. Mental illness, mass shootings, and the politics of American firearms. Am J Public Health 2015 Feb;105(2):240-249. [doi: 10.2105/AJPH.2014.302242] [Medline: 25496006]
51. How does gun violence impact the communities you care about? EveryStat. URL: https://everystat.org/#Pennsylvania [accessed 2025-04-18]
52. Nestadt PS, Triplett P, Fowler DR, Mojtabai R. Urban-rural differences in suicide in the state of Maryland: the role of firearms. Am J Public Health 2017 Oct;107(10):1548-1553. [doi: 10.2105/AJPH.2017.303865] [Medline: 28817331]
53. Kaufman EJ, Song J, Xiong R, Seamon MJ, Delgado MK. Fatal and nonfatal firearm injury rates by race and ethnicity in the United States, 2019 to 2020. Ann Intern Med 2024 Sep;177(9):1157-1169. [doi: 10.7326/M23-2251] [Medline: 39074371]
54. Morgan ER, Rowhani-Rahbar A, Azrael D, Miller M. Public perceptions of firearm- and non-firearm-related violent death in the United States: a national study. Ann Intern Med 2018 Nov 20;169(10):734-737. [doi: 10.7326/M18-1533] [Medline: 30383115]
55. Genovesi AL, Donaldson AE, Morrison BL, Olson LM. Different perspectives: a comparison of newspaper articles to medical examiner data in the reporting of violent deaths. Accid Anal Prev 2010 Mar;42(2):445-451. [doi: 10.1016/j.aap.2009.09.006] [Medline: 20159065]
56. Schildkraut J. Mass shootings, legislative responses, and public policy: an endless cycle of inaction. Emory Law J 2020 Sep 9;69(5):1043 [FREE Full text]
57. Ethical reporting guidelines for media. American Foundation for Suicide Prevention. 2025. URL: https://afsp.org/ethicalreporting/ [accessed 2025-04-30]
58. Beard J. Better gun violence reporting: a toolkit for minimizing harm. Philadelphia Center for Gun Violence Reporting. 2024 May. URL: https://www.pcgvr.org/wp-content/uploads/2025/07/Better-Gun-Violene-Reporting-Toolkit.pdf [accessed 2025-10-28]
59. Yildirim TM. Emotions in the aisles: unpacking the use of emotive language in the UK House of Commons. European J Political Res 2025 May;64(2):943-959. [doi: 10.1111/1475-6765.12733]
60. Chou WYS, Iles IA, Gaysynsky A, Klein WMP. Public health communication approaches for building common ground. Am J Public Health 2025 Apr;115(4):511-518. [doi: 10.2105/AJPH.2024.308003] [Medline: 39913869]
61. Communicating for connection: values-based messaging for primary prevention in multicultural and faith-based communities. Safe+Equal. 2022. URL: https://safeandequal.org.au/resources/communicating-for-connection/ [accessed 2025-05-01]
62. Grene K, Dharani A, Siegel M. A common ground gun violence prevention policy package. Psychol Rep 2025 Feb;128(1):334-349. [doi: 10.1177/00332941241248602] [Medline: 38684128]
63. Gatewood J, Monks SL, Singletary CR, Vidrascu E, Moore JB. Social media in public health: strategies to distill, package, and disseminate public health research. J Public Health Manag Pract 2020;26(5):489-492. [doi: 10.1097/PHH.0000000000001096] [Medline: 32732723]
64. Shenker-Osorio A. Messaging this moment: a handbook for progressive communicators. : Centre for Community Change; 2017 URL: https://www.communitychange.org/wp-content/uploads/C3-Messaging-This-Moment-Handbook.pdf [accessed 2025-11-28]
65. Ojo A, Guntuku SC, Zheng M, Beidas RS, Ranney ML. How health care workers wield influence through Twitter hashtags: Retrospective cross-sectional study of the gun violence and covid-19 public health crises. JMIR Public Health Surveill 2021 Jan 6;7(1):e24562. [doi: 10.2196/24562] [Medline: 33315578]
66. Patrick SW, Loch SF, McNeer E, Davis MM. Voter support for policies associated with child health as national campaign priorities. JAMA Health Forum 2024 Sep 6;5(9):e243305. [doi: 10.1001/jamahealthforum.2024.3305] [Medline: 39331370]
67. Schildkraut J. Recommended practices for media outlets working with mass shooting survivors. : The Rockefeller Institute; 2024 URL: https://rockinst.org/wp-content/uploads/2025/03/Recommended-Best-Practices-for-Media.pdf [accessed 2025-10-28]

XSL•FO
RenderX

68. Betz ME, Harkavy-Friedman J, Dreier FL, Pincus R, Ranney ML. Talking about “firearm injury” and “gun violence”: words matter. Am J Public Health 2021 Dec;111(12):2105-2110. [doi: 10.2105/AJPH.2021.306525] [Medline: 34878863]
69. Pallin R, Siry B, Azrael D, et al. “Hey, let me hold your guns for a while”: a qualitative study of messaging for firearm suicide prevention. Behav Sci Law 2019 May;37(3):259-269. [doi: 10.1002/bsl.2393] [Medline: 30706954]
70. Agree to Agree. 2025. URL: https://agreetoagree.org/about [accessed 2025-10-31]
71. New campaign encourages hispanic parents to store their firearms safely. The Ad Council. 2024. URL: https://www.adcouncil.org/learn-with-us/press-releases/new-campaign-encourages-hispanic-parents-to-store-firearms-safely [accessed 2024-06-11]
72. Vivekaashok-ops/smgv. GitHub. URL: https://github.com/vivekaashok-ops/smgv.git [accessed 2025-11-28]

Abbreviations

MP: members of parliament
MSF: Multiple Streams Framework
REDCap : Research Electronic Data Capture

Edited by A Mavragani; submitted 09.07.25; peer-reviewed by J Beard; revised version received 10.11.25; accepted 11.11.25; published 16.12.25.

Please cite as:
Ashok VA, Vervilles WJK, Kellom KS, Chatterjee A, Ntigbu I, Boms O, Szabo A, Fein JA, Richmond TS, Purtle J, Kearney MD, Meisel ZF
Firearm Violence and Health in Policymaker Discourse: Mixed Methods Social Media Analysis
JMIR Form Res 2025;9:e80397
URL: https://formative.jmir.org/2025/1/e80397
doi:10.2196/80397

© Vivek A Ashok, William J K Vervilles, Katherine S Kellom, Anyun Chatterjee, Isabella Ntigbu, Okechi Boms, Andrea Szabo, Joel A Fein, Therese S Richmond, Jonathan Purtle, Matthew D Kearney, Zachary F Meisel. Originally published in JMIR Formative Research (https://formative.jmir.org), 16.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Quality Assessment of Shock Videos on Video Sharing Platforms: Cross-Sectional Study

Wenxin Wang[1], MMed; Luping Cheng[1], MMed; Xia Hu[1], MMed; Chuanliang Pan[2], MMed

[1]North Sichuan Medical University, Nanchong, China

[2]Chengdu Third People's Hospital, No. 82 Qinglong Street, Qingyang District, Chengdu, Sichuan Province, China

Corresponding Author:
Chuanliang Pan, MMed
Chengdu Third People's Hospital, No. 82 Qinglong Street, Qingyang District, Chengdu, Sichuan Province, China

Abstract

Background: As a highly lethal circulatory failure syndrome, the pathophysiological mechanisms of shock can lead to multiple organ dysfunction syndrome (MODS), which significantly increases the demand for intensive care and the length of hospitalization. There is therefore an urgent need for the public to be informed about health-related issues. In recent years, videos have become a significant medium for health education, and this study aimed to evaluate shock-related videos on video sharing platforms.

Objective: The objective of this study is to identify the top 100 videos related to impact on TikTok, Bilibili, and Xiaohongshu. These videos will then be assessed in terms of their effectiveness and credibility. Following this evaluation, relevant recommendations will be provided.

Methods: The study included a search for videos related to shock on the three video-sharing platforms: TikTok, Bilibili, and Xiaohongshu. The Global Quality Score (GQS) and mDISCERN tools were used to evaluate the credibility and quality of the videos, in addition to employing the Patient Education Materials Evaluation Tool for Audiovisual Content (PEMAT-A/V). Finally, the video was evaluated by examining disease definitions, clinical manifestations, risk factors, assessment, management, and outcomes.

Results: A total of 244 videos (TikTok:87, Bilibili:80, Xiaohongshu:77） were retrieved from the three platforms. The overall video quality was found to be moderately low. The majority of videos were uploaded by health advocates (n=102, 41.8%) and health professionals (n=98, 40.1%). The individual video sources of the GQS were of lower quality (1-3), the mDISCERN scores were moderate (2-4), and the quality of individual users is higher than that of organizational users. The PEMAT A/V scores were as follows: in the overall comprehensibility evaluation, 91% (220) videos of the scores were above 70%; in the actionability evaluation, 65% (157) videos of the scores were below 70%. It should be noted that the actionability scores for different video sources were generally low. In 172 videos (70.4%), the definition of shock and its clinical manifestations were explained in detail, while in 137 videos (56.1%), the definition of shock and its clinical manifestations were also clearly explained. The majority of videos provided a relatively comprehensive explanation of the definition of shock and its clinical signs and symptoms.

Conclusions: Our study have demonstrated that the content and information quality of shock videos is unsatisfactory, as a general rule. This underscores the necessity for pertinent regulatory bodies to oversee the caliber of health-related videos, and for content creators to enhance the quality of their content.

(JMIR Form Res 2025;9:e76715) doi:10.2196/76715

KEYWORDS

quality assessment; shock; video; GQS; mDISCERN; PEMAT A/V; Global Quality Score; Patient Education Materials Evaluation Tool for Audiovisual Content

Introduction

Background

Shock is defined as a systemic hypoxic crisis triggered by a breakdown of circulatory system function. The central paradox is a serious imbalance between tissue oxygen demand and blood oxygen supply capacity. If this is not corrected in time, irreversible organ damage will ensue. The underlying causes of shock are multifactorial in nature, encompassing a wide spectrum of etiologies, including hypovolemic, cardiogenic, obstructive, and distributive causes, which may manifest in isolation or in conjunction [1]. Irrespective of the nature of the shock, its consequences are invariably fatal. For instance, in cases of cardiogenic shock, the heart's capacity to pump blood is significantly compromised, resulting in systemic organ dysfunction. The mortality rate remains high, estimated at approximately 50% [2]. Research indicates that key elements of optimized treatment for cardiogenic shock include early assessment and timely initiation of acute mechanical circulatory

XSL•FO
RenderX

support devices [3], and early diagnosis and treatment of pediatric sepsis and infectious shock can improve patient outcomes [4]. Therefore, early recognition and rapid response are critical for improving patient outcomes. The timely recognition of shock and subsequent prompt emergency treatment (including on-site resuscitation, rapid hospitalization, and other related measures) is contingent upon public health awareness. It is imperative to emphasize that enhancing the public's awareness of shock-related health issues is of great significance.

In recent years, videos have attracted mounting interest as a medium for health information dissemination. Empirical evidence has demonstrated the efficacy of brief audiovisual presentations in enhancing viewers' health knowledge and awareness, with a particular emphasis on the popularization of first aid knowledge [5]. The rapid dissemination and high acceptance of videos renders them a significant source of health information, particularly in the context of disseminating knowledge about complex diseases such as shock [6]. The dissemination of such information has the potential to enhance public awareness of diseases, facilitate the timely treatment of patients, and improve patient outcomes. Such studies can provide a valuable reference point for enhancing public awareness of shock and offer data-driven support for the development of suitable health education strategies [5,7]. Nevertheless, it is imperative to direct particular attention to the quality of the information presented in these videos. The increasing prevalence of video content has the capacity to disseminate health-related health information, which may consequently mislead patients in their disease management [8,9]. While videos may facilitate timely access to medical information for the general population, they may also result in delays to treatment. Consequently, it is imperative to evaluate the quality of shock-related videos. However, to the best of the present author's knowledge, there are currently no studies evaluating the quality of shock-related short videos. The objective of this study is to assess the health information in videos related to shock, provide data support for existing health information, and formulate relevant recommendations.

The approach adopted in this study is advantageous in that it facilitates the rapid collection of a substantial amount of data, enables efficient statistical analysis, and ensures the reliability and validity of the results [10]. The present study will analyze the content of videos with a view to revealing their role and existing issues in the dissemination of health information related to shock. The study will provide a scientific basis for the effective use of short videos in public health education. Key issues that will be examined include the accuracy of information in short videos, their professionalism, and their influence on viewers' behavior [11,12]. Finally, the relevant results will be analyzed, and appropriate recommendations will be proposed.

Objective

TikTok, Bilibili, and Xiaohongshu have demonstrated significant efficacy in disseminating educational video content in prior studies [13,14], which informed our selection of these platforms for the current research. The objective of this study is to identify and collect the top 100 videos related to shock on the TikTok, Bilibili, and Xiaohongshu platforms. These videos will then be assessed in terms of their effectiveness and credibility, and recommendations will be provided based on the findings.

Methods

Ethical Considerations

This project's research content is publicly disseminated video content on public network platforms, and it does not involve the collection or use of any personal privacy information. The research process did not involve any intervention, investigation or potential risk to individuals, nor did it involve the use of data relating to human subjects or patients, so it did not affect any individuals. Consequently, this study is not subject to ethical review.

Collection of Data

In order to guarantee the objectivity of data collection, the research team strictly implemented the following standardized preprocessing procedures prior to implementing the retrieval operation. In order to rectify the issue, the following three steps must be taken: (1) the factory settings of all mobile terminals must be restored; (2) the browsing history and cache data of the devices must be completely cleared; and (3) all the platform accounts must be logged out of. In order to circumvent the possibility of potential impact on the results, the study proceeded to retrieve the top 100 videos from TikTok, Bilibili, and Xiaohongshu by conducting a search for the Chinese keyword "休克." TikTok and Xiaohongshu have opted for comprehensive sorting, while Bilibili has opted for default sorting. The entire search was conducted over the course of seven days, from 5th March to 11th March 2025. The videos were then categorized by three qualified doctors with extensive experience working in tertiary care hospitals in the Department of Critical Care Medicine. The integrity of the video is based on classification, with preliminary decisions made by three clinicians. Any discrepancies are resolved through consensus among the three clinicians. The consistency scores (Cohen κ) for GQS, mDISCERN, PEMAT-U, and PEMAT-A, completed by two doctors, were 0.766, 0.627, 0.839, and 0.802, respectively. These results indicate that GQS and mDISCERN have strong consistency, while PEMAT-U and PEMAT-A have very strong consistency. The following attributes are contained in each video data set: publisher's profile (comprising account nicknames, personal profiles and official authentication logos); the content release time, the current fan base, the video duration information, and the user interaction metrics (including the number of likes, favorites, share spreads and comments). The data were recorded using Microsoft Excel.

Inclusion and Exclusion Criteria

Videos that provide knowledge about shock-related topics. The exclusion criteria were as follows: (1) videos with repetitive content; (2) videos containing advertisements or commercials; and (3) videos unrelated to the topic.

Assessment Methodology and Process

The video publishing body is divided into two major categories : individual users and institutional accounts. In accordance with

XSL•FO
RenderX

the findings of preceding studies by Cuiet al [15] and Liu et al [16], the following professional evaluation system is used for multidimensional quality testing. The credibility of the videos was evaluated using the modified DISCERN criteria. The information quality of the videos was evaluated using the Global Quality Score (GQS) criteria. The videos were evaluated using the Patient Education Materials Evaluation Tool for Audiovisual Content (PEMAT-A/V). The mDISCERN tool was originally developed for the purpose of evaluating the quality of written health information concerning treatment regimens [17]. The mDISCERN tool from DISCERN was used to analyze the credibility and quality of the video. The efficacy of this approach has been demonstrated in the evaluation of health-related video material on diverse platforms, including *Helicobacter pylori*, tachycardia, thyroid cancer, and other disorders in assessing video use [18-20]. The mDISCERN tool is composed of five inquiries: (1) Are the objectives clear and realized?; (2) Were reliable sources of information used?; (3) Was the information presented balanced and fair?; (4) Are alternative sources of information listed? Are other sources of information listed for patient reference? (5) Are areas of uncertainty mentioned? The GQS is another widely used tool for assessing the quality of health information in videos, with response selection based on a 5-point scale from 1 (poor quality) to 5 (good quality) [21,22]. The PEMAT-A/V [23] was used to evaluate the comprehensibility and actionability of the video material. The PEMAT-A/V comprises 17 questions; 13 assess the patient's comprehension of the health information and 4 evaluate the actionability of the recommendations in the video.

Finally, the video was evaluated by examining disease definitions, clinical signs and symptoms, risk factors, assessment, management, and outcomes. The allocation of scores was categorized into three primary sections: no content (0 points), content (1 point), and more content (2 points). Consequently, the content, reliability, and quality of the videos were assessed by two clinicians based on the mDISCERN, GQS scoring systems, and PEMAT-A/V. Continuous variables are expressed as mean and SD or median and IQR. Categorical variables were expressed as numbers with corresponding percentages. Prior to the evaluation, the two clinicians familiarized themselves with the formal descriptions of the utilization tools and discussed the use of the tools to assess the content and quality of the videos.

Data Analysis

Data were analyzed using SPSS (version 30.0; IBM Corp) for statistical correlation and analysis. Data are presented as numbers (percentage), mean (SD), or median (IQR). Kolmogorov-Smirnov test (n>50) or Shapiro-Wilk test (n≤50) were used to assess whether the data followed a normal distribution. The Mann-Whitney *U* test was used to compare two groups of data. The Kruskal-Wallis H test is used to compare three or more groups of data, with Dunn test subsequently used as post-hoc test for pairwise comparisons. The interrater reliability of the two evaluators' scores was assessed using Cohen χ test.

Results

A total of 244 videos (TikTok:87, Bilibili:80, Xiaohongshu:77）were retrieved from the three platforms and included in the subsequent data analysis (Figure 1). The categorization of video sources according to account information from the three platforms revealed two categories: individual users and institutional accounts. A further categorization of individual users is possible, distinguishing between health professionals, health advocates, and patients. A similar classification system was applied to institutional accounts, which were divided into three categories: news organizations, nonprofit organizations, and for-profit organizations. The results of the video analysis (Table 1) demonstrated that health professionals were the most prolific contributors, uploading a total of 102 videos (41.8%).

In the runner-up position were health advocates a total of 98 videos (40.1%), and patients, who uploaded a total of 12 videos (4.9%). Amongst institutional accounts, nonprofit organizations uploaded the most videos related to shock, with a total of 18 videos (7.4%). Subsequent to this, for-profit organizations uploaded 10 videos (4.1%) , while news organizations uploaded 4 videos (1.7%).

XSL•FO
RenderX

Figure 1. Patient flowchart.

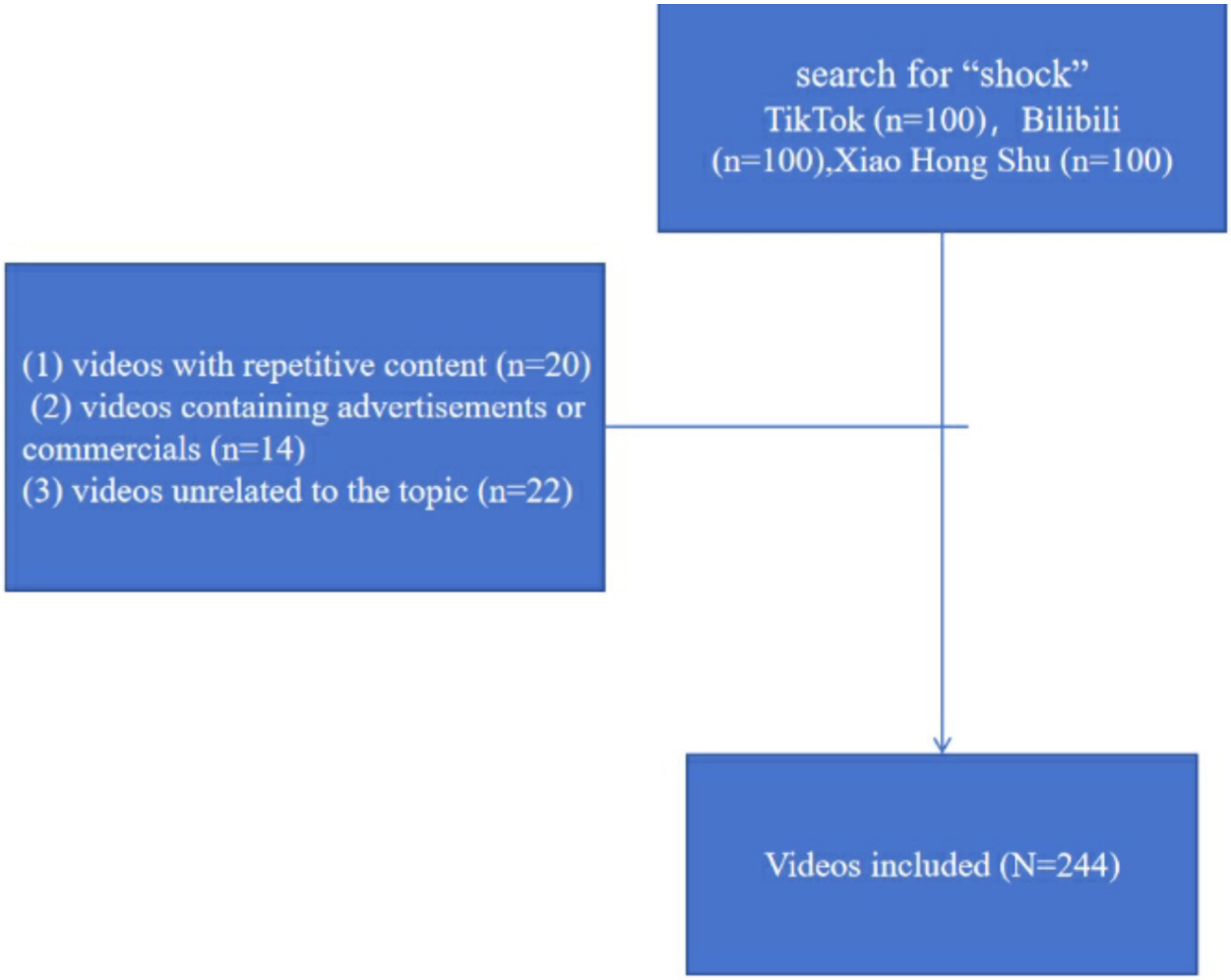

Search strategy for videos on shock. TikTok excluded 5 videos with repetitive content, 3 videos containing advertisements or commercials, and 5 videos unrelated to the topic. Bilibili excluded 6 videos with repetitive content, 4 videos containing advertisements or commercials, and 10 videos unrelated to the topic. Xiaohongshu excluded 9 videos with repetitive content, 7 videos containing advertisements or commercials, and 7 videos unrelated to the topic.

Table . Descriptions of video sources.[a]

Source type	Description	Video, n (%)
Health professionals	Individuals who have received specialized education in healthcare (eg, doctors, nurses)	41.8 (102)
Heath advocate	Transform knowledge that one understands personally into knowledge that can be comprehended by the general public.	40.1 (98)
Patient	Have suffered from the corresponding disease before	4.9 (12)
Nonprofit organizations	Organizations that exist to serve a specific charitable, educational purpose (eg, hospitals, colleges)	7.4 (18)
For-profit organizations	Companies that operate to generate income	4.1 (10)
News organizations	Organizations whose main purpose is to gather and distribute news and information (TV account, newspaper)	1.7 (4)

[a]The following section provides a detailed explanation of each video source category, along with their respective percentage shares.

As demonstrated in Table 2, the duration of videos uploaded on video sharing platforms varied significantly across the sample. The median length of upload dates for health advocates was the longest at 562 days, followed by news organizations at

XSL•FO
RenderX

501 days, and the shortest for non-profit organizations at 146 days. The median duration of videos was the longest for patients at 146 seconds, followed by non-profit organizations at 103.5 seconds, and the shortest for health advocates at 40.5 seconds. The most prevalent engagement patterns were observed among news organizations, with 'liking', 'favouriting' and commenting. In contrast,, for-profit organizations demonstrated the least engagement, with the least popular patterns being 'liking', 'favouriting' and commenting. Health professionals had the highest number of followers, with a median value of 6654. Subsequent to this, news organizations followed, patients and health advocates, with 2173, 2158, and 2060 followers, respectively.

Table . Video characteristics.

Source type	Account fans	Length	Video likes	Collections	Video comment	Days on platform
Health professionals, median (IQR)	6654（613 to 135200）	76（7 to −3648）	110（0 to 8700）	30（0 to 31280）	6（1 to 3497）	299（8 to 1650）
Heath advocate, median (IQR)	2060（219 to 9658）	40.5（8 to 7020）	35（0 to 414000）	61（0 to 15000）	2（0 to 10000）	562（1 to 8654）
Patient, median (IQR)	2158（180 to 8983）	146（12 to 255）	29（1 to 4706）	6（1 to 645）	5（0 to 1341）	170（1 to 770）
Nonprofit organizations, median (IQR)	1624（490 to 16750）	65.5（10 to −3628）	55（0 to 3648）	47（0 to 2089）	1.5（0 to 41）	199（46 to 1577）
For-profit organizations, median (IQR)	700（190 to 7207）	103.5（8 to 356.5）	10（2 to 52）	7（0 to 106）	0（0 to 10）	146（3 to 947）
News organizations, median (IQR)	2173（341 to 12,405,804）	97（7 to 1500）	4069（0 to 38,000）	533（3 to 2364）	180（44 to 2535）	501（329 to −1430）

Differences between groups were analyzed based on GQS and mDISCERN scores. Figures A and D were analyzed using the Mann-Whitney U test. Figures B, C, E, and F were analyzed using the Kruskal-Wallis H test, followed by Dunn's test subsequently used for post-hoc comparisons with $P<.05$.

As illustrated in Table 3 and Figure 2, the GQS scores for all video sources are relatively low, with health advocates demonstrating the highest quality, health professionals the second-highest, and news organizations the lowest. The mDISCERN scores were moderate; with health advocates exhibiting the highest credibility; health professionals, the second-highest; and news organizations the lowest. Moreover, the mDISCERN frequency distribution indicates that patients, news organizations, nonprofit organizations, and for-profit organizations have low credibility. Figure 3 compares the reliability and quality of information from different video sources. A subsequent comparison of individual users with organizational users revealed that both GQS and mDISCERN scores were higher for individual users than for organizational users ($P<.05$). Within the individual user category, health advocates and health professionals demonstrated higher information quality, while patients exhibited lower information quality ($P<.05$). However, no statistically significant differences were observed within the organizational category. The overall quality level of mDISCERN is moderate, while the quality level of news organizations is lower. Within the individual user category, health advocates and health professionals were found to have higher credibility than patients ($P<.05$), while no statistically significant differences were observed within the organizational user category.

Table . GQS[a] and mDISCERN scores. Mean (SD) and median (IQR) values of GQS and mDISCERN scores based on video classification.

Source type	GQS		mDISCERN scores	
	Mean (SD)	Median (IQR）	Mean (SD)	Median (IQR)
Health professionals	2.65 (0.79)	3 (2 - 3)	3.56 (0.76)	4 (3 - 4)
Heath advocates	2.87 (0.70)	3 (2.75 - 3)	3.87 (0.70)	4 (3 - 4)
Patients	2.25 (0.87)	2.5 (1.25 - 3)	3.17 (0.84)	3 (2.25 - 4)
Non-profit organizations	2.44 (0.92)	3 (1.75 - 3)	3.5 (0.99)	4 (2.75 - 4)
For-profit organizations	2.2 (0.79)	2 (1.75 - 3)	3.2 (0.79)	3 (2.75 - 4)
News organizations	1.75 (0.96)	1.5 (1 - 2.75)	2.75 (0.96)	2.5 (2 - 3.75)

[a]GQS: Global Quality Score.

XSL•FO
RenderX

Figure 2. Frequency distribution of GQS and mDISCERN scores. GQS: Global Quality Score.

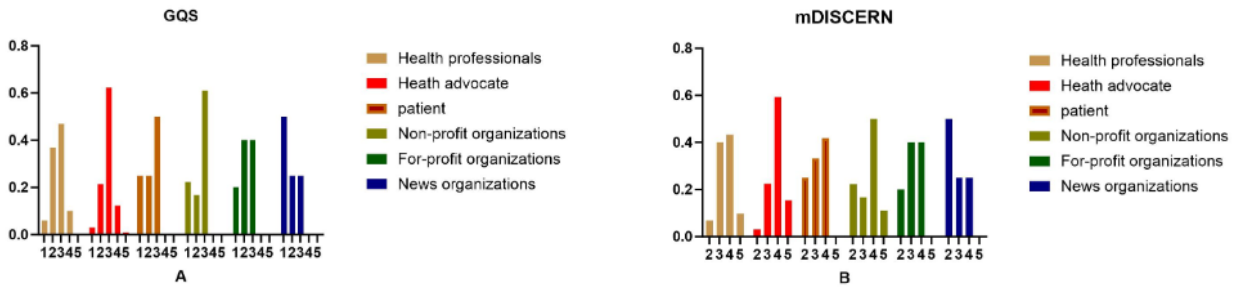

Figure 3. Comparison between mDISCERN scores and global quality scores (GQS) across video sources.

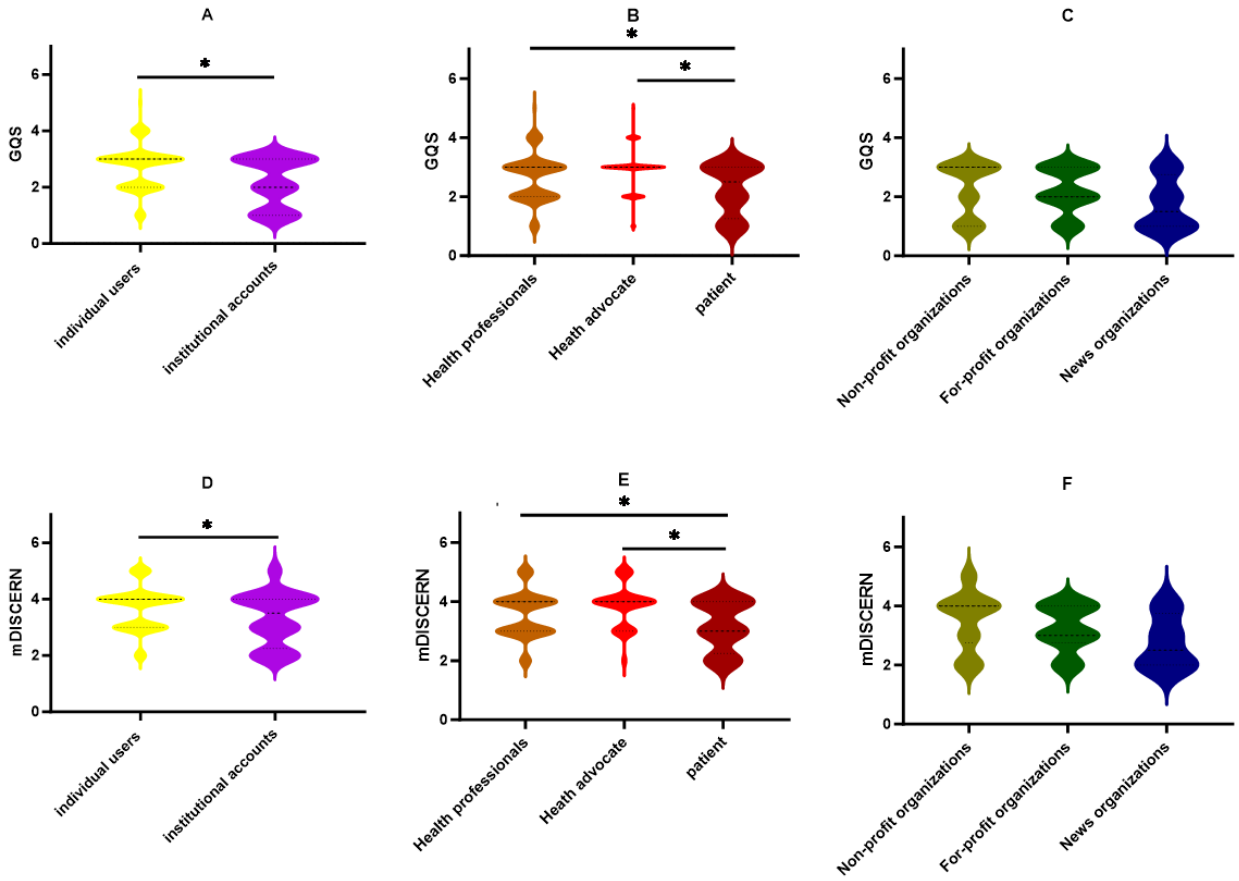

As demonstrated in Table 4, a considerable proportion of the PEMATA/V scores surpassed the overall understandability threshold of 70%, with 91% (n=220) videos of the scores falling into this category. Furthermore, 65% (n=159) videos of the scores did not exceed 70% in terms of actionability, while 73% (n=178) videos of the scores exceeded 70% at the overall level. As demonstrated in Figure 4, there is no statistically significant difference between individuals and organizations in terms of understandability scores. However, among individual users, health professionals and health advocates demonstrated the highest levels of understanding, in comparison to patients (P<.05). The study revealed no statistically significant differences in actionability scores between individuals and organizations, nor within each group.

XSL•FO
RenderX

Table . PEMAT scores presented in the video.

Scale, score	Rate n (%)
PEMAT-U[a]	
0 - 60	22 (9)
61 - 80	66 (27)
81 - 100	156 (64)
PEMAT-A[b]	
0 - 50	86 (35)
67 - 75	141 (58)
100	17 (7)
PEMAT-T[c]	
0 - 60	42 (17)
61 - 80	156 (64)
81 - 100	46 (19)

[ab] PEMAT-U: Patient Education Materials Assessment Tool-Understandability

[b]PEMAT-A: Patient Education Materials Assessment Tool for Audio Visual Materials.

[c]PEMAT-T: Patient Education Materials Assessment Tool for Telehealth.

Figure 4. Comparison between PEMAT-U scores and PEMAT-A scores (GQS) across video sources. GQS: Global Quality Score; PEMAT-A: Patient Education Materials Assessment Tool for Audio Visual Materials; PEMAT-U: Patient Education Materials Assessment Tool-Understandability.

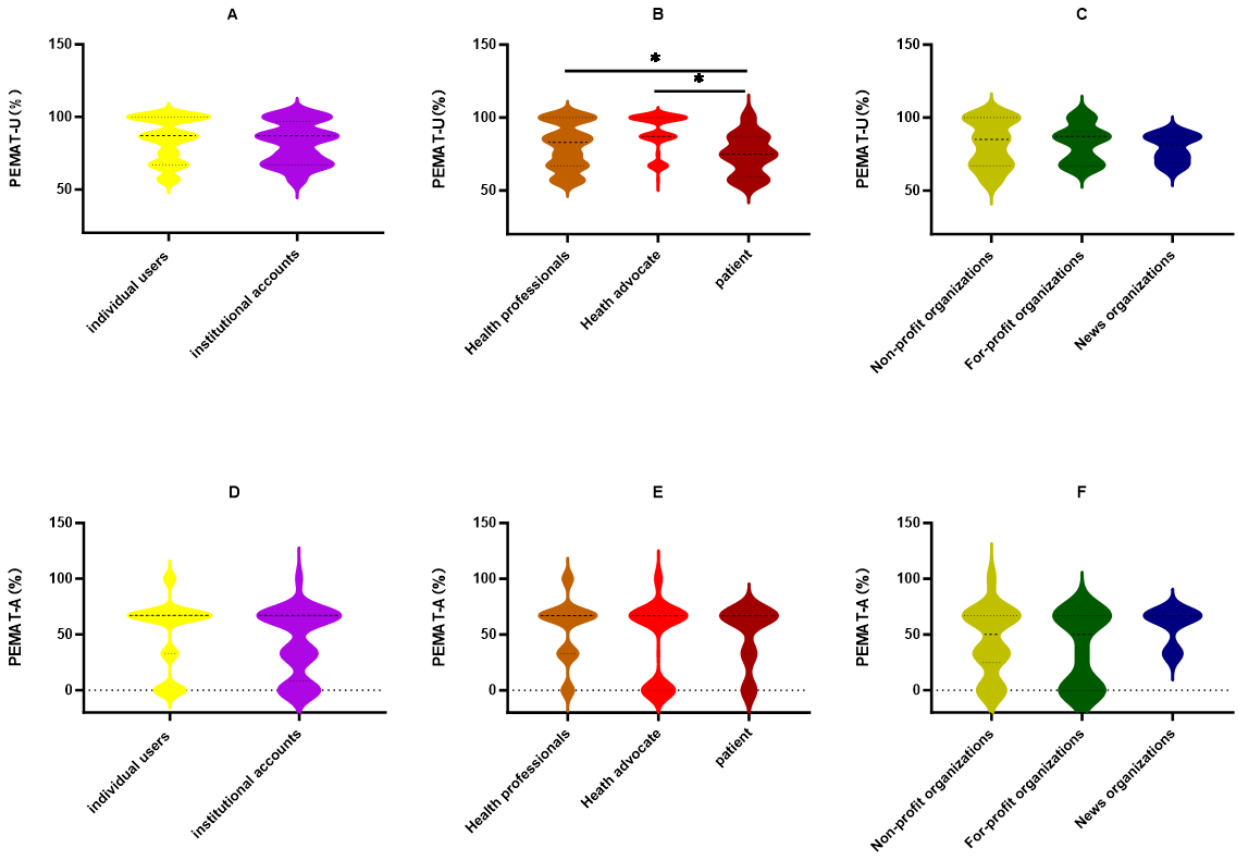

Differences between groups were analyzed based on PEMAT-U scores and PEMAT-A Scores. (Figures 4A and 4D) were analyzed using the Mann-Whitney U test. Figures (Figure 4B, 4C, 4E, 4F) were analyzed using the Kruskal-Wallis H test, followed by Dunn's test subsequently used for post-hoc comparisons (P<.05).

XSL•FO
RenderX

The frequency of myopic shock, as outlined in Table 5, encompasses six fundamental domains: disease definition, clinical manifestations, risk factors, assessment, management, and outcome. The results demonstrated that the definition of shock and its clinical manifestation were clearly explained in 172 videos (70.4%) and 137 videos (56.1%), respectively, while a few videos (26%, 10.7% and 12, 4.9%) had no content. The analysis revealed that approximately half of the videos contained content on risk factors (n=146, 59.8%), management (n=122, 50%), and assessment (n=142, 58.1%). In contrast, approximately half of the videos (n=124, 50.8%) did not include content on outcomes.

Table . Completeness of video content. The six aspects of the video content are indicated by percentages.

Content	Definition, N (%)	Clinical manifestation, n (%)	Risk factors, n (%)	Evaluation, n (%)	Management, n (%)	Outcomes, n (%)
No content	26 (10.7)	12 (4.9)	86 (35.2)	95 (38.9)	112 (45.9)	124 (50.8)
Some content	46(18.9)	95 (38.9)	146 (59.8)	142 (58.1)	122 (50.0)	109 (42.2)
Extensive content	172 (70.4)	137 (56.1)	12 (4.9)	7 (2.9)	10 (4.1)	11 (4.5)

Discussion

Principal Findings

The physiological impact of shock on patients is extremely severe, leading to multiple organ failure, and the economic burden, including hospitalization and long-term rehabilitation costs, significantly increases the strain on the healthcare system [24]. Despite the advances in the diagnosis and management of shock, including pharmacological treatment, fluid resuscitation, and surgical intervention, there are still deficiencies in the early recognition and rapid response, resulting in a poor prognosis for patients [25]. Research also indicates that early diagnosis can improve patient outcomes [7,8]. Early diagnosis and rapid treatment are critical, but timely identification of shock and swift implementation of measures depend on public health awareness, with short videos playing a key role as an important source of health information [8,9]. Consequently, the study of health information dissemination for shock, particularly the use of short videos, is of particular importance in enhancing public awareness of shock [26]. This study used an observational review design with the aim of examining the current status of videos as a shock health information resource and assessing their effectiveness and potential impact in public health education. The content and quality of videos were analyzed to understand the role and problems of videos in health information dissemination.

The current user scale of video applications shows a rapid growth trend. This emerging mode of communication in the field of health knowledge popularization shows a unique advantage: video will transform complex medical concepts into dynamic images, three-dimensional demonstrations and scenario-based explanations. The convenience for users to obtain health information has been significantly improved. While this medium has undoubtedly facilitated the dissemination of health-related knowledge, it also increases the risk of the public obtaining incorrect health information [8,9]. To date, no studies have evaluated videos related to shock, the necessity for rigorous evaluation through relevant studies is therefore paramount. The impact of shock-related videos on social media platforms such as TikTok, BiliBili, and Xiaohongshu remains to be elucidated. This study comprehensively assessed 244 shock videos on TikTok, BiliBili, and Xiaohongshu, using rigorous content and quality evaluation methodologies.

The findings of the present study demonstrated that these 244 videos amassed a total of 129,664,761 followers, 879,855 likes, 127,345 collections, and 35,226 comments. Video sharing platform, represented by TikTok (Shake), Bilibili and Xiaohongshu, have transformed professional content into visualizations and become an important channel for accessing information. The results of the study indicate that health professionals and health advocates dominate the content, a phenomenon that differs from previous studies [11,27], some videos uploaded from patients, news media, nonprofits, and for-profit organizations, suggesting that people are willing to share their health knowledge on the Internet. Despite the fact that health professionals boast a greater number of followers than news organizations, news outlets receive a greater number of likes, saves and comments, and are more disseminative. This finding is consistent with the conclusions of Cui et al [15] and Jun Qiu et al [27]. It is therefore vital that health professionals enhance their ability to promote health knowledge. However, the study's findings reveal that the median upload duration for the majority of videos exceeded 100 days, with some sources displaying an upload duration of over 500 days. This suggests a tendency for videos pertaining to shock to undergo less frequent updates. The dissemination of video recordings that have been in circulation for an extended period may result in the dissemination of outdated knowledge, thereby necessitating the prompt updating of such content by the public. The platform has the capacity to provide relevant incentives to encourage individuals to disseminate health knowledge related to shock.

According to previous studies by Cui et al and Liu et al [15,16], we selected mDISCERN, GQS, and PEMAT-A/V to evaluate video quality. The Mann-Whitney U test is used to compare two sets of data, while the Kruskal-Wallis test is used to compare three or more sets of data. The Dunnet multiple comparison test is used for two-way comparisons between two groups.

Initially, the credibility of the video content was assessed using mDISCERN. The majority of video sources exhibited mDISCERN scores that approximated the medium quality threshold. The credibility of individual users was found to exceed that of organizational accounts. Additionally, within the cohort of individual users, health professionals and health advocates were found to possess greater credibility than patients. This finding is consistent with the observations reported by Lai

XSL•FO
RenderX

Y [19]. It is recommended that the platform enhance its content review process for patient and organizational users. This is attributable to the capacity for non-professionals to disseminate misinformation, which may occasion delays in treatment.

The GQS validation tool was utilized to assess the quality of the videos. It was found that individual users exhibited higher quality than organizational accounts. Individual users have been found to demonstrate higher quality than organizational accounts. Furthermore, among individual users, health professionals and health advocates have been found to demonstrate higher quality than patients. However, the mean score for the majority of videos is 3 or below, indicating that video quality is at a moderately low level, consistent with the findings of Collà Ruvolo et al [28]. This finding suggests a need for further development and refinement of trauma-related videos to enhance their effectiveness. Based on our analysis of the videos, we have identified two potential issues: (1) health professionals and health advocates use technical terms that are difficult for the general public to understand and (2) Patients and individuals lacking professional knowledge may produce videos containing inaccurate information. Therefore, health professionals and health advocates must share their knowledge in a way that is easier for the general public to understand and access. Additionally, relevant platforms must conduct more rigorous content reviews to ensure the public can access accurate information.

In conclusion, the third instrument utilized was PEMAT-A/V, which was used to evaluate the educational pertinence of the videos. As posited by Shoemaker et al [23], the definition of video content as both understandable and actionable was proposed. Furthermore, it was indicated that a PEMAT score in excess of 70% would be indicative of a professional video. This finding demonstrated that shock-related videos were comprehensible but not directly applicable. The two groups of " health professionals" and "health advocates" demonstrated a significantly higher level of understanding of the information compared to patients, consistent with previous research [18]. The study's findings suggest that the comprehensibility of the information meets established standards; however, the actionable content is found to be deficient. As demonstrated by Kanner et al [29], actionable content is relatively restricted. Morra et al [30] also noted that the study topic may lead to differences in related scores. The relatively low GQS scores mentioned above may also be related to the actionable content of the reviewed videos. It is evident that a significant proportion of videos offer merely a cursory introduction to the subject, with a paucity of actionable content. This is exemplified by the absence of emergency procedures, such as lying down, elevating the legs, maintaining body temperature, monitoring breathing, and performing cardiopulmonary resuscitation when necessary. It is imperative to enhance the practicality of these systems, a necessity that is especially pronounced in trauma-related scenarios where immediate attention is paramount. The practicality of video content from various sources should be enhanced, for example by incorporating demonstrations by health professionals and health advocates.

The study results indicate that the majority of videos (172, accounting for 70.4%; 137, accounting for 56.1%) provided clear explanations of the definition of shock and its clinical manifestations, covering all aspects of the disease, which is consistent with un Qiu et al [27]. The dissemination of knowledge regarding the mechanisms of shock is predominantly facilitated by the diversification of science communication channels, with online media assuming a particularly prominent role in this regard. Nevertheless, the dissemination of knowledge pertaining to risk factors, management, and outcomes continues to necessitate ongoing enhancement. Platforms have the capacity to engage in collaborative endeavors with pertinent professionals, thereby facilitating the dissemination of such knowledge.

The present study is subject to several limitations. First, it is important to note that this is a cross-sectional study, which is limited in its ability to reflect results at a single point in time. This inherent limitation makes it difficult to replicate and unable to conduct dynamic analysis modeling of user interaction behavior. Furthermore, it cannot perform causal analysis. Second, there is a possibility of bias in the classification of videos and diseases across the six dimensions examined in this study. Third, the present study systematically examined the quality of content themed around 'shock' on video sharing platform such as TikTok, Bilibili, and Xiaohongshu. However, the review mechanisms of these platforms differ, which may have influenced the study results. Fourth, due to the limitations of the platform itself, it is impossible to obtain accurate viewing figures for videos, which may result in the videos included not being sufficiently disseminated, thereby affecting the results. Fifth, owing to defects inherent in the platform, a significant proportion of user identities are incomplete, thus impeding the execution of subgroup analyses, including those pertaining to the classification of doctors.

Conclusion

The present study analyzed a total of 244 videos (TikTok:87, Bilibili:80, and Xiaohongshu:77) related to shock, obtained from the three video sharing platforms. The analysis indicated that the content and information quality of the videos was, overall, not satisfactory. The quality of the video was found to be contingent upon the source of the videos in question. With regard to the reliability and quality of the videos, the study found that overall, these were moderately low, although videos from health advocates and health professionals ranked first and second, respectively. This highlights the necessity for the first relevant regulatory authority to oversee the quality of health-related short videos, thereby assisting the public in achieving early prevention and treatment of diseases. The majority of videos fail to incorporate actionable elements, and it is imperative that they are improved in terms of their presentation of content in order to enhance their actionability. In addition, health professionals and health advocates should intensify their efforts in dissemination and enhance the actionability of the content.

XSL•FO
RenderX

Acknowledgments

We extend our gratitude to all contributing researchers and relevant funding bodies for their support of this article.

Funding

Funding from the Sichuan Provincial Natural Science Foundation is gratefully acknowledged (Project Number: 2024NSFSC0488).

Authors' Contributions

WW contributed to the design and implementation of the research; CP and WW analyzed the results; WW, LC, and HX wrote the manuscript; LC conceived and supervised the project.

Conflicts of Interest

None declared.

References

1. Leech C, Turner J. Shock in trauma. Emerg Med Clin North Am 2023 Feb;41(1):1-17. [doi: 10.1016/j.emc.2022.09.007]
2. Dumont R, Tridetti J, Ancion A, Maréchal P, Lancellotti P. Cardiogenic shock : etiology and management. Rev Med Liege 2021 Feb;76(2):88-92. [Medline: 33543853]
3. Sarma D, Jentzer JC. Cardiogenic shock: pathogenesis, classification, and management. Crit Care Clin 2024 Jan;40(1):37-56. [doi: 10.1016/j.ccc.2023.05.001] [Medline: 37973356]
4. Hilarius KWE, Skippen PW, Kissoon N. Early recognition and emergency treatment of sepsis and septic shock in children. Pediatr Emerg Care 2020 Feb;36(2):101-106. [doi: 10.1097/PEC.0000000000002043] [Medline: 32011568]
5. Puchała S, Kasprzak W, Piwowarski P. Human interaction classification in sliding video windows using skeleton data tracking and feature extraction. Sensors (Basel) 2023 Jul 10;23(14):6279. [doi: 10.3390/s23146279] [Medline: 37514573]
6. Zhang J, Han P, Tang Y, Xi W, Xiao X, Yang F. Popular science and education of cosmetic surgery in China: quality and reliability evaluation of Douyin short videos. Health Expect 2023 Jun;26(3):1221-1226. [doi: 10.1111/hex.13737] [Medline: 36807969]
7. Sun Y, Liu X, Zhang X, Xu Q, Li A. Health information analysis of cryptorchidism-related short videos: analyzing quality and reliability. Digit Health 2025;11:20552076251317578. [doi: 10.1177/20552076251317578] [Medline: 39877853]
8. Yeung A, Ng E, Abi-Jaoude E. TikTok and attention-deficit/hyperactivity disorder: a cross-sectional study of social media content quality. Can J Psychiatry 2022 Dec;67(12):899-906. [doi: 10.1177/07067437221082854] [Medline: 35196157]
9. Yeung AWK, Tosevska A, Klager E, et al. Medical and health-related misinformation on social media: bibliometric study of the scientific literature. J Med Internet Res 2022 Jan 25;24(1):e28152. [doi: 10.2196/28152] [Medline: 34951864]
10. Journault AA, Cernik R, Charbonneau S, et al. Learning to embrace one's stress: the selective effects of short videos on youth's stress mindsets. Anxiety, Stress, & Coping 2024 Jan 2;37(1):29-44. [doi: 10.1080/10615806.2023.2234309]
11. Song J, Liu J. A data-driven short video international communication model based on indicator system communication network and attention BiLSTM neural network. Sci Rep 2024;14(1):14532. [doi: 10.1038/s41598-024-65098-x]
12. Liao M. Analysis of the causes, psychological mechanisms, and coping strategies of short video addiction in China. Front Psychol 2024;15:1391204. [doi: 10.3389/fpsyg.2024.1391204] [Medline: 39165759]
13. Zheng S, Tong X, Wan D, Hu C, Hu Q, Ke Q. Quality and reliability of liver cancer–related short Chinese videos on TikTok and Bilibili: cross-sectional content analysis study. J Med Internet Res 2023 Jul 5;25:e47210. [doi: 10.2196/47210]
14. Guan JL, Xia SH, Zhao K, et al. Videos in short-video sharing platforms as sources of information on colorectal polyps: cross-sectional content analysis study. J Med Internet Res 2024 Oct 29;26:e51655. [doi: 10.2196/51655] [Medline: 39470708]
15. Cui N, Lu Y, Cao Y, Chen X, Fu S, Su Q. Quality assessment of TikTok as a source of information about mitral valve regurgitation in China: cross-sectional study. J Med Internet Res 2024 Aug 20;26:e55403. [doi: 10.2196/55403] [Medline: 39163110]
16. Liu Z, Chen Y, Lin Y, et al. YouTube/ Bilibili/ TikTok videos as sources of medical information on laryngeal carcinoma: cross-sectional content analysis study. BMC Public Health 2024 Jun 14;24(1):1594. [doi: 10.1186/s12889-024-19077-6]
17. Charnock D, Shepperd S, Needham G, Gann R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. J Epidemiol Community Health 1999 Feb;53(2):105-111. [doi: 10.1136/jech.53.2.105] [Medline: 10396471]
18. Lai Y, Liao F, He Z, et al. The status quo of short videos as a health information source of Helicobacter pylori: a cross-sectional study. Front Public Health 2024 Jan 8;11:1344212. [doi: 10.3389/fpubh.2023.1344212]
19. Li Z, Yan C, Lyu X, Li F, Zeng R. Assessing quality and reliability of online videos on tachycardia: a YouTube video-based study. BMC Public Health 2024 Sep 27;24(1). [doi: 10.1186/s12889-024-20062-2]
20. Yang S, Zhan J, Xu X. Is TikTok a high-quality source of information on thyroid cancer? Endocrine 2023 Aug;81(2):270-276. [doi: 10.1007/s12020-023-03332-8] [Medline: 36840912]

XSL•FO
RenderX

21. He Z, Wang Z, Song Y, et al. The reliability and quality of short videos as a source of dietary guidance for inflammatory bowel disease: cross-sectional study. J Med Internet Res 2023;25:e41518. [doi: 10.2196/41518]
22. Mueller SM, Jungo P, Cajacob L, Schwegler S, Itin P, Brandt O. The absence of evidence is evidence of non-sense: cross-sectional study on the quality of psoriasis-related videos on YouTube and their reception by health seekers. J Med Internet Res 2019 Jan 16;21(1):e11935. [doi: 10.2196/11935]
23. Shoemaker SJ, Wolf MS, Brach C. Development of the Patient Education Materials Assessment Tool (PEMAT): a new measure of understandability and actionability for print and audiovisual patient information. Patient Educ Couns 2014 Sep;96(3):395-403. [doi: 10.1016/j.pec.2014.05.027] [Medline: 24973195]
24. Auger N, Carrier FM, Waechter J, et al. Long-term outcomes of patients with toxic shock syndrome: a matched cohort study. J Infect 2024 Aug;89(2):106213. [doi: 10.1016/j.jinf.2024.106213] [Medline: 38944286]
25. Ali Baig S, Malhotra K, Banerjee AJ, et al. Assessment of the quality, content, and reliability of YouTube videos on diabetes mellitus and polycystic ovary syndrome: a systematic review with cross-sectional analysis comparing peer-reviewed videos. Endocr Connect 2024 Jul 1;13(7):e240059. [doi: 10.1530/EC-24-0059] [Medline: 38856005]
26. Dong Z, Xie T. How to know who you are through your short video selfies?-Capturing personality via short video selfies. Front Psychol 2023;14:1072344. [doi: 10.3389/fpsyg.2023.1072344] [Medline: 36949915]
27. Qiu J, Zhou YL. Quality assessment of heatstroke videos on TikTok. Front Public Health 2024;12:1446003. [doi: 10.3389/fpubh.2024.1446003] [Medline: 39296850]
28. Collà Ruvolo C, Califano G, Tuccillo A, et al. YouTube as a source of information on placenta accreta: a quality analysis. European Journal of Obstetrics & Gynecology and Reproductive Biology 2022 May;272:82-87. [doi: 10.1016/j.ejogrb.2022.03.015]
29. Kanner J, Waghmarae S, Nemirovsky A, Wang S, Loeb S, Malik R. TikTok and YouTube videos on overactive bladder exhibit poor quality and diversity. Urol Pract 2023 Sep;10(5):493-500. [doi: 10.1097/UPJ.0000000000000423] [Medline: 37347790]
30. Morra S, Collà Ruvolo C, Napolitano L, et al. YouTube as a source of information on bladder pain syndrome: a contemporary analysis. Neurourol Urodyn 2022 Jan;41(1):237-245. [doi: 10.1002/nau.24802] [Medline: 34559920]

Abbreviations

GQS: Global Quality Score
MODS: multiple organ dysfunction syndrome
PEMAT-A/V: Patient Education Materials Evaluation Tool for Audiovisual Content
PEMAT-U: Patient Education Materials Assessment Tool-Understandability

Edited by J Sarvestan; submitted 29.04.25; peer-reviewed by Z Liu; revised version received 10.10.25; accepted 13.10.25; published 30.12.25.

Please cite as:
Wang W, Cheng L, Hu X, Pan C
Quality Assessment of Shock Videos on Video Sharing Platforms: Cross-Sectional Study
JMIR Form Res 2025;9:e76715
URL: https://formative.jmir.org/2025/1/e76715
doi:10.2196/76715

© Wenxin Wang, Luping Cheng, Xia Hu, Chuanliang Pan. Originally published in JMIR Formative Research (https://formative.jmir.org), 30.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Didactic and Content Quality of Basic Life Support Videos on YouTube: Cross-Sectional Study

Jasmina Sterz[1], Dr Med; Yvonne Beaugé[1], PhD; Pia Tueckmantel[1], Dr Med; Lena Bepler[1], Dr Med; Armin N Flinspach[2], Dr Med; Yves Gramlich[3], Prof. Dr. med; René Verboket[4], Dr Med; Philip Bintaro[5], Dr Med; Maren Janko[4], Dr Med; Mairen H Flinspach[6], Dr Med; Michael Merker[7], Dr Med; Sven Bepler[8], Dr Med; Jan T Vollrath[4], Dr Med; Sebastian H Voß[9], Dr Med; Miriam Rüsseler[1], Prof Dr med

[1]Medical Faculty, Institute for Medical Education and Clinical Simulation, Goethe University, Theodor Stern Kai 7, Frankfurt am Main, Germany

[2]Department of Anaesthesiology, University Hospital Frankfurt, Intensive Care Medicine and Pain Therapy, Frankfurt am Main, Germany

[3]Agaplesion Markus Krankenhaus, Frankfurt am Main, Germany

[4]Department of Trauma Hand and Reconstructive Surgery, University Hospital Frankfurt, Frankfurt am Main, Germany

[5]Nephrologische Praxis und Dialyse im Klinikum Peine, Peine, Germany

[6]Department of Anaesthesiology, Intensive Care Medicine and Pain Therapy, Sana Clinic Offenbach GmbH, Offenbach/Main, Germany

[7]Department for Children and Adolescents, Division for Stem Cell Transplantation, Immunology and Intensive Care Medicine, University Hospital Frankfurt, Frankfurt am Main, Germany

[8]Department of Anaesthesiology, BG Unfallklinik Frankfurt am Main, Frankfurt am Main, Germany

[9]VOSS – doctor‘s office, Aschaffenburg, Germany

Corresponding Author:
Miriam Rüsseler, Prof Dr med
Medical Faculty, Institute for Medical Education and Clinical Simulation, Goethe University, Theodor Stern Kai 7, Frankfurt am Main, Germany

Abstract

Background: Cardiopulmonary resuscitation (CPR) is vital for improving patient outcomes in medical emergencies. Both laypersons and health care professionals often seek guidance on performing CPR. In today’s digital age, many turn to easily accessible platforms such as YouTube for practical skills.

Objective: This study evaluates the didactic and content quality of CPR videos on YouTube using comprehensive checklists and investigates the association between the assigned quality scores and type of publisher, view count, and video rankings.

Methods: Videos were included based on defined search terms and exclusion criteria. Two emergency physicians rated each video independently using validated checklists concerning content and didactic quality. Linear regression analysis was performed to assess the relationships between video quality scores and view counts, as well as video rankings.

Results: Of the 250 videos identified, 74 (29.6%) met the inclusion criteria. On the content checklist, videos scored an average of 56.5% (SD 19.2%), and on the didactic checklist, they scored 66.6% (SD 14.3%); none achieved the maximum score. Videos from official medical institutions scored significantly higher in content quality compared to nonofficial sources (P=.04). Video quality scores were not associated with video rankings or view counts.

Conclusions: The study highlights substantial variability in the didactic and content quality of CPR-related videos on YouTube. For medical educators, this underlines the need to curate and recommend reliable online resources or to develop new high-quality content aligned with established checklists. For the general public, the findings caution against relying on popularity metrics as indicators of accuracy and emphasize the importance of guidance from trusted institutions.

(JMIR Form Res 2025;9:e69103) doi:10.2196/69103

KEYWORDS

cardiopulmonary resuscitation; CPR; basic life support; YouTube; quality assessment; didactic quality; content quality; medical education

Introduction

Ischemic heart disease, the leading global cause of death, presents a substantial challenge to health care systems worldwide [1,2]. In Europe, the incidence of out-of-hospital cardiac arrest ranges from 67 to 179 cases per 100,000 inhabitants [3]. Alarmingly, data from the German resuscitation

XSL•FO
RenderX

register show that only 11.1% of individuals who experience out-of-hospital cardiac arrest survive to hospital discharge [4].

Prompt and proficient cardiopulmonary resuscitation (CPR) substantially improves patient outcomes [5-8]. However, gaps in knowledge and performance persist among both laypersons and health care professionals, particularly in essential aspects such as compression depth, hand placement, and compression rate [9-12]. These deficits are further compounded by delays between arrest recognition, CPR initiation, and the arrival of professional medical support [4].

To address these challenges, widespread CPR education is essential, not only in formal training environments but also through accessible, scalable resources that support independent learning. In recent years, especially during the COVID-19 pandemic, online video platforms such as YouTube have become increasingly popular sources for acquiring or enhancing practical medical skills, including CPR [13]. While this accessibility offers great potential, concerns remain regarding the quality, accuracy, and didactic effectiveness of CPR videos freely available online [14,15].

Multiple studies have identified quality deficiencies in YouTube videos on resuscitation [16-27]. For instance, Yaylaci et al [28] reported that only 11.5% of resuscitation videos correctly demonstrated the necessary steps. Similarly, Elicabuk et al [29] reported that 75.3% of Turkish-language CPR videos failed to adhere to current guidelines. However, most previous evaluations relied on limited or nonvalidated assessment criteria [30,31]. This highlights the central problem: learners are widely exposed to CPR videos of uncertain quality but lack reliable indicators of which resources provide accurate and pedagogically sound instruction.

To address this gap, 2 validated checklists were developed: the didactic quality checklist by Rüsseler et al [32] and the content quality checklist by Sterz et al [33], the latter aligned with American Heart Association guidelines [31]. These instruments offer a structured way to evaluate both the educational effectiveness and the technical accuracy of resuscitation videos [32,33].

This study is the first to apply these validated checklists to a sample of English- and German-language YouTube CPR videos, while also examining associations between video quality and publisher type, view count, and search ranking. By integrating didactic and content analysis with visibility metrics, it provides a systematic evaluation of freely available CPR videos and highlights implications both for educators and for the general public.

Methods

Study Design and Setting

A retrospective, cross-sectional study design was used. The dataset includes videos uploaded between February 2010 and October 2018 and reflects the state of available content during that period. The video assessment is designed to address a broad target audience, including medical laypersons, professionals, and students.

The evaluation focuses on didactic quality, defined as the appropriateness of educational styles used in the videos, and content quality, which encompasses the procedural correctness of depicted resuscitation techniques and the accuracy of explanations regarding the resuscitation algorithm. Differences in assigned didactic and content quality scores are examined between videos published by official medical institutions or organizations and those from nonofficial sources. In addition, the analysis investigates potential associations between video quality scores and platform metrics such as view counts and search rankings, providing insights into the relationship between video quality and audience engagement within that historical context.

Video Selection Process

The video selection followed a multistage procedure. In October 2018, a search was conducted on YouTube using a German IP address and 10 relevant keywords in both English and German, including chest compression (“Thoraxkompression” and “Herzdruckmassage”), CPR, basic life support, cardiac arrest first aid (“Herzstillstand erste Hilfe”), heart first aid (“Herz erste Hilfe”), and resuscitation (“Wiederbelebung”). These terms were chosen to ensure thematic relevance and to capture common linguistic variations used by both laypersons and professionals.

To reflect typical user behavior, where users predominantly engage with results from the first search page, only the first 25 videos returned for each term were screened. Videos uploaded between 2010 and 2018 were included, acknowledging the ongoing visibility and ranking of older content within YouTube’s algorithm. For each video, metadata such as view count, number of likes and dislikes (noting that public dislike counts were removed in 2021), uploader identity, upload date, video duration, and channel subscriber count were documented.

Inclusion and Exclusion Criteria

Videos were included regardless of their explicitly stated target audience, as this information is often missing from video titles or descriptions. Instructional videos were also considered, even if not explicitly labeled as such.

Exclusion criteria comprised videos related to pediatric or animal resuscitation, content in languages other than English or German, videos lacking visual content (eg, audio only), demonstrations of mechanical compression devices, real-life or intraoperative resuscitation footage, and content focusing solely on cardiac arrest prevention without instructional guidance. Additionally, parody, satire, comedy, entertainment content, and promotional material for CPR training courses were excluded. Duplicate videos were removed based on identical URLs. An overview of the selection process is provided in the Results section.

Reviewer Selection

In total, 16 experienced emergency physicians participated as reviewers in the evaluation process. These individuals were selected based on their extensive expertise and experience in the medical field. The panel included 10 (63%) male physicians and 6 (37%) female physicians, representing a slightly skewed

XSL•FO
RenderX

gender distribution toward male physicians. The panel consisted of emergency physicians specialized in various fields, including pediatrics, internal medicine, oral and maxillofacial surgery, trauma care, and anesthesiology. All reviewers actively practice within emergency department settings. Addition, emergency physicians from different German cities, both from university-affiliated and nonuniversity settings, were intentionally included to align with training regulations. This ensures that individuals responsible for resuscitation have relevant experiences and knowledge, regardless of their practice location or institutional affiliation.

Didactic and Content Checklists

Two separate reviewers assessed each video using 2 distinct checklists: the didactic checklist (Checklist 1), developed by Rüsseler et al [32], and the content checklist (Checklist 2), developed by Sterz et al [33]. The mean of both ratings was used for subsequent analyses. In the event of markedly divergent ratings, cases would have been reexamined to ensure plausibility.

Scoring Methodology

The didactic checklist consisted of 21 items, each rated on a 5-point Likert scale (1=strongly disagree to 5=strongly agree), yielding a maximum possible score of 105 points. Didactic quality was evaluated based on various checklist aspects, including the title; learning goals; content and technique; content; text, graphics, and images; logical sequencing; aspects of hygiene; target audience; video length; readability of text, graphics, and images displayed; camera perspective; and the quality of auditory and visual elements.

The content checklist comprised 25 items across 4 domains: initial measures (7 items), chest compressions (8 items), automated external defibrillator (AED) use (6 items), and ventilation (4 items). Each item was scored on a 3-point scale (0=not mentioned, 1=incomplete or incorrect, and 2=correct), resulting in a maximum possible score of 50 points.

Applicability and Score Normalization

Not all checklist sections were applicable to every video. For example, many videos intended for layperson training did not include AED use or ventilation. In such cases, nonapplicable items were excluded entirely from both the maximum possible points and the score. Each video was thus assessed only on relevant items.

To allow for comparison across videos with varying scopes, scores were normalized. The achieved score was divided by the maximum applicable score for that specific video and multiplied by 100 to yield a percentage. For instance, a layperson-focused video with 20 applicable items (maximum score=40) that achieved all 40 points would receive 100%, just like a more comprehensive video with 25 applicable items and a maximum score of 50 that also achieved full marks.

Data Analysis

The data were recorded using Microsoft Excel. Statistical analysis was conducted using Minitab (Minitab Inc) and SPSS (version 26; IBM Corp). Videos were categorized by language, duration, view count, and publisher type (official vs nonofficial sources). Content and didactic quality scores were analyzed descriptively. For continuous variables, results are reported as means and SDs; categorical variables are presented as frequencies and percentages. Missing values were excluded. The 5 highest-rated, 5 lowest-rated, and 5 most-watched videos were identified to illustrate score extremes.

For group comparisons, Student *t* tests were applied to examine differences in quality scores (content and didactic) between official and nonofficial publishers. Assumptions of independence, normality, and equality of variances were considered: independence was given by design, normality was assessed using the Shapiro-Wilk and Kolmogorov-Smirnov tests, and equality of variances was evaluated with the Levene test. Cohen *d* was calculated as a measure of effect size.

Associations between video metrics (view count, YouTube ranking, and publisher type) and quality scores were examined using regression analyses. Linear regression was selected because the outcome variables (checklist scores) were continuous. Model assumptions were systematically evaluated using the diagnostic output from Minitab: scatterplots of predictors against outcomes (linearity), residuals-versus-fitted plots (homoscedasticity), and Q-Q plots of residuals (normality). Outliers and influential observations were identified through scatterplots and regression diagnostics, and sensitivity analyses were conducted with and without these data points to assess robustness. R^2 values were reported as measures of explained variance (effect size). Regression analyses were conducted in Minitab using a 95% confidence level. A *P* value of .05 or less was considered indicative of statistical significance.

Ethical Considerations

The Ethics Committee of the Faculty of Medicine, Goethe University Frankfurt, confirmed that no formal ethics approval was required for this type of educational research, as it does not constitute a biomedical research project in the sense of the Declaration of Helsinki and the requirements of §15 of the Professional Code of Conduct for Physicians in Hesse. The study was conducted in accordance with the Declaration of Helsinki.

Results

Overview

Initially, a total of 250 videos were identified based on the search terms. Among these, 120 videos were excluded due to meeting at least one exclusion criterion. Subsequently, 51 duplicate videos were identified and excluded, while an additional 5 videos were removed from the online platform before the rating process. This led to the final analysis comprising 74 videos (Figure 1).

XSL•FO
RenderX

Figure 1. Flowchart of YouTube video selection (retrospective cross-sectional study; YouTube videos accessed via German IP: 2010 - 2018; N=74). A total of 250 YouTube videos were screened. After applying predefined exclusion criteria and removing duplicates, 74 unique videos were included in the final analysis.

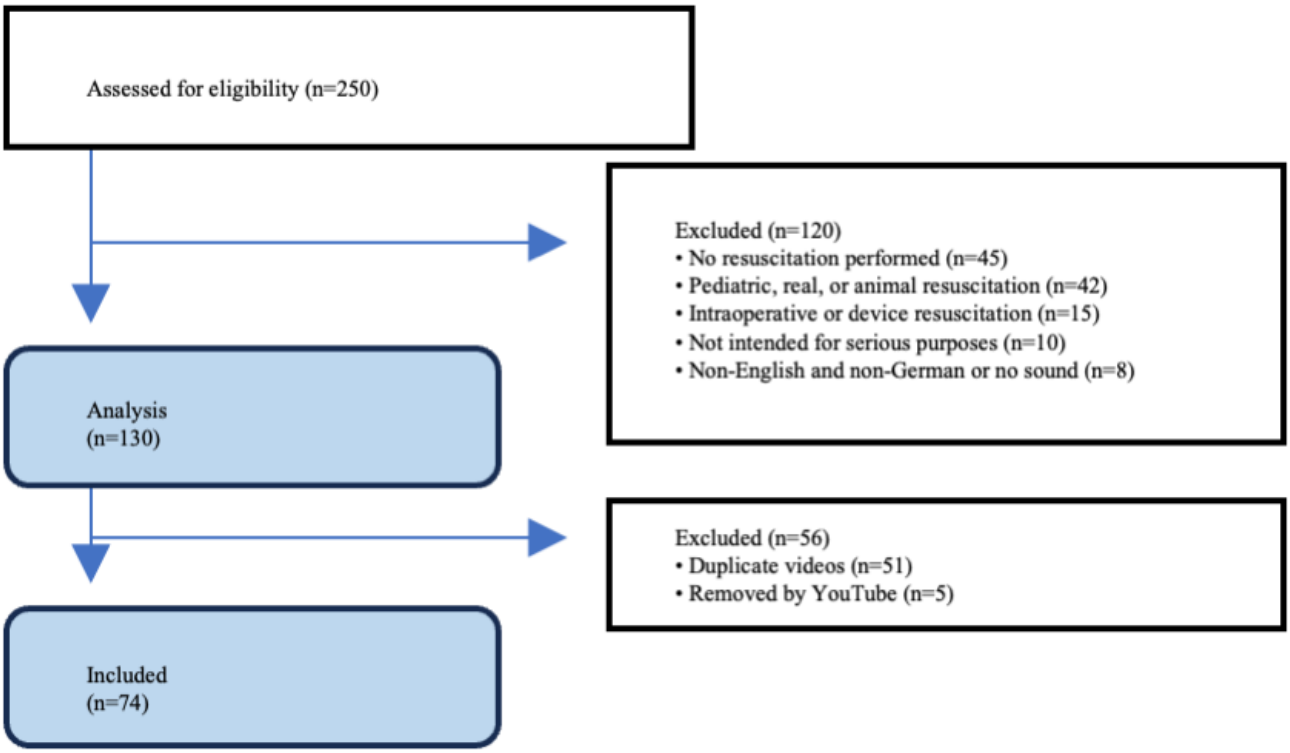

Video Characteristics

The timeline of video uploads spanned from February 1, 2010, to October 26, 2018. Among the included videos, 38 were in English, whereas 36 were in German. Video durations ranged from 13 seconds to 1 hour and 49 minutes. View counts for these videos varied considerably, ranging from 59 to 3,202,821 views. Notably, English-language videos generally garnered more views than their German counterparts, as of November 30, 2018. Table 1 provides a summary of the key characteristics of the videos.

XSL•FO
RenderX

Table . Key characteristics of the YouTube cardiopulmonary resuscitation videos (retrospective cross-sectional study; YouTube videos accessed via German IP: February 1, 2010-October 26, 2018; N=74).

Characteristics	Videos
Language, n (%)	
English	36 (49)
German	38 (51)
View count, mean (range; SD)	126,002 (59-3,202,821; 395,913)
≤60, n (%)	1 (1)
61 - 60,000, n (%)	50 (68)
60,001 - 120,000, n (%)	7 (10)
120,001 - 300,000, n (%)	11 (15)
300,001 - 3,000,000, n (%)	4 (5)
>3,000,000, n (%)	1 (1)
Duration (minutes:seconds), mean (range)	7:05 (0:13-60:49)
≤1:00, n (%)	1 (1)
1:01-5:00, n (%)	45 (61)
5:01-10:00, n (%)	17 (23)
10:01-20:00, n (%)	6 (8)
>20:00, n (%)	5 (7)
Publisher, n (%)	
Official medical institution	13 (18)
Other sources	61 (82)

Didactic Quality Evaluation

None of the videos achieved the maximum score of 100% on the didactic-related checklist. The average score on the didactic checklist was 66.6% (SD 14.3%) with one video achieving the highest possible score of 94.7%. Conversely, the lowest didactic quality score observed was 33.6% (Table 2). The didactic quality evaluation revealed significant disparities across single items, with higher ratings for visual (62.3% strongly agree) and audio (52.3% strongly agree) aspects indicating effective presentation, contrasted starkly by critical deficiencies in hygiene (57.7% strongly disagree) and in-depth reading (64.6% strongly disagree), underscoring urgent areas for improvement in the didactic approach. Table 3 presents the top 5 highest-rated videos, the 5 lowest-rated videos, and the 5 most-watched videos, along with their specific didactic quality scores and characteristics.

Table . Distribution of didactic quality scores (retrospective cross-sectional study; YouTube videos accessed via German IP: 2010 - 2018; N=74).

Scores	Videos[a]
≤50.00, n (%)	11 (15)
50.01 - 60.00, n (%)	16 (22)
60.01 - 70.00, n (%)	14 (19)
70.01 - 80.00, n (%)	16 (22)
80.01 - 90.00, n (%)	14 (19)
≥90.01, n (%)	3 (4)

[a]Mean 66.6% (SD 14.4%); range 33.6% - 94.7%.

XSL•FO
RenderX

Table . Top 5, bottom 5, and most-viewed YouTube videos by didactic quality score (retrospective cross-sectional study; YouTube videos accessed via German IP, 2010 - 2018; N=74).

Category and rank	Reference	Channel	Didactic score	Content score	Language	Year	Views, n	Duration (h:min:sec)
Top 5 videos by didactic quality score								
1	[34]	DRK Rettungsdienst Mittelhessen	94.7	92.5	German	2018	18,119	0:09:26
2	[35]	SIKANA English	92.6	77.5	English	2016	5240	0:02:55
3	[36]	Arbeitsgemeinschaft für Notfallmedizin	92.2	51.4	German	2015	4911	0:20:16
4	[37]	Das Weltrettungsforum im Namen der Wahrheit	88	55.2	German	2017	59	0:43:17
5	[38]	Saxe Healthcare Communications	86.7	11.9	English	2017	122,466	1:00:49
Bottom 5 videos by Didactic quality score								
1	[39]	nordbayern.de	33.6	48.1	German	2016	5522	0:00:13
2	[40]	LearnEngg	33.8	61	English	2016	314,273	0:03:25
3	[41]	American Heart Assoc.	38.7	56.3	English	2013	82,275	0:03:32
4	[42]	Conny X	44.3	25.8	German	2013	2406	0:04:19
5	[43]	H1 Fernsehen	45.3	34	German	2013	38,117	0:04:24
5 most-watched videos								
1	[44]	CPRCertified	74.9	82.7	English	2014	3,202,821	0:04:58
2	[45]	Weisbrod Imaging	86.3	93.6	English	2012	1,204,871	0:08:59
3	[46]	ProCPR	78.7	64.6	English	2011	600,031	0:06:30
4	[47]	tracy	76.2	59.6	English	2013	378,706	0:01:44
5	[40]	LearnEngg	33.8	61	English	2016	314,273	0:03:25

Content Quality Evaluation

Overview

None of the videos achieved the maximum score of 100% on the content-related checklist, with an average total score of 56.5%, about 10% less than the average didactic score. This indicates that videos demonstrated slightly higher didactic than content quality. The highest content checklist score observed for a video was 95.2%, while the lowest recorded score was 6.7% (Table 4). Upon detailed analysis of specific procedures within CPR instructional videos, it is observed that “Set pain stimulus” is mentioned in only 14.9% of the videos, whereas “Shortest possible hands-off time” is mentioned in 22.3%, and “Allowing complete chest recoil” is mentioned in only 29.7% of the videos. Conversely, “Make an emergency call,” “Frequency of 100 - 120/min,” and “Correct pressure point” were correct in 70.3%, 68.9%, and 66.9% of videos, respectively. Table 5 includes a selection of the top 5 highest-rated and lowest-rated videos, along with the 5 most watched videos and their related content quality score.

XSL•FO
RenderX

Table . Distribution of content quality scores (retrospective cross-sectional study; YouTube videos accessed via German IP: 2010 - 2018; N=74).

Score range	Videos[a]
<10.00, n (%)	1 (1)
10.01 - 30.00, n (%)	5 (7)
30.01 - 50.00, n (%)	20 (27)
50.01 - 70.00, n (%)	34 (46)
70.01 - 90.00, n (%)	9 (12)
≥90.01, n (%)	5 (7)

[a]Mean 56.5% (SD 19.3%); range 6.7% - 95.2%.

Table . Top 5, bottom 5, and most-viewed YouTube videos by content quality score in percent (retrospective cross-sectional study; YouTube videos accessed via German IP: 2010 - 2018; N=74).

Category and rank	Video reference	Channel	Didactic score	Content score	Language	Year	Views, n	Duration (h:min:sec)
Top 5 videos by content quality score								
1	[48]	LearningInn	73.6	95.2	English	2013	151,401	00:04:53
2	[49]	heartcom UG	86.3	93.6	German	2015	52,113	00:06:25
3	[34]	DRK Rettungsdienst Mittelhessen	94.7	92.5	German	2018	18,119	00:09:26
4	[50]	ercEuroResus-Council	77.1	92.4	English	2013	42,339	00:13:31
5	[51]	Thieme	80.5	90.5	German	2015	52,937	00:03:32
Bottom 5 videos by content quality score								
1	[52]	Dr. Heart	59.5	6.7	German	2016	4330	00:02:29
2	[38]	Saxe Healthcare Communications	86.7	11.9	English	2017	122,466	01:00:49
3	[53]	Liverpool John Moores University	56.0	19.6	English	2013	7407	00:01:19
4	[54]	CPR Council	54.3	22.4	English	2015	20,391	00:03:24
5	[55]	SAT.1 Regional	52.1	24.3	German	2015	265	00:02:05
5 most watched videos								
1	[44]	CPRCertified	74.9	82.7	English	2014	3,202,821	00:04:58
2	[45]	Weisbrod Imaging	86.3	93.6	English	2012	1,204,871	00:08:59
3	[46]	ProCPR	78.7	64.6	English	2011	600,031	00:06:30
4	[47]	tracy	76.2	59.6	English	2013	378,706	00:01:44
5	[40]	LearnEngg	33.8	61.0	English	2016	314,273	00:03:25

Differences in Quality Score by Publisher Type

Among the 74 videos analyzed, 13 (18%) were published by official medical institutions or organizations, while the remaining 61 (82%) originated from nonofficial sources. Videos published by official medical institutions achieved a content score of 67.5% (SD 20.4%) compared to those from nonofficial sources (mean 54.1%, SD 18.4%). This difference was statistically significant (t_{16}=−2.18, P=.04; 95% CI –26.4% to −0.4%). The effect size was moderate (Cohen d=0.71). For didactic quality, official videos scored slightly higher on average (mean 68.9%, SD 15.5%) than nonofficial videos (mean 66.1%, SD 14.3%), but this difference was not statistically significant (t_{14}=−0.23, P=.82; 95% CI −12.7% to 7.1%). The effect size was small (Cohen d=0.19).

XSL•FO
RenderX

Regression Analysis: Views, YouTube Ranking, and Quality Score

Regression analyses showed that videos with more views appeared to have slightly higher content quality scores (β=.2, P=.02, R^2=7.2%). However, this association was influenced by 3 outlier videos with view counts of 600,031; 1,204,871; and 3,202,821. After excluding these outliers, the association was no longer significant (β=.2, P=.08, R^2=4.3%). No association was found between view count and didactic quality (β=.2, P=.25, R^2=1.9%). Similarly, YouTube ranking was not associated with either content quality (β=.2, P=.59, R^2=0.4%) or didactic quality (β=.2, P=.08, R^2=4.3%).

Discussion

Deficiencies in Didactic and Content Quality of YouTube CPR Videos

This retrospective analysis of CPR-related YouTube videos revealed notable deficiencies in both didactic structure and content accuracy. The average didactic score was 66.6% (SD 14.3%), while the average content score was even lower at 56.5% (SD 19.2%). Importantly, many of the most viewed videos contained inaccurate information or omitted essential steps of resuscitation, such as correct compression technique or emergency call initiation. These results highlight a critical gap in the educational value of widely accessed CPR content and underscore the need for improved quality control in this domain.

Popularity Versus Quality: The Algorithmic Mismatch

Among the educational videos examined, it became evident that popularity, as measured by view counts, did not consistently align with the quality of didactic instruction or content accuracy. For example, the video with the highest didactic score had only 5200 views, while the video with the highest content score had 151,401 views, neither of which was among the most viewed videos overall. These observations raise questions about the role of algorithms in promoting content to a broader audience, potentially exposing viewers to inadequate or misleading information. Regression analyses confirmed this mismatch: although some associations between popularity and quality reached statistical significance, all models yielded very low R^2 values (<10%), indicating that views and rankings explain little of the variation in video quality and are therefore not reliable indicators of educational accuracy. Better curation and dissemination of accurate, high-quality resuscitation material is essential to ensure that the public has access to reliable information in this crucial field.

Incomplete Demonstration of Critical CPR Components

The findings align with previous studies by Katipoglu et al [15] and Ferhatoglu and Kudsioglu [14], which also analyzed CPR videos on YouTube and emphasized their poor quality. However, this study differs in its approach, as the checklists encompass a wider range of resuscitation aspects, including the often-overlooked complete chest recoil, revealing significant variations in video quality. While some critical actions, such as making an emergency call, were generally performed correctly, others, such as ensuring complete chest recoil, were frequently inadequately depicted or omitted. Favorable outcomes and high-quality CPR depend on the correct execution of all actions [56]. For instance, even with adequate chest compression depth and frequency, a positive outcome becomes less likely in the absence of chest recoil, which eliminates the diastolic filling phase [57]. In addition, only 22.3% of the videos adequately explained the shortest possible hands-off time, despite its pivotal role in maintaining continuous chest compression, a critical factor for a favorable outcome [58,59].

Instructional Gaps and the Concept of Conscious Competence

These gaps in video content may be attributed to the concept of "conscious competence" in teaching. Experienced educators often possess unconscious competence [60], automatically performing numerous details correctly without being able to explicitly articulate them. Creators of medical educational videos may also internalize critical aspects, such as chest recoil and minimal hands-off time, treating them as self-evident. To address this challenge, the study used checklists that were developed and validated by medical education experts to meticulously assess didactic and content-related aspects in detail [32,33].

Interestingly, when considering valid outliers, results showed a statistically significant association between view count and content quality score. However, after removing the outliers, no statistically significant association existed between video ranking, view counts, and assigned didactic or content quality scores. This implies again that viewers should exercise caution when relying solely on top-ranked videos to acquire practical medical skills, as these videos may not consistently offer the most accurate or comprehensive information. The recent removal of the dislike button on YouTube further complicates viewers' ability to accurately assess video quality.

While videos produced by official medical organizations received higher-quality ratings compared to nonofficial sources, it is vital to acknowledge that their videos still achieved an average content checklist score of only 67.5% (SD 20.4%). Therefore, while they may provide valuable insights, they should be approached with discernment rather than being unquestionably recommended for medical education purposes.

Comparison With Previous Work

The findings echo those of previous research on resuscitation videos and extend to studies on the quality of other instructional videos across a range of medical topics found on YouTube [22,24,28]. For example, Yoo et al [27] in 2020 found no difference in the quality of videos related to knee examinations from professional and nonprofessional organizations. Similarly, Flinspach et al [23] demonstrated a lack of association between the content parameters endorsed by YouTube and the overall quality of videos pertaining to epidural catheterization in obstetrics.

XSL•FO
RenderX

Implications for Future Practice and Educational Strategy

This study highlights the urgent need for more rigorous standards in the development and dissemination of medical educational videos. Validated checklists, such as those used in this analysis, offer a practical and evidence-based framework for both guiding content creation and evaluating instructional quality. Their integration into the production process can help ensure that videos are pedagogically sound, clinically accurate, and aligned with defined learning objectives. In light of the tendency among experienced practitioners to omit critical explanatory detail due to unconscious competence, content creators, particularly those without formal training in medical education, may benefit from targeted support in didactic design and instructional clarity. Moreover, given that video popularity does not reliably reflect content quality, professional organizations and academic institutions have a responsibility to curate, recommend, or endorse high-quality resources. Such measures are essential to help learners navigate an oversaturated digital landscape and access trustworthy materials.

Study Limitations

This study has limitations that warrant discussion. First, the retrospective design and focus on videos from 2010 to 2018 narrows the applicability of findings to the current YouTube landscape. However, many of these older videos remain widely accessible and are frequently viewed, and the core CPR principles remain, supporting the continued relevance of the dataset. Second, the study focused solely on freely available content from YouTube, justified by its widespread use as a source of resuscitation information [61]. This choice, while providing a comprehensive dataset, excludes potential insights from other platforms or sources. Additionally, the study included videos without explicit classification as instructional, recognizing that viewers often prioritize popularity metrics over instructional labels. However, this approach introduces variability in viewer interpretation. The restriction to English- and German-language videos reduces the transferability of findings to other linguistic and cultural contexts. Furthermore, reliance on YouTube's search and ranking system, combined with the decision to analyze only the first 25 results per term to reflect typical user behavior, may have introduced a visibility bias by favoring popularity over educational merit and potentially omitting higher-quality videos less prominently ranked. Interrater reliability was not formally calculated; however, the use of independent dual ratings and validated checklists provided safeguards against individual reviewer bias. Finally, the variation in the applicability of checklist items across different videos, especially regarding aspects related to AED and ventilation, represents an inherent limitation. Given the hands-only approach recommended for laypersons, certain video segments did not pertain to these specific categories, leading to uneven scoring across checklist items. To address this, scores were normalized by excluding nonapplicable items from the total possible points, allowing for a more accurate and comparable assessment.

Future Research

Future research should build on these findings in 2 ways. First, newer YouTube videos should be analyzed to determine whether content and didactic quality have improved in response to evolving CPR guidelines and changes in the platform's algorithms. This would provide an updated picture of the educational value currently available to learners. Second, research should move beyond descriptive analyses to test practical interventions aimed at improving video quality and visibility. Examples include integrating validated checklists into the production process, evaluating whether institutional peer review or endorsement increases viewer trust, and examining whether algorithmic adjustments can direct users toward higher-quality content. Together, these approaches would not only monitor the current state of online CPR education but also help identify strategies to actively enhance its reliability and reach.

Conclusions

This research reveals deficiencies in both the didactical quality and content accuracy of the CPR-related videos available on YouTube. Despite the potential for these videos to disseminate life-saving knowledge, many of them failed to meet fundamental criteria for effective CPR guidance. This educational gap is particularly worrisome, given the crucial roles that laypeople, health care professionals, and medical students play in emergencies. In the rapidly evolving landscape of online education, it is imperative to prioritize the widespread availability of high-quality, accurate, and accessible instructional materials, especially in critical domains like CPR.

Moving forward, medical educators, content creators, and professional organizations should take an active role in ensuring that online CPR education adheres to pedagogical and scientific standards. This may include the broader adoption of validated quality checklists, which inherently support structured video creation processes, as well as institutional endorsement of high-quality content. These measures are essential to empower viewers to make informed decisions about the accuracy of online content and, most importantly, to strengthen CPR training effectiveness and improve patient survival in real emergencies [62].

Disclaimer

ChatGPT (OpenAI) was used to assist with language editing. The authors take the ultimate responsibility for the content of this publication.

Data Availability

The datasets generated and analyzed in this study are available from the corresponding author upon reasonable request.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Checklist 1
Didactic quality checklist.
[DOCX File, 23 KB - formative_v9i1e69103_app1.docx]

Checklist 2
Content quality checklist.
[DOCX File, 18 KB - formative_v9i1e69103_app2.docx]

References

1. Kiguchi T, Okubo M, Nishiyama C, et al. Out-of-hospital cardiac arrest across the world: first report from the International Liaison Committee on Resuscitation (ILCOR). Resuscitation 2020 Jul;152:39-49. [doi: 10.1016/j.resuscitation.2020.02.044] [Medline: 32272235]
2. The top 10 causes of death. World Health Organization. 2024. URL: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death [accessed 2025-10-15]
3. Gräsner JT, Bein B. [Resuscitation - adult advanced life support]. Anästhesiol Intensivmed Notfallmed Schmerzther 2016 Mar;51(3):188-195. [doi: 10.1055/s-0041-103638] [Medline: 27022698]
4. Fischer M, Wnent J, Gräsner JT, et al. Jahresbericht des Deutschen reanimationsregisters: außerklinische reanimation 2021 [Article in German]. Anasthesiol Intensivmed 2022;63(6):V116-V122. [doi: 10.19224/ai2022.v116]
5. Rajan S, Wissenberg M, Folke F, et al. Association of bystander cardiopulmonary resuscitation and survival according to ambulance response times after out-of-hospital cardiac arrest. Circulation 2016 Dec 20;134(25):2095-2104. [doi: 10.1161/CIRCULATIONAHA.116.024400] [Medline: 27881566]
6. Hasselqvist-Ax I, Riva G, Herlitz J, et al. Early cardiopulmonary resuscitation in out-of-hospital cardiac arrest. N Engl J Med 2015 Jun 11;372(24):2307-2315. [doi: 10.1056/NEJMoa1405796] [Medline: 26061835]
7. Abella BS, Aufderheide TP, Eigel B, et al. Reducing barriers for implementation of bystander-initiated cardiopulmonary resuscitation: a scientific statement from the American Heart Association for healthcare providers, policymakers, and community leaders regarding the effectiveness of cardiopulmonary resuscitation. Circulation 2008 Feb 5;117(5):704-709. [doi: 10.1161/CIRCULATIONAHA.107.188486] [Medline: 18195177]
8. Yan S, Gan Y, Jiang N, et al. The global survival rate among adult out-of-hospital cardiac arrest patients who received cardiopulmonary resuscitation: a systematic review and meta-analysis. Crit Care 2020 Feb 22;24(1):61. [doi: 10.1186/s13054-020-2773-2] [Medline: 32087741]
9. Chaudhary A, Parikh H, Dave V. Current scenario: knowledge of basic life support in medical college. Natl J Med Res 2011;1(2):80-82 [FREE Full text]
10. Baldi E, Contri E, Bailoni A, et al. Final-year medical students' knowledge of cardiac arrest and CPR: we must do more!. Int J Cardiol 2019 Dec 1;296:76-80. [doi: 10.1016/j.ijcard.2019.07.016] [Medline: 31375334]
11. Roshana S, Kh B, Rm P, Mw S. Basic life support: knowledge and attitude of medical/paramedical professionals. World J Emerg Med 2012;3(2):141-145. [doi: 10.5847/wjem.j.issn.1920-8642.2012.02.011] [Medline: 25215053]
12. Swor R, Khan I, Domeier R, Honeycutt L, Chu K, Compton S. CPR training and CPR performance: do CPR-trained bystanders perform CPR? Acad Emerg Med 2006 Jun;13(6):596-601. [doi: 10.1197/j.aem.2005.12.021] [Medline: 16614455]
13. Bashir A, Bashir S, Rana K, Lambert P, Vernallis A. Post-COVID-19 adaptations; the shifts towards online learning, hybrid course delivery and the implications for biosciences courses in the higher education setting. Front Educ 2021 Aug 12;6. [doi: 10.3389/feduc.2021.711619]
14. Yilmaz Ferhatoglu S, Kudsioglu T. Evaluation of the reliability, utility, and quality of the information in cardiopulmonary resuscitation videos shared on open access video sharing platform YouTube. Australas Emerg Care 2020 Sep;23(3):211-216. [doi: 10.1016/j.auec.2020.05.005] [Medline: 32563667]
15. Katipoğlu B, Akbaş İ, Koçak AO, Erbay MF, Turan E, Kasali K. Assessment of the accuracy of cardiopulmonary resuscitation videos in English on YouTube according to the 2015 AHA resuscitation guidelines. Emerg Med Int 2019 May 2;2019:1272897. [doi: 10.1155/2019/1272897] [Medline: 31186962]
16. Spence AD, Derbyshire S, Walsh IK, Murray JM. Does video feedback analysis improve CPR performance in phase 5 medical students? BMC Med Educ 2016 Aug 12;16(1):203. [doi: 10.1186/s12909-016-0726-x] [Medline: 27519273]
17. Lehmann R, Lutz T, Helling-Bakki A, Kummer S, Huwendiek S, Bosse HM. Animation and interactivity facilitate acquisition of pediatric life support skills: a randomized controlled trial using virtual patients versus video instruction. BMC Med Educ 2019 Jan 5;19(1):7. [doi: 10.1186/s12909-018-1442-5] [Medline: 30611273]
18. Nomura O, Irie J, Park Y, Nonogi H, Hanada H. Evaluating effectiveness of YouTube videos for teaching medical students CPR: solution to optimizing clinician educator workload during the COVID-19 pandemic. Int J Environ Res Public Health 2021 Jul 2;18(13):7113. [doi: 10.3390/ijerph18137113] [Medline: 34281050]

XSL•FO
RenderX

19. Duncan I, Yarwood-Ross L, Haigh C. YouTube as a source of clinical skills education. Nurse Educ Today 2013 Dec;33(12):1576-1580. [doi: 10.1016/j.nedt.2012.12.013] [Medline: 23332710]
20. ReFaey K, Tripathi S, Yoon JW, et al. The reliability of YouTube videos in patients education for Glioblastoma treatment. J Clin Neurosci 2018 Sep;55:1-4. [doi: 10.1016/j.jocn.2018.07.001] [Medline: 30075980]
21. Fischer J, Geurts J, Valderrabano V, Hügle T. Educational quality of YouTube videos on knee arthrocentesis. J Clin Rheumatol 2013 Oct;19(7):373-376. [doi: 10.1097/RHU.0b013e3182a69fb2] [Medline: 24048116]
22. Helming AG, Adler DS, Keltner C, Igelman AD, Woodworth GE. The content quality of YouTube videos for professional medical education: a systematic review. Acad Med 2021 Oct 1;96(10):1484-1493. [doi: 10.1097/ACM.0000000000004121] [Medline: 33856363]
23. Flinspach AN, Raimann FJ, Schalk R, et al. Epidural catheterization in obstetrics: a checklist-based video assessment of free available video material. J Clin Med 2022 Mar 20;11(6):1726. [doi: 10.3390/jcm11061726] [Medline: 35330050]
24. Murugiah K, Vallakati A, Rajput K, Sood A, Challa NR. YouTube as a source of information on cardiopulmonary resuscitation. Resuscitation 2011 Mar;82(3):332-334. [doi: 10.1016/j.resuscitation.2010.11.015] [Medline: 21185643]
25. Hassounah MM, AlOwaini HS, Diab CN, Khamis NN. YouTube videos teaching Arabic speaking population how to perform cardiopulmonary resuscitation: the gap between the need and quality!. Resuscitation 2018 Oct;131:e13-e14. [doi: 10.1016/j.resuscitation.2018.08.009] [Medline: 30114448]
26. Şaşmaz MI, Akça AH. Reliability of trauma management videos on YouTube and their compliance with ATLS® (9th edition) guideline. Eur J Trauma Emerg Surg 2018 Oct;44(5):753-757. [doi: 10.1007/s00068-017-0803-9] [Medline: 28573427]
27. Yoo M, Hong J, Jang CW. Suitability of YouTube videos for learning knee stability tests: a cross-sectional review. Arch Phys Med Rehabil 2020 Dec;101(12):2087-2092. [doi: 10.1016/j.apmr.2020.05.024] [Medline: 32593549]
28. Yaylaci S, Serinken M, Eken C, et al. Are YouTube videos accurate and reliable on basic life support and cardiopulmonary resuscitation? Emerg Med Australas 2014 Oct;26(5):474-477. [doi: 10.1111/1742-6723.12274] [Medline: 25168312]
29. Elicabuk H, Yaylacı S, Yilmaz A, Hatipoglu C, Kaya FG, Serinken M. The reliability of Turkish “Basic Life Support” and “Cardiac Massage” videos uploaded to websites. Eurasian J Med 2016 Feb;48(1):15-19. [doi: 10.5152/eurasianjmed.2015.61] [Medline: 27026758]
30. Kleinman ME, Goldberger ZD, Rea T, et al. 2017 American Heart Association focused update on adult basic life support and cardiopulmonary resuscitation quality: an update to the American Heart Association guidelines for cardiopulmonary resuscitation and emergency cardiovascular care. Circulation 2018 Jan 2;137(1):e7-e13. [doi: 10.1161/CIR.0000000000000539] [Medline: 29114008]
31. Leitlinien-aktualisierung 2015 für HLW und kardiovaskuläre notfallmedizin: zusammenfassung der American Heart Association [Report in German]. : American Heart Association; 2015 URL: https://www.crit.cloud/uploads/2/7/6/1/27612891/2015-aha-guidelines-highlights-deutsch.pdf [accessed 2025-08-19]
32. Rüsseler M, Sterz J, Kalozoumi-Paisi P, et al. Qualitätssicherung in der lehre – entwicklung und analyse von checklisten zur beurteilung von lehrvideos zum erlernen praktischer fertigkeiten [Article in German]. Zentralbl Chir 2017;142(1):32-38. [doi: 10.1055/s-0042-124416]
33. Sterz J, Tückmantel PR, Bepler L, et al. [Development and validation of a checklist for evaluating videos for learning resuscitation measures]. Med Klin Intensivmed Notfmed 2022 Oct;117(7):525-530. [doi: 10.1007/s00063-021-00857-6] [Medline: 34468770]
34. Leitstellensymposium 2018 - reanimation. YouTube. 2018. URL: https://www.youtube.com/watch?v=soh9s4LD4Gc [accessed 2024-08-19]
35. Learn first aid gestures: adult cardiopulmonary resuscitation. YouTube. 2016. URL: https://www.youtube.com/watch?v=ZJVGPcd_BIM [accessed 2024-08-19]
36. Basic life support - Stefan Heschl. YouTube. 2015. URL: https://www.youtube.com/watch?v=bSQkXXd8Dxc [accessed 2024-08-19]
37. 4:56 Erste hilfe: was tun bei atem- und/oder herzstillstand? (ARD-buffet 28.09.2016). YouTube. 2017. URL: https://www.youtube.com/watch?v=nWBvvsR8Mn0 [accessed 2024-08-19]
38. Whats new in cardiac resuscitation AHA guidelines for ACLS and BLS. YouTube. 2017. URL: https://www.youtube.com/watch?v=6lCztv_3sMM [accessed 2024-08-19]
39. How to operate an automated external defibrillator (AED). YouTube. 2013. URL: https://www.youtube.com/watch?v=lKEL7rONx5c [accessed 2024-08-20]
40. How to do CPR | First aid treatments #learnengg #medicalemergency. YouTube. 2016. URL: https://www.youtube.com/watch?v=soq9CiuX67w [accessed 2024-08-19]
41. High-quality CPR and in-hospital adult resuscitation. YouTube. 2013. URL: https://www.youtube.com/watch?v=eXmAzsRQi9I [accessed 2024-08-20]
42. Herzdruckmassage flashmob. YouTube. 2013. URL: https://www.youtube.com/watch?v=78xSD7qyrx8 [accessed 2024-08-20]
43. Anleitung zur wiederbelebung: herz-lungen-massage. YouTube. 2013. URL: https://www.youtube.com/watch?v=M_05VPtRb8Q [accessed 2024-08-20]
44. How to perform CPR video. YouTube. 2014. URL: https://www.youtube.com/watch?v=cosVBV96E2g [accessed 2024-08-19]

XSL•FO
RenderX

45. CPR/AED refresher course (2012). YouTube. 2012. URL: https://www.youtube.com/watch?v=CuUXdQI5LLs [accessed 2024-08-19]
46. Adult CPR. YouTube. 2011. URL: https://www.youtube.com/watch?v=OaSovqEimyA [accessed 2024-08-19]
47. Vinnie-jones-hard-and-fast-hands-only-CPR. YouTube. 2013. URL: https://www.youtube.com/watch?v=tD2qTmDsiHk [accessed 2024-08-19]
48. Basic life support in 5 minutes. YouTube. 2013. URL: https://www.youtube.com/watch?v=_zQpiDhTc-0 [accessed 2024-08-19]
49. Herz-lungen-wiederbelebung und der einsatz eines AED retten leben. YouTube. 2015. URL: https://www.youtube.com/watch?v=Boj82HsDD60 [accessed 2014-08-19]
50. Cardiopulmonary resuscitation (CPR) and automated external defibrillation (AED) on an Ambu mannequin. YouTube. URL: https://www.youtube.com/watch?v=eNr9x3VJZyM [accessed 2024-08-19]
51. Thieme. Sofortmassnahmen reanimation. YouTube. 2018. URL: https://www.youtube.com/watch?v=D_C3hFhXXyY [accessed 2018-11-22]
52. Herzdruckmassage und wiederbelebung - allgemeines. YouTube. 2016. URL: https://www.youtube.com/watch?v=Eklew0tJb_o [accessed 2024-08-19]
53. How to perform CPR (cardiopulmonary resuscitation). YouTube. 2013. URL: https://www.youtube.com/watch?v=EGMSH7uz8kM [accessed 2024-08-19]
54. RC (UK) guidelines 2015 - professor Gavin Perkins. YouTube. 2015. URL: https://www.youtube.com/watch?v=D7qbWeEqd1E [accessed 2024-08-19]
55. “Hamburg shocks”: first aid for sudden cardiac arrest. YouTube. 2016. URL: https://www.youtube.com/watch?v=nUawgUynDsM [accessed 2024-08-19]
56. Harris AW, Kudenchuk PJ. Cardiopulmonary resuscitation: the science behind the hands. Heart 2018 Jul;104(13):1056-1061. [doi: 10.1136/heartjnl-2017-312696] [Medline: 29353251]
57. Yannopoulos D, McKnite S, Aufderheide TP, et al. Effects of incomplete chest wall decompression during cardiopulmonary resuscitation on coronary and cerebral perfusion pressures in a porcine model of cardiac arrest. Resuscitation 2005 Mar;64(3):363-372. [doi: 10.1016/j.resuscitation.2004.10.009] [Medline: 15733767]
58. Brouwer TF, Walker RG, Chapman FW, Koster RW. Association between chest compression interruptions and clinical outcomes of ventricular fibrillation out-of-hospital cardiac arrest. Circulation 2015 Sep 15;132(11):1030-1037. [doi: 10.1161/CIRCULATIONAHA.115.014016] [Medline: 26253757]
59. Cheskes S, Common MR, Byers PA, Zhan C, Morrison LJ. Compressions during defibrillator charging shortens shock pause duration and improves chest compression fraction during shockable out of hospital cardiac arrest. Resuscitation 2014 Aug;85(8):1007-1011. [doi: 10.1016/j.resuscitation.2014.05.001] [Medline: 24830868]
60. Broadwell MM. Teaching for learning (XVI). Gospel Guard 1969 Feb 20;20(41):1a-3a [FREE Full text]
61. Stollak MJ, Vandenberg A, Burklund A, Weiss S. Getting social: the impact of social networking usage on grades amongst college students. Presented at: Proceedings from ASBBS Annual Conference; Feb 22-27, 2011; Las Vegas, NV URL: https://www.researchgate.net/publication/303818592_Getting_Social_The_Impact_of_Social_Networking_Usage_on_Grades_Amongst_College_Students [accessed 2025-10-16]
62. Marino PL, Galvagno SM, Sing RF. Marino’s the Little ICU Book: Lippincott Williams & Wilkins; 2016.

Abbreviations

AED: automated external defibrillator
CPR: cardiopulmonary resuscitation

Edited by A Mavragani; submitted 28.11.24; peer-reviewed by ET Cakmak, R D'Ambrosi, V Reis; revised version received 15.09.25; accepted 16.09.25; published 05.11.25.

Please cite as:
Sterz J, Beaugé Y, Tueckmantel P, Bepler L, Flinspach AN, Gramlich Y, Verboket R, Bintaro P, Janko M, Flinspach MH, Merker M, Bepler S, Vollrath JT, Voß SH, Rüsseler M
Didactic and Content Quality of Basic Life Support Videos on YouTube: Cross-Sectional Study
JMIR Form Res 2025;9:e69103
URL: https://formative.jmir.org/2025/1/e69103
doi:10.2196/69103

XSL•FO
RenderX

© Jasmina Sterz, Yvonne Beaugé, Pia Tueckmantel, Lena Bepler, Armin N Flinspach, Yves Gramlich, René Verboket, Philip Bintaro, Maren Janko, Mairen H Flinspach, Michael Merker, Sven Bepler, Jan T Vollrath, Sebastian H Voß, Miriam Rüsseler. Originally published in JMIR Formative Research (https://formative.jmir.org), 5.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Quality Analysis of Stroke-Related Videos on Video Platforms: Cross-Sectional Study

Shao-Jie Nie[1*], MS; Shuai-Nan Ning[2,3*], MS; Qi-Chao Ding[4], MM; Jie Hou[5], MM; Yue-Yang Luo[6], MMS; Hai-Yang Wang[3], MBBS

[1]Department of General Surgery, The 966th Hospital of PLA Joint Logistics Support Force, Dandong, China

[2]Chinese PLA Medical School, Beijing, China

[3]Medical Service Division, The 966th Hospital of PLA Joint Logistics Support Force, 19 Shijing Street, Zhenxing District, Dandong, China

[4]The Second Navy Hospital of Southern Threater Command of PLA, Sanya, China

[5]Changzheng Hospital, Naval Medical University, Shanghai, China

[6]Sichuan Provincial Center for Mental Health, Sichuan Provincial People's Hospital, School of Medicine, University of Electronic Science and Technology of China, Chengdu, China

[*]these authors contributed equally

Corresponding Author:
Hai-Yang Wang, MBBS
Medical Service Division, The 966th Hospital of PLA Joint Logistics Support Force, 19 Shijing Street, Zhenxing District, Dandong, China

Abstract

Background: Stroke has become a global public health problem due to its high incidence, disability, and mortality. In China, TikTok and Bilibili, as mainstream video-sharing platforms, serve as key sources of getting stroke-related information for people, yet their videos' content quality and reliability remain insufficiently evaluated.

Objective: This cross-sectional study aimed to analyze the content and quality of stroke-related videos on Chinese video-sharing platforms.

Methods: In March 2025, stroke-related videos were retrieved from TikTok and Bilibili using the search term “卒中” (Chinese for stroke). Eligible videos were analyzed for metadata and content indicators. Researchers assessed video quality using validated tools: the Global Quality Scale (GQS), modified DISCERN (mDISCERN), and Patient Education Materials Assessment Tool (PEMAT). Statistical analyses were performed with Python, including descriptive statistics, group comparisons (Kruskal–Wallis tests), and Spearman’s rank correlation to evaluate variable associations, with all *P* values adjusted for multiple comparisons using the Bonferroni method. A binary classification predictive model was developed using the random forest algorithm, accompanied by feature importance analysis.

Results: Among the stroke-related videos from Bilibili (n=157) and TikTok (n=149), popular science education content predominated (204/306, 66.7%). Bilibili videos were primarily categorized as professional lectures (83/157, 52.9%), while most TikTok videos were popular science education (139/149, 93.3%). TikTok videos demonstrated significantly higher median likes and comments ($P<.001$) and shorter durations compared to Bilibili ($P<.001$). No significant differences were observed in median GQS (4) or mDISCERN scores (3) between platforms ($P>.05$). Videos produced by professional teams exhibited significantly higher GQS and PEMAT-A/V scores than those created by independent content creators ($P<.05$). Popular science education videos achieved higher PEMAT-A/V actionability scores than professional lectures ($P<.001$), while videos addressing treatment options scored lowest in GQS ($P<.05$). Strong positive correlations were identified among user engagement parameters (likes, shares, comments; $\rho=0.81$ - 0.90, $P<.001$), but only weak correlations were found between engagement and quality scores ($\rho<0.3$). Machine learning modeling (AUC=0.58) identified video duration (importance score: 0.15) and uploader subscriber count (importance score: 0.13) as key predictors of content quality.

Conclusions: The quality of stroke-related videos on TikTok and Bilibili remains suboptimal. Content uploaded by certified physicians and institutions received higher GQS/mDISCERN scores, confirming that medical authority is a key quality indicator. Our exploratory random-forest model, which used only basic metadata (duration, likes, subscriber count), achieved an area under the curve of 0.58, indicating that surface engagement metrics alone are insufficient to discriminate high- from low-quality material. Consequently, future screening algorithms should incorporate content-based features (eg, captions, medical keywords, visual cues) and creator credentials rather than relying solely on readily available interaction parameters. Multi-platform, larger-scale datasets are warranted to develop clinically useful prediction tools.

XSL•FO
RenderX

(JMIR Form Res 2025;9:e80458) doi:10.2196/80458

KEYWORDS

stroke; short videos; TikTok; video quality; PEMAT-A/V; quality assessment; Patient Education Materials Assessment Tool

Introduction

Stroke, an acute cerebrovascular disease, is pathophysiologically characterized by abrupt focal neurological deficits resulting from cerebral circulation dysfunction. Current classification delineates two entities: (1) ischemic stroke (60%‑70% of incident cases), primarily caused by atherosclerotic thrombosis or cardioembolic occlusion and (2) hemorrhagic stroke (30%‑40%), encompassing hypertensive parenchymal hemorrhage and aneurysmal subarachnoid hemorrhage [1]. Globally, stroke remained the second leading cause of death (11.6% of total deaths) and the third leading cause of death and disability combined (5.7% of total disability-adjusted life-years [DALYs]). According to the Global Burden of Disease Study 2019, there were 12.2 million incident stroke cases worldwide. Despite a 17% decline in age-standardized incidence rates from 1990 to 2019,, the absolute number of stroke cases demonstrated alarming growth—newly diagnosed cases surged by 70%, while the age-standardized prevalence rate among people under 70 years rose by 22% [2]. According to the Global Burden of Disease 2019, China maintained the highest stroke burden globally, reporting 3.94 million (95% uncertainty interval 3.43‑4.58) incident stroke cases in 2019. Alarmingly, the age-standardized incidence rate surged by 86% (73.2‑99.0) from 1990 to 2019, reaching 276.7 (241.3‑322.0) per 100,000 population by 2019 [3,4]. Notably, the mean age of patients with stroke and stroke-related fatalities in China clusters in the mid-60s, nearly a decade younger than their counterparts in high-income countries, where peak incidence occurs in the mid-70s [4]. Metabolic disorders and suboptimal lifestyle patterns, as key modifiable determinants, drive the escalating stroke burden; therefore, lifestyle modification coupled with targeted health education constitutes a cornerstone intervention for primary prevention [5,6]. Globally, approximately 50%‑70% of stroke cases are attributable to hypertension [1]. In 2019, hypertension contributed to 79.6 million DALYs, accounting for 55.5% of total stroke-related DALYs [2]. In China, the population-attributable risk of hypertension for stroke-related mortality demonstrated a sustained increase from 48.31% (1990) to 57.18% (2021) [5]. The proportion of patients with hypertension being well controlled is under 20% in China, markedly lower than the corresponding rates in the United Kingdom or the United States [4,6]. The suboptimal persistence with and adherence to secondary preventive therapies is associated with a 1-year recurrence rate of 13.2% in Chinese patients with stroke [5,7]. Therefore, given these epidemiological characteristics, systematic health education programs to improve public awareness of stroke are critically necessary. Through evidence-based health communication, such initiatives can effectively enhance preventive awareness, improve early symptom recognition, standardize prehospital emergency protocols, clarify different clinical interventions, and guide personalized rehabilitation strategies [6,8,9].

With the rapid development of mobile internet technology, video platforms—emerging communication media—have become a crucial channel for the public to access health information due to their visual presentation and fragmented dissemination features. As the world's largest internet user market, China has seen its domestic short-video platforms, TikTok (Chinese Version) and Bilibili, surpass 700 million and 300 million monthly active users, respectively, collectively accounting for over 65% of the domestic video traffic market share [10,11]. Video-based educational interventions on digital platforms significantly enhance layperson recognition of stroke symptoms [12]. Implementation of evidence-based health education initiatives, such as the "1-2-0" stroke recognition protocol (a simplified detection method analogous to the FAST mnemonic), is associated with reduced prehospital emergency delay, thereby improving clinical outcomes and alleviating the burden among patients with stroke [13]. However, official health authorities have not fully utilized video platforms to promote health education initiatives [14], and nowadays, these open platforms lack stringent content review mechanisms, leading to inconsistent quality of disease education videos. Patients and their caregivers frequently access health-related videos for medical information. However, exposure to inaccurate content may lead to misunderstandings regarding disease diagnosis and treatment, potentially compromising personal health decision-making [15,16]. Existing studies have systematically evaluated the quality of video content on various diseases, including colorectal polyps, mitral regurgitation, and steatotic liver disease [17-19]. While studies have documented improvements in the quality of stroke-related videos on YouTube in 2023, they still fall short of serving as reliable educational resources for accurate stroke symptom recognition [20]. We conducted this cross-sectional investigation of stroke-related videos on TikTok and Bilibili to analyze their uploaded sources, content, and features, aiming to systematically evaluate videos from two platforms for their informational quality and educational value, with usage of validated quantitative scoring frameworks—including a modified version of the DISCERN instrument (mDISCERN), the Global Quality Scale (GQS), and the Patient Education Materials Assessment Tool for Audiovisual Materials (PEMAT-A/V) [21].

Methods

Ethical Considerations

Ethical approval was not required since this study exclusively analyzed publicly available data from short-form video-sharing platforms and involved no experimentation on human subjects. All potentially identifiable information was anonymized through strict confidentiality protocols, with data access restricted solely to the research team.

XSL•FO
RenderX

Search Strategy and Data Collection

The search was performed on TikTok and Bilibili on March 12, 2025. To minimize bias introduced by personalized recommendation algorithms, we conducted searches using newly registered accounts on the web interface of each platform without applying any filters. The keyword "stroke" (卒中) was systematically searched across both platforms. We extracted metadata from all 196 videos displayed on the search page for TikTok. For Bilibili, we collected video metadata from the first six pages of search results, resulting in 212 videos.

Our database comprised 306 videos (157 from Bilibili and 149 from TikTok) meeting predefined inclusion criteria: stroke-related content presented in Chinese or English, excluding duplicates and irrelevant materials. Each video was recorded and analyzed across multiple parameters, including Title, Likes, Comments, Favorites, Shares, Days Published, Duration, Author Identification, Subscribers, Disease Knowledge Dimensions, and Content Typology. Platform-specific variations were noted: TikTok inherently restricts access to view counts and danmu comments (real-time bullet–screen interactions). Bilibili uniquely tracks viewer support through "Coins" donations alongside metrics like views and danmu frequency. The composite indicators of Likes, Comments, Favorites, and Shares were collectively operationalized as audience engagement metrics.

Visual Classification

The classification framework was structured around three primary dimensions: disease knowledge dimensions, content typology, and author identification, which were subsequently treated as categorical features. The disease knowledge dimensions were categorized into six groups: symptom identification, disease prevention, pathogenesis and mechanisms, diagnostic methods, treatment options, and posttreatment rehabilitation. Content typology was systematically classified into four major types: popular science education, professional lectures, case reports or patient narratives, and other content types. For author categorization, uploader identities were divided into six hierarchical groups: certified medical professionals (platform-verified physicians), official entities (government departments, public hospitals, and state-affiliated media institutions), nonprofit science communicators (dedicated to science popularization), independent content creators, patients and family members, and other contributors.

Assessment of Content and Quality of Videos

Video credibility was assessed using the mDISCERN, while information quality was evaluated through the GQS. The educational impact on public health literacy was quantified using the PEMAT-A/V. These three validated instruments were systematically applied to all eligible videos identified through the predefined inclusion criteria, ensuring standardized quality appraisal across the dataset.

To evaluate the reliability and quality of the videos, we employed the mDISCERN score—an adaptation of the DISCERN tool [22]. This instrument has been validated for assessing health-related video content on platforms like TikTok [23-26]. The mDISCERN score consists of 5 questions: (1) Are the aims of the video clear and achieved? (2) Are reliable sources of information used? (3) Is the information presented in a balanced and unbiased manner? (4) Are additional sources of information listed for patient reference? (5) Are areas of uncertainty mentioned? Each question is scored as 1 for "yes" and 0 for "no." Higher scores indicate that the video is more reliable.

We applied the GQS to evaluate the quality of information in the videos. The GQS is widely recognized in assessing the quality of health information on online video platforms [23,25,26]. The GQS framework includes five key criteria: (1) poor quality (poor flow, most information missing, and not useful for patients), (2) generally poor quality (poor flow, some information provided but many important topics missing, and of very limited use to patients), (3) moderate quality (suboptimal flow, some important information adequately discussed while others are poorly covered, and somewhat beneficial for patients), (4) good quality (generally good flow, most of the relevant information is listed, though some topics are not covered and useful for patients, and (5) excellent quality (excellent flow and very useful information for patients). Higher scores indicate higher-quality videos.

To assess the educational effectiveness of the video materials for the public, we employed the PEMAT-A/V tool [27,28]. Specifically designed to evaluate audiovisual resources, the PEMAT-A/V comprises 17 items: 13 items measure the understandability of health information, while 4 items assess the actionability of recommendations. Responses are categorized as "agree" (1 point), "disagree" (0 points), or "N/A." Scores are calculated using the formula "total points obtained divided by total possible points×100%," generating overall scores and separate scores for understandability and actionability. Higher scores reflect superior clarity in comprehension and practical applicability of the content. The PEMAT actionability and understandability scores are also called PEMAT-A and PEMAT-U scores, respectively.

All authors are senior physicians with extensive clinical expertise in neurological specialties. One investigator (SNN) collected and archived all video data. Two independent raters (SNN and SJN) assessed the videos using the mDISCERN instrument, GQS, and PEMAT-A/V, with ratings finalized through iterative consensus discussions. Should discrepancies arise during the evaluation process, an arbiter (QCD) adjudicated the final decision, which was subsequently reviewed and calibrated by all authors to ensure unanimous approval of each rating.

The inter-rater reliability was rigorously assessed using the intraclass correlation coefficient (ICC), with reliability defined as poor (ICC<0.50), moderate (ICC=0.50 - 0.75), good (ICC=0.75 - 0.90), or excellent (ICC>0.90) [29]. The two researchers achieved good agreement across all evaluated domains: the GQS showed an ICC of 0.88 (95% CI 0.85 - 0.90), the mDISCERN score yielded an ICC of 0.86 (95% CI 0.82 - 0.88), while comprehensibility and actionability demonstrated particularly high consistency with ICC values of 0.94 (95% CI 0.93 - 0.95) and 0.97 (95% CI 0.97 - 0.98), respectively.

XSL•FO
RenderX

Data Preprocessing

Systematic recoding was performed to consolidate classification variables with category frequencies below 10 into an aggregated "Others" group, then transformed into binary features via one-hot encoding. Continuous numerical variables such as Subscribers, Days Published, and Likes were retained in their original scales without normalization. The mDISCERN score (range 1 - 5) served as the outcome variable, dichotomized into two quality classes: Class 0 (poor quality) for scores ≤3 and Class 1 (good quality) for scores >3.

Statistical Analysis

This study employed the Shapiro-Wilk test to assess the normality of data derived from the included videos. As all data followed nonparametric distributions, descriptive analyses were conducted using the median. The Kruskal–Wallis test was utilized for multigroup comparisons, while the Mann-Whitney *U* test was applied for comparisons between two groups. Chi-square tests analyzed nominal categorical variables across platforms, with the Monte Carlo simulation used to calculate *P* values for categories with expected frequencies of less than five. A Bonferroni correction was applied to account for multiple hypothesis testing, with adjusted *P* values <.05 considered significant. Given the categorical nature of our dataset and non-normal distribution, Spearman's rank correlation coefficient was adopted to evaluate parameter associations. Correlation strength was categorized as follows: <0.25 (weak), 0.25 - 0.50 (moderate), 0.50 - 0.75 (strong), and 0.75 - 1.00 (very strong). A random forest classifier was implemented via RandomForestClassifier (scikit-learn). The dataset was partitioned into training (70%) and testing (30%) sets using stratified sampling to preserve class distribution. Data analysis and model construction were performed using Python (version 3.12.7).

Results

The General Characteristics of Videos

After removing duplicate and irrelevant videos, our analysis identified a final cohort of 306 eligible videos, comprising 157 from Bilibili and 149 from TikTok. Descriptive metrics demonstrated wide variability: the median video duration was 210 seconds (5 - 7651 s), with median likes at 89 (0 - 276,000) and median shares at 41 (0 - 137,000). Comparative analyses revealed significant platform-specific disparities. Bilibili videos exhibited longer durations than their TikTok counterparts (*P*<.001). Conversely, TikTok videos achieved markedly higher median likes and comments (all *P*<.001), suggesting superior audience engagement on this platform. Temporal analysis showed TikTok videos had a median publication recency of 137 days (range 0 - 1965 d), significantly shorter than Bilibili's timeframe (*P*<.001), indicating more recent content dissemination on TikTok (Figure 1). The detailed characteristics of stroke-related videos across different platforms are summarized in Table 1.

XSL•FO
RenderX

Figure 1. Flow diagram depicting the selection process of stroke-related videos in this analysis.

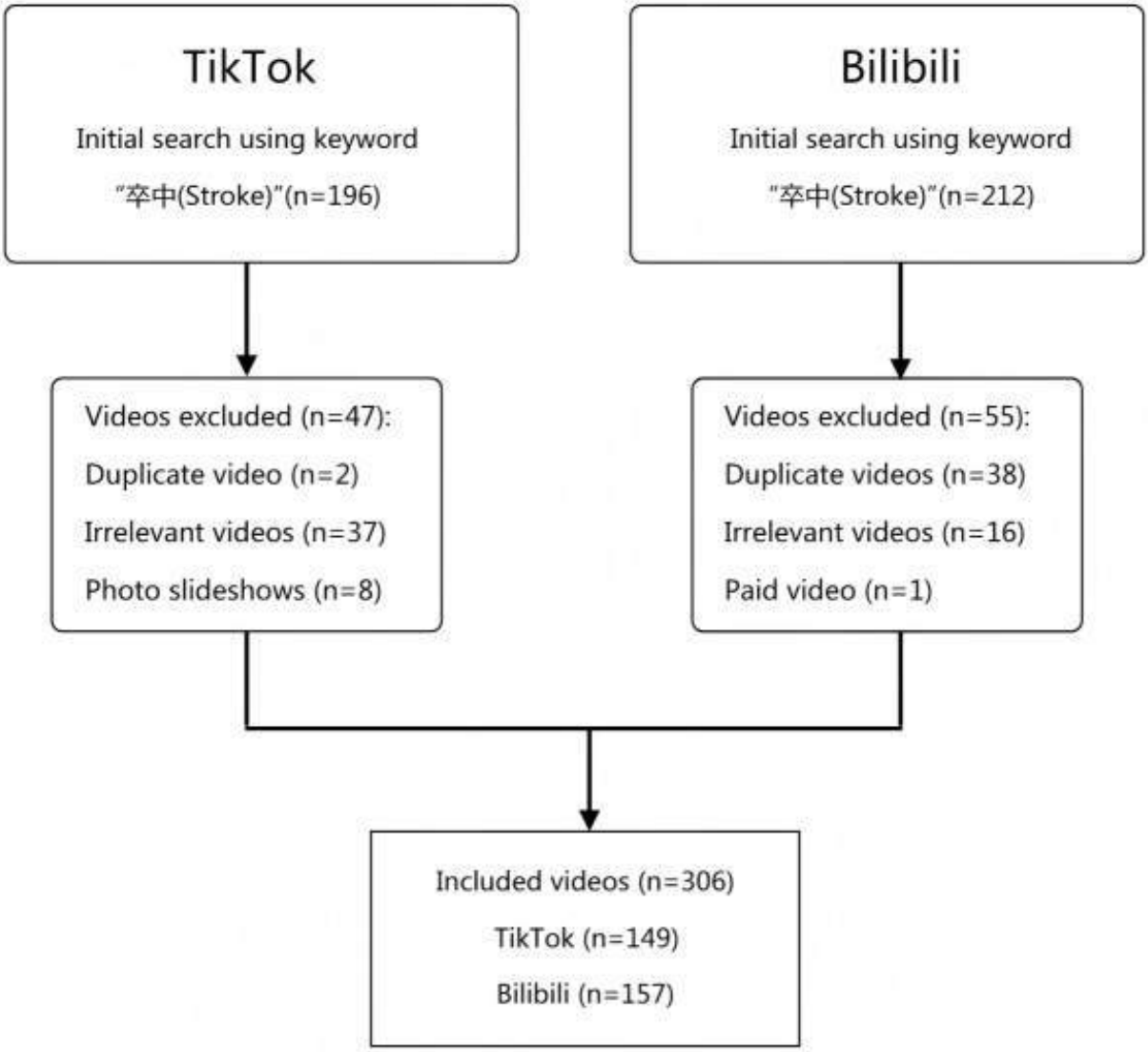

XSL•FO
RenderX

Table . The general characteristics and scores of the stroke-related videos.

Parameters	Total (N=306)	Bilibili (n=157)	TikTok (n=149)	*P* values
Likes	89 (0 - 2.76×105)	46 (0 - 44,000)	210 (2 - 2.76×105)	<.001
Coins	8 (0 - 10,000)	8 (0 - 10,000)	_[a]	_
Comments	5 (0 - 8433)	2 (0 - 973)	12 (0 - 8433)	<.001
Favorites	86 (0 - 90,000)	133 (0 - 33,000)	64 (0 - 90,000)	.140
Shares	41 (0 - 1.37×105)	36 (0 - 9553)	53 (0 - 1.37×105)	.026
Views	2,627 (1 - 15.67×105)	2,627 (1 - 15.67×105)	_	_
Danmaku	0 (0 - 1,117)	0 (0 - 1,117)	_	_
Fans of video uploaders	8,930 (0 - 2.6×108)	5,526 (0 - 28.09×105)	3.4×104 (7 - 2.6×108)	<.001
Duration (s)	210 (5 - 7651)	803 (37 - 7651)	84 (5 - 684)	<.001
Days since upload	334 (0 - 2887)	612 (0 - 2887)	137 (0 - 1965)	<.001
GQS[b] score (0 - 5)	4 (1-5)	4 (1-5)	4 (2-5)	.319
Modified DISCERN score (0 - 5)	3 (1-5)	3 (1-5)	3 (1-5)	.443
PEMAT[c] understandability, %	75 (18 - 100)	72.73 (18 - 100)	81.82 (33.33 - 100)	<.001
PEMAT actionability, %	33.33 (0 - 100)	33.33 (0 - 100)	33.33 (0 - 100)	.025

[a]Not available.

[b]GQS: Global Quality Scale.

[c]PEMAT: Patient Education Materials Assessment Tool.

The videos were categorized and statistically analyzed according to the classification criteria of author identification, content typology, and disease knowledge dimensions, with the results comprehensively tabulated in Table 2. Regarding author distribution, 102 videos were uploaded by certified medical professionals verified by the platforms, 109 by independent content creators, and 69 by official entities. On TikTok, medical professionals and official institutions dominated, accounting for 94.6% (141/149) of its videos, whereas independent content creators contributed 68.2% (107/157) of Bilibili videos. Regarding content typology, popular science education and professional lectures (including academic conferences, clinical courses, and guideline interpretations) constituted 66.7% (204/306) and 27.8% (85/306) of all videos, respectively. Notably, 52.9% (83/157) of Bilibili videos featured professional lectures, compared to only 1.3% (2/149) on TikTok, where popular science education dominated with 93.3% (139/149) of its content. Analysis of disease knowledge dimensions revealed that TikTok videos primarily focused on disease prevention and symptom identification (71.8% combined), while Bilibili videos predominantly addressed posttreatment rehabilitation and treatment methods (82%). Temporal and duration analyses showed that 71.8% (107/149) of TikTok videos were published after 2024, with 77.9% (116/149) lasting less than 180 seconds, whereas 60.5% (95/157) of Bilibili videos exceeded 600 seconds in duration. These findings indicate that TikTok videos are more recent and have a shorter duration than Bilibili.

XSL•FO
RenderX

Table . Description of stroke video sources, content, and characteristics.

Variable	Total (N=306), n (%)	TikTok (n=149), n (%)	Bilibili (n=157), n (%)	*P* values
Author Identification				<.001
Certified medical professionals	102 (33.3)	85 (57)	17 (10.8)	
Official entities	69 (22.5)	56 (37.6)	13 (8.3)	
Nonprofit science communicators	19 (6.2)	5 (3.4)	14 (8.9)	
Independent content creators	109 (35.6)	2 (1.3)	107 (68.2)	
Patients and family members	3 (1)	1 (0.7)	2 (1.3)	
Other contributors	4 (1.3)	0 (0)	4 (2.5)	
Content typology				<.001
Popular science education	204 (66.7)	139 (93.3)	65 (41.4)	
Professional lectures	85 (27.8)	2 (1.3)	83 (52.9)	
Case reports/patient narratives	9 (2.9)	4 (2.7)	5 (3.2)	
Other content types	8 (2.6)	4 (2.7)	4 (2.5)	
Disease knowledge dimensions				<.001
Disease prevention	78 (25.5)	54 (36.2)	24 (15.3)	
Symptom identification	64 (20.9)	53 (35.6)	11 (7)	
Posttreatment rehabilitation	56 (18.3)	10 (6.7)	46 (29.3)	
Pathogenesis and mechanisms	48 (15.7)	22 (14.8)	26 (16.6)	
Treatment options	46 (15)	10 (6.7)	36 (22.9)	
Diagnostic methods	14 (4.6)	0 (0)	14 (8.9)	
Publication date				<.001
Post-2025	70 (22.9)	53 (35.6)	17 (10.8)	
2024 - 2025	98 (32)	54 (36.2)	44 (28)	
2023 - 2024	63 (20.6)	25 (16.8)	38 (24.2)	
Pre-2023	75 (24.5)	17 (11.4)	58 (36.9)	
Duration				<.001
0-60 s	57 (18.6)	51 (34.2)	6 (3.8)	
60-180 s	82 (26.8)	65 (43.6)	17 (10.8)	
180-300 s	46 (15)	26 (17.4)	20 (12.7)	
300-600 s	25 (8.2)	6 (4)	19 (12.1)	
>600 s	96 (31.4)	1 (0.7)	95 (60.5)	

Video Quality and Reliability Assessments

The videos were assessed using the mDISCERN, GQS, and PEMAT-A/V quantitative assessment tools, with detailed scoring metrics in Table 1 and Figure 2. The median scores for GQS and mDISCERN across all videos were 4 (range 1 - 5) and 3 (range 1 - 5), respectively, with no significant differences observed between the two platforms (P>.05), indicating comparable video quality across platforms but highlighting persistent limitations in content reliability. However, significant disparities were identified in PEMAT-A/V understandability (P<.001) and actionability (P=.03), demonstrating that TikTok videos were more comprehensible and actionable compared to Bilibili content. The median PEMAT-A/V actionability score across all videos was 33.33% (0% - 100%), while the median understandability score was 75% (18.18% - 100%). Despite superior understandability metrics, the markedly lower actionability scores indicate that the videos prioritized theoretical comprehension over practical implementation strategies,

XSL•FO
RenderX

resulting in deficient guidance on actionable clinical steps for viewers.

Figure 2. The GQS, mDISCERN scores, and PEMAT scores of the videos. GQS: Global Quality Scale; mDISCERN: modified version of the DISCERN instrument; PEMAT-A/V: Patient Education Materials Assessment Tool for Audiovisual Materials.

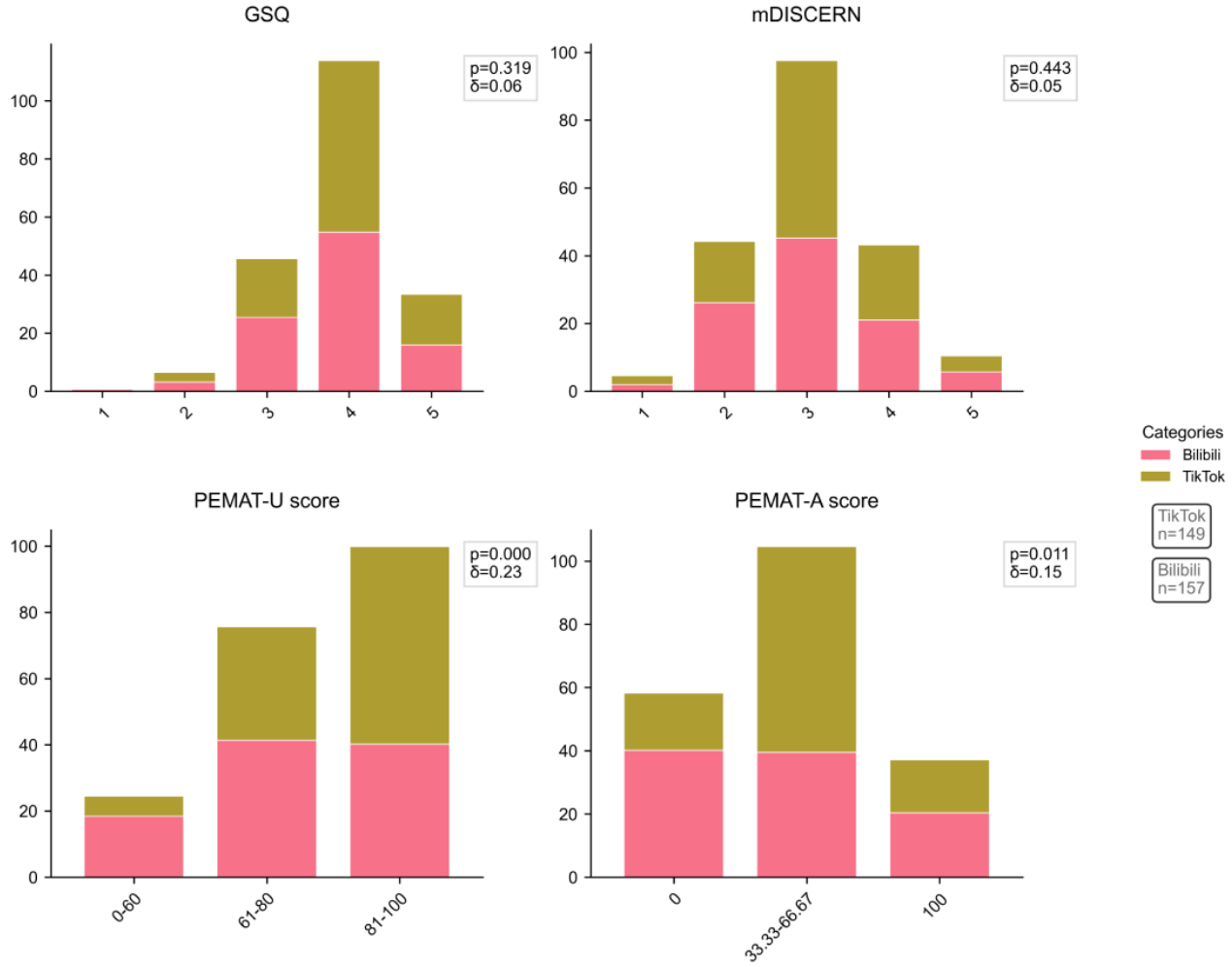

Comparison of the General Characteristics of Videos

Supplementary Table S1 and S2 in Multimedia Appendix 1 comprehensively present cross-platform intergroup comparisons of video characteristics categorized by author identification, content typology, and disease knowledge dimensions. On the TikTok platform, significant differences were observed in the publication timing of videos across various disease knowledge dimensions (P=.01). Videos specifically covering posttreatment rehabilitation exhibited significantly shorter publication timelines (ie, smaller video age values) compared to those focusing on pathogenesis (P=.04), disease prevention (P=.02), and symptom identification (P=.004). For stroke-related videos on Bilibili, videos about disease prevention were significantly shorter than those covering treatment options and diagnostic methods (P=.003 and P=.01, respectively). Additionally, significant differences were observed in shares and video duration across content typologies on Bilibili (all P<.001). Popular science education videos received significantly more likes and shares than professional lectures (P=.01 and P<.001, respectively), while professional lectures were significantly longer in duration (P<.001).

Regarding author identity on Bilibili, videos from nonprofit science communicators had significantly higher subscriber counts than those from independent content creators (P=.02), and videos from independent content creators were significantly longer than those from official entities (P=.04). On TikTok, significant differences were found in subscriber counts, likes, favorites, shares, and video duration across creators (P<.03). Nonprofit science communicators had significantly higher subscriber counts than certified medical professionals (P=.03). Furthermore, videos from nonprofit science communicators received significantly more likes, comments, favorites, and shares than those from certified medical professionals and official entities (all P<.006). Notably, videos from certified medical professionals on TikTok were published significantly more recently than those from official entities (P=.01).

Across both platforms, videos with different publication timelines exhibited significant differences in likes, favorites, comments, and shares (all P≤.001). On Bilibili, video duration varied significantly by publication timeline (P=.02). Additionally, videos with durations between 180 and 300 seconds on Bilibili received significantly more shares than those exceeding 600 seconds (P=.02).

XSL•FO
RenderX

Comparative Analysis of Video Quality Assessment Scores

As shown in Supplementary Table S1 and S2 in Multimedia Appendix 1, significant differences in PEMAT actionability scores were observed across disease knowledge dimensions on both platforms (all $P<.001$). Pairwise comparisons revealed that videos addressing pathogenesis and mechanisms had significantly lower actionability scores than those covering posttreatment rehabilitation, with platform-specific significance levels ($P=.001$ for Bilibili; $P=.004$ for TikTok). On Bilibili, popular science education videos exhibited markedly higher actionability scores than professional lectures ($P<.001$).

Temporal and duration-based analyses showed significant variations in PEMAT-A/V understandability scores across video durations on both platforms (Bilibili: $P=.0004$; TikTok: $P=.03$). Specifically, Bilibili videos lasting 300 - 600 seconds achieved higher actionability scores than those exceeding 600 seconds. Notably, on TikTok, videos shorter than 60 seconds demonstrated significantly higher understandability scores but lower mDISCERN reliability scores than 60 - to 180-second videos ($P<.01$ for all comparisons), indicating a trade-off between brevity-driven accessibility and content credibility.

After integrating video data from both platforms, comprehensive statistical analyses were conducted based on author identification, content typology, and disease knowledge dimensions, with subgroup comparisons detailed in Figures 3-5.

Figure 3. Quality assessment of videos based on author identification. (A) The GQS score. (B) The mDISCERN score. (C) The PEMAT-A/V understandability score. (D) The PEMAT-A/V actionability score. GQS: Global Quality Scale; mDISCERN: modified version of the DISCERN instrument; PEMAT-A/V: Patient Education Materials Assessment Tool for Audiovisual Materials.

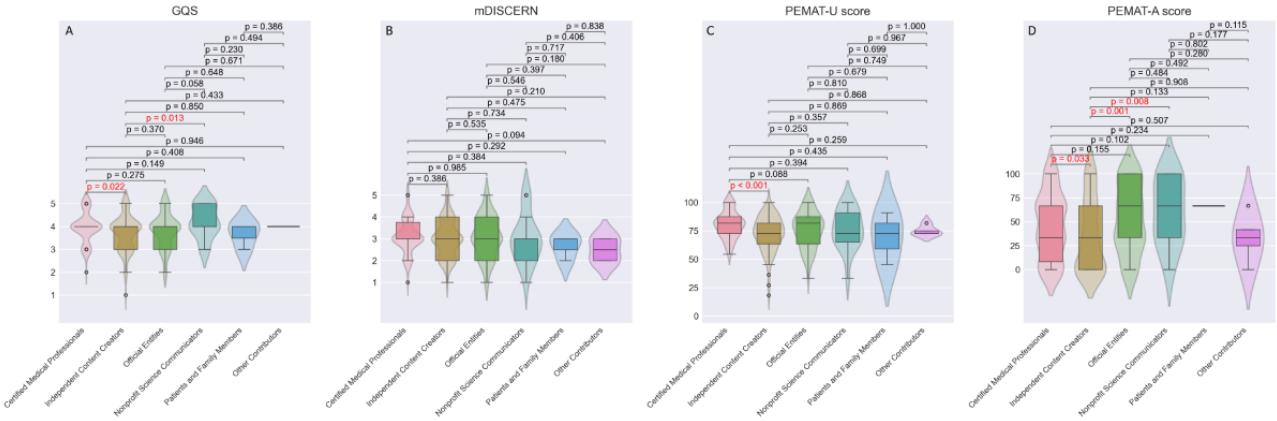

Figure 4. Quality assessment of videos based on content typology. (A) The GQS score. (B) The mDISCERN score. (C) The PEMAT-A/V understandability score. (D) The PEMAT-A/V actionability score. GQS: Global Quality Scale; mDISCERN: modified version of the DISCERN instrument; PEMAT-A/V: Patient Education Materials Assessment Tool for Audiovisual Materials.

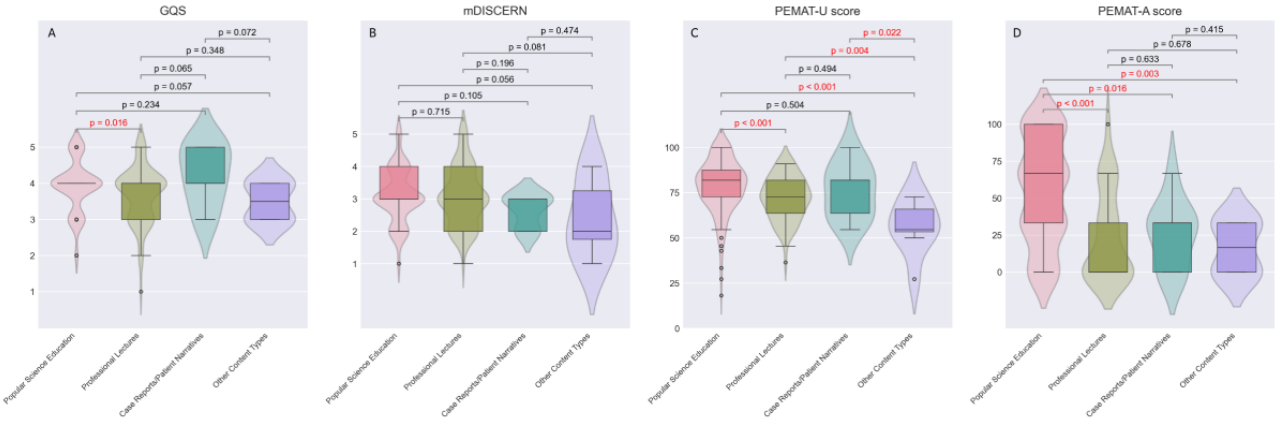

XSL•FO
RenderX

Figure 5. Quality assessment of videos based on disease knowledge. (A) The GQS score. (B) The mDISCERN score. (C) The PEMAT-A/V understandability score. (D) The PEMAT-A/V actionability score. GQS: Global Quality Scale; mDISCERN: modified version of the DISCERN instrument; PEMAT-A/V: Patient Education Materials Assessment Tool for Audiovisual Materials.

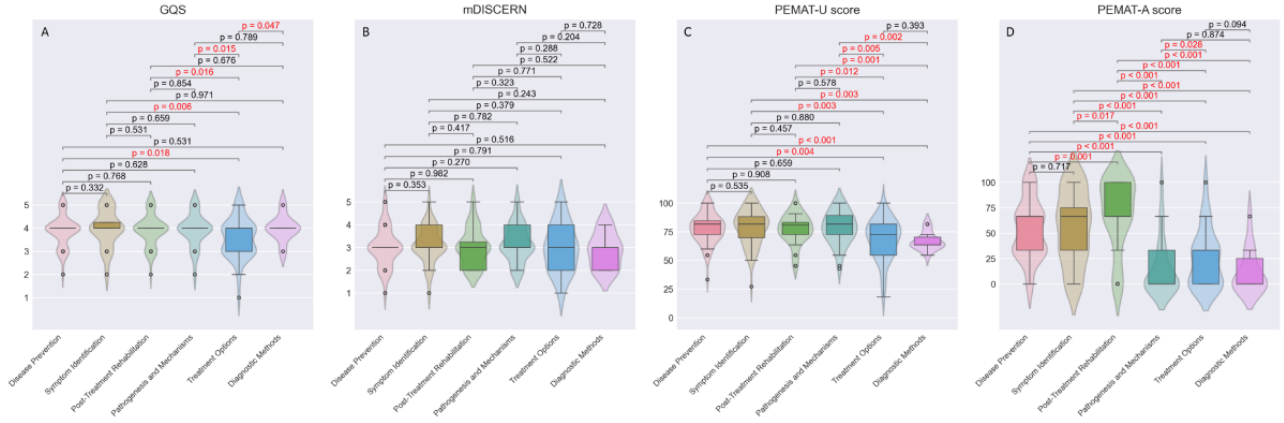

Certified medical professionals and nonprofit science communicators demonstrated significantly higher GQS scores than independent content creators (medical professionals versus independent creators: *P*=.02; nonprofit science communicators versus independent creators: *P*=.01). Videos produced by certified medical professionals also exhibited superior PEMAT understandability and actionability scores relative to independent content creators (*P*<.001 and *P*=.03, respectively). Similarly, videos from official entities and nonprofit science communicators achieved significantly higher PEMAT actionability scores compared to those by independent content creators (official entities versus independent creators: *P*=.001; nonprofit science communicators versus independent creators: *P*=.01). This suggests that content produced by professional teams or individuals with organizational backing provides more actionable guidance for audiences than videos created by independent contributors.

Regarding disease knowledge dimensions, videos addressing treatment methods exhibited significantly lower GQS scores than other disease-related content (all *P*<.05). Posttreatment rehabilitation videos achieved higher PEMAT actionability scores than other categories (all *P*<.05). Treatment options and diagnostic methods videos consistently scored lower in PEMAT understandability (all *P*<.05) than other disease knowledge domains.

From the perspective of content typology, popular science education videos significantly outperformed professional lectures in both GQS and PEMAT-A/V scores (all *P*<.02), particularly excelling in actionability metrics. Notably, no significant differences were observed in mDISCERN scores across subgroups.

Correlation Analysis

Spearman correlation analysis revealed the relationship between video engagement metrics and quality indicators (Figure 6). Strong positive correlations were observed between likes, shares, comments, favorites, and uploader subscriber counts (all *P*<.001), with particularly robust correlations between favorites and shares (ρ=0.90, *P*<.001), likes and shares (ρ=0.85, *P*<.001), likes and comments (ρ=0.84, *P*<.001), and likes and favorites (ρ=0.81, *P*<.001). In contrast, weaker correlations (ρ<0.3) were found between these engagement metrics and quality scores, such as GQS versus likes (ρ=0.16, *P*<.01), PEMAT-A versus likes (ρ=0.19, *P*=.001), and PEMAT-A versus shares (ρ=0.22, *P*<.001).

A moderate positive correlation was identified between video duration and publication date (ρ=0.31, *P*<.001), suggesting shorter video durations in more recent uploads. Significant negative correlations primarily involved video duration, including platform (TikTok/Bilibili) versus duration (ρ=−0.70, *P*<.001), likes versus duration (ρ=−0.33, *P*<.001), comments versus duration (ρ=−0.31, *P*<.001), PEMAT-U vs duration (ρ=−0.25, *P*<.001), and PEMAT-A versus duration (ρ=−0.17, *P*<0.001).

XSL•FO
RenderX

Figure 6. Spearman correlation between video features. GQS: Global Quality Scale; mDISCERN: modified version of the DISCERN instrument; PEMAT-A/V: Patient Education Materials Assessment Tool for Audiovisual Materials.

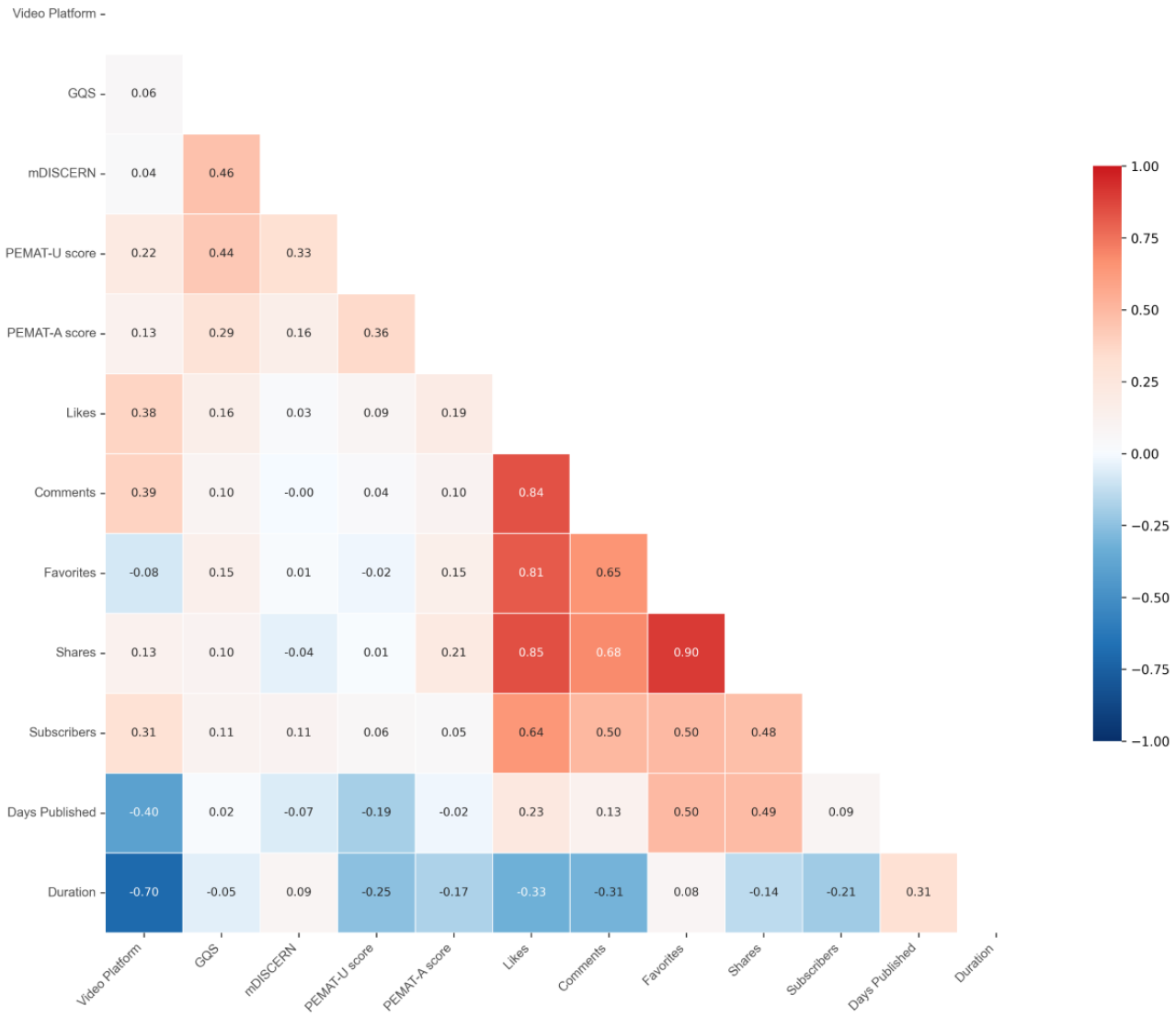

Machine Learning Model

The model demonstrated moderate discriminative capacity, with an accuracy of 0.663 and an area under the curve (AUC)-receiver operating characteristic of 0.577, indicating limited yet nonnegligible predictive utility for video quality classification (high vs low quality). Feature importance analysis derived from the random forest model identified the top 10 predictors associated with the outcome variable (mDISCERN), with video duration (importance score: 0.151) and uploader subscribers (importance score: 0.130) emerging as the most influential factors (Figure 7).

XSL•FO
RenderX

Figure 7. Top 10 feature importance ranking for video quality prediction model.

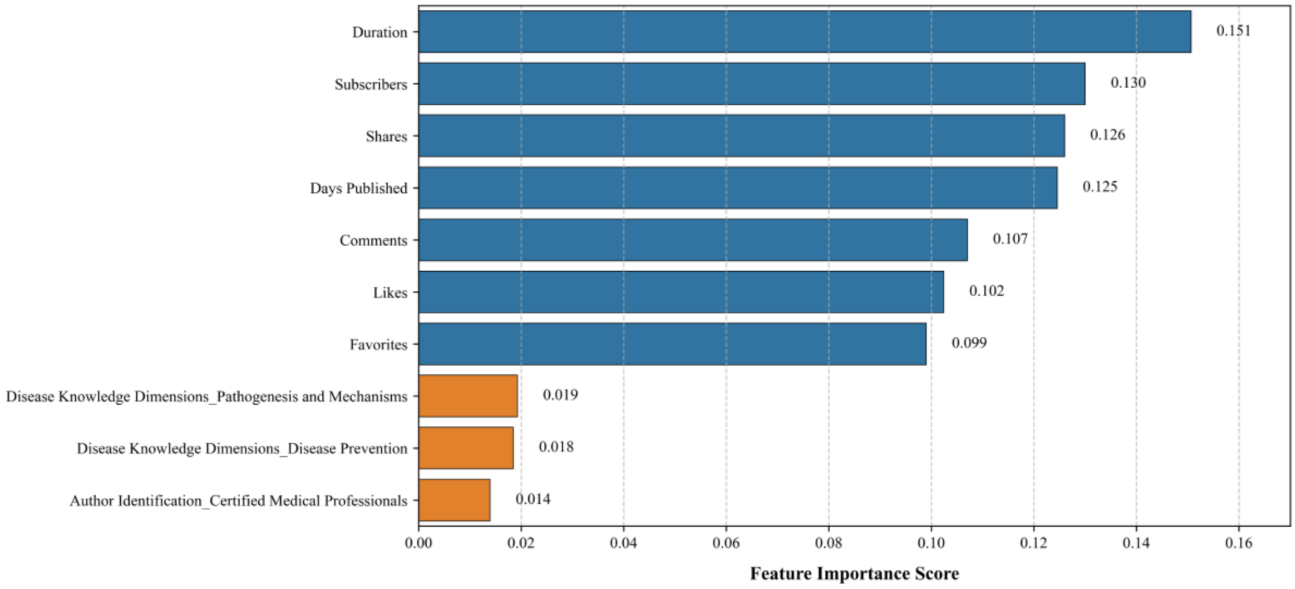

Discussion

Principal Findings

Social media platforms (YouTube, TikTok, and Bilibili) have become the main channel through which the public obtains health information; yet the quality of medically themed videos is generally low, especially in specialized fields such as stroke, brain tumors, and breast cancer [30-34]. Previous studies about stroke have taken a more focused, almost exclusively on, English-language YouTube content [35], and comparative quality data for Chinese short-video platforms (TikTok and Bilibili) are lacking, although their combined user base exceeds one billion. We therefore conducted the present study.

This cross-sectional study of stroke-related videos on Bilibili and TikTok revealed several critical findings. Most videos adopted a popular science education format, which was mainly created by certified medical professionals on TikTok and independent content creators on Bilibili. Notably, many Bilibili-based independent content creators were medical students and not platform-certified physicians. Quality assessments demonstrated relatively satisfactory GQS scores, whereas mDISCERN scores remained suboptimal, indicating gaps in content reliability. Evaluation of educational utility via PEMAT metrics revealed high understandability scores but low actionability benchmarks, suggesting that while these videos are broadly comprehensible to audiences, their practical applicability remains limited. Popular science education videos demonstrated superior audience comprehension and acceptance; those targeting posttreatment rehabilitation were more likely to provide actionable procedural guidance.

Characterization of Video Attributes and Viewer Engagement Patterns

Our study found that among the 306 analyzed videos, the highest engagement metrics reached 276,000 likes, 137,000 shares, and 90,000 favorites. This high level of audience interaction underscores the substantial public interest in stroke-related content. Additionally, the annual increase in stroke videos since 2022 reflects the growing recognition of stroke as a life-threatening condition, accompanied by gradual improvements in video quality [21].

Compared to Bilibili, TikTok videos generally achieved higher engagement metrics (likes, favorites, comments, and shares). While TikTok content predominantly originated from certified medical professionals and official entities, Bilibili videos were primarily authored by independent creators, indicating diverse stakeholder participation in stroke knowledge dissemination.

Studies have shown that short-form health education videos integrating infographics and plain-language explanations demonstrate higher audience engagement [36]. Content analysis revealed that popular science education videos dominated TikTok, whereas professional lecture videos (eg, academic conferences, guideline interpretations) prevailed on Bilibili. Notably, Bilibili videos were significantly longer than TikTok videos, aligning with its role as a long-form platform for specialized content. Thematic coverage spanned disease prevention, symptom identification, posttreatment rehabilitation, pathogenesis and mechanisms, and treatment options. This diversity fulfills the informational needs of the population at different disease stages, particularly enhancing pre-onset prevention awareness [13].

Furthermore, intergroup analyses based on classification criteria revealed significant differences in video characteristics across categories. Notably, posttreatment rehabilitation videos on TikTok had shorter days published than other disease knowledge categories, indicating a prioritization of post-stroke rehabilitation in current public health initiatives.

Science communicators demonstrated significantly higher engagement metrics (likes, favorites, comments, and shares) than certified medical professionals and official entities, though their content quality showed no statistical superiority. This discrepancy may be attributed to their larger subscriber base. A representative example includes a viral science communication video achieving 276,000 likes, 90,000 favorites, and 137,000 shares. Its success likely stems from an empathic

XSL•FO
RenderX

narrative approach that contextualizes stroke risks and prevention strategies within daily life scenarios [37].

Video Ratings and Quality Evaluation

This study utilized three validated assessment tools to evaluate the quality of all videos. Although most videos were published by certified medical professionals and official entities (eg, hospitals), the content's overall reliability, understandability, and actionability remained suboptimal. The GQS was employed to assess comprehensive quality, yielding a median score of 4 (range 1 - 5), which is higher than prior GQS evaluations of stroke-related videos [21].

The mDISCERN tool revealed moderate reliability scores (median 3), indicating acceptable but unsatisfactory credibility. Compared to previous studies analyzing videos associated with urinary tract infections, inflammatory bowel disease, and nonalcoholic fatty liver disease, stroke-related content demonstrated relatively better performance in these metrics [23-26]. Notably, TikTok videos shorter than 60 seconds exhibited significantly lower mDISCERN scores. Content analysis identified a critical gap: only 24.8% of videos cited additional information sources, resulting in zero scores for the mDISCERN criterion "Are additional sources listed for patient reference?" Adding authoritative references could enhance content credibility.

Furthermore, science communicators frequently disclosed information sources in video credits, possibly partially explaining their higher sharing rates. For instance, one science communication video achieved 137,000 shares alongside explicit citations of clinical guidelines.

The PEMAT-A/V was utilized to evaluate the educational value of stroke-related videos. The median PEMAT understandability score was 75%, while the actionability score was 33.3%. TikTok videos demonstrated significantly higher PEMAT understandability scores compared to Bilibili. Although a PEMAT score (PEMAT-U or/and PEMAT-A) above 70% indicates the content is generally reckoned to be comprehensible and actionable [27], both platforms exhibited critical deficiencies in actionability—a finding consistent with prior studies on mitral valve regurgitation-related videos [18].

Subgroup analyses revealed that videos from certified physicians achieved higher PEMAT actionability and understandability scores than those by independent content creators. Notably, videos focusing on posttreatment rehabilitation outperformed other disease knowledge categories in actionability. Additionally, popular science education videos scored significantly higher in GQS and PEMAT-A/V metrics than professional lectures.

Furthermore, some uploaders may incorporate inaccurate or misleading information to attract views. However, disseminating erroneous health content in short videos increases viewers' risk of making health decisions based on unverified claims [15]. Therefore, audiences should exercise caution when accessing health information online and prioritize content from verified medical sources.

Correlation Analysis and Model Prediction

Spearman's rank correlation analysis investigated the relationships between video characteristics and quality metrics. This method enabled a comprehensive assessment of how video quality correlates with fundamental features such as engagement and duration. Strong positive correlations were observed among likes, comments, favorites, shares, and uploader subscriber counts. Conversely, significant negative correlations existed between video duration and metrics, including likes, comments, and PEMAT understandability scores. While prior studies suggest that audiences gravitate toward authoritative health content with sustained engagement, our findings revealed no robust correlations between quality scores (GQS, mDISCERN, and PEMAT) and video features (eg, duration, engagement metrics) [37]. For instance, some high-quality videos exhibited below-median engagement, potentially attributable to uploaders' limited subscriber bases. Existing literature further implicates presentation style (eg, narrative framing) and emotional resonance as drivers of engagement, independent of content quality [37,38]. Machine learning models identified the top seven predictive features (all continuous variables: duration, likes, shares, etc), accounting for 83.7% cumulative importance in quality classification. Content attributes contributed minimally. Although baseline metrics partially reflect quality trends, neither feature-based nor content-driven models achieved clinically meaningful discriminative power. These results underscore the limitations of relying solely on engagement metrics or content taxonomies to evaluate stroke video quality.

Recommendations Based on Our Results

Undoubtedly, health-related videos often face greater challenges in maintaining audience engagement than entertainment content due to the cognitive demands of their specialized content. Previous studies have demonstrated compelling titles paired with accurate, high-quality content can significantly enhance viewer retention [37,39]. To amplify the impact of stroke-related videos, a viable strategy involves leveraging certified medical professionals and science communication teams with substantial followings to disseminate validated disease information. Platforms should prioritize distributing content from these verified sources to broader audiences, improving public stroke awareness.

Implementing standardized content review protocols and collaboration with neurology and rehabilitation specialists could further elevate the quality and reliability of health information [40]. Such measures would establish a trustworthy repository of actionable knowledge, strengthening public understanding of stroke prevention and management. Technologically, platforms could adopt AI-powered moderation systems driven by medical knowledge graphs, akin to the Dutch model, which achieved 89% accuracy in detecting misinformation during pilot testing [41].

Future research should focus on developing interdisciplinary evaluation frameworks that integrate the GQS with information diffusion dynamics models, enabling multidimensional assessments of both medical accuracy and societal impact [21,42]. Additionally, enhancing creator authentication mechanisms would bolster audience trust in professional content.

XSL•FO
RenderX

To bridge the pervasive actionability gap (PEMAT-A median=33 %), we propose a creator-side, platform-supported intervention bundle that can be implemented immediately without altering the existing subgroup taxonomy: (i) 30-second rule: open the video with a demonstrable action (eg, FAST test) that viewers can perform in real time; (ii) Check-box animation: end with an interactive tick-list of key treatment steps; viewers tap to complete, providing creators with instant engagement analytics; (iii) Time-critical QR code: prominently display a QR code linking to a county-level map of stroke or chest-pain centers (an application programming interface can be obtained by submitting a data-use request to the National Health Commission), ensuring that audiences know where to go within the golden time window; (iv) Certified medical template: platforms should offer an optional "Medical Science" upload wizard preloaded with (a) a simplified diagnostic and treatment flowchart validated by professional societies and (b) an embedded hospital map, reducing the risk of error by nonmedical creators; (v) Social shareable: automatically generate a "learning check" completion card that the audience can post to social media, amplifying reach and reinforcing behavioral memory. Platforms should strengthen their oversight of medical science videos: they could, for example, give greater traffic weight to content that achieves ≥60% PEMAT-A, steering viewers toward videos that are both popular and behavior-changing.

Optimal video duration should match its content complexity: excessively brief videos risk oversimplification, while lengthy ones may hinder comprehension. Beyond posttreatment rehabilitation, creators should incorporate tangible, step-by-step guidance to address pervasive actionability gaps observed across other stroke knowledge domains.

Practical Significance

To cope with stroke, one of the leading causes of mortality, we need to enhance public awareness of prevention and early recognition. In the digital era, individuals increasingly rely on video platforms to access health information and share experiences. TikTok and Bilibili, as dominant video-sharing platforms in China, have immense potential in disseminating health-related knowledge. This study represents the first comprehensive evaluation of stroke-related video content quality and reliability using multiple validated tools (GQS, mDISCERN, and PEMAT-A/V).

Our findings underscore that videos authored by certified medical professionals and official entities provide scientifically accurate and accessible content, which may improve public understanding of stroke prevention and disease management. However, persisting challenges remain, including insufficient actionable guidance and inadequate citation of credible sources, which collectively limit the practical utility of these materials. To further advance stroke literacy, platforms should prioritize distributing high-quality content while fostering collaborations among health care institutions, certified clinicians, and science communication teams to optimize content creation and dissemination strategies.

Limitations

This study has several limitations. First, the video search and data collection were exclusively conducted on Chinese-language video-sharing platforms (TikTok and Bilibili), which may limit the generalizability of findings to other platforms or cultural contexts. Second, the analysis included only top-ranked search results for the target keyword rather than the entire corpus of available content. Furthermore, TikTok's algorithm-driven content curation mechanisms introduce inherent variability in search result rankings, potentially compromising the representativeness of the sampled videos. Third, the cross-sectional design restricted the scope to a specific time window, failing to capture the temporal dynamics of content evolution and shifting engagement patterns on social media platforms. This temporal constraint may affect the long-term applicability and robustness of the conclusions. Future longitudinal studies are warranted to address these limitations. Fourth, the random-forest model achieved an AUC of only 0.577, indicating that readily available metadata (duration, likes, subscriber count, etc) are insufficient to predict video quality. This suboptimal discriminative performance, however, does not undermine our primary findings, which were derived from validated assessment tools (GQS, mDISCERN, and PEMAT-A/V) rather than from algorithmic prediction. The poor AUC actually reinforces our central message that engagement metrics should not serve as proxies for medical reliability. Future work will incorporate

Natural Language Processing–derived content features and creator-authority indicators to develop more accurate hybrid models. Additionally, as data were collected exclusively from Chinese-language videos on TikTok and Bilibili, our findings may not generalize to other linguistic or cultural contexts where platform algorithms, health-literacy levels, and clinical pathways differ substantially.

Conclusions

Stroke-related videos on TikTok and Bilibili exhibit variable characters, but both show unsatisfactory quality, particularly when delivering actionable guidance for audiences. Our subgroup analysis implies that certified medical professionals convey complex medical concepts more effectively than independent creators. In addition, videos about posttreatment rehabilitation have the best actionability compared to those about other knowledge dimensions. In the further correlation analysis, we find no significant correlation between basic characteristics and video quality. Our machine learning models also fail to establish accurate predictions for quality, so we suggest not relying solely on engagement metrics or content taxonomies to judge the video's credibility. For the public, click-heat does not equate to quality; for platforms, narrowing the gap between "what audiences watch" and "what audiences can actually learn" requires strengthened oversight of medical-science videos and traffic prioritization for videos that meet ≥60% PEMAT-A actionability. Only through such coordinated efforts—platform regulation, creator accountability—can short-video ecosystems contribute meaningfully to stroke popularization of science.

XSL•FO
RenderX

Acknowledgments

This research received no external funding. The APC was covered by the authors' institutional Academic Development Fund.

Disclaimer

No generative-AI tools (including but not limited to ChatGPT, Copilot, or Gemini) were used for drafting, data analysis, figure creation, or language generation. Because access to such services is restricted in our region, we employed GeenMedical [42], a commercial language-polishing platform, to improve readability. All changes were human-verified and further edited by the authors, who accept full responsibility for the final content.

Data Availability

The data for this study, derived from the TikTok platform, have been anonymized to protect privacy, with URLs and titles retained for research integrity. Because of privacy concerns, these data are not publicly accessible. However, we are committed to sharing the data upon request for legitimate research purposes in accordance with privacy protection principles and data-sharing policies. Interested researchers should contact the corresponding author for access.

Authors' Contributions

All authors contributed to the study's conception and design. SNN, SJN, and QCD were involved in the acquisition and interpretation of data. YYL and SJN were involved in the data analysis. SJN and JH drafted the manuscript. QCD critically revised the manuscript. SJN and QCD were involved in the study design and supervision. All authors read and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Video characteristics and quality assessment on TikTok and Bilibili across different categories.
[DOC File, 144 KB - formative_v9i1e80458_app1.doc]

References

1. Hilkens NA, Casolla B, Leung TW, de Leeuw FE. Stroke. Lancet 2024 Jun 29;403(10446):2820-2836. [doi: 10.1016/S0140-6736(24)00642-1] [Medline: 38759664]
2. Feigin VL, Stark BA, Johnson CO, et al. Global, regional, and national burden of stroke and its risk factors, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet Neurol 2021 Oct;20(10):795-820. [doi: 10.1016/S1474-4422(21)00252-0]
3. Ma Q, Li R, Wang L, et al. Temporal trend and attributable risk factors of stroke burden in China, 1990–2019: an analysis for the Global Burden of Disease Study 2019. Lancet Public Health 2021 Dec;6(12):e897-e906. [doi: 10.1016/S2468-2667(21)00228-0] [Medline: 34838196]
4. Wu S, Wu B, Liu M, et al. Stroke in China: advances and challenges in epidemiology, prevention, and management. Lancet Neurol 2019 Apr;18(4):394-405. [doi: 10.1016/S1474-4422(18)30500-3] [Medline: 30878104]
5. Ji C, Ge X, Zhang J, Tong H. The stroke burden in China and its long-term trends: insights from the Global Burden of Disease (GBD) Study 1990-2021. Nutr Metab Cardiovasc Dis 2025 Jul;35(7):103848. [doi: 10.1016/j.numecd.2025.103848] [Medline: 39948019]
6. Pandian JD, Gall SL, Kate MP, et al. Prevention of stroke: a global perspective. The Lancet 2018 Oct;392(10154):1269-1278. [doi: 10.1016/S0140-6736(18)31269-8]
7. Tan Y, Pan Y, Liu L, et al. One-year outcomes and secondary prevention in patients after acute minor stroke: results from the China National Stroke Registry. Neurol Res 2017 Jun;39(6):484-491. [doi: 10.1080/01616412.2017.1322804] [Medline: 28476094]
8. Yang J, Zheng M, Cheng S, et al. Knowledge of stroke symptoms and treatment among community residents in Western Urban China. J Stroke Cerebrovasc Dis 2014;23(5):1216-1224. [doi: 10.1016/j.jstrokecerebrovasdis.2013.10.019] [Medline: 24274934]
9. Zhong X, Wang J, He L, Xu R. Recognition of stroke-related knowledge among community residents and the improvement after intensive health education: a cross-sectional study. BMC Neurol 2020 Oct 12;20(1):373. [doi: 10.1186/s12883-020-01951-6] [Medline: 33045995]
10. Review of the new media ecosystem. : QuestMobile; 2024 URL: https://www.questmobile.com.cn/research/report/1871401635281997826 [accessed 2025-10-29]
11. Bilibili's Q4 and full-year financial report. : Bilibili; 2024 URL: https://ir.bilibili.com/media/s1dhpz1w/bilibili-inc-announces-fourth-quarter-and-fiscal-year-2024-financial-results.pdf [accessed 2025-10-27]

XSL•FO
RenderX

12. Lee C, Chow CM, Yan AT, Moe GW, Tu JV, Chu JY. Awareness of warning symptoms of heart disease and stroke: results of a follow-up study of the chinese canadian cardiovascular health project. CJC Open 2021 Jun;3(6):741-750. [doi: 10.1016/j.cjco.2021.01.014] [Medline: 34169253]
13. Yuan J, Li M, Liu Y, et al. Analysis of time to the hospital and ambulance use following a stroke community education intervention in China. JAMA Netw Open 2022 May 2;5(5):e2212674. [doi: 10.1001/jamanetworkopen.2022.12674] [Medline: 35579896]
14. Zhu C, Xu X, Zhang W, Chen J, Evans R. How health communication via Tik Tok makes a difference: a content analysis of Tik Tok accounts run by Chinese Provincial Health Committees. Int J Environ Res Public Health 2019 Dec 27;17(1):192. [doi: 10.3390/ijerph17010192] [Medline: 31892122]
15. McMullan RD, Berle D, Arnáez S, Starcevic V. The relationships between health anxiety, online health information seeking, and cyberchondria: systematic review and meta-analysis. J Affect Disord 2019 Feb 15;245:270-278. [doi: 10.1016/j.jad.2018.11.037] [Medline: 30419526]
16. Zhu Z, Liu S, Zhang R. Examining the persuasive effects of health communication in short videos: systematic review. J Med Internet Res 2023 Oct 13;25:e48508. [doi: 10.2196/48508] [Medline: 37831488]
17. Guan JL, Xia SH, Zhao K, et al. Videos in short-video sharing platforms as sources of information on colorectal polyps: cross-sectional content analysis study. J Med Internet Res 2024 Oct 29;26:e51655. [doi: 10.2196/51655] [Medline: 39470708]
18. Cui N, Lu Y, Cao Y, Chen X, Fu S, Su Q. Quality assessment of TikTok as a source of information about mitral valve regurgitation in China: cross-sectional study. J Med Internet Res 2024 Aug 20;26:e55403. [doi: 10.2196/55403] [Medline: 39163110]
19. Ding R, Kong Q, Sun L, et al. Health information in short videos about metabolic dysfunction-associated steatotic liver disease: analysing quality and reliability. Liver Int 2024 Jun;44(6):1373-1382. [doi: 10.1111/liv.15871] [Medline: 38441405]
20. Özdemir Z, Acar E. YouTube as a source of recognizing acute stroke; progress in 2 years. BMC Public Health 2024 Aug 13;24(1):2208. [doi: 10.1186/s12889-024-19710-4] [Medline: 39138572]
21. Moorhead SA, Hazlett DE, Harrison L, Carroll JK, Irwin A, Hoving C. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. J Med Internet Res 2013 Apr 23;15(4):e85. [doi: 10.2196/jmir.1933] [Medline: 23615206]
22. Kılınç DD. Is the information about orthodontics on Youtube and TikTok reliable for the oral health of the public? A cross sectional comparative study. J Stomatol Oral Maxillofac Surg 2022 Oct;123(5):e349-e354. [doi: 10.1016/j.jormas.2022.04.009] [Medline: 35470122]
23. Tam J, Porter EK, Lee UJ. Examination of information and misinformation about urinary tract infections on TikTok and YouTube. Urology 2022 Oct;168:35-40. [doi: 10.1016/j.urology.2022.06.030] [Medline: 35809700]
24. He Z, Wang Z, Song Y, et al. The reliability and quality of short videos as a source of dietary guidance for inflammatory bowel disease: cross-sectional study. J Med Internet Res 2023 Feb 9;25:e41518. [doi: 10.2196/41518] [Medline: 36757757]
25. Lai Y, He Z, Liu Y, et al. The quality and reliability of TikTok videos on non-alcoholic fatty liver disease: a propensity score matching analysis. Front Public Health 2023;11:1231240. [doi: 10.3389/fpubh.2023.1231240] [Medline: 37860803]
26. Shoemaker SJ, Wolf MS, Brach C. Development of the patient education materials assessment tool (PEMAT): a new measure of understandability and actionability for print and audiovisual patient information. Patient Educ Couns 2014 Sep;96(3):395-403. [doi: 10.1016/j.pec.2014.05.027] [Medline: 24973195]
27. Yeung A, Ng E, Abi-Jaoude E. TikTok and attention-deficit/hyperactivity disorder: a cross-sectional study of social media content quality. Can J Psychiatry 2022 Dec;67(12):899-906. [doi: 10.1177/07067437221082854] [Medline: 35196157]
28. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016 Jun;15(2):155-163. [doi: 10.1016/j.jcm.2016.02.012] [Medline: 27330520]
29. Liu H, Peng J, Li L, et al. Assessment of the reliability and quality of breast cancer related videos on TikTok and Bilibili: cross-sectional study in China. Front Public Health 2023;11. [doi: 10.3389/fpubh.2023.1296386]
30. Zheng S, Tong X, Wan D, Hu C, Hu Q, Ke Q. Quality and reliability of liver cancer-related short Chinese videos on TikTok and Bilibili: cross-sectional content analysis study. J Med Internet Res 2023 Jul 5;25:e47210. [doi: 10.2196/47210] [Medline: 37405825]
31. Zhang R, Zhang Z, Jie H, et al. Analyzing dissemination, quality, and reliability of Chinese brain tumor-related short videos on TikTok and Bilibili: a cross-sectional study. Front Neurol 2024;15:1404038. [doi: 10.3389/fneur.2024.1404038] [Medline: 39494168]
32. Liu Z, Chen Y, Lin Y, et al. YouTube/ Bilibili/ TikTok videos as sources of medical information on laryngeal carcinoma: cross-sectional content analysis study. BMC Public Health 2024;24(1):1594. [doi: 10.1186/s12889-024-19077-6]
33. Qiu J, Zhou YL. Quality assessment of heatstroke videos on TikTok. Front Public Health 2024;12:1446003. [doi: 10.3389/fpubh.2024.1446003] [Medline: 39296850]
34. Szmuda T, Alkhater A, Albrahim M, et al. YouTube as a source of patient information for stroke: a content-quality and an audience engagement analysis. J Stroke Cerebrovasc Dis 2020 Sep;29(9):105065. [doi: 10.1016/j.jstrokecerebrovasdis.2020.105065] [Medline: 32807469]
35. Xiao L, Min H, Wu Y, et al. Public's preferences for health science popularization short videos in China: a discrete choice experiment. Front Public Health 2023;11:1160629. [doi: 10.3389/fpubh.2023.1160629] [Medline: 37601206]

XSL•FO
RenderX

36. Tian K, Hao L, Xuan W, Phongsatha T, Hao R, Wei W. The impact of perceived value and affection on Chinese residents' continuous use intention of mobile health science information: an empirical study. Front Public Health 2023;11:1034231. [doi: 10.3389/fpubh.2023.1034231] [Medline: 36844825]
37. Amini F, Riche NH, Lee B, Leboe-McGowan J, Irani P. Hooked on data videos. 2018 May 29 Presented at: AVI '18; May 29, 2018; Castiglione della Pescaia Grosseto Italy URL: https://dl.acm.org/doi/proceedings/10.1145/3206505 [doi: 10.1145/3206505.3206552]
38. Dobransky K, Hargittai E. Inquiring minds acquiring wellness: uses of online and offline sources for health information. Health Commun 2012;27(4):331-343. [doi: 10.1080/10410236.2011.585451] [Medline: 21932982]
39. Van de Belt TH, Engelen L, Berben SAA, Teerenstra S, Samsom M, Schoonhoven L. Internet and social media for health-related information and communication in health care: preferences of the Dutch general population. J Med Internet Res 2013 Oct 2;15(10):e220. [doi: 10.2196/jmir.2607] [Medline: 24088272]
40. Charnock D, Shepperd S, Needham G, Gann R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. J Epidemiol Community Health 1999 Feb;53(2):105-111. [doi: 10.1136/jech.53.2.105] [Medline: 10396471]
41. Peng Y, Wu X, Atkins S, et al. Internet-based health education in China: a content analysis of websites. BMC Med Educ 2014 Jan 27;14(1):16. [doi: 10.1186/1472-6920-14-16] [Medline: 24467710]
42. GeenMedical. URL: http://www.geenmedical.com [accessed 2025-10-27]

Abbreviations

AUC: area under the curve
DALY: disability-adjusted life-years
GQS: Global Quality Scale
ICC: intraclass correlation coefficient
mDISCERN: modified DISCERN
PEMAT-A/V: Patient Education Materials Assessment Tool for Audio-Visual Content.

Edited by A Mavragani; submitted 10.07.25; peer-reviewed by J Zhang, T Xia; revised version received 09.10.25; accepted 09.10.25; published 03.11.25.

Please cite as:
Nie SJ, Ning SN, Ding QC, Hou J, Luo YY, Wang HY
Quality Analysis of Stroke-Related Videos on Video Platforms: Cross-Sectional Study
JMIR Form Res 2025;9:e80458
URL: https://formative.jmir.org/2025/1/e80458
doi:10.2196/80458

© Shao-Jie Nie, Shuai-Nan Ning, Qi-Chao Ding, Jie Hou, Yue-Yang Luo, Hai-Yang Wang. Originally published in JMIR Formative Research (https://formative.jmir.org), 3.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Associations Between Online Search Trends and Outpatient Visits for Common Medical Symptoms in the United States from 2004 to 2019: Time Series Ecological Study

Ryan Heumann[1,2], MS; Steven R Steinhubl[2], MD

[1]Indiana University School of Medicine, 340 W 10th St, Indianapolis, IN, United States

[2]Weldon School of Biomedical Engineering, Purdue University West Lafayette, West Lafayette, IN, United States

Corresponding Author:
Ryan Heumann, MS
Indiana University School of Medicine, 340 W 10th St, Indianapolis, IN, United States

Abstract

Background: The seemingly endless amount of information available on the internet at the touch of a few buttons has increasingly served as a resource for individuals to find health information over the last 20+ years. Previous research in the field of infodemiology suggests that this change in use of the internet by the public to find health information has an impact on their interaction with the health care system. However, to this date, no study has directly juxtaposed the growth of internet searching and the number of visits to ambulatory care centers for the same concerns over the last two decades of internet growth.

Objective: This study aimed to explore the relationship between online searches for three common primary care concerns and ambulatory visits for those same concerns during 2004 - 2019.

Methods: The National Ambulatory Medical Care Survey and National Hospital Ambulatory Medical Care Survey data were searched between the years 2004 and 2019 in order to estimate the number of annual ambulatory care visits for cough, sore throat, and abdominal pain. Google Trends data were explored to analyze the number of people searching for information about the same complaints over the same years. Cross correlation, time series regression, and Granger causality analysis were performed to analyze patterns and causality.

Results: Google Trends data showed that the number of searches for the common primary care symptoms "cough," "sore throat," and "stomach pain" in the United States grew by 208%, 290%, and 490%, respectively, between 2004 and 2019. However, over the same time, United States population-adjusted outpatient visits for cough and sore throat decreased by 41.5% and 40%, respectively, and stomach pain visits remained unchanged. According to conservative estimation, analysis concluded that an increase in internet searches for cough can predict a decline in ambulatory visits for that symptom.

Conclusions: This supports that, on a population level, exploring online health information about some common, acute symptoms did not lead to increase in care seeking, but instead, suggests that a substantial portion of the population found the information to be reassuring or informative enough to not feel the need to seek care from a primary health care provider. With the rapid evolution and availability of more detailed and personalized information from various large language models, it is likely that internet search habits of users will continue to grow, and with it, continue to transform interactions with the health care system.

(JMIR Form Res 2025;9:e77274) doi:10.2196/77274

KEYWORDS

internet; health information; patient education; ambulatory care; primary care; infodemiology; digital health

Introduction

The term "infodemiology" was created in 2002 by Gunther Eysenbach in an attempt to describe an approach for revolutionizing public health by using data monitoring and mining [1-3]. It is defined as "the science of distribution and determinants of information in an electronic medium, specifically the internet, or in a population, with the ultimate aim to inform public health and public policy" [3]. The underlying theory is that changes in information and communication patterns may either reflect or cause changes in population health [3]. The field is further divided into supply-based infodemiology, that is the study of the amount and quality of information on the internet, and demand-based infodemiology, that is the study of search patterns in a population [3]. This discussion will center around the latter and its relationship with ambulatory care seeking.

One of the original infodemiology studies, by Eysenbach [3] in 2006, showed that internet searches preceded visits to a primary care physician by one week, suggesting that people often consulted the internet first before seeing a physician. Since then,

XSL•FO
RenderX

the number of global internet users and with it, the volume of searches has grown exponentially—from 200 million a day in 2004 to a projected 13.6 billion in 2025—and so has the ability of the internet to affect health care [4,5]. A study by Finney et al [6] found that the number of individuals who had used the internet first in their most recent search for health information grew from 61.2% in 2008, to 69.6% in 2011, and to 74.9% in 2017. Furthermore, the Health Information National Trends Survey (HINTS) in 2022 found that, of respondents who use the internet, 85.3% have used it to look for health information in the past year [7].

What qualifies as health information in this context is not well-defined, in part because the motivations individuals have for using the internet in order to inform a health-affecting decision are varied. While the data are limited, several interview-based studies find that some (~60% according to one study [8]) turn to the internet first to determine whether their symptoms seem serious enough to warrant visiting a doctor, with the hope being that they can find reassurance and save time and money [8,9]. If they fall into the ~40% of people who are typically reassured by an online health information search, the road might end there [10]. However, if they are not reassured, or even fall into the ~30% of people who tend to be more worried after a search, then they typically make an appointment with a physician [9,10]. In this case, a separate motivation to search for information might be to explore possible diagnoses, symptoms, treatments, or better understand medical jargon in order to be best prepared for their visit [9,11]. After a visit, individuals have cited clarifying physician instruction, finding a second opinion, and reading forums of others with their diagnosis as further motivations for searching the internet [9,12]. In fact, in 2022, 54% of survey respondents found internet information helpful to understanding information given to them by a physician [8]. No matter what the motivation, only 8% of people believe that using the internet has an adverse effect on their relationship with their physician and one of every four people think it has a positive impact [13].

It has been theorized that the internet empowers patients by giving them the ability to actively seek information, which improves their agency and future interactions with their health care provider [14]. This sense of empowerment and control has been shown to lead to more discussion with physicians and better ability to elicit information during consultations [14]. This generally leads to better treatment outcomes and higher levels of patient satisfaction. Patient-physician communication can also improve from this re-balanced division of knowledge, which can additionally lead to more recall and understanding of information and improved emotional state [14].

We sought to explore the relationship between the known increase in internet search volume for health information and outpatient care seeking over the last two decades in the United States. To do this, we compared the trajectory in the yearly number of online searches for specific acute symptoms with that of the number of people seeking outpatient care for that same symptom that same year. Though previous research discussed here offers motivations for internet searching as well as occasional estimates of impact on care, none analyze the change in behaviors throughout the evolution of the internet in society. We hypothesized that the increasing use of online health information could potentially impact peoples' care-seeking behavior in one of several ways.

First, it is possible that as online search volume goes up, so does care seeking. While this could potentially be due to an increase in the prevalence of a symptom over time, we selected three symptoms—cough, sore throat and stomach pain—that despite yearly fluctuations would not be expected to increase on a year-to-year basis. That would leave an alternative explanation for a potential yearly increase in outpatient visits for that symptom, that online information seeking heightened a majority of users' concerns and prompted them to seek care when they might not otherwise have done so. Second, it is also possible that there could be an inverse relationship between the search volume for a specific symptom and care seeking for that same symptom. This might suggest that the information found online reassured the user enough to decide not to seek care.

Methods

United States Outpatient Visit Data

The National Ambulatory Medical Care Survey (NAMCS) and National Hospital Ambulatory Medical Care Survey (NHAMCS) from the Division of Health Care Statistics of the National Center for Health Statistics, Centers for Disease Control and Prevention were used to determine common primary care reasons for visit. These surveys are sent out to a sample of US-based physicians, and responses are used to create country-wide estimations, with standard error reported, in various domains. The NAMCS analyzes office-based health care providers while the NHAMCS analyzes emergency departments (EDs) and hospital-based outpatient centers. Taken together, the aim is to provide a comprehensive picture of ambulatory care use in the United States for any given year [15].

For the purposes of this analysis, data collection was done only until 2019 in order to not analyze data that might have been skewed by factors attributable to the COVID-19 pandemic that began in 2020 such as quarantine and decreased availability of ambulatory care sites. In the NAMCS, the top 20 patient-reported reasons for visit were available for every year from 2000 to 2019 with the exception of 2017. In 2017, a change in collection methods led to difficulty in processing the data, which has indefinitely postponed the release of data [16]. In all 19 years of data available, cough, symptoms referable to the throat, and stomach pain, cramps, and spasms were listed in the top 20 reasons for visit.

The NHAMCS provides separate reports for ED visits and hospital-based outpatient visits. Data for each year from 2000 to 2019 were available for ED visits, but only until 2011 for hospital-based outpatient visits. The hospital-based outpatient data from 2012 to 2017 was not released due to quality assurance issues while the survey was not conducted in 2018 - 2019 because of budget constraints [16]. The top 20 patient-reported reasons for visit were collected until 2007 when ED data started only recording the top 10.

XSL•FO
RenderX

Internet Search Data

To provide data, Google Trends finds the percentage of all searches for a certain time frame that were for the input term. It then scales these percentages to the same time period in which that input term had the highest percentage of searches, giving an array of values from 0 to 100, with 100 being the most popular month and 0 being the least [17]. Google Trends was searched for the terms "cough," "sore throat," and "stomach pain" from January 1st, 2004 (the earliest data available) to December 31, 2019 (to eliminate most COVID-19 data) in the United States across all categories. The results were displayed as a graph with relatively popularity of the term available monthly from 2004 to 2019.

A secondary resource, Glimpse, which provides data similar to Google Trends, was used in attempt to convert relative popularity to gross number of individual searches for the same time period. Glimpse was used to find the gross number of searches in the month that was the most popular according to Google Trends (relative popularity of 100). Then Google Trends data was used to calculate the gross number of searches for the rest of the months. Because Glimpse provides very limited information about their methods, graphs were compared to Google Trends data, which provides much more transparency. When comparing relative popularity of search terms reported by Glimpse and Google Trends, the Pearson correlation coefficients were 0.99 for cough, sore throat, and stomach pain, indicating that the data were consistent enough between platforms to justify using Glimpse in a very limited capacity.

To decide the query term analyzed, the goal was to choose a term that a patient would reasonably search when feeling the given symptom. The query "cough" was searched to correspond to cough for obvious reasons. For symptoms referable to the throat, this was taken to mean sore throat, so that was the query analyzed. For stomach pain, cramps, and spasms, the decision was made to choose "stomach pain" because it seemed the most general and colloquial.

Statistical Analysis

NAMCS and NHAMCS visits were adjusted to estimate the number of visits that would have occurred given the 2019 population in order to more equally compare them with internet search data that is already adjusted for population change. The number of visits according to the NAMCS and NHAMCS data were aggregated for each year to provide one measure of total yearly ambulatory visits. Each month of Google Trends data was aggregated for a given year as well. The result was a list of yearly internet searches and a list of yearly ambulatory visits for cough, sore throat, and abdominal pain for each year from 2004 - 2019.

In the cases in which one of the three symptoms of interest were not reported in the top 10 or top 20 symptoms of NHAMCS data, one of two paths was taken. In the first method, maximum estimation, the number of visits for the symptom of interest would be recorded as the number of visits for the least common symptom that was still listed. Because the interest is in analyzing the decreasing trend in visits, using this number, that represents the maximum number of visits the symptom of interest could have been, provides a conservative estimate. The other method consisted of using linear interpolation to estimate the missing data. Analysis was done using both sets of data and the more conservative method was selected in order to prevent spurious conclusions. To determine significance in the change of number of ambulatory visits from 2004 to 2019 for each symptom, a two-tailed *z* test was performed. Standard error was reported along with the visits from the NAMCS and NHAMCS data and was used during analysis.

For both the data using maximum estimation and linear interpolation estimation, stationarity was determined for each symptom for search data and ambulatory visit data. Once stationarity was achieved (using differencing or double differencing), cross correlation, time series regression, and Granger causality tests were performed to analyze the relationship between internet searches and ambulatory visits for each of the three symptoms of interest.

Ethical Considerations

All information used in this manuscript was publicly available and deidentified. Because we did not perform any human or animal subjects research, ethics board approval was not applied for per the Institutional Review Board of Purdue University.

Results

Findings from NAMCS and NHAMCS Data

For cough and sore throat, there was a significant decrease over time in the population-adjusted number of outpatient visits reported by NAMCS and NHAMCS data (Figure 1). Between 2004 and 2019, there was a 41.5% decrease in the number of visits for cough (~35 million to ~20.4 million, 95% CI −22,050,103 to −7,030,485, $P<.001$) and a 40% decrease in the number of visits for sore throat (~20.7 million to ~12.4 million, 95% CI −14,654,695 to −1,921,239, $P<.01$). During this same period, the number of searches reported by Google Trends increased for cough by 208% from ~15.7 million in 2004 to ~48.4 million in 2019 and by 290% for sore throat from ~6.7 million to ~26.2 million. While online searches for stomach pain increased by 490%, from ~3.2 million to ~19 million, visits remained relatively consistent, with a non-significant decrease of 6.3% from ~24.8 million in 2004 to ~23.2 million in 2019 (95% CI −7,494,982 to 4,372,021, $P=.61$; Figure 1).

XSL•FO
RenderX

Figure 1. Estimated visits for cough (**A**), sore throat (**B**), and stomach pain (**C**) based on NAMCS/NHAMCS data compared to Google Trends cumulative popularity for the corresponding query terms. NAMCS: National Ambulatory Medical Care Survey; NHAMCS: National Hospital Ambulatory Medical Care Survey.

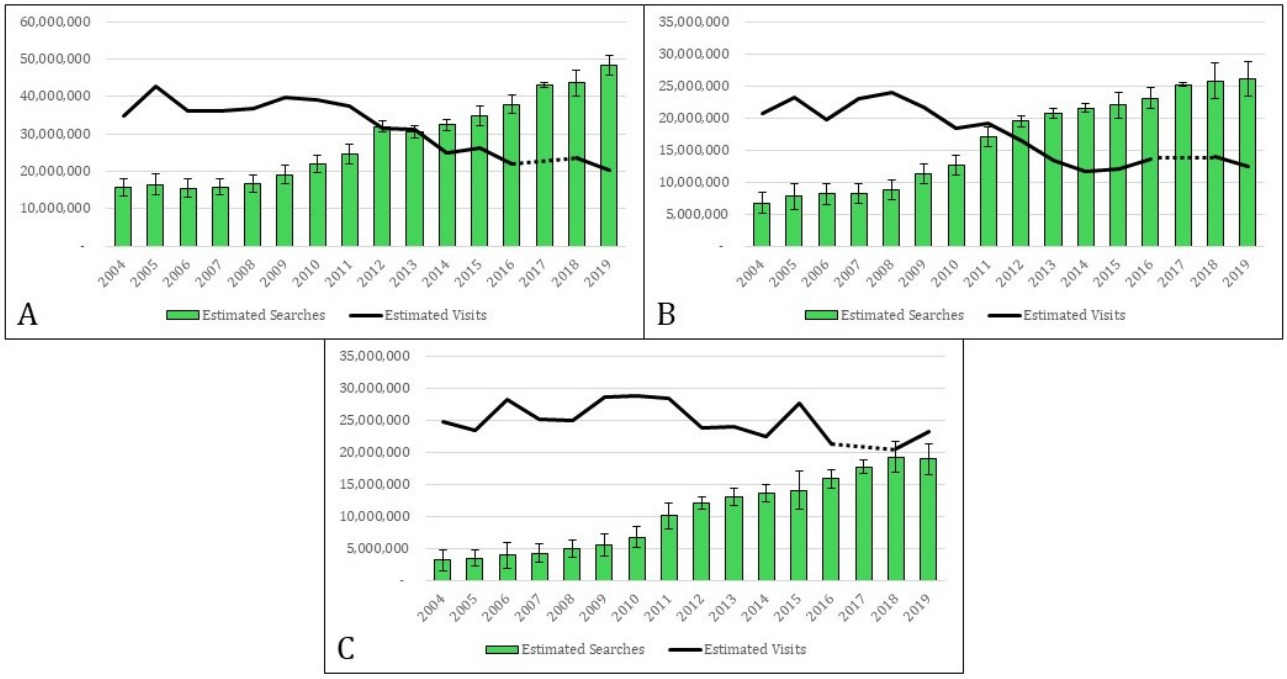

Cross Correlation

When using linear interpolation estimation, cross correlation was insignificant for cough, but significant for abdominal pain with an R value of −0.595 (*P*=.025) at a lag of 5 years (the pattern in internet searches followed the pattern of clinical encounters 5 years prior). However, when using maximum estimation, cross correlation became insignificant for all three symptoms (Table 1).

Table . Summary of cross correlation, time series regression, and Granger causality testing for each method of estimating missing data.

Symptom	Cross correlation	Regression coefficient	Granger causality
Linear interpolation estimation			
Cough	No (*P*=.061)	−0.84 (*P*<.01[a])	Yes (*P*=.020)[a]
Sore throat	No (*P*=.09)	−1.32 (*P*=.043)[a]	Yes (*P*=.045)[a]
Abdominal pain	Yes (*r*=−.595, *P*=.025)[a]	−2.49 (*P*=.014)[a]	No (*P*>.99)
Maximum estimation			
Cough	No (*P*=.064)	−0.84 (*P*=.011)[a]	Yes (*P*=.026)[a]
Sore throat	No (*P*=.098)	−1.48 (*P*=.111)	No (*P*=.139)
Abdominal pain	No (*P*=.079)	−2.67 (*P*=.042)[a]	No (*P*>.99)

[a]Significant at α=.05

Time Series Regression

The linear interpolation estimates led to significant time series regression models for all three symptoms. For cough, sore throat, and abdominal pain, the model coefficients were −0.84,−1.32, and −2.49, respectively. Using maximum estimates, sore throat became insignificant, but cough and abdominal pain remained significant with coefficients of −0.84 and −2.67, respectively (Table 1).

Granger Causality

Linear interpolation estimation led to Granger causality (the number of searches can help predict number of visits) for cough and sore throat, but not abdominal pain. The maximum estimation also led to Granger causality for cough, but not for sore throat or abdominal pain (Table 1).

Discussion

Principal Findings and Comparison With Previous Works

For all three acute symptomatic concerns studied, after adjusting for the ~35.5 million person population increase in the United States between 2004 and 2019, the yearly internet search volume in the United States increased by 2- to 5-fold over that 15-year

XSL•FO
RenderX

period. During that same period, the number of people seeking outpatient care for that same primary complaint significantly decreased by ~40% for two of the three symptoms and remained unchanged for the third.

For cough, there was no significant cross correlation, but there was significant time series regression and Granger causality when using maximum estimation. Therefore, while internet searches and ambulatory visits did not quite change together over time significantly (P=.064), there is still evidence that an increase in internet searches for cough can predict a decline in ambulatory visits for that symptom.

For sore throat, cross correlation, time series regression, and Granger causality were all insignificant for the maximum estimation method. This indicates that there is not enough evidence using this data that internet searches for sore throat correlate with or predict ambulatory visits for the same concern.

Finally, for abdominal pain, time series regression was significant, but cross correlation and Granger causality were not. From this, it can be said that there may be a weak association between internet searches and ambulatory visits, especially when considering that the cross correlation was close to significance (P=.079), but there is not evidence presented to say definitively that one predicts the other.

Because of the discrepancy between the "eyeball test" that shows internet searches for a symptom increasing over the same time period that ambulatory visits for that symptom are decreasing, and the statistical analysis that had mixed results, it seems evident that there are other variables interacting that have not been taken into account.

It is still clear that internet search volume for these common health symptoms has increased over time from 2004 to 2019, which is not surprising in and of itself. Prior research has shown that a higher percentage of people are turning to the internet first when they have a health question because it is viewed as accessible, direct, effective, and quick [9]. They often view it as a way to save time and money by not having to see a physician [9]. The concurrent decrease in ambulatory visits, at least for cough and sore throat, seem to indicate that these searches are reassuring the general public and not driving an increase in seeking outpatient care, but the analysis done here does not quite prove such a straightforward picture. Possibly as evidence of this, internet searches for stomach pain increase nearly 5-fold, but the population-adjusted number of outpatient visits for that complaint remained stable.

One alternative explanation is that the inverse trends are not being driven by reassurance provided by internet searches, but by necessity to overcome financial and accessibility barriers. In other words, ambulatory visits are not decreasing because people are finding satisfactory answers online, but instead because they do not have a practical ability to see a doctor anyway. In fact, in 2019, 33% of respondents of a Gallup survey said that in the last year, they or a family member of theirs put off medical treatment because of the cost associated with it [18]. People with this barrier, or a similar barrier like insurance coverage or access to a close care center, may be using the internet because it is the only option they have. In this case, we would expect to see the trend seen here, but not necessarily with statistical analysis showing a direct predictive relationship between internet searches and clinical visits.

Overall, there is an undeniable inverse trend in internet searches and ambulatory visits recorded by NAMCS and NHAMCS data, at least for cough and sore throat. However, the statistical analysis indicates that, while there is some evidence that searches themselves are impacting ambulatory visits, possibly through patient reassurance, the entire story is not captured here. It is likely that there are other influencing factors that have not been accounted for and could feasibly play a role.

Future Directions

Just as the last two decades saw rapid change in internet tools and use, the current search habits of individuals are in constant flux. Today, the digital landscape is being populated more and more with artificial intelligence (AI) tools, like large language models (LLM) or AI "chatbots," that are already being widely used by patients to find health care information [12,19]. According to one survey from 2024, 64% of adults have used AI and 17% claim to use AI chatbots at least once a month to find health information [20]. The ability of LLMs to understand human queries, access a huge body of information, then synthesize findings into accurate natural language responses in easy-to-understand terms allows the delivery of customized content to users [21,22]. Patients are able to find information about conditions and treatments in a way that mimics human interaction, but without worry of physician availability or judgment [23]. The pursuit of agentic AI models that are able to essentially "think ahead" take these already impressive abilities one step further. They would be capable of long-term memory and longer reasoning chains that allow initiative taking instead of just response to user input [24]. It seems likely that LLM generated answers to health queries, which in some studies have already been shown to be of higher quality and show more empathy than physician's online answers to questions, will drive even greater change in internet search habits of users, and potentially completely transform interactions with the health care system [25].

Limitations

Google Trends accounts for individual IP addresses, but it is still not possible to prove that each search is due to one individual feeling a symptom for themself. It is entirely possible that they are searching on behalf of another person or simply out of curiosity alone.

Moreover, it is possible that not all clinical outpatient visits were recorded. Though the two surveys seek to encompass office-based, hospital-based office visits, and ED visits, it is subject to sampling bias. Additionally, although certain tables in the NAMCS dataset mention "urgicenters", urgent care visits seem largely unaccounted for in the data. Telemedicine visits are also not mentioned explicitly in NAMCS and NHAMCS data, leading to ambiguity about their inclusion. Missing NAMCS data for 2017 and NHAMCS data for 2012 - 2017 could additionally skew analysis. Another limitation exists in not capturing all search queries made for the related symptoms.

XSL•FO
RenderX

Differences in phrasing could lead to exclusion from this Google Trends data.

Seasonal variability is to be expected in both search volume and clinical encounters for all three symptoms. However this cyclical nature should not have any impact in comparing year to year trends.

Finally, coding and delayed care could be altering data in ways that are not apparent. As coding practices change over time, it is possible that NAMCS and NHAMCS data have changed with it.

Conclusions

The internet as a whole has greatly evolved over the last 20+ years and today is at the beginning of undergoing its greatest change since its inception with the incorporation of LLMs. With that change comes change in how users interact with it, including when it comes to searching for health information. There are motivations for internet searches at every step in the health care visit process, which includes even before deciding to go see a physician at all. As internet search volume for health information has increased exponentially over the last two decades, it had the potential to drive either an increase or a decrease in visits to health care professionals. What has been illustrated here is that, in the case of cough, stomach pain, and sore throat in the United States in 2004 - 2019, a 2- to 5-fold increase in internet search volume was associated with either a decrease or no change in outpatient visits. This lends evidence to the theory that the internet has provided a place for individuals, on a population-level, to find reassurance for their symptoms to the point where they feel it is unnecessary to see a health care provider. However, it is likely that there are other influencing factors affecting these inverse trends that have not been fully elucidated by this analysis and more work should be done to highlight these. As the ability and availability of AI continues to grow, the impact of these tools on the use of health care resources must be addressed to offer patients the best possible care.

Data Availability

All analysis is based on publicly available, de-identified, aggregate data.

Authors' Contributions

Conceptualization, investigation, writing - original draft: RAH
Conceptualization, investigation, project administration, supervision, writing - review & editing: SRS

Conflicts of Interest

RAH declares no conflicts. SRS is an advisor to Eko Health, Tempus Labs, and Google Health.

References

1. Bragazzi NL. Infodemiology and infoveillance of multiple sclerosis in Italy. Mult Scler Int 2013;2013:924029. [doi: 10.1155/2013/924029] [Medline: 24027636]
2. Mackey T, Baur C, Eysenbach G. Advancing infodemiology in a digital intensive era. JMIR Infodemiology 2022;2(1):e37115. [doi: 10.2196/37115] [Medline: 37113802]
3. Eysenbach G. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the internet. J Med Internet Res 2009 Mar 27;11(1):e11. [doi: 10.2196/jmir.1157] [Medline: 19329408]
4. Cardillo A. Exploding topics. How Many Google Searches Are There Per Day. 2025 Mar. URL: https://explodingtopics.com/blog/google-searches-per-day [accessed 2025-04-29]
5. Kemp S. More than 5 billion people now use the internet. We Are Soc. 2022. URL: https://wearesocial.com/hk/blog/2022/04/more-than-5-billion-people-now-use-the-internet/ [accessed 2025-04-24]
6. Finney Rutten LJ, Blake KD, Greenberg-Worisek AJ, Allen SV, Moser RP, Hesse BW. Online health information seeking among US adults: measuring progress toward a healthy people 2020 objective. Public Health Rep 2019;134(6):617-625. [doi: 10.1177/0033354919874074] [Medline: 31513756]
7. National Institutes of Health. Health information national trends survey. : National Cancer Institute; 2022 URL: https://hints.cancer.gov/ [accessed 2025-10-28]
8. NORC at the university of chicago. GSS Data Explorer. URL: https://gssdataexplorer.norc.org/variables/vfilter [accessed 2024-12-27]
9. Huisman M, Joye S, Biltereyst D. Searching for health: doctor Google and the shifting dynamics of the middle-aged and older adult patient-physician relationship and interaction. J Aging Health 2020 Oct;32(9):998-1007. [doi: 10.1177/0898264319873809] [Medline: 31517558]
10. Van Riel N, Auwerx K, Debbaut P, Van Hees S, Schoenmakers B. The effect of Dr Google on doctor-patient encounters in primary care: a quantitative, observational, cross-sectional study. BJGP Open 2017 May 17;1(2):bjgpopen17X100833. [doi: 10.3399/bjgpopen17X100833] [Medline: 30564661]

XSL•FO
RenderX

11. Ling R, Lee J. Disease monitoring and health campaign evaluation using Google search activities for HIV and AIDS, stroke, colorectal cancer, and marijuana use in Canada: a retrospective observational study. JMIR Public Health Surveill 2016 Oct 12;2(2):e156. [doi: 10.2196/publichealth.6504] [Medline: 27733330]
12. Klimiuk KB, Balwicki Ł. What is infodemiology? An overview and its role in public health. Przegl Epidemiol 2024 Jun 7;78(1):81-89. [doi: 10.32394/pe/188119] [Medline: 38904314]
13. Tan SSL, Goonawardene N. Internet health information seeking and the patient-physician relationship: a systematic review. J Med Internet Res 2017 Jan 19;19(1):e9. [doi: 10.2196/jmir.5729] [Medline: 28104579]
14. Broom A. Virtually he@lthy: the impact of internet use on disease experience and the doctor-patient relationship. Qual Health Res 2005 Mar;15(3):325-345. [doi: 10.1177/1049732304272916] [Medline: 15761103]
15. Cherry DK, Woodwell DA. National ambulatory medical care survey: 2000 summary. Adv Data 2002 Jun 5(328):1-32. [Medline: 12661586]
16. Centers for disease control. National Center for Health Statistics Notices for NAMCS and NHAMCS Public Use Data File Users. 2024. URL: https://archive.cdc.gov/www_cdc_gov/nchs/ahcd/notice-archive.htm [accessed 2025-04-21]
17. FAQ about google trends data. SupportGoogle.com. URL: https://support.google.com/trends/answer/4365533?hl=en [accessed 2025-01-24]
18. Gallup. Healthcare System. URL: https://news.gallup.com/poll/4708/Healthcare-System.aspx [accessed 2024-12-22]
19. Kohane IS. Compared with what? Measuring AI against the health care we have. N Engl J Med 2024 Oct 31;391(17):1564-1566. [doi: 10.1056/NEJMp2404691] [Medline: 39465890]
20. Presiado M, Montero A, Lopes L, Published LH. KFF health misinformation tracking poll: artificial intelligence and health information. KFF. 2024. URL: https://www.kff.org/health-misinformation-and-trust/poll-finding/kff-health-misinformation-tracking-poll-artificial-intelligence-and-health-information/ [accessed 2024-10-06]
21. Pan A, Musheyev D, Bockelman D, Loeb S, Kabarriti AE. Assessment of artificial intelligence chatbot responses to top searched queries about cancer. JAMA Oncol 2023 Oct 1;9(10):1437-1440. [doi: 10.1001/jamaoncol.2023.2947] [Medline: 37615960]
22. Rosner S, Vassileva M. How thoughtful AI implementation can rebuild trust in healthcare. 2025. URL: https://hitconsultant.net/2025/01/24/how-thoughtful-ai-implementation-can-rebuild-trust-in-healthcare/ [accessed 2025-01-29]
23. Tandar CE, Bajaj SS, Stanford FC. Social media and artificial intelligence-understanding medical misinformation through Snapchat's new artificial intelligence chatbot. Mayo Clin Proc Digit Health 2024 Jun;2(2):252-254. [doi: 10.1016/j.mcpdig.2024.04.004] [Medline: 38962215]
24. Zou J, Topol EJ. The rise of agentic AI teammates in medicine. The Lancet 2025 Feb;405(10477):457. [doi: 10.1016/S0140-6736(25)00202-8]
25. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023 Jun 1;183(6):589-596. [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]

Abbreviations

AI: artificial intelligence
ED: emergency department
HINTS: Health Information National Trends Survey
LLM: large language model
NAMCS: National Ambulatory Medical Care Survey
NHAMCS: National Hospital Ambulatory Medical Care Survey

Edited by A Mavragani; submitted 10.05.25; peer-reviewed by A Hungbo, J Igboanugo; revised version received 27.07.25; accepted 23.09.25; published 04.11.25.

Please cite as:
Heumann R, Steinhubl SR
Associations Between Online Search Trends and Outpatient Visits for Common Medical Symptoms in the United States from 2004 to 2019: Time Series Ecological Study
JMIR Form Res 2025;9:e77274
URL: https://formative.jmir.org/2025/1/e77274
doi:10.2196/77274

© Ryan Heumann, Steven R Steinhubl. Originally published in JMIR Formative Research (https://formative.jmir.org), 4.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License

XSL•FO
RenderX

(https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Medical Information Provided by Transgender and Gender-Diverse Content Creators on YouTube: Descriptive Content Analysis

Lydia Bliss[1], BA; Qianqian Zhao[2], MS; Irene Chao[1], BS; Oliver L Haimson[3], PhD; Ellen Selkie[1], MD, MPH

[1]Department of Pediatrics, School of Medicine and Public Health, University of Wisconsin–Madison, 800 University Bay Drive, Office 300-21, Madison, WI, United States

[2]Department of Biostatistics and Medical Informatics, School of Medicine and Public Health, University of Wisconsin–Madison, Madison, WI, United States

[3]School of Information, University of Michigan, Ann Arbor, MI, United States

Corresponding Author:
Lydia Bliss, BA
Department of Pediatrics, School of Medicine and Public Health, University of Wisconsin–Madison, 800 University Bay Drive, Office 300-21, Madison, WI, United States

Abstract

Background: Transgender and gender-diverse (TGD) individuals frequently turn to social media to find community, express their identities, and access essential information. These platforms are easily accessible to TGD people and enable health information–seeking in anonymous, identity-affirming spaces outside of traditional health care systems. As a result, social media has become a critical source of health information on topics like gender-affirming care for TGD individuals, specifically for TGD youth. YouTube, one of the most widely used social media platforms, is especially popular for its long-form videos made by content creators who have built dedicated followings on the platform. Among them are TGD content creators, many of whom make content documenting their medical transition and gender identity journey and provide general information about TGD topics. TGD creator content therefore makes YouTube an important platform for health education for TGD individuals.

Objective: This study aims to describe the health-related content shared by TGD content creators on YouTube. Specifically, we characterize the medical topics addressed, the frameworks used to discuss these topics, and the valence of creators' health care experiences.

Methods: A descriptive content analysis was performed on 2485 videos posted by 42 self-identified TGD YouTube content creators. Videos were systematically evaluated for mentions of gender-affirming care and other health-related topics. We also examined whether creators framed medical information using personal narratives or an informational approach and if they characterized their medical experiences as positive, negative, or neutral.

Results: Most videos (n=1724, 69.4%) created by TGD content creators did not include discussions related to gender identity or transitioning. However, among the videos that did address gender identity (n=761, 30.6%), mentions of medical topics were prevalent (n=554, 72.8%). Of videos that discussed medical topics, gender-affirming surgeries (n=356, 64.3%) and hormone replacement therapy (n=307, 55.4%) were the most frequently discussed. Other commonly discussed medical topics included mental health (n=131, 23.6%) and sexual health (n=96, 17.3%). Videos covering medical topics primarily centered on personal experiences (n=411, 74.2%), with content creators often characterizing these experiences positively (n=224, 73.2%).

Conclusions: This study highlights the breadth of health-related information shared by TGD content creators on YouTube. Our findings underscore the role of long-form video content on YouTube as an educational resource for TGD people, offering health information that is both easy to access and grounded in lived experience. Clinicians can use these findings to better understand the health information that their TGD clients are likely to encounter online, fostering more informed and supportive conversations about gender-affirming care.

(JMIR Form Res 2025;9:e75787) doi:10.2196/75787

KEYWORDS

transgender; transgender youth; gender-affirming care; YouTube; social media; content creators; influencers; social media health information

Introduction

Transgender and gender-diverse (TGD) individuals have been known to use social media to fulfill a variety of social, emotional, and informational needs. TGD individuals use social media as a vital space for seeking support, connecting with others, and developing their identity, often curating their online environments for these purposes [1-6]. These behaviors are

XSL•FO
RenderX

especially significant for TGD youth, who use the internet and social media more than their cisgender peers, particularly as a resource for exploring their gender identity [7-10]. Importantly, online spaces improve information accessibility, enabling individuals to seek advice about gender identity that may be difficult to obtain through traditional means or uncomfortable to discuss with offline support networks [11,12].

Given the significant barriers that TGD individuals often encounter in accessing medical care, social media plays a critical role in providing a place to address health-related questions, specifically about gender-affirming care [10]. Seeking health information online has become a preferred method for many TGD individuals due to its accessibility, convenience, affordability, and the option for anonymity [12-14]. Social media provides safe spaces for community members to share information while offering diverse formats for consuming it, such as real-time answers in online forums or archived content on platforms like YouTube [13]. Furthermore, TGD individuals have highlighted the value of social media for accessing firsthand accounts of those who have undergone gender-affirming treatments [13].

Seeking health-related information online is particularly common among TGD youth. Herrmann et al [4] found that all TGD adolescent participants used social media or the internet to seek information on transgender-related topics, most commonly gender-affirming medical interventions. Selkie et al [15] reported similar findings, with TGD adolescents identifying informational support as a helpful aspect of social media, specifically for topics like gender-affirming care. Importantly, in both studies, YouTube was identified as the most common web-based platform for TGD adolescents to seek transgender-specific content [4,15].

YouTube, a global video-sharing platform, is the most commonly used social media platform among Americans, used by 83% of adults and 93% of teens [16,17]. As the second-most visited site in the world (behind only Google), YouTube has proven to be a dominant and enduring platform in the social media landscape [18]. Compared to other social media platforms, YouTube is uniquely suited to serve as an educational resource. YouTube videos tend to be significantly longer than videos shared on other platforms and can range from a few minutes to over an hour. These long-form videos allow for the sharing of more detailed and in-depth videos where creators can share their experiences, opinions, and knowledge through a variety of genres, including video blogs, story times, and commentary. This extended format supports the sharing of health-related content through a variety of informative styles, often centered on sharing personal experiences [19].

Much of the popular content available on YouTube is made by those who have gained a substantial online following. In this paper, they are referred to as “content creators” or simply “creators,” although they are also commonly known as “YouTubers” or “influencers.” Content creators build a following through their self-branded content and are important figures online: they are drivers of internet discourse, create spaces for like-minded viewers, and disseminate content that is seen as relatable to their audience. As such, content creators are known to cultivate strong fan communities, frequently made up of younger viewers [20].

A content creator’s ability to remain both authentic and personable, while simultaneously maintaining the reputation of that of a celebrity figure, serves to amplify their influence among their audience [21]. Familiar content creators have been shown to be perceived as more trustworthy and similar to oneself than traditional celebrities, showing that they may be particularly able to influence behaviors among audiences seeking validation and guidance [22]. Because these creators are often regarded as important role models to their viewers, they are uniquely positioned to serve as influential health educators [23]. Bandura’s [24] Social Cognitive Theory of Mass Communication posits that individuals learn by observing media figures and that health media messaging is most effective when it promotes self-efficacy—the belief in one’s ability to exercise control over their own health. As personable and relatable figures, content creators may promote a sense of empowerment in viewers, influencing their health behaviors.

TGD content creators have a considerable presence on YouTube and are known to make content about gender-affirming care. A popular format for sharing this information is the transition vlog—a video blog where creators document their physical and emotional changes resulting from gender-affirming care interventions such as hormone replacement therapy or surgery [25]. This content has been shown to be valuable to TGD viewers, who watch these transition vlogs to learn about transition processes and experiences [26].

Although a number of scholars have examined TGD content and experiences on YouTube, only one study to date has systematically analyzed the content made by popular TGD content creators [27-38]. In an analysis of the content made by 8 TGD creators, Miller [38] found that 53.5% of their videos provided educational insights on transgender topics, with 34.2% of videos discussing elements of physical transition, most commonly hormones or surgery. However, the breadth and frequency of medical topics covered by TGD content creators remains unexplored at a large scale. Further research is needed to assess the health-related information shared by these creators, given the diversity of TGD contributors and the vast range of content on the platform. It is also important to examine how this information is framed and how creators portray their medical experiences as this may shape how viewers interpret and respond to their content.

Although online content is recognized as an important resource for TGD individuals, limited research has examined the actual content provided. Content analysis is a commonly used method in social media research to characterize and describe health-related information online [39]. It is especially useful for examining the rhetoric, topics, and emerging themes in online content [10,38,40,41]. This approach allows for the direct observation of material that social media users encounter; as it does not rely on participants’ memories or interpretations, content analysis offers a more accurate and unobtrusive method for evaluating online health information.

Given YouTube’s popularity and the influence of its creators, understanding the platform’s medical content is essential. This

XSL•FO
RenderX

exploratory study aims to describe the health-related content shared by TGD content creators on YouTube by analyzing the medical topics they address, the frameworks they use to discuss these topics, and the valence of their health care experiences.

Methods

Study Design

A descriptive content analysis study was conducted between June 2021 and October 2024. This analysis was part of a larger study evaluating the videos posted by TGD content creators discussing gender identity and transitioning. This analysis focused specifically on videos discussing medical topics.

Content Creator Identification

Recommendations for TGD content creators were first obtained from a large convenience sample of TGD young people, researchers, and health professionals serving TGD individuals in June 2021. This approach prioritized community relevance in the selection of creators and avoided reliance on YouTube's recommendations, which risked introducing algorithmic bias. Recommended content creators were individually evaluated for their internet presence and following across YouTube, Instagram, Twitter (now X), and/or TikTok. Accounts on each platform were evaluated against 4 criteria: content in English, identification as TGD by the creator, at least 10,000 followers, and a minimum of 10 posts within the two years prior to data collection. Content creators with a YouTube channel and at least 1 social media account on any platform that met all 4 criteria were selected for further investigation. We included creators with large followings on other social media platforms, even if they did not meet the follower threshold on their specific YouTube channels. This preserved our intent to examine long-form video content while acknowledging that popular influencers do not limit their presence to one platform. Moreover, research indicates that even content creators with relatively small followings (commonly referred to as micro-influencers) can still have a substantial influence over their audience, comparable to those with larger followings [42,43].

The research team was comprised of academic researchers and gender-affirming care clinicians representing diverse gender identities, including lived TGD experience. The team met regularly to evaluate the initial group of YouTube accounts meeting the inclusion criteria. The accounts were reviewed for post history, content type, and the overarching narrative of the channel. The research team reviewed each account to reach consensus on whether to include or exclude the creator from the final analysis. Since our objective was to examine content specific to gender identity and transitioning, creators were determined as fit for inclusion if they posted regularly about these topics. Excluded creators focused on other topics within the LGBTQ+ community (such as advocacy), posted about other topics entirely (such as music), or posted minimally in recent years. This process yielded a total of 42 eligible YouTube content creators.

Ethical Considerations

Because the data were publicly accessible, this study did not constitute human subjects research and did not require review by the University of Wisconsin–Madison Institutional Review Board [44,45]. However, analyzing public social media data raises important ethical considerations. Although content posted to social media may be technically public, Buck and Ralston [46] argue that this does not automatically make online content "public data" in a research context. They note that social media users hold different expectations of privacy depending on how or where they share their content. Accordingly, privacy should instead be understood in terms of the user's intended audience and their expected reach, rather than their content's technical availability [46]. Such expectations vary significantly among different users; while average social media users may reasonably expect their posts to remain within their personal networks, content creators typically create videos with the specific intention of reaching a broader audience, suggesting a lower expectation of privacy.

Still, while content creators often aim to reach larger audiences, they may reasonably expect their content to stay within TGD communities or among supportive viewers. When amplified beyond these circles, they can become targets of significant online harassment, especially within online spaces that have been known to be hostile toward the TGD community [47-49]. Publicly identifying a list of individual TGD creators in this paper could increase risk for targeted harassment; therefore, anonymizing their identities is essential to minimizing potential harm.

To protect creators' privacy, we took several precautions. First, we avoided using automated scraping tools for data collection, instead using human coders to manually collect metadata and analyze video content. Second, all identifiable information, such as creator names, channel names, and video titles, were deidentified for reporting. Example quotes reported in this paper are paraphrases rather than direct quotes from creators. Lastly, our analysis focused on characterizing the types of content present across the full body of videos, rather than on the individual creators themselves.

Codebook Development

A team-based, iterative approach was used to develop a structured codebook for content analysis, consistent with prior content analyses [50-52]. As to not interfere with the coding of chosen accounts, a sample of videos were pulled from the channels of creators that were not selected for final analysis but whose content closely resembled the content made by the included creators. The core research team evaluated these posts to identify variables related to gender identity development and transitioning in both medical and nonmedical contexts. Once the preliminary codebook was created, members of the research team independently applied the codebook to an additional set of nonstudy posts. The research team regularly convened to review coding agreements and disagreements, refining the codebook with emergent themes and clarifying existing codes until reaching consensus. The full, expanded codebook (including nonmedical topics not included in this analysis) is available in Multimedia Appendix 1.

XSL•FO
RenderX

Coder Training and Agreement

Five coders were trained by evaluating 20 nonstudy posts and comparing their coding to that of a lead researcher [50]. Coding of study data began after each of the 5 coders reached 80% agreement or higher with the lead researcher across all codes and posts. All coding disagreements were discussed to refine coder interpretation and ensure accurate application of the codebook.

During active periods of study coding, paired indexing and weekly team meetings were conducted to periodically assess interrater agreement and resolve questionable codes by consensus [53-55]. Any necessary refinements to the codebook were discussed and implemented. During data collection, interrater agreement was assessed between the 2 primary coders who analyzed the majority of the dataset, coding 534 of 761 (70.2%) of the included videos. Agreement was calculated based on a subsample of 10 jointly coded videos. Results indicated strong reliability: prevalence-adjusted and bias-adjusted κ=0.93 (SE=0.0087), 95% (CI 0.9118-0.9460); Gwet AC1=0.96 (SE=0.0051), 95% CI (0.9497-0.9697). Multimedia Appendix 2 presents the percent observed agreement, prevalence-adjusted and bias-adjusted κ, and Gwet AC1 for each code.

Content Coding

Overview

Coding of YouTube videos was conducted from January 2023 to February 2024. Coders reviewed one channel at a time from the 42 YouTube content creators selected for analysis. Each channel was reviewed by 1 coder, which ensured continuity and context in coding this large dataset. Videos from each channel were reviewed in order of popularity, using YouTube's sorting feature to organize the videos by view count. To ensure theoretical saturation, coders reviewed every video on the account until 20 videos that met inclusion criteria were coded. To meet inclusion criteria, a video had to explicitly mention topics related to gender identity or the medical, social, or legal aspects of transitioning. The number of likes and views were collected for videos meeting inclusion criteria. The number of subscribers of each channel was recorded after the content coding of all channels was complete.

Videos were analyzed based on the creators' personal mentions of their gender identity, the medical topics that they addressed, the frameworks used to discuss medical topics, and the valence of their medical experiences. Coders used Qualtrics to record their evaluations of each video. Codes were marked as present or not present; for codes marked as present, coders explained their decisions in a corresponding textbox.

Medical Topics

Discussions of medical topics were categorized into subtopics: puberty and androgen blockers, hormone replacement therapy, menstruation, surgery, mental health, sexual health, fertility and family planning, treatment regrets, and general mentions of medical topics (Table 1). Medical topics were coded if they were discussed in the video at least once.

XSL•FO
RenderX

Table . Codebook of medical topics identified in YouTube videos made by selected transgender and gender-diverse content creators.

Code	Definition	Subcodes[a]
Puberty and androgen blockers	Discussions about medications to delay puberty or block androgens. Includes personal experience and/or general information.	N/A[b]
Hormone replacement therapy	Discussions about hormone replacement therapy. Includes personal experience and/or general information.	Estrogen, testosterone, not specified
Surgery	Discussions about surgical interventions for gender-affirming care. Includes personal experience and/or general information.	Top surgery (gender-affirming chest surgery), bottom surgery (gender-affirming genital surgery), facial feminization surgery, Brazilian butt lift, general mentions, other
Mental health	Discussions about mental health diagnoses, treatments, and therapies. Includes personal experience and/or general information.	Medication, treatment, and therapy; therapists/psychologists; reparative/conversion therapy; suicidality; eating disorders; substance use disorder/addiction; posttraumatic stress disorder; other diagnoses; general mentions
Menstruation	Discussions about the management of periods and symptoms associated with menstruation. Includes personal experience and/or general information.	N/A
Sexual health	Discussions about safe sex practices, sexual boundaries, sexual function, sexually transmitted infections, and consent. Includes personal experience and/or general information.	N/A
Fertility and family planning	Discussions about the ability to conceive or have biological children or alternative plans for creating a family, such as adoption or surrogacy. Includes personal experience and/or general information.	N/A
Treatment regrets	Expressing regret about undergoing a certain treatment or regretting a certain aspect of that treatment. Includes personal experience and/or general information.	N/A
General mentions	General mentions of medically transitioning and gender-affirming health care, without naming a specific treatment. Includes personal experience and/or general information.	N/A

[a]Codes were marked as present or not present.

[b]N/A: not applicable.

Medical Discussion Frameworks

All discussions about medical topics were categorized by framework. Discussions were categorized as “personal” when the creator’s statements were self-referential, usually by sharing personal experiences, feelings, and opinions. Discussions that presented facts, advice, or explanations were categorized as “informational.” Discussions that could not be categorized as personal or informational were categorized as “other” (Table 2).

XSL•FO
RenderX

Table . Codebook of the framework and valence of medical topics discussed by selected transgender and gender-diverse YouTube content creators.

Framework	Definition	Valence (if applicable)[a]
Personal	Statements are self-referential. Creator shares personal experiences, feelings, and opinions.	Positive (affirming experiences), negative (disaffirming experiences), or neutral/not applicable (neither positive nor negative)
Informational	Statements are not self-referential. Creator presents information as facts, explanations, or advice.	N/A[b]
Other	Statements are neither personal nor informational.	N/A

[a]Valence assessed for creators' personal discussions of surgical and pharmacological interventions directly related to gender-affirming care.
[b]N/A: not applicable.

Valence of Medical Experiences

Valence was assigned to creators' personal discussions of surgical and pharmacological interventions directly related to gender-affirming care. Valence refers to the creators' characterization of their medical experiences, with positive indicating desirable or affirming experiences, and negative indicating undesirable or disaffirming experiences. Experiences were categorized as being positive, negative, or neutral/not applicable (Table 2). Figure 1 illustrates example quotes coded according to our codebook.

Figure 1. Example quotes from transgender and gender-diverse content creator videos with corresponding codes.

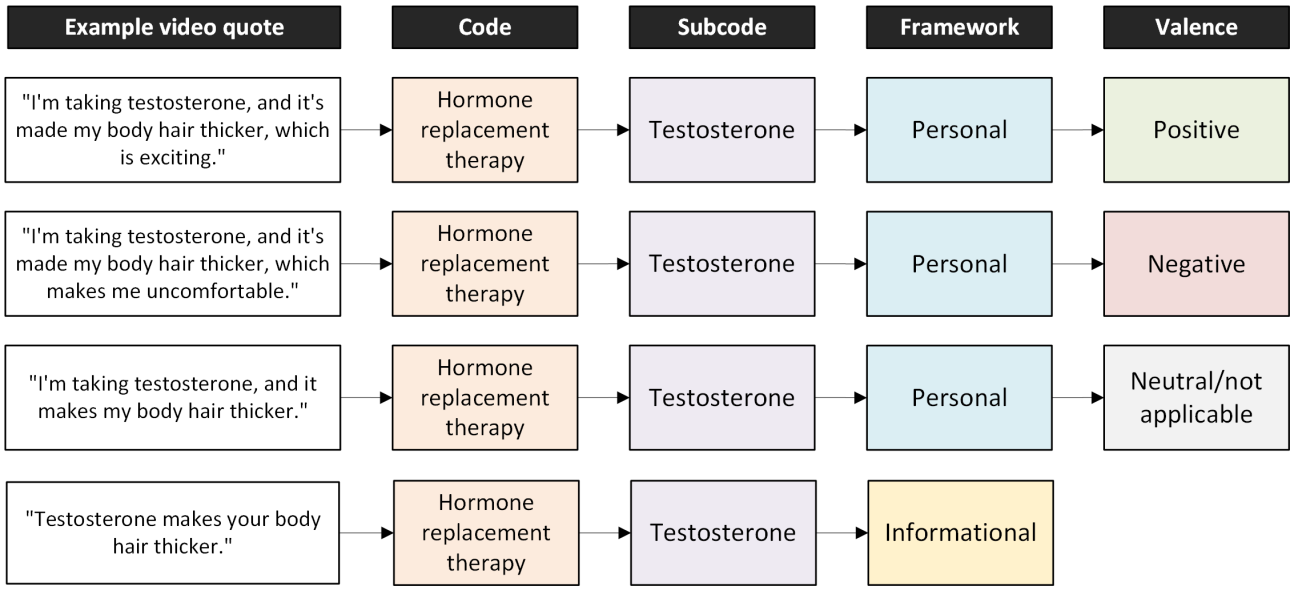

Results

A total of 42 content creators were included in the analysis. Figure 2 presents a flowchart of the creator selection process.

XSL•FO
RenderX

Figure 2. Transgender and gender-diverse content creator selection process.

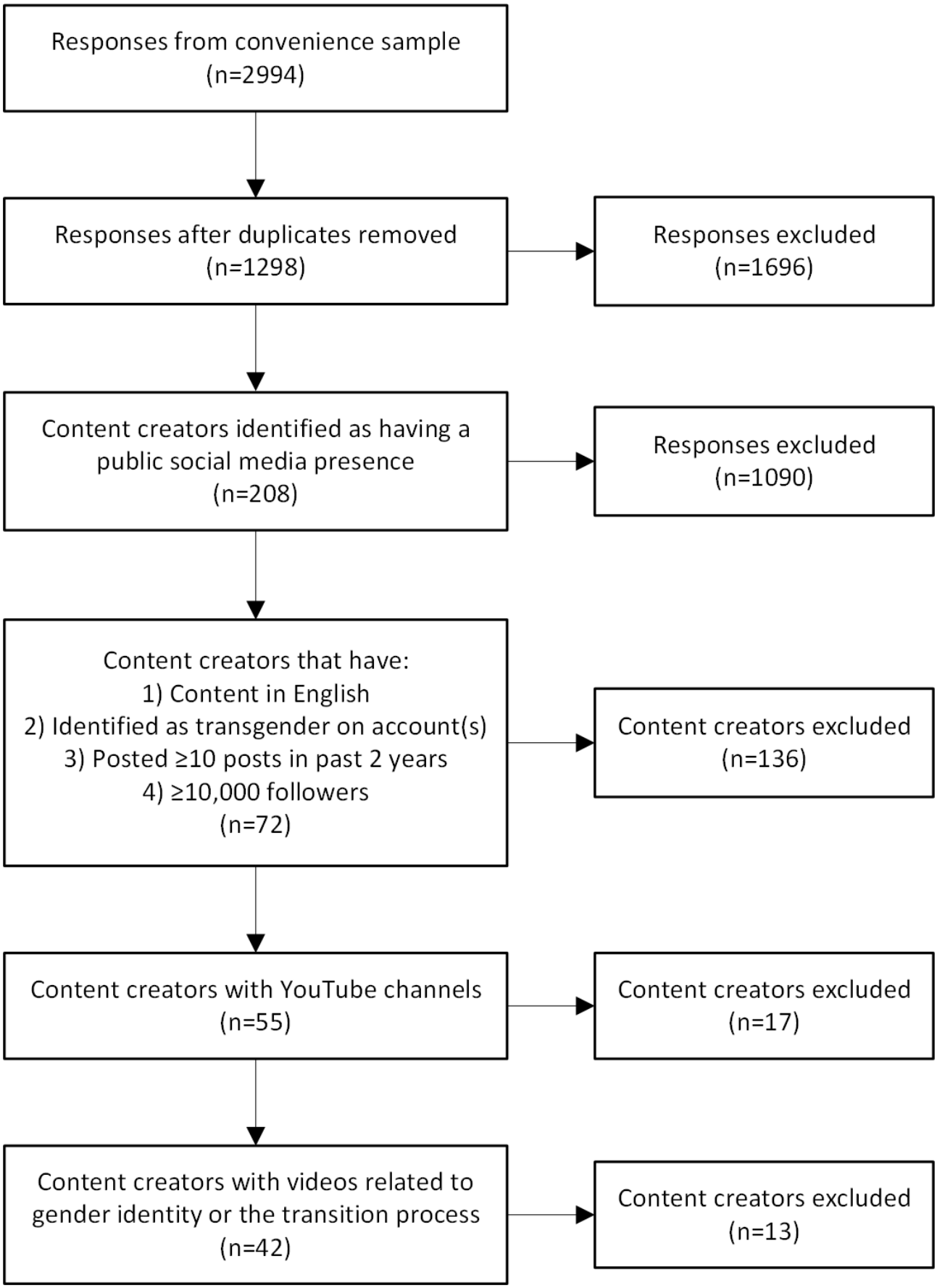

XSL•FO
RenderX

Creator Gender and Channel Engagement

Gender Identity

Gender identity labels that creators used to describe themselves in more than 1 video were used in analysis. Gender identities reported are not mutually exclusive, to reflect creators who identified with multiple genders. Creators described their personal gender identities as 45.2% transgender male (n=19), 21.4% transgender female (n=9), 23.8% transgender (n=10), 21.4% nonbinary (n=9), 7.1% gender queer (n=3), and 9.5% additional gender identities (n=4). In addition, 9.5% of creators did not specify their gender identities more than once (n=4).

Subscribers

The number of subscribers to the creators' YouTube channels varied widely, ranging from 160 to 3,410,000. The mean number of subscribers was 492,976 (SD 799,317), while the median was 117,000 (IQR 618,800).

Video Engagement

The average number of likes and views for all videos in the sample (n=761) was 24,997 (SD 42,192) and 702,961 (SD 2,303,544), respectively. Given the skewed distribution of engagement metrics, medians and IQRs are also reported (Tables 3,4). Patterns were largely consistent when comparing videos with and without discussions of medical topics, with medical topic videos showing slightly higher average engagement but similar medians. A full comparison of likes and views for all videos, medical-topic videos, and nonmedical videos is available in Tables 3,4.

Table . Likes across all videos, medical-topic videos, and nonmedical videos made by selected transgender and gender-diverse YouTube content creators.

Video type	Mean (SD)	Median (IQR)
All videos in sample (n=761)	24,997 (42,192)	8650 (29,400)
Nonmedical (n=207)	24,521 (31,706)	9300 (40,700)
Medical (n=554)	25,176 (45,518)	8600 (25,700)

Table . Views across all videos, medical-topic videos, and nonmedical videos made by selected transgender and gender-diverse YouTube content creators.

Video type	Mean (SD)	Median (IQR)
All videos in sample (n=761)	702,961 (2,303,544)	221,711 (702,974)
Nonmedical (n=207)	560,002 (784,340)	237,785 (768,292)
Medical (n=554)	756,377 (2,655,731)	216,342 (703,509)

Medical Topics

Of the videos analyzed (n=2485), 30.6% (n=761) of videos included discussions directly related to gender identity or transitioning. Of these videos, 72.8% (n=554) discussed medical topics at least once. Videos that did not discuss medical topics addressed other topics related to gender identity and the transition process, including nonmedical affirming activities (eg, clothing, makeup), social experiences, and legal sex and name changes (Multimedia Appendix 1). Further analysis included videos that discussed a medical topic at least once.

The most frequently discussed medical topics were surgery (n=356, 64.3%) and hormone replacement therapy (n=307, 55.4%; Table 5). Testosterone was the most frequently discussed type of hormone replacement therapy (n=205, 37%), while top surgery was the most frequently discussed type of surgical intervention (n=193, 34.8%). Other commonly discussed medical topics include mental health (n=131, 23.6%) and sexual health (n=96, 17.3%).

XSL•FO
RenderX

Table . Proportion of videos made by selected transgender and gender-diverse YouTube content creators regarding specific medical topics (N=554).

Code and subcode		Number of videos[a]	Percent of videos[b]
Puberty and androgen blockers		48	8.7
Hormone replacement therapy		307	55.4
	Estrogen	31	5.6
	Testosterone	205	37
	Not specified	92	16.6
Menstruation		24	4.3
Surgery		356	64.3
	Top surgery (chest surgery)	193	34.8
	Bottom surgery (genital surgery)	111	20
	Facial feminization surgery	34	6.1
	Brazilian butt lift	7	1.3
	General mentions	116	20.9
	Other	13	2.3
Mental health		131	23.6
	Medication, treatment, and therapy	36	6.5
	Therapists/psychologists	5	0.9
	Reparative/conversion therapy	10	1.8
	Suicidality	29	5.2
	Eating disorders	3	0.5
	Substance use disorder/addiction	5	0.9
	Depression	40	7.2
	Anxiety	21	3.8
	Other diagnoses	5	0.9
	General mentions	33	6
Sexual health		96	17.3
Fertility and family planning		36	6.5
Treatment regrets		9	1.6
General mentions		69	12.5

[a]Medical topics were not mutually exclusive. Videos discussing more than one topic are included in each relevant category. Videos that discussed a medical topic at least once were counted.

[b]Compared to all videos discussing medical topics (N=554).

Medical Discussion Frameworks

Of all videos where creators discussed medical topics (n=554), 74.2% of videos (n=411) included discussions categorized as personal, 55.8% of videos (n=309) included discussions categorized as informational, and 70.9% of videos (n=393) included discussions categorized as other. Medical frameworks were not mutually exclusive. Videos using more than one framework were included in each relevant category.

Valence of Medical Experiences

Of all videos where creators shared their personal experiences with gender-affirming care (n=306), 73.2% of videos (n=224) discussed positive experiences, 33.7% (n=103) discussed negative experiences, and 61.1% (n=187) discussed experiences that were neutral or not applicable. Positive, negative, and neutral categories were not mutually exclusive, since videos could include discussions of multiple experiences with different valences. Videos that discussed medical experiences with different valences were included in each relevant category.

Discussion

Relevance

Given the unique health needs and challenges faced by TGD individuals, access to health information is essential to support their well-being and empower them to make informed decisions about their care. However, TGD individuals experience significant barriers to accessing health services and are less

XSL•FO
RenderX

likely to use both primary and specialty care services compared to cisgender individuals [56-58]. These challenges are even greater for the estimated 113,900 transgender youth (over one-third of transgender youth in the United States) who live in a state that has enacted a ban on gender-affirming care [59]. Additionally, an executive order issued by the Trump administration that aims to end federal funding for health care organizations providing gender-affirming care reflects a broader effort to restrict access to such care for transgender youth in the United States [60]. Without accessible systems in place, TGD individuals may turn to alternative sources for health information, such as social media. It is therefore important to understand the medical content available on popular platforms such as YouTube, which is widely used among TGD youth [4,15].

Principal Results

This study described the health-related content made by TGD content creators on YouTube. This study found that most content created by transgender creators did not center on transgender-specific topics. Instead, these creators showcased multiple aspects of their lives beyond their gender identity across their videos. This variety of content may allow TGD followers to connect with creators over additional shared interests, fostering a stronger sense of kinship with the creator, enhancing their relatability and influence.

When TGD content creators discussed their gender identity, medical topics frequently emerged. These findings indicate that creators commonly highlight the role of medical care in affirming their gender identity. This suggests that TGD individuals engaging with such content, particularly for gender identity development purposes, are likely to encounter information about gender-affirming health care.

Surgery and hormone replacement therapy were common topics, with testosterone and top surgery discussed most frequently. Other commonly discussed medical topics included mental health and sexual health. This range of topics highlights a holistic perspective of gender-affirming care, with discussions extending beyond physical interventions to include broader aspects of health.

Most videos including discussions of medical topics focused on personal experiences, aligning with previous research that highlights social media's role in sharing firsthand medical narratives [13,19]. Compared to the smaller sample sizes of previous studies, our findings indicate that this trend is evident across the wider genre of TGD medical content. Such videos offer insights grounded in lived experience, providing information from those who have undergone gender-affirming treatments, which may be especially helpful for informing viewers' medical decisions. Notably, the content often reflected positive experiences, suggesting that content creators tend to report or emphasize favorable outcomes from gender-affirming care.

Implications

Although it is known that TGD YouTubers discuss medical topics, prior studies have largely relied on small samples. To our knowledge, no study has systematically categorized this content at scale. Our study contributes new insights by using a larger dataset and a more granular, structured codebook to identify the frequency and framing of medical discussions across a wide range of videos.

These findings may be especially insightful for clinicians working with TGD individuals. Clinicians are likely aware that their clients are viewing online videos related to their experiences. Providers may find it useful to ask clients about specific stories they may have seen on social media and use their responses to build a discussion about individual goals and expectations for treatment. By understanding the types of narratives their clients are exposed to, clinicians can better contextualize their clients' perspectives and tailor their guidance accordingly.

Limitations

The large number of YouTube videos included in this study enhances its generalizability and depth of findings across a diverse range of content creators. However, while the study expands on our understanding of creators on YouTube, this focus also excludes insights from emerging creators present on other platforms like Instagram and TikTok. Furthermore, this study did not include short-form content (ie, YouTube Shorts), which has risen in popularity in recent years. Possible differences between long-form and short-form health content across social media platforms should be examined in subsequent studies.

This study also used a convenience sample to identify content creators, aiming to reflect the viewing habits of TGD individuals, rather than relying on algorithm-driven recommendations on the YouTube platform. Although this strategy helped prioritize community relevance in our sampling, it also introduces selection bias. We also limited our sample to English-speaking creators, which limits cultural context for non-English speaking TGD individuals. Additionally, our analysis of channel subscriber and engagement metrics (ie, likes and views) did not account for time-dependent factors, such as how long individual videos have been online or the age of the channel. Future analyses may account for these variables to better contextualize engagement patterns over time.

This study also did not assess the accuracy of health information made by content creators. Although these content creators have lived experience with gender-affirming medical treatments, they are not medical professionals, and their content is not held to clinical standards. As a result, there is the potential for this content to contain inaccurate or misleading information. Assessing the accuracy of health information shared by TGD content creators is a vital area for future study given broader concerns about medical misinformation online.

Conclusions

This study offers insights into the health-related information shared by TGD content creators on YouTube, emphasizing the range of topics covered and the approaches used to share medical information. Although most videos by TGD content creators are unrelated to gender identity, videos that do address this topic frequently mention gender-affirming medical care. Our findings underscore the potential of social media sites like

XSL•FO
RenderX

YouTube to serve as a vital resource for TGD individuals seeking accessible and relatable health information, particularly in the face of systemic barriers to accessing inclusive health care.

Acknowledgments

This work was supported by the Eunice Kennedy Shriver National Institute of Child Health and Human Development of the National Institutes of Health under grant R03HD102472 and grant K23HD093815. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Thank you to Megan Gour, LCSW; Victoria Adkins, MSW; and Angela Calvin, PhD, for critical feedback during codebook creation. Special thanks to Ami Lieberman, Gabby Schreiber, Zoe Lash, and Julie Valdivia for vital assistance with data collection, which made this study possible. Artificial intelligence tools were used in the proofreading of this manuscript. Responsibility for the scientific integrity, accuracy, and validity of this work lies solely with the authors.

Data Availability

The dataset generated and analyzed during this study is not publicly available in order to protect the privacy of the content creators whose videos were included. The dataset is available from the corresponding author upon reasonable request.

Authors' Contributions

Supervision: ES. Funding acquisition: ES. Project administration: LB. Conceptualization: ES and OH. Methodology: ES and LB. Software: LB. Validation: LB. Investigation: LB and IC. Resources: ES. Formal analysis: QZ. Data curation: LB and QZ. Visualization: LB. Writing – original draft: LB and IC. Writing – review & editing: LB, QZ, IC, OH, and ES.

Conflicts of Interest

ES is on a scientific advisory board for Google. Google did not fund this research, nor were any data from the study shared with Google. The other authors declare no conflicts.

Multimedia Appendix 1
Full codebook of medical and nonmedical topics identified in YouTube videos made by selected transgender and gender-diverse content creators.
[DOCX File, 29 KB - formative_v9i1e75787_app1.docx]

Multimedia Appendix 2
Summary of interrater reliability statistics.
[DOCX File, 20 KB - formative_v9i1e75787_app2.docx]

References

1. Bates A, Hobman T, Bell BT. “Let me do what I please with it... don’t decide my identity for me”: LGBTQ+ youth experiences of social media in narrative identity development. J Adolesc Res 2020 Jan;35(1):51-83. [doi: 10.1177/0743558419884700]
2. Cannon Y, Speedlin S, Avera J, Robertson D, Ingram M, Prado A. Transition, connection, disconnection, and social media: examining the digital lived experiences of transgender individuals. J LGBT Issues Couns 2017 Apr 3;11(2):68-87. [doi: 10.1080/15538605.2017.1310006]
3. Buss J, Le H, Haimson OL. Transgender identity management across social media platforms. Media, Culture & Society 2022 Jan;44(1):22-38. [doi: 10.1177/01634437211027106]
4. Herrmann L, Bindt C, Hohmann S, Becker-Hebly I. Social media use and experiences among transgender and gender diverse adolescents. Int J Transgend Health 2024;25(1):36-49. [doi: 10.1080/26895269.2023.2252410] [Medline: 38328588]
5. Coyne SM, Weinstein E, Sheppard JA, et al. Analysis of social media use, mental health, and gender identity among US youths. JAMA Netw Open 2023 Jul 3;6(7):e2324389. [doi: 10.1001/jamanetworkopen.2023.24389] [Medline: 37486631]
6. Fisher CB, Tao X, Ford M. Social media: a double-edged sword for LGBTQ+ youth. Comput Human Behav 2024 Jul;156:108194. [doi: 10.1016/j.chb.2024.108194]
7. Allen BJ, Zhao Q, Moreno MA. Digital media use differs between transgender and cisgender youth. Pediatrics 2021 Mar 1;147(3_MeetingAbstract):214-215. [doi: 10.1542/peds.147.3MA3.214]
8. Charmaraman L, Hodes R, Hernandez JM. Marginalized and understudied populations using digital media. In: Nesi J, Prinstein MJ, Telzer EH, editors. Handbook of Adolescent Digital Media Use and Mental Health: Cambridge University Press; 2022:188-214. [doi: 10.1017/9781108976237.011]

XSL•FO
RenderX

9. Berger MN, Taba M, Marino JL, Lim MSC, Skinner SR. Social media use and health and well-being of lesbian, gay, bisexual, transgender, and queer youth: systematic review. J Med Internet Res 2022 Sep 21;24(9):e38449. [doi: 10.2196/38449] [Medline: 36129741]
10. Stoody VB, Glick HR, Murphey AC, Sturza JM, Selkie EM. A content analysis of transgender health and wellness themes shared through social media. Clin Pediatr (Phila) 2024 Oct;63(10):1352-1363. [doi: 10.1177/00099228231219499] [Medline: 38135918]
11. Berger MN, Taba M, Marino JL, et al. Social media's role in support networks among LGBTQ adolescents: a qualitative study. Sex Health 2021 Nov;18(5):421-431. [doi: 10.1071/SH21110] [Medline: 34706814]
12. Blotner C, Rajunov M. Engaging transgender patients: using social media to inform medical practice and research in transgender health. Transgend Health 2018;3(1):225-228. [doi: 10.1089/trgh.2017.0039] [Medline: 30596150]
13. Augustaitis L, Merrill LA, Gamarel KE, Haimson OL. Online transgender health information seeking: facilitators, barriers, and future directions. 2021 May 6 Presented at: CHI '21; May 8-13, 2021; Yokohama, Japan. [doi: 10.1145/3411764.3445091]
14. Pohjanen AM, Kortelainen TAM. Transgender information behaviour. Journal of Documentation 2016 Jan 11;72(1):172-190. [doi: 10.1108/JD-04-2015-0043]
15. Selkie E, Adkins V, Masters E, Bajpai A, Shumer D. Transgender adolescents' uses of social media for social support. J Adolesc Health 2020 Mar;66(3):275-280. [doi: 10.1016/j.jadohealth.2019.08.011] [Medline: 31690534]
16. Anderson M, Faverio M, Gottfried J. Teens, social media and technology 2023. Pew Research Center. 2023. URL: https://www.pewresearch.org/internet/2023/12/11/teens-social-media-and-technology-2023/ [accessed 2025-08-21]
17. Gottfried J. Americans' social media use. Pew Research Center. 2024. URL: https://www.pewresearch.org/internet/2024/01/31/americans-social-media-use/ [accessed 2025-08-21]
18. Top websites in the United States (all industries). SEMRUSH. URL: https://www.semrush.com/trending-websites/us/all [accessed 2024-10-17]
19. Madathil KC, Rivera-Rodriguez AJ, Greenstein JS, Gramopadhye AK. Healthcare information on YouTube: a systematic review. Health Informatics J 2015 Sep;21(3):173-194. [doi: 10.1177/1460458213512220] [Medline: 24670899]
20. Tolbert AN, Drogos KL. Tweens' wishful identification and parasocial relationships with YouTubers. Front Psychol 2019;10:2781. [doi: 10.3389/fpsyg.2019.02781] [Medline: 31920829]
21. Gleason TR, Theran SA, Newberg EM. Parasocial interactions and relationships in early adolescence. Front Psychol 2017;8:255. [doi: 10.3389/fpsyg.2017.00255] [Medline: 28280479]
22. Gräve JF. Exploring the perception of influencers vs. traditional celebrities: are social media stars a new type of endorser? Presented at: Proceedings of the 8th International Conference on Social Media & Society; Jul 28-30, 2017; New York, NY. [doi: 10.1145/3097286.3097322]
23. Zimmermann D, Noll C, Gräßer L, et al. Influencers on YouTube: a quantitative study on young people's use and perception of videos about political and societal topics. Curr Psychol 2022 Oct;41(10):6808-6824. [doi: 10.1007/s12144-020-01164-7]
24. Bandura A. Social cognitive theory of mass communication. Media Psychol 2001 Aug;3(3):265-299. [doi: 10.1207/S1532785XMEP0303_03]
25. Raun T. Video blogging as a vehicle of transformation: exploring the intersection between trans identity and information technology. International Journal of Cultural Studies 2015 May;18(3):365-378. [doi: 10.1177/1367877913513696]
26. Rothbaum B, Etengoff C, Uribe E. Transgender community resilience on YouTube: constructing an informational, emotional, and sociorelational support exchange. J Community Psychol 2022 Jul;50(5):2366-2384. [doi: 10.1002/jcop.22781] [Medline: 34942028]
27. Eckstein AJ. Out of sync: complex temporality in transgender men's YouTube transition channels. QED 2018 Feb 1;5(1):24-47. [doi: 10.14321/qed.5.1.0024]
28. Horak L. Trans on YouTube: intimacy, visibility, temporality. TSQ 2014;1(4):572-585. [doi: 10.1215/23289252-2815255]
29. Miller JF MA. YouTube as a site of counternarratives to transnormativity. J Homosex 2019;66(6):815-837. [doi: 10.1080/00918369.2018.1484629] [Medline: 29893635]
30. O'Neill MG. Transgender youth and youtube videos: self-representation and five identifiable trans youth narratives. In: Pullen C, editor. Queer Youth and Media Cultures: Palgrave Macmillan UK; 2014:34-45. [doi: 10.1057/9781137383556_3]
31. Raun T. DIY therapy: exploring affective self-representations in trans video blogs on YouTube. In: Karatzogianni A, Kuntsman A, editors. Digital Cultures and the Politics of Emotion: Feelings, Affect and Technological Change: Palgrave Macmillan UK; 2012:165-180. [doi: 10.1057/9780230391345_10]
32. Raun T. Archiving the wonders of testosterone via YouTube. TSQ 2015 Nov 1;2(4):701-709. [doi: 10.1215/23289252-3151646]
33. Raun T. Out online: trans self-representation and community building on YouTube. In: Gender, Bodies and Transformation, 1st edition: Routledge; 2016.
34. Raun T. Capitalizing intimacy: new subcultural forms of micro-celebrity strategies and affective labour on YouTube. Convergence: The International Journal of Research into New Media Technologies 2018;24(1):99-113. [doi: 10.1177/1354856517736983]
35. Reinke R. Embodying resistance online: trans youth reconfigure discursive space(s) of visibility on YouTube. Spectator 2017;37(2) [FREE Full text]

XSL•FO
RenderX

36. Tortajada I, Willem C, Platero R, Arauna N. Lost in transition? Digital trans activism on YouTube. Inf Commun Soc 2020;0:1-17. [doi: 10.1080/1369118X.2020.1797850]
37. Schier HE, Linsenmeyer WR. Nutrition-related messages shared among the online transgender community: a netnography of YouTube vloggers. Transgend Health 2019;4(1):340-349. [doi: 10.1089/trgh.2019.0048] [Medline: 31872063]
38. Miller B. YouTube as educator: a content analysis of issues, themes, and the educational value of transgender-created online videos. Social Media + Society 2017 Apr;3(2):1-12. [doi: 10.1177/2056305117716271]
39. Fu J, Li C, Zhou C, et al. Methods for analyzing the contents of social media for health care: scoping review. J Med Internet Res 2023 Jun 26;25:e43349. [doi: 10.2196/43349] [Medline: 37358900]
40. Beatini JR, Sun NY, Coleman JK, et al. #GenderAffirmingHormoneTherapy and health information on TikTok: thematic content analysis. JMIR Infodemiology 2025 Apr 29;5:e66845. [doi: 10.2196/66845] [Medline: 40300180]
41. Walsh-Buhi E, Houghton RF, Lange C, Hockensmith R, Ferrand J, Martinez L. Pre-exposure prophylaxis (PrEP) information on Instagram: content analysis. JMIR Public Health Surveill 2021 Jul 27;7(7):e23876. [doi: 10.2196/23876] [Medline: 34061759]
42. Park J, Lee JM, Xiong VY, Septianto F, Seo Y. David and Goliath: when and why micro-influencers are more persuasive than mega-influencers. J Advert 2021 Oct 20;50(5):584-602. [doi: 10.1080/00913367.2021.1980470]
43. Walter N, Föhl U, Zagermann L. Big or small? Impact of influencer characteristics on influencer success, with special focus on micro- versus mega-influencers. Journal of Current Issues & Research in Advertising 2025 Apr 3;46(2):160-182. [doi: 10.1080/10641734.2024.2366198]
44. Human research protection program plan (HRP-101). University of Wisconsin–Madison. 2025. URL: https://irb.wisc.edu/manual/investigator-manual/introduction/human-research-protection-program/ [accessed 2025-08-14]
45. Moreno MA, Goniu N, Moreno PS, Diekema D. Ethics of social media research: common concerns and practical considerations. Cyberpsychol Behav Soc Netw 2013 Sep;16(9):708-713. [doi: 10.1089/cyber.2012.0334] [Medline: 23679571]
46. Buck AM, Ralston DF. I didn't sign up for your research study: the ethics of using "public" data. Computers and Composition 2021 Sep;61:102655. [doi: 10.1016/j.compcom.2021.102655]
47. Exposed: the scale of transphobia online. Brandwatch. URL: https://www.brandwatch.com/reports/transphobia/ [accessed 2025-08-26]
48. Tracking anti-transgender rhetoric online, offline, and in our legislative chambers. Anti-Defamation League. 2021. URL: https://www.adl.org/resources/report/tracking-anti-transgender-rhetoric-online-offline-and-our-legislative-chambers [accessed 2025-08-26]
49. Unsafe: Meta fails to moderate extreme anti-trans hate across Facebook, Instagram, and Threads. GLAAD. 2024. URL: https://glaad.org/smsi/report-meta-fails-to-moderate-extreme-anti-trans-hate-across-facebook-instagram-and-threads/ [accessed 2025-08-26]
50. Beresford M, Wutich A, du Bray MV, et al. Coding qualitative data at scale: guidance for large coder teams based on 18 studies. Int J Qual Methods 2022 Apr;21:1-15. [doi: 10.1177/16094069221075860]
51. Moreno MA, Egan KG, Brockman L. Development of a researcher codebook for use in evaluating social networking site profiles. J Adolesc Health 2011 Jul;49(1):29-35. [doi: 10.1016/j.jadohealth.2011.04.015] [Medline: 21700153]
52. Moreno MA, Ton A, Selkie E, Evans Y. Secret society 123: understanding the language of self-harm on Instagram. J Adolesc Health 2016 Jan;58(1):78-84. [doi: 10.1016/j.jadohealth.2015.09.015] [Medline: 26707231]
53. Olson J, McAllister C, Grinnell L, Gehrke Walters K, Appunn F. Applying constant comparative method with multiple investigators and inter-coder reliability. TQR 2016;21:26-42. [doi: 10.46743/2160-3715/2016.2447]
54. Macqueen K, McLellan-Lemal E, Bartholow K, Milstein B. Team-based codebook development: structure, process, and agreement. Handbook for Team-Based Qualitative Research 2008:119-135 [FREE Full text]
55. O'Connor C, Joffe H. Intercoder reliability in qualitative research: debates and practical guidelines. Int J Qual Methods 2020 Jan 1;19:1609406919899220. [doi: 10.1177/1609406919899220]
56. Call J, Holloway B. Intersectional model of service use: understanding transgender and nonbinary healthcare access. JSBHS 2024;18(1):97-116. [doi: 10.5590/JSBHS.2024.18.1.07]
57. Clark DBA, Veale JF, Greyson D, Saewyc E. Primary care access and foregone care: a survey of transgender adolescents and young adults. Fam Pract 2018 May 23;35(3):302-306. [doi: 10.1093/fampra/cmx112] [Medline: 29177485]
58. Ehrenfeld JM, Zimmerman DR, Gonzales G. Healthcare utilization among transgender individuals in California. J Med Syst 2018 Mar 16;42(5):77. [doi: 10.1007/s10916-018-0923-8] [Medline: 29546483]
59. Redfield E, Conron KJ, Mallory C. The impact of 2024 anti-transgender legislation on youth. UCLA School of Law Williams Institute. 2024. URL: https://williamsinstitute.law.ucla.edu/publications/2024-anti-trans-legislation/ [accessed 2025-06-25]
60. Executive order 14187: protecting children from chemical and surgical mutilation. Fed Regist. 2025. URL: https://www.federalregister.gov/documents/2025/02/03/2025-02194/protecting-children-from-chemical-and-surgical-mutilation [accessed 2025-08-26]

XSL•FO
RenderX

Abbreviations

TGD: transgender and gender-diverse

Edited by A Mavragani; submitted 10.04.25; peer-reviewed by N Döring, S Mohanadas; revised version received 10.07.25; accepted 11.07.25; published 29.08.25.

Please cite as:
Bliss L, Zhao Q, Chao I, Haimson OL, Selkie E
Medical Information Provided by Transgender and Gender-Diverse Content Creators on YouTube: Descriptive Content Analysis
JMIR Form Res 2025;9:e75787
URL: https://formative.jmir.org/2025/1/e75787
doi:10.2196/75787

© Lydia Bliss, Qianqian Zhao, Irene Chao, Oliver L Haimson, Ellen Selkie. Originally published in JMIR Formative Research (https://formative.jmir.org), 29.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Public Information Needs and Interest in Specific Food and Drug Allergy Disorders in Germany (2022–2024): Google Search Engine Analysis

Tobias Fuchs[1]; Michael Hindelang[1,2,3], MSc; Sebastian Sitaru[1], MD; Alexander Zink[1], MD, MPH, MBA, PhD; Julia Welzel[4], MD

[1]Department of Dermatology and Allergy, TUM School of Medicine and Health, Technical University of Munich, Biedersteiner Str. 29, Munich, Germany

[2]Pettenkofer School of Public Health, Munich, Germany

[3]Institute for Medical Information Processing, Biometry and Epidemiology (IBE), Faculty of Medicine, Ludwig-Maximilian University, LMU Munich, Munich, Germany

[4]Department of Dermatology and Allergology, University Hospital Augsburg, Augsburg, Germany

Corresponding Author:
Tobias Fuchs
Department of Dermatology and Allergy, TUM School of Medicine and Health, Technical University of Munich, Biedersteiner Str. 29, Munich, Germany

Abstract

Background: The prevalence of food and drug allergies has been steadily increasing in Germany. These conditions not only impair the quality of life of those affected but also place an additional burden on the health care system. At the same time, an increasing number of people are using the internet and other digital sources to seek health-related information.

Objective: This study aimed to use the Google Ads Keyword Planner to identify the information needs and knowledge gaps of the internet-using population in Germany and to provide a foundation for future prevention and educational strategies regarding food and drug allergies.

Methods: Relevant keywords related to selected food and drug allergies were extracted using the Google Ads Keyword Planner and analyzed according to predefined criteria. The observation period was from September 2022 to October 2024.

Results: A total of 633 keywords related to specific types of food and drug allergies were identified, generating a combined search volume of 3,649,390 queries. The most frequently searched terms nationwide were “histamine allergy” (368,980/3,649,390, 10.1%), “penicillin allergy” (266,410/3,649,390, 7.3%), and “nut allergy” (103,850/3,649,390, 2.8%). Although “histamine allergy” was the most frequently searched term in this analysis, most searches for “histamine allergy” likely referred to an intolerance rather than a true immunoglobulin E–mediated allergy. Seasonal patterns were also observed, with increased searches for the categories “nut” and “penicillin” in the winter months and for “histamine” in the spring months.

Conclusions: This study demonstrates the potential of Google search query data analysis in a medical context and, in particular, underscores its relevance for understanding the public interest in food and drug allergies in Germany. The findings highlight the need for improved, easily accessible educational resources and for implementing allergy-specific, socially relevant health campaigns to address the unmet information needs of the population living in Germany regarding food and drug allergies.

(JMIR Form Res 2025;9:e75395) doi:10.2196/75395

KEYWORDS

food allergies; drug allergies; public health informatics; web search analysis; digital health literacy

Introduction

Background

The global prevalence of food allergies has been steadily increasing since the early 21st century [1-4], making them a widespread public health concern. It is estimated that at least 2% to 4% of all children and adults are affected [5,6], with current projections suggesting that approximately 520 million individuals worldwide have food allergies. This corresponds to a prevalence of 3% to 10% among children and up to 10% among adults [7]. The prevalence of drug allergies is of a similar magnitude. Self-reported data indicate that 7% to 8% of the general population has a medication allergy, and approximately 15% of all hospitalized patients are reportedly allergic to beta-lactam antibiotics [8-10]. These allergies not only impair the quality of life of those affected and their families [11] but also impose substantial costs on health care systems [12].

XSL•FO
RenderX

In Germany, the prevalence of allergic diseases has also risen steadily since the 1970s. According to the German Gesundheit in Deutschland aktuell 2019-2020–European Health Interview Survey study conducted between April 2019 and September 2020, among respondents, 31% reported having an allergic disease, excluding allergic asthma [13,14]. Women were found to be more frequently affected than men [13]. Data from the Commission for Environmental Medicine of the Robert Koch Institute indicate that the lifetime prevalence of food allergies among adults in Germany is 4.7% [15]. According to the most recent guideline, the prevalence is estimated at 3.7% [16], with common triggers including nuts, soy, and dairy products [16]. Reliable data on the prevalence of drug allergies in Germany remain scarce. However, point prevalence rates for penicillin allergy—one of the most important drug allergies—were reported at 17% in both 2019 and 2022 at the University Medical Center Hamburg-Eppendorf [17]. Both food and drug allergies can cause mild symptoms, such as urticaria or exanthema, as well as potentially life-threatening anaphylactic reactions [18].

The prevalence figures and the potential severity of symptoms underscore the public health relevance of these types of allergies and the urgency of addressing them—at both the clinical and patient levels.

However, the sometimes fluctuating data also highlight the challenges in accurately determining the true number of individuals affected. Reported figures are often based solely on patients' self-reported histories or on general population surveys. The already complex medical differentiation between a confirmed allergy, an intolerance reaction, and an adverse drug reaction is further complicated by the growing influence of digital nutrition "experts," lifestyle coaches, and self-help groups providing lifestyle advice. As a result, patients may assume that they have certain allergies and avoid the relevant allergens—with all associated consequences—even when allergological diagnostics fail to confirm the suspected allergy [19]. The goal should be to ensure effective treatment for individuals who are truly affected through early and targeted diagnostics, while simultaneously promoting consistent "delabeling" of nonexistent allergies [17]. Furthermore, public awareness of this issue in general and the clear differentiation between allergic and nonallergic reactions in particular should be strengthened through improved public health education.

In addition, with the ongoing digitalization of the health care sector and the expanding range of telemedical information sources, patients can now independently access information about their illnesses and symptoms online, without relying solely on medical professionals. This trend has been further accelerated by the COVID-19 pandemic, and the public's willingness to seek health information digitally continues to grow [20].

Objectives

To better understand and contextualize the current level of knowledge and the unmet information needs of the German population regarding food and drug allergies, a Google search query analysis was conducted as part of this study. The results are presented in detail herein.

A Google search query analysis is a method in which user-generated search query data are collected using tools such as the Google Ads Keyword Planner and then systematically analyzed. Initially, this approach was primarily used to optimize corporate advertising campaigns [21]. However, because this form of web analysis can reflect the various interests of the internet-using population, it has increasingly been adopted in scientific research [22-25]. Topics relevant to allergic diseases have also been analyzed with this tool [26-29]. For the first time, to the best of our knowledge, this study applies this method to the context of food and drug allergies. In Germany, such analyses are particularly effective when conducted via the Google search engine, which remains the market leader with a share of 88.2% as of November 2024 [30], making it the primary platform for the majority of the country's internet users. Overall, 94% of Germany's total population—nearly 80 million people—currently use the internet [31,32]. Therefore, a Google-based analysis can capture data reflecting the search behavior of a substantial proportion of the German population.

Methods

Overview

This Google-based analysis was conducted as a scientific extension of an ongoing research project evaluating the benefits of the What's in My Meds medication app for patients with allergies. Accordingly, the same food and drug allergies included as selectable options in the What's in My Meds smartphone app were considered in this analysis. The geographic focus was similarly aligned, targeting the Munich and Augsburg regions. This complementary project enables a meaningful supplementation of the app-based data, thereby enhancing the specificity and strengthening the relevance of both studies within a comparable regional context.

The Google Ads Keyword Planner was used to analyze search volumes in the field of allergies. Search volume refers to the number of search queries generated by internet users for a specific term (keyword) [21]. The Keyword Planner generates a list of search terms related to predefined topics relevant to the study. These search terms may consist of single words or context-related phrases entered by users in the Google search engine [20]. In addition, the Keyword Planner provides an estimate of the average monthly search volume for each relevant term. Collected data can be retrieved retrospectively for up to 48 months from the time of the query [20].

Statistics

This study primarily analyzed search queries related to allergies to selected medications and foods, as well as seasonal and regional variations in the most frequently searched keywords. Keyword data were retrieved using the Google Ads Keyword Planner application programming interface under a registered institutional account. After retrieval, an initial screening of the keywords was conducted, during which their clinical and contextual relevance was discussed within the research team. Final keyword inclusion was determined by consensus.

The analysis considered allergies to beta-lactam antibiotics (eg, penicillins, aminopenicillins, and cephalosporins), analgesics

XSL•FO
RenderX

(eg, aspirin, ibuprofen, metamizole, and diclofenac), histamine, nuts, soy, egg proteins (albumin and lecithin), macrogol, and parabens. These allergens were selected because they corresponded to those examined in the related What's in My Meds app study. In addition to the allergy terms themselves, search queries regarding allergy-related symptoms or potential cross-allergies were included.

While the analysis covered the whole of Germany, a particular focus was placed on the southern regions, especially Munich and Augsburg. Keywords in German were primarily examined, although relevant English terms were also included. The evaluation period spanned from September 2022 to October 2024.

The data were analyzed descriptively, and the most frequent keywords were qualitatively categorized. This categorization into thematic groups ("histamine," "nuts," "soy," and "penicillin") was performed by the first author and independently reviewed for consistency and completeness by other members of the study team. Data analysis was conducted using SPSS software (version 29.0; IBM Corp).

Results

Overview

A total of 633 distinct keywords related to "food allergy" and "drug allergy" were identified, generating a combined search volume of 3,649,390 queries during the observation period from September 2022 to October 2024. Distributed by year, the 3,649,390 queries corresponded to 489,630 (13.4%) in 2022, a total of 1,666,690 (45.7%) in 2023, and 1,493,070 (40.9%) in 2024. The lower figure for 2022 is explained by the fact that only data from September to December were included, whereas 2023 and 2024 were fully or nearly fully represented. When the 3,649,390 search queries were broadly assigned to the categories "food allergy" and "drug allergy," the respective volumes were 3,380,720 (92.6%) and 268,670 (7.4%), indicating that queries about food allergies accounted for a vast majority of all searches. Regarding language, 3,567,950 (97.8%) of the 3,649,390 searches were conducted in German and 81,440 (2.2%) in English. The 16 most frequently searched terms are presented in Table 1. In the Google AdWords Keyword Planner, synonyms (eg, "nut allergy" and "nut allergies") are listed separately, even when they share identical search volumes; therefore, synonymous terms with identical counts were consolidated for this analysis. Across the entire observation period, the most frequently searched term was "histamine allergy" (368,980/3,649,390, 10.1%), followed by "penicillin allergy" (266,410/3,649,390, 7.3%) and "nut allergy" (103,850/3,649,390, 2.8%). It should be noted that searches for "histamine" likely reflect lay terminology that conflates histamine intolerance with histamine allergy, highlighting common public misunderstandings. Overall, the top 5 keywords accounted for 25.1% (917,070/3,649,390) of the total search volume. Allergy-related terms concerning analgesics were searched far less frequently, with "ibuprofen allergy" being the most common (18,900/3,649,390, 0.5%).

Table . Google users' top keywords in Germany regarding specific food and drug allergies listed by search volume (September 2022-October 2024; N=3,649,390).

Keywords (a consolidation of synonymous search terms was performed)	Search volume, n (%)
Histamine allergy (IgE[a] mediated)	368,980 (10.1)
Penicillin allergy	266,410 (7.3)
Nut allergy, nut allergies, nut allergy sufferers	103,850 (2.8)
Allergy to soy	92,680 (2.5)
Histamine intolerance symptoms (non-IgE mediated)	85,150 (2.3)
Nut allergy symptoms, symptoms of a nut allergy, symptoms of nut allergy, symptoms nut allergy	75,700 (2.1)
Soy allergy, allergy soy, allergens soy, soya allergy	46,340 (1.3)
Intolerance to soy, soy intolerance	39,680 (1.1)
Peanut allergy symptoms	33,330 (0.9)
Allergy to histamine (IgE mediated)	33,140 (0.9)
Hazelnut allergy, allergy hazelnuts, allergy to hazelnuts	32,740 (0.9)
Symptoms of histamine allergy, allergy histamines symptoms, allergy histamine symptoms, histamine allergy symptoms (IgE mediated)	31,170 (0.9)
Cross allergy hazelnuts	29,120 (0.8)
Walnut allergen	28,680 (0.8)
Allergy hazelnut	26,160 (0.7)
Allergy soy symptoms	24,470 (0.7)

[a]IgE: immunoglobulin E.

XSL•FO
RenderX

Categorization of Content

In addition, the top 16 keywords were categorized by content. While search queries for these keywords were evaluated individually, a thematic aggregation was also performed. Accordingly, all search terms within the top 16 were grouped into the categories “nuts,” “histamine,” “soy,” and “penicillin.” The total search volumes for each thematic group were then analyzed. The resulting search volumes for the categorized top 16 keywords across the entire observation period are described in more detail below.

Searches related to “nuts” accounted for 20.1% (733,527/3,649,390) of the total search volume, followed by “histamine” (613,098/3,649,390, 16.8%), “soy” (383,186/3,649,390, 10.5%), and “penicillin” (266,405/3,649,390, 7.3%).

Search terms concerning analgesics were also categorized by content. The category “ibuprofen” accounted for 2.4% (86,820/3,649,390) of the total search volume, followed by “diclofenac” (8050/3,649,390, 0.2%), “acetylsalicylic acid” (2260/3,649,390, 0.1%), and “naproxen” (330/3,649,390, 0.0%).

Regional Differences

Of the 3,649,390 searches, 153,500 (4.2%) and 40,460 (1.1%) were recorded in Munich and Augsburg, respectively, over the entire observation period, whereas, for the rest of Germany, the search volume amounted to 3,455,430 (94.7%) queries (Figure 1).

XSL•FO
RenderX

Figure 1. Regional distribution of Google search volumes regarding specific food and drug allergy search terms in Germany (September 2022-October 2024).

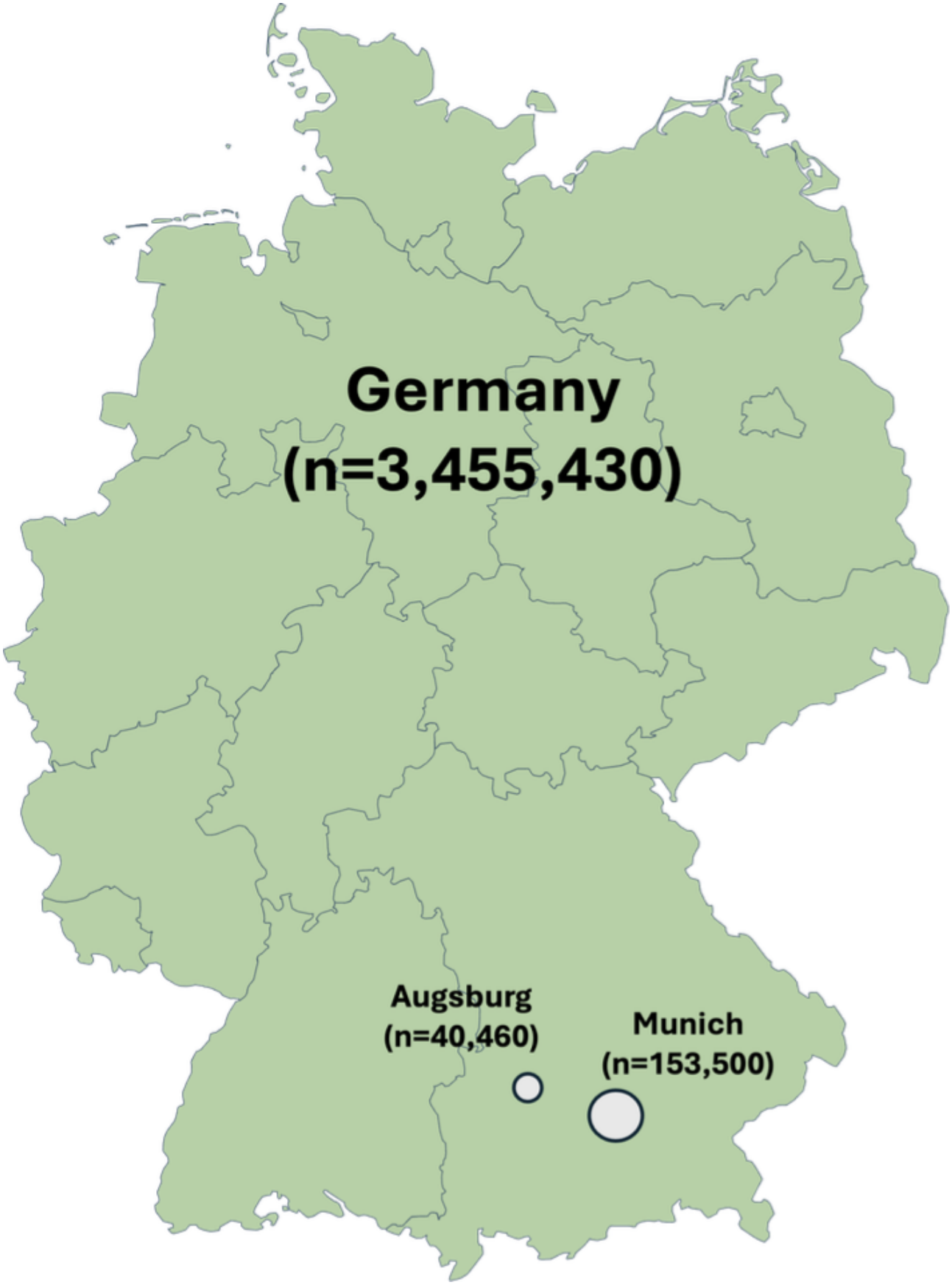

The linguistic distribution of search queries in the respective regions was as follows: of the 3,455,430 search queries from the rest of Germany, 3,377,950 (92.6%) were in German and 77,480 (2.1%) in English. In Munich, of the 153,500 queries,

XSL•FO
RenderX

1,50,450 (98%) were in German and 3050 (2%) in English, while in Augsburg, of the 40,460 queries, 39,550 (97.8%) were in German and 910 (2.2%) in English.

A qualitative classification of all keywords into the main categories of “histamine,” “nuts,” “penicillin,” and “soy” was also performed to assess regional differences. All remaining search terms were grouped under the heading “Other.” Table 2 presents the regional differences in search behavior for the top 4 categories—expressed per 100,000 inhabitants—for Germany as a whole as well as Munich and Augsburg.

Table . Regional differences in Google search volumes for food and drug allergy categories, including the top 4 and “Other,” per 100,000 inhabitants in Germany (September 2022-October 2024).

Categories	Germany (except Munich and Augsburg)	Munich	Augsburg
Other	571.8	2066.9	3342.2
Histamine	1008.8	2451.3	2804.1
Nuts	1568.0	3541.9	5010.0
Penicillin	326.7	559.1	364.3
Soy	677.8	1695.3	2002.0

Seasonal Trends

To provide an overview of seasonal trends, the most frequently searched categories over the entire study period were analyzed for all regions examined. These categories included all search queries related to “nuts,” “histamine,” “soy,” and “penicillin.” Figure 2 illustrates the monthly variation in search volumes for these top 4 categories throughout the observation period.

Figure 2. Monthly variation in Google users’ top 4 search categories regarding specific food and drug allergies in Germany (September 2022 to October 2024).

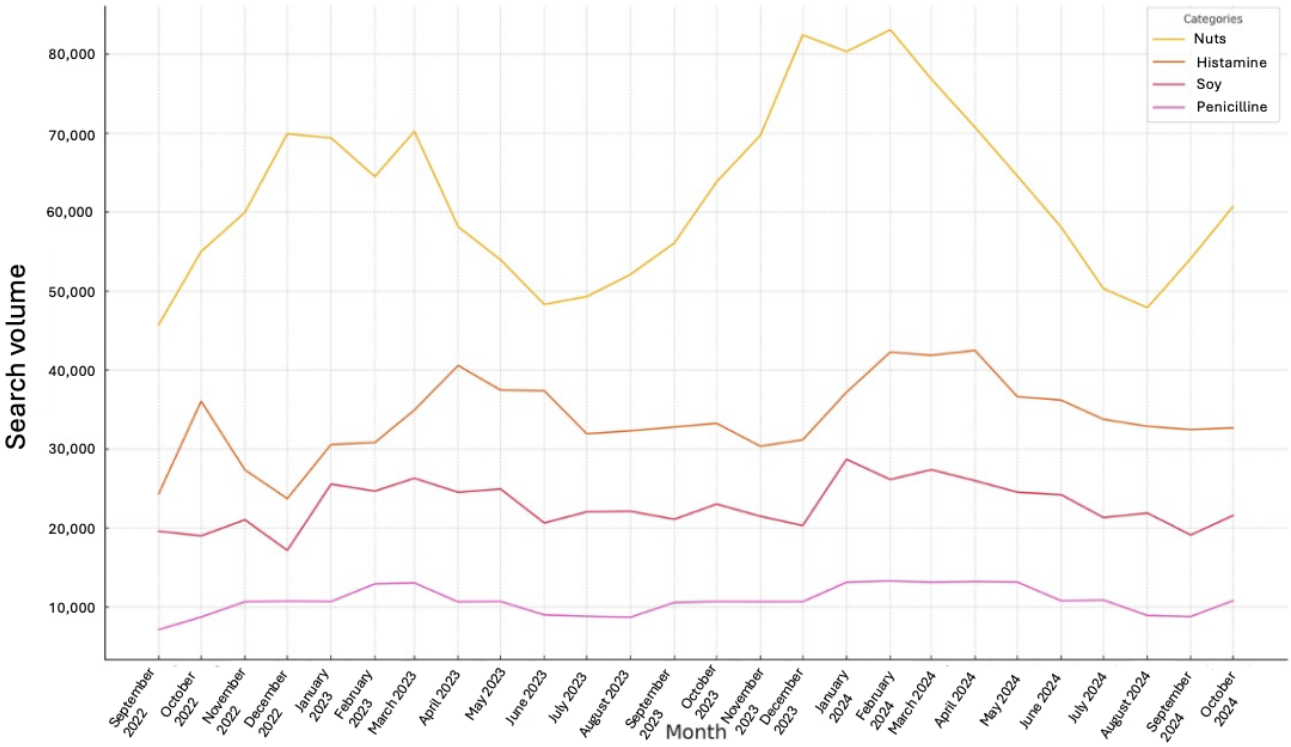

The seasonal distribution of search volumes for the categories “nuts,” “histamine,” “soy,” and “penicillin” revealed clear patterns and recurring fluctuations across 2022, 2023, and 2024. For “nuts,” an increase in search volume was consistently observed during the winter months of each year, peaking in February 2024 over the entire observation period. From April onward, search volumes declined sharply, reaching their lowest levels during the summer months, particularly between May and August. By contrast, the “histamine” category exhibited the opposite pattern: peak search volumes occurred mainly in March and April, followed by a steady decline continuing into the winter months. The only deviation from this trend was a secondary peak in October 2022, which cannot be fully contextualized due to the absence of data from earlier months that year. The “soy” category reached its highest search volumes in 2023 and 2024, especially between January and May, with a peak in January 2024. Overall, fluctuations in “soy” search volumes were less pronounced than those observed for “nuts” and “histamine.”

The “penicillin” category showed a moderate increase during the winter months of each study year, extending into May in

XSL•FO
RenderX

2024. From spring onward, search volumes gradually declined—sometimes with a slight delay—before stabilizing at a lower level during the summer months. These observed seasonal patterns were consistent across all years analyzed.

Discussion

Principal Findings

The Google-based analysis conducted in this study provides valuable insights into the search behavior and distribution of interest within the German population regarding "food allergy" and "drug allergy." In the context of this study, the term "German population" refers to individuals living in Germany who use the internet, regardless of country of origin or citizenship. With a total of 3,649,390 search queries, it is evident that allergy-related topics are of substantial public interest and are regularly researched via the Google search engine. The distribution of search queries suggests that conditions such as nut allergy and histamine intolerance are subjects of active public discussion, potentially driving the observed high search volumes. However, this association requires validation through future studies, as current evidence is lacking. In the case of nut allergies, heightened interest may be linked to the potential severity of reactions—members of the nut family account for 70% to 90% of all food-related anaphylactic deaths [33]. The public relevance of histamine intolerance may be explained by the presence of histamine in commonly consumed foods and beverages such as cheese, fish, and red wine [34]. Notably, searches for "histamine" were often framed as "histamine allergy," the single most searched term in the study period. Scientifically, such cases more likely represent intolerance reactions—caused by reduced histamine breakdown due to potential diamine oxidase deficiency—rather than true immunoglobulin E–mediated allergies [35]. The existence of histamine intolerance itself is still debated among experts [36,37]. These findings highlight the need for improved medical education, as the distinction between allergy and intolerance can have significant therapeutic and prognostic implications.

A similar situation applies to "penicillin allergy," the second most frequently searched term. Although its public relevance is clear, the diagnosis should be approached with caution. Differentiating between adverse effects and true allergic reactions is challenging; for instance, aminopenicillin administration during an Epstein-Barr virus infection may trigger a rash resembling an allergy but without life-threatening implications [38]. Many older patients report penicillin allergies dating back to childhood and, consequently, avoid all derivatives. This avoidance can be problematic, as penicillins remain the first-line therapy for many bacterial infections. Uncritical acceptance of self-reported allergies often leads to the use of broad-spectrum reserve antibiotics, further exacerbating antimicrobial resistance [39]. Given the public interest indicated by our findings, there is a strong need to raise awareness in clinical practice and to promote systematic "delabeling" of inaccurately assigned allergies [17]. By contrast, searches for allergic reactions to analgesics were far less frequent, despite being among the most common drug-induced causes of anaphylaxis [40]. This may reflect a general lack of public awareness of analgesics as potential allergens, lower perceived social relevance, or the fact that painkillers are typically used intermittently or by specific patient groups (eg, patients with chronic pain or those with cancer) [41]. Older adults, who often require analgesics for rheumatic or degenerative diseases [42], are also less likely to seek information online, limiting the representativeness of Google search data for this topic. Nonetheless, health care providers should increase public awareness of analgesics as possible allergens, with family physicians ideally leading such efforts, as they are often the first point of contact for patients and have the most comprehensive overview of their long-term and as-needed medications. Educational strategies could include printed information materials, short instructional videos, and other outreach initiatives, as described later in this paper. Seasonal analysis revealed distinct trends. The winter peak in nut allergy searches may relate to increased consumption of nuts during the Christmas period, while the spring and summer rise in histamine-related searches could reflect the seasonal availability of histamine-rich foods such as strawberries and citrus fruits. The winter increase in penicillin allergy searches aligns with the documented patterns of antibiotic overprescription in colder months [43]. At present, these seasonal interpretations remain hypotheses, as the available data are limited, underscoring the need for further investigation.

The regional analysis revealed notable differences. As expected, Munich had a higher total search volume than Augsburg, which can be attributed to its larger population. When comparing the search volumes for the top 4 categories per 100,000 inhabitants in Munich and Augsburg with those for Germany overall, it becomes evident that urban search volumes are substantially higher than the national average. One possible explanation is greater medical awareness in cities due to higher educational levels or more direct access to digital media compared with rural areas. To support this hypothesis, future studies should compare data across federal states and include other major German cities. Nevertheless, it remains unclear to what extent factors such as educational level, cultural differences, or access to medical information influenced the results, as such aspects are difficult to assess through search engine analysis alone.

The methodology applied—particularly the use of the Google Ads Keyword Planner—allowed for systematic recording and categorization of search behavior but also carries limitations. First, the analysis includes only searches conducted via Google. While Google remained the market leader in Germany in 2024 with a market share of 88.2%, other search engines still account for a portion of the market. Consequently, the search behavior of approximately 10 million people is not represented. Second, it is questionable whether the analyzed queries are fully representative of the entire German population, as groups considered vulnerable and with limited digital engagement—such as older adults or rural populations with restricted internet access—may be underrepresented. According to an annual online study conducted by German public broadcasters in 2022, internet use declines significantly among those aged more than 70 years [44]. Given that this age group comprises approximately 15 million people, or roughly 19% of the German population [45], this likely represents a major

XSL•FO
RenderX

limitation of the study. It also remains unclear whether the analyzed search queries were generated by individuals personally affected, by professionals, or on behalf of others (eg, parents searching for information for their children). As children are frequently affected by food allergies [2], it is plausible that a proportion of the relevant searches originated from parents or guardians. In addition, while keyword categorization was essential for analysis, it risks oversimplifying the data and overlooking nuanced or highly specific issues.

Despite these limitations, the results indicate a substantial public demand for information about food and drug allergies via online search platforms. Further research is needed to determine whether this demand will continue to grow. As patients increasingly seek information online before consulting a health care professional [20], it is critical that search engine providers present medically accurate and scientifically sound content. At the same time, the medical community must provide evidence-based, patient-oriented information to strengthen health literacy and enable individuals to critically evaluate online medical resources—especially in the context of digital infodemics [46].

This study highlights the need to improve the public's ability to distinguish between true allergies and intolerances and to encourage consistent "delabeling" of nonexistent allergies, such as inaccurately reported penicillin allergies. It also underscores the importance of raising awareness about other potent allergens, such as analgesics. Multiple strategies could be used to address these needs; for example, targeted educational campaigns in hospitals and outpatient clinics could be aligned with frequently searched terms and seasonal patterns, providing printed materials or posters focusing on specific allergens—for instance, nut allergies in winter or histamine intolerance in spring. Media campaigns involving collaborations between clinics, medical practices, health insurers, and allergy associations, as well as organizing allergy awareness events, could further increase public engagement. Establishing official allergy awareness days might also enhance visibility. Finally, expanding the range of digital information tools—such as regular use of smartphone apps or short educational videos on platforms such as Instagram Reels or TikTok—could help disseminate allergy-related knowledge to the general public, independent of direct physician contact. The key to these initiatives lies in balancing individual needs with societal priorities and in addressing all population groups and generations as inclusively as possible. If implemented strategically, these measures could sustainably improve awareness and understanding of food and drug allergies in Germany.

Conclusions

The findings of this study underscore the considerable value of Google search query analysis for identifying the information needs of the German population regarding food and drug allergies. Such data can serve as a foundation for developing improved, evidence-based educational strategies and enhancing public awareness of the allergological topics highlighted in this study. The results demonstrate that individuals with allergies commonly use internet search engines such as Google to seek health-related information. Building on this, the future expansion of digital information tools—such as social media platforms, smartphone apps, or wearable technologies—could further support patients with allergies.

The analysis also revealed that histamine intolerance is still frequently perceived by the German public as a manifest allergy and that "penicillin allergy" and "nut allergy" remain highly prominent topics in public discourse.

Consequently, incorporating search engine analyses into health care research is advisable, both to gain an overview of public interest in specific health topics and to use the resulting data as a valuable complement to traditional epidemiological studies.

The informative value of such data could be further enhanced by integrating additional insights from social media analyses or qualitative interviews on specific issues. Supplementary studies addressing demographic and socioeconomic factors could deepen the analysis and help reduce potential biases in the research findings.

Acknowledgments

Figures 1 and 2 were prepared independently by our team, with supportive assistance from the generative artificial intelligence tool ChatGPT (OpenAI) [47].

Data Availability

The datasets generated or analyzed during this study are included in this paper. All aggregate data are available from the corresponding author upon reasonable request.

Authors' Contributions

We would like to thank SS for providing the data. Special thanks to MH, AZ, and JW for assistance with proofreading.

Conflicts of Interest

None declared.

References

XSL•FO
RenderX

1. Bahna SL. History of food allergy and where we are today. World Allergy Organ J 2024 May 16;17(5):100912. [doi: 10.1016/j.waojou.2024.100912] [Medline: 38800498]
2. Gupta RS, Springston EE, Warrier MR, et al. The prevalence, severity, and distribution of childhood food allergy in the United States. Pediatrics 2011 Jul;128(1):e9-e17. [doi: 10.1542/peds.2011-0204] [Medline: 21690110]
3. Gupta R, Holdford D, Bilaver L, Dyer A, Holl JL, Meltzer D. The economic impact of childhood food allergy in the United States. JAMA Pediatr 2013 Nov;167(11):1026-1031. [doi: 10.1001/jamapediatrics.2013.2376] [Medline: 24042236]
4. Gupta RS, Warren CM, Smith BM, et al. Prevalence and severity of food allergies among US adults. JAMA Netw Open 2019 Jan 4;2(1):e185630. [doi: 10.1001/jamanetworkopen.2018.5630] [Medline: 30646188]
5. Muraro A, de Silva D, Halken S, et al. Managing food allergy: GA2LEN guideline 2022. World Allergy Organ J 2022 Sep 7;15(9):100687. [doi: 10.1016/j.waojou.2022.100687] [Medline: 36119657]
6. Sicherer SH, Sampson HA. Food allergy: a review and update on epidemiology, pathogenesis, diagnosis, prevention, and management. J Allergy Clin Immunol 2018 Jan;141(1):41-58. [doi: 10.1016/j.jaci.2017.11.003] [Medline: 29157945]
7. Gupta RS, Warren CM, Smith BM, et al. The public health impact of parent-reported childhood food allergies in the United States. Pediatrics 2018 Dec;142(6):e20181235. [doi: 10.1542/peds.2018-1235] [Medline: 30455345]
8. Demoly P, Castells M. Important questions in drug allergy and hypersensitivity: consensus papers from the 2018 AAAAI/WAO international drug allergy symposium. World Allergy Organ J 2018 Dec 19;11(1):42. [doi: 10.1186/s40413-018-0224-1] [Medline: 30598723]
9. van Dijk SM, Gardarsdottir H, Wassenberg MW, Oosterheert JJ, de Groot MC, Rockmann H. The high impact of penicillin allergy registration in hospitalized patients. J Allergy Clin Immunol Pract 2016;4(5):926-931. [doi: 10.1016/j.jaip.2016.03.009] [Medline: 27131826]
10. Eigenmann PA, Akdis C, Bousquet J, Grattan CE, Hoffmann-Sommergruber K, Jutel M. Food and drug allergy, and anaphylaxis in EAACI journals (2018). Pediatr Allergy Immunol 2019 Dec;30(8):785-794. [doi: 10.1111/pai.13125] [Medline: 31539176]
11. Nowak-Wegrzyn A, Hass SL, Donelson SM, et al. The Peanut Allergy Burden Study: impact on the quality of life of patients and caregivers. World Allergy Organ J 2021 Feb 15;14(2):100512. [doi: 10.1016/j.waojou.2021.100512] [Medline: 33664934]
12. Bilaver LA, Chadha AS, Doshi P, O'Dwyer L, Gupta RS. Economic burden of food allergy: a systematic review. Ann Allergy Asthma Immunol 2019 Apr;122(4):373-380. [doi: 10.1016/j.anai.2019.01.014] [Medline: 30703439]
13. Heidemann C, Scheidt-Nave C, Beyer AK, et al. Gesundheitliche lage von erwachsenen in Deutschland – ergebnisse zu ausgewählten indikatoren der studie GEDA 2019/2020-EHIS [Article in German]. J Health Monit 2021;6(3). [doi: 10.25646/8456]
14. Bergmann KC, Brehler R, Endler C, et al. Auswirkungen des klimawandels auf allergische erkrankungen in Deutschland [Article in German]. J Health Monit 2023;8(S4). [doi: 10.25646/11648]
15. Bergmann KC, Heinrich J, Niemann H. Current status of allergy prevalence in Germany: position paper of the Environmental Medicine Commission of the Robert Koch Institute. Allergo J Int 2016;25:6-10. [doi: 10.1007/s40629-016-0092-6] [Medline: 27069844]
16. Worm M, Reese I, Ballmer-Weber B, et al. Update Leitlinie zum Management IgE-vermittelter Nahrungsmittelallergien – S2k-Leitline der DGAKI [Article in German]. Allergologie 2021;44(7):488-541. [doi: 10.5414/ALX02257]
17. Koch T, Leubner H, Brehm TT, Witte J. Penicillinallergie: sicher und effektiv ausschließen [Article in German]. Deutsches Ärzteblatt 2023;120(18):A822-A824 [FREE Full text]
18. Pawankar R, Baena-Cagnani CE, Bousquet J, et al. State of world allergy report 2008: allergy and chronic respiratory diseases. World Allergy Organ J 2008 Jun;1(6 Suppl):S4-S17. [doi: 10.1097/WOX.0b013e31817ff995] [Medline: 23282447]
19. Oberhofer E. Sieben Mythen rund ums Thema Allergie [Article in German]. MMW - Fortschritte der Medizin 2013 Dec;155(21-22):6-6. [doi: 10.1007/s15006-013-2468-6]
20. Ziehfreund S, Tizek L, Zink A. [Web search data as health data? : Geographic differences, temporal trends, and topics of interest from internet search engine analyses in Germany]. Hautarzt 2022 Jan;73(1):53-60. [doi: 10.1007/s00105-021-04918-x] [Medline: 34812913]
21. Zink A, Rüth M, Schuster B, Darsow U, Biedermann T, Ständer S. Pruritus in Germany-a Google search engine analysis [Article in German]. Hautarzt 2019 Jan;70(1):21-28. [doi: 10.1007/s00105-018-4215-5] [Medline: 29876613]
22. Zink A, Schuster B, Rüth M, et al. Medical needs and major complaints related to pruritus in Germany: a 4-year retrospective analysis using Google AdWords Keyword Planner. J Eur Acad Dermatol Venereol 2019 Jan;33(1):151-156. [doi: 10.1111/jdv.15200] [Medline: 30067284]
23. Seidl S, Schuster B, Rüth M, Biedermann T, Zink A. What do Germans want to know about skin cancer? A nationwide google search analysis from 2013 to 2017. J Med Internet Res 2018 May 2;20(5):e10327. [doi: 10.2196/10327] [Medline: 29698213]
24. Wehner MR, Nead KT, Linos E. Correlation among cancer incidence and mortality rates and internet searches in the United States. JAMA Dermatol 2017 Sep 1;153(9):911-914. [doi: 10.1001/jamadermatol.2017.1870] [Medline: 28658470]
25. Mei YJ, Mao YM, Cao F, Wang T, Li ZJ. Using internet search data to explore the global public concerns in ankylosing spondylitis. Postgrad Med J 2021 Feb;97(1144):93-96. [doi: 10.1136/postgradmedj-2019-137407] [Medline: 32094142]

XSL•FO
RenderX

26. Tizek L, Schielein MC, Tizek L, Zink A. [Atopic dermatitis-identifying needs in the German population by internet search queries]. Hautarzt 2022 Jun;73(6):475-484. [doi: 10.1007/s00105-022-04974-x] [Medline: 35294564]
27. Sitaru S, Tizek L, Buters J, Ekebom A, Wallin JE, Zink A. Assessing the national burden of allergic asthma by web-search data, pollen counts, and drug prescriptions in Germany and Sweden. World Allergy Organ J 2023 Feb 23;16(2):100752. [doi: 10.1016/j.waojou.2023.100752] [Medline: 36896457]
28. Schober A, Tizek L, Johansson EK, et al. Monitoring disease activity of pollen allergies: what crowdsourced data are telling us. World Allergy Organ J 2022 Nov 14;15(12):100718. [doi: 10.1016/j.waojou.2022.100718] [Medline: 36438191]
29. King C, Judge C, Byrne A, Conlon N. Googling allergy in Ireland: content analysis. J Med Internet Res 2020 May 13;22(5):e16763. [doi: 10.2196/16763] [Medline: 32401220]
30. Search engine market share Germany. StatCounter GlobalStats. URL: https://gs.statcounter.com/search-engine-market-share/all/germany [accessed 2024-12-12]
31. Anteil der internetnutzer in deutschland in den jahren 2001 bis 2023 [Web page in German]. Statista. URL: https://de.statista.com/statistik/daten/studie/13070/umfrage/entwicklung-der-internetnutzung-in-deutschland-seit-2001/ [accessed 2024-12-12]
32. Bevölkerung - einwohnerzahl von deutschland von 1990 bis 2024 [Web page in German]. Statista. URL: https://de.statista.com/statistik/daten/studie/2861/umfrage/entwicklung-der-gesamtbevoelkerung-deutschlands/ [accessed 2024-12-12]
33. Walnuss: allergene – fakten, symptome und behandlung [Web page in German]. Thermo Fisher. URL: https://www.thermofisher.com/allergy/de/de/allergen-fact-sheets/walnut.html [accessed 2024-12-15]
34. Hattori Y, Seifert R. Histamine and Histamine Receptors in Health and Disease: Springer; 2017.
35. Maintz L, Novak N. Histamine and histamine intolerance. Am J Clin Nutr 2007 May;85(5):1185-1196. [doi: 10.1093/ajcn/85.5.1185] [Medline: 17490952]
36. Fischer A. Histaminunverträglichkeit – reale erkrankung oder akademisches Luftschloss? [Article in German]. Hautnah Dermatologie 2013;29:307. [doi: 10.1007/s15012-013-1678-z]
37. von Grätz PG. Kleines molekül, große frage – histaminintoleranz: pathologie oder psychopathologie? : DGVS; 2012 URL: https://www.springermedizin.de/dgvs-2012/allgemeinmedizin/histaminintoleranz-pathologie-oder-psychopathologie/9939986 [accessed 2025-10-22]
38. Chovel-Sella A, Ben Tov A, Lahav E, et al. Incidence of rash after amoxicillin treatment in children with infectious mononucleosis. Pediatrics 2013 May;131(5):e1424-e1427. [doi: 10.1542/peds.2012-1575] [Medline: 23589810]
39. Weber A, Arenz L. Penicillinallergie – wirklich [Article in German]. Infection&more 2024:26-31 [FREE Full text]
40. Worm M, Eckermann O, Dölle S, et al. Triggers and treatment of anaphylaxis: an analysis of 4,000 cases from Germany, Austria and Switzerland. Dtsch Arztebl Int 2014 May 23;111(21):367-375. [doi: 10.3238/arztebl.2014.0367] [Medline: 24939374]
41. Schmerzmitteltherapie bei ambulanten patientinnen und patienten mit und ohne tumordiagnose [Web page in German]. BARMER Institut für Gesundheitssystemforschung. URL: https://www.bifg.de/daten-und-analysen/arzneimittel-und-apotheken/arzneimittel/ambulante-schmerzmitteltherapie [accessed 2024-12-15]
42. Taghizadeh H, Benrath J. Schmerztherapie bei besonderen patientengruppen [Book in German]. In: Pocket Guide Schmerztherapie: Soforthilfe Bei Den Häufigsten Schmerzzuständen: Springer; 2024:491-584. [doi: 10.1007/978-3-662-67658-5_8]
43. Holstiege J, Schink T, Molokhia M, et al. Systemic antibiotic prescribing to paediatric outpatients in 5 European countries: a population-based cohort study. BMC Pediatr 2014 Jul 5;14:174. [doi: 10.1186/1471-2431-14-174] [Medline: 24997585]
44. Beisch N, Koch W, Schäfer C. ARD/ZDF-onlinestudie 2019: mediale internetnutzung und video-on-demand gewinnen weiter an bedeutung [Article in German]. Media Perspektiven 2019:374-388 [FREE Full text]
45. Bevölkerung nach Altersgruppen 2011 bis 2023 in ProzentDeutschland [Web page in German]. Statistisches Bundesamt. URL: https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bevoelkerung/Bevoelkerungsstand/Tabellen/bevoelkerung-altersgruppen-deutschland.html [accessed 2025-10-12]
46. Borges do Nascimento IJ, Pizarro AB, Almeida JM, et al. Infodemics and health misinformation: a systematic review of reviews. Bull World Health Organ 2022 Sep 1;100(9):544-561. [doi: 10.2471/BLT.21.287654] [Medline: 36062247]
47. Introducing GPT-5. OpenAI. URL: https://openai.com/index/introducing-gpt-5/ [accessed 2025-10-12]

Abbreviations

XSL•FO
RenderX

Edited by A Mavragani; submitted 02.04.25; peer-reviewed by R Poss-Doering, Y Wu; revised version received 18.08.25; accepted 21.08.25; published 12.11.25.

Please cite as:
Fuchs T, Hindelang M, Sitaru S, Zink A, Welzel J
Public Information Needs and Interest in Specific Food and Drug Allergy Disorders in Germany (2022–2024): Google Search Engine Analysis
JMIR Form Res 2025;9:e75395
URL: https://formative.jmir.org/2025/1/e75395
doi:10.2196/75395

© Tobias Fuchs, Michael Hindelang, Sebastian Sitaru, Alexander Zink, Julia Welzel. Originally published in JMIR Formative Research (https://formative.jmir.org), 12.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Quality and Dissemination of Uterine Fibroid Health Information on TikTok and Bilibili: Cross-Sectional Study

Lan Wang[1], BM; Yiwen Chen[2], BM; Duo Zhao[3], BM; Tao Xu[1], BM; Fu Hua[4], MD

[1]Department of Obstetrics and Gynecology, Huai'an 82 Hospital, Huai'an, China

[2]Department of Endocrinology, Nanjing University of Chinese Medicine, Nanjing Drum Tower Hospital, Nanjing, China

[3]Department of Orthopedics, The First People's Hospital of Guannan: Lianyun, Lianyungang, China

[4]Department of Gynecology, Nanjing Medical University, Huaian First People's Hospital, No. 1 West Huanghe Road, Huaiyin District, Huai'an, China

Corresponding Author:
Fu Hua, MD
Department of Gynecology, Nanjing Medical University, Huaian First People's Hospital, No. 1 West Huanghe Road, Huaiyin District, Huai'an, China

Abstract

Background: The rise of short-video platforms, such as TikTok (Douyin in China) and Bilibili, has significantly influenced how health information is disseminated to the public. However, the quality, reliability, and effectiveness of health-related content on these platforms, particularly regarding uterine fibroids, remain underexplored. Uterine fibroids are a common medical condition that affects a substantial proportion of women worldwide. While these platforms have become vital sources of health education, misinformation and incomplete content may undermine their efficacy.

Objective: This study aims to address these gaps by evaluating the quality and dissemination effectiveness of uterine fibroid–related health information on TikTok and Bilibili.

Methods: A total of 200 uterine fibroid-related videos (100 from TikTok and 100 from Bilibili) were selected through a keyword search. The videos were evaluated by 2 trained gynecological experts using the Global Quality Score (GQS) and a modified DISCERN (mDISCERN) tool. In addition, the Patient Education Materials Assessment Tool for Audio and Visual Materials was used to assess the understandability and actionability of the videos. Statistical analyses, including the Mann-Whitney *U* test, Spearman rank correlation, and stepwise regression analysis, were used to assess differences between platforms and identify predictors of video quality.

Results: The results indicated that TikTok outperformed Bilibili in terms of user engagement metrics, such as likes, comments, shares, and followers (all $P<.001$). However, Bilibili videos were generally longer than those on TikTok ($P<.001$). The videos on both platforms demonstrated suboptimal overall quality and reliability, reflected by median GQS score of 3 (IQR 3-4) for TikTok and the median GQS score of Bilibili is 3 (IQR 2-4). The median modified DISCERN scores were also low: 2 (IQR 2 - 2) for TikTok and 2 (IQR 2 - 2) for Bilibili, with no significant differences between the 2 platforms ($P=.62$ for GQS; $P=.18$ for mDISCERN). The videos on both platforms yielded comparable median scores for Patient Education Materials Assessment Tool-Understandability (PEMAT-U) and Patient Education Materials Assessment Tool-Actionability (PEMAT-A). The median score of PEMAT-U was 77% (IQR 69%-83%) for TikTok and 77% (IQR 69%-85%) for Bilibili. The PEMAT-A yielded a median score of 67% (IQR 33%-67%) for TikTok and 67% (IQR 0-67%) for Bilibili. Videos uploaded by medical professionals on TikTok had significantly higher quality scores compared to those uploaded by nonprofessionals. A moderate positive correlation was observed between the GQS and mDISCERN scores ($r=0.41$, $P<.01$), indicating an interrelationship between quality and reliability. Stepwise regression analysis identified "completeness score," "source," and "PEMAT scores" as significant predictors of video quality.

Conclusions: This study highlights the generally low quality of uterine fibroid–related health information on short-video platforms, although TikTok showed better performance in terms of engagement and quality. The involvement of medical professionals was found to enhance video quality. These findings underscore the need for improved oversight of health content on social media platforms and greater involvement of health care professionals to ensure the dissemination of accurate and reliable health information.

(JMIR Form Res 2025;9:e75120) doi:10.2196/75120

KEYWORDS

uterine fibroids; videos quality; TikTok; Bilibili; modified DISCERN; PEMAT-A/V; Spearman correlation analysis; stepwise regression analysis; Patient Education Materials Assessment Tool-Audiovisual Materials; Global Quality Score

XSL•FO
RenderX

Introduction

Uterine fibroids are the most prevalent benign tumors of the female reproductive system, affecting approximately 70% of women of reproductive age worldwide, particularly women of color, with an increasing incidence rate [1]. Approximately 30% of uterine fibroids manifest clinical symptoms, predominantly including abnormal uterine bleeding, compression symptoms affecting surrounding organs and tissues (such as constipation and urinary frequency), pelvic pain, and infertility [2]. Furthermore, uterine fibroids exert a significant impact on women's reproductive health, influencing fertility potential and pregnancy outcomes [3]. Primary risk factors for the condition include age, race, endogenous and exogenous hormonal factors, obesity, uterine infections, and lifestyle factors (including diet, caffeine and alcohol consumption, physical activity, stress, and smoking) [4]. Ultrasound is the preferred imaging modality, supplemented by magnetic resonance imaging when necessary [5]. Principal treatment modalities include pharmacological therapy, surgical intervention, and interventional procedures; surgical options comprise hysterectomy, myomectomy, uterine artery embolization, and focused ultrasound surgery [5]. Individualized and long-term management is paramount.

With the burgeoning ubiquity of the internet and the expanding reach of digital information, the internet has become a prevalent public resource for acquiring health knowledge and investigating personal health conditions [6]. As early as 2005, research indicated that the internet constituted one of the primary sources for adolescents seeking health information [7]. With advancements in technology, short videos centered on patients' diseases have been used in medical education, garnering positive evaluations [8]. Online short videos, as an emerging form of video media, have achieved substantial dissemination reach [9]. The public and health care professionals can engage in dialogue on health issues via social media, potentially enhancing health outcomes [10]. However, misinformation and disinformation on social media amplify the risks inherent in digital health communication, posing a challenge to the efficacy of public health measures [11].

Short-video platforms, such as TikTok/Douyin and Bilibili, have become significant channels through which the public accesses health information. While platforms like TikTok/Douyin (the international version of Douyin) and Bilibili have achieved immense popularity and reach [9,12], their role in disseminating health content necessitates a critical examination of the quality of this information. Understanding how users interact with and evaluate health content on these platforms is crucial, particularly in light of concepts such as digital health literacy and the potential for misinformation dissemination.

Prior studies have investigated the quality of health–related videos on these platforms concerning various conditions, such as adolescent vision health [13], lung cancer [14], liver cancer [15], brain tumor [16], breast cancer [17], heart failure [18], thyroid nodules [19], gallstone disease [20], chronic obstructive pulmonary disease [21], etc. Topics related to the female reproductive system, including oral contraceptives [22], contraception-related videos [23], and cervical cancer [24], have also been subjects of research. However, studies specifically focusing on the particular condition of uterine fibroids remain notably scarce. Given that misinformation in the field of reproductive health can adversely affect health outcomes, weaken public trust in medical systems, and potentially lead to inappropriate policy restrictions [25], evaluating the quality of information regarding uterine fibroids on these high-traffic platforms is of pressing importance.

While short-video platforms have rapidly become significant conduits for the public to access gynecological health knowledge, a critical gap exists in the systematic evaluation of the quality and reliability of content specifically pertaining to uterine fibroids on these platforms. Addressing this crucial research gap, this study aims to rigorously assess the quality, reliability, and dissemination characteristics of health information on uterine fibroids across TikTok and Bilibili. By analyzing differences between platforms and publisher types, this research seeks to provide a scientific foundation for improving the effective and accurate dissemination of uterine fibroids–related health information, thereby contributing to better public health literacy and outcomes in this area.

Methods

Data Collection and Screening

In this cross-sectional study, new accounts were registered on the TikTok and Bilibili platforms between March 3 and March 7, 2025. Video searches were conducted using the Chinese keyword "子宫肌瘤" (uterine fibroids), and the top 100 comprehensively ranked relevant videos from each platform were selected (Figure 1). Inclusion criteria included videos addressing uterine fibroids, presented in Chinese, with durations ranging from 20 seconds to 15 minutes. Exclusion criteria encompassed videos unrelated to uterine fibroids, duplicates or substantially similar videos, promotional advertisements, news reports, surgical or pathology-specific videos, outpatient consultation or ward round recordings, medical exam instructional videos, non–Chinese-language videos, videos with missing audio or poor sound quality, and those uploaded by users with unverified identities. Ultimately, 200 videos were included (100 from TikTok and 100 from Bilibili). The analysis was restricted to the top 100 videos per platform, as prior studies have demonstrated that videos beyond this threshold exert no significant influence on the analysis [15,26].

XSL•FO
RenderX

Figure 1. Flowchart for searching short videos on uterine fibroids (March 3 - 7, 2025; cross-sectional study).

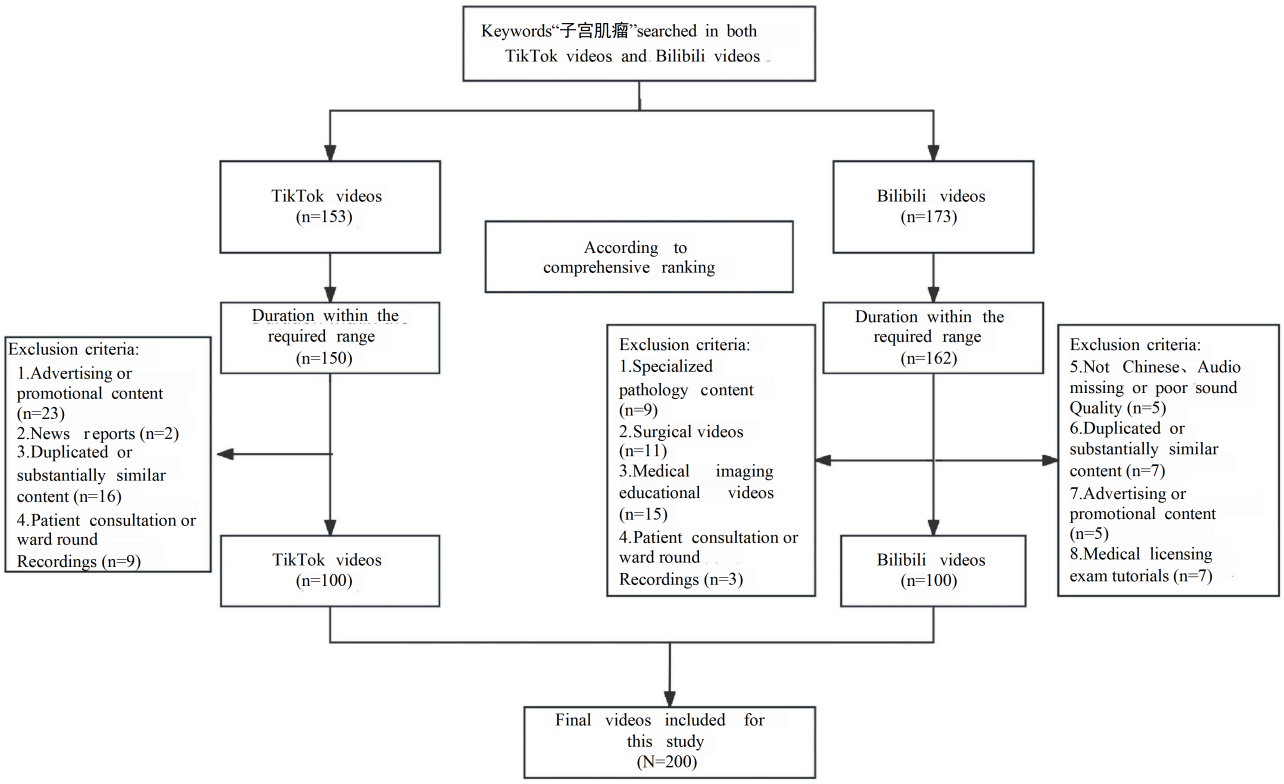

Data Collection Content

The uploader's identity, follower count, video duration, number of likes, comments, favorites, and shares were recorded. All extracted data were documented in Excel (Microsoft Corp). The dataset is available in Multimedia Appendix 1.

Platforms were designated as 1=TikTok and 2=Bilibili. Videos were categorized into 4 groups based on source and 5 groups based on content. Video sources were classified as follows: (1) professional individuals, including gynecologists, radiologists, and oncologists; (2) nonprofessional individuals, encompassing practitioners of traditional Chinese medicine, physicians from other disciplines, individuals with medical backgrounds, and those without medical backgrounds; (3) professional institutions; and (4) nonprofessional institutions. Video content was classified by theme as follows: (1) disease-related knowledge, (2) etiology, (3) symptoms, (4) treatment, and (5) prevention and prognosis.

Quality Assessment Tools

Global Quality Score

Ranging from 1 to 5 points, with higher scores indicating superior quality [27], this metric evaluates accuracy, authority, completeness, and educational value (Table S1 in Multimedia Appendix 2).

Modified DISCERN Tool

Comprising 5 questions adapted from the original DISCERN tool, it is used to assess the reliability of health information [28], evaluating clarity, relevance, traceability, scientific basis, and impartiality. Each dimension is scored 1 point for "yes" and 0 points for "no" (Textbox 1).

Textbox 1. Modified DISCERN quality criteria for assessing the reliability of video. (1 point for answer "yes," 0 point for answer "no").

Reliability score

- Is the video clear, concise, and understandable?
- Are valid sources cited?
- Is the content presented balanced and unbiased?
- Are additional sources of content listed for patient reference?
- Are areas of uncertainty mentioned?

The Patient Education Materials Assessment Tool

Using the Patient Education Materials Assessment Tool (PEMAT) [29] for audiovisual materials (PEMAT-A/V) with an automated scoring form, which assesses the understandability and actionability of videos.

XSL•FO
RenderX

Statistical Analysis

Data analysis was performed using SPSS version 27.0 (IBM Corp). Descriptive statistics were presented as median (IQR). The Mann-Whitney *U* test was used to compare quality differences between platforms (TikTok vs Bilibili), publisher, and across video content categories. Spearman's rank correlation coefficient (ρ) and stepwise regression analysis were calculated to evaluate the association between quality scores and engagement metrics (likes, shares, and comments). The Cohen κ coefficient was used to assess interrater reliability among experts. Statistical significance was set at $P<.05$ for all tests.

Ethical Considerations

This study did not use clinical data, human specimens, or laboratory animals. All information was derived from publicly released videos on TikTok and Bilibili platforms, with no data involving personal privacy issues. Furthermore, this research did not involve any interaction with users; therefore, ethical review was not required.

Results

Videos Characteristics

The overall characteristics of the videos are summarized in Table 1. TikTok videos exhibit significantly higher numbers of likes, comments, favorites, shares, and publisher follower counts compared to Bilibili videos (all $P<.001$), whereas Bilibili videos are notably longer in duration than TikTok videos ($P<.001$).

Table . Characteristics of uterine fibroid videos on TikTok and Bilibili.

Variable	TikTok (n=100), median (IQR)	Bilibili (n=100), median (IQR)	Wilcoxon rank-sum test	
			Z score	*P* value
Likes	431 (249 - 1238)	44 (9 - 267)	−7.68	<.001
Comments	35 (12 - 93)	4 (0 - 28)	−6.5	<.001
Collections	129 (67 - 437)	30 (4 - 150)	−5.71	<.001
Shares	72 (32 - 220)	14 (2 - 108)	−4.95	<.001
Duration	66 (46 - 91)	138 (80 - 294)	−7.2	<.001
Followers	115,000 (27,750 - 511,750)	14,000 (1212 - 65,000)	−6.76	<.001
GQS[a]	3 (3 - 4)	3 (2 - 4)	−0.49	.62
mDIS[b]	2 (2 - 2)	2 (2 - 2)	−1.35	.18
PEMAT-U[c], %	77 (69 - 83)	77 (69 - 85)	−0.72	.47
PEMAT-A[d], %	67 (33 - 67)	67 (0 - 67)	−0.29	.77

[a]GQS: Global Quality Score.

[b]mDIS: modified DISCERN.

[c]PEMAT-U: Patient Education Materials Assessment Tool-Understandability score.

[d]PEMAT-A: Patient Education Materials Assessment Tool-Actionability score.

The median Global Quality Score (GQS) for TikTok videos is 3 (IQR 3 - 4), with a median modified DISCERN (mDISCERN) score of 2 (IQR 2 - 2). For Bilibili videos, the median GQS and mDISCERN scores are 3 (IQR 2 - 4) and 2 (IQR 2 - 2), showing no statistically significant differences between the 2 platforms ($P=.62$ for GQS and $P=.18$ for mDISCERN), respectively. Interrater agreement between the 2 evaluators was quantified using Cohen κ. According to the criteria proposed by Landis and Koch [30], a κ value >0.8 indicates excellent agreement, values between 0.6 and 0.8 denote substantial agreement, values between 0.4 and 0.6 reflect moderate agreement, and values <0.4 suggest poor agreement. The Cohen κ values for GQS and mDISCERN scores were 0.727 and 0.725, indicating substantial interrater agreement, respectively.

The median PEMAT-U (understandability) score for TikTok videos is 77% (IQR 69% - 83%), with a median PEMAT-A (actionability) score of 67% (IQR 33% - 67%). For Bilibili videos, the median PEMAT-U and PEMAT-A scores are 77% (IQR 69% - 85%) and 67% (IQR 0% - 67%), with no statistically significant differences between the platforms ($P=.47$ for PEMAT-U and $P=.77$ for PEMAT-A), respectively. Tables S2-S4 in Multimedia Appendix 2 along with Table 2, illustrate the sources and content types of videos on TikTok and Bilibili.

XSL•FO
RenderX

Table . Percentage distribution of uterine fibroid videos by source content and category on TikTok and Bilibili.

		Distribution of uterine fibroid videos, %
Sources of TikTok videos		
	Nonprofessional individuals	1
	Nonprofessional institutions	0
	Professional individuals	83
	Professional institutions	16
Sources of Bilibili videos		
	Nonprofessional individuals	6
	Nonprofessional institutions	3
	Professional individuals	52
	Professional institutions	39
Content categories of TikTok videos		
	Disease knowledge	31
	Etiology	5
	Prevention and prognosis	11
	Symptoms	5
	Treatment	48
Content categories of Bilibili videos		
	Disease knowledge	32
	Etiology	3
	Prevention and prognosis	30
	Symptoms	4
	Treatment	31

On TikTok, professional individuals uploaded the majority of videos, accounting for 83% (83/100), a significantly high proportion, followed by nonprofessional individuals (16/100, 16%) and professional institutions (1/100, 1%). Regarding video content on TikTok, the majority focused on treatment (48/100, 48%), followed by disease knowledge (31/100, 31%), prevention and prognosis (11/100, 11%), etiology (5/100, 5%), and symptoms (5/100, 5%). On Bilibili, professional individuals accounted for 52% of video sources (52/100), a proportion significantly lower than on TikTok, while non-professional individuals comprised 39% (39/100). This disparity is closely linked to the publisher verification policies of the two platforms: TikTok enforces stricter credential verification for medical personnel and imposes rigorous restrictions on medical advertisements and pharmaceutical promotions. On Bilibili, content pertaining to disease knowledge (32%), treatment (31%), and prevention and prognosis (30%) is represented in roughly equal proportions.

Video Content Distribution

We further analyzed the content discussed in the included short videos (Table S5 in Multimedia Appendix 2) and evaluated their completeness (Multimedia Appendix 3) across 7 domains: epidemiology, etiology, symptoms, diagnosis, treatment, prevention, and prognosis. Each domain was scored from 0 to 2 points, with a total possible score of 14 points. A subdomain mentioned once received 1 point (partial explanation), more than 3 mentions received 2 points (complete explanation), and no mention received 0 points. The scores for each domain are presented in Figure 2. We found that the treatment of uterine fibroids was the most frequently discussed topic, addressed in 70% (140/200) of the total 200 videos (123/200, 61.5% partial explanation and 17/200, 8.5% complete explanation). On TikTok, 77% (77/100) of videos provided partial or full explanations of treatment, compared to 63% (63/100) on Bilibili. Symptoms followed as the second most common topic, appearing in 49% (98/200) of the total videos. Etiology and diagnosis of uterine fibroids also garnered relative public interest, with similar proportions (30.5% vs 32%). However, few videos offered detailed health information on the epidemiology of uterine fibroids, with only 12% (12/100) of TikTok videos providing partial (11/100, 11%) or full (1/100, 1%) explanations. Notably, no videos provided a detailed explanation of prevention strategies for uterine fibroids.

XSL•FO
RenderX

Figure 2. Percentage distribution of explanatory coverage across different domains in uterine fibroid video content.

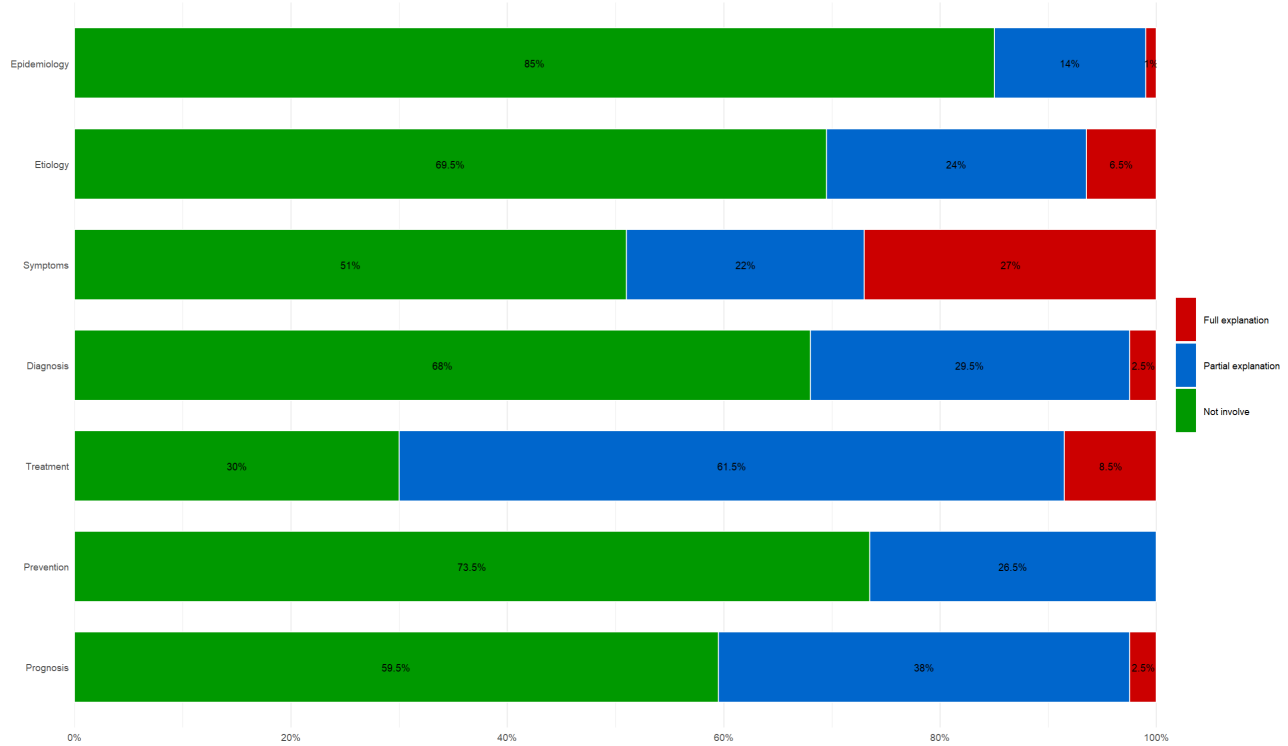

Videos covered the 7 predefined content domains of the completeness scoring to varying degrees (Table S6 in Multimedia Appendix 2, Figure 3). Nonprofessional institutions scored highest in the “symptoms” (mean 1.50, SD 0.84) and “prognosis” (mean 0.83, SD 0.41) domains but recorded a score of 0 in the “prevention” domain, indicating a complete absence of prevention-related information from this source. Professional individuals performed well in the “treatment” (mean 0.83, SD 0.58) and “symptoms” (mean 0.75, SD 0.86) domains but scored lowest in “epidemiology” (mean 0.10 SD 0.32), suggesting limited coverage in this area. Professional institutions achieved higher scores in the “etiology” and “symptoms” domains (mean 0.75, SD 0.96) but scored lower in “prevention” and “prognosis” (mean 0.25, SD 0.50), possibly due to a focus on academic rigor at the expense of practical utility. The mean score for “symptoms” from non-professional institutions (1.50, SD 0.84) was significantly higher than that of other sources, likely because their content aligns more closely with public needs. Based on Figure 4, professional institutions emerge as the most reliable sources for comprehensive information, demonstrating broad coverage across various disease topics coupled with higher completeness. Conversely, information originating from nonprofessional individuals exhibits notable limitations in overall comprehensiveness. The radar plot reveals weaker coverage in several critical domains, including epidemiology, diagnosis, and prognosis, while the violin plot indicates a wider distribution of their completeness scores. While the radar chart emphasizes the breadth of topics addressed, the violin plot focuses on the completeness score. The congruence and discrepancies between these visualizations provide valuable insights into the strengths and weaknesses of diverse information sources.

XSL•FO
RenderX

Figure 3. Multidimensional content quality assessment of uterine fibroid videos using radar charts, stratified by source.

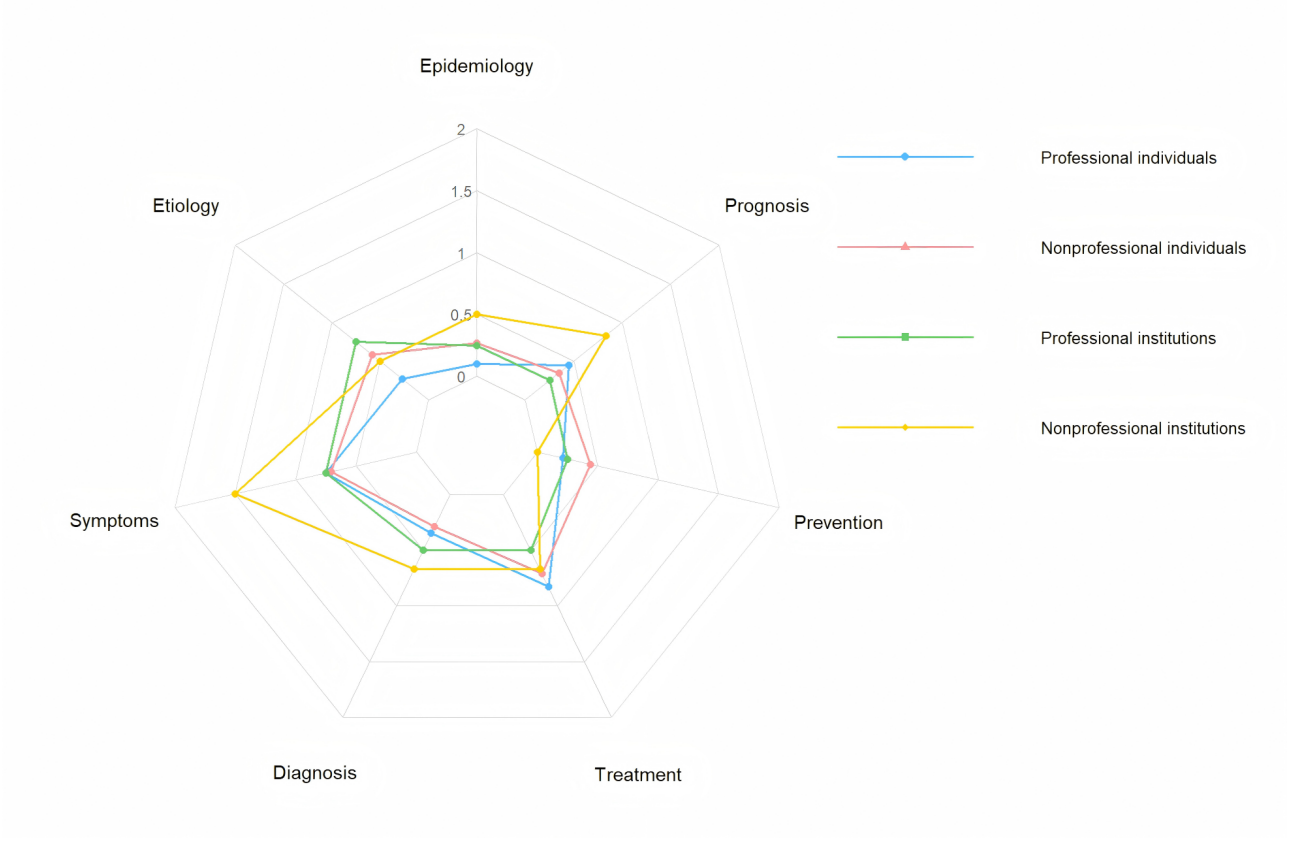

Figure 4. Probability density distribution comparison of content integrity composite scores for uterine fibroid videos.

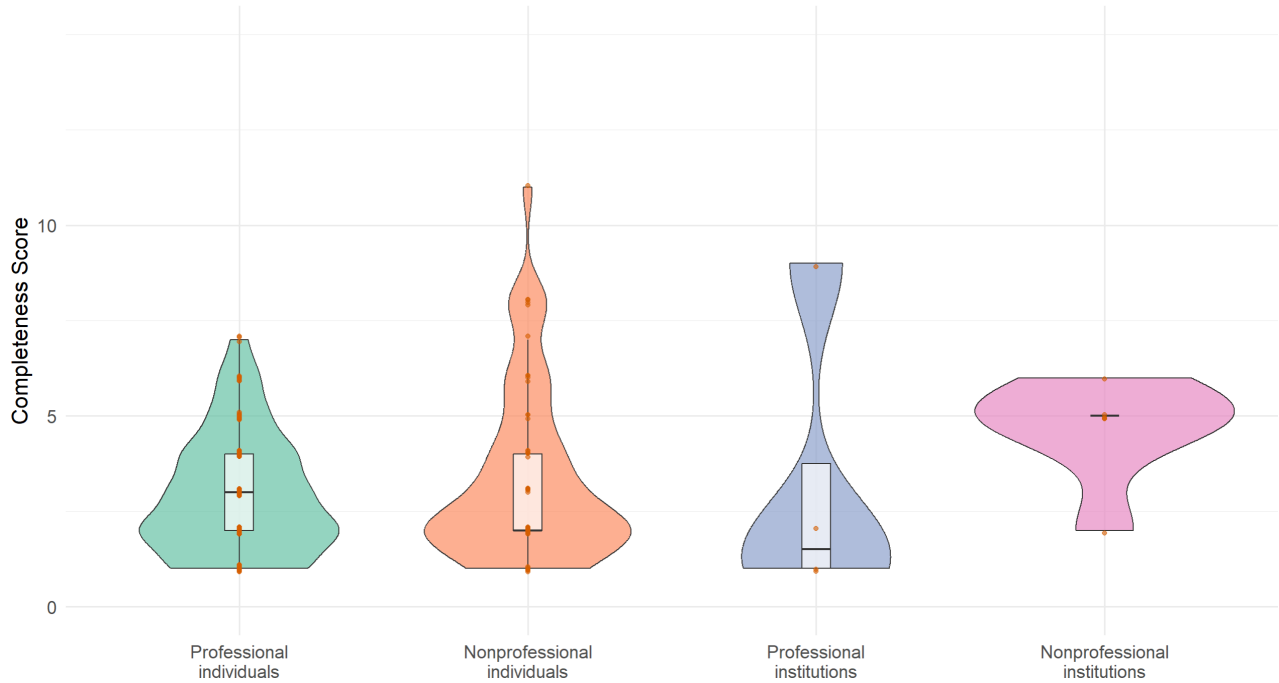

Video Quality and Reliability Assessments

The GQS and mDISCERN scores were compared across different platforms, video sources, and content types. The GQS scores for TikTok and Bilibili (Figure 5A) are predominantly clustered around 3, indicating moderate video quality. Although Bilibili videos exhibit a concentration of higher scores, a greater number of videos are clustered at the lower end, which drags down the overall quality level. The mDISCERN scores (Figure 5B) reveal moderate reliability for both platforms. There is no significant difference in quality and reliability between the 2 platforms, with TikTok slightly outperforming Bilibili.

XSL•FO
RenderX

Figure 5. Global Quality Score and modified DISCERN for Uterine Fibroid Videos. (A) Comparison of Global Quality Scores between TikTok and Bilibili videos. (B) Comparison of modified DISCERN scores between TikTok and Bilibili videos. (C) Global Quality Scores across different video sources. (D) Modified DISCERN scores across different video sources. (E) Global Quality Scores across different content categories. (F) Modified DISCERN scores across different content categories.GQS: Global Quality Score; mDIS: modified DISCERN.

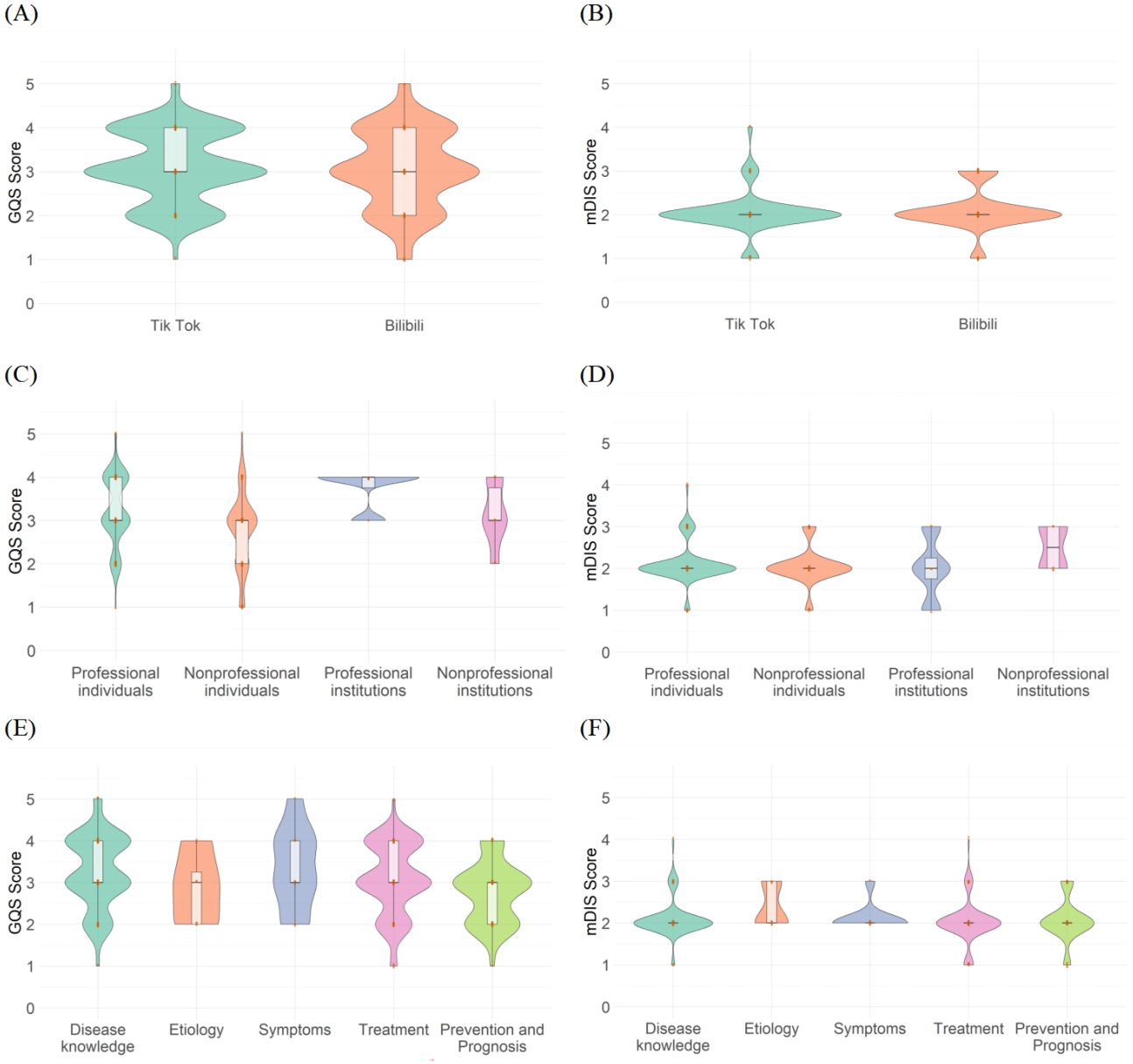

Regarding video sources, GQS scores for professional individuals and professional institutions show a concentration of higher scores, reflecting superior quality. Nonprofessional institutions, however, achieve a median GQS score of 4 (Figure 5C), significantly higher than other groups (professional groups and nonprofessional individuals, both at 3), and a median mDISCERN score of 3 (Figure 5D), surpassing the median of 2 for other groups. This suggests that content from nonprofessional institutions excels in both quality and reliability, challenging the traditional assumption that professional sources consistently dominate. However, this finding is tempered by the limited number of videos from nonprofessional institutions, rendering it unrepresentative of the broader dataset. The score distribution for nonprofessional individuals displays considerable variability, underscoring the heterogeneity in content quality and reliability within this group.

In terms of content domains, the GQS scores for disease knowledge (Figure 5E) exhibit a concentration of high scores with low dispersion, indicating high quality and strong information authority. Similarly, in mDISCERN scores (Figure 5F), disease knowledge demonstrates greater reliability, outperforming other categories.

Video Understandability and Actionability

A comparison of Patient Education Materials Assessment Tool for Understandability (PEMAT-U) scores across video sources on different platforms (Figure 6A) reveals that TikTok videos exhibit a concentration of higher scores, outperforming Bilibili in understandability. Although the highest understandability score was derived from Bilibili. This, combined with the shorter duration of TikTok videos, suggests that their content is concise, with strong visual aids. However, lower scores stem from oversimplification and insufficient information completeness.

XSL•FO
RenderX

In terms of PEMAT-A (actionability) scores (Figure 6B), TikTok similarly surpasses Bilibili.

Figure 6. Comparative analysis of video understandability (PEMAT-U) and Actionability (PEMAT-A) of Uterine Fibroid Videos. (A) Patient Education Materials Assessment Tool-Understandability scores of videos across TikTok and Bilibili. (**B**) Patient Education Materials Assessment Tool-Actionability scores of videos across TikTok and Bilibili. (**C**) Patient Education Materials Assessment Tool-Understandability scores of videos across different sources. (**D**) Patient Education Materials Assessment Tool-Actionability scores of videos across different sources. (**E**) Patient Education Materials Assessment Tool-Understandability scores of videos across different content categories. (**F**) Patient Education Materials Assessment Tool-Actionability scores of videos across different content categories. PEMAT-A: Patient Education Materials Assessment Tool-Actionability. PEMAT-U: Patient Education Materials Assessment Tool-Understandability.

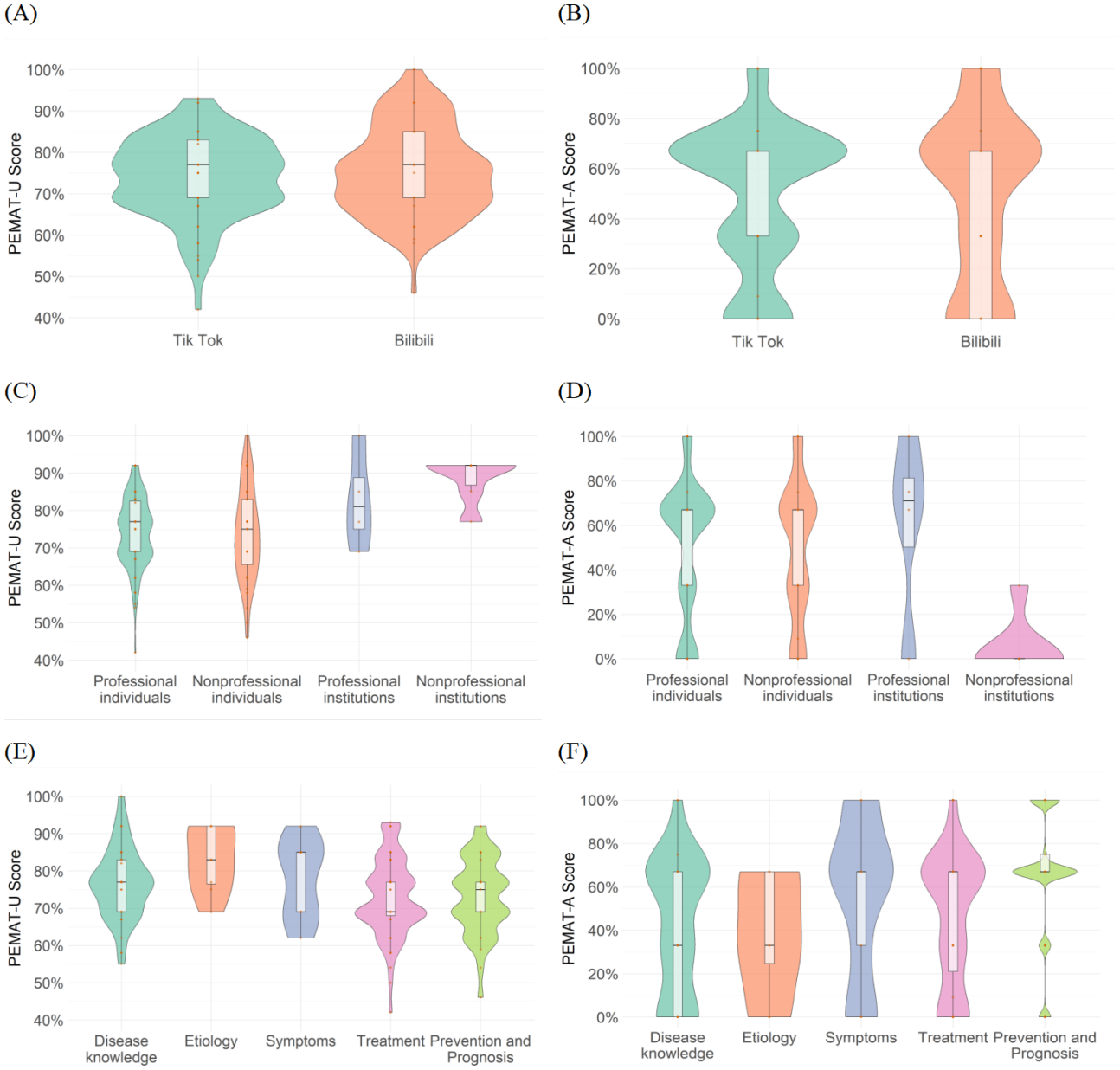

Regarding understandability across video sources, nonprofessional individuals demonstrate a stronger advantage (Figure 6C), likely due to their use of more colloquial language, while professional individuals show significant variation; the scores ranged from a minimum of 42% to a maximum of 100%. For actionability, professional individuals, nonprofessional individuals, and professional institutions all have a median score of 67%, with no notable differences; however, professional institutions exhibit marked polarization. In contrast, videos from nonprofessional institutions consistently show lower actionability, indicating poor feasibility (Figure 6D).

Concerning video content, the “symptoms” domain achieves the highest understandability while the “prevention and prognosis” domain scores lower (Figure 6E). Overall, no understandability scores fall below 42%, whereas actionability scores include a notable proportion of low values, even reaching 0% (Figure 6F).

Correlation and Stepwise Regression Analysis

Spearman rank correlation analysis was used to assess the relationships among quality scores, interaction data, and video content, as shown in Figure 7. The results indicate strong positive correlations between likes and comments, favorites

XSL•FO
RenderX

(0.88<r<0.94), and shares, while follower count exhibits moderate to strong positive correlations with interaction data (0.45<r<0.63). The correlation between GQS and mDISCERN scores is r=0.41, with a significance level of P<.01, suggesting a moderate positive relationship between the two. This finding implies that higher quality scores may be associated with improved content distribution characteristics. Etiology demonstrates weak positive correlations (0.19<r<0.27) with interaction data (likes, comments, favorites, and shares), indicating that viewers may have a heightened interest in etiological content. Video interaction metrics exhibit only a slight positive correlation with both GQS and mDISCERN scores (0.05<r<0.10).

Video understandability (PEMAT-U) shows weak positive correlations with quality (GQS) and reliability (mDISCERN), with specific r values of 0.21 and 0.29, suggesting that content with higher quality scores tends to exhibit greater user-friendliness, reflecting a stronger alignment between content design and user needs. The actionability of videos, as assessed by the Patient Education Materials Assessment Tool for Audiovisual Materials (PEMAT-A), shows only a moderately low positive correlation with the prevention domain (r=0.30), while exhibiting a weak negative correlation with diagnosis (r=−0.14). In addition, other negative correlations among video parameters were identified, such as those between prevention and treatment (r=−0.38), symptoms (r=−0.18), diagnosis (r=−0.17), and prognosis (r=−0.21), as well as between etiology and treatment (r=−0.21).

Figure 7. Correlation matrix of variables assessing the characteristics, quality, understandability, and actionability of uterine fibroid videos. DIS:mDISCERN; GQS: Global Quality Score; PEMAT-A: Patient Education Materials Assessment Tool-Actionability; PEMAT-U: Patient Education Materials Assessment Tool-Understandability.

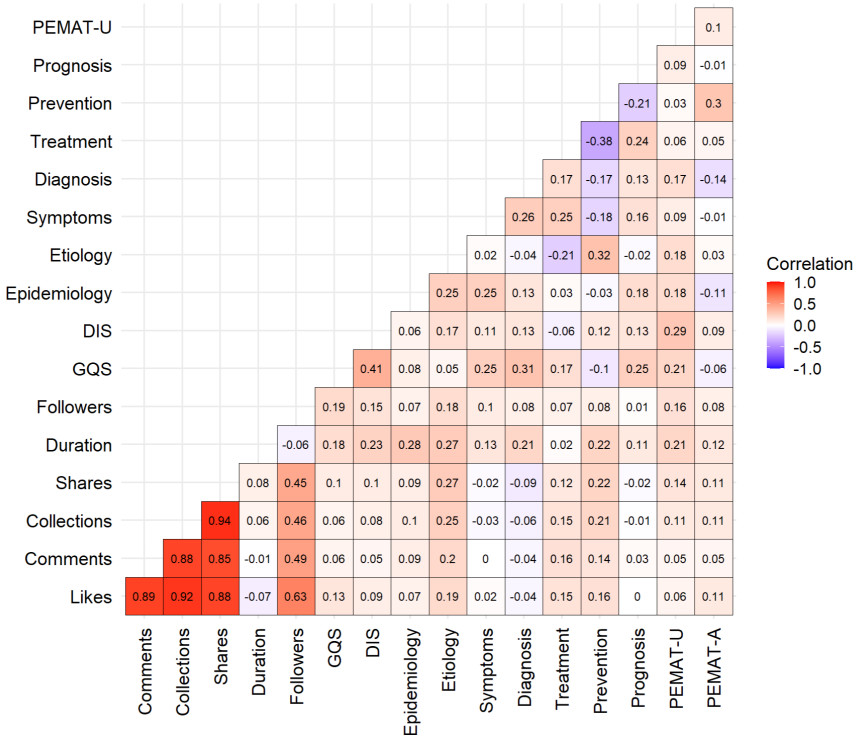

Stepwise regression analysis for GQS yielded 4 models, with each subsequent model incorporating additional predictors (Tables 3-6). The final model (model 4) demonstrated statistical significance (F_4=18.346; P<.001), accounting for 25.9% of the variance in GQS (adjusted R^2=0.259). This model, exhibiting the strongest predictive capacity, included "Completeness Score" (β=0.368; t_{195}=5.732; P<.001), "Source" (β=–0.342; t_{195}=−5.385; P<.001), "PEMAT-U" (β=0.219; t_{195}=3.389; P<.001), and "PEMAT-A" (β=−0.154; t_{195}=−2.475; P<.05) as predictors. The 95% CIs for the unstandardized coefficients (B) confirmed the significance of each predictor: Completeness Score (0.111-0.228), Source (−0.588 to −0.273), PEMAT-U (0.008-0.029), and PEMAT-A (−0.007 to −0.001). Tolerance values for the predictors ranged from 0.895 to 0.964, and variance inflation factor values ranged from 1.000 to 1.118, indicating the absence of multicollinearity and supporting the stability of the regression coefficients.

XSL•FO
RenderX

Table . Stepwise regression analysis predicting Global Quality Scores of uterine fibroid videos on TikTok and Bilibili.

Model	Predictors included	R^2	Adjusted R^2	SE of the estimate	*F* change (*df*)	*P* value	ANOVA	
							F test (*df*)	*P* value
1	Constant and completeness score	0.133	0.129	0.799	30.44 (1, 198)	<.001	30.44 (1, 198)	<.001
2	Constant and completeness score and source	0.213	0.205	0.763	20.095 (1, 197)	<.001	26.735 (2, 197)	<.001
3	Constant and completeness score, source, and PEMAT-U[a]	0.251	0.239	0.747	9.709 (1, 196)	.002	21.848 (3,196)	<.001
4	Constant and completeness score, source, and PEMAT-U and PEMAT-A[b]	0.273	0.259	0.737	6.217 (1, 195)	.01	18.346 (4,195)	<.001

[a]PEMAT-U: Patient Education Materials Assessment Tool-Understandability.

[b]PEMAT-A: Patient Education Materials Assessment Tool-Actionability.

XSL•FO
RenderX

Table . Stepwise regression coefficients, statistical significance, and collinearity assessment for the prediction of Global Quality Scores of uterine fibroid videos.

Model and predictors	Unstandardized coefficients		Standardized coefficients	*t* test (*df*)	*P* value	Collinearity statistics	
	B (95% CI)	SE	β			Tolerance	VIF[a]
1							
Constant	2.554 (2.238 to 2.771)	0.110	__[b]	23.228 (198)	<.001	—	—
Completeness score	0.168 (0.108 to 0.228)	0.030	0.365	5.517 (198)	<.001	1.000	1.000
2							
Constant	2.992 (2.709 to 3.275)	0.143	—	20.867 (197)	<.001	—	—
Completeness score	0.191 (0.133 to 0.249)	0.030	0.415	6.462 (197)	<.001	.970	1.031
Source	–0.362 (–0.521 to –0.203)	0.081	–0.288	–4.483 (197)	<.001	0.970	1.031
3							
Constant	1.841 (1.062 to 2.620)	0.395		4.661 (196)	<.001	0.906	1.103
Completeness score	0.167 (0.108 to 0.226)	0.030	0.363	5.582 (196)	<.001	0.952	1.051
Source	–0.396 (–0.553 to –0.239)	0.080	–0.315	-4.969 (196)	<.001	0.904	1.106
PEMAT-U	0.017 (0.006 to 0.028)	0.005	0.203	3.116 (196)	.002	—	—
4							
Constant	1.982 (1.204 to 2.579)	0.394	—	5.028 (195)	<.001	0.905	1.104
Completeness score	0.169 (0.111 to 0.228)	0.030	0.368	5.732 (195)	<.001	0.923	1.083
Source	–0.430 (–0.588 to –0.273)	0.080	–0.342	–5.385 (195)	<.001	0.895	1.118
PEMAT-U[c]	0.018 (0.008 to 0.029)	0.005	0.219	3.389 (195)	<.001	0.964	1.038
PEMAT-A[d]	–0.004 (–0.007 to –0.001)	0.002	–0.154	–2.475 (195)	.01	—	—

[a]VIF: variance inflation factor.

[b]Not applicable.

[c]PEMAT-U: Patient Education Materials Assessment Tool-Understandability.

[d]PEMAT-A: Patient Education Materials Assessment Tool-Actionability.

XSL•FO
RenderX

Table . Stepwise regression analysis predicting modified DISCERN scores of uterine fibroid videos on TikTok and Bilibili.

Model	Predictors included	R^2	Adjusted R^2	SE of the estimate	F change (*df*)	*P* value	ANOVA	
							F test (*df*)	*P* value
1	Constant and PEMAT-U[a]	0.09	0.086	0.469	19.648 (1, 198)	<.001	19.648 (1, 198)	<.001
2	Constant, PEMAT-U, and completeness score	0.113	0.104	0.464	5.077 (1, 197)	.03	12.565 (2, 197)	<.001
3	Constant, PEMAT-U, completeness score and treatment	0.131	0.118	0.461	4.019 (1, 196)	.05	9.844 (3, 196)	<.001
4	Constant, PEMAT-U, completeness score, treatment, and epidemiology	0.152	0.135	0.456	4.86 (1, 195)	.03	8.744 (4, 195)	<.001

[a]PEMAT-U: Patient Education Materials Assessment Tool-Understandability.

XSL•FO
RenderX

Table . Stepwise regression coefficients, statistical significance, and collinearity assessment for the prediction of modified DISCERN scores of uterine fibroid videos.

Model (Predictors)	Unstandardized coefficients		Standardized coefficients	*t* (*df*)	*P* value	Collinearity statistics	
	B (95% CI)	SE	β			Tolerance	VIF[a]
1							
Constant	0.993 (0.508 to 1.479)	0.246	—[b]	4.035 (198)	<.001		—
PEMAT-U[c]	0.014 (0.008 to 0.021)	0.003	0.300	4.433 (198)	<.001	1.000	1.000
2							
Constant	1.023 (0.542 to 1.505)	0.244	—	4.193 (197)	<.001	—	
PEMAT-U	0.012 (0.006 to 0.019)	0.003	0.256	3.669 (197)	<.001	0.922	1.085
Completeness score	0.042 (0.005 to 0.078)	0.018	0.157	2.253 (197)	.03	0.922	1.085
3							
Constant	1.084 (0.603 to 1.566)	0.244	—	4.442 (196)	<.001	—	—
PEMAT-U	0.012 (0.006 to 0.019)	0.003	0.252	3.632 (196)	<.001	0.921	1.086
Completeness score	0.058 (0.019 to 0.098)	0.020	0.221	2.897 (196)	.004	0.763	1.311
Treatment	–0.124 (–0.246 to –0.002)	0.062	–0.147	–2.005 (196)	.04	0.821	1.218
4							
Constant	1.025 (0.545 to 1.505)	0.243	—	4.213 (195)	<.001	—	
PEMAT-U	0.013 (0.006 to 0.019)	0.003	0.260	3.781 (195)	<.001	0.918	1.089
Completeness score	0.087 (0.040 to 0.134)	0.024	0.330	3.655 (195)	<.001	0.534	1.872
Treatment	–0.154 (–0.277 to –0.030)	0.063	–0.183	–2.453 (195)	.02	0.783	1.278
Epidemiology	–0.221 (–0.419 to –0.023)	0.100	–0.177	–2.205 (195)	0.03	0.671	1.491

[a]VIF: variance inflation factor.

[b]Not applicable.

[c]PEMAT-U: Patient Education Materials Assessment Tool-Understandability.

Similarly, stepwise regression analysis for mDISCERN generated 4 models, with each building upon the previous by adding more predictors (Tables 5 and 6). The final model (model 4) was statistically significant (F_4=8.744; P<.001) and explained 13.5% of the variance in mDISCERN scores (Adjusted R^2=0.135). This model, which showed the greatest predictive power, incorporated PEMAT-U (β=0.260; t_{195}=3.781; P<.001), "Completeness Score" (β=0.330; t_{195}=3.655; P<.001), "Treatment" (β=−0.183; t_{195}=−2.453; P<.05), and "Epidemiology" (β=−0.177; t_{195}=−2.205; P<.05) as predictors. The 95% CIs for the unstandardized coefficients (B) corroborated the significance of each predictor: PEMAT-U (0.006-0.019), Completeness Score (0.040-0.134), Treatment (−0.277 to −0.030), and Epidemiology (−0.419 to −0.023). Variance inflation factor values for all models remained below 2, confirming the absence of significant multicollinearity. PEMAT-U and "Completeness Score" positively influenced mDISCERN scores, while "Treatment" and "Epidemiology" exhibited a negative association.

XSL•FO
RenderX

Discussion

Principal Findings

This study reveals that the quality of health-related content on uterine fibroids across short video platforms is generally inadequate, aligning with previous research on liver cancer [15]. No significant quality differences were observed between TikTok and Bilibili, although TikTok showed a slight edge, particularly in videos produced by experts, which tended to be of higher quality.

During data collection, Bilibili excluded numerous in-depth instructional videos on surgery, pathology, and medical examinations, likely due to their specialized focus on medical professionals rather than the general public. In contrast, TikTok lacked any comparable content, suggesting Bilibili's greater propensity for hosting academically oriented, in-depth medical resources. However, platform policies may prioritize broader appeal over highly specialized educational material. For platforms, striking a balance between commercial attractiveness and educational value is crucial for fostering a healthier, more accurate, and sustainable content ecosystem.

An interesting pattern emerged when comparing the sources of videos on both platforms. On TikTok, the majority of video creators were professionals with medical expertise, such as gynecologists, radiologists, oncologists, and practitioners of traditional Chinese medicine who also engage in modern health care. The only exception was a fitness blogger who posted a video about uterine fibroids, but this content was excluded from the analysis due to its relatively low overall ranking. In contrast, Bilibili featured a broader range of video creators, including not only medical professionals but also a small number of medical institutions, nonmedical organizations, and individuals without formal medical backgrounds. This difference in content creators reflects, in part, the distinct demographic profiles of the 2 platforms' audiences.

The completeness scores of video content are generally low, typically below 7 points, and show a moderate positive correlation with the length of short videos. Creating high-quality content that is both thematically relevant and logically coherent within a brief time frame has become a key area of focus for content creators. The advantages of professional individuals and institutions, stemming from their resources, experience, and expertise, likely contribute to their better performance in this area. In contrast, nonprofessional creators tend to show more variation in their results, potentially influenced by factors such as their knowledge base, available resources, and level of training. Future research could investigate how these variables affect the performance of different groups in producing such content, providing more insightful guidance for practice and policy development in related domains.

Correlation and Stepwise Regression Analysis of Video Quality, Reliability, and Video Characteristics

The weak correlation observed between video interaction metrics and GQS/mDISCERN scores suggests that audience engagement, while significant, is not a reliable indicator of video content quality and reliability. This finding corroborates the results reported by Cui et al [31], but diverges from those of Li et al [32] concerning osteoporosis. These discrepancies highlight the need for enhanced public discernment in differentiating between high-quality and low-quality video resources.

The primary preventive measures for uterine fibroids include regular check-ups (for early detection and intervention), weight management, avoidance of prolonged exogenous estrogen exposure, a healthy diet (eg, reducing red meat intake), and chronic disease management. These preventive measures are relatively feasible for the general public, thereby accounting for the positive correlation between video actionability and prevention. In contrast, diagnosis requires specialized medical knowledge, rendering it less practicable for the general public and thus explaining the observed negative correlation.

The stepwise regression analysis revealed a significant positive predictive relationship between "Information Completeness Score" and GQS scale scores, while "Video Source" exhibited a negative association. Specifically, videos originating from nonprofessionals tended to receive lower GQS ratings from viewers. This could be attributed to potential limitations in the accuracy, professional rigor, and presentation quality of content from nonprofessional sources compared to that from professionals and specialized medical institutions, potentially failing to meet viewers' demands for high-quality information and consequently leading to lower evaluations. Furthermore, even videos from professional channels might receive lower GQS ratings from viewers with nonspecialized backgrounds if the content is overly technical and uses obscure terminology. Future research could further investigate the specific content characteristics of videos from different sources and their impact on viewers' GQS evaluations. Conversely, the PEMAT-U scale and Information Completeness Score positively predicted mDISCERN scores, whereas variables related to "Treatment" and "Epidemiology" showed negative associations. It is crucial to interpret the negative coefficients between variables and scores with caution, as these results do not necessarily imply causation.

Practical Significance

Short videos, known for their conciseness, effective visual communication, ease of production, and high interactivity, address the varied preferences of different audience groups. To stand out in a highly competitive environment, both individual creators and corporate organizations must focus on the 3 key principles: "visual appeal, emotional connection, and actionable guidance." However, in response to market pressures, clickbait titles have become increasingly common, often paired with content that is shallow or misleading. This trend highlights the urgent need to improve the overall quality of video production.

Globally, TikTok has emerged as the fastest-growing social media platform among children and young people [33], with substantial potential as a public health tool [34]. Personal health advice disseminated via social media appears to exert a profound influence on adolescents [35], and misinformation in some videos can rapidly proliferate, adversely affecting public health [36]. A study on tonsil stones revealed that children attempting untested remedies from videos suffered significant harm [35].

XSL•FO
RenderX

Platform regulation and expert engagement are crucial for enhancing the quality of online health information. Based on the Elaboration Likelihood Model (ELM), the Heuristic-Systematic Model (HSM), and the Media Trust Model, the following actionable recommendations are proposed: (1) Establish a "Health Video Tagging Mechanism": implement a standardized system enabling creators to categorize health-related videos with specific tags (eg, symptoms, treatment, and prevention) and indicate the level of evidence or source type (eg, expert opinion, personal experience, and research-based); (2) enhance expert verification methods: implement more rigorous and standardized procedures, requiring proof of licensure and establishing a transparent process for users to report potentially misleading or unqualified "expert" accounts; (3) prioritize verified expert content algorithmically: content originating from accounts confirmed as "experts" should receive preferential algorithmic weighting to increase its visibility and reach; this directly addresses the need for expert involvement and leverages platform algorithms to promote credible information dissemination; and (4) promote evidence-based content structures: provide creators with guidelines and templates for structuring their videos to clearly and concisely emphasize essential, evidence-based information. Encourage the use of interactive features (eg, quizzes and polls) to enhance engagement while reinforcing accurate information.

Strengths and Limitations

This research scrutinized video material sourced from 2 major Chinese short-form video applications, covering a wide spectrum of age groups. Qualified health care experts appraised the videos, judging their quality, reliability, understandability, and actionability through the usage of the GQS, mDISCERN, and PEMAT-A/V rating methodologies. To strengthen the robustness of the research, Spearman's rank correlation coefficient and stepwise regression techniques were implemented. The outcomes present valuable direction for developers of content endeavoring to improve the standard of health-related information sharing.

Nonetheless, several constraints merit attention. First, the confinement to videos in the Chinese language restricts the extent to which the results can be broadly applied. Subsequent studies ought to delve into how well the sample reflects the overall population and the restrictions imposed by choosing a specific language and platform. Second, choosing the top 100 most popular videos could potentially introduce bias caused by algorithms. Third, the appropriateness of using the GQS and mDISCERN instruments for evaluating video content demands further confirmation. Finally, the restricted number of videos from both professional and nonprofessional institutions may not adequately depict the wider range of videos available. Subsequent investigations should strive to enlarge the number of samples within these classifications.

While the selection of the top 100 videos via a comprehensive ranking intended to identify prevalent content, it is crucial to recognize the potential impact of platform-specific algorithms on video visibility and ranking. TikTok's algorithm, known to favor concise and visually stimulating content, may have elevated videos exhibiting these attributes in the top ranks, potentially leading to an underrepresentation of lengthier, more in-depth explanations. Similarly, Bilibili's algorithm, with its emphasis on community engagement and potentially longer-form content, might have highlighted a distinct set of videos. These inherent algorithmic biases could affect the generalizability of our findings to the entirety of uterine fibroid information present on these platforms and may have influenced the representativeness of the analyzed content. Subsequent research using alternative sampling strategies is necessary to achieve a more holistic understanding of the online information landscape concerning uterine fibroids.

Conclusion

This research evaluated the quality of health-related information about uterine fibroids on the TikTok and Bilibili platforms, revealing that the overall quality is generally inadequate, with minimal differences between the 2 platforms. However, TikTok videos tend to show slightly higher quality compared to those on Bilibili, while content created by medical experts stands out with significantly better quality. The growing availability of online educational resources has highlighted increasing disparities in content quality and professionalism across platforms. While Bilibili hosts a higher percentage of videos with an academic focus, users must still apply discernment to ensure the information they encounter is accurate. Future studies could investigate strategies to improve video quality on these platforms and enhance users' ability to critically assess and use the available content. Furthermore, research may explore ways to balance professionalism and accessibility to better meet the diverse needs of various audiences. It is therefore recommended that platforms strengthen oversight and encourage more active involvement from medical professionals in content production. In addition, medical experts should address common misconceptions and inaccuracies in health-related videos, ensuring that the public receives more reliable and scientifically sound health information.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.The raw video and comment data analyzed are publicly available on the Bilibili and Douyin (TikTok) platforms, subject to their terms of service and content availability.

XSL•FO
RenderX

Authors' Contributions

LW conceived and designed the study. LW, TX, and FH were responsible for reviewing and scoring the videos. DZ and YC collected and analyzed the data. LW wrote the original draft. FH reviewed the manuscript and provided critical revisions to the intellectual content.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Uterine Fibroid dataset.
[XLSX File, 39 KB - formative_v9i1e75120_app1.xlsx]

Multimedia Appendix 2
Evaluation of uterine fibroid–related videos on TikTok and Bilibili by quality, engagement, understandability, actionability, and content completeness across sources and content categories
[DOCX File, 27 KB - formative_v9i1e75120_app2.docx]

Multimedia Appendix 3
Uterine Fibroid Completeness Scoring Table.
[DOC File, 37 KB - formative_v9i1e75120_app3.doc]

References

1. Yang Q, Ciebiera M, Bariani MV, et al. Comprehensive review of uterine fibroids: developmental origin, pathogenesis, and treatment. Endocr Rev 2022 Jul 13;43(4):678-719. [doi: 10.1210/endrev/bnab039] [Medline: 34741454]
2. Giuliani E, As‐Sanie S, Marsh EE. Epidemiology and management of uterine fibroids. Intl J Gynecology Obste 2020 Apr;149(1):3-9. [doi: 10.1002/ijgo.13102]
3. Alkhrait S, Malasevskaia I, Madueke-Laveaux OS. Fibroids and fertility. Obstet Gynecol Clin North Am 2023 Dec;50(4):663-675. [doi: 10.1016/j.ogc.2023.08.006] [Medline: 37914486]
4. Pavone D, Clemenza S, Sorbi F, Fambrini M, Petraglia F. Epidemiology and risk factors of uterine fibroids. Best Pract Res Clin Obstet Gynaecol 2018 Jan;46:3-11. [doi: 10.1016/j.bpobgyn.2017.09.004] [Medline: 29054502]
5. De La Cruz MSD, Buchanan EM. Uterine fibroids: diagnosis and treatment. Am Fam Physician 2017 Jan 15;95(2):100-107. [Medline: 28084714]
6. Swire-Thompson B, Lazer D. Public health and online misinformation: challenges and recommendations. Annu Rev Public Health 2020 Apr 2;41:433-451. [doi: 10.1146/annurev-publhealth-040119-094127] [Medline: 31874069]
7. Gray NJ, Klein JD, Noyce PR, Sesselberg TS, Cantrill JA. Health information-seeking behaviour in adolescence: the place of the internet. Soc Sci Med 2005 Apr;60(7):1467-1478. [doi: 10.1016/j.socscimed.2004.08.010]
8. Adam M, Chen SF, Amieva M, et al. The use of short, animated, patient-centered springboard videos to underscore the clinical relevance of preclinical medical student education. Acad Med 2017 Jul;92(7):961-965. [doi: 10.1097/ACM.0000000000001574] [Medline: 28121656]
9. Zheng C. Research on the flow experience and social influences of users of short online videos. A case study of DouYin. Sci Rep 2023 Feb 27;13(1):3312. [doi: 10.1038/s41598-023-30525-y] [Medline: 36849531]
10. Moorhead SA, Hazlett DE, Harrison L, Carroll JK, Irwin A, Hoving C. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. J Med Internet Res 2013 Apr 23;15(4):e85. [doi: 10.2196/jmir.1933] [Medline: 23615206]
11. Schüz B, Jones C. Mis- and disinformation in social media: mitigating risks in digital health communication. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2024 Mar;67(3):300-307. [doi: 10.1007/s00103-024-03836-2] [Medline: 38332143]
12. Harness J, Getzen H. TikTok's sick-role subculture and what to do about it. J Am Acad Child Adolesc Psychiatry 2022 Mar;61(3):351-353. [doi: 10.1016/j.jaac.2021.09.312] [Medline: 34534625]
13. Ming S, Han J, Li M, Liu Y, Xie K, Lei B. TikTok and adolescent vision health: content and information quality assessment of the top short videos related to myopia. Front Public Health 2022;10:1068582. [doi: 10.3389/fpubh.2022.1068582] [Medline: 36684892]
14. Zhao X, Yao X, Sui B, Zhou Y. Current status of short video as a source of information on lung cancer: a cross-sectional content analysis study. Front Oncol 2024;14:1420976. [doi: 10.3389/fonc.2024.1420976] [Medline: 39650058]
15. Zheng S, Tong X, Wan D, Hu C, Hu Q, Ke Q. Quality and reliability of liver cancer-related short Chinese videos on TikTok and Bilibili: cross-sectional content analysis study. J Med Internet Res 2023 Jul 5;25:e47210. [doi: 10.2196/47210] [Medline: 37405825]

XSL•FO
RenderX

16. Zhang R, Zhang Z, Jie H, et al. Analyzing dissemination, quality, and reliability of Chinese brain tumor-related short videos on TikTok and Bilibili: a cross-sectional study. Front Neurol 2024;15:1404038. [doi: 10.3389/fneur.2024.1404038] [Medline: 39494168]
17. Liu H, Peng J, Li L, et al. Assessment of the reliability and quality of breast cancer related videos on TikTok and Bilibili: cross-sectional study in China. Front Public Health 2023;11:1296386. [doi: 10.3389/fpubh.2023.1296386] [Medline: 38317686]
18. Gong X, Dong B, Li L, Shen D, Rong Z. TikTok video as a health education source of information on heart failure in China: a content analysis. Front Public Health 2023;11:1315393. [doi: 10.3389/fpubh.2023.1315393] [Medline: 38146471]
19. Chen Y, Wang Q, Huang X, et al. The quality and reliability of short videos about thyroid nodules on BiliBili and TikTok: cross-sectional study. Digit Health 2024;10:20552076241288831. [doi: 10.1177/20552076241288831] [Medline: 39381823]
20. Sun F, Zheng S, Wu J. Quality of information in gallstone disease videos on tiktok: cross-sectional study. J Med Internet Res 2023 Feb 8;25:e39162. [doi: 10.2196/39162] [Medline: 36753307]
21. Song S, Xue X, Zhao YC, Li J, Zhu Q, Zhao M. Short-video apps as a health information source for chronic obstructive pulmonary disease: information quality assessment of TikTok videos. J Med Internet Res 2021 Dec 20;23(12):e28318. [doi: 10.2196/28318] [Medline: 34931996]
22. Shackleford M, Horvath A, Repetto M, et al. An analysis of oral contraceptive related videos on TikTok. AJOG Glob Rep 2024 Aug;4(3):100364. [doi: 10.1016/j.xagr.2024.100364] [Medline: 39077679]
23. Stoddard RE, Pelletier A, Sundquist EN, et al. Popular contraception videos on TikTok: an assessment of content topics. Contraception 2024 Jan;129:110300. [doi: 10.1016/j.contraception.2023.110300] [Medline: 37802460]
24. Zhang J, Yuan J, Zhang D, et al. Short video platforms as sources of health information about cervical cancer: a content and quality analysis. PLoS ONE 2024;19(3):e0300180. [doi: 10.1371/journal.pone.0300180] [Medline: 38457419]
25. John JN, Gorman S, Scales D, Gorman J. Online misleading information about women's reproductive health: a narrative review. J Gen Intern Med 2025 Apr;40(5):1123-1131. [doi: 10.1007/s11606-024-09118-6] [Medline: 39511120]
26. Mueller SM, Hongler VNS, Jungo P, et al. Fiction, falsehoods, and few facts: cross-sectional study on the content-related quality of atopic eczema-related videos on YouTube. J Med Internet Res 2020 Apr 24;22(4):e15599. [doi: 10.2196/15599] [Medline: 32329744]
27. Kyarunts M, Mansukhani MP, Loukianova LL, Kolla BP. Assessing the quality of publicly available videos on MDMA-assisted psychotherapy for PTSD. Am J Addict 2022 Nov;31(6):502-507. [doi: 10.1111/ajad.13325] [Medline: 36000281]
28. Uzun O. Assessment of reliability and quality of videos on medial epicondylitis shared on YouTube. Cureus 2023 Apr;15(4):e37250. [doi: 10.7759/cureus.37250] [Medline: 37168186]
29. PEMAT tool for audiovisual materials (PEMAT-a/v). AHRQ. 2025. URL: https://www.ahrq.gov/health-literacy/patient-education/pemat-av.html [accessed 2025-06-27]
30. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159-174. [Medline: 843571]
31. Cui N, Lu Y, Cao Y, Chen X, Fu S, Su Q. Quality assessment of TikTok as a source of information about mitral valve regurgitation in China: cross-sectional study. J Med Internet Res 2024 Aug 20;26:e55403 [FREE Full text] [doi: 10.2196/55403] [Medline: 39163110]
32. Li A, Xing Q, Zhang Y, et al. Evaluation of the information quality related to osteoporosis on TikTok. BMC Public Health 2024 Oct 18;24(1):2880. [doi: 10.1186/s12889-024-20375-2] [Medline: 39425064]
33. McCashin D, Murphy CM. Using TikTok for public and youth mental health - A systematic review and content analysis. Clin Child Psychol Psychiatry 2023 Jan;28(1):279-306. [doi: 10.1177/13591045221106608] [Medline: 35689365]
34. Chochol MD, Gandhi K, Elmaghraby R, Croarkin PE. Harnessing youth engagement with mental health TikTok and its potential as a public health tool. J Am Acad Child Adolesc Psychiatry 2023 Jul;62(7):710-712. [doi: 10.1016/j.jaac.2022.11.015] [Medline: 36773699]
35. Sulibhavi A, Isaacson G. TikTok Tonsils. Ear Nose Throat J 2024 Mar;103(3):NP183-NP184. [doi: 10.1177/01455613211038340] [Medline: 34569296]
36. Aktas BK, Demirel D, Celikkaleli F, et al. YouTubeTM as a source of information on prostatitis: a quality and reliability analysis. Int J Impot Res 2024 May;36(3):242-247. [doi: 10.1038/s41443-023-00666-9] [Medline: 36646834]

Abbreviations

ELM: Elaboration Likelihood Model
GQS: Global Quality Score
HSM: Heuristic-Systematic Model
mDISCERN: modified DISCERN
PEMAT: Patient Education Materials Assessment Tool
PEMAT-A: Patient Education Materials Assessment Tool-Actionability
PEMAT-A/V: Patient Education Materials Assessment Tool for Audio/Visual Materials

XSL•FO
RenderX

PEMAT-U: Patient Education Materials Assessment Tool-Understandability

Edited by A Mavragani; submitted 28.03.25; peer-reviewed by SB Guo, S Saponara; revised version received 20.05.25; accepted 08.06.25; published 01.08.25.

Please cite as:
Wang L, Chen Y, Zhao D, Xu T, Hua F
Quality and Dissemination of Uterine Fibroid Health Information on TikTok and Bilibili: Cross-Sectional Study
JMIR Form Res 2025;9:e75120
URL: https://formative.jmir.org/2025/1/e75120
doi:10.2196/75120

© Lan Wang, Yiwen Chen, Duo Zhao, Tao Xu, Fu Hua. Originally published in JMIR Formative Research (https://formative.jmir.org), 1.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Quality and Reliability of Transarterial Chemoembolization Videos on TikTok and Bilibili: Cross-Sectional Content Analysis Study

Yushuo Niu[1*], MD; Guilan Song[1*], BD; Zheyu Niu[1*], PhD; Sijian Xiao[2*], MD; Cuicui Li[1], MD; Na Han[1], BD; Hao Wan[3], MD; Xiaohong Hou[1], MD

[1]Department of Hepatobiliary Surgery, Shandong Provincial Hospital Affiliated to Shandong First Medical University, 9677 Jingshi Road, Lixia District, Jinan, China

[2]Qilu Hospital of Shandong University, Jinan, China

[3]The Second Affiliated Hospital of Shandong First Medical University, Taian, China

[*]these authors contributed equally

Corresponding Author:
Xiaohong Hou, MD
Department of Hepatobiliary Surgery, Shandong Provincial Hospital Affiliated to Shandong First Medical University, 9677 Jingshi Road, Lixia District, Jinan, China

Abstract

Background: Transarterial chemoembolization (TACE) is a widely used treatment for advanced, unresectable hepatocellular carcinoma, often requiring multiple sessions for optimal efficacy. TikTok and Bilibili have gained widespread popularity as easily accessible sources of health information.

Objective: This study aims to assess the quality of the information in Chinese short videos on TACE shared on TikTok and Bilibili.

Methods: In November 2024, the top 100 TACE-related Chinese-language short videos on TikTok and Bilibili (a total of 200 videos) were assessed and reviewed. Initially, basic information about the videos was recorded and analyzed. Subsequently, Global Quality Score and the DISCERN tool were used to evaluate the information quality and reliability of the videos on both platforms. Finally, multifactorial analysis was used to identify potential factors influencing the quality of the videos.

Results: TikTok is more popular than Bilibili, despite its videos being shorter in length ($P<.001$). The quality of short videos on TACE found on both platforms was of low quality, with average Global Quality Score scores of 2.31 (SD 0.81) on TikTok and 2.48 (SD 0.80) on Bilibili, as well as DISCERN scores of 1.86 (SD 0.40) on TikTok and 2.00 (SD 0.44) on Bilibili. The number of saves ($\beta=.184$, $P=.008$; $\beta=.176$, $P=.01$) and days ($\beta=.214$, $P=.002$; $\beta=.168$, $P=.01$) since publication were identified as closely related variables to video quality and reliability. Furthermore, the duration of the video was closely related to its reliability ($\beta=.213$, $P=.002$).

Conclusions: This study indicates that the quality of TACE-related health information in the top 100 short videos on both Bilibili and TikTok platforms is suboptimal. Patients should exercise caution when relying on health-related information from these platforms. Social media companies should establish review teams with basic medical knowledge. It is essential for the platforms to enhance their recommendation algorithms and implement measures for video quality assessment. Health care professionals should be aware of the limitations of these videos and work to improve their quality.

(JMIR Form Res 2025;9:e73855) doi:10.2196/73855

KEYWORDS

transarterial chemoembolization; TACE; hepatocellular carcinoma; health education; short videos; quality analysis

Introduction

Hepatocellular carcinoma (HCC) ranks as the fourth most common cause of cancer-related deaths globally and poses a significant challenge to health care systems worldwide [1,2]. According to the Barcelona Clinic Liver Cancer staging system, transarterial chemoembolization (TACE) is the first-line treatment for intermediate and advanced HCC, including unresectable, multinodular HCC without extrahepatic spread [3]. TACE has been widely accepted and recommended across various clinical settings in Asian countries, demonstrating strong acceptability and favorable treatment outcomes [4]. However, the complexity of the procedure, potential adverse effects, and need for repeated interventions often leave patients with limited understanding and high emotional distress [5-8]. Therefore, enhancing public awareness and patient education about TACE is critical.

XSL•FO
RenderX

With the increasing prevalence of internet technology, the importance and ubiquity of social media use have risen, with an estimated global user base of 4.9 billion [8]. In China, TikTok and Bilibili are the top two video-based social media platforms, boasting hundreds of millions of users. According to the latest data from QuestMobile, TikTok's monthly active users surpassed the 1 billion mark as of September 2024 [9]. This figure highlights the widespread acceptance and influence of short videos, emphasizing the public's strong preference for fast and intuitive methods of information access. Short videos can be easily created and shared on these platforms, enabling users to interact through likes, comments, and bookmarks, which fosters broader participation across social groups, regardless of gender or age. Health care professionals and organizations have also leveraged these platforms to disseminate health-related information, promote healthy habits, and encourage disease prevention. These platforms present an unparalleled opportunity to improve public health literacy, enhance self-efficacy, and promote treatment adherence [10].

A recent systematic review of health information systems underscores accuracy and completeness as foundational data quality elements critical for informed decision-making [11]. While videos related to TACE are available on TikTok and Bilibili, providing accessible information to patients, the quality of these videos remains uncertain. Prior studies suggest that misinformation is common on social media and that users often lack the digital health literacy needed to evaluate content critically [12,13]. Therefore, understanding the sources of information, the characteristics of social media channels, and the content and framing of health information is crucial for navigating the dynamic landscape of TACE-related content on these platforms.

Content analysis achieves an objective and replicable examination of communication materials by encoding the content into predefined categories. This method enables quantitative and qualitative interpretation of media content, making it particularly suitable for evaluating health information on web-based platforms. Moreover, the Global Quality Scores (GQS) and DISCREN systems are widely used to assess the content, quality, and reliability of short videos related to various diseases on social media platforms [12,14-16]. Therefore, this study aims to investigate the top 100 videos related to TACE on TikTok and Bilibili, evaluating their content, quality, and reliability. Additionally, the correlation between video quality and factors such as video source, content, length, likes, comments, shares, and saves was analyzed and discussed. This analysis can help foster the development of social media platforms, offering the general public more comprehensive and appropriate ways to access TACE-related information. Ultimately, our goal is to promote the correct treatment of HCC.

Methods

Ethical Considerations

Shandong Provincial Hospital affiliated to Shandong First Medical University did not conduct ethical approval. The information in this study does not involve clinical data, human specimens, or experimental animals. All data were obtained from publicly available videos on TikTok and Bilibili. We complied with informed consent guidelines and have adhered to Chinese law and regulations regarding the protection of personal information, privacy, and human rights.

Study Design

Search Strategy

Videos were obtained from two platforms, Bilibili [17] and TikTok (Chinese version [18] on November 18, 2024. The search was conducted using the keyword "TACE" (Figure 1). The search was conducted through a mobile phone to ensure consistency and to avoid algorithm adjustments specific to desktop browsers. To minimize bias from personalized recommendation algorithms, we registered and logged into new accounts on each platform. Furthermore, we have disabled personalized recommendations in the platform settings.

The following filters were applied: (1) language restriction: Chinese only; (2) upload date: no restriction; (3) sorting method: "Comprehensive ranking," which prioritizes videos with video completion rate, high engagement (likes, comments, and shares), and recent uploads. The exclusion criteria were as follows: (1) duplicate videos, (2) advertisement content, and (3) irrelevant content.

XSL•FO
RenderX

Figure 1. Search strategy for short videos on TACE. TACE: transarterial chemoembolization.

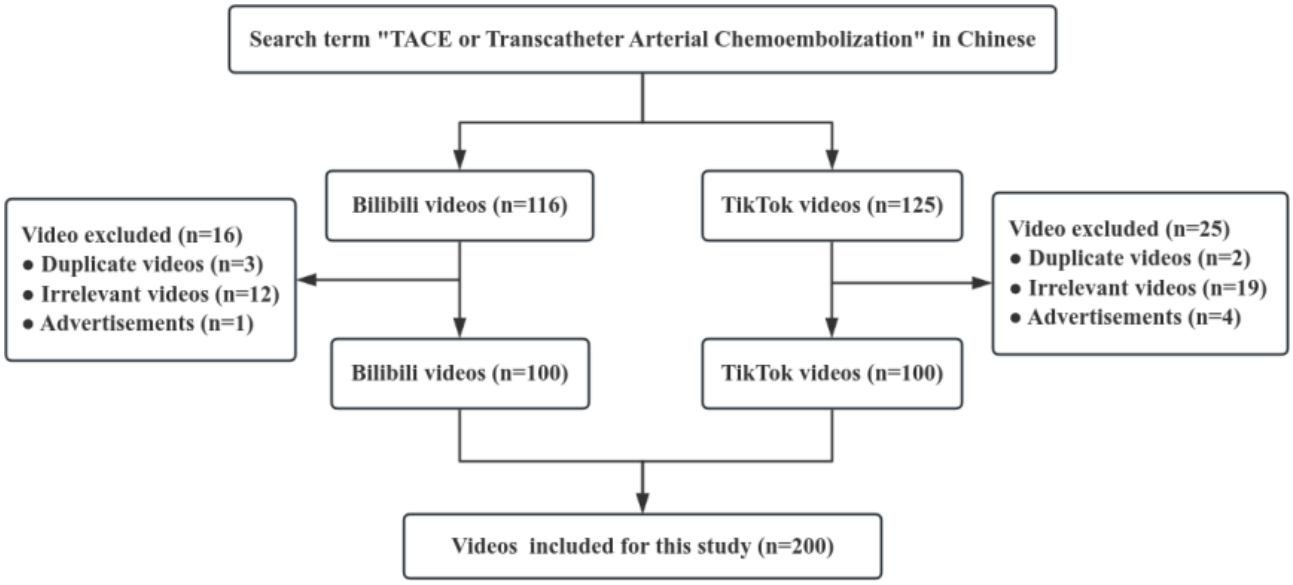

Sampling Procedure

We retrieved the top 100 videos based on overall rankings. This cutoff was chosen for three reasons: (1) it reflects common user experience, and focusing on the mainstream content recommended by the platform algorithm, and thus reflecting the current situation of health information dissemination in a more representative manner; (2) it balances depth and feasibility of manual coding across platforms; and (3) it is consistent with prior social media health content analyses, improving comparability [13,19,20]. There were no duplicate videos between the Bilibili and TikTok datasets.

Data Extraction

Two trained reviewers independently recorded. Data collected included the title, number of likes, number of comments, number of saves, number of shares, video duration, upload date, number of days since posting, uploader, video content, and video source. Disagreements were resolved by discussion or a third reviewer.

Video Quality Assessments

Two widely adopted standardized scales were used. The GQS was used to assess the quality of information in the videos. The GQS consists of five criteria, with scores ranging from 1 to 5. The higher the score, the better the quality [21]. The DISCERN questionnaire was used to evaluate their reliability. In addition, we used the full version of the DISCERN tool [22]. The questionnaire consists of 16 questions, scored on a 5-point scale (totaling 80 points), and is divided into three sections (the reliability of the video, the quality of treatment options, and the overall score). The total score and average value of 16 items are used to distinguish between high-quality and low-quality health information. The specific items can be found in Multimedia Appendix 1. Two independent reviewers with a background in medicine received training and independently conducted the review and evaluation. In the event of any discrepancies or disagreements between the two reviewers, a third expert was included in the discussion to determine the final score. The intraclass correlation coefficient values evaluated by the assessors were 0.901 and 0.887, indicating better reliability.

Classification of Videos

Three groups were classified according to the source of the video, and five groups according to the content. The video sources were categorized as follows: (1) professional individuals, (2) nonprofessional individuals, and (3) professional institutions. The video content was classified into the following categories: (1) treatment description, (2) image analysis, (3) treatment suggestion, (4) precautions after intervention, and (5) experience and feeling. Specific classification criteria are shown in Multimedia Appendix 2.

Statistical Analysis

Data analysis was performed using SPSS (IBM Corp) software. Normally distributed continuous data were expressed as ($\bar{\chi}$±[s]), while skewed data were presented as median (IQR). Categorical data were presented as frequency and percentage. The normality of GQS and DISCERN scores was assessed using the Shapiro-Wilk test. For normally distributed continuous data, comparisons between groups were conducted using the independent samples 2-tailed *t* test or ANOVA. For nonnormally distributed continuous data, comparisons between groups were made using the Mann-Whitney *U* test or Kruskal-Wallis H test. Spearman correlation analysis was performed to assess the relationships between variables. Variables with $P<.05$ were selected for inclusion in the multivariate analysis. Multiple linear regression was applied to identify the factors influencing video quality. GraphPad Prism (version 9.0.2; GraphPad Software, LLC) software was used for plotting the graphs.

Results

Overview of the Video Screening Process

To obtain the top 100 eligible videos from Bilibili and TikTok, we removed 16 noncompliant videos from Bilibili and 25 from TikTok. In the end, a thorough evaluation was carried out on a total of 200 eligible videos, comprising content from both

XSL•FO
RenderX

platforms. Each video was unique in content and uploader across the two platforms (Figure 1).

Landscape of the TACE Videos on TikTok and Bilibili

TikTok and Bilibili are public platforms in China, with a wide variety of uploaded content and types. In this study, we categorized the videos based on the uploader's level of expertise. The results showed that videos from professionals dominated, accounting for 85.5%. The remaining sources were as follows: professional institutions (8.5%) and nonprofessional individuals (6%; Figure 2A). Overall, both TikTok and Bilibili have a large number of videos uploaded by professionals (Figure 2B). Notably, in terms of video content, "Treatment description" was the most prevalent category (36.5%), followed by "Treatment suggestions" (33%), while "Experience and feeling" had the lowest proportion (3.5%; (Figure 2C). A separate comparison of content types on TikTok and Bilibili showed similar results (Figure 2D).

Figure 2. General information on TACE videos from TikTok and Bilibili. (A) The doughnut chart shows the sources of all included videos. (B) The bar chart shows the sources of videos from TikTok and Bilibili. (C) The doughnut chart shows the content of all included videos. (D) The bar chart shows the content of videos from TikTok and Bilibili. TACE: transarterial chemoembolization.

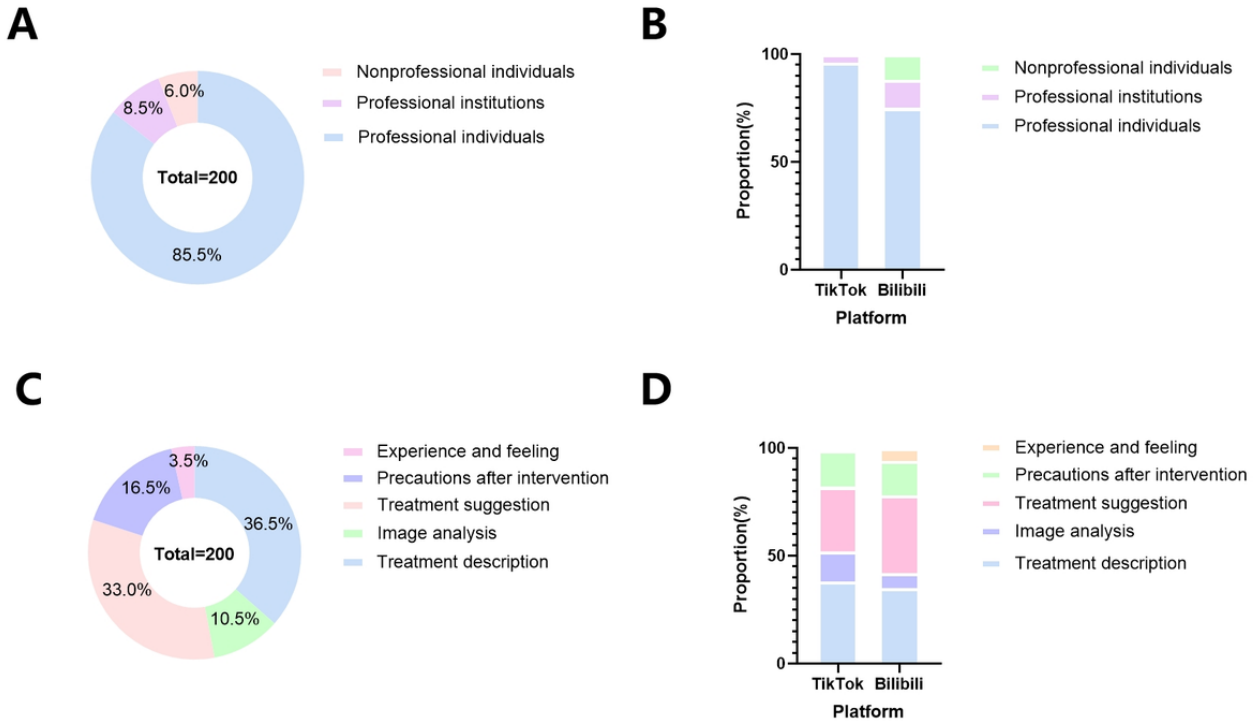

General Information and Index of Web-Based Videos

The number of views, likes, comments, shares, and other metrics on TikTok and Bilibili serve as indicators of the attention these videos receive and are suitable tools for assessing their influence. First, we compared these metrics between TikTok and Bilibili. In contrast, related videos on TikTok received more likes, comments, shares, and saves compared to those on Bilibili (P<.001). However, videos on Bilibili had longer durations and were uploaded earlier (P<.001; Table 1).

Table . Basic index of videos about TACE[a] on Bilibili and TikTok.

Variable	TikTok (n=100), median (IQR)	Bilibili (n=100), median (IQR)	z [b]	P value
Likes	76.00 (37.5-143)	6 (2-16)	−8.874	<.001
Comments	6 (2.25-12.75)	1 (0-3)	−6.677	<.001
Saves	17 (4.25-56.50)	4 (1-19)	−4.594	<.001
Shares	10 (2-42)	3 (1-10.75)	−3.951	<.001
Duration	1.02 (0.19-1.58)	1.54 (1.09-3.26)	−5.159	<.001
Days since published	145.50 (61-325)	582.50 (312.25-890.25)	−7.208	<.001

[a]TACE: transarterial chemoembolization.

[b]Wilcoxon rank sum test was used.

Next, we further compared the differences in the number of likes, comments, shares, saves, video duration, and upload time of web-based videos based on their source and content type (Tables 2 and 3). In Bilibili, although nonprofessional

XSL•FO
RenderX

individuals uploaded a smaller number of videos, their videos had a high median number of saves and longest durations, suggesting a potential preference among viewers for longer, personal, or experience-based content. Notably, videos from professional institutions had significantly higher median save counts and share counts, suggesting these videos may be more valued as reference material. On TikTok, the number of comments from professionals is relatively high. To our surprise, "Experience and feeling" videos showed the highest quartile value in terms of likes, though differences did not reach statistical significance. In addition, when comparing across content types, videos classified as "Treatment description" received the most saves and shares in Bilibili and TikTok. This indicates that people have a high level of attention to information related to treatment.

Table . Characteristics of the videos across sources and content in Bilibili.

Variable	Likes	Comments	Saves	Shares	Duration	Days since published
Video sources (n=100), median (IQR)						
Professional individuals (n=75)	5 (2-15)	1 (0-3)	3 (1-9)	1 (0-9)	1.46 (1.06-2.11)	505 (311-967)
Professional institutions (n=13)	8 (1.5-36)	0 (0-2)	21 (2.5-86)	13 (3-41.5)	4.38 (1.34-20.09)	584 (281.5-721)
Nonprofessional individuals (n=12)	12.5 (4-68.75)	0 (0-16)	21.5 (2-38.25)	5.5 (1.5-32.25)	12.43 (1.90-19.40)	735.5 (412.75-1039.50)
H[a]	2.135	1.303	11.863	10.827	11.774	0.365
P value	.34	.003	.003	.004	.003	.83
Video content (n=100), median (IQR)						
Treatment description (n=35)	11.5 (3.50-26.75)	1 (0-3)	11.5 (2.75-42.5)	9 (1.75-35.5)	2.19 (1.28-10.22)	600 (373.75-867)
Image analysis (n=7)	3 (1-6)	1 (0-2)	1 (1-2)	0 (0-1)	1.55 (0.22-3.59)	434 (141-741)
Treatment suggestion (n=36)	5 (2-13.25)	1 (0-3)	3.5 (1-13.5)	2 (1-7)	1.54 (1.14-2.06)	450.50 (130-974.5)
Precautions after intervention (n=16)	4 (2-17.25)	0 (0-3)	1 (0-6.5)	1 (0-8.75)	1.25 (0.67-4.10)	585.5 (431-813.75)
Experience and feeling (n=6)	14 (3.5-474.75)	1 (0.75-95)	2 (1-168.5)	1 (0.75-68.5)	1.29 (0.85-23.11)	584 (104.5-1245.25)
H[a]	6.908	3.368	13.933	16.210	4.164	3.611
P value	.14	.49	.008	.003	.38	.46

[a]Kruskal-Wallis H test.

XSL•FO
RenderX

Table . Characteristics of the videos across sources and content in TikTok.

Variable	Likes	Comments	Saves	Shares	Duration	Days since published
Video sources (n=100), median (IQR)						
Professional individuals (n=96)	80.5 (39.25-150.50)	6 (3-13)	18 (5-59.25)	10.5 (2-46.5)	1.02 (0.19-1.56)	145.5 (61-317.75)
Professional institutions (n=4)	25.5 (13.25-71.50)	1.5 (0.25-2.75)	9 (1.25-16.75)	3 (1-16.25)	1.54 (0.44-2.30)	305.5 (55.75-1216.75)
Nonprofessional individuals (n=0)	—[a]	—	—	—	—	—
z [b]	−1.785	−2.257	−1.497	−0.669	−1.355	−1.566
P value	.07	.02	.14	.52	.19	.12
Video content (n=100), median (IQR)						
Treatment description (n=38)	98 (40.50-156.25)	6 (3-13.25)	28.25 (7.5-60.75)	22 (6-87.5)	0.83 (0.12-1.50)	279.5 (122-630)
Image analysis (n=14)	45 (17.75-68.5)	4 (1.75-7)	3 (2-7)	1.5 (1-3.5)	0.19 (0.14-0.33)	115.5 (83.5-154.75)
Treatment suggestion (n=30)	65 (36-124.25)	6.5 (2.75-13)	14.5 (6.75-53.75)	8 (3-34.5)	1.41 (0.39-3.24)	69 (26-183.25)
Precautions after intervention (n=17)	100 (30.5-535)	7 (2-42)	25 (3.5-218)	15 (1.5-131)	1.20 (0.24-2.05)	111 (67-287)
Experience and feeling (n=1)	—	—	—	—	—	—
H[c]	9.668	5.761	15.106	16.830	14.347	18.150
P value	.05	.22	.004	.002	.006	.001

[a]Views were not available on TikTok.

[b]Mann-Whitney *U* test.

[c]Kruskal-Wallis H test.

Quality Analysis of Web-Based Videos

To objectively assess the quality of web-based videos, we conducted a thorough evaluation using the internationally recognized GQS and the DISCREN scoring system. For GQS scores, the majority of videos received a score of 2, making up 47% of the total, followed by scores of 3 (34.5%), 1 (11%), 4 (6.5%), and 5 (1%). Regarding DISCREN scores, the range of 17 - 32 represented the largest proportion at 53%, while scores of 33 - 48 accounted for 44.5%, and 49 - 64 for 2.5%. The quality scores of videos from both platforms were distributed primarily in the low to medium range, as shown in Figure 3 These results suggest that the quality of TACE-related videos is unsatisfactory and needs further improvement.

Figure 3. Quality analysis of web-based videos. (A) The bar chart shows the proportion of video editing software types within each GQS category (Bilibili and TikTok). (B) The bar chart shows the proportion of different software in the DISCERN scores (Bilibili and TikTok). GQS: Global Quality Score.

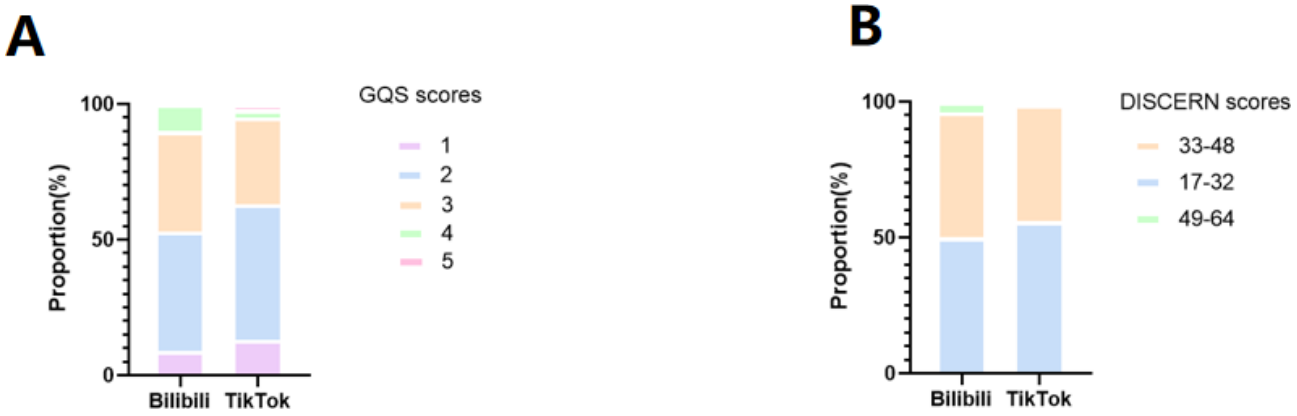

XSL•FO
RenderX

Quality Comparison Across Platforms and Formats

Considering the unique characteristics of Bilibili and TikTok, we performed a comparative analysis of video quality between the two platforms. Regarding TikTok videos, the average GQS score was 2.31 (SD 0.81), and the average DISCERN score was 1.86 (SD 0.40), indicating that TikTok videos generally have moderate quality and low reliability. For Bilibili videos, the average GQS score was 2.48 (SD 0.80), and the average DISCERN score was 2.00 (SD 0.44), suggesting that Bilibili videos also have moderate quality and reliability. Furthermore, there was a significant difference in DISCERN scores between Bilibili and TikTok, with Bilibili videos being more reliable (*P*=.02). However, no significant difference was found in the GQS scores between the two platforms (*P*=.14).

To investigate whether the type of uploader affects video quality, we further compared the GQS and DISCERN scores of videos uploaded by different types of creators. The results showed no statistically significant differences in GQS scores or DISCERN ratings between different uploaders (*P*=.06; *P*=.48). However, there were differences in GQS scores and DISCERN scores between different content (*P*=.04; *P*=.04). Nevertheless, pairwise comparisons revealed no differences between the two groups (Table 4).

Table . Analysis and comparison of the quality of different sources and contents.

Variable	n	GQS[a], mean (SD)	DISCREN, mean (SD)	*F* test (*df*)[b]	*F* test (*df*)[c]	*P* value[b]	*P* value[c]
Video sources (n=200)				2.924 (2,197)	0.730 (2,197)	.06	.48
Professional individuals	171	2.34 (0.78)	2.00 (0.48)				
Professional institutions	17	2.76 (0.75)	2.14 (0.29)				
Nonprofessional individuals	12	2.67 (1.07)	2.03 (0.38)				
Video content (n=200)				2.504 (4,195)	2.494 (4,195)	.04	.04
Treatment description	73	2.49 (0.82)	2.00 (0.44)				
Image analysis	21	2.14 (0.79)	1.74 (0.43)				
Treatment suggestion	66	2.36 (0.74)	1.94 (0.40)				
Precautions after intervention	33	2.55 (0.87)	1.96 (0.39)				
Experience and feeling	7	1.71 (0.76)	2.67 (1.07)				

[a]GQS: Global Quality Score.

[b]ANOVA test for the differences in GQS among different projects.

[c]ANOVA test for the differences in DISCREN among different projects.

Correlation Analysis

Spearman correlation analysis was conducted to explore the relationships between various video variables. Positive correlations were found among the following variables: likes and comments (*r*=0.864; *P*<.001), likes and shares (*r*=0.631; *P*<.001), likes and saves (*r*=0.921; *P*<.001), saves and comments (*r*=0.734; *P*<.001), comments and shares (*r*=0.594; *P*<.001), and saves and shares (*r*=0.750; *P*<.001). Additionally, saves, shares, duration, and days since publication were positively correlated with both GQS and DISCERN. Furthermore, a strong correlation between GQS and DISCERN was observed, confirming their significant consistency and reinforcing the robustness of our scoring results (Table 5).

XSL•FO
RenderX

Table . Spearman correlation analysis between the video variables and GQS[a] scores and DISCERN scores.

Variable	Likes	Comments	Saves	Shares	Duration	Days since published	GQS	DISCERN scores
Likes	1	0.864[b]	0.921[b]	0.631[b]	−0.015	−0.024	0.148	0.137
Comments	0.864[b]	1	0.734[b]	0.594[b]	0.032	0.074	0.130	0.120
Saves	0.921[b]	0.734[b]	1	0.750[b]	0.010	0.030	0.189[b]	0.222[c]
Shares	0.631[b]	0.594[b]	0.750[b]	1	−0.018	0.076	0.194[b]	0.209[b]
Duration	−0.015	0.032	0.010	−0.018	1	0.141	0.164[c]	0.267[b]
Days since published	−0.024	0.074	0.030	0.076	0.141	1	0.216[b]	0.238[b]
GQS scores	0.148	0.130	0.189[b]	0.194[b]	0.164[c]	0.216[b]	1	0.802[b]
DISCERN scores	0.137	0.120	0.222[c]	0.209[b]	0.267[b]	0.238[b]	0.802[b]	1

[a]GQS: Global Quality Score.
[b]Significant at $P<.001$
[c]Significant at $P<.05$

An Investigation Into Determinants of Video Quality

By incorporating meaningful factors from univariate and correlation analyses into multivariate analysis, the results showed that the number of saves and the number of days since upload were closely related to GQS. Specifically, videos with a higher number of saves and longer time since upload tended to have higher GQS quality. Additionally, the number of saves, video duration, and number of days since upload were closely related to DISCERN. Specifically, videos with more saves, longer duration, and earlier upload times demonstrated better reliability (Table 6).

Table . Multiple linear regression analysis of GQS[a] scores and DISCERN scores.

Variable	B	SE	β	*t* test (df)	*P* value
GQS					
Constants	2.149	0.084	—[b]	26.304 (197)	<.001
Days since published	<.001	<0.001	0.214	3.127 (197)	.002
Saves	.001	<0.001	0.184	2.685 (197)	.008
DISCERN					
Constants	1.804	0.043	—	41.487 (196)	<.001
Duration	.005	0.001	0.213	3.133 (196)	.002
Days since published	<.001	<0.001	0.168	2.50 (196)	.01
Saves	<.001	<0.001	0.176	2.595 (196)	.01

[a]GQS: Global Quality Score.
[b]Not applicable.

Discussion

Overview

TACE treatment usually requires multiple surgeries for patients. Therefore, the patient's knowledge, attitude, and treatment compliance are crucial for the success of the treatment. In this regard, educational videos on social media provide a convenient means of accessing relevant information, helping to overcome the traditional barriers to obtaining medical support and resources. Several studies, including those on attention deficit [23], laryngeal cancer [24], thyroid cancer [25], and gallstones [20], provide evidence that short videos facilitate easy access to and positive dissemination of disease-related information. However, the overall quality of these videos is unsatisfactory, raising concerns about misinformation on social media. Inaccurate or misleading information spreads rapidly on these platforms, potentially having a negative impact on health-related attitudes and behaviors. To evaluate the quality and identify the factors influencing TACE-related videos, we performed a comprehensive content analysis of these videos.

Principal Findings

This study is the first to comprehensively evaluate TACE-related videos across major video platforms. We reviewed the top 100 short videos on TACE from two popular Chinese platforms,

XSL•FO
RenderX

TikTok and Bilibili. In this cross-sectional study, the general characteristics of these videos were analyzed, the DISCERN and GQS tools were used to assess their quality and reliability, and the factors influencing video quality were investigated. Our findings reveal that the overall quality of short videos on both platforms was lower than anticipated. Furthermore, we identified that factors such as the number of saves and the number of days since upload were associated with video quality. Additionally, the number of saves, video duration, and days since upload were linked to video reliability.

Video Characteristics

We found that TikTok is more popular than Bilibili. To be precise, the number of likes, comments, shares, and favorites is all high. However, videos on Bilibili tend to have longer durations and earlier upload times. Research indicated that shorter videos were generally more engaging and spread more quickly, which helped explain why TikTok videos received more likes despite having shorter durations [26]. Additionally, we analyzed the types of video creators and found that most TACE-related videos on both platforms were uploaded by professionals. On the other hand, we analyzed the different types of TACE-related content on both platforms. Treatment descriptions were the most prevalent. This underscores the strong demand for treatment-related knowledge among patients with HCC. We observed "experience and feeling" elicited higher levels of interaction, including shares and saves, compared to more technical formats like "image analysis." This implies that emotionally engaging or explanatory content may be more effective in attracting user attention and encouraging dissemination. However, videos describing treatment experiences and personal feelings were the least prevalent. This highlights a gap in TACE-related content on TikTok and Bilibili. Therefore, we encourage health care professionals to expand and diversify video content to include aspects such as prevention, treatment, and precautions, ensuring that the public receives comprehensive and reliable medical information.

Quality of the Short Videos on TACE

The quality and reliability of TACE-related short videos on both platforms were found to be low, with overall performance primarily falling within the moderate to poor range. Research has shown that inaccurate and misleading videos on social media often receive more views from users compared to accurate content, which is similar to our research findings [27,28]. While watching the videos, we noticed that in order to draw attention, some nonprofessionals pretended to be professionals and promoted unproven and unreliable treatment methods. What surprised us was that these videos received a high number of likes and saves. Due to the complexity of the surgery and the psychological burden associated with liver cancer treatment, inaccurate or misleading video content regarding TACE may have serious consequences [29,30]. The patient may thus develop unrealistic expectations regarding such matters, underestimate potential complications, or improperly delay seeking clinical advice. Furthermore, incorrect information may erode people's trust in health care professionals, especially when professional sources are perceived as less appealing than nonprofessional ones [30]. This highlights the urgent need for a balanced health communication strategy that combines accuracy, clarity, and engagement.

Improving health content quality requires a multilevel strategy. First, the platform can modify its algorithm to enhance the visibility of content provided by verified medical professionals and certification organizations, especially in searches related to disease terms. Second, social media companies can establish professional health content review teams, and these teams should possess basic medical knowledge and the ability to judge health literacy. Third, the platforms lack scientific oversight from experts in relevant fields and are unable to conduct peer reviews of the videos being published and shared. It is regrettable that implementing peer reviews may not be feasible. They can assist these platforms in developing effective tagging systems to flag potentially inaccurate or harmful health content. Finally, it is recommended that health care professionals provide evidence-based information when posting videos, including links or citations to credible sources.

We further analyzed the factors related to the quality and reliability of the video content. Interestingly, we found that video duration was closely related to content reliability. This result is likely related to platform limitations. Short video platforms have time constraints, making it difficult to provide the necessary background or depth to convey a complete picture or concept within such a short period. The lack of sufficient information somewhat reduces the quality of the content. Moreover, viewers are often unable to distinguish between high-quality and low-quality videos [13,20]. The audience for these short videos is not composed of professional health care workers, and they often find specialized knowledge unengaging, which leads to content being simpler and more superficial. Additionally, although there were no significant differences in video quality between the two platforms, there were notable differences in reliability, with Bilibili's content being rated as more reliable. This result may also be related to video duration. Bilibili allows for more flexibility in video length, enabling clearer presentation of treatment-related information, while TikTok typically has shorter video durations. Therefore, given the complexity and specialization of medical content, the length of medical-related videos should be extended accordingly to ensure that sufficient information is provided. Furthermore, health care professionals should seek diverse forms of video content to increase its richness and engagement. Interestingly, we found that the longer the video has been published, the better the content quality and reliability. This serves as a warning: the quality of content tended to decrease. Given the widespread focus on entertainment and viral trends, the importance of video content should not be overlooked.

Strengths and Limitations

This study has several strengths. First, it is the first study to analyze the quality of short videos related to TACE on social media platforms, offering guidance for the public in accessing health information and providing valuable insights for content creators and platforms. Second, our research focused on China's two largest short video platforms, overcoming the limitations of studying a single platform, which enhances the reliability of our results. Finally, we used widely recognized tools, the GQS

XSL•FO
RenderX

and DISCERN, to assess the quality and reliability of the videos, ensuring a robust evaluation of our findings.

However, this study has some limitations. First, the videos included in our analysis were limited to those from China, and global platforms like YouTube, which dominate the web-based video space, were excluded due to access restrictions. Second, we focused on the top 100 videos from each platform. This algorithm does not equate to relevance or quality. It may systematically overlook content with lower visibility but potentially higher accuracy. Future research should consider broader sampling strategies to capture less visible but potentially high-quality content. Third, due to the limited sample size within certain categories, we did not perform subgroup or interaction analyses, which might have missed some subtle differences. We identified it as a priority for future, larger-scale analyses.

Conclusions

The overall quality and reliability of the top 100 videos related to TACE videos on TikTok and Bilibili were relatively low. Higher numbers of saves and longer time since upload were positively associated with better video quality and reliability. These findings highlight the need for caution when using social media as a health information source. The quality of the video should be further improved. Platforms should adjust their algorithms to prioritize content from verified medical professionals. Establishing review teams with basic medical literacy can help flag misleading information. Health care professionals are also encouraged to share evidence-based information and cite credible sources when publishing videos.

Acknowledgments

We thank all uploaders of relevant videos. The authors declare that no financial support was received for the research, authorship, and publication of this article.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

YN designed the study and drafted the manuscript. GS, CL, and NH collected all relevant data and assisted in results interpretation. ZN and HW carried out data analysis. ZN drew the picture. XH conceived the study. SX participated in the design and coordination. All authors contributed to the article and approved the submitted version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Full list of questions.
[DOCX File, 16 KB - formative_v9i1e73855_app1.docx]

Multimedia Appendix 2
Classification of videos.
[DOCX File, 17 KB - formative_v9i1e73855_app2.docx]

References

1. Ganesan P, Kulik LM. Hepatocellular carcinoma: new developments. Clin Liver Dis 2023 Feb;27(1):85-102. [doi: 10.1016/j.cld.2022.08.004] [Medline: 36400469]
2. Vogel A, Meyer T, Sapisochin G, Salem R, Saborowski A. Hepatocellular carcinoma. Lancet 2022;400:1345-1362. [doi: 10.1016/S0140-6736(22)01200-4]
3. Chang Y, Jeong SW, Young Jang J, Jae Kim Y. Recent updates of transarterial chemoembolilzation in hepatocellular carcinoma. IJMS 2020;21(21):8165. [doi: 10.3390/ijms21218165]
4. Lee M, Shin HP. Efficacy of transarterial chemoembolization (TACE) for early-stage hepatocellular carcinoma. Med Bogota Colomb 2023;59(12):2174. [doi: 10.3390/medicina59122174]
5. Zhou N, Pan L, Xu AM, Gong SJ, Xing DJ, Chen JY. Study on the influencing factors of fear of cancer recurrence in patients with primary liver cancer after TACE surgery and the therapeutic effect of cognitive behavioral therapy. Progress Mod Biomed 2023;23:3037-3042. [doi: 10.13241/j.cnki.pmb.2023.16.007]
6. Li N, Wang G, Duan G, et al. Clinical observation of transcatheter arterial chemoembolization in super-aged patients with advanced gastric cancer. Support Care Cancer 2022 Feb;30(2):1441-1450. [doi: 10.1007/s00520-021-06565-3] [Medline: 34529138]

XSL•FO
RenderX

7. Tümen D, Heumann P, Gülow K, et al. Pathogenesis and current treatment strategies of hepatocellular carcinoma. Biomedicines 2022 Dec 9;10(12):10. [doi: 10.3390/biomedicines10123202] [Medline: 36551958]
8. Koh GK, Ow Yong JQY, Lee A, et al. Social media use and its impact on adults' mental health and well-being: a scoping review. Worldviews Evid Based Nurs 2024 Aug;21(4):345-394. [doi: 10.1111/wvn.12727] [Medline: 38736207]
9. Mobile internet autumn report in China. QuestMobile. 2024. URL: https://www.questmobile.com.cn/research/report/1853641593686429698 [accessed 2025-08-28]
10. Suarez-Lledo V, Alvarez-Galvez J. Prevalence of health misinformation on social media: systematic review. J Med Internet Res 2021 Jan 20;23(1):e17187. [doi: 10.2196/17187] [Medline: 33470931]
11. Ghalavand H, Shirshahi S, Rahimi A, Zarrinabadi Z, Amani F. Common data quality elements for health information systems: a systematic review. BMC Med Inform Decis Mak 2024;24(1):243. [doi: 10.1186/s12911-024-02644-7]
12. Ming S, Han J, Li M, Liu Y, Xie K, Lei B. TikTok and adolescent vision health: content and information quality assessment of the top short videos related to myopia. Front Public Health 2022;10:1068582. [doi: 10.3389/fpubh.2022.1068582] [Medline: 36684892]
13. Zheng S, Tong X, Wan D, Hu C, Hu Q, Ke Q. Quality and reliability of liver cancer-related short chinese videos on TikTok and Bilibili: cross-sectional content analysis study. J Med Internet Res 2023 Jul 5;25:e47210. [doi: 10.2196/47210] [Medline: 37405825]
14. Lai Y, Liao F, He Z, et al. The status quo of short videos as a health information source of Helicobacter pylori: a cross-sectional study. Front Public Health 2023;11:1344212. [doi: 10.3389/fpubh.2023.1344212]
15. Gong X, Dong B, Li L, Shen D, Rong Z. TikTok video as a health education source of information on heart failure in China: a content analysis. Front Public Health 2023;11:1315393. [doi: 10.3389/fpubh.2023.1315393] [Medline: 38146471]
16. Lukić S, Petrović J. The quality of information provided by the most popular dementia videos on TikTok. Front Public Health 2023;11:1266415. [doi: 10.3389/fpubh.2023.1266415] [Medline: 38089039]
17. Bilibili. URL: https://www.bilibili.com/ [accessed 2025-09-03]
18. Douyin. URL: https://www.douyin.com/?recommend=1 [accessed 2025-09-03]
19. Guan JL, Xia SH, Zhao K, et al. Videos in short-video sharing platforms as sources of information on colorectal polyps: cross-sectional content analysis study. J Med Internet Res 2024 Oct 29;26:e51655. [doi: 10.2196/51655] [Medline: 39470708]
20. Sun F, Zheng S, Wu J. Quality of information in gallstone disease videos on TikTok: cross-sectional study. J Med Internet Res 2023 Feb 8;25:e39162. [doi: 10.2196/39162] [Medline: 36753307]
21. Bernard A, Langille M, Hughes S, Rose C, Leddin D, Veldhuyzen van Zanten S. A systematic review of patient inflammatory bowel disease information resources on the World Wide Web. Am J Gastroenterol 2007 Sep;102(9):2070-2077. [doi: 10.1111/j.1572-0241.2007.01325.x] [Medline: 17511753]
22. Charnock D, Shepperd S, Needham G, Gann R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. J Epidemiol Community Health 1999 Feb;53(2):105-111. [doi: 10.1136/jech.53.2.105] [Medline: 10396471]
23. Yeung A, Ng E, Abi-Jaoude E. TikTok and attention-deficit/hyperactivity disorder: a cross-sectional study of social media content quality. Can J Psychiatry 2022 Dec;67(12):899-906. [doi: 10.1177/07067437221082854] [Medline: 35196157]
24. Liu Z, Chen Y, Lin Y, et al. YouTube/ Bilibili/ TikTok videos as sources of medical information on laryngeal carcinoma: cross-sectional content analysis study. BMC Public Health 2024;24(1):1594. [doi: 10.1186/s12889-024-19077-6]
25. Yang S, Zhan J, Xu X. Is TikTok a high-quality source of information on thyroid cancer? Endocrine 2023;81(2):270-276. [doi: 10.1007/s12020-023-03332-8]
26. Lin IT, Shen YM, Shih MJ, Ho CC. Short video addiction on the interaction of creative self-efficacy and career interest to innovative design profession students. Healthcare (Basel) 2023 Feb 15;11(4):11. [doi: 10.3390/healthcare11040579] [Medline: 36833113]
27. Lai Y, He Z, Liu Y, et al. The quality and reliability of TikTok videos on non-alcoholic fatty liver disease: a propensity score matching analysis. Front Public Health 2023;11:1231240. [doi: 10.3389/fpubh.2023.1231240]
28. Liang J, Wang L, Song S, et al. Quality and audience engagement of takotsubo syndrome-related videos on tiktok: content analysis. J Med Internet Res 2022 Sep 26;24(9):e39360. [doi: 10.2196/39360] [Medline: 36155486]
29. Wang Y, McKee M, Torbica A, Stuckler D. Systematic literature review on the spread of health-related misinformation on social media. Soc Sci Med 2019 Nov;240:112552. [doi: 10.1016/j.socscimed.2019.112552]
30. Loeb S, Langford AT, Bragg MA, Sherman R, Chan JM. Cancer misinformation on social media. CA Cancer J Clin 2024;74(5):453-464. [doi: 10.3322/caac.21857] [Medline: 38896503]

Abbreviations

GQS: Global Quality Score
HCC: hepatocellular carcinoma
TACE: transarterial chemoembolization

XSL•FO
RenderX

Edited by J Sarvestan; submitted 12.03.25; peer-reviewed by Q Wang, S Lukić, Y Li; revised version received 02.08.25; accepted 04.08.25; published 17.09.25.

Please cite as:
Niu Y, Song G, Niu Z, Xiao S, Li C, Han N, Wan H, Hou X
Quality and Reliability of Transarterial Chemoembolization Videos on TikTok and Bilibili: Cross-Sectional Content Analysis Study
JMIR Form Res 2025;9:e73855
URL: https://formative.jmir.org/2025/1/e73855
doi:10.2196/73855

© Yushuo Niu, Guilan Song, Zheyu Niu, Sijian Xiao, Cuicui Li, Na Han, Hao Wan, Xiaohong Hou. Originally published in JMIR Formative Research (https://formative.jmir.org), 17.9.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Investigating the Association of Subjective Numeracy, Interpersonal Communication, and Perceived Discrimination With Watching Health-Related Videos on Social Media Platforms: Cross-Sectional Analysis

Katerina Andreadis[1], MS; Nancy Buderer[2], MS; Aisha Langford[3], PhD

[1]Department of Population Health, NYU Grossman School of Medicine, 227 E 30th Street, 6th floor, New York, NY, United States

[2]Nancy Buderer Consulting, LLC, Oak Harbor, OH, United States

[3]Department of Family Medicine and Public Health Sciences, Wayne State University, Detroit, United States

Corresponding Author:
Katerina Andreadis, MS
Department of Population Health, NYU Grossman School of Medicine, 227 E 30th Street, 6th floor, New York, NY, United States

Abstract

Background: Over the past two decades, use of social media has grown among US adults. Common social media platforms include Facebook, YouTube, Instagram, X, LinkedIn, and TikTok. People proactively use social media for a variety of purposes including searching for health information, peer-to-peer social support, and health-related information sharing. As these platforms often serve as sources of health information, understanding how, if at all, people use them may inform future behavioral interventions delivered via social media. Additionally, a better understanding of social engagement may have implications for public health messaging and patient-centered communication.

Objective: Using a nationally representative sample of US adults, we explored how factors including subjective numeracy (ie, ease of understanding medical statistics), interpersonal communication with family and friends, and perceived discrimination influence whether people ever watched versus never watched health-related videos on social media platforms.

Methods: We analyzed the National Cancer Institute's Health Information National Trends Survey data, which were collected from March to November 2022 (n=6252). After excluding participants who did not have complete data for all variables of interest, we analyzed responses from 4543 participants. Respondents were asked, "In the past 12 months, how often did you watch a health-related video on a social media site (eg, YouTube)?" Response options included: almost every day, at least once a week, a few times a month, less than once a month, and never. We collapsed answers into ever or never watched. Odds ratios (OR), 95% CIs, and *P* values were calculated. A multivariate logistic regression model was considered using all factors that were univariately significant ($P<.10$). Using backward elimination, factors that were not significant with $P>.05$ were removed one by one until remaining factors were all significant collectively ($P<.05$).

Results: Of 4543 adults analyzed, 61.5% reported watching at least one health-related video in the past 12 months, whereas 38.5% had never watched one. In the multivariable analysis, all age group categories over 50 years were less likely to watch health-related videos compared to those aged 18‑34 years, with respondents aged ≥75 years having the lowest odds of all groups for watching a health-related video (OR 0.16, $P<.001$). Higher odds of watching health-related videos were observed among respondents who were Black (OR 1.59, $P<.01$), Hispanic (OR 1.54, $P=.01$), and from "Other" minority groups (OR 2.07, $P=.01$) compared to White respondents. College graduates (OR 1.71, $P<.01$) and those who found medical statistics easy to understand (OR 1.29, $P=.04$), talked about health with friends or family (OR 1.68, $P<.01$), or experienced racial discrimination in medical care (OR 1.59, $P=.02$) also had higher odds of watching health-related videos on social media.

Conclusions: Findings from this study may help target health communication campaigns on social media designed to improve screening, lifestyle changes, medication adherence, and disease management.

(JMIR Form Res 2025;9:e71971) doi:10.2196/71971

KEYWORDS

social media; health communication; perceived discrimination; information dissemination; health-related videos

XSL•FO
RenderX

Introduction

In 2023, eight in ten US adults engaged with platforms like YouTube, nearly half of US adults used Instagram, and over one-third of US adults used TikTok [1]. Health information is a major component of this engagement, with 85% of adults seeking health information on social media and nearly 40% watching health-related YouTube videos [2,3]. Since 2018, social media has overtaken traditional print news as the primary source of news and continues to expand its influence on public perceptions and knowledge about health [4].

The internet has allowed for anyone to disseminate health information, which has both benefits and drawbacks [5]. On the one hand, patients and community groups can share helpful information without going through gatekeepers. Online health communities enable peer-to-peer support and patient-generated health information sharing, providing emotional support and experiential knowledge that complements traditional health care [6]. On the other hand, democratization and ease of information sharing may increase the quantity of misinformative and inaccurate health information on social media platforms. Systematic reviews have documented a substantial prevalence of health misinformation across social media platforms, with studies finding significant rates of inaccurate health content on X (formerly Twitter), YouTube, and Facebook [7,8].

In recent years, a growing number of health care professionals and health systems have been sharing health-related content on social media [9-11]. A systematic review in 2023 found that researchers and clinicians use social media to facilitate research, for professional development, and to enhance patient-provider communication and offline services (eg, medication adherence), while the public uses it to seek and share health information, exchange social support, and document their health experiences [11]. Professional guidelines on the responsible use of social media advise clinicians to keep personal and professional identities separate and to protect patient privacy and confidentiality online [12]. Although posts about medical tests from physicians are more likely to mention harms and less likely to have an overall promotional tone, health misinformation remains prevalent, with most posts being misleading, not reporting medical harms, or not citing scientific evidence [13]. In a newer analysis of TikTok transcripts on diabetic foot care, 42% of posts contained misleading or false claims, with advice that could delay treatment or worsen outcomes [14]. Finally, algorithmic recommendation systems shape what health content people see: audits of YouTube report that recommendations are conditioned by users' watch histories, and analyses of X have observed that posts from influential accounts with high engagement, including low-credibility content, receive increased visibility [15,16]. These trends highlight the need to identify which populations engage with health-related video content on social media.

Accordingly, this exploratory study focuses on how various sociodemographic and health-related variables influence engagement with health-related videos on social media platforms. By understanding these factors, we aim to provide contemporary data on the use of social media for health-related purposes, particularly for populations who may use social media more often.

Methods

Data Sources

This study used self-reported data from the Health Information National Trends Survey (HINTS 6), collected in 2022. HINTS collects data from a nationally representative sample of the US population, focusing on trends in health communication and health information technology. The HINTS 6 dataset originally included 6252 respondents, who completed the survey online or by paper [17].

Variables of Interest

Main Outcome

Participants were asked, "In the past 12 months, how often did you watch a health-related video on a social media site (for example, YouTube)?" Responses were combined into two groups: *ever watched* (a combination of almost every day, at least once a week, a few times a month, less than once a month) and *never watched*.

Sociodemographic and Health-Related Covariates

We collected sociodemographic and health-related information including age, birth gender, race/ethnicity, education level, self-rated general health, BMI, and history of chronic conditions such as diabetes, high blood pressure, depression, and cancer. Additional covariates included confidence in completing medical forms, confidence in taking care of one's health, confidence in finding health resources online, ease of understanding medical statistics, discussions about health with friends and family, trust in health care systems, and experiences of perceived racial discrimination when obtaining medical care.

Statistical Methods

Odds ratios (OR), 95% CIs, and *P* values were calculated using SAS (Statistical Analysis System) proc surveylogistic with jackknife weighting, including the overall weight and 50 replicate weights, and the Newton-Raphson algorithm [17]. Results are presented as the OR (95% CI) for the odds of yes (ever watched video) and *P* value. When interpreting ORs, confidence intervals that do not include the value 1 were considered significant. Data were analyzed using SAS (version 9.4; SAS Institute Inc) and SAS/STAT (version 14.1; SAS Institute Inc). HINTS 6 contained 6252 responses. Of these, there were 6158 with a response to the primary question of interest. We further eliminated respondents if they were missing data on any of the variables in the study. Therefore, analyses are based on a sample size of 4543.

Ethical Considerations

The HINTS 6 data collection protocol received institutional review board (IRB) approval by the Westat IRB (project number 6632.03.51) [18]. This secondary analysis study used only publicly available, deidentified data from the HINTS 6 survey, and therefore did not undergo IRB review as it fits the secondary research exemption under federal guidelines (45 Code of Federal

XSL•FO
RenderX

Regulations 46.104) [18]. Data were accessed under the HINTS public use policy, with no additional permissions required.

Results

As shown in Table 1, 61.5% of respondents watched a health-related video on a social media website in the past 12 months, while 38.5% never watched a video. In the multivariable analysis (Table 2), respondents aged 50 years and older were less likely to watch health-related videos compared to those aged 18 - 34 years. Specifically, compared to those aged 18 - 24 years, those aged 50 - 64 years had an OR of 0.58 (95% CI 0.38 - 0.88); those aged 65 - 74 had an OR of 0.28 (95% CI 0.17 - 0.45), and those aged 75 and older had an OR of 0.16 (95% CI 0.10 - 0.25).

XSL•FO
RenderX

Table . Univariable associations with watching health-related videos on social media, with data from the Health Information National Trends Survey 6, United States, 2022 (N=4543).[a]

	Watched (n=2671, 61.5%), % of row (SE)[b]	Never watched (n=1872, 38.5%), % of row (SE)	Odds ratio for Watched (95% CI)	*P* value
Age (mean years)	45.8 (0.45)	55.7 (0.83)	0.97 (0.96-0.97)	<.001
Age group				<.001
18 - 34	74.3 (3.5)	25.7 (3.5)	Ref[c]	
35 - 49	70.7 (2.8)	29.3 (2.8)	0.84 (0.50-1.40)	.49
50 - 64	61.3 (2.3)	38.7 (2.3)	0.55 (0.37-0.81)	.003
65 - 74	42.9 (2.2)	57.1 (2.2)	0.26 (0.17-0.41)	<.001
≥75	28.5 (2.6)	71.5 (2.6)	0.14 (0.09-0.22)	<.001
Birth gender				
Female	63.0 (1.4)	37.0 (1.4)	1.15 (0.94-1.40)	.17
Male	59.7 (2.0)	40.3 (2.0)	Ref	
Race/ethnicity				<.001
Non-Hispanic White	56.5 (1.6)	43.5 (1.6)	Ref	
Non-Hispanic Black or African American	68.1 (3.1)	31.9 (3.1)	1.65 (1.24-2.19)	.001
Hispanic	69.0 (2.6)	31.0 (2.6)	1.72 (1.29-2.28)	<.001
Non-Hispanic Asian	77.3 (5.0)	22.7 (5.0)	2.62 (1.43-4.82)	.003
Non-Hispanic other	77.1 (4.0)	22.9 (4.0)	2.60 (1.60-4.21)	<.001
Education				
College graduate	70.2 (1.4)	29.8 (1.4)	1.80 (1.47-2.19)	<.001
Not college graduate	56.8 (1.8)	43.2 (1.8)	Ref	
General health				.21
Excellent or very good	61.2 (2.1)	38.8 (2.1)	1.20 (0.83-1.74)	.32
Good	63.9 (1.9)	36.1 (1.9)	1.35 (0.95-1.90)	.09
Fair or poor	56.8 (3.7)	43.2 (3.7)	Ref	
Diabetes				
Yes	56.5 (3.3)	43.5 (3.3)	0.78 (0.59-1.03)	.08
No	62.5 (1.3)	37.5 (1.3)	Ref	
High blood pressure				
Yes	55.0 (2.1)	45.0 (2.1)	0.65 (0.54-0.78)	<.001
No	65.4 (1.3)	34.6 (1.3)	Ref	
Heart condition				
Yes	46.4 (4.3)	53.6 (4.3)	0.51 (0.36-0.73)	<.001
No	62.7 (1.3)	37.3 (1.3)	Ref	
Lung disease				
Yes	59.5 (3.3)	40.5 (3.3)	0.91 (0.66-1.25)	.55
No	61.8 (1.4)	38.2 (1.4)	Ref	
Depression				
Yes	67.6 (1.9)	32.4 (1.9)	1.47 (1.18-1.82)	<.001
No	58.7 (1.6)	41.3 (1.6)	Ref	
Ever had cancer				

XSL•FO
RenderX

	Watched (n=2671, 61.5%), % of row (SE)[b]	Never watched (n=1872, 38.5%), % of row (SE)	Odds ratio for Watched (95% CI)	*P* value
Yes	50.3 (2.8)	49.7 (2.8)	0.60 (0.47-0.77)	<.001
No	62.8 (1.3)	37.2 (1.3)	Ref	
BMI group				.37
Underweight or normal weight (<25)	60.9 (2.3)	39.1 (2.3)	1.06 (0.81-1.39)	.65
Overweight (25 - 29.9)	64.1 (2.3)	35.9 (2.3)	1.22 (0.92-1.60)	.16
Obese (>30)	59.5 (2.2)	40.5 (2.2)	Ref	
Confident with filling out medical forms				
Very	63.2 (1.7)	36.8 (1.7)	1.20 (0.97-1.49)	.10
Somewhat, a little, not at all	58.9 (1.9)	41.1 (1.9)	Ref	
Confident to take good care of your health				
Completely	57.1 (2.6)	42.9 (2.6)	0.78 (0.63-0.97)	.03
Very, somewhat, a little, not at all	63.1 (1.2)	36.9 (1.2)	Ref	
Confident to find internet health resources				
Completely	74.2 (4.0)	25.8 (4.0)	1.96 (1.24-3.09)	.005
Very, somewhat, a little, not at all	59.5 (1.4)	40.5 (1.4)	Ref	
Understand medical statistics				
Very easy, easy	63.4 (1.4)	36.6 (1.4)	1.38 (1.09-1.75)	.01
Hard, very hard	55.6 (2.6)	44.4 (2.6)	Ref	
Patient-centered communication scale (mean linearized score on a 0 - 100 scale)	76.1 (0.52)	78.5 (0.85)	1.00 (0.99-1.00)	.04
Talk about health with friends				.01
Yes	63.1 (1.3)	36.9 (1.3)	1.49 (1.10-2.01)	
No	53.5 (3.5)	46.5 (3.5)	Ref	
Trust the health care system				
Very	58.6 (2.0)	41.4 (2.0)	0.82 (0.69-0.98)	.03
Somewhat, a little, not at all	63.3 (1.3)	36.7 (1.3)	Ref	
Racially discriminated against when getting medical care				
Yes	77.5 (3.3)	22.5 (3.3)	2.26 (1.50-3.40)	<.001
No	60.3 (1.3)	39.7 (1.3)	Ref	

[a]Cross-sectional, survey-weighted logistic regressions estimating unadjusted odds ratios with 95% CIs for “ever watched” (≥1 time in past 12 mo) versus “never” in response to the question “In the past 12 months, how often did you watch a health-related video on a social media site (for example, YouTube)?” Row percentages are weighted.

[b]SE: weighted standard error.

[c]Ref: reference category of the independent variable.

XSL•FO
RenderX

Table . Multivariable associations with watching health-related videos on social media, with data from the Health Information National Trends Survey 6, United States, 2022.[a]

	Adjusted odds ratio for Watched versus Never (95% CI)	*P* value
Age group		<.001
18 - 34	Ref	
35 - 49	0.82 (0.50-1.35)	.43
50 - 64	0.58 (0.38-0.88)	.01
65 - 74	0.28 (0.17-0.45)	<.001
≥75	0.16 (0.10-0.25)	<.001
Race/ethnicity		.001
Non-Hispanic White	Ref	
Non-Hispanic Black or African American	1.59 (1.19-2.23)	.002
Hispanic	1.54 (1.11-2.11)	.01
Non-Hispanic Asian	1.72 (0.90-3.25)	.10
Non-Hispanic other	2.07 (1.19-3.60)	.01
Education		
College graduate	1.71 (1.36-2.16)	<.001
Not college graduate	Ref	
Confident to take good care of your health		
Completely	0.71 (0.57-0.89)	.004
Very, somewhat, a little, not at all	Ref	
Understand medical statistics		
Very easy, easy	1.29 (1.01-1.65)	.04
Hard, very hard	Ref	
Talk about health with friends		
Yes	1.68 (1.23-2.31)	.002
No	Ref	
Racially discriminated against when getting medical care		
Yes	1.59 (1.08-2.34)	.02
No	Ref	

[a]Survey-weighted logistic regression modeling the odds of "ever watched" (≥1 time in past 12 months mo) versus "never" in response to the question "In the past 12 months, how often did you watch a health-related video on a social media site (for example, YouTube)?" Factors that were univariately significant with $P<.10$ were attempted in the model. Using backward elimination, factors were removed one by one until all of the remaining factors in the model were significant with $P<.05$. Adjusted odds ratios with 95% CIs are shown, reference categories are indicated by "Ref," and overall *P* values are reported for multilevel predictors.

Respondents who identified as Black (OR 1.59, 95% CI 1.19 - 2.23), Hispanic (OR 1.54, 95% CI 1.11 - 2.11), and "Other" races (OR 2.07, 95% CI 1.19 - 3.60) were more likely to have watched health-related videos compared to White respondents. College graduates (OR 1.71, 95% CI 1.36 - 2.16), those who found medical statistics easy to understand (OR 1.29, 95% CI 1.01 - 1.65), those who discussed health with friends and family (OR 1.68, 95% CI 1.23 - 2.31), and those who experienced racial discrimination in health care settings (OR 1.59, 95% CI 1.08 - 2.34) also had higher odds of watching health-related videos. Conversely, respondents who were "completely confident" in their ability to take good care of their health were less likely to watch health-related videos on social media (OR 0.71, 95% CI 0.57 - 0.89).

Discussion

Principal Findings

The purpose of this exploratory study was to examine the associations between various sociodemographic and health-related factors and watching health-related videos on social media platforms. Our results indicate that age, race/ethnicity, education, subjective numeracy, interpersonal communication, and perceived discrimination are significantly associated with watching health-related videos on social media.

XSL•FO
RenderX

Comparison With Prior Work

To date, several studies have used HINTS data to explore different aspects of health information–seeking and communication, including watching health-related videos on social media platforms [19,20]. For example, Garg et al [20] examined patterns in the consumption of health-related videos on social media among urban and rural populations and explored how this consumption may be linked to awareness of specific health topics. They found that 59.6% of US adults watched health videos, with younger adults and college graduates being more likely to do so [20]. However, our study is the first to report associations between watching health-related videos and perceived discrimination, subjective numeracy, and interpersonal communication, thereby highlighting the unique ways these factors may influence health information–seeking behaviors on social media.

Other research has highlighted that younger adults are more likely to use social media for health information–seeking purposes, often engaging with content passively or actively seeking it out through groups and pages [21]. Our findings are consistent with prior findings that demonstrate higher engagement on social media among younger adults. According to the Pew Research Center, there are racial/ethnic differences in social media use, with Hispanic users having the highest use across platforms [1]. Similarly, in our study, the higher engagement among Black and Hispanic individuals may reflect a broader trend of these groups using social media to access health information, possibly due to perceived gaps in traditional health care services, disparities in health insurance coverage, and limited access to high-quality health education.

Our study also sheds light on the novel finding that perceived racial discrimination is significantly associated with increased viewing of health-related videos on social media. This suggests that individuals who feel marginalized by health care systems may turn to social media as an alternative source of health information; this finding warrants further exploration. Additionally, participants who reported talking about health with friends or family had significantly higher odds of watching health-related videos on social media than those who did not, suggesting that interpersonal health discussions may encourage individuals to engage in digital health information–seeking and facilitate the cross-platform circulation of health content within social networks. Finally, our results indicate that individuals with higher subjective numeracy (ie, those who self-report understanding medical statistics with ease) and educational attainment are more likely to watch health-related videos. This may be because these individuals are better equipped to search for and interpret complex health information (ie, they may have higher health literacy).

Implications for Practice

These findings have important implications for public health professionals, clinicians, and policymakers. Understanding the demographics and behaviors of those who seek health information on social media can help in designing targeted interventions and communication strategies. For example, health care providers can guide patients toward credible online resources [3,22]. Additionally, efforts to improve social media literacy among the general population could enhance the quality of health information consumed. For example, Polanco-Levicán et al [23] conducted a systematic review analyzing definitions and competencies related to social media literacy. Their findings emphasized critical skills such as evaluating the reliability of information, understanding the context of social interactions on digital platforms, and managing privacy settings—all crucial for discerning credible health information [24]. Notably, the National Academy of Medicine and the World Health Organization have created guidelines for identifying credible sources on social media, which can be integrated into patient education initiatives and ongoing communication training for clinicians [24].

Strengths, Limitations, and Future Directions

This study used a nationally representative dataset, which included a comprehensive range of health- and technology-related variables. Despite these strengths, several limitations must be acknowledged. First, the data do not specify the type of health videos watched. Second, we do not know how, if at all, watching a video changed behavior or influenced specific medical decision-making. Third, the potential sharing of these videos among friends and family, an important aspect of interpersonal communication, remains unexplored. Finally, because we merged the 5 original frequency categories into a single “ever watched” measure, we could not evaluate whether viewing frequency varied with the factors under study. Future research should further evaluate social media content sharing within one’s interpersonal communication network and the impact of health-related video content on viewer behavior and decision-making processes.

Conclusion

This study highlights the roles of various sociodemographic and health-related factors on watching health-related videos on social media platforms. By understanding social media engagement patterns, health professionals can design targeted health communication interventions on social media while dually enhancing health and social media literacy.

Data Availability

The analyses used the public-use, deidentified Health Information National Trends Survey 6 dataset available at [25]. Analytic scripts are not publicly posted but are available from the corresponding author upon request.

Conflicts of Interest

None declared.

XSL•FO
RenderX

References

1. Americans' social media use. Pew Research Center. 2024 Jan 31. URL: https://www.pewresearch.org/internet/2024/01/31/americans-social-media-use/ [accessed 2025-04-29]
2. Marar SD, Al-Madaney MM, Almousawi FH. Health information on social media: perceptions, attitudes, and practices of patients and their companions. Saudi Med J 2019;40(12):1294-1298. [doi: 10.15537/smj.2019.12.24682]
3. Langford A, Loeb S. Perceived patient-provider communication quality and sociodemographic factors associated with watching health-related videos on YouTube: a cross-sectional analysis. J Med Internet Res 2019 May 17;21(5):e13512. [doi: 10.2196/13512] [Medline: 31102372]
4. Social media and news fact sheet. Pew Research Center. 2024 Jan 31. URL: https://www.pewresearch.org/journalism/fact-sheet/social-media-and-news-fact-sheet/ [accessed 2025-10-29]
5. Hendl T, Shukla A. Can digital health democratize health care? Bioethics 2024 Jul;38(6):491-502. [doi: 10.1111/bioe.13266] [Medline: 38193584]
6. Johansson V, Islind AS, Lindroth T, Angenete E, Gellerstedt M. Online communities as a driver for patient empowerment: systematic review. J Med Internet Res 2021 Feb 9;23(2):e19910. [doi: 10.2196/19910] [Medline: 33560233]
7. Suarez-Lledo V, Alvarez-Galvez J. Prevalence of health misinformation on social media: systematic review. J Med Internet Res 2021 Jan 20;23(1):e17187. [doi: 10.2196/17187] [Medline: 33470931]
8. Wang Y, McKee M, Torbica A, Stuckler D. Systematic literature review on the spread of health-related misinformation on social media. Soc Sci Med 2019 Nov;240:112552. [doi: 10.1016/j.socscimed.2019.112552] [Medline: 31561111]
9. Jeyaraman M, Ramasubramanian S, Kumar S, et al. Multifaceted role of social media in healthcare: opportunities, challenges, and the need for quality control. Cureus 2023 May. [doi: 10.7759/cureus.39111]
10. Anderer S. Patients are turning to TikTok for health information—here's what clinicians need to know. JAMA 2024 Apr 16;331(15):1262. [doi: 10.1001/jama.2024.1280]
11. Chen J, Wang Y. Social media use for health purposes: systematic review. J Med Internet Res 2021 May 12;23(5):e17917. [doi: 10.2196/17917] [Medline: 33978589]
12. Parsi K, Elster N. Why can't we be friends? A case-based analysis of ethical issues with social media in health care. AMA J Ethics 2015 Nov 1;17(11):1009-1018. [doi: 10.1001/journalofethics.2015.17.11.peer1-1511] [Medline: 26595241]
13. Nickel B, Moynihan R, Gram EG, et al. Social media posts about medical tests with potential for overdiagnosis. JAMA Netw Open 2025 Feb 3;8(2):e2461940. [doi: 10.1001/jamanetworkopen.2024.61940] [Medline: 40009378]
14. Andrikopoulou E, Talam N, Kanta A. MedTok or MythTok? Classifying health misinformation on TikTok with AI. 2025 Presented at: EFMI Special Topic Conference 2025; Oct 20-22, 2025; Osnabrück, Germany p. 67-71. [doi: 10.3233/SHTI251497]
15. Ng YMM, Hoffmann Pham K, Luengo-Oroz M. Exploring YouTube's recommendation system in the context of COVID-19 vaccines: computational and comparative analysis of video trajectories. J Med Internet Res 2023 Sep 15;25:e49061. [doi: 10.2196/49061] [Medline: 37713243]
16. Corsi G. Evaluating Twitter's algorithmic amplification of low-credibility content: an observational study. EPJ Data Sci 2024;13(1). [doi: 10.1140/epjds/s13688-024-00456-3]
17. Westat. Health Information National Trends Survey 6 (HINTS 6) methodology report. : National Cancer Institute; 2023 Apr 28 URL: https://hints.cancer.gov/docs/methodologyreports/HINTS_6_MethodologyReport.pdf [accessed 2025-10-29]
18. Institutional review board (IRB) approvals for the Health Information National Trends Survey (HINTS). National Cancer Institute. URL: https://hints.cancer.gov/about-hints/institutional-review-board.aspx [accessed 2025-10-29]
19. Finney Rutten LJ, Blake KD, Greenberg-Worisek AJ, Allen SV, Moser RP, Hesse BW. Online health information seeking among US adults: measuring progress toward a Healthy People 2020 objective. Public Health Rep 2019;134(6):617-625. [doi: 10.1177/0033354919874074] [Medline: 31513756]
20. Garg A, Nyitray AG, Roberts JR, et al. Consumption of health-related videos and human papillomavirus awareness: cross-sectional analyses of a US national survey and YouTube from the urban-rural context. J Med Internet Res 2024 Jan 15;26(1):e49749. [doi: 10.2196/49749] [Medline: 38224476]
21. Lim MSC, Molenaar A, Brennan L, Reid M, McCaffrey T. Young adults' use of different social media platforms for health information: insights from web-based conversations. J Med Internet Res 2022 Jan 18;24(1):e23656. [doi: 10.2196/23656] [Medline: 35040796]
22. Langford AT, Roberts T, Gupta J, Orellana KT, Loeb S. Impact of the internet on patient-physician communication. Eur Urol Focus 2020 May 15;6(3):440-444. [doi: 10.1016/j.euf.2019.09.012] [Medline: 31582312]
23. Polanco-Levicán K, Salvo-Garrido S. Understanding social media literacy: a systematic review of the concept and its competences. Int J Environ Res Public Health 2022 Jul 20;19(14):8807. [doi: 10.3390/ijerph19148807] [Medline: 35886657]
24. Kington RS, Arnesen S, Chou WYS, Curry SJ, Lazer D, Villarruel AM. Identifying credible sources of health information in social media: principles and attributes. NAM Perspect 2021;2021. [doi: 10.31478/202107a] [Medline: 34611600]
25. Download data | HINTS. National Cancer Institute. URL: https://hints.cancer.gov/data/download-data.aspx [accessed 2025-10-31]

XSL•FO
RenderX

Abbreviations

HINTS: Health Information National Trends Survey
IRB: institutional review board
OR: odds ratio
SAS: Statistical Analysis System

Edited by A Mavragani, J Sarvestan; submitted 31.01.25; peer-reviewed by R Ganjoo, S Thawillarp; revised version received 10.10.25; accepted 13.10.25; published 05.11.25.

Please cite as:
Andreadis K, Buderer N, Langford A
Investigating the Association of Subjective Numeracy, Interpersonal Communication, and Perceived Discrimination With Watching Health-Related Videos on Social Media Platforms: Cross-Sectional Analysis
JMIR Form Res 2025;9:e71971
URL: https://formative.jmir.org/2025/1/e71971
doi:10.2196/71971

© Katerina Andreadis, Nancy Buderer, Aisha Langford. Originally published in JMIR Formative Research (https://formative.jmir.org), 5.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Self-Disclosure and Social Support in a Web-Based Opioid Recovery Community: Machine Learning Analysis

Yu Chi[1], PhD; Huai-yu Chen[2], PhD; Khushboo Thaker[3], PhD

[1]School of Information, College of Information, Data and Society, San Jose State University, 1 Washington Sq, San Jose, CA, United States
[2]Department of Communication, College of Communication and Information, University of Kentucky, Lexington, KY, United States
[3]School of Computing and Information, University of Pittsburgh, Pittsburgh, PA, United States

Corresponding Author:
Yu Chi, PhD
School of Information, College of Information, Data and Society, San Jose State University, 1 Washington Sq, San Jose, CA, United States

Abstract

Background: The opioid crisis remains a critical public health challenge, with opioid use disorder (OUD) imposing significant societal and health care burdens. Web-based communities, such as the Reddit community r/OpiatesRecovery, provide an anonymous and accessible platform for individuals in recovery. Despite the increasing use of Reddit for substance use research, limited studies have explored the content and interactions of self-disclosure and social support within these communities.

Objective: This study aims to address the following research questions: (1) What content do users disclose in the community?; (2) What types of social support do users receive?; and (3) How does the content disclosed relate to the type and extent of social support received?

Methods: We analyzed 32,810 posts and 324,224 comments from r/OpiatesRecovery spanning 8 years (2014 - 2022) using a mixed method approach. Posts were coded for recovery stages, self-disclosure, and goals, while comments were categorized into informational and emotional support types. A machine learning–based classifier was used to scale the analysis. Regression analyses were conducted to examine the relationship between post content and received support.

Results: The majority of posts were made by individuals using opioids (7225/32,810, 22.0%) or in initial recovery stages (less than 1 mo of abstinence; 27.7%). However, posts by individuals in stable recovery (abstinence for more than 5 years) accounted for only 1.8%. Informational self-disclosure appeared in 88.3% (n=28,977) of posts, while emotional self-disclosure was present in 75.6% (n=24,816). Posts seeking informational support (19,153/32,810, 58.4%) were far more common than those seeking emotional support (779/32,810, 2.4%). On average, each post received 9.88 (SD 11.36) comments. The most frequent types of support were fact and situational appraisal (mean 5.62, SD 6.82) and personal experience (mean 4.88, SD 5.98), while referral was the least common (mean 0.61, SD 0.50). Regression analyses revealed significant relationships between self-disclosure and received support. Posts containing informational self-disclosure were more likely to receive advice (β=0.17, P<.001), facts (β=0.30, P<.001), and opinions (β=0.11, P<.001). Emotional self-disclosure predicted higher levels of emotional support (β=0.17, P<.001) and personal experiences (β=0.07, P<.001). Posts from individuals in the addiction stage received more advice (β=−0.06, P<.001) but less emotional support (β=−0.05, P<.001) compared with posts from individuals in later recovery stages.

Conclusions: This study highlights the role of self-disclosure in fostering social support within web-based OUD recovery communities. Findings suggest a need for increasing engagement from individuals in stable recovery stages and improving the diversity and quality of social support. By uncovering interaction patterns, this study provides valuable insights for leveraging online support groups as complementary resources to traditional recovery interventions.

(JMIR Form Res 2025;9:e71207) doi:10.2196/71207

KEYWORDS

substance use; opioid; recovery; Reddit; social media; peer support; online health community; natural language processing

Introduction

Background

The opioid crisis remains one of the most pressing public health emergencies globally, with opioid use disorder (OUD) contributing to significant mortality, societal burden, and health care challenges, particularly in the United States [1-3]. Addressing OUD, a chronic and relapsing condition, requires sustained access to effective recovery support systems [4-6]. Peer-based programs like Narcotics Anonymous have been shown to reduce substance use, increase treatment completion, and promote long-term recovery [7]. However, barriers such as stigma, financial difficulties, or logistic issues prevent many

XSL•FO
RenderX

from accessing offline, in-person resources [8], highlighting the increasing importance of web-based communities as accessible, anonymous, and scalable recovery support platforms [9,10].

Reddit, a widely used social media platform, hosts numerous topic-focused communities, or "subreddits," where users can connect, share experiences, and offer support. Among these, r/OpiatesRecovery has emerged as the largest recovery-focused subreddit for individuals with OUD [11]. As of December 2024, it had over 50,000 subscribers, with even nonsubscribed users able to post. With its title line, "You are not alone in this fight," the subreddit encourages individuals to engage in self-disclosure, the act of sharing personal thoughts, emotions, or experiences [12]. Self-disclosure plays a central role in the recovery process, particularly in digital spaces where anonymity reduces fears of stigma and encourages open dialogue.

Self-disclosure in web-based communities not only facilitates personal expression but also elicits social support from peers who have undergone similar experiences [13,14]. According to social penetration theory (SPT), interpersonal relationships develop through progressive layers of self-disclosure, where deeper and more personal revelations foster greater intimacy and supportive exchange [15]. The relationship between self-disclosure and social support has been examined in other vulnerable or stigmatized contexts in web-based communities. For example, De Choudhury and De [16] found that Reddit users disclosing mental health struggles using emotionally candid ("low inhibition") language received more social support. Similarly, Andalibi et al [17] revealed that Instagram posts tagged with #depression and disclosed personal narratives, illness, or self-appearance concerns tended to receive more positive social support.

However, in the context of OUD recovery, self-disclosure and social support as well as the relationship between them remain largely underexamined. Existing Reddit-based substance use studies have primarily focused on high-level trending topics and language features [18-20]. A recent review found that only 6 out of 60 studies examined user-to-user interactions, such as seeking and providing support, with even fewer studies focusing specifically on self-disclosure or social support [21]. Moreover, existing studies often rely on small datasets or lack in-depth analysis, restricting their ability to produce generalizable findings applicable to broader OUD recovery communities [22,23].

In recent years, natural language processing (NLP) methods have been widely adopted for analyzing social media data to study health-related web-based communication, particularly at scale [24,25]. In the context of OUD, recent research by Balsamo et al [26] examined peer support within r/OpiatesRecovery, highlighting the importance of recognition, acknowledgment, and knowledge exchange in sustained engagement by analyzing 2125 recovering users. Their work demonstrated the potential of Reddit as a complementary intervention for individuals with OUD through the application of NLP approaches. However, the relationship between the content of self-disclosure and the type of support received is still underexplored.

To address these gaps, we conduct a large-scale, machine learning-driven NLP analysis of r/OpiatesRecovery. Specifically, we analyze 32,810 posts and 324,224 comments spanning 8 years (2014 - 2022). Our findings revealed that most posts were made by individuals in addiction or early stages of recovery, while contributions from individuals in long-term recovery were limited. Posts containing informational and emotional self-disclosure were more likely to elicit support from the community, although the type and amount of support varied, underscoring the importance of disclosure in fostering engagement.

This formative research contributes empirical, large-scale evidence to the feasibility of web-based communities like Reddit as peer-driven support systems. The contributions of this study to the field of OUD recovery research are threefold. First, it characterizes the populations using web-based recovery communities, patterns of self-disclosures, and the types of social support they exchange. Second, it contributes empirical evidence to the relationship between self-disclosure and social support for the OUD recovery community. Third, it also discusses design insights for improving web-based recovery communities, such as addressing limited engagement from long-term recovery users and the types of support less prevalent.

Objectives

This study aims to explore self-disclosure and social support in web-based OUD recovery communities. By leveraging the open-access Reddit data, we seek to answer 3 research questions (RQs): (RQ1) What content do users disclose in a web-based community for opioid recovery?, (RQ2) What types of social support do users receive within the community?, and (RQ3) How does the content disclosed relate to the type and extent of social support received? By addressing these questions, this study enhances our understanding of the role of web-based communities in supporting individuals on their journey of recovery from OUD.

Methods

Data Collection

The study used data from r/OpiatesRecovery. Data collection used the Python Reddit Application Programming Interface Wrapper [27] to retrieve all the posts, comments, and associated metadata from January 1, 2014, to May 5, 2022. A line chart of the number of posts and comments from 2014 to 2022 is included in Multimedia Appendix 1.

Data Preprocessing

Prior to analysis, posts that were deleted or removed at the time of analysis were excluded to ensure the dataset reflected active and accessible content during the study period. In addition, entries with missing or malformed timestamps or text were filtered out. For model training, all text data were tokenized using the WordPiece tokenizer, as implemented by the pretrained BERT-large-uncased model (bidirectional encoder representations from transformers) in the transformers library [28]. While our analysis was conducted at the post and comment level, we also calculated the number of unique users (n=12,375) who contributed posts, to provide additional context on user

XSL•FO
RenderX

engagement. However, we did not perform a user-level analysis, as the focus of this study was on the characteristics of individual posts and comments rather than user trajectories. As a result, the dataset consists of 32,810 initial posts and 324,224 corresponding comments.

Analysis of Initial Posts

Semiopen coding was guided by an initial codebook derived from prior literature and refined through pilot coding. The finalized codebook aimed to identify 5 categories of information: demographic information, informational self-disclosure, emotional self-disclosure [29], opioid use and recovery stages [30,31], and the goal of the post [14]. The descriptions and deidentified examples of each category are provided in Table 1. To ensure reliability, 200 randomly selected posts were independently coded by the first 2 authors, followed by weekly discussions to reconcile any disagreements. This process resulted in a good to a substantial level of agreement across all categories between the 2 annotators, as measured by Cohen κ ranging from 0.55 to 0.78 [32,33]. We then applied the codebook to manually annotate 462 randomly selected posts.

Table . Codebook for annotating initial posts.

Category, subcategory, and description	Example
Demographic information	
Age, gender, employment if disclosed	"Hey guys. I am a 36 y old man from the US."
Opioid use and recovery stages	
Addiction: actively using	"i still take codeine."
Initial recovery: abstinence lasts for less than 1 mo	"Been clean nearly two weeks now…"
Sustained recovery: abstinence lasts for 1 mo to 5 y	"I have 14 mo of sober time from heroin/opiates."
Stable recovery: abstinence lasts for more than 5 y	"I am an ex-opiate user. abstaining completely for 6 y."
Unknown stage: cannot tell or not disclosed	__[a]
Informational self-disclosure	
Whether the post contains disclosure of personal information, such as using history, financial information, or trauma experience, etc	"The first time I took hydrocodone I was 13, smoked oxy at 14, injected heroin for the first time at 16."
Emotional self-disclosure	
Whether the post contains disclosure of personal feelings, such as concerns, frustrations, humiliation, agony, or fears, etc	"I'm terrified of the withdrawal again…"
Goal of the post	
Seeking informational support: whether the post contains the goal of seeking informational support	"What are some methods you guys in recovery use to manage/shape your digestive systems?"
Seeking emotional support: whether the post contains the goal of seeking emotional support	"How long until I turn a corner and feel better? I just need to woman up...pls help me thx."
Providing informational support: whether the post contains the goal of providing informational support	"The XXX [a medication] did wonders for me made the monster much more bearable. Just wanted to share a different way."
Providing emotional support: whether the post contains the goal of providing emotional support	"I'm truly blessed and just wanted to share my hope with everyone."

[a]No quote provided; the recovery stage could not be inferred from the post.

Given the impracticality of manually annotating all 32,810 posts, we used a machine learning approach by developing BERT-based classifiers for each category using annotated posts, except for demographic information, which was excluded from machine learning due to the limited and sparse data available. We opted for BERT as our foundational NLP model due to its demonstrated superiority in handling the intricate linguistic nuances prevalent in web-based communication, particularly within the specialized context of the r/OpiatesRecovery community. Unlike traditional machine learning models such as term frequency-inverse document frequency or simpler neural networks (eg, long short-term memory networks), which process text sequentially or rely on bag-of-words assumptions, BERT's bidirectional nature allows it to pretrain on vast amounts of text data, capturing deep contextual relationships between words [28]. This capability is critical for accurately interpreting the often informal, emotionally charged, and jargon-filled language characteristic of web-based recovery forums, where the meaning of a word can be heavily dependent on its surrounding context. Furthermore, while other transformer-based models exist, BERT's robust pretraining on a large corpus and its proven effectiveness across a wide range of natural language understanding tasks—particularly when fine-tuned for specific classification objectives [34]—made it a well-suited and efficient choice for scaling our analysis from a manually annotated subset to the full dataset of over 32,000 posts. Details of model architecture, training configuration, and hyperparameters are provided in the Multimedia Appendix 2. Our approach maintains interpretability by directly predicting predefined, human-interpretable categories (eg, recovery stage,

XSL•FO
RenderX

informational self-disclosure), which allows for intuitive post hoc analysis of community discourse.

Regarding informational and emotional self-disclosure, we exclusively relied on our annotated datasets. To develop the opioid use and recovery stage classifier, we leveraged the Drug Abuse Reddit dataset from Ghosh et al [31], which provided a more balanced distribution of recovery stages. This dataset contains 3151 Reddit posts categorized into 5 recovery stages: "Addicted," "E-Recovery" (Early), "M-Recovery" (Maintaining), "A-Recovery" (Advanced), and "Others." We selectively incorporated examples from the "Addicted," "M-Recovery," and "A-Recovery" categories to align with our target classification schema: addiction, sustained recovery, and stable recovery. The inclusion of this external data allowed us to address the class imbalance and ensure that the classifier was exposed to sufficient samples from each stage during training, ultimately enhancing the classifier's performance. In addition, to bolster our annotated dataset for providing informational support and emotional support, we integrated data from prior work, which annotated information and emotional support within 500 comments from a web-based community for individuals with ovarian cancer [35]. Opioid recovery and ovarian cancer represent distinct health contexts, which may influence specific language use and communication norms. However, prior research suggests that core categories of social support, particularly informational and emotional support, are broadly applicable across web-based health communities [36,37]. These support types reflect general communicative functions that transcend disease-specific contexts. This approach enabled us to expand our dataset and improve the robustness of our analysis. Table 2 contains comprehensive information regarding the performance of each model, alongside interrater agreement between human annotators.

XSL•FO
RenderX

Table . Interrater agreement between human annotators and performance of automatic classifiers for classifying initial posts.

Category	Interrater agreement		Classifiers' performance		
	Percent agreement	Cohen κ	Precision	Recall	F_1-score
Opioid use and recovery stages	74.3	0.65			
Addiction			0.95	0.83	0.89
Initial recovery			0.64	0.75	0.69
Sustained recovery			0.85	0.93	0.89
Stable recovery			0.91	0.88	0.89
Unknown stage			0.64	0.85	0.73
Informational self-disclosure	96.6	0.78			
Yes			0.94	1.00	0.97
No			1.00	0.67	0.8
Emotional self-disclosure	84.9	0.63			
Yes			0.59	0.68	0.63
No			0.88	0.83	0.85
Providing informational support	91.1	0.60			
Yes			0.88	0.95	0.91
No			0.97	0.92	0.94
Providing emotional support	95.0	0.64			
Yes			0.88	0.93	0.90
No			0.97	0.92	0.94
Seeking informational support	80.4	0.58			
Yes			0.81	0.89	0.85
No			0.93	0.86	0.89
Seeking emotional support	77.7	0.55			
Yes			0.77	0.66	0.71
No			0.77	0.74	0.76

Analysis of Comments

The analysis of comments focused on identifying 2 types of social support: informational support and emotional support, as these are the most commonly exchanged types of social support in web-based communities [36,37]. Additionally, informational support was further classified into 5 subcategories based on the content of the information: advice, referral, fact and situational appraisal, personal experience, and opinion [38]. The descriptions and corresponding deidentified example comments are provided in Table 3.

To identify support types in comments, we leveraged models trained on annotated data from an ovarian cancer web-based community [35]. Each comment could provide 0, 1, or multiple types of social support. Model details are provided in Multimedia Appendix 2 and an open repository [39]. To assess the models' transferability to the OUD recovery context, we conducted a manual evaluation on 50 randomly sampled comments per social support category. The first 2 authors, one with expertise in information science and the other in health communication, independently annotated the samples, achieving good to substantial interrater agreement. Discrepancies were resolved through discussion, and model predictions were evaluated against the finalized annotations. All models demonstrated high performance, with F_1-scores ranging from 0.78 to 0.94, indicating reasonable generalizability to the OUD recovery domain. Table 4 shows the interrater agreement between human annotators and model performance across all the social support categories.

XSL•FO
RenderX

Table . Codebook for annotating social support in comments.

Social support type	Description	Example
Advice	Whether the comment offers ideas and suggestions for coping with challenges	"Find a group/groups and connect with someone who can help. Ask your friends and family (if they know; if not, you might consider telling them, if you think they'd respond in a supportive manner)…"
Referral	Whether the comment refers to information sources, eg, books, websites, contacts.	" Here is one I found very quickly explaining the issue with [a substance] and [a medication] introduction. https://___"
Fact and situational appraisal	Whether the comment offers facts or reassesses the situation.	"For that specific situation, [a medication] implant might be most effective…"
Personal experience	Whether the comment shares personal stories or incidents.	"I am on [a medication] now successfully for at least a year."
Opinion	Whether the comment offers a view or judgment formed about something. It is not necessarily based on fact or knowledge.	"Also Physical therapy works. It is literally the best thing you can do to manage long term pain"
Emotional support	Whether the comment offers empathy, encouragement, or appreciation.	"Good luck and stay strong"

Table . Interrater agreement between human annotators and performance of automatic classifiers for classifying social support types in comments.

Category	Interrater agreement		Classifiers' performance		
	Percent agreement	Cohen κ	Precision	Recall	F_1-score
Advice	96.0	0.91	0.94	0.94	0.94
Referral	98.0	0.90	1.00	0.80	0.89
Fact and situational appraisal	80.0	0.52	0.66	0.95	0.78
Personal experience	94.0	0.88	0.85	0.96	0.90
Opinion	74.0	0.48	0.85	0.88	0.86
Emotional support	96.0	0.92	0.88	0.92	0.90

Regression Analysis

Pearson correlations were first conducted in our preliminary analysis to identify the relevant variables for inclusion in our model. After confirming adherence to statistical assumptions, we proceeded with multiple regression analysis for each outcome variable, including advice, referral, fact and situational appraisal, personal experience, opinion, and emotional support. As we were interested in comparing the frequency of different types of social support received across various posts, we divided each outcome variable by the total comment count of the respective posts. This normalization process ensures that we can consistently evaluate and compare different types of social support received while mitigating the impact of varying comment counts across posts. Additionally, to facilitate comparisons between each OUD use and recovery stage, we used a dummy coding scheme for the stages, designating the first stage (addiction stage) as our reference group. We selected the addiction group as the reference category, rather than the overall mean, to enable direct comparisons across recovery stages. A sensitivity analysis was conducted to assess potential bias from outliers, and the results showed no significant changes in the findings. Furthermore, a natural log transformation was applied to the referral variable to address the violation of the normality assumption.

Ethical Considerations

This study was determined to be not human research and received an exemption from review by the Institutional Review Board at the University of Kentucky. The study involved no interaction or intervention with human participants. All data were collected from publicly accessible information on Reddit, and no account or login procedures were required. Data collection procedures adhered to Reddit's terms of service, ensuring ethical compliance throughout the study. To better protect user privacy, all usernames have been excluded from this paper. In addition, no post titles or URLs are reported, and we avoid quoting content that could reasonably be traced back to specific users.

Results

Overview

Among the 462 human-annotated posts, only 5.6% (n=26) of posts disclosed the age of the individual using opioids, which ranged from 14 to 52 years (mean 27.85, SD 7.43). A total of 8.0% (n=37) of posters identified themselves as men (19/462,

XSL•FO
RenderX

4.1%), women (17/462, 3.7%), or nonbinary (1/462, 0.2%). Regarding employment, 16.6% (n=79) of users disclosed their employment status, with 13.6% (n=63) being employed and 6.9% (n=16) unemployed. Due to the low prevalence of demographic self-disclosure, we did not attempt to infer or model the relationship between user demographics and social support patterns. The remaining results presented are based on the full dataset of 32,810 posts and 324,224 comments.

What Content Do Users Disclose in a Web-Based Community for Opioid Recovery?

Figure 1 presents the frequency distribution of posts, categorized by the content disclosed within them. Among the 32,810 posts, 22.0% of them (n=7225) were posted by individuals actively using opioids, indicating the addiction stage. Another 27.7% (n=9080) were attributed to the initial recovery stage (abstinence of less than 1 month). Additionally, 18.1% (n=5944) represented the sustained recovery stage (abstinence lasting from 1 month to 5 years), while only 1.8% (n=600) pertained to the stable recovery stage (abstinence exceeding 5 years). Notably, 30.4% (n=9961) of posts did not explicitly disclose their recovery stage. Most posts included informational self-disclosure (28,977/32,810, 88.3%) and emotional self-disclosure (24,816/32,810, 75.6%). However, most posts did not provide informational support (28,290/32,810, 86.2%) or emotional support (27,481/32,810, 83.8%). It is noteworthy that a higher percentage of posts sought informational support (19,153/32,810, 58.4%), while only a few posts explicitly sought emotional support (779/32,810, 2.4%).

Figure 1. Frequency distribution of posts categorized by content disclosed. OUD: opioid use disorder.

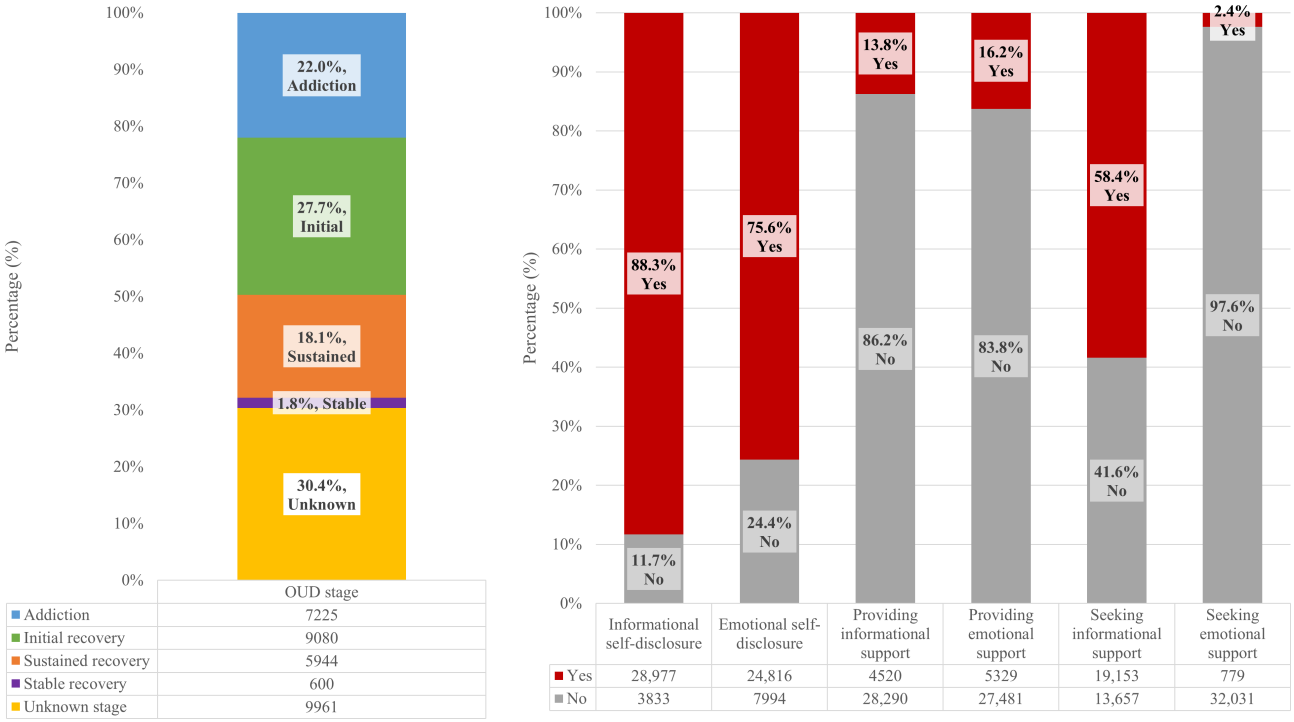

What Types of Social Support Do Users Receive Within the Community?

Table 5 presents the average number of comments per post, the distribution of different types of social support embedded within those comments, as well as correlations among the various types of social support. On average, each post receives approximately 9.88 (SD 11.36) comments. Fact and situational appraisal emerged as the most common form of support, with an average of 5.62 (SD 6.82) comments per post. This was followed by the sharing of personal experiences, averaging 4.88 (SD 5.98) comments per post. Opinions, which reflect individual viewpoints, accounted for an average of 4.36 (SD 5.60) comments per post. Emotional support constituted an average of 3.85 (SD 4.81) comments, while advice was present in approximately 2.61 (SD 3.45) comments on average. Notably, referrals appear less frequently, with posts typically containing fewer than one comment providing referral details, averaging 0.61 (SD 0.50) comments. Table 5 also depicts Pearson correlations among the number of comments and different types of social support. Positively significant correlations were observed for all pairs of variables.

XSL•FO
RenderX

Table . Pearson correlations among the number of comments and the number of different types of social support.

	Mean (SD)	Comment number	Advice	Referral	Fact and situational appraisal	Personal experience	Opinion	Emotional support
Comment number	9.88 (11.36)	—[a]	0.725[b]	0.357[b]	0.893[b]	0.918[b]	0.891[b]	0.837[b]
Advice	2.61 (3.45)	0.725[b]	—	0.267[b]	0.845[b]	0.722[b]	0.796[b]	0.639[b]
Referral	0.16 (0.50)	0.357[b]	0.267[b]	—	0.323[b]	0.313[b]	0.335[b]	0.266[b]
Fact and situational appraisal	5.62 (6.82)	0.893[b]	0.845[b]	0.323[b]	—	0.910[b]	0.927[b]	0.692[b]
Personal experience	4.88 (5.98)	0.918[b]	0.722[b]	0.313[b]	0.910[b]	—	0.903[b]	0.803[b]
Opinion	4.36 (5.60)	0.891[b]	0.796[b]	0.335[b]	0.927[b]	0.903[b]	—	0.770[b]
Emotional support	3.85 (4.81)	0.837[b]	0.639[b]	0.266[b]	0.692[b]	0.803[b]	0.770[b]	—

[a]Not applicable.

[b]All the correlations are significant at $P<.001$.

How Does the Content Disclosed Relate to the Type and Extent of Social Support Received?

Results from multiple regression analysis for the number of comments with advice revealed that all predictors together explained 11.9% of the variance in the number of comments with advice ($R^2=0.12$, $F_{10,\ 30,470}=412.06$, $P<.001$). For the number of referrals, the regression analysis showed that stable recovery, informational self-disclosure, emotional self-disclosure, providing informational support, providing emotional support, and seeking informational support collectively accounted for 3.0% of the variance ($R^2=0.03$, $F_{10,\ 3971}=12.16$, $P<.001$). Similarly, the analysis for fact and situational appraisal indicated that initial recovery, sustained recovery, stable recovery, unknown stage, informational self-disclosure, emotional self-disclosure, providing informational support, providing emotional support, and seeking informational support jointly explained 18.7% of the variance ($R^2=0.19$, $F_{10,\ 30,470}=699.48$, $P<.001$). The analysis further revealed that stable recovery, emotional self-disclosure, providing informational support, providing emotional support, and seeking informational support together explained 3.0% of the variance in the number of personal experiences ($R^2=0.03$, $F_{10,\ 30,470}=93.32$, $P<.001$). Regarding the number of opinions, the regression analysis demonstrated that initial recovery, sustained recovery, informational self-disclosure, emotional self-disclosure, providing emotional support, seeking informational support, and seeking emotional support jointly accounted for 4.8% of the variance ($R^2=0.05$, $F_{10,\ 30,470}=153.59$, $P<.001$). Last, the analysis for the number of comments with emotional support indicated that all the predictors except for informational self-disclosure collectively explained 13.2% of the variance ($R^2=0.13$, $F_{10,\ 30,470}=463.18$, $P<.001$). Table 6 contains a summary of the regression. A detailed statistical report is in Multimedia Appendix 1.

XSL•FO
RenderX

Table . Summary of hierarchical regression analysis for variables predicting satisfaction.

Variables	Advice	Referral[a]	Fact and situational appraisal	Personal experience	Opinion	Emotional support
Initial recovery[b]						
β	−0.06	−0.03	−0.06	0.01	−0.11	0.03
P value	<.001	.21	<.001	.35	<.001	<.001
Sustained recovery[b]						
β	−0.08	−0.01	−0.10	−0.01	−0.03	0.08
P value	<.001	.74	<.001	.11	<.001	<.001
Stable recovery[b]						
β	−0.05	0.04	−0.05	−0.01	−0.01	0.02
P value	<.001	.01	<.001	.04	.21	<.01
Unknown stage[b]						
β	−0.10	0.02	−0.07	−0.02	0.00	−0.05
P value	<.001	.29	<.001	.06	.70	<.001
Informational self-disclosure						
β	0.02	0.11	0.04	0.00	0.02	−0.01
P value	<.01	<.001	<.001	.73	<.01	.38
Emotional self-disclosure						
β	0.12	−0.08	0.05	0.07	0.14	0.17
P value	<.001	<.001	<.001	<.001	<.001	<.001
Providing informational support						
β	−0.08	−0.10	−0.07	0.04	−0.01	0.03
P value	<.001	<.001	<.001	<.001	.39	<.001
Providing emotional support						
β	−0.09	0.09	−0.13	−0.10	−0.07	0.10
P value	<.001	<.001	<.001	<.001	<.001	<.001
Seeking informational support						
β	0.17	0.09	0.30	0.12	0.11	−0.23
P value	<.001	<.001	<.001	<.001	<.001	<.001
Seeking emotional support						
β	0.03	0.01	0.01	0.00	0.01	0.05
P value	<.001	.60	.23	.62	.02	<.001
Adjusted R^2	0.12	0.03	0.19	0.03	0.05	0.13
R^2	0.12	0.03	0.19	0.03	0.05	0.13

[a]Natural log transformation was applied to address the violation of the normality assumption.

[b]The initial recovery, sustained recovery, stable recovery, and unknown stage were dummy-coded to compare with the addition stage.

Discussion

Content Disclosed in the Web-Based Opioid Recovery Community

Even though results based on the randomly sampled 462 posts revealed that the majority of posters did not disclose personal details such as age, gender, or employment status, other aspects of informational self-disclosure, such as financial status and history, were prevalent across all the analyzed posts (28,977/32,810, 88.3%). Emotional self-disclosure was also markedly present (24,816/32,810, 75.6%). The finding aligns with the inherent anonymity of web-based platforms, which often encourage personal self-disclosure [40]. Notably, an insubstantial fraction of posts (779/32,810, 2.4%) conveyed a quest for emotional support, while more than half

XSL•FO
RenderX

(19,153/32,810, 58.4%) aimed at seeking informational support from the community. These patterns suggest a preference for seeking information—such as ideas, insights, or advice—within the community, rather than seeking emotional or moral solidarity. However, the results raise questions about whether web-based communities like Reddit are a good resource for OUD-related advice. Are there sufficient experienced and knowledgeable members providing informational support in the community?

Our findings regarding the trajectory of the recovery stage further underscore this concern. Notably, the vast majority of posts were centered on initial recovery (27.7%, abstinence lasts for less than 1 mo) or addiction (22.0%, actively using). In contrast, posts from individuals in stable recovery, those who have sustained abstinence for over 5 years, accounted for only 1.8% of all posts. This imbalance highlights the pressing need for support, especially in the vulnerable initial and early stages of recovery. These stages are often marked by uncertainty, intense challenges, and the need for consistent guidance [41,42]. However, the scarcity of contributions from individuals in long-term recovery, those who have successfully navigated the extended journey of recovery, suggested a potential lack of experienced guidance within the r/OpiatesRecovery subreddit.

Several strategies can be devised to bridge the gap and enrich the community's collective experience and knowledge base. First, dedicated threads for sharing success stories could be created and pinned to encourage long-term users to share their experiences and continue their engagement [43]. Reddit's social norms inherently promote mutual support and experience sharing [44]. In addition, health advocates are encouraged to use Reddit as a valuable channel for disseminating educational resources, articles, and interventions. For example, Clendennen et al [45] conducted a longitudinal investigation involving 5482 students, revealing a positive relationship between exposure to social media tobacco messages and self-reported tobacco use. Their findings suggest that exposure to health-related messages, including posts on Reddit, might shape social media users' health behavior. In light of these findings, public health agents can play a pivotal role in enhancing Redditors' awareness of accessible resources and potentially influencing their health-related behaviors.

Social Support in the Web-Based Opioid Recovery Community

Regarding the amount of each type of social support, a large number of fact and situational appraisals were first noticed. This might indicate a collective understanding that recovery is not just an emotional journey, but also one requiring concrete steps, knowledge, and understanding [46,47]. The sharing of facts and knowledge about OUD and recovery can potentially empower peer users in the community, providing them with tangible guidance they can apply in their own recovery journey. However, particular caution should be paid to the quality and accuracy of the information, including facts and advice, shared in web-based communities like Reddit. For example, ElSherief et al [48] discovered that the myth that medication treatment for OUD, such as agonist therapy or medication-assisted treatment, merely substitutes one drug for another is present in 4 out of 1000 posts on web-based health communities like Reddit and Drugs-Forum. Users should be encouraged to cross-reference the information they acquire with credible sources and consult professionals to ensure informed decisions in their recovery.

Furthermore, the prevalence of personal experiences underscores the value of storytelling in recovery. Efforts should be made to enhance and promote storytelling among individuals dealing with OUD. This could involve implementing digital storytelling tools, organizing workshops, and providing resources that empower individuals to share their experiences and journeys toward recovery [49,50]. In addition, the limited number of referrals provided is noteworthy. Similar results were revealed by D'Agostino et al [22]. Through analyzing 100 posts related to OUD on Reddit, they found that only 15% of the posts direct individuals with OUD issues to health care professionals. Resources shared by peers in web-based communities could potentially promote resource dissemination and benefit members [51]. However, the limited sharing of resources among individuals with OUD raises questions: Is it due to the lack of user-preferred credible resources for individuals with OUD or the lack of awareness of available resources? Further exploration is required to understand the reasons behind the limited sharing of resources among users with OUD.

Relationship Between Disclosure and Received Social Support

The multiple regression results offer a nuanced understanding of how post content interacts with the types of social support provided within comments in the web-based community. We first found a differential pattern of support across the recovery stages. Specifically, posts indicating a current addiction stage tend to receive significantly more advice and facts from the community, but less emotional support compared with posts that reveal other stages of recovery. This pattern may reflect the community's focus on offering tangible solutions to individuals actively struggling with addiction. However, the reduced emotional support for this group suggests a potential area of improvement, as individuals in the addiction might also benefit from empathetic interactions [52].

The findings also emphasize the critical role of self-disclosure in shaping the dynamics of community interactions. Our results indicate that posts containing self-disclosures generally receive more comments offering social support than those without such disclosures. Such observations align with recent Reddit-based research in the context of female infertility [53]. This pattern is also consistent with SPT [15], which posits that individuals tend to respond to others' disclosures based on the norm of reciprocity. That is, as one person reveals personal information, others feel encouraged to reciprocate with similarly meaningful responses [54]. In the context of r/OpiatesRecovery, self-disclosure functions as a relational cue, signaling openness and vulnerability that invite deeper engagement from community members.

More importantly, the type of self-disclosure affects the type and variety of social support received. Our results indicated emotional self-disclosure posts prompt more diverse and affectively rich responses (eg, emotional support and personal

XSL•FO
RenderX

experience) than those focused solely on factual or situational information. This finding aligns with SPT, which suggests that emotional self-disclosures represent a deeper and more intimate layer of self-revelation, which tends to foster stronger interpersonal resonance and encourage others to respond with similarly deep responses [55]. Previous experimental research in both web-based and in-person contexts has also shown that emotional content promotes reciprocal self-disclosure, encouraging others to respond more openly and supportively [56].

Interestingly, posts with emotional self-disclosure tend to gather fewer referrals than those without emotional self-disclosure. This may reflect the nature of emotional disclosures as relational rather than instrumental cues, as suggested by SPT, which may lower the salience of problem-solving responses like referrals and increase the likelihood of affective support[55]. When Reddit users share emotionally vulnerable content, it may evoke a sense of sympathy, closeness, or relational depth among community members. As a result, responders might prioritize empathy and subjective perspectives over pointing to external resources.

It is also important to consider how Reddit differs from other social media platforms in supporting individuals with OUD. Reddit's affordances (eg, anonymous, long-form posting, and topic-specific subreddits) may foster an environment where users often feel more comfortable engaging in candid self-disclosure and sustained peer-to-peer support. In contrast, Twitter is a public platform where visibility is shaped by algorithms and brevity is prioritized. Prior research indicates that Twitter tends to be institution-driven, with a significant presence in addiction treatment marketing and professional accounts. For example, Russell et al [57] found that 27.4% of opioid-related tweets were associated with addiction treatment marketing, some of which promoted scientifically unsupported claims. Similarly, Tofighi et al [58] reported that 24.6% of opioid-related tweets came from private residential or detoxification programs, while only 4.7% were authored by people who use opioids. In our Reddit dataset (462 human-annotated posts), none were authored by institutions. These findings suggest that Twitter serves both as a platform for recovery support and as a channel for institutional messaging and commercial promotion—unlike Reddit, which fosters greater narrative depth and peer engagement among individuals with lived experience.

Facebook, though more community-oriented, differs in that it typically requires some form of identity verification to join private groups. This partial transparency may encourage more Facebook users to offer social support, as the perceived accountability can foster a stronger sense of community than the fully anonymous environment on Reddit. For instance, Ellway et al [59] analyzed over 14,000 posts from the Australian Facebook page "Never Give Up Giving Up Ice, Drugs" and found that 24.2% of these posts provided informational support while 42.1% offered emotional support. In contrast, our findings from Reddit showed lower rates of support provided, with 13.5% of posts offering informational support and 16.3% providing emotional support. These distinctions highlight how different platform designs shape the nature of recovery dialogue and support. To conclude, Reddit's anonymity may invite more vulnerable disclosures, while Facebook's identity-linked structure may encourage more expert-driven support. Twitter, by contrast, appears better suited for rapid information dissemination, public awareness-raising, and promotional messaging. Future research should consider triangulated cross-platform comparisons to better understand how structural and functional differences across social media platforms influence recovery experiences and outcomes.

Limitations and Future Directions

This study has several limitations. First, our data was exclusively collected from the r/OpiatesRecovery subreddit, which introduces potential sampling bias. The dataset represents a distinct subset of individuals in opioid recovery who actively engage with Reddit communities, potentially differing from those who do not participate in such web-based communities. Therefore, the generalizability of our findings beyond the Reddit community is constrained. Caution should be noted when generalizing these insights to the broader population of individuals in opioid recovery. Moreover, the data gathered from open Reddit discussions is self-reported and lacks clinical verification, limiting the reliability of our dataset [21]. Second, there is a scarcity of demographic information, such as age, gender, and location information, as well as long-term opioid use and recovery. While we manually coded a small set of posts for basic demographic information, only a small percentage of users disclosed such information, which limited our ability to analyze how support patterns may vary by demographics. Access to such demographic and long-term recovery data could provide deeper insights into the opioid recovery landscape and enhance future studies. Third, while we conducted a manual evaluation to confirm the applicability of support-type classification models trained on non-OUD forums, we acknowledge the risk of domain shift. The language used in OUD communities may include slang or emotionally intense narratives that differ from those in cancer-related forums. As such, the models may underperform in capturing OUD-specific linguistic nuances. Future work should consider domain-adaptive training or fine-tuning approaches to enhance model validity within substance use recovery contexts. Fourth, another limitation concerns the detection of emotional content. While the informational self-disclosure classifier performed well (F_1=0.97), the emotional self-disclosure classifier showed lower performance (F_1=0.63), potentially leading to the underrepresentation of emotional content in our analyses. Similarly, the low rate of posts classified as seeking emotional support (2.4%) may reflect both actual user behavior and the model's difficulty detecting subtle or indirect emotional expressions. For example, a post describing feelings of isolation might implicitly seek empathy and connection, even if it is framed as a request for advice on finding local support groups. Such cases may lead to an underestimation of the true prevalence of emotional support-seeking in the dataset. Future work should explore improved emotional modeling, such as fine-tuned lexicons, affective embeddings, or human-in-the-loop methods to better capture emotional nuance in web-based recovery discussions [60,61]. Last, we acknowledge that deleted Reddit posts may lead to an incomplete representation of self-disclosure and

XSL•FO
RenderX

support patterns, particularly for vulnerable or stigmatized groups, which may potentially impact our findings. These limitations should be noted when interpreting our findings and highlight the need for future research to provide a more comprehensive understanding of the complex dynamics within web-based communities focusing on opioid recovery.

Conclusions

In conclusion, this study used a machine-learning approach to analyze a large-scale dataset, focusing on self-disclosure and social support interactions within a web-based opioid recovery community. The findings revealed that most posts were made by individuals in early recovery stages, with informational and emotional self-disclosure being prevalent and often eliciting more support. Most posts seek informational support rather than emotional support. Fact and situational appraisal was the most common type of social support provided. These results offer valuable insights into the recovery experiences and peer interactions of individuals with OUD in web-based communities. While highlighting the potential of these platforms to complement traditional recovery interventions, the study also identifies areas for improvement, such as fostering long-term engagement and enhancing the quality of shared information. Future research should address these gaps to maximize the effectiveness of online support groups.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Authors' Contributions

YC contributed to the conceptualization, data curation, formal analysis, and methodology of the study, and was involved in writing the original draft and reviewing and editing the manuscript. HC contributed to the formal analysis and methodology and participated in writing the original draft, reviewing, and editing. KT also contributed to the formal analysis and methodology and helped draft, reviewed, and edited the original manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary results.
[DOCX File, 98 KB - formative_v9i1e71207_app1.docx]

Multimedia Appendix 2
Model configuration and hyperparameters.
[DOCX File, 29 KB - formative_v9i1e71207_app2.docx]

References

1. The opioid crisis: an overview. MedlinePlus. 2024. URL: https://magazine.medlineplus.gov/article/the-opioid-crisis-an-overview [accessed 2024-01-05]
2. Dowell D, Brown S, Gyawali S, et al. Treatment for opioid use disorder: population estimates - United States, 2022. MMWR Morb Mortal Wkly Rep 2024 Jun 27;73(25):567-574. [doi: 10.15585/mmwr.mm7325a1] [Medline: 38935567]
3. World drug report 2024. United Nations Office on Drugs and Crime. 2024. URL: https://www.unodc.org/unodc/en/data-and-analysis/world-drug-report-2024.html [accessed 2025-01-14]
4. McGaffin BJ, Deane FP, Kelly PJ, Blackman RJ. Social support and mental health during recovery from drug and alcohol problems. Addict Res Theory 2018 Sep 3;26(5):386-395. [doi: 10.1080/16066359.2017.1421178]
5. Brousseau NM, Karpyn A, Laurenceau JP, et al. The impacts of social support and relationship characteristics on commitment to sobriety among people in opioid use disorder recovery. J Stud Alcohol Drugs 2022 Sep;83(5):646-652. [doi: 10.15288/jsad.21-00225] [Medline: 36136434]
6. Rokiyah R, Zuanda N, Alrefi A, Akbari A. The role of social support in substance addiction recovery: a systematic review. EDUSOSHUM 2024;4(2):213-222. [doi: 10.52366/edusoshum.v4i2.112]
7. Costello MJ, Li Y, Remers S, et al. Effects of 12-step mutual support and professional outpatient services on short-term substance use outcomes among adults who received inpatient treatment. Addict Behav 2019 Nov;98:106055. [doi: 10.1016/j.addbeh.2019.106055] [Medline: 31357071]
8. Bremer W, Plaisance K, Walker D, et al. Barriers to opioid use disorder treatment: a comparison of self-reported information from social media with barriers found in literature. Front Public Health 2023;11:1141093. [doi: 10.3389/fpubh.2023.1141093] [Medline: 37151596]
9. Alambo A, Padhee S, Banerjee T, Thirunarayan K. COVID-19 and mental health/substance use disorders on Reddit: a longitudinal study. arXiv. Preprint posted online on Nov 20, 2020. [doi: 10.48550/arXiv.2011.10518]

XSL•FO
RenderX

10. Cline RJW, Haynes KM. Consumer health information seeking on the Internet: the state of the art. Health Educ Res 2001 Dec;16(6):671-692. [doi: 10.1093/her/16.6.671] [Medline: 11780707]
11. R/OpiatesRecovery. Reddit. URL: https://www.reddit.com/r/OpiatesRecovery/ [accessed 2024-05-05]
12. Wheeless LR, Grotz J. Conceptualization and measurement of reported self-disclosure. Human Comm Res 1976 Jun;2(4):338-346. [doi: 10.1111/j.1468-2958.1976.tb00494.x]
13. Wang YC, Kraut RE, Levine JM. Eliciting and receiving online support: using computer-aided content analysis to examine the dynamics of online social support. J Med Internet Res 2015 Apr 20;17(4):e99. [doi: 10.2196/jmir.3558] [Medline: 25896033]
14. Yang D, Kraut R, Smith T, Mayfield E, Jurafsky D. Seekers, providers, welcomers, and storytellers: modeling social roles in online health communities. Presented at: CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems; May 4-9, 2019; Glasgow, Scotland, UK. [doi: 10.1145/3290605.3300574]
15. Altman I, Taylor DA. Social Penetration: the Development of Interpersonal Relationships: Holt, Rinehart & Winston; 1973.
16. De Choudhury M, De S. Mental health discourse on Reddit: self-disclosure, social support, and anonymity. ICWSM 2014;8(1):71-80. [doi: 10.1609/icwsm.v8i1.14526]
17. Andalibi N, Ozturk P, Forte A. Sensitive self-disclosures, responses, and social support on Instagram: the case of #depression. Presented at: Proceedings of the 2017 ACM Conference on Computer-supported Cooperative Work and Social Computing; Feb 25 to Mar 1, 2017; Portland, OR, USA. [doi: 10.1145/2998181.2998243]
18. Graves RL, Perrone J, Al-Garadi MA, et al. Thematic analysis of Reddit content about buprenorphine-naloxone using manual annotation and natural language processing techniques. J Addict Med 2022;16(4):454-460. [doi: 10.1097/ADM.0000000000000940] [Medline: 34864788]
19. Balsamo D, Bajardi P, Salomone A, Schifanella R. Patterns of routes of administration and drug tampering for nonmedical opioid consumption: data mining and content analysis of Reddit discussions. J Med Internet Res 2021 Jan 4;23(1):e21212. [doi: 10.2196/21212] [Medline: 33393910]
20. Preiss A, Baumgartner P, Edlund MJ, Bobashev GV. Using named entity recognition to identify substances used in the self-medication of opioid withdrawal: natural language processing study of Reddit data. JMIR Form Res 2022 Mar 30;6(3):e33919. [doi: 10.2196/33919] [Medline: 35353047]
21. Chi Y, Chen HY. Investigating substance use via Reddit: systematic scoping review. J Med Internet Res 2023 Oct 25;25:e48905. [doi: 10.2196/48905] [Medline: 37878361]
22. D'Agostino AR, Optican AR, Sowles SJ, Krauss MJ, Escobar Lee K, Cavazos-Rehg PA. Social networking online to recover from opioid use disorder: a study of community interactions. Drug Alcohol Depend 2017 Dec 1;181:5-10. [doi: 10.1016/j.drugalcdep.2017.09.010] [Medline: 29024875]
23. Andy A. Self-disclosure in opioid use recovery forums. In: In International Workshop on Health Intelligence: Springer International Publishing; 2021:65-74. [doi: 10.1007/978-3-030-93080-6_6]
24. Le Glaz A, Haralambous Y, Kim-Dufor DH, et al. Machine learning and natural language processing in mental health: systematic review. J Med Internet Res 2021 May 4;23(5):e15708. [doi: 10.2196/15708] [Medline: 33944788]
25. Coppersmith G, Leary R, Crutchley P, Fine A. Natural language processing of social media as screening for suicide risk. Biomed Inform Insights 2018;10:1178222618792860. [doi: 10.1177/1178222618792860] [Medline: 30158822]
26. Balsamo D, Bajardi P, De Francisci Morales G, Monti C, Schifanella R. The pursuit of peer support for opioid use recovery on Reddit. ICWSM 2023 Jun;17:12-23. [doi: 10.1609/icwsm.v17i1.22122]
27. PRAW 7.8.2.dev0 documentation. PRAW. URL: https://praw.readthedocs.io/en/latest/ [accessed 2024-05-05]
28. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. Presented at: 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; Jun 2-7, 2019; Minneapolis, Minnesota.
29. Barak A, Gluck-Ofri O. Degree and reciprocity of self-disclosure in online forums. Cyberpsychol Behav 2007 Jun;10(3):407-417. [doi: 10.1089/cpb.2006.9938]
30. What is recovery? a working definition from the Betty Ford Institute. J Subst Abuse Treat 2007 Oct;33(3):221-228. [doi: 10.1016/j.jsat.2007.06.001]
31. Ghosh S, Misra J, Ghosh S, Podder S. Utilizing social media for identifying drug addiction and recovery intervention. Presented at: 2020 IEEE International Conference on Big Data (Big Data); Dec 10-13, 2020; Atlanta, GA, USA. [doi: 10.1109/BigData50022.2020.9378092]
32. Engel L, Henderson C, Fergenbaum J, Colantonio A. Medical record review conduction model for improving interrater reliability of abstracting medical-related information. Eval Health Prof 2009 Sep;32(3):281-298. [doi: 10.1177/0163278709338561] [Medline: 19679636]
33. Gwet KL. Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement among Raters: Advanced Analytics, LLC; 2014.
34. Sun C, Qiu X, Xu Y, Huang X. How to fine-tune BERT for text classification. Presented at: Chinese Computational Linguistics CCL 2019 Lecture Notes in Computer Science; Oct 18-20, 2019; Kunming, China. [doi: 10.1007/978-3-030-32381-3_16]

XSL•FO
RenderX

35. Chi Y, Thaker K, He D, et al. Knowledge acquisition and social support in online health communities: analysis of an online ovarian cancer community. JMIR Cancer 2022 Sep 13;8(3):e39643. [doi: 10.2196/39643] [Medline: 36099015]
36. Wang X, High A, Wang X, Zhao K. Predicting users' continued engagement in online health communities from the quantity and quality of received support. Asso for Info Science & Tech 2021 Jun;72(6):710-722. [doi: 10.1002/asi.24436]
37. Wang YC, Kraut R, Levine JM. To stay or leave?: the relationship of emotional and informational support to commitment in online health support groups. Presented at: Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work; Feb 11-15, 2012; Seattle, WA, USA. [doi: 10.1145/2145204.2145329]
38. Chuang KY, Yang CC. Informational support exchanges using different computer‐mediated communication formats in a social media alcoholism community. J Assoc Inf Sci Technol 2014 Jan;65(1):37-52. [doi: 10.1002/asi.22960]
39. HELPeR-codes/post support type prediction. GitHub. URL: https://github.com/HELPeR-codes/Post-Support-Type-Prediction [accessed 2024-05-05]
40. Clark-Gordon CV, Bowman ND, Goodboy AK, Wright A. Anonymity and online self-disclosure: a meta-analysis. Commun Rep 2019 May 4;32(2):98-111. [doi: 10.1080/08934215.2019.1607516]
41. Kelly JF, White WL. Addiction Recovery Management Theory, Research, and Practice: Humana Press; 2011.
42. Lawton MJ. Challenges in recovery from addiction. Alcohol Treat Q 1985 Jan;1(4):55-66. [doi: 10.1300/J020V01N04_02]
43. Pater JA, Phelan C, Cornet VP, et al. User-centered design of a mobile app to support peer recovery in a clinical setting. Proc ACM Hum-Comput Interact 2021 Apr 13;5(CSCW1):1-31. [doi: 10.1145/3449186]
44. Bunting AM, Frank D, Arshonsky J, Bragg MA, Friedman SR, Krawczyk N. Socially-supportive norms and mutual aid of people who use opioids: an analysis of Reddit during the initial COVID-19 pandemic. Drug Alcohol Depend 2021 May 1;222:108672. [doi: 10.1016/j.drugalcdep.2021.108672] [Medline: 33757708]
45. Clendennen SL, Loukas A, Vandewater EA, Perry CL, Wilkinson AV. Exposure and engagement with tobacco-related social media and associations with subsequent tobacco use among young adults: a longitudinal analysis. Drug Alcohol Depend 2020 Aug 1;213:108072. [doi: 10.1016/j.drugalcdep.2020.108072] [Medline: 32563115]
46. Silverstein SM, Daniulaityte R, Miller SC, Martins SS, Carlson RG. On my own terms: motivations for self-treating opioid-use disorder with non-prescribed buprenorphine. Drug Alcohol Depend 2020 May 1;210:107958. [doi: 10.1016/j.drugalcdep.2020.107958] [Medline: 32203863]
47. Meacham MC, Nobles AL, Tompkins DA, Thrul J. "I got a bunch of weed to help me through the withdrawals": naturalistic cannabis use reported in online opioid and opioid recovery community discussion forums. PLOS ONE 2022;17(2):e0263583. [doi: 10.1371/journal.pone.0263583] [Medline: 35134074]
48. ElSherief M, Sumner SA, Jones CM, et al. Characterizing and identifying the prevalence of web-based misinformation relating to medication for opioid use disorder: machine learning approach. J Med Internet Res 2021 Dec 22;23(12):e30753. [doi: 10.2196/30753] [Medline: 34941555]
49. Judd H, Meier CL, Yaugher AC, Campbell E, Atismé-Bevins K. Opioid use disorder stigma reduction through story telling narrative and sharing: a qualitative review and lessons learned. Int J Ment Health Addiction 2023 Feb;21(1):468-483. [doi: 10.1007/s11469-021-00606-y]
50. Paterno MT, Leckenby D, Pachucki MC. Challenges and lessons learned from digital storytelling with parenting women in recovery. Health Promot Pract 2020 May;21(3):340-343. [doi: 10.1177/1524839920902751] [Medline: 32046517]
51. Thaker K, Chi Y, Birkhoff S, et al. Exploring resource-sharing behaviors for finding relevant health resources: analysis of an online ovarian cancer community. JMIR Cancer 2022 Apr 12;8(2):e33110. [doi: 10.2196/33110] [Medline: 35258465]
52. Stanojlović M, Davidson L. Targeting the barriers in the substance use disorder continuum of care with peer recovery support. Subst Abuse 2021;15:1178221820976988. [doi: 10.1177/1178221820976988] [Medline: 34211275]
53. Zou W, Tang L, Zhou M, Zhang X. Self-disclosure and received social support among women experiencing infertility on Reddit: a natural language processing approach. Comput Human Behav 2024 May;154:108159. [doi: 10.1016/j.chb.2024.108159]
54. Catona D, Greene K. Self‐disclosure. Int Encycl Interpersonal Commun 2015:1-5. [doi: 10.1002/9781118540190]
55. Taylor DA, Altman I. Communication in interpersonalrelationships: social penetration processes. In: Roloff ME, Miller GR, editors. Interpersonal Processes: New Directions in Communication Research: Sage Publication; 1987:257-277.
56. Forgas JP. Affective influences on self-disclosure: mood effects on the intimacy and reciprocity of disclosing personal information. J Pers Soc Psychol 2011 Mar;100(3):449-461. [doi: 10.1037/a0021129] [Medline: 21171788]
57. Russell AM, Valdez D, Wang M, et al. Content analysis of substance use disorder recovery discourse on Twitter: from personal recovery narratives to marketing of addiction treatment. Alcohol, Clinical and Experimental Research 2025 Mar;49(3):629-640. [doi: 10.1111/acer.15531]
58. Tofighi B, El Shahawy O, Segoshi A, et al. Assessing perceptions about medications for opioid use disorder and naloxone on Twitter. J Addict Dis 2021;39(1):37-45. [doi: 10.1080/10550887.2020.1811456] [Medline: 32835641]
59. Ellway D, Reilly R, Le Couteur A, Ward J. Exploring how people affected by methamphetamine exchange social support through online interactions on Facebook: content analysis. JMIR Ment Health 2019 Oct 1;6(10):e14011. [doi: 10.2196/14011] [Medline: 31573926]
60. Hartmann J, Heitmann M, Siebert C, Schamp C. More than a feeling: accuracy and application of sentiment analysis. Int J Res Mark 2023 Mar;40(1):75-87. [doi: 10.1016/j.ijresmar.2022.05.005]

XSL•FO
RenderX

61. Sharma NA, Ali A, Kabir MA. A review of sentiment analysis: tasks, applications, and deep learning techniques. Int J Data Sci Anal 2025 Apr;19(3):351-388. [doi: 10.1007/s41060-024-00594-x]

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
NLP: natural language processing
OUD: opioid use disorder
RQ: research question
SPT: social penetration theory

Edited by J Sarvestan; submitted 12.01.25; peer-reviewed by M Elbattah, X Ding; revised version received 05.06.25; accepted 06.06.25; published 17.07.25.

Please cite as:
Chi Y, Chen HY, Thaker K
Self-Disclosure and Social Support in a Web-Based Opioid Recovery Community: Machine Learning Analysis
JMIR Form Res 2025;9:e71207
URL: https://formative.jmir.org/2025/1/e71207
doi:10.2196/71207

© Yu Chi, Huai-yu Chen, Khushboo Thaker. Originally published in JMIR Formative Research (https://formative.jmir.org), 17.7.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Digital Media Coverage of Respiratory Syncytial Virus-Related News in India: Mixed Methods Content Analysis of Disease Burden and Intervention

Rhythm Hora[1*], BDS, MPH; Arindam Ray[2*], MBBS, MD; Amrita Kumari[2*], MBBS, MD; Rashmi Mehra[1*], BDS, MPH, MSc; Amanjot Kaur[1*], BDS, MBA; Syed F Quadri[1], MBBS, MPH; Bodhisatwa Ray[3*], MSc; Seema Singh Koshal[1*], MPH; Shyam Kumar Singh[1], MCA; Abida Sultana[1*], BHMS, MPH; Arup Deb Roy[1*], MBBS, MD

[1]John Snow India, Plot No.5 & 6, LSC Shopping Complex, Nelson Mandela Marg, Vasant Kunj, New Delhi, India

[2]Gates Foundation, Seattle, WA, United States

[3]University of Exeter, Exeter, United Kingdom

[*]these authors contributed equally

Corresponding Author:
Rhythm Hora, BDS, MPH
John Snow India, Plot No.5 & 6, LSC Shopping Complex, Nelson Mandela Marg, Vasant Kunj, New Delhi, India

Abstract

Background: Respiratory syncytial virus (RSV) is a leading cause of lower respiratory tract infections in children younger than 5 years of age. Given the high morbidity and mortality associated with RSV in India, the introduction of a vaccine against RSV will potentially reduce the disease's burden. However, vaccine acceptance is influenced by public perception, which is shaped by information disseminated through media sources. This study aims to explore the landscape of RSV–related news coverage in India's digital media.

Objective: This study aims to conduct a comprehensive content analysis to explore the landscape of RSV–related news coverage in India's digital media.

Methods: Media content analysis was retrospectively conducted by a digital search for all related news pieces in the trustworthy brands of 4 trusted newspapers (Hindustan Times, The Hindu, The Indian Express, and The Times of India) and 3 news channel websites (India Today, NDTV news, and News 18), between November 1, 2022, and October 31, 2023. A total of 58 news pieces were retrieved using selected keywords, with inclusion criteria encompassing English-language news pieces with RSV–specific content. Two reviewers compiled, coded, and analyzed the content. Quantitative data were analyzed descriptively, while qualitative content analysis assessed the emotional tone and sentiment of the pieces.

Results: The findings revealed significant digital media coverage on RSV infection and the potential vaccines. The majority of news pieces (53/58, 91%) discussed RSV signs and symptoms, with 64% (37/58) addressing the disease severity and 36% (21/58) highlighting its seasonal surge. However, only 5% (3/58) focused on diagnostic aids. Additionally, 41% (24/58) of news pieces discussed RSV in the context of COVID-19. Regarding the vaccine, 29% (17/58) of news pieces mentioned it, with 26% (15/58) highlighting manufacturers such as Pfizer and GlaxoSmithKline (GSK). Positive sentiment was found in 35% (20/58) of news pieces, while 43% (25/58) exhibited negative sentiment, often related to the disease burden and severity. Emotional tone analysis revealed that 74% (43/58) of news pieces contained emotional elements, with 58% (25/43) expressing negative emotions (eg, concern and anxiety), particularly about hospitalizations and deaths. In contrast, a positive tone was emulated in the frequent mentions of the RSV vaccines as safe, effective, and approved.

Conclusions: The analysis revealed significant coverage of RSV–related news in India's digital media, with a focus on disease severity and hospitalizations. While positive sentiment was expressed in coverage of the RSV vaccine, negative sentiments dominated discussions on the disease burden. However, considering the limited number of news pieces, the study highlights the need for improved media coverage to raise awareness about the disease and its preventive strategies. Further research should explore the implications of the overlap between RSV and COVID-19 in media coverage and the limited focus on RSV diagnostics, with a focus on understanding how these factors impact public health outcomes.

(JMIR Form Res 2025;9:e70322) doi:10.2196/70322

KEYWORDS

vaccine content analysis; respiratory syncytial virus; RSV; respiratory syncytial virus vaccine; digital media; public health

XSL•FO
RenderX

Introduction

Respiratory syncytial virus (RSV) stands as a leading cause of lower respiratory tract infections in children younger than 5 years of age, posing a significant public health challenge worldwide [1,2]. The estimates for 2019 show 33 million episodes of RSV-associated lower respiratory infection in this vulnerable age group each year worldwide, with a striking 95% of these cases occurring in lower-middle-income countries [3]. This underscores the disproportionate burden faced by these regions in managing RSV-related health challenges.

In India alone, the situation is terrible, with the National Health Portal report (2019) stating 41.9 million cases of respiratory infections in 2018 [4]. These infections not only led to a significant number of hospitalizations but also contributed to 0.003 million fatalities during that year [4].

To control the growing burden of RSV, adopting preventive measures such as vaccination is sought to be the most promising and cost-effective public health intervention [5]. RSV vaccines can significantly reduce the incidence and severity of infectious diseases, preventing hospitalizations, reducing health care costs, and saving lives [6]. However, the success of these vaccination programs goes beyond the mere availability of vaccines and hinges significantly on community acceptance, which is greatly influenced by the information disseminated [7]. Moreover, the community perception is largely influenced by the various sources of information the individuals come across, such as health care providers, print media, digital media, and social media [5,8,9]. Recent research indicates the significant role of media sources in shaping or influencing people's perceptions and attitudes toward public health services [5,10,11]. Concurrently, digital media, including digital news platforms, have emerged as prime avenues for the rapid and widespread dissemination of information related to diseases or infections and their preventive measures [11-13]. Given this critical role, it is imperative or critical to assess how RSV-related news is presented in the digital landscape, particularly in a country like India where media consumption patterns are rapidly evolving [14-17]. This context underscores the critical need for a comprehensive analysis of news coverage related to RSV in India's digital media. Therefore, this study aims to conduct an in-depth content analysis to assess the portrayal of RSV-related information in news coverage across digital media platforms in India.

Methods

The study was done in accordance with the SRQR (Standards for Reporting Qualitative Research) guidelines (Checklist 1) [18,19].

Study Design

Media content analysis was retrospectively conducted by digital search for news pieces published on the leading newspapers' websites and news channel websites appearing over 1 year (November 1, 2022, to October 31, 2023). National newspapers and news channel websites were included in this study due to their wide audience reach, credibility, consistency, and ability to effectively capture the dominant narratives and trends that influence public opinion. Furthermore, the 1-year time frame was selected to capture relevant trends and events, ensuring a comprehensive analysis of current developments. Additionally, this period offered a balanced perspective on seasonal variations and emerging patterns in RSV-related news coverage within the digital media landscape.

Data Extraction Strategy

This study used conceptual content analysis of the most trustworthy brands of national newspapers and news channel websites [20,21]. Based on the brand trust scores and SCImago media rankings, 4 trusted national newspapers (Hindustan Times, The Hindu, The Indian Express, and The Times of India) and 3 trusted national news channel websites (India Today, NDTV News, and News 18) in cyberspace were selected [20,21]. A standardized extraction method using the selected keywords was prepared for retrieving the desired information. The dataset included all the news pieces retrieved after entering the selected keywords: "RSV" or "RSV vaccine," and "RSV" and "vaccine." Only news pieces containing these exact words were extracted. The search was limited only to news pieces in English with all RSV-specific content, whether in the Indian or global context. Focusing on mainstream media in a dominant language (eg, English) was chosen to ensure consistency in content, audience, and reach to obtain more reliable conclusions. This resulted in the retrieval of 62 news pieces from the newspapers and news channel websites included in the study. News pieces were excluded if they were duplicates (news pieces published around the same time, with the same headline and text in the same publication). The news pieces in the vernacular language were excluded from the study due to the potential for inaccuracies or misinterpretations arising from the researchers' limited understanding of regional languages. Besides, no editorials or feature stories were included in the study, as they are subjective opinion pieces that could skew the findings, which are focused on objective reporting. Furthermore, health blogs and independent media were not included in the study to maintain a focus on reliable, widely recognized sources that influence mainstream discourse. After excluding the irrelevant news pieces, 58 news pieces were yielded for analysis.

Data Analysis

To begin with, the news pieces were numbered for ease of reference. Later, the data were analyzed in two steps: first, a quantitative analysis of the news pieces' characteristics and the nature of the information presented, including the sentimental and emotional tonality. Later, this was followed by the qualitative analysis to further account for the sentimental and emotional content in the news pieces.

For the quantitative analysis, a codebook was developed through an iterative process by reviewing a few arbitrarily selected news pieces to identify predominant themes and generate codes. In total, 10% (n=6) of the news pieces were coded independently by two skilled coders (RH and RM) using NVivo (version 14; Lumivero) qualitative data analysis software. Intercoder reliability was measured to ensure agreement between the coders while coding the same dataset. The reliability criterion set for this study was κ _0.60, which is considered ample according

XSL•FO
RenderX

to recommended sources [22]. Disagreements were discussed with another author (AKaur), and consensus was reached. The codebook generated captured information on 4 main categories: information on RSV (including diagnostics, seasonal surge, and co-mention with COVID-19), information on RSV vaccine (including the manufacturers, potential recipients, approval status, and concerns around efficacy or safety), sentiment in title or content, and emotional tone. The data were statistically analyzed using SPSS (version 21.0; IBM Corp) and results were expressed in numbers and percentages.

For qualitative analysis, in-depth content analysis of the news pieces was done to thematically explore sentiment analysis (negative, neutral, or positive) and emotion analysis (record the emotional tone). In this context, negative sentiment refers to expressions or views that convey unfavorable attitudes, disapproval, sadness, or dissatisfaction toward a subject. On the other hand, positive sentiment refers to expressions or views that convey favorable attitudes, approval, happiness, or satisfaction toward the subject. Additionally, emotional tone conveys the overall emotive quality expressed in a piece of text, reflecting the author's or speaker's feelings toward the subject matter. Emotional tone was assessed using methods like the NRC Emotion Lexicon, which associates English words with 8 core emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust). For instance, words such as "concern" or "anxiety" were associated with emotions like fear or sadness, while terms like "hope" or "joy" were connected to positive emotions [23].

This analysis used the initially generated codebook to code all news pieces using NVivo. The coding was primarily done to delineate content specific to the sentiment conveyed in the title or content of all news pieces. Coded data were then read and reread to arrive at a consensus.

Ethical Considerations

No ethical approval was sought for the study, as it used publicly available data of a nonidentifiable nature.

Results

A total of 58 news pieces were included in the analysis. Our findings underscore the various media sources (the media source and the news source), the nature of the information presented, and the emotional tones conveyed in the reporting.

Characteristics of News Stories

Media Source

Of the 58 news pieces included in the analysis, 37 (64%) were identified from the leading national newspapers and 21 (36%) from the leading news channel websites selected for the study (Figure 1).

Figure 1. Comprehensive overview of the sources of retrieved news pieces in the digital media landscape (N=58).

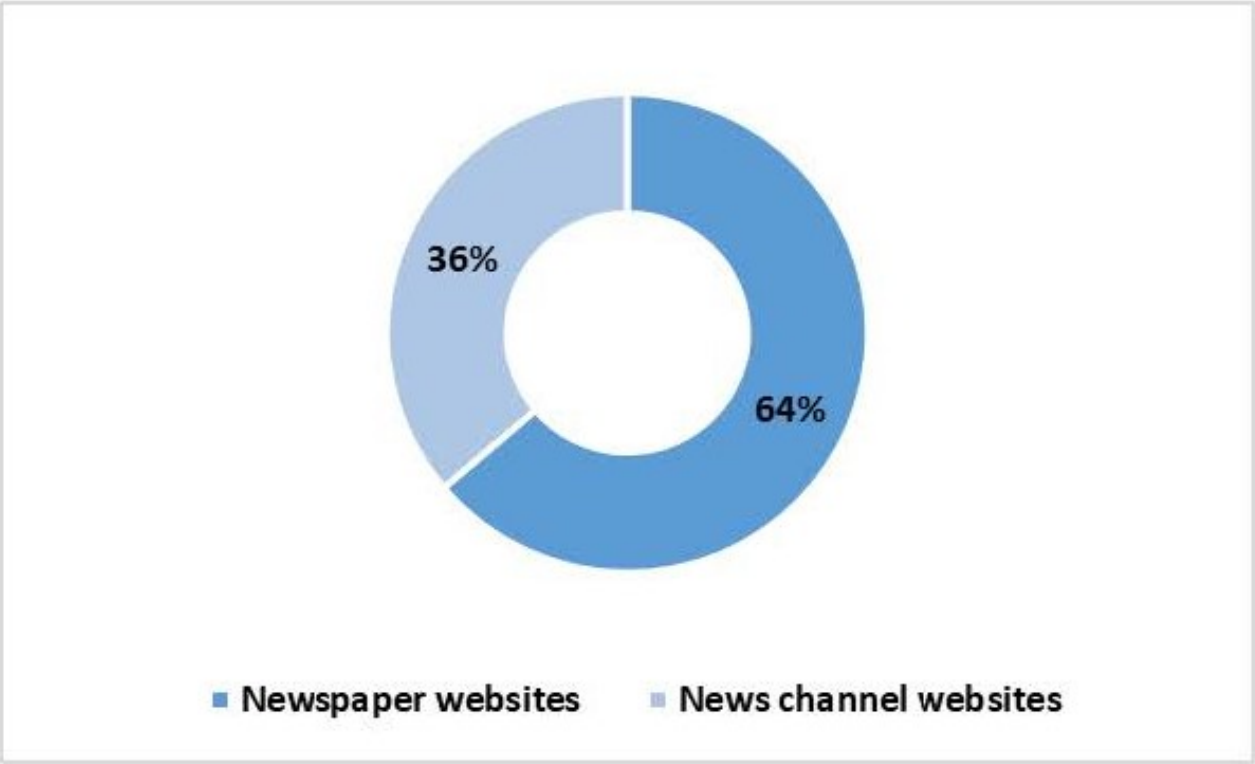

News Source

The news pieces (N=58) relied on various sources (Table 1); however, in almost one-third, that is, 20/58 (35%), the news source was the news or web desk (Hindu Bureau, Times News Network (TNN), and Express-News). The other most common news source was Reuters (n=9, 15%).

XSL•FO
RenderX

Table . Description of the identified news sources featured in the retrieved digital media news (N=58).

News source	Frequency (N=58), n (%)
Press Trust of India (PTI)	4 (7)
Reuters	9 (15)
Agence France-Presse (AFP)	5 (9)
Asian News International (ANI)	2 (3)
Associated Press (AP)	4 (7)
News desk or web desk (Hindu bureau, TNN[a], and Express News)	20 (35)
Others: Individual person name or lifestyle desk or local circles	14 (24)

[a]TNN: Times News Network

Nature of Information in News Pieces

Information on RSV

Of the total news pieces (N=58) retrieved, 53 (91%) news pieces had a detailed description of the signs and symptoms of RSV infection, with 37 (64%) mentioning the severity of the infection and 21 (36%) discussing the seasonal surge of the RSV infection (Table 2). Besides, 24 (41%) news pieces reporting on the COVID-19 infection also discussed the RSV infection. In total, 1 (2%) news piece stated the modes of transmission of RSV infection, while 12 (21%) mentioned the potential complications and the modes of prevention of RSV infection. Only 3 (5%) news pieces discussed the diagnostic aids available for detecting RSV infection.

XSL•FO
RenderX

Table . News stories on RSV[a] infection and RSV vaccine featured in the digital media landscape.

Categories and codes identified	Frequency (N=58), n (%)
RSV	
Signs and symptoms of disease	53 (91)
RSV burden (epidemiology)	10 (17)
Modes of transmission	14 (24)
Severity of disease	37 (64)
Potential complications	12 (21)
Modes of prevention	12 (21)
RSV diagnostics	3 (5)
Seasonal RSV surge	21 (36)
RSV and COVID-19	24 (41)
RSV vaccine	
Overview about vaccine	17 (29)
Vaccine products and manufacturers	15 (26)
Vaccine approval status	8 (14)
Vaccine trials	5 (9)
Potential recipients of vaccine	15 (25)
Vaccine efficacy	9 (15)
Vaccine safety	6 (10)
Cost or price	2 (3)
Sentiment in title or content	
Positive	20 (35)
Neutral	13 (22)
Negative	25 (43)
Emotional tone	
Yes	43 (74)
No	15 (26)

[a]RSV: respiratory syncytial virus.

Information on RSV Vaccine

Out of all the news pieces (N=58) included in the study, 17 (29%) pieces had a mention of the RSV vaccine. In total, 15/58 (26%) news pieces had a clear mention of the globally available vaccine products and their manufacturers (Table 1) with 7 (47%) referencing only the Pfizer vaccine product (Abrysvo) and 5 (33%) of the GSK vaccine product (Arexvy), and 1 mentioned AstraZeneca and Sanofi's product (Nirsevimab). Only 2 news pieces reported on both the leading vaccine manufacturers (Pfizer and GSK) and their products, and that they competed with each other.

Furthermore, 15 (26%) news pieces specifically mentioned the potential recipients of the RSV vaccine. Of the 15 news pieces, 9 (60%) referenced older adults aged 60 years and older as potential recipients, while 6 (40%) news pieces mentioned pregnant women as potential recipients (Table 2).

Besides, 9 (15%) and 6 (10%) news pieces discussed the RSV vaccine efficacy and safety, respectively. Only 2 (3%) new pieces covered the cost/ price of the RSV vaccine.

Sentiment in Title or Content

Most (n=25, 43%) of the news pieces had a negative sentiment in the title or content. However, 20 (35%) news pieces had a positive and 13 (22%) had a neutral sentiment in the title or content (Table 2). Furthermore, in all 17 news pieces (Table 2) that reported the RSV vaccine, there was a positive sentiment in the title and the content.

Emotional Tone

The emotional tone was also analyzed for all the news pieces (n=58). Nearly 43 of 58 (74%) of the news pieces contained emotional elements either around the RSV infection or the RSV vaccine. However, only 25 of 43 (58%) news pieces stated negative emotions that were entirely related to RSV disease (Table 2).

XSL•FO
RenderX

Qualitative Content Analysis

Sentiment in Title or Content

Negative Sentiments Toward the Vaccine

Among 25 of 58 (43%) news pieces with negative sentiments, the most recurring negative sentiment was the increased burden of respiratory illnesses and RSV infection, with its seasonal surge found in 21 of 25 news pieces (Table 2). For example, the medical director of Bharati Vidyapeeth Medical Hospital and Research Center said

> *Viruses like influenza and parainfluenza are now the dominant ones. Both in children and adults, we are witnessing an increasing incidence of respiratory viral illnesses, as the humidity levels have been quite high. The admission pattern across hospitals shows that ICU admissions have increased by nearly 25% - 30%.* [News piece 27]

The other negative theme that emerged was the severity of the infection, resulting in hospitalizations and deaths. For example

> *"This was the hardest week of my life," wrote American standup comedian and actor Amy Schumer on Instagram when her three-year-old son was admitted to the hospital, adding that her son was rushed to ER and admitted for RSV.* [News piece 42]

The negative theme related to the RSV vaccine was a limitation in vaccine introduction and implementation, owing to the cost of the vaccine (2 news pieces). News piece 52 mentions that "the vaccine would be priced above $120 per shot, provided, expected soon, shows that it offers protection for two RSV seasons."

Positive Sentiments Toward the Vaccine

The most recurring positive theme was the approval of vaccines (17 of the 20 news pieces with positive sentiment) to prevent RSV-related disease burden. Almost all these news pieces assured that the approved vaccines are safe and effective, with common and mild side effects. Referring to the declaration by the medical director of the National Foundation for Infectious Diseases, news piece 29 mentioned, "This is a great first step ... to protect older persons from serious RSV disease."

Another positive theme (3 of 20 news pieces) was continued screening and diagnostics for RSV and other viral or flu packages. For example, "Tamil Nadu to continue fever screening!" is mentioned in the title of news piece 31.

Emotional Tone

Of the total news pieces (n=43) with emotions included, negative emotions were reflected in 25 (58%) news pieces. The negative emotions expressed included adjectives such as concern, anxiety, agony, and grief due to disease burden, hospitalizations, severity of infection, and deaths. These negative emotive patterns were chiefly around the RSV infection only. For example, citing the statement of a pediatrician, one news piece mentioned:

> *We are seeing a significant number of pediatric cases in the outpatient department, testing positive for RSV. While RSV has the same symptoms as influenza, including cold, cough, and fever in infants, wheezing is also seen in newborns. Some are coming with bronchiolitis and have to be put on respiratory support and ICU care.* [News piece 24]

Discussion

Principal Findings and Interpretation

The burden of RSV infection in the younger than 5 years age group worldwide necessitates the need to unveil increasing concerns around a booming public health challenge. As digital media plays a crucial role in communicating public health issues and policy information, this study explored the news content related to RSV in the trusted brands of newspapers and news channels in the digital environment.

The study findings revealed that digital news media have provided quite a reasonable amount of information on the signs and symptoms, modes of transmission and prevention, and severity of RSV infection. This is crucial for generating public awareness and could potentially drive early recognition and treatment [24]. However, less information on diagnostic aids (n=3, 5%) in the news pieces reflects a notable gap in comprehensive health communication and management. Previous literature also highlights that a thorough understanding of information on diagnostics makes a ground for bureaucratic interactions and influences public policy formulations [25,26]. Further, the overlap of RSV and COVID-19 across news pieces (n=24, 41%) suggests that the COVID-19 pandemic continues to influence the narratives surrounding respiratory illnesses. Discussions in the existing literature emphasize that other respiratory diseases should not be overshadowed, in light of the COVID-19 pandemic, as the consequences could be dramatic [27].

This study also found a decent amount of information on vaccines available to prevent RSV infection, as a promising strategy to alleviate RSV-related morbidity and mortality. Additionally, the news pieces have discussed the vaccine manufacturers and potential recipients of the vaccine, including the approval status of the vaccines. Besides, there have been deliberations in the news pieces regarding the efficacy or safety of the vaccine. A content analysis of newspaper coverage of the COVID-19 pandemic in 2021 demonstrated that updates on the progress of vaccines, in general, might often provide better hope to the people and bring them psychological harmony [28].

Furthermore, our analysis showed that most news coverage expressed negative opinions regarding the burden of respiratory illnesses, including RSV infections and associated hospitalizations. This aligns with previous research indicating that media reporting often emphasizes negative aspects of public health crises, which can heighten public anxiety [29,30]. Only 2 news pieces discussed the cost of the RSV vaccine, which was a limitation in vaccine introduction and implementation. Moreover, negative emotions elicited by news pieces, such as personal narratives from affected families, amplify the emotional tone. This is consistent with other studies demonstrating that people often respond more strongly to the negative elements in their environment, reflecting the media's power to shape public perception [28,31]. Besides, the approval of safe and efficacious

XSL•FO
RenderX

RSV vaccines, which may be viewed as a beacon of optimism, contributed to the positive tone of all news pieces reporting on the RSV vaccine. Previous studies mentioned that media narratives can drive policy-level solutions and influence public willingness to accept the vaccine [28,32].

Limitations and Recommendations

The study also has some limitations. Although the news pieces over the past 1 year had RSV coverage, considering the time frame included in the study, the number of news pieces retrieved was limited. Furthermore, this analysis focused on news pieces but did not include editorials or feature stories that may have contained additional information. Besides, news pieces in the vernacular language were not explored for this study. Another potential limitation of the study was that the data source was not exhaustive. Thus, there may be relative under or overestimation of data as the study only selected newspapers and news channel websites for data collection for practical reasons, while other forms of media, including digital media (including social media), electronic media (including television, radio, and cinema), and magazines, were not used.

This study did not examine the implications of the overlap between RSV and COVID-19 in media coverage, nor the limited focus on RSV diagnostics. Future research to explore how this overlap may cause confusion, affect risk perceptions, and influence health priorities, as well as how the gap in coverage of RSV diagnostics impacts public awareness and health outcomes, is sought. Furthermore, the study recommends that policy makers enhance public health communication and encourage media outlets to rely on credible health sources for clearer, more consistent reporting, especially during overlapping health crises, to improve public understanding and reduce confusion.

Conclusions

This study sought to analyze the portrayal of RSV-related news within the digital media landscape in India. The study highlighted a substantial presence of RSV and RSV-related information in digital media. Content analysis of screened news pieces revealed that negative sentiments were often found in news pieces capturing the burden of respiratory illnesses, including RSV infections and associated hospitalizations. On the other hand, there was a consistent positive emotion in the news pieces around the approval of safe and effective vaccines to prevent RSV-related disease burden. As the media plays a crucial role in communicating public health and policy information, the study elicits the need for improved media coverage. This will aid in communicating appropriate messages and generating awareness about the disease, including its preventive strategies. In addition, the study highlights the need for improved public health communication by policy makers, particularly during overlapping health crises. Furthermore, a larger study can be undertaken to include news in the vernacular languages, especially in the Indian context. Future research to explore the implications of the overlap between RSV and COVID-19, along with the gap in coverage of RSV diagnostics on public health outcomes, is recommended.

Acknowledgments

Funding for this study was provided by the Gates Foundation under grant INV-030655.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

RH conceptualized the study, designed the methodology, analyzed the findings, and drafted and edited the manuscript. BR, AKumari, and AS supported the development of the manuscript throughout. SFQ, SKS, and SSK verified the analysis. RM and AKaur reviewed and edited the manuscript. ADR and AR supervised and guided the study from inception to culmination. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Checklist 1
SRQR (Standards for Reporting Qualitative Research) checklist.
[DOCX File, 20 KB - formative_v9i1e70322_app1.docx]

References

1. Driscoll AJ, Arshad SH, Bont L, et al. Does respiratory syncytial virus lower respiratory illness in early life cause recurrent wheeze of early childhood and asthma? Critical review of the evidence and guidance for future studies from a World Health Organization-sponsored meeting. Vaccine (Auckl) 2020 Mar 4;38(11):2435-2448. [doi: 10.1016/j.vaccine.2020.01.020] [Medline: 31974017]
2. Munro APS, Martinón-Torres F, Drysdale SB, Faust SN. The disease burden of respiratory syncytial virus in Infants. Curr Opin Infect Dis 2023;36(5):379-384. [doi: 10.1097/QCO.0000000000000952]

XSL•FO
RenderX

3. Li Y, Wang X, Blau DM, et al. Global, regional, and national disease burden estimates of acute lower respiratory infections due to respiratory syncytial virus in children younger than 5 years in 2019: a systematic analysis. Lancet 2022 May 28;399(10340):2047-2064. [doi: 10.1016/S0140-6736(22)00478-0] [Medline: 35598608]
4. Waghmode R, Jadhav S, Nema V. The burden of respiratory viruses and their prevalence in different geographical regions of India: 1970-2020. Front Microbiol 2021;12:723850. [doi: 10.3389/fmicb.2021.723850] [Medline: 34531842]
5. Roy DN, Huda MN, Azam MS. Factors influencing COVID-19 vaccine acceptance and hesitancy among rural community in Bangladesh: a cross-sectional survey based study. Hum Vaccin Immunother 2022 Nov 30;18(5):2064685. [doi: 10.1080/21645515.2022.2064685] [Medline: 35580319]
6. Andreoni M, Bonanni P, Gabutti G, Maggi S, Siliquini R, Ungar A. RSV vaccination as the optimal prevention strategy for older adults. Infez Med 2024;32(4):478-488. [doi: 10.53854/liim-3204-6] [Medline: 39660164]
7. Dhaliwal BK, Chandrashekhar R, Rattani A, et al. Community perceptions of vaccination among influential stakeholders: qualitative research in rural India. BMC Public Health 2021 Nov 18;21(1):2122. [doi: 10.1186/s12889-021-12188-4] [Medline: 34794415]
8. Kumar D, Chandra R, Mathur M, Samdariya S, Kapoor N. Vaccine hesitancy: understanding better to address better. Isr J Health Policy Res 2016;5(1):2. [doi: 10.1186/s13584-016-0062-y] [Medline: 26839681]
9. Wu X, Li Z, Xu L, Li P, Liu M, Huang C. COVID-19 vaccine-related information on the WeChat public platform: topic modeling and content analysis. J Med Internet Res 2023 Apr 14;25(1):e45051. [doi: 10.2196/45051] [Medline: 37058349]
10. Anwar A, Malik M, Raees V, Anwar A. Role of mass media and public health communications in the COVID-19 pandemic. Cureus 2020 Sep 14;12(9):e10453. [doi: 10.7759/cureus.10453] [Medline: 33072461]
11. Recio-Román A, Recio-Menéndez M, Román-González MV. Influence of media information sources on vaccine uptake: the full and inconsistent mediating role of vaccine hesitancy. Computation 2023 Oct;11(10):208. [doi: 10.3390/computation11100208]
12. Anwar A, Malik M, Raees V, Anwar M, Anwar A. Role of mass media and public health communications in COVID-19 vaccination. MRAJ 2023;11(2). [doi: 10.18103/mra.v11i2.3544]
13. Fernández PP, Aríztegui MM. COVID-19 and social media. In: COVID-19 Pandemic: Elsevier; 2022:85-97. [doi: 10.1016/B978-0-323-82860-4.00009-4]
14. Gupta M, Keshri VR, Konwar P, Cox KL, Jagnoor AJ. Media coverage of COVID-19 health information in India: a content analysis. Health Promot Int 2022 Apr 29;37(2):daab116. [doi: 10.1093/heapro/daab116] [Medline: 34297832]
15. Jeyaraman M, Ramasubramanian S, Kumar S, et al. Multifaceted role of social media in healthcare: opportunities, challenges, and the need for quality control. Cureus 2023 May;15(5):e39111. [doi: 10.7759/cureus.39111] [Medline: 37332420]
16. Melton CA, White BM, Davis RL, Bednarczyk RA, Shaban-Nejad A. Fine-tuned sentiment analysis of COVID-19 vaccine-related social media data: comparative study. J Med Internet Res 2022 Oct 17;24(10):e40408. [doi: 10.2196/40408] [Medline: 36174192]
17. Ngai CSB, Singh RG, Lu W, Koon AC. Grappling with the COVID-19 health crisis: content analysis of communication strategies and their effects on public engagement on social media. J Med Internet Res 2020 Aug 24;22(8):e21360. [doi: 10.2196/21360] [Medline: 32750013]
18. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014 Sep;89(9):1245-1251. [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]
19. Ha NT, Chi NTY, Van Nuil J, Thwaites L, Chambers M. Media portrayal of vaccine: a content analysis of Vietnam online news about a pentavalent vaccine in the Expanded Program of Immunization. Wellcome Open Res 2022;7:271. [doi: 10.12688/wellcomeopenres.18457.1] [Medline: 39220241]
20. Newman N, Fletcher R, Eddy, K. Reuters Institute digital news report 2023. : Reuters Institute for the Study of Journalism; 2023. [doi: 10.60625/RISJ-P6ES-HB13]
21. Media rankings. SCImago Media Rankings. URL: https://www.scimagomedia.com/rankings.php?country=India&edition=2023_03 [accessed 2024-12-08]
22. Habel MA, Liddon N, Stryker JE. The HPV vaccine: a content analysis of online news stories. J Womens Health (Larchmt) 2009 Mar;18(3):401-407. [doi: 10.1089/jwh.2008.0920]
23. Park S, Suh YK. A comprehensive analysis of COVID-19 vaccine discourse by vaccine brand on Twitter in Korea: Topic and sentiment analysis. J Med Internet Res 2023 Jan 31;25(1):e42623. [doi: 10.2196/42623] [Medline: 36603153]
24. Parvin GA, Ahsan R, Rahman M, Abedin M. Novel coronavirus (COVID-19) pandemic: the role of printing media in Asian countries. Front Commun 2020 Nov;5. [doi: 10.3389/fcomm.2020.557593]
25. Balogh EP. The diagnostic process. In: Improving Diagnosis in Health Care: National Academies Press (US); 2015. [doi: 10.17226/21794]
26. Rosenberg CE. The tyranny of diagnosis: specific entities and individual experience. Milbank Q 2002;80(2):237-260. [doi: 10.1111/1468-0009.t01-1-00003] [Medline: 12101872]
27. Di Mattia G, Nenna R, Mancino E, et al. During the COVID-19 pandemic where has respiratory syncytial virus gone? Pediatr Pulmonol 2021 Oct;56(10):3106-3109. [doi: 10.1002/ppul.25582] [Medline: 34273135]

XSL•FO
RenderX

28. Zafri NM, Afroj S, Nafi IM, Hasan MMU. A content analysis of newspaper coverage of COVID-19 pandemic for developing a pandemic management framework. Heliyon 2021 Mar;7(3):e06544. [doi: 10.1016/j.heliyon.2021.e06544] [Medline: 33758786]
29. Altheide DL. The news media, the problem frame, and the production of fear. Sociol Q 1997 Sep 1;38(4):647-668. [doi: 10.1111/j.1533-8525.1997.tb00758.x]
30. Garfin DR, Silver RC, Holman EA. The novel coronavirus (COVID-2019) outbreak: amplification of public health consequences by media exposure. Health Psychol 2020 May;39(5):355-357. [doi: 10.1037/hea0000875] [Medline: 32202824]
31. Feng GC. Effects of narratives and information valence in digital headlines on user responses. Asian J Commun 2024 Mar 3;34(2):156-177. [doi: 10.1080/01292986.2024.2317308]
32. Oh SH, Lee CJ, Park A. Trust matters: the effects of social media use on the public's health policy support through (mis)beliefs in the context of HPV vaccination. Health Commun 2023 Dec;38(12):2628-2639. [doi: 10.1080/10410236.2022.2096985] [Medline: 35850554]

Abbreviations

RSV: respiratory syncytial virus
SRQR: Standards for Reporting Qualitative Research

Edited by A Mavragani; submitted 19.12.24; peer-reviewed by G Gourzoulidis; revised version received 05.04.25; accepted 23.04.25; published 05.08.25.

Please cite as:
Hora R, Ray A, Kumari A, Mehra R, Kaur A, F Quadri S, Ray B, Singh Koshal S, Kumar Singh S, Sultana A, Deb Roy A
Digital Media Coverage of Respiratory Syncytial Virus-Related News in India: Mixed Methods Content Analysis of Disease Burden and Intervention
JMIR Form Res 2025;9:e70322
URL: https://formative.jmir.org/2025/1/e70322
doi:10.2196/70322

© Rhythm Hora, Arindam Ray, Amrita Kumari, Rashmi Mehra, Amanjot Kaur, Syed F Quadri, Bodhisatwa Ray, Seema Singh Koshal, Shyam Kumar Singh, Abida Sultana, Arup Deb Roy. Originally published in JMIR Formative Research (https://formative.jmir.org), 5.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Analyzing the Contents of a Large, Public Online Peer Support Forum for Obsessive-Compulsive Disorder: Thematic Analysis

Nora Yanyi Sun[1]; Christopher Pittenger[2], MD, PhD; Terence Ching[2], PhD

[1]Harvard University, 58 Plympton St, Cambridge, MA, United States

[2]Department of Psychiatry, Yale University School of Medicine, New Haven, CT, United States

Corresponding Author:
Nora Yanyi Sun
Harvard University, 58 Plympton St, Cambridge, MA, United States

Abstract

Background: Obsessive-compulsive disorder (OCD) is a debilitating chronic anxiety disorder with low rates of remission. r/OCD is a peer support forum hosted by the Reddit website with over 180,000 users and 100 - 200 new posts daily. While peer support has been associated with improved treatment adherence and outcomes, online forums can also be an outlet for performing and accommodating compulsions (eg, seeking and receiving reassurance), which can ultimately exacerbate OCD.

Objective: This study aims to inductively assess what types of content are posted on r/OCD, an online peer support forum for individuals with OCD, to better understand the prevalent topics, dynamics, and norms of such online communities.

Methods: To describe the explicit content of the forum, 132 new posts with a total of 739 comments added to the subreddit within a 24-hour period from August 4, 2023, to August 5, 2023, were manually downloaded and coded using inductive thematic analysis. Though posts beyond the first 24 hours were downloaded, saturation of ideas was observed after the first 110 posts, and thus, no posts after the first 24 hours were included in the analysis. Through inductive thematic analysis, codes were organized into overarching themes.

Results: Four main themes emerged during analysis: (1) validating OCD symptoms (n=69 total posts and comments), (2) connecting with peers with similar OCD themes (n=512), (3) coping with symptoms (n=280), and (4) sharing treatment experiences and advice (n=118). Many posts involved users describing their symptoms, questioning if a particular symptom was OCD, and asking other users if they had similar experiences. Users frequently categorized their OCD into subtypes using abbreviations (eg, “ROCD” [relationship OCD]) to seek others with similar experiences. A minority of posts solicited and provided advice on therapy and medication. Users were supportive and encouraging of each other’s recovery journeys.

Conclusions: Online peer support forum users demonstrated substantial knowledge of OCD and treatments; these forums enable users to receive social support and exchange helpful information and peer experiences related to seeking medical treatment. However, many users use the platform to connect with others experiencing similar subtypes of OCD or to seek confirmation that their symptoms are OCD, which is promoted by community norms and may suggest they are reassurance-seeking. Close moderation by health care professionals and restricting detailed sharing of OCD symptoms to prevent reassurance-seeking may be important for ensuring a net positive impact of online peer support forums for OCD. In addition, clinicians should consider if their patients may be reassurance-seeking on this and similar OCD forums, given their popularity. Future studies should conduct interviews with users and investigate patterns of user ability to further understand the potential harms and benefits of online peer support forums.

(JMIR Form Res 2025;9:e60899) doi:10.2196/60899

KEYWORDS

obsessive-compulsive symptoms; reassurance-seeking; peer support group; thematic analysis; Reddit; online peer support forum; online support; obsessive disorder; anxiety disorder; debilitating chronic anxiety disorder; anxiety; stress; depression; website; treatment; online forums; obsession; patients

Introduction

Background

Obsessive-compulsive disorder (OCD) is characterized by recurrent, unwanted, distressing thoughts, images, or impulses termed obsessions. These obsessions are highly diverse and can often be idiosyncratic [1]. In response to obsessional distress, individuals with OCD typically engage in repetitive or highly ritualized physical or mental behaviors aimed at reducing or preventing the distress, termed compulsions [2]. Though highly heterogeneous, a majority of OCD symptoms can be categorized into four dimensions: (1) obsessions about contamination, accompanied with cleaning rituals; (2) obsessions about causing

XSL•FO
RenderX

harm, accompanied with checking or reassurance-seeking rituals; (3) obsessions about incompleteness, accompanied with ordering or rearranging rituals; (4) obsessions about morally unacceptable violent, sexual, or religious behaviors, accompanied with mental rituals intended to negate these thoughts [1,3].

OCD is a debilitating condition resulting in significant impairment of quality of life [4] that affects around 2% of adults in the United States, with women being 1.5 times more likely to experience OCD [5]. Current first-line treatments for OCD include cognitive-behavioral therapy using exposure and response prevention (ERP) and pharmacotherapy with selective serotonin reuptake inhibitors [6,7]. Though these therapies are effective, an estimated 57% of people with OCD do not access treatment due to various barriers, such as financial or geographical constraints, a deficit of available therapists, and mental health stigma [8,9].

Peer support groups may encourage and help people with OCD better understand their condition and motivate them to identify, seek, and continue engaging in treatment [10]. Social media is a common platform for social interaction in modern societies, and its use generates a substantial impact on anxiety disorders [11]. While in-person OCD support groups overseen by health care professionals tend to be beneficial for OCD patients [12], such support groups may be expensive or geographically inaccessible for some individuals. However, OCD peer support groups, particularly on Facebook (Meta), may also have mixed or negative effects on patients [13]; thus, the content of these online communities should be further investigated.

An online forum focused on OCD may provide particular benefits. OCD patients may be reluctant to discuss symptoms with physicians due to their taboo subject matter [6], but may be more willing to do so on an anonymous online forum. Interacting with other individuals with OCD in an online community may encourage individuals to recognize and avoid performing compulsions, as well as learn about and seek medical treatments—for example, users may use the platform to exchange helpful information about treatment experiences and assist other users in enrolling in treatments they were previously unaware of.

However, an online forum with OCD may also be harmful in certain circumstances. Misinformed or underqualified users or “trolls” may perpetuate counterproductive or harmful information about OCD or send rude messages in the community. In addition, many individuals with OCD engage in pathological and compulsive reassurance-seeking, which provides temporary relief but ultimately aggravates their condition [14]: for example, an individual with symptoms in the unacceptable thoughts dimension of OCD may use an online forum to confess their OCD obsessions to allay their fears [15]. Symptom accommodation may occur when other users comment to reassure the original poster that they are not a pedophile, that what they fear they did was a false memory, and that such reactions are a symptom of OCD. Although such reassurance provides temporary relief, it ultimately paradoxically reinforces obsessions and prompts future reassurance-seeking and other compulsions [16].

As no research has been done on public Reddit-based forums specifically for OCD, despite their 100s of 1000s of active users, a robust qualitative understanding of the structures and conventions of these forums would be valuable for clinician awareness and provide important foundations for further investigations on potential harms and risks posed by these forums.

Objective

The primary purpose of this study is to explore what types of content are posted in r/OCD, a large, public online support forum for OCD hosted on Reddit, to build our understanding of the dynamics and norms of such online support communities.

Reddit is a free online social media platform accessible in most nations in the world that enables registered users (Redditors) to add posts and comments that may include text, images, videos, links, and GIFs in a variety of themed communities known as “subreddits.” Reddit, which enables users to post anonymously and hold discussions in the comments, hosts over 100,000 active subreddits on a wide range of topics [17], including psychopathology, in subreddits such as r/Anxiety, r/Depression, and r/OCD. Reddit is predominantly used by men (78%) between the ages of 18 and 29 (59%) [18].

The r/OCD subreddit is an online forum dedicated to users with a specific interest in OCD, including but not limited to individuals living with, or family, friends, and acquaintances of people living with OCD. Posts in this forum are divided into categories, known as flares, including “questions about OCD,” “crisis,” “articles,” “discussion,” and “sharing a win.” The forum is moderated by 6 users, one of whom is a licensed therapist. It is currently the largest public online forum targeting users with OCD, with 180,000 users and 100 - 200 new posts added every day, as of June 2023. The usage of this forum, or any other online forum for OCD, has not been previously studied.

Thematic analysis has previously been used to study posted content in other online forums used by individuals with mental illness, such as eating disorder support forums [19], forums dedicated to other potentially taboo topics, such as men’s sexual health [20], and forums focused on response to treatment and care practices, such as tinnitus treatment [21]. Using inductive thematic analysis, we captured the main types of posts and exchanges taking place in the subreddit to better understand what posters use this platform for.

Methods

Data Collection

Reddit has a 1000-item listing limit: users can view the 1000 most recent posts and comments within a subreddit. To capture as much data as possible, on August 8, 2023, the lead author (NYS) manually downloaded every post and comment added to the r/OCD subreddit over a 72-hour period, from August 4, 2023, to August 7, 2023. The coders felt that saturation of ideas (defined in Data Analysis) was reached after coding the posts collected over the first 24-hour period; this comprised a total of 132 new posts with a total of 739 comments that were added to the r/OCD subreddit over this period. Similar Reddit content analyses have examined 100 - 200 posts [22,23].

XSL•FO
RenderX

Data Analysis

To most directly capture the content in the subreddit, we conducted inductive thematic analysis using a postpositivist "scientifically descriptive" small-q approach [24]. Analysis was performed inductively at the explicit level—a method used to describe sets of previously uncharacterized data [25]—as there were no suitable pre-existing frameworks characterizing online peer support groups for people with OCD.

The lead author (NYS), a female student and research assistant who has previously led inductive thematic analyses using data from social media, read all posts in the data set and developed an initial codebook focused on characterizing the categories of content being posted (eg, describing a symptom and asking for advice on a medication) rather than the details of the content (eg, individual worries about being a pedophile or having eaten contaminated food). To enhance dependability, the initial codebook, which contained 58 codes, was independently reviewed by the last author (TC), a male licensed clinical psychologist and instructor who was also familiar with the data and had expertise in OCD research and qualitative analytical methods. The 2 authors discussed and reflected on potential oversights and biases until they reached a consensus on the initial codebook.

The lead author coded the data by thread (ie, a post and its comments) in a spreadsheet, with each post or comment being labeled with at least one code. During the coding process, the lead author iteratively discarded codes that were not used, combined codes that were highly similar, and added codes to describe new patterns observed. The last author regularly reviewed coding and discussed any potential biases or overinterpretation of not explicitly present phenomena with the first author until consensus was reached. Saturation of ideas, which was defined as no new codes or themes emerging in new threads [26], was observed within the first 110 posts coded.

After the coding of the entire data set was completed, and agreement on all codes was reached between the first and last author, the coding sheet contained 78 codes. Due to the higher number of resulting codes, both authors agreed to eliminate codes that did not appear a substantial number of times in the data set. The number of times each code was applied was tabulated, and there appeared a clear divide with a cluster of codes that appeared 2 - 3 times versus codes that appeared more than 10 times. The codes that appeared less than 10 times were removed from the data set, leaving 38 codes.

Finally, through discussion and a consensual decision-making process [27] between the first and last author, these 38 codes were grouped into overarching themes. The themes were iteratively refined by discarding ones that were too heterogeneous or without enough support from the data, combining ones that were too similar, and so forth, until the authors arrived at a set of 4 themes to capture the text [24].

Ethical Considerations

All data collected from the r/OCD subreddit were posted publicly online and did not require any permission to access. All posts and comments analyzed in this study were posted by anonymous Reddit users and are not identifiable. This study (IRB Protocol ID: 2000036062) received a "Not Human Subjects Research Determination" from the Yale Institutional Review Board.

Results

Overview

We analyzed 132 new posts with a total of 739 comments added to r/OCD, an online support forum focused on OCD, during a 24-hour period. The user-selected category (ie, flare) breakdown of the posts were as follows: 7 were tagged as "Discussion," 11 as "I just need to vent," 69 as "I need support," 30 posts as "Question about OCD," 6 as "Crisis," 4 as "Art Film Media," and 5 as "Sharing a Win!"

During the coding process, we focused on characterizing the types of content being posted, and 38 codes were developed (refer to Data Analysis subsection). The 38 codes were subsequently and consensually organized into 4 main themes that capture the types of discussions taking place in the r/OCD online support forum: (1) validating OCD symptoms (69 total posts and comments), (2) connecting with peers with similar OCD themes (512 total posts and comments), (3) coping with symptoms (280 total posts and comments), and (4) sharing treatment experiences and advice (118 total posts and comments). Posts and comments could be categorized into multiple themes. Quotes provided below are lightly edited for clarity and grammar.

Theme 1: Validating OCD Symptoms

Under this first theme, which includes 69 posts and comments, Redditors on r/OCD posted descriptions of their own fears and behaviors, and asked the subreddit community whether their behaviors could be caused by or related to OCD. Some Redditors posting content that fell under this theme appeared to be somewhat unfamiliar with OCD, expressing uncertainty about whether their concerns are descriptive of or related to OCD.

> ***Post title***: *Am i a [expletive] pedo?*
>
> ***Post text***: *I was watching Reddit and some baby reacting to a video. My groinal part had a weird feeling then like it moved or something. I do get that out of no where but I'm scared that what if I'm a pedo? Is ocd the cause of this? Please help.*

However, many others posting content that fell under this theme seemed familiar with OCD; some mention already having an OCD diagnosis, while others state that they have family members or close friends with OCD. They appeared to be aware that their symptoms were "probably OCD" and/or used terminology associated with OCD to describe their symptoms, but subsequently seemed to express doubt about whether the label is valid. This matches clinical descriptions of "OCD about OCD," in which individuals worry obsessively about whether their symptoms are truly OCD. While this behavior was likely reassurance-seeking, and it typically elicited reassuring or clarifying responses from others, it was not directly labeled as such.

> ***Post title***: *Scared of getting cold sores*

XSL•FO
RenderX

> ***Post text***: *I have pretty intense contamination OCD. Someone I'm interested in gets cold sores. I've never had them myself but I fear them giving me herpes and having to disclose that for the rest of my life. I can't tell if its just my OCD, or if im thinking rationally and it's my OCD just dramatizing it? I know that you can pass it without having an outbreak which is what has me spiraling. Should I cut things off? I really don't want to*
>
> ***Post title***: *Is this even an obsession?*
>
> ***Post text***: *I am obsessed over thinking about whether people hate me and want the worst for me and then I try and go ask for reassurance. And then I do it all over again.*

In response to these posts, other Reddit users commented on whether they believed the OP has OCD. In most posts, users affirmed that the individual has OCD. These responses often overlapped with theme 2 (connecting with peers with similar OCD themes); after agreeing that the OP's experience was a symptom of OCD, they would detail their own experience with a similar symptom. In very few cases, Reddit users suggested that the symptom was normal or attributed it to another diagnosis.

> ***Comment***: *Yes [this is a symptom of OCD]!! It's called magical thinking in OCD land*

Theme 2: Connecting With Peers With Similar OCD Themes

Connecting over shared experiences related to OCD was by far the most common type of post or comment that appeared in r/OCD, with a total of 512 posts and comments. Shared experiences could include obsessions, compulsions, and triggers, as well as frustrations relating to OCD. Causal attributions for OCD were also common, concerning events or experiences in users' youth. Most users who posted content falling under this theme appeared somewhat familiar with OCD. In almost all cases, individuals appeared to understand what their obsessions and compulsions were; they correctly used medical terms, such as "groinal responses," to describe their experiences.

Furthermore, there are common abbreviations on the forum that these users consistently use, such as "ROCD" (relationship OCD, or obsessive-compulsive concerns related to one's personal relationships), "POCD" (pedophilia-themed OCD, or obsessive doubt about whether one is a pedophile). Interestingly, the term "Pure O" (compulsions that are primarily or exclusively mental, which are often mistaken as obsessions) was also used frequently and erroneously [28]. This and the other terms do not always perfectly correspond to the scientific literature's understanding of OCD symptom dimensions; rather, they seem to be popularized by the r/OCD community, within the community itself.

Sometimes, the user used strongly emotional language, venting about symptoms, and detailing how OCD was "ruining [their] life," but more often, the user appeared to be succinctly and calmly recalling their experiences.

Finally, this type of content also often ended with the Reddit user asking if anyone else had a similar experience to them, as an attempt at seeking validation from other users. While this behavior was likely reassurance-seeking, it was not directly coded as such, as in some cases, individuals appeared to be seeking individuals who related to their symptoms to vent to, and it is impossible to distinguish from a single post whether individuals were seeking reassurance, searching for peer support, or both.

> ***Post title***: *Fear of being delusional*
>
> ***Post text***: *Hey guys! I have pure ocd for 5 years and mostly my main theme is guess what becoming a schizophrenic. Couple years ago i read here about fear of being delusional. And now i also got this i have weird intrusive thoughts-what if my parents dont love or what if the food is poisoned. I dont believe in this thoughts but i am afraid of them they make my a little of anxious. I am afraid what if i will one day become delusional. I must say that i havent got this fear since 2020. I mostly have other type of schiz theme. Yesterday i was bored and now because my brain cant be calm it suddenly repeat this fear from 2020. Any similar experiences? Thanks :))*

Oftentimes, other users replied to their post or comment, relating to their symptoms, triggers, and frustrations.

> ***Comment***: *I can totally relate. I have an extreme phobia about sending horrible messages to everyone in my contacts or downloading very illegal content on my devices when I am not conscious. I also had a long-time suspicion that devices were recording me so I would need to constantly make sure they were totally turned off, closed, or placed face down when not using them. This really got on my nerves. I honestly don't know the cause of this either, but I feel like false memory OCD attacks our values.*

This type of content was not limited to OCD-specific symptoms, but also involved expressing frustrations regarding public perception and misconceptions of OCD. Among the most popular threads in the forum were posts that discussed misconceptions related to OCD, bringing many users into the comment section to vent.

> ***Comment***: *I saw a post on Instagram (Meta) recently that was a response to a person who posted a video that was saying intrusive thoughts are your intuition.*
>
> ***Reply***: *well guess I'm a future serial killer because I saw a dead possum. And I'm an irredeemable monster worse than hilter because my room was messy. Man, Ugh, the idea of intrusive thoughts being intuition and being you is horrible.*

Reddit users also shared their OCD experiences through referencing art and media, such as songs they found expressed the anxiety they experienced due to their OCD.

> ***Post title***: *Has anybody watched monk*
>
> ***Post text***: *I've been watching it on peacock and I absolutely love it. Idk [if] that's a[n] accurate representation of some people with ocd either way I*

XSL•FO
RenderX

relate to him so much when it comes to my own problems.

Some Reddit users created art of their own to share with the community, such as writing out abstract metaphors that convey their experience of OCD.

> ***Comment**: OCD being egodystonic is why it's such a torture. It's like, I'm battling an alternate universe version of myself inside my mind.*

Theme 3: Sharing Adaptive and Maladaptive Coping Mechanisms

Under the third theme, which includes 280 posts and comments, Reddit users described and recommended using a variety of both adaptive and maladaptive coping mechanisms to deal with OCD symptoms on their own. Much of this content took the form of a user describing their obsessions and compulsions (overlapping with theme 2) and asking for advice on how to alleviate their symptoms.

> ***Post title**: How do you just not let your intrusive thoughts bother you?*
>
> ***Post text**: I'm really sick of my thoughts and it doesn't help that I suffer from POCD and real event OCD. Sometimes it bothers me but I just ignore it and sometimes it bothers me too much to ignore it. I definitely think I should get therapy but I'm scared to tell my parents that I need therapy. So in the meantime, what are some tips and tricks to not let the thoughts bother you?*

Other users then suggested coping mechanisms, sometimes describing their own experiences applying them.

> ***Comment**: I just started doing this recently after another bad night of intrusive thoughts. I was getting extremely frustrated because I couldn't sleep, so I came up with this "game" to do in my head. I don't really have a name for it. What I do is choose a topic. Literally any topic, like fruits & vegetables, animals, rabbit breeds, TV shows, anything that I can make a list of. Then, I try to name something for every letter of the alphabet. So if I'm doing fruits and vegetables, I go: apple, broccoli, cucumber, date, eggplant... And if I get stuck on a letter, I try to think about it as much as possible. If the intrusive thoughts try to come in, I think to myself, "Not now, I'm trying to think of something that starts with this letter. " If I finally come up with something, I restart the alphabet and work my way up to Z. If I can't think of anything for that letter, I just go to the next letter and continue on. Once I finish that topic, I start a new one. I finished the food one, so now I'm going to make a list of animals. Then I start: alpaca, bunny, capybara, deer, elephant... I keep doing this game either until my head stops buzzing or I pass the hell out in the middle of it. It has helped me so much, and I am a lot calmer now because of it. Apologies if this isn't helpful. It really helps me out so I figured I would share:)*

Some of these coping mechanisms are in line with clinical practice and appear to be suggested by individuals who are well-informed about OCD and the nature of obsessions and compulsions. Users commonly suggested recognizing thoughts as being a symptom of OCD, avoiding performing compulsions, and exposing oneself to their fears. Users also often linked resources, such as blogs or YouTube (Google LLC) channels, that offer coaching for individuals with OCD.

> ***Comment**: I found this article by Anxiety Canada [link] helpful. Scroll down past the diagram – there are a few things you could start trying.*

Other coping mechanisms, however, may have inadvertently reinforced the original poster's OCD symptoms. Users frequently and perhaps unknowingly offered reassurance in their comments, stating that the original poster's fears will not come true. In some cases, this was done by users who appeared well-informed about OCD and had linked to high-quality clinical resources as well, making it appear that users may not be aware that they are offering reassurance.

The forum is partially moderated by a licensed therapist, who occasionally alerts a user that they are giving harmful reassurance and removes that user's comment, but due to the high volume of traffic received by this forum, most reassurance-giving comments are not removed. Users also sometimes suggested avoidance techniques, such as avoiding a dirty location if an individual has OCD obsessions relating to contamination, which is a maladaptive response that serves to reinforce the individual's obsessive fears [29].

> ***Comment**: I think you should go to the doctor just in case given how you've described the bite. Ik it's reinforcing OCD but in this one particular situation I think it's important to make sure [that you don't have rabies, which is the original poster's OCD fear].*

In addition to sharing coping mechanisms, many users also expressed sympathy and encouragement for other users, providing social support to each other. In return, the users expressed gratitude for the resources and advice offered to them to cope with their symptoms. A similar supportive social culture has been observed on other subreddit communities for individuals battling addiction [30].

> ***Comment**: Are you talking about Acceptance and Commitment therapy? I am reading about self compassion in relation to therapy and I love the concepts. Thank you for your comment!*

Theme 4: Sharing Treatment Experiences and Advice

Under the fourth theme, which includes 118 posts and comments, Reddit users shared their experiences in and advice regarding a variety of professional treatments, including ERP therapies and cognitive-behavioral therapies, and taking various prescribed medications. Some users did so unprompted, while others did so in response to users asking about others' experiences with certain treatments that they were considering for themselves.

> ***Post title**: How do I get over my OCD regarding public toilet fumes?*
>
> ***Post text:** All this started over quarantine; I never used to get overly OCD until I started watching those*

XSL•FO
RenderX

videos about germs. It hasn't gone away since then, and it's gotten to the point where I can barely use toilets in my own home. It seems to be naturally going away a bit but I still get extreme anxiety when it comes to using public restrooms because of the toilet flushing fumes, and the fact that there's no toilet cover. And help?

Some users used the platform to express criticism of their psychiatrist or therapist, and asked others for advice on whether to stop or change their treatment routine. The most common criticism involved their therapist not seeming to understand how to treat OCD, or concerns about side effects of medication.

***Post title**: Clomipramine and weight gain*

***Post text**: Hello everyone I hope you all are doing okay, I just got prescribed anafranil (clomipramine) for ocd I had ocd since I was a kid I just didn't know that it was a disorder but it got really bad as I grew up, I've been on sertraline for three months now and I just got prescribed anafranil and I want to know how does it affect weight gain is it just by increasing your appetite or does it do something to your body that make you produce more fat?*

Other users used the platform to describe improvements in their OCD following treatment and offer encouragement to others that it is possible to get better. This content was rare but typically well-met, with many users congratulating the original poster on their progress.

***Post title**: Ate food that made me anxious. I feel sick to my stomach, but also accomplished.*

***Post text**: A few days ago, my neighborhood had an unexpected power outage (no storm or anything to indicate it was going to happen, there was just a bang and everything was off) for a little over four hours. I have major anxiety about food going bad – when I am at the grocery, I put refrigerated food into an insulated bag and keep it in the bag until I am at home putting the groceries away and I still feel like the food has gone bad. So, the power going out and cutting power to the freezers and fridge is like my worst nightmare. It's caused a lot of stress to both myself and my family the past few days because they don't think it's any problem and think I'm worried about nothing. Now, I know that logically, the stuff in the fridge could maybe have gone bad, but the stuff in the freezer should be totally fine. Still, I have been so stressed about eating it. Just now I warmed up something that was in the chest freezer and ate it. I honestly feel like I'm going to vomit, but I did it and I know that everything is going to be okay.*

Discussion

Principal Findings and Comparison With Previous Work

To our knowledge, this study is the first study specific to a Reddit forum dedicated to supporting individuals with OCD. OCD is a debilitating disorder, and less than half of people with OCD are able to access medical treatments, such as ERP [8,9]. Online support communities may provide a free and more accessible pathway for receiving social support and valuable information on OCD treatments; however, they may also provide an outlet for individuals with OCD to receive misinformation and seek reassurance anonymously online from other users, aggravating their condition.

We examined the content of the subreddit r/OCD, the largest online support community for individuals with OCD, using thematic analysis. This research extends previous studies that have examined in-person support groups for individuals with OCD [13,31] by exploring a potential internet-based, highly accessible support group. This study draws upon the popular and highly active pre-existing r/OCD subreddit with the goal of providing generalizable observations on the dynamics and norms of online support forums for OCD.

In the most common type of post on this subreddit, users described OCD symptoms, and either inquired whether their symptoms could be classified as OCD and asked if anyone else had had similar experiences and asked other users to suggest coping mechanisms. Then, other users indirectly or directly affirmed that this was a symptom of OCD and shared their own experiences with similar symptoms and/or a combination of beneficial and potentially harmful coping mechanisms for these symptoms.

Subreddit users frequently titled their posts with a subtype of OCD (eg, pedophilia, relationship, sexuality, and cleanliness) based on the theme of the obsessions, which enabled users to rapidly connect with other individuals suffering from the same theme. Notably, "Pure O" was commonly perpetuated by users as a subtype of OCD "without compulsions"—in reality, the compulsions are simply purely mental—suggesting that the subreddit is vulnerable to the spread of misinformation.

While these linguistic features suggested that the user base is overall familiar with OCD, users often expressed lingering doubt about whether their symptoms were due to OCD. Many users made posts asking other users if their symptoms are "probably due to OCD, right?" Previous studies have found that a major motivation that drives participants to join health-related peer support groups is to seek more information on their condition. Users asking these questions may become dissatisfied by information they judge to be of poor quality, which can lead to extensive mental review of their symptoms and the information provided, and ultimately generate greater worry and obsessions about their disorder [13]. Indeed, a previous study by Tan et al [13] examined 10 Facebook groups and 3 Reddit forums and found that internal preoccupation with symptoms was a significant predictor of negative experiences in these groups.

Users' desire to frequently question their symptoms online is described by the emerging concept of cyberchondria: compulsive internet-checking behaviors of individuals with health-related obsessions due to intolerance of uncertainty, a core trait associated with OCD [13,32].

Though the r/OCD user base demonstrated substantial knowledge about common OCD therapies, such as ERP, users troublingly failed to refrain from or recognize and call out

XSL•FO
RenderX

potentially reassuring messages to other users. This resembles the phenomenon of symptom accommodation, in which individuals, especially family members, provide reassurance or otherwise assist an individual with OCD in performing their compulsions with the goal of offering temporary relief, oftentimes despite being aware that this behavior is deleterious [33]. Symptom accommodation has been shown to delay and decrease treatment response [34]; thus, interactions on this widely and frequently visited forum may present a previously undocumented cause for poorer treatment outcomes among individuals with OCD. This counterproductive behavior is unfortunately condoned by the forum moderators: though the subreddit guidelines explain that reassurance-seeking is counterproductive, this guideline begins by stating that "some reassurance is allowed" (as of December 2024).

The forum norm of using subtypes of OCD to connect with other users with very similar experiences may make it especially difficult for users to refrain from accommodating reassurance-seeking, as the community is structured for users to easily find other users whose experiences they may most easily empathize with. In a study of patients' caregivers, empathy with the patient with OCD was suggested as a driving factor for the caregivers (often family members) who understood that reassurance-seeking was unhelpful in the context of OCD, to provide reassurance to the patient anyway [35].

A smaller number of posts shared therapy, medication, and recovery experiences. Users successfully established camaraderie with other users, shared useful resources on treatment options, and bonded over shared symptom experiences, provided sympathy and encouragement for each other's recovery, and expressed gratitude. This positive atmosphere was ubiquitous, no matter the type of post or dimension of symptoms being discussed. Past studies have found that individuals with OCD have lower perceived social support [36], but that social support is positively associated with symptom improvement [37] and negatively associated with symptom severity [38]. Platforms such as this one may be an ideal way for isolated individuals affected with OCD to obtain peer support.

Observations of the community's supportive culture is in agreement with a previous study on a subreddit for opioid addiction recovery, which also found many instances of social support and mutual aid [30], as well as a study of in-person group therapy for individuals with OCD that found that the participants naturally formed their own support group to support each other and exchange recovery strategies [31]. However, our findings contradict a previous study of online peer support groups for individuals with obsessive-compulsive and related disorders, which focused primarily on online trichotillomania peer support groups on Facebook; this work found that participants had a variety of negative social experiences, including cyberbullying and being blocked by the group. This difference may be due to the fact that Facebook groups are smaller and more personal, whereas Reddit is anonymous, and this OCD forum consists of over 180,000 users who seem to visit the forum as needed, making cyberbullying less feasible.

Overall, community norms appear to promote seeking out other users with similar OCD subtypes and experiences and asking for confirmation that certain symptoms are considered OCD. The community culture of reassurance-seeking emphasizes the importance of having knowledgeable moderators on online support forums. While a few reassurance-providing comments were removed by the licensed therapist moderator with an explanation of why offering reassurance is harmful, the volume of posts and comments was too large for a single moderator to parse. However, though a minority of posts discussed treatment-seeking, these posts appeared highly productive for users, who received helpful information and encouragement on their recovery journeys.

A potential improvement to the forum may be restricting the detailed sharing and discussion of specific themes and symptoms or limiting the discussion of specific OCD subtypes, so reassurance cannot be offered. Along this line, the forum can be restructured to promote the exchanging of treatment information, as those posts appeared to be the most beneficial for users, but only represented a minority of content posted. Currently, none of the flairs are dedicated to seeking and receiving advice about medical treatment—creating flairs, such as "Treatment–Medication" and "Treatment–Therapy," may encourage users to share such posts.

Strengths and Limitations

This study has several strengths—it examines a large volume of qualitative data from a very well-visited subreddit with over 180,000 users and 100 - 200 new posts daily. Individuals from any demographic background may contribute anonymously, and thus the data is less vulnerable to some forms of stigma or underrepresentation compared to other studies that do not use naturally occurring data. Overall, this study provides an important basic understanding of the content posted on this community, which future studies may build off of. As no prior studies have focused on such an online forum despite its large user base, an understanding of the content of such forums is valuable to clinicians, whose patients may be visitors or participants in the forum.

However, study results must be considered in the context of their limitations. First, the goal of this study is to understand the forum's community rather than individual users; thus, it was beyond the scope of this preliminary study to determine through interviews and user analytics whether most users are first-time visitors or recurring, if users used the platform to seek reassurance, if users truly benefited from exchanging treatment advice, etc. Second, thematic analysis cannot quantify the extent of benefit or harm provided in Reddit forums, which should be further investigated through quantitative analyses that were beyond the scope of this study. Third, while anonymous platforms, such as Reddit, enable users to discuss potentially taboo OCD symptoms with decreased fear of repercussions, they also prevent the collection of user demographics. As noted as a limitation in similar studies, Reddit users are primarily male young adults and may not be representative of all patients with OCD [39]. Fourth, only English posts were included in this study. Thus, the themes identified may not reflect the

XSL•FO
RenderX

experiences of users in OCD online support forums hosted on other platforms or in other languages.

Future Directions

To build upon the results of this study and address the aforementioned limitations, future studies may analyze the usage patterns of frequent visitors to the subreddit (ie, characterizing what type of posts and comments individuals who may be compulsively returning to this subreddit are making) to provide clinicians with better tools to determine the impact OCD patients' subreddit use may have on their treatment course–whether online forums are helping or delaying it. Future studies may also interview and quantitatively survey subreddit visitors, especially those who make posts that appear to be reassurance-seeking, to assess at the user-level the benefits and harms of visiting the subreddit.

Conclusions

In conclusion, we observed that most posts on r/OCD involve users sharing their OCD symptoms, finding others with similar symptoms, and potentially seeking reassurance that their symptoms are OCD; forum norms promote using OCD subtypes to seek users with similar experiences. Secondarily, forums are used to exchange information about treatment and provide emotional and social support on each other's recovery journeys. Overall, our study suggests that individuals with OCD may benefit from participating in online support forums to receive valuable information and social support, but that such forums benefit from moderators with health care backgrounds to prevent establishing a norm of symptom accommodation. Given the over 180,000 users and hundreds of daily new posts every day, we recommend that clinicians should cautiously evaluate their patients' use of this subreddit or similar internet-based OCD forums to determine if patients may be displaying this pattern of reassurance-seeking.

Acknowledgments

This work was supported by the State of Connecticut through its support of the Ribicoff Research Facilities at the Connecticut Mental Health Center. The content reflects the views of the authors and not of the State of Connecticut.

Data Availability

The data analyzed during this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

References

1. Abramowitz JS, Deacon BJ, Olatunji BO, et al. Assessment of obsessive-compulsive symptom dimensions: development and evaluation of the Dimensional Obsessive-Compulsive Scale. Psychol Assess 2010 Mar;22(1):180-198. [doi: 10.1037/a0018260] [Medline: 20230164]
2. Olatunji BO, Christian C, Brosof L, Tolin DF, Levinson CA. What is at the core of OCD? A network analysis of selected obsessive-compulsive symptoms and beliefs. J Affect Disord 2019 Oct 1;257(45-54):45-54. [doi: 10.1016/j.jad.2019.06.064] [Medline: 31299404]
3. Wheaton MG, Abramowitz JS, Berman NC, Riemann BC, Hale LR. The relationship between obsessive beliefs and symptom dimensions in obsessive-compulsive disorder. Behav Res Ther 2010 Oct;48(10):949-954. [doi: 10.1016/j.brat.2010.05.027] [Medline: 20800750]
4. Asnaani A, Kaczkurkin AN, Alpert E, McLean CP, Simpson HB, Foa EB. The effect of treatment on quality of life and functioning in OCD. Compr Psychiatry 2017 Feb;73(7-14):7-14. [doi: 10.1016/j.comppsych.2016.10.004] [Medline: 27838572]
5. Fawcett EJ, Power H, Fawcett JM. Women are at greater risk of OCD than men: a meta-analytic review of OCD prevalence worldwide. J Clin Psychiatry 2020 Jun 23;81(4):19r13085. [doi: 10.4088/JCP.19r13085] [Medline: 32603559]
6. Law C, Boisseau CL. Exposure and response prevention in the treatment of obsessive-compulsive disorder: current perspectives. Psychol Res Behav Manag 2019;12(1167-1174):1167-1174. [doi: 10.2147/PRBM.S211117] [Medline: 31920413]
7. Kellner M. Drug treatment of obsessive-compulsive disorder. Dialogues Clin Neurosci 2010;12(2):187-197. [doi: 10.31887/DCNS.2010.12.2/mkellner] [Medline: 20623923]
8. Kohn R, Saxena S, Levav I, Saraceno B. The treatment gap in mental health care. Bull World Health Organ 2004 Nov;82(11):858-866. [Medline: 15640922]
9. Senter MS, Patel SR, Dixon LB, Myers RW, Simpson HB. Defining and addressing gaps in care for obsessive-compulsive disorder in the United States. PS 2021 Jul 1;72(7):784-793. [doi: 10.1176/appi.ps.202000296]
10. Harandi TF, Taghinasab MM, Nayeri TD. The correlation of social support with mental health: a meta-analysis. Electron Physician 2017 Sep;9(9):5212-5222. [doi: 10.19082/5212] [Medline: 29038699]

XSL•FO
RenderX

11. Vannucci A, Flannery KM, Ohannessian CM. Social media use and anxiety in emerging adults. J Affect Disord 2017 Jan 1;207(163-6):163-166. [doi: 10.1016/j.jad.2016.08.040] [Medline: 27723539]
12. Jónsson H, Hougaard E. Group cognitive behavioural therapy for obsessive-compulsive disorder: a systematic review and meta-analysis. Acta Psychiatr Scand 2009 Feb;119(2):98-106. [doi: 10.1111/j.1600-0447.2008.01270.x] [Medline: 18822090]
13. Tan YT, Rehm IC, Stevenson JL, De Foe A. Social media peer support groups for obsessive-compulsive and related disorders: understanding the predictors of negative experiences. J Affect Disord 2021 Feb 15;281:661-672. [doi: 10.1016/j.jad.2020.11.094] [Medline: 33234279]
14. Stewart SE, Hezel D, Stachon AC. Assessment and medication management of paediatric obsessive-compulsive disorder. Drugs (Abingdon Engl) 2012 May 7;72(7):881-893. [doi: 10.2165/11632860-000000000-00000] [Medline: 22564131]
15. Abramowitz JS, Jacoby RJ. Scrupulosity: a cognitive–behavioral analysis and implications for treatment. J Obsessive Compuls Relat Disord 2014 Apr;3(2):140-149. [doi: 10.1016/j.jocrd.2013.12.007]
16. Parkin R. Obsessive-compulsive disorder in adults. Int Rev Psychiatry 1997 Jan;9(1):73-82. [doi: 10.1080/09540269775600]
17. Reddit. 2023. URL: https://www.redditinc.com/press
18. Barthel M, Stocking G, Holcomb J, Mitchell A. Reddit news users more likely to be male, young and digital in their news preferences. Pew Research. 2016. URL: https://www.pewresearch.org/journalism/2016/02/25/reddit-news-users-more-likely-to-be-male-young-and-digital-in-their-news-preferences
19. McCormack A. Individuals with eating disorders and the use of online support groups as a form of social support. Comput Inform Nurs 2010;28(1):12-19. [doi: 10.1097/NCN.0b013e3181c04b06] [Medline: 19940616]
20. Uberoi M, Abdulcadir J, Ohl DA, Santiago JE, Rana GK, Anderson FWJ. Potentially under-recognized late-stage physical and psychosexual complications of non-therapeutic neonatal penile circumcision: a qualitative and quantitative analysis of self-reports from an online community forum. Int J Impot Res 2023 May;35(3):234-241. [doi: 10.1038/s41443-022-00619-8] [Medline: 36274189]
21. Assouly KKS, Shabbir M, van Dijk B, et al. The impact of tinnitus on adult cochlear implant recipients: a mixed-method approach. PLOS ONE 2023;18(4):e0284719. [doi: 10.1371/journal.pone.0284719] [Medline: 37079589]
22. Park JYE, Howren AM, Davidson E, De Vera MA. Insights on mental health when living with rheumatoid arthritis: a descriptive qualitative study of threads on the Reddit website. BMC Rheumatol 2020 Nov 26;4(1):62. [doi: 10.1186/s41927-020-00163-2] [Medline: 33292866]
23. Sowles SJ, McLeary M, Optican A, et al. A content analysis of an online pro-eating disorder community on Reddit. Body Image 2018 Mar;24:137-144. [doi: 10.1016/j.bodyim.2018.01.001] [Medline: 29414146]
24. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
25. Braun V, Clarke V. Toward good practice in thematic analysis: avoiding common problems and be(com)ing a knowing researcher. Int J Transgend Health 2023;24(1):1-6. [doi: 10.1080/26895269.2022.2129597] [Medline: 36713144]
26. Guest G, Namey E, Chen M. A simple method to assess and report thematic saturation in qualitative research. PLoS One 2020;15(5):e0232076. [doi: 10.1371/journal.pone.0232076] [Medline: 32369511]
27. Schweiger DM, Sandberg WR, Rechner PL. Experiential effects of dialectical inquiry, devil's advocacy and consensus approaches to strategic decision making. Academy of Management Journal 1989 Dec 1;32(4):745-772. [doi: 10.2307/256567]
28. Williams MT, Farris SG, Turkheimer E, et al. Myth of the pure obsessional type in obsessive--compulsive disorder. Depress Anxiety 2011 Jun;28(6):495-500. [doi: 10.1002/da.20820] [Medline: 21509914]
29. Starcevic V, Berle D, Brakoulias V, et al. The nature and correlates of avoidance in obsessive-compulsive disorder. Aust N Z J Psychiatry 2011 Oct;45(10):871-879. [doi: 10.3109/00048674.2011.607632] [Medline: 21875307]
30. Bunting AM, Frank D, Arshonsky J, Bragg MA, Friedman SR, Krawczyk N. Socially-supportive norms and mutual aid of people who use opioids: an analysis of Reddit during the initial COVID-19 pandemic. Drug Alcohol Depend 2021 May;222(108672):108672. [doi: 10.1016/j.drugalcdep.2021.108672]
31. Himle JA, Van Etten M, Fischer DJ. Group cognitive behavioral therapy for obsessive‐compulsive disorder: a review. Brief Treatment and Crisis Intervention 2003 Jun;3(2):217-230. [doi: 10.1093/brief-treatment/mhg016]
32. Vismara M, Caricasole V, Starcevic V, et al. Is cyberchondria a new transdiagnostic digital compulsive syndrome? A systematic review of the evidence. Compr Psychiatry 2020 May;99:152167. [doi: 10.1016/j.comppsych.2020.152167] [Medline: 32146315]
33. Wu MS, McGuire JF, Martino C, Phares V, Selles RR, Storch EA. A meta-analysis of family accommodation and OCD symptom severity. Clin Psychol Rev 2016 Apr;45:34-44. [doi: 10.1016/j.cpr.2016.03.003] [Medline: 27019367]
34. Boeding SE, Paprocki CM, Baucom DH, et al. Let me check that for you: symptom accommodation in romantic partners of adults with obsessive-compulsive disorder. Behav Res Ther 2013 Jun;51(6):316-322. [doi: 10.1016/j.brat.2013.03.002] [Medline: 23567474]
35. Halldorsson B, Salkovskis PM, Kobori O, Pagdin R. I do not know what else to do: caregivers' perspective on reassurance seeking in OCD. J Obsessive Compuls Relat Disord 2016 Jan;8:21-30. [doi: 10.1016/j.jocrd.2015.11.003]
36. Subramaniam M, Abdin E, Vaingankar J, et al. Obsessive-compulsive disorder in Singapore: prevalence, comorbidity, quality of life and social support. Ann Acad Med Singap 2020 Jan;49(1):15-25. [Medline: 32200393]

XSL•FO
RenderX

37. Wilhelm S, Steketee G, Fama JM, Buhlmann U, Teachman BA, Golan E. Modular cognitive therapy for obsessive-compulsive disorder: a wait-list controlled trial. J Cogn Psychother 2009;23(4):294-305. [doi: 10.1891/0889-8391.23.4.294] [Medline: 21072138]
38. Palardy V, El-Baalbaki G, Fredette C, Rizkallah E, Guay S. Social support and symptom severity among patients with obsessive-compulsive disorder or panic disorder with agoraphobia: a systematic review. Eur J Psychol 2018 Mar;14(1):254-286. [doi: 10.5964/ejop.v14i1.1252] [Medline: 29899808]
39. Osadchiy V, Mills JN, Eleswarapu SV. Understanding patient anxieties in the social media era: qualitative analysis and natural language processing of an online male infertility community. J Med Internet Res 2020 Mar 10;22(3):e16728. [doi: 10.2196/16728]

Abbreviations

OCD: obsessive-compulsive disorder
ERP: exposure and response prevention

Edited by A Mavragani; submitted 24.05.24; peer-reviewed by AD Wiese, J Latorre, S Atanasova; revised version received 21.03.25; accepted 01.04.25; published 26.05.25.

Please cite as:
Sun NY, Pittenger C, Ching T
Analyzing the Contents of a Large, Public Online Peer Support Forum for Obsessive-Compulsive Disorder: Thematic Analysis
JMIR Form Res 2025;9:e60899
URL: https://formative.jmir.org/2025/1/e60899
doi:10.2196/60899

© Nora Yanyi Sun, Christopher Pittenger, Terence Ching. Originally published in JMIR Formative Research (https://formative.jmir.org), 26.5.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Internet Health Information–Seeking Trend of Urinary Incontinence in Mainland China: Infodemiology Study

Shuangquan Lin[1], MD; Lingxing Duan[1], MD; Xiongbing Lu[1], MD; Haichao Chao[1], MD; Xi Wen[2], MPH; Shanzun Wei[1], MD, PhD

[1]Department of Urology, The Second Affiliated Hospital, Mingde Rd 1, Nanchang, China

[2]Laboratory of Urology, The Second Affiliated Hospital, Nanchang, China

Corresponding Author:
Shanzun Wei, MD, PhD
Department of Urology, The Second Affiliated Hospital, Mingde Rd 1, Nanchang, China

Abstract

Background: Urinary incontinence (UI) is a series of clinical episodes featuring involuntary urine leakage. UI affects people in terms of their physical, emotional, and cognitive functioning, and the negative perceptions and impact on patients are not fully understood. In addition, the true demand for the treatment of UI and related issues is yet to be revealed.

Objective: The aim of this study is to examine the online search trend, user demand, and encyclopedia content quality related to UI on a national and regional scale on Baidu search, the major search engine in Mainland China.

Methods: The Baidu Index was queried using UI-related terms for the period from January 2011 to August 2023. The search volume for each term was recorded to analyze the search trend and demographic distributions. For user interest, the demand graph data and trend data were collected and analyzed.

Results: Three search topics were identified with the 18 available UI search keywords. The total Baidu search index for all UI topics was 11,472,745. The annual percent changes (APCs) for the topic Complaint were 1.7% ($P<.05$) from 2011-2021 and −7.9% ($P<.05$) from 2021-2023, and the average annual percent change (AAPC) was 0.1% ($P<.05$). For the topic Inquiry, the APCs were 16% ($P<.05$) from 2011 to 2016, −27.00% from 2016 to 2019, and 21.2% ($P<.05$) from 2019 to 2023, with an AAPC of 4.8%. Regarding the topic of Treatment, the APC was 20.3% from 2011-2018 ($P<.05$), −36.9% from 2018-2021 ($P>.05$), and 2.2% from 2021-2023, with a −0.4% overall AAPC. The age distribution of the population of each UI search topic inquiry shows that the search inquiries for each topic were mainly made by the population aged 30 to 39 years. People from the eastern part of China made up around 30% of each search query.

Conclusions: Web-based searching for UI topics has been continuous and traceable since January 2011. Different categorized themes within the UI topic highlight specific demands from various populations, necessitating tailored responses. Although online platforms can offer answers, medical professionals' involvement is crucial to avoid misdiagnosis and delayed treatment.

(JMIR Form Res 2025;9:e55670) doi:10.2196/55670

KEYWORDS

urinary incontinence; Baidu; Baidu index; Baidu encyclopedia; infodemiology; public interest; patients' concern

Introduction

Urinary incontinence is a series of clinical episodes that involve involuntary urine leakage [1]. Based on the different mechanisms and symptomatic discrepancies, major phenotypes of UI are (1) stress UI (SUI), (2) urgency UI, (3) mixed UI, and (4) overflow incontinence, along with other defined categories [1,2]. The frequency of UI varies greatly depending on demography, gender, and definitional differences. The surveyed prevalence ranges from 30.3% to 45% in females and 3.5% to 6% in males [3]. In China, a survey of 518,465 people revealed the UI prevalence ranges from 37.1% to 45.1% [4]. In men over 60, the prevalence ranges from 35% to 56.3% [5,6]. Additionally, the yearly expense of UI ranges from US $7.6 billion (€7 billion) to US $66 billion; it is estimated to rise to $82.6 billion by 2020 [7,8].

UI affects people's physical, emotional, and cognitive functioning [9]; the negative perceptions of urinary leakage problems negatively impact patients and affect their psychological status and social roles, impairing their overall quality of life and well-being [10]. It was reported that individuals with UI were more likely to conceal their distress and be hesitant to seek medical attention out of shame and stigma [11,12]. As a result, the known frequency of UI may be the tip of the iceberg [13,14].

Infodemiology, the study of infodemics, information distribution, and demand for online health information and

XSL•FO
RenderX

misinformation, has been credited with revealing aspects of public health, social politics, and public perception, as well as guiding health professionals and patients to quality health information [15-17]. Infodemic research was carried out in the field of urology for both premature ejaculation and lower urinary tract symptoms, which were identified as the most widespread and concerning urologic issues [18,19]. Considering the possible adverse effects of UI on the well-being, quality of life, and social involvement of patients, the utilization of infodemiology as a metric may shed light on the unrecognized needs and demands of these populations through an alternative approach. Hence, this study aims to assess the online demand for UI information by analyzing search trends and inquiry distribution and user demand. Additionally, we aim to identify the most prominent issues pertaining to real-world geographic databases at present.

Methods

Keyword Selection and Data Retrieval

The UI keywords used for the search were determined by referring to the International Urogynecological Association and International Continence Society published terminology report [1,20]. Based on these reports, various UI-describing keywords for both genders and multiple specific conditions were listed to avoid omission. All synonyms or complex derivatives were screened and selected to minimize language habits derived from ambiguity and bias, as previously described [18,21]. All available UI keywords with established temporal search trend data were categorized into 3 topics according to their connotation and are listed in Multimedia Appendix 1.

As described in previous studies [18,21], three modules on the Baidu Index platform—(1) the searching trend, (2) user demand, and (3) the demographic portrait module—enabled the analysis of search demand by examining trends in popularity, topic-related concerns, and the geodemographic features of each topic. The Baidu search index (BSI) value is the key index that numerically scales the popularity of each search keyword. It is updated daily and is recorded at the national and subnational level [18,21]; therefore, further quantifying the BSI could allow for the analysis of the search trend, regional distribution, demographic preferences, and user interest. For each search keyword, the Baidu Index platform provides the BSI record for up to a decade, depending on the module, when required. Hence, the provincial and national trend data were collected from January 1, 2011, to August 31, 2023. In addition, the most recent within-range data from the geodemographic and search demand modules were collected [18,21].

The Baidu Index user demand platform delivers to users the most pertinent search results (“related terms”) for each keyword search on a weekly basis. These randomly generated related terms reflect users’ interests and queries. These data can be analyzed and ranked based on their frequency of appearance and popularity in searches.

To review the content from the Baidu Encyclopedia, the UI-related content on this popular science platform was also collected. All UI content was scored and sorted with a preliminary global quality scale assessment [22]. The number of visits, word count, and other essential information were documented.

Data Analysis

For each UI topic, each public demand trend (the specific question proposed by the users) was presented with data sequentially plotted from BSI data in chronological order starting from January 1, 2011.

The overtime trend change for each domain was determined by the annual percent change (APC) model. The APC and average annual percentage change (AAPC) are statistical measures used in epidemiology and public health to quantify trends over time in rates or proportions of events, such as disease incidence or mortality rates. They are typically calculated using regression methods, such as joinpoint regression analysis.

APC measures the average rate of change in percentage terms over a single year. It provides insights into how a particular rate or proportion has been changing annually, that is, whether it has been increasing, decreasing, or remaining stable [23].

AAPC takes into account trends over multiple years, providing a summary measure of the average rate of change per year over a specified period. It gives a more comprehensive view of trends compared to APC, especially when trends are not linear. These models are designed to examine the change in popularity over a specified fixed interval of time [21,24,25]. $P<.05$ suggests that the trend is statistically significant.

For geographic disparities, the regional distribution of the demand for each UI term was computed annually based on provincial BSI values and sorted by the 7 Chinese administrative divisions as previously reported [21].

In the user demand section, the top 10 most frequently appearing related words for each search keyword were listed and sorted by their weekly popularity. This shows the frequency and popularity of each search-related issue on the Baidu index. Similar to previous findings, these user demand module data were retrieved and categorized into 13 categories to clarify the users’ main concerns and implied intentions [21]. These categories are (1) complaints, (2) inquiries, (3) treatment decisions, (4) health issues, (5) diagnoses, (6) hospitals and services, (7) symptom confirmation, (8) tests and examinations, (9) prognoses, (10) traditional Chinese medicine (TCM) complaints, (11) TCM inquiries, (12) TCM ingredients, and (13) TCM regimens, along with unrelated, random, or off-topic words (ie, irrelevant) [21].

Statistical Analysis

The entire database was constructed using Excel 2019 (Microsoft Corp). The APC was calculated using the joinpoint regression model (version 4.7.0.0; Statistical Research and Applications Branch, National Cancer Institute). The intergroup differences were compared with the Fisher exact test or the Kruskal-Wallis test in applicable situations. A $P<.05$ was considered statistically significant. We used Prism 9 for macOS (version 9.5.0 [525]; GraphPad Software) to conduct statistical analysis and create figures.

XSL•FO
RenderX

Ethical Considerations

We used publicly available, anonymized data that can be accessed without requiring any special permissions. Since the data are aggregated and publicly accessible, institutional review board approval was waived.

Results

The Data on Trends in UI Topics Are Currently Accessible

We summarized the total UI search requests on Baidu over the past 12 years with the described search keywords. Based on the analysis of these 18 keywords, three major topics were recognized for initial classification: (1) complaint with 1 search keyword, (2) inquiry with 3 search keywords, and (3) treatment with 4 search keywords (see Multimedia Appendix 1). The cumulative BSI of the UI keywords inside the trend module amounted to 11,472,745. Across all topics, the total value of BSI in 13 years was 8,174,724 in the complaint category, 1,671,387 in the inquiry category, and 1,626,634 in the treatment category.

Based on the annual average BSI count for each UI topic, it was observed that search requests in each topic underwent a comparable period of growth, culminating in a baseline returning trend by the conclusion of 2023. The particular APCs for the topic Complaint were 1.7% ($P<.05$) from 2011 to 2021 and −7.9% ($P<.05$) from 2021 to 2023, while AACP was 0.1% ($P<.05$). For the topic Inquiry, the APCs were 16.0% ($P<.05$) from 2011 to 2016, –27.00% from 2016 to 2019, and 21.2% ($P<.05$) from 2019 to 2023, with an AAPC of 4.8%. Regarding the topic Treatment, the APC was 20.3% from 2011 to 2018 ($P<.05$), –36.9% from 2018 to 2021 ($P>.05$), and 2.2% from 2021 to 2023, with an overall AAPC of −0.4%. The APC trend, the joint-point, and the AAPC over time were calculated and demonstrated in Figure 1.

Figure 1. Online interest in urinary incontinence topics since 2011. (A) Search trend of topics about urinary incontinence. (B) Sum BSI proportion of each topic about urinary incontinence. APC: annual percentage change; BSI: Baidu search index.

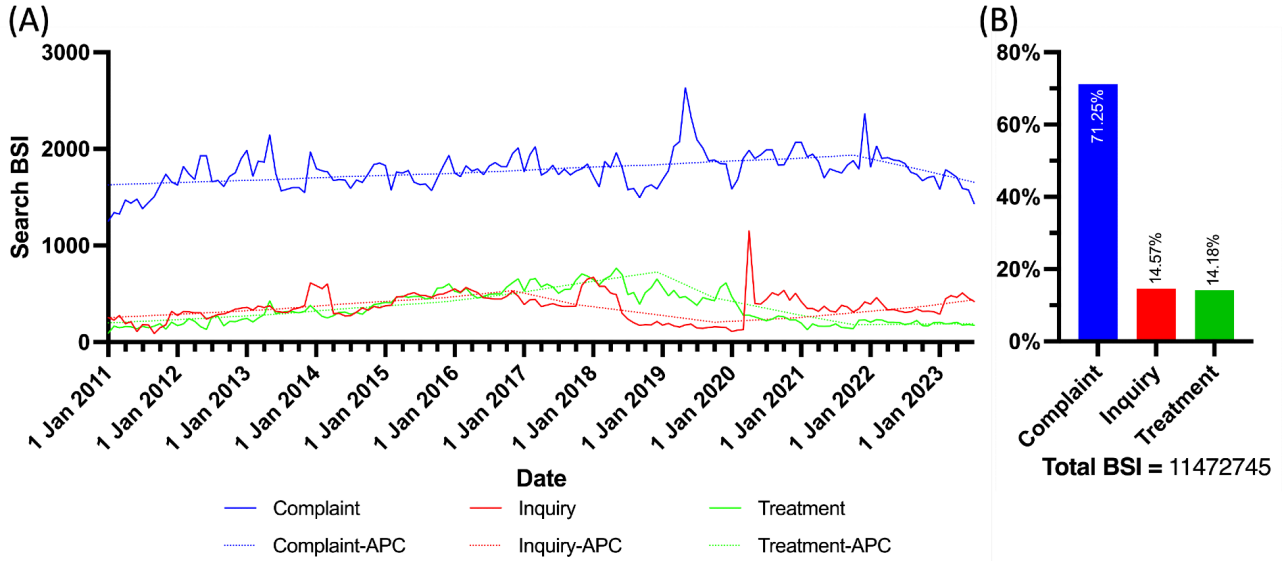

Geographic Disparities

The 13-year regional BSI proportions for each UI topic accompanied by a valid search record are shown in Figure 2. East Chinese users comprised the majority of search query contributors, accounting for 33.18%, 31.22%, and 32.07% in the Complaint, Inquiry, and Treatment categories, respectively. The regions of North China and Central China were next in succession. The demand levels in the Northeast, West China, and South China were comparable, each contributing 12%. Northwest demand rates were last, contributing 6.7%, 7.21%, and 7.19% to each subject.

XSL•FO
RenderX

Figure 2. Online interest in urinary incontinence topics since 2011. (A) Annual trend of the BSI for each region for the topic of Complaint. (B) Total search rates of each area for the topic of Complaint. (C) Annual trend of the BSI for each region for the topic of Inquiry. (D) Total search rates for each area for the topic of Inquiry. (E) Annual trend of the BSI for each region for the topic of Treatment. (F) Total search rates for each area for the topic of Treatment. BSI: Baidu search index.

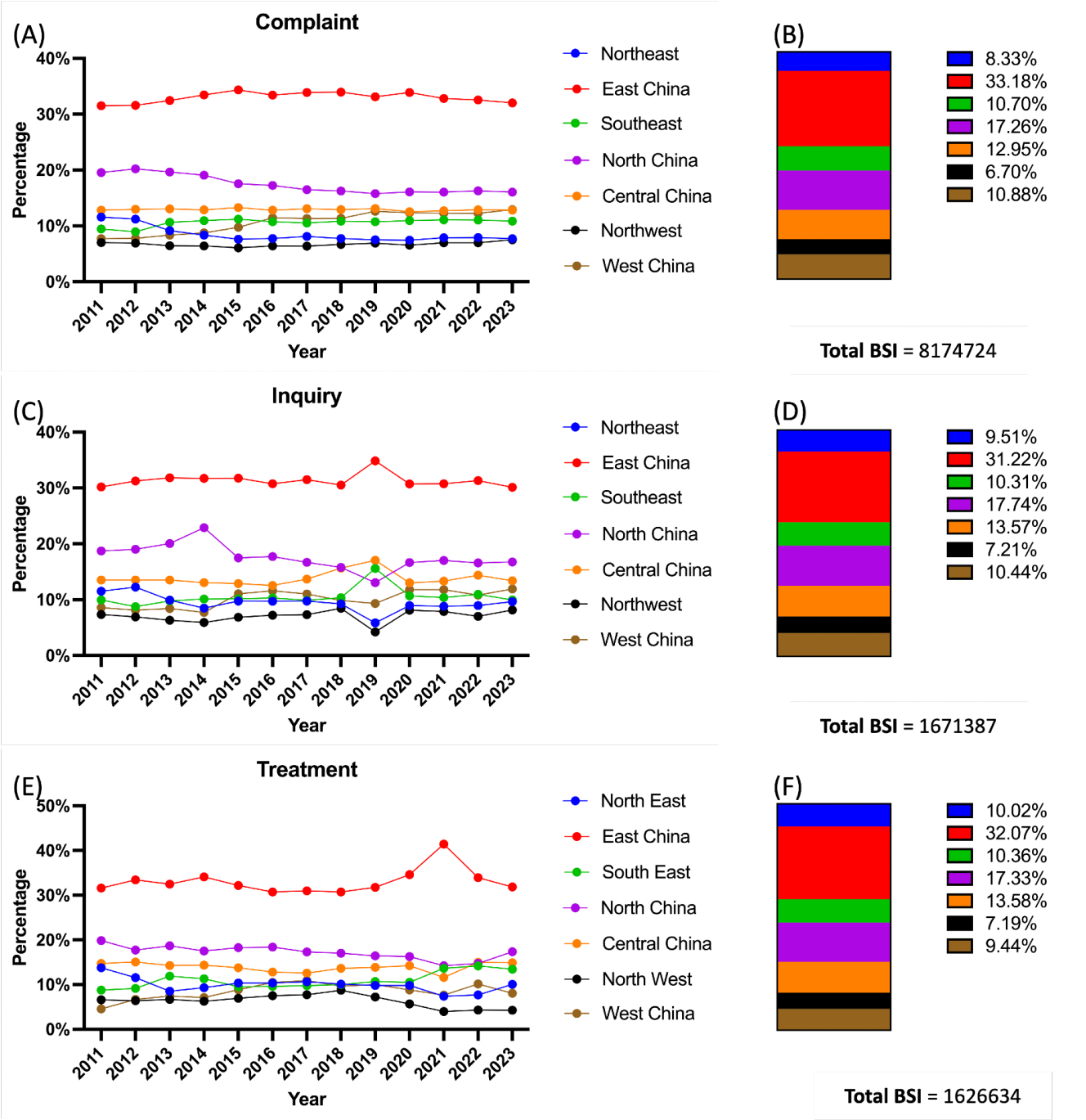

Demographic Differences

From the topic-specific demographic distribution analysis, searches varied in age and gender distribution. Regarding the Complaint topic, users aged 20 - 29 and 30-39 years made 36.01% and 32.59% of requests. Similarly, most of the requests in the Inquiry topic came from people aged 30-39 years (33.23%), followed by people aged 20-29 years (22.72%), and finally people aged 40-49 years (21.38%). In the Treatment topic, more than 51.79% of requests originated from users over 50, and this trend continued, with only 7.3% of requests coming from users under 19 years of age. In contrast, requests from this age group were below 10% in other topics. Regarding gender disparities, women made the majority of the total requests for each subject. Although the demand for Treatment was very low, around 60% of the requests were still made by female users (Figure 3).

Furthermore, we discovered that some gender-specific and older adult–restricting information was indicated in particular search keywords. Herein, an additional topic in the female, geriatric, and male categories was sorted for examination. Most requests (40.09%) for the female topic came from the 30 - 39 year old demographic. For the male issue, requests from the 30 - 39 year old group ranked highest, but they also shared a similar

XSL•FO
RenderX

proportion, at 20% - 30%, with the second- and third-highest groups. Regarding the geriatric subject, it was seen that a majority of the requests, namely 60%, were made by users aged between 20-29 and 30-39 years. Additionally, it was found that female users made 71.47% of these requests. In contrast, requests from the 40-49 and ≥50 age groups were just 13.91% and 13.91%, respectively (Figure 4).

Figure 3. Population disparities in search preference for urinary incontinence topics. (A) Age disparities in search preference for urinary incontinence topics. (B) Gender disparities in search preference for urinary incontinence topics. BSI: Baidu search index.

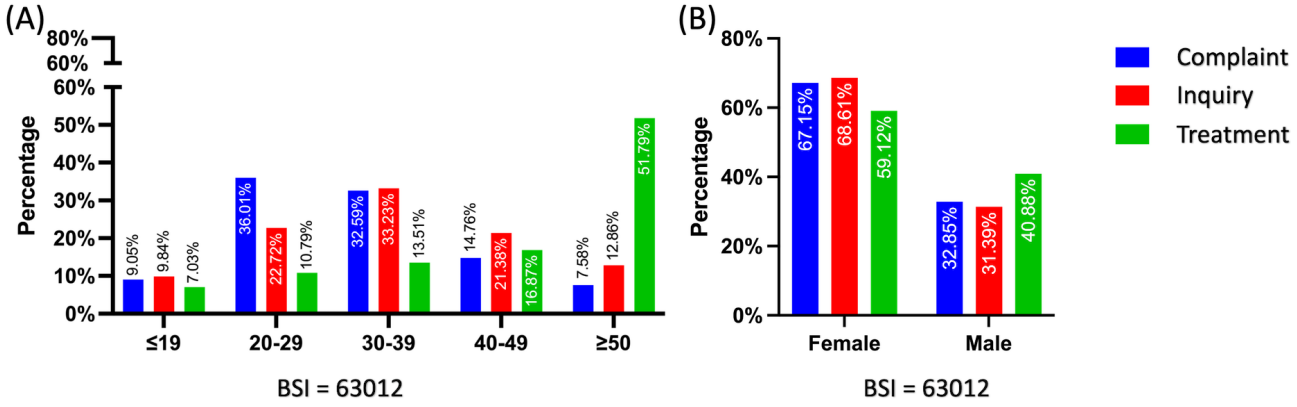

Figure 4. Population disparities in search preference for urinary incontinence topics specific to the female, male, and geriatric populations. (A) Age disparities in search preference for urinary incontinence topics specific to different populations. (B) Gender disparities in search preference for urinary incontinence topics specific to different populations. BSI: Baidu search index.

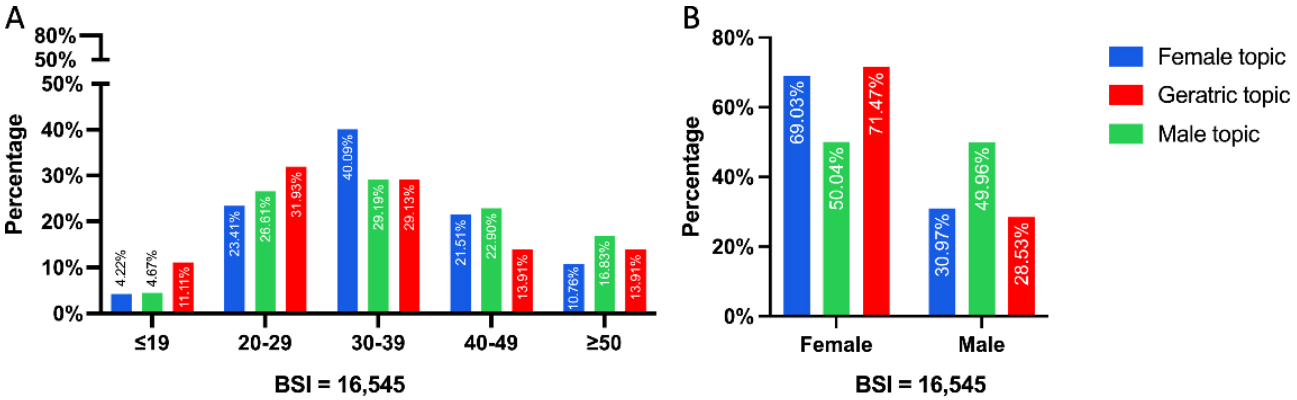

Keywords, Related Terms, and Search Frequency

During the studied period, 7436 of the 9310 words that appeared were classified as valid user demands related to the topic. The total BSI for these user demand terms was 190,037,698, accounting for only 15.74% (190,037,698/1,207,203,466) of the overall user interest in the topic of UI. Diagnosis-related issues were the most frequent, making up 38.22% (2842/7436) of the total. Treatment- and decision-related issues, as well as complaints, followed. In terms of popularity, health issues received the most attention, accounting for 24.09%. The popularity ratios for the top 3 frequent themes were 14.67%, 14.90%, and 5.60%. This suggests that while UI requests are steady, their value might be overshadowed by other health issues that spike in search popularity. Detailed distributions of these related terms and their search frequencies are depicted in Figure 5. Tables 1 and 2 present the top 3 related terms that were most frequently searched.

XSL•FO
RenderX

Figure 5. Term categories related to urinary incontinence in the Baidu index demand graph. (A) The distribution of most-appearing related words (word units) that users inquired about in the Baidu index related to urinary incontinence. (B) The distribution of most-searched related words that users inquired about in the Baidu index related to urinary incontinence. BSI: Baidu search index; TCM: traditional Chinese medicine.

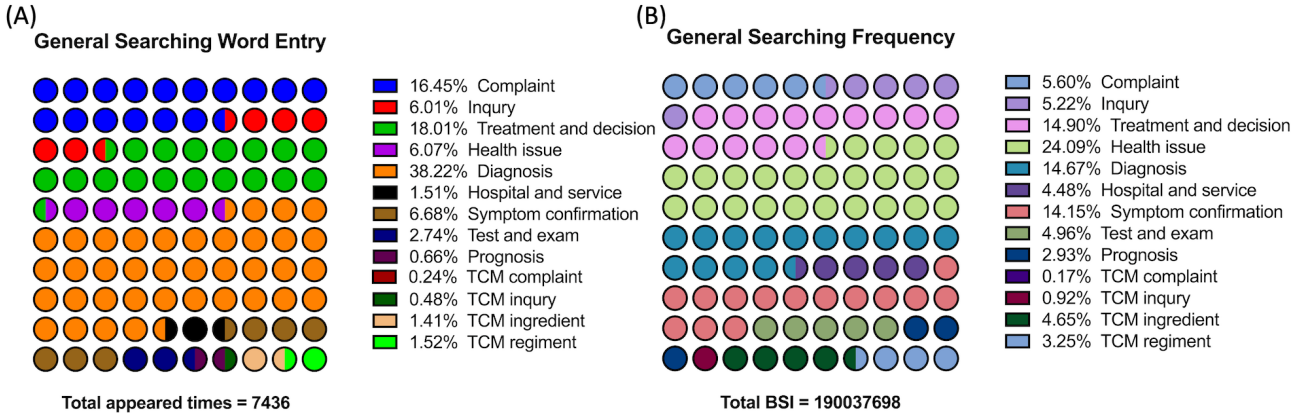

XSL•FO
RenderX

Table . The top 3 most-appearing related words (word units) that users inquired about in the Baidu index related to urinary incontinence.

Category[a]	Term 1	Number of times term appears	Term 2	Number of times term appears	Term 3	Number of times term appears
A	漏尿 (Urinary leakage)	203	尿潴留 (Urinary retention)	112	尿急 (Urinary urgency)	67
B	漏尿是什么原因造成的 (What is the cause for urinary leakage?)	101	尿失禁的原因 (What is the cause for urinary incontinence?)	60	女性漏尿是什么原因 (What is the cause for female incontinence?)	32
C	尿失禁怎么治疗 (How to treat urinary incontinence)	79	凯格尔运动 (Kegel exercise)	50	米多君 (Midodrine)	39
D	盆底肌 (Pelvic muscle)	37	尿道括约肌 (Sphincter urethra)	34	子宫 (Uterus)	14
E	尿失禁 (Urinary incontinence)	421	压力性尿失禁 (Stress incontinence)	258	急迫性尿失禁 (Urge incontinence)	198
F	泌尿外科 (Department of Urology)	4	浙江省中医院 (Zhejiang Provincial Hospital of TCM[b])	3	北京大学第一医院 (The First Affiliated Hospital of Peking University)	2
G	尿路感染的症状 (Symptoms of urinary tract infection)	33	前列腺增生的症状 (Symptoms of benign prostate hyperplasia)	26	子宫脱垂的症状 (Symptoms of uterus prolapse)	21
H	PSA (Prostate-specific antigen)	5	尿常规能检查出什么 (Clinical indications from urine routine test)	5	膀胱镜 (Cystoscope)	5
I	前列腺炎有什么症状和危害 (Hazards and symptoms of prostatitis)	21	尿酸高有什么严重后果 (Severe consequence of hyperuricemia)	2	憋尿的危害 (Hazard of pee holding)	2
J	伤寒 (Febrile disease)	2	舌苔黄腻 (Yellow and greasy tongue fur)	2	女性肾虚 (”Shen” essence insufficiency of ”Yin” in female)	1
K	肾虚的表现症状有哪些 (Top 10 manifestations of ”Shen” essence insufficiency)	4	中医养生项目 (Body maintenance project of TCM)	1	中医怎么减肥 (How to lose weight with TCM)	1
L	枇杷叶煮水的功效与作用 (Efficacy of folium eriobotryae decoct)	6	五味子的功效与作用 (Efficacy of *Schisandra chinensis*)	4	南瓜子的作用与功效 (Efficacy of cushaw seed)	3
M	补中益气丸的功效和作用 (Efficacy of “BuZhongYiQiWan” pill)	9	金匮肾气丸 (“JinKuiShenQiWan” pill)	9	金匮肾气丸的作用与功效 (Efficacy of “JinKuiShenQiWan”)	5

[a]A: complaint; B: etiology and causes; C: treatment and pharmaceutical; D: health care–related terms; E: diagnosis; F: health care services and commodities; G: diagnosis confirmation; H: test and exam; I: prognosis; J: TCM diagnosis; K: TCM diagnosis confirmation; L: TCM regimen; M: TCM remedy and materials.

[b]TCM: traditional Chinese medicine.

XSL•FO
RenderX

Table . The top 3 most searched related words that users inquired about in the Baidu index related to urinary incontinence.

	Term 1	BSI	Term 2	BSI	Term 3	BSI
A	漏尿 (Urinary leakage)	1,317,720	尿潴留 (Urinary leakage)	1,260,048	阴道前壁膨出 (Anterior vaginal wall prolapse)	651,482
B	尿频繁是什么原因 (What are the causes of urinary frequency?)	1,454,404	尿频尿急尿不尽是什么原因造成的 (What are the causes of urinary frequency, urgency, and incomplete voiding?)	1,323,360	漏尿是什么原因造成的 (What are the causes of urinary leakage?)	862,462
C	凯格尔运动 (Kegel exercise)	1,919,912	甲钴胺片的功效与作用 (The efficacy of Mecobalamin tablet)	1,590,308	提肛运动怎么做才正确 (How to do anus lifting [Kegel] correctly)	1,461,760
D	知网 (CKNI.com)	18,351,618	各城市首轮感染高峰期预测 (Prediction of pandemic peak in each city)	3,739,154	新冠感染5次必死 (Certain death after 5 times of covid-19 infection)	2,141,998
E	前列腺炎 (Prostatitis)	2,585,588	尿失禁 (Urinary incontinence)	2,257,708	尿路感染 (Urinary tract infection)	1,631,582
F	京东 (JD.com)	7,659,590	好大夫在线 (Haodaifu.com)	92,324	医生 (Doctors)	82,880
G	尿路感染的症状 (What are the symptoms of urinary tract infections?)	4,361,232	尿毒症的早期症状 (What are the early symptoms of uremia?)	2,467,396	糖尿病的症状有哪些 (What are the symptoms of diabetes mellitus?)	1,947,950
H	抗原检测怎么做 (How to do antigen testing?)	3,788,050	核酸检测多久出结果 (How long does it take for nucleic acid testing?)	1,006,772	HPV (HPV [human papillomavirus])	315,724
I	前列腺炎有什么症状和危害性 (What are the hazards and symptoms of prostatitis?)	5,092,960	肛门口摸到个软软的肉要紧嘛 (Is it consequential to notice a palpable perianal mass?)	130,732	肝囊肿是怎么回事?有危险吗 (What is a hepatic cyst, is it dangerous?)	84,436
J	脾虚 (Pi [spleen] insufficiency)	64,268	伤寒 (Typhoid fever)	29,578	肾阳虚 (‘Shen’ essence insufficiency of ‘Yang’)	26,398
K	肾虚的表现症状有哪些 (What are the symptoms of insufficiency in Shen essence?)	608,124	风热感冒和风寒感冒的症状区别 (Symptom differentiation of FengRe [heat wind origin] fever and FengHan [cold wind origin] cold)	433,660	湿热疹怎么治疗能除根 (How to eradicate “humid heat” rashes?)	84,050
L	陈皮的功效与作用 (The efficacy of tangerine peel)	853,524	石斛的功效与作用 (The efficacy of dendrobium)	808,556	黄芪的功效与主治 (The efficacy and indication of using astragalus membranaceus)	790,142
M	补中益气丸的功效和作用 (The efficacy of “BuQiYiZhongWan” pill)	935,942	金匮肾气丸 (“JinKuiShenQiWan” pill)	647,778	六味地黄丸 (“LiuWeiDiHuangWan” pill)	559,834

[a]A: complaint; B: etiology and causes; C: treatment and pharmaceutical; D: health care–related terms; E: diagnosis; F: health care services and commodities; G: diagnosis confirmation; H: test and exam; I: prognosis; J: traditional Chinese medicine diagnosis; K: traditional Chinese medicine

XSL•FO
RenderX

diagnosis confirmation; L: traditional Chinese medicine regimen; M: traditional Chinese medicine remedy and materials.

Assessment of the Content Quality of the Baidu Encyclopedia

On the Baidu Encyclopedia platform, 40 chapters were identified and titled with UI. There were 32 chapters accessible for content reference, and each of these chapters could be matched to one of the 5-point scale standards of the global quality score system based on their content quality. There were 8 chapters removed from the analysis due to their book-introducing nature. Disparities in the accessibility of the English title and video link were reported. The number of sections, word count, and number of page visits differed significantly (Table 3).

Table . Characteristics of the urinary incontinence content on Baidu Encyclopedia.

	GQS[a] 1 (N=13)	GQS 2 (N=7)	GQS 3 (N=4)	GQS 4 (N=3)	GQS 5 (N=5)	*P* value
English title						.01[b]
Yes	5	0	0	0	4	
No	8	7	4	3	1	
Video link						.04[b]
Yes	1	1	0	1	4	
No	7	6	4	2	1	
Number of sections, median (IQR)	0 (0-0)	2 (2-2.5)	5 (3-7.75)	6 (6-7)	6 (6-7)	<.001[c]
Word count, median (IQR)	105 (50-252)	707 (610-780)	1334 (931.25 - 1788.5)	3002 (2110 - 3339.5)	1909 (1906-4633)	.002[c]
Number of page visits, median (IQR)	4789 (1228-7592)	27,114 (15,587.5 - 68,407)	8203 (6480.25 - 20,747)	21,288 (14,095.5 - 131,130)	329,725 (127,846-473,125)	<.001[c]

[a]GQS: global quality score.

[b]Fisher exact test.

[c]Kruskal-Wallis test.

Discussion

Principal Findings

Using the most popular search engine, Baidu, made it possible to conduct research with the broadest conceivable national scope and depth [21]. With billions of Chinese users submitting requests daily, it is possible to identify and address the care-seeking demand and decision-making conundrum quickly. To our knowledge, this study represents the initial foray into infodemiology research to elucidate patients' perceptions and awareness of UI from a clinical and health care standpoint, drawing upon data from billions of users [19,26]. By analyzing a comprehensive dataset spanning 13 years at a national level, it is possible to uncover valuable insights into the prevailing trends, topics of interest, user search behavior, disease recognition, and demographic characteristics.

Within UI issues, topics can be created focusing solely on complaints, inquiries, and treatments using current search keywords. Interest in symptom complaints made up 71.25% of the total search volume, involving over 11 UI-related keywords. Inquiries and treatment requests comprised 3 words (14.57%) and 4 words (14.18%), respectively. Despite the popularity of UI complaint searches, it is important to note that there were specific types of UI complaints, such as those related to postpartum, urgency, older adult patients, and gender-specific conditions. This indicates that the Chinese population has a detailed and specific recognition of UI, requiring professional attention in each subfield. For inquiries and treatments, the search keywords were broader, focusing on exploring causes and potential remedies.

There has been less research focused on exploring the prognosis of UI compared to earlier studies on other subjects [21,26,27]. Several variables may underpin this finding. First, compared to other commonly experienced lower urinary tract symptoms such as frequency, urgency, and incomplete urination, UI exhibits a higher degree of predictability and controllability regarding its onset and occurrence. By definition, the most dominant UI types—SUI and mixed urinary incontinence—addressed the feature of "involuntary urine leakage when straining or during exerting efforts" [28,29]. Dealing with a UI is contingent upon and subordinate to a specific preceding action.

Further, in other UI types, the symptoms tended to be overwhelmed by other stronger senses, such as urgency in urgent UI or being mostly asymptomatic as with neurogenic UI or true incontinence [28,29]. Hence, acquiring knowledge about the diagnosis, etiology, and treatment options for UI is essential for addressing the concerns of those affected by this condition. Consequently, medical practitioners must prioritize raising awareness and promoting the recognition of UI prognosis.

In this study, several search keywords were identified as inquiries about etiology and treatment options, but these words were mere variations of a simple open-ended question. Though self-assisted cyber diagnosis predisposing individuals to delay treatment has been a concern for a long time [30], the limited

XSL•FO
RenderX

popularity of searching for these 2 topics shows that the subjective treatment goal of UI has yet to be recognized and established [31]. Therefore, local urologists still have the opportunity to fill these shoes.

The search trends for the 3 UI topics showed some similarities with an initial growth phase. The complaint topic maintained an increasing trend for 10 years, with an APC of only 1.7%, but a steady demand due to high initial search volumes and an average annual increase of 0.1%. UI inquiries and treatments were more variable. Inquiries showed sharp growth from 2016, peaking and then gradually declining by 2019. Treatment searches saw a strong increase over 8 years with an APC of 20.3%, but then the APC dropped sharply by 36.9% from 2018 to 2021. Fluctuations in search popularity may be linked to public reactions to controversies and criticisms of Baidu's marketing strategies [30,32]. However, during the initial COVID-19 outbreak in Wuhan, China, Baidu remained a key source of reliable information [10,18]. Hence, daily searches for UI topics remained steady at 2000 BSI, indicating a stable interest in general UI information.

Considering the geographic distribution, East China led UI searches across all 3 topics, averaging 30% of the total requests over the past 13 years. North China followed, with an average of 17%. The Northwest region consistently had the lowest average, accounting for 7% of total UI requests, which is lower than other regions, which averaged around 10%. The high number of requests from East China and North China aligns with our previous findings on the premature ejaculation issue [21] due to higher economic levels and population density reported by [33-35]. According to Xue et al, the reported UI frequency in the female population ranged from 8.7% to 69.8%, representing approximately 43‑439 million people [4]. Additionally, the prevalence of UI in males ranges from 17.3% to 35%, according to the existing literature [36]. However, it is important to note that the available data may only provide a limited understanding of the overall prevalence of male UI nationally. The focus of medical professionals on this issue is significantly determined by the problematic UI situation revealed in [2], and there is a shortage of comprehensive surveys on this topic. Less than 25% of survey participants sought medical consultation for UI [36]. Therefore, traditional epidemiological methods may not adequately capture the full scope of morbidity among individuals affected by UI, which often remains overlooked as a clinical issue.

UI is a prevalent clinical condition that impacts individuals of both genders. The primary symptoms of UI vary between genders and age groups, as influenced by the underlying mechanisms associated with each clinical subtype [28,29,37]. However, these clinical conditions share similarities and primarily involve involuntary urine leakage, which is not widely recognized by the general public. This is evident in the users' search behavior, as indicated by the prevalence of related search terms. Our research shows that the majority of UI inquiries come from women, with the proportions of complaints, inquiries, and treatment-related themes being 67.15%, 68.61%, and 59.12%, respectively. The research focus among individuals aged 20‑29 and 30‑39 years primarily centers around urinary system disorders and their underlying mechanisms, comprising about 60% of research efforts within these age groups. Yet, most searches for disease treatment were from those over 50, accounting for 51.79% of all searches. This result suggests that disparities and inequalities in sociodemographic factors are a significant concern. Although relevant domestic research in Mainland China is lacking, a recent population-comparable study found that sociodemographic characteristics such as age, education, family income, and parity may have had a role in the reported occurrence rate of UI [38,39].

In simpler terms, varying levels of education and socioeconomic status affect how people perceive beliefs, tolerate UI symptoms, and seek health care. Additionally, aging exacerbates UI and is widely recognized as a significant risk factor [40]. Moreover, socioeconomic factors greatly influence web-based search behavior on platforms such as Baidu. People from higher socioeconomic backgrounds generally have better internet access and digital literacy skills, enabling them to search for health information more effectively [41,42]. They tend to use precise medical terminology and look for reputable sources and research. Conversely, individuals from lower socioeconomic backgrounds face barriers like limited internet access, lower digital literacy, and less trust in web-based health information [12]. They may rely on less accurate sources, and their search queries might be more symptom-focused than diagnostic. Cultural and societal influences also shape which health issues people are willing to search for and discuss openly. Higher educational attainment often leads to more proactive information-seeking behaviors and a better understanding of complex health issues. In contrast, those with lower education levels may struggle with medical terminology and have difficulty conducting effective searches [42,43].

Further analysis revealed differences in UI preferences and understanding among various groups. The retrieved search keywords indicated gender- and population-specific demands, showing terms related to men, women, postpartum, and geriatric UI issues. The demographic distribution of these specific UI concerns was also highlighted. As expected, female users made the most inquiries across all demographic-specific UI topics, including those concerning men. In terms of age distribution, the largest group of users was aged 30‑39 years, making up 40.09% of the total. This was followed by those aged 20‑29 and 40‑49 years, who accounted for 29.19% and 29.13%, respectively.

Unlike our previous study on premature ejaculation, which is a male issue with concerns coming mainly from men [21], UI concerns and demands come mostly from women. Intuitively, one would expect patients to be most concerned about their conditions. However, the significant efforts and attention from women within the family context have not been fully recognized. Women play a prominent role in family health care decisions, making about 80% of these choices [44]. Therefore, concerns and demands from women, who may be the wives of husbands or daughters of older adult patients, should not be overlooked. As a result, even content related to male UI issues should consider a women-friendly perspective.

Diagnosis-related issues were the most frequent, making up 38.22% of the total. Treatment- and decision-related issues, as

XSL•FO
RenderX

well as complaints, followed. In terms of popularity, health issues received the most attention, accounting for 24.09%. The popularity percentages for the top 3 most frequent themes were 14.67%, 14.90%, and 5.60%, respectively. This suggests that while UI requests are steady, their value might be overshadowed by other health issues that spike in search popularity.

Among the most related words, "urinary leakage," "urinary retention," and "urinary urgency" were the most common complaints. In terms of search popularity, "anterior vaginal wall prolapse" ranked third among complaint words. Although these terms relate to UI symptoms, it is unclear if they reflect users' inquiries or complaints about a symptom. Since different types of UI have similar symptoms but different causes, accurate and thorough UI assessments are needed to prevent misdiagnosis and delayed treatment.

Detrusor hyperactivity with impaired contractile function is a UI-mimicking condition [45]. Characterized by involuntary urine leakage and urgency, the primary cause of this condition is a lack of detrusor contractility. The treatment for this condition is fundamentally different from that of other UI phenotypes. According to a prior prevalence survey on UI misdiagnosis, 24% of older adult nursing home residents have SUI misdiagnosed [46]. Although these facilities offer professional health care services, such conditions cannot be confirmed without implementing cystometry or multichannel videourodynamic evaluation. Therefore, a worse situation of more diagnosis delay and mistreatment could occur if the potential patient only seeks out a doctor via the internet based on their perception of symptoms.

In the treatment topic, we noticed the related terms of Kegel exercise and anus lifting (Tables 1 and 2) were most frequently inquired about with the highest popularity. Pelvic floor muscle training is considered the primary treatment option for UI since it has been demonstrated to effectively reduce episodes of urine leakage and enhance quality of life [47]. Kegel exercises are considerably more popular than other therapies [47] due to their adaptability, simplicity of implementation, and low cost. However, the pelvic strengthening exercises could only moderate the weakened pelvic muscle for better stabilization and hence could mostly be effective in lower-grade UI with moderate symptoms or alleviate severe symptoms of a particular phenotype [31,47,48].

Users who search for information on Baidu can find answers to their questions in Baidu's official encyclopedia [49]. For UI issues, there were 40 pages describing UI issues ranging from general aspects to certain specific treatment options. Though the content quality varies, the "must knows" of UI, such as its phenotypes, etiologic features, and population-specific issues, were included. Although the convenience of an online encyclopedia platform is well-recognized, the quality and dependability have yet to be sufficiently accurate in resolving confusion and providing health advice [49,50]. Similar to the findings of Handler et al [51], we found that the material on the Baidu platform of popular science is characterized by a high level of detail. However, it is essential to note that this information lacks proper categorization and even includes some inaccurate titles. As a result, not only does this situation impose additional difficulties, but it also has the potential to mislead users, with potentially severe repercussions.

Given these considerations, medical professionals should anticipate patients' needs and identify relevant information. Adjustments in therapeutic sessions, communication, and public education are essential to address patients' issues effectively.

There are several limitations to this study. Although Baidu is the largest search platform in Mainland China, other search engines and social media platforms are increasingly competing, which means the daily search-based BSI may not capture all online user demands. Additionally, regional differences in user preferences and political restrictions may exclude search queries from local citizens who use international search platforms legally. Furthermore, as a Chinese-based internet service, requests from foreign languages or non-Chinese minority groups are likely to be systematically excluded. Additionally, all data from the Baidu Index platform were presented in a fixed and finalized format, limiting the ability to further manipulate or reorganize the data for more flexible structural customization.

To the best of our knowledge, this study represents the initial exploration of infodemiology to address the issue of assessing public concern in the Chinese-speaking population around UI. We chose to investigate the popularity of UI because of its complex nature and the potential for a lack of interest, which can lead to hesitation and delayed treatment. By examining current trends and demand, we can understand prevailing patterns and identify patients' main concerns. This insight allows for adjustments to health care strategies and responses to enhance Medicare services.

Conclusion

Web-based UI searching is continuous and traceable since January 2011. Different categorized themes within the UI topic highlight specific demands from various populations, necessitating tailored responses. Although web-based platforms can offer answers, medical professionals' involvement is crucial to avoid misdiagnosis and delayed treatment.

Acknowledgments

Artificial intelligence or chatbot tools were solely used for language editing in this manuscript. No content generation or ideation were originated by the artificial intelligence or chatbot.

XSL•FO
RenderX

Data Availability

The datasets generated and analyzed during this study, including search trend and demographic, are available from Baidu index platform and the corresponding author upon reasonable request.

Authors' Contributions

This study was conceptualized by SW. The methodology was designed by SW. The investigation was carried out by LS, DL, and LX. CH, DL, and LS wrote the original draft. SW, LX, and WX reviewed and edited the draft. Funding was acquired by SW. WX and CH contributed resources, and SW and LX supervised the study.

Conflicts of Interest

None declared.

Multimedia Appendix 1
List of keywords used in composite search index.
[DOCX File, 16 KB - formative_v9i1e55670_app1.docx]

References

1. Bo K, Frawley HC, Haylen BT, et al. An International Urogynecological Association (IUGA)/International Continence Society (ICS) joint report on the terminology for the conservative and nonpharmacological management of female pelvic floor dysfunction. Int Urogynecol J 2017 Feb;28(2):191-213. [doi: 10.1007/s00192-016-3123-4] [Medline: 27921161]
2. Gandi C, Sacco E. Pharmacological management of urinary incontinence: current and emerging treatment. Clin Pharmacol 2021;13:209-223. [doi: 10.2147/CPAA.S289323] [Medline: 34858068]
3. Dominguez-Antuña E, Diz JC, Ayán C, Suárez-Iglesias D, Rodríguez-Marroyo JA. Prevalence and severity of urinary incontinence among male and female competitors and recreational CrossFit practitioners. Eur J Obstet Gynecol Reprod Biol 2022 Sep;276:144-147. [doi: 10.1016/j.ejogrb.2022.07.014] [Medline: 35905542]
4. Xue K, Palmer MH, Zhou F. Prevalence and associated factors of urinary incontinence in women living in China: a literature review. BMC Urol 2020 Oct 14;20(1):159. [doi: 10.1186/s12894-020-00735-x] [Medline: 33054777]
5. Kozomara-Hocke M, Hermanns T, Poyet C. Male urinary incontinence--a taboo issue. Praxis (Bern 1994) 2016 Mar 2;105(5):269-277. [doi: 10.1024/1661-8157/a002297] [Medline: 26934011]
6. Luo Y, Zou P, Wang K, Li X, Wang J. Prevalence and risk factors of urinary incontinence among elderly adults in rural China: a cross-sectional survey. J Wound Ostomy Continence Nurs 2022;49(1):78-86. [doi: 10.1097/WON.0000000000000829] [Medline: 35040817]
7. Coyne KS, Wein A, Nicholson S, Kvasz M, Chen CI, Milsom I. Economic burden of urgency urinary incontinence in the United States: a systematic review. J Manag Care Pharm 2014 Feb;20(2):130-140. [doi: 10.18553/jmcp.2014.20.2.130] [Medline: 24456314]
8. Milsom I, Coyne KS, Nicholson S, Kvasz M, Chen CI, Wein AJ. Global prevalence and economic burden of urgency urinary incontinence: a systematic review. Eur Urol 2014 Jan;65(1):79-95. [doi: 10.1016/j.eururo.2013.08.031] [Medline: 24007713]
9. Coyne KS, Wein A, Nicholson S, Kvasz M, Chen CI, Milsom I. Comorbidities and personal burden of urgency urinary incontinence: a systematic review. Int J Clin Pract 2013 Oct;67(10):1015-1033. [doi: 10.1111/ijcp.12164] [Medline: 24073974]
10. Osman NI, Chapple CR. The burden of urgency urinary incontinence on health and wellbeing. Int J Clin Pract 2013 Nov;67(11):1072-1074. [doi: 10.1111/ijcp.12206] [Medline: 24165420]
11. Wadensten T, Nyström E, Franzén K, Lindam A, Wasteson E, Samuelsson E. A mobile app for self-management of urgency and mixed urinary incontinence in women: randomized controlled trial. J Med Internet Res 2021 Apr 5;23(4):e19439. [doi: 10.2196/19439] [Medline: 33818395]
12. Elenskaia K, Haidvogel K, Heidinger C, Doerfler D, Umek W, Hanzal E. The greatest taboo: urinary incontinence as a source of shame and embarrassment. Wien Klin Wochenschr 2011 Oct;123(19-20):607-610. [doi: 10.1007/s00508-011-0013-0] [Medline: 21935649]
13. Hammad FT. Prevalence, social impact and help-seeking behaviour among women with urinary incontinence in the Gulf countries: a systematic review. Eur J Obstet Gynecol Reprod Biol 2021 Nov;266:150-156. [doi: 10.1016/j.ejogrb.2021.10.005] [Medline: 34653920]
14. Minassian VA, Yan X, Lichtenfeld MJ, Sun H, Stewart WF. The iceberg of health care utilization in women with urinary incontinence. Int Urogynecol J 2012 Aug;23(8):1087-1093. [doi: 10.1007/s00192-012-1743-x] [Medline: 22527544]
15. Gao GF. Infodemiology: the science studying infodemic and inforus. China CDC Wkly 2022 Dec 30;4(52):1181-1182. [doi: 10.46234/ccdcw2022.237] [Medline: 36779173]

XSL•FO
RenderX

16. Eysenbach G. Infodemiology: the epidemiology of (mis)information. Am J Med 2002 Dec 15;113(9):763-765. [doi: 10.1016/s0002-9343(02)01473-0] [Medline: 12517369]
17. Jiang L, Wei S, Sam Saji A, Li J, Che G. An analysis of public perception and concern toward electronic cigarettes: exploring attitudes and profiles. Cureus 2023 Oct;15(10):e47983. [doi: 10.7759/cureus.47983] [Medline: 38034205]
18. Jiang L, Ma Q, Wei S, Che G. Online public attention of COVID-19 vaccination in Mainland China. Digit Health 2022;8:20552076211070454. [doi: 10.1177/20552076211070454] [Medline: 35096408]
19. Guérin S, Khene ZE, Peyronnet B, Bensalah K, Zimmern PE. Google searches for stress urinary incontinence treatment options: a worldwide trend analysis in the synthetic mesh controversy era. World J Urol 2023 Aug;41(8):2217-2223. [doi: 10.1007/s00345-023-04474-0] [Medline: 37358599]
20. D'Ancona C, Haylen B, Oelke M, et al. The International Continence Society (ICS) report on the terminology for adult male lower urinary tract and pelvic floor symptoms and dysfunction. Neurourol Urodyn 2019 Feb;38(2):433-477. [doi: 10.1002/nau.23897] [Medline: 30681183]
21. Wei S, Ma M, Wen X, Wu C, Zhu G, Zhou X. Online public attention toward premature ejaculation in Mainland China: infodemiology study using the Baidu index. J Med Internet Res 2021 Aug 26;23(8):e30271. [doi: 10.2196/30271] [Medline: 34435970]
22. Langille M, Bernard A, Rodgers C, Hughes S, Leddin D, van Zanten SV. Systematic review of the quality of patient information on the internet regarding inflammatory bowel disease treatments. Clin Gastroenterol Hepatol 2010 Apr;8(4):322-328. [doi: 10.1016/j.cgh.2009.12.024] [Medline: 20060070]
23. SEER*Stat trends exercise 1a: examine trends in rates. National Cancer Institute. 2021. URL: https://seer.cancer.gov/seerstat/tutorials/trends1a/webprint [accessed 2025-05-01]
24. Fay MP, Tiwari RC, Feuer EJ, Zou Z. Estimating average annual percent change for disease rates without assuming constant change. Biometrics 2006 Sep;62(3):847-854. [doi: 10.1111/j.1541-0420.2006.00528.x] [Medline: 16984328]
25. Trend algorithms (for percent change and annual percent change). National Cancer Institute. URL: https://seer.cancer.gov/seerstat/WebHelp/Trend_Algorithms.htm [accessed 2025-05-01]
26. Wei S, Ma M, Wu C, et al. Using search trends to analyze web-based interest in lower urinary tract symptoms-related inquiries, diagnoses, and treatments in Mainland China: infodemiology study of Baidu index data. J Med Internet Res 2021 Jul 6;23(7):e27029. [doi: 10.2196/27029] [Medline: 34255683]
27. Frutos-Reoyo EJ, Luque-Linero P, Cantalapiedra-Puentes E, Mendi-Gabarain I, Bermejo-de la Fuente P, Candau-Pérez ED. Prognostic factors and treatment outcomes for female urinary incontinence rehabilitation. Actas Urol Esp (Engl Ed) 2023;47(6):376-381. [doi: 10.1016/j.acuroe.2023.01.007] [Medline: 36842707]
28. Gacci M, Sakalis VI, Karavitakis M, et al. European Association of Urology guidelines on male urinary incontinence. Eur Urol 2022 Oct;82(4):387-398. [doi: 10.1016/j.eururo.2022.05.012] [Medline: 35697561]
29. Nambiar AK, Arlandis S, Bø K, et al. European Association of Urology guidelines on the diagnosis and management of female non-neurogenic lower urinary tract symptoms. Part 1: diagnostics, overactive bladder, stress urinary incontinence, and mixed urinary incontinence. Eur Urol 2022 Jul;82(1):49-59. [doi: 10.1016/j.eururo.2022.01.045] [Medline: 35216856]
30. Hu X, Pan H, He W, Hua H, Yan Z. Evaluation of the content quality of websites for recurrent aphthous ulcers and oral lichen planus. BMC Oral Health 2017 Dec 29;17(1):170. [doi: 10.1186/s12903-017-0467-1] [Medline: 29284468]
31. Casteleijn FM, Zwolsman SE, Kowalik CR, Roovers J. Patients' perspectives on urethral bulk injection therapy and mid-urethral sling surgery for stress urinary incontinence. Int Urogynecol J 2018 Sep;29(9):1249-1257. [doi: 10.1007/s00192-018-3644-0] [Medline: 29675556]
32. Pogacar FA, Ghenai A, Smucker MD, Clarke CLA. The positive and negative influence of search results on people's decisions about the efficacy of medical treatments. 2017 Presented at: ICTIR '17; Oct 1-4, 2017; Amsterdam, The Netherlands. [doi: 10.1145/3121050.3121074]
33. Xu C, Wang Y, Yang H, et al. Association between cancer incidence and mortality in web-based data in China: infodemiology study. J Med Internet Res 2019 Jan 29;21(1):e10677. [doi: 10.2196/10677] [Medline: 30694203]
34. Hou Z, Du F, Zhou X, et al. Cross-country comparison of public awareness, rumors, and behavioral responses to the COVID-19 epidemic: infodemiology study. J Med Internet Res 2020 Aug 3;22(8):e21143. [doi: 10.2196/21143] [Medline: 32701460]
35. Wang T, Xia Q, Chen X, Jin X. Use of Baidu index to track Chinese online behavior and interest in kidney stones. Risk Manag Healthc Policy 2020;13:705-712. [doi: 10.2147/RMHP.S245822] [Medline: 32694922]
36. Cheng MC, Liu SP, Chuang YC, Hsu KCP, Chow PM. Prevalence and impacts of male urinary incontinence on quality of life, mental health, work limitation, and health care seeking in China, Taiwan, and South Korea (LUTS Asia): results from a cross-sectional, population-based study. Investig Clin Urol 2022 Jan;63(1):71-82. [doi: 10.4111/icu.20210259] [Medline: 34983125]
37. Arlandis S, Bø K, Cobussen-Boekhorst H, et al. European Association of Urology guidelines on the management of female non-neurogenic lower urinary tract symptoms. Part 2: Underactive bladder, bladder outlet obstruction, and nocturia. Eur Urol 2022 Jul;82(1):60-70. [doi: 10.1016/j.eururo.2022.01.044] [Medline: 35181193]

XSL•FO
RenderX

38. Sundqvist C, Li X, Sundquist K, Jansåker F. Sociodemographic disparities and parity in relation to urinary incontinence: a nationwide primary healthcare cohort study (1997-2018). J Clin Med 2022 Jan 19;11(3):496. [doi: 10.3390/jcm11030496] [Medline: 35159948]
39. Shaw C, Rajabali S, Tannenbaum C, Wagg A. Is the belief that urinary incontinence is normal for ageing related to older Canadian women's experience of urinary incontinence? Int Urogynecol J 2019 Dec;30(12):2157-2160. [doi: 10.1007/s00192-019-03906-z] [Medline: 30796477]
40. Fritel X, Panjo H, Varnoux N, Ringa V. The individual determinants of care-seeking among middle-aged women reporting urinary incontinence: analysis of a 2273-woman cohort. Neurourol Urodyn 2014 Sep;33(7):1116-1122. [doi: 10.1002/nau.22461] [Medline: 23818427]
41. Khan AS, Rahman AU, Qazi LT. The relationship between internet usage, socioeconomic status, subjective health and social status. BER 2016 Oct 30;8(Special Edition):67-82. [doi: 10.22547/BER/8.SE.5]
42. Willems S, De Maesschalck S, Deveugele M, Derese A, De Maeseneer J. Socio-economic status of the patient and doctor-patient communication: does it make a difference? Patient Educ Couns 2005 Feb;56(2):139-146. [doi: 10.1016/j.pec.2004.02.011] [Medline: 15653242]
43. Faber JS, Al-Dhahir I, Reijnders T, et al. Attitudes toward health, healthcare, and eHealth of people with a low socioeconomic status: a community-based participatory approach. Front Digit Health 2021;3:690182. [doi: 10.3389/fdgth.2021.690182] [Medline: 34713165]
44. Matoff-Stepp S, Applebaum B, Pooler J, Kavanagh E. Women as health care decision-makers: implications for health care coverage in the United States. J Health Care Poor Underserved 2014 Nov;25(4):1507-1513. [doi: 10.1353/hpu.2014.0154] [Medline: 25418222]
45. Yamamoto T, Sakakibara R, Uchiyama T, et al. Neurological diseases that cause detrusor hyperactivity with impaired contractile function. Neurourol Urodyn 2006;25(4):356-360. [doi: 10.1002/nau.20224] [Medline: 16532465]
46. Resnick NM, Brandeis GH, Baumann MM, DuBeau CE, Yalla SV. Misdiagnosis of urinary incontinence in nursing home women: prevalence and a proposed solution. Neurourol Urodyn 1996;15(6):599-618. [doi: 10.1002/(SICI)1520-6777(1996)15:6<599::AID-NAU2>3.0.CO;2-A]
47. Dumoulin C, Cacciari LP, Hay-Smith EJC. Pelvic floor muscle training versus no treatment, or inactive control treatments, for urinary incontinence in women. Cochrane Database Syst Rev 2018 Oct 4;10(10):CD005654. [doi: 10.1002/14651858.CD005654.pub4] [Medline: 30288727]
48. Cavkaytar S, Kokanali MK, Topcu HO, Aksakal OS, Doğanay M. Effect of home-based Kegel exercises on quality of life in women with stress and mixed urinary incontinence. J Obstet Gynaecol 2015 May;35(4):407-410. [doi: 10.3109/01443615.2014.960831] [Medline: 25264854]
49. Ma M, Yin S, Zhu M, et al. Evaluation of medical information on male sexual dysfunction on Baidu Encyclopedia and Wikipedia: comparative study. J Med Internet Res 2022 Aug 9;24(8):e37339. [doi: 10.2196/37339] [Medline: 35943768]
50. Sun F, Yang F, Zheng S. Evaluation of the liver disease information in Baidu Encyclopedia and Wikipedia: longitudinal study. J Med Internet Res 2021 Jan 18;23(1):e17680. [doi: 10.2196/17680] [Medline: 33459597]
51. Handler SJ, Eckhardt SE, Takashima Y, Jackson AM, Truong C, Yazdany T. Readability and quality of Wikipedia articles on pelvic floor disorders. Int Urogynecol J 2021 Dec;32(12):3249-3258. [doi: 10.1007/s00192-021-04776-0] [Medline: 33797592]

Abbreviations

AAPC: average annual percent change
APC: annual percent change
BSI: Baidu search index
SUI: stress urinary incontinence
TCM: traditional Chinese medicine
UI: urinary incontinence

Edited by A Mavragani; submitted 20.12.23; peer-reviewed by M Ma, T Kinoshita; revised version received 16.03.25; accepted 21.03.25; published 23.06.25.

Please cite as:
Lin S, Duan L, Lu X, Chao H, Wen X, Wei S
Internet Health Information–Seeking Trend of Urinary Incontinence in Mainland China: Infodemiology Study
JMIR Form Res 2025;9:e55670
URL: https://formative.jmir.org/2025/1/e55670
doi:10.2196/55670

XSL•FO
RenderX

© Shuangquan Lin, Lingxing Duan, Xiongbing Lu, Haichao Chao, Xi Wen, Shanzun Wei. Originally published in JMIR Formative Research (https://formative.jmir.org), 23.6.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Medical Misinformation in AI-Assisted Self-Diagnosis: Development of a Method (EvalPrompt) for Analyzing Large Language Models

Troy Zada[1*], BASc; Natalie Tam[1*], BASc; Francois Barnard[1], BASc; Marlize Van Sittert[2], BS; Venkat Bhat[3,4], MSc, MD; Sirisha Rambhatla[1], PhD

[1]Department of Management Sciences and Engineering, University of Waterloo, 200 University Avenue West, Waterloo, ON, Canada
[2]Faculty of Law, University of Toronto, Toronto, ON, Canada
[3]Department of Psychiatry, University of Toronto, Toronto, ON, Canada
[4]Interventional Psychiatry Program, St. Michael's Hospital, Unity Health Toronto, Toronto, ON, Canada
[*]these authors contributed equally

Corresponding Author:
Sirisha Rambhatla, PhD
Department of Management Sciences and Engineering, University of Waterloo, 200 University Avenue West, Waterloo, ON, Canada

Abstract

Background: Rapid integration of large language models (LLMs) in health care is sparking global discussion about their potential to revolutionize health care quality and accessibility. At a time when improving health care quality and access remains a critical concern for countries worldwide, the ability of these models to pass medical examinations is often cited as a reason to use them for medical training and diagnosis. However, the impact of their inevitable use as a self-diagnostic tool and their role in spreading health care misinformation has not been evaluated.

Objective: This study aims to assess the effectiveness of LLMs, particularly ChatGPT, from the perspective of an individual self-diagnosing to better understand the clarity, correctness, and robustness of the models.

Methods: We propose the comprehensive testing methodology evaluation of LLM prompts (EvalPrompt). This evaluation methodology uses multiple-choice medical licensing examination questions to evaluate LLM responses. Experiment 1 prompts ChatGPT with open-ended questions to mimic real-world self-diagnosis use cases, and experiment 2 performs sentence dropout on the correct responses from experiment 1 to mimic self-diagnosis with missing information. Humans then assess the responses returned by ChatGPT for both experiments to evaluate the clarity, correctness, and robustness of ChatGPT.

Results: In experiment 1, we found that ChatGPT-4.0 was deemed correct for 31% (29/94) of the questions by both nonexperts and experts, with only 34% (32/94) agreement between the 2 groups. Similarly, in experiment 2, which assessed robustness, 61% (92/152) of the responses continued to be categorized as correct by all assessors. As a result, in comparison to a passing threshold of 60%, ChatGPT-4.0 is considered incorrect and unclear, though robust. This indicates that sole reliance on ChatGPT-4.0 for self-diagnosis could increase the risk of individuals being misinformed.

Conclusions: The results highlight the modest capabilities of LLMs, as their responses are often unclear and inaccurate. Any medical advice provided by LLMs should be cautiously approached due to the significant risk of misinformation. However, evidence suggests that LLMs are steadily improving and could potentially play a role in health care systems in the future. To address the issue of medical misinformation, there is a pressing need for the development of a comprehensive self-diagnosis dataset. This dataset could enhance the reliability of LLMs in medical applications by featuring more realistic prompt styles with minimal information across a broader range of medical fields.

(JMIR Form Res 2025;9:e66207) doi:10.2196/66207

KEYWORDS

ChatGPT; health care; LLM; misinformation; self-diagnosis; large language model

XSL•FO
RenderX

Introduction

Background

Large language models (LLMs) have grown in popularity with an ever-expanding list of applications due to their efficiency and accessibility [1,2]. With their increased usage, LLMs are gaining user trust [3], partly due to the anthropomorphic responses produced by models such as GPT-4o, even though they can generate misinformation at scale [4,5]. Recent reports highlight the inability of differentiating truth from misinformation, and the potential collapse of health care systems, as major disruptors on the horizon [6]. This emphasizes the urgent need to develop solutions to ensure the delivery of factual information.

In health care, self-diagnosis through web searches has become a widespread practice and is especially important for underserved communities [7-10], which means that the prospective usage of LLMs in this domain is inevitable. However, relying solely on online searches for health information can result in severe misinformation as content on social media often spreads more rapidly than scientific knowledge [11]. Inaccurate content, conspiracy theories, and false claims are all forms of misinformation which can impact public perceptions, alter behaviors, and reduce trust in health care systems [12]. Moreover, the ongoing global shortage of health care workers [13-16] has driven government entities and health care organizations to explore the use of LLMs as health care assistants and expertise replacements for diagnosis and education [17-23]. Thus, there is a need to study the quality and reliability of LLM-generated responses to health care–related questions.

Prior Work

Recent work has focused on analyzing ChatGPT across various industries, including its application within health care. ChatGPT is a natural language processing model distinctive for its narrative response style to user input [24,25]. Studies have assessed its performance on examinations [20,24,26] and its utility as a self-studying tool [24,27,28], leveraging its ability to provide tailored responses and immediate feedback. Furthermore, ChatGPT has demonstrated potential in assisting research and academic writing by enhancing efficiency and mitigating gaps in researcher knowledge [29]. However, the increased usage of ChatGPT raises significant ethical concerns regarding plagiarism, bias, transparency, inaccuracy, and health equity [30-34].

ChatGPT has also demonstrated superior performance in answering medical questions compared to other LLMs [35]. Research has explored its applications in medical education, including its effectiveness on licensing examinations, tailored learning experiences, and comprehension of complex medical concepts and clinical reasoning [20,24,27,33,36]. Other areas of study have focused on identifying inefficiencies and inaccuracies within clinical workflows, medical research, and diagnoses, with the objective of integrating LLMs to optimize documentation, triage, and clinical data management procedures [33,37-44]. Additionally, investigations into diagnostic assistance have integrated patient questionnaires and medical imaging with LLMs [43,45-48]. Despite the apparent high performance of LLMs in health care, they remain inferior compared to the judgment of human clinicians [49].

In summary, prior work has generally evaluated the trustworthiness of LLMs [50], along with specifically examining their performance in medical situations. These methods have used the multiple-choice questions from the United States Medical Licensing Exam (USMLE) [51] to evaluate the capability of LLMs in attaining scores near the passing threshold of 60% [20,52]. However, these evaluations inaccurately depict the capability of LLMs for self-diagnosis. In practical situations where individuals use technology to self-diagnose, they would not include answers when posing questions and would not provide the same level of information as the examination questions.

Goal of This Study

This study critically examines the performance of LLMs in responding to health care–related questions. To achieve this, we propose evaluation of LLM prompts. This evaluation procedure contains detailed guidelines to assess ChatGPT's response to open-ended questions and validate the robustness of these responses using a sentence dropout method. This 2-staged approach, to our knowledge, is the first comprehensive strategy aimed at better understanding LLM responses and their implications for medical misinformation. We hypothesize that LLMs, particularly ChatGPT-4.0 (referred to as GPT-4.0), are currently unsuitable for self-diagnosis purposes since a significant portion of responses will be ambiguous or incorrect. In particular, this hypothesis is validated if GPT-4.0 surpasses a minimum threshold of 60% [20] for each of the following three questions:

1. Are the responses clear? This question can be answered by analyzing the response consistency.
2. Are the responses genuinely correct? This question can be answered by identifying the responses classified as correct by all assessors.
3. Are the responses robust? This question can be answered by conducting an ablation study on the correct responses.

Methods

Study Design

In this section, the considerations and preparations for the decided dataset are first specified. Then, the assessor procedures and guidelines are discussed along with the process used to analyze the output answers from ChatGPT. Finally, the complete testing methodology is introduced. The overall process uses both nonmedical and medical experts for assessment and is segmented into two experiments: (1) ChatGPT responses on USMLE step 1 open-ended prompts and (2) ChatGPT robustness and ablation study analysis. The GPT-4.0 responses were generated for each of the USMLE questions using Python, with the code and datasets available in the EvalPrompt (evaluation of large language model prompts) repository [53].

XSL•FO
RenderX

Dataset Considerations and Preparations

Overview

The USMLE [51] dataset consists of 3 steps undertaken by medical students throughout their program. Each step is a test consisting of multiple-choice, single-answer, and no-justification questions. Particularly, in this work, the questions are extracted from step 1 since the step 2 and step 3 questions are medically complex for the general population. Furthermore, only textual questions were kept, resulting in a dataset containing 94 single-answer questions that would be used for prompting ChatGPT. Textbox 1 [51] displays a sample of an extracted step 1 question.

Textbox 1. Sample question 3.1 is directly extracted from the United States Medical Licensing Exam step 1 test. This question is a multiple-choice question with a correct answer of (D).

Question: In a sample of 100 individuals, the mean leukocyte count is 7500/mm³, with a standard deviation of 1000/mm³. If the leukocyte counts in this population follow a normal (gaussian) distribution, approximately 50% of individuals will have which of the following total leukocyte counts?

(A) 5500–9500/mm³

(B) <6500/mm³ or >8500/mm³

(C) 6500–8500/mm³

(D) <7500/mm³

(E) >9500/mm³

Answer: (D) <7500/mm³

Baseline ChatGPT Answer Analysis

The first experiment established the foundation for all testing. From the initial multiple-choice questions, each question was transformed into an open-ended question to accurately simulate the circumstances of an individual interacting with ChatGPT. This process was accomplished by removing the multiple-choice options and replacing any instance of "which of the following" with "what." An example transformation is provided in Textbox 2 [51], where the original question from Textbox 1 was transformed into an open-ended question.

Textbox 2. Transformed United States Medical Licensing Exam step 1 test question based on the original Textbox 1 question. The transformation involves removing the (A)-(E) options and replacing the text "which of the following" with "what."

Question: In a sample of 100 individuals, the mean leukocyte count is 7500/mm³, with a standard deviation of 1000/mm³. If the leukocyte counts in this population follow a normal (gaussian) distribution, approximately 50% of individuals will have what total leukocyte counts?

ChatGPT Robustness and Ablation Study Analysis

The second experiment was built upon the foundation established in the first experiment by using only the questions that were categorized as correct (C) by all assessors. Using this subset of data, new questions were iteratively generated by removing a sentence from the question and prompting ChatGPT with the modified question. This process is repeated across all the sentences of a question, excluding the final question sentence indicated by the "what...?" structure. For instance, if a question has 3 sentences, then 3 versions of the same question would be generated, where each question variation has 2 sentences. The first version would remove the first sentence, the second version would remove the second sentence, and the third version would remove the third sentence. The ablation study aimed to simulate how a self-diagnosing individual may overlook information within their health-related searches.

Each of the open-ended modified questions was then provided to ChatGPT using the OpenAI application programming interface to aggregate the LLM responses for the questions. Specifically, the *gpt-4* model was used for GPT-4.0, along with specific parameters to ensure consistency among the responses and that each question would be answered independently. For instance, the model's temperature was set to zero to ensure the responses were focused and deterministic. Details of the complete GPT-4.0 model and corresponding parameters are provided in Multimedia Appendix 1, along with Textbox 3 [51] displaying the GPT-4.0 response to the open-ended Textbox 2 question.

Textbox 3. GPT-4.0 response to United States Medical Licensing Exam step 1 test question based on the transformed Textbox 2 question. The assessor can then compare the GPT-4.0 answer with the original multiple-choice options to evaluate the correctness of the response.

Answer: Approximately 50% of individuals will have total leukocyte counts of 7500/mm³ or less. Explanation: In a normal distribution, the mean value is the point at which approximately 50% of the data falls below and 50% falls above. Therefore, in this case, where the mean leukocyte count is 7500/mm³, approximately 50% of individuals will have leukocyte counts of 7500/mm³ or less.

Assessor Procedures and Guidelines

The 94 open-ended questions served as prompts for ChatGPT, where 3 nonexperts and 3 experts evaluated the generated responses. Nonexperts were used to simulate the self-diagnosis process, while experts were used to assess the response quality.

Each assessor was provided with detailed procedures and guidelines as outlined in Multimedia Appendix 2 to ensure the response categorization process was standardized among the

XSL•FO
RenderX

assessors. Assessors were instructed to categorize the responses as either C, partially correct (PC), incorrect (I), or ambiguous (A), and were provided with detailed examples to distinguish between the categories. Moreover, assessors were directed to evaluate the responses independently to ensure no bias.

Testing Methodology

We develop the testing methodology EvalPrompt, as summarized in Figure 1. In total, Nq=94 questions were selected from the USMLE step 1 test. Each question contained a question portion xj, answer options oj, and a correct answer yj, altogether forming the dataset of multiple-choice questions D=xj,oj,yjj=1Nq. Subsequently, each question, xj, was extracted and transformed into an open-ended variation which was then presented to ChatGPT.

Figure 1. EvalPrompt summary. (1) A subset of 94 USMLE step 1 [51] questions consisting of multiple-choice, single-answer questions are selected. (2) These 94 questions are modified to produce open-ended prompts. (3) The open-ended prompts are processed through the ChatGPT API (4) ChatGPT's answers are recorded and presented to *k* independent assessors to categorize as either correct, partially correct, incorrect, or ambiguous. (5) The categorizations classified as correct by all assessors are aggregated to formulate a new dataset for sensitivity analysis. (6) The prompts produced from the iterative sentence dropout are processed through ChatGPT. (7) The independent assessors categorize ChatGPT's answers. These categorizations are then aggregated based on the agreement where all assessors categorized the answer as correct. The resulting dataset, W, is the subset of USMLE question variations that ChatGPT answered correctly. USMLE: United States Medical Licensing Exam.

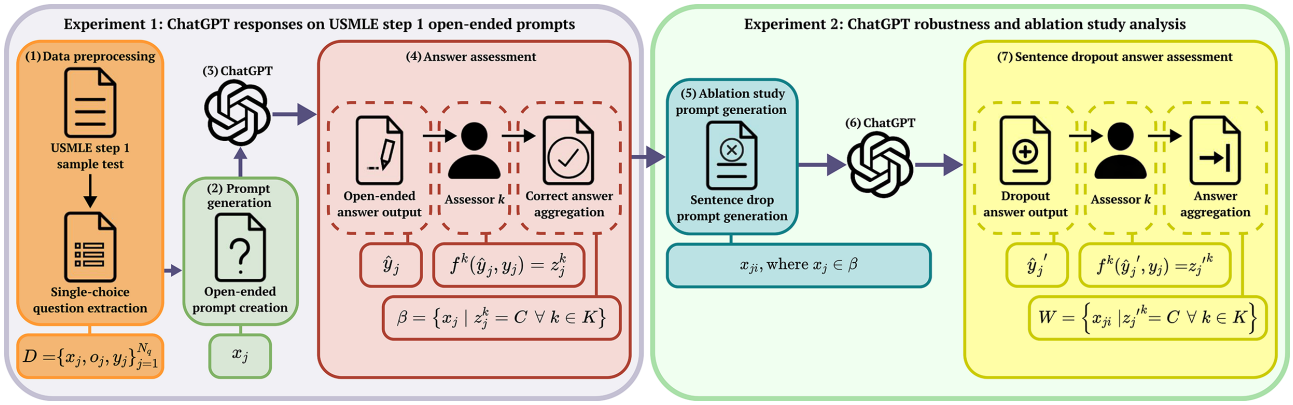

After collecting and processing the answers from ChatGPT, 3 nonexpert and 3 expert assessors independently evaluated each ChatGPT answer, y^j. In total, K=6 assessors were used. The assessors were instructed to provide a label z∈Z;Z=C,PC,I,A denoting whether the answer, in comparison to the ground truth, yj, was C, PC, I, or A, respectively. The assessment process can be expressed as a function fky^j,yj=zjk for the kth assessor, where the assessor would use the correct answer to the question along with ChatGPT's answer to categorize the response.

After the k assessors finished the categorization process, the questions categorized as C by all k assessors were aggregated to define a subset dataset, β. Once the resulting dataset, β=xj|zjk=C∀k∈K was generated, the first experiment was concluded.

The second experiment was built upon the dataset and assessments completed during the first experiment. An ablation study was conducted over the questions within dataset β via an iterative sentence dropout by performing i-1 iterations over an open-ended prompt xj, where i is the number of sentences in the prompt. For the ith iteration, the ith sentence was removed from the prompt before running the prompt through ChatGPT, as expressed by xji, where xj∈β. Each ith sentence was iteratively removed and processed except for the final sentence which contained the question sentence. The final sentence of the question was mandatory to include to ensure ChatGPT provided an appropriate response.

After processing the sentence dropout questions through ChatGPT, the same 6 assessors evaluated the responses, y^j`. The assessment process can again be expressed as a function fky^j`,yj=zj`k for the kth assessor, where the assessor would use the correct answer to the question along with ChatGPT's answer to categorize the response. This generated the resulting dataset, W, defined as W=xji|zj`k=C∀k∈K, being the subset of USMLE question variations for the certainly correctly prompts.

Ethical Considerations

Since the aim of this study is to analyze ChatGPT and not human subjects, a research ethics board review was not required. The evaluation of ChatGPT's responses was carried out on a volunteer basis, and all assessors were informed that contributing to the experiments would not result in any safety or privacy risks.

Results

Are the Responses Clear?

To determine if the GPT-4.0 responses are clear, response categorizations across assessors were compared. As displayed in Figure 2, categorization disparities exist across the nonexpert and expert assessors with few alignments across assessors as depicted in Table 1. Namely, the number of C categorizations ranges from 39 to 51, while the number of I categorizations ranges from 12 to 39. This wide range of values for the categories immediately depicts the uncertainty across assessors due to each assessor having varying backgrounds and levels of medical expertise.

Analyzing the categorization for each question independently, many discrepancies exist among the assessors. As detailed in Table 2, on average 52% ([(52+46)/2]/94, SD of 3) of the

XSL•FO
RenderX

responses in experiment 1 and 73% ([(119+104)/2]/152, SD of 7.5) of the responses in experiment 2 were categorized identically by the assessors. For example, USMLE question 54.1 was categorized by the nonexperts as PC, A, and C, and by the experts as I, PC, and PC, respectively. The inconsistency in categorizations depicts that LLM responses do not yet have a single apparent answer, but rather are still subject to interpretation depending on the individual.

Figure 2. Experiment 1: ChatGPT responses on USMLE step 1 open-ended prompts. Individual categorizations for the nonexpert and expert assessors, where each bar represents an individual's categorizations. The categorizations drastically vary across assessors, where some assessors categorized more responses as correct while other assessors categorized more responses as incorrect on the same dataset. USMLE: United States Medical Licensing Exam.

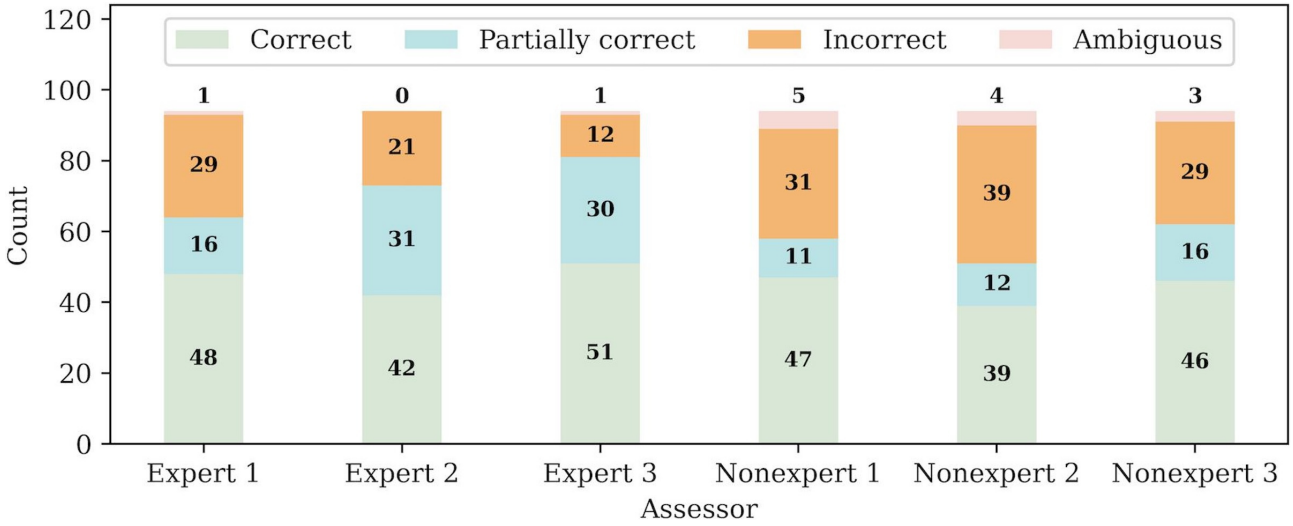

Table . Experiment 1: ChatGPT responses on USMLE[a] step 1 open-ended prompts. Overall response categorization between nonexpert and expert assessors. The top left section indicates the number of instances, regardless of correctness, where all nonexpert and expert assessors categorized a question identically. The bottom right section indicates the total number of questions categorized as correct by all nonexpert and expert assessors.

	Expert			
	Identical	Mismatch	Correct	Other
Nonexpert				
Identical	32	20	—[b]	—
Mismatch	14	28	—	—
Correct	—	—	29	4
Other	—	—	6	55

[a]USMLE: United States Medical Licensing Exam.

[b]Not applicable.

Table . Number of identical categorizations among all individuals within the 2 groups for the 2 experiments. Experiment 1 had 94 questions in total, with 52 and 46 of the questions being categorized the same among the nonexpert and expert assessors, respectively. Similarly, experiment 2 had 152 responses in total spanning 29 unique questions, with 119 and 104 of the questions being categorized the same among the nonexpert and expert assessors, respectively.

	Experiment 1		Experiment 2	
	Nonexpert	Expert	Nonexpert	Expert
Correct	33	35	111	96
Partially correct	1	4	0	1
Incorrect	18	7	8	7
Ambiguous	0	0	0	0
Overall agreement	52	46	119	104

The nonexpert and expert assessors reached an identical conclusion for 52 and 46 of the questions for experiment 1, respectively, as detailed in Table 2. However, the groups collectively could only reach an identical conclusion for 32 of the questions as listed in Table 1. Since only 34% (32/94) of the responses were consistent across all assessors, the 60% threshold could not be met indicating that the responses are unclear. The decrease in the collective number of identically

XSL•FO
RenderX

classified responses suggests that the GPT-4.0 responses are still too ambiguous for assessors to reach appropriate conclusions. The responses are not obvious enough such that anyone, regardless of their background and expertise, can reach the same conclusion.

Are the Responses Genuinely Correct?

To determine the number of GPT-4.0 responses that are genuinely correct, the responses categorized as C by all nonexpert and expert assessors were analyzed and compared. As provided in Table 1, the experiment 1 categorizations where all assessors agreed are listed. In particular, 29 of the 94 responses were categorized as C, meaning that GPT-4.0 is certainly correct 31% (29/94) of the time.

The limited number of C responses indicates that GPT-4.0 is not often factual. As portrayed in Table 2, nonexpert and expert assessors classified 33 and 35 of the responses as C, respectively, even though collectively 29 responses were considered C. Since only 31% (29/94) of the responses were considered C across all assessors, the 60% threshold could not be met indicating that the responses are mostly incorrect. Hence, even though both groups approximately categorized the same number of responses as C, there are still many responses that are either A or I.

Are the Responses Robust?

To determine if the GPT-4.0 responses are robust, an ablation study was conducted on the 29 correct responses. Reprompting GPT-4.0 with similar variations of the correct questions tested its ability to attain the correct answer even with information missing. This process aimed to simulate the self-diagnosis process since each individual would prompt LLMs with varying levels of information; some individuals would provide extensive details, while others may provide limited information. Thus, assessing GPT-4.0's robustness.

Table 2 provided the categorization details of the ablation study, consisting of 29 unique questions with 152 question variations. On average 68% ([(111+96)/2]/152, SD of 7.5) of the responses continue to be categorized as C even after removing information for each group. These results are far greater than GPT-4.0's accuracy on the initial 94 questions, which was correct only 31% (29/94) of the time. Moreover, Figure 3 depicts the categorizations for each of the 6 assessors. Assessors categorized the experiment 2 dataset C much more frequently than the experiment 1 dataset. On average, 80% ([(127+120+105+116+133+129)/6]/152, SD of 9.34) of the assessors categorized the questions as C, depicting GPT-4.0's robustness on answers that are certainly correct.

Table 3 also displays the assessment similarities between the nonexpert and expert assessors for experiment 2. A total of 92 of the 152 questions were categorized C by all nonexpert and expert assessors, meaning that 61% (92/152) of the responses are certainly correct. Since 61% (92/152) of the responses for the sentence dropout experiment were categorized as C across all assessors, the 60% threshold is met indicating that the responses are robust. In other words, if GPT-4.0 correctly answers a question, it is likely to correctly answer a similar variation of the question again, even if some information is missing.

Figure 3. Experiment 2: ChatGPT robustness and ablation study analysis. Individual categorization for the nonexpert and expert assessors, where each bar represents an individual's categorizations. Out of the 152 questions, the assessors on average categorized 122 of the questions as correct, depicting its robustness.

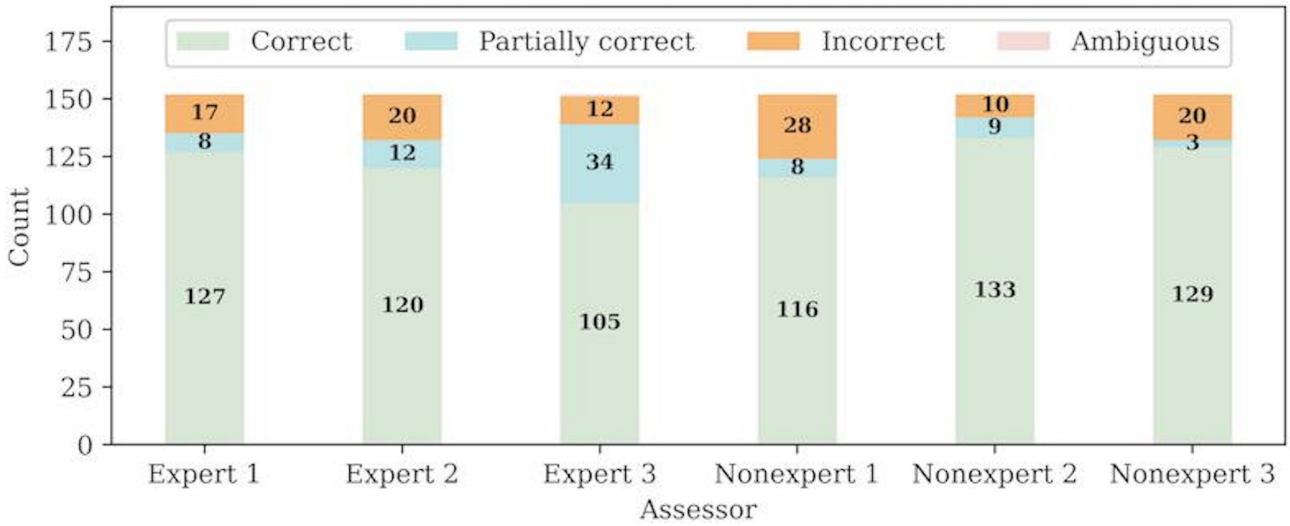

XSL•FO
RenderX

Table . Experiment 2: ChatGPT robustness and ablation study analysis. Overall response categorization between nonexpert and expert assessors. The top left section indicates the number of instances, regardless of correctness, where all nonexpert and expert assessors categorized a question identically. The bottom right section indicates the total number of questions categorized as correct by all nonexpert and expert assessors.

	Expert			
	Identical	Mismatch	Correct	Other
Nonexpert				
Identical	95	27	—[a]	—
Mismatch	9	21	—	—
Correct	—	—	92	22
Other	—	—	4	34

[a]Not applicable.

Discussion

Principal Results

The hypothesis that ChatGPT is currently unsuitable for self-diagnosis is proved. From each of the 3 addressed assessments with a passing threshold of 60%, GPT-4.0 marginally only passed one. The analysis indicates that GPT-4.0 is generally unclear and incorrect when providing medical information. However, when GPT-4.0 provides correct responses, it remains robust enough to continue answering these questions accurately even when some information is missing. Table 4 summarizes the results of the evaluation procedure for GPT-4.0.

Table . Results summarization of the evaluation procedure for GPT-4.0. With a minimum threshold of 60%, the GPT-4.0 responses are deemed generally unclear and incorrect, while exhibiting robustness when providing correct answers.

	Definition	Success rate (%)	Pass or fail
Are the responses clear?	Agreement reached between all assessors	34	Fail
Are the responses genuinely correct?	Responses categorized as correct by all assessors	31	Fail
Are the responses robust?	Responses continue to be categorized as correct by all assessors	61	Pass

Implications

Recognizing that the GPT-4.0 responses are often ambiguous, its medical advice should be accepted with caution since these responses can vary widely in interpretation depending on the individual. Additionally, since only a small number of responses were found to be genuinely correct, LLMs still require improvement before they can reliably be used in a medical setting. Without these improvements, LLMs risk misinforming individuals.

Even though ChatGPT cannot currently be used for self-diagnosis, substantial evidence indicates that LLMs are continuously improving, suggesting their potential future use in health care systems. As EvalPrompt was also conducted on an earlier GPT-3.5 version, as detailed in Multimedia Appendix 3, the significant difference between the model performances proves that LLMs continue to improve as time progresses.

With ChatGPT being highly acclaimed for its success in passing medical examinations, researchers have proposed using ChatGPT in areas such as medical education and medical report creation [17,20]. However, ChatGPT's ability to answer examination questions does not inherently equate to genuine medical comprehension and proficiency. Instead, using ChatGPT in these medical settings can undermine health care systems since ChatGPT's overconfidence can result in misinforming individuals.

Unlike practicing clinicians, ChatGPT lacks formal testing and accreditation for its abilities. It has not undergone accredited medical education or licensing, lacks approval for clinical practice, and has not demonstrated the necessary understanding or skill set to support its claims. While clinicians face severe consequences for errors, such as medical malpractice charges or license revocation, ChatGPT lacks such liability. Thus, relying on ChatGPT before it becomes accurate, consistent, and robust poses a significant risk of misleading health care practitioners and the general public.

Limitations

This study has signified a substantial advancement in the quality of medical advice that LLMs can provide and their potential utility in the health care industry. Although the USMLE questions provide a solid foundation for simulating the self-diagnosis process via LLMs, the prompts contain extensive detail. In other words, an ordinary individual typically would not input paragraphs of information when self-diagnosing and may lack explicit technical health care knowledge. Therefore, future research can explore using a dataset of self-diagnosis questions that contains less detailed information to accurately assess the capability of LLMs. As a result, developing larger and more realistic self-diagnosing datasets can enhance the

XSL•FO
RenderX

training of LLMs, in turn improving its performance. Additionally, LLMs are inclined to excel in tasks well-represented in the training data, potentially leading to lower performance in niche problems [17]. This raises equity concerns, as questions concerning underrepresented groups may endure poor performance. Although not explicitly examined in this study, future research could generate datasets with diverse question designs across medical specializations to ensure all fields are represented in training.

Conclusions

While LLMs make headlines for passing medical licensing examinations and are consequently being considered as candidates to train the next generation of health care professionals, it is evident that LLMs' capabilities are (understandably) modest at this time. More importantly, this misplaced trust in these systems can lead to reliance on their use for self-diagnosis by the public. In constructing EvalPrompt for assessing the capabilities of LLMs in medical contexts, the extent of misinformation inversely correlates to the LLMs' performance. LLMs that offer unclear and inaccurate responses are more likely to misinform individuals. Ultimately, this investigation presents challenges for machine learning researchers to build more transparent artificial intelligence–powered models capable of reasoning and responding responsibly, while also highlighting the need for a dataset that incorporates more realistic prompt styles.

Acknowledgments

Sirisha Rambhatla, Ph.D., would like to acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant (RGPIN-2022-03512).

Data Availability

The step 1 United States Medical Licensing Exam dataset along with the GPT-4.0 answers analyzed during this study are available in the EvalPrompt repository [53]. The datasets for assessor categorizations analyzed during this study are not publicly available due to ChatGPT only being analyzed in aggregate, but are available from the corresponding author on reasonable request.

Conflicts of Interest

VB is supported by an Academic Scholar Award from the University of Toronto Department of Psychiatry and has received research funding from the Canadian Institutes of Health Research, Brain & Behavior Foundation, Ontario Ministry of Health Innovation Funds, Royal College of Physicians and Surgeons of Canada, Department of National Defence (Government of Canada), New Frontiers in Research Fund, Associated Medical Services Inc Healthcare, American Foundation for Suicide Prevention, Roche Canada, Novartis, and Eisai.

Multimedia Appendix 1
The selected ChatGPT-4.0 model along with the corresponding parameters.
[DOCX File, 24 KB - formative_v9i1e66207_app1.docx]

Multimedia Appendix 2
Procedures and guidelines were provided to the assessors to evaluate the LLM responses. LLM: large language model.
[DOCX File, 143 KB - formative_v9i1e66207_app2.docx]

Multimedia Appendix 3
Categorization results for ChatGPT-3.5 in comparison to ChatGPT-4.0.
[DOCX File, 413 KB - formative_v9i1e66207_app3.docx]

References

1. Shahsavar Y, Choudhury A. User intentions to use ChatGPT for self-diagnosis and health-related purposes: cross-sectional survey study. JMIR Hum Factors 2023 May 17;10:e47564. [doi: 10.2196/47564] [Medline: 37195756]
2. Taecharungroj V. "What can ChatGPT do?" analyzing early reactions to the innovative AI Chatbot on Twitter. BDCC 2023 Mar;7(1):35. [doi: 10.3390/bdcc7010035]
3. Salah M, Alhalbusi H, Ismail MM, Abdelfattah F. Chatting with ChatGPT: decoding the mind of Chatbot users and unveiling the intricate connections between user perception, trust and stereotype perception on self-esteem and psychological well-being. Curr Psychol 2024 Mar;43(9):7843-7858. [doi: 10.1007/s12144-023-04989-0]
4. Pan Y, Pan L, Chen W, Nakov P, Kan MY, Wang WY. On the risk of misinformation pollution with large language models. arXiv. Preprint posted online on Oct 26, 2023. [doi: 10.48550/arXiv.2305.13661]
5. Weidinger L, Uesato J, Rauh M, et al. Taxonomy of risks posed by language models. Presented at: FAccT '22; Jun 21-24, 2022; Seoul, Republic of Korea. [doi: 10.1145/3531146.3533088]

XSL•FO
RenderX

6. Disruptions on the horizon. Policy Horizons Canada. 2024. URL: https://horizons.service.canada.ca/en/2024/disruptions/index.shtml [accessed 2025-02-18]
7. Goyder C, McPherson A, Glasziou P. Self diagnosis. BMJ 2009 Nov 11;339:b4418. [doi: 10.1136/bmj.b4418]
8. Jacobs W, Amuta AO, Jeon KC. Health information seeking in the digital age: an analysis of health information seeking behavior among US adults. Cogent Social Sciences 2017 Jan 1;3(1):1302785. [doi: 10.1080/23311886.2017.1302785]
9. Swire-Thompson B, Lazer D. Public health and online misinformation: challenges and recommendations. Annu Rev Public Health 2020 Apr 2;41:433-451. [doi: 10.1146/annurev-publhealth-040119-094127] [Medline: 31874069]
10. White RW, Horvitz E. Experiences with web search on medical concerns and self diagnosis. AMIA Annu Symp Proc 2009 Nov 14;2009:696-700. [Medline: 20351943]
11. El Mikati IK, Hoteit R, Harb T, et al. Defining misinformation and related terms in health-related literature: scoping review. J Med Internet Res 2023 Aug 9;25:e45731. [doi: 10.2196/45731] [Medline: 37556184]
12. Okoro YO, Ayo-Farai O, Maduka CP, Okongwu CC, Sodamade OT. A review of health misinformation on digital platforms: challenges and countermeasures. Int J Appl Res Soc Sci 2024 Mar 8;6(1):23-36. [doi: 10.51594/ijarss.v6i1.689]
13. Boniol M, Kunjumen T, Nair TS, Siyam A, Campbell J, Diallo K. The global health workforce stock and distribution in 2020 and 2030: a threat to equity and "universal" health coverage? BMJ Glob Health 2022 Jun;7(6):e009316. [doi: 10.1136/bmjgh-2022-009316] [Medline: 35760437]
14. Kuehn BM. Clinician shortage exacerbates pandemic-fueled "mental health crisis". JAMA 2022 Jun 14;327(22):2179. [doi: 10.1001/jama.2022.8661]
15. Michel JP, Ecarnot F. The shortage of skilled workers in Europe: its impact on geriatric medicine. Eur Geriatr Med 2020 Jun;11(3):345-347. [doi: 10.1007/s41999-020-00323-0] [Medline: 32328964]
16. Turale S, Nantsupawat A. Clinician mental health, nursing shortages and the COVID-19 pandemic: crises within crises. Int Nurs Rev 2021 Mar;68(1):12-14. [doi: 10.1111/inr.12674] [Medline: 33891772]
17. Garg RK, Urs VL, Agarwal AA, Chaudhary SK, Paliwal V, Kar SK. Exploring the role of ChatGPT in patient care (diagnosis and treatment) and medical research: a systematic review. Health Promot Perspect 2023;13(3):183-191. [doi: 10.34172/hpp.2023.22] [Medline: 37808939]
18. Horesh A. Using ChatGPT to study medicine: learn the basics. FutureDoctorAI. 2023 Mar. URL: https://futuredoctor.ai/chatgpt [accessed 2025-02-18]
19. Iftikhar L, Iftikhar MF, Hanif MI. DocGPT: impact of ChatGPT-3 on health services as a virtual doctor. EC Paediatrics 2023 Mar;12(3):45-55 [FREE Full text]
20. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
21. Lee H. The rise of ChatGPT: Exploring its potential in medical education. Anat Sci Educ 2024;17(5):926-931. [doi: 10.1002/ase.2270] [Medline: 36916887]
22. Primack D. Here come the robot doctors. Axios. 2023 Jan. URL: https://www.axios.com/2023/01/18/chatgpt-ai-health-care-doctors [accessed 2025-02-18]
23. Sedaghat S. Early applications of ChatGPT in medical practice, education and research. Clin Med (Lond) 2023 May;23(3):278-279. [doi: 10.7861/clinmed.2023-0078] [Medline: 37085182]
24. Gilson A, Safranek C, Huang T, et al. How does ChatGPT perform on the medical licensing exams? The implications of large language models for medical education and knowledge assessment. medRxiv. Preprint posted online on Dec 26, 2022. [doi: 10.1101/2022.12.23.22283901]
25. Scott K. Microsoft teams up with openai to exclusively license GPT-3 language model. The Official Microsoft Blog. 2020 Sep. URL: https://blogs.microsoft.com/blog/2020/09/22/microsoft-teams-up-with-openai-to-exclusively-license-gpt-3-language-model [accessed 2025-02-18]
26. Choi JH, Hickman KE, Monahan A, Schwarcz DB. ChatGPT goes to law school. SSRN Journal 2022;71(3):387-400. [doi: 10.2139/ssrn.4335905]
27. Sallam M, Salim NA, Barakat M, Al-Tammemi AB. ChatGPT applications in medical, dental, pharmacy, and public health education: a descriptive study highlighting the advantages and limitations. Narra J 2023 Apr;3(1):e103. [doi: 10.52225/narra.v3i1.103] [Medline: 38450035]
28. van Dis EAM, Bollen J, Zuidema W, van Rooij R, Bockting CL. ChatGPT: five priorities for research. Nature New Biol 2023 Feb;614(7947):224-226. [doi: 10.1038/d41586-023-00288-7] [Medline: 36737653]
29. Zhai X. ChatGPT user experience: implications for education. SSRN Journal 2022. [doi: 10.2139/ssrn.4312418]
30. Arora A, Arora A. The promise of large language models in health care. The Lancet 2023 Feb;401(10377):641. [doi: 10.1016/S0140-6736(23)00216-7]
31. Biswas SS. Role of Chat GPT in public health. Ann Biomed Eng 2023 May;51(5):868-869. [doi: 10.1007/s10439-023-03172-7] [Medline: 36920578]
32. Liebrenz M, Schleifer R, Buadze A, Bhugra D, Smith A. Generating scholarly content with ChatGPT: ethical challenges for medical publishing. Lancet Digit Health 2023 Mar;5(3):e105-e106. [doi: 10.1016/S2589-7500(23)00019-5] [Medline: 36754725]

XSL•FO
RenderX

33. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023 Mar 19;11(6):887. [doi: 10.3390/healthcare11060887] [Medline: 36981544]
34. Ufuk F. The role and limitations of large language models such as ChatGPT in clinical settings and medical journalism. Radiology 2023 May;307(3):e230276. [doi: 10.1148/radiol.230276] [Medline: 36880943]
35. Li J, Dada A, Puladi B, Kleesiek J, Egger J. ChatGPT in healthcare: a taxonomy and systematic review. Comput Methods Programs Biomed 2024 Mar;245:108013. [doi: 10.1016/j.cmpb.2024.108013] [Medline: 38262126]
36. Holmes J, Liu Z, Zhang L, et al. Evaluating large language models on a highly-specialized topic, radiation oncology physics. Front Oncol 2023;13:1219326. [doi: 10.3389/fonc.2023.1219326] [Medline: 37529688]
37. Agrawal M, Hegselmann S, Lang H, Kim Y, Sontag D. Large language models are few-shot clinical information extractors. Presented at: The 2022 Conference on Empirical Methods in Natural Language Processing; Dec 7-11, 2022; Abu Dhabi, United Arab Emirates. [doi: 10.18653/v1/2022.emnlp-main.130]
38. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. J Med Syst 2023 Mar 4;47(1):33. [doi: 10.1007/s10916-023-01925-4] [Medline: 36869927]
39. Harrer S. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. EBioMedicine 2023 Apr;90:104512. [doi: 10.1016/j.ebiom.2023.104512] [Medline: 36924620]
40. Jayakumar P, Moore MG, Furlough KA, et al. Comparison of an artificial intelligence-enabled patient decision aid vs educational material on decision quality, shared decision-making, patient experience, and functional outcomes in adults with knee osteoarthritis: a randomized clinical trial. JAMA Netw Open 2021 Feb 1;4(2):e2037107. [doi: 10.1001/jamanetworkopen.2020.37107] [Medline: 33599773]
41. Patel SB, Lam K. ChatGPT: the future of discharge summaries? Lancet Digit Health 2023 Mar;5(3):e107-e108. [doi: 10.1016/S2589-7500(23)00021-3] [Medline: 36754724]
42. Tan XJ, Cheor WL, Lim LL, Ab Rahman KS, Bakrin IH. Artificial intelligence (AI) in breast imaging: a scientometric umbrella review. Diagnostics (Basel) 2022 Dec 9;12(12):3111. [doi: 10.3390/diagnostics12123111] [Medline: 36553119]
43. Xue VW, Lei P, Cho WC. The potential impact of ChatGPT in clinical and translational medicine. Clin Transl Med 2023 Mar;13(3):e1216. [doi: 10.1002/ctm2.1216] [Medline: 36856370]
44. Yang X, Chen A, PourNejatian N, et al. A large language model for electronic health records. NPJ Digit Med 2022 Dec 26;5(1):194. [doi: 10.1038/s41746-022-00742-2] [Medline: 36572766]
45. Horvat N, Veeraraghavan H, Nahas CSR, et al. Combined artificial intelligence and radiologist model for predicting rectal cancer treatment response from magnetic resonance imaging: an external validation study. Abdom Radiol (NY) 2022 Aug;47(8):2770-2782. [doi: 10.1007/s00261-022-03572-8] [Medline: 35710951]
46. Pun FW, Leung GHD, Leung HW, et al. Hallmarks of aging-based dual-purpose disease and age-associated targets predicted using PandaOmics AI-powered discovery engine. Aging (Albany NY) 2022 Mar 29;14(6):2475-2506. [doi: 10.18632/aging.203960] [Medline: 35347083]
47. Rao A, Kim J, Kamineni M, Pang M, Lie W, Succi MD. Evaluating ChatGPT as an adjunct for radiologic decision-making. medRxiv. Preprint posted online on Feb 7, 2023. [doi: 10.1101/2023.02.02.23285399] [Medline: 36798292]
48. Wang S, Zhao Z, Ouyang X, Liu T, Wang Q, Shen D. Interactive computer-aided diagnosis on medical image using large language models. Commun Eng 2024 Sep 17;3(1):133. [doi: 10.1038/s44172-024-00271-8] [Medline: 39284899]
49. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature New Biol 2023 Aug;620(7972):172-180. [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
50. Huang Y, Sun L, Wang H, Wu S, Zhang Q, Li Y, et al. TrustLLM: trustworthiness in large language models. arXiv. Preprint posted online on Jan 10, 2024. [doi: 10.48550/arXiv.2401.05561]
51. Step 1 sample test questions. USMLE. 2022 Jun. URL: https://www.usmle.org/exam-resources/step-1-materials/step-1-sample-test-questions [accessed 2025-02-18]
52. Ziaei R, Schmidgall S. Language models are susceptible to incorrect patient self-diagnosis in medical applications. arXiv. Preprint posted online on Sep 17, 2023. [doi: 10.48550/arXiv.2309.09362]
53. Zada T, Tam N, Rambhatla S. EvalPrompt: analyzing large language models for self-diagnosis. GitHub. 2022. URL: https://github.com/criticalml-uw/EvalPrompt [accessed 2025-02-18]

Abbreviations

A: ambiguous
C: correct
EvalPrompt: evaluation of large language model prompts
I: incorrect
LLM: large language model
PC: partially correct
USMLE: United States Medical Licensing Exam

XSL•FO
RenderX

Edited by A Mavragani, D Chartash; submitted 06.09.24; peer-reviewed by D Lungeanu, GK Gupta, Y Huang; revised version received 29.01.25; accepted 29.01.25; published 10.03.25.

Please cite as:
Zada T, Tam N, Barnard F, Van Sittert M, Bhat V, Rambhatla S
Medical Misinformation in AI-Assisted Self-Diagnosis: Development of a Method (EvalPrompt) for Analyzing Large Language Models
JMIR Form Res 2025;9:e66207
URL: https://formative.jmir.org/2025/1/e66207
doi:10.2196/66207

© Troy Zada, Natalie Tam, Francois Barnard, Marlize Van Sittert, Venkat Bhat, Sirisha Rambhatla. Originally published in JMIR Formative Research (https://formative.jmir.org), 10.3.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Online Health Information–Seeking Behaviors Among the Chongqing Population: Cross-Sectional Questionnaire Study

Honghui Rong[1*], MSc; Lu Lu[1*], MSc; Miao He[2], MSc; Tian Guo[2], MSc; Xian Li[2], MSc; Qingliu Tao[2], MSc; Yixin Li[2], MSc; Chuanfen Zheng[1], MSc; Ling Zhang[1], MSc; Fengju Li[1], MSc; Dali Yi[1], MSc; Enyu Lei[1], BS; Ting Luo[1], MSc; Qinghua Yang[2], MSc; Ji-an Chen[1], PhD

[1]Department of Health Education, School of Military Preventive Medicine, Army Medical University (Third Military Medical University), Gaotanyan Street No 30, Shapingba District, Chongqing, China

[2]Chongqing Health Education Institute, Chongqing, China

[*]these authors contributed equally

Corresponding Author:
Ji-an Chen, PhD
Department of Health Education, School of Military Preventive Medicine, Army Medical University (Third Military Medical University), Gaotanyan Street No 30, Shapingba District, Chongqing, China

Related Article:

This is a corrected version. See correction statement: https://formative.jmir.org/2025/1/e77500

Abstract

Background: With the rapid development of the internet and its widespread use, online health information–seeking (OHIS) has become a popular and important research topic. Various benefits of OHIS are well recognized. However, OHIS seems to be a mixed blessing. Research on OHIS has been reported in Western countries and in high-income regions in eastern China. Studies on the population in the western region of China, such as Chongqing, are still limited.

Objective: The aim of the study was to identify the prevalence, common topics, and common methods of health information–seeking and the factors influencing these behaviors among the Chongqing population.

Methods: This cross-sectional questionnaire study was conducted from September to October 2021. A web-based questionnaire was sent to users aged 15 years and older in Chongqing using a Chinese web-based survey hosting site (N=14,466). Data on demographics, web-based health information resources, and health topics were collected. Factors that may influence health literacy were assessed using the chi-square test and multivariate logistic regression models.

Results: A total of 67.1% (9704/14,466) of the participants displayed OHIS behaviors. Participants who were younger, had a higher educational level, and worked as medical staff or teachers were more likely to engage in OHIS, while those living in rural areas, ethnic minorities, and farmers were less likely to seek health information on the web ($P<.01$). Among the Chongqing population, the most common topic searched on the internet was health behavior and literacy (87.4%, 8483/9704), and the most popular method of seeking health information on the web was through WeChat (77.0%, 7468/9704).

Conclusions: OHIS is prevalent in Chongqing. Further research could be performed based on the influencing factors identified herein and high-priority, effective ways of improving the OHIS behaviors of the Chongqing population.

(JMIR Form Res 2025;9:e56028) doi:10.2196/56028

KEYWORDS

online health information seeking; health behavior; Chongqing; China; Internet

Introduction

With rapid development, the internet has become a major source of health information worldwide. According to Internet World Stats [1], there are 5 billion internet users, accounting for more than 60% of the world population. China also has a large population of internet users. In 2020, approximately 9.89 million Chinese people (70.4% of the Chinese population) had access to the internet [2]. The internet hosts a tremendous amount and variety of health-related information that can be accessed at conveniently, anonymously, and at relatively low cost [3,4]. Among all health information sources, the internet has the highest usage rate [5-7]. Although certain groups still rely on traditional sources such as books and printed journals for health information [8], web-based sources of health information are increasingly growing in popularity.

XSL•FO
RenderX

Various benefits of online health information–seeking (OHIS) are well recognized. Patients turning to the internet before seeking medical consultations or diagnosis may help improve patient-doctor relationships, and patients are more inclined to trust their physicians' advice when they are able to discuss the information they found on the web with their doctors [9-12]. As individuals aim to change their lifestyle or health behavior, the frequency of the use of the internet to retrieve health information is likely to increase [13]. For people with chronic diseases, OHIS may help them manage their health condition [14,15]. However, OHIS seems to be a mixed blessing. Studies have reported that the overall quality of web-based health information is relatively low [16-18]. Information seekers are at risk of making hasty or dangerous health decisions based on questionable web-based information [19]. In contrast to the information available in Western countries, low-quality web-based health information seems to be particularly prevalent in Asian countries [20].

With the rapid development of the internet and its widespread use, OHIS has become a popular and important research topic. Research on OHIS has been reported in Western countries [8,14,21] as well as low- and middle-income countries [22-24]. Various factors, such as age, gender, education, and internet usage, have been shown to affect the prevalence and extent of OHIS in previous studies [13,22,25]. In addition, researchers have been interested in the topic of OHIS. A previous study also reported that various health topics are searched by web-based health information seekers. The most common topics searched on the internet are likely to fall under 2 categories: health behavior, such as nutrition or diet, exercise, and body maintenance; and medical concerns, including information related to disease, medications, and treatments [11,14,22,26]. In recent years, several academic studies on OHIS have collected evidence from high-income regions in eastern China, such as Zhejiang, Guangdong, and Hong Kong [26-30]. Western China, as a transitional region—home to 27% of the national population with distinct socioeconomic and health care system characteristics—remains critically underresearched [31]. This knowledge gap limits our understanding of the situation of OHIS thoroughly in this vast country with uneven development. Chongqing, as the largest municipality in western China, embodies the region's characteristic "urban-rural dual structure" with more than a 60% urbanization rate and significant health resource disparities between metropolitan and rural areas [32]. This metropolis is an ideal context for understanding situation of OHIS in western China.

The study objectives were to (1) determine the prevalence, common topics, and methods of OHIS in the Chongqing population and (2) identify the factors that influence the OHIS behaviors of this population. This research may help improve ways of promoting efficient and appropriate OHIS for users and harnessing the benefits of the internet as a source of health information.

Methods

Study Design

From September to October 2021, Chongqing Health Education Institution and the Army Medical University carried out a web-based study to assess residents' health care needs in Chongqing municipality. The target participants were Chongqing residents. A survey QR code was disseminated on popular Chinese social media applications such as WeChat for voluntary participation in our web-based survey, which was hosted on the Chinese Sojump site. The first page of the survey was a web-based consent form including study information. After reading the consent forms and indicating consent to participate in the survey, respondents were allowed to proceed. To avoid multiple submissions, only one submission per IP address was allowed. Ultimately, 14,466 participants were included in this study.

Measures

This questionnaire included 2 parts: demographic characteristics and OHIS behaviors. Demographic characteristics included age, gender, education, occupation, ethnicity, and area of residence, gender, education, and ethnicity. Age was measured by asking participants to indicate their numeric age, and other variables were measured with multiple-choice questions. The part on OHIS behaviors were included three questions. The first one dealt addressed having experience with OHIS, wherein participants' OHIS behavior was measured with a single question, "Which sources are your main sources for seeking health information?" The response options included books and journals, broadcasts, television, PCs (desktops and laptops), mobile phones, lectures, professional staff, and advertisements. Participants who chose PCs and/or mobile phones as their main sources for seeking health information were considered to have experience with OHIS. The second one addressed major health topics in OHIS: health topics that the participants searched on the internet were captured by a multiple-choice question with response options including health behavior and literacy (such as diet, fitness, exercise, and drug usage), infectious diseases, chronic disease, first aid, and health policies (such as medical insurance). Participants could select one or more answers as their major health topics when searching on the internet. The third one addressed the main method of OHIS: the main method of seeking health information on the web was measured with a single item with response options including WeChat, Weibo, search engines (such as Baidu and Google), websites, short-video apps (such as TikTok), and others. Participants could select one or more answers as their main method of seeking health information on the web.

Ethical Considerations

The study was carried out in accordance with ethical principals and was approved by the ethics review board of the Army Medical University (2023-5-02). Participants provided informed consent. All participants' information was anonymized. There was no financial compensation for patients or researchers nor any source of funding that could lead to a conflict of interest for the study.

XSL•FO
RenderX

Statistical Analysis

All data were input into an Epidata database (version 3.1) after checking and correcting errors. SPSS (version 22.0; IBM Corp) was used for analyses. A descriptive analysis (frequencies, percentages, and means with SDs) of the participant characteristics was performed. The chi-square test was used to compare OHIS behaviors among groups. Multiple logistic regression models were used to assess the influencing factors associated with OHIS. Statistical significance was set to a *P* value of <.05 (2-sided).

Results

Sociodemographic Characteristics of the Study Sample

The demographic characteristics of the study sample are listed in Table 1. In total, 67.1% (9704/14,466) of the participants had OHIS experience. The average age was 46.2 (SD 18.0) years, while approximately half (51.8%, 7495/14,466) of the participants were younger than 45 years. More than half (52.0%, 7520/14,466) of the participants were female. Most participants (95.3%, 13,793/14,466) were of Han Chinese ethnicity. The percentage of participants residing in urban areas was 69.7% (10,090/14,466). Overall, 47.1% (6813/14,466) of participants were college graduates or had a higher level of education. The participants spanned all occupation groups.

As indicated in Table 1, OHIS experience significantly differed by age, ethnicity, area of residence, education and occupation (*P*<.05) but not by gender.

XSL•FO
RenderX

Table . Differences in the OHIS[a] characteristics of the respondents and their sociodemographic characteristics (N=14,466).

Characteristics	Participants, n (%)	Respondents without OHIS experience, n (%)	Respondents with OHIS experience, n (%)	*P* value
Age (years)				
15 - 45	7495 (51.8)	1208 (25.4)	6287 (64.8)	<.001
46 - 60	3036 (21.0)	790 (16.6)	2246 (23.1)	
61 or older	3935 (27.2)	2764 (58.0)	1171 (12.1)	
Gender				
Male	6946 (48.0)	2308 (48.5)	4638 (47.8)	.46
Female	7520 (52.0)	2454 (51.5)	5066 (52.2)	
Ethnicity				
Han Chinese	13,793 (95.3)	4513 (94.8)	9280 (95.6)	.02
Ethnic minority	673 (4.7)	249 (5.2)	424 (4.4)	
Area of residence				
Urban	10,090 (69.7)	2587 (54.3)	7503 (77.3)	<.001
Rural	4376 (30.3)	2175 (45.7)	2201 (22.7)	
Education level				
Primary school or less	2998 (20.7)	2301 (48.3)	697 (7.2)	<.001
Junior high school	2467 (17.1)	1020 (21.4)	1447 (14.9)	
Senior high school	2188 (15.1)	514 (10.8)	1674 (17.3)	
College graduate	6289 (43.5)	855 (18.0)	5434 (56.0)	
Postgraduate	524 (3.6)	72 (1.5)	452 (4.7)	
Occupation				
Civil servants	1669 (11.5)	253 (5.3)	1416 (14.6)	<.001
Teachers	1208 (8.4)	150 (3.1)	1058 (10.9)	
Medical staff	1603 (11.1)	223 (4.7)	1380 (14.2)	
Staff in public institutions	969 (6.7)	175 (3.7)	794 (8.2)	
Students	887 (6.1)	171 (3.6)	716 (7.4)	
Farmers	3317 (22.9)	2266 (47.6)	1051 (10.8)	
Workers	1180 (8.2)	407 (8.5)	773 (8.0)	
Enterprise personnel	1250 (8.6)	238 (5.0)	1012 (10.4)	
Others	2383 (16.5)	879 (18.5)	1504 (15.5)	

[a]OHIS: online health information–seeking.

Multivariate Logistic Regression Analysis of Risk Factors Associated With the Rate of Health Literacy Knowledge

The variables with statistical significance in the chi-square test (Table 1) were analyzed using multivariate logistic regression. As shown in Table 2, participants aged 46 - 60 years (odds ratio [OR] 0.782, 95% CI 0.695-0.880) or more than 61 years (OR 0.298, 95% CI 0.260-0.341) were less likely to have OHIS experience than those younger than 20 years. Participants who belonged to ethnic minorities (OR 0.621, 95% CI 0.513-0.752) were less likely to have OHIS experience than Han Chinese participants. Participants from rural areas (OR 0.815, 95% CI 0.734-0.906) were less likely to have OHIS experience than urban participants. Compared to the respondents with primary school education and below, those with junior high school education (OR 2.290, 95% CI 2.000-2.622), senior high school education (OR 3.274, 95% CI 2.765-3.877), college graduate degrees (OR 5.012, 95% CI 4.163-6.033), and postgraduate education (OR 4.809, 95% CI 3.442-6.720) were more likely to have OHIS experience. Based on the participants' occupation, compared to civil servants, teachers (OR 1.407, 95% CI 1.118-1.770) and medical staff (OR 1.359, 95% CI 1.098-1.683) were more likely to seek health information on the web, while farmers (OR 0.656, 95% CI 0.525-0.820) were less likely to seek health information on the internet.

XSL•FO
RenderX

Table . Multivariate logistic regression analysis of the factors associated with OHIS[a] in participants.

Characteristics	OR[b]	95% CI	*P* value
Age (years)			
15 - 45	1[c]	—[d]	—
46 - 60	0.782	0.695 - 0.880	<.001
61 or older	0.298	0.260 - 0.341	<.001
Gender			
Male	1	—	—
Female	0.950	0.871 - 1.036	.24
Ethnicity			
Han Chinese	1	—	—
Ethnic minority	0.621	0.513 - 0.752	<.001
Area of residence			
Urban	1	—	—
Rural	0.815	0.734 - 0.906	<.001
Education			
Primary school or less	1	—	—
Junior high school	2.290	2.000 - 2.622	<.001
Senior high school	3.274	2.765 - 3.877	<.001
College graduate	5.012	4.163 - 6.033	<.001
Postgraduate	4.809	3.442 - 6.720	<.001
Occupation			
Civil servants	1	—	—
Teachers	1.407	1.118 - 1.770	.004
Medical staff	1.359	1.098 - 1.683	.005
Staff in public institution	1.100	0.872 - 1.837	.421
Students	1.279	0.979 - 1.672	.072
Farmers	0.656	0.525 - 0.820	<.001
Workers	0.905	0.721 - 1.135	.385
Enterprise personnel	1.052	0.843 - 1.313	.652
Others	1.035	0.843 - 1.272	.740

[a]OHIS: online health information–seeking.
[b]OR: odds ratio.
[c]Reference variable.
[d]Not applicable.

Health Topics Searched on the Internet by Participants With OHIS Experience

Table 3 shows the health topics searched on the internet by the study participants. Most of the participants (8483/9704, 87.4%) indicated that they sought information about health behavior and literacy. More than three-quarters of the participants used the internet to find information about infectious diseases (78.6%, 7628/9704), chronic diseases (76.0%, 7375/9704), and first aid (75.0%, 7277/9704). More than half (58.6%, 5688/9704) of the participants searched for health policy information on the web.

XSL•FO
RenderX

Table . Common health topics searched on the internet by participants with OHIS[a] experience (N=9704).

Topics	Participants, n (%)
Health behavior and literacy	8483 (87.4)
Infectious diseases	7628 (78.6)
Chronic disease	7375 (76.0)
First aid	7277 (75.0)
Health policies	5688 (58.6)

[a]OHIS: online health information–seeking.

Sources of Web-Based Health Information Among Participants With OHIS Experience

Participants reported 1 or more different sources of health information that they sought on the web. The majority of participants (7468/9704, 77.0%) used WeChat. More than half of the participants indicated that search engines (5547/9704, 57.2%) and short-video apps (5359/9704, 55.2%) were their main sources. A total of 4079 out of 9704 (42.0%) participants used Weibo. Websites and web-based courses were used for seeking health information on the web by 36.3% (3527/9704) and 31.3% (3034/9704) of participants, respectively. The prevalence of the source of web-based health information among the participants with OHIS experience is shown in Table 4.

Table . Source of online health information among the participants with OHIS experience (N=9704).

Sources	Participants, n (%)
WeChat	7468 (77.0)
Weibo	4079 (42.0)
Search engines	5547 (57.2)
Websites	3527 (36.3)
Web-based courses	3034 (31.3)
Short-video apps	5359 (55.2)
Others	418 (4.3)

[a]OHIS: online health information–seeking.

Discussion

Principal Findings

In this study, more than 67% of participants sought health information on the internet, which is similar to recent studies reporting extensive internet use and highly prevalent OHIS [22,24,26,33]. The internet has the highest usage rate among health information sources [5-7]. The results from our study suggested that among the Chongqing population, the internet has become a major source for seeking health information, which is consistent with other reports. How to provide more high-quality health information on the web may be a key objective for public health policies and practices in Chongqing.

In this study, younger participants were more likely to seek health information on the web than older participants. Similar results were also reported in some previous studies [22,30,33,34]. IT, including health IT, is usually accessible for younger generations [35]. Therefore, younger populations may more frequently seek health information on the internet.

We also found evidence of disparities in OHIS by ethnicity in Chongqing. This result was similar to those of previous studies showing that racial or ethnic minorities were less likely to use web-based resources to seek health information [33,36,37]. A possible reason may be that di erences in cultural values, care preferences, and perceived benefits of web-based health information likely contributed to these di erential rates of use [36-38]. Previous research has shown that health communication in a multicultural society mainly takes the dominant culture into account, often neglecting those of nondominant groups [39-41]. In China, Han is the dominant ethnicity. This is likely secondary in part to the limited availability of web-based sources such as health-related websites and patient-provider portals in ethnic minorities other than the majority [42]. In Chongqing, the population of ethnic minorities is more than 2 million [32]. Considering the benefits of OHIS, providing web-based health information that is suitable for racial or ethnic minorities may contribute to health improvement in the multiethnic region.

Participants who lived in urban areas were more likely to use the internet to seek health information. This result corroborates previous findings that the urban population had a higher rate of OHIS [33,34,43,44]. A possible reason may be the digital divide between the urban and rural participants in this study. Although there are more than 1 billion netizens in China, people living in urban areas more easily access the internet [45]. Goldner et al [44] reported that lower internet access was associated with fewer web-based health behaviors. Therefore, eliminating the digital divide may help the rural population benefit from these resources.

XSL•FO
RenderX

In this study, a higher education level was associated with the highest odds of OHIS among the Chongqing population. This finding was in line with those of previous studies [22,27,29,30,33,34,38]. Seeking health information on the web requires not only access to technology but also the ability to retrieve, understand, and use information [46]. In addition, the vast majority of web-based patient resources contain health information that is above the reading level of most users [47,48]. Although a substantial proportion of lesser educated individuals have significant health care needs, they often encounter di culties in finding acceptable information on the web [49]. Continued e orts to ensure that web-based health information is easy to read, understand, and retrieve are needed.

We also found that occupation was associated with the rate of OHIS behavior. In this study, participants who were medical staff or teachers were more likely to use the internet to seek health information. The reason may be that medical staff and teachers more often have higher education levels, higher socioeconomic status, and easier access to web-based resources. In addition, due to their duties, medical staff and teachers often conduct health education for patients and students, and the internet has a tremendous amount and variety of health-related information [3,4]. Therefore, they may have a higher ability and willingness to seek health information on the internet [50]. Farmers in our study were less likely to use web-based sources for seeking health information. This finding was similar to that of a previous study in Zhejiang province [29]. Farmers always live in rural areas. Due to the digital divide between urban and rural areas, farmers may have more difficulty accessing the internet [44,45].

In this study, participants reported seeking web-based health information for a broad range of health topics, including health behavior and literacy, infectious diseases, chronic disease, first aid, and health policies. Health behavior and literacy are the most common topics searched on the internet. This finding is similar to those of previous studies in which health behavior, lifestyle, and health science popularization were most commonly searched for types of web-based health information among the Chinese general or younger population [27,28]. In recent years, the Chinese government has tended to use social media to improve public health literacy and health status among Chinese citizens and has encouraged the dissemination of health science popularization information in various ways [51-53]. The increasing popularity of social media and the ever-growing number of official accounts of health science popularization might have attracted many Chinese netizens to use such information to improve their health behavior and literacy.

We found that WeChat was the most commonly used source for OHIS on the internet. A previous study reported that the number of WeChat customers has exceeded 900 million, with 150 million customers using the web for at least 2 hours every day [54]. A nationwide survey in China found that one-third of participants regularly read health information articles on WeChat, and more than 90% of the participants chose to use WeChat for health information seeking, indicating that a WeChat account is the most popular platform for acquiring health information in China [55]. However, due to growing OHIS behaviors, increasing numbers of nonauthorized social media accounts share biased or inaccurate health information; continued e orts are needed to improve the quality of health information on social media [56-58]. Improving perceived eHealth literacy among netizens and feedback-seeking behavior in digital environments would be useful to increase OHIS, and may finally help to improve public health among Chongqing Population [59,60].

Limitations

This study has several limitations. First, this study was conducted in Chongqing, and the results may not be generalizable to the general population in China. Second, due to the nature of a cross-sectional survey, it is difficult to draw causal conclusions. Third, as a hot topic of OHIS in this study, failing to examine the effect of the COVID-19 on OHIS was a significant study limitation.

Conclusion

In summary, OHIS is prevalent in Chongqing. We found that participants who were younger, lived in an urban area, had a higher educational level, and worked as medical staff or teachers were more likely to engage in OHIS, while ethnic minorities and farmers were less likely to seek health information on the web. Among the Chongqing population, the most common topics sought on the internet were health behavior and literacy, and the most popular method of OHIS was through WeChat. According to the identified influencing factors, future research could focus on bridging the digital divide between urban and rural areas, providing higher-quality web-based health information, and examining cultural barriers to health information access among ethnic minority groups. These efforts may help to enhance Chongqing residents' ability to obtain web-based health resources and ultimately improve public health outcomes.

Acknowledgments

Ji-an Chen and Qinghua Yang are co-corresponding authors, the latter of which can be contacted using the following information: phone number 86 02367168619 and email address 914543700@qq.com. We acknowledge all participants in the study.

Data Availability

The datasets used and analyzed during this study are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Authors' Contributions

Conceptualization: HR, J-aC, and QY
Data collection: TG, XL, YL, and HR
Data curation: HR, LL, and DY
Formal analysis: LL, HR, and LZ
Investigation: TG, XL, YL, HR, and LL
Methodology: HR, J-aC, and LL
Software: EL, LZ, FL, and DY
Supervision: J-aC
Validation: TG, QT, and FL
Visualization: LL and HR
Writing—original draft: LL and HR
Writing—review and editing: LL, CZ, FL, and TL

Conflicts of Interest

None declared.

References

1. World internet users and 2021 population stats. Internet World Stats. 2021. URL: https://www.internetworldstats.com/stats.htm [accessed 2025-04-03]
2. The 47th statistical report on China's internet development. China Internet Network Information Center. 2021. URL: https://www.cnnic.com.cn/IDR/ReportDownloads/202104/P020210420557302172744.pdf [accessed 2025-04-03]
3. Solberg LB. The benefits of online health communities. Virtual Mentor 2014 Apr 1;16(4):270-274. [doi: 10.1001/virtualmentor.2014.16.04.stas1-1404] [Medline: 24735575]
4. Cline RJ, Haynes KM. Consumer health information seeking on the internet: the state of the art. Health Educ Res 2001 Dec;16(6):671-692. [doi: 10.1093/her/16.6.671] [Medline: 11780707]
5. Cherrez Ojeda I, Vanegas E, Torres M, et al. Ecuadorian cancer patients' preference for information and communication technologies: cross-sectional study. J Med Internet Res 2018 Feb 20;20(2):e50. [doi: 10.2196/jmir.8485] [Medline: 29463492]
6. Cash T, Desbrow B, Leveritt M, Ball L. Utilization and preference of nutrition information sources in Australia. Health Expect 2015 Dec;18(6):2288-2295. [doi: 10.1111/hex.12198] [Medline: 24798108]
7. Khoong EC, Le GM, Hoskote M, Rivadeneira NA, Hiatt RA, Sarkar U. Health information-seeking behaviors and preferences of a diverse, multilingual urban cohort. Med Care 2019 Jun;57 Suppl 6 Suppl 2(Suppl 6 2):S176-S183. [doi: 10.1097/MLR.0000000000001050] [Medline: 31095058]
8. Jacobs W, Amuta AO, Jeon KC. Health information seeking in the digital age: an analysis of health information seeking behavior among US adults. Cogent Social Sciences 2017 Jan 1;3(1):1302785. [doi: 10.1080/23311886.2017.1302785]
9. Peng Y, Yin P, Deng Z, Wang R. Patient-physician interaction and trust in online health community: the role of perceived usefulness of health information and services. Int J Environ Res Public Health 2019 Dec 24;17(1):1-13. [doi: 10.3390/ijerph17010139] [Medline: 31878145]
10. Tan SSL, Goonawardene N. Internet health information seeking and the patient-physician relationship: a systematic review. J Med Internet Res 2017 Jan 19;19(1):e9. [doi: 10.2196/jmir.5729] [Medline: 28104579]
11. Rider T, Malik M, Chevassut T. Haematology patients and the internet--the use of on-line health information and the impact on the patient-doctor relationship. Patient Educ Couns 2014 Nov;97(2):223-238. [doi: 10.1016/j.pec.2014.06.018] [Medline: 25053474]
12. Moreland J, French TL, Cumming GP. The prevalence of online health information seeking among patients in Scotland: a cross-sectional exploratory study. JMIR Res Protoc 2015 Jul 15;4(3):e85. [doi: 10.2196/resprot.4010] [Medline: 26177562]
13. Ayers SL, Kronenfeld JJ. Chronic illness and health-seeking information on the internet. Health (London) 2007 Jul;11(3):327-347. [doi: 10.1177/1363459307077547.11/3/327] [Medline: 17606698]
14. Madrigal L, Escoffery C. Electronic health behaviors among US adults with chronic disease: cross-sectional survey. J Med Internet Res 2019 Mar 5;21(3):e11240. [doi: 10.2196/11240] [Medline: 30835242]
15. Jin Y, Jing M, Zhang L, Song S, Ma X. Internet access and hypertension management among the elderly population: a nationally representative cross-sectional survey in China. J Med Internet Res 2019 Jan 31;21(1):e11280. [doi: 10.2196/11280] [Medline: 30702439]
16. Sun Y, Zhang Y, Gwizdka J, Trace CB. Consumer evaluation of the quality of online health information: systematic literature review of relevant criteria and indicators. J Med Internet Res 2019 May 2;21(5):e12522. [doi: 10.2196/12522] [Medline: 31045507]
17. Garfinkle R, Wong-Chong N, Petrucci A, et al. Assessing the readability, quality and accuracy of online health information for patients with low anterior resection syndrome following surgery for rectal cancer. Colorectal Dis 2019 May;21(5):523-531. [doi: 10.1111/codi.14548] [Medline: 30609222]

XSL•FO
RenderX

18. Storino A, Castillo-Angeles M, Watkins AA, et al. Assessing the accuracy and readability of online health information for patients with pancreatic cancer. JAMA Surg 2016 Sep 1;151(9):831. [doi: 10.1001/jamasurg.2016.0730] [Medline: 27144966]
19. Schulz PJ, Nakamoto K. Patient behavior and the benefits of artificial intelligence: the perils of “dangerous” literacy and illusory patient empowerment. Patient Educ Couns 2013 Aug;92(2):223-228. [doi: 10.1016/j.pec.2013.05.002] [Medline: 23743214]
20. Varady NH, Dee EC, Katz JN. International assessment on quality and content of internet information on osteoarthritis. Osteoarthr Cartil 2018 Aug;26(8):1017-1026. [doi: 10.1016/j.joca.2018.04.017] [Medline: 29758353]
21. Europeans becoming enthusiastic users of online health information. European Commission. URL: https://ec.europa.eu/digital-single-market/en/news/europeans-becoming-enthusiastic-users-online-healthinformation [accessed 2022-12-22]
22. Ghweeba M, Lindenmeyer A, Shishi S, Abbas M, Waheed A, Amer S. What predicts online health information-seeking behavior among Egyptian adults? A cross-sectional study. J Med Internet Res 2017 Jun 22;19(6):e216. [doi: 10.2196/jmir.6855] [Medline: 28642216]
23. Al-Shorbaji N. E-health in the Eastern Mediterranean Region: a decade of challenges and achievements. East Mediterr Health J 2008;14 Suppl:S157-S173. [Medline: 19205616]
24. AlSaadi MM. Evaluation of internet use for health information by parents of asthmatic children attending pediatric clinics in Riyadh, Saudi Arabia. Ann Saudi Med 2012;32(6):630-636. [doi: 10.5144/0256-4947.2012.630] [Medline: 23396028]
25. Wang X, Shi J, Kong H. Online health information seeking: a review and meta-analysis. Health Commun 2021 Sep;36(10):1163-1175. [doi: 10.1080/10410236.2020.1748829] [Medline: 32290679]
26. Wong DKK, Cheung MK. Online health information seeking and eHealth literacy among patients attending a primary care clinic in Hong Kong: a cross-sectional survey. J Med Internet Res 2019 Mar 27;21(3):e10831. [doi: 10.2196/10831] [Medline: 30916666]
27. Xiong Z, Zhang L, Li Z, Xu W, Zhang Y, Ye T. Frequency of online health information seeking and types of information sought among the general Chinese population: cross-sectional study. J Med Internet Res 2021 Dec 2;23(12):e30855. [doi: 10.2196/30855] [Medline: 34860676]
28. Zhang D, Zhan W, Zheng C, et al. Online health information-seeking behaviors and skills of Chinese college students. BMC Public Health 2021 Apr 15;21(1):736. [doi: 10.1186/s12889-021-10801-0] [Medline: 33858389]
29. Qiu Y, Ren W, Liu Y, Yin P, Ren J. Online health information in a rural residential population in Zhejiang Province, China: a cross-sectional study. BMJ Open 2019 May 5;9(5):e026202. [doi: 10.1136/bmjopen-2018-026202] [Medline: 31061032]
30. Liu L. Medical information seeking behavior of urban patients in Zhejiang Province, China: a cross-sectional study. BMC Public Health 2022 Aug 21;22(1):1591. [doi: 10.1186/s12889-022-14017-8] [Medline: 35987622]
31. China statistical yearbook [Article in Chinese]. National Bureau of Statistics of China. 2023. URL: https://www.stats.gov.cn/sj/ndsj/2023/indexch.htm [accessed 2025-04-03]
32. Chongqing statistical yearbook. Chongqing Municipal Bureau of Statistics. 2023. URL: https://tjj.cq.gov.cn/zwgk_233/tjnj/tjnj.html?url=https://tjj.cq.gov.cn/cslm/tjsjzl/202409/P020240913357156130170.pdf [accessed 2025-04-03]
33. Nguyen A, Mosadeghi S, Almario CV. Persistent digital divide in access to and use of the Internet as a resource for health information: results from a California population-based study. Int J Med Inform 2017 Jul;103:49-54. [doi: 10.1016/j.ijmedinf.2017.04.008] [Medline: 28551001]
34. Bujnowska-Fedak MM, Waligóra J, Mastalerz-Migas A. The internet as a source of health information and services. Adv Exp Med Biol 2019;1211:1-16. [doi: 10.1007/5584_2019_396] [Medline: 31273574]
35. Powell J, Inglis N, Ronnie J, Large S. The characteristics and motivations of online health information seekers: cross-sectional survey and qualitative interview study. J Med Internet Res 2011 Feb 23;13(1):e20. [doi: 10.2196/jmir.1600] [Medline: 21345783]
36. Yamin CK, Emani S, Williams DH, et al. The digital divide in adoption and use of a personal health record. Arch Intern Med 2011 Mar 28;171(6):568-574. [doi: 10.1001/archinternmed.2011.34] [Medline: 21444847]
37. Hsu J, Huang J, Kinsman J, et al. Use of e-Health services between 1999 and 2002: a growing digital divide. J Am Med Inform Assoc 2005;12(2):164-171. [doi: 10.1197/jamia.M1672] [Medline: 15561786]
38. Graetz I, Gordon N, Fung V, Hamity C, Reed ME. The digital divide and patient portals: internet access explained differences in patient portal use for secure messaging by age, race, and income. Med Care 2016 Aug;54(8):772-779. [doi: 10.1097/MLR.0000000000000560] [Medline: 27314262]
39. Gao H, Dutta M, Okoror T. Listening to Chinese immigrant restaurant workers in the Midwest: application of the culture-centered approach (CCA) to explore perceptions of health and health care. Health Commun 2016;31(6):727-737. [doi: 10.1080/10410236.2014.989383] [Medline: 26575110]
40. Dutta MJ. Culture-centered approach in addressing health disparities: communication infrastructures for subaltern voices. Commun Methods Meas 2018 Oct 2;12(4):239-259. [doi: 10.1080/19312458.2018.1453057]
41. Dutta MJ. A culture-centered approach to listening: voices of social change. Int J List 2014 May 4;28(2):67-81. [doi: 10.1080/10904018.2014.876266]
42. Morahan-Martin JM. How internet users find, evaluate, and use online health information: a cross-cultural review. Cyberpsychol Behav 2004 Oct;7(5):497-510. [doi: 10.1089/cpb.2004.7.497] [Medline: 15667044]

XSL•FO
RenderX

43. Singh S, Banerjee A. Internet and doctor-patient relationship: cross-sectional study of patients' perceptions and practices. Indian J Public Health 2019;63(3):215-219. [doi: 10.4103/ijph.IJPH_392_18] [Medline: 31552851]
44. Goldner M, Hale TM, Cotten SR, Stern MJ, Drentea P. The intersection of gender and place in online health activities. J Health Commun 2013;18(10):1235-1255. [doi: 10.1080/10810730.2013.778364] [Medline: 23886026]
45. The 50th statistical report on China's internet development [Article in Chinese]. China Internet Network Information. URL: http://www.gov.cn/xinwen/2022-08/31/content_5707608.htm [accessed 2025-04-03]
46. Zach L, Dalrymple PW, Rogers ML, Williver-Farr H. Assessing internet access and use in a medically underserved population: implications for providing enhanced health information services. Health Info Libr J 2012 Mar;29(1):61-71. [doi: 10.1111/j.1471-1842.2011.00971.x] [Medline: 22335290]
47. Ibrahim AMS, Vargas CR, Koolen PGL, Chuang DJ, Lin SJ, Lee BT. Readability of online patient resources for melanoma. Melanoma Res 2016 Feb;26(1):58-65. [doi: 10.1097/CMR.0000000000000210] [Medline: 26479217]
48. Vargas CR, Chuang DJ, Ganor O, Lee BT. Readability of online patient resources for the operative treatment of breast cancer. Surgery 2014 Aug;156(2):311-318. [doi: 10.1016/j.surg.2014.03.004] [Medline: 24953268]
49. McCloud RF, Okechukwu CA, Sorensen G, Viswanath K. Beyond access: barriers to internet health information seeking among the urban poor. J Am Med Inform Assoc 2016 Nov;23(6):1053-1059. [doi: 10.1093/jamia/ocv204] [Medline: 27206459]
50. Wang Z, Wang S, Zhang Y, Jiang X. Social media usage and online professionalism among registered nurses: a cross-sectional survey. Int J Nurs Stud 2019 Oct;98:19-26. [doi: 10.1016/j.ijnurstu.2019.06.001] [Medline: 31255853]
51. Notification of the National Health Commission on printing the guidelines of generating and spreading health science popularization information [Article in Chinese]. National Health Commission of the People's Republic of China. 2015. URL: http://www.nhc.gov.cn/cms-search/xxgk/getManuscriptXxgk.htm?id=5fe32b5a1a8243e2bd819f9eeebfd8b1 [accessed 2025-04-03]
52. Li W, Han LQ, Guo YJ, Sun J. Using WeChat official accounts to improve malaria health literacy among Chinese expatriates in Niger: an intervention study. Malar J 2016 Nov 24;15(1):567. [doi: 10.1186/s12936-016-1621-y] [Medline: 27881122]
53. Zhou J, Liu F, Zhou H. Understanding health food messages on Twitter for health literacy promotion. Perspect Public Health 2018 May;138(3):173-179. [doi: 10.1177/1757913918760359] [Medline: 29513075]
54. Ma X, Lu J, Liu W. Influencing factors on health information to improve public health literacy in the official WeChat account of Guangzhou CDC. Front Public Health 2021;9:657082. [doi: 10.3389/fpubh.2021.657082] [Medline: 34414152]
55. Zhang X, Wen D, Liang J, Lei J. How the public uses social media WeChat to obtain health information in China: a survey study. BMC Med Inform Decis Mak 2017 Jul 5;17(Suppl 2):66. [doi: 10.1186/s12911-017-0470-0] [Medline: 28699549]
56. A Rahim AI, Ibrahim MI, A Salim FN, Ariffin MAI. Health information engagement factors in Malaysia: a content analysis of Facebook use by the Ministry of Health in 2016 and 2017. Int J Environ Res Public Health 2019 Feb 18;16(4):591. [doi: 10.3390/ijerph16040591] [Medline: 30781691]
57. Ahmed Shaikh Z, Ali Sathio A, Ali Laghari A, Ahmed Memon M, Hussain Mirani I. Study of the role of new technologies in pharmaceutical industry. J Pharm Res Int 2019;31:1-11. [doi: 10.9734/jpri/2019/v31i630313]
58. Miller LMS, Bell RA. Online health information seeking: the influence of age, information trustworthiness, and search challenges. J Aging Health 2012 Apr;24(3):525-541. [doi: 10.1177/0898264311428167] [Medline: 22187092]
59. Zhang X, Wang X, Tian F, Xu D, Fan L. Anticipating the antecedents of feedback-seeking behavior in digital environments: a socio-technical system perspective. Internet Res 2023 Mar 28;33(1):388-409. [doi: 10.1108/INTR-09-2021-0672]
60. Zhang X, Yan X, Cao X, Sun Y, Chen H, She J. The role of perceived e-health literacy in users' continuance intention to use mobile healthcare applications: an exploratory empirical study in China. Inf Technol Dev 2018 Apr 3;24(2):198-223. [doi: 10.1080/02681102.2017.1283286]

Abbreviations

OHIS: online health information–seeking
OR: odds ratio

Edited by A Mavragani; submitted 06.01.24; peer-reviewed by PCI Pang, X Zhang; revised version received 08.02.25; accepted 20.02.25; published 05.05.25.

Please cite as:
Rong H, Lu L, He M, Guo T, Li X, Tao Q, Li Y, Zheng C, Zhang L, Li F, Yi D, Lei E, Luo T, Yang Q, Chen JA
Online Health Information–Seeking Behaviors Among the Chongqing Population: Cross-Sectional Questionnaire Study
JMIR Form Res 2025;9:e56028
URL: https://formative.jmir.org/2025/1/e56028
doi:10.2196/56028

XSL•FO
RenderX

© Honghui Rong, Lu Lu, Miao He, Tian Guo, Xian Li, Qingliu Tao, Yixin Li, Chuanfen Zheng, Ling Zhang, Fengju Li, Dali Yi, Enyu Lei, Ting Luo, Qinghua Yang, Ji-an Chen. Originally published in JMIR Formative Research (https://formative.jmir.org), 5.5.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

YouTube User Traffic to Paired Epilepsy Education Videos in English and Spanish: Comparative Study

Luna Kimahri Varela, MS; Stephanie Horton, MHA; Ahmed Abdelmoity, MD; Jean-Baptiste Le Pichon, MD, PhD; Mark A Hoffman, PhD

Children's Mercy Hospital Kansas City, Children's Mercy Research Institute, 2401 Gillham Road, Kansas City, MO, United States

Corresponding Author:
Mark A Hoffman, PhD
Children's Mercy Hospital Kansas City, Children's Mercy Research Institute, 2401 Gillham Road, Kansas City, MO, United States

Abstract

Background: Effectively managing epilepsy in children necessitates the active engagement of parents, a factor that is reliant on their understanding of this neurological disorder. Widely available, high-quality, patient-focused, bilingual videos describing topics important for managing epilepsy are limited. YouTube Analytics is a helpful resource for gaining insights into how users of differing backgrounds consume video content.

Objective: This study analyzes traffic to paired educational videos of English and Spanish versions of the same content. By examining the use patterns and preferences of individuals seeking information in different languages, we gained valuable insights into how language influences the use of clinical content.

Methods: Physician experts created epilepsy management videos for the REACT (Reaching Out for Epilepsy in Adolescents and Children Through Telemedicine) YouTube channel about 17 subjects, with an English and Spanish version of each. The Children's Mercy Kansas City neurology clinic incorporated these into the department's educational process. YouTube Analytics enabled analysis of traffic patterns and video characteristics between September 2, 2021, and August 31, 2023.

Results: The Spanish group had higher engagement and click-through rates. The English versions of all videos had 141,605 total impressions, while impressions for the Spanish versions totaled 156,027. The Spanish videos had 11,339 total views, while the English videos had 3366. The views per month were higher for the Spanish videos (mean 472, SD 292) compared to the English set (mean 140, SD 91; $P<.001$). The two groups also differed in search behavior and external traffic sources, with WhatsApp driving more traffic to the Spanish videos than the English versions (94 views compared to 1). The frequency of search terms used varied by language. For example, "tonic clonic" was the most frequent term (n=372) resulting in views for English videos, while "tipos de convulsiones" (types of convulsions) was the most common expression (n=798) resulting in views for Spanish videos. We noted increased monthly views for all videos after adding tags on YouTube. Before tagging, the mean number of views per month for the English-language group was 61 (SD 28), which increased to 220 (SD 53) post tagging. A similar trend can be observed in the Spanish-language group as well. Before tagging, the mean number of monthly views was 201 (SD 71), which increased to 743 (SD 144) after tagging.

Conclusions: This study showed high traffic for Spanish video content related to epilepsy in a set of paired English/Spanish videos. This highlights the importance of bilingual health content and optimizing video content based on viewer preferences and search behavior. Understanding audience engagement patterns through YouTube Analytics can further enhance the dissemination of clinical video content to users seeking content in their primary language, and tagging videos can have a substantial impact on views.

(JMIR Form Res 2025;9:e56720) doi:10.2196/56720

KEYWORDS

epilepsy; patient education; informatics; social media; biomedical research; social determinants of health; accessibility; engagement; comparative analysis; clinical videos; English; Spanish; neurological disorder; YouTube; bilingual; audience engagement; clinical knowledge

XSL•FO
RenderX

Introduction

Video Health Content

Access to informative and reliable clinical content is crucial for managing complex clinical conditions, especially for individuals in rural areas or with limited access to specialist care [1]. Epilepsy, a common chronic condition, highlights the challenges faced by patients needing access to clinical specialists. Children and youth with epilepsy and their families need accurate and trustworthy information to manage their condition effectively [2].

Social media platforms, particularly YouTube, have become significant sources of information for many people. With billions of users monthly, YouTube has established its presence in over 100 countries worldwide and offers localized versions in 80 languages to serve a diverse global population [3]. However, poor-quality health videos on YouTube can be misleading, but they are still promoted via YouTube's search algorithm that prioritizes popularity, relevance, and view history over content quality and accuracy [4]. This presents a significant challenge for individuals who seek health information without proper guidance or supervision, as they are increasingly exposed to unverified and partially misleading content. The consequences of such exposure can include the promotion of unhealthy habits and activities [5]. A 2023 study about video-sharing platforms revealed a strong negative correlation between high-quality content and popularity regarding videos providing accurate information on cancer and nutrition [6]. Out of 1200 videos, the most reliable videos on YouTube were from universities, professional organizations, and nonprofit physician groups, while unreliable videos included content related to treatments that scored low on the Journal of American Medical Association scale for quality medical content [7]. Analysis of video content using the DISCERN instrument yielded similar results for diabetes content [8]. High-quality health care educational resources developed by experts on widely accessible platforms such as YouTube offer an approach to mitigating these concerns. Through analytics available through YouTube, we can understand the factors influencing viewership and engagement with clinical videos.

REACT

Through the REACT (Reaching Out for Epilepsy in Adolescents and Children Through Telemedicine) project, Children's Mercy Kansas City has developed online educational resources for families and children and youth with epilepsy [9]. One of these resources includes online videos covering different aspects of managing epilepsy and seizures, such as general content about seizures, types of seizures, seizure precautions, alternative treatments, and frequently asked questions. The goal was to provide access to patients and their families, primarily those who spoke English or Spanish. This design provided an opportunity to compare the use of videos with the same clinical content. A preliminary analysis of YouTube Analytics traffic patterns for one of these videos ("What is an absence seizure?") showed high use for the Spanish version [7]. This paper presents a comprehensive traffic analysis of the REACT YouTube channel, comparing the paired English and Spanish versions of otherwise identical content.

Methods

Epilepsy Video Development and Deployment

Videos for epilepsy-related topics prioritized by the clinical team were developed in English and Spanish. A multistep approach was taken when translating the videos from English to Spanish. First, the bilingual research team translated the scripts into Spanish. The scripts were reviewed by bilingual and culturally competent staff before recording the videos. The resulting videos were uploaded to the REACT YouTube channel. The videos for each topic are considered as episodes with two versions.

YouTube Analytics

YouTube Analytics for the REACT channel were accessed to evaluate online traffic patterns to the video content [10]. No personal identifiers were available or sought from YouTube. We did not attempt to correlate patient-prescribed views with the traffic patterns evaluated in this work. GraphPad Prism 9.5.1 (GraphPad Software, Inc) was used to quantify video characteristics, with continuous variables presented as means with SDs and ranges.

Tags are descriptive keywords used on YouTube to improve the discoverability of a video. In August 2022, the study team added tags to each of the videos. To analyze the impact of these tags, total monthly views were separated into a pretagging period (September 2021 to August 2022) and a posttagging period (September 2022 to August 2023).

Impressions are the number of times a video's thumbnail is presented to a user on YouTube. The click-through rate was provided by YouTube Advanced Analytics and represents the percentage of REACT channel thumbnails that were clicked by viewers.

Ethical Considerations

All study data used were anonymous.

Results

Epilepsy Video Development and Deployment

Clinical experts developed video content for 17 topics related to treating epilepsy and seizures. The Children's Mercy Kansas City neurology clinic adopted these videos as part of the patient departure educational process. Children and youth with epilepsy and their families were "prescribed" to view at least one video after each visit. The information and prescribed video were printed in the educational department and given to patients so they could scan a QR code that led them to a REDCapÒ survey. This survey included demographic information, a link to the prescribed YouTube video, and a pre- and postsurvey. Children and youth with epilepsy and families also had the option to view more than one video if desired. The first videos were uploaded to the REACT YouTube channel on August 7, 2020, and all videos were made publicly available by September 2, 2021 [11].

XSL•FO
RenderX

YouTube Analytics

We evaluated traffic patterns between September 2, 2021, when all videos were available, and August 31, 2023. In the aggregate, views per day for the Spanish videos surpassed the English versions (Figure 1A). Compared to the English videos, traffic to the Spanish-language videos was significantly higher ($P<.001$) based on click-through rate (Figure 1B). Some episodes, such as 4 and 14, exhibited different click-through rates between the two languages.

XSL•FO
RenderX

Figure 1. Daily views, by language, of epilepsy YouTube channel. (1A) The total daily views of the REACT (Reaching Out for Epilepsy in Adolescents and Children Through Telemedicine) YouTube channel from September 2, 2021, to August 31, 2023. The x-axis represents the total daily views, and a smoothing spline curve is included. (1B) Average daily click-through rate per video grouped by language. ROSA: robotic stereotactic assistance.

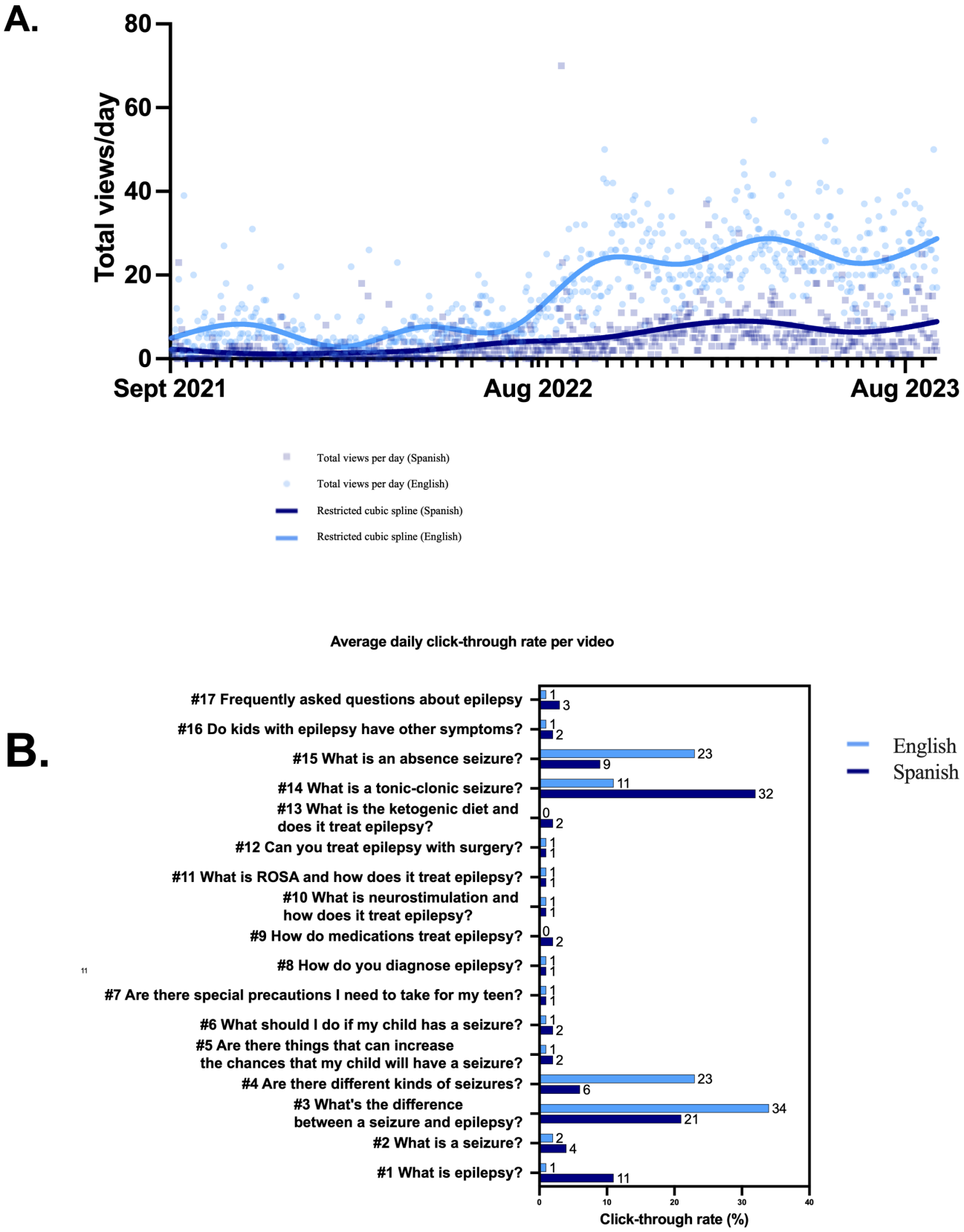

The addition of tags correlates with higher traffic (Figure 2A). The highest total monthly views for the English-language group of videos was 320 in January 2023, while the Spanish-language group of videos' highest monthly views were 939 in March 2023. To further assess the impact of the tags, the total monthly views were divided into two periods: pretagging period (September 2021 to August 2022) and posttagging period (September 2022 to August 2023). In the pretagging period, the

XSL•FO
RenderX

views per month for the English-language group increased after tagging (pretagging mean 61, SD 28; posttagging mean 220, SD 53). A similar trend was observed in the Spanish-language group as well (pretagging mean 201, SD 71; posttagging mean 743, SD 114). Both the English- and Spanish-language videos showed significant increases in traffic after the addition of tags ($P<.001$; Figure 2B).

Figure 2. Impact of tagging: Total views per month grouped by language. (2A) Total views per month. Monthly views for all videos on the REACT (Reaching Out for Epilepsy in Adolescents and Children Through Telemedicine) YouTube channel were collected from September 2, 2021, to August 31, 2023, and grouped by language. The dotted line represents the addition of tags. Each data point indicates the total views per language group for that month. (2B) Mean monthly views pre- versus posttagging periods. Comparison of mean total monthly views before and after adding tags in August 2022 for English- and Spanish-language group videos. Paired *t* tests showed a significant difference in views. ****$P<.001$.

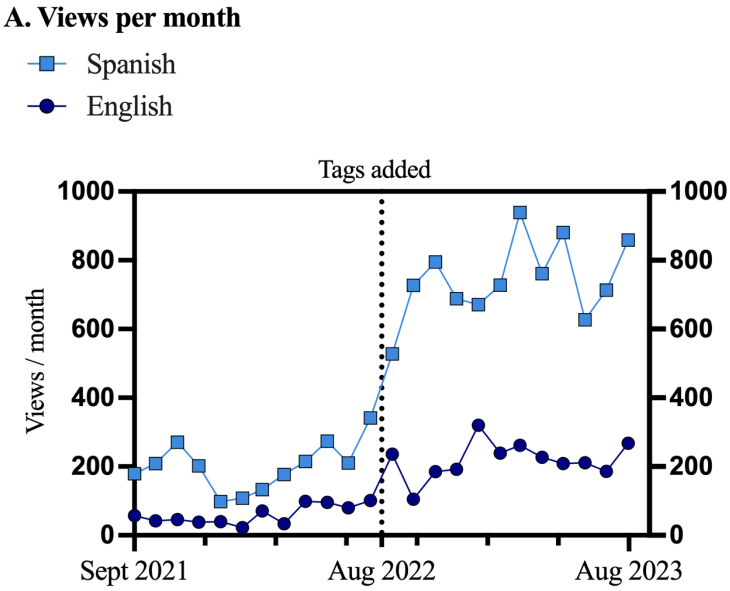

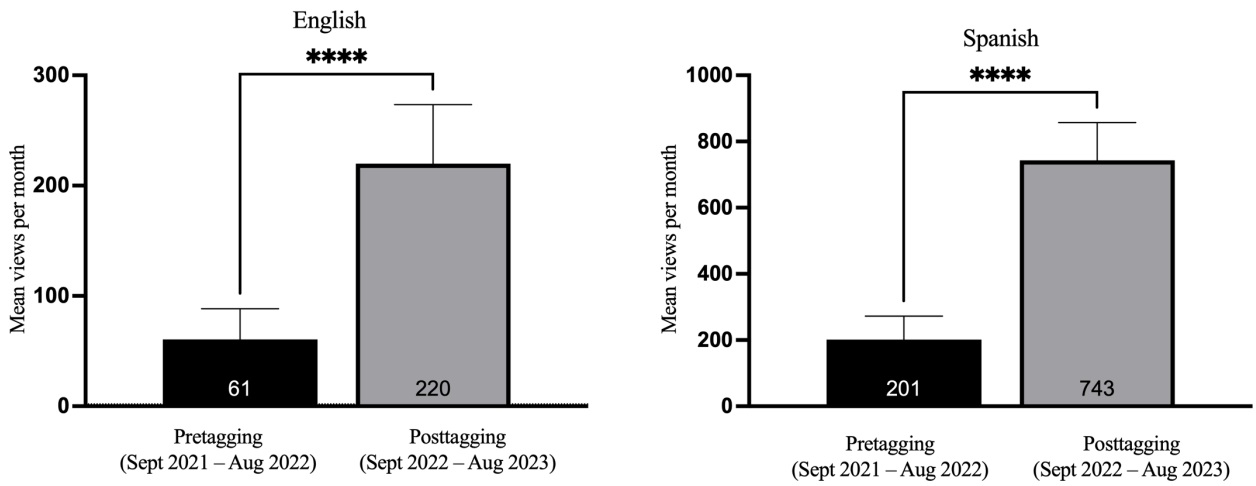

Traffic to individual episodes demonstrated differing patterns. The Spanish versions of videos 3 and 4 showed a strong association between impressions and views (Figure 3A-D). For video 14, impressions and views were similar for the Spanish version, but for the English version, impressions increased earlier (Figure 3E-F). In the English category, video 14 had the highest impressions (n=106,936) and views (n=1087; Table 1). In the Spanish category, video 15 had the highest impressions (n=49,039), followed closely by video 3 (n=41,892). However, video 3 had more views (n=3839) than video 14.

XSL•FO
RenderX

Figure 3. Monthly impressions and views for videos 3, 4, and 14. (3A, C, and E) Monthly views per month across time for videos 3, 4, and 14 from September 2021 to August 2023. (3B, D, and F) Monthly impressions across time for videos 3, 4, and 14.

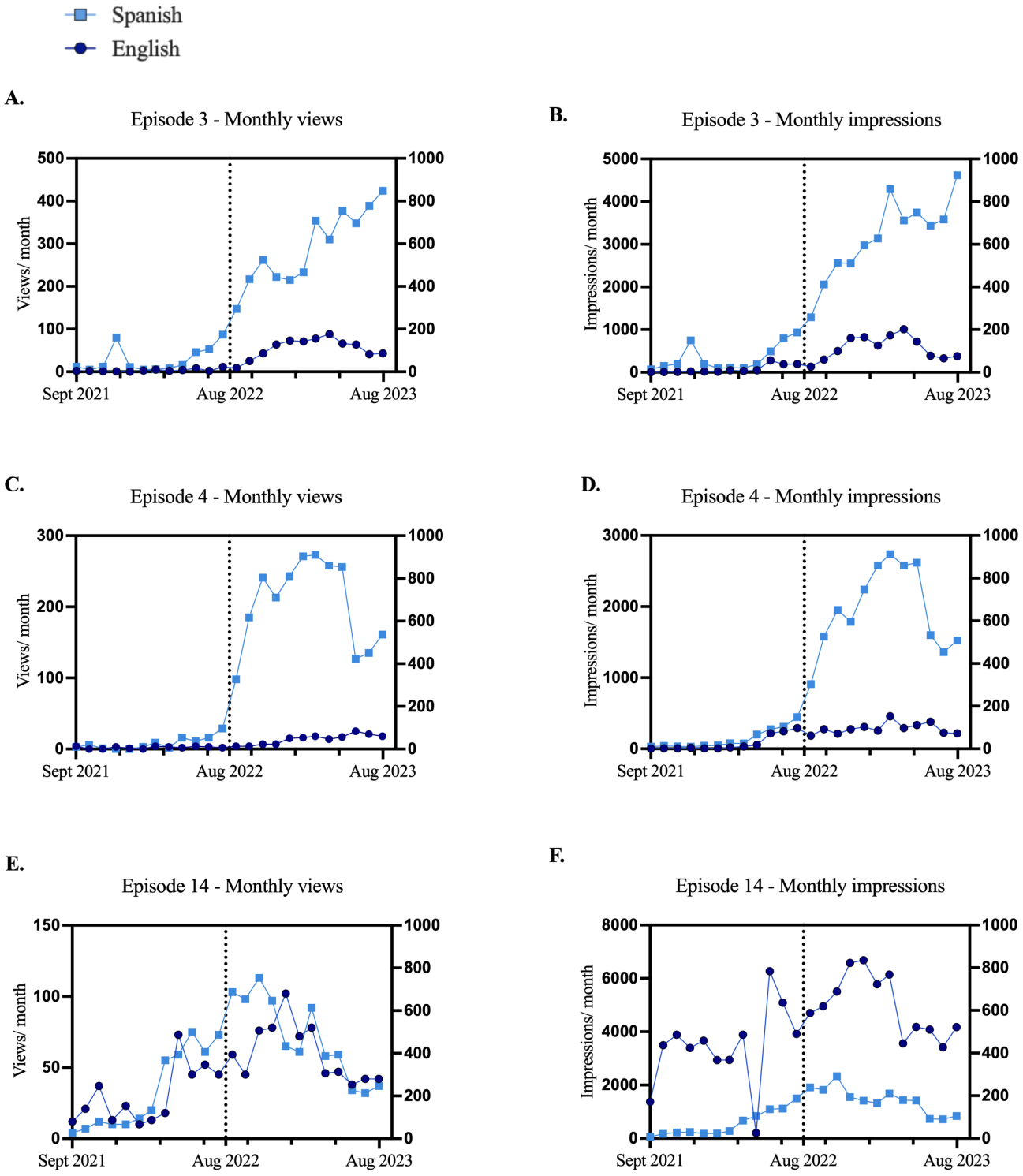

XSL•FO
RenderX

Table . Impressions, click-through rate, and views for each epilepsy video.

	Video title		Total impressions, n		Click-through rate		Total views, n	
	English title	Spanish title	English	Spanish	English	Spanish	English	Spanish
1	What is epilepsy?	¿Que es la epilepsia?	454	806	4	3	352	106
2	What is a seizure?	¿Qué es una convulsión?	718	3523	4	5	144	255
3	What's the difference between a seizure and epilepsy?	¿Cuál es la diferencia entre la epilepsia y una convulsión?	7727	41,892	7	7	706	3839
4	Are there different kinds of seizures?	¿Hay diferentes tipos de convulsiones?	4327	25,084	2	8	192	2557
5	Are there things that can increase the chances that my child will have a seizure?	¿Hay cosas que pueden aumentar las posibilidades de que tenga una convulsión?	341	902	5	4	51	66
6	What should I do if my child has a seizure?	¿Qué debo hacer si mi hijo tiene una convulsión?	1434	781	1	5	73	75
7	Are there special precautions I need to take for my teen?	¿Se deben tomar precauciones si mi hijo adolescente tiene epilepsia?	276	945	3	6	35	80
8	How do you diagnose epilepsy?	¿Cómo se diagnostica la epilepsia?	318	1519	3	5	33	111
9	How do medications treat epilepsy?	¿Cómo tratan los medicamentos la epilepsia?	1894	543	2	2	62	25
10	What is neurostimulation and how does it treat epilepsy?	¿Qué es la neuroestimulación y cómo trata la epilepsia?	355	2241	7	5	47	131
11	What is ROSA[a] and how does it treat epilepsy?	¿Qué es ROSA y como trata la epilepsia?	326	606	4	4	25	35
12	Can you treat epilepsy with surgery?	¿Se puede tratar la epilepsia con cirugía?	214	1481	4	4	22	74
13	What is the ketogenic diet and does it treat epilepsy?	¿Qué es la dieta Ketogénica?	1106	439	4	4	79	26

XSL•FO
RenderX

	Video title		Total impressions, n		Click-through rate		Total views, n	
	English title	Spanish title	English	Spanish	English	Spanish	English	Spanish
14	What is a tonic-clonic seizure?	¿Cómo cuido a mi hijo con convulsiones tónico´-clónicas?	106,936	23,727	1	4	1087	1249
15	What is an absence seizure?	¿Qué es una crisis de ausencia?	13,319	49,039	2	4	304	2562
16	Do kids with epilepsy have other symptoms?	Otros síntomas de epilepsia	950	1968	2	4	65	115
17	Frequently Asked Questions about Epilepsy	Nuestra experta responde a sus preguntas frecuentes	910	531	6	4	89	33
Totals	__[b]	—	141,605	156,027	1	6	3366	11,339

[a]ROSA: robotic stereotactic assistance.

[b]Not applicable.

While the location of the viewer was only available for 2096 (14.3%) of the 14,705 views, we observed that countries with predominantly non–English-speaking populations tended to have higher view counts reported by YouTube than primarily English-speaking countries. Users in the United States generated the most views (n=884) for the Spanish language, followed by Mexico (n=859) and Argentina (n=110; Multimedia Appendix 1).

The results from the analysis of YouTube search terms used to find English and Spanish videos are presented in Table 2. The top searches that resulted in views for English videos were “tonic,” “clonic,” or “tonic-clonic.” The term “absence seizure” also generated strong interest, resulting in 118 views. Spanish videos had more views for the comparable search term “crisis de ausencia” or “absence seizure,” with 527 views. In comparison, the terms “convulsiones tonico clonico” or “tonic clonic seizures” only garnered 118 views. The search term “tipos de convulsiones” or “types of seizures” resulted in the most views for the Spanish-language group, with 798 views. These findings suggest differences in viewers’ search behavior when seeking content in different languages on YouTube.

XSL•FO
RenderX

Table . Most frequent YouTube search terms leading users to REACT (Reaching Out for Epilepsy in Adolescents and Children Through Telemedicine) content.

Search terms	Views, n
English language	
“tonic clonic” or “clonic” or tonic”	372
“Absence Seizure”	64
“what is a seizure”	17
“difference between seizure and epilepsy”	16
“seizure”	15
“types of seizures”	10
“epilepsy seizure”	8
“different types of seizures”	7
“difference between epilepsy and seizure”	5
“what is epilepsy”	5
Spanish language	
“tipos de convulsiones” (types of seizures)	798
“crisis de ausencia” (absence seizures)	527
“convulsiones tonico clonicas” (tonic-clonic seizures)	118
“que es una convulsión” (what is a seizure)	43
“convulsiones” (seizures)	40
“epilepsia” (epilepsy)	31
“que es la epilepsia” (what is epilepsy)	25
“como es una convulsión” (how is a seizure)	13
“convulsiones en adultos” (seizure’s in adults)	13
“convulsiones y epilepsia” (seizures and epilepsy)	13

Table 3 presents a comprehensive list of external traffic sources used to find video content on the REACT YouTube channel. English viewers predominantly accessed the content through the Children’s Mercy website, with 443 views, compared to the Spanish count of 151. Spanish-language content was embedded on an epilepsy education website maintained by the government of Chile and generated 121 views. Some of the Spanish-language video views originated from a Spanish-language occupational health education consultancy group’s website. Spanish-language viewers relied more heavily on WhatsApp as their preferred means of finding content, with 94 views compared to only 1 among English-language viewers.

XSL•FO
RenderX

Table . Traffic sources used by viewers to find REACT (Reaching Out for Epilepsy in Adolescents and Children Through Telemedicine) content.

Traffic source	Views, n	
	English language	Spanish language
Children's Mercy website	443	161
Google	201	73
YouTube	42	35
Yahoo	2	3
Instagram	1	0
WhatsApp	1	94
bing.com	1	0
live.com	1	5
telegram.org	1	0
revistachilenadeepilepsia.cl	0	121
WhatsApp Business	0	2
precoinprevencion.com	0	10
Facebook	0	12

Much of the traffic to the REACT channel originated from sources outside the YouTube platform. These include internet searches, direct URL entries, bookmarks, or embedded links. External traffic sources accounted for 38,172 (12.8%) of the 297,632 impressions. However, despite the relatively lower number of impressions, these external sources, especially WhatsApp, contributed significantly to the overall viewership. In contrast, videos directly delivered by YouTube accounted for 259,460 (87.2%) impressions. Further analysis examined views by internal versus external sources, grouped by language. The Spanish-language groups contributed 8697 views internally and 2642 views externally to the YouTube platform, while the English videos generated 2508 views from within YouTube and 858 from external sources.

We investigated the impact of YouTube algorithmic suggestions on driving views to REACT content. These suggestions were presented as pop-ups after a video was completed or while users were browsing the website. Within the English group, we observed that videos discussing atonic seizures, tonic-clonic seizures, and the differences between seizures and epilepsy received the highest number of views, with 61, 31, and 14 views, respectively (Multimedia Appendix 2). The Spanish videos addressing convulsions and absence seizures garnered the most attention from these algorithmic recommendations, with 146 and 96 views, respectively.

Discussion

YouTube reaches a larger audience than videos delivered through self-publication to isolated websites, for example, a university website [12]. This study evaluated the viewership of paired Spanish and English educational videos on the YouTube platform and noted substantial traffic to the Spanish-language versions relative to most of the English versions. One notable aspect of our work is the access to detailed analytics for videos with identical clinical content in English and Spanish, a distinguishing feature from other studies evaluating YouTube content for chronic disease [13]. The availability of such analytics provides valuable insights into the use patterns for clinical video content, allowing for informed content design and development strategies. YouTube analytics has substantial, yet underused, value to the developers of clinical content. While the primary intended use of YouTube metrics is to monitor and optimize ad views, we demonstrate that there is utility to these metrics for health education. By leveraging data-driven approaches, the quality and relevance of clinical content can be improved. This analysis enables the promotion of health equity by tailoring content and using search engine optimization features such as tagging to increase visibility.

Our analysis demonstrates a distinct difference in audience engagement between English- and Spanish-language videos providing the same content. Engagement is quantified by the number of views a video generates as well as other metrics, including view time, which we did not evaluate in this study. Overall, the Spanish-language group consistently exhibited higher levels of engagement, as evidenced by increased views and impressions. The modes of interaction differed; for example, Spanish speakers were more likely to use WhatsApp to access the content, consistent with the popularity of this messaging app among Spanish-speaking groups [14,15]. This finding highlights the importance of developing bilingual clinical content to meet the needs of diverse populations.

Furthermore, our study explored the correlation between impressions and views per month, providing insights into variations in monthly engagement across different videos.

The videos with the highest viewing rates were related to the clinical appearance of a seizure ("What's the difference between a seizure and epilepsy?" "Are there different kinds of seizures?" "What is a tonic-clonic seizure?" "What is an absence seizure?"). This is in line with the public misconceptions of what a seizure looks like [16] and further emphasizes the importance of

XSL•FO
RenderX

education regarding the many possible phenomenology of epileptic seizures [17].

While data about audience demographics was limited, our analysis showed that users in countries with predominantly non–English-speaking populations tended to have higher reported view counts of the Spanish versions of a video than users in primarily English-speaking countries. This suggests a global demand for clinical content in languages other than English, highlighting the need for bilingual or multilingual approaches to reach diverse audiences. The search behavior of viewers also differed between languages, as reflected in the top search terms leading to views for English and Spanish videos. Understanding these differences can inform content creators about the keywords and topics that resonate with their audiences, enabling them to optimize their content accordingly. For example, search terms most often leading users to a video can be added as tags.

External traffic sources (non-YouTube) significantly drove viewership. This highlights the importance of leveraging external platforms, such as websites and social media, to promote and distribute clinical video content posted on YouTube. WhatsApp was a frequent source of traffic to the Spanish videos, consistent with common use of this app in non–English-speaking countries [18]. Tailoring strategies to effectively reach and engage different target audiences based on language preferences and behaviors is crucial for maximizing the impact of clinical content dissemination. Our findings also demonstrate that the implementation of tags significantly improved the visibility and accessibility of the videos, leading to a substantial increase in total monthly views. This is consistent with previous research that has shown the effectiveness of tags in enhancing the discoverability of content on YouTube [19].

Finally, our analysis of suggestions generated by the YouTube platform showed variations in viewer preferences between the English and Spanish groups. This underscores the influence of recommendations in directing viewers toward specific content related to seizures and epilepsy. Content creators should consider these recommendations to increase viewership and engagement when developing and promoting their videos.

This study has known limitations. We focused solely on YouTube to evaluate clinical video viewership, and we acknowledge that other platforms may exhibit different use patterns and levels of audience engagement. The analysis was limited to English and Spanish videos, excluding other languages that may generate significant viewership. User privacy settings influence whether their location is visible in analytics, limiting access to granular geographic data and other demographic data (race, gender, age). Furthermore, the study did not explore the long-term effects of video viewership on health outcomes or behavior change. Lastly, we did not consider the influence of external factors, such as marketing or promotion efforts, on video viewership.

We demonstrated the effectiveness of paired educational videos in reaching a global audience well beyond the intended focus of the in-clinic recommendations. Techniques such as tags improve the discoverability and viewership of clinical videos on YouTube. The availability of detailed analytics provides valuable insights into audience engagement and use patterns, enabling informed content design and development strategies. Bilingual or multilingual approaches are essential for reaching diverse populations and promoting health equity. Leveraging external (outside of the YouTube platform) traffic sources and optimizing video content based on search behavior and viewer preferences can further enhance the impact of clinical video dissemination, for example, top YouTube searches that resulted in views related to the differences between types of seizures. The development of paired videos enabled us to perform side-by-side comparisons of traffic patterns and video characteristics on a specific health topic. Through this work, we elucidated practices to improve clinical content available on the YouTube platform for future health content creators seeking to support non–English-speaking users.

Acknowledgments

This work was funded in part by the Health Resources and Services Administration Awareness and Access to Care for Children and Youth with Epilepsy award H98MC332390100. We appreciate the contributions of the Division of Neurology experts who developed the video content and the marketing team of Children's Mercy Kansas City for the video editing. We thank Rose Reynolds for the helpful discussions.

Data Availability

The datasets generated or analyzed during this study are available at [20].

Conflicts of Interest

None declared.

Multimedia Appendix 1
Total view count by country.
[DOCX File, 14 KB - formative_v9i1e56720_app1.docx]

Multimedia Appendix 2

XSL•FO
RenderX

Top 10 traffic sources to REACT (Reaching Out for Epilepsy in Adolescents and Children Through Telemedicine) from other YouTube content grouped by language.
[DOCX File, 16 KB - formative_v9i1e56720_app2.docx]

References

1. Cassidy L, Quirke MB, Alexander D, et al. Integrated care for children living with complex care needs: an evolutionary concept analysis. Eur J Pediatr 2023 Apr;182(4):1517-1532. [doi: 10.1007/s00431-023-04851-2] [Medline: 36780041]
2. Ellis N, Upton D, Thompson P. Epilepsy and the family: a review of current literature. Seizure 2000 Jan;9(1):22-30. [doi: 10.1053/seiz.1999.0353] [Medline: 10667959]
3. YouTube by the numbers secondary YouTube by the numbers 2023. YouTube Blog. URL: https://blog.youtube/press [accessed 2025-03-03]
4. Osman W, Mohamed F, Elhassan M, Shoufan A. Is YouTube a reliable source of health-related information? A systematic review. BMC Med Educ 2022 May 19;22(1):382. [doi: 10.1186/s12909-022-03446-z] [Medline: 35590410]
5. Nour MM, Nour MH, Tsatalou OM, Barrera A. Schizophrenia on YouTube. Psychiatr Serv 2017 Jan 1;68(1):70-74. [doi: 10.1176/appi.ps.201500541] [Medline: 27524369]
6. Basch CH, Menafro A, Mongiovi J, Hillyer GC, Basch CE. A content analysis of YouTube videos related to prostate cancer. Am J Mens Health 2017 Jan;11(1):154-157. [doi: 10.1177/1557988316671459] [Medline: 27694554]
7. Demirci A, Başar H. YouTube is an unreliable source of information about delayed ejaculation treatment. Urologia 2024 May;91(2):403-412. [doi: 10.1177/03915603231204077] [Medline: 37818833]
8. Birch EM, Leziak K, Jackson J, Dahl E, Niznik CM, Yee LM. Content quality of YouTube videos about gestational diabetes: systematic evaluation. JMIR Diabetes 2022 Apr 7;7(2):e30156. [doi: 10.2196/30156] [Medline: 35389355]
9. Le Pichon JB, Horton S, Abdelmoity O, et al. The use of virtual tools in narrowing the impact of health disparities in neurology. Front Pediatr 2022 Oct 14;10:1028833. [doi: 10.3389/fped.2022.1028833] [Medline: 36313873]
10. AlHasan A. The power of YouTube videos for surgical journals. Surgery 2023 Sep;174(3):744-746. [doi: 10.1016/j.surg.2023.05.042] [Medline: 37419760]
11. Children's Mercy KC Neurology-REACT YouTube page. URL: https://www.youtube.com/@reactcmh [accessed 2025-02-28]
12. Topps D, Helmer J, Ellaway R. YouTube as a platform for publishing clinical skills training videos. Acad Med 2013 Feb;88(2):192-197. [doi: 10.1097/ACM.0b013e31827c5352] [Medline: 23269305]
13. Zaila KE, Osadchiy V, Anderson AS, Eleswarapu SV, Mills JN. Popularity and worldwide reach of targeted, evidence-based internet streaming video interventions focused on men's health topics. Transl Androl Urol 2020 Jun;9(3):1374-1381. [doi: 10.21037/tau-20-580] [Medline: 32676422]
14. Anderson E, Koss M, Castro Luque AL, Garcia D, Lopez E, Ernst K. WhatsApp-based focus groups among Mexican-origin women in Zika risk area: feasibility, acceptability, and data quality. JMIR Form Res 2021 Oct 28;5(10):e20970. [doi: 10.2196/20970] [Medline: 34709185]
15. Díaz EN, Le Q, Campos D, et al. Central American parents' preferences for content and modality for a family-centered intervention to promote healthful energy balance-related behaviors of their preschool-age children. Int J Environ Res Public Health 2022 Apr 21;19(9):5080. [doi: 10.3390/ijerph19095080] [Medline: 35564475]
16. Kerson JF, Kerson TS, Kerson LA. The depiction of seizures in film. Epilepsia 1999 Aug;40(8):1163-1167. [doi: 10.1111/j.1528-1157.1999.tb00836.x] [Medline: 10448833]
17. Kilinç S, Campbell C. "It shouldn't be something that's evil, it should be talked about": a phenomenological approach to epilepsy and stigma. Seizure 2009 Dec;18(10):665-671. [doi: 10.1016/j.seizure.2009.09.001] [Medline: 19819737]
18. Rosen AO, Holmes AL, Balluerka N, et al. Is social media a new type of social support? Social media use in Spain during the COVID-19 pandemic: a mixed methods study. Int J Environ Res Public Health 2022 Mar 26;19(7):3952. [doi: 10.3390/ijerph19073952] [Medline: 35409634]
19. Zhou R, Xia D, Wan J, Zhang S. An intelligent video tag recommendation method for improving video popularity in mobile computing environment. IEEE Access 2019 Dec;8. [doi: 10.1109/ACCESS.2019.2961392]
20. Data from: youtube user traffic to paired epilepsy education videos in english and spanish: comparative study. SHARE. 2024. URL: https://scholarlyexchange.childrensmercy.org/data_sets/3/ [accessed 2025-03-13]

Abbreviations

REACT: Reaching Out for Epilepsy in Adolescents and Children Through Telemedicine

XSL•FO
RenderX

Edited by A Mavragani; submitted 24.01.24; peer-reviewed by J Huber; revised version received 08.11.24; accepted 10.11.24; published 13.03.25.

Please cite as:
Varela LK, Horton S, Abdelmoity A, Le Pichon JB, Hoffman MA
YouTube User Traffic to Paired Epilepsy Education Videos in English and Spanish: Comparative Study
JMIR Form Res 2025;9:e56720
URL: https://formative.jmir.org/2025/1/e56720
doi:10.2196/56720

© Luna Kimahri Varela, Stephanie Horton, Ahmed Abdelmoity, Jean-Baptiste Le Pichon, Mark A Hoffman. Originally published in JMIR Formative Research (https://formative.jmir.org), 13.3.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Stigma of Dementia on Social Media During World Alzheimer's Awareness Month: Thematic Analysis of Posts

Juanita-Dawne Bacsu[1], PhD; Jasmine Cassy Mah[2], PhD; Ali Akbar Jamali[3], PhD; Christine Conanan[1], MN; Samantha Lautrup[1]; Corinne Berger[3], PhD; Dylan Fiske[4], BSc; Sarah Fraser[5], PhD; Anila Virani[1], PhD; Florriann Fehr[6], PhD; Alison L Chasteen[7], PhD; Zahra Rahemi[8], PhD; Shirin Vellani[9], PhD; Melissa K Andrew[10], PhD; Allison Cammer[11], PhD; Katherine S McGilton[12], PhD; Rory Gowda-Sookochoff[13], BCom; Kate Nanson[1], BSN, RN; Karl S Grewal[14], PhD; Raymond J Spiteri[3], PhD

[1]School of Nursing, Population Health and Aging Rural Research (PHARR) Centre, Thompson Rivers University, 805 TRU Way, Kamloops, BC, Canada

[10]Division of Geriatric Medicine, Dalhousie University, Halifax, NS, Canada

[11]College of Pharmacy and Nurtrition, University of Saskatchewan, Saskatoon, SK, Canada

[12]Knowledge, Innovation, Talent, Everywhere (KITE) Research Institute, Toronto Rehabilitation Institute, University Health Network, Toronto, ON, Canada

[13]Department of Psychology, University of Saskatchewan, Saskatoon, SK, Canada

[14]Department of Psychology, Glenrose Rehabilitation Hospital, Edmonton, AB, Canada

[2]Department of Medicine, Dalhousie University, Halifax, NS, Canada

[3]Department of Computer Science, University of Saskatchewan, Saskatoon, SK, Canada

[4]Department of Science, Thompson Rivers University, Kamloops, BC, Canada

[5]Interdisciplinary School of Health Sciences, Faculty of Health Sciences, University of Ottawa, Ottawa, ON, Canada

[6]School of Nursing, Thompson Rivers University, Kamloops, BC, Canada

[7]Department of Psychology, University of Toronto, Toronto, ON, Canada

[8]School of Nursing, Clemson University, Clemson, SC, United States

[9]Virtual Behavioural Medicine Program, Toronto Rehabilitation Institute, University Health Network, Toronto, ON, Canada

Corresponding Author:
Juanita-Dawne Bacsu, PhD
School of Nursing, Population Health and Aging Rural Research (PHARR) Centre, Thompson Rivers University, 805 TRU Way, Kamloops, BC, Canada

Abstract

Background: Dementia-related stigma is a significant global health concern. However, public awareness and education about dementia-related stigma remain limited, especially on social media. Examining dementia-related stigma on social media is critical because it impacts how the public perceives people living with dementia. By understanding dementia-related stigma on social media, we can develop educational strategies to target false stereotypes, beliefs, and misinformation to improve the quality of life of people living with dementia.

Objective: This study examines dementia-related stigma on the X platform (formerly Twitter) during World Alzheimer's Month to identify opportunities for intervention to address dementia-related stigma.

Methods: A total of 266,211 posts were scraped from X during the World Alzheimer's Awareness Month from September 1 - 30, 2022, a global advocacy campaign organized by Alzheimer's Disease International. We used filters to exclude non-English content, duplicate posts, and reply posts with missing content. To ensure rigor and trustworthiness in the research, several measures were employed, ranging from peer debriefing sessions to documenting the research process.

Results: After filtering the data, 1981 posts were examined using thematic analysis. A total of four main themes were identified including: (i) dementia stereotypes: "a burden to society"; (ii) discrimination and denied dignity: "discrimination exists in public spaces"; (iii) devaluing the lives of people with dementia: "society should legalize euthanasia"; and (iv) countering dementia-related stigma: "break down the stigma." Although the World Alzheimer's Awareness Month is helpful for raising awareness, more research is needed to address dementia-related stigma, stereotypes, and discrimination on social media.

Conclusions: By analyzing how stigma manifests on social media, our study sheds light on the dementia education and information needed to address false beliefs, misinformation, and dementia-related stigma. The findings from our study have important

XSL•FO
RenderX

implications for policymakers, health professionals, and community advocates working to design awareness campaigns to reduce dementia-related stigma on social media.

(JMIR Form Res 2025;9:e72775) doi:10.2196/72775

KEYWORDS

stigma; stereotypes; misinformation; dementia; Alzheimer disease; awareness; social media; Alzheimer's disease

Introduction

Dementia is a significant global health concern. Worldwide, there are an estimated 10 million new cases of dementia annually, with one new diagnosis every 3.2 seconds [1]. In 2019, the global financial cost of dementia was approximately US$ 1.3 trillion [2]. It is estimated that 57 million people live with dementia, with this number projected to triple to 153 million by 2050 with population growth and aging [3].

Although dementia is a global issue, there is limited knowledge and awareness about dementia. A survey from Alzheimer's International [1] found that over 70% of respondents mistakenly view dementia as a normal part of aging rather than a neurodegenerative disorder. Studies show that dementia-related stigma (negative attitudes, stereotypes, and discrimination) results in social exclusion and self-internalization, and it contributes to delayed dementia diagnoses [4-6]. Moreover, dementia-related stigma can lead to health inequity [7], negative stereotypes [8], and health care discrimination [9-12].

Emerging research has started to examine dementia discourse on social media. More specifically, recent studies have reported stigmatizing language on social media [13-17]. Cheng and colleagues [17] identified stigmatization as the main theme of most of the tweets on X (formerly Twitter) regarding dementia. Another study examined dementia-related stigma on social media during the COVID-19 pandemic and identified issues of misinformation, ageism, and dementia being used as an insult for political ridicule [6]. However, no existing studies have examined the impact of global awareness campaigns on dementia-related stigma.

This study examines dementia-related stigma on X during World Alzheimer's Month, a global advocacy campaign organized by Alzheimer's Disease International [18]. Our study used data from X because it is a widely used platform for raising dementia awareness [19] and advocacy [20], but it has also been used to spread misinformation [21]. The X platform provides diverse insight with approximately 586 million active monthly users [22]. Moreover, emerging studies on social media awareness campaigns on health issues such as mental illness may be effective in increasing awareness, promoting help-seeking behavior [23], and reducing stigma [24]. Accordingly, X provides an effective platform to analyze dementia-related discourse during World Alzheimer's Month to identify opportunities for intervention to reduce dementia-related stigma.

By analyzing how stigma manifests on social media, our study sheds light on the dementia education and information needed to address false beliefs, misinformation, and dementia-related stigma. Additionally, this study sheds light on the role of social media in shaping public perceptions of dementia to promote more inclusive dementia discourse. Findings from our study have important implications for policymakers, health professionals, and community advocates working to design awareness campaigns to reduce dementia-related stigma on social media.

Methods

Ethical Considerations

This research did not require ethics board review because it only used secondary data collected from publicly available posts, which are not classified as human subject research. The Panel on Research Ethics [25] suggests that social media research does not require research ethics board review when the study focuses solely on information that is in the public domain and the people posting the information have no reasonable belief or expectation of privacy. Similarly, the Journal of Medical Internet Research [26] notes that the analysis of large volumes of text (ie, tweets) that are publicly available on X are not typically deemed human subject research. However, we omitted the usernames and Twitter handles from our findings to help support the anonymity of the X users.

Search Strategy

We bought a developer account for the X platform to gain access to the application programming interface (API). More specifically, the API enables developers to use data to generate material from the X platform. Using the Tweepy library [27] in Python and X's API, we collected posts published during the annual World Alzheimer's Awareness Month from September 1, 2022, to September 30, 2022.

The keyword search strategy was developed by examining commonly used terms and phrases on X during the World Alzheimer's Awareness Month. We also organized team meetings to develop, review, and finalize the phrases based on their usage on the X platform. For example, we ran test searches in real time to observe the usage of specific phrases related to the World Alzheimer's Awareness Month. This process enabled us to select terms that were being shared on X. This search strategy was based on observational patterns and trending discussions rather than using a specific tool or software.

Regarding inclusion and exclusion criteria, we included commonly used phrases and terms related to World Alzheimer's Awareness and excluded terms that were not specific to the awareness context. More specifically, to ensure that our search captured posts related to World Alzheimer's Month, our search phrases included the following: "World Alzheimers Month" OR "WorldAlzMonth" OR "World Alzheimer's Month" OR "World Alzheimers Awareness Month" OR "World Alzheimer's

XSL•FO
RenderX

Awareness Month'' OR "World Alzheimer's Day" OR "Alzheimer's Awareness" OR "Dementia Awareness."

We identified approximately 266,211 posts through our search. We removed 14,502 posts that were not written in English, and we excluded duplicate posts and posts that only included hashtags. We also excluded reply posts because they are often missing important background information, which left us with a total of 1,981 tweets.

Data Analysis

Thematic analysis was used to analyze the data by following the guidelines outlined by Braun and Clarke [28]. More specifically, our thematic analysis was conducted in the following stages: (1) a thorough review of the collected data; (2) code creation and data coding; (3) extraction of the initial themes; (4) theme refinement; and (5) naming of the final themes.

Our approach to analyzing the data involved creating a codebook development team (consisting of five coauthors) to thoroughly review the posts to draft and refine the codebook. Specifically, the creation of our codebook involved the familiarization with the data through a detailed review [29] by reading and rereading the posts to develop the codebook. The codebook consisted of nine codes: (1) stopping stigma; (2) sponsorship, related to fundraising activities; (3) simple or basic posts, lacking educational content; (4) promotional and marketing content; (5) lived experiences, sharing perspectives from family, friends, or individuals with dementia; (6) fostering stigma; (7) resources, providing support information; (8) informative posts, delivering educational messages; and (9) exploitation.

Once the codebook was completed, pilot coding exercises were held to familiarize our full team of coauthors with the codebook. The first exercise included a practice coding session where the full team collaboratively coded 50 posts. Following this, more advanced exercises were undertaken, including the independent coding of 75 posts that were compared to an answer key with extensive details about why the answer was chosen. Additionally, a collective coding exercise was held where the team collaboratively coded the posts and discussed discrepancies and questions to support intercoder reliability. Once the team felt comfortable with coding, the coauthors each received approximately 208 posts to independently code and then compare the findings with a coding partner, to reconcile any differences through pair discussions and group consensus when necessary. The data analysis concluded with a full team meeting to review and finalize the naming of each theme.

Rigor and Trustworthiness

In this study, trustworthiness and rigor were established by adhering to Lincoln and Guba's [30] guidelines to ensure credibility, dependability, and confirmability. For example, the coding process included coding by a multidisciplinary team (with expertise in nursing, geriatrics, psychology, computer science, and biology), iterative refinement of the codebook, and independent and group coding exercises. This meticulous approach ensured that various perspectives were considered, enhancing the reliability and validity of the data analysis [28,29].

To further achieve credibility, the study employed methods such as peer debriefing, facilitated through coding exercises [31], and triangulation, using the expertise of our research team [32]. For example, our approach included partnering less experienced coders with senior coders to provide coding mentorship to further enhance intercoder reliability [33]. Dependability and confirmability were supported by developing a thorough audit trail to outline our research process (fieldnotes, codebook, and thematic map) to ensure that the study's methodology is replicable [34].

Regular team meetings during the coding and theme development processes enhanced the study's rigor. These sessions were foundational in addressing potential discrepancies to enhance clarity in our interpretation of the data [32]. For example, post examples were added to the codebook to enhance the clarity of the codes. Finally, providing rich descriptions of the study's themes and supporting them with direct quotes helped to support the validity of our analysis and contribute to the study's trustworthiness.

Results

Analysis Results

The data screening process used to filter the posts is illustrated in Figure 1. Based on our thematic analysis, four main themes were identified including (1) dementia stereotypes; (2) discrimination and denied dignity; (3) devaluing the lives of people with dementia; and (4) countering dementia-related stigma.

XSL•FO
RenderX

Figure 1. Data screening process.

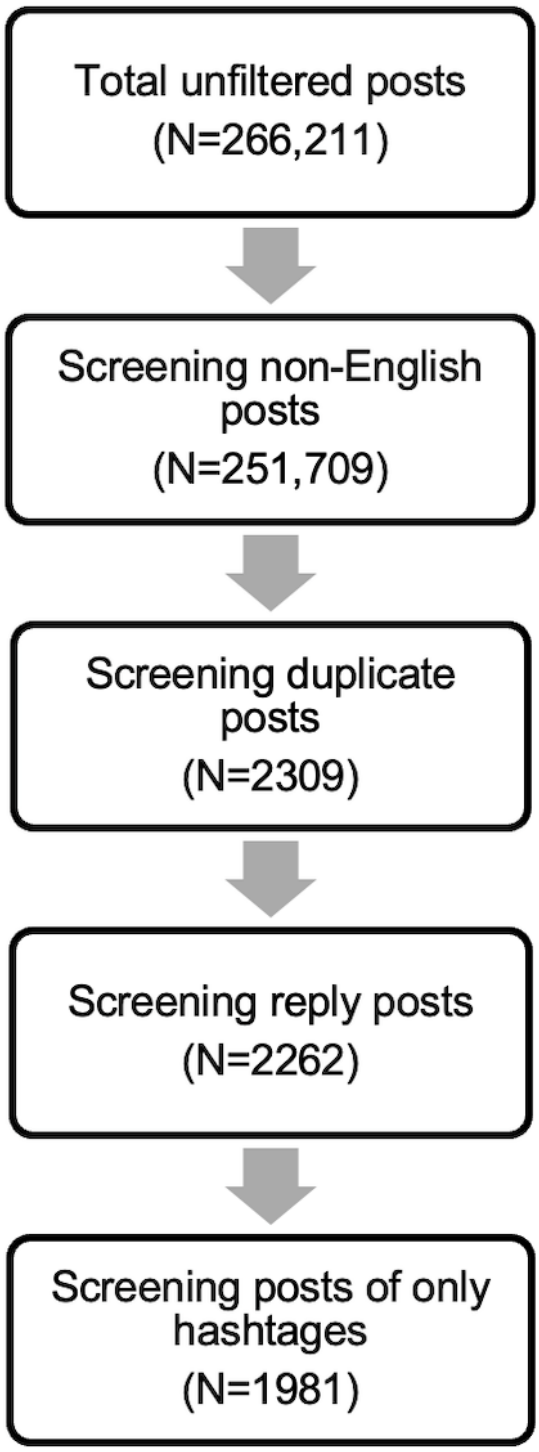

Theme 1: Dementia Stereotypes: "A Burden to Society"

Stereotype propagation of dementia was evident in many of the posts. Ironically, many of these posts may have been intended to be supportive but used negative language to fuel stereotypes. These posts often described people living with dementia as a burden to the economy and the health care system. These burden-laden posts are presented below:

More than 6 million Americans are living with Alzheimer's. This #WorldAlzheimersMonth, learn about the burden of Alzheimer's and dementia on individuals, caregivers, government, and the nation's health care system…

It's #WorldAlzheimersMonth. Alzheimers is not only a devastating disease, but a costly one. In 2020, total #healthcarecosts for Alzheimers treatment were estimated at $305 billion.

XSL•FO
RenderX

Posts also referred to people with dementia as "patients" rather than people. Many of these posts described people with dementia as "suffering" from the disease. However, this type of post contributes to negative stereotypes that depict all people living with dementia as suffering patients in the end-of-life stage, which further fuels the burden rhetoric.

> *On the eve of #WorldAlzheimersDay, which falls on September 21 every year, we look at how caregivers can support patients suffering from the condition.*
>
> *…Dementia can change patients' behavior and swing their moods, which builds up stress among others while caring for their dear parents or spouse…*

It is important to recognize that people living with dementia are people who contribute to society rather than a burden as portrayed in the stereotypes.

Theme 2: Discrimination and Denied Dignity: "Discrimination Exists in Public Spaces"

A predominant theme described discrimination and denied dignity as experienced by people living with dementia. More specifically, family members of people living with dementia shared stories of disparate treatment and public discrimination. The issues of discrimination and denied dignity are demonstrated in the following posts:

> *It's Dementia Awareness week and my Mum has Alzheimer's Dementia. She lives with us so our family sees first hand how hard and unfair this is but also how discrimination exists in public places. We are also making sure everyday counts…*
>
> *Our family member lost her battle to Vascular dementia. Let's together learn how best to support them. Stories about locking, tying them up, must stop. They need our understanding, support & love. We miss her already, very much #dementia #dementiaawareness #vasculardementia*

Posts also described how stigma reduces the dignity and quality of life of people living with dementia. For example, posts highlighted that stigma is often a reality faced by people living with dementia. Stigma was often described as one of the greatest challenges impacting people living with dementia. These posts demonstrate stigma and deny dignity:

> *A wonderful quote from our friend... Dementia does not rob someone of their dignity. It's our reaction to them that does. #caregiving #caregivers #caregiver #homecare #health #heathcare #respite*
>
> *#Dementia is one of the biggest challenges we face, and stigma is one of the main barriers for people w/ dementia to live their lives fully with dignity…*

Theme 3: Devaluing the Lives of People With Dementia: "Society Should Legalize Euthanasia"

Another theme was devaluing the lives of people living with dementia. These tweets were typically in the form of false assumptions, patronizing sentiments, and spreading detrimental information. For example, these posts often described people with dementia as incapable of living meaningful and high-quality lives. Some posts even argued that people living with dementia were better off ending their lives than continuing to live. The following posts highlight these devaluing sentiments:

> *September is Alzheimer's awareness month. I don't wish it on anyone, and I think society should legalise euthanasia for those who don't want to be stripped of their dignity and forced to sink into deep and upsetting confusion until their death.*
>
> *Yesterday she died. It was a painful year, she was suffering so I'm glad it ended…*

There was a strong focus on detailing the negative aspects of the disease. Posts described issues of people with dementia slowly slipping away and physical deterioration, rather than sharing narratives of living well with dementia.

> *Those who care for relatives with dementia often describe how a loved one slowly slips away, physically intact but unrecognisable bit by bit by bit…*
>
> *World Alzheimer's day Alzheimer's is a devastating disease. It was painful to watch my person deteriorate. #mentalhealthawareness #ptsd #psychosis #mentalhealthmatters #psychology #psychiatrist*

However, these posts can be harmful because they may contribute to self-stigma among people living with dementia as well as societal myths that all people with dementia are unable to live meaningful lives.

Theme 4: Countering Dementia-Related Stigma: "Break Down the Stigma"

During World Alzheimer's Awareness Month, many posts called for unity to dispel the stigma surrounding dementia. This theme often called for advocacy and actions to address stigmatization. For example, these posts encouraged readers to promote dementia awareness and education and "join the fight" to challenge the stigma of dementia.

> *#WorldAlzheimersDay is on September 21st—a time when people unite to raise awareness and attempt to break down the stigma of all forms of #dementia We can also involve children in age-appropriate caregiving…*
>
> *This #WorldAlzMonth, we aim to raise awareness of and challenge the stigma around Alzheimer's disease and other dementias.*
>
> *September is #WorldAlzheimersMonth! People unite from around the world to raise awareness and challenge the stigma of Alzheimer's disease & dementia.*
>
> *#WorldAlzheimersMonth is dedicated to raising awareness and challenging the stigma surrounding dementia - join the fight today!*

Other advocacy posts highlighted the need to respect people living with dementia. Some of these posts also made reference to recognizing the personhood and the human rights of people living with dementia.

XSL•FO
RenderX

…Respect our right to freedom, the right to have a say, the right to make a choice in matters related to us. #NothingAboutUsWithoutUs #KnowDementia #TogetherWeCanDoMore...

World Alzheimer's Day - Dementia Experts Call for Care to be Recognised as a Human Right…

Dementia awareness includes understanding that changes in cognitive function or processing data does not mean a person is less intelligent, or less sensitive, or less human. #Respect still matters.

Discussion

Principal Findings and Comparison With Previous Works

Our study analyzed dementia-related stigma on X during World Alzheimer's Month to identify opportunities for intervention to address the stigma of dementia. Examining the stigma of dementia is critical to understanding how stigma manifests on social media, especially during global awareness campaigns. Based on our thematic analysis, four main themes were identified, ranging from dementia stereotypes to devaluing the lives of people with dementia.

Findings from our study suggest that negative stereotypes and false information about dementia continue to persist during the global awareness campaign. For example, our study identified harmful stereotypes that depicted all people with dementia as suffering and being a burden on society. Moreover, posts often homogenize people with dementia as being incapable of living meaningful lives. There was also a strong focus on emphasizing the end-of-life stage, with some posts even suggesting that people with dementia were better off dead. Our findings are consistent with COVID-19 literature on dementia-related stigma, which reported that social media fuels misinformation, assumptions, and discrimination towards people with dementia [6]. However, dementia stereotypes and false information are harmful because they can lead to devaluing the lives of people with dementia, health inequities, and discrimination within the health care system [7,11,12].

An emerging body of literature indicates that people with dementia often self-internalize stigma and are highly concerned about being a burden on society [35-37]. Literature suggests that this concern of being a burden, combined with fears of suffering and loss of autonomy, often fuels the requests for physician-assisted suicide [36-40]. This literature on physician-assisted suicide is especially troubling when compared to our study's findings that shed light on the dissemination of dementia-related stigma on social media in terms of associating dementia diagnosis with the end-of-life and loss of meaning and purpose of life. However, it is important to note that dementia impacts everyone differently, and not everyone's dementia journey will be the same [41]. Accordingly, more dementia education and awareness strategies are needed to combat false information, stereotypes, and dementia-related stigma on social media.

It is important to note that organizations, such as the Alzheimer's Association [42] and the Alzheimer Society of Canada [43], are working to address dementia-related stigma by providing accurate information. More specifically, the Alzheimer Society of Canada provides tips to reduce dementia-related stigma, ranging from using person-centered language to avoiding assumptions about people with dementia [44]. They also provide educational resources on dementia, including information on dementia screening, diagnosis, treatment, and living well with dementia [45].

Limitations

Although our study was conducted in a rigorous manner, this research is not without limitations. More specifically, X users' sociodemographics may not represent the general population. For example, people with lower socioeconomic status may not have access to computers or technology to access the X platform. Moreover, our study only included English-language posts, which reduces our ability to make any generalizations about dementia-related stigma across different cultures. Consequently, additional research is needed to examine the roles of sociodemographic status and culture on dementia-related stigma.

Another weakness is that using data from X may contribute to biases related to the platform's algorithms. For example, recent literature suggests that X prioritizes and amplifies posts with higher engagement, which may increase the visibility of more sensational and low-credibility content [46]. Additionally, our study included data from all users, including nonverified accounts (possibly including chatbots and trolls), which may further contribute to bias and low-credibility discourse. However, Zhao and colleagues [47] note that there is a paucity of studies examining biases related to social media research. Accordingly, further research is needed to address the potential impact of biases related to low-credibility content and algorithms on social media research.

An additional limitation of our research may include the restricted timeframe. For example, our study focused on tweets from September 1, 2022, to September 30, 2022, to coincide with the World Alzheimer's Awareness Month. However, this limited timeframe may not fully capture discourse on dementia-related stigma given the extensive awareness raising efforts. Accordingly, further research is needed to examine dementia-related stigma outside of World Alzheimer's Month.

Conclusion

Our study examined dementia-related stigma on X during the World Alzheimer's Awareness Month. Our analysis of the data identified four main themes, ranging from dementia stereotypes to challenging stigma of dementia. Examining dementia-related stigma is essential to understanding how stigma manifests and persists on social media, especially during global awareness campaigns. By analyzing how stigma manifests on social media, our study sheds light on the dementia information and education needed to combat false beliefs, misinformation, and dementia-related stigma. Findings from our study have important implications for policymakers, health professionals, and community advocates working to design awareness campaigns to reduce dementia-related stigma on social media.

XSL•FO
RenderX

Acknowledgments

This work was supported by the Canada Research Chairs Program [Grant Number- 451097]. This work was also supported by Team 15 in the Canadian Consortium on Neurodegeneration in Aging (CCNA). CCNA is supported by a grant from the Canadian Institutes of Health Research with funding from several partners including the Saskatchewan Health Research Foundation, the Centre for Aging and Brain Health, and the Alzheimer Society of Canada. Funding was provided by the Internal Research Grant and the Undergraduate Research Apprenticeship Fund at Thompson Rivers University.

Data Availability

Supplementary materials (codebook) are available to other researchers for replication purposes; 2) supplementary data (X excel sheet) are available to other researchers for replication purposes by contacting the lead author; and 3) the study reported in the manuscript was not preregistered.

Authors' Contributions

JDB and RJS conceived the idea for the manuscript. JDB, AAJ, & RJS devised the X scraping approach, and AAJ scraped for tweets. JDB, RGS, AAJ, CB, and SAF devised the codebook. All of the coauthors coded the posts, contributed example tweets, and participated in the thematic analysis. JDB wrote the first draft of the manuscript with theme writing contributions from DF, CC, SV, KM, JM, AV, SAF, AC, ZR, FF, KN, and CB. All authors reviewed the final manuscript.

Conflicts of Interest

MKA received grants from Sanofi, GSK, Prizer, Merck, and Icosavac, all unrelated to the present study.

References

1. Alzheimer's Disease International. Dementia Statistics. 2025. URL: https://www.alzint.org/about/dementia-facts-figures/dementia-statistics/ [accessed 2025-02-15]
2. World Health Organization. Dementia: Fact Sheets. 2023. URL: https://www.who.int/news-room/fact-sheets/detail/dementia [accessed 2025-02-12]
3. Nichols E, Steinmetz JD, Vollset SE. Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the Global Burden of Disease Study 2019. Lancet Public Health 2022 Feb;7(2):e105-e125. [doi: 10.1016/S2468-2667(21)00249-8] [Medline: 34998485]
4. Philip G, Savundranayagam MY, Kothari A, Orange JB. Exploring stigmatizing perceptions of dementia among racialized groups living in the anglosphere: a scoping review. Aging and Health Research 2024 Mar;4(1):100170. [doi: 10.1016/j.ahr.2023.100170]
5. Vernooij-Dassen MJFJ, Moniz-Cook ED, Woods RT, et al. Factors affecting timely recognition and diagnosis of dementia across Europe: from awareness to stigma. Int J Geriatr Psychiatry 2005 Apr;20(4):377-386. [doi: 10.1002/gps.1302] [Medline: 15799080]
6. Bacsu JD, Fraser S, Chasteen AL, et al. Using Twitter to examine stigma against people with dementia during COVID-19: infodemiology study. JMIR Aging 2022 Mar 31;5(1):e35677. [doi: 10.2196/35677] [Medline: 35290197]
7. Bacsu JDR, O'Connell ME, Wighton MB. Improving the health equity and the human rights of Canadians with dementia through a social determinants approach: a call to action in the COVID-19 pandemic. Can J Public Health 2022 Apr;113(2):204-208. [doi: 10.17269/s41997-022-00618-8]
8. Rewerska-Juśko M, Rejdak K. Social stigma of people with dementia. J Alzheimers Dis 2020;78(4):1339-1343. [doi: 10.3233/JAD-201004] [Medline: 33185610]
9. Ménard A, O'Sullivan T, Mulvey M, Belanger C, Fraser S. Perceptions of hospital care for persons with dementia during the COVID-19 pandemic: a social media sentiment analysis. Gerontologist 2024 Jul 1;64(7):37943714. [doi: 10.1093/geront/gnad155] [Medline: 37943714]
10. Werner P, Giveon SM. Discriminatory behavior of family physicians toward a person with Alzheimer's disease. Int Psychogeriatr 2008 Aug;20(4):824-839. [doi: 10.1017/S1041610208007060] [Medline: 18341751]
11. Bacsu J, Mateen FJ, Johnson S, Viger MD, Hackett P. Improving dementia care among family physicians: from stigma to evidence-informed knowledge. Can Geriatr J 2020 Dec;23(4):340-343. [doi: 10.5770/cgj.23.426] [Medline: 33282053]
12. Bacsu JDR, Spiteri RJ, Nanson K, et al. Understanding stigma of dementia during COVID-19: a scoping review. Front Psychiatry 2024 Mar 27;15:1261113. [doi: 10.3389/fpsyt.2024.1261113]
13. Oscar N, Fox PA, Croucher R, Wernick R, Keune J, Hooker K. Machine learning, sentiment analysis, and tweets: an examination of Alzheimer's disease stigma on Twitter. J Gerontol B Psychol Sci Soc Sci 2017 Sep 1;72(5):742-751. [doi: 10.1093/geronb/gbx014] [Medline: 28329835]

XSL•FO
RenderX

14. Creten S, Heynderickx P, Dieltjens S. The stigma toward dementia on Twitter: a sentiment analysis of Dutch language tweets. J Health Commun 2022 Oct 3;27(10):697-705. [doi: 10.1080/10810730.2022.2149904] [Medline: 36519829]
15. Alvarez-Mon MA, Llavero-Valero M, Sánchez-Bayona R, et al. Areas of interest and stigmatic attitudes of the general public in five relevant medical conditions: thematic and quantitative analysis using Twitter. J Med Internet Res 2019 May 28;21(5):e14110. [doi: 10.2196/14110] [Medline: 31140438]
16. Bartmess M, Talbot C, O'Dwyer ST, Lopez RP, Rose KM, Anderson JG. Using Twitter to understand perspectives and experiences of dementia and caregiving at the beginning of the COVID-19 pandemic. Dementia (London) 2022 Jul;21(5):1734-1752. [doi: 10.1177/14713012221096982] [Medline: 35549466]
17. Cheng TYM, Liu L, Woo BK. Analyzing Twitter as a platform for Alzheimer-related dementia awareness: thematic analyses of tweets. JMIR Aging 2018 Dec 10;1(2):e11542. [doi: 10.2196/11542] [Medline: 31518232]
18. Alzheimer's disease international. World Alzheimer's Month. 2022. URL: www.alzint.org/get-involved/world-alzheimers-month/ [accessed 2025-05-12]
19. Bacsu JD, Cammer A, Ahmadi S, et al. Examining the Twitter discourse on dementia during Alzheimer's awareness month in Canada: infodemiology study. JMIR Form Res 2022 Oct 26;6(10):e40049. [doi: 10.2196/40049] [Medline: 36287605]
20. Bacsu JD, Fraser SA, Jamali AA, et al. Navigating awareness and strategies to support dementia advocacy on social media during world Alzheimer's month: infodemiology study. JMIR Infodemiology 2024 Dec 27;4:e63464. [doi: 10.2196/63464] [Medline: 39729354]
21. Bacsu JDR, Andrew MK, Azizi M, et al. Using Twitter to understand COVID-19 vaccine-related ageism during the pandemic. Gerontologist 2024 Feb 1;64(2):gnad061. [doi: 10.1093/geront/gnad061] [Medline: 37267449]
22. Statista. Global social networks ranked by number of users. 2025. URL: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/#:~:text=In%202025%2C%20social%20networking%20sites,traction%20in%20previously%20underserved%20markets [accessed 2025-03-27]
23. Draganidis A, Fernando AN, West ML, Sharp G. Social media delivered mental health campaigns and public service announcements: A systematic literature review of public engagement and help-seeking behaviours. Social Science & Medicine 2024 Oct;359:117231. [doi: 10.1016/j.socscimed.2024.117231]
24. Hong EL, Slay PD, Hampton M, et al. Illustrating best practices in optimizing social media strategy for a campaign targeting military mental health stigma. J technol behav sci 2021 Jun;6(2):427-435. [doi: 10.1007/s41347-020-00183-2]
25. Government of Canada. Does research using social media platforms require research ethics board review? guidance to support the implementation of the tri-council policy statement: ethical conduct for research involving humans – TCPS 2. 2022 URL: https://ethics.gc.ca/eng/documents/reb-cer_social-sociaux-en.pdf [accessed 2025-02-15]
26. JMIR. Do I need ethics approval for social media research?. 2022. URL: https://support.jmir.org/hc/en-us/articles/115001620728-Do-I-need-ethics-approval-for-social-media-research [accessed 2025-02-17]
27. Tweepy. URL: https://www.tweepy.org/ [accessed 2025-02-12]
28. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
29. Crabtree BF, Miller WL. A template approach to text analysis: developing and using codebooks. In: Crabtree BF, Miller WL, editors. Doing Qualitative Research: Sage Publications; 1992:93-109. [doi: 10.4135/9781446287958]
30. Lincoln YS, Guba EG. But is it rigorous? Trustworthiness and authenticity in naturalistic evaluation. New Directions for Program Evaluation 1986 Jun;1986(30):73-84. [doi: 10.1002/ev.1427]
31. Thomas DR. A general inductive approach for analyzing qualitative evaluation data. American Journal of Evaluation 2006 Jun;27(2):237-246. [doi: 10.1177/1098214005283748]
32. Raskind IG, Shelton RC, Comeau DL, Cooper HLF, Griffith DM, Kegler MC. A review of qualitative data analysis practices in health education and health behavior research. Health Educ Behav 2019 Feb;46(1):32-39. [doi: 10.1177/1090198118795019] [Medline: 30227078]
33. O'Connor C, Joffe H. Intercoder reliability in qualitative research: debates and practical guidelines. Int J Qual Methods 2020 Jan 1;19. [doi: 10.1177/1609406919899220]
34. Tobin GA, Begley CM. Methodological rigour within a qualitative framework. J Adv Nurs 2004 Nov;48(4):388-396. [doi: 10.1111/j.1365-2648.2004.03207.x] [Medline: 15500533]
35. Alzheimer Society of Canada. Alzheimer Society: 2017 Awareness Survey Executive Summary. URL: https://coilink.org/20.500.12592/rrx9hw [accessed 2025-02-12]
36. Johnstone MJ. Alzheimer's Disease, Media Representations and the Politics of Euthanasia: Routledge; 2013. [doi: 10.4324/9781315566634]
37. Low LF, Purwaningrum F. Negative stereotypes, fear and social distance: a systematic review of depictions of dementia in popular culture in the context of stigma. BMC Geriatr 2020 Nov 17;20(1):477. [doi: 10.1186/s12877-020-01754-x] [Medline: 33203379]
38. Rajkumar RP. Physician-assisted suicide in dementia: paradoxes, pitfalls and the need for prudence. Front Sociol 2021;6:815233. [doi: 10.3389/fsoc.2021.815233] [Medline: 35004941]

XSL•FO
RenderX

39. Tomlinson E, Spector A, Nurock S, Stott J. Euthanasia and physician-assisted suicide in dementia: a qualitative study of the views of former dementia carers. Palliat Med 2015 Sep;29(8):720-726. [doi: 10.1177/0269216315582143] [Medline: 25881624]
40. Jakhar J, Ambreen S, Prasad S. Right to Life or Right to Die in Advanced Dementia: Physician-Assisted Dying. Front Psychiatry 2020;11:622446. [doi: 10.3389/fpsyt.2020.622446] [Medline: 33551882]
41. Alzheimer Society of Canada. What to expect as the persons dementia progresses. 2025. URL: https://alzheimer.ca/en/help-information/im-caring-person-living-dementia/what-expect-persons-dementia-progresses [accessed 2025-02-15]
42. Alzheimer's Association. 2025. URL: https://www.alz.org/ [accessed 2025-02-15]
43. Alzheimer Society of Canada. 2025. URL: https://alzheimer.ca/en [accessed 2025-02-16]
44. Alzheimer Society of Canada. 10 Easy ways to fight stigma against dementia. 2025. URL: https://alzheimer.ca/en/about-dementia/stigma-against-dementia/10-easy-ways-fight-stigma-against-dementia [accessed 2025-02-14]
45. Alzheimer Society of Canada. Living well with dementia. 2025. URL: https://alzheimer.ca/en/help-information/im-living-dementia/living-well-dementia [accessed 2025-02-17]
46. Corsi G. Evaluating Twitter's algorithmic amplification of low-credibility content: an observational study. EPJ Data Sci 2024;13(1):18. [doi: 10.1140/epjds/s13688-024-00456-3]
47. Zhao Y, Yin P, Li Y, et al. Data and model biases in social media analyses: a case study of COVID-19 tweets. AMIA Annu Symp Proc 2021;2021:1264-1273. [Medline: 35308985]

Abbreviations

API: application programming interface

Edited by A Mavragani; submitted 17.02.25; peer-reviewed by E Dal, J Ogunsakin, M Gasmi, T Adebambo; revised version received 17.04.25; accepted 24.04.25; published 02.06.25.

Please cite as:
Bacsu JD, Mah JC, Jamali AA, Conanan C, Lautrup S, Berger C, Fiske D, Fraser S, Virani A, Fehr F, Chasteen AL, Rahemi Z, Vellani S, Andrew MK, Cammer A, McGilton KS, Gowda-Sookochoff R, Nanson K, Grewal KS, Spiteri RJ
Stigma of Dementia on Social Media During World Alzheimer's Awareness Month: Thematic Analysis of Posts
JMIR Form Res 2025;9:e72775
URL: https://formative.jmir.org/2025/1/e72775
doi:10.2196/72775

© Juanita-Dawne Bacsu, Jasmine Cassy Mah, Ali Akbar Jamali, Christine Conanan, Samantha Lautrup, Corinne Berger, Dylan Fiske, Sarah Fraser, Anila Virani, Florriann Fehr, Alison L Chasteen, Zahra Rahemi, Shirin Vellani, Melissa K Andrew, Allison Cammer, Katherine S McGilton, Rory Gowda-Sookochoff, Kate Nanson, Karl S Grewal, Raymond J Spiteri. Originally published in JMIR Formative Research (https://formative.jmir.org), 2.6.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Privacy Concerns Versus Personalized Health Content—Pregnant Individuals' Willingness to Share Personal Health Information on Social Media: Survey Study

Haijing Hao[1], PhD; Yang W Lee[2], PhD; Marianne Sharko[3], MD; Qilu Li[3], MS; Yiye Zhang[3], PhD

[1]Bentley University, Boston, MA, United States

[2]Northeastern University, Boston, MA, United States

[3]Weill Cornell Medicine, 575 Lexington Ave, New York, NY, United States

Corresponding Author:
Yiye Zhang, PhD
Weill Cornell Medicine, 575 Lexington Ave, New York, NY, United States

Abstract

Background: Often lacking immediate access to care providers, pregnant individuals frequently turn to web-based sources for information to address their evolving physical and mental health needs. Social media has gained increasing prominence as a source of news and information despite privacy concerns and unique risks posed to the pregnant population.

Objectives: This study investigated the extent to which patients may be willing to disclose personal health information to social media companies in exchange for more personalized health content.

Methods: We designed and deployed an electronic survey to pregnant individuals worldwide electronically in 2023. We used the classical Internet Users' Information Privacy Concerns (IUIPC) model to examine how privacy concerns modulate pregnant individuals' behaviors and beliefs regarding risk and trust when using social media for health purposes. Results were analyzed using partial least squares structural equation modeling.

Results: Among 317 respondents who initiated the survey, 84% (265/317) of the respondents remained in the study, providing complete responses. Among them, 54.7% (145/265) indicated willingness to provide their personalized health information for receiving personalized health content via social media, while 26% (69/265) were uncertain and 19.3% (51/265) were opposed. Our estimated IUIPC model results are statistically significant and qualitatively align with the classic IUIPC model for the general population, which was previously found in an e-commerce context. The structural model revealed that privacy concerns (IUIPC) negatively affected trusting beliefs ($\beta=-0.408$; $P<.001$) and positively influenced risk beliefs ($\beta=0.442$; $P<.001$). Trusting beliefs negatively impacted risk beliefs ($\beta=-o.362$; $P<.001$) and positively affected the intention to disclose personal health information ($\beta=o.266$; $P<.001$). Risk beliefs negatively influenced the intention to disclose ($\beta=-0.281$; $P<.001$). The model explained 41.5% of the variance in the intention to disclose personal health information ($R^2=0.415$). In parallel with pregnant individuals' willingness to share, we find that they have heightened privacy concerns and their use of social media for information seeking is largely impacted by their trust in the platforms. This heightened concern significantly affects both their trusting beliefs, making them less inclined to trust social media companies, and their risk beliefs, leading them to perceive greater risks in sharing personal health information. However, within this population, an increase in trust toward social media companies leads to a more substantial decrease in perceived risks than what has been previously observed in the general population.

Conclusions: We find that more than half of the pregnant individuals are open to sharing their personal health information to receive personalized content about health via social media, although they have more privacy concerns than the general population. This study emphasizes the need for policy regarding the protection of health data on social media for the pregnant population and beyond.

(JMIR Form Res 2025;9:e60862) doi:10.2196/60862

KEYWORDS

privacy concerns; trust; pregnancy; health information seeking; pregnant women; maternal; maternity; childbearing; web-based information; health information; mental health; internet; social support; technology; mobile health; mHealth; digital health; health informatics; social media

XSL•FO
RenderX

Introduction

Background

Advancements in web-based technology have significantly enhanced opportunities for digital social support and web-based health information seeking [1,2]. The trend of searching for health information on the web has gained momentum [3-5], particularly since the COVID-19 pandemic where health care services were overburdened to address individual questions [6,7]. Instead, patients gathered on the web, using social media in particular, for guidance and resources to empower themselves, cope with stress, and strengthen positive social support. This health information seeking behavior is particularly prevalent among pregnant patients [8]. Pregnancy is a vulnerable period that exposes patients to heightened anxiety, depression, and stress, leading to adverse maternal, infant, and family outcomes, disproportionately affecting disadvantaged families [9,10]. The negative impact can be mitigated by interventions from health care providers [11,12]. However, disparities in access to health care, health literacy, socioeconomic status, and neighborhood characteristics strangle equitable access to clinical interventions [12-14]. As a result, many patients resort to web-based information seeking and social support to relieve stressors and resolve individual needs during pregnancy [15]. In fact, people of childbearing age are one of the most active users in the digital space in the United States [16,17]. An estimated three-quarters of the US pregnant population seek pregnancy and birth information on the web, and there is increasing IT acceptance for data sharing [18,19]. Social media is an overwhelmingly popular source for pregnancy information, with myriad of social media accounts and influencers distributing a wide range of information and offering venues for discussion [8,20]. For example, research showed that web-based support and interventions improved health outcomes among pregnant patients with gestational diabetes mellitus, and individualized health interventions resulted in lower odds of developing gestational diabetes mellitus [21].

Yet, privacy and risk concerns are important to consider when using digital health technologies for pregnancy [22,23]. Research has shown that individuals often possess both risk and privacy concerns when turning to social media for health information [24-28]. For example, pregnancy privacy concerns have heightened since the overturn of Dobbs v Jackson by the Supreme Court in the United States, which had protected an individual's right to abortion. This has resulted in variable US state laws criminalizing certain reproductive health care, amplifying privacy concerns for pregnant individuals seeking web-based information about abortions and storing reproductive health information on personal applications [29]. However, despite significant privacy concerns, given patients' overwhelming health needs and the lack of clinical resources to address them, social media remains as a popular source for seeking health information and digital social support.

Objectives

Given the increasing need for information and limited access to prompt responses from health care providers regarding pregnancy-related inquiries, there is a notable tension between privacy concerns and the demand for personalized health communication. Our study aims to fill a significant research gap by applying the Internet Users' Information Privacy Concerns (IUIPC) model to understand pregnant individuals' privacy concerns when seeking health information and social support on social media. While prior studies have used the IUIPC model to examine privacy issues across various digital technologies and populations, none have specifically focused on pregnant individuals—a group that is particularly active on the web and yet vulnerable due to heightened health needs and privacy risks. By investigating how privacy concerns, trusting beliefs, and risk perceptions influence pregnant individuals' willingness to disclose personal health information for personalized content, our research contributes to the literature by providing new insights into the information disclosure behaviors of this unique population on social media. This enhances the understanding of privacy dynamics in digital health contexts and informs strategies for health care providers, policy makers, and social media companies to better support pregnant individuals' needs.

Methods

Study Design

This is a survey study. We developed and deployed a survey on Prolific in March 2023. Prolific is a web-based platform that connects researchers with a pool of prescreened study participants globally. Earlier studies have shown that participants recruited from Prolific provide high-quality data regarding users' perceptions of software and digital platforms [30,31]. The survey questions were hosted on Qualtrics (Qualtrics), a survey software. The survey comprises 41 questions: 22 questions focus on the conceptual model, with 7-point Likert scales, and the remaining questions focus on demographic information, social media questions related to health, and 3 quality-check questions. The survey questions are listed in Table S1 in Multimedia Appendix 1.

Settings

Participants were recruited using Prolific's recruitment method based on inclusion criteria: (1) aged at least 18 years or older, and (2) currently or recently pregnant within 2 years prior to the time of the survey. We used Prolific's prescreening filters to identify eligible participants, including age, gender, and pregnancy status to align with our inclusion criteria. Eligible participants identified through the prescreening process received an invitation to participate in our study via Prolific's messaging system.

The invitation included a brief overview of the study's purpose, estimated time commitment, and the compensation offered. Participants were excluded if their responses in the survey were incomplete or of poor quality, as determined by the quality-check questions. The first quality check is based on age, using the following questions: Q29_1 "What is your age?—Age" and Q14_1 "What is your birth year?—Year," which were placed in 2 different sections of the survey. If the difference between the answers to these 2 questions is more than 2 years, we excluded that participant. The second quality check is based on the Collection construct from the IUIPC model, using the

XSL•FO
RenderX

following questions: Q18_1 “It usually bothers me when health-conscious social media companies ask me for personal health information.” and Q18_5 “It usually doesn’t bother me when health-conscious social media companies ask me for personal health information.” These 2 opposite questions were placed in the survey, and if participants provide similar responses to both (eg, strongly disagree or strongly agree for both), they were excluded. However, if both answers are neutral (neither agree nor disagree), the participants were retained. The third quality check is based on the Risk-Belief construct from the IUIPC model, using the following questions: Q20_1 “In general, it would be risky to give my health information to health-conscious social media companies.” and Q20_5 “I would feel safe giving my health information to health-conscious social media companies.” The same strategy as the second quality check applies to this pair of questions.

Ethical Considerations

Our study was approved as exempt under institutional review board category II by Bentley University (no. 2023011). Upon accessing the survey link, participants were presented with an informed consent form detailing the study’s objectives, procedures, potential risks, benefits, and the voluntary nature of participation. Participants had to electronically agree to the consent form before proceeding to the survey questions. The study data are fully anonymous. Each participant received a US $1 reward for answering the questions per the site’s rules. Compensation was managed through Prolific’s system to ensure timely and secure payment.

Variables and Data Sources

The survey draws on a conceptual model, the IUIPC model by Malhotra et al [32]. IUIPC model was designed to evaluate whether concerns about information privacy sway an individual’s intent to adopt digital technology in the e-commerce context [32]. Since then, numerous studies have delved into various populations’ privacy concerns across various digital technologies based on the IUIPC model [26,33-38]. However, limited knowledge exists on the perspective of pregnant individuals’ privacy concerns when seeking health information on social media. Thus, we aim to uncover how information privacy concerns may influence pregnant individuals’ intentions to reveal personal health data to receive personalized pregnant-related health content through social media using IUIPC.

Figure 1 depicts the structure of the basic theoretical IUIPC model, consisting of 5 major constructs: collection, control, awareness, trusting beliefs, and risk beliefs [32]. Collection, control, and awareness are 3 first-order constructs, which are directly measured through survey items. IUIPC is a second-order construct, or a latent construct, which is estimated by the 3 first-order constructs. In this study, the collection construct refers to the extent of an individual’s privacy concerns about whether it is worth revealing their personal information to social media companies. The control construct measures the degree to which individuals feel they have control over their personal data and know how the data will be collected, used, and shared. The awareness construct gauges the level of a user’s privacy concern regarding their awareness of social media companies’ transparency and the proper handling of their information. Trusting beliefs refer to the degree to which people believe an entity is trustworthy in protecting their personal information, whereas risk beliefs refer to the expectation of a potential loss when sharing personal information with the entity. In this study, the personal information is personal health information and the entity is social media companies. IUIPC integrates trusting beliefs and risk beliefs to elucidate why a user may have an intention to adopt a new technology for certain benefits. In the context of this study, IUIPC reflects social media users’ concerns about social media companies’ collection of personal information, the users’ control over the collected information, and the users’ awareness of how the collected information is used, which in turn affects users’ intentions to reveal personal health information to receive customized health information through social media.

Figure 1. Internet Users’ Information Privacy Concerns (IUIPC) with 5 major constructs: collection, control, awareness, trusting beliefs, and risk beliefs.

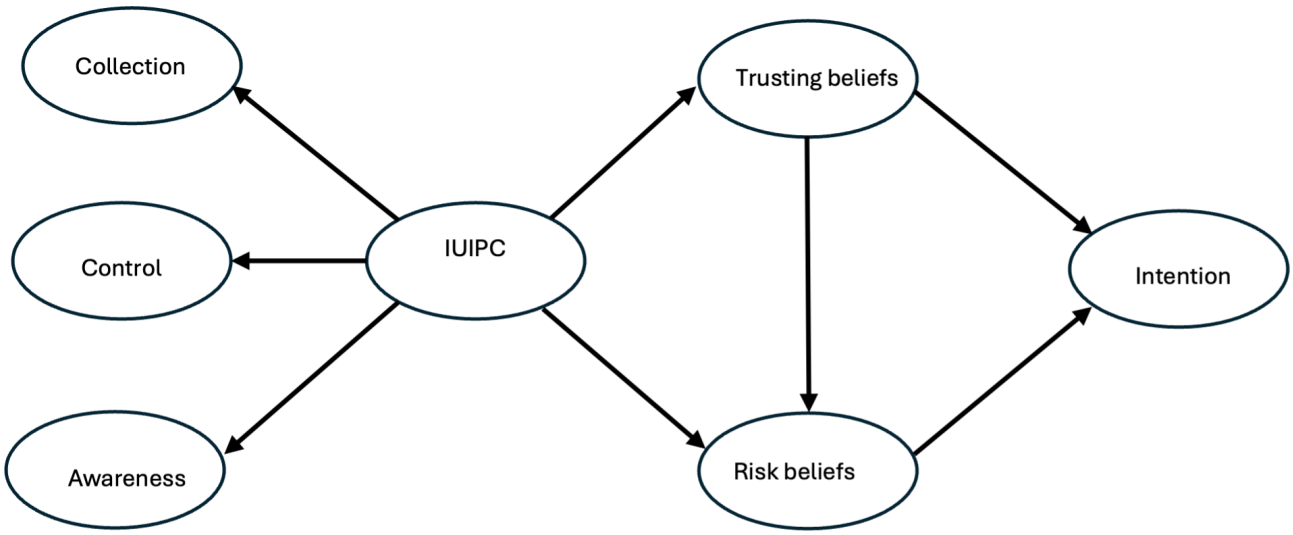

XSL•FO
RenderX

Research Hypotheses

Previous research has demonstrated that users with greater privacy concerns regarding social media tend to have lower trust toward these platforms [36,38]. Therefore, we hypothesize the following:

- Hypothesis 1: For pregnant individuals, IUIPC will have a negative effect on their trusting beliefs.

There is often an inverse relationship between trusting beliefs and risk beliefs because trust implies a firm belief in the reliability of something, while risk indicates the possibility of negative outcomes. When users exhibit higher privacy concerns toward social media, it implies that they perceive a greater risk associated with its use [36,38]. Thus, we propose the following 2 hypotheses:

- Hypothesis 2: For pregnant individuals, IUIPC will have a positive effect on their risk beliefs.
- Hypothesis 3: For pregnant individuals, their trusting beliefs will have a negative effect on their risk beliefs.

Research has shown that trust in a specific technology positively influences individuals' intention to explore that technology [39]. Pregnant individuals will share their personal information in exchange for health information through social media if they trust the companies and find the specific health content valuable enough to justify the privacy trade-off. Previous studies have also shown that trust plays a crucial role in adopting social media [33,34,40]. Therefore, we further hypothesize the following:

- Hypothesis 4: For pregnant individuals, trusting beliefs have a positive effect on their intention to reveal personal health information to social media companies to receive personalized health content through social media.

A previous study on risk beliefs concerning social media suggested that a perception of risk by users made them less likely to adopt social media [38]. If pregnant individuals perceive that sharing personal health information with social media companies for health purposes is risky, then this perception will result in a lower intention to use. Thus, we hypothesize the following:

- Hypothesis 5: For pregnant individuals, risk beliefs have a negative effect on their intention to reveal personal health information to social media companies in exchange for personalized health content through social media.

Measurement Model and Structural Equation Model Evaluation

We used partial least squares structural equation modeling to estimate the IUIPC model and test our hypotheses. Partial least squares structural equation modeling is a variance-based structural equation modeling (SEM) technique suitable for exploratory research and complex models with latent constructs, particularly when the sample size is small and data that may not meet the strict assumptions of covariance-based SEM [41]. We estimated the IUIPC model and assessed the measurement model results according to the following 4 criteria: unidimensionality, item reliability, construct reliability, and convergent validity [39,40]. We used Stata SE 17, and the Stata add-on package, plssem (StataCorp) [42] for the analysis. We reported the *P* values associated with each path coefficient in the structural model to indicate the statistical significance of the hypothesized relationships with significance at the .05 level ($P<.05$). CIs were determined using bootstrapping with 5000 resamples. Dimensionality is the number of constructs that a group of items or a group of survey questions reflects in a formative measurement model. We conducted principal component factor analysis to examine eigenvalues and kept only the constructs whose eigenvalues are >1. Item reliability refers to the extent to which an observed item can reliably measure its corresponding construct. We assume reliability when the standardized loading of an item on their constructs is >0.7. Construct reliability refers to the degree to which the items of a construct consistently and accurately measure that construct. Two commonly used metrics to evaluate construct reliability are Cronbach α and Dillon-Goldstein ρ, and their generally acceptable values are >0.7. Convergent validity evaluates the extent to which a set of items designed to measure the same construct are strongly related to one another. Average variance extracted (AVE) measures the average amount of variance in the items explained by the construct. AVE values should be >0.5, meaning that the construct explains 50% or more of the variance in its indicators. We used the R^2 value and the goodness of fit to evaluate the predictive power of IUIPC, which is a structural model. The acceptable threshold for goodness of fit is 0.36 [43]. Finally, variance inflation factor values are used to check for multicollinearity, with a common threshold of 2.5.

Results

Descriptive Statistics

The data collection lasted for 2 days until we reached the goal of 300 responses. A total of 317 respondents answered the survey. After removing incomplete surveys and responses that did not meet our quality criteria based on the 3 quality check questions, 83.6% (265/317) of the respondents remained in the study. Table 1 shows the sociodemographic information of the respondents. The largest racial group is White (164/265, 61.9%), and the second largest racial group is African American (82/265, 30.9%). Nearly 75% (198/265) of the respondents have a bachelor's degree or above, which matches our expectations and is consistent with previous studies showing that participants on Prolific tend to have higher levels of education than the general population [44].

Most participants (256/265, 96.7%) responded that they receive their pregnancy health information from their doctors, and 69.4% (184/265) of the participants mentioned the web or social media as pregnancy health information sources. Regarding specific social media platforms, 74.7% (198/265) used Instagram, 69.4% (184/265) used Facebook, 55.5% (147/265) used TikTok, and 30.1%(81/265) used Twitter (Now X). A small percentage (9/265, 3.4%) responded that they used PatientsLikeMe, a health-focused social media platform. An additional 4.2% (11/265) named other social media platforms including Reddit, LinkedIn, YouTube, Pinterest, and WhatsApp. Finally, 0.74% (2/265) of participants did not use social media.

XSL•FO
RenderX

Overall, 99.2% (263/265) of respondents used 1 or more social media platforms. Among all 265 participants, 54.7% (145/265) answered that they were comfortable sharing their personal health information in order to receive customized health content on social media, whereas 26% (69/265) were uncertain and 19.3% (51/265) responded that they were not comfortable sharing.

Table . Descriptive statistics of the 265 participants who were enrolled in the survey study.

Participant characteristics		Values, n (%)
Race		
	Asian	8 (3)
	African American	82 (30.9)
	Multirace	5 (1.9)
	White	164 (61.9)
	Other	6 (2.3)
Ethnicity		
	Hispanic	14 (5.3)
	Non-Hispanic	228 (86)
	Prefer not to say	23 (8.7)
Marital status		
	Married	126 (47.6)
	In a relationship	111 (41.9)
	Single[a]	28 (10.6)
Education		
	High school	23 (8.7)
	Some college	44 (16.6)
	Bachelor's degree	141 (53.2)
	Master's degree	47 (17.7)
	Doctoral degree	2 (0.8)
	Professional degree	8 (3)

[a]Single includes separated, divorced, widowed, and single.

Outcome Data

Of those who were willing to share their personal health information to receive personalized health content, all of them used 1 or more popular social media platforms such as Facebook, Instagram, TikTok, or Twitter. Of those who responded that they were not comfortable sharing, 47% (24/51) cited privacy concerns, and equally, 47% (24/51) cited a lack of trust in social media companies and influencers. One respondent commented on the political environment as a concern for sharing. There were no significant differences in ethnicity, race, or education level (bachelor's degree or more vs up to high school) in the responses. Among users of the top four social media platforms, 56.1% (111/198) of Instagram users, 56.5% (104/184) of Facebook users, 64% (94/147) of TikTok users, and 65.4% (53/81) of Twitter (now X) users were comfortable sharing personal health information in exchange for more customized social media content. There was a significant difference in willingness between people who use and do not use TikTok, and between those who do and do not use Twitter (now X).

Measurement Model Result

Overall, the measurement model results demonstrate that our constructs exhibit strong reliability and validity, confirming their appropriateness for use in the structural analysis. The collection, control, and awareness constructs all had eigenvalues >1. Thus, they were kept in the model as unidimensional constructs. The loadings and constructs' AVE values are shown in Tables S2 and S3 in Multimedia Appendix 1. As Table S2 in Multimedia Appendix 1 shows, most of the bolded diagonal block loadings are >0.7 for the corresponding constructs, which demonstrate the items' strong reliability. Although a few items on the second-order construct, IUIPC, are slightly <0.7, we still accept this model result because these loadings are not trivial, ranging from 0.44 to 0.58. In addition, they are still statistically significant. Furthermore, as Figure 2 shows, among the estimated coefficients for the structural model, all the path coefficients are statistically significant and are in the directions we expected, which confirms that those items are meaningful for our model. In Table S2 in Multimedia Appendix 1, we can see that the 2 commonly used metrics for evaluating construct reliability, Cronbach α and Dillon-Goldstein ρ, are >0.7, above

XSL•FO
RenderX

the acceptable threshold value, indicating good construct reliability of our model. As Table S3 in Multimedia Appendix 1 shows, all constructs' AVE values are >0.5 except for the IUIPC, the second order construct. This AVE value is <0.5 because of the lower loadings of the control and awareness items on IUIPC, as shown in Table S2 in Multimedia Appendix 1. Finally, the model's composite reliability scores, as measured by Dillon-Goldstein ρ of all constructs (Table S2 in Multimedia Appendix 1), are >0.7, indicating good internal consistency of the IUIPC model.

Figure 2. Internet Users' Information Privacy Concerns (IUIPC) model with estimated path coefficients all statistically significant and are in the directions we expected, confirming that those items are meaningful for our model. The asterisks denote statistical significance based on *P* values. **P*<.05, ***P*<.01. ****P*<.001

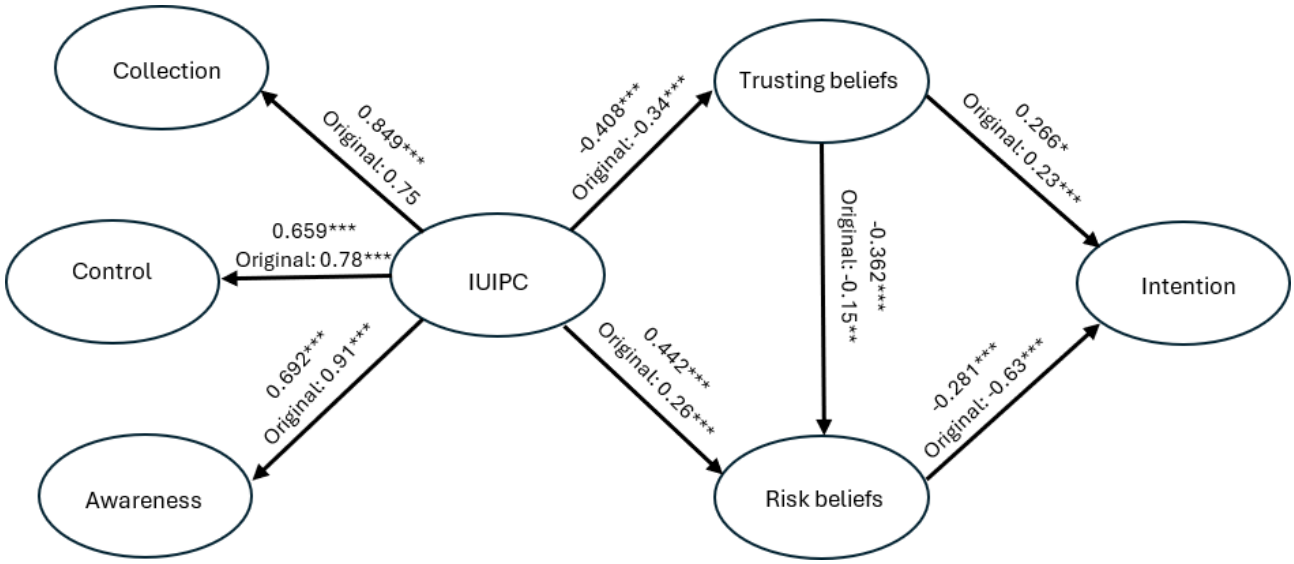

Structural Model Result

Our estimated IUIPC model results and the 5 path coefficients of the structural model support all 5 hypotheses that we constructed. The IUIPC model's average R^2 value is 0.4146, indicating a moderate effect. The absolute goodness of fit is 0.4923, which is >0.36, the acceptable threshold [43]. All variance inflation factor values were <2.5, the commonly suggested threshold value (Table S4 in Multimedia Appendix 1), indicating little concern for multicollinearity in our structural model. As Figure 2 shows, all path coefficients are statistically significant and in the directions as we expected, and most of them are qualitatively consistent with the original IUIPC model estimates from the study by Malhotra et al [32].

Discussion

Interpretation

Our survey of 265 pregnant individuals globally finds that more than half of pregnant individuals are open to sharing their personal health information to receive personalized content about health via social media, although they have more privacy concerns than the general population. In the original study by Malhotra et al [32], the coefficient between the collection construct and the IUIPC was not statistically significant (β=0.75). However, we observed significance in this coefficient in our model (β=0.849; *P*<.001). This finding was supported by another study by Zeng et al [38], which examined the privacy concerns on social media platforms among the general population and found this coefficient to be at a similar significant level to ours (β=0.36; *P*<.001), albeit smaller. This may indicate that pregnant individuals are more careful about sharing their health-related information with social media companies, reflecting their privacy concerns. In this study, the loadings for both the control and awareness constructs are slightly lower than those in the original study by Malhotra et al [32]. Both our study and the study by Zeng et al [38] have loadings of control and awareness that are lower than results from the classic model by Maholtra et al [32]. This potentially indicates that today, social media users' information control and social media users' awareness of the social media company's practices have a strong impact on their IUIPC concerns, although their quantitative effects are lower than those previously found in the e-commerce setting.

Our model results suggest that the impact of trusting beliefs on pregnant individuals' intention to adopt social media for health purposes is greater than that found in the study by Malhotra et al [32]. The IUIPC in our model negatively affects trusting beliefs, evidenced by a path coefficient of β=−0.408 (*P*<.001), which is consistent with but stronger than the original model's path coefficient of β=−0.34 (*P*<.001). This difference may suggest that pregnant individuals' privacy concerns regarding trust are more pronounced than those found in the classical model [32]. The IUIPC construct in our model positively influences risk beliefs, with a coefficient of β=0.442 (*P*<.001), which is larger than the coefficient in the original model (β=0.26; *P*<.001). This suggests that, in addition to trust, pregnant individuals' privacy concerns also have a stronger influence on their risk beliefs than those of the general population in the e-commerce context. In our model, trusting beliefs negatively impact risk beliefs, with a coefficient of β=−0.362 (*P*<.001), which is stronger than the original model's coefficient of β=−0.15 (*P*<.001). This suggests that if a pregnant individual's trust in a social media company increases, their

XSL•FO
RenderX

perceived risks will decrease more than in the general population in the e-commerce context as previously found [32]. Trusting beliefs positively influence pregnant individuals' intention to reveal health information to receive customized health content through social media with a coefficient of β=0.266 (*P*<.001), which is larger than the original model (β=0.23; *P*<.001). Conversely, risk beliefs have a negative effect on this intention with a coefficient of β=−0.281 (*P*<.001), which is smaller in magnitude than the original model (β=−0.63; *P*<.001). This is consistent with a previous study, which showed that trust is positively associated with adoption intention while privacy concerns are negatively associated with the adoption intention of technology in health care [45].

Our study adds to our current understanding that pregnant individuals are increasingly turning to social media for health information and support, despite significant privacy and risk concerns. For health care providers, this underscores the importance of integrating digital tools into their practice to meet patients where they are seeking information. By establishing a trustworthy web-based presence and providing accurate, evidence-based health information through social media, providers can better address the informational needs of pregnant patients. Educating patients about the potential risks associated with sharing personal health information on the web is also crucial. This includes guidance on maintaining privacy, recognizing credible sources, and understanding how personal data might be used by third parties. Enhancing the patient-provider relationship through open communication about privacy concerns can reduce the likelihood of patients seeking information from less reliable sources. For policy makers, our findings confirm a pressing need to safeguard personal health information on the web and in social media. Developing robust regulatory frameworks that govern how social media companies collect, use, and share personal health data is essential. Updating and enforcing privacy laws to encompass social media platforms can provide additional protections for sensitive health information. Policy makers can also educate individuals about digital privacy, helping pregnant individuals make informed decisions about sharing personal health information on the web.

Our results suggest that pregnant individuals are concerned about their personal health data being collected by social media companies, about losing control of their health information collected by social media companies, and about how their health information will be used. Privacy and confidentiality are paramount, as pregnant individuals may inadvertently disclose sensitive information that could be misused and harm them. Therefore, prior to collecting personal health information, social media companies with health-related content should explain clearly that the purpose of data collection is to create personalized health content for each user's benefit. Furthermore, social media companies should guarantee that patients can easily access any information they have provided to ensure users' rights to control their data. Users should also be informed, in clear language, of the companies' privacy policies, including if and how their health data will be used. This transparency and control over their data may help build trust between pregnant individuals and the social media companies.

Limitations

This study has limitations. First, since we distributed the survey to participants globally, various cultural differences, religions, and government policies may influence the perceptions of the participants. Second, our sample size was small and limited to web-based survey participants, which may have introduced a built-in bias toward the digital space. Our sample, recruited from Prolific, likely comprises individuals who are more digitally savvy compared with the general population, as reflected in their higher education levels and potentially greater web-based literacy, which may introduce selection bias and affect the generalizability of our findings. Future studies will recruit larger samples while specifically target participants with low web-based health literacy to ensure adequate representation of their perspectives and investigate whether there is any difference in the types of personal health data that participants are willing to share. Relatedly, while we did ask each participant which social media platforms they use, our small sample size did not allow for extensive platform-specific comparisons. In addition, our survey did not specify any social media or web-based health forum specifically but referred to general social media for health. Considering that privacy policies of various health-related social media platforms differ, users' trusting beliefs and risk beliefs may differ across social media platforms. Furthermore, we acknowledge while our SEM allows for the testing of directional hypotheses, causality cannot be definitively established due to the cross-sectional nature of the data. Future studies should investigate additional factors such as sociodemographic moderating effects that may influence pregnant individuals' intentions to use social media or web-based health platforms for health information seeking and communication. Moreover, future studies may investigate the benefit or cost of health information social media despite privacy concerns and consider extending our study by exploring how other vulnerable populations such as patients with chronic illnesses or mental health conditions navigate privacy and trust issues in the digital space. Finally, future research can consider designing and developing customized constructs to add to the classical IUIPC model. This would allow for examining how these tailored constructs impact pregnant patients' privacy concerns on social media.

Conclusions

In this study, we examined pregnant individuals' willingness to disclose personal health information in exchange for more personalized health content on social media. Based on a classical model of IUIPC, we designed and deployed a survey for pregnant patients on the web. As found in previous studies such as those by Malhotra et al [32] and Zeng et al [38], our findings align with this consensus that higher privacy concerns negatively impact trust and positively influence risk perceptions in digital contexts. We found that pregnant individuals exhibit more pronounced social privacy concerns than those identified in the classic IUIPC model. This heightened concern significantly affects both their trusting beliefs, making them less inclined to trust social media companies, and their risk beliefs, leading them to perceive greater risks in sharing personal health information. However, within this population, an increase in trust toward social media companies leads to a more substantial

XSL•FO
RenderX

decrease in perceived risks than what has been previously observed in the general population. Thus, social media companies delivering personalized-related health content should prioritize building trust with their users. This involves clearly informing patients about what health data are collected, providing them with easy access to their own health data, keeping them informed about how their health data are used, and including information about privacy protection policies. Given pregnant individuals' needs for web-based health information seeking and the level of openness toward information sharing found from this study, we argue that more regulations are needed to mitigate users' privacy and risk concerns and protect their data against exploitation on social media.

Acknowledgments

We sincerely thank Maya Daiter, Yiliang Zhou, and Haiyu Zheng for deploying the surveys used in the study. We sincerely thank Dr Sergio Venturini for his patience and thorough explanation of the usage of the plssem package, developed by Dr Sergio Venturini and Dr Mehmet Mehmetoglu, for our model estimation.

Data Availability

The datasets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

HH initiated the study, designed the survey, analyzed the data, conducted model estimation and analyses, wrote and revised the manuscript. YWL provided advice on the paper's structure and assisted with revising the model and writing the paper. MS contributed a clinical perspective, reviewed and revised the manuscript, and provided references for related health care studies. QL assisted in data collection, data cleaning, literature review, and paper format polishing. YZ initiated the study, provided funding for data collection, designed and deployed the survey, analyzed data and model, and wrote and revised the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary tables.
[DOCX File, 26 KB - formative_v9i1e60862_app1.docx]

References

1. Hao H, Levkoff S, Wang W, et al. Studying online support for caregivers of patients with Alzheimer's disease in China: a text-mining approach to online forum in China. Int J Healthc Inf Syst Inform 2020;15:1-17. [doi: 10.4018/IJHISI.2020100101]
2. Hao H, Zhang K. The voice of Chinese health consumers: a text mining approach to web-based physician reviews. J Med Internet Res 2016 May 10;18(5):e108. [doi: 10.2196/jmir.4430] [Medline: 27165558]
3. Jia X, Pang Y, Liu LS. Online health information seeking behavior: a systematic review. Healthcare (Basel) 2021;9(12):1740. [doi: 10.3390/healthcare9121740]
4. Zhao Y, Zhang J. Consumer health information seeking in social media: a literature review. Health Info Libr J 2017 Dec;34(4):268-283. [doi: 10.1111/hir.12192] [Medline: 29045011]
5. Chen J, Wang Y. Social media use for health purposes: systematic review. J Med Internet Res 2021 May 12;23(5):e17917. [doi: 10.2196/17917] [Medline: 33978589]
6. Chatwin J, Butler D, Jones J, James L, Choucri L, McCarthy R. Experiences of pregnant mothers using a social media based antenatal support service during the COVID-19 lockdown in the UK: findings from a user survey. BMJ Open 2021 Jan 17;11(1):e040649. [doi: 10.1136/bmjopen-2020-040649] [Medline: 33455927]
7. Neely S, Eldredge C, Sanders R. Health information seeking behaviors on social media during the COVID-19 pandemic among American social networking site users: survey study. J Med Internet Res 2021 Jun 11;23(6):e29802. [doi: 10.2196/29802] [Medline: 34043526]
8. Lee JY, Lee E. What topics are women interested in during pregnancy: exploring the role of social media as informational and emotional support. BMC Pregnancy Childbirth 2022 Jun 26;22(1):517. [doi: 10.1186/s12884-022-04842-5] [Medline: 35754031]
9. Dunkel Schetter C, Tanner L. Anxiety, depression and stress in pregnancy: implications for mothers, children, research, and practice. Curr Opin Psychiatry 2012 Mar;25(2):141-148. [doi: 10.1097/YCO.0b013e3283503680] [Medline: 22262028]
10. Walsh K, McCormack CA, Webster R, et al. Maternal prenatal stress phenotypes associate with fetal neurodevelopment and birth outcomes. Proc Natl Acad Sci U S A 2019 Nov 26;116(48):23996-24005. [doi: 10.1073/pnas.1905890116] [Medline: 31611411]

XSL•FO
RenderX

11. Peahl AF, Zahn CM, Turrentine M, et al. The Michigan plan for appropriate tailored health care in pregnancy prenatal care recommendations. Obstet Gynecol 2021;138:593. [doi: 10.1097/AOG.0000000000004531]
12. Panaite V, Bowersox NW, Zivin K, Ganoczy D, Kim HM, Pfeiffer PN. Individual and neighborhood characteristics as predictors of depression symptom response. Health Serv Res 2019 Jun;54(3):586-591. [doi: 10.1111/1475-6773.13127] [Medline: 30829408]
13. Zhang Y, Tayarani M, Wang S, et al. Identifying urban built environment factors in pregnancy care and maternal mental health outcomes. BMC Pregnancy Childbirth 2021 Dec;21(1):599. [doi: 10.1186/s12884-021-04056-1]
14. Adhikari K, Patten SB, Williamson T, et al. Neighbourhood socioeconomic status modifies the association between anxiety and depression during pregnancy and preterm birth: a community-based Canadian cohort study. BMJ Open 2020 Feb 10;10(2):e031035. [doi: 10.1136/bmjopen-2019-031035] [Medline: 32047008]
15. Bedaso A, Adams J, Peng W, Sibbritt D. The relationship between social support and mental health problems during pregnancy: a systematic review and meta-analysis. Reprod Health 2021 Jul 28;18(1):162. [doi: 10.1186/s12978-021-01209-5] [Medline: 34321040]
16. Harpel T. Pregnant women sharing pregnancy-related information on Facebook: web-based survey study. J Med Internet Res 2018 Mar 22;20(3):e115. [doi: 10.2196/jmir.7753] [Medline: 29567636]
17. Lupton D. The use and value of digital media for information about pregnancy and early motherhood: a focus group study. BMC Pregnancy Childbirth 2016 Jul 19;16(1):171. [doi: 10.1186/s12884-016-0971-3] [Medline: 27435182]
18. Sayakhot P, Carolan-Olah M. Internet use by pregnant women seeking pregnancy-related information: a systematic review. BMC Pregnancy Childbirth 2016 Mar 28;16:65. [doi: 10.1186/s12884-016-0856-5] [Medline: 27021727]
19. Rahdar S, Montazeri M, Mirzaee M, Ahmadian L. The relationship between e-health literacy and information technology acceptance, and the willingness to share personal and health information among pregnant women. Int J Med Inform 2023 Oct;178:105203. [doi: 10.1016/j.ijmedinf.2023.105203] [Medline: 37688834]
20. Smith M, Mitchell AS, Townsend ML, Herbert JS. The relationship between digital media use during pregnancy, maternal psychological wellbeing, and maternal-fetal attachment. PLoS One 2020;15(12):e0243898. [doi: 10.1371/journal.pone.0243898] [Medline: 33326465]
21. Cheng Z, Hao H, Tsofliou F, Katz MD, Zhang Y. Effects of online support and social media communities on gestational diabetes: a systematic review. Int J Med Inform 2023 Dec;180:105263. [doi: 10.1016/j.ijmedinf.2023.105263] [Medline: 37907014]
22. Wanni Arachchige Dona S, Angeles MR, Nguyen D, Cooper P, Sweet L, Peeters A. An assessment of ten popular pregnancy applications (apps) available for women in Australia. Int J Med Inform 2024 Apr;184:105354. [doi: 10.1016/j.ijmedinf.2024.105354] [Medline: 38325121]
23. Valencia SA, Barrientos Gómez JG, Gómez Ramirez MC, et al. Evaluation of a telehealth program for high-risk pregnancy in a health service provider institution. Int J Med Inform 2023 Nov;179:105234. [doi: 10.1016/j.ijmedinf.2023.105234] [Medline: 37776668]
24. Wijayanti RP, Handayani PW, Azzahro F. Intention to seek health information on social media in Indonesia. Procedia Comput Sci 2022;197:118-125. [doi: 10.1016/j.procs.2021.12.125]
25. Petersen C, Lehmann CU. Social media in health care: time for transparent privacy policies and consent for data use and disclosure. Appl Clin Inform 2018 Oct;9(4):856-859. [doi: 10.1055/s-0038-1676332] [Medline: 30485880]
26. Li Y, Wang X. Seeking health information on social media: a perspective of trust, self-determination, and social support. J Organ End User Comput 2018;30:1-22. [doi: 10.4018/JOEUC.2018010101]
27. Fan K, Jiang W, Li H, Yang Y. Lightweight RFID protocol for medical privacy protection in IoT. IEEE Trans Ind Inf 2018;14(4):1656-1665. [doi: 10.1109/TII.2018.2794996]
28. Grande D, Mitra N, Iyengar R, et al. Consumer willingness to share personal digital information for health-related uses. JAMA Netw Open 2022 Jan 4;5(1):e2144787. [doi: 10.1001/jamanetworkopen.2021.44787] [Medline: 35072717]
29. Clayton EW, Embí PJ, Malin BA. Dobbs and the future of health data privacy for patients and healthcare organizations. J Am Med Inform Assoc 2022 Dec 13;30(1):155-160. [doi: 10.1093/jamia/ocac155] [Medline: 36048014]
30. Douglas BD, Ewell PJ, Brauer M. Data quality in online human-subjects research: comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. PLoS One 2023;18(3):e0279720. [doi: 10.1371/journal.pone.0279720] [Medline: 36917576]
31. Peer E, Rothschild D, Gordon A, Evernden Z, Damer E. Data quality of platforms and panels for online behavioral research. Behav Res 2021;54(4):1643-1662. [doi: 10.3758/s13428-021-01694-3]
32. Malhotra NK, Kim SS, Agarwal J. Internet Users' Information Privacy Concerns (IUIPC): the construct, the scale, and a causal model. Inf Syst Res 2004 Dec;15(4):336-355. [doi: 10.1287/isre.1040.0032]
33. Adhikari K, Panda RK. Users' information privacy concerns and privacy protection behaviors in social networks. J Glob Mark 2018 Mar 15;31(2):96-110. [doi: 10.1080/08911762.2017.1412552]
34. Mohamed N, Ahmad IH. Information privacy concerns, antecedents and privacy measure use in social networking sites: evidence from Malaysia. Comput Human Behav 2012 Nov;28(6):2366-2375. [doi: 10.1016/j.chb.2012.07.008]
35. Harborth D, Pape S. JonDonym users' information privacy concerns. In: Janczewski LJ, Kutyłowski M, editors. ICT Systems Security and Privacy Protection: Springer International Publishing; 2018:170-184.

XSL•FO
RenderX

36. Harborth D, Pape S. How privacy concerns, trust and risk beliefs, and privacy literacy influence users' intentions to use privacy-enhancing technologies: the case of Tor. SIGMIS Database 2020 Jan 21;51(1):51-69. [doi: 10.1145/3380799.3380805]
37. Pape S, Ivan A, Harborth D, et al. Re-evaluating internet users' information privacy concerns: the case in Japan. AIS Trans Replic Res 2020. [doi: 10.17705/1atrr.00061]
38. Zeng M, Lin S, Armstrong D. Are all internet users' information privacy concerns (IUIPC) created equal? Trans Rep Res 2020;6. [doi: 10.17705/1ATRR.00046]
39. Mcknight DH, Carter M, Thatcher JB, et al. Trust in a specific technology: an investigation of its components and measures. ACM Trans Manage Inf Syst 2011;2:1-25. [doi: 10.1145/1985347.1985353]
40. Cain JA, Imre I. Everybody wants some: collection and control of personal information, privacy concerns, and social media use. New Media Soc 2022 Dec;24(12):2705-2724. [doi: 10.1177/14614448211000327]
41. Mehmetoglu M, Venturini S. Structural Equation Modelling with Partial Least Squares Using Stata and R: CRC Press; 2021.
42. Venturini S, Mehmetoglu M. plssem: a Stata package for structural equation modeling with partial least squares. J Stat Soft 2019;88(8). [doi: 10.18637/jss.v088.i08]
43. Wetzels MGM, Odekerken-Schröder GJ, van Oppen C. Using PLS path modeling for assessing hierarchical construct models: guidelines and empirical illustration. MIS Q 2009;33(1):177. [doi: 10.2307/20650284]
44. Samiotis IP, Qiu S, Lofi C, Yang J, Gadiraju U, Bozzon A. An analysis of music perception skills on crowdsourcing platforms. Front Artif Intell 2022;5:828733. [doi: 10.3389/frai.2022.828733] [Medline: 35774636]
45. Dhagarra D, Goswami M, Kumar G. Impact of trust and privacy concerns on technology acceptance in healthcare: an Indian perspective. Int J Med Inform 2020 Sep;141:104164. [doi: 10.1016/j.ijmedinf.2020.104164] [Medline: 32593847]

Abbreviations

AVE: average variance extracted
IUIPC: Internet Users' Information Privacy Concerns
SEM: structural equation modeling

Edited by A Mavragani; submitted 23.05.24; peer-reviewed by A Jain, L Ahmadian, RS Gomaa Mahmoud, U Sinha, Z Hou; revised version received 19.11.24; accepted 27.11.24; published 10.02.25.

Please cite as:
Hao H, Lee YW, Sharko M, Li Q, Zhang Y
Privacy Concerns Versus Personalized Health Content—Pregnant Individuals' Willingness to Share Personal Health Information on Social Media: Survey Study
JMIR Form Res 2025;9:e60862
URL: https://formative.jmir.org/2025/1/e60862
doi:10.2196/60862

© Haijing Hao, Yang W Lee, Marianne Sharko, Qilu Li, Yiye Zhang. Originally published in JMIR Formative Research (https://formative.jmir.org), 10.2.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Instagram Posts Promoting Colorectal Cancer Awareness: Content Analysis of Themes and Engagement During Colorectal Cancer Awareness Month

Aditi Srivastava[1]; Jim P Stimpson[2], PhD

[1]Frisco High School, Frisco, TX, United States

[2]Department of Health Economics, Systems, and Policy, Peter O'Donnell Jr. School of Public Health, The University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd, Dallas, TX, United States

Corresponding Author:
Jim P Stimpson, PhD
Department of Health Economics, Systems, and Policy, Peter O'Donnell Jr. School of Public Health, The University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd, Dallas, TX, United States

Abstract

Background: Colorectal cancer (CRC) is a leading cause of cancer-related deaths worldwide, with early detection and screening being critical for reducing mortality. Social media platforms like Instagram offer a unique opportunity to raise awareness about CRC, particularly during designated awareness months. However, there is limited research on the effectiveness of CRC-related content on Instagram.

Objective: This study aims to examine how Instagram is used to raise awareness about CRC during Colorectal Cancer Awareness Month by analyzing the thematic content and engagement metrics of related posts. The research seeks to identify the prevalent themes, assess audience interaction with these messages, and highlight areas for improvement in leveraging Instagram as a tool for cancer awareness campaigns.

Methods: A total of 150 Instagram posts were collected based on their use of specific hashtags related to CRC awareness (#colorectalcancer, #colorectalcancerawareness, #colorectalcancerawarenessmonth) during March 2024. The text and images in the posts were categorized into themes such as screening and early detection, symptoms, general awareness, risk factors, individual's experiences, representation of racial and ethnic minoritized communities, and representation of women. Engagement metrics, including the number of likes and comments, were also analyzed. Two researchers independently coded the posts, achieving high interrater reliability (Cohen κ=0.93).

Results: Organizational accounts were more active, contributing 82% (n=123) of the 150 posts, compared to 18% (n=27) from individual users. The most frequently mentioned theme was screening and early detection, which made up 37.3% (n=56) of all posts. General awareness came in second at 19.3% (n=29), and risk factors came in third at 12% (n=18). Posts about individual experiences and general awareness received the highest engagement, indicating the effectiveness of personal narratives and broad informational content. Themes related to symptoms and representation of racial and ethnic minoritized communities and women were underrepresented.

Conclusions: This study highlights the potential of Instagram as a platform for promoting CRC awareness, particularly through posts about screening and early detection and personal experiences. However, there is a need for more inclusive and diverse content to ensure a broader reach and impact.

(JMIR Form Res 2025;9:e63344) doi:10.2196/63344

KEYWORDS

social media; colorectal neoplasms; early detection of cancer; public health; health inequities; harnessing; Instagram; colorectal cancer; colorectal cancer awareness; content analysis; cancer-related deaths; detection; screening; mortality; post; early detection

Introduction

Colorectal cancer (CRC) remains one of the most prevalent and deadly forms of cancer worldwide [1,2]. Despite the clear benefits of early detection and screening for reducing mortality rates, many individuals remain unaware of the importance of regular screenings and the risk factors associated with CRC [3-5]. March is designated as Colorectal Cancer Awareness Month, a period dedicated to increasing public knowledge about CRC, promoting early detection and screening, and ultimately reducing the incidence and mortality of this disease [6]. During this month, various health organizations, advocacy groups, and individuals spread awareness through different communication channels, including social media.

XSL•FO
RenderX

Instagram, a visual-centric platform with over a billion users, offers a unique opportunity to disseminate health information widely and engage with a diverse audience [7-9]. The platform's use of hashtags allows for the categorization and amplification of posts, making it easier for users to find and share relevant content. In the context of health awareness campaigns, hashtags such as #colorectalcancer, #colorectalcancerawareness, and #colorectalcancerawarenessmonth can help consolidate posts under a common theme, enhancing the visibility of key messages [10].

Instagram has been shown to be instrumental in disseminating cancer prevention material. A study of Brazilian Instagram accounts during a nationwide cervical cancer campaign revealed an emphasis on secondary prevention measures, such as Pap screenings, and primary prevention strategies, like the human papillomavirus (HPV) vaccine, received less attention [11]. A descriptive study of Instagram posts related to breast cancer found that the most common posts were individual stories or posts discussing treatment [12]. A separate investigation examined the portrayal of skin cancer narratives on Instagram and underscored the platform's potential for sharing personal experiences and fostering prevention efforts and awareness [13]. A study that searched specific hashtags on Instagram for posts about acute myeloid leukemia found that the most frequently posted type of content by patients were health updates, and a quarter of these posts described symptom experiences [14]. Finally, a content analysis was conducted on information related to CRC on Instagram using the hashtag #colorectalcancer and found the most common posts supported someone with CRC, provided an individual story, or discussed treatment [15]. Collectively, these studies highlight Instagram's potential as a tool for raising cancer awareness and promoting preventive behaviors, while emphasizing the need for further research to identify the most impactful types of cancer prevention messaging and to optimize social media platforms for effective dissemination of cancer prevention information [16-20].

Despite the potential of Instagram to raise awareness about CRC, there is limited research on how effectively the platform is used for this purpose, particularly during dedicated cancer awareness months. This study aims to fill this gap by examining the engagement of Instagram posts related to CRC during Colorectal Cancer Awareness Month. By analyzing the engagement metrics of these posts, this research seeks to understand the current state of CRC awareness on Instagram and identify areas to improve cancer communication strategies. By analyzing Instagram posts based on specific hashtags and thematic content, the findings offer insights into the types of messages that resonate most with the audience and highlight the strengths and weaknesses of current CRC awareness efforts on social media. Ultimately, this research contributes to the broader goal of enhancing public health campaigns and increasing the effectiveness of social media as a tool for health promotion.

Methods

Data Collection

The data for this study were collected from Instagram, focusing specifically on public posts related to CRC published from March 1 to 31, 2024, which is Colorectal Cancer Awareness Month. These dates were chosen to capture the peak of awareness activities and messages. Using Instagram's search functionality, the top 150 posts containing hashtags related to CRC were identified and collected. The hashtags used for this search included #colorectalcancer, #colorectalcancerawareness, #colorectalcancerawarenessmonth, and combinations of these hashtags. To ensure comprehensive coverage, a supplementary search was conducted using additional terms such as "CRC," "Colon cancer," and "Colorectal Cancer." This supplementary search confirmed that the top posts identified through these terms consistently included 1 or more of the primary hashtags, indicating that the initial hashtag-based approach effectively captured relevant posts. This process minimized the risk of excluding significant content while aligning with the study's focus on thematic content and engagement during Colorectal Cancer Awareness Month. The number of likes and comments on the posts served as a gauge of their popularity, ensuring a representative sample of both highly engaged content and content with little engagement.

Given that institutional accounts often repost the same content multiple times during awareness campaigns, we implemented a systematic approach to handle duplicate posts. If a post was identical in both text and hashtags and originated from the same account, only the first instance was included in the dataset to prevent overrepresentation. However, if a repost contained minor variations in wording or hashtag use, it was retained to capture potential differences in engagement strategies. Engagement metrics (likes and comments) were analyzed separately for each retained post, rather than aggregating them across duplicates, as each post had the potential to generate unique interactions. This approach ensured that our dataset reflected a diverse range of content while minimizing redundancy and maintaining the integrity of engagement analysis.

Coding Process

This study employed a deductive content analysis approach to analyze both the text and images of Instagram posts related to CRC awareness. The thematic categories used for coding were predefined based on existing literature and theoretical frameworks in health communication [21-23]. The deductive approach was intentionally chosen to assess the prevalence and resonance of these theoretically supported themes, as reflected by engagement metrics (eg, likes, comments). This structured framework allowed for a systematic evaluation of whether established health communication messages were effectively disseminated through Instagram posts during Colorectal Cancer Awareness Month.

The coding framework was developed based on a review of existing literature on health communication and social media and common themes in CRC awareness [11-15]. Screening and early detection posts emphasized the importance of regular

XSL•FO
RenderX

screenings and early detection methods. Symptom posts provided information on the symptoms of CRC.

General awareness posts were aimed at raising general awareness about CRC through encouraging others to post about it or stating general statistics. Risk factor posts discussed factors that increase the risk of developing CRC and, conversely, preventative lifestyle measures one can take to reduce their risk. Individuals' experience posts shared personal stories and experiences of individuals affected by CRC. Because racial and ethnic minoritized groups and women face significant barriers to screening and care due to socioeconomic issues, limited health care access, and cultural or systemic biases, we examined posts on racial and ethnic minoritized communities and women to uncover messaging gaps and improve future awareness efforts' reach and impact [1,3]. Therefore, the representation of racial and ethnic minoritized community posts highlighted the vulnerability of racial and ethnic minoritized communities to CRC, and the representation of women posts focused on how CRC affects women and the risks it poses. Irrelevant posts were not related to CRC or awareness activities. Finally, other posts contained relevant information about CRC but did not fit into the predefined categories.

Captions and images contained in the Instagram posts were coded to align with the predefined thematic categories. Each post was examined holistically, with its image and caption assessed collectively to determine the dominant theme. Images were categorized based on their primary visual content and captions were analyzed alongside images to provide context and ensure consistency in thematic classification. If the image and caption conveyed different themes, coders determined which element carried the primary message, prioritizing captions for text-heavy posts and images for visually dominant content. In cases where thematic classification was unclear, discrepancies were resolved through discussion between researchers to ensure consistency and accuracy in coding.

To ensure reliability and minimize bias, a systematic coding process was implemented. One researcher (AS) initially coded all posts in the dataset, applying the predefined thematic categories based on existing literature and the objectives of the study. To validate the coding and minimize potential bias, a second researcher (JPS) independently coded a random sample of 10% of the posts. Discrepancies in coding between the 2 researchers were resolved through discussion and consensus, refining the thematic framework as needed. The level of agreement between the 2 researchers for the text and image content of the posts was measured using the κ reliability statistic, which resulted in a high agreement rate of 93%. This high level of interrater reliability confirms the consistency and accuracy of the thematic classification.

Analysis of Content and Engagement

Hashtags and post content were analyzed separately as they serve distinct functions in social media communication. Hashtag analysis was conducted to understand how users label and promote their posts within the broader discourse of CRC awareness, which helped identify trending topics, user engagement strategies, and the reach of campaign-related messaging. In contrast, analyzing the textual content (captions) and images within posts to determine the thematic focus of the message itself allowed us to assess whether the themes promoted through hashtags aligned with the actual messages in the post captions and images.

Our analysis began by describing the distribution of hashtags used for our search during Colorectal Cancer Awareness Month and the engagement of Instagram posts in terms of posts, likes, comments, and hashtag counts. Then, we described the content of the Instagram posts by hashtag, thematic category, and user profile (individual versus organizational account). This study adheres to the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines for reporting observational studies.

Ethical Considerations

The University of Texas Southwestern Medical Center Institutional Review Board determined that the study does not meet the definition of human subjects research and therefore does not require institutional review board approval or oversight (Y1-23-0390). As the data used are publicly available on Instagram, no anonymization or deidentification was required.

Results

This study analyzed a total of 150 Instagram posts related to CRC during Colorectal Cancer Awareness Month. These posts were categorized based on specific hashtags and the captions and image content of each post were classified into various thematic categories to understand the messaging and engagement trends.

Table 1 presents the distribution and engagement metrics of the posts across different hashtags. Posts tagged with #colorectalcancer received the highest engagement among the posts that were singularly tagged, amassing 9751 likes and 560 comments from 16 posts. This indicates a substantial interaction level, suggesting that this hashtag effectively attracts attention and engagement from users. The hashtag #colorectalcancerawarenessmonth was the most frequently used, appearing in 45 posts. These posts collectively garnered 6065 likes and 188 comments, reflecting their widespread use and moderate engagement. #colorectalcancerawareness was utilized in 31 posts, which received 3712 likes and 66 comments. This shows a lower engagement compared to the other hashtags, despite its considerable use. Posts with multiple tags were the most common, with 58 instances. These posts achieved the highest overall engagement, accumulating 17,729 likes and 120 comments. The use of multiple hashtags likely expanded the reach and visibility of the posts.

XSL•FO
RenderX

Table . Hashtag distribution and engagement of 150 Instagram posts during Colorectal Cancer Awareness Month, March 1-31, 2024.

Hashtag	Posts, n	Likes, n	Comments, n	Hashtag count, n
#colorectalcancer	16	9751	560	107
#colorectalcancerawareness-month	45	6065	188	132
#colorectalcancerawareness	31	3712	66	100
Multiple tags	58	17,729	120	414

Table 2 classifies the content of the Instagram posts by hashtag and thematic category. Screening and early detection was the most prevalent theme, featuring in 56 posts. This underscores a strong focus on promoting regular screenings and the importance of early detection in preventing the severity and mortality associated with CRC. General awareness was the theme in 29 posts, which highlighted efforts to spread general knowledge about CRC, including statistics and awareness activities. Risk factors were discussed in 18 posts, educating users about the various factors that increase the risk of developing CRC. Posts about the symptoms of CRC appeared in 12 instances, providing crucial information on recognizing the signs of the disease. Individual's experiences were shared in 16 posts, offering personal narratives and testimonies from individuals affected by CRC, which can be powerful in humanizing the disease and encouraging others to take preventive actions. The representation of racial and ethnic minoritized communities and women was notably limited, with only 3 and 2 posts, respectively. This indicates a gap in addressing the specific risks and experiences of these groups, suggesting a need for more inclusive messaging. The other category included 14 posts that were relevant but did not fit into the predefined categories, reflecting a diversity of information and perspectives related to CRC. Only 1 post was classified as irrelevant, demonstrating that most posts were pertinent to the topic.

Table . Content classification of 150 Instagram posts by hashtag during Colorectal Cancer Awareness Month, March 1-31, 2024.

Hashtag	Screening or early detection (n=56), n	Symptoms (n=12), n	General awareness (n=29), n	Risk factors (n=18), n	Individual's experience (n=16), n	Representation of racial and ethnic minority communities (n=3), n	Representation of women (n=2), n	Other (n=14), n	Irrelevant (n=1), n
#colorectal-cancer (n=16)	5	1	1	1	4	0	1	2	1
#colorectal-canceraware-nessmonth (n=45)	21	3	4	5	5	1	1	5	0
#colorectal-canceraware-ness (n=31)	13	3	6	5	2	1	0	1	0
Multiple tags (n=59)	17	5	18	7	5	1	0	6	0

Table 3 categorizes the 150 Instagram posts by user profile type—individual users (denoted as "Person") and organizational accounts (denoted as "Organization"). The distribution of these categories was analyzed across the 2 user profile types. Organizational accounts were more active overall, contributing 82% (n=123) of the total posts, compared to 18% (n=27) from individual users. Screening and early detection emerged as the predominant category overall, with 37.3% (56/150) of all posts falling into this category. This category was significantly more common among organizational accounts, constituting 41.4% (51/123) of their posts, compared to 19% (5/27) of posts from individual users. Posts related to symptoms accounted for 8% (12/150), with a nearly equal distribution between individual and organizational accounts. The general awareness category comprised 19% (n=29) of the total posts, with individual users and organizational accounts contributing 19% (5/27) and 19.5% (24/123) of their posts, respectively. Risk factors were addressed in 12% (18/150) of the posts. Both individual users and organizational accounts showed interest in this category, with 19% (5/27) of individual users' posts and 10.5% (13/123) of organizational accounts' posts mentioning risks. Posts discussing individuals' experiences accounted for 10.6% (16/150) of the total, with greater representation from individual users at 19% (5/27) compared to organizational accounts. Posts representing racial and ethnic minority communities were notably sparse, comprising only 2% (3/150) of the total posts. Individual users contributed 7% (2/27) and organizational accounts just 0.8% (1/123). Similarly, posts focusing on the representation of women were minimal, representing 1.3% (2/150) of the total posts, with no contributions from individual users and 1.6% (2/123) from organizational accounts. The other category included 9.3% (14/150) of the total posts, with individual users

XSL•FO
RenderX

contributing 7% (2/27) and organizational accounts contributing 9.7% (12/123). Lastly, irrelevant posts constituted 0.6% (1/150) of the total, all from individual users at 4% (1/27), with no contributions from organizational accounts.

Table . Content classification of 150 Instagram posts by user profile during Colorectal Cancer Awareness Month, March 1-31, 2024.

	Person (n=27), n (%)	Organization (n=123), n (%)	Total (N=150), n (%)
Screening and early detection	5 (19)	51 (41.4)	56 (37.3)
Symptoms	2 (7)	10 (8.1)	12 (8)
General awareness	5 (19)	24 (19.5)	29 (19.3)
Risk factors	5 (19)	13 (10.5)	18 (12)
Individual's experience	5 (19)	10 (8.1)	15 (10)
Representation of racial and ethnic minoritized communities	2 (7)	1 (0.8)	3 (2)
Representation of women	0 (0)	2 (1.6)	2 (1.3)
Other	2 (7)	12 (9.7)	14 (9.3)
Irrelevant	1 (4)	0 (0)	1 (0.6)

Discussion

Principal Findings

The goal of our research study was to explore the thematic content and engagement patterns of Instagram posts shared during Colorectal Cancer Awareness Month, with a focus on understanding how key messages about CRC prevention and awareness are conveyed and received. We analyzed the top 150 Instagram posts, and this methodological approach ensured a robust analysis of the content and engagement of CRC-related posts on Instagram, providing insights into the effectiveness of social media as a tool for health awareness campaigns. Our findings highlighted the diverse approaches used to raise awareness about CRC on Instagram.

A key finding was the dominance of themes related to screening and early detection, which reflects the goal of Colorectal Cancer Awareness Month to raise awareness about CRC and motivate individuals to participate in screening, and also suggests that public health campaigns are effectively leveraging Instagram to promote these critical aspects of CRC prevention [2,3,15]. The significant engagement metrics, particularly the high number of likes and comments associated with the hashtag #colorectalcancer, suggest that this specific tag resonates well with the Instagram audience. Additionally, the use of multiple hashtags in posts appeared to enhance engagement, suggesting that combining hashtags may increase the visibility and interaction of the posts.

While the themes of general awareness and risk factors were well represented, the relatively limited number of posts addressing symptoms and individual experiences points to potential gaps in the current messaging strategy. Personal stories can be particularly powerful in motivating behavior change and should be utilized more extensively [22-24]. Furthermore, the underrepresentation of posts targeting racial and ethnic minoritized communities and women highlights a critical gap in the current approach. Given the disparities in CRC incidence and outcomes among different demographic groups, it is essential to tailor messages to address these disparities and ensure more inclusive outreach [1,3].

This study supports the existing literature on the effectiveness of visual content in attracting user interaction [10-15]. However, discrepancies may exist regarding the types of messages that are most effective. For instance, while some studies suggest that statistical information is persuasive, the current findings highlight the potential impact of personal narratives, which may be underutilized in existing campaigns [24].

Implications for Public Health Practice

The results of this study have several implications for public health practice. First, there is a need for more tailored campaigns that specifically address the risks and experiences of marginalized groups and women [25]. Developing culturally sensitive and inclusive content can help bridge the current gaps in outreach [26]. Second, incorporating more individual experiences and testimonies in posts can create more relatable and impactful messages, potentially increasing awareness and encouraging preventive behaviors [27]. Lastly, public health campaigns should consider using multiple and specific hashtags to maximize reach and engagement [28].

Limitations and Future Research

This study has several strengths, including a comprehensive analysis of the top posts on a significant awareness month and high interrater reliability in content classification, which ensures consistency in the findings. However, there are also limitations. The study is limited to 1 social media platform (Instagram) and 1 specific month, which may not capture the full spectrum of awareness activities and engagement throughout the year on other social media platforms. Instagram was chosen due to its visual-centric nature, which is particularly well suited for health awareness campaigns, and its algorithmic features that facilitate content discovery and engagement through hashtags. Additionally, the overlap between Instagram and Facebook, both owned by Meta, often results in similar campaign content across the 2 platforms, providing a representative snapshot of Meta's ecosystem. Moreover, the cross-sectional nature of the

XSL•FO
RenderX

study limits the ability to assess changes over time or the long-term impact of these posts.

We did not assess the credibility and quality of the Instagram posts, which was a deliberate decision, as the study aimed to explore thematic content and user engagement rather than verify the accuracy or reliability of the posts. The credibility of posts is important for assessing the public health value of messages [29]. Future research should consider assessing the quality and credibility of health-related social media content to better understand its reliability and influence on public health outcomes.

Future research should consider longitudinal studies to evaluate the impact of social media campaigns over time. Comparative studies across different social media platforms would also be beneficial to identify the most effective channels for health promotion. Furthermore, research focusing on the development and testing of tailored messages for underrepresented groups is necessary to enhance the inclusivity and effectiveness of these campaigns. Finally, while this study focused on predefined themes, we recognize the value of inductive approaches, such as topic modeling, for identifying emergent themes or uncovering patterns directly from the data. Future research could complement our findings by employing such methods to explore new dimensions of social media health communication.

Conclusion

In conclusion, this study highlights the potential of Instagram as a platform for promoting CRC awareness, particularly for screening and early detection. While certain themes effectively engage users and align with the goals of awareness campaigns, critical gaps remain in inclusivity and representation, particularly for marginalized communities. To maximize the impact of future campaigns, messaging strategies should be refined and incorporate a broader range of themes, ensuring that social media platforms like Instagram are leveraged more effectively to raise awareness, reduce disparities, and promote preventive health behaviors.

Acknowledgments

The American Cancer Society provided funding for JPS' research (award DBG-23-1155771-01-HOPS). The content is solely the responsibility of the authors and does not necessarily represent the official views of the funder. The funder had no role in the study design, data analysis, decision to publish, or preparation of the manuscript.

Data Availability

The datasets analyzed during this study are publicly available on Instagram.

Authors' Contributions

All authors contributed to the study's conception and design. Data collection and analysis were performed by AS. JPS wrote the first draft of the manuscript. All authors contributed to subsequent drafts of the manuscript. All authors read and approved the final version of the manuscript.

Conflicts of Interest

None declared.

References

1. Tan JY, Yeo YH, Ng WL, Fong ZV, Brady JT. How have US colorectal cancer mortality trends changed in the past 20 years? Int J Cancer 2024 Aug 1;155(3):493-500. [doi: 10.1002/ijc.34926] [Medline: 38525799]
2. Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2024;74(3):229-263. [doi: 10.3322/caac.21834] [Medline: 38572751]
3. Cronin KA, Scott S, Firth AU, et al. Annual report to the nation on the status of cancer, part 1: national cancer statistics. Cancer 2022 Dec 15;128(24):4251-4284. [doi: 10.1002/cncr.34479] [Medline: 36301149]
4. Doubeni CA, Fedewa SA, Levin TR, et al. Modifiable failures in the colorectal cancer screening process and their association with risk of death. Gastroenterology 2019 Jan;156(1):63-74. [doi: 10.1053/j.gastro.2018.09.040] [Medline: 30268788]
5. Smith RA, Fedewa S, Siegel R. Early colorectal cancer detection-current and evolving challenges in evidence, guidelines, policy, and practices. Adv Cancer Res 2021;151:69-107. [doi: 10.1016/bs.acr.2021.03.005] [Medline: 34148621]
6. National Colorectal Cancer Awareness Month. Colorectal Cancer Alliance. URL: https://colorectalcancer.org/get-involved/advocate/colorectal-cancer-awareness-month [accessed 2025-02-12]
7. Pantel HJ, Kleiman DA, Kuhnen AH, Marcello PW, Stafford C, Ricciardi R. Has National Colorectal Cancer Awareness Month increased endoscopy screening rates and public interest in colorectal cancer? Surg Endosc 2021 Jan;35(1):398-405. [doi: 10.1007/s00464-020-07413-x] [Medline: 32016518]
8. Smith K. 50 incredible Instagram statistics you need to know. Brandwatch. 2019. URL: https://www.brandwatch.com/blog/instagram-stats/ [accessed 2025-02-12]

XSL•FO
RenderX

9. Dixon SJ. Number of monthly active Instagram users from January 2013 to December 2021. Statista. 2024. URL: https://www.statista.com/statistics/253577/number-of-monthly-active-instagram-users/ [accessed 2025-02-12]
10. Gatewood J, Monks SL, Singletary CR, Vidrascu E, Moore JB. Social media in public health: strategies to distill, package, and disseminate public health research. J Public Health Manag Pract 2020;26(5):489-492. [doi: 10.1097/PHH.0000000000001096] [Medline: 32732723]
11. Vicente EP, De Faria SEE, Almeida ABL, et al. Cervical cancer prevention on Instagram: content and social interaction analysis of Brazilian accounts. Asian Pac J Cancer Prev 2022 Sep 1;23(9):3043-3049. [doi: 10.31557/APJCP.2022.23.9.3043] [Medline: 36172667]
12. Basch CH, MacLean SA. Breast cancer on Instagram: a descriptive study. Int J Prev Med 2019;10(1):166. [doi: 10.4103/ijpvm.IJPVM_36_19] [Medline: 32133084]
13. Gomaa B, Houghton RF, Crocker N, Walsh-Buhi ER. Skin cancer narratives on Instagram: content analysis. JMIR Infodemiology 2022;2(1):e34940. [doi: 10.2196/34940] [Medline: 37113805]
14. Parker C, Zomer E, Liew D, Ayton D. Characterising experiences with acute myeloid leukaemia using an Instagram content analysis. PLoS ONE 2021;16(5):e0250641. [doi: 10.1371/journal.pone.0250641] [Medline: 33939746]
15. Basch CH, MacLean SA. Colorectal cancer on Instagram: a content analysis. J Consum Health Internet 2019 Oct 2;23(4):378-383. [doi: 10.1080/15398285.2019.1684149]
16. Pellino G, Simillis C, Qiu S, et al. Social media and colorectal cancer: a systematic review of available resources. PLoS ONE 2017;12(8):e0183031. [doi: 10.1371/journal.pone.0183031] [Medline: 28832603]
17. Plackett R, Kaushal A, Kassianos AP, et al. Use of social media to promote cancer screening and early diagnosis: scoping review. J Med Internet Res 2020 Nov 9;22(11):e21582. [doi: 10.2196/21582] [Medline: 33164907]
18. Xu S, Markson C, Costello KL, Xing CY, Demissie K, Llanos AA. Leveraging social media to promote public health knowledge: example of cancer awareness via Twitter. JMIR Public Health Surveill 2016;2(1):e17. [doi: 10.2196/publichealth.5205] [Medline: 27227152]
19. Moorhead SA, Hazlett DE, Harrison L, Carroll JK, Irwin A, Hoving C. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. J Med Internet Res 2013 Apr 23;15(4):e85. [doi: 10.2196/jmir.1933] [Medline: 23615206]
20. Laranjo L, Arguel A, Neves AL, et al. The influence of social networking sites on health behavior change: a systematic review and meta-analysis. J Am Med Inform Assoc 2015 Jan;22(1):243-256. [doi: 10.1136/amiajnl-2014-002841] [Medline: 25005606]
21. Elo S, Kyngäs H. The qualitative content analysis process. J Adv Nurs 2008 Apr;62(1):107-115. [doi: 10.1111/j.1365-2648.2007.04569.x] [Medline: 18352969]
22. Wu J, Benjamin EJ, Ross JC, Fetterman JL, Hong T. Health messaging strategies for vaping prevention and cessation among youth and young adults: a systematic review. Health Commun 2024 May 14:1-19. [doi: 10.1080/10410236.2024.2352284] [Medline: 38742648]
23. Zhao YC, Zhao M, Song S. Online health information seeking among patients with chronic conditions: integrating the health belief model and social support theory. J Med Internet Res 2022 Nov 2;24(11):e42447. [doi: 10.2196/42447] [Medline: 36322124]
24. Hinyard LJ, Kreuter MW. Using narrative communication as a tool for health behavior change: a conceptual, theoretical, and empirical overview. Health Educ Behav 2007 Oct;34(5):777-792. [doi: 10.1177/1090198106291963] [Medline: 17200094]
25. Agarwal V. Editorial: communicating for social justice in health contexts: creating opportunities for inclusivity among marginalized groups. Front Commun 2022 Sep 30;7:1008533. [doi: 10.3389/fcomm.2022.1008533]
26. Häfliger C, Diviani N, Rubinelli S. Communication inequalities and health disparities among vulnerable groups during the COVID-19 pandemic - a scoping review of qualitative and quantitative evidence. BMC Public Health 2023 Mar 6;23(1):428. [doi: 10.1186/s12889-023-15295-6] [Medline: 36879229]
27. de Graaf A, Sanders J, Hoeken H. Characteristics of narrative interventions and health effects: a review of the content, form, and context of narratives in health-related narrative persuasion research. RCR 2016;4:88-131. [doi: 10.12840/issn.2255-4165.2016.04.01.011]
28. Ghahramani A, Prokofieva M, de Courten MP. Health promotion campaigns using social media: association rules mining and co-occurrence network analysis of Twitter hashtags. BMC Public Health 2025 Jan 7;25(1):67. [doi: 10.1186/s12889-024-21255-5] [Medline: 39773431]
29. Etta RE, Babatunde AO, Okunlola PO, et al. The assessment of TikTok as a source of quality health information on human papillomavirus: a content analysis. Cureus 2024 Dec;16(12):e75419. [doi: 10.7759/cureus.75419] [Medline: 39791048]

Abbreviations

CRC: colorectal cancer
HPV: human papillomavirus
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

XSL•FO
RenderX

Edited by A Mavragani; submitted 17.06.24; peer-reviewed by D Surian, Y Sarikhani; revised version received 22.01.25; accepted 29.01.25; published 19.02.25.

Please cite as:
Srivastava A, Stimpson JP
Instagram Posts Promoting Colorectal Cancer Awareness: Content Analysis of Themes and Engagement During Colorectal Cancer Awareness Month
JMIR Form Res 2025;9:e63344
URL: https://formative.jmir.org/2025/1/e63344
doi:10.2196/63344

© Aditi Srivastava, Jim P Stimpson. Originally published in JMIR Formative Research (https://formative.jmir.org), 19.2.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Quality Assessment of Medical Institutions' Websites Regarding Prescription Drug Misuse of Glucagon-Like Peptide-1 Receptor Agonists by Off-Label Use for Weight Loss: Website Evaluation Study

Rie Oyama[1], MD; Tsuyoshi Okuhara[1], MPH, PhD; Emi Furukawa[2], MD, PhD; Hiroko Okada[1], MPH, PhD; Takahiro Kiuchi[1], MD, PhD

[1]Department of Health Communication, School of Public Health, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan
[2]University Hospital Medical Information Network, The University of Tokyo Hospital, Tokyo, Japan

Corresponding Author:
Rie Oyama, MD
Department of Health Communication
School of Public Health
Graduate School of Medicine, The University of Tokyo
7-3-1 Hongo
Bunkyo-ku
Tokyo, 113-8655
Japan
Phone: 81 3 5800 6549
Fax: 81 3 5689 0726
Email: rooyama-oss@umin.ac.jp

Abstract

Background: Misuse of glucagon-like peptide-1 receptor agonists (GLP-1RAs) has emerged globally as individuals increasingly use these drugs for weight loss because of unrealistic and attractive body images advertised and shared on the internet.

Objective: This study assesses the quality of information and compliance with Japan's medical advertising guidelines on the websites of medical institutions that prescribe GLP-1RAs off-label for weight loss.

Methods: Websites were identified by searching Google and Yahoo! by using keywords related to GLP-1RAs and weight loss in August 2024. The quality of information on these websites was assessed using the DISCERN instrument. To comply with Japan's medical advertising guidelines, we evaluated whether the 5 mandatory items for advertisements of self-paid medical treatments involving the off-label use of drugs were stated and whether there were any exaggerated claims. The content of the exaggerated advertisements was categorized into themes.

Results: Of the 87 websites included, only 1 website stated all 5 mandatory items. Websites listing "ineligible for the relief system for sufferers from adverse drug reactions" had the lowest percentage at 9% (8/87), while 83% (72/87) of the websites listed exaggerated advertisements. Approximately 69% (60/87) of the websites suggested that no exercise or dietary therapy was required, 24% (21/87) suggested that using GLP-1RAs is a natural and healthy method, and 31% (27/87) of the websites provided the author's personal opinions on the risks of using GLP-1RAs. The mean total DISCERN score for all 87 websites was 32.6 (SD 5.5), indicating low quality. Only 1 website achieved a good rating, and 9 websites were rated as fair. The majority of the websites were rated as poor (72 websites) or very poor (5 websites).

Conclusions: We found that the quality of information provided by the websites of medical institutions prescribing GLP-1RAs off-label for weight loss was very low and that many websites violated Japan's medical advertising guidelines. The prevalence of exaggerated advertisements, which may lead consumers to believe that they can lose weight without dietary or exercise therapy, suggests the risk of GLP-1RA misuse among consumers. Public institutions and health care providers should monitor and regulate advertisements that violate guidelines and provide accurate information regarding GLP-1RAs, obesity, and weight loss.

(JMIR Form Res 2025;9:e68792) doi:10.2196/68792

XSL•FO
RenderX

KEYWORDS

prescription drug misuse; GLP-1 receptor agonists; off-label use; weight loss; information quality; DISCERN; web-based information; information provision; misinformation; advertising guidelines; exaggerated advertisements

Introduction

Prescription drug misuse is an emerging global public health issue [1-3]. According to a previous study [4], misuse is defined as the therapeutic use of legal drugs in a manner other than as prescribed, and abuse is defined as the use of illegal drugs for nontherapeutic purposes such as psychological pleasure and weight loss. The misuse and abuse of drugs can lead to health problems. In the United States and European countries, misuse of opioid analgesics, benzodiazepines, over-the-counter cough suppressants containing dextromethorphan and codeine, and psychostimulants have been observed [1-3,5]. Further, in Japan, misuse of over-the-counter cough suppressants, pain relievers, antianxiety medications, and sedative drugs has been observed [6,7]. However, misuse of glucagon-like peptide-1 receptor agonists (GLP-1RAs) has become a problem in recent years [8-10].

GLP-1RAs are the most effective drugs for obesity treatment [11]. Semaglutide (Wegovy), a type of GLP-1RA, was approved by the US Food and Drug Administration (FDA) as an obesity treatment drug in 2021 [12]. Since then, its market share has rapidly expanded [13]. Conversely, multiple news media outlets have reported concerns about the misuse of GLP-1RAs [14,15]. Due to unrealistic and attractive body images that are advertised by celebrities on social media, people without obesity allegedly obtain and use GLP-1RAs illicitly for weight loss purposes [16]. Adverse events of GLP-1RA misuse have also been reported [8,9]. A previous study showed that the use of semaglutide in patients with overweight/obesity may be associated with the incidence of nonarteritic anterior ischemic optic neuropathy, a severe ophthalmologic condition that can lead to blindness [17]. Another previous study showed that GLP-1RA users for weight management had an increased risk of serious gastrointestinal disorders compared to bupropion-naltrexone users, with adjusted hazard ratios for pancreatitis, bowel obstruction, and gastroparesis reported as 9.09, 4.22, and 3.67, respectively [18]. However, the use of GLP-1RAs for weight management is a relatively new therapeutic approach, and the long-term side effects associated with their use remain unknown. Additionally, the effectiveness and potential adverse effects of using GLP-1RAs for further weight loss in individuals without obesity also remain unknown. This increased demand has caused a global shortage of GLP-1RAs [16]. The European Medicines Agency has announced recommendations by manufacturers, health care professionals, and citizens for the proper use of GLP-1RAs [19].

In Japan, Wegovy was approved for the treatment of obesity on March 27, 2023. According to the Optimal Use Promotion Guideline by the Japan's Ministry of Health, Labor, and Welfare, patients eligible for obesity treatment are "those with BMI ≥35 kg/m² or those with BMI ≥27 kg/m² who also have 2 or more health conditions related to obesity and who do not achieve sufficient results from diet and exercise therapy" [20]. The guidelines also strictly prohibit the use of GLP-1RAs for weight loss purposes without obesity as well as institutional criteria for prescribing GLP-1RAs under insurance-covered medical care [20,21]. Japan's Ministry of Health, Labor, and Welfare released a "Request for cooperation due to tight inventory of GLP-1 receptor agonist" [22], and the Japan Diabetes Society [23] and Japan Society for the Study of Obesity [24] have made recommendations regarding the off-label use of GLP-1RAs.

However, several private clinics in Japan prescribe GLP-1RAs for off-label weight loss. Japan's National Consumer Affairs Center and Consumer Affairs Agency warns citizens against the use of GLP-1RAs for weight loss purposes because of the possibility of medical institutions prescribing GLP-1RAs through advertisements that are prohibited by guidelines, inadequate physician consultations and explanations, and inadequate systems to manage adverse events [25,26]. Websites of medical institutions in Japan are regulated by the Medical Care Act [27] and the Ministry of Health, Labor and Welfare's "Regulations on Advertisements of Hospitals, etc under the Medical Care Act" (Japan's medical advertising guidelines) [28], which prohibit false or exaggerated advertisements. The reason for this is "because false advertisements may cause patients to lose the opportunity to receive appropriate medical consultation or receive inappropriate medical treatments by giving them information that is significantly different from the facts, etc"[28].

In previous studies on health information quality, these issues were assessed as part of the reliability of health information. The DISCERN instrument developed by Charnock et al [29] has been used in many studies to assess the publication reliability and the quality of the written health information (Multimedia Appendix 1). The validity and reliability of the DISCERN instrument were ensured, and in addition to the original English language version, Chinese [30], Spanish [31], Brazilian Portuguese [32], and Japanese [33] versions have been developed [29]. Previous studies using the DISCERN instrument have examined the quality of information on websites regarding posttraumatic stress disorder [34], breast cancer [35], nasopharyngeal carcinoma [36], and cosmetic injectability [37]. These studies showed that the quality of individual websites varied considerably: the quality of information was higher for academic and government institutions that operate websites and lower for commercial institutions [34-37]. A systematic review showed that in research using the DISCERN instrument, none found the mean DISCERN score of all websites to be excellent, and the information quality of the majority of websites ranged from fair to very good [38].

In recent years, most people have searched the internet for health information [39]. The internet usage rate for individuals in Japan in 2022 was 84.9% and exceeded 90% when limited to the 10-50 years age range [40]. For cosmetic treatments, 95% of the patients search for information on the internet before entering a doctor's office [41]. Hence, if medical institutions' websites

XSL•FO
RenderX

on GLP-1RAs contain misleading, exaggerated, or unreliable information, these websites should be immediately assessed and improved. However, to our knowledge, no study has investigated the quality of information that people without obesity obtain from websites when they consider obtaining GLP-1RAs for weight loss. This study aims to assess compliance with medical advertising guidelines and the quality of information on the websites of medical institutions that prescribe GLP-1RAs off-label for weight loss in Japan.

Methods

Data Collection

Search Terms

We used a Japanese-language search string input into Google Japan [42] and Yahoo Japan [43], which has the largest search engine market shares in Japan, in August 2024 [44]. The GLP-1RAs approved by the FDA for obesity treatment are liraglutide, semaglutide, and tirzepatide that are sold in Japan under the product names Rybelsus, Ozempic, Victoza, Saxenda, and Mounjaro. The search keywords comprised 6 in total (Rybelsus diet, Ozempic diet, Victoza diet, Saxenda diet, Mounjaro diet, and GLP-1 diet), which included each product name combined with diet (used specifically to mean weight loss in Japan), as well as the comprehensive term GLP-1 diet. Wegovy was excluded from the search terms because it was the newest drug and was not available in the general market in Japan. Saxenda is approved in Japan for the treatment of diabetes under the name Victoza, but private clinics that use the drug off-label advertise it under the name Saxenda; so, it was included in the search terms.

All words were entered in the search window, one keyword at a time, and 20 websites were reviewed per search engine for each term. This is because a previous study showed that 97.2% of the users only view the first 1-10 websites of search results, and 2% click on links on the second page of the search results and beyond [45]. Google Chrome's incognito browser that would have all history deleted was used, and personal accounts were not used.

Inclusion and Exclusion Criteria

The inclusion criteria were as follows: websites written in Japanese; websites created by medical institutions (with medical department labels); websites clearly stating that GLP-1RAs were being sold; and the medical institutions operating the websites were not accredited as facilities capable of treating obesity using GLP-1RAs by the Japanese Circulation Society, Japan Diabetes Society, or Japan Endocrine Society. The exclusion criteria were as follows: duplicate websites, government or academic statements, PDF documents, news reports, affiliate advertisements, and websites that were unclear whether they prescribe GLP-1RAs. The first author (RO) reviewed all the websites in the search results and extracted eligible websites.

Data Extraction

RO recorded the names and URLs of all the websites, names of the medical institutions operating the websites, the medical departments they represented, and the types of GLP-1RAs they sold. If not listed, it was recorded as unknown. Medical institutions can represent multiple medical departments in Japan. Therefore, if several departments were listed, all were recorded. All data were recorded in a Microsoft Excel 2021 spreadsheet by RO.

Data Assessment

First, all identified websites were assessed by an internal medicine physician (RO) for compliance with the medical advertising guidelines in Japan and quality by using the DISCERN instrument. Second, the third author (EF, an internal physician) was trained through a 1-hour meeting and a 1-hour pilot assessment, and then EF independently assessed approximately 20% (20/87) of the total websites. These 20 websites were randomly selected using Microsoft Excel to generate random numbers. Finally, they assessed the agreement between the coders of the 20 websites. The results were recorded using a Microsoft Excel 2021 spreadsheet. More precise information is provided below.

Assessment of Compliance With Medical Advertising Guidelines

In the case of noninsured medical treatments involving off-label use of drugs, it is prohibited to advertise on websites without explicitly stating that "the drugs are not approved, the acquisition route, information on whether there are other domestically approved drugs with the same ingredients and performance, information on the safety of the products in other countries, and they are ineligible for the relief system for sufferers from adverse drug reactions," in Japan's medical advertising guidelines [28]. All the websites were assessed based on whether these 5 items were stated.

Japan's medical advertising guidelines also prohibit exaggerated advertisements [28]. The guidelines [28] define exaggerated advertisements as advertisements that are not necessarily false but that misleads citizens by misrepresenting facts such as about the size of the facility, staffing, or the contents of the medical care provided. Moreover, "misleading citizens" is defined as "it is sufficient to say, as a matter of common sense, that there is a difference between the 'impression' or 'expectation' that citizens perceive from the advertisements and the actual contents, and it is not necessary to prove that the advertisements mislead or to show that the advertisements have actually misled citizens" [27]. Therefore, since taking GLP-1RAs is considered an auxiliary therapy for exercise and dietary therapy in the treatment of obesity, statements such as "you can lose weight with GLP-1RAs alone," "you can lose weight even if you cannot exercise," or "you do not need dietary therapy" are considered exaggerated advertisements. The coders carefully read the guidelines and assessed each statement by using a binary (0/1) scale to indicate the presence or absence of exaggerated advertisements. RO then analyzed the contents of the exaggerated advertisements and categorized them into themes according to the specific examples in the guidelines [28].

Quality Assessment Using the DISCERN Instrument

The DISCERN instrument in the Japanese version [33] was used for the quality assessment of information. The DISCERN

XSL•FO
RenderX

instrument consists of 16 items, each of which is assessed on a 5-point Likert scale (1=not at all, 5=completely). Section 1 (questions 1-8) addresses website reliability, section 2 (questions 9-15) addresses the quality of information about treatment options, and section 3 (question 16) presents the overall assessment. The total score ranged from 16 to 80, with higher scores indicating higher quality information. Several previous studies ranked the quality of websites based on their total score as follows: excellent=80-63, good=62-51, fair=50-39, poor=38-27, very poor=26-16 [35,46].

Statistical Analysis

Data are presented as means for continuous variables and counts with frequency percentages for categorical variables. Cohen κ statistic was used to measure the intercoder reliability for compliance with medical advertising guidelines and DISCERN scores between the first and second coders. Linear regression models were used to investigate the association between the specialties of the clinic operating the website and the total DISCERN score. P values were 2-sided, and $P<.05$ was considered statistically significant. All analyses were conducted using the R version 4.3.1 (June 16, 2023; R Foundation for Statistical Computing).

Ethical Considerations

This study did not involve human or animal participants, and all data were collected from publicly available web-based sources. This study was granted an exemption from ethics approval by the ethics review committee of the Graduate School of Medicine, University of Tokyo.

Results

Characteristics of the Included Websites

In total, 240 websites were identified. After the removal of intrasearch term duplicates and those that met the exclusion criteria, 87 websites were used for the analysis (Figure 1). A total of 67 clinics operated the 87 websites. This was because 8 clinics had multiple different GLP-1RA diet-related articles. Table 1 shows that the most common departments, as far as could be determined from the websites, were dermatology (22/120, 18.3%), cosmetic dermatology (21/120, 17.5%), and internal medicine (19/120, 15.8%); 17 clinics were unaware of their departments. The percentages of advertisements with Ozempic, Rybelsus, Mounjaro, Victoza, and Saxenda were 18.5% (30/162), 29.1% (47/162), 12.3% (20/162),14.2% (23/162), and 25.9% (42/162), respectively. Approximately 18% (16/87) of the websites mentioned selling sodium-glucose cotransporter 2 inhibitor along with GLP-1RAs, and 12% (10/87) mentioned selling metformin.

Figure 1. Search process flowchart. GLP-1: glucagon-like peptide-1; GLP-1RA: glucagon-like peptide-1 receptor agonist.

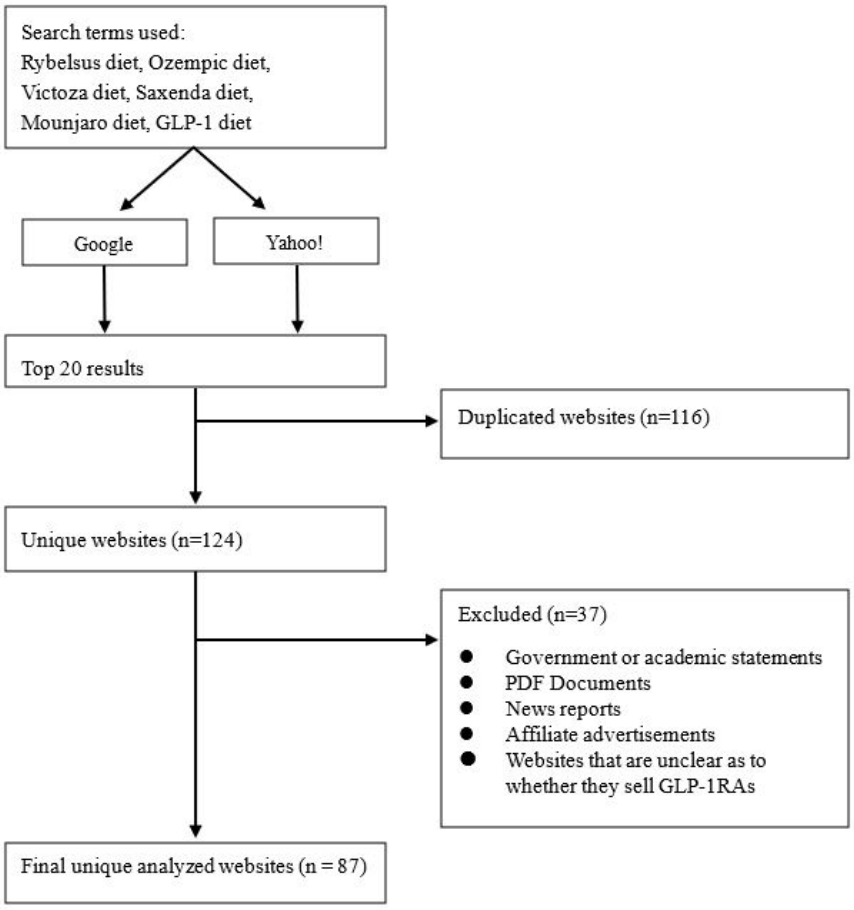

XSL•FO
RenderX

Table 1. Characteristics of the websites regarding prescription drug misuse of glucagon-like peptide-1 receptor agonists by off-label use for weight loss.

	Values, n (%)
Specialty[a] (n=120[b])	
Dermatology	22 (18.3)
Cosmetic dermatology	21 (17.5)
Internal medicine	19 (15.8)
Cosmetic surgery	6 (5)
Urology	6 (5)
Allergology	5 (4.2)
Other	24 (20)
Unknown	17 (14.2)
Types of advertised drugs[c] (n=162[d])	
Ozempic	30 (18.5)
Rybelsus	47 (29)
Mounjaro	20 (12.3)
Victoza	23 (14.2)
Saxenda	42 (25.9)

[a]Medical departments represented by medical institutions.

[b]Total number of mentions of specialty.

[c]Types of glucagon-like peptide-1 receptor agonists that medical institutions advertised for sale.

[d]Total number of mentions of drugs.

Compliance With Medical Advertising Guidelines

The intercoder reliability was acceptable (average Cohen κ=0.667). Table 2 shows the compliance with medical advertising guidelines. Only one of the 87 websites stated the 5 required items. The items with the highest percentage of listing were "acquisition route" and "information on the safety of the drugs in other countries," both at 79% (69/87). The item with the lowest percentage of listing was "ineligible for the relief system for sufferers from adverse drug reactions" at 9% (8/87).

Table 2. Compliance with advertising guidelines of the websites regarding prescription drug misuse of glucagon-like peptide-1 receptor agonists by off-label use for weight loss (N=87).

	Values, n (%)
Content that should be included[a]	
The drug is unapproved	43 (49)
Acquisition route	69 (79)
Information on whether there are other domestically approved drugs with the same ingredients and performance	24 (28)
Information on the safety of the drug in other countries	69 (79)
Ineligible for relief system for sufferers from adverse drug reactions	8 (9)
Exaggerated advertisements[b]	72 (83)

[a]Explicit statements required under Japan's medical advertising guidelines for noninsured medical treatments involving off-label drug use.

[b]Advertisements containing expressions that could mislead citizens, as determined by coders under Japan's medical advertising guidelines.

Table 2 shows that 83% (72/87) of the websites listed exaggerated advertisements. Table 3 shows the content of the exaggerated advertisements: of the 87 websites, 69% (60/87) conveyed messages indicating no exercise and dietary therapy were required, 24% (21/87) conveyed messages indicating that using GLP-1RAs is a natural and healthy method, and 31% (27/87) provided the website author's subjective expression of the risks of using GLP-1RAs.

XSL•FO
RenderX

Table 3. Content of exaggerated advertising regarding prescription drug misuse of glucagon-like peptide-1 receptor agonists by off-label use for weight loss (N=87)[a].

Themes	Illustrative quotes	Values, n (%)
Induction by nonscientific information (no diet or exercise therapy required)	*…Exercise is also unnecessary. Exercise is less effective in losing weight.* *…Medical dieting is a treatment that suppresses appetite and promotes weight loss by taking a medicine called GLP-1 receptor agonist, which is used in the treatment of diabetes, without diet restrictions or hard exercise regimen.*	60 (69)
Induction by nonscientific information (natural and healthy methods)	*…By managing portion sizes, you can reduce calorie intake and engage in a natural and controlled dietary regimen.* *…To achieve weight loss in a healthy manner.* *…It also has fat-burning and pancreas-protecting effects, making it a healthy way to lose weight.*	21 (24)
Subjective expressions of risks	*…Since the delayed gastric emptying effect is not long-lasting, there is little need to consider it dangerous.* *…I, as the director, have also used it and had no side effects other than slight heartburn, which may have caused mild anorexia and slight weight loss.*	27 (31)
Statements that you can always get the drug only through web-based medical consultation	*…Everything from making an appointment and consultation to prescribing a drug is done online, so there is no need to go out.* *…You can see us from anywhere in the country via our web-based clinic. [...] There is no consultation fee; GLP-1 injections will be mailed to you.*	20 (23)

[a]Advertisements containing expressions that could mislead citizens, as determined by coders under Japan's medical advertising guidelines.

Quality Assessment Using DISCERN Instrument

Intercoder reliability was acceptable (average Cohen κ=0.663). Figure 2 shows the DISCERN scores for all 87 websites across each item. Table 4 presents the mean score for each of the 16 DISCERN instrument criteria and the total DISCERN score. The mean total DISCERN score for all 87 websites was 32.6 (SD 5.5). Only 1 website scored more than 51 points, which is good; 9 websites scored between 39 and 50 points, which is fair; 72 websites were poor; and 5 websites were very poor. The items that scored a mean of at least 3 points were criterion 9 (does it describe how each treatment works?) with a mean score of 3.96 (SD 1.10), criterion 2 (does it achieve its aims?) with a mean score of 3.18 (SD 0.70), criterion 1 (are the aims clear?) with a mean score of 3.10 (SD 0.57), and criterion 10 (does it describe the benefits of each treatment?) with a mean score of 3.01 (SD 0.42). The items that scored a mean of less than 2 points were criterion 15 (does it provide support for shared decision-making?), criterion 12 (does it describe what would happen if no treatment was used?), criterion 7 (does it provide details of additional sources of support and information?), criterion 8 (does it refer to the areas of uncertainty?), criterion 14 (is it clear that there may be more than one possible treatment choice?), criterion 6 (is this balanced or unbiased?), criterion 4 (is it clear what sources of information were used to compile the publication?), and criterion 5 (is it clear when the information used or reported in the publication was produced?). The mean score for criteria 16, which assessed the overall quality, was 1.49 (SD 0.73). No statistically significant association was found between the specialties and the total DISCERN score (all $P>.05$).

XSL•FO
RenderX

Figure 2. DISCERN scores for each item: #1 to #16 represent the criteria numbers of the DISCERN instrument. No: identification numbers for individual websites for analysis.

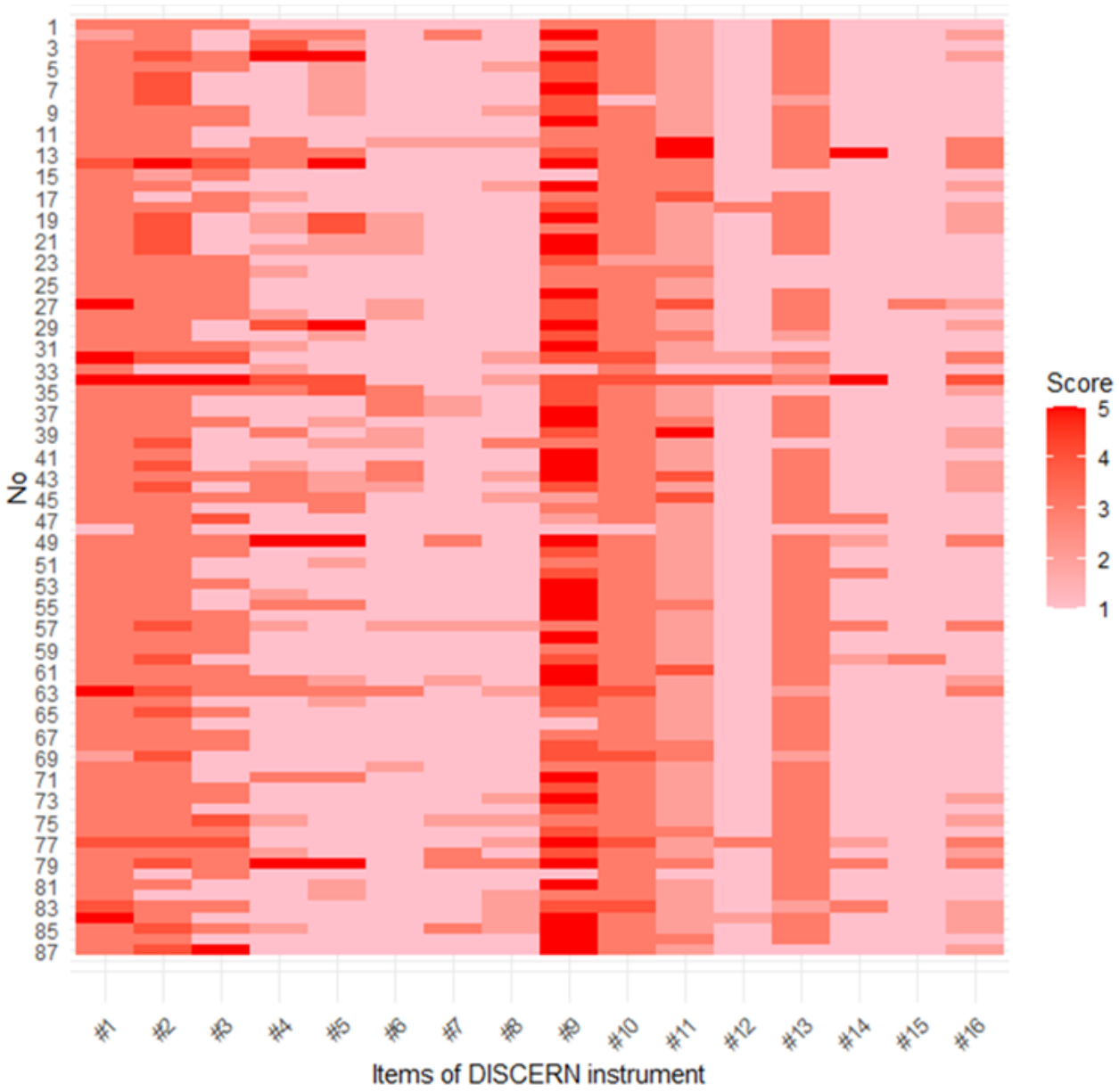

XSL•FO
RenderX

Table 4. Mean score for each of the 16 DISCERN instrument criteria[a].

Criterion number	Criteria	Mean score (SD)
1	Provides clear aims	3.10 (0.57)
2	Achieves its aims	3.18 (0.71)
3	Provides relevant information	2.21 (1.15)
4	Provides sources of information	1.70 (1.07)
5	Provides information production date	1.73 (1.16)
6	Balanced and unbiased	1.28 (0.58)
7	Provides additional sources of information	1.18 (0.52)
8	Refers to areas of uncertainty	1.24 (0.48)
9	Describes how each treatment works	3.97 (1.10)
10	Describes benefits of each treatment	3.01 (0.42)
11	Describes risks of each treatment	2.36 (0.79)
12	Describes what would happen if no treatment is used	1.10 (0.46)
13	Describes how treatment affects quality of life	2.70 (0.67)
14	Makes clear that there may be more than one possible treatment choice	1.24 (0.76)
15	Supports shared decision-making	1.05 (0.30)
16	Overall quality	1.49 (0.73)

[a]Each DISCERN item is assessed on a 5-point Likert scale (1=not at all, 5=completely). Total score = 32.55 (SD 5.53).

Discussion

Principal Findings

This study assesses compliance with Japanese medical advertising guidelines, exaggerated advertisement classification, and DISCERN scores of Japanese websites of medical institutions that prescribe GLP-1RAs off-label for weight loss, which Japanese consumers may refer to when considering weight loss with GLP-1RAs.

Many websites potentially violated Japan's medical advertising guidelines, and very few websites listed all the items required for advertising GLP-1RAs off-label use. In particular, only 9% (8/87) of the websites stated that using GLP-1RAs off-label was not covered by the Adverse Reactions Relief System for Drugs. In recent years, problems related to cosmetic therapy for weight loss have increased in Japan [47], so it is problematic that consumers have limited exposure to information indicating that the off-label use of GLP-1RAs is not covered by the Adverse Reactions Relief System for Drugs. Furthermore, concerns have been raised that GLP-1RAs may be obtained without a prescription and are at risk of misuse and abuse in inappropriate quantities [8]. However, research on who is actually misusing and what kinds of adverse effects occur because of misuse remains scarce. Generally, it has been reported that distinguishing between therapeutic errors, unintentional misuse, and intentional abuse of drugs is challenging, and the content of the reports tends to vary [5]. As mentioned above, in Japan, because it is not covered by the Adverse Reactions Relief System for Drugs, it may be difficult for consumers who have experienced adverse events due to the misuse of GLP-1RAs to claim compensation.

Approximately 83% (72/87) of the websites included exaggerated advertisements, and much of their content may mislead consumers into assuming that one can lose weight by using GLP-1RAs without exercise or dietary therapy. However, since participants in clinical trials evaluating the efficacy of obesity treatment drugs received medication therapy as an auxiliary to lifestyle interventions, the message that the use of GLP-1RAs for weight loss does not require exercise or dietary therapy is unscientific [48]. Additionally, according to another previous study, such messages may evoke a "compensatory health belief" in consumers that "if an alternative behavior is adopted, it is not necessary to engage in healthy behaviors or refrain from unhealthy ones." [49,50]

In this study, the mean of the total DISCERN scores of the analyzed websites was 32.56, which qualified as poor. Individually, most of the websites qualified as poor, only 1 website qualified as good, 9 websites qualified as fair, 72 websites qualified as poor, and 5 websites qualified as very poor. The mean score of criterion 16, the overall quality rating, was also low (score=1.49). These results were extremely low compared to those of previous studies that used the DISCERN instrument to assess the quality of health information [34-36,38]. For each question, criterion 15 (shared decision-making), criterion 12 (if no treatment is used), criterion 7 (additional information sources), criterion 8 (uncertainty), criterion 14 (possible treatment choice), and criterion 6 (balanced and unbiased) scored low, and criterion 9 (how each treatment works), criterion 2 (achieves its aims), and criterion 1 (provides clear aim) scored high. This trend was consistent with that of previous studies [34-36]. Unlike previous studies that compared DISCERN scores by the type of website operator (academic, government, or commercial agencies), only private clinic

XSL•FO
RenderX

websites that prescribe GLP-1RAs for off-label use were assessed in this study. Therefore, the DISCERN score was significantly lower in this study, possibly because the purpose of the websites was to sell GLP-1RAs and not to provide accurate information about obesity or obesity treatment to the public. In other words, these websites were advertisements designed to appeal to consumers about the effectiveness of GLP-1RAs for weight loss and not to draw their attention to other treatment options or shared decision-making with other health care providers or family members. Nevertheless, a previous study on consumer judgments of the reliability of web-based health information showed that website owners' authority had a positive impact on reliability and credibility [51]. Accordingly, if medical institutions representing specialties such as internal medicine were the owners of these websites, many consumers might trust the content of the websites. In summary, the problems with the information on medical institution websites identified by this study are not only unscientific and in violation of medical advertising guidelines but are also problematic from the perspective of public health and health promotion.

In this study, we observed that there were few public institution websites providing information on obesity treatment for consumers. To the best of our knowledge, no study has comprehensively compared the overall quality of web-based medical information between Japan and other countries. However, a previous study comparing palliative care websites in the United States and Japan reported that Japanese websites were of lower quality than the American ones [52]. A study focusing on the quality of web-based medical information within Japan has generally reported low quality [53]. This issue may be influenced by differences in the medical advertising regulations between Japan and other countries. Future studies should aim to provide a more comprehensive and international assessment of the quality of medical information.

Practical Implications

The results of this study have several implications. Regarding future studies, researchers should assess the reliability of media content, including medical institutions' websites that prescribe drugs for off-label use to contribute to better control of prescription drug misuse in the future. Many studies have investigated the consumer risk factors and preventive interventions for drug misuse. For example, a study investigating the misuse of over-the-counter drugs showed that health literacy, low education level, and misunderstanding of over-the-counter medicine instructions were correlated with the incidence of misuse [5]. Another study showed that the misuse of prescription stimulants for weight loss was positively associated with being female, lacking insurance, and having a history of lifetime misuse but not past-year misuse of prescription sedatives or tranquilizers [54]. Limited health literacy has been reported to be correlated with overweight/obesity in the first place [55], greater difficulty in the medication compliance, and higher mortality [56]. Thus, efforts to enhance consumer health literacy may be equally important as well as the content reliability of medical institutions' websites. In particular, studies on the risk of GLP-1RA misuse or off-label use are scarce and should be investigated.

Regarding future practices, this study suggests that governments should continuously monitor and provide corrective actions on websites, including unscientific content, to promote drug sales. A previous study in Japan showed that the number and volume of newspaper advertisements for dietary supplements that might mislead consumers decreased following the collective action of Japan's Consumer Affairs Agency [57]. Therefore, government monitoring and regulations are also important to address misleading information provided by medical institutions concerning prescription drug misuse. Furthermore, there are no uniform international guidelines for medical advertising. The permissibility of over-the-counter drug advertisements and advertisements for medical services differs from country to country. However, with the recent proliferation of social networking service platforms, misinformation or information with commercial intent about medicine can spread around the world in a short period of time. Therefore, it is also important to establish international medical advertising guidelines. Second, medical institutions and health care providers should recognize the current problem of GLP-1RA misuse and provide accurate information on the internet and through in-person explanations to individuals who are interested in weight loss. Additionally, medical institutions, academic organizations, and ministries should improve their websites with a focus on search engine optimization to ensure that high-quality medical information appears at the top of search engine results [58]. Internet-based companies, including Google, should be encouraged to prioritize the display of high-quality medical information. Third, medical institutions, health care providers, academic institutions, and ministries should inform citizens that current information on GLP-1RAs on websites is unreliable and that there is no miracle drug that is risk-free and absolutely effective. If individuals without obesity wish to lose weight, they should consult a trusted specialist such as a primary care physician to understand the risks and proceed with the therapy best suited for them.

Limitations and Strengths

This study has several limitations. First, this study only analyzed Japanese-language websites. The approval status of GLP-1RAs and medical advertising regulations differ from country to country. Thus, the generalizability of the results of this study to other countries is limited. Second, this study assesses the search results for web browsers and websites when viewed with web browsers. The results and website display contents may differ when using a smartphone. The results may have been affected by websites and search algorithms that were optimized for smartphones. Third, this study only used the DISCERN instrument for quality assessment. Therefore, this study was unable to discuss the reliability items other than those evaluated by DISCERN. There are multiple other quality assessment indicators such as the Health on the Net Foundation Code of Conduct certification and the Journal of the American Medical Association (JAMA) benchmark criteria. However, the Health on the Net Foundation Code of Conduct is not well known in Japan, and few websites have obtained it; therefore, it was not used in this study. The JAMA benchmark has only 4 criteria. Since most of the medical institutions of this study were private clinics, it was assumed that the 2 criteria of JAMA benchmark, authorship and attribution, would not differ. Additionally,

XSL•FO
RenderX

previous studies that evaluated web-based health information reported a positive correlation between JAMA benchmark scores and DISCERN scores [59,60]. Therefore, only DISCERN was used in this study. Additionally, the DISCERN instrument was developed to enable patients to assess the quality of written information related to treatment choices and to promote the creation of high-quality health information for documents intended for patients. Thus, assessing quality in terms of medical accuracy is not the primary focus of the DISCERN instrument. However, in this study, 2 internal medicine physician coders assessed the medical accuracy when evaluating questions 6, 9, and 12 of the DISCERN instrument, thereby reflecting medical accuracy in the DISCERN score. Fourth, this study relies on a single reviewer to screen eligible websites, which raises the possibility that some potentially eligible websites may have been excluded from the analyzed dataset. Fifth, the intercoder reliability between the 2 coders was ensured to be acceptable; however, coder bias may not have been completely eliminated. Finally, the impact of the websites analyzed in this study on viewers' cognition and behavior was outside the scope of this study; future studies should examine this aspect.

Despite these limitations, to the best of our knowledge, this is the first study on the quality of information on the websites of medical institutions that prescribe GLP-1RAs for weight loss in Japan. This study provides important implications for future studies and practice on better communication to decrease prescription drug misuse of GLP-1RAs, as discussed above. The quality of the websites of medical institutions that prescribe GLP-1RAs for weight loss was very low, and many of them did not comply with the medical advertising guidelines in Japan. This study shows that many websites, including exaggerated advertisements, may promote the misconception that exercise and dietary therapy are unnecessary. This study also shows that many websites provide biased information that may lead consumers to use GLP-1RAs without presenting alternative weight loss methods other than GLP-1RAs or the option of shared decision-making with health care providers or family members. Consumers and citizens who read such information may misuse GLP-1RAs because of insufficient understanding of their side effect risks and proper usage. Public institutions must monitor and regulate advertising content that violates guidelines and misleads individuals. Health care providers need to inform consumers about the risks and proper usage of GLP-1RAs and provide accurate information related to weight loss. Beneficial medications developed for people's health should not pose health risks because of the inappropriate information provided by health care providers.

Acknowledgments

This work was supported by the Japan Society for the Promotion of Science KAKENHI (23K09595).

Data Availability

The data presented in this study are available upon reasonable request to the corresponding author.

Authors' Contributions

RO and TO conceptualized this study and performed the methodology. RO and EF performed the formal analysis. RO wrote the original draft. HO acquired the funds. TK supervised this study. All authors contributed to reviewing and editing the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
DISCERN instrument.
[DOC File , 29 KB - formative_v9i1e68792_app1.doc]

References

1. Casati A, Sedefov R, Pfeiffer-Gerschel T. Misuse of medicines in the European Union: a systematic review of the literature. Eur Addict Res 2012;18(5):228-245 [FREE Full text] [doi: 10.1159/000337028] [Medline: 22572594]
2. Schifano F, Chiappini S, Miuli A, Mosca A, Santovito MC, Corkery JM, et al. Focus on over-the-counter drugs? misuse: A systematic review on antihistamines, cough medicines, and decongestants. Front Psychiatry 2021;12:657397 [FREE Full text] [doi: 10.3389/fpsyt.2021.657397] [Medline: 34025478]
3. Martins SS, Sampson L, Cerdá M, Galea S. Worldwide prevalence and trends in unintentional drug overdose: A systematic review of the literature. Am J Public Health 2015 Nov;105(11):e29-e49. [doi: 10.2105/AJPH.2015.302843] [Medline: 26451760]
4. Smith SM, Dart RC, Katz NP, Paillard F, Adams EH, Comer SD, et al. Classification and definition of misuse, abuse, and related events in clinical trials: ACTTION systematic review and recommendations. Pain 2013 Nov;154(11):2287-2296 [FREE Full text] [doi: 10.1016/j.pain.2013.05.053] [Medline: 23792283]

XSL•FO
RenderX

5. Algarni M, Hadi MA, Yahyouche A, Mahmood S, Jalal Z. A mixed-methods systematic review of the prevalence, reasons, associated harms and risk-reduction interventions of over-the-counter (OTC) medicines misuse, abuse and dependence in adults. J Pharm Policy Pract 2021 Sep 13;14(1):76 [FREE Full text] [doi: 10.1186/s40545-021-00350-7] [Medline: 34517925]
6. Hashimoto T, Kaneda Y, Ozaki A, Hori A, Tsuchiya T. Eleven-year trend of drug and chemical substance overdose at a local emergency hospital in Japan. Cureus 2022 Dec;14(12):e32475 [FREE Full text] [doi: 10.7759/cureus.32475] [Medline: 36644086]
7. Matsumoto T. Distributed research report: survey of drug-related psychiatric disorders in psychiatric hospitals in Japan. Department of Drug Dependence Research.: National Institute of Mental Health; 2020. URL: https://www.ncnp.go.jp/nimh/yakubutsu/report/pdf/J_NMHS_2020.pdf [accessed 2024-12-22]
8. Chiappini S, Vickers-Smith R, Harris D, Papanti Pelletier GD, Corkery JM, Guirguis A, et al. Is there a risk for semaglutide misuse? focus on the Food and Drug Administration's FDA Adverse Events Reporting System (FAERS) pharmacovigilance dataset. Pharmaceuticals (Basel) 2023 Jul 11;16(7):994 [FREE Full text] [doi: 10.3390/ph16070994] [Medline: 37513906]
9. Butuca A, Dobrea CM, Arseniu AM, Frum A, Chis AA, Rus LL, et al. An assessment of semaglutide safety based on real world data: from popularity to spontaneous reporting in Eudravigilance database. Biomedicines 2024 May 18;12(5):1124 [FREE Full text] [doi: 10.3390/biomedicines12051124] [Medline: 38791086]
10. Mahase E. Semaglutide: FDA publishes warning amid reports of patients overdosing. BMJ 2024 Jul 30;386:q1696. [doi: 10.1136/bmj.q1696] [Medline: 39079728]
11. Han SH, Safeek R, Ockerman K, Trieu N, Mars P, Klenke A, et al. Public interest in the off-label use of glucagon-like peptide 1 agonists (Ozempic) for cosmetic weight loss: a Google trends analysis. Aesthet Surg J 2023 Dec 14;44(1):60-67. [doi: 10.1093/asj/sjad211] [Medline: 37402640]
12. FDA approves new drug treatment for chronic weight management, first since 2014. USFDA. URL: https://www.fda.gov/news-events/press-announcements/fda-approves-new-drug-treatment-chronic-weight-management-first-2014 [accessed 2024-09-08]
13. Suran M. As Ozempic's popularity soars, here's what to know about semaglutide and weight loss. JAMA 2023 May 16;329(19):1627-1629. [doi: 10.1001/jama.2023.2438] [Medline: 37099334]
14. Goodman B. Poison centers see nearly 1,500% increase in calls related to injected weight-loss drugs as people accidentally overdose. CNN Health. URL: https://www.cnn.com/2023/12/13/health/semaglutide-overdoses-wellness [accessed 2024-10-21]
15. Brafman N. Ozempic: French authorities issue alert for anti-diabetic drug misused for weight loss. Le Monde 2023 Mar 2. URL: https://www.lemonde.fr/en/health/article/2023/03/02/ozempic-french-authorities-issue-alert-for-anti-diabetic-drug-misused-for-weight-loss_6017913_14.html [accessed 2024-10-21]
16. Mahase E. GLP-1 shortages will not resolve this year, EMA warns, amid concern over off-label use. BMJ 2024 Jun 28;385:q1448. [doi: 10.1136/bmj.q1448] [Medline: 38942431]
17. Hathaway JT, Shah MP, Hathaway DB, Zekavat SM, Krasniqi D, Gittinger JW, et al. Risk of nonarteritic anterior ischemic optic neuropathy in patients prescribed semaglutide. JAMA Ophthalmol 2024 Aug 01;142(8):732-739. [doi: 10.1001/jamaophthalmol.2024.2296] [Medline: 38958939]
18. Sodhi M, Rezaeianzadeh R, Kezouh A, Etminan M. Risk of gastrointestinal adverse events associated with glucagon-like peptide-1 receptor agonists for weight loss. JAMA 2023 Nov 14;330(18):1795-1797 [FREE Full text] [doi: 10.1001/jama.2023.19574] [Medline: 37796527]
19. Recommendations of the executive steering group on shortages and safety of medicinal products on shortage of glucagon-like peptide-1 (GLP-1) receptor agonists. European Medicines Agency. URL: https://www.quotidianosanita.it/allegati/allegato1719409581.pdf [accessed 2024-09-08]
20. Notes on the establishment of guidelines for the promotion of appropriate use of semaglutide (genetic recombination) preparation for the indication of obesity document in Japanese. Ministry of Health, Labour and Welfare, Japan. URL: https://www.pmda.go.jp/files/000265450.pdf [accessed 2024-09-08]
21. Hakariya H, Ohnishi M, Tanimoto T. Japan initiates market authorization of weight-loss drug semaglutide under universal health coverage, but with stringent prescription restrictions. Diabetes Obes Metab 2024 Aug;26(8):3006-3008. [doi: 10.1111/dom.15638] [Medline: 38747140]
22. Request for cooperation due to tight inventory of GLP-1 receptor agonist (Part 2) document in Japanese. Ministry of Health, Labour, and Welfare, Japan. URL: https://www.mhlw.go.jp/content/001165743.pdf [accessed 2024-11-09]
23. Position on off-label use of GLP-1 receptor agonists and GIP/GLP-1 receptor agonists document in Japanese. The Japan Diabetes Society. URL: https://www.jds.or.jp/uploads/files/document/info/jds_statement_GLP-1.pdf [accessed 2024-10-28]
24. Statement on the safe and appropriate use of obesity drugs document in Japanese. Japan Society for the Study of Obesity. URL: http://www.jasso.or.jp/data/Introduction/pdf/academic-information_statement_20231127.pdf [accessed 2024-09-08]
25. Telemedicine trouble for slimming purposes: Subscription trouble? where diabetes drugs are prescribed off-label for several months for the purpose of weight loss document in Japanese. National Consumer Affairs Center of Japan. URL: https://www.kokusen.go.jp/news/data/n-20231220_1.html [accessed 2024-09-08]

XSL•FO
RenderX

26. Before you get a cosmetic procedure, recheck to make sure you've been properly briefed and understand it again document in Japanese. Consumer Affairs Agency, Government of Japan. URL: https://www.caa.go.jp/policies/policy/consumer_policy/information/information_002/assets/information_002_240202_02.pdf [accessed 2024-09-08]
27. Medical Care Act (Japanese Law Translation System). Ministry of Justice, Japan. 1948. URL: https://www.japaneselawtranslation.go.jp/ja/laws/view/4006 [accessed 2024-09-08]
28. Regulations on advertisements of hospitals under the Medical Care Act (in Japanese). Ministry of Health, Labour and Welfare, Japan. URL: https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/kenkou_iryou/iryou/kokokukisei/index.html [accessed 2024-09-08]
29. Charnock D, Shepperd S, Needham G, Gann R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. J Epidemiol Community Health 1999 Feb;53(2):105-111 [FREE Full text] [doi: 10.1136/jech.53.2.105] [Medline: 10396471]
30. Shan Y, Xing Z, Dong Z, Ji M, Wang D, Cao X. Translating and adapting the DISCERN instrument into a simplified Chinese version and validating its reliability: Development and usability study. J Med Internet Res 2023 Feb 02;25:e40733 [FREE Full text] [doi: 10.2196/40733] [Medline: 36729573]
31. Montoya A, Llopis N, Gilaberte I. Validation of the translation of an instrument to measure reliability of written information on treatment choices: a study on attention deficit/hyperactivity disorder (ADHD). Educ Health (Abingdon) 2011 Dec;24(3):577. [Medline: 22267353]
32. Logullo P, Torloni MR, de O C Latorraca C, Riera R. The Brazilian Portuguese version of the DISCERN instrument: Translation procedures and psychometric properties. Value Health Reg Issues 2019 Dec;20:172-179 [FREE Full text] [doi: 10.1016/j.vhri.2019.09.001] [Medline: 31622803]
33. Batchelor JM, Ohya Y. Use of the DISCERN instrument by patients and health professionals to assess information resources on treatments for asthma and atopic dermatitis. Allergol Int 2009 Mar;58(1):141-145 [FREE Full text] [doi: 10.2332/allergolint.08-SC-0022] [Medline: 19153541]
34. Killip SC, Kwong NKR, MacDermid JC, Fletcher AJ, Carleton NR. The quality, readability, completeness, and accuracy of PTSD websites for firefighters. Int J Environ Res Public Health 2020 Oct 19;17(20):7629 [FREE Full text] [doi: 10.3390/ijerph17207629] [Medline: 33086772]
35. Li Y, Zhou X, Zhou Y, Mao F, Shen S, Lin Y, et al. Evaluation of the quality and readability of online information about breast cancer in China. Patient Educ Couns 2021 Apr;104(4):858-864 [FREE Full text] [doi: 10.1016/j.pec.2020.09.012] [Medline: 32988687]
36. Tan DJY, Ko TK, Fan KS. The readability and quality of web-based patient information on nasopharyngeal carcinoma: quantitative content analysis. JMIR Form Res 2023 Nov 27;7:e47762 [FREE Full text] [doi: 10.2196/47762] [Medline: 38010802]
37. Patel AA, Joshi C, Varghese J, Hassan AM, Janis JE, Galiano RD. Do websites serve our patients well? a comparative analysis of online information on cosmetic injectables. Plast Reconstr Surg 2022 Apr 01;149(4):655e-668e. [doi: 10.1097/PRS.0000000000008921] [Medline: 35139065]
38. Daraz L, Morrow AS, Ponce OJ, Beuschel B, Farah MH, Katabi A, et al. Can patients trust online health information? a meta-narrative systematic review addressing the quality of health information on the internet. J Gen Intern Med 2019 Sep;34(9):1884-1891 [FREE Full text] [doi: 10.1007/s11606-019-05109-0] [Medline: 31228051]
39. Weaver JB, Mays D, Weaver SS, Hopkins GL, Eroglu D, Bernhardt JM. Health information-seeking behaviors, health indicators, and health risks. Am J Public Health 2010 Aug;100(8):1520-1525. [doi: 10.2105/AJPH.2009.180521] [Medline: 20558794]
40. Telecommunications usage trends survey (in Japanese). Ministry of Internal Affairs and Communications. URL: https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/r05/html/datashu.html#f00281 [accessed 2024-10-12]
41. Montemurro P, Porcnik A, Hedén P, Otte M. The influence of social media and easily accessible online information on the aesthetic plastic surgery practice: literature review and our own experience. Aesthetic Plast Surg 2015 Apr;39(2):270-277. [doi: 10.1007/s00266-015-0454-3] [Medline: 25697277]
42. Google Japan. URL: https://www.google.co.jp/ [accessed 2024-12-23]
43. Yahoo! Japan. URL: https://www.yahoo.co.jp/ [accessed 2024-12-23]
44. Search engine share in Japan. Ministry of Internal Affairs and Communications. URL: https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/r05/html/datashu.html#f00228 [accessed 2024-09-08]
45. Eysenbach G, Köhler C. How do consumers search for and appraise health information on the world wide web? Qualitative study using focus groups, usability tests, and in-depth interviews. BMJ 2002 Mar 09;324(7337):573-577 [FREE Full text] [doi: 10.1136/bmj.324.7337.573] [Medline: 11884321]
46. Ved R, Cobbold N, Igbagiri K, Willis M, Leach P, Zaben M. Online patient information on Vagus Nerve Stimulation: How reliable is it for facilitating shared decision making? Seizure 2017 Aug;50:125-129 [FREE Full text] [doi: 10.1016/j.seizure.2017.06.009] [Medline: 28651125]
47. Trouble with online medical treatment for slimming purposes, etc document in Japanese. National Consumer Affairs Center, Japan. URL: https://www.kokusen.go.jp/pdf/n-20231220_1.pdf [accessed 2024-10-01]

XSL•FO
RenderX

48. Jastreboff AM, Aronne LJ, Ahmad NN, Wharton S, Connery L, Alves B, SURMOUNT-1 Investigators. Tirzepatide once weekly for the treatment of obesity. N Engl J Med 2022 Jul 21;387(3):205-216. [doi: 10.1056/NEJMoa2206038] [Medline: 35658024]
49. Hartmann C, Keller C, Siegrist M. Compensatory beliefs, nutrition knowledge and eating styles of users and non-users of meal replacement products. Appetite 2016 Oct 01;105:775-781. [doi: 10.1016/j.appet.2016.07.013] [Medline: 27421243]
50. Iye R, Okuhara T, Okada H, Goto E, Furukawa E, Kiuchi T. The effect of exposure to "exemption" video advertisements for functional foods: a randomized control study in Japan. Healthcare (Basel) 2022 Feb 11;10(2):345 [FREE Full text] [doi: 10.3390/healthcare10020345] [Medline: 35206959]
51. Zhang Y, Kim Y. Consumers' evaluation of web-based health information quality: Meta-analysis. J Med Internet Res 2022 Apr 28;24(4):e36463 [FREE Full text] [doi: 10.2196/36463] [Medline: 35482390]
52. Tanabe K, Fujiwara K, Ogura H, Yasuda H, Goto N, Ohtsu F. Quality of web information about palliative care on websites from the United States and Japan: comparative evaluation study. Interact J Med Res 2018 Apr 03;7(1):e7 [FREE Full text] [doi: 10.2196/ijmr.9574] [Medline: 29615388]
53. Yamamura A, Watanabe S, Yamaguchi S, Iwata K, Kimura S, Mikami Y, et al. Readability and quality of online patient resources regarding knee osteoarthritis and lumbar spinal stenosis in Japan. J Orthop Sci 2024 Sep;29(5):1313-1318. [doi: 10.1016/j.jos.2023.08.003] [Medline: 37599135]
54. Compton WM, Han B, Blanco C, Johnson K, Jones CM. Prevalence and correlates of prescription stimulant use, misuse, use disorders, and motivations for misuse among adults in the United States. Am J Psychiatry 2018 Aug 01;175(8):741-755 [FREE Full text] [doi: 10.1176/appi.ajp.2018.17091048] [Medline: 29656665]
55. Michou M, Panagiotakos DB, Costarelli V. Low health literacy and excess body weight: a systematic review. Cent Eur J Public Health 2018 Sep;26(3):234-241 [FREE Full text] [doi: 10.21101/cejph.a5172] [Medline: 30419628]
56. Fan Z, Yang Y, Zhang F. Association between health literacy and mortality: a systematic review and meta-analysis. Arch Public Health 2021 Jul 01;79(1):119 [FREE Full text] [doi: 10.1186/s13690-021-00648-7] [Medline: 34210353]
57. Okuhara T, Yokota R, Shirabe R, Iye R, Okada H, Kiuchi T, et al. Japanese newspaper advertisements for dietary supplements before and after COVID-19: a content analysis. BMJ Open 2021 Nov 23;11(11):e050898 [FREE Full text] [doi: 10.1136/bmjopen-2021-050898] [Medline: 34815281]
58. Modave F, Shokar NK, Peñaranda E, Nguyen N. Analysis of the accuracy of weight loss information search engine results on the internet. Am J Public Health 2014 Oct;104(10):1971-1978. [doi: 10.2105/AJPH.2014.302070] [Medline: 25122030]
59. Foster B, Grampp S, Ozdag Y, Tang A, Vazquez F, Mercuri J. Assessing the quality, content, and readability of online patient resources on viscosupplementation. Journal of Orthopaedic Experience & Innovation Charter Services New York d/b/a Journal of Orthopaedic Experience and Innovation; 2024 Oct 4 2024 Oct 4;5(2):1-8. [doi: 10.60118/001c.120353]
60. Ozduran E, Büyükçoban S. Evaluating the readability, quality and reliability of online patient education materials on post-covid pain. PeerJ 2022;10:e13686 [FREE Full text] [doi: 10.7717/peerj.13686] [Medline: 35880220]

Abbreviations

FDA: Food and Drug Administration
GLP-1: glucagon-like peptide-1
GLP-1RA: glucagon-like peptide-1 receptor agonist
JAMA: Journal of the American Medical Association

Edited by A Mavragani; submitted 13.11.24; peer-reviewed by E Gray, N Misra, M Agbede; comments to author 30.11.24; revised version received 08.12.24; accepted 16.12.24; published 01.01.25.

Please cite as:
Oyama R, Okuhara T, Furukawa E, Okada H, Kiuchi T
Quality Assessment of Medical Institutions' Websites Regarding Prescription Drug Misuse of Glucagon-Like Peptide-1 Receptor Agonists by Off-Label Use for Weight Loss: Website Evaluation Study
JMIR Form Res 2025;9:e68792
URL: https://formative.jmir.org/2025/1/e68792
doi:10.2196/68792
PMID:

©Rie Oyama, Tsuyoshi Okuhara, Emi Furukawa, Hiroko Okada, Takahiro Kiuchi. Originally published in JMIR Formative Research (https://formative.jmir.org), 01.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The

XSL•FO
RenderX

complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Public Health Discussions on Social Media: Evaluating Automated Sentiment Analysis Methods

Lisa M Gandy[1*], PhD; Lana V Ivanitskaya[2*], PhD; Leeza L Bacon[3*], DHA; Rodina Bizri-Baryak[2*], DHA

[1]Department of Computer Science, College of Sciences and Liberal Arts, Kettering University, Flint, MI, United States

[2]Department of Health Administration, The College of Health Professions, Central Michigan University, Mt Pleasant, MI, United States

[3]Department of Healthcare Management, Northwood University, Midland, MI, United States

[*]all authors contributed equally

Corresponding Author:
Lisa M Gandy, PhD
Department of Computer Science
College of Sciences and Liberal Arts
Kettering University
1700 University Ave
2300 AB
Flint, MI, 48504
United States
Phone: 1 9898547001
Email: lgandy@kettering.edu

Abstract

Background: Sentiment analysis is one of the most widely used methods for mining and examining text. Social media researchers need guidance on choosing between manual and automated sentiment analysis methods.

Objective: Popular sentiment analysis tools based on natural language processing (NLP; VADER [Valence Aware Dictionary for Sentiment Reasoning], TEXT2DATA [T2D], and Linguistic Inquiry and Word Count [LIWC-22]), and a large language model (ChatGPT 4.0) were compared with manually coded sentiment scores, as applied to the analysis of YouTube comments on videos discussing the opioid epidemic. Sentiment analysis methods were also examined regarding ease of programming, monetary cost, and other practical considerations.

Methods: Evaluation methods included descriptive statistics, receiver operating characteristic (ROC) curve analysis, confusion matrices, Cohen κ, accuracy, specificity, precision, sensitivity (recall), F_1-score harmonic mean, and the Matthews correlation coefficient. An inductive, iterative approach to content analysis of the data was used to obtain manual sentiment codes.

Results: A subset of comments were analyzed by a second coder, producing good agreement between the 2 coders' judgments (κ=0.734). YouTube social media about the opioid crisis had many more negative comments (4286/4871, 88%) than positive comments (79/662, 12%), making it possible to evaluate the performance of sentiment analysis models in an unbalanced dataset. The tone summary measure from LIWC-22 performed better than other tools for estimating the prevalence of negative versus positive sentiment. According to the ROC curve analysis, VADER was best at classifying manually coded negative comments. A comparison of Cohen κ values indicated that NLP tools (VADER, followed by LIWC's tone and T2D) showed only fair agreement with manual coding. In contrast, ChatGPT 4.0 had poor agreement and failed to generate binary sentiment scores in 2 out of 3 attempts. Variations in accuracy, specificity, precision, sensitivity, F_1-score, and MCC did not reveal a single superior model. F_1-score harmonic means were 0.34-0.38 (SD 0.02) for NLP tools and very low (0.13) for ChatGPT 4.0. None of the MCCs reached a strong correlation level.

Conclusions: Researchers studying negative emotions, public worries, or dissatisfaction with social media face unique challenges in selecting models suitable for unbalanced datasets. We recommend VADER, the only cost-free tool we evaluated, due to its excellent discrimination, which can be further improved when the comments are at least 100 characters long. If estimating the prevalence of negative comments in an unbalanced dataset is important, we recommend the tone summary measure from LIWC-22. Researchers using T2D must know that it may only score some data and, compared with other methods, be more time-consuming and cost-prohibitive. A general-purpose large language model, ChatGPT 4.0, has yet to surpass the performance of NLP models, at least for unbalanced datasets with highly prevalent (7:1) negative comments.

XSL•FO
RenderX

(JMIR Form Res 2025;9:e57395) doi:10.2196/57395

KEYWORDS

ChatGPT; VADER; valence aware dictionary for sentiment reasoning; LIWC-22; machine learning; social media; sentiment analysis; public health; population health; opioids; drugs; pharmacotherapy; pharmaceuticals; medications; YouTube

Introduction

The Pew Research Center [1] reports that as of 2021, 72% of Americans used social media. On a global scale, as of 2019, social media platforms were used by 1 in 3 people worldwide and by more than two-thirds of all internet users [2]. With social media users rapidly increasing, user-generated content has grown exponentially. Generated by a broad swath of global citizens, the data provides insights into a wide array of human experiences, for example, the effects of the opioid crisis on physical, mental, and social well-being.

Analysis of the opioid epidemic revealed the struggles of opioid victims, their families, and communities. This information adds value to health policy analysis.

This study examines the use of YouTube, a unique social media platform. It is a diverse medium because its primary purpose is content sharing and educating its users about relevant topics [3]. Users seek out content of interest and, comment and react. When examining health care crises and the relay of medical information, sentiment analysis may be used to gauge the response to the stimulus. By analyzing the sentiments expressed in comments, researchers can perform qualitative content analyses to explore how these reactions influence reputation, potentially affect individuals, and impact the communities involved. Furthermore, understanding the sentiment toward health-related content or issues like human suffering and societal concerns can aid policy makers in developing strategies to enhance public health.

Sentiment analysis is the most popular artificial intelligence used to mine and examine all text types in various fields of study. Sentiment analysis is a computational method that extracts sentiment from a text. Some sentiment analysis methods use rule-based lexicons such as Linguistic Inquiry and Word Count (LIWC)-22 [4]. Other sentiment analysis methods use traditional machine learning approaches such as Support Vector Machines [5] and Naive Bayes classification [6], while others use deep learning [7]. Sometimes, a hybrid approach is used between 2 or more sentiment analysis methods [8].

This study compares 3 popular sentiment analysis methods on social media data: Valence Aware Dictionary for Sentiment Reasoning (VADER), TEXT2DATA (T2D), and LIWC, and also uses the ChatGPT large language model (LLM) for sentiment analysis. A total of 2 methods (VADER and LIWC) were picked due to their previous validation and use in many published studies by scholars from different disciplines; other methods were chosen because of their user-friendly interface and no requirement of prerequisite programming skills (LIWC-22 and T2D). ChatGPT was chosen as LLM represent a game-changing technological leap in natural language processing (NLP), including sentiment analysis.

VADER [9] is a rule and lexicon-based sentiment analysis tool that, when analyzing text, returns a numeric valence (polarity) score between –1 (extremely negative) and +1 (extremely positive). The VADER lexicon is primarily built using pre-existing lexicons, but these lexicons were extended to include emoticons, acronyms, and slang commonly used in social media. Each lexical feature is assigned a score; then, the score is shifted based on the presence of punctuation marks, capital letters, and negations. The lexical scores are then averaged. VADER has been validated on multiple data types, including product and movie reviews, and has been used as a quality benchmark in numerous studies [9-11]. It has been used for sentiment analysis in a wide variety of areas, such as customer reviews and opinion mining [12,13], political discourse analysis [14,15], and mental health studies [16,17]. Although VADER has been used extensively, concerns about VADER's sensitivity to text length have been expressed [18,19]. For example, as each word in a text is assigned a polarity score, words with strong sentiment can unduly influence the overall sentiment of a short text. There are proposed methods to mitigate this phenomenon [9], such as considering the text of an entire document about a sentence. Another possible issue is that VADER is primarily used as a Python-based program, and users must be moderately proficient in Python programming to use its full capabilities.

T2D [20] is a Microsoft Excel or Google Sheets add-in with a sentiment analysis application programming interface (API). It classifies text into 5 categories: very negative, negative, neutral, positive, and very positive. The T2D website contains scant information about the corpus or the methodology used for sentiment classification. The website states that the API is "based on an NLP engine" and that the "system also contains specially prepared classification models for Twitter (rebranded as X) and other social media content, trained on billions of manually verified entries" [20]. It is a paid subscription model, but users can make 1000 API calls a month for free. The next tier is 10,000 API calls for US $27 a month, then a variety of tiers exist, with the highest tier being Enterprise at US $351 a month and Unlimited API calls. As a Microsoft Excel add-in and a user-friendly tool for those familiar with Microsoft Excel, T2D has been applied to studies that examined tourist experiences [21], compared Twitter comments with polling results [22], and analyzed digital mental health interventions [23].

LIWC-22 is a dictionary-based sentiment analysis tool with a representative score in over 100 categories. The score in each category indicates the percentage of words in the text corresponding to the particular category and, therefore, can range from 0% to 100%. Many LIWC-22 categories are organized in a hierarchical structure. The same word may be categorized into multiple categories. For instance, "celebrate" is in both the positive emotion and achievement categories.

XSL•FO
RenderX

Each category is represented internally in LIWC-22 by a dictionary with words and emoticons related to that category. A complete list of LIWC-22 categories can be found in [24]. The creators of LIWC-22 selected a vast pool of words that represented a wide range of linguistic categories and psychological dimensions, such as emotions, cognitive processes, social terms, pronouns, prepositions, and other linguistic constructs. LIWC-22 is the fifth iteration of LIWC-22 (LIWC 2001, 2007, 2015, and 2019). To create LIWC, experts in linguistics and psychology first gathered a large corpus of texts and tagged each word into a linguistic category. The resulting dictionaries underwent validation studies and then passed through a refinement phase when researchers added new vocabulary and, as meanings shifted, modified existing word-category pairings. LIWC-22 is a subscription model. Currently, users pay US $129.95 for a 3-year subscription, and shorter-term subscriptions are offered at lower prices. LIWC-22 has both a web-based app and an app for download. The users upload their file (JSON, CSV, and EXCEL), choose the LIWC-22 dictionary they would like to use, and then select the label in their data that denotes the text to be classified. Users also have the option to choose all or specific LIWC-22 categories to be evaluated.

Similar to VADER and T2D, LIWC-22 has been used as a sentiment analysis tool in many studies, such as education [25], public discourse analysis [26], and brand perception [27] among others. Many categories in LIWC-22 could represent sentiment, including positive emotion, negative emotion, anger, sadness, and anxiety. LIWC-22 includes positive tone and negative tone dimensions but also includes another Tone variable that merges the 2 dimensions into a single summary variable. The higher the number, the more positive the tone; numbers below 50 are counted as negative tone. This paper uses the LIWC-22 tone composite (summary) measure as its dictionary was more extensive than other sentiment-based LIWC-22 categories. We will refer to it as LIWC tone.

Previous efforts to evaluate sentiment analysis tools did not compare LIWC tone, VADER, and T2D. For example, Boukes et al [28] compared LIWC, SentiStrength, Pattern, Polyglot, and DANEW on economic news in Danish. These sentiment analysis tools were chosen as they were off the shelf and supported the Danish language. Hartmann et al [29] compared LIWC-22 with a host of machine learning-based sentiment analysis methods (Support Vector Machines, Neural Networks, K Nearest Neighbors, and Random Forests) but did not include VADER or T2D and focus on the marking aspect of social media data.

Recent advances in NLP have led to the development of powerful language models, such as the GPT series, including ChatGPT (GPT–3.5 and GPT–4) [30]. These models, pretrained on vast amounts of text data, have demonstrated strong performance across tasks like language translation, text summarization, and question-answering [31]. A LLM ChatGPT, in particular, has shown promise in education, health care, reasoning, text generation, human-machine interaction, and scientific research [30,32]. Despite these opportunities, challenges and ethical concerns remain, particularly regarding accuracy. The accuracy of these models depends heavily on the quality, diversity, and complexity of the training data, as well as the quality of user input [30,31]. Previous research has highlighted the importance of developing higher-order thinking skills in education, but NLP systems may struggle with the nuances of human language, leading to potential errors [30-32].

We compare the sentiment scores produced by VADER, T2D, LIWC_tone, and ChatGPT 4.0 to manually created codes. The data source is YouTube comments on videos that discussed the opioid epidemic. This data source was chosen as it was easily accessible to the authors, but also because each of the sentiment analysis classifiers mentioned has been heavily used to classify social media data in the past. As social media (and the data it creates) grows, researchers will likely analyze it using the sentiment tools discussed in this paper.

In the remainder of the paper, we discuss the methods by which data was collected and coded and will then compare the results given by VADER, T2D, LIWC_tone, and ChatGPT 4.0 to human coding. This paper reports which methods give accurate sentiment scores and discusses other practical considerations when using these tools. We will end with a discussion of the results and considerations for future research.

Methods

Data Sources

To evaluate sentiment analysis methods, we used secondary data from a qualitative study about the opioid crisis in the United States [33], which included manually coded sentiment from YouTube comments. To collect YouTube comments (N=8761), the term “opioid” epidemic was searched on YouTube with a date range between January 1, 2017, and December 31, 2018. The majority of the videos collected were located on CNN’s (Cable News Network) YouTube channel and the Fox News YouTube Channel. Subsequently, videos were ranked by the number of views, and the 20 most watched videos by CNN (10 newscasts) and Fox News (10 newscasts) were kept for further analysis. As Google Trends indicates (Figure 1), the chosen dates coincide with a particularly high interest in the opioid epidemic. The comments for each video were downloaded using the Netlytic website. Comments were deidentified by deleting email addresses and assigning codes such as a1, a2, and so on, to track comments on comments. All other comment information was left unaltered.

XSL•FO
RenderX

Figure 1. Google Trends results for the search term "opioid epidemic." The search was constrained to dates between February 2015 and December 2019. The y-axis indicates percent popularity.

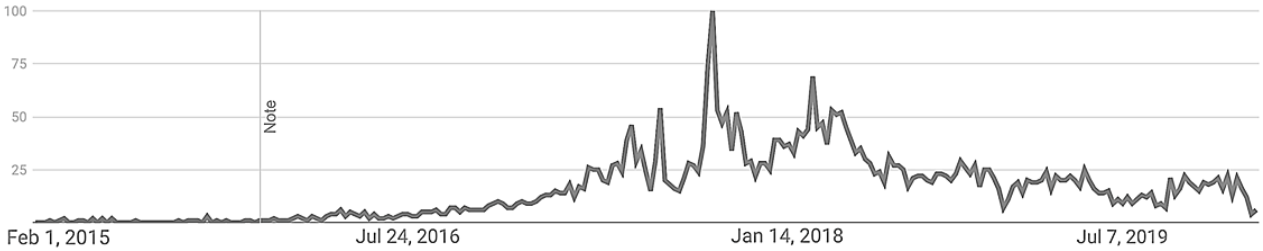

Sentiment Measures

An inductive, iterative approach to content analysis of the data was used to obtain manual sentiment codes. Bacon [33] created a codebook to determine the sentiment of comments as positive, neutral, or negative based on the commenter's attitude toward the video or another comment. The comments were sorted by time of post and video, then coded in the context of the broader discussion. Video transcripts also informed the coding. A subset of comments was analyzed by a second coder, producing good agreement between the 2 coders' judgments (κ=0.734; 95% CI 0.57-0.89; P<.001).

To begin testing automated sentiment tools, comments with manual codes indicating neutral or unclear sentiment were excluded from the data. The remaining comments were scored using VADER, T2D's Google Sheets add-in, LIWC, and ChatGPT 4.0. Specifically, the VADER compound score, T2D sentiment score, and LIWC-22 tone score (LIWC tone), a composite measure for positive and negative tone dimensions, were used. ChatGPT 4.0, VADER, and T2D were centered around 0 and ranged from –1 to 1. Negative values indicated a negative sentiment and positive values showed a positive sentiment. LIWC tone scores ranged from 1 to 100, representing percentiles based on standardized scores from large comparison corpora [34]. They are calculated using a dictionary with words, word stems, phrases, and select emoticons built for text analysis. LIWC tone overall mean for 15 diverse corpora was 47.81 (SD 26.39). Bacon's codebook was adapted to prompt the general-purpose LLM to generate positive and negative sentiment classification (Multimedia Appendix 1 for the ChatGPT 4.0 prompt). The classification was generated on the third try after two failed attempts.

Analyses

We performed the receiver operating characteristic (ROC) curve analysis and computed descriptive statistics, confusion matrices, Cohen , accuracy, specificity, precision, sensitivity (recall), F_1-score harmonic mean, and the Matthews correlation coefficient (MCC). Relying solely on the ROC curve without considering precision and negative predictive value can lead to a misleading assessment of a model's success [35]. Although widely used in machine learning, the F_1-score also has limitations; it can vary when positive and negative classes are exchanged, potentially distorting its interpretation [36]. In addition, the F_1-score does not account for correctly classified negative and positive samples, drawing criticism for diverging from more intuitive metrics like accuracy and losing significance when class labels are reversed [36]. In contrast, MCC provides a more balanced evaluation, achieving its highest values of –1 or +1 only when the classifier performs well across all 4 key rates of the confusion matrix: sensitivity, specificity, precision, and negative predictive value [35,36].

Finally, to reveal misclassification patterns by model, we analyzed the content of the comments marked as false positives and false negatives from both NLP and LLM models and compared them with true positives and true negatives identified through manual coding. We summarized possible reasons behind misclassifications and provided representative comments as illustrations.

Ethical Considerations

Data were collected from a social media platform (YouTube) where data are publicly available. However, all sentiment evaluation methods were performed at a macro scale and not at the user level. In addition, social media profile information is not shared in the data provided in this article.

Results

Manual coding placed 63.2% of comments into either positive or negative categories; the remainder were neutral, or their sentiment could not be ascertained. Only positively or negatively classified comments were used for further analyses (N=5533). Positive comments were much less common (79/662, 12%) than negative (4286/4,871, 88%) comments.

VADER, LIWC tone, and ChatGPT 4.0 were able to classify all comments. However, for unknown reasons, T2D sentiment scores could not be calculated for 514 comments, resulting in 9% of missing values. T2D is a "black box" system, and documentation has not been released. As shown in Figure 2, LIWC tone analysis most closely matched the high prevalence of negative comments (88%) in manually coded data: 82% of LIWC tone scores fell below the mean of 47.81. ChatGPT 4.0 overestimated negative sentiment, classifying 98% of the comments as negative. In comparison, only 56% of VADER and 66% of T2D scores were negatively scored. VADER score distribution had a mode around 2 and was the most continuous compared with T2D and LIWC. VADER assigned near-0 scores to 15% of comments, T2D had 1% of near-zero data but did not score 9% of comments, and LIWC tone scores were very unevenly distributed with 3 modes (at scale's endpoints and 23.23).

XSL•FO
RenderX

Figure 2. A comparison of Valence Aware Dictionary for Sentiment Reasoning, TEXT2DATA, and Linguistic Inquiry and Word Count tone score distributions for 5533 comments classified as either negative or positive using manual coding. *LIWC tone scores below 47.81 are considered negative. T2D: TEXT2DATA.

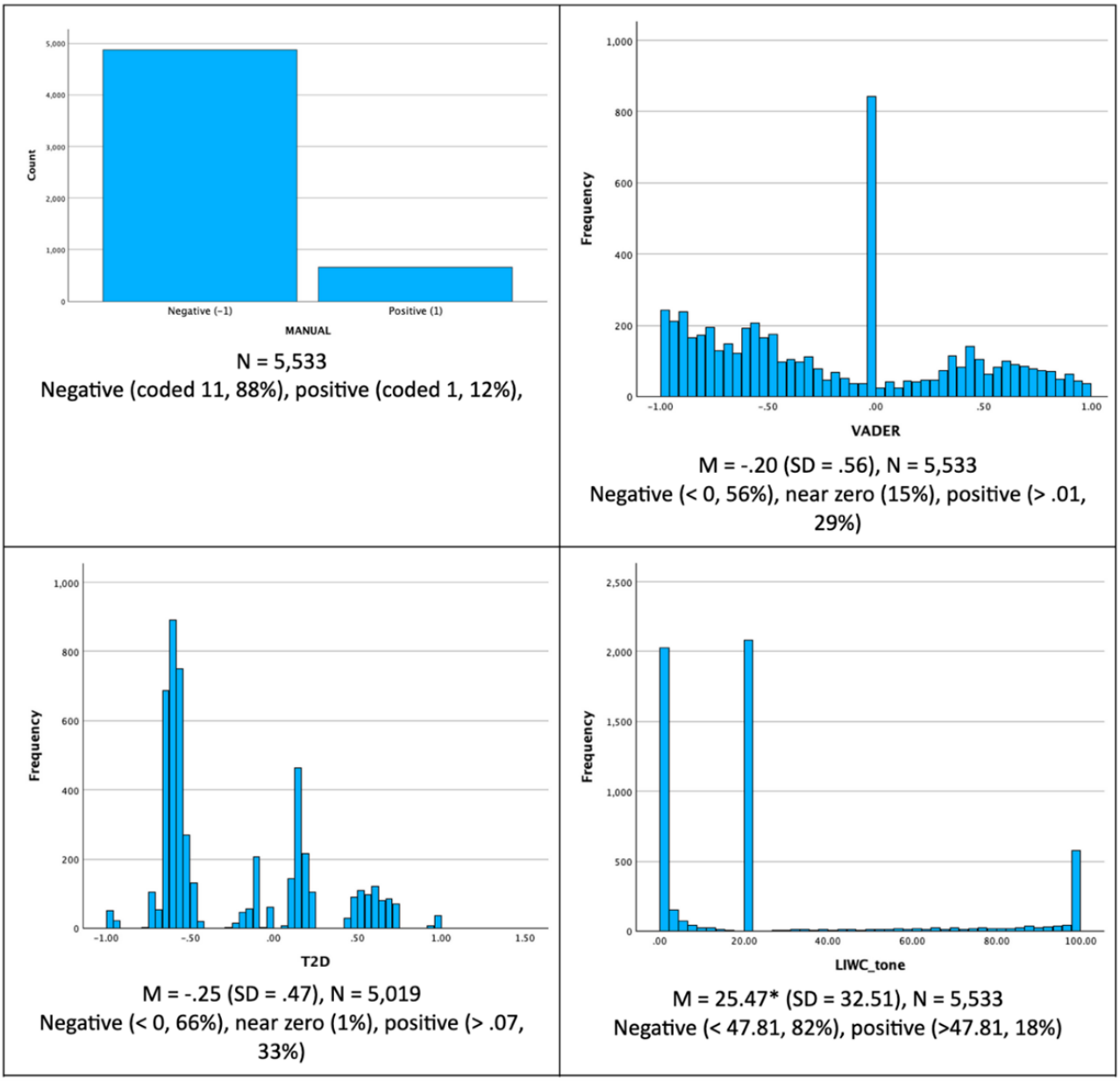

As an overall measure of discrimination, we used the ROC curve [37]. Discrimination is the ability of a measure to discern between social media comments that are manually coded as negative or positive. We were interested in evaluating continuously measured VADER, T2D, and LIWC tone scores as predictors of positive and negative sentiment in YouTube comments; manual coding is an outcome against which the 3 models were assessed. ROC curve analysis was not performed for ChatGPT 4.0 because it was only generated as a binary measure. The area under the ROC curve is a measure of the overall discriminatory ability of the binomial logistic regression model, that is, the ability of a chosen sentiment scoring method to classify comments into the 2 groups of our dichotomous dependent variable, a manually coded sentiment where negative is assigned a value –1 and positive is given a value of 1. ROC curve analysis is most suitable for balanced analyses. While we are interested in correctly classifying positives and negatives, our dataset is unbalanced due to the high prevalence of negative comments.

As shown in Figure 3, ROC curves with red lines above the blue straight reference line indicate discrimination; the further above the reference line, the better. The area under the ROC curve is equivalent to the concordance probability [38]. Figure 2 shows an excellent level of discrimination, according to Hosmer et al [39], for VADER's ability to classify manually coded negative comments (the area under the ROC curve was 0.800, 95% CI 0.78-0.82), followed by T2D and LIWC tone. T2D and LIWC tone demonstrated acceptable discrimination, with the areas under the curve of 0.770 and 0.747, respectively.

LIWC tone, according to Figure 4, performs the same or worse across the entire range of true positive rates (sensitivity), except at the short stretch of higher sensitivity rates where it is superior to T2D. Overall, VADER performs better than the 2 other

XSL•FO
RenderX

sentiment analysis systems. Next, we tested if VADER performed better for longer strings of text. Figure 4 compares VADER results or all comments, regardless of length, and for longer comments (>100 characters and >200 characters long). VADER indeed performed better when short comments were excluded. Confusion matrices are given in Tables 1-4.

Table 5 shows 7 measures of model performance compared with manual coding that are Cohen , accuracy, specificity, precision, sensitivity (recall), F_1-score, and MCC.

Figure 3. Receiver operating characteristic analyses: a comparison of Valence Aware Dictionary for Sentiment Reasoning, TEXT2DATA, and Linguistic Inquiry and Word Count tone scores' ability to classify comments that were manually coded as either negative (–1) or positive (1).

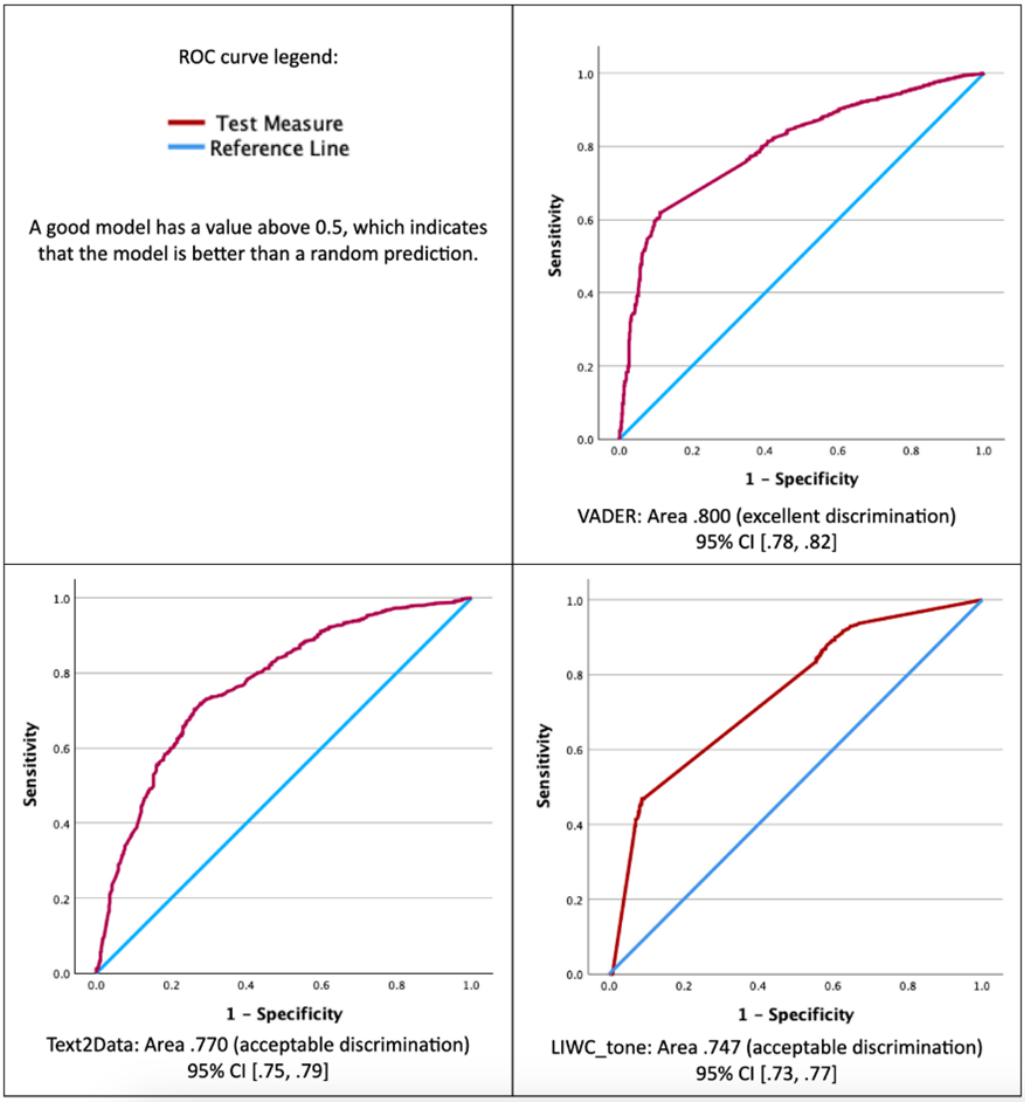

XSL•FO
RenderX

Figure 4. Receiver operating characteristic curve comparisons for Valence Aware Dictionary for Sentiment Reasoning, TEXT2DATA, and Linguistic Inquiry and Word Count Tone.

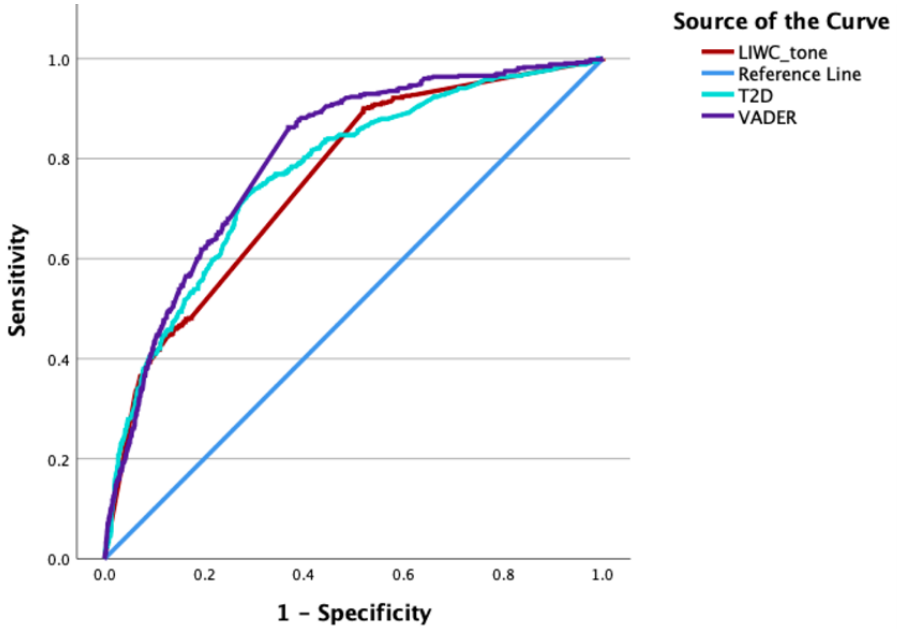

Table 1. A confusion matrix for binary coded Valence Aware Dictionary for Sentiment Reasoning, which are negative (–1) and positive (1) sentiments.

VADER[a]	Manual coding, n		Total, N
	–1	1	
–1	3693	230	3923
1	1178	432	1610
Total	3923	662	5533

[a]VADER: Valence Aware Dictionary for Sentiment Reasoning.

Table 2. A confusion matrix for binary coded t2d: negative (–1) and positive (1) sentiment.

T2D[a]	Manual coding, n		Total, n
	–1	1	
–1.00	3218	145	3363
1.00	1277	379	1656
Total	4495	524	5019

[a]T2D: TEXT2DATA.

Table 3. A confusion matrix for binary coded LIWC_tone: negative (–1) and positive (1) sentiment.

LIWC_tone[a]	Manual coding, n		Total, n
	–1	1	
–1	4214	373	4515
1	729	289	1018
Total	4871	662	5533

[a]LIWC: Linguistic Inquiry and Word Count.

XSL•FO
RenderX

Table 4. A confusion matrix for binary coded ChatGPT 4.0: negative (–1) and positive (1) sentiments.

ChatGPT 4.0	Manual coding, n, negative (–1); positive (1)		Total, n
	–1	1	
–1	4814	610	5424
1	57	52	109
Total	4871	662	5533

Table 5. A comparison of Valence Aware Dictionary for Sentiment Reasoning, TEXT2DATA, Linguistic Inquiry and Word Count tone, and ChatGPT 4.0 against manual coding.

Model	Cohen κ (95% CI; *P* value)	Accuracy	Specificity	Precision	Sensitivity (recall)	F_1-score	MCC[a]
VADER[b]	0.254 (95% CI 0.23- 0.28; <.001)	0.74	0.94	0.26	0.65	0.38	0.29
T2D[c]	0.225 (95% CI (0.2-.25; <.001)	0.72	0.96	0.23	0.72	0.35	0.3
LIWC[d]_tone	0.233 (95% CI 0.20-0.26; <.001)	0.80	0.92	0.28	0.47	0.34	0.24
ChatGPT 4.0	0.105 (95% CI 0.1-0.31; <.001)	0.88	0.88	0.50	0.08	0.13	0.16

[a]MCC: Matthews correlation coefficient.

[b]VADER: Valence Aware Dictionary for Sentiment Reasoning.

[c]T2D: TEXT2DATA.

[d]LIWC: Linguistic Inquiry and Word Count.

First, Cohen is used to examine the classification of YouTube comments into positive and negative sentiments. κ Values show fair agreement of manual coding with VADER (κ=0.254, 95% CI 0.23-0.28; *P*<.001), LIWC tone (κ=0.233, 95% CI 0.20-0.26; *P*<.001), and T2D (κ=0.26, 95% CI 0.20-0.25; *P*<.001). Overall, NLP demonstrated better agreement than LLM (κ=0.105, 95% CI 0.1-0.31; *P*<.001) for ChatGPT 4.0 indicating especially poor agreement of manual coding [40].

Second, ChatGPT 4.0 achieved the highest accuracy score of 88%, followed by LIWC tone (80%), VADER (74%), and T2D (72%). Third, the proportion of accurately detected negatives or specificity varied from 96% for T2D to 94% for VADER, 92% for LIWC tone, and 88% for ChatGPT 4.0. Fourth, precision reveals how well a model accurately makes positive predictions [41]. Precision values indicate that positive predictions were unlikely to be very accurate: ChatGPT 4.0 was at 50%, as compared with even lower values for LIWC tone (28%), VADER (26%), and T2D (23%).

Fifth, sensitivity, recognized as recall, detects true positives within a confusion matrix, with T2D achieving 72%, VADER achieving 65%, LIWC tone reaching 47%, and ChatGPT 4.0 achieving only 8% [41]. Sixth, the F_1-score is the harmonic mean of precision and sensitivity (recall), offering an evaluation of true positives and positive predictive values was below 50% for all models: VADER (at 0.38) was closely followed byT2D (0.35), LIWC tone (0.34), whereas ChatGPT's F_1-score stood out as the lowest (0.13). Finally, MCC was calculated by measuring the sensitivity, specificity, precision, and negative predictive value to evaluate the model performance [36]. None of the MCCs reported in Table 5 reached a strong correlation level; they varied from 0.30 (T2D) to 0.29 (VADER), 0.24 (LIWC tone), and the lowest value of 0.16 for ChatGPT 4.0.

Large discrepancies observed between the automated classifications and human coding may be explained by specific linguistic features, contextual nuances, model limitations, and other factors. To uncover any underlying patterns contributing to misclassifications and determine whether specific errors are more prevalent in 1 model over another, we analyzed comments classified as false positives and false negatives by LLM and NLP models.

Table 6 shows how sentiment misclassification varied by model. None of the models stood out as being excellent at avoiding both false positives and false negatives; however, ChatGPT 4.0 performed especially poorly. It classified even the most obviously positive comments as negative sentiment, leading to a very high rate (92%) of false negatives.

An analysis of false positives did not produce model-specific patterns. Sarcastic comments expressing feigned empathy for drug users were frequently misidentified by all models as false positives, for example, when the commenters talked about people dying but also added phrases such as "*White addiction, i truly love it. Now lets see what happens. Lmbao!*" [comment ID: 756]. Other common false positives involved discussions about marijuana as a gateway drug, criticisms of both Democrats and Republicans, racist remarks, and the attribution of blame to drug users.

Across all 4 models, discussions about the legalization of marijuana were misclassified as false negatives, as illustrated by this comment,

> *The opioid crisis was huge here in Florida. Look at what happened to it when they legalized marijuana for medical purposes, opioid deaths dropped! Look at the states that have it as recreational, opioid use*

XSL•FO
RenderX

and crime drop 20% in just the first year! [comment ID: 3795]

Our examination of false negatives also revealed model-specific, thematic patterns. VADER tended to misclassify political posts and comments on drug legalization as false negatives. T2D had difficulties with mentions of safe injection sites and support for Donald Trump. LIWC tone was prone to misclassifying short comments, those with emojis, as well as references to kratom, CBD (Cannabidiol) oil, and marijuana.

Finally, Table 7 summarizes the observed differences between sentiment analysis methods applied to our unbalanced social media dataset.

As shown in Table 7, T2D operates as a black box with minimal technical documentation, posing challenges for academic analysis. Unlike the other methods, VADER requires a moderate level of programming skill. All methods have low or no monetary cost (VADER is free) except for T2D, which charges per transaction, potentially making it cost-prohibitive for large datasets. Based on our findings, Table 7 provides other important considerations when selecting a model.

Table 6. Misclassified comments by model: false positives, false negatives, and representative comments.

Method, compared with manual coding	False positives, %	False negatives, %	Representative comments
VADER[a]	30.02	34.74	FP[b]: *Doctor's have become legalized dope dealers, they tried to get my 17 nephew to take opioids. They gave him a 2 month supply. Luckily he didn't take it because he'd seen what it can do. He's still playing football no thanks to the doc and big pharma* [comment ID: 70]; FN[c]: *Yes! It [kratom] got me off pain meds... pain medication* [comment ID: 2470]
T2D[d]	28.40	27.67	FP: *Big Phama! Big Insurance! Doctors get a cut for each pill script filled! We're worth more dead, than alive*! [comment ID 71]; FN: *Wish this war on opioids started earlier so many people gone. Very grateful for President Trump* [comment ID: 1483]
LIWC[e]_tone	14.97	56.34	FP: *Sue the opioid companies? What like the CIA[f]? LMFAO #CNNISFAKENEWS* [comment ID: 5872]; FN: K*ratom is the best maintenance program you could ever discover. Was on oxy for years, then methadone for years, then I found Kratom. It helps with pain, cravings, and has a bonus effect of reducing the craving for alcohol...* [comment ID: 2478]
ChatGPT 4.0	1.17	92.14	FP: *This isn't helping anyone. The medical industry is already terrified to prescribe these drugs, you people who don't take opioids have no idea of what damage you are doing to your medical system until you need to use it…* [comment ID: 909]; FN: *I have used MJ In the past to get off of Opioids and Xanax and Vicodin.. Yes, it does work and it also helps with Pain very well. It should be LEGAL even for NonMedical Use! It helps MANY problems and Various Ailments!* [comment ID: 3071]

[a]VADER: Valence Aware Dictionary for Sentiment Reasoning.

[b]FP: false positive.

[c]FN: false negative.

[d]T2D: TEXT2DATA.

[e]LIWC: Linguistic Inquiry and Word Count.

[f]CIA: Central Intelligence Agency.

XSL•FO
RenderX

Table 7. Considerations for selecting sentiment analysis methods when using social media datasets that are unbalanced toward negatives.

Method	Implementation type	Programming skill	Cost	Additional cons	Additional pros
Manual coding	Coding by a trained researcher	None	Time for creating codebook, manual coding	Retraining for intercoder reliability	Reaches good agreement when 2 humans code comments, accurate coding of ambiguous comments (sarcasm, etc)
VADER[a]	Rule-based dictionary	Moderate	Free	Not user-friendly for beginners, may code sarcastic comments as false positives.	Low runtime, fair agreement with manual codes, excellent discrimination, and performance can be improved by excluding short comments (<100 characters)
T2D[b]	Black box	None	Paid API[c]: Free up to 1000 transactions, US $27/month for 10,000 transactions.	Might not code all data, implementation is not defined (black box), may code sarcastic comments as false positives.	Low learning curve, self-contained within the same spreadsheet, high runtime, fair agreement with manual codes, acceptable discrimination
LIWC[d]_tone	Rule-based dictionary across multiple LIWC dimensions	None	Academic license: US $55 (1-year license) US $129 (3-year license)	May misclassify sarcastic comments as false positives and short comments as false negatives.	Includes a contextualizer that highlights words in reported dimensions, low learning curve, low runtime, accurate estimation of prevalence of negative versus positive sentiment, fair agreement with manual codes, acceptable discrimination
ChatGPT 4.0	Large language model	None	US $20/month	Requires prompt design, might not be consistent between iterations, may not be responsive due to high API usage, poor agreement with manual codes, inaccurate estimation of prevalence of negative vs positive sentiment, low MCC[e] may code sarcastic comments as false positives, hallucination possible.	Low runtime through the OpenAI API can provide reasoning for classification

[a]VADER: Valence Aware Dictionary for Sentiment Reasoning.

[b]T2D: TEXT2DATA.

[c]API: application programming interface.

[d]LIWC: Linguistic Inquiry and Word Count.

[e]MCC: Matthews correlation coefficient.

Discussion

Principal Findings

This study involved a comparison of manual sentiment coding to 4 automated sentiment analysis methods, namely VADER, T2D, LIWC_tone, and ChatGPT 4.0. We aimed to assess the efficacy of these sentiment analysis techniques in categorizing comments as either positive or negative sentiment in YouTube comments.

YouTube and other social media platforms are valuable repositories of comments and reviews on topics relevant to various organizations and stakeholders, such as businesses, public policy analysis, and politicians. Our corpus, comments on the US opioid crisis, was manually analyzed to reveal the struggles of opioid epidemic victims, their families, and communities, the issues of value to health policy analysis [33]. Like many other datasets of interest to social media researchers, it was skewed toward negative sentiment (7:1) and contained relatively few positive comments. Social media discussions often lean in the negative direction, such as Facebook posts on vaccine hesitancy [42] or long Covid discussions on Twitter (rebranded as X) [43]. Moreover, identifying social media content with negative effects is particularly valuable because it is more likely to be shared than positive content [44,45].

Our analysis showed that accurate positive sentiment classification can be important and policy-relevant when applied to negatively skewed data. The discussion of the opioid epidemic was primarily about human suffering but it was not all negative. Some YouTube users, for example, praised kratom as a means of overcoming opioid addiction. A subset of comments highlighted positive experiences of overcoming addiction and mixed reactions to health policies, such as safe injection sites and marijuana legalization. Marijuana, for example, can be discussed as a gateway drug (negative sentiment) and also from a harm reduction perspective as a substitute for opioids and

XSL•FO
RenderX

other harmful street drugs (positive sentiment). Classifying the data by sentiment enables researchers to explore diverse perspectives of the digital public, potentially leading to health policy insights. Given the importance of sentiment analysis in unbalanced datasets, this study offers valuable guidance for social media researchers on the pros and cons of several available methods.

Overall, VADER performed best on the ROC curve analysis, demonstrating excellent discriminatory capabilities compared with LIWC tone (acceptable) and T2D (acceptable). VADER's performance improved when short comments were excluded, a finding that verified VADER's sensitivity to text length suggested by Nair et al [18] and Tymann et al [19]. However, the ROC analysis has limitations when applied to data with an unbalanced count of negative and positive comments. Also, it could not be computed for binary ChatGPT 4.0 data. LIWC tone performed better than other automated models when estimating the prevalence of negative and positive sentiment in our data.

A comparison of Cohen κ values indicated that the NLP models (VADER, followed by LIWC and T2D) showed only fair agreement of manual coding, whereas ChatGPT 4.0 had poor agreement. While all models performed better than chance when predicting the dominant class, leading to higher precision, their level of agreement with manual coding is not exceptionally high. Moreover, variations in accuracy, specificity, precision, and sensitivity (recall), F_1-score, and MCC did not suggest a single superior model. All models evaluated had relatively low F_1–scores, which serve as overall prediction performance measures that combine precision and recall. F_1-score was below 50% for all automated models and especially low for ChatGPT 4.0. The same pattern was observed for MCC: none of the models resulted in a significant correlation (T2D's value was the highest at 0.30), but ChatGPT 4.0 performed the worst. While we cannot endorse a particular best model, our analysis suggests caution when using ChatGPT 4.0 to classify sentiment in an unbalanced dataset.

A total of 5 sets of instructions were uploaded to train ChatGPT 4.0 to conduct sentiment analysis, incorporating the codebook used by manual coders. The file upload process could have been more convenient, and 2 out of 3 attempts resulted in errors during file reading. LLM sentiment analysis was only accessible through a paid subscription, as the free version could not handle uploading and coding thousands of comments or formatting the results for further statistical analysis and comparison. Despite this, ChatGPT 4.0, like NLP models, seemed to rely on identifying words labeled as positive or negative in the codebook to classify comments, which might have led to misclassifications of sarcastic comments.

ChatGPT 4.0 outperformed the NLP models as an LLM in only 2 measures: accuracy and precision. It required specific coding instructions expressed as prompts, which we derived from the manually created codebook. Even though the same codebook was used for manual coding and designing ChatGPT 4.0 prompts, their level of agreement could have been better. Considering generation failures, poor agreement with manual coding, and the need for a paid subscription, ChatGPT 4.0 may be different from the model of choice for social media researchers looking to perform sentiment analysis on unbalanced datasets. Our comparison of ChatGPT 4.0 to 3 NLP models indicated that the general-purpose LLM has yet to surpass the performance of traditional NLP models, at least for unbalanced datasets with highly prevalent (7:1) negative comments.

Across all models, false negatives were associated with discussions on the legalization of marijuana and the observed reduction in mortality in states with more permissive drug policies. The primary issue with NLP models may stem from their reliance on pre-existing dictionaries to classify sentiment in a way that is not target specific. They cannot interpret the nuances of certain words within specific contexts the way a human can. To mitigate this, LIWC-22 and similar NLP models may require the creation of tailored dictionaries to better grasp the particular meanings of words in relevant contexts. Even then, NLP models may never be able to differentiate between negative statements of fact and negative sentiment with a specific target.

Not only NLP models but perhaps also LLM, tended to link any drug-related vocabulary with negative sentiment, failing to consider the context or nuance, particularly in discussions about overcoming the opioid crisis. This highlights a significant limitation of dictionary-based NLP models: their inability to accurately classify positive comments or recognize positive aspects of complex, contentious discussions compared with the accuracy of manual coding. Misclassification was observed for comments with sarcasm, leading models to mistake feigned empathy for genuine concern. Sentiment classification is complex and requires a deep understanding of the issues to interpret social media discourse accurately.

Manual coding remains the most reliable method for detecting sentiments when analyzing complex topics on social media, especially for unbalanced datasets. In addition to being time-intensive, it has other limitations. According to Krippendorff [46], texts often have multiple meanings. Manual coding is an interpretive process that may only sometimes match the commenter's intent, even when there is a good interrater agreement. On the other hand, automated sentiment classification may not consistently align with human judgments due to sentiment's inherently subjective and context-dependent nature, lowering reliability [47]. Fair reliability can be considered appropriate in cases involving complex or subjective tasks, exploratory research, resource constraints, qualitative analysis, or research contexts where a high level of agreement is not a primary objective [47]. Researchers must carefully weigh the trade-offs between achieving higher reliability and the practical constraints specific to their research.

Future research should explore whether general-purpose and fine-tuned LLMs and NLP models demonstrate comparable discriminatory performance in social media samples where positive comments are more prevalent than negative ones. Researchers should also test different methods for prompting general-purpose LLMs to follow complex manual codebook instructions.

XSL•FO
RenderX

Conclusions and Recommendations

We offer suggestions for VADER, T2D, LIWC tone, and ChatGPT 4.0 applications in the semantic classification of social media, specific to an unbalanced dataset with a high prevalence of negative comments. None of the automated models emerged as a clear leader. With caution, we recommend a no-cost tool, VADER, due to its excellent discrimination, according to our ROC curve analysis, which improves when the comments are at least 100 characters long. VADER requires some programming skills and may underestimate the prevalence of negative comments in unbalanced datasets. LIWC tone may be useful for social media researchers studying negative emotions, public worries, or dissatisfaction when they need to accurately estimate the prevalence of positive versus negative comments in their datasets. Researchers using T2D must know that it may only score some data and, compared with other NLP methods, can be time-consuming and cost prohibitive. ChatGPT 4.0 did not demonstrate superior performance. While the use of general-purpose LLMs is promising, it remains to be determined how to translate manual codebook instructions into prompts best to achieve superior classification results.

Data Availability

A codebook with representative comments has been added through Multimedia Appendix 1, as well as a copy of the ChatGPT 4.0 prompts used in sentiment analysis, and a list of videos that evoked YouTube comments.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Representative code book and the prompt used for ChatGPT sentiment analysis. A list of videos from which YouTube comments were extracted are also included.
[PDF File (Adobe PDF File), 228 KB - formative_v9i1e57395_app1.pdf]

References

1. Social Media Fact Sheet. URL: https://www.pewresearch.org/internet/fact-sheet/social-media/ [accessed 2024-10-02]
2. Ortiz-Ospina E. The rise of social media. Our World in Data. URL: https://ourworldindata.org/rise-of-social-media [accessed 2024-10-02]
3. Khan ML, Malik A. Researching Youtube: Methods, Tools, and Analytics. In: The Sage Handbook of Social Media Research Methods. Los Angeles, Washington DC: Sage Publications; 2022:651-663.
4. Lyu S, Ren X, Du Y, Zhao N. Detecting depression of Chinese microblog users text analysis: Combining Linguistic Inquiry Word Count (LIWC) with culture and suicide related lexicons. Front Psychiatry 2023;14:1121583 [FREE Full text] [doi: 10.3389/fpsyt.2023.1121583] [Medline: 36846219]
5. Liu JX, Leu JS, Holst S. Stock price movement prediction based on Stocktwits investor sentiment using FinBERT and ensemble SVM. PeerJ Comput Sci 2023;9:e1403 [FREE Full text] [doi: 10.7717/peerj-cs.1403] [Medline: 37346695]
6. Xu F, Pan Z, Xia R. E-commerce product review sentiment classification based on a naïve Bayes continuous learning framework. Information Processing & Management 2020;57(5):102221 [FREE Full text] [doi: 10.1016/j.ipm.2020.102221]
7. Subramanian M, Easwaramoorthy Sathiskumar V, Deepalakshmi G, Cho J, Manikandan G. A survey on hate speech detection and sentiment analysis using machine learning and deep learning models. Alexandria Engineering Journal 2023;80:110-121 [FREE Full text] [doi: 10.1016/j.aej.2023.08.038]
8. Appel O, Chiclana F, Carter J, Fujita H. A hybrid approach to the sentiment analysis problem at the sentence level. Knowledge-Based Systems 2016;108:110-124 [FREE Full text] [doi: 10.1016/j.knosys.2016.05.040]
9. Hutto C, Gilbert E. VADER: A parsimonious rule-based model for sentiment analysis of social media text. ICWSM 2014;8(1):216-225 [FREE Full text] [doi: 10.1609/icwsm.v8i1.14550]
10. Alaei AR, Becken S, Stantic B. Sentiment analysis in tourism: capitalizing on big data. J Travel Res 2019;58:175-191 [FREE Full text] [doi: 10.1177/0047287517747753]
11. Ribeiro FN, Araújo M, Gonçalves P, André Gonçalves M, Benevenuto F. SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods. EPJ Data Sci 2016;5(1):1-29 [FREE Full text] [doi: 10.1140/epjds/s13688-016-0085-1]
12. Botchway RK, Jibril AB, Kwarteng MA, Chovancova M, Oplatková ZK. A review of social media posts from UniCredit bank in Europe: a sentiment analysis approach. 2019 Presented at: Proceedings of the 3rd International Conference on Business and Information Management; 2020 July 17th-19th; Ireland. [doi: 10.1145/3361785.3361814]
13. Fasha EF, Keikhosrokiani P, Asl MP. Opinion mining using sentiment analysis: a case study of readers' response on Long Litt Woon's The Way Through the Woods in Goodreads. In: Advances on Intelligent Informatics and Computing. Cham: Springer International Publishing; 2021.

XSL•FO
RenderX

14. Bestvater SE, Monroe BL. Sentiment is not stance: target-aware opinion classification for political text analysis. Polit Anal 2022;31(2):235-256 [FREE Full text] [doi: 10.1017/pan.2022.10]
15. Nithyanand R, Schaffner B, Gill P. Online political discourse in the Trump era. ArXiv Preprint posted online on November 17, 2017 [FREE Full text] [doi: 10.5860/choice.45-0602]
16. Loveys K, Crutchley P, Wyatt EA, Coppersmith G. Small but mighty: affective micropatterns for quantifying mental health from social media language. 2017 Presented at: Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology - from Linguistic Signal to Clinical Reality; 2016 June 16; California, USA URL: https://doi.org/10.18653/v1/w17-3110 [doi: 10.18653/v1/w17-3110]
17. Valdez D, Ten Thij M, Bathina K, Rutter LA, Bollen J. Social media insights into US mental health during the COVID-19 pandemic: longitudinal analysis of Twitter data. J Med Internet Res 2020;22(12):e21418 [FREE Full text] [doi: 10.2196/21418] [Medline: 33284783]
18. Nair AJ, Veena G, Vinayak A. Comparative study of twitter sentiment on covid-19 tweets. : IEEE; 2021 Presented at: Proceedings of 5th International Conference on Computing Methodologies and Communication; 2021 April 08-10; Erode, India URL: https://doi.org/10.1109/ICCMC51019.2021.9418320 [doi: 10.1109/iccmc51019.2021.9418320]
19. Tymann K, Lutz M, Palsbröker P, Gips C. GerVADER-A German Adaptation of the VADER Sentiment Analysis tool for Social Media Texts. 2021 Presented at: Lernen, Wissen, Daten, Analysen; 2024 September 23-25; Germany p. 178-189.
20. TEXT2DATA. URL: https://text2data.com/sentiment-analysis-api [accessed 2024-10-02]
21. Gao S, Hao J, Fu Y. The application and comparison of web services for sentiment analysis in tourism. 2015 Presented at: Proceedings of the 12th International Conference on Service Systems and Service Management (ICSSSM); 2015 June 22-24; Guangzhou p. 1-6 URL: https://doi.org/10.1109/ICSSSM.2015.7170341 [doi: 10.1109/icsssm.2015.7170341]
22. Ramesh D. The correlation between Twitter sentiments and polling results for the 2016 Presidential race. South Carolina Junior Academy of Science. 2016. URL: https://scholarexchange.furman.edu/scjas/2016/all/273 [accessed 2024-10-02]
23. Tan B, Lim S, Qi Y, Miao C. A comparison of web services for sentiment analysis in digital mental health interventions. : Springer International Publishing; 2022 Presented at: Proceedings of International Conference on Human-Computer Interaction; 2023 July 23–28; Copenhagen, Denmark p. 389-407. [doi: 10.1007/978-3-031-05061-9_28]
24. Boyd RL, Ashokkumar A, Seraj S, Pennebaker JW. The development and psychometric properties of LIWC-22. The University of Texas at Austin. 2022. URL: https://www.liwc.app [accessed 2024-10-02]
25. Sell J, Farreras IG. LIWC-ing at a century of introductory college textbooks: have the sentiments changed? Procedia Computer Science 2017;118:108-112 [FREE Full text] [doi: 10.1016/j.procs.2017.11.151]
26. Smirnova A, Laranetto H, Kolenda N. Ideology through sentiment analysis: a changing perspective on Russia and Islam in NYT. Discourse & Communication 2017;11(3):296-313 [FREE Full text] [doi: 10.1177/1750481317699347]
27. Olagunju T, Oyebode O, Orji R. Exploring key issues affecting African mobile eCommerce applications using sentiment and thematic analysis. IEEE Access 2020;8:114475-114486 [FREE Full text] [doi: 10.1109/access.2020.3000093]
28. Boukes M, van de Velde B, Araujo T, Vliegenthart R. What's the Tone? Easy doesn't do it: analyzing performance and agreement between off-the-shelf sentiment analysis tools. Commun Methods Meas 2019;14(2):83-104 [FREE Full text] [doi: 10.1080/19312458.2019.1671966]
29. Hartmann J, Huppertz J, Schamp C, Heitmann M. Comparing automated text classification methods. Int J Res Mark 2019;36(1):20-38 [FREE Full text] [doi: 10.1016/j.ijresmar.2018.09.009]
30. Liu Y, Han T, Ma S, Zhang J, Yang Y, Tian J, et al. Summary of ChatGPT-related research and perspective towards the future of large language models. Meta-Radiology 2023;1(2):100017 [FREE Full text] [doi: 10.1016/j.metrad.2023.100017]
31. Baskara R, Mukarto M. Exploring the implications of chatgpt for language learning in higher education. IJELTAL 2023;7:343-358 [FREE Full text]
32. Fuchs K. Exploring the opportunities and challenges of NLP models in higher education: is Chat GPT a blessing or a curse? Front Educ 2023;8:1166682 [FREE Full text] [doi: 10.3389/feduc.2023.1166682]
33. Bacon LL. Opioid epidemic: CNN and Fox News Youtube viewers' stories, claims, and solutions. URL: https://tinyurl.com/yemxye4f [accessed 2024-10-02]
34. Cohn MA, Mehl MR, Pennebaker JW. Linguistic markers of psychological change surrounding September 11, 2001. Psychol Sci 2004;15(10):687-693. [doi: 10.1111/j.0956-7976.2004.00741.x] [Medline: 15447640]
35. Chicco D, Jurman G. The matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. BioData Min 2023;16(1):4 [FREE Full text] [doi: 10.1186/s13040-023-00322-4] [Medline: 36800973]
36. Chicco D, Jurman G. The advantages of the matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics 2020;21(1):6 [FREE Full text] [doi: 10.1186/s12864-019-6413-7] [Medline: 31898477]
37. Hilbe JM. Logistic Regression Models. New York, NY: Chapman and Hall/CRC; 2009.
38. Gonen M. Analyzing receiver operating characteristic curves with SAS. United States: SAS Institute; 2007:362.
39. Hosmer D, Lemeshow S, Sturdivant R. Applied Logistic Regression. Chicester, UK: Wiley; 2013.
40. Laerd Statistics. URL: https://statistics.laerd.com/premium/spss/ck/cohens-kappa-in-spss-7.php [accessed 2002-10-24]

XSL•FO
RenderX

41. Sharma DK, Chatterjee M, Kaur G, Vavilala S. Deep learning applications for disease diagnosis. Deep Learning for Medical Applications with Unique Data 2022;1:31-51 [FREE Full text] [doi: 10.1016/b978-0-12-824145-5.00005-8]
42. Aleksandric A, Anderson HI, Melcher S, Nilizadeh S, Wilson G. Spanish facebook posts as an indicator of COVID-19 vaccine hesitancy in Texas. Vaccines (Basel) 2022;10(10):1713 [FREE Full text] [doi: 10.3390/vaccines10101713] [Medline: 36298580]
43. Martin S, Chepo M, Deom N, Khalid AF, Vindrola-Padros C. #LongCOVID affects children too: a Twitter analysis of healthcare workers sentimentdiscourse about Long COVID in childrenyoung people in the UK. medRxiv Preprint posted online on July 25, 2022 [FREE Full text] [doi: 10.1101/2022.07.20.22277865]
44. Schöne JP, Garcia D, Parkinson B, Goldenberg A. Negative expressions are shared more on twitter for public figures than for ordinary users. PNAS Nexus 2023;2(7):pgad219. [doi: 10.1093/pnasnexus/pgad219] [Medline: 37457891]
45. Schöne J, Parkinson B, Goldenberg A. Negativity spreads more than positivity on twitter after both positive and negative political situations. Affect Sci 2020;2(4):379-390 [FREE Full text] [doi: 10.31234/osf.io/x9e7u]
46. Krippendorff K. Commentary: A dissenting view on so-called paradoxes of reliability coefficients. Annals of the International Communication Association 2016;36(1):481-499 [FREE Full text] [doi: 10.1080/23808985.2013.11679143]
47. Provoost S, Ruwaard J, van Breda W, Riper H, Bosse T. Validating automated sentiment analysis of online cognitive behavioral therapy patient texts: an exploratory study. Front Psychol 2019;10:1065 [FREE Full text] [doi: 10.3389/fpsyg.2019.01065] [Medline: 31156504]

Abbreviations

API: application programming interface
LIWC: Linguistic Inquiry and Word Count
LLM: large language model
MCC: Matthews correlation coefficient
NLP: natural language processing
ROC: receiver operating characteristic
T2D: TEXT2DATA
VADER: Valence Aware Dictionary and Sentiment Reasoning

Edited by A Mavragani; submitted 15.02.24; peer-reviewed by AR Ashraf, DB Margolin; comments to author 25.06.24; revised version received 31.08.24; accepted 20.09.24; published 08.01.25.

Please cite as:
Gandy LM, Ivanitskaya LV, Bacon LL, Bizri-Baryak R
Public Health Discussions on Social Media: Evaluating Automated Sentiment Analysis Methods
JMIR Form Res 2025;9:e57395
URL: https://formative.jmir.org/2025/1/e57395
doi:10.2196/57395
PMID:

©Lisa M Gandy, Lana V Ivanitskaya, Leeza L Bacon, Rodina Bizri-Baryak. Originally published in JMIR Formative Research (https://formative.jmir.org), 08.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Understanding Health-Related Discussions on Reddit: Development of a Topic Assignment Method and Exploratory Analysis

Garrett J Chan[1], MD; Mark Fung[2], MD, PhD; Jill Warrington[2], MD, PhD; Sarah A Nowak[2], PhD

[1]Departments of Pathology and Laboratory Medicine, University of California, San Francisco, San Francisco, CA, United States
[2]Larner College of Medicine, University of Vermont, Burlington, VT, United States

Corresponding Author:
Sarah A Nowak, PhD
Larner College of Medicine
University of Vermont
89 Beaumont Ave
Burlington, VT, 05405
United States
Phone: 1 802 656 0359
Email: sarah.nowak@med.uvm.edu

Abstract

Background: Social media has become a widely used way for people to share opinions about health care and medical topics. Social media data can be leveraged to understand patient concerns and provide insight into why patients may turn to the internet instead of the health care system for health advice.

Objective: This study aimed to develop a method to investigate Reddit posts discussing health-related conditions. Our goal was to characterize these topics and identify trends in these social media–based medical discussions.

Methods: Using an initial query, we collected 1 year of Reddit posts containing the phrases “get tested” and “get checked.” These posts were manually reviewed, and subreddits containing irrelevant posts were excluded from analysis. This selection of posts was manually read by the investigators to categorize posts into topics. A script was developed to automatically assign topics to additional posts based on keywords. Topic and keyword selections were refined based on manual review for more accurate topic assignment. Topic assignment was then performed on the entire 1-year Reddit dataset containing 347,130 posts. Related topics were grouped into broader medical disciplines. Analysis of the topic assignments was then conducted to assess condition and medical topic frequencies in medical condition–focused subreddits and general subreddits.

Results: We created an automated algorithm to assign medical topics to Reddit posts. By iterating through multiple rounds of topic assignment, we improved the accuracy of the algorithm. Ultimately, this algorithm created 82 topics sorted into 17 broader medical disciplines. Of all topics, sexually transmitted infections (STIs), eye disorders, anxiety, and pregnancy had the highest post frequency overall. STIs comprised 7.44% (5876/78,980) of posts, and anxiety comprised 5.43% (4289/78,980) of posts. A total of 34% (28/82) of the topics comprised 80% (63,184/78,980) of all posts. Of the medical disciplines, those with the most posts were psychiatry and mental health; genitourinary and reproductive health; infectious diseases; and endocrinology, nutrition, and metabolism. Psychiatry and mental health comprised 26.6% (21,009/78,980) of posts, and genitourinary and reproductive health comprised 13.6% (10,741/78,980) of posts. Overall, most posts were also classified under these 4 medical disciplines. During analysis, subreddits were also classified as *general* if they did not focus on a specific health issue and *topic-specific* if they discussed a specific medical issue. Topics that appeared most frequently in the top 5 in general subreddits included addiction and drug anxiety, attention-deficit/hyperactivity disorder, abuse, and STIs. In topic-specific subreddits, most posts were found to discuss the topic of that subreddit.

Conclusions: Certain health topics and medical disciplines are predominant on Reddit. These include topics such as STIs, eye disorders, anxiety, and pregnancy. Most posts were classified under the medical disciplines of psychiatry and mental health, as well as genitourinary and reproductive health.

(JMIR Form Res 2025;9:e55309) doi:10.2196/55309

XSL•FO
RenderX

KEYWORDS

digital health; internet; open data; social networking; social media

Introduction

Background

With over 3.8 billion people regularly sharing information about their lives on social media platforms [1], social media has become a ubiquitous part of daily life. In addition to building relationships with other people, social media helps users find information on disciplines that they are concerned about or interested in. Social media platforms have become a commonly used way to discuss problems and seek advice [2-4].

Given its prevalent use, social media content may provide insight into individual attitudes and opinions about health. The anonymous nature of certain social media platforms allows users to choose to discuss potentially embarrassing subjects without an expectation of privacy. Deidentified information can be used for research purposes to understand social behavior as it is impossible to associate a specific comment with a single person.

Analysis of social media content has other strengths compared to more traditional methods such as surveys. For example, surveys may be influenced by social desirability bias, and respondents may not be as open when discussing sensitive or potentially embarrassing conditions [5] such as drug use or sexually transmitted infections (STIs). Several studies have demonstrated that internet-based platforms can circumvent or mitigate this limitation. For example, as Google search history reflects an active process (seeking information) rather than stated beliefs, it may be more accurate than surveys for some stigmatized beliefs in particular [6]. People have a variety of motivations for using social media, including information seeking, information sharing, and expression of opinion [2]. While previous work has also shown that online content becomes more shared, or viral, when it invokes strong, activating emotions, either positive or negative [7], some studies have suggested that individuals may be more honest with sharing their thoughts online.

An advantage of the use of social media to learn about people's behaviors and sentiments is that we can obtain large volumes of data in a short period and user-generated content can be processed quickly [8,9]. Social media websites count among the most trafficked ones in the United States. Facebook is the 6th most popular website in the United States at the time of writing, Reddit is the 7th, Instagram is the 18th, and Twitter (subsequently rebranded X) is the 33rd. The variety of topics discussed on social media is wide-ranging, and these topics include information about people's health concerns as users can pose and answer questions about health-related conditions on internet forums.

Reddit has over 270 million monthly users [10,11] who can post first-person accounts of illness and interactions with the health care system and leave comments on other users' posts about their experiences with similar illnesses. At the time of writing, Reddit ranks 17th in global engagement [12]. In addition, 42.8% of Reddit site traffic comes from search engines rather than direct visits to the site. For comparison, only 12.4% of Twitter traffic comes from searches, which may mean that a greater proportion of visits to Reddit comes from internet users doing a general search for information rather than choosing to go to a specific website. Therefore, many internet users reach Reddit from search engines, making the site an important source of information to both Reddit users and nonusers.

While the spectrum of discussion about health on social media is broad, studies of sites such as Reddit, Twitter, and Facebook have been able to examine disease- and health concern–specific discussions. This includes detection of opioid addiction based on Twitter posts [13] and discussion of colorectal cancer, breast cancer, and diabetes through analysis of Twitter posts and Facebook groups [14]. One study also categorized user data from Twitter posts into 4 specific disease-related topics: *flu*, *depression*, *pregnancy*, and *eating disorders* [15]. Other studies on Reddit have sought to detect depression [16], detect suicidality among opioid users [17], and characterize weight loss discussions [18]. These works have demonstrated the ability to select specific topics from social media data and draw conclusions about how these topics are encountered in the social media sphere. Multiple studies have characterized language markers to evaluate discussions on medical topics, giving each medical topic a unique pattern of markers [19]. In all, each medical topic can be thought of as having a social media "gene" that conceptualizes it as having a recognizable identity on social media that can be tracked and analyzed [20].

The specific system architecture and features of different social media platforms make each more optimally suited for different types of studies (Table 1). For example, differences in information organization, post length, and rules for content creation can influence the analysis. The use of Reddit, for example, is well suited for evaluation of stigmatized topics and topic-based conversations. On Reddit, posts are organized into groups by topic, and users can interact through specific forums, called subreddits, to find discussions related to the subject they are interested in. A post's visibility in a subreddit is determined by user upvotes and downvotes. In addition, posts may be removed by moderators if they violate subreddit-specific community rules [21]. Therefore, subreddits are a dedicated space for people to address a topic with the expectation of pertinent discussion. Previous work has noted that the topic-based organization of Reddit content is appropriate for studying specific health topics in greater detail [22]. Furthermore, some subreddits are question and answer–based. On the subreddit *r/askdocs*, users can seek answers about medical topics from health professionals who have been verified by subreddit moderators. While these subreddits do not coalesce around a certain topic, they allow users to ask questions to content experts anonymously, a function difficult to find on other social media platforms.

XSL•FO
RenderX

Table 1. Comparison of select social media platforms by content, organization, level of anonymity, moderation, and number of articles on PubMed.

	Reddit	Twitter	Facebook
Content length	40,000-character limit	280-character limit	No limit
Content organization and sharing	Users search for posts of interest; posts organized in topic-based subreddits	On the basis of accounts that a user follows; users can search for common themes using hashtags	Often based on real-life connections
Anonymity	May be completely anonymous	May or may not be anonymous	Real-name culture encouraged
Moderation	User moderated; subreddits have content guidelines	Platform moderated	Platform moderated
PubMed articles, N	277	5769	5030

Despite discussion forums such as Reddit having features that distinguish them from platforms such as Twitter and Facebook in important ways, comparatively little research has been conducted investigating health beliefs and discussions on Reddit. A number of text analysis studies on social media platforms have been conducted—at the time of writing, there are 5769 articles on PubMed discussing Twitter and 5030 about Facebook. However, only 277 articles have been published discussing Reddit [23]. For an internet user searching for information on a specific subject, Reddit represents a useful starting point because of its topic-focused organization.

Objectives

We report a text analysis of health concerns and topics reflected in users' Reddit posts during 2019. Given the previous studies on using social media to study health and medicine, we wanted to continue this work while looking at medical topics on another platform. In addition to the relative lack of study compared to other social media platforms, we chose to investigate Reddit because it has features that differentiate it from heretofore more frequently studied social media platforms: specific, topic-based discussions; greater anonymity; and user moderators. We wanted to capture posts from a variety of health-relevant subreddits discussing seeking or receiving testing and diagnostic advice.

Methods

Ethical Considerations

The RAND Corporation was consulted regarding institutional review board approval. The RAND Human Subject Protection Committee (the RAND Institutional Review Board) determined that the project did not involve human participants as defined by the regulations of Title 45 of the Code of Federal Regulations Part 46.102(f) and, therefore, was not subject to further review. We did not explicitly obtain a waiver of informed consent; however, the determination of the absence of human participants implies a waiver of the need for documented informed consent. In addition, all data were anonymous, and obtaining informed consent would require identifying individuals, which would also be a violation of the Reddit terms of service.

Overview

We conducted our analysis in 4 steps (Figure 1) designed to balance false positives and false negatives. In step 1, we developed and refined a query to obtain posts about medical testing and advice about seeking diagnoses. In step 2, topics were created to classify the posts. Subreddits that clearly contained irrelevant posts were identified and excluded to improve the efficiency of the topic creation process. In step 3, posts assigned to the developed topics were reviewed, and the keywords used for topic assignment were refined. The process for identifying subreddits with relevant posts to be included in the analysis was refined as well. In step 4, keywords were used to do a final automatic assignment to a set of posts from a 12-month period. The 4 steps are shown in more detail in Figure 2 and are described in the following sections.

Figure 1. Overview of analysis steps—the overall methodology for gathering posts, creating topics, and automatically assigning posts to topics.

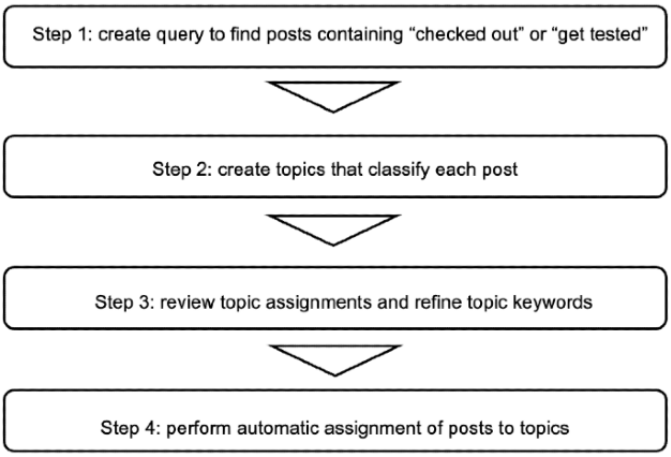

XSL•FO
RenderX

Step 1: Query Creation and Refinement

First, we captured posts from Reddit that were relevant to testing and diagnosis. We developed a query in Crimson Hexagon's Monitor application programming interface to obtain posts that contained variations of the phrases "getting tested" or "getting checked out" and excluded posts from subreddits that would be unlikely to be relevant to medicine or health care (Figure 2A and Table S1 in Multimedia Appendix 1). The query run was as follows: "((get OR got) AND ((test AND tested)OR(check OR checked)) AND for)~4" (Figure 2A). Posts gathered using this query included either the words "get" or "got" in conjunction with "test," "tested," "check," or "checked" appearing within 4 words of each other. The posts collected were created between January 1, 2019, and December 31, 2019, and included the text content of the post, its title, and its date. A *post* could be the body of text created by the original poster or a comment from a user on an original post. In both cases, the content of a post during our analysis included the title of the original post to provide additional context.

Figure 2. Detailed flow diagrams of analysis steps 1 to 4. (A) Step 1: selection of Reddit posts. Posts from January 1, 2019, to December 31, 2019, were considered in this study. (B) Step 2: identification of topic keywords. Topics were manually created from three 100-post samples, and keywords were assigned to each topic. (C) Step 3: topic assignment. Topics were manually assigned to additional 100-post samples based on the topic keyword file. (D) Step 4: automatic assignment of posts for analysis. The algorithm using the final topic keyword file was run on 78,980 posts chosen from the overall corpus of 347,130 posts.

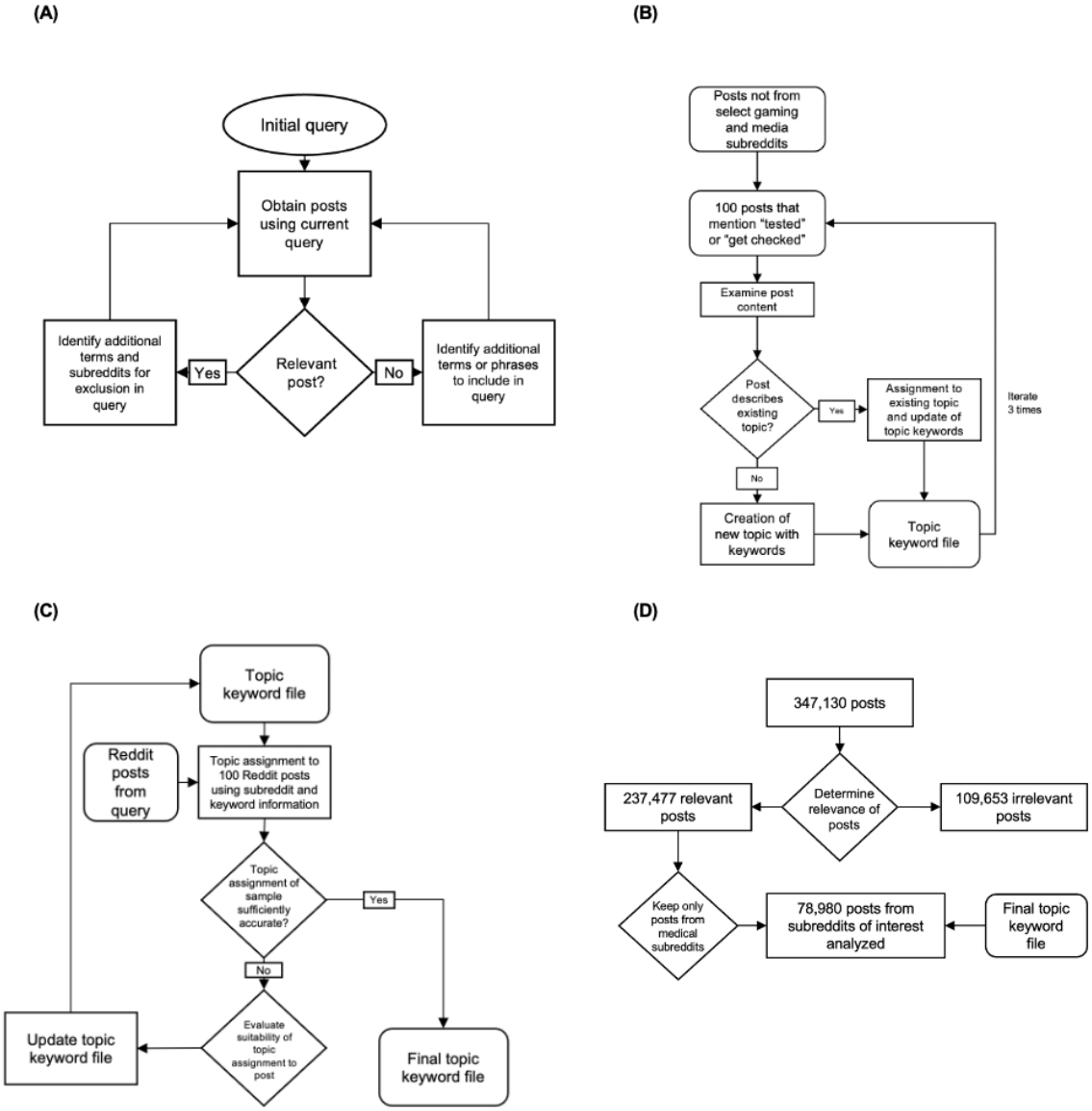

To limit our false-positive rate, we first excluded subreddits tabulated on *r/listofsubreddits* (accessed November 18, 2019) that would be most irrelevant to health-related discussions (Table S1 in Multimedia Appendix 1). Posts not from these subreddits were included in the topic creation step.

Step 2: Topic Creation

In this step, we manually read through 100 posts and labeled each with a de novo topic (Figure 2B). Topic creation was guided by the existing taxonomy in the *International Classification of Diseases, 10th Revision* (*ICD-10*) [24]. Topics were sorted into medical disciplines to optimize the analysis by clustering topics into larger groups.

XSL•FO
RenderX

These disciplines corresponded to chapter titles in the *ICD-10*. We then designated keywords that corresponded to each of these topics. For example, for the topic of STIs, keywords included "antibiotics," "gonorrhea," "chlamydia," and "bacterial vaginosis." Additional keywords were chosen from the corresponding article topic on UpToDate (Wolters Kluwer N.V.) [25], a reference for health care providers covering the diagnosis, pathophysiology, and treatment of a broad spectrum of medical conditions. In addition, several topics related to the health care system but not to a disease or topic were created, such as medical billing. Finally, for posts that did not fit with any of the aforementioned topics, a *nonmedical* topic was devised. Posts that corresponded to an existing topic were assigned to that topic.

The topics and disciplines that we formulated corresponded to the diagnoses and chapters in the *ICD-10*. A list of topics with their corresponding category is available in Table S2 in Multimedia Appendix 1. This categorization system is similar to the categorization of diseases and topics that health care providers follow. Therefore, our topics and medical disciplines were validated by an existing and accepted hierarchy of disease classification. We chose to use the *ICD-10* over other controlled vocabulary systems such as the Medical Subject Headings thesaurus of the National Library of Medicine and the Logical Observation Identifiers Names and Codes database because the *ICD-10* is the system that most physicians are in contact with and is the most targeted at covering the diagnosis of diseases. In contrast, the Medical Subject Headings thesaurus is a broader system for indexing publications that encompasses anatomy, organisms, and drugs, among many other topics.

During this process, additional irrelevant subreddits and terms were excluded from the query. In total, 3 iterations of keyword selection, each performed on a 100-post set, were completed. This generated our initial topic keyword file.

Step 3: Topic Refinement and Determination of False-Positive Rate as Means of Validating Topic Assignment

To begin automatically assigning topics to posts (Figure 2C), we created a script in R (R Foundation for Statistical Computing) that first put each post into tidy format using the *tidyverse* package [26]. The script then assigned a single topic to each post based on words found in the post and on the keyword file. With further manual analyses of more posts, we added more keywords to each topic with the goal of improving future topic assignment in later iterations.

We read through 100-post samples to add additional keywords to topics. After each run, we manually validated the accuracy of the topic assignments and iterated through this review process for 15 runs. Our topic assignments were reviewed by SAN and GC. Keyword classifications were reviewed by SAN and GC, as well as by MF and JW, 2 clinical pathologists. Accuracy was determined by reading the title and content of the posts and determining whether the topic captured the essence of the post. Topic assignment was designated as either correct or incorrect. We maintained objectivity by blinding the topic reviewer to each post's topic as assigned by the script before manually assigning a topic to that post. An example of a post and its assigned topic is as follows:

> *POST: I find that when I go through bouts of anxiety the mornings are usually the hardest, no appetite, tired etc, but as the day goes on I find I feel better, even to the extent by evening time I feel "normal."*
>
> *ASSIGNED TOPIC: Anxiety*

We then determined how many posts' topics had been correctly assigned per run and calculated an accuracy rate. During this process, we improved our script by allowing multi-word keywords and lemmatizing keywords. In addition, more subreddits were excluded to improve accuracy (Table S3 in Multimedia Appendix 1).

In addition, our algorithm excluded irrelevant posts based on irrelevant terms and subreddits determined in step 2. During the manual review process, we reviewed which posts were determined to be irrelevant by the algorithm to determine the accuracy and efficacy of using those exclusion methods [27-29]. Then, in each iteration, we calculated the false-positive rate, defined as the number of relevant posts divided by the total number of posts subtracted from 1 [30].

Posts were defined as relevant to health care if the main subject of the post related to seeking a diagnosis for a medical condition, discussion of treatment and management of a medical condition, or medical billing. An example of such a post is as follows:

> *Licensed skincare practitioner here. Get it checked. SKs [seborrheic keratoses] can itch and flake. Also may be BCC.*

An example of an irrelevant post is as follows:

> *Him going around checking if his wife cleaned properly is not something that should be rationalised....*

Step 4: Analysis of Reddit Data From January 1, 2019, to December 31, 2019, and Subreddit Analysis

To restrict our final analysis to posts mostly likely to be relevant to health-related discussion, we further limited the posts in the final analysis to those from certain subreddits. We specifically selected the 77 health-related subreddits posted on *r/listofsubreddits* and manually read through these subreddits to check for relevance to health discussions. In addition, we chose 10 subreddits with the most posts from our data sample to gather a broad range of Reddit discussion and another 8 health-related subreddits not posted on *r/listofsubreddits* that we determined contained relevant posts. In total, we obtained 95 subreddits.

We used the topic keyword file to automatically assign topics and medical disciplines to the set of Reddit posts that matched the query described previously (Figure 2D). Of note, we found that only posts from 89% (85/95) of the subreddits were assigned to a topic, indicating that the other 10 subreddits did not contain any health-related discussion as determined by the algorithm. We manually read through 300 posts to determine the accuracy rate. The topic assignments then underwent further analysis to determine the distribution of topics and medical disciplines,

XSL•FO
RenderX

distribution of posts in subreddits, and distribution of topics in each subreddit. Finally, we selected quotes from our quantitative results to provide human-relatable insights that illustrated what was discussed in each discipline, as previously done in other social media–mining research [27,31,32].

Results

Step 1: Determination of Relevant Posts, Subreddit Limitation, and Query Refinement

In total, 347,130 posts from January 1, 2019, to December 31, 2019, were downloaded from Reddit. Of those 347,130 posts, 109,653 (31.59%) belonged to subreddits that were determined to be irrelevant, and 237,477 (68.41%) came from relevant subreddits. Irrelevant subreddits were found to be those pertaining to games, popular media, and technology (Table S1 in Multimedia Appendix 1).

In step 1 (Figure 2A), the exclusion of irrelevant subreddits from our query decreased the false-positive rate of irrelevant posts from 59% to 29% (calculated from a random sample of 100 posts). We did not attempt to further decrease the false-positive rate in this step to avoid overexcluding subreddits and posts in the manual review step.

Steps 2 and 3: Topic Creation and Validation

A total of 82 topics were created through the inductive workflow outlined previously and sorted into 17 medical disciplines. Each topic was assigned to only 1 medical discipline. We validated our choices of topics and medical disciplines by comparing these categorizations to the diseases and medical topics delineated in the *ICD-10*. Of the 82 topics we created, 80 (98%) had a corresponding topic in the *ICD-10*; the exceptions were the *medical billing* and *nonmedical* topics. The accuracy of the topic assignments in this step was 56.3% (SD 4.92%; SEM 4.62%).

Step 4: Topics and Medical Disciplines

In step 4, subreddits to include in the final analysis were identified. While our objective in step 1 was to reduce the false-positive rate without being too restrictive (eliminating only subreddits least likely to contain medical discussion), our objective in this step was to keep only subreddits most likely to contain medical discussions. In total, 95 subreddits were chosen. These subreddits contained 78,980 posts.

Limiting posts to only relevant subreddits further decreased the false-positive rate from 29% to 15.5%. The accuracy rate in this step was 62.5% (SD 1.59%; SEM 0.793%), which was higher than in steps 1 and 3. Therefore, limiting posts to mainly medically relevant subreddits improved the relevance and accuracy compared to only excluding irrelevant subreddits.

The relative frequency of each topic was calculated (Figure 3). STIs, eye disorders, anxiety, and pregnancy had the highest frequency at 7.44% (5876/78,980), 5.6% (4423/78,980), 5.43% (4289/78,980), and 5.09% (4020/78,890) of all posts, respectively. Interestingly, 10 topics had a 0% relative frequency: asthma, body dysmorphic disorder, cystic fibrosis, dementia, gestational diabetes, infectious mononucleosis, Lyme disease, testicular cancer, thyroid (unspecified), and vaccination. Nonmedical posts made up 3.87% (3057/78,890) of all posts. If including the *nonmedical* topic, 80% (63,184/78,890) of posts were categorized within the first 34% (28/82) of topics. The mean relative frequency of all topics was 1.22% (SD 0.15%).

Figure 3. Relative frequency of all topics in descending order as determined by the topic assignment algorithm. The cumulative frequency of all topics is also shown. ADHD: attention-deficit/hyperactivity disorder; HPV: human papillomavirus; LGBTQ: lesbian, gay, bisexual, transgender, and queer; PCOS: polycystic ovary syndrome; PTSD: posttraumatic stress disorder; STI: sexually transmitted infection.

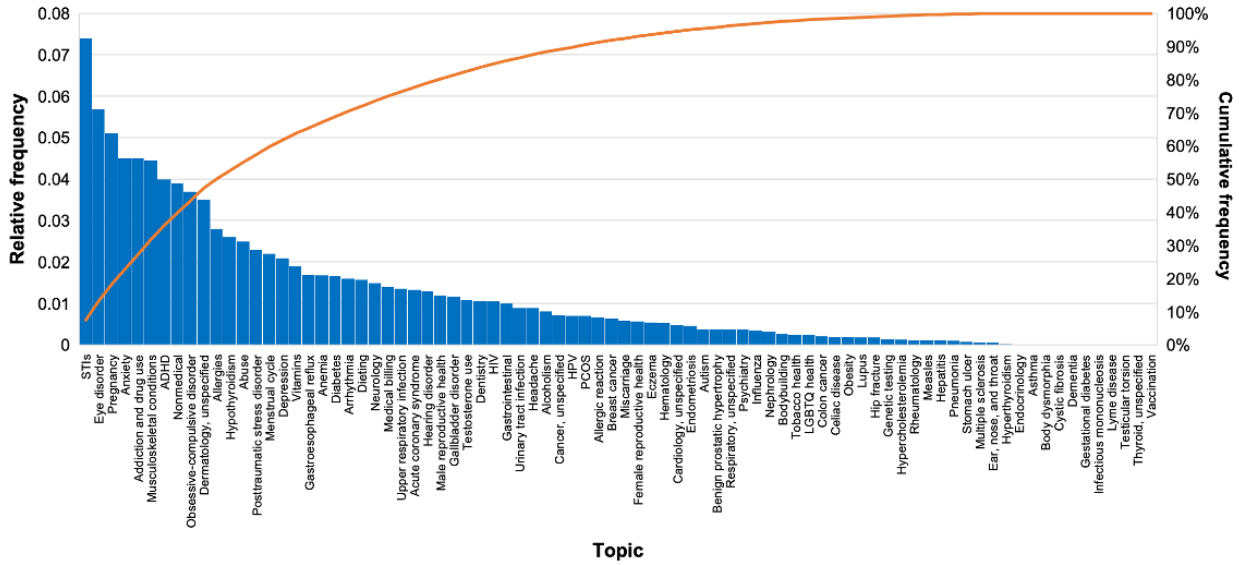

As 34% (28/82) of the topics comprised 80% (63,184/78,890) of posts, this points to a skewed distribution of topics. In total, 61% (17/28) of those topics belonged to the top 4 medical disciplines (psychiatry and mental health; genitourinary and reproductive health; infectious diseases; and endocrinology, nutrition, and metabolism).

Topics were grouped into 17 broader medical disciplines (Table 2) based on the categorization scheme of the chapter codes in

XSL•FO
RenderX

the *ICD-10*, as described previously. The most discussed disciplines were psychiatry and mental health; genitourinary (related to the urinary and reproductive systems) and reproductive health; infectious diseases; and endocrinology, nutrition, and metabolism, whose posts, respectively, constituted 26.6% (21,009/78,980), 13.6% (10,741/78,980), 8.95% (7069/78,980), and 8.91% (7037/78,980) of all analyzed posts. Psychiatry and mental health comprised most posts. Interestingly, most of the other medical disciplines were discussed with similar relative frequencies. The next most common medical discipline was hematology and immunology with a 5.88% (4644/78,890) relative frequency. The medical discipline with the lowest relative frequency was otolaryngology (related to medicine dealing with the ear, nose, and throat) at 0.57% (450/78,890).

Table 2. The relative frequency of the broader medical disciplines, into which the medical topics were sorted, shown in descending order.

Medical discipline	Relative frequency
Psychiatry and mental health	0.266
Genitourinary and reproductive health	0.136
Infectious diseases	0.0895
Endocrinology, nutrition, and metabolism	0.0891
Hematology and immunology	0.0588
Ophthalmology	0.0572
Musculoskeletal and connective tissue conditions	0.0476
Gastroenterology	0.0426
Dermatology	0.0410
Nonmedical	0.0387
Neurology	0.0373
Cardiovascular	0.0353
Pulmonology	0.0183
Medical; other	0.0160
Neoplasms	0.0157
Dentistry	0.0095
Otolaryngology	0.0022

The most common topics in the 4 most discussed medical disciplines are shown in Table 3. Within the medical discipline of psychiatry and mental health, the 3 most common topics were anxiety, addiction and drug use, and attention-deficit/hyperactivity disorder (ADHD) with relative frequencies of 20.42% (4289/21,009), 17.3% (3635/21,009), and 15.1% (3172/21,009), respectively. In the medical discipline of infectious diseases, the most common topic by a significant margin was STIs, making up 83.12% (5876/7069) of infectious disease posts. HIV and influenza made up 11.01% (778/7069) and 3.3% (233/7069) of posts, respectively. In genitourinary and reproductive health, pregnancy was the most common topic, comprising 37.43% (4020/10,741) of genitourinary and reproductive health posts, followed by menstrual cycle (1665/10,741, 15.5%) and male reproductive health (591/10,741, 5.5%). Finally, posts within the medical discipline of endocrinology, nutrition, and metabolism were more evenly distributed across topics. Hypothyroidism had the greatest number of posts (2013/7037, 28.61%), followed by vitamins (1534/7037, 21.8%) and diabetes (1316/7037, 18.7%).

XSL•FO
RenderX

Table 3. Relative frequency of topics on Reddit, as determined by our algorithm, within medical disciplines. The breakdown of the relative frequency of topics in the 4 most prevalent medical disciplines is as follows: psychiatry and mental health; genitourinary and reproductive health; infectious diseases; and endocrinology, nutrition, and metabolism. The medical disciplines are also shown in descending order by prevalence.

Medical discipline and topic	Relative frequency
Psychiatry and mental health	
Anxiety	0.203
Addiction and drug use	0.169
ADHD[a]	0.151
Obsessive-compulsive disorder	0.139
Abuse	0.094
PTSD[b]	0.087
Depression	0.079
Alcoholism	0.031
Autism	0.010
Psychiatry	0.0090
Tobacco use	0.0085
LGBTQ[c] health	0.0051
Body dysmorphic disorder	0.0015
Dementia	0.0013
Genitourinary and reproductive health	
Pregnancy	0.376
Menstrual cycle	0.160
Male reproductive health	0.088
Testosterone use	0.081
Urinary tract infection	0.068
HPV[d]	0.053
Miscarriage	0.044
Female reproductive health	0.042
Endometriosis	0.035
Benign prostatic hypertrophy	0.028
Nephrology; unspecified	0.024
Gestational diabetes	0
Testicular torsion	0
Infectious diseases	
STIs[e]	0.831
HIV	0.118
Influenza	0.039
Measles	0.013
Infectious mononucleosis	0
Lyme disease	0
Vaccination	0
Endocrinology, nutrition, and metabolism	
Hypothyroidism	0.286
Vitamins	0.216

XSL•FO
RenderX

Medical discipline and topic	Relative frequency
Diabetes	0.188
Dieting	0.178
PCOS[f]	0.078
Bodybuilding	0.029
Obesity	0.022
Hyperthyroidism	0.002
Endocrinology	0.001
Thyroid; unspecified	0

[a]ADHD: attention-deficit/hyperactivity disorder.
[b]PTSD: posttraumatic stress disorder.
[c]LGBTQ: lesbian, gay, bisexual, transgender, and queer.
[d]HPV: human papillomavirus.
[e]STI: sexually transmitted infection.
[f]PCOS: polycystic ovary syndrome.

Posts per Subreddit

Of the 95 subreddits identified in step 4, a total of 10 (11%) did not contain any posts that were assigned to our topics and were excluded, leaving 85 final subreddits. These 85 subreddits contained 78,980 posts (Table 4). The subreddits with the most posts were those that we describe as *general*, or non–topic-specific, subreddits. The top 10 subreddits with the most posts included 6 general subreddits: *r/askreddit*, *r/askdocs*, *r/amitheasshole*, *r/medical*, *r/legaladvice*, and *r/advice*. The most popular subreddit was *r/askreddit* with 26.7% (21,086/78,980) of posts, followed by *r/askdocs* with 12.02% (9496/78,980) of posts. A total of 61.93% (48,912/78,890) of the posts analyzed were found in the top 5 subreddits by number of posts, and 79.91% (63,112/78,890) were found in the top 15 subreddits.

XSL•FO
RenderX

Table 4. Subreddits by number of posts analyzed showing the relative differences in the number of posts in each subreddit (N=78,980).

Subreddit	Posts analyzed, n (%)
r/askreddit	21,086 (26.7)
r/askdocs	9496 (12.02)
r/amitheasshole	5473 (6.93)
r/relationship_advice	5322 (6.74)
r/relationships	2775 (3.51)
r/medical	2495 (3.16)
r/skincareaddiction	2268 (2.87)
r/legaladvice	2227 (2.82)
r/advice	2150 (2.72)
r/teenagers	1974 (2.5)
r/ADHD	1876 (2.38)
r/babybumps	1775 (2.25)
r/Testosterone	1746 (2.21)
r/STD	1345 (1.7)
r/Anxiety	1215 (1.54)
r/Depression	1193 (1.51)
r/pregnant	1166 (1.48)
r/birthcontrol	1102 (1.4)
r/childfree	966 (1.22)
r/Hypothyroidism	880 (1.11)
r/fitness	778 (0.99)
r/MentalHealth	741 (0.94)
r/Diabetes	699 (0.89)
r/Dentistry	602 (0.76)
r/science	489 (0.62)
r/diagnoseme	467 (0.59)
r/Infertility	446 (0.56)
r/SuicideWatch	429 (0.54)
r/Optometry	410 (0.52)
r/Ibs	391 (0.5)
r/CrohnsDisease	376 (0.48)
r/Cancer	349 (0.44)
r/Fibromyalgia	318 (0.4)
r/ChronicPain	300 (0.38)
r/MultipleSclerosis	279 (0.35)
r/Aspergers	272 (0.34)
r/bpd	235 (0.3)
r/SleepApnea	230 (0.29)
r/BipolarReddit	201 (0.25)
r/cfs	188 (0.24)
r/askdoctorsmeeee	175 (0.22)
r/askdoctors	150 (0.19)

XSL•FO
RenderX

Subreddit	Posts analyzed, n (%)
r/Rheumatoid	127 (0.16)
r/medicine	125 (0.16)
r/Eczema	120 (0.15)
r/Asthma	118 (0.15)
r/InterstitialCystitis	117 (0.15)
r/Nutrition	115 (0.15)
r/Dementia	86 (0.11)
r/askscience	84 (0.11)
r/Healthcare	80 (0.1)
r/StopSmoking	79 (0.1)
r/Dysautonomia	77 (0.1)
r/Psoriasis	75 (0.09)
r/Keratoconus	72 (0.09)
r/Pharmacy	59 (0.07)
r/Vegetarian	52 (0.07)
r/Massage	48 (0.06)
r/KidneyStones	47 (0.06)
r/GravesDisease	40 (0.05)
r/Menieres	32 (0.04)
r/Pilonidalcyst	31 (0.04)
r/Posture	30 (0.04)
r/Paleo	27 (0.03)
r/Gastroparesis	24 (0.03)
r/Hemophilia	22 (0.03)
r/Psychology	17 (0.02)
r/TroubledTeens	16 (0.02)
r/PancreaticCancer	13 (0.02)
r/Amblyopia	11 (0.01)
r/LactoseIntolerant	9 (0.01)
r/AlternativeHealth	8 (0.01)
r/NaturalBeauty	7 (0.01)
r/PublicHealth	6 (0.01)
r/Transplant	5 (0.01)
r/CareGiverSupport	3 (<0.01)
r/hepc	3 (<0.01)
r/chd	2 (<0.01)
r/Flu	2 (<0.01)
r/Juicing	1 (<0.01)
r/Pescetarian	1 (<0.01)
r/UKHealthcare	1 (<0.01)
r/Dystonia	1 (<0.01)
r/ItsNeverLupus	1 (<0.01)
r/MacularDegeneration	1 (<0.01)

XSL•FO
RenderX

General Subreddits Versus Topic-Specific Subreddits

Subreddits that we defined as *general* were those that did not focus on a specific health topic. These 15 subreddits were *r/askreddit*, *r/askdocs*, *r/amitheasshole*, *r/relationship_advice*, *r_relationships*, *r/medical*, *r/legaladvice*, *r/advice*, *r/teenagers*, *r/fitness*, *r/science*, *r/diagnoseme*, *r/askdoctorsmeeee*, *r/askdoctors*, and *r/medicine*. Compared to topic-specific subreddits, a wider variety of topics was represented in these subreddits, and no one topic dominated over other topics in each subreddit. We identified the top 5 topics by number of posts in each subreddit. Among the general subreddits, these topics comprised, on average, 41.1% of posts (SD 12.9%; 95% CI 28.2%-54%). For comparison, among topic-specific subreddits, the top 5 topics on average comprised 70.3% of posts (SD 38.3%; 95% CI 32%-100%).

In 67% (10/15) of the general subreddits, STIs was one of the top 5 topics (Table 5). Notably, the second most prevalent topic in *r/science* was human papillomavirus, whereas the most prevalent topic in *r/healthcare* (not one of the top 15 general subreddits) was medical billing. Interestingly, we also noted that the top 5 topics in *r/askreddit*, which is not specifically targeted toward health issues, were musculoskeletal conditions, eye disorders, ADHD, addiction and drug use, and allergies. For comparison, in *r/askdocs*, the most prevalent topics were STIs, dermatology, musculoskeletal conditions, allergies, and anxiety. Mental health topics were common in the general subreddits. Anxiety appeared as a top-5 topic in 53% (8/15) of the general subreddits. In those 8 subreddits, an average of 6.27% (SD 0.017%) of all posts were assigned to the topic of anxiety. Addiction and drug use appeared as a top-5 topic in 33% (5/15) of the general subreddits, and an average of 10.9% (SD 5.30%) of all posts in those 5 subreddits were assigned to that topic. On the other hand, we found that topic-specific subreddits often had the topic of the same name as the most assigned topic (eg, the topic of ADHD would be the most assigned topic in the subreddit *r/ADHD*; Table 6).

XSL•FO
RenderX

Table 5. The top 5 topics by number of posts in the top 15 general subreddits (subreddits without a specific health-related topic). A description of the subreddit and the total number of posts in the subreddit are provided (N=78,980).

Subreddit	Description	Posts, n (%)	Top 5 topics in subreddit
r/askreddit	Open question-and-answer forum	21,086 (26.7)	• Musculoskeletal conditions (n=1673 posts) • Eye disorder (n=1556 posts) • ADHD[a] (n=1043 posts) • Addiction and drug use (n=1001 posts) • Allergies (937 n=posts)
r/askdocs	Questions and answers with verified medical professionals	9496 (12.02)	• STIs[b] (n=748 posts) • Dermatology; unspecified (n=646 posts) • Musculoskeletal conditions (n=598 posts) • Allergies (n=465 posts) • Anxiety (n=435 posts)
r/amItheasshole	Users post stories to “find out if [they] were wrong”	5473 (6.93)	• STIs (n=637 posts) • Anxiety (n=461 posts) • Abuse (n=400 posts) • Nonmedical (n=353 posts) • Pregnancy (n=334 posts)
r/relationship_advice	Relationship-based discussions	5322 (6.74)	• STIs (n=1190 posts) • Abuse (n=530 posts) • Dating (n=497 posts) • Obsessive-compulsive disorder (n=374 posts) • Anxiety (n=358)
r/relationships	Interpersonal relationship–based discussion	2775 (3.51)	• STIs (n=420 posts) • Abuse (n=390 posts) • Obsessive-compulsive disorder (n=281 posts) • Dating (n=248 posts) • Anxiety (n=222 posts)
r/medical	Medical question–based discussion	2495 (3.16)	• Dermatology; unspecified (n=235 posts) • STIs (n=175 posts) • Musculoskeletal conditions (n=173 posts) • Allergies (n=151 posts) • Eye disorder (n=114 posts)
r/legaladvice	Forum to ask “simple legal questions”	2227 (2.82)	• Addiction and drug use (n=401 posts) • Medical billing (n=373 posts) • Musculoskeletal conditions (n=110 posts) • PTSD[c] (n=109 posts) • STIs (n=102 posts)
r/advice	Advice forum	2134 (2.7)	• Obsessive-compulsive disorder (n=186 posts) • STIs (n=184 posts) • Anxiety (n=168 posts) • Addiction and drug use (n=163 posts) • ADHD (n=144 posts)
r/teenagers	Intended for users aged 13-19 years to discuss age-appropriate subjects	1974 (2.5)	• Dermatology; unspecified (n=591 posts) • ADHD (n=154 posts) • Eye disorder (n=101 posts) • Depression (n=91 posts) • Anxiety (n=73 posts)
r/fitness	Physical fitness–based discussion	778 (0.99)	• Musculoskeletal conditions (n=152 posts) • Dieting (n=77 posts) • Bodybuilding (n=69 posts) • Gallbladder disorder (n=46 posts) • Testosterone use (n=37 posts)

XSL•FO
RenderX

Subreddit	Description	Posts, n (%)	Top 5 topics in subreddit
r/science	Forum for new scientific research	489 (0.62)	• Addiction and drug use (n=48 posts) • HPV[d] (n=29 posts) • Anxiety (n=27 posts) • STIs (n=23 posts) • Obsessive-compulsive disorder (n=19 posts)
r/diagnoseme	Medical question–based discussion	479 (0.61)	• Dermatology; unspecified (n=52 posts) • STIs (n=42 posts) • Allergies (n=38 posts) • Musculoskeletal conditions (n=28 posts) • Hypothyroidism (n=25 posts)
r/askdoctorsmeeee	Forum for medical questions	175 (0.22)	• Dermatology; unspecified (n=14 posts) • Musculoskeletal conditions (n=14 posts) • Allergies (n=12 posts) • Arrhythmia (n=9 posts) • Hypothyroidism (n=9 posts)
r/askdoctors	Forum for medical advice	150 (0.19)	• STIs (n=13 posts) • Dermatology; unspecified (n=11 posts) • Allergies (n=8 posts) • Anxiety (n=8 posts) • Gallbladder disorder (n=8 posts)
r/medicine	Forum for physicians to discuss medicine	125 (0.16)	• Addiction and drug use (n=18 posts) • Pregnancy (n=8 posts) • Anemia (n=6 posts) • Gallbladder disorder (n=5 posts) • HIV (n=5 posts)

[a]ADHD: attention-deficit/hyperactivity disorder.

[b]STI: sexually transmitted infection.

[c]PTSD: posttraumatic stress disorder.

[d]HPV: human papillomavirus.

XSL•FO
RenderX

Table 6. The top 5 topics by number of posts in the top 15 topic-specific subreddits (describing a certain health-related topic). A description of the subreddit and the total number of posts in the subreddit are provided (N=78,980).

Subreddit	Description	Posts, n (%)	Top 5 topics in each subreddit
r/skincareaddiction	Broad forum to discuss skin	2268 (2.87)	• Eye disorder (n=1226 posts) • Dermatology; unspecified (n=273 posts) • PCOS[a] (n=225 posts) • Allergies (n=87 posts) • Eczema (n=73 posts)
r/testosterone	Testosterone replacement therapy and testosterone levels	1746 (2.21)	• Vitamins (n=370 posts) • Testosterone use (n=264 posts) • Hypothyroidism (n=171 posts) • Pregnancy (n=105 posts) • Anemia (n=103 posts)
r/babybumps	Pregnancy-related discussion	1775 (2.25)	• Pregnancy (n=819 posts) • Musculoskeletal conditions (n=69 posts) • Anxiety (n=58 posts) • Menstrual cycle (n=58 posts) • STIs[b] (n=49 posts)
r/ADHD	Forum for people with ADHD[c] to exchange stories and strategies	1807 (2.29)	• ADHD (n=944 posts) • Addiction and drug use (n=199 posts) • Obsessive-compulsive disorder (n=112 posts) • Depression (n=91 posts) • Anxiety (n=87 posts)
r/STD	Forum to discuss STDs[d]	1611 (2.04)	• STIs (n=1158 posts) • HIV (n=75 posts) • Anxiety (n=54 posts) • HPV[e] (n=48 posts) • Respiratory infection (n=32 posts)
r/Pregnant	Forum to discuss pregnancy	1166 (1.48)	• Pregnancy (n=494 posts) • Menstrual cycle (n=92 posts) • Hematology (n=50 posts) • STIs (n=47 posts) • Anxiety (n=40 posts)
r/Depression	Support for those struggling with a depressive disorder	1193 (1.51)	• Depression (n=255 posts) • Obsessive-compulsive disorder (n=236 posts) • Anxiety (n=129 posts) • Addiction and drug use (n=940 posts) • ADHD (n=63 posts)
r/Anxiety	Support for those struggling with anxiety	1215 (1.54)	• Anxiety (n=628 posts) • Obsessive-compulsive disorder (n=165 posts) • Arrhythmia (n=55 posts) • PTSD[f] (n=53 posts) • Eye disorder (n=31 posts)
r/birthcontrol	Forum to discuss birth control	1102 (1.4)	• Menstrual cycle (n=521 posts) • Pregnancy (n=180 posts) • STIs (n=70 posts) • Anxiety (n=26 posts) • Musculoskeletal conditions (n=24 posts)
r/Childfree	Forum for those who choose to not have children	966 (1.22)	• Pregnancy (n=580 posts) • Menstrual cycle (n=58 posts) • Male reproductive health (n=32 posts) • Abuse (n=22 posts) • Anxiety (n=18 posts)

XSL•FO
RenderX

Subreddit	Description	Posts, n (%)	Top 5 topics in each subreddit
r/Hypothyroidism	Forum for those with hypothyroidism	779 (0.99)	• Hypothyroidism (n=519 posts) • Anemia (n=69 posts) • Dieting (n=29 posts) • Vitamins (n=23 posts) • Anxiety (n=19 posts)
r/MentalHealth	Moderated forum to discuss mental health issues	741 (0.94)	• Obsessive-compulsive disorder (n=208 posts) • Anxiety (n=126 posts) • Depression (n=73 posts) • ADHD (n=57 posts) • Addiction and drug use (n=43 posts)
r/Dentistry	Forum for dental professionals	683 (0.86)	• Dentistry (n=418 posts) • Dermatology; unspecified (n=28 posts) • Allergies (n=24 posts) • STIs (n=23 posts) • Musculoskeletal conditions (n=19 posts)
r/Diabetes	Forum for those living with diabetes	702 (0.89)	• Diabetes (n=491 posts) • Eye disorder (n=35 posts) • Anemia (n=23 posts) • Dieting (n=19 posts) • Vitamins (n=15 posts)
r/Infertility	Forum for those dealing with infertility for all reasons	446 (0.56)	• Pregnancy (n=110 posts) • Menstrual cycle (n=34 posts) • Vitamins (n=33 posts) • Hypothyroidism (n=29 posts) • Endometriosis (n=22 posts)

[a]PCOS: polycystic ovary syndrome.

[b]STI: sexually transmitted infection.

[c]ADHD: attention-deficit/hyperactivity disorder.

[d]STD: sexually transmitted disease.

[e]HPV: human papillomavirus.

[f]PTSD: posttraumatic stress disorder.

Medical Disciplines Divided by Subreddit

We also analyzed the breakdown of posts in each subreddit differentiated by medical discipline (Figure S1 in Multimedia Appendix 1). The relative frequency of medical disciplines in each subreddit was calculated by dividing the number of posts in a medical discipline by the total number of analyzed posts in that subreddit. Our analysis showed that many subreddits included posts about psychiatry and mental health. This mirrors the breakdown of medical disciplines overall as most posts belonged to that discipline. Some subreddits predominantly contained posts belonging to other medical disciplines. In those cases, the medical discipline often correlated with the subreddit (eg, in *r/hypothyroidism*, most posts belonged to the endocrinology, nutrition, and metabolism medical discipline). However, most subreddits comprised posts from multiple medical disciplines.

Discussion

Most Prevalent Medical Topics and Disciplines

By using markers of medical topics identified from broader samples of user data to categorize social media posts into specific diseases, previous studies have been able to predict diagnoses using social media information [13,14,20]. We accomplished a similar goal in our study by identifying topic keywords and using those to assign topics to posts. From there, our analysis revealed which disciplines were most prevalent across Reddit. Similarly to previous studies, we also used keywords to categorize this form of data into medical topics, but in this case, we did this on a broader scale, sorting data into all possible medical disciplines and generally not excluding posts that did not meet predetermined categories.

The 4 medical disciplines with the greatest proportion of posts were psychiatry and mental health; genitourinary and reproductive health; infectious diseases; and endocrinology, nutrition, and metabolism. We note the fact that psychiatry and mental health had approximately twice as many posts as the next highest medical discipline, genitourinary and reproductive health (21,009/79,980, 26.6% vs 10,741/78,980, 13.6% of posts) as another sign that discussion of mental health is common on Reddit.

Psychiatry and mental health, as a discipline, was the most prevalent. In addition to posts about anxiety, this medical discipline comprised posts about depression, addiction, alcoholism, and tobacco use. There were many posts about drug use, drug testing at work, and how drugs would affect other

XSL•FO
RenderX

health conditions. Overall, the medical discipline of psychiatry and mental health emphasized the amount of concern that people have about mental health topics and that they were willing to discuss their issues online. The large proportion of posts being about mental health raises the question of whether there is insufficient access to mental health services, whether online discussion of mental health is predominant over other health topics because of anonymity, or both.

Within this discipline, anxiety was another medical topic that frequently appeared, making up 5.43% (4289/78,980) of all posts and being in the top 5 topics in both general and topic-specific subreddits. Interestingly, anxiety was one of the most frequent topics not only in subreddits related to mental health but also in *r/PancreaticCancer*, *r/STD*, and *r/juicing*. Anxiety is a broad topic. Posts in this topic ranged from those about sharing experiences fighting melanoma to anxiety due to marijuana-triggered hallucinations to broad descriptions of "head pressure, headache, jaw pain, throat tension." Because the description of anxiety is open to interpretation, capturing which posts discuss anxiety is difficult and will result in a variety of posts that mention anxious feelings and worry, but the main subject of the post may not necessarily be anxiety.

We note that the most prevalent medical disciplines and topics can entail feelings of shame and embarrassment, be highly sensitive, or evoke a need for secrecy [33-35]. STIs was the most common topic, making up 7.44% (5876/78,980) of all posts. STIs had a significantly higher relative frequency than the next most common topic, eye disorders. Posts on this topic included many situational descriptions of suspected STIs. An example of an STI post is as follows:

> *[These] past couple of weeks, my girlfriend (we are in a monogamous relationship) found out that she was having light stringy brown discharge midway through her period cycles...So she went to the doctor and the doctor has prescribed her 1g Azithromycin (one time) and a 1 week course of Metrogyl 400 gms...The doctor also adviced [sic] her that I should take only the Azithromycin as I could have the infection as well (no blood tests or similar were done, the doctor did a physical examination)...*

The medical discipline of endocrinology, nutrition, and metabolism, which included posts about hypothyroidism, vitamins, and diabetes, among others, presented an interesting category. Posts discussed checking vitamin D levels and supplements. However, many posts also overlapped with the topic of pregnancy, such as checking vitamin levels during pregnancy and hormone levels. Therefore, it was difficult to parse whether all these posts were truly about endocrinology, nutrition, and metabolism. We also noticed that there were posts mentioning thyroid problems or diseases that were not specific to hypo- or hyperthyroidism. While these posts were more often categorized under hypothyroidism, the endocrine, nutrition, and metabolism medical discipline more accurately captured them under a broader category.

Overall, Reddit holds advantages over other social media platforms because its culture trends toward anonymity. Authors have discussed how the anonymity of communication over the internet reduces social desirability bias and social anxiety. Users are more disinhibited to discuss things openly [5]. Other studies have suggested that the internet is useful for collecting self-reports on risky health behaviors such as alcohol use, drug use, and smoking. Respondents accurately self-reported behaviors online, even ones that were undesirable, because of the perception of privacy [36]. It is established that social media sites with a culture of anonymity allow people to talk about sensitive issues online without fear of social desirability bias. Our study leverages that culture to gather a broader look at all medical disciplines rather than a predetermined set of medical topics. Our methods classified user data into those topics iteratively and broadly into all possible medical disciplines. Therefore, our findings expand on the literature by highlighting the most prevalent disciplines of conversation in a platform-wide study.

Interestingly, the top 5 topics may not be considered top health topics by health care providers. In a recent metric published by Blue Cross Blue Shield (Blue Cross Blue Shield Health Index), the top 10 health topics that had the greatest impact on quality of life in the United States did not include any of the topics that we mentioned previously. Therefore, our study may show that there is a different set of online topics, or it may reinforce that mining internet data reveals a different set of health concerns from those traditionally discussed or considered to be major health issues.

The Distribution of Topics

The aforementioned skewed distribution of topics suggests to us that conversation on Reddit is concentrated in a minority of discussion topics. This correlates with our supposition that users, when discussing health topics, overwhelmingly post about issues related to mental health, reproductive health, and infectious diseases, as well as endocrine, metabolic, and nutritional issues. The relative paucity of other topics may suggest that these are not discussed either in the general subreddits we focused on or in the topic-specific subreddits in the selection of health-related subreddits provided by Reddit but may be discussed in other subreddits.

As previously noted, 10 topics had no posts assigned to them. This meant that some topics that were created inductively during manual post review did not have any posts assigned to them during automatic topic assignment in the 1-year dataset. This invites us to investigate why certain topics would not have posts assigned to them. One reason is that posts in the subreddits included in the final analysis may not have been assigned to those topics. Another reason is that posts about a certain topic, for example, vaccination, may have also discussed measles, influenza, or human papillomavirus and, subsequently, were assigned to other topics, such as upper respiratory infection and STIs.

In general, users appeared more willing to talk about subjects that would otherwise be uncomfortable to discuss in person. This is in line with other studies that have examined how people are more willing to talk about chronic diseases on social media. They are also willing to share their disease experience in detail. Previous work has found that, on certain disease-oriented subreddits, users freely discuss disease-anatomy relationships,

XSL•FO
RenderX

such as cancer-lymph and asthma-lung relationships. The authors write that this openness on subreddits may stem from a desire to find others who share their experience of the disease [24] and users feel reassured by the subreddits' shared experiences.

Giving Advice

A subset of subreddits include those where users can post a question with the expectation that a subject matter expert will provide an answer while also being open to public input. On subreddits such as *r/askdocs* and *r/askdoctorsmeeee*, the experts are stated to be health professionals verified by the subreddit moderators. In a way, these accounts are analogous to the quantified influence described in the study by Koo et al [37] and the power accounts in the study by Sugawara et al [38], showing that, similarly to Twitter, medical advice about testing and diagnosis may be heavily influenced by several accounts on Reddit.

r/askdocs has approximately 270,000 members, whereas *r/askdoctorsmeeee* currently has approximately 14,400 members. In these subreddits, the most prevalent topics were similar to those found in other subreddits. We believe that many people turn to these forums for advice and that these subreddits are representative of the kinds of conversations that users have elsewhere on the site. However, because of the availability of experts on certain subreddits, Reddit users may be more discerning about the type of advice they expect to receive.

Limitations

We initially wanted to determine whether there was any correlation between medical topics that were mentioned on Reddit and medical tests that were mentioned on Reddit. However, our review of the Reddit posts demonstrated that users, in general, did not always mention specific medical tests, instead referring or alluding to a medical condition or a broad class of diseases, such as autoimmune diseases or STIs. At times, the content of a post would simply talk about "something being wrong," making it difficult to classify into a topic.

We noted other limitations, some of which we were able to address. Some posts were included that only had phrases such as "you should get that checked out," which were too vague to categorize under any topic. However, this problem was mitigated by using posts only from the 95 subreddits, where the frequency of such comments was lower. This was because those select subreddits, being more focused on their specific diseases or topics, had fewer vague comments and tended to have more thoughtful, directed discussion.

It is possible that keyword selection introduced bias into our topic assignments. For example, for the topic of benign prostatic hypertrophy, keywords included "difficulty urinating" and "difficulty peeing," but these keywords could have also been applicable to a topic such as nephrology (unspecified). However, we mitigated these biases by having multiple reviewers with backgrounds in clinical medicine and correlating our topic assignments with words in UpToDate. This ensured that the keywords we chose would lead to the least overlap between topics.

We also noted that some posts had too much jargon to be easily classifiable into any one topic. In future studies, they should be categorized in the *nonmedical* topic. Some posts, upon manual review, covered multiple health topics. These were the most difficult posts of all. As our algorithm did not have a way of tagging posts with multiple topics, these posts were assigned to only 1 topic. Future refinement of our algorithm could create a tagging system to identify all the major topics in a post, thereby circumventing this issue.

Future Applications

Our study provides insights into health discussions on Reddit. Our results are generalizable to other social media platforms. Facebook groups and Twitter hashtags allow for topic-based forums where people can find community discussion on health issues similarly to subreddits. Previous studies on Facebook and Twitter topic analyses have shown that social media data can be collected to better understand public sentiment on specific health topics [13-15,19,39,40]. The topic analysis methods we outlined could be applied to other social media platforms, which may show differences in what topics are discussed. Therefore, future platform-wide studies are needed.

We found that topics such as STIs and mental health issues are especially prevalent. These are potentially subjects that patients are not often discussing with their health care providers. As previously discussed, people tend to avoid potentially stigmatizing, embarrassing, and uncomfortable conversation topics, and STIs and mental health often entail such feelings. Therefore, our findings reinforce the need for health care systems to have open discussions about these issues. Health care providers establish a framework for these conversations through currently available questionnaires. The Patient Health Questionnaire–9 is a tool that providers can use to assess for depressive symptoms and mood. Asking for a sexual history during patient visits using the 5 P approach (partners, practices, protection from STIs, past history of STIs, and pregnancy) is a systematic way of starting a conversation about a patient's sexual history and concerns about STIs and pregnancy, topics that were found to be prevalent in our study. In addition, prior work has shown that, when patients seek health information online, the patient-physician relationship can be improved if the physician asks about this information and discusses it with the patient [41]. Our study supports that physicians and patients should discuss health information that the patient finds online.

Health-related discussions on social media and their effects on health policy remain an active area of research. Broadly, health-related discussions on social media have been cited as a potential impactor on disease surveillance and health care policy, practices, and delivery [42,43]. More recently, the relationship between social media and health policies has been extensively discussed in contexts such as COVID-19 [39,40,44] and mental health [45-49]. Further studies on the impact of social media on health interventions and health outcomes are needed [40]. It has been suggested that social media can play a role in health policy beyond increasing awareness of health issues [50]. Given the volume and wide variety of health topics discussed on Reddit, future health policies may need to address prevalent beliefs promulgated by health-related discussions on social

XSL•FO
RenderX

media sites. Guarding against health misinformation remains an important issue on social media and an active area of study [51-53]. Therefore, we support policies that encourage informed use of social media and discourage the spread of misinformation. Governance and education about obtaining health information online is crucial.

Conclusions

To investigate how people use Reddit to talk about seeking medical advice, we conducted a targeted review of posts that talked about "getting tested" and "checked out." Using an inductive approach to sort these posts into different medical topics, we found that discussion centered on several medical topics. They were further grouped into medical disciplines, and the infectious disease, reproductive, mental health, and endocrine categories had the greatest prevalence.

Reddit proved to be conducive to gathering information on what people discuss when talking about getting checked or tested. We took a platform-wide approach to this study, aiming to examine posts from a spectrum of subreddits related to health care and medicine. We believe that this approach is unique as it aimed to classify posts into all possible medical disciplines. In addition, based on our review of the literature, this is also one of the few studies conducted on Reddit to understand health discussions. Reddit provides similar advantages to those of other social media platforms in its large user base, wealth of information, and exchange of ideas. However, Reddit is additionally useful for studying communication about health concerns because of several of its features. Subreddits provide structured forums to further sort where discussions are happening on the platform. Several subreddits feature "expert" users, fostering a question-and-answer culture. Most importantly, users may favor Reddit because of its anonymity, which allows them to ask about and discuss potentially embarrassing topics in detail. In this way, Reddit's unique features prove it to be a viable source of research information that is worth exploring.

This study was a preliminary examination of the discussions occurring on Reddit. Future work on Reddit would likely take a more focused approach. Further exploration would include answering specific questions about certain topics encountered on Reddit and their prevalence in certain subreddits. Ultimately, these findings could be translated into actionable items, such as improving physician-patient interactions and laboratory management practices. Overall, the culture, structure, and wealth of information on Reddit make it a valuable source for continued investigation of health-related discussions.

Acknowledgments

The authors would like to thank the Department of Pathology and Laboratory Medicine at the University of Vermont Robert Larner College of Medicine and the RAND Corporation for their support. No generative artificial intelligence was used in the writing of this manuscript.

Data Availability

The datasets generated and analyzed during this study are available at the Reddit_public_data repository [54].

Conflicts of Interest

None declared.

Multimedia Appendix 1
Analysis of our Reddit dataset.
[DOCX File , 156 KB - formative_v9i1e55309_app1.docx]

References

1. We Are Social. URL: https://wearesocial.com/blog/2020/01/digital-2020-3-8-billion-people-usesocial-media [accessed 2020-01-30]
2. Kim Y, Sohn D, Choi SM. Cultural difference in motivations for using social network sites: a comparative study of American and Korean college students. Comput Hum Behav 2011 Jan;27(1):365-372. [doi: 10.1016/j.chb.2010.08.015]
3. Whiting A, Williams D. Why people use social media: a uses and gratifications approach. Qual Mark Res 2013 Aug 30;16(4):362-369. [doi: 10.1108/QMR-06-2013-0041]
4. Moturu ST, Yang J, Liu H. Quantifying utility and trustworthiness for advice shared on online social media. In: Proceedings of the International Conference on Computational Science and Engineering. 2009 Presented at: CSE 2009; August 29-31, 2009; Vancouver, BC. [doi: 10.1109/CSE.2009.461]
5. Joinson A. Social desirability, anonymity, and internet-based questionnaires. Behav Res Methods Instrum Comput 1999 Sep;31:433-438. [doi: 10.3758/BF03200723]
6. Stephens-Davidowitz S. Everybody Lies: Big Data, New Data, and What the Internet Can Tell Us About Who We Really Are. New York City, NY: William Morrow; 2017.
7. Berger J, Milkman KL. What makes online content viral? J Mark Res 2012 Apr 01;49(2):192-205. [doi: 10.1509/jmr.10.0353]

XSL•FO
RenderX

8. Shatz I. Fast, free, and targeted: Reddit as a source for recruiting participants online. Soc Sci Comput Rev 2016 May 19;35(4):537-549. [doi: 10.1177/0894439316650163]
9. Kubina M, Varmus M, Kubinova I. Use of big data for competitive advantage of company. Procedia Econ Financ 2015;26:561-565. [doi: 10.1016/s2212-5671(15)00955-7]
10. Robards B. Belonging and neo-tribalism on social media site Reddit. In: Hardy A, Bennett A, Robards B, editors. Neo-Tribes. Cham, Switzerland: Palgrave Macmillan; Sep 29, 2018.
11. Reddit statistics For 2024 (Demographics, usage and traffic data). Foundation Marketing. URL: https://foundationinc.co/lab/reddit-statistics/ [accessed 2022-09-14]
12. Amazon Alexa. Amazon Web Services. URL: https://www.alexa.com/siteinfo/reddit.com#card_overlap [accessed 2022-09-13]
13. Fan Y, Zhang Y, Ye Y, Li X, Zheng W. Social media for opioid addiction epidemiology: automatic detection of opioid addicts from Twitter and case studies. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. 2017 Presented at: CIKM '17; November 6-10, 2017; Singapore, Singapore. [doi: 10.1145/3132847.3132857]
14. De la Torre-Díez I, Díaz-Pernas FJ, Antón-Rodríguez M. A content analysis of chronic diseases social groups on Facebook and Twitter. Telemed J E Health 2012;18(6):404-408. [doi: 10.1089/tmj.2011.0227] [Medline: 22650380]
15. Prieto VM, Matos S, Álvarez M, Cacheda F, Oliveira JL. Twitter: a good place to detect health conditions. PLoS One 2014 Jan 29;9(1):e86191 [FREE Full text] [doi: 10.1371/journal.pone.0086191] [Medline: 24489699]
16. Ren L, Lin H, Xu B, Zhang S, Yang L, Sun S. Depression detection on Reddit with an emotion-based attention network: algorithm development and validation. JMIR Med Inform 2021 Jul 16;9(7):e28754 [FREE Full text] [doi: 10.2196/28754] [Medline: 34269683]
17. Yao H, Rashidian S, Dong X, Duanmu H, Rosenthal RN, Wang F. Detection of suicidality among opioid users on Reddit: machine learning-based approach. J Med Internet Res 2020 Nov 27;22(11):e15293 [FREE Full text] [doi: 10.2196/15293] [Medline: 33245287]
18. Liu Y, Yin Z. Understanding weight loss via online discussions: content analysis of Reddit posts using topic modeling and word clustering techniques. J Med Internet Res 2020 Jun 08;22(6):e13745 [FREE Full text] [doi: 10.2196/13745] [Medline: 32510460]
19. Eichstaedt JC, Smith RJ, Merchant RM, Ungar LH, Crutchley P, Preoţiuc-Pietro D, et al. Facebook language predicts depression in medical records. Proc Natl Acad Sci U S A 2018 Oct 30;115(44):11203-11208 [FREE Full text] [doi: 10.1073/pnas.1802331115] [Medline: 30322910]
20. Merchant RM, Asch DA, Crutchley P, Ungar LH, Guntuku SC, Eichstaedt JC, et al. Evaluating the predictability of medical conditions from social media posts. PLoS One 2019 Jun 17;14(6):e0215476 [FREE Full text] [doi: 10.1371/journal.pone.0215476] [Medline: 31206534]
21. Ways to become a moderator. Reddit. URL: https://mods.reddithelp.com/hc/en-us/articles/360001745332-Ways-to-become-a-moderator [accessed 2023-02-28]
22. Park A, Conway M, Chen AT. Examining thematic similarity, difference, and membership in three online mental health communities from reddit: a text mining and visualization approach. Comput Human Behav 2018 Jan;78:98-112 [FREE Full text] [doi: 10.1016/j.chb.2017.09.001] [Medline: 29456286]
23. reddit - search results - PubMed. PubMed. URL: https://pubmed.ncbi.nlm.nih.gov/?term=reddit&sort_order=asc&size=200 [accessed 2022-09-14]
24. The ICD-10 classification of mental and behavioural disorders : diagnostic criteria for research. World Health Organization. URL: https://iris.who.int/handle/10665/37108 [accessed 2022-12-03]
25. UpToDate: trusted, evidence-based solutions for modern healthcare. Wolters Kluwer. URL: https://www.wolterskluwer.com/en/solutions/uptodate [accessed 2023-07-15]
26. Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, et al. Welcome to the Tidyverse. J Open Source Softw 2019 Nov 21;4(43):1686. [doi: 10.21105/joss.01686]
27. Foufi V, Timakum T, Gaudet-Blavignac C, Lovis C, Song M. Mining of textual health information from Reddit: analysis of chronic diseases with extracted entities and their relations. J Med Internet Res 2019 Jun 13;21(6):e12876 [FREE Full text] [doi: 10.2196/12876] [Medline: 31199327]
28. Nowak SA, Chen C, Parker AM, Gidengil CA, Matthews LJ. Comparing covariation among vaccine hesitancy and broader beliefs within Twitter and survey data. PLoS One 2020 Oct 8;15(10):e0239826 [FREE Full text] [doi: 10.1371/journal.pone.0239826] [Medline: 33031405]
29. Eldh AC, Årestedt L, Berterö C. Quotations in qualitative studies: reflections on constituents, custom, and purpose. Int J Qual Methods 2020 Nov 02;19:1-6. [doi: 10.1177/1609406920969268]
30. Patton MQ. Qualitative Research & Evaluation Methods: Integrating Theory and Practice. Thousand Oaks, CA: SAGE Publications; 2014.
31. Roman P, Rodriguez-Arrastia M, Molina-Torres G, Márquez-Hernández VV, Gutiérrez-Puertas L, Ropero-Padilla C. The escape room as evaluation method: a qualitative study of nursing students' experiences. Med Teach 2020 Apr;42(4):403-410. [doi: 10.1080/0142159X.2019.1687865] [Medline: 31738615]

XSL•FO
RenderX

32. Du C, Lee W, Moskowitz D, Lucioni A, Kobashi KC, Lee UJ. I leaked, then I Reddit: experiences and insight shared on urinary incontinence by Reddit users. Int Urogynecol J 2020 Feb;31(2):243-248. [doi: 10.1007/s00192-019-04165-8] [Medline: 31832718]
33. Stuart GS, Grimes DA. Social desirability bias in family planning studies: a neglected problem. Contraception 2009 Aug;80(2):108-112. [doi: 10.1016/j.contraception.2009.02.009] [Medline: 19631784]
34. Morisky DE, Ang A, Sneed CD. Validating the effects of social desirability on self-reported condom use behavior among commercial sex workers. AIDS Educ Prev 2002 Oct;14(5):351-360. [doi: 10.1521/aeap.14.6.351.24078] [Medline: 12413181]
35. Davis CG, Thake J, Vilhena N. Social desirability biases in self-reported alcohol consumption and harms. Addict Behav 2010 Apr;35(4):302-311. [doi: 10.1016/j.addbeh.2009.11.001] [Medline: 19932936]
36. Crutzen R, Peters GJ, Portugal SD, Fisser EM, Grolleman JJ. An artificially intelligent chat agent that answers adolescents' questions related to sex, drugs, and alcohol: an exploratory study. J Adolesc Health 2011 May;48(5):514-519. [doi: 10.1016/j.jadohealth.2010.09.002] [Medline: 21501812]
37. Koo PK, Majdandzic A, Ploenzke M, Anand P, Paul SB. Global importance analysis: an interpretability method to quantify importance of genomic features in deep neural networks. PLoS Comput Biol 2021 May 13;17(5):e1008925 [FREE Full text] [doi: 10.1371/journal.pcbi.1008925] [Medline: 33983921]
38. Sugawara Y, Narimatsu H, Hozawa A, Shao L, Otani K, Fukao A. Cancer patients on Twitter: a novel patient community on social media. BMC Res Notes 2012 Dec 27;5:699 [FREE Full text] [doi: 10.1186/1756-0500-5-699] [Medline: 23270426]
39. Monselise M, Chang CH, Ferreira G, Yang R, Yang CC. Topics and sentiments of public concerns regarding COVID-19 vaccines: social media trend analysis. J Med Internet Res 2021 Oct 21;23(10):e30765 [FREE Full text] [doi: 10.2196/30765] [Medline: 34581682]
40. Valdez D, Ten Thij M, Bathina K, Rutter LA, Bollen J. Social media insights into US mental health during the COVID-19 pandemic: longitudinal analysis of Twitter data. J Med Internet Res 2020 Dec 14;22(12):e21418 [FREE Full text] [doi: 10.2196/21418] [Medline: 33284783]
41. Tan SS, Goonawardene N. Internet health information seeking and the patient-physician relationship: a systematic review. J Med Internet Res 2017 Jan 19;19(1):e9 [FREE Full text] [doi: 10.2196/jmir.5729] [Medline: 28104579]
42. Chen J, Wang Y. Social media use for health purposes: systematic review. J Med Internet Res 2021 May 12;23(5):e17917 [FREE Full text] [doi: 10.2196/17917] [Medline: 33978589]
43. Patrick M, Venkatesh RD, Stukus DR. Social media and its impact on health care. Ann Allergy Asthma Immunol 2022 Feb;128(2):139-145. [doi: 10.1016/j.anai.2021.09.014] [Medline: 34555532]
44. Lin CY, Broström A, Griffiths MD, Pakpour AH. Investigating mediated effects of fear of COVID-19 and COVID-19 misunderstanding in the association between problematic social media use, psychological distress, and insomnia. Internet Interv 2020 Sep;21:100345 [FREE Full text] [doi: 10.1016/j.invent.2020.100345] [Medline: 32868992]
45. Latha K, Meena KS, Pravitha MR, Dasgupta M, Chaturvedi SK. Effective use of social media platforms for promotion of mental health awareness. J Educ Health Promot 2020 May 28;9:124 [FREE Full text] [doi: 10.4103/jehp.jehp_90_20] [Medline: 32642480]
46. Mazuz K, Yom-Tov E. Analyzing trends of loneliness through large-scale analysis of social media postings: observational study. JMIR Ment Health 2020 Apr 20;7(4):e17188 [FREE Full text] [doi: 10.2196/17188] [Medline: 32310141]
47. Onyeaka H, Firth J, Ajayi KV, Muoghalu C, Holmes K, Nkemjika S, et al. Association between social media use and health promotion among individuals with depression and anxiety: insights from the 2017–2020 Health Information National Trends Survey. J Mood Anxiety Disord 2023 Jun;1:100006. [doi: 10.1016/j.xjmad.2023.100006]
48. Plackett R, Blyth A, Schartau P. The impact of social media use interventions on mental well-being: systematic review. J Med Internet Res 2023 Aug 11;25:e44922 [FREE Full text] [doi: 10.2196/44922] [Medline: 37565693]
49. Vandenbosch L, Fardouly J, Tiggemann M. Social media and body image: recent trends and future directions. Curr Opin Psychol 2022 Jun;45:101289. [doi: 10.1016/j.copsyc.2021.12.002] [Medline: 35030460]
50. Ghahramani A, de Courten M, Prokofieva M. "The potential of social media in health promotion beyond creating awareness: an integrative review". BMC Public Health 2022 Dec 21;22(1):2402 [FREE Full text] [doi: 10.1186/s12889-022-14885-0] [Medline: 36544121]
51. Allcott H, Boxell L, Conway J, Gentzkow M, Thaler M, Yang D. Polarization and public health: partisan differences in social distancing during the coronavirus pandemic. J Public Econ 2020 Nov;191:104254 [FREE Full text] [doi: 10.1016/j.jpubeco.2020.104254] [Medline: 32836504]
52. Suarez-Lledo V, Alvarez-Galvez J. Prevalence of health misinformation on social media: systematic review. J Med Internet Res 2021 Jan 20;23(1):e17187 [FREE Full text] [doi: 10.2196/17187] [Medline: 33470931]
53. Wang ML, Togher K. Health misinformation on social media and adolescent health. JAMA Pediatr 2024 Feb 01;178(2):109-110. [doi: 10.1001/jamapediatrics.2023.5282] [Medline: 38079180]
54. Reddit_data_public. Google Drive. URL: https://drive.google.com/drive/folders/1gkyl8COxEeumbMZP-CFa2cbBnV1pNGt1 [accessed 2025-01-16]

XSL•FO
RenderX

Abbreviations

ADHD: attention-deficit/hyperactivity disorder
***ICD-10*:** *International Classification of Diseases, 10th Revision*
STI: sexually transmitted infection

Edited by A Mavragani; submitted 08.12.23; peer-reviewed by J Mistry; comments to author 26.08.24; revised version received 10.10.24; accepted 28.10.24; published 29.01.25.

Please cite as:
Chan GJ, Fung M, Warrington J, Nowak SA
Understanding Health-Related Discussions on Reddit: Development of a Topic Assignment Method and Exploratory Analysis
JMIR Form Res 2025;9:e55309
URL: https://formative.jmir.org/2025/1/e55309
doi:10.2196/55309
PMID:39879094

©Garrett J Chan, Mark Fung, Jill Warrington, Sarah A Nowak. Originally published in JMIR Formative Research (https://formative.jmir.org), 29.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Person-Specific Analyses of Smartphone Use and Mental Health: Intensive Longitudinal Study

Merve Cerit[1], MS; Angela Y Lee[2], MA; Jeffrey Hancock[2,3], PhD; Adam Miner[4], MS, PsyD; Mu-Jung Cho[5], PhD; Daniel Muise[6], PhD; Anna-Angelina Garròn Torres[2], BA; Nick Haber[1,7], PhD; Nilam Ram[2,8], PhD; Thomas N Robinson[9], MD, MPH; Byron Reeves[2], PhD

[1]Graduate School of Education, Stanford University, Stanford, CA, United States
[2]Department of Communication, Stanford University, Stanford, CA, United States
[3]Cyber Policy Center, Stanford University, Stanford, CA, United States
[4]Department of Psychiatry and Behavioral Science, Stanford University, Stanford, CA, United States
[5]Center for Survey Research, Research Center for Humanities and Social Sciences, Academia Sinica, Taipei, Taiwan
[6]Screenlake Inc, San Mateo, CA, United States
[7]Department of Computer Science, Stanford University, Stanford, CA, United States
[8]Department of Psychology, Stanford University, Stanford, CA, United States
[9]Departments of Pediatrics, Medicine and Epidemiology & Population Health, Stanford University, Stanford, CA, United States

Corresponding Author:
Merve Cerit, MS
Graduate School of Education
Stanford University
520 Galvez Mall
Stanford, CA, 94305
United States
Phone: 1 650 723 21 46
Email: mervecer@stanford.edu

Abstract

Background: Contrary to popular concerns about the harmful effects of media use on mental health, research on this relationship is ambiguous, stalling advances in theory, interventions, and policy. Scientific explorations of the relationship between media and mental health have mostly been found null or have small associations, with the results often blamed on the use of cross-sectional study designs or imprecise measures of media use and mental health.

Objective: This exploratory empirical demonstration aims to answer whether mental health effects are associated with media use experiences by (1) redirecting research investments to granular and intensive longitudinal recordings of digital experiences to build models of media use and mental health for single individuals over the course of 1 year, (2) using new metrics of fragmented media use to propose explanations of mental health effects that will advance person-specific theorizing in media psychology, and (3) identifying combinations of media behaviors and mental health symptoms that may be more useful for studying media effects than single measures of dosage and affect or assessments of clinical symptoms related to specific disorders.

Methods: The activity on individuals' smartphone screens was recorded every 5 seconds when devices were in use over 1 year, resulting in a dataset of 6,744,013 screenshots and 123 fortnightly surveys from 5 adult participants. Each participant contributed between 0.8 and 2.7 million screens. Six media use metrics were derived from smartphone metadata. Fortnightly surveys captured symptoms of depression, attention-deficit/hyperactivity disorder, state anxiety, and positive affect. Idiographic filter models (p-technique canonical correlation analyses) were applied to explore person-specific associations.

Results: Canonical correlations revealed substantial person-specific associations between media use and mental health, ranging from r=0.82 (P=.008) to r=0.92 (P=.03). The specific combinations of media use metrics and mental health dimensions were different for each person, reflecting significant individual variability. For instance, the media use canonical variate for 1 participant was characterized by higher loadings for app-switching, which, in combination with other behaviors, correlated strongly with a mental health variate emphasizing anxiety symptoms. For another, prolonged screen time, alongside other media use behaviors, contributed to a mental health variate weighted more heavily toward depression symptoms. These within-person correlations are among the strongest reported in this literature.

XSL•FO
RenderX

Conclusions: Results suggest that the relationships between media use and mental health are highly individualized, with implications for the development of personalized models and precision smartphone-informed interventions in mental health. We discuss how our approach can be extended generally, while still emphasizing the importance of idiographic approaches. This study highlights the potential for granular, longitudinal data to reveal person-specific patterns that can inform theory development, personalized screening, diagnosis, and interventions in mental health.

(JMIR Form Res 2025;9:e59875) doi:10.2196/59875

KEYWORDS

media use; mental health; mHealth; uHealth; digital health; precision mental health; idiographic analysis; person-specific modeling; p-technique; longitudinal study; precision interventions; smartphones; idiosyncrasy; psychological well-being; canonical correlation analysis; United States

Introduction

Over the last decade, researchers have conducted extensive studies involving hundreds of thousands of participants to understand the relationship between media use and mental health. Despite significant investments, findings remain ambiguous, often reporting null or small associations [1-5]. This lack of clarity undermines progress in theory development, public policy, and clinical interventions. Researchers have proposed solutions, including granular tracking of media use [6-8], longitudinal study designs [9,10], the inclusion of at-risk populations [11-13], and person-specific analytics to account for individual differences [14,15].

A central challenge in this research is the problem of idiosyncrasy—the substantial variability in how individuals use media and how it affects their mental health [16,17]. In psychology, the opportunity to reclaim idiosyncratic variance has been defined as one of identifying idiographic filters for nomothetic interests [18]. The core argument is that the conventional practice of averaging out idiosyncrasy is neither theoretically satisfactory nor practically useful. We explore if and how an idiographic filter approach might enhance this research about media use and mental health. While researchers often seek universal rules to inform public policies or clinical interventions, such generalizations obscure the reality that individuals' media experiences are inherently personal and shaped by unique combinations of content and context. Aggregate analyses, which average out these differences, are insufficient and often misleading. Similarly, mental health symptom clusters frequently cut across traditional diagnostic boundaries, further emphasizing the need for individualized approaches. Addressing this challenge requires an idiographic perspective, which preserves individual variability and analyzes person-specific patterns rather than treating them as noise. By adopting an idiographic approach, research can uncover meaningful insights into the unique configurations of media use and mental health for each person.

To address this challenge, we used p-technique canonical correlation analysis (pCCA) [19], a method designed to analyze within-person relationships over time. This approach creates idiographic filters [20] that highlight how core constructs, such as media use and mental health metrics, manifest differently for each participant. Unlike traditional methods that assume consistency across individuals, pCCA uncovers the specific patterns and dynamics unique to each person based on the common set of variables. For example, media use heavily characterized by frequent app-switching may correlate with symptoms of hyperactivity for 1 individual, while prolonged social screen time may have significant weight in describing the changes in symptoms of depression for another. By prioritizing individual-level insights, this method provides a pathway to more personalized interventions (Multimedia Appendix 1 [2-4,7,15-19,21-50]).

This study uses granular longitudinal data collected through the Human Screenome Project [21,51], which recorded smartphone activity every 5 seconds for 1 year, resulting in over 6.7 million screenshots and 123 surveys across 5 participants. Mental health metrics, including symptoms of depression [52], attention-deficit/hyperactivity disorder (ADHD) [22], state anxiety [53], and positive affect were assessed fortnightly. Using pCCA, we explored person-specific associations between media use and mental health to address the following questions.

- Can granular longitudinal recordings reveal associations between media use and mental health that differ from the null or small effects reported in aggregated studies?
- Can new metrics, such as fragmentation of media use, explain person-specific mental health effects and advance media psychology theory?
- Can combinations of mental health symptoms provide better insights into media effects than single measures of dosage or individual clinical symptoms?

By focusing on individual experiences and variability over time, this study offers a novel approach to understanding the relationship between media use and mental health, emphasizing the potential of person-specific data to inform theory development and precision interventions.

Methods

Procedure and Participants

This study was conducted in the United States, leveraging a longitudinal design to assess smartphone use and its relationships with mental health. Data were collected via a custom smartphone app and biweekly participant surveys, enabling the integration of behavioral data with mental health measures. Participants were recruited via the Qualtrics research panel as part of a larger research project on assessing digital experiences with smartphone data [51]. Eligibility criteria included residing in the United States, being the sole user of an Android phone, and providing informed consent and Health

XSL•FO
RenderX

Insurance Portability and Accountability Act (HIPAA) authorization.

Participants installed a custom study app on their smartphones, enabling an unobtrusive collection of smartphone screens. The app required no active user management and operated transparently, with a status bar icon indicating its activity. Screenshots, along with operating system metadata like phone battery state and foreground app were captured and recorded automatically every 5 seconds when the screen was active.

Our study selected a focused sample that satisfied the minimum requirements for robust within-person analysis and allowed for a demonstration of our methodology, prioritizing long time series from a small number of participants over larger aggregate datasets so that we might uncover detailed person-specific insights and enable future scalability through computational methods [7,21,23]. We conducted a detailed analysis of 5 participants, randomly chosen from a group of 36 individuals that met several eligibility requirements. First, they had contributed at least 800,000 screenshots, reflecting an average of at least 3 hours of daily smartphone use over the year. This measure was chosen to capture use data that is close to the national average [54-56]. The goal was to ensure that participants had sufficient smartphone use to demonstrate relationships with mental health. Second, the selected participants had completed at least 20 biweekly surveys, creating a sufficient longitudinal record for each participant that exceeded the number of variables in the pCCA [57]. Finally, the selected participants had reported at least 1 biweekly period where their mental health symptoms reached clinically significant thresholds [22,52,58,59].

Ethical Considerations

This study adhered to national regulations and institutional policies, including the Declaration of Helsinki (2013), and was approved by Stanford University's institutional review board under protocol (#56430). All procedures were reviewed and classified as minimal risk.

Participants provided electronic informed consent and HIPAA authorization before enrollment. The consent process included detailed information about the study's goals, procedures, and privacy protections, and participants confirmed their consent using an e-signature interface. To ensure privacy and confidentiality, all data were encrypted during transfer and securely stored on Stanford's high-risk personal identifiable information and protected health information–compliant servers. Screenshots and metadata were deidentified prior to analysis, with access to raw screenshots restricted to a small subset of researchers under strict institutional protocols. Only aggregated, deidentified data were analyzed and reported to prevent the identification of individual participants. Participants were compensated US $15 for each fortnightly survey and data submission, with potential earnings of up to US $390 for completing the 1-year study.

Data were collected unobtrusively using a custom Screenomics app that captured screenshots every 5 seconds when the device was active. The app included an icon on the status bar to ensure transparency. Although nearly all screenshots were processed through automated methods, a small subset could be viewed by researchers in specific analyses. To mitigate risks, research staff were trained as mandated reporters under Stanford University and California guidelines. Any concerning content signaling potential legal, medical, or social risks was escalated to the Principal Investigators for appropriate action. Protocols were in place to refer participants to professionals or services as needed, ensuring their privacy while addressing potential risks to their well-being.

No images or supplementary materials in this manuscript include identifiable participant information. All visual data were anonymized to ensure privacy, and no consent was required for image inclusion as all images were deidentified.

All research staff completed mandatory HIPAA and human subjects training. Data security measures included strong passwords, restricted database access, and frequent reminders of privacy policies. These procedures ensured the highest level of data integrity and participant confidentiality.

Measures

Mental Health Measures

Participants completed fortnightly surveys over a year to assess symptoms of depression, and ADHD, in addition to state anxiety and positive affect. This extended time frame allowed the study of fluctuations across weeks or months [24-29].

Depression symptoms were measured using the 10-item Center for Epidemiologic Studies Depression scale [52]. Participants rated the frequency of symptoms (eg, "I felt I could not shake off the blues") on a 4-point Likert scale. Scores range from 0 to 60, with scores above 16 indicating clinical risk [59]. These scores were used to calculate the number of fortnight participants who were at clinical risk for depression (Table 1). The CES-D scale is in the public domain.

State anxiety was assessed with the 20-item State Anxiety subscale of the State-Trait Anxiety Inventory [60]. Items (eg, "I am tense; I am worried") were rated on a 4-point Likert scale. Scores range from 20 to 80, with scores over 40 indicating moderate to high anxiety [58]. This threshold was used to calculate fortnights at clinical risk (Table 1). A license was obtained from Mind Garden, Inc.

Positive affect was measured using a single item: "Looking back over the last two weeks, I felt happy," rated on a 100-point slider scale from "strongly disagree" to "strongly agree."

Hyperactivity and inattention symptoms were measured fortnightly using the 18-item Adult ADHD Self-Report Scale [22,61]. Participants rated hyperactivity and inattention on a 5-point Likert scale. Subscale scores range from 0 to 36. A score of 9 or more on the first 6 items (4 inattention and 2 hyperactivity) indicates clinical risk for adult ADHD [62]. These thresholds were used to calculate risk fortnights (Table 1). A license was obtained from NYU TOV Licensing, New York University.

XSL•FO
RenderX

Table 1. Participant data and demographics[a].

	Participant A (n=26)	Participant B (n=25)	Participant C (n=20)	Participant D (n=25)	Participant E (n=27)
Total screens	824,111	1,191,958	897,514	1,106,164	2,724,266
Age (year)	62	22	37	20	24
Sex	Female	Female	Female	Female	Female
Race	White	White	Multiracial Hispanic	White	Black
Region	Midwest	Midwest	Northeast	Pacific	Northeast
Residence type	Rural	Suburban	Suburban	Suburban	Urban
Highest educational level	2-year college	High school or GED[b]	Some college	Some college	Some College
Marital status	Divorced or separated	Single or never married	Married	Single or never married	Married
Income level (US $)	15,000-24,999	14,999 or less	25,000-29,999	75,000-99,999	14,999 or less

[a]The table presents the demographic characteristics of participants in the study, as collected through an initial onboarding survey. The survey was administered before the commencement of the study to gather pertinent information regarding participant backgrounds. “n” is the count of fortnightly surveys completed by each participant throughout the study year, while “Total screens” is the cumulative number of screenshots recorded from each participant’s smartphone during the study year.

[b]GED: General Educational Development.

Media Use Metrics

Six metrics for smartphone media use were derived from the screenshots and operating system metadata recorded continuously at 5-second intervals over the study year. Metadata enabled us to map app names and categories using Google Play Market classifications [63,64]. All metrics were averaged daily within biweekly periods, defined as the 14 days preceding each survey, except session duration calculated as the average screen-on interval. Missing data were addressed under the “missing at random” assumption.

Screen time was measured as the daily average number of screenshots captured during screen-on periods. The data collection app captured screenshots every 5 seconds when the screen was on. For each fortnight, the total screenshots were divided by the number of active data collection days.

Social screen time was measured as the daily average of screenshots recorded while using social apps. Social apps were those classified as “social” by Google Play Market categorization [63,64]. Social apps included common social media apps (eg, TikTok and Instagram), certain messaging apps (eg, Telegram and Text Free), and social network or dating apps (eg, Tagged and Grindr). Similar to the screen time measure, we calculated this metric by dividing the total number of social app screenshots by the number of active data collection days.

The number of sessions of smartphone use indicated average daily sessions, defined as intervals between screen on or off events. For each fortnight, we calculated the daily average number of sessions by dividing the total number of sessions by the number of days the participant actively used their phone. A higher number of sessions (eg, ten 30-second sessions) relative to the duration of each session (eg, one 5-minute session) indicates a more fragmented digital experience.

Session duration was calculated as the time interval between the initiation of a session and its termination. For each fortnight, we summed the duration of all sessions that occurred within that period and then computed the mean. This provided an average session duration metric for the specified time frame.

The number of unique apps was calculated as the daily average of distinct apps used, calculated by dividing the count of unique apps within a period by the number of active days.

The number of app switches indicated the daily average number of times individuals switched between apps and was calculated as a measure of fragmentation. For each biweekly period, we counted how many times individuals moved from app to app (ie, from Facebook to Zoom=1 switch) and divided that number by the number of active smartphone data collection days. Please see Multimedia Appendix 1 [2-4,7,15-19,21-50] for more discussion on theories and measures examining media use and mental health over time and the origins of our measures.

Data Analysis

Canonical correlation analysis (CCA) is a classical technique used to obtain the maximal correlation between objects that are represented by 2 sets of variables [65]. Adopting the Nesselroade et al idiographic filter approach [18], we use pCCA to model each individual’s multivariate repeated measures data. The observed correlation between an individual’s media use and mental health is not static; rather, it depends on the chosen coordinate system, that is, how observed variables are projected onto latent constructs of media use and mental health. Unlike traditional factorial analyses that assume a uniform variable projection across individuals, the idiographic filter approach allows person-specific rotations of the coordinate system.

pCCA does this by locating the specific rotation where the linear correlation between the latent media use and mental health canonical variates is mutually maximized. The size and significance of the maximum correlation (ρ) are used to answer our first research question: Does the analysis of single individuals offer different conclusions than the small or null effects reported in the across-individual analyses dominating the literature? The configurations of the basis vectors x and y

XSL•FO
RenderX

are interpreted to address our second and third questions: Can we identify combinations of media use metrics and mental health metrics that might advance theories about media effects?

Given A_x (27 occasions × 6 media use variables) and A_y (27 occasions × 5 mental health variables) for an individual, the data are projected onto canonical variates, $z_x = A_x x$ and $z_y = A_y y$, the correlation of which is given by the following equation.

where x and y are chosen to maximize ρ, achieved via singular value decomposition [19,66]. We use permutation tests to ensure observed correlations are not simply artifacts of random chance. Specifically, for each participant, we perform 1000 permutations by randomly shuffling the values within the columns of A_x or A_y, disrupting any temporal ordering between media use and mental health. This process generates a null distribution for ρ, representing the range of chance correlations expected in the absence of a systematic relationship. Comparing the observed ρ to this null distribution provides an exact P value based on the empirical data, thereby avoiding reliance on theoretical distributions that could inflate P values. Histograms of these permutations are presented in Multimedia Appendix 2. All pCCA and permutation test analyses were done using the *cca* package [67] in R (R Foundation for Statistical Computing).

Results

The following results include an analysis of a full year of media use and mental health data (6,744,013 screenshots and 123 fortnightly surveys) provided by 5 individuals. The background of each participant and descriptive statistics about media use and mental health are in Tables 1-3.

Table 2. Descriptive statistics for participants' mental health measures throughout 1 year[a].

Participants	Depression symptoms (0-60)		State anxiety (20-80)		ADHD[b] symptoms (0-72)			Positive affect (0-100), mean (SD)
	Mean (SD)	Clinical risk periods, n	Mean (SD)	Clinical risk periods, n	Inattention (0-36), mean (SD)	Hyperactivity (0-36), mean (SD)	Clinical risk periods, n	
A	9.23 (5.31)	1	44.23 (4.43)	22	1.88 (3.24)	2.04 (3.32)	0	53.85 (17.54)
B	24.80 (5.07)	24	51.84 (10.15)	21	12.72 (1.99)	8.72 (1.79)	13	62.12 (19.81)
C	12.40 (4.77)	4	45.05 (14.25)	13	10.10 (2.25)	5.70 (0.80)	3	65.35 (5.43)
D	12.80 (6.10)	7	45.84 (8.76)	17	5.36 (2.27)	2.76 (1.39)	0	57.00 (15.21)
E	14.78 (7.06)	4	55.00 (8.98)	27	4.52 (4.37)	4.34 (4.36)	0	16.85 (27.98)

[a]The table shows the means and SD for each mental health measure, as well as the clinical risk associated with symptomatology exceeding certain thresholds. Clinical risk periods are calculated as the number of fortnightly surveys in which participants' responses exceeded the clinical threshold for each measure.

[b]ADHD: attention-deficit/hyperactivity disorder.

Table 3. Descriptive statistics for participants' media use measures throughout 1 year.

Participant	Screen time (hours per day), mean (SD)	Social screen time (hours per day), mean (SD)	Number of sessions (per fortnight), mean (SD)	Session duration (minutes per day), mean (SD)	Number of unique apps (per day), mean (SD)	Number of app switches (per day), mean (SD)
A	2.99 (0.51)	0.41 (0.17)	57.84 (8.12)	3.10 (0.49)	4.08 (0.39)	210.27 (32.79)
B	4.15 (1.10)	0.37 (0.12)	32.76 (3.18)	7.59 (1.86)	3.64 (0.37	206.87 (18.96)
C	4.18 (0.95)	0.15 (0.23)	57.81 (8.89)	4.33 (0.85)	7.29 (2.50)	324.88 (52.73)
D	4.18 (1.05)	0 (0)	90.90 (25.28)	2.87 (0.89)	4.54 (0.61)	273.22 (54.38)
E	9.55 (2.03)	0 (0)	9.45 (2.10)	62.62 (14.75)	1.50 (0.59)	39.51 (14.48)

The main objective of this study was to assess whether analysis of media use metrics derived from super-intensive observational data passively collected from the smartphones individuals use in their daily lives could unearth relationships between media use and mental health within individuals that have previously been difficult to observe using analyses aggregated across individuals. Related to that objective, we first report the within-person associations between the media use and mental health latent variables that were obtained separately for each participant using pCCA.

The main results of this study are reported in Figure 1, which shows correlations and plots of the fortnightly media use and mental health variables across 1 year for each participant. The observed p-technique canonical correlations are large, ranging from 0.78 to 0.92. These values provide a compelling answer to our primary research question: for 4 of the 5 participants in

XSL•FO
RenderX

this study, fluctuations in canonical media use variate were significantly and tightly coupled with fluctuations in a canonical mental health variate, more so than for any previous associations observed in this literature [3,5].

Figure 1. Canonical variate plot for each participant (A-E). This figure presents the results of pCCA, highlighting within-person associations between media use and mental health metrics for five female Android phone users in the United States. The data were collected longitudinally as part of the Human Screenome Project, spanning 1 year (from July 2020 to October 2021). For each participant, the figure shows the temporal alignment of canonical media use variate (red solid line) and mental health variate (blue solid line), derived from granular measures recorded fortnightly. Individual metrics of media use (dotted red lines, eg, app-switching frequency, social screen time) and mental health (dotted blue lines, eg, depression, state anxiety) are also plotted in the background for comparison. Canonical correlations (r) for each participant range from 0.78 to 0.92, indicating strong covariation. Wilks statistic (Λ), *P* values, and sample size (n) are displayed for statistical significance. Permutation tests were conducted to validate correlations by generating null distributions based on shuffled data. Participant A shows a strong positive association between mental health and media use, with both variates generally moving together over time. Participant B exhibits a similar pattern, though with more pronounced fluctuations. Participant C has the highest correlation, indicating tightly coupled trends in media use and mental health across the year. Participant D also shows a strong association, though with some deviations at certain time points. In contrast, Participant E presents a weaker and statistically non-significant correlation, suggesting that media use and mental health follow more independent patterns in this case. These findings illustrate the unique within-person dynamics of media use and mental health over time, emphasizing the importance of person-specific models in understanding this relationship.

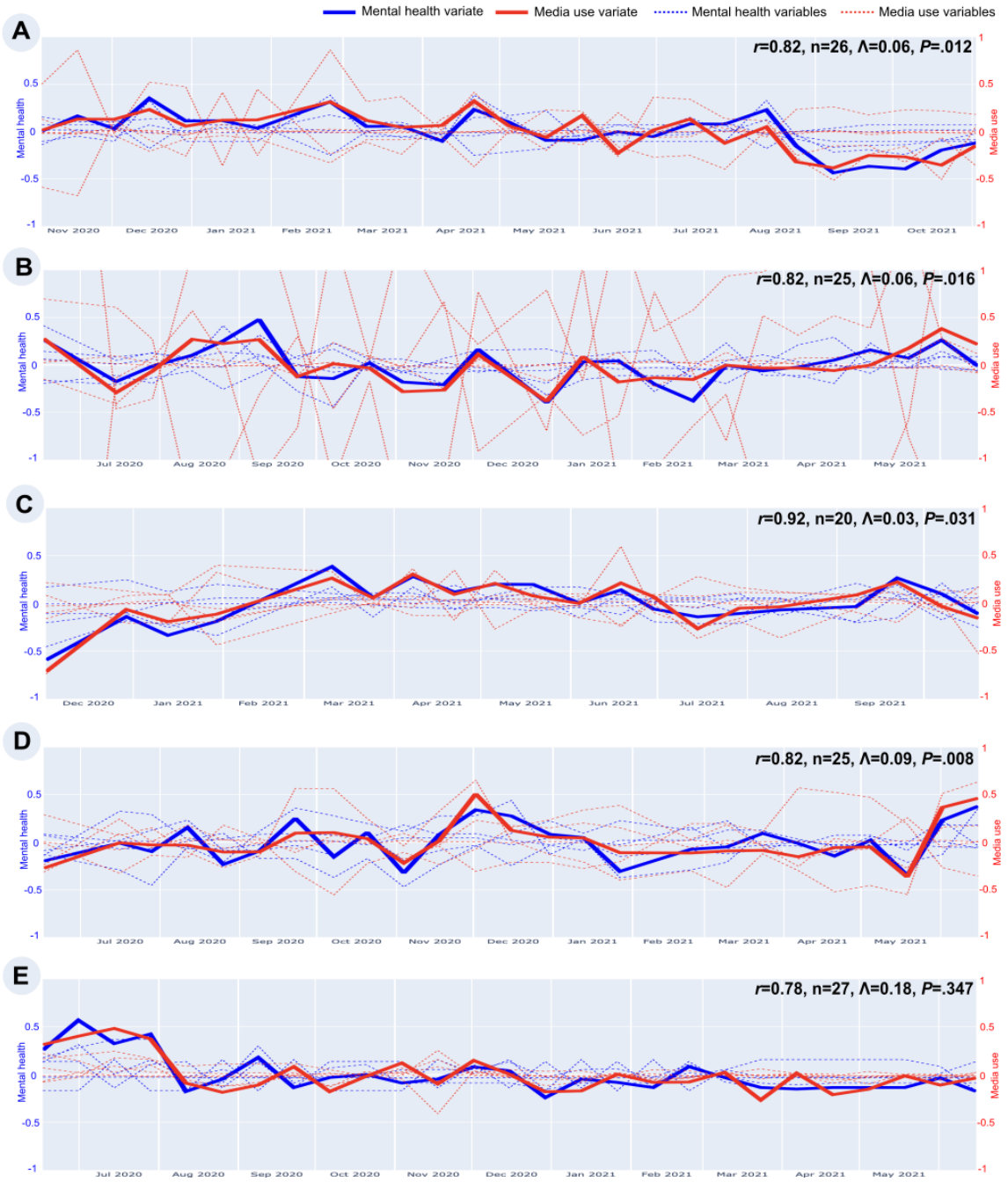

The composition of media use and mental health variates was unique for each participant, reflecting person-specific relationships and highlighting the heterogeneity in how media use impacts mental health. While individual metrics (eg,

XSL•FO
RenderX

app-switching, screen time) varied considerably over time, the canonical variates showed consistent and strong covariation for four of the 5 participants. These findings underscore the use of pCCA in identifying tightly coupled, within-person associations that are otherwise obscured in aggregated analyses, paving the way for more individualized theoretical insights and interventions.

The figures below highlight the unique person-specific relationships between media use metrics and mental health measures. These findings demonstrate the variability in how media use correlates with mental health for each participant, emphasizing the value of individualized approaches.

For participant A (Figure 2), the media use behavior most related to greater mental health symptoms over time was primarily shorter sessions (–0.93), and secondarily a higher number of sessions (0.51) and shorter screen time (–0.48). This suggests that the fluctuations in media use most related to mental health for this person are quick-checking behaviors on the smartphone. In complement, the configuration of mental health symptoms and positive affect that was most related to fluctuations in media use was characterized by lower state anxiety (–0.85) and lower levels of ADHD symptoms for both inattention (–0.59) and hyperactivity (–0.52), and lower positive affect (–0.71). Generally, for participant A, quick checking on the smartphone was associated with lower positive affect, and lower state anxiety and ADHD symptoms.

Figure 2. Canonical correlation analysis of participant A's media use and mental health. This figure illustrates the results of pCCA for participant A, based on 1 year of longitudinal data. Media use (left column) includes 6 variables (eg, session duration, number of sessions, and screen time), while mental health (right column) includes 5 variables (eg, state anxiety, depression symptoms, and positive affect). Arrows from the canonical variates (center) to observed variables indicate the direction (positive or negative) and strength of associations, with values closer to ±1.0 indicating stronger relationships. The canonical correlation (r=0.82; P=.01) highlights a significant within-person relationship, showing that quick-checking behaviors (eg, shorter session duration and more frequent sessions) were associated with reduced anxiety and ADHD symptoms but also diminished positive affect.

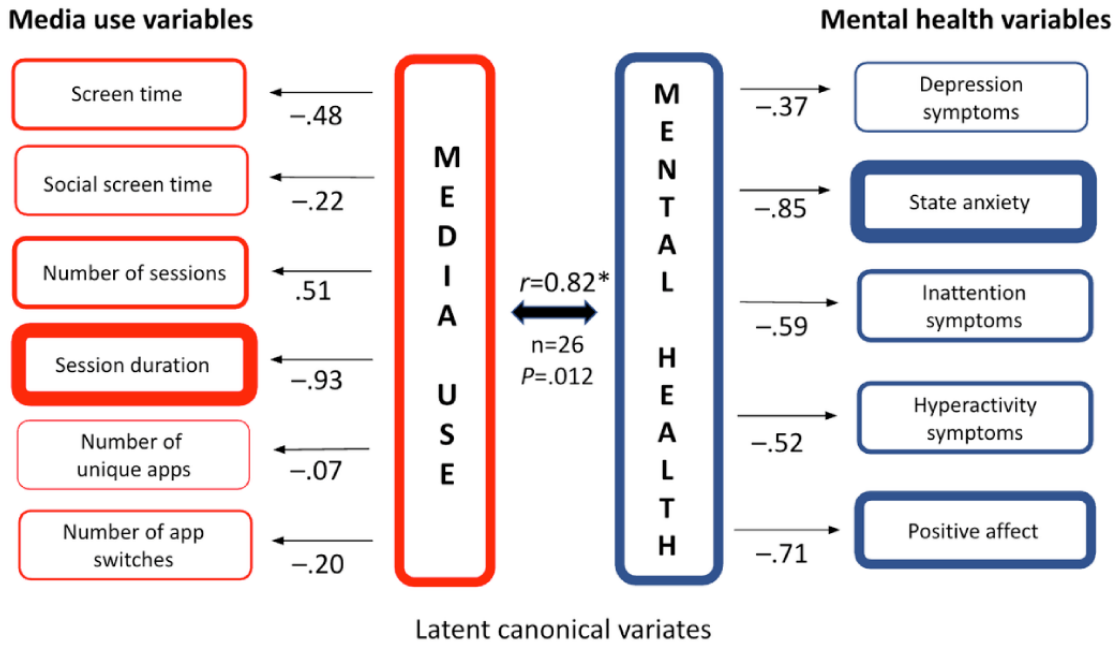

Figure 3 shows that for participant B, the media use behaviors most related to greater mental health symptoms over time were characterized by longer sessions (0.66) and secondarily by more screen time (0.53) and little app switching (–0.55). We characterize this combination as focused smartphone use. The mental health symptoms most related to media use were characterized by depression symptoms (0.61) and lower positive affect (–0.53). It is also notable that this participant's mental health changes, especially true for depression symptoms, had a stronger association with media use changes than for state anxiety. The media use canonical variate for participant B was driven by fluctuations in total screen time rather than social app use, specifically.

XSL•FO
RenderX

Figure 3. Canonical correlation analysis of participant B's media use and mental health. This figure illustrates the results of pCCA for participant B, based on 1 year of longitudinal data. Media use (left column) includes 6 variables (eg, session duration, screen time, and app-switching intensity), while mental health (right column) includes 5 variables (eg, depression symptoms, state anxiety, and positive affect). Arrows from the canonical variates (center) to observed variables indicate the direction (positive or negative) and strength of associations, with values closer to ±1.0 indicating stronger relationships. The canonical correlation (*r*=0.82; *P*=.02) highlights a significant within-person relationship, showing that focused smartphone use—characterized by longer session duration, higher screen time, and minimal app-switching—was associated with increased depression symptoms and reduced positive affect.

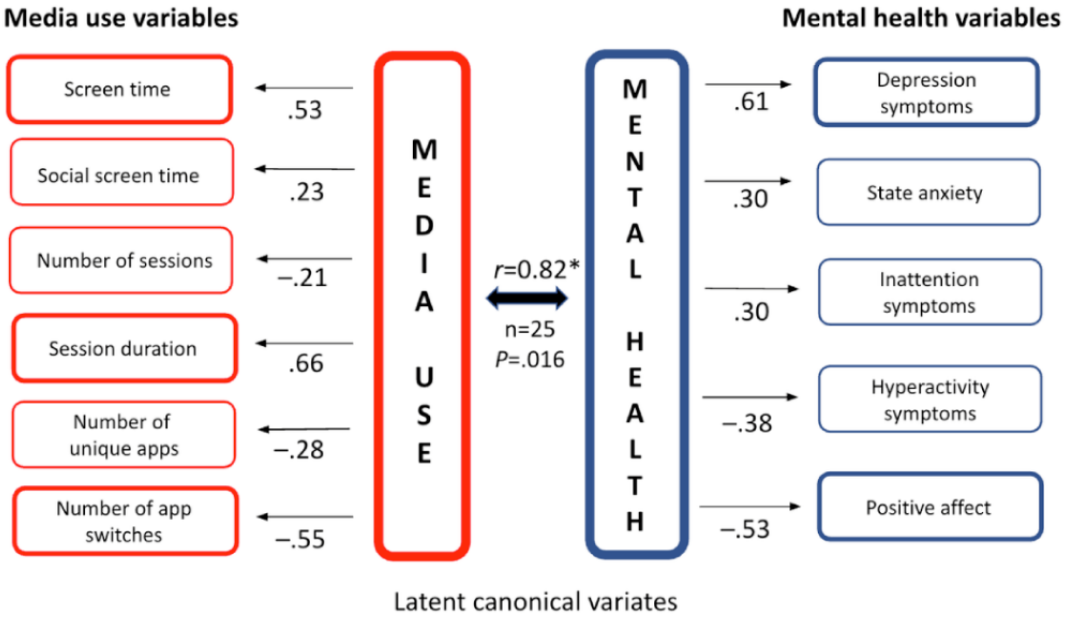

For participant C (Figure 4), the media use canonical variate shows the strongest loadings for the number of social screen time (distinct from total screen time) at 0.79. Secondarily, there is a high loading for the use of unique apps (0.53). The mental health canonical variate shows a strong loading associated with depression symptoms (0.75). Additionally, there is a noteworthy negative loading in relation to hyperactivity symptoms (–0.60). The much larger loading for social screen time in comparison to total screen time shows the value of separating these uses when examining relationships with mental health. This distinction is often part of discussions about media effects (ie, more worry about intensive social app use and depression than about the effects of all media use). As for many media effects, however, the value of the distinction is person-specific.

Figure 4. Canonical correlation analysis of participant C's media use and mental health. This figure illustrates the results of pCCA for participant C, based on 1 year of longitudinal data. Media use (left column) includes 6 variables (eg, social screen time, number of unique apps, session duration), while mental health (right column) includes 5 variables (eg, depression symptoms, hyperactivity symptoms, and positive affect). Arrows from the canonical variates (center) to the observed variables indicate the direction (positive or negative) and strength of associations, with values closer to ±1.0 reflecting stronger relationships. The canonical correlation (*r*=0.92; *P*=.03) underscores a significant within-person relationship, with social screen time (0.79) emerging as the strongest media use factor, closely tied to increased depressive symptoms (0.75) and reduced hyperactivity symptoms (–0.60).

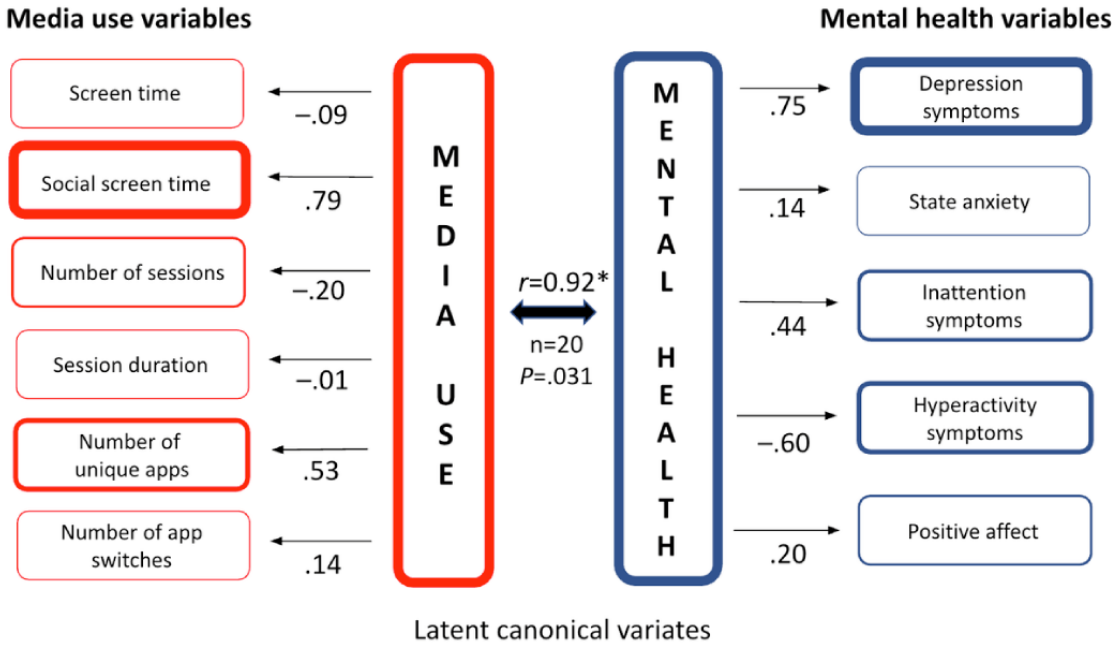

XSL•FO
RenderX

Participant D (Figure 5) did not use any social apps during the entire year (ie, there are no screens from any of the hundreds of social apps in the Google Play store). Their media use canonical variate is characterized by fewer app switches (–0.77), extended session durations (0.71), and a reduced number of sessions overall (–0.57). This pattern indicates longer, focused interactions. The mental health canonical variate was characterized primarily by reduced positive affect (–0.58), with no large loadings associated with any of the mental health symptoms. For participant D, longer and more focused smartphone sessions had a more singular definition (diminished positive affect). However, the number and duration of the sessions were still highly correlated with smartphone use features (r=0.82; P=.008).

There was no statistically significant association between the media and mental health canonical variates for Participant E (Figure 6).

Figure 5. Canonical correlation analysis of participant D's media use and mental health. This figure presents the results of pCCA for participant D, based on longitudinal data collected over 1 year. Media use (left column) includes 6 variables (eg, session duration, number of sessions, and app switching), while mental health (right column) includes 5 variables (eg, positive affect, state anxiety, and ADHD symptoms). The arrows between latent canonical variates (center) and observed variables represent the direction (positive or negative) and strength of associations, with values closer to ±1.0 indicating stronger relationships. The canonical correlation (r=0.82; P=.008) underscores a significant within-person association, highlighting that longer session durations and reduced app-switching behaviors were most associated with lower positive affect for participant D.

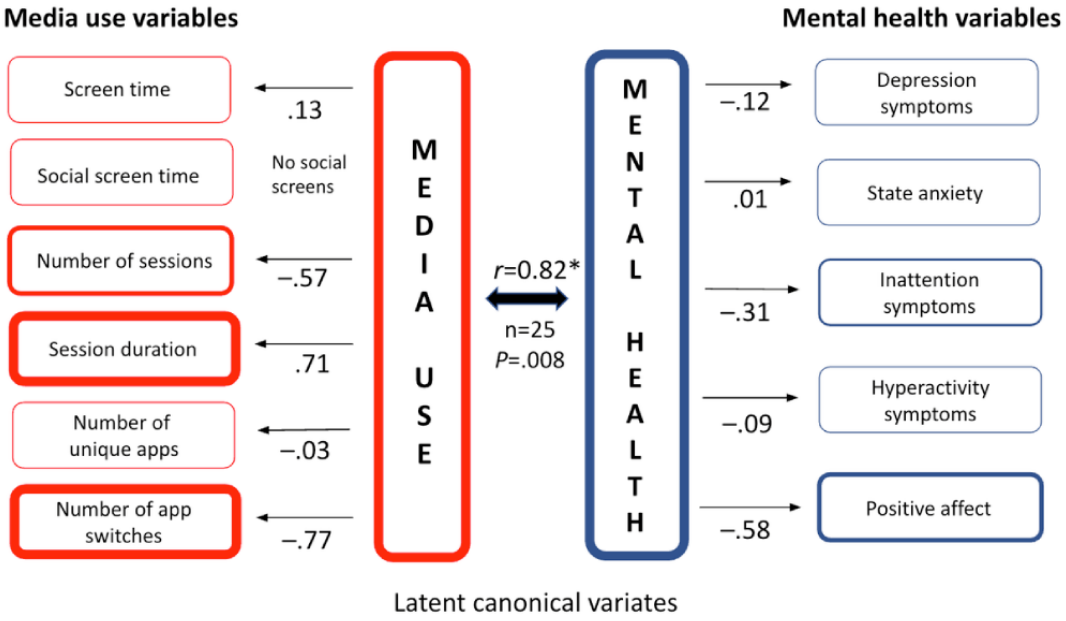

Figure 6. Canonical correlation analysis of participant E's media use and mental health. This figure illustrates the results of pCCA for participant E, based on 1 year of longitudinal data. Media use (left column) is represented by 6 variables, such as screen time, session duration, and number of app switches, while mental health (right column) is represented by 5 variables, including depression symptoms, state anxiety, and positive affect. Arrows connecting latent canonical variates (center) to observed variables indicate the direction (positive or negative) and strength of associations, with values closer to ±1.0 reflecting stronger associations. The canonical correlation (r=0.78; P=.35) indicates a nonsignificant association.

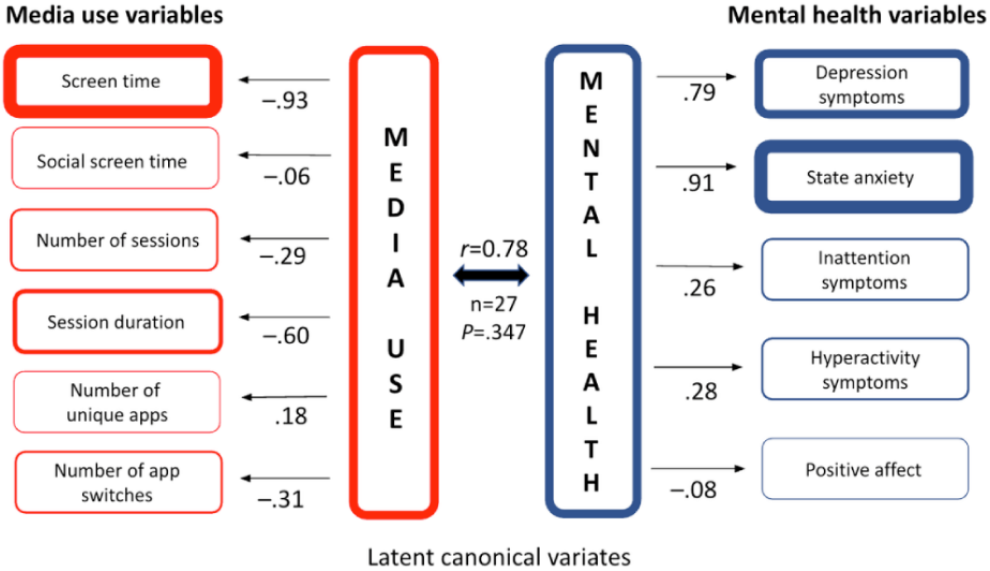

XSL•FO
RenderX

Discussion

Principal Findings

The results highlight the person-specific associations between media use and mental health. Canonical correlations revealed substantial covariation (0.82 to 0.92) over time for 4 out of 5 individuals. Simultaneously, as hypothesized, the specific combinations of media dosage metrics and dimensions of mental health that defined this covariation were different for each person (Multimedia Appendix 1 [2-4,7,15-19,21-50]. The approach supports the idea of a common association between media use and mental health while accommodating idiosyncrasy in how that association manifests for individuals. The couplings between media use and mental health for each person offer insights into the complexity of this relationship, directly addressing the research questions posed in the introduction.

In summary, our conclusions are the following.

- As research investments shift toward intensive, granular longitudinal data about individuals' digital experiences, this study demonstrates the potential to construct distinct, person-specific models of media use and mental health, revealing insights that transcend traditional aggregated analyses.
- The study's innovative approach, combining various dosage metrics of fragmented media use, advances person-specific theorizing in media psychology, offering potential explanations for the impact on mental health and emphasizing the significance of tailored approaches.
- By identifying individualized combinations of mental health symptoms, the study underscores the necessity of considering nuanced factors beyond single measures of dosage and affect or clinical symptoms associated with specific disorders.

Causality and Theory

The substantial p-technique canonical correlations observed in this study reflect within-person associations between media use and mental health but do not establish causality. Media use metrics may act as markers responding to mental health changes, antecedents influencing mental health, or reflections of confounding variables. A causal explanation likely involves complex mechanisms interacting over different time domains and individual histories, which this analysis does not aim to disentangle.

Past research has explored single mechanisms over varying time frames, such as the introduction of Facebook on campuses (months/years) [68], ecological momentary assessment studies linking self-reported media use with mental health (days/weeks) [69], and cohort analyses over decades [70]. Additionally, momentary exposures, like a single unflattering social media post, might amplify depression symptoms, particularly during periods of heightened vulnerability. Interactions between short- and long-term domains—for example, a negative social media post during an extended depressive episode—could further complicate causal pathways.

Combining mechanisms into a single causal conclusion for media use and mental health is beyond the scope of this project and likely any single study, given the heterogeneity observed across individuals. The participants in this study suggest that no universal explanation applies across diverse media behaviors, content, and time domains. Instead, multiple mechanisms, each tailored to specific contexts, are likely needed. Controlled experiments targeting distinct media experiences [71,72] and idiographic approaches like ours can help identify mechanisms for further exploration.

The within-person associations in this project emphasize the need for new theories addressing the accumulation of similar media experiences over time, regardless of specific content. Such mechanisms echo the concept of resonance in social reality theory [73,74], where outcomes arise from the alignment of digital and real-life environments. For instance, increased screen use during a depressive episode may reinforce social isolation, creating a feedback loop that intensifies mental health symptoms [75]. These findings highlight the importance of reexamining how repetitive media experiences influence mental health over extended periods.

Interventions

The p-technique canonical correlations observed in this project (Figure 1) suggest that results obtained for a participant over 1 year may generalize to future months. Acute changes in media use could serve as markers for triggering mental health interventions without requiring frequent or intrusive direct assessments. Interventions could integrate patient and clinician access to data through dashboards alerting users to mental health risks and offering tailored recommendations. For instance, systems might limit social app usage for some individuals or promote it for others based on identified risks. Person-specific thresholds for intervention provide a more accurate approach than generalized thresholds derived from population averages [76].

The key to these interventions is precision timing and targeted actions that connect continuous monitoring of screen activity with information designed to improve mental health. Screen-based systems might rival the diagnostic accuracy of traditional methods for some disorders. Studies estimate, for example, that up to 60% of people with depression are never diagnosed [77,78], and even those connected to care often face delays in appointments, diagnosis, and treatment initiation [79]. Screen summaries might accelerate care by offering behavioral insights prior to intake or as indicators of treatment effectiveness.

Determining the clinical use of idiographic tracking will require research to establish optimal baseline collection periods, which may vary by individual. While longer data collection can enhance accuracy, even shorter periods could yield valuable insights. For example, smartphone monitoring during the weeks between initial clinician contact and the first appointment, or in critical postdischarge periods, could inform preliminary evaluations and adaptive interventions. These monitoring systems could feed data into clinician dashboards and support the implementation of just-in-time adaptive interventions [80]. Importantly, such tools are intended to augment rather than replace direct patient-clinician interactions, which require robust support resources.

XSL•FO
RenderX

Limitations of the Study

This longitudinal study, conducted with a small, nonclinical sample of female Android phone users, was not intended to provide population-level conclusions about the prevalence of media and mental health relationships nationally or within subgroups. Avoiding generalization at this stage helps circumvent the small sample fallacy [81]. Our intent was to first understand the unique media dosage and mental health indicators important for individuals without prematurely investing in a research strategy aimed at extensive between-subject variability that we believe has resulted in considerable ambiguity in this literature. This bottom-up inductive approach celebrates idiosyncrasy and prioritizes individual models as a first step toward building models that might apply to larger groups of people. We have successfully demonstrated that it is possible to find strong, individualized relationships with granular measures and longitudinal data.

Our analysis focused on a limited set of variables: 6 media use metrics and 5 mental health measures. Many additional factors could influence within-person associations between media use and mental health. Notably, data collection coincided with the height of the COVID-19 pandemic. Although none of the participants contracted the virus, the pandemic's broader effects may have shaped their mental health and media use behaviors. Future analyses could examine how contextual factors, such as global or personal life events, moderate these relationships. For example, exploring interactions between media use, sleep patterns, and mental health outcomes [82,83], or identifying whether media use acts as a supportive resource [84,85] or an additional stressor [86,87] could inform context-aware models that better capture individual contingencies.

Future Work and Conclusions

While new computational methods enable the collection and analysis of large datasets, they often focus on aggregated measures, obscuring person-specific longitudinal insights and yielding ambiguous population-level conclusions. Our method emphasizes the importance of a person-specific approach applicable to new datasets. A key recommendation for extending this investigation is to maintain this individual focus and compare people based on similarities between models derived for single individuals, rather than relying solely on priori groupings based on demographics or other psychosocial characteristics. When idiographic models are available for many individuals, researchers can use clustering, latent class, and mixture models to quantify and examine between-person similarities and differences [88-90]. One such method suggested by Ram et al [88] involves calculating Tucker congruence coefficients [91] that compare parameter estimates (eg, the canonical variates) for each pair of individuals in a larger sample. Importantly, this approach entails specifying and modeling each person's data separately, as done in this study, before identifying shared patterns. Rather than assuming a priori commonalities (eg, based on sex or personality), this bottom-up paradigm uncovers the full range of person-specific solutions, which can then inform theory, policy, and interventions [18,88,92,93].

Future work will expand on the metrics presented here, integrating richer variables through our computational assay [20] and identifying screen content characteristics—such as words, objects, emotional arousal, and environmental features—using advanced artificial intelligence models. These granular data can help refine our understanding of how different media use patterns related to mental health. While traditional screen time measures, often derived from self-reports, have been criticized for their imprecision [94], this study suggests that accurate, granular recordings of screen behavior can reveal strong associations for some individuals. Similar increments in the precision of assessing media use have paid off in the study of other media effects; for example, exposure to political information [95], advertising [94], and food content in digital media [30]. Our ongoing work also includes open-sourcing the most up-to-date Screenomics app and toolset [96].

The idiographic filter approach, producing individual models 1 person at a time, may initially seem to risk overfocusing on small samples. However, there may not be a shorter route to useful answers about media use and mental health. This is especially true if we want to retain media use and mental health as important and psychologically complex categories of digital and non-digital life. The present results suggest that to advance research addressing media use and mental health problems generally, we may need to seek answers applicable to one person at a time; that is, we must embrace idiosyncrasy, not ignore it.

Acknowledgments

This project was partially funded by an internal Stanford University seed grant from the Stanford Medical School (Spectrum SPADA).

Data Availability

All deidentified data used and reported in this study, along with the code for analysis and visualizations, are available from the corresponding author upon reasonable request.

Authors' Contributions

TNR, NR, and BR did the inception and design of the study; MC, AYL, and BR drafted the manuscript and did data interpretation; NR, MC, and NH did data analysis and strategy; AM and JH advised on literature review, mental health assessments, and

XSL•FO
RenderX

interpretations; DM and AAGT organized data acquisition and data management. All authors have edited the manuscript in multiple versions and accept responsibility for the entire content of this manuscript.

Conflicts of Interest

DM is a founder of Screenlake Inc, a startup company focusing on smartphone data collection and analysis for the purpose of market intelligence. His involvement in the present study was the design and management of data collection during his doctoral studies, which took place prior to the founding of the company. JH consults for YouTube and Google on issues related to social media and receives compensation for these services. However, his contribution to this paper is solely his own work. The other authors declare no competing interests.

Multimedia Appendix 1
Supplementary Information.
[DOCX File , 31 KB - formative_v9i1e59875_app1.docx]

Multimedia Appendix 2
Permutation distribution histograms. This figure displays histograms of the permutation distributions of 1,000 permuted canonical correlations for each participant (A to E). To assess the statistical significance of the observed canonical correlations, we conducted permutation tests using Wilks' lambda by randomly permuting each individual's data, creating a null distribution. This null distribution represents correlations expected by chance—meaning the values that would arise if there were no systematic relationship between media use and mental health variables and any observed associations were purely random. The histograms display these chance-based distributions under the null hypothesis, with observed canonical correlations indicated by red dashed lines to evaluate their statistical significance. The test statistic shown in each histogram is Wilks' lambda. For Participant E, the result is nonsignificant ($P > 0.05$), suggesting that the observed correlation cannot be confirmed as statistically significant. For the other participants, the observed correlations are confirmed by the permutation test.
[PNG File , 5001 KB - formative_v9i1e59875_app2.png]

References

1. Sohn SY, Rees P, Wildridge B, Kalk NJ, Carter B. Prevalence of problematic smartphone usage and associated mental health outcomes amongst children and young people: a systematic review, meta-analysis and GRADE of the evidence. BMC Psychiatry 2019;19(1):356 [FREE Full text] [doi: 10.1186/s12888-019-2350-x] [Medline: 31779637]
2. Kushlev K, Leitao MR. The effects of smartphones on well-being: theoretical integration and research agenda. Curr Opin Psychol 2020;36:77-82. [doi: 10.1016/j.copsyc.2020.05.001] [Medline: 32563707]
3. Hancock J, Liu SX, Luo M, Mieczkowski H. Psychological well-being and social media use: a meta-analysis of associations between social media use and depression, anxiety, loneliness, eudaimonic, hedonic and social well-being. SSRN J 2022:84. [doi: 10.2139/ssrn.4053961]
4. Marciano L, Ostroumova M, Schulz PJ, Camerini AL. Digital media use and adolescents' mental health during the Covid-19 pandemic: a systematic review and meta-analysis. Front Public Health 2021;9:793868 [FREE Full text] [doi: 10.3389/fpubh.2021.793868] [Medline: 35186872]
5. Nawi A, Zameran FA. Exploring the relationship between smartphone use and adolescent well-being: a systematic literature review. In: Rajakumar G, Du KL, Vuppalapati C, Beligiannis GN, editors. Intelligent Communication Technologies and Virtual Mobile Networks. Singapore: Springer; 2023.
6. Harari GM, Müller SR, Aung MS, Rentfrow PJ. Smartphone sensing methods for studying behavior in everyday life. Curr Opin Behav Sci 2017;18:83-90. [doi: 10.1016/j.cobeha.2017.07.018]
7. Brinberg M, Bodie GD, Solomon DH, Jones SM, Ram N. Examining individual differences in how interaction behaviors change over time: a dyadic multinomial logistic growth modeling approach. Psychol Methods 2023. [doi: 10.1037/met0000605] [Medline: 37561491]
8. Mao C, Bayer JB, Ross MQ, Rhee L, Le HTK, Mount J, et al. Perceived vs. observed mHealth behavior: a naturalistic investigation of tracking apps and daily movement. Mobile Media Commun 2023;11(3):526-548. [doi: 10.1177/20501579221149823]
9. Barnett ML, Gonzalez A, Miranda J, Chavira DA, Lau AS. Mobilizing community health workers to address mental health disparities for underserved populations: a systematic review. Adm Policy Ment Health 2018;45(2):195-211 [FREE Full text] [doi: 10.1007/s10488-017-0815-0] [Medline: 28730278]
10. Chancellor S, De Choudhury M. Methods in predictive techniques for mental health status on social media: a critical review. NPJ Digit Med 2020;3:43 [FREE Full text] [doi: 10.1038/s41746-020-0233-7] [Medline: 32219184]
11. Torous J, Choudhury T, Barnett I, Keshavan M, Kane J. Smartphone relapse prediction in serious mental illness: a pathway towards personalized preventive care. World Psychiatry 2020;19(3):308-309 [FREE Full text] [doi: 10.1002/wps.20805] [Medline: 32931109]

XSL•FO
RenderX

12. Orben A, Przybylski AK, Blakemore SJ, Kievit RA. Windows of developmental sensitivity to social media. Nat Commun 2022;13(1):1649 [FREE Full text] [doi: 10.1038/s41467-022-29296-3] [Medline: 35347142]
13. Orben A, Blakemore SJ. How social media affects teen mental health: a missing link. Nature 2023;614(7948):410-412. [doi: 10.1038/d41586-023-00402-9] [Medline: 36788371]
14. Beyens I, Pouwels JL, van Driel II, Keijsers L, Valkenburg PM. Social media use and adolescents' well-being: developing a typology of person-specific effect patterns. Commun Res 2021;51(6):691-716. [doi: 10.1177/00936502211038196]
15. Valkenburg PM, Meier A, Beyens I. Social media use and its impact on adolescent mental health: an umbrella review of the evidence. Curr Opin Psychol 2022;44:58-68 [FREE Full text] [doi: 10.1016/j.copsyc.2021.08.017] [Medline: 34563980]
16. Bryant J, Oliver MB. Media Effects: Advances in Theory and Research. New York, NY: Routledge; 2009.
17. McLeod JM, Reeves B. On the nature of mass media effects. In: Withey SB, Abeles RP, editors. Television and Social Behavior: Beyond Violence and Children/A Report of the Committee on Television and Social Behavior, Social Science Research Council. Oxfordshire, United Kingdom: Routledge; 2013.
18. Nesselroade JR, Gerstorf D, Hardy SA, Ram N. Focus article: idiographic filters for psychological constructs. Meas Interdiscip Res Perspect 2007;5(4):217-235. [doi: 10.1080/15366360701741807]
19. Thompson B. Canonical Correlation Analysis: Uses and Interpretation. Washington, DC: Sage Publications; 1984.
20. Ram N, Haber N, Robinson TN, Reeves B. Binding the person-specific approach to modern AI in the human screenome project: moving past generalizability to transferability. Multivariate Behav Res 2024;59(6):1211-1219. [doi: 10.1080/00273171.2023.2229305] [Medline: 37439508]
21. Reeves B, Robinson T, Ram N. Time for the Human Screenome Project. Nature 2020;577(7790):314-317. [doi: 10.1038/d41586-020-00032-5] [Medline: 31942062]
22. Kessler RC, Adler L, Ames M, Demler O, Faraone S, Hiripi E, et al. The World Health Organization Adult ADHD Self-Report Scale (ASRS): a short screening scale for use in the general population. Psychol Med 2005;35(2):245-256. [doi: 10.1017/s0033291704002892] [Medline: 15841682]
23. Chen R, Mias GI, Li-Pook-Than J, Jiang L, Lam HYK, Chen R, et al. Personal omics profiling reveals dynamic molecular and medical phenotypes. Cell 2012;148(6):1293-1307 [FREE Full text] [doi: 10.1016/j.cell.2012.02.009] [Medline: 22424236]
24. Rozgonjuk D, Levine JC, Hall BJ, Elhai JD. The association between problematic smartphone use, depression and anxiety symptom severity, and objectively measured smartphone use over one week. Comput Hum Behav 2018;87:10-17. [doi: 10.1016/j.chb.2018.05.019]
25. Alloy LB, Uroševič S, Abramson LY, Jager-Hyman S, Nusslock R, Whitehouse WG, et al. Progression along the bipolar spectrum: a longitudinal study of predictors of conversion from bipolar spectrum conditions to bipolar I and II disorders. J Abnorm Psychol 2012;121(1):16-27 [FREE Full text] [doi: 10.1037/a0023973] [Medline: 21668080]
26. McInnis MG, Andreassen OA, Andreazza AC, Alon U, Berk M, Brister T, et al. Strategies and foundations for scientific discovery in longitudinal studies of bipolar disorder. Bipolar Disord 2022;24(5):499-508 [FREE Full text] [doi: 10.1111/bdi.13198] [Medline: 35244317]
27. Keller MB, Lavori PW, Mueller TI, Endicott J, Coryell W, Hirschfeld RM, et al. Time to recovery, chronicity, and levels of psychopathology in major depression. a 5-year prospective follow-up of 431 subjects. Arch Gen Psychiatry 1992;49(10):809-816. [doi: 10.1001/archpsyc.1992.01820100053010] [Medline: 1417434]
28. Crocetti E, Klimstra T, Keijsers L, Hale WW, Meeus W. Anxiety trajectories and identity development in adolescence: a five-wave longitudinal study. J Youth Adolesc 2009;38(6):839-849. [doi: 10.1007/s10964-008-9302-y] [Medline: 19636785]
29. Buecker S, Luhmann M, Haehner P, Bühler JL, Dapp LC, Luciano EC, et al. The development of subjective well-being across the life span: a meta-analytic review of longitudinal studies. Psychol Bull 2023;149(7-8):418-446. [doi: 10.1037/bul0000401]
30. Ram N, Yang X, Cho MJ, Brinberg M, Muirhead F, Reeves B, et al. Screenomics: a new approach for observing and studying individuals' digital lives. J Adolesc Res 2020;35(1):16-50 [FREE Full text] [doi: 10.1177/0743558419883362] [Medline: 32161431]
31. Humphreys L. The Qualified Self: Social Media and the Accounting of Everyday Life. Cambridge, MA: MIT Press; 2018.
32. Conway CC, Forbes MK, Forbush KT, Fried EI, Hallquist MN, Kotov R, et al. A hierarchical taxonomy of psychopathology can transform mental health research. Perspect Psychol Sci 2019;14(3):419-436 [FREE Full text] [doi: 10.1177/1745691618810696] [Medline: 30844330]
33. Nock MK, Kessler RC, Franklin JC. Risk factors for suicide ideation differ from those for the transition to suicide attempt: the importance of creativity, rigor, and urgency in suicide research. Clin Psychol Sci Pract 2016;23(1):31-34. [doi: 10.1037/h0101734]
34. Nock MK, Banaji MR. Prediction of suicide ideation and attempts among adolescents using a brief performance-based test. J Consult Clin Psychol 2007;75(5):707-715 [FREE Full text] [doi: 10.1037/0022-006X.75.5.707] [Medline: 17907852]
35. Robinaugh DJ, Brown ML, Losiewicz OM, Jones PJ, Marques L, Baker AW. Towards a precision psychiatry approach to anxiety disorders with ecological momentary assessment: the example of panic disorder. Gen Psychiatr 2020;33(1):e100161 [FREE Full text] [doi: 10.1136/gpsych-2019-100161] [Medline: 32175524]

XSL•FO
RenderX

36. Chahal R, Ho TC, Miller JG, Borchers LR, Gotlib IH. Sex-specific vulnerability to depressive symptoms across adolescence and during the COVID-19 pandemic: the role of the cingulum bundle. JCPP Adv 2022;2(1):e12061 [FREE Full text] [doi: 10.1002/jcv2.12061] [Medline: 35572852]
37. Hardoon DR, Szedmak S, Shawe-Taylor J. Canonical correlation analysis: an overview with application to learning methods. Neural Comput 2004;16(12):2639-2664. [doi: 10.1162/0899766042321814] [Medline: 15516276]
38. Orben A. Teenagers, screens and social media: a narrative review of reviews and key studies. Soc Psychiatry Psychiatr Epidemiol 2020;55(4):407-414. [doi: 10.1007/s00127-019-01825-4] [Medline: 31925481]
39. Schimmack U, Krause P, Wagner GG, Schupp J. Stability and change of well being: an experimentally enhanced latent state-trait-error analysis. Soc Indic Res 2009;95(1):19-31. [doi: 10.1007/s11205-009-9443-8]
40. Liu X, Ping S, Gao W. Changes in undergraduate students' psychological well-being as they experience university life. Int J Environ Res Public Health 2019;16(16):2864 [FREE Full text] [doi: 10.3390/ijerph16162864] [Medline: 31405114]
41. Odgers CL, Jensen MR. Annual research review: adolescent mental health in the digital age: facts, fears, and future directions. J Child Psychol Psychiatry 2020;61(3):336-348 [FREE Full text] [doi: 10.1111/jcpp.13190] [Medline: 31951670]
42. Yoon S, Kleinman M, Mertz J, Brannick M. Is social network site usage related to depression? A meta-analysis of Facebook-depression relations. J Affect Disord 2019;248:65-72. [doi: 10.1016/j.jad.2019.01.026] [Medline: 30711871]
43. Houben M, Van Den Noortgate W, Kuppens P. The relation between short-term emotion dynamics and psychological well-being: a meta-analysis. Psychol Bull 2015;141(4):901-930. [doi: 10.1037/a0038822] [Medline: 25822133]
44. Meier A, Reinecke L. Computer-mediated communication, social media, and mental health: a conceptual and empirical meta-review. Commun Res 2020;48(8):1182-1209. [doi: 10.1177/0093650220958224]
45. Vanden Abeele MMP, Halfmann A, Lee EWJ. Drug, demon, or donut? Theorizing the relationship between social media use, digital well-being and digital disconnection. Curr Opin Psychol 2022;45:101295. [doi: 10.1016/j.copsyc.2021.12.007] [Medline: 35123383]
46. Goodman LS, Gilman A. In: Brunton LL, Knollman BC, editors. Goodman and Gilman's The Pharmacological Basis of Therapeutics. New York, NY: McGraw Hill; 2022.
47. Schleidgen S, Friedrich O, Gerlek S, Assadi G, Seifert J. The concept of “interaction” in debates on human-machine interaction. Humanit Soc Sci Commun 2023;10(1). [doi: 10.1057/s41599-023-02060-8]
48. Orben A, Meier A, Dalgleish T, Blakemore SJ. Mechanisms linking social media use to adolescent mental health vulnerability. Nat Rev Psychol 2024;3(6):407-423. [doi: 10.1038/s44159-024-00307-y]
49. Lee AY, Katz R, Hancock J. The role of subjective construals on reporting and reasoning about social media use. Soc Media Soc 2021;7(3):205630512110353. [doi: 10.1177/20563051211035350]
50. Yeykelis L, Cummings JJ, Reeves B. The fragmentation of work, entertainment, e-mail, and news on a personal computer: motivational predictors of switching between media content. Media Psychol 2017;21(3):377-402. [doi: 10.1080/15213269.2017.1406805]
51. Reeves B, Ram N, Robinson TN, Cummings JJ, Giles CL, Pan J, et al. Screenomics: A framework to capture and analyze personal life experiences and the ways that technology shapes them. Hum Comput Interact 2021;36(2):150-201 [FREE Full text] [doi: 10.1080/07370024.2019.1578652] [Medline: 33867652]
52. Radloff L. The CES-D Scale: A self-report depression scale for research in the general population. Appl Psychol Meas 1977;1(3):385-401 [FREE Full text] [doi: 10.1177/014662167700100306]
53. Marteau TM, Bekker H. The development of a six-item short-form of the state scale of the Spielberger State-Trait Anxiety Inventory (STAI). Br J Clin Psychol 1992;31(3):301-306. [doi: 10.1111/j.2044-8260.1992.tb00997.x] [Medline: 1393159]
54. Wilkinson D. Screen time trends in the age of COVID-19. SimpleTexting. 2021. URL: https://simpletexting.com/blog/screen-time-survey [accessed 2023-10-22]
55. Flynn J. 20 Vital Smartphone Usage Statistics (2023): facts, data, and trends on mobile use in the U.S. Zippia.com. Zippia. 2023. URL: https://www.zippia.com/advice/smartphone-usage-statistics/ [accessed 2024-02-03]
56. Kerai A. Cell phone usage statistics: mornings are for notifications. Reviews.org. URL: https://www.reviews.org/mobile/cell-phone-addiction/ [accessed 2024-02-03]
57. Pezeshki A, Scharf LL, Azimi-Sadjadi MR, Lundberg M. Empirical canonical correlation analysis in subspaces. 2004 Presented at: Conference Record of the Thirty-Eighth Asilomar Conference on Signals, Systems and Computers,; November 7-10, 2004; Pacific Grove, CA p. 994-997. [doi: 10.1109/acssc.2004.1399288]
58. Julian LJ. Measures of anxiety: State-Trait Anxiety Inventory (STAI), Beck Anxiety Inventory (BAI), and Hospital Anxiety and Depression Scale-Anxiety (HADS-A). Arthritis Care Res 2011;63:S467-S472 [FREE Full text] [doi: 10.1002/acr.20561] [Medline: 22588767]
59. Vargas-Terrones M, Nagpal TS, Perales M, Prapavessis H, Mottola MF, Barakat R. Physical activity and prenatal depression: going beyond statistical significance by assessing the impact of reliable and clinical significant change. Psychol Med 2021;51(4):688-693. [doi: 10.1017/S0033291719003714] [Medline: 32102723]
60. Spielberger CD, Gorsuch RL, Lushene RE. State-Trait Anxiety Inventory (STAI): Test Manual for Form X. Washington, DC: Consulting Psychologists Press; 1968.
61. Adler L, Cohen J. Diagnosis and evaluation of adults with attention-deficit/hyperactivity disorder. Psychiatr Clin North Am 2004;27(2):187-201. [doi: 10.1016/j.psc.2003.12.003] [Medline: 15063992]

XSL•FO
RenderX

62. Kessler RC, Berglund P, Demler O, Jin R, Merikangas KR, Walters EE. Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the national comorbidity survey replication. Arch Gen Psychiatry 2005;62(6):593-602. [doi: 10.1001/archpsyc.62.6.593] [Medline: 15939837]
63. JoMingyu. Google Play scraper. Github. URL: https://github.com/JoMingyu/google-play-scraper [accessed 2024-02-05]
64. Rhee L, Bayer JB, Lee DS, Kuru O. Social by definition: how users define social platforms and why it matters. Telematics Inform 2021;59:101538. [doi: 10.1016/j.tele.2020.101538]
65. Hotelling H. Relations Between Two Sets of Variates. In: Kotz S, Johnson NL, editors. Breakthroughs in Statistics: Methodology and Distribution. New York, NY: Springer; 1992.
66. Weenink D. Canonical correlation analysis. 2003 Presented at: Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam ;25?99; 2003; University of Amsterdam p. 81-99 URL: https://uvafon.hum.uva.nl/archive/2003/2003-ifaproc25-Weenink.pdf
67. González I, Déjean S. Package CCA: Canonical correlation analysis. R Foundation for Statistical Computing. 2021. URL: https://cran.r-project.org/package=CCA [accessed 2024-04-04]
68. Braghieri L, Levy R, Makarin A. Social media and mental health. Am Econ Rev 2022;112(11):3660-3693. [doi: 10.1257/aer.20211218]
69. Müller SR, Peters H, Matz SC, Wang W, Harari GM. Investigating the relationships between mobility behaviours and indicators of subjective well–being using smartphone–based experience sampling and GPS tracking. Eur J Pers 2020;34(5):714-732. [doi: 10.1002/per.2262]
70. Johnson AL. Changes in mental health and treatment, 1997-2017. J Health Soc Behav 2021;62(1):53-68. [doi: 10.1177/0022146520984136] [Medline: 33480305]
71. Bailey ER, Matz SC, Youyou W, Iyengar SS. Authentic self-expression on social media is associated with greater subjective well-being. Nat Commun 2020;11(1):4889 [FREE Full text] [doi: 10.1038/s41467-020-18539-w] [Medline: 33024115]
72. Brailovskaia J, Delveaux J, John J, Wicker V, Noveski A, Kim S, et al. Finding the "sweet spot" of smartphone use: reduction or abstinence to increase well-being and healthy lifestyle?! An experimental intervention study. J Exp Psychol Appl 2023;29(1):149-161. [doi: 10.1037/xap0000430] [Medline: 35389685]
73. Gerbner G, Gross L, Signorielli N, Morgan M. Aging with television: images on television drama and conceptions of social reality. J Commun 1980;30(1):37-47. [doi: 10.1111/j.1460-2466.1980.tb01766.x] [Medline: 7372841]
74. Clark JL, Green MC. Self-fulfilling prophecies: perceived reality of online interaction drives expected outcomes of online communication. Personal Individual Differ 2018;133:73-76. [doi: 10.1016/j.paid.2017.08.031]
75. Boers E, Afzali MH, Newton N, Conrod P. Association of screen time and depression in adolescence. JAMA Pediatr 2019;173(9):853-859 [FREE Full text] [doi: 10.1001/jamapediatrics.2019.1759] [Medline: 31305878]
76. Molenaar PCM, Campbell CG. The new person-specific paradigm in psychology. Curr Dir Psychol Sci 2009;18(2):112-117. [doi: 10.1111/j.1467-8721.2009.01619.x]
77. Williams SZ, Chung GS, Muennig PA. Undiagnosed depression: a community diagnosis. SSM Popul Health 2017;3:633-638 [FREE Full text] [doi: 10.1016/j.ssmph.2017.07.012] [Medline: 29349251]
78. Handy A, Mangal R, Stead TS, Coffee RL, Ganti L. Prevalence and impact of diagnosed and undiagnosed depression in the United States. Cureus 2022;14(8):e28011 [FREE Full text] [doi: 10.7759/cureus.28011] [Medline: 36134073]
79. Boerema AM, Ten Have M, Kleiboer A, de Graaf R, Nuyen J, Cuijpers P, et al. Demographic and need factors of early, delayed and no mental health care use in major depression: a prospective study. BMC Psychiatry 2017;17(1):367 [FREE Full text] [doi: 10.1186/s12888-017-1531-8] [Medline: 29145820]
80. Nahum-Shani I, Smith SN, Spring BJ, Collins LM, Witkiewitz K, Tewari A, et al. Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. Ann Behav Med 2018;52(6):446-462 [FREE Full text] [doi: 10.1007/s12160-016-9830-8] [Medline: 27663578]
81. Poulton EC. Small sample fallacy. In: Behavioral Decision Theory: A New Approach. Cambridge, MA: Cambridge University Press; 1994.
82. Li X, Buxton OM, Lee S, Chang AM, Berger LM, Hale L. Sleep mediates the association between adolescent screen time and depressive symptoms. Sleep Med 2019;57:51-60 [FREE Full text] [doi: 10.1016/j.sleep.2019.01.029] [Medline: 30897456]
83. Guerrero MD, Barnes JD, Chaput JP, Tremblay MS. Screen time and problem behaviors in children: exploring the mediating role of sleep duration. Int J Behav Nutr Phys Act 2019;16(1):105 [FREE Full text] [doi: 10.1186/s12966-019-0862-x] [Medline: 31727084]
84. Zhang K, Kim K, Silverstein NM, Song Q, Burr JA. Social media communication and loneliness among older adults: the mediating roles of social support and social contact. Gerontologist 2021;61(6):888-896. [doi: 10.1093/geront/gnaa197] [Medline: 33284972]
85. Lisitsa E, Benjamin KS, Chun SK, Skalisky J, Hammond LE, Mezulis AH. Loneliness among young adults during COVID-19 pandemic: the mediational rolea of social media use and social support seeking. J Soc Clin Psychol 2020;39(8):708-726. [doi: 10.1521/jscp.2020.39.8.708]

XSL•FO
RenderX

86. Zhao N, Zhou G. Social media use and mental health during the COVID-19 pandemic: moderator role of disaster stressor and mediator role of negative affect. Appl Psychol Health Well Being 2020;12(4):1019-1038 [FREE Full text] [doi: 10.1111/aphw.12226] [Medline: 32945123]
87. Seo J, Lee CS, Lee YJ, Bhang SY, Lee D. The type of daily life stressors associated with social media use in adolescents with problematic internet/smartphone use. Psychiatry Investig 2021;18(3):241-248 [FREE Full text] [doi: 10.30773/pi.2020.0060] [Medline: 33735548]
88. Ram N, Brose A, Molenaar PCM. Dynamic factor analysis: modeling person-specific process. In: The Oxford Handbook of Quantitative Methods. New York, NY: Oxford University Press; 2013:441.
89. van der Nest G, Lima Passos V, Candel MJJM, van Breukelen GJP. An overview of mixture modelling for latent evolutions in longitudinal data: modelling approaches, fit statistics and software. Adv Life Course Res 2020;43:100323. [doi: 10.1016/j.alcr.2019.100323] [Medline: 36726256]
90. Gates KM, Lane ST, Varangis E, Giovanello K, Guskiewicz K. Unsupervised classification during time-series model building. Multivariate Behav Res 2017;52(2):129-148 [FREE Full text] [doi: 10.1080/00273171.2016.1256187] [Medline: 27925768]
91. Lorenzo-Seva U, ten Berge JMF. Tucker's congruence coefficient as a meaningful index of factor similarity. Methodology 2006;2(2):57-64. [doi: 10.1027/1614-2241.2.2.57]
92. Brose A, Ram N. Within-person factor analysis: Modeling how the individual fluctuates changes across time. In: Mehl MR, editor. Handbook of Research Methods for Studying Daily Life. New York, NY: Guilford Press; 2012:459.
93. Yang X, Ram N, Lougheed JP, Molenaar PCM, Hollenstein T. Adolescents' emotion system dynamics: network-based analysis of physiological and emotional experience. Dev Psychol 2019;55(9):1982-1993 [FREE Full text] [doi: 10.1037/dev0000690] [Medline: 31464499]
94. Lee J, Hamilton JT, Ram N, Roehrick K, Reeves B. The psychology of poverty and life online: natural experiments on the effects of smartphone payday loan ads on psychological stress. Inf Commun Soc 2022;26(14):2775-2796. [doi: 10.1080/1369118x.2022.2109982]
95. Muise D, Hosseinmardi H, Howland B, Mobius M, Rothschild D, Watts DJ. Quantifying partisan news diets in web and TV audiences. Sci Adv 2022;8(28):eabn0083 [FREE Full text] [doi: 10.1126/sciadv.abn0083] [Medline: 35857498]
96. Yee AZH, Yu R, Lim SS, Lim KH, Dinh TTA, Loh L, et al. ScreenLife Capture: an open-source and user-friendly framework for collecting screenomes from android smartphones. Behav Res Methods 2023;55(8):4068-4085 [FREE Full text] [doi: 10.3758/s13428-022-02006-z] [Medline: 36289177]

Abbreviations

ADHD: attention-deficit/hyperactivity disorder
CCA: canonical correlation analysis
HIPAA: Health Insurance Portability and Accountability Act
pCCA: p-technique canonical correlation analysis

Edited by A Mavragani; submitted 24.04.24; peer-reviewed by A AL-Asadi, T Seibers; comments to author 20.08.24; revised version received 15.10.24; accepted 17.12.24; published 26.02.25.

Please cite as:
Cerit M, Lee AY, Hancock J, Miner A, Cho MJ, Muise D, Garròn Torres AA, Haber N, Ram N, Robinson TN, Reeves B
Person-Specific Analyses of Smartphone Use and Mental Health: Intensive Longitudinal Study
JMIR Form Res 2025;9:e59875
URL: https://formative.jmir.org/2025/1/e59875
doi:10.2196/59875
PMID:39808832

©Merve Cerit, Angela Y Lee, Jeffrey Hancock, Adam Miner, Mu-Jung Cho, Daniel Muise, Anna-Angelina Garròn Torres, Nick Haber, Nilam Ram, Thomas N Robinson, Byron Reeves. Originally published in JMIR Formative Research (https://formative.jmir.org), 26.02.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Analyzing Patient Complaints in Web-Based Reviews of Private Hospitals in Selangor, Malaysia, Using Large Language Model–Assisted Content Analysis: Mixed Methods Study

Muhammad Hafiz Sulaiman[1,2*], MD, MS; Nora Muda[1*], PhD; Fatimah Abdul Razak[1*], PhD

[1]Department of Mathematical Sciences, Faculty of Science and Technology, National University of Malaysia, Bangi, Malaysia

[2]Quality Unit, Clinical Management Section, Hospital Sultan Idris Shah, Serdang, Malaysia

[*]all authors contributed equally

Corresponding Author:
Muhammad Hafiz Sulaiman, MD, MS
Department of Mathematical Sciences
Faculty of Science and Technology
National University of Malaysia
Lingkungan Ilmu
Bangi, 43600
Malaysia
Phone: 60 197763962
Email: hafizmuar@gmail.com

Abstract

Background: Large language model (LLM)–assisted content analysis (LACA) is a modification of traditional content analysis, leveraging the LLM to codevelop codebooks and automatically assign thematic codes to a web-based reviews dataset.

Objective: This study aims to develop and validate the use of LACA for analyzing hospital web-based reviews and to identify themes of issues from web-based reviews using this method.

Methods: Web-based reviews for 53 private hospitals in Selangor, Malaysia, were acquired. Fake reviews were filtered out using natural language processing and machine learning algorithms trained on yelp.com validated datasets. GPT-4o mini model application programming interface (API) was then applied to filter out reviews without any quality issues. In total, 200 of the remaining reviews were randomly extracted and fed into the GPT-4o mini model API to produce a codebook validated through parallel human-LLM coding to establish interrater reliability. The codebook was then used to code (label) all reviews in the dataset. The thematic codes were then summarized into themes using factor analysis to increase interpretability.

Results: A total of 14,938 web-based reviews were acquired, of which 1121 (9.3%) were fake, 1279 (12%) contained negative sentiments, and 9635 (88%) did not contain any negative sentiment. GPT-4o mini model subsequently inducted 41 thematic codes together with their definitions. Average human-GPT interrater reliability is perfect (κ=0.81). Factor analysis identified 6 interpretable latent factors: “Service and Communication Effectiveness,” “Clinical Care and Patient Experience,” “Facilities and Amenities Quality,” “Appointment and Patient Flow,” “Financial and Insurance Management,” and “Patient Rights and Accessibility.” The cumulative explained variance for the six factors is 0.74, and Cronbach α is between 0.88 and 0.97 (good and excellent) for all factors except factor 6 (0.61: questionable). The factors identified follow a global pattern of issues identified from the literature.

Conclusions: A data collection and processing pipeline consisting of Python Selenium, the GPT-4o mini model API, and a factor analysis module can support valid and reliable thematic analysis. Despite the potential for collection and information bias in web-based reviews, LACA of web-based reviews is cost-effective, time-efficient, and can be performed in real time, helping hospital managers develop hypotheses for further investigations promptly.

(JMIR Form Res 2025;9:e69075) doi:10.2196/69075

KEYWORDS

large language model; hospital quality; patient satisfaction; big data; web-based review

XSL•FO
RenderX

Introduction

Quality Improvement Activities in Hospitals

Getting feedback from patients and families is important to continuously improve patient care and ensure patient and family satisfaction within health care settings [1,2]. Patient and family satisfaction is crucial for repeat visits and the economic sustainability of the health care provider; therefore, management teams need to empathize with their patients' and families' understandings, feelings, and behaviors in order to thrive in a competitive health care market.

Web-Based Reviews

The use of traditional surveys like SERVQUAL to measure quality in Malaysian health care is documented by Butt and de Run [3], Aliman and Mohamad [4], and Abd et al [5]. Observations, formal interviews, and surveys such as SERVQUAL are standard methods for collecting feedback from patients and families, but all require a significant amount of time for data collection [6,7]. Since these methods are resource-intensive, obtaining unbiased results will require researchers to invest more time and money, otherwise, the number of respondents is limited, which could lead to low study power and false negative results. Additionally, ethnographic studies often fall short in their ability to observe patients before, during, and after hospital stays due to privacy concerns from both the clinicians' side and the patients' side [8].

The use of web-based reviews to gather feedback on patients' experiences and opinions can help hospital managers address the limitations mentioned above. Unlike ethnographic studies and interviews, web-based reviews by patients and families have no spatial limitations, meaning that patients and families can share their feelings, experiences, and opinions throughout their entire journey—before, during, and after their hospital stay or visit. Web-based reviews are also readily available on the internet, and with the aid of a large language model (LLM)–assisted content analysis (LACA), we can include a larger number of respondents, thereby reducing the risk of type II error (false negatives) caused by an insufficient sample size.

Ranard et al [9] mentioned the advantages of using web-based reviews for hospital quality improvements, including the diversity of domains reported in web-based reviews. Traditional surveys such as Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS) have fixed domains in which the questions are based and have origins that derive from 1995. Since patients' indications and experience for hospitalization have changed greatly, using a fixed set of surveys would set a barrier to fully understanding current patients' needs. His study found 12 web-based review domains not otherwise reflected in HCAHPS.

Rahim et al [10], in their paper, however, raised the need for health care organizations to change in accordance with Industrial Revolution 4.0 by using web-based reviews to understand patients' and families' interests, desires, and values. He too mentioned that the use of traditional surveys like HCAHPS and SERVQUAL is restrictive in the ways that these surveys are fixed, time-intensive, lengthy, fail to identify the causes of concern, and are subject to response and selection bias. The author suggested using web-based reviews, such as Facebook web-based reviews on hospitals' pages, as new sources for quality monitoring in hospitals and using supervised machine learning (ML) to train ML models to classify these reviews into SERVQUAL domains.

Publicly Available Web-Based Data

The exponential growth of digital communication channels has transformed health care feedback mechanisms in Malaysia and globally. Malaysia's internet adoption rate reached 96.8% in 2023 [11], with social media users aged 18 years and older exceeding 24.8 million, representing 99.8% of the population [12]. This digital transformation has generated substantial health care–related user-generated content.

While global data use was projected to grow from 33 ZB in 2018 to 175 ZB by 2025 [13], Malaysia's health care sector has seen its own surge in digital footprint, with an estimated 54.7% of all patients using the internet to search for health information [14] and 91.2% health care workers have good eHealth literacy [15]. Web-based data reflecting patients' experiences in Malaysian health care settings is increasingly available through various digital channels, including social media platforms, hospital review websites, health care forums, and patient blogs. The Health White Paper for Malaysia further emphasizes the importance of leveraging these digital data sources for health care quality improvement [16].

Study Designs

This study is exploratory in nature, and our research questions are as follows: (1) How to apply LACA satisfactorily on hospital web-based reviews? (2) What are the themes of issues identified by LACA on hospital web-based reviews? The purposes of this study are (1) to develop and recommend a method for analyzing hospital web-based reviews to serve as an alternative to direct observations or interviews on patients and families in hospital settings and (2) to then identify themes of current issues in private hospitals in Selangor, Malaysia. For these purposes, we developed two hypotheses: (1) LACA developed based on our methods produces satisfactory coding works equivalent to a human coder with Cohen $\kappa>0.80$; (2) the themes identified from factor analysis produce a Cronbach $\alpha>0.70$ on all factors with interpretable items.

The population of the study is patients or families who have posted their reviews on a web-based review platform for private hospitals in the state of Selangor, Malaysia, from January 1, 2023, to December 31, 2023. All 53 private hospitals in the state were included in this study. This population was chosen because researchers and experts involved in this study have good knowledge of local private hospitals as compared to hospitals located somewhere else, therefore, enabling them to contribute to the qualitative inputs needed in this research. This study includes all reviews posted on a web-based review platform for all private hospitals inside Selangor using universal sampling so that we could get as much diversity of reviews as possible, a point of advantage over traditional observations or interviews.

XSL•FO
RenderX

The web-based reviews are then filtered to exclude reviews that are not accompanied by any comments. Since web-based reviews are subject to manipulation by hospitals, we also exclude fake reviews using natural language processing (NLP) and ML algorithms to make sure that the results represent the real patients' suggestions, opinions, and experiences. Detailed explanations of this method are discussed below.

Methods

Thematic Analysis

Thematic analysis is a qualitative research method used to identify, analyze, and report patterns or themes within data. It begins with researchers immersing themselves in the data to gain a deep understanding, which involves repeatedly reading and reviewing the material. Following this, they generate initial codes by systematically tagging relevant sections of the data with short labels that capture key aspects. These codes are then organized into potential themes—broader patterns that reflect significant features related to the research question.

Researchers review and refine the themes created to ensure they accurately represent the data and fit together cohesively. Each theme is clearly defined and named, and the final step involves writing up the findings to provide a comprehensive interpretation of the data, weaving together the themes to offer insightful conclusions. Thematic analysis is valued for its flexibility and ability to uncover patterns within complex qualitative data. Most of the thematic analysis framework used in this study is based on Braun and Clarke [17].

Inductive Coding in Thematic Analysis

Inductive coding in thematic analysis is a process where researchers develop codes directly from the data, rather than applying predetermined categories or theoretical frameworks (which apply in deductive coding). It begins with a thorough examination of the data to gain an in-depth understanding. Researchers then identify significant segments of text and create codes based on the content and meaning of these segments. These initial codes are descriptive and reflect the language and concepts used by the participants.

As coding progresses, similar codes are grouped together to form broader themes, which emerge naturally from the data itself. This approach allows for a more grounded analysis, as themes are developed from the participants' perspectives rather than imposed by external theories. The themes are then reviewed and refined to ensure they accurately represent the data and provide a coherent interpretation. Inductive coding is particularly useful for exploring new research areas and gaining insights that are deeply rooted in the data [18].

LLM

Traditional methods of analyzing text data from surveys and reviews often pose significant challenges, including time-consuming manual processes and resource-intensive endeavors [19]. Moreover, the sheer volume and unstructured nature of textual data available on the internet further exacerbate the complexity of analyzing and extracting actionable insights using conventional methodologies. Given the overwhelming amount of data deriving from web-based review platforms, attention is increasingly turning toward automated content analysis instead of pure qualitative content analysis [20].

With the emergence of LLMs, such as the GPT-4o mini model, health care institutions now possess a powerful tool to navigate and extract valuable insights from the vast expanse of unstructured text data available on the web. Hassani et al [21] confirm the fact that text mining in big data analytics is emerging as a powerful tool for harnessing the power of unstructured textual data by analyzing it to extract new knowledge and identify significant patterns and correlations hidden in the data.

LLMs are sophisticated artificial intelligence (AI) models trained on large corpora of text data [22], enabling them to understand and generate human-like language with remarkable accuracy and fluency. GPT architecture incorporates attention mechanisms and feed-forward neural networks to predict the next word in a sequence, which has improved LLM functionality significantly [23]. Leveraging advanced NLP techniques, LLMs excel in tasks such as text summarization, thematic analysis [24], and sentiment analysis [25], making them well-suited for analyzing qualitative data in health care contexts.

LLM is increasingly experimented with in the health care industry. For example, recent studies were conducted to see the impact of LLM in drug discovery [26], extraction of medical notes [27], prediction of diagnosis-related groups [28], and diagnostics [29]. A study by Lin and Kuo [30] highlighted the huge opportunities to leverage LLM in clinical decision-making systems. The advancement of LLM is parallel to the world's movements toward using AI in health care. The Ministry of Health, Malaysia [31], for example, mentioned the use of AI and big data as essential tools for improved health care delivery in the near future. The Ministry of Health Malaysia also made digitalization, advanced data analytics, and AI as the main agenda for the country's health reformation [16].

The use of LLMs in quality improvement activities offers several notable advantages over traditional methodologies. First, LLMs enable the automated processing and analysis of large volumes of unstructured text data, significantly reducing the time and resources required for data collection and analysis. This scalability allows health care institutions to extract insights from diverse sources of patient feedback in a timely and efficient manner, facilitating rapid response to emerging trends or issues.

Additionally, LLMs can identify nuanced patterns, sentiments, and themes within diverse textual data, providing deeper insights into patients' perceptions of health care quality and identifying areas for improvement that may have been overlooked using manual methods.

Furthermore, the integration of LLM-driven insights into quality improvement initiatives has the potential to enhance the patient-centeredness of health care delivery. By capturing and analyzing patients' experiences, opinions, and preferences as expressed in their own words, health care institutions can gain a more comprehensive understanding of patient needs and priorities. This patient-centric approach enables tailored quality improvement interventions that address specific patient

XSL•FO
RenderX

concerns, ultimately leading to improved patient satisfaction and outcomes.

LACA

LACA is a term coined by Chew et al [32] that describes the use of LLMs to enhance and streamline qualitative content analysis. Researchers begin by inputting qualitative data—such as text from interviews or documents—into an LLM, which processes and summarizes the content to provide an initial understanding. The LLM aids in coding and categorizing the text by suggesting themes and patterns based on its advanced NLP capabilities. This helps in identifying and organizing key topics and underlying themes more efficiently [33].

LLMs offer several advantages in the coding process, starting with consistency and standardization. By using sophisticated algorithms, they apply uniform criteria across datasets, minimizing variability and subjectivity that can arise from different human coders. This ensures coding reliability, especially when working with large datasets, where human inconsistencies can become more pronounced. Such standardization is crucial for maintaining high-quality and reliable results [24,34].

Another key benefit of LLMs is their scalability. These models can process and analyze vast amounts of text data much more quickly than manual methods, addressing significant resource constraints in qualitative research. The ability to handle large datasets efficiently makes LLMs an ideal solution for large-scale projects, where traditional methods may not be feasible due to time and personnel limitations [21].

LLMs also improve efficiency by automating the coding process, reducing the time required for annotation compared to manual coding. This can lead to substantial time savings, although the exact extent of these savings still requires further empirical investigation. The automation of this process helps streamline workflows and allows researchers to focus on other aspects of their projects [35].

In addition to efficiency, LLMs contribute to error reduction by adhering to predefined coding rules. This minimizes inconsistencies often caused by human fatigue or differences in interpretation. While LLMs can reduce errors, careful validation against human coding is still necessary to ensure accuracy and reliability in the results [36,37].

The potential for replicability and cost-effectiveness is another strength of LLM-based coding. When using the same models and algorithms, LLMs can produce more consistent results, making it easier to replicate studies or analyses. However, long-term reproducibility remains an area of ongoing research, as models and datasets evolve over time [38,39]. In terms of cost-effectiveness, LLM-assisted coding can reduce the need for large human teams, particularly for large-scale projects. However, it is important to consider the initial implementation costs and the resources required for model training. Despite these costs, the long-term savings in human resources and time can make LLM-based coding a more economical solution for large research projects [40,41].

Finally, LLMs have the capability for complex pattern detection. Trained on extensive datasets, they can identify nuanced patterns that manual coding might overlook. While this ability enhances the depth of analysis, it also requires rigorous validation to ensure that the patterns detected are meaningful and accurate [42,43]. Each of these factors—consistency, scalability, efficiency, error reduction, replicability, cost-effectiveness, and complex pattern detection—demonstrates the potential of LLMs to significantly enhance coding processes in qualitative research.

LACA uses an LLM, and in this study, is specifically using a GPT in which prompts were transformed to trigger the model to generate responses, that is, codes or attribute labels. Other than this GPT method, latent Dirichlet allocation is widely used to classify documents into topics. GPT is advantageous over latent Dirichlet allocation because GPT is context-aware and able to directly produce textual descriptions [44]. In summary, LACA offers improved accuracy, efficiency, and scalability, making it a powerful tool for handling extensive and intricate content analysis tasks.

Ethical Considerations

As this study involves secondary analysis of publicly available web-based reviews, which were anonymized and deidentified during the results visualization stage, no ethical review/approval was needed, per the Malaysian Medical Review and Ethics Committee (MREC) Ministry of Health guidelines [45]. No personal or hospital identifiers were associated with the reviews in the final results. Data processing was done only by authorized researchers, while data filtering was performed to ensure the exclusion of fake reviews using NLP and ML algorithms, which were trained on publicly available datasets such as Yelp.com. Protections were implemented during the presentation of results to minimize any potential risk of identifying individuals or revealing sensitive information. The study methodology adheres to ethical standards concerning the use of publicly available web-based content for research purposes.

Flow of the Study

The flow of the study is shown in Figure 1.

XSL•FO
RenderX

Figure 1. Flow of research (summarized).

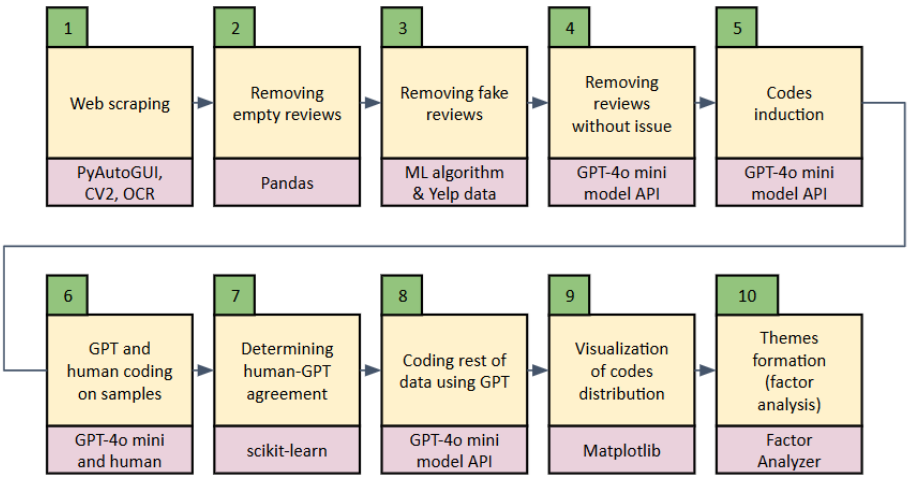

Web Scraping

We begin our research by getting a list of private hospitals in the state of Selangor sourced from the Ministry of Health Malaysia's website. The list was used to search for the hospital's name on Google Search. This allowed us to locate Google Reviews page for each of these hospitals. We develop a program using the Python PyAutoGUI module to automate data scraping. Ratings were extracted using computer vision (CV2) since star ratings in Google Reviews are presented in .jpg format and not in text format. Optical character recognition techniques were used to detect and locate the word "Newest" on tabs so that we could click the "Newest" button and ensure that the data were sorted from newest to oldest. There were many reviews written too long, so part of the reviews were hidden and readers had to click on the "more" button to reveal the hidden message. We automated the process of recognizing the "more" button by using CV2 and automatically expanding the text using the PyAutoGUI click function. The process of copying all text to the computer's memory is done using PyAutoGUI's select, scroll, and copy functions. The copied text was stored in a .txt file with a highly specific separator between reviews.

Removing Empty Reviews

Web-based reviews that are not followed by comments are removed, since we are doing qualitative data analysis on texts. This is done using the Python Pandas module by excluding documents with empty text.

Removing Fake Reviews

Since web-based reviews are susceptible to manipulation by individuals from the same institution (in the case of fake positive reviews) or competitors (in the case of fake negative reviews), these fake reviews do not represent real opinions, experiences, and suggestions. To do this, a program was developed based on previous studies [46-49] that incorporates NLP methodologies with ML algorithms. All data for training and testing the algorithm was acquired from yelp.com, a web-based review platform for hotels, restaurants, and hospitals, among others, that separates fake reviews from real reviews. The dataset was fed into a natural language preprocessing pipeline, which included the process of standardization, punctuation removal, numerical removal, tokenization, stop word removal, and formation of trigrams. Each trigram is now forming a single column in a matrix of term frequency-inverse document frequency, a vector matrix that becomes an input for a support vector machine and logistic regression ML algorithm. The ML model was trained and tested using the preprocessed yelp.com dataset and achieved a precision of 0.87, a recall of 0.89, a high F_1-score, and an accuracy of 0.88. Using the same NLP pipeline, each of our hospital review documents is preprocessed and then transformed into vector form (term frequency-inverse document frequency) before being fed into our validated ML model to filter out as many fake reviews as possible.

Removing Comments Without Issue

A comment can have good (positive) sentiment, bad (negative) sentiment, or can have both. This study focuses on identifying issues (negative sentiment) related to quality in hospitals. A previous study on bias in web-based reviews by Roh and Yang [50] showed that bad reviews are the most meaningful to help readers make decisions about a hospital. Other than that, we also tried to avoid having to make tuples for each label or code, that is, having to label or code ("communication," "negative") and ("communication," "positive") instead of just "communication" because tuples will produce double the number of variables, a phenomenon seen in the paper by Zaman et al [51]. A long list of variables also makes it more difficult during the theme formation (dimension reduction) phase. We maintain reviews containing negative sentiment and filter out the remaining by calling the GPT-4o mini model application programming interface (API) to respond "Yes" if a review contains issues and "No" otherwise. Full prompts are given in Multimedia Appendix 1.

Codes Induction

This stage begins with extracting issues present in each review. GPT-4o mini model API was used to identify, summarize, and list issues raised by each customer based on their web-based customer review. Then the next step involves randomly sampling 200 of these extracts to feed the GPT-4o mini model to produce lists of codes together with their definitions (codebook). A total of 5 iterations were performed to create a comprehensive list of codes using the GPT-4o mini model API calls. The final code list was reviewed by health care management experts and used

XSL•FO
RenderX

as a codebook to code issues in the dataset. Full prompts are provided in Multimedia Appendices 2 and 3.

GPT and Human Sample Coding

A random 200 reviews were selected. GPT-4o mini model API was tasked to use the codebook and label the 200 reviews. A human researcher was also tasked to label the same 200 reviews. Both GPT and human coders were instructed to code each review by iterating each item in the codebook and answer 0 if the item is not an issue in the review, 1 if the item is a small issue, 2 if the item is a moderate issue, 3 if the item is a serious issue, and 4 if the item is an extremely serious issue in the review (Multimedia Appendix 4). If there are n_i items or codes in the codebook, the number of API calls will be $200n_i$. Interrater reliability was evaluated between a human coder and a GPT coder across the 200 reviews or documents using Cohen κ, a statistical measure that accounts for chance agreement.

Determining GPT-Human Agreement

At this stage, we will have a matrix of $200 \times n_i$ where the rows are reviews or documents and the columns are item variables, *i*. Each cell contains integers from 0 (indicating no issue) to 4 (indicating extremely serious issue). Interrater reliability is a crucial aspect of ensuring the reliability of coding and categorization in our research. We do this by converting the scalar data type to binary (0 if no issue, and 1 if there is at least one issue related to the thematic code). Cohen κ was calculated between the GPT and the human coder for each document. The average κ score across the 200 documents was calculated to assess overall agreement. An average Cohen κ of more than 0.8 is acceptable to proceed to the next stage.

Coding the Rest of the Data Using GPT

As GPT coding tasks are proven to be reliable, the process of coding (described above) can then be continued solely by the GPT-4o mini model on the rest of our data. This way, we can label all our current data (and other data in the future) easily without the need for human coders. This is the advantage of our current method as compared to manually coding the web-based reviews.

Visualization of Codes Distribution

During this stage, we visualized the code distribution in our dataset, including the prevalence of each of the codes. Co-occurrences of codes are visualized through heat maps of correlation between individual codes. Pearson correlation was conducted to see the strength of the association between thematic codes.

Themes Formation by Factor Analysis

Thematic codes used to label the review were reduced to single-digit latent factors. Latent factors are essentially the underlying variables that explain the patterns of correlations among observed variables. The formation of single-digit themes is important so that we can focus our efforts on these themes. Although this process can be done manually through rearrangements of codes into themes, we automate the process so that any future analysis will also be done automatically for efficiency. The use of factor analysis to reduce attributes or codes into themes is documented by Sovacool [52]. We decided to include or exclude factors based on cumulative explained variance, Cronbach α [53], and qualitative assessment—content validity [54].

Overall, the methodology involves a systematic approach to collect, integrate, analyze, and interpret data from web-based reviews to understand the factors influencing demand for private hospitals in Selangor. Advanced techniques like ML and NLP were used to filter out fake reviews and extract meaningful insights from large datasets.

Results

Overview

Exactly 14,938 Google Review data points were scraped by our program, developed using a graphical user interface. These data include 53 private hospitals in the state of Selangor. The data collected consists of data from one year prior to the date the data were collected (January 2023-December 2023). Of the 14,938 data collected, a total of 12,035 (81%) evaluations were accompanied by comments, while 2903 (19%) evaluations were not accompanied by comments. Among all reviews with comments, 1121 (9.3%) reviews were fake and excluded from the data. There are 1279 evaluations that have issues (negative sentiments), and 9635 evaluations do not have issues (positive or neutral sentiments). The full list of thematic codes generated by LLM and its definition, that is, the codebook, can be accessed in Multimedia Appendix 5. An example of actual web-based reviews and the codes labeled to them by LLM is shown in Table 1 and more in Multimedia Appendix 4.

Cohen κ revealed a perfect level of agreement between the human and GPT coders. Across the 200 random reviews or documents, the lowest κ score between human and GPT coder was 0.44, and the highest κ score was 1.00. The average Cohen κ score was found to be exceptionally high (0.81), indicating a very strong level of agreement beyond what would be expected by chance. These results demonstrate that the GPT coder exhibited a high degree of consistency in its coding, with minimal discrepancies across the dataset. As GPT coding tasks are validated, the process of coding was then continued solely by the GPT-4o mini model to the rest of our 1279 data points (reviews), which gives us the following finding.

XSL•FO
RenderX

Table 1. Example of actual web-based reviews and the themes assigned by LLM[a].

No.	Actual web-based review	Codes assigned by LLM
1	Staff R****i very helpful but waiting time is too long[1]. 2 hours though. I'm patient no. 4.	1. Waiting time
2	Very friendly staff. We were the regular there since my new born daughter always went there check up and vaccines. It's fine. But when comes to serious illness, something emergency[2,3], they are really lack of experienced staff[1,4].	1. Work load 2. Emergency services 3. Patient safety and hygiene 4. Doctor's qualification and doctor's change
3	On 2nd Feb 2023 this S**a nurse said will check for me for an available appointment for this specialist doctor, whom I want to see. She didn't revert back to me at all[1,2,3,4]. I called again on 28th February this nurse A***a said my appointment was slot on 23rd March @ 5 pm so on 22nd March I call up to confirmed and they said my name was not in the system and I have been waited for almost 1 month. Nurse M*****h help me to rebooked but didn't inform me that the date has been postponed to 24th March[1,2] instead she told me is 3 pm so I thought is in 23rd March at 3 pm. Such a private hospital so incompetent and inefficient[4] the nurses here. I just wants to make an appointment as this doctor specialist is always full. Simple tasks can't do it well. How you expect people will come to this hospital?	1. Communication 2. Staff responsiveness 3. Staff attitude 4. Inefficient and disorganized processes

[a]LLM: large language model.

Distribution of Themes Found From Web-Based Reviews

The most common themes based on web-based reviews include the themes of "Service Quality and Professionalism" (n=511), "Communication" (n=506), "Waiting Time" (n=382), "Staff Attitude" (n=284), and the theme of "Responsiveness" (n=192). The least common themes based on web-based reviews include the themes of "Workload" (n=9), "Insurance Billing Errors" (n=9), "Communication Barriers" (n=10), "Comparison Between Races" (n=11), and the theme of "Support for Breastfeeding Mothers" (n=11). The complete list of distributions can be seen in the distribution in Multimedia Appendix 6.

Correlation Between Thematic Codes

The correlation was done between 2 thematic codes at a time to observe the relationship between them (please refer to Multimedia Appendix 5 for codes definition). We used Pearson correlation, and our analysis found correlation ranges from –0.0 to 0.9 between each thematic code (Figure 2). The correlation between thematic codes is shown in the correlation heat map below. After the Kaiser-Meyer-Olkin test (0.98) and the Bartlett test of sphericity ($P<.05$), we accepted our data size as adequate and the code matrix as significantly different from the identity matrix and proceeded to the next step—factor analysis.

The heat map shows the correlation between thematic codes. As we can see, most thematic codes have a majority of red boxes, indicating a high correlation with most of the other thematic codes. There are, however, thematic codes that have the majority of blue boxes indicating low correlation with most of the other thematic codes, including these thematic codes: Financial Concerns, Facility Maintenance, Cafeteria and Facilities, Room and Amenities, Patient Privacy, Accessibility for Individuals with Disabilities, Breastfeeding Support, Lift and Equipment Issues, and Insurance Billing Mistakes. The heat map also shows some columns having similar patterns (of red and blue boxes), for example, columns 26, 27, 31, and 33, indicating that they measure an item inside the same theme and should probably be included under the same latent factor when factor analysis is done.

XSL•FO
RenderX

Figure 2. Heat map of correlations between thematic codes.

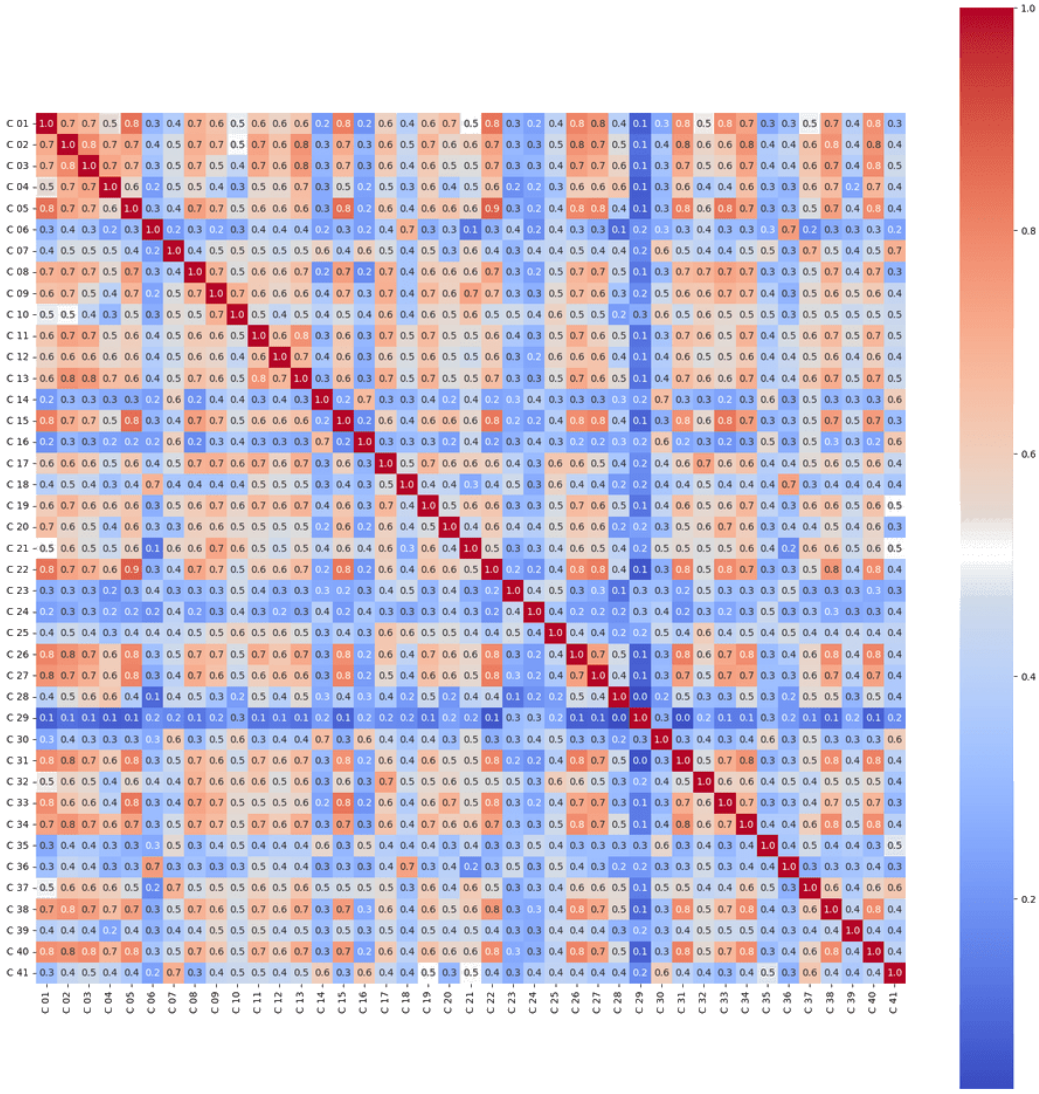

Factor Analysis

We were using the Factor Analyzer module from Python to derive latent factors from our variables. We determined the appropriate number of latent factors based on eigenvalues and the scree test [55]. Based on the total number of factors with eigenvalues more than 1, the total number of factors to include is 6. Using the scree test, where a straight line is drawn from the least eigenvalue to the largest eigenvalue, the suggested number of factors is also 6 (Figure 3). The Cumulative Explained Variance was 0.74 for the first 6 factors (factors with eigenvalues more than 1). For the full list of cumulative explained variance, please refer to Multimedia Appendix 7. Latent factors, together with their suggested factor labels, codes, factor loadings, and Cronbach α, are summarized in Table 2.

Figure 3. Scree plot of eigenvalue against factor number.

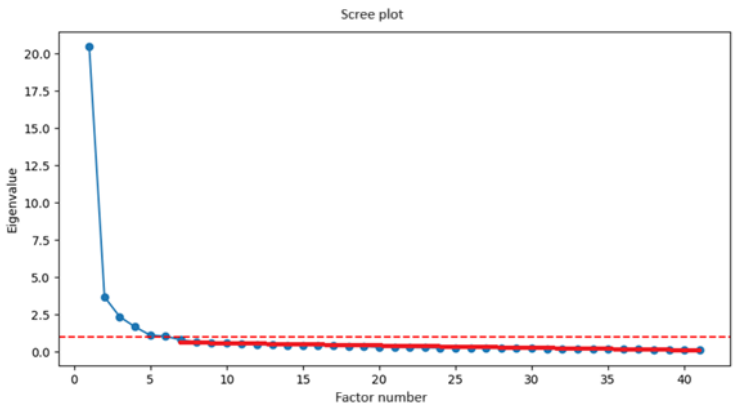

XSL•FO
RenderX

Table 2. Latent factors together with their suggested factor labels, codes, factor loadings, and Cronbach α.

Code and code names	Factor loading	Cronbach α
Factor 1: Service and Communication Effectiveness		0.97
C01: Communication Issues	0.95	
C05: Service Quality and Professionalism	1.00	
C08: Doctor's Behavior and Tardiness	0.51	
C15: Customer Service Training	0.97	
C20: Language and Communication Barriers	0.60	
C22: Staff Responsiveness	0.98	
C26: Departmental Coordination	0.71	
C27: Front Desk Service	0.85	
C31: Organizational Efficiency	0.80	
C33: Staff Attitude	0.98	
Factor 2: Clinical Care and Patient Experience		0.92
C09: Nursing Care	0.77	
C10: Patient Safety and Hygiene	0.64	
C11: Electronic Health Info Management	0.47	
C17: Diagnosis and Treatment	0.86	
C21: Patient Rest and Comfort	0.53	
C25: Medication Issues	0.65	
C32: Doctor Changes and Qualifications	0.72	
Factor 3: Facilities and Amenities Quality		0.90
C07: Facility Maintenance	0.61	
C14: Cafeteria and Facilities	0.89	
C16: Room and Amenities	0.87	
C30: Hospital Facilities and Food Quality	0.71	
C35: Lift and Equipment Issues	0.46	
C41: Amenities Adequacy	0.70	
Factor 4: Appointment and Patient Flow		0.89
C03: Appointment System Inconsistency	0.42	
C04: Waiting Time	0.82	
C13: Test and Result Processing	0.50	
C28: Punctuality	0.76	
Factor 5: Financial and Insurance Management		0.88
C06: Financial Concerns	0.89	
C18: Insurance Issues	0.76	
C36: Insurance Billing Mistakes	0.83	
Factor 6: Patient Rights and Accessibility		0.61
C23: Patient Privacy	0.50	
C24: Accessibility for Individuals with Disabilities	0.71	
C29: Breastfeeding Support	0.45	

XSL•FO
RenderX

Discussion

Principal Findings

The analysis of review content identified a substantial portion of evaluations (n=12,035, 81%) accompanied by comments, with 19% (n=2903) lacking comments. Of the reviews with comments, 9.3% (n=1121) were deemed fake and excluded, leaving 9594 evaluations without issues and 1279 with issues. The use of the GPT-4o mini model for coding these reviews showed high interrater reliability, with an average Cohen κ score of 0.81, indicating strong agreement between human and AI coders. This high level of consistency supports the validity of the coding process and the reliability of the insights derived from the data.

Granularity Issues

We acknowledge the lack of granularity of the codes produced by the GPT-4o mini model, that is, some of the codes are subsets of the other, and some of the codes contain more than 1 topic, making it nongranular. We noticed these results in the early phase of our research and decided to move forward with these codes. The first reason was that, in this study, we used inductive coding rather than deductive coding. Inductive coding is when codes and codebooks are created from scratch based on the available data (reviews), while deductive coding, on the other hand, is developing codebooks and coding the text using established concepts or theories. We did inductive coding and intentionally gave the freedom to the LLM to decide on the codes during codebook formation.

The second reason was that LLM was trained on a vast body of knowledge, which is not limited to just Medicine and Management. This includes Politics, Economics, Social, Technology, Legal, Environment, and Demographic domains (and this list is not exhaustive). Thus, the knowledge of LLM is not restricted to one or two theories only. Since the knowledge of LLM is cross-disciplinary, there are possibilities that we humans could not see the relevance of a code that is understandable only to the AI (LLM). This phenomenon is called the black box problem, in which mechanisms leading to an outcome are difficult to explain to humans but are effective in determining the outcome [56].

Despite the granularity issues, LLM has proven that the whole process is effective and reliable. Despite some of the codes looking nongranular to us humans, the LLM effectively uses the codes to label each review consistently across our data. For that reason, we were able to get excellent Cronbach α in almost all factors during factor analysis. If the lack of granularity causes confusion to the LLM itself, the coding phase will be inconsistent or chaotic and will then result in poor factor analysis. This is proof that the producer of the codes (the LLM itself) does understand the codes precisely, despite some of them looking nongranular to us humans.

Other than that, the freedom given to LLM to decide codes without any human intervention will also make it possible for us to replicate the system in the future without human intervention using the same pipeline. Thus, this kind of analysis can be adapted instantaneously to any industry and any web-based review platform that holds web-based review data. High replicability will make this study's impact greater.

Factor Analysis

Factor analysis identified 6 interpretable latent factors: "Service and Communication Effectiveness," "Clinical Care and Patient Experience," "Facilities and Amenities Quality," "Appointment and Patient Flow," "Financial and Insurance Management," and "Patient Rights and Accessibility." These factors encompass the key areas influencing patient satisfaction, as reflected in the items and their factor loadings. The cumulative explained variance for the seven variables is 0.74.

As we add (or remove) factors, we face a tradeoff between cumulative explained variance, reliability measured by Cronbach α, and interpretability. Adding more factors increases the cumulative explained variance but compromises Cronbach α on the additional factors and their interpretability. Removing factors, on the other hand, maintains Cronbach α high on all remaining factors together with interpretability, but reduces cumulative explained variance. Since any trade-off requires a decision maker to decide based on preference [57], we find it helpful to explain our preference below.

We prefer interpretable factors over higher explained variance because, unlike ML algorithms—for example, artificial neural networks, which can process the entire dataset to predict outcomes without fully understanding the underlying mechanisms—we humans emphasize understanding the factors behind observed variables as a base for future studies. The need for us to understand the mechanisms behind predictions is justified by Vamathevan et al [58], who raised concerns about the black box phenomenon in ML algorithms, how it leads to a lack of mechanism understanding, and mistrust. The cumulative variance of 0.74 for our 6 factors is considered very good since this study is behavioral, as noted by Williams et al [55].

Implications

The themes we identified using web-based reviews, GPT-4o mini model, and factor analysis show strong alignment with global patterns identified in previous research. The systematic review of health care quality literature by Ferreira et al [59] revealed similar priorities in patient satisfaction assessment. Their analysis showed medical care appearing in 34% of studies, communication in 31%, doctor's characteristics in 28%, accommodations in 23%, admission and discharge in 13%, nurse characteristics in 11%, appointment in 8%, environment in 8%, medical expenditure in 8%, and organization in 8%. The prominence of service quality and communication factors in our analysis mirrors these global patterns, though with some notable differences. For instance, our findings suggest a higher emphasis on appointment and waiting time issues in the Malaysian private health care context.

The high reliability of Service and Communication Effectiveness (α=0.97), coupled with strong factor loadings (0.51-1.00), demonstrates how tightly interconnected communication and service quality are in health care delivery. The strong correlation between communication issues and staff responsiveness (r=0.78) suggests that these aspects cannot be effectively addressed in

XSL•FO
RenderX

isolation. This aligns with previous findings about the integrated nature of health care service quality (Ferreira et al [59]).

The emergence of distinct factors for both operational aspects (Appointment and Patient Flow, α=0.89) and support services (Financial and Insurance Management, α=0.88) indicates that patient satisfaction in private hospitals extends beyond clinical care. The moderate correlation between waiting time and staff attitude (r=0.41) reveals how operational inefficiencies can impact interpersonal aspects of care delivery. The questionable reliability of the Patient Rights and Accessibility factor (α=0.61) points to potential measurement challenges in these areas, suggesting a need for refined assessment tools. This finding reflects the broader challenge in health care research of quantifying and standardizing measurements of patient rights and accessibility (Williams et al [55]).

Limitations

Despite providing valuable insights into patient experiences, this study has several important limitations that should be considered when interpreting the results. First, the reliance on web-based reviews as the primary data source may introduce sampling bias, as this method potentially excludes perspectives from patients who are less likely to share feedback on web-based review platforms, particularly older patients or those with limited digital access. While more granular categorization could have been achieved through human modulation, which might have improved the specificity and interpretability of our analysis, it also limits the true potential of LLMs, which are trained on vast bodies of knowledge. Future research could compare LACA outcomes based on codebooks entirely produced by LLMs versus those modulated by humans.

Second, the geographical scope of this study, being limited to private hospitals in Selangor, Malaysia, may not fully represent patient experiences in other regions or health care settings. This regional focus, while providing depth in local context, limits the generalizability of our findings to other geographical areas or health care systems. Third, variations in review platforms and user demographics could affect the comprehensiveness of our findings. Different web-based review platforms may attract distinct user demographics and encourage varying styles of feedback, potentially skewing our understanding of patient experiences [9]. Future research should address these limitations by incorporating multiple data sources, expanding geographical coverage, and considering diverse patient populations to provide a more comprehensive understanding of hospital quality issues.

Conclusions

This study underscores the effectiveness of LACA for processing large-scale patient feedback, achieving high reliability (κ=0.81) between human and automated content analysis. The identification of six distinct factors explaining 74% of the variance provides a structured framework for understanding patient satisfaction in private hospitals.

Our findings suggest several actionable improvements for hospital management: (1) investing in integrated staff training programs focusing on communication skills and service delivery, given the strong factor loadings in these areas; (2) implementing advanced appointment systems and patient tracking technologies to address waiting time concerns, which appeared in 29.9% of negative reviews; and (3) developing integrated financial service units to handle billing, insurance, and payment issues cohesively, addressing the high-severity impact of these issues when they occur.

For policy makers, our results indicate the need for more structured guidelines in patient rights and accessibility standards. Health care regulators might consider developing comprehensive frameworks that address these aspects more systematically. Future research should expand this analysis to other geographical regions and health care contexts. Additionally, investigating the long-term impact of implementing LACA on web-based reviews could provide valuable insights into their effectiveness for continuous quality improvement in health care settings.

Acknowledgments

The project is self-funded by the researchers.

Data Availability

The datasets analyzed during the factor analysis phase of this study are available in Google Drive [60]. The datasets generated during the large language model–assisted content analysis phase of this study are not made publicly available because they may contain sensitive information (name of hospitals or doctor) but are available from the corresponding author upon reasonable request or from public web-based review platforms at any time.

Authors' Contributions

MHS contributed to data scraping, programming, data analysis, and initial manuscript writing. NM and FAR contributed to supervising the project, methodology, statistical analysis, and editing of the paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1

XSL•FO
RenderX

API (application programming interface) call for detecting presence of issue(s) in an online review.
[DOCX File , 16 KB - formative_v9i1e69075_app1.docx]

Multimedia Appendix 2
API (application programming interface) call for induction of thematic codes from a list of 200 random reviews.
[DOCX File , 16 KB - formative_v9i1e69075_app2.docx]

Multimedia Appendix 3
API (application programming interface) calls for justifying inclusion of each thematic code to develop a codebook.
[DOCX File , 16 KB - formative_v9i1e69075_app3.docx]

Multimedia Appendix 4
Sample of actual online comments and the thematic codes assigned by the large language model.
[DOCX File , 24 KB - formative_v9i1e69075_app4.docx]

Multimedia Appendix 5
Codebook for guiding content analysis.
[DOCX File , 27 KB - formative_v9i1e69075_app5.docx]

Multimedia Appendix 6
Frequency of mentions of each thematic code.
[XLSX File (Microsoft Excel File), 10 KB - formative_v9i1e69075_app6.xlsx]

Multimedia Appendix 7
Full list of cumulative explained variance from factor analysis.
[DOCX File , 15 KB - formative_v9i1e69075_app7.docx]

References

1. Lambert M, Shimokawa K. Collecting Client Feedback. Washington DC: Americal Psychological Association; 2016:203-223.
2. Gondek D, Edbrooke-Childs J, Fink E, Deighton J, Wolpert M. Feedback from outcome measures and treatment effectiveness, treatment efficiency, and collaborative practice: a systematic review. Adm Policy Ment Health 2016;43(3):325-343 [FREE Full text] [doi: 10.1007/s10488-015-0710-5] [Medline: 26744316]
3. Butt MM, de Run EC. Private healthcare quality: applying a SERVQUAL model. Int J Health Care Qual Assur 2010;23(7):658-673 [FREE Full text] [doi: 10.1108/09526861011071580] [Medline: 21125961]
4. Aliman NK, Mohamad WN. Linking service quality, patients' satisfaction and behavioral intentions: an investigation on private healthcare in Malaysia. Procedia Soc Behav Sci 2016;224:141-148. [doi: 10.1016/j.sbspro.2016.05.419]
5. Abd RM, Mansor A, Hamzah M. Service quality and patients' satisfaction in healthcare service in Malaysia. Int J Customer Serv Manage 2011;1(1):41-49 [FREE Full text]
6. Greaves F, Laverty AA, Cano DR, Moilanen K, Pulman S, Darzi A, et al. Tweets about hospital quality: a mixed methods study. BMJ Qual Saf 2014;23(10):838-846 [FREE Full text] [doi: 10.1136/bmjqs-2014-002875] [Medline: 24748372]
7. Hawkins JB, Brownstein JS, Tuli G, Runels T, Broecker K, Nsoesie EO, et al. Measuring patient-perceived quality of care in US hospitals using Twitter. BMJ Qual Saf 2016;25(6):404-413 [FREE Full text] [doi: 10.1136/bmjqs-2015-004309] [Medline: 26464518]
8. Lee YU, Chung SH, Park JY. Online review analysis from a customer behavior observation perspective for product development. Sustainability 2024;16(9):3550. [doi: 10.3390/su16093550]
9. Ranard BL, Werner RM, Antanavicius T, Schwartz HA, Smith RJ, Meisel ZF, et al. Yelp reviews Of hospital care can supplement and inform traditional surveys of the patient experience of care. Health Aff (Millwood) 2016;35(4):697-705 [FREE Full text] [doi: 10.1377/hlthaff.2015.1030] [Medline: 27044971]
10. Rahim AIA, Ibrahim MI, Musa KI, Chua S, Yaacob NM. Patient satisfaction and hospital quality of care evaluation in Malaysia using SERVQUAL and facebook. Healthcare (Basel) 2021;9(10):1369 [FREE Full text] [doi: 10.3390/healthcare9101369] [Medline: 34683050]
11. Malaysia ranks 3rd in world for spending most time online. New Straits Times. URL: https://tinyurl.com/5c2cvhhb [accessed 2025-05-13]
12. Digital 2023: Malaysia. datareportal. 2023. URL: https://datareportal.com/reports/digital-2023-malaysia [accessed 2025-05-13]
13. 175 zettabytes By 2025. Forbes. URL: https://www.forbes.com/sites/tomcoughlin/2018/11/27/175-zettabytes-by-2025/ [accessed 2025-06-02]

XSL•FO
RenderX

14. Lim H, Wong SS, Yip KC, Chang FWS, Chin AJZ, Teo CH, et al. Online health information-seeking behaviour of patients attending a primary care clinic in Malaysia: a cross-sectional study. Fam Pract 2022;39(1):38-45. [doi: 10.1093/fampra/cmab099] [Medline: 34423368]
15. Nor NAM, Zaberi ZHS, Nazari NSM. Internet use, online health information seeking behavior and electronic health literacy among dental auxiliary personnel in Malaysia. The Southeast Asian Journal of Tropical Medicine and Public Health. URL: https://journal.seameotropmednetwork.org/index.php/jtropmed/article/view/667 [accessed 2025-05-13]
16. Health White Paper for Malaysia. Malaysia: Ministry of Health Malaysia; 2023.
17. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
18. Frith H, Gleeson K. Qualitative data collection: asking the right questions. In: Qualitative Research Methods in Mental Health and Psychotherapy: A Guide for Students and Practitioners. Hoboken, New Jersey: Wiley; 2011:55-67.
19. Merriam SB, Tisdell EJ. Qualitative Research: A Guide to Design and Implementation. Hoboken, New Jersey: John Wiley & Sons; 2015.
20. Tuan A, Grandi S. Emerging trends in qualitative research: a focus on Social Media. Mercati Competitività 2018;4:17-26. [doi: 10.3280/mc2018-004002]
21. Hassani H, Beneki C, Unger S, Mazinani MT, Yeganegi MR. Text mining in big data analytics. BDCC 2020;4(1):1. [doi: 10.3390/bdcc4010001]
22. Stammbach D, Antoniak M, Ash E. Heroes, villains, and victims, and GPT-3: automated extraction of character roles without training data. 2022 Presented at: Proceedings of the 4th Workshop of Narrative Understanding (WNU2022); July, 2022; Seattle, United States p. 47-56. [doi: 10.18653/v1/2022.wnu-1.6]
23. Vaswani A. Attention is all you need. 2017 Presented at: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems; December 4-9, 2017; Long Beach California, USA p. 6000-6010 URL: https://cir.nii.ac.jp/crid/1370849946232757637
24. Xiao Z, Yuan X, Liao QV, Abdelghani R, Oudeyer PY. Supporting qualitative analysis with large language models: combining codebook with GPT-3 for deductive coding. 2023 Presented at: IUI '23 Companion: Companion Proceedings of the 28th International Conference on Intelligent User Interfaces; March 27-31, 2023; Sydney NSW, Australia p. 75-78. [doi: 10.1145/3581754.3584136]
25. Lubis AR. Balancing the equation: investigating AI advantages, challenges, and ethical considerations in the context of GPT-3, natural language processing, and researcher roles. SAR J 2023;6(4):257-262. [doi: 10.18421/sar64-05]
26. Oniani D, Hilsman J, Zang C, Wang J, Cai L, Zawala J, et al. Emerging opportunities of using large language models for translation between drug molecules and indications. Sci Rep 2024;14(1):10738 [FREE Full text] [doi: 10.1038/s41598-024-61124-0] [Medline: 38730226]
27. Chiang CC, Luo M, Dumkrieger G, Trivedi S, Chen YC, Chao CJ, et al. A large language model-based generative natural language processing framework fine-tuned on clinical notes accurately extracts headache frequency from electronic health records. Headache 2024;64(4):400-409. [doi: 10.1111/head.14702] [Medline: 38525734]
28. Wang H, Gao C, Dantona C, Hull B, Sun J. DRG-LLaMA : tuning LLaMA model to predict diagnosis-related group for hospitalized patients. NPJ Digital Med 2024;7(1):16 [FREE Full text] [doi: 10.1038/s41746-023-00989-3] [Medline: 38253711]
29. Gandomi A, Wu P, Clement DR, Xing J, Aviv R, Federbush M, et al. ARDSFlag: an NLP/machine learning algorithm to visualize and detect high-probability ARDS admissions independent of provider recognition and billing codes. BMC Med Inform Decis Mak 2024;24(1):195 [FREE Full text] [doi: 10.1186/s12911-024-02573-5] [Medline: 39014417]
30. Lin C, Kuo CF. Roles and potential of large language models in healthcare: a comprehensive review. Biomed J 2025 Apr 29:100868 [FREE Full text] [doi: 10.1016/j.bj.2025.100868] [Medline: 40311872]
31. Regional Action Agenda on Harnessing e-Health for Improved Health Service Delivery in the Western Pacific. Geneva: World Health Organization; 2019.
32. Chew R, Bollenbacher J, Wenger M, Speer J, Kim A. LLM-assisted content analysis: Using large language models to support deductive coding. ArXiv. Preprint posted online on June 23, 2023 2023. [doi: 10.48550/arXiv.2306.14924]
33. Khan AH, Kegalle H, D'Silva R, Watt N, Whelan-Shamy D, Ghahremanlou L. Automating thematic analysis: how LLMs analyse controversial topics. ArXiv. Preprint posted online on May 11, 2024 2024. [doi: 10.48550/arXiv.2405.06919]
34. Lewis SC, Zamith R, Hermida A. Content analysis in an era of big data: a hybrid approach to computational and manual methods. J Broadcast Electron Media 2013;57(1):34-52. [doi: 10.1080/08838151.2012.761702]
35. King N, Brooks J. Thematic analysis in organisational research. In: The SAGE Handbook of Qualitative Business and Management Research Methods: Methods and Challenges. Thousand Oaks CA: Sage Publications; 2018:219-236.
36. Krippendorff K. Content Analysis: An Introduction to Its Methodology. Thousand Oaks CA: Sage Publications; 2018.
37. McDonald N, Schoenebeck S, Forte A. Reliability and inter-rater reliability in qualitative research. Proc ACM Hum-Comput Interact 2019;3(CSCW):1-23. [doi: 10.1145/3359174]
38. O'Connor C, Joffe H. Intercoder reliability in qualitative research: debates and practical guidelines. Int J Qual Methods 2020;19:1609406919899220. [doi: 10.1177/1609406919899220]

XSL•FO
RenderX

39. Roberts ME, Stewart BM, Tingley D. Stm: an R package for structural topic models. J Stat Software 2019;91(2):1-40. [doi: 10.18637/jss.v091.i02]
40. Shah DV, Cappella JN, Neuman WR. Big data, digital media, and computational social science. Ann Am Acad Polit Soc Sci 2015;659(1):6-13. [doi: 10.1177/0002716215572084]
41. Grimmer J, Stewart BM. Text as data: the promise and pitfalls of automatic content analysis methods for political texts. Polit Anal 2017;21(3):267-297. [doi: 10.1093/pan/mps028]
42. Brown T, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners advances. 2020 Presented at: NIPS '20: Proceedings of the 34th International Conference on Neural Information Processing Systems; December 6-12, 2020; Vancouver BC, Canada p. 1877-1901 URL: https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html
43. Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, et al. On the opportunities and risks of foundation models. ArXiv. Preprint posted online on August 16, 2021 2021. [doi: 10.48550/arXiv.2108.07258]
44. Namita N. Comparison between traditional topic modeling and generative AI topic modeling. IJAEM 2023;5(10):97-102 [FREE Full text]
45. Guidelines for Ethical Review of Clinical Research or Research involving human subjects. Medical Review & Ethics Committee (MREC): Ministry of Health Malaysia. 2025 Jun 26. URL: https://tinyurl.com/3a9dmxyp
46. Elmogy AM, Tariq U, Mohammed A, Ibrahim A. Fake reviews detection using supervised machine learning. IJACSA 2021;12(1). [doi: 10.14569/ijacsa.2021.0120169]
47. Asaad WH, Allami R, Ali YH. Fake review detection using machine learning. RIA 2023;37(5):1159-1166. [doi: 10.18280/ria.370507]
48. Alsubari SN. Data analytics for the identification of fake reviews using supervised learning. CMC 2022;70(2):3189-3204. [doi: 10.32604/cmc.2022.019625]
49. Zhang D, Zhou L, Kehoe JL, Kilic IY. What online reviewer behaviors really matter? Effects of verbal and nonverbal behaviors on detection of fake online reviews. J Manage Inf Syst 2016;33(2):456-481. [doi: 10.1080/07421222.2016.1205907]
50. Roh M, Yang S. Exploring extremity and negativity biases in online reviews: evidence from Yelp. com. Soc Behav Pers 2021;49(11):1-15. [doi: 10.2224/sbp.10825]
51. Zaman N, Goldberg DM, Abrahams AS, Essig RA. Facebook hospital reviews: automated service quality detection and relationships with patient satisfaction. Decis Sci 2020;52(6):1403-1431. [doi: 10.1111/deci.12479]
52. Sovacool BK. A qualitative factor analysis of renewable energy and sustainable energy for all (SE4ALL) in the Asia-Pacific. Energy Policy 2013;59:393-403. [doi: 10.1016/j.enpol.2013.03.051]
53. George D, Mallery P. SPSS for Windows Step by Step: A Simple Guide and Reference. Boston: Allyn & Bacon; 2003.
54. Yusoff MSB. ABC of content validation and content validity index calculation. EIMJ 2019;11(2):49-54. [doi: 10.21315/eimj2019.11.2.6]
55. Williams B, Onsman A, Brown T. Exploratory factor analysis: a five-step guide for novices. Australas J Paramed 2010;8:1-13. [doi: 10.33151/ajp.8.3.93]
56. Wang Y. Deciphering the enigma: a deep dive into understanding and interpreting llm outputs. TechRxiv. 2023. URL: https://www.techrxiv.org/users/691297/articles/682897-deciphering-the-enigma-a-deep-dive-into-understanding-and-interpreting-llm-outputs [accessed 2025-05-13]
57. Knetsch J. The endowment effect and evidence of nonreversible indifference curves. Am Econ Rev 1989;79(5):1277-1284. [doi: 10.1017/cbo9780511803475.010]
58. Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discovery 2019;18(6):463-477 [FREE Full text] [doi: 10.1038/s41573-019-0024-5] [Medline: 30976107]
59. Ferreira DC, Vieira I, Pedro MI, Caldas P, Varela M. Patient satisfaction with healthcare services and the techniques used for its assessment: a systematic literature review and a bibliometric analysis. Healthcare (Basel) 2023;11(5):639 [FREE Full text] [doi: 10.3390/healthcare11050639] [Medline: 36900644]
60. Google drive. URL: https://drive.google.com/file/d1Pw1eRAzjOmm7-wcP1YtfBguqHPtnA8YR [accessed 2025-06-05]

Abbreviations

AI: artificial intelligence
API: application programming interface
CV2: computer vision
HCAHPS: Hospital Consumer Assessment of Healthcare Providers and Systems
LACA: large language model–assisted content analysis
LLM: large language model
ML: machine learning
NLP: natural language processing

XSL•FO
RenderX

Edited by A Mavragani; submitted 21.11.24; peer-reviewed by Yanto, GSA Mala; comments to author 26.12.24; revised version received 15.01.25; accepted 07.04.25; published 27.06.25.

Please cite as:
Sulaiman MH, Muda N, Abdul Razak F
Analyzing Patient Complaints in Web-Based Reviews of Private Hospitals in Selangor, Malaysia, Using Large Language Model–Assisted Content Analysis: Mixed Methods Study
JMIR Form Res 2025;9:e69075
URL: https://formative.jmir.org/2025/1/e69075
doi:10.2196/69075
PMID:

©Muhammad Hafiz Sulaiman, Nora Muda, Fatimah Abdul Razak. Originally published in JMIR Formative Research (https://formative.jmir.org), 27.06.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Examining How Oral Nicotine Pouches Are Trending on TikTok: A Qualitative Descriptive Study

Ashmeet Mand[1]; Karlee Fonteyne[1], RN, BScN, MSN; Laura Struik[1], RN, BScN, MSN, PhD

University of British Columbia Okanagan, Kelowna, BC, Canada

Corresponding Author:
Laura Struik, RN, BScN, MSN, PhD
University of British Columbia Okanagan
1147 Research Rd
Kelowna, BC, V1V1V7
Canada
Phone: 1 2508079972
Email: laura.struik@ubc.ca

Abstract

Background: Oral nicotine pouches (ONPs) have become a popular alternative nicotine product in recent years, especially among youth. This is concerning given the substantial health risks associated with nicotine use at an early age. Social media channels that are popular among youth (eg, TikTok) are being used to promote ONPs. Understanding the ways that individuals communicate about ONPs on popular social media, like TikTok, is critical for informing public health efforts aimed at protecting youth from the harms of nicotine.

Objective: This study aimed to explore how ONPs (specifically Zyns) are represented on TikTok.

Methods: Through a qualitative descriptive design, TikTok videos were collected and analyzed. A total of 250 TikTok videos were screened under the hashtags #zyn, #zyns, and #nicotinepouch, with 132 included for final analysis. Analysis was guided by the Braun and Clarke thematic analysis approach. Collaborative coding ensured reliability and supported the identification of key themes in the videos.

Results: The following five themes emerged: (1) the Zyn movement (58/132, 43.9%), which relays how ONPs have become a popular culture trend; (2) “boy heaven” (28/132, 21.2%), which showcases the dominance of male use; (3) unintended negative consequences (18/132, 13.6%), which reveals how negative side effects are downplayed; (4) product design: life does not need to stop (13/132, 9.8%), which highlights how ONPs enable nicotine consumption anywhere at any time; and (5) physical benefits: “It’s like IcyHot for your mouth” (6/132, 4.5%), which brings forward positive sensory experiences with ONP use. Overall, the content heavily condoned ONP use and normalized it, with male individuals disproportionately represented as the primary users of ONPs.

Conclusions: This study reveals TikTok’s influence in shaping the perceptions of ONPs as trendy and relatively harmless, especially among youth. The findings highlight how these products can quickly become embedded in youth culture through popular social media channels. The findings also underscore the need for more restrictions on the content that is so readily available to youth.

(JMIR Form Res 2025;9:e73032) doi:10.2196/73032

KEYWORDS

qualitative research; oral nicotine pouch; ONP; TikTok; Zyn; youth; social media

Introduction

Background

In recent years, novel forms of nicotine consumption have surged in popularity. A relatively new addition to the landscape of nicotine products is oral nicotine pouches (ONPs), which has grabbed the attention of youth. A recent scoping review based in the United States estimated that adolescent lifetime use of ONPs was 3.5% to 4.1% and current use was 1.5% to 2.0% [1]. It is likely that these numbers will climb within this demographic as ONP use continues to infiltrate the lives of today’s youth [1,2].

ONPs are slim microfiber pouches filled with white, finely granular nicotine salts [3-5]. ONPs typically contain between

XSL•FO
RenderX

2 mg and 10 mg of nicotine per pouch, subject to country-specific regulations; however, some pouches have been reported to contain up to 50 mg of nicotine per individual pouch [2-7]. These pouches may come in a myriad of flavors, such as berry, citrus, and mint, making them more enticing to users [8,9]. Similar to traditional snus (a smokeless tobacco-based product), ONPs are inserted between the gum line and the cheek. However, unlike snus, ONPs are touted as "tobacco free" and are frequently marketed as a healthier option for nicotine users [1,8,10]. There is not enough evidence to date to definitively determine whether ONPs are less harmful than their electronic cigarette (e-cigarette) counterparts [1,2,11]. That said, ample research has emphasized the negative health implications of acute and chronic nicotine exposure, particularly among youth, including neurodevelopmental and cardiovascular effects, as well as a relationship with mental health issues [12-14]. Despite the negative health implications of nicotine use, several studies have found that users of ONPs perceived them to be less harmful compared with other available nicotine products [1,2,10,15].

Brands (eg, market leader Zyn, a subsidiary of Philip Morris International) [7] selling and marketing ONPs have pushed the idea that these products are discreet, convenient, and trendy, and use buzzwords like "freedom," "flavor," and "innovation" [8,15,16]. Furthermore, these companies have capitalized on the reach and power of social media to market ONPs, particularly targeting the youth demographic [8]. Research has shown that social media plays a role in influencing trends among youth, including the use of nicotine products, by increasing their social acceptability and desirability and lowering their harm perception [17-20]. Unlike traditional advertising channels, social media platforms facilitate peer-to-peer sharing, amplify exposure via algorithms, and create participatory cultures where behaviors can quickly become normalized. This makes them a particularly influential force in shaping youth perceptions of nicotine products.

One of the most popular social media platforms used among youth, where ONPs are being promoted, is TikTok [17,21,22]. TikTok was first introduced in 2017 and has become one of the fastest-growing social media platforms [23]. The hallmark feature of TikTok is the ability to create short videos with different effects (eg, overlaid music or text, filters) that can then be posted and shared via hashtags. As of 2024, there were an estimated 2 billion TikTok users worldwide. According to Pew Research Center [24], 63% of teenagers (aged 14-17 years) in the United States use TikTok, and more than half of these users report using it daily. TikTok's format of highly visual, short-form content encourages the rapid dissemination of trends and narratives, making it a key hub for examining how ONPs are portrayed and engaged with in real time.

This Study

Two recent content analyses examined the presence and promotion of ONPs on the TikTok platform [21,22]. Both Donalson et al [21] and Zenone et al [22] highlighted the normalization of ONP use, the reliance on humor, and the minimization of associated risks, noting a lack of cessation-related content. While these content analyses provided timely and valuable insights into the promotional strategies surrounding ONPs and the presence of ONP content on TikTok, they captured a limited picture of the meaning embedded in user-generated posts. Given that engagement with nicotine-related social media content has been linked to an increased likelihood of product use [20], it is critical to further examine how ONPs are represented on TikTok. Therefore, this study used a thematic analysis to move toward an in-depth understanding of how users express their experiences and perspectives on ONPs. By doing so, it offers additional insights into the ways ONP use is constructed, communicated, and normalized in digital spaces, extending the current literature and informing future prevention, policy, and cessation efforts.

Methods

Study Design

Given the lack of research in this emerging area, we used qualitative description, which is well suited for exploring and describing phenomena where research is lacking [25,26]. This study focused on surfacing the meaning of ONP use conveyed through TikTok videos without excessive theoretical abstraction. This methodological approach ultimately led to actionable findings for public health efforts.

This study involved 3 coauthors: a senior researcher, an emerging researcher, and a youth coresearcher. This authorship enabled blending of the expertise of the senior and emerging researchers with the lived experience and cultural immersion of the youth coresearcher as it related to TikTok and nicotine use trends within the youth demographics. The entire study process, from conceptualization to knowledge translation, was a collaborative endeavor.

Ethical Consideration

TikTok has multiple options for users to restrict their content through privacy settings. For example, users can choose between a public and private account and can also select "who can view" each video. Public accounts can set each TikTok video to be viewed by friends, everyone, or only themselves. Therefore, ethical clearance was deemed unnecessary, as TikTok videos posted from public accounts with no viewing restrictions indicate that the creators intended their content to be widely accessible. This approach is in line with previous studies examining TikTok videos [27].

Only posts that were publicly accessible and downloadable were included in the thematic analysis, with the assumption that creators had no expectation of privacy and that their videos were part of the public domain. To minimize the potential for harm, all usernames and profile images were removed, and direct quotes were paraphrased when possible or used only when necessary to illustrate a theme. Paraphrasing was applied when the exact wording was not essential to the meaning, which reduced the likelihood of reverse identification through search engines. However, when specific phrasing, slang, or tone was central to understanding the theme, direct quotes were retained. This approach balanced the need to preserve interpretive fidelity with the imperative to protect user anonymity.

We also acknowledge the ethical implications of studying youth behavior on social media, particularly given the sensitivity of

XSL•FO
RenderX

nicotine use among minors. Our analysis was strictly observational, with no interaction or engagement with users. The TikTok account used for data collection was created solely for research purposes, with the age set to 15 years to ensure the algorithm displayed content representative of what midadolescent users might encounter.

Data Collection

TikTok videos containing the hashtags #zyn, #zyns, and #nicotinepouch were analyzed. These hashtags were selected because, at the time of data collection (May 13, 2024), they were the most frequently used and most frequently engaged tags related to ONPs, collectively encompassing approximately 45,000 posts. Content under these hashtags included personal accounts of ONP use, as well as discussions from significant others about a partner's use. A total of 250 TikTok videos were selected from a combined pool of content under the 3 hashtags. This combined approach was used because many TikTok videos contained more than one of the selected hashtags, with some including 2 or even all 3 hashtags. Pooling avoided duplicate content and ensured the sample represented the overall range of popular ONP content on the platform. The 250-video threshold was chosen because TikTok's algorithm organizes videos based on engagement metrics such as likes, views, and interactions; therefore, the first 250 videos represented the most popular content available at that time. From the initial pool, we excluded any TikTok videos that (1) were not in English, (2) were not downloadable, (3) did not mention ONPs or Zyns, or (4) were posted before January 2023. After applying these exclusion criteria, 132 TikTok videos remained for analysis. The number of likes, comments, and shares for each TikTok was noted and recorded on a Microsoft Excel sheet.

Data Analysis

The first author (AM) and the third author (LS) analyzed the compiled TikTok videos guided by the Braun and Clarke 6-step approach for thematic analysis [28]. For the first step (familiarization), we viewed the collected TikTok videos and discussed our observations and how to approach the coding process. For the second step (initial code generation), the analysis of these videos began with individually coding the first 20 TikTok videos. Each TikTok was watched at least 3 times. Then, we met and discussed our findings to ensure there was consistency within the analysis. This process was repeated with the first 50 TikTok videos to ensure consistency in the coding process. The first author (AM) then completed coding the remaining TikTok videos. It is noteworthy that the familiarity of the youth coresearcher with the lingo used on TikTok and the social trends associated with nicotine use among younger demographics was critical to the analytic process because this familiarity enabled an accurate depiction of the dominant themes. This process helped to develop a codebook that was accurately representative of the content and increased interrater reliability. During step 3 (development of themes), after all the videos were coded, AM and LS engaged in several collaborative meetings to discuss how these codes naturally group together to form themes. For example, while we had codes relating to the ability to conceal use, the ability to use ONPs anywhere, and the ability to engage in sports or military training while using them, it became evident that there was an underlying discourse that life does not need to stop consuming nicotine with ONPs. Therefore, our primary theme became "life doesn't have to stop." During step 4 (theme refinement), we collaboratively reviewed themes and ensured that our themes and associated subthemes were capturing the video content appropriately. For step 5 (theme naming), we built a diagram to showcase how each theme and subtheme related to each other (Figure 1). Within our findings, we indicated the number of videos associated with each theme to showcase the prominence of the themes, as well as the degree to which videos conveyed ONPs in a positive or negative way.

XSL•FO
RenderX

Figure 1. Themes identified across content posted under #zyn, #zyns, and #nicotinepouch on TikTok. The five main themes are represented by larger circles, with their size reflecting the relative prominence of the theme within the dataset. These themes include: (1) the Zyn movement, (2) "boy heaven", (3) product design: life doesn't have to stop, (4) unintended negative consequences, and (5) physical benefit: "It's like IcyHot for your mouth." Subthemes are shown in smaller circles connected to their corresponding theme. Within the Zyn movement, subthemes include (a) #zynlifestyle, (b) Zyns make life great, and (c) artistic expression. Within unintended negative consequences, subthemes include (a) nic sick and (b) long-term side effects. Lines connecting shapes indicate conceptual relationships observed during coding.

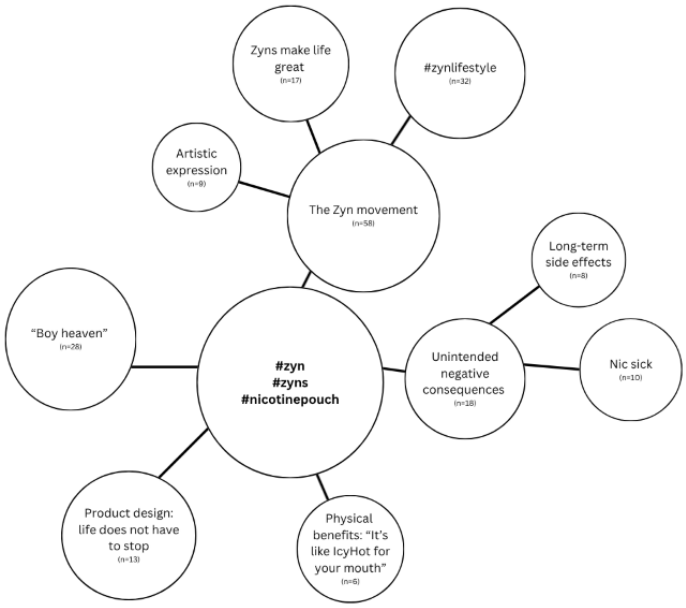

Results

Overview

The authors (AM and LS) interpreted the data to form 5 themes (two of which have subthemes), as shown in Figure 1. The authors understood that each individual TikTok varied greatly despite sharing the same hashtag, so each video was coded according to the central message the creator was trying to deliver. These TikTok videos generated a sum of 16,488,662 likes; 114,125 comments; and 1,930,115 shares. From the most common theme to the least, the themes are (1) the Zyn movement, (2) "boy heaven", (3) product design: life does not need to stop, (4) unintended negative consequences, and (5) physical benefits: "It's like Icyhot for your mouth."

The Zyn Movement

Overview

Creators (58/132, 43.9%) showcased their alliance with using Zyns by posting content in favor of Zyns and ONPs. This theme captured how creators framed Zyn use as empowering, exclusive, and socially desirable, often presenting participation in the "movement" as part of a distinct social identity. The Zyn movement was a common hashtag under videos where Zyns were mentioned, signaling membership in a shared identity group. Within this theme, 3 subthemes were identified: (1) #Zynlifestyle, (2) Zyns make life great, and (3) artistic expression.

#Zynlifestyle

This subtheme showcased how Zyns could become a part of someone's lifestyle. A lifestyle describes a way in which a person chooses to live their life. Creators (32/132, 24.2%) posted videos portraying Zyns as a seamless part of their day, whether at work, watching sports, playing sports, or on vacation. One creator even called Zyns the "perfect ball marker" while playing golf, illustrating how the product was framed as both practical and socially embedded within leisure activities. Creators also showed Zyns being used as tools, such as a car jack and a line marker, reinforcing the portrayal of Zyns as adaptable to many contexts. Collectively, these examples depicted Zyn use as normalized and compatible with diverse lifestyles.

Zyns Make Life Great

This subtheme promoted the use of Zyns as something that makes life great. Creators (17/132, 12.8%) made TikTok videos expressing how great Zyns are and how they have made their lives better. For example, one creator made a 2-minute video solely expressing his love for Zyns, directly reflecting the *Zyns make life great* narrative. Another showcased a soda alongside a Zyn with the caption, "Just copped the 'Why do I feel so good all the time? starter pack'," implying that Zyns are part of a recipe for happiness. Similarly, a creator posted, "How are you sad in a generation where wintergreen zyns, coors banquets, and girls in 7s exist," framing Zyns as central to maintaining a positive mood. Together, these posts position Zyn use as a simple and reliable pathway to a better life.

Artistic Expression

This subtheme highlighted the creativity of creators who incorporated Zyns into their TikTok videos to showcase the product's versatility. Zyns were not only presented as a source of nicotine but also as a medium for artistic expression. Creators (9/132, 6.8%) featured Zyn cans in decorations, music, toys, and participation in TikTok trends. For example, the hashtag #zynmas appeared under videos displaying Zyn cans as Christmas tree ornaments, demonstrating seasonal creativity tied to the product. One creator made her boyfriend a cake

XSL•FO
RenderX

identical to a Zyn can, while others incorporated the cans into trending TikTok formats. TikTok trends are typically short-lived periods when a particular song or dance becomes widely popular, and participating can help creators remain visible and engaged with audiences. Within this subtheme, Zyns were integrated into these trends in ways that positioned them as a creative prop rather than solely a nicotine product.

"Boy Heaven"

Creators (28/132, 21.2%) portrayed Zyns as a source of pleasure and indulgence, often tied to male identity or relationships. The phrase "boy heaven" appeared in a video highlighting how cheap and accessible Zyns were at a festival, directly capturing this theme's essence. Many of these videos were made by what appeared to be males or their significant others, with some partners commenting that they "just couldn't get mad" because the Zyn "rewards" were so good. One creator assembled a care package for her military boyfriend containing multiple Zyn cans, presenting them as thoughtful and valued gifts. Captions such as "perfect boy's night" and "just a guy lunch" paired Zyns with alcohol or energy drinks, reinforcing the idea that Zyns were central to the ideal male social experience, akin to a personal "heaven on Earth."

Unintended Negative Consequences

Overview

Creators (18/132, 13.6%) often reached out to social media to share their experiences with using ONPs. Whether it was their first day of use or consistent use, there was a lot of testimonial-style content. One user, who presented their content through a podcast on TikTok, labeled the side effects of nicotine consumption, specifically through ONPs, as "unintended negative consequences", presenting them as mere coincidences. This phrase was a strong representation of how the harms associated with nicotine consumption are minimized, with many creators choosing to dismiss them altogether. This theme was divided into 2 subthemes: (1) nic sick and (2) long-term side effects.

Nic Sick

This subtheme highlighted the negative physical reactions to nicotine, particularly among first-time users of Zyns or other ONPs. Commonly referred to as "nic sick," these symptoms include nausea, sweating, and vomiting. Creators (10/132, 7.5%) shared "storytimes" about their severe first-day experiences, often framing them with humor or dramatic storytelling. One creator, confident from prior vaping experience, tried a Zyn and reported they "felt like they won't make it till tomorrow" and "slept in the fetal position," illustrating the intensity of nicotine's acute effects despite previous exposure. In some cases, the severity of symptoms led creators to describe fearing for their well-being, while still presenting these incidents as simultaneously cautionary tales and entertaining content.

Long-Term Side Effects

This subtheme highlighted the health issues associated with prolonged ONP use. Creators (8/132, 6%) featured content describing impacts such as gum recession, tooth decay, and, in severe cases, oral cancer, often showing visible dental changes on camera. Some posts also referenced nicotine's potential impact on cardiac health. These videos presented ONPs, like Zyns, as capable of causing damage to the user's physical health.

Product Design: Life Does Not Need to Stop

Creators (13/132, 9.8%) produced videos highlighting the convenience, discreteness, and shareability of ONPs, as well as the ability to use multiple at once. Content showed individuals using Zyns while working out, at work, and during everyday activities. Athletes such as golfers, hockey players, and baseball players were seen using Zyns midgame or interacting with fans, with some videos showing fans offering Zyns and athletes accepting them, illustrating how easily the product could be distributed and shared. Other posts emphasized discreet use in restricted environments, such as soldiers using Zyns while on duty. In addition, 2 creators demonstrated using multiple pouches simultaneously, showcasing how nicotine intake can be readily increased and customized. Overall, these videos presented Zyns as a product that integrated seamlessly into both professional and leisure activities without interrupting daily routines.

Physical Benefits: "It's Like IcyHot for Your Mouth."

Creators (6/132, 4.5%) produced videos promoting perceived physical and cognitive benefits of ONPs. Reported benefits included nicotine being described as a nootropic, which, in lay terms, is a work enhancer and a substance that increases mental acuity. For male individuals, it was linked to boosting testosterone and referred to as a "male enhancer." Some posts suggested it could aid weight management, with one creator recommending its use for severe weight problems. Statements such as "I don't think nicotine hurts anything" and "Zyn is not a sin" reflected a minimization of potential harms in favor of perceived benefits. One creator compared ONPs to the popular muscle relief topical medication IcyHot (drug class: menthol and methyl salicylate), which is known to provide significant relief. This comparison further solidified how these creators portrayed ONPs as beneficial.

Discussion

Principal Findings

This study revealed themes around how ONPs are experienced, perceived, and promoted among TikTok users. The implications of these findings are critical for advancing health promotion efforts so that they resonate with individuals who use these platforms, particularly youth. A primary finding is the resoundingly positive way in which ONPs, and Zyns in particular, were portrayed in TikTok videos (eg, "boy heaven," "unintended negative consequences," and "life doesn't have to stop"). This echoed findings from previous works examining how nicotine products popular among youth were engaged with on TikTok, such as vaping [17-20], as well as the two previously mentioned content analyses of ONPs on TikTok [21,22]. For example, Sun et al [19] found that, of a sample of 808 TikTok videos that were collectively viewed more than 1.5 billion times, the majority (n=509, 62.9%) portrayed e-cigarette use in a positive way. Taken together, the findings not only underscore how novel nicotine products are being positively portrayed on

XSL•FO
RenderX

popular youth-led social media platforms but also explain why these products have become so widely proliferated in youth cultures.

A second key finding was the way in which ONPs are not promoted as a cessation product, for which they were supposedly intended, but rather portrayed as a lifestyle product. Some videos directly referred to the use of ONPs as a movement (eg, #zynlifestyle, #zynbabwe, and #zyngang), and other videos associated the use of ONPs as something that improves everyday life (eg, "Zyns make life great; product design: life does not have to stop"). This is similar to how vaping became a lifestyle product through similar social media posts and hashtags (eg, #vapenation and #vapefam), thereby normalizing the use of e-cigarettes as a part of everyday life and positioning them as a communal experience [18,29]. In this regard, social media serves as a powerful tool for the tobacco industry in normalizing the use of their products among youth, a prime demographic to secure profits from because of the developmental propensity of youth to become addicted to their products for the long term [12,30].

Another important finding was that individuals who appeared to be male created most of the TikTok videos about ONP use, or showed ONPs being used by what appeared to be male individuals. This is reflected in our theme, "boy heaven." This emphasized that the use of ONPs encompassed possible gender-based dynamics whereby the use of these products is portrayed as a dominantly masculine endeavor. For example, male role models, including athletes, military personnel, and influencers, were represented in these TikTok videos, which not only reinforced and promoted the use of pouches among men, but also showed that perceived "healthy men" and "masculine men" are using these products. Furthermore, while male individuals presented as the primary users of ONPs on TikTok, female individuals were presented as supporters of their male counterparts' use, which foregrounded stereotypical gender roles playing out. For example, in many of the videos, the female partners, while they may not have agreed with the behavior, still showed support for ONP use by creating art with the tins (eg, a Zyn birthday cake for her boyfriend), sharing that they could not become angry at their partner because they received free items through the rewards and by remaining silent when their partners announced that they are running out to obtain pouches. Not only did these findings align with studies indicating that more male individuals use ONPs compared with female individuals to date [1,31,32] but they also provided a glimpse into how gender is influencing and reinforcing these trends.

Implications

A major implication of this study's findings was the need to address ONPs as lifestyle products, given their widespread normalization among youth as represented in the analyzed TikTok videos. Once a trend, like ONP use, becomes embedded in youth contexts and is accelerated via social media, it becomes normalized, which inherently makes such trends difficult to counter. This brings forward the need to shift focus from presenting ONP use as solely a health issue to acknowledging that it is also a lifestyle trend embedded into youth cultures. Efforts to educate youth about the harms of these products may fall flat if we only focus on this issue for its health implications. This was exemplified in the way that the negative health impacts of ONP use were downplayed or presented as a fluke, or as an "unintended" side effect of using them. Therefore, there is an urgent need to design prevention messaging that resonates with youth cultures and to empower youth to believe that they do not need to engage in ONP use to fit in or to feel good. Designing youth-driven and peer-based interventions, disseminating targeted prevention and cessation messaging via popular social media platforms, and harnessing the power of social media influencers to counter and challenge these trends are promising ways in which to positively engage youth on this issue.

There currently appears to be a gender gap when it comes to ONP use, whereby men are overwhelmingly represented as pouch users on TikTok. This is similar to how smoking and vaping began, but through clever marketing, the gender gap began to close. One can expect that the same gender gap will close shortly without due attention to how women may be particularly targeted.

Social media regarding Zyns capitalizes on the concept of social bonding among users. Users establish small communities named #zyngang, #zymbabwe, and #zynlifestyle on social media platforms, which represent a safe space for expressing their passion for ONPs (Zyns). These communities can become particularly appealing to young people as they increasingly value and develop social connections and social capital. As per the Erikson theory [33] of psychosocial development, youth aged 12 to 18 years belong in the "identity versus role confusion" stage. This stage is where youth are figuring out their identity and what their values are, ultimately driving their decision-making around where they belong. This is why Zyn communities on social media can further increase the uptake of these products among youth. Therefore, stricter policies are needed for hashtags and content. TikTok has community guidelines for products such as cannabis, which aim to block the content from being viewed by anyone who is underage; however, there is currently no policy in place for ONPs.

Furthermore, education needs to be provided to individuals who interact with youth daily, such as parents, teachers, and coaches. They should be provided with education and resources on ONPs, as well as how to engage in productive conversations with youth about ONPs. Because the product design is so slick and easily concealable, schools will need support to address its use in classrooms. In addition, health care professionals, such as pharmacists, need to be educated on the actual consequences of these products. There needs to be an increase in awareness of the negative consequences of this product, so they cannot continue to be dismissed as "unintended."

Finally, research is needed to further understand the use of ONPs, particularly among youth. Future studies that qualitatively explore the subjective experiences among diverse ONP users are needed, as well as quantitative explorations of how social media portrayals may be linked to uptake among younger demographics. Health promotion efforts also need to be

XSL•FO
RenderX

developed and tested so that more proactive protection of youth health can be achieved.

Strengths and Limitations

A major strength of this study was that, through a thematic analysis, it surfaced the meanings underpinning ONP-related posts on the most popular social media platform used among youth, TikTok. Another strength of this study was that it was co-led by a youth coresearcher alongside an emerging researcher and a senior researcher. This is a major strength because the lead author was familiar with the platform and with the lingo and the ways in which youth cultures are represented on this platform, which facilitated rigorous theme development. Finally, this study was strengthened by the collaborative coding and theme development process that characterized data analysis.

In relation to limitations, it is important to note that this study was conducted in 2024 and that rapid shifts in social media policies (eg, the temporary ban on TikTok in the United States), as well as ONP policies (eg, US Food and Drug Administration approval of Zyn marketing), may alter the ways in which ONPs are communicated. In addition, TikTok trends are highly temporal; hashtags, sounds, and viral formats can rise and fade within days or weeks. As such, the popularity, tone, and framing of ONP-related content captured in this study reflect a specific moment in time. While future analyses conducted months or years later may identify different creators or engagement patterns, the thematic insights derived from this dataset offer enduring relevance for understanding adolescent exposure and risk messaging in social media environments. This temporal volatility reinforces the importance of continuous monitoring. In addition, most TikTok users are based in the United States, which limits Canadian representation in this study. Because only English-language content was included, our findings may not be generalizable to non–English-speaking populations. While European creators were not intentionally excluded, their representation in the dataset was low because of the English-language focus and TikTok's algorithm prioritizing content from US-based creators.

This study also excluded content on snus, which limits the applicability of these findings to other oral-based nicotine products that may be popular and portrayed on social media. Furthermore, this content is representative of Zyn and does not include content shared in relation to other ONP brands (eg, Rogue and Zonnic), which is likely due to the higher proliferation of Zyn in the United States than other brands to date. However, this focus on Zyn may overlook other ONP brands that are popular in different regions. Finally, TikTok is a constantly growing platform, with millions of videos posted daily from all over the world. TikTok is algorithm focused, meaning it tracks the type of videos users watch to curate their experience. However, to make our findings unbiased, we created a new account on a new device with no previous data to be tracked and used. In addition, because the TikTok account used for data collection was set to a user age of 15 years, the algorithm may have prioritized content tailored to younger users, which could limit the generalizability of the findings to other age groups.

Conclusions

In this study, we analyzed TikTok content related to ONPs, revealing how the trend has become embedded in youth cultures through positive portrayals that resonate strongly with youth. The themes identified in our analysis illustrated how marketing and peer-to-peer promotion highlighted perceived benefits of Zyn while subtly leveraging masculinity and the male identity to normalize and glamorize use. This gendered dimension of ONP promotion is both novel and concerning, underscoring a pressing public health need to address how such dynamics may shape attitudes and behaviors. Policy and prevention efforts, ideally co-designed with youth, are urgently required to counter these messages and support informed decision-making. Further research is also essential to guide interventions and monitor how emerging nicotine products continue to evolve within rapidly shifting social media landscapes.

Acknowledgments

This study was supported by funding awarded to LS through a Canadian Cancer Society Scholar Award (prevention; grant 707156) and a Michael Smith Health Research Scholar Award (grant SCH-2023-3195).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

Study conceptualization, design, data collection, and analysis were completed by LS and AM. KF assisted with drafting and editing of the final manuscript.

Conflicts of Interest

None declared.

References

XSL•FO
RenderX

1. Travis N, Warner KE, Goniewicz ML, Oh H, Ranganathan R, Meza R, et al. The potential impact of oral nicotine pouches on public health: a scoping review. Nicotine Tob Res 2025 Mar 24;27(4):598-610. [doi: 10.1093/ntr/ntae131] [Medline: 38880491]
2. Felicione NJ, Ozga JE, Eversole A, Hart JL, Tackett A, Hrywna M, et al. Oral nicotine pouches: rising popularity and state of the science. Public Health Rep 2025 Apr 28:333549251313668. [doi: 10.1177/00333549251313668] [Medline: 40293136]
3. Kent JT, Mok G, Austin E. Nicotine toxicity from repeat use of nicotine pouches. Nicotine Tob Res 2025 Mar 24;27(4):767-768. [doi: 10.1093/ntr/ntae111] [Medline: 39087735]
4. Mallock N, Schulz T, Malke S, Dreiack N, Laux P, Luch A. Levels of nicotine and tobacco-specific nitrosamines in oral nicotine pouches. Tob Control 2024 Feb 20;33(2):193-199 [FREE Full text] [doi: 10.1136/tc-2022-057280] [Medline: 38378209]
5. Stanfill S, Tran H, Tyx R, Fernandez C, Zhu W, Marynak K, et al. Characterization of total and unprotonated (free) nicotine content of nicotine pouch products. Nicotine Tob Res 2021 Aug 18;23(9):1590-1596. [doi: 10.1093/ntr/ntab030] [Medline: 34233354]
6. Duren M, Atella L, Welding K, Kennedy RD. Nicotine pouches: a summary of regulatory approaches across 67 countries. Tob Control 2023 Feb 07;33(e1):e32-e40. [doi: 10.1136/tc-2022-057734] [Medline: 36750358]
7. Majmundar A, Okitondo C, Xue A, Asare S, Bandi P, Nargis N. Nicotine pouch sales trends in the US by volume and nicotine concentration levels from 2019 to 2022. JAMA Netw Open 2022 Nov 01;5(11):e2242235 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.42235] [Medline: 36378312]
8. Duan Z, Henriksen L, Vallone D, Rath JM, Evans WD, Romm KF, et al. Nicotine pouch marketing strategies in the USA: an analysis of Zyn, On! and Velo. Tob Control 2024 Feb 20;33(2):154-163. [doi: 10.1136/tc-2022-057360] [Medline: 35817549]
9. Vogel EA, Tackett AP, Unger JB, Gonzalez MJ, Peraza N, Jafarzadeh NS, et al. Effects of flavour and modified risk claims on nicotine pouch perceptions and use intentions among young adults who use inhalable nicotine and tobacco products: a randomised controlled trial. Tob Control 2025 May 15;34(3):315-322. [doi: 10.1136/tc-2023-058382] [Medline: 38148143]
10. Morean ME, Bold KW, Davis DR, Kong G, Krishnan-Sarin S, Camenga DR. "Tobacco-free" nicotine pouches: risk perceptions, awareness, susceptibility, and use among young adults in the United States. Nicotine Tob Res 2023 Jan 01;25(1):143-150 [FREE Full text] [doi: 10.1093/ntr/ntac204] [Medline: 36000776]
11. Jenssen BP, Camenga D. Oral nicotine products, nicotine pouches, and adolescent vaping: a public health perspective. Pediatrics 2025 Jun 01;155(6):e2025070742 [FREE Full text] [doi: 10.1542/peds.2025-070742] [Medline: 40383536]
12. Dorotheo EU, Arora M, Banerjee A, Bianco E, Cheah NP, Dalmau R, et al. Nicotine and cardiovascular health: when poison is addictive - a WHF policy brief. Glob Heart 2024;19(1):14 [FREE Full text] [doi: 10.5334/gh.1292] [Medline: 38312998]
13. Hackworth EE, Vidaña-Pérez D, O'Neal R, Kim M, Fillo J, Hammond D, et al. Trends in mental health symptoms, nicotine product use, and their association over time among adolescents in Canada, England, and the US: findings from the ITC adolescents tobacco and vaping survey, 2020-2023. Nicotine Tob Res 2025 Jun 23;27(7):1256-1264. [doi: 10.1093/ntr/ntaf015] [Medline: 39820468]
14. Yuan M, Cross SJ, Loughlin SE, Leslie FM. Nicotine and the adolescent brain. J Physiol 2015 Aug 15;593(16):3397-3412 [FREE Full text] [doi: 10.1113/JP270492] [Medline: 26018031]
15. Han DH, Cho J, Harlow AF, Tackett AP, Vogel EA, Wong M, et al. Young adults' beliefs about modern oral nicotine products: implications for uptake in nonvapers, dual use with e-cigarettes, and use to reduce/quit vaping. Exp Clin Psychopharmacol 2023 Apr;31(2):455-463 [FREE Full text] [doi: 10.1037/pha0000595] [Medline: 36048111]
16. Hardin A, Ahn S, Ma S, Thai V, Berman M, Mays D, et al. The association between product attributes and oral nicotine pouch (ONP) preference - evidence from online stores. Nicotine Tob Res 2025 Jul 01:ntaf137. [doi: 10.1093/ntr/ntaf137] [Medline: 40590502]
17. Marynak K, Robichaud M, Puryear T, Kennedy R, Moran M. #Nicotineaddiction on TikTok: a quantitative content analysis of top-viewed posts. Tob Induc Dis 2022 Aug 5;20(August):69-69 [FREE Full text] [doi: 10.18332/tid/151868] [Medline: 35979474]
18. Morales M, Fahrion A, Watkins SL. #NicotineAddictionCheck: puff bar culture, addiction apathy, and promotion of e-cigarettes on TikTok. Int J Environ Res Public Health 2022 Feb 05;19(3):1820 [FREE Full text] [doi: 10.3390/ijerph19031820] [Medline: 35162846]
19. Sun T, Lim CC, Chung J, Cheng B, Davidson L, Tisdale C, et al. Vaping on TikTok: a systematic thematic analysis. Tob Control 2023 Mar 26;32(2):251-254. [doi: 10.1136/tobaccocontrol-2021-056619] [Medline: 34312317]
20. Vassey J, Vogel EA, Unger JB, Cho J, Bae D, Donaldson SI, et al. Impact of Instagram and TikTok influencer marketing on perceptions of e-cigarettes and perceptions of influencers in young adults: a randomised survey-based experiment. Tob Control 2025 Feb 13:tc-2024-059021 [FREE Full text] [doi: 10.1136/tc-2024-059021] [Medline: 39947698]
21. Donaldson SI, La Capria K, DeJesus A, Ganz O, Delnevo CD, Allem J. Describing ZYN-related content on TikTok: content analysis. Nicotine Tob Res 2025 Jan 17:ntaf016. [doi: 10.1093/ntr/ntaf016] [Medline: 39823221]
22. Zenone M, Harries B, Hartwell G. The promotion of oral nicotine pouches for non-smoking cessation purposes on TikTok. Nicotine Tob Res 2025 Jan 29:024. [doi: 10.1093/ntr/ntaf024] [Medline: 39879071]
23. Statista. URL: https://www.statista.com/topics/6077/tiktok/#topicOverview [accessed 2025-09-13]

XSL•FO
RenderX

24. Teens and social media fact sheet. Pew Research Center. URL: https://www.pewresearch.org/internet/fact-sheet/teens-and-social-media-fact-sheet/ [accessed 2025-09-13]
25. Hall S, Liebenberg L. Qualitative description as an introductory method to qualitative research for master's-level students and research trainees. Int J Qual Methods 2024 Mar 25;23:33 [FREE Full text] [doi: 10.1177/16094069241242264]
26. Sandelowski M. What's in a name? Qualitative description revisited. Res Nurs Health 2010 Feb;33(1):77-84 [FREE Full text] [doi: 10.1002/nur.20362] [Medline: 20014004]
27. Herrick SS, Hallward L, Duncan LR. "This is just how I cope": an inductive thematic analysis of eating disorder recovery content created and shared on TikTok using #EDrecovery. Int J Eat Disord 2021 Apr;54(4):516-526 [FREE Full text] [doi: 10.1002/eat.23463] [Medline: 33382136]
28. Terry G, Hayfield N, Clarke V, Braun V. Thematic analysis. In: Willig C, Stainton-Rogers W, editors. The SAGE Handbook of Qualitative Research in Psychology. 2nd edition. Thousand Oaks, CA: Sage Publication; 2017:17-37.
29. Jancey J, Leaver T, Wolf K, Freeman B, Chai K, Bialous S, et al. Promotion of e-cigarettes on TikTok and regulatory considerations. Int J Environ Res Public Health 2023 May 09;20(10):5761 [FREE Full text] [doi: 10.3390/ijerph20105761] [Medline: 37239490]
30. Nicotine and the young brain. The Truth Initiative. URL: https://truthinitiative.org/research-resources/harmful-effects-tobacco/nicotine-and-young-brain [accessed 2025-05-29]
31. Patel M, Kierstead EC, Kreslake J, Schillo BA. Patterns of oral nicotine pouch use among U.S. adolescents and young adults. Prev Med Rep 2023 Aug;34:102239 [FREE Full text] [doi: 10.1016/j.pmedr.2023.102239] [Medline: 37228836]
32. Couch ET, Halpern-Felsher B, Werts M, Chaffee BW. Use of emerging and conventional oral tobacco among adolescent and young adult e-cigarette users. Subst Use Misuse 2023 Jan 16;58(2):283-288 [FREE Full text] [doi: 10.1080/10826084.2022.2161314] [Medline: 36645845]
33. Orenstein GA, Lewis L. Erikson's stages of psychosocial development. StatPearls. URL: https://www.ncbi.nlm.nih.gov/books/NBK556096/ [accessed 2025-05-17]

Abbreviations

e-cigarette: electronic cigarette
ONP: oral nicotine pouch

Edited by A Mavragani; submitted 23.02.25; peer-reviewed by SP Uppu, C Keke, Q Ng, R Dhillon; comments to author 04.04.25; revised version received 08.08.25; accepted 17.09.25; published 14.11.25.

Please cite as:
Mand A, Fonteyne K, Struik L
Examining How Oral Nicotine Pouches Are Trending on TikTok: A Qualitative Descriptive Study
JMIR Form Res 2025;9:e73032
URL: https://formative.jmir.org/2025/1/e73032
doi:10.2196/73032
PMID:41236804

©Ashmeet Mand, Karlee Fonteyne, Laura Struik. Originally published in JMIR Formative Research (https://formative.jmir.org), 14.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Association Between eHealth Literacy and Mental Health Literacy: Cross-Sectional Study

Efrat Neter[1], PhD; Refael Youngmann[2], PhD; Naama Gruper[2], MA

[1]Department of Behavioral Sciences, Ruppin Academic Center, 3 Bait, Emeq Hefer, Israel
[2]Clinical Psychology Graduate Program, Ruppin Academic Center, Emeq Hefer, Israel

Corresponding Author:
Efrat Neter, PhD
Department of Behavioral Sciences, Ruppin Academic Center, 3 Bait, Emeq Hefer, Israel

Abstract

Abstract: Associations between eHealth literacy and mental health literacy were examined; no significant association was identified between overall eHealth and mental health literacy and only weak associations between specific skills were recorded. Results are interpreted in lieu of a difference between perceived ability and actual performance.

(JMIR Form Res 2025;9:e76812) doi:10.2196/76812

KEYWORDS
eHealth literacy; mental health literacy; stigma; perceived; performance; mental health

Introduction

As digital technology advances, eHealth literacy (eHL) has become increasingly important for citizens. Defined as the ability to seek, understand, appraise, and apply health information from electronic sources [1], eHL was found to be associated with better health processes [2,3]. Still, it remains unknown whether the skills of using digital resources in health contexts are translated into enhanced mental health (MH) knowledge, such as better identification of MH needs or more accepting attitudes toward seeking support.

Mental health literacy (MHL) refers to knowledge about MH disorders and their recognition, management, and prevention [4]. It encompasses maintaining positive MH, recognizing mental disorders and their treatments, reducing stigma, and enhancing help-seeking efficacy. Low MHL (including stigma) has been linked to the global treatment gap in MH care, persisting across income levels and health care availability [5].

This study thus examined the association between the above two literacies, hypothesizing a positive association, considering potentially overlapping skill sets needed to navigate digital health information and to understand MH.

Methods

Participants

The study recruited 269 Hebrew-speaking participants aged ≥18 years; 241 were recruited via an online panel (Panel4All) and 28 using WhatsApp snowball sampling, the latter with no monetary compensation. This study is part of a bigger study assessing MHL in ethnocultural groups in Israel.

Measures

EHL Assessment

eHL was measured using the HLS_{19}-DIGI Instrument (Health Literacy Scale 19 - Digital Instrument) [6], comprising 8 items on eHL (Cronbach α=.89). We also analyzed the skills separately: access (3 items, Cronbach α=.69), understand (1 item), appraise (3 items, Cronbach α=.72), and apply (1 item).

MHL Assessment

MHL was assessed using items adapted from the Australian National Survey of MHL and Stigma Questionnaire [7], with case vignettes for depression, posttraumatic stress disorder, and schizophrenia, addressing the following dimensions: identification of MH distress, knowledge (potential helpers, appropriate intervention, first aid, all also summed into a general knowledge index), and stigma (9 items, 5-point Likert scale, Cronbach α=0.71 - 0.78 for the different vignettes). For more details, see [8].

Demographic Characteristics

Information on gender, age, country of birth, ethnicity, education, perceived income adequacy, religiosity, marital status, having children, and self-rated health was collected.

Ethical Considerations

This cross-sectional survey was conducted online in 2022 following explicit informed consent. The study was approved by the institutional review board of Ruppin Academic Center (approval number 2022 - 203) and was registered at OSF (OSF.IO/XMQPC).

XSL•FO
RenderX

Data Analysis

The characteristics of the participants and the primary variables of the research were described, including bivariate associations (Pearson or Spearman, depending on variable type) between MHL dimensions and eHL.

Results

Participants' Characteristics

Study participants included 151 women (56.1%) and 118 men (43.9%), with a mean age of 42.7 (SD 14.47) years, ranging from 18 to 80 years. They were predominantly Jewish (n=261, 97%) and primarily secular (n=152, 56.7%), the majority had attained higher education (n=154, 57.3%), and financially, they mostly perceived themselves as either living comfortably or managing (n=214, 79.5%). A majority (n=170, 63.1%) rated their health as good or very good.

Bivariate Associations Between MHL Indices and eHL

Pearson coefficients were computed to examine associations between eHL and MHL indices (Table 1). No significant association was found between the eHL index and indices of MHL (range −0.08 to 0.12), and we therefore computed correlations between items of eHL (assessing, accessing, understanding, applying, and appraising) and MHL indices. A significant positive correlation was found between the perceived skill of *understanding* and knowledge of appropriate interventions and MHL knowledge index (r_{267}=0.18, P=.004 and r_{267}=0.16, P=.01, respectively) and a significant negative correlation between *understanding* and stigma (r_{267}=−0.15, P=.02). Lastly, *applying* was significantly negatively associated with correctly identifying MH conditions. No other association was found between MHL indices and other cognitive aspects of eHL. Controlling for age did not change the results.

Table . Pearson correlations between mental health literacy indices and digital health literacy skills.

	Access	Understand	Apply	Appraise	Digital health literacy index
Potential helpers	0.04	0.06	0.06	0.02	0.03
Appropriate intervention	0.18	0.18[a]	−0.05	0.00	0.07
First aid	0.11	0.11	0.16[b]	0.05	0.12
MHL[c] Knowledge Index	0.12	0.16[b]	0.01	0.03	0.09
Stigma	−0.07	−0.15[b]	0.05	−0.06	−0.07
Identification[d]	0.04	0.02	−0.20[a]	−0.08	−0.08

[a]P<.01

[b]P<.05.

[c]MHL: mental health literacy.

[d]Spearman correlation in this variable.

Discussion

We found no significant association between eHL as an index and the different dimensions of MHL, yet we found weak positive associations between the digital skills of understanding and applying and dimensions of knowledge and stigma, while a weak negative association was found between applying and identification of MH conditions.

These findings are consistent with the content of these two measures: the eHL assesses the perceived ease of cognitive activities, whereas the MHL assesses actual performance, that is, the identification of MH conditions and specific knowledge (eg, potential helpers, appropriate interventions), judged in terms of true or false. Indeed, perceptions differ from actual performance, even within the same domain, as evident from modest to nil associations between perceived and performed eHL [9] or in numeracy.

The generalizability of the study is limited by a specific sample, which consisted mostly of highly educated, Jewish, Hebrew-speaking adults with relatively high perceived health and financial stability. Future studies could test the association in broader and more diverse populations and tease apart whether the mostly nil associations are due to measurement (assessing perceptions versus performance) or whether they reflect a genuine lack of association by using different tools.

Acknowledgments

No generative artificial intelligence or artificial intelligence–assisted technologies were used in the conception, analysis, or writing of this manuscript.

XSL•FO
RenderX

Funding

Funding was provided from Ruppin Academic Center's researcher grants to EN and RY.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

EN, NG, and RY conceived and planned the study. NG coordinated data collection, and EN and NG conducted the data analysis and wrote the first draft; all authors commented on the drafts.

Conflicts of Interest

None declared.

References

1. Norman CD, Skinner HA. eHealth literacy: essential skills for consumer health in a networked world. J Med Internet Res 2006 Jun 16;8(2):e9. [doi: 10.2196/jmir.8.2.e9] [Medline: 16867972]
2. Kim K, Shin S, Kim S, Lee E. The relation between eHealth Literacy and health-related behaviors: systematic review and meta-analysis. J Med Internet Res 2023 Jan 30;25:e40778. [doi: 10.2196/40778] [Medline: 36716080]
3. Xie L, Zhang S, Xin M, Zhu M, Lu W, Mo PKH. Electronic health literacy and health-related outcomes among older adults: a systematic review. Prev Med 2022 Apr;157:106997. [doi: 10.1016/j.ypmed.2022.106997] [Medline: 35189203]
4. Jorm AF. Mental health literacy: empowering the community to take action for better mental health. Am Psychol 2012 Apr;67(3):231-243. [doi: 10.1037/a0025957] [Medline: 22040221]
5. Andrade LH, Alonso J, Mneimneh Z, et al. Barriers to mental health treatment: results from the WHO World Mental Health surveys. Psychol Med 2014 Apr;44(6):1303-1317. [doi: 10.1017/S0033291713001943] [Medline: 23931656]
6. The HLS 19-DIGI instruments for measuring digital health literacy development of the instruments. M-POHL. 2024. URL: https://m-pohl.net [accessed 2025-12-12]
7. Reavley NJ, Jorm AF. National survey of mental health literacy and stigma. : Department of Health and Ageing; 2011.
8. Knaifel E, Youngmann R, Neter E. Immigrant generation, acculturation, and mental health literacy among former Soviet Union immigrants in Israel. Int J Soc Psychiatry 2023 May;69(3):724-734. [doi: 10.1177/00207640221134236] [Medline: 36409068]
9. Neter E, Brainin E. Perceived and performed eHealth literacy: survey and simulated performance test. JMIR Hum Factors 2017 Jan 17;4(1):e2. [doi: 10.2196/humanfactors.6523] [Medline: 28096068]

Abbreviations:

eHL: electronic health literacy
MH: mental health
MHL: mental health literacy

Edited by J Sarvestan; submitted 01.05.25; peer-reviewed by C Smith, P Matthews, R Menhas; revised version received 13.11.25; accepted 13.11.25; published 19.12.25.

Please cite as:
Neter E, Youngmann R, Gruper N
Association Between eHealth Literacy and Mental Health Literacy: Cross-Sectional Study
JMIR Form Res 2025;9:e76812
URL: https://formative.jmir.org/2025/1/e76812
doi:10.2196/76812

© Efrat Neter, Refael Youngmann, Naama Gruper. Originally published in JMIR Formative Research (https://formative.jmir.org), 19.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Move – Mix It Up – Make It Fun: Summarizing and Visualizing Physical Activity Guidance

Julian Brummer[1,2], BSc, MSc; Florian Herbolsheimer[1], Dr; Mona Illmann[1,2], BSc, MSc; Duc Truong Nguyen[1], BSc, MSc; Karen Steindorf[1], Prof Dr

[1]Division of Physical Activity, Cancer Prevention and Survivorship, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 581, Heidelberg, Germany

[2]Medical Faculty, Heidelberg University, Heidelberg, Germany

Corresponding Author:
Karen Steindorf, Prof Dr
Division of Physical Activity, Cancer Prevention and Survivorship, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 581, Heidelberg, Germany

Abstract

We derived a slogan summarizing contemporary physical activity guidance, “Move – Mix it up – Make it fun,” and created an infographic designed to illustrate these recommendations and support their use in physical activity campaigns.

(JMIR Form Res 2025;9:e84225) doi:10.2196/84225

KEYWORDS

physical activity; infographic; public health; slogan; AI; artificial intelligence

Introduction

Lifestyle factors like physical activity (PA) have a profound impact on human health. Hence, national and international public health authorities publish guidelines on these factors. Concise, easy-to-understand, actionable, and accurate public health messages derived from these extensive guidelines are desirable to improve their communication to the general public [1]. The simple three-line verse “Eat food. Not too much. Mostly plants” [2] sets an example here, providing a simplified summary of dietary guidance. To date, no similar slogan exists to summarize PA guidance. We aimed to close this gap and derive such a slogan for PA, based on, but not limited to, the 2020 PA guidelines of the World Health Organization [3]. In a second step, we aimed to design an appealing infographic to visualize the recommendations and support their use in PA campaigns.

Methods

Ethical Considerations

Ethics approval was not applicable as no human participants were involved in this research.

Slogan

Our starting point was our own take on the most central public health messages arising from the scientific literature on PA for health. A first draft was discussed among the authors, aiming to improve its content and language. To complement our own ideas, we prompted various large language models (ie, Perplexity AI, ChatGPT, Claude AI) to generate ten ideas each for summarizing contemporary PA guidelines [4-6]. The suggestions of the large language models contained a wide range of summaries. We prioritized understandable, readable, memorable, and actionable messages primarily targeting adults and taking a meta-level perspective focusing on the essence instead of details. We also aimed to address the behavioral aspect of maintaining PA over time. Most importantly, the slogan had to be consistent with the literature on PA and health [3-7]. Taking these objectives into consideration, we finalized the slogan.

Textbox 1 shows the suggestions made by three different large language models when asked to summarize contemporary PA guidance to optimize the PA slogan. For Perplexity AI and ChatGPT, the following prompt was used: “Based on the attached reports, can you give me ten ideas on how to summarize the core guidance regarding physical activity? Please consider the following criteria for deriving your ideas: (1) three brief statements each, similar to the dietary guidance “Eat food. Not too much. Mostly plants,” (2) easy to understand, (3) concise, and (4) most importantly: scientifically accurate-please be very rigorous.” Due to input restrictions, for Claude AI, an adapted prompt without any attachments was used: “Based on the US, World Health Organization, and Canadian physical activity guidelines from 2018 - 2020, can you give me ten ideas on how to summarize the core guidance regarding physical activity? Please consider the following criteria for deriving your ideas: (1) three brief statements each, similar to the dietary guidance “Eat food. Not too much. Mostly plants,” (2) easy to understand,

XSL•FO
RenderX

(3) concise, and (4) most importantly: "scientifically accurate-please be very rigorous."

Textbox 1. Overview of suggestions made by the large language models.

Perplexity AI

1. Move more. Sit less. Every bit counts.
2. Be active. Be regular. Mix intensities.
3. Aim for 150. Add strength. Break it up.
4. Start now. Progress gradually. Make it routine.
5. Any activity helps. More is better. Avoid none.
6. Move daily. Challenge muscles. Support balance.
7. Walk often. Lift things. Play safely.
8. Be active together. Enjoy movement. Keep it varied.
9. Reduce sitting. Increase steps. Find your way.
10. Move with purpose. Rest enough. Stay consistent.

ChatGPT with reasoning feature

1. Move 150. Lift 2×. Sit less.
2. Any move counts. More is better. Keep going.
3. 150 min. 75 min. Your choice.
4. Light steps. Big gains. Sit less.
5. Lift twice. Balance twice. Thrive.
6. Break up sitting. Move more. Live better.
7. Sleep well. Move well. Age well.
8. Kids on the move. Habits for life.
9. One bout today. One less drug tomorrow.
10. Go outside. Get stronger. Feel brighter.

Claude AI

1. Move daily. Build strength. Limit sitting.
2. Every minute counts. Strengthen twice weekly. Break up sitting.
3. Be active daily. Build muscle. Stand more.
4. Move more. Lift regularly. Sit less.
5. Start somewhere. Progress gradually. Stay consistent.
6. Sweat often. Strengthen muscles. Reduce screens.
7. Play daily. Build strength. Move frequently.
8. Exercise weekly. Strengthen regularly. Stand hourly.
9. Move your body. Challenge your muscles. Limit lounging.
10. Activity first. Strength second. Sitting last.

Infographic

Aiming to create a visually appealing infographic that prioritized intuitive understandability and inclusivity, we consulted DALL-E to garner initial inspiration and an expert on visual communication of science who provided feedback on an early draft. Extensive work on various drafts among the multidisciplinary author group and retrieving the opinions of uninvolved individuals resulted in the final infographic.

Results

Ultimately, the discussions resulted in the final slogan: "Move–Mix it up–Make it fun." Beyond meeting all of the above criteria—most importantly, being in line with current scientific evidence [3-9]—it allowed us to use an alliteration, which is a powerful linguistics tool [10]. The accompanying infographic is shown in Figure 1.

XSL•FO
RenderX

Figure 1. Infographic for the slogan "Move – Mix it up – Make it fun," which was derived to summarize contemporary physical activity guidance by combining evidence from the scientific literature and the use of large language models to optimize its content and phrasing. The depicted infographic was designed to illustrate these recommendations and support their use in physical activity campaigns. The human drawings were sourced from The Noun Project originally [11] and were modified to depict more diversity and match the slogan; the attributions for the specific drawings are as follows: "pet walk", "Cycling", "weightlifting", "Exercise", "Dancing", "Volleyball", "rugby playing", and "Kick scooter" icons by Brickclay from thenounproject.com CC BY 3.0 and "heart health muscle" icon by Md Ahasan Habib from thenounproject.com CC BY 3.0. Edited by D.T. Nguyen. The infographic is published under a Creative Commons Attribution 4.0 International License [12].

XSL•FO
RenderX

Discussion

The present work presents a three-line verse that summarizes contemporary PA guidance—"Move – Mix it up – Make it fun."—and an accompanying infographic that visualizes its contents.

Our slogan is in line with the current scientific literature on PA for health. Its first line, "Move," encourages any type and amount of PA and therefore focuses on the great benefits that can be gained by going from no PA to some PA [3,4]. This acknowledges that health-promoting PA can also be done outside of the exercise domain (eg, active transport) [3], many different activities confer health benefits [3,8], and even light-intensity and lower volumes of PA can have substantial health benefits [3,6]. The second line gets more granular since it advises individuals to "Mix it up" and therefore include variation in their PA. This pertains primarily to the recommendation to engage in a combination of aerobic PA (eg, running, cycling) and muscle-strengthening activities (eg, resistance training) [3]. The third line, "Make it fun," emphasizes the relevance of finding and engaging in activities and contexts (eg, those involving social interaction) that one personally enjoys. Unlike the preceding statements, this one is not a PA recommendation in itself, but rather a behavioral strategy designed to support the long-term maintenance of PA [9].

Our work illustrates the potential of combining large language models with human expertise for creating public health PA messaging. Yet, there are also limitations. First, our slogan is only one potential way of summarizing PA guidance; by taking a meta-level approach, specifics from PA guidelines were not incorporated. Second, this report presents an initial step. The effectiveness of the slogan and the accompanying infographic in changing PA attitudes, intentions, or behavior has not been tested yet; the present work serves as the basis for such future investigations.

After future evaluation, the slogan "Move – Mix it up – Make it fun." may be used as a concise, memorable, easy-to-understand, and scientifically accurate tool to communicate actionable PA advice to the general public, for example in social media campaigns or related public health messaging—especially when longer messaging is not feasible or desired.

Acknowledgments

Perplexity AI [13], ChatGPT [14], and Claude AI [15] were used to optimize the slogan. The ChatGPT companion DALL-E [16] was used to garner initial inspiration for creating the infographic.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability

Data sharing is not applicable to this article as no datasets were generated or analyzed during this study.

Authors' Contributions

Conceptualization: DTN, FH, JB, KS, MI
Investigation: DTN, FH, JB, KS, MI
Methodology: DTN, FH, JB, KS, MI
Supervision: FH, KS
Visualization: DTN, FH, JB, KS, MI
Writing–original draft: DTN, JB, MI
Writing–review & editing: DTN, FH, JB, KS, MI

Conflicts of Interest

None declared.

References

1. Centers for Disease Control and Prevention (U.S.), Office of the Associate Director for Communication, Strategic and Proactive Communication Branch. Simply put; a guide for creating easy-to-understand materials. 2010 URL: https://stacks.cdc.gov/view/cdc/11938 [accessed 2025-11-23]
2. Pollan M. Unhappy meals. The New York Times Magazine. 2007 URL: https://www.nytimes.com/2007/01/28/magazine/28nutritionism.t.html [accessed 2025-11-23]
3. Bull FC, Al-Ansari SS, Biddle S, et al. World Health Organization 2020 guidelines on physical activity and sedentary behaviour. Br J Sports Med 2020 Dec;54(24):1451-1462. [doi: 10.1136/bjsports-2020-102955] [Medline: 33239350]

XSL•FO
RenderX

4. U.S. Department of Health and Human Services, 2018. 2018 Physical Activity Guidelines Advisory Committee Scientific Report. 2018 URL: https://odphp.health.gov/sites/default/files/2019-09/PAG_Advisory_Committee_Report.pdf [accessed 2025-11-23]
5. WHO guidelines on physical activity and sedentary behaviour. : World Health Organization; 2020 URL: https://www.who.int/publications/i/item/9789240015128 [accessed 2025-11-23]
6. Ross R, Chaput JP, Giangregorio LM, et al. Canadian 24-Hour Movement Guidelines for adults aged 18-64 years and adults aged 65 years or older: an integration of physical activity, sedentary behaviour, and sleep. Appl Physiol Nutr Metab 2020 Oct;45(10 (Suppl. 2):S57-S102. [doi: 10.1139/apnm-2020-0467] [Medline: 33054332]
7. Ekelund U, Sanchez-Lastra MA, Dalene KE, Tarp J. Dose-response associations, physical activity intensity and mortality risk: a narrative review. J Sport Health Sci 2024 Jan;13(1):24-29. [doi: 10.1016/j.jshs.2023.09.006] [Medline: 37734548]
8. Oja P, Memon AR, Titze S, et al. Health benefits of different sports: a systematic review and meta-analysis of longitudinal and intervention studies including 2.6 million adult participants. Sports Med Open 2024 Apr 24;10(1):46. [doi: 10.1186/s40798-024-00692-x] [Medline: 38658416]
9. Madigan CD, Grimmett C, Sweet SN, et al. Understanding adult's experiences and perceptions of how to maintain physical activity: a systematic review and qualitative synthesis. Int J Behav Med 2024 Nov 18. [doi: 10.1007/s12529-024-10335-w] [Medline: 39557793]
10. Boers F, Lindstromberg S. Finding ways to make phrase-learning feasible: the mnemonic effect of alliteration. System 2005 Jun;33(2):225-238. [doi: 10.1016/j.system.2004.12.007]
11. The Noun Project. URL: https://thenounproject.com/ [accessed 2025-11-25]
12. Attribution 4.0 international (CC BY 4.0). Creative Commons. URL: https://creativecommons.org/licenses/by/4.0/ [accessed 2025-11-24]
13. Perplexity AI. URL: https://www.perplexity.ai/ [accessed 2025-11-26]
14. ChatGPT. URL: https://chatgpt.com/ [accessed 2025-11-26]
15. Claude AI. URL: https://claude.ai/ [accessed 2025-11-26]
16. DALL-E AI Image Generatorchat. URL: https://chat.ai-pro.org/chat/dall-e/ [accessed 2025-11-26]

Abbreviations

AI: artificial intelligence
PA: physical activity

Edited by A Mavragani; submitted 16.09.25; peer-reviewed by CTV Swain, L Lafave; accepted 03.11.25; published 28.11.25.

Please cite as:
Brummer J, Herbolsheimer F, Illmann M, Nguyen DT, Steindorf K
Move – Mix It Up – Make It Fun: Summarizing and Visualizing Physical Activity Guidance
JMIR Form Res 2025;9:e84225
URL: https://formative.jmir.org/2025/1/e84225
doi:10.2196/84225

© Julian Brummer, Florian Herbolsheimer, Mona Illmann, Duc Truong Nguyen, Karen Steindorf. Originally published in JMIR Formative Research (https://formative.jmir.org), 28.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Effectiveness of Text Message Outreach to Promote Enrollment in a Government Food Assistance Program: Pilot Randomized Trial

Emily M Melnick[1,2*], MPH, PhD; Francesco Acciai[1*], PhD; Nicole Vaudrin O'Reilly[3], MS, RD; Mindy Jossefides[3], RD; A Bea Ronan[1], MNSP; Punam Ohri-Vachaspati[1], PhD, RD

[1]College of Health Solutions, Arizona State University, 425 N 5th St, Phoenix, AZ, United States

[2]Department of Psychology, University of Denver, Denver, CO, United States

[3]Inter Tribal Council of Arizona, Phoenix, AZ, United States

[*]these authors contributed equally

Corresponding Author:
Emily M Melnick, MPH, PhD
College of Health Solutions, Arizona State University, 425 N 5th St, Phoenix, AZ, United States

Abstract

Findings from a pilot randomized trial of 916 households indicated that outreach text messages were not an effective strategy to increase enrollment in the United States Special Supplemental Nutrition Program for Women, Infants, and Children (WIC) among eligible households in Arizona.

(JMIR Form Res 2025;9:e78907) doi:10.2196/78907

KEYWORDS

United States; WIC; SNAP; text messaging; intervention; women, infants, and children; Supplemental Nutrition Assistance Program

Introduction

The Special Supplemental Nutrition Program for Women, Infants, and Children (WIC) provides nutritious food packages, nutrition education, breastfeeding support, and health care referrals for pregnant and postpartum women and children aged 0 - 4 years from low-income households in the United States [1]. WIC participation improves dietary and health outcomes and reduces health disparities [2]. However, many eligible households do not participate. Among those eligible for both WIC and the Supplemental Nutrition Assistance Program (SNAP) [3], another US food assistance program, about half participate in only SNAP and miss the additional benefits WIC provides [4]. In response, US agencies have called for outreach efforts to boost enrollment among SNAP-participating households eligible but not participating in WIC [5]. Increasing WIC participation is particularly important among American Indians who face higher rates of food insecurity [6], alongside declining WIC participation [7].

Text message outreach may offer a low-cost, scalable approach to encourage WIC enrollment among eligible SNAP-participating households [2]. No studies have evaluated impacts of such outreach in American Indian populations. This study evaluated a pilot randomized text message intervention aimed at increasing enrollment in the Inter Tribal Council of Arizona, Inc. (ITCA) WIC program. We hypothesized that households receiving persuasive outreach texts would be more likely to enroll in ITCA WIC than those who did not.

Methods

Study Overview

We conducted a pilot randomized trial among likely-WIC-eligible SNAP households in the ITCA WIC service area, including urban areas and tribal lands. The Arizona Department of Economic Security, which administers SNAP, provided the target population list. After eligibility screening, households were randomized in a 1:1 ratio to intervention and delayed control groups using a computerized random number generator. Participants were blinded to trial arm allocation, while evaluators were not. The study adhered to CONSORT (Consolidated Standards of Reporting Trials) guidelines (Checklist 1).

In February 2024, intervention households were sent a persuasive outreach text message via Teletask (Fair Oaks) that included the message, "Want healthy foods for your family? WIC is just a click away. With SNAP, you may already qualify." The text message was developed by the study team in partnership with a communication specialist then field-tested with ITCA WIC local agency staff and low-income adults to assess persuasiveness and relevance for the target audience. Control households did not receive any texts during the 3-month

XSL•FO
RenderX

(February-April 2024) evaluation period. To assess intervention effects, we conducted a per-protocol analysis, excluding 177 households in the intervention group that did not successfully receive the text message. Excluded households were demographically similar to those that successfully received the message, which, in turn, were similar to the delayed control group.

All analyses were run in Stata 16 (StataCorp). We used the command "proportion" to obtain group-specific enrollment proportions, estimated standard errors, and logit-transformed confidence intervals. A Pearson χ^2 test (α=.05, two-tailed) compared enrollment rates across groups. The final analytical sample included 916 likely-WIC-eligible households with no missing data.

Ethical Considerations

The intervention was implemented by ITCA WIC using data shared through an agreement with the Arizona Department of Economic Security. Because the text messages were sent as a part of programmatic improvement efforts and data shared were per a data sharing agreement protocol, consent to participate in this study was not required. All data were de-identified prior to analysis. Households did not receive compensation. The Arizona State University Institutional Review Board approved study protocols.

Results

A CONSORT diagram (Figure 1) outlines intervention allocation and participant flow. Over half of the sample (516/916, 56.3%) identified as American Indian and 85.2% (780/916) lived in urban areas. During the 3-month follow-up, 54 of the 916 households (5.9%) enrolled in ITCA WIC. Enrollment was 4.3% (95% CI 2.6% - 7.0%) among households that received the text, while it was 6.9% (95% CI 5.1% - 9.3) among delayed controls. The difference was not statistically significant (*P*=.097, Pearson χ^2 test).

XSL•FO
RenderX

Figure 1. CONSORT flow diagram.

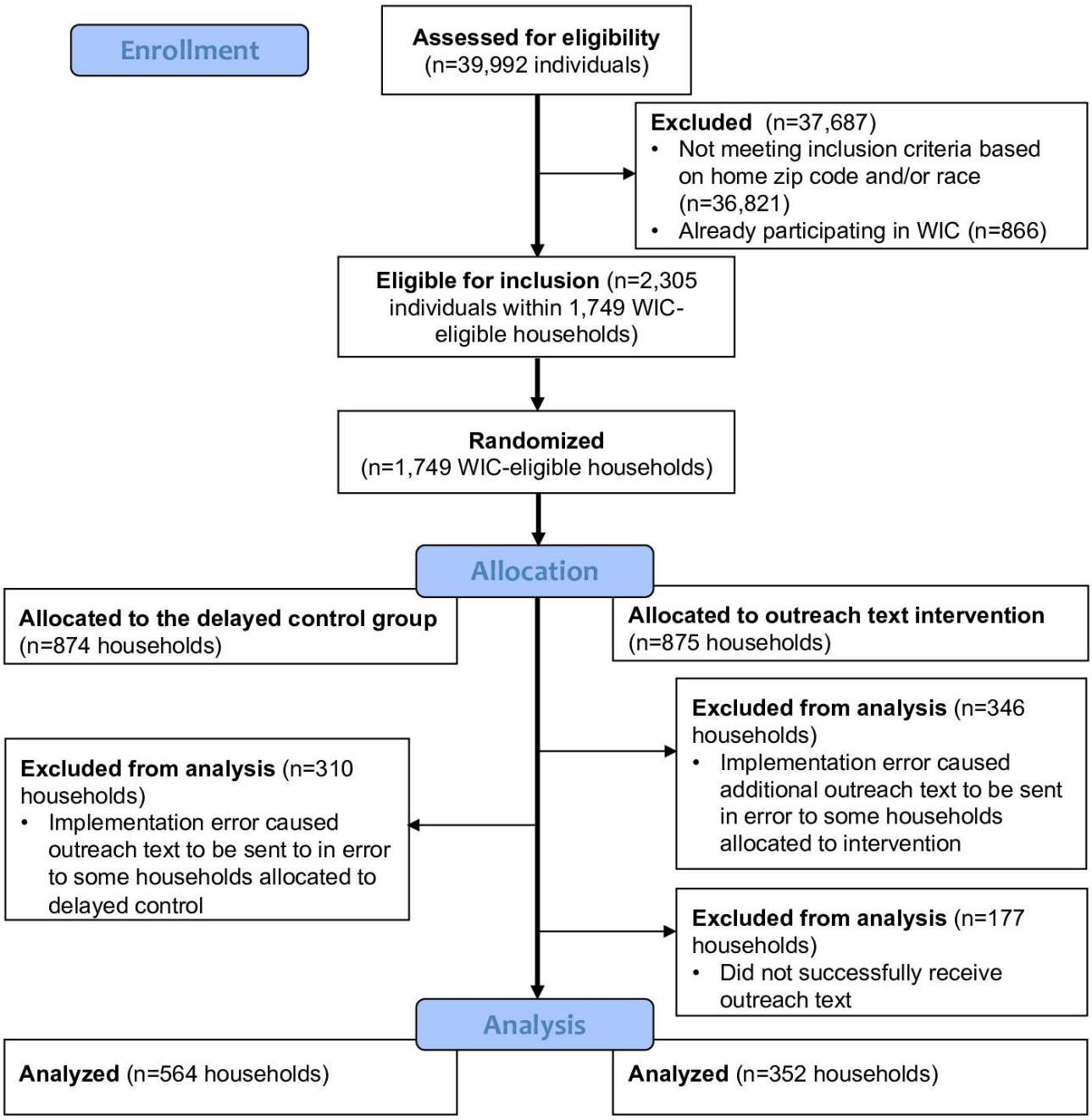

Discussion

This pilot study found no evidence that a text message increased WIC enrollment among likely-eligible SNAP households. This finding is consistent with results from pilot studies in Colorado, Montana, and Virginia [8]. While text interventions have improved outcomes such as retention of current WIC participants [9] and enrollment among WIC-eligible Medicaid participants [10], more intensive or multi-pronged approaches may be needed to increase WIC enrollment among eligible non-participating SNAP households. Notably, enrollment rates were low in both intervention and delayed control households. Possible reasons include children aging out of WIC eligibility before receiving the text and families already receiving services from another WIC agency in the state. Given this study's focus on American Indian communities, future research should explore whether similar text-based interventions are effective in other populations and countries offering similar programs. We did not conduct a power calculation prior to the trial as we enrolled all eligible households; post-hoc analyses confirm the study was adequately powered to detect group differences. We also cannot rule out the possibility that intervention households shared messages with others. Future randomized trials should address the potential for cross-contamination between study arms. Trials with longer follow-up periods and those that test effects of multiple messages are also warranted.

XSL•FO
RenderX

Acknowledgments

This work was supported by federal funds from the U.S. Department of Agriculture. The contents of this publication do not necessarily reflect the view or policies of the U.S. Department of Agriculture, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government. The study funder had no role in the design, execution, or interpretation of the research.

Data Availability

Data associated with this report are not able to be shared because they belong to a tribal entity.

Authors' Contributions

NVO'R, MJ, FA, and POV conceived the study and designed the statistical methods. FA performed the data analysis. FA and EMM wrote the first manuscript draft. NVO'R, MJ, ABR, and POV critically reviewed and edited the manuscript. All authors helped to interpret the results, critically revised the manuscript, and provided final approval of the version for publication.

Conflicts of Interest

None declared.

Checklist 1
CONSORT checklist
[PDF File, 158 KB - formative_v9i1e78907_app1.pdf]

References

1. Special supplemental nutrition program for women, infants, and children (WIC). US Department of Agriculture, Food and Nutrition Service. URL: https://www.fns.usda.gov/wic [accessed 2023-01-21]
2. Colman S, Nichols-Barrer IP, Redline JE, Devaney BL, Ansell SV, Joyce T. Effects of the special supplemental nutrition program for women, infants, and children (WIC): a review of recent research. : US Department of Agriculture, Food and Nutrition Service, Office of Research and Analysis; 2012 Jan WIC-12-WM URL: https://fns-prod.azureedge.us/sites/default/files/WICMedicaidLitRev.pdf [accessed 2025-10-20]
3. Supplemental nutrition assistance program (SNAP). US Department of Agriculture, Food and Nutrition Service. URL: https://www.fns.usda.gov/snap [accessed 2023-01-21]
4. Neuberger Z. WIC coordination with medicaid and SNAP. Center on Budget and Policy Priorities. URL: https://www.cbpp.org/research/food-assistance/wic-coordination-with-medicaid-and-snap-1 [accessed 2025-02-25]
5. WIC policy memorandum #2023-5: data sharing to improve outreach and streamline certification in WIC. US Department of Agriculture, Food and Nutrition Service. URL: https://www.fns.usda.gov/wic/data-sharing [accessed 2025-07-15]
6. Nikolaus CJ, Johnson S, Benally T, et al. Food Insecurity among American Indian and Alaska Native People: A Scoping Review to Inform Future Research and Policy Needs. Adv Nutr 2022 Oct 2;13(5):1566-1583. [doi: 10.1093/advances/nmac008] [Medline: 35092417]
7. WIC participant and program characteristics 2018. US Department of Agriculture, Food and Nutrition Service. 2018. URL: https://www.fns.usda.gov/research/wic/participant-program-characteristics-2018 [accessed 2025-02-25]
8. Maneely J, Neuberger Z. Using data matching and targeted outreach to enroll families with young children in WIC. Center on Budget and Policy Priorities. URL: https://www.cbpp.org/research/food-assistance/using-data-matching-and-targeted-outreach-to-enroll-families-with-young [accessed 2025-02-25]
9. Davis RA, Leavitt HB, Chau M. A review of interventions to increase WIC enrollment and participation. J Community Health 2022 Dec;47(6):990-1000. [doi: 10.1007/s10900-022-01131-2] [Medline: 35962868]
10. Maneely J, Neuberger Z. Targeted text message outreach can increase WIC enrollment, pilots show. Center on Budget and Policy Priorities. URL: https://www.cbpp.org/research/food-assistance/targeted-text-message-outreach-can-increase-wic-enrollment-pilots-show [accessed 2025-01-14]

Abbreviations

ITCA: Inter Tribal Council of Arizona, Inc.
SNAP: Supplemental Nutrition Assistance Program
WIC: Special Supplemental Nutrition Program for Women, Infants, and Children

XSL•FO
RenderX

Edited by A Mavragani; submitted 11.06.25; peer-reviewed by M Zimmer, M Estrade; revised version received 26.09.25; accepted 28.09.25; published 31.10.25.

Please cite as:
Melnick EM, Acciai F, Vaudrin O'Reilly N, Jossefides M, Ronan AB, Ohri-Vachaspati P
Effectiveness of Text Message Outreach to Promote Enrollment in a Government Food Assistance Program: Pilot Randomized Trial
JMIR Form Res 2025;9:e78907
URL: https://formative.jmir.org/2025/1/e78907
doi:10.2196/78907

© Emily M Melnick, Francesco Acciai, Nicole Vaudrin O'Reilly, Mindy Jossefides, A Bea Ronan, Punam Ohri-Vachaspati. Originally published in JMIR Formative Research (https://formative.jmir.org), 31.10.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Use First, Trust Later? Exploring How Health Care Providers View the Gaps Between AI's Regulation and Its Implementation

Aviad Raz[1], PhD; Yael Inbar[2], PhD; Ziv Paz[3], MD; Barkan Hofman[1], MA; Zalman Itzhakov[3], MD; Orly Weinstein[3], MD

[1]Department of Sociology and Anthropology, Ben-Gurion University of the Negev, Be'er-Sheva, Be'er-Sheva, Israel
[2]Technology Management and Information Systems, Tel Aviv University, Tel-Aviv, Israel
[3]Clalit Health Services, Tel-Aviv, Israel

Corresponding Author:
Aviad Raz, PhD
Department of Sociology and Anthropology, Ben-Gurion University of the Negev, Be'er-Sheva, Be'er-Sheva, Israel

Abstract

The regulatory focus on premarket approval, overlooking and limiting the life cycle, real-world adaptiveness potential of artificial intelligence in health care, requires ongoing oversight and quality assurance by health care providers–cum-deployers, especially during early adoption.

(JMIR Form Res 2025;9:e74038) doi:10.2196/74038

KEYWORDS

artificial intelligence; regulation; implementation; health care management; radiology; content analysis; qualitative

Introduction

This study explores how the specific nature of regulating health care artificial intelligence (AI) influences managing its deployment. Food and Drug Administration and European Commission regulations focus on premarket approval yet fail to address monitoring of AI postmarketing [1]. Better regulations for "adaptive" AI systems are needed. These are considered with precaution as "high-risk" and prohibited from fully autonomous diagnosis. As AI models are exposed to new data in clinical settings, their performance may degrade or alter over time, necessitating ongoing oversight [2,3]. This often means that health care providers must step into the regulatory ambiguity zone to develop local protocols for quality assurance during early adoption phases.

Methods

Overview

In-depth semistructured interviews (30‑80 min) were conducted with 65 (29 females, 36 males) managers and physicians (mainly radiologists) from 4 different Israeli hospitals. Participants, identified by hospital management, were voluntarily included as deployers of AI for medical imaging analysis. Using a grounded theory approach [4], preliminary codes were identified based on the literature review, with additional themes identified through interview analysis. The research team did the coding in parallel to ensure interrater reliability and to discuss further emerging codes. Data analysis was completed upon agreement that thematic saturation was reached.

Ethical Considerations

This research was approved by Clalit Health Services Ethical Review Board (number 0001‑22-CNT-C). Participants signed an informed consent form; all details were anonymized to maintain privacy and confidentiality; no compensation was provided to participants.

Results

Overview

Many of our respondents criticized the absence of clear regulations for AI implementation, seen as challenging and at times confusing, yet also enabling experimentation.

The Precautionary Nature of AI Regulation in Health Care and Its Postdeployment Implications

Current premarket approval processes require AI-based medical devices to stop their retraining postdeployment. This reflects precautionary regulation regarding health care AI. Our respondents expressed worries about having to become the "AI's QA [quality assurance] testers": "I cannot always rely on the AI in trauma cases where the injection is not good enough and it does not detect heart failure...At nights I filter the hemorrhages alerts because it is over-sensitive...I wish I could teach it, talk to it, like a real team member" (N8).

AI systems were considered "overalerting," and managers instructed physicians to customize the alert threshold according

XSL•FO
RenderX

to their needs: “We got used to all sorts of false alerts. So, we rely more on our experience and knowledge, and we combine that with the system” (M6). Managers and local “champions” (“superuser” radiologists) had to become AI supervisors: ”I make everyone here provide feedback regarding the AI system” (M5). Playing an active role in shaping the use of AI systems was nevertheless seen positively: “We feel like we are part of the development team. It feels good. It makes me even more engaged” (R1).

Discussion

Our main study findings show how regulating health care AI influences managing its deployment by highlighting the need for customization of the AI tools that require experimentation. Strategies included vendor-appointed “champions” and on-site feedback loops [5]. Nevertheless, such customization requires further monitoring of bias metrics in model outputs and selective use by physicians.

There are some limitations to consider. Findings were drawn from interviews with radiologists, considered early adaptors of technologies, whose experiences may vary from that of other health providers. Although the study captures early adoption phases, future research should explore how this might shape long-term AI integration. In conclusion, the broader implications are that to become a sustainable approach in medical AI implementation, the “use first, trust later” mindset cannot be left unmonitored. Regulation should address adaptive AI and consider how transparency (rather than explainability) of the initial and postdeployment training, as well as bias monitoring, can enable oversight. Additionally, hospital administrators should work with AI developers to monitor biases in customization [6], as well as manage effective teaming routines with AI.

Acknowledgments

The authors acknowledge support from the Israel Science Foundation grant number 1325/23. We are grateful to all our team members at the AI in Healthcare Organizations Lab [7] and to the interviewees who participated in the study.

Data Availability

Anonymized transcripts are available from the corresponding author.

Authors' Contributions

AR: conceptualization, writing, formal analysis. YI: writing, formal analysis. ZP: review & editing, funding acquisition. BH: investigation, formal analysis, review & editing. ZI and OW: project administration.

Conflicts of Interest

None declared.

References

1. Agarwal R, Dugas M, Gao G(. Augmenting physicians with artificial intelligence to transform healthcare: Challenges and opportunities. Economics Manag Strategy 2024 Mar;33(2):360-374. [doi: 10.1111/jems.12555]
2. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Med 2021;27(1):9-20.
3. Lee P, Kummerfeld E, Yu KH. Regulatory oversight of artificial intelligence and machine learning in healthcare. J Healthc Inform Res 2020;4(1):32-54.
4. Gioia DA, Corley KG, Hamilton AL. Seeking Qualitative Rigor in Inductive Research. Organ Res Methods 2013 Jan;16(1):15-31. [doi: 10.1177/1094428112452151]
5. Faulconbridge J, Sarwar A, Spring M. How Professionals Adapt to Artificial Intelligence: The Role of Intertwined Boundary Work. J Management Studies 2025 Jul;62(5):1991-2024. [doi: 10.1111/joms.12936]
6. Cross JL, Choma MA, Onofrey JA. Bias in medical AI. PLOS Digit Health 2024 Jul;3(11):e0000651. [doi: 10.1371/journal.pdig.0000651]
7. AI in Healthcare Organizations Lab. URL: https://aihealthcarelab.com/ [accessed 2025-11-03]

Abbreviations

AI: artificial intelligence

XSL•FO
RenderX

Edited by A Mavragani, J Sarvestan; submitted 16.03.25; peer-reviewed by A Adekola, C Onah, S Wagle; revised version received 23.09.25; accepted 24.09.25; published 10.11.25.

Please cite as:
Raz A, Inbar Y, Paz Z, Hofman B, Itzhakov Z, Weinstein O
Use First, Trust Later? Exploring How Health Care Providers View the Gaps Between AI's Regulation and Its Implementation
JMIR Form Res 2025;9:e74038
URL: https://formative.jmir.org/2025/1/e74038
doi:10.2196/74038

© Aviad Raz, Yael Inbar, Ziv Paz, Barkan Hofman, Zalman Itzhakov, Orly Weinstein. Originally published in JMIR Formative Research (https://formative.jmir.org), 10.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

A Video-Based Intervention Supporting Spanish-Speaking Fathers of Infants in the Neonatal Intensive Care Unit: Open-Label Pilot Study

Janine Bernardo[1], MD, MPH; Elise Belkin[1], BS; Suzanne Hally[1], RN; Sara Boyle[1], RN; Eric Van[1], MS; Jose Tascon-Arcila[2], MD; Emily Herzberg[1], MD; Juan Matute[3], MD; Paul Lerou[1], MD

[1]Division of Newborn Medicine, Department of Pediatrics, Massachusetts General Hospital, 55 Fruit Street, BUL 2, Boston, MA, United States

[2]Department of Pediatrics, Boston Children's Hospital, Boston, United States

[3]Division of Neonatology, Department of Pediatrics, Perelman School of Medicine at the University of Pennsylvania, Children's Hospital of Philadelphia, Philadelphia, United States

Corresponding Author:
Janine Bernardo, MD, MPH
Division of Newborn Medicine, Department of Pediatrics, Massachusetts General Hospital, 55 Fruit Street, BUL 2, Boston, MA, United States

Abstract

This open-label pilot study showed the feasibility and acceptability of an educational video intervention for Spanish-speaking fathers in the neonatal intensive care unit to increase knowledge and confidence, decrease anxiety, and improve parenting competency.

(JMIR Form Res 2025;9:e72520) doi:10.2196/72520

KEYWORDS

neonatal intensive care unit; NICU; language other than English; Spanish-speaking; fathers; infants; videos; pilot study; feasibility; accessibility

Introduction

Parental involvement for infants admitted to the neonatal intensive care unit (NICU) encapsulates the earliest, critical moments of interaction with a child, yet can be accompanied with stress, anxiety, and uncertainty [1]. Parents who speak languages other than English are at increased risk of disrupted communication, stress, and suboptimal care [2]. NICU fathers who are from minority racial and ethnic groups have higher rates of depression and anxiety than fathers who are not [3,4], in part due to ineffective communication and targeted support/resources, resulting in feeling undervalued [3,5,6].

Various interventions have been implemented to support the well-being and involvement of NICU fathers including psychosocial support [7], written materials [8], enhanced nursing support [9], narrative writing [10], skin-to-skin, and increased physical interaction [11,12]. Despite being inexpensive and distributable, there is a paucity of video-based interventions [12] targeting Spanish-speaking NICU fathers. Prior work demonstrates that NICU fathers find staff education, particularly from nurses, pivotal in overcoming low knowledge and uncertainty, leading to improved confidence and overall NICU experience [13]. Thus, a video-based, educational intervention for Spanish-speaking fathers would provide an opportunity to better support fathers and the child-father relationship at its earliest stages.

The Bronfenbrenner's ecological systems theory posits an individual's health and development is shaped by varying layers of environmental and social relations from the micro to macro scale [14]. Over the course of a lifetime, an individual's development evolves through shared, increasingly complex interactions between the individual and their environment. This project aimed to support the early child-father relationship to improve fathers' knowledge and confidence such that the reciprocal, foundational nature of their relationship may flourish.

We performed a single-arm intervention feasibility study using short videos on high-yield NICU topics. Our objective was to assess whether the intervention "Los Padres de los Pajaritos" ("the fathers of little birds") would be feasible and acceptable to Spanish-speaking NICU fathers to assess improved knowledge and confidence, decreased anxiety, and improved parenting competency to lay the groundwork for future, larger-scale efficacy trials.

Methods

Participants

The study population included Spanish-speaking and Spanish-reading fathers with infants admitted to a Level IV

XSL•FO
RenderX

NICU or Level II special care nursery (SCN) at a single, urban center from July 1, 2023, to March 31, 2024. Of note, in 2021, 7% of NICU/SCN parents identified Spanish as their preferred language. Infants who were transferred were followed after transfer, as able. The sample size was 10 fathers to identify a mean difference of 2 with ±2 standard deviations for a 5-point pre/post scale.

Video Intervention

Three short videos (5 - 10 min each) were written, edited, and recorded by neonatologists, nurses, and lactation specialists on the following topics: NICU Orientation, Nursing Care, and Lactation/Breastfeeding Support. Videos were translated into Spanish by native Spanish-speaking physicians and recorded by qualified interpreters. Videos were filmed by a trained medical videographer.

NICU Orientation introduced NICU team members, toured the unit, explained rounds, and provided information on interpretive services. The Nursing Care video demonstrated bedside skills including changing diapers, taking temperatures, swaddling, and performing skin-to-skin. The Lactation/Breastfeeding Support video discussed the benefits of breast milk using Centers for Disease Control and Prevention–generated resources, support for lactating mothers, and cleaning pump parts [15]. Fathers of patients admitted directly to the SCN were only offered Nursing Care and Lactation/Breastfeeding Support videos.

Recruitment and Data Collection

Eligible fathers were identified through electronic medical record notifications for mother's language preference and confirmed with the medical team before being approached for enrollment (Table 1). Preintervention surveys were administered within a week of admission by a research assistant and included age, employment, education, relationship status, insurance type, country of origin, and neonatal descriptors. Videos were distributed by a research assistant who demonstrated the use of QR codes to access videos. Fathers were instructed to view all videos in a preferred location either on a hospital iPad or personal device on their own time. Postintervention surveys were administered after fathers watched the videos and prior to discharge home. Upon completion, fathers were given remuneration including a gift card, books, and toys. Surveys were collected on iPads or paper and data were stored in REDCap (Research Electronic Data Capture). Standard of care, provided to all study participants, included teaching by bedside nurses, therapists, and providers. Social workers were available in the case of emotional distress, of which there were none.

XSL•FO
RenderX

Table . Study respondent sociodemographics.

Characteristic	Values
Average age (years)	31
Employed, n (%)	12 (92)
Highest level of education, n (%)	
Never graduated high school	5 (38)
High school diploma or GED[a]	7 (54)
Graduated college	1 (8)
Trade/vocational school	0 (0)
Advanced degree	0 (0)
Relationship status, n (%)	
Not in a relationship	0 (0)
In a relationship but not living with partner	1 (8)
In a relationship and living with partner but not married	7 (54)
Married and living together	5 (38)
Hispanic origin, n (%)	
Central American	6 (46)
Mexican, Mexican American, Chicano	1 (8)
Dominican Republic	2 (15)
South American	4 (31)
Languages spoken at home, n (%)	
Spanish	10 (77)
English and Spanish	3 (23)
Number of children (N=12 respondents), n (%)	
One	8 (62)
Two	3 (23)
Three	1 (8)
Four or more	0 (0)
Admission to, n (%)	
NICU[b]	11 (85)
SCN[c]	2 (15)
Previous child in NICU/SCN	0 (0)
Gestational age of child in NICU/SCN, n (%)	
<28 weeks	3 (23)
29 - 32 weeks	3 (23)
33 - 37 weeks	1 (8)
37 weeks and above	6 (46)
Insurance of child, n (%)	
Private	0 (0)
Public	9 (69)
No insurance	1 (8)
Not sure	3 (23)
Average visitation to NICU per week, n (%)	

XSL•FO
RenderX

Characteristic	Values
I do not visit	0 (0)
1 - 2 times a week	2 (15)
2 - 5 times a week	3 (23)
More than 5 times a week	8 (62)
Transportation to the hospital, n (%)	
Drive	10 (77)
Public transit	1 (8)
Driver service (Uber, Lyft, taxi, etc)	2 (15)
Walk	0 (0)
I do not visit	0 (0)

[a]GED: general education development.

[b]NICU: neonatal intensive care unit.

[c]SCN: special care nursery.

Ethical Considerations

This study was approved by the institutional review board of Mass General Brigham (number 2022P003021); all procedures were in accordance with the ethical standards of the institutional review board. Verbal, informed consent was obtained in Spanish and participants had an opportunity to ask questions in their preferred language. Confidentiality and privacy were strictly maintained. Participants were compensated with a US $30 gift card and age-appropriate toys and books.

Measures

Feasibility was based on enrollment, with a goal of >70% to be considered good and >80% to be excellent [16]. Acceptability was defined as participant satisfaction with materials, with the above goals [17].

Preintervention surveys included demographics, a 5-point Likert scale on confidence/knowledge in infant care (1 being very low confidence/knowledge; 5 being very high), the Generalized Anxiety Disorder-7 (GAD-7) scale, and the Parenting Sense of Competence scale (PSOC). Postintervention surveys included 5-point Likert scales on confidence/knowledge in infant care, the GAD-7 scale, the PSOC, and feedback.

Analysis

Demographic characteristics were described using frequencies and percentages. Preintervention and postintervention Likert scales, GAD-7 scores, and PSOC scores were analyzed using paired 1-sided *t* tests. An α level of 0.05 was used to signify statistical significance. Data were analyzed with IBM SPSS Statistics for Windows (version 28.0; IBM Corp).

Results

A total of 42 fathers met inclusion criteria; 27 [2] were approached and 21 (78%) were enrolled; 13 (62%) completed all study procedures (Figure 1). Demographics and clinical variables are reported in Table 1.

XSL•FO
RenderX

Figure 1. Schematic of eligibility, enrollment, and follow-up, showing the numbers of individuals who met study criteria, were approached and enrolled, and were lost to follow-up. Deceased refers to neonates who died during hospitalization. LOS: length of stay.

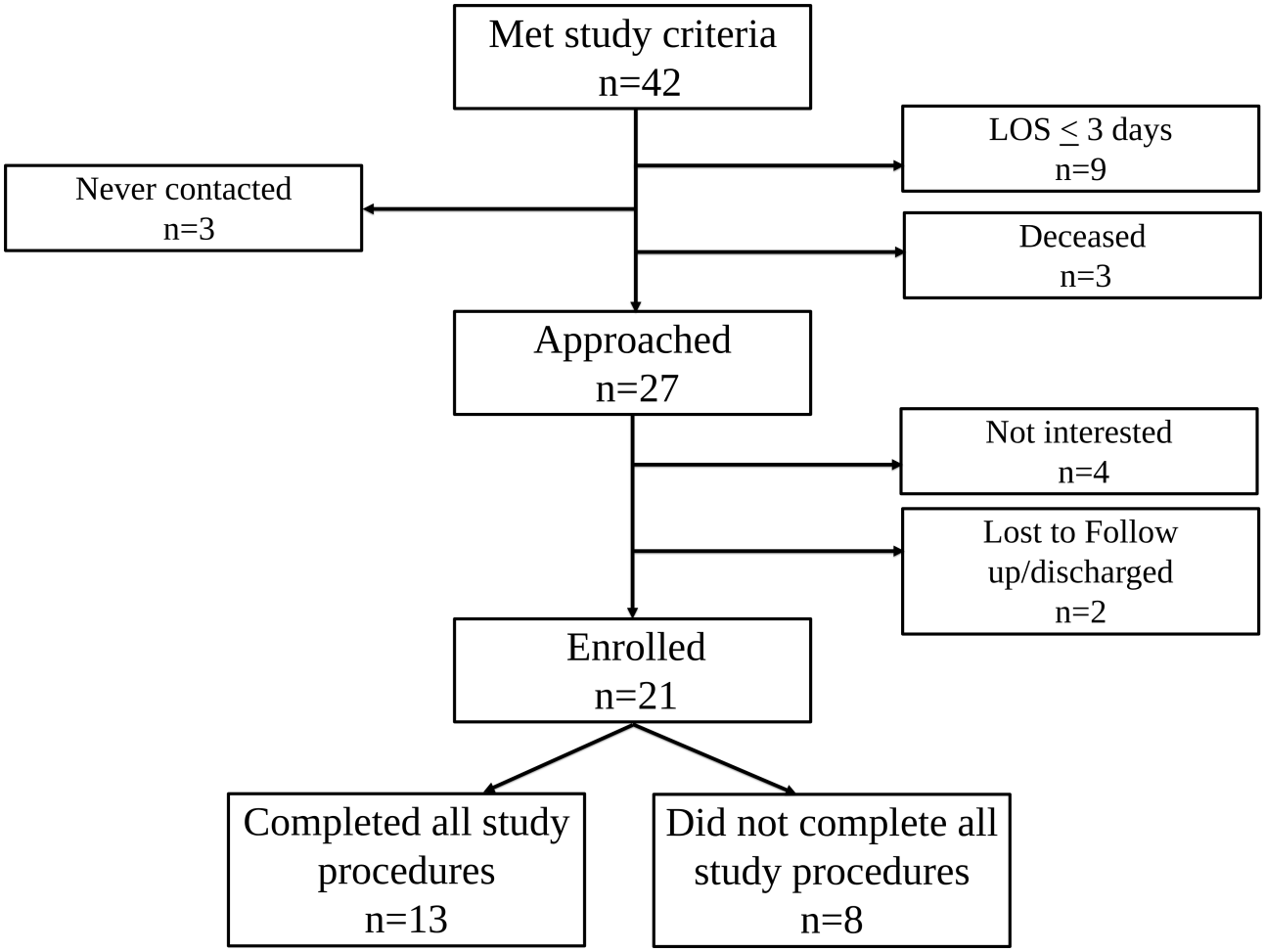

All fathers viewed at least 2 videos; 85% (11/13) viewed all videos. The Nursing Care video was most informative, while the Lactation Support video was the least watched. Likert scores on video feedback, including understandability, length, and quality, were all greater than 4.5 (Table 2). Confidence in caring for their infant showed a significant increase from 4.2 (95% CI 3.66-4.74) to 4.8 (95% CI 4.51-5.09), *P*=.02, while knowledge increased (not significantly) from 4.09 (95% CI 3.46-4.72) to 4.69 (95% CI 4.29-4.97), *P*=.08. The average GAD-7 and PSOC scores did not change significantly (GAD-7: 4.75, 95% CI 2.56-6.92 and 3.33, 95% CI 1.49-5.17; *P*=.07; PSOC: 68.5, 95% CI 60.48-76.52 and 77.2, 95% CI 64.16-76.5; *P*=.31). Select comments were positive (Table 3). Qualitative video feedback included the following: “I found all the videos very interesting and educational,” “The videos speak very clearly and precisely...it is very important to know who you are dealing with and the trust that one places in them,” “I’m looking forward to seeing more videos,” and “Don’t stop sharing them because they are very important and you really do learn a lot.”

XSL•FO
RenderX

Table . Video usage descriptors. Self-reported video usage and feedback, suggestions for future interventions, and descriptions of timing of video completion are presented.

	Values
Video watch, n/N (%)	
Introduction to the NICU[a]	11/11 (100)
Orientation to nursing care	12/13 (92)
Lactation support	9/13 (69)
Most informative video, n/N (%)	
Introduction to the NICU	5/11 (45)
Orientation to nursing care	6/13 (46)
Lactation support	2/13 (15)
Least informative video, n/N (%)	
Introduction to the NICU	1/9 (11)
Orientation to nursing care	4/9 (44)
Lactation support	4/9 (44)
Suggested means of discussing care, n/N (%)	
Pamphlets/papers	5/13 (38)
Conferences	4/13 (31)
Bedside teaching	10/13 (77)
Other	
From other providers	1/13 (7)
Suggested topics for education	
Oxygen	—[b]
Reasons for babies to cry	—
Everything	—
Average days of viewing video following enrollment, n	4.08
Average days between viewing video and completing survey, n	7.79
Average days between completing preintervention and postintervention surveys, n	13.5

[a]NICU: neonatal intensive care unit.

[b]Not applicable.

XSL•FO
RenderX

Table . Video acceptability and qualitative feedback. Preintervention and postintervention data points are described, including knowledge and confidence Likert scales, Generalized Anxiety Disorder-7 scores, and Parenting Sense of Competence Scale scores. Measures of video acceptability on a 5-point Likert scale are shown.

	Preintervention Likert scores, mean	Postintervention Likert scores, mean	*P* value
Knowledgeable on care of infant (n=11)	4.09	4.69, mean difference: 0.55 (95% CI −0.27 to 1.36)	.08
Confident in caring for infant (n=10)	4.20	4.80, mean difference: 0.30 (95% CI 0.05 to 1.40)	.02
Generalized Anxiety Disorder-7 (n=12)	4.75	3.33, mean difference: 0.90 (95% CI −3.40 to 0.56)	.07
Parenting Sense of Competence Scale (n=12)	68.5	77.2, mean difference: 3.62 (95% CI −6.12 to 9.79)	.31
Understood material discussed	—[a]	4.8	
Interested in material discussed	—	4.8	
Satisfied with watching videos	—	5	
Satisfied with length of videos	—	4.9	
Videos were well made	—	5	

[a]Not applicable.

Discussion

We performed an open-label pilot study evaluating the feasibility and acceptability of videos to increase knowledge and confidence, decrease anxiety, and increase parenting competency of Spanish-speaking fathers of NICU infants. Enrollment feasibility was good (21/27, 78%), but completion of study material could be improved upon (13/21, 62%). Based on high Likert scales on video feedback, the project was accessible and well-accepted. Confidence increased significantly following the intervention, but not anxiety or parenting competence. Reasons for suboptimal study recruitment/retention could include fathers preferring alternative means of communication or poor digital literacy. Ways to improve completion rates of study material include greater incentivization and repeated follow-up (phone calls, emails) [18].

This study has several strengths. The project targets a vulnerable population that is undersupported [2]. We sought an intervention that was portable, flexible, accessible, and did not require NICU presence as fathers visit less than mothers [19,20]. The intervention was low-cost (a couple thousand dollars) and is reproducible for other units, languages, and topics. Fathers gave positive feedback on the videos and suggested other topics as well. This affirmative feedback to improve knowledge is consistent with other data on videos administered in a hospital-based setting that have had significant positive effects on outcomes including knowledge and emotional and behavioral outcomes [21].

The study has several limitations. First, the study sample size was small; significance may have been overestimated and the study was underpowered to detect changes in GAD-7 and PSOC scores or associations between outcomes while controlling for generalized anxiety and clinical characteristics. Second, the study lacked a control group and depended on precomparisons and postcomparisons, making it difficult to establish causality. Second, Latinos are a heterogeneous group and cultural differences may exist that impact study engagement or benefit. Of great importance, multiple factors affect the knowledge, confidence, anxiety, and competence of a father in the NICU; perhaps a single intervention will not influence these factors and a multifactorial approach is necessary. Additionally, the utilized Likert scales were not validated Spanish tools and recall bias may have influenced outcomes. Given this was a feasibility study to elucidate procedures for larger trials, future work will include a control group to allow for external validity. Lastly, fathers self-reported the timing of video watching and number of videos watched; future interventions can confirm videos were watched entirely in a consistent manner.

This preliminary work lends itself to future steps, including improving retention and follow-up through the exploration of barriers to enrollment and retention, as well as a larger, randomized controlled trial powered to assess longer-term changes in outcomes. Due to staffing limitations, we were unable to recruit at all times. Tailoring recruitment during visitation, particularly given that fathers visit less frequently [20], could improve recruitment feasibility and will be included in future work. Finally, the perspective of fathers on their preferred mode of receiving information and support is unknown; we are surveying fathers on their preferred means of communication and content to better target their needs.

XSL•FO
RenderX

Acknowledgments

Thank you to Kate Flock, Medical Videographer at the Massachusetts General Hospital Photo Lab, for her time, dedication, and excitement for the project. Study data were collected and managed using REDCap electronic data capture tools hosted at Mass General Brigham.

Funding

All phases of this study were supported by an American Academy of Pediatrics Section on Early Career Physicians Health Equity Grant.

Data Availability

Data are available upon request from the corresponding author.

Conflicts of Interest

None declared.

References

1. Stefana A, Padovani EM, Biban P, Lavelli M. Fathers' experiences with their preterm babies admitted to neonatal intensive care unit: a multi-method study. J Adv Nurs 2018 May;74(5):1090-1098. [doi: 10.1111/jan.13527] [Medline: 29350787]
2. Sigurdson K, Morton C, Mitchell B, Profit J. Disparities in NICU quality of care: a qualitative study of family and clinician accounts. J Perinatol 2018 May;38(5):600-607. [doi: 10.1038/s41372-018-0057-3] [Medline: 29622778]
3. Cyr-Alves H, Macken L, Hyrkas K. Stress and symptoms of depression in fathers of infants admitted to the NICU. J Obstet Gynecol Neonatal Nurs 2018 Mar;47(2):146-157. [doi: 10.1016/j.jogn.2017.12.006] [Medline: 29428520]
4. Gutierrez-Galve L, Stein A, Hanington L, Heron J, Ramchandani P. Paternal depression in the postnatal period and child development: mediators and moderators. Pediatrics 2015 Feb;135(2):e339-e347. [doi: 10.1542/peds.2014-2411] [Medline: 25560437]
5. Arockiasamy V, Holsti L, Albersheim S. Fathers' experiences in the neonatal intensive care unit: a search for control. Pediatrics 2008 Feb;121(2):e215-e222. [doi: 10.1542/peds.2007-1005] [Medline: 18182470]
6. Obregon E, Martin CR, Frantz Iii ID, Patel P, Smith VC. Neonatal Intensive Care Unit discharge preparedness among families with limited English proficiency. J Perinatol 2019 Jan;39(1):135-142. [doi: 10.1038/s41372-018-0255-z] [Medline: 30341402]
7. Ocampo MJ, Tinero JA, Rojas-Ashe EE. Psychosocial interventions and support programs for fathers of NICU infants - a comprehensive review. Early Hum Dev 2021 Mar;154:105280. [doi: 10.1016/j.earlhumdev.2020.105280] [Medline: 33221029]
8. Chen YL, Lee TY, Gau ML, Lin KC. The effectiveness of an intervention program for fathers of hospitalized preterm infants on paternal support and attachment 1 month after discharge. J Perinat Neonatal Nurs 2019;33(2):160-169. [doi: 10.1097/JPN.0000000000000406] [Medline: 31021942]
9. LeDuff LD 3rd, Carter BM, Cunningham CA, Braun LA, Gallaher KJ. NICU fathers: improving the quality of paternal support in the NICU. Adv Neonatal Care 2021 Oct 1;21(5):387-398. [doi: 10.1097/ANC.0000000000000796] [Medline: 33009159]
10. Akbari N, Moradi Z, Sabzi Z, Mehravar F, Fouladinejad M, Asadi L. The effect of narrative writing on fathers' stress in neonatal intensive care settings. J Matern Fetal Neonatal Med 2021 Feb;34(3):403-408. [doi: 10.1080/14767058.2019.1609926] [Medline: 31204537]
11. Dongre S, Desai S, Nanavati R. Kangaroo father care to reduce paternal stress levels: a prospective observational before-after study. J Neonatal Perinatal Med 2020;13(3):403-411. [doi: 10.3233/NPM-180190] [Medline: 32538877]
12. Filippa M, Saliba S, Esseily R, Gratier M, Grandjean D, Kuhn P. Systematic review shows the benefits of involving the fathers of preterm infants in early interventions in neonatal intensive care units. Acta Paediatr 2021 Sep;110(9):2509-2520. [doi: 10.1111/apa.15961] [Medline: 34053115]
13. Hearn G, Clarkson G, Day M. The role of the NICU in Father involvement, beliefs, and confidence: a follow-up qualitative study. Adv Neonatal Care 2020 Feb;20(1):80-89. [doi: 10.1097/ANC.0000000000000665] [Medline: 31567277]
14. Bronfenbrenner U. Ecological models of human development. In: International Encyclopedia of Education, 2nd edition: Oxford: Elsevier Sciences, Ltd; 1994:1643-1647 URL: https://impactofspecialneeds.weebly.com/uploads/3/4/1/9/3419723/ecologial_models_of_human_development.pdf [accessed 2025-11-25]
15. Breastfeeding fast facts. Centers for Disease Control and Prevention. 2025 May 20. URL: https://www.cdc.gov/infant-toddler-nutrition/breastfeeding/index.html [accessed 2025-09-28]
16. Teresi JA, Yu X, Stewart AL, Hays RD. Guidelines for designing and evaluating feasibility pilot studies. Med Care 2022 Jan 1;60(1):95-103. [doi: 10.1097/MLR.0000000000001664] [Medline: 34812790]

XSL•FO
RenderX

17. Gooding K, Phiri M, Peterson I, Parker M, Desmond N. Six dimensions of research trial acceptability: how much, what, when, in what circumstances, to whom and why? Soc Sci Med 2018 Sep;213:190-198. [doi: 10.1016/j.socscimed.2018.07.040] [Medline: 30142500]
18. Booker QS, Austin JD, Balasubramanian BA. Survey strategies to increase participant response rates in primary care research studies. Fam Pract 2021 Sep 25;38(5):699-702. [doi: 10.1093/fampra/cmab070] [Medline: 34213547]
19. Garten L, Maass E, Schmalisch G, Bührer C. O father, where art thou? Parental NICU visiting patterns during the first 28 days of life of very low-birth-weight infants. J Perinat Neonatal Nurs 2011;25(4):342-348. [doi: 10.1097/JPN.0b013e318233b8c3] [Medline: 22071618]
20. Harris LM, Shabanova V, Martinez-Brockman JL, et al. Parent and grandparent neonatal intensive care unit visitation for preterm infants. J Perinatol 2024 Mar;44(3):419-427. [doi: 10.1038/s41372-023-01745-x] [Medline: 37573462]
21. Dahodwala M, Geransar R, Babion J, de Grood J, Sargious P. The impact of the use of video-based educational interventions on patient outcomes in hospital settings: a scoping review. Patient Educ Couns 2018 Dec;101(12):2116-2124. [doi: 10.1016/j.pec.2018.06.018] [Medline: 30087021]

Abbreviations

GAD-7: Generalized Anxiety Disorder-7
NICU: neonatal intensive care unit
PSOC: Parenting Sense of Competence Scale
REDCap: Research Electronic Data Capture
SCN: special care nursery

Edited by J Sarvestan; submitted 11.02.25; peer-reviewed by DB McCarty; revised version received 14.10.25; accepted 15.10.25; published 09.12.25.

Please cite as:
Bernardo J, Belkin E, Hally S, Boyle S, Van E, Tascon-Arcila J, Herzberg E, Matute J, Lerou P
A Video-Based Intervention Supporting Spanish-Speaking Fathers of Infants in the Neonatal Intensive Care Unit: Open-Label Pilot Study
JMIR Form Res 2025;9:e72520
URL: https://formative.jmir.org/2025/1/e72520
doi:10.2196/72520

© Janine Bernardo, Elise Belkin, Suzanne Hally, Sara Boyle, Eric Van, Jose Tascon-Arcila, Emily Herzberg, Juan Matute, Paul Lerou. Originally published in JMIR Formative Research (https://formative.jmir.org), 9.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Changes in Skeletal Muscle Mass Index With Personalized Exercise and Meal Photo Analysis via Nutrition Applications: Single-Arm Pilot Study

Hidenori Onishi[1*], PhD; Yasutaka Mizukami[1*], PhD; Yuki Niida[2], PhD; Kazue Fujita[1]; Osamu Yamamura[1], MD, PhD

[1]Department of Community Medicine, Faculty of Medical Sciences, University of Fukui, 23-3 Matsuokashimoaizuki Eiheiji-cho, Yoshida-gun, Fukui, Japan

[2]Department of Health and Nutrition, Faculty of Human Life, Jin-ai University, Echizen-City, Fukui, Japan

[*]these authors contributed equally

Corresponding Author:
Hidenori Onishi, PhD
Department of Community Medicine, Faculty of Medical Sciences, University of Fukui, 23-3 Matsuokashimoaizuki Eiheiji-cho, Yoshida-gun, Fukui, Japan

Abstract

This study evaluated changes in the skeletal muscle mass index (SMI) in community residents who received a photos-based dietary guidance system, personalized exercise plans, and meal photography feedback to address sarcopenia. These findings indicated a possible upward trend in SMI within 3 months.

(JMIR Form Res 2025;9:e71807) doi:10.2196/71807

KEYWORDS

Sarcopenia; Skeletal Muscle Index; Exercise; Meal Photos; Dietary Guidance System

Introduction

Sarcopenia is characterized by the progressive loss of skeletal muscle mass and function and has become an important health challenge among older adults [1]. Nutrition and exercise have been shown to be effective in preventing and improving sarcopenia [2]. To address regional challenges, remote measurements using Internet of Things (IoT) technology are essential. Our research team is collaborating with private companies to develop an application that provides personalized exercise and nutrition management using tablet devices. This study aimed to evaluate the effectiveness of providing nutrition guidance and exercise programs through IoT technology to improve the skeletal muscle mass index (SMI).

Methods

Study settings

Since June 2022, health screenings have been conducted for community-dwelling older adults in Wakasa Town (Fukui, Japan), with 226 participants to date. This retrospective interventional study invited screening participants (June and December 2022, and December 2023) via email [3].

The exclusion criteria included lack of informed consent, hospitalization or residence in care facilities, investigators' determination that participation was inappropriate, or unsuitability for nutritional management.

The screening procedure included study explanation and obtaining informed consent, medical interview, physical function testing, and body composition analysis.

Assessments included body composition measured using a multifrequency analyzer (MC-780A-N; Tanita). Sarcopenia was classified according to the AWGS 2019 criteria [4] into seven categories. Participants received dietary guidance via an application and personalized exercise plans based on their functional categories. The intervention included tablet training, program provision, and home-based implementation. The application was run on Android 12 tablets.

Prepost comparisons used the Wilcoxon signed-rank test (version 1.54; EZR), with significance set at $P<.05$.

Ethical Considerations

This study was approved by the Fukui University Medical Research Ethics Review Committee (approval number: 20220046). All participants provided informed consent, and all data used for analysis was anonymized and used under the original consent.

Results

This study included 31 participants (14 men and 17 women; mean age: 72.8 y). After excluding 11 participants who met the exclusion criteria, 20 participants were included in this study. The pilot study was conducted from March to June 2024 [3].

XSL•FO
RenderX

Seven exercise programs were implemented at a frequency of 2‐3 times per week. The low-impact resistance training program included video guides [3] (Figure 1A and B). Dietary guidance was provided by having the participants upload photos of their meals and receive remote feedback from registered dietitians. Dietary guidance was defined as "utilized" if the application was used three or more times per week. Additionally, group assignment criteria were based on continuous application use for 7 weeks or more, and participants were categorized into three groups: (1) dropout group with four participants, (2) diet-only group with eight participants, and 3) diet-plus-exercise group with eight participants. No significant differences were observed between the dropout and control groups before or after the intervention. In the diet-only group, significant differences were observed in SMI (P=.03) in men and total body muscle mass (P=.03) and limb skeletal muscle mass (P=.03) in women. In the diet-plus-exercise group, significant differences were observed in total muscle mass (P=.009) in women, and in limb skeletal muscle mass (men: P=0.01; women: P=.009) and SMI (men: P=.004; women: P=.02) (Table 1).

Figure 1. Overview of the developed exercise and nutrition application (A) Automatic analysis of meals using AI and comments from a registered dietitian (B) Proposing personalized videos based on physical ability [3].

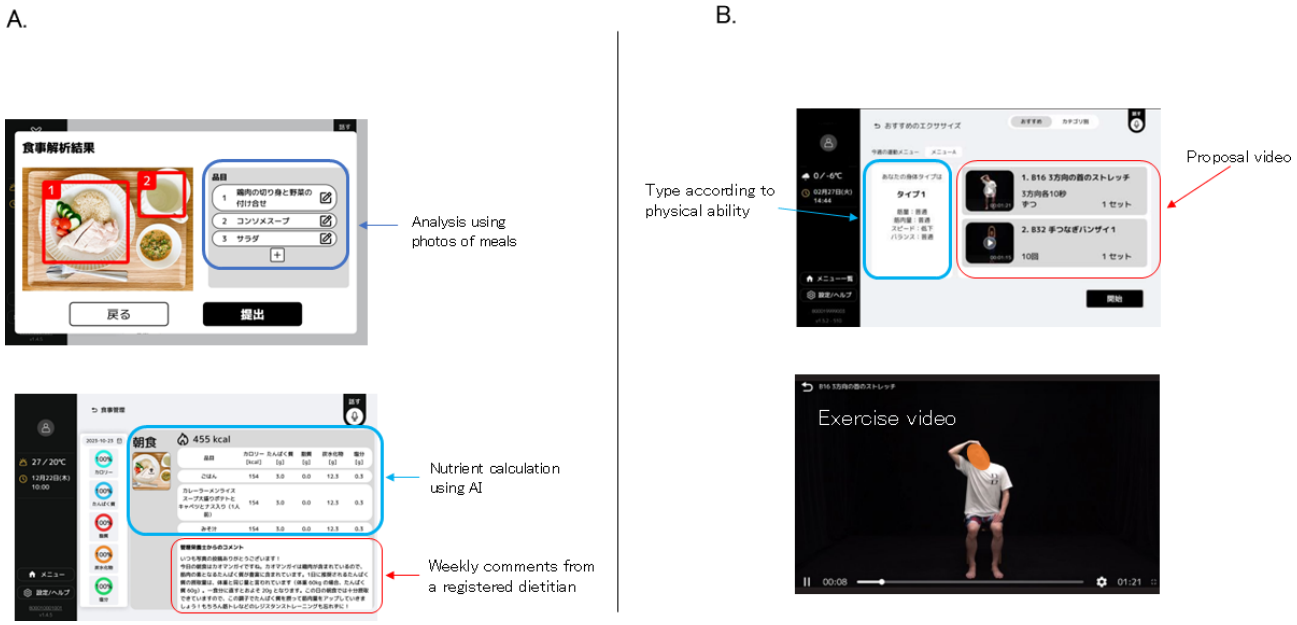

XSL•FO
RenderX

Table . Comparison of outcomes before and after using the developed exercise and nutrition application.

Study outcome	Participants (N=20)									Men (n=9)									Women (n=11)								
	Discontinuation group[a] (n=4)			Diet-only group[b] (n=8)			Diet-plus-exercise group[c] (n=8)			Discontinuation group (n=2)			Diet-only group (n=4)			Diet-plus-exercise group (n=3)			Discontinuation group (n=2)			Diet-only group (n=4)			Diet-plus-exercise group (n=5)		
	Before use	After use	*P* value[d]	Before use	After use	*P* value[d]	Before use	After use	*P* value[d]	Before use	After use	*P* value[d]	Before use	After use	*P* value[d]	Before use	After use	*P* value[d]	Before use	After use	*P* value[d]	Before use	After use	*P* value[d]	Before use	After use	*P* value[d]
Age (years), mean (SD)	77.0 (1.8)	77.5 (2.1)	.35	70.4 (4.2)	70.9 (4.6)	.02	75.4 (8.9)	76.0 (8.8)	.001	77.5 (2.1)	78.5 (2.1)	.35	71.3 (4.3)	72.0 (4.7)	.04	78.0 (8.5)	78.7 (8.5)	.01	76.5 (2.1)	76.5 (2.1)	≥.99	70.3 (4.6)	70.5 (5.0)	≥.99	73.8 (9.7)	74.4 (9.5)	.07
Height, cm	156.0 (8.5)	153.3 (8.3)	.42	167.7 (8.5)	164.4 (8.7)	.04	158.0 (10.1)	158.1 (10.2)	.28	167.7 (2.8)	167.7 (2.8)	≥.99	166.6 (5.9)	166.5 (5.9)	.27	169.9 (10.1)	167.3 (10.1)	≥.99	145.3 (6.7)	145.9 (7.2)	.50	154.9 (5.8)	154.4 (5.8)	.69	156.6 (7.4)	156.6 (7.4)	.72
Body weight, kg	54.8 (9.9)	53.7 (10.9	.13	61.4 (8.7)	61.5 (9.0)	.95	55.2 (12.7)	55.4 (12.8)	.48	63.8 (7.6)	65.3 (9.5)	≥.99	68.9 (4.8)	69.0 (5.0)	≥.99	66.8 (15.9)	66.9 (16.4)	≥.99	45.8 (3.4)	44.0 (3.4)	.50	54.0 (5.5)	54.1 (6.7)	.44	48.1 (5.8)	48.4 (5.8)	.97
BMI, kg/m^2	23.2 (1.8)	22.6 (2.4)	.25	23.3 (1.6)	23.5 (1.7)	.78	21.8 (2.7)	21.9 (2.6)	.93	24.7 (2.1)	24.6 (2.8)	≥.99	24.3 (2.0)	24.3 (2.0)	.92	23.7 (3.0)	23.6 (3.3)	.73	21.7 (0.4)	20.7 (0.5)	.50	22.5 (0.8)	22.6 (1.3)	.31	20.7 (2.3)	20.8 (2.2)	.76
Muscle mass-whole body, kg	37.2 (7.5)	37.8 (7.5)	.13	41.9 (7.5)	42.6 (7.8)	.03	38.0 (8.8)	38.2 (8.8)	.11	44.3 (2.9)	44.9 (3.6)	.50	48.9 (2.9)	50.0 (2.6)	.09	46.9 (9.9)	47.2 (10.0)	.08	30.2 (3.9)	30.7 (3.5)	.50	34.9 (3.7)	35.3 (3.6)	.03	32.6 (2.9)	32.9 (2.9)	.009
ASM,[e] kg	15.9 (3.4)	16.4 (3.3)	.13	19.3 (3.8)	19.9 (4.1)	.02	17.1 (4.7)	17.5 (5.0)	.04	18.9 (1.7)	19.4 (2.2)	.50	22.5 (2.1)	23.5 (1.9)	.06	21.3 (6.4)	21.9 (6.6)	.01	12.9 (2.7)	13.5 (2.1)	.50	16.0 (2.6)	16.3 (2.7)	.03	14.7 (1.9)	14.8 (1.9)	.009
SMI,[f] kg/m^2	6.7 (0.7)	6.9 (0.7)	.18	7.3 (0.8)	7.5 (0.8)	.01	6.7 (1.0)	6.9 (1.0)	.02	7.3 (0.4)	7.5 (0.6)	.50	7.9 (0.3)	8.3 (0.3)	.03	7.5 (1.5)	7.7 (1.5)	.004	6.1 (0.7)	6.3 (0.4)	≥.99	6.6 (0.6)	6.8 (0.6)	.06	6.3 (0.3)	6.3 (0.3)	.009
Five-times sit-to-stand test, s	7.3 (1.8)	8.1 (3.5)	.88	7.5 (1.3)	6.6 (1.5)	.11	6.7 (0.8)	7.0 (1.3)	.95	7.3 (0.6)	7.6 (0.4)	≥.99	7.9 (1.7)	6.5 (1.8)	.16	7.2 (1.0)	7.7 (1.1)	.43	7.4 (3.5)	8.7 (6.8)	≥.99	7.0 (0.9)	6.6 (1.6)	≥.99	6.5 (0.7)	6.6 (1.5)	.83
Standing on one leg with eyes open, s	31.4 (28.6)	28.8 (18.8)	.63	42.9 (18.8)	37.5 (20.5)	.30	39.4 (22.2)	42.2 (23.2)	.58	32.0 (39.6)	27.3 (27.6)	≥.99	38.9 (24.5)	26.2 (23.2)	.09	46.3 (28.8)	38.0 (12.6)	.11	30.8 (41.3)	30.3 (28.8)	≥.99	46.9 (17.1)	48.8 (15.4)	.86	35.4 (25.4)	39.8 (27.8)	.35
Maximum grip strength, kg	27.1 (7.9)	28.3 (8.0)	.63	36.5 (9.9)	35.4 (9.1)	.20	31.6 (8.2)	32.6 (7.4)	.73	31.7 (11.3)	33.8 (11.5)	.50	45.6 (5.6)	43.2 (7.1)	.56	38.0 (12.6)	37.1 (11.4)	.36	22.5 (6.1)	22.9 (1.6)	≥.99	27.5 (3.3)	27.5 (1.9)	≥.99	27.8 (2.3)	29.9 (4.6)	.70

[a]discontinuation group (hypertension, 2 persons; dyslipidemia, 1 person)

[b]diet-only group (diabetes, 1 person; hypertension, 4 persons; dyslipidemia, 1 person).

[c]diet-plus-exercise group (diabetes, 1 person; hypertension, 4 persons; dyslipidemia, 3 persons; heart disease, 1 person).

[d]Before-and-after comparison: Wilcoxon signed-rank test.

XSL•FO
RenderX

[e]ASM: appendicular skeletal mass.

[f]SMI: skeletal muscle mass index.

Discussion

In this study, a substantial increase in the SMI was observed in both the diet-only and diet-plus-exercise groups. Previous studies have demonstrated the effectiveness of personalized nutrition plans [5] and interventions that combine exercise and nutrition [6,7]. These studies primarily used nutritional supplements for their experiments. In contrast, this study confirmed the effectiveness of providing regular dietary guidance. These results suggest that interventions based on appropriately customized indirect applications can increase SMI. However, no significant increase in total muscle mass was observed in men in the diet plus exercise group, and no improvement in physical function was observed in either sex. Possible factors include sex differences in trunk and limb muscle strength [8], the small number of participants, and the fact that the exercise program of the application focused on the limbs. Furthermore, low- to moderate-intensity endurance training for older adults typically shows physiological responses after 4 months [9] and is recommended to be performed two to three times per week [10], suggesting that the trial period of the application was too short. Future challenges include expanding the sample size, investigating the intervention period, frequency, and its effects on muscle quality. Although this was a small-scale study without a control group, the results suggest that the program is useful for preventing or improving sarcopenia in older adults.

Acknowledgments

This study was conducted in collaboration with Naoki Okuyama, Eriko Masui, Atsushi Tsuchiya, Kenichi Hashiguchi, and Keiji Kajitani of Aflac Life Insurance Co., Ltd. We are grateful to Mr. Yasushi Matsumura of Xenera Co. Ltd. for his technical support.

Data Availability

The datasets generated or analyzed during this study are not publicly available due, as permission for data sharing with third parties outside the research team has not been obtained, but are available from the corresponding author on reasonable request.

Conflicts of Interest

HO entered confidentiality agreements with Aflac Life Insurance Co., Ltd. and Xenera Co., Ltd.
Other authors do not have conflicts to declare.

References

1. Papadopoulou SK. Sarcopenia: a contemporary health problem among older adult populations. Nutrients 2020 May 1;12(5):1293. [doi: 10.3390/nu12051293] [Medline: 32370051]
2. Anton SD, Hida A, Mankowski R, et al. Nutrition and exercise in sarcopenia. Curr Protein Pept Sci 2018;19(7):649-667. [doi: 10.2174/1389203717666161227144349] [Medline: 28029078]
3. Onishi H, Mizukami Y, Fujita K, Yamamura O. Developing individualized exercise programs using exercise videos. Jpn Soc Phys Ther Prev 2024;4(2):43-52. [doi: 10.57304/jptp.JPTP-D-24-00020]
4. Chen LK, Woo J, Assantachai P, et al. Asian Working Group for Sarcopenia: 2019 consensus update on sarcopenia diagnosis and treatment. J Am Med Dir Assoc 2020 Mar;21(3):300-307. [doi: 10.1016/j.jamda.2019.12.012] [Medline: 32033882]
5. Calvani R, Picca A, Coelho-Júnior HJ, Tosato M, Marzetti E, Landi F. Diet for the prevention and management of sarcopenia. Metab Clin Exp 2023 Sep;146:155637. [doi: 10.1016/j.metabol.2023.155637] [Medline: 37352971]
6. Yamada M, Nishiguchi S, Fukutani N, Aoyama T, Arai H. Mail-based intervention for sarcopenia prevention increased anabolic hormone and skeletal muscle mass in community-dwelling Japanese older adults: The INE (Intervention by Nutrition and Exercise) study. J Am Med Dir Assoc 2015 Aug 1;16(8):654-660. [doi: 10.1016/j.jamda.2015.02.017] [Medline: 25858281]
7. Lu Y, Niti M, Yap KB, et al. Assessment of sarcopenia among community-dwelling at-risk frail adults aged 65 years and older who received multidomain lifestyle interventions: a secondary analysis of a randomized clinical trial. JAMA Netw Open 2019 Oct 2;2(10):e1913346. [doi: 10.1001/jamanetworkopen.2019.13346] [Medline: 31617926]
8. Iwase H, Murata S, Kubo A, et al. Gender differences in back muscle strength and physical functions among community-dwelling elderly. Japanese Journal of Health Promotion and Physical Therapy 2013;3(3):97-101. [doi: 10.9759/hppt.3.97]
9. Posner JD, Gorman KM, Windsor-Landsberg L, et al. Low to moderate intensity endurance training in healthy older adults: physiological responses after four months. J Am Geriatr Soc 1992 Jan;40(1):1-7. [doi: 10.1111/j.1532-5415.1992.tb01820.x] [Medline: 1727835]

XSL•FO
RenderX

10. Rodrigues F, Domingos C, Monteiro D, Morouço P. A review on aging, sarcopenia, falls, and resistance training in community-dwelling older adults. Int J Environ Res Public Health 2022 Jan 13;19(2):874. [doi: 10.3390/ijerph19020874] [Medline: 35055695]

Abbreviations

IoT: Internet of Things
SMI: Skeletal Muscle Mass Index

Edited by A Mavragani; submitted 27.01.25; peer-reviewed by C Ochoa-Zezzatti, T Shimamoto; revised version received 12.08.25; accepted 12.08.25; published 25.09.25.

Please cite as:
Onishi H, Mizukami Y, Niida Y, Fujita K, Yamamura O
Changes in Skeletal Muscle Mass Index With Personalized Exercise and Meal Photo Analysis via Nutrition Applications: Single-Arm Pilot Study
JMIR Form Res 2025;9:e71807
URL: https://formative.jmir.org/2025/1/e71807
doi:10.2196/71807

© Hidenori Onishi, Yasutaka Mizukami, Yuki Niida, Kazue Fujita, Osamu Yamamura. Originally published in JMIR Formative Research (https://formative.jmir.org), 25.9.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Readability of Online Patient Educational Materials for Rosacea: Systematic Web Search and Analysis

Derek Nguyen[1], BA; Jennifer Javaheri[2], BS; Daniel Nguyen[2], BS; Vy Han[2], MD

[1]University of California Riverside, 900 University Avenue, Riverside, CA, United States
[2]California University of Science and Medicine, Colton, CA, United States

Corresponding Author:
Derek Nguyen, BA
University of California Riverside, 900 University Avenue, Riverside, CA, United States

Abstract

Most online educational materials about rosacea exceed recommended readability levels, often requiring at least a high school education to understand, with content authored by physicians being significantly more difficult to read than that written by nonphysicians.

(JMIR Form Res 2025;9:e67916) doi:10.2196/67916

KEYWORDS

rosacea; websites; Google; readability; patient information; health education; dermatology; health literacy

Introduction

Rosacea is a common skin condition characterized by persistent facial flushing and redness, which can progress to visible blood vessels and pus-filled bumps. Patients with rosacea often struggle to manage their symptoms and the psychosocial burdens associated with the condition [1]. As a result, many seek information beyond their health care providers to understand the disease better [1]. With increased internet accessibility, dermatological patients often consult online health resources when making decisions on treating illnesses [2]. Since the average reading ability of a US adult is at the eighth-grade level, the American Medical Association recommends patient educational material be written at a sixth-grade reading level [2]. Adhering to these readability standards is essential for improving patient comprehension and ensuring that dermatological resources are accessible to a wider audience. In this study, we assess the readability of available online educational material dedicated to the management and treatment of rosacea.

Methods

A web search was conducted on April 25, 2025, using Google with the query “Rosacea Patient Information.” The top 50 results were selected for the study as past research suggests that internet users generally do not search beyond this point [3]. The contents of these websites were then evaluated by 2 reviewers for relevance to patient education, with non-English sites and advertisements excluded. In total, 43 websites passed the inclusion criteria. The readability of their contents was assessed using an online scoring tool [4]. Seven established reading tools (Flesch Reading Ease, Flesch-Kincaid Grade Level, Gunning Fog Index, Coleman-Liau Readability Index, SMOG [Simple Measure of Gobbledygook] index, Automated Readability Index [ARI], and Linsear Write Readability Formula) were used in this process [5]. These tests assess various factors, including word count, character count, syllable count, and complexity, to generate a composite score that corresponds to a specific reading level. Finally, we compared group differences in readability between doctor of medicine (MD) and non-MD authors. Bachelor of medicine, bachelor of surgery, and other equivalent medical degrees were included in the MD category.

Readability scores were analyzed using 2-tailed *t* tests, with a *P* value <.05 considered statistically significant.

Results

Despite guidelines recommending a sixth-grade reading level for patient materials, few websites on rosacea met this standard across commonly used readability formulas: 3 for ARI, 5 for Flesch-Kincaid Grade Level, 4 for SMOG, 0 for the Coleman-Liau Index, 0 for the Gunning Fog Index, 4 for the Linsear Write Readability Formula, and 0 for Flesch Reading Ease. Furthermore, online educational material on rosacea exceeded this guideline by an average of 4.2 grade levels (Table 1). The differences between articles written by MD authors and non-MD authors were further analyzed, with articles on rosacea written by MD authors (n=18) found to require higher literacy skills than articles written by non-MD authors (n=25) across all readability formulas (Table 2). Previous studies have found a similar effect, highlighting a trend in which materials authored by health care professionals tend to be more complex and less accessible to the general public [6,7]. Interestingly, we observed a diverse range of authorship across the articles, with contributions from various organizations, including medical

XSL•FO
RenderX

societies, academic centers, and online platforms, featuring both MD and non-MD authors.

Table . Readability of rosacea online materials according to website category.

Readability	All websites (n=43), mean score (SD; range)	Online resources (n=21), mean score (SD; range)	Academic (n=12), mean score (SD; range)	Medical societies (n=10), mean score (SD; range)
Automated Readability Index	9.7 (2.9; 4.8 - 17.1)	9.9 (3.2; 4.8 - 17.1)	8.9 (3.0; 5.3 - 16.3)	10.4 (2.0; 7.3 - 14.2)
Gunning Fog Index	11.1 (2.5; 6.1 - 16.3)	11.1 (2.9; 6.1 - 16.3)	10.4 (2.3; 6.5 - 14.8)	11.9 (1.7; 9.6 - 14.9)
Coleman-Liau Readability Index	11.6 (2.6; 7.2 - 17.8)	11.6 (2.8; 7.2 - 17.6)	11.3 (2.8; 7.6 - 17.8)	12.1 (2.1; 10.0 - 17.1)
Simple Measure of Gobbledygook index	8.9 (2.0; 5.2 - 13.9)	9.0 (2.3; 5.2 - 13.9)	8.2 (1.9; 5.4 - 12.2)	9.5 (1.1; 7.7 - 11.2)
Linsear Write Readability Formula	9.9 (3.5; 4.1 - 16.8)	10.1 (3.6; 4.1 - 16.8)	9.6 (4.4; 4.1 - 15.6)	9.6 (1.9; 7.4 - 12.1)
Flesch-Kincaid Grade Level	9.2 (2.7; 5.0 - 15.4)	9.4 (3.1; 5.0 - 15.4)	8.4 (2.6; 5.2 - 14.0)	9.8 (1.8; 7.3 - 13.3)
Flesch Reading Ease	52 (15.6; 14 - 75)	52 (17.5; 14 - 73)	55.2 (14.6; 25 - 75)	48.9 (13.1; 19 - 62)

Table . Differences in readability of rosacea content: doctor of medicine (MD) versus non-MD authors.

Readability	MD (n=18), mean score (SD; range)	Non-MD (n=25), mean score (SD; range)	*P* value
Automated Readability Index	11.3 (3.0; 5.3 - 17.1)	8.4 (2.1; 4.8 - 13.7)	.002
Gunning Fog Index	12.8 (2.3; 8.2 - 16.3)	9.7 (1.9; 5.0 - 12.8)	<.001
Coleman-Liau Readability Index	13.1 (2.8; 7.6 - 17.8)	10.5 (1.9; 7.2 - 13.4)	.002
Simple Measure of Gobbledygook index	10.1 (1.8; 6.3 - 13.9)	7.9 (1.6; 5.2 - 11.7)	.007
Linsear Write Readability Formula	11.0 (3.3; 4.3 - 16.8)	9.0 (3.5; 4.1 - 14.6)	.03
Flesch-Kincaid Grade Level	10.8 (2.7; 5.3 - 15.4)	7.9 (2.0; 5.0 - 12.8)	.005
Flesch Reading Ease	42.6 (16.1; 14 - 73)	59.4 (10.6; 41 - 75)	.008

Discussion

The results obtained from 7 well-known and validated readability scales suggest that readers need to have a minimum of a high school level understanding of the English language to understand the majority of available online material on rosacea. Patients encounter an additional barrier when seeking credible health information from physician authors, as many of these authors lack the knowledge to address health literacy concerns and are therefore unable to explain these topics to readers effectively [8]. Previous studies have shown that patients with lower health literacy are more likely to distrust information from physicians [9]. Consequently, as the internet becomes a primary source for health information, this gap in readability not only hinders health decision-making but also increases the risk of incorrect self-diagnosis and self-treatment, which can worsen health outcomes and contribute to unnecessary strain on health care resources [10]. The primary limitation of our study was that we did not assess the content quality or understandability of the websites. As a result, the findings are limited to the readability aspect alone and do not provide a comprehensive evaluation of the usefulness or accuracy of the information presented. Further studies can evaluate patient comprehension of online dermatology patient educational material. Effective communication of medical knowledge to the general public is essential to bridging the readability gap. Readability should be facilitated through interactions between physicians and readers online and better physician understanding of patients' health literacy needs.

Data Availability

The raw data from this study, in the form of the websites used and the scores generated for each of them across the 7 scales, are available from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

XSL•FO
RenderX

References

1. Alinia H, Moradi Tuchayi S, Farhangian ME, et al. Rosacea patients seeking advice: qualitative analysis of patients' posts on a rosacea support forum. J Dermatolog Treat 2016;27(2):99-102. [doi: 10.3109/09546634.2015.1133881] [Medline: 26815357]
2. Prabhu AV, Gupta R, Kim C, et al. Patient education materials in dermatology: addressing the health literacy needs of patients. JAMA Dermatol 2016 Aug 1;152(8):946. [doi: 10.1001/jamadermatol.2016.1135] [Medline: 27191054]
3. Skrzypczak T, Skrzypczak A, Szepietowski JC. Readability of patient electronic materials for atopic dermatitis in 23 languages: analysis and implications for dermatologists. Dermatol Ther (Heidelb) 2024 Mar;14(3):671-684. [doi: 10.1007/s13555-024-01115-1] [Medline: 38402338]
4. Readability Formulas. URL: https://readabilityformulas.com/ [accessed 2025-08-04]
5. Shedlosky-Shoemaker R, Sturm AC, Saleem M, Kelly KM. Tools for assessing readability and quality of health-related web sites. J Genet Couns 2009 Feb;18(1):49-59. [doi: 10.1007/s10897-008-9181-0] [Medline: 18937063]
6. Kirchner A, Kulkarni V, Rajkumar J, Usman A, Hassan S, Lee EY. Readability assessment of patient-facing online educational content for pyoderma gangrenosum. J Am Acad Dermatol 2022 May;86(5):1127-1128. [doi: 10.1016/j.jaad.2021.04.023] [Medline: 33882279]
7. Jia JL, Nguyen B, Sarin KY. Assessment of readability and content of patient-initiated Google search results for epidermolysis bullosa. Pediatr Dermatol 2019 Nov;36(6):1004-1006. [doi: 10.1111/pde.13975] [Medline: 31468562]
8. Sprogell A, Casola AR, Cunningham A. Health literacy in primary care: reflections and suggestions for physicians, researchers, and administrators. Am J Lifestyle Med 2022;16(3):408-411. [doi: 10.1177/15598276211041283] [Medline: 35706590]
9. Chen X, Hay JL, Waters EA, et al. Health literacy and use and trust in health information. J Health Commun 2018;23(8):724-734. [doi: 10.1080/10810730.2018.1511658] [Medline: 30160641]
10. Daraz L, Morrow AS, Ponce OJ, et al. Readability of online health information: a meta-narrative systematic review. Am J Med Qual 2018;33(5):487-492. [doi: 10.1177/1062860617751639] [Medline: 29345143]

Abbreviations

ARI: Automated Readability Index
MD: doctor of medicine
SMOG: Simple Measure of Gobbledygook

Edited by A Mavragani; submitted 24.10.24; peer-reviewed by E Okoyeocha, M Gasmi, Y Cui; revised version received 03.07.25; accepted 05.07.25; published 08.09.25.

Please cite as:
Nguyen D, Javaheri J, Nguyen D, Han V
Readability of Online Patient Educational Materials for Rosacea: Systematic Web Search and Analysis
JMIR Form Res 2025;9:e67916
URL: https://formative.jmir.org/2025/1/e67916
doi:10.2196/67916

© Derek Nguyen, Jennifer Javaheri, Daniel Nguyen, Vy Han. Originally published in JMIR Formative Research (https://formative.jmir.org), 8.9.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Coach-Supported Internet-Based Cognitive Therapy for Social Anxiety Disorder in Hong Kong: Development, Training, and Pilot Clinical Outcomes

Graham R Thew[1,2], DClinPsy, DPhil; Candice LYM Powell[3], PsyD; Lok Sum Chan[4], BSocSc; Ezmond Siu Long Cheung[4], BPsySc; Jeff Lau Tsz Chung[4], BSc; Chun Lap Li[4], BSSc; Melody Miriam So[4], BSc; Yik Kwan Hsu[4], MSSc; Man Ying Ng[4], MSSc; Sze Hang Siu[4], MSSc; Katja Beardsell[2], MSc; Amy PL Kwok[5], PhD; Mandy H Lissillour Chan[5], PsyD; David M Clark[2], DPhil; Patrick WL Leung[4], PhD

[1]Oxford Health National Health Service Foundation Trust, Oxford, United Kingdom

[2]Department of Experimental Psychology, University of Oxford, The Old Rectory, Paradise Square, Oxford, United Kingdom

[3]Mind Hong Kong, Hong Kong, China

[4]Department of Psychology, Chinese University of Hong Kong, Hong Kong, China

[5]Department of Clinical Psychology, Hong Kong East Cluster, Hospital Authority, Hong Kong, China

Corresponding Author:
Graham R Thew, DClinPsy, DPhil
Oxford Health National Health Service Foundation Trust, Oxford, United Kingdom

Abstract

Internet-based Cognitive Therapy for Social Anxiety Disorder (iCT-SAD), a treatment that is normally supported by therapists, has been modified for delivery by trained coaches without mental health practitioner qualifications and has shown promising efficacy for patients in Hong Kong.

(JMIR Form Res 2025;9:e69818) doi:10.2196/69818

KEYWORDS

social anxiety disorder; cognitive behavioral therapy; training; supervision; internet interventions

Introduction

Internet-based Cognitive Therapy for Social Anxiety Disorder (iCT-SAD), a therapist-guided treatment based on the social anxiety disorder (SAD) model of Clark and Wells [1], was designed to replicate the face-to-face treatment protocol and has shown good efficacy in the United Kingdom [2,3] and Hong Kong [4,5]. Clinical psychologists provided guidance in these studies. Other internet interventions have shown positive outcomes using guidance from trained coaches [6]. Coach support may be less resource-intensive and more scalable, but it is unknown whether coaches can effectively deliver iCT-SAD, which typically has close therapist support.

This letter describes the development of a modified version of iCT-SAD, optimised for delivery by coaches, along with coach training and supervision, and pilot patient outcomes.

Methods

Design

This single-cohort pilot study measured social anxiety symptoms weekly using the self-reported Liebowitz Social Anxiety Scale (LSAS) [7].

Intervention

iCT-SAD includes core modules on key SAD treatment principles and a range of optional modules for tailoring treatment to patients' specific concerns [2,3]. Support is provided via telephone, messaging, and SMS texts. Four calls are scheduled over the first 2 weeks, then weekly thereafter until week 14. Three monthly booster calls are then scheduled to monitor and consolidate progress. Behavioral experiments, constituting practical exercises to test fearful beliefs in anxiety-provoking situations [8], are used throughout treatment. The English-language program was translated into Chinese following a similar process for a Japanese translation [9].

We significantly modified 3 of 27 iCT-SAD modules for coach support by replacing or removing elements requiring expert therapist support. Typically, therapists support patients in completing exercises speaking to a stranger via webcam, demonstrating experientially the disadvantages of self-focused attention and safety behaviors [10], with subsequent video feedback [11]. These two modules were replaced with a demonstration video and guidance on implementing this experiment independently. The module addressing past socially traumatic memories was edited to focus solely on stimulus discrimination, removing imagery rescripting, which typically

XSL•FO
RenderX

requires closer therapist guidance. We added new videos (reducing safety behaviors, generating experiment ideas, behavioral experiment tips) and made minor text changes to reflect the coach role (removing mention of the therapist administering surveys, which was not within the coach role). The frequency and length of coach phone calls were identical to therapist calls in previous studies.

Coach Recruitment, Training, and Supervision

Five psychology graduates with no professional qualifications in mental health care (2 female, 3 male; mean age 25.0, SD 1.22 years) were recruited as coaches. Training consisted of a 1-day in-person workshop, followed by two 2-hour online training sessions given by GRT. These were adapted from previous iCT-SAD training [4] and covered key principles of SAD and cognitive therapy, the treatment procedure, the website, and patient communication methods. The training summarized each module, focusing on how coaches support patients within their role. It emphasized taking queries to their supervisor and fidelity to the treatment content and training.

Coaches then received weekly group supervision with CLYMP (a clinical psychologist) as they saw initial cases. Each saw 2 to 3 cases between September 2023 and July 2024.

Participants

Twelve participants received treatment (10 female, 2 male; mean age 26.3, SD 4.1 years). All were referred from clinical services and experienced SAD as their main concern, confirmed at assessment by a clinical psychologist. All had sufficient literacy, time, and technology to undertake treatment.

Ethical Considerations

The Joint Chinese University of Hong Kong-New Territories East Cluster Clinical Research Ethics Committee approved this study (2022.086-T). All participants provided informed consent, which included anonymized reporting of outcomes.

Results

Ten participants completed the 14-week treatment phase and 9 completed the monthly booster phase. One participant withdrew in week 2, reporting treatment did not suit them, 1 withdrew in week 13 due to work commitments, and 1 was uncontactable during the booster phase. Participants spent a mean of 22.2 (SD 14.4) hours on the site. They were granted access to a mean of 19.6 (SD 4.7) modules, and completed a mean of 18.2 (SD 5.8;93%), plus a mean of 12.6 (SD 12.3) behavioral experiments. Figure 1 shows weekly LSAS scores [7], indicating substantial reductions in SAD symptoms. The mean scores were 78.8 (SD 19.5) at baseline and 23.2 (SD 15.5) at week 14. A pre-post paired 2-tailed t test was significant (t_{11}=7.18; P<.001). The effect size estimate (Cohen d) was 2.86, which is considered large.

XSL•FO
RenderX

Figure 1. Liebowitz Social Anxiety Scale (LSAS): weekly (1-14) and monthly booster follow-up [FU] 1-3) scores for each participant in the pilot treatment cohort.

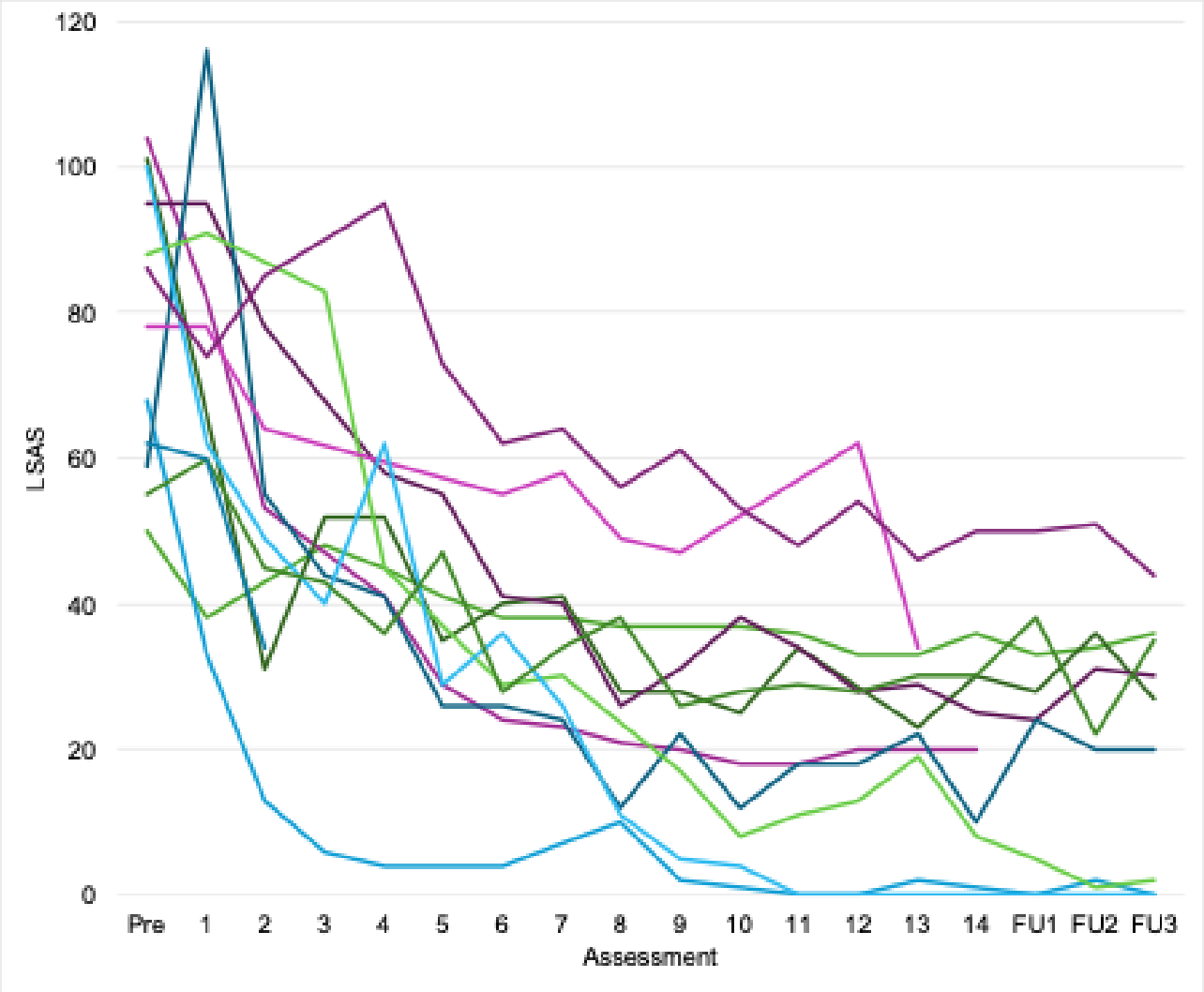

Discussion

The findings provide initial evidence that coach-supported iCT-SAD can achieve positive outcomes broadly consistent with Thew et al [5]. Further text modifications and updates to coach training (eg, to clarify task instructions and emphasize video feedback) have been made where indicated. Findings are limited by the lack of a control group and the small sample, but progression to a larger randomized controlled trial is indicated to evaluate it against waitlist and active controls.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

References

1. Clark DM, Wells A. A cognitive model of social phobia. In: Heimberg RG, Liebowitz M, Hope DA, Schneier FR, editors. Social Phobia: Diagnosis, Assessment, and Treatment: Guilford Press; 1995:69-93.
2. Clark DM, Wild J, Warnock-Parkes E, et al. More than doubling the clinical benefit of each hour of therapist time: a randomised controlled trial of internet cognitive therapy for social anxiety disorder. Psychol Med 2023 Aug;53(11):5022-5032. [doi: 10.1017/S0033291722002008] [Medline: 35835726]
3. Stott R, Wild J, Grey N, et al. Internet-delivered cognitive therapy for social anxiety disorder: a development pilot series. Behav Cogn Psychother 2013 Jul;41(4):383-397. [doi: 10.1017/S1352465813000404] [Medline: 23676553]

XSL•FO
RenderX

4. Thew GR, Powell CL, Kwok AP, et al. internet-based cognitive therapy for social anxiety disorder in Hong Kong: therapist training and dissemination case series. JMIR Form Res 2019 May 15;3(2):e13446. [doi: 10.2196/13446] [Medline: 31094320]
5. Thew GR, Kwok APL, Lissillour Chan MH, et al. Internet-delivered cognitive therapy for social anxiety disorder in Hong Kong: a randomized controlled trial. Internet Interv 2022 Apr;28:100539. [doi: 10.1016/j.invent.2022.100539] [Medline: 35493437]
6. Baumeister H, Reichler L, Munzinger M, Lin J. The impact of guidance on Internet-based mental health interventions — a systematic review. Internet Interv 2014 Oct;1(4):205-215. [doi: 10.1016/j.invent.2014.08.003]
7. Baker SL, Heinrichs N, Kim HJ, Hofmann SG. The Liebowitz Social Anxiety Scale as a self-report instrument: a preliminary psychometric analysis. Behav Res Ther 2002 Jun;40(6):701-715. [doi: 10.1016/s0005-7967(01)00060-2] [Medline: 12051488]
8. Warnock-Parkes E, Thew GR, Murray H, et al. Driving cognitive change: a guide to behavioural experiments in cognitive therapy for anxiety disorders and PTSD. Cogn Behav Ther 2025 Jun 26:1-27. [doi: 10.1080/16506073.2025.2518427] [Medline: 40569017]
9. Yoshinaga N, Thew GR, Kobori O, Hayashi Y, Clark DM. Lost in translation? Cultural adaptation of treatment content for Japanese internet-based cognitive therapy for social anxiety disorder. J Behav Cogn Ther 2021 Dec;31(4):363-368. [doi: 10.1016/j.jbct.2021.05.004]
10. McManus F, Clark DM, Grey N, et al. A demonstration of the efficacy of two of the components of cognitive therapy for social phobia. J Anxiety Disord 2009 May;23(4):496-503. [doi: 10.1016/j.janxdis.2008.10.010] [Medline: 19081225]
11. Wild J, Warnock-Parkes E, Stott R, et al. Video feedback to update negative self-perceptions in social anxiety disorder: a comparison of internet-delivered vs face-to-face cognitive therapy formats. J Affect Disord 2023 Jun 15;331:139-144. [doi: 10.1016/j.jad.2023.03.017] [Medline: 36907460]

Abbreviations

iCT-SAD: Internet-based Cognitive Therapy for Social Anxiety Disorder
LSAS: Liebowitz Social Anxiety Scale
SAD: social anxiety disorder

Edited by A Mavragani; submitted 09.12.24; peer-reviewed by A Debrot; revised version received 03.07.25; accepted 05.07.25; published 26.08.25.

Please cite as:
Thew GR, Powell CLYM, Chan LS, Cheung ESL, Chung JLT, Li CL, So MM, Hsu YK, Ng MY, Siu SH, Beardsell K, Kwok APL, Lissillour Chan MH, Clark DM, Leung PWL
Coach-Supported Internet-Based Cognitive Therapy for Social Anxiety Disorder in Hong Kong: Development, Training, and Pilot Clinical Outcomes
JMIR Form Res 2025;9:e69818
URL: https://formative.jmir.org/2025/1/e69818
doi:10.2196/69818

© Graham R Thew, Candice LYM Powell, Lok Sum Chan, Ezmond Siu Long Cheung, Jeff Lau Tsz Chung, Chun Lap Li, Melody Miriam So, Yik Kwan Hsu, Man Ying Ng, Sze Hang Siu, Katja Beardsell, Amy PL Kwok, Mandy H Lissillour Chan, David M Clark, Patrick WL Leung. Originally published in JMIR Formative Research (https://formative.jmir.org), 26.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Expressions of “Ikizurasa” in Posts on X (Formerly Twitter) in Japan in 2023: Descriptive Analysis

Kanami Tsuno[1], MPH, PhD; Azusa Arimoto[2], PhD; Yuki Imamatsu[3], PhD; Yumiko Kobayashi[2], MPH, PhD; Miho Satoh[4], PhD; Tomoyuki Miyazaki[3], PhD

[1]School of Health Innovation, Kanagawa University of Human Services, Yokosuka, Japan
[2]Department of Community Health Nursing, Graduate School of Medicine, Yokohama City University, 3-9 Fukura Kanazawa-ku, Yokohama, Japan
[3]Center for Promotion of Research and Industry-Academic Collaboration, Yokohama City University, Yokohama, Japan
[4]Department of Fundamental Nursing, Graduate School of Medicine, Yokohama City University, Yokohama, Japan

Corresponding Author:
Azusa Arimoto, PhD
Department of Community Health Nursing, Graduate School of Medicine, Yokohama City University, 3-9 Fukura Kanazawa-ku, Yokohama, Japan

Abstract

This study analyzed all public X (formerly Twitter) posts in Japan in 2023 and identified a substantial number mentioning *ikizurasa* (pain of living), with notable fluctuations over time. The findings suggest a link between *ikizurasa* and stress in minority groups, particularly minority ethnic and gender groups.

(JMIR Form Res 2025;9:e70613) doi:10.2196/70613

KEYWORDS

Japan; social media; social networking; suicide; young adult; youth

Introduction

In Japan, deaths from suicide declined during the early phase of the COVID-19 pandemic but later increased, especially among women and people younger than 40 years [1]. Depression and hopelessness are key predictors of suicidal ideation and behavior [2,3].

Recently, *ikizurasa* (“pain of living”) has drawn attention as a factor in adolescent mental health issues, including *hikikomori* (“being confined,” a term used to describe deep social isolation) and suicide [4]. *Ikizurasa* was defined by the Japanese Neuropsychiatric Society as “impairment and suppression of independent social relationship formation.” It has also been described as involving solitude, anxiety, low self-esteem, hopelessness, and anger [5]. However, quantitative research focusing on *ikizurasa* remains scarce.

Social media behavior has been linked to mental health, particularly among youth. A cross-sectional study in Japan found that posting phrases like “want to die” on X (formerly Twitter) was significantly associated with suicidal thoughts and behaviors [6]. A meta-analysis also reported that frequent social media and smartphone use correlates with suicidal tendencies [7].

This study aimed to examine how often *ikizurasa* was expressed and when these expressions peaked on X in Japan, which has the second-largest X user base after the United States. In 2023, X was used by 65.7% of teenagers, and 81.6% of people in their 20s—much higher rates than among older generations or on other platforms [8]. As X allows anonymous, text-based posting, it offers a space where users may express their emotional distress openly.

Methods

Overview

All X public account data from January 1, 2023, through December 31, 2023, were obtained using the official X application programming interface (version 2). The search query was 生きづら (*ikizura*) or 生きにくい (*ikinikui*; both mean “pain of living” or "hard to live"). The analysis was performed in Python.

Ethical Considerations

This study used only secondary data without identifiable individuals.

Results

In 2023, 1,399,746 posts mentioned *ikizurasa*; 397,517 were organic (ie, excluding reposts or quotes). Figure 1 shows the daily number of posts. The overall average was 3835, with a peak of 26,063 on April 29. Organic posts averaged 1089, peaking at 3414 on July 13, with another peak in late February.

XSL•FO
RenderX

Figure 1. Time series of X posts. The blue line shows the total number of posts (including original posts and reposts), and the green line shows the number of original posts.

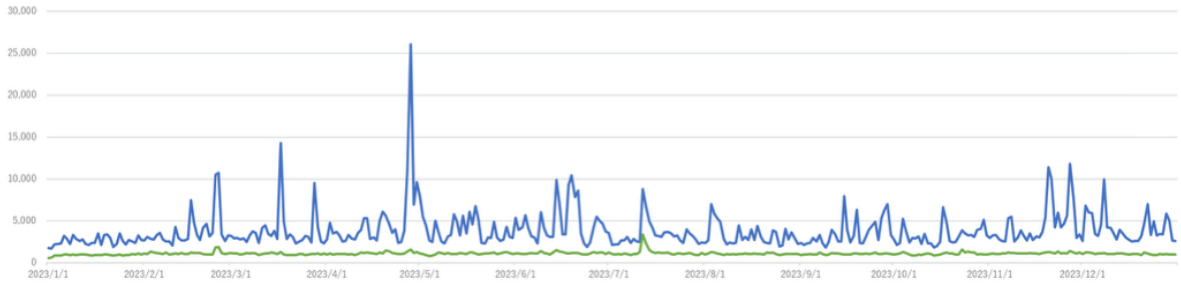

On April 28, an interview article titled "Difficulties in Life as a Minority" [9] was published, reporting on the experiences of discrimination of an ethnically Uyghur member of the House of Representatives, Eri Arfiya, and many X users mentioned the article. On July 12, the news of the death (by presumed suicide) of a well-known transgender celebrity, Ryuchell, was shared with the X community. Many mentioned the news by discussing *ikizurasa* as a transgender person and expressing their shock at Ryuchell's sudden death that day.

The late February peak was triggered by a post from a US celebrity in Japan, who noted that the lack of English on train tickets may cause foreigners to feel *ikizurasa*.

Table 1 shows the age distribution of the X users who posted using the term *ikizurasa*. Most users who posted *ikizurasa* were in their 20s and 30s.

Table . Age distribution of X users who posted using the terms *ikizura* or *ikinikui*.

Age (years)	Men, %	Women, %
≤19	1.19	9.66
20 - 29	28.57	39.20
30 - 39	33.33	32.95
40 - 49	22.62	15.34
≥50	10.71	2.84

Discussion

This study aimed to examine the frequency and timing of posts using the term *ikizurasa* on X in Japan, where the platform is particularly popular among younger generations. Our results showed that over 1.3 million posts containing the term *ikizurasa* were recorded in 2023, with notable peaks in late April and mid-July. These peaks appeared to coincide with news events involving ethnic and gender minorities, suggesting that expressions of *ikizurasa* may be linked to the stress of being part of a minority group.

Previous studies have documented strong associations between minority status—such as ethnicity or gender—and poor mental health outcomes [10,11]. Consistent with this literature, our findings indicate that individuals from minority backgrounds may be more likely to experience and publicly express feelings of *ikizurasa*, reflecting psychological distress tied to discrimination and marginalization.

This study has limitations. It focused only on posts using specific Japanese terms, possibly excluding other expressions of distress. As the analysis relied on public social media data, user authenticity, intent, and demographics could not be verified, limiting interpretation.

Even so, the findings suggest that social media can serve as a real-time indicator of psychological distress, especially among younger and marginalized groups. Integrating such monitoring into public health strategies may help detect emerging mental health issues and support timely, targeted responses.

Acknowledgments

This work was supported by the Japan Science and Technology Agency (JST; grant JPMJPF2203) and YCU Future Co-Creation Project. We also thank D4DR Inc and Missile Inc for supporting the scraping and analyzing of X posts for this study.

XSL•FO
RenderX

Data Availability

The datasets generated and analyzed during the current study were obtained via the public X API. In accordance with X's Developer Policy, raw tweet content cannot be redistributed. However, tweet IDs used in this study are available from the corresponding author upon reasonable request.

Authors' Contributions

AA, YI, and YK conceived and planned the study. KT took the lead in writing the manuscript. AA and TM supervised the entire study project. All authors provided critical feedback and helped shape the research, analysis, and manuscript.

Conflicts of Interest

None declared.

References

1. Ueda M, Nordström R, Matsubayashi T. Suicide and mental health during the COVID-19 pandemic in Japan. J Public Health (Oxf) 2022 Aug 25;44(3):541-548. [doi: 10.1093/pubmed/fdab113] [Medline: 33855451]
2. Ribeiro JD, Huang X, Fox KR, Franklin JC. Depression and hopelessness as risk factors for suicide ideation, attempts and death: meta-analysis of longitudinal studies. Br J Psychiatry 2018 May;212(5):279-286. [doi: 10.1192/bjp.2018.27] [Medline: 29587888]
3. Hawton K, Casañas i Comabella C, Haw C, Saunders K. Risk factors for suicide in individuals with depression: a systematic review. J Affect Disord 2013 May;147(1-3):17-28. [doi: 10.1016/j.jad.2013.01.004] [Medline: 23411024]
4. Nae N. Quo Vadis? Japanese youth and the burden of living. Euromentor J 2018;IX(2):15-25 [FREE Full text]
5. Kido R. Engaging the angst of unemployed youth in post‐industrial Japan: a narrative self‐help approach [thesis]. University of Adelaide. 2018. URL: https://digital.library.adelaide.edu.au/dspace/bitstream/2440/112673/2/Kido2018_PhD.pdf [accessed 2025-08-01]
6. Sueki H. The association of suicide-related Twitter use with suicidal behaviour: a cross-sectional study of young internet users in Japan. J Affect Disord 2015 Jan 1;170:155-160. [doi: 10.1016/j.jad.2014.08.047] [Medline: 25240843]
7. Macrynikola N, Auad E, Menjivar J, Miranda R. Does social media use confer suicide risk? A systematic review of the evidence. Comput Hum Behav Rep 2021 Jan;3:100094. [doi: 10.1016/j.chbr.2021.100094]
8. Survey on information and communication media usage time and information behavior in FY2023. Ministry of Internal Affairs and Communications. 2024. URL: https://www.soumu.go.jp/main_content/000953020.pdf [accessed 2025-08-01]
9. Kato T. 英利アルフィヤさん 選挙で体験した「マイノリティーの生きづらさ」語る ヘイト被害「民主主義を脅かす」. Tokyo Shimbun. 2023 Apr 27 URL: https://www.tokyo-np.co.jp/article/246442 [accessed 2025-08-12]
10. Pascoe EA, Smart Richman L. Perceived discrimination and health: a meta-analytic review. Psychol Bull 2009 Jul;135(4):531-554. [doi: 10.1037/a0016059] [Medline: 19586161]
11. Meyer IH. Prejudice, social stress, and mental health in lesbian, gay, and bisexual populations: conceptual issues and research evidence. Psychol Bull 2003 Sep;129(5):674-697. [doi: 10.1037/0033-2909.129.5.674] [Medline: 12956539]

Edited by A Mavragani; submitted 27.12.24; peer-reviewed by A Minoura, S Arce-Garcia; revised version received 10.07.25; accepted 11.07.25; published 22.08.25.

Please cite as:
Tsuno K, Arimoto A, Imamatsu Y, Kobayashi Y, Satoh M, Miyazaki T
Expressions of "Ikizurasa" in Posts on X (Formerly Twitter) in Japan in 2023: Descriptive Analysis
JMIR Form Res 2025;9:e70613
URL: https://formative.jmir.org/2025/1/e70613
doi:10.2196/70613

© Kanami Tsuno, Azusa Arimoto, Yuki Imamatsu, Yumiko Kobayashi, Miho Satoh, Tomoyuki Miyazaki. Originally published in JMIR Formative Research (https://formative.jmir.org), 22.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Examining the Relationship Between Assertiveness and Anxiety in First- and Second-Year US Medical Students

Jonathan Shaw[1*], BS; James Hagerty[1*], BS; Kristen Masada[1], BA; Angelene Eunji Won[1], BA; Ashley Lai[1], BS; Jisu Shin[1], BS; Van Le[1], BA; Brenton Phung[1], BA; Charles Lai[1], BS; Peter Bota[1], BA; Aaron Jacobs[2], PhD

[1]School of Medicine, California University of Science and Medicine, 1501 Violet St, Colton, United States

[2]Medical Education, California University of Science and Medicine, United States

[*]these authors contributed equally

Corresponding Author:
Jonathan Shaw, BS
School of Medicine, California University of Science and Medicine, 1501 Violet St, Colton, United States

Abstract

In this single-institution remote survey study of a California medical school, anxiety and assertiveness were found to inversely correlate in preclinical medical students.

(JMIR Form Res 2025;9:e73394) doi:10.2196/73394

KEYWORDS

assertiveness; anxiety; medical student; preclinical; medical education; well-being

Introduction

Confident and assertive physicians' development is essential to effective patient care [1]. Medical training shapes students' knowledge, skills, and interpersonal abilities, including confidence and assertiveness, which evolve throughout education [2]. Assertiveness influences clinical decision-making, patient communication, and professional success [3]. Further, assertiveness has been shown to improve feelings of anxiety, stress, and depression [4]. Although literature confirms that training can improve assertiveness, factors that affect assertiveness and its gender-based differences have been underexplored in medical students, who face uniquely stressful hierarchical environments [1]. A better understanding of assertiveness and anxiety correlations can help educators tailor interventions to support students' well-being and competence. To guide future curricular improvements, we investigate this relationship through the Simple Rathus Assertiveness Scale-Short Form (SRAS-SF) and the General Anxiety Disorder 7-item (GAD-7) assessment.

Methods

Participants and Recruitment

First-year and second-year medical students (n=120 and n=126, respectively) from a California allopathic school with a pass/fail curriculum completed an anonymous Google Forms survey from January 31 to February 29, 2024 (invited via institutional email). Submission of multiple responses was prevented via Google Forms' single-response function. For study inclusion, students had to be in their preclinical years of the school's Doctor of Medicine program; else, they were excluded. A convenience sample of 30 responses (23 and 7 responses from first-year and second-year medical students, respectively) was collected (Table 1).

XSL•FO
RenderX

Table . Descriptive statistics (responses: N=30).

Survey items	Value
Demographics, n (%)	
Which school year are you?	
First year	23 (77)
Second year	7 (23)
What gender do you identify as?	
Male	13 (43)
Female	17 (57)
SRAS-SF[a,b] score, median (IQR)	
When I am eating out and the food I am served is not cooked the way I like it, I complain to the person serving it	−2.00 (−2.25 to 1.00)
There are times when I look for a good strong argument	1.00 (−2.00 to 2.00)
I try as hard in life to get ahead as most people like me do	1.00 (0.50 to 2.00)
If a famous person were talking in a crowd and I thought he/she was wrong, I would get up and say what I thought	−2.00 (−3.00 to −1.00)
If someone has been telling false and bad stories about me, I see him or her as soon as possible to “have a talk” about it	1.00 (−2.00 to 2.00)
I complain about poor service when I am eating out or in other places	−2.00 (−2.00 to −1.00)
If a couple near me in the theater were talking rather loudly, I would ask them to be quite or to go somewhere else and talk	−1.00 (−2.25 to 1.00)
I am quick to say what I think	0 (−1.00 to 1.00)
Most people stand up for themselves more than I do	0 (−2.00 to 2.00)
At times I have not made or gone on dates because of my shyness	−1.00 (−2.25 to 1.00)
If a person serving in a store has gone to a lot of trouble to show me something which I do not really like, I have a hard time saying, “No.”	−1.00 (−2.00 to 1.25)
To be honest, people often get the better of me	−1.00 (−2.00 to 1.00)
I do not like making phone calls to businesses or companies	1.50 (0.25 to 2.00)
I feel silly if I return things I don’t like to the store that I bought them from	−1.50 (−2.00 to 1.00)
If a close relative that I like was upsetting me, I would hide my feelings rather than say that I was upset	−1.00 (−1.25 to 2.00)
I have sometimes not asked questions for the fear of sounding stupid	2.00 (−1.25 to 2.25)
During an argument, I am sometimes afraid that I will get so upset that I will shake all over	−2.00 (−3.00 to −0.50)
I often have a hard time saying, “No.”	1.00 (−2.00 to 2.00)
When someone says I have done well, I sometimes just don’t know what to say	1.00 (−2.00 to 1.25)
SRAS-SF total score	−0.11 (−0.76 to 0.84)
GAD-7[c,d] assessment score, median (IQR)	
Feeling nervous, anxious, or on edge	1.00 (0.75 to 2)
Not being able to stop or control worrying	0 (0 to 1.00)
Worrying too much about different things	1.00 (0 to 2.00)
Trouble relaxing	1.00 (0 to 2.00)
Being so restless that it’s hard to sit still	0 (0 to 1.00)
Becoming easily annoyed or irritable	1.00 (0 to 2.00)
Feeling afraid as if something awful might happen	0.50 (0 to 1.00)

XSL•FO
RenderX

Survey items	Value
GAD-7 total score	4.50 (2.00 to 11.25)

[a]SRAS-SF: Simple Rathus Assertiveness Scale-Short Form.

[b]SRAS-SF scoring: The SRAS-SF consists of 19 statements that participants indicate their agreement with, using a 6-point Likert-scale (−3="very much unlike me"; 3="very much like me"). Responses are averaged, resulting in total scores between −3 (less assertive) and 3 (more assertive).

[c]GAD-7: General Anxiety Disorder 7-item.

[d]GAD-7 scoring: The GAD-7 consists of 7 statements about anxiety symptoms, with participants indicating how often they experienced these symptoms within the last 2 weeks by using a 4-point Likert scale (0="not at all"; 3="nearly every day"). These scores are added together to determine anxiety severity: 0 - 4 (minimal), 5 - 9 (mild), 10 - 14 (moderate), and 15 - 21 (severe anxiety).

Measures

The survey included demographic questions (school year and gender), 19 SRAS-SF items [5], and the GAD-7. The SRAS-SF and GAD-7 were presented in separate, randomized sections.

Statistical Analysis

IBM SPSS Statistics 28.0.1.0 (IBM Corp) was used for analysis. Due to the small sample size (n=30), a Shapiro-Wilk test was used to assess if data were normally distributed [6]. Parametric (2-tailed independent samples *t* test and Pearson correlation) and nonparametric (Kruskal-Wallis test and Spearman correlation) statistical tests were used based on data distribution normality. Gender and school year were used as grouping variables.

Ethical Considerations

This study received ethical approval from the California University of Science and Medicine Institutional Review Board (approval: HS-2024 - 03) on January 22, 2024. Informed consent for primary data collection and secondary analyses of the data was obtained from all participants. Participants received no compensation for participation.

Results

The Shapiro-Wilk test indicated that SRAS-SF scores (P=.07) were normally distributed, while GAD-7 scores (P=.01) and all individual survey items (P<.05) were not normally distributed.

Per the Kruskal-Wallis test for examining differences between responses by school year and gender, first-year medical students were more likely to feel uncomfortable when returning purchases (P=.03), and female participants were more likely to ask loud theater couples to be quiet (P=.05). No differences in GAD-7 scores by school year (P=.67) or gender (P=.52) were noted.

As the overall SRAS-SF scores were normally distributed, an independent *t* test was used; it found no significant differences in SRAS-SF scores by school year (P=.95) or gender (P=.62).

A Pearson correlation revealed a strong negative correlation between SRAS-SF and GAD-7 scores (n=30, r=−0.624; P<.001; Figure 1).

XSL•FO
RenderX

Figure 1. Scatterplot of GAD-7 scores by SRAS-SF scores. This figure was generated by using SPSS, and it visualizes the line of best fit between the GAD-7 and SRAS-SF scores. The R^2 value is 0.39, and the r value is −0.62, indicating a strong negative correlation between GAD-7 and SRAS-SF scores. GAD-7: General Anxiety Disorder 7-item; SRAS-SF: Simple Rathus Assertiveness Scale-Short Form.

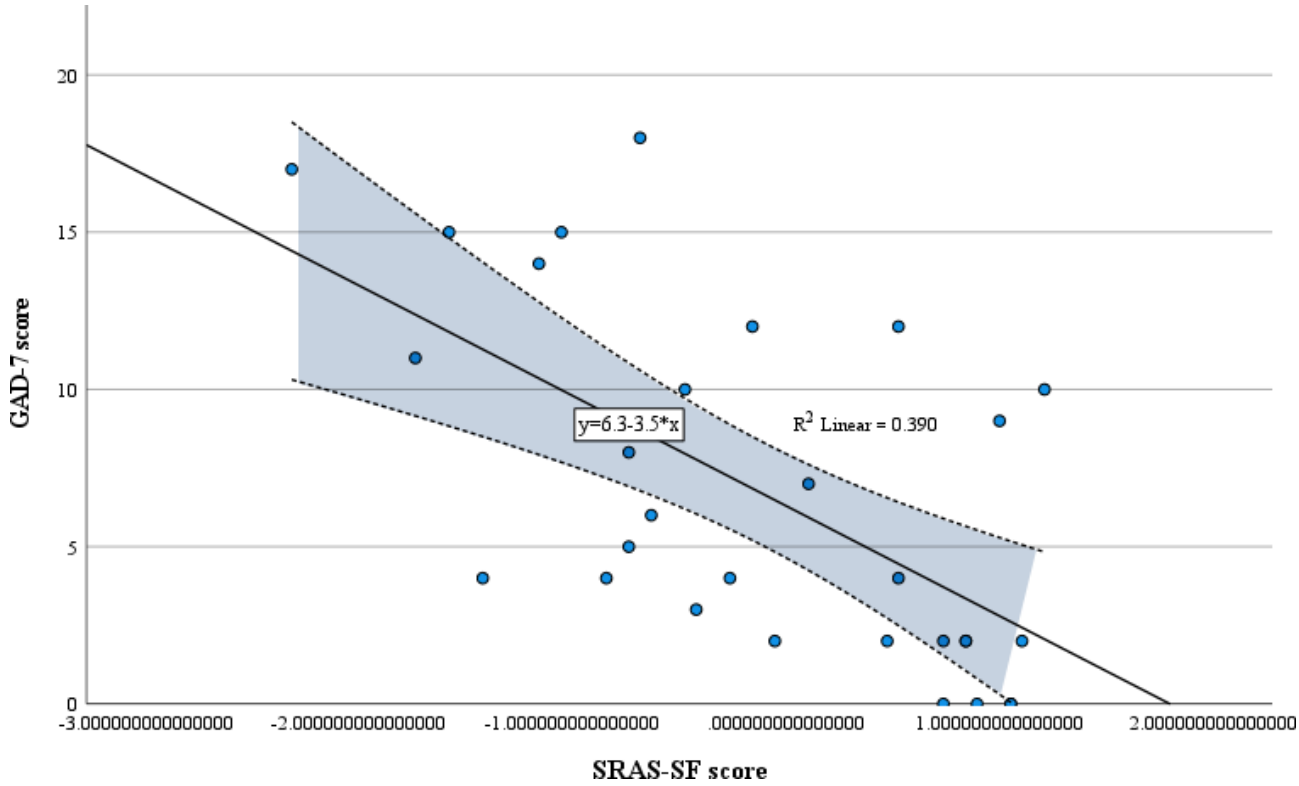

Discussion

Principal Findings

The differences between first-year and second-year medical students' confidence toward returning purchases may reflect cohort-based personality variation. Female participants' greater likelihood of addressing disruptive behavior may reflect gender differences in assertiveness, warranting further study on socialization and professional behaviors in medical education.

The strong negative correlation between assertiveness and anxiety aligns with research indicating medical students' high anxiety levels [7], raising the possibility that greater assertiveness is linked to lower anxiety, though causality cannot be inferred [8]. To determine if assertiveness development mitigates anxiety, future studies should explore whether interventions targeting assertiveness influence students' well-being. Given anxiety's impact on academic performance and mental health, tailored strategies could help students in their training.

This study focuses on preclinical students, limiting applicability to clinical training or residency students. However, medical students face increasing anxiety due to high-stakes evaluations and residency match competitiveness [9,10]. Larger longitudinal studies could clarify this relationship and better inform future interventions.

Limitations

Our small, single-institution sample limits generalizability and statistical power. Future studies should include multiple institutions to account for educational and cultural variations. Additionally, this study focuses on preclinical students, limiting relevance to clinical training or residency students.

Conclusions

Our findings support the existing literature and suggest that assertiveness is inversely associated with preclinical medical students' anxiety [8]. Although confidence typically improves with training, faculty and administrators can implement proactive strategies and training to support students' interpersonal and professional development. Future research should explore longitudinal trends to refine educational interventions that enhance assertiveness and mental well-being.

Acknowledgments

We thank Coretta Jenerette, PhD, RN, for permitting us to use the Simple Rathus Assertiveness Scale-Short Form (SRAS-SF).

XSL•FO
RenderX

Data Availability

The data used to support our conclusions were not acquired from a public repository. The raw data and statistical analyses can be accessed through openICPSR [11].

Authors' Contributions

Conceptualization: J Shaw, JH, AL, KM, AEW, J Shin, VL, BP, CL, PB, AJ
Data curation: J Shaw
Formal analysis: J Shaw
Investigation: J Shaw, JH, AL, KM, AEW, J Shin, VL, BP, CL, PB, AJ
Methodology: J Shaw, AJ
Project administration: J Shaw
Supervision: AJ
Visualization: J Shaw, AL
Writing – original draft: J Shaw, JH, KM, AEW, J Shin, VL, BP, CL, PB
Writing – review & editing: J Shaw, JH, AL, AJ

Conflicts of Interest

None declared.

References

1. Gutgeld-Dror M, Laor N, Karnieli-Miller O. Assertiveness in physicians' interpersonal professional encounters: a scoping review. Med Educ 2024 Apr;58(4):392-404. [doi: 10.1111/medu.15222] [Medline: 37725417]
2. Ben Cherifa D, Saguem BN, Chelbi S, Braham A, Ben Nasr S, Ben Saad H. Predictors of assertive behaviors among a sample of first-year Tunisian medical students. Libyan J Med 2022 Dec;17(1):2095727. [doi: 10.1080/19932820.2022.2095727] [Medline: 35775812]
3. Uy RC, Sarmiento RF, Gavino A, Fontelo P. Confidence and information access in clinical decision-making: an examination of the cognitive processes that affect the information-seeking behavior of physicians. AMIA Annu Symp Proc 2014 Nov 14;2014:1134-1140. [Medline: 25954424]
4. Eslami AA, Rabiei L, Afzali SM, Hamidizadeh S, Masoudi R. The effectiveness of assertiveness training on the levels of stress, anxiety, and depression of high school students. Iran Red Crescent Med J 2016 Jan 2;18(1):e21096. [doi: 10.5812/ircmj.21096] [Medline: 26889390]
5. Jenerette C, Dixon J. Developing a short form of the simple Rathus assertiveness schedule using a sample of adults with sickle cell disease. J Transcult Nurs 2010 Oct;21(4):314-324. [doi: 10.1177/1043659609360712] [Medline: 20592057]
6. Mishra P, Pandey CM, Singh U, Gupta A, Sahu C, Keshri A. Descriptive statistics and normality tests for statistical data. Ann Card Anaesth 2019;22(1):67-72. [doi: 10.4103/aca.ACA_157_18] [Medline: 30648682]
7. Kaiser H, Grice T, Walker B, Kaiser J. Barriers to help-seeking in medical students with anxiety at the University of South Carolina School of Medicine Greenville. BMC Med Educ 2023 Jun 21;23(1):463. [doi: 10.1186/s12909-023-04460-5] [Medline: 37344832]
8. ElBarazi AS, Mohamed F, Mabrok M, et al. Efficiency of assertiveness training on the stress, anxiety, and depression levels of college students (randomized control trial). J Educ Health Promot 2024 Jul 5;13:203. [doi: 10.4103/jehp.jehp_264_23] [Medline: 39268439]
9. LaPaglia D, Robiner WN, Yozwiak JA, Brosig C, Cubic B, Leventhal G. A shortage of medical residency positions: parallels with psychology. Acad Psychiatry 2015 Dec;39(6):706-712. [doi: 10.1007/s40596-015-0324-y] [Medline: 26022617]
10. Lefebvre C, Hartman N, Tooze J, Manthey D. Determinants of medical specialty competitiveness. Postgrad Med J 2020 Sep;96(1139):511-514. [doi: 10.1136/postgradmedj-2019-137160] [Medline: 31780597]
11. Examining the relationship between assertiveness and anxiety in 1st and 2nd year US medical students. openICPSR. 2025. URL: https://www.openicpsr.org/openicpsr/project/221464/version/V2/view [accessed 2025-06-13]

Abbreviations

GAD-7: General Anxiety Disorder 7-item
SRAS-SF: Simple Rathus Assertiveness Scale-Short Form

XSL•FO
RenderX

Edited by J Sarvestan; submitted 03.03.25; peer-reviewed by A Andricioaei, T Kotyk; revised version received 27.05.25; accepted 29.05.25; published 20.06.25.

Please cite as:
Shaw J, Hagerty J, Masada K, Won AE, Lai A, Shin J, Le V, Phung B, Lai C, Bota P, Jacobs A
Examining the Relationship Between Assertiveness and Anxiety in First- and Second-Year US Medical Students
JMIR Form Res 2025;9:e73394
URL: https://formative.jmir.org/2025/1/e73394
doi:10.2196/73394

© Jonathan Shaw, James Hagerty, Kristen Masada, Angelene Eunji Won, Ashley Lai, Jisu Shin, Van Le, Brenton Phung, Charles Lai, Peter Bota, Aaron Jacobs. Originally published in JMIR Formative Research (https://formative.jmir.org), 20.6.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Relationship Between Medical Student Neuroticism and Preferred Medical Specialty: Cross-Sectional Study

Jonathan Shaw[1*], BS; Brenton Phung[1*], BA; Ashley Lai[1], BS; James Hagerty[1], BS; Van Le[1], BA; Seung Rim Yoo[1], BA; Jerome Torres[1], BA; Angelene Won[1], BA; Charles Lai[1], BS; Peter Bota[1], BA; Aaron Jacobs[2], PhD

[1]School of Medicine, California University of Science and Medicine, 1501 Violet St, Colton, CA, United States

[2]Medical Education, California University of Science and Medicine, Colton, CA, United States

[*]these authors contributed equally

Corresponding Author:
Jonathan Shaw, BS
School of Medicine, California University of Science and Medicine, 1501 Violet St, Colton, CA, United States

(JMIR Form Res 2025;9:e73507) doi:10.2196/73507

KEYWORDS

neuroticism; medical students; preclinical; specialty interest; medical education

Introduction

Choosing a specialty is a key decision for US medical students, finalized in the fourth year through the National Resident Matching Program. Influencing factors include experience, gender, lifestyle, finances, and personality [1]. Aligning personality with specialty can boost job satisfaction and patient care [2]. Further, 51% of preclerkship students eye competitive fields, making early insights valuable [3,4].

The 5-factor model includes extraversion, agreeableness, openness, conscientiousness, and neuroticism [5,6]. Neuroticism—linked to anxiety and emotional instability—affects decision-making [7], but its role in specialty choice is underexplored [2]. Studies show that surgeons tend to have low neuroticism [6,8], while future obstetricians and psychiatrists show higher levels [9,10].

This study examines the relationship between neuroticism and specialty choice among medical students at a US allopathic medical school. The goal is to provide insights for personalized medical career counseling and interventions tailored to students' needs.

Methods

Participants and Recruitment

First-year and second-year medical students (n=120 and n=126, respectively) from a California allopathic school with a pass/fail curriculum were invited to complete an anonymous survey from January 11 to 25, 2024, via institutional email. A convenience sample of 32 responses (22 and 10 responses from first-year and second-year medical students, respectively) was collected.

Measures

The survey included demographic questions (school year and gender), 8 questions from the Big Five Inventory assessing neuroticism, and specialty-related questions (specialty interests, perceived competitiveness of chosen specialties, and self-perceived preparedness for matching). Question order was randomized.

Statistical Analysis

Data analysis was conducted using IBM SPSS Statistics 28.0.1.0 (IBM Corp). Normality was assessed via Kolmogorov-Smirnov tests. Independent samples *t* tests (2-tailed) were used for normal data; Kruskal-Wallis tests were used for nonnormal data. The grouping variables were school year, gender, and specialty interest type. Statistical significance was determined as $P \leq .05$.

Ethical Considerations

This study received ethical approval from the California University of Science and Medicine Institutional Review Board (approval: HS-2023 - 59) on November 27, 2023. Informed consent for primary data collection and secondary analyses of the research data was obtained from all individual participants included in this study. Participants received no compensation for participation.

Results

Kolmogorov-Smirnov tests indicated normal distributions for self-reported surgical, medical, and overall competitiveness, as well as neuroticism scores. Independent *t* tests revealed no significant differences based on gender, school year, or specialty interest type. However, students interested in both surgical and medical specialties had lower neuroticism scores ($t_{30}=-2.039$; $P=.03$) than those of students who focused on only one field.

Second-year medical students reported feeling tenser ($H_1=4.867$; $P=.03$) and moodier ($H_1=10.364$; $P=.001$) than first-year medical students. Students interested in surgical specialties perceived their specialties as more competitive ($H_1=4.080$; $P=.04$). Participants interested in both surgery and medicine reported

XSL•FO
RenderX

feeling more relaxed (H_1=5.882; *P*=.02) than those interested in only one field. Further survey data are presented in Table 1.

XSL•FO
RenderX

Table . Descriptive statistics for survey items from this cross-sectional, web-based survey study, which was conducted at a California allopathic medical school via institutional email from January 11 to 25, 2024.

Survey items	Responses, n	Minimum score	Maximum score	Score, mean	Score, SD
Which school year are you?[a]	32	1	2	1.31	0.471
What gender do you identify as?[b]	32	1	2	1.41	0.499
Surgery[c]	32	1	2	1.59	0.499
Medicine[c]	32	1	2	1.19	0.397
Both[c]	32	1	2	1.69	0.471
Surgical competitiveness[d]	15	1	5	3.27	1.123
Medical competitiveness[d]	25	1	5	2.83	1.197
Overall competitiveness[d]	29	1	5	2.91	1.058
Neuroticism vs emotional stability[e]	32	6	26	15.19	6.061
Is depressed, blue[f]	32	1	4	2.25	1.218
Is relaxed, handles stress well[f]	32	1	5	3.72	1.301
Can be tense[f]	32	1	5	3.34	1.153
Worries a lot[f]	32	1	5	3.28	1.35
Is emotionally stable, not easily upset[f]	32	1	5	4.03	1.15
Can be moody[f]	32	1	5	2.84	1.247
Remains calm in tense situations[f]	32	2	5	3.84	1.11
Gets nervous easily[f]	32	1	5	3.34	1.181
Medical school feels overwhelming[f]	32	1	5	3.16	1.37
I feel I am very prepared academically[f]	32	1	5	3.5	1.164
I am a competitive applicant for the specialties I have in mind[f]	32	1	5	3.44	1.216
I worry a lot about matching into my preferred specialty[f]	32	1	5	3.25	1.27
I am confident in my ability to succeed as a medical student[f]	32	1	5	3.91	1.027
The specialties I am considering are not that hard to match into[f]	32	1	5	2.78	1.289

XSL•FO
RenderX

Survey items	Responses, n	Minimum score	Maximum score	Score, mean	Score, SD
I have a plan for how I'm going to match into the specialty that I want[f]	32	1	5	3.5	1.164
I have access to enough resources to become a competitive applicant for my specialties of interest[f]	32	1	5	3.69	1.176
I'm worried about not having enough research/publications for my specialty of interest[f]	32	1	5	3.66	1.335
I'll have to take a gap/research year to match into my specialties of choice[f]	32	1	5	1.84	1.194

[a]1=first year; 2=second year.

[b]1=woman; 2=man.

[c]Participants' interest in surgical, medical, and mixed/both specialties was indicated in their demographic section by checking either one or both types of specialties (1=no; 2=yes).

[d]Competitiveness was determined by participants self-reporting their perceived competitiveness on a scale of 1 to 5, with 1 being "not very competitive," 3 being "average competitiveness," and 5 being "very competitive." Participants indicated which specific specialties they wished to pursue (urology, emergency medicine, etc), and these were used to determine which group (surgical vs medical) their responses were analyzed with.

[e]Neuroticism vs emotional stability was determined by summing participant responses to a series of 5-point Likert questions from the 5-factor personality test, with 1 being "disagree strongly" and 5 being "agree strongly." Participants scoring 6 were seen as more emotionally stable, while those scoring 26 were seen as more neurotic.

[f]Scored on a Likert scale from 1 to 5, with 1 being "I disagree strongly" and 5 being "I agree strongly."

Discussion

We found that surgery-interested students viewed their field as more competitive, but their neuroticism levels were similar to those of peers interested in other fields. They may need more academic support, and mental health resources should be cohort-specific, as participants showed moderately elevated neuroticism (mean 15.19, SD 6.061) when compared to literature [6].

Surgery-interested students viewed fields like orthopedics and neurosurgery as more competitive, which is consistent with prior research [2]. Unlike existing literature, this study found similar neuroticism levels between surgical and medical students (F_{30}=0.220; P=.64) [9,10]. Students interested in both surgical and medical specialties reported lower neuroticism and better stress management (F_{30}=5.142; P=.03), possibly due to having interest in less competitive fields or having backup plans. No gender differences were found, though second-year medical students showed higher tension (H_1=4.867; P=.03) and moodiness (H_1=10.364; P=.001), likely due to board examination stress. The absence of gender effects may be due to the small sample size of 32 participants.

These discrepancies may have resulted from the small sample size or the early stage of participants' medical education. Since this study was not longitudinal, it could not capture changes in specialty interests that may occur during participants' clinical years. Although the medical program's pass/fail curriculum may have helped with reducing stress and neuroticism, its impact remains unclear due to this study's single-institution scope.

This study found a link between neuroticism and specialty interest among medical students. The pressure to choose and prepare for a specialty so early in medical school may explain the neuroticism and stress found in preclinical participants, highlighting the need for early experiences with specialties, so that students can select their field and begin tailoring their activities as early as possible.

Data Availability

The data used to support our conclusions were not acquired from a public repository. The raw data and statistical analyses can be accessed through openICPSR [11].

XSL•FO
RenderX

Conflicts of Interest

None declared.

References

1. Sarikhani Y, Ghahramani S, Bayati M, Lotfi F, Bastani P. A thematic network for factors affecting the choice of specialty education by medical students: a scoping study in low-and middle-income countries. BMC Med Educ 2021 Feb 10;21(1):99. [doi: 10.1186/s12909-021-02539-5] [Medline: 33568113]
2. Kase J, Doolittle B. Job and life satisfaction among emergency physicians: a qualitative study. PLoS One 2023 Feb 24;18(2):e0279425. [doi: 10.1371/journal.pone.0279425] [Medline: 36827313]
3. Radulovich NP, Burke S, Brown NJ, et al. The importance of research experience with a scoreless step 1: a student survey at a community-based medical school. Cureus 2023 Aug 14;15(8):e43476. [doi: 10.7759/cureus.43476] [Medline: 37711915]
4. Lefebvre C, Hartman N, Tooze J, Manthey D. Determinants of medical specialty competitiveness. Postgrad Med J 2020 Sep;96(1139):511-514. [doi: 10.1136/postgradmedj-2019-137160] [Medline: 31780597]
5. Sultan S, Labban OM, Hamawi AM, et al. Relationship of big five personality traits and future specialty preference among undergraduate medical students: a cross-sectional study. Egypt J Neurol Psychiatr Neurosurg 2023 Jul 24;59(1):103. [doi: 10.1186/s41983-023-00699-3]
6. Stienen MN, Scholtes F, Samuel R, Weil A, Weyerbrock A, Surbeck W. Different but similar: personality traits of surgeons and internists-results of a cross-sectional observational study. BMJ Open 2018 Jul 7;8(7):e021310. [doi: 10.1136/bmjopen-2017-021310] [Medline: 29982214]
7. Savchenko OV, Muzychko LV, Kolesnichenko LA. The influence of unhealthy neuroticism on the peculiarities of personal decision making. Wiad Lek 2021;74(11 cz 1):2695-2699. [Medline: 35023477]
8. Kalywis AL, Samuel R, Scholtes F, et al. Distribution of psychological instability among surgeons. World Neurosurg 2023 Jul;175:e531-e541. [doi: 10.1016/j.wneu.2023.03.137] [Medline: 37028482]
9. Borracci RA, Ciambrone G, Arribalzaga EB. Tolerance for uncertainty, personality traits and specialty choice among medical students. J Surg Educ 2021;78(6):1885-1895. [doi: 10.1016/j.jsurg.2021.03.018] [Medline: 34001460]
10. Tobiaszewska M, Koweszko T, Jurek J, et al. Personality types of medical students in terms of their choice of medical specialty: cross-sectional study. Interact J Med Res 2024 Dec 31;13:e60223. [doi: 10.2196/60223] [Medline: 39740218]
11. The relationship between medical student neuroticism and preferred medical specialty. openICPSR. 2025. URL: https://www.openicpsr.org/openicpsr/project/221542/version/V2/view [accessed 2025-06-10]

Edited by A Mavragani; submitted 05.03.25; peer-reviewed by A Andricioaei, F Quarshie, S Dream; revised version received 07.05.25; accepted 13.05.25; published 18.06.25.

Please cite as:
Shaw J, Phung B, Lai A, Hagerty J, Le V, Yoo SR, Torres J, Won A, Lai C, Bota P, Jacobs A
The Relationship Between Medical Student Neuroticism and Preferred Medical Specialty: Cross-Sectional Study
JMIR Form Res 2025;9:e73507
URL: https://formative.jmir.org/2025/1/e73507
doi:10.2196/73507

© Jonathan Shaw, Brenton Phung, Ashley Lai, James Hagerty, Van Le, Seung Rim Yoo, Jerome Torres, Angelene Won, Charles Lai, Peter Bota, Aaron Jacobs. Originally published in JMIR Formative Research (https://formative.jmir.org), 18.6.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Seasonal Variation in Conjunctivitis in Saudi Arabia and Reduction of Searches During the COVID-19 Pandemic: Analysis of Google Trends Data

Abdulaziz S AlHarthi, MD

Ophthalmology Department, College of Medicine, Majmaah University, Al-Majmaah, Saudi Arabia

Corresponding Author:
Abdulaziz S AlHarthi, MD
Ophthalmology Department, College of Medicine, Majmaah University, Al-Majmaah, Saudi Arabia

Abstract

Analysis of Google Trends data for the Arabic term for "conjunctivitis" found significant seasonality, with the highest interest in February and the lowest in August (*P*<.001); moreover, during COVID-19 restrictions, relative search volume decreased by 53.3%, with a disruption of the seasonal pattern.

(JMIR Form Res 2025;9:e73845) doi:10.2196/73845

KEYWORDS

Google Trends; conjunctivitis; COVID-19; seasonality

Introduction

Conjunctivitis is one of the most common eye-related infections worldwide, and it frequently results in nonemergency ophthalmic visits to health care facilities [1]. Previous studies have observed seasonal patterns in the incidence of cases. In the United States, cases have been reported to be predominant in the winter and spring [2].

During the COVID-19 pandemic, preventive measures to decrease transmission were implemented worldwide. In the Kingdom of Saudi Arabia, this included the obligation to wear a face mask, social distancing, and the suspension of gatherings in educational institutions. A curfew was imposed in late March 2020; all measures were lifted entirely in March 2022.

To date, few studies have explored the impact of COVID-19 restrictions on the spread of other infectious diseases. Here, we explore Google Trends data to determine whether public search volumes for conjunctivitis in Saudi Arabia exhibit seasonality and how this was impacted during the COVID-19 period.

Methods

Overview

Monthly normalized relative search volume data (0 - 100) were extracted as CSV files from Google Trends in September 2024 for the search term " " (Arabic for "conjunctivitis"), with the location set to Saudi Arabia and the period from January 2011 to October 2024, without a category selection. Statistical analyses were conducted in R (version 4.4.2; R Foundation) with the *Forecast*, *seatest*, and *season* packages.

Poisson cosinor analysis was performed to assess seasonality during the pre–COVID-19 period (2011 - 2019), avoiding pattern distortions due to COVID-19–related search behavior, with the significance level set at .025, consistent with previous studies [3]. To estimate the impact of COVID-19 restrictions, a log-transformed autoregressive integrated moving average (ARIMA) model (1,0,0) was fitted on data from 2011 to 2023, including binary variables for the intervention period (March 2020 to March 2022), monthly dummy variables, lag-12, and linear and quadratic trends (significance at *P*<.05). Residual diagnostics indicated approximate normality and passed the Ljung-Box test and passed the Ljung-Box test (χ^2=16.19; *P*=.06) which indicates white noise in the residuals. Seasonality disruption was assessed using the Friedman test across the years before and during COVID-19.

Ethical Considerations

The data are public source, so no institutional review board approval was needed.

Results

Cosinor analysis confirmed a statistically significant seasonal pattern in the relative search volume for conjunctivitis (amplitude=7.57, phase month=2.7, and low-point month=8.7; *P*<.001; Figure 1). A relative search peak was observed in February and the lowest volume during summer, in August. An interrupted model using ARIMA showed an estimated 53.3% reduction (95% CI −43.2% to 61.6%; *P*<.001) in the relative search volume for conjunctivitis during the COVID-19 quarantine compared to the pre–COVID-19 period. Notably, the Friedman test showed that the prepandemic seasonal pattern (when the value was *P*=.03) was disrupted during pandemic

XSL•FO
RenderX

restriction policies from 2020 to 2022 (when the value changed to P=.20).

Figure 1. Cosinor model plot for conjunctivitis seasonality in Google Trends search volume in Saudi Arabia for the Arabic word for "conjunctivitis" for the period from January 2011 to December 2019.

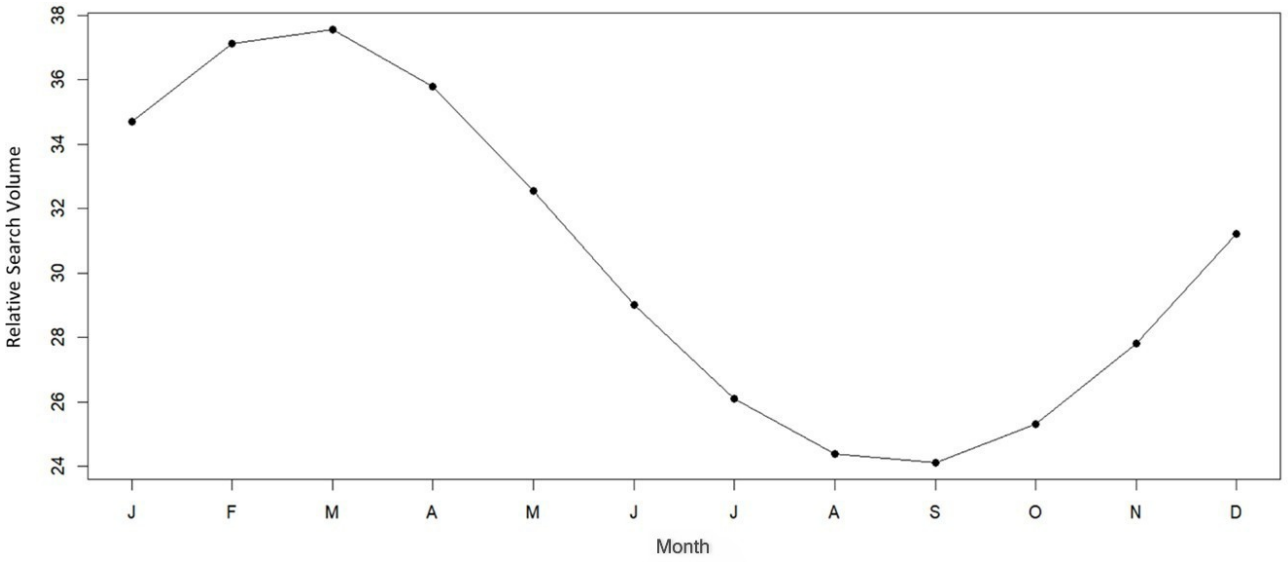

Discussion

Principal Findings

This study shows that there is significant seasonality in the search volume for conjunctivitis. Moreover, there was a significant reduction in and disruption of seasonality during the pandemic, providing a line of evidence to complement traditional methods; the finding suggests that the preventive strategies taken to lower the risk of COVID-19 transmission may have also reduced the incidence of conjunctivitis or at least online search interest.

Previous studies from the United States have explored the seasonality of searches for conjunctivitis and obtained similar findings, that is, that search increases from fall to spring [2]. Several factors lead to an increased risk of conjunctivitis during winter, including indoor crowdedness and an increased risk of contact with infected people, especially in schools. Additionally, conjunctivitis cases increase with other infections, like the flu, that exhibit seasonal variation.

Adenoviral conjunctivitis has a similar transmission mechanism as coronaviruses [4], with transmission rates of 30% in nursing homes and 76% in schools [5,6]. Recently, Deiner et al [7] showed that there was a sustained reduction in searches for communicable diseases, including conjunctivitis, during the pandemic in the United States. In a study from Spain, there was a decrease in the incidence of conjunctivitis by 48.5% during the pandemic in 2020 [8]. Additionally, Lavista Ferres et al [9] published evidence of a decrease in nonallergic conjunctivitis cases in the emergency department of 37.3%.

Limitations

These findings are based on Google Trends search volume, which may be influenced by other factors, such as the media, not only by the actual incidence of conjunctivitis. Therefore, further studies to elaborate on the seasonality and impact of COVID-19 restrictions are needed.

Conclusions

These findings highlight the potential seasonality of conjunctivitis. Understanding disease patterns and the role of public health measures against COVID-19 in reducing conjunctivitis spread has public health implications for the reduction of conjunctivitis prevalence.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

References

1. Channa R, Zafar SN, Canner JK, Haring RS, Schneider EB, Friedman DS. Epidemiology of eye-related emergency department visits. JAMA Ophthalmol 2016 Mar;134(3):312-319. [doi: 10.1001/jamaophthalmol.2015.5778] [Medline: 26821577]

XSL•FO
RenderX

2. Deiner MS, Lietman TM, McLeod SD, Chodosh J, Porco TC. Surveillance tools emerging from search engines and social media data for determining eye disease patterns. JAMA Ophthalmol 2016 Sep 1;134(9):1024-1030. [doi: 10.1001/jamaophthalmol.2016.2267] [Medline: 27416554]
3. Kardeş S. Seasonal variation in the internet searches for psoriasis. Arch Dermatol Res 2019 Aug;311(6):461-467. [doi: 10.1007/s00403-019-01921-0] [Medline: 31025101]
4. Li K, Li L, Wang X, et al. Comparative analysis of clinical features of SARS-CoV-2 and adenovirus infection among children. Virol J 2020 Dec 10;17(1):193. [doi: 10.1186/s12985-020-01461-4] [Medline: 33302983]
5. Madurapandian Y, Rubeshkumar P, Raju M, et al. Case report: an outbreak of viral conjunctivitis among the students and staff of visually impaired school, Tamil Nadu, India, 2020. Front Public Health 2022;10:978200. [doi: 10.3389/fpubh.2022.978200] [Medline: 35991078]
6. Domínguez-Berjón MF, Hernando-Briongos P, Miguel-Arroyo PJ, Echevarría JE, Casas I. Adenovirus transmission in a nursing home: analysis of an epidemic outbreak of keratoconjunctivitis. Gerontology 2007;53(5):250-254. [doi: 10.1159/000101692] [Medline: 17429213]
7. Deiner MS, Seitzman GD, Kaur G, et al. Sustained reductions in online search interest for communicable eye and other conditions during the COVID-19 pandemic: infodemiology study. JMIR Infodemiology 2022;2(1):e31732. [doi: 10.2196/31732] [Medline: 35320981]
8. Conde Bachiller Y, Puente Gete B, Gil Ibáñez L, Esquivel Benito G, Asencio Duran M, Dabad Moreno JV. COVID-19 pandemic: impact on the rate of viral conjunctivitis. Arch Soc Esp Oftalmol (Engl Ed) 2022 Feb;97(2):63-69. [doi: 10.1016/j.oftale.2022.01.001] [Medline: 35152950]
9. Lavista Ferres JM, Meirick T, Lomazow W, Lee CS, Lee AY, Lee MD. Association of public health measures during the COVID-19 pandemic with the incidence of infectious conjunctivitis. JAMA Ophthalmol 2022 Jan 1;140(1):43-49. [doi: 10.1001/jamaophthalmol.2021.4852] [Medline: 34792555]

Abbreviations

ARIMA: autoregressive integrated moving average

Edited by A Mavragani; submitted 12.03.25; peer-reviewed by T Porco; revised version received 09.06.25; accepted 11.06.25; published 21.07.25.

Please cite as:
AlHarthi AS
Seasonal Variation in Conjunctivitis in Saudi Arabia and Reduction of Searches During the COVID-19 Pandemic: Analysis of Google Trends Data
JMIR Form Res 2025;9:e73845
URL: https://formative.jmir.org/2025/1/e73845
doi:10.2196/73845

© Abdulaziz S AlHarthi. Originally published in JMIR Formative Research (https://formative.jmir.org), 21.7.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

How Medical Students Manage Depression, Anxiety, and Stress: A Cross-Sectional Study

Jonathan Shaw[1*], BS; Ashley Lai[1*], BS; Sasha Singh[1], BS; Seung Rim Yoo[1], BA; Maha Fathali[1], BS; Laura Stuck[1], BS; James Hagerty[1], BS; Van Le[1], BA; Jisu Shin[1], BS; Charles Lai[1], BS; Peter Bota[1], BA; Aaron Jacobs[2], PhD

[1]School of Medicine, California University of Science and Medicine, 1501 Violet St, Colton, CA, United States

[2]Medical Education, California University of Science and Medicine, Colton, CA, United States

[*]these authors contributed equally

Corresponding Author:
Jonathan Shaw, BS
School of Medicine, California University of Science and Medicine, 1501 Violet St, Colton, CA, United States

Abstract

Given the established challenging nature of medical school, first- and second-year medical students were surveyed on their feelings of stress alongside their coping mechanisms; the study found a higher level of stress among second-year medical students, indicating a need for support during difficult transition periods.

(JMIR Form Res 2025;9:e74218) doi:10.2196/74218

KEYWORDS

medical student; stress; depression; anxiety; coping

Introduction

Due to the intense pressure and stress of the training process, medical students face a higher risk of burnout and depression than the general US population [1]. They must pursue extracurriculars, like research, to be competitive for residency applications. Common stressors include academics, peer-to-peer relationships, and finances [2], which can impact sleep quality, physical and mental health, and academic performance [3]. Stress can lead to burnout, with some even considering leaving medicine [4]. Medical students may cope with stress through healthy habits like exercise or harmful ones like alcohol abuse [5,6].

This study explores how first- and second-year medical students manage and reduce stress, to improve medical student wellness efforts.

Methods

Participants and Recruitment

Overall, 120 first-year and 126 second-year students at a California allopathic medical school were contacted via institutional email to complete an anonymous survey. One survey round was sent, with 27 responses collected.

Measures

The survey comprised 2 demographic questions (school year and gender); 24 multiple-choice questions on stress management; 1 free-response question for any unlisted stress management activities and their frequency; and 21 randomly shuffled questions from the Depression, Anxiety and Stress Scale–21 Items (DASS-21) [7]. The hobbies listed were based on community observation.

Statistical Analysis

Data were analyzed using IBM SPSS Statistics (version 28.0.1.0). Normality was assessed via the Kolmogorov-Smirnov test. For nonnormally distributed data, a Kruskal-Wallis test (via *k* independent samples) was conducted, with gender and school year as the grouping variables and the range being 1-2. Additionally, Spearman correlations were obtained through SPSS's bivariate correlation function.

Ethical Considerations

This study received ethical approval from the California University of Science and Medicine Institutional Review Board (HS-2023‑56) on November 27, 2023. Informed consent was obtained from all participants included in the study for data collection and analyses. Identifying information was not collected. Participants were not compensated.

Results

The Kolmogorov-Smirnov test found that neither the questions nor the DASS-21 subscores were normally distributed. The Kruskal-Wallis results are presented in Table 1. Spearman correlations showed that school year was positively and moderately correlated with stress (r_{27}=.699; P<.001), anxiety (r_{27}=.585; P<.001), and depression (r_{27}=.408; P=.03).

XSL•FO
RenderX

Second-year medical students had mild depression (13.67), with moderate anxiety (11.17) and stress (20.50; Figure 1). In contrast, first-year medical students were in the "normal" range across all categories. This demonstrates that second-year medical students experience higher levels of distress.

Table . The Kruskal-Wallis results. DASS-21[a] scoring was conducted by summing the Likert-scale responses (0=did not apply to me at all, 3=applied to me very much or most of the time) for each subscale (Depression, Anxiety, and Stress) and multiplying this sum by 2 to determine the participant severity per subscale. Depression thresholds included normal (0 - 9), mild (10-13), moderate (14-20), severe (21-27), and extremely severe (28+). Anxiety thresholds included normal (0 - 7), mild (8-9), moderate (10-14), severe (15-19), and extremely severe (20+). Stress thresholds included normal (0 - 14), mild (15-18), moderate (19-25), severe (26-33), and extremely severe (34+). All other questions used a Likert scale (0=not at all, 3=nearly every day), asking participants if they had partaken in a stress management behavior within the last 2 weeks. This included a free-response question, which asks participants to fill in any unlisted activities and self-report the frequency of the activity in the same manner as the Likert scale.

Question or topic	Kruskal-Wallis *H* (*df*=1)	*P* value	Group 1, mean (SD)	Group 2, mean (SD)
Read books to relax	4.994	.02	.67 (.900)[b]	.00 (.000)[c]
Talk with friends	5.888	.02	2.28 (.895)[b]	1.44 (.726)[c]
Go out with friends in person	5.081	.02	1.14 (.910)[d]	.20 (.447)[e]
Painting	4.000	.046	.00 (.000)[d]	.75 (1.500)[e]
Find it harder to wind down	12.222	.001	.62 (.590)[d]	2.50 (.837)[e]
Aware of having a dry mouth	10.300	.001	.10 (.301)[d]	1.50 (1.225)[e]
Could not experience positive feelings	7.502	.006	.14 (.359)[d]	1.33 (1.366)[e]
Difficulty breathing	5.234	.02	.10 (.301)[d]	.67 (.816)[e]
Overreacting to situations	6.131	.01	.19 (.402)[d]	1.17 (1.169)[e]
Using a lot of nervous energy	13.211	.001	.57 (.676)[d]	2.33 (.516)[e]
Getting more agitated	6.911	.009	.57 (.870)[d]	2.00 (1.095)[e]
Difficult to relax	9.453	.002	.81 (.873)[d]	2.50 (.837)[e]
Felt downhearted and blue	4.374	.04	.48 (.602)[d]	1.33 (1.033)[e]
Felt close to panic	8.576	.003	.14 (.478)[d]	1.17 (1.169)[e]
Feeling scared without any good reason	5.565	.02	.24 (.539)[d]	1.17 (1.169)[e]
Stress levels (DASS-21)	12.715	.001	5.48 (4.665)[d]	20.50 (4.416)[e]
Anxiety levels (DASS-21)	8.894	.003	2.43 (3.091)[d]	11.17 (7.139)[e]
Depression levels (DASS-21)	4.336	.04	4.67 (5.677)[d]	13.67 (11.535)[e]

[a]DASS-21: Depression, Anxiety and Stress Scale–21 Items.

[b]Female group.

[c]Male group.

[d]First-year group.

[e]Second-year group.

XSL•FO
RenderX

Figure 1. DASS-21 scoring differences across the Depression, Anxiety, and Stress sections between first- and second-year medical students. The DASS-21 can be divided into 3 subsections: Depression, Anxiety, and Stress. Depression thresholds included normal (0 - 9), mild (10-13), moderate (14-20), severe (21-27), and extremely severe (28+). Anxiety thresholds included normal (0 - 7), mild (8-9), moderate (10-14), severe (15-19), and extremely severe (20+). Stress thresholds included normal (0 - 14), mild (15-18), moderate (19-25), severe (26-33), and extremely severe (34+). Averaged DASS-21 scores are displayed, with the bars representing the SD. DASS-21: Depression, Anxiety and Stress Scale–21 Items.

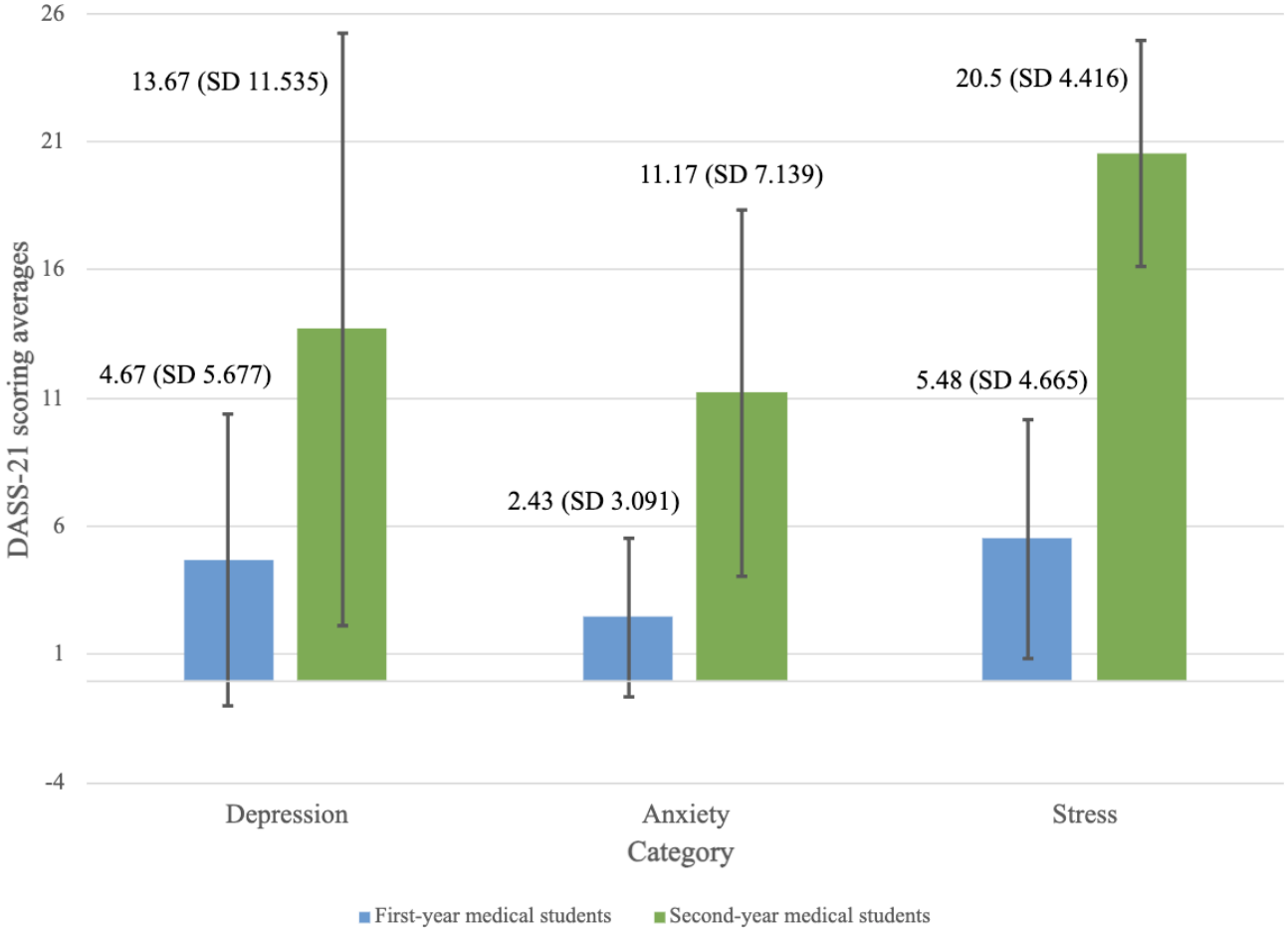

Discussion

Female participants were more likely to read (H_1=4.994; P=.02) and talk with friends (H_1=5.888; P=.02). This aligns with the existing literature, as women have been shown to read more than men, but this could also be due to our limited sample size [8]. Exercise habits did not significantly differ by school year (P=.09) or gender (P=.71). Prior studies have shown that exercise reduces burnout and improves quality of life [6], suggesting that medical schools should encourage physical activity, especially for students studying remotely.

First-year medical students were more likely to go out with friends (H_1=5.081; P=.02). Second-year medical students had varied schedules, due to rotations or professional development semesters, which at this institution are a dedicated 6-month period for self-directed activities such as research projects or studying for the United States Medical Licensing Examination (USMLE) Step exams. This may limit socialization opportunities compared to first-year medical students, who remain on a shared schedule. On the DASS-21, second-year medical students had moderate stress levels compared to first-year medical students—who had "normal" stress levels (a 15-point difference)—potentially due to clinical rotations, professional development semesters, or Step 1 exam preparation. The 6 - to 8-week "dedicated" study period is highly stressful, and despite Step 1's transition to a pass-or-fail grade, concerns remain about its impact on mental health. However, it is important to acknowledge that due to the limited sample size (n=27), it would be difficult to make broad generalizations. Similarly, this raises the question of validity, as students at other medical schools may have different experiences.

These findings highlight the need to focus on increasing social support through events such as mug painting, potlucks, or board game nights. This will help students build a network, better cope with stress, and reduce the negative feelings they experience [9]. Future research should focus on optimizing such interventions not only for second-year medical students but also for third- and fourth-year medical students, who were not included in this study. Future studies should also measure stress management and stress levels longitudinally across cohorts of various institutions so that standardized approaches can be developed, which can further be adjusted to fit the cultures of individual institutions.

XSL•FO
RenderX

Acknowledgments

The authors declare that they did not use generative artificial intelligence in the writing of this research letter.

Data Availability

The data used to support the conclusions were not acquired from a public repository. The raw data and statistical analyses can be accessed on the web [10].

Authors' Contributions

Conceptualization: J Shaw, PB, CL
Data curation: J Shaw, MF
Formal analysis: J Shaw
Supervision: AJ
Visualization: J Shaw, CL
Writing - original draft: AL, SS, SRY, LS, MF, JH, VL, J Shin
Writing - reviewing & editing: J Shaw, SS, AL, MF, SRY

Conflicts of Interest

None declared.

References

1. Dyrbye LN, West CP, Satele D, et al. Burnout among U.S. medical students, residents, and early career physicians relative to the general U.S. population. Acad Med 2014 Mar;89(3):443-451. [doi: 10.1097/ACM.0000000000000134] [Medline: 24448053]
2. Hill MR, Goicochea S, Merlo LJ. In their own words: stressors facing medical students in the millennial generation. Med Educ Online 2018 Dec;23(1):1530558. [doi: 10.1080/10872981.2018.1530558] [Medline: 30286698]
3. Almojali AI, Almalki SA, Alothman AS, Masuadi EM, Alaqeel MK. The prevalence and association of stress with sleep quality among medical students. J Epidemiol Glob Health 2017 Sep;7(3):169-174. [doi: 10.1016/j.jegh.2017.04.005] [Medline: 28756825]
4. Rajapuram N, Langness S, Marshall MR, Sammann A. Medical students in distress: the impact of gender, race, debt, and disability. PLoS One 2020 Dec 3;15(12):e0243250. [doi: 10.1371/journal.pone.0243250] [Medline: 33270759]
5. Jackson ER, Shanafelt TD, Hasan O, Satele DV, Dyrbye LN. Burnout and alcohol abuse/dependence among U.S. medical students. Acad Med 2016 Sep;91(9):1251-1256. [doi: 10.1097/ACM.0000000000001138] [Medline: 26934693]
6. Taylor CE, Scott EJ, Owen K. Physical activity, burnout and quality of life in medical students: a systematic review. Clin Teach 2022 Dec;19(6):e13525. [doi: 10.1111/tct.13525] [Medline: 36052814]
7. Lovibond SH, Lovibond PF. Depression Anxiety and Stress Scales. American Psychological Association. 1995. URL: https://psycnet.apa.org/doi/10.1037/t39835-000 [accessed 2025-06-25]
8. Cerbara L, Ciancimino G, Tintori A. Are we still a sexist society? primary socialisation and adherence to gender roles in childhood. Int J Environ Res Public Health 2022 Mar 14;19(6):3408. [doi: 10.3390/ijerph19063408] [Medline: 35329095]
9. Drageset J. Social support. In: Haugan G, Eriksson M, editors. Health Promotion in Health Care – Vital Theories and Research: Springer; 2021:137-144. [doi: 10.1007/978-3-030-63135-2_11]
10. Shaw J, Lai A, Singh S, et al. Depression, anxiety, and stress: how do medical students manage them? Inter-university Consortium for Political and Social Research. URL: https://doi.org/10.3886/E221562V1 [accessed 2025-06-26]

Abbreviations

DASS-21: Depression, Anxiety and Stress Scale–21 Items
USMLE: United States Medical Licensing Examination

Edited by A Mavragani; submitted 19.03.25; peer-reviewed by A Andricioaei, M Adnan; revised version received 15.05.25; accepted 03.06.25; published 07.07.25.

Please cite as:
Shaw J, Lai A, Singh S, Yoo SR, Fathali M, Stuck L, Hagerty J, Le V, Shin J, Lai C, Bota P, Jacobs A
How Medical Students Manage Depression, Anxiety, and Stress: A Cross-Sectional Study
JMIR Form Res 2025;9:e74218
URL: https://formative.jmir.org/2025/1/e74218
doi:10.2196/74218

XSL•FO
RenderX

© Jonathan Shaw, Ashley Lai, Sasha Singh, Seung Rim Yoo, Maha Fathali, Laura Stuck, James Hagerty, Van Le, Jisu Shin, Charles Lai, Peter Bota, Aaron Jacobs. Originally published in JMIR Formative Research (https://formative.jmir.org), 7.7.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

The Digital Availability of US Departments of Corrections' Research Policies: Cross-Sectional Analysis

Jesse Martinez-Kratz[1], MPH, AB; David Manning[2], BS; Noel Vest[3], PhD; Jason Glenn[4], PhD; Lauren Brinkley-Rubinstein[5], PhD; Alysse Wurcel[6,7], MD, MS

[1]School of Medicine, Tufts University, 145 Harrison Ave, Boston, MA, United States
[2]School of Arts and Sciences, Tufts University, Boston, MA, United States
[3]Department of Community Health Science, School of Public Health, Boston University, Boston, MA, United States
[4]Department of History and Philosophy of Medicine, University of Kansas Medical Center, Kansas City, KS, United States
[5]Department of Population Health Sciences, Duke University, Durham, NC, United States
[6]Department of Geographic Medicine and Infectious Diseases, Tufts Medical Center, Boston, MA, United States
[7]Section of General Internal Medicine, Boston Medical Center, Boston, MA, United States

Corresponding Author:
Jesse Martinez-Kratz, MPH, AB
School of Medicine, Tufts University, 145 Harrison Ave, Boston, MA, United States

Abstract

This cross-sectional survey of US Departments of Correction websites found significant heterogeneity in the availability and content of prison-based research policy, procedure, and contact information.

(JMIR Form Res 2025;9:e76835) doi:10.2196/76835

KEYWORDS

prison; consent; websites; Department of Corrections; incarceration; research policy

Introduction

Mass incarceration poses a public health crisis in the United States as a major structural determinant of health [1]. Exposure to incarceration heightens risks of chronic conditions, including cardiovascular disease and cancer, while significantly increasing all-cause mortality [2]. Prior unethical research resulted in federal restrictions on research conducted on criminal-legal involved populations [3]. Varied interpretations of federal regulations have led to heterogenous operationalization of research systems, evidenced in varied state Department of Corrections (DOC) policies in state prisons. Ethical research is needed to measure health-related outcomes for those in prisons; however, these varied interpretations of federal regulations create barriers. We evaluated the accessibility of research-related information on publicly available websites of state DOCs.

Methods

Overview

We iteratively developed an 8-point rubric for evaluating the digital imprint and internet-accessible completeness of each state's DOC's research policies. Two research team members (JM-K and DM) independently located DOC-based websites (June-August 2023). If the state had a website, the research team sought the presence of a DOC-based research web page. States lacking any DOC website or web page were excluded. For states lacking either a DOC website or a DOC-based research web page, researchers searched "[STATE] Research Department of Corrections" to assess for external websites containing information about prison-based research policies and procedures that appeared within 25 search results.

If a DOC research website or web page was found for a given state, we assigned up to 8 points, analyzing for the (1) presence of contact information provided for a DOC's research-related representative and responsiveness within 3 weeks to initiated contact, (2) identification of an institutional review board (IRB) and available policy information. Of available policy information, we evaluated for the presence of policies for (3) staff recruitment, (4) incarcerated people's recruitment, (5) staff consent, (6) incarcerated people's consent, (7) staff stipends/payment, and (8) incarcerated people's stipends/payment. Criteria were piloted on 15 states to ensure interrater reliability (>75%) before assessing the remaining 35, with a final interrater reliability of 94.4%. All data were stored in Qualtrics [4]. Eight-point scores are visualized in Figure 1, and descriptive univariate analysis for categorical variables is reported in Table 1.

XSL•FO
RenderX

Figure 1. Choropleth map of policy indicators from state DOC websites, generated with MapChart [5]. DOC: Department of Corrections.

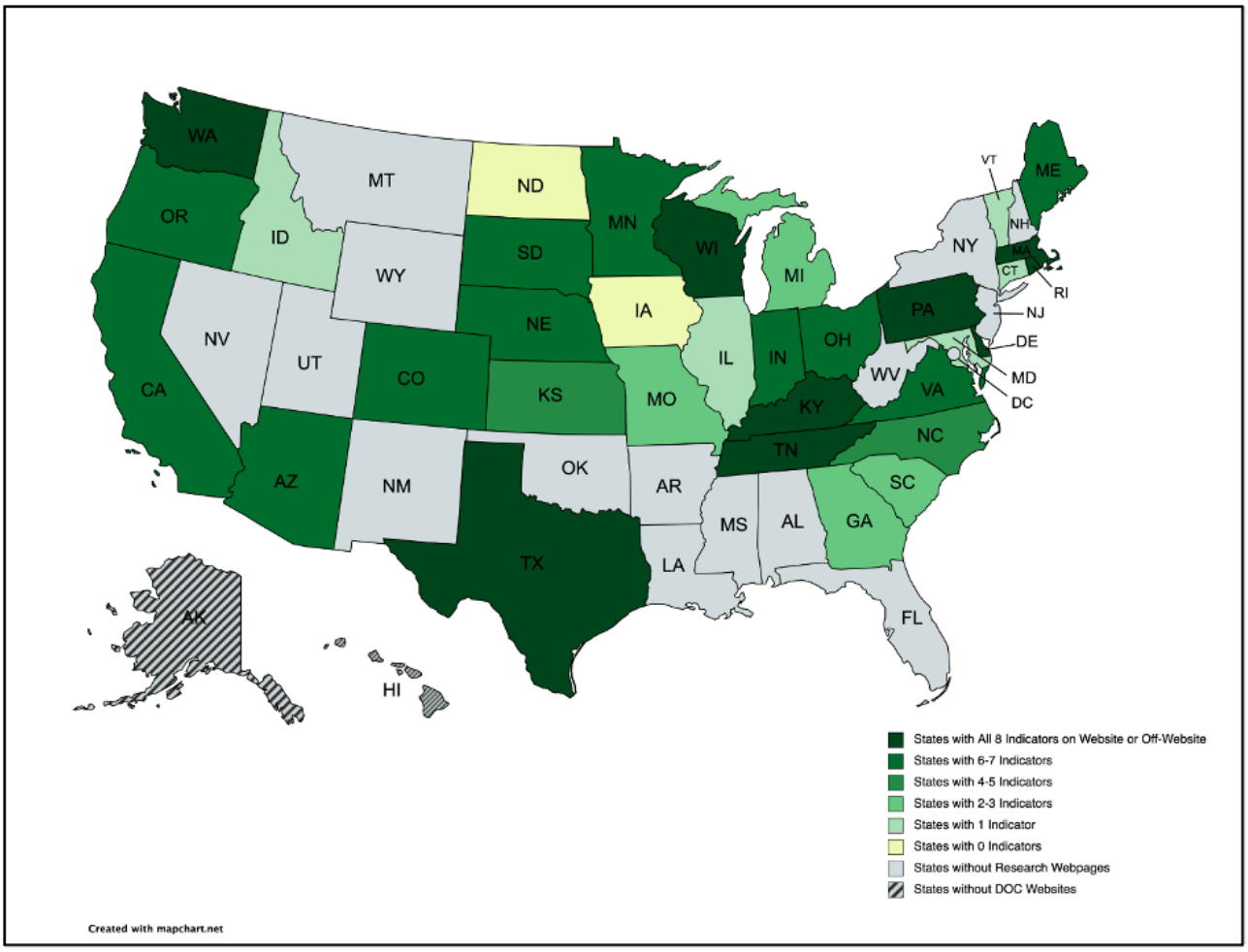

XSL•FO
RenderX

Table . Summary of the availability of websites, availability of policy, contact information, and content of policy information. A web page is referred to as such to delineate it from a website in the same statement.

			States, n/N (%)
Total states			
	States with a DOC[a]-specific[a] website		48/50 (96)
	Sates with a research web page on the website		33/50 (66)
		If no research web page, research resources found through Google search in the first 25 results	10/50 (20)
		No research web page or easily found information from Google searches	5/50 (10)
Subanalysis of research policies on research web pages			
	Availability of research policies on research web pages		
		Research policies downloadable from web pages	29/33 (88)
		Research policies/procedures directly available on web pages without download	7/33 (21)
		Reference to research policies but not linked to web pages/required separate searches	1/33 (3)
		No reference to policies or procedures	3/33 (9)
		Contact email/phone available	25/33 (76)
	Content of policies on research web pages (available both directly and/or downloadable)		
		Recruitment of staff	22/29 (76)
		Consent of staff	24/29 (83)
		Payment of staff	14/29 (48)
		Recruitment of people who are incarcerated	23/29 (79)
		Consent of people who are incarcerated	24/29 (83)
		Payment of people who are incarcerated	18/29 (62)
		External institutional review board requirements	20/29 (69)

[a]DOC: Department of Corrections.

Ethical Considerations

This qualitative study was conducted on publicly available data and thus was deemed nonhuman subject research by the Tufts University Health Sciences IRB.

Results

Two state DOCs' (Alaska and Hawaii) websites were unavailable; only 33 of the remaining 48 states' research web pages were available. Less than half of states had web page–accessible policies available about consent, recruitment, or use of incentives for staff or incarcerated participants (Table 1). We received a response via phone or email from only 23 of the 25 DOC web pages that provided contact information. While similar in number, the states that provided information about incarcerated participants and those that provided information about involved staff only partially overlapped, and several only provided policies for group or the other.

Discussion

We observed significant heterogeneity in the availability and content of state DOC websites and research web pages, with most DOCs having little information readily available. For instance, only 20 DOC web pages explicitly mentioned IRB requirements, and as few as 14 states had policies available describing staff payments. Barriers to accessing DOC-specific policies have impeded researchers interested in developing prison-based research questions and studies aiming to improve

XSL•FO
RenderX

conditions and health outcomes in these settings [6,7]. By extension, the limited accessibility demonstrated herein could hinder future research. Furthermore, limited access to research can be harmful, as demonstrated by HIV's impact on incarcerated populations [8]. The COVID-19 pandemic reenforced this and displayed how research partnership across carceral, academic, and public sectors are needed to mitigate harm [9,10]. Decreasing barriers and increasing the web-based availability of prison research policies is therefore a reasonable first step to support impactful and feasible research.

Our evaluation was cross-sectional, and websites may change rapidly, reflecting funding or policy, as evidenced by current DOC websites for Alaska and Hawaii that were not available for analysis in 2023. Repeat analysis may reveal changes to other DOC websites as well. Furthermore, public-facing information on prison-based research does not necessarily translate into actual collaboration between researchers and state DOCs, and assessments of virtual information do not capture existing relationships among researchers, prison-based populations, and DOCs. Further analysis should focus on the cascade of research projects (eg, submitted, approved, and published) in each state to better understand how this accessibility translates to research partnerships. Conceivably, increasing the digital availability of DOCs' research policies will facilitate research initiation and partnerships to improve health outcomes for incarcerated individuals.

Acknowledgments

This study received funding from the Tufts University School of Medicine Harold Williams, M.D. Summer Research Fellowship. The funders approved the design and conduct of this study and did not administer any of the collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication. Generative artificial intelligence was not used in any portion of manuscript writing.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: AW
Data curation: JM-K and DM
Formal analysis: JM-K and DM
Investigation: JM-K and DM
Methodology: AW, JM-K, DM, LB-R, NV, and JG
Supervision: AW
Visualization: JM-K and DM
Writing—original draft: AW, JM-K, DM, LB-R, NV, and JG
Writing—review and editing: AW, JM-K, DM, LB-R, NV, and JG

Conflicts of Interest

None declared.

References

1. Brinkley-Rubinstein L, Cloud DH. Mass incarceration as a social-structural driver of health inequities: a supplement to AJPH. Am J Public Health 2020 Jan;110(S1):S14-S15. [doi: 10.2105/AJPH.2019.305486] [Medline: 31967896]
2. Nosrati E, Kang-Brown J, Ash M, McKee M, Marmot M, King LP. Incarceration and mortality in the United States. SSM Popul Health 2021 Sep;15:100827. [doi: 10.1016/j.ssmph.2021.100827] [Medline: 34150979]
3. Glenn JE. Dehumanization, the symbolic gaze, and the production of biomedical knowledge. In: Black Knowledges/Black Struggles: Essays in Critical Epistemology: Liverpool University Press; 2015.
4. Qualtrics. july 2023 ed. : Qualtrics; 2025.
5. MapChart. 2023. URL: https://www.mapchart.net/index.html [accessed 2025-08-07]
6. Pascoe KJ, Walsh E, Gage BC, Van Son C. Health research in a jail: methodological challenges. Nurs Sci Q 2022 Oct;35(4):477-482. [doi: 10.1177/08943184221115129] [Medline: 36171705]
7. Neher TL, Udochi AL, Wilson KA, Guillory DM, Zaller ND, Zielinski MJ. Conducting health research in carceral systems: considerations and recommendations. Am J Public Health 2020 Jan;110(S1):S52-S55. [doi: 10.2105/AJPH.2019.305449] [Medline: 31967891]
8. Wurcel AG, Kraus C, Johnson O, et al. Stakeholder-engaged research is necessary across the criminal-legal spectrum. J Clin Transl Sci 2023;7(1):e5. [doi: 10.1017/cts.2022.501] [Medline: 36755540]
9. Wang EA, Zenilman J, Brinkley-Rubinstein L. Ethical considerations for COVID-19 vaccine trials in correctional facilities. JAMA 2020 Sep 15;324(11):1031-1032. [doi: 10.1001/jama.2020.15589] [Medline: 32808972]

XSL•FO
RenderX

10. Saloner B, Kramer C, Song M, et al. COVID-19 restrictions in jails and prisons: perspectives from carceral leaders. Health Aff (Millwood) 2023 Jun;42(6):841-848. [doi: 10.1377/hlthaff.2022.01473] [Medline: 37276483]

Abbreviations

DOC : Department of Corrections
IRB: institutional review board

Edited by N Cahill; submitted 01.05.25; peer-reviewed by JM Brenner, LB Strick; revised version received 02.07.25; accepted 03.07.25; published 20.08.25.

Please cite as:
Martinez-Kratz J, Manning D, Vest N, Glenn J, Brinkley-Rubinstein L, Wurcel A
The Digital Availability of US Departments of Corrections' Research Policies: Cross-Sectional Analysis
JMIR Form Res 2025;9:e76835
URL: https://formative.jmir.org/2025/1/e76835
doi:10.2196/76835

© Jesse Martinez-Kratz, David Manning, Noel Vest, Jason Glenn, Lauren Brinkley-Rubinstein, Alysse Wurcel. Originally published in JMIR Formative Research (https://formative.jmir.org), 20.8.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Guideline-Incorporated Large Language Model-Driven Evaluation of Medical Records Using MedCheckLLM

Marc Cicero Schubert; Stella Soyka, MD; Wolfgang Wick, MD; Varun Venkataramani, MD, PhD

[1]Department of Neurology, University Hospital Heidelberg, Im Neuenheimer Feld 400, Heidelberg, Germany

Corresponding Author:
Varun Venkataramani, MD, PhD

Abstract

Abstract: The study introduces MedCheckLLM, a large language model–driven framework that enhances medical record evaluation through a guideline-in-the-loop approach by integrating evidence-based guidelines.

(JMIR Form Res 2025;9:e53335) doi:10.2196/53335

KEYWORDS

large language models; AI; electron medical records; checklists; LLM; language model; NLP; natural language processing; records; documentation; documents; framework; conceptual; machine learning; artificial intelligence; evidence; evaluate; evaluation; guideline; health care

Introduction

Large language models (LLMs) have demonstrated enormous potential in assessing complex datasets in health care across many applications [1,2]. One underexplored area is their application for the reliable evaluation of medical documents. The automated evaluation of these documents has the potential to enhance patient safety. The system's reasoning process must be (1) transparent and comprehensible to human evaluators and (2) guided by established medical guidelines proven to increase patient safety [3].

In this study, we introduce a framework that consists of a multistep approach for medical record evaluation that incorporates guidelines into the evaluation process (ie, guideline-in-the-loop). Our proposed algorithm, MedCheckLLM, is an LLM-driven, structured reasoning mechanism designed to automate the evaluation of medical records against evidence-based guidelines. The guidelines are deterministically accessed and returned to the LLM as input without further model fine-tuning. This strict separation of LLM and guidelines is expected to increase the validity and interpretability of the evaluations. The approach's step-by-step structure could improve transparency in clinical applications. The primary objective of this research is to introduce the conceptual framework and assess its feasibility.

Methods

The MedCheckLLM algorithm begins by extracting a patient's diagnosis from the medical report (Figure 1). Based on the diagnosis, it suggests an appropriate guideline. A human medical expert makes the final guideline selection. Guidelines are then accessed independently of the LLM's mechanisms using programmatically built interfaces for guideline retrieval. Subsequently, guidelines are provided as input to the LLM and are either identified as already formatted in a usable checklist or converted into a checklist. This diagnosis-specific checklist is used to assess the medical report by the LLM, with a final verification by a human medical expert. To test this approach, we used expert-validated simulated medical reports (simulated dataset) and physician-generated medical reports (physician dataset). Performance was analyzed for patient histories with headaches using guidelines from the International Headache Society and the physician dataset for four further neurological diagnoses (ie, border zone infarction, meningitis, neuromyelitis optica, and subarachnoid haemorrhage). The validity of this method was further analyzed by evaluating generated doctor's notes with a correct diagnosis compared to doctor's notes with a false diagnosis. The LLMs, GPT-4 and Claude-3 were used for testing (see Multimedia Appendix 1).

XSL•FO
RenderX

Figure 1. Workflow of MedCheckLLM. A medical report including medical history, diagnosis and treatment is provided as input. First, the LLM identifies the given diagnosis. Second, it suggests a medical guideline for evaluation of the medical report, with a human medical expert making the final selection. Then, independently of the LLM, the selected guideline is accessed, and diagnosis-specific text is extracted and inputted into the LLM. Subsequently, the LLM determines whether the input guideline text is in checklist format; if not, it extracts a checklist. Using this diagnosis-specific checklist, the LLM evaluates the medical report based on the diagnosis-specific checklist. Finally, a human expert assesses the LLM evaluation. Dashed arrow: Checklist extraction instead of detection of checklist format. Blue box: Component uses an LLM. Green Box: Components do not use an LLM.

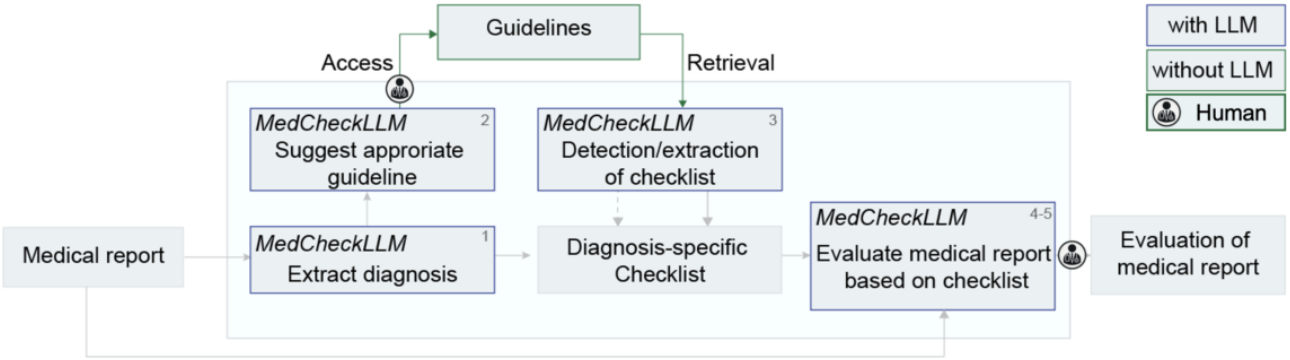

Results

We evaluated the medical report analysis conducted by MedCheckLLM for various headache diagnoses. In the simulated dataset, MedCheckLLM (based on GPT-4 and Claude-3, Table 1) extracted the specified diagnosis correctly in 100% of cases from a list of 61 possible diagnoses from The International Classification of Headache Disorders-3 [4]. The model suggested existing evidence-based guidelines in 70.59% (12/17) of medical reports and detected the format of the guidelines as checklists in 100% of the cases (N=17). MedCheckLLM accurately evaluated 87% (67/ 77) of checklist items. Performance on the physician dataset showed an accurate evaluation in 77.4% (24/ 31) of checklist items (Table 2). It identified incorrect diagnoses where the stated diagnosis did not align with the content of the doctor's letters in 94.1% (16/17) of cases, while it correctly recognized 100% (N=17) of letters with matching diagnoses.

Table . Performance of MedCheckLLM on the simulated dataset.

Elements of algorithmic structure	GPT-4 performance, % (n/N)	Claude-3 performance, % (n/N)	Explanation of specific task of each element
Extracting stated diagnosis	100 (17/17)	100 (17/17)	Extract the diagnosis that is stated in the medical report
Suggestion of existing guidelines	70.6 (12/17)	58.8 (10/17)	Suggest a guideline that should be used to evaluate the medical report
Detection of checklist	100 (17/17)	100 (17/17)	Detect whether the accessed guidelines are in a structured checklist-criteria format
Evaluation of diagnostic criteria (checklist items)	87 (67/77)	83.8 (62/74)	Assess whether the criteria listed in the checklist are met in the medical report
Evaluation of letters with correct diagnosis (clinical descriptions and diagnosis match)	100 (17/17)	94.1 (16/17)	Assess whether the diagnosis stated in the medical report aligns with the clinical descriptions
Evaluation of letters with false diagnosis (clinical descriptions and diagnosis do not match)	91.4 (16/17)	91.4 (16/17)	Evaluate whether the diagnosis that is stated in the medical report fits the clinical descriptions

XSL•FO
RenderX

Table . Performance of MedCheckLLM on the physician dataset.

Element of algorithmic structure	Stroke	Meningitis	Neuromyelitis optica	Subarachnoid hemorrhage
Extracting stated diagnosis	Yes	Yes	Yes	Yes
Suggestion of existing guidelines[a]	Yes, applicable	Yes, partially applicable	Yes, applicable	Yes, partially applicable
Creation of checklist, level of detail[a]	Yes, moderate detail	Yes, moderate detail	Yes, thorough detail	Yes, minimal detail
Evaluation of diagnostic criteria, % (n/N)	100 (7/7)	66.7 (4/6)	87.5 (7/8)	60 (6/10)

[a]The responses were classified as yes, and partially applicable,applicable, or minimal, moderate, or thorough detail.

Discussion

The framework of MedCheckLLM represents a promising approach for a comprehensive, guideline-anchored review of electronic health records. It holds the potential to function as a quality assurance framework throughout patient care due to its advantages of separate partitioning of the LLM and the guidelines, rather than training guidelines into an LLM. The flexibility of this approach allows for immediate implementation of guideline updates or the option to implement customized protocols for subgroups of patients. Due to the checklist-based approach, each item can be verified individually, thus increasing the algorithm's interpretability, which is crucial in health care settings [5]. Due to the LLM's subpar guideline suggestion capability, medical experts are integrated at this step to ensure that established guidelines are used. Further research is essential to advance the development of LLM-driven methods for extracting checklists from unstructured guidelines, as well-structured guidelines are crucial for detailed, high-quality checklists. Further, this framework facilitates improved data mining practices in electronic health records [6]. In the future, it is crucial to address privacy concerns to ensure the ethical application of these powerful tools in real-world clinical settings [7-9].

Conflicts of Interest

None declared.

Multimedia Appendix 1
Large language models used in this study.
[PDF File, 68 KB - formative_v9i1e53335_app1.pdf]

References

1. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
2. Schubert MC, Wick W, Venkataramani V. Performance of Large Language Models on a Neurology Board-Style Examination. JAMA Netw Open 2023 Dec 1;6(12):e2346721. [doi: 10.1001/jamanetworkopen.2023.46721] [Medline: 38060223]
3. Thomassen O, et al. The effects of safety checklists in medicine: a systematic review. Acta Anaesthesiol Scand 2014 Jan;58(1):5-18. [doi: 10.1111/aas.12207] [Medline: 24116973]
4. Headache Classification Committee of the International Headache Society (IHS) The International Classification of Headache Disorders, 3rd edition. Cephalalgia 2018 Jan;38(1):1-211. [doi: 10.1177/0333102417738202] [Medline: 29368949]
5. Amann J, et al. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak 2020 Nov 30;20(1):310. [doi: 10.1186/s12911-020-01332-6] [Medline: 33256715]
6. Jiang LY, Liu XC, Nejatian NP, et al. Health system-scale language models are all-purpose prediction engines. Nature New Biol 2023 Jul;619(7969):357-362. [doi: 10.1038/s41586-023-06160-y] [Medline: 37286606]
7. Meskó B. The Impact of Multimodal Large Language Models on Health Care's Future. J Med Internet Res 2023 Nov 2;25:e52865. [doi: 10.2196/52865] [Medline: 37917126]
8. Dorr DA, Adams L, Embí P. Harnessing the Promise of Artificial Intelligence Responsibly. JAMA 2023 Apr 25;329(16):1347-1348. [doi: 10.1001/jama.2023.2771] [Medline: 36972068]
9. Wang C, Liu S, Yang H, Guo J, Wu Y, Liu J. Ethical Considerations of Using ChatGPT in Health Care. J Med Internet Res 2023 Aug 11;25:e48009. [doi: 10.2196/48009] [Medline: 37566454]

XSL•FO
RenderX

Abbreviations

LLM: large language model

Edited by A Mavragani; submitted 03.10.23; peer-reviewed by D Chrimes, P Han; revised version received 02.11.24; accepted 17.11.24; published 24.04.25.

Please cite as:
Schubert MC, Soyka S, Wick W, Venkataramani V
Guideline-Incorporated Large Language Model-Driven Evaluation of Medical Records Using MedCheckLLM
JMIR Form Res 2025;9:e53335
URL: https://formative.jmir.org/2025/1/e53335
doi:10.2196/53335

© Marc Cicero Schubert, Stella Soyka, Wolfgang Wick, Varun Venkataramani. Originally published in JMIR Formative Research (https://formative.jmir.org), 24.4.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

The AI Reviewer: Evaluating AI's Role in Citation Screening for Streamlined Systematic Reviews

Jamie Ghossein[1*], MD; Brett N Hryciw[2*], MD; Tim Ramsay[3,4], PhD; Kwadwo Kyeremanteng[2,4,5], MD

[1]Interdepartmental Division of Critical Care Medicine, University of Toronto, Toronto, ON, Canada
[2]Division of Critical Care, Department of Medicine, University of Ottawa, 501 Smyth Road, Ottawa, ON, Canada
[3]Faculty of Medicine, University of Ottawa, Ottawa, ON, Canada
[4]Clinical Epidemiology, Ottawa Hospital Research Institute, Ottawa, ON, Canada
[5]Institute du Savoir Montfort, Montfort Hospital, Ottawa, ON, Canada
[*]these authors contributed equally

Corresponding Author:
Brett N Hryciw, MD
Division of Critical Care, Department of Medicine, University of Ottawa, 501 Smyth Road, Ottawa, ON, Canada

(JMIR Form Res 2025;9:e58366) doi:10.2196/58366

KEYWORDS

article screening; artificial intelligence; systematic review; AI; large language model; LLM; screening; analysis; reviewer; app; ChatGPT 3.5; chatbot; dataset; data; adoption

Introduction

Systematic reviews are regarded as one of the highest form of evidence in medical research and are vital for answering clinical questions [1]. However, the conventional systematic review methodology is time-consuming, particularly the manual screening of articles for pertinence [2]. The exponential increase in biomedical literature presents a challenge for researchers to remain updated. Artificial intelligence (AI) has shown promise in various fields [3], with large language models (LLMs) specifically offering capabilities to interpret complex text, which can be leveraged in the systematic review process [4]. We conducted a pilot feasibility study evaluating 5 distinct LLMs in an existing systematic review dataset.

Methods

Overview

We compared 5 commonly used LLMs to screen citations from a previously published systematic review on trauma hemorrhage, originally screened by two human reviewers [5]. Of the 1186 total citations, 21 (1.8%) were included for full-text review and 1165 (98.2%) were excluded. We randomly selected 100 excluded citations using Microsoft Excel. Hence, 121 citations (n=21, 17.4% included and n=100, 82.6% excluded) were tested against predefined eligibility criteria using ChatGPT 3.5 (version September 25, 2023), ChatGPT 4 (version September 25, 2023), Google Bard (version 1.15; released on September 2, 2023), Meta Llama 2 (70b parameters, version 2.1.1; released on October 10, 2023), and Claude AI 2 (version 1.3; released on July 11, 2023). We used descriptive statistics to evaluate sensitivity, specificity, and overall accuracy.

Ethical Considerations

All citations were taken from publicly available, previously published literature. No personal or patient-level data were used, and no identifiers were included. Formal research ethics board approval was therefore not required.

Results

Among the 121 total citations, the LLMs' sensitivity (correctly identifying included citations) ranged from 57% to 100%, and specificity (correctly excluding noneligible citations) ranged from 18% to 79%. ChatGPT 3.5 achieved the highest sensitivity (100%) and the highest specificity (79%). Full results are shown in Table 1.

Table . Performance metrics of large language models in citation screening for systematic reviews, including sensitivity, specificity, and accuracy.

Large language model	Sensitivity, %	Specificity, %	Accuracy, %
ChatGPT 3.5	100	79	83
ChatGPT 4	95	66	72
Google Bard	100	71	77
Meta Llama 2 (70b parameters)	95	18	34
Claude AI 2	57	77	73

XSL•FO
RenderX

Discussion

In this pilot assessment, selected LLMs demonstrated high sensitivity for identifying relevant studies, with ChatGPT 3.5 and Google Bard reaching 100%. Notably, the specificity varied widely, ranging from as low as 18% for Meta Llama 2 to 79% for ChatGPT 3.5. While some LLMs can be remarkably sensitive for screening articles within our sample, excluding irrelevant citations remains a challenge for certain LLMs. These findings suggest that AI-driven LLMs could be poised to support the screening phase, potentially replacing the second human reviewer and streamlining the often labor-intensive study screening process.

The sample size of 121 citations is a limitation, and findings may not be generalizable to other systematic reviews or inclusion and exclusion criteria. Larger studies, ideally with multiple runs of the same citations, are necessary to capture the probabilistic variability inherent to LLMs. As we only ran each citation through a given LLM once, multiple runs or "prompt engineering" strategies could yield more consistent or refined outcomes when evaluating LLMs. Nonetheless, our study offers a novel approach by directly comparing the performance of multiple LLMs, thus providing insight into how different architectures perform on the same dataset. Future research should explore repeated runs to assess LLM consistency, implement advanced prompt engineering, and investigate the explainability of LLM results.

LLMs have previously been demonstrated to effectively generate Boolean queries for a systematic review literature search [1]. As LLMs evolve further, it is conceivable that they could entirely manage the title and abstract screening. This progress can eventually lead to a fully automated review process, where AI might oversee the search strategy, title and abstract screening, full-text review, data analysis and synthesis, and even drafting and publication. Such automation would epitomize a living systematic review, ensuring evidence is continuously updated as soon as new research is published. As transparency and accountability concerns may arise, a robust ethical framework will be paramount as we navigate the advancements of this technology [6].

Conflicts of Interest

None declared.

References

1. Wang S, Scells H, Koopman B, Zuccon G. Can ChatGPT write a good boolean query for systematic review literature search? Presented at: SIGIR '23: The 46th International ACM SIGIR Conference on Research and Development in Information Retrieval; Jul 23-27, 2023; Taipei, Taiwan p. 1426-1436. [doi: 10.1145/3539618.3591703]
2. Tsafnat G, Glasziou P, Karystianis G, Coiera E. Automated screening of research studies for systematic reviews using study characteristics. Syst Rev 2018 Apr 25;7(1):64. [doi: 10.1186/s13643-018-0724-7] [Medline: 29695296]
3. Hryciw BN, Fortin Z, Ghossein J, Kyeremanteng K. Doctor-patient interactions in the age of AI: navigating innovation and expertise. Front Med (Lausanne) 2023 Aug 30;10:1241508. [doi: 10.3389/fmed.2023.1241508] [Medline: 37711734]
4. Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. JAMA 2023 Jul 3;330(1):78-80. [doi: 10.1001/jama.2023.8288] [Medline: 37318797]
5. Ghossein J, Fernando SM, Rochwerg B, Inaba K, Lampron J, Tran A. A systematic review and meta-analysis of sample size methodology for traumatic hemorrhage trials. J Trauma Acute Care Surg 2023 Jun 1;94(6):870-876. [doi: 10.1097/TA.0000000000003944] [Medline: 36879398]
6. Jha D, Durak G, Sharma V, Keles E, Cicek V, Zhang Z, et al. A conceptual algorithm for applying ethical principles of AI to medical practice. arXiv. Preprint posted online on Jan 3, 2025. [doi: 10.48550/arXiv.2304.11530]

Abbreviations

AI: artificial intelligence
LLM: large language model

Edited by A Mavragani; submitted 13.03.24; peer-reviewed by J Deng, J Mistry, Q Khraisha; revised version received 26.01.25; accepted 29.01.25; published 28.03.25.

Please cite as:
Ghossein J, Hryciw BN, Ramsay T, Kyeremanteng K
The AI Reviewer: Evaluating AI's Role in Citation Screening for Streamlined Systematic Reviews
JMIR Form Res 2025;9:e58366
URL: https://formative.jmir.org/2025/1/e58366
doi:10.2196/58366

XSL•FO
RenderX

©Jamie Ghossein, Brett N Hryciw, Tim Ramsay, Kwadwo Kyeremanteng. Originally published in JMIR Formative Research (https://formative.jmir.org), 28.3.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Engaging Stakeholders With Professional or Lived Experience to Improve Firearm Violence Lexicon Development

Nicole Cook[1], MPA, PhD; Frances M Biel[1], MPH; Kerry Ann Bet[1], MPH; Marion R Sills[1], MPH, MD; Ali Al Bataineh[2], PhD; Pedro Rivera[1], MS; Anna R Templeton[1], DNP; Natalie Cartwright[2], PhD

[1]OCHIN, PO Box 5426, Portland, OR, United States

[2]Norwich University, Northfield, VT, United States

Corresponding Author:
Nicole Cook, MPA, PhD
OCHIN, PO Box 5426, Portland, OR, United States

Abstract

Framing the public health burden of firearm violence should include people with secondary exposure to firearm violence beyond acute bodily injury, yet such data are limited. Electronic health record clinical notes, when leveraged through natural language processing (NLP), are a potential source of data on firearm exposure. As part of NLP lexicon development, diverse stakeholders were engaged to identify keywords, and our findings demonstrated that engaging diverse stakeholders adds valuable input to NLP development.

(JMIR Form Res 2025;9:e68105) doi:10.2196/68105

KEYWORDS

firearm violence; firearm injury; stakeholder engagement; natural language processing; electronic health record

Introduction

Exposure to firearm violence includes witnessing a shooting, being threatened by a firearm, losing a loved one to gun violence, and sustaining injuries from firearms [1]. Such exposure is associated with adverse physical and behavioral impacts. To better understand health sequelae following firearm violence exposure, some researchers recommend including these exposures in ongoing surveillance and research [2,3]. Although structured data (eg, diagnostic codes) for exposure to firearm violence are largely unavailable, unstructured electronic health record (EHR) data represent a potential source of information collected in the clinical care process [4]. To explore the applicability of such data in surveillance and research, we developed a natural language processing (NLP) pipeline for using unstructured EHR fields, including clinical notes, to identify those with exposure to firearm violence and subsequently understand more about the health impacts of such exposure. A crucial part of formative work on NLP lexicon development is effectively engaging a broad representation of stakeholders to contribute lexicon terms that might not otherwise be known to study teams [5]. In this research letter, we describe the process and outcomes of engaging stakeholders in NLP lexicon development.

Methods

Study Design

We assembled a stakeholder advisory committee (SAC) comprised of patients, advocates, clinicians, researchers, and others with lived and/or professional experience that includes exposure to firearm violence. The SAC structure was designed to include community advocates, patients with lived experience, practicing clinicians, clinical researchers, data scientists, and clinical informaticists. SAC members were recruited via prior collaboration with a study team member, word of mouth, and presentations to patients and clinicians at web-based workgroup meetings. The final SAC composition included 12 members—5 community advocates and/or patients with lived experience, 1 physician, 2 clinician researchers, 2 clinical informaticists, and 2 data scientists.

In April 2024, SAC members reviewed the lexicon of identified keywords that was developed by the study team [4] and informed by MacPhaul et al's [6] NLP study, which investigated firearm injury intent. Afterward, SAC members suggested additional terms from their lived or professional experience that may indicate exposure to firearm violence. The review of terms and suggestion of new terms were done via an asynchronous, interactive SAC meeting.

New terms identified by the SAC were searched across EHR clinical notes from 7,103,301 patients who were receiving primary or behavioral health care at community-based health centers in the OCHIN multistate network [7]. A maximum of

XSL•FO
RenderX

30 random notes per keyword were reviewed by a study team member to determine whether a note indicated exposure to firearm violence from a powder mechanism [8]. Descriptive statistics were used to summarize results in a table for review, discussion, and adjudication by the study team and SAC.

Ethical Considerations

The Norwich University institutional review board deemed this study exempt. Stakeholders agreed to guide this study and received compensation for participating in monthly meetings; they were not research participants. Stakeholders completed a professional services agreement, which outlined roles and responsibilities, and a nondisclosure agreement prior to joining the SAC. The EHR study data used by the study team for NLP development were a limited dataset. SAC members were not provided access to the limited dataset.

Results

The initial lexicon of identified keywords provided by the study team to the SAC included 19 terms, of which 13 were used in pilot work and 6 were additionally identified by the study team [4]. SAC members identified 35 additional keywords that were not included in the first iteration of the lexicon. Of the 35 terms, 27 had at least one mention in at least one clinical note. Of the 585 clinical notes reviewed, 5 contained 4 SAC-identified terms—*bbs*, *buckshot*, *firing*, and *metal pole*—and possibly indicated exposure to firearm violence. The study team met to perform the final adjudication for including or excluding each term. Two terms (*buckshot* and *firing*) were determined to have sufficient contextual information to indicate true exposure (Table 1). Notes that included these terms were added to the NLP testing and training dataset (Figure 1).

XSL•FO
RenderX

Table . Manual review of electronic health record (EHR) clinical notes (N=585) containing stakeholder advisory committee (SAC)–identified firearm lexicon terms from 300 ambulatory health care clinics.

SAC-identified term	Notes in EHR with keyword, n	Clinical notes reviewed[a,b], n	Notes with possible indication of exposure to firearm violence, n	Text portion reviewed by study team to determine exposure	Final determination that reviewed clinical notes indicated firearm violence exposure
buckshot	1	1	1	• "My dad had just been shot in the back, he had buckshot in him."	Yes
firing	13	13	2	• "He reports there was a firing of fire arms at his apartment complex this last weekend." • "A few days later, the same brother was apparently responsible for firing a gun into the house; the bullet traversed two rooms and came to rest very close to the patient's son."	Yes
metal pole	2	2	1	• "Pt believes she is being tracked to be murdered, pt sees others are potential threats, as people who are trying to kill her-…reports pt threatening another wielding a metal pole"	No
bbs	30	30	1	• "Cl said he had a gun and he was not going to jail. (He had a BB gun, broken, threw it in the woods)"	No

[a]All notes for terms that had less than 30 notes were reviewed. For terms with >30 notes, a random sample of notes was reviewed.

[b]Additional stakeholder advisory committee–identified terms with clinical notes that did not indicate firearm violence exposure include *Banger*, *draco*, *semiautomatic*, *gat*, *toaster*, *drill*, *stick*, *strap*, *trigger*, *heat*, *heater*, *iron*, *metal*, *nina*, *nine*, *piece*, *pole*, *rod*, *burner*, *cap*, *carrying*, *69*, and *ammunition*.

XSL•FO
RenderX

Figure 1. Diagram of SAC keyword term review for firearm violence exposure. NLP: natural language processing; SAC: stakeholder advisory committee.

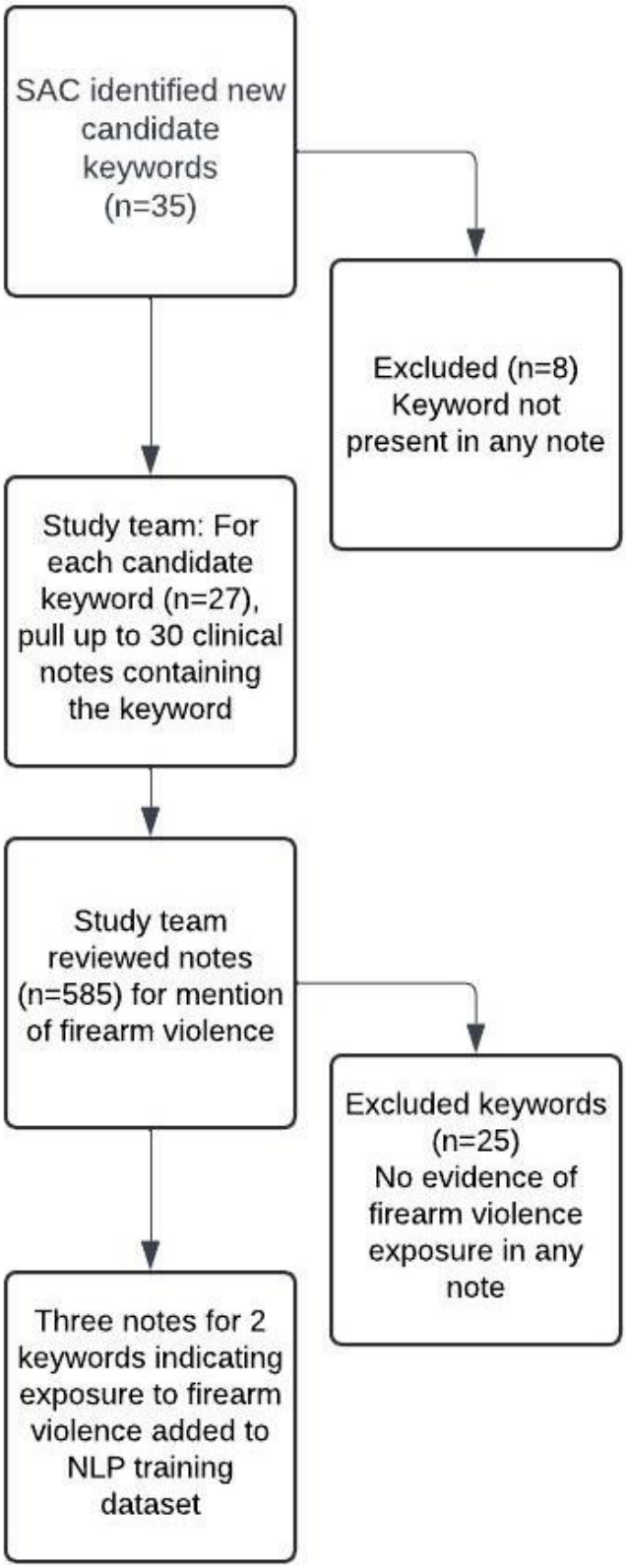

Discussion

Effectively engaging varied stakeholders in NLP lexicon development led to the identification of 2 additional terms (*buckshot* and *firing*) that were previously not considered by the study team. Novel clinical notes with these keywords were input into the training and testing dataset to enhance NLP model performance. Although most words suggested by the SAC did not indicate true exposure, the exercise demonstrated that periodic engagement of different stakeholder advisors in artificial intelligence and machine learning research is an important strategy to reduce data bias [9,10].

XSL•FO
RenderX

Acknowledgments

The authors are grateful to Angela Williams, Patricia Devine Wilder, Jennifer Obermeyer, Jelena MacLeod, and the rest of the stakeholder advisory committee for their work driving efforts toward trauma-informed care that supports collection and identification of firearm violence exposure data in patient electronic health records.

The research reported in this work was powered by PCORnet. PCORnet has been developed with funding from the Patient-Centered Outcomes Research Institute (PCORI) and conducted with the Accelerating Data Value Across a National Community Health Center Network (ADVANCE) Clinical Research Network (CRN). ADVANCE is a clinical research network in PCORnet led by OCHIN in partnership with Health Choice Network, Fenway Health, University of Washington, and Oregon Health & Science University. ADVANCE's participation in PCORnet is funded through the PCORI Award RI-OCHIN-01-MC.

This work was supported by the AIM-AHEAD Coordinating Center, funded by the National Institutes of Health (NIH). The research reported in this publication was supported by the Office of the Director, NIH Common Fund, under award number 1OT2OD032581, OTA: 1OT2OD032581. The work is solely the responsibility of the authors and does not necessarily represent the official view of AIM-AHEAD or the NIH.

Conflicts of Interest

None declared.

References

1. Schumacher S, Kirzinger A, Presiado M, Valdes I, Mollyann B. Americans' experiences with gun-related violence, injuries, and deaths. KFF. 2023 Apr 11. URL: https://www.kff.org/other/poll-finding/americans-experiences-with-gun-related-violence-injuries-and-deaths/ [accessed 2025-04-07]
2. Kaufman EJ, Delgado MK. The epidemiology of firearm injuries in the US: the need for comprehensive, real-time, actionable data. JAMA 2022 Sep 27;328(12):1177-1178. [doi: 10.1001/jama.2022.16894] [Medline: 36166012]
3. Cook N, Sills M. Tracking all injuries from firearms in the US. JAMA 2023 Feb 14;329(6):514. [doi: 10.1001/jama.2022.21997] [Medline: 36786795]
4. Cook N, Biel FM, Cartwright N, Hoopes M, Al Bataineh A, Rivera P. Assessing the use of unstructured electronic health record data to identify exposure to firearm violence. JAMIA Open 2024 Nov 4;7(4):ooae120. [doi: 10.1093/jamiaopen/ooae120] [Medline: 39498385]
5. Wan R, Kim J, Kang D. Everyone's voice matters: quantifying annotation disagreement using demographic information. In: Proceedings of the 37th AAAI Conference on Artificial Intelligence: Association for the Advancement of Artificial Intelligence; 2023:14523-14530. [doi: 10.1609/aaai.v37i12.26698]
6. MacPhaul E, Zhou L, Mooney SJ, et al. Classifying firearm injury intent in electronic hospital records using natural language processing. JAMA Netw Open 2023 Apr 3;6(4):e235870. [doi: 10.1001/jamanetworkopen.2023.5870] [Medline: 37022685]
7. Health IT and EHR solutions - community health care. OCHIN. URL: https://ochin.org/ [accessed 2025-04-07]
8. Pulcini CD, Goyal MK, De Souza HG, et al. A firearm violence research methodologic pitfall to avoid. Acad Emerg Med 2022 Sep;29(9):1140-1145. [doi: 10.1111/acem.14491] [Medline: 35332976]
9. Zou J, Schiebinger L. Ensuring that biomedical AI benefits diverse populations. EBioMedicine 2021 May;67:103358. [doi: 10.1016/j.ebiom.2021.103358] [Medline: 33962897]
10. Vishwanatha JK, Christian A, Sambamoorthi U, Thompson EL, Stinson K, Syed TA. Community perspectives on AI/ML and health equity: AIM-AHEAD nationwide stakeholder listening sessions. PLOS Digit Health 2023 Jun 30;2(6):e0000288. [doi: 10.1371/journal.pdig.0000288] [Medline: 37390116]

Abbreviations

EHR: electronic health record
NLP: natural language processing
SAC: stakeholder advisory committee

XSL•FO
RenderX

Edited by A Mavragani; submitted 29.10.24; peer-reviewed by M Elbattah, R Khoury; revised version received 18.03.25; accepted 19.03.25; published 21.04.25.

Please cite as:
Cook N, Biel FM, Bet KA, Sills MR, Al Bataineh A, Rivera P, Templeton AR, Cartwright N
Engaging Stakeholders With Professional or Lived Experience to Improve Firearm Violence Lexicon Development
JMIR Form Res 2025;9:e68105
URL: https://formative.jmir.org/2025/1/e68105
doi:10.2196/68105

© Nicole Cook, Frances M Biel, Kerry Ann Bet, Marion R Sills, Ali Al Bataineh, Pedro Rivera, Anna R Templeton, Natalie Cartwright. Originally published in JMIR Formative Research (https://formative.jmir.org), 21.4.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Incidence, Prognostic Factors, and Treatment Impact on Survival in Natural Killer/T-Cell Lymphoma: Population-Based Study in the United States

Yi Zhao[1*], MD, PhD; Yi Zhong[1*], MD; Mengqi Xiong[1], MD; Linlin Huang[1], MD; Xiujin Ye[1,2], MD, PhD; Jingsong He[1,2], MD, PhD

[1]First Affiliated Hospital Zhejiang University, No.79 Qingchun Road, Hangzhou, China

[2]Institute of Hematology, Zhejiang University, Hangzhou, China

[*]these authors contributed equally

Corresponding Author:
Jingsong He, MD, PhD
First Affiliated Hospital Zhejiang University, No.79 Qingchun Road, Hangzhou, China

Abstract

Background: Natural killer/T-cell lymphoma (NKTL) is a rare malignancy of mature natural killer/T-cells, predominantly found in Asian and South/Central American populations, with limited studies conducted in Europe and the United States.

Objective: The aim of this study is to present an overview of the incidence rate, demographic and clinical characteristics, treatment options, overall survival (OS), and factors influencing OS of NKTL in the United States.

Methods: We used data from the Surveillance, Epidemiology, and End Results 17 database to analyze NKTL cases recorded between 2000 and 2020. In a cohort of 1162 patients with NKTL, we calculated the incidence rates and performed statistical analyses to evaluate OS, the effect of radiotherapy and chemotherapy on survival, and lymphoma-specific survival.

Results: The mean annual incidence rate of NKTL in the United States was 0.067 per 100,000, with higher rates observed in men compared to women, and an increase noted with age. However, there has been no significant rise in incidence over recent years. Significant racial disparities were observed, with higher incidence rates in non-Hispanic Asian or Pacific Islanders and Hispanic people. The median survival time for patients with NKTL was 21 months, with a 5-year OS rate of 39.5%, which has shown improvement in recent years. Key independent prognostic factors impacting patient survival included age at diagnosis, clinical stage, nasal type presentation, presence of systemic symptoms, and treatment modality. Patients receiving combined radiotherapy and chemotherapy exhibited the best outcomes, with a median OS of 138 months and a 5-year OS rate of 58%. This survival benefit remained consistent even in patients with stage I/localized nasal type lymphoma, achieving a 5-year OS rate of 73.3%.

Conclusions: The incidence of NKTL has remained stable in recent years. Patients with the nasal type generally experience better survival outcomes. The use of combined radiotherapy and chemotherapy appears to enhance survival, though further validation through prospective multicenter clinical trials is necessary.

(JMIR Form Res 2025;9:e70129) doi:10.2196/70129

KEYWORDS

NK/T-cell lymphoma; incidence; prognostic factors; chemoradiotherapy; SEER

Introduction

Natural killer (NK)/T-cell lymphoma (NKTL) is an aggressive form of non-Hodgkin lymphoma that originates from mature T-cells and NK cells. It is distinct from other peripheral T-cell lymphomas, primarily affecting the nasal cavity and nasopharynx, and is closely associated with Epstein-Barr virus (EBV) infection [1]. Previously classified in the 4th and earlier editions of the *WHO Classification of Haematolymphoid Tumours* as “NK/T-cell lymphoma, nasal-type,” the latest (5th) edition has dropped the qualifier “nasal-type” to acknowledge that the disease can also manifest in other locations, such as the lungs, gastrointestinal tract, and skin [1-3]. Due to rapid local progression and distant metastasis, NKTL is typically characterized by a poor prognosis [4], with the extranasal type being particularly aggressive [5]. The clinical management of this disease remains highly challenging. National Comprehensive Cancer Network guidelines recommend radiotherapy (RT) in the first-line treatment of early-stage extranodal NK/T-cell lymphoma (ENKL). For advanced-stage nasal type and the extranasal type with a poor prognosis, the primary treatment is chemotherapy.

XSL•FO
RenderX

NKTL constitutes approximately 10% of all mature T/NK cell lymphomas globally, with significantly higher prevalence rates in Asia and Central or South America [5,6]. Data from the International Peripheral T-cell Lymphomas database indicate that out of 1153 peripheral T-cell lymphoma cases, ENKL represents 11.8%, with a much higher relative frequency in Asian countries compared to Western nations (22% vs 5%) [5]. Studies have shown that variations in EBV strains (eg, LMP1 30-base-pair deletion) and genetic profiles (eg, STAT3 mutations) may differentially influence the pathogenesis mechanisms across different ethnic groups [7,8]. Most published literature on NKTL consists of case reports, with larger cohort studies mainly originating from China and other Asian regions [1-3,5,9,10]. In contrast, data on NKTL in Western countries remain scarce [11,12]. Using the Surveillance, Epidemiology, and End Results (SEER) database, this study investigates the incidence, demographic and clinical characteristics, treatment options, patient survival, and prognostic factors of NKTL in the United States, with a particular focus on the impact of radiotherapy and chemotherapy on survival, aiming to provide a comprehensive overview of NKTL management and outcomes in this region.

Methods

Data Source, Patient Inclusion Criteria, and Study Objectives

This study conducted a retrospective analysis using cohort data from the National Cancer Institute's SEER 17 registries [13]. The patient cohort was accessed via the SEER website [13] using the SEER*Stat software (version 8.4.3; National Cancer Institute). The selected dataset was "SEER Database: Incidence - SEER Research Data, 17 Registries, Nov 2022 Sub (2000 - 2020)," which was released in April 2023, based on submissions from November 2022. The dataset encompasses population-based patients with cancer from various US regions including San Francisco-Oakland Standard Metropolitan Statistical Area, Connecticut, Hawaii, Iowa, New Mexico, Seattle (Puget Sound), Utah, Atlanta (Metropolitan), San Jose-Monterey, Los Angeles, Alaska Natives, Rural Georgia, California excluding San Francisco-Oakland Standard Metropolitan Statistical Area/San Jose-Monterey/Los Angeles, Kentucky, Louisiana, New Jersey, and Greater Georgia, covering approximately 26.5% of the US population (based on the 2020 census).

Patient inclusion criteria were as follows: (1) diagnosis of NK/T-cell lymphoma, nasal-type, classified under the International Classification of Diseases for Oncology, Third Edition (ICD-O-3) code 9719/3; (2) diagnosis made between the years 2000 and 2020; and (3) diagnosis confirmed by positive histopathology. Patients were excluded if they had missing survival information, nonhistological diagnostic confirmation, or were diagnosed solely by autopsy or death certificate.

The objective of this study was to analyze the incidence rate (IR), demographic and clinical characteristics, treatment modalities, survival outcomes, and factors influencing the prognosis of patients with NKTL across 17 US regions from 2000 to 2020. This comprehensive overview aims to enhance our understanding of the diagnosis, treatment, and survival of this rare disease in the United States, particularly evaluating the impact of radiotherapy and chemotherapy on the survival of patients with NKTL.

Patient Cohort Variables

The variables collected for each patient in this study included the year of diagnosis, gender, race, age at diagnosis (up to ≥90 y), marital status, median household income, and area of residence. Patients were stratified by the year of diagnosis into 4 groups (2000 - 2005, and subsequent 5-year intervals), by age into 3 groups (≤39 y, 40 - 69 y, and ≥70 y), and by median household income at diagnosis into 3 groups (≥US $70,000, US $50,000 - 69,999, and <US $50,000). Racial categories comprised non-Hispanic White (NHW), non-Hispanic Black (NHB), Hispanic, non-Hispanic Asian or Pacific Islander (NHAPI), and non-Hispanic American Indian/Alaska Native (NHAIAN). Marital status was categorized as married or unmarried, with unmarried encompassing separated, widowed, divorced, single, unmarried, or domestic partner. The area of residence was divided into counties in metropolitan areas and nonmetropolitan counties.

Patient disease characteristics included the site of tumor involvement, clinical stage, presence of systemic symptoms (referred to as B symptoms), whether NKTL was the first tumor, as well as treatments received, such as radiotherapy and chemotherapy, and survival outcomes. Sites of tumor involvement were classified as nodal or extranodal, with further details on specific sites. Staging was based on the Ann Arbor Stage and Combined Summary Stage, grouping patients into Ann Arbor stage I (localized), stage II (regional), and stages III or IV (distant). Systemic symptoms were identified through the presence of fever, night sweats, or weight loss, classified under the variable B symptoms recode (B symptoms). Treatment modalities included radiotherapy, categorized by type as beam radiation, radiation NOS (ie, method or source not otherwise specified), and radioactive implants (including brachytherapy), with patients classified as receiving radiotherapy (yes) or not (no/unknown). Chemotherapy was recorded as yes or no/unknown. Survival data encompassed "vital status" and "survival months." In cases where NKTL was the primary tumor, lymphoma-specific deaths were analyzed using the "SEER cause-specific death classification."

Ethical Considerations

The SEER database is a public cancer database, which is authorized by the National Cancer Institute, National Institutes of Health, US Department of Health and Human Services, and USA.gov. The data in the SEER database is deidentified and publicly available, containing no personal identifying information. According to the exemption criteria outlined in 45 CFR 46.104(d)(4) published by the US Department of Health and Human Services, the use of deidentified public data does not require institutional review board approval. In our study, we have been granted permission to use the SEER data and this analysis was exempt from institutional review board review.

XSL•FO
RenderX

Statistical Analysis

Incidence rates (IR; cases per 100,000) were calculated using data from 2000 to 2020 obtained from the SEER database and were age-adjusted to the US standard population. These calculations were performed using the SEER*Stat 8.4.3 software, which provided the rate and the corresponding 95% CI. Counts and percentages were used for categorical variables, while medians and interquartile ranges (IQR) were used for continuous variables to illustrate baseline patient characteristics. The Pearson χ^2 test and the Wilcoxon rank-sum test were used for comparing categorical and continuous variables between groups, respectively. Overall survival (OS) was analyzed up to 2020 using the Kaplan-Meier estimator, and survival outcomes were compared using the log-rank test. Both univariate and multivariate Cox proportional hazards regression models were applied to identify predictors that may influence OS, presenting hazard ratios (HRs) and 95% CIs. For patients with NKTL as the initial tumor, lymphoma-specific OS was also analyzed. Deaths attributable to NKTL were treated as events, while all other deaths or survivals were considered censored. A *P* value of less than .05 was considered statistically significant. Statistical analyses and graphical representations were conducted using IBM SPSS (version 26.0; IBM Corp) and the R programming language (version 4.4.0; R Foundation for Statistical Computing).

Results

Incidence Rate

From 2000 to 2020, the mean annual IR of NKTL was 0.067 per 100,000 (95% CI 0.063 - 0.071). The lowest IR occurred in 2000, at 0.014 per 100,000 (95% CI 0.007 - 0.025), and the highest in 2010, at 0.086 per 100,000 (95% CI 0.067 - 0.109). The IR among males was significantly higher at 0.093 per 100,000 (95% CI 0.086 - 0.100) compared to females at 0.044 per 100,000 (95% CI 0.040 - 0.049; *P*<.05), resulting in a male-to-female ratio of 2.11:1. Figure 1A displays the annual IRs for males and females from 2000 to 2020. IRs also showed a gradual increase with age, as depicted in Figure 1B. There were significant differences in IRs across the age groups: ≤39 years at 0.031 per 100,000 (95% CI 0.028 - 0.035), 40 - 69 years at 0.100 per 100,000 (95% CI 0.093 - 0.109), and ≥70 years at 0.165 per 100,000 (95% CI 0.145 - 0.187; *P*<.05) (Figure 1C). When analyzing IRs by race, NHAPI had the highest rate at 0.157 per 100,000 (95% CI 0.139 - 0.177), similar to Hispanic people at 0.132 per 100,000 (95% CI 0.119 - 0.146) and NHAIAN at 0.097 per 100,000 (95% CI 0.052 - 0.166). However, these rates were significantly higher than those of NHW and NHB, which were 0.036 per 100,000 (95% CI 0.033 - 0.040) and 0.030 per 100,000 (95% CI 0.022 - 0.039), respectively, with no significant difference noted between NHW and NHB (*P*<.05) (Figure 1D).

Figure 1. Annual incidence of NKTL in 17 regions of the United States between 2000 and 2020. (**A**) Annual incidence rates for males and females, with lighter shading showing 95% CI. (**B**) Incidence rates for populations of different ages. (**C**) Incidence rates for populations in age groups ≤39 years, 40-69 years, and ≥70 years. (**D**) Incidence rates for different races. NHAIAN: non-Hispanic American Indian/Alaska Native; NHAPI: non-Hispanic Asian or Pacific Islander; NHB: non-Hispanic Black; NHW: non-Hispanic White; NKTL: natural killer/T-cell lymphoma.

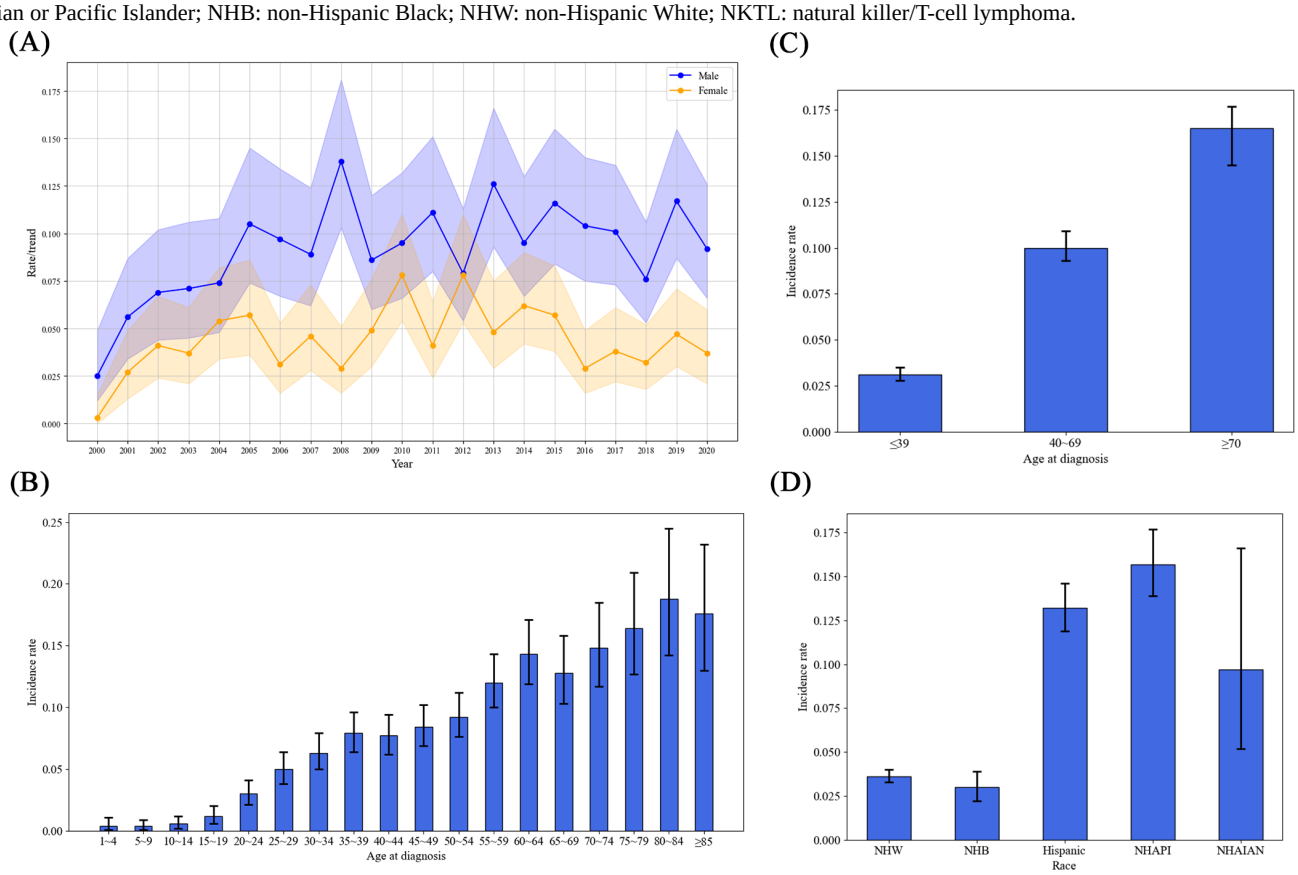

XSL•FO
RenderX

Demographic and Clinical Characteristics of Patients With NKTL

The study encompassed 1162 participants. Details regarding the demographic and clinical characteristics of the participants are presented in Table 1. Of these, 764 patients (65.7%) were male, resulting in a male-to-female ratio of 1.92, and the median age was 54 (IQR 49 - 72) years. The racial demographic was predominated by Hispanic participants (36.5%), followed by NHW (34.6%) and NHAPI (22.9%). The median ages varied significantly across racial groups (P<.001): 60.5 years for NHW (IQR 49 - 72 y), 51.0 years for NHB (IQR 39 - 60 y), 43.0 years for Hispanic people (IQR 31 - 56.8 y), 59.0 years for NHAPI (IQR 46.8 - 71.0 y), and 44.0 years for other races (IQR 25.5 - 57.5 y). A majority of the patients (91.1%) resided in metropolitan areas, 52.7% reported household incomes above US $70,000, and 57.6% were married.

In our cohort, a significant majority of the patients (1016 cases, 87.4%) experienced primary invasion of extranodal tissues. The most frequently affected extranodal site was the nasal cavity (581 cases, 50%), followed by the oropharynx, nasopharynx, pharynx, and throat (164 cases, 14.1%), with the paranasal sinus being the third most common site (88 cases, 7.6%). Patients with invasions in these 3 primary locations, totaling 833 cases (71.7%), were classified as having the nasal type. Other cases involving the gastrointestinal tract, lungs, bone marrow, and skin soft tissues were categorized as extranasal type. Based on the Ann Arbor staging system, when combined with the summary stage, patients were classified as follows: stage I (localized) comprised 490 patients (37.8%), stage II (regional) included 231 patients (17.8%), and stages III and IV (distant) accounted for 489 patients (37.7%). Patients diagnosed after 2010 had records of B symptoms, with 201 (17.4%) exhibiting these symptoms. Regarding treatment, 829 patients (71.3%) underwent chemotherapy, 615 (42.3%) received radiotherapy, and 477 (41%) were treated with a combination of radiotherapy and chemotherapy.

XSL•FO
RenderX

Table . Demographic and clinical characteristics of patients with NKTL (United States; 2000 - 2020; N=1162).

Variable	Number of participants (%)
Year of diagnosis	
2000-2005	226 (19.4)
2006-2010	294 (25.3)
2011-2015	338 (29.1)
2016-2020	304 (26.2)
Gender	
Female	398 (34.3)
Male	764 (65.7)
Age group (years)	
≤39	294 (25.3)
40-69	624 (53.7)
≥70	244 (21)
Race	
NHW[a]	402 (34.6)
NHB[b]	53 (4.6)
Hispanic	424 (36.5)
NHAPI[c]	266 (22.9)
NHAIAN[d]	14 (1.2)
Unknown	3 (0.3)
Marital status at diagnosis	
Married	669 (57.6)
Unmarried[e]	429 (36.9)
Unknown	64 (5.5)
Median household income (US $)	
>70,000	612 (52.7)
50,000-69,999	483 (41.6)
<50,000	67 (5.8)
Rural-urban continuum	
Counties in metropolitan areas[f]	1067 (91.1)
Nonmetropolitan counties[g]	91 (7.9)
Unknown	4 (0.3)
Sites of tumor involvement	
Nodal or extranodal	
Nodal	146 (12.6)
Extranodal	1016 (87.4)
Specific sites	
Nasal cavity	581 (50)
Paranasal sinus	88 (7.6)
Oropharynx, nasopharynx, pharynx, and throat	164 (14.1)
Gastrointestinal tract	44 (3.8)
Respiratory tract, pleura, and mediastinum	18 (1.5)

XSL•FO
RenderX

Variable	Number of participants (%)
Bone, bone marrow	10 (0.9)
Lymph nodes, spleen, and Waldeyer's ring	146 (12.6)
Skin, soft tissue	62 (5.3)
Others[h]	36 (3.1)
Unknown[i]	13 (1.1)
Stage	
Stage I/localized	492 (42.3)
Stage II/regional	231 (19.9)
Stage III or IV/distant	373 (32.1)
Unknown	66 (5.7)
B symptoms	
No	383 (33)
Yes	201 (17.3)
Unknown	578 (49.7)
First malignant or not	
Yes	1049 (90.3)
No	113 (9.7)
Radiation	
Yes[j]	615 (42.3)
No/unknown[k]	547 (47.1)
Chemotherapy	
Yes	829 (71.3)
No/unknown	333 (28.7)
Survival status	
Survival	460 (39.6)
Death	702 (60.4)

[a]NHW: non-Hispanic White.

[b]NHB: non-Hispanic Black.

[c]NHAPI: non-Hispanic Asian or Pacific Islander.

[d]NHAIAN: non-Hispanic American Indian/Alaska Native.

[e]Single (never married), widowed, divorced, separated, unmarried, or domestic partner.

[f]Counties in metropolitan areas of 1 million population, counties in metropolitan areas of 250,000 to 1 million population, and counties in metropolitan areas of 250,000 population.

[g]Nonmetropolitan counties adjacent to a metropolitan area, and nonmetropolitan counties not adjacent to a metropolitan area.

[h]Mammary gland, male/female reproductive system, kidney, lacrimal gland, orbit, central nervous system, thyroid gland, adrenal gland.

[i]Head, face or neck, not otherwise specified and unknown.

[j]Beam radiation, radiation, not otherwise specified method, source not specified, or radioactive implants.

[k]Refused (these data have been recorded by SEER since 1988), recommended, unknown if administered, and none/unknown.

Survival Analysis

By the end of the follow-up period in December 2020, the median follow-up duration was 12.5 months (IQR 3.0 - 59.0 mo), with a total of 702 deaths recorded, representing 60.4% of the cohort. The median survival time was 21.0 months (95% CI 14.5 - 27.5). The OS rates were 56.3% at 12 months, 44.5% at 35 months, and 39.5% at 60 months, as shown in Figure 2A. Survival analyses for different subgroups of patients are detailed in Table 2 and Table S1 in Multimedia Appendix 1. When grouped by year of diagnosis, there was a significant improvement in survival over the years (P<.001). The 5-year OS rates for patients diagnosed in the periods 2000 - 2005, 2006 - 2010, and 2011 - 2015 were 31.5%, 37.5%, and 41.0%, respectively (Figure 2B). Age also played a crucial role in influencing survival outcomes. Compared to patients aged ≤39

XSL•FO
RenderX

years and those aged 40 - 69 years, older patients (≥70 y) exhibited markedly poorer survival, with a 5-year OS of only 26.3% (*P*<.001 for both comparisons); the 5-year OS rates for the younger age groups were 48.7% and 40.5%, respectively; there was no significant difference observed between them (*P*=.07), as illustrated in Figure 2C. No significant differences in survival were observed based on gender or race (*P*=.61 and *P*=.44, respectively). Additionally, factors such as marital status and area of residence at the time of diagnosis did not significantly impact survival outcomes. However, household income at diagnosis appeared to influence survival to some extent (*P*=.04).

Survival was significantly better in patients diagnosed with NKTL as their initial tumor compared to those who had other types of tumors previously. The median OS for patients with NKTL as their first tumor was 26.0 months (95% CI 18.1 - 33.9 mo), whereas it was only 11.0 months for patients with prior tumors (95% CI 6.1 - 15.9 mo; *P*<.001). Further analysis revealed a notable age difference between the two groups: the median age for patients with NKTL as their first tumor was 52 years, compared to 71 years for those with nonfirst NKTL, indicating a significant disparity (*P*<.001). This variation in survival rates may be influenced by the patients' ages.

Patients with localized disease exhibited significantly better OS than those with regional or distant disease (*P*<.001). There was also a marked survival difference between localized, regional, and distant stages (*P*<.001), with 5-year OS rates of 55.3%, 39.8%, and 19.6%, respectively (Figure 2D). Furthermore, patients with primary extranodal invasion had significantly better survival compared to those with nodal invasion (*P*<.001), featuring a median OS of 27.0 months (95% CI 17.9 - 36.1 mo) versus 6.0 months (95% CI 4.1 - 7.9 mo). Patients diagnosed with the nasal type of the disease showed significantly better survival outcomes than those with the extranasal type, as detailed in Table 2, Table 1S in Multimedia Appendix 1, and Figure 2E. The 3-year and 5-year OS rates for nasal type were 52.5% and 46.4%, respectively, whereas for extranasal type, they were substantially lower at 24.4% and 22.2%, respectively.

Figure 2. OS for different subgroups of patients with NKTL in 17 regions of the United States between 2000 and 2020. (**A**) OS of 1162 patients with NKTL. (**B**) OS of patients based on different diagnostic years. (**C**) OS of patients with different age groups. (**D**) OS of patients based on different clinical stages. (**E**) OS of patients with nasal type and extranasal type. (**F**) OS of patients based on different treatment modalities. NKTL: natural killer/T-cell lymphoma; OS: overall survival.

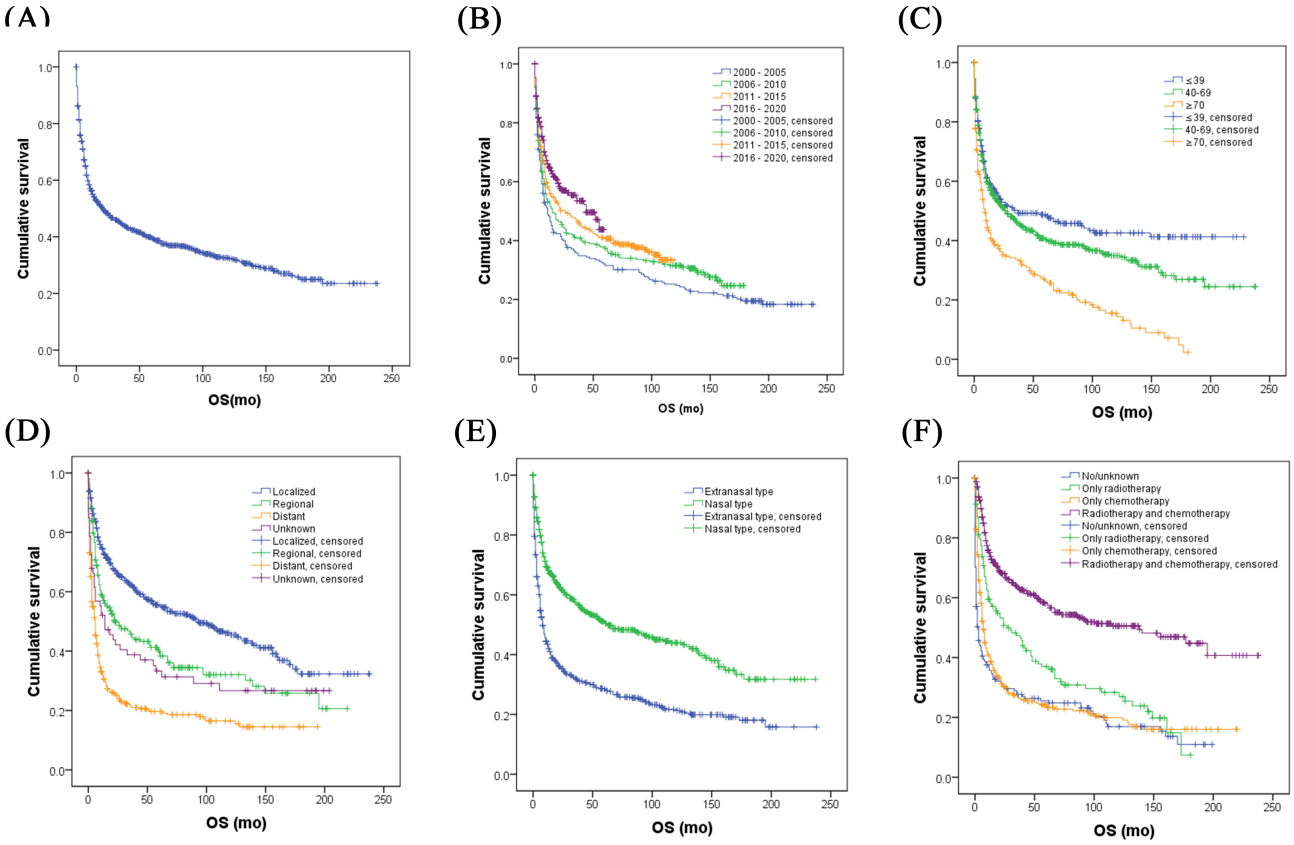

XSL•FO
RenderX

Table . Survival analysis for different subgroups of patients with natural killer/T-cell lymphoma (United States; 2000 - 2020; N=1162).

Variable	Median overall survival	1-year overall survival rate, %	3-year overall survival rate, %	*P* value
Year of diagnosis				<.001
2000-2005	11.0 (6.7 - 15.3)	48.3	35.8	
2006-2010	15.0 (7.4 - 22.6)	52.9	40.8	
2011-2015	26.0 (12.4 - 39.7)	57.2	46.7	
2016-2020	44.0 (28.4 - 59.6)	65.1	53.5	
Gender				.61
Female	20.0 (8.8 - 31.2)	56.7	45.6	
Male	22.0 (14.6 - 29.4)	56.1	43.9	
Age (years)				<.001
≤39	37.0 (0 - 79.8)	60.9	50.2	
40-69	27.0 (17.5 - 36.5)	59.1	46	
≥70	9.0 (6.5 - 11.5)	43.5	33.7	
Race (n=1159)				.44
NHW[a]	18.0 (8.3 - 27.7)	54.2	43.9	
NHB[b]	22.0 (0 - 45.5)	56.1	40	
Hispanic	23.0 (11.4 - 34.6)	58.4	45.5	
NHAPI[c]	22.0 (7.9 - 36.1)	57.4	45.5	
NHAIAN[d]	4.0 (0 - 13.2)	35.7	28.6	
Marital status at diagnosis (n=1098)				.98
Married	24.0 (14.8 - 33.2)	56.7	45.8	
Unmarried[e]	18.0 (9.6 - 26.4)	55.5	43.4	
Unknown	21.0 (5.8 - 36.2)	56.4	42	
Median household income (US $)				.04
>70,000	27.0 (15.6 - 38.4)	59.3	46.5	
50,000-69,999	18.0 (10.9 - 25.1)	53.6	42.5	
<50,000	12.0 (3.1 - 20.9)	48.8	41	
Rural-urban continuum (n=1158)				.16
Counties in metropolitan areas[f]	22.0 (15.5 - 28.5)	56.9	44.7	
Nonmetropolitan counties[g]	13.0 (0 - 29.4)	50.7	42.6	
First malignant or not				<.001
Yes	26.0 (18.1 - 33.9)	57.2	46.1	
No	11.0 (6.1 - 15.9)	48	30.1	
Clinical stage				<.001
Stage I/localized	93.0 (61.2 - 124.8)	74.4	62.4	
Stage II/regional	23.0 (9.6 - 36.4)	58.5	45.7	
Stage III or IV/distant	6.0 (4.8 - 7.2)	31.8	21	
Unknown	14.0 (2.2 - 25.8)	52.1	38.8	
Nodal or extranodal				<.001
Nodal	6.0 (4.1 - 7.9)	35	23.3	

XSL•FO
RenderX

Variable	Median overall survival	1-year overall survival rate, %	3-year overall survival rate, %	*P* value
Extranodal	27.0 (17.9 - 36.1)	59.3	47.5	
Nasal type (n=1149)				<.001
No	6.0 (4.5 - 7.5)	36.5	24.4	
Yes	44.0 (30.8 - 57.2)	64	52.5	
B symptoms				<.001
No	67.0 (40.0 - 94.0)	69.4	57.6	
Yes	11.0 (3.4 - 18.6)	49.1	41	
Unknown	13.0 (9.5 - 16.5)	50.2	37.6	
Radiotherapy				<.001
Yes[h]	67.0 (38.3 - 95.7)	71.6	59.8	
No/unknown[i]	6.0 (4.7 - 7.3)	38.6	26.9	
Chemotherapy				<.001
Yes	30.0 (18.2 - 41.8)	60.5	47.9	
No/unknown	9.0 (5.2 - 12.8)	45.7	36	
Treatment				<.001
No/unknown	3.0 (1.4 - 4.6)	35.8	27.7	
Only radiation	27.0 (10.7 - 43.3)	59.4	47.4	
Only chemotherapy	7.0 (5.6 - 8.4)	40.2	26.6	
Radiation and chemotherapy	138.0 (76.0 - 200.0)	75.2	63.5	

[a]NHW: non-Hispanic White.

[b]NHB: non-Hispanic Black.

[c]NHAPI: non-Hispanic Asian or Pacific Islander.

[d]NHAIAN: non-Hispanic American Indian/Alaska Native.

[e]Single (never married), widowed, divorced, separated, unmarried, or domestic partner.

[f]Counties in metropolitan areas of 1 million population, counties in metropolitan areas of 250,000 to 1 million population, and counties in metropolitan areas of 250,000 population.

[g]Nonmetropolitan counties adjacent to a metropolitan area, and nonmetropolitan counties not adjacent to a metropolitan area.

[h]Beam radiation, radiation, not otherwise specified method, source not specified, or radioactive implants.

[i]Refused (these data have been recorded by SEER since 1988), recommended, unknown if administered, and none/unknown.

Among the localized patients, those with the nasal type (420 patients, 85.5%) had a notably higher median OS of 121.0 months (95% CI 89.1 - 152.9 mo), with 3-year and 5-year OS rates of 67.4% and 59.7%, respectively. In contrast, the 71 patients (14.5%) with the extranasal type experienced a significantly shorter median OS of only 13.0 months (95% CI 6.6 - 19.3 mo), with 3-year and 5-year OS rates of 31.7% and 28.0%, respectively. In stages III and IV (distant), patients with the nasal type also showed superior survival outcomes compared to those with the extranasal type. The median OS for the nasal type was 7.0 months (95% CI 5.2 - 8.8 mo), significantly better than the median OS for the extranasal type, which was 5.0 months (95% CI 3.6 - 6.4 mo; *P*=.04). The 3-year and 5-year OS rates for the nasal type were 24.9% and 24.0%, respectively, while those for the extranasal type were lower at 17.5% and 15.6%, respectively.

Further analysis of the survival of patients with specific extranodal invasion sites revealed that those with nasal cavity involvement exhibited the best median OS of 63.0 months (95% CI 33.8 - 92.2 mo), significantly outperforming patients with other types of extranodal involvement (*P*<.05). This group was followed by patients with involvement of the oropharynx, nasopharynx, pharynx, and throat, who had a median OS of 14.0 months (95% CI 6.9 - 21.1 mo). Patients with paranasal sinus involvement had the third highest median OS at 12.0 months (95% CI 0 - 38.1 mo), as shown in Table S2 and Figure S1 in Multimedia Appendix 1. Patients with nodal invasion, including involvement of splenic and Weil's rings, also experienced poor outcomes, with a median OS of only 6.0 months and a 3-year OS rate of 23.3%. The median OS was even lower in patients presenting with B symptoms at 11.0 months (95% CI 3.4 - 18.6 mo), compared to those without B symptoms, who had a median OS of 67.0 months (95% CI 40.0 - 94.0 m, *P*<.001).

Radiotherapy and chemotherapy are primary treatments for NKTL. The analysis indicated that patients who underwent either radiotherapy or chemotherapy exhibited improved survival (*P*<.001). Specifically, 477 patients received both treatments

XSL•FO
RenderX

concurrently and demonstrated the highest OS, with a median OS of 138.0 months (95% CI 76.0 - 200.0 mo). This outcome was significantly superior to those who underwent only radiotherapy (27.0 mo, 95% CI 10.7 - 43.3 mo) or chemotherapy (7.0 mo, 95% CI 5.6 - 8.4 mo), and to patients categorized as no/unknown (3.0 mo, 95% CI 1.4 - 4.6 mo; *P*<.001) (Table 2, Table S1 in Multimedia Appendix 1, Figure 2F). Five-year OS rates for these groups were 58.0%, 37.0%, 23.3%, and 25.6%, respectively. Additionally, survival for patients treated with only radiotherapy was also significantly higher than for those who received only chemotherapy or no treatment (*P*=.001 and *P*<.001, respectively). Survival rates for patients undergoing chemotherapy alone were also significantly higher than for those receiving no treatment (*P*=.02). Multivariate Cox regression analysis incorporated variables such as year of diagnosis, age at diagnosis, annual household income, whether it was the first tumor, clinical stage, nasal type, presence of B symptoms, and treatment types. Results identified several independent prognostic factors for NKTL. Age at diagnosis ≥70 years (HR=1.839; *P*<.001), advanced clinical stages II (HR=1.699; *P*<.001) and III/IV (HR=2.365; *P*<.001), nasal type (HR=0.815; *P*=.02), presence of B symptoms (HR=1.418; *P*=.005), radiotherapy (HR=0.623; *P*=.001), chemotherapy (HR=0.707; *P*=.003), and combined radiotherapy and chemotherapy (HR=0.373; *P*<.001) were determined to be significant predictors of survival (Table S1 in Multimedia Appendix 1, Figure 3).

Figure 3. Multivariate Cox proportional hazards regression model to analyze the factors influencing the prognosis of patients with natural killer/T-cell lymphoma. HR: hazard ratio.

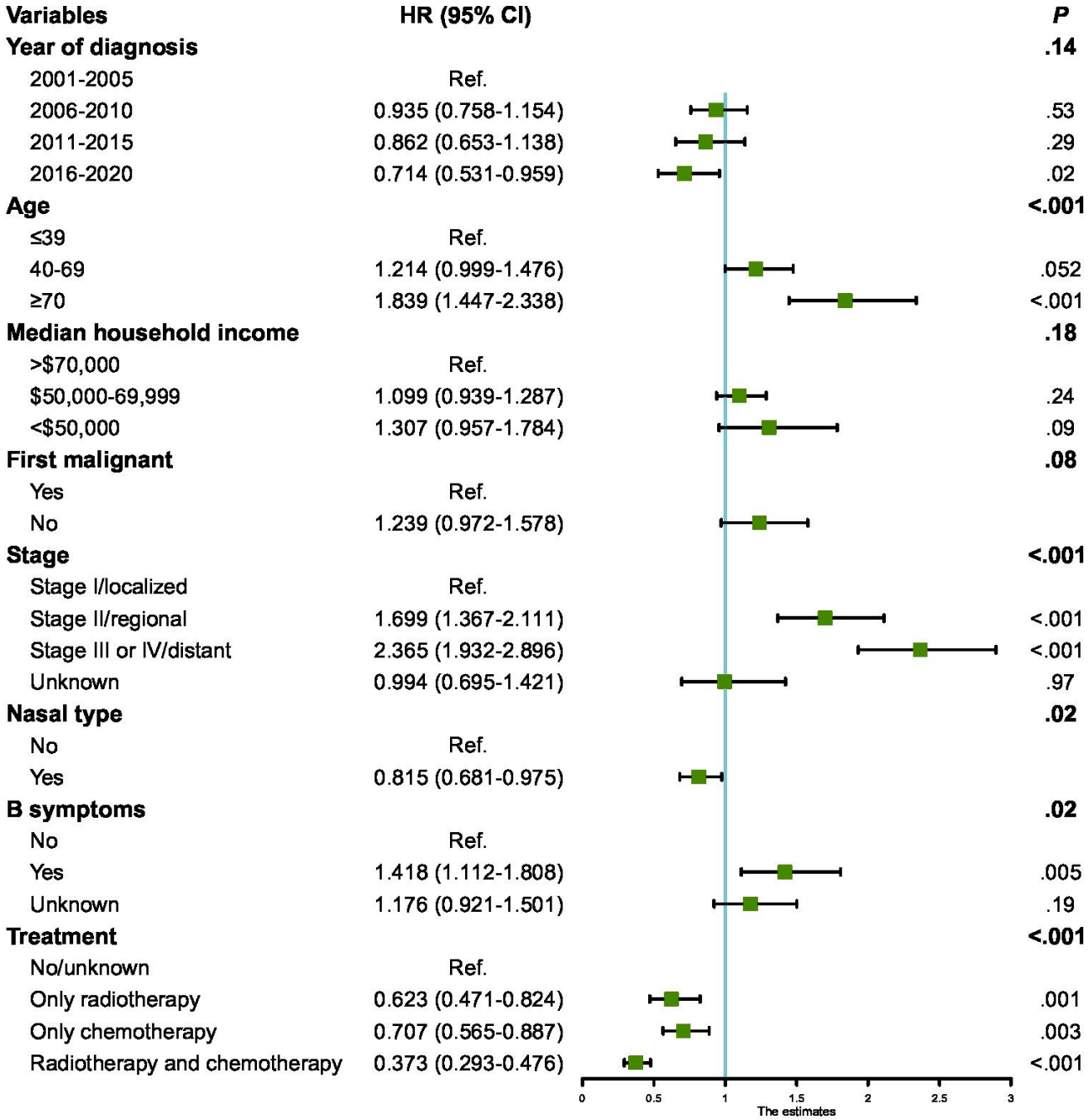

XSL•FO
RenderX

Clinical Characteristics and Subgroup Survival Analysis of Patients Receiving Radiotherapy and Chemotherapy

Patients receiving both radiotherapy and chemotherapy demonstrated the highest survival rates, with an HR of 0.293 (95% CI 0.237 - 0.364). Treatment with radiotherapy or chemotherapy alone also significantly improved survival, with HRs of 0.563 (95% CI 0.435 - 0.728) for radiotherapy alone and 0.808 (95% CI 0.659 - 0.989) for chemotherapy alone. Further analysis of clinical characteristics revealed significant variations based on the type of treatment received. Differences were observed in the year of diagnosis, age at diagnosis, marital status, whether the involvement was extranodal or nodal, nasal type, clinical stage, presence of B symptoms, and whether it was the first tumor (Table 3).

We created subgroups of patients based on factors that might influence the clinical treatment decisions of hematologists treating malignancies, examining the impact of receiving radiotherapy, chemotherapy, or a combination of both on survival across these subgroups. These subgroups were defined by patients' age, clinical stage, and nasal type status. The results are detailed in Table 4. Patients with varying clinical stages, different age groups, and whether they were of the nasal type or not, who received a combination of radiotherapy and chemotherapy, exhibited the best survival (P<.001). Notably, among stage I/localized patients with nasal type—a total of 420 patients—those receiving combination therapy (231 patients, 55%), radiotherapy alone (81 patients, 19.3%), chemotherapy alone (56 patients, 13.3%), and neither therapy (52 patients, 12.4%) had median OS rates of not reached, 47.0 months (95% CI 30.2 - 63.7 mo), 22.0 months (95% CI 0 - 61.1 mo), and 40.0 months (95% CI 0 - 153.2 mo), respectively. These differences were statistically significant (P<.001). Survival was significantly superior in the combination therapy group compared to the other 3 groups. However, there were no significant differences in survival among the latter 3 groups (Figure 4). Five-year OS rates were 73.3% in the combination therapy group, 44.9% in the radiotherapy alone group, and 40.8% in the chemotherapy alone group.

Among the 373 patients with advanced, distant-stage disease, treatment modalities were distributed as follows: 85 patients (22.8%) received combined radiotherapy and chemotherapy, 22 patients (5.9%) received radiotherapy alone, 195 patients (52.3%) underwent chemotherapy alone, and 71 patients (19.0%) received neither treatment. The median overall survival (OS) for these groups was 11.0 months (95% CI 7.0 - 15.0 mo), 3.0 months (95% CI 0 - 7.5 mo), 6.0 months (95% CI 4.4 - 7.6 mo), and 1.0 month (95% CI 0.4 - 1.6 mo), respectively, demonstrating significant differences (P<.001). The 1-year OS rates were 48.4% for combined therapy, 20.2% for radiotherapy alone, 31.4% for chemotherapy alone, and 16.3% for no treatment. Patients receiving combination therapy exhibited significantly better survival compared to those in the other 3 groups. The 5-year OS for the combination therapy group was 34.7%. There was no significant difference in survival among the other 3 groups (Figure 4).

XSL•FO
RenderX

Table . Demographic and clinical characteristics of patients with natural killer/T-cell lymphoma based on the clinical treatment decisions (United States; 2000 - 2020; N=1162).

Variable	Radiotherapy and chemotherapy, n (%)	Only radiotherapy, n (%)	Only chemotherapy, n (%)	No/unknown, n (%)	*P* value
Year of diagnosis					<.001
2000-2005	80 (35.4)	18 (8)	82 (36.3)	46 (20.4)	
2006-2010	97 (33)	54 (18.4)	86 (29.3)	57 (19.4)	
2011-2015	149 (44.1)	45 (13.3)	90 (26.6)	54 (16)	
2016-2020	151 (49.7)	21 (6.9)	94 (30.9)	38 (12.5)	
Gender					.28
Female	161 (40.5)	51 (12.8)	110 (27.6)	76 (19.1)	
Male	316 (41.4)	87 (11.4)	242 (31.7)	119 (15.6)	
Age group (years)					<.001
≤39	144 (49)	21 (7.1)	90 (30.6)	39 (13.3)	
40-69	277 (44.4)	50 (8)	222 (35.6)	75 (12)	
≥70	56 (23)	67 (27.5)	40 (16.4)	81 (33.2)	
Race (n=1159)					.08
NHW[a]	169 (42)	49 (12.2)	111 (27.6)	73 (18.2)	
NHB[b]	14 (16.4)	4 (7.5)	25 (47.2)	10 (18.9)	
Hispanic	186 (43.9)	49 (11.6)	133 (31.4)	56 (13.2)	
NHAPI[c]	101 (38)	35 (13.2)	80 (30.1)	50 (18.8)	
NHAIAN[d]	7 (50)	1 (7.1)	2 (14.3)	4 (28.6)	
Marital status at diagnosis (n=1098)					<.001
Married	300 (44.8)	83 (12.4)	205 (30.6)	81 (12.1)	
Unmarried[e]	163 (38)	48 (11.2)	136 (31.7)	82 (19.1)	
Median household income (US $)					.43
>70,000	266 (43.5)	68 (11.1)	188 (30.7)	90 (14.7)	
50,000-69,999	185 (38.3)	62 (12.8)	144 (29.8)	92 (19)	
<50,000	26 (38.8)	8 (11.9)	20 (29.9)	13 (19.4)	
Rural-urban continuum (n=1158)					.79
Counties in metropolitan areas[f]	437 (41)	127 (11.9)	327 (30.6)	176 (16.5)	
Nonmetropolitan counties[g]	38 (41.8)	11 (12.1)	24 (26.4)	18 (19.8)	
Nodal or extranodal					<.001
Nodal	32 (21.9)	3 (2.1)	76 (52.1)	35 (24)	
Extranodal	445 (43.8)	135 (13.3)	276 (27.2)	160 (15.7)	
Nasal type (n=1149)					<.001
No (n=316)	59 (18.7)	14 (4.4)	158 (50)	85 (26.9)	
Yes (n=833)	416 (49.9)	124 (14.9)	190 (22.8)	103 (12.4)	
Stage (n=1096)					<.001
Stage I/localized	250 (50.8)	90 (18.3)	85 (17.3)	67 (13.6)	
Stage II/regional	134 (59)	23 (10)	56 (24.2)	18 (7.8)	
Stage III or IV/distant	85 (22.8)	22 (5.9)	195 (52.3)	71 (19)	

XSL•FO
RenderX

Variable	Radiotherapy and chemotherapy, n (%)	Only radiotherapy, n (%)	Only chemotherapy, n (%)	No/unknown, n (%)	*P* value
B symptoms					<.001
No	213 (55.6)	48 (12.5)	81 (21.1)	41 (10.7)	
Yes	79 (39.3)	14 (7)	81 (40.3)	27 (13.4)	
Unknown	185 (32)	76 (13.1)	190 (32.9)	127 (22)	
First malignant or not					<.001
Yes	444 (42.3)	327 (31.2)	114 (10.9)	164 (15.6)	
No	33 (29.2)	25 (22.1)	24 (21.2)	31 (27.4)	

[a]NHW: non-Hispanic White.

[b]NHB: non-Hispanic Black.

[c]NHAPI: non-Hispanic Asian or Pacific Islander.

[d]NHAIAN: non-Hispanic American Indian/Alaska Native.

[e]Single (never married), widowed, divorced, separated, unmarried, or domestic partner.

[f]Counties in metropolitan areas of 1 million population, counties in metropolitan areas of 250,000 to 1 million population, and counties in metropolitan areas of 250,000 population.

[g]Nonmetropolitan counties adjacent to a metropolitan area, and nonmetropolitan counties not adjacent to a metropolitan area.

Table . Survival analysis for different subgroups of patients with natural killer/T-cell lymphoma based on the clinical treatment decisions (United States; 2000 - 2020; N=1162).

Variable		Radiotherapy and chemotherapy, median (95% CI)	Only radiotherapy, median (95% CI)	Only chemotherapy, median (95% CI)	No/unknown, median (95% CI)	*P* value
Clinical stage						
	Stage I/localized	Not reached	47.0 (37.5 - 56.5)	13.0 (5.6 - 20.4)	15.0 (0 - 48.8)	<.001
	Stage II/regional	52.0 (25.9 - 78.1)	9.0 (0.5 - 17.5)	10.0 (3.0 - 17.0)	1.0 (0 - 2.4)	<.001
	Stage III or IV/distant	11.0 (7.0 - 15.0)	3.0 (0 - 7.5)	6.0 (4.4 - 7.6)	1.0 (0.4 - 1.6)	<.001
	Unknown	57.0 (0 - 162.8)	6.0 (0 - 14.0)	21.0 (0 - 42.1)	6.0 (0 - 22.5)	.22
Age group (years)						
	≤39	Not reached	93.0 (0 - 199.6)	8.0 (3.1 - 12.9)	7.0 (0 - 33.8)	<.001
	40-69	138.0 (77.2 - 198.8)	18.0 (0 - 37.5)	8.0 (6.3 - 9.7)	4.0 (0 - 9.2)	<.001
	≥70	59.0 (20.0 - 98.0)	38.0 (12.9 - 63.1)	6.0 (3.7 - 8.3)	1.0 (0.2 - 1.8)	<.001
Site of involvement						
	Nasal type	156.0 (100.6 - 211.4)	35.0 (17.5 - 52.5)	9.0 (4.7 - 13.3)	7.0 (0 - 15.3)	<.001
	Extranasal type	15.0 (0.1 - 29.9)	7.0 (1.5 - 12.5)	6.0 (4.2 - 7.8)	1.0 (0 - 2.0)	<.001

XSL•FO
RenderX

Figure 4. OS of patients with different treatment modalities (combination of radiotherapy and chemotherapy, radiotherapy alone, chemotherapy alone, or neither treatment). OS: overall survival.

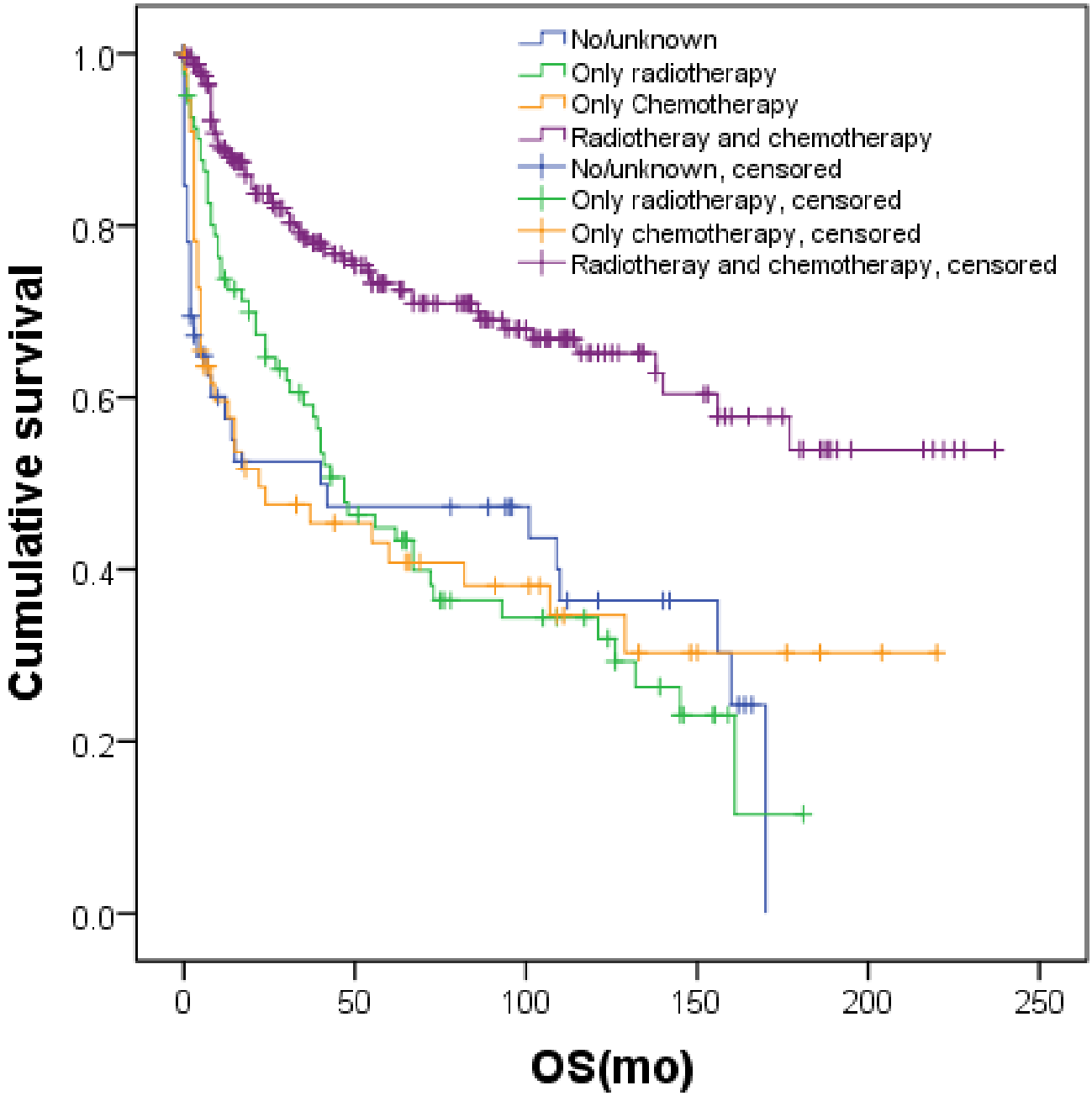

Lymphoma-Specific Survival

NKTL was the initial diagnosis for 1049 patients, of whom 17 had unknown causes of death. Consequently, 1032 patients were included in the lymphoma-specific mortality analysis. By the end of the follow-up period, 491 (47.6%) of these patients had succumbed to NKTL, with a median lymphoma-specific OS of 50.0 months (95% CI 27.3 - 72.7 mo). An analysis of patient survival based on their treatment modalities—radiotherapy and chemotherapy—revealed that those who received combination therapy had superior survival compared to those treated with radiotherapy alone, chemotherapy alone, or neither treatment (P=.011, P<.001, and P<.001, respectively). The median OS for these groups was not reached, 72.0 months (95% CI 0 - 153.5 mo), 8.0 months (95% CI 6.0 - 10.0 mo), and 10.0 months (95% CI 2.1 - 17.8 mo), respectively. Furthermore, patients who underwent radiotherapy alone experienced significantly better survival compared to those who underwent chemotherapy alone and those who did not receive either treatment (P<.001 for both comparisons). However, there was no significant difference in survival between patients who received chemotherapy alone and those who did not receive any treatment (P=.73) (Figure 5).

XSL•FO
RenderX

Figure 5. Lymphoma-specific survival in patients with different treatment modalities (combination of radiotherapy and chemotherapy, radiotherapy alone, chemotherapy alone, or neither treatment). OS: overall survival.

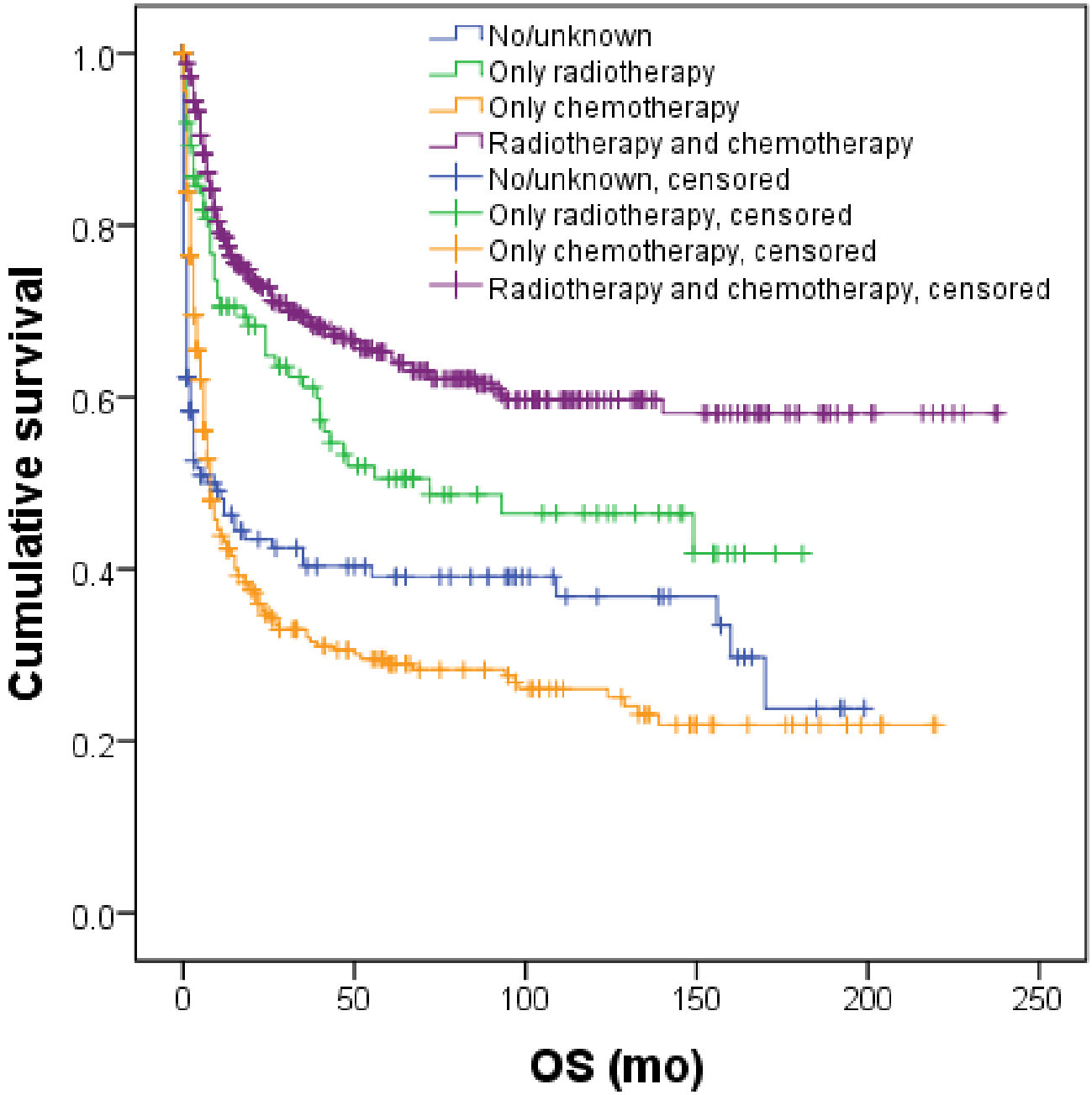

Discussion

Principal Findings

In this study, we observed that the incidence rate of NKTL increased with age, and there was no significant change in the incidence of the disease between 2000 and 2020, although notable differences in incidence among different racial groups were identified. The use of combined radiotherapy and chemotherapy markedly improved survival rates, even in patients with limited nasal type or advanced disease. To the best of our knowledge, this study is the first to successfully present a comprehensive overview of NKTL management and outcomes in the United States.

NKTL is a rare lymphoma originating from mature T/NK cells. According to data from the SEER 17 database, the average annual incidence rate across 17 regions in the United States was 0.067 per 100,000 between 2000 and 2020, with the rate remaining relatively stable over the past 2 decades. The incidence rate was significantly higher in males than in females, with a ratio of approximately 2.11:1. Additionally, the incidence rate gradually increased with age, without showing peaks in any specific age groups. Racial disparities in incidence rates were noted; NHAPI and Hispanic populations had higher rates ranging from 0.132 to 0.157 per 100,000, whereas NHW and NHB populations had lower rates around 0.03 per 100,000, consistent with previous reports [1,2,5,14]. The underlying reasons for these differences are not well understood but are hypothesized to be related to variations in EBV subtypes, genetic susceptibility, or both [7,15-17]. There are 2 main strains or branches of EBV currently identified globally, each with multiple variants. Notably, the strains found in Asia differ

XSL•FO
RenderX

significantly from those in other regions [18]. Transcriptome analyses have indicated that certain EBV variants are more likely to cause NKTL or other malignancies [19,20]. A recent collaborative effort involving experts from China, the United States, and Mexico conducted comprehensive genomic analyses and related clinical studies on 209 patients with extranodal NKTL. This collaboration led to the identification of 7 distinct gene clusters associated with varying clinical outcomes, revealing notable genetic similarities between Asian and Hispanic patients that suggest a possible shared pathogenesis and tumor development [8]. Furthermore, polymorphisms in genes related to the host immune-inflammatory responses have been linked to NKTL development [16,21]. However, further research is needed to deepen our understanding of the genetic susceptibility associated with NKTL.

Like many cancers, NKTL predominantly affects adults, with most studies reporting a median age of diagnosis between 44 and 54 years. In our cohort, the median age was 54 years. Notably, we observed significant racial differences in the median age at first diagnosis; Hispanic and NHAIAN patients were generally younger, with a median age of less than 45 years, whereas NHW and NHAPI patients had a median age of about 60 years. Our survival analysis identified that being aged 70 years or older at diagnosis was an independent adverse factor affecting patient survival. Moreover, patients diagnosed in more recent years have demonstrated improved survival outcomes. Specifically, for those diagnosed after 2016, the median survival was 44.0 months and the 3-year OS rate was 53.5%. This improvement is likely due to the introduction and use of more effective treatments for NK/T lymphomas, such as L-asparaginase and immune checkpoint inhibitors [1,3].

The Ann Arbor staging system, commonly used for clinical staging of NKTL, suggests significant prognostic implications. In our study, a majority of patients (723 cases, 62.2%) were classified as stage I/II with relatively limited invasion. However, this staging system primarily assesses the extent of lesion involvement and fails to account for the significant prognostic impact of primary invasion of the upper aerodigestive tract by NKTL [22]. To address this, the extranodal NKTL (Chinese-Asia or CA) staging system, developed byChina and Asia Lymphoma Study Group, incorporates factors such as the site of involvement, regional lymph node involvement, and distant metastasis. This system has been validated in patients treated with asparaginase [23]. Unfortunately, due to data acquisition limitations, we were unable to validate this prognostic staging system in our cohort.

The upper aerodigestive tract—including the nasal cavity, nasopharynx, sinuses, and palate—is the most frequent site of NKTL involvement. Patients with primary invasion of these areas are categorized as having ENKL, nasal type. In contrast, primary extranasal invasion sites such as the skin, soft tissues, gastrointestinal tract, and testes, are associated with the extranasal type. This type presents more adverse prognostic clinical features and poorer OS, which has been confirmed by multiple clinical studies [5,6,8,24]. Aggressive NK-cell leukemia was once considered an aggressive subtype of extranodal NKTL with vascular involvement [25,26]. However, in the *5th edition of the WHO Classification of Haematolymphoid Tumours*, it is now recognized as a distinct entity of peripheral T-cell lymphoma, not otherwise specified [27]. This rare lymphoma frequently involves the skin and central nervous system and is associated with a grave prognosis [28]. Due to its unclear pathological nature, this subtype has been excluded from this analysis. Notably, limited nasal or nasopharyngeal involvement is considered the most favorable prognosis group (stage I) within the CA extranodal NKTL staging system. In our patient cohort, the nasal type comprised 71.7%. Recent data from 14 and 19 medical centers in China, involving more than 1000 patients each, showed that the nasal type represented 87.7% (1393/1589) and 86.1% (1006/1168) of cases [3,23]. This prevalence suggests that the nasal type may be more common in Asian populations.

Survival analysis revealed that the median OS of our patients was 21 months, with 3-year and 5-year OS rates of 44.5% and 39.5%, respectively. Patients diagnosed with the nasal type exhibited significantly better prognoses compared to those with the nonnasal type, achieving 5-year OS rates of 46.4% and 22.2%, respectively. Specifically, among patients with localized disease and those with stage III/IV disease, which indicates extensive involvement at other sites, 5-year OS rates for the nasal type were 59.7% and 24.0%, respectively. This contrasts with 5-year OS rates of only 28% and 15.7% for patients with the nonnasal type. Comparatively, literature indicates that the median OS in the Asian population was 76 months, with a 3-year OS of 59% [2]. For patients with limited disease (stage I/II) of the nasal type, the 5-year OS was approximately 53% [22], whereas those with nasal type localized disease achieved a 5-year OS of up to 76% [24], suggesting better survival rates than those observed in our study group.

The present treatment for NKTL still relies on radiotherapy and chemotherapy. A significant portion (71.3%) of patients in this group underwent chemotherapy, while fewer than half received radiotherapy. Among the 477 patients (41%) who underwent both radiotherapy and chemotherapy, the best survival outcomes were observed, with a median OS of 138 months and a 5-year OS rate of 58%. In contrast, patients who received either radiotherapy or chemotherapy alone had lower median survival times of 27 months and 7 months, respectively. A notable difference in survival was observed among these 3 patient groups. Further analysis of patient survival across different treatments was conducted based on age subgroup, clinical stage, and nasal type. The findings indicated that patients who underwent combined radiotherapy and chemotherapy consistently exhibited the best survival outcomes across all subgroups. Even among patients with localized (stage I) nasal type disease, the combination therapy significantly enhanced survival rates. The 5-year OS rate was 73.3% in the combination therapy group, compared to 44.9% and 40.8% in the radiotherapy and chemotherapy alone groups, respectively. Recent studies have shown that in patients with ENKL, particularly those with early limited disease (Ann Arbor stage I/II confined to the nasal cavity), the primary treatment modality involving radiotherapy (either alone or in combination with chemotherapy) resulted in a 5-year OS rate of 70%. Low-risk patients treated solely with radiotherapy achieved particularly favorable outcomes, with a 5-year OS rate approaching 90%, while the addition of chemotherapy did not yield further

XSL•FO
RenderX

improvements in patient survival [23,29-31]. Nevertheless, for early-stage patients at moderate to high risk, combining radiotherapy with chemotherapy led to significant enhancements in both OS and progression-free survival compared to radiotherapy alone, with a 5-year OS and progression-free survival of 73.2% versus 60.9% (P<.001) and 63.5% versus 54.2% (P<.001), respectively. Hence, it is suggested that chemotherapy could be adjunctive to radiotherapy solely for early-stage patients with NKTL facing a grim prognosis [29-33]. Current National Comprehensive Cancer Network guidelines suggest that stage I/II patients with extranodal nasal-type NKTL first be assessed for their fitness for chemotherapy, followed by treatment with radiotherapy alone or through clinical trials. Radiotherapy combined with chemotherapy is only recommended for patients with advanced-stage nasal disease and all patients with extranasal involvement. In light of the above, incorporating risk-adapted therapy into the guidelines is essential to optimize the benefits of combined chemoradiotherapy for early-stage disease. However, additional confirmation through extensive prospective clinical studies with large sample sizes may be warranted.

In patients presenting with nonnasal type or advanced (stage III/IV) disease, treatment regimens containing L-asparaginase, such as SMILE (dexamethasone, methotrexate, ifosfamide, L-asparaginase, and etoposide) and P-GEMOX (pembrolizumab, gemcitabine, and oxaliplatin), have significantly enhanced outcomes when combined with consolidation radiotherapy. The 5-year OS rates for these regimens ranged from 34% to 50% [10,34]. In this patient cohort, those with distant metastasis exhibited a 5-year OS of 19.6%. However, patients who received both radiotherapy and chemotherapy demonstrated markedly improved survival, with a 5-year OS of 34.7%. The integration of radiotherapy and chemotherapy can be administered either simultaneously or sequentially, using a sandwich approach [32,35]. Nevertheless, patients with advanced NKTL who have extensive disease primarily undergo chemotherapy [10,34,36,37].

Other factors that may affect patient survival include annual household income at the time of diagnosis, with patients from higher-income households generally exhibiting better survival rates. Conversely, patients with prior diagnoses of other tumors or those experiencing systemic symptoms tend to have poorer survival outcomes. The multivariate Cox regression analysis identified several independent prognostic factors for patients with NKTL, including patient age, clinical stage, nasal type, systemic symptoms, and the receipt of both radiotherapy and chemotherapy. EBV viral load is considered to have a significant prognostic impact on survival in NKTL. A meta-analysis of 11 studies found that a high pretreatment EBV viral load was associated with worse survival outcomes (HR=3.45; P=.001). Furthermore, patients whose EBV DNA levels returned to normal ranges after treatment demonstrated better survival than those with persistently positive EBV DNA levels [35,38,39]. Consequently, measuring EBV DNA viral load through quantitative polymerase chain reaction is deemed valuable for both diagnostic and prognostic purposes and is recommended by the National Comprehensive Cancer Network as part of the initial workup. Unfortunately, the SEER program does not provide this information.

This retrospective study has several inherent limitations. The SEER program lacks critical data such as patients' physical status, serum lactate dehydrogenase levels, results from other prognostic scoring systems, EBV viral load, and specific treatment regimens. Despite these constraints, the analysis presented in this paper yielded some valuable insights. However, to substantiate these findings, prospective, multicenter studies are necessary.

Conclusions

Overall, this study indicated that the incidence of NKTL has remained stable in recent years. Patients with the nasal type or the use of combined radiotherapy and chemotherapy generally experience better survival outcomes. The results presented an overview of the incidence rate, demographic and clinical characteristics, treatment regimens, and outcomes of NKTL in the United States, which may provide data support for the precise diagnosis and treatment of this rare disease. Further validation through prospective multicenter clinical trials can be done to confirm these findings.

Acknowledgments

We are grateful to the Surveillance, Epidemiology, and End Results (SEER) Program cancer registries in the creation of the SEER database. This research was funded by National Natural Science Foundation of China (number 82270206), Zhejiang Provincial Natural Science Foundation (number LZ22H160009) and Key Project of the Zhejiang Higher Education Association's Special Project on "Smart Education and Digital Transformation of Teaching" (KT2023416).

Data Availability

The database is publicly available and can be obtained upon user authentication from the Surveillance, Epidemiology, and End Results Program at the National Cancer Institute. All data used and analyzed are available from the corresponding authors on reasonable request.

Authors' Contributions

The work reported in the paper has been performed by the authors, unless clearly specified in the text.
Conceptualization: Y Zhao.
Data curation and methodology: Y Zhao, Y Zhong.

XSL•FO
RenderX

Formal analysis and investigation: Y Zhao, Y Zhong, MX, LH.
Supervision: XY, JH.
Writing – original draft: Y Zhao, Y Zhong.
Writing – review & editing: all authors.
Author XY is the co-corresponding author and can be reached at yxjsunny@zju.edu.cn.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary Tables S1-S2 and Figures S1-S2.
[DOCX File, 125 KB - formative_v9i1e70129_app1.docx]

References

1. Wang H, Fu BB, Gale RP, Liang Y. NK-/T-cell lymphomas. Leukemia 2021 Sep;35(9):2460-2468. [doi: 10.1038/s41375-021-01313-2] [Medline: 34117356]
2. Kim SJ, Yoon DH, Jaccard A, et al. A prognostic index for natural killer cell lymphoma after non-anthracycline-based treatment: a multicentre, retrospective analysis. Lancet Oncol 2016 Mar;17(3):389-400. [doi: 10.1016/S1470-2045(15)00533-1] [Medline: 26873565]
3. Chen Z, Huang H, Huang H, et al. Genomic features reveal potential benefit of adding anti-PD-1 immunotherapy to treat non-upper aerodigestive tract natural killer/T-cell lymphoma. Leukemia 2024 Apr;38(4):829-839. [doi: 10.1038/s41375-024-02171-4] [Medline: 38378844]
4. Harabuchi Y, Imai S, Wakashima J, et al. Nasal T-cell lymphoma causally associated with Epstein-Barr virus: clinicopathologic, phenotypic, and genotypic studies. Cancer 1996;77:2137-2149. [doi: 10.1002/(SICI)1097-0142(19960515)77:103.0.CO;2-V] [Medline: 8640683]
5. Au WY, Weisenburger DD, Intragumtornchai T, et al. Clinical differences between nasal and extranasal natural killer/T-cell lymphoma: a study of 136 cases from the International Peripheral T-Cell Lymphoma Project. Blood 2009 Apr 23;113(17):3931-3937. [doi: 10.1182/blood-2008-10-185256]
6. Julie V, James A, Dennis W. International peripheral T-cell and natural killer/T-cell lymphoma study: pathology findings and clinical outcomes. JCO 2008 Sep 1;26(25):4124-4130. [doi: 10.1200/JCO.2008.16.4558]
7. Montes-Mojarro IA, Chen BJ, Ramirez-Ibarguen AF, et al. Mutational profile and EBV strains of extranodal NK/T-cell lymphoma, nasal type in Latin America. Mod Pathol 2020 May;33(5):781-791. [doi: 10.1038/s41379-019-0415-5] [Medline: 31822801]
8. Dong G, Liu X, Wang L, et al. Genomic profiling identifies distinct genetic subtypes in extra-nodal natural killer/T-cell lymphoma. Leukemia 2022 Aug;36(8):2064-2075. [doi: 10.1038/s41375-022-01623-z] [Medline: 35697790]
9. Yang Y, Cao JZ, Lan SM, et al. Association of improved locoregional control with prolonged survival in early-stage extranodal nasal-type natural killer/T-cell lymphoma. JAMA Oncol 2017 Jan 1;3(1):83-91. [doi: 10.1001/jamaoncol.2016.5094] [Medline: 27893001]
10. Li X, Cui Y, Sun Z, et al. DDGP versus SMILE in newly diagnosed advanced natural killer/T-cell lymphoma: a randomized controlled, multicenter, open-label study in China. Clin Cancer Res 2016 Nov 1;22(21):5223-5228. [doi: 10.1158/1078-0432.CCR-16-0153] [Medline: 27060152]
11. Gaal K, Sun NC, Hernandez AM, Arber DA. Sinonasal NK/T-cell lymphomas in the United States. Am J Surg Pathol 2000 Nov;24(11):1511-1517. [doi: 10.1097/00000478-200011000-00006] [Medline: 11075852]
12. Vargo JA, Patel A, Glaser SM, et al. The impact of the omission or inadequate dosing of radiotherapy in extranodal natural killer T-cell lymphoma, nasal type, in the United States. Cancer 2017 Aug 15;123(16):3176-3185. [doi: 10.1002/cncr.30697] [Medline: 28380259]
13. SEER*Stat database. Incidence - SEER Research Data, 17 Registries, Nov 2022 Sub (2000-2020) - Linked to County Attributes - Time Dependent (1990-2021) Income/Rurality, 1969-2021 Counties. Surveillance, Epidemiology, and End Results (SEER) Program. URL: https://seer.cancer.gov/data-software/documentation/seerstat/nov2022/ [accessed 2025-05-06]
14. Adams SV, Newcomb PA, Shustov AR. Racial patterns of peripheral T-cell lymphoma incidence and survival in the United States. J Clin Oncol 2016 Mar 20;34(9):963-971. [doi: 10.1200/JCO.2015.63.5540] [Medline: 26962200]
15. Saleem A, Natkunam Y. Extranodal NK/T-cell lymphomas: the role of natural killer cells and EBV in lymphomagenesis. Int J Mol Sci 2020 Feb 22;21(4):1501. [doi: 10.3390/ijms21041501] [Medline: 32098335]
16. Lin GW, Xu C, Chen K, et al. Genetic risk of extranodal natural killer T-cell lymphoma: a genome-wide association study in multiple populations. Lancet Oncol 2020 Feb;21(2):306-316. [doi: 10.1016/S1470-2045(19)30799-5] [Medline: 31879220]
17. Okuno Y, Murata T, Sato Y, et al. Defective Epstein-Barr virus in chronic active infection and haematological malignancy. Nat Microbiol 2019 Mar;4(3):404-413. [doi: 10.1038/s41564-018-0334-0] [Medline: 30664667]

XSL•FO
RenderX

18. Palser AL, Grayson NE, White RE, et al. Genome diversity of Epstein-Barr virus from multiple tumor types and normal infection. J Virol 2015 May;89(10):5222-5237. [doi: 10.1128/JVI.03614-14] [Medline: 25787276]
19. Peng RJ, Han BW, Cai QQ, et al. Genomic and transcriptomic landscapes of Epstein-Barr virus in extranodal natural killer T-cell lymphoma. Leukemia 2019 Jun;33(6):1451-1462. [doi: 10.1038/s41375-018-0324-5] [Medline: 30546078]
20. Chen ZH, Yan SM, Chen XX, et al. The genomic architecture of EBV and infected gastric tissue from precursor lesions to carcinoma. Genome Med 2021 Sep 7;13(1):146. [doi: 10.1186/s13073-021-00963-2] [Medline: 34493320]
21. Li Z, Xia Y, Feng LN, et al. Genetic risk of extranodal natural killer T-cell lymphoma: a genome-wide association study. Lancet Oncol 2016 Sep;17(9):1240-1247. [doi: 10.1016/S1470-2045(16)30148-6] [Medline: 27470079]
22. Kim TM, Park YH, Lee SY, et al. Local tumor invasiveness is more predictive of survival than International Prognostic Index in stage I(E)/II(E) extranodal NK/T-cell lymphoma, nasal type. Blood 2005 Dec 1;106(12):3785-3790. [doi: 10.1182/blood-2005-05-2056] [Medline: 16109779]
23. Hong H, Li Y, Lim ST, et al. A proposal for a new staging system for extranodal natural killer T-cell lymphoma: a multicenter study from China and Asia Lymphoma Study Group. Leukemia 2020 Aug;34(8):2243-2248. [doi: 10.1038/s41375-020-0740-1] [Medline: 32066865]
24. Lee J, Suh C, Park YH, et al. Extranodal natural killer T-cell lymphoma, nasal-type: a prognostic model from a retrospective multicenter study. J Clin Oncol 2006 Feb 1;24(4):612-618. [doi: 10.1200/JCO.2005.04.1384] [Medline: 16380410]
25. Kuo TT, Chen MJ, Kuo MC. Cutaneous intravascular NK-cell lymphoma: report of a rare variant associated with Epstein-Barr virus. Am J Surg Pathol 2006 Sep;30(9):1197-1201. [doi: 10.1097/01.pas.0000213263.99973.09] [Medline: 16931967]
26. Cerroni L, Massone C, Kutzner H, Mentzel T, Umbert P, Kerl H. Intravascular large T-cell or NK-cell lymphoma: a rare variant of intravascular large cell lymphoma with frequent cytotoxic phenotype and association with Epstein-Barr virus infection. Am J Surg Pathol 2008 Jun;32(6):891-898. [doi: 10.1097/PAS.0b013e31815d29c9] [Medline: 18425045]
27. Alaggio R, Amador C, Anagnostopoulos I, et al. The 5th edition of the World Health Organization Classification of Haematolymphoid Tumours: lymphoid neoplasms. Leukemia 2022;37(9). [doi: 10.1038/s41375-023-01962-5]
28. El Hussein S, Medeiros LJ, Khoury JD. Aggressive NK cell leukemia: current state of the art. Cancers (Basel) 2020 Oct 9;12(10):2900. [doi: 10.3390/cancers12102900] [Medline: 33050313]
29. Yang Y, Zhu Y, Cao JZ, et al. Risk-adapted therapy for early-stage extranodal nasal-type NK/T-cell lymphoma: analysis from a multicenter study. Blood 2015 Sep 17;126(12):1424-1432. [doi: 10.1182/blood-2015-04-639336] [Medline: 26109206]
30. Liu X, Wu T, Zhu SY, et al. Risk-dependent conditional survival and failure hazard after radiotherapy for early-stage extranodal natural killer/T-cell lymphoma. JAMA Netw Open 2019 Mar 1;2(3):e190194. [doi: 10.1001/jamanetworkopen.2019.0194] [Medline: 30821826]
31. Qi SN, Yang Y, Zhang YJ, et al. Risk-based, response-adapted therapy for early-stage extranodal nasal-type NK/T-cell lymphoma in the modern chemotherapy era: a China Lymphoma Collaborative Group study. Am J Hematol 2020 Sep;95(9):1047-1056. [doi: 10.1002/ajh.25878] [Medline: 32449800]
32. Kim SJ, Yoon SE, Kim WS, Jin KS, Eun YS, Seog KW. Treatment of localized extranodal NK/T cell lymphoma, nasal type: a systematic review. J Hematol Oncol 2018 Dec 20;11(1):140. [doi: 10.1186/s13045-018-0687-0] [Medline: 30567593]
33. Xu PP, Xiong J, Cheng S, et al. A phase II study of methotrexate, etoposide, dexamethasone and pegaspargase sandwiched with radiotherapy in the treatment of newly diagnosed, stage IE to IIE extranodal natural-killer/T-cell lymphoma, nasal-type. EBioMedicine 2017 Nov;25:41-49. [doi: 10.1016/j.ebiom.2017.10.011] [Medline: 29056540]
34. Kwong YL, Kim WS, Lim ST, et al. SMILE for natural killer/T-cell lymphoma: analysis of safety and efficacy from the Asia Lymphoma Study Group. Blood 2012 Oct 11;120(15):2973-2980. [doi: 10.1182/blood-2012-05-431460] [Medline: 22919026]
35. Qi SN, Li YX, Specht L, et al. Modern radiation therapy for extranodal nasal-type NK/T-cell lymphoma: risk-adapted therapy, target volume, and dose guidelines from the International Lymphoma Radiation Oncology Group. Int J Radiat Oncol Biol Phys 2021 Jul 15;110(4):1064-1081. [doi: 10.1016/j.ijrobp.2021.02.011] [Medline: 33581262]
36. Yamaguchi M, Kwong YL, Kim WS, et al. Phase II study of SMILE chemotherapy for newly diagnosed stage IV, relapsed, or refractory extranodal natural killer (NK)/T-cell lymphoma, nasal type: the NK-Cell Tumor Study Group study. J Clin Oncol 2011 Nov 20;29(33):4410-4416. [doi: 10.1200/JCO.2011.35.6287] [Medline: 21990393]
37. Liang JH, Wang L, Peter Gale R, et al. Efficacy of pegaspargase, etoposide, methotrexate and dexamethasone in newly diagnosed advanced-stage extra-nodal natural killer/T-cell lymphoma with the analysis of the prognosis of whole blood EBV-DNA. Blood Cancer J 2017 Sep 15;7(9):e608. [doi: 10.1038/bcj.2017.88] [Medline: 29016569]
38. Zhao Q, Fan S, Chang Y, et al. Clinical efficacy of cisplatin, dexamethasone, gemcitabine and pegaspargase (DDGP) in the initial treatment of advanced stage (stage III–IV) extranodal NK/T-cell lymphoma, and its correlation with Epstein-Barr virus. Cancer Manag Res 2019;11:3555-3564. [doi: 10.2147/CMAR.S191929] [Medline: 31118779]
39. Liu ZL, Bi XW, Liu PP, Lei DX, Jiang WQ, Xia Y. The clinical utility of circulating Epstein-Barr virus DNA concentrations in NK/T-cell lymphoma: a meta-analysis. Dis Markers 2018;2018:1961058. [doi: 10.1155/2018/1961058] [Medline: 30581497]

XSL•FO
RenderX

Abbreviations

EBV: Epstein-Barr virus
ENKL: extranodal natural killer/T-cell lymphoma
HR: hazard ratio
IR: incidence rate
NHAIAN: non-Hispanic American Indian/Alaska Native
NHAPI: non-Hispanic Asian or Pacific Islanders
NHB: non-Hispanic Black
NHW: non-Hispanic White
NK: natural killer
NKTL: natural killer/T-cell lymphoma
OS: overall survival
P-GEMOX: pembrolizumab, gemcitabine, and oxaliplatin
RT: radiotherapy
SEER: Surveillance, Epidemiology, and End Results
SMILE: dexamethasone, methotrexate, ifosfamide, L-asparaginase, and etoposide

Edited by A Mavragani; submitted 16.12.24; peer-reviewed by JCL Chow, M Larkins; revised version received 21.03.25; accepted 07.04.25; published 15.05.25.

Please cite as:
Zhao Y, Zhong Y, Xiong M, Huang L, Ye X, He J
Incidence, Prognostic Factors, and Treatment Impact on Survival in Natural Killer/T-Cell Lymphoma: Population-Based Study in the United States
JMIR Form Res 2025;9:e70129
URL: https://formative.jmir.org/2025/1/e70129
doi:10.2196/70129

© Yi Zhao, Yi Zhong, Mengqi Xiong, Linlin Huang, Xiujin Ye, Jingsong He. Originally published in JMIR Formative Research (https://formative.jmir.org), 15.5.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Research Letter

Exploring Public Sentiment on the Repurposing of Ivermectin for COVID-19 Treatment: Cross-Sectional Study Using Twitter Data

Angga Prawira Kautsar[1,2,3*], MBA; Rano Kurnia Sinuraya[1,3,4*], PhD; Jurjen van der Schans[1,5], PhD; Maarten Jacobus Postma[1,3,5], PhD; Auliya A Suwantika[3,4,6], PhD

[1]Unit of Global Health, Department of Health Sciences, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands
[2]Department of Pharmaceutics and Pharmaceutical Technology, Faculty of Pharmacy, Universitas Padjadjaran, Sumedang, Indonesia
[3]Center of Excellence in Higher Education for Pharmaceutical Care Innovation, Universitas Padjadjaran, Bandung, Indonesia
[4]Department of Pharmacology and Clinical Pharmacy, Faculty of Pharmacy, Universitas Padjadjaran, Sumedang, Indonesia
[5]Department of Economics, Econometrics and Finance, Faculty of Economics and Business, University of Groningen, Groningen, Indonesia
[6]Center for Health Technology Assessment, Universitas Padjadjaran, Universitas Padjadjaran, Bandung, Indonesia
[*]these authors contributed equally

Corresponding Author:
Angga Prawira Kautsar, MBA
Unit of Global Health, Department of Health Sciences
University Medical Center Groningen
University of Groningen
Antonius Deusinglaan 1
Groningen, 9713 AV
The Netherlands
Phone: 31 0503611111
Email: angga.prawira@unpad.ac.id

Abstract

A sentiment analysis of 5051 Twitter posts from January 2022 found that 53.4% of them expressed positive views on ivermectin as a COVID-19 treatment, 35.6% of them were neutral, and 11% of them were negative, highlighting the polarized public perception and the need for careful interpretation of social media data in health communication.

(JMIR Form Res 2025;9:e50536) doi:10.2196/50536

KEYWORDS

COVID-19; ivermectin; sentiment analysis; Twitter; social media; public health; misinformation; geolocation analysis

Introduction

As the COVID-19 pandemic evolves, the scientific community confronts the limitations of vaccines due to emerging viral mutations that potentially decrease vaccine efficacy [1], thus necessitating a parallel investigation into additional therapeutic agents. Ivermectin, a well-established, safe drug, has emerged as a repurposed drug candidate due to preliminary studies suggesting its antiviral properties against SARS-CoV-2 in vitro [1,2]. Nonetheless, the scientific debate remains vigorous, with discussions on the drug's appropriate formulation and dosing for potential COVID-19 prophylaxis and treatment [3].

Simultaneously, ivermectin's role has garnered widespread attention on social media, reflecting the public's quest for alternative treatments. Twitter (now X), a hub for real-time public discourse, has become a fertile ground for divergent views on COVID-19 treatment [4,5]. This sentiment analysis focuses on Twitter discussions about ivermectin, showing public opinion that, while not devoid of misinformation risks, these discussions offer an alternative lens to understand the societal pulse on this contentious topic [6]. By examining the sentiments expressed on Twitter, we aim to add nuance to the ongoing discourse, acknowledging the platform's influence on public perception and its implications for health communication strategies.

Methods

Overview

This cross-sectional examined the COVID-19–related sentiments on ivermectin using Twitter data. Primary data were collected from Twitter posts about ivermectin worldwide from January 15 to 22, 2022. We searched posts with the keyword

XSL•FO
RenderX

"ivermectin" and retrieved raw data with various variables. Data cleaning involved lowercasing capital letters, eliminating retweet symbols, removing punctuation marks, and other preprocessing steps to ensure data accuracy [6]. People's sentiments were determined by creating a corpus (body of text) and loading a lexicon dictionary based on the positive and negative words. The sentiment score was calculated with the Bing method using a range of –6 to +6 and considered 3 sentiment types: positive, neutral, and negative [7,8]. The Bing lexicon was chosen for its proven effectiveness and simplicity in extensive dataset analysis. Frequency analysis of single terms (unigrams), word pairs (bigrams), and bigram networks helped identify frequently mentioned terms and explore relationships between words [7,8]. Despite its simplicity, this approach is as robust as more complex methods, as it accurately categorizes sentiments and identifies patterns through unigram and bigram frequency analysis [6]. This computational analysis–focused methodology offers valuable insights into public sentiment, complementing traditional clinical evaluation without extensive statistical validation. Data were mined using Twitter's limited application programming interface (version 2) and analyzed using RStudio (R Foundation) and relevant packages for visualization.

Ethical Considerations

There was no direct connection with Twitter users. We removed all individual data, usernames, IDs, and tweets in the manuscript and supporting material. Therefore, ethical approval was not required.

Results

In total, 5051 ivermectin-related tweets underwent sentiment analysis, revealing a prevailing positive sentiment (53.4%), followed by neutral (35.6%) and negative (11%) sentiments among the analyzed tweets (Figure 1). The analysis identified frequently mentioned positive terms such as "medication," "potentially," and "treatment," while negative terms like "poised," "smuggling," and "toxic" were less common. Notably, analysis of word pairings unveiled strong associations with positive sentiments, further supporting the notion of ivermectin's potential efficacy in COVID-19 treatment. Examining these word pairings facilitated in-depth exploration of sentiment patterns within the bigram network, revealing significant connections and relationships in the analyzed tweets (Figure 2).

Figure 1. Distribution of sentiment score.

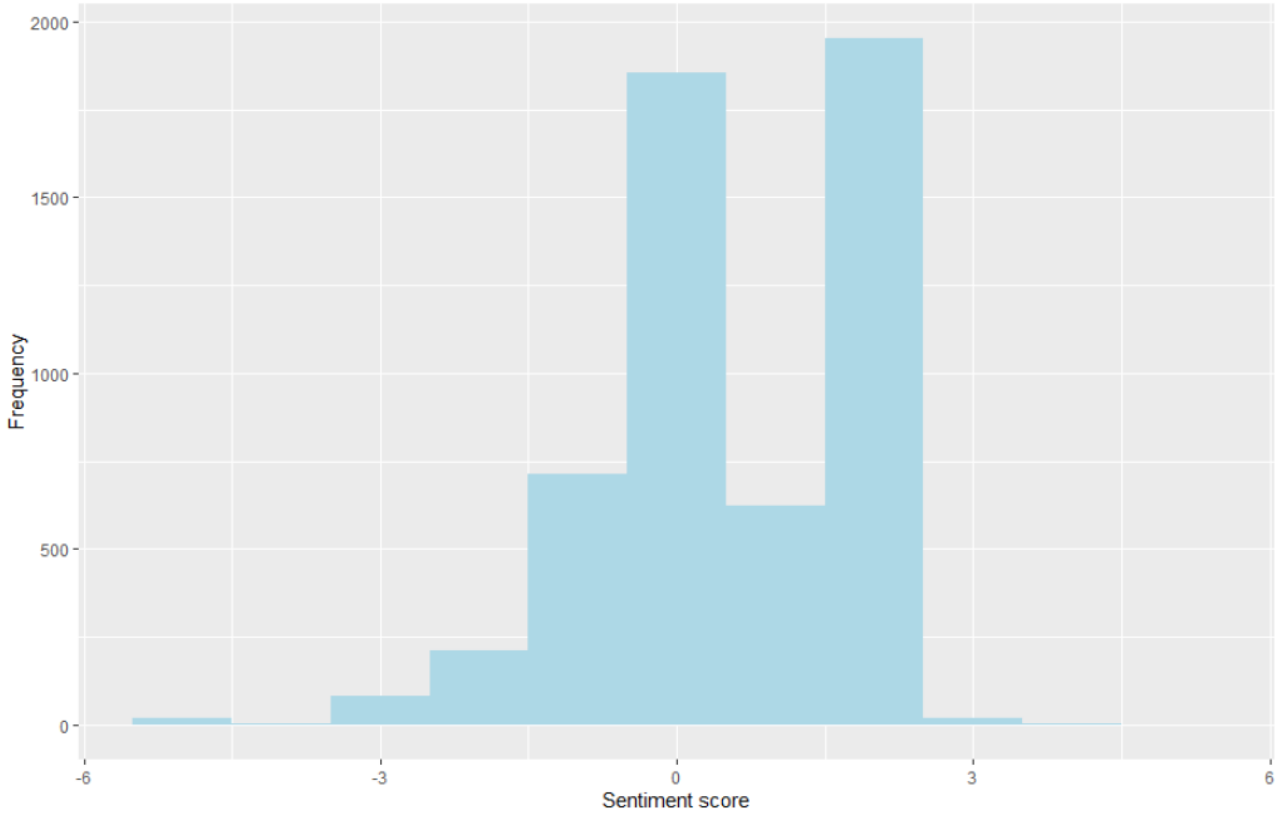

XSL•FO
RenderX

Figure 2. Bigram count network related to tweets ivermectin.

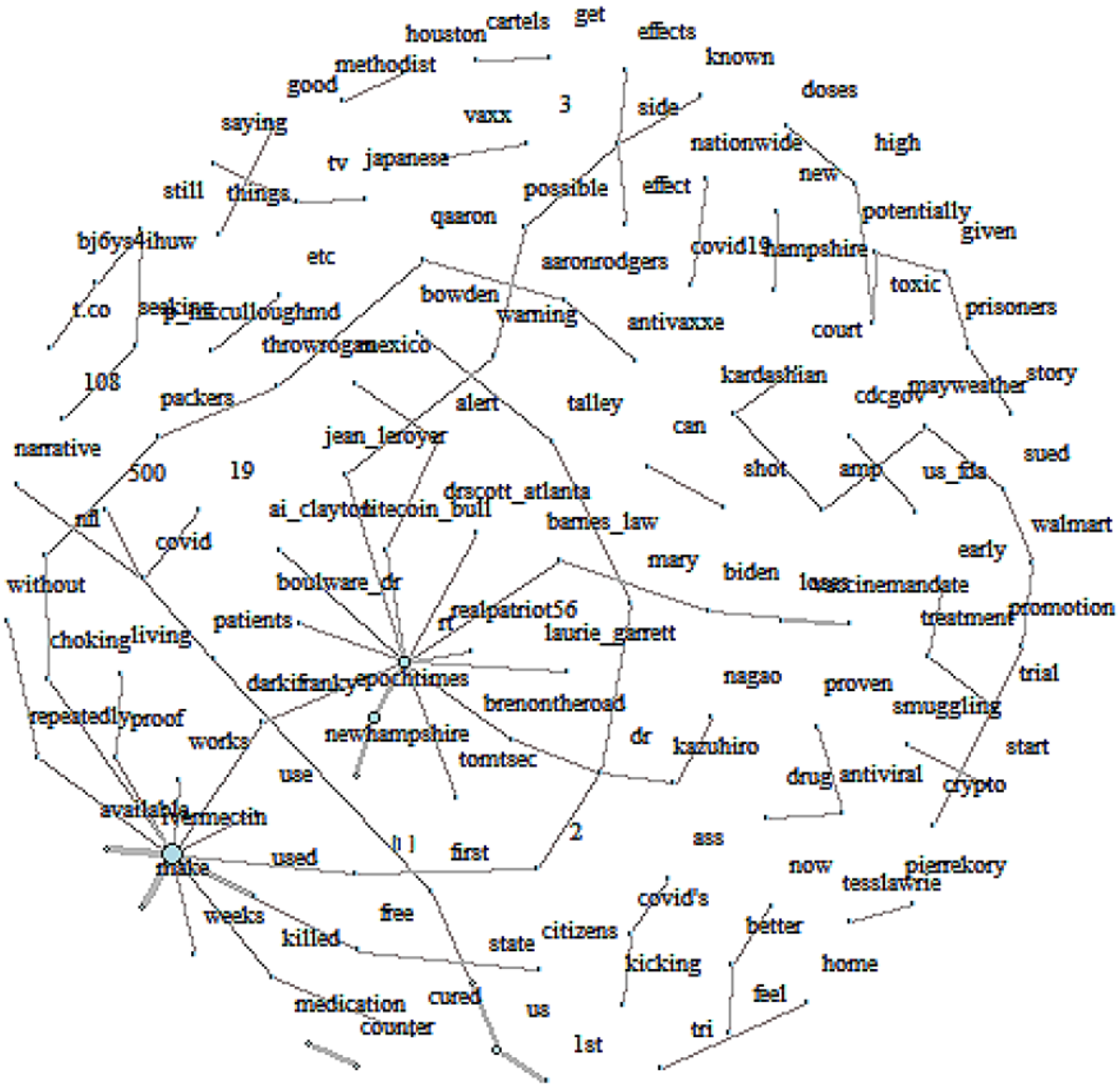

Discussion

Principal Findings

Our analysis shows that tweets about repurposing ivermectin for COVID-19 treatment are predominantly positive. Terms like “medication,” “potentially,” and “treatment” frequently appeared in positive contexts, reinforcing this finding. Phrases such as “ivermectin works,” “available ivermectin,” and “ivermectin medication” were strongly associated with positive sentiment. These results align with previous findings showing ivermectin’s popularity on Twitter for COVID-19 treatment [4,5]. However, some negative sentiment was observed, particularly concerning warnings from the Food and Drug Administration and the limited clinical evidence supporting ivermectin’s efficacy in COVID-19 prevention and treatment [9].

Twitter data have proven valuable in monitoring public responses to the COVID-19 pandemic, as evidenced by a study of millions of SARS-CoV-2–related tweets [10]. This study not only identified dominant topics like new cases, death rates, and preventive measures, but also explored the geographic distribution of sentiments through tweet-embedded geolocation data (code available in Multimedia Appendix 1).

Limitations

Of note, the findings are based on Twitter data, which may not represent the entire population’s sentiments. Therefore, researchers conducted cross-validation of the sources.

Conclusions

This sentiment analysis highlights the polarized perception of ivermectin in COVID-19–related discourse and reflects broader public health debates. Given the regulatory advisories and the limited clinical evidence, these public sentiments must be interpreted cautiously.

XSL•FO
RenderX

Acknowledgments

We acknowledge receiving the Ministry of Higher Education, Science, and Technology Overseas Postgraduate Education scholarship (DIKTI BPP-LN grant T/915/D3.2/KD.02.01/2019) awarded to AK. Despite this financial support, we have diligently maintained transparency and objectivity in our work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Code in R.
[ZIP File (Zip Archive), 5 KB - formative_v9i1e50536_app1.zip]

References

1. Jiang S. Don't rush to deploy COVID-19 vaccines and drugs without sufficient safety guarantees. Nature 2020 Mar 16;579(7799):321-321. [doi: 10.1038/d41586-020-00751-9] [Medline: 32179860]
2. Caly L, Druce JD, Catton MG, Jans DA, Wagstaff KM. The FDA-approved drug ivermectin inhibits the replication of SARS-CoV-2 in vitro. Antiviral Res 2020 Jun;178:104787 [FREE Full text] [doi: 10.1016/j.antiviral.2020.104787] [Medline: 32251768]
3. The WP. NIH trial may settle debate over ivermectin as a covid-19 treatment. The Washington Post. URL: https://www.washingtonpost.com/health/2021/04/08/ivermectin-covid-drug/ [accessed 2023-03-06]
4. de Melo T, Figueiredo CMS. Comparing news articles and tweets about COVID-19 in Brazil: sentiment analysis and topic modeling approach. JMIR Public Health Surveill 2021 Feb 10;7(2):e24585 [FREE Full text] [doi: 10.2196/24585] [Medline: 33480853]
5. Garcia K, Berton L. Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA. Appl Soft Comput 2021 Mar;101:107057 [FREE Full text] [doi: 10.1016/j.asoc.2020.107057] [Medline: 33519326]
6. Hill S, Scott R. Developing an Approach to Harvesting, Cleaning, and Analyzing Data from Twitter Using R. Inf Sys Educ J 2017;15(3):15.
7. Liu B, Hu M, Cheng J. Opinion observer: analyzing and comparing opinions on the Web. 2005 Presented at: WWW '05: Proceedings of the 14th international conference on World Wide Web; May 10-14, 2005; Chiba, Japan p. 342-351. [doi: 10.1145/1060745.1060797]
8. Liu B. Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. Cambridge, United Kingdom: Cambridge University Press; 2020.
9. Why You Should Not Use Ivermectin to Treat or Prevent COVID-19. Food and Drug Administration. URL: https://nycourts.gov/reporter/webdocs/Why-You-Should-Not-Use-Ivermectin-to-Treat-or-Prevent-COVID-19.pdf [accessed 2022-03-01]
10. Xue J, Chen J, Hu R, Chen C, Zheng C, Su Y, et al. Twitter discussions and emotions about the COVID-19 pandemic: machine learning approach. J Med Internet Res 2020 Nov 25;22(11):e20550 [FREE Full text] [doi: 10.2196/20550] [Medline: 33119535]

Edited by A Mavragani; submitted 05.07.23; peer-reviewed by R Gore, J Plasek, A Naser, B Cao; comments to author 16.11.23; revised version received 12.02.24; accepted 09.09.24; published 27.03.25.

Please cite as:
Kautsar AP, Sinuraya RK, van der Schans J, Postma MJ, Suwantika AA
Exploring Public Sentiment on the Repurposing of Ivermectin for COVID-19 Treatment: Cross-Sectional Study Using Twitter Data
JMIR Form Res 2025;9:e50536
URL: https://formative.jmir.org/2025/1/e50536
doi:10.2196/50536
PMID:

©Angga Prawira Kautsar, Rano Kurnia Sinuraya, Jurjen van der Schans, Maarten Jacobus Postma, Auliya A Suwantika. Originally published in JMIR Formative Research (https://formative.jmir.org), 27.03.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Research Letter

Identifying Themes for an Initial Beta Version of a Mobile Health App for Latino and Native Hawaiian/Pacific Islander Communities: Co-Design and Community-Based Participatory Research in a Code to Community Study

Christina K Holub[1*], MPH, PhD; Amy L Barrera[1*], MPH; Rosalva Romero Gonzalez[2*], BA; Diane Hoang[1*], MPH; Luna Prieto[2*], BA; Samuelu Fesili[1*], BA; Tiana Smith[1*], BA; Harleen Kaur[1*], MPH; Cassandra Surban[1*], MPH; Michael Markidis[*], BS; Tana Lepule[3*], BA; Konane Martinez[2*], PhD

[1]Department of Public Health, California State University, San Marcos, San Marcos, CA, United States
[2]National Latino Research Center, California State University, San Marcos, San Marcos, CA, United States
[3]Pacific Islander Collective San Diego, San Diego, CA, United States
[*]all authors contributed equally

Corresponding Author:
Christina K Holub, MPH, PhD
Department of Public Health
California State University, San Marcos
333 S Twin Oaks Valley Road
San Marcos, CA, 92096
United States
Phone: 1 760 750 4295
Email: cholub@csusm.edu

Related Article:

This is a corrected version. See correction statement: https://formative.jmir.org/2025/1/e79194

Abstract

Through a co-design and community-based participatory approach with Latino and Native Hawaiian/Pacific Islander communities, we identified key themes for an initial beta version of a mobile health app, including priorities regarding mental health, access to resources, and chronic disease prevention. Social and cultural connectedness and generational approaches emerged as important strategies for successful intervention design and long-term adoption.

(JMIR Form Res 2025;9:e76178) doi:10.2196/76178

KEYWORDS

community co-design; community-based participatory research; CBPR; mobile health; mHealth

Introduction

Adopting innovations has become increasingly important, driven by postpandemic growth in the use of technology for wellness and health care [1]. New social norms, attitudes, and behavior patterns impact health and participation in health behavior interventions. Latino and Native Hawaiian/Pacific Islander (NHPI) populations, among the fastest-growing minority US populations [2], experience exacerbated health disparities. Heart disease, cancer, and diabetes are leading causes of death in these populations, and both have high rates of obesity [3]. Latino and NHPI populations are less likely to receive mental health services [3], even though stigmas have decreased [4].

Formative research is essential to informing culturally tailored and innovative strategies to reduce health disparities and address community priorities. Preliminary analysis of PIC Health (Pacific Islander Community Health/Elige Salud) study data will inform the first phase of developing a beta version of a mobile health (mHealth) app. This will enable collection of detailed feedback on the app's acceptability, feasibility, and general design from community leaders, members, and other

XSL•FO
RenderX

stakeholders. This study aims to identify community priorities and general themes to build the mHealth beta app.

Methods

Through a co-design and community-based participatory research (CBPR) approach, we partnered with Latino and NHPI communities to conduct 12 key informant interviews, 4 focus groups, and 2 community stakeholder meetings, divided equally across the groups. Participants were recruited by convenience sampling at cultural organizations, clubs, and community events. Latino-identifying participants (8 female, 3 male) were primarily of Mexican origin and NHPI participants (8 female, 7 male) included Samoan, Native Hawaiian, and Chamorro people. All participants resided in North County, San Diego, and were older than 18 years (range: 18-69 years).

Key informant interviews and focus groups were recorded, transcribed, and uploaded into ATLAS.ti for qualitative coding and analysis. At least two research team members coded, verified, and extracted themes from the transcripts, which were analyzed using directed content analysis to identify and categorize meaningful information about participants' opinions, perceptions, and experiences [5]. Coding strategies included the selection of responses for emergent themes and a priori categories [6]. Emergent themes were verified through team discussions and the coding framework. Once themes were identified, the research team worked with a senior software engineer and computer science students to design a progressive web application (PWA; a website that behaves like a mobile app) as an initial beta mHealth app. The research was deemed exempt by the California State University, San Marcos institutional review board, with informed consent waived (1858351). Data were anonymized and participants received US $30 gift cards for interviews and US $20 gift cards for focus groups.

Results

Across all participants, three major themes were identified for community health priorities: (1) mental health, (2) access to resources and health education, and (3) attention to chronic disease prevention (eg, diabetes, cancer, heart disease). Latino participants emphasized access to health resources and NHPI participants emphasized mental health. In response to questions about improving intervention success, both communities identified two major themes: (1) social and cultural connectedness and (2) attention to generational approaches (Table 1). For the latter, Latino participants identified a need for intergenerational strategies (across generations), while some NHPI participants mentioned multigenerational strategies (tailoring to specific generations).

Based on community stakeholder meetings, theme prioritization, and adaptability to different cultures, the beta mHealth app was created in the PWA format. Figure 1 depicts the alignment between key themes and their corresponding app components.

XSL•FO
RenderX

Table 1. Themes, definitions, and participant quote examples.

Theme	Definition; participant describes...	Example
Community priorities		
Mental health	Psychological and emotional well-being; mental health	• "[W]hen I'm thinking about the young people that we've worked with is, um, mental health and access to mental health." [NHPI[a] participant] • "Also, I feel like there's just like a stigma with like mental health, you know, sometimes with the community, with like behavioral health and stuff like that..." [Latino participant]
Access to health resources	Need for access to health services or options to improve approaches to health services; not only the first access but also continuous service	• "If we just had access to education or if we just had better trust with some of these things – [family member] would not have passed away." [NHPI participant] • "It's still a lack of access. Just a lack of knowledge of what, like, what services are out there." [Latino participant]
Chronic disease prevention	Diseases or conditions like hypertension, diabetes, obesity, and other chronic diseases; could be mentioned as underlying conditions	• "I see a lot of diabetes related complications." [Latino participant] • "I would say they struggle with like obesity and diabetes and gout heart disease are the big ones." [NHPI participant]
Intervention success		
Social and cultural connectedness	Specific ways to connect to culture; cultural protocols	• "Definitely reminding people, at least on the Latino side for me with my family, like that connection back to a little more of like our roots and more like, because real, traditional, healthy, organic Mexican food, it's not bad. It's all the additional everything that we end up with, particularly in the US." [Latino participant] • "I love it when I see...programs where they can connect to traditions.... You know this is what our ancestors did, and this is why they did it, or this was their philosophy, like I love it when they can take something that's you know...contemporary. You know, but still connecting it to the past." [NHPI participant]
Generational Approaches		
Intergenerational and multigenerational	Approaches for the different age groups or generations	• "And I think I'd agree with that really having the younger generation, you know…They're the ones that are being exposed to…nutrition classes in school… and really encouraging students to go home and share this with your parents. Go home and share this with…mom and dad or grandma and grandpa." [Latino participant] • "…If you had exclusively for that age group meeting. I don't think putting them all together in a forum would be the ideal." [NHPI participant]

[a]Native Hawaiian/Pacific Islander.

XSL•FO
RenderX

Figure 1. Mobile app theme alignment and components, including social connectedness (community calendar), cultural connection (app design, language, and culturally tailored resources), and community resources (separate tabs for health, education, and a combined resource section for culture, arts, and language).

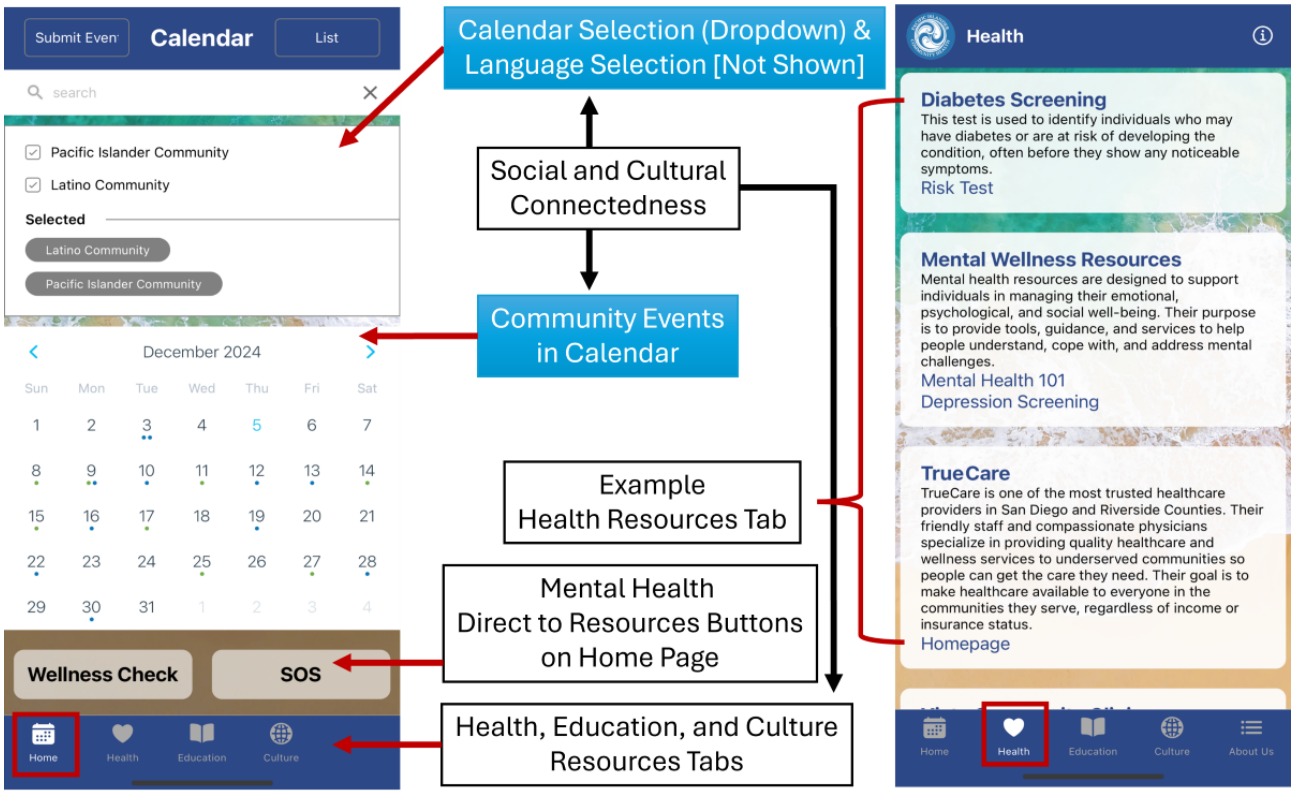

Discussion

Our research identified major themes from qualitative data that translated into components of an initial mHealth app beta designed for the Latino and NHPI communities. The themes identified in the preliminary analysis for community health priorities include mental health, access to resources, and chronic disease prevention. Themes surrounding intervention success include social and cultural connectedness and generational approaches. Many themes overlap between the Latino and NHPI communities, but cultural details and tailoring should differ.

The initial beta incorporates both communities; however, future work may lead to the need for separate mHealth apps or other advanced strategies to tailor information in real time, such as the use of artificial intelligence to populate appropriate resources, although challenges related to artificial intelligence in mHealth must be addressed [7]. The next phase of app development will include gathering additional community insights on usability, acceptability, and design improvements through a mixed methods approach. Feedback and ongoing co-design collaboration in this next phase will result in a final mHealth app for beta testing. Overall, to address health equity, co-design and CBPR approaches with the communities who benefit are essential for adoption of new technologies and greater mHealth intervention success.

Acknowledgments

This work could not be possible without community wisdom, knowledge, and an unwavering passion to improve Latino and Native Hawaiian and Pacific Islander health. We are grateful to Ana Paula Muraro (visiting faculty scholar); the California State University, San Marcos National Latino Research Center; our PIC Health (Pacific Islander Community Health/Elige Salud) team, including our nonauthor team members who supported this initial phase of research (Ana Ardon, Alyssa Lane, Amy Gonzalez, Maja Loba, and Lorenzo Cataquiz); and our first computer science student team (Adam Salter, Nathan Potraz, Billy Vo, Kane Svelan, and Michael Duggan). Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health (award R16GM145519). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author MM is not affiliated with any institution, but is an independent Senior Software Engineer Consultant in San Diego, CA, United States.

XSL•FO
RenderX

Conflicts of Interest

None declared.

References

1. Qiu Y, Huang H, Gai J, De Leo G. The effects of the COVID-19 pandemic on age-based disparities in digital health technology use: secondary analysis of the 2017-2022 Health Information National Trends Survey. J Med Internet Res 2024 Dec 04;26:e65541 [FREE Full text] [doi: 10.2196/65541] [Medline: 39631070]
2. Improved race and ethnicity measures reveal United States population is much more multiracial. United States Census Bureau. 2021. URL: https://www.census.gov/library/stories/2021/08/improved-race-ethnicity-measures-reveal-united-states-population-much-more-multiracial.html [accessed 2025-05-29]
3. Minority population profiles. Office of Minority Health. URL: https://minorityhealth.hhs.gov/minority-population-profiles [accessed 2025-05-29]
4. Pescosolido BA, Halpern-Manners A, Luo L, Perry B. Trends in public stigma of mental illness in the US, 1996-2018. JAMA Netw Open 2021 Dec 01;4(12):e2140202 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.40202] [Medline: 34932103]
5. Hsieh H, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005 Nov;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
6. Ryan GW, Bernard HR. Techniques to identify themes. Field Methods 2003 Feb 01;15(1):85-109. [doi: 10.1177/1525822x02239569]
7. Deniz-Garcia A, Fabelo H, Rodriguez-Almeida AJ, Zamora-Zamorano G, Castro-Fernandez M, Alberiche Ruano MDP, et al. Quality, usability, and effectiveness of mHealth apps and the role of artificial intelligence: current scenario and challenges. J Med Internet Res 2023 May 04;25:e44030 [FREE Full text] [doi: 10.2196/44030] [Medline: 37140973]

Abbreviations

CBPR: community-based participatory research
DCA: directed content analysis
mHealth: mobile health
NHPI: Native Hawaiian/Pacific Islander
PIC Health: Pacific Islander Community Health (Elige Salud)
PWA: progressive web application

Edited by A Mavragani; submitted 17.04.25; peer-reviewed by J Gai; comments to author 21.05.25; revised version received 23.05.25; accepted 25.05.25; published 12.06.25.

Please cite as:
Holub CK, Barrera AL, Romero Gonzalez R, Hoang D, Prieto L, Fesili S, Smith T, Kaur H, Surban C, Markidis M, Lepule T, Martinez K
Identifying Themes for an Initial Beta Version of a Mobile Health App for Latino and Native Hawaiian/Pacific Islander Communities: Co-Design and Community-Based Participatory Research in a Code to Community Study
JMIR Form Res 2025;9:e76178
URL: https://formative.jmir.org/2025/1/e76178
doi:10.2196/76178
PMID:40424151

©Christina K Holub, Amy L Barrera, Rosalva Romero Gonzalez, Diane Hoang, Luna Prieto, Samuelu Fesili, Tiana Smith, Harleen Kaur, Cassandra Surban, Michael Markidis, Tana Lepule, Konane Martinez. Originally published in JMIR Formative Research (https://formative.jmir.org), 12.06.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Research Letter

Use of and Attitudes Toward Technology Among Young People Living With HIV in San Francisco: Cross-Sectional Study

Sean Arayasirikul[1], PhD; Caitlin Turner[2], MPH; Dillon Trujillo[3], MPH; Jarett Maycott[4], MPH

[1]The Legacy Center, Joe C Wen School of Population and Public Health, University of California, Irvine, Irvine, CA, United States
[2]Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, CA, United States
[3]Fielding School of Public Health, Department of Community Health Sciences, University of California, Los Angeles, Los Angeles, CA, United States
[4]Center for Public Health Research, San Francisco Department of Public Health, San Francisco, CA, United States

Corresponding Author:
Sean Arayasirikul, PhD
The Legacy Center
Joe C Wen School of Population and Public Health
University of California, Irvine
856 Health Sciences Quad
Irvine, CA, 92697-3957
United States
Phone: 1 949 824 2680
Email: sean.arayasirikul@uci.edu

Abstract

Digital technology is an important tool for engaging and delivering care to and for young people living with HIV. This research letter examines how 120 out-of-care young people living with HIV in San Francisco use digital technology, the attitudes they have toward technology, and the anxiety they feel about being without technology. Our findings emphasize the importance of being aware of the unintended consequences of digital health interventions.

(JMIR Form Res 2025;9:e81845) doi:10.2196/81845

KEYWORDS

HIV/AIDS; young people living with HIV; digital health; mental health; technology use

Introduction

The COVID-19 pandemic catalyzed a digital shift in our work and social engagements. Our online world has expanded immensely, with smartphone use increasing from 15% before the pandemic to 36% and tablet/PC use from 2% to 22% [1]. Concerns about this increased screentime and adverse mental health outcomes, especially among young adults, persist despite research in the 1990s and early 2000s [2]. The scientific literature around technology and mental health offers mixed support [2]. Research on the effectiveness of digital and mobile health interventions continues to increase, and the results are promising for many health conditions, especially among young adults with chronic conditions [3]. In this research letter, we present findings on digital technology use among a sample of young people living with HIV and describe their attitudes toward technology and anxiety about being without technology.

Methods

Study Design and Participants

We conducted a secondary data analysis of baseline data from a digital HIV navigation intervention study that used text messaging to improve HIV care linkage, retention, and viral suppression among 120 out-of-care young people living with HIV in San Francisco, California. Participants were eligible if they identified as a man who has sex with men or a trans woman, were aged between 18 and 34 years, resided in San Francisco, and were not engaged in HIV care. A digital navigator delivered motivational interviewing, social support, health education, and referrals to participants through text messaging over a period of 6 months. Study procedures were described in a prior manuscript [4].

Measures and Analysis

We analyzed the following sociodemographic information: age, gender identity, race/ethnicity, education, employment status, and current living situation. To assess use of the internet, mobile

XSL•FO
RenderX

phones, and social networking sites [5], participants were asked how often (once a week or less, several times a week, once a day, or several times a day or more) they used email, text messaging, or apps; searched for information; checked social media; or sought out sexual partners or relationships on social networking sites. Positive and negative attitudes toward technology and dependence on technology [5] were assessed using 6, 4, and 4 dichotomous (yes/no) items, respectively. For each of these 3 constructs, items were summed (1/0) and high and low categories created using the median score as a cutoff. To measure psychological distress [6], a composite score was created by summing responses to items ranging from "none of the time" (0) to "all of the time" (5), detailing the frequencies of the following in the last month: feeling calm and peaceful, having a lot of energy, feeling downhearted and blue, and physical health or emotional problems interfering with social activities. Descriptive statistics were used to characterize variables. No missing data were recorded.

Ethical Considerations

The study protocol was approved by the Institutional Review Board of the University of California, San Francisco (16-19675). Participants provided signed informed consent and were remunerated US $50 for completing the baseline assessment. Data for this secondary analysis were deidentified to protect participants' privacy and confidentiality.

Results

The sample was racially and ethnically diverse, with a majority of participants reporting some college or more, unstable housing, and unemployment (56.7%, 67.5%, and 64.2%, respectively). A majority of participants (87.5%) reported using text messaging several times a day or more, followed by mobile phone apps, searching for information on a mobile phone, checking social media, and seeking sexual partners and or relationships on social networking sites (80.8%, 80.0%, 61.7%, and 25.8%, respectively). While most participants (93.3%) reported high levels of positive attitudes toward technology, more than a third (39.2%) reported high levels of negative attitudes toward technology. Many (63.3%) reported high levels of anxiety being without technology. A majority of participants reported being anxious without their cell phone and without internet access (74.2% and 66.7%, respectively). Slightly more than half (53.3%) reported feeling anxious about their personal health information being available online, and nearly two-thirds (65.0%) reported being dependent on technology.

XSL•FO
RenderX

Table 1. Demographic characteristics and use of the internet, mobile phones, and social networking sites of a study sample of young people living with HIV in San Francisco, 2016-2019 (N=120; mean age 27.8, SD 4.1 years).

	Participants, n (%)
Demographics	
Gender identity	
Trans woman	17 (14.2)
Cisgender man	103 (85.8)
Race/ethnicity	
Black, non-Hispanic/Latinx	22 (18.3)
Hispanic/Latinx	38 (31.7)
Multiple races, non-Hispanic/Latinx	28 (23.3)
White, non-Hispanic/Latinx	32 (26.7)
Education	
High school/GED[a] or less	52 (43.3)
Some college or more	68 (56.7)
Employment status	
Unemployed	77 (64.2)
Part-time employed	24 (20.0)
Full-time employed	19 (15.8)
Current living situation	
Unstable	81 (67.5)
Stable	39 (32.5)
Use of the internet, mobile phones, and social networking sites	
How often do you send, receive and read e-mails (not including spam or junk mail)? (Choose one)	
Once a week or less	24 (20.0)
Several times a week	23 (19.2)
Once a day	14 (11.7)
Several times a day or more	59 (49.2)
How often do you send and receive text messages on a mobile phone?	
Once a week or less	8 (6.7)
Several times a week	3 (2.5)
Once a day	4 (3.3)
Several times a day or more	105 (87.5)
How often do you use apps (for any purpose) on a mobile phone?	
Once a week or less	12 (10.0)
Several times a week	6 (5.0)
Once a day	5 (4.2)
Several times a day or more	97 (80.8)
How often do you search for information with a mobile phone?	
Once a week or less	9 (7.5)
Several times a week	10 (8.3)
Once a day	5 (4.2)
Several times a day or more	96 (80.0)
How often do you check your social media page from your smartphone?	

XSL•FO
RenderX

	Participants, n (%)
Once a week or less	27 (22.5)
Several times a week	12 (10.0)
Once a day	7 (5.8)
Several times a day or more	74 (61.7)
How often do you seek sexual partners and or relationships on social networking sites?	
Once a week or less	72 (60.0)
Several times a week	16 (13.3)
Once a day	1 (0.8)
Several times a day or more	31 (25.8)

[a]General Educational Development.

XSL•FO
RenderX

Table 2. Attitudes toward technology and anxiety about being without technology among young people living with HIV in San Francisco, 2016-2019 (N=120).

	Participants, n (%)
Positive attitudes toward technology	
I feel it is important to be able to find any information whenever I want online.	
No	15 (12.5)
Yes	105 (87.5)
I feel it is important to be able to access my personal health information whenever I want online.	
No	16 (13.3)
Yes	104 (86.7)
I feel it is important to be able to access the Internet any time I want.	
No	12 (10.0)
Yes	108 (90.0)
I think it is important to keep up with the latest trends in technology.	
No	29 (24.2)
Yes	91 (75.8)
Technology will provide solutions to many of our problems.	
No	34 (28.3)
Yes	86 (71.7)
With technology anything is possible.	
No	34 (28.3)
Yes	86 (71.7)
I feel that I get more accomplished because of technology.	
No	31 (25.8)
Yes	89 (74.2)
Mean composite score (mean score 5.6, SD 1.8)	
Low	8 (6.7)
High	112 (93.3)
Negative attitudes toward technology	
New technology makes people waste too much time.	
No	68 (56.7)
Yes	52 (43.3)
New technology makes life more complicated.	
No	80 (66.7)
Yes	40 (33.3)
New technology makes people more isolated.	
No	56 (46.7)
Yes	64 (53.3)
Mean composite score (mean score 1.3, SD 1.2)	
Low	73 (60.8)
High	47 (39.2)
Anxiety about being without technology	
I get anxious when I don’t have my cell phone.	
No	31 (25.8)

XSL•FO
RenderX

	Participants, n (%)
Yes	89 (74.2)
I get anxious when I don't have the Internet available to me.	
No	40 (33.3)
Yes	80 (66.7)
I am anxious about having my personal health information available online.	
No	64 (53.3)
Yes	56 (46.7)
I am dependent on my technology.	
No	42 (35.0)
Yes	78 (65.0)
Mean composite score (mean score 2.5, SD 1.5)	
Low	44 (36.7)
High	76 (63.3)

Discussion

We found that many young people living with HIV used mobile phones, social media, apps, and social networking sites on a regular basis. This high use may be a factor in participants' experience of anxiety being without these digital technologies. Other research has found that technology dependence is an important consideration to shape the care of young people [7]. Furthermore, technology can exacerbate isolation, negative social comparison, loneliness, and other poor mental health outcomes [8]. This study has limited generalizability due to its study design and its focus on HIV care engagement among young people living with HIV. Despite this, it is critical to consider not only how digital technology can improve health, but also its influence on technology dependence and anxiety as unintended consequences [9].

Acknowledgments

The authors would like to thank all participants in the study. This work was funded by the Health Resources and Services Administration (award H97HA28895).

Data Availability

The datasets generated or analyzed during this study are not publicly available as they contain details that could be used to identify participants but are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

References

1. Werling AM, Walitza S, Drechsler R. Impact of the COVID-19 lockdown on screen media use in patients referred for ADHD to child and adolescent psychiatry: an introduction to problematic use of the internet in ADHD and results of a survey. J Neural Transm (Vienna) 2021 Jul;128(7):1033-1043 [FREE Full text] [doi: 10.1007/s00702-021-02332-0] [Medline: 33885969]
2. George MJ, Russell MA, Piontak JR, Odgers CL. Concurrent and subsequent associations between daily digital technology use and high-risk adolescents' mental health symptoms. Child Dev 2018 Jan;89(1):78-88 [FREE Full text] [doi: 10.1111/cdev.12819] [Medline: 28466466]
3. Rathbone AL, Prescott J. The use of mobile apps and SMS messaging as physical and mental health interventions: systematic review. J Med Internet Res 2017 Aug 24;19(8):e295 [FREE Full text] [doi: 10.2196/jmir.7740] [Medline: 28838887]
4. Arayasirikul S, Trujillo D, Turner CM, Le V, Wilson EC. Implementing a digital HIV care navigation intervention (Health eNav): protocol for a feasibility study. JMIR Res Protoc 2019 Nov 08;8(11):e16406 [FREE Full text] [doi: 10.2196/16406] [Medline: 31702561]
5. Rosen LD, Whaling K, Carrier LM, Cheever NA, Rokkum J. The Media and Technology Usage and Attitudes Scale: An empirical investigation. Comput Human Behav 2013 Nov 01;29(6):2501-2511 [FREE Full text] [doi: 10.1016/j.chb.2013.06.006] [Medline: 25722534]

XSL•FO
RenderX

6. Ware J, Kosinski M, Keller SD. A 12-Item Short-Form Health Survey: construction of scales and preliminary tests of reliability and validity. Med Care 1996 Mar;34(3):220-233. [doi: 10.1097/00005650-199603000-00003] [Medline: 8628042]
7. Toly VB, Blanchette JE, Al-Shammari T, Musil CM. Caring for technology-dependent children at home: problems and solutions identified by mothers. Appl Nurs Res 2019 Dec;50:151195 [FREE Full text] [doi: 10.1016/j.apnr.2019.151195] [Medline: 31668894]
8. Ponnusamy S, Iranmanesh M, Foroughi B, Hyun S. Drivers and outcomes of Instagram addiction: psychological well-being as moderator. Comput Human Behav 2020 Jun;107:106294 [FREE Full text] [doi: 10.1016/j.chb.2020.106294]
9. Cao W, Cao X, Sutherland AD. Planning for the unexpected and unintended effects of mHealth interventions: systematic review. J Med Internet Res 2025 Aug 07;27:e68909 [FREE Full text] [doi: 10.2196/68909] [Medline: 40774342]

Edited by A Mavragani; submitted 04.08.25; peer-reviewed by J Abua, L Meena; comments to author 11.08.25; revised version received 12.08.25; accepted 30.08.25; published 15.09.25.

Please cite as:
Arayasirikul S, Turner C, Trujillo D, Maycott J
Use of and Attitudes Toward Technology Among Young People Living With HIV in San Francisco: Cross-Sectional Study
JMIR Form Res 2025;9:e81845
URL: https://formative.jmir.org/2025/1/e81845
doi:10.2196/81845
PMID:40952781

©Sean Arayasirikul, Caitlin Turner, Dillon Trujillo, Jarett Maycott. Originally published in JMIR Formative Research (https://formative.jmir.org), 15.09.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Research Letter

Feasibility of a Virtual Reality Intervention Protocol to Improve Cognitive and Behavioral Skills in Older Adults at Increased Risk of Developing Dementia

Oshadi Jayakody[1], PhD; Mirnova Ceïde[2], MD; Joe Verghese[3], MD; Robert Carrera[4], MS; Hussain Doriwala[4], BSc; Sunil Agrawal[4], PhD; Judy Pa[5], PhD; Helena Blumen[3], PhD

[1]Department of Medicine, Albert Einstein College of Medicine, Bronx, NY, United States
[2]Department of Psychiatry, Albert Einstein College of Medicine, Bronx, NY, United States
[3]Department of Neurology, Stony Brook Medicine, Stony Brook, NY, United States
[4]Department of Mechanical Engineering, Columbia University, New York, NY, United States
[5]Department of Neurosciences, University of California, San Diego, San Diego, CA, United States

Corresponding Author:
Oshadi Jayakody, PhD
Department of Medicine
Albert Einstein College of Medicine
1225 Morris Park Avenue
Bronx, NY, 10461
United States
Phone: 1 718 430 7081
Email: shanika.jayakody@einsteinmed.edu

Abstract

This pilot study offers preliminary evidence that a virtual meal-preparation task is feasible for older adults and highlights that the community engagement studios are an effective approach to generate community-informed strategies to enhance intervention designs and reach.

(JMIR Form Res 2025;9:e77111) doi:10.2196/77111

KEYWORDS

pilot study; community engagement studio; Alzheimer disease; clinical trials; skills training

Introduction

Virtual reality (VR) offers immersive, interactive environments that simulate real-world tasks—providing opportunities for cognitive stimulation through increasing task complexity and real-time feedback. While prior studies report feasibility and cognitive benefits of VR interventions in older adults (ie, improved memory and attention [1-3]), transfer of benefits to real-world functioning is limited [4,5]. Additionally, despite facing disproportionately high dementia risk, older adults from minoritized backgrounds were underrepresented in prior interventions [6]. We are developing a VR intervention to promote cognitive and daily functioning in Black and Hispanic older adults. This study aimed to generate preliminary data to support the proposed large-scale feasibility study by testing the feasibility and safety of an ecologically valid virtual meal-preparing task (involving planning, memory, and visuospatial skills) and obtaining feedback from community experts to inform intervention refinement.

Methods

Overview

Eight adults aged ≥65 years at risk for dementia (AD8 Dementia Screening Interview score ≥1 [7]) and 8 young adults (aged 18-30 years) completed a single-session virtual meal-preparing task, developed in Unity 3D (2022 LTD version; Unity Technologies; Figure 1). Seated participants wore an HTC Vive headset and used wireless hand controllers. To assist navigation, 3D cues (eg, a stovetop) were placed. Personalized training sessions were provided to reduce usability barriers. The task involved 5 sequential steps engaging goal-setting, working memory, planning, visuospatial functions, and memory—mimicking a real-world kitchen. Task completion time, errors, motion sickness [8], simulator sickness [9],

XSL•FO
RenderX

acceptability (Likert ratings on headset comfort and instruction clarity), and self-efficacy were compared across groups to understand older adults' ability to engage in VR. Given the small sample, we report descriptive statistics to summarize outcomes.

Figure 1. The virtual kitchen.

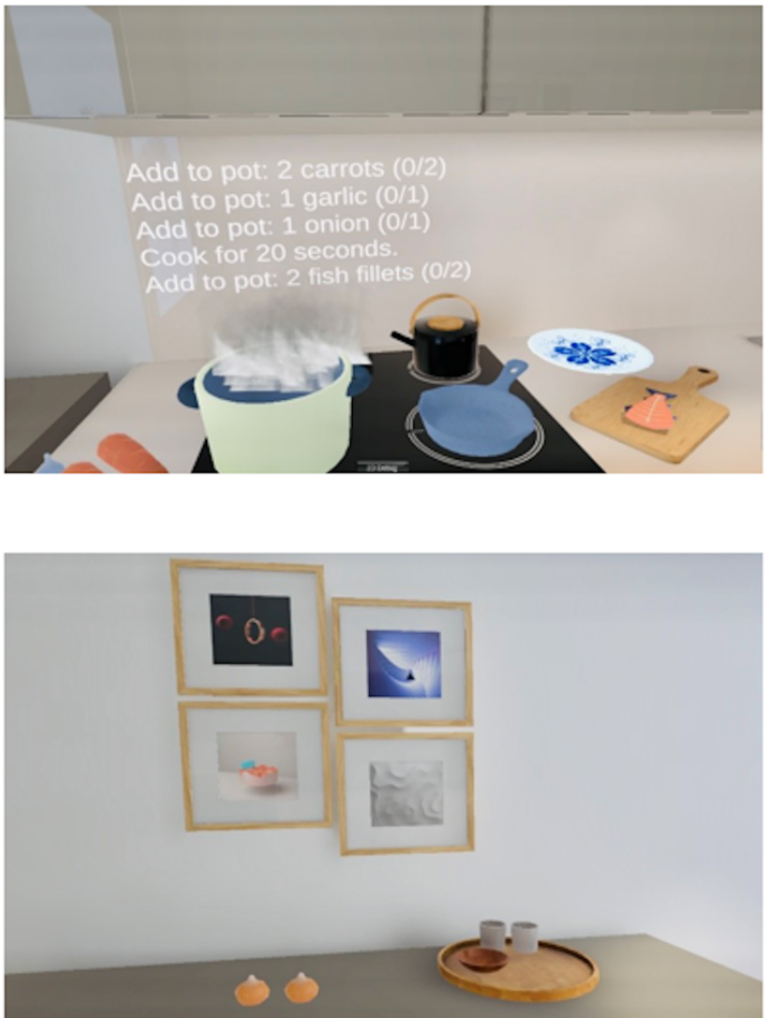

A 1-hour community engagement studio (CES) was held at a community center with 16 older Black/Hispanic adults. The project lead hosted an information session to invite participants and delivered a 10-minute, jargon-free presentation outlining the design and goals of the proposed intervention. A bilingual facilitator guided a prestructured discussion to obtain feedback on the study's significance, recruitment strategies, eligibility criteria, and duration. Responses were documented in real time and thematically coded by 2 investigators to identify key insights.

Ethical Considerations

Ethics approval was obtained from the institutional review board of Albert Einstein College of Medicine, New York (2023-14856). Written informed consent was obtained from all participants prior to enrollment. Participant confidentiality was maintained through the use of deidentified data.

Results

This study included 8 older adults (mean age 76.2, SD 7.9 years) and 8 young adults (mean age 25.0, SD 2.6 years), comprising, respectively, 5 (62%) and 4 (50%) women, 3 (38%) and 1 (12%) Hispanic individuals, and 2 (25%) and 1 (12%) non-Hispanic Black individuals. Older adults had a mean 12.8 (SD 8.3) years of education and young adults had 17.1 (SD 1.6) years. Older adults showed longer completion times (235.0, SD 99.6 vs 74.7, SD 18.9 seconds) and more errors (1.2, SD 1.3 vs 0.1, SD 0.3). Simulator sickness was minimal: pretest scores were 0.03 (SD 0.05) and 0.16 (SD 0.15); posttest scores were 0.02 (SD 0.03) and 0.07 (SD 0.06) for older and young adults. No safety concerns were reported except minor sweating due to the headset.

Acceptability scores (range 0-16) averaged 13.3 (SD 1.5) for older adults and 14.3 (SD 1.18) for young adults, and 6 (75%) older and 7 (88%) young adults reported enjoying the task. All participants expressed high self-efficacy for future VR use. The virtual environment was rated as realistic by 5 (62%) older adults; 2 (25%) somewhat agreed, and 1 (12%) disagreed.

Textbox 1 shows thematic insights from the CES. Community experts viewed the intervention as highly relevant. Health fairs and educational talks were identified as effective recruitment strategies. The bilingual flyer was well received, with comments to improve visual appeal. Eligibility for magnetic resonance

XSL•FO
RenderX

imaging was considered a potential barrier. The proposed intervention duration was not seen as burdensome, though interventions conducted at community centers were preferred.

Textbox 1. A summary of key insights from the community engagement studio.

Significance of the intervention

- Participants felt that the proposed intervention addresses an important question (improving cognitive and behavioral skills to delay the onset of Alzheimer disease and related dementias), considering their lived experience of family members and friends affected by Alzheimer disease and related dementias.

Recruitment strategies and material

- Suggested recruitment strategies included community outreach through educational talks.
- Magnetic resonance imaging at health fairs (which is an outcome assessment in the proposed intervention) might limit recruitment due to contraindications.
- Participants provided input to make the study flyer more appealing; they appreciated that the information is available in both English and Spanish.

Study duration

- Committing to the intervention duration (twice weekly for 4 weeks) was not considered burdensome.
- Suggestions to increase retention and adherence included providing free transportation, regular reminders, and refreshments.
- Most participants preferred that the intervention be conducted in a community setting rather than an academic study center.

Use of technology and intervention appropriateness

- Community experts expressed interest in using virtual reality and the use of an electronic pillbox (a suggested outcome assessment in the proposed intervention).
- The experts also agreed that the proposed virtual tasks were appropriate, culturally relevant, and reflected their daily activities.

Discussion

Principal Findings

This pilot study provides preliminary evidence supporting the feasibility of a virtual meal-preparing task in older adults. While older adults required more guidance, they completed the task in <4 minutes and reported high self-efficacy. The CES allowed us to obtain valuable, community-driven insights to improve recruitment, adherence, and delivery of the proposed intervention.

Prior VR interventions lasting 6 to 12 weeks (2-3 sessions/week; 30-45 minutes each), have shown cognitive benefits in older adults [2,3]. Our findings support progressing toward a similar structure, informed by user data and local community input. The brief training period was crucial in optimizing task completion and building confidence. The absence of simulator sickness—likely due to the seated setup and spatial cues—supports our safety approaches. Minor headset discomfort highlights the need for lighter headsets in future studies. CES insights reaffirm the value of involving community experts early in research development to enhance relevance and effective delivery.

Limitations include the small sample and single-session format, restricting statistical inference and long-term cognitive outcome assessments.

Conclusions

These findings directly informed the next phase of our intervention while highlighting the promise of developing community-informed VR interventions for older adults.

Acknowledgments

We acknowledge the contribution of the study participants, Jewish Association for Services for the Aged staff, Ms Samantha Johnson, and Ms Dachel Sanchez.

Funding

This pilot data collection was supported by an intramural Resnick Gerontology Center Early Career Research Award offered by the Division of Geriatrics, Department of Medicine, Albert Einstein College of Medicine.

Data Availability

Data from the pilot testing of the virtual task are available upon request from the corresponding author.

XSL•FO
RenderX

Conflicts of Interest

None declared.

References

1. Gómez-Cáceres B, Cano-López I, Aliño M, Puig-Perez S. Effectiveness of virtual reality-based neuropsychological interventions in improving cognitive functioning in patients with mild cognitive impairment: A systematic review and meta-analysis. Clin Neuropsychol 2023 Oct;37(7):1337-1370. [doi: 10.1080/13854046.2022.2148283] [Medline: 36416175]
2. Kim O, Pang Y, Kim JH. The effectiveness of virtual reality for people with mild cognitive impairment or dementia: a meta-analysis. BMC Psychiatry 2019 Jul 12;19(1):219 [FREE Full text] [doi: 10.1186/s12888-019-2180-x] [Medline: 31299921]
3. Zhu S, Sui Y, Shen Y, Zhu Y, Ali N, Guo C, et al. Effects of virtual reality intervention on cognition and motor function in older adults with mild cognitive impairment or dementia: a systematic review and meta-analysis. Front Aging Neurosci 2021;13:586999 [FREE Full text] [doi: 10.3389/fnagi.2021.586999] [Medline: 34025384]
4. Kang J, Kim N, Lee S, Woo S, Park G, Yeon B, et al. Effect of cognitive training in fully immersive virtual reality on visuospatial function and frontal-occipital functional connectivity in predementia: randomized controlled trial. J Med Internet Res 2021 May 06;23(5):e24526 [FREE Full text] [doi: 10.2196/24526] [Medline: 33955835]
5. Thapa N, Park H, Yang J, Son H, Jang M, Lee J, et al. The effect of a virtual reality-based intervention program on cognition in older adults with mild cognitive impairment: a randomized control trial. J Clin Med 2020 Apr 29;9(5):1283 [FREE Full text] [doi: 10.3390/jcm9051283] [Medline: 32365533]
6. Hughes T, Flatt J, Fu B, Butters M, Chang C, Ganguli M. Interactive video gaming compared with health education in older adults with mild cognitive impairment: a feasibility study. Int J Geriatr Psychiatry 2014 Sep;29(9):890-898 [FREE Full text] [doi: 10.1002/gps.4075] [Medline: 24452845]
7. Galvin JE, Roe CM, Powlishta KK, Coats MA, Muich SJ, Grant E, et al. The AD8: a brief informant interview to detect dementia. Neurology 2005 Aug 23;65(4):559-564. [doi: 10.1212/01.wnl.0000172958.95282.2a] [Medline: 16116116]
8. Golding JF. Motion sickness susceptibility. Auton Neurosci 2006 Oct 30;129(1-2):67-76. [doi: 10.1016/j.autneu.2006.07.019] [Medline: 16931173]
9. Kennedy R, Lane N, Berbaum K, Lilienthal M. Simulator sickness questionnaire: an enhanced method for quantifying simulator sickness. Int J Aviat Psychol 1993 Jul;3(3):203-220. [doi: 10.1207/s15327108ijap0303_3]

Abbreviations

CES: community engagement studio
VR: virtual reality

Edited by A Mavragani, J Sarvestan; submitted 08.05.25; peer-reviewed by B Lai; comments to author 22.08.25; revised version received 28.08.25; accepted 03.10.25; published 11.12.25.

Please cite as:
Jayakody O, Ceïde M, Verghese J, Carrera R, Doriwala H, Agrawal S, Pa J, Blumen H
Feasibility of a Virtual Reality Intervention Protocol to Improve Cognitive and Behavioral Skills in Older Adults at Increased Risk of Developing Dementia
JMIR Form Res 2025;9:e77111
URL: https://formative.jmir.org/2025/1/e77111
doi:10.2196/77111
PMID:41379535

©Oshadi Jayakody, Mirnova Ceïde, Joe Verghese, Robert Carrera, Hussain Doriwala, Sunil Agrawal, Judy Pa, Helena Blumen. Originally published in JMIR Formative Research (https://formative.jmir.org), 11.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Implementing a Cross-Border Next-Generation Personal Health Record in the Philippines and Taiwan: An Implementation Case Report Using Health Level 7 International Fast Healthcare Interoperability Resources

Hsiu-An Lee[1,2,3], PhD; Jui-Chun Huang[4], MS; Shih-Wun Huang[4], MS; Wei-Han Chen[4], MS; Alvin B Marcelo[2,5], PhD; Miguel Sandino O Aljibe[6*], MD, LME; Chien-Yeh Hsu[2,3,7*], PhD

[1]National Health Research Institutes, The National Institute of Cancer Research, Tainan, Taiwan
[2]Asia eHealth Information Network, Hong Kong, China (Hong Kong)
[3]Standards and Interoperability Lab, Smart Healthcare Center of Excellence, Taipei, Taiwan
[4]Department of Information Management, National Taipei University of Nursing and Health Science, Taipei, Taiwan
[5]Medical Informatics Unit, University of the Philippines, Manila, Philippines
[6]Standards and Interoperability Lab, University of the Philippines, Manila, Philippines
[7]Department of Information Management, National Taipei University of Nursing and Health Science, 365 Mingde Road, Peitou District, Taipei, Taiwan
[*]these authors contributed equally

Corresponding Author:
Chien-Yeh Hsu, PhD
Asia eHealth Information Network, Hong Kong, China (Hong Kong)

Abstract

Background: Disparities in electronic health record systems hinder cross-border continuity of care, particularly where labor mobility and tourism intersect (eg, between the Philippines and Taiwan). Both nations collect claim data, yet neither fully aligns with international standards such as the Health Level 7 International, International Patient Summary (IPS).

Objective: This implementation report aimed to convert health insurance data from Taiwan's My Health Bank (MHB) and the Philippine Health Insurance Corporation's Claim Form 4 (CF4) into a cross-border personal health record (PHR) aligned with the IPS using (Fast Healthcare Interoperability Resources) FHIR standards.

Methods: We mapped each data element from CF4 (n=7 main sections) and MHB (n=12 major data items) to 17 sections of the IPS. We analyzed whether these elements matched IPS requirements (required or recommended or optional) and identified missing fields (eg, device use, social history, and advanced directives). We also designed a FHIR-based integration architecture, addressing system security with OAuth 2.0/SMART on FHIR and proposing a national uptake strategy for accelerating cross-border PHR implementation.

Results: Of the 17 IPS sections, MHB covered 14 sections (82.4%), while CF4 covered 12 sections (70.6%). Both systems lacked sufficient data elements for medical devices, social history (eg, alcohol or tobacco), and advanced directives. We developed an implementation plan focusing on data interoperability, standardization, and privacy or security protocols. We propose a multiphase approach—beginning with the stakeholder engagement and pilot testing in both countries.

Conclusions: Aligning CF4 and MHB data with IPS standards via FHIR can facilitate a robust cross-border next-generation PHR ecosystem. This approach may enhance patient safety, continuity of care, and policy development for the Philippines and Taiwan. Further collaboration, regulatory updates, and public awareness are vital to sustain integration and maximize patient benefits.

(JMIR Form Res 2025;9:e56272) doi:10.2196/56272

KEYWORDS

health insurance; personal health records; international patient summary; HL7 FHIR; cross-border healthcare; implementation study

XSL•FO
RenderX

Introduction

Background

In recent decades, numerous eHealth technologies have become available, as countries have embarked on eHealth initiatives to promote health education and patient-centered care objectives [1]. The usage of personal health records (PHRs) has been associated with several benefits, including enhanced patient-provider relationships, increased patient engagement, improved medication adherence, positive health outcomes (such as better blood pressure and glycemic control), and greater organizational efficiency [2,3].

Under the aim of promoting patient safety, PHR facilitates the continuity of care [4]. When patients move to different countries or seek medical treatment abroad, having access to their health records ensures that health care providers can make informed decisions based on their medical history, allergies, medications, and previous treatments. In emergencies, such as accidents or sudden illnesses while traveling, having access to PHRs can be a matter of life and death. Medical professionals need immediate access to a patient's health information to provide appropriate and timely treatment.

The evolution of PHRs into the next generation represents a significant shift in how individuals engage with and manage their health information. Unlike traditional PHRs, which often mimic digital versions of paper-based records, next-gen PHRs harness advanced technologies and data sources to offer a far more comprehensive and dynamic view of an individual's health. Central to these next-gen PHRs is their emphasis on interoperability, designed to seamlessly integrate and share data with various health care providers, electronic health record (EHR) systems, wearable devices, and other health-related apps [5]. This interoperability ensures that health information is no longer confined within silos but can be readily accessed and shared with the authorized parties. Real-time data updates are a hallmark of next-gen PHRs. They provide continuous streams of health data, including vital signs, lab results, and medication adherence, often integrating with wearable devices and IoT (Internet of Things) sensors for ongoing monitoring [6,7]. Furthermore, these PHRs excel in data aggregation, pulling in health information from a multitude of sources such as EHRs, pharmacies, insurance claims, and even patient-generated data. This holistic view empowers users to track and manage their health more effectively. PHR should allow users to set health goals, track progress, and receive motivational feedback and reminders [8]. Within the broader health care ecosystem, next-generation PHRs often function as central hubs connecting patients with health care providers, pharmacies, laboratories, and insurers. Personalization is a key aspect. These personal health management tailor health recommendations and content to an individual's specific health needs and goals [9]. In addition, users have greater control and ownership over their health data, deciding who can access their information and for what purposes. The next-generation personal health record is designed to empower individuals with greater control, access, and understanding of their health data while fostering improved communication and collaboration within the health care ecosystem. Through the integration of advanced technologies, it introduces a more dynamic and user-centric approach to personal health management.

The data sources for PHRs are diverse, including self-reports, long-term monitoring data, nutritional information, medical records, and more. Currently, common PHRs often separate into 2 kinds of data, self-reported or questionnaire data or clinical medical data. However, in practical applications, integrating this information holds great potential for providing more comprehensive and valuable medical information. In Taiwan, the Health Bank Book serves as a significant data source for PHRs, primarily deriving its content from health insurance reimbursement data [10]. These data are provided by validated sources and typically encompass crucial clinical diagnoses and medical service-related information. Therefore, the release of Health Bank Book data primarily follows an application framework focused on clinical diagnosis and medical services, providing robust reference information.

In recent years, there has been increasing collaboration across the Asia-Pacific region. From the perspectives of population and labor mobility, Taiwan has approximately 12 million workers, among which more than 700,000 are foreign laborers from Thailand, the Philippines, Indonesia, and Vietnam. Among them, around 150,000 are from the Philippines, primarily working in manufacturing (accounting for about 60% of foreign laborers) and domestic services or caregiving (approximately 36%).

Health insurance coverage in Taiwan is one of the labor rights granted to foreign workers, enabling effective cross-border exchange of medical data. This not only ensures the quality of medical care provided abroad but also safeguards the rights and well-being of workers from both sides.

According to the Philippines' Department of Tourism, between January and August 2024, approximately 155,086 Taiwanese tourists visited the Philippines, representing a 17.62% increase compared to the 131,849 visitors during the same period last year. Taiwan ranked sixth among the sources of international tourists to the Philippines, accounting for about 4.19% of the total number of international visitors.

In an increasingly globalized world, people frequently move across borders for work, education, and leisure. International health management ensures that individuals can access health care services and manage their health records seamlessly, regardless of their location. This is vital for expatriates, travelers, and immigrants. After the COVID-19 pandemic, changes in the epidemic have made us realize the importance of international cooperation. In the past, countries developed medical systems and provided technical support, knowledge exchange, and cooperative research with each other.

Studies show that design and adoption vary globally, with standardization efforts (including Health Level 7 International [HL7], Fast Healthcare Interoperability Resources [FHIR]) promising better interoperability [11]. From 2023 onward, new frameworks and policies (eg, 21st Century Cures Act in the United States) further encourage multinational data sharing [12].

XSL•FO
RenderX

However, future developments should move toward more comprehensive and integrated PHRs. This will facilitate the provision of diverse health information, enabling more forward-looking medical decisions while better meeting individual health needs. In this developmental process, we can draw from the experience of the Health Bank to ensure the accuracy and reliability of data sources for PHRs, providing enhanced medical and health management support for individuals.

Objectives

The purpose of this study is to discuss the possibility of converting health insurance declaration data into PHRs and implement them in an internationally common format, and to explore the future adoption model of the international personal health management framework using health insurance data from Taiwan and the Philippines as examples.

Methods

Study Design

This implementation report followed a comparative design to evaluate whether Claim Form 4 (CF4; the Philippines) and My Health Bank (MHB; Taiwan) align with or can be adapted to the International Patient Summary (IPS). We systematically mapped all data elements from both forms to IPS's 17 sections (required, recommended, and optional). We then proposed an architecture based on FHIR to enable cross-border interoperability. Finally, we formulated a multistakeholder national promotion strategy.

Discover the Key Elements of the Next Generation of PHRs

According to changes in international cooperation and medical information architecture, PHRs are becoming more and more important. In addition to data content, there are many issues between system architectures that need to be paid attention to. Some key features and characteristics of next-generation PHRs are given in Textbox 1.

Textbox 1. Key features and characteristics of next-generation personal health records (PHRs).

- Interoperability: PHRs are designed to seamlessly integrate and share data with various health care providers, electronic health record (EHR) systems, wearables, and other health-related applications. This interoperability ensures that health information is not siloed and can be easily accessed and shared with authorized parties [13].
- Real-time data: Provide real-time updates of health data, including vital signs, lab results, and medication adherence. Integration with wearable devices and IoT (Internet of Things) sensors enables continuous data collection and monitoring [14].
- Data aggregation: These PHRs aggregate health data from multiple sources, such as EHRs, pharmacies, insurance claims, and patient-generated data. This comprehensive view allows users to track their health more effectively [15].
- Artificial intelligence (AI) and predictive analytics: Next-gen PHRs may leverage AI and machine learning algorithms to analyze health data. This can help users identify trends, receive personalized health recommendations, and even predict potential health issues [16].
- Security and privacy: Robust security measures are in place to protect sensitive health information. This includes encryption, 2-factor authentication, and compliance with health care privacy regulations like HIPAA (Health Insurance Portability and Accountability Act in the United States).
- Telehealth integration: Next-generation PHRs often include telehealth capabilities, allowing users to schedule virtual appointments with health care providers and access medical advice and consultations remotely [17].
- Health care ecosystem integration: These PHRs may serve as a central hub within a broader health care ecosystem, connecting patients with health care providers, pharmacies, laboratories, and insurers.

Overall, the next-generation personal health record aims to empower individuals with greater control, access, and understanding of their health data while facilitating better communication and collaboration within the health care ecosystem. It leverages advanced technologies to create a more dynamic and user-centric approach to personal health management.

Implementation Framework

This study evaluates and inventories the feasibility of international data interoperability and explores collaborative measures for data application with the Philippines. The overall framework adopts a deductive approach, focusing on the data architecture. The selection process begins by identifying hospital-level or national-level datasets. The design prioritizes mapping essential fields to corresponding fields in international data formats, with all selections and analyses jointly inferred by the 2 research teams.

Our process emphasized (1) content mapping to IPS sections, (2) structural alignment with FHIR resources, (3) designing a security architecture using OAuth 2.0 or SMART on FHIR, and (4) drafting a national adoption strategy. We consulted HL7's official IPS Implementation Guide [11] and relevant FHIR documentation for best practices.

International PHR Adoption

In order to achieve international data exchange and health record sharing, the adoption of internationally open standards and data content will accelerate development. This approach allows different systems and countries to have a common architecture for implementation, and reduces differential communication and comparison between different suppliers.

The adoption of FHIR has gained significant momentum in the health care industry due to several key factors, including interoperability, simplicity, modularity, community support,

XSL•FO
RenderX

government mandates, and scalability. FHIR was developed to address the interoperability challenges in health care. Its standardized approach to data exchange allows health care systems and applications to communicate and share data seamlessly. This is crucial for improving patient care, reducing medical errors, and enhancing the overall efficiency of health care processes.

The straightforward and modern RESTful application programming interface approach was used to implement it. Its resource-based model is based on common web standards like HTTP, JSON, and XML, making it accessible to a wide range of developers and organizations. All structures are designed with a modular approach, where health care data is organized into discrete resources (eg, patients, medications, and observations). This modular approach enables the gradual adoption of FHIR resources based on specific needs, making it flexible and adaptable for different health care environments. The fundamental nature of FHIR allows it to accommodate both small clinics and large health care systems. This makes it an attractive choice for a wide range of health care providers.

Since there is a large group of volunteers and international organizations supporting it, FHIR is backed by a strong and active international community. Mainly developed and maintained by HL7, a global authority in health care standards. The collaborative nature of FHIR development ensures that it remains up-to-date and relevant. In terms of policy, many countries have recognized the importance of health care interoperability and have implemented regulations that encourage or mandate the use of FHIR. For example, the United States has included FHIR as part of the interoperability rules through the 21st Century Cures Act.

An IPS will be adopted in this study. The IPS is a standardized and structured health summary document designed to improve the exchange of essential patient information across borders and different health care systems. It aims to ensure that vital patient data is accessible to health care providers, particularly in emergency situations or when patients are seeking medical care in foreign countries. An IPS document serves as an EHRs extract containing crucial health care information about an individual's care. It adheres to the standards outlined in EN 17269 and ISO 27269 [18], primarily designed to support scenarios involving "unplanned, cross-border care," while its utility extends beyond this scope. The IPS document is crafted to be globally applicable, transcending regional or national boundaries. The IPS dataset is intentionally concise and not exhaustive, making it suitable for any medical specialty and independent of specific health conditions while remaining clinically pertinent.

IPS follows a standardized format and structure, making it easy for health care providers to find and understand critical patient information quickly. The essential health information such as allergies, current medications, existing medical conditions, recent procedures, immunization history, and contact details is included. The whole document is designed to be multilingual, ensuring that health care providers from different linguistic backgrounds can access and comprehend the patient's health information. IPS is particularly valuable in emergency situations where patients may not be able to provide their complete medical history, allowing health care providers to make informed decisions about their care.

For data exchange and transformation, IPS should can be integrated into various health information systems and EHRs. The secure and privacy compliant should be included into system which designed for IPS. IPS adheres to privacy and security standards, ensuring that patient data remains confidential and protected during exchange.

The IPS is part of broader efforts to improve health information exchange at the international level, promoting better patient care and safety, especially for individuals who receive medical treatment in foreign countries or travel extensively. It contributes to the standardization of health data formats and facilitates more seamless health care delivery across borders.

PHR Data Source

Taiwan MHB

Taiwan MHB is a government-initiated online platform in Taiwan that serves as a personal health record system for citizens. It is designed to provide individuals with easy access a wide range of their health-related data which includes diagnosis, medication, treatment from outpatient, clinic, dental, traditional Chinese medicine, allergy information, inspection (check) result, imaging or pathological examination (examination) report information, summary of discharge medical records, organ donation or palliative care wishes, adult preventive, vaccination, and 4 cancer screening results. All the data is collected from various health care sources for health insurance claims which include hospitals, clinics, pharmacies, and laboratories. The platform aggregates health data from multiple health care providers and sources, providing users with a centralized and holistic view of their health history. This integration helps individuals and healthcare professionals make informed decisions about their health care.

Users can view their prescription history and medication information, helping them keep track of their prescribed drugs and potential drug interactions. Taiwan MHB offers health promotion and education materials to help users make informed decisions about their health and well-being. It provides information on various health topics and encourages preventive health care measures.

An individual's National Health Insurance card and a personal identification number can be used to secure access Taiwan MHB. This ensures that only authorized users can access their health records. It also provides mobile applications to make it convenient for users to check their health information on-the-go.

Philippine Health Insurance Corporation

Established in 1995, the Philippine Health Insurance Corporation (PhilHealth) plays a pivotal role in realizing universal health coverage within the Philippines. As a tax-exempt, government-owned and controlled corporation attached to the Department of Health, PhilHealth is the primary governmental agency entrusted with the administration of the nation's health insurance program. Its fundamental mission is to furnish health insurance coverage and benefits to the Filipino populace,

XSL•FO
RenderX

guaranteeing access to vital medical services and financial security against the often-burdensome expenses associated with healthcare.

The purview of PhilHealth encompasses a broad spectrum of individuals, including employees, self-employed persons, and sponsored members such as indigent families and senior citizens. Its comprehensive coverage extends to a diverse array of health care services, encompassing hospitalization, outpatient care, preventive health measures, and maternity benefits, among other critical provisions. The overarching goal is to enhance accessibility to health care, particularly for those in vulnerable circumstances, achieved through strategic partnerships with various health care providers, including both public and private institutions.

Recognizing the need for equitable health care access, PhilHealth extends its coverage to different categories of beneficiaries, ensuring inclusivity across society. The “Formal” sector encompasses individuals employed by public and private entities, while the “Indigents,” also known as “PhilHealth Ng Masa,” benefit from national government subsidies via the National Household Targeting System for Poverty Reduction. “Sponsored Members” receive support from their respective Local Governments, while “Lifetime” members, comprising retirees and pensioners with at least 120 months of premium payments, are accorded nonpaying status.

In addition, the “Senior Citizen” category, as stipulated under RA 10645, offers free PhilHealth coverage to all Filipino citizens aged 60 and above. Meanwhile, the “Informal Economy” classification accommodates informal sectors, self-earning individuals, organized groups, Filipinos with dual citizenship, natural-born citizens, and even overseas Filipino workers within its fold.

PhilHealth’s enduring commitment to fostering equitable access to health care services continues to drive its initiatives, striving to safeguard the health and well-being of the Filipino population at large.

PhilHealth CF4 is used by organizations to apply for health insurance reimbursement, and it contains 7 main sections. Since PhilHealth has not yet provided people with PHRs, CF4 data will be used in this study to evaluate whether it is suitable to develop a PHR framework for public use and compare it with IPS.

Ethical Considerations

This study did not involve any human subject research. The system development and scenario testing were conducted using simulated data solely generated for this research purpose. Therefore, the study did not require institutional review board approval. The data used in this study were fully synthetic and created by the research team. As such, there were no issues related to informed consent, data authorization, or secondary data usage. Since the dataset was entirely simulated and did not include any real individual or patient information, there were no privacy or confidentiality concerns associated with the research. No human participants were involved in this study; hence, there was no compensation provided. Finally, the study did not include any identifiable images or visual materials derived from real individuals. All figures and visual content are based on simulated or generic representations, and there is no risk of identification.

In ethics review or approval, the data sets examined (CF4 and MHB) involve health insurance claim information. No identifiable patient-level data were directly used in this implementation study. Because only de-identified, aggregated, and publicly or institutionally authorized data were analyzed, the study qualified for exemption per local institutional review board guidelines. In informed consent was not required from individual patients as data were anonymized and retrospective.

In context with privacy and confidentiality, no patient names, addresses, or unique identifiers are presented or shared. The study team adhered to data protection protocols and removed any potentially identifiable information before analysis. Compensation was not provided, as no direct participant enrollment was required. This manuscript does not include identifiable patient images.

Results

Overview

In this study, the FHIR international standard format is used to implement the content of IPS.

This study investigates the development of the medical information industry in the United States, Singapore, Japan, and Indonesia. Many manufacturers have developed electronic medical record systems that use FHIR as the data exchange standard, as described in Table 1.

XSL•FO
RenderX

Table . Major health information standards and prominent health IT vendors by country.

Country	Standard	Manufacturer
Unite States	HL7,[a] FHIR[b]	Epic Systems, Cerner, Allscripts Healthcare Solutions, Athenahealth, McKesson
Singapore	HL7, DICOM. IHE	Integrated Health Information Systems, Philips Healthcare, InterSystems, NCS Group
Japan	HL7, DICOM, JAMI	HITACHI, Fujitsu, NEC, NTT DATA, Mitsubishi Electric
Indonesia	HL7, DICOM. IHE	Jireh Group, BMSC, HSC Medical Center, MAKNA Cancer Research Institute

[a]HL7: Health Level 7 International.

[b]FHIR: Fast Healthcare Interoperability Resources.

[c]DICOM: Digital Imaging and Communications in Medicine

[d]IHE: Integrating the Healthcare Enterprise

Among them, Epic Systems and Cerner in the United States have outstanding achievements. In the Newsweek world best smart hospital 2023 survey, EPIC's system was used in 6 out of the 10 top-ranking hospitals, and it was used in 14 out of the amount top 20 hospitals.

Oracle Cerner leads the EHR global market according to the 2021 - 2022 Global Growth Report presented by Cerner. The KLAS 2022 global (non-US) electronic medical record market report states that Oracle Cerner's customers include 2389 acute inpatient hospitals, of which 1052 are non-US customers, including medical institutions in the Middle East, the United Kingdom, Canada, and Australia.

FHIR International Standard Adoption

FHIR is adopted to enhance interoperability among health care systems, enabling different health care providers, systems, and applications to seamlessly exchange and share patient data. This interoperability is crucial for delivering better patient care, reducing errors, and improving efficiency. It offers a standardized way to represent and exchange health care information. Through the standardization simplifies data integration and reduces the complexity of data exchange, making it easier for health care organizations to work together. Through the FHIR, patient-centric care can be easier reached. This leads to more informed decision-making and better patient outcomes. For the technology and system development part, the uses modern web standards which include JSON and RESTful application programming interfaces, making it compatible with today's technology stack. This compatibility encourages its adoption as it aligns with current IT trends. Furthermore, many regulations discuss the application of FHIR, which all require data exchange and operational protection to be completed, such as the Health Insurance Portability and Accountability Act (HIPAA) and the 21st Century Cures Act in the United States.

There are many international companies across the healthcare spectrum that have integrated FHIR into their systems to enhance interoperability and patient care.

Epic Systems, Cerner, and Allscripts have embraced FHIR to improve data exchange in EHRs systems, And health insurance companies like Aetna and UnitedHealthcare use FHIR to enable secure data exchange with health care providers. In Europe, Orion Health and InterSystems are implementing FHIR to promote cross-border data exchange and interoperability.

Data Content of IPS FHIR

A team has already defined the FHIR IPS implementation guild based on the IPS specifications. This implementation guide is used to compare and use data fields and select necessary items. Include necessary items such as medications, allergies, and disease status; it is recommended to include information such as vaccinations, treatment history, medical equipment, diagnostic reports, etc; additional options include vital signs, disease history, pregnancy status, smoking and alcohol habits, care plan, functional status, and advance directives (Figure 1).

XSL•FO
RenderX

Figure 1. The structure of the International Patient Summary (IPS).

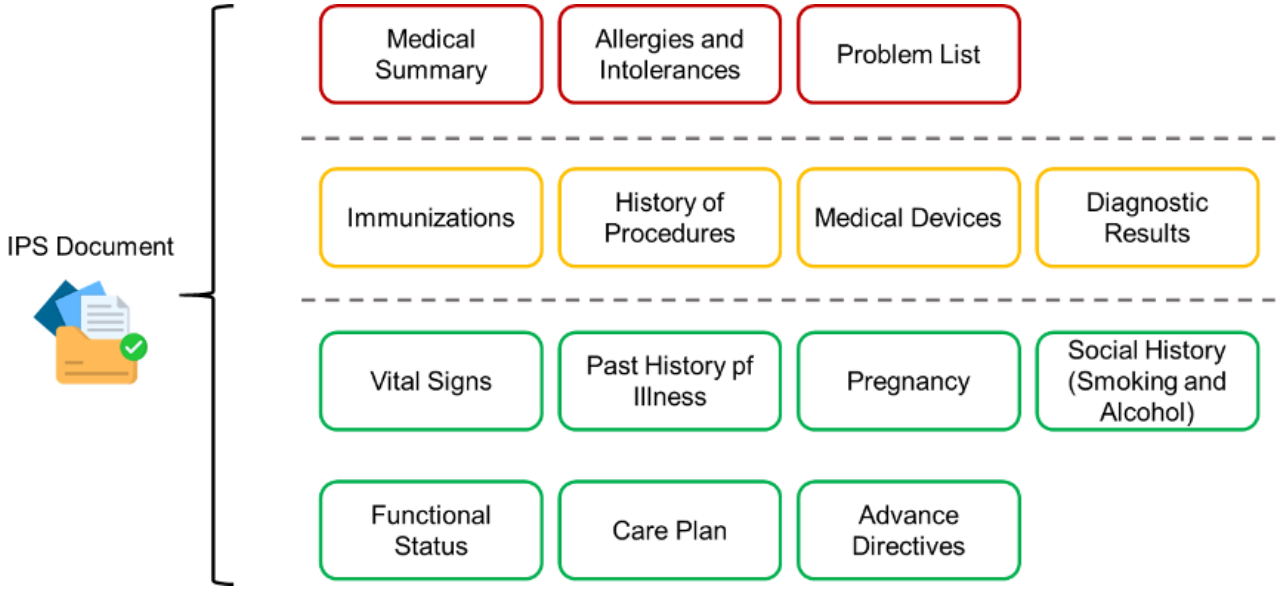

Data Mapping Outcomes

IPS defines 17 sections (5 required, 7 recommended, and 5 optional). Table 2 details how CF4 and MHB compare. In summary, MHB fully or partially covered 82.4% (14/17) IPS sections. Missing elements primarily related to device usage, social history (tobacco and alcohol), functional status, and advance directives.

CF4 covered 12/17 (70.6%), with additional limitations regarding unstructured text fields, past history of illnesses, radiology or pathology results, and social history.

Consequently, substantial modifications or additional data capture would be needed to meet all IPS-required or recommended sections.

XSL•FO
RenderX

Table . Comparison of IPS, CF4, and MHB data content and architecture (This table compares how each data element in IPS aligns—or does not align—with CF4 and MHB. It includes required, recommended, and optional sections. [O=presentor easily mapped, X=absentor limited.]).

Section	Type	IPS[a]	PhilHealth CF4[b]	Taiwan MHB[c]
Header	Required	Patient	O	O
Header	Required	Encounter	O	O
Medication summary				
	Required	Medication request	O	O
	Required	Medication statement	O	O
	Required	Medication	O	O
Allergies and intolerances	Required	Allergy intolerance	O	O
Problem list	Required	Condition	O	O
Immunizations	Recommended	Immunization	O	O
History of procedures				
	Recommended	Procedure	O	O
	Recommended	Organization	O	O
	Recommended	Device	X	X
Medical devices				
	Recommended	Device use statement	X	X
	Recommended	Device	X	X
Diagnostic results				
	Recommended	Observation (results)	O	O
	Recommended	Diagnostic report	O	O
	Recommended	Organization	O	O
Laboratory results				
	Optional	Observation results: laboratory	O	O
	Optional	Specimen	X	X
	Optional	Media observation (results: laboratory and media)	X	O
Radiology results				
	Optional	Observation (results: radiology)	X	O
	Optional	Device	X	X
	Optional	Imaging study	X	O
	Optional	Practitioner	X	O
Pathology results				
	Optional	Observation results: pathology	X	O
	Optional	Specimen	X	X
	Optional	Media observation	X	O
Vital signs	Optional	Observation	O	O
Past history of illnesses	Optional	Condition	X	O
Pregnancy				

XSL•FO
RenderX

Section	Type	IPS[a]	PhilHealth CF4[b]	Taiwan MHB[c]
	Optional	Observation (pregnancy: EDD)	O	O
	Optional	Observation (pregnancy: outcome)	O	O
	Optional	Observation (pregnancy: status)	O	O
Social history (SH)				
	Optional	Observation (SH: tobacco use)	X	X
	Optional	Observation (SH: alcohol use)	X	X
Functional status (autonomy or invalidity)				
	Optional	Condition	X	X
	Optional	Clinical impression	X	X
Plan of care	Optional	Care plan	X	X
Advance directives	Optional	Consent	X	X

[a]IPS: International Patient Summary.

[b]CF4: Claim Form 4.

[c]MHB: My Health Bank.

[d]EDD: Estimated date of delivery.

Security and Architecture Implementation

We adopted OAuth 2.0 as the primary authorization mechanism, with SMART on FHIR providing user-level permissions and token-based access to patient data. This approach ensures robust privacy controls consistent with HIPAA-aligned protocols. Both CF4 and MHB would require re-engineering to issue FHIR resources, including (1) FHIR resource mapping, that is patient resource aligns well with CF4's identification data and MHB's beneficiary info; (2) medication and allergies from MHB integrate readily with FHIR's medication statement or allergy intolerance resources, that is, CF4 has partial coverage requiring structured fields; and (3) cross-border data flow, that is, a pilot exchange between Taiwanese and Philippine health care providers could test read or write operations on a shared FHIR server.

PHR Management Structure

To accelerate the development of personal health record management globally and facilitate international data integration, we propose adopting a personal health record content based on the IPS and using FHIR as the international standard. Throughout the development process in various regions, we have also compiled recommendations from WHO (World Health Organization) guidelines and international organizations, emphasizing the following key areas mentioned in Textbox 2.

Textbox 2. Key areas.

- Standardization: Establishing standardized formats and protocols for health data to ensure interoperability across different systems and countries.
- Data security and privacy: Implementing robust security measures and privacy safeguards to protect sensitive health information.
- Interoperability: Ensuring seamless integration with various health information systems, both at the national and international levels.
- Inclusion of key health information: Incorporating essential patient information, medical history, medications, allergies, immunizations, and relevant diagnostic results.
- Patient access and control: Enabling patients to access and control their health records, allowing them to share information securely with health care providers.
- Cross-border health data exchange: Facilitating the exchange of health information across borders to support continuity of care for individuals who travel or receive medical services internationally.
- Integration with existing health systems: Ensuring compatibility and integration with existing EHR systems, health information exchanges, and other healthcare infrastructure.
- Technological standards: Adhering to international standards for health informatics to enhance consistency and compatibility.

XSL•FO
RenderX

National Participation Process

In the promotion of national strategies, there are various facets to consider. Drawing insights from cases in the United States, Europe, and international corporations, key implementation points for national initiatives can be summarized as follows:

In the planning and development process, it is imperative to engage various stakeholders such as health care providers, government agencies, insurers, and patient advocacy groups. These entities should be invited to join the promotion group, contributing their expertise to provide critical solutions for the market. Regulatory compliance is essential, involving alignment with national and international regulations and standards related to health data management and interoperability. Furthermore, a comprehensive approach includes public awareness and education through campaigns that inform the public about the benefits and implications of the comprehensive international PHR.

User feedback and input play a crucial role, with efforts focused on soliciting opinions from end-users, including patients and health care professionals, to ensure the system aligns with their needs and expectations. Practical testing is facilitated through the implementation of pilot programs, evaluating the functionality, security, and usability of the comprehensive international PHR in real-world settings. Finally, policy development is integral, requiring the formulation of policies and guidelines governing the use, access, and sharing of international health records to establish a robust legal framework. This holistic approach ensures the involvement of key stakeholders, regulatory compliance, public awareness, user input, practical testing, and the establishment of necessary policies for the successful development and implementation of the comprehensive international PHR.

Discussion

Principal Findings

This research focuses on advancing the development of an international PHR by analyzing health insurance claims records. The study investigates the differences between the health insurance claim form used by PhilHealth in the Philippines and the MHB in Taiwan. By comparing these 2 schemas to the IPS, the research is able to identify gaps and overlaps that are important when transforming CF4 and MHB into a comprehensive PHR.

Using a comparative analysis, the study identifies key data elements present in international patient summaries that are lacking or underrepresented in the current systems. The aim is to optimize the effectiveness of CF4 and MHB by incorporating essential messages that align them with the functionalities of a PHR. The findings reveal gaps in converting health insurance claims into PHRs, emphasizing the importance of addressing missing components for holistic care, patient safety, and continuity of care.

Comparison to Previous Work

In terms of data fields, Taiwan's MHB has undergone years of adjustments and now possesses a comprehensive structured framework that can be effectively converted into the FHIR IPS format. In contrast, the data structure of the Philippines' CF4 is relatively less developed, with many fields stored in unstructured text formats, leading to significant limitations in data usability.

Each country should conduct a thorough inventory of its medical data to ensure maximum consistency and redundancy. This approach would give greater significance to data exchange and usage.

For addressing international medical needs, it is essential to conduct an in-depth assessment of data usability and standardization. By converting data into international formats, greater interoperability can be achieved [19,20].

Strengths and Limitations

Strengths include a direct comparison of 2 national frameworks (CF4 and MHB) to internationally recognized standards. Limitations involve the largely conceptual nature of the cross-border PHR architecture and the reliance on partial unstructured data in CF4. Further pilot testing and real-world implementation remain future work.

Future Directions

This study has confirmed that the IPS is a sufficient reference standard for health data exchange in international applications. The next phase should involve engaging with more countries to discuss feasible data sources, standardized data preprocessing workflows, details of textual data extraction, and reliable methods for data exchange. Furthermore, more advanced discussions should focus on data integration and consistency standards.

Conclusion

By mapping CF4 and MHB to IPS sections, this implementation study highlights the feasibility of developing a shared FHIR-based next-generation PHR for the Philippines and Taiwan. Such an approach can improve health care coordination, especially for foreign workers and tourists, through seamless exchange of medical data. Adoption hinges on robust policy, stakeholder collaboration, and structured data collection to fill existing gaps in coverage.

Acknowledgments

The work presented in this paper was carried out in collaboration among all authors. H-AL and C-YH formed the conception and study design, and also did the architecture design of the system. H-AL and MSOA did the literature review, and system analysis.S-WH, W-HC and J-CH put a lot of effort on the design the system. H-AL drafted the manuscript and C-YH made significant revisions. C-YH and ABM supervised the methods of the implementation on cross-border architecture and supplied

XSL•FO
RenderX

valuable improvement suggestions. All authors provided approval of the final version. This project has received funding from the National Science and Technology Council, Taiwan.

Data Availability

This study primarily used deidentified health insurance claim information from Taiwan's MHB and the Philippine Health Insurance Corporation's CF4. Due to confidentiality obligations and national regulations, the full data set cannot be made publicly available. Aggregated or anonymized data may be shared upon reasonable request to the corresponding author, contingent on approval by the relevant authorities in Taiwan and the Philippines.

Conflicts of Interest

None declared.

Checklist 1
iCHECK-DH: Guidelines and Checklist for the Reporting on Digital Health Implementations.
[DOCX File, 31 KB - formative_v9i1e56272_app1.docx]

References

1. Yousef CC, Salgado TM, Farooq A, et al. Health care providers' acceptance of a personal health record: cross-sectional study. J Med Internet Res 2021 Oct 21;23(10):e31582. [doi: 10.2196/31582] [Medline: 34569943]
2. Tapuria A, Porat T, Kalra D, Dsouza G, Xiaohui S, Curcin V. Impact of patient access to their electronic health record: systematic review. Inform Health Soc Care 2021 Jun 2;46(2):192-204. [doi: 10.1080/17538157.2021.1879810] [Medline: 33840342]
3. Gazzarata R, Almeida J, Lindsköld L, et al. HL7 Fast healthcare interoperability resources (HL7 FHIR) in digital healthcare ecosystems for chronic disease management: scoping review. Int J Med Inform 2024 Sep;189:105507. [doi: 10.1016/j.ijmedinf.2024.105507] [Medline: 38870885]
4. Han HR, Gleason KT, Sun CA, et al. Using patient portals to improve patient outcomes: systematic review. JMIR Hum Factors 2019 Dec 19;6(4):e15038. [doi: 10.2196/15038] [Medline: 31855187]
5. Lee YL, Lee HA, Hsu CY, Kung HH, Chiu HW. Implement an international interoperable PHR by FHIR—a Taiwan innovative application. Sustainability 2020;13(1):198. [doi: 10.3390/su13010198]
6. Bateja R, Dubey SK, Bhatt A. Integration of EHR and PHR leveraging cloud services for approving treatments. ijhs :4940-4959 [FREE Full text] [doi: 10.53730/ijhs.v6nS8.13322] [Medline: 26219539]
7. Singh B, et al. Wearable sensors assimilated with internet of things (iot) for advancing medical imaging and digital healthcare: real-time scenario. In: Inclusivity and Accessibility in Digital Health: IGI Global; 2024:275-297. [doi: 10.4018/979-8-3693-1463-0.ch018]
8. Harahap NC, Handayani PW, Hidayanto AN. Functionalities and issues in the implementation of personal health records: systematic review. J Med Internet Res 2021 Jul 21;23(7):e26236. [doi: 10.2196/26236] [Medline: 34287210]
9. Charishma T, et al. PHR and personalized health recommendations using rule-based approach. In: International Conference on Intelligent Vision and Computing: Springer Nature Switzerland; 2022:433-455. [doi: 10.1007/978-3-031-31164-2_37]
10. Liu P, Yeh LL, Wang JY, Lee ST. Relationship between levels of digital health literacy based on the Taiwan digital health literacy assessment and accurate assessment of online health information: cross-sectional questionnaire study. J Med Internet Res 2020 Dec 21;22(12):e19767. [doi: 10.2196/19767] [Medline: 33106226]
11. Tabari P, Costagliola G, De Rosa M, Boeker M. State-of-the-art fast healthcare interoperability resources (FHIR)-based data model and structure implementations: systematic scoping review. JMIR Med Inform 2024 Sep 24;12(1):e58445. [doi: 10.2196/58445] [Medline: 39316433]
12. Mandl KD, Gottlieb D, Mandel JC. Integration of AI in healthcare requires an interoperable digital data ecosystem. Nat Med 2024 Mar;30(3):631-634. [doi: 10.1038/s41591-023-02783-w] [Medline: 38291298]
13. Torab-Miandoab A, Samad-Soltani T, Jodati A, Rezaei-Hachesu P. Interoperability of heterogeneous health information systems: a systematic literature review. BMC Med Inform Decis Mak 2023 Jan 24;23(1):18. [doi: 10.1186/s12911-023-02115-5] [Medline: 36694161]
14. Allam S, Nadikattu AKR. AI economical wearable smart device to alert real time health reports to doctors. IJCRT 2020;8(12):2320-2882 [FREE Full text]
15. Subbiah V. The next generation of evidence-based medicine. Nat Med 2023 Jan;29(1):49-58. [doi: 10.1038/s41591-022-02160-z] [Medline: 36646803]
16. Amiri Z, Heidari A, Darbandi M, et al. The personal health applications of machine learning techniques in the internet of behaviors. Sustainability 2023;15(16):12406. [doi: 10.3390/su151612406]
17. Integrating telemedicine and AI to improve healthcare access in rural settings. Int J Life Sci Res Arch 2024;7(1):059-077 [FREE Full text] [doi: 10.53771/ijlsra.2024.7.1.0061]

XSL•FO
RenderX

18. Organization, I.S. Health Informatics — International Patient Summary: ISO; 2021. URL: https://www.iso.org/standard/79491.html
19. Spanakis EG, Sfakianakis S, Bonomi S, Ciccotelli C, Magalini S, Sakkalis V. Emerging and established trends to support secure health information exchange. Front Digit Health 2021;3:636082. [doi: 10.3389/fdgth.2021.636082] [Medline: 34713107]
20. Kiourtis A, Mavrogiorgou A, Mavrogiorgos K, et al. Electronic health records at people's hands across Europe: the InteropEHRate protocols. Stud Health Technol Inform 2022 Nov 3;299:145-150. [doi: 10.3233/SHTI220973] [Medline: 36325855]

Abbreviations

CF4: Claim Form 4
EHR: electronic health record
FHIR: Fast Healthcare Interoperability Resources
HIPAA: Health Insurance Portability and Accountability Act
HL7: Health Level 7 International
IPS: International Patient Summary
MHB: My Health Bank
PHR: personal health record
WHO: World Health Organization

Edited by A Mavragani; submitted 11.01.24; peer-reviewed by S Shin; revised version received 27.04.25; accepted 02.05.25; published 02.07.25.

Please cite as:
Lee HA, Huang JC, Huang SW, Chen WH, Marcelo AB, Aljibe MSO, Hsu CY
Implementing a Cross-Border Next-Generation Personal Health Record in the Philippines and Taiwan: An Implementation Case Report Using Health Level 7 International Fast Healthcare Interoperability Resources
JMIR Form Res 2025;9:e56272
URL: https://formative.jmir.org/2025/1/e56272
doi:10.2196/56272

© Hsiu-An Lee, Jui-Chun Huang, Shih-Wun Huang, Wei-Han Chen, Alvin B Marcelo, Miguel Sandino O Aljibe, Chien-Yeh Hsu. Originally published in JMIR Formative Research (https://formative.jmir.org), 2.7.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Sprint-Intensity Arm Interval Training May Improve Cardiorespiratory Fitness and Cardiometabolic Health Among Children With Mobility Disabilities: Case Report

Natalie Dean[1*], DO; Tanvee Sinha[1*], BS; Larsen Bright[2]; Katie M Ellison[3*], LDN, RDN, PhD; Drew Sayer[3*], PhD; Raven Young[2*], BS; Drew Davis[2*], MD; James H Rimmer[4*], PhD; Byron Lai[2*], PhD

[1]Department of Physical Medicine and Rehabilitation, University of Alabama at Birmingham, 1600 7th Avenue South 5 Dearth, McWane, Birmingham, AL, United States

[2]Division of Pediatric Rehabilitation Medicine, Department of Pediatrics, University of Alabama at Birmingham, Birmingham, AL, United States

[3]Division of Family and Community Medicine, Department of Medicine, University of Alabama at Birmingham, Birmingham, AL, United States

[4]Dean's Office, School of Health Professions, University of Alabama at Birmingham, Birmingham, AL, United States

[*]these authors contributed equally

Corresponding Author:
Natalie Dean, DO
Department of Physical Medicine and Rehabilitation, University of Alabama at Birmingham, 1600 7th Avenue South 5 Dearth, McWane, Birmingham, AL, United States

Abstract

Background: There are limited options for aerobic exercise options that improve cardiorespiratory fitness and manage cardiometabolic health that are also age-appropriate and suitable for children with mobility disabilities. Children with disabilities require exercise programs that incorporate adapted movements to meet various functional needs, which offer brief training durations to accommodate busy schedules and use remote training methods at home to bypass logistical transportation barriers.

Objective: The aim of this study is to test the potential effects and safety of a sprint-intensity arm-exercise interval training program, combined with music and telehealth, on cardiorespiratory fitness and cardiometabolic health in a child with cerebral palsy.

Methods: This study was a 12-week exercise intervention from pretrial to posttrial for a single child with cerebral palsy (male, age 17 y). The intervention was conducted at the participant's homes. The participant exercised 3 times per week while following along with YouTube exercise videos. Videos included 4-second maximal sprint bouts followed by periods of rest, which were repeated 30 times during a single exercise session (total of ~2 minutes of maximal exercise). Exercise sessions were supervised by research staff using videoconferencing. Cardiorespiratory fitness was indicated by peak oxygen consumption (pVO_2), which was measured by a portable metabolic cart during a graded exercise test. Cardiometabolic health outcomes included body composition by dual-energy x-ray absorptiometry scan and a cardiometabolic blood profile by a dried blood spot test. Outcomes were descriptively analyzed.

Results: The participant achieved a 33.6% increase in pVO_2 (14.6 to 19.5 $mL/kg^{-1}/min^{-1}$), a 37.8% improvement in blood triglycerides (82 to 51 mg/dL), and a 15.4% improvement in the total cholesterol to high-density lipoprotein ratio (6.5 to 5.5). Additionally, he had a 5.9% reduction in body weight (171 to 161 lbs) and a 9.6% reduction in total body fat (61.35 to 55.48 lbs) from the arms, legs, and trunk. The participant experienced no adverse events or problems during the intervention. After completing the program, the participant elicited a maximal intensity of exercise using armbands, as demonstrated through pVO_2.

Conclusions: Sprint-intensity interval training that uses the arms may be safe and potentially effective for enhancing cardiorespiratory fitness and cardiometabolic health in children with physical disabilities. Further research is needed to verify the outcomes of this case report.

(JMIR Form Res 2025;9:e64440) doi:10.2196/64440

KEYWORDS

adapted physical activity; high-intensity interval training; telehealth; teleexercise; cerebral palsy; pediatric rehabilitation; cardiorespiratory fitness; cardiometabolic; case report; aerobic exercise; cerebral palsy; physical disabilities

XSL•FO
RenderX

Introduction

Cerebral palsy (CP) is a common motor and physical disability in childhood [1], which can be described as a group of "permanent, non-progressive disorders of the development of movement and posture" and is often accompanied by severe intellectual, conceptual, and social limitations [2]. CP develops from any event that can affect the fetal and neonatal brain and can be categorized by (1) predisposing pathological processes (such as placental vascular disorders), (2) peripartum events (such as chorioamnionitis or abruption), and/or (3) neonatal complications (sepsis, intraventricular hemorrhage) [1]. Although it is not a progressive disease, the clinical presentation of CP can vary as the child grows, requiring ongoing monitoring as well as updated assessments of health risks with subsequent treatments.

Physical impairment increases the difficulty of performing everyday activities, which puts an individual with CP at risk of a sedentary or physically inactive lifestyle [3]. Given their various limitations, individuals with CP have an increased risk of cardiovascular disease, metabolic syndrome, and cardiovascular disease mortality compared to age-matched individuals [4]. Since children with CP who exercise regularly are twice as likely to continue exercising as adults, there is an urgent need to find ways to promote exercise adoption in youth to reduce the health risks associated with this condition as they grow older [5]. Notably, one-third of children with CP are nonambulant and rely on wheelchair mobility [6], and the remaining two-thirds have varying degrees of impairments primarily affecting the lower extremities, which significantly impacts their ability to walk [7]. This makes conventional exercise modalities like running, cycling, and walking generally not feasible for a large portion of people with CP. Currently, there are few proven, evidence-based options for aerobic exercise other than wheelchair propulsion or arm ergometry for nonambulatory individuals or those who require accommodations for their lower limb limitations [8-10].

Sprint interval training (SIT) is a form of high-intensity interval training (HIIT), which is a time-efficient and evidence-based approach to increase cardiorespiratory fitness through its structure of sets of short "sprints" of activity, with periods of rest in between [11]. SIT involves repeated maximal efforts typically lasting less than 30 seconds, designed to reach near-maximal intensity in short bursts [12,13]. In contrast, HIIT consists of intervals performed at submaximal intensities around 80% - 95% of an individual's maximal heart rate, with longer activity durations [12,13]. Compared to HIIT, SIT may be more practical for this population since it produces significant metabolic and muscular adaptations with seemingly less overall physiological stress [13,14]. SIT uses primarily anaerobic type II fast-twitch muscle fibers and features shorter exercise intervals, which in theory allows for greater replenishment of anaerobic energy pathways compared to HIIT [15,16]. Considering that SIT using the arms has been found to elicit similar benefits to programs that use the legs due to musculature differences [17], a SIT arm program may be ideal for people with mobility disabilities. Additionally, access to the program can be enhanced by combining telehealth technology [18], which bypasses barriers related to transportation and a lack of nearby accessible facilities for people with disabilities [19].

This study modified an existing movement-to-music telehealth program [20] to incorporate SIT training using the arms for one adolescent with CP. The previous program included long durations of low-to-moderate intensity exercises using rhythmic movements to music, along with adult actors. This study aimed to test a SIT program with brief exercise sessions to make it more convenient for families to perform, combined with child-appropriate video themes and music. To the best of our knowledge, interventions related to this form of training have not been published previously.

The purposes of this study were to:

1. Examine the potential effects of the movement-to-music SIT program on the health of a child with CP. Health outcomes included cardiorespiratory fitness, body composition, and cardiometabolic health.
2. Evaluate the safety of a SIT program in a person with CP through adverse events and problems.
3. Evaluate the intensity of exercise that can be achieved from SIT by analyzing oxygen consumption during exercise.

Methods

Recruitment of Participant

The participant included in this study was a member of the research team. Eligibility criteria included wireless internet access at home, being between the ages of 10 and 17 years, a medical diagnosis of CP, the ability to exercise using the arms, self-reported mobility disability, and a caregiver to support the study procedures of the child. The participant recruited for this study was a 17-year-old male with spastic diplegic CP. He had mild-to-moderate hemiparesis, intellectual disability, and ambulated independently (Gross Motor Function Classification System level II). He received regular botulinum toxin injections for the management of spasticity in his lower limbs.

Ethical Considerations

The case study was not considered research by the university review board; thus, review and approval by the board was not required. As per university guidelines for case studies, written informed consent and assent were still obtained from the participant and their caregiver prior to participation. Study data were deidentified using alpha-numeric codes, and participation was confidential. No compensation was provided for participation. No artificial intelligence was used in any portion of the manuscript.

Procedures

The participant came to the laboratory for a baseline data collection visit, participated in the 12-week intervention, and then came back to the laboratory for a post–data collection visit. Health outcomes were measured predata and postdata. On the postvisit, he conducted an evaluative exercise assessment for our third study purpose. The participant completed the intervention in May 2022. This case study informed the development of a pilot randomized controlled trial [21].

XSL•FO
RenderX

Intervention

The intervention prescription included a 12-week SIT program. The SIT program included 3 sessions per week that each lasted 15 - 25 minutes. Each SIT session included a warm-up followed by 30 repeated sequences of 4 seconds of all-out arm exercise and 30 seconds of rest (totaling 2 minutes of all-out exercise over the entire session), as well as a cooldown. Rest periods included active or passive recovery that progressed from 30 seconds in week 1 to as low as 16 seconds in week 7 of the program, guided by the participants' needs. The exercise prescription was based on a previous SIT study [16]. Regarding specific movements, the warm-up primarily included low-intensity repetitive movements that were opposite to those used in the exercise sequence. For example, backward "arm jogging" (shoulder circles backward emphasizing the posterior deltoids) was a common warm-up movement, given that most exercise movements used the anterior deltoids. Exercise movements coincided with the theme of the exercise video. Movements included gross shoulder and elbow movements, such as climbing motions, rope swinging, lightsaber chopping, web shooting, and vertical and straight forward punching. Videos informed the viewer to mirror or modify the movements to what felt "good" to them (eg, use only arm). Videos included verbal cues to encourage high-intensity exercise through verbal cues: "Try your very best!" "Don't hold back," and "Use all of your energy." Videos were based on age-appropriate themes, specifically, 3 popular fictional superhero genres: Spider-Man, Star Wars, and Super Mario. The videos included 2 actors, 1 disability exercise expert (BL), and 1 child with CP (LB; Figure 1).

Figure 1. Demonstration of a sprint-intensity arm interval training video used in a 12-week, home-based, upper-body movement-to-music program designed to make high-intensity exercise more accessible and enjoyable for children with mobility disabilities. The program was performed by a 17-year-old male participant with spastic diplegic cerebral palsy.

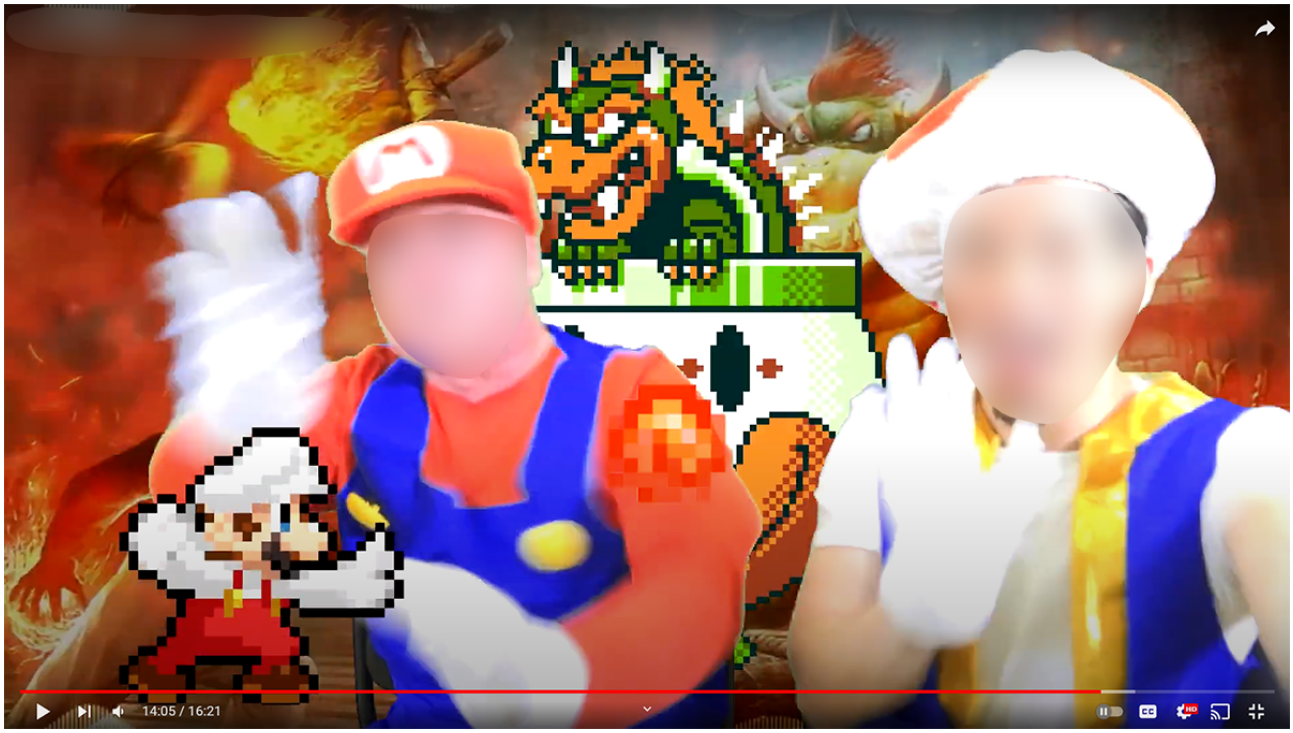

Each session was guided by a video that was played from a cloud-based server (YouTube). The intervention used a web-based platform, Zoom (Zoom Communications Inc), to allow videoconference supervision by a research staff member who was stationed at the research laboratory. The participant wore a Garmin Vivosmart 3 activity monitor (Garmin), which was used for a gross approximation of heart rate during exercise. Periodically during exercise, the research staff member would ask the participant what their heart rate reading was on their activity monitor, as well as their rating of perceived exertion using a 0 - 10 scale [22]. At the end of each session, the participant was asked to rate their perceived exertion to gauge fatigue from the session as a whole. If exertion was lower than a score of 7 for 3 consecutive sessions, resistance bands were to be introduced as a form of exercise progression. Resistance bands were added in the tenth week of the program. The participant was consistently asked to keep his diet and nutrition habits stable throughout the intervention period.

Exercise Intensity Evaluation Procedure

At the postvisit, the participant was provided with 20 minutes of rest after his graded exercise test, after which he underwent an exercise evaluation of the SIT program. For the evaluation, the participant was equipped with the portable metabolic cart and exercised along with one of the SIT videos taken from the later weeks of the program (ie, a video with less rest). The purpose of this was to assess whether arm SIT could elicit a high intensity of exercise. According to the American College of Sport Medicine guidelines for determining exercise intensity, low intensity is characterized as 57% - 63% of the age-predicted maximum heart rate, moderate as 64% - 76%, vigorous as 77% - 95%, and near maximal or maximal as ≥96%, respectively [23]. In addition to the American College of Sports

XSL•FO
RenderX

Medicine guidelines, we utilized the Tanaka equation (208 – 0.7 × age) for live target heart rate zones during exercise [24]. For the evaluation, the participant was asked to rest for 5 minutes, perform a warm-up of 3 minutes, and then do 3 stages of arm SIT: 3 minutes with no resistance band, followed by a 1-minute rest period, 3 minutes with a low-grade resistance band and a 1-minute rest period, and then 3 minutes with a midgrade band. Oxygen consumption was measured continuously throughout the evaluation.

Adverse events were monitored and recorded during the supervised sessions by the exercise coach. Additionally, during each call, the participant was asked how they were feeling since their last session, with a specific focus on whether their shoulders or arms were too sore to begin exercise and whether they experienced any injuries or problems since their last session.

Measures

Outcome measures were obtained before and after the 12-week program (at weeks 0 and 13). The primary outcome of interest was cardiorespiratory fitness, indicated by peak oxygen consumption (pVO_2), which was measured by a portable metabolic cart (COSMED K5, COSMED) during a graded exercise test using an arm ergometer. The test increases in resistance, and the individual performs the exercise until they can no longer continue. Test values were expressed as milliliters per kilogram per minute (relative measure).

Cardiometabolic health measures included a dual-energy x-ray absorptiometry scan to evaluate body composition (total mass, total lean mass, percent lean tissue, and percent fat tissue), a dried blood spot test to evaluate the lipoprotein profile (triglycerides, cholesterol, and low-density and high-density lipoprotein), glucose and glycated hemoglobin A_{1c} tests, and body weight measurement using a scale. To evaluate safety, the research staff consistently monitored the participant for adverse events (eg, muscle soreness/injury, overexertion, and illness).

Results

Feasibility Data

The participant completed all 36 prescribed exercise sessions (100% attendance) with no adverse events or problems. An exercise band was introduced in week 10 to maintain an increased level of exertion. Heart rate monitoring of exercise intensity did not work as planned and was not recorded during the sessions, despite being planned. The data did not appear accurate for two reasons: (1) the participant was taking medications that blunt heart rate and (2) there was a notable delay between stopping the exercise and the participant reporting the heart rate to the coach (primarily due to difficulty reading the small values on the wrist monitor because the participant had substantial visual impairment).

Health Outcomes

At baseline, the participant had a BMI of 31.4 kg/m^2 and dyslipidemia (high-density lipoprotein <40, elevated total cholesterol:high-density lipoprotein ratio; Table 1). The participant achieved a 33.6% increase in pVO_2 (14.6 to 19.5 mL/kg^{-1}/min^{-1}), a 37.8% improvement in blood triglycerides (82 to 51 mg/dL), and a 15.4% improvement in their total cholesterol to high-density lipoprotein ratio (6.5 to 5.5). Glucose and glycated hemoglobin A_{1c} remained stable (88% to 80% and 5.2% to 5.1%, respectively). Additionally, he had a 5.9% reduction in body weight (171 to 161 lbs), and a 9.6% reduction in total body fat (61.35 to 55.48 lbs) from the arms, legs, and trunk (14.9%, 9.6%, and 8.72% reduction, respectively).

Table . Pre and post outcomes.

Outcome	Baseline	Postintervention	Percent change
BMI (kg/m^2)	31.4	29.6	–5.7
Dyslipidemia (high-density lipoprotein <40; total cholesterol:high-density lipoprotein ratio)	6.5	5.5	–15.4
Peak oxygen consumption (mL/kg/min)	14.6	19.5	33.6
Blood triglycerides (mg/dL)	82	51	–37.8
Glucose (mg/dL)	88	80	–9.1
Glycated hemoglobin A_{1c} (%)	5.2	5.1	–1.9
Body weight (lbs)	171	161	–5.9
Total body fat (lbs)	61.35	55.48	–9.6
Arm fat (%)	34.9	30.3	–13.2
Leg fat (%)	39.4	38.8	–1.5
Trunk fat (%)	40.9	39.6	–3.18

XSL•FO
RenderX

Exercise Intensity Evaluation

Results for the exercise evaluation are shown in Figure 2. At rest, the participant had a pVO_2 of 6.53 (SD 1.6) mL/kg^{-1}/min^{-1}. During warm-up, he had a pVO_2 of 8.78 (SD 3.5) mL/kg^{-1}/min^{-1}. In stage 1 (no band), he exercised at a moderate intensity of 61.8% of his pVO_2 (12.04/19.5 mL/kg^{-1}/min^{-1}). In stage 2 (low-grade resistance band), he exercised at a near-maximal intensity of 90.8% pVO_2 (17.71/19.5 mL/kg^{-1}/min^{-1}). In stage 3 (midgrade resistance band), he exercised at a near-maximal intensity of 90.8% pVO_2 (17.61/19.5 mL/kg^{-1}/min^{-1}) [19]. Visually, the participant was performing faster movements during stage 2 (low-grade band) than he was in stage 3 (midgrade band). The rating of perceived exertion was 3 for the warm-up, 7 for exercise with no resistance band, and 10 at the end of each band stage.

Figure 2. Mean oxygen utilization across graded sprint-intensity arm intervals in a 17-year-old male with spastic diplegic cerebral palsy. Increases in resistance (no, low, or medium) led to near-maximal oxygen uptake, demonstrating the potential for high-intensity cardiovascular output. Error bars indicate SD.

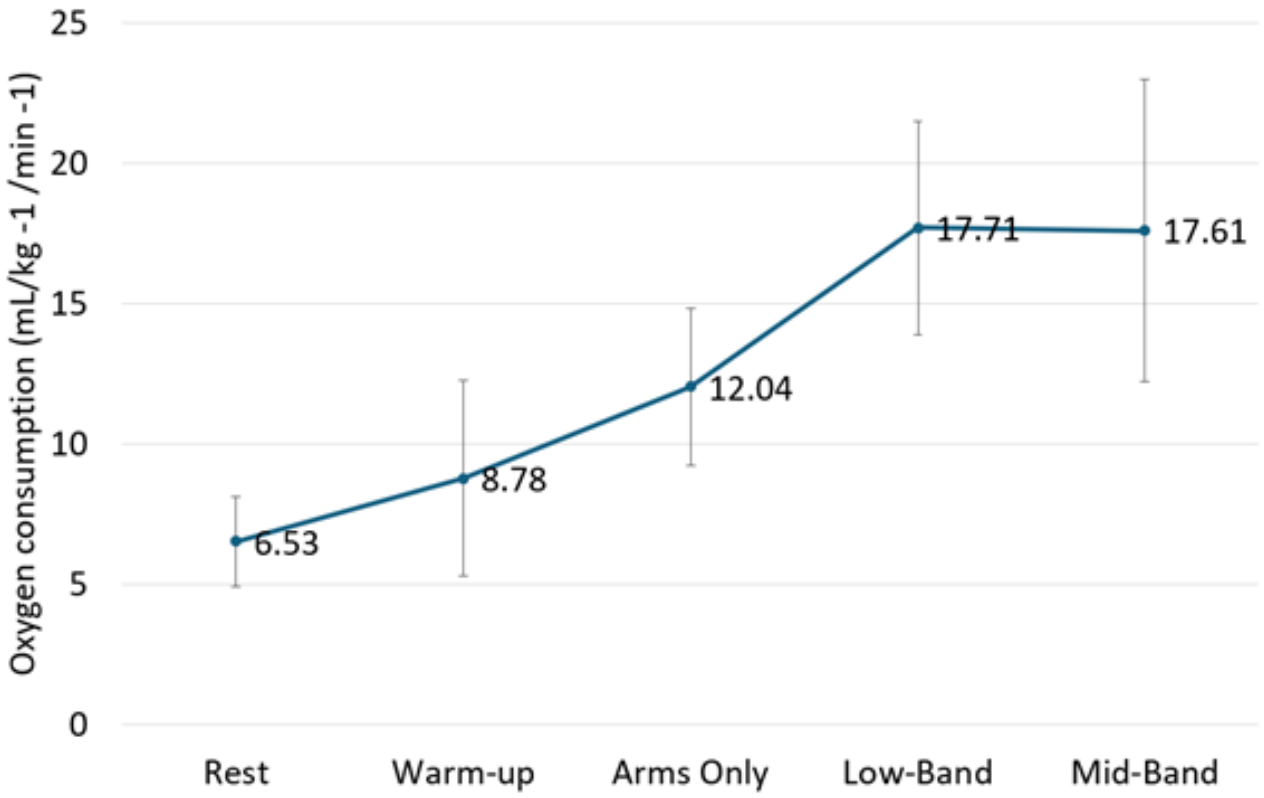

Discussion

This case study tested a unique arm-exercise protocol as a method for cardiovascular exercise that could be used by children with mobility disabilities. The 12-week protocol included repetitive maximal-intensity arm movements in short intervals (4 s on and 16 - 30 s rest), which resulted in a brief exercise duration for each session. The study aimed to explore the potential benefits of SIT exercise on indicators of cardiometabolic health. This case study demonstrated that SIT may be a potentially beneficial method for improving cardiovascular and cardiometabolic health among children with mobility disabilities. Additionally, the study suggests that even an experienced exerciser who completed 12 weeks of the intervention can achieve near-maximal intensities of exercise while following along with a SIT music video.

The participant experienced a 33.6% increase in pVO2. This increase was slightly lower than what has been observed from SIT using the arms in the general adult population [17]. However, this was slightly higher than the 18% - 22% that has been observed from short-term aerobic exercise studies conducted with children with CP [25]. Additionally, this study identified benefits to cardiometabolic health, namely reduced body weight and improved body composition, as well as some health-related blood biomarkers. Improvements to cardiometabolic health are critical, considering that few studies have demonstrated improvements to cardiometabolic health, particularly from seated arm exercises [9]. These findings are important considering that children with disabilities are at a two- to three-fold risk of mortality from cardiovascular disease and increased risk of metabolic syndrome compared to age-matched individuals secondary to physical inactivity and higher prevalence of dyslipidemia, hypertension, and fatigue [26,27]. In a recent population-based study, 17% of people with CP were classified as having metabolic syndrome and were at a 20% - 40% increased risk for cardiovascular disease (39.7% BMI-based and 26.55% lipid-based) [4]. Nevertheless, further research is necessary to begin to confirm these findings, considering that case studies typically have larger effects than randomized controlled trials.

XSL•FO
RenderX

A strength of the intervention is that it can be widely distributed through video cloud servers (like YouTube), allowing for easy replication and implementation in a large randomized controlled trial.

As this was a case study, there were several limitations. First, the findings should be interpreted with caution, considering that there was a single participant. Second, the participant was a member of the study team, which exposes the findings to a high level of bias. Third, we found that measuring exercise intensity using a wrist-worn monitor was not suitable because of heart rate blunting medications taken by the participant and difficulties with obtaining a real-time heart rate through verbal communication using the web-based platform (due to a long delay in the participant checking their watch and reporting the result to the coach). Fourth, the participant's outside physical activity participation and nutritional habits were not monitored, which could have confounded the results.

This study identified a convenient and potentially beneficial protocol for improving cardiorespiratory and cardiometabolic health among children with CP. As the intervention uses exercise videos with minimal supplies and only arm exercises, it is a potentially scalable intervention for a large population of children with mobility disabilities. However, further research is warranted to examine the effects of arm SIT with confirmatory sample sizes and study designs.

Acknowledgments

The laboratory equipment and space used in this project were supported by the Center for Engagement in Disability Health and Rehabilitation Sciences – Wellness, Health and Research Facility at the University of Alabama at Birmingham.

Data Availability

All data generated or analyzed during this study are included in this published article.

Conflicts of Interest

None declared.

References

1. Stavsky M, Mor O, Mastrolia SA, Greenbaum S, Than NG, Erez O. Cerebral palsy-trends in epidemiology and recent development in prenatal mechanisms of disease, treatment, and prevention. Front Pediatr 2017;5:21. [doi: 10.3389/fped.2017.00021] [Medline: 28243583]
2. McGuire DO, Tian LH, Yeargin-Allsopp M, Dowling NF, Christensen DL. Prevalence of cerebral palsy, intellectual disability, hearing loss, and blindness, National Health Interview Survey, 2009-2016. Disabil Health J 2019 Jul;12(3):443-451. [doi: 10.1016/j.dhjo.2019.01.005] [Medline: 30713095]
3. Ryan JM, Crowley VE, Hensey O, Broderick JM, McGahey A, Gormley J. Habitual physical activity and cardiometabolic risk factors in adults with cerebral palsy. Res Dev Disabil 2014 Sep;35(9):1995-2002. [doi: 10.1016/j.ridd.2014.03.051] [Medline: 24864052]
4. Heyn PC, Tagawa A, Pan Z, Thomas S, Carollo JJ. Prevalence of metabolic syndrome and cardiovascular disease risk factors in adults with cerebral palsy. Dev Med Child Neurol 2019 Apr;61(4):477-483. [doi: 10.1111/dmcn.14148] [Medline: 30663044]
5. Waltersson L, Rodby-Bousquet E. Physical activity in adolescents and young adults with cerebral palsy. Biomed Res Int 2017;2017:8080473. [doi: 10.1155/2017/8080473] [Medline: 29423412]
6. Rodby-Bousquet E, Hägglund G. Use of manual and powered wheelchair in children with cerebral palsy: a cross-sectional study. BMC Pediatr 2010 Aug 16;10(59):59. [doi: 10.1186/1471-2431-10-59] [Medline: 20712899]
7. van Ravesteyn NT, Scholtes VA, Becher JG, Roorda LD, Verschuren O, Dallmeijer AJ. Measuring mobility limitations in children with cerebral palsy: content and construct validity of a mobility questionnaire (MobQues). Dev Med Child Neurol 2010 Oct;52(10):e229-e235. [doi: 10.1111/j.1469-8749.2010.03729.x] [Medline: 20646033]
8. Lai B, Lee E, Kim Y, et al. Leisure-time physical activity interventions for children and adults with cerebral palsy: a scoping review. Dev Med Child Neurol 2021 Feb;63(2):162-171. [doi: 10.1111/dmcn.14751] [Medline: 33241561]
9. Martin Ginis KA, van der Ploeg HP, Foster C, et al. Participation of people living with disabilities in physical activity: a global perspective. Lancet 2021 Jul 31;398(10298):443-455. [doi: 10.1016/S0140-6736(21)01164-8] [Medline: 34302764]
10. Lai B, Young HJ, Bickel CS, Motl RW, Rimmer JH. Current trends in exercise intervention research, technology, and behavioral change strategies for people with disabilities: a scoping review. Am J Phys Med Rehabil 2017 Oct;96(10):748-761. [doi: 10.1097/PHM.0000000000000743] [Medline: 28398967]
11. Wood KM, Olive B, LaValle K, Thompson H, Greer K, Astorino TA. Dissimilar physiological and perceptual responses between sprint interval training and high-intensity interval training. J Strength Cond Res 2016 Jan;30(1):244-250. [doi: 10.1519/JSC.0000000000001042] [Medline: 26691413]
12. Zwinkels M, Verschuren O, de Groot JF, et al. Effects of high-intensity interval training on fitness and health in youth with physical disabilities. Pediatr Phys Ther 2019 Jan;31(1):84-93. [doi: 10.1097/PEP.0000000000000560] [Medline: 30507854]

XSL•FO
RenderX

13. Buchheit M, Laursen PB. High-intensity interval training, solutions to the programming puzzle: part I: cardiopulmonary emphasis. Sports Med 2013 May;43(5):313-338. [doi: 10.1007/s40279-013-0029-x] [Medline: 23539308]
14. Stöggl TL, Strepp T, Wiesinger HP, Haller N. A training goal-oriented categorization model of high-intensity interval training. Front Physiol 2024;15:1414307. [doi: 10.3389/fphys.2024.1414307] [Medline: 38957216]
15. Ito S. High-intensity interval training for health benefits and care of cardiac diseases - the key to an efficient exercise protocol. World J Cardiol 2019 Jul 26;11(7):171-188. [doi: 10.4330/wjc.v11.i7.171] [Medline: 31565193]
16. Satiroglu R, Lalande S, Hong S, Nagel MJ, Coyle EF. Four-second power cycling training increases maximal anaerobic power, peak oxygen consumption, and total blood volume. Med Sci Sports Exerc 2021 Dec 1;53(12):2536-2542. [doi: 10.1249/MSS.0000000000002748] [Medline: 34310498]
17. Zinner C, Morales-Alamo D, Ørtenblad N, et al. The physiological mechanisms of performance enhancement with sprint interval training differ between the upper and lower extremities in humans. Front Physiol 2016;7:426. [doi: 10.3389/fphys.2016.00426] [Medline: 27746738]
18. Rimmer JH, Lai BW, Young HJ. Bending the arc of exercise and recreation technology toward people with disabilities. Arch Phys Med Rehabil 2016 Sep;97(9 Suppl):S247-S251. [doi: 10.1016/j.apmr.2016.02.029] [Medline: 27267858]
19. Martin Ginis KA, Ma JK, Latimer-Cheung AE, Rimmer JH. A systematic review of review articles addressing factors related to physical activity participation among children and adults with physical disabilities. Health Psychol Rev 2016 Dec;10(4):478-494. [doi: 10.1080/17437199.2016.1198240] [Medline: 27265062]
20. Lai B, Vogtle L, Young R, et al. Telehealth movement-to-music to increase physical activity participation among adolescents with cerebral palsy: pilot randomized controlled trial. JMIR Form Res 2022 Oct 28;6(10):e36049. [doi: 10.2196/36049] [Medline: 36306154]
21. Lai B, Davis D, Young R, et al. The effects of virtual reality tele-exergaming on cardiometabolic indicators of health among youth with cerebral palsy: protocol for a pilot randomized controlled trial. JMIR Res Protoc 2022 Aug 17;11(8):e40708. [doi: 10.2196/40708] [Medline: 35976192]
22. Borg G. Borg's Perceived Exertion and Pain Scales: Human Kinetics 1998.
23. American College of Sports Medicine. ACSM's Guidelines for Exercise Testing and Prescription, 11th edition: Wolters Kluwer; 2021.
24. Cicone ZS, Holmes CJ, Fedewa MV, MacDonald HV, Esco MR. Age-based prediction of maximal heart rate in children and adolescents: a systematic review and meta-analysis. Res Q Exerc Sport 2019 Sep;90(3):417-428. [doi: 10.1080/02701367.2019.1615605] [Medline: 31157608]
25. Butler JM, Scianni A, Ada L. Effect of cardiorespiratory training on aerobic fitness and carryover to activity in children with cerebral palsy: a systematic review. Int J Rehabil Res 2010 Jun;33(2):97-103. [doi: 10.1097/MRR.0b013e328331c555] [Medline: 19770667]
26. Whitney DG. 5-year risk of "adult-onset" chronic diseases during childhood and adolescent transitioning for individuals with cerebral palsy. Prev Med Rep 2022 Oct;29:101933. [doi: 10.1016/j.pmedr.2022.101933] [Medline: 35928595]
27. Peterson MD, Lin P, Kamdar N, et al. Cardiometabolic morbidity in adults with cerebral palsy and spina bifida. Am J Med 2020 Dec;133(12):e695-e705. [doi: 10.1016/j.amjmed.2020.05.032] [Medline: 32687812]

Abbreviations

CP: cerebral palsy
HIIT: high-intensity interval training
pVO_2: peak oxygen consumption
SIT: sprint interval training

Edited by A Mavragani; submitted 17.07.24; peer-reviewed by A Woodward, KL Weston; revised version received 15.07.25; accepted 28.08.25; published 08.10.25.

Please cite as:
Dean N, Sinha T, Bright L, Ellison KM, Sayer D, Young R, Davis D, Rimmer JH, Lai B
Sprint-Intensity Arm Interval Training May Improve Cardiorespiratory Fitness and Cardiometabolic Health Among Children With Mobility Disabilities: Case Report
JMIR Form Res 2025;9:e64440
URL: https://formative.jmir.org/2025/1/e64440
doi:10.2196/64440

© Natalie Dean, Tanvee Sinha, Larsen Bright, Katie M Ellison, Drew Sayer, Raven Young, Drew Davis, James H Rimmer, Byron Lai. Originally published in JMIR Formative Research (https://formative.jmir.org), 8.10.2025. This is an open-access article

XSL•FO
RenderX

distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Applying the Nonadoption, Abandonment, Scale-Up, Spread, and Sustainability (NASSS) Framework to Adapt the CHAMP App for Pediatric Feeding Tube Weaning: Application and Case Report

Dana M Bakula[1,2,3], PhD; Alexandra Zax[1,4], MA; Sarah Edwards[1,2,3], DO; Kristina Nash[1], MSN, CPNP-AC; April Escobar[1], MS, RD; Rachel Graham[1], MSN, RN, CNP; Amy Ricketts[5,6], DNP, MSN, CPNP-AC; Ryan Thompson[5], MS; Sarah Bullard[3], MS, RD; Julianne Brogren[5], MSN, RN, CPPS; Leah Shimmens[1], MSN, RN; Lori A Erickson[5,6], PhD, CPNP-PC

[1]Department of Pediatrics, Children's Mercy Kansas City, 2401 Gillham Rd, Kansas City, MO, United States
[2]School of Medicine, University of Missouri-Kansas City, Kansas City, MO, United States
[3]Center for Children's Healthy Lifestyles and Nutrition, Kansas City, MO, United States
[4]University of Kansas, Lawrence, KS, United States
[5]Remote Health Solutions, Children's Mercy Hospital, Kansas City, MO, United States
[6]School of Nursing, University of Missouri-Kansas City, Kansas City, MO, United States

Corresponding Author:
Dana M Bakula, PhD
Department of Pediatrics, Children's Mercy Kansas City, 2401 Gillham Rd, Kansas City, MO, United States

Abstract

Background: Evidence-based tube feeding (TF) weaning involves reducing the volume of tube feeds to induce hunger, with interdisciplinary monitoring to allow for proactive medical, nutritional, and behavioral intervention as needed. This can be done outpatient; however, the current standard of care requires a high degree of medical monitoring and care coordination, which can be challenging to implement. The CHAMP App is a mobile app designed for remote patient monitoring of children born with congenital heart conditions who are at high risk for medical morbidity and mortality. The CHAMP App remote patient monitoring program would be ideally suited to improve medical monitoring and care coordination.

Objective: This study aims to determine the feasibility of adapting the CHAMP App for children ready to wean from TF.

Methods: We used the Non-adoption, Abandonment, Scale-up, Spread, and Sustainability (NASSS) framework as a formative tool and conducted a case study beta test.

Results: The level of complexity for the digital innovation's adaptation supported a high likelihood of success for the TF population. Most issues were simple, such as expanding the types of data that could be entered into the app, and some were more complicated, for instance, training all relevant staff to use and maintain the technology. The case study beta test was conducted with "Greyson", a 10-month old child weaning from TF (name changed for confidentiality). Once a week, the team reviewed the parent-entered data and communicated with Greyson's parents, recommending a 25% reduction in tube feeding each week. With the CHAMP App facilitating 2-way communication between the family and the team, Greyson successfully transitioned from receiving 30% of his feeds orally and 70% via tube feeding to 100% oral feedings over the course of 1 month in a home setting.

Conclusions: The CHAMP App is likely to be incredibly valuable in TF weaning. The NASSS framework helped identify key considerations for adapting the CHAMP App for TF weaning at a Midwestern children's hospital. Interviews with the health care team highlighted issues like data entry expansion and staff training. The framework confirmed TF weaning as a suitable application with no major barriers. The CHAMP App successfully supported a test patient, Greyson, in weaning from his feeding tube. It may improve access, communication efficiency, and satisfaction among families and health care teams while reducing costs and enhancing safety monitoring. The app could also make TF weaning more accessible to families with lower health literacy.

(JMIR Form Res 2025;9:e67398) doi:10.2196/67398

KEYWORDS

gastric feeding tubes; mobile applications; interdisciplinary communication; feeding and eating disorders of childhood; case report

XSL•FO
RenderX

Introduction

Enteral tube feeding (TF) is a lifesaving measure, but the prolonged use of TFs is taxing on the child's and family's well-being [1]. TF requires complicated medical devices, access to formulas or blended foods that must be prepared and stored in specific ways, and 24/7 management, resulting in a high burden on the family of TF children and limitations to the TF child's independence [1-3]. Caregivers report that TF comes with the "loss of a normal life," negative psychological impacts, practical challenges, and the need to become the "expert" [1]. Thus, the goal is to wean from TF once a child is able to achieve weight stabilization and improved safety in eating by mouth [2,4].

Evidence-based TF weaning involves gradually reducing the volume of tube feeds to induce hunger under the guidance of a multidisciplinary team so that behavioral, medical, and nutrition interventions can be implemented as needed to support tube weaning [5]. It is particularly critical to prevent dehydration and excessive weight loss during TF weaning, as these can be dangerous and lead to other health problems [6]. With the appropriate monitoring and interdisciplinary support, outpatient TF weaning is incredibly effective (67% - 69% successful) [5]. However, the current standard of care requires a high degree of care coordination and communication between families and the medical team. This can be challenging to implement using traditional models of outpatient care that rely on parent-initiated contact between outpatient clinic visits [5,6].

Mobile health (mHealth) apps and software platforms that allow for remote home monitoring have become common strategies to monitor weight and growth in other pediatric specialties, such as cardiology [7]. The CHAMP App is a proprietary software platform developed in 2014, with key stakeholder input from parents and health care providers, for asynchronous data transfer by parents at home to their specialty health care teams, starting with the pediatric cardiac population [8]. Through this model, pediatric cardiology researchers have reported improved communication, pediatric survival through staged surgeries, growth, and fewer intensive care days [9]. We hypothesize that expansion of the CHAMP App adaptation to pediatric TF weaning would be an optimal way to improve the TF weaning process.

Our study objective is to describe the use of the Non-adoption, Abandonment, Scale-up, Spread, and Sustainability (NASSS) framework as a formative tool in identifying the potential fit between the CHAMP App and TF weaning. The NASSS framework is an implementation science framework aimed at predicting and evaluating the success of mHealth interventions [10]. We will also describe a case example of CHAMP for TF weaning, which we have labeled "CHAMP for the Feeder."

Method

Study Design

We used the NASSS framework to determine the feasibility of adapting the CHAMP App for children ready to wean from TF. The NASSS framework is organized by seven domains: (1) condition; (2) technology; (3) value proposition; (4) adopter system (5) organization; (6) broader context; and (7) embedding [10,11]. Each domain has a set of questions to answer and review in a mixed-methods approach from primary or secondary data sources, expert knowledge, and interviews with key stakeholders [10]. Each domain is categorized by (1) simple: straightforward, predictable, and have few components, (2) complicated: multiple interacting components or issues, or (3) complex: dynamic, unpredictable, and prove difficult to disaggregate into constituent components [11].

Setting

The organization is a Midwest, free-standing tertiary children's hospital with multiple inpatient and outpatient locations. The hospital has 390 inpatient beds over two locations, and multiple ambulatory care clinics across the region. Nearly 9000 employees care for children before birth to 18+, from all 50 states, 49.6% medicaid payor source, and over 14 countries internationally [12].

Interdisciplinary Feeding and Swallowing Program

The organization's Interdisciplinary Feeding and Swallowing Program (IFSP) is a specialized outpatient clinical team within the Gastroenterology Department [13]. The IFSP specializes in providing integrated care, from multiple specialists in one location for families of children with medical complexities and feeding difficulties. The program has a team of more than 30 specialists including gastroenterology physicians, nurse practitioners, nurse coordinators, speech-language pathologists, occupational therapists, dietitians, psychologists, and social workers [13]. This program conducts more than 2000 patient visits per year, and each year sees over 300 new patients.

Remote Health Solutions Team

The Remote Health Solutions (RHS) team is a department within the Division of Strategy, Innovation, and Partnerships. Their focus is on digital health models that can apply a personalized, proactive approach to remote asynchronous monitoring. This team includes nurse researchers, research nurses, clinical research coordinators, and program coordinators.

Description of the NASSS Framework Process

The NASSS framework is applied through planning and feasibility stages to guide conversations, and identify implementation barriers and facilitators. Semi-structured meetings were completed to evaluate feedback reflexively to increase answers to the questions from the 7 domains(Textbox 1). We met with stakeholders with varying expertise inside the organization. Interviews were conducted as virtually in small groups. Stakeholders included leaders with knowledge of remote health and mHealth platforms, care of children with tube feedings, pediatric providers, administration, strategy, innovation, ambulatory nursing, information systems, software development, patient and family partners, and research. Administrative data were used to evaluate feasibility and planning via patient volumes and incidence.

XSL•FO
RenderX

Textbox 1. Non-adoption, Abandonment, Scale-up, Spread, and Sustainability (NASSS) framework for planning and feasibility management by the research leadership team.

The questions were used to guide semi-structured interviews, and the summarized themes were then coded to reflect simple, complicated, or complex issues.

Condition or Illness

• What is the nature of the condition or illness?

• What are the relevant sociocultural factors and co-morbidities?

Technology

• What are the key features of the technology?

• What kind of knowledge does the technology bring into play?

• What knowledge or support is required to use the technology?

• What is the technology supply model?

Value Proposition

• What is the developer's business case for the technology (supply-side value)?

• What is its desirability, efficacy, safety, and cost-effectiveness (demand-side value)?

Adopter System

• What changes in staff roles, practices, and identities are implied?

• What is expected of the patient (and or immediate caregiver)- and is this achievable by and acceptable to them?

• What is assumed about the extended network of lay caregivers?

Organization

• What is the organization's capacity to innovate?

• How ready is the organization for this technology-supported change?

• How easy will the adoption and funding decision be?

• What changes will be needed in team interactions and routines?

• What work is involved in implementation and who will do it?

Wider Context

What is the political, economic, regulatory, professional (eg medicolegal), and sociocultural context for program rollout?

Embedding and Adaptation Over Time

How much scope is there for adapting and coevolving the technology and the service over time?

Analyses

Feedback from leaders, field notes from interviews, and administrative data were compiled. Interview themes were collated and key points were identified based on core components of the NASSS framework. Project leaders from the IFSP and RHS teams from the Gastroenterology and Strategy, Innovation, and Partnerships departments collaboratively reviewed participant responses via typed transcripts and meeting notes and conducted an inductive thematic analysis, guided by Morgan and Krueger methodology, to identify themes within the responses [14]. The project leaders met as a group to reach a consensus on theme content areas and to develop the final wording of themes. Questions and notes for each domain were answered by the project leaders and confirmed with division and executive leadership at the organization for agreement on the planned program's complexity. Generative artificial intelligence (AI) models were not used for any portion of the research project or writing of the study findings.

Ethical Considerations

Interviews with key stakeholders as part of the NASSS formative research process were deemed not to be human participants research and therefore did not involve consent procedures. Stakeholders were recruited to participate in semi-structured meetings through email and phone communication starting in late 2020. For the beta case testing of the CHAMP App for tube weaning, we received institutional review board (IRB) approval (Study #00001436) before study activities commencing in 2021. The parent was provided with all relevant study information and educated on the study procedures by a trained study staff member and given the option to opt in or out of the study. Upon agreeing to participate, the parent signed the IRB-approved informed consent to allow their child's data to be used in a clinical repository. In addition, the parent agreed to a secondary analysis case study report centered on their experience using the CHAMP App for TF weaning, as well as consented to share deidentified study data. All data presented here has been deidentified. There was no compensation provided to human participants research participants in this study.

XSL•FO
RenderX

Results

Domain 1: Condition

In line with the first domain of the NASSS framework, we began by considering the nature of the condition (ie, weaning from TF) to determine whether the CHAMP App might be appropriate for children ready for TF weaning. We determined that TF weaning was an appropriate population for the CHAMP App, and that the level of complexity needed to use the app to this population was simple based on the following key points: (1) The population who receive TF is well characterized and understood (26.99 - 44.25 per 1000 children <5 y of age require TF [15,16]). (2) Children with other medical conditions, or who are born prematurely, are more likely to need TF to support their health and development, and therefore, TF is often well understood by medical teams across most major pediatric tertiary care settings. (3) The TF weaning process requires close medical monitoring, which maps well to the capabilities of the CHAMP App. Specifically, TF weaning involves slowly reducing the amount of nutrition delivered via TF so that the child can eat more by mouth. Medical monitoring ensures that the child maintains age-appropriate growth and does not experience adverse consequences such as dehydration or malnutrition. (4) TF weaning in outpatient and home settings is ideal, with interdisciplinary oversight from physicians or nurse practitioners, nurses, psychologists, dieticians, and feeding therapists (speech and occupational therapists) [2]. This approach aligns well with the original use of the CHAMP App for use in the home setting.

Domain 2: Technology

We determined that TF weaning was an appropriate fit for the CHAMP App technology, and that the level of complexity needed to adapt the technology to this population was simple based on the following key points: (1) First, is the successful implementation and dissemination of the CHAMP App in cardiology. The CHAMP App was developed for data transfer with a pediatric cardiac population and provides families and the clinical team with a new model of care leveraged by a mobile app integrated with a software platform [17]. Data options include feeding intake, output, vital signs, weights, videos, and concerns (Figure 1). Since 2014, the CHAMP App in the pediatric cardiac population has been used in over 1000 patients, 28 states, and across 12 pediatric hospitals and is available in 11 languages for families to use on their mobile phones in research formats (NCT 15030113CMH). Nurses review the data intermittently and communicate with families using proactive data review in place of standard parent-driven communication in a reactive format in ambulatory clinics. Data is transferred to a health care team web portal which allows for early identification and feeding and growth intervention [9]. The CHAMP App is free-standing and dependable as an app with inclusion of more than 1100 patients across 12 pediatric hospitals in a pediatric research registry. The ability for proactive data review has led to reports of improved survival outcomes, no instances of malnutrition, and fewer episodes of morbidity with less intensive care days [9]. (2) The CHAMP App has easy access in mobile version and web-based health care team dashboard, simple sets of instructions, verified training methods, and manuals developed in multiple languages for smooth implementation. This would allow easy usability of the pp for new purposes. (3) The cloud-based technology supply model for CHAMP App is customizable and easy to adapt for new purposes. The app is designed to be generic with “plug and play” features on both the parent-facing and health care team dashboards and reports. To tailor the CHAMP App for the TF weaning population, minimal adjustments to the mHealth technology were needed, such as the ability to report additional feeding measures (Figure 2), red flags for parents of children being TF weaned (Figure 3).

XSL•FO
RenderX

Figure 1. CHAMP App user interface screenshot.

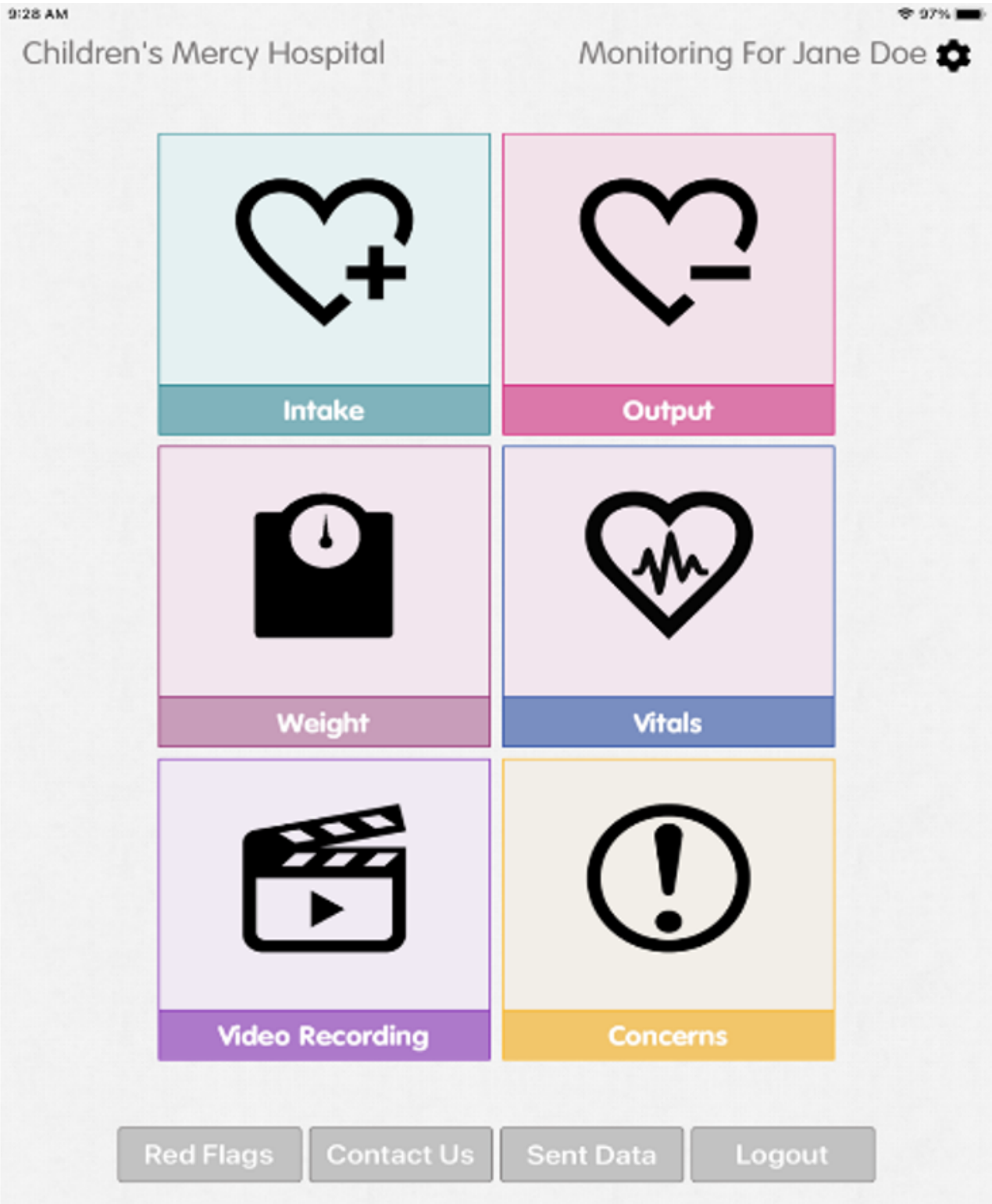

XSL•FO
RenderX

Figure 2. CHAMP App red flags function screenshot.

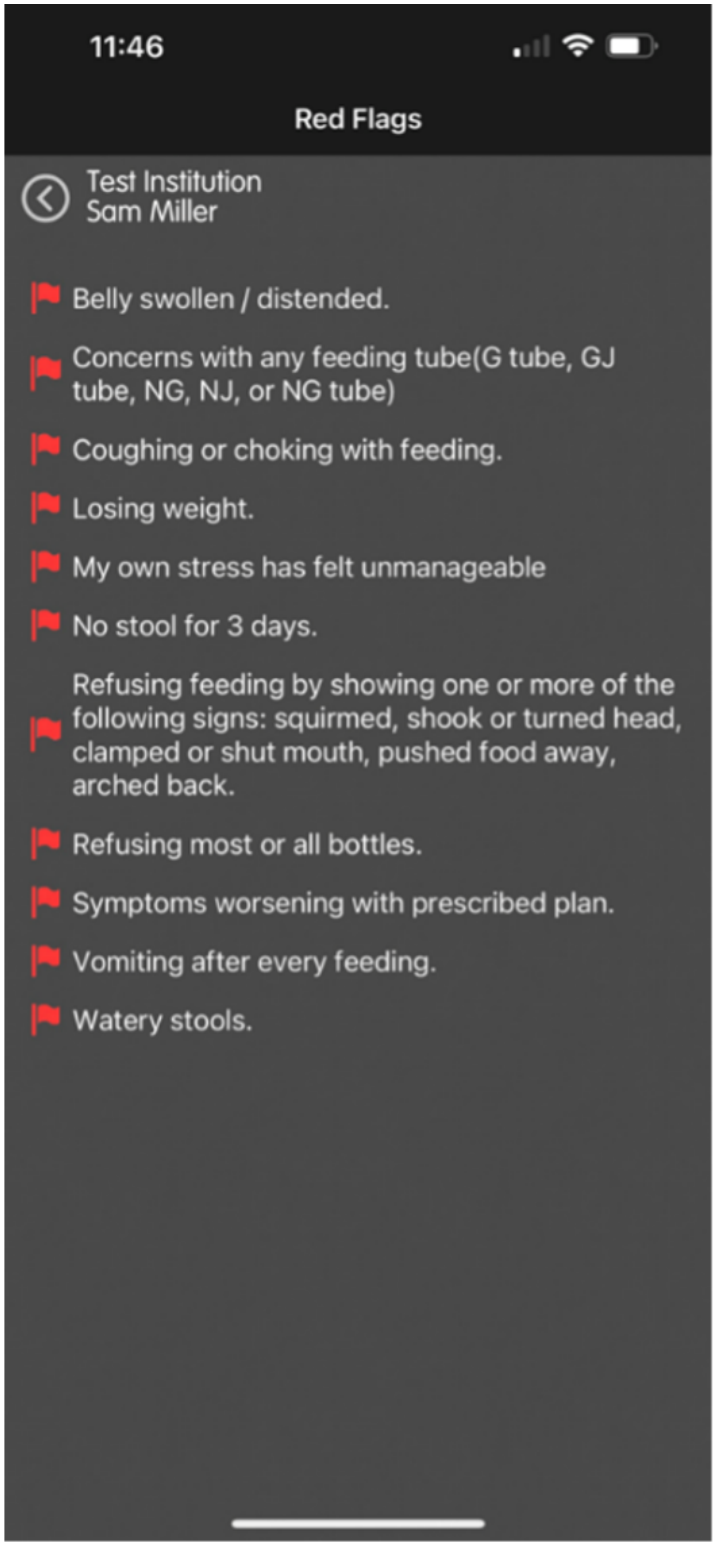

XSL•FO
RenderX

Figure 3. CHAMP App provider dashboard screenshot.

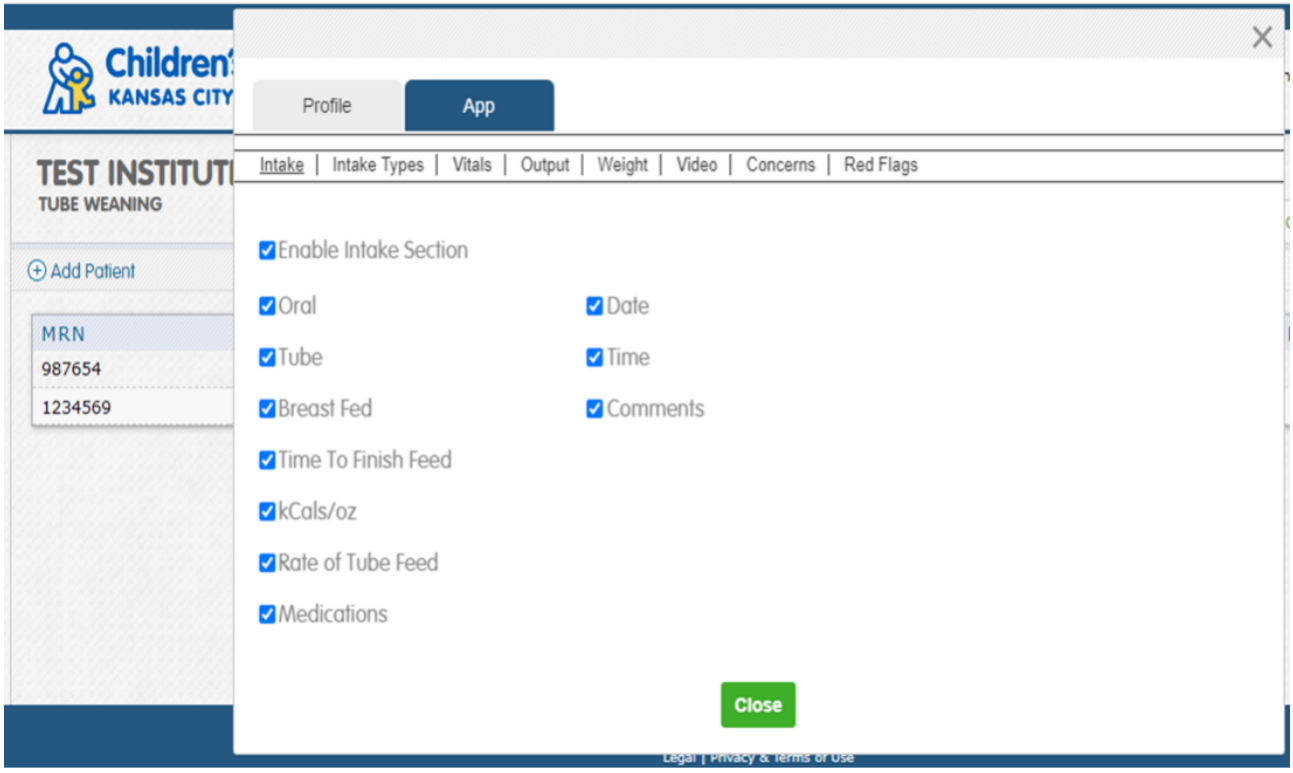

Domain 3: Value Proposition

During the value proposition phase, we evaluated the developers' business case for the supply-side and demand-side values of the technology and determined that the level of complexity needed to supply use of the CHAMP App for children undergoing TF weaning across multiple clinics was simple based on the following key points: (1) since the CHAMP App is a proprietary software platform developed by the organization, the hospital derives benefits from the use of the app and dissemination of the app. The return on this investment with improved patient outcomes reduced health care costs related to maintaining TF technology, family experience, and innovation in remote home monitoring internally and externally far outweighed the cost of expansion with staff time and financial instances. (2) The demand-side value is high because the app is desirable, efficacious, safe, and cost-effective. CHAMP App had a good adherence rate with nearly 75% for families of children with cardiac disease [17]. Over 90% of caregivers offered, consent to use the CHAMP App over paper documentation by parents [8]. (3) This project will fill an existing demand at our institution. The IFSP team at our institution has a high demand for appointments for new and follow-up patients and more children are being discharged each year from the organization with feeding needs. From administrative data, approximately 200 new children under the age of 5 years have gastrostomy tubes across all outpatient clinics per year. The incidence and value of an innovative support model for children with TF were displayed.

Domain 4: Adopter System

In line with the adopter system phase of evaluation, we considered the current standard of care for TF weaning and determined that the level of complexity to adapt the app within the adopter system was complicated based on the following key points, (1) when evaluating critical changes in staff roles, practices, and identities, it was determined that a front-line nutrition team member was necessary for successful adoption of the CHAMP App. This team member would need to be readily available to review data, communicate with families during the 4 - 6-week protocol wean, and be responsible for early interventions when weaning questions and problems arose. Further, the existing nutrition and feeding team staff would need to learn how to use CHAMP health care web portal. This change was informed by the standard of care which involves frequent and systematic communication between the health care team and family to identify readiness with oral intake, receive support from speech and feeding therapists, and acquire clinical approval from the specialty health care teams. During the tube wean itself, families rely on communication with the health care team for guidance on reduction of calories from the feeding tube over a short period of time, with the intent to increase hunger and subsequently oral intake. In outpatient settings for TF weaning, parents communicate with the health care team at clinic visits and through reactive health care communications initiated by parents to the health care team. In some cases, practices report inpatient admissions for weaning so that patients are guaranteed access to holistic team support; however, the disruption for families and high cost of readmission is limiting [18]. Therefore, a front-line nutrition team member would guarantee that families had support and access to the health care team while also minimizing disruption and cost for families.

It was noted that parent engagement would be key to successful implementation. The organization's cardiac parent family advisory board had specifically asked for an innovation like this for their children. Given the success of adoption and adherence among parents of the pediatric cardiac group [9,19],

XSL•FO
RenderX

it was expected that login to the CHAMP App and the patient portal would also be acceptable for parents of children prepared for TF weaning. (3) Finally, we considered the adoptees' network of lay caregivers and determined that this was not particularly relevant to the success of the app with this population given that data entry would be asynchronous and would initially follow a research protocol and not an extended care network. Further, larger scale implementation may also increase the use of extended caregivers, such as primary care providers, and a report system is built into the web portal for instances when parents require additional support.

To address these complicated issues, we determined that some work would be needed to build a shared vision, engage staff, enact new practices, and monitor the impact of the CHAMP App. TF weaning specific handouts, research protocols, and grant submissions for a project manager nurse were submitted and completed to address existing gaps. Further, weekly meetings continued with the RHS and IFSP teams and included short bursts of projects and long-term planning for implementation.

Domain 5: The Organization

In line with Domain 5 of the NASSS framework, we considered our organization's capacity to innovate and determined that the level of complexity for adaptation within said organization was simple-complicated based on the following key points: (1) The organization is known for a culture of pediatric innovation with a strong support model for technology innovations. Therefore, the risk involved in customizations of the CHAMP App for TF weaning was deemed appropriate at an organizational level. Notably, the organization has a high tension for change, good innovation-system fit, widespread internal support, and supported the research team's desire to submit for external funding with cost-sharing support for the endeavor. (2) Within this domain, we also considered factors such as adoption and funding. Accordingly, our organization had sufficient resources, anticipated cost savings, and required no new infrastructure or recurrent costs to adopt the technology for TF weaning. It was noted that new team routines would need to be aligned with established protocols if teams were not already using the weaning protocol.

Domain 6: The Broader Context

We considered the broader context that would be relevant to the implementation of the CHAMP App for children undergoing TF weaning and determined that the level of complexity for successful adaptation within the broader context was *simple* based on the following key points: (1) Professional and lay stakeholders were generally supportive and remote patient monitoring for weight gain is in place nationally from the Centers for Medicaid and Medicare Services guidelines from 2023 [20]. (2) While sense-making, collective reflection, and adaptive actions across multiple service lines are more complex compared to other areas within this domain, we were confident that collective reflection with IFSP and RHS teams would allow us to appropriately handle critical events and adapt to unforeseen eventualities.

Domain 7: Embedding and Initial Implementation

In line with the embedding and initial implementation domain of the NASSS framework, we evaluated the likelihood that app use would spread locally and, eventually, regionally. We determined that the level of complexity was *simple-complicated* based on the following key points: (1) ongoing use cases that have emerged for the technology, as well as continued organizational support for its adaptation and implementation. This suggests that app is generalizable outside of pediatric cardiology alone and may provide potential opportunities locally and even beyond our institution. (2) Further, the application may provide benefits across multiple pediatric specialties, as children with a range of medical comorbidities use enteral feeding for short or long-term needs. Indeed, there are over 15 service lines at our institution that wean children from TF. (3) Finally, the CHAMP App as a software platform continues to have collective reflection and support to continue the innovations, maintenance, and growth in areas that can follow the organization's mission and values specific to innovation for all children.

In sum, the level of complexity for the digital innovation's success across domains supported a high likelihood of success. The areas of focus would include primarily the adopter system as it was the domain that had the highest need for support for the implementation and change models.

Beta Testing: Case Study

Finally, we beta-tested this model of CHAMP App for usability with "Greyson," a 10-month-old child with congenital heart disease and reliance on TF who showed signs of TF weaning readiness including safe oral eating and weight stabilization (name changed for confidentiality). The family, who were familiar with CHAMP from their previous monitoring using the CHAMP App for Greyson's cardiac monitoring elected to use the CHAMP App for TF weaning. Parents were given a home scale, and entered weights, feeding information, and vital signs multiple times per week, which the IFSP team reviewed in real time. Once weekly after data review, the IFSP team communicated with Greyson's parents to recommend a TF feed reduction, over a 4-week span (25% volume reduction per wk). With the CHAMP App facilitating 2 way communication between the family and the IFSP team, Greyson transitioned from 30% of feeds orally (green dots in Figure 4) and 70% by TF to 100% oral feedings (purple dots in Figure 4) over one month in the home setting. In addition, his growth was maintained throughout the tube wean with no age-adjusted growth failures when the health care team completed proactive data review and nutritional interventions. The family also reported strong satisfaction with the technology and overall process. Upon 2 year follow up, Greyson still remains on 100% feedings by mouth, demonstrating the promising potential for the expansion of CHAMP in this population.

Following this beta test with Greyson our team identified opportunities to better customize the CHAMP App to optimize it's use for TF weaning. These changes include allowing for more detailed daily diet record entries, designing TF weaning specific alerts that the family can send to the IFSP team, and modifying the back-end infrastructure to optimally highlight

XSL•FO
RenderX

data that is important for TF weaning. These modifications will require collaboration with software teams to implement.

Figure 4. CHAMP tube feeding weaning case study data for "Greyson." Greyson transitioned from 30% of feeds orally (green dots) and 70% by TF to 100% oral feedings (purple dots).

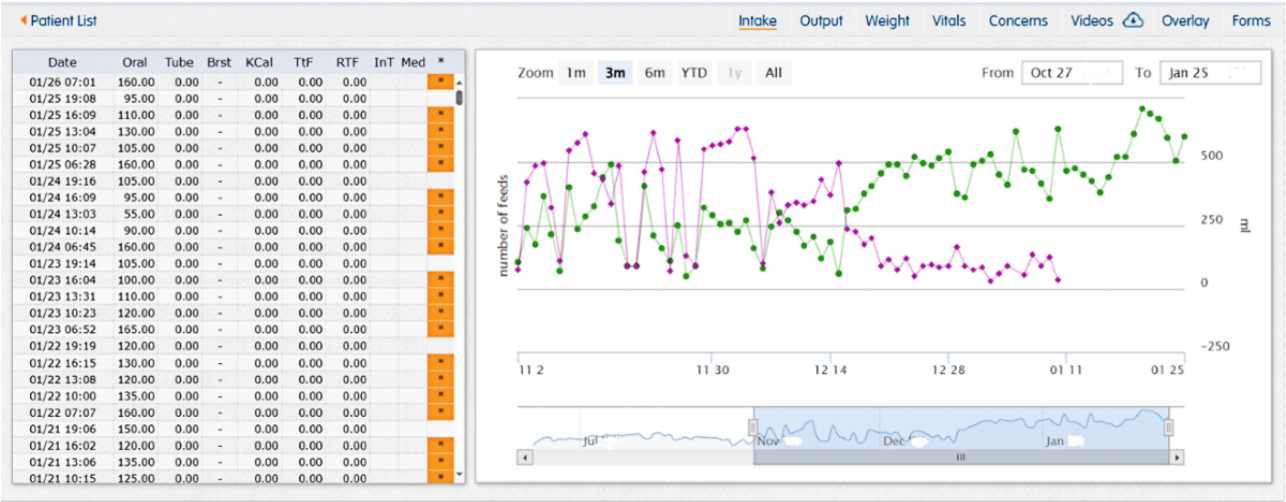

Discussion

Principal Findings

The NASSS framework successfully identified critically important considerations for applying the CHAMP App to a new population and conducted a successful pilot of the app for this new purpose with a test participant [10,11]. The NASSS framework was chosen to support our project with a strong background in sustainability and use in more than 1900 research and clinical evaluations of mHealth, including pediatric, remote patient monitoring, and nutrition focused projects [11,21-23]. Using interviews with key multi-disciplinary stakeholders of the health care team involved in TF weaning at a Midwestern children's hospital, we identified several critical considerations for adaptation of the CHAMP App across domains of condition, technology, value proposition, adopter system, organization, wider context, and embedment/adaptation over time [10,11]. Analysis and thematic understanding of the project were supported by clinical and technology experts to strengthen the data gathered and support the cooperative leadership of the clinical and remote monitoring teams. Executive leadership at the organization, had a strategic position in supporting the project, and provided strong insights into the organizational and broader context of implementation and adaptation internally and externally of the RPM model.

The NASSS-informed interviews allowed our team to recognize that TF weaning was an ideal opportunity for CHAMP App expansion, identified no insurmountable adaptation barriers for TF weaning, and allowed us to develop a roadmap toward successful adaptation and pilot testing. In our initial implementation with a test patient, Greyson, the CHAMP App successfully supported Greyson in weaning from his feeding tube and was acceptable to both the parents and providers involved in weaning. Most barriers were simple, such as expanding the types of data that could be entered into the app, and some were more complicated, for instance, training all relevant staff to use and maintain the technology. Similar to other area pediatric projects, uptake of innovations in an already burdened clinical team takes straightforward and easy to complete training with superior usability of the health care portion of mHealth technologies to encourage adoption [21].

The CHAMP App may have the opportunity to improve access to TF weaning in several ways. The CHAMP App can improve communication efficiency, which may allow for more satisfaction among families and the health care team. As this project moves beyond the research team alone into other areas of the organization, the barriers of teams not yet tube weaning can be addressed through clear steps in the care model to combat drift of sustainability outside the research team [24]. The CHAMP App allows us the opportunity to serve more families than could be served without RPM. Further, the current standard of care for outpatient TF weaning relies heavily on family health literacy and proactive communication with the health care team [2]. The CHAMP App's RPM system has the potential to be more accessible to families with lower health literacy and allows the health care team to be more proactive in tracking and reaching out to families, which may improve equitable access to TF weaning for all families [25]. In addition, the improved efficiency of TF weaning through RPM may improve outpatient workflow by reducing the number of in-clinic follow-ups needed for children who have successfully weaned from TF through the CHAMP App and, therefore, improve access to the outpatient clinic for new patients.

The CHAMP App may also improve the quality of TF weaning and pediatric nutrition care. Specifically, technology supported models of nutrition care have been identified as a way to allow for more sophisticated safety monitoring as a way to provide high quality nutrition care with appropriate health care team members [23]. As demonstrated by our test case, parent satisfaction is also likely to be high, as this app allows for more seamless support with TF weaning and may help parents to feel more supported and confident in approaching TF weaning. While this was a minimal cost for the test case since they were already used to a technology supported model of care, ongoing costs of implementation in a digital world will need to be followed [23,25]. Further, this app may reduce health care costs,

XSL•FO
RenderX

as maintaining TF is expensive [18], and alternatives to outpatient TF weaning are often incredibly expensive as they include inpatient or intensive outpatient treatment which is costly [5].

Limitations

This study highlights the use of the NASSS framework in identifying a novel opportunity for expansion of a RPM program and is an important first step towards expansion of the CHAMP App. However, several limitations should be taken into consideration when interpreting the findings. While many years of human-centered innovation design and feedback from stakeholders were completed to support this project, we did not conduct a formal qualitative analysis so these findings may be biased by interviewer interpretation. Further, this is a single site study, and involves only a single pilot case. Our team intends to engage in a larger scale implementation of the CHAMP App for TF weaning to better understand the use of this app for TF weaning and to test our indicators of barriers and enablers in scaling beyond the initial championing innovative team with feedback from key stakeholders [22,24].

Conclusion

The NASSS framework was valuable for identifying important considerations for expanding the CHAMP App to a new population [11]. Health care leaders, innovators, and clinicians may similarly find this framework valuable when planning new mobile technology projects and identifying potential areas of complexity. Based on the outcome of our project, we identified that the CHAMP App is likely to be incredibly valuable and useful in TF weaning, and our team intends to use the information learned in this study to inform the adaptation of the CHAMP App for TF weaning. We will be piloting the CHAMP App for TF weaning in a large cohort of children at a large Midwestern hospital (ClinicalTrials.gov Identifier: NCT06052891).

Acknowledgments

Funding for this project was funded in part by The Gerber Foundation. This work was supported by a CTSA grant from NCATS awarded to Frontiers Clinical and Translational Science Institute # KL2TR002367. The contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH or NCATS.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

SE and LAE are the primary Children's Mercy Hospital innovators developing the CHAMP App GI module discussed in this manuscript.

References

1. Serjeant S, Tighe B. A meta-synthesis exploring caregiver experiences of home enteral tube feeding. J Hum Nutr Diet 2022 Feb;35(1):23-32. [doi: 10.1111/jhn.12913] [Medline: 33934418]
2. Edwards S, Davis AM, Bruce A, et al. Caring for tube-fed children: a review of management, tube weaning, and emotional considerations. JPEN J Parenter Enteral Nutr 2016 Jul;40(5):616-622. [doi: 10.1177/0148607115577449] [Medline: 25791833]
3. Spratling R, Chambers R, Lawrence P, Faulkner MS. Best practices with use of feeding tubes for children at home. PNJ 2021;47(1):7. [doi: 10.62116/PNJ.2021.47.1.7]
4. Slater N, Spader M, Fridgen J, et al. Weaning from a feeding tube in children with congenital heart disease: a review of the literature. Prog Pediatr Cardiol 2021 Sep;62:101406. [doi: 10.1016/j.ppedcard.2021.101406]
5. Killian HJ, Bakula DM, Wallisch A, et al. Pediatric tube weaning: a meta-analysis of factors contributing to success. J Clin Psychol Med Settings 2023 Dec;30(4):753-769. [doi: 10.1007/s10880-023-09948-2] [Medline: 36856955]
6. Wright CM, McNair S, Milligan B, et al. Weight loss during ambulatory tube weaning: don't put the feeds back up. Arch Dis Child 2022 Aug;107(8):767-771. [doi: 10.1136/archdischild-2021-323592] [Medline: 35351738]
7. Marcolino MS, Oliveira JAQ, D'Agostino M, et al. The impact of mHealth interventions: systematic review of systematic reviews. JMIR Mhealth Uhealth 2018 Jan 17;6(1):e23. [doi: 10.2196/mhealth.8873] [Medline: 29343463]
8. Shirali G, Erickson L, Apperson J, et al. Harnessing teams and technology to improve outcomes in infants with single ventricle. Circ Cardiovasc Qual Outcomes 2016 May;9(3):303-311. [doi: 10.1161/CIRCOUTCOMES.115.002452] [Medline: 27166202]
9. Bingler M, Erickson LA, Reid KJ, et al. Interstage outcomes in infants with single ventricle heart disease comparing home monitoring technology to three-ring binder documentation: a randomized crossover study. World J Pediatr Congenit Heart Surg 2018 May;9(3):305-314. [doi: 10.1177/2150135118762401] [Medline: 29692236]
10. Greenhalgh T, Abimbola S. The NASSS framework - a synthesis of multiple theories of technology implementation. Stud Health Technol Inform 2019 Jul 30;263:193-204. [doi: 10.3233/SHTI190123] [Medline: 31411163]

XSL•FO
RenderX

11. Greenhalgh T, Wherton J, Papoutsi C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-Up, spread, and sustainability of health and care technologies. J Med Internet Res 2017 Nov 1;19(11):e367. [doi: 10.2196/jmir.8775] [Medline: 29092808]
12. Children's mercy facts. Children's Mercy. 2024. URL: https://www.childrensmercy.org/about-us/childrens-mercy-facts [accessed 2025-05-17]
13. Interdisciplinary feeding and swallowing program. Children's Mercy. 2024. URL: https://www.childrensmercy.org/departments-and-clinics/gastroenterology/feeding-clinic/ [accessed 2025-05-17]
14. Morgan DL, Krueger RA, King JA. The Focus Group Guidebook: Sage; 1998. URL: https://books.google.com/books?hl=en&lr=&id=YzjzGJd-OtoC&oi=fnd&pg=PR9&dq=Morgan+DL,+Krueger+RA,+King+JA.+The+Focus+Group+Guidebook.+Sage%3B+1998.&ots=C_OQk4BIvb&sig=1FrQYvc5bXPc-VBGvT5aA-oaSWM#v=onepage&q&f=false [accessed 2025-06-13]
15. Kovacic K, Rein LE, Szabo A, et al. Pediatric feeding disorder: a nationwide prevalence study. J Pediatr 2021 Jan;228(126-131):126-131. [doi: 10.1016/j.jpeds.2020.07.047] [Medline: 32702429]
16. Edwards ST, Glynn EF, Slogic M, et al. Demographics of children with feeding difficulties from a large electronic health record database. JPEN J Parenter Enteral Nutr 2022 Jul;46(5):1022-1030. [doi: 10.1002/jpen.2379] [Medline: 35383982]
17. Erickson LA. A Descriptive Correlational Study of Rate and Determinants of Parental mHealth Adherence to Symptom Home Monitoring for Infants with Congenital Heart Disease during the Single Ventricle Interstage Period: The DOMAIN Study: University of Missouri-Kansas City; 2020. URL: https://www.proquest.com/openview/5c89f4067f74ce50c64283d65777d2f7/1?pq-origsite=gscholar&cbl=18750&diss=y
18. Williams KE, Riegel K, Gibbons B, et al. Intensive behavioral treatment for severe feeding problems: a cost-effective alternative to tube feeding? J Dev Phys Disabil 2007 Jun;19(3):227-235. [doi: 10.1007/s10882-007-9051-y]
19. Erickson LA, Emerson A, Russell CL. Parental mobile health adherence to symptom home monitoring for infants with congenital heart disease during the single ventricle interstage period: a concept analysis. J Spec Pediatr Nurs 2020 Oct;25(4):e12303. [doi: 10.1111/jspn.12303] [Medline: 32662243]
20. Physician fee schedule CMS-1784. Centers for Medicare & Medicaid Services. 2024. URL: https://www.cms.gov/medicare/medicare-fee-service-payment/physicianfeesched/pfs-federal-regulation-notices/cms-1784-f [accessed 2025-05-17]
21. Castor C, Lindkvist RM, Hallström IK, et al. Health care professionals' experiences and views of eHealth in pediatric care: qualitative interview study applying a theoretical framework for implementation. JMIR Pediatr Parent 2023 Oct 18;6:e47663. [doi: 10.2196/47663] [Medline: 37851500]
22. Rudin RS, Perez S, Rodriguez JA, et al. User-centered design of a scalable, electronic health record-integrated remote symptom monitoring intervention for patients with asthma and providers in primary care. J Am Med Inform Assoc 2021 Oct 12;28(11):2433-2444. [doi: 10.1093/jamia/ocab157] [Medline: 34406413]
23. Barnett A, Kelly JT, Wright C, et al. Technology-supported models of nutrition care: perspectives of health service providers. Digit Health 2022;8:20552076221104670. [doi: 10.1177/20552076221104670] [Medline: 35677784]
24. Mazzocato P, Luckhaus JL, Malmqvist Castillo M, et al. A patient-driven mobile health innovation in cystic fibrosis care: comparative cross-case study. J Med Internet Res 2024 Jul 31;26:e50527. [doi: 10.2196/50527] [Medline: 39083342]
25. Portz J, Moore S, Bull S. Evolutionary trends in the adoption, adaptation, and abandonment of mobile health technologies: viewpoint based on 25 years of research. J Med Internet Res 2024 Sep 27;26:e62790. [doi: 10.2196/62790] [Medline: 39331463]

Abbreviations

IFSP: Interdisciplinary Feeding and Swallowing Program
IRB: institutional review board
mHealth: mobile health
NASSS: Non-adoption, Abandonment, Scale-up, Spread, and Sustainability
RHS: Remote Health Solutions
TF: tube feeding

XSL•FO
RenderX

Edited by A Mavragani; submitted 01.11.24; peer-reviewed by F Smith; revised version received 27.02.25; accepted 05.03.25; published 16.06.25.

Please cite as:
Bakula DM, Zax A, Edwards S, Nash K, Escobar A, Graham R, Ricketts A, Thompson R, Bullard S, Brogren J, Shimmens L, Erickson LA
Applying the Nonadoption, Abandonment, Scale-Up, Spread, and Sustainability (NASSS) Framework to Adapt the CHAMP App for Pediatric Feeding Tube Weaning: Application and Case Report
JMIR Form Res 2025;9:e67398
URL: https://formative.jmir.org/2025/1/e67398
doi:10.2196/67398

© Dana M Bakula, Alexandra Zax, Sarah Edwards, Kristina Nash, April Escobar, Rachel Graham, Amy Ricketts, Ryan Thompson, Sarah Bullard, Julianne Brogren, Leah Shimmens, Lori A Erickson. Originally published in JMIR Formative Research (https://formative.jmir.org), 16.6.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Development of an Eco-Friendly Smartphone-Assisted Nystagmus Recording System for Recording Vertigo Attacks Anytime, Anywhere: Pilot App Development Study

Yumi Dobashi[1], DipMLT; Masao Noda[1,2], MD, MBA, PhD; Tatsuaki Kuroda[1,3], MD; Noriaki Miyata[1], BLA; Makoto Ito[2], MD, PhD; Reiko Tsunoda[1], MD, PhD; Hiroaki Fushiki[1], MD, PhD

[1]Department of Otolaryngology, Mejiro University Ear Institute Clinic, 320 Ukiya, Iwatsuki-ku, Saitama-shi, Saitama, Japan

[2]Department of Otolaryngology and Head and Neck Surgery, Jichi Medical University, Shimotsuke, Japan

[3]Kuroda ENT Clinic, Yatsushiro-shi, Kumamoto, Japan

Corresponding Author:
Masao Noda, MD, MBA, PhD
Department of Otolaryngology, Mejiro University Ear Institute Clinic, 320 Ukiya, Iwatsuki-ku, Saitama-shi, Saitama, Japan

Abstract

Background: The widespread adoption of smartphones and tablet devices, along with advancements in data communication technology, has resulted in a paradigm shift in the treatment of dizziness. External factors, such as the spread of COVID-19, have accelerated this transformation in recent years. We have been pursuing telemedicine and web-based medical care to treat dizziness and have developed different products and services necessary for each treatment process stage. Several patients face difficulties in accessing medical facilities during severe vertigo episodes. Furthermore, clinical findings, such as nystagmus or other symptoms, may be absent when symptoms subside by the time of their appointment.

Objective: This study aimed to develop a smartphone app for capturing eye movements and head positions during vertigo attacks, enabling recordings anywhere, even at home or work.

Methods: We developed an app named “iCapNYS” that uses the iPhone’s front camera and gyro sensor to record eye movements and head positions. The app incorporates features designed to encourage spontaneous eye movements, minimizing nystagmus suppression caused by fixation. Additionally, we designed lightweight, recyclable cardboard goggles to securely hold the smartphone and block visual stimuli from the surrounding environment, optimizing the recording conditions.

Results: The “iCapNYS” system successfully captured subtle peripheral vestibular nystagmus in a patient with vertigo. The recorded nystagmus characteristics are comparable to those obtained using traditional infrared CCD (charge-coupled device) cameras.

Conclusions: This app is an effective tool for treating vertigo and is easy for older adults to use, as it can be recorded with only 3 taps. We expect that the introduction of this nystagmus-monitoring system will improve vertigo treatment quality, promote medical collaboration, and provide patients with peace of mind in their care.

(JMIR Form Res 2025;9:e73811) doi:10.2196/73811

KEYWORDS

dizziness; telemedicine; vertigo; “iCapNYS” system; mobile health; smartphone

Introduction

Vertigo and dizziness are symptoms of various disorders, including peripheral vestibular, central, cardiovascular, and psychogenic disorders. The 1-year prevalence and incidence of vertigo in adults are 4.9%, and 1.4%, respectively [1]. These conditions, which lead to episodic vertigo attacks, reduce the patients’ quality of life by causing nausea and vomiting, accompanied by headaches or an increased risk of falls [2]. Among diseases, benign paroxysmal positional vertigo (BPPV) is one of the most frequent causes of episodic vertigo; its 1-year prevalence and incidence were 1.6% and 0.6%, respectively [3]. The 1-year prevalence of vestibular migraine is estimated to be 1% to 2.7% [4] and that of definite Meniere disease is estimated to be 0.034% to 0.19% [5].

Despite the established diagnostic criteria, the transient nature of vertigo often precludes a definitive diagnosis during clinical assessments owing to the absence of observable oculomotor abnormalities. Eye movement observations during vertigo attacks are extremely useful for differentiating between these diseases [6,7].

Recently, the widespread use of smartphones and tablets has fueled rapid advancements in information and data

XSL•FO
RenderX

communication technologies. These developments have enabled the growth of digital medical care and telemedicine [8]. Furthermore, external factors, most notably the COVID-19 pandemic, have accelerated this transformation in recent years [9,10]. Previous studies have attempted to record eye movements during attacks using the recording function of cell phones for diagnosing Meniere disease [11]. Shah et al [12] reported 7 cases of BPPV diagnosed with high specificity and sensitivity using eye movement videos captured with a smartphone. Nonetheless, challenges remain in capturing recordings under low-light conditions using smartphones, which limits the universal applicability of these methods [13]. In addition to smartphone-based recordings, various other technologies have been explored in the field of vestibular medicine, including wearable goggles and home-use devices with guided instructions [14]. We recently devised a simple eye movement–recording system that combines a commercially available mini-infrared camera with 3D-printed goggles to record eye movements at home using smartphones [15]. This system allows patients to record their eye movements during vertigo attacks in the dark and provides recorded information to their doctors from home. These approaches reflect the increasing diversity of remote vestibular assessment technologies. However, its widespread adoption and ease of use were limited because the device had to be distributed to patients.

This study devised an eye movement–recording system that uses an iPhone to record eye movements during a vertigo attack. We report the development of an iPhone app that enables the observation and recording of eye movements and head position, along with eco-friendly specialized goggles made of cardboard, specially designed to hold the iPhone. This system is expected to improve vertigo treatment quality by enabling nystagmus monitoring, irrespective of the location and time.

Methods

The ICapNYS App That Captures Eye Movements and Head Position

We developed the iCapNYS app to record eye movements and head positions using the iPhone's front camera and built-in gyro sensor. The app allows for adjustment of focus, exposure, and zoom. The app is available for download from the Apple App Store under the title "iCapNYS." Since launching the app, the recorded eye movements can be checked with only 3 taps: first, after launching the app, the red circular button at the center of the screen should be tapped to start recording immediately (Figures 1A and 2A). Next, the center of the screen is to be tapped again to end the recording (Figures 1B and 2A). Third, the last recorded video can be played back by tapping the thumbnail in the upper-right corner (Figures 1C and 2A).

Figure 1. Instructions on how to use the "iCapNYS" application. (a) Tap the red circle button in the center of the screen to start recording. (b) Tap the center of the screen again to end the recording. (c) Tap a thumbnail to check the recorded video. (d) Display the video list. (e) Select and playback. (f) Swipe left to delete. (g) Send the video.

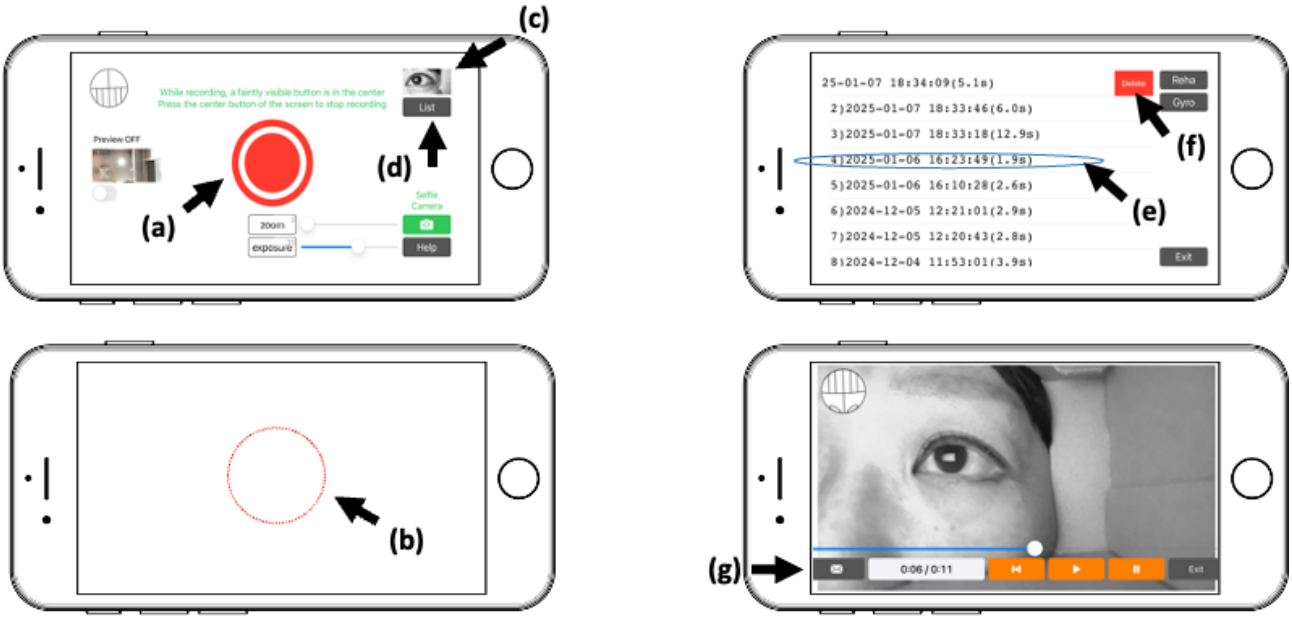

XSL•FO
RenderX

Figure 2. Eye movement video recording during vertigo attacks. (A) If one wants to record easily or can darken the room. (B) If one wants to record more reliably or if the room is brightly lit. The cardboard goggles and iPhone holder must be assembled. The iPhone must be placed in the lower-right corner of the holder and then the two parts must be attached together.

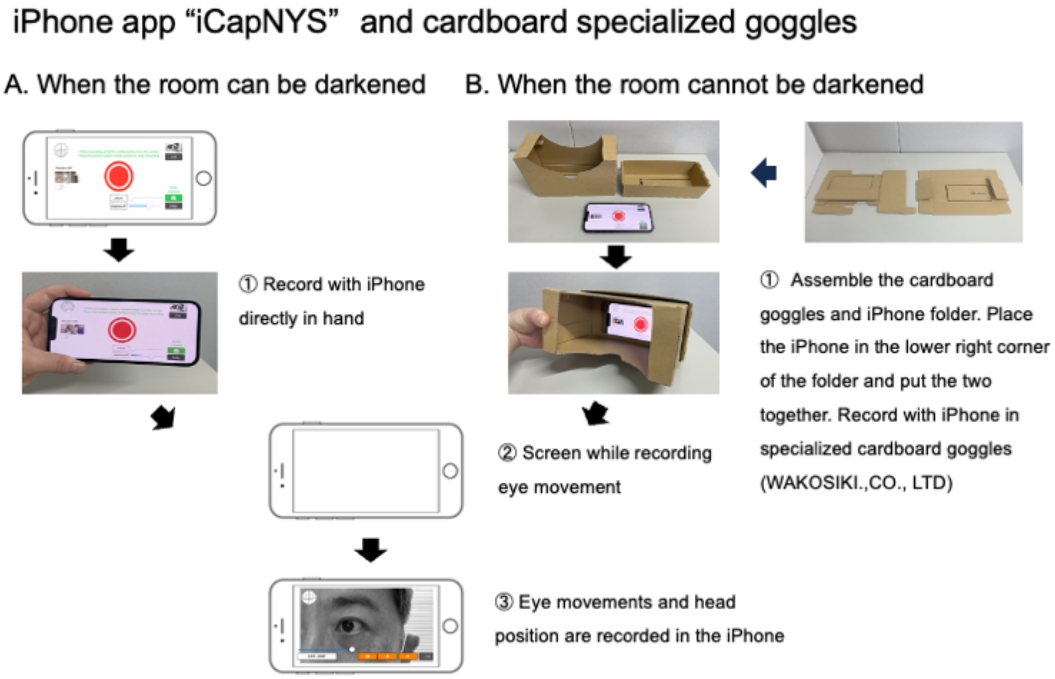

The list of videos can be displayed by tapping the button below (Figure 1D). Next, an item can be tapped to play it (Figure 1E), or the video may be deleted by swiping left (Figure 1F). Additionally, videos are saved in "iCapNYS" albums and can be managed in the Photos app. The recorded videos could be sent to medical professionals using the mail button in the lower-left corner of the video playback screen (Figure 1G).

"Cardboard Goggles" for iPhone

Specialized goggles made of recyclable cardboard (Wako Shiki Co, Ltd) were used to maintain an optimal positional relationship between the eyeball and the iPhone, ensuring more reliable recording of eye movements. Wearing these specialized goggles prevents fixation suppression caused by the surrounding field of vision. The cardboard goggles consist of 2 parts: a holder for the iPhone and the goggles themselves. These can be easily assembled by the user by folding them along the creases. The iPhone running the app was placed in the lower-right corner of the holder, and the goggles were placed on top for recording. The goggles hold the iPhone in the proper position, enabling more accurate recording of head position changes than direct handheld recording, and allowing for recording eye movements in bright light (Figure 2B). The assembly time was approximately 1 minute (Video 1 in Multimedia Appendix 1). Patients can record their eye movements during head position changes by following the automatic voice guidance by switching to the "automatic 90-seconds" mode with video commentary from the "selfie camera" in the lower-right corner of the iCapNYS app. The voice prompts the user to record eye movements for 15 seconds in the sitting position, followed by supine, right-side supine, left-side supine, and sitting positions, with eye movements recorded for 15 seconds in each position. The system supports audio in Japanese and English.

Ethical Considerations

This study was conducted with approval from the Medical Research Ethics Committee of Mejiro University (approval number 24 medicine-007). The patient was fully informed of the purpose of the study, its contents, and the handling of the findings, and written informed consent was obtained. All data were treated confidentially and reported anonymously. This study was conducted in accordance with the tenets of the Declaration of Helsinki. Consent to publish their image has been obtained from the individuals pictured in the figures and multimedia appendices. The participant received a JPY 2000 (approximately US $15) gift voucher as compensation.

Results

To assess the system's ability to detect pathological nystagmus, we compared nystagmus recordings obtained under infrared video Frenzel glasses with those captured using the iPhone app iCapNYS housed in specialized cardboard goggles (iCapNYS with goggles). The patient was an 82-year-old woman who presented with a vertigo attack attributed to left-sided Meniere disease.

Initially, no pathological nystagmus was observed during bedside examination under direct visual inspection with the naked eye. However, when the same patient was evaluated using either the infrared video Frenzel glasses or iCapNYS with goggles, fine pathological horizontal nystagmus was clearly visualized. Video recordings were performed in the supine position for approximately 30 seconds using both iCapNYS with goggles and the infrared video Frenzel glasses. Each video was recorded independently (Video 2: infrared video Frenzel glasses in Multimedia Appendix 2 and Video 3: iCapNYS with

XSL•FO
RenderX

goggles in Multimedia Appendix 3). With the patient's cooperation, the pathological nystagmus obtained with both devices was presented as an electronystagmography (ENG) waveform (Figure 3). Recordings obtained from infrared video Frenzel goggles show a distinct leftward horizontal nystagmus with a slow-phase velocity of approximately 12.6°/second (Figure 3A). Comparable waveforms were captured using the "iCapNYS" app under specialized cardboard goggles, with a similar slow-phase velocity of 12.0°/second (Figure 3B). The average number of nystagmus beats during the 30-second recording was 23 for the infrared video Frenzel glasses and 22 for iCapNYS with goggles. These findings suggest that the iCapNYS system can reliably detect subtle peripheral nystagmus and demonstrates performance comparable to that of standard infrared video systems. The specialized goggles with a white screen provided sufficient visual isolation and contributed to the successful detection of nystagmus.

Figure 3. Spontaneous nystagmus in an 82-year-old woman with vertigo. (A) Infrared CCD (charge-coupled device) camera, (B) the "iCapNYS" iPhone app in which nystagmus was simultaneously recorded with electronystagmography in the supine position. Upper panel: eye position. Lower panel: eye velocity waveform.

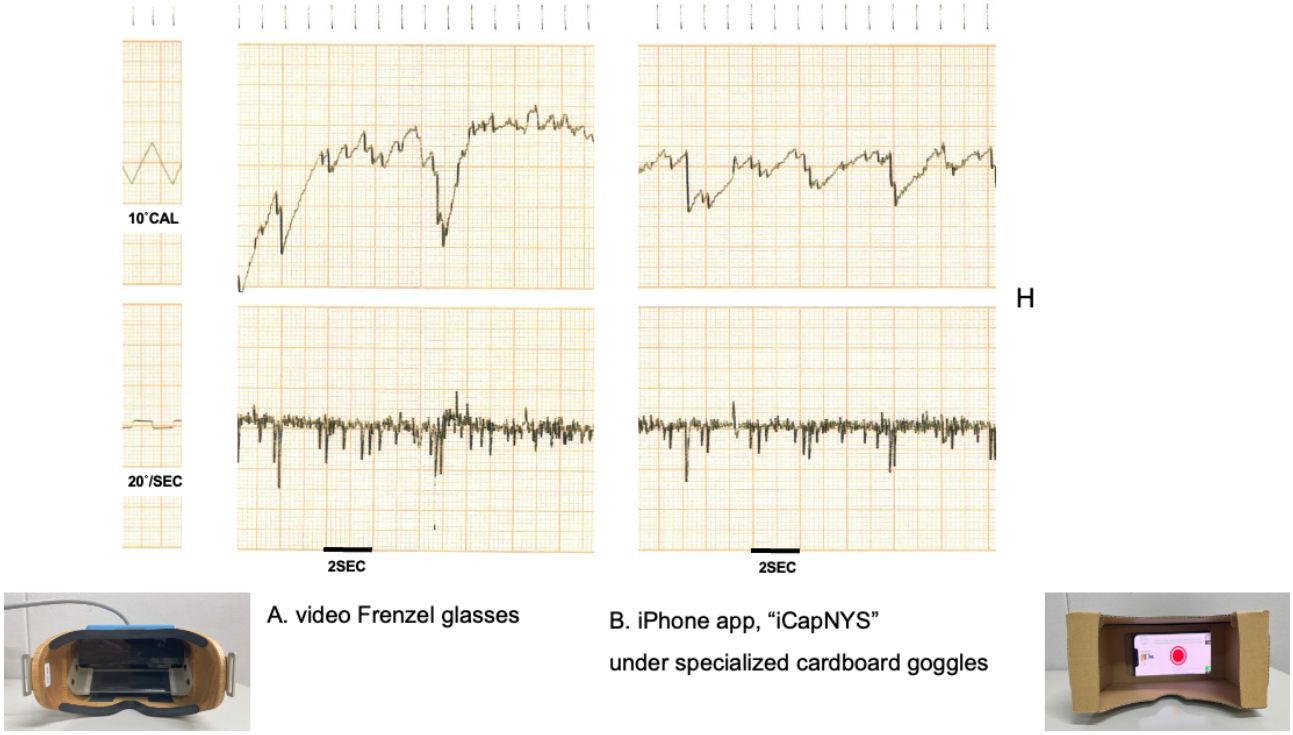

Discussion

Principal Findings

Episodic vertigo attacks can occur at any time and in any location. However, dizziness and abnormal results are not always present in patients with vertigo, and several cases exist that are difficult to differentiate or do not result in a definitive diagnosis. In such cases of episodic attacks, it is necessary to differentiate among Meniere disease, vestibular migraine, and other conditions. Therefore, it is extremely useful for patients to record their eye movements during vertigo attacks using this system and provide this information to their physicians for diagnosis and differentiation.

One approach to capturing eye movements during a vertigo attack is to use video recordings from a cell phone or smartphone. For instance, Kıroğlu and Dağkıran [11] showed that recordings made with a cell phone can be used to diagnose Meniere disease. Similarly, Young et al [14] showed that patients could record nystagmus during vertigo episodes using a device attached to an audio or video recorder that incorporated 2 infrared lights within lightweight swimming goggles. They observed that if the head position was simultaneously recorded, distinguishing between normal vestibulo-ocular reflexes and the pathological nystagmus seen in BPPV would be possible, thereby increasing diagnostic accuracy. Our system addresses these difficulties by leveraging a geolocation sensor already integrated into an iPhone, eliminating the requirement for patients to manually record their head positions.

In the diagnosis of vertigo, confirming spontaneous nystagmus under nonattentive conditions is especially critical [16]. iCapNYS was specifically designed to reduce gaze-induced nystagmus suppression, enabling patients to record spontaneous eye movements in a nonfixating state during a vertigo attack. iCapNYS does not require being in a dark room and can record even when directly held in front of the patient's eyes. Moreover, recording stable eye movements is possible using specialized cardboard goggles. Recent studies have underscored the potential of deep learning systems as well as other smartphone-based diagnostic tools for detecting nystagmus [17,18]. Shah et al [12] showed that remote specialists can use smartphone-recorded eye movements to diagnose BPPV. Our system is designed to be operated as easily and accurately as possible by patients themselves during a vertigo attack. Patients can perform practice recordings to familiarize themselves with the procedure, enabling them to record their eye movements

XSL•FO
RenderX

during vertigo attacks regardless of location or time. Additionally, the ability to monitor nystagmus enhances the quality of medical care and the accuracy of diagnosis. Although ENG and infrared video Frenzel glasses are typically used by physicians and medical personnel in clinical settings, the proposed device is designed for patient use during actual episodes of vertigo. If a medical system were established in which smartphone-captured eye movements could be emailed to health care providers, appropriate measurements could be taken based on the diagnosis. For example, if a diagnosis of BPPV is confirmed, the patient could be instructed to adopt a head position that does not trigger vertigo and be reassured not to worry. In the future, it may be possible to advise self-treatment through home-based maneuvers in a digital consultation. The introduction of this technology into telemedicine is expected to help build a sense of security and trust between patients and physicians (Figure 4).

Figure 4. iCapNYS used in episodic vertigo of unknown cause.

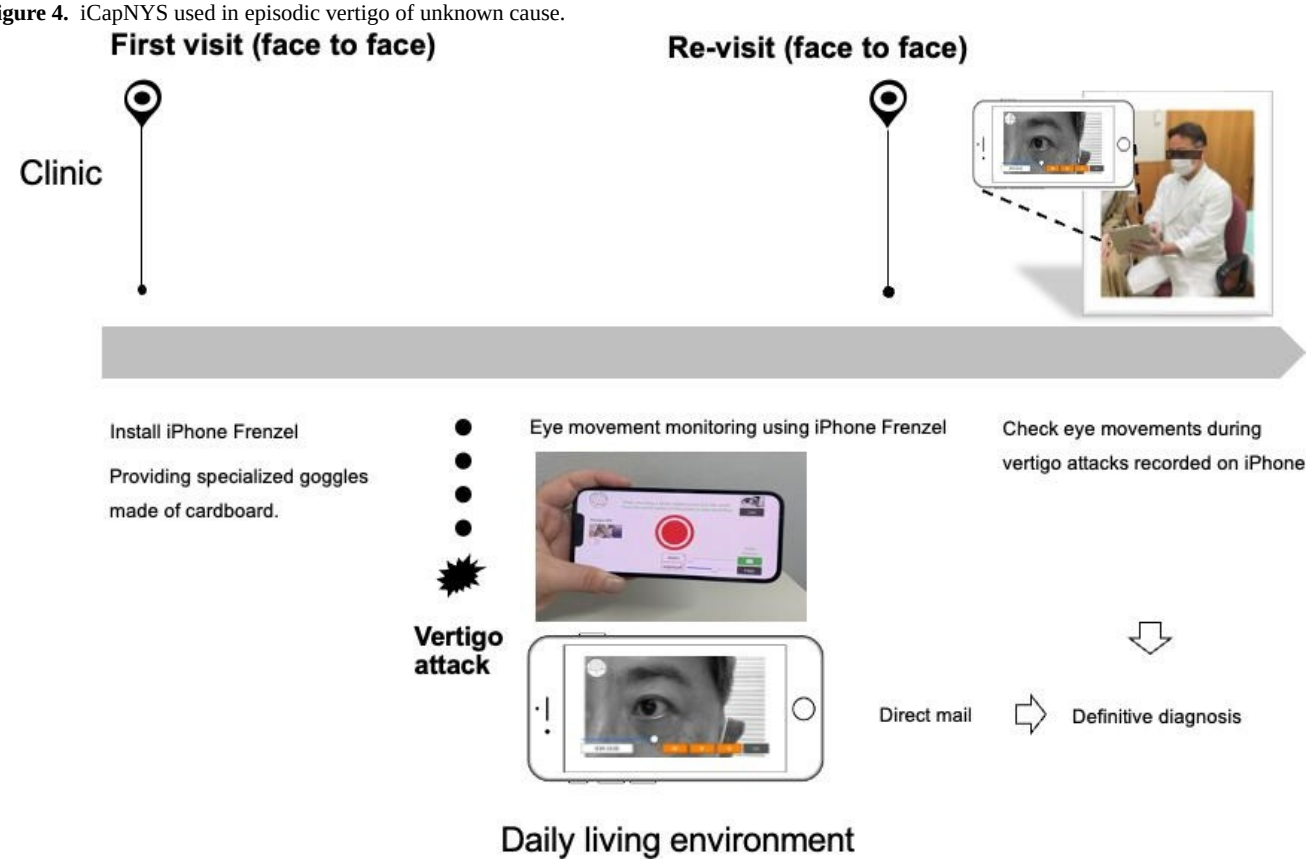

Limitations

Despite these promising developments, the proposed system has certain limitations. The present assessment was based on a single case study, which may not completely represent a broader patient population. Additionally, variations in smartphone hardware and ambient lighting conditions can influence the recording quality, and relying on patients to operate the system during an acute vertigo episode may be challenging, especially for individuals with limited technological proficiency.

The iCapNYS iPhone app, housed in specialized cardboard goggles, has the potential to record subtle peripheral nystagmus in patients with vertigo. However, the lack of infrared capabilities in most smartphones limits their use in complete darkness. Although the cardboard housing is designed to block out ambient light, some leakage may still occur.

Conclusions

Herein, we developed an innovative eye movement–recording system that uses a patient's iPhone during a vertigo attack. The system comprises an "iCapNYS" app that captures eye movements and head positions, along with a pair of specialized cardboard goggles specifically designed for the iPhone. We recorded the nystagmus of a patient during a vertigo attack and compared the results with those obtained using conventional infrared video Frenzel glasses to examine the system's effectiveness. Our findings indicate that the nystagmus frequency and slow-phase velocity recorded using the iCapNYS system were comparable to those captured using traditional methods, with similar visual characteristics observed in both sets of recordings. This system is highly beneficial for treating vertigo, as it enables patients to easily record their eye movements and share data with their physicians at any time and from any location. We anticipate that the widespread adoption of this nystagmus-monitoring system, particularly when integrated into telemedicine, will increase vertigo treatment quality and ultimately foster greater patient confidence and trust.

XSL•FO
RenderX

Acknowledgments

We thank Editage for the English language editing. This work was supported by a Japan Science and Technology Agency grant (JPMJPF2403) and by a Japan Society for the Promotion of Science KAKENHI grant (23K08970). The authors did not use any generative artificial intelligence tools in the preparation or writing of this paper.

Data Availability

The data that support the findings of this study are available from the corresponding author (MN) upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Video 1.
[AVI File, 11556 KB - formative_v9i1e73811_app1.avi]

Multimedia Appendix 2
Video 2.
[AVI File, 545 KB - formative_v9i1e73811_app2.avi]

Multimedia Appendix 3
Video 3.
[AVI File, 436 KB - formative_v9i1e73811_app3.avi]

References

1. Neuhauser HK. The epidemiology of dizziness and vertigo. Handb Clin Neurol 2016;137:67-82. [doi: 10.1016/B978-0-444-63437-5.00005-4] [Medline: 27638063]
2. Mendel B, Bergenius J, Langius A. Dizziness symptom severity and impact on daily living as perceived by patients suffering from peripheral vestibular disorder. Clin Otolaryngol Allied Sci 1999 Aug;24(4):286-293. [doi: 10.1046/j.1365-2273.1999.00261.x] [Medline: 10472462]
3. von Brevern M, Radtke A, Lezius F, et al. Epidemiology of benign paroxysmal positional vertigo: a population based study. J Neurol Neurosurg Psychiatry 2007 Jul;78(7):710-715. [doi: 10.1136/jnnp.2006.100420] [Medline: 17135456]
4. Lempert T, Olesen J, Furman J, et al. Vestibular migraine: diagnostic criteria1. J Vestib Res 2022;32(1):1-6. [doi: 10.3233/VES-201644] [Medline: 34719447]
5. Lopez-Escamez JA, Carey J, Chung WH, et al. Diagnostic criteria for Menière's disease. J Vestib Res 2015;25(1):1-7. [doi: 10.3233/VES-150549] [Medline: 25882471]
6. Sung KB, Lee TK, Furman JM. Abnormal eye movements in dizzy patients. Neurol Clin 2005 Aug;23(3):675-703. [doi: 10.1016/j.ncl.2005.01.006] [Medline: 16026671]
7. Kim JS, Newman-Toker DE, Kerber KA, et al. Vascular vertigo and dizziness: diagnostic criteria. J Vestib Res 2022;32(3):205-222. [doi: 10.3233/VES-210169] [Medline: 35367974]
8. Noda M, Kuroda T, Nomura A, Ito M, Yoshizaki T, Fushiki H. Smartphone-assisted medical care for vestibular dysfunction as a telehealth strategy for digital therapy beyond COVID-19: scoping review. JMIR Mhealth Uhealth 2023 Sep 11;11:e48638. [doi: 10.2196/48638] [Medline: 37695671]
9. Barahim Bastani P, Saber Tehrani AS, Badihian S, et al. Self-recording of eye movements in amyotrophic lateral sclerosis patients using a smartphone eye-tracking app. Digit Biomark 2024;8(1):111-119. [doi: 10.1159/000538992] [Medline: 39015513]
10. Parker TM, Badihian S, Hassoon A, et al. Eye and head movement recordings using smartphones for telemedicine applications: measurements of accuracy and precision. Front Neurol 2022;13:789581. [doi: 10.3389/fneur.2022.789581] [Medline: 35370913]
11. Kıroğlu M, Dağkıran M. The role of mobile phone camera recordings in the diagnosis of Meniere's disease and pathophysiological implications. J Int Adv Otol 2020 Apr;16(1):18-23. [doi: 10.5152/iao.2019.6605] [Medline: 32066548]
12. Shah MU, Lotterman S, Roberts D, Eisen M. Smartphone telemedical emergency department consults for screening of nonacute dizziness. Laryngoscope 2019 Feb;129(2):466-469. [doi: 10.1002/lary.27424] [Medline: 30478898]
13. Melliti A, van de Berg M, van de Berg R. Capturing nystagmus during vertigo attacks using a smartphone: adherence, characteristics, pearls and pitfalls. J Neurol 2023 Dec;270(12):6044-6056. [doi: 10.1007/s00415-023-11965-y] [Medline: 37653139]

XSL•FO
RenderX

14. Young AS, Lechner C, Bradshaw AP, et al. Capturing acute vertigo: A vestibular event monitor. Neurology 2019 Jun 11;92(24):e2743-e2753. [doi: 10.1212/WNL.0000000000007644] [Medline: 31092626]
15. Sakazaki H, Noda M, Dobashi Y, Kuroda T, Tsunoda R, Fushiki H. Monitoring nystagmus in a patient with vertigo using a commercial mini-infrared camera and 3D printer: cost-effectiveness evaluation and case report. JMIR Form Res 2025 Feb 27;9:e70015. [doi: 10.2196/70015] [Medline: 40014039]
16. Karlsen EA, Goetzinger CP, Hassanein R. Effects of six conditions of ocular fixation on caloric nystagmus. Arch Otolaryngol 1980 Aug;106(8):474-476. [doi: 10.1001/archotol.1980.00790320026007] [Medline: 7396793]
17. Wagle N, Morkos J, Liu J, et al. aEYE: A deep learning system for video nystagmus detection. Front Neurol 2022;13:963968. [doi: 10.3389/fneur.2022.963968] [Medline: 36034311]
18. van Bonn SM, Behrendt SP, Pawar BL, Schraven SP, Mlynski R, Schuldt T. Smartphone-based nystagmus diagnostics: development of an innovative app for the targeted detection of vertigo. Eur Arch Otorhinolaryngol 2022 Dec;279(12):5565-5571. [doi: 10.1007/s00405-022-07385-9] [Medline: 35451613]

Abbreviations

BPPV: benign paroxysmal positional vertigo
CCD: charge-coupled device
ENG: electronystagmography

Edited by A Mavragani; submitted 12.03.25; peer-reviewed by A Melliti, T Miwa; revised version received 14.05.25; accepted 21.05.25; published 18.06.25.

Please cite as:
Dobashi Y, Noda M, Kuroda T, Miyata N, Ito M, Tsunoda R, Fushiki H
Development of an Eco-Friendly Smartphone-Assisted Nystagmus Recording System for Recording Vertigo Attacks Anytime, Anywhere: Pilot App Development Study
JMIR Form Res 2025;9:e73811
URL: https://formative.jmir.org/2025/1/e73811
doi:10.2196/73811

© Yumi Dobashi, Masao Noda, Tatsuaki Kuroda, Noriaki Miyata, Makoto Ito, Reiko Tsunoda, Hiroaki Fushiki. Originally published in JMIR Formative Research (https://formative.jmir.org), 18.6.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

A Brief Technique to Reduce Flashbacks of Sexual Trauma in an Adolescent: Proof-of-Concept Case Study Using Imagery Interference

Alex Lau-Zhu[1,2,3,4], PhD, DClinPsy; Carmen Chan[3], DClinPsy

[1]Department of Experimental Psychology, University of Oxford, Life and Mind Building, Oxford, United Kingdom
[2]Imperial College London, London, United Kingdom
[3]Oxford Health NHS Foundation Trust, Oxford, United Kingdom
[4]Linacre College, Oxford, United Kingdom

Corresponding Author:
Alex Lau-Zhu, PhD, DClinPsy
Department of Experimental Psychology, University of Oxford, Life and Mind Building, Oxford, United Kingdom

Abstract

Background: Trauma exposure, including sexual harm, is prevalent in adolescents. A key resulting symptom relates to reexperiencing mental images of trauma, such as intrusive memories and flashbacks. Established treatments are used to address flashback memories but are hard to access, often leave remaining symptoms, and require extensive exposure to traumatic materials. An emerging approach with adult populations suggests intrusive imagery symptoms can be precisely targeted with simple cognitive tasks.

Objective: We describe a first proof-of-concept demonstration of an imagery interference technique in a 15-year-old to target residual flashback symptoms after a course of treatment for posttraumatic stress disorder (PTSD) following sexual trauma, to lay the groundwork for further evaluations.

Methods: A case study (Mia) was presented, drawing from routine clinical practice within the United Kingdom's National Health Service. After 23 sessions of trauma-focused cognitive behavioral therapy, Mia received the imagery interference technique at session 24 and returned a month later for session 25 (the final session). The technique involved playing Tetris within a wider protocol informed by the science of memory malleability (eg, including brief memory recall and working memory taxation). Memory ratings (frequency, vividness, and distress) were assessed before and immediately after training on the technique and a month later. Symptoms of PTSD, anxiety, and depression were gathered at the first and final sessions. Views from Mia and her parents were also obtained.

Results: For the specific flashback targeted by Mia, vividness reduced within the session (40% to 15%), distress reduced within the session (40% to 15%) and a month later (then to 10%), and frequency reduced a month later (once to zero times per week; 100% reduction). Nontargeted flashbacks also reduced in frequency (from 4 times to 1 time per month; 75% reduction). Mia described the memories as more "distant." Symptoms of PTSD, depression, and anxiety reduced overall.

Conclusions: Pending further rigorous testing beyond this single case, the imagery interference approach has potential as a low-intensity and early intervention for adolescents to address intrusive imagery of trauma, such as sexual harm, and also in other clinical contexts (eg, anxiety).

(JMIR Form Res 2025;9:e79708) doi:10.2196/79708

KEYWORDS

adolescent; trauma; PTSD; flashback; intrusive memories; intrusions; imagery; posttraumatic stress disorder

Introduction

One in 20 young people in the United Kingdom report sexual abuse, with global data confirming higher rates in girls than boys and in older than younger teens [1-3]. A key symptom after such traumatic events is reexperiencing in the form of intrusive memories and flashbacks—persistent, distressing, and predominantly visual mental images of the trauma that pop back into mind [4-6]. They are cardinal symptoms of posttraumatic stress disorder (PTSD) and also drive distress in conditions such as anxiety and depression across the lifespan [7,8]. The highest burden of PTSD is associated with sexual traumas in youth and adult samples [9-11]. Intrusive memory symptoms can bring substantial impairment even when diagnostic criteria for PTSD are not met (eg, concentration) [7].

XSL•FO
RenderX

Established interventions for flashbacks and intrusive memories include trauma-focused cognitive behavioral therapies (CBT) and eye movement desensitization and reprocessing (EMDR) [4]. These powerful treatments aim to treat the full constellation of PTSD symptoms, rather than just flashbacks, but they also have limitations. They require extensive exposure to traumatic materials, including talking in detail about the trauma within CBT approaches. This is a key reason adolescents are often reluctant to begin treatment after trauma [12]. Residual symptoms are also common after these treatments despite meeting recovery criteria [13,14]. Many young people struggle to access these treatments (eg, due to long waiting lists) [15]. An accessible and effective intervention to specifically target intrusions may be beneficial.

A novel idea has been proposed from memory science, which can precisely address flashbacks and intrusive memories, without requiring extensive verbal recounting of trauma; for review, see [16]. This involves using a brief task (such as the shape-fitting game Tetris) to tax working memory [17] during memory malleability (ie, when a memory trace is thought to be labile and susceptible to update or interference), for example, upon retrieval [18]. This task thus interferes with capacity-limited cognitive resources needed to maintain the image-based memory, rendering it less vivid and emotional [19]. This disrupted memory is hypothesized to be re-stored [18]. The predicted effects are supported by preclinical, experimental research conducted over almost 2 decades [7,20-22]. Notably, the underpinning theory parallels a proposed explanation for the effectiveness of EMDR [23].

Initial intervention development studies have demonstrated that this imagery interference protocol can reduce the incidence of intrusive memories after real-life traumas in adults [16], regardless of the time since trauma. Critically, the approach appears to leave memory intact—an ability to recall details of the traumatic event at will seems to be retained [21,22]. This is important because professionals are often concerned about the impact of psychological therapies on trauma memory credibility within legal investigations, which frequently follow disclosures of sexual harm [24]. However, this intervention approach has yet to be translated for adolescents in experimental or clinical contexts.

Inspired by the application of this approach in adults, we present a case study as its first proof of concept for adolescents. The intervention was used predischarge to target residual flashbacks following established PTSD treatment for sexual traumas. Our aim was to gather initial information on feasibility and preliminary signal of effectiveness (ie, a reduction in the intensity and frequency of the targeted flashback) to ensure the approach is acceptable and appealing to the young person to lay the groundwork for further formal testing.

Methods

Case Presentation

Mia (pseudonym) is a 15-year-old adolescent girl referred to a specialist team supporting young people affected by sexual harm and their families. The referral was made by a community mental health team following an attempted overdose. At assessment, she no longer experienced suicidal intent, but reported recurring and distressing reexperiencing symptoms of an abusive relationship, including intrusive memories, flashbacks, and nightmares. She coped by pushing those thoughts away and avoiding reminders, such as music she listened to a lot during the relationship. The reexperiencing symptoms affected her self-confidence, as she continued to appraise the events as being her fault. She also reported occasional urges to self-harm by cutting, although these were generally fleeting. Her main therapy goal was to experience fewer intrusive memories and flashbacks, so that her concentration in school could be improved due to upcoming exams.

Measures

For the specific flashback targeted as part of the imagery interference technique, vividness and distress ratings were obtained ("how vivid/distressing was the flashback?" anchored from "not at all [0%]" to "extremely [100%]"), based on prior experimental [25] and intervention protocols [26]; frequency estimates (for the prior week) were obtained retrospectively.

For the overall course of treatment, symptoms of PTSD were assessed using the Child PTSD Symptom Scale (CPSS; sessions 1, 12, 23, and 25) [27], symptoms of anxiety and depression with the Revised Child Anxiety and Depression Scale (RCADS; sessions 1 and 25) [28], and negative trauma-related appraisals were assessed with the Child Post-Traumatic Cognitions Inventory (CPTCI; sessions 1 and 25) [29].

Intervention

Overview of Full Treatment

Mia received 23 one-hour sessions of trauma-focused CBT [30] over the course of 6 months. These sessions involved trauma processing by focusing on 5 key hotspots, all linked to distinct episodes of harm. She described these as highly vivid and distressing, often with strong physical sensations. She then took a 1-month break from therapy due to a busy period with exams. When she returned, she reported still experiencing some residual flashback symptoms (despite marked improvement on overall PTSD symptoms), which were the main clinical target for this report.

Targeting Residual Intrusion Symptoms

The imagery interference technique (Figure 1) was delivered on session 24. Mia identified one specific hotspot memory (the specific key moment of the trauma that flashed back) that popped up in the last week and that in her view would make the most difference if it were to be addressed. She labeled this the "shock memory." This was a mental image of a hard object approaching to hit her. She first estimated the frequency of this flashback (and others) in the last week. Then, she was instructed to gently picture this hotspot in her mind—she was specifically told to do so just for a few seconds until she had a clear picture of it in her head, but without it needing to be as emotionally intense as done during reliving sessions of trauma-focused CBT. She provided ratings of vividness and distress for this hotspot.

XSL•FO
RenderX

Figure 1. Schematic of the imagery interference technique (key steps).

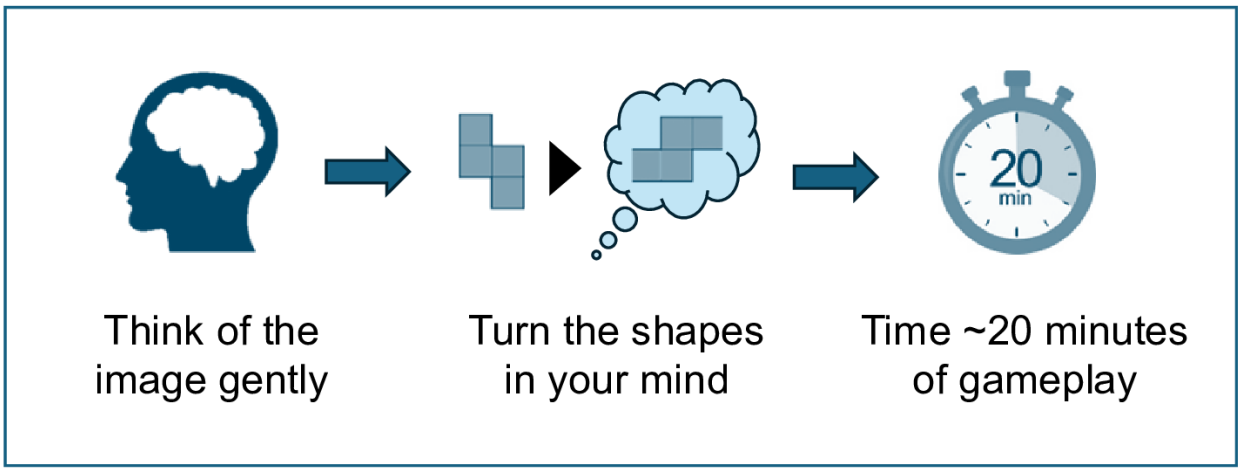

Afterward, the freely available game Tetris (Marathon version; Tetris Holding) was downloaded on her smartphone. The clinician (ALZ) emphasized instructions to play this by focusing on mental rotation of the upcoming shape and trying her best, rather than on performance or scores. This practice stage took around 10 minutes, and then Mia continued to play uninterrupted for another 10 minutes. Mia then provided ratings of vividness and distress again.

Mia was encouraged to use this technique on days whenever she had this flashback over the next few weeks and to apply this technique to other flashbacks she wanted to address. She could apply the technique immediately after having a flashback if it was convenient or schedule it for later within the same day, as long as the memory was briefly reactivated as part of the intervention (Figure 1). For the final session (session 25), Mia returned and estimated the frequency of the targeted flashbacks (and others) in the prior week. She also was instructed to gently picture the key hotspot in her mind as she did in the last session and provided ratings of vividness and distress. This intervention protocol delivered to an adolescent was based on prior work with adults [31,32].

Ethical Considerations

The intervention ran as part of routine clinical care within the United Kingdom National Health Service, thus ethical approval was not needed. Parental and child consent were obtained for the study and publication. To maintain patient confidentiality, a pseudonym has been used in this report. No compensation was provided to the patient. We have abided by the Ethical Principles of Psychologists and Code of Conduct as set out by the British Association of Behavioral and Cognitive Therapies and the British Psychological Society.

Results

Flashback Ratings

Mia reported that the “shock memory” intruded once a week prior to the imagery interference intervention session (equivalent to 4 times in a 1-month period) and rated it as 40% vivid and 40% distressing when she gently brought it to mind as part of the first step of the intervention. She also reported flashbacks of other content, but these nontargeted ones were less frequent overall and together occurred about once a week (equivalent to 4 times in a 1-month period). Note these initial ratings were collected after finishing the preceding 23 sessions of trauma-focused CBT.

After the intervention, she pictured the memory in her mind again and rated it as 15% vivid (65% reduction) and 15% distressing (65% reduction).

A month later, at the follow-up, Mia explained that she managed to engage with the intervention 3 times the week immediately after the initial training. She reported the “shock memory” had not intruded at all in the prior week, nor in the whole period since the intervention session (100% reduction). She gently pictured the memory and rated it as 40% vivid (no reduction) and 10% distressing (75% reduction). She also reported that other flashbacks, which were not targeted in the session, intruded only once overall since the last session (75% reduction)—when this happened, it was “not powerful enough” to cause her significant disruption.

Mia also reported that she had used the technique in contexts unrelated to trauma memories. She applied it to intrusive mental images of social situations, which triggered feelings of “embarrassment.” She found that the technique reduced the emotional intensity of the image afterward. She also said she enjoyed having it as a “grounding” strategy for dealing with anxiety-provoking situations in general.

Clinical Measures

Sessions 1 - 23 corresponded to the trauma-focused CBT phase, while sessions 24 - 25 (final) corresponded to the imagery interference phase.

On sessions 1, 12, and 23, scores on the CPSS were 54, 32, and 23, respectively, demonstrating a progressive decline in PTSD symptoms toward under the clinical cutoff (>31) during the CBT phase. The final score on session 25 was 22, suggesting a drop of only 1 point during the imagery interference phase.

As the imagery interference phase specifically targeted flashbacks and intrusive memories (rather than full PTSD as with the CBT phase), one relevant item was identified from the

XSL•FO
RenderX

intrusion subscale of the CPSS, specific to intrusive images. Item 1 described "Having upsetting thoughts or pictures about it that came into your head when you didn't want them to." Before the imagery interference phase, Mia's score on this item was 2 (referring to 2 - 3 times per week or sometimes), but 1 (referring to once a week or less or a little) at the end. Although she only experienced a single flashback in the month prior (see Flashback Ratings), a score of 0 was not appropriate, as this would indicate the high threshold of "not at all" for the entire month.

Scores on the RCADS at session 1 were 63 for anxiety (below clinical cutoff) and 73 for depression (above cutoff); at session 25 (final), scores were 55 for anxiety and 56 for depression (both below cutoff). The score on the CPTCI declined from 69 at session 1 to 59 at session 25 (final), largely driven by Mia's reduced conviction that trauma had permanently damaged the self and future.

Qualitative Feedback

Mia said she found the imagery interference technique "helpful" for managing her remaining intrusion symptoms. A week after practicing the technique in session, she said that the memory no longer had the sense of "nowness." Instead, she experienced more sadness, like she was "watching a TV show" as if "it happened to a friend; seeing someone I care about rather than myself." She noted how strange it felt to be able to deliberately recall the memory without feeling overwhelmed.

Mia also described feeling able to concentrate more and experiencing a renewed sense of self-confidence. She was able to enjoy her time spent with friends and to return to her hobbies, such as playing the guitar. She said "I'm not a broken person" and "if I were to have a crash, I can get out of it," for example, by using the technique. The gaming element of Tetris was also highlighted, as it was a "light-hearted" way to deal with difficult material and was easily accessible on her smartphone. She has also been free of self-harm for 87 days (which she was self-tracking), which had been her longest period in the last 2 years. Her parents also noted a change in Mia's self-confidence, such as observing that she talks to adults more proactively and can comfortably make jokes.

Discussion

Summary of Findings

We presented the first application of an imagery interference technique to address symptoms of flashbacks and intrusive memories in the adolescence context, inspired by initial clinical demonstrations in adult populations. The intervention protocol involved playing the game Tetris on a smartphone, including brief memory recall and sustained working memory engagement, informed by memory science. Thus, the intervention is not just about playing Tetris. The clinical aim was to reduce residual reexperiencing symptoms following full PTSD treatment for sexual harm.

Our findings showed that the adolescent (Mia) found the intervention appealing and was willing to try it; she reported changes in key properties of flashback memories (including frequency and distress) within the intervention session and at a 1-month follow up, with subjective accounts of perceived benefits. Overall improvements in symptoms of PTSD, depression, anxiety, trauma-related appraisals, and self-harm were also observed in relation to the full therapeutic arc, which included preceding trauma-focused CBT prior to introducing the imagery interference technique.

Clinical Implications

These preliminary adolescent data align with emerging laboratory [21,22] and clinical findings with adults [16]. The brief and precise nature of the protocol meant that minimal adaptations to the adolescent context were needed, especially as it does not require detailed discussions of trauma (eg, sexual abuse). The game- and phone-based nature of the intervention made it particularly accessible to the adolescent.

As this case illustrated, the aim of the technique was not to replace an established treatment approach, which we used as first-line treatment (ie, trauma-focused CBT). A relatively simple technique could provide an additional tool for managing distressing symptoms, such as residual ones common after trauma-focused treatments [13,14]. Given problems with access in child and adolescent mental health services [15], the technique presented here holds potential as an early intervention; for example, while young people are waiting for treatment or in the form of a scalable, single-session intervention [33]. Many adolescents are often reluctant to initiate trauma-focused therapies [12]; thus, this technique may also facilitate a gentler initial engagement with trauma memory.

Although clinical guidelines recommend CBT—often delivered by highly-trained therapists—within the first 3 months after trauma [34], many young people's first professional contact after disclosures of sexual harm is with nonspecialists, such as teachers and social workers. Interventions that can be more easily deployed have potential for scalability, including for high-risk and low-resource contexts. However, despite the seemingly simple nature of the technique, appropriate training and supervision may still be important as with other low-intensity interventions [35]. A game-based technique like this may also be harnessed as an adjunct within established treatment for out-of-session practice, given the frequent challenges with therapy homework with adolescents [36].

Limitations

The changes observed may not be solely attributed to the imagery interference technique. Other explanations for the observed effects on intrusive memories should be considered (eg, a continuation of improvement after 23 sessions of CBT). The single case design was limited by the routine setup, in collaboration with the adolescent, rather than designed as a research study. Building on these initial data supporting acceptability, stronger designs for causal inference should be used in the future, such as single-case experimental designs, which can better account for time trends prior to introducing the intervention.

The reduction in intrusive memory indices was not easily captured by the CPSS, an established questionnaire for PTSD symptoms. However, this only captured intrusive memories with a single item, which may not be sufficiently sensitive.

XSL•FO
RenderX

Novel measurement approaches to capture changes in intrusive memories could be considered, such as diary methods and validated scales.

Conclusions

Early intrusive symptoms are important to ameliorate to support key developmental tasks, such as school learning. Our preliminary data lay the groundwork to evaluate the use of this imagery interference technique in adolescents, so we can better attribute any observed benefits to the technique itself rather than other therapeutic elements. Beyond sexual harm, the benefits on flashbacks and intrusions are predicted to extend to other types of trauma experienced by young people, as the focus is on shared memory mechanisms. Finally, as Mia used the technique for addressing anxiety, where intrusive imagery is also prominent [8], the technique merits additional explorations in a range of clinical contexts relevant to youth mental health.

Acknowledgments

We are grateful for Mia for her bravery in sharing her story with us and her generosity in allowing us to share these findings.

Funding

ALZ is funded by a Medical Research Council Clinician Scientist Fellowship (MR/Y009460/1).

Data Availability

The data are available from the corresponding author on reasonable request.

Authors' Contributions

ALZ: conceptualization, formal analysis, investigation, writing – original draft; CC: supervision, writing – reviewing & editing.

Conflicts of Interest

None declared.

References

1. Radford L, Corral S, Bradley C, Bassett C, Collishaw S. Child abuse and neglect in the UK today. NSPCC Learning. 2011. URL: https://learning.nspcc.org.uk/research-resources/pre-2013/child-abuse-neglect-uk-today [accessed 2025-06-26]
2. Karsna K, Kelly L. The Scale and Nature of Child Sexual Abuse: Review of Evidence (Revised Edition): Barkingside; 2021. [doi: 10.47117/OBKC1345]
3. Piolanti A, Schmid IE, Fiderer FJ, Ward CL, Stöckl H, Foran HM. Global prevalence of sexual violence against children: a systematic review and meta-analysis. JAMA Pediatr 2025 Mar 1;179(3):264-272. [doi: 10.1001/jamapediatrics.2024.5326] [Medline: 39804632]
4. Brewin CR, Atwoli L, Bisson JI, Galea S, Koenen K, Lewis-Fernández R. Post-traumatic stress disorder: evolving conceptualization and evidence, and future research directions. World Psychiatry 2025 Feb;24(1):52-80. [doi: 10.1002/wps.21269] [Medline: 39810662]
5. Steil R, Fischer A, Gutermann J, Rosner R. Mental imagery in adolescent PTSD patients after child abuse: a comparison with matched healthy controls. BMC Psychiatry 2022 Jan 27;22(1):64. [doi: 10.1186/s12888-022-03706-8] [Medline: 35086493]
6. McKinnon AC, Nixon RDV, Brewer N. The influence of data-driven processing on perceptions of memory quality and intrusive symptoms in children following traumatic events. Behav Res Ther 2008 Jun;46(6):766-775. [doi: 10.1016/j.brat.2008.02.008] [Medline: 18405880]
7. Iyadurai L, Visser RM, Lau-Zhu A, et al. Intrusive memories of trauma: a target for research bridging cognitive science and its clinical application. Clin Psychol Rev 2019 Apr;69:67-82. [doi: 10.1016/j.cpr.2018.08.005] [Medline: 30293686]
8. Schwarz S, Grasmann D, Schreiber F, Stangier U. Mental imagery and its relevance for psychopathology and psychological treatment in children and adolescents: a systematic review. J Cogn Ther 2020 Dec;13(4):303-327. [doi: 10.1007/s41811-020-00092-5]
9. Kessler RC, Aguilar-Gaxiola S, Alonso J, et al. Trauma and PTSD in the WHO World Mental Health Surveys. Eur J Psychotraumatol 2017;8(sup5):1353383. [doi: 10.1080/20008198.2017.1353383] [Medline: 29075426]
10. McLaughlin KA, Koenen KC, Hill ED, et al. Trauma exposure and posttraumatic stress disorder in a national sample of adolescents. J Am Acad Child Adolesc Psychiatry 2013 Aug;52(8):815-830. [doi: 10.1016/j.jaac.2013.05.011] [Medline: 23880492]
11. Nöthling J, Simmons C, Suliman S, Seedat S. Trauma type as a conditional risk factor for posttraumatic stress disorder in a referred clinic sample of adolescents. Compr Psychiatry 2017 Jul;76:138-146. [doi: 10.1016/j.comppsych.2017.05.001] [Medline: 28521252]

XSL•FO
RenderX

12. Hanson RF, Gros KS, Davidson TM, et al. National trainers' perspectives on challenges to implementation of an empirically-supported mental health treatment. Adm Policy Ment Health 2014 Jul;41(4):522-534. [doi: 10.1007/s10488-013-0492-6] [Medline: 23605292]
13. Larsen SE, Bellmore A, Gobin RL, Holens P, Lawrence KA, Pacella-LaBarbara ML. An initial review of residual symptoms after empirically supported trauma-focused cognitive behavioral psychological treatment. J Anxiety Disord 2019 Apr;63:26-35. [doi: 10.1016/j.janxdis.2019.01.008] [Medline: 30825720]
14. Harpviken AN, Jensen TK, Johnson SU, Ormhaug SM, Birkeland MS. Assessing change and persistence of specific post-traumatic stress symptoms among youth in trauma treatment. Eur J Psychotraumatol 2025 Dec;16(1):2515683. [doi: 10.1080/20008066.2025.2515683] [Medline: 40631373]
15. Radez J, Reardon T, Creswell C, Lawrence PJ, Evdoka-Burton G, Waite P. Why do children and adolescents (not) seek and access professional help for their mental health problems? A systematic review of quantitative and qualitative studies. Eur Child Adolesc Psychiatry 2021 Feb;30(2):183-211. [doi: 10.1007/s00787-019-01469-4] [Medline: 31965309]
16. Highfield J, Iyadurai L, Holmes EA. A summary review of the development of using a brief imagery-competing task intervention (ICTI) for reducing intrusive memories of psychological trauma: applications in healthcare settings for both staff and patients. Discov Ment Health 2025 May 27;5(1):78. [doi: 10.1007/s44192-025-00205-6] [Medline: 40425895]
17. Lau-Zhu A, Holmes EA, Butterfield S, Holmes J. Selective association between tetris game play and visuospatial working memory: a preliminary investigation. Appl Cogn Psychol 2017;31(4):438-445. [doi: 10.1002/acp.3339] [Medline: 29540959]
18. Visser RM, Lau-Zhu A, Henson RN, Holmes EA. Multiple memory systems, multiple time points: how science can inform treatment to control the expression of unwanted emotional memories. Philos Trans R Soc Lond B Biol Sci 2018 Mar 19;373(1742):20170209. [doi: 10.1098/rstb.2017.0209] [Medline: 29352036]
19. Baddeley AD, Andrade J. Working memory and the vividness of imagery. J Exp Psychol Gen 2000 Mar;129(1):126-145. [doi: 10.1037//0096-3445.129.1.126] [Medline: 10756490]
20. James EL, Lau-Zhu A, Clark IA, Visser RM, Hagenaars MA, Holmes EA. The trauma film paradigm as an experimental psychopathology model of psychological trauma: intrusive memories and beyond. Clin Psychol Rev 2016 Jul;47:106-142. [doi: 10.1016/j.cpr.2016.04.010] [Medline: 27289421]
21. Lau-Zhu A, Henson RN, Holmes EA. Intrusive memories and voluntary memory of a trauma film: differential effects of a cognitive interference task after encoding. J Exp Psychol Gen 2019 Dec;148(12):2154-2180. [doi: 10.1037/xge0000598] [Medline: 31021150]
22. Lau-Zhu A, Henson RN, Holmes EA. Selectively interfering with intrusive but not voluntary memories of a trauma film: accounting for the role of associative memory. Clin Psychol Sci 2021 Nov;9(6):1128-1143. [doi: 10.1177/2167702621998315] [Medline: 34777922]
23. van den Hout MA, Engelhard IM. How does EMDR work? J Exp Psychopathol 2012 Dec;3(5):724-738. [doi: 10.5127/jep.028212]
24. Otgaar H, Howe ML, Muris P, Merckelbach H. Dealing with false memories in children and adults: recommendations for the legal arena. Policy Insights Behav Brain Sci 2019 Mar;6(1):87-93. [doi: 10.1177/2372732218818584]
25. Rackham LA, Lau-Zhu A. Taxing working memory to modulate mental imagery of the 9/11 terrorist attacks following media exposure during childhood: a pilot study in young adult UK residents. Anxiety Stress Coping 2021 Jul;34(4):423-436. [doi: 10.1080/10615806.2020.1870107] [Medline: 33412932]
26. Lau-Zhu A, Tuxen N, Roerne ML, Di Simplicio M. Imagery-based Cognitive Therapy for comorbid anxiety in bipolar disorder: two case studies in Denmark. Psychiatry Research Case Reports 2023 Jun;2(1):100124. [doi: 10.1016/j.psycr.2023.100124]
27. Foa EB, Johnson KM, Feeny NC, Treadwell KRH. The child PTSD Symptom Scale: a preliminary examination of its psychometric properties. J Clin Child Psychol 2001 Sep;30(3):376-384. [doi: 10.1207/S15374424JCCP3003_9] [Medline: 11501254]
28. Chorpita BF, Moffitt CE, Gray J. Psychometric properties of the Revised Child Anxiety and Depression Scale in a clinical sample. Behav Res Ther 2005 Mar;43(3):309-322. [doi: 10.1016/j.brat.2004.02.004] [Medline: 15680928]
29. Meiser-Stedman R, Smith P, Bryant R, et al. Development and validation of the Child Post-Traumatic Cognitions Inventory (CPTCI). J Child Psychol Psychiatry 2009 Apr;50(4):432-440. [doi: 10.1111/j.1469-7610.2008.01995.x] [Medline: 19338628]
30. Smith P, Perrin S, Yule W, Clark DM. Post Traumatic Stress Disorder: Cognitive Therapy with Children and Young People, 1st edition: Taylor and Francis; 2014. [doi: 10.4324/9780203087190]
31. Iyadurai L, Hales SA, Blackwell SE, Young K, Holmes EA. Targeting intrusive imagery using a competing task technique: a case study. Behav Cogn Psychother 2020 Nov;48(6):739-744. [doi: 10.1017/S1352465820000296] [Medline: 32594968]
32. Holmes EA, Hales S, Young K. Imagery-Based Cognitive Therapy for Bipolar Disorder and Mood Instability: The Guildford Press; 2019.
33. Schleider JL, Zapata JP, Rapoport A, et al. Single-session interventions for mental health problems and service engagement: umbrella review of systematic reviews and meta-analyses. Annu Rev Clin Psychol 2025 May;21(1):279-303. [doi: 10.1146/annurev-clinpsy-081423-025033] [Medline: 39874601]
34. Post-traumatic stress disorder. National Institute for Health and Care Excellence. URL: https://www.nice.org.uk/guidance/ng116 [accessed 2025-11-06]

XSL•FO
RenderX

35. Roach A, Cullinan S, Shafran R, Heyman I, Bennett S. Implementing brief and low-intensity psychological interventions for children and young people with internalizing disorders: a rapid realist review. Br Med Bull 2023 Apr 5;145(1):120-131. [doi: 10.1093/bmb/ldad001] [Medline: 36715209]
36. Gaynor ST, Lawrence PS, Nelson-Gray RO. Measuring homework compliance in cognitive-behavioral therapy for adolescent depression: review, preliminary findings, and implications for theory and practice. Behav Modif 2006 Sep;30(5):647-672. [doi: 10.1177/0145445504272979] [Medline: 16894234]

Abbreviations

CBT: cognitive behavioral therapy
CPSS: Child PTSD Symptom Scale
CPTCI: Child Post-Traumatic Cognitions Inventory
EMDR: eye movement desensitization and reprocessing
PTSD: posttraumatic stress disorder
RCADS: Revised Children's Anxiety and Depression Scale

Edited by A Mavragani; submitted 24.08.25; peer-reviewed by AJ Mussap, O Abasilim; revised version received 27.10.25; accepted 28.10.25; published 12.11.25.

Please cite as:
Lau-Zhu A, Chan C
A Brief Technique to Reduce Flashbacks of Sexual Trauma in an Adolescent: Proof-of-Concept Case Study Using Imagery Interference
JMIR Form Res 2025;9:e79708
URL: https://formative.jmir.org/2025/1/e79708
doi:10.2196/79708

© Alex Lau-Zhu, Carmen Chan. Originally published in JMIR Formative Research (https://formative.jmir.org), 12.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Monitoring Nystagmus in a Patient With Vertigo Using a Commercial Mini-Infrared Camera and 3D Printer: Cost-Effectiveness Evaluation and Case Report

Hiroyuki Sakazaki[1], SLP, MPAS; Masao Noda[1,2*], MD, MBA, PhD; Yumi Dobashi[1], MT; Tatsuaki Kuroda[1,3], MD; Reiko Tsunoda[1], MD, PhD; Hiroaki Fushiki[1*], MD, PhD

[1]Department of Otolaryngology, Mejiro University Ear Institute Clinic, 320 Ukiya, Iwatsuki-ku, Saitama-shi, Saitama, Japan

[2]Department of Otolaryngology and Head and Neck Surgery, Jichi Medical University, Tochigi, Japan

[3]Kuroda ENT Clinic, Yatsushiro-shi, Kumamoto, Japan

[*]these authors contributed equally

Corresponding Author:
Masao Noda, MD, MBA, PhD
Department of Otolaryngology, Mejiro University Ear Institute Clinic, 320 Ukiya, Iwatsuki-ku, Saitama-shi, Saitama, Japan

Abstract

Background: Observing eye movements during episodic vertigo attacks is crucial for accurately diagnosing vestibular disorders. In clinical practice, many cases lack observable symptoms or clear findings during outpatient examinations, leading to diagnostic challenges. An accurate diagnosis is essential for timely treatment, as conditions such as benign paroxysmal positional vertigo (BPPV), Ménière's disease, and vestibular migraine require different therapeutic approaches.

Objective: This study aimed to develop and evaluate a cost-effective diagnostic tool that integrates a mini-infrared camera with 3D-printed goggles, enabling at-home recording of nystagmus during vertigo attacks.

Methods: A commercially available mini-infrared camera (US $25) was combined with 3D-printed goggles (US $13) to create a system for recording eye movements in dark conditions. A case study was conducted on a male patient in his 40s who experienced recurrent episodic vertigo.

Results: Initial outpatient evaluations, including oculomotor and vestibular tests using infrared Frenzel glasses, revealed no spontaneous or positional nystagmus. However, with the proposed system, the patient successfully recorded geotropic direction-changing positional nystagmus during a vertigo attack at home. The nystagmus was beating distinctly stronger on the left side down with 2.0 beats/second than the right side down with 1.2 beats/second. Based on the recorded videos, a diagnosis of lateral semicircular canal-type BPPV was made. Treatment with the Gufoni maneuver effectively alleviated the patient's symptoms, confirming the diagnosis. The affordability and practicality of the device make it particularly suitable for telemedicine and emergency care applications, enabling patients in remote or underserved areas to receive accurate diagnoses.

Conclusions: The proposed system demonstrates the feasibility and utility of using affordable, accessible technology for diagnosing vestibular disorders outside of clinical settings. By addressing key challenges, such as the absence of symptoms during clinical visits and the high costs associated with traditional diagnostic tools, this device offers a practical solution for real-time monitoring and accurate diagnosis. Its potential applications extend to telemedicine, emergency settings, and resource-limited environments. Future iterations that incorporate higher-resolution imaging and automated analysis could further enhance its diagnostic capabilities and usability across diverse patient populations.

(JMIR Form Res 2025;9:e70015) doi:10.2196/70015

KEYWORDS

dizziness; vertigo; smartphone; BPPV; telemedicine; 3D-printer

Introduction

Vertigo is a common ailment experienced at all ages, with an estimated lifetime prevalence of 3% - 10% [1]. It is symptomatic of peripheral vestibular, central, and psychogenic disorders and circulatory dysfunction, as well as life-threatening conditions, such as cerebrovascular disease, making accurate diagnosis clinically significant [2,3]. Even when not life-threatening, episodic vertigo attacks accompanied by nausea and vomiting can substantially impair an individual's quality of life [4].

Recently, the use of telemedicine has expanded across various fields owing to the rapid spread of information and communication technology, the widespread use of smartphones and tablets, and advancements in data communication

XSL•FO
RenderX

technology [5]. The COVID-19 epidemic further accelerated the adoption of telemedicine in clinical care [6]. Consequently, the use of smartphones for telemedicine in vertigo treatment has increased [7-10]. For example, Shah et al [11] used smartphone-recorded eye movement videos to diagnose benign paroxysmal positional vertigo (BPPV). Additionally, we developed a recording and analysis system for video head impulse testing using smartphones [12]; this system was found to be effective when field physicians collaborated with remote specialists, especially in emergencies.

Observing eye movements during vertigo attacks is critical for accurately diagnosing recurrent vestibular disorders [13-15]. Young et al [15] demonstrated that patient-initiated monitoring of vestibular events is feasible and can lead to rapid and accurate diagnoses of episodic vestibular disorders. The characteristics of nystagmus observed in conditions such as Ménière's disease, vestibular migraine, and BPPV can aid in diagnosis, even in remote settings. This approach is also valuable in emergency departments, bridging the gap between frontline physicians and specialists [12]. Continuous monitoring of nystagmus may help differentiate between Ménière's disease and vestibular migraine based on variations in nystagmus duration [14]. However, in clinical practice, symptoms often resolve before examination, making diagnosis challenging. Independent at-home recording of eye movements during vertigo attacks provides critical diagnostic information, benefiting both patients and physicians. Although smartphone recordings have been used to diagnose conditions such as Ménière's disease [16], they may not always adequately capture nystagmus due to suppression by the patient's gaze. Phillips et al [17] demonstrated a case in which the continuous ambulatory vestibular assessment (CAVA) device, which does not require a smartphone, successfully recorded nystagmus throughout an entire Ménière's attack. Their study highlighted the device's utility in predicting vertigo attack onset and diagnosing Ménière's disease [17].

By integrating a commercially available mini-infrared camera with 3D-printed goggles, we devised a method that allows patients to record eye movements during vertigo attacks in darkness at home and share the recordings with their physicians. Herein, we report a case demonstrating the diagnostic utility of this method.

Methods

Ethical Considerations

This study was approved by the Medical Research Ethics Committee of Mejiro University (approval number Medical 21 - 005). Written informed consent was obtained from the patient after providing a sufficient explanation regarding the protection of personal information. All data collected in this study were anonymized, ensuring that no personally identifiable information was included in the analysis or publication. CARE (Case Report) EQUATOR (Enhancing the Quality and Transparency of Health Research) guidelines were followed. No financial or material compensation was provided to the participant.

Recruitment Procedures

Participants for this study were recruited from the otolaryngology outpatient clinic at Mejiro University, a tertiary center specializing in the diagnosis and treatment of vestibular disorders. The patient included in this case report was referred to the clinic after experiencing recurrent episodes of vertigo that had gone undiagnosed in a primary care setting. During the recruitment process, patients presenting with vertigo as their primary symptom were screened based on their medical history, physical examination, and routine vestibular tests. Those who were unable to obtain a definitive diagnosis through standard diagnostic procedures, such as infrared video Frenzel glasses or vestibular function tests, were considered eligible for inclusion. The patient selected for this study provided written informed consent after receiving a detailed explanation of the study's objectives, the operation of the device, and measures to protect personal information.

Development of the Device

A mini-infrared camera equipped with an infrared imaging mode was selected to capture spontaneous nystagmus under low-light conditions (Wireless Mini Camera Model: Q15; Shenzhen Aobo Technology Co, Ltd; Figure 1A). The mini-infrared camera, costing approximately US $25, boasts a resolution of 1920×1080 pixels, a shooting angle of 150°, video encoding of H.264, and dimensions of 32.5 mm in width, 33.0 mm in height, and 34.2 mm in depth. Using Wi-Fi to start and stop recordings, patients can simultaneously and conveniently produce video and audio recordings using their smartphones.

XSL•FO
RenderX

Figure 1. The mini-infrared camera and handmade goggles designed for recording eye movements during vertigo attacks. This system was developed as part of a study to enhance diagnostic support for patients with episodic vertigo at Mejiro University Ear Institute Clinic. (A) A mini-infrared camera designed for portability and affordability. (B) Goggles featuring slots to securely attach the wireless mini-infrared camera on each side, ensuring stable recording during vertigo episodes. (C) A computer-aided design (CAD) diagram of the goggles, optimized for 3D printing compatibility. (D) The final output of the goggles produced using a 3D printer with a filament.

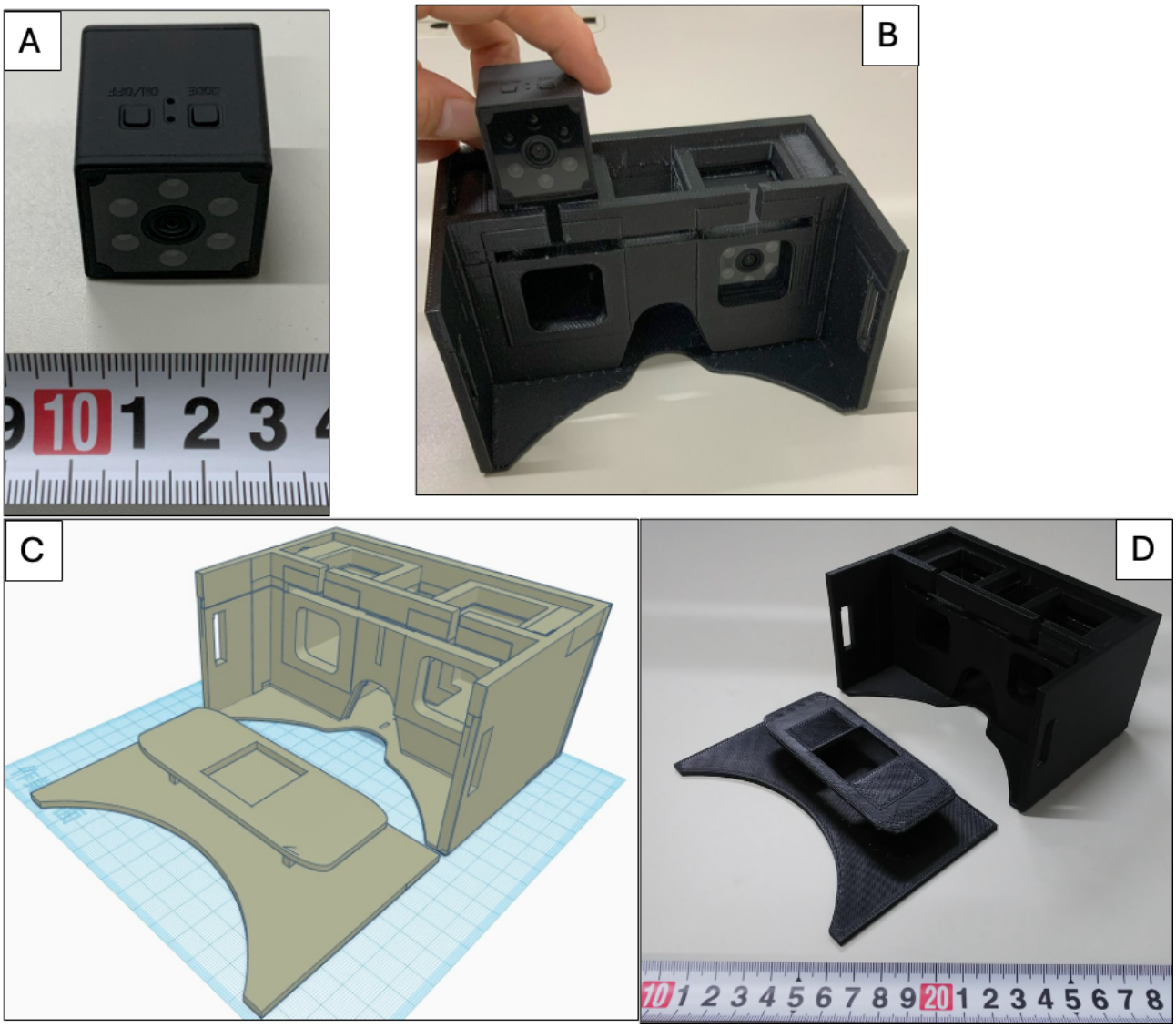

Slots were placed inside the goggles on the left and right sides, and the center of the mini-infrared camera was inserted in the slot aligned with the line of sight (Figure 1B). Both eyes were simultaneously photographed using 2 mini-infrared cameras. We designed the goggles using Tinkercad, a web-based 3D modeling tool, and an AFINIA H+1 3D printer (Microboards Technology Inc; Figure 1C). Black filament AFINIA ABS Premium Plus (Microboards Technology Inc) was used for printer output to preserve the darkness (Figure 1D). Eighteen pairs of goggles were produced, with filament costs of approximately US $13 per pair. The construction data for the 3D printer are available for download [18]. These data can be used to create goggles compatible with other mini-infrared cameras.

Operation of the Device

To operate the mini-infrared camera, the patient installed an application (HDSPCAM APP) on his smartphone or tablet. He paired the smartphone/tablet with the mini-infrared camera and selected the infrared imaging mode. During vertigo attacks, he placed the mini-infrared camera in the goggles, pressed the start button on the paired smartphone or tablet, and placed the goggles on his face. The captured images were automatically stored via Wi-Fi on the patient's smartphone or tablet (Figure 2).

The handmade goggles were used by the patient to record eye movements during vertigo episodes at home under dark conditions. The videos were wirelessly transferred to and stored on a smartphone for later analysis by physicians. This system was used in a case study of a male patient in his 40s with recurrent episodic vertigo.

XSL•FO
RenderX

Figure 2. Use of the handmade goggles.

Results

Case Presentation

A man in his 40s with episodic vertigo visited our clinic. Initial assessments using infrared video Frenzel glasses showed no spontaneous or positional nystagmus, and oculomotor-vestibular function tests revealed no abnormalities; therefore, a definitive diagnosis was not reached. We rented him our self-made 3D-printed goggles and a mini-infrared camera to record the patient's eye movements during vertigo. We also asked him to verbally record his posture when he changed positions (eg, supine, lower left, supine, and lower right).

Assessment Results

On reviewing the videos, geotropic direction-changing positional nystagmus was observed, and lateral semicircular canal-type BPPV was diagnosed (Figure 3). The nystagmus in the first 5 seconds after the change in head position was beating distinctly stronger on the left side down with 2.0 beats/second than on the right side down with 1.2 beats/second. The affected ear was identified from the nystagmus's appearance, direction, and intensity in different positions using the audio recordings. Treatment with the Gufoni maneuver alleviated the vertigo and nystagmus.

XSL•FO
RenderX

Figure 3. Eye movements recorded by the patient during a vertigo attack. Geotropic direction-changing positional nystagmus was observed. (**A**) A still image extracted from the video recording, captured in a dark environment to avoid nystagmus suppression. (**B**) Because video clips cannot be shown in this paper, we instead present the waveforms of the eye movements. The ocular position was measured using the corneal reflex method to illustrate the recorded nystagmus. OpenCV (Intel Corp) was used for image processing. H: horizontal component. V: vertical component.

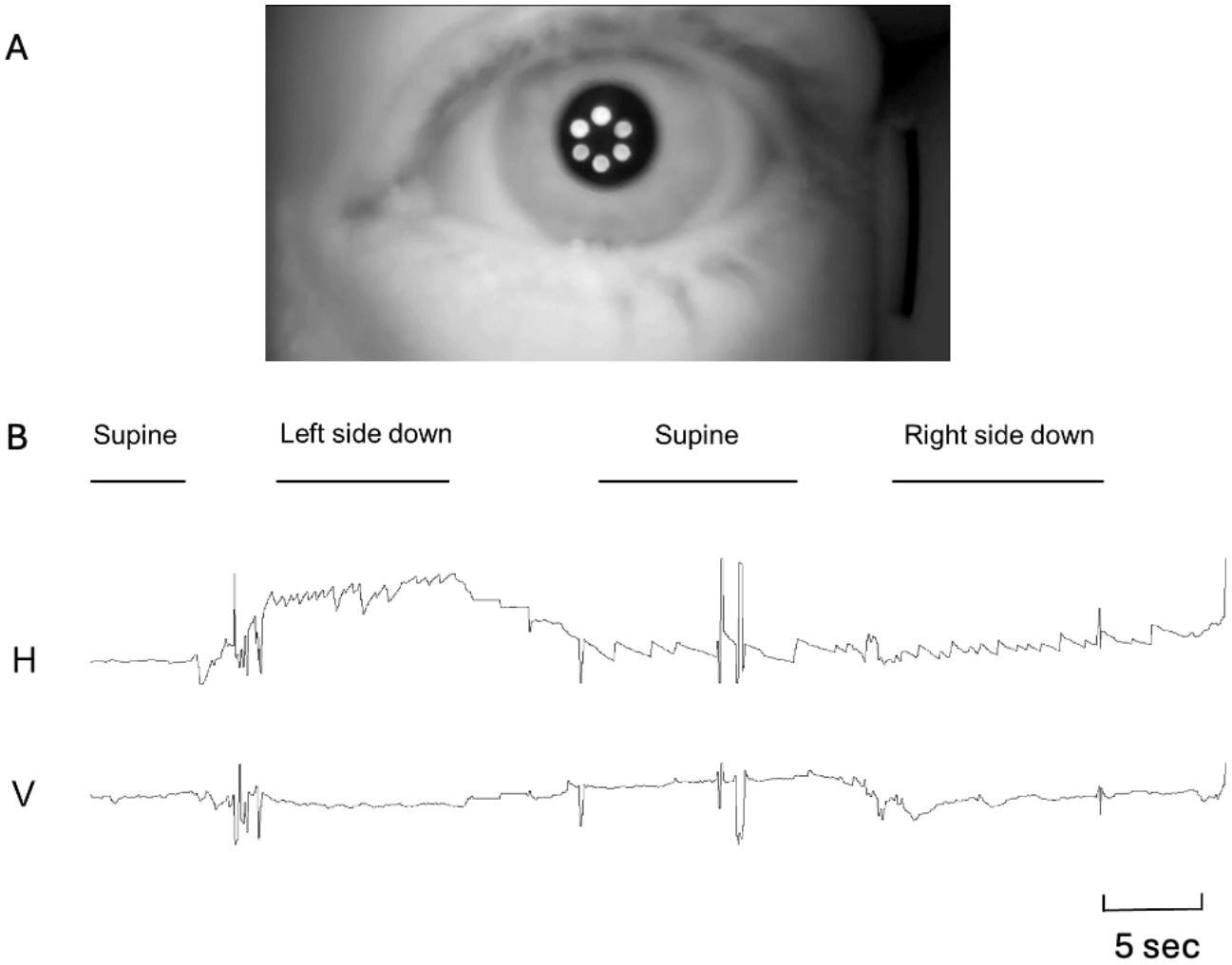

Discussion

Principal Findings

Episodic vertigo attacks often pose diagnostic challenges due to their transient nature and the absence of symptoms during clinical examinations. In this study, we developed a cost-effective diagnostic tool that combines a mini-infrared camera with 3D-printed goggles, enabling patients to record eye movements during vertigo attacks in dark conditions at home. Using this system, we successfully diagnosed lateral semicircular canal-type BPPV in a patient with recurrent vertigo, which had previously gone undiagnosed in standard outpatient evaluations. This approach demonstrates the feasibility and utility of our device in providing accurate diagnostic support for patients experiencing episodic vertigo. The system is also effective for patients whose vertigo symptoms resolve before clinical visits, allowing for the capture of critical diagnostic information at home during active episodes.

The goggles were printed with an adhesive filament to secure the mini-infrared camera and maintain their shape despite their complex structure. However, each printing required approximately 10 hours. Optimizing the design of the goggles' interior and improving printer performance might shorten production time. Alternatively, eye movements during vertigo attacks can be captured via video recording using a smartphone. Using this method, Kıroğlu and Dağkıran [16] diagnosed Ménière's disease in patients with vertigo attacks. For vertigo diagnosis, confirming spontaneous nystagmus under nongazing conditions is crucial [19]. However, when recording with a smartphone, nystagmus may be suppressed in bright conditions, as people tend to gaze at ambient light or the lens of the mini-infrared camera, making it difficult to capture eye movements in the dark. Melliti et al [20] attempted to record nystagmus using a smartphone lens adapter but reported that only approximately half (49%) of the patients successfully recorded nystagmus during the study period. Additionally, the study highlighted that significant time and effort were required to adequately explain the device's usage to patients. In contrast, the device used in this study allows for recording in dark conditions, overcoming the limitations of ambient light and enabling more reliable recordings for a wider range of patients. This feature holds promise for improving diagnostic outcomes by providing clear recordings of nystagmus during vertigo attacks. While Phillips et al [14] introduced an excellent system for continuous monitoring to investigate the pathophysiology

XSL•FO
RenderX

of Ménière's disease and vestibular migraine, practical clinical settings often require affordable and accessible tools, such as our system, to differentiate between BPPV and spontaneous vertigo.

In addition to its affordability compared to conventional diagnostic tools, the system demonstrated in this study offers significant benefits in terms of accessibility and usability. The mini-infrared camera used in this device costs approximately US $25, while the 3D-printed goggles can be produced for US $13 per unit. These low costs make the device not only economical for health care systems but also user-friendly for patients, allowing them to independently record diagnostic findings during vertigo attacks. This accessibility enables the capture of observations that might not have been possible with traditional equipment, particularly for patients who face challenges in attending repeated clinical visits. Such findings are crucial for early diagnosis and timely treatment, improving patient outcomes and reducing the burden on health care systems. Similarly, studies on comparable systems, such as the CAVA system, have reported cost savings of between £631 (US $791) and £1305 (US $1635.90) per patient by eliminating unnecessary tests and streamlining the diagnostic process [21]. This highlights the potential of community-based, patient-centered diagnostic tools to transform vertigo care, making it more efficient and accessible for all.

Previous studies used a deep learning system [22] and a smartphone application [23] to detect nystagmus. Several innovative applications for the diagnosis and management of vertigo have been developed [23], and combining these applications with other devices has the potential to enhance diagnostic accuracy and usability. Advances in these technologies and their integration with other devices will facilitate the assessment of smartphone-captured eye movements at home and improve the detection of nystagmus.

Future developments could further optimize the system for broader patient populations, such as older individuals or those with manual dexterity challenges, by integrating automated features and caregiver support options. These enhancements would not only expand the usability of the device but also strengthen its role in community-based and remote medical care, making it a versatile tool for vertigo diagnostics. As a potential adaptation to telemedicine, remote specialists have used eye movements captured on a smartphone to evaluate BPPV. Compared to infrared cameras, smartphones are more portable and less expensive. Commercially available infrared cameras with Wi-Fi and 3D-printed goggles can be used for web-based medical care by dizziness specialists and physicians in emergency medicine.

Limitations

Despite its promising utility, this study has several limitations. First, the findings are based on a single case, which limits the generalizability of the results. Second, the low resolution of the mini-infrared camera made it difficult to analyze the torsional components of nystagmus due to unclear iris patterns. However, the direction and intensity of nystagmus were easily determined, making the device a practical diagnostic tool.

Third, the system requires patients to operate the device during vertigo attacks, which may pose challenges for older individuals or those with impaired dexterity. Additionally, it relies on the user's ability to keep their eyes open and to ensure that the device is adequately charged and ready for use during an episode of vertigo. These operational requirements may limit its applicability for certain patient populations, particularly those who experience severe vertigo or lack manual dexterity.

Conclusion

In this study, we demonstrate the feasibility of using a low-cost, commercially available mini-infrared camera combined with 3D-printed goggles to accurately diagnose vestibular disorders such as BPPV during vertigo attacks. This approach highlights the potential for integrating affordable and accessible technologies into telemedicine, providing a valuable tool for remote, real-time diagnostic support. By enabling patients to record eye movements at home, this system bridges gaps in traditional outpatient care and empowers both patients and physicians to achieve timely and precise diagnoses. Future advancements in device performance and usability could further expand its applications in diverse health care settings, particularly for older populations or those in resource-limited regions.

Acknowledgments

We thank Editage (www.editage.com) for English language editing. This work was supported by the JST (Japan Science and Technology Agency; grant JPMJPF2403).

Data Availability

The data that support the findings of this study are available from the corresponding author (MN) upon reasonable request.

Conflicts of Interest

None declared.

References

1. Murdin L, Schilder AGM. Epidemiology of balance symptoms and disorders in the community: a systematic review. Otol Neurotol 2015 Mar;36(3):387-392. [doi: 10.1097/MAO.0000000000000691] [Medline: 25548891]

XSL•FO
RenderX

2. Kim JS, Newman-Toker DE, Kerber KA, et al. Vascular vertigo and dizziness: diagnostic criteria. J Vestib Res 2022;32(3):205-222. [doi: 10.3233/VES-210169] [Medline: 35367974]
3. Strupp M, Brandt T. Diagnosis and treatment of vertigo and dizziness. Dtsch Arztebl Int 2008 Mar;105(10):173-180. [doi: 10.3238/arztebl.2008.0173] [Medline: 19629221]
4. Mendel B, Bergenius J, Langius A. Dizziness symptom severity and impact on daily living as perceived by patients suffering from peripheral vestibular disorder. Clin Otolaryngol Allied Sci 1999 Aug;24(4):286-293. [doi: 10.1046/j.1365-2273.1999.00261.x] [Medline: 10472462]
5. Haleem A, Javaid M, Singh RP, Suman R. Telemedicine for healthcare: capabilities, features, barriers, and applications. Sens Int 2021;2:100117. [doi: 10.1016/j.sintl.2021.100117] [Medline: 34806053]
6. Ganjali R, Jajroudi M, Kheirdoust A, Darroudi A, Alnattah A. Telemedicine solutions for clinical care delivery during COVID-19 pandemic: a scoping review. Front Public Health 2022;10:937207. [doi: 10.3389/fpubh.2022.937207] [Medline: 35937265]
7. Noda M, Kuroda T, Nomura A, Ito M, Yoshizaki T, Fushiki H. Smartphone-assisted medical care for vestibular dysfunction as a telehealth strategy for digital therapy beyond COVID-19: scoping review. JMIR Mhealth Uhealth 2023 Sep 11;11:e48638. [doi: 10.2196/48638] [Medline: 37695671]
8. Hoppes CW, Lambert KH, Whitney SL, Erbele ID, Esquivel CR, Yuan TT. Leveraging technology for vestibular assessment and rehabilitation in the operational environment: a scoping review. Bioengineering (Basel) 2024 Jan 25;11(2):117. [doi: 10.3390/bioengineering11020117] [Medline: 38391603]
9. Barahim Bastani P, Saber Tehrani AS, Badihian S, et al. Self-recording of eye movements in amyotrophic lateral sclerosis patients using a smartphone eye-tracking app. Digit Biomark 2024;8(1):111-119. [doi: 10.1159/000538992] [Medline: 39015513]
10. Parker TM, Badihian S, Hassoon A, et al. Eye and head movement recordings using smartphones for telemedicine applications: measurements of accuracy and precision. Front Neurol 2022;13:789581. [doi: 10.3389/fneur.2022.789581] [Medline: 35370913]
11. Shah MU, Lotterman S, Roberts D, Eisen M. Smartphone telemedical emergency department consults for screening of nonacute dizziness. Laryngoscope 2019 Feb;129(2):466-469. [doi: 10.1002/lary.27424] [Medline: 30478898]
12. Kuroda T, Kuroda K, Fushiki H. Development of a prototype video head impulse test system using an iPhone for screening of peripheral vestibular dysfunction. Digit Biomark 2023;7(1):150-156. [doi: 10.1159/000534543] [Medline: 37928503]
13. Sung KB, Lee TK, Furman JM. Abnormal eye movements in dizzy patients. Neurol Clin 2005 Aug;23(3):675-703. [doi: 10.1016/j.ncl.2005.01.006] [Medline: 16026671]
14. Phillips JS, Newman JL, Cox SJ. Clinical techniques and technology: vestibular telemetry. Otolaryngol Head Neck Surg 2021 Nov;165(5):751-753. [doi: 10.1177/0194599821993411] [Medline: 33650899]
15. Young AS, Lechner C, Bradshaw AP, et al. Capturing acute vertigo: a vestibular event monitor. Neurology (ECronicon) 2019 Jun 11;92(24):e2743-e2753. [doi: 10.1212/WNL.0000000000007644] [Medline: 31092626]
16. Kıroğlu M, Dağkıran M. The role of mobile phone camera recordings in the diagnosis of Meniere's disease and pathophysiological implications. J Int Adv Otol 2020 Apr;16(1):18-23. [doi: 10.5152/iao.2019.6605] [Medline: 32066548]
17. Phillips JS, Newman JL, Cox SJ, FitzGerald J. Nystagmus during an acute Ménière's attack: from prodrome to recovery. Int J Audiol 2021 Jan;60(1):70-74. [doi: 10.1080/14992027.2020.1799252] [Medline: 32731785]
18. Handmade goggles designed for recording eye movements. OSF. URL: https://osf.io/rvxw7/ [accessed 2025-02-19]
19. Halmagyi GM, McGarvie LA, Strupp M. Nystagmus goggles: how to use them, what you find and what it means. Pract Neurol 2020 Dec;20(6):446-450. [doi: 10.1136/practneurol-2020-002513] [Medline: 33115786]
20. Melliti A, van de Berg M, van de Berg R. Capturing nystagmus during vertigo attacks using a smartphone: adherence, characteristics, pearls and pitfalls. J Neurol 2023 Dec;270(12):6044-6056. [doi: 10.1007/s00415-023-11965-y] [Medline: 37653139]
21. Phillips JS, Newman J. Quantifying the direct cost benefits of vestibular telemetry using the CAVA system to diagnose the causes of dizziness. Cost Eff Resour Alloc 2023 Jan 13;21(1):2. [doi: 10.1186/s12962-022-00413-9] [Medline: 36639653]
22. Wagle N, Morkos J, Liu J, et al. aEYE: a deep learning system for video nystagmus detection. Front Neurol 2022;13:963968. [doi: 10.3389/fneur.2022.963968] [Medline: 36034311]
23. van Bonn SM, Behrendt SP, Pawar BL, Schraven SP, Mlynski R, Schuldt T. Smartphone-based nystagmus diagnostics: development of an innovative app for the targeted detection of vertigo. Eur Arch Otorhinolaryngol 2022 Dec;279(12):5565-5571. [doi: 10.1007/s00405-022-07385-9] [Medline: 35451613]

Abbreviations

BPPV: benign paroxysmal positional vertigo
CARE: Case Report
CAVA: continuous ambulatory vestibular assessment
EQUATOR: Enhancing the Quality and Transparency of Health Research

XSL•FO
RenderX

Edited by A Mavragani; submitted 12.12.24; peer-reviewed by JS Phillips, T Miwa; revised version received 31.01.25; accepted 03.02.25; published 27.02.25.

Please cite as:
Sakazaki H, Noda M, Dobashi Y, Kuroda T, Tsunoda R, Fushiki H
Monitoring Nystagmus in a Patient With Vertigo Using a Commercial Mini-Infrared Camera and 3D Printer: Cost-Effectiveness Evaluation and Case Report
JMIR Form Res 2025;9:e70015
URL: https://formative.jmir.org/2025/1/e70015
doi:10.2196/70015

© Hiroyuki Sakazaki, Masao Noda, Yumi Dobashi, Tatsuaki Kuroda, Reiko Tsunoda, Hiroaki Fushiki. Originally published in JMIR Formative Research (https://formative.jmir.org), 27.2.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Case Report

Concurrent Treatment of Opioid and Tobacco Use Disorder in a Telemedicine Clinic: Case Report of Breaking Through Barriers

Darcy Michero[1], BS; Laura Monico[1], PhD; Peyton Pielsticker[1], LADC, LMSW; Larissa J Mooney[2], MD; Suzette Glasner[1,2], PhD

[1]Department of Clinical Affairs, Digital Therapeutics, New York, NY, United States

[2]Department of Psychiatry, University of California, Los Angeles, CA, United States

Corresponding Author:
Darcy Michero, BS
Department of Clinical Affairs
Digital Therapeutics
276 Fifth Avenue
Suite 608
New York, NY, 10001
United States
Phone: 1 (518) 545 5428
Email: darcy@pelagohealth.com

Abstract

There are a myriad of obstacles faced by individuals who take medication for opioid use disorder, including high out-of-pocket costs, stigma, logistical challenges associated with traveling to treatment facilities, and negative experiences with treatment providers. The proliferation of illicit fentanyl use and associated overdose risk underscores the urgent need to broaden access to lifesaving pharmacotherapy for opioid use disorder. This case report describes the safe and effective induction of a 32-year-old male onto sublingual buprenorphine in a digital care setting, with integrated, counselor-facilitated psychosocial treatment delivered via a smartphone app targeting relapse prevention. The digital care modality enabled this patient to overcome 2 major treatment barriers: limited access to evidence-based treatment given his residence in a rural area, and heightened overdose risk given his use of fentanyl, a potent synthetic opioid. In addition, we describe tobacco cessation treatment delivered digitally concurrently with opioid use disorder care, combining pharmacotherapy and behavioral skills training. Given the disproportionately high rate of co-occurring opioid use disorder and tobacco use and associated mortality rates, the effective treatment of these conditions simultaneously in this case advances our knowledge of potentially efficacious treatment approaches for underserved individuals with opioid use disorder who present with polysubstance use.

International Registered Report Identifier (IRRID): RR2-10.1177/20552076241258400

(JMIR Form Res 2025;9:e72872) doi:10.2196/72872

KEYWORDS

opioid use disorder; tobacco use disorder; buprenorphine; nicotine replacement therapy; case report

Introduction

Buprenorphine, a medication used to treat opioid use disorder (OUD), reduces relapse and overdose risk and improves long-term treatment outcomes in individuals with opioid addiction. Studies have established the safety and efficacy of buprenorphine induction in both office-based and unobserved (ie, home) settings. Despite the wealth of evidence supporting their efficacy, only half or fewer of individuals diagnosed with OUD receive buprenorphine, and treatment dropout rates remain high, with an average treatment duration of less than 6 months [1]. Historically, treatment programs in rural areas were significantly less likely to offer buprenorphine than in urban areas because many practitioners lacked a Drug Enforcement Agency waiver to prescribe buprenorphine [2]. However, in efforts to boost rural accessibility to medications for opioid use disorder (MOUD), the United States Department of Health and Human Services updated its guidelines in 2021, which granted physicians expanded prescribing capabilities and reduced certification requirements [3]. Initially, though the elimination of the waiver requirement was expected to increase MOUD access, recent studies have highlighted that despite increases in buprenorphine prescribers, the number of patients who receive buprenorphine has not proportionally changed [4]. Furthermore,

XSL•FO
RenderX

patients in rural areas may face unique challenges to buprenorphine access due to travel distance from specialists [5], and rural providers have cited concerns about misuse and diversion as barriers to prescribing buprenorphine [6]. Recent research indicates that a lack of adequate training resources for prescribers hinders the successful implementation of greater buprenorphine prescribing capacity [7,8]. As such, an urgent need remains to improve access to MOUD in nonmetropolitan areas.

Digital treatment modalities have been highlighted as a potential solution to overcome geographical barriers to accessing traditional care, including attending in-person visits, picking up medication from the pharmacy, or arranging transportation [9,10]. Recent evidence suggests that increasing access to buprenorphine via telemedicine facilitates treatment engagement and reduces overdoses [11]. Although some providers have fears of diversion or risk associated with an unobserved, home-based induction [12], both unobserved and observed buprenorphine inductions have been found to be safe and effective [13,14]. In light of the central role of potent synthetic opioids such as fentanyl in the opioid overdose crisis, telemedicine approaches have the potential to be transformative among individuals who are using fentanyl if implemented effectively. However, there is little guidance in the literature concerning telehealth-based buprenorphine induction for individuals who are actively using fentanyl and its analogs, for whom concerns about precipitated withdrawal, relapse, and overdose risk may pose barriers [15,16].

Polysubstance use is another critical public health problem among individuals with OUD, for whom nicotine co-use is highly prevalent, reported among over 80%-90% of the population [17]. In light of the disproportionately high mortality rate due to smoking-related illness among those with substance use disorders, coupled with findings that individuals with co-occurring OUD and tobacco use disorder face unique obstacles to smoking cessation, including limited access to treatment, low motivation to quit, and concerns about how cessation could impact their ability to manage opioid withdrawal symptoms (eg, negative affect) without relapsing [18], novel and scalable approaches to addressing nicotine use in this population are urgently needed.

Although evidence-based pharmacotherapies such as nicotine replacement therapy (NRT) increase cessation rates among those with OUD, research to inform digital care models addressing opioid and nicotine use concurrently is lacking [19]. This case report discusses the clinical characteristics, course, and outcomes of an individual with co-occurring OUD and tobacco use disorder who participated in a clinical trial evaluating the feasibility and preliminary efficacy of Pelago Opioid (Pel-O), developed by Digital Therapeutics, [20], a commercially available digital outpatient treatment program that combines MOUD, NRT, and evidence-based behavioral treatment for substance use disorders (SUDs; ie, cognitive behavioral therapy (CBT) and motivational enhancement therapy [MET]) in a digital care modality.

Methods

Overview

The patient was a 32-year-old Caucasian male with no significant past medical or surgical history. The patient was unemployed and lived with his significant other in a rural area in the southern Appalachian region. He had a family history of opioid addiction, indicating that he “watched (his) mom and dad use opioids” during his teenage years. The patient reported a history of an unspecified anxiety disorder, initiation of smoking a pack of cigarettes per day since he was 16 years old, and intermittent use of methamphetamine approximately once per month beginning at the age of 19 years. He reported initiating opioid use at 12 years of age when he experimented with recreational use of prescription opioids including Percocet and hydrocodone. At 14 years of age, he reported initiating the use of oxycodone 40 mg tablets regularly, and by the age of 16 years, he began to experience withdrawal symptoms in between episodes of opioid use. The patient continued to use illicit prescription opioids until the age of 28 years when he transitioned to using heroin and fentanyl exclusively via oral and intranasal routes of administration, a pattern he continued until 2 weeks before treatment initiation.

The patient’s treatment history includes 2 previous pharmacotherapy treatment episodes for OUD: (1) at 19 years of age, he obtained methadone from an opioid treatment program for 1 year, but discontinued treatment due to accessibility barriers including the cost associated with driving 200 miles per day to and from the clinic, as well as losing access to a driver’s license; and (2) at 26 years of age, buprenorphine was prescribed by an outpatient provider for 1 year but was also discontinued due to the same obstacles that impeded his first treatment episode (ie, transportation and cost).

Two weeks before enrolling in the Pel-O clinical trial, the patient’s opioid supplier was incarcerated, leading him to lose access to fentanyl and heroin. To avoid entering withdrawal, the patient purchased nonprescribed buprenorphine and reported illicit sublingual use of buprenorphine every other day, with 14 days of abstinence reported in the 30 days before treatment entry.

Pel-O is a digital SUD treatment approach with 2 essential components: (1) telemedicine-delivered medication management with a licensed nurse practitioner, doctor of medicine, or doctor of osteopathic medicine with expertise in addiction medicine for medication management, and behavioral therapy visits targeting relapse prevention with a licensed drug and alcohol counselor, using synchronous video visits via a smartphone app and (2) digital psychoeducation and evidence-based therapy skills training (eg, CBT and MET), delivered via the Pel-O and supplemented with counseling sessions (refer to Monico et al [20] for a detailed description of Pel-O).

Ethical Considerations

The individual described in this case study was part of a larger clinical trial for which informed consent was provided for the provision of clinical care, as well as quantitative and qualitative interview-based data collection. The original informed consent

XSL•FO
RenderX

as part of participation in the primary study allows for secondary analysis without additional consent. Oversight for this study was provided by the Ethical and Independent Review Services institutional review board. All study data are deidentified. Data for this case report were based both on quantitative data gathered over the course of the patient's participation in the larger trial for which he was able to gain US $195 in gift cards as compensation, as well as a qualitative interview for which he was compensated with a US $50 gift card. Written informed consent was obtained from the patient for the use and publication of his clinical data for research following a detailed explanation of the objectives and protocol of the study.

At the patient's first telehealth physician visit, he was assessed using a semistructured diagnostic interview based on the American Society of Addiction Medicine's criteria for placement and treatment planning [21] and met *DSM-5* (*Diagnostic and Statistical Manual of Mental Disorders* [Fifth Edition]) criteria for severe OUD. He was determined to have a high risk for opioid relapse, necessitating outpatient care combining MOUD with behavioral skills training for relapse prevention. The patient subsequently attended eleven 30-minute video-based medical management appointments over a 5-month duration with his Pel-O physician; 8 visits over the first 3 months and 3 visits over months 4 and 5. Medical management visits included discussion of dose adjustments, side effects, and limited counseling as routinely provided to patients in outpatient settings. The patient's initial prescription was issued to his local pharmacy to facilitate immediate buprenorphine initiation, whereas subsequent prescriptions were mailed directly to the patient's residence according to his preferences, for convenience. Since the patient reported current use of nonprescribed buprenorphine at the time of treatment initiation, he was first prescribed sublingual buprenorphine with naloxone in the film formulation (hereafter, buprenorphine) at a dosage of 16 mg daily, with dosing evenly split twice daily. At the first and second follow-up visits, the patient reported continued cravings for opioids; accordingly, the prescribing physician adjusted his dosage to 20 mg, and subsequently to 24 mg, respectively, split evenly twice daily. The only notable side effect of buprenorphine reported by the patient was mild constipation that improved over time and with the aid of increased hydration and an over-the-counter stool softener. According to his report, the patient was prescribed hydroxyzine, quetiapine, and mirtazapine, each on an as-needed basis for sleep disturbance by an outside psychiatrist. At his final medication management follow-up in Pel-O, the patient reported that he was currently taking only mirtazapine regularly for sleep, though he did not specify the dosage.

The patient attended thirteen 45-minute video-based visits with a licensed drug and alcohol counselor over the course of 5 months of his Pel-O treatment participation. The counselor delivered manualized CBT and MET, with between-session asynchronous messaging through the chat function in the Pel-O smartphone app. Over the 6-month duration, the counselor sent 133 chat messages, and the patient sent 116 chat messages, demonstrating a high level of engagement in care. The patient also learned coping skills and exercises to help identify triggers pertaining to his cravings for both opioids and tobacco during these sessions.

Approximately 6 weeks after Pel-O treatment initiation, the patient expressed a desire to quit using tobacco, and after discussing nicotine replacement pharmacotherapy options, his Pel-O physician provider prescribed nicotine nasal spray. However, due to nasal mucosa irritation, the patient discontinued the use of the nasal spray and transitioned to alternative forms of NRT (ie, nicotine gum and patches). The Pel-O program provided 4 mg nicotine gum for as-needed use, as well as 21 mg and 14 mg nicotine replacement patches. After 4 weeks at 21 mg, the patient transitioned to the use of the 14 mg patches every 24 hours and subsequently titrated down to 7 mg per 24 hours. As instructed, the patient wore the patches over the course of 16 weeks. Concurrently, his CBT and MET counseling sessions incorporated his treatment objectives for smoking cessation to enable the use of cognitive and behavioral coping skills to prevent relapse to tobacco use.

After stabilization on buprenorphine, a total of 3 self-administered urine drug screens (UDS) were gathered at a monthly cadence. The patient was mailed a US Food and Drug Administration–approved 1-step test for the UDS and was asked to gather the sample privately while the clinician waited digitally for the patient to return to the camera to display the results. The UDS screened for the presence of opiates (heroin, morphine, and prescription drugs), fentanyl, amphetamines, benzodiazepines, methadone, cocaine, methamphetamine, and buprenorphine with a threshold >10 ng/mL considered indicative of medication adherence.

Results

The patient's first UDS gathered 8 weeks postinduction, was positive for opiates and negative for fentanyl. Other substance use detected included amphetamines, opiates, tetrahydrocannabinol (THC), and synthetic marijuana. Concurrently, the test result was positive for buprenorphine, indicating treatment adherence to the MOUD regimen. The second UDS, performed 26 days after the first, was negative for opiates and fentanyl and positive for buprenorphine. In terms of other substance use, the second UDS was positive for THC only. The third UDS performed a month later, was negative for opiates and fentanyl positive for buprenorphine, indicating continued adherence to MOUD care. The test also detected THC, while negative for all other substances. The patient's depression, anxiety, craving for opioids, and cigarette use were measured at baseline, 4-, 8-, and 12- weeks post treatment entry. The patient evidenced a decrease in self-reported depression severity, measured with the Patient Health Questionnaire-9, from the moderate range at baseline (total score=12) to the mild range at week 12 (total score=8).

Likewise, the patient evidenced a decrease in self-reported anxiety severity, measured with the Generalized Anxiety Disorder-7 Scale, from the severe range at baseline (total score=15) to the mild range at week 12 (total score=5).

In addition to the UDS assay, adherence to buprenorphine was assessed via monthly telephone-based buprenorphine film counts

XSL•FO
RenderX

over the 12-week study period, adapted from an empirically validated unannounced pill count procedure [22]. Trained research staff called the patient once per month to conduct a film count. Study staff completed the first film count 1-month post buprenorphine stabilization as determined by the study clinician. Each subsequent count allowed for an adherence score, ranging from 0 to 1, to be calculated by the ratio of films remaining relative to films prescribed, taking into account the number of films dispensed [23].

Opioid use frequency in the past 30 days was assessed monthly using a timeline follow-up interview [24]. At baseline, the patient reported 16 days of illicit opioid use in the previous 30 days. At 1-month follow-up, following his induction onto buprenorphine, he reported 0 days of opioid use in the previous 30 days, with 98% adherence to medication based on the unannounced film counts. At week 8, the patient reached stabilization on buprenorphine and reported 0 days of opioid use in the previous 30 days; however, this was inconsistent with his UDS, which was positive for opiates. Nevertheless, his unannounced film count revealed 80% medication adherence, which is clinically consistent with a temporary return to opioid use. At week 12, the patient's opioid- and fentanyl-negative UDS was consistent with his self-reported abstinence from opioid use over the past 30 days, coupled with his film count which demonstrated 100% MOUD adherence. In addition, the patient's craving for opioids, measured using a Visual Analog Scale (with a subjective craving intensity score ranging from 0 to 100), decreased by 98% between baseline and week 12.

The patient's self-reported cigarette use was also captured during the timeline follow-up interview every 4 weeks, which showed a 75% reduction from 20 cigarettes per day at baseline, to 5 cigarettes per day at week 12. The patient was stable in his recovery from opioid and tobacco use disorder at the time of discharge from Pel-O. His discharge was prompted by the conclusion of the clinical trial intervention period, at which time he was referred, in accordance with his preferences, to another digital OUD care provider to continue his MOUD treatment regimen.

Discussion

Principal Findings

Despite living in a rural area where access to evidence-based treatment for OUD was scarce, the patient was able to receive care for co-occurring OUD and tobacco use disorder via a digital care modality. This promising treatment strategy enabled the patient to overcome barriers including transportation costs, financial burden, and stigma associated with receiving care and participating in UDS procedures. Although research to date has demonstrated the feasibility of digital treatment for OUD [10,20,25,26], evidence-based clinical guidelines to inform the practice of digital buprenorphine induction for individuals using highly potent synthetic opioids such as fentanyl are lacking [13,14]. As such, evidence from this case demonstrates the feasibility, safety, and potential efficacy of this highly scalable approach to making MOUD care accessible to individuals using fentanyl who would otherwise face numerous obstacles to receiving care.

Of note, consistent with research demonstrating that nonprescribed buprenorphine use is frequently observed among individuals who are attempting to taper themselves off opioids [27], the patient's buprenorphine induction was preceded by nonprescribed buprenorphine use over a 2-week period, which likely contributed, in part, to his successful induction in the digital clinic setting. In addition, despite the challenges inherent to engaging and motivating individuals with OUD and co-occurring tobacco use disorder, the use of combination NRT coupled with behavioral treatment was effective in facilitating tobacco cessation simultaneously with recovery from OUD, which adds to the existing literature underscoring the need for concurrent treatment [28].

The Patient's Perspective

The major themes that emerged from the patient's qualitative interview as well as interviews with other participants in the larger study are summarized in detail in the study by Monico et al (2025) [29]. The patient interviewed for this case report expressed that he felt cared for by his Pel-O physician and counselor:

> *[Prescriber] is a really, really good doctor. He really cares. I mean, he takes the time to sit down with you and ask you how you are.*

The patient was most motivated by the digital doctor's visits and counseling aspect of the Pel-O program, highlighting the contrast between his experience at Pel-O when compared with the difficulty of commuting at inconvenient times to a methadone clinic:

> *Well, I went to the methadone clinic about probably about eight years ago. And, you know, it kind of worked. But the way that they make you do it, you know, you have to drive 60 miles there every day, and pick it up at 6:00 in the morning and then drive 60 miles back every day. It just made it too hard.*

The patient also expressed gratitude for the flexibility and convenience afforded by the digital care modality:

> *I actually loved it. I loved how convenient it was. You know, it made it where it was easier to talk to her, you know, to talk to my counselor because of how convenient it was. And it made me want to open up more, to [prescriber], to about, you know, and because it is so convenient. But I never thought I would be that honest with a counselor or a doctor, you know?*

Limitations

Several limitations of this report warrant comment, particularly in regard to the time-limited treatment duration imposed by the clinical trial context of this patient's care. Specifically, the patient continued to use cannabis throughout his treatment, and the impact of this use on his short- and longer-term recovery from OUD warrants further examination in light of research examining the association between OUD treatment outcomes and cannabis use among individuals receiving MOUD [30]. Second, given his transition out of Pel-O approximately 2 months after reducing his use of tobacco, the stability of changes

XSL•FO
RenderX

in his tobacco use is unknown. Future research elucidating the longer-term outcomes of individuals who concurrently discontinue or reduce their use of opioids and tobacco will have great use in identifying gaps in care where further innovation is needed to address polysubstance use.

Conclusion

In conclusion, the case presented herein expands the growing body of literature supporting the safety and efficacy of MOUD delivered via telemedicine, extending the potential indication of this treatment modality to address polysubstance use and populations who present with recent use of potent synthetic opioids including fentanyl. The success of this case highlights the need for greater resource allocation to telehealth modalities for treating severe OUD, combining psychotherapy and MOUD to alleviate barriers associated with traditional, face-to-face treatment.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to the sensitive and highly confidential nature of substance use disorder data under 42 Code of Federal Regulations Part 2 but are available from the corresponding author on reasonable request subject to approval by the appropriate institutional review board and compliance with applicable redisclosure restrictions.

Conflicts of Interest

SG and DM are employed by and hold stock options from Digital Therapeutics. LM and PP were employed by and held stock options from Digital Therapeutics at the time of contribution. LJM has received research support from Aelis Farma.

References

1. Shulman M, Wai JM, Nunes EV. Buprenorphine treatment for opioid use disorder: an overview. CNS Drugs 2019 Jun;33(6):567-580 [FREE Full text] [doi: 10.1007/s40263-019-00637-z] [Medline: 31062259]
2. Jones KF, O'Reilly Jacob M, Spetz J, Hailer L, Tierney M. Eliminate the buprenorphine DEA X waiver: Justification using a policy analysis approach. J Nurs Scholarsh 2023 May;55(3):655-664 [FREE Full text] [doi: 10.1111/jnu.12871] [Medline: 36624606]
3. Jones CM, Olsen Y, Ali MM, Sherry TB, Mcaninch J, Creedon T, et al. Characteristics and prescribing patterns of clinicians waivered to prescribe buprenorphine for opioid use disorder before and after release of new practice guidelines. JAMA Health Forum 2023 Jul 07;4(7):e231982 [FREE Full text] [doi: 10.1001/jamahealthforum.2023.1982] [Medline: 37477926]
4. Chua K, Bicket MC, Bohnert ASB, Conti RM, Lagisetty P, Nguyen TD. Buprenorphine dispensing after elimination of the waiver requirement. N Engl J Med 2024 Apr 25;390(16):1530-1532. [doi: 10.1056/NEJMc2312906] [Medline: 38657250]
5. Lofwall MR, Fanucchi LC. Long-acting buprenorphine injectables: Opportunity to improve opioid use disorder treatment among rural populations. Prev Med 2021 Nov;152(Pt 2):106756. [doi: 10.1016/j.ypmed.2021.106756] [Medline: 34352306]
6. Andrilla CHA, Moore TE, Patterson DG. Overcoming barriers to prescribing buprenorphine for the treatment of opioid use disorder: recommendations from rural physicians. J Rural Health 2019 Jan;35(1):113-121. [doi: 10.1111/jrh.12328] [Medline: 30339720]
7. Franz B, Dhanani LY, Hall OT, Brook DL, Simon JE, Miller WC. Differences in buprenorphine prescribing readiness among primary care professionals with and without X-waiver training in the US. Harm Reduct J 2023 Dec 21;20(1):180 [FREE Full text] [doi: 10.1186/s12954-023-00918-3] [Medline: 38129903]
8. Fenstemaker C, Abrams EA, Obringer B, King K, Dhanani LY, Franz B. Primary care professionals' perspectives on tailoring buprenorphine training for rural practice. J Rural Health 2024 Sep;40(4):671-680. [doi: 10.1111/jrh.12832] [Medline: 38486066]
9. Weintraub E, Seneviratne C, Anane J, Coble K, Magidson J, Kattakuzhy S, et al. Mobile telemedicine for buprenorphine treatment in rural populations with opioid use disorder. JAMA Netw Open 2021 Aug 02;4(8):e2118487 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.18487] [Medline: 34448869]
10. Maricich YA, Xiong X, Gerwien R, Kuo A, Velez F, Imbert B, et al. Real-world evidence for a prescription digital therapeutic to treat opioid use disorder. Curr Med Res Opin 2021 Feb;37(2):175-183 [FREE Full text] [doi: 10.1080/03007995.2020.1846023] [Medline: 33140981]
11. Jones CM, Shoff C, Hodges K, Blanco C, Losby JL, Ling SM, et al. Receipt of telehealth services, receipt and retention of medications for opioid use disorder, and medically treated overdose among Medicare beneficiaries before and during the COVID-19 pandemic. JAMA Psychiatry 2022 Oct 01;79(10):981-992 [FREE Full text] [doi: 10.1001/jamapsychiatry.2022.2284] [Medline: 36044198]
12. Srivastava A, Kahan M, Leece P, McAndrew A. Buprenorphine unobserved "home" induction: a survey of Ontario's addiction physicians. Addict Sci Clin Pract 2019 May 01;14(1):18 [FREE Full text] [doi: 10.1186/s13722-019-0146-4] [Medline: 31039821]

XSL•FO
RenderX

13. Samuels EA, Khatri UG, Snyder H, Wightman RS, Tofighi B, Krawczyk N. Buprenorphine telehealth treatment initiation and follow-up during COVID-19. J Gen Intern Med 2022 Apr;37(5):1331-1333 [FREE Full text] [doi: 10.1007/s11606-021-07249-8] [Medline: 34981357]
14. Gunderson EW, Wang X, Fiellin DA, Bryan B, Levin FR. Unobserved versus observed office buprenorphine/naloxone induction: a pilot randomized clinical trial. Addict Behav 2010 May;35(5):537-540 [FREE Full text] [doi: 10.1016/j.addbeh.2010.01.001] [Medline: 20106601]
15. Silverstein SM, Daniulaityte R, Martins SS, Miller SC, Carlson RG. "Everything is not right anymore": Buprenorphine experiences in an era of illicit fentanyl. Int J Drug Policy 2019 Dec;74:76-83 [FREE Full text] [doi: 10.1016/j.drugpo.2019.09.003] [Medline: 31563098]
16. Antoine D, Huhn AS, Strain EC, Turner G, Jardot J, Hammond AS, et al. Method for successfully inducting individuals who use illicit fentanyl onto buprenorphine/naloxone. Am J Addict 2021 Jan;30(1):83-87 [FREE Full text] [doi: 10.1111/ajad.13069] [Medline: 32572978]
17. Custodio L, Malone S, Bardo MT, Turner JR. Nicotine and opioid co-dependence: Findings from bench research to clinical trials. Neurosci Biobehav Rev 2022 Mar;134:104507 [FREE Full text] [doi: 10.1016/j.neubiorev.2021.12.030] [Medline: 34968525]
18. Kathuria H, Seibert RG, Cobb V, Herbst N, Weinstein ZM, Gowarty M, et al. Perceived barriers to quitting cigarettes among hospitalized smokers with substance use disorders: A mixed methods study. Addict Behav 2019 Aug;95:41-48. [doi: 10.1016/j.addbeh.2019.02.017] [Medline: 30836208]
19. Vlad C, Arnsten JH, Nahvi S. Achieving smoking cessation among persons with opioid use disorder. CNS Drugs 2020 Apr;34(4):367-387 [FREE Full text] [doi: 10.1007/s40263-020-00701-z] [Medline: 32107731]
20. Monico LB, Eastlick M, Michero D, Pielsticker P, Glasner S. Feasibility and acceptability of a novel digital therapeutic combining behavioral and pharmacological treatment for opioid use disorder. Digit Health 2024;10:20552076241258400 [FREE Full text] [doi: 10.1177/20552076241258400] [Medline: 38812851]
21. Mee-Lee D, Shulman GD, Fishman MJ, Gastfriend DR, Miller MM, Provence SM. In: Chevy Chase MD, editor. The ASAM criteria: Treatment for addictive, substance-related, and co-occurring conditions. ASAM: American Society of Addiction Medicine; 2013.
22. Kalichman SC, Amaral CM, Stearns H, White D, Flanagan J, Pope H, et al. Adherence to antiretroviral therapy assessed by unannounced pill counts conducted by telephone. J Gen Intern Med 2007 Jul;22(7):1003-1006 [FREE Full text] [doi: 10.1007/s11606-007-0171-y] [Medline: 17390095]
23. Kalichman SC, Pellowski J, Kegler C, Cherry C, Kalichman MO. Medication adherence in people dually treated for HIV infection and mental health conditions: test of the medications beliefs framework. J Behav Med 2015 Aug;38(4):632-641 [FREE Full text] [doi: 10.1007/s10865-015-9633-6] [Medline: 25835435]
24. Hjorthøj CR, Hjorthøj AR, Nordentoft M. Validity of timeline follow-back for self-reported use of cannabis and other illicit substances--systematic review and meta-analysis. Addict Behav 2012 Mar;37(3):225-233. [doi: 10.1016/j.addbeh.2011.11.025] [Medline: 22143002]
25. Hailu R, Mehrotra A, Huskamp HA, Busch AB, Barnett ML. Telemedicine use and quality of opioid use disorder treatment in the US during the COVID-19 pandemic. JAMA Netw Open 2023 Jan 03;6(1):e2252381 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.52381] [Medline: 36692880]
26. Meyer B, Utter G, Hillman C. A personalized, interactive, cognitive behavioral therapy-based digital therapeutic (MODIA) for adjunctive treatment of opioid use disorder: development study. JMIR Ment Health 2021 Oct 08;8(10):e31173 [FREE Full text] [doi: 10.2196/31173] [Medline: 34623309]
27. Lofwall MR, Walsh SL. A review of buprenorphine diversion and misuse: the current evidence base and experiences from around the world. J Addict Med 2014;8(5):315-326 [FREE Full text] [doi: 10.1097/ADM.0000000000000045] [Medline: 25221984]
28. Morris CD, Garver-Apgar CE. Nicotine and opioids: a call for co-treatment as the standard of care. J Behav Health Serv Res 2020 Oct;47(4):601-613 [FREE Full text] [doi: 10.1007/s11414-020-09712-6] [Medline: 32495248]
29. Monico LB, Eastlick M, Michero D, Pielsticker P, Glasner S. Overcoming barriers to traditional care delivery and pharmacy challenges: a qualitative study of buprenorphine, telehealth, and a digital therapeutic for opioid use disorder. Subst Abuse Treat Prev Policy 2025 Feb 18;20(1):8 [FREE Full text] [doi: 10.1186/s13011-024-00631-9] [Medline: 39966974]
30. Lake S, St Pierre M. The relationship between cannabis use and patient outcomes in medication-based treatment of opioid use disorder: A systematic review. Clin Psychol Rev 2020 Dec;82:101939. [doi: 10.1016/j.cpr.2020.101939] [Medline: 33130527]

Abbreviations

CBT: cognitive behavioral therapy
***DSM-5*:** *Diagnostic and Statistical Manual of Mental Disorders (Fifth Edition)*
MET: motivational enhancement therapy
MOUD: medications for opioid use disorder

XSL•FO
RenderX

NRT: nicotine replacement therapy
OUD: opioid use disorder
Pel-O: Pelago Opioid
SUD: substance use disorder
THC: Tetrahydrocannabinol
UDS: urine drug screens

Edited by A Mavragani; submitted 19.02.25; peer-reviewed by EP Greenup; comments to author 20.03.25; revised version received 28.03.25; accepted 23.04.25; published 15.05.25.

Please cite as:
Michero D, Monico L, Pielsticker P, Mooney LJ, Glasner S
Concurrent Treatment of Opioid and Tobacco Use Disorder in a Telemedicine Clinic: Case Report of Breaking Through Barriers
JMIR Form Res 2025;9:e72872
URL: https://formative.jmir.org/2025/1/e72872
doi:10.2196/72872
PMID:40372777

©Darcy Michero, Laura Monico, Peyton Pielsticker, Larissa J Mooney, Suzette Glasner. Originally published in JMIR Formative Research (https://formative.jmir.org), 15.05.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Development and Health System Deployment of an Electronic Health Record–Integrated Chatbot Intervention for Connecting Fall Risk Screening to Community Resources After Emergency Department Visits: Implementation Study

Audrey Keleman[1,2], MSCI, PhD; Megan Bounds[3], MPH; Maxwell Lunt[4], BS; Jennifer Portz[5], MSW, PhD; Bucky Ferozan[6], DO; Jonathan Gomez Picazo[3], BS; Kelly Bookman[3], MD; Hillary D Lum[7], MD, PhD; Elizabeth M Goldberg[3], ScM, MD

[1]Eastern Colorado Geriatric Research, Education, and Clinical Center, United States Department of Veterans Affairs, Aurora, CO, United States
[2]Physical Therapy, University of Colorado Anschutz Medical Campus, Aurora, CO, United States
[3]Emergency Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO, United States
[4]Experience and Innovation, UCHealth, Aurora, CO, United States
[5]Internal Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO, United States
[6]Emergency Medicine, UCHealth Northern Colorado, Fort Collins, CO, United States
[7]Division of Geriatric Medicine, Department of Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO, United States

Corresponding Author:
Elizabeth M Goldberg, ScM, MD
Emergency Medicine
University of Colorado Anschutz Medical Campus
12401 E. 17th Ave
Campus Box B-215
Aurora, CO, 80045
United States
Phone: 1 720 848 6777
Email: elizabeth.goldberg@cuanschutz.edu

Abstract

Background: Emergency departments (EDs) routinely screen for fall risk, but patients are rarely notified of their results or referred to preventive resources. There is a critical need for an intervention that notifies patients when they are at risk for falls and automates referrals to fall prevention programs without increasing clinician workload. Chatbots can be used to provide patient education and community resources in a conversational, friendly manner. We developed and implemented an automated intervention using our health system's electronic health record (EHR) and an artificial intelligence chatbot, Livi, to address this gap in fall prevention across 17 EDs.

Objective: This study aimed to share how we developed our fall risk notification and referral intervention and iteratively improved it based on end-user feedback.

Methods: We collaborated with the EHR and ED operations teams to automate patient notification of fall risk and referral. First, we leveraged existing fall risk screening questions in nursing documentation to identify patients at risk for falls. We then developed an EHR workflow that delivers a QR code in the after-visit summary for all high-risk patients at ED discharge. Scanning the QR code launches a conversation with Livi, guiding users to physician-vetted, evidence-based, free or low-cost fall prevention resources in their area. In this workflow, only ED patients who are screened as high risk receive linkage to Livi, and clinicians do not need to manually place referrals or enter specific fall prevention resources at discharge. We conducted rapid, iterative usability testing of the Livi falls chatbot with 93 community members during the development process at 3 community fairs in distinct settings.

Results: Rapid iterative testing led to enhancements in the intervention, such as increased font size, an option for Spanish language, additional geographic locations for fall prevention resources, home modification resources, the ability to self-assess for fall risk, fall prevention tips, and the ability for patients to leave feedback on the Livi chatbot. Because all EDs in the health

XSL•FO
RenderX

system use the same instance of Epic, the EHR workflow was instantaneously deployed system-wide. The use of a QR code linked to the Livi chatbot also allows for the rapid updating of prevention resources.

Conclusions: This study describes the formative development and system-wide implementation of the intervention. This scalable, EHR-integrated intervention demonstrates a novel and pragmatic approach to improving population health by capitalizing on existing clinical workflows and automating both risk notification and personalized resource referral for older adults without increasing clinician burden. The next steps include conducting a randomized controlled trial to assess the impact of the screening and referral tool on recurrent fall-related health care use compared with routine care in the ED. Formal evaluation of the implementation outcomes will be conducted in the planned trial.

(JMIR Form Res 2025;9:e77237) doi:10.2196/77237

KEYWORDS

falls; older adults; digital health; referral pathway; high fall risk screening; fall prevention; emergency department

Introduction

Falls are the leading cause of injury-related emergency department (ED) visits among older adults [1]. Community-dwelling older adults who visit the ED after a fall have a 30% increased risk of falling again within 6 months compared to age-matched peers, and they often experience declines in functional abilities, such as balance, following discharge [2]. Evidence-based fall prevention programs can reduce fall risk by up to 30% [3,4], yet few older adults are successfully connected to these interventions after an ED visit. One persistent barrier is the absence of scalable digital health interventions to notify patients of their fall risk and refer them to appropriate, accessible programs.

Although ED nurses routinely screen older adults for fall risk, many patients are not informed of their screening results or provided with clear, actionable next steps [5,6]. Clinicians recognize the importance of fall prevention; however, acute medical needs typically take precedence during ED visits, limiting the time and attention available for preventive counseling [5,7]. Existing ED-based strategies have largely focused on automating risk identification using electronic health record (EHR) data [8] and manually referring high-risk ED patients to fall prevention clinics [9]. However, we are unaware of any intervention that can automate both patient notification of fall screening results and referral to fall prevention clinics. Furthermore, many ED patients lack access to local fall prevention clinics, and generic discharge instructions often lack the location-specific guidance that older adults need to access services close to home.

This paper describes the pre-evaluation development of an EHR-embedded intervention that automates fall risk notification and referral using a health system–integrated chatbot (ie, a virtual conversational agent). The intervention directs high-risk patients to free or low-cost, evidence-based fall prevention programs that are localized to their geographic area. Designed to scale across EDs and integrate easily into clinical workflows, this approach aims to reduce clinician burden, improve patient outcomes, and serve as a model for embedding scalable digital health interventions into routine emergency care.

Methods

Overview of Intervention Design

To create the intervention, we first developed a fall prevention resource library of evidence-based interventions available in the state. We then collaborated with our health system's Epic team, Livi chatbot team, and ED clinical teams to automate fall risk notification and referral to the Livi chatbot to provide location-specific fall prevention resources. Finally, we optimized the Livi chatbot intervention by incorporating feedback and experiences from end users. Each stage is described in detail in the subsequent sections.

Ethical Considerations

The Colorado Multiple Institutional Review Board approved this study and granted exempt status (COMIRB; 24-1835). The phases described in this paper did not require patient data or consent. Community members who participated in the usability testing did not provide any identifying information.

Fall Prevention Resources

Our team compiled a list of current fall prevention resources in the state by searching online, calling identified programs, and gathering the current resources that EDs in our system provide to patients. Resources included information about local evidence-based fall prevention classes (eg, Stepping On), and how to access help with home modifications (eg, Volunteers of America Handyperson Program). We consulted the National Council on Aging's approved list of evidence-based programs for structured class and program resources. We included only free or low-cost resources to ensure broad accessibility. Our health care system's legal and compliance departments vetted these resources to ensure that there were no Stark law [10] violations. This law specifies that patient referrals must be based on patient needs rather than clinician profit. Because our health system encompasses many regions in the state, we grouped resources by location, including the Denver Metro area, Colorado's North region (eg, Fort Collins and Steamboat Springs), and the South region (eg, Colorado Springs).

Collaboration With Key Partners

Industry Partner

Our health care system is in partnership with Avaamo's generative artificial intelligence (AI) conversational platform

XSL•FO
RenderX

to deliver a self-service patient experience called Livi. Livi can help patients by answering clinical and logistical questions, refilling prescriptions, resetting passwords, and performing other Epic EHR-based tasks. The team responsible for deploying the Livi chatbot at our health system selected our team to conduct a pilot project that extends Livi's capabilities for ED patients.

The Avaamo platform includes advanced generative AI-based natural language understanding. This enables computers to better understand the meaning and intent of human language from patients. Within the Avaamo platform, the Livi team created intents and responses for fall prevention topics, then trained the platform by feeding it sample queries (eg, "Where can I find fall prevention resources?") and contextual clues (eg, cities and health care specialties). In the version we developed, the Livi chatbot did not use its full large language model (LLM) capabilities. Instead, we limited it to providing rule-based answers to common fall prevention questions and specific resources that were reviewed and selected by our team. Future versions of the intervention are likely to deploy the full range of Livi's capabilities, but we opted to deploy a more limited version at this proof-of-concept stage to reduce the risk of "hallucinations" and to better control the accuracy and appropriateness of the content provided to patients.

Health System and Clinical Partners

We collaborated with our health system's EHR team (the ED-designated Epic team) and the Livi chatbot team to create the new notification and referral workflow. Creating an automated link to Livi in the after-visit summary (AVS) of older adult patients at high fall risk required communication and approval from the EHR and ED operations teams. Our team was the first within the health system to suggest an automated notification and referral process in the ED. This process required review and approval by the ED vice chair of clinical operations, consultation with the IT service management team, and coordination of the development, testing, and planning for the rollout of the EHR change. This coordination, from the initial planning to the launch, occurred over the course of 11 months. It involved synchronous meetings with multiple teams (ED-designated Epic team, Livi chatbot team, and ED clinical operations) that were scheduled as needed for each stage of development and occurred approximately every 3 months. Collaboration between meetings occurred approximately monthly via email with task lists and regular updates.

User Input and Usability

We used rapid usability testing and an iterative design with end users during the development process. This user-centered design method involves gathering real-time feedback during the design process and making quick adjustments to improve functionality and experience [11,12]. Engaging end users early in the prototyping process enhances usability, ensures alignment with user needs, and increases adoption rates [13,14]. This approach aligns with implementation science frameworks that emphasize user engagement and iterative adaptation to optimize intervention effectiveness [15,16].

We engaged 93 users across 3 community events (known as research roadshows) to test usability and elicit suggestions for refinement. These sessions were held at a religious organization, an assisted living facility, and an academic medical campus in the greater Denver area. Participants included older adults, caregivers, and experts in aging and fall prevention. Using their own devices, participants scanned the QR code and navigated the Livi chatbot modules. They were encouraged to think aloud while interacting with the intervention, providing real-time feedback on its usability [17]. We also solicited specific suggestions for improvement, which guided multiple iterative updates to the Livi chatbot before implementation.

Results

Livi Chatbot Intervention Workflow

Screening for Fall Risk—Routine Care

The workflow incorporating Livi is illustrated in Figure 1. In this workflow, the ED nurse first administers a fall risk screen and documents the results in the EHR nursing flow sheet as part of routine care. This fall risk screen used is the health system's Geriatric Fall Risk Score (8 questions; total possible score range, 0-24), as shown in Multimedia Appendix 1.

Figure 1. Flowchart illustrating fall risk screening and Livi chatbot resource referral within the emergency department (ED) workflow. EHR: electronic health record.

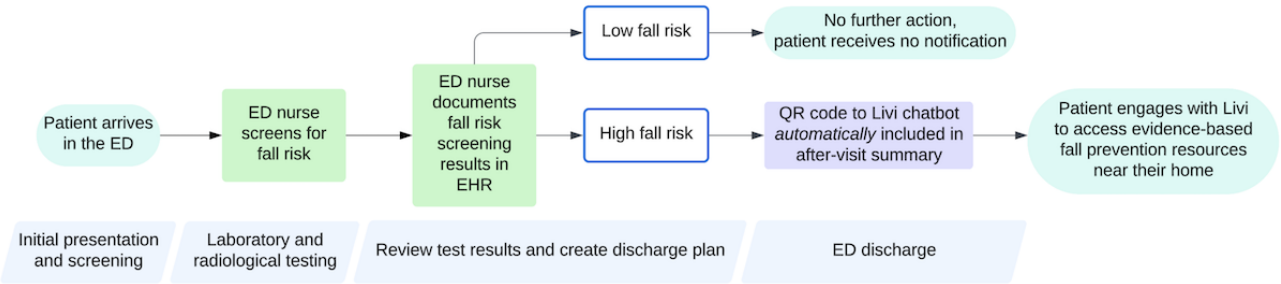

Notification of Risk and Fall Resource Referral—Livi Intervention

In the intervention, only those patients who score ≥3 (high fall risk) receive a standardized fall risk message in their AVS indicating that they screened at high fall risk and are directed to scan a QR code linking them to Livi. The AVS is printed for every patient discharged from the ED. Figure 2 illustrates the communication provided to the patient within the AVS. The AVS can be viewed by the patient in the patient portal, and the patient also receives a paper version upon ED discharge.

XSL•FO
RenderX

Figure 2. Screenshot of a message that appears in the after-visit summary for emergency department patients who screened at high fall risk. The QR code directs the patient to the Livi chatbot for fall prevention.

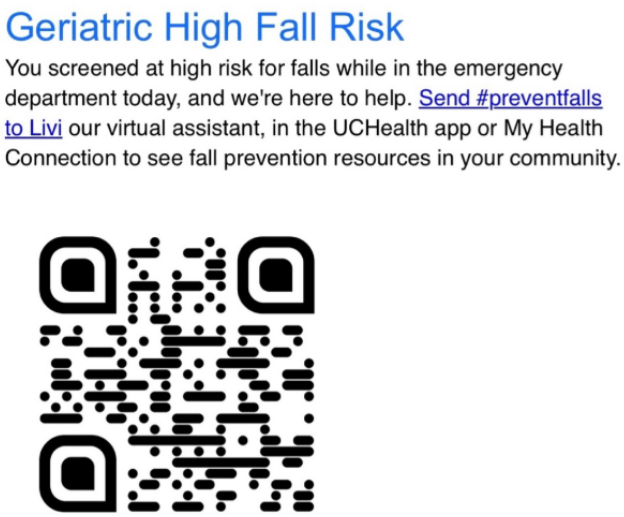

Patients deemed at low risk (score<3) do not receive this communication in their AVS. This process is fully automated and capitalizes on the existing ED screening results in the nursing flow sheet. The automated process reduces the reliance on clinicians to inform the patient of their fall risk and to remember to insert fall prevention referrals into the AVS.

Livi

Screenshots of the Livi chatbot are shown in Figures 3-7. To connect patients with appropriate resources, Livi first prompts users to select the city closest to them (Figure 3). On the basis of this selection, the chatbot provides a curated list of local fall prevention resources, each accompanied by a brief description and contact information, including an email address, phone number, and website (Figures 4-6). In addition to location-specific recommendations, Livi can answer general questions about fall risk through interactive chat responses or user-selected options to learn more (Figure 7). Patients are also given the opportunity to provide feedback on the resources offered (Figures 4-6).

Figure 3. Screenshot of the Livi chatbot displaying location options. Patients select the area closest to their home, after which Livi displays fall prevention resources available in that area.

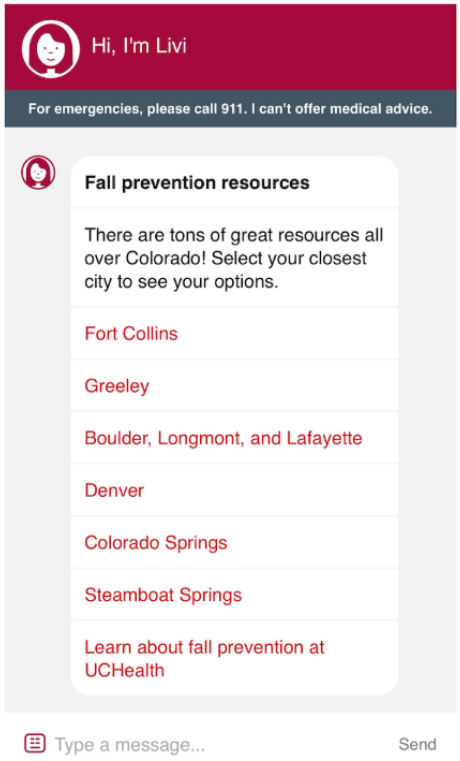

XSL•FO
RenderX

Figure 4. Screenshot of the Livi chatbot displaying a fall prevention resource available in the area selected by the patient (Fort Collins). The Livi chatbot describes the resource and contact information (including an email address and phone number), as well as a link to the organization's website.

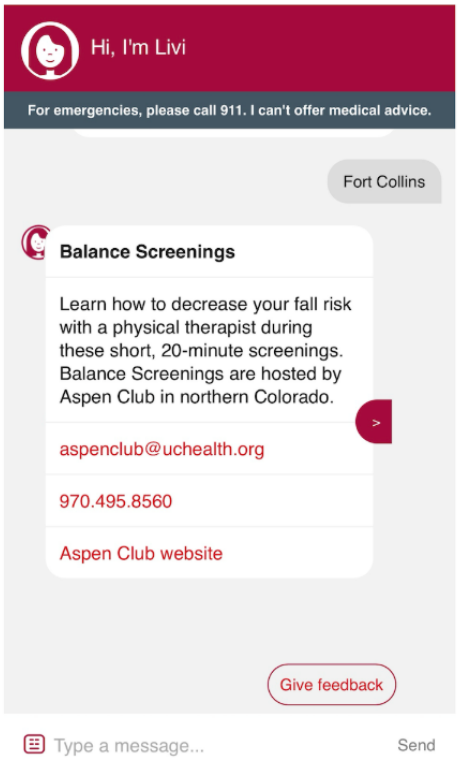

Figure 5. Screenshot of the Livi chatbot displaying a fall prevention resource available in the area selected by the patient (Denver). The Livi chatbot provides a description of the resource and contact information (including a phone number) and a link to the organization's website.

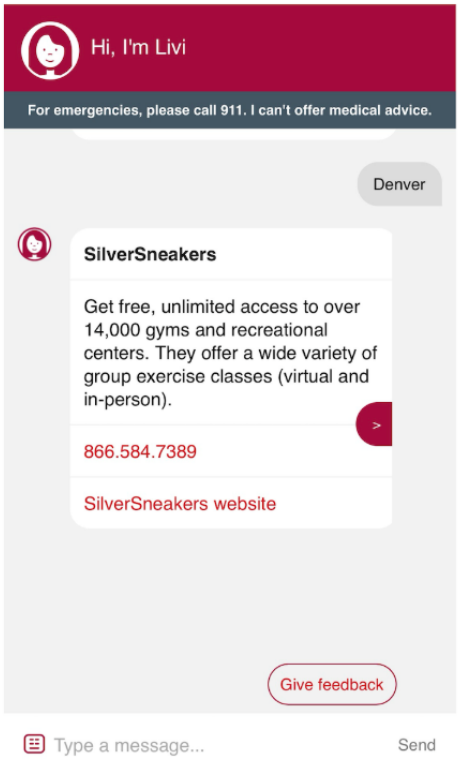

XSL•FO
RenderX

Figure 6. Screenshot of the Livi chatbot displaying a fall prevention resource. The Livi chatbot describes the resource and contact information (including a phone number), and a link to the organization's website. In this example, the geographic area contains several resources (Boulder, Longmont, and Lafayette); therefore, Livi provides additional options to display resources by specific region. The option for the patient to provide feedback on Livi fall prevention is displayed in the bottom right corner.

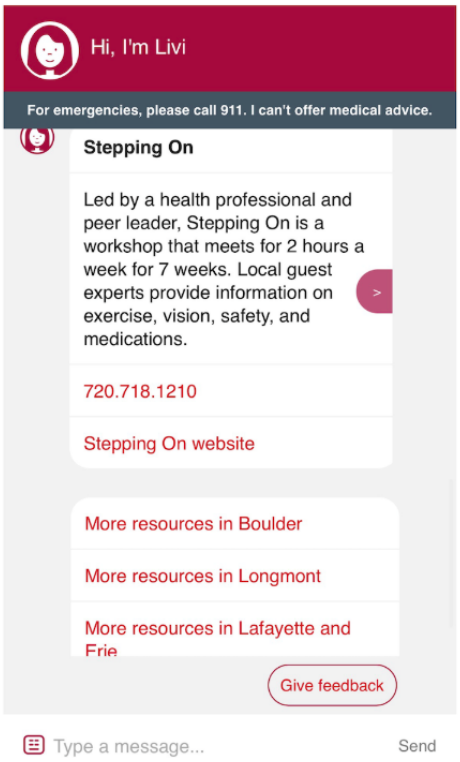

Figure 7. Screenshot of the Livi chatbot displaying information on the fall risk screening used in the emergency department, and additional information in response to a patient's question about fall risk screening.

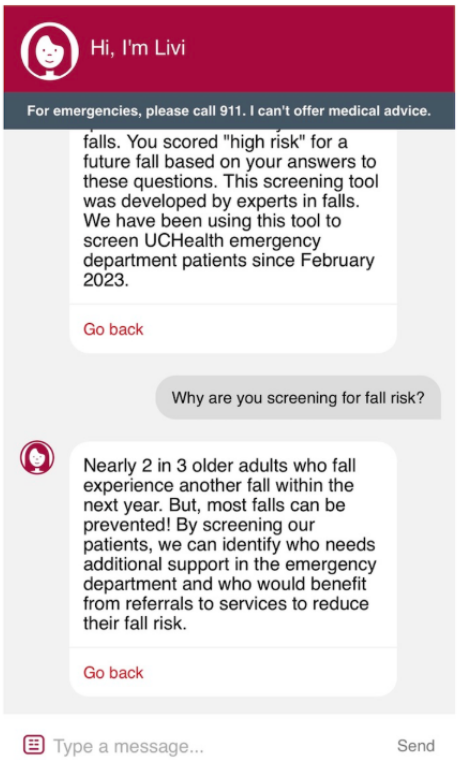

The final version of Livi was refined through rapid iterative feedback, as described subsequently. For this study, we did not use Livi's natural language understanding capabilities to initiate conversations. Instead, patients accessed Livi directly through a QR code or link that launched the Livi fall prevention dialogue. From there, structured button-based navigation guided the users through tailored conversation branches based on their selections.

XSL•FO
RenderX

Usability Testing—Iterative Design

The Livi chatbot underwent multiple iterations based on feedback from potential users (Textbox 1). Participants suggested several enhancements, including increasing font size, alternative language options, expanding geographic coverage across Colorado, and adding content on medications that increase fall risk, home modification resources, self-assessment tools, and answers to frequently asked questions about falls. Although some older adults experienced difficulty accessing Livi using the QR code, most expressed their support for the intervention.

In response, our team implemented several updates: we increased the font size, enabled language selection between English and Spanish, expanded resource coverage to include northern and southern Colorado, and incorporated the National Council on Aging 13-question fall risk self-assessment tool. Additional improvements included fall prevention tips and functionalities that allowed users to provide feedback directly within Livi.

In Table 1, we list several example questions that users could choose, along with responses from the Livi chatbot.

Textbox 1. Recommended modifications to Livi from usability testing with community members, and the resultant changes implemented.

Modifications suggested

- Larger font
- Spanish language
- More locations
- Medication risks
- Home modification resources
- Ability to self-assess fall risk
- Definition of “high fall risk”

Changes implemented

- Increased font size
- Enabled a Spanish version
- Added Colorado Springs and Steamboat Springs
- Currently, we are developing a plan to incorporate information about medications and fall risk into Livi
- Added “fall prevention tips” linking to a community health pamphlet on making the home safer
- Added “check my risk for falls” linking to the National Council on Aging validated fall risk survey
- Added a statement explaining how patients were screened for fall risk during their emergency department visit (Table 1)

Table 1. Example questions patients may ask the Livi chatbot and corresponding responses.

Prompt	Livi chatbot’s response
What does “high fall risk” mean?	During today’s emergency department visit, nurses asked you questions to determine your risk for falls. You scored “high risk” for a future fall based on your answers to these questions. This screening tool was developed by experts in fall risk and prevention. We have been using this tool to screen UCHealth emergency department patients since February 2023.
Why are you screening for fall risk?	Nearly 2 in 3 older adults who fall experience a recurrent fall within the next year, but most falls can be prevented! By screening our patients, we can identify who needs additional support in the emergency department and who would benefit from referral to services to reduce their fall risk.
What can I do to reduce my risk?	Select the city nearest to you to see fall prevention resources in your community. These services were vetted by fall prevention experts, and we only recommend programs proven to reduce your fall risk. You may also receive brochures with specific fall prevention classes, referrals to home or outpatient physical therapy, community paramedic home-based referrals, or home health aide referrals from our staff.
What is UCHealth’s goal for fall prevention?	We hope to reduce repeat falls and injuries and empower our patients to start a dialogue about their fall risk with their primary care provider. Because we screen you for fall risk, we want to ensure you have the tools to prevent falls after you leave our emergency departments.

Livi Go-Live

After incorporating modifications from iterative testing at community events, we launched Livi in all 17 EDs associated with our health care system (November 2024). We communicated the planned launch date, the purpose of the intervention, and shared a link to Livi with nursing staff. We also disseminated information about Livi by posting flyers in high-traffic areas within the ED. The flyer is included in

XSL•FO
RenderX

Multimedia Appendix 1. We encouraged ED nurses to point out the QR code and assist patients with accessing Livi at the point of ED discharge.

Implementation Barriers and Facilitators

We identified several informatics barriers during implementation. For example, PDF links within Livi would not open on Android devices and required reconfiguration of the fall-related resources. We also encountered challenges in integrating the QR code into the AVS, requiring collaboration with additional teams (acute and emergency business analytics) to automate Livi within the AVS of the EHR.

Facilitators of Livi implementation included the fact that all EDs within the health system use the same version of Epic, allowing the intervention to be rapidly implemented at all sites. Using a QR code that links to Livi, we can rapidly update information about the fall prevention programs when new information becomes available (eg, new contact number or location), ensuring that the information provided to patients remains timely and accurate.

Discussion

Principal Findings

We developed and implemented a novel digital health intervention that leverages existing fall risk screening in the ED to automate both patient notification and referral to community-based prevention resources. This approach addresses a critical gap in current care processes and exemplifies how digital tools can translate routine clinical data into automated, actionable insights. Early and ongoing engagement with end users during the design process—including older adults, caregivers, and community partners—was essential to ensure the intervention's accessibility, usability, and relevance.

This initiative required collaboration among clinical, operational, and informatics teams. Although not all health systems have invested in chatbot infrastructure, this experience highlights a replicable model: integrating validated screeners with localized, vetted resources and embedding the pathway into standard ED workflows. By sharing our process, we aim to support other health systems in developing contextually appropriate, scalable solutions for digital referral.

Iterative usability testing demonstrated a strong community interest in the Livi chatbot and surfaced critical design improvements, such as the need for a larger font size, multilingual options, and alternate access pathways. Offering multiple modes of entry (eg, QR code in AVS, access through the electronic patient portal, or clinical staff drawing attention to the intervention) may be essential to maximizing reach and reducing disparities in access due to the digital divide, whereby older adults may not have the same digital health literacy as younger populations.

ED visits are increasingly recognized as opportunities to improve population health by screening patients for health conditions [18-21]. Automating notification to patients about their high-risk status and referring them to resources could reduce reliance on health care workers at the point of care to remember to refer eligible patients and may improve efficiency. Such digital solutions are crucial, as they may also reduce the health system's exposure to legal, ethical, financial, and reputational risk when high-risk patients are not properly notified of their health risks and referred. Digital interventions embedded in EHR systems are recommended to expedite the integration of evidence-based interventions into practice [22].

Comparison to Prior Work

Prior studies have provided evidence that digital interventions can be used to successfully refer patients to outpatient or specialty services [23,24], enhance patient activation [25], and help address social determinants of health [26]. Our automated screening and referral intervention shares a few commonalities with other fall risk interventions in the acute care setting. One study found that automated referral to inpatient physical therapy based on routine mobility screenings was effective in reducing 30-day readmission and death [27]. Our intervention differs in that it targets ED patients who are discharged, and it refers patients to outpatient resources using an interactive tool, Livi. Another study also used fall risk screening questions to identify patients' fall risk and made clinicians aware of the patient's risk status with an EHR-embedded alert. The alert included language that explained why the patient was considered at risk, and allowed the clinician to select a button to order a referral to a fall clinic within their health care system [9]. Unlike our intervention, this solution interrupts the clinician and requires them to take action to place the referral. Our intervention does not interrupt or require action from the clinician, and it refers patients to free or low-cost community resources rather than a specialty clinic. Our automated intervention that connects patients to community resources not affiliated with the hospital system may be more easily implemented and adopted in other EDs. Further study is necessary to determine whether referral to a specific fall clinic may be more effective for fall prevention than a personalized referral that allows patients to peruse and select from local resources.

Other benefits of the Livi chatbot include its automatic linkage to the AVS of high fall risk ED patients, allowing patients to return to the chatbot using their AVS printout or through their online patient portal. Settings on a patient's phone or tablet, such as accessibility settings (eg, magnification or zoom in, bold font, screen reader capabilities, and color and contrast settings), allow the Livi information to be presented in a way that they are used to seeing and interacting with. Finally, an additional advantage of Livi is its ability to quickly update resource information, thereby ensuring that the information remains current.

Limitations

There are limitations to this work that may impact its scalability and effectiveness. First, less technologically advanced health systems may lack the capability to operationalize the intervention. Our health system may be unique in that it has a virtual health center, a central telehealth command center, with specific resources dedicated to deploying novel EHR-embedded interventions across the health system's Epic instance [28]. For hospitals lacking this capability, possible adaptations could include clinicians initiating the pathway to a linked website

XSL•FO
RenderX

based on screening results and making fall prevention resources available in a website format rather than a chatbot. This would add steps to the clinician workflow and may be less user-friendly, but could still enable fall risk screening and prevention resources to reach high-risk patients. Other health systems with limited technological abilities may choose to develop resources that can automatically be appended to the AVS in PDF format, consistent with more typical discharge paperwork. Second, the intervention is algorithmic and requires manual programming of responses rather than taking full advantage of the potential of LLMs, which offer more natural conversation dialogue capabilities. Third, patients still need to engage with the Livi chatbot and its resources to benefit from fall prevention programs. We recognize that the adaptability and impact of the intervention are highly dependent on patients' digital literacy and device access. Fourth, free and low-cost community-based fall prevention programs alone may not be enough to prevent falls; more personalized approaches may be needed in some cases that involve paid services, such as occupational therapy-led home safety evaluations and modifications, physical therapy services, or pharmacist-led medication reconciliation and management.

Future Directions

We would like to expand Livi's conversational capacity using AI while maintaining the validity of its output and full confidentiality of protected health information. The current Livi build is designed to algorithmically provide answers to common fall-related questions and vetted fall prevention resources for patients. The introduction of additional LLM features to Livi could aid in the conversational style, but these would need to be carefully balanced to prevent the incorrect output of information or "hallucinations." Although older patients have been found to trust the health care system chatbot privacy [29], communicating patient privacy protections remains a priority, especially as expanded capabilities and chatbot access to data are being considered.

Looking ahead, we plan to expand Livi's language options and continue enhancing its usability and feasibility for both older adults and ED staff. We are training nurses to demonstrate AVS use and plan to leverage available caregivers to help older adults use these tools. We aim to improve navigation by allowing users to browse resources by category, such as exercise programs, home modifications, and support for fear of falling, to better match individual needs. To further support continuity of care, we intend to incorporate a template that helps patients initiate a conversation with their primary care physician about fall risk. This feature could empower patients to share their screening results and more easily engage in preventive care discussions.

Finally, we will assess the effectiveness of the intervention in a future randomized controlled trial, evaluating its impact on clinical outcomes, including fall prevention behaviors, fall-related knowledge, and fall incidence. We will also collect more detailed use data and patient perspectives on the follow-up use of resources. We will gather demographic data on users to better understand who the tool is most useful for and how to further adapt it to increase scalability and inclusivity. We will measure implementation outcomes using the reach, effectiveness, adoption, implementation, and maintenance framework [30]. Our planned trial of Livi will include important clinical and usability outcomes, such as reduction in recurrent ED visits for falls, rates of participant attendance at community fall prevention programs, and likelihood of recommending the tool to others. These findings will inform future enhancements and guide broader implementation.

Conclusions

This digital intervention offers a pragmatic, scalable solution to the longstanding gap in emergency care, ensuring that older adults identified as having a high fall risk are screened, notified, and connected to relevant, accessible fall prevention resources. By automating this process through existing clinical workflows, we aim to reduce the burden on frontline clinicians while advancing equity, efficiency, and population health. Our model is readily adaptable to other health systems seeking to transform screening into meaningful action, setting the stage for the broader implementation of digital tools that bridge acute care with community-based prevention.

Funding

Research reported in this publication was supported by the National Institute On Aging of the National Institutes of Health under Award Numbers K76AG059983 and K24AG084999. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The geriatric fall risk assessment tool and flyer used to inform emergency department staff about Livi.
[DOCX File , 374 KB - formative_v9i1e77237_app1.docx]

References

1. Bergen G, Stevens MR, Burns ER. Falls and fall injuries among adults aged ≥65 years - United States, 2014. MMWR Morb Mortal Wkly Rep 2016 Sep 23;65(37):993-998 [FREE Full text] [doi: 10.15585/mmwr.mm6537a2] [Medline: 27656914]

XSL•FO
RenderX

2. Salter AE, Khan KM, Donaldson MG, Davis JC, Buchanan J, Abu-Laban RB, et al. Community-dwelling seniors who present to the emergency department with a fall do not receive guideline care and their fall risk profile worsens significantly: a 6-month prospective study. Osteoporos Int 2006 Feb 21;17(5):672-683. [doi: 10.1007/s00198-005-0032-7] [Medline: 16491323]
3. Gillespie LD, Robertson MC, Gillespie WJ, Sherrington C, Gates S, Clemson LM, et al. Interventions for preventing falls in older people living in the community. Cochrane Database Syst Rev 2012 Sep 12;2012(9):CD007146 [FREE Full text] [doi: 10.1002/14651858.CD007146.pub3] [Medline: 22972103]
4. Burns E, Kakara R, Moreland B. A CDC compendium of effective fall interventions: what works for community-dwelling older adults. Centers for Disease Control and Prevention. 2022. URL: https://www.cdc.gov/falls/pdf/steadi_compendium_2023_508.pdf [accessed 2025-11-04]
5. Davenport K, Alazemi M, Sri-On J, Liu S. Missed opportunities to diagnose and intervene in modifiable risk factors for older emergency department patients presenting after a fall. Ann Emerg Med 2020 Dec;76(6):730-738. [doi: 10.1016/j.annemergmed.2020.06.020] [Medline: 33010956]
6. Shankar KN, Treadway NJ, Taylor AA, Breaud AH, Peterson EW, Howland J. Older adult falls prevention behaviors 60 days post-discharge from an urban emergency department after treatment for a fall. Inj Epidemiol 2017 Dec 19;4(1):18 [FREE Full text] [doi: 10.1186/s40621-017-0114-y] [Medline: 28626848]
7. Patterson BW, Engstrom CJ, Sah V, Smith MA, Mendonça EA, Pulia MS, et al. Training and interpreting machine learning algorithms to evaluate fall risk after emergency department visits. Med Care 2019 Jul;57(7):560-566 [FREE Full text] [doi: 10.1097/MLR.0000000000001140] [Medline: 31157707]
8. Haimovich AD, Shah MN, Southerland LT, Hwang U, Patterson BW. Automating risk stratification for geriatric syndromes in the emergency department. J Am Geriatr Soc 2024 Jan 09;72(1):258-267. [doi: 10.1111/jgs.18594] [Medline: 37811698]
9. Jacobsohn GC, Leaf M, Liao F, Maru AP, Engstrom CJ, Salwei ME, et al. Collaborative design and implementation of a clinical decision support system for automated fall-risk identification and referrals in emergency departments. Healthc (Amst) 2022 Mar;10(1):100598 [FREE Full text] [doi: 10.1016/j.hjdsi.2021.100598] [Medline: 34923354]
10. Huttinger R, Aeddula NR. Stark law. In: StatPearls. Treasure Island, FL: StatPearls Publishing; Oct 06, 2022.
11. Huguet A, Rozario S, Wozney L, McGrath PJ. An online psychological program for adolescents and young adults with headaches: iterative design and rapid usability testing. JMIR Hum Factors 2023 Dec 12;10:e48677 [FREE Full text] [doi: 10.2196/48677] [Medline: 38085567]
12. Kushniruk AW, Borycki EM. Integrating low-cost rapid usability testing into agile system development of healthcare IT: a methodological perspective. Stud Health Technol Inform 2015;210:200-204. [Medline: 25991130]
13. van Kessel K, Babbage DR, Kersten P, Drown J, Sezier A, Thomas PW, et al. Design considerations for a multiple sclerosis fatigue mobile app MS Energize: a pragmatic iterative approach using usability testing and resonance checks. Internet Interv 2021 Feb 02;24:100371 [FREE Full text] [doi: 10.1016/j.invent.2021.100371] [Medline: 33614414]
14. Yardley L, Morrison L, Bradbury K, Muller I. The person-based approach to intervention development: application to digital health-related behavior change interventions. J Med Internet Res 2015 Jan 30;17(1):e30 [FREE Full text] [doi: 10.2196/jmir.4055] [Medline: 25639757]
15. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. Implement Sci 2009 Aug 07;4(1):1-15 [FREE Full text] [doi: 10.1186/1748-5908-4-50] [Medline: 19664226]
16. Rolland B, Resnik F, Hohl SD, Johnson LJ, Saha-Muldowney M, Mahoney J. Applying the lessons of implementation science to maximize feasibility and usability in team science intervention development. J Clin Transl Sci 2021 Jul 22;5(1):e197 [FREE Full text] [doi: 10.1017/cts.2021.826] [Medline: 34888066]
17. Nielsen J. Thinking aloud: the #1 usability tool. Nielsen Norman Group. 2012 Jan 15. URL: https://www.nngroup.com/articles/thinking-aloud-the-1-usability-tool/ [accessed 2025-11-04]
18. Ball JR, Balogh E. Improving diagnosis in health care: highlights of a report from the National Academies of Sciences, Engineering, and Medicine. Ann Intern Med 2016 Jan 05;164(1):59-61. [doi: 10.7326/M15-2256] [Medline: 26414299]
19. Greenwood-Lee J, Jewett L, Woodhouse L, Marshall DA. A categorisation of problems and solutions to improve patient referrals from primary to specialty care. BMC Health Serv Res 2018 Dec 20;18(1):986 [FREE Full text] [doi: 10.1186/s12913-018-3745-y] [Medline: 30572898]
20. Tinazzi M, Gandolfi M, Menaspà Z, Sandri A, Landi S, Leardini C. Reducing healthcare costs by timely diagnosis and management in functional motor disorders. Neurol Sci 2025 Mar;46(3):1191-1200. [doi: 10.1007/s10072-024-07865-2] [Medline: 39532797]
21. Weber W, Heins A, Jardine L, Stanford K, Duber H. Principles of screening for disease and health risk factors in the emergency department. Ann Emerg Med 2023 May;81(5):584-591. [doi: 10.1016/j.annemergmed.2022.06.015] [Medline: 35940988]
22. Trinkley KE, An R, Maw AM, Glasgow RE, Brownson RC. Leveraging artificial intelligence to advance implementation science: potential opportunities and cautions. Implement Sci 2024 Feb 21;19(1):17 [FREE Full text] [doi: 10.1186/s13012-024-01346-y] [Medline: 38383393]

XSL•FO
RenderX

23. Straus SG, Chen AH, Yee H, Kushel MB, Bell DS. Implementation of an electronic referral system for outpatient specialty care. AMIA Annu Symp Proc 2011;2011:1337-1346 [FREE Full text] [Medline: 22195195]
24. Odisho AY, Lui H, Yerramsetty R, Bautista F, Gleason N, Martin E, et al. Design and development of referrals automation, a SMART on FHIR solution to improve patient access to specialty care. JAMIA Open 2020 Oct 28;3(3):405-412 [FREE Full text] [doi: 10.1093/jamiaopen/ooaa036] [Medline: 33215075]
25. Lightfoot CJ, Wilkinson TJ, Sohansoha GK, Gillies CL, Vadaszy N, Ford EC, et al. The effects of a digital health intervention on patient activation in chronic kidney disease. NPJ Digit Med 2024 Nov 12;7(1):1-13 [FREE Full text] [doi: 10.1038/s41746-024-01296-1] [Medline: 39533053]
26. Buitron de la Vega P, Losi S, Sprague Martinez L, Bovell-Ammon A, Garg A, James T, et al. Implementing an EHR-based screening and referral system to address social determinants of health in primary care. Med Care 2019 Jun;57 Suppl 6 Suppl 2:S133-S139. [doi: 10.1097/MLR.0000000000001029] [Medline: 31095052]
27. Chou A, Johnson JK, Jones DB, Euloth T, Matcho BA, Bilderback A, et al. Effects of an electronic health record-based mobility assessment and automated referral for inpatient physical therapy on patient outcomes: a quasi-experimental study. Health Serv Res 2023 Feb 20;58(Suppl 1):51-62 [FREE Full text] [doi: 10.1111/1475-6773.14087] [Medline: 36271503]
28. Goldberg E, Kao D, Kwan B, Patel H, Hassell A, Zane R. UCHealth's virtual health center: how Colorado's largest health system creates and integrates technology into patient care. NPJ Digit Med 2024 Jul 11;7(1):187 [FREE Full text] [doi: 10.1038/s41746-024-01184-8] [Medline: 38992097]
29. Ellis JR, Dellavalle NS, Hamer MK, Akerson M, Andazola M, Moore AA, et al. The halo effect: perceptions of information privacy among healthcare chatbot users. J Am Geriatr Soc 2025 May 12;73(5):1472-1483. [doi: 10.1111/jgs.19393] [Medline: 39936842]
30. Glasgow RE, Vogt TM, Boles SM. Evaluating the public health impact of health promotion interventions: the RE-AIM framework. Am J Public Health 1999 Sep;89(9):1322-1327. [doi: 10.2105/ajph.89.9.1322] [Medline: 10474547]

Abbreviations

AI: artificial intelligence
AVS: after-visit summary
ED: emergency department
EHR: electronic health record
LLM: large language model

Edited by J Sarvestan; submitted 12.05.25; peer-reviewed by M Sharma, A Gupta, V Ajibade, S Mohanadas, C Obianyo; comments to author 20.06.25; revised version received 23.07.25; accepted 14.10.25; published 18.11.25.

Please cite as:
Keleman A, Bounds M, Lunt M, Portz J, Ferozan B, Gomez Picazo J, Bookman K, Lum HD, Goldberg EM
Development and Health System Deployment of an Electronic Health Record–Integrated Chatbot Intervention for Connecting Fall Risk Screening to Community Resources After Emergency Department Visits: Implementation Study
JMIR Form Res 2025;9:e77237
URL: https://formative.jmir.org/2025/1/e77237
doi:10.2196/77237
PMID:41252184

©Audrey Keleman, Megan Bounds, Maxwell Lunt, Jennifer Portz, Bucky Ferozan, Jonathan Gomez Picazo, Kelly Bookman, Hillary D Lum, Elizabeth M Goldberg. Originally published in JMIR Formative Research (https://formative.jmir.org), 18.11.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Case Report

Enhancing Parenting Using AI: Exploratory Hackathon

Peter Woods[1*], BA, MBChB, PGCert(Genomic Med), PGCert(Med Ed), MRCP; Stephanie Donohoe[1*], MA; Louise Turtle[1], PGCE, MA; Udit Agrawal[1], BSc; Joshua Humphriss[1]; Niel Cordes[1], MSci; Nathan Hodson[1], MBBS, MPH, MRCPsych

Warwick Medical School, University of Warwick, Coventry, England, United Kingdom

[*]these authors contributed equally

Corresponding Author:
Peter Woods, BA, MBChB, PGCert(Genomic Med), PGCert(Med Ed), MRCP
Warwick Medical School
University of Warwick
Gibbet Hill Road
Coventry, England, CV4 7AL
United Kingdom
Phone: 44 2476 574 ext 880
Email: peter.woods@warwick.ac.uk

Abstract

Background: Parenting skills programs are the primary intervention for conduct disorders in children. The Pause app enhances these programs by providing digital microinterventions that reinforce learning between sessions and after program completion. The potential of artificial intelligence (AI) in this context remains untapped. Hackathons have proven effective for health care innovation and can facilitate collaborative development in this space.

Objective: We aimed to rapidly build AI-powered features in the Pause app to enhance parenting skills.

Methods: We undertook a 1-day hackathon that included an ideation phase drawing on the Design Council's double diamond framework and a development phase using microsprints based on agile and scrum approaches. The interdisciplinary participants included medical professionals, developers, and product managers.

Results: Participants identified 3 core problems: generating age-appropriate distractions, receiving feedback on parenting efforts, and effectively using the journal function. During the solution phase, a wide range of options were explored, resulting in 3 key solutions: AI-assisted idea generation, a tool for summarizing parenting interactions, and a weekly journal roundup. During the development phase, participants completed 4 microsprints. Teams focused on 3 workstreams: building a "weekly roundup" module, creating an AI-based distraction generator, and developing a summarizer for active play sessions. These prototypes were integrated into the preproduction environment, with each workstream producing a functional component. Participant feedback (n=4) was unanimously positive, with all participants rating the event as "excellent" and highlighting the value of in-person collaboration.

Conclusions: This 1-day hackathon used the double diamond approach to develop AI-powered features for parenting programs. Three solutions were explored across workstreams, resulting in 2 fully functioning and 1 near-functioning app component. The rapid problem-solving approach mirrors other health technology hackathons and highlights the untapped potential of AI in digital parenting support, surpassing traditional e-learning or video-based methods. This work suggests broader applications of AI-driven coaching in fields like social care. Despite a small team, the hackathon was focused and productive, generating relevant solutions based on prior engagement with parents and practitioners. Future research will assess the impact of the app's AI-powered features on parenting outcomes.

(JMIR Form Res 2025;9:e68780) doi:10.2196/68780

KEYWORDS

hackathon; mental health; parenting; psychiatry; large language models

XSL•FO
RenderX

Introduction

Background

Parenting skills programs are the first-line treatment recommended for conduct disorder, the most common mental health disorder of childhood [1,2]. The Pause app enhances parenting skills programs by providing digital support between parenting program sessions [3,4]. Specifically, the Pause app contains a series of digital microinterventions that map to the core skills parents learn; between sessions, parents are prompted to use these digital microinterventions to strengthen their skills and reflect on their success [3,5].

The rapid increase in available artificial intelligence (AI) tools provides opportunities and challenges for children and families. For instance, in the educational domain, concerns have been raised regarding the impact of large language models (LLMs) on academic integrity and critical thinking, highlighting the disruptive potential of AI [6] alongside its benefits. Simultaneously, although AI presents novel avenues for support, it has not yet been systematically leveraged to strengthen parenting skills programs, which are crucial for the child's well-being.

Hackathons have previously provided an effective "stepping stone for health care innovation" by creating space for "blue sky thinking" and collaboration [7]. They are structured using design-thinking approaches to address specific pain points [8]. The Pause app provides a platform through which the potential of AI can be directed toward solving parenting challenges. This paper reports on an exploratory hackathon, the central theme of which was to rapidly co-design and prototype AI-powered features to enhance the support offered by the Pause app, thereby addressing specific parenting needs identified through prior user engagement.

Objective

We conducted a hackathon with the aim of rapidly building AI-powered features to enhance parenting skills within the framework of the Pause app.

Methods

Theoretical Approach

This hackathon was designed in keeping with the best practices reported by Poncette et al [7]. The design approach used was the Design Council's double diamond framework [9]. In this approach, participants come up with as many problems as possible and then narrow them down to the most important problems (the first diamond), then they come up with as many solutions as possible before narrowing down to the best solutions (the second diamond). The selected development approach was a series of microagile development sprints [10].

Figure 1 presents a schematic outlining the structure of the 1-day exploratory hackathon. The hackathon was designed to rapidly develop AI-powered features for the Pause parenting app and comprised a design phase using the Design Council's double diamond framework, followed by an iterative development phase using microagile sprints.

Figure 1. A diagrammatic illustration of the hackathon.

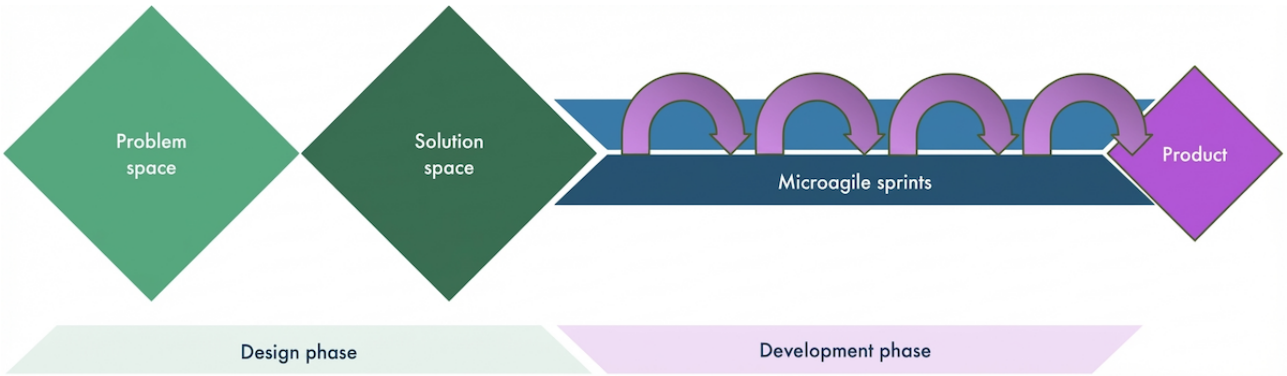

Community Building and Stakeholder Involvement

We undertook a bootstrapping-style hackathon, not a corporate-style event, to enable a small team to work on specific issues [7]. Participants in this hackathon were drawn from the digital mental health teams at the University of Warwick and several of them had previously worked together. Overall, 2 medical doctors, 3 developers, and 2 product managers participated.

Hackathon Theme

The hackathon theme was set through development work with parents and parenting program practitioners who have used the Pause app [3,4]. The theme was announced to participants in advance. Three of the participants had been involved in analyzing the user data.

Venue and Timing

The selected venue was a collaboration space at the University of Warwick, due to its available facilities for individual and group work. The development day took place on a Sunday so those with clinical and industry commitments in the usual working week could collaborate.

Announcement

The hackathon was announced 1 month before the main event. Participants knew they would be building on the Pause app, and all participants had access to the Pause app before launch. All developers had access to the Pause development area before coding.

XSL•FO
RenderX

Ethical Considerations

The hackathon activities described in this manuscript focused on formative development work by the study team and did not involve direct recruitment or new data collection from human subjects. However, the hackathon drew on anonymized patient and practitioner themes obtained from a previous foundational study [3], which had received ethics approval from the University of Warwick's Biomedical Sciences Research Ethics Committee (BSREC 14/23-24). For that foundational study, informed consent was obtained from participants for their feedback to be used for improving digital mental health provisions, including the Pause app. For the hackathon itself, team members were invited to contribute based on their professional roles.

No new personal data were collected from hackathon participants. Data from the previous research were fully anonymized, and only nonidentifiable themes were used during the hackathon. Parents and practitioners from the foundational study were compensated with a US $32 voucher [3]. Hackathon team members contributed voluntarily as honorary fellows or as part of their compensated professional duties. No images in this manuscript identify individuals, and any future illustrative images will ensure anonymity or be accompanied by explicit consent.

Hackathon Plan

The hackathon development day lasted from 10 AM to 6 PM with a design phase and a development phase. At 10 AM, there was a presentation that included icebreakers, inspirational talks, and the acknowledgment of participant expertise. Then, the 3 core problems areas were presented (idea generation, feedback on parenting encounters, and journal fatigue). Next, participants had a 30-minute group discussion to explore the problems in the widest sense before narrowing down on definitions of specific problems. Then, participants broke into smaller groups to consider the widest range of possible solutions, spending 1 hour on this. Another whole-group discussion was undertaken to select the most promising solutions.

The remainder of the day was broken into augmented microagile sprints [10]. Participants had 1 hour to conduct development microsprints during which time participants worked on different solutions. After every 60 minutes, a 5-minute microreview was conducted, where participants reviewed the progress of each other. Microreviews followed the agile scrum "stand-up" and "sprint review" approach of asking (1) what they had done; (2) what they were about to do; and (3) whether there were any barriers that needed to be unblocked [11].

Lessons Learned and Follow-Up

The hackathon day ended with a presentation of the working code produced [10]. Functioning solutions were shared with all participants. Next steps for development were identified. Feedback was shared verbally and in writing, with space for a "miniretrospective" to improve future ways of working [11].

Hackathon Process and Prototyping Outputs

Broadening the Problem Space

Three core ideas were presented: (1) parents need more ideas; (2) parents need more feedback; and (3) parents struggle with journaling. A wide range of problems were considered around these ideas.

Narrowing the Problem Space

Three problems were selected. The first was that parents find it hard to come up with age-appropriate distractions for their children. The second was that parents do not obtain feedback on their efforts to encourage their children. The third was that parents do not use the journal function effectively.

Broadening the Solution Space

A wide range of solutions were considered, some of which addressed multiple areas in the problem space. Participants considered adding new prompts for ways in which parents can spend time together with their children. They considered whether to provide real-time audio coaching to parents while they played with their child. Participants considered creating journal voice notes or journal entries through text and using these to create a coaching conversation with an AI chatbot. They also considered creating a function to summarize the content of time together and provide feedback. Participants considered allowing parents to select and rate from pregenerated ideas, generating prompts for parents, and providing parents with a weekly roundup of journaling activity over the week, inspired by Spotify Wrapped.

Narrowing the Solution Space

Participants settled on 3 solutions, one matching each of the initial problems. The first was AI-assisted idea generation for distracting children, the second was a tool to summarize and encourage time together with children, and the third was a weekly roundup to gamify the journaling process. These 3 ideas formed the basis for 3 workstreams during the development phase.

It maps the 3 core parenting challenges presented to participants, the specific problem definitions narrowed down during the first diamond of the Design Council's double diamond framework, examples of solutions explored, and the final AI-powered solutions selected for prototype development during the hackathon held at the University of Warwick.

Development Phase

The development phase proceeded through four 1-hour microsprints covering the 3 workstreams. These are described in Table 1. The table details the specific tasks undertaken and outputs achieved by the interdisciplinary teams (medical doctors, developers, and product managers) in each workstream.

XSL•FO
RenderX

Table 1. Outcomes of development microsprints.

	Workstream 1	Workstream 2	Workstream 3
Microsprint 1	• Team 1 mapped out a new "weekly roundup" module. • It will pop-up every week, ideally the day before their next session. • It will ask parents questions about what they think their main section of the app is. • Then, it will also give them some data on their parenting. • This was sketched, and no code was made.	• Team 2 started making an Android APIa to pull information through to the app. • Team 2 had some difficulty making it work, so they asked colleagues to identify the bug.	• Team 3 made a "time together" summarizer, which can transcribe a short conversation and then summarize it, including positives and negatives.
Microsprint 2	• Team 1 finished sketching out the module and prepared it in the development space. • A new module was made within the app codebase.	• Team 2 successfully made the API generate 5 personalized distractions based on a prompt word.	• Team 3 made a Time Together summarizer which can listen to a conversation and then summarize it, including positives and negatives.
Microsprint 3	• Team 1 started drafting code to pull together star ratings and gather parents' thoughts.	• Team 2 incorporated the AIb distraction generator into the distract module. • Parents can add the AI-generated distractions into their list of distractions. • Parents can delete distractions they do not like.	• Team 3 incorporated the summarizer into the app so that when they hit play on "Time Together" it listens and then gives them feedback. • Team strengthened the relevance of the parenting information.
Microsprint 4 (end of hackathon)	• Team 1 mapped out the module flow and prepared the back-end code.	• Team 2 added a "location" input as well as an "interests" input. • This meant that the app generated highly targeted distractions for the situation people were in.	• Team 3 completed a module where a time together session was summarized and positive and negative feedback were given. • This feedback was then added directly to the journal and parents could add a star rating out of 5 for their experience.

[a]API: application programming interface.

[b]AI: artificial intelligence.

Results

Of the 7 participants who attended the hackathon, 4 (57%) completed a postevent feedback survey. The results were unanimously positive and are presented in subsequent sections.

Participant Ratings (Quantitative Results)

All respondents (4/4, 100%) rated the hackathon as "excellent." Furthermore, all respondents agreed that it would be "worth doing another hackathon," with comments, including "Yes!," "Definitely," and "When there's a need for new features."

Participant Experience (Qualitative Results)

We analyzed qualitative feedback, which fell into 3 themes: positive aspects, suggestions for improvement, and ideas for future events.

Positive Aspects

The most valued aspects of the hackathon were in-person collaboration and rapid development. Comments included the following:

> *Being face to face.*
>
> *Getting together to try out fresh new ideas.*
>
> *Coming up with ideas and making them quickly.*

Suggestions for Improvement

The main suggestion for future events was to have more time. One participant suggested it would be beneficial to do the following:

> *Have more than one day to try and get a fully working prototype and flesh out more ideas.*
>
> *Test with users and do some iteration.*

Ideas for Future Events

When asked about future events, participants suggested including more diverse expertise, such as,

> *Design...to think about interaction.*
>
> *Combining it [the hackathon] with a conference.*
>
> *Maybe lots of teams and prizes.*

XSL•FO
RenderX

Discussion

Principal Findings

The primary objective of this study was to rapidly build AI-powered features to enhance parenting skills via the Pause app. This objective was successfully met through a 1-day hackathon, which resulted in the development of 2 fully functioning and 1 near-functioning app component designed to address core parenting challenges identified from a previous stakeholder engagement. These outcomes confirm that an intensive, structured hackathon approach—using frameworks such as the double diamond and agile microsprints—is a viable methodology for the swift creation and initial integration of AI-powered tools in this domain. This methodological success is further supported by our new quantitative and qualitative participant feedback (see the Results section), where all respondents (4/4, 100%) rated the event as "excellent" and highlighted the value of in-person, rapid collaboration for "coming up with ideas and making them quickly." These prototypes are now prepared for subsequent testing and validation.

The AI-powered features developed during this hackathon offer potential advantages over traditional methods of delivering parenting skills support, such as static e-learning modules or periodic video calls. The AI-assisted idea generator for distractions offers a substantial step beyond static lists by providing suggestions tailored to the child's age, with the underlying AI capability allowing for future enhancements incorporating more sophisticated contextual prompts, thereby reducing parental cognitive load in real time. The AI-powered summarization tool for parenting interactions, leveraging multimodal input (specifically audio analysis), provides parents with timely and specific feedback, and offers a more objective basis for self-assessment and reinforcement of positive parenting techniques—a feature not readily achievable with traditional non-AI tools. Furthermore, the AI-driven weekly journal roundup aims to combat journal fatigue and enhance reflection by transforming logged experiences into an engaging, summarized format, potentially identifying patterns and highlighting successes. These capabilities highlight the potential of AI to deliver more interactive, responsive, and individualized support, thereby deepening the engagement and impact of digital parenting interventions. Our exploration of rapidly prototyped AI features aligns with emerging evidence in this field. For instance, Entenberg et al [12] found that an AI-based chatbot microintervention for parents demonstrated meaningful engagement, learning, and efficacy, underscoring the promise of AI in delivering accessible parenting support. This paper builds on such insights by demonstrating a rapid and collaborative methodology—the hackathon—for developing a broader suite of AI-driven tools tailored to specific parenting challenges, such as idea generation for distractions and reflective journaling, thereby expanding the repertoire of potential AI applications within digital parenting platforms.

The development of these interactive and personalized features was facilitated by the selection of LLMs, favored for their advanced text generation, summarization, and multimodal audio analysis capabilities. Specifically, publicly available models accessed via their respective APIs (application programming interfaces), with versions current as of October 2024, were used. This included GPT-4o by OpenAI for text-based processing and Gemini 1.5 Pro by Google for its multimodal capabilities, particularly audio analysis for the summarization tool. This choice reflects the rapidly evolving landscape of accessible AI tools. The rapid problem-solving approach achieved in this 1-day event mirrors successes reported in other health technology hackathons [7,8], demonstrating an effective pathway from identified user needs to functional prototypes. The immediate measure of success for the prototypes developed within this hackathon was their functional integration into the preproduction environment of the Pause app. Formal preliminary user feedback sessions or expert evaluations were not conducted during the hackathon itself, but are planned as crucial next steps for development and validation.

Existing digital parenting interventions have primarily used e-learning or video calls [13]. This project suggests that current AI technology has much deeper implications for supporting parents, particularly within a platform like the Pause app. The availability of LLMs also has wider implications for mental health beyond parenting, such as improving therapy processes [14]. Indeed, the use of AI to support, coach, and encourage carers could have profound implications in various domains, for example, in adult social care, an area yet to be fully explored.

The integration of AI into parenting supports, particularly tools that analyze personal interactions or provide advice, raises important ethical considerations beyond the conduct of this specific hackathon. These include ensuring data privacy and security for sensitive family information processed by LLMs, obtaining informed consent from parents for AI analysis of their inputs, and mitigating potential biases in AI-generated advice. The psychological impact of AI-generated feedback, such as the risk of overreliance or misinterpretation leading to parental anxiety, also warrants careful consideration and the development of clear guidelines for responsible deployment, emphasizing AI as a supportive tool rather than a replacement for human judgment or professional guidance.

Limitations

There are several limitations. The 1-day hackathon format, while fostering rapid development, inherently limits the depth of exploration and the complexity of the solutions that can be built. The development work with parents and parenting practitioners that preceded this hackathon meant we could prime participants to pursue relevant ideas; however, the small number of hackathon participant feedback (n=4), although conducive to a focused "bootstrapping" approach, may restrict the diversity of ideas and solutions generated compared to a larger event and introduces potential bias from a specialized group primarily from one institution. A larger, more diverse group of participants might generate a broader array of solutions but could also introduce logistical complexities in a single-day format. Furthermore, while this hackathon demonstrated technical feasibility, the integration of AI into sensitive contexts like parenting presents considerable challenges. These include ensuring the ethical use of data, managing user expectations,

XSL•FO
RenderX

preventing overreliance on AI-generated advice, addressing potential biases in AI models, and ensuring equitable access for all parents, which were only touched upon thematically during this event and require ongoing dedicated research.

Conclusions

The primary objective of this study, to rapidly build AI-powered features to enhance parenting skills via the Pause app, was successfully achieved. This 1-day hackathon resulted in the development of 2 fully functional and 1 near-functional app component, demonstrating the feasibility of this approach for swift prototyping and initial integration.

The key conclusions from this work are 2-fold: first, that intensive, structured co-design events similar to hackathons, using frameworks such as the double diamond and agile microsprints, serve as a valuable and efficient methodology for exploring and actualizing the potential of AI in applied digital health settings. Second, this exploration highlights that AI, particularly LLMs, offers novel avenues for enhancing digital parenting interventions, moving beyond traditional e-learning or video-based methods to provide more interactive and potentially personalized support.

We therefore recommend hackathons to other mental health teams considering the integration of AI into their digital platforms, as this format allowed our team to focus on underlying user problems and generate innovative solutions outside of routine work pressures. Future research will be essential to evaluate the uptake, user engagement, and ultimately the clinical impact of these AI-powered parenting support features on parenting skills and child outcomes.

Acknowledgments

The authors wish to thank Dr Joshua Cartwright for his support in organizing this hackathon.

Generative AI (Google Gemini) was used to assist with proofreading and grammar of this manuscript. The authors reviewed and edited all content and take full responsibility for the final manuscript.

Funding

The authors received funding from the Medical Research Council for this hackathon.

Data Availability

The hackathon reported in this manuscript focused on the development of software prototypes and did not generate traditional research datasets. Inquiries regarding the conceptual outputs or specific aspects of the prototypes described can be directed to the corresponding author, and access may be considered on a case-by-case basis, subject to institutional policies and the stage of development.

Conflicts of Interest

The Pause app is freely available online; however, two of the authors (NH and PW) have a division of proceeds arrangement with the University of Warwick with respect to any future commercialization.

References

1. Antisocial behaviour and conduct disorders in children and young people: recognition, intervention and management. National Institute for Health and Care Excellence. 2017. URL: https://pubmed.ncbi.nlm.nih.gov/26065062/ [accessed 2025-05-29]
2. No authors listed. Conduct disorder. Nat Rev Dis Primers 2019 Jun 27;5(1):44. [doi: 10.1038/s41572-019-0100-5] [Medline: 31249311]
3. Hodson N, Woods P, Solano JL, Talbot C, Giacco D. Evaluating a mobile app supporting evidence-based parenting skills: thematic analysis of parent experience. JMIR Pediatr Parent 2024 Sep 05;7:e53907 [FREE Full text] [doi: 10.2196/53907] [Medline: 39235353]
4. Hodson N, Woods P, Sobolev M, Giacco D. A digital microintervention supporting evidence-based parenting skills: development study using the agile scrum methodology. JMIR Form Res 2024 Jun 28;8:e54892 [FREE Full text] [doi: 10.2196/54892] [Medline: 38941594]
5. Baumel A, Fleming T, Schueller SM. Digital micro interventions for behavioral and mental health gains: core components and conceptualization of digital micro intervention care. J Med Internet Res 2020 Oct 29;22(10):e20631 [FREE Full text] [doi: 10.2196/20631] [Medline: 33118946]
6. Lee VR, Pope D, Miles S, Zárate RC. Cheating in the age of generative AI: a high school survey study of cheating behaviors before and after the release of ChatGPT. Comput Educ Artif Intell 2024 Dec;7:100253. [doi: 10.1016/j.caeai.2024.100253]
7. Poncette AS, Rojas P, Hofferbert J, Valera Sosa A, Balzer F, Braune K. Hackathons as stepping stones in health care innovation: case study with systematic recommendations. J Med Internet Res 2020 Mar 24;22(3):e17004 [FREE Full text] [doi: 10.2196/17004] [Medline: 32207691]

XSL•FO
RenderX

8. Silver JK, Binder DS, Zubcevik N, Zafonte RD. Healthcare hackathons provide educational and innovation opportunities: a case study and best practice recommendations. J Med Syst 2016 Jul 8;40(7):177 [FREE Full text] [doi: 10.1007/s10916-016-0532-3] [Medline: 27277278]
9. 11 lessons: a study of the design process. Design Council. 2022. URL: https://www.designcouncil.org.uk/our-resources/archive/reports-resources/11-lessons-managing-design-global-brands/ [accessed 2025-05-29]
10. The Agile Alliance. The agile manifesto. In: King D, Lawley S, editors. Cases in Organizational Behaviour. Oxford, UK: Oxford University Press; 2022.
11. Schwaber K, Sutherland J. The scrum guide. ScrumGuides. URL: https://scrumguides.org/docs/scrumguide/v2020/2020-Scrum-Guide-US.pdf#zoom=100 [accessed 2025-05-29]
12. Entenberg GA, Mizrahi S, Walker H, Aghakhani S, Mostovoy K, Carre N, et al. AI-based chatbot micro-intervention for parents: meaningful engagement, learning, and efficacy. Front Psychiatry 2023 Jan 20;14:1080770 [FREE Full text] [doi: 10.3389/fpsyt.2023.1080770] [Medline: 36741110]
13. Baumel A, Pawar A, Kane JM, Correll CU. Digital parent training for children with disruptive behaviors: systematic review and meta-analysis of randomized trials. J Child Adolesc Psychopharmacol 2016 Oct;26(8):740-749. [doi: 10.1089/cap.2016.0048] [Medline: 27286325]
14. Hodson N, Williamson S. Can large language models replace therapists? Evaluating performance at simple cognitive behavioral therapy tasks. JMIR AI 2024 Jul 30;3:e52500 [FREE Full text] [doi: 10.2196/52500] [Medline: 39078696]

Abbreviations

AI: artificial intelligence
API: application programming interface
LLM: large language model

Edited by A Mavragani; submitted 13.11.24; peer-reviewed by D Stojanovic, S Kheirinejad; comments to author 02.05.25; revised version received 10.11.25; accepted 11.11.25; published 17.12.25.

Please cite as:
Woods P, Donohoe S, Turtle L, Agrawal U, Humphriss J, Cordes N, Hodson N
Enhancing Parenting Using AI: Exploratory Hackathon
JMIR Form Res 2025;9:e68780
URL: https://formative.jmir.org/2025/1/e68780
doi:10.2196/68780
PMID:

©Peter Woods, Stephanie Donohoe, Louise Turtle, Udit Agrawal, Joshua Humphriss, Niel Cordes, Nathan Hodson. Originally published in JMIR Formative Research (https://formative.jmir.org), 17.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Adapted Virtual Reality Exergaming Using Off-the-Shelf Supplies for Poststroke Hemiparetic Arm Rehabilitation: Case Study

Sadie Hare[1*], OTD; Jason Vice[1*], PhD; Mary H Bowman[1*], BS OT; Ashley Wright[2*], MS; Raven Young[2*], BS; Mahmoud Ahmad[2*], MD; Maggie Logan[1*], BS; Byron Lai[2*], PhD

[1]Department of Occupational Therapy, University of Alabama at Birmingham, Birmingham, AL, United States

[2]Division of Pediatric Rehabilitation Medicine, Department of Pediatrics, University of Alabama at Birmingham, Birmingham, AL, United States

[*]all authors contributed equally

Corresponding Author:
Byron Lai, PhD
Division of Pediatric Rehabilitation Medicine
Department of Pediatrics
University of Alabama at Birmingham
701 20th Street South
Birmingham, AL, 35233
United States
Phone: 1 1 205 638 9790 ext 8
Email: byronlai@uab.edu

Abstract

Background: Virtual reality (VR) can supplement exercise therapy for poststroke upper-arm hemiparesis, but treatments have been largely limited by specialized or costly equipment, hindering replicability and generalizability.

Objective: This study examined the feasibility of using a commercially available bundle of VR supplies to improve hemiparetic arm function before and after an exergaming program in an individual post stroke.

Methods: We conducted a pre-post case study (male, aged 72 years, chronic stroke) of a 20-day VR exergaming program (1-hour session per day) using a head-mounted display (Meta Quest 2), with adaptive software (WalkinVR) to boost and adjust in-game movements. Measures of upper-arm function were performed at preintervention (day 0), midintervention (day 10), and postintervention (day 21) and included the Wolf Motor Function Test (WMFT) and Disabilities of the Arm, Shoulder, and Hand Questionnaire (DASH). Data were descriptively analyzed.

Results: The participant demonstrated improvement in the mean time to complete tasks of the WMFT by 70.5% (δ=11.73 s; preintervention mean time 16.63, SD 31 s; effect size=0.54) from preintervention to midintervention and 78% (δ=12.96 s; effect size=0.59) from preintervention to postintervention. WMFT mean functional ability score demonstrated an improvement of 18% (δ=0.46 points; preintervention mean score 2.67, SD 0.87 points; effect size=0.59) from preintervention to midintervention and 23% (δ=0.6 points; effect size=0.79) from preintervention to postintervention. Range of motion improved in all joints by an average of 35.64% (SD 20%) from preintervention to postintervention. DASH scores demonstrated minimal improvements across the intervention.

Conclusions: VR exergaming with adaptive software could be an easy-to-adopt method for improving the functional ability of the hemiparetic arm among people post stroke. Improvements were potentially meaningful but warrant confirmation in more rigorous study designs.

(JMIR Form Res 2025;9:e80721) doi:10.2196/80721

KEYWORDS

active video game; augmented reality; disability; exercise; extended reality; physical activity; rehabilitation

Introduction

A stroke results from occlusion of the blood supply to a part of the brain and often leads to many impairment-related functional limitations for those affected. Poststroke functional limitations hinder a person's ability to perform their activities of daily living and other meaningful daily tasks. Specifically, damage in motor control areas or associated pathways can lead to decreased

XSL•FO
RenderX

unilateral upper extremity function (ie, hemiparesis) [1,2]. Hemiparesis of an upper limb typically results in weakness of the affected limb, resulting in a limited range of motion and impaired ability to complete functional tasks. Upper-limb hemiparesis is a common impairment, making it a high-priority target for therapeutic interventions.

An effective method of regaining function in an affected limb is mirror therapy [3]. Mirror therapy can be described as involving the use of a mirror to hide the impaired limb of an individual while reflecting the unaffected one [4]. Asking the individual to move both limbs reflects the motion in the unaffected limb, providing visual feedback implying full movements in both arms. A specific mechanism for this feedback therapy is the activation of the mirror neuron system (MNS) [5-7]. The MNS is a theoretical compilation of neurons that are used when an individual performs a specific action. Most importantly, the system can be activated when simply observing a task being completed, particularly from a first-person view [8-10]. For example, mirror therapy is designed to facilitate neuroplasticity in motor areas affected by a stroke through visual presentation of actions that increase activity in the MNS [7,9]. However, since mirror therapy generally requires mass repetition of a simple task over a long period, individuals may struggle with maintaining adherence to the therapy prescription [5].

Game-based rehabilitation therapies are recognized for fostering high engagement among participants [11], and these interventions have been transformed by innovations in virtual reality (VR) technology [5]. VR is “a computer-generated environment in which people can feel and interact with various situations in three dimensions” [5]. There have been studies that have investigated the plausibility and effectiveness of combining MNS-based training with VR systems to create functional gains in upper extremity function among stroke survivors [5,9]. One study created a mirror-based VR rehabilitation system that provided unilateral and bilateral limb mirroring exercises in a fully immersive virtual environment [12]. After that intervention, the participants demonstrated significant improvements in their Fugl-Meyer Upper Extremity Assessment (FM-UE) scores. One of the measures on the FM-UE is proprioception, defined as the ability to perceive or be aware of the position and movement of the body [13]. In the case of VR mirror therapy for arm abduction, proprioception in the shoulder joint could be perceptually recalibrated in order to convince people post stroke that their motions are larger (or even smaller) than in reality [4]. Nevertheless, it is important to note that the precise mechanism (ie, MNS) underlying improvements following VR mirror therapy is still unclear. Meta-analysis has demonstrated that Mu suppression, an outcome often used to assess the effects of mirror therapy, was a valid indicator of the MNS [14]. However, recent investigations suggest the strength of Mu suppression as a mechanism could be confounded by other factors [15,16].

VR mirror therapy research for improving poststroke upper-arm hemiparesis appears promising but has been limited by complex setups, costly equipment, and a supplementation of nonspecified physical or occupational therapy treatment, hindering replicability and generalizability [17]. A recent meta-analysis found that VR mirror therapy was more effective than control groups [17]. For example, one study by Jo et al [18] demonstrated that a 360-degree immersive VR intervention was more effective than conventional mirror therapy versus a control group. However, the meta-analysis called for further investigations with larger randomized controlled trials, because the review included only 5 randomized controlled trials, totaling 148 people post stroke (a sample size of 30 people per study). To conduct large-scale trials, VR interventions will require easy-to-use methodologies that can facilitate adoption across a large variety of settings.

The goal of this study was to evaluate the feasibility of a low-cost, mirror neuron–based VR gaming protocol to improve upper extremity function in a person with chronic stroke. The protocol combined “off-the-shelf” supplies and software with an intervention dose that matched the average prescription given in mirror therapy. The VR headset included the Meta Quest 2, an all-in-one headset with built-in tracking, and a rhythmic movement-to-music game (Beat Saber), along with a computer and freely downloadable recalibration software (WalkinVR) to adapt conventional gaming for people with physical disabilities. The ease of obtaining study supplies (obtainable in any major electronics retailer) makes the intervention protocol capable of being easily implemented in various clinical and community settings.

Study purposes:

1. The primary purpose of this study was to examine the potential benefits of a short, intensive VR-adaptive therapeutic gaming intervention on hemiparetic stroke function and range of motion in an individual post stroke at midintervention and postintervention.
2. The second purpose of this study was to explore the feasibility of the study through intervention adherence, safety, and technical issues.
3. The third aim was to qualitatively describe the participant’s satisfaction with the program (acceptability of the program) through perceptions of intervention benefits, likes and dislikes, and recommendations for improvement.

Methods

Design and Overview

This case study was a pre-post trial design for a single participant. The intervention and data collection were conducted on-site at a university laboratory. The study lasted 4 weeks, which totaled 20 days of intervention. The study included both quantitative and qualitative procedures, analyzed separately (ie, not mixed methods). The study included quantitative outcomes that were measured before (day 0), in the middle (day 10), and after the 20-day intervention (total duration of 4 weeks, with intervention conducted only on weekdays). The study also included a qualitative postintervention interview. This case study followed the CARE (Case Report) guidelines while incorporating selected Single-Case Reporting Guideline in Behavioural Interventions (SCRIBE) elements relevant to behavioral intervention research (eg, description of context, participant flow, and intervention fidelity). This hybrid approach

XSL•FO
RenderX

was used to maximize transparency and ethics (no multiple baseline or follow-up to minimize participant burden and no withdrawal phase to test a washout effect on outcomes).

Participants

The participant was chosen based on the following inclusion criteria: (1) aged ≥18 years, (2) chronic stroke (time since onset ≥6 months), (3) hemiparesis with a minimal degree of functional ability of the affected arm (manual muscle test score ≥2), (4) access to transportation to the laboratory, and (5) ability to communicate in English. Exclusion criteria included (1) having a health condition that prevented participation in moderate-intensity exercise and (2) susceptibility to VR-induced motion sickness. The participant was recruited via word of mouth through a rehabilitation research collaborative at the university. The rationale for including only a single participant was due to time and effort constraints, since the project needed to be completed within a university semester.

Equipment and Setup

Study equipment included a Meta Quest 2 that was connected to a desktop gaming computer through a high-speed USB cable (Meta Link Cable). The gaming computer included VR recalibration software, WalkinVR, through the Steam gaming platform. WalkinVR is a free-to-download application on Steam that allows VR gaming to be fully adapted for a person with a disability. When used for research, WalkinVR requires a fee for use and technical support. WalkinVR features include, but are not limited to, physical boosts or projections of small movements in real life to big movements in the VR environment and repositioning of the headset or handheld controllers in the digital space. Additional software included the VR game Beat Saber, which was installed on the desktop computer. The participant wore the Quest 2 while seated in an armless chair (shown in Figure 1).

Figure 1. Demonstration of the participant post stroke playing Beat Saber.

XSL•FO
RenderX

Intervention

The intervention prescription consisted of active gaming for 60 minutes per day, 4 times a week, for a total of 5 weeks. The intensity (ie, difficulty) of the gameplay ranged from "easy" to "hard." The intensity and the level were selected by the participant. At each intervention visit, 2 research staff members helped the participant equip the VR headset, set up the gaming system, and supervise the exercise session. Research staff also provided occasional verbal praise for successful task completion and encouraged the participant to minimize downtime between completing one level and selecting another to play. On the first day of exercise, research staff calibrated the WalkinVR software for the needs of the participant. Calibration included 3 steps. First, the participant was instructed to maintain a neutral arm position, with elbow joints held to the side of their body, in a comfortable resting position, with the forearm pointed forward as straight as possible. Second, research staff used the WalkinVR software on the computer to adjust the controller position of the affected arm to match the position of the unaffected arm in the digital VR environment. Third, the participant was asked to play an easy level in Beat Saber. During calibration, if the participant had insufficient reach in the game to successfully complete the box-chopping tasks, real-time adjustments of controller position were made in WalkinVR until successful gameplay was achieved. As a means of intervention progression, positional assistance in the affected arm from WalkinVR was gradually reduced over time. Research staff were instructed to make minor reductions in positional assistance every few days, without making box-chopping too difficult (ensuring that the participant is chopping blocks with at least a 70% success rate). To avoid boredom from repeating the same levels that were provided in the base version of the game, the participant was allowed to request that the staff purchase additional levels in Beat Saber (song packs).

Measures and Analyses

Quantitative

Quantitative data were collected on-site at preintervention (day 0), midintervention (day 10), and postintervention (after day 20). Primary aim outcomes included (1) the Wolf Motor Function Test (WMFT) and (2) the Disabilities of the Arm, Shoulder, and Hand Questionnaire (DASH). Secondary outcomes were the upper extremity range of motion of the hemiparetic side as measured by goniometry. The WMFT includes 15 tasks, and participants are instructed to complete tasks as quickly as possible. The first 6 tasks of the WMFT include timed joint-segment movements, whereas tasks 7 to 15 include integrative functional movements. For this study, task 7 (the maximum weight a person can lift onto the table) and task 14 (grip strength) were excluded because musculoskeletal strength measures were not relevant to the study goals. The WMFT provides a total score based on average task completion time, capped at 120 seconds per task. It can also be rated on a functional ability scale from 0 to 5—"0" means the participant "does not attempt with the involved arm," and "5" means the "arm does participate and the movement appears to be normal." Primary and secondary outcome measures were conducted by an occupational therapy graduate student, who was instructed by a senior occupational therapist. Functional ability scores were rated by a senior therapist not involved in the intervention.

Because this study involved a single participant, quantitative data were analyzed descriptively. Means, SDs, percent change, and standardized difference scores (effect size) were calculated to describe the magnitude of change across time points. The standardized difference was computed as the mean change divided by the pooled SD between 2 time points. This approach was used only to provide a descriptive indication of change, rather than for statistical inference. Minimal clinically important difference (MCID) and minimal detectable change (MDC) values from prior validation studies were reported alongside the descriptive metrics to support interpretation of the clinical relevance of observed changes.

Tertiary feasibility outcomes included attendance, technology issues, and unanticipated problems. Attendance was defined as the number of sessions attended and completed divided by the number of prescribed sessions. An arbitrary value of 70% attendance was considered "acceptable." Technology issues that were encountered by the participant during each session were documented by research staff. Problems and adverse events were recorded and reported in accordance with university procedures.

Qualitative

After completion of the intervention, the participant underwent a semistructured interview (in-person) with the lead interventionist (SH) at the end of the program. The qualitative component of this study was an intrinsic case study design [19,20] to explore the lived experience of a single participant who completed the program. The specific philosophical assumptions that underpinned the qualitative study methods were critical realism, from the ontological perspective, and interpretivism, from the epistemological perspective [21]. These assumptions allowed the researchers to recognize that the participant will perceive a reality when reporting responses, but this reality could be subtly enhanced by the interaction between participant and interviewer and the act of interpreting the data. Underpinned by an interpretivist epistemology, our aim was to understand how the participant makes meaning of exercising in VR post stroke, rather than to quantify frequency of codes or seek a single "true" account. The interview included 10 overarching questions with several follow-up questions. The questions probed participants' overall views of the program, likes and dislikes, experiences using the VR equipment for upper extremity exercise, comparison with previous therapy experience, barriers and facilitators to meeting the intervention prescription, preferences, and recommendations to enhance the protocol for a future trial. The interviewer was trained by a qualitative expert in disability and exercise (BL), who had conducted more than 400 interviews related to exercise and disability. The interview was audio recorded and transcribed for analysis by a member of the research team.

Qualitative analysis was framed by Braun and Clarke's [22,23] 6-step latent thematic analysis approach. First, the analyst familiarized himself with the data by reading the transcript multiple times and noting early impressions and reflexive insights about tone, emotion, and context. Second, initial codes

XSL•FO
RenderX

were generated using artificial intelligence (ChatGPT) on a segment-by-segment basis to identify relevant units of meaning to describe actions, perceptions, or feelings. Third, codes were aggregated and clustered using ChatGPT to form one level of overarching categories (ie, themes). The rationale for including only one level of themes was due to only having one interview; an attempt to provide fewer, more concentrated themes [24] rather than abundant but watered-down themes. Fourth, the qualitative expert analyst cross-checked and revised the resultant codes and themes with the case transcription for accuracy (ie, resultant codes similar to what the analyst would have generated) and relevancy (ie, ensuring that codes and themes aligned with philosophical assumptions and the study objective). Fifth, themes were revised again to more clearly define the central concept of the supporting codes, and additional participant quotes were sought to strengthen concepts. Quotations and codes were selected for their richness and relevance, rather than frequency. Sixth, the results were presented in an interpretive commentary with participant quotations to support the concept. To enhance rigor throughout this process, reflexivity was maintained through memoing and keeping an audit trail for iterations of themes and codes. The transcription was performed by ChatGPT and double-checked for accuracy by an author.

Ethical Considerations

This study was conducted in accordance with case study guidelines set by the Institutional Review Board for Human Use at the University of Alabama at Birmingham, and approval for the study was granted (IRB-300010445). Written informed consent was obtained from the participant prior to participation. Identifying information (eg, name, initials, hospital numbers) of the case participant was omitted from publication, and data were deidentified and stored using an alpha numeric code.The participant was compensated at the end of the study with a US $50 gift card.

Results

The participant, referred to as “Chop,” was a 72-year-old White male, 5’11” tall and weighing 220 lb. He was 3 years post stroke and used a cane as a primary mobility device and a power wheelchair during prolonged daily activities. He had hemiparesis affecting his left, nondominant side. He had no cognitive, visual, or hearing impairment preventing him from completing the intervention.

Feasibility Metrics (Adherence, Safety and Technical Issues)

Chop attended and completed 20 intervention sessions (20/20, 100% attendance), with only one additional song pack purchase in Beat Saber. The participant rescheduled 5 sessions (5/20, 25%). No adverse events or problems were experienced by the participant during the intervention. As for technical issues, the participant reported the following equipment-related problems: (1) the VR headset had a slight weight, exerting mild pressure on the participant’s face, despite using a head-strap accessory, and (2) the controller was challenging to grasp during exercises for the nondominant (affected) hand, even with the hand-knuckle strap accessory.

Wolf Motor Function Test

Figures 2 and 3 display the results for the average time to complete tasks and the functional ability score for the WMFT, respectively. The results for each task of the WMFT are shown in Table 1 (time to complete) and Table 2 (task score). Results demonstrated that the average time to complete tasks improved by 70.5% (δ=11.73 s; preintervention mean time 16.63, SD 31 s; effect size=0.54) from preintervention to midintervention and 78% (δ=12.96 s; effect size=0.59) from preintervention to postintervention. A 25% (δ=1.23 s; midintervention mean time 4.9, SD 4 s; effect size=0.33) improvement was observed from midintervention to postintervention. All of these changes exceeded an $MDC_{95\%}$ of 0.7 seconds [25], as well as an MCID of small effect. Results for functional ability score demonstrated improvements of 18% (δ=0.46 points; preintervention mean score 2.67, SD 0.82 points; effect size=0.59) from preintervention to midintervention and 23% (δ=0.6 points; effect size=0.79) from preintervention to postintervention. A 4% (δ=0.14 points; midintervention mean score 3.13, SD 0.74 points; effect size=0.19) improvement was observed from midintervention to postintervention. All of these changes surpassed an $MDC_{95\%}$ of 0.1 points [25]. Changes from preintervention to midintervention and preintervention to postintervention exceeded an MCID of small effect, but the change from midintervention to postintervention did not.

XSL•FO
RenderX

Figure 2. Average time to complete tasks for the Wolf Motor Function Test (WMFT).

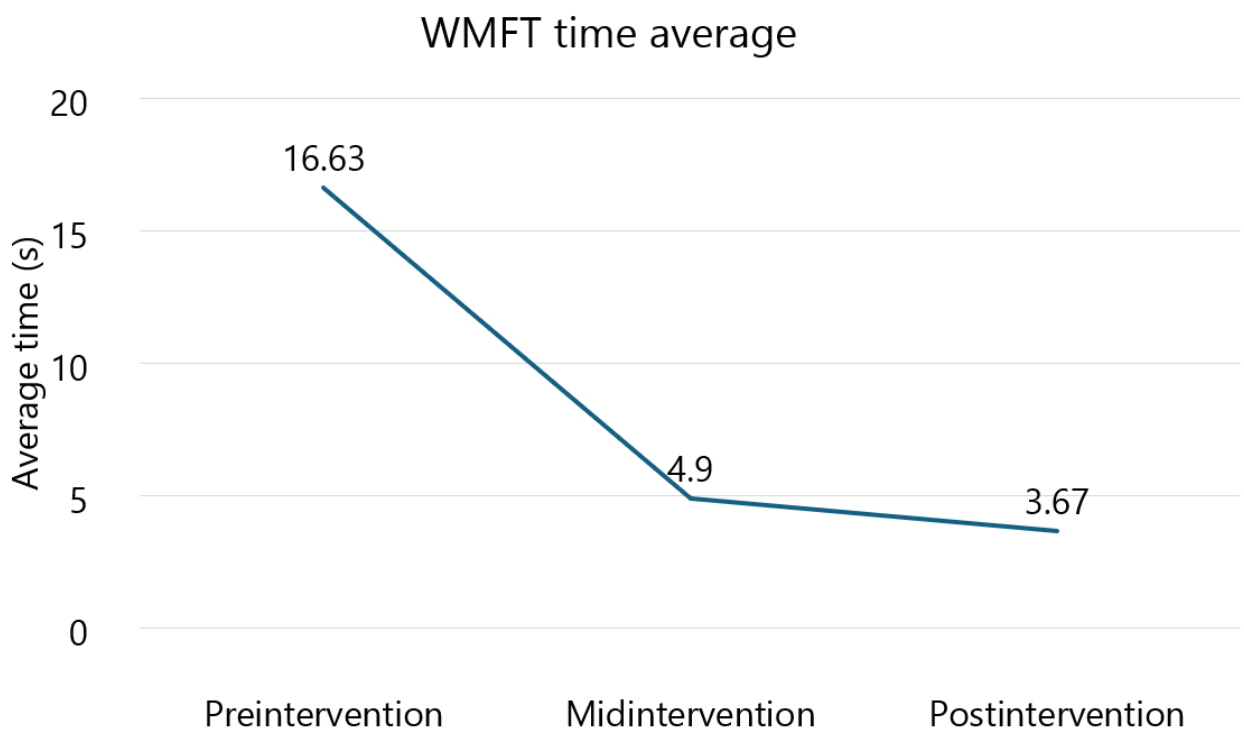

Figure 3. Average Wolf Motor Function Test (WMFT) functional ability score.

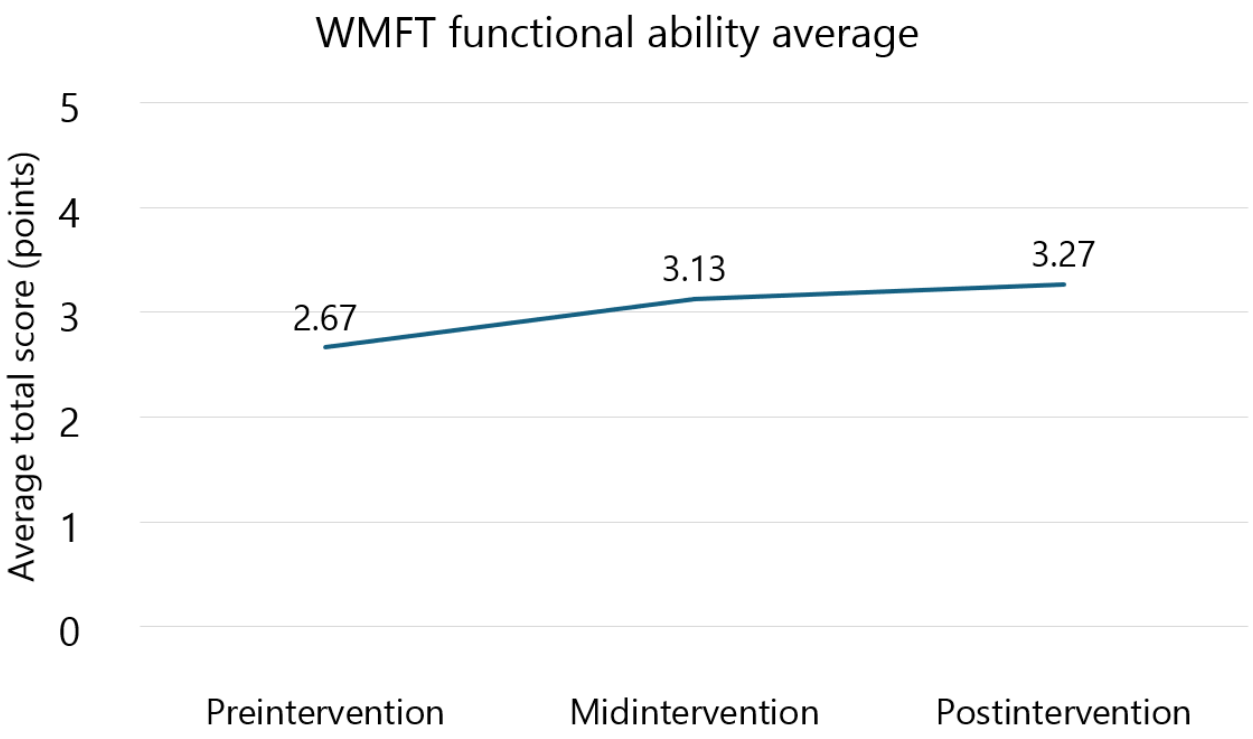

XSL•FO
RenderX

Table 1. Wolf Motor Function Test time.

Item	Preintervention (seconds)	Midpoint (seconds)	Postintervention (seconds)
1	2.75	1.25	1.11
2	17.56	4.78	1.34
3	120	12.31	14.03
4	43.16	9.2	2.88
5	1.38	0.84	0.79
6	2.34	1.82	1.37
7	N/A[a]	N/A	N/A
8	1.09	1.03	0.85
9	2.78	2.68	2.09
10	2.91	1.71	1.88
11	4.19	2.03	1.87
12	20.34	5.87	4.87
13	9	7.88	6.25
14	N/A	N/A	N/A
15	6	6.25	5.16
16	10.79	11.6	8.22
17	5.1	4.25	2.38
Mean (SD)	16.63 (30.7)	4.9 (3.9)	3.67 (3.6)

[a]N/A: not applicable.

Table 2. Wolf Motor Function Test score.

Item	Preintervention	Midpoint	Postintervention
1	3	3	3
2	2	3	3
3	1	2	2
4	2	2	3
5	3	4	4
6	3	4	4
7	N/A[a]	N/A	N/A
8	2	3	3
9	4	3	4
10	3	4	4
11	3	4	4
12	2	4	4
13	3	3	2
14	N/A	N/A	N/A
15	4	3	3
16	3	2	3
17	2	3	3
Mean (SD)	2.67 (0.82)	3.13 (0.74)	3.27 (0.70)

[a]N/A: not applicable.

XSL•FO
RenderX

Disabilities of the Arm, Shoulder, and Hand Questionnaire

Results demonstrated that the total DASH score, from preintervention (41.67 points) to midintervention (36.68 points), improved by 11.98% (δ=4.99 points; effect size=0.23) and 10% (δ=4.17 points; effect size=0.18) from preintervention to postintervention (37.5 points). From midintervention to postintervention, a change of 2.25% (δ=0.83 points; effect size=0.03) was observed. None of these changes surpassed an $MDC_{95\%}$ of 12.75 total score points [26].

Range of Motion

The results for Chop's affected arms' range of motion preintervention and postintervention can be found in Table 3. Results demonstrated that range of motion improved in all joints by an average of 35.67% (SD 20%). Favorable changes in range of motion (ie, improvements toward the normal range of motion for a given joint) were observed in all joints. Notable changes were observed for shoulder flexion (67%), elbow flexion (50%), hand supination (40%), and elbow extension (35%). Minimal improvements were observed for arm abduction (9%). Hand pronation improved minimally (10%), but a normal range of 90 degrees was achieved at postintervention.

Table 3. Range of motion for hemiparetic arm joints.

	Preintervention (degree)	Postintervention (degree)	Change, degree (%)
Left shoulder flexion	55	92	37 (67)
Left elbow extension	40	26	–14 (–35)[a]
Left elbow flexion	80	120	40 (50)
Left arm abduction	110	120	10 (9)
Left hand pronation	80	90	10 (13)
Left hand supination	50	70	20 (40)

[a]Negative values indicate a reduction in joint range of motion from preintervention to postintervention.

Qualitative Results

Case Description

Chop was a 72-year-old White male living in a suburban area in Alabama, United States. Chop used a standard cane as a primary means of mobility and used a power wheelchair when traveling in new or high-activity areas. Chop did not participate in any exercise before this study began. He completed inpatient occupational and physical therapy immediately following his stroke and outpatient therapy periodically during the first year post stroke. In the year before this study, Chop did not participate in formal therapy, feeling he had reached a plateau in functional ability. As a grandparent, he spent much of his time with and entertained his grandchildren. He saw himself as a handyman, enjoying fixing things, and spent his free time working on his car and fixing appliances. After reaching a functional plateau, Chop relied on compensatory strategies and adaptive equipment for daily tasks like dressing, bathing, toileting, socializing, and driving. Qualitative analysis resulted in 4 themes that are displayed in Table 4 and described below.

XSL•FO
RenderX

Table 4. Qualitative results.

Theme	Codes	Supporting quotes
Curiosity into confidence: "Not sure what it was, but it was exciting"	• Initial curiosity creates feelings of excitement • Confidence is questionable prior to participation • Early physical exertion leads to thoughts of withdrawal • Persistence was supported by enjoyment with the program and his commitment to the study team • Confidence bolstered by 2 weeks of successful repetition	• "I really wasn't sure. I think I didn't really know for sure what it was... I mean, it was exciting." • "I think I was confident I could do it. I just wasn't sure... Like anything, I think I can do them all. Everything I think I can do, but I can't always do it." • "It did surprise me a little. It surprised me that I was doing that. Really, you know, somebody don't really play them kind of games and stuff, you know? That was surprising, but yeah, I think it's pretty good." • Interviewer: Was it difficult to do 60 minutes of it four days a week? • "No, it wasn't difficult, you know. It just, no, not really. I mean, I guess I did get a little tired." • "[first week of the program] I said, I ain't going back to see these girls I've had enough of. I got three more weeks left? No, I mean, I just thought, this is wild. This is wild to me, you know... I ain't played a lot of these games, you know. But it was fun."
Redefining rehabilitation through play: "It wasn't boring like therapy"	• Gamification transforms past perceptions of therapy • Virtual reality is innovative for its higher cognitive demand and difficulty	• "I guess it [conventional therapy] worked, but you know, it's always boring. Yeah. You know, and I mean, comparing the two [conventional versus virtual reality], it's just like night and day... I think it causes you to have more mental reactions to help you move." • "Well, I mean, I liked it all... I mean, I just enjoyed it. I didn't mind going over and doing it. I mean, it was a lot of fun."
Reclaiming independence: "I could tie my boots again"	• Improved function and range of motion of the affected arm • Functional improvements fostered benefits to daily activities • Newly found modifications for daily activities • Early improvements	• "I was putting on my socks a lot easier than I was." • "I put on some boots... Well, half boots. Laced them and actually tied them... It's been a while since I put on anything. Most of my stuff is all slip-on stuff now." • "After that first week, I might have tested a little more movement. I would notice things. I'm doing this [demonstrates movement], where it normally would have been hard to do." • "One thing I did get out of... Getting out of the bed, when I'd get over on my side, I would push myself up more with this arm than I have been doing. • "Seems like I have been doing better with this arm [affected arm] on things I'm working on, where normally it's just running all over the place. You know, be still while you do this."

Theme 1: Curiosity Into Confidence

"Not sure what it was, but it was exciting." Chop began the program with initial curiosity and uncertainty, which gradually progressed to growing confidence and engagement. Before starting the program, he perceived VR to be mysterious and exciting, a novelty period, which drove him to pursue the program despite internal struggles. However, initial excitement was coupled with internal uncertainty regarding his confidence to successfully engage in the program using his affected arm. Uncertainty was also affected by a lack of experience playing video games, and this uncertainty lingered through the first week of the program. Moreover, the first week of the program brought additional internal tension between the desire to persist and physical fatigue from the exercise. Chop described early physical exertion, particularly related to his affected arm, which led to thoughts of withdrawing from the program: "I said, I ain't going back to see these girls. I've had enough of it." These thoughts were outweighed by the enjoyment of the program, his perceived commitment to the study team, and growing confidence after 2 weeks of successful completion of his exercises. As repetition accumulated over the first 2 weeks, the participant reported improved confidence and described surprise at his own competence and adaptation to the virtual environment: "It did surprise me a little. It surprised me that I was doing that. Really, you know, somebody don't really play them kind of games and stuff, you know?"

Theme 2: Redefining Rehabilitation Through Play

"It wasn't boring like therapy." This theme reveals how Chop's understanding of therapy shifted from monotonous, clinical exercise to stimulating and playful engagement. He framed conventional therapy as "boring" and repetitive, but perceived VR as "like night and day." The gamified environment demanded higher cognitive processing—reacting to visual and auditory cues, anticipating movement, and achieving performance feedback. This experience transformed his perception of rehabilitation from a medical obligation into a mentally engaging activity. When asked about conventional therapy, he explained:

"Well, I guess it [conventional therapy] worked but, you know, it was always boring. If I compare the therapy I've had to this

XSL•FO
RenderX

VR kind, it's like night and day. I think it [VR] causes you to have more of a mental reaction to help you move, move of a natural reaction rather than a repetitive and conscious effort to try to do something difficult."

Theme 3: Reclaiming Independence

"I could tie my boots again." Chop reported several improvements in the functional ability of his affected arm, which he noticed while performing daily tasks. Improvements included range of motion and control of his affected arm, which would previously be "running all over the place." He reported accomplishing tasks he hadn't been able to do since the stroke onset, such as independently tying his shoes with laces, easier bed mobility, working on his car, and dressing. After being able to tie his laced boots for the first time in years, the participant told his wife, "Look here, I put these on and tied them!" Additionally, he discovered new and efficient modifications to daily activities: "I would push myself up more with this [affected] arm than I have been doing." Chop reported that he could do more at home and that those benefits were observed as early as the second week of the intervention.

Theme 4: Motivation Through Flow and Mastery

"It was a challenge, and that was good." Chop repeatedly described VR exercise as an enjoyable challenge that kept him engaged. Successful gameplay demanded constant visual and cognitive attention to quickly react to in-game tasks, creating a mental state of performance without thought (referred to as "flow"): "It was happening so fast you didn't have time to think about it, you know? You'd just done it." Chop's full attention was placed on gameplay performance, and this was supported by visual immersion in a virtual world and a high level of enjoyment in gameplay. These components distracted Chop from real-world movements of his affected arm, which allowed him to push past his limits:

"I feel like it was more helpful that I couldn't see my arm in real life because it makes you feel like you're moving your arm all the way. I think it's because I could do a little bit more with it based on what I was seeing on the screen. It sort of tricks your brain."

Chop reported that gameplay was difficult due to the physical fatigue of his affected arm. His main challenge was achieving complete accuracy in hitting all in-game boxes. His primary goal was to enhance his accuracy to hit blocks and maintain a high streak over time (repeated successful hits on blocks without failing). Chop explained why maintaining accuracy was the biggest challenge: "getting my arm up high enough and keep from getting so tired after playing for a while and after moving for so long. I never really move my left arm that much, especially not all in one hour." Instead of deterring participation, gameplay difficulty was met by Chop's firm resolve to persevere and master the challenges. Perseverance was supported by functional improvements he observed throughout the program, where Chop would engage in motivational self-talk: "I felt like even after the end of two weeks I was getting more function in my left arm. Each week I would tell myself that I could do better than the week before."

Discussion

Overview and Principal Results

This study investigated the preliminary effect of a 20-day VR program to improve the affected upper limb function of a person post stroke with hemiparesis. The novel aspect of this study was the inclusion of adaptive VR software (free to download) to simulate a mirror therapy intervention, as well as the inclusion of low-cost technology that can be purchased at any major electronics retailer.

Quantitative findings demonstrated moderate, clinically meaningful improvements in the functional ability of the participant as measured by the WMFT, as well as a small, likely nonmeaningful change in self-reported functional ability as measured by the DASH. For the WMFT, a remarkable 70.5% (δ=11.73 s; preintervention mean time 16.63, SD 31 s) improvement in task-affected arm performance was observed at the midpoint of the intervention, approximately 10 days of intervention. Without a control group and a larger sample, it is difficult to compare the absolute improvement change to previous investigations. Nevertheless, moderate effects indicated by the WMFT were comparable to the moderate effects observed from conventional mirror therapy interventions [3]. Similar interpretations can be made for the WMFT functional ability findings [27-29]. WMFT improvements for this study appeared lower than the strong effect (effect size=0.84) on upper extremity function, measured by the FM-UE, which was reported in a meta-analysis of 4 other VR mirror therapy interventions [17]. These differences could potentially be explained by a few factors. First, past studies often supplemented VR intervention with physical or occupational therapy. Second, past VR interventions often conducted task-specific practice using specialized games that targeted upper extremity function, as opposed to using an "off-the-shelf" game that was not designed for therapeutic purposes. Third, differences could be due to the different measures of upper extremity function that were used (WMFT vs FM-UE). Fourth, previous interventions used different VR systems (eg, addition of specialized controllers to a head-mounted display or interactive television applications). Fifth, the participant in this study was not in the acute stage of recovery (<2 years post stroke), which could potentially dampen gains in physical function.

Of note, this study found improvements in range of motion, which indicated that WMFT improvements stemmed from the shoulder, elbow, and forearm joints. Interestingly, improvements in performance task time and functional ability score seemed much smaller from midintervention to postintervention. Collectively, these findings indicate that the intervention has a potentially moderate strength effect on affected arm functional ability at an even shorter intervention duration than originally planned. A moderate effect can likely be experienced following a brief 10-day intervention. A shorter intervention duration has obvious benefits, including reduced cost and burden, but this warrants investigation in a randomized controlled efficacy trial.

The qualitative findings expanded upon quantitative findings and, most importantly, identified potential mechanisms that led to the success of the intervention. The participant reported a

XSL•FO
RenderX

physiological burden from the intervention, which was overcome by a few factors. The participant reported greater engagement in the exercise movements of the intervention compared to what he had previously experienced in conventional therapy. This was due to immersion within a digital environment, the enjoyment of the game box–chopping activity, the accompanied music, and the performance feedback reward system that put him in a state of "flow." Moreover, the participant reported noticeable gains in the function of the affected arm, which led to favorable changes in his daily activities.

While these findings are encouraging, a notable limitation of VR mirror-simulated exergaming is that it requires moderate function of the affected arm to complete the in-game tasks successfully. Whereas conventional mirror therapy is likely beneficial when impairment of the affected arm is severe, and simply observing the nonaffected arm can stimulate small movements in the affected arm. Moreover, the study findings had numerous limitations, which warrant further investigation in a randomized controlled trial.

Limitations

This proof-of-concept case study had limitations. First, the study included only a single case, which impedes the findings from any form of representativeness or generalizability. Second, the outcome assessor was not blinded, which could have affected the results. Third, the study did not have a control group. Including a control group or other kinds of control would help determine whether the observed improvements were due to the intervention itself rather than natural recovery, motivation, or other external factors. Fourth, this study did not include a repeated baseline design to allow stabilization of baseline outcomes prior to testing effects. Fifth, this study did not include a washout phase to test either outcome sustainability or a repeated intervention effect. Therefore, study findings should be interpreted with caution.

Conclusion

This study identified a low-resource protocol for VR exergaming with adaptive software to improve the hemiparetic arm function of people post stroke. The identification of an easy-to-adopt protocol could help address a key limitation in VR mirror therapy research: small sample sizes and nonconfirmatory randomized clinical trials. Although favorable findings were observed in this case study, these findings need to be further explored in a pilot randomized controlled trial before larger investigations into efficacy or effectiveness are warranted.

Acknowledgments

Laboratory space and supplies were provided by the Center for Engagement in Disability Health and Rehabilitation Sciences at the University of Alabama at Birmingham.

The authors declare the use of generative artificial intelligence (GAI) in the research and writing process. According to the Generative AI Delegation Taxonomy (2025), data analysis was delegated to GAI tools under full human supervision. The GAI tool used was ChatGPT 5.0. Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes. The declaration was submitted by BL.

Data Availability

All data generated or analyzed during this study are included in this published paper.

Conflicts of Interest

None declared.

References

1. Langhorne P, Coupar F, Pollock A. Motor recovery after stroke: a systematic review. Lancet Neurol 2009;8(8):741-754. [doi: 10.1016/S1474-4422(09)70150-4] [Medline: 19608100]
2. Nijland RH, van Wegen EE, Harmeling-van der Wel BC, Kwakkel G, EPOS Investigators. Presence of finger extension and shoulder abduction within 72 hours after stroke predicts functional recovery: early prediction of functional outcome after stroke: the EPOS cohort study. Stroke 2010;41(4):745-750. [doi: 10.1161/STROKEAHA.109.572065] [Medline: 20167916]
3. Thieme H, Morkisch N, Mehrholz J, Pohl M, Behrens J, Borgetto B, et al. Mirror therapy for improving motor function after stroke. Cochrane Database Syst Rev 2018;7(7):CD008449 [FREE Full text] [doi: 10.1002/14651858.CD008449.pub3] [Medline: 29993119]
4. Spiss S, Siess M, Kim Y, Harders M. Effect of touch stimuli on proprioceptive recalibration during upper-limb rotation in virtual reality mirror therapy. 2018 Presented at: 7th IEEE International Conference on Biomedical Robotics and Biomechatronics (Biorob); 2018 August 26; the Netherlands. [doi: 10.1109/biorob.2018.8488133]
5. Bao X, Mao Y, Lin Q, Qiu Y, Chen S, Li L, et al. Mechanism of Kinect-based virtual reality training for motor functional recovery of upper limbs after subacute stroke. Neural Regen Res 2013;8(31):2904-2913 [FREE Full text] [doi: 10.3969/j.issn.1673-5374.2013.31.003] [Medline: 25206611]

XSL•FO
RenderX

6. Lee M, Cho HY, Song CH. The mirror therapy program enhances upper-limb motor recovery and motor function in acute stroke patients. Am J Phys Med Rehabil 2012;91(8):689-96, quiz 697. [doi: 10.1097/PHM.0b013e31824fa86d] [Medline: 22469877]
7. Chang WK, Lim H, Park SH, Lim C, Paik N, Kim W, et al. Effect of immersive virtual mirror visual feedback on Mu suppression and coherence in motor and parietal cortex in stroke. Sci Rep 2023;13(1):12514 [FREE Full text] [doi: 10.1038/s41598-023-38749-8] [Medline: 37532803]
8. Rizzolatti G, Craighero L. The mirror-neuron system. Annu Rev Neurosci 2004;27:169-192. [doi: 10.1146/annurev.neuro.27.070203.144230] [Medline: 15217330]
9. Modroño C, Navarrete G, Rodríguez-Hernández AF, González-Mora JL. Activation of the human mirror neuron system during the observation of the manipulation of virtual tools in the absence of a visible effector limb. Neurosci Lett 2013;555:220-224. [doi: 10.1016/j.neulet.2013.09.044] [Medline: 24080372]
10. Matamala-Gomez M, Frisco F, Guidali G, Lega C, Beacco A, Bolognini N, et al. Virtual body continuity during action observation affects motor cortical excitability. Sci Rep 2025;15(1):13364 [FREE Full text] [doi: 10.1038/s41598-025-97695-9] [Medline: 40247078]
11. Hung Y, Huang PC, Chen KT, Chu WC. What do stroke patients look for in game-based rehabilitation: a survey study. Medicine (Baltimore) 2016;95(11):e3032 [FREE Full text] [doi: 10.1097/MD.0000000000003032] [Medline: 26986120]
12. Mekbib DB, Zhao Z, Wang J, Xu B, Zhang L, Cheng R, et al. Proactive motor functional recovery following immersive virtual reality-based limb mirroring therapy in patients with subacute stroke. Neurotherapeutics 2020;17(4):1919-1930 [FREE Full text] [doi: 10.1007/s13311-020-00882-x] [Medline: 32671578]
13. Kiper P, Baba A, Agostini M, Turolla A. Proprioceptive Based Training for stroke recovery. Proposal of new treatment modality for rehabilitation of upper limb in neurological diseases. Arch Physiother 2015;5:6 [FREE Full text] [doi: 10.1186/s40945-015-0007-8] [Medline: 29340175]
14. Fox NA, Bakermans-Kranenburg MJ, Yoo KH, Bowman LC, Cannon EN, Vanderwert RE, et al. Assessing human mirror activity with EEG mu rhythm: a meta-analysis. Psychol Bull 2016;142(3):291-313 [FREE Full text] [doi: 10.1037/bul0000031] [Medline: 26689088]
15. Hobson HM, Bishop DVM. Mu suppression--a good measure of the human mirror neuron system? Cortex 2016;82:290-310 [FREE Full text] [doi: 10.1016/j.cortex.2016.03.019] [Medline: 27180217]
16. Köster M, Meyer M. Down and up! Does the mu rhythm index a gating mechanism in the developing motor system? Dev Cogn Neurosci 2023;60:101239 [FREE Full text] [doi: 10.1016/j.dcn.2023.101239] [Medline: 37030147]
17. Okamura R, Nakashima A, Moriuchi T, Fujiwara K, Ohno K, Higashi T, et al. Effects of a virtual reality-based mirror therapy system on upper extremity rehabilitation after stroke: a systematic review and meta-analysis of randomized controlled trials. Front Neurol 2023;14:1298291 [FREE Full text] [doi: 10.3389/fneur.2023.1298291] [Medline: 38259644]
18. Jo S, Jang H, Kim H, Song C. 360° immersive virtual reality-based mirror therapy for upper extremity function and satisfaction among stroke patients: a randomized controlled trial. Eur J Phys Rehabil Med 2024;60(2):207-215 [FREE Full text] [doi: 10.23736/S1973-9087.24.08275-3] [Medline: 38483333]
19. Tisdell E, Merriam SB, Stuckey-Peyrot HL. Stuckey-Peyrot, Qualitative Research: A Guide to Design and Implementation. Hoboken, New Jersey: John Wiley & Sons; 2025.
20. Stake R. Case Study Research. Cham: Springer; 1995.
21. Fletcher AJ. Applying critical realism in qualitative research: methodology meets method. Int J Soc Res Methodol 2017;20(2):181-194. [doi: 10.1080/13645579.2016.1144401]
22. Braun V, Clarke V, Weate P. Using thematic analysis in sport and exercise research. In: Smith B, Sparkes AC, editors. Routledge Handbook of Qualitative Research in Sport and Exercise. London: Routlegde; 2016:191-2025.
23. Braun V, Clarke V. Thematic Analysis. Washington, D.C: American Psychological Association; 2012.
24. O'Reilly M, Parker N. 'Unsatisfactory saturation': a critical exploration of the notion of saturated sample sizes in qualitative research. Qual Res 2012;13(2):190-197. [doi: 10.1177/1468794112446106]
25. Fritz SL, Blanton S, Uswatte G, Taub E, Wolf SL. Minimal detectable change scores for the Wolf Motor Function Test. Neurorehabil Neural Repair 2009;23(7):662-667. [doi: 10.1177/1545968309335975] [Medline: 19498013]
26. Beaton DE, Katz JN, Fossel AH, Wright JG, Tarasuk V, Bombardier C. Measuring the whole or the parts? Validity, reliability, and responsiveness of the disabilities of the arm, shoulder and hand outcome measure in different regions of the upper extremity. J Hand Ther 2001;14(2):128-146. [Medline: 11382253]
27. Chan WC, Au-Yeung SS. Recovery in the severely impaired arm post-stroke after mirror therapy: a randomized controlled study. Am J Phys Med Rehabil 2018;97(8):572-577. [doi: 10.1097/PHM.0000000000000919] [Medline: 29521681]
28. Colomer C, NOé E, Llorens R. Mirror therapy in chronic stroke survivors with severely impaired upper limb function: a randomized controlled trial. Eur J Phys Rehabil Med 2016;52(3):271-278 [FREE Full text] [Medline: 26923644]
29. Chen YW, Li KY, Lin CH, Hung PH, Lai HT, Wu CY. The effect of sequential combination of mirror therapy and robot-assisted therapy on motor function, daily function, and self-efficacy after stroke. Sci Rep 2023;13(1):16841 [FREE Full text] [doi: 10.1038/s41598-023-43981-3] [Medline: 37803096]

XSL•FO
RenderX

Abbreviations

CARE: Case Report
DASH: Disabilities of the Arm, Shoulder, and Hand Questionnaire
FM-UE: Fugl-Meyer Upper Extremity Assessment
MCID: minimal clinically important difference
MDC: minimal detectable change
MNS: mirror neuron system
SCRIBE: Single-Case Reporting Guideline in Behavioural Interventions
VR: virtual reality
WMFT: Wolf Motor Function Test

Edited by J Sarvestan; submitted 15.07.25; peer-reviewed by D Poenaru; comments to author 17.09.25; revised version received 12.11.25; accepted 13.11.25; published 24.12.25.

Please cite as:
Hare S, Vice J, Bowman MH, Wright A, Young R, Ahmad M, Logan M, Lai B
Adapted Virtual Reality Exergaming Using Off-the-Shelf Supplies for Poststroke Hemiparetic Arm Rehabilitation: Case Study
JMIR Form Res 2025;9:e80721
URL: https://formative.jmir.org/2025/1/e80721
doi:10.2196/80721
PMID:

©Sadie Hare, Jason Vice, Mary H Bowman, Ashley Wright, Raven Young, Mahmoud Ahmad, Maggie Logan, Byron Lai. Originally published in JMIR Formative Research (https://formative.jmir.org), 24.12.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Feasibility of an Electronic Patient-Reported Outcome System in People Living With HIV: Retrospective Analysis of a Mobile App-Based Pilot Study

Yusuke Yoshino[1,2,3], MD, PhD; Yoshitaka Kimura[1], MD, PhD; Yoshitaka Wakabayashi[2], MD, PhD; Takatoshi Kitazawa[2], MD, PhD

[1]Department of Microbiology and Immunology, School of Medicine, Teikyo University, 2-11-1, Kaga, Itabashi, Tokyo, Japan

[2]Infectious Disease Group, Internal Medicine, Teikyo University Hospital, Tokyo, Japan

[3]Asia International Institute of Infectious Disease Control, Teikyo University, Tokyo, Japan

Corresponding Author:
Yusuke Yoshino, MD, PhD
Department of Microbiology and Immunology, School of Medicine, Teikyo University, 2-11-1, Kaga, Itabashi, Tokyo, Japan

Abstract

Background: Advances in antiretroviral therapy have transformed HIV into a manageable chronic condition, allowing people living with HIV to live longer. However, with aging, people living with HIV face increased risks of lifestyle-related diseases and unaddressed psychosocial issues, including stigma, discrimination, and mental health concerns. Patient-reported outcomes are essential tools in person-centered care and have demonstrated clinical utility in oncology and rheumatology; yet remains underutilized in infectious disease settings. Recently, the European AIDS Clinical Society has recommended electronic patient-reported outcome (ePRO) systems to support HIV care.

Objective: This study aimed to evaluate the feasibility and clinical utility of a smartphone-based ePRO system specifically developed for people living with HIV in Japan.

Methods: A retrospective study was conducted among people living with HIV who attended the HIV outpatient clinic at Teikyo University Hospital between July and September 2022. Participants who consented to use the ePRO system installed a smartphone app and completed the HIV symptom index prior to their clinic visit. Physicians reviewed the responses during consultations; following the visits, patients completed a usability survey addressing clarity, response time, satisfaction, communication quality, and intention for future use. Medical records were reviewed to determine any new symptoms, findings, or medical actions that had not been documented in the previous two years.

Results: A total of 27 people living with HIV (median age, 46 years; 100% male) used the ePRO app, and 25 (93%) completed the postuse questionnaire. Of these, 19/25 (76%) completed the Symptom Index within 5 minutes, while one participant required more than 15 minutes. Regarding usability, 76% (n=19) reported being satisfied or very satisfied, and 76% (n=19) found the system useful in improving communication with their provider. Additionally, 76% (n=19) expressed willingness to use the system again, while 5 (20%) participants indicated interest only if improvements were made. Medical record analysis revealed that 17/27 patients (63%) had new clinical information documented, including mental health symptoms (n=7, 26%), skin problems (n=7, 26%), and new diagnoses or treatment changes in 6 (22%) cases. Over 40% (n=11) of patients reported issues such as anxiety, insomnia, dermatological symptoms, or concerns related to body image.

Conclusions: This pilot study demonstrated that a smartphone-based ePRO system for people living with HIV is feasible and well-accepted in real-world clinical practice. It facilitated early detection of psychosocial and physical issues that may otherwise be overlooked in routine care and improved patient–provider communication. These findings support the integration of ePRO systems into HIV care and underscore the need for further refinement of the app and prospective studies to assess long-term impact on patient outcomes and quality of care.

(JMIR Form Res 2025;9:e71493) doi:10.2196/71493

KEYWORDS

patient reported outcome; smartphone app; quality of life; feasibility test; electronic patient reported outcome

XSL•FO
RenderX

Introduction

Advances in HIV treatment have transformed it from incurable to a manageable disease, requiring continuous use of medication to suppress viral growth. However, as people living with HIV live longer, the risk of lifestyle diseases such as diabetes increases with age [1]. Additionally, previous studies have shown that stigma and discrimination remain prevalent, potentially leading to psychological issues [2]. This has increased the demand for patient-centered medicine beyond viral suppression.

Patient-reported outcome (PROs) are gaining attention as tools for patient-centered care. PROs are based on patients' direct responses to questionnaires. Their validity has been evaluated in various studies, and PROs can assess symptoms, functioning, health status, and quality of life (QOL) [3]. PROs can be disease-specific or general, applicable across conditions. Clinicians can select appropriate PROs for care and research.

PROs are increasingly used for diseases affecting long-term QOL, including malignancies. Studies have revealed gaps between patients and providers in assessing treatment side-effects. Patient-centered evaluations are crucial for improving health-related QOL (HRQOL) in cancer care [4]. In collagen diseases, PROs are used to evaluate symptoms like pain and improve care quality and HRQOL. Thus, PROs are valuable for enhancing care and supporting patient-centered medicine.

PROs have rarely been evaluated in infectious diseases, particularly acute infections, which often resolve with short-term treatment. We previously assessed PROs to measure QOL in patients with community-acquired influenza [5]. After antiviral treatment for influenza, mental QOL worsened during the course. Physical QOL was higher in vaccinated than in nonvaccinated patients [6]. These findings suggest PROs may benefit HIV care, as HIV has become a chronic disease.

In Europe and the United States, PROs are used in HIV trials to evaluate efficacy and adverse effects. Experts report that PROs improve HRQOL in people living with HIV when applied in practice. Recently, an electronic patient-reported outcome (ePRO) system was developed for real-time health assessment via tablets or computers in clinics. The European AIDS Clinical Society 2023 guidelines recommend this ePRO system [7]. In 2022, we collaborated with Integrity Healthcare Corporation and Gilead Sciences Inc. to develop a home-based mobile ePRO app. Here, we present a feasibility study of our prototype mobile ePRO app and its utility.

Methods

Study Design and Setting

The responses of a postuse questionnaire and the impact of the ePROs on clinical practice were retrospectively evaluated among outpatients who visited the Teikyo University Hospital, Tokyo, Japan, between July and September 2022. Participants had agreed to use the ePRO app, particularly during their clinic visits. First, the participants were asked to install the newly developed mobile app on their phones. Informed consent to use the app was obtained from all participants. They responded to the PRO through the app during the week preceding their clinic visit. During the visit, the clinician reviewed the PRO results and treated the participants. All people living with HIV visiting the clinic during the observation period were introduced to the app, and those who agreed to use it on the same day were included in the analysis.

After the clinic visit, the PROs answered via the app were evaluated through a survey of questions concerning the clarity of content, time required to answer, satisfaction with use, changes in quality of communication with the clinician, and desire to use PROs in the future. Satisfaction was assessed with the following question, rated on a 5-point scale (very satisfied, satisfied, neutral, somewhat dissatisfied, dissatisfied): "How was your experience with the consultation after answering the HIV-SI via the app and sharing the results with your doctor in advance?" In addition, medical records from the consultation when the PRO was used were reviewed to identify any new symptoms or findings that had not existed in the past two years, or if any additional medical procedures initiated. The ePRO results were also summarized.

The ePRO content was based on the HIV symptom index [8], comprising 20 items, including those on physical findings, psychological problems, sexual activity, and body image. Each item was rated on a five-point scale: (1) I did not have this symptom, (2) I was not bothered, (3) I was a little bothered, (4) I was indeed bothered, and (5) I was bothered a lot. Responses to the questions were available to the responsible physician immediately after the people living with HIV submitted them online.

The number of participants (n) and corresponding percentages (%) are reported in this paper.

Ethical Considerations

This study was conducted in accordance with the Declaration of Helsinki and approved by the Teikyo University Ethics Committee (approval number: 23‐009). Due to the retrospective nature of the study and the use of anonymized data, the requirement for written informed consent was waived. Instead, an opt-out method was used to inform patients about the study via the institution's website, providing them with the opportunity to decline participation. All data were fully anonymized prior to analysis, and no identifying information was included. No financial compensation or other incentives were provided to participants.

Results

Twenty-seven patients used the mobile app in clinical practice. All patients were male, with a median age of 46 years. All were taking antiretroviral treatment, had maintained a CD4-positive lymphocyte count >200/μL for the past six months, and had a viral load <200 copies.

Twenty-five patients returned the postuse questionnaire. Nineteen of 25 (76%) respondents reported it took 5 minutes to complete; 5 (20%) reported 5-10 minutes, and one (4%) reported >15 minutes. Regarding satisfaction, 19 (76%) patients

XSL•FO
RenderX

were very satisfied or satisfied, 5 (20%) were neutral, and one (4%) was somewhat dissatisfied. Nineteen (76%) patients found the questionnaire very or mostly useful for communication; 6 (24%) patients reported it to be neutral. Regarding future use, 19 (76%) patients wished to continue, 5 (20%) preferred use with improvements, and 1 did not.

Medical record review showed that 17/27 (63%) patients had additional information not recorded in the past two years. Of these, 7 (26%) reported mental health–related symptoms and 7 (26%) had skin problems. In 6 (22%) patients, PRO usage led to new treatments or actions not taken in the previous two years. Three patients were referred to dermatology, and 1 was newly diagnosed with chronic obstructive pulmonary disease following a complaint of respiratory symptoms.

Over 40% (n=11) patients reported symptoms including anxiety, sleep issues, rashes, dry or itchy skin, and appearance changes from weight or fat gain.

Discussion

This study demonstrated the feasibility and clinical utility of a newly developed ePRO smartphone app for people living with HIV. The system was well accepted, with most users completing the questionnaire in under 5 minutes and reporting high satisfaction. The ePRO facilitated improved communication between patients and clinicians and led to documentation of new clinical information in nearly two-thirds of participants. These findings support the hypothesis that an ePRO system tailored for people living with HIV can enhance patient-centered care and identify unrecognized issues.

The short completion time and favorable user responses suggest that the app's design featuring fewer questions and a simplified interface, contributed to its high acceptability. This aligns with studies emphasizing minimal patient burden when implementing PROs in chronic disease care [3]. The enhanced documentation of mental and dermatological symptoms highlights how ePROs can uncover patient concerns that may be under-reported during standard consultations. These findings align with literature showing PROs improve HRQOL by better capturing subjective symptoms [9].

Notably, mental health issues common among people living with HIV often go undiagnosed due to their subjective nature and limited psychiatric training among providers. Similarly, skin-related concerns such as rashes or appearance changes are frequently under-recognized in routine care. ePROs appear to bridge this gap by allowing patients to express symptoms that may not surface during clinician-led interviews. Compared to paper- or tablet-based PROs, smartphone apps offer greater accessibility and real-time symptom reporting from any location. As smartphone use increases globally, especially among younger generations, app-based ePROs hold promise for broader implementation in HIV care and beyond.

This study has limitations. First, the small sample from a single center may limit generalizability. Second, the retrospective design and lack of a control group prevent causal inference. Third, only the HIV symptom index was used; the utility of other PRO tools remains unexplored. While most participants expressed satisfaction, some suggested improvements, indicating room for optimization. Finally, the short observation period limits assessment of long-term utility and sustainability.

The findings of this study underscore the feasibility and clinical relevance of incorporating smartphone-based ePRO systems into HIV outpatient care. By enabling the identification of otherwise undocumented symptoms and improving communication between patients and providers, ePROs may contribute to more holistic and person-centered care models. These findings support the integration of digital health tools in HIV management and highlight the potential for ePROs to address unmet needs in clinical communication and symptom assessment. Future research should refine the app, broaden use, and evaluate long-term outcomes of routine ePRO implementation.

Acknowledgments

We acknowledge Integrity Healthcare Co., Ltd. for providing the system used for data collection and analysis. This study was supported in part by Gilead Sciences Inc., which provided financial support for inventing the application.

YY is affiliated with the Japan AIDS Society and serves as a member of the International Cooperation Committee. YY also plays a pivotal role in introducing the use of patient-reported outcomes in Japan by drawing from experiences abroad.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Methodology: YY
Data curation: TK, YK, YW, YY
Writing – original draft: YY

Conflicts of Interest

None declared.

XSL•FO
RenderX

References

1. Schouten J, Wit FW, Stolte IG, et al. Cross-sectional comparison of the prevalence of age-associated comorbidities and their risk factors between HIV-infected and uninfected individuals: the AGEhIV cohort study. Clin Infect Dis 2014 Dec 15;59(12):1787-1797. [doi: 10.1093/cid/ciu701] [Medline: 25182245]
2. Chenneville T, Kosyluk K, Gabbidon K, Franke M, Serpas D, Galea JT. Positive, Open, Proud: an adapted disclosure-based intervention to reduce HIV stigma. Front Glob Womens Health 2024;5:1469465. [doi: 10.3389/fgwh.2024.1469465] [Medline: 39664652]
3. McKenna SP. Measuring patient-reported outcomes: moving beyond misplaced common sense to hard science. BMC Med 2011 Jul 14;9:1-12. [doi: 10.1186/1741-7015-9-86] [Medline: 21756344]
4. Basch E, Iasonos A, McDonough T, et al. Patient versus clinician symptom reporting using the National Cancer Institute Common Terminology Criteria for Adverse Events: results of a questionnaire-based study. Lancet Oncol 2006 Nov;7(11):903-909. [doi: 10.1016/S1470-2045(06)70910-X] [Medline: 17081915]
5. Yoshino Y, Kitazawa T, Ota Y. Clinical efficacy of baloxavir marboxil in the treatment of seasonal influenza in adult patients: a prospective observational study. Int J Gen Med 2020;13:735-741. [doi: 10.2147/IJGM.S273266] [Medline: 33061543]
6. Yoshino Y, Wakabayashi Y, Kitazawa T. The clinical effect of seasonal flu vaccination on health-related quality of life. Int J Gen Med 2021;14:2095-2099. [doi: 10.2147/IJGM.S309920] [Medline: 34079350]
7. Ambrosioni J, Levi L, Alagaratnam J, et al. Major revision version 12.0 of the European AIDS Clinical Society guidelines 2023. HIV Med 2023 Nov;24(11):1126-1136. [doi: 10.1111/hiv.13542] [Medline: 37849432]
8. Justice AC, Holmes W, Gifford AL, et al. Development and validation of a self-completed HIV symptom index. J Clin Epidemiol 2001 Dec;54 Suppl 1:S77-S90. [doi: 10.1016/s0895-4356(01)00449-8] [Medline: 11750213]
9. Yoshino Y, Kimura Y. Clinical value of patient-reported outcomes for people living with HIV: mini-review. Re:GEN Open 2024 Jan 1;4(1):34-38. [doi: 10.1089/regen.2024.0005]

abbreviations

ePRO: electronic PRO
HRQOL: health-related quality of life
PRO: patient-reported outcome
QOL: quality of life

Edited by A Mavragani; submitted 19.01.25; peer-reviewed by M Suzuki; revised version received 07.04.25; accepted 10.04.25; published 10.06.25.

Please cite as:
Yoshino Y, Kimura Y, Wakabayashi Y, Kitazawa T
Feasibility of an Electronic Patient-Reported Outcome System in People Living With HIV: Retrospective Analysis of a Mobile App-Based Pilot Study
JMIR Form Res 2025;9:e71493
URL: https://formative.jmir.org/2025/1/e71493
doi:10.2196/71493

© Yusuke Yoshino, Yoshitaka Kimura, Yoshitaka Wakabayashi, Takatoshi Kitazawa. Originally published in JMIR Formative Research (https://formative.jmir.org), 10.6.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Short Paper

Identifying the Minimal Clinically Important Difference in Emotion Regulation Among Youth Using the JoyPop App: Survey Study

Jaidyn Charlton[1], MA; Ishaq Malik[1], MA; Angela M Ashley[1], HBSc; Amanda Newton[2], PhD; Elaine Toombs[1], PhD; Fred Schmidt[1], PhD; Janine V Olthuis[3], PhD; Kristine Stasiuk[4], HBSW; Tina Bobinski[4], MSW; Aislin Mushquash[1], PhD

[1]Department of Psychology, Lakehead University, Thunder Bay, ON, Canada
[2]Department of Pediatrics, University of Alberta, Edmonton, AB, Canada
[3]Department of Psychology, University of New Brunswick, Fredericton, NB, Canada
[4]Dilico Anishinabek Family Care, Fort William First Nation, ON, Canada

Corresponding Author:
Aislin Mushquash, PhD
Department of Psychology
Lakehead University
955 Oliver Road
Thunder Bay, ON
Canada
Phone: 1 8073438010 ext 8771
Email: aislin.mushquash@lakeheadu.ca

Abstract

Background: The minimal clinically important difference (MCID) is an important threshold to consider when evaluating the meaningfulness of improvement following an intervention. The JoyPop app is an evidence-based smartphone app designed to improve resilience and emotion regulation. Information is needed regarding the JoyPop app's MCID among culturally diverse youth.

Objective: This study aims to calculate the MCID for youth using the JoyPop app and to explore how the MCID may differ for a subset of Indigenous youth.

Methods: Youth (N=36; aged 12-18 years) were recruited to use the JoyPop app for up to 4 weeks as part of a larger pilot evaluation. Results were based on measures completed after 2 weeks of app use. The MCID was calculated using emotion regulation change scores (Difficulties in Emotion Regulation–Short Form [DERS-SF]) and subjective ratings on the Global Rating of Change Scale (GRCS). This MCID calculation was completed for youth overall and separately for Indigenous youth only.

Results: A significant correlation between GRCS scores and change scores on the DERS-SF supported face validity (r=–0.37; P=.04). The MCID in emotion regulation following the use of the JoyPop app for youth overall was 2.80 on the DERS-SF. The MCID for Indigenous youth was 4.29 on the DERS-SF. In addition, most youth reported improved emotion regulation after using the JoyPop app.

Conclusions: These MCID findings provide a meaningful threshold for improvement in emotion regulation for the JoyPop app. They provide potential effect sizes and can aid in sample size estimations for future research with the JoyPop app or e-mental health technologies in general. The difference between overall youth and Indigenous youth MCID values also highlights the importance of patient-oriented ratings of symptom improvement as well as cultural considerations when conducting intervention research and monitoring new interventions in clinical practice.

(JMIR Form Res 2025;9:e64483) doi:10.2196/64483

KEYWORDS

mHealth; mobile health; app; psychometrics; emotion regulation; Indigenous mental health; Indigenous youth; mental health interventions; resilience; clinical psychology; adolescent mental health; mental health; JoyPop; pediatrics; mobile phone

XSL•FO
RenderX

Introduction

The minimal clinically important difference (MCID) is the smallest change in an outcome (eg, symptom or ability) that patients consider meaningful to continue using an intervention [1]. These patient-derived scores serve as thresholds for outcome improvement when evaluating new interventions [1] and inform researchers on potential effect sizes in clinical trials.

The JoyPop app is a mobile health (mHealth) tool with a growing evidence base supporting its use among youth (Figure 1) [2]. JoyPop app use is associated with improved emotion regulation [2], and service providers, adolescents, and young adults view it as an acceptable, useful, and feasible tool to support existing mental health services [3,4]. Research to determine how much change in emotion regulation is clinically meaningful to app users is needed. Additionally, identifying the MCID among Indigenous youth, who experience greater mental health difficulties [5], is required.

Figure 1. JoyPop app features.

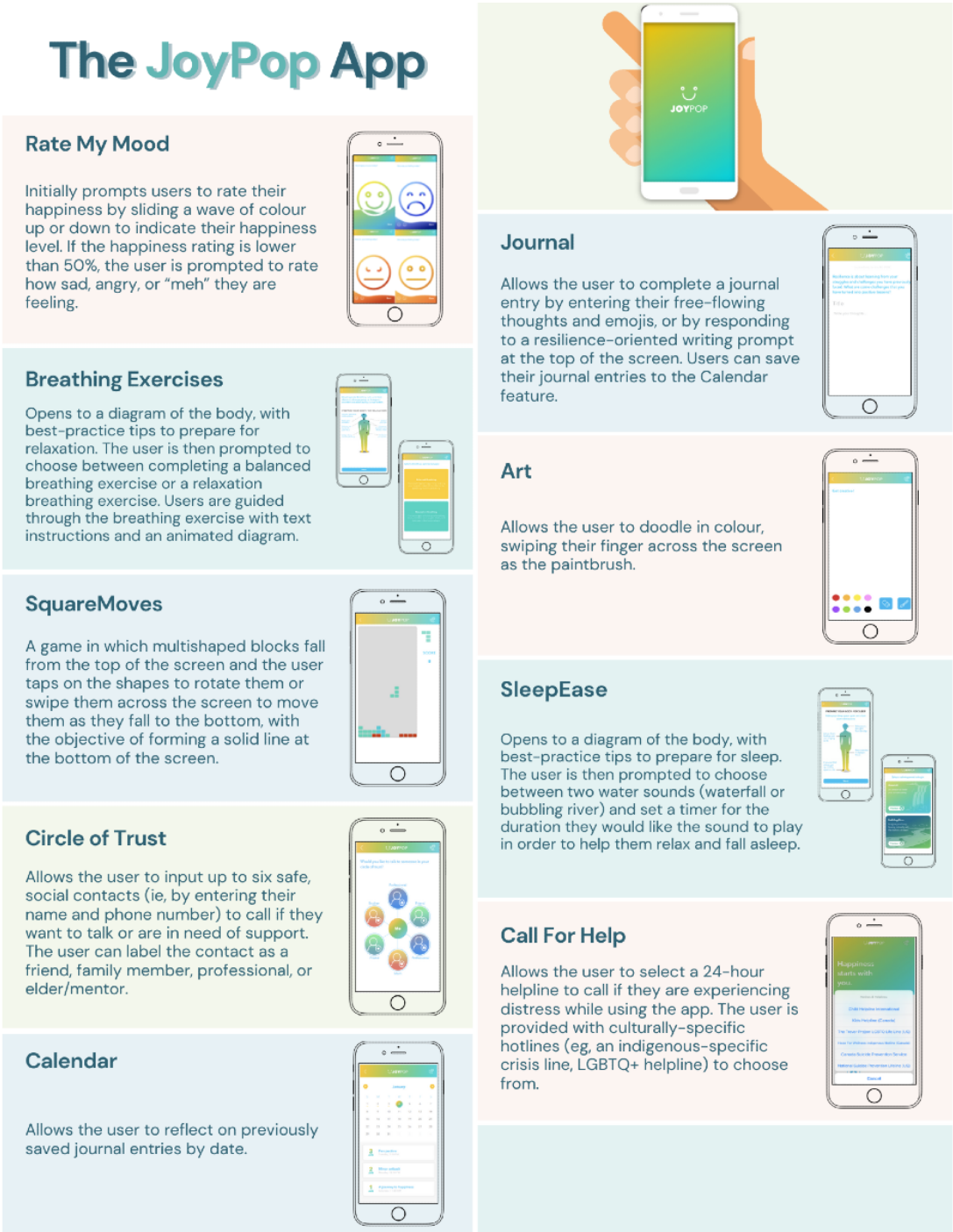

We asked youth about the emotion regulation change that was important to them before and after using the JoyPop app. We calculated the MCID for all youth and Indigenous youth. We also explored the importance of patient perspectives and how MCIDs may differ between populations to provide patient-oriented recommendations for future research. Given

XSL•FO
RenderX

this study's exploratory nature, we did not have a priori hypotheses.

Methods

Participants and Procedure

All youth (aged 12-18 years) who used the JoyPop app while receiving mental health services from the 2 largest mental health organizations in Northwestern Ontario were eligible to participate; 36 participated. Staff (ie, the youth's clinician) from the organizations informed them about the study. Staff sent the research team contact information for youth who were interested. Youth were contacted to receive further information. Interested youth attended an orientation session and downloaded the app onto their own device or were provided with a refurbished phone (containing just the JoyPop app) to use during the study. Participants were encouraged to use the app at least twice daily for 4 weeks and were sent 2 daily reminders to use the app through email and text. Participants completed measures for a larger pilot study at baseline (time 1; N=36), after 2 weeks (time 2; n=31), and after 4 weeks (time 3; n=25). Compensation was provided for each time point completed (up to CAD $80 [US $54]). If the JoyPop app was downloaded to participants' devices, they could continue using it after study completion. Given the pilot study's small sample size and attrition by time 3, we used time-1 and time-2 data for calculating the MCID.

Measures

GRCS

The Global Rating of Change Scale (GRCS) is a single-item measure assessing the degree of change individuals felt occurred during or after an intervention [6]. The GRCS asked participants "As a result of using the JoyPop app, has your ability to understand and manage your emotions been..." Participants responded using an 11-point Likert scale (–5="a very great deal worse" to 0="no change" to +5="a very great deal better") at time 2. At time 1, we asked participants how much change in their ability to regulate their emotions they "would like to see" to deem using the app worthwhile.

DERS-SF

The Difficulties in Emotion Regulation Scale–Short Form (DERS-SF) [7,8] is an 18-item questionnaire assessing emotion regulation abilities (eg, "When I was upset, I became out of control") rated on a 5-point Likert scale (1="almost never" to 5="almost always"). Total scores range from 18 to 80, with higher scores indicating greater difficulties. The DERS-SF has satisfactory reliability and validity [7,8]. The reliability among participants in this study was excellent (time-1 α=.90; time-2 α=.94).

Statistical Analysis

We used the anchor-based method: patient-reported scores on outcome measures were compared with a more subjective measure of change ("anchor" or cutoff) [1]. Time-1 GRCS scores showed that most participants reported wanting to feel +2 "somewhat better" emotion regulation from using the app (Table 1). Therefore, +2 was used as the "anchor" at time 2 when determining the MCID. We calculated the mean change score on the DERS-SF (outcome measure) for participants who rated +2 on the GRCS ("anchor")—this became the MCID. We ran descriptive statistics to calculate the proportion of participants who scored at or above the MCID. We completed these analyses for the total sample and separately for Indigenous participants. Correlation tested face validity.

Table 1. Frequencies of Global Rating of Change Scale (GRCS) ratings.

GRCS rating[a]	Time 1[b] (N=36), n (%)	Time 2 (n=31), n (%)
Emotion regulation worsened (rated from –1 to –5)	4 (11)	2 (6)
No change in emotion regulation (rated 0)	4 (11)	5 (16)
Emotion regulation was almost the same, hardly better at all (rated +1)	3 (8)	8 (26)
Emotion regulation was somewhat better (rated +2)	13 (36)	10 (32)
Emotion regulation was much better (rated from +3 to +5)	12 (33)	6 (19)

[a]The percentage of youth that rated each item on the GRCS (N=36). Participants (aged 12-18 years) were recruited for the pilot study from 2 mental health organizations in Northwestern Ontario.

[b]GRCS ratings at time 1 (N=36) reflect how much change youth prospectively wanted to see from using the JoyPop app. GRCS ratings at time 2 (n=31) reflect how much change youth actually experienced after 2 weeks of using the JoyPop app.

Ethical Considerations

The study was reviewed and approved by research ethics boards at Lakehead University (ROMEO #1468701) and the collaborating mental health organizations. Study information was reviewed with all youth. All participants provided informed consent. Data only contained the ID number provided at study onset. The list linking participant ID numbers to participant names was kept separately from the study dataset and was destroyed after data collection. Youth could withdraw at any time without penalty.

Results

Descriptive Statistics

Participant demographics (Table 2) and descriptive statistics for emotion regulation and global ratings of change (Table 3) are shown. Time-2 GRCS scores showed that many participants reported improved emotion regulation after using the app for 2 weeks (Table 1). A significant negative correlation between time-2 GRCS ratings and DERS-SF change scores provided face validity that the GRCS can indicate a meaningful patient-rated change in emotion regulation (r=–0.37; P=.04).

XSL•FO
RenderX

Table 2. Youth[a] participant demographics.

Variables	Value (N=36)
Ethnicity, n (%)	
White	9 (25)
Indigenous	25 (69)
Mixed	1 (3)
South Asian	1 (3)
Age (years), mean (SD)	14.47 (1.75)
Gender identity, n (%)	
Girl	22 (61)
Boy	5 (14)
Agender or nonbinary	7 (19)
Transgender	2 (6)

[a]Participants (N=36; aged 12-18 years) were recruited for the pilot study from 2 mental health organizations in Northwestern Ontario.

Table 3. Mean and SD values of the Difficulties in Emotion Regulation Scale–Short Form (DERS-SF) and Global Rating of Change Scale (GRCS) overall and for Indigenous youth.

Participants[a] and time point	Participants, n	DERS-SF[b], mean (SD)	GRCS[c], mean (SD)
All youth			
Time 1	36	59.64 (14.03)	1.75 (1.54)
Time 2	31	55.84 (16.11)	1.48 (1.44)
Indigenous youth			
Time 1	25	58.44 (15.92)	1.64 (1.29)
Time 2	21	53.19 (17.49)	1.43 (1.36)

[a]Participants (aged 12-18 years) were recruited for the pilot study from 2 mental health organizations in Northwestern Ontario.

[b]Higher scores indicate greater emotional dysregulation and lower scores indicate improvement.

[c]Higher scores indicate higher subjective improvement. GRCS ratings at time 1 reflect how much change youth prospectively wanted to see from using the JoyPop app. GRCS ratings at time 2 reflect how much change youth actually experienced after 2 weeks of using the JoyPop app.

MCID

The mean change score on the DERS-SF for participants who rated +2 “somewhat better” on the GRCS at time 2 was –2.80 (SD 9.22), translating to an MCID of 2.80. Using this threshold, 45% (14/31) of participants met or exceeded this change after 2 weeks, regardless of their time-2 GRCS rating (Table 4). For Indigenous youth who rated +2 “somewhat better” on the GRCS, the mean change score on the DERS-SF was –4.29 (SD 8.96), indicating that they experienced a 4.29-point improvement in emotion regulation after using the app for 2 weeks and had an MCID of 4.29. Using this threshold, 43% (9/21) of Indigenous youth met or exceeded this change after 2 weeks.

Table 4. Frequencies of Global Rating of Change Scale (GRCS) ratings of those who met or exceeded the minimal clinically important difference on the Difficulties in Emotion Regulation Scale–Short Form.

GRCS rating[a]	Youth (n=14), n (%)
Emotion regulation worsened (rated from –1 to –5)	0
No change in emotion regulation (rated 0)	1 (7)
Emotion regulation was almost the same, hardly better at all (rated +1)	5 (36)
Emotion regulation was somewhat better (rated +2)	5 (36)
Emotion regulation was much better (rated from +3 to +5)	3 (21)

[a]Participants (aged 12-18 years) were recruited for the pilot study from 2 mental health organizations in Northwestern Ontario. This table represents only the youth who reached the minimal clinically important difference of 2.80 or exceeded it (14/31, 45%) and what they rated on the GCRS at time 2.

XSL•FO
RenderX

Discussion

The MCID in emotion regulation was 2.80 for all youth and 4.29 for Indigenous youth. These provide preliminary effect sizes for future research as benchmarks for meaningful patient-oriented change, rather than solely basing findings on statistical significance, which lacks practicality or patient-oriented relevance [9]. Future study designs could recognize that youth generally need at least a 2.80-point improvement in their emotion regulation on the DERS-SF to consider the JoyPop app worthwhile. These findings can also inform future sample size calculations for JoyPop research [10,11].

Evaluating the MCID among Indigenous youth allowed us to consider their unique cultural experiences and provided insight into the effectiveness and appropriateness of the app for this subgroup. Our results suggest that Indigenous youth need to see a greater improvement in emotion regulation to deem the JoyPop app worthwhile and to start feeling "somewhat better." Conducting MCID research with different groups, beyond ethnicity, can provide information on distinct change thresholds that interventions should aim for as opposed to "one-size-fits-all" approaches.

The findings also highlight the value of pairing outcome measures with patient-rated measures of change [12]. Not all participants who met or exceeded the MCID threshold subjectively rated an improvement in their emotion regulation (Table 4). This suggests that many, but not all, youth notice improvements in their emotion regulation, and that some youth require a higher threshold of change (>2.80) to notice improvement. This underscores the importance of measuring how individuals perceive their symptoms and abilities in relation to an intervention. If a young person does not perceive improvement from using an intervention, they are unlikely to continue, even if outcome data show improvement. We recommend subjective experience ratings (eg, GRCS) alongside outcome measures (eg, DERS-SF) when evaluating mHealth interventions.

MCID research typically involves participants reporting on what happened following an intervention. Previous research has relied on clinician judgment and clinical theory to determine their "anchor" on which to base the MCID [13]. We included a novel prospective question at time 1 to explore what youth would want to experience to inform the choice of anchor for our sample. Most youth wanted to see a minimum "somewhat better" improvement in their emotion regulation, aligning with previous research using +2 as an anchor [13]. This is useful in promoting the JoyPop app to youth and service providers as it highlights continuity between clinician and client judgment.

Despite encouraging preliminary findings, this study has limitations. The GRCS has received criticism as a single-item measure, which lacks robust psychometric properties [14]. Single-item measures do not capture the complexity of experience and can introduce differences in interpretation between participants [15]. However, single-item measures are informative and valid when used appropriately, such as concurrently with multi-item measures (eg, alongside the DERS-SF) or when they ask straightforward questions that do not require multiple items (eg, how much change in emotion regulation youth experienced) [14]. Another limitation is that the MCIDs were calculated based on a small sample, which limits generalizability. Nevertheless, these findings provide valuable contributions to the JoyPop app's evidence base. Through cumulative longitudinal evidence, the MCIDs can be improved and expanded on for different populations (eg, adolescent vs young adult; differing ethnicities), thus demonstrating a client-centered approach to research and mHealth app implementation [16].

Acknowledgments

We would like to thank the youth who informed the initial design of the JoyPop app and those who continue to inform its evolution through participation in past and current research. We would also like to acknowledge our community partner organizations for their ongoing support of our collaborations.

Data Availability

The dataset generated and analyzed is publicly available due to our team's adherence to the Ownership, Control, Access, and Possession (OCAP) standards set by the First Nations Information Governance Centre. Questions about study datasets can be directed to the corresponding author and principal investigator (ARM).

Conflicts of Interest

The principal investigator (ARM) recently acquired intellectual property ownership rights for the JoyPop app (in June 2024). Before this time, intellectual property rights were owned by other researchers and institutions. ARM did not own any intellectual property rights when the study was developed, when data were collected, when data were analyzed, or when the initial draft of the manuscript was written.

References

1. Copay AG, Subach BR, Glassman SD, Polly DW, Schuler TC. Understanding the minimum clinically important difference: a review of concepts and methods. Spine J 2007;7(5):541-546. [doi: 10.1016/j.spinee.2007.01.008] [Medline: 17448732]

XSL•FO
RenderX

2. MacIsaac A, Mushquash AR, Mohammed S, Grassia E, Smith S, Wekerle C. Adverse childhood experiences and building resilience with the JoyPop app: evaluation study. JMIR Mhealth Uhealth 2021;9(1):e25087 [FREE Full text] [doi: 10.2196/25087] [Medline: 33393908]
3. Malik I, Perez A, Toombs E, Schmidt F, Olthuis JV, Charlton J, et al. Female youth and mental health service providers' perspectives on the JoyPop™ app: a qualitative study. Front Digit Health 2023;5:1197362 [FREE Full text] [doi: 10.3389/fdgth.2023.1197362] [Medline: 37829596]
4. Mushquash AR, Pearson ES, Waddington K, MacIsaac A, Mohammed S, Grassia E, et al. User perspectives on a resilience-building app (JoyPop): qualitative study. JMIR Mhealth Uhealth 2021;9(7):e28677 [FREE Full text] [doi: 10.2196/28677] [Medline: 34255696]
5. Firestone M, Smylie J, Maracle S, McKnight C, Spiller M, O'Campo P. Mental health and substance use in an urban first nations population in Hamilton, Ontario. Can J Public Health 2015;106(6):e375-e381 [FREE Full text] [doi: 10.17269/cjph.106.4923] [Medline: 26680428]
6. Kamper SJ, Maher CG, Mackay G. Global rating of change scales: a review of strengths and weaknesses and considerations for design. J Man Manip Ther 2009;17(3):163-170 [FREE Full text] [doi: 10.1179/jmt.2009.17.3.163] [Medline: 20046623]
7. Gratz KL, Roemer L. Multidimensional assessment of emotion regulation and dysregulation: development, factor structure, and initial validation of the difficulties in emotion regulation scale. Journal of Psychopathology and Behavioral Assessment 2004;26(1):41-54. [doi: 10.1023/B:JOBA.0000007455.08539.94]
8. Kaufman EA, Xia M, Fosco G, Yaptangco M, Skidmore CR, Crowell SE. The Difficulties in Emotion Regulation Scale-Short Form (DERS-SF): validation and replication in adolescent and adult samples. J Psychopathol Behav Assess 2015;38(3):443-455. [doi: 10.1007/s10862-015-9529-3]
9. Anvari F, Lakens D. Using anchor-based methods to determine the smallest effect size of interest. Journal of Experimental Social Psychology 2021;96:104159. [doi: 10.1016/j.jesp.2021.104159]
10. Neely JG, Karni RJ, Engel SH, Fraley PL, Nussenbaum B, Paniello RC. Practical guides to understanding sample size and minimal clinically important difference (MCID). Otolaryngol Head Neck Surg 2007;136(1):14-18. [doi: 10.1016/j.otohns.2006.11.001] [Medline: 17210326]
11. McGlothlin AE, Lewis RJ. Minimal clinically important difference: defining what really matters to patients. JAMA 2014;312(13):1342-1343. [doi: 10.1001/jama.2014.13128] [Medline: 25268441]
12. Cook CE. Clinimetrics corner: the minimal clinically important change score (MCID): a necessary pretense. J Man Manip Ther 2008;16(4):E82-E83 [FREE Full text] [doi: 10.1179/jmt.2008.16.4.82E] [Medline: 19771185]
13. Radomski AD, Bagnell A, Curtis S, Hartling L, Newton AS. Examining the usage, user experience, and perceived impact of an internet-based cognitive behavioral therapy program for adolescents with anxiety: randomized controlled trial. JMIR Ment Health 2020;7(2):e15795 [FREE Full text] [doi: 10.2196/15795] [Medline: 32022692]
14. Allen MS, Iliescu D, Greiff S. Single item measures in psychological science. European Journal of Psychological Assessment 2022;38(1):1-5. [doi: 10.1027/1015-5759/a000699]
15. Sedaghat AR. Understanding the minimal clinically important difference (MCID) of patient-reported outcome measures. Otolaryngol Head Neck Surg 2019;161(4):551-560. [doi: 10.1177/0194599819852604] [Medline: 31159641]
16. McGrath P, Wozney L, Rathore SS, Notarianni M, Schellenberg P. Toolkit for e-mental health implementation. Mental Health Commission of Canada. 2018. URL: https://mentalhealthcommission.ca/resource/toolkit-for-e-mental-health-implementation/ [accessed 2024-09-28]

Abbreviations

DERS-SF: Difficulties in Emotion Regulation Scale–Short Form
GRCS: Global Rating of Change Scale
MCID: minimal clinically important difference
mHealth: mobile health

Edited by A Mavragani; submitted 18.07.24; peer-reviewed by A Serlachius; comments to author 16.08.24; revised version received 28.08.24; accepted 24.09.24; published 23.01.25.

Please cite as:
Charlton J, Malik I, Ashley AM, Newton A, Toombs E, Schmidt F, Olthuis JV, Stasiuk K, Bobinski T, Mushquash A
Identifying the Minimal Clinically Important Difference in Emotion Regulation Among Youth Using the JoyPop App: Survey Study
JMIR Form Res 2025;9:e64483
URL: https://formative.jmir.org/2025/1/e64483
doi:10.2196/64483
PMID:39847426

XSL•FO
RenderX

©Jaidyn Charlton, Ishaq Malik, Angela M Ashley, Amanda Newton, Elaine Toombs, Fred Schmidt, Janine V Olthuis, Kristine Stasiuk, Tina Bobinski, Aislin Mushquash. Originally published in JMIR Formative Research (https://formative.jmir.org), 23.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Short Paper

Assessing the Impact of a Virtual Reality Cognitive Intervention on Tennis Performance in Junior Tennis Players: Pilot Study

Joaquin A Anguera[1], PhD; Aleem Choudhry[2], BS, MBA; Michael Seaman[2], BS, MBA; Dominick Fedele[2], BS

[1]Neuroscape, Department of Neurology and Psychiatry, Weill Institute for Neurosciences, University of California, San Francisco, San Francisco, CA, United States

[2]Mastermind Sports, Naperville, IL, United States

Corresponding Author:
Joaquin A Anguera, PhD
Neuroscape
Department of Neurology and Psychiatry, Weill Institute for Neurosciences
University of California, San Francisco
675 Nelson Rising Lane
San Francisco, CA, 94158
United States
Phone: 1 415 702 4322
Email: joaquin.anguera@ucsf.edu

Abstract

Background: There is evidence that cognitive training interventions can positively impact executive functions, and that some studies have demonstrated that athletes typically exhibit greater accuracy and faster response times on select cognitive tasks. While the engagement of executive functions is suggested to be part of high-level sporting activities, it is unclear whether such training approaches could directly benefit athletic performance.

Objective: The objective of this study was to evaluate the impact of a combined virtual reality (VR)– and tablet-based cognitive training intervention on adolescent tennis players' performance. Here, we examined differences in Universal Tennis Rating (UTR) between players who supplemented their regular tennis training with a cognitive training intervention and a group that continued regular tennis training alone. This custom cognitive training program targeted specific cognitive control abilities including attention, working memory, and goal management.

Methods: Data were collected from a cohort of tennis players in a randomized controlled trial design led by the dedicated research team. Participants (N=23, age: mean 14.8, SD 2.4 years) from the Czech Lawn Tenis Klub (Prague, Czech Republic) were invited to participate in this study. These individuals were randomized into an intervention + training-as-usual group (n=13) or training-as-usual group (control group; n=10), with the change in UTR score being the primary metric of interest.

Results: There was no difference in UTR between the 2 groups at baseline (intervention: mean 8.32, SD 2.7; control: mean 7.60, SD 2.3). Following the treatment period, individuals in the intervention group showed a significant improvement in their UTR (an increase of 0.5; t_{12}=4.88, P<.001) unlike the control group (an increase of 0.02; t_9=1.77, P=.12). On comparing the change in UTR (posttraining UTR minus pretraining UTR) attained by each group, we found that the intervention group had a 38% greater improvement in UTR than the control group. An analysis of covariance revealed a significantly greater improvement in UTR for the intervention group than for the control group ($F_{1,20}$=8.82, P=.008).

Conclusions: The present findings suggest that training cognitive abilities through an immersive visual platform may benefit athletic performance, including tennis. Such a result warrants careful consideration, given the known difficulties in evidencing far transfer not only in cognitive studies but also in athletic activities. These preliminary pilot findings suggest that the Mastermind Cognitive Training program may be a viable tool for supplementing athletic training practices, although this result warrants further investigation and replication. However, many questions remain unanswered, and further work is needed to better understand the potential utility and mechanisms underlying potential effects of such a platform.

(JMIR Form Res 2025;9:e66979) doi:10.2196/66979

XSL•FO
RenderX

KEYWORDS

executive function; serious games; cognitive training; performance enhancement; athletes; sport; pilot study; VR; virtual reality; serious game; tennis; adolescents; teenagers; youth; randomized controlled trial; players

Introduction

The engagement of cognitive control abilities, defined here as attention, working memory, and multitasking [1], is thought to be a critical part of high-level athletic performance. This idea stems from research showing that both amateur and elite athletes typically exhibit greater accuracy and faster response times on select cognitive tasks [2,3], and that one's underlying cognitive abilities can predict future athletic achievement [4-6]. However, the existing literature on whether enhancing these abilities through cognitive training can improve athletic performance is limited [7-9]. Numerous examples in the cognitive training literature provide reasons why so-called "far transfer" (eg, the ability to apply learning from one context to a different, dissimilar context [10]) is rarely observed [11,12]. However, those positive findings following the use of specific cognitive training platforms suggest that such an approach could possibly be beneficial [13-16]. Such positive effects have been observed in those platforms using closed-loop algorithms, an approach that changes the level of difficulty on a trial-by-trial basis to challenge a given individual appropriately for a given training experience [1,17-19]. This approach has shown generalizable benefits to not only untrained tasks assessing cognitive abilities (including attention, working memory, and multitasking) [16,20-24] but also so-called "real-world" constructs including mood [14,15,25-27] and even academic abilities [28]. Here, we describe the preliminary findings of a proof-of-concept randomized controlled study involving advanced adolescent tennis players who supplemented their regular tennis training with a cognitive training intervention in comparison to a group that continued regular tennis training alone.

Methods

Ethical Considerations

For this project, all study procedures were conducted in accordance with protocols approved by the Fakultní Nemocnice Brno (Brno, Czech Republic; 14-110123/EK). Participant assent and parental consent were obtained prior to study participation in person by the study team. All study data were deidentified to ensure privacy and confidentiality protections were in place. No compensation was given for participating in this study.

Participants

A total of 59 tennis players (age: mean 16.7, SD 3.6 years; 34 male) from the Czech Lawn Tenis Klub (Prague, Czech Republic) were invited to participate in this study (and provided a full overview of the study procedures and time requirements). See Multimedia Appendix 1 for more details.

Intervention

The Mastermind Cognitive Training program used a virtual reality (VR) headset (Oculus Quest 2 by Meta) to deliver a custom cognitive training program targeting specific cognitive control abilities including attention, working memory, and goal management. The VR component comprised 3 modules, each focusing on a different aspect of cognitive control: a multiple object–tracking module aimed at engaging attention (with increasing distraction), a spatial span module to engage working memory, and a visual search module with task switching requirements to engage multitasking abilities. In each case, participants used VR hand-held controllers (Figure 1A) to respond to stimuli presented in a video game–like environment, in some cases "shooting" specific target stimuli while avoiding others (Figure 1B), or trying to catch certain colored targets in a given order stimuli while avoiding nontargets (Figure 1C). Each module was adaptive, adjusting task difficulty based on performance. In addition, participants completed 20 minutes of tablet-based training (iPad) on tasks related to rhythmic timing (where participants were asked to tap on the screen to the beat of music being played [28-31]; Figure 1D). The intervention group was asked to complete 24 program sessions over the course of 10 weeks, with each session consisting of 15 minutes of VR exercises and 20 minutes of tablet-based exercises, totaling approximately 14 hours of training. See Multimedia Appendix 1 for more details.

XSL•FO
RenderX

Figure 1. The Mastermind Cognitive Training program. (A) Participant training during the Mastermind Cognitive Training program using virtual reality. Participants were tennis players from the Czech Lawn Tenis Klub (Prague, Czech Republic) who trained between January and November 2023. Participants used VR hand-held controllers to respond to stimuli presented in a video game–like environment, in some cases "shooting" specific target stimuli while avoiding others or trying to catch certain colored targets in a given order stimuli while avoiding nontargets. (B) the virtual reality attention module. (C) The virtual reality multitasking module. (D) The tablet rhythm training module where participants tapped in a rhythmic fashion to stimuli presented on the screen.

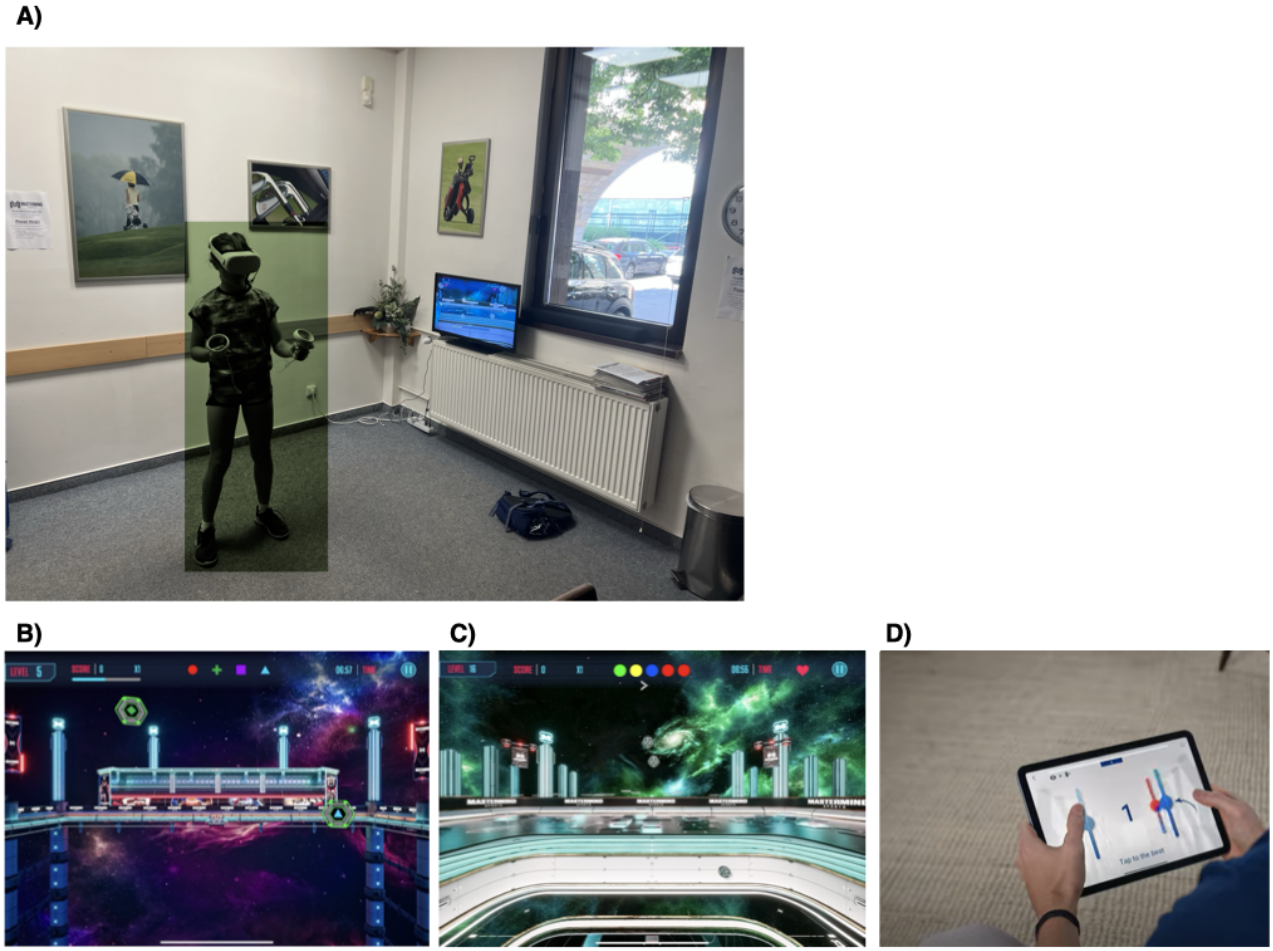

Measure

The primary outcome measure was the change in Universal Tennis Rating (UTR), a scale from 1.00 (eg, a beginner player) to 16.50 (eg, a professional player; for context, the current #1 player in the world is rated 16.05), which promotes fair and competitive play [32]. Each player's UTR was collected before training and after the completion of the 24-session training period, which an interassessment interval of approximately 4 months. See Multimedia Appendix 1 for more details about this measure.

Results

Of the 59 consenting individuals, 9 withdrew upon learning of the study requirements, 15 failed to complete the stipulated training schedule and posttraining assessments due to scheduling conflicts with travel and academic commitments, 8 opted out due to moving out of the area or changing tennis clubs, and 4 discontinued participation due to unrelated injuries requiring physical rehabilitation. Thus, the final sample that completed the stipulated training and outcome assessments consisted of 23 participants: 13 in the intervention group (3 female, aged mean 15.0, SD 2.4) and 10 (7 female, aged mean 14.6, SD 2.5) in the training-as-usual group (see Figure 2 for the CONSORT [Consolidated Standards of Reporting Trials] diagram and Multimedia Appendix 2 for CONSORT checklist). To test for training effects on UTR (Table 1), we conducted an analysis of covariance (ANCOVA) with posttraining UTR as the dependent variable, pretraining UTR as the covariate, and the group as the fixed factor as in our previous work [14,20,33,34]. The analysis revealed a significant group difference in UTR improvement over time ($F_{1,20}$=8.82, P=.008; Figure 3), indicating that the UTR of the intervention group improved significantly more than that of the control group. On comparing the change in UTR (posttraining UTR minus pretraining UTR) attained by each group, the intervention group had a 38% greater improvement in UTR than the control group. Post hoc analysis of the within-group changes in UTR using 2-tailed, paired-samples t tests for each group separately to test for significant differences between each testing session revealed a significant improvement by the intervention group (increase of 0.50; t_{12}=4.88, P<.001) unlike the control group (increase of 0.20; t_9=1.77, P=.12). Critically, there was no difference in UTR at baseline between the 2 groups (independent-samples t test: t_{21}=.68, P=.50).

XSL•FO
RenderX

Figure 2. CONSORT (Consolidated Standards of Reporting Trials) flow diagram.

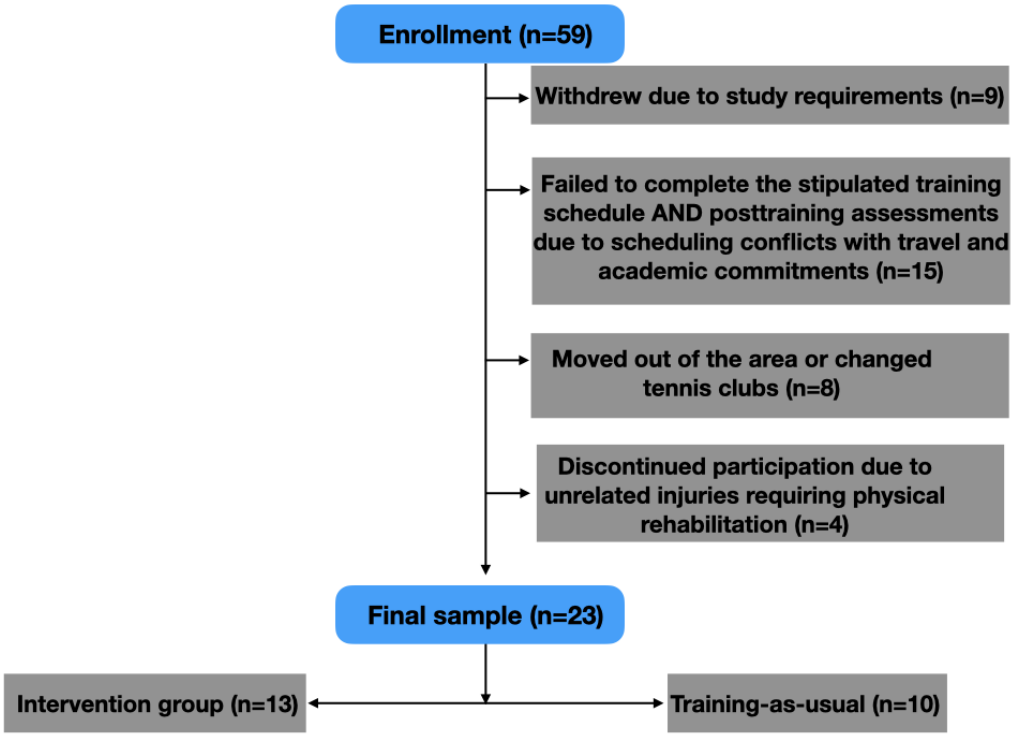

Table 1. Universal Tennis Rating (UTR) by group and timepoint.

	Pretraining UTR, mean (SD)	Posttraining UTR, mean (SD)	Change (posttraining UTR–pretraining UTR)
Intervention	8.32 (2.7)	8.83 (2.4)	0.50 (0.37)
Treatment-as-usual	7.60 (2.3)	7.79 (2.2)	0.20 (0.37)

Figure 3. Universal Tennis Rating (UTR) by group. The UTR is a scale from 1.00 (eg, a beginning player) to 16.50 (eg, a professional player), which promotes fair and competitive play. Each player's UTR was recorded before training and after the completion of the 24-session training period (approximately 4 months). Posttraining UTR scores for the intervention and training-as-usual groups visualized via a raincloud plot with difference-adjusted 95% CIs of the mean. An analysis of covariance (controlling for pretraining performance) revealed that the improvement observed by the intervention group was significantly greater than that of the training-as-usual group ($F_{1,20}$=8.82, P=.008).

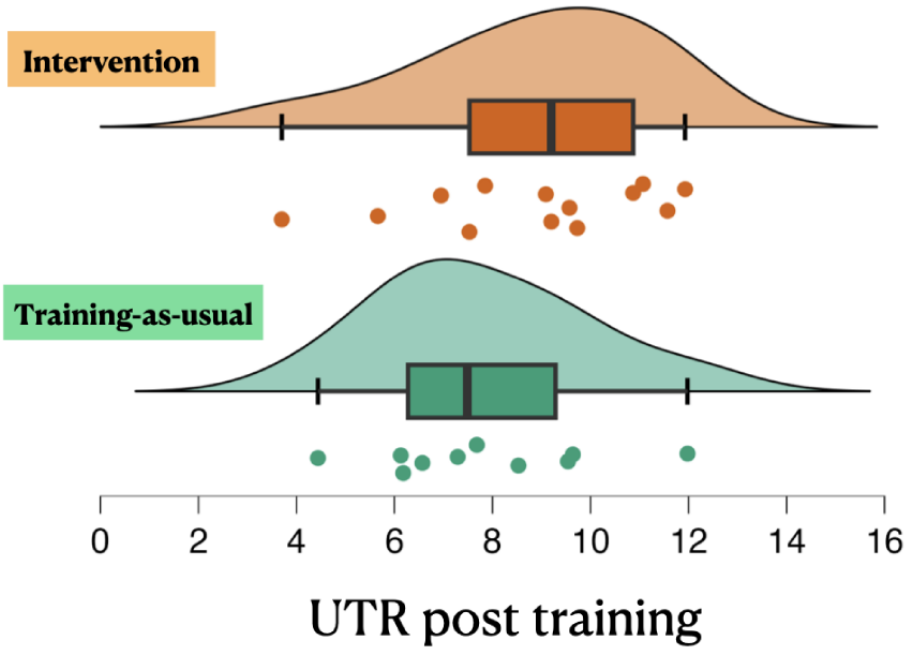

XSL•FO
RenderX

Discussion

Principal Findings

The primary goal of this proof-of-concept pilot study was to investigate the effects of the Mastermind Cognitive Training program in adolescent tennis players. These findings suggest that supplementing regular training with cognitive training can lead to meaningful improvements in real-world athletic performance. It has been suggested that visual-based training (including the use of VR and/or tablet platforms) is more likely to benefit trained athletes than novices [35]. Indeed, this theory has empirical support: “real-world” improvements have been observed, including performance in college-level baseball players [36,37] as well as table tennis performance [38] when coupling such training with their regular sports training [38]. In conjunction with those studies, the present findings suggest that training cognitive abilities through an immersive visual platform may benefit performance on athletic endeavors, including tennis. Such a result warrants careful consideration, given the known difficulties in evidencing far transfer not just in cognitive studies, but especially in athletic activities [7]. That said, our findings do align with those of the few positive studies demonstrating that cognitive training can result in some form of on-field improvements [8,9]. See Multimedia Appendix 1 for a further discussion of practical considerations.

Limitations

Several limitations should be considered with the present findings. First, the most important factor to consider with this study is the small sample size, especially considering the large number of individuals who did not complete the study. It should be noted that this study was not designed to be anything more than a pilot study intended to gauge the feasibility of implementing the Mastermind Cognitive Training program following the unexpected opportunity to work with adolescent tennis players. Future work must involve a larger trial of participants to replicate these findings to provide supporting evidence for the present interpretations. Second, the very real possibility exists that participation in the intervention generated a placebo effect through the novel use of a VR platform. Given the pervasive problem of placebo effects in such work [39], it is critical that future endeavors account for these effects in a carefully designed manner. Finally, it is unknown what aspect of the training program led to the improvement in UTR. The absence of cognitive control outcome measures to assess whether, and to what extent, this training program led to cognitive changes hinders a deeper understanding of what possibly drove the differential improvement in UTR. Future work is warranted to better understand the potential underlying mechanisms driving such change.

Conclusions

This exploratory proof-of-concept study investigated the potential benefits of a novel cognitive intervention platform in a sample of healthy adolescent tennis athletes. It represents the first step toward a number of data-driven studies demonstrating the utility and benefits of the use of this platform. These pilot findings suggest that this platform may be beneficial for adolescent tennis athletes looking to improve their sporting performance by supplementing their existing training protocols with a targeted cognitive training experience. However, many questions remain unanswered, and further work is needed to better understand the potential utility and mechanisms underlying potential effects of such a platform.

Acknowledgments

We would like to thank the participants from the Czech Lawn Tenis Klub, their coaches, and their families for being a part of this study. We thank C Gallen for her insights with the study design and suggestions regarding data preprocessing, C Walsh for providing suggestions for the manuscript, and P Wais and T Zanto for insights on figure creation and use. Note that no generative artificial intelligence was used in any portion of the manuscript writing.

Data Availability

The datasets generated during and/or analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization and design: JAA, MS, AC, and DF
Data collection: AC
Data processing and analysis: JAA and AC
Writing—original draft: JAA
Writing—review and editing: JAA, MS, AC, and DF

Conflicts of Interest

JAA is a paid consultant for MasterMind Sports, MS and DF are employees of MasterMind Sports. AC is a board member of MasterMind Sports.

Multimedia Appendix 1

XSL•FO
RenderX

Additional material.
[DOCX File , 22 KB - formative_v9i1e66979_app1.docx]

Multimedia Appendix 2
CONSORT-EHEALTH (V 1.6.1) checklist.
[PDF File (Adobe PDF File), 2950 KB - formative_v9i1e66979_app2.pdf]

References

1. Anguera JA, Gazzaley A. Video games, cognitive exercises, and the enhancement of cognitive abilities. Curr Opin Behav Sci 2015 Aug;4:160-165. [doi: 10.1016/j.cobeha.2015.06.002]
2. Mann D, Williams AM, Ward P, Janelle CM. Perceptual-cognitive expertise in sport: a meta-analysis. J Sport Exerc Psychol 2007 Aug;29(4):457-478. [doi: 10.1123/jsep.29.4.457] [Medline: 17968048]
3. Voss MW, Kramer AF, Basak C, Prakash RS, Roberts B. Are expert athletes 'expert' in the cognitive laboratory? A meta - analytic review of cognition and sport expertise. Appl Cogn Psychol 2009 Jun 17;24(6):812-826. [doi: 10.1002/acp.1588]
4. Vestberg T, Gustafson R, Maurex L, Ingvar M, Petrovic P. Executive functions predict the success of top-soccer players. PLoS One 2012 Apr 4;7(4):e34731 [FREE Full text] [doi: 10.1371/journal.pone.0034731] [Medline: 22496850]
5. Vestberg T, Reinebo G, Maurex L, Ingvar M, Petrovic P. Core executive functions are associated with success in young elite soccer players. PLoS One 2017 Feb 8;12(2):e0170845 [FREE Full text] [doi: 10.1371/journal.pone.0170845] [Medline: 28178738]
6. Mangine G, Hoffman JR, Wells AJ, Gonzalez AM, Rogowski JP, Townsend JR, et al. Visual tracking speed is related to basketball-specific measures of performance in NBA players. J Strength Cond Res 2014 Sep;28(9):2406-2414. [doi: 10.1519/JSC.0000000000000550] [Medline: 24875429]
7. Walton CC, Keegan RJ, Martin M, Hallock H. The potential role for cognitive training in sport: more research needed. Front Psychol 2018 Jul 3;9:1121 [FREE Full text] [doi: 10.3389/fpsyg.2018.01121] [Medline: 30018585]
8. Romeas T, Guldner A, Faubert J. 3D-Multiple Object Tracking training task improves passing decision-making accuracy in soccer players. Psychol Sport Exerc 2016 Jan;22:1-9. [doi: 10.1016/j.psychsport.2015.06.002]
9. Hirao T, Masaki H. Modulation of spatial attentional allocation by computer-based cognitive training during lacrosse shooting performance. Front Psychol 2017 Jan 4;8:2271 [FREE Full text] [doi: 10.3389/fpsyg.2017.02271] [Medline: 29354085]
10. Sala G, Gobet F. Cognitive training does not enhance general cognition. Trends Cogn Sci 2019 Jan;23(1):9-20. [doi: 10.1016/j.tics.2018.10.004] [Medline: 30471868]
11. Zelinski EM. Far transfer in cognitive training of older adults. Restor Neurol Neurosci 2009 Jan 01;27(5):455-471. [doi: 10.3233/rnn-2009-0495]
12. Simons DJ, Boot WR, Charness N, Gathercole SE, Chabris CF, Hambrick DZ, et al. Do "Brain-Training" programs work? Psychol Sci Public Interest 2016 Oct;17(3):103-186. [doi: 10.1177/1529100616661983] [Medline: 27697851]
13. Anguera JA, Brandes-Aitken AN, Antovich AD, Rolle CE, Desai SS, Marco EJ. A pilot study to determine the feasibility of enhancing cognitive abilities in children with sensory processing dysfunction. PLoS One 2017 Apr 5;12(4):e0172616 [FREE Full text] [doi: 10.1371/journal.pone.0172616] [Medline: 28380008]
14. Anguera JA, Gunning FM, Areán PA. Improving late life depression and cognitive control through the use of therapeutic video game technology: A proof-of-concept randomized trial. Depress Anxiety 2017 Jun 03;34(6):508-517 [FREE Full text] [doi: 10.1002/da.22588] [Medline: 28052513]
15. Arean PA, Hallgren KA, Jordan JT, Gazzaley A, Atkins DC, Heagerty PJ, et al. The use and effectiveness of mobile apps for depression: results from a fully remote clinical trial. J Med Internet Res 2016 Dec 20;18(12):e330 [FREE Full text] [doi: 10.2196/jmir.6482] [Medline: 27998876]
16. Ziegler DA, Simon AJ, Gallen CL, Skinner S, Janowich JR, Volponi JJ, et al. Closed-loop digital meditation improves sustained attention in young adults. Nat Hum Behav 2019 Jul 03;3(7):746-757 [FREE Full text] [doi: 10.1038/s41562-019-0611-9] [Medline: 31160812]
17. Mishra J, Gazzaley A. Closed-loop cognition: the next frontier arrives. Trends Cogn Sci 2015 May;19(5):242-243 [FREE Full text] [doi: 10.1016/j.tics.2015.03.008] [Medline: 25818939]
18. Mishra J, Anguera J, Gazzaley A. Video games for neuro-cognitive optimization. Neuron 2016 Apr 20;90(2):214-218 [FREE Full text] [doi: 10.1016/j.neuron.2016.04.010] [Medline: 27100194]
19. Mishra J, Anguera JA, Ziegler DA, Gazzaley A. A cognitive framework for understanding and improving interference resolution in the brain. Prog Brain Res 2013;207:351-377 [FREE Full text] [doi: 10.1016/B978-0-444-63327-9.00013-8] [Medline: 24309262]
20. Anguera JA, Volponi JJ, Simon AJ, Gallen CL, Rolle CE, Anguera-Singla R, et al. Integrated cognitive and physical fitness training enhances attention abilities in older adults. NPJ Aging 2022 Aug 30;8(1):12 [FREE Full text] [doi: 10.1038/s41514-022-00093-y] [Medline: 36042247]

XSL•FO
RenderX

21. Mishra J, Sagar R, Parveen S, Kumaran S, Modi K, Maric V, et al. Closed-loop digital meditation for neurocognitive and behavioral development in adolescents with childhood neglect. Transl Psychiatry 2020 May 18;10(1):153 [FREE Full text] [doi: 10.1038/s41398-020-0820-z] [Medline: 32424253]
22. Mishra J, de Villers-Sidani E, Merzenich M, Gazzaley A. Adaptive training diminishes distractibility in aging across species. Neuron 2014 Dec 03;84(5):1091-1103 [FREE Full text] [doi: 10.1016/j.neuron.2014.10.034] [Medline: 25467987]
23. Rolle CE, Anguera JA, Skinner SN, Voytek B, Gazzaley A. Enhancing spatial attention and working memory in younger and older adults. J Cogn Neurosci 2017 Sep;29(9):1483-1497 [FREE Full text] [doi: 10.1162/jocn_a_01159] [Medline: 28654361]
24. Wais PE, Arioli M, Anguera-Singla R, Gazzaley A. Virtual reality video game improves high-fidelity memory in older adults. Sci Rep 2021 Jan 28;11(1):2552 [FREE Full text] [doi: 10.1038/s41598-021-82109-3] [Medline: 33510315]
25. Gunning FM, Anguera JA, Victoria LW, Areán PA. A digital intervention targeting cognitive control network dysfunction in middle age and older adults with major depression. Transl Psychiatry 2021 May 04;11(1):269 [FREE Full text] [doi: 10.1038/s41398-021-01386-8] [Medline: 33947831]
26. Pratap A, Homiar A, Waninger L, Herd C, Suver C, Volponi J, et al. Real-world behavioral dataset from two fully remote smartphone-based randomized clinical trials for depression. Sci Data 2022 Aug 27;9(1):522 [FREE Full text] [doi: 10.1038/s41597-022-01633-7] [Medline: 36030226]
27. Pratap A, Renn BN, Volponi J, Mooney SD, Gazzaley A, Arean PA, et al. Using mobile apps to assess and treat depression in Hispanic and Latino populations: fully remote randomized clinical trial. J Med Internet Res 2018 Aug 09;20(8):e10130 [FREE Full text] [doi: 10.2196/10130] [Medline: 30093372]
28. Zanto TP, Giannakopoulou A, Gallen CL, Ostrand AE, Younger JW, Anguera-Singla R, et al. Digital rhythm training improves reading fluency in children. Dev Sci 2024 May 09;27(3):e13473 [FREE Full text] [doi: 10.1111/desc.13473] [Medline: 38193394]
29. Johnson V, Hsu W, Ostrand AE, Gazzaley A, Zanto TP. Multimodal sensory integration: Diminishing returns in rhythmic synchronization. J Exp Psychol Hum Percept Perform 2020 Oct;46(10):1077-1087. [doi: 10.1037/xhp0000833] [Medline: 32730071]
30. Nandi B, Ostrand A, Johnson V, Ford TJ, Gazzaley A, Zanto TP. Musical training facilitates exogenous temporal attention via delta phase entrainment within a sensorimotor network. J Neurosci 2023 Mar 28;43(18):3365-3378. [doi: 10.1523/jneurosci.0220-22.2023]
31. Zanto TP, Padgaonkar NT, Nourishad A, Gazzaley A. A tablet-based assessment of rhythmic ability. Front Psychol 2019 Nov 1;10:2471 [FREE Full text] [doi: 10.3389/fpsyg.2019.02471] [Medline: 31736843]
32. How UTR Rating Works. UTR Sports. 2023. URL: https://www.utrsports.net/blogs/news/how-utr-works [accessed 2024-09-25]
33. Anguera JA, Boccanfuso J, Rintoul JL, Al-Hashimi O, Faraji F, Janowich J, et al. Video game training enhances cognitive control in older adults. Nature 2013 Sep 05;501(7465):97-101 [FREE Full text] [doi: 10.1038/nature12486] [Medline: 24005416]
34. Hsu W, Rowles W, Anguera JA, Zhao C, Anderson A, Alexander A, et al. Application of an adaptive, digital, game-based approach for cognitive assessment in multiple sclerosis: observational study. J Med Internet Res 2021 Jan 20;23(1):e24356 [FREE Full text] [doi: 10.2196/24356] [Medline: 33470940]
35. Wilkins L, Gray R. Effects of stroboscopic visual training on visual attention, motion perception, and catching performance. Percept Mot Skills 2015 Aug 01;121(1):57-79. [doi: 10.2466/22.25.pms.121c11x0]
36. Deveau J, Ozer DJ, Seitz AR. Improved vision and on-field performance in baseball through perceptual learning. Curr Biol 2014 Feb 17;24(4):R146-R147 [FREE Full text] [doi: 10.1016/j.cub.2014.01.004] [Medline: 24556432]
37. Clark JF, Ellis JK, Bench J, Khoury J, Graman P. High-performance vision training improves batting statistics for University of Cincinnati baseball players. PLoS One 2012 Jan 19;7(1):e29109 [FREE Full text] [doi: 10.1371/journal.pone.0029109] [Medline: 22276103]
38. Rezaee M, Ghasemi A, Momeni M. Visual and athletic skills training enhance sport performance. Eur J Exp Biol 2012;2(6):2243-2250 [FREE Full text]
39. Boot WR, Simons DJ, Stothart C, Stutts C. The pervasive problem with placebos in psychology: why active control groups are not sufficient to rule out placebo effects. Perspect Psychol Sci 2013 Jul 09;8(4):445-454. [doi: 10.1177/1745691613491271] [Medline: 26173122]

Abbreviations

ANCOVA: analysis of covariance
CONSORT: Consolidated Standards of Reporting Trials
UTR: Universal Tennis Rating
VR: virtual reality

XSL•FO
RenderX

Edited by A Mavragani; submitted 02.10.24; peer-reviewed by OR Runswick; comments to author 31.10.24; revised version received 08.01.25; accepted 16.01.25; published 07.02.25.

Please cite as:
Anguera JA, Choudhry A, Seaman M, Fedele D
Assessing the Impact of a Virtual Reality Cognitive Intervention on Tennis Performance in Junior Tennis Players: Pilot Study
JMIR Form Res 2025;9:e66979
URL: https://formative.jmir.org/2025/1/e66979
doi:10.2196/66979
PMID:39918854

©Joaquin A Anguera, Aleem Choudhry, Michael Seaman, Dominick Fedele. Originally published in JMIR Formative Research (https://formative.jmir.org), 07.02.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Short Paper

Association Between Alcohol Consumption and Psoriasis: Exploratory Analysis of Crowdsourced Web Search Data in Sweden

Anna Katharina Schober[1,2], MSc; Klas Nordlind[1], MD, PhD; Alexander Zink[2], MBA, MPH, MD, PhD

[1]Department of Medicine Solna, Division of Dermatology and Venereology, Karolinska Institutet, Stockholm, Sweden

[2]Department of Dermatology and Allergy, Technical University of Munich, Munich, Germany

Corresponding Author:
Anna Katharina Schober, MSc
Department of Dermatology and Allergy
Technical University of Munich
Biedersteiner Str 29
Munich, 80802
Germany
Phone: 49 41403184
Email: anna.schober@tum.de

Abstract

Background: The nature of the relationship between psoriasis and alcohol consumption has been the topic of discussion for many years. Some studies have found that a higher intake of alcohol may be associated with a more severe manifestation of the disease. At the same time, patients with psoriasis often demonstrate elevated levels of alcohol consumption. It has not yet been fully established whether alcohol abuse serves as a trigger for psoriasis or if patients with psoriasis are simultaneously more prone to alcohol abuse.

Objective: The objective of this study was to employ Google Trends as a tool for crowdsourcing data on a national level to explore the relationship between psoriasis and the consumption of alcohol in Sweden.

Methods: This study examines crowdsourced web search data related to psoriasis and other skin disease–related search terms (such as utslag [rash]) as well as search interest in different types of alcohol. The analysis covers a 5-year period from 2018 to 2023 in Sweden, focusing on search behavior and correlations across this period.

Results: The search behavior regarding psoriasis and alcohol-related search terms showed seasonal variations throughout the year. The relative search volume for psoriasis peaked in February, while alcohol-related searches, particularly Systembolaget and vodka, peaked in December and June. Our statistical analysis revealed relationships between the search interest regarding psoriasis and terms related to alcohol consumption, with disparities between different types of alcohol. The term “psoriasis” was negatively correlated with “Systembolaget” (r=–0.210), “vitt vin” (r=–0.224), and “vodka” (r=–0.220) (all P<.001), while the term “utslag” showed positive correlations with these same alcohol-related terms (r=0.278-0.347; P<.001).

Conclusions: Crowdsourced data can offer valuable insights into population-level behavior. The observed negative correlations between psoriasis and alcohol-related searches suggest complex interactions, possibly reflecting reduced disease awareness or care during periods of higher alcohol consumption. The direction and strength of the correlations with psoriasis were not consistent across the different types of alcohol investigated in this study, which poses the question whether the relationship might be influenced by the type of beverage consumed. Further research is warranted to explore underlying mechanisms and validate these findings in clinical populations.

(JMIR Form Res 2025;9:e71992) doi:10.2196/71992

KEYWORDS

addiction; alcohol; crowdsourcing; psoriasis; skin disease; epidemiology; infodemiology; alcohol abuse; web search data; internet

XSL•FO
RenderX

Introduction

Around 3% of Europeans experience psoriasis, a chronic inflammatory skin disease characterized by papulosquamous plaques, often accompanied by itching [1]. As psoriasis frequently has a significant negative impact on patients' quality of life and may lead to stigmatization and psychosocial impairment, mental illnesses represent a concern among such patients [2,3], and they tend to have relatively highly addictive behaviors, especially toward alcohol [4]. The relationship between psoriasis and alcohol consumption has been discussed for many years. Higher intake of alcohol may be associated with a more severe manifestation of the disease, which is unusual compared to other skin diseases [4]. Concurrently, patients with psoriasis often demonstrate elevated levels of alcohol consumption [4,5].

To approximate disease patterns and health care–seeking behaviors in the general population, crowdsourced data such as web search data have increasingly been employed in chronic and acute skin diseases for various approaches, including disease monitoring and the assessment of unmet needs and national disease [6,7]. With the data being publicly available and derived from the information-seeking behavior of large proportions of the general population [7], crowdsourcing represents a powerful tool in public health research.

In Sweden, the exclusive right to sell alcoholic beverages stronger than 3.5% alcohol by volume is held by Systembolaget—Sweden's government-owned chain of liquor stores, which holds the state monopoly [8]. Thus, engagement with Systembolaget represents a unique opportunity to capture drinking behaviors at a population level.

This study aims to use web data related to skin diseases and alcohol consumption to further examine the interplay between psoriasis and alcohol in Sweden by crowdsourcing web search data at a national level.

Methods

Google Trends was accessed in Sweden to analyze the public interest in search terms related to skin diseases and alcohol. This tool provides a relative search volume (RSV) metric, which quantifies the search interest for a given input within a defined region and timespan. RSV values range from 0 to 100, with 100 indicating the highest search volume during the specified period. The user input may be recognized as terms or topic, with terms referring to the exact words entered. Search topics are broader categories that Google identifies when it detects related phrases or concepts associated with a popular query. The latter can also be compared across different countries, as its identification is irrespective of language. We used only search terms, as they offer more transparency and control over the specific query content included in the analysis. Since this study was based on publicly available terms, institutional review board approval was not needed, and informed consent was not applicable.

The data collection spanned a 5-year period from June 2018 to June 2023 in Sweden. In this study, RSV data were collected for the search terms "psoriasis" and eczematous skin diseases (neurodermatit [atopic dermatitis], eksem [eczema], and utslag [rash]) as control. Neurodermatit and eksem represent the most common chronic inflammatory skin disease aside from psoriasis [9], providing a clinically relevant comparator. In contrast, utslag serves as a generic umbrella term encompassing acute and chronic skin eruptions of diverse etiologies, thus functioning as a broader population-level reference point. To capture general interest in alcohol purchase rather than preference for a specific beverage, the search term Systembolaget was included. Additionally, the selection of specific alcohol-related search terms was included to act as a proxy for patterns of alcohol interest, and thus, indirectly, consumption. These were vitt vin (white wine), rött vin (red wine), öl (beer), and vodka, chosen to represent both common beverage categories and a mix of lower-alcohol and higher-alcohol content drinks. All included search terms are listed in Table 1.

The statistical analysis encompassed the Wilcoxon and Kruskal-Wallis tests as well as Spearman correlations to examine the relationship between the search volumes of the selected terms. The analysis was conducted using the R software (version 4.3.0; R Foundation for Statistical Computing).

Table 1. Search terms.

	Minimum monthly RSV[a]	Maximum monthly RSV	Range
Utslag (rash)	56.3	83.5	27.2
Vitt vin (white wine)	25.8	54	28.2
Vodka (vodka)	29.5	62.5	33
Psoriasis (psoriasis)	36.3	69.8	33.5
Rött vin (red wine)	26.4	61	34.6
Öl (beer)	50.6	89.5	38.9
Eksem (eczema)	45.3	86.3	41
Systembolaget	19.3	65.5	46.2

[a]RSV: relative search volume.

XSL•FO
RenderX

Results

RSV examination for the selected terms revealed noteworthy variations in search trends, both within and between years over the observed 5-year period. Notably, the yearly interest in the search terms psoriasis, eksem, utslag, Systembolaget, vitt vin, and vodka exhibited inconsistency. Statistical analysis confirmed significant differences in yearly RSVs among the 4 complete years included in the study (2019-2022) for all terms (P<.05). Similarly, the comparison of mean volumes by month yielded significant differences (P<.001) in mean RSVs throughout the year (Table 1). Specifically, the interest in psoriasis reached its zenith in February, followed by March and April. Likewise, the peak interest in eczema occurred in February, while the highest number of queries for rash was recorded in June, trailed by July and May. Regarding alcohol-related terms, the highest RSV was observed for Systembolaget and vodka in December succeeded by vitt vin in June or July. Notably, the interest in rash (utslag) exhibited the smallest overall range, with RSVs ranging from 56.3 to 83.5, while the interest in Systembolaget displayed the highest range, with RSVs ranging from 19.3 to 65.5.

Several associations were revealed in the correlation analysis. Negative correlations were obtained between psoriasis and Systembolaget (r=–0.21; P=.001; Figure 1), vitt vin (r=–0.22; P<.001), and vodka (r=–0.22; P<.001); the correlations with beer and red wine were nonsignificant (r<–0.1; P>.05).

In contrast, utslag was positively correlated with alcohol-related search terms such as Systembolaget (r=0.28), vitt vin (r=0.33), and vodka (r=0.35). These correlations were highly significant (P<.001). For eksem and neurodermatit, there was no evidence for correlations with the alcohol-related terms (r<0.1; P>.05).

Figure 1. Relative search volume for Systembolaget and psoriasis (June 2018 to June 2023). Dark red dotted lines indicate selected time points where opposing patterns in search interest are most apparent, serving as a visual guide to illustrate periods of inverse search behavior.

Discussion

RSV analysis for alcohol-related and skin disease–related search terms in Sweden through crowdsourcing provides a unique perspective into the relationship between alcohol and psoriasis on a population-based level. The differences in RSVs between years and seasons underscore the dynamic nature of public interest in these topics [10]. The distinct seasonal variations in interest, as evidenced by the peak months for psoriasis, eczema, rash, and alcohol-related terms, might indicate potential links between these factors and specific periods of the year [11,12].

The negative correlations between psoriasis and alcohol-related terms such as Systembolaget, vitt vin, and vodka raise questions about the relationship between alcohol consumption and the development or exacerbation of psoriasis. One possible explanation may be the temporary neglect of the disease during elevated alcohol intake, which may lead to impaired adherence to therapy or promote the development of long-term drinking habits [13-15]. Conversely, an individual with an existing addiction might not be affected by the seasonal variability in drinking habits. Other seasonal or environmental factors should also be considered, such as weather, which may influence both alcohol consumption and psoriasis severity independently, potentially confounding the observed associations. For example, reduced sunlight exposure in winter exacerbates psoriasis [16], which may contribute more directly to the observed peaks in search volume than behavioral changes alone. Other contributors to disease severity, such as dietary triggers or stress, could not be captured but may shape public interest over time.

Interestingly, the direction and strength of the correlations were not consistent across different types of alcohol. Regarding the

XSL•FO
RenderX

effects of different alcohol types, the *Nordic Quality of Life Study* [17] found a negative correlation between psoriasis and the consumption of wine but not beer. To our knowledge, other than that, there is limited evidence on the effect of different types of alcoholic beverages on psoriasis, as research has focused on total alcohol consumption [4].

However, these correlations must be interpreted cautiously, as they do not establish a causal relationship between psoriasis and alcohol consumption. The underlying mechanisms and potential confounding factors that may contribute to these associations must be elucidated. Additionally, our reliance on web search data introduces inherent limitations such as the inability to ascertain the specific motivations and demographics of the individuals conducting the searches. Search behavior may be influenced by external factors such as media coverage, public health campaigns, or temporary events, which are difficult to control and may introduce bias. Further, there may be temporal or geographical inaccuracies in the data aggregation, which can affect the stability of the results. Due to the relative scaling of search volumes in Google Trends, direct comparisons between different search terms are less straightforward. Accordingly, we have limited cross-term comparisons and focused primarily on within-term trends such as seasonal patterns and the range of RSVs over time to ensure cautious interpretation.

Despite this, we underscore the value of crowdsourced web search data as a tool for investigating public interest and potentially identifying trends and patterns related to health conditions and behaviors. This approach opens new avenues for exploring the connection between psoriasis and alcohol consumption, complementing traditional research methods.

In conclusion, the analysis of search trends in Sweden provides intriguing insights into the relationship between psoriasis and alcohol consumption. The observed correlations between search volumes for alcohol-related terms and psoriasis-related terms warrant further investigation to enhance our understanding of the potential links between these factors. Future studies employing a multidimensional approach, including clinical data and patient-reported outcomes, can provide a more comprehensive understanding of the complex interplay between different alcohol sorts, their consumption, psoriasis, and associated inflammatory processes.

Data Availability

The websearch data used in this study are publicly available on the Google Trends website [18]. The processed datasets generated during the current study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization and project administration: AZ, KN
Data preparation and analysis: AKS, AZ
Writing – original draft: AKS, AZ
Supervision, review, and editing: KN, AZ
All authors read and approved the manuscript.

Conflicts of Interest

None declared.

References

1. Griffiths CEM, Armstrong AW, Gudjonsson JE, Barker JNWN. Psoriasis. The Lancet 2021 Apr;397(10281):1301-1315. [doi: 10.1016/s0140-6736(20)32549-6]
2. Van Beugen S, Schut C, Kupfer J, Bewley AP, Finlay AY, Gieler U, European Society for DermatologyPsychiatry (ESDaP) Study collaborators. Perceived stigmatization among dermatological outpatients compared with controls: an observational multicentre study in 17 European countries. Acta Derm Venereol 2023 Jun 22;103:adv6485 [FREE Full text] [doi: 10.2340/actadv.v103.6485] [Medline: 37345973]
3. Mahé E, Tubach F, Jullien D, Tran D, Paul C, Beylot-Barry M, PSOBIOTEQ Registry investigators. Impact of childhood onset psoriasis on addictive behaviours, socioeconomic and educational data in adulthood. Acta Derm Venereol 2022 Jun 09;102:adv00733 [FREE Full text] [doi: 10.2340/actadv.v102.2484] [Medline: 35506359]
4. Adamzik K, McAleer MA, Kirby B. Alcohol and psoriasis: sobering thoughts. Clin Exp Dermatol 2013 Dec;38(8):819-822. [doi: 10.1111/ced.12013] [Medline: 24252076]
5. Svanström C, Lonne-Rahm S, Nordlind K. Psoriasis and alcohol. Psoriasis (Auckl) 2019;9:75-79 [FREE Full text] [doi: 10.2147/PTT.S164104] [Medline: 31687362]
6. Tizek L, Schielein MC, Tizek L, Zink A. [Atopic dermatitis-identifying needs in the German population by internet search queries]. Hautarzt 2022 Jun;73(6):475-484 [FREE Full text] [doi: 10.1007/s00105-022-04974-x] [Medline: 35294564]
7. Tizek L, Wecker H, Schneider S, Johansson EK, Girmay Y, Zink A. Eczema-related web search data in sweden: investigating search patterns and the influence of weather. Acta Derm Venereol 2022 Nov 07;102:adv00810 [FREE Full text] [doi: 10.2340/actadv.v102.2937] [Medline: 36065747]

XSL•FO
RenderX

8. Centrum för Näringslivshistoria. Systembolaget Historia (2024). URL: https://systembolagethistoria.se/teman/ursprunget/1955-systembolaget-grundas/ [accessed 2025-08-25]
9. Pezzolo E, Naldi L. Epidemiology of major chronic inflammatory immune-related skin diseases in 2019. Expert Review of Clinical Immunology 2020 Jan 28;16(2):155-166. [doi: 10.1080/1744666x.2020.1719833]
10. Wallnöfer F, Erbas M, Tizek L, Schuster B, Wecker H, Biedermann T, et al. Leveraging web search data to explore public interest in psoriasis in Germany. JEADV Clinical Practice 2022 Jun 16;1(3):254-263. [doi: 10.1002/jvc2.37]
11. Park BS, Youn JI. Factors influencing psoriasis: an analysis based upon the extent of involvement and clinical type. J Dermatol 1998 Feb;25(2):97-102. [doi: 10.1111/j.1346-8138.1998.tb02357.x] [Medline: 9563276]
12. Brenaut E, Horreau C, Pouplard C, Barnetche T, Paul C, Richard M, et al. Alcohol consumption and psoriasis: a systematic literature review. J Eur Acad Dermatol Venereol 2013 Aug;27 Suppl 3:30-35. [doi: 10.1111/jdv.12164] [Medline: 23845150]
13. Krantz, Carrero JJ, Yang Y, Schalin L, Lysell J, Mazhar F. Psoriasis/psoriatic arthritis patients' long-term treatment patterns and adherence to systemic treatments monitoring recommendations. Acta Derm Venereol 2023 Jun 15;103:adv6505 [FREE Full text] [doi: 10.2340/actadv.v103.6505] [Medline: 37318074]
14. Zou L, Lonne-Rahm S, Helander A, Stokkeland K, Franck J, Nordlind K. Alcohol intake measured by phosphatidylethanol in blood and the lifetime drinking history interview are correlated with the extent of psoriasis. Dermatology 2015;230(4):375-380. [doi: 10.1159/000380818] [Medline: 25823412]
15. Schielein M, Tizek L, Knobloch L, Maaßen D, Biedermann T, Zink A. Psoriasis and addiction: assessing mental health based on a cross-sectional study in Germany. Eur J Dermatol 2021 Dec 01;31(6):722-729. [doi: 10.1684/ejd.2021.4146] [Medline: 34726161]
16. Yang J, Li G, Yue L, Dang E, Qiao P. The impacts of seasonal factors on psoriasis. Exp Dermatol 2025 Mar;34(3):e70078. [doi: 10.1111/exd.70078] [Medline: 40103264]
17. Davidsson S, Blomqvist K, Molin L, Mørk C, Sigurgeirsson B, Zachariae H, et al. Lifestyle of Nordic people with psoriasis. Int J Dermatol 2005 May;44(5):378-383. [doi: 10.1111/j.1365-4632.2005.01925.x] [Medline: 15869535]
18. Google Trends. URL: https://trends.google.com/ [accessed 2025-08-27]

Abbreviations

RSV: relative search volume

Edited by A Mavragani; submitted 31.01.25; peer-reviewed by M Walker, MW Huang, O Abasilim; comments to author 13.04.25; revised version received 25.07.25; accepted 21.08.25; published 27.08.25.

Please cite as:
Schober AK, Nordlind K, Zink A
Association Between Alcohol Consumption and Psoriasis: Exploratory Analysis of Crowdsourced Web Search Data in Sweden
JMIR Form Res 2025;9:e71992
URL: https://formative.jmir.org/2025/1/e71992
doi:10.2196/71992
PMID:

©Anna Katharina Schober, Klas Nordlind, Alexander Zink. Originally published in JMIR Formative Research (https://formative.jmir.org), 27.08.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org

XSL•FO
RenderX